ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE AND TECHNOLOGY

ELTRUN: THE E-BUSINESS RESEARCH CENTER

# Ph.D. Thesis

## Data-driven Innovation in Shopper Marketing: A Business Analytics Approach for Visit Segmentation in the Retail Industry

*by*

**Anastasia Griva**

*A thesis submitted for the degree of Doctor of Philosophy (Ph. D.)*

*Thesis supervisors:*

Assoc. Prof. Katerina Pramatari
Prof. Georgios Doukidis
Lecturer, Dr. Dimitris Papakiriakopoulos

**Athens,**

**June 2019**

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ELTRUN: ΕΡΓΑΣΤΗΡΙΟ ΗΛΕΚΤΡΟΝΙΚΟΥ ΕΜΠΟΡΙΟΥ ΚΑΙ ΕΠΙΧΕΙΡΕΙΝ

# Διδακτορική Διατριβή

## Δεδομενοκεντρική καινοτομία στο Μάρκετινγκ: Μια προσέγγιση βασισμένη σε τεχνικές επιχειρηματικής αναλυτικής για την κατάτμηση των επισκέψεων στο λιανεμπόριο

*της*

**Αναστασίας Γρίβα**

*Η παρούσα διδακτορική διατριβή υποβλήθηκε για την απονομή του τίτλου του Διδάκτορα*

*Επιβλέποντες:*

Αναπλ. Καθ. Κατερίνα Πραματάρη
Καθ. Γεώργιος Δουκίδης
Λέκτορας, Δρ. Δημήτρης Παπακυριακόπουλος

**Αθήνα**
**Ιούνιος 2019**

*Πέντε χρόνια πέρασαν,*
*μ' έρευνα κ' αναλύσεις.*
*Ποιο είναι το ερώτημα;*
*Τι άραγε θα λύσεις;*

*Φορές αναρωτήθηκα,*
*εάν άξιζε τον κόπο,*
*και η Κατερίνα έλεγε*
*«Μη σκας θα βρεις τον τρόπο»*

*Μια start-up και μια έρευνα,*
*τι άραγε μου ταιριάζει;*
*«Θα συνδυάσεις και τα δυο...»,*
*έλεγε, «...μη σε νοιάζει!».*

*Σε όλους όσους προσπαθούν να «συνδυάσουν και τα δυο»*

*S*

# PUBLICATIONS

## Journal Papers

1. Griva, A., Bardaki, C., Pramatari K., Papakiriakopoulos, D., (2018), Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data, *Expert Systems with Applications, 100,* 1–16.
2. Triantafyllou S., Koutsokera L., Stavrou V., Griva A., (2018), Enrich customer experience and support decision making using IoT technologies in a grocery retail store, *Astrolavos*, *Scientific Journal of New Technologies (Hellenic Mathematical Society),* 28, 60-71.
3. Kalaidopoulou, K., Griva, A. (2016), Extract purchasing patterns for a focal product category using sales data: The case of skincare products, *Astrolavos*, *Scientific Journal of New Technologies (Hellenic Mathematical Society)*, 25, 3-18, (in Greek).

## Papers in Edited Volumes

4. Griva, A., Pramatari, K. (2019), Applying innovative business analytics approaches to analyze shopper behavior in retail, In Doukidis G. (Ed.) *The Digital Future,* Sideris I. (in Greek, in Press).

## Refereed Conference Papers

5. Griva, A., Bardaki, C., Pramatari, K., Doukidis G. (2018), Design of shopper segmentation systems in retail. Evidence from 2 heterogeneous retail cases, *Proceedings of pre-International Conference on Information Systems (pre-ICIS 2018), Special Interest Group on Decision Support and Analytics (SIGDSA)*

*symposium on Decision Analytics Connecting People, Data & Things,* December 12-16, San Francisco.

6. Kalaidopoulou, K., Triantafyllou, S., Griva, A., Pramatari, K. (2017), Identifying Customer Satisfaction Patterns via Data Mining: The Case of Greek E-Shops, *Proceedings of 11$^{th}$ Mediterranean Conference on Information Systems (MCIS 2017)*, September 4-5, Genoa, Italy.
7. Griva, A., Bardaki, C., Pramatari, K. (2016), Customer Visit Segmentation Using Market Basket Data, *Proceedings of pre-International Conference on Information Systems (pre-ICIS 2016), Special Interest Group on Decision Support and Analytics (SIGDSA)/ IFIP WG8.3 symposium: Innovations in Data Analytics*, 11-14 December 2016, Dublin, Ireland (Best Paper Award).
8. *Sarantopoulos, P., Griva, A., Papakiriakopoulos, D., Giovanis A. (2016), Identifying complex and multi-step customer projects: A graph mining market basket approach, Proceeding of AMBS Big Data Forum, 28 September 2016, Manchester, United Kingdom*.
9. Griva, A., Bardaki, C., Pramatari, K., Doukidis G. (2016), Mapping moving object events into a network of object flows to support decisions, *Proceedings of 24$^{th}$ European Conference on Information Systems (ECIS 2016)*, 12-15 June 2016, Istanbul, Turkey.
10. Griva, A., Bardaki, C., Pramatari, K., (2015), RFID-Enabled Visualization of Product Flows: A Data Analytics Approach, *Proceedings of 9$^{th}$ Mediterranean Conference on Information Systems (MCIS 2015)*, October 3-5, Samos, Greece.
11. Griva, A., Bardaki, C., Sarantopoulos, P., Papakiriakopoulos, D. (2014), A Data Mining-based Framework to Identify Shopping Missions, *Proceedings of 8$^{th}$ Mediterranean Conference on Information Systems (MCIS 2014)*, September 3-5, Verona, Italy.

**Papers in Non-Refereed Conferences and Workshops**

12. Batziakoudi K., Griva A., Stavrou V., Pramatari K., (2019), The value of collaborative analytics in contemporary retail. Evidence from a real-life case study. *Proceedings of 16th Student Conference of Management Science and Technology*, 14 May 2019, Athens, Greece. (Abstract only).
13. Triantafyllou, S., Koutsokera, L., Stavrou V., Griva, A., (2017), Enhance shopping experience and support decision making leveraging BLE beacons in a grocery retail store, *Proceedings of 14th Student Conference of Management Science and Technology*, 27 April 2017, Athens, Greece.
14. Griva, A., Pramatari, K., Bardaki, C., Doukidis G. (2016), Segmentation of Shopper Visits based on Shopper Interaction data in Different Retail Contexts, *1st EURO Working Group on Retail Operations*, 3-5 June 2016, Munich, Germany. (Abstract only).
15. Kalaidopoulou, K., Griva, A., Sarantopoulos, P., (2016), How shoppers buy a specific product category in retail stores? The case of face care products, *Proceedings of 13th Student Conference of Management Science and Technology*, 12 May 2016, Athens, Greece.
16. Kalaidopoulou, K., Koutsokera L., Stavrou V., Griva, A., (2016), Investigating shopping visits patterns across different store types: The case of a grocery retail chain, *Proceedings of 13th Student Conference of Management Science and Technology*, 12 May 2016, Athens, Greece. (Abstract only)
17. Griva, A., Bardaki, C., Papakiriakopoulos, D., (2015), Shopping Goals Detection: Mining POS data from a Do-it-Yourself (DIY) retailer, *Proceedings of 12th Student Conference of Management Science and Technology*, 14 May 2015, Athens, Greece.

18. Kalaidopoulou, K., Kanellopoulos I., Griva, A., (2015), Identification of Customer Segments via Data Mining, *Proceedings of $12^{th}$ Student Conference of Management Science and Technology,* 14 May 2015, Athens, Greece.
19. Griva, A., Bardaki, C., Pramatari, K., (2014), A Data Mining-based Framework to Identify Shoppers' Missions, *Proceedings of $11^{th}$ Student Conference of Management Science and Technology*, 15 May 2014, Athens, Greece.

### Doctoral Consortium

1. $10^{th}$ Mediterranean Conference on Information Systems (MCIS 2016), September 4-6, 2016, Paphos, Cyprus.

### Working Papers and Work under Review

1. Griva, A., Bardaki, C., Pramatari, K., Doukidis G. "Designing Shopper Segmentation Systems in Contemporary Retail. Evidence from three Heterogenous Retail Cases" (Submitted to *Decision Support Systems)*.
2. Griva, A., Theotokis A., Pramatari, K. "Transforming Retail: Moving from Category Management to Shopping Mission Management and Analytics".
3. Stavrou, V., Bardaki, C., Griva A., Pramatari, K. "Deploying a Retail Location-based Coupon Recommendation Application: Guidelines and Lessons Learnt". (Submitted to *ICIS 2019*).
2. Papakiriakopoulos D., Griva, A., Stavrou V. "Graph mining to extract shopping missions: A case from DIY industry".

### Research-related Awards

- 2017: $1^{st}$ Award among 120 teams at the research stream of ennovation contest, representing ShopMind: a retail analytics platform.

- 2017: Self Service Excellence Awards, Best Practices in the FMCG - for the project AB ShopMate – Beacon enabled store, with cooperation to AB Vassilopoulos, Ahold Delhaize Group.
- 2016: Best Paper Award, Pre-ICIS SIGDSA/IFIP WG8.3 Symposium Innovations in Data Analytics, Dublin, Ireland for the paper titled "Customer Visit Segmentation using market basket data".

# ACKNOWLEDGEMENTS

Για καιρό άφηνα αυτή τη σελίδα κενή. Αφενός, γιατί θεωρούσα πιο σημαντικό να εκφράσω τις ευχαριστίες μου λεκτικά. Αφετέρου, γιατί δεν ήξερα τι να πρώτο-γράψω, δε χωρά άλλωστε μια σελίδα για να ευχαριστήσεις τόσους ανθρώπους! Μέσα στη σχεδόν πενταετή πορεία μου πολλοί ήταν αυτοί που στάθηκαν δίπλα μου, με βοήθησαν, συμβούλευσαν, άκουσαν τις ανησυχίες μου και με στήριξαν στις όποιες αποτυχίες. Από τους επιβλέποντες καθηγητές μου και τους συνεργάτες μου, μέχρι τους φίλους και την οικογένειά μου.

Αρχικά θα εστιάσω στους επιβλέποντες καθηγητές μου. Θα ήθελα από καρδιάς να ευχαριστήσω την Αναπλ. Καθ. Κατερίνα Πραματάρη, η οποία μου στάθηκε, και στέκεται σε όλους τους συνεργάτες της, σα να είναι οικογένειά της. Μας συμβουλεύει τόσο στα ακαδημαϊκά και τα επαγγελματικά, όσο και στα προσωπικά ζητήματα. Είναι ένας άνθρωπος που λειτουργεί με αμέριστο αλτρουισμό, παρακινεί τους γύρω της και τους βοηθά να κάνουν πραγματικότητα τα όνειρά τους, προσφέροντας παράλληλα λίγο από την αισιοδοξία της. Σε όλα αυτά τα χρόνια, αποτέλεσε και αποτελεί ένα πρότυπο για εμένα.

Θα ήθελα επίσης, να εκφράσω τις ιδιαίτερες ευχαριστίες μου στον Καθ. Γεώργιο Δουκίδη. Είναι αξιοσημείωτη η όρεξη που έδειξε κατά την εκτέλεση της διδακτορικής μου διατριβής. Με τα σχόλια και τις συμβουλές του με βοήθησε να δω τα πράγματα έξω από το «ακαδημαϊκό κουτί», και να καταλάβω πως αυτά εφαρμόζονται και δίνουν λύσεις σε πραγματικά επιχειρηματικά προβλήματα. Εδώ δε θα μπορούσα να μη σχολιάσω το γεγονός ότι έχει ιδρύσει ένα ερευνητικό κέντρο πρότυπο για τα Ελληνικά δεδομένα, μέσου του οποίου μας δίνεται η δυνατότητα να πραγματοποιούμε εφαρμοσμένη έρευνα.

Παράλληλα, θα ήθελα να ευχαριστήσω τον Δρ. Δημήτρη Παπακυριακόπουλο. Είναι ένας άνθρωπος που με ενέπνευσε και με τα σχόλια με τα οποία με βομβάρδιζε σε κάθε συνάντησή μας, συνέβαλε σημαντικά στην εξέλιξη της έρευνάς μου.

Θα ήθελα επίσης να ευχαριστήσω την Δρ. Κλεοπάτρα Μπαρδάκη, που παρότι δεν είναι στην επταμελή μου, μου στάθηκε σαν επιβλέπουσα. Κλείνοντας, ευχαριστώ όλους τους συνεργάτες μου στο ELTRUN, αλλά και ιδιαίτερα τους Βασίλη Σταύρου και Λέων Γαβαλά. Χωρίς αυτή την ομάδα αυτή η έρευνα δε θα είχε καταφέρει να διαγράψει μια σημαντική πορεία και στον επιχειρηματικό χώρο.

Έπειτα, θα ήθελα να ευχαριστήσω τις εταιρίες και τους ανθρώπους από τον επιχειρηματικό κόσμο, οι οποίες μας εμπιστεύτηκαν τα δεδομένα τους για τους σκοπούς της έρευνας. Αυτές είναι οι: AB Vassilopoulos - Delhaize Group, Carrefour-Marinopoulos, Diamantis Masoutis S.A., MyMarket, Nestle Hellas, L'Oréal, Johnson & Johnson, Adler Modemärkte, Diffusione Tessile, Plaisio, the Fortune 500 Specialty retailer. Τέλος, ευχαριστώ το Ινστιτούτο Έρευνας Λιανεμπορίου Καταναλωτικών Αγαθών (ΙΕΛΚΑ) που μας βοήθησε σημαντικά στην έρευνα για τους αγοραστές.

Δε θα μπορούσα να παραλείψω τις πολύ καλές μου φίλες Αγγελική, Άννα, Ελένη και Κωνσταντίνα, που μου στάθηκαν σε όλο αυτό το ταξίδι και μου θύμιζαν πόσο σημαντική και όμορφή είναι η ζωή έξω από αυτό. Κλείνοντας, Δώρα, Αφροδίτη και Αντρέα σας ευχαριστώ επίσης από καρδιάς! Τέλος, θα ήθελα από καρδιάς να ευχαριστήσω για την αμέριστη στήριξη, συμπαράσταση και ενθάρρυνση που μου προσέφεραν όλα αυτά τα χρόνια τους γονείς μου και την αδερφή μου.

Αναστασία

# ΕΠΙΤΕΛΙΚΗ ΣΥΝΟΨΗ

Οι αυξανόμενες δυνατότητες των εργαλείων και τεχνικών επιχειρηματικής αναλυτικής και η δεδομενοκεντρική λήψη των αποφάσεων έχουν πλέον μπει στην ατζέντα πολλών επιχειρήσεων. Οι έμποροι λιανικής πώλησης συλλέγουν και αποθηκεύουν καθημερινά πολλά και διαφορετικά είδη δεδομένων σχετικά με τους πελάτες τους. Μέσα σε αυτό το περιβάλλον, μια από τις μεγαλύτερες φιλοδοξίες τους είναι να βρεθούν καινοτόμοι τρόποι αξιοποίησης των συλλεγόμενων δεδομένων.

Εάν λοιπόν, όλα αυτά τα δεδομένα αναλυθούν σωστά, μπορούν να βοηθήσουν τους λιανέμπορους να έρθουν πιο κοντά με τους πελάτες τους, να εντοπίσουν τα διάφορα τμήματα των πελατών τους (shopper segments), να κατανοήσουν τη συμπεριφορά τους και να καθοδηγήσουν τόσο τις μελλοντικές στρατηγικές, όσο και τις καθημερινές δραστηριότητές τους (Sharma, Mithas and Kankanhalli, 2014). Η κατάτμηση των αγοραστών (shopper segmentation) είναι μια παραδοσιακή και θεμελιώδης έννοια στο μάρκετινγκ (Wilkie, 1978). Οι αγοραστές σε κάθε τμήμα έχουν τα ίδια ή παρόμοια χαρακτηριστικά και μπορούν να ικανοποιηθούν από παρόμοια μίγματα μάρκετινγκ (Hong and Kim, 2012).

Ωστόσο, στις μέρες μας ο σύγχρονος αγοραστής έχει αλλάξει, αναζητά συνεχώς νέες, βελτιωμένες εμπειρίες στα λιανεμπορικά καταστήματα. Πλέον, ο αγοραστής επισκέπτεται καθημερινά διαφορετικά κανάλια και εκτελεί ένα περίπλοκο ταξίδι με σκοπό να ικανοποιήσει τις αυξανόμενες απαιτήσεις του. Η συμπεριφορά του σύγχρονου αγοραστή δεν είναι πια προβλέψιμη. Μεταβάλλεται διαχρονικά και ακόμη και μεταξύ των επισκέψεων του στο ίδιο κατάστημα (Sorensen et al., 2017). Αυτό έχει ως αποτέλεσμα, οι έμποροι λιανικής πώλησης να έχουν αρχίσει να συνειδητοποιούν ότι οι τεχνικές κατάτμησης των αγοραστών δεν είναι πλέον κατάλληλες, και δε μπορούν να περιγράψουν τις νέες, ασταθείς συνήθειες και προτιμήσεις των αγοραστών.

Από την άλλη πλευρά, και οι ερευνητές (Walters and Jamil, 2003; Bell, Corsten and Knox, 2011) αναγνωρίζουν αυτή την ανάγκη και υποδεικνύουν ότι πρέπει να δώσουμε προσοχή σε κάθε επίσκεψη ενός αγοραστή και να μην εστιάζουμε στο πώς εκείνος αντιδρά σε όλη του την αγοραστική ιστορία π.χ. μέσα σε ένα χρόνο. Εστιάζοντας και δίνοντας αξία σε κάθε επίσκεψη ενός αγοραστή, αντί της συνολικής αγοραστικής του συμπεριφοράς (όπως λειτουργούν οι προσεγγίσεις κατάτμησης πελατών) έχουμε τη δυνατότητα να εξασφαλίσουμε μια πιο ακριβή εικόνα των αναγκών του αγοραστή, που αλλάζουν καθημερινά λόγω της αφθονίας των προϊόντων, των καναλιών και των προσφερόμενων υπηρεσιών. Κατά συνέπεια, τόσο στον ακαδημαϊκό, όσο και τον επιχειρηματικό χώρο, καταδεικνύεται η ανάγκη για την κατάτμηση των επισκέψεων (visit segmentation) για να κατανοήσουμε τις μεταβαλλόμενες ανάγκες των αγοραστών που διαφέρουν σε κάθε επίσκεψη.

Σε αυτό το πλαίσιο στόχος της παρούσας διατριβής είναι να μελετήσει την έννοια της κατάτμησης των επισκέψεων (visit segmentation) στο λιανεμπόριο. Ορίζουμε κατάτμηση των επισκέψεων (visit segmentation) ως:

*«Τη διαδικασία κατάταξης των επισκέψεων των πελατών σε ομογενείς ομάδες που αποκαλύπτουν τις βαθύτερες ανάγκες, τις προτιμήσεις και τις αγοραστικές αποστολές (shopping missions) των πελατών, όπως αντικατοπτρίζονται στην αγοραστική συμπεριφορά τους κατά τις επισκέψεις στα φυσικά ή τα ηλεκτρονικά καταστήματα»*

Αναφερόμενοι στην αγοραστική συμπεριφορά εννοούμε στο πώς αυτή αντικατοπτρίζεται με βάση:

(Α) Τα περιεχόμενα ενός καλαθιού, όπως ορίζονται με βάση τις κατηγορίες προϊόντων που αυτό περιέχει π.χ. αυτή είναι μια επίσκεψη αγοράς πρωινού.

(Β) Τα χαρακτηριστικά του καλαθιού (όπως αξία, τεμάχια, ποικιλία), π.χ. αυτή η επίσκεψη είναι μια «επίσκεψη αναπλήρωσης», που περιλαμβάνει έναν μεγάλο όγκο τεμαχίων από μια μεγάλη ποικιλία προϊόντων.

(Γ) Τα χαρακτηριστικά της επίσκεψης, π.χ. στην επίσκεψη αυτή ο πελάτης έχει ως στόχο να περάσει γρήγορα τους διαδρόμους που εμφανίζουν επαγγελματικά ρούχα πιθανόν για να εξετάσει εάν τον εξυπηρετεί η ποικιλία του καταστήματος κλπ.

Όμως, καθώς οι νέες τεχνολογίες όπως το Διαδίκτυο των πραγμάτων (IoT) ενισχύουν τα δεδομένα που καταγράφουν την αγοραστική συμπεριφορά των πελατών, σε αυτά προσθέτουμε όλες τις αλληλεπιδράσεις κατά τη διάρκεια ενός αγοραστικού ταξιδιού, π.χ. τι αγοράζει ένας πελάτης σε ένα φυσικό κατάστημα ή ένα ηλεκτρονικό κατάστημα, τι τοποθετεί σε ένα εικονικό καλάθι στο διαδίκτυο αλλά τελικά δεν αγοράζει, ποια προϊόντα επιλέγει από τα έξυπνα ράφια ενός καταστήματος, ποια προϊόντα προσθέτει σε μια wish-list κοκ.

Στην παρούσα διατριβή η κατάτμηση των επισκέψεων πραγματοποιείται με τη χρήση δεδομένων πωλήσεων.

Αναφορικά με τη βιβλιογραφία, όπως αυτή παρουσιάζεται αναλυτικά στο κεφάλαιο 2, εντοπίσαμε ότι οι ερευνητές εστιάζουν κυρίως στην κατάτμηση των αγοραστών (shopper segmentation) π.χ. εντοπίζουν αγοραστές που αγοράζουν προϊόντα ρουτίνας, ή/και αγοραστές που ξοδεύουν πολλά χρήματα, και όχι στην κατάτμηση των επισκέψεων (visit segmentation). Το άλλο πιο κοντινό στην παρούσα έρευνα είναι μελέτες που εστιάζουν στις συσχετίσεις των προϊόντων (market basket analysis) που αγοράζει ένας πελάτης σε κάθε του επίσκεψη π.χ. όσοι αγόρασαν πάνες, αγόρασαν και μπύρες, χωρίς αυτές να πραγματοποιούν κατάτμηση των επισκέψεων (visits segmentation). Οι παραπάνω μελέτες αγνοούν το σκοπό επίσκεψης των αγοραστών στο κατάστημα, τις βαθύτερες προθέσεις τους και τις αγοραστικές τους αποστολές (shopping mission) οι οποίες δεν παραμένουν σταθερές στο πλαίσιο της κάθε επίσκεψης (φυσικής ή διαδικτυακής). Παράλληλα, εάν εστιάσουμε στη σχετική βιβλιογραφία, υπάρχει μια έλλειψη προσεγγίσεων που βασίζονται σε δεδομένα και εστιάζουν στην κατάτμηση επισκέψεων. Συνεπώς, διαφαίνεται ότι η σύγχρονη λιανική απαιτεί μετασχηματισμό παραδοσιακών συστημάτων και προσεγγίσεων κατάτμησης. Εμπνευσμένοι από τα παραπάνω, σε αυτή τη διδακτορική διατριβή προτείνουμε μία επέκταση των κλασσικών και παραδοσιακών προσεγγίσεων κατάτμησης των αγοραστών (shopper segmentation) προς την ενσωμάτωση και εκμετάλλευση της κατάτμησης των επισκέψεων (visit segmentation), σε μία προσπάθεια εκμαίευσης των αγοραστικών αποστολών (shopping mission) και των αναγκών που ώθησαν τον πελάτη να επισκεφθεί ένα κατάστημα π.χ. για να αγοράσει προϊόντα για το πρωινό του ή για να προμηθευτεί υλικά για την ανακαίνιση του μπάνιου του κλπ.

Σε αυτό το πλαίσιο, στην παρούσα διατριβή εξετάζονται τα ακόλουθα ερωτήματα:

- Ε1. Πώς μπορούμε να κάνουμε κατάτμηση των επισκέψεων των πελατών χρησιμοποιώντας δεδομένα που χαρακτηρίζουν τη συμπεριφορά τους;

  Σε αυτή τη λογική προκύπτουν τα παρακάτω υπο-ερωτήματα:

    - Μπορούμε να εξαγάγουμε τις αγοραστικές αποστολές (shopping missions) των πελατών από τις αντίστοιχες ομάδες (visit segments);
    - Μπορούμε να αναπτύξουμε μια προσέγγιση/μεθοδολογία βασισμένη σε τεχνικές επιχειρηματικής αναλυτικής (business analytics), για να επιτύχουμε την κατάτμηση των επισκέψεων (visit segmentation);

- Ε2. Ποιοι είναι οι παράγοντες που επηρεάζουν το σχεδιασμό των συστημάτων κατάτμησης των επισκέψεων;

Για να απαντήσουμε στα παραπάνω ερωτήματα, στην παρούσα διατριβή συνδυάζουμε τρία διαφορετικά γνωστικά πεδία: Πληροφοριακά Συστήματα (Information Systems), Επιχειρηματική Αναλυτική (Business Analytics), και Shopper

Marketing, και υιοθετούμε ως μεθοδολογική προσέγγιση τη μεθοδολογία Design Science (Hevner et al., 2004). Ως εκ τούτου, η θεωρητική συμβολή της παρούσας διατριβής εκτείνεται και σε αυτούς τους τρεις κλάδους. Επιπλέον, λόγω της έλλειψης προηγούμενης συστηματικής έρευνας για το θέμα της κατάτμησης των επισκέψεων η παρούσα έρευνα βασίζεται σε σχεδιασμό πολλαπλών μελετών περίπτωσης (multiple case study design).

Απατώντας στο πρώτο ερώτημα, αναλυτικότερα, στο κεφάλαιο 4 αναπτύξαμε μια προσέγγιση (approach) κατάτμησης της επίσκεψης, η οποία φιλοδοξεί να καλύψει το κενό της βιβλιογραφίας. Η προσέγγιση αυτή περιλαμβάνει τις ακόλουθες φάσεις: (α) κατανόηση και προετοιμασία δεδομένων, (β) μοντελοποίηση των δεδομένων και αξιολόγηση του μοντέλου, (γ) μετάφραση αποτελεσμάτων και αναγνώριση των αγοραστικών αποστολών. Η προσέγγιση αυτή δίνει μια ευρύτερη προοπτική για τον τρόπο με τον οποίο εξετάζουμε τις αγοραστικές επισκέψεις και αναζητούμε την πρόθεση/αποστολή του πελάτη σε κάθε επίσκεψή του.

Όπως αναφέρθηκε παραπάνω, η αναγνώριση των αγοραστικών αποστολών βασίζεται στα προϊόντα με τα οποία αλληλοεπιδρά ένας πελάτης σε κάθε του επίσκεψη. Κατά συνέπεια, ένα πολύ σημαντικό θέμα είναι ο σωστός ορισμός των κατηγοριών. Για το λόγο αυτό η παρούσα έρευνα προτείνει μια ημι-εποπτευμένη προσέγγιση επιλογής χαρακτηριστικών (semi-supervised feature selection method) από προϊοντικές κατηγορίες του λιανεμπόρου. Η προσέγγιση αυτή δέχεται ως είσοδο της το δέντρο ιεραρχίας των προϊοντικών κατηγοριών ενός λιανεμπόρου και εξάγει κατάλληλες προϊοντικές κατηγορίες έτοιμες να χρησιμοποιηθούν για την ανάλυση. Ουσιαστικά η παραπάνω μέθοδος χρησιμοποιείται για την ισοστάθμιση (balancing) του αρχικού δένδρου ταξινόμησης των προϊόντων και λαμβάνει υπόψη της τόσο τη συχνότητα των αγορών προϊόντων, όσο και τη σημασιολογία (semantics) των προϊόντων. Αυτό έχει ως αποτέλεσμα τη δημιουργία ενός πιο απλού χώρου για να πραγματοποιηθεί η συσταδοποίηση (clustering).

Σε αυτή την έρευνα δείχνουμε επιπροσθέτως ότι το επίπεδο ανάλυσης (unit of analysis) που χρησιμοποιείται στη βιβλιογραφία, δηλαδή η μοναδική επίσκεψη ή όλες οι επισκέψεις των αγοραστών, δεν καταδεικνύουν σε κάθε λιανικό περιβάλλον την αγοραστική αποστολή ενός πελάτη. Αντιθέτως, υπάρχουν περιπτώσεις στις οποίες θα πρέπει να εξετάσουμε «Χ» διαδοχικές επισκέψεις για να κατανοήσουμε την πραγματική αγοραστική αποστολή του πελάτη. Η τιμή του «Χ» διαφέρει ανάλογα με το πεδίο από το οποίο προέρχονται τα προς ανάλυση δεδομένα.

Στο κεφάλαιο 5 εφαρμόσαμε και επικυρώσαμε την προτεινόμενη προσέγγιση που εξάγει τις αγοραστικές αποστολές μέσω τριών ετερογενών μελετών περίπτωση. Με αυτό τον τρόπο καταδεικνύουμε τη γενίκευση και την εφαρμοσιμότητα της προσέγγισης αυτής σε διαφορετικά πεδία. Η πρώτη μελέτη περίπτωσης αφορά δεδομένα πωλήσεων από διαφορετικά κανάλια και καταστήματα ενός μεγάλου ευρωπαίου λιανοπωλητή προϊόντων ταχείας κυκλοφορίας (FMCG retailer). Αντίστοιχα, στη δεύτερη περίπτωσης, χρησιμοποιήσαμε δεδομένα από καταστήματα ενός Fortune 500 λιανοπωλητή που πουλά προϊόντα βελτίωσης σπιτιού και ιδιοκατασκευής- γνωστός και ως DIY (do-it-yourself) λιανέμπορος. Η τρίτη περίπτωση αφορά δεδομένα από ένα φυσικό και το ηλεκτρονικό κατάστημα ενός μεγάλου Ευρωπαίου λιανοπωλητή μόδας.

Εκτός από την αξιολόγηση της προτεινόμενης προσέγγισης που βασίζεται σε τεχνικές επιχειρηματικής αναλυτικής (business analytics), εφαρμόζοντάς την στις διάφορες περιπτώσεις λιανικής, αξιολογούμε επίσης τα αποτελέσματα της προσέγγισής μας. Πιο συγκεκριμένα, στο κεφάλαιο 6, πραγματοποιήσαμε ημι-δομημένες ομάδες εστίασης (focus groups) για να συζητήσουμε με αγοραστές (που ψωνίζουν στο κατάστημα από το οποίο προέκυψαν τα δεδομένα) και ρωτήσαμε την άποψή τους για τις αγοραστικές αποστολές που αναγνωρίσαμε από τα δεδομένα πωλήσεων. Επίσης, αξιολογήσαμε στο πλαίσιο μίας μελέτης πεδίου (field study) μέσα στο κατάστημα τις προκύπτουσες αγοραστικές αποστολές και την εγκυρότητά τους. Για το λόγο αυτό, χρησιμοποιήσαμε μια εφαρμογή για έξυπνα κινητά τηλέφωνα, μέσω της οποίας διανείμαμε κουπόνια. Στόχος μας ήταν να διερευνήσουμε πως η αξιοποίηση της γνώσης από την κατάτμηση των επισκέψεων μπορεί να έχει επίδραση στα αποτελέσματα μιας προωθητικής ενέργειας.

Μετά την αξιολόγηση των αποτελεσμάτων, στο ίδιο κεφάλαιο, παρουσιάζουμε διαφορετικούς τρόπους αξιοποίησης της εξαγόμενης γνώσης από την πλευρά του μάρκετινγκ. Πιο συγκεκριμένα, η αξία μιας τέτοιας καινοτόμας προσέγγισης ανάλυσης δεδομένων, διαφαίνεται όταν χρησιμοποιούμε την εξαγόμενη γνώση για τη στήριξη των επιχειρηματικών αποφάσεων. Για παράδειγμα, μια τέτοια μεθοδολογία θα μπορούσε να εξελιχθεί σε ένα εργαλείο σχεδιασμού καινοτόμων εκστρατειών μάρκετινγκ, προωθήσεων και πωλήσεων. Παρομοίως, μπορούμε να δημιουργήσουμε καταλόγους προϊόντων για συγκεκριμένες αγοραστικές αποστολές. Η εξαγόμενη γνώση θα μπορούσε να είναι επίσης πολύτιμη για διαφημιστικούς σκοπούς. π.χ. διαφημίσεις προϊόντων πρωινού. Επιπλέον, η κατάτμηση των επισκέψεων μπορεί να οδηγήσει και σε μια νέα επανασχεδιασμένη διάταξη του καταστήματος όπου οι κατηγορίες προϊόντων που ανήκουν στην ίδια αγοραστική αποστολή τοποθετούνται σε κοντινούς διαδρόμους και ράφια (Vrechopoulos, O'Keefe, Doukidis and Siomkos, 2004; Cil, 2012; Sarantopoulos, Theotokis, Pramatari and Doukidis, 2016).

Σε αυτή τη λογική , η παρούσα διατριβή προτείνει τη μετάβαση από την παραδοσιακή τεχνική διαχείρισης των κατηγοριών (Category Management) στη διαχείριση των αγοραστικών αποστολών (Shopping Mission Management), ανοίγοντας ένα νέο πεδίο στη βιβλιογραφία της διαχείρισης των κατηγοριών.

Επιπροσθέτως, σε αυτή τη διατριβή προσδιορίζουμε και συζητούμε όλους αυτούς τους παράγοντες που, όπως φαίνεται από τη βιβλιογραφία, αναμένεται να επηρεάσουν τα παραδοσιακά συστήματα κατάτμησης των αγοραστών. Επίσης, επισημαίνουμε εκείνους τους παράγοντες που επηρεάζουν την προσέγγιση κατάτμησης των επισκέψεων. Τέλος, η παρούσα έρευνα φιλοδοξεί να γεφυρώσει τους ερευνητές και τους διευθυντές μάρκετινγκ με τους επιστήμονες δεδομένων (data scientists) και τους σχεδιαστές συστημάτων κατάτμησης των επισκέψεων και των αγοραστών. Για το σκοπό αυτό, κλείνοντας αυτή τη διατριβή, παρουσιάζουμε λεπτομερώς τους παράγοντες που όλοι οι παραπάνω πρέπει να εξετάζουν αν θέλουν να σχεδιάσουν συστήματα και προσεγγίσεις κατάτμησης των επισκέψεων. Έτσι, η έρευνα αυτή θέτει επίσης τις βάσεις για τις αρχές ανάπτυξης σχετικών εργαλείων.

# ABSTRACT

In contemporary retail, both practitioners and researchers agree that old-school shopper segmentation is not enough and cannot describe the new, volatile shopper habits and preferences. They suggest that retail nowadays, demands a transformation of shopper segmentation systems and approaches. This happens since the modern shopper has changed. Nowadays, the shopper flits between shopping channels and performs a complex shopper journey with the purpose to satisfy his/her increasing demands for quality and value (Wood, 2018). Shopper behavior is no longer predictable; it is changing through time and, even, between shopping visits in the same store (Sorensen et al., 2017). Thus, there is a need to focus on shoppers' visits and perform visit segmentation to cope with shoppers' changing behavior. In this spirit, the goal of this research is to study and advance the understanding of visit segmentation in retail.

Practitioners have coined the term "shopping mission" to refer to the intention behind a shopper's visit (ECR Europe, 2011). Similarly, researchers (Walters and Jamil, 2003; Bell et al., 2011) agree with practitioners and suggest that we should pay attention to each shopper visit as it carries valuable insight on the shopper needs. Looking at each specific shopper visit, instead of a shopper's total buying behavior over many visits, provides a more accurate view of the shopper desires that change frequently due to an abundance of new products, shopping channels and services offered every day. Approaches and studies that exploit shopper heterogeneity and take into account the changing behavior of customers are becoming all the more important (Rust and Huang, 2014).

Looking into the segmentation literature (in chapter 2), we can classify the pertinent studies in two broad categories: those focusing on shopper segmentation, and those that analyze shopper data and focus on the associations between the products a shopper purchases during a visit (also known as "market basket analysis"). The first group of studies examine everything a shopper has purchased in bulk, and overlook the shopping purpose, intentions and missions of the shopper, which are not the same in every store visit. Regarding the second group of studies, although they examine shoppers' visits, they mostly focus on the association between specific products (market basket analysis) (Srikant and Agrawal, 1995; Boztuğ and Reutterer, 2008; Cil, 2012; Beck and Rygl, 2015) e.g. those who bought diapers also bought beer. Thus, they still overlook the shopping purpose of each shopper visit.

In this dissertation, we examine each shopper visit individually in order to acquire a more accurate view of the shopper needs and understand the underlying needs that boosted a customer visit a store e.g. to purchase products for a light meal, or to procure materials to renovate his/her bathroom etc. These needs and missions can be extracted using various datasets reflecting customers' behavior e.g. product purchases, interactions, preferences etc. We further segment shoppers based on the different visit profiles and coin the term "visit segmentation".

To this end, in the current dissertation the following questions are addressed:

- Q1. How can we derive visit segments from shopper data?
    - Can we extract the customers' shopping missions from the identified visit segments?
    - Can we develop a business analytics-informed approach to perform visit segmentation?

- Q2. What are the factors that affect the design of visit segmentation systems?

To address these questions, this dissertation interweaves three different disciplines: Information Systems (IS), Business Analytics (BA), Shopper Marketing and adopts the design science paradigm (Hevner, March, Park and Ram, 2004). In design science, the researcher creates and evaluates IT (Information Technology) artifacts and/or theories intended to solve identified organizational problems. In the current thesis we consider a business analytics approach that performs visit segmentation as outcome of this study. Additionally, due to the lack of prior systematic research on the topic of visit segmentation, this research is based on multiple case studies design.

Delving deeper into the segmentation literature, we identified that there is a lack of data-driven approaches that perform visit segmentation. Thus, addressing the first research question, in chapter 4 we developed a visit segmentation approach that aspires to fill the literature gap. This includes the following phases: (a) data understanding and preparation (where the data are pre-processed, cleaned and prepared for the data analysis purposes), (b) data modelling and model evaluation (where the data mining model is created and the results are evaluated in both business and technical terms), (c) results interpretation (where the visit segments are extracted, interpreted and translated into shopping missions).

Our proposed segmentation approach moves the attention from the purchased products to the shopping needs that motivate the shopper's shopping visits. We adopt a broader perspective on how we examine shopping visits and we seek the shopping intention(s), motive(s) and mission(s) of each visit. As the visit segmentation is based on the products a customer interacted with (e.g. purchased, during a shopping visit) it is important to correctly define the product dimension. Thus, during the insight generation process we do not overlook the significant effect of the product taxonomy on the effectiveness and validity of our clustering results (Albadvwe & Shahbazi, 2009; Cho, Kim, & Kimb, 2002). More specifically, product taxonomies are often unbalanced and have characteristics hindering the performance of data mining algorithms (Srikant and Agrawal, 1995; Cho et al., 2002; Cho and Kim, 2004; Hung, 2005; Albadvi and Shahbazi, 2009; Han, Ye, Fu and Chen, 2014). Thus, it matters for example whether we should refer to a can of sparkling orange juice of brand XYZ as sparkling beverage, as beverage, or as orange juice. For that reason, this research also suggests a semi-supervised feature selection approach that uses the product taxonomy as input and extracts the features (product categories) as output. This approach considers both the frequency of product purchases and the product semantics to adjust and balance the original product taxonomy tree.

In this research, we also revealed that the unit of analysis used in the literature, i.e. product items in a single visit, or all shopper visits, are not applicable in every retail context, but there are cases where we should examine groups of “x” sequential visits. The value of “x” differs according to the features of the domain the data derived from. We devise and test a new unit of analysis where we examine groups of x continuous visits. This intermediate unit of analysis is dictated by the particularity of some retail domains that demand many store visits during small time windows.

In chapter 5, we applied and validated our approach through three heterogeneous retail cases to demonstrate its generalizability. The first case concerns sales data from different channels and stores of a major European fast-moving consumer goods (FMCG) retailer. Respectively, in the second case, we produced the visit segments for the physical stores of a Fortune 500 specialty retailer of home improvement and construction products – also known as do-it-yourself (DIY) retailer. The third case concerns data from a physical and web store of a major European fashion retailer.

Apart from evaluating the proposed business analytics approach by applying it to the different retail cases, we also evaluate the impact of our approach. For that reason, in chapter 6, we conducted semi-structured focus groups to discuss with actual shoppers (shopping at the store the data derived) and ask for their view on the resulting visit segments/shopping missions. We also designed an in-store field study to evaluate the resulting data-driven shopping missions and asses their validity in the context of a specific in-store promotion using a mobile app. We demonstrate that the shopping mission-related disseminated coupons achieve higher redemption rate and are claimed by shoppers in less time than the non-related coupons.

We further present data-driven innovations in shopper marketing that the resulting visit segments could support ranging from marketing campaigns per visit segment and redesign of a store's layout to cross-selling strategies and product recommendations. To this end, this dissertation also proposes moving from category to shopping mission management and opens a new chapter in the category management (CM) literature.

In parallel, we identify and discuss the factors that, according to the literature, are expected to affect shopper segmentation systems. This research aspires to bridge marketing researchers and managers with data scientists and shopper segmentation designers. Hence, we conclude by presenting the various factors (data, shopper, marketing and retail-specific) that managers in the retail industry, as well as marketing researchers and data scientists should consider when designing visit-segmentation systems, setting the basis for the development of IS tools for visit segmentation.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| ABBREVIATIONS | TERMS |
|---|---|
| 3PL | Third Party Logistics |
| 4Ps | Place, Product, Price, Promotion |
| ANNs | Artificial Neural Networks |
| ANOVA | Analysis of variance |
| BA | Business Analytics |
| BI | Business Intelligence |
| BLE | Bluetooth Low Energy |
| BOGO | Buy one, get one free |
| CLV (or LTV) | Customer Lifetime Value |
| CM | Category Management |
| CRISP-DM | Cross Industry Standard Process for Data Mining |
| Data 4Vs | Volume, Velocity, Variety, Value |
| DGX | Discrete Gaussian Exponential |
| DIY | Do-it-yourself |
| DM | Data Mining |
| DSR | Design Science Research |
| ECR | Efficient Consumer Response |
| EDLP | Everyday low price |
| EM | Expectation Maximization |
| FMCG | Fast Moving Consumers Goods |
| GPS | Global Positioning System |
| IoT | Internet of Things |
| IS | Information Systems |
| IT | Information Technology |
| KPI | Key Performance Indicator |
| MBA | Market Basket Analysis |
| PCA | Principal Component Analysis |
| PL | Private Label |
| POS | Point-of-Sale |
| RFID | Radio Frequency Identification |
| RFM | Recency, Frequency, Monetary |
| Shopper 4Vs | Visits, Variety, Value, Volume |
| SQL | Structured Query Language |

# 1. INTRODUCTION

This opening chapter begins by laying the motivation for undertaking this research and by positioning its topic within its research context. Subsequently, it pinpoints pertinent research gaps and questions. Then it shortly presents the research methodology and concludes by providing the dissertation outline.

## 1.1. Research motivation

The increasing capabilities of business intelligence (BI), business analytics (BA) tools and techniques and data-driven decision making have risen in the agenda of many businesses (McKinsey Global Institute, 2011). Companies in various domains are trying to become data-driven and cultivate a data-oriented culture (McAfee and Brynjolfsson, 2012). Similarly, retailers collect and store voluminous and several types of data about their customers daily, ranging from customer demographics, to data that indicate how customers move into the physical or web stores, what products they put in their baskets or try on in the fitting rooms, what products they purchase etc. Since the data volume, variety and velocity have far outstripped the capacity of manual analysis (Chang, Kauffman and Kwon, 2014), one of the greatest aspiration of retailers is to find innovative ways to exploit the collected datasets. They have long recognized that data-driven decision making can improve decision quality (Kowalczyk and Buxmann, 2015). Since customer satisfaction affects profitability, i.e. the key to business success, retailers want to embrace a more customer-centric approach and find out innovative ways to understand their customers and satisfy them (Anderson, Jolly and Fairhurst, 2007; Linoff and Berry, 2011). At the same time, marketeers and category managers want to incorporate the extracted customer knowledge into the current category management (CM) practices and embrace a more consumer-centric CM approach (Han et al., 2014; Nielsen, Karolefski and Heller, 2015).

Consequently, they seek for means to exploit the collected customer data (e.g. what they purchase, how they move in the stores etc.) and extract knowledge that facilitates effective decisions and offer extra value to their demanding customers. Taking advantage of business analytics, their aim is to acquire new non-trivial knowledge and from the accumulated data and, ultimately, contribute to more efficient decisions and more satisfied consumers. Additionally, another important fact is to identify interesting patterns. According to Silberschatz and Tuzhilin (1996) a pattern is interesting if it satisfies two measures i.e. actionability (e.g. can be used to support various actions) and unexpectedness (e.g. it is a "surprising" outcome). Such interesting patterns and new knowledge are an opportunity for companies to generate more reliable shopper segments and provide to their shoppers targeted services tailored to their needs and preferences (Boone & Roehm, 2002).

Shopper segmentation is an old concept that is rapidly revived in contemporary retail, due to the data revolution and explosion. Nowadays, large volume of data emerges every day from various devices and interactions. Thus, in the retail environment, massive amounts of data are gathered daily from different channels e.g. physical, web, mobile (Bradlow, Gangwar, Kopalle and Voleti, 2017). The derived shopper datasets are ranging from transactions, loyalty schemes and legacy systems, to RFID (Radio Frequency Identification) tracking technologies and Bluetooth low energy (BLE) beacons. These data, if properly analyzed, can connect retailers with people, help identify the different shopper segments visiting their stores, understand their behavior, and guide both future strategies and daily operations (Sharma et al., 2014).

In contemporary retail, apart from the data explosion, researchers (Walters and Jamil, 2003; Bell et al., 2011) have also detected a transformation in shopper behavior. Modern shoppers are shifting their behavior over time and push retailers to become increasingly agile. Shopper behavior is no longer predictable; it is changing through time and, even, between shopping visits in the same store (Sorensen et al., 2017). Hence, the retailers have begun to realize that the traditional, old-school shopper segmentation is not enough and cannot describe the new, volatile shopper habits and preferences. These facts demand changes in contemporary shopper segmentation systems and approaches.

Data explosion and shoppers' changing behavior impose the **need to focus on each shopper visit to better understand their needs and missions** in contemporary retail. Marketing **researchers** (Bell, Corsten, & Knox, 2011; Walters & Jamil, 2003), have stressed the need to put the shopper visit into the spot (rather than the shopper) and perform **visit segmentation**, to better understand shopper needs and design more efficient shopper marketing activities. Alike **practitioners** have coined the term "**shopping mission**" to refer to the intention(s) that boosted a shopper's visit (ECR Europe, 2011).

Retailers seek for their shoppers deeper shopping missions and motives to offer more suitable services, such as personalized promotions, cross-selling coupons, shopping mission-based layouts etc. For instance, retailers are interested in identifying that there are store visits which shopping mission to purchase products for "sushi", or "pastry making", or "breakfast", or "food-to-go" etc. Similarly, other examples of shopping missions could be "kitchen renewal" in a DIY (do-it-yourself) store, or "professional clothes" in a fashion store.

At the same time, big data and business analytics offer the opportunity to analyze massive data volumes and extract insights to understand modern shoppers and acclimate into this new era. However, practitioners (Bean and Davenport, 2019) highlight that there is a data-centric "fatigue" and companies are failing to become data-driven. Thus, literature and guidelines are needed to guide both practitioners and researchers where and how to deep dive in the data to develop sharp hypotheses that can be tested (Bradlow et al., 2017; Delen and Zolbanin, 2018). Therefore, it is critical to design business-analytics informed approaches to identify the different visit segments and understand shoppers' deeper needs, preferences and missions.

## 1.2. Research objective & questions

Considering the above, as visit segmentation research is still infancy, the objective of this research is to advance the understanding of visit segmentation. Visit segmentation focuses on the underlying needs that boosted a customer visit a store e.g. to purchase products for a light meal, or to procure materials to renovate their bathroom etc. These needs and missions can be extracted using various shopper-related datasets reflecting customers' behavior e.g. product purchases, interactions, preferences etc.

Thus, the first research question is formulated as follows:

- Q1. How can we derive visit segments from shopper data?

Via performing desk research, we identified that business people translate visit segmentation as "shopping mission" (ECR Europe, 2011). Thus, another objective of this research is to answer whether we can identify shoppers' intentions and missions from the available visit segments. Thus, the first question is reformulated as follows:

- Q1. How can we derive visit segments from shopper data?
    - Can we extract the customers' shopping missions from the identified visit segments?

At the same time, looking more thoroughly into the literature, we identified that Editorials (Agarwal and Dhar, 2014; Goes, 2014), other academic papers (Abbasi, Sarker and Chiang, 2016; Müller, Junglas, Brocke and Debortoli, 2016; Delen and Zolbanin, 2018) and practitioners (McKinsey Global Institute, 2011) emphasized the need to develop data-driven approaches, systems and frameworks to better understand shoppers' behavior and shoppers' changing needs (Pick, Turetken, Deokar and Sarkar, 2017). However, delving deeper into the rest segmentation literature, there are is a lack of business analytics-informed and data-driven approaches to identify the various visit segments and understand shoppers' deeper needs, preferences and missions. Thus, our first question is further reformed as follows:

- Q1. How can we derive visit segments from shopper data?
    - Can we extract the different shopping missions of customers from the identified visit segments?
    - Can we develop a business analytics-informed approach to perform visit segmentation?

Delving deeper into the segmentation literature, this research also seeks for factors that affect the input, the design and the results of such systems. Our goal is to pinpoint factors, that (A) prospective designers of segmentation systems should consider if they want to produce valid segments, (B) data scientist should consider when manipulating and modeling data and (C) marketeers should consider interpreting the segmentation results. Closing, we identify several research gaps related to the visit segmentation concept. Thus, a second question is also formulated:

- Q2. What are the factors that affect the design of visit segmentation systems?

Here, we must admit that studying segmentation literature various other questions may arise. For instance:

- Do shopper segmentation systems serve contemporary retail?
- How we define visit segmentation?
- How can we derive the various visit segments?
- What kind of data do we need to do so?
- Can we use the visit segments to support marketing strategies?
- Are the visit segmentation-informed marketing actions more effective than the traditional actions?
- What is the conceptual relation between the visit segments and the shopping missions?
- Shoppers identify the shopping mission they entered the store for?

These questions might concern not only business analytics and IS, but also marketing researchers. Therefore, this research also performs an overall discussion regarding the aforementioned questions.

## 1.3. Overview of the research methodology

Given our research objective and the aforementioned research questions, we adopt as methodological backbone the design science research (DSR) paradigm (Hevner et al.,

2004) and we consider a business analytics approach that performs visit segmentation as outcome of this study. In design science, the researcher creates and evaluates IT (Information Technology) artefacts and/or theories intended to solve identified organizational problems. The knowledge base is composed of foundations and methodologies used to develop the artefact.

Owing the lack of prior systematic research on the visit segmentation topic this research is based on multiple case studies design. Below we present the basic components of design science research and how these are addressed in the current dissertation (Figure 1-1). Afterwards, we shortly explain and translate the basic components of DSR according to our research objectives. An extended description of this dissertations' research methodology is also presented in Chapter 3.

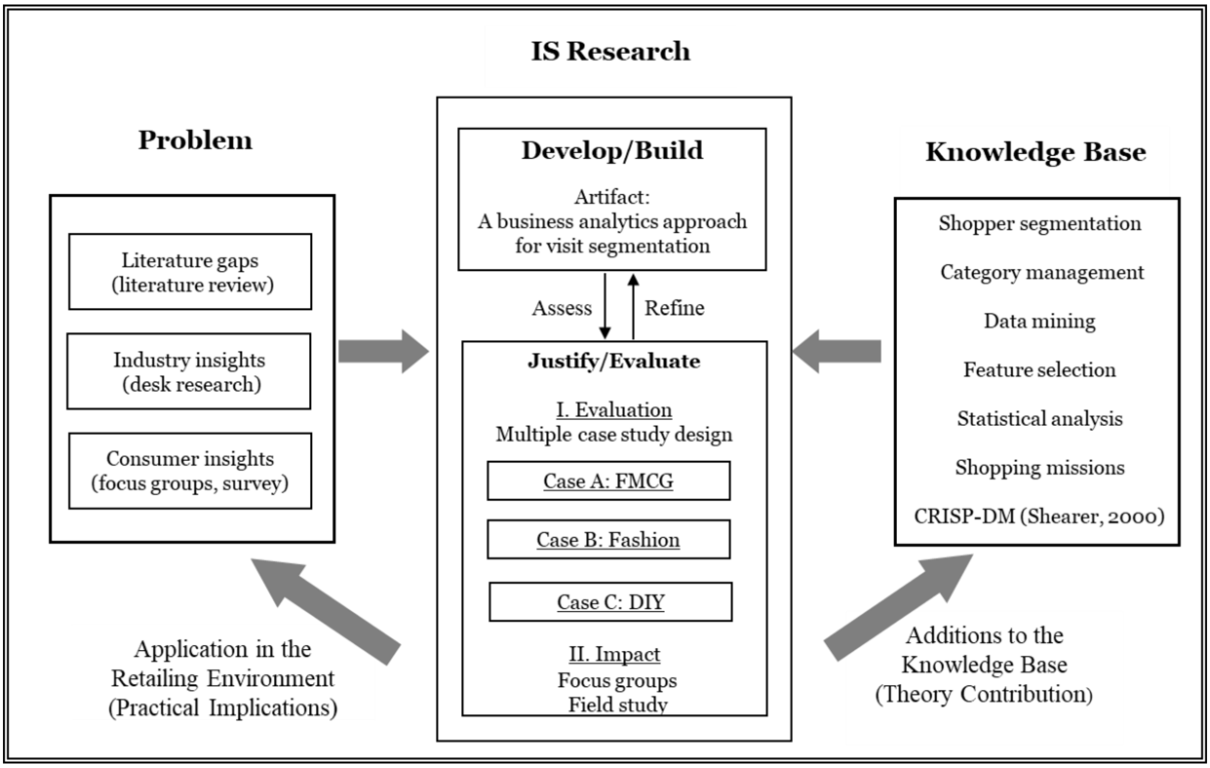



**Figure 1-1. Research methodology (adapted from Hevner, 2004)**

## (A) Problem definition

This dissertation aims to solve a business problem/need in the retailing environment, which is to perform visit segmentation and identify the underlaying shopping needs and missions of customers. To better define this problem, it follows the steps below:

- **Literature gaps**: To set the research setting firstly we conducted a review of the pertinent literature. This way we specified the research questions which is related with the visit segmentation concept and we pinpoint the research gaps and the purpose of this research.
- **Industry insights**: Apart from defining the foundations upon which this doctoral research is grounded, we also investigated various open issues and business problems industry people face, when they try to better understand and satisfy their demanding customers. In more detail, we identified that business people translate visit segmentation as "shopping mission". Thus, we use the term "visit segmentation" to more precisely describe the "shopping mission" term which widely utilized in the industry literature. Visit segmentation focuses on the underlying needs that boosted a customer visit a store e.g. to purchase products for a light meal, or to procure materials to renovate their bathroom etc.
- **Consumer insights**: Afterwards, to better understand and conceptually define shopping missions in the FMCG domain, we conducted a series of eight semi-structured focus group discussions with 71 shoppers. These discussions confirmed that contemporary shoppers entering the store having in mind a specific shopping mission. In addition, survey is used to investigate shoppers' behavior and perception regarding the shopping mission concept.

**(B) Develop/Build**

Then we develop and evaluate the solution that is relevant to the above research problem. In this research the **developed artefact is an approach**, providing a certain manner to handle the appropriate data aiming **to extract the visit segments.**

**(C) Justify/Evaluate**

Then, we put the approach in practice to evaluate it and realize if it can solve the original problem. This phase includes two steps: (i) Evaluation and (ii) Impact. Regarding the first one owing the lack of prior systematic research on the visit segmentation topic, to address this objective the research is based on multiple case studies design. In more detail, the proposed approach is evaluated by applying it to real data derived from three case studies (fast moving consumer goods retailing - FMCG, Do-it-yourself retailing - DIY, fashion retailing). By applying the approach in three different retail cases we evaluate it and we confirm its generatability. Afterwards, to evaluate the results of our approach and asses their impact we designed a series of focus groups and a field study using a mobile app for a store of a major Greek FMCG retailer.

**(D) Knowledge base**

To build the proposed approach, we used both theoretical foundations and methodologies. Theory regarding shopper segmentation and category management and CRISP-DM (Chapman et al., 2000) which is a Cross-Industry Standard Process for Data Mining, were used as the basic knowledge inputs in the developed approach. In more detail in our approach we follow and alter the CRISP-DM steps. In addition, data mining techniques such as clustering and classification, data mining algorithms such as k-means and feature selection methods were also used to develop the approach. Statistical analysis and measures such as ANOVA, Pearson correlation, Jaccard similarity etc. were used to evaluate the field study results and to analyze the shopper survey. Likewise, qualitative analysis used to analyze the focus group transcripts during the various research faces.

Closing, the theory contribution and the practical implications are detailed. Regarding theory contribution this dissertation, develops a business analytics approach that performs visit segmentation. To the best of our knowledge this is the first data-driven attempt to identify visit segments and explore the underlying customers' shopping missions. In a nutshell, the practical value of this work is stressed when considering the consumer-oriented business decisions it can support e.g. shopping-mission based store layout, or product catalogues, or promotions etc.

## 1.4. Thesis outline

There are seven (7) chapters that constitute this dissertation as follows:

- Chapter 1 (Introduction)

  This chapter introduces the readers to the main concepts of this research, i.e. visit segmentation. It communicates the research's motivation, as well as the key research question. Last, but not least, this chapter shortly describes the research approach according to which the main research objective and question are answered.

- Chapter 2 (Research background)

  It is critical to be cognizant of the rationale for the relevance of the work. Therefore, in this chapter an extensive literature review is presented to pinpoint the research gaps. Also, these research gaps are translated into industry open issues and business problems. Then, customer insights via focus groups and via a survey are used as tools for highlighting the significance of this study and for validating the main research objective. Finally, we conclude with the main research gaps.

- Chapter 3 (Research methodology)

  The aim of this chapter is to present the research methodology employed to address the research objectives and answer the research questions. Given our research objective and the aforementioned research questions, we adopt as methodological backbone the Design Science Research (DSR) paradigm (Hevner et al., 2004) and we consider a business analytics approach that performs visit segmentation as outcome of this study. For collecting data for the various steps of DSR, three different cases studies are selected and presented (multiple case study design). Thus, firstly we present and explain DSR, and then we explain multiple case study rationale. Afterwards, we describe in detail how we adopt these two research approaches into the design of this dissertation.

- Chapter 4 (A business analytics approach for visit segmentation)

  In this chapter we describe in detail the proposed data-driven approach that could be used to perform visit segmentation. In high level it includes the following phases/layers: : (a) data understanding and preparation (where the data are pre-processed, cleaned and prepared for the data analysis purposes), (b) data modeling and model evaluation (where the data mining model is created and the results are evaluated in both business and technical terms), (c) results interpretation (where the visit segments are extracted and interpreted).

The major input of our approach is data related to shopping behavior (e.g. content of a basket and basket characteristics). The output is the final visit segments that are translated and interpreted into shopping missions.

- Chapter 5 (Application of the proposed approach in three heterogenous retail cases)

  Here, we put our proposed business analytics approach in practice demonstrating how it achieves the original goal, i.e. to segment the shoppers' visits. We applied and validated our approach through three heterogeneous retail cases to demonstrate its generalizability. The first case concerns sales data from different channels and stores of a major European fast-moving consumer goods (FMCG) retailer. Respectively, in the second case, we produced the visit segments for the physical stores of a Fortune 500 specialty retailer of home improvement and construction products – also known as do-it-yourself (DIY) retailer. The third case concerns data from a physical and the web store of a major European fashion retailer.

- Chapter 6 (The impact of visit segmentation on shopper marketing)

  At the beginning of this chapter we evaluate the results of our approach and examine their impact on shopper marketing actions. For that reason, we conducted semi-structured focus groups to discuss with the actual store shoppers and ask for their view on the resulting visit segments/shopping missions. Also, we designed an in-store field study to evaluate the resulting data-driven shopping missions and asses their validity. For that reason, we utilized a mobile app and we distributed coupons. We demonstrate that the shopping mission-related disseminated coupons achieve higher redemption rate and are claimed by a shopper into less time than the non-related coupons. After the results evaluation, to showcase the impact of the visit segmentation, we present data-driven innovations in shopper marketing that these resulting visit segments could support. Closing, we present an alternative way from shopper segmentation using the resulting visit segments. Thus, we use real datasets to illustrate how these visit segments could be used as the cornerstone to perform shopper segmentation for more effective shopper marketing.

- Chapter 7 (Conclusions and Discussion)

  This final chapter overviews the main outcomes of this research. Then, it presents and discusses the research's contribution to theoretical knowledge along with its practical value. Then, the research limitations are pointed out and avenues for further research are recommended. At the end of this chapter, we present thoughts for visit segmentation systems designers in contemporary retail. Closing this dissertation, we present in detail the data, shopper, marketing and retailer's factors that designers should take into consideration when designing visit segmentation systems.

  Finally, the thesis includes a set of four (4) Appendices A to D that complement the chapters.

To support reading comprehension, the following figure (Figure 1-2) presents thesis' structure according to the research methodology, as described in the previous sub-section.

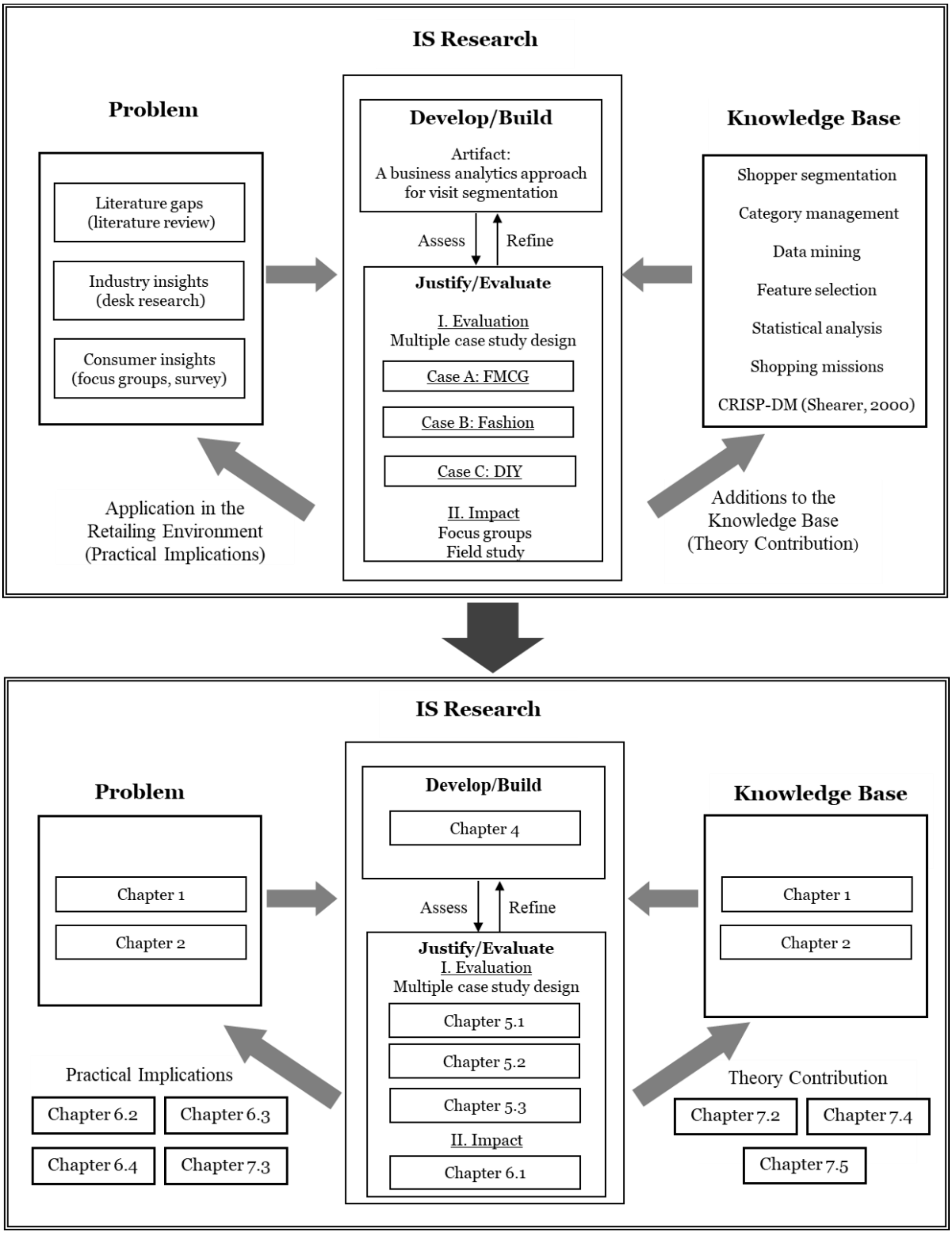



**Figure 1-2. Research outline**

# 2. BACKGROUND

It is critical to be cognizant of the rationale for the relevance of the work. Therefore, an extensive literature review and shopper insights have been used as tools for highlighting the significance of this study. In more detail, firstly we delved deeper into the related literature to identify the research gaps. Then, we asked for consumer's opinion and point of view via focus groups and via a survey. Closing this chapter, we conclude with the research gaps.

## 2.1. Literature review

Retailers want to embrace a more customer-centric approach and find out innovative ways to understand the specific needs, shopping missions and preferences of their customers (Anderson et al., 2007; Linoff and Berry, 2011). At the same time, marketeers and category managers want to incorporate the extracted customer knowledge into the current category management (CM) practices and embrace a more consumer-centric CM approach (Han et al., 2014; Nielsen et al., 2015). Thus, the first objective of the literature review is to explore Category Management (CM) literature.

The biggest change in CM, is that over the years it becoming more customer-centric (Han et al., 2014). Consumer-centric CM can be achieved by analyzing shopper data and applying segmentation techniques (Nielsen et al., 2015). Hence, afterwards we discuss the relevant literature. In more detail, we can classify this pertinent literature in two broad categories: those focusing on shopper segmentation, and those that analyze shopper data and focusing on the associations between the products a shopper purchases during a visit (also known as market basket analysis). Hence, our second objective is to examine and understand existing shopper segmentation systems. And our third goal is to examine market basket analysis (MBA) literature and spot if such practices can be utilized to serve visit segmentation. Then, our fourth goal is to delve deeper into the existing literature and seek for the factors that affect the input, the design and the results of such segmentation systems and approaches.

### 2.1.1. Category management

Category management (CM) has its roots in early 1990s and up until now it is a fundamental concept in retail. There are various definitions for CM (American Marketing Association.; Nielsen Marketing Research, 1992), a mainstream definition comes from Nielsen that defines CM as “the process that involves managing product categories as business units and customizing them on a store-by-store basis to satisfy customer needs” (Nielsen Marketing Research, 1992).

CM plays a critical role in retailing as it is designed to aid retailers give the right product, at the right place and time, having the right promotion at the right price (Gruen and Shah, 2000). CM is the starting point between retailer and consumer interaction (Aastrup, Grant and Bjerre, 2007). In recent years, as customers are becoming far more demanding, there is a shift in attitudes between suppliers and retailers; thus, CM is viewed as a joint process between these two parties (Dupre and Gruen, 2004; Aastrup et al., 2007). In today’s competitive environments, they develop collaborative plans to boost categories, maximize profits and ensure a good long-term customer relationship (Dhar, Hoch and Kumar, 2001; Han et al., 2014). This happens as suppliers have more insights than retailers and stronger capabilities to structure CM plans, and it eliminates inefficiencies.

The CM planning process includes the following steps (Grossi, Harris, Joint Industry Project on Efficient Consumer Response and Partnering Group, 1995; Desrochers and Nelson, 2006). (A) Category definition to determine the products that will constitute the category and the formulation of the different sub-categories (Aanen, Vandic and Frasincar, 2015). (B) Identification of category role, to assign a purpose to the category, (C) category assessment to define opportunities, (D) category scorecard to establish category's goals (E) declaration of category strategies to support marketing and supply chain decisions. The last phase is critical and includes the (F) establishment of category tactics and then the implementation of the plan. CM tactics may include (Hübner and Kuhn, 2012) assortment planning, store layout planning, space allocation, pricing, promotional activities and logistics planning (Lindblom and Olkkonen, 2008).

CM can be divided into product-centric CM and consumer-centric CM (Han et al., 2014). Product-centric CM focus on the product and its attributes, whereas, consumer-centric is focusing on shoppers and their needs (Han et al., 2014; Nielsen et al., 2015). In the early 1990s, category managers focused solely into product-centric CM, they investigated the data and what the numbers revealed looking into their product category under examination (Nielsen et al., 2015). For example, in the fast-moving consumer goods (FMCG) domain there are different category managers e.g. for cereals, diary product, chocolates, coffee etc. and each category is treated separately. As a result, both practice and research focused on product-centric CM tactics. For instance, there are many examples for product-centric assortment (Chernev, 2003), assortment planning (Lotfi & Torabi, 2011) and optimization (Papakiriakopoulos, Pramatari and Doukidis, 2009). Likewise, many product promotions are also based on product-centric concepts e.g. "buy one, get one free" (BOGO) from the same category.

Here, it is notable to mention that suppliers also play a critical role in successful CM practices, as the product characteristics and attributes are shaped by them (Lindblom and Olkkonen, 2008). Similarly, many BOGO actions are a result of suppliers' promotional activities.

There are a few researchers that highlight the need to manage categories based on cross-category relations as a shopper typically purchases from multiple categories. For example, Song and Chintagunta (2006) and Kamakura and Kang (2007) perform analysis on sales data to identify cross-category effects. Similarly, Cil (2012) and Beck and Rygl (2015) perform market basket analysis to identify associations between products e.g. diapers → beers. However, the biggest change in CM, is that over the years it is moving from product-centric to consumer-centric (Han et al., 2014).

Consumer-centric CM could be achieved by analyzing shopper data and applying segmentation techniques (Nielsen et al., 2015). Shopper/customer, segmentation is the process of dividing customers into groups having similar behavior, and it is used to manage customer preferences more efficiently (Hong and Kim, 2012) (see section 2.1.2 for more details). In this spirit, some researchers, pinpoint that we should manage categories based on shopper needs and behaviors (Dhar et al., 2001; Desrochers and Nelson, 2006; Han et al., 2014; Nielsen et al., 2015; Griva, Bardaki, Pramatari and Papakiriakopoulos, 2018). For instance, Hesková (2006) proposes that the actual determination of the category should start via defining customer needs. Consumer-centric CM enhances the right selection of products that are used to shape effective CM tactics based on consumer understanding and needs (Nielsen et al., 2015). For example, Desrochers and Nelson (2006) propose to add customer behavioral insights into CM to improve assortment planning.

As shopper segmentation seems to be the cornerstone in consumer-centric CM, below we discuss shopper segmentation.

### 2.1.2. Shopper segmentation

Retailers like Tesco, Metro and Wal-Mart have recognized the need of data-driven decision-making. Mainly, they utilize business analytics tools to gain a competitive advantage in areas such as marketing e.g. cross-selling, in-store behavior analysis, customer segmentation and multi-channel experience (European Commission, 2014). Among their greatest endeavors is to identify the different customer groups visiting their stores, understand the specific needs and preference of each segment, and offer suitable services with a view to satisfy them e.g. by tailoring their marketing mixes (Boone and Roehm, 2002).

Shopper segmentation is a traditional and fundamental concept in marketing (Wilkie, 1978) and it is defined as the process of splitting heterogeneous shoppers, into homogeneous groups. Shoppers within each segment have the same, or similar characteristics and can be satisfied by similar marketing mixes (Hong and Kim, 2012). Shopper segmentation is vital nowadays, as consumers have become more demanding asking for personal retail services tailored to their needs and desires and not to the mass market. Thus, retail companies should become more shopper-centric and precisely reach their audience via providing services that suit the specific needs and preferences of the different shoppers (Anderson et al., 2007; Griva et al., 2018). There are plenty of studies using different datasets to segment shoppers into groups and the availability of new sources of consumer data (e.g. sensed data, social media posts etc.) forward the shopper segmentation research.

Researchers have responded to the retailers' interest for effective customer segmentation and many studies have appeared that utilize various kinds of data. Customer segmentation is the process of dividing heterogeneous customers into homogeneous groups on the basis of common attributes and is essential for handling a variety of customers with rich sets of diverse customer preferences more efficiently (Hong and Kim, 2012).

Until now, researchers performed shopper segmentation either using data related to (A) shopper characteristics (Cui, Wong and Lui, 2006; Hong and Kim, 2012; Miguéis, Camanho and Falcão e Cunha, 2012) such as: (i) Demographic data, e.g. gender, age, marital status, household size etc.; (ii) Geographic data, e.g. city/country of residence, or shopping etc.; (iii) Psychographic data, e.g. social class, lifestyle and personality characteristics etc.; (iv) Attitudinal data, i.e. perceived data gathered from surveys that capture information about what people say they do in order to understand and interpret shoppers' behavior (Woodside, 1973; Konuş, Verhoef and Neslin, 2008);

Or using (B) data that indicate shopping behavior (Griva et al., 2018) i.e. behavioral data. We classify the behavioral data in the following categories: (i) data that indicate the content of a basket; for example, the product categories it contains e.g. 1-liter milk, 300 grams cheese, or the attributes of the contained products e.g. biological, gluten free, diet or light. (ii) Basket characteristics such as volume i.e. the number of items a basket contains, value i.e. the cost of a basket, and the variety of products it contains e.g. milk, cheese. (iii) visit characteristics e.g. visit frequency, duration, or even sources of visit and aisles or pages visited.

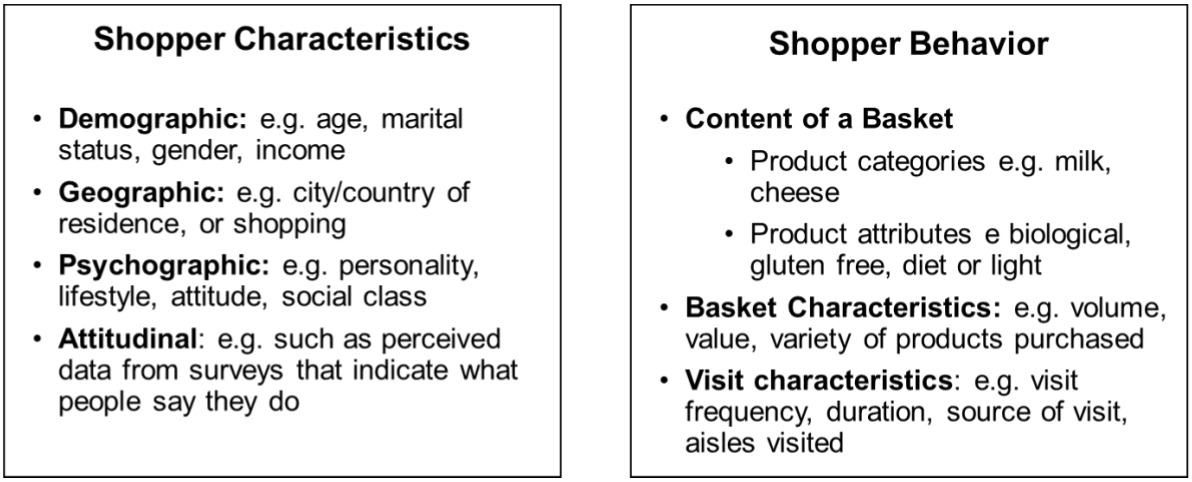



**Figure 2-1. Shopper segmentation data**

Researchers utilize different data mining models for customer segmentation using sales data, such as models based on associations (e.g. association rules, Markov chains), classification (e.g. neural-networks, decision trees), clustering, sequence discovery, forecasting (e.g. neural-networks) (Ngai, Xiu and Chau, 2009). In all these empirical works, researchers utilize customer-level sales data in order to segment shoppers and examine their purchase behavior. In other words, they either examine their basket and visit characteristics (e.g. basket volume, visit frequency etc.) or the content of their basket i.e. the mix of the products that shoppers have purchased in their whole purchase history e.g. during all their visits in a physical or web store of a retailer. For instance, a stream of studies that focuses on shopper behaviors, utilize sales data to segment shoppers based on their LTV or CLV (Customer Lifetime Value), mainly using RFM (Recency, Frequency, Monetary) and clustering analysis (Dwyer, 1989; Cheng and Chen, 2009; Chen, Kuo, Wu and Tang, 2009; Khajvand, Zolfaghar, Ashoori and Alizadeh, 2011; Aeron, Kumar and Moorthy, 2012).

On the other hand, there are also studies that focus on the mix of the products or product categories that shoppers have purchased in their whole purchase history (content of a basket). For instance, (Park, Park and Schweidel, 2014) propose a modeling framework for customer base analysis in a multi-category context to predict customer purchase patterns. To this end, a beauty care company in Korea provided sales data that concern both shopping behavior and categories mix. Furthermore, statistical methods e.g. Markov chains, Euclidean distances, are utilized to model the time between a customer's purchases (interarrival time) at the firm and the product categories that comprise a shopping basket. In another study, (Miguéis et al., 2012) propose a method for market segmentation in retailing based on a customer's lifestyle, supported by information extracted from a large transactional database. They analyze the product categories shoppers have purchased from a European retailing company. Using clustering, they propose promotional policies tailored to the customers of each segment, with the purpose to support loyal relationships and increase sales. In addition, (Han et al., 2014) showcasing the role of categories in customer segmentation, they compared different techniques and performed clustering using k-means in customer-level sales data to segment shoppers. In their segmentation approach they identified customer segments e.g. customers who purchase routine, seasonal or convenience categories.

From a different perspective, Liao, Chen and Hsieh (2011) utilized sales data and data collected via questionnaires from shoppers who purchase skincare and cosmetic products to segment customers into clusters, according to their lifestyle habits and purchasing behavior. By adopting clustering and association rules, they provide suggestions and solutions for direct marketing to design possible new services and sales for each customer segment. (Boone and Roehm, 2002) utilize sales data that concern shopping behaviors (e.g. orders, spending, days since last and first purchase etc.) provided by a retailer, and other artificial data, in order to examine the use of artificial neural networks (ANNs) as an alternative mean of segmenting retail databases. Their results indicate that ANNs may be more useful to retailers for segmenting markets because they provide more homogeneous segmentation solutions than mixture models and k-means clustering algorithms. In addition, (Kitts, Freed and Vrieze, 2000) developed an algorithm to analyze a customer's purchasing history provided by an on-line and catalogue hardware retailer, in order to provide item-level recommendations and promotions. (Liao and Chen, 2004) combine various kinds of data, such as sales data regarding categories mix, demographics, and attitudinal data collected via questionnaires, and use a business analytics approach to segment customers, to enhance the effectiveness of direct marketing and sales management in retailing, and more specifically to format electronic catalogues.

### 2.1.3. Market basket analysis

The second group of studies focus on the baskets of the shoppers and it looks for associations between the items/products a shopper purchases during a visit. A famous example is that of diapers and beers in Wal-Mart stores. These studies perform market basket analysis (also known as association rule mining or MBA), which is a data mining method that examines large transactional databases to determine which items are most frequently purchased jointly (Agrawal, Imieliński and Swami, 1993; Srikant and Agrawal, 1995). Although, MBA is originated in the marketing field, many extensions of this method have been proposed and it has been widely used in various fields varying from bioinformatics, nuclear science and immunology to strategic management and organizational behavior etc. (Aguinis, Forcum and Joo, 2013). At the same time, MBA has been applied in various domains, such as finance, telecommunications, retailing etc. (Chen, Tang, Shen, & Hu, 2005). For instance, in the retail domain, Cil (2012) introduces a framework that identifies the associations among the purchased categories in a supermarket. These associations between product categories reveal “consumption universes” and are utilized to change the store’s layout. For analysis purposes, he utilizes sales data and the categories mix provided by a Turkish supermarket. Furthermore, Tang et al. (2008) introduce a new approach to perform market basket analysis in a multiple-store and multiple-period environment. They use sales data provided by twenty stores of a supermarket chain in Taiwan and propose purchasing pattern analysis at a detailed level of time and place, such as a combination of days and stores. Although variations of association rule mining have been proposed, certain characteristics of real world data hinders their performance when the algorithms have been designed and evaluated with artificial data sets (Zheng, Kohavi and Mason, 2001); thus, making the applicability in real world settings is crucial. The next section addresses significant issues and factors that affect such studies and are common in various retail contexts.

### 2.1.4. Business analytics and retail segmentation systems

Looking into the segmentation literature, we identified various factors of the data (e.g. basket variety and volume) and the retail case itself (e.g. shopping channel) that (A) prospective designers of segmentation systems should consider if they want to produce valid segments, (B) data scientist should take into account when manipulating and modeling data and (C) marketeers should take into account interpreting the shopper segmentation results. Below we discuss all these identified factors that affect business analytics and retail segmentation systems, and could be utilized to derive design principles for shopper segmentation in retail. Here, we should note that the direction and the magnitude of the factors is not examined.

#### 2.1.4.1. Shopper volume, variety, value, visits (shopper 4Vs)

***Volume***

Basket/visit size (or volume) is defined as the number of products a shopper purchases in a visit (Noble, Lee, Zaretzki and Autry, 2017). Volume could be influenced by various variables e.g. income, product, promotions (Kahn and Schmittlein, 1992; Noble et al., 2017), but it also affects segmentation results. Literature that examines basket volume often notes that there are data skewness and sparsity issues that require the elimination of some data from consideration (Cho et al., 2002; Cho and Kim, 2004; Griva et al., 2018). When we eliminate data and excluding outliers based on the volume feature, we should not overlook the impact that these exclusion on the quality of the segmentation results (Cho et al., 2002). Volume is also combined with variety (see next paragraph) and there are many studies utilizing basket volume, along with basket variety to perform segmentation. For instance, Yao, Sarlin, Eklund and Back (2012) used basket volume, basket variety and value along with other data such as demographics from a retail company to perform temporal customer segmentation.

***Variety***

Variety shows the total number of distinct product categories a shopper purchased or interacted with during his/her shopping visit. Thus, it is important to indicate how a product is defined. For example, in item/SKU level, or parent product category, or sub-category etc. Many researchers (Srikant and Agrawal, 1995; Cho and Kim, 2004; Videla-Cavieres and Ríos, 2014) imply poor results in their segmentation and market basket analytics approaches due to sparsity issues caused by the variety feature; thus, they have tried to tackle this issue using various methods. For example, Cho and Kim

(2004) merge sub-categories based on their purchases and re-execute their analysis in higher product granularity level to increase basket variety. Likewise, Srikant and Agrawal (1995), produce association rules for every product granularity level and prune the redundant ones. Closing, here we should mention that, visit volume and variety are often correlated. Also, when they have low values, they are causing data skewness and sparsity issues; thus, these factors affect the customer segmentation results and quality (Cho et al., 2002).

### *Value*

Much work has been focused on defining value and how it affects shopper segmentation approaches. Value is either considered as shopping trip/basket value, or as shopper value. Different studies utilize shopper value i.e. total value spent by a customer e.g. CLV (Customer Lifetime Value) (Gupta et al., 2006; Homburg, Steiner and Totzek, 2009; Aeron et al., 2012), to perform segmentation. Another well-known example is RFM (Recency, Frequency, Monetary value) (Khajvand et al., 2011), where monetary value is combined with other variables to segment shoppers. Closing, the basket value is utilized and affects the results of customer segmentation approaches and sometimes it is used for outlier elimination purposes (Griva et al., 2018).

### *Visits*

Visit frequency, recency, number of customer visits and the time between them are also important and affect customer segmentation approaches (Griva et al., 2018). For instance, Park et al. (2014) segment shoppers using statistical methods to model the time between purchases. Visits is utilized along with other factors for segmentation purposes; for example, Boone and Roehm (2002) utilize sales data e.g. orders, value, days since last and first visit etc. to examine the use of artificial neural networks (ANNs) as an alternative mean of segmenting shoppers.

#### 2.1.4.2. Loyalty programs and cards

Retailers consider loyalty programs and cards as tools to develop marketing strategies. Loyalty cards are not only utilized for customer retention purposes, but also are viewed as an additional mean to collect data about shoppers' behavior (Demoulin and Zidda, 2009). Loyalty programs play a vital role in retailing as are utilized to monitor and influence consumer choices. Via these programs, retailers offer benefits and encourage consumers use the service and/or continue shopping to receive rewards and reach a higher level. Likewise, firms that can potentially gain more repeat businesses and, gather detailed consumer insights that allow them to target customers with tailored marketing activities (Yuping Liu, 2007; Breugelmans et al., 2015). An important fact is shoppers card adoption i.e. percentage of shoppers use the card. Low rates might be a result of badly designed programs, that require much effort from customers. In addition, equally important is the card penetration rate in each shopper's visits i.e. percentage of purchases made using loyalty card. Plenty of studies utilize loyalty card data and customer loyalty id as input into the segmentation analysis (Kitts et al., 2000; Reutterer, Mild, Natter and Taudes, 2006; Liao et al., 2011; Miguéis et al., 2012; Yao et al., 2012; Han et al., 2014; Park et al., 2014); as a result, the existence and the adoption of loyalty cards and programs seems that may affect, amplify or could be a barrier in segmentation approaches. In many cases, (e.g. Chen, Chiu and Chang, 2005) researchers are not able to perform customer segmentation and mine changes in customer behavior without having loyalty cards data, and thus customer ids. On the other side, there are researchers (Griva, Bardaki, Sarantopoulos and

Papakiriakopoulos, 2014) who claimed that significantly low percentage of loyalty card penetration in the period covered by the given sales data lead to exclude a large amount of shoppers; thus they doubt about the quality of the segmentation results. Finally, there are studies (e.g. Cil, 2012), in which the existence of loyalty cards data seems that it could really enrich the results.

#### 2.1.4.3. 4Ps (product, price, promotion, place)

The marketing researchers and practitioners, being the ones mostly concerned with consumer behavior, have devised the fundamental model "marketing mix" – 4Ps (product, price, promotion, place) prescribing what factors should be considered when studying consumer behavior and segments.

***Product***

*Product, brand (and its price)* are important factors that should be taken into consideration in segmentation approaches (Lockshin, Spawton and Macintosh, 1997). Lots of segmentation studies are based on the product mixes customers purchase. The product is either treated in brand, or item, or in parent product category, or sub-category level etc. Researchers have claimed poor results, when there is not a right selection on the granularity level we define the product. Thus, **we should not overlook the significant role of the product taxonom**y (see Figure 2-2 for an example) in such data analytics studies, since it may affect the knowledge discovery process and the data mining results (Cho et al., 2002). The study of the impact of product taxonomy on data mining is mainly found in the recommendation systems literature and in the semantic web literature. Many researchers emphasize that it is critical to find the right product category granularity level, because it could affect association rule results and, thus, the whole recommendation system (Albadvwe & Shahbazi, 2009; Cho & Kim, 2004; Cho et al., 2002; Han et al., 2014; Hung, 2005; Kim, Cho, Kim, Kim, & Suh, 2002; Srikant & Agrawal, 1995). Existing approaches handle this issue by examining (A) the product items a customer purchases or interacts with at a stock-keeping-unit (SKU)/item level (Kim, Kim, & Chen, 2012). However, selecting a low grain in a product taxonomy tree, state high dimensionality issues and problematic results (Kimball and Ross, 2013); or (B) examining product categories (e.g. beverages, breads, orange juices) based on the granularity level as indicated in the product taxonomy (e.g. level/height=3 in Figure 2-2) (Cil, 2012; Videla-Cavieres and Ríos, 2014). However, selecting a higher-level grain, they limit their study to fewer and less detailed dimensions (Kimball and Ross, 2013).

In addition, others utilize a cross-category level as indicated by marketers or domain experts (e.g. shaded nodes in Figure 2-2) (Albadvwe & Shahbazi, 2009). To the best of our knowledge, only Cho & Kim (2004) and Srikant & Agrawal, (1995) propose an algorithmic logic to define the right granularity level of product taxonomy. On the one side, Cho & Kim (2004) define the right granularity level by selecting cross-category levels and merging some categories based solely on product purchases (e.g. merging socks and skincare). On the other side, Srikant & Agrawal, (1995) propose producing associations between items at any level of the taxonomy and pruning redundant rules in order to address issues in the product taxonomy.

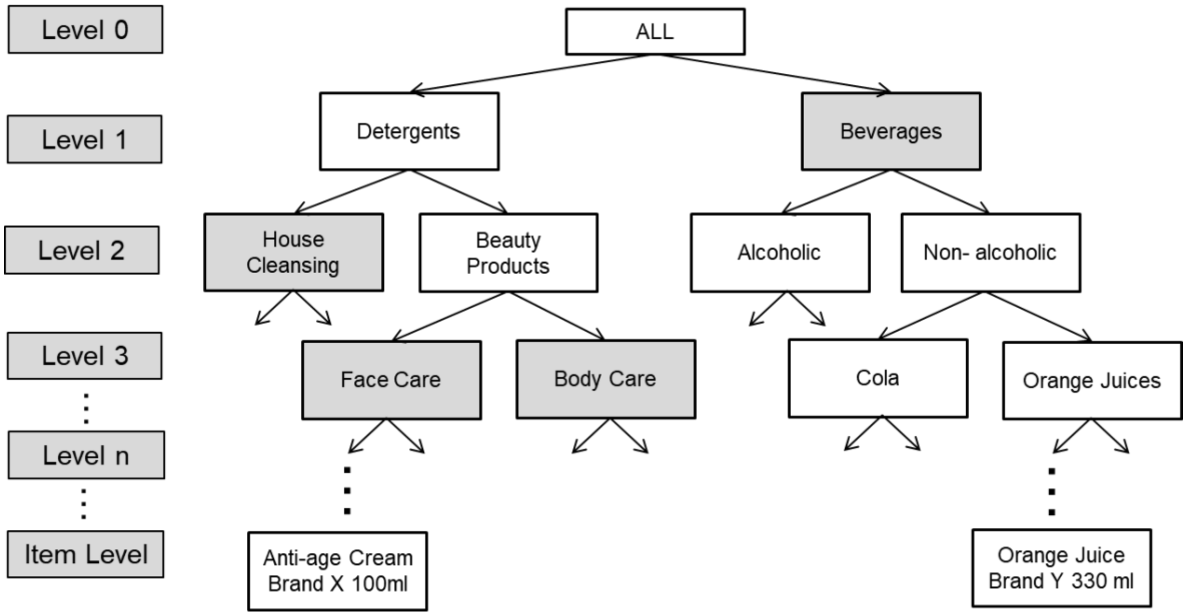



**Figure 2-2. Product taxonomy example**

Apart from the recommendation systems literature, the problem of defining the right granularity level is also met in the semantic web literature. In this case, an ontology merging and mapping on products over the different product classification taxonomies is required. This is based solely on product semantics (e.g. merging of books and humor books) and it could be vital for product-comparison sites and recommender systems (Park et al., 2014; Aanen et al., 2015).

***Price***

Variations in household incomes leads to segmenting some markets along a price feature. Price segmentation most commonly is met in markets with particular products (e.g. durable products such as cars, premium products). For example, Thach and Olsen (2015) performed price segmentation for wine shoppers. Similarly, Liu, Liao, Huang and Liao (2018) propose a multicriteria segmentation that uses factors such as customer preferences and product factors e.g. price to perform segmentation for a car seller. Additionally, price factor is important when we have significant variations in prices between the products a retail store sells. For instance, a store selling both expensive (e.g. televisions) and cheap products (e.g. USB sticks). To effectively segment shoppers, we should combine this factor with others e.g. variety. Moreover, segmenting shoppers based on price factor is important for shopper marketing purposes. For example, Bell and Lattin (1998) revealed that shoppers purchasing more expensive products prefer EDLP (everyday low price) strategies. Closing, in existing literature it seems that the price of a product it is usually combined with other factors and it is commonly utilized in more particular markets.

***Promotion***

Many researches have highlighted the relationship between promotions and shopping behavior (Kahn and Schmittlein, 1992). Results have indicated that promotions affect shopper segmentation. For instance, Lockshin et al. (1997) highlight that response to promotions is different among the derived shopper segments. Thus, when we examine a promo-oriented retail context we should examine and consider any data indicating promotional behavior. The lack of such data may lead to misleadingly segments.

***Place***

We translate the place into the channel and the store type or format in shopper segmentation literature.

*Channel*: Shopping via multiple channels is a rapidly growing phenomenon. With companies continually adding new channels, retailers face constraints such as channel integration difficulties (Beck and Rygl, 2015). However, multiple channels assist retailers to augment their core product offerings and expand service outputs (Sands, Ferraro, Campbell and Pallant, 2016). This has lead consumers facing a two-dimensional decision at their path to purchase: which firm and which channel to interact with (Neslin et al., 2006). There are studies that segment consumers based on their multichannel behavior and on channel usage, providing different segments across several stages of the buying process. For example, Konuş et al. (2008) proposed distinct multichannel consumer segments based on the importance of stores, the Internet and catalogs at the search and purchase stages. Likewise, Nakano and Kondo (2018) utilized purchase scan panel data from physical and web stores. Both these works extracted different shopper segments within each channel. As literature indicates that diverse shopper segments could arise in different channels, data scientist should build distinct shopper segmentation models per channel and examine each channel separately.

*Store type/ format*: Apart from the different behavior of shopper across different channels, shopping behavior of consumers also differs across different store types and formats depending on shopping situations i.e., fill-in or major trips (Bell et al., 2011).; thus, different shopping patterns may result from different store formats (Gijsbrechts, Campo and Nisol, 2008). Store format also affects other important segmentation factors such as basket value (Klein and Schmitz, 2016). Although, store format is an important factor, research on cross-format shopping patterns, and more specifically on the distribution of consumers' shopping basket among different retail formats, has been largely ignored (Skallerud, Korneliussen and Olsen, 2009; Baltas, Argouslidis and Skarmeas, 2010). The different cross-format shopping behaviors of consumers may affect the customer segments and their characteristics; hence, store format seems to be an important factor that should be taken into consideration in shopper segmentation. For example, there are studies (Sarantopoulos et al., 2016; Griva et al., 2018) that declared different shopping visit segments within different store types such as convenience stores, supermarkets and hypermarkets. Hence, data scientists, should build different shopper segmentation models per store format/type.

Figure 2-3 includes the factors that researchers have implied that affect the shopper segmentation process and results. These factors are either related to the retailer (i.e. 4Ps), or to the shopper (i.e. shopper 4Vs), or to both (i.e. loyalty programs). We inspired the term shopper 4Vs from the data 4Vs i.e. variety, volume, velocity, veracity.

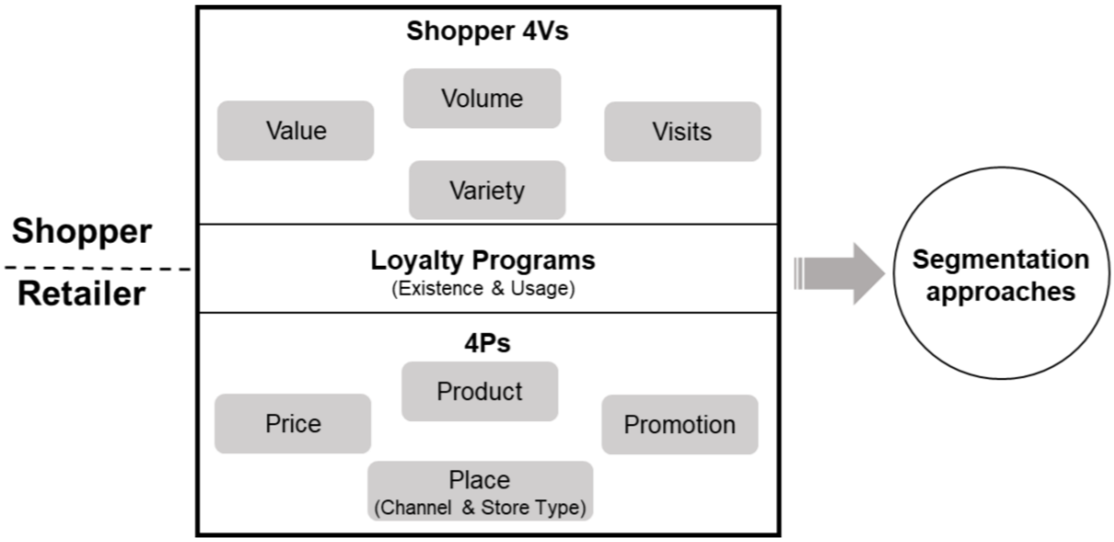



**Figure 2-3. Factors affecting segmentation approaches**

Table 2-1 presents the definition of each factor reflecting the available literature and Table 2-2 presents each factor, and how (according to our research) it affects the various shopper segmentation studies.

| Factor | Definition |
|---|---|
| Volume | Is defined as the number of products a shopper interacts with e.g. purchases in a single visit (or during his/her purchase history). |
| Variety | Shows the total number of distinct product categories purchased in each basket/or purchased by each shopper. |
| Value | Is either considered as shopping trip/basket value, or as customer value. |
| Visits | Has two definitions: First, it refers to the number of visits a shopper performs at a retailer's physical or web store. Secondly, it refers to the time between each shopper's visits. Each time we use this feature we clearly declare how it is defined. |

| | |
|---|---|
| Loyalty Program | It refers to the existence of loyalty cards in retailer's store and the card penetration percentage i.e. the percentage of purchases made using a loyalty card. |
| Price | Refers to the price a product is sold. |
| Product | The product is either treated in brand, or item, or parent product category, or sub-category level etc. Some researchers define product in a customized product category level. |
| Promotion | Refers to any type of marketing communication used either in or outside of the retail store, using several means. |
| Place | Is either defined as different shopping channel e.g. web, physical store, or as different retailer's store type/format. |

**Table 2-1. Factors' definitions**

| Factor | How it affects shopper segmentation | Related/affected works |
|---|---|---|
| Volume, Variety | Volume is often correlated with variety feature. Low variety cause data sparsity and skewness issues and lead to poor results in segmentation and market basket approaches. | Srikant and Agrawal, 1995; Cho and Kim, 2004; Yao et al., 2012; Videla-Cavieres and Ríos, 2014 |
| Value | It mainly affects the input of the CLV and RFM approaches and the output of the segmentation results. | Gupta et al., 2006; Cheng and Chen, 2009; Y. L. Chen et al., 2009; Khajvand et al., 2011; Aeron et al., 2012; Yao et al., 2012 |
| Visits | It affects approaches using as input visits frequency, recency, number of customer visits and the time between them (interarrival time). Also, visit recency is used and affect RFM approaches. | Boone and Roehm, 2002; Cheng and Chen, 2009; Y. L. Chen et al., 2009; Khajvand et al., 2011; Park et al., 2014 |
| Loyalty Program | Is viewed as an additional mean to collect data about shoppers' behavior. Many segmentation approaches require shopper's loyalty card. The lack of this feature or low card penetration either causes poorer results or makes shopper segmentation approaches unable to produce results. | Kitts et al., 2000; Liao and Chen, 2004; Reutterer et al., 2006; Liao et al., 2011; Cil, 2012; Miguéis et al., 2012; Yao et al., 2012; Griva et al., 2014; Han et al., 2014; Park et al., 2014 |
| Price | It mainly affects segmentations in markets with particular products, e.g. cars, wines, more premium products. It seems that the price factor it is usually combined with other factors. | Thach and Olsen, 2015; J. Liu et al., 2018 |
| Product | The results of market basket analysis and similar segmentation approaches are affected by the granularity level we define the product. Wrong product definition may result data | Albadvwe & Shahbazi, 2009; Cho & Kim, 2004; Cho et al., 2002; Han et al., 2014; Hung, 2005; Kim, Cho, Kim, Kim, & |

| | | |
|---|---|---|
| | skewness issues, poor data mining model performance, and poor results. | Suh, 2002; Liao et al., 2011; Miguéis et al., 2012; Tang et al., 2008 |
| Promotion | Results have indicated that response to promotions and marketing actions affect shopper segmentation and differentiates between the derived clusters of shoppers. The lack of such data in promo-driven retail contexts may result false positive results. | Lockshin et al., 1997 |
| Place | Different channels and store types result dissimilar shopper segments. Also, literature has proved the shopping visit segments across different store types vary significantly. | Konuş et al., 2008; Sarantopoulos et al., 2016; Griva et al., 2018; Nakano and Kondo, 2018 |

**Table 2-2. Identified factors VS affected works**

#### 2.1.4.4. Data 4Vs: volume, velocity, variety and veracity

Here, we should also admit that volume, velocity, variety and veracity of data affect every data analytics process (Abbaswe et al., 2016). Hence, the same happens in shopper segmentation research. Velocity is the rate at which new data is generated. According to (McAfee and Brynjolfsson, 2012) “Walmart collects more than 2.5 petabytes of data every hour from its customer transactions”. Thus, as volume and velocity has far outstripped the capacity of manual analysis (Chang et al., 2014) many technical issues arise and sophisticated data infrastructures and techniques are required to manage the enormous data volumes (Goes, 2014). Additionally, require dynamic and more sophisticated segmentation approaches. To tackle such issues Reutterer et al. (2006) propose a dynamic segmentation of shoppers enrolled in a loyalty program of a “do-it-yourself ” retailer.

Regarding variety, data no longer come from one single source and in one format they could be structured, semi-structured, and unstructured (Abbaswe et al., 2016). From the other hand, using different data sources aid businesses obtain a multifaceted view about their customers. Thus, as mentioned before, in retail context different datasets such as sales data from the omnichannel environment, loyalty cards data, demographics, geographic, attitudinal and behavioral data could be utilized and combined to identify customer segments  (Boone and Roehm, 2002; Liao and Chen, 2004; Boztuğ and Reutterer, 2008; Chen et al., 2009; Khajvand et al., 2011; Aeron et al., 2012; Miguéis et al., 2012; Yao et al., 2012; Rust and Huang, 2014). As explain above, the utilization of these sources (e.g. loyalty cards data) affects the segmentation results.

Validating the veracity of the data, sorting out the noise from valid information has been and will continue to be a major issue in  big data research (Goes, 2014). Data are not always clean and complete; thus, they must be consolidated and cleaned to analyze them, extract insights and make the right decision. In the retail context, different data issues might arise. Apart from inconsistencies in the sales data, several data issues should be tackled. For example, demographics data might be inaccurate, as shoppers might deliberately provide wrong data (Chahal, 2015) for instance in attributes such as age, income, household size etc. For instance, Griva et al. (2014) claimed inconsistences and poor quality in the given demographics data such as age and household size, so they omit them from their segmentation and utilized solely sales data. Additionally, these features even if the current quality is adequate, could be declared as “slowly changing features”, as they are altering over the time e.g. people are getting married, salaries may grow etc. (Kohavi, Mason, Parekh and Zheng, 2004); thus, a segmentation approach that will be conducted in the future, it might incorporate false data.

## 2.2. Consumer Insights

In this section, our goal is to better understand and conceptually define shopping mission concept. For that reason, we asked for consumer's opinion and point of view via a series of focus groups and via a survey.

### 2.2.1. Focus groups

Our exploratory study included a series of eight focus group discussions with 71 shoppers. Our goal was to better understand and conceptually define shopping missions in the FMCG domain. Each session included 8-10 shoppers and lasted 45 minutes. Discussions were designed to elicit insights from participants in relation to how they schedule and organize their shopping trips, how they perceive the organization of product categories in the stores they visit. Focus groups are generally more suited for exploratory research (Belk, 2007; Calder, 2011); thus, this study attempts to build a holistic understanding of the shopping mission concept.

Discussants were randomly contacted via telephone and after an initial screening, we offered a voucher for their participation. Participants provided consent to videotape the discussions, and all recordings were subsequently transcribed. Discussions were guided by a semi-structured group interview guide and were moderated by a group leader. We structured the focus groups into three generic sections: (A) discussion over shopper profile asking questions such as age, marital status, the main shopper of the household, times visiting the store during a week etc., (B) discussion regarding the usage of product list during a store visit and (C) discussion related to shopping mission concept. Concerning the latter, we asked participants to recall the last time they visited a grocery store. Afterwards, we asked them to indicate the products they purchased and to denote whether they group these products based on their shopping needs.

Regarding the demographics of the participants the 30,3% of them were men and the rest (69,7%) were women. According to experts' opinion, this percentage follows the typical ratio between women and men shoppers in these FMCG stores. Regarding the age range, we interviewed shoppers from 22 to 68 years old. We weighed our sample according to experts' suggestions who are aware of the age distribution of the typical shoppers.

Almost all the participants confirmed that they visit the store three to four times each week and they purchase a narrow variety of products to cover their short terms needs. A married woman in her early forties said, *"I visit the supermarket 3-4 times a week. I visit retailer X for cleaners and stuff like that and retailer Z for fresh products"*. Similarly, a single man in his thirties stated that: *"I visit a store next to my work 2-3 times a week according to my needs. Usually, I am going there after work at around 7 to 8 during the afternoon"*. These findings are a first indication of the existence of the shopping mission concept, as it seems that modern shoppers enter the stores purchasing a narrow variety of products that satisfy their temporary needs.

Regarding the shopping missions that shoppers execute in the various store types, shoppers stated that they visit the larger stores such as hyper stores for their weekly and more abstract visits. They visited these stores mainly during Saturdays. A married man said, *"I visit the store once a week mainly during Saturdays... with my wife we purchase everything for the whole week"*. Whereas, they execute more targeted visits in the smaller stores such as supermarkets and/or convenience stores. In more detail they stated that they visit the convenience stores mainly for immediate consumption needs such as “breakfast”, “snack”, “heathy snack”, “light meal”, “food-to-go”, “soft drinks and alcohols”. A young man mentioned the following, “*In the cases where I do not have home cooked meal in my work and I am in a hurry, I visit the convenience store in the next corner and buy a wrap or a sandwich and something to drink…I would call it food-to-go*”.

Also, shoppers stated that they visit larger stores e.g. supermarkets or hyper stores to purchase products for “sweet preparation”, “gourmet meal”, “house cleansing” and “personal hygiene”, “semi-prepared food”, “biological products”, “baby products”, and “meal preparation”. A woman in her early thirties mentioned “*Sometimes after work I visit the store nearby to buy products to prepare dinner for today or meal for tomorrow”.* Similarly, a married woman said, “almost once every two months I visit the X supermarket at the suburbs to purchase product for myself such as face care*, body care, make up, hair colorants etc. … I would call these visits ‘my beauty visits’ (chuckles)”.*

Regarding the usage or not of a shopping list, we noticed that shoppers who claimed that they visit the store having a shopping list, tend to perform more abstract shopping missions, or they do not confirm the existence of the shopping mission concept e.g. *“I do not visit the store for something specific, I just want to purchase everything for the week”, “I have plenty of time* (a retired man said), *almost every day I write down what is missing and I visit the store at the next corner (of my house)…I could buy tissues, cheese and fruits or just one product each time…but not for something specific”.*

### 2.2.2. Survey

To identify the percentage of shoppers that visit a retail store having a specific shopping mission in mind, we used a survey. In more detail, we included three relevant questions in the context of a wider survey which was distributed in different retail stores all around Greece. In the first question, we asked shoppers about the shopping missions they execute in the different retail stores. The shopping missions we utilized were those that the focus group discussions indicated us i.e. “breakfast”, “snack”, “soft drinks and alcohols, “sweet preparation”, “food-to-go”, “house cleansing”, “personal hygiene”, “semi-prepared food”, “baby products”, and “meal preparation”. Apart from the available shopping missions, shoppers also had the opportunity to select that they do not visit the store having a specific mission in mind. In the second question, we asked shoppers about the how frequently (rating from 1 to 7) they use or not of a shopping list during their visits.

We randomly distributed the questionnaires to 1903 shoppers, visiting various supermarket stores all around Greece. Results indicated that the clear majority of shoppers (85,5%) answered that they visit the supermarket stores having a specific mission in mind. Results also revealed that there is a statistically significant difference regarding the usage of shopping list between those shoppers that are visiting the store having a shopping mission and mind, and those who do not. In more detail, via running ANOVA we identified that the significance value is 0,022 (i.e., $p = 0{,}022 < 0{,}05$). Thus, we validated shopper’s statement during the focus groups that indicated that those who do not confirm the existence of shopping mission, tend to use a shopping list.

## 2.3. Research Gaps

Below we present all the research-related gaps according to the literature review conducted in the previous sections.

**Shopper segmentation**

Overall, the studies presented in literature review section show that researchers have applied different business analytics approaches to shopper-level data to produce shopper segments. They divide the customers into groups based either (A) on their complete shopping behavior in terms of basket volume, visit frequency etc. (basket and visit characteristics), or (B) on the mix of products or product categories (contents of a basket) recorded in their total purchase history (Aeron et al., 2012; Boone & Roehm, 2002; Chen et al., 2009; Cheng & Chen, 2009; Han et al., 2014; Khajvand et al., 2011; Kitts et al., 2000; Liao & Chen, 2004; Liao et al., 2011, Park et al., 2014).

However, modern shoppers are changing their behavior over time, so we cannot talk any more about shopper segmentation. We state that the aforementioned studies overlook the holistic shopping purpose, intentions and missions of shoppers, which are not the same in every (physical or web) store visit. Sharing other researchers (Walters and Jamil, 2003; Bell et al., 2011) concerns, in this new era we should put the shopper visit on the spot, instead of the shopper behavior that changes over time and that traditional shopper segmentation relies on.

**Market basket analysis**

In contrast, there is a group of scholars analyzing sales data per visit (basket-level) to identify associations between products (e.g. Agrawal et al., 1993; Cil, 2012; Srikant & Agrawal, 1995; Tang et al., 2008). In other words, they do not look for shopper segments, but they focus on pairs of products the customers purchase together more frequently (e.g. diapers and beers in the famous Wal-Mart stores' study, or in another example eggs →milk). Although, these studies examine the product association in basket/visit level, still they overlook the shopping purpose of each shopper visit.

**Product taxonomy**

The right product category level, i.e. the right level of analysis in the product taxonomy tree, is crucial to the results of the study, it may affect the knowledge discovery process and the data mining results (Cho et al., 2002). At the same time, each retailer has its own product taxonomy and this taxonomy serves other purposes e.g. store replenishment, shelf space allocation, product assortment selection. Researchers who selected an existing level in retailer's product taxonomy, claimed very poor results in both the algorithms' accuracy and the business evaluation (Cho and Kim, 2004; Videla-Cavieres and Ríos, 2014). In more detail, on the one hand researches selecting a low grain in a product taxonomy tree, state high dimensionality issues and problematic results. On the other hand, those selecting a higher-level grain, limit their study to fewer and less detailed dimensions (Kimball and Ross, 2013). Therefore, the selection of grain affects the data mining results and it is important for the design science to choose the right level of analysis.

Studies that tackle this issue are divided into two groups: (A) Those that defined the right granularity level by selecting cross-category levels and merging some categories, based solely on product purchases (e.g. merging socks and skincare). (B) Those utilized in the semantic web, which take into consideration solely product semantics. However, the available studies show that there is no generic rule; the researchers select the product taxonomy level that better serves their research purposes (Cil, 2012;

Videla-Cavieres and Ríos, 2014). In the literature there is lack of algorithmic approaches that take into account both product semantics and product purchases to formulate custom categories that serve the data mining purposes. In other words, there is lack of approached to manage the feature space problem and absorb any anomalies with respect to identify a friendly context in order to undertake a data mining task.

Figure 2-4 depicts the research gap concerning the above aspects i.e. the scope of the analysis (shopper segmentation and MBA), and the product taxonomy. The shaded areas declare the research gaps, and the dark grey rectangle in the middle is the area that our research contributes the most. More specifically, the scope of analysis describes the extent to which market baskets are utilized to study a specific issue. On the one hand, researchers study the associations between products that customers purchase during a visit. On the other hand, they study and group baskets using the entirety of a customer's shopping visits. In our perspective, this dimension is shaped with a view to study the shopping purpose/mission of a single customer visit. Also, regarding the product taxonomy, as mentioned earlier, researchers utilize the trees' internal nodes (product categories) or the tree leaves (SKU level) depending on the scope of analysis. In our work, we adjust the original product taxonomy, often defined by a retailer for operational purposes, and produce customized product-categories, which can adequately support the visit segmentation.

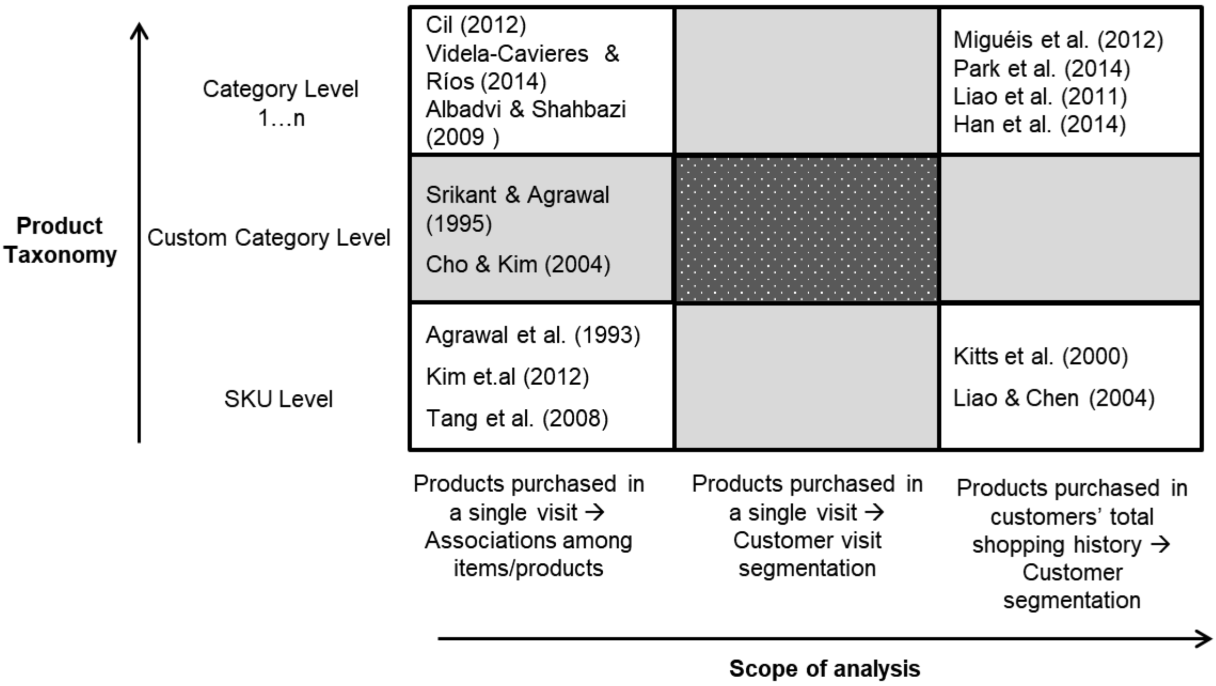



**Figure 2-4. Research gap**

## Category Management

Existing category management practices focus separately in each category e.g. milk, cereals, coffee etc. Business executives recognize the need to incorporate the shopper behavior, needs and missions into CM practices. Similarly , there are a few researchers (e.g. Song and Chintagunta, 2006; Kamakura and Kang, 2007, Han et al., 2014; Nielsen et al., 2015) that highlight the need to manage categories based on shoppers and their needs (consumer-centric CM).

However, in existing category management literature there are no such practices. Even consumer-centric CM is focusing merely on cross-category relations and not on shopper needs. Hence, retailers are still losing potential revenue due to their failure to get the right goods to the right places at the right price. As a result, incorporating shopper needs, missions and behaviors in CM, it is still an open issue in business and a research gap.

### Factors affecting segmentation approaches

Delving deeper into the segmentation literature we identified that there are studies mainly in the marketing domain, that discuss several factors that affect big data analytics systems in general. However, they do not present evidence of how these factors affected relevant segmentation cases. Also, in the IS literature there is a great majority of papers that perform shopper segmentation. Though to the best of our knowledge, authors describe their own case and not “the bigger” picture i.e. how system inputs and factors (e.g. data) affect and alter the segmentation process, system and results/outputs. It is only implied, and it is not discussed how different factors affected segmentation results. As it is obvious, there is a need to identify factors affecting segmentation approaches and to present evidence on how that happens.

### Business analytics approaches

Editorials (Agarwal and Dhar, 2014; Goes, 2014), other academic papers (Abbasi et al., 2016; Müller et al., 2016; Delen and Zolbanin, 2018) and practitioners (McKinsey Global Institute, 2011; McAfee and Brynjolfsson, 2012; Bean and Davenport, 2019) highlighted that we should take full advantage of the possibilities created by the availability of big data and relevant technologies; also, they emphasize the need to develop data-driven approaches, systems and frameworks to better understand and form the insight generation processes (Pick et al., 2017). However, delving deeper into the rest segmentation literature, there are is a lack of data-driven segmentation approaches.

Below, (Table 2-1) we present an overview of both business and research gaps. Also, we shortly present how we address each gap in the current dissertation.

| Concept | Gap | How we address it |
|---|---|---|
| Shopper segmentation | Practitioners suggest that contemporary retail demands a transformation of traditional shopper segmentation systems and approaches. Old-school shopper segmentation is not enough and cannot describe the new, volatile shopper habits and preferences.<br>Current research on customer segmentation utilizes all shopping visits to identify customer groups. These studies examine shoppers' behavior via looking at the entirety of the products a shopper has purchased, regardless of whether this took place in one or more visits and try to segment shoppers based on this behavior.<br>The aforementioned studies overlook the shopping purpose of a single customer visit. However, marketing researchers who talk about different shopping trip types, e.g. fast refilling trip or major monthly trip (Walters and Jamil, 2003; Bell et al., 2011), have stressed the need to understand a single customer visit. | We suggest that we should put the shopper visit on the spot, instead of the shopper behavior that changes over time and that traditional shopper segmentation relies on.<br><br>We coin the term "visit segmentation", to pinpoint this need.<br><br>We use real data from three different case studies and we generate segments of customer visits. Then, we attribute to each segment the shopping intention behind the visits. |
| Market basket analysis (MBA) | Although, market basket analysis practices focus on the associations between the purchased products in basket/visit level. Still, they overlook the shopping purpose of each shopper visit. | The visit segmentation that we propose, focuses on the underlying needs that boosted a customer visit a store e.g. to purchase products for a light meal, or to procure materials to renovate their bathroom etc. These needs and missions can be extracted using various datasets reflecting customers' behavior e.g. product purchases, interactions, preferences etc. |
| Product taxonomy | Lots of segmentation studies are based on the product mixes customers purchase. Researchers performing either market basket analysis (MBA), or similar segmentation approaches have claimed poor results, when there is not a right selection on the granularity level we define the product. | We propose formulating a customized product category level, via balancing a retailer's product taxonomy. In more detailer, we suggest a semi-supervised feature selection method that uses a |

| | | |
|---|---|---|
| | Regarding product taxonomies there are two group of studies that try to tackle this issue: (A) those that formulate categories based solely on product purchases without considering product semantics (e.g. merging socks and skincare, (B) those utilized in the semantic web, which take into consideration solely product semantics.<br>However, in the literature there is a lack of algorithmic approaches that consider both product semantics and product purchases to formulate custom categories that serve the data mining purposes. | product taxonomy as an input and suggests the features/custom categories as an output. This approach is used to balance retailer's product taxonomy tree, and it considers both the frequency of product purchases and the product semantics. |
| Category management (CM) | Business executives recognize the need to incorporate the shopper behavior, needs and missions into CM practices. Similarly, researchers pinpoint that we should manage categories based on shopper needs and behaviors. However, this is still an open issue in business and a research gap in literature. | We propose not only to focus separately on each category e.g. milk, cereals, coffee etc. as traditional CM does, but to move from CM to Shopping Mission Management. This way we will treat categories collaboratively under the shopping mission they participate. |
| Factors affecting segmentation approaches | Marketing literature: studies that discuss several factors that affect big data analytics systems in general→ they do not present evidence of how these factors affected relevant segmentation cases.<br>IS literature: there is a great majority of papers that perform shopper segmentation→ authors describe their own case and it is only implied, and it is not discussed how different factors affected segmentation, system and results/outputs.<br>There is a need to identify factors affecting segmentation approaches and to present evidence on how that happens. | We pinpoint factors, that (A) prospective designers of segmentation systems should consider if they want to produce valid segments, (B) data scientist should consider when manipulating and modeling data and (C) marketeers should consider interpreting the segmentation results. We do so, by presenting three heterogeneous case studies from the retail domain. |
| Business analytics approaches | Lack of data-driven approaches, methods, frameworks in general and thus, lack of data-driven approaches that perform visit segmentation. | We propose, develop and evaluate a business analytics approach that performs visit segmentation. |

**Table 2-3. Overview of business and research gaps**

# 3. RESEARCH METHODOLOGY

The aim of this chapter is to present the research methodology employed to address the research objectives and answer the research questions. Thus, in this chapter firstly we present the research questions that are formulated based on the identified research gaps and business problems. Then, given our research objective and the research questions, we adopt as methodological backbone the design science paradigm (Simon, 1996; Hevner et al., 2004) and **we consider a business analytics approach that performs visit segmentation as outcome of this study**. For collecting data for the various steps of Design Science Research, three different cases studies are selected and presented (multiple case study design). Here we should declare that, **multiple case study design serves Design Science Research** (DSR) approach. Below both Design Science Research approach and Multiple Case Study design are presented. Closing, we describe in detail how we adopt these approaches into the research methodology and design of this dissertation.

## 3.1. Formulating and explaining the research questions

Retailers have begun to realize that the traditional, old-school shopper segmentation is not enough and cannot describe the new, volatile shopper habits and preferences. This happens since the modern shopper has changed and looks constantly for new, improved value-added retail experiences. The shopper flits between shopping channels and performs a complex shopper journey with the purpose to satisfy his/her increasing demands for quality and value (Wood, 2018). Shopper behavior is no longer predictable; it is changing through time and, even, between shopping visits in the same store (Sorensen et al., 2017). Thus, from a methodological perspective it is meaningless to analyze as a bulk, all the visits of a shopper to understand his/her behavior.

To cope with the changing behavior of shoppers, both researchers (Walters and Jamil, 2003; Bell et al., 2011) and practitioners (ECR Europe, 2011) have stressed the need to focus on each single customer visit. Putting the shopper visit on the spot, instead of the shopper total buying behavior that shopper segmentation relies on, has the potential to ensure a more accurate view of the shopper desires that change frequently due to an abundance of new products, shopping channels and services offered every day. Hence, from a methodological perspective, there is a need to analyze each visit a shopper performs, separately.

Based on the above we formulate the first research question as follows:

- How can we derive visit segments from shopper data?

On the other hand, business people translate visit segmentation as "shopping mission". In more detail, practitioners have coined the term "shopping mission" to refer to the intention that initiated a shopper's visit (ECR Europe, 2011). At the same time, marketing researchers talk about different shopping trip types, e.g. fast refilling trip or major monthly trip (Walters and Jamil, 2003; Bell et al., 2011), to refer to shoppers' intentions, missions and deeper motives when visiting a store.

Based on this we enhance the first question as follows:

- Can we extract the different shopping missions of customers from the identified visit segments?

In parallel, delving deeper into the segmentation literature, there are is a lack of business analytics-informed and data-driven approaches to identify the various visit segments and understand shoppers' deeper needs, preferences and missions.

Based on this, the following questions is formed.

- Can we develop a business analytics-informed approach to perform visit segmentation?

Similarly, in existing segmentation literature, we identified that there are different data, retailer and shopper factors and characteristics that affect the input, the analysis and the results of segmentation systems and approaches. Thus, another question we seek to answer is whether these factors affect our proposed visit segmentation approach. Thus, another question is also formed:

- What are the factors that affect the design of visit segmentation systems?

Below we present the aforementioned research questions this dissertation seeks to answer:

- Q1. How can we derive visit segments from shopper data?
    - Can we extract the different shopping missions of customers from the identified visit segments?
    - Can we develop a business analytics-informed approach to perform visit segmentation?
- Q2. What are the factors that affect the design of visit segmentation systems?

As it is obvious, to address these questions, first we should clearly define visit segmentation.

### Defining visit segmentation

We define visit segmentation as: *the process of dividing customers' visits into homogenous groups that unveil customers' deeper needs, preferences and missions as reflected in their behavior during the store visits.*

By referring to behavior, we mean the purchase behavior as reflected in shopper-related data such as:

(A) The contents of a basket, to extract insights e.g. this visit's goal is to purchase products such as rice, salmon, shrimps, soya sauce and seaweed, to prepare sushi.

(B) The basket characteristics, e.g. this visit is a "stock-out visit", including a large volume from a wide variety of products.

(C) The visit characteristics, e.g. this visit was intent to take a quick look at the aisles displaying professional clothes etc.

Here we should admit that, new technologies such as Internet of Things (IoT), boost the (shopper) data that capture customer behavior. Thus, nowadays behavior covers all the interactions during the shopper visits e.g. what a customer purchases in a physical or a web store, puts in a virtual web basket but finally doesn't bought, tries on in a sensor-enabled fitting room, grabs from the smart shelves of an Amazon Go store, puts in the wish-list etc.

## 3.2. Design science research approach

Design Science Research (DSR) is one of the two research paradigms that (Hevner et al., 2004) have recognized. The other research paradigm, called as behavioral-science paradigm, has its roots in natural science research methods and focuses on identifying and explaining the underlying regularities of phenomena or on interpreting human experiences and discourse (Romme, 2003). It seeks to develop and justify theories that explain or predict organizational and human phenomena surrounding the analysis, design, implementation, management, and use of information systems. On the other hand, the design-science paradigm has its roots in engineering and the sciences of the artificial (Simon, 1996) guidelines, design principles and technical capabilities through which the analysis, design, implementation and use of information systems can be effectively and efficiently accomplished (Denning, 1997). Such artifacts are not exempt from natural laws or behavioral theories. On the contrary, their creation relies on existing kernel theories that are applied, tested, modified, and extended through the experience, creativity, intuition, and problem-solving capabilities of the researcher (Markus, Majchrzak and Gasser, 2002). Such artifacts vary from constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems) (Hevner et al., 2004; Hevner and Chatterjee, 2016). Goes (2014) highlights the absence of design science research in Top Journals and subscribes to the notion that the IS field needs more design science research. The design science research paradigm increasingly diffuses into the IS community and has gained increasing recognition over the last years (Baskerville, 2008).

## 3.3. Multiple case studies design

Theory building from multiple case studies gained respect as it is suitable for unexplored research areas where it is critical to bring the researcher in close proximity, both conceptually and physically, to the underlying phenomenon, allowing for deeper engagement with the social settings (Fendt and Sachs, 2008). As Eisenhardt and Graebner (2007) highlight, "a major reason for the popularity and relevance of theory building from case studies is that it is one of the best (if not the best) of the bridges from rich qualitative evidence to mainstream deductive research.". Papers that build theory from cases are often regarded as the "most interesting" research (Bartunek, Rynes and Ireland, 2006).

Selecting cases is an important but difficult aspect of case research. Literature provides some insight into this process (Yin, 1994; Stake, 1995) recommending that the cases should be easy and willing subjects, maximizing what can be learned within limited time. Based on the assertion of Stake n(1995) "a good instrumental case does not have to defend its typicality". A good practice in multiple case study design is the cases to follow replication logic. In this regard, although each individual case study represents a "whole" study, in which information is gathered from various sources and conclusions drawn on those facts, the outcomes from one case are compared with the conclusions from the other cases. This indicates that we talk about literal replication expecting that each case shows the same results. Yin (1994) proposes the usage of around 2-3 cases for literal replication. The first case can be considered as the pilot case that will help us in deciding the final data collection protocols to be used and the design as a whole. Finally, all the cases can be considered as embedded case studies, as they try to draw conclusions by analysing sub-units of the study object and not the phenomenon as a whole.

## 3.4. Research design

In design science, the researcher creates and evaluates IT (Information Technology) artifacts and/or theories intended to solve identified organizational problems. The knowledge base is composed of foundations and methodologies used to develop the artifact. Below we present the basic components of design science research and how are addressed in the current dissertation (Figure 3-1). Afterwards, we shortly explain and translate the basic components of DSR according to our research objectives.

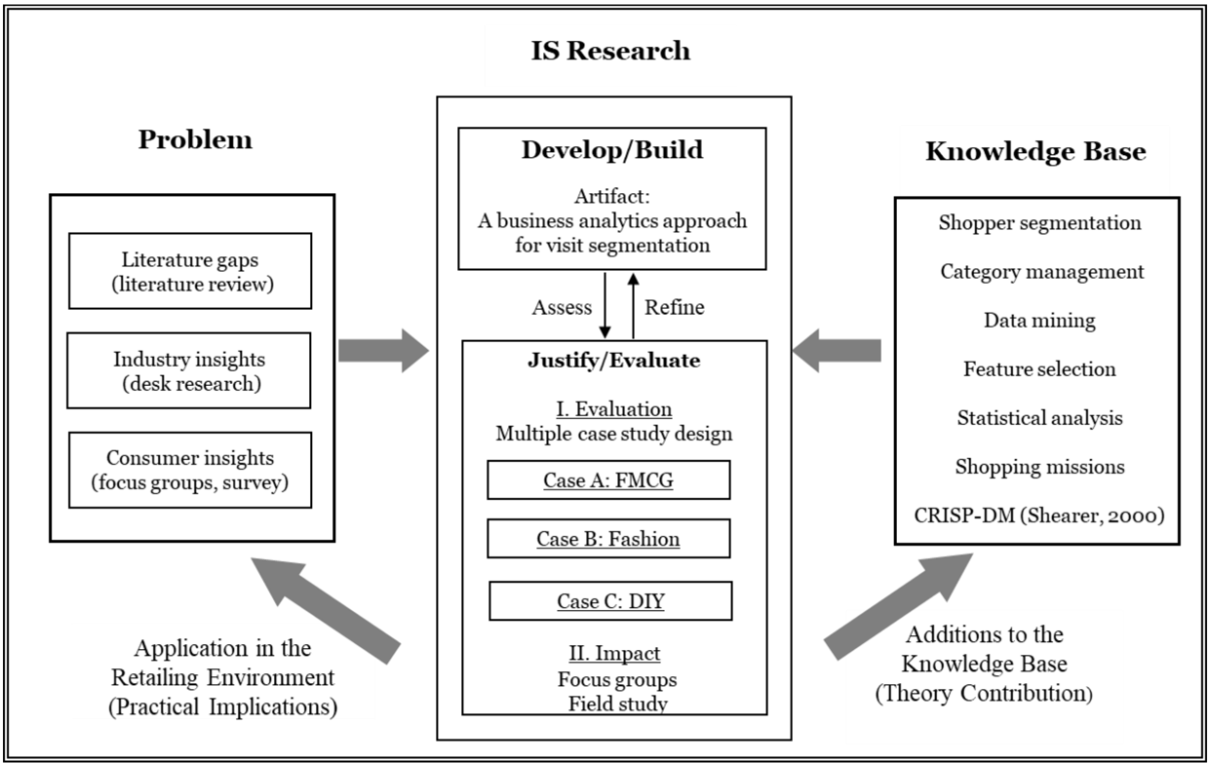



**Figure 3-1. Research approach (adapted from Hevner, 2004)**

### (A) Problem definition

This dissertation aims to solve a business problem/need in the retailing environment, which is to perform visit segmentation and identify the underlaying shopping needs and missions of customers. To better define this problem, it follows the below steps:

- **Literature gaps**: To set the research setting firstly we conducted a review of the pertinent literature. This way we specified the research questions which is related with the visit segmentation concept and we pinpoint the research gaps and the purpose of this research.
- **Industry insights**: Having laid the foundations upon which this doctoral research will be grounded, then we investigated various open issues and business problems industry people face, when try to better understand and satisfy their demanding customers. In more detail, as we identified that

business people translate “visit segmentation” as “shopping mission”. Thus, we use the term “visit segmentation” to more precisely describe the “shopping mission” term which widely utilized in the industry literature.

- **Consumer insights**: Afterwards, to better understand and conceptually define shopping missions in the FMCG domain, we conducted a series of eight semi-structured focus group discussions with 71 shoppers. These discussions confirmed that contemporary shoppers entering the store having in mind a specific shopping mission. In addition, a survey was used to investigate shoppers’ behavior and perception regarding the shopping mission concept.

**(B) <u>Develop/Build</u>**

Then we develop and evaluate a technology-based solution that is relevant to the above research problem (visit segmentation). In this research the **developed artefact is an approach**, providing a certain manner to handle the appropriate data aiming to extract the visit segments.

**(C) <u>Justify/Evaluate</u>**

Then, we put the approach in practice to evaluate it, asses its impact and realize if it can solve the original problem. This phase includes two steps: (i) Approach evaluation and (ii) impact. Regarding the first one owing the lack of prior systematic research on the visit segmentation topic, to address this objective the research is based **on multiple case studies design**. Regarding the second one (impact), we designed a series of focus groups and a field study to assess the value, the impact and the validity of the resulting visit segments. Below we explain both (i) evaluation and (ii) impact in detail.

## I. Evaluation

In more detail, **the proposed approach is been evaluated** by **applying it into** real data derived from **three case studies. The three cases are chosen based on our involvement in industry projects** within the context of analytics. By applying the approach in three different cases we evaluate it and prove its generalizability. Below, we present the characteristics of each case and we discuss the data 4 Vs (variety, volume, veracity, velocity) for each study (Table 3-1).

**Case A**: The first case concerns sales data from different channels and stores of two major Greek fast-moving consumer goods (FMCG) retailers. Regarding data volume and variety, the first FMCG retailer has provided one-year point-of-sale (POS) data from two representative physical mini-hyper markets, two supermarkets, two convenience stores and the web store. Similarly, the second retailer provided one-year POS data from a supermarket. Apart from the POS data, we also received data regarding the product taxonomy and loyalty cards data. In more detail, we received loyalty cards data and customer demographics solely from the web store of the first retailer, and we had information regarding the declared card holder's age, gender, and household size. Loyalty cards usage in the web store is increased, as the retailer has set a beneficial points system. Likewise, the second retailer provided us with loyalty and cardholders data. We performed ad-hoc analysis based on historical data, thus data velocity didn't affect us. Regarding data veracity, although in the retail context, different data issues may arise; we didn't identify significant imprecisions to our datasets. We faced imprecisions in the demographics data, as a few shoppers declared wrong information, for instance, in attributes such as age, households etc.

**Case B**: Respectively, in the second case, we produced the visit segments for all the physical stores of a Fortune 500 specialty retailer of home improvement and construction products – also known as do-it-yourself (DIY) retailer. In more detail, regarding dataset volume and variety, we received two-year POS data, of various stores of the retailer. Each visit was associated with a cardholder; hence, we could identify all the baskets a shopper had purchased through his/her history. Apart from the POS and loyalty data, we also received the product taxonomy retailer use to categorize the available products. Data veracity in this case was low, as the POS data was already cleaned and curated by retailer's team. Also, here we didn't receive demographics data.

**Case C**: The third case concerns data from a physical and the web store of a major German fashion retailer. Regarding dataset volume we received one-year POS data from one physical store of a European fashion retailer and the transactions of the web store. Concerning data variety, we received: POS data, data for the product taxonomy tree, cardholders' demographics e.g. gender and age, data regarding promotions e.g. we could track whether each transaction was promo-driven, garments' data e.g. color, size. Data veracity in this case was medium as we faced imprecisions in the demographics data, as a few shoppers declared wrong information.

| | **Case A** | **Case B** | **Case C** |
|---|---|---|---|
| Domain | FMCG | Home Improvement - DIY | Fashion |
| Data Variety | -POS<br>-Loyalty<br>-Product taxonomy<br>-Cardholders' demographics | -POS<br>-Loyalty<br>-Product taxonomy | -POS<br>-Loyalty<br>-Product taxonomy<br>-Cardholders' demographics<br>-Product characteristics (e.g. color, size)<br>-Promotions |

| | | | |
|---|---|---|---|
| Data Volume | 5.5 GB | 0.5GB | 0.8GB |
| Data Veracity | Medium due to cardholders' demographics | Low | Medium due to cardholders' demographics |
| Data Velocity | Ad-hoc analysis based on historical data | Ad-hoc analysis based on historical data | Ad-hoc analysis based on historical data |
| Stores | 3 different store types, 6 physical stores, web store of a Major Greek retailer, 1 supermarket of another Major Greek retailer | All physical stores of an American retailer | 1 physical store, web store of a German retailer |

**Table 3-1. Case studies overview**

**II. Impact**

To **examine the impact of our approach** we designed a series of **focus groups** and a **field study** for one supermarket store of an FMCG retailer. This process included three phases (Figure 5-3):


**Figure 3-2. Shopping mission evaluation phases in the FMCG environment**

i. Firstly, we analyzed one-year transactional/POS data from one grocery store to identify the shopping missions that shoppers perform during visiting each store.

ii. Then we conducted **semi-structured focus groups** to discuss with the actual store shoppers and ask for their view on the resulting shopping missions.

iii. Afterwards, we designed a field study in the store to evaluate the resulting data-driven shopping missions and asses their validity. To achieve this, we exploited two different means i.e. a **mobile app** and a **survey** using hardcopy questionnaires. While users shopped and navigated in this store, they used a custom mobile application which distributed various coupons. Then, at the store exit they filled a short questionnaire. Via this case study we proved that shoppers confirm the identified data-driven shopping missions. Also, to enhance shopping mission's validity we demonstrate that the shopping mission-related disseminated coupons achieve higher redemption rate and are claimed by a shopper into less time than the non-related coupons.

### (D) Knowledge base

To build the proposed approach, we used both theoretical foundation and methodologies. Theory regarding shopper segmentation and category management and CRISP-DM (Chapman et al., 2000) which is a Cross-Industry Standard Process for Data Mining were used as the basic knowledge inputs in the developed approach. In more detail in our approach we follow and alter the CRISP-DM steps.

In addition, data mining techniques such as clustering and classification, data mining algorithms such as k-means and feature selection methods were used to develop the approach. Statistical analysis and measures such as ANOVA, Pearson correlation, Jaccard similarity etc. were used to evaluate the field study results and to analyze the shopper questionnaires. Likewise, qualitative analysis was used to analyze the focus group transcripts during the various research faces.

Closing, the theory contribution and the practical implications are detailed. Regarding theory contribution this dissertation, develops a business analytics approach that performs visit segmentation. To the best of our knowledge this is the first data-driven attempt to identify visit segments and explore the underlying customers' shopping missions. Also, this dissertation opens a new chapter in category management. In a nutshell, the practical value of this work is stressed when considering the consumer-oriented business decisions it can support e.g. shopping-mission based store layout, or product catalogues, or promotions etc.

# 4. A BUSINESS ANALYTICS APPROACH FOR VISIT SEGMENTATION

Our goal is to explore visit segmentation concept in a way to comprehend customers' shopping behavior and intentions and identify their deeper shopping needs that motivated the shopper's visits. To do so, we propose a data-driven approach to identify the different visit segments. In brief, we generate clusters of visits based on the content of a basket and the basket characteristics (see 2.1.2 for more details). The resulting mix of purchased product categories guides us to identify the original shopping purpose of each shopper visit and, thus, characterize each cluster of visits based on the shopping intentions and missions that motivated the visits e.g. if we identify that in a visit segments shoppers purchase rice, salmon, shrimps, soya sauce and noodles, then we assume that this visit segment entered the store to purchase products for an ethnic meal.

To develop and design the visit segmentation approach, we have adjusted CRISP-DM (Shearer, 2000), a cross industry standard process for data mining. Our approach includes the following phases/layers: (a) data understanding and preparation (where the data are pre-processed, cleaned and prepared for the data analysis purposes), (b) data modeling and model evaluation (where the data mining model is created and the results are evaluated in both business and technical terms), (c) results interpretation (where the visit segments are extracted, interpreted and translated into shopping missions).

The major input of our approach is data recording to shopping behavior (e.g. content of a basket and basket characteristics). Also, other data sources e.g. demographics, loyalty cards data, product information etc. are used to enrich the analysis. Different factors and data that are either related to the retailer (i.e. 4Ps), or to the shopper (i.e. shopper 4Vs), or to both (i.e. loyalty programs) could affect the input of the visit segmentation approach. In more detail these factors are twofold, as on the one hand they shape the initial data set, and on the other hand they have a mediating role in explaining the results.

The output is the final interpreted and translated visit segments into shopping missions that can be used by marketeers and decision makers to support decision making. The originality of our approach is embodied to the last phase/layer where we interpret the visits' clustering results to communicate them to the experts, and on the modeling phase which is affected by specific factors e.g. variety, product. We highlight that the effectiveness of our approach and the generation of meaningful visit segments, depends on the afore-described factors of the data (e.g. basket volume) and the retail case itself (e.g. shopping channel and product). The values that these factors take in each retail case should guide both the execution of the data analysis, as well as the translation of visit segments to shopping mission per visit.

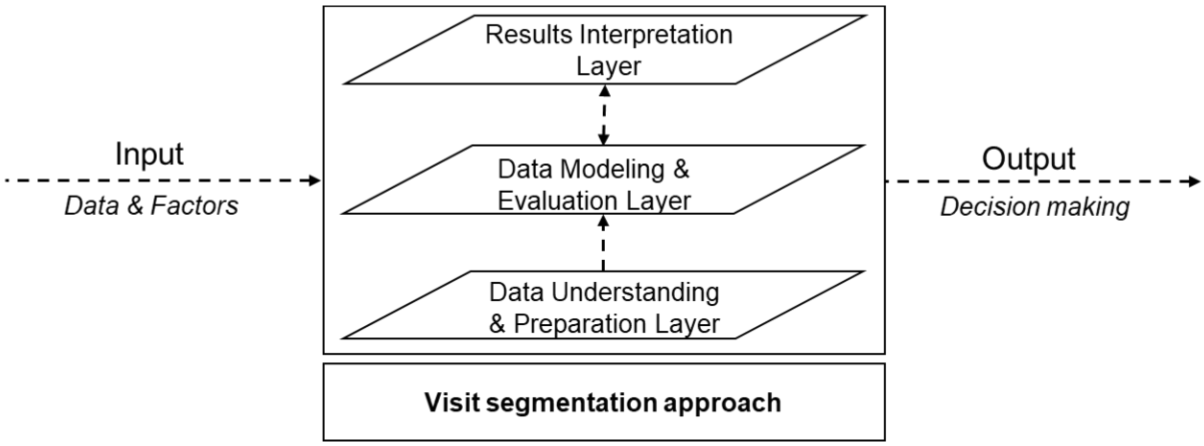



**Figure 4-1. Visit segmentation system**

Next, we analyse more thoroughly each layer of the visit segmentation approach (Figure 4-2). As mentioned above, the originality of our approach is embodied to the last phase/layer where we interpret the clustering/segmentation and in the "Modeling" phase (marked with red in Figure 4-2) where we employ clustering for the customer visits segmentation. This phase includes: (a) product taxonomy adjustment, (b) cluster sampling and (c) adjustment of the input data to produce valid customer segments. Next, we summarize the steps of our approach.

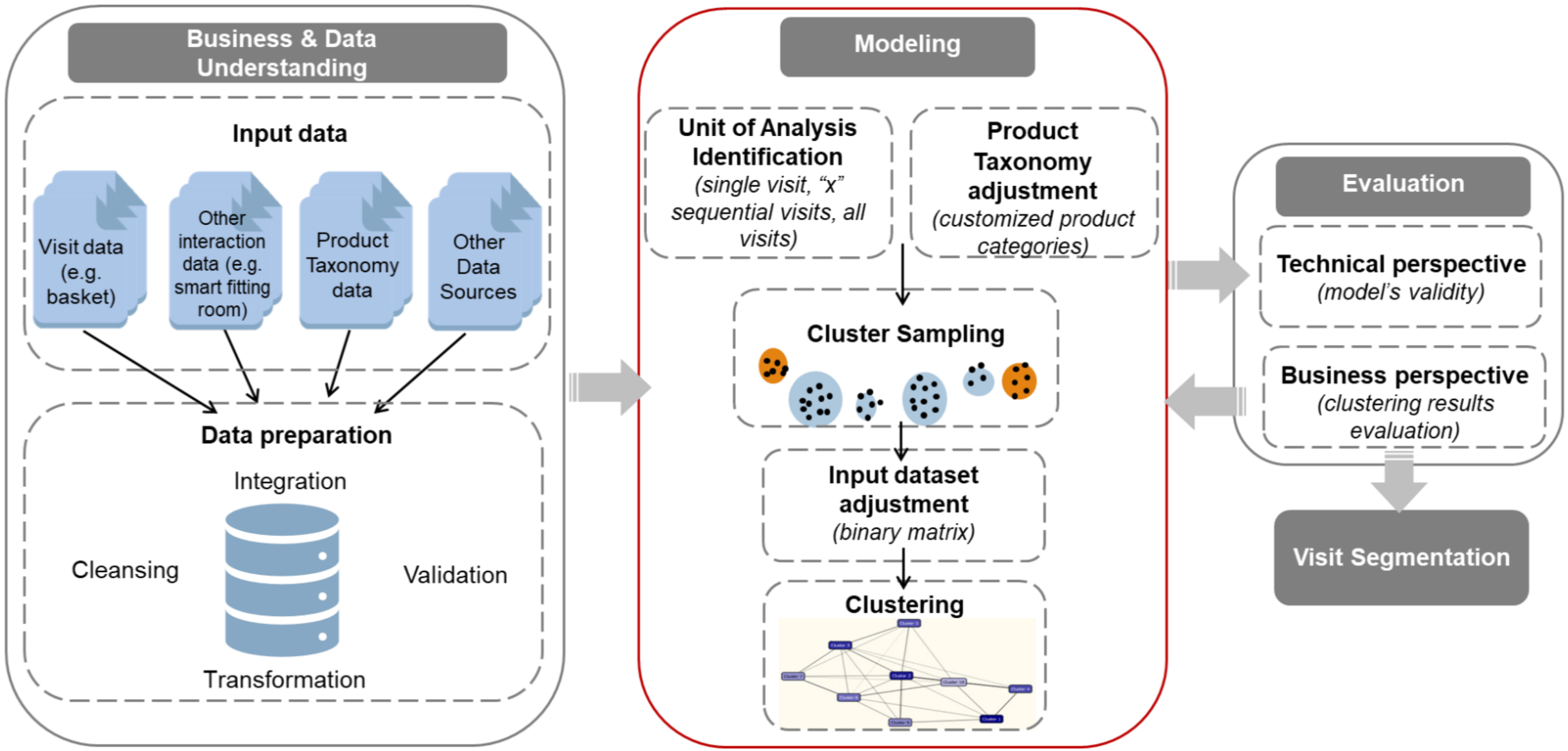



**Figure 4-2. Business analytics approach for visit segmentation**

## 4.1. Business, data understanding and preparation layer

The business goal is to identify the different segments of customer shopping visits which reflect the specific shopping missions, i.e. needs and preferences of the corresponding customers that motivated these shopping trips/visits. Our goal is to offer them the appropriate service mix. We perform the segmentation by examining the product categories the customers purchase during their visits in physical or web retail stores.

**Input dataset:** Apart from data referring to the product purchases per visit (i.e. basket data), the relative input dataset includes the product category tree and the product barcodes of the retailers' product assortment. More input data might be other interactions between shoppers and products during store visits. For instance, products that customers put in their physical or electronic basket, store aisles they pass by, products putting in their wish list, or garments that they try on in fitting rooms in fashion retail stores, but they do not purchase them. Such data may be captured by the standard point-of-sales devices or by RFID sensors, Bluetooth Low Energy (BLE) tracking devices (e.g. beacons), or by navigation data (e.g. google analytics) in the web environment. Extra data sources, e.g. customer demographics data, could enrich and enlighten the resulting visit segmentations.

**Data preparation:** Given the heterogeneity and the noisy nature of the data, it is not enough to just collect them and throw them into a data repository (Jagadish et al., 2014). Synchronizing and integrating the datasets derived from various sources for establishing data consistency is a major challenge. Thus, data preparation is required to support the comprehension of the data sources and the business context they originate from. In other words, firstly wet perform data integration, which involves combining data residing in different sources (Lenzerini, 2002). Then, we apply data cleansing for detecting and correcting or removing errors and inconsistencies of the data to improve data quality (Rahm, 2000); and we transform the data in a way to be ready for the modeling. Finally, we end up with data validation after each of the above steps to consolidate the data integrity of the available datasets based in ad hoc criteria selected by the researchers.

## 4.2. Data modeling and evaluation layer

Essentially, this phase performs the following three prerequisite tasks: (a) unit of analysis identification, (b) product taxonomy adjustment, (c) cluster sampling and (d) input data adjustment and clustering, which ensure that the clustering analysis will produce meaningful results.

### 4.2.1. Modeling

#### 4.2.1.1. Unit of analysis identification

It is critical to identify the unit of analysis we will use to identify customer's shopping intention. Extracting shoppers' mission might not require zooming into a single store visit. In more detail, there could be retail cases where we can perform visit segmentation and identify shoppers' deeper intentions in visit level e.g. this visit is performed to purchase breakfast. There could also exist cases where we need to examine "x" sequential visits to identify the shopping mission. For instance, a shopper usually visits a retail store that sells products for home improvement many times and purchases few materials each time (Wolf and McQuitty, 2011). Hence, to understand his/her shopping mission we need to examine his/her continuous in time visits. Lastly, there could also be cases where we should examine all shoppers' store visits to identify the shopping intention e.g. in a fashion store. This way we move from visit segmentation to traditional segmentation approaches (Figure 4-3). The factors and the particularity of each retail domain will help us on this decision.

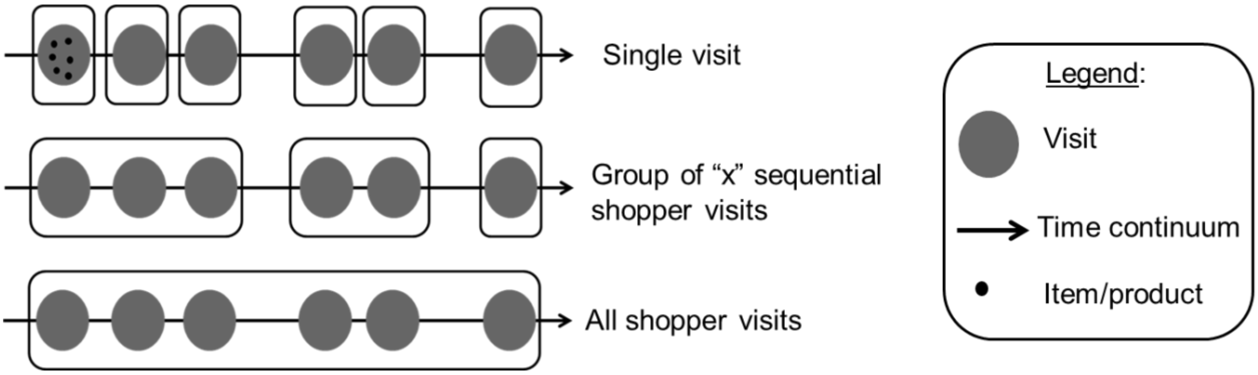



**Figure 4-3. Possible units of analysis**

#### 4.2.1.2. Product taxonomy adjustment

Each retail chain has designed and maintains a product hierarchy (often referred to as product-categories tree) that is necessary to conduct various business processes (e.g. store replenishment, shelf space allocation, product assortment selection). This tree corresponds to the product variety and market specialization to facilitate the operational activities in the best way possible. However, we suggest that it is not suitable "as-is" for data analytics purposes because it is often unbalanced and has characteristics hindering the performance of data mining algorithms. These characteristics, which we also came across in our study, are: (a) the height of the sub-trees is significantly different, indicating high product specialization in some product categories (sub-trees), (b) the product tree might be a forest from a data structure perspective meaning that the product categories are expanding separately and managed independently, and (c) the node's degree is varying significantly especially at the SKU/item level. Hence, we suggest that the underlying characteristics of the dataset might affect data mining activities due to the utilization of highly skewed data sets.

To see into the characteristics of the dataset and discover any signs of skewness, we examined the relationship between two variables, namely the number of SKUs/items classified at every branch of the Product Taxonomy (product variety) and the participation percentage of a branch in the baskets (basket frequency). In the next scatter plot (Figure 4-4), we depict that the x-axis depicts the former variable and y-axis the latter for a product taxonomy tree with height=3. Every point of the plot represents a single product's taxonomy branch and different colors are used to discriminate paths belonging to different product taxonomies (forest). The plot suggests significant positive skewness in both variables; therefore, we had to manage the dispersion either by merging nodes relying at the bottom right area or by splitting nodes found at the top left corner and produce an efficient balanced product taxonomy. According to Aggarwal (2016), our problem domain requires extreme-value analysis as it suffers from outliers and we adopted Aggrawal's suggestion that "the choice of the model depends highly on the analyst's understanding of the natural data patterns in that particular domain". In this spirit, we initially utilized relevant techniques (e.g. Box-Cox transformation) to manage outliers, but we finally came up with a semi-supervised feature (product category) selection approach that gets the product taxonomy as input and suggests the features as output.

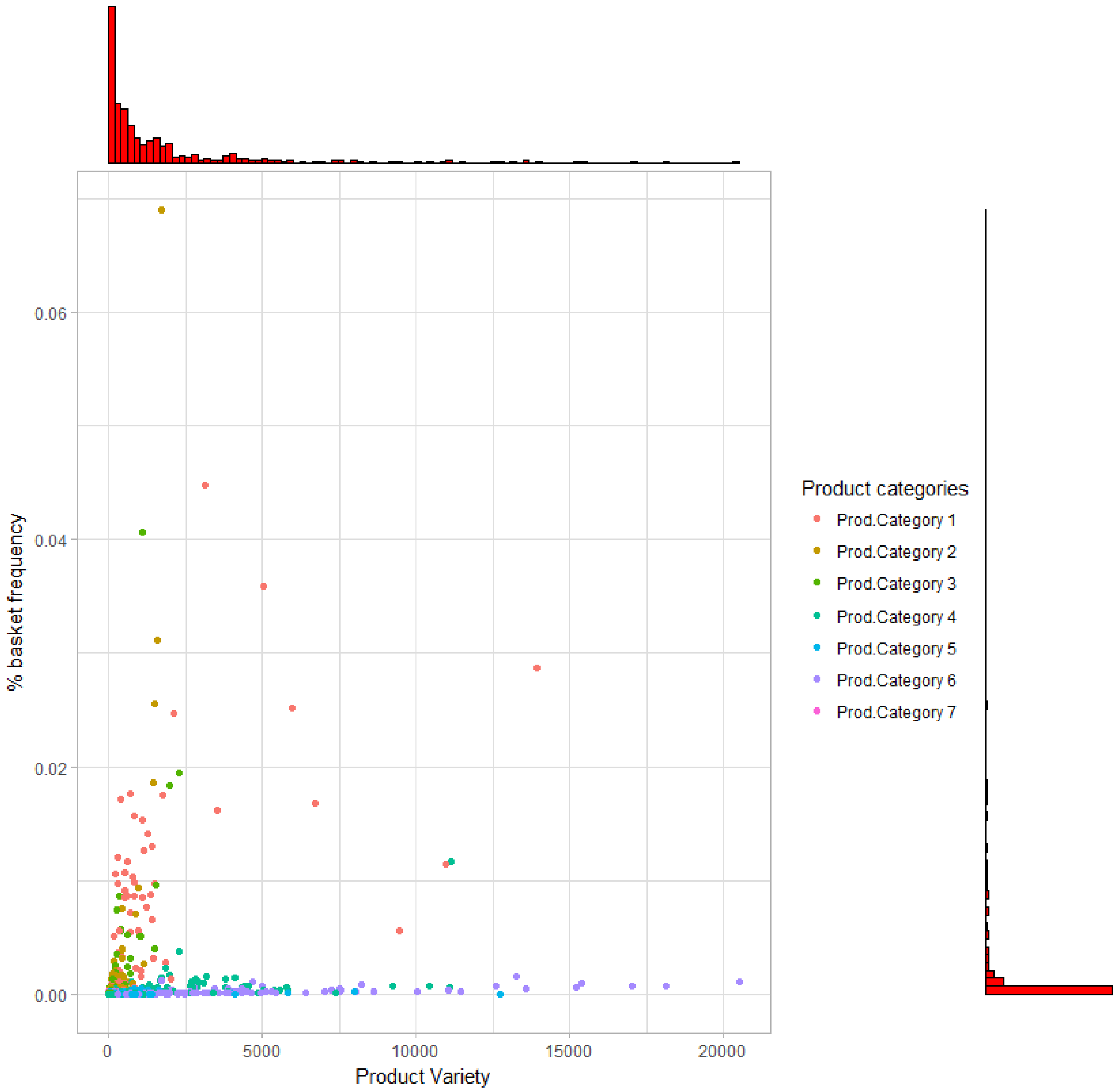



**Figure 4-4. The relationship between product variety and basket frequency**

More specifically, we propose an approach relying on the variety of the product categories in a shopping basket or visit (product variety) that adjusts the retailer's original product taxonomy and produces a customized product-categories tree, which can adequately support the clustering analysis and the identification of the customer visit segments. The logic behind the balancing of the product-categories tree is mainly quantitative. The steps we follow to balance the product taxonomy tree are:

- **a.** First, we identify the main product categories (e.g. initial level n in *Figure* 4-5). Then, e utilize the other researchers' proposition (e.g. Bi, Faloutsos, & Korn, 2001) that retail sales data could be represented as a Discrete Gaussian Exponential (DGX) Distribution; thus, a relative small percentage of product categories contributes in most of sales (or Basket Frequency in our case). The role of DGX is to isolate product categories into two disjoint sets: (i) the *green* set includes product categories with high Basket Frequency and (ii) the *red* set assembles the remainder product categories. The proportion between green and red product categories empirically was found around 1:10.
- **b.** Secondly, we adopt a bottom-up iterative approach and focus on the *red* set of product categories and merge nodes sharing the same parent. In other words, we shrink a sub-tree of red nodes and replace it with the parent node with respect to manage the long tails negative effects and the skewness of the data.
- **c.** Regarding the height of the sub-trees, we revised the new product taxonomy and if a tree branch is shallow, e.g. see 'level n-1" of *Figure* 4-5, the last available nodes will also become *green*.
- **d.** Finally, we reconsider the merged product categories in a qualitative manner taking into account the business context, the analysts' acquired knowledge of the context and experts' opinions (e.g. suppliers, retail managers etc.). Ultimately, we determine if we keep these *red* categories that had been merged as the algorithm indicated, or we split them, or we devise completely new categories that serve the data mining purposes by merging selected nodes.

We emphasize that we merge or split categories that belong to the same parent node, considering the experts' opinion and the product semantics in a way that we avoid merging unrelated categories e.g. skincare and socks.

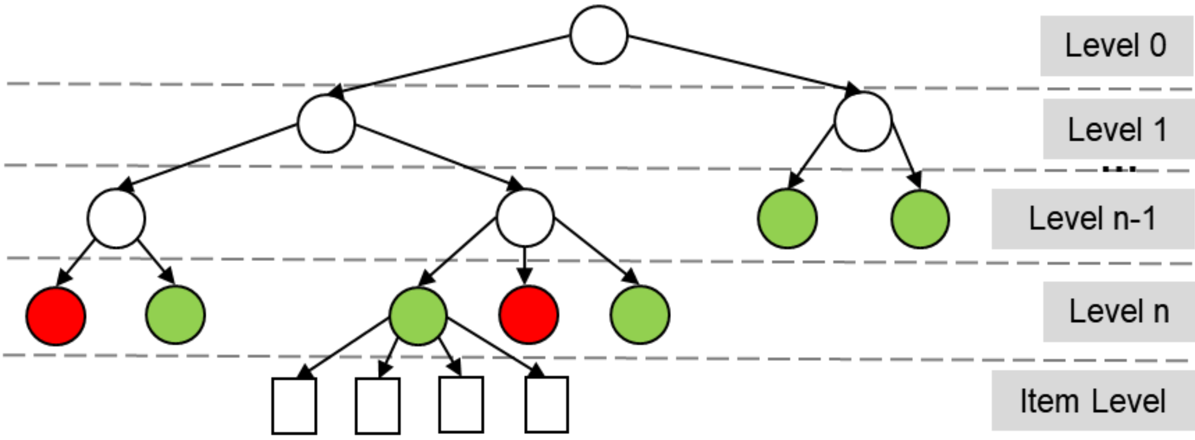



**Figure 4-5. Balancing product categories tree to formulate new product categories**

### 4.2.1.3. Cluster sampling

We use cluster sampling, with equal sampling weights, to eliminate the outliers. we consider the visits during which a customer has purchased very few or too many products in terms of variety to be outliers. These visits correspond to too concrete or too abstract shopping trips (Bell et al., 2011). The concrete shopping trips are too targeted to extract any product affinities, whereas the abstract ones contain such a wide variety of purchased products, e.g. a monthly stock-out visit in a supermarket, which cannot highlight a specific shopping mission.

The cluster sampling technique groups a finite population into subpopulations-groups called clusters; then, a subset of these clusters is selected (Särndal, Swensson and Wretman, 2003). We select the final meaningful clusters considering a basic criterion i.e. the percentage of baskets belonging to each cluster, as well as other relevant descriptive statistics, e.g. revenues per cluster. For instance, in the case of a specific grocery retail store, we can eliminate the baskets that contain only one product (8% of total baskets) and those baskets with more than 80 products (2%). We will utilize the rest of the baskets which reflect 95% of the total revenues.

### 4.2.1.4. Input dataset adjustment and clustering

To perform clustering, we need to adjust the dataset in order to form the fact table (Table 4-1) (Shearer, 2000), which represents the learning dataset of the clustering model and includes all the information about a customer shopping visit. Each row of our data table represents a visit (or basket) and the columns correspond to our customized product categories, as well as the visit attributes. The product categories' columns are filled with a binary flag, one (1) or zero (0), indicating that the respective basket contains products of this product category or not, respectively. These categories/columns are the input to the data-mining model. In our case, we have selected clustering as the data mining technique to segment visits. More specifically, we have selected k-means. The basic idea of k-means is to discover k clusters, such that the objects within each cluster are similar to each other and dissimilar from the objects in other clusters. K-means is an iterative algorithm; thus, an initial set of clusters is defined, and the clusters are repeatedly updated until no further improvement is possible (You et al., 2015; Huerta-Muñoz, Ríos-Mercado and Ruiz, 2017). The accuracy of this algorithm and the quality of the results depend also on the initial number of clusters (Mesforoush and Tarokh, 2013). Thus, it is critical to define a mechanism to determine the optimal number of clusters. Well-known methods to determine the optimal number of k are elbow, silhouette and gap statistic (Milligan and Cooper, 1985; Tibshirani, Walther and Hastie, 2001). Here, we should mention that the proposed approach seems to be independent and free from any clustering method, as other clustering algorithms e.g. EM (Expectation Maximization).

| Visit/Basket UID | Custom Category 1 | Custom Category 2 | … | Custom Category n | Other relative data e.g. demographics and Meta- data e.g. basket size, revenue etc. |
|---|---|---|---|---|---|
| 1 | 1 | 1 | | 0 | |
| 2 | 0 | 0 | | 1 | |
| … | | | | | |

**Table 4-1. Fact table structure**

### 4.3.2. Evaluation

Here, we suggest that the resulted visit segments should be assessed in both business and technical terms. On the one hand, a group of industry experts should assess the validity of the results based on their accumulated experience. If they defy them, we should re-execute the analysis after changing the input dataset. For communicating the results of our approach to the business experts, we translate each found segment of visits to a shopping intention that motivated the segment's visits. More specifically, we characterize each group of shopping visits/ trips by examining the prevailing product categories the customers purchased during the shopping visits of each segment. For example, if a cluster includes baskets that mainly include categories such as milk, cereals, coffee, sugar etc., we call this segment of visits as "breakfast", declaring that customers have visited the store to buy goods for their breakfast.

If we need to make changes to the original input data based on the experts' comments, we usually delete, merge, or split some of the customized product categories. Empirically we identified that merging or splitting contiguous product categories is a practice for increasing the internal consistency of clusters. Thus, after a first trial, it is more effective to reconsider custom categories level. For example, in some cases, merging two or more product categories has resulted in a generic category. On the contrary, disjoining results is the split of a custom category in its children categories/ nodes. From a data mining perspective, this decision decreases a sample's variability and consequently yields better performance results. Thus, the business evaluation constitutes a dialectic process between the experts and the data mining techniques. The researcher should calibrate the cluster model, to satisfy important data mining metrics and, at the same time, deliver a readable abstraction of the cluster to the experts.

On the other hand, in terms of technical evaluation, we need to test the model's validity. Since clustering is based on the similarity of the contained objects, metrics such as a cluster's compactness (e.g. how closely related the objects within the same cluster are) and separation (e.g. how separated the clusters are) could be calculated for the internal validation of the clusters (Liu et al., 2013).

## 4.3. Results interpretation layer – Visit segments and shopping mission identification

Then, for communicating the results to the business people, we interpret and translate each found segment of visits to a shopping mission and intention that motivated these visits. This interpretation is are a result of the product taxonomy adjustment. As the shopping mission naming is based on the parent nodes/categories that participate in each cluster. To extract wisdom from the data, we need experts' opinion. Experts will not only examine the tangible/quantitative features e.g. basket volume, value, but also intangible elements such as their domain knowledge and accumulate experience. This step is critical as business people should understand the results to support decision making. Here, the clustering results are extracted, and the final visit segments are shaped with the objective to give retailers new knowledge for decision support purposes. We suggest calculating some extra descriptive statistics/Key Performance Indicators (KPIs) per cluster, proposed by the experts, e.g. the basket variety and volume (i.e. the number of product categories and the average number of items it contains, respectively) and the revenues per cluster etc. (Figure 4-6). Such measures can support the characterization of the final visit segments.

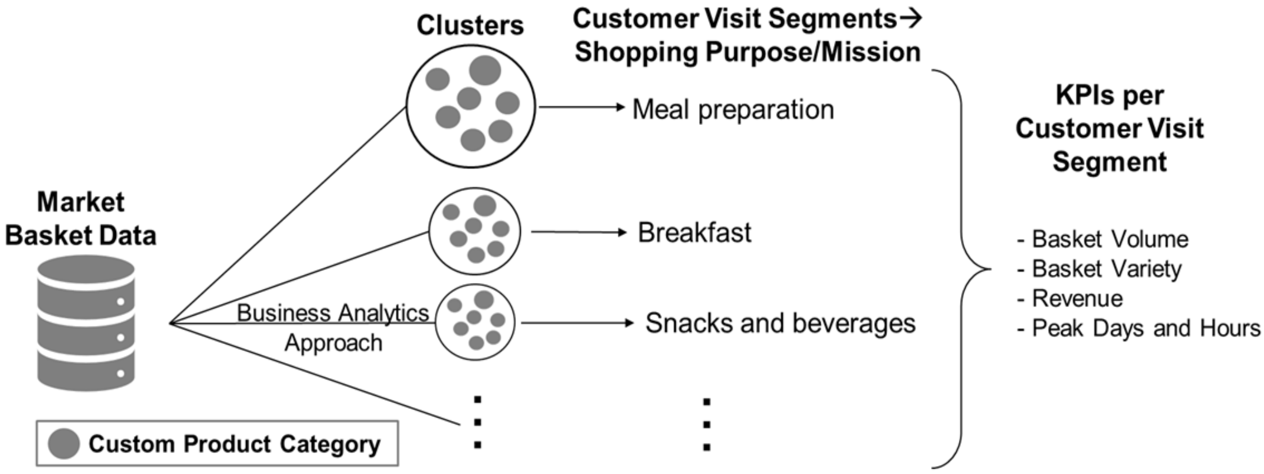



**Figure 4-6. Conceptualization of visit segmentation and characterization**

A drill-down analysis can further be applied to clusters that contain more abstract visits, namely to perform clustering within a single cluster. Then, an abstract cluster may contain more than one sub-cluster. For example, if we apply drill-down to a cluster with many products and product categories, such as wine, beverages, beers, chips, nuts, chocolates, ice, biscuits, orange juice etc., then the original cluster may split into two sub-clusters. Figure 4-7 shows two new clusters, one with “beverages” and one with “snack” products. In other words, drilling-down can highlight hidden shopping purposes. However, the resulting sub-clusters are often the same with the original ones. An alternative option is executing a similarity function (e.g. Jaccard similarity) between the well-defined and formed clusters with those that are more abstract. In the same spirit, apart from drilling-down, it is also possible to roll-up and merge some of the resulting clusters. The involvement of the decision maker and the opinion of the experts will help researchers choose between roll-up and drill-down operations.

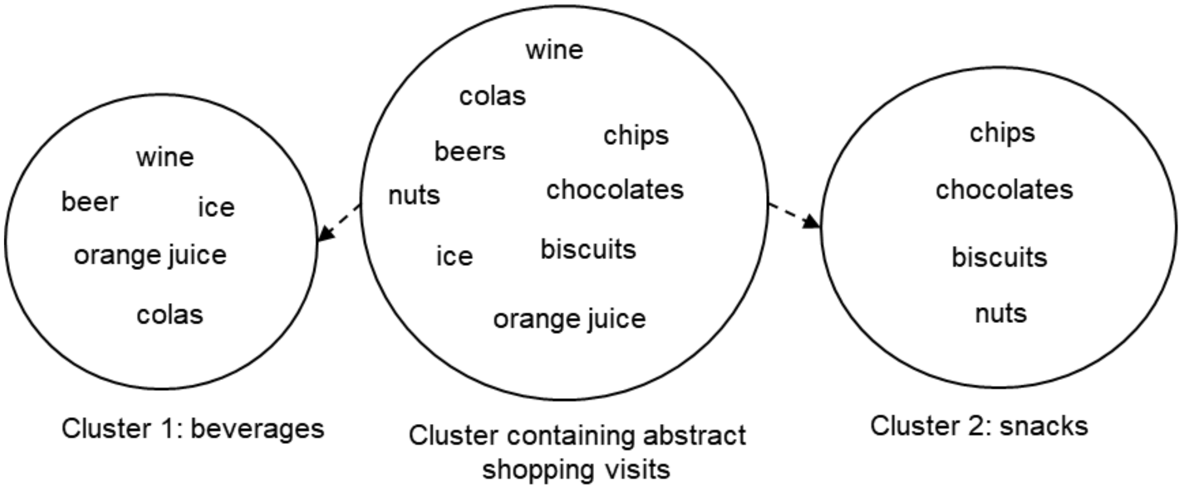



**Figure 4-7. Drill-Down analysis in a cluster containing abstract shopping visit**

# 5. APPLICATION OF THE PROPOSED APPROACH IN THREE HETEROGENOUS RETAIL CASES

Here, we put our proposed business analytics approach in practice demonstrating how it achieves the original goal, i.e. to segment the customers' visits. We applied and validated our approach through three heterogeneous retail cases in terms of both factors and retail context to demonstrate its generalizability. The first case concerns sales data from different channels and stores of a major European fast-moving consumer goods (FMCG) retailer. Respectively, in the second case, we produced the visit segments for the physical stores of a Fortune 500 specialty retailer of home improvement and construction products – also known as do-it-yourself (DIY) retailer. The third case concerns data from a physical and the web store of a major European fashion retailer.

## 5.1. Case A: Application of visit segmentation in FMCG retailing

Here, we present our proposed business analytics approach in practice demonstrating how it achieves the original goal, i.e. to segment the customers' visits. We utilized original sales data from one European FMCG retailer with more than 300 stores, one of the major retailers in the national market.

### 5.1.1. Business and data understanding

In this case we have high visit variety. Yearly number of visits seems to be indifferent to the results, as here we examine separately each single store visit. Also, this case concerns different channels and store types. In more detail, regarding data volume and variety, the FMCG retailer has provided one-year point-of-sale (POS) data (see

Appendix A: Indicative structure of the analyzed retail datasets and for an indicative grocery payment receipt) from two representative physical mini-hyper markets, two supermarkets, two convenience stores and the web store. We analyzed 25.887.925 records that correspond to 1.835.174 baskets or store visits. Apart from the POS data, we also received data regarding the product taxonomy and loyalty cards data. In more detail, we received loyalty cards data only from the web store, and we had information regarding the declared card holder's age, gender, and household size. Loyalty cards usage in the web store is increased, as the retailer has set a beneficial points system. Table 5-1 shows more details about the given dataset. After integrating and cleansing the dataset, we kept 97% of the initial data, as the given data was already cleaned by retailer's team. We only had to eliminate product returns, seasonal items and services provided by the retailer, e.g. product transfers to a shopper's home.

| | Mini-hyper | Supermarket | Convenience | Web store |
|---|---|---|---|---|
| Unique SKUs-Barcodes | 180.620 | 126.402 | 15.917 | 21.240 |
| Dataset Volume (in records) | 11.645.232 | 7.075.445 | 4.678.820 | 2.488.428 |
| No of Baskets/Visits | 862.241 | 476.729 | 339.832 | 156.372 |
| Basket Volume | 15,3 | 13,3 | 9,6 | 42 |
| Basket Variety (in SKU level) | 9,4 | 7 | 6 | 15 |
| Average Basket Value (in €) | 38,4 | 32,6 | 23 | 72 |
| Average no of visits (per cardholder) | N/A | N/A | N/A | 5,3 |
| No of cardholders | N/A | N/A | N/A | 28.943 |

| | | | | |
|---|---|---|---|---|
| Loyalty card usage | N/A | N/A | N/A | 96,5% |

**Table 5-1. FMCG dataset identity**

### 5.1.2. Modeling

We studied each store separately, as practitioners and business experts suggested so. We utilized cluster sampling with equal sampling weights to eliminate outliers, namely very concrete (low limit) or very abstract (upper limit) shopping visits. Regarding the web store, the retailer has already set the lower limit, as the value of each order should be more than 50€. After cluster sampling (using k-means), we calculated the actual number of baskets and the corresponding revenues per cluster to help us with the outliers' extraction Table 5-2 summarizes the final dataset analyzed, according to cluster sampling results. The second column represents the range (from, to) of the basket size per store type and the last two columns include the percentage of baskets and their corresponding revenues that we have finally utilized to mine the customer segments (see Appendix C: Cluster sampling results – FMCG case for more details).

Additionally, we performed product taxonomy adjustment beginning with rough balancing of the product tree on quantitative criteria. We balanced the product category tree by examining the participation of each tree node in the total purchases. Thus, we generated a first set of 110 customized product categories (see Appendix B: Product taxonomies structure per case study for more details). Then, we consulted experts of the domain for the final fine-tuning of the product categories. Ultimately, in order to create a more balance product taxonomy, we created 90 new-customized categories by merging some product categories-node. For example, the quantitative criteria highlighted that "lager" beers should be examined separately due to their high participation to the total beer purchases in all stores. Hence, we concluded that the other types of beer (such as stout, bock, ale etc.) should be grouped in one product category named "other beers". However, the experts prescribed to us that we should handle both "lager beers" and "other beers" as one product category named "beers", because the results should also correspond to how the retailers and suppliers handle and understand such products in reality.

| Store Type | Basket Size Range Sample | Percentage of the total baskets used | Percentage of revenue used |
|---|---|---|---|
| Convenience | 2-24 | 78,6% | 86,0% |
| Supermarket | 3-40 | 75,4% | 83,4% |
| Mini-hyper | 3-51 | 79,7% | 94,5% |
| Web store | 4-35 | 83,2% | 85,7% |

**Table 5-2. Summarized results of cluster sampling**

We executed the k-means clustering method resulting in seven models. We analyzed each store separately because identical visit segments will not necessary result from the same store type. For this reason, we developed Java code to create seven fact tables, one per store.

### 5.1.3. Evaluation

One common method of choosing the appropriate cluster solution is to compare the sum of squared error (SSE) for various numbers of clusters (i.e. different numbers of K). SSE is defined as the sum of the squared distance between each object of a cluster and its cluster centroid. Hence, SSE is a global measure of error. It is common that the more the clusters are the smaller the SSE is. Thus, a plot of the SSE against several values of *k* can provide a useful graphical way to choose an appropriate number of clusters. A suitable "K" value could be defined as the one at which the reduction in SSE slows dramatically. This produces an "elbow" in the SSE plot against cluster solutions

(Ketchen & Shook, 1996; Likas, Vlassis, & Verbeek, 2003). Figure 5-1 depicts the Elbow Method for the fact table (i.e. the table will be used at the input in our data mining model) of a supermarket of our dataset. In our case, this plot doesn't show a very strong elbow. We do not have a substantial impact on the total SSE for "K" values between 6 and 10. Thus, we performed clustering several times experimenting with different "K" values ranging from 6 to 10. Again, we consulted domain experts to evaluate the results and depict the optimal number of clusters "K" from a business perspective.

Assessing the first clustering results, the industry people noticed an important product category absence. None of the clusters, in all the different trials, included product categories related to "meat". For that reason, we stepped back at the product taxonomy adjustment phase and we modified the feature space via merging meat-related categories, such as pork, beef, lamb etc., into one, to deliver a readable abstraction of the cluster to the experts. Finally, we ended up with 90 out of the initial 110 product categories. After re-executing clustering for different "K" values, the experts indicated to produce 10 clusters (K=10) for the supermarket of Figure 5-1. Alike, we found suitable K values for the rest of the stores.

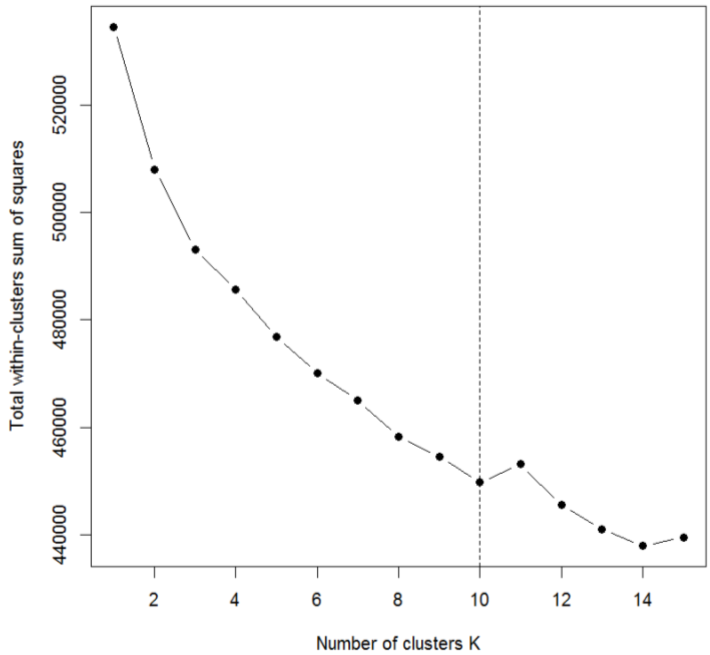



**Figure 5-1. Elbow Method to determine the number of clusters for a supermarket**

Closing, we performed clustering using SQL Server data tools of Visual studio and we utilized R programming language[1] to compute the SSE and determine the optimal number of clusters to split the dataset.

### 5.1.4. Visit segmentation

Figure 5-2 shows the final cluster diagram for a supermarket. The more densely populated clusters have darker color. The intensity of the line's shading that connects one cluster to another represents the strength of the similarity of the clusters.

[1] https://www.r-project.org/

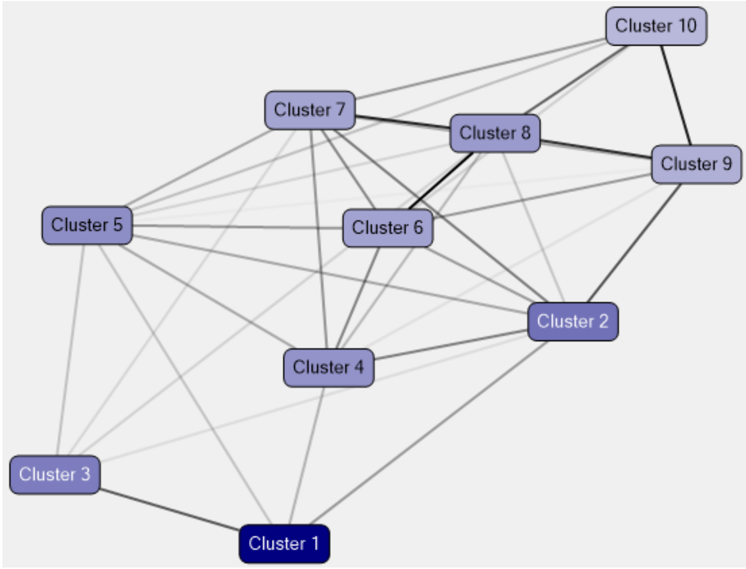



**Figure 5-2. Cluster diagram for a supermarket**

We have calculated the following descriptive statistics per visit segment and visit for translating the findings in a business meaningful way: (a) percentage of baskets (visits) per cluster (visit segment size), (b) average basket size in terms of items (visit volume), (c) average number of distinct product categories per basket (visit variety); and (d) average value in Euros per basket (visit value). For instance, Table 5-3 shows that cluster 2 includes shopper visits with 7,88 products (visit volume) that belong to 4,6 product categories (visit variety); and cost 14,67€ on average (visit value). Moreover, this cluster contains 12,04% (visit segment size) of the total shopping visits in this supermarket in a year and a half. Also, Table 5-3 depicts the percentage of visits that took place during each part of the day (morning, afternoon, evening) per each visit segment. For instance, 43,12% of the shopping visits, where customers entered the store to buy breakfast, took place in the morning. The darker the "part of day" column is, the highest percentage of baskets/visits it contains in contrast to the other segments. Similarly, Table 5-4 is a heatmap depicting the percentage of shopping visits per weekday, per each visit segment. For example, 23,47% of the shopping visits with the intention to buy snacks and beverages happens during Friday. At this point, we would like to mention that the percentage of baskets regarding Sunday is low, since stores are usually closed this day apart from some exceptions e.g. before public holidays.

| Cluster/ Segment | Visit Segment Name | Visit Segment Size | Visit Volume | Visit Variety | Visit Value | Part of Day | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | Morning | Afternoon | Evening |
| 1 | Food and drink on-the-go | 21,66% | 6,3 | 3,3 | 13,66 € | 27,75% | 40,15% | 32,10% |
| 2 | Meal preparation | 12,04% | 7,88 | 4,66 | 14,67 € | 23,09% | 36,92% | 40,00% |
| 3 | Breakfast | 11,06% | 7,63 | 4,42 | 14,61 € | 43,12% | 23,01% | 33,86% |
| 4 | Snacks and beverages | 9,10% | 9,47 | 5,31 | 17,59 € | 26,43% | 33,90% | 39,67% |
| 5 | Detergents and hygiene | 9,60% | 10,03 | 5,77 | 20,59 € | 29,22% | 39,65% | 31,12% |
| 6 | Sandwich with packed products | 7,76% | 11,97 | 7,3 | 24,53 € | 28,52% | 31,00% | 40,48% |
| 7 | Light meal | 7,58% | 11,36 | 6,51 | 19,37 € | 28,82% | 39,08% | 32,10% |
| 8 | Sandwich with fresh-cut products | 8,50% | 12,53 | 7,93 | 25,71 € | 38,31% | 33,66% | 28,02% |
| 9 | Extended visits around food | 6,69% | 19,66 | 11,54 | 36,43 € | 29,62% | 38,27% | 32,11% |
| 10 | Extended visits around non-food | 6,02% | 26,01 | 15,06 | 49,66 € | 32,10% | 38,14% | 29,76% |

**Table 5-3 Clustering descriptive statistics of a supermarket**

Cluster 2 mainly contains product categories related to fresh vegetables, red meat, chicken, white cheese, pasta, eggs, bread, oil, vinegar etc. According to the contribution of these categories (i.e. frequency of appearance in the baskets), we infer that this cluster represents visits where the shopper's mission is "meal preparation". According to Table 5-3 shoppers enter the store to purchase products for meal preparation mainly during afternoon and evening. In addition, this visit segment (as shown Table 5-4) is purchased almost equally each weekday. Alike, cluster 3 contains dominant categories, such as milk, baked goods, juice, coffee, tea, cereals, and oral care products. Thus, we can attribute these store visits to shoppers wishing to purchase for "breakfast". We see that according to the baskets in this cluster, shoppers purchase together the products to make their breakfast (e.g. coffee, cereals etc.) and the ones to wash their teeth (oral care category as a daily morning habit) in the same shopping trip. Such outcomes reveal hidden shopper behavior insights that can be useful for marketing purposes. In addition, according to Table 5-3 shoppers enter the store to buy breakfast mainly during morning. Also, based on Table 5-4 Monday is the weekday that this visit segment scores the highest percentage. This is an interesting outcome as the marketing team of the collaborative retail chain informed as that each Monday they make discounts on milk, and thus, we realize that the discount on this category increased the rest visit segment.

Similarly, biscuits, chocolates, beverages, ice creams, beers, soft drinks, chips, nuts are the prominent categories in cluster 4. These shoppers visit the store to buy their "snacks and beverages". Moreover, by examining the days that shoppers visit the store for "snacks and beverages", we found that Friday and Saturday evening are the prevailing days. Also, this visit segment scores high at Sundays, as the stores are usually open during these days before public holidays (e.g. Christmas, Easter, Carnival). Cluster 1 also contains biscuits, chocolates, chips, coffee, soft drinks and water. The first impression is that it resembles a lot with the "breakfast" and the "snacks and beverages" visit segments, but a more thorough examination showed that it contains products only from 3,3 product categories on average. The domain experts came again in our assistance and we, finally, recognized in this segment shoppers that pass by and pick up some food products for immediate consumption. The descriptive statistics in Table 5-3 shows that this cluster has the biggest size. We can attribute this fact to the position of this supermarket; it is nearby many companies and, perhaps, many employees buy products that can eat and drink quickly during breaks. In addition, according to Table 5-3 and Table 5-4 the most visits regarding food and drink on-the-go segment, take place during working days and mostly during noon probably at employee's lunch break.

Cluster 5 represents store visits with non-food products, mostly "detergents and hygiene". More specifically, the dominant categories are powders, dish washing, bathroom cleaners, paper rolls, shampoos, body creams, oral care etc., and these visits happens mostly during afternoon. Cluster 6 appears to involve visits where shoppers look for products to prepare a sandwich with "packed products", as the dominant categories are packed cheese, packed cold cuts and packed bakery products. These visits take place mainly during evening and with a more thorough examination we found out that there is a peak the hour before the stores close. Cluster 8 contains almost the same categories with cluster 6, but this time the products are fresh-cut instead of packed. Thus, we refer to this cluster as "sandwich with fresh-cut products". These two segments look a lot alike, but their shoppers have distinct shopping behaviors. In the first one, we may assume that shoppers have time restrictions and they choose not to spend time at the deli counter. On the contrary, the second segment concerns shoppers who value freshness more and are willing to wait at the queues. Also, these visits happen mostly during the morning, thus, probably here WE have shoppers that have more time e.g. more elder housewives. Cluster 7 represents shoppers who visit the store with the intention to buy products for a “light meal”. More precisely, they pick pasta, rice, pulses, vegetables, white cheese and canned food, but not meat. These visits take place mainly during working days and mostly during afternoon. Finally, clusters 9 and 10 indicate more abstract shopper visits. The first one concerns visits for food products, meaning that they visit the store to purchase and store food in general, and the second one visits containing many non-food products. So, both segments refer to more abstract shopping missions/purposes of visits that take place mainly during Saturday afternoon. We performed drill-down in both clusters, but the results did not reveal any further hidden visit segments. In more detail, the occurring sub-clusters either contained the same segments as those that were mentioned before, or they didn't indicate a certain shopping purpose.

Overall, the visit segments per store type shared similarities. We observed that the visit segments –and, thus, the customer shopping missions- are becoming more abstract, as the size of the store grows. This way we confirmed shoppers’ statement during the explanatory focus groups research. Hence, the segments of the mini hyper-store type were more abstract than those of the other two store types. In the mini-hyper stores, we mined many shopping visits related to food-oriented missions, such as “meal preparation”, “breakfast”, “snack” etc. Still, we identified some segments with different shopping purpose, such as “biological products”, “sweet preparation”, “snacks and animal feed” and “semi-prepared food”. In turn, the convenience store gave us smaller visit segments in terms of items and revenues, and the visits were more targeted, mainly around food, and there was a lack of extended visits. Some of the identified segments were snacks, soft drinks and alcohols, snacks and beverages, sandwich with packed products, light meal, breakfast, food-to-go, house cleansing and personal hygiene etc.

| Visit Segment Name | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|---|
| Food and drink on-the-go | 16,50% | 16,40% | 16,87% | 17,49% | 18,96% | 12,37% | 0,41% |
| Meal preparation | 16,37% | 15,87% | 15,89% | 16,29% | 16,28% | 18,87% | 0,42% |
| Breakfast | 19,76% | 16,77% | 14,94% | 15,83% | 16,21% | 16,28% | 0,20% |
| Snacks and beverages | 15,03% | 13,93% | 14,24% | 15,38% | 23,47% | 17,42% | 0,52% |
| Detergents and hygiene | 17,43% | 15,97% | 16,15% | 16,37% | 15,92% | 17,86% | 0,30% |
| Sandwich with packed products | 17,87% | 16,30% | 14,85% | 15,36% | 16,70% | 18,55% | 0,37% |
| Light meal | 17,61% | 17,48% | 15,91% | 17,16% | 16,21% | 15,26% | 0,37% |
| Sandwich with fresh-cut products | 18,49% | 16,12% | 15,02% | 14,86% | 16,32% | 18,83% | 0,37% |
| Extended visits around food | 17,38% | 14,63% | 13,85% | 13,97% | 17,37% | 22,44% | 0,35% |
| Extended visits around non-food | 15,94% | 13,65% | 12,94% | 13,38% | 17,87% | 25,68% | 0,54% |

**Table 5-4. Percentage of visits that take place per visit segment per weekday**

Similarly, the visit segments per channel shared similarities. However, in the web store we didn't identified patterns in the days related to each segment, whereas we detected such patterns in the physical stores. Table 5-5 shows the resulting segments and descriptive statistics/KPIs regarding the web store. In more detail, cluster 1 mainly contains product categories related to baby foods, kids' food, diapers, fresh fruits, fresh vegetables, fresh milk, yogurt, cheese, eggs, fresh fish, fresh meat etc. According to the contribution of these categories (i.e. frequency of appearance in the baskets), we infer that this cluster represents visits where shoppers visit the web store to buy "kids and babies' products and fresh food". Alike, cluster 2 contains dominant categories, such as prepared meals, canned vegetables, nuts, canned fish, eggs, frozen sea food, cold cuts, semi-prepared meals. Thus, we can attribute these store visits to shoppers wishing to purchase "semi-prepared meal". Cluster 3 appears to involve visits where shoppers look for products to prepare their meal, as it contains product categories related to fresh vegetables, cheese, fruits, eggs, fresh meat, chicken, yogurt etc. Cluster 6 also contains mostly "light meal", including products such as pasta, tomato products, rice, cheese, hot beverages, but not meat as the previous one. Similarly, fresh milk, muesli cereals, cold cuts, toast bread, yogurt, pastry, juices, salty snack, sugar confectionery are the prominent categories in cluster 4. These shoppers visit the store to buy their "breakfast and snacks". Cluster 5 represents store visits with non-food products, mostly "detergents and hygiene" e.g. paper, surface cleaners, body care, dish washing, bathroom cleaners, paper rolls, fabric cleaners, oral care etc. Similarly, cluster 7 contains water bottles, spirits, wine, refreshments, traditional desserts, party equipment etc. According to the contribution of these categories these shoppers enter the web store to purchase "spirits and beverages".

| Cluster | Name | Size | Volume | Variety | Value |
|---|---|---|---|---|---|
| 1 | Kids and babies' products and fresh food | 15,56% | 32,43 | 9,53 | 60,91€ |
| 2 | Ready/semi-prepared meal | 13,66% | 38,81 | 11,34 | 70,32€ |
| 3 | Main course | 17,37% | 51,97 | 22,18 | 90,85€ |
| 4 | Breakfast and snacks | 12,07% | 37,90 | 16,45 | 66,02€ |
| 5 | Detergents and hygiene | 19,92% | 44,62 | 17,95 | 74,49€ |
| 6 | Light meal | 9,01% | 69,84 | 18,56 | 119,14€ |
| 7 | Spirits and beverages | 12,41% | 31,43 | 5,00 | 50,53€ |

**Table 5-5. Clustering descriptive statistics – Web grocery store**

## 5.2. Case B: A Fortune 500 specialty retailer - DIY retailing

Here, we have a low value in visit variety combined with a high value at the yearly visits. These two factors lead us to form an intermate unit of analysis as described below in detail. As well, visit factors i.e. number of visits, and time between visits affected our analysis. Also, in this case we only have data from physical stores, and we examine all stores together, as a limitation was that we didn't receive a unique store identifier. In more detail, regarding dataset volume and variety, we received two-year POS data, of various stores of a Fortune 500 specialty retailer of home improvement, and construction products – also known as do-it-yourself (DIY). We analyzed 1.590.649 records. Each visit was associated with a cardholder; hence, we could identify all the baskets a shopper had purchased through his history. Table 5-6 shows details of the available dataset.

Data veracity didn't affect us neither in this case, as company's experts had already cleaned the dataset; thus, we only cleaned the 2% of the initial dataset. As we can observe at Table 5-6, shoppers purchase on average, 3,16 unique stock keeping units

(SKUs) in a single store visit. Thus, the unit of analysis could not be shoppers' single visit, as a visit involves a narrow variety of products, and consequently we will not recognize sufficient product affinities, to identify the shoppers' goals. Thus, we also checked out as unit of analysis all shopper's visits through their purchase history, to examine whether we should apply a traditional shopper segmentation approach. However, a shopper performs on average 12,59 store visits within a year. This means that on average a shopper has purchased at about 39,78 SKUs during his yearly history. The wide variety of products combined with the particularity of the domain; that implies that shoppers need to perform multiple visits to both procure materials and obtain ideas for what materials are available and how they could use them to a project/mission (e.g. to paint their house) (Wolf and McQuitty, 2011), lead us to identify that we cannot consider all shopper's visit as the unit of analysis; but, it needs an intermediate one, which contains shoppers' "x" sequential in time visits. This unit simulates the project(s)/mission(s) shoppers seek to accomplish.

| | |
|---|---|
| Unique SKUs-barcodes | 111.916 |
| No. of records | 1.590.649 |
| No. of baskets/visits | 503.857 |
| Average basket volume | 6,76 |
| Average basket variety (in SKU) | 3,16 |
| Average no. of yearly visits | 12,59 |
| Average basket value (in $) | 87,87 |
| No. of cardholders | 20.000 |
| Loyalty card usage | 100% |

**Table 5-6. DIY dataset identity**

The logic behind the identification of the intermediate unit of analysis is the following. Each shopper visits the store with his own visiting rate to accomplish a project/mission (e.g. to decorate his garden). If we could calculate shopper's visiting rate and find out the cases this visiting rate diverges from the average shopper's rate, then we could identify bundles of customer's "x" sequential visits, named in our case as super-visits. Let's assume that a shopper visits the store every ten days, then he interrupts his visits; and he starts going to the store again after one month. This means that before the pause event he had in mind a specific project to accomplish and then another project/mission, so he made two super-visits. The graphical representation of the above logic is shown in Figure 5-3.

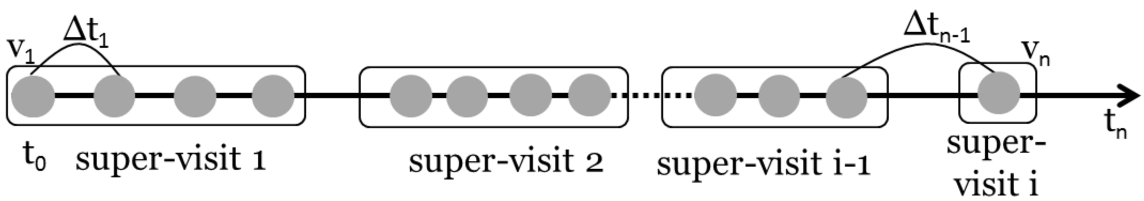



**Figure 5-3. Super-visits creation**

More specifically, in this figure we can notice all the visits of a shopper makes during the available time span. This shopper (S1) has bought from the store for the first time at date $t_0$, and for the last time at date tn. Each dot in the time axis represents a visit; thus, S1 has made n visits. For each shopper's visit we calculate the time difference (Δt) from his prior visit, so n-1 Δts will occur. Then, by eliminating Δt outliers (ΔTmin, ΔTmax), we calculate the average visiting rate for each shopper (AvgRateS1). The super-visits are created as follows: for each shopper we incur all his visits, if the Δt of the current visit is less or equal to AvgRateS1, then this visit could be viewed as a bundle with the previous one, and with the x precedent visits that satisfy this restriction. Else, we create a new super-visit. Furthermore, there are some exceptions derived from the above logic. These are customers that are not a lot of time active; for example, they visited only 4 times the store in one month ($\Delta t_n$-$t_o$=30), for a specific reason, and then they didn't visit the store again. We identified these outliers and merge all visits per shopper which falls in this case into one super-visit.

Following the above logic, 141.154 super-visits had been formed from the 503.857 actual visits. Thus, for each customer, all his visits have been grouped in 7,1 super-visits on average. Moreover, for each shopper, we calculate his active time span (Δtn-t0) and then we eliminated the outliers and create one super-visit for each customer that his Δtn-t0 less than 104 days. Also, we formed an initial set of 94 new-customized categories product categories and afterwards we increased these categories to 117. According to the cluster sampling, we kept the units of analysis (super-visits) with 2 to 50 custom categories. This means that 79,5% of the created super-visits were kept, that correspond to 91,01% of the total revenues. Here we should mention that in this case we didn't had a unique identification of each store; thus, we analyzed all stores together. According to the first clustering results, and we reconsidered the 94 custom categories and we created 117 custom categories (see Appendix B: Product taxonomies structure per case studyfor more details). K-means algorithm has been executed and we split the dataset into 5 clusters. Figure 5-4 depicts the cluster diagram.

In this case we have 5 generic visit segments (Table 5-7). This means that in the DIY stores, shoppers make more targeted visits and have more core projects/missions to accomplish. According to the clustering results, cluster 1 contains the 16,04% of the total units of analysis. It contains in high percentages, product categories related to flowers, soils and mulch, garden chemicals, fertilizers, planters, seeds, watering and lawn accessories. So, according to the contribution of the percentages that these categories have, this cluster refers to visits that shoppers want to do their "Gardening". Moreover, by exploring the other categories that are contained in this cluster we could notice that with the gardening products there are also other categories, such as BBQ grills, garden hard shapes, products for exterior decoration, outdoor furniture, and fencing, that supplements the “Gardening” project; thus, we can say that shoppers make 11,5 continuous store visits to do the “Gardening and Exterior Decoration” project and these visits happens between 19,33 days.

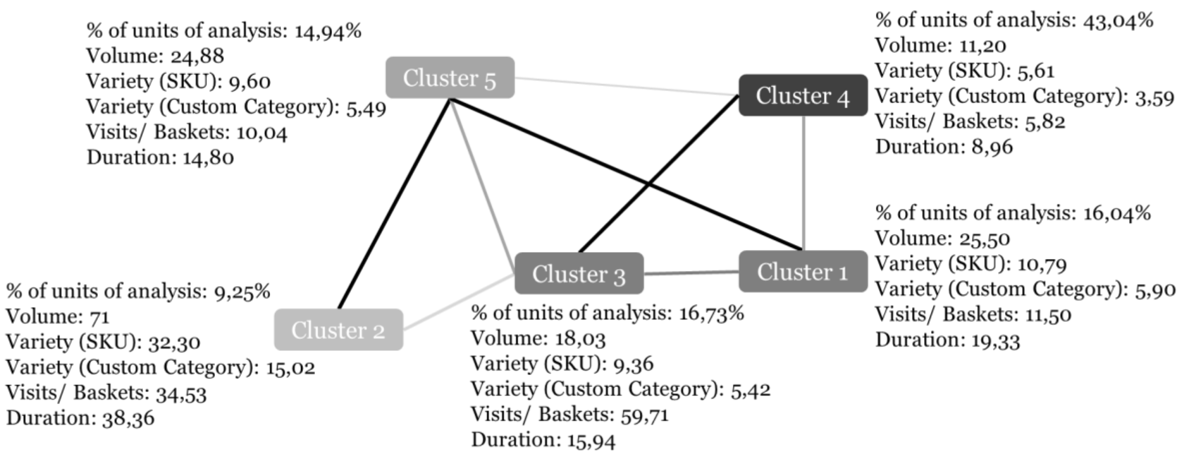



**Figure 5-4. Cluster diagram - DIY**

Cluster 2 contains visits whose target is to accomplish a “paint” project. In this cluster there are collected in high participation percentages products categories such as paint applicators, interior paints, paint tapes, pain buckets and tarps, caulks, concrete and gypsum, paint adhesives, wall abrasives, paint spray, paint safety and paint tools. Shoppers visit the store on average 32,53 times to accomplish this painting project. Examining the descriptive statistics, we can say that this is an intensive project that demands lots of voluminous store visits. Additionally, in this cluster the snack and beauty products hit their highest participation percentage. In the same context, cluster 3 contains all the tool equipment for special house works. The dominant categories are hand and electrical tools, security electrical devices, fasteners, power tools, plumbing pipes, plumbing and watering accessories, builder’s hardware, lightning, wiring, and conduit boxes. Thus, we can assume that the underlying project of shopper is "Tool Equipment for Electrical and Plumbing Works". Cluster 4 contains lots of categories in low percentages (lightning, cleaning, chemicals, security equipment, plumbing pipes and fitting, plumbing accessories, watering, builder's hardware, wiring devices). The above is explained from the fact that, super-visits that exist in this cluster contain a narrow variety of product categories, only just 3,59, thus it is considered as a cluster that represents targeted visits. Possible projects that this cluster implies are "lightning and plumbing". Finally, cluster 5 includes products related to electrical tools and equipment for woodworking. More specifically, it contains in high percentages categories such as fasteners, woodwork, wood boards in different sizes, sandpapers, woodwork prime, woodwork tools and equipment and security equipment.

| Cluster | Name | Size | Volume | Variety | Variety (SKU level) | Visits | Project Duration |
|---|---|---|---|---|---|---|---|
| 1 | Gardening and exterior decoration | 16,04% | 25,50 | 5,90 | 10,79 | 11,50 | 13,33 |
| 2 | Paint | 9,25% | 71, 00 | 15,02 | 32,30 | 34,53 | 38,36 |
| 3 | Tool equipment (electrical and plumbing) | 16,73% | 18,03 | 5,42 | 9,36 | 59,71 | 15,94 |
| 4 | Lightning and plumbing | 43,04% | 11,20 | 3,59 | 5,61 | 5,82 | 8,96 |
| 5 | Woodworking | 14,94% | 24,88 | 5,49 | 9,60 | 10,04 | 14,80 |

**Table 5-7. Clustering descriptive statistics – DIY stores**

## 5.3. Case C: Fashion retailing

In contrast to the previous cases, here we have low basket variety and visit feature is too low. Though, visit is critical, as it affects the unit of analysis i.e. all shopper's visits. Also, we can examine and compare cross-channel behaviors. Giving some more details, regarding dataset volume we received one-year POS data (1.590.649 records), from one physical store of a European fashion retailer and the transactions of his web store. Concerning data variety, we received: POS data, data for the product taxonomy tree, cardholders' demographics e.g. gender and age, data regarding promotions e.g. we could track whether each transaction was promo-driven, garments' data e.g. color, size. Data veracity didn't affect our analysis as we only cleaned the 5% of the initial records. Table 5-8 depicts some descriptive statistics to better understand the dataset.

| | Physical Store | Web store |
|---|---|---|
| Unique SKUs-barcodes | 240.075 | 239.734 |
| Dataset Volume (in records) | 1.541.339 | 493.101 |
| No. of Baskets/Visits | 407.972 | 183.320 |
| Basket Volume | 3,04 | 4 |
| Basket Variety (in SKU level) | 2,10 | 1,9 |
| No. of visits (per cardholder) | 4 | 3 |
| No. of cardholders | 10.000 | 52.736 |
| Loyalty card usage | 94% | 60% |

**Table 5-8. Fashion retail dataset identity**

As shown, shoppers buy on average 1,9 and 2,10 unique product categories per visit. Hence, the unit of analysis could not be shoppers' single visit, as a visit involves a narrow variety of products, and thus we will not recognize sufficient product affinities, to identify the visit segments. After rejecting the single visit as unit of analysis, we examined whether we will form an intermediate unit of analysis, as happened in the previous case. However, the narrow variety of each store visit didn't allow us to form valid super-visits. In more detail, shoppers had purchase on average 8,20 product categories throughout their purchase history in physical store (5,7 in web store). Thus, in this case we present a limitation of the visit segmentation approach, were we should move into traditional shopper segmentation, using all shoppers' visits (or else shopper) as unit of analysis. Still, even by proving that in this case we cannot talk about visit segmentation; we follow the rest steps proposed in chapter 4.

Examining the given product taxonomy, we formed 120 new-customized categories. According to the cluster sampling, we kept the units of analysis/shoppers that have purchased more than 2 custom categories. This way we eliminated the 9% of the records of the physical store (18% of the web store). Thus, the respective variety of purchased products, was increased. Additionally, we eliminated those baskets that were not associated with a cardholder id. Finally, after receiving the first clustering results, due to the high level of abstraction chosen from the previous phases, and due to the special factors of the domain, it was difficult to interpret them. Thus, based on experts' opinion, we decrease the level of abstraction and via disjoining some custom categories we formed 160 new categories.

Concerning the physical store, the algorithm split the dataset in 7 clusters. Descriptive statistics (Table 5-9) have been calculated to aid us with the characterization of shoppers' visits. Moreover, we calculated descriptive regarding the gender of the cardholders, the colors of the garments, and the percentage of items that have been purchased in each segment and was in promotion. The first shopper segment corresponds to the 20,19% of the shoppers in the physical store. Examining the clothes, they purchase, it is a women prevailing segment. It contains women that buy mainly casual clothing, and they complete the outfit with scarfs and different accessories and they also prefer to purchase some products for their house. Also, according to the distribution of the ages this is the “youngest” women segment and it is the only women segment that have not bought throughout their purchase history any men clothes. Furthermore, we can notice that the 27% of the purchased items is a result of promotions. This percentage may seem high, but in contrast to the other customer segments, this one is not so prone to marketing actions. Likewise, we calculated that the 11,21% of the cardholders in this cluster are men, or probably women using their husband’s or father’s loyalty card.

Cluster 5 contains a lot of garments in large sizes. Shoppers buy garments only for themselves, and products for their house. Moreover, the only men clothing in they purchase are socks. As well, in this cluster the percentage of cardholders that are men is the lowest of all the other segments. Also, they are not so prone to promotions. Cluster 3 is a men prevailing segment. Shoppers buy garments for their whole year outfit. They visit the store at about twice a year, but they buy more products than the other segments per visit. They are prone to promotions and they also buy some house products, but not as much as the previous segments. When they buy women clothes, it is mostly shirts, underwear and leather accessories. In cluster 2 we have another young segment. The driver of their store visits seems to be children and baby clothing; however, shoppers also buy underwear, nightwear and accessories. The rest identified segments are shown at Table 5-9. Finally, regarding the colors of the garments per customer segment, we observed that colors like denim, black, grey, blue etc. score high in men prevailing segments. In contrast, colors like emerald, pink/lilac, silver score high in women prevailing clusters. Likewise, large size women prefer extravagant colors like emerald, silver, multicolor purple, fashion colors, bright green, petrol, mint etc. Also, they prefer black colored garments.

| Cluster | Name | Size | Volume | Variety | Visits | Men Cardholder | % of promotions |
|---|---|---|---|---|---|---|---|
| 1 | Younger woman casual outfit | 20,19% | 5,0 | 3,2 | 2,0 | 11,21% | 27,3% |
| 2 | Children clothing | 17,40% | 5,5 | 3,1 | 1,9 | 10,15% | 28,4% |
| 3 | Men all year clothing | 17,69% | 8,0 | 3,5 | 2,1 | 40,38% | 51,8% |
| 4 | Married woman professional outfit | 13,02% | 14,7 | 7,5 | 5,0 | 24,00% | 34,2% |
| 5 | Large size woman | 12,70% | 20,5 | 9,5 | 7,0 | 7,39% | 27,8% |
| 6 | Couple's draft clothing | 10,06% | 32,4 | 13,5 | 9,6 | 32,35% | 37,6% |
| 7 | Elder woman casual outfit | 8,94% | 49,5 | 18,9 | 14,0 | 14,24% | 28,0% |

**Table 5-9. Clustering descriptive statistics – Physical fashion store**

Likewise, we formed the different shopper segments derived from the e-shop. Regarding the identified segments, shoppers do not complete the outfit with accessories, scarfs, belts, hats etc., as it happens in physical store, but it could be a great potential for promotions and recommendations. Moreover, women and men do not buy clothes for their partners. Further, we cannot identify couples' segment, neither woman that buy professional outfit, nor elder women outfit. Lastly, baby and kids clothing have too low penetration. At this point we should mention that we can detect that shoppers purchase more products per visit in web than in the physical store. But, their basket variety is lower than the physical store, thus as is also implied by the segmentation results shoppers appear to make more targeted purchases in the web channel.

# 6. THE IMPACT OF VISIT SEGMENTATION ON SHOPPER MARKETING

At the beginning of this chapter we evaluate the resulting visit segments and thus the shopping missions. To evaluate the impact of the results of our approach we conducted semi-structured focus groups to discuss with the actual store shoppers and ask for their view on the resulting shopping missions. Also, we designed a field study in the store to evaluate the resulting data-driven shopping missions and asses their validity. For that reason, we utilized a mobile app and we distributed coupons. We demonstrate that the shopping mission-related disseminated coupons achieve higher redemption rate and are claimed by a shopper into less time than the non-related coupons.

The last step of a category management is CM tactics. CM tactics may include (Hübner and Kuhn, 2012) assortment planning, store layout planning, space allocation, pricing, promotional activities and logistics planning (Lindblom and Olkkonen, 2008). All these shopper marketing-related decisions can be revamped using the visit segmentation results which indicate shoppers' missions. Thus, after examining the impact of the visit segments in the field study, we highlight the need to move from traditional category management practices to shopping mission management.

Then, we present a series of data-driven innovations in shopper marketing the resulting shopping missions could support. Closing, we present an alternative way from shopper segmentation using the resulting visit segments. Thus, we illustrate how these visit segments could be used as the cornerstone to perform shopper segmentation for more effective shopper marketing. To achieve this, we utilize customer loyalty data and we combine them with the resulting visit segments. This alternative way of segmentation could be exploited by the marketeers for more targeted and innovative marketing actions.

## 6.1. Shopping missions' evaluation via a field study

To evaluate the impact of the results of our approach we designed a field study for a major grocery European retail chain. Firstly, we analyzed one-year transactional/POS data from all the shopper of one grocery store to identify the shopping missions that shoppers perform during visiting each store. Then we conducted semi-structures focus groups to discuss with the actual store shoppers and ask for their view on our resulting shopping missions. Then, we designed a field study in the store to evaluate the resulting data-driven shopping missions and asses their validity. To achieve this, we exploited two different means i.e. a mobile app and a survey using hardcopy questionnaires. In more detail, we conducted a pilot study and we approached store shoppers to participate in it. While users shopped and navigated in this store, they used a custom mobile application which disseminated various coupons. Then, at the store exit they filled a short questionnaire. Via this field study we prove that shoppers confirm the identified data-driven shopping missions. Also, to enhance shopping mission's validity we demonstrate that the shopping mission-related disseminated coupons achieve higher redemption rate.


**Figure 6-1. Shopping mission evaluation phases**

The pilot run for one month, 87 shoppers participated, the 45% of them were men and the rest (55%) were women. Only the 21% lived alone and the 68% were married. The age distribution is shown in Table 6-1. Here we should admit that since a smart phone was a prerequisite for the pilot we weren't able to gather more shoppers that were more than 56 years old. However, the distribution of the rest pilot shopper ages follows the actual distribution of the typical store shoppers.

| Age range | Percentage of shoppers |
|---|---|
| 18-24 | 15,38% |
| 25-35 | 42,31% |
| 36-44 | 29,49% |
| 45-55 | 11,54% |
| 56+ | 1,28% |

**Table 6-1. Pilot shoppers - age distribution**

### 6.1.1. Resulting shopping missions

The categories that discriminate (i.e. are statistically significant) the first cluster are fresh vegetables, red meat, poultry and fish from the counter, bread slices, pasta, packed salad, bread, cheese, deli etc. (Table 6-2). By examining these product categories, we can assume that these shopping visits happened for the mission "main course". Also, we can extract some descriptive statistics for this mission (see Table 6-3).; for instance, the 15,4% of the total store baskets happens to satisfy this need, these baskets contain on average 15,8 products (basket volume), from 7,4 product categories (basket variety) and cost 30 euros (basket value). Also, by delving deeper into the prevailing days and hours, we can observe that these visits are stronger on business days especially during Tuesday and Thursday during the afternoon (5PM – 9PM). By combining loyalty cards data, we can identify that most of the shoppers (82%) purchasing this mission are married women more than 45 years old. Here, we must admit that each resulting mission contains more products than those we denote in Table 6-2; however, in this table we only present those product categories that are statistically significant and discriminate each cluster.

The second cluster contains categories such as beverages, salty snack, chips, beers, biscuits etc. Thus, we assume that these shopping visits happened for the mission "snacks and beverages". The 11% of the total store baskets happens to satisfy this need, these baskets contain on average 7,9 products and cost 18,4 euros. Also, by delving deeper into the prevailing days and hours, we can observe that these visits mostly take place during Friday at seven to nine during the afternoon. Another important finding is that this shopping mission has a peak during winter months and more specifically there is an extremely high peak during Valentine's day. The most the shoppers (60%) purchasing this mission are men. This percentage is high as on average the 70% of the shoppers in this store were women and the rest were men. Furthermore, the most shoppers of the "snacks and beverages" mission (85%) are not married as they have declared in their loyalty card and their age ranges from 18 to 33 years old.

"Pastry making" is the third mission we identified. It includes products such as kit desserts, sugar, fresh milk, flour, confectionary, cocoa, coffee, culinary aids etc. The baskets that constitute this mission cost on average 22 euros and contain 10,4 products. The prevailing visit slot is 9 to 12 and there is a peak before celebration days. Regarding shoppers, they are mainly married women, more than 55 years old. "Personal care and hygiene", as well as "house cleaning and maintenance" shopping missions, are purchased by both man and women, and there exists a peak during summer months. Their difference is that the first one is purchased mostly by younger and mostly unmarried shoppers (18-33) and the latter by older and married shoppers (45-64).

The "breakfast" shopping mission is mostly purchased during Friday at 6PM-9PM and during Saturday. It appears to drop during holidays and the age group ranges from 25 to 34 years old. “Sandwich” mission is mostly purchased by married men from 33 to 54 years old, at “12PM-2PM” and 6PM-9PM during business days. Lastly, we identified two other more abstract shopping missions containing more than 24 products (see Table 6-3). The first one contained mostly detergents, tissues, rolls, beauty products and some canned food, and the second contained products around food and in low percentages detergents and hygiene products. These missions appear to be purchased by married shoppers mostly during Saturday.

| **Main course** | | **Snacks and beverages** | | **Pastry making** | | **Health and beauty** | | **Breakfast** | | **House cleaning and maintenance** | | **Sandwich** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| fresh vegetables | 62 % | **beverages** | **35 %** | **kit desserts** | **27 %** | **shower gel** | **63 %** | **coffee** | 50 % | **house cleaning** | **65 %** | **counter cheese** | 63 % |
| **counter red meat** | **54 %** | **salty snacks** | 24 % | **sugar** | **25 %** | **shampoo** | **50 %** | cereals | 31 % | **paper tissue and rolls** | **61 %** | **deli counter** | 61 % |
| **counter poultry** | **50 %** | **chips** | 22 % | fresh milk | 25 % | oral care | 45 % | marmalade | 27 % | laundry | 59 % | **bread slices** | 47 % |
| **counter- fishmonger** | **47 %** | **beers** | **21 %** | **flour** | **24 %** | women haircare | 41 % | toast bread | 23 % | dishwashers | 56 % | pies | 19 % |
| bread slices | 44 % | biscuits | 20 % | confectionary | 24 % | facecare | 29 % | packed sliced cheese | 22 % | food storage | 55 % | packed sliced cheese | 16 % |
| pasta | 27 % | spirits | 20 % | cocoa | 23 % | deodorants | 25 % | yogurt - desserts | 20 % | house and garden | 14 % | packed ham slices | 15 % |
| packed salad | 26 % | sweets | 19 % | coffee | 23 % | sanitary protection | 13 % | cookies | 19 % | linen | 13 % | bread | 10 % |
| bread | 20 % | deserts | 18 % | culinary aids | 21 % | bodycare | 13 % | fresh milk | 17 % | paper- schools | 13 % | packed salads | 5 % |
| counter cheese | 18 % | watter | 17 % | margarine | 20 % | baby care | 12 % | long-life milk | 16 % | DIY and car | 12 % | | |
| fresh fruits | 18 % | juices & smoothies | 16 % | eggs | 20 % | conditioner | 9 % | bakery sweets | 15 % | insecticides | 11 % | | |
| deli counter | 15 % | wines | 16 % | butter | 20 % | clothing | 8 % | juices & smoothies | 14 % | | | | |
| tinned tomatoes | 14 % | fresh milk | 15 % | spices and herbs | 17 % | make up | 7 % | honey | 14 % | | | | |
| biological vegetables | 12 % | fresh vegetables | 13 % | milk cream - sandy | 17 % | men haircare | 7 % | | | | | | |
| table sauces | 10 % | beverages | 11 % | sweeteners | 16 % | perfumes | 5 % | | | | | | |
| frozen vegetables | 10 % | processed fruits | 8 % | chocolates | 16 % | accessories | 4 % | | | | | | |
| butter | 9 % | empty cans | 6 % | powder milk | 7 % | vitamins | 3 % | | | | | | |
| oils and fats | 9 % | party equipment | 5 % | | | | | | | | | | |
| dressing | 9 % | | | | | | | | | | | | |
| rice | 8 % | | | | | | | | | | | | |
| seafood | 7 % | | | | | | | | | | | | |
| ethnic food | 5 % | | | | | | | | | | | | |
| white cheese | 4 % | | | | | | | | | | | | |
| flour | 3 % | | | | | | | | | | | | |

**Table 6-2. Resulting shopping missions**

### 6.1.2. Focus groups with pilot store shoppers

A series of four focus group discussions with 32 store shoppers was conducted. The structure of the focus groups was the same as described in section 2.2.1. The discussion over shopper profile and the usage of product list during a store visit extracted almost the same insights as those described in section 2.2.1. Regarding the shopping missions that shoppers execute in the store, shoppers confirmed all the shopping missions we identified analyzing the POS data. It is remarkable that even we didn't share any insights with them they named almost all these mission in the same way as we did. We solely asked them to recall the last time they visited a grocery store and indicate the purchased products. For instance, a 30-years-old man said: "*I visit the store for salty snacks and beverages … spirits, chips, beers, cola and nuts … almost every two Saturdays to watch football with my friends*. "*I often visit the store for personal care products for urgent situations … I will only buy these and that's all, nothing more*", said, a 25-years-old woman. Also, shoppers confirmed that they perform their stock visits during Saturdays, as the actual data indicated.

Here we should mention that there were times where shoppers described "correctly" the purchased the products contained in a shopping mission, but they gave a different naming e.g. "I buy cheese, sliced turkey and bread slices for my breakfast". In our identified missions we call these shopping visit "sandwich" and there is another mission related to breakfast. However, this naming seems to be influenced by the consumption time.

In addition to the shopping missions we identified using the actual POS data, shoppers indicated additional missions such as light meal, food-to-go, baby exclusive, semi-prepared meals, bio visits, visits from promo products. For instance, a married woman mentioned, *"Beyond the weekly stock out visit, lots of times I visit the store for a light meal some fruits, light yogurt, and in general for dietary products…this happens mainly before summer* (chuckles)". An elder woman said, *"I visit the store sometime to purchase the promotions I showed in retailer's weekly promo brochure or in the television*".

Here, we should mention, that the fact that the key users (i.e. shoppers) validate the results of the proposed visit segmentation approach, indicates the interestingness of the identified patterns according to Silberschatz and Tuzhilin (1996).

### 6.1.3. Field study

#### 6.1.3.1. Setting

For the prevailing categories of each shopping mission (bold categories in Table 6-2) we created coupons. For example, we associated pastry making shopping mission with coupons related to product categories such as kit desserts, sugar and flour (Table 6-3). For one month, via a custom developed app we disseminated these coupons. Researchers were approaching customers in the entrance of the retailer's store, they aid them download the app and explained the process. Also, after each store trip/visit the researchers disseminated to the shoppers a short survey.

| Mission Name | Basket volume | Basket value | Basket variety | Baskets % | Promotional message |
|---|---|---|---|---|---|
| Main course | 15,8 | 30,0 | 8,4 | 15,4 | Buying 500gr fresh meat or fish: 30 points |
| | | | | | Buying 1kg fresh vegetables: 20 points |
| Snacks and beverages | 7,9 | 18,4 | 5,0 | 13,2 | Buying a Beer or Refreshments multi-pack: 25 points |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | Buying a pack of chips or salty snacks: 15 points |
| Pastry making | 10,4 | 22,0 | 6,5 | 9,7 | Buying 2 packs of sugar or flour: 15 points<br>Buying a pack of kit desserts: 25 points |
| Personal care and hygiene | 7,1 | 21,3 | 5,3 | 5,4 | Buying a bottle of shower gel: 20 points<br>Buying a bottle of shampoo: 20 points |
| Breakfast | 6,3 | 16,1 | 5,2 | 11,0 | Buying a coffee package: 15 points |
| House cleaning and maintenance | 9,4 | 24,9 | 5,3 | 5,5 | Buying a liquid detergent bottle: 25 points<br>Buying a pack of paper tissues or rolls: 15 points |
| Sandwich | 5,5 | 12,2 | 4,2 | 12,2 | Buying a package of sliced bread: 15 points<br>Buying 300gr counter cheese or deli: 20 points |
| Abstract detergent visits | 24,3 | 56,7 | 13,8 | 9,0 | N/A |
| Abstract food visits | 37,3 | 89,8 | 19,8 | 18,6 | N/A |

**Table 6-3. Promotional message per shopping mission and descriptive statistics**

Bellow we present the app flow. Each shopper scanned his/her loyalty card into the app. Each time a customer visited retailer's store the app identified this visit leveraging GPS (Global Positioning System) capabilities. Then, a question appeared related to the shopping mission the customer entered the store for (second screen in Figure 6-2). The shopper was able to select one or more missions. We gave this option to the shoppers as we wanted to simulate the more abstract shopping purposes. This question didn't have any effect on the coupons we displayed to the shoppers, it was merely informative to better understand shoppers' behavior and to compare it with the actual shopping mission s/he performed.

After answering to this question, the shopper was able to navigate to all the available coupons. The coupons were shorted per shopping mission (third screen in Figure 6-2), however, each shopper was able to see all the available coupons regardless the selected shopping mission. By clicking to a shopping mission, the shopper was able to view all the available coupons related to the selected mission (fourth screen in Figure 6-2). Each coupon was associated with a promotional message (see Table 6-3). For example, “buying 500gr fresh meat or fish you will earn 30 points”. The giveaway points had been defined by the retailer. According to retailer’s loyalty program, when a shopper gathered 200 points s/he earned a 6-euro gift card. For each coupon, a shopper was able to mark the claim button. Then, this coupon was transferred to the “claimed coupons” and it was active to be redeemed at the cashier only for the current in-store visit (last screen in Figure 6-2).

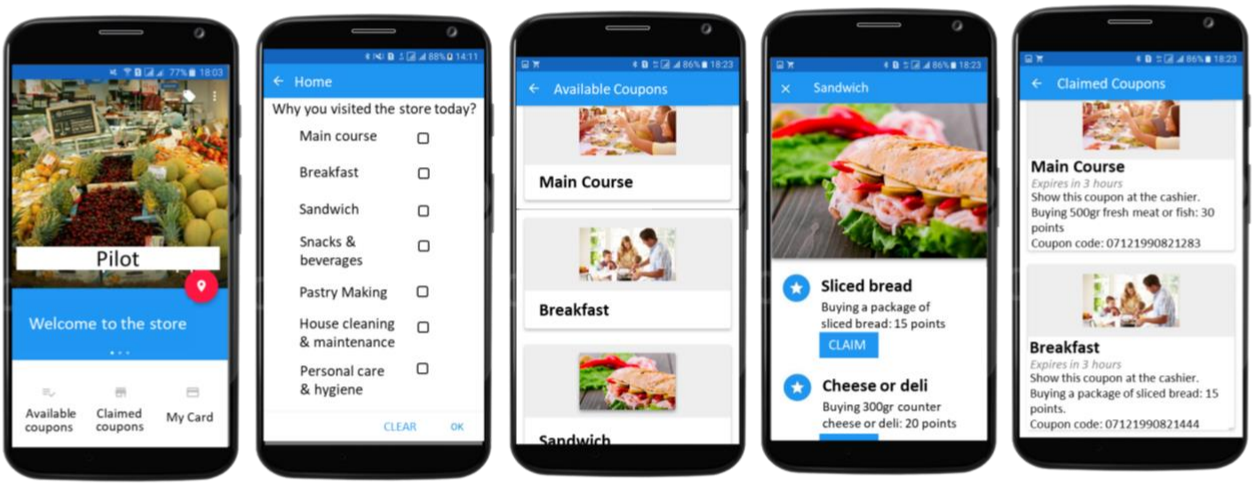



**Figure 6-2. Pilot app flow**

After paying at the cashier, the shopper participated in a survey to better understand his/her shopping behavior. Questionnaire questions are shown in Table 6-4. Here, should note that the questionnaire was distributed in hardcopy.

| |
|---|
| In your current visit, did you buy more of the products you had planned? If yes, fill in the number. |
| In your current visit, did you buy more products than you have planned due to the app coupons? If yes, fill in the number. |
| In your current visit did you have a shopping list? |
| For which of the following reasons you were planning to visit the store today? (Free selection between the identified shopping missions as shown in Table 6-3) |

| |
|---|
| After all, for which of the following reasons you visited the store today? (Free selection between the identified shopping missions as shown in Table 6-3) |

**Table 6-4. Survey questions**

#### 6.1.3.2. Data collection

For the field study purposes, we analyzed data from four different data sources (A) app analytics (B) new POS data and (C) loyalty data for those shoppers who participated in the pilot and (D) survey data. The loyalty card of each shopper was a prerequisite to use the app. Also, the researchers denoted which shopper filled which questionnaire; thus, we were able to combine all these datasets.

From the loyalty data we gathered shopper demographics, such as marital status, household size, gender and age. Using the app data, we were able to:

- track whether a shopper has claimed a coupon
- calculate the minutes passed from the app launch till a shopper claim a coupon
- identify the shopping mission(s), s/he state that entered the store for
- extract the duration of the visit

From the POS data we were able to:

- track whether a shopper has redeemed or not a claimed coupon
- calculate descriptive statistics for the basket e.g. value, variety, volume
- identify the final shopping mission

Regarding the latter, we exploited the initial model we created to identify the shopping missions and using predictive analytics we classified each new visit/transaction into an existing cluster/mission. This way we were able to identify the actual shopping mission of the customers as derived by the POS data.

### 6.1.4. Findings

#### 6.1.4.1. Relation between actual and declared shopping mission

We were able to gather information regarding the shopping mission of the shoppers in three stages:

(A) at the store entrance where shoppers declared the initial mission into the app,

(B) at the store exit (after the cashier) where shoppers declared their initial and final shopping mission via questionnaire,

(C) at the cashier analyzing the pilot POS data (this is called actual mission).

According to the results, users select on average 3,5 missions in the app during entering the store. Then, in the questionnaire, they choose on average 2,5 missions, as their initial shopping goal and 2,9 missions, as their final shopping goal. Whereas, shoppers perform only one shopping mission at the cashier. This disproportion could be explained by the fact that shoppers, in the app, selected more missions than they wanted to perform as they probably believed that the available coupons are related to this selection. Also, after their purchases they selected more than one missions, as they cannot perceive that the "abstract" shopping mission selections includes the rest.

To further analyze the relation between the initial, final, declared and actual shopping mission, we examined similarities between the different answers and the actual shopping mission. First, for each user (Ui) we created three binary vectors (Vj).

$$U\ iVj\ \forall i \in \{1, \dots ,87\}, \forall j \in \{1, 2, 3\}$$

The first vector (j=1) depicted the initial missions as selected by the user in the questionnaire. For example, the vector below $U_1V_1$= (0,0,0,0,1,0,1,0,0), means that the first user at the questionnaire selected breakfast and sandwich as his/her initial shopping mission. The order of these missions is the same as in Table 6-7.

The second vector (j=2) declared the missions as derived from the questionnaire. For instance, $U_1V_2$= (0,0,0,0,1,0,0,0,0) which means that user 1 answered only breakfast in the final shopping mission in the questionnaire.

Regarding the vector (j=3) depicting the missions declared in the app we faced a limitation i.e. the two abstract missions weren't available for selection. Thus, we performed data extrapolation to estimate the value of these not explicitly stated missions from existing information. For instance, in the cases where a user (n) had selected all the available missions i.e. $U_nV_3$= (1,1,1,1,1,1,1), we transformed this vector into $U_nV_3$= ($U_nV_3$=0,0,0,0,0,0,0,1,1), as the user has selected all the available missions which declared that s/he entered the store having an abstract shopping purpose.

Then we examined the relation between the actual shopping mission and the declared missions in the different trip stages. The 60,1% of shoppers have selected correctly their actual mission via the app when entering the store. At the 73% of shoppers, the mission declared as "initial" in the questionnaire at the store exit, was recognized as their actual mission according to the POS data. Whereas, the 75,7% of shoppers declared correctly their "final" mission in the questionnaire after the cashier. By examining these results, we confirm that most of the shoppers confirmed the identified shopping missions and this enhances their validity.

To enhance this statement, in Table 6-5 we have calculated the percentage of shoppers who have selected correctly their actual shopping mission in the different trip stages. For instance, we can observe that only the 27,3% of shoppers selected correctly that they entered the store for the "main course" mission. Most of them mainly selected breakfast, and or/ snacks and beverages. This might happen as the shoppers probably expanded their initial mission due to the coupon recommendations. However, at the store exit all these shoppers identified "main course" as their final shopping mission and almost the 91% of them as their initial mission. Likewise, in Table 6-5 we can observe that even those users that didn't identify correctly their actual mission during the different trip stages, they declared similar/relative missions e.g. instead of breakfast they chose snacks and beverages or sandwich; thus, this confirms the validity of the results.

| **Actual shopping mission (cashier/ POS data)** | **j=3 (initial shopping mission/ app/ store entrance)** | **j=1 (initial shopping mission/ questionnaire / store exit)** | **j=2 (final shopping mission/ questionnaire store exit)** | **Most selected shopping missions** |
|---|---|---|---|---|
| Main course | 27,3% | 90,9% | 100,0% | Snacks and beverages, breakfast, abstract food visits |
| Snacks and beverages | 60,0% | 60,0% | 60,0% | Breakfast |
| Pastry making | 0,0% | 50,0% | 50,0% | Breakfast, main course |
| Personal care and hygiene | 100,0% | 100,0% | 100,0% | Abstract detergent visits |
| Breakfast | 75,0% | 75,0% | 75,0% | Snacks and beverages, sandwich |
| House cleaning and maintenance | 100,0% | 100,0% | 100,0% | Abstract detergent visits |
| Sandwich | 66,7% | 55,6% | 66,7% | Breakfast |
| Abstract detergent visits | 73,1% | 59,6% | 63,5% | Personal care and hygiene, House cleaning |

| | | | | |
|---|---|---|---|---|
| | | | | and maintenance |
| Abstract food visits | 66,1% | 73,8% | 78,2% | Main course, breakfast, sandwich |

**Table 6-5. Percentage of shoppers selecting correctly their actual mission in the different trip stages**

The results above indicated the relation between actual mission and declared missions. However, the actual mission is only one, whereas shoppers declare on average from 2,5 to 3,5 shopping missions in the app and in the questionnaire. Hence, to further examine the similarity between the j=1, j=2 and j=3, we calculated Jaccard similarity to compare these binary vectors per user. Results indicated that similarity between j=1 and j=2 is estimated at 94,23%. This means that there is a little discrepancy between these two sets. Whereas, similarity between j=1 and j=3 is 78,5% and between j=2 and j=3 is 79,9%. This confirms that shoppers altered more intensively their shopping mission from the entrance and to the store exit.

The Jaccard similarity is more interesting when computing the results between j=1, j=2 and j=3 per actual shopping mission (Table 6-6). For instance, regarding "snacks and beverages", similarity between j=3 (app mission) and j=2 (questionnaire's final mission) is 80,18%, which means that shoppers entering the store for this mission seem to be more determined for their shopping goal in contrast to those entering the store for other missions.

| **Actual shopping mission** | **j=3, j=1** | **j=3, j=2** | **j=1, j=2** |
|---|---|---|---|
| Main course | 76,22% | 75,52% | 96,10% |
| Snacks and beverages | 79,23% | 80,18% | 94,29% |
| Pastry making | 73,08% | 69,23% | 96,43% |
| Personal care and hygiene | 76,92% | 73,08% | 96,43% |
| Breakfast | 77,88% | 79,81% | 91,96% |
| House cleaning and maintenance | 84,62% | 76,92% | 85,71% |
| Sandwich | 75,21% | 74,36% | 92,86% |
| Abstract detergent visits | 78,63% | 78,63% | 94,44% |

| | | | |
|---|---|---|---|
| Abstract food visits | 79,29% | 80,00% | 94,51% |

**Table 6-6. Similarity between actual and declared shopping missions in the different trip stages**

#### 6.1.4.2. Effect on impulse buying and shopping missions

During the pilot shoppers purchased more products than the average. In more detail, the average pilot basket contained 22,9 products from 12,9 different product categories that costed on 59,1€. In contrast to the average non-pilot baskets which during which contained 15,8 products (9,2 product categories) that costed 36,7€. Additionally, by examining the baskets per each shopping mission, we can observe that basket volume, variety and value were increased during the pilot almost for every mission. The above could be easily extracted by comparing the results of Table 6-7 with those of Table 6-3 -which depicts the shopping mission descriptive statistics regarding the one-year store visits.

| Shopping Mission Name | Basket volume | Basket value | Basket variety | Baskets % | Average visit duration |
|---|---|---|---|---|---|
| Main course | 13,6 | 26,9 | 8,1 | 13,79% | 35,8 |
| Snacks and beverages | 8,1 | 18,8 | 5,0 | 12,64% | 30,9 |
| Pastry making | 13,5 | 37,4 | 11,5 | 3,45% | 38,0 |
| Personal care and hygiene | 7,5 | 26,4 | 5,5 | 3,45% | 29,0 |
| Breakfast | 9,5 | 27,3 | 6,8 | 10,34% | 33,8 |
| House cleaning and maintenance | 11,3 | 27,6 | 6,2 | 1,15% | 14,0 |
| Sandwich | 10,9 | 20,2 | 7,9 | 11,49% | 34,7 |
| Abstract detergent visits | 25,4 | 68,6 | 15,9 | 11,49% | 38,4 |
| Abstract food visits | 43,7 | 116,5 | 21,8 | 32,18% | 43,0 |

**Table 6-7. Pilot Shopping missions and descriptive statistic**

The above happened probably due to two reasons: (A) shoppers who agreed to participate in the pilot were not in a hurry and were in a more spending mindset in contrast to the average shopper, (B) the app coupons urged shoppers to spend more (or make more impulse purchases) than in average. Shoppers via the questionnaire confirmed that during the pilot they purchased more products in their store visit than they have planned (see Table 6-8 below).

| Number of products | Shoppers percentage/ questionnaire question: "In your current visit, did you buy more of the products you had planned? If yes, fill in the number" | Shoppers percentage/ questionnaire question: "In your current visit, did you buy more products than you have planned due to the app coupons? If yes, fill in the number" |
|---|---|---|
| 0 | 10,0% | 21,8% |
| 1 to 4 | 48,3% | 52,7% |
| 5 to 9 | 13,3% | 10,9% |
| 10 to 14 | 16,7% | 7,3% |
| 15 to 40 | 11,7% | 7,3% |

**Table 6-8. Questionnaire questions regarding impulse product purchases**

Additionally, regarding the pilot shopping missions (Table 6-8), we can observe that the 44,86% of the shoppers executed more abstract missions (i.e. abstract detergent visits, abstract food visits) which might be also explained by (A) and (B). In addition, here we should mention that we do not have lots of visits concerning "pasty making" mission. This probably happens as shoppers who purchase this mission are mainly more than 55 years old (see section 6.1.1) and these shoppers didn't participate in our pilot (see Table 6-1). Likewise, we do not have many visits regarding "personal care and hygiene", which also makes sense, as this mission has a peak during summer months and the pilot ran during winter months. Closing, we should denote that the average visit lasts 33 minutes, however this includes 3 to 5 minutes to download the app and explain the study to the shopper. Also, in Table 6-7 it is obvious that there is a statistically significant linear relation between the average visit duration and the basket volume (r=+0,632 according to Pearson Correlation).

#### 6.1.4.3. Relation between actual shopping mission and coupons

Via the app we achieved 194 coupon claims, the distribution of these claims in the various coupons is shown in Table **6-9**. The 96,9% of these claims led to coupon redemptions. To examine the relation between the actual shopping mission and the coupon redemptions, we split our sample into two sets:

- Set A (mission-related): Those observations where the coupon redeemed is related to their actual shopping mission (as shown in Table 6-3). For example, a shopper that executed the shopping mission “Pastry making” and redeemed the coupon “Buying 2 packs of sugar or flour: 15 points” or the coupon “Buying a pack of kit desserts: 25 points”, belongs in this group.
- Set B (mission-unrelated): Those observations where the coupon redeemed is not related to the actual shopping mission. For instance, a shopper who executed the mission “Personal care and hygiene” and redeemed the “Buying a pack of kit desserts: 25 points”, belongs in this second group.

Using the app analytics, we were able to spot the minutes passed from the app launch till each shopper claimed a coupon (coupon time-to-claim). Running ANOVA between Set A and Set B regarding coupon time-to-claim we identified that the significance value is 0,0005 (i.e., $p = 0{,}001 < 0{,}05$). This means that there is a statistically significant difference in coupon time-to-claim between these two sets ($F(1,234)=47.559$, $p=0{,}001$).

Delving deeper into the data we identified that shoppers spend less minutes when the coupon they claimed was related to their actual shopping mission as derived from the POS data (see Table **6-9**). For instance, when the actual shopping mission was "sandwich" s/he claimed the coupon "sliced bread" during the first minute of the shopping trip. Whereas, shoppers spent more minutes to claim coupons unrelated to their actual shopping mission e.g. they spend on average 6.70 minutes to claim the coupon "sliced bread", when their mission wasn't sandwich. The coupons that do not follow this trend are those related to the non-food products (i.e. liquid detergent, shower gel, paper tissues or rolls) except shampoo (see Table **6-9**). This means that these coupons are purchased across the various shopping missions and show a great potential for cross-selling actions.

In the last column of Table **6-9** it is denoted the p-value, after running ANOVA between Set A and Set B based on time-to-claim per coupon. Results confirmed our aforementioned statement i.e. the non-food products (i.e. liquid detergent, shower gel, paper tissues or rolls) except shampoo, are purchased across the various missions and there is no statistically significant difference in time-to-claim between related and non-related mission-coupons.

| Coupon product | Average coupons time-to-claim (in minutes) when the actual shopping mission is related to the coupon | Average coupons time-to-claim (in minutes) when the actual shopping mission is not related to the coupon | Coupon claims distribution | Percentage of claimed coupons redeemed | ANOVA (Set A VS Set B) p -value |
|---|---|---|---|---|---|
| Fresh meat or fish | 10,6 | 18,2 | 7,7% | 100% | 0,0204 |
| Fresh vegetables | 4,8 | 11,5 | 11,9% | 100% | 0,0005 |
| Beer or refreshments | 2,1 | 7,8 | 5,7% | 100% | 0,0010 |
| Chips or salty snacks | 2,0 | 6,6 | 8,8% | 94,1% | 0,0200 |
| Sugar or flour | 5,1 | 5,5 | 4,1% | 100% | 0,0203 |

| | | | | | |
|---|---|---|---|---|---|
| Kit desserts | 6,0 | 10,0 | 3,6% | 71,4% | 0,0320 |
| Shower gel | 3,3 | 3,3 | 7,2% | 100% | 0,1650 |
| Shampoo | 1,2 | 5,5 | 3,1% | 85,7% | 0,0070 |
| Coffee | 4,3 | 9,1 | 10,3% | 100% | 0,0010 |
| Liquid detergents | 3,1 | 3,0 | 6,2% | 100% | 0,1330 |
| Paper tissues or rolls | 1,4 | 1,2 | 6,7% | 92,3% | 0,2610 |
| Sliced bread | 1,0 | 6,7 | 11,9% | 95,6% | 0,0020 |
| Cheese or deli | 4,0 | 8,2 | 12,9% | 100% | 0,0005 |

**Table 6-9. Claimed coupons - descriptive statistics**

To further examine the relation between the actual shopping mission and coupons, we calculated a matrix (Table 6-10) containing as rows each available coupon and as columns each mission. The percentage in this matrix depicts the percentage of coupons redeemed compared to the total coupons redeemed per shopping mission. The darker the table column is, the highest percentage of redeemed coupons it contains in contrast to the other shopping missions. For example, we may observe that the 22,22% of the coupons redeemed when the actual shopping mission of a shopper was “main course”, concerned the coupon fresh meat or fish. Also, the 33,33% of the rest redeemed coupons in this mission were for fresh vegetables. The rest redemptions concerned mainly other food-related coupons such as sliced bread, cheese or deli, sugar or flour, coffee and the only non-food coupon that was redeem is liquid detergent. As we can observe in Table 6-2, which presents the products that constitute each mission, all the food related products redeemed by “main course” shoppers, except coffee, are included in the respective mission. Thus, we can see a clear relation between the redeemed coupons and the actual shopping mission. Regarding “snacks and beverages” mission, the 2/3 of the redemptions concerned beer or refreshments and chips or salty snacks coupons. By examining the rest coupon redemptions in Table 6-10 we can conclude that the redemption rate is increased when we offer to the shoppers, coupons related to their shopping mission (bold values in Table 6-10). We consider as relative mainly those coupons/missions that are matched according to Table 6-3. However, as shown in Table 6-10 there are coupons mainly associated with non-food products e.g. shower gel, liquid detergents, that are redeemed in high percentages regardless shopper's mission. This validates our aforementioned conclusion that according to Table **6-9** the coupons that are related to the non-food products (i.e. liquid detergent, shower gel, paper tissues or rolls) except shampoo have no significant difference in time-to-claim between set A (mission related) and set B (mission unrelated).

At Appendix D: Relation between shopping list and shopping mission we also present the findings regarding the relation between the shopping list and the shopping mission, which is over the scope of this dissertation.

| | Main course | Snacks and beverages | Pastry making | Personal care and hygiene | Breakfast | House cleaning and maintenance | Sandwich | Abstract detergent visits | Abstract food visits |
|---|---|---|---|---|---|---|---|---|---|
| Fresh meat or fish | **22,22%** | 0,00% | 0,00% | 0,00% | 7,14% | 0,00% | 0,00% | 4,88% | 8,43% |
| Fresh vegetables | **33,33%** | 0,00% | 12,50% | 0,00% | 0,00% | 0,00% | 0,00% | 7,32% | 13,25% |
| Beer or refreshments | 0,00% | **33,33%** | 0,00% | 0,00% | 0,00% | 0,00% | 0,00% | 7,32% | 7,23% |
| Chips or salty snacks | 5,56% | **33,33%** | 0,00% | 0,00% | 7,14% | 0,00% | 14,29% | 4,88% | 8,43% |
| Sugar or flour | 5,56% | 0,00% | **12,50%** | 0,00% | 0,00% | 0,00% | 0,00% | 4,88% | 6,02% |
| Kit desserts | 0,00% | 0,00% | **25,00%** | 9,09% | 0,00% | 0,00% | 0,00% | 7,32% | 0,00% |
| Shower gel | 0,00% | 0,00% | 12,50% | **36,36%** | 7,14% | 18,18% | 7,14% | 9,76% | 7,23% |
| Shampoo | 0,00% | 0,00% | 0,00% | **27,27%** | 0,00% | 18,18% | 0,00% | 9,76% | 2,41% |
| Coffee | 5,56% | 22,22% | 25,00% | 9,09% | **28,57%** | 0,00% | 14,29% | 7,32% | 7,23% |
| Liquid detergents | 5,56% | 0,00% | 0,00% | 9,09% | 7,14% | **27,27%** | 0,00% | 14,63% | 6,02% |
| Paper tissues or rolls | 0,00% | 11,11% | 0,00% | 9,09% | 7,14% | **36,36%** | 7,14% | 7,32% | 7,23% |
| Sliced bread | 11,11% | 0,00% | 0,00% | 0,00% | 14,29% | 0,00% | **28,57%** | 7,32% | 13,25% |
| Cheese or deli | 11,11% | 0,00% | 12,50% | 0,00% | 21,43% | 0,00% | **28,57%** | 7,32% | 13,25% |

**Table 6-10. Relation between actual shopping mission and redeemed coupons**

## 6.2. From category management to shopping mission management

In contemporary retail there are various customer touchpoints through a customer journey. Thus, retailers daily gather massive amounts of data regarding their customers transactions, preferences, demographics, shoppers in-store movements etc. Simultaneously, shoppers are becoming far more demanding. The consumer-packaged goods (CPGs) marketplace is the one that facing major issues, as retailers have become largely substitutable in shoppers' minds due to offering similar merchandise (Pepe and Pepe, 2012). Thus, retailers must work harder than ever to differentiate themselves. Despite heavy investment in business analytics infrastructures that could aid more effective CM, retailers are still losing potential revenue due to their failure to get the right goods to the right places at the right price.

In this everchanging environment old CM practices seem to be cumbersome. Business executives recognize the need to incorporate the shopper behavior and needs into CM practices. However, even customer-centric CM should be revamped. This alteration is required in contemporary retail, as we observe a changing behavior of shoppers i.e. a shopper might swift his/her behavior even when visiting the same retail store. To cope with the changing behavior of shoppers, both researchers (Walters and Jamil, 2003; Bell et al., 2011) and practitioners (ECR Europe, 2011) have stressed the need to focus on each single customer visit. This way, we will capture each customer's shopping occasion (Desrochers and Nelson, 2006), need and mission (Sarantopoulos et al., 2016; Griva et al., 2018) e.g. that a shopper entered two times in a supermarket to purchase products to prepare breakfast, and then for a gourmet meal etc.

Hence, we propose not only to focus separately on each category e.g. milk, cereals, coffee etc. as traditional CM does, but to move from CM to Shopping Mission

Management (Figure 6-3). This way we will treat categories collaboratively under the shopping mission they participate. Thus, a new chapter in category management is unfolding and more effective assortment, store layout planning, space allocation and promotional activities could be achieved.

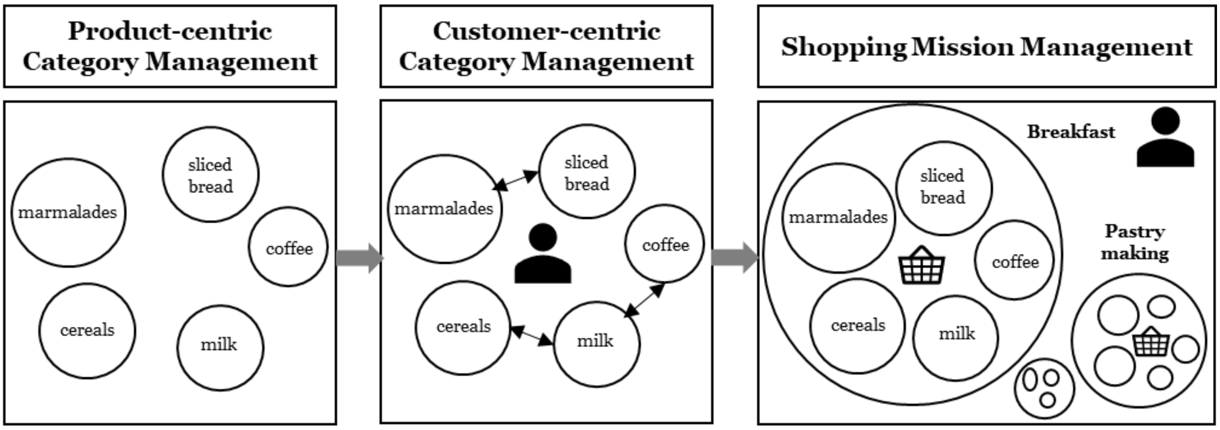



**Figure 6-3. From category to shopping mission management**

Via moving from traditional CM to Shopping Mission Management, we also propose to alter the classic CM process (right part of Figure 6-4) in a way that each shopping mission is run as a business unit in its own right, with its own set of turnovers, profitability targets and strategies. The “role of the category” step is replaced by the shopping mission identification step. Its goal is to detect the mission(s) each category belongs to and structure the role of each category in the different mission it belongs. The rest steps remain the same, though the category term is replaced by the shopping mission concept.

The last step of a category management is CM tactics. CM tactics may include (Hübner and Kuhn, 2012) assortment planning, store layout planning, space allocation, pricing, promotional activities and logistics planning (Lindblom and Olkkonen, 2008). All these shopper marketing-related decisions can be revamped using the resulting shopping missions. Thus, in the next section we present data-driven innovations in shopper marketing that the resulting shopping missions can support.

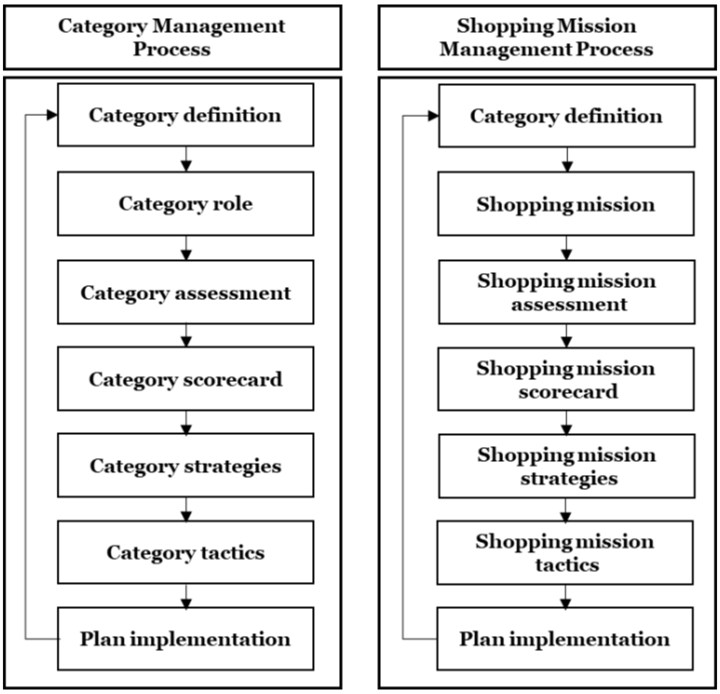



**Figure 6-4. Category management VS shopping mission management process**

## 6.3. Data-driven innovations in shopper marketing

Regarding the output and the value of such a system, it is stressed when considering the consumer-oriented business decisions it can support. Our approach could be evolved into to a tool for designing innovative marketing campaigns, bundled promotions and **cross-coupon** programs for product categories that belong to the same visit segment. For instance, ABC analysis can be applied to classify the products belonging into each segment into three classes based on product purchases. Then, alternative product **combos** and **bundling** strategies could be tested. Alternatively, our approach may become the cornerstone of a recommendation system for real-time purchases in retail stores. It will propose to the customers more products that they may have forgotten to buy, considering their prior or current visit(s). Apart from recommending products from the same cluster/visit segment, marketing managers could also exploit the knowledge extracted from the more abstract clusters to make cross-cluster promotions that fit the specific needs of shoppers. For instance, considering a shopper that visits the store to buy beverages and the recommender promotes to her/him products from the snacks cluster aiming to increase the basket value and variety. This promotion is based on the detected new knowledge that beverages and snacks are sometimes co-purchased in a broader shopping visit.

Likewise, we can create offline and online **product catalogues**. For instance, we have detected women that enter a fashion store to purchase professional clothes and baby clothes. Thus, to promote the new collection, it could be more effective to send them product catalogues that meet their specific preferences, instead of including all the available new garments. The extracted knowledge could be also valuable for advertising purposes; e.g. breakfast products advertisements. Current advertising strategies in retailing are brand, or category-oriented; thus, shopping mission-oriented advertising can be truly disruptive for the retail environment, as it reflects shoppers' deeper needs. For instance, instead of making advertisements of specific product categories, retailers could advertise bundled product categories that correspond to a shopping mission, e.g. gourmet products advertisements.

The visit segments can may support more business decisions that have a more indirect, but not less significant, impact on customer satisfaction. For example, the visit segments may dictate a new **redesigned retail store layout** where product categories in the same visit segment are positioned in nearby store aisles and shelves. Considering the bigger picture, we can move from a category-based layout to a mission-based layout that can help customers locate products in the store more easily and buy more in less time (Vrechopoulos et al., 2004; Cil, 2012; Sarantopoulos et al., 2016; Sarantopoulos, Theotokis, Pramatari and Roggeveen, 2019). Similarly, the same logic could be applied on retailer's web store. Also, as the **web** environment is more flexible **alternative layouts** based on the resulting visit segments can be tested for different seasons, or customers or even days. For example, each shopper when signs in the web store a custom layout can be depicted, based on his/her preferences.

In the same spirit, **second in-store placement** spots can be detected and really increase purchases, e.g. a mouthwash next to chocolates etc. Additionally, personalized product recommendations based on the extracted visit segments could also be designed. For example, either disseminating these **recommendations** using a mobile app, or using e-mail marketing, or even by printing coupons during the checkout process at the cashier. Also, the store managers could utilize the extracted visit segments to monitor and **benchmark** his/her store's performance with other similar stores e.g. the gourmet meal shopping mission performs 5% better in both value and volume terms in other similar stores in contrast to mine. Similarly, by comparing the products a shopper purchases during a shopping visit, with those products contained in the typical shopping mission this visit belongs to, marketeers could **detect missing sales**.

Moreover, our approach could be even evolved into a **collaborative analytics platform** that gathers transactional data from various retailers and provide them insights regarding the visit segments derived from their data and enable performance benchmarking to competitors. Simultaneously, it could disseminate to other parties, such as suppliers, insights regarding the categories and brands they sell e.g. in the context of which mission are purchased. This way collaborative analytics in demand-chain management could enhance a more Efficient Consumer Response (ECR).

Additionally, **predicting future behaviors** and missions based on historical data can support several **supply-oriented operations** e.g. product replenishment based on the identified visit segments or prediction of possible out-of-self-situations, based on visit peak days and hours. In more detail, the store manager could reengineer store operations management and **replenishment strategies** by ordering groups of products based on the identified visit segments. Or even change the shelves replenishment by time of day and day of week given taking into consideration the peaks of each visit segment as shown for example in Table 5-3.

Lastly, this approach could be even utilized to **rearrange** and modify a retailer's **warehouse**, by placing in nearby aisles products matching online orders to decrease order-picking time. This kind of rearrangement has been previously examined in the literature using solely association rule mining (Chan and Pang, 2011; Chuang, Chia and Wong, 2014).

We acknowledge that large companies, such as SAS, IBM, SAP etc., have developed commercial tools and suites e.g. IBM WATSON, IBM COGNOS, SAP HANA etc. to ease companies perform different data analyses varying from reporting to data mining (e.g. clustering). Companies utilize these suites to perform customized analyses e.g. produce customer segmentations (Chen, 2014). Our approach is complementary to such solutions. It may be treated as an additional layer of functionality on top of such software tools, for generating visit segmentations and consequently deducing customers' shopping intentions per visit.

## 6.4. From visit to shopper segmentation for more effective shopper marketing

Apart from the aforementioned data-driven innovations, the value of such a system could be further enhanced when we use the resulting visit segments to identify shopper segments. By looking into the shopping missions that each shopper performs in all the stores of a retail chain, we can boost shopper marketing activities. Shopper segmentation based on the identified visit segments/shopping missions can aid retailers identify selling gaps and opportunities and enhance personalization. To achieve this, we need to obtain access to all the POS data from all the various stores a retailer has, otherwise, any conclusion could be misleading.

To pinpoint the value of shopper segmentation based on the identified visit segments, we used point-of-sales (POS) data from a small retail chain named "XYZ" having 222 stores in the urban areas of Greece country. Retailer provided us with all the transactions that had been performed within a year by retailer's shoppers using loyalty card. Thus, we received more than 120 billion product swipes that correspond to 15 billion transactions/baskets performed by 1.120.021 shoppers. Here, we should admit that almost the 96% of the total retail chain transactions happen using loyalty card; thus, sample is representative. The average basket costed 16,7€ and contained 8,1 products from 4,9 different product categories.

Firstly, we began the analysis by identifying the different shopping missions of the shoppers as shown in Table 6-11.

| Cluster | Name | Size | Volume | Variety | Value |
|---|---|---|---|---|---|
| 1 | Sandwich | 8,8% | 4,80 | 3,70 | 8 |
| 2 | House cleaning & maintenance | 15,3% | 10,00 | 5,80 | 25,3 |
| 3 | Personal Care & Hygiene | 10,7% | 8,50 | 4,10 | 16,8 |
| 4 | Main Course | 16,4% | 14,20 | 6,90 | 25,1 |
| 5 | Fruits & Vegetables | 12,8% | 6,20 | 5,00 | 9,8 |

| | | | | | |
|---|---|---|---|---|---|
| 6 | Breakfast | 11,2% | 7,20 | 4,40 | 13,2 |
| 7 | Snack & Beverages | 11,9% | 6,50 | 4,80 | 12,3 |

**Table 6-11. Shopping missions for all the stores of a Greek retail chain**

The stores that this retailer has mainly belong to store typologies such as convenience stores and smaller supermarkets. Thus, we didn't identify abstract shopping visits, as happened in the previous case. Also, these missions are more generic in contrast to the missions identified in the previous cases. Thus, by treating the stores as a bulk we lose the unique shopping missions resulting in each store according to the store characteristics and location (e.g. next to a fleet market, next to a gym etc.).

Then we exploited the loyalty cards data, to detect the shopping missions a shopper performed during his/her purchase history. We rated each shopper (Table 6-12) with a value ranging from 1 (low) to 5 (high) according to the s/he visited the store for each mission during the whole year weeks (weeks of presence).

| | Sandwich | House cleaning & maintenance | Personal Care & Hygiene | Main Course | Fruits & Vegetables | Breakfast | Snack & Beverages | Toiletries |
|---|---|---|---|---|---|---|---|---|
| Shopper 1 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 5 |
| ... | 3 | 3 | 5 | 3 | 3 | 1 | 2 | 3 |
| Shopper N | 1 | 4 | 1 | 5 | 3 | 1 | 3 | 5 |

**Table 6-12. Shoppers' fact table**

In more detail, to calculate this value we used as benchmark the weeks of presence of all the shoppers per mission. Thus, this value is different for the various shopping missions. For example, a value equals 5 at the breakfast mission is not the same with a 5 into house cleansing and maintenance, as a shopper purchases more often breakfast than house cleansing products.

We use text mining to extract more meta-data from the available product descriptions. Thus, we identified whether a product is premium, private label (PL), and/or biological. Also, we classified the products based on their descriptions into children or elder usage. In addition, we added a binary flag in the cases that the product sold was in promo. Hence, we enriched the shopper's fact table (Table 6-12) by adding more meta-data per customer. Then, we utilized shoppers' fact table as input in the data mining model, we executed clustering with k-means algorithm and we segmented shoppers based on the missions they had performed during their yearly visits.

We shaped ten shopper clusters i.e. ten shopper segments. Each shopper is assigned to one shopper segment. Table 6-13 depicts the descriptive statistics for each resulting segment. Whereas, in Figure 6-5 is shown the cluster diagram. The most density populated a cluster is, the more shoppers it contains. Also, the lines between the clusters declare their similarity. These 10 segments could be grouped into 5 more generic segments according to shoppers' behavior and loyalty in each identified shopping mission. These hyper-segments are also denoted in Figure 6-5 and are explained in detail below.

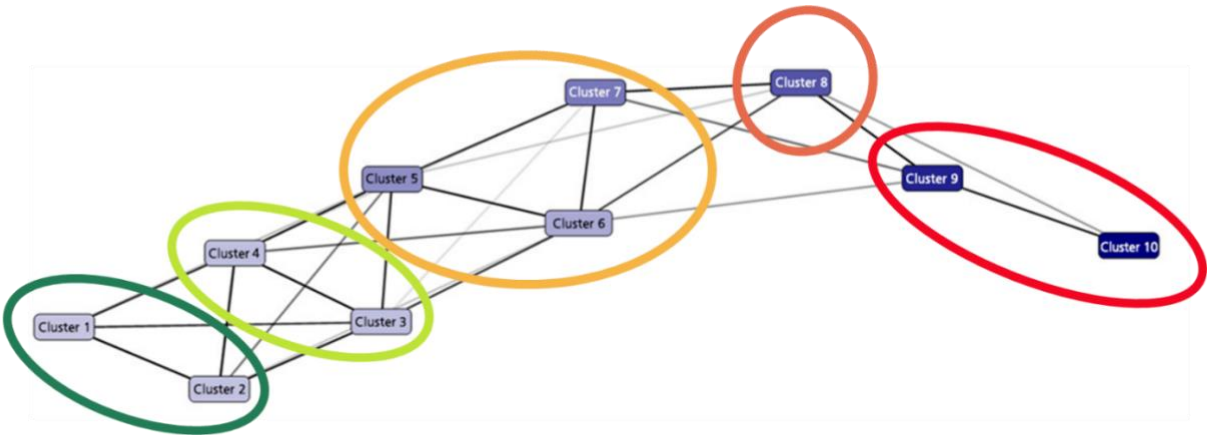



**Figure 6-5. Shopper segments based on visit segmentation**

Segment/cluster 1 contains the 4,6% of shoppers. These shoppers visited the store for all the identified shopping missions as they are rated with five in all of them. Thus, here we have the more loyal shoppers that spend more than 281,83€ per month. These shoppers purchase many private label (PL), baby and biological products. Similarly, segment 2 contains loyal shoppers, as they are rated with five in almost all the missions. However, occasionally they seem to visit other stores for missions such as fruits and vegetables and main course. This lead them spend almost 100€ less (197,58€) than the previous segment. In addition, shoppers in this segment purchase more premium products. Closing, segments 1 and 2 could be declared as "the most loyal retailer's shoppers" that seem to have this retail chain as primary for their purchases. Thus, customer retention strategies could be more appropriate for these two segments.

Segments 3 and 4 contain the 5,4% and 5,1% of shoppers respectively. Shoppers in segment 3 purchase all the available shopping missions, having a rate equals 4 in the non-food related missions and a rate equals 5 in the food related missions. The average monthly basket is 121€ and shoppers purchase lots of PL and elderly products. Shoppers in segment 4 have the opposite behavior of those in cluster 4 (i.e. rating equals 4 in food and 5 in non-food). Their average monthly basket and their characteristics are almost the same with those of segment 3. Closing these two clusters could be characterized as "the loyal shopper segments" having lower budget than the two previous. Probably these shoppers visit occasionally a second store for to satisfy the lower rated shopping missions. Hence, both customer retention and development strategies seem to be more efficient for these segments.

Segments 5, 6 and 7 are considered as "the medium-loyal shoppers". According to the results the "XYZ" retailer is not the primary retail store they visit to satisfy their needs. It is probably shoppers' second or third choice. In more detail, shoppers in segment 5 spend at about 85,60€ per month in various shopping missions, as they are rated with

3 in all of them. Also, these shoppers purchase many baby and biological products. Likewise, shoppers in segment 6 visit retailer's stores to purchase occasionally (rate value=3) shopping missions such as breakfast, sandwich, fruits and vegetables, snacks and beverages. On the contrary, they visit the store more rarely (value=2 or 1), for missions such as main course, house cleaning and maintenance, personal care and hygiene, and toiletries. Shoppers in this segment spend on average 59,48€ per month. Likewise, segment 7 contains shopper visiting the store occasionally (value =3) for non-food shopping missions such as, house cleaning and maintenance, personal care and hygiene, and toiletries. The rest missions are rated with 2 and main course is rated with 1. These shoppers spend 61,65€ per month and they purchase more premium products and many products in promo. Closing these three segments seem to be the more valuable segments the retailer needs to attract. Shopper attraction strategies are important to boost these shoppers make retailer's stores as their primary choice.

Segment 8 includes the 14% of retailer's shoppers. They mainly purchase missions such as breakfast, sandwich and personal care and hygiene. These shoppers spend on average 53€ per month. According to retailer's input the prevailing stores that these shoppers visit are close to student campus. Thus, probably students visit the "XYZ" stores as their main retail chain. The lower rating in the rest shopping missions, does not definitely indicate that these shoppers visit another retail chain, as according to experts' opinion students tend to purchase only these shopping missions in their store visits. This finding is important for the retailer in order not to spend marketing budget to promote the lower shopping missions to these shoppers as they are not related to students' real needs.

Additionally, segment 9 contains the 18,1% of the shoppers. They visit the store on average once every two months and they purchase products in promotion from various shopping missions. Almost in all shopping missions they are rated with 1. Likewise, segment 10 contains the 21% of total shoppers that have visited retailer's store only 2 times per year and purchase premium products across all missions that where in promo. Segments 9 and 10 are "the non-loyal shopper segments" that visit retailer's stores only for promo purposes.

| Segment No, | Size | Average monthly value | Average basket value |
|---|---|---|---|
| 1 | 4,6% | 281,83 | 25,86 |
| 2 | 4,9% | 197,58 | 23,71 |
| 3 | 5,4% | 121 | 23,51 |
| 4 | 5,1% | 128,2 | 23,99 |
| 5 | 9,1% | 85,6 | 23,45 |
| 6 | 6,9% | 59,48 | 17,89 |
| 7 | 10,9% | 61,65 | 22,06 |
| 8 | 14,0% | 53,95 | 15,55 |
| 9 | 18,1% | 24,33 | 7,33 |
| 10 | 21,0% | 26,05 | 5,22 |

**Table 6-13. Descriptive statistics for the identified shopper segment**

# 7. DISCUSSION AND CONCLUSIONS

This final chapter overviews the main outcomes of this research. Then, it presents and discusses the research's contribution to theoretical knowledge along with its practical value. Afterwards, we present our thoughts for visit segmentation systems designers in contemporary retail. We present in detail the data, shopper, marketing and retailer's factors that designers should take into consideration when designing visit segmentation systems. At the end of this chapter, the research limitations are pointed out and avenues for further research are recommended.

## 7.1. Research outcomes

This research journey began with the aim of advancing the understanding of visit segmentation in retail. More specifically, the following questions have been addressed:

- Q1. How can we derive visit segments from shopper data?
    - Can we extract the different shopping missions of customers from the identified visit segments?
    - Can we develop a business analytics-informed approach to perform visit segmentation?
- Q2. What are the factors that affect the design of visit segmentation systems?

To address these questions, we adopt as methodological backbone the design science paradigm (Hevner et al., 2004). Additionally, owing the lack of prior systematic research on the visit segmentation topic this research is based on multiple case studies design. Thus, initially, we started our research via revealing the need of visit segmentation in contemporary retail. Both practitioners and researchers agree that old-school shopper segmentation is not enough and cannot describe the new, volatile shopper habits and preferences. This happens since the modern shopper has changed. The shopper flits between shopping channels and performs a complex shopper journey with the purpose to satisfy his/her increasing demands for quality and value (Wood, 2018). Shopper behavior is no longer predictable; it is changing through time and, even, between shopping visits in the same store (Sorensen et al., 2017). Hence, there is a need for visit segmentation to cope with shoppers' changing behavior.

Looking into the segmentation literature, researchers view shoppers as a bulk of all shopper visits (e.g. shoppers who purchase routine products) or as associations between the products/ items purchased during a shopper's single visit (e.g. bread→ milk). We state that the aforementioned studies overlook the holistic shopping purpose, intentions and missions of shoppers, which are not the same in every (physical or web) store visit.

In this dissertation, we suggest putting the shopper visit on the spot, instead of the shopper total buying behavior that shopper segmentation relies on. Putting the visit on the spot has the potential to ensure a more accurate view of the shopper desires. To this end, we coined the **term "visit segmentation" (the first outcome)**. Visit segmentation focuses on the underlying needs that boosted a customer visit a store e.g. to purchase products for a light meal, or to procure materials to renovate their bathroom etc. These needs and missions can be extracted using various datasets reflecting shopper behavior e.g. product purchases, interactions, preferences etc.

Afterwards, to perform visit segmentation we developed **a business analytics-based approach (the second outcome)** which could be applied into various shopper interaction data. We applied, validated and refined the proposed approach through three heterogeneous retail cases. The first case concerns sales data from different channels and stores of a major European fast-moving consumer goods (FMCG) retailer. Respectively, in the second case, we produced the visit segments for the physical stores of a Fortune 500 specialty retailer of home improvement and construction products – also known as do-it-yourself (DIY) retailer. The third case concerns data from a physical and the web store of a major European fashion retailer. We analyze retail basket data from these cases and we produce groups of visits based on the product categories the customers have purchased during each visit to a physical retail or web store.

We suggest that the resulting mix of product categories that prevails each visit segment reflects the shopping intentions of the customers that held the baskets included in each visit segment. In other words, we generate segments of visits and, then, we **identify the shopping intention and missions that boosted these visits (the third outcome)**. Let's assume that the prevailing product categories purchased during the shopping visits of a mined segment are biscuits, chocolates, beverages, ice creams, beers, soft drinks and chips. Then, we conclude that the shopping intention of the respective customers was to buy "snacks and beverages".

In addition, during the application of the proposed approach in the different cases we did not overlook the significant effect of the product taxonomy in the effectiveness and validity of our data mining results. Other data analytics studies have highlighted that product taxonomies may seriously affect the knowledge discovery process and the data mining results (Albadvi & Shahbazi, 2009; Cho, Kim, & Kimb, 2002). More specifically, product taxonomies are often unbalanced and have characteristics hindering the performance of data mining algorithms. Thus, it matters for example whether we should refer to a can of sparkling orange juice of brand XYZ as sparkling beverage, as beverage, or as orange juice. For that reason, this research also **suggests a semi-supervised feature selection approach (the forth outcome)** that uses the product taxonomy as input and extracts the features (product categories) as output. This approach is used **to adjust and balance the original product taxonomy**, and it considers both the frequency of product purchases and the product semantics to tackle with data skewness problems.

We also suggest that the units of analysis used in the literature, i.e. product items in a single visit, or all shopper visits, are not applicable in every retail context but there are cases where we should examine groups of "x" sequential visits. Thus, we also suggest **the creation of an intermediate unit of analysis (the fifth outcome)**, which is required in some retail domains where shoppers perform many store visits during small time windows e.g. in DIY retailing.

Apart from the aforementioned outcomes, an equally significant outcome, is related to the **data-driven innovations (the sixth outcome) in shopper marketing** that the knowledge derived from the proposed approach may support. Indicatively, the proposed approach extracts knowledge that may support several decisions ranging from marketing campaigns per shopping mission, redesign of a store's layout to product recommendations.

To prove the effectiveness and the validity of the identified shopping missions, we conducted a field study using a smart mobile app. We **demonstrated that the resulting shopping missions effectively support innovative marketing actions (the seventh outcome).** We conclude that shopping mission-related coupons achieve higher redemption rates and are claimed by a shopper into less time than the non-related coupons. Likewise, we enhanced shopping mission's validity via conducting focus groups and discussing the identified missions with the shoppers.

Closing, another important outcome of this dissertation is **the identification of the factors** that **affect** the input and the results of shopper **segmentation systems and approaches (the eighth outcome).** Similarly, this research also pinpoints the **factors** that the **designers** of **visit segmentation systems** should **take into consideration** in contemporary retail (see section 7.4 below). Thus, it also sets the bases for generic IS tools for visit segmentation **(the ninth outcome)**.

## 7.2. Theoretical contribution

This thesis has an interdisciplinary nature, as it interweaves three different disciplines: Information Systems (IS), Business Analytics (BA), Shopper Marketing. Therefore, the contribution of this thesis from a theoretical perspective is found across these three disciplines.

The following paragraphs summarize and discuss the theoretical contribution of this research:

**Defines the concept of visit segmentation.**

Shopper segmentation is a traditional concept that is flourishing in contemporary retail due to the data explosion and the transformation of modern shoppers. From the one side, IoT technologies aid us capture more data regarding customer interactions during their shopping journey. From the other side, customers repetitively shift their behavior during time, even when they visit the same channel and the same store. In this research, we question whether shopper segmentation approaches can serve this new reality. Researchers (Walters and Jamil, 2003; Bell et al., 2011) suggest that we should pay attention to each shopper visit as it carries valuable insight on the shopper needs. Putting the shopper visit on the spot, instead of the shopper total buying behavior that shopper segmentation relies on, has the potential to ensure a more accurate view of the shopper desires that change frequently due to an abundance of new products, shopping channels and services offered every day.

Sharing other researchers (Walters and Jamil, 2003; Bell et al., 2011) concerns, we propose that in this new era we should put the shopper visit on the spot, instead of the shopper behavior that changes over time and that traditional shopper segmentation relies on. Thus, we coin the term "visit segmentation", to pinpoint this need. Visit segmentation focuses on the underlying needs that boosted a customer visit a store e.g. to purchase products for a light meal, or to procure materials to renovate their bathroom etc. These needs and missions can be extracted using various datasets reflecting customers' behavior e.g. product purchases, interactions, preferences etc.

**Proposes a behavioral segmentation and characterization of shopper visits that reveal shoppers' missions and intentions.**

Current research on shopper segmentation utilizes all shopping visits to identify customer groups. These studies examine shoppers' behavior via looking at the entirety of the products a shopper has purchased, regardless of whether this took place in one or more visits and try to segment shoppers based on this behavior. In more detail, an extensive review of the relevant literature has revealed that researchers have analyzed sales data per customer (customer-level data) utilizing different methods e.g. clustering, Markov chains, etc. (e.g. Chen et al., 2009; Cheng & Chen, 2009; Han et al., 2014; Kitts et al., 2000; Liao et al., 2011). They examine the complete sales history per customer in terms of sales volumes, visit frequency, mix of products or product categories etc. and, thus, generate customer segments that provide a generic characterization of consumers in terms of the products they prefer along with other characteristics such as available budget, demography etc. Indicatively, in these works researchers identify shopper segments who purchase routine, seasonal or convenience categories (Han, Ye, Fu, & Chen, 2014), or segment shoppers based in high spenders and frequent buyers (Aeron et al., 2012). However, the aforementioned studies, overlook the shopping purpose of a customer visit which carries valuable insights about shopper motives and shopper missions.

Here, we should admit that there are researchers who focus on customers' visits and not on the entirety of shoppers' behavior (e.g. during a year), as the aforementioned stream of studies does. They focus on the association between the purchased products in basket/visit level (also known as market basket analysis) (Srikant and Agrawal, 1995; Boztuğ and Reutterer, 2008; Cil, 2012; Beck and Rygl, 2015) e.g. those how bought diapers also bought beer. However, market basket analysis still overlooks the shopping purpose of each shopper visit.

In turn, we analyze visit/basket-level data and extract neither shopper segments, nor pairs of product categories customers prefer to purchase together. We focus on each shopping visit separately and we assign visits to groups, which are characterized by the product categories they contain. The resulting mix of purchased product categories is not random but reflects the shopping purpose, the shopping mission and the shopping intention that motivated each visit. Different mixes of product categories per visit segment reflect different shopping needs. Ultimately, we start with groups of shopping visits to infer distinct customer shopping missions e.g. light meal, breakfast, snacks and beverages, food and drink on-the-go.

**Develops a business analytics approach that performs visit segmentation.**

The most relevant study to visit segmentation comes from the marketing domain and is those of Bell et al, (2011), which identifies the need to segment each visit and identified different shopping trip types. However, in the rest segmentation literature, there are is a lack of business analytics driven approaches that perform visit segmentation.

Editorials (Agarwal and Dhar, 2014; Goes, 2014), other academic papers (Abbasi et al., 2016; Delen and Zolbanin, 2018) and practitioners (McKinsey Global Institute, 2011) emphasize the need to develop data-driven approaches, systems and frameworks to better understand and form the insight generation processes (Pick et al., 2017). However, there is limited literature that proposes innovative business analytics frameworks, approaches, systems and methods that pinpoint steps on how to delve deeper in the data to better understand the behavior of modern shoppers. To the best of our knowledge this is the first attempt to develop a business analytics approach that performs visit segmentation.

Towards that end, we develop such an approach that utilizes clustering techniques to identify segments of visits. The input of this approach are POS/transactional data in visit level from various retail stores. This approach extracts groups of visits based on the product categories the customers have purchased during each visit. We suggest that the resulting mix of product categories in each visit segment reflects shoppers' missions e.g. that they enter the store for "snacks and beverages" shopping mission.

We applied, validated and refined the proposed approach through three heterogeneous retail cases. This way we demonstrated its generalizability. The first case concerns sales data from different channels and stores of a major European fast-moving consumer goods (FMCG) retailer. Respectively, in the second case, we produced the visit segments for the physical stores of a Fortune 500 specialty retailer of home improvement and construction products – also known as do-it-yourself (DIY) retailer. The third case concerns data from a physical and the web store of a major European fashion retailer. The application of the proposed approach in the different retail contexts, also contributes to the generalization of the results.

**Proposes a semi-supervised feature selection method to extract custom product categories.**

We suggest that identifying the right product category level, i.e. the right level of analysis in the product taxonomy tree, is crucial to the results of the study, it may affect the knowledge discovery process and the data mining results (Cho et al., 2002). Thus, we chose to work on visit data at a customized product category level and not at item/SKU level. Researchers that have utilized a retailer's existing product taxonomy have often claimed very poor results in both the algorithms' accuracy and the business evaluation (Cho and Kim, 2004; Videla-Cavieres and Ríos, 2014). With a typical retail store, having more than 10.000 SKU's in its assortment, it is rather impossible to identify significant patterns at an SKU level and working at a higher level of analysis is required to avoid data sparsity problems. Besides, the main store retail activities (e.g. store replenishment, shelf space allocation, product assortment selection) and the relevant decisions mainly refer to product categories, as the shopper needs are often expressed at the category level (e.g. 'I need to buy milk') rather than at a specific SKU level (e.g. 'I need to buy this specific milk in a 250ml bottle'). In addition, by working at the product category level we ensure that the results are more generic and may also apply to new products of a category.

In our proposed approach, the adjustment and the balancing of retailer's product taxonomy contributes significantly to shaping the feature space of the problem; as it determines the main input (i.e. fact table) of the clustering model. Thus, it is an important research decision to select the level of analysis at the product taxonomy that is efficient to obtain meaningful clustering results (Albadvi & Shahbazi, 2009; Cho & Kim, 2004; Cho et al., 2002; Han et al., 2014; Hung, 2005; Kim, et al., 2002; Srikant & Agrawal, 1995). For example, whether we should refer to a can of sparkling orange juice of brand XYZ as sparkling beverage, as beverage, or as orange juice etc. Utilizing existing techniques regarding feature selection (or more precise dimension reduction) for unsupervised learning (e.g. Principal Component Analysis - PCA) it was rather straightforward. Although this technique is highly adaptive, it appeared having two major drawbacks in our case. Firstly, the produced principal components (product categories in our case) were not comprehensible by the experts and secondly -and most important from a technical perspective- the proposed components had a poor performance in terms of variance explanation, indicating that either the skewed data (existence of latent variables) or the binary values of product categories feature space (distance metric) are performing poorly. To this end, these findings confirm the role of dimensionality reduction in clustering, as well as the capabilities of PCA as discussed in the existing literature (e.g Lawrence, 2005).

Hence, too avoid poor, not representative customer segmentation results, we propose formulating a customized product category level, via balancing a retailer's product taxonomy. We decided to adjust the product taxonomy adopting the suggestions provided by Dy & Brodley (2004) and Guyon & Elisseeff (2003) regarding feature engineering and the role of the domain problem on the features. In the current study we designed, developed and employed a semi-supervised feature selection approach that uses the product taxonomy as input and suggests the features as output. Specifically, it parses bottom-up the product categories tree and acts in a two-fold manner: (i) it merges nodes (product categories) with low percentage in total baskets and (ii) splits a node with high percentage in total baskets. The proposed approach has not yet been thoroughly optimized and compared to similar methods, yet it supports the extraction of high-quality clusters regarding visit segmentation and, thus, we consider it as a first step towards contributing to the existing feature engineering literature. Moreover, we consider the major advantage of our suggested approach to be that we preserve the semantic information of features (product categories) because we deploy it according to expert's intervention.

Our method differs from those utilized in the semantic web, which take into consideration only product semantics. In addition, our method differs from Cho & Kim (2004) that formulates categories based solely on product purchases without taking into account product semantics, leading to merging unrelated products such as skincare and socks. Last but not least, we also differ from Srikant & Agrawal (1995), as in their approach they produce association rules for any product taxonomy level, and prune the redundant ones; however, they do not formulate customized product categories, also this approach is dependent on the data mining technique e.g. association rules, and hasn't been tested in other techniques such as clustering.

**Identifies factors affecting segmentation approaches and systems, and sets the basis for the development of IS tools for visit segmentation.**

This research highlights retailer and shopper characteristics and factors affecting traditional shopper segmentation systems and approaches. As well, it discusses how these factors affect the input, the processing approach and the results of such segmentation systems. In our interdisciplinary study we identify all these factors that marketing literature admits their existence and we examine how they affected segmentation approaches in the Information Systems (IS) field. In more detail we identified factors related to marketing 4Ps (place, product, price, promotion), loyalty programs, data 4Vs (variety, volume, velocity, veracity) and shopper 4Vs (volume, variety, value, visit) - we inspired this term from the data 4Vs. We propose that the consumer segmentation analysis and results should be conducted and translated respectively considering the “marketing” characteristics of the shoppers and the retailers.

Studying segmentation literature, we identified that there are works mainly in the marketing domain (e.g. Bradlow et al., 2017), that discuss several factors that affect big data analytics systems in general. However, they do not present evidence of how these factors affected relevant segmentation cases. Also, in the IS literature there is a great majority of papers (e.g. Boone and Roehm, 2002; Boztuğ and Reutterer, 2008; Aeron et al., 2012; Miguéis et al., 2012) that perform shopper segmentation. Though to the best of our knowledge, authors describe their own case and not “the bigger” picture i.e. how system inputs and factors (e.g. data) affect and alter the segmentation process, system and results/outputs. As, it is only implied, and it is not discussed how different factors affected segmentation results. Thus, to the best of our knowledge this is the first effort to sketch thoroughly the segmentation era. In addition, this research highlights how these factors affect not only shopper segmentation, but also new visit segmentation approaches and systems.

**Proposes moving from category to shopping mission management and opens a new chapter in the category management (CM) literature.**

In contemporary retail, we observe a changing behavior of shoppers i.e. a shopper might swift his/her behavior even when visiting the same retail store. We suggest that CM practices seem to be cumbersome and need to be revamped in order to cope with the changing behavior of shoppers. In existing literature, there are a few researchers (e.g. Song and Chintagunta, 2006; Kamakura and Kang, 2007, Han et al., 2014; Nielsen et al., 2015) that highlight the need to manage categories based on shoppers and their needs (consumer-centric CM). However, in existing category management literature there are no such practices. Even consumer-centric CM is focusing merely on cross-category relations and not on shopper needs. We propose not only to focus separately on each category e.g. milk, cereals, coffee etc. as traditional CM does, but to move from CM to Shopping Mission Management. This way we will treat categories collaboratively under the shopping mission they participate. Hence, a new chapter in category management is unfolding and more effective assortment, store layout planning, space allocation and promotional activities could be achieved.

**Proposes an intermediate unit of analysis.**

The application of the approach in these cases also revealed that the units of analysis used in the literature, i.e. product items in a single visit, or all shopper visits, are not sufficient and applicable in every retail context, but there are cases where we should examine groups of “x” sequential visits. The value of “x” differs according to the domain the data derived from. As we proved and as other researches support (Wolf and McQuitty, 2011) a shopper usually visits a retail store that sells products for home improvement many times and purchases few materials each time. We devise and test a new unit of analysis where we examine groups of “x” continuous visits. This intermediate unit of analysis is dictated by the particularity of some retail domains that demand many store visits during small time windows. The value of some of the factors identified in the literature review, can aid the researcher determine the unit of analysis.

## 7.3. Practical implications

In a nutshell as we described in detail in chapter 6, the practical value of this work is stressed when considering the shopper-oriented business decisions it can support. More specifically, our approach can be evolved to a tool for designing innovative marketing campaigns and bundled promotions for product categories that belong to the same shopping visit segment. For example, retailers may plan cross-coupon programs for addressing the needs of customers visiting the store with a specific purpose in mind. Alternatively, our approach may become the cornerstone of a recommendation system for real-time purchases in retail stores. It will propose to the customers more products that they may have forgotten to buy, considering their prior or current visit(s).

In the same spirit, we can create offline and online product catalogues. For instance, we have detected women that enter a fashion store to purchase professional clothes and baby clothes. Thus, to promote the new collection, it could be more effective to send them product catalogues that meet their specific preferences, instead of including all the available new garments. The extracted knowledge could also be valuable for advertising purposes; for instance, instead of making advertisements of specific product categories, retailers could advertise bundled product categories that correspond to a shopping mission, e.g. breakfast products advertisements.

On the other hand, the customer visit segments can dictate a new redesigned retail store physical or web store layout. For example, the product categories in the same visit segment could be positioned in nearby store aisles and shelves. Considering the bigger picture, we can move from a category-based layout to a mission-based layout that can help customers locate products in the store more easily and buy more in less time (Cil, 2012; Sarantopoulos et al., 2016, 2019). Alternatively, second in-store placement spots can be detected.

Moreover, our approach can be even evolved into a collaborative analytics platform that gathers transactional data from various retailers and provide them insights regarding the visit segments derived from their data and enable performance benchmarking to competitors.

Further, the value of such a system could be further enhanced when we use the resulting visit segments to perform shopper segmentation. By looking into the shopping missions that each shopper performs in all the stores of a retail chain, we can boost shopper marketing activities. Shopper segmentation based on the identified shopping missions can aid retailers identify selling gaps and opportunities and enhance personalization.

Additionally, the store manager could reengineer store operations management and replenishment strategies by ordering groups of products based on the identified visit segments. Additionally, this approach could be even utilized to rearrange and modify a retailer's warehouse, by placing in nearby aisles products matching online orders to decrease order-picking time.

Closing, our research outcomes may assist system engineers in designing segmentation systems and data scientists in modifying the data manipulation and modeling processes. Also, our study can aid marketers to understand and interpret the shopper segmentation results in a way that approaches better the shoppers' behavior and, thus, take more effective decisions concerning the customers' treatment and experience. This research aspires to bridge marketing researchers and managers with data scientists and shopper segmentation designers to obtain a more thorough understanding of shopper behavior.

## 7.4. Thoughts for visit segmentation systems designers in contemporary retail: Factors to take into consideration

Here, we discuss how the various factors identified in the literature review (see section 2.1.4), affected, or not the segmentation execution and results in the three presented retail cases. In our case studies, these factors are twofold, as on the one hand they shape the initial data set, and on the other hand they have a mediating role in explaining the results. Table 7-2 summarizes the results. This section contributes both in literature and practice as it sets the bases for generic IS tools for visit segmentation.

### 7.4.1. Shopper 4Vs

- *Visits*: This factor enriched our segmentation results, but it wasn't available in all cases, as in the physical stores of the grocery retailer we didn't had a loyalty card to associate each visit with a customer and identify her/his visit history. Though, it didn't affect the analysis in this case. However, as described below (in the variety feature) it affected the unit of analysis. Also, the time between a shopper's visits was very crucial to identify and calculate the intermediate unit of analysis i.e. super-visits in the DIY case.

- *Variety*: played an important role in our analysis. Firstly, in all the cases it was used to identify the outliers in our dataset. Also, variety enriched our results and aid us with their interpretation. As well, as discussed below in the product feature it also affected the product granularity level, selected in the modeling phase. This way we confirmed the problems existing literature admits (Srikant and Agrawal, 1995; Cho and Kim, 2004; Videla-Cavieres and Ríos, 2014) and we managed to outstripped poor results and data sparsity issues. Lastly, it was used to identify in which unit we will perform the visit segmentation (visit, all shopper's visits, group of continuous visits etc.). In the FMCG case where a customer purchases many product categories, i.e. the variety is high, the unit of analysis is this single visit. Thus, we can perform the segmentation and identify the shopping mission and intention of customers in each visit. In the fashion retail case as visit variety was low, we also examined all customers' visits. The low number of visits leads us to analyze customers' all visiting history in the store. Thus, examining their entire purchase history we moved to traditional segmentation approaches, extracting shopper segments. Compared to the two other cases, in DIY we had a too high number of visits combined with a low variety of products (per visit). Thus, neither the single visit, nor all shopper's visits were selected as the unit of analysis. On the contrary, a customized/intermediate unit of analysis was formed (super-visits). Hence, this unit is formed via the merging operation of "x" contiguous/sequential visits, made by the same shopper. This happened because a shopper has in mind a specific construction project(s), e.g. to renew his garden, and thus s/he executed "x" contiguous in time visits, until accomplishes this construction project. Testing this intermediate unit, we expand existing literature and prove that the existing unit of analysis i.e. product items in a single visit, or all shopper visits, are not sufficient and applicable in every retail context. Below (Table 7-1) we summarize our results:

| Case | Variety | Visits | Unit of analysis |
|---|---|---|---|
| FMCG | High | Indifferent | visit |
| Fashion | Low | Medium | all shopper's visits |
| DIY | Low | High | "x" contiguous in time visits |

**Table 7-1. Relation between variety, visits and unit of analysis**

- *Value*: In the fashion and DIY case this feature wasn't available. As proved in the FMCG case the value of each basket indeed enriched the resulting segments. For example, we identified that "spirits and beverages" segment had the lowest basket value and it is only 50 cents higher than the allowed order level of the web store. Also, value had an impact at the outliers' extraction in the cluster sampling phase.

- *Volume*: Was used in all cases as an additional descriptive statistic to enrich the results. Also, in combination with the variety it highlighted us differences in shopping behavior. For instance, in the fashion case we identified that men purchase more products in a single visit than women. Also, shoppers tend to have higher basket volume in retailer's web store than in the physical. Lastly, along with variety feature aid us with the product granularity selection.

### 7.4.2. Marketing 4Ps

- *Product:* Our segmentation approach was based on the product characteristics. We analyzed retail shopper data and produce groups of visits based on the product categories the customers have purchased during each visit to a physical retail or web store. The resulting mix of product categories in each visit segment reflects the shopping intentions of the respective customers that visited the stores. Confirming and expanding existing research, in all our cases we proved that the product and the product granularity level should be correctly defined because it affects the segmentation results. This happens because product taxonomies are often unbalanced and have characteristics hindering the performance of data mining algorithms. For that reason, we devised customized product categories based on a semi-supervised feature selection method using both product taxonomy and variety feature, and we reconsider these categories after receiving the first clustering results. Also, the product itself e.g. consumables products in the FMCG case, or durable products in the DIY case affected the business interpretation layer.

- *Place:* In the FMCG case we received data from different channels and store types. Confirming and expanding existing literature, some dissimilar segments with different behaviors derived across the different channels and stores. Also, in the fashion case WE had data from both web and physical store; similarly, in this case different shopper behaviors and segments identified across different channels.

- *Promotion:* In the data provided by the fashion retailer we were able to track whether each transaction was a result of a promotion. This enriched our results, as we could detect those segments that we or not prone to marketing actions. These results are valuable as they can help marketeers decide the segments targeted for specific actions. In the other cases this feature wasn't available.

- *Price:* We should mention that price feature didn't affect our segmentation approach. First, it wasn't available in all our cases, secondly even in the FMCG case, that was available it didn't influence the core phases of our segmentation. Hence, we partially confirm existing literature that admits that price feature plays an important role in more particular products e.g. cars. Lastly, it seems that this factor could somehow enrich the results interpretation layer.

Closing, we propose that the consumer segmentation analysis and results should be conducted and translated respectively considering the “marketing” characteristics of the shoppers and the retailers.

### 7.4.3. Loyalty programs

Fashion retailer run loyalty programs in all his channels. From the one hand these programs seem to be effective in the physical store as it has high card usage penetration, on the other in the web store only 60% of their shoppers use their card. As the unit to be analyzed was all shopper visit, we had to exclude those the visits weren’t associated with a shopper. Thus, we eliminated a vast amount of the web transactions. DIY retailer every visit was associated with a shopper, as loyalty card was prerequisite of each purchase. We should mention that in this case, this feature was vital for our analysis, because without having such data to identify the shopper performed each visit, we could not perform the segmentation. In the FMCG case retailer maintained a loyalty program in the web store also there was a high card adoption. Though, this was indifferent for our analysis, as in this case the unit to be analyzed was visits. But, we could use these additional data source to further explore the dataset e.g. identify patterns of shopper’s behavior regarding the identified shopping visit segments.

| Factors | FMCG | Fashion | DIY | How the proposed approach was affected |
|---|---|---|---|---|
| Visits | ✓(Indifferent) | ✓ (Medium) | ✓ (High) | Data preparation (outliers’ extraction, unit of analysis), data modeling, results |

| | | | | |
|---|---|---|---|---|
| | | | | interpretation, results enrichment |
| Variety | ✓ (High) | ✓ (Low) | ✓ (Low) | Data preparation (outliers' extraction, unit of analysis), data modeling (skewness problems, product granularity selection) results interpretation, results enrichment |
| Value | ✓ | N/A | N/A | Data preparation (outliers' extraction), results interpretation, results enrichment, |
| Volume | ✓ | ✓ | ✓ | Results interpretation, results enrichment, data modeling (product granularity selection), high correlation with variety factor |
| Product | ✓ | ✓ | ✓ | Results interpretation, data modeling (product granularity selection) |
| Place | Web and physical channel, various store types | Web and physical | Physical stores | Results interpretation |
| Promotion | N/A | Available | N/A | Results interpretation, results enrichment, |
| Price | ✓ | N/A | N/A | Results interpretation |
| Loyalty Program | ✓ (All channels) | ✓ (Web store) | ✓ (Physical stores) | Data preparation, in the cases where the unit of analysis is not the single visit, without having this factor we are not able to perform the segmentation |

**Table 7-2. Factors VS approach phases affected**

### 7.4.4. Data 4Vs

Apart from the aforementioned factors we admit that a feature that also affected our segmentation is the variety of the given data. This feature enriched our results; also, the different data sources affected the data preparation phase. In all cases we received data regarding POS and product taxonomy. For the web store of the FMCG retailer we also received loyalty cards. Regarding the fashion retailer e also received, loyalty cards data, demographics, other product characteristics and promotion data. These additional datasets indeed enriched our results and aid us with our results interpretation. Lastly, regarding the DIY retailer apart from the POS data, we solely received loyalty cards data. In this case more data sources could scientifically have helped us to interpret the resulting segments. Hence, we confirm existing literature which admits that the utilization of different data sources aid businesses obtain a multifaceted view about their customers and as a result data variety seems to affect significantly every analytics process (Goes, 2014).

Apart from the data variety feature we claim that the rest data 4Vs (volume, velocity, veracity) may affect future segmentation approaches. Retail data volume and velocity may cause technical issues and require sophisticated data infrastructures to manage them (Goes, 2014). Thus, they will affect the first layer (preparation) of our proposed approach. Additionally, dynamic segmentation models will be required to analyze the data on the fly and support real time segmentations issues (Reutterer et al., 2006). Hence, the data modeling layer will be affected. However, in our cases we didn’t receive massive data volumes, thus we analyzed the dataset using common tools e.g. R studio and SQL Server. Also, we performed ad-hoc analysis based on historical data, thus data velocity didn't affect us. Regarding data veracity, although we admit that in the retail context, different data issues may arise; we didn't identify significant imprecisions to our datasets. However, in the literature researchers admit that data inconsistencies affected their analysis. As, for example, shoppers often declare wrong information, for instance, in attributes such as age, income etc. Additionally, these factors are "slowly changing"; thus, a segmentation that will be conducted in the future, it might incorporate false data e.g. people are getting married, salaries grow etc. (Kohavwe et al., 2004).

At, Figure 7-1 we provide a graphical representation of the factors (Shopper 4Vs, Marketing 4Ps, Loyalty programs, Data 4Vs), affecting each phase of our proposed visit segmentation approach. As shown the interpretation layer is the one that is affected by most of the factors. At this phase to extract wisdom from the results, we need experts' opinion that know the market. Experts not only consider the tangible, quantitative factors (e.g. value, volume etc.) to identify shoppers' missions and motives, but also intangible elements such as their domain knowledge and accumulate experience. Likewise, we could claim that variety is the most important feature that affects not only all the phases, but also almost every sub-phase/step of our approach; from the outliers' elimination, and the product taxonomy calibration, to the identification of the unit of analysis and the interpretation of the results and their translation into insights.

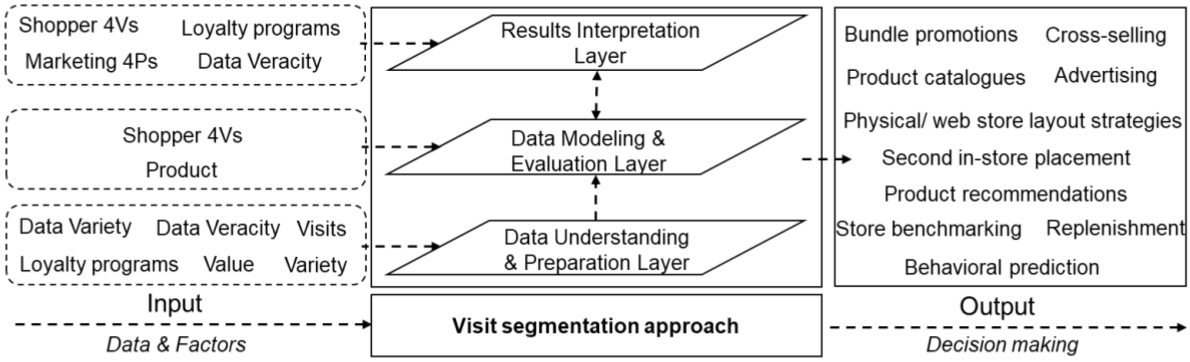



**Figure 7-1. Visit segmentation system**

## 7.5. Limitations and future research

Further research may address some limitations of this study. We can use more complex shopper interaction data derived from alternative technological means (e.g. RFID, BLE beacons) from other retail contexts to evaluate and validate the proposed approach. For instance, data that indicate the products a customer puts in his RFID-enabled shopping cart during a shopping visit in a grocery store, or the aisles a shopper visits during his/her in-store journey. It would also be a challenge to use different interaction data of the same retailer and compare the results. For instance, we can examine the visit segments derived via combining different interaction data of the same retailer. In addition, we can identify the selling gaps via comparing the visit segments stemming from data of products in the shopping carts and products that are finally purchased. Further research may also study other shopping occasions and visit segments e.g. cases where the purpose of the visit is to return items, or buying as a gif etc. Additionally, future research may examine datasets from other contexts e.g. third-party logistics (3PL) companies to identify the ordering purpose of consumers.

Furthermore, future research, may address other interesting questions that arise in Section 1.2 regarding visit segmentation e.g. “Do traditional shopper segmentation systems serve contemporary retail?”, or “Are the visit segmentation-informed marketing actions more effective than traditional actions?". Regarding the latter, shopping mission-based promotions, could be compared with traditional product-centric promotions e.g. "buy one, get one free" (BOGO) from the same category. Also, shopping mission-informed product catalogues, can be compared with old-school product catalogue display. In general, the value and effectiveness of shopping mission management versus the category management can be measured. Similarly, further research may examine the impact of collaborative shopping mission-driven analytics practices between retailers and suppliers.

From another perspective, it could be also interesting not only to perform shopper segmentation based on the visit segments, but also to leverage this knowledge and design marketing actions. For instance, different strategies could be tested to move the medium-loyal shoppers of Figure 6-5 into a better segment e.g. by recommending shopping missions they tend to purchase at retailer's competitor.

Additionally, in the current thesis the direction and the magnitude of the factors affecting visit segmentation systems is not examined. However, this is a great era for future research. Closing, a limitation of this study that future research may address is related to the fact that other tangible and in-tangible factors may affect the resulting visit segments. For instance, store characteristics such as the geographical area the store belongs to, the existence or not of a parking space (which is common in the FMCG domain) etc. As well, shopper characteristics e.g. age, weight, strength etc., or product characteristics e.g. weight, volume etc. may limit the resulting visit segments. For example, in the DIY case the weight and the volume of the products in combination with the shopper's characteristics and the capacity of his/her car may alter the resulting visit segments.

From a technical perspective, we can apply more data mining techniques, such as association rules, and compare the resulting visit segments with those that have been derived from clustering. Or even other techniques and algorithms could also be examined to cope with the difficulty of identifying core visit segments at the hyper-stores. A limitation of the proposed approach is that it works when the factor "visit variety" is sufficient to identify interesting patterns and visit segments; then, the rest of the visits are considered as outliers. Hence, alternative techniques such as graph mining could also be utilized to further analyze each resulting segment. This way we can cope with the difficulty to identify more detailed segments in the DIY case study.

Closing, future research may apply the semi-supervised feature selection method in other taxonomies and hierarchies, e.g. to regions, sub-regions etc. Also, further research may improve the semi-supervised feature selection method. For example, via automating the identification of product semantics which is currently supervised by experts. Additionally, future research may extent the developed feature selection method for self-balancing binary search trees such as B-trees and AVL trees.

# REFERENCES

Aanen, S. S., D. Vandic and F. Frasincar. (2015). "Automated product taxonomy mapping in an e-commerce environment." *Expert Systems with Applications*, *42*(3), 1298–1313.

Aastrup, J., D. B. Grant and M. Bjerre. (2007). "Value Creation and Category Management through Retailer–Supplier Relationships." *The International Review of Retail, Distribution and Consumer Research*, *17*(5), 523–541.

Abbasi, A., S. Sarker and R. H. L. Chiang. (2016). "Big data research in information systems: Toward an inclusive research agenda." *Journal of the Association of Information Systems*, *17*(2), 1–32.

Aeron, H., A. Kumar and J. Moorthy. (2012). "Data mining framework for customer lifetime value-based segmentation." *Journal of Database Marketing and Customer Strategy Management*, *19*(1), 17–30.

Agarwal, R. and V. Dhar. (2014). "Editorial —Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research." *Information Systems Research*, *25*(3), 443–448.

Aggarwal, C. C. (2016). *Outlier Analysis Second Edition*. Springer, Cham.

Agrawal, R., T. Imieliński and A. Swami. (1993). "Mining association rules between sets of items in large databases." In: *ACM SIGMOD international conference on Management of data* (pp. 207–216).

Aguinis, H., L. E. Forcum and H. Joo. (2013). "Using Market Basket Analysis in

Management Research." *Journal of Management*, *39*(7), 1799–1824.

Albadvi, A. and M. Shahbazi. (2009). "A hybrid recommendation technique based on product category attributes." *Expert Systems with Applications*, *36*(9), 11480–11488.

American Marketing Association and Nielsen Marketing Research (Firme). (1992). *Nielsen category management: Positioning your organization to win.* Lincolnwood, Ill. : NTC Business Books.

Anderson, J. L., L. D. Jolly and A. E. Fairhurst. (2007). "Customer relationship management in retailing: A content analysis of retail trade journals." *Journal of Retailing and Consumer Services*, *14*(6), 394–399.

Baltas, G., P. C. Argouslidis and D. Skarmeas. (2010). "The Role of Customer Factors in Multiple Store Patronage: A Cost-Benefit Approach." *Journal of Retailing*, *86*(1), 37–50.

Bartunek, J. M., S. L. Rynes and R. D. Ireland. (2006). "What Makes Management Research Interesting, and Why Does It Matter?" *The Academy of Management Journal*, *49*(1), 9–15.

Baskerville, R. (2008). "What design science is not." *European Journal of Information Systems*, *17*(5), 441–443.

Bean, R. and T. H. Davenport. (2019). "Companies Are Failing in Their Efforts to Become Data-Driven" *Harvard Business Review*.

Beck, N. and D. Rygl. (2015). "Categorization of multiple channel retailing in Multi-, Cross-, and Omni-Channel Retailing for retailers and retailing." *Journal of*

*Retailing and Consumer Services*, *27*, 170–178.

Belk, R. W. (2007). *Handbook Of Qualitative Research Methods In Marketing*. (C. Russell W. Belk, Kraft Foods Canada Chair of Marketing, Schulich School of Business, York University, Toronto, Ed.). Edward Elgar Pub.

Bell, D. R., D. Corsten and G. Knox. (2011). "From Point of Purchase to Path to Purchase: How Preshopping Factors Drive Unplanned Buying." *Journal of Marketing*, *75*(1), 31–45.

Bell, D. R. and J. M. Lattin. (1998). "Shopping Behavior and Consumer Preference for Store Price Format: Why "Large Basket" Shoppers Prefer EDLP." *Marketing Science*, *17*(1), 66–88.

Bi, Z., C. Faloutsos and F. Korn. (2001). "The "DGX" distribution for mining massive, skewed data." *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '01*, 17–26.

Boone, D. S. and M. Roehm. (2002). "Retail segmentation using artificial neural networks." *International Journal of Research in Marketing*, *19*(3), 287–301.

Boztuğ, Y. and T. Reutterer. (2008). "A combined approach for segment-specific market basket analysis." *European Journal of Operational Research*, *187*(1), 294–312.

Bradlow, E. T., M. Gangwar, P. Kopalle and S. Voleti. (2017). "The Role of Big Data and Predictive Analytics in Retailing." *Journal of Retailing*, *93*(1), 79–95.

Breugelmans, E., T. H. A. Bijmolt, J. Zhang, L. J. Basso, M. Dorotic, P. Kopalle, ... N. V. Wünderlich. (2015). "Advancing research on loyalty programs: a future research agenda." *Marketing Letters*, *26*(2), 127–139.

Calder, B. J. (2011). "Focus Groups and the Nature of Qualitative Marketing Research." *Journal of Marketing*, *14*(3), 353–364.

Chahal, M. (2015). "Consumers are "dirtying" databases with false details." Retrieved from https://www.marketingweek.com/2015/07/08/consumers-are-dirtying-databases-with-false-details

Chan, H. L. and K. W. Pang. (2011). "Association Rule Based Approach for Improving Operation Efficiency in a Randomized Warehouse." *Methodology*, 704–710.

Chang, R. M., R. J. Kauffman and Y. Kwon. (2014). "Understanding the paradigm shift to computational social science in the presence of big data." *Decision Support Systems*, *63*, 67–80.

Chen, M. C., A. L. Chiu and H. H. Chang. (2005). "Mining changes in customer behavior in retail marketing." *Expert Systems with Applications*, *28*(4), 773–781.

Chen, Y.-L., K. Tang, R.-J. Shen and Y.-H. Hu. (2005). "Market basket analysis in a multiple store environment." *Decision Support Systems*, *40*(2), 339–354.

Chen, Y. L., M. H. Kuo, S. Y. Wu and K. Tang. (2009). "Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data." *Electronic Commerce Research and Applications*, *8*(5), 241–251.

Cheng, C. H. and Y. S. Chen. (2009). "Classifying the segmentation of customer value via RFM model and RS theory." *Expert Systems with Applications*, *36*(3 PART 1), 4176–4184.

Chernev, A. (2003). "Product Assortment and Individual Decision Processes." *Journal of Personality and Social Psychology*, *85*(1), 151–162.

Cho, Y. H. and J. K. Kim. (2004). "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce." *Expert Systems with Applications*, *26*(2), 233–246.

Cho, Y. H., S. H. Kim and J. K. Kimb. (2002). "A personalized recommender system based on web usage mining and decision tree induction." *Expert Systems with Applications*, *23*, 329–342.

Chuang, Y. F., S. H. Chia and J. Y. Wong. (2014). "Enhancing order-picking efficiency through data mining and assignment approaches." *WSEAS Transactions on Business and Economics*, *11*(1), 52–64.

Cil, I. (2012). "Consumption universes based supermarket layout through association rule mining and multidimensional scaling." *Expert Systems with Applications*, *39*(10), 8611–8625.

Cui, G., M. Wong and H.-K. Lui. (2006). "Machine Learning for Direct Marketing Response Models: Bayesian Networks with Evolutionary Programming." *ManagementScience*, *52*(4), 597–612.

Delen, D. and H. M. Zolbanin. (2018). "The analytics paradigm in business research." *Journal of Business Research*, *90*(April), 186–195.

Demoulin, N. T. M. and P. Zidda. (2009). "Drivers of Customers' Adoption and Adoption Timing of a New Loyalty Card in the Grocery Retail Market." *Journal of Retailing*, *85*(3), 391–405.

Denning, P. J. (1997). “A new social contract for research.” *Communications of the ACM*, *40*(2), 132–134.

Desrochers, D. M. and P. Nelson. (2006). “Adding consumer behavior insights to category management: Improving item placement decisions.” *Journal of Retailing*, *82*(4), 357–365.

Dhar, S. K., S. J. Hoch and N. Kumar. (2001). “Effective category management depends on the role of the category.” *Journal of Retailing Journal of Retailing,* 77(2), 65–184

Dupre, K. and T. W. Gruen. (2004). “The use of category management practices to obtain a sustainable competitive advantage in the fast-moving-consumer-goods industry.” *Journal of Business and Industrial Marketing*, *19*(7), 444–459.

Dwyer, F. R. (1989). “Customer Lifetime Valuation Decision Making.” *Journal of Direct Marketing*, *11*(4), 6–13.

Dy, J. G. and C. E. Brodley. (2004). “Feature Selection for Unsupervised Learning .” *Journal of Machine Learning Research*, *5*, 845–889.

Eisenhardt, K. M. and M. E. Graebner. (2007). “Theory Building From Cases: Opportunities And Challenges.” *Academy of Management Journal*, *50*(1), 25–32.

ECR Europe (2011). *The Consumer & Shopper Journey Framework*. Retrieved from https://www.ecrireland.ie/uploadedfiles/shopper/projects-ecr-eu/the-consumer-and-shopper-journey-framework-.pdf

European Commission, D.-G. for R. and I. (2014). *Six perspectives on Retail Innovation, Expert Group on retail sector innovation*. *Business*.

Fendt, J. and W. Sachs. (2008). “Grounded theory method in management research: Users’ perspectives.” *Organizational Research Methods*, *11*(3), 430–455.

Gijsbrechts, E., K. Campo and P. Nisol. (2008). “Beyond promotion-based store switching: Antecedents and patterns of systematic multiple-store shopping.” *International Journal of Research in Marketing*, *25*(1), 5–21.

Goes, P. B. (2014). “Big Data and IS Research.” *MIS Quarterly*, *38*(3), iii–viii.

Griva, A., C. Bardaki, K. Pramatari and D. Papakiriakopoulos. (2018). “Retail business analytics: Customer visit segmentation using market basket data.” *Expert Systems with Applications*, *100*.

Griva, A., C. Bardaki, P. Sarantopoulos and D. Papakiriakopoulos. (2014). “A Data Mining-Based Framework To Identify Shopping Missions.” *MCIS 2014 Proceedings*.

Grossi, F., B. Harris, Joint Industry Project on Efficient Consumer Response and Partnering Group. (1995). *Category management report*. United States: Joint Industry Project on Efficient Consumer Response.

Gruen, W. T. and H. R. Shah. (2000). “Determinants and Outcomes of Plan Objectivity and Implementation in Category Management Relationships.” *Journal of Retailing*, *76*(4), 483–510.

Gupta, S., D. Hanssens, B. Hardie, W. Kahn, V. Kumar, N. Lin, ... S. Sriram. (2006). “Modeling customer lifetime value.” *Journal of Service Research*, *9*(2), 139–155.

Guyon, I. and A. Elisseeff. (2003). “An Introduction to Variable and Feature Selection.” *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182.

Han, S., Y. Ye, X. Fu and Z. Chen. (2014). "Category role aided market segmentation approach to convenience store chain category management." *Decision Support Systems*, *57*(1), 296–308.

Hesková, M. (2006). *Category management*. *Profess Consulting*.

Hevner, A. R. and S. Chatterjee. (2016). "Design research in information systems: Theory and practice." *Reference Syllabi, Association for Information Systems (AIS)*, 337–342.

Hevner, A. R., S. T. March, J. Park and S. Ram. (2004). "Design science in information systems research." *MIS Quartely*, *28*(1), 75–105.

Homburg, C., V. V Steiner and D. Totzek. (2009). "Managing Dynamics in a Customer Portfolio." *Journal of Marketing*, *73*(5), 70–89.

Hong, T. and E. Kim. (2012). "Segmenting customers in online stores based on factors that affect the customer's intention to purchase." *Expert Systems with Applications*, *39*(2), 2127–2131.

Hübner, A. H. and H. Kuhn. (2012). "Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management." *Omega*, *40*(2), 199–209.

Huerta-Muñoz, D. L., R. Z. Ríos-Mercado and R. Ruiz. (2017). "An Iterated Greedy Heuristic for a Market Segmentation Problem with Multiple Attributes." *European Journal of Operational Research*, (September 2014).

Hung, L. P. (2005). "A personalized recommendation system based on product taxonomy for one-to-one marketing online." *Expert Systems with Applications*, *29*(2), 383–392.

Jagadish, H. V., J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan and C. Shahabi. (2014). "Big data and its technical challenges." *Communications of the ACM*, *57*(7), 86–94.

Kahn, B. E. and D. C. Schmittlein. (1992). "The relationship between purchases made on promotion and shopping trip behavior." *Journal of Retailing*, *68*(3), 294–315.

Kamakura, W. A. and W. Kang. (2007). "Chain-wide and store-level analysis for cross-category management." *Journal of Retailing*, *83*(2), 159–170.

Ketchen, D. J. and L. Shook. (1996). "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique." *Strategic Management Journal*, *17*(6), 441–458.

Khajvand, M., K. Zolfaghar, S. Ashoori and S. Alizadeh. (2011). "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study." *Procedia Computer Science*, *3*, 57–63.

Kim, H. K., J. K. Kim and Q. Y. Chen. (2012). "A product network analysis for extending the market basket analysis." *Expert Systems with Applications*, *39*(8), 7403–7410.

Kim, J. K., Y. H. Cho, W. J. Kim, J. R. Kim and J. H. Suh. (2002). "A personalized recommendation procedure for Internet shopping support." *Electronic Commerce Research and Applications*, *1*(3–4), 301–313.

Kimball, R. and M. Ross. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.). Wiley Publishing.

Kitts, B., D. Freed and M. Vrieze. (2000). "Cross-Sell: A Fast Promotion-Tunable Customer-Item Recommendation Method Based on Conditionally Independent Probabilities." *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '00*, 437–446.

Klein, A. and G. Schmitz. (2016). "Main-format dominance in consumers' FMCG cross-format shopping patterns." *Journal of Retailing and Consumer Services*, *30*, 105–115.

Kohavi, R., L. Mason, R. Parekh and Z. Zheng. (2004). "Lessons and challenges from mining retail e-commerce data." *Machine Learning*, *57*(1-2 SPEC. ISS.), 83–113.

Konuş, U., P. C. Verhoef and S. A. Neslin. (2008). "Multichannel Shopper Segments and Their Covariates." *Journal of Retailing*, *84*(4), 398–413.

Kowalczyk, M. and P. Buxmann. (2015). "An ambidextrous perspective on business intelligence and analytics support in decision processes: Insights from a multiple case study." *Decision Support Systems*, *80*, 1–13.

Lawrence, N. (2005). "Probabilistic non-linear Principal Component Analysis with Gaussian Process Latent Variable Models." *Journal OfMachine Learning Research*, *6*, 1783–1816.

Lenzerini, M. (2002). "Data Integration: A Theoretical Perspective." *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 233–246.

Liao, S. H. and Y. J. Chen. (2004). "Mining customer knowledge for electronic catalog marketing." *Expert Systems with Applications*, *27*(4), 521–532.

Liao, S. H., Y. J. Chen and H. H. Hsieh. (2011). "Mining customer knowledge for direct selling and marketing." *Expert Systems with Applications*, *38*(5), 6059–6069.

Likas, A., N. Vlassis and J. J. Verbeek. (2003). "The global k-means clustering algorithm." *Pattern Recognition*, *36*(2), 451–461.

Lindblom, A. and R. Olkkonen. (2008). "An analysis of suppliers' roles in category management collaboration." *Journal of Retailing and Consumer Services*, *15*(1), 1–8.

Linoff, G. and M. Berry. (2011). "Wiley: Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd Edition -." Retrieved from http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470650931.html

Liu, J., X. Liao, W. Huang and X. Liao. (2018). "Market segmentation: A multiple criteria approach combining preference analysis and segmentation decision." *Omega (United Kingdom)*.

Liu, Yanchi, Z. Li, H. Xiong, X. Gao, J. Wu and S. Wu. (2013). "Clustering Validation Measures," *43*(3), 982–994.

Liu, Yuping. (2007). "The Long-Term Impact of Loyalty Programs on Consumer Purchase Behavior and Loyalty." *Journal of Marketing*, *71*(4), 19–35.

Lockshin, L. S., A. L. Spawton and G. Macintosh. (1997). "Using product, brand and purchasing involvement for retail segmentation." *Journal of Retailing and Consumer Services*, *4*(3), 171–183.

Lotfi, M. M. and S. A. Torabi. (2011). "A fuzzy goal programming approach for mid-term assortment planning in supermarkets." *European Journal of Operational*

*Research*, *213*(2), 430–441.

Markus, M. L., A. Majchrzak and L. Gasser. (2002). “A Design Theory for Systems That Support Emergent Knowledge Processes.” *MIS Quartely*, *26*(3), 179–212.

McAfee, A. and E. Brynjolfsson. (2012). “Big Data: The Management Revolution.”, *Harvard Business Review*.

McKinsey Global Institute. (2011). “Big data : The next frontier for innovation, competition, and productivity”. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx

Mesforoush, A. and M. J. Tarokh. (2013). “Customer profitability segmentation for SMEs case study: Network equipment company.” *International Journal of Research in Industrial Engineering*, *2*(1), 30–44.

Miguéis, V. L., A. S. Camanho and J. Falcão e Cunha. (2012). “Customer data mining for lifestyle segmentation.” *Expert Systems with Applications*, *39*(10), 9359–9366.

Milligan, G. W. and M. C. Cooper. (1985). “An examination of procedures for determining the number of clusters in a data set.” *Psychometrika*, *50*(2), 159–179.

Müller, O., I. Junglas, J. Vom Brocke and S. Debortoli. (2016). “Utilizing big data analytics for information systems research: Challenges, promises and guidelines.” *European Journal of Information Systems*, *25*(4), 289–302.

Nakano, S. and F. N. Kondo. (2018). “Customer segmentation with purchase channels and media touchpoints using single source panel data." *Journal of Retailing and Consumer Services*, *41*(November 2017), 142–152.

Neslin, S. A., D. Grewal, R. Leghorn, V. Shankar, M. L. Teerling, J. S. Thomas and P. C. Verhoef. (2006). "Challenges and opportunities in multichannel customer management." *Journal of Service Research*, *9*(2), 95–112.

Ngai, E. W. T., L. Xiu and D. C. K. Chau. (2009). "Application of data mining techniques in customer relationship management: A literature review and classification." *Expert Systems with Applications*, *36*(2), 2592–2602.

Nielsen, A., J. Karolefski and A. Heller. (2015). *Consumer-Centric Category Management: How to Increase Profits by Managing Categories Based on Consumer Needs*. Wiley.

Noble, S. M., K. B. Lee, R. Zaretzki and C. Autry. (2017). "Coupon clipping by impoverished consumers: Linking demographics, basket size, and coupon redemption rates." *International Journal of Research in Marketing*, *34*(2), 553–571.

Papakiriakopoulos, D., K. Pramatari and G. Doukidis. (2009). "A decision support system for detecting products missing from the shelf based on heuristic rules." *Decision Support Systems*, *46*(3), 685–694.

Park, C. H., Y.-H. Park and D. A. Schweidel. (2014). "A multi-category customer base analysis." *International Journal of Research in Marketing*, *31*(3), 266–279.

Pepe, M. R. and M. S. Pepe. (2012). "Using Point of Sale (POS) Data to Deliver Customer Value in the Supermarket Industry through Category Management

Practices." *Journal of Marketing Development & Competitiveness*, *6*(1), 69–73.

Pick, J. B., O. Turetken, A. V. Deokar and A. Sarkar. (2017). "Location analytics and decision support: Reflections on recent advancements, a research framework, and the path ahead." *Decision Support Systems*, *99*, 1–8.

Rahm, E. (2000). "Data Cleaning : Problems and Current Approaches." *Informatica*, 1–11.

Reutterer, T., A. Mild, M. Natter and A. Taudes. (2006). "A dynamic segmentation approach for targeting and customizing direct marketing campaigns." *Journal of Interactive Marketing*, *20*(3–4), 43–57.

Romme, A. G. L. (2003). "Making a Difference: Organization as Design." *Organization Science*, *14*(5), 558–573.

Rust, R. T. and M.-H. Huang. (2014). "The Service Revolution and the Transformation of Marketing Science." *Marketing Science*, *33*(2), 206–221.

Sands, S., C. Ferraro, C. Campbell and J. Pallant. (2016). "Segmenting multichannel consumers across search, purchase and after-sales." *Journal of Retailing and Consumer Services*, *33*, 62–71.

Sarantopoulos, P., A. Theotokis, K. Pramatari and G. Doukidis. (2016). "Shopping Missions: An Analytical Method for the Identification of Shopper Need States." *Journal of Business Research*, *69*(3), 1043–1052.

Sarantopoulos, P., A. Theotokis, K. Pramatari and A. L. Roggeveen. (2019). "The Impact of a Complement-Based Assortment Organization on Purchases." *Journal of Marketing Research*, 1–20.

Särndal, C.-E., B. Swensson and J. Wretman. (2003). *Model Assisted Survey Sampling (Springer Series in Statistics)*. Springer.

Sharma, R., S. Mithas and A. Kankanhalli. (2014). "Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations." *European Journal of Information Systems*, *23*(4), 433–441.

Shearer, C. (2000). "The CRIS-DM model: The New Blueprint for Data Mining." *Journal of Data Warehousing14*, *5*(4), 13–22.

Silberschatz, A. and A. Tuzhilin. (1996). "What makes patterns interesting in knowledge discovery systems." *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 970–974.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). MIT Press.

Skallerud, K., T. Korneliussen and S. O. Olsen. (2009). "An examination of consumers' cross-shopping behaviour." *Journal of Retailing and Consumer Services*, *16*(3), 181–189.

Song, I. and P. K. Chintagunta. (2006). "Measuring Cross-Category Price Effects with Aggregate Store Data." *Management Science*, *52*(10), 1594–1609.

Sorensen, H., S. Bogomolova, K. Anderson, G. Trinh, A. Sharp, R. Kennedy, ... M. Wright. (2017). "Fundamental patterns of in-store shopper behavior." *Journal of Retailing and Consumer Services*, *37*(February), 182–194.

Srikant, R. and R. Agrawal. (1995). "Mining Generalized Association Rules." In: *VLDB '95 Proceedings of the 21th International Conference on Very Large Data Bases* (pp. 407–419).

Stake, R. E. (1995). *The art of case research* (Vol. 80). Thousand Oaks, CA: Sage Publications.

Tang, K., Y.-L. Chen and H.-W. Hu. (2008). "Context-based market basket analysis in a multiple-store environment." *Decision Support Systems*, *45*(1), 150–163.

Thach, L. and J. Olsen. (2015). "Profiling the high frequency wine consumer by price segmentation in the US market." *Wine Economics and Policy*, *4*(1), 53–59.

Tibshirani, R., G. Walther and T. Hastie. (2001). "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *63*(2), 411–423.

Videla-Cavieres, I. F. and S. a. Ríos. (2014). "Extending market basket analysis with graph mining techniques: A real case." *Expert Systems with Applications*, *41*(4 PART 2), 1928–1936.

Vrechopoulos, A. P., R. M. O'Keefe, G. I. Doukidis and G. J. Siomkos. (2004). "Virtual store layout: An experimental comparison in the context of grocery retail." *Journal of Retailing*, *80*(1), 13–22.

Walters, R. G. and M. Jamil. (2003). "Exploring the relationships between shopping trip type, purchases of products on promotion, and shopping basket profit." *Journal of Business Research*, *56*(1), 17–29.

Wilkie, W. L. (1978). "Issues and advances in segmentation research." *Journal of Marketing Research*, *15*(3), 317–337.

Wolf, M. and S. McQuitty. (2011). "Understanding the do-it-yourself consumer: DIY motivations and outcomes." *AMS Review*, *1*, 154–170.

Wood, S. (2018). “The evolution of shopper behaviour.” Ipsos.

Woodside, A. G. (1973). “Patronage motives and Marketing Strategies.pdf.” *Journal of Retailing*.

Yao, Z., P. Sarlin, T. Eklund and B. Back. (2012). “Temporal customer segmentation using the self-organizing time map.” *Proceedings of the International Conference on Information Visualisation*, 234–240.

Yin, R. K. (1994). *Case study research: Design and Methods (second edition)*. Beverly Hills, CA: Sage Publishing.

You, Z., Y. W. Si, D. Zhang, X. Zeng, S. C. H. Leung and T. Li. (2015). “A decision-making framework for precision marketing.” *Expert Systems with Applications*, *42*(7), 3357–3367.

Zheng, Z., R. Kohavi and L. Mason. (2001). “Real world performance of association rule algorithms.” *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’01*, 401–406.

# APPENDIX

## Appendix A: Indicative structure of the analyzed retail datasets and indicative grocery payment receipt

### Dataset 1: POS data

| Store id | Date | Basket id | Barcode | Units | Price | Loyalty Card ID | Product in promo |
|---|---|---|---|---|---|---|---|
| **2** | 1/4/2015 | 11 | 1234121123413 | 2 | 10,22 | 1 | 1 |
| **2** | 1/4/2015 | 11 | 4221232222212 | 3 | 20,99 | 1 | 0 |
| **3** | 2/3/2015 | 22 | 1234121123433 | 1 | 2,99 | N/A | 1 |

### Dataset 2: Product taxonomy

| Level 1 Category | Level 2 Category | Level 3 Category | Item Description | Barcode |
|---|---|---|---|---|
| Bakery Products | Packed | Rye toast bread | 20 Slices Rye Toast Brand X | 1234121123413 |
| Beverages | Non-Alcohol | Cola | 1-liter Y cola light | 4225555212423 |

### Dataset 3: Loyalty cards data

| Loyalty card id | Gender | Marital status | Household size | Birthday |
|---|---|---|---|---|
| 1 | Female | Single | 2 | 7/12/1990 |
| 2 | Male | Married | 4 | 17/2/1959 |

## Indicative grocery payment receipt

Example Supermarket
Address | Phone
VAT

Store: 423
Transaction: 89483
Cashier: 02/Sveti

------------

| | |
|---|---|
| 8762198211212 Milk Free lact Brand X 1,5LT | 1.98 |
| 7382947293412 Muesli with Choco 2x 2.26 | 4.52 |
| 0003483928134 Soya Milk discount -0.38 | 2.74 |
| 32443454568690 Instant Coffee 50gr coupon -0.50 | 3.13 |

---------

**TOTAL:** 12.35€

Number of items: *5*

-----------------------------------------------------

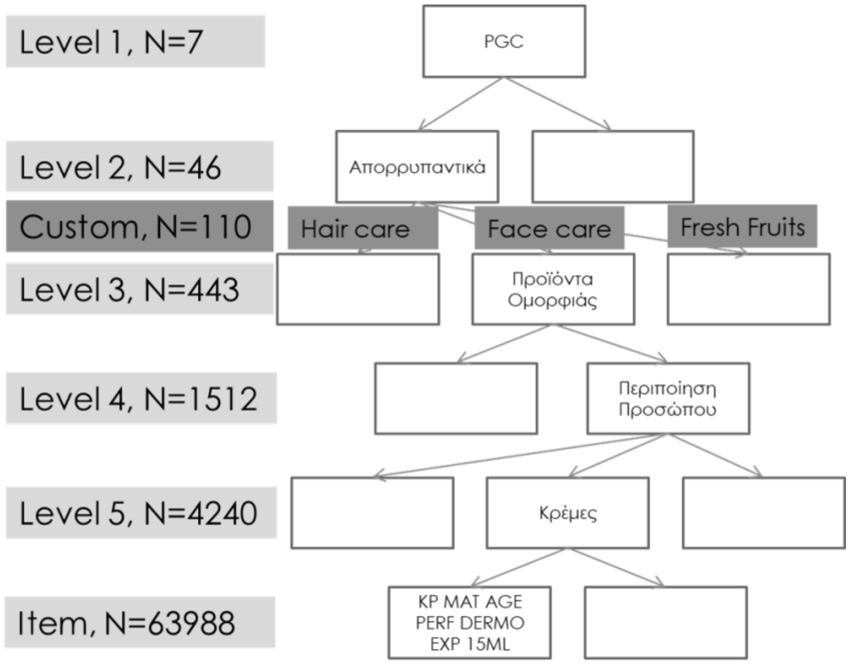
Date: 7/12/2018 | Time: 16:43:15

---------------------------------------------

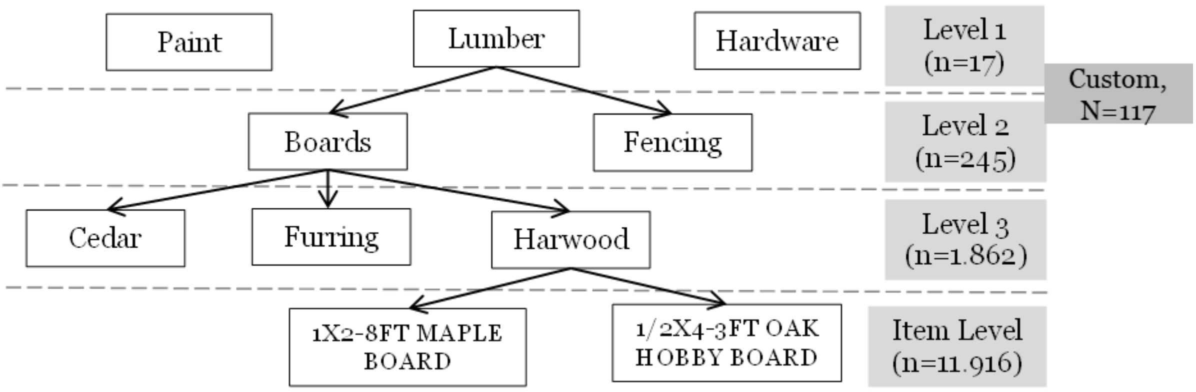
You earned 66 points
Loyalty card id: 4113922650

## Appendix B: Product taxonomies structure per case study

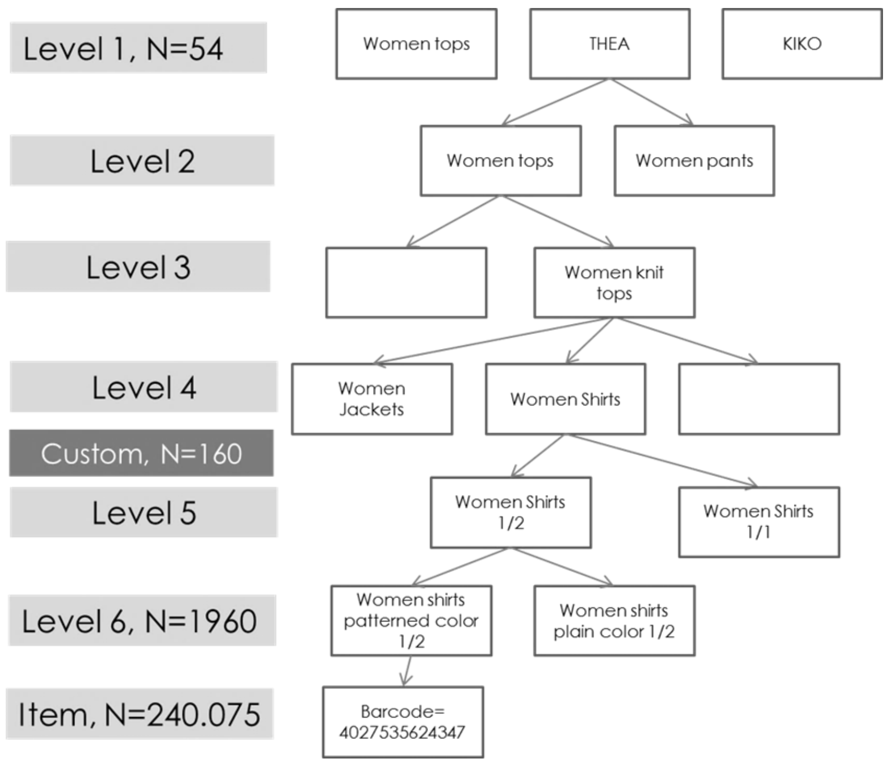
### FMCG retailer




### DIY retailer



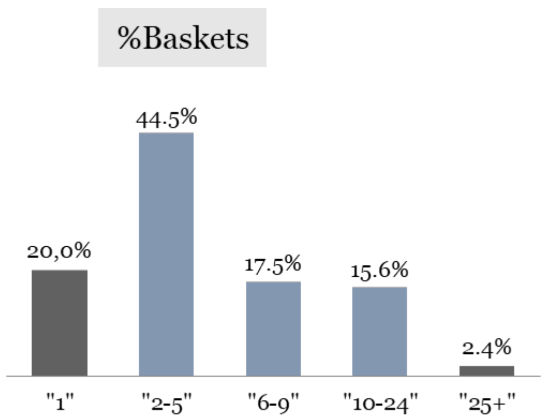
## Fashion retail

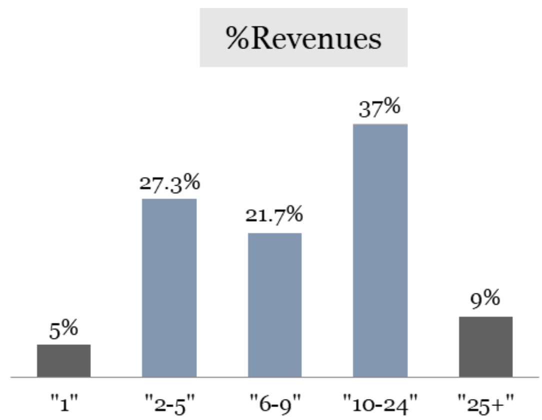


## Appendix C: Cluster sampling results – FMCG case

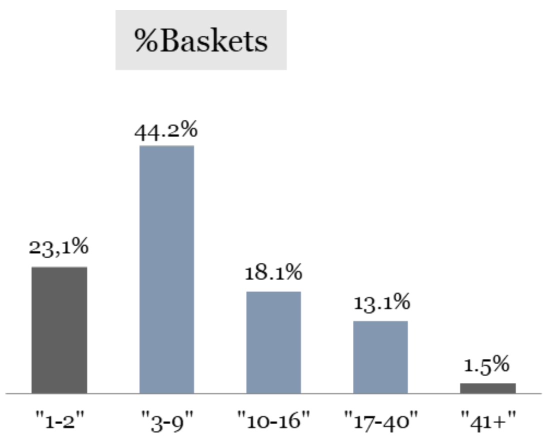
Below we present the cluster sampling results per store type of the FMCG retailer. The charts on the left depict information about the percentage of baskets contained per basket size range (volume). Similarly, the charts on the right, depict information regarding the revenues (value). Each bar depicts a cluster i.e. a visit/basket size range. Examining the charts below, the blue-color clusters were used for the analysis. Whereas, the grey bars were excluded.

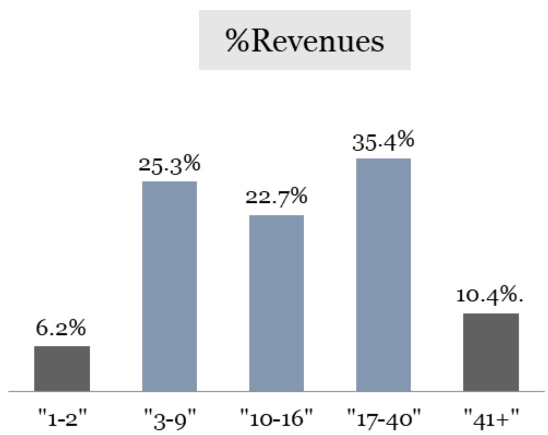
### Convenience stores



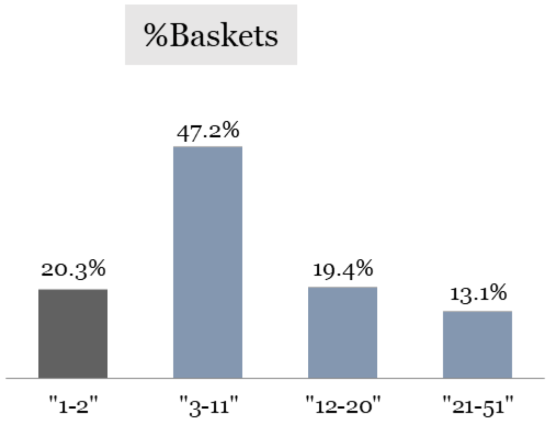



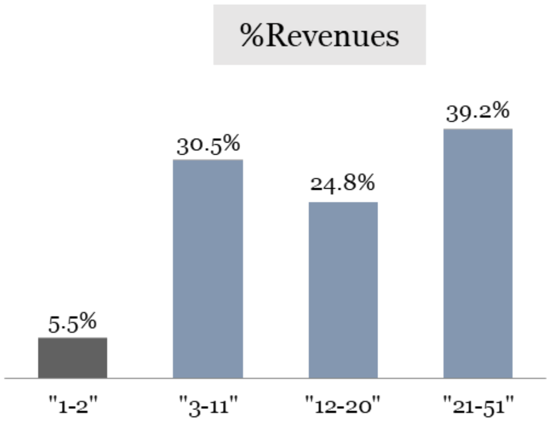

### Supermarket stores



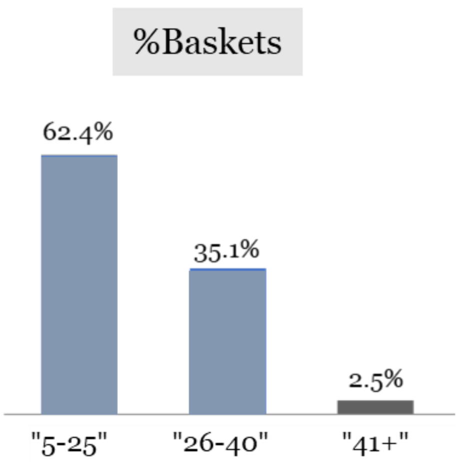



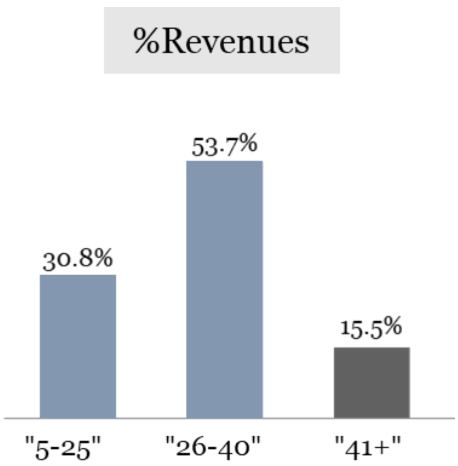
## Mini-hyper stores




## Web store



## Appendix D: Relation between shopping list and shopping mission

In the context of the field study, via the questionnaires at the store exit, we were able to track whether each shopper visited the store having a shopping list during the pilot. Results indicated that the 41,7% of the shoppers that visited the store to purchase products for their main course, had a shopping list (see Appendix Table 1). Likewise, all the shoppers that entered the store for pastry making had also a shopping list. This can be explained by the fact that this shopping mission demands to be more precise and not forget any ingredients from the pastry recipe. Comparing Appendix Table 1 with Table 6-3 which depicts the average basket size per mission, we can admit that there is a positive relation between the existence of a shopping list and the basket volume. Also, we confirm the statement of our focus groups participants who admit that they use a shopping list to perform more abstract shopping missions. However, still we cannot confirm that shoppers using shopping list do not perceive the existence of the shopping mission concept.

| Shopping mission | Shoppers percentage |
|---|---|
| Main course | 41.7% |
| Snacks and beverages | 18.2% |
| Pastry making | 100.0% |
| Personal care and hygiene | 33.3% |
| Breakfast | 22.2% |
| House cleaning and maintenance | 0.0% |
| Sandwich | 10.0% |
| Abstract detergent visits | 35.0% |
| Abstract food visits | 45.7% |

**Appendix Table 1. Percentage of shoppers having a shopping list vs shopping mission**

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ