ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΙΩΑΝΝΑΣ ΨΥΧΟΓΥΙΟΥ

ΤΙΤΛΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ
Τμηματοποίηση και Στόχευση της πελατειακής βάσης με Clustering και RFM Analysis.

Επιβλέπων : Γιώργος Λεκάκος – Αναπληρωτής Καθηγητής

Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Οκτώβριος 2019


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΙΩΑΝΝΑΣ ΨΥΧΟΓΥΙΟΥ

ΤΙΤΛΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ
Τμηματοποίηση και Στόχευση της πελατειακής βάσης με Clustering και RFM Analysis.

Επιβλέπων : Γιώργος Λεκάκος – Αναπληρωτής Καθηγητής

Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Οκτώβριος 2019

Βεβαίωση εκπόνησης Διπλωματικής εργασίας

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη μεταπτυχιακή εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών του ΠΜΣ στη Διοικητική Επιστήμη και Τεχνολογία του Τμήματος Διοικητικής Επιστήμης και Τεχνολογίας του Οικονομικού Πανεπιστημίου Αθηνών έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών στην Ελλάδα ή το εξωτερικό. Η εργασία αυτή έχοντας εκπονηθεί από εμένα, αντιπροσωπεύει τις προσωπικές μου απόψεις επί του θέματος. Οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης διπλωματικής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο».

(Υπογραφή)

.....

ΙΩΑΝΝΑ ΨΥΧΟΓΥΙΟΥ

Φοιτητής MSc στη Διοικητική Επιστήμη και Τεχνολογία



Περίληψη

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι να δείξει πως μπορούν να αξιοποιηθούν οι δυνατότητες που προσφέρει η ανάλυση και η εξόρυξη δεδομένων, προκειμένου να μπορέσουν οι επιχειρήσεις και οι οργανισμοί να αναγνωρίσουν και να εντοπίσουν τα τμήματα των καταναλωτών στα οποία απευθύνονται και τα ιδιαίτερα χαρακτηριστικά αυτών. Με αυτόν τον τρόπο θα είναι σε θέση οι επιχειρήσεις να τους προσφέρουν ανταγωνιστικά προϊόντα που να ανταποκρίνονται καλύτερα στις ανάγκες τους.

Στην σύγχρονη εποχή το περιβάλλον των επιχειρήσεων γίνεται όλο και πιο ανταγωνιστικό, ενώ ταυτόχρονα οι εταιρείες, λόγω της ανάπτυξης της τεχνολογίας έχουν στα χέρια τους πληθώρα δεδομένων που αφορούν τους πελάτες τους. Μόνο οι επιχειρήσεις που θα μπορέσουν να αναλύσουν και να αξιοποιήσουν σωστά αυτά τα δεδομένα θα καταφέρουν να επιβιώσουν και ενδεχομένως να κυριαρχήσουν στην αγορά. Γι' αυτό τον λόγο είναι επιτακτική η ανάγκη της αξιοποίησης των τεχνικών data mining, προκειμένου να αναγνωριστούν μοτίβα στην συμπεριφορά του αγοραστικού κοινού.

Στην παρούσα διπλωματική εργασία θα χρησιμοποιηθεί ο αλγόριθμος k-means, ο πιο γνωστός από τους αλγορίθμους συσταδοποίησης, πάνω σε πραγματικά δεδομένα πωλήσεων, προκειμένου να δείξουμε πως μπορεί να εφαρμοστεί στην πράξη και ποια η συμβολή του κατά την διαδικασία της τμηματοποίησης της αγοράς.

Λέξεις Κλειδιά: << Clustering, K-means, Customer Segmentation>>

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Abstract

The scope of this thesis is to show the way that the potentials of data analysis and mining can be exploit, in order to help organizations and enterprises to recognize and find the segments of the market in which they are targeting and their specific features. As a result, the companies will be able to offer them competitive products, which will be able to fulfil better customers' needs.

Nowadays, corporate environment is becoming more and more competitive, while companies, due to the technological growth, have access to a large dataset of variables, concerning their customers. Only those enterprises that will be able to analyze and take advantage of this information, will be able to survive and maybe conquer the market. So, it is critical companies to take advantage of the techniques of data mining, in order to find out patterns in the behavior of their target group.

On this thesis, we are going to show how algorithm k-means, the most famous of clustering algorithms, can be applied on real sales data and how can contribute in the procedure of market segmentation.

Key words: <<Clustering, K-means, Customer Segmentation>>

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Πίνακας Περιεχομένων

1. Εισαγωγή Σελ.2

1.1. Το σύγχρονο περιβάλλον των επιχειρήσεων Σελ.2

1.2. Τμηματοποίηση της Αγοράς και Clustering Σελ.4

1.2.1. Συνεισφορά Σελ.4

1.3. Οργάνωση Κειμένου.....Σελ.5

2. Θεωρητικό Υπόβαθρο..... Σελ.6

2.1. Customer Segmentation.....Σελ.6

2.2. RFM Analysis.....Σελ.9

2.3. Αλγόριθμοι Clustering.....Σελ.12

2.3.1.K-Means Algorithm.....Σελ**.13**

3. Ο κλάδος της χημικής βιομηχανίας..... Σελ.18

4. Περιγραφή της εταιρείας..... Σελ.23

5. Ανάλυση dataset και Υλοποίηση Clustering..... Σελ.26

5.1. Περιγραφή Dataset.....Σελ.26

5.2. Customer Segmentation.....Σελ.43

5.3. RFM Analysis.....Σελ.54

5.4. Προτάσεις Σελ.63

6. Τεχνικές Λεπτομέρειες..... Σελ.64

6.1. Λεπτομέρειες Υλοποίησης.....Σελ.64

6.2. Πλατφόρμες και Προγραμματιστικά Εργαλεία Σελ.65

7. Επίλογος..... Σελ.66

8. Βιβλιογραφία Σελ.67

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

1

ΕΙΣΑΓΩΓΗ

1.1 Το σύγχρονο περιβάλλον των επιχειρήσεων

Σε ένα περιβάλλον έντονου ανταγωνισμού για τις επιχειρήσεις, η ικανοποίηση του πελάτη είναι καθοριστικής σημασίας για την επιβίωση της επιχείρησης και για την απόκτηση συγκριτικού πλεονεκτήματος. Προκειμένου ένας οργανισμός όχι μόνο να καταφέρει να διατηρήσει τους πελάτες του αλλά και να τους αυξήσει, είναι απαραίτητο να μπορεί να τους παρέχει προϊόντα και υπηρεσίες υψηλής προστιθέμενης αξίας που να ανταποκρίνονται στις ανάγκες τους με τον καλύτερο δυνατό τρόπο. Γι' αυτόν ακριβώς τον λόγο είναι ιδιαίτερα σημαντικό οι επιχειρήσεις να μπορούν να αναγνωρίζουν τα τμήματα των καταναλωτών στα οποία απευθύνονται, τα χαρακτηριστικά καθενός από αυτά και τις ιδιαίτερες ανάγκες τους, έτσι ώστε να μπορούν να προσαρμόζουν κατάλληλα τα στοιχεία του μείγματος μάρκετινγκ. Να διευκρινίσουμε στο σημείο αυτό ότι το μείγμα μάρκετινγκ είναι ένα εργαλείο που χρησιμοποιείται από τους μαρκετίερς και τις επιχειρήσεις γενικότερα και

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

σχετίζεται με αποφάσεις που αφορούν το προϊόν (Product), την τιμή (Price), την διανομή (Place) και την προώθηση του προϊόντος (Promotion). Τα στοιχεία αυτά είναι γνωστά και ως 4P's στην ορολογία του μάρκετινγκ.

Ενώ όπως αναφέρθηκε είναι ιδιαίτερα σημαντικό οι επιχειρήσεις να μπορούν να αναγνωρίζουν τους πελάτες στους οποίους απευθύνονται, δηλαδή την αγορά στόχο, και τις ανάγκες τους, στις μέρες μας το έργο τους αυτό γίνεται όλο και πιο δύσκολο. Η πρόοδος της τεχνολογίας έχει οδηγήσει σε ιδιαίτερα μεγάλη αύξηση των δεδομένων που έχει ένας οργανισμός στα χέρια του και καλείται να αναλύσει προκειμένου να λάβει τις κατάλληλες αποφάσεις. Όμως, ο όγκος των δεδομένων που συλλέγονται καθώς και η ταχύτητα συλλογής τους μπορεί να αποτελέσουν ανασταλτικούς παράγοντες στη διαδικασία λήψης των αποφάσεων. Προκειμένου να ξεπεραστεί αυτό το πρόβλημα είναι απαραίτητο η διαδικασία της λήψης των αποφάσεων να υποστηρίζεται από κατάλληλα συστήματα ανάλυσης δεδομένων. Η ανάγκη αυτή οδήγησε στην ανάπτυξη της Εξόρυξης Δεδομένων. Με τον όρο εξόρυξη δεδομένων αναφερόμαστε στη διαδικασία ανακάλυψης χρήσιμων μοτίβων και σχέσεων μέσα από μεγάλες βάσεις δεδομένων. Μέσα από την ανάλυση μεγάλων ποσοτήτων δεδομένων, προσπαθεί να εξάγει κάποιο ενδιαφέρον πρότυπο, που μέχρι εκείνη την στιγμή δεν ήταν γνωστό. Πρόκειται ουσιαστικά για την εξαγωγή γνώσης μέσα από τα δεδομένα που είναι αποθηκευμένα στις βάσεις δεδομένων των επιχειρήσεων. Ανάμεσα στις τεχνικές που χρησιμοποιούνται βρίσκουμε τους Κανόνες Συσχέτισης (Association Rules), την Συσταδοποίηση (Clustering), την Κατηγοριοποίηση (Classification) και τις Προβλέψεις (Predictions).

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

1.2 Τμηματοποίηση της Αγοράς και Clustering

Η παρούσα διπλωματική εργασία θα επικεντρωθεί σε μια από τις τεχνικές της εξόρυξης δεδομένων, αυτή που ονομάζεται Συσταδοποίηση (Clustering). Με τον όρο Συσταδοποίηση αναφερόμαστε στην διαδικασία κατά την οποία τα δεδομένα ομαδοποιούνται σε σύνολα (clusters), έτσι ώστε τα δεδομένα κάθε cluster να είναι παρόμοια, αλλά να διαφοροποιούνται αρκετά από τα δεδομένα των υπόλοιπων clusters.

Στόχος της παρούσας διπλωματικής είναι να δείξει πως μπορεί να αξιοποιηθεί η τεχνική της Συσταδοποίησης από τις επιχειρήσεις, προκειμένου να μπορέσουν να εντοπίσουν τα τμήματα των καταναλωτών στα οποία απευθύνονται και τις ιδιαίτερες ανάγκες αυτών. Με αυτό τον τρόπο είναι πιθανό οι επιχειρήσεις να μπορέσουν να ανταποκριθούν καλύτερα στον αυξανόμενο ανταγωνισμό, και να ικανοποιήσουν αποτελεσματικά τις ανάγκες της αγοράς – στόχου αποκτώντας ανταγωνιστικό πλεονέκτημα.

Για την ανάλυση που πραγματοποιήθηκε χρησιμοποιήθηκε ο αλγόριθμος k – means. Πρόκειται για έναν αλγόριθμο που προτάθηκε για πρώτη φορά από τον MacQueen, το 1967 και ο οποίος είναι ιδιαίτερα δημοφιλής λόγω της απλότητας και της ευκολίας στην εφαρμογή του και της αποτελεσματικότητας του. Τέλος, η ανάλυση μας πραγματοποιήθηκε με χρήση της γλώσσας προγραμματισμού Python.

1.2.1 Συνεισφορά

Στην παρούσα διπλωματική αρχικά μελετήσαμε εκτενώς την έννοια του Customer Segmentation και της RFM Analysis, στην συνέχεια αναφερθήκαμε στους αλγορίθμους συσταδοποίησης και εκτενέστερα στον αλγόριθμο K-means. Τέλος, δείξαμε πως μπορούν να αξιοποιηθούν οι τεχνικές αυτές σε πραγματικά δεδομένα πωλήσεων που προέρχονται από μια ελληνική εταιρεία που δραστηριοποιείται στον κλάδο της χημικής βιομηχανίας.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

1.3 Οργάνωση Κειμένου

Στο κεφάλαιο 2 εξηγούμε εκτενώς τις έννοιες που θα μελετήσουμε παρακάτω και τον αλγόριθμο που θα χρησιμοποιήσουμε για την ανάλυση μας. Στο κεφάλαιο 3 περιγράφουμε τον κλάδο στον οποίο δραστηριοποιείται η υπό εξέταση επιχείρηση. Ακολουθεί στο κεφάλαιο 4 η παρουσίαση της επιχείρησης. Στο κεφάλαιο 5 προχωράμε με την ανάλυση των δεδομένων μας προκειμένου να τα κατανοήσουμε σε βάθος. Στη συνέχεια δείχνουμε πως μπορεί να χρησιμοποιηθεί ο αλγόριθμος k-means για την τμηματοποίηση της πελατειακής βάσης και προχωράμε με την RFM Analysis, προκειμένου να καταλήξουμε σε συμπεράσματα και προτάσεις προς την επιχείρηση. Στο κεφάλαιο 6 εξηγούμε τις τεχνικές λεπτομέρειες και δυσκολίες της ανάλυσης μας. Και στο κεφάλαιο 7 κάνουμε μια σύνοψη όσων αναφέρθηκαν. Τέλος στο κεφάλαιο 8 παρουσιάζεται η σχετική βιβλιογραφία.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

2

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Customer Segmentation

Η διαδικασία της τμηματοποίησης της αγοράς, η στόχευση των τμημάτων αυτών και η τοποθέτηση του προϊόντος αποτελούν βασικές λειτουργίες του μάρκετινγκ. Η έννοια της τμηματοποίησης της αγοράς προέκυψε από την διαπίστωση ότι όλοι οι αγοραστές δεν είναι ίδιοι, δεν έχουν τις ίδιες ανάγκες ούτε τις ίδιες αντιλήψεις για τα προϊόντα, και επίσης, οι επιχειρήσεις έχουν περιορισμένους πόρους στην διάθεσή τους και επομένως δεν μπορούν να ικανοποιήσουν εξίσου όλους τους καταναλωτές.

Με τον όρο τμηματοποίηση της αγοράς αναφερόμαστε ουσιαστικά στην μελέτη της αγοράς, προκειμένου να εντοπιστούν τα τμήματα της, να υπολογιστεί το μέγεθος καθενός από αυτά, να αναγνωριστούν οι ανάγκες κάποιων από αυτών των τμημάτων και ο τρόπος με τον οποίο είναι δυνατόν αυτές να ικανοποιηθούν, και στην πορεία να προσεγγιστούν αυτά τα τμήματα, ώστε να ικανοποιηθούν οι ανάγκες τους με το κατάλληλο μίγμα μάρκετινγκ. Πρόκειται για την συστηματική

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

διαίρεση της συνολικής αγοράς έτσι ώστε να εντοπιστούν οι ομάδες καταναλωτών που έχουν παρόμοιες ανάγκες και που αντιδρούν στις μεταβλητές του μίγματος μάρκετινγκ με παρόμοιο τρόπο, όπως αναφέρει ο Γεώργιος Ι. Σιώμκος, στο βιβλίο του Συμπεριφορά Καταναλωτή και Στρατηγική Μάρκετινγ..

Για να θεωρηθεί επιτυχής μια τμηματοποίηση, σύμφωνα με τον Πέτρο Γ. Μάλλιαρη, στο βιβλίο του Εισαγωγή στο Μάρκετινγκ θα πρέπει οι αγοραστές κάθε τμήματος να έχουν αρκετά όμοια χαρακτηριστικά, να συμπεριφέρονται δηλαδή και να αντιδρούν με τον ίδιο τρόπο. Ακόμη, θα πρέπει οι αγοραστές κάθε τμήματος να διαφοροποιούνται αρκετά από τους αγοραστές των υπόλοιπων τμημάτων, να αποτελούν δηλαδή μια ετερογενή ομάδα. Επιπλέον, θα πρέπει τουλάχιστον ένα από τα τμήματα της αγοράς που έχουν εντοπιστεί, να είναι αρκετά μεγάλο ώστε να παρουσιάζει ενδιαφέρον από οικονομική άποψη, να μπορεί να αποτελέσει δηλαδή την αγορά στόχο και ταυτόχρονα το τμήμα αυτό να είναι δυνατόν να προσεγγιστεί μέσω του μίγματος μάρκετινγκ και κυρίως μέσω της προβολής και της διανομής.

Για να είναι επιτυχής η επιλογή του τμήματος, δηλαδή η επιλογή της αγοράς στόχου, σύμφωνα με τον Abratt R. (1993), θα πρέπει το τμήμα αυτό να έχει προοπτικές ανάπτυξης και μεγέθυνσης, τα άτομα που το αποτελούν να αναμένεται να αυξηθούν με το πέρασμα του χρόνου και να μην ελκύει ή να μην αναμένεται να έλξει στο μέλλον ανταγωνιστές. Ακόμη, θα πρέπει τα έσοδα από την συγκεκριμένη αγορά - στόχο να είναι τέτοιου ύψους που να μεγιστοποιούν την αποδοτικότητα και το κόστος ικανοποίησης των καταναλωτών να μην είναι δυσανάλογο, με βάση την υπάρχουσα στρατηγική που ακολουθεί η επιχείρηση. Και τέλος θα πρέπει η ικανοποίηση των αναγκών του συγκεκριμένου τμήματος της αγοράς να συμβαδίζει με την αποστολή και το όραμα της εταιρείας.

Η αγορά θα μπορούσαμε να υποστηρίξουμε ότι διακρίνεται σε καταναλωτική (B2C) και σε βιομηχανική (B2B), με αρκετά διαφορετικά χαρακτηριστικά η κάθε μία. Το βασικό κριτήριο διαφοροποίησης τους είναι το κίνητρο και ο σκοπός αγοράς του προϊόντος. Στην καταναλωτική αγορά ο πελάτης – αγοραστής είναι ο τελικός καταναλωτής του προϊόντος και προβαίνει στην αγορά με σκοπό να το χρησιμοποιήσει ο ίδιος. Από την άλλη στην βιομηχανική αγορά, πελάτης είναι κάποια επιχείρηση, που αγοράζει ένα προϊόν είτε με σκοπό να το μεταπωλήσει είτε να το χρησιμοποιήσει στην παραγωγική της διαδικασία.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Υπάρχουν διάφορα κριτήρια που μπορούν να χρησιμοποιήσουν οι επιχειρήσεις για την τμηματοποίηση της αγοράς των καταναλωτών. Θα μπορούσαμε να υποστηρίξουμε ότι αυτά τα κριτήρια μπορούν να χωριστούν σε τέσσερις κατηγορίες. Οι κατηγορίες αυτές είναι γεωγραφικά κριτήρια, δημογραφικά κριτήρια, ψυχογραφικά κριτήρια και προϊοντικά κριτήρια. Με τον όρο γεωγραφικά κριτήρια αναφερόμαστε κυρίως στην τοποθεσία διαμονής του καταναλωτή και περιλαμβάνει επίσης το κλίμα, την περιοχή (π.χ. αστική, αγροτική), το μέγεθος της πόλης, την πυκνότητα του πληθυσμού και την διαμόρφωση εδάφους (π.χ. πεδινό, ορεινό, νησί). Τα κριτήρια αυτά συνήθως έχουν μικρή επίδραση στην συμπεριφορά των καταναλωτών, με εξαίρεση ενδεχομένως σε ορισμένα προϊόντα (όπως πχ. παλτά, αδιάβροχα). Το μειονέκτημα αυτών των κριτηρίων είναι ότι δεν παρέχουν αρκετές πληροφορίες για τον πελάτη, αν και είναι εύκολα στην εφαρμογή. Όσον αφορά τα δημογραφικά κριτήρια αποτελούν τα πιο συχνά χρησιμοποιούμενα και περιλαμβάνουν την ηλικία, το φύλο, το εισόδημα, το επάγγελμα, το επίπεδο μόρφωσης, την καταγωγή, την θρησκεία, την κοινωνική τάξη και την οικογενειακή κατάσταση. Οι πωλήσεις πολλών προϊόντων εξαρτώνται από τα δημογραφικά χαρακτηριστικά των καταναλωτών. Ακόμα όμως και αν δύο πελάτες έχουν ακριβώς τα ίδια δημογραφικά χαρακτηριστικά και γεωγραφικά, δεν σημαίνει ότι θα συμπεριφερθούν με τον ίδιο ακριβώς τρόπο. Το κενό αυτό έρχονται να συμπληρώσουν τα ψυχογραφικά κριτήρια, τα οποία περιλαμβάνουν τον τρόπο ζωής και την προσωπικότητα του καταναλωτή. Μερικά από τα στοιχεία που χρησιμοποιούνται για να μελετηθούν αυτά τα χαρακτηριστικά είναι η δραστηριότητα, η ανεξαρτησία, η κοινωνικότητα, η προσαρμοστικότητα, η σοβαρότητα και ο έλεγχος των συγκινήσεων. Με βάση αυτά τα χαρακτηριστικά μπορούν να δημιουργηθούν τύποι καταναλωτών που θα χρησιμοποιηθούν ως βάση για την τμηματοποίηση της αγοράς. Τέλος, στην τέταρτη κατηγορία ανήκουν τα προϊοντικά κριτήρια, κριτήρια δηλαδή που περιγράφουν την σχέση καταναλωτή – προϊόντος. Συγκεκριμένα περιλαμβάνει την χρήση του προϊόντος και τον τρόπο χρήσης του, τις προσδοκώμενες ωφέλειες από την χρήση του προϊόντος, την μάρκα του προϊόντος και την τιμή αυτού, τον τρόπο απόκτησης του προϊόντος, τον βαθμό στον οποίο οι καταναλωτές γνωρίζουν το προϊόν καθώς επίσης και τον βαθμό ανάμιξης με το προϊόν.

Όσον αφορά την βιομηχανική αγορά, αυτή διαφέρει αρκετά σε σχέση με την καταναλωτική αγορά. Στην βιομηχανική αγορά, οι αγορές δεν γίνονται για την άμεση ικανοποίηση των αναγκών των

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

καταναλωτών αλλά για την έμμεση ικανοποίηση αναγκών τρίτων. Ουσιαστικά οι αγορές αποτελούν εισροές μιας παραγωγικής διαδικασίας, η οποία θα οδηγήσει σε ένα τελικό προϊόν, το οποίο με την σειρά του θα ικανοποιήσει ανάγκες. Κύριο χαρακτηριστικό της συμπεριφοράς των βιομηχανικών αγοραστών είναι ο ορθολογισμός. Γι' αυτόν ακριβώς το λόγο η τμηματοποίηση της θεωρείται πιο εύκολη σε σχέση με την καταναλωτική αγορά. Κριτήρια για την τμηματοποίηση της βιομηχανικής αγοράς που χρησιμοποιούνται συχνά είναι η γεωγραφική τοποθεσία του βιομηχανικού αγοραστή, ο κλάδος της δραστηριότητας του και το μέγεθος του, το μέγεθος της παραγγελίας και η συχνότητα παραγγελιών καθώς και η χρήση του προϊόντος.

Η τμηματοποίηση της πελατειακής βάσης προσφέρει πολλά πλεονεκτήματα στις επιχειρήσεις και τους δίνει την δυνατότητα να διαφοροποιηθούν. Τους επιτρέπει να εντοπίσουν τους πιο κερδοφόρους και τους λιγότερους κερδοφόρους πελάτες, τους δίνει την δυνατότητα μέσω της ανάλυσης δεδομένων να προβλέψουν μελλοντικές αγοραστικές συμπεριφορές, να βελτιώσουν την αποτελεσματικότητα του μάρκετινγκ και να δημιουργήσουν ισχυρές σχέσεις με τους πελάτες τους, να ακολουθήσουν διαφορετική τιμολογιακή πολιτική και να δημιουργήσουν βελτιωμένα προϊόντα με εξειδικευμένα χαρακτηριστικά που να ανταποκρίνονται καλύτερα στις ανάγκες της αγοράς – στόχου.

Οι δυνατότητες και τα πλεονεκτήματα που προσφέρει ο κλάδος των analytics και της ανάλυσης δεδομένων, μπορούν να αξιοποιηθούν από τις επιχειρήσεις ώστε να τους διευκολύνουν στην διαδικασία της τμηματοποίησης της αγοράς, ιδίως στις μέρες μας όπου σχεδόν όλα τα απαιτούμενα στοιχεία υπάρχουν και είναι εύκολο να βρεθούν ηλεκτρονικά.

2.2 RFM Analysis

Η ανάλυση RFM αποτελεί μια από τις πλέον συχνά χρησιμοποιούμενες τεχνικές του μάρκετινγκ για την τμηματοποίηση της πελατειακής βάσης. Το ακρωνύμιο RFM προέρχεται από τα αρχικά των λέξεων Recency, Frequency και Monetary και ουσιαστικά χρησιμοποιείται για να χωρίσει τους πελάτες με βάση τρία κριτήρια: πόσο συχνά ο πελάτης αγοράζει από την εταιρεία μας (frequency), πόσο πρόσφατα έχει ο πελάτης προβεί σε κάποια αγορά από την εταιρεία μας (recency) και ποιο το

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ποσό που ξοδεύει ο πελάτης στις αγορές του από την εταιρεία μας (monetary). Υιοθετώντας αυτό το μοντέλο, γίνεται πιο εύκολο για τις επιχειρήσεις να αναγνωρίσουν τους πιο πολύτιμους – σημαντικούς πελάτες τους και να αναπτύξουν κατάλληλες στρατηγικές μάρκετινγκ, παρέχοντας τους προσωποποιημένες και εξατομικευμένες υπηρεσίες. Ακόμη, είναι πιο εύκολο να εντοπίσουν τα τμήματα εκείνα των καταναλωτών που θα ανταποκριθούν καλύτερα σε μια προωθητική ενέργεια (D.Birant, 2011). Όπως χαρακτηριστικά αναφέρουν οι Sohrabi και Khanlari (2007) εφόσον δεν είναι όλοι οι πελάτες το ίδιο επικερδείς για τις επιχειρήσεις, είναι ιδιαίτερα σημαντικό να εντοπίσουν πρώτα την κερδοφορία του καθενός και μετά να αξιοποιήσουν τους πόρους που έχουν στην διάθεση τους, προκειμένου να τους προσεγγίσουν σύμφωνα με την αξία που προσδίδουν στην επιχείρηση. Επίσης, προκειμένου οι επιχειρήσεις να μπορέσουν να κρατήσουν τους πελάτες τους και να βελτιώσουν την αποδοτικότητα τους, χρειάζεται να προσαρμόζουν την στρατηγική μάρκετινγκ που ακολουθούν και να ανταποκρίνονται σε διαφορετικές ανάγκες των καταναλωτών, αξιοποιώντας κατάλληλα τους πόρους που έχουν στην διάθεσή τους (Huang et al., 2009 , Chang et al., 2010).

Το μοντέλο RFM είναι ένα συμπεριφορικό μοντέλο που χρησιμοποιείται για να αναλύσει την συμπεριφορά των πελατών βασιζόμενο στις τρεις προαναφερθείσες μετρήσεις: regency, frequency και monetary, οι οποίες θεωρείται πως συνδέονται με την πιθανότητα οι πελάτες να προβούν σε μελλοντικές αγορές. Ο όρος recency αναφέρεται στο χρονικό διάστημα που έχει περάσει από την τελευταία αγορά του πελάτη έως σήμερα. Πολλοί μαρκετίερς πιστεύουν ότι αυτοί που έχουν προβεί πρόσφατα σε αγορά είναι πιο πιθανό να ξανά – αγοράσουν. Η συχνότητα (frequency) μετράει τον αριθμό των συναλλαγών που έχει πραγματοποιήσει ο πελάτης με την εταιρεία μας μέσα σε ένα δεδομένο χρονικό διάστημα. Αυτή η μέτρηση βασίζεται στην θεώρηση ότι οι πελάτες που αγοράζουν συχνά είναι πιο πιθανό να αγοράσουν προϊόντα σε σχέση με αυτούς που έχουν λιγότερες αγορές. Τέλος, με τον όρο Monetary υπολογίζεται το άθροισμα των συνολικών χρημάτων που έχει ξοδέψει ένα συγκεκριμένος πελάτης στις αγορές του.

Η αυξημένη εφαρμογή της τεχνικής RFM για την τμηματοποίηση των πελατών από το μάρκετινγκ μπορεί να δικαιολογηθεί καθώς υπάρχει μια πληθώρα πλεονεκτημάτων που απορρέουν από την χρήση της. Σύμφωνα με τους Wei et al. (2010), η χρήση αυτού του μοντέλου είναι ιδιαίτερα αποδοτική από οικονομικής άποψης, καθώς πλέον είναι ιδιαίτερα εύκολο να αποκτήσεις και να ποσοτικοποιήσεις

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

δεδομένα της συμπεριφοράς των καταναλωτών, εφόσον πλέον τα δεδομένα αυτά είναι εύκολο να αποθηκευτούν σε ηλεκτρονική μορφή. Επίσης, με την χρήση αυτού του μοντέλου είναι εύκολο να εξάγεις συμπεράσματα για την καταναλωτική συμπεριφορά με χρήση ενός μικρού αριθμού μεταβλητών. Ακόμη, η τεχνική αυτή είναι ιδιαίτερα αξιόπιστη στην πρόβλεψη μελλοντικών συμπεριφορών και μπορεί να συμβάλει στην αύξηση της κερδοφορίας. Επιπλέον, είναι ιδιαίτερα αποδοτική στην στόχευση από πλευράς μάρκετινγκ και πωλήσεων συγκεκριμένων τμημάτων των καταναλωτών. Και τέλος, πρόκειται για ένα μοντέλο που είναι ιδιαίτερα αξιόπιστο στην εκτίμηση της σχέσης της εταιρείας με τον καταναλωτή και στο πόσο ισχυρή είναι αυτή, καθώς μπορεί να εντοπίσει αρκετά αποτελεσματικά τους πιο αξιόπιστους πελάτες για την επιχείρηση.

Από την άλλη βέβαια, το γεγονός ότι δίνει τόση έμφαση στους πιο πολύτιμους πελάτες της εταιρείας θα μπορούσε να θεωρηθεί ως ένα από τα μειονεκτήματα αυτού του μοντέλου, καθώς οι πληροφορίες που δίνει όταν ένας πελάτης δεν έχει αγοράσει πρόσφατα, δεν αγοράζει συχνά και έχει δαπανήσει ένα μικρό ποσό για τις αγορές του, δεν είναι εύκολο να ερμηνευθούν. Ακόμη, η απλότητα του μοντέλου με την χρήση λίγων μεταβλητών ταυτόχρονα αποτελεί και μειονέκτημα του, καθώς υπάρχουν πολλά χαρακτηριστικά που επιδρούν στις αγοραστικές αποφάσεις των καταναλωτών. Επίσης, η τεχνική αυτή δίνει έμφαση στους υπάρχοντες πελάτες της επιχείρησης χωρίς να μπορεί να εφαρμοστεί σε μελλοντικούς, καθώς η εταιρεία δεν διαθέτει δεδομένα συναλλαγών γι’ αυτούς. Επιπλέον, προβλέπει έναν τρόπο αντίδρασης για όλους τους πελάτες της, θεωρώντας ότι υπάρχει ομοιογένεια ανάμεσα τους, αγνοώντας την ετερογένεια που συναντάμε ανάμεσα στους καταναλωτές. Τέλος, δεν μπορεί να θεωρηθεί ως ένα απόλυτα ακριβές ποσοτικό μοντέλο, καθώς η σημασία κάθε μέτρησης της τεχνικής RFM διαφέρει από κλάδο σε κλάδο. (Wei et al, 2010)

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

2.3 Αλγόριθμοι Clustering

Η συσταδοποίηση (clustering) αποτελεί μια από τις πλέον συχνά χρησιμοποιούμενες τεχνικές που χρησιμοποιούνται στην ανάλυση δεδομένων προκειμένου να αποκτήσουμε μια πρώτη αντίληψη της δομής τους. Μέσω της τεχνικής αυτής, είναι εύκολο να αναγνωριστούν υποσύνολα (συστάδες) δεδομένων, έτσι ώστε τα δεδομένα που ανήκουν στην ίδια συστάδα να έχουν αρκετά παρόμοια χαρακτηριστικά, ενώ αυτά που ανήκουν σε διαφορετικές συστάδες να διαφέρουν αρκετά. Ουσιαστικά, εντοπίζονται ομοιογενή υποσύνολα έτσι ώστε τα δεδομένα κάθε υποσυνόλου να είναι όσο πιο όμοια γίνεται με βάση το χαρακτηριστικό που έχουμε ορίσει ως κριτήριο (Huang Z., 1998).

Η τεχνική της συσταδοποίησης ανήκει στην κατηγορία της μη καθοδηγούμενης μάθησης (unsupervised learning). Διαχειρίζεται ουσιαστικά δεδομένα τα οποία δεν έχουν κάποιο label. Το clustering συνήθως αποτελεί το πρώτο βήμα στη τεχνική της εξόρυξης δεδομένων, που μπορεί να χρησιμοποιηθεί σαν το σημείο εκκίνησης για την ανακάλυψη άγνωστων σχέσεων ανάμεσα στα δεδομένα μας. Επιπλέον ανάλυση μπορεί να συμβάλει στην ανακάλυψη επιπλέον χαρακτηριστικών που σχετίζονται με αυτές τις ομάδες. Το clustering αποτελεί μια τεχνική εξόρυξης δεδομένων που τοποθετεί στοιχεία δεδομένων σε διάφορες ομάδες χωρίς προηγούμενη γνώση των ομάδων που θα δημιουργηθούν. Έχουν αναπτυχθεί διάφορες μέθοδοι και αλγόριθμοι που μπορούν να χρησιμοποιηθούν κατά την διαδικασία της συσταδοποίησης (clustering).

Οι αλγόριθμοι clustering διακρίνονται σε ιεραρχικούς (hierarchical), σε διαμεριστικούς (partitional) και σε αλγόριθμους που βασίζονται στην πυκνότητα (Rai & Singh, 2010). Η ιεραρχική συσταδοποίηση δημιουργεί μια ιεραρχία από clusters. Οι αλγόριθμοι αυτής της κατηγορίας όπως ο Lance – Williams και ο SLINK, δημιουργούν clusters βαθμιαία. Η διαδικασία της συσταδοποίησης μπορεί να παρουσιαστεί γραφικά με ένα διάγραμμα δύο διαστάσεων, γνωστό ως δενδρόγραμμα, το οποίο αναπαριστά τις ενώσεις και τις διαιρέσεις σε κάθε διαδοχικό στάδιο της διαδικασίας. Οι αλγόριθμοι αυτής της κατηγορίας διακρίνοντα σε σωρευτικούς (agglomerative) και διαιρετικούς (divisive). Οι διαιρετικοί, όπως δηλώνει και το όνομα τους διαιρούν ένα σύνολο n αντικειμένων σε

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

μικρότερες ομάδες, ενώ οι σωρευτικοί, που είναι και πιο συχνά χρησιμοποιούμενοι, δημιουργούν ενώσεις n αντικειμένων

Απ' την άλλη, η διαμεριστική συσταδοποίηση καταλήγει σε ένα σύνολο Ν συστάδων (clusters), όπου κάθε στοιχείο δεδομένων ανήκει σε ένα cluster. Σε κάθε cluster αντιστοιχεί ένα κέντρο (centroid), που αντιπροσωπεύει το σύνολο των στοιχείων της συγκεκριμένης συστάδας. Ο πλέον χαρακτηριστικός αλγόριθμος της συγκεκριμένης κατηγορίας είναι ο K-Means, τον οποίο θα αναλύσουμε περισσότερο παρακάτω. Άλλοι αλγόριθμοι που ανήκουν σε αυτή την κατηγορία είναι ο IPAM (Partitioning Around Medoids) και ο CLARA (Clustering Large Applications).

Τέλος, οι βασιζόμενοι στην πυκνότητα αλγόριθμοι είναι ικανοί να δημιουργήσουν clusters τυχαίων μορφών. Αυτού του τύπου οι αλγόριθμοι ομαδοποιούν τα στοιχεία με βάση το πλήθος των δεδομένων που υπάρχει σε κάποια "γειτονιά" και συνεχίζουν να αυξάνονται όσο ο αριθμός των αντικειμένων υπερβαίνει κάποια οριζόμενη τιμή. Πλεονέκτημα της μεθόδου αυτής είναι ότι προστατεύει από τις ακραίες τιμές που μπορεί να υπάρχουν στα δεδομένα, δηλαδή δεν επηρεάζεται από αυτές.

Στην συνέχεια αναφερόμαστε αναλυτικότερα στον αλγόριθμο K- Means.

2.3.1 K- Means Algorithm

Ένας από τους πιο γνωστούς και ευρέως χρησιμοποιούμενους αλγόριθμους που χρησιμοποιείται κατά την συσταδοποίηση είναι ο αλγόριθμος k-means. Πρόκειται για έναν επαναληπτικό αλγόριθμο που προσπαθεί να χωρίσει τα δεδομένα σε k προκαθορισμένα groups, μη αλληλό-επικαλυπτόμενα, έτσι ώστε κάθε στοιχείο να ανήκει σε μια μόνο ομάδα. Προσπαθεί να κάνει τα στοιχεία κάθε συστάδας όσο πιο όμοια γίνεται, ενώ ταυτόχρονα διαφοροποιεί κάθε ομάδα από τις υπόλοιπες. Αναθέτει στοιχεία σε κάθε συστάδα έτσι ώστε να μειώνεται στο ελάχιστο η τετραγωνική απόσταση ανάμεσα στα στοιχεία του cluster και στο κέντρο του.

Αρχικά ορίζουμε τον αριθμό k των clusters. Στην συνέχεια επιλέγουμε τυχαία το κέντρο (centroid) κάθε συστάδας και αναθέτουμε κάθε στοιχείο των δεδομένων μας στο κοντινότερο cluster με βάση συνήθως την Ευκλείδεια Απόσταση. Αφού έχουν ανατεθεί όλα τα στοιχεία σε ένα cluster, ξανά – υπολογίζουμε το κέντρο κάθε cluster και ξανά – αναθέτουμε τα στοιχεία στο πιο κοντινό cluster. Η

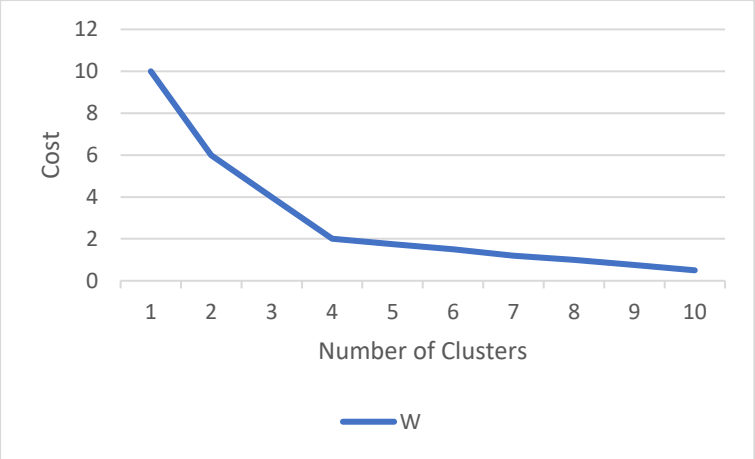
ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

διαδικασία αυτή συνεχίζεται μέχρι να μην υπάρχει καμία αλλαγή στα στοιχεία κάθε συστάδας. Συνήθως ορίζουμε η διαδικασία να ολοκληρώνεται μετά από κάποιο αριθμό επαναλήψεων που θεωρούμε ότι θα αποδώσει ικανοποιητική συσταδοποίηση. (Satish G. et al, 2014)

Ο αλγόριθμος αυτός παράγει ακριβώς k clusters, που διαφοροποιούνται όσο το δυνατόν περισσότερο. Πρόκειται για έναν αλγόριθμο που είναι εύκολος να εφαρμοστεί και παράγει αξιόπιστα αποτελέσματα. Ακόμη έχει την δυνατότητα να διαχειριστεί μεγάλα αρχεία δεδομένων. Το μεγαλύτερο μειονέκτημα του είναι ότι πρέπει να προκαθορίσεις τον αριθμό των clusters που θα παράγει, γεγονός που απαιτεί αρκετά καλή γνώση των διαθέσιμων δεδομένων. Ακόμη, ο αλγόριθμος μπορεί να πέσει σε τοπικό ελάχιστο και όχι σε καθολικό, καθώς διαφορετικές αρχικοποιήσεις μπορεί να οδηγήσουν σε διαφορετικά clusters. Ως εκ τούτου, συνήθως χρησιμοποιούνται διαφορετικές αρχικοποιήσεις των κέντρων των clusters κατά την εκτέλεση του αλγορίθμου και επιλέγεται εκείνη με την μικρότερη τετραγωνική απόσταση. Τέλος, επειδή ο αλγόριθμος χρησιμοποιεί για την αξιολόγηση του αποτελέσματος μετρήσεις που βασίζονται στην απόσταση, θεωρείται σκόπιμη η τυποποίηση των δεδομένων στην κανονική κατανομή Z(0,1) , καθώς συνήθως τα δεδομένα που περιλαμβάνονται σε ένα dataset έχουν διαφορετικές μονάδες μέτρησης (Steinley & Brusco, 2008).

Ένα από τα μειονεκτήματα του αλγορίθμου K means όπως ήδη αναφέρθηκε, αποτελεί το γεγονός ότι πρέπει ο χρήστης να ορίσει εξ αρχής τον αριθμό k των clusters που θα χρησιμοποιήσει. Έχουν αναπτυχθεί διάφορες προσεγγίσεις που μπορούν να χρησιμοποιηθούν για τον καθορισμό του αριθμού των clusters. Μια από αυτές είναι η Μέθοδος Elbow. Η μέθοδος αυτή καταλήγει σε μια γραφική απεικόνιση του αριθμού των clusters συναρτήσει του κόστους. Σύμφωνα με τους Kodinariya & Makwana (2013), ο αλγόριθμος εκτελείται αρχικά για k=2 και σε κάθε επανάληψη αυξάνεται ο αριθμός των clusters κατά μια μονάδα, υπολογίζοντας την σχέση που υπάρχει ανάμεσα στο αριθμό των συστάδων και στο κόστος που προκύπτει. Μέχρι κάποια τιμή του k το κόστος μειώνεται με αύξοντα ρυθμό και μετά από εκείνο το σημείο το κόστος μειώνεται με φθίνοντα ρυθμό. Από αυτό το σημείο και μετά κάθε cluster που προκύπτει είναι πολύ κοντά στο προηγούμενο. Στο σημείο αυτό που αλλάζει ο ρυθμός μείωσης του κόστους βρίσκεται ο βέλτιστος αριθμός συστάδων. Στο παρακάτω γράφημα αυτό γίνεται για k=4.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ



Διάγραμμα 1: Line Plot of Cost according to the number of Clusters

Το μειονέκτημα αυτής της μεθόδου είναι ότι δεν είναι πάντα εύκολο να αναγνωριστεί το σημείο αλλαγής του ρυθμού μείωσης του κόστους, καθώς και ότι σε ένα γράφημα μπορεί να έχουμε παραπάνω από ένα τέτοιο σημείο. Μια άλλη προσέγγιση που χρησιμοποιείται για τον καθορισμό του αριθμού των clusters είναι αυτή του Silhouette Score. Ουσιαστικά αναθέτει κάθε στοιχείο των δεδομένων στο cluster με τα περισσότερα κοινά χαρακτηριστικά, μετρώντας την ανομοιογένεια κάθε στοιχείου με τα υπόλοιπα στοιχεία (Hruschka & Covoes, 2005, Lleti et al, 2004). Σύμφωνα με τους Kaufman και Rousseeuw (1990), έστω ότι έχουμε ένα δεδομένο i που ανήκει στο cluster A. Η μέση ανομοιογένεια του στοιχείου i από όλα τα υπόλοιπα στοιχεία της συστάδας A ορίζεται ως a_i. Στην συνέχεια υπολογίζεται η μέση ανομοιογένεια του στοιχείου i από τα στοιχεία του cluster C, για κάθε C ≠ A ($d_{i,C}$) και επιλέγεται η μικρότερη από αυτές, δηλαδή b_i = min ($d_{i,C}$). Η τιμή αυτή αντιπροσωπεύει την ανομοιογένεια του i από τα γειτονικά του cluster και το silhouette score δίνεται από τον εξής τύπο:

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

Είναι σαφές ότι η τιμή κυμαίνεται από -1 έως 1 και όσο πιο υψηλή και θετική η τιμή, τόσο πιο καλά διαχωρισμένα είναι τα clusters.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Προκειμένου να μετρήσουμε την ομοιογένεια των στοιχείων των clusters, όπως έχει ήδη αναφερθεί, χρησιμοποιούνται δείκτες μέτρησης της απόστασης. Η πιο δημοφιλής μέτρηση που χρησιμοποιείται συνήθως στον αλγόριθμο k means είναι η Ευκλείδεια Απόσταση, η οποία ισούται με την τετραγωνική ρίζα της απόστασης κάθε στοιχείου του cluster από το κέντρο του (Singh A et al., 2013).

$$d_{xy} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})}$$

Επίσης, προτείνεται και η χρήση της απόστασης Manhattan, η οποία ισούται με την απόλυτη τιμή της απόστασης κάθε στοιχείου του cluster από το centroid της συστάδας (Singh A et al., 2013). Δηλαδή,

$$d_{xy} = | x_{ik} - x_{jk} |$$

Τέλος, ένας άλλος δείκτης είναι ο Davies – Bouldin Index, ο οποίος συνδυάζει την έξω-συσταδική απόσταση με την έσω-συσταδική απόσταση. Για να θεωρηθεί επιτυχημένη μια τμηματοποίηση, θα πρέπει η έσω-συσταδική απόσταση να είναι όσο το δυνατόν πιο μικρή και η έξω-συσταδική όσο το δυνατόν πιο μεγάλη, δηλαδή, τα στοιχεία του κάθε cluster να βρίσκονται όσο το δυνατόν πιο κοντά, να είναι όσο το δυνατόν πιο όμοια και κάθε cluster να απέχει, να διαφέρει όσο το δυνατόν περισσότερο από τα υπόλοιπα.

Κλείνοντας, να αναφέρουμε επίσης ότι ο αλγόριθμος k-means μπορεί να χρησιμοποιηθεί εξίσου και με κατηγορικές μεταβλητές, καθώς και σε text analytics. Αντιμετωπίζει την όλη διαδικασία της συσταδοποίησης σαν μια διαδικασία βελτιστοποίησης, η οποία έχει ως σκοπό να ομαδοποιήσει δεδομένα σε k clusters, έτσι ώστε να ικανοποιείται κάποιο κριτήριο μεγιστοποίησης ή ελαχιστοποίησης (Singh V.K. & et al, 2011). Κατά την διαδικασία του text clustering, συνήθως χρησιμοποιείται κάποιο διάνυσμα (vector), όπου ουσιαστικά μεταφέρεται το σύνολο του κειμένου. Δημιουργείται ένας πίνακας όπου κάθε γραμμή αντιπροσωπεύει ένα κείμενο και κάθε στήλη έναν όρο – λέξη του κειμένου. Πάνω σε αυτού του είδους τα δεδομένα είναι δυνατό να χρησιμοποιηθούν αλγόριθμοι όπως ο k-means και οι παραλλαγές του, καθώς και οι ιεραρχικού τύπου αλγόριθμοι συσταδοποίησης. Σύμφωνα με τους Jing & Huang (2005), σε πραγματικά δεδομένα οι λέξεις και ο αριθμός των κειμένων μπορεί να είναι αρκετά μεγάλος και αυτό έχει ως αποτέλεσμα αρκετά μεγάλους πίνακες. Ακόμη, τα κείμενα μπορεί να έχουν διαφορετικά υποσύνολα όρων και αυτό έχει ως

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

αποτέλεσμα έναν αρκετά αραιό πίνακα. Προκειμένου να μειωθούν οι όροι πάνω στους οποίους θα εφαρμοστεί ο αλγόριθμος πρέπει να γίνει κάποιου είδους επιλογή. Η τεχνική TF – IDF χρησιμοποιείται για τον περιορισμό των πιο κοινών λέξεων και την εξαγωγή των πιο σχετικών όρων του κειμένου (Forsati et al., 2013, Luo & Chung, 2009). Ο όρος TF_i (term frequency) αντιπροσωπεύει την συχνότητα του όρου i στο κείμενο, δηλαδή το πόσες φορές εμφανίζεται ο όρος i στο σύνολο των όρων του κειμένου. Απ' την άλλη ο όρος IDF (inverse document frequency) βασίζεται στην άποψη ότι οι λέξεις που εμφανίζονται πιο συχνά στο σύνολο των εγγράφων είναι μικρής σημασίας και γι' αυτό δεν πρέπει να τους δίνεται έμφαση. Ουσιαστικά υπολογίζεις στο σύνολο των εγγράφων N πόσα έγγραφα είναι αυτά που περιέχουν τον όρο i και το idf ισούται με τον λογάριθμο της διαίρεσης, N/d_i , ,όπου d_i είναι το σύνολο των εγγράφων που περιέχουν τον όρο i. Επομένως, προκειμένου να αποδυναμωθεί ο όρος i που εμφανίζεται συχνά σε πολλά έγγραφα θα πρέπει να πολλαπλασιαστούν αυτοί οι δύο όροι. Αυτός ο πίνακας που προκύπτει από αυτή την διαδικασία, είναι αυτός πάνω στον οποίο θα εφαρμοστεί ο αλγόριθμος clustering.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

3

Ο ΚΛΑΔΟΣ ΤΗΣ ΧΗΜΙΚΗΣ ΒΙΟΜΗΧΑΝΙΑΣ

Η χημική βιομηχανία αποτελεί έναν από τους μεγαλύτερους κλάδους της παγκόσμιας οικονομίας, καθώς απορροφά πρώτες ύλες και παράγει προϊόντα για ένα πλήθος εφαρμογών, που προορίζονται είτε για τον τελικό καταναλωτή, είτε για περαιτέρω επεξεργασία από άλλες επιχειρήσεις. Η πορεία της χημικής βιομηχανίας είναι στενά συνδεδεμένη με την εξέλιξη της παγκόσμιας οικονομίας και με τις τιμές των α' υλών που θεωρούνται χρηματιστηριακά είδη, όπως είναι το πετρέλαιο και το φυσικό αέριο. Επομένως επηρεάζεται έντονα από τις διακυμάνσεις τόσο της διεθνούς αγοράς όσο και των τιμών των βασικών α' υλών. Η παραγωγή των περισσότερων προϊόντων γίνεται σε μεγάλες ποσότητες ή απαιτεί ειδική τεχνογνωσία. Επομένως, η παραγωγική διαδικασία θα μπορούσαμε να πούμε ότι είναι εντάσεως κεφαλαίου και απαιτεί σημαντικές επενδύσεις.

Σύμφωνα με τον οργανισμό American Chemistry Council, ο κλάδος των χημικών περιλαμβάνει τους εξής πέντε τομείς: φαρμακευτικά, βασικά χημικά, ειδικά χημικά (specialty chemicals), αγροχημικά (λιπάσματα, φυτοφάρμακα και εντομοκτόνα) και καταναλωτικά χημικά (απορρυπαντικά, σαπούνια, καλλυντικά κτλ.). Στο σημείο αυτό να διευκρινίσουμε ότι με τον όρο βασικά χημικά χαρακτηρίζονται

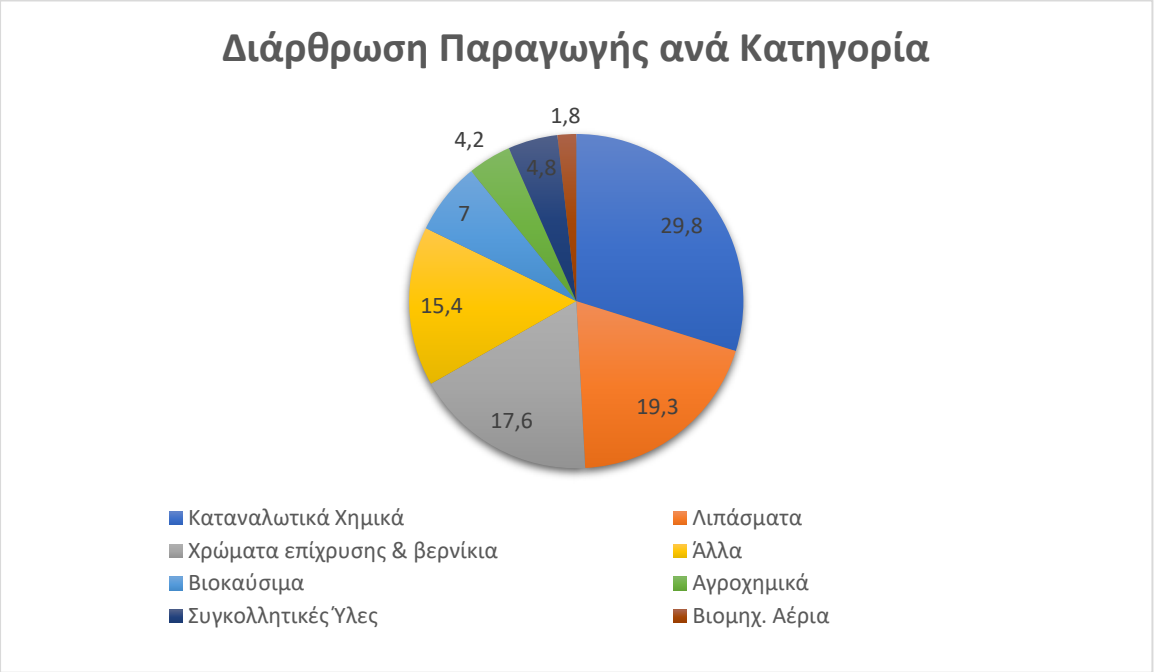
ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

τα προϊόντα εκείνα που παράγονται σε μεγάλες ποσότητες (bulk) και κάποια από αυτά είναι τα ανόργανα χημικά, τα πετροχημικά, τα οργανικά ενδιάμεσα χημικά, οι τεχνητές ίνες και οι χρωστικές. Ενώ με τον όρο ειδικά χημικά αναφερόμαστε σε εκείνα τα προϊόντα που παράγονται σε μικρές σχετικά ποσότητες αλλά προσδίδουν υψηλή προστιθέμενη αξία με βάση την αποτελεσματικότητα τους. Στην κατηγορία αυτή ανήκουν προϊόντα όπως τα αιθέρια έλαια, οι αρωματικές ύλες, τα χημικά νερού και λυμάτων και τα χημικά εξόρυξης πετρελαίου.

Η ζήτηση για χημικά προϊόντα B2B, δηλαδή για προϊόντα που διατίθενται στην βιομηχανία, με σκοπό την παραγωγή ή την περαιτέρω επεξεργασία τους πριν διατεθούν στον τελικό καταναλωτή, επηρεάζεται άμεσα από την ζήτηση των εταιρειών των επιμέρους κλάδων εφαρμογής στους οποίους απευθύνεται. Οι κλάδοι αυτοί περιλαμβάνουν τα πλαστικά προϊόντα και κυρίως υλικά συσκευασίας και σωληνώσεις, την χάρτινη συσκευασία, τα απορρυπαντικά και τα σαπούνια, τα καλλυντικά, τις γραφικές τέχνες, τον οικοδομικό και κατασκευαστικό κλάδο και τα αγροχημικά. Αναφορικά με τα αγροχημικά να σημειώσουμε ότι τα εγχώρια παραγόμενα λιπάσματα καλύπτουν περίπου το 55% των συνολικών πωλήσεων, ενώ το υπόλοιπο κομμάτι καλύπτεται από εισαγωγές. Ο κλάδος περιλαμβάνει τόσο παραγωγικές όσο και εμπορικές – εισαγωγικές επιχειρήσεις. Σε ορισμένες περιπτώσεις και προκειμένου οι επιχειρήσεις να καλύψουν μεγαλύτερη ποικιλία προϊόντων, συνδυάζουν την παραγωγική τους δραστηριότητα με εισαγωγές προϊόντων. Επίσης, πολλές πολυεθνικές επιλέγουν να διανέμουν τα προϊόντα τους στην Ελλάδα μέσω αντιπροσώπων. Με βάση στοιχεία της ΕΛΣΤΑΤ, το 2015 στην βιομηχανία των χημικών προϊόντων, δραστηριοποιούνταν 244 επιχειρήσεις και παρατηρήθηκε αύξηση στην παραγωγή κατά 3,8% σε σχέση με το προηγούμενο έτος.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Το 2015 εκτιμάται ότι στον κλάδο της χημικής βιομηχανίας απασχολούνταν περίπου 12.900 άνθρωποι, εκ των οποίων περίπου το 80% ήταν στους τομείς των ειδικών και καταναλωτικών χημικών. Το αντίστοιχο ποσοστό στην Ευρωπαϊκή Ένωση είναι 30,5%. Από άποψη διάρθρωσης παραγωγής, το μεγαλύτερο μέρος της απασχολείται στα καταναλωτικά χημικά, ενώ ακολουθεί ο κλάδος των λιπασμάτων.



Διάγραμμα 2: Διάρθρωση Παραγωγής της Χημικής Βιομηχανίας

Πηγή: ΕΛΣΤΑΤ (PRODCOM)

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Το χρονικό διάστημα 2006-2016, η εγχώρια αγορά των χημικών επηρεάστηκε έντονα από την κρίση, παρουσιάζοντας διακυμάνσεις, Για την ακρίβεια παρουσιάστηκε μείωση στο μέγεθος της αγοράς με ετήσιο ρυθμό πτώσης περίπου 5%. Τα πρώτα σημάδια ανάκαμψης παρουσιάστηκαν το 2014 και προς το παρών υπάρχει μια μικρή αύξηση στο μέγεθος της αγοράς. Το μεγαλύτερο τμήμα της εγχώριας αγοράς φαίνεται ότι καταλαμβάνουν τα ειδικά χημικά προϊόντα, τα οποία μαζί με τα καταναλωτικά χημικά προϊόντα καταλαμβάνουν το 55% της αγοράς, ενώ στα πετροχημικά – πολυμερή αντιστοιχεί το 45%. (Ίδρυμα οικονομικών & βιομηχανικών ερευνών)

Αναφορικά με την εξαγωγική δραστηριότητα της χώρας μας στον τομέα των χημικών προϊόντων, από τις αρχές του 2000 έως και το 2008 είχε παρουσιάσει αξιόλογη ανοδική πορεία. Όμως το 2009, εξαιτίας της διεθνούς οικονομικής κρίσης, οι εξαγωγές της Ελλάδας υποχώρησαν κατά 18%. Κατάφερε όμως να επανακάμψει σχετικά άμεσα, με τα σημερινά επίπεδα εξαγωγών να είναι υψηλότερα σε σχέση με τα αντίστοιχα του παρελθόντος. Τα χημικά προϊόντα αποτελούν μια από τις πιο σημαντικές κατηγορίες εξαγωγών της χώρας μας, καταλαμβάνοντας την 4^{η} θέση με ποσοστό 5% επί του συνόλου των εξαγωγών. (Ίδρυμα οικονομικών & βιομηχανικών ερευνών)

Από την άλλη οι εισαγωγές χημικών κατά το χρονικό διάστημα 2000-2008 παρουσίασαν σημαντική αύξηση, φτάνοντας στο υψηλότερο τους σημείο το 2008. Κατά την διάρκεια της κρίσης, το ποσοστό τους μειώθηκε κατά 20% και από τότε έχει παραμείνει σε σχετικά σταθερά επίπεδα. Συνολικά το εμπορικό ισοζύγιο της χώρας μας στο κλάδο των χημικών ουσιών είναι ελλειμματικό. Σύμφωνα με στοιχεία του 2017, υπάρχει αύξηση στις εξαγωγές της χώρας μας σε χημικά προϊόντα κατά 12%. Όμως αντίστοιχη αύξηση παρουσιάζουν και οι εισαγωγές κατά 6,5%. Γενικότερα θα μπορούσαμε να υποστηρίξουμε ότι τα τελευταία χρόνια οι εξαγωγές σε χημικά προϊόντα παρουσιάζουν αξιόλογη αύξηση, ισορροπώντας ως ένα βαθμό τις απώλειες από την εγχώρια αγορά, ενώ ταυτόχρονα οι αντίστοιχες εισαγωγές έχουν ελαφρώς περιοριστεί. Το γεγονός αυτό έχει σημαντική επίδραση στο εμπορικό ισοζύγιο, το οποίο όμως εξακολουθεί να είναι ελλειμματικό.

Στις κυριότερες χώρες προς τις οποίες εξάγει η Ελλάδα, βρίσκουμε τα Βαλκάνια, την Τουρκία και την Ιταλία αλλά και χώρες με ανεπτυγμένη βιομηχανία όπως η Γερμανία και η Ισπανία. Από την άλλη, το μεγαλύτερο μέρος των εισαγωγών προέρχεται από χώρες της Ευρωπαϊκής Ένωσης με ανεπτυγμένες χημικές βιομηχανίες, όπως η Γερμανία, η Ισπανία, η Γαλλία και η Μεγάλη Βρετανία.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Σύμφωνα με το Ίδρυμα οικονομικών & βιομηχανικών ερευνών, όσον αφορά την παγκόσμια αγορά των χημικών, το μεγαλύτερο μερίδιο το κατέχει η Κίνα με ποσοστό 39,6%. Ενώ ακολουθούν αρκετά πιο κάτω οι ΗΠΑ, η Γερμανία, η Ιαπωνία και η Νότια Κορέα. Σχετικά με την ευρωπαϊκή αγορά της χημικής βιομηχανίας δραστηριοποιούνται περίπου 30.000 επιχειρήσεις που απασχολούν περίπου 1.5εκ. προσωπικό. Η χημική βιομηχανία είναι ο πέμπτος μεγαλύτερος βιομηχανικός κλάδος στην Ευρώπη, με σημαντικότερο παίχτη την Γερμανία. Ιδιαίτερα αναπτυγμένος είναι ο κλάδος των χημικών και στην Γαλλία, στην Ιταλία, στην Ολλανδία, στην Ισπανία, στο Ηνωμένο Βασίλειο και στο Βέλγιο. Τα τελευταία χρόνια, η αύξηση της παραγωγής βασικών χημικών στην Ασία και στην Αμερική με χαμηλότερο κόστος παραγωγής, καθώς και η αυξημένη ζήτηση στις αναπτυσσόμενες οικονομίες της Ασίας και της Λατινικής Αμερικής, έχουν προκαλέσει μείωση του μεριδίου αγοράς της Ευρώπης. Παρόλα αυτά η διεθνής θέση της στον κλάδο της χημικής βιομηχανίας παραμένει αρκετά σταθερή και ισχυρή, λόγω των ευνοϊκών συνθηκών του περιβάλλοντος της, που διευκολύνει την ανάπτυξη καινοτομιών, της μεγάλης εξειδίκευσης του ανθρώπινου δυναμικού και της εξειδίκευσης στα ειδικά χημικά, καθώς και της ύπαρξης υποδομών που στηρίζουν την χημική βιομηχανία

Συνοψίζοντας θα μπορούσαμε να πούμε ότι ο κλάδος της χημικής βιομηχανίας είναι άμεσα συνδεδεμένος με την πορεία της ελληνικής οικονομίας και της ανάκαμψης της. Όπως είναι λογικό, η κρίση των τελευταίων χρόνων επηρέασε σημαντικά και τις πωλήσεις των βιομηχανικών κλάδων στους οποίους απευθύνονται τα χημικά προϊόντα και οι χημικές πρώτες ύλες. Μεγαλύτερες προοπτικές φαίνεται ότι έχουν οι παραγωγικές εταιρείες που πραγματοποιούν εξαγωγές καθώς και εμπορικές εταιρείες που διαθέτουν ευρεία γκάμα προϊόντων και άρα μπορούν να προσεγγίσουν περισσότερα τμήματα καταναλωτών.

Γενικότερα προβλέπεται ότι στα επόμενα χρόνια θα υπάρξει αύξηση του μεγέθους της αγοράς των χημικών, περίπου 2%. Για να μπορέσουν οι επιχειρήσεις να ανταποκριθούν καλύτερα στις αλλαγές του περιβάλλοντος τους θα ήταν σκόπιμο να φροντίσουν για την ενίσχυση της εξωστρέφειας τους και ταυτόχρονα να ενισχύσουν την γκάμα των προϊόντων τους με καινοτόμα προϊόντα υψηλής προστιθέμενης αξίας για τους πελάτες τους.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

4

ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΕΤΑΙΡΕΙΑΣ

Για την ανάλυση που θα ακολουθήσει, έχουν χρησιμοποιηθεί δεδομένα από μια ελληνική εταιρεία που δραστηριοποιείται στον κλάδο των χημικών. Πρόκειται για μια χημική βιομηχανία με έντονη παρουσία τόσο στην εγχώρια όσο και στην διεθνή αγορά. Η εταιρεία ιδρύθηκε το 1946 ως εργαστήριο για την μελέτη των φυσικών προϊόντων, δίνοντας ιδιαίτερη σημασία στην έρευνα πάνω στις παραγωγικές διαδικασίες και στις εφαρμογές των αιθέριων ελαίων. Μετά από περίπου δέκα πέντε χρόνια, δημιουργούνται οι μονάδες παραγωγής βιομηχανικών αρωμάτων και αρωμάτων τροφίμων και στην πορεία, με διάφορες επενδύσεις κυρίως σε τεχνολογικό εξοπλισμό, η δραστηριότητα της επεκτάθηκε σε διάφορους τομείς. Αυτή την στιγμή η εταιρεία αναπτύσσει δράση στους εξής: Βιομηχανικά Αρώματα, Αρώματα Τροφίμων, Πρώτες Ύλες Αρωματοποιίας, Προϊόντα Θρέψης Φυτών και Φυτοπροστασίας και Απωθητικά Εντόμων (Καταναλωτικά Προϊόντα).

Η υπό εξέταση εταιρεία έχει καταφέρει να εδραιώσει την θέση της στην παγκόσμια αγορά και αυτή την στιγμή η δράση της επεκτείνεται εκτός της Ευρώπης, στην Μέση Ανατολή, στην Αφρική, στην Άπω Ανατολή και στην Βόρεια Αμερική. Περίπου το 50% του κύκλου εργασιών της προέρχεται από

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

τις εξαγωγές, κυρίως αρωμάτων και ειδικών πρώτων υλών, ενώ στοχεύει στην ανάπτυξη του κύκλου εργασιών της μέσω της επέκτασης σε νέες αγορές, με την βοήθεια των αντιπροσώπων της.

Σπουδαίο ρόλο στην επιτυχία της εταιρείας παίζει η φιλοσοφία και η κουλτούρα που την διαπνέει. Ιδιαίτερη έμφαση δίνεται στον πελάτη, στις ανάγκες του και στην ικανοποίηση αυτών με το καλύτερο δυνατό τρόπο. Γι' αυτό τον λόγο η εταιρεία φροντίζει να δημιουργεί προϊόντα καινοτόμα, ασφαλή και υψηλής ποιότητας που να ανταποκρίνονται στις απαιτήσεις των πελατών της, ενώ ταυτόχρονα φροντίζει να παρέχει υπηρεσίες που να καλύπτουν μια μεγάλη γκάμα αιτημάτων τους. Σημαντικό ρόλο στην επίτευξη αυτού του στόχου παίζει η Έρευνα και η Ανάπτυξη. Η εταιρεία φροντίζει και επενδύει στην επιστημονική έρευνα και στην καινοτομία, προκειμένου όχι μόνο να παραμένει ενημερωμένη για τις σύγχρονες εξελίξεις αλλά και να τις δημιουργεί. Με αυτό τον τρόπο, καταφέρνει να είναι σε θέση να προτείνει εναλλακτικές και λύσεις στους πελάτες της, που να τους ικανοποιούν με τον καλύτερο δυνατό τρόπο. Η έρευνα επικεντρώνεται στην σύνθεση νέων χημικών ενώσεων, στη βελτιστοποίηση των παραγωγικών διαδικασιών, στη μελέτη των φυσικών προϊόντων, στην απομόνωση και την εφαρμογή χημικών ενώσεων υψηλής προστιθέμενης αξίας για τον κλάδο της αρωματοποιίας ή και άλλους βιομηχανικούς τομείς, ενώ ακόμη ιδιαίτερη σημασία δίνεται στην ανάπτυξη προϊόντων θρέψης και προστασίας των φυτών, μέσω της διερεύνησης της φυσιολογίας τους. Τέλος, το ανθρώπινο δυναμικό της εταιρείας αποτελεί βασικό πυλώνα για την επιτυχία της. Ιδιαίτερη σημασία δίνεται στην προσωπική ανάπτυξη και εξέλιξη των εργαζομένων της, με βάση τις ικανότητες και τα προσόντα τους, ενώ ταυτόχρονα φροντίζει να τους παρέχει όσο το δυνατόν καλύτερες συνθήκες εργασίας. Ακόμα, ενισχύει το ομαδικό πνεύμα, την συνεργασία, την αμοιβαία εμπιστοσύνη και την υπευθυνότητα ανάμεσα στους εργαζομένους της, συμβάλλοντας στην επιτυχία τόσο των ίδιων όσο και της επιχείρησης.

Όπως αναφέρθηκε και παραπάνω, η δράση της επιχείρησης επεκτείνεται σε διάφορους τομείς. Πιο συγκεκριμένα, αναφορικά με τον κλάδο των αρωμάτων, η εταιρεία φροντίζει να σχεδιάζει και να προμηθεύει τους πελάτης της με αρώματα που έχουν εφαρμογή σε ένα ευρύ φάσμα προϊόντων, όπως είναι οι κολόνιες, τα προϊόντα περιποίησης μαλλιών και σώματος, τα προϊόντα καθαρισμού σπιτιού και φροντίδας ρούχων. Αντιλαμβανόμενη την τάση της αγοράς για οικολογικά προϊόντα και φυσικά καλλυντικά, έχει αναπτύξει τα τελευταία χρόνια μια γκάμα προϊόντων φιλικών ως προς το

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

περιβάλλον. Ακόμη, συγγενείς κλάδος με αυτόν των αρωμάτων (των βιομηχανικών) θα μπορούσαμε να πούμε πως είναι αυτός των αρωμάτων τροφίμων. Η επιχείρηση φροντίζει και δημιουργεί αρώματα για διάφορους τύπους τροφίμων, όπως τα ποτά, τα είδη ζαχαροπλαστικής και αρτοποιίας, τα σνακ και τα καρυκεύματα.

Η δραστηριοποίηση της στον κλάδο των αρωμάτων γενικότερα, της έδωσε το πάτημα ώστε να μπορέσει να επεκταθεί και στην παραγωγή Α' υλών. Η επιχείρηση έχει την δυνατότητα πλέον να παράγει μια μεγάλη ποικιλία φυσικών και συνθετικών συστατικών. Στην γκάμα των προϊόντων της περιλαμβάνονται φυσικά φουρανοειδή, πυραζίνες, ακόρεστα οξέα και αλδεΰδες. Αρκετά διακριτός τομέας δραστηριοποίησης θα μπορούσαμε να υποστηρίξουμε ότι είναι αυτός της Υγείας των Φυτών. Εδώ εντοπίζουμε προϊόντα που σχετίζονται με την θρέψη και την προστασία των φυτών από επιβλαβή έντομα, κατάλληλων για εφαρμογή στην συμβατική αλλά και στην βιολογική καλλιέργεια. Ακόμη διαθέτει προϊόντα για την φροντίδα και την προστασία των φυτών κήπου, φυτών εσωτερικού χώρου, λουλουδιών και καλλωπιστικών φυτών. Τέλος, τα τελευταία χρόνια έχει επεκτείνει την δραστηριότητα της σε έναν ακόμα τομέα, αυτόν των εντομοαπωθητικών προϊόντων. Τα προϊόντα αυτά ανήκουν στην κατηγορία των βιοκτόνων και ήδη εξάγονται και σε άλλες ευρωπαϊκές χώρες, εκτός της Ελλάδος. Πρόκειται για προϊόντα φυσικής εντομοαπωθητικής προστασίας, τόσο για τον άνθρωπο όσο και για τα κατοικίδια ζώα (γάτα και σκύλος).

Από όλα τα παραπάνω γίνεται φανερό ότι πρόκειται για μια εταιρεία που έχει καταφέρει να αναπτυχθεί σε διάφορους τομείς, προσεγγίζοντας εν μέρει και εντελώς διαφορετικά τμήματα πελατών, με διαφορετικές ανάγκες και απαιτήσεις. Ως εκ τούτου γίνεται ακόμα πιο επιτακτική η ανάγκη του σωστού εντοπισμού των τμημάτων της αγοράς, στα οποία απευθύνεται η επιχείρηση. Με αυτόν τον τρόπο θα είναι σε θέση να εντοπίσει τις ανάγκες κάθε τμήματος και να εφαρμόσει τις κατάλληλες στρατηγικές μάρκετινγκ και πωλήσεων, προκειμένου να ικανοποιήσει τους πελάτες της με τον καλύτερο δυνατό τρόπο, μετατρέποντας τους σε loyal customers.

Στην συνέχεια και για να προχωρήσουμε με την ανάλυση μας, έχουν χρησιμοποιηθεί δεδομένα πωλήσεων δύο ετών από την συγκεκριμένη εταιρεία. Αρχικά θα προχωρήσουμε με μια ανάλυση των δεδομένων μας και έπειτα θα ακολουθήσει η τμηματοποίηση των πελατών και η ανάλυση RFM.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

5

ΑΝΑΛΥΣΗ DATASET ΚΑΙ ΥΛΟΠΟΙΗΣΗ CLUSTERING

5.1 ΠΕΡΙΓΡΑΦΗ DATASET

Προκειμένου να δείξουμε την χρησιμότητα του customer segmentation και τον τρόπο με τον οποίο μπορεί να χρησιμοποιηθεί αυτή η τεχνική για την καλύτερη προσέγγιση της πελατειακής βάσης, έχουν συλλεχθεί δεδομένα πωλήσεων από την χημική βιομηχανία που περιγράψαμε παραπάνω, που όπως ήδη έχουμε δει, δραστηριοποιείται σε αρκετά διακριτούς τομείς.

Τα δεδομένα αυτά αφορούν χρονικό διάστημα περίπου δύο ετών, από 01/02/2017 έως 31/12/2018 και περιλαμβάνουν τα εξής στοιχεία:

- ✓ Salesman: ο πωλητής στον οποίο ανήκει ο συγκεκριμένος πελάτης
- ✓ Customer Type: η μεταβλητή αυτή δηλώνει εάν πρόκειται για πελάτη εσωτερικού ή εξωτερικού

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

- ✓ Country: η χώρα στην οποία βρίσκεται ο πελάτης
- ✓ Customer Code: ο μοναδικός κωδικός του πελάτη
- ✓ Customer Name (Hash)
- ✓ Item Code: κάθε προϊόν χαρακτηρίζεται από ένα μοναδικό 8-ψήφιο κωδικό
- ✓ Item Name (Hash)
- ✓ Sales Category I: η μεταβλητή αυτή δηλώνει την προϊοντική κατηγορία στην οποία ανήκει το κάθε προϊόν, πχ. αρώματα, αγροχημικά αγοραζόμενα. Παρατηρούμε ότι η συγκεκριμένη μεταβλητή περιλαμβάνει 8 μοναδικές τιμές και αυτές είναι: Αρώματα, Αρώματα Τροφίμων, Υλικά Συσκευασίας, Παραγόμενες Α' ύλες, Αγοραζόμενες Α' ύλες, Αγροχημικά Παραγόμενα, Αγροχημικά Αγοραζόμενα, Καταναλωτικά Αγαθά.
- ✓ Invoice: το νούμερο του τιμολογίου
- ✓ Invoice Type: η μεταβλητή αυτή δείχνει εάν πρόκειται για τιμολόγιο εσωτερικού ή εξωτερικού
- ✓ Invoice Date
- ✓ Invoice Year
- ✓ Quantity: η ποσότητα πώλησης
- ✓ Amount: συνολική αξία πώλησης

Για λόγους προστασίας των προσωπικών δεδομένων, έχουν αντικατασταθεί τόσο τα ονόματα των πελατών όσο και των προϊόντων με χρήση της συνάρτησης hash. Κάθε μοναδικό στοιχείο των μεταβλητών αυτών αντιστοιχεί σε μια μοναδική hash τιμή.

Οι μεταβλητές μας διακρίνονται σε κατηγορικές και αριθμητικές. Αριθμητικές μεταβλητές είναι οι εξής: Customer Code, Item Code, Invoice, Invoice Year, Quantity και Amount, ενώ κατηγορικές είναι οι Salesman, Customer Type, Country, Customer Name (Hash), Item Name (Hash), Sales Category I, Invoice Type και Invoice Date.

Παρακάτω παρουσιάζουμε ένα preview των πρώτων δέκα γραμμών του dataset, καθώς και κάποιες πληροφορίες για τα δεδομένα μας.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

	Salesman	CustomerType	Country	CustomerCode	CustomerNameHash	ItemCode	ItemNameHash	SalesCatI	Invoice	InvoiceType	InvoiceDate	InvoiceYear	Quantity	Ammount
0	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23080900	0x37C534BB3E38D7B2336E8E9B17AE9858	ΑΡΩΜΑΤΑ	2035955	ΤΞ	21-06-2017	2017	20.0	354.0
1	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23090700	0xB191B1BEF3CD945CE1AEA6BC49F83584	ΑΡΩΜΑΤΑ	2033017	ΤΞ	26-04-2017	2017	20.0	364.0
2	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23090700	0xB191B1BEF3CD945CE1AEA6BC49F83584	ΑΡΩΜΑΤΑ	2046636	ΤΞ	19-06-2018	2018	20.0	382.0
3	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23090700	0xB191B1BEF3CD945CE1AEA6BC49F83584	ΑΡΩΜΑΤΑ	2039089	ΤΞ	29-09-2017	2017	20.0	364.0
4	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23097300	0x52870B1D8567B91BE9CE50F97E6474D3	ΑΡΩΜΑΤΑ	2040290	ΤΞ	17-11-2017	2017	20.0	212.0
5	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23097300	0x52870B1D8567B91BE9CE50F97E6474D3	ΑΡΩΜΑΤΑ	2033017	ΤΞ	26-04-2017	2017	20.0	212.0
6	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23099000	0x2F8068DF87E754431EA38012C5E7229E	ΑΡΩΜΑΤΑ	2035955	ΤΞ	21-06-2017	2017	40.0	1904.0
7	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23099000	0x2F8068DF87E754431EA38012C5E7229E	ΑΡΩΜΑΤΑ	2041251	ΤΞ	12-01-2018	2018	20.0	952.0
8	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23099000	0x2F8068DF87E754431EA38012C5E7229E	ΑΡΩΜΑΤΑ	2050952	ΤΞ	02-11-2018	2018	20.0	952.0
9	NO SALESMAN	ΕΞ	ΑΓΓΛΙΑ	36282	0x851089DA90C57B2E36383E216CB835FC	23099000	0x2F8068DF87E754431EA38012C5E7229E	ΑΡΩΜΑΤΑ	2039089	ΤΞ	29-09-2017	2017	20.0	952.0

Πίνακας 1, Πρώτες δέκα γραμμές του dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78890 entries, 0 to 78889
Data columns (total 14 columns):
Salesman            78890 non-null object
CustomerType        78890 non-null object
Country             78890 non-null object
CustomerCode        78890 non-null int64
CustomerNameHash    78890 non-null object
ItemCode            78890 non-null int64
ItemNameHash        78890 non-null object
SalesCatI           78890 non-null object
Invoice             78890 non-null int64
InvoiceType         78890 non-null object
InvoiceDate         78890 non-null object
InvoiceYear         78890 non-null int64
Quantity            78890 non-null float64
Ammount             78890 non-null float64
dtypes: float64(2), int64(4), object(8)
memory usage: 8.4+ MB
```

Εικόνα 1, Πληροφορίες για τα δεδομένα του dataset

Όπως βλέπουμε το dataset περιλαμβάνει 78890 γραμμές και δεν υπάρχουν καθόλου null τιμές. Κάθε γραμμή αντιστοιχεί σε ένα προϊόν που έχει πωληθεί, χωρίς όμως να είναι ομαδοποιημένα ανά τιμολόγιο. Επομένως, ένα συγκεκριμένο νούμερο τιμολογίου θα εμφανίζεται τόσες φορές στο αρχείο μας όσες και τα προϊόντα που περιέχει.

Στη συνέχεια και για να καταλάβουμε καλύτερα τόσο το dataset όσο και την υπό εξέταση εταιρεία, προχωρήσαμε με κάποιες οπτικοποιήσεις των δεδομένων μας. Για την ανάλυση μας χρησιμοποιήθηκε η διανομή της Python από την Anaconda και πιο συγκεκριμένα το JupyterLab.

Για αρχή επιλέξαμε να χωρίσουμε το dataset μας σε δύο επιμέρους, με βάση το έτος όπου πραγματοποιήθηκε η εκάστοτε συναλλαγή. Το dataset για το έτος 2017 περιλαμβάνει 38507 γραμμές ενώ αυτό του 2018 περιλαμβάνει 40383 γραμμές. Ακόμη καταμετρώντας τον μοναδικό αριθμό των μεταβλητών Country, ItemCode και CustomerCode, παρατηρούμε ότι στα στοιχεία του έτους 2017

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

περιλαμβάνονται 47 χώρες, 1625 πελάτες και 5145 προϊόντα, και αντίστοιχα στα δεδομένα του έτους 2018 περιλαμβάνονται 44 χώρες, 1628 πελάτες και 5339 προϊόντα. Παρακάτω παραθέτουμε σχετικό πίνακα.

	2017	**2018**
Country	47	44
CustomerCode	1625	1628
ItemCode	5145	5339

Πίνακας 2, Σύνολο Χωρών, Πελατών και Προϊόντων για το 2017 και 2018

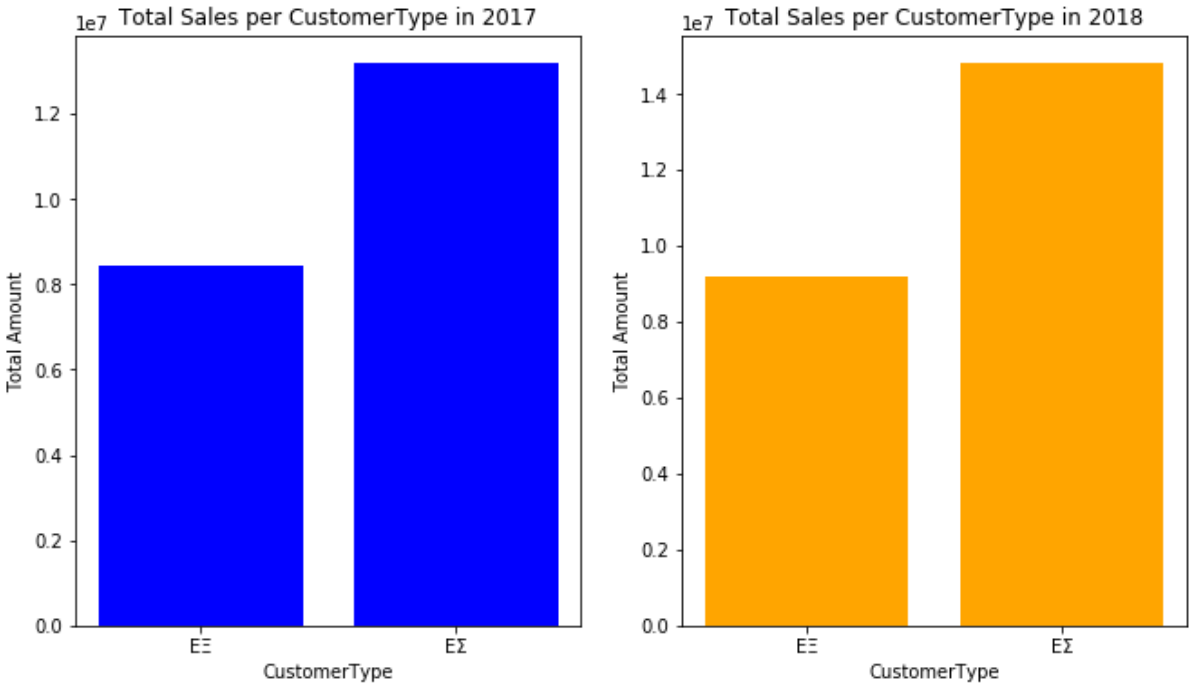
Βλέπουμε ότι παρόλο που υπάρχει μια μικρή μείωση (6,38%) στις χώρες όπου δραστηριοποιήθηκε η εταιρεία το 2018 σε σχέση με το 2017, έχουμε αύξηση στο πλήθος των πελατών κατά 3,77% ενώ ο αριθμός των διαφορετικών προϊόντων που πουλήθηκαν παρέμεινε σχετικά σταθερός (αύξηση κατά 0,18%). Τέλος, οι συνολικές πωλήσεις για το 2018 ήταν υψηλότερες συγκριτικά με αυτές του 2017 κατά 10.9%

```
InvoiceYear
2017    21610938.78
2018    23964279.58
Name: Ammount, dtype: float64
```

Εικόνα 2, Συνολικές Πωλήσεις 2017 και 2018

Εφόσον πρόκειται για μια εταιρεία με έντονη παρουσία στις διεθνείς αγορές, επιλέξαμε να δούμε τις συνολικές πωλήσεις που πραγματοποιήθηκαν στην εγχώρια αγορά σε σχέση με αυτές στις αγορές του εξωτερικού. Η μεταβλητή CustomerType παίρνει την τιμή ΕΣ για πελάτες του εσωτερικού και ΕΞ για πελάτες του εξωτερικού. Στα παρακάτω bar charts βλέπουμε το ύψος των συνολικών πωλήσεων ανά κατηγορία πελάτη.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ



Διάγραμμα 3, Συνολικές Πωλήσεις ανά Τύπο Πελάτη για τα έτη 2017 και 2018

Παρατηρούμε ότι η πλειοψηφία των εσόδων της εταιρείας προέρχεται από την ελληνική αγορά. Όμως, δεν γίνεται να μην αναφέρουμε ότι η εταιρεία παρουσιάζει και έντονη εξαγωγική δραστηριότητα, το ύψος της οποίας ξεπερνά το 50% των πωλήσεων στην Ελλάδα. Οι πωλήσεις στο εξωτερικό για τα έτη 2017 και 2018 αγγίζουν σχεδόν το 40% του συνολικού κύκλου εργασιών της, όπως φαίνεται και στα παρακάτω donut charts.



Διάγραμμα 4, Ποσοστό πωλήσεων σε εσωτερικό και εξωτερικό για τα έτη 2017 και 2018

Στην συνέχεια επιλέξαμε να δούμε το ύψος των πωλήσεων ανά χώρα για τα δύο έτη. Ταυτόχρονα υπολογίσαμε για κάθε χώρα το ποσοστό των πωλήσεων της στο σύνολο του κύκλου εργασιών. Παρακάτω παρουσιάσουμε σχετικούς πίνακες με τις πωλήσεις ανά χώρα ταξινομημένες κατά φθίνουσα σειρά, καθώς και αντίστοιχα bar charts.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

2017		
Country	**Amount**	**Percentage**
Ελλάδα	12985978.40	60.09
Νιγηρία	3002898.00	13.90
Ελβετία	1775744.65	8.22
Σερβία	382694.00	1.77
Ρωσία	366533.40	1.70
Αλβανία	265461.00	1.23
Ρουμανία	262703.70	1.22
Κύπρος	248555.80	1.15
Ιταλία	224258.10	1.04
Ν. Κορέα	224056.77	1.04
Βουλγαρία	212690.10	0.98
Γερμανία	208308.30	0.96
Τουρκία	204477.90	0.95
Ελλάδα (EU)	186503.40	0.86
Αγγλία	183580.10	0.85
Πολωνία	154629.60	0.72
Ταϊλάνδη	105234.14	0.49
Γαλλία	82917.50	0.38
Συρία	82162.50	0.38
Ινδονησία	77431.45	0.36
Αίγυπτος	66825.22	0.31
Ταιβάν	52908.82	0.24
Σαουδική Αραβία	37762.51	0.17
Αυστρία	33511.80	0.16
Ισραήλ	22470.60	0.10
Λίβανος	21659.08	0.10
Σιγκαπούρη	20115.40	0.09
ΗΠΑ	16670.00	0.08
Ν/Α	16397.80	0.08
Κροατία	15892.20	0.07
Βέλγιο	15230.00	0.07
FYROM	14485.50	0.07
Μαλαισία	13155.21	0.06
Ουγγαρία	5417.80	0.03
Τσεχία	3807.50	0.02
Ισπανία	3197.40	0.01
Ομάν	2700.00	0.01
Βραζιλία	2053.97	0.01
Γεωργία	1948.00	0.01
Ιρλανδία	1270.00	0.01
Ινδία	1417.00	0.01
Ιορδανία	1270.00	0.01
Κίνα	1225.00	0.01
Ντουμπάι	1149.16	0.01
Σουηδία	593.00	0.00
Κουβέιτ	275.00	0.00
Μπαγκλαντές	200.00	0.00

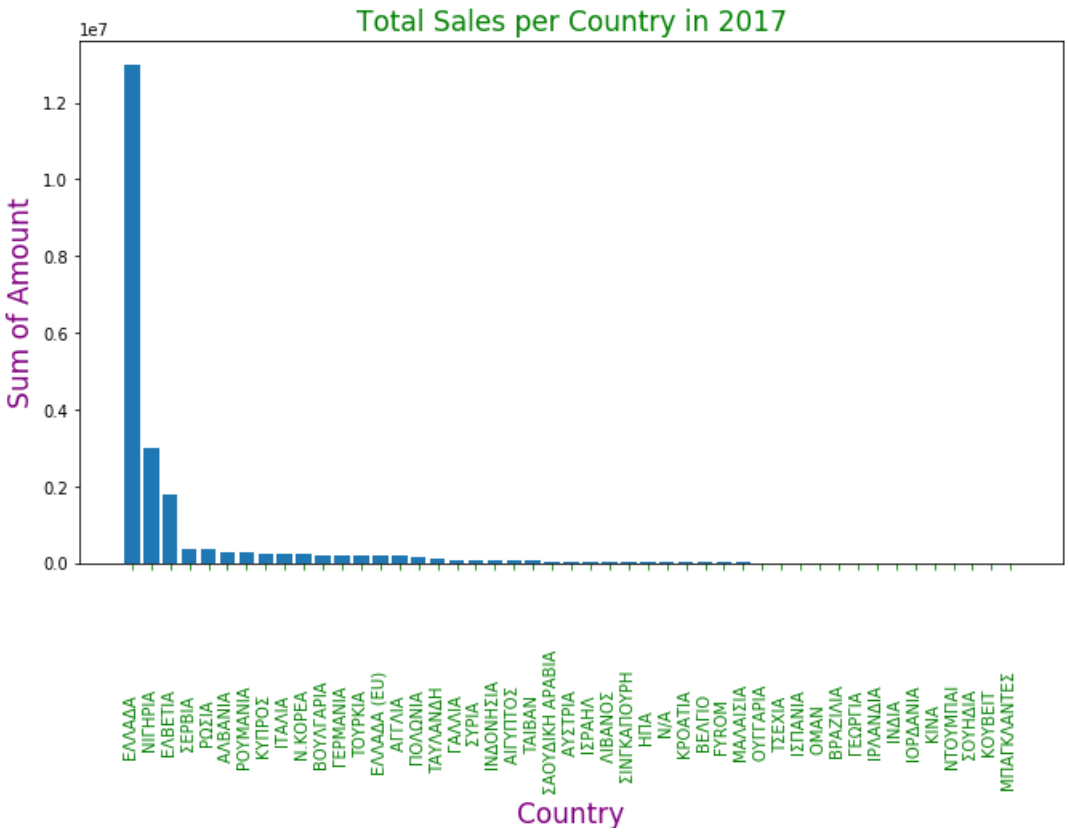
2018		
Country	**Amount**	**Percentage**
Ελλάδα	14697499.14	61.33
Νιγηρία	3200210.82	13.35
Ελβετία	1738349.67	7.25
Γερμανία	483974.80	2.02
Ρωσία	472961.00	1.97
Πολωνία	334429.65	1.40
Κύπρος	320233.98	1.34
Ν. Κορέα	299173.50	1.25
Σερβία	280608.50	1.17
Βουλγαρία	279963.30	1.17
Αγγλία	273002.00	1.14
Ρουμανία	257067.60	1.07
Ιταλία	213561.00	0.89
ΗΠΑ	162594.00	0.68
Τουρκία	116114.00	0.48
Αίγυπτος	99032.83	0.41
Συρία	92568.75	0.39
Ταϊλάνδη	90880,00	0.38
Ελλάδα (EU)	74484.50	0.31
Ταιβάν	70431.93	0.29
Ινδονησία	69020.90	0.29
Ισραήλ	48389.85	0.20
Λίβανος	45115.94	0.19
Ν/Α	38867.22	0.16
Αυστρία	38011.20	0.16
Γαλλία	34005.50	0.14
Αλβανία	30668.00	0.13
Ιορδανία	29540.00	0.12
Σουδάν	13610.00	0.06
FYROM	10936.40	0.05
Ισπανία	8357.50	0.03
Κροατία	7636.60	0.03
Ντουμπάι	7289.78	0.03
Ουγγαρία	5876.80	0.02
Σιγκαπούρη	5511.39	0.02
Τσεχία	2901.00	0.01
Γεωργία	2387.00	0.01
Ινδία	2021.50	0.01
Ιρλανδία	1782.00	0.01
Σουηδία	1444.00	0.01
Αυστραλία	1300.00	0.01
Βραζιλία	1001.21	0.00
Ουκρανία	970.00	0.00
Μαλαισία	494.82	0.00

.

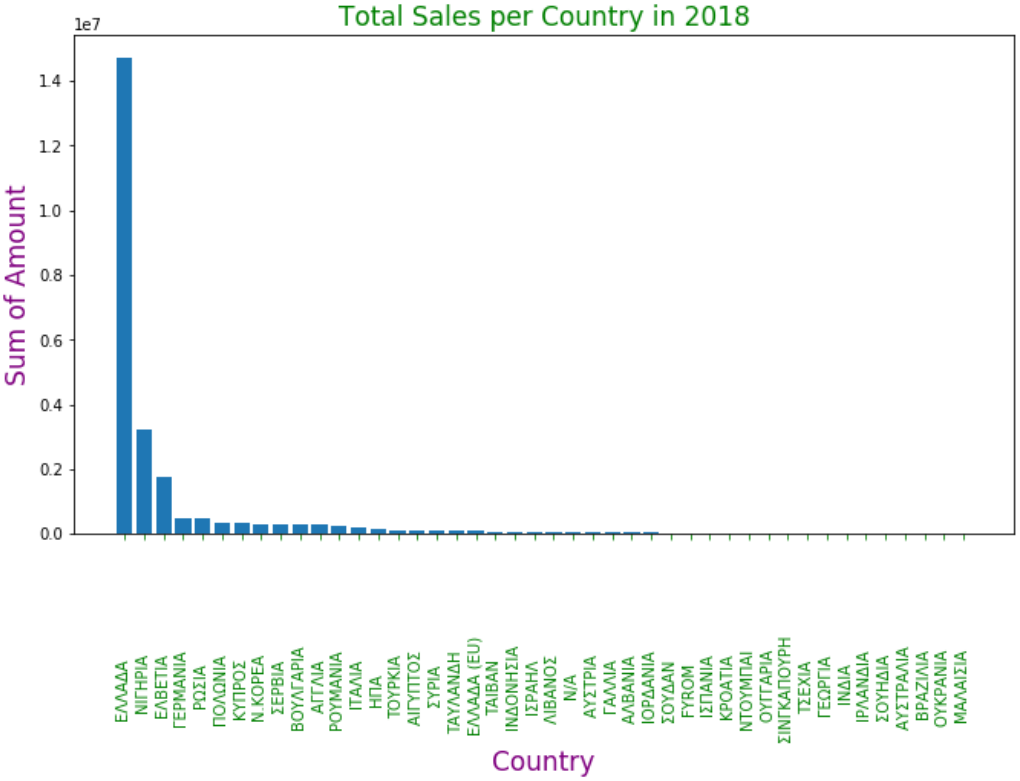
Πίνακας 3, Συνολικές πωλήσεις και ποσοστό πωλήσεων ανά χώρα, για το 2017 και 2018

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

BAR CHARTS



Διάγραμμα 5, Συνολικές Πωλήσεις ανά Χώρα για το 2017



Διάγραμμα 6, Συνολικές Πωλήσεις ανά Χώρα για το 2018

Όπως βλέπουμε, οι υψηλότερες πωλήσεις της εταιρείας μετά την Ελλάδα, όπου βρίσκεται και η έδρα της επιχείρησης εντοπίζονται στην Νιγηρία και στην Ελβετία, ενώ ακολουθούν για το 2017 η Σερβία και η Ρωσία και για το 2018 η Γερμανία και η Ρωσία. Αξιοσημείωτο είναι ότι η διαφορά της τρίτης χώρας σε πωλήσεις (Ελβετία) από την τέταρτη ξεπερνά τόσο για το 2017 όσο και για το 2018, σε απόλυτα νούμερα το 1.000.000,00€, περίπου 5%-6%.

Ακόμη, παρατηρούμε ότι σχεδόν το 80% του κύκλου εργασιών της προέρχεται από την Ελλάδα, την Νιγηρία και την Ελβετία. Το γεγονός αυτό μας κάνει να υποπτευθούμε ότι στις χώρες αυτές είτε υπάρχουν αρκετοί πελάτες είτε υπάρχουν σημαντικοί πελάτες στους οποίους αντιστοιχούν μεγάλα ποσοστά του τζίρου της εταιρείας. Για να δούμε το πλήθος των πελατών που εντοπίζονται σε κάθε χώρα, κάναμε count στον μοναδικό αριθμό των πελατών εφαρμόζοντας group by country στο dataset και επιλέξαμε να δούμε το πλήθος των πελατών γι' αυτές τις τρεις χώρες. Παρακάτω βλέπουμε το σχετικό output.

Country	CustomerCode
ΕΛΛΑΔΑ	1753
ΝΙΓΗΡΙΑ	3
ΕΛΒΕΤΙΑ	2

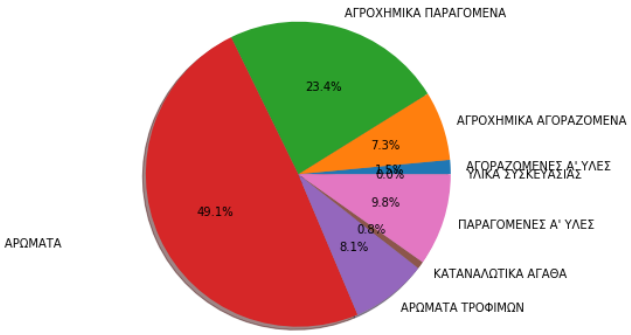
Πίνακας 4, Πλήθος Πελατών

Όπως ήταν αναμενόμενο στην Ελλάδα συγκεντρώνεται ο υψηλότερος αριθμός πελατών, ενώ στην Νιγηρία και στην Ελβετία μόλις 3 και 2 αντίστοιχα. Αυτό μας κάνει να συμπεραίνουμε ότι στις δύο αυτές χώρες συναντάμε σημαντικούς πελάτες για την εταιρεία, οι οποίοι δαπανούν μεγάλα ποσά για να αγοράσουν τα προϊόντα της.

Εφόσον η εταιρεία δεν δραστηριοποιείται μόνο σε έναν τομέα, ενδιαφέρον θα είχε να δούμε το ποσοστό του κύκλου εργασιών που αποφέρει κάθε κλάδος δραστηριοποίησης της. Για να γίνει αυτό, υπολογίσαμε το σύνολο του Amount ανά κατηγορία πώλησης, τόσο για το 2017 όσο και για το 2018 και τα οπτικοποιήσαμε ως ποσοστό σε δύο pie charts.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

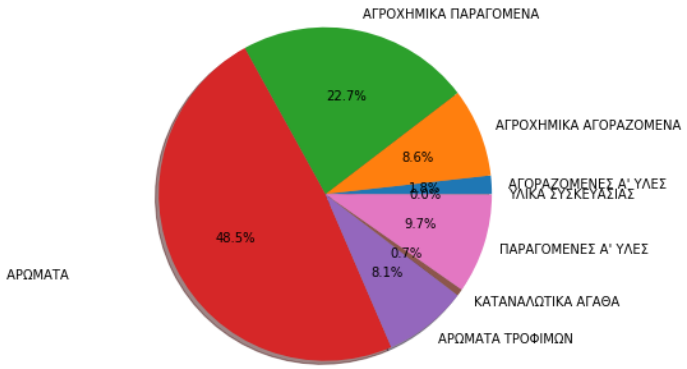
Για το 2017:



SalesCatI	Ammount
ΑΓΟΡΑΖΟΜΕΝΕΣ Α' ΥΛΕΣ	324995.29
ΑΓΡΟΧΗΜΙΚΑ ΑΓΟΡΑΖΟΜΕΝΑ	1587605.30
ΑΓΡΟΧΗΜΙΚΑ ΠΑΡΑΓΟΜΕΝΑ	5059763.09
ΑΡΩΜΑΤΑ	10611434.45
ΑΡΩΜΑΤΑ ΤΡΟΦΙΜΩΝ	1743979.83
ΚΑΤΑΝΑΛΩΤΙΚΑ ΑΓΑΘΑ	166131.95
ΠΑΡΑΓΟΜΕΝΕΣ Α' ΥΛΕΣ	2115184.96
ΥΛΙΚΑ ΣΥΣΚΕΥΑΣΙΑΣ	1843.91

Διάγραμμα 7, Ποσοστό του κύκλου εργασιών ανά κατηγορία Πώλησης

Για το 2018:



SalesCatI	Ammount
ΑΓΟΡΑΖΟΜΕΝΕΣ Α' ΥΛΕΣ	428298.96
ΑΓΡΟΧΗΜΙΚΑ ΑΓΟΡΑΖΟΜΕΝΑ	2059200.40
ΑΓΡΟΧΗΜΙΚΑ ΠΑΡΑΓΟΜΕΝΑ	5428966.47
ΑΡΩΜΑΤΑ	11610915.43
ΑΡΩΜΑΤΑ ΤΡΟΦΙΜΩΝ	1950216.98
ΚΑΤΑΝΑΛΩΤΙΚΑ ΑΓΑΘΑ	158323.12
ΠΑΡΑΓΟΜΕΝΕΣ Α' ΥΛΕΣ	2321071.65
ΥΛΙΚΑ ΣΥΣΚΕΥΑΣΙΑΣ	7286.57

Διάγραμμα 8, Ποσοστό του κύκλου εργασιών ανά κατηγορία Πώλησης

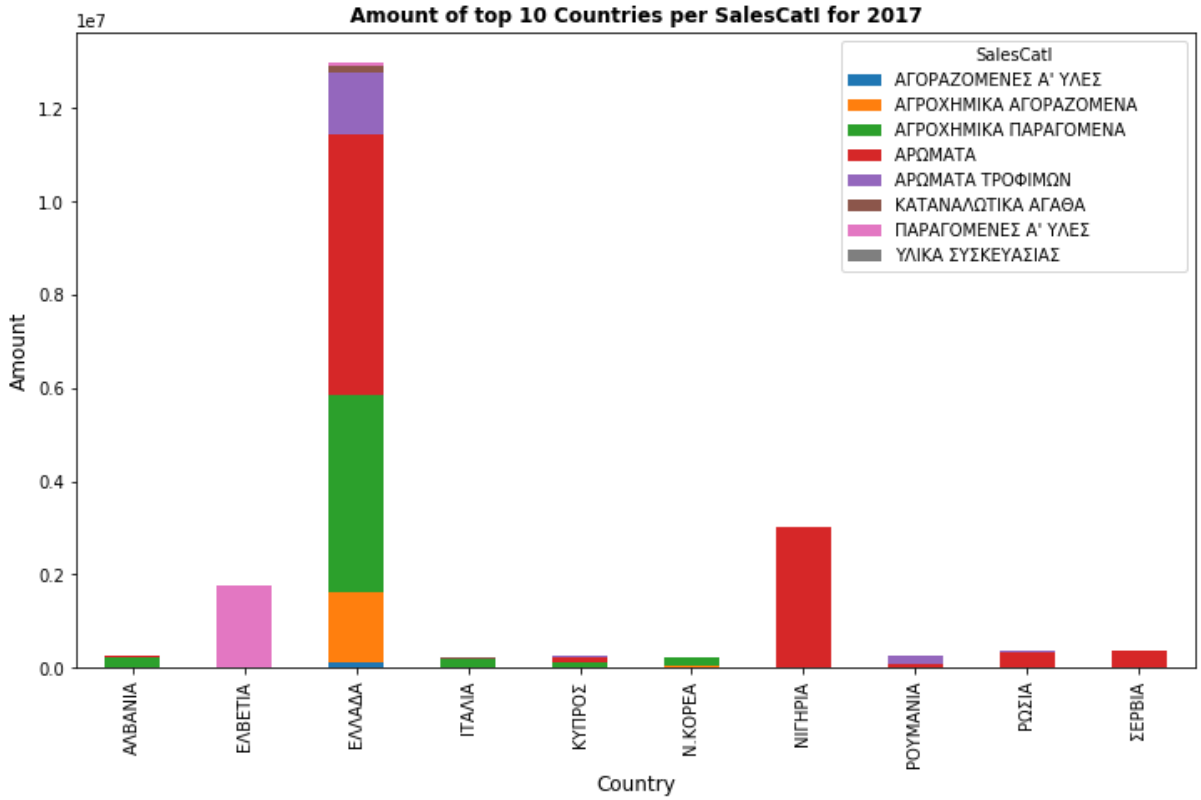
Παρατηρούμε ότι περίπου το 50% των εσόδων της εταιρείας προέρχεται από τον κλάδο των αρωμάτων, ενώ ακολουθεί στο 22%-23% ο κλάδος των αγροχημικών (παραγόμενων). Στην τελευταία θέση βρίσκεται ο κλάδος των καταναλωτικών προϊόντων, με ποσοστό κάτω από 1%. Ο συγκεκριμένος κλάδος έχει αρχίσει να αναπτύσεται στην εταιρεία τα τελευταία έξι χρόνια, επομένως είναι εύλογο να μην έχει καταφέρει ακόμα να προσεγγίσει όλα τα πιθανά τμήματα των πελατών και να αναπτυχθεί σε όλες τις πιθανές αγορές. Ακόμη, παρατηρούμε ότι ανάμεσα στις προϊοντικές κατηγορίες υπάρχει και μια που δεν αναφέρθηκε νωρίτερα, αυτή των υλικών συσκευασίας, της οποίας η συνεισφορά στο κύκλο εργασιών της εταιρείας είναι ελάχιστη. Στο σημείο αυτό θα πρέπει να διευκρινήσουμε ότι η επιχείρηση δεν εμπορεύεται υλικά συσκευασίας.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Όμως σε κάποιες περιπτώσεις η συσκευασία χρεώνεται επιπλέον στον πελάτη και γι' αυτό τον λόγο την συναντάμε και ως κατηγορία πώλησης στο dataset και με μικρή συνεισφορά στον τζίρο της εταιρείας. Τέλος θα μπορούσαμε να υποστηρίξουμε ότι ο τομέας των αρωμάτων θα μπορούσε να ενωθεί με αυτών των αρωμάτων τροφίμων, καθώς πρόκειται για παραπλήσιους κλάδους με αρκετά κοινή τεχνογνωσία. Αντίστοιχα, στην περίπτωση των Αγροχημικών υπάρχει η διάκριση σε Παραγόμενα και Αγοραζόμενα. Στα Αγοραζόμενα ανήκουν εκείνα τα προϊόντα που η εταιρεία αγοράζει με σκοπό να τα μεταπωλήσει, και επί της ουσίας απευθύνονται στο ίδιο target group με τα Αγροχημικά Παραγόμενα. Το ίδιο ακριβώς σκεπτικό έχει ακολουθηθεί και στην περίπτωση των Α' υλών (παραγόμενες και αγοραζόμενες).

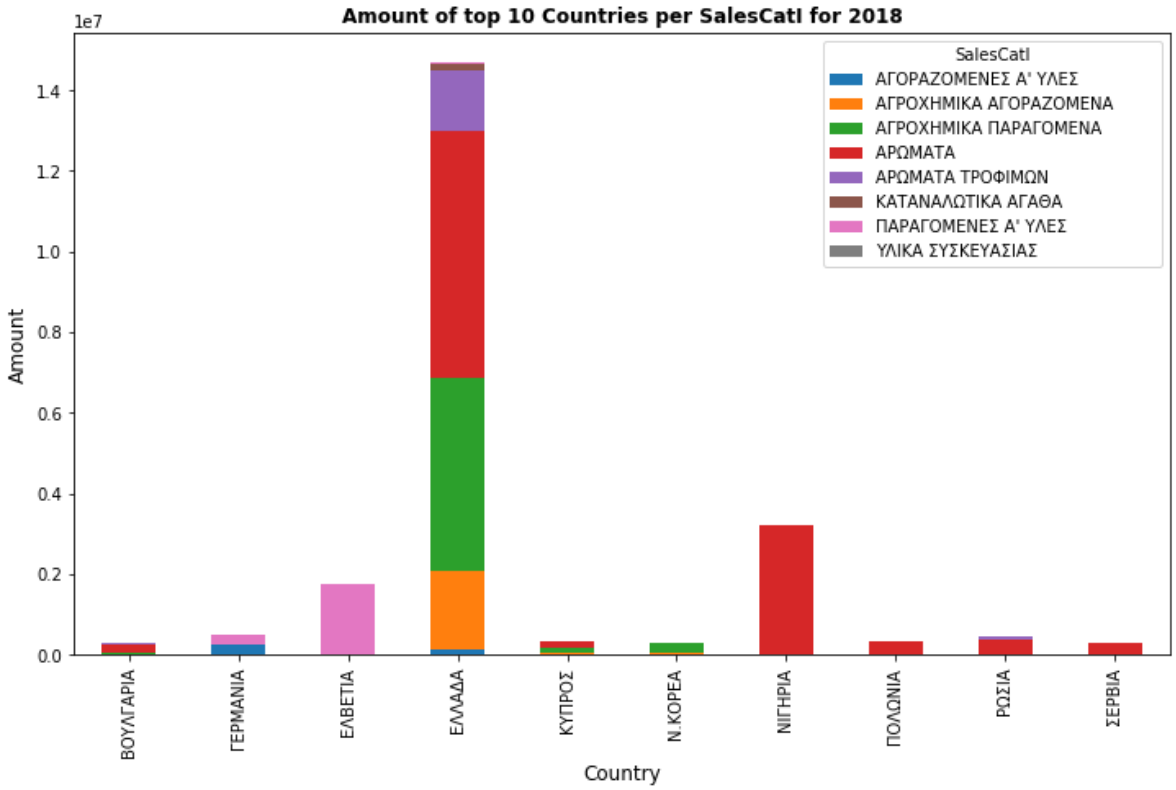
Ακόμη, ενδιαφέρον θα είχε να δούμε για τις πρώτες δέκα χώρες με τις υψηλότερες πωλήσεις, πως διαρθρώνονται τα έσοδα της εταιρείας ανά προϊόντική κατηγορία.

Παρακάτω παρουσιάζουμε σχετικά stacked bar charts για τα δύο έτη.



Διάγραμμα 9, Έσοδα ανά προϊόντική κατηγορία για τις top 10 χώρες για το 2017

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ



Διάγραμμα 10, Έσοδα ανά προϊόντική κατηγορία για τις top 10 χώρες για το 2017

Παρατηρούμε πως η χώρα στην οποία συναντάμε έσοδα από όλους τους κλάδους δραστηριοποίησης της εταιρείας είναι η Ελλάδα, γεγονός αναμενόμενο εφόσον πρόκειται για την χώρα που αποτελεί την έδρα της επιχείρησης, με τους περισσότερους πελάτες και είναι εύλογο η δράση της να επεκτείνεται σε όλους τους τομείς. Ακόμη βλέπουμε ότι οι επόμενες δύο χώρες σε πωλήσεις αφορούν αποκλειστικά έναν κλάδο της εταιρείας έκαστη, των αρωμάτων και των παραγόμενων Α' υλών. Ενώ υπάρχει μια μικρή παρουσία και στα δύο έτη του κλάδου των Αγροχημικών. Το 2017 όμως συναντάμε σε πολύ μικρό ποσοστό τον τομέα των Αρωμάτων Τροφίμων, ενώ το 2018 τον τομέα των Αγοραζόμενων Α' υλών.

Στην συνέχεια επιλέξαμε να εξετάσουμε για τους top δέκα πελάτες μας, την διάρθρωση του τζίρου τους ανά κλάδο δραστηριοποίησης της εταιρείας. Προκειμένου να βρούμε τους top δέκα πελάτες της εταιρείας υπολογίσαμε το σύνολο (sum) του Amount ανά κωδικό πελάτη και τους ταξινομήσαμε σε φθίνουσα σειρά. Οι top δέκα πελάτες για τα έτη 2017 και 2018 φαίνονται στους ακόλουθους πίνακες.

2017

CustomerCode	Ammount	pctAmount
15742	2442350.69	11.30
20442	1773264.65	8.21
15436	626665.92	2.90
24139	610384.15	2.82
15744	491811.55	2.28
23829	382126.00	1.77
19629	366533.40	1.70
16838	290054.00	1.34
15409	274480.50	1.27
36498	266662.20	1.23

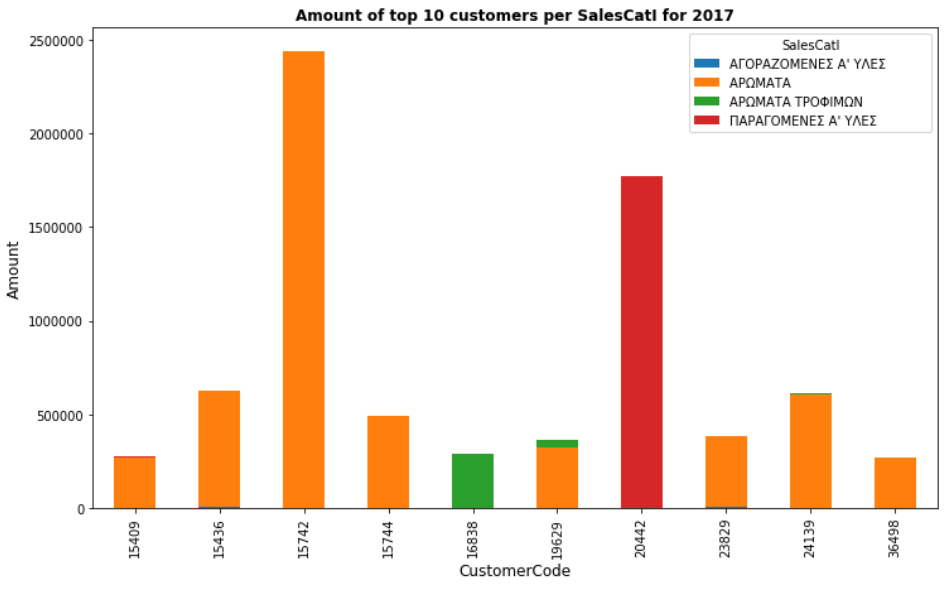
2018

CustomerCode	Ammount	pctAmount
15742	2257921.13	9.42
20442	1738349.67	7.25
15744	909241.35	3.79
15436	666271.41	2.78
24139	529979.80	2.21
19629	472961.00	1.97
23043	410957.30	1.71
22564	331000.00	1.38
19051	308785.75	1.29
17578	294253.50	1.23

Πίνακας 5, Διάρθρωση τζίρου ανά κατηγορία top 10 πελατών για το 2017 και 2018

Στους παραπάνω πίνακες εμφανίζεται επίσης και το ποσοστό του κύκλου εργασιών της εταιρείας που οφείλεται στους συγκεκριμένους πελάτες. Παρατηρούμε ότι τα ποσοστά των δύο πρώτων πελατών είναι αρκετά πιο υψηλά σε σχέση με τους υπόλοιπους πελάτες, καθώς επίσης και ότι σχεδόν το 33% - 34% του κύκλου εργασιών της εταιρείας οφείλεται στους πρώτους 10 πελάτες της.

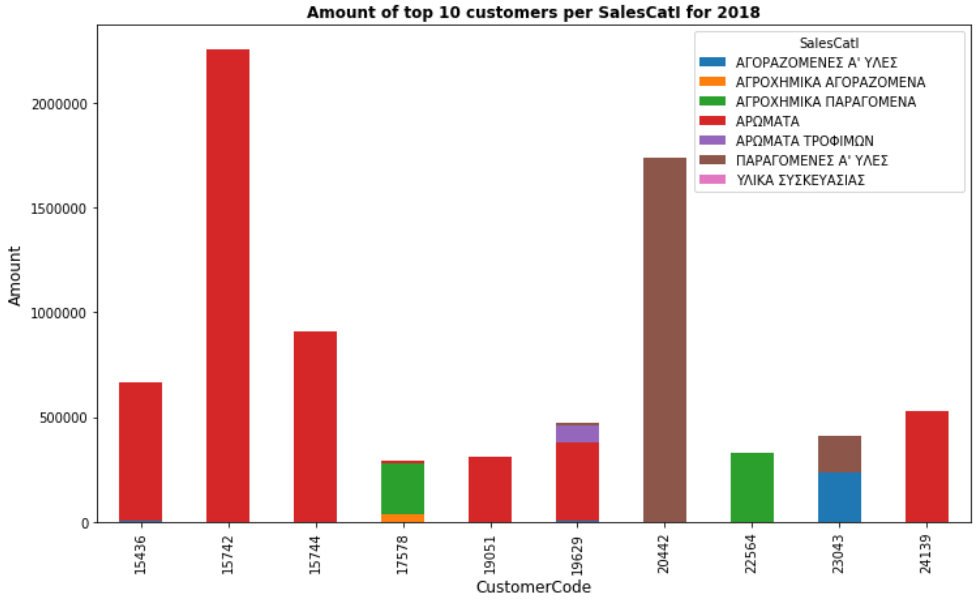
Στα παρακάτω stacked bar charts φαίνεται το ύψος του ποσού του κύκλου εργασιών που οφείλεται σε κάθε κλάδο δραστηριοποίησης της εταιρείας για τους δέκα πρώτους πελάτες της και για τις δύο χρήσεις 2017 και 2018.



Διάγραμμα 11, Έσοδα ανά κατηγορία πωλήσεων για τους top 10 πελάτες για το 2017

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Όπως φαίνεται από το παραπάνω bar chart για το έτος 2017, στους κορυφαίους δέκα πελάτες της εταιρείας κυριαρχεί ο κλάδος των Αρωμάτων, γεγονός αρκετά αναμενόμενο αν λάβουμε υπόψη μας ότι είναι και ο τομέας εκείνος που της αποφέρει τα περισσότερα έσοδα. Ενώ ο δεύτερος σημαντικότερος πελάτης ανήκει στον κλάδο των Παραγόμενων Α' υλών. Στο σημείο αυτό επίσης να αναφέρουμε ότι υπάρχει κάποια μικρή παρουσία των Αρωμάτων Τροφίμων, ενώ η παρουσία των Αγοραζόμενων Α' υλών είναι σχεδόν μηδαμινή. Παρακάτω παρουσιάζουμε αντίστοιχο stacked bar chart για το έτος 2018.



Διάγραμμα 12, Έσοδα ανά κατηγορία πωλήσεων για τους top 10 πελάτες για το 2018

Αντίστοιχα όπως φαίνεται και από το γράφημα για το 2018, και πάλι πιο έντονη είναι η παρουσία του κλάδου των αρωμάτων. Σχεδόν έξι στους δέκα (6/10) είναι αποκλειστικά πελάτες του κλάδου των αρωμάτων, με δύο να έχουν ελάχιστη παρουσία και στον τομέα των Αγοραζόμενων Α' υλών, ενώ ο ένας εκ των δύο και στα Αρώματα Τροφίμων. Παρατηρούμε ότι σε σχέση με το 2017 υπάρχει μια αύξηση στον τομέα των Α' υλών, καθώς εμφανίζεται και άλλος πελάτης στους πρώτους 10 πέρα από τον δεύτερο σημαντικότερο που να ενδιαφέρεται για τον συγκεκριμένο κλάδο. Τέλος να αναφέρουμε ότι βλέπουμε πως στους πρώτους δέκα πελάτες πλέον εμφανίζονται και πελάτες από τον τομέα των Αγροχημικών (παραγόμενων και αγοραζόμενων), γεγονός που μπορεί να υποδηλώνει τις προσπάθειες της εταιρείας να αναπτύξει περαιτέρω τον συγκεκριμένο τομέα.

Επίσης, παρατηρούμε ότι έξι στους δέκα πελάτες είναι κοινοί και στις δύο χρονιές και μάλιστα στις υψηλότερες θέσεις, γεγονός που δείχνει την προσπάθεια της εταιρείας να δημιουργεί και να διατηρεί σταθερές και αξιόπιστες σχέσεις με τους πελάτες της. Ακόμη, αν επιλέξουμε να δούμε τις χώρες από τις οποίες προέρχονται οι δέκα κορυφαίοι πελάτες της εταιρείας, θα διαπιστώσουμε ότι οι τέσσερις πρώτοι προέρχονται από την Νιγηρία, την Ελβετία και την Ελλάδα, γεγονός που επιβεβαιώνει τον ισχυρισμό που αναφέρθηκε παραπάνω ότι στις χώρες αυτές υπάρχουν σημαντικοί και μεγάλοι πελάτες για την εταιρεία. Παρακάτω παρουσιάζουμε σχετικά previews που δείχνουν τις χώρες των δέκα κορυφαίων πελατών της εταιρείας.

2017

```
Cus17.groupby(['Country'])['CustomerCode'].unique()
```

```
Country
ΕΛΒΕΤΙΑ                                          [20442]
ΕΛΛΑΔΑ                  [15436, 36498, 24139, 16838, 15409]
ΝΙΓΗΡΙΑ                                  [15744, 15742]
ΡΩΣΙΑ                                            [19629]
ΣΕΡΒΙΑ                                           [23829]
Name: CustomerCode, dtype: object
```

Εικόνα 3, Χώρες top 10 Πελατών για το 2017

2018

```
Cus18.groupby(['Country'])['CustomerCode'].unique()
```

```
Country
ΓΕΡΜΑΝΙΑ                                         [23043]
ΕΛΒΕΤΙΑ                                          [20442]
ΕΛΛΑΔΑ                  [15436, 19051, 22564, 24139]
Ν.ΚΟΡΕΑ                                          [17578]
ΝΙΓΗΡΙΑ                                  [15744, 15742]
ΡΩΣΙΑ                                            [19629]
Name: CustomerCode, dtype: object
```

Εικόνα 4, Χώρες top 10 Πελατών για το 2017

Επιπλέον, επιλέξαμε να εξετάσουμε πως κυμαίνεται ο κύκλος εργασιών της εταιρείας σε χρονικό διάστημα ενός έτους, με βάση τον μήνα. Προκειμένου να γίνει αυτό έπρεπε να εξάγουμε από την στήλη InvoiceDate τον μήνα. Έτσι, αρχικά μετατρέψαμε την στήλη InvoiceDate σε datetime object (μέσω της εντολής to_datatime) και για τα δύο dataset και εκχωρήσαμε την μεταβλητή σε μια καινούργια στήλη InvoiceDateN, την οποία θα χρησιμοποιήσουμε για την ανάλυση μας, προκειμένου να εξάγουμε τον μήνα της εκάστοτε συναλλαγής. Για τον σκοπό αυτό εφαρμόσαμε το function dt.month στη στήλη InvoiceDateN και το output το καταχωρήσαμε σε μια καινούργια στήλη, με όνομα InvoiceMonth, που δείχνει τον μήνα που πραγματοποιήθηκε η εκάστοτε συναλλαγή. Στην

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

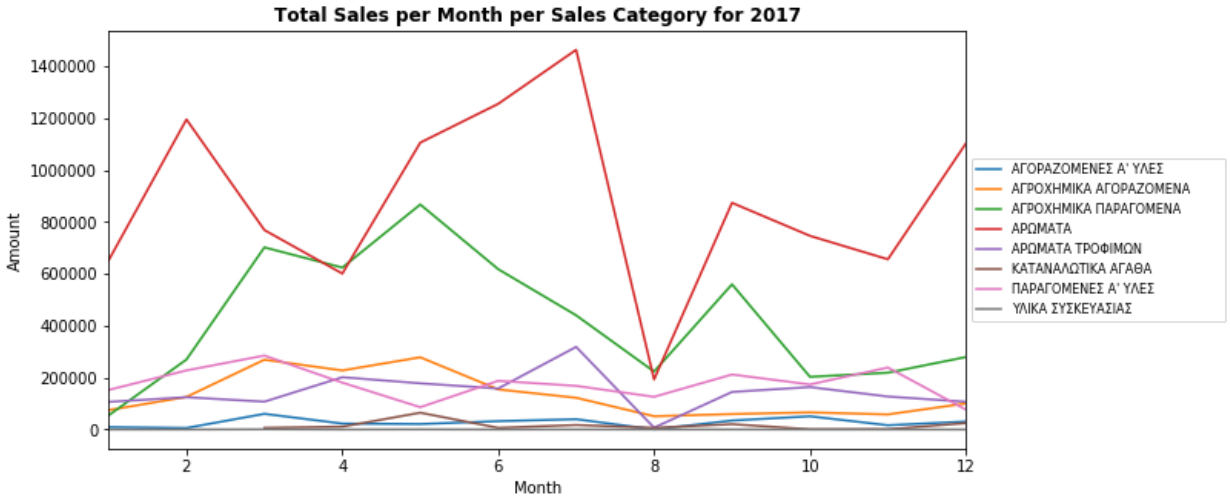
συνέχεια και προκειμένου να δούμε πως κυμαίνονται οι πωλήσεις ανά κατηγορία προϊόντος μέσα στους μήνες του έτους, υπολογίσαμε το σύνολο των πωλήσεων (sum of Amount) ανά μήνα και ανά κλάδο δραστηριότητας της εταιρείας.

Παρακάτω παρουσιάζουμε σχετικό πίνακα για το έτος 2017 και 2018 και αντίστοιχα line plots.

2017

SalesCatl / InvoiceMonth	ΑΓΟΡΑΖΟΜΕΝΕΣ Α' ΥΛΕΣ	ΑΓΡΟΧΗΜΙΚΑ ΑΓΟΡΑΖΟΜΕΝΑ	ΑΓΡΟΧΗΜΙΚΑ ΠΑΡΑΓΟΜΕΝΑ	ΑΡΩΜΑΤΑ	ΑΡΩΜΑΤΑ ΤΡΟΦΙΜΩΝ	ΚΑΤΑΝΑΛΩΤΙΚΑ ΑΓΑΘΑ	ΠΑΡΑΓΟΜΕΝΕΣ Α' ΥΛΕΣ	ΥΛΙΚΑ ΣΥΣΚΕΥΑΣΙΑΣ
1	9321.50	74589.68	54244.05	651367.78	106397.38	6207.36	152413.11	5.00
2	5939.50	125598.53	269838.48	1195088.46	124482.50	NaN	227814.40	2.75
3	60437.20	269099.12	701954.15	768689.84	107328.30	6496.00	285076.72	744.23
4	22879.36	227667.61	624098.73	600904.56	201560.44	10912.00	180659.06	8.50
5	21116.00	278375.98	867296.13	1105658.40	178065.86	64788.84	86018.36	454.00
6	32179.20	154557.31	617967.50	1255100.39	158109.76	6446.27	187545.96	22.65
7	39810.93	122364.87	440100.88	1463429.07	318467.54	17382.04	168191.32	307.75
8	430.00	51205.74	223373.68	192913.12	7054.90	6504.00	126058.70	8.75
9	34684.80	59552.57	559692.88	873688.15	144792.93	20602.56	212001.31	253.00
10	50991.58	65995.40	203045.32	746443.34	163545.74	1033.92	173516.84	4.26
11	16512.22	58180.82	218923.72	655968.10	127125.40	1101.12	238971.17	1.75
12	30693.00	100417.67	279227.57	1102183.24	107049.08	24657.84	76918.01	31.27

Πίνακας 6, Διακύμανση των πωλήσεων ανά προϊόντική κατηγορία και ανά μήνα για το 2017

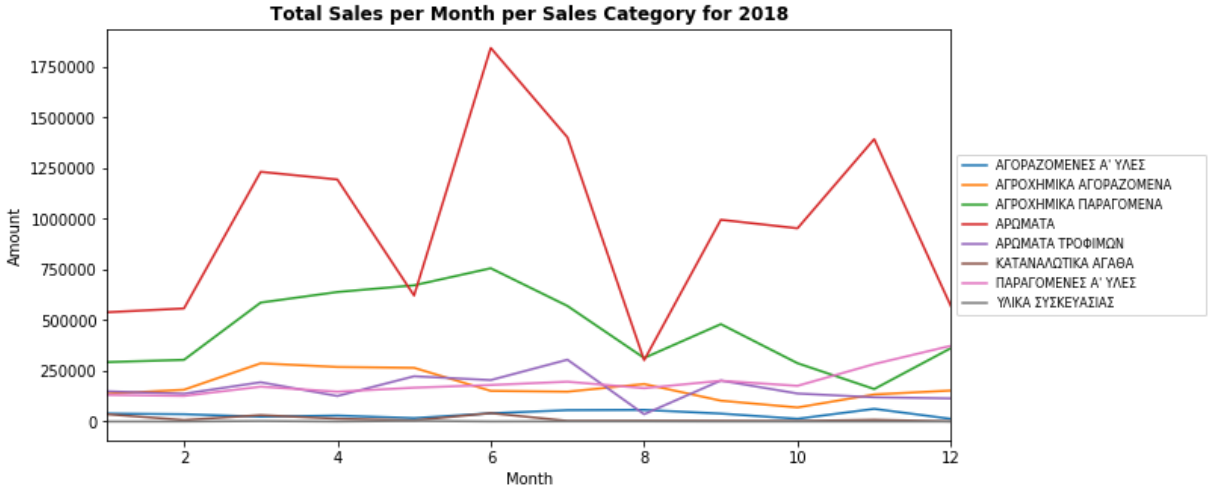


Διάγραμμα 13, Διακύμανση των πωλήσεων ανά προϊόντική κατηγορία και ανά μήνα για το 2017

2018

SalesCatl InvoiceMonth	ΑΓΟΡΑΖΟΜΕΝΕΣ Α' ΥΛΕΣ	ΑΓΡΟΧΗΜΙΚΑ ΑΓΟΡΑΖΟΜΕΝΑ	ΑΓΡΟΧΗΜΙΚΑ ΠΑΡΑΓΟΜΕΝΑ	ΑΡΩΜΑΤΑ	ΑΡΩΜΑΤΑ ΤΡΟΦΙΜΩΝ	ΚΑΤΑΝΑΛΩΤΙΚΑ ΑΓΑΘΑ	ΠΑΡΑΓΟΜΕΝΕΣ Α' ΥΛΕΣ	ΥΛΙΚΑ ΣΥΣΚΕΥΑΣΙΑΣ
1	39395.70	137298.90	293350.49	539469.92	149146.61	35920.88	131452.71	12.25
2	36073.66	157047.86	304509.10	558254.11	137345.18	7327.04	126446.44	102.00
3	23917.06	287348.02	587142.31	1232761.98	193942.10	31615.92	172009.32	2261.40
4	29477.34	269263.27	639369.04	1194826.57	126038.00	13938.24	147355.42	6.75
5	16886.30	265103.99	672379.14	621905.60	223475.64	5797.92	167053.34	3028.50
6	40590.36	151363.46	756682.75	1844322.13	204609.99	41350.40	180347.12	259.68
7	56112.90	147367.17	570565.97	1402836.50	305260.36	3420.72	196639.14	828.74
8	57392.30	185234.26	313680.02	302744.52	35934.70	3730.56	164579.05	768.25
9	39177.30	102939.16	480765.78	996030.97	202406.98	3392.16	201241.11	5.25
10	13540.54	69446.99	287842.63	953895.62	138279.16	2521.92	176308.12	5.25
11	62609.80	133804.38	160379.00	1394126.24	119794.83	8666.56	283685.36	2.75
12	13125.70	152982.94	362300.24	569741.27	113983.43	640.80	373954.52	5.75

Πίνακας 7, Διακύμανση των πωλήσεων ανά προϊόντική κατηγορία και ανά μήνα για το 2018



Διάγραμμα 14, Διακύμανση των πωλήσεων ανά προϊόντική κατηγορία και ανά μήνα, για το 2018

Από τα παραπάνω line plots παρατηρούμε ότι, όπως ήταν αναμενόμενο, οι πωλήσεις των αρωμάτων βρίσκονται πιο ψηλά ενώ ακολουθούν αυτές των αγροχημικών παραγόμενων. Ακόμη, βλέπουμε ότι υπάρχει έντονη διακύμανση στους δύο αυτούς κλάδους στο ύψος των πωλήσεων. Επειδή πρόκειται για εταιρεία που έχει ως πελάτες της εταιρείες (B2B αγορά), η απότομη αυτή διακύμανση μπορεί να υποδηλώνει ότι τοποθετήθηκε μια μεγάλη παραγγελία από κάποιον πελάτη της, τουλάχιστον όσον αφορά τον τομέα των αρωμάτων. Βλέπουμε ότι και στα δύο έτη οι υψηλότερες πωλήσεις στα αρώματα συναντώνται τους καλοκαιρινούς μήνες, κυρίως τον Ιούνιο και τον Ιούλιο, γεγονός που δείχνει μια περιοδικότητα από άποψη παραγγελιών. Απ' την άλλη όσον αφορά τον κλάδο των αγροχημικών (παραγόμενων και αγοραζόμενων) παρατηρούμε ότι υπάρχει ανοδική πορεία στις

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

πωλήσεις από τον Φεβρουάριο έως και τον Ιούλιο περίπου, και τον Σεπτέμβριο. Το γεγονός αυτό μπορεί να δικαιολογηθεί και από την φύση των προϊόντων που περιλαμβάνει ο συγκεκριμένος κλάδος, καθώς πρόκειται για λιπάσματα και φυτοφάρμακα και το συγκεκριμένο χρονικό διάστημα υπάρχει αυξημένη ανάγκη γι' αυτού του είδους τα προϊόντα από γεωπονικά καταστήματα και γεωργούς. Ακόμα, στο σημείο αυτό να αναφέρουμε ότι οι πωλήσεις των καταναλωτικών προϊόντων, παραμένουν σχετικά σταθερές σε όλη την διάρκεια του έτους, αν και θα περιμέναμε να υπάρχει κάποια αύξηση στις πωλήσεις τους καλοκαιρινούς μήνες, που υπάρχει αυξημένη ζήτηση για εντομοαπωθητικά προϊόντα. Το γεγονός αυτό μπορεί να οφείλεται στο ότι πρόκειται για έναν κλάδο στον οποίο η εταιρεία έχει αρχίσει να αναπτύσσεται τα τελευταία χρόνια. Τέλος, παρατηρούμε ότι και στα δύο έτη, σε κάθε κλάδο που επεκτείνεται η δράση της εταιρείας, οι χαμηλότερες πωλήσεις συναντώνται τον μήνα Αύγουστο, χρονικό διάστημα στο οποίο η εταιρεία παραμένει κλειστή.

Στην συνέχεια προχωρήσαμε σε τμηματοποίηση της πελατειακής βάσης της εταιρείας προκειμένου να εντοπίσουμε τα segments των πελατών στα οποία απευθύνεται και σε μια ανάλυση RFM (Regency-Frequency-Monetary), προκειμένου να ταξινομήσουμε τους πελάτες μας, με βάση την συχνότητα των αγορών τους, το πότε πραγματοποίησαν την τελευταία αγορά τους και το σύνολο των χρημάτων που έχουν δαπανήσει για να αγοράσουν προϊόντα από την εταιρεία μας.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

5.2 CUSTOMER SEGMENTATION

Η τμηματοποίηση της πελατειακής βάσης είναι η πρακτική του να διαιρείς την πελατειακή βάση σε ομάδες με βάση παρόμοια χαρακτηριστικά. Η διαίρεση αυτή διευκολύνει τις επιχειρήσεις στην καλύτερη στόχευση και προσέγγιση των πελατών τους, έτσι ώστε να τους μεταδίδουν εξατομικευμένα μηνύματα και να τους προσφέρουν λύσεις που να ικανοποιούν όσο το δυνατόν καλύτερα τις ανάγκες τους.

Όπως έχουμε ήδη αναφέρει το dataset που έχουμε, περιλαμβάνει δέκα τέσσερις στήλες και κάθε γραμμή αντιστοιχεί σε ένα προϊόν που έχει αγοραστεί. Αρχικά επιλέξαμε να δούμε ανά πελάτη το σύνολο των συναλλαγών που είχε με την εταιρεία μας, καταμετρώντας τον μοναδικό αριθμό των τιμολογίων και το σύνολο της αξίας των τιμολογίων που σχετίζονται με εκείνον. Ακόμη, δημιουργήσαμε μια καινούργια στήλη, η οποία δείχνει τον μέσο όρο της αξίας των τιμολογίων που σχετίζονται με τον εκάστοτε πελάτη.

Παρακάτω παρουσιάζουμε τις πρώτες πέντε γραμμές του καινούργιου dataset που δημιουργήσαμε και πάνω στο οποίο θα βασιστούμε προκειμένου να προχωρήσουμε με την τμηματοποίηση της πελατειακής μας βάσης.

CustomerCode	Count of Invoices	Total Sales	avgInvAmount
12147	3	556.4	185.47
12199	7	4354.0	622.00
13035	4	80.0	20.00
13091	7	15960.0	2280.00
13107	2	208.0	104.00

Πίνακας 8, Σύνολο τιμολογίων και αξίας τους και Μ.Ο αξίας/τιμολόγιο

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Ο αλγόριθμος K-means όπως και οι περισσότεροι αλγόριθμοι clustering είναι ευαίσθητος στην κλίμακα μέτρησης των μεταβλητών. Ακόμα και εάν οι μεταβλητές έχουν την ίδια μονάδα μέτρησης αλλά παρουσιάζουν διαφορετικές διακυμάνσεις, θα πρέπει να προχωρήσουμε με κανονικοποίηση των δεδομένων, καθώς ο αλγόριθμος τείνει να δημιουργεί clusters με βάση τις μεταβλητές εκείνες που έχουν την μεγαλύτερη διακύμανση.

Προκειμένου να γίνει αυτό, εφαρμόσαμε το function rank πάνω στο dataset και από το output που προέκυψε αφαιρέσαμε τον μέσο όρο αυτού και το διαιρέσαμε με την τυπική του απόκλιση. Παρακάτω παρουσιάζουμε τις πρώτες γραμμές του dataset μας έτσι όπως προέκυψαν από την παραπάνω διαδικασία.

CustomerCode	Count of Invoices	Total Sales	avgInvAmount
12147	-0.733867	-0.895567	-0.904451
12199	0.028431	0.305630	0.604152
13035	-0.495760	-1.636542	-1.723611
13091	0.030208	1.035943	1.446411
13107	-1.091028	-1.373558	-1.428642

Πίνακας 9, Κανονικοποιημένος Πίνακας

Στον αλγόριθμο k means ο αριθμός των clusters είναι προκαθορισμένος και ο αλγόριθμος επαναληπτικά αναθέτει κάθε ένα από τα δεδομένα σε ένα cluster με βάση την ομοιότητα των χαρακτηριστικών. Υπάρχουν διάφοροι τρόποι που μπορούν να χρησιμοποιηθούν για τον καθορισμό του αριθμού των clusters.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Silhouette

Η μέθοδος αυτή καθορίζει τον αριθμό των clusters με βάση την ομοιότητα του κάθε στοιχείου με τα υπόλοιπα στοιχεία του cluster, συγκρινόμενο με τα δεδομένα των υπόλοιπων clusters. Η τιμή του Silhouette score κυμαίνεται από -1 έως 1, όπου μια υψηλή τιμή δηλώνει ότι ένα στοιχείο είναι καλά ενσωματωμένο στο συγκεκριμένο cluster και ταιριάζει ελάχιστα με τα γειτονικά του cluster.

Επιλέξαμε να τρέξουμε τον αλγόριθμο k means για clusters από τέσσερα έως και οχτώ και υπολογίσαμε σε κάθε περίπτωση το αντίστοιχο Silhouette Score. Στο ακόλουθο output φαίνονται οι αντίστοιχες τιμές του δείκτη.

```
Silhouette Score for 4 Clusters: 0.4190
Silhouette Score for 5 Clusters: 0.3822
Silhouette Score for 6 Clusters: 0.3900
Silhouette Score for 7 Clusters: 0.3918
Silhouette Score for 8 Clusters: 0.3747
```

Εικόνα 5, Silhouette Score for k=4 - 8

Βλέπουμε ότι η υψηλότερη τιμή του δείκτη εντοπίζεται στα τέσσερα clusters και ισούται με 0.4190. Το γεγονός αυτό υποδηλώνει ότι ο βέλτιστος αριθμός clusters για το dataset μας είναι τέσσερα.

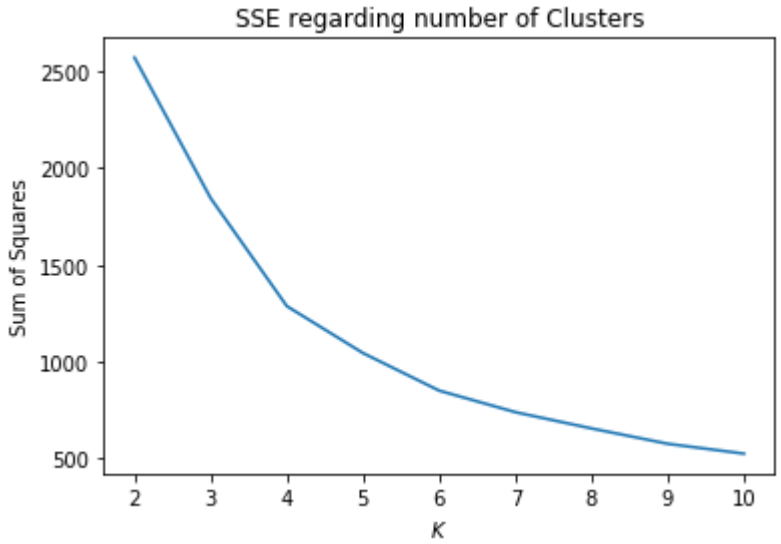
Μια άλλη μέθοδος που χρησιμοποιείται για να βρούμε τον καλύτερο αριθμό clusters είναι η Elbow.

Elbow Method

Στην μέθοδο αυτή τρέχουμε τον αλγόριθμο k means για ένα εύρος τιμών clusters και για κάθε τιμή του k υπολογίζουμε τον δείκτη SSE (Sum of square errors). Στην συνέχεια δημιουργούμε ένα line plot που απεικονίζει τον δείκτη SSE για τις διάφορες τιμές k των clusters. Στόχος μας είναι να βρούμε εκείνη την τιμή του SSE που να είναι αρκετά χαμηλή, διατηρώντας και έναν σχετικά μικρό αριθμό clusters. Όσο αυξάνεται ο αριθμός των clusters, τόσο μειώνεται και η τιμή του δείκτη SSE, καθώς κάθε στοιχείο τείνει όλο και πιο κοντά στον εαυτό του. Πρόκειται ουσιαστικά για εκείνο το σημείο στο γράφημα, όπου αλλάξει η κλίση της καμπύλης και από εκεί που ο δείκτης SSE μειωνόταν με αύξοντα ρυθμό, συνεχίζει να μειώνεται αλλά με φθίνοντα ρυθμό.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Στο παρακάτω γράφημα φαίνεται η καμπύλη του SSE σε σχέση με τον αριθμό των clusters.



Διάγραμμα 15, Line Plot of SSE according to number of clusters

Και πάλι παρατηρούμε ότι ο βέλτιστος αριθμός clusters είναι αυτός των τεσσάρων, αποτέλεσμα που ταυτίζεται και με αυτό που προέκυψε από την ανάλυση που προηγήθηκε με βάση το Silhouette Score.

Στην πορεία και εφόσον καταλήξαμε στον βέλτιστο αριθμό των clusters, ξανατρέξαμε τον αλγόριθμο για k=4, προκειμένου να εντοπίσουμε τα segments των πελατών, έτσι όπως προκύπτουν και να αναθέσουμε κάθε έναν πελάτη μας σε ένα από τα τέσσερα segments. Έτσι, δημιουργήσαμε ένα αντίγραφο του αρχείου δεδομένων πάνω στο οποίο τρέξαμε την ανάλυση μας και σε αυτό δημιουργήσαμε μια επιπλέον στήλη, όπου εκχωρήσαμε το cluster στο οποίο ανήκει ο κάθε πελάτης. Παρακάτω παρουσιάζουμε τις πέντε πρώτες γραμμές του καινούργιου μας αρχείου, όπου βλέπουμε στο τέλος να έχει προστεθεί και η καινούργια μας στήλη .

CustomerCode	Total Sales	Count of Invoices	avgInvAmount	Cluster
12147	-0.895567	-0.733867	-0.904451	0
12199	0.305630	0.028431	0.604152	1
13035	-1.636542	-0.495760	-1.723611	0
13091	1.035943	0.030208	1.446411	3
13107	-1.373558	-1.091028	-1.428642	0

Πίνακας 10, Νέο dataset με column cluster

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Το μέγεθος κάθε cluster παρουσιάζεται στον πίνακα που ακολουθεί. Ακόμη στον ίδιο πίνακα έχει συμπεριληφθεί το "είδος" των πελατών που περιλαμβάνει κάθε cluster, με βάση την αξία που προσδίδουν στην επιχείρηση από άποψη κύκλου εργασιών, καθώς και το χρώμα που αντιστοιχεί σε κάθε segment και το οποίο έχει χρησιμοποιηθεί στα scatter plots που ακολουθούν.

Clusters	Size	Type of Customers	Color in plots
Cluster 0	597	Low value customers	Μπλε Χρώμα
Cluster 1	368	Mid value customers_1	Κόκκινο Χρώμα
Cluster 2	438	Mid value customers_2	Πορτοκαλί Χρώμα
Cluster 3	546	High value customers	Πράσινο Χρώμα

Πίνακας 11, Clusters

Στο σημείο αυτό θα μπορούσαμε να πούμε ότι τα αποτελέσματα από την υλοποίηση του αλγορίθμου είναι ικανοποιητικά, οι πελάτες είναι καλά διαχωρισμένοι, τα clusters ισορροπημένα και ο αλγόριθμος δεν φαίνεται να έχει επηρεαστεί από τυχόν ακραίες τιμές των μεταβλητών – γεγονός στο οποίο συνέβαλε και η τυποποίηση στην κανονική κατανομή Z(0,1).

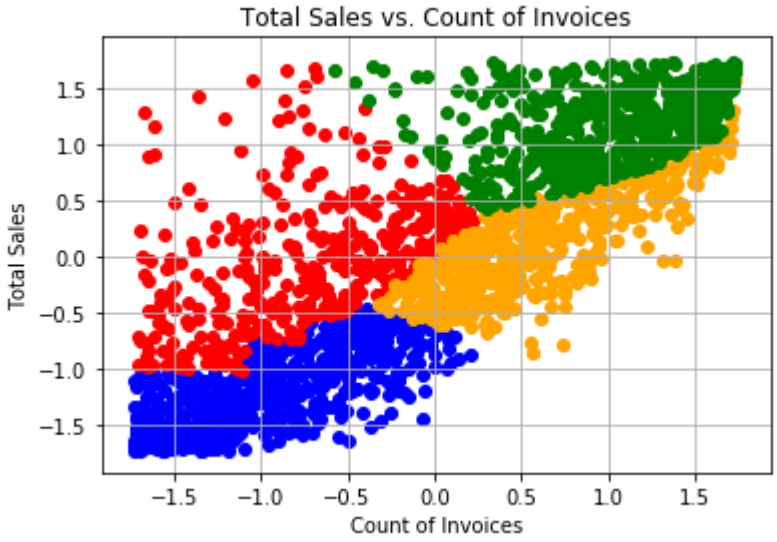
Στην συνέχεια και προκειμένου να μπορέσουμε να εξάγουμε κάποια χρήσιμα συμπεράσματα για τα segments των πελατών στα οποία απευθυνόμαστε υπολογίσαμε για κάθε cluster το κέντρο και για τις τρεις μεταβλητές του dataset μέσω της εντολής kmeans.cluster_centers. Το output αυτών των εντολών το εισαγάγαμε σε ένα dataframe, το οποίο παρουσιάζουμε παρακάτω.

	Total Sales	Count of Invoices	avgInvAmount
0	-1.180147	-0.940918	-0.991568
1	-0.015394	-0.733008	0.858942
2	0.129906	0.620015	-0.540455
3	1.196546	1.025474	0.938818

Πίνακας 12, Centroids of Cluster

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Στο επόμενο βήμα επιλέξαμε να οπτικοποιήσουμε σε scatter plots, τα clusters που προέκυψαν, συναρτήσει των μεταβλητών Total Sales, Count of Invoices και avgInvoiceAmount.

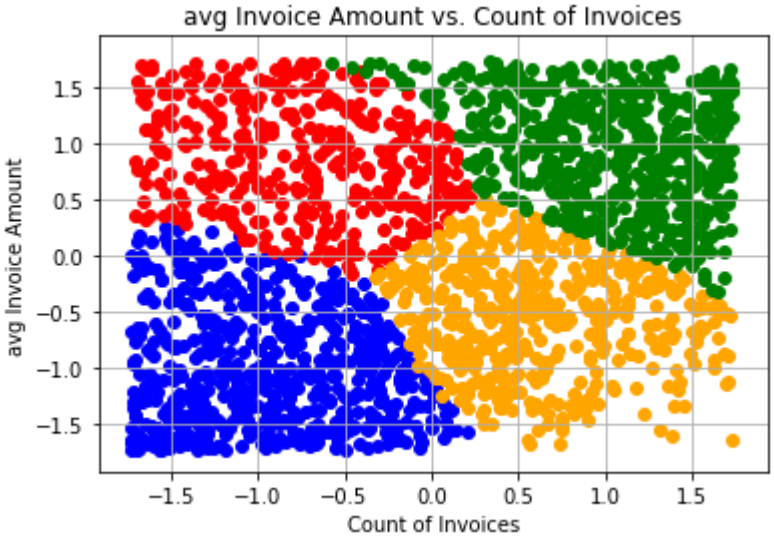


Διάγραμμα 16, Συνολικές Πωλήσεις – Σύνολο τιμολογίων

Στο παραπάνω γράφημα βλέπουμε το σύνολο της αξίας των αγορών των πελατών σε σχέση με τον αριθμό των τιμολογίων που έχουν εκδοθεί. Το cluster με το πράσινο χρώμα είναι εκείνο με τις υψηλότερες πωλήσεις σε αξία και με τις περισσότερες παραγγελίες. Επομένως θα μπορούσε να θεωρηθεί ως εκείνο το segment που περιλαμβάνει τους πιο valuable customers. Ενώ αντίθετα το cluster με το μπλε χρώμα περιλαμβάνει εκείνους τους πελάτες με το μικρότερο ύψος τόσο σε συνολικές πωλήσεις όσο και σε πλήθος συναλλαγών με την εταιρεία.

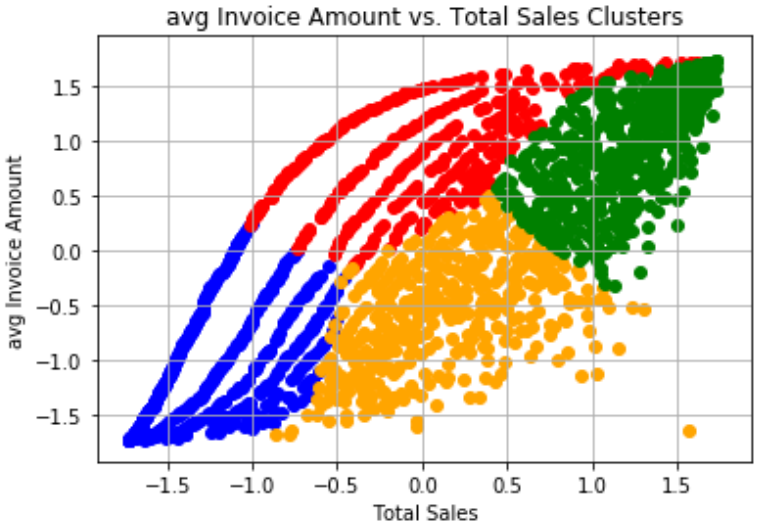
Στο επόμενο scatter plot βλέπουμε τον μέσο όρο της αξίας του τιμολογίου συναρτήσει του αριθμού των τιμολογίων. Και πάλι οι πελάτες που ανήκουν στο cluster με το πράσινο χρώμα, δηλαδή στο cluster 3 είναι οι high value customers για την εταιρεία, ενώ αυτοί που ανήκουν στο cluster 0, με το μπλε χρώμα. είναι οι low value customers.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ



Διάγραμμα 17, Μέση αξία τιμολογίων – Σύνολο τιμολογίων

Τέλος, στο γράφημα που ακολουθεί παρουσιάζουμε την σχέση ανάμεσα στην συνολική αξία πωλήσεων για κάθε cluster σε σχέση με την μέση τιμή της αξία των τιμολογίων. Και πάλι βλέπουμε πως επιβεβαιώνεται ο ισχυρισμός ότι οι πελάτες που ανήκουν στο πράσινο cluster αποτελούν για την επιχείρηση τους πιο επικερδείς, ενώ αυτοί που ανήκουν στο cluster 0 με το μπλε χρώμα της προσδίδουν την μικρότερη αξία.



Διάγραμμα 18, Μέση αξία τιμολογίων - Συνολικές Πωλήσεις

Η επιχείρηση προκειμένου να αυξήσει την κερδοφορία της και να ανταποκριθεί καλύτερα στις ανάγκες των πελατών της θα μπορούσε να επικεντρωθεί στα clusters 1 και 2, με το κόκκινο και το πορτοκαλί χρώμα αντίστοιχα. Στην περίπτωση του cluster 1, έχουμε υψηλές πωλήσεις αλλά μικρό σχετικά αριθμό συναλλαγών, υψηλότερες από το cluster 0 (μπλε χρώμα) αλλά χαμηλότερες από το

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

cluster 3 (πράσινο χρώμα). Ενώ στην περίπτωση του cluster 2 έχουμε πιο χαμηλές πωλήσεις από άποψη αξίας, αλλά περισσότερες συναλλαγές με την επιχείρηση μας. Και στα δύο άρα, αυτά clusters υπάρχουν περιθώρια ανάπτυξης για την εταιρεία. Επομένως, θα ήταν σκόπιμο η επιχείρηση να αναλύσει εις βάθος τις ανάγκες τους, προκειμένου να μπορέσει να τους προτείνει λύσεις και εναλλακτικές που να εξυπηρετούν όσο το δυνατόν καλύτερα την ζήτηση τους.

Ακόμη, πλέον είναι εύκολο να εντοπίσουμε ποια προϊόντα είναι τα πλέον αντιπροσωπευτικά σε κάθε segment. Έτσι, στο segment 3 που είναι και το πιο κερδοφόρο για την επιχείρηση μας παρουσιάζονται τα πέντε προϊόντα με τις περισσότερες πωλήσεις για την επιχείρηση μας.

ItemCode	ItemNameHash
21928205	1086
21928209	977
21928206	632
21928203	596
21928239	536

Πίνακας 11, Top 5 προϊόντα με περισσότερες πωλήσεις για cluster 3

Στο σημείο αυτό να αναφέρουμε ότι πρόκειται για τέσσερα προϊόντα που ανήκουν στον κλάδο των αγροχημικών και μάλιστα τα τέσσερα πρώτα αποτελούν διαφορετικές συσκευασίες του ίδιου προϊόντος.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Από άποψη ανάπτυξης θα ήταν χρήσιμο για την εταιρεία να επικεντρωθεί στα clusters 1 και 2, με το κόκκινο και κίτρινο χρώμα αντίστοιχα, όπου για το πρώτο όπως είπαμε έχουμε μικρό αριθμό συναλλαγών με υψηλή συνολική αξία, ενώ για το δεύτερο cluster, έχουμε μεγαλύτερο αριθμό συναλλαγών αλλά με μικρότερη αξία. Ας δούμε στη συνέχεια γι' αυτά τα δύο segments ποια είναι τα προϊόντα εκείνα με τις υψηλότερες πωλήσεις.

Cluster 2

ItemCode	ItemNameHash
20109002	1205
20110501	772
21928205	619
21928203	412
21646300	406

Πίνακας 124, Προϊόντα με περισσότερες πωλήσεις για cluster 2

Cluster 1

ItemCode	ItemNameHash
21646300	55
21928205	37
24217702	33
21928209	33
21928235	33

Πίνακας 15, Προϊόντα με περισσότερες πωλήσεις για cluster 1

Παρατηρούμε ότι στο Cluster 2 όπου έχουμε υψηλό σχετικά αριθμό συναλλαγών αλλά μικρής σχετικής αξίας βρίσκονται πρώτα σε πωλήσεις δύο προϊόντα του καταναλωτικού τομέα, καθώς επίσης και δύο προϊόντα αγροχημικών που συναντάμε τόσο στο cluster 3 όσο και στο 1. Αυτό σημαίνει ότι περιλαμβάνει πελάτες που ναι μεν αγοράζουν συχνά από την επιχείρηση μας αλλά η συνολική αξία των προϊόντων που αγοράζουν είναι σχετικά μικρή. Σε αυτούς τους πελάτες θα μπορούσαμε να προτείνουμε κάποια επιπλέον προϊόντα, είτε συμπληρωματικά είτε υποκατάστατα αυτών που ήδη αγοράζουν, που όμως να αποφέρουν για την επιχείρηση μας μεγαλύτερο όφελος. Τέλος αντίστοιχα και στο cluster 1 όπου έχουμε μικρό αριθμό συναλλαγών αλλά μεγαλύτερης αξίας, παρατηρούμε ότι όλα τα προϊόντα ανήκουν στα αγροχημικά και μάλιστα τέσσερα εκ των οποίων τα έχουμε ήδη συναντήσει και στα δύο προηγούμενα clusters. Σε αυτούς τους πελάτες θα μπορούσε να γίνει κάποια προωθητική ενέργεια μέσω email που να τους "υπενθυμίζει" την αξιοποίηση και τα πλεονεκτήματα που προκύπτουν από την χρήση συγκεκριμένων προϊόντων, ανάλογα την χρονική περίοδο, δεδομένου ότι η ζήτηση για αγροχημικά προϊόντα δεν είναι ίδια σε όλη την διάρκεια του χρόνου.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Στην συνέχεια προχωρήσαμε σε τμηματοποίηση της πελατειακής μας βάσης έχοντας ως κριτήριο τα προϊόντα που προμηθεύονται από την επιχείρηση μας, με βάση την κατηγορία πώλησης στην οποία ανήκει το κάθε προϊόν. Προκειμένου να γίνει αυτό αξιοποιήσαμε τις στήλες CustomerCode, SalesCatI,. Quantity και Amount. Στην συνέχεια δημιουργήσαμε μια λίστα με τις προϊοντικές κατηγορίες από τις οποίες έχει προμηθευτεί ο κάθε πελάτης και με βάση αυτήν δημιουργήθηκε ένας πίνακας με βάρη για κάθε πελάτη, όπου η κάθε μια μεταβλητή αντιπροσωπεύει την σπουδαιότητα κάθε κατηγορίας για τον συγκεκριμένο πελάτη, καθώς και την σπουδαιότητα της κατηγορίας στο σύνολο των πελατών. Πάνω σε αυτόν τον πίνακα εφαρμόστηκε ο αλγόριθμος k-means, ο οποίος υλοποιήθηκε για k=4.Έτσι προέκυψαν τέσσερα clusters, τα βασικά χαρακτηριστικά των οποίων παρουσιάζονται στον πίνακα που ακολουθεί.

	Clients	**Popular Item Code**	**Amount**
Cluster0	1027	21928210	1347130.00
		21646300	1020156.30
		21928209	956121.13
		21928205	697635.58
Cluster1	601	22298800	1368672.69
		21522200	807857.88
		23007300	510651.44
		22548200	505935.47
Cluster2	222	22527200	466620.00
		23153300	205890.00
		23618100	153400.00
		21224100	77995.50
Cluster3	99	21902800	3616217.32
		24235800	446233.12
		21904700	200921.13
		21901300	163011.30

Πίνακας 136, Popular Item Code and its Amount per Cluster

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Σε κάθε cluster συμμετέχουν πελάτες που έχουν συναλλαγές με όλες τις προϊοντικές κατηγορίες της εταιρείας, όμως είναι μια αυτή που κυριαρχεί. Για την ακρίβεια στο cluster 0 κυριαρχεί ο κλάδος των αγροχημικών, στο cluster 1 επικρατούν οι πελάτες των αρωμάτων, στο cluster 2 η πλέον αντιπροσωπευτική κατηγορία πώλησης είναι αυτή των αρωμάτων τροφίμων και τέλος στο cluster 3 συναντάμε κυρίως τους πελάτες των πρώτων υλών. Στον παραπάνω πίνακα τα προϊόντα που αναφέρονται, που είναι και τα υψηλότερα σε πωλήσεις για τον συγκεκριμένο κλάδο, είναι αντιπροσωπευτικά προϊόντα αυτών των κατηγοριών.

Στον πίνακα που ακολουθεί δείχνουμε τι ποσοστό των πωλήσεων του κάθε cluster αντιστοιχεί σε κάθε κατηγορία πωλήσεων από την οποία απαρτίζεται το συγκεκριμένο segment.

SalesCatI	Cluster 0 %	Cluster 1 %	Cluster 2 %	Cluster 3 %
Αγροχημικά Παραγόμενα	73.75	1.85	-	-
Αγροχημικά Αγοραζόμενα	26.15	0.33	-	0.02
Αρώματα	0.03	95.93	0.55	0.06
Αρώματα Τροφίμων	0.00	1.42	96.72	1.89
Α’ ύλες Παραγόμενες	-	0.08	0.11	81.55
Α’ ύλες Αγοραζόμενες	0.05	0.40	1.81	10.95
Καταναλωτικά Αγαθά	0.00	0.00	0.81	5.48
Υλικά Συσκευασίας	0.02	0.02	-	0.05

Πίνακας 17, Ποσοστό πωλήσεων ανά κατηγορία πώλησης και ανά cluster

Όπως φαίνεται και από τον παραπάνω πίνακα σε κάθε segment το μεγαλύτερο μέρος του κύκλου εργασιών προέρχεται από μια προϊοντική κατηγορία που αποτελεί και την πλέον χαρακτηριστική του cluster.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

5.3 RFM ANALYSIS

Όπως έχουμε ήδη αναφέρει, η ανάλυση RFM είναι μια τεχνική που χρησιμοποιείται από τις επιχειρήσεις προκειμένου να εντοπίσουν και να μετρήσουν την αξία του πελάτη. Ουσιαστικά τμηματοποιεί τους πελάτες σε ομάδες με κοινά χαρακτηριστικά με βάση το πόσο πρόσφατα αγόρασαν από την επιχείρηση μας (recency), το πόσο συχνά πραγματοποιούν συναλλαγές με την εταιρεία μας (frequency) και το σύνολο των χρημάτων που δαπανούν προκειμένου να αγοράσουν τα προϊόντα μας (monetary).

Για την ανάλυση μας χρησιμοποιήθηκε το σύνολο των δεδομένων που έχουμε στην διάθεση μας και αφορά χρονικό διάστημα από 01/02/2017 έως 31/12/2018. Όπως είδαμε και παραπάνω, η πλειοψηφία των πελατών της εταιρείας συγκεντρώνεται στην Ελλάδα. Για την ακρίβεια, σχεδόν το 90% αυτών βρίσκεται στον ελλαδικό χώρο, ως εκ τούτου επιλέξαμε η ανάλυση μας να επικεντρωθεί σε αυτούς του πελάτες.

	Country	CustomerCode	pctCustomers
49	ΕΛΛΑΔΑ	1753	89.943561
48	N/A	51	2.616727
47	ΚΥΠΡΟΣ	16	0.820934
46	ΙΤΑΛΙΑ	12	0.615700
45	ΑΛΒΑΝΙΑ	6	0.307850

Πίνακας 148, Πελάτες ανά Χώρα

Επομένως, συγκεντρώσαμε σε ένα ξεχωριστό dataset τους πελάτες που προέρχονται από την Ελλάδα, ορίζοντας στο αρχικό dataset φίλτρο στη χώρα. Το αρχείο που προέκυψε αποτελείται από 72259 γραμμές (από τις 78890 του αρχικού dataset), χωρίς καμία null τιμή, όπως φαίνεται και στο preview που ακολουθεί.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
greece_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 72259 entries, 530 to 78889
Data columns (total 14 columns):
Salesman            72259 non-null object
CustomerType        72259 non-null object
Country             72259 non-null object
CustomerCode        72259 non-null int64
CustomerNameHash    72259 non-null object
ItemCode            72259 non-null int64
ItemNameHash        72259 non-null object
SalesCatI           72259 non-null object
Invoice             72259 non-null int64
InvoiceType         72259 non-null object
InvoiceDate         72259 non-null object
InvoiceYear         72259 non-null int64
Quantity            72259 non-null float64
Ammount             72259 non-null float64
dtypes: float64(2), int64(4), object(8)
memory usage: 8.3+ MB
```

Εικόνα 6, Πληροφορίες στηλών Greece_data

Στο παρακάτω output παρουσιάζονται μέσω της εντολής describe κάποια στατιστικά δεδομένα για το dataset μας. Παρουσιάζονται το πλήθος, ο μέσος όρος, η τυπική απόκλιση, τα τεταρτημόρια και η μέγιστη και ελάχιστη τιμή των δεδομένων του dataset.

	CustomerCode	ItemCode	Invoice	InvoiceYear	Quantity	Ammount
count	72259.000000	7.225900e+04	7.225900e+04	72259.000000	72259.000000	72259.000000
mean	25729.942831	2.267044e+07	1.853027e+06	2017.509348	18.577861	383.114595
std	7635.242099	1.008991e+06	3.975454e+05	0.499916	146.800328	1322.277554
min	12147.000000	2.010800e+07	1.009626e+06	2017.000000	0.000000	0.000000
25%	18565.000000	2.193240e+07	2.031366e+06	2017.000000	1.000000	76.000000
50%	24139.000000	2.238320e+07	2.038767e+06	2018.000000	5.000000	153.000000
75%	32840.000000	2.345340e+07	2.045331e+06	2018.000000	10.000000	337.500000
max	39333.000000	2.465100e+07	2.052380e+06	2018.000000	24900.000000	134400.000000

Εικόνα 7, Βασικοί στατιστικοί δείκτες

Προκειμένου να προχωρήσουμε στην ανάλυση RFM, θα χρειαστούμε τις εξής στήλες: CustomerCode, Invoice, InvoiceDate, Quantity και Amount.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Για αρχή μετατρέψαμε την στήλη InvoiceDate σε datetime object μέσω της εντολής to_datetime και ορίσαμε ως τρέχουσα ημερομηνία για την ανάλυση μας την 30/09/2019. Στην συνέχεια δημιουργήσαμε ένα καινούργιο dataset, όπου συγκεντρώσαμε ανά πελάτη το σύνολο των χρημάτων που έχει δαπανήσει για να αγοράσει προϊόντα της εταιρείας μας (sum of amount, monetary), την συχνότητα με την οποία πραγματοποιεί συναλλαγές με την εταιρεία μας, καταμετρώντας τον μοναδικό αριθμό των τιμολογίων που έχουν κοπεί (frequency) και τέλος υπολογίσαμε το χρονικό διάστημα που έχει περάσει από την πραγματοποίηση της τελευταίας συναλλαγής του με την εταιρεία μας μέχρι σήμερα (recency). Για να γίνει αυτό, αφαιρέσαμε από την τρέχουσα ημερομηνία την πιο κοντινή στο σήμερα ημερομηνία τιμολογίου. Έτσι δημιουργήθηκε ένα αρχείο δεδομένων όπου γραμμές είναι οι πελάτες και στήλες οι τιμές των Recency, Frequency και Monetary.

Στην συνέχεια χωρίσαμε τις αξίες κάθε μιας στήλης σε τεταρτημόρια και κάναμε ανάθεση σε κάθε τεταρτημόριο των τιμών από ένα έως τέσσερα, όπου το ένα αντιστοιχεί στην καλύτερη μέτρηση. Οι τιμές αυτές αποθηκεύτηκαν ως στήλες στο dataset μας με τα εξής ονόματα: Recency Score, Frequency Score και Monetary Score. Οι πελάτες οι οποίοι έχουν τις υψηλότερες τιμές σε συχνότητα αγορών και ύψος χρημάτων που έχουν δαπανήσει για την εταιρεία μας καθώς και τις χαμηλότερες τιμές σε regency, δηλαδή έχουν πραγματοποιήσει κάποια συναλλαγή με την εταιρεία μας πιο πρόσφατα θεωρούνται οι καλύτεροι πελάτες μας. Έτσι όσον αφορά τις τιμές των frequency και monetary, στους πελάτες που ανήκουν στο πρώτο τεταρτημόριο, δηλαδή έχουν τις πιο χαμηλές τιμές σε frequency και monetary θα αντιστοιχηθεί η τιμή τέσσερα έως Frequency και Monetary Score, ενώ σε αυτούς που ανήκουν στο τέταρτο τεταρτημόριο θα αντιστοιχηθεί η τιμή ένα. Απ' την άλλη στους πελάτες που ανήκουν στο πρώτο τεταρτημόριο αναφορικά με την αξία της μεταβλητής Recency, θα αντιστοιχηθεί η τιμή ένα, καθώς πρόκειται για αυτούς που έχουν πραγματοποιήσει πιο πρόσφατα συναλλαγή με την εταιρείας μας, οι μέρες που έχουν περάσει από όταν πραγματοποίησαν την τελευταία τους συναλλαγή με την εταιρεία μας είναι οι λιγότερες στο σύνολο του dataset μας. Ενώ αντίστοιχα για τους πελάτες που ανήκουν στο τελευταίο τεταρτημόριο το Recency Score θα πάρει την τιμή τέσσερα.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Παρακάτω παρουσιάζουμε ένα preview των πρώτων πέντε γραμμών του dataset μας, έτσι όπως έχει προκύψει από την μέχρι τώρα ανάλυσή μας.

CustomerCode	Regency	Frequency	Monetary	r_quantile	f_quantile	m_quantile
12147	424	3	556.4	3	3	4
12199	283	7	4354.0	1	3	2
13035	452	4	80.0	3	3	4
13091	340	7	15960.0	2	3	1
13107	591	2	208.0	4	4	4

Πίνακας 159, Dataset με επιμέρους RFM Scores

Στη συνέχεια υπολογίσαμε το συνολικό RFM Score, που ουσιαστικά αποτελεί συνδυασμό των επιμέρους scores. Για παράδειγμα για τον πελάτη με κωδικό 12147, το συνολικό RFM Score ισούται με 334, που προκύπτει από το άθροισμα των επιμέρους scores ως string μεταβλητές. Έτσι στο παραπάνω dataset προστίθεται μια επιπλέον στήλη, η RFM Score, όπως παρουσιάζεται στο παρακάτω output.

CustomerCode	Regency	Frequency	Monetary	r_quantile	f_quantile	m_quantile	RFM Score
12147	424	3	556.4	3	3	4	334
12199	283	7	4354.0	1	3	2	132
13035	452	4	80.0	3	3	4	334
13091	340	7	15960.0	2	3	1	231
13107	591	2	208.0	4	4	4	444

Πίνακας 20, Dataset με συνολικό RFM Score

Όπως είναι εύλογο η χαμηλότερη τιμή του RFM Score είναι η 111. Σε αυτή την τιμή εντοπίζουμε τους καλύτερους πελάτες για την εταιρεία μας: αυτούς που έχουν αγοράσει πρόσφατα, αγοράζουν συχνά και στους οποίους αποδίδεται μεγάλο μέρος του κύκλου εργασιών της εταιρείας.

Παρακάτω παρουσιάζουμε στα πλέον χαρακτηριστικά segments που προκύπτουν από την ανάλυση μας, το πλήθος των πελατών που περιλαμβάνουν.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
Best Customers:  219
Loyal Customers:  424
Big Spenders:  438
Almost Lost:  433
Almost Lost Best Customers:  17
Lost Customers:  438
Lost Best Customers:  6
Lost Cheap Customers:  225
```

Εικόνα 8, Πλήθος κάθε segment

Best Customers

Όπως ήδη αναφέρθηκε στην κατηγορία αυτή ανήκουν οι πελάτες εκείνοι με RFM Score ίσο με 111 και περιλαμβάνει σύνολο 219 πελάτες. Πρόκειται για πελάτες που έχουν αγοράσει πρόσφατα, αγοράζουν συχνά και έχουν δαπανήσει αρκετά χρήματα στις συναλλαγές τους με την επιχείρηση μας. Γι' αυτούς τους πελάτες, η εταιρεία θα μπορούσε να υιοθετήσει κάποιες εκπτώσεις τζίρου στην περίπτωση που ο τζίρος των συγκεκριμένων πελατών έχει ξεπεράσει κάποιο επίπεδο.

Loyal Customers

Ως loyal πελάτες θεωρούνται αυτοί που αγοράζουν συχνά από την επιχείρηση μας, αυτοί δηλαδή για τους οποίους ισχύει Frequency Score = 1. Στο segment αυτό ανήκουν συνολικά 424 πελάτες. Στους πελάτες αυτούς η επιχείρηση θα μπορούσε να εφαρμόσει κάποιες εκπτώσεις στην τιμή προϊόντος, εάν η συχνότητα αγοράς τους ξεπερνάει κάποιο επίπεδο.

Big Spenders

Στην κατηγορία αυτή ανήκουν οι πελάτες εκείνοι στους οποίους οφείλεται μεγάλο μέρος του κύκλου εργασιών της εταιρείας, που έχουν δαπανήσει αρκετά χρήματα στις συναλλαγές τους με την εταιρεία μας και επομένως έχουν Monetary Score = 1. Το πλήθος αυτών των πελατών ισούται με 438. Για τους πελάτες αυτούς η εταιρεία θα μπορούσε στο τέλος του έτους και εφόσον έχουν δαπανήσει για τις αγορές τους ένα αξιοσημείωτο ποσό, να εκδώσει κάποια πιστωτικά τζίρου τιμολόγια.

Όπως είναι κατανοητό οι Best Customers ανήκουν τόσο στους Loyal Customers όσο και στους Big Spenders, αφού όλα τα επιμέρους scores της RFM Analysis που έχει προηγηθεί ισούνται με ένα.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Almost Lost Best Customers

Στο segment με τους Almost Lost Customer συμπεριλαμβάνονται όλοι εκείνοι οι πελάτες για τους οποίους ισχύει Recency Score = 3, ανεξάρτητα από τις τιμές των Frequency και Monetary Score και περιλαμβάνει συνολικά 433 πελάτες. Από αυτούς επιλέξαμε να επικεντρωθούμε σε αυτούς που προσδίδουν περισσότερη αξία στην επιχείρηση, από άποψη συχνότητας και αξίας αγορών. Στην κατηγορία αυτή, των Almost Lost Best Customers ανήκουν οι πελάτες εκείνοι οι οποίοι ναι μεν έχουν δαπανήσει πολλά λεφτά για τις αγορές τους από την εταιρεία μας και οι οποίοι συναλλάσσονταν συχνά με την επιχείρηση μας, όμως έχουν αρκετό καιρό να πραγματοποιήσουν κάποια συναλλαγή. Στην περίπτωση αυτή τόσο το frequency score όσο και το monetary score ισούται με ένα, όμως το recency score ισούται με τρία. Η κατηγορία αυτή περιλαμβάνει 17 πελάτες. Επειδή πρόκειται για μια επιχείρηση που δραστηριοποιείται σε μια B2B αγορά, σημαντικό ρόλο παίζει ο πωλητής και η προσωπική πώληση. Όπως φαίνεται, πρόκειται για πελάτες που μέχρι στιγμής συναλλάσσονταν συχνά με την επιχείρηση μας και που δαπανούσαν σημαντικά ποσά στις αγορές τους, επομένως εφόσον η εταιρεία δεν επιθυμεί να τους χάσει, θα μπορούσε μέσω του εκάστοτε πωλητή να εντοπίσει ποιο είναι ακριβώς το πρόβλημα και ποιος ο λόγος απομάκρυνσης του πελάτη από την επιχείρηση, έτσι ώστε να προβεί σε βελτιωτικές κινήσεις.

Lost Best Customers

Στην κατηγορία των Lost Customers ανήκουν εκείνοι οι πελάτες για τους οποίους έχουμε υψηλό Recency Score, είναι εκείνοι οι πελάτες που δεν έχουν πραγματοποιήσει συναλλαγή με την επιχείρηση μας για μεγάλο χρονικό διάστημα (Recency Score = 4), ανεξαρτήτως των τιμών frequency και monetary score. Στο cluster αυτό ανήκουν 438 πελάτες. Από αυτούς οι έξι πελάτες θα μπορούσαμε να υποστηρίξουμε ότι ανήκουν στους high value customers, καθώς τόσο το frequency όσο και το monetary score ισούται με ένα. Στην κατηγορία των Lost Best Customers ανήκουν πελάτες που αγοράζουν συχνά προϊόντα από την επιχείρηση μας και που έχουν δαπανήσει αρκετά λεφτά στις συναλλαγές τους, όπως και στην προηγούμενη κατηγορία των Almost Lost Best Customers, με την μόνη διαφορά ότι έχουν να πραγματοποιήσουν κάποια συναλλαγή με την εταιρεία για ακόμα μεγαλύτερο χρονικό διάστημα. Για την προσέγγιση των πελατών αυτών, θα μπορούσε να

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

χρησιμοποιηθεί αντίστοιχη πολιτική με αυτή των Almost Lost High Value Customers, μέσω του εκάστοτε πωλητή, προκειμένου η επιχείρηση να αναγνωρίσει τους λόγους απομάκρυνσης τους.

Σχετικά με τους Almost Lost Best Customers όσο και με τους Lost Best Customers, θα είχε ενδιαφέρον να εξετάσουμε και την συνολική αξία αγορών τους, προκειμένου να εντοπίσουμε σε τι ποσοστό συμβάλουν στον συνολικό κύκλο εργασιών της επιχείρησης και κατ' επέκταση κατά πόσο είναι πραγματικά σημαντικοί για την εταιρεία. Στην συνέχεια παρουσιάζονται σχετικοί πίνακες.

Να διευκρινίσουμε ότι ο συνολικός κύκλος εργασιών για την Ελλάδα είναι 27.683.477,54€

Almost Lost Best Customers		
Customer Code	**Monetary**	**Percentage**
34263	146771.40	0.53
17521	74356.00	0.27
33018	55528.60	0.20
32259	51979.20	0.19
24200	45091.12	0.16
36273	37756.58	0.14
19363	34376.70	0.12
32677	31473.14	0.11
25655	21441.30	0.08
32133	21376.70	0.08
30515	20383.72	0.07
18976	17362.16	0.06
22508	17013.99	0.06
17399	14758.58	0.05
35668	13821.03	0.05
18989	12535.59	0.05
18200	11692.40	0.04

Πίνακας 21, Almost Lost Best Customers – Monetary Value

Lost Best Customers		
Customer	**Monetary**	**Percentage**
18565	67371.70	0.24
32999	24301.61	0.09
35954	21347.50	0.08
17303	18416.30	0.07
17052	16089.06	0.06
22889	15628.25	0.06

Πίνακας 22, Lost Best Customers – Monetary Value

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Από τα παραπάνω είναι φανερό ότι παρόλο που οι συγκεκριμένοι πελάτες έχουν Monetary Score ίσο με ένα, ανήκουν δηλαδή στο τέταρτο τεταρτημόριο με τις υψηλότερες monetary values, δεν συμβάλουν με ιδιαίτερα μεγάλο ποσοστό στον συνολικό τζίρο της επιχείρησης. Το γεγονός αυτό μπορεί να εξηγηθεί από την διαπίστωση που έγινε στην αρχική ανάλυση του dataset ότι περίπου το 33% - 34% του συνολικού κύκλου εργασιών της, ανεξαρτήτως χώρας, προέρχεται από τους δέκα πρώτους σε πωλήσεις πελάτες. Σε αντίστοιχη διαπίστωση καταλήξαμε και για τον κύκλο εργασιών που προέρχεται μόνο από την Ελλάδα, όπου περίπου το 23% προέρχεται από τους πρώτους δέκα πελάτες. Ενδεικτικά να αναφέρουμε ότι ο πρώτος σε πωλήσεις πελάτης για την Ελλάδα συμβάλλει στον τζίρο σε ποσοστό 4,67% - συνολικές πωλήσεις 1.292.937,33€ και αντίστοιχα ο δέκατος σε πωλήσεις πελάτης σε ποσοστό 1,35% - συνολικές πωλήσεις 373.980,00€.

Lost Cheap Customers

Στην κατηγορία αυτή ανήκουν πελάτες, που όπως φαίνεται και από το όνομα της, έχουν δαπανήσει μικρά ποσά για τις αγορές τους από την επιχείρηση μας, ενώ ταυτόχρονα η συχνότητα αγορών τους είναι μικρή και δεν έχουν προβεί σε κάποια συναλλαγή με την εταιρεία πρόσφατα. Πρόκειται δηλαδή για πελάτες όπου όλα τα επιμέρους score του RFM Score ισούνται με τέσσερα. Το πλήθος αυτών των πελατών είναι 225 επιχειρήσεις. Επειδή πρόκειται για πελάτες οι οποίοι αποφέρουν το μικρότερο ποσοστό εσόδων στον κύκλο εργασιών της επιχείρησης, η εταιρεία θα μπορούσε να μην δώσει ιδιαίτερη προσοχή στην επικοινωνία μαζί τους και στην προσπάθεια να τους ανακτήσει. Παρόλα αυτά θα μπορούσε μέσω εξατομικευμένων ηλεκτρονικών μηνυμάτων να τους παρουσιάζει νέα προϊόντα, παρεμφερή με όσα αγόραζαν στο παρελθόν, ανοίγοντας έτσι ένα δίαυλο επικοινωνίας μαζί τους και αυξάνοντας τις πιθανότητες να απευθυνθούν ξανά σε εκείνη για την κάλυψη των αναγκών τους.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Παρακάτω παρουσιάζουμε για κάθε RFM Score που υπολογίσαμε τον μέσο όρο της μεταβλητής Monetary.

RFM Score	
111	71581.278493
112	7982.382157
121	30172.886286
122	6399.338929
123	1581.896667
131	28205.182500
132	4751.987692
133	1244.867419
134	403.353333
142	7837.500000
143	1071.800000
144	299.910714
211	51057.932895
212	7352.860250
213	2441.925000
221	41356.610000
222	5395.960256
223	1621.745588
231	21032.100000
232	4845.616800
233	1320.409403
234	403.741250
241	12520.000000
243	1399.169231
244	245.483333
311	36901.106471
312	7692.030000
321	28917.444167
322	5521.300462
323	1761.901190
331	42538.311429
332	4663.394857
333	1294.379560
334	362.332051
342	3022.500000
343	1186.893000
344	234.195000
411	27192.403333
421	32293.600000
422	5237.995789
423	1802.011429
431	15871.610000
432	4153.552414
433	1184.454490
434	386.295484
441	28000.000000
442	4226.598750
443	1160.967241
444	230.954622

Εικόνα 16, M.O monetary value for each RFM Score

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

5.4 Προτάσεις

Από την ανάλυση που προηγήθηκε παρατηρούμε ότι το μεγαλύτερο κομμάτι του κύκλου εργασιών της εταιρείας προέρχεται από τον κλάδο των αρωμάτων, ενώ αμέσως μετά ακολουθεί ο κλάδος των αγροχημικών, στον οποίο ανήκουν και οι περισσότεροι πελάτες της επιχείρησης. Το γεγονός αυτό είναι δικαιολογημένο καθώς ο κλάδος των αγροχημικών είναι πιο διευρυμένος σε σχέση με αυτόν των αρωμάτων, λόγω της φύσης των προϊόντων που προσφέρει η κάθε κατηγορία. Το άρωμα απευθύνεται σε μικρότερη αγορά από ότι τα αγροχημικά.

Ανάμεσα στα προϊόντα με τον υψηλότερο αριθμό πωλήσεων συναντάμε πολλά που ανήκουν στην κατηγορία των αγροχημικών, όπως επίσης τα συναντάμε και σε εκείνα τα τμήματα των πελάτων που έχουν περιθώρια ανάπτυξης, σε εκείνα δηλαδή τα τμήματα που είτε δεν αγοράζουν συχνά από την εταιρεία είτε η αξία των αγορών τους είναι σχετικά μικρή. Σε αυτούς τους πελάτες θα μπορούσε να εφαρμοστεί κάποια προωθητική ενέργεια μέσω ενός newsletter το οποίο ανάλογα την χρονική περίοδο θα τους προέτρεπε να χρησιμοποιήσουν συγκεκριμένα προϊόντα, τα οποία θα βελτίωναν την απόδοση της παραγωγής τους. Ακόμη, για τους πελάτες εκείνους που συνεισφέρουν σε μεγάλο ποσοστό στον κύκλο εργασιών της εταιρείας, η επιχείρηση θα μπορούσε να ακολουθήσει κάποιες εκπτωτικές πολιτικές, όπου για παράδειγμα να τους έκανε κάποια έκπτωση τζίρου, αν οι πωλήσεις σε αξία είχαν ξεπεράσει μια συγκεκριμένη τιμή.

Ακόμη, κατά την παραπάνω διαδικασία της τμηματοποίησης της αγοράς παρατηρήσαμε ότι σε ένα από τα segments που έχει προοπτικές ανάπτυξης, ανάμεσα στα προϊόντα με τις υψηλότερες πωλήσεις βρίσκονται δύο από τον κλάδο των καταναλωτικών προϊόντων. Το γεγονός αυτό μπορεί να σημαίνει ότι στην συγκεκριμένη αγορά υπάρχει χώρος για την επιχείρηση προκειμένου να διεκδικήσει μεγαλύτερο μερίδιο της. Ταυτόχρονα σε αυτούς τους πελάτες μπορεί να προτείνει παραπλήσια προϊόντα που να της αποφέρουν μεγαλύτερα έσοδα. Απ' την άλλη βέβαια το γεγονός ότι πρόκειται για προϊόντα που ναι μεν πωλούνται συχνά αλλά η συνολική αξία πώλησης είναι μικρή, μπορεί να σημαίνει ότι η επιχείρηση λειτουργεί με μικρό περιθώριο κέρδους στην συγκεκριμένη κατηγορία προϊόντος και άρα να πρέπει να εφαρμόσει βελτιωτικές κινήσεις προς αυτή την κατεύθυνση.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

6

ΤΕΧΝΙΚΕΣ ΛΕΠΤΟΜΕΡΕΙΕΣ

6.1 Λεπτομέρειες Υλοποίησης

Εφαρμόσαμε τον αλγόριθμο k-means πάνω σε δεδομένα πωλήσεων σχεδόν δύο ετών όπως περιγράψαμε παραπάνω. Καθώς πραγματοποιήθηκαν δύο διαφορετικές υλοποιήσεις με τον αλγόριθμο k-means, η μια με ποσοτικά δεδομένα και η άλλη με κατηγορικές μεταβλητές, τα δεδομένα πάνω στα οποία εφαρμόστηκε ο αλγόριθμος είχαν τροποποιηθεί κατάλληλα. Στην πρώτη περίπτωση των ποσοτικών δεδομένων υπολογίσαμε ανά πελάτη το σύνολο των τιμολογίων που έχουν εκδοθεί στο όνομα του, την συνολική αξία αυτών των τιμολογίων και τον μέσο όρο της αξίας ανά τιμολόγιο και πάνω σε αυτά τα δεδομένα υλοποιήθηκε ο αλγόριθμος. Στην δεύτερη περίπτωση των κατηγορικών μεταβλητών, η συσταδοποίηση έγινε με βάση την προϊοντική κατηγορία, στις οποίες απευθύνεται ο κάθε πελάτης. Στο σημείο αυτό να αναφερθεί ότι έγινε προσπάθεια να υλοποιηθεί ο αλγόριθμος λαμβάνοντας υπόψη του τον κωδικό του είδους και όχι την κατηγορία πώλησης. Όμως το αποτέλεσμα

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

που προέκυψε δεν ήταν ικανοποιητικό, καθώς δημιουργούσε πολλά μικρά clusters και ένα μεγάλο στο οποίο άνηκε η πλειοψηφία των πελατών και όλες οι κατηγορίες πώλησης, ενώ στα μικρότερα clusters υπήρχαν λιγότερες προϊοντικές κατηγορίες. Το μεγαλύτερο μέρος των πελατών της επιχείρησης προέρχεται από τον κλάδο των Αγροχημικών, περίπου το 56,7%. Ο αλγόριθμος ξεκινάει να δημιουργεί τα clusters, όπου όπως είναι λογικό στα περισσότερα συναντάμε την κατηγορία των Αγροχημικών (Παραγόμενα ή/και Αγοραζόμενα), ώσπου κάποια στιγμή δεν έχει άλλα clusters για να σπάσει και επομένως εντάσσει όλα τα υπόλοιπα σε ένα cluster. Από άποψη αλγόριθμού ενδεχομένως να έβγαζε ένα ικανοποιητικό output για έναν υπερβολικά μεγάλο αριθμό clusters. Όμως από επιχειρησιακή σκοπιά, αυτή η κίνηση δεν θα είχε νόημα για την επιχείρηση, καθώς δεν μπορεί να απευθύνεται και να στοχεύει σε υπερβολικά μεγάλο αριθμό τμημάτων πελατών. Ως εκ τούτου, επιλέχθηκε να γίνει η τμηματοποίηση με βάση την κατηγορία πώλησης, με την διαδικασία που αναφέρθηκε παραπάνω.

6.2 Πλατφόρμες και προγραμματιστικά εργαλεία

Όπως έχει ήδη αναφερθεί η ανάλυση των δεδομένων έγινε με χρήση της γλώσσας προγραμματισμού Python. Ακόμη, για την υλοποίηση του αλγορίθμου και της ανάλυσης χρησιμοποιήθηκε η διανομή της Python από την Anaconda (https://www.anaconda.com) και συγκεκριμένα το JupyterLab. Χρησιμοποιήθηκε η version της Pyrhon 3.7 για 64-bit λειτουργικό σύστημα.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

7

ΕΠΙΛΟΓΟΣ

Η παρούσα εργασία χρησιμοποίησε έναν από τους πλέον διαδεδομένους αλγορίθμους συσταδοποίησης, τον k-means, προκειμένου να δείξει πως μπορεί να συμβάλλουν οι τεχνικές data mining στην τμηματοποίηση της πελατειακής βάσης των επιχειρήσεων. Ο αλγόριθμος εφαρμόστηκε πάνω σε πραγματικά δεδομένα πωλήσεων χρονικού διαστήματος σχεδόν δύο ετών.

Κατά την υλοποίηση της ανάλυσης, προέκυψε ότι ο αλγόριθμος είναι απλός στην εφαρμογή του και εύχρηστος, ενώ η ταχύτητα υλοποίησης και εκτέλεσης του είναι χαμηλή. Επιπλέον τα αποτελέσματα που προέκυψαν θεωρήθηκαν αξιόπιστα και μπορούν να αξιοποιηθούν από την υπό εξέταση εταιρεία για την καλύτερη τοποθέτηση της απέναντι στους καταναλωτές και στόχευσης των κατάλληλων τμημάτων της αγοράς.

Στο μέλλον θα μπορούσε να εξεταστεί η υλοποίηση και απόδοση του αλγορίθμου, έχοντας ως κριτήριο παραπάνω από μια κατηγορική μεταβλητή. Και τέλος θα μπορούσαν να συγκριθούν τα αποτελέσματα που προέκυψαν από την χρήση του αλγορίθμου k-means με αυτά άλλων αλγορίθμων συσταδοποίησης, έτσι ώστε να αξιολογηθεί η απόδοση τους συγκριτικά.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

8

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ίδρυμα οικονομικών & βιομηχανικών ερευνών, Η Χημική Βιομηχανία στην Ελλάδα, Συνεισφορά στην οικονομία και προοπτικές ανάπτυξης, Απρίλιος 2018

Ελληνική Στατιστική Αρχή, Ελληνική Δημοκρατία, Έρευνα Παραγωγής και Πωλήσεων Βιομηχανικών Προϊόντων (PRODCOM), Έτους 2017

Πέτρος Κ. Μάλλιαρης, Καθηγητής Πανεπιστημίου Πειραιώς, Εισαγωγή στο Μάρκετινγκ, Γ ’έκδοση, Εκδόσεις Αθ. Σταμούλης

Συμπεριφορά Καταναλωτή & Στρατηγική Μάρκετινγκ, Γεώργιος Ι. Σιώμκος, Ph. D, Καθηγητής Μάρκετινγκ, Οικονομικό Πανεπιστήμιο Αθηνών, Β’ έκδοση, Εκδόσεις Αθ. Σταμούλης

Abratt R., Profesor of Marketing, Graduate School of Business, University of the Witwatersrand, Johanesburg, South Africa, Market Segmentation Practices of Industrial Marketers, Industrial Marketing Management, 1993, Elsevier Science Publishing Co.

Birant Deyya, Dokuz Eylul University Turkey, Data Mining Using RFM Analysis, Knowledge-Oriented Applications in Data Mining, 2011

Forsati R, Faculty of Electrical and Computer Engineering, Shahid Beheshti Univertity, Tehran, Iran, Mahdavi M., Department of Computer Engineering, Sharif University of Technology, Tehran, Iran,

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Shamsfard M., Department of Computer Engineering and IT, Amirkabir University of Technology. Tehran, Iran, Meybody M. R., Institute of Studies in Theoretical Physics and Mathematics, School of Computer Science, Tehran, Iran, Efficient Stohastic algorithms for document clustering, Information Science, 2013, elsevier.com

Hruschka E. R. & Covoes T. F., Catholic University of Santos (UniSantos), Brazil, Feature Selection for Cluster Analysis, an approach Based on the Simplified Silhouette Criterion, International Conference on Computational Intelligence for Modeling, Control and Automation & International Conference on Intelligent Agents, Wrb Technologies and Internet Commerce (doi:10.1109/cimca.2005.1631238)

Huang Z., ACSys CRC., CSIRO Mathematical and Information Sciences, Extensions to the k-Mean Algorithm for clustering large data sets with categorical Values, Data Mining and Knowledge Discovery, 1998, Kluwer Academic Publishers

Jing L., Ng M. K., Xu J, Huang J. Z., Subspace Clustering of Text Documents with Feature Weighting K-Means Algorithm, Department of Mathematics, The University of Hong Kong, E-Business Technology Institute, The University of Hong – Kong, Hong – Kong, China (doi:10.1007/11430919_94)

Kaufman L.,Vrije Universiteit Brussel, Brussels, Belgium, Rousseeuw P. , Universitaire Instelling Antwerpen, Antwerp, Belgium, Finding groups in data, An introduction to cluster analysis, A John Wiley & Sons, Inc., Publication, 1990

Kodinariya M. Trupti, Research Scholar in JJT University, Department of Computer Engineering, Jhunjhunu, Rajasthan, India, Dr. Makwana R. Prashant, Director – Research, GRMECT Research Center, Rajkot, India, Review on determining number of Cluster in K-Means Clustering, International Journal of Advance Research in Computer Science and Management Studies, Volume 1, Issue 6, November 2013

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Lleti R, Ortiz M. C., Department of Chemisty, Faculty of Sciences, University of Burgos, Spain, Sarabia L. A., Sanchez M. S., Department of Mathematics and Computer Science, University of Burgos, Spain, Selecting Variables for k-means cluster analysis by using a genetic algorithm that optimizes the silhouettes, Analytica Chemica Acta 515, 2004

Luo C, Li Y, M. S., Teradata Corporation, San Diego, CA, Li Y., Depatment of Computer Science and Information Science, Fordham University, Bronx, NY, USA, Chung S. M., Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA, Text Document Clustering Based on Neighbors, 2009, google scholar (doi:10.1016/j.datak.2009.06.007)

Pradeep Rai, Asst. Prof., CSE Department, Kanpur Institute of Technology (India), Shubha Singh, Asst. Prof., MCA Department, Kanpur Institute of Technology (India), A survey of Clustering Techniques, International Journal of Computer Application, October 2010

Satish Garla, Goutam Chakraborty, Oklahoma State University, Stillwater, OK, US Gary Gaeth, University of Iowa, Iowa City, Iowa, US, Comparison of K-means, Normal Mixtures and Probabilistic-D Clustering for B2B Segmentation using Customers' Perceptions, 2014

Singh V. K., Department of Computer Science, South Asian University, New Delhi – India, Tiwari N., Garg S., Department of Computer Science, Banaras Hindu University, Varanassi, India, Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means, 2011 International Conference on Computational Intelligence and Communication Systems

Singh A., Amity University Uttar Pradesh, India, Yadav A., Amity University Uttar Pradesh, India, Rana A., Amity University Uttar Pradesh, India, K-means with three different distance metrics, International Journal of Computer Application, Volume 67, April 2013

Sohrabi B., Assistant Professor, School of Management, University of Tehran, Khanlari A., PhD student of Marketing, School of Management, University of Tehran, Customer Lifetime Value (CLV) Measurement based on RFM Model, Iranian Accounting & Auditing Review, Vol. 14, Spring 2007

Steinley D., University of Missouri – Columbia, Brusco M.J., Florida State University, A new variable weighting and selection procedure for K-Means Cluster Analysis, Multivariate Behavioral Research, 2008 (doi: 10.1080/00273170701836695)

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Wei Jo-Ting, Department of Business Management, National Sun Yat-Sen University, Taiwan, Lin Shih-Yen, Department of Leisure Studies and Tourism Management, National Chi Nan University, Taiwan, Wu Hsin-Hung, Department of Business Administration, National Changhua University of Education, Taiwan, A review of the application of RFM Model, African Journal of Business Management, December Special Review 2010

Yada J., PG Student, CSE Department, M.D.U Rohtak, Haryana, India, Sharma M., Assistant Professor, IT Department, M.D.U Rohtak, Haryana, India, A review of K – means algorithm, International Journal of Engineering Trends and Technology (IJETT), Volume 4, Issue 7, July 2013

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΑΡΑΡΤΗΜΑ

ΚΩΔΙΚΑΣ

1. Ανάλυση Δεδομένων

```
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import MinMaxScaler
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
import os

dataset = pd.read_csv('C:/JoanDB.csv', sep =';' , decimal =',')

#εύρεση min & max της στήλης InvoiceDate
dataset.InvoiceDate.min()
dataset.InvoiceDate.max()

dataset17=dataset.loc[dataset.InvoiceYear == 2017]
dataset18 = dataset[dataset['InvoiceYear'] == 2018]

dataset.groupby(['InvoiceYear'])['Ammount'].sum()

dt17CT=dataset17.groupby(['CustomerType'])['Ammount'].sum().reset_index()
dt18CT=dataset18.groupby(['CustomerType'])['Ammount'].sum().reset_index()

fig, (ax1 , ax2)=plt.subplots(1 , 2, figsize=(11,6))
ax1.bar(dt17CT.CustomerType, dt17CT.Ammount, color = 'blue')
ax1.set_xlabel('CustomerType')
ax1.set_ylabel('Total Amount')
ax1.set_title('Total Sales per CustomerType in 2017')
ax2.bar(dt18CT.CustomerType, dt18CT.Ammount, color = 'orange')
ax2.set_xlabel('CustomerType')
ax2.set_ylabel('Total Amount')
ax2.set_title('Total Sales per CustomerType in 2018')
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
#percentage of Amount per customerType for 2017 & 2018 in donut charts

plt.pie(dt17CT.Ammount, labels = dt17CT.CustomerType, startangle=45, shadow = True, autopct='%.1f%%')
plt.title('2017')
circle=plt.Circle( (0,0), 0.7, color='white')
p=plt.gcf()
p.gca().add_artist(circle)

plt.pie(dt18CT.Ammount, labels = dt17CT.CustomerType, startangle=45, shadow = True, autopct='%.1f%%')
plt.title('2018')
circle=plt.Circle( (0,0), 0.7, color='white')
p=plt.gcf()
p.gca().add_artist(circle)

dt17Country=dataset17.groupby(['Country'])['Ammount'].sum().reset_index().sort_values(by='Ammount', ascending = False )
dt18Country = dataset18.groupby(['Country'])['Ammount'].sum().reset_index().sort_values(by='Ammount', ascending = False )

#ποσοστό του τζίρου της εταιρείας που οφείλεται σε κάθε χώρα

dt17Country['percentage'] = (dt17Country.Ammount / dt17Country.Ammount.sum() * 100).round(2)
dt18Country['percentage'] = (dt18Country.Ammount / dt18Country.Ammount.sum() * 100).round(2)

plt.figure(figsize=(11,6))
plt.bar(dt17Country.Country, dt17Country.Ammount)
plt.title('Total Sales per Country in 2017',  color = 'green' , fontsize='17', horizontalalignment='center')
plt.xlabel('Country' , color = 'purple' , fontsize='17', horizontalalignment='center')
plt.ylabel('Sum of Amount', color = 'purple' , fontsize='17', horizontalalignment='center')
plt.tick_params(axis='x', colors='green', rotation=90)
```

```
plt.figure(figsize=(11,6))
plt.bar(dt17Country.Country, dt17Country.Ammount)
plt.title('Total Sales per Country in 2017',  color = 'green' , fontsize='17', horizontalalignment='center')
plt.xlabel('Country' , color = 'purple' , fontsize='17', horizontalalignment='center')
plt.ylabel('Sum of Amount', color = 'purple' , fontsize='17', horizontalalignment='center')
plt.tick_params(axis='x', colors='green', rotation=90)

plt.figure(figsize=(11,6))
plt.bar(dt18Country.Country, dt18Country.Ammount)
plt.title('Total Sales per Country in 2018',  color = 'green' , fontsize='17', horizontalalignment='center')
plt.xlabel('Country' , color = 'purple' , fontsize='17', horizontalalignment='center')
plt.ylabel('Sum of Amount', color = 'purple' , fontsize='17', horizontalalignment='center')
plt.tick_params(axis='x', colors='green', rotation=90)

Customers_Country=dataset.groupby(['Country'])['CustomerCode'].nunique().reset_index().sort_values(by='CustomerCode', ascending=False)
Customers_Country.Country = Customers_Country.Country.str.strip()
Customers_Country.loc[(Customers_Country.Country == 'ΕΛΛΑΔΑ') | (Customers_Country.Country == 'ΕΛΒΕΤΙΑ') | (Customers_Country.Country == 'ΝΙΓΗΡΙΑ')]
Customers_Country.groupby(['CustomerCode'])['Country'].unique()
dt17Category = dataset17.groupby(['SalesCatI'])['Ammount'].sum().reset_index()

#ποσοστό τζίρου ανά προϊοντική κατηγορία σε pie charts για 2017 & 2018

plt.figure(figsize=(11,6))
plt.pie(dt17Category.Ammount, labels = dt17Category.SalesCatI, shadow = True, autopct='%.1f%%' )

dt18Category = dataset18.groupby(['SalesCatI'])['Ammount'].sum().reset_index()
plt.figure(figsize=(11,6))
plt.pie(dt18Category.Ammount, labels = dt18Category.SalesCatI, shadow = True, autopct='%.1f%%' )
```

```
dt17Customer=dataset17.groupby(['CustomerCode'])['Ammount'].sum().reset_index().sort_values(by='Ammount', ascending = False)
TSales17=sum(dataset17.Ammount)
#ποσοστό του τζίρου της εταιρείας που οφείλεται σε κάθε πελάτη για το 2017
dt17Customer['pctAmount'] = ((dt17Customer.Ammount / TSales17)*100).round(2)

Cus17 = dataset17[(dataset17.CustomerCode==15742) | (dataset17.CustomerCode==20442) | (dataset17.CustomerCode==15436) | (dataset17.CustomerCode==24139)
        | (dataset17.CustomerCode==15744) | (dataset17.CustomerCode==23829) | (dataset17.CustomerCode==19629) | (dataset17.CustomerCode==16838)
        | (dataset17.CustomerCode==15409) | (dataset17.CustomerCode==36498)]
top17=Cus17.groupby(['CustomerCode' , 'SalesCatI'])['Ammount'].sum().reset_index()
top17.pivot(index='CustomerCode' , columns='SalesCatI', values='Ammount').fillna(0).plot.bar(stacked=True, figsize=(12,7))
plt.xlabel('CustomerCode', fontsize='12')
plt.ylabel('Amount', fontsize='12')
plt.title('Amount of top 10 customers per SalesCatI for 2017', fontweight='bold', fontsize='12')

TSales18=sum(dataset18.Ammount)
#ποσοστό του τζίρου της εταιρείας που οφείλεται σε κάθε πελάτη για το 2018
dt18Customer['pctAmount'] = ((dt18Customer.Ammount / TSales18)*100).round(2)
dt18Customer=dataset18.groupby(['CustomerCode'])['Ammount'].sum().reset_index().sort_values(by='Ammount', ascending = False)
Cus18 = dataset18[(dataset18.CustomerCode==15742) | (dataset18.CustomerCode==20442) | (dataset18.CustomerCode==15744) | (dataset18.CustomerCode==15436)
        | (dataset18.CustomerCode==24139) | (dataset18.CustomerCode==19629) | (dataset18.CustomerCode==23043) | (dataset18.CustomerCode==22564)
        | (dataset18.CustomerCode==19051) | (dataset18.CustomerCode==17578)]
top18=Cus18.groupby(['CustomerCode' , 'SalesCatI'])['Ammount'].sum().reset_index()
top18.pivot(index='CustomerCode' , columns='SalesCatI', values='Ammount').fillna(0).plot.bar(stacked=True, figsize=(12,7))
plt.xlabel('CustomerCode', fontsize='12')
plt.ylabel('Amount', fontsize='12')
plt.title('Amount of top 10 customers per SalesCatI for 2018', fontweight='bold', fontsize='12')
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
Cus17.groupby(['Country'])['CustomerCode'].unique()
Cus18.groupby(['Country'])['CustomerCode'].unique()

#απαλοιφή των κενών στη στήλη country στα dataset17 & dataset18

dataset17.Country = dataset17.Country.str.strip()
dataset18.Country = dataset18.Country.str.strip()

#επιλογή των πρώτων 10 χωρών από τα datasets

TCountries17=dataset17.loc[(dataset17.Country == 'ΕΛΛΑΔΑ')| (dataset17.Country == 'ΝΙΓΗΡΙΑ') |(dataset17.Country == 'ΕΛΒΕΤΙΑ') | (dataset17.Country == 'ΣΕΡΒΙΑ')
                    |(dataset17.Country == 'ΡΩΣΙΑ') |(dataset17.Country == 'ΑΛΒΑΝΙΑ') |(dataset17.Country == 'ΡΟΥΜΑΝΙΑ') |(dataset17.Country == 'ΚΥΠΡΟΣ')
                    |(dataset17.Country == 'ΙΤΑΛΙΑ') |(dataset17.Country == 'Ν.ΚΟΡΕΑ')]

TCountries18=dataset18.loc[(dataset18.Country == 'ΕΛΛΑΔΑ')| (dataset18.Country == 'ΝΙΓΗΡΙΑ') |(dataset18.Country == 'ΕΛΒΕΤΙΑ') | (dataset18.Country == 'ΓΕΡΜΑΝΙΑ')
                    |(dataset18.Country == 'ΡΩΣΙΑ') |(dataset18.Country == 'ΠΟΛΩΝΙΑ') |(dataset18.Country == 'ΚΥΠΡΟΣ') |(dataset18.Country == 'Ν.ΚΟΡΕΑ')
                    |(dataset18.Country == 'ΣΕΡΒΙΑ') |(dataset18.Country == 'ΒΟΥΛΓΑΡΙΑ')]

Country_Cat17=TCountries17.groupby(['Country' , 'SalesCatI'])['Ammount'].sum().reset_index()
Country_Cat17.pivot(index='Country' , columns='SalesCatI', values='Ammount').fillna(0).plot.bar(stacked=True, figsize=(12,7))
plt.xlabel('Country', fontsize='12')
plt.ylabel('Amount', fontsize='12')
plt.title('Amount of top 10 Countries per SalesCatI for 2017', fontweight='bold', fontsize='12')
```

```
Country_Cat18=TCountries18.groupby(['Country' , 'SalesCatI'])['Ammount'].sum().reset_index()
TCountries18.groupby(['Country' , 'SalesCatI'])['Ammount'].sum()
Country_Cat18.pivot(index='Country' , columns='SalesCatI', values='Ammount').fillna(0).plot.bar(stacked=True, figsize=(12,7))
plt.xlabel('Country', fontsize='12')
plt.ylabel('Amount', fontsize='12')
plt.title('Amount of top 10 Countries per SalesCatI for 2018', fontweight='bold', fontsize='12')

#convert column InvoiceDate to datetime object - add as different column to datasets

dataset17['InvoiceDateN'] = pd.to_datetime(dataset17['InvoiceDate'] , infer_datetime_format=True)
dataset18['InvoiceDateN'] = pd.to_datetime(dataset18['InvoiceDate'] , infer_datetime_format=True)

#extract month from column InvoiceDateN for each dataset

dataset17['InvoiceMonth'] = dataset17['InvoiceDateN'].dt.month
dataset18['InvoiceMonth'] = dataset18['InvoiceDateN'].dt.month

dt17M=dataset17.groupby(['SalesCatI' , 'InvoiceMonth'])['Ammount'].sum().reset_index()
dt17Month=dt17M.pivot(index = 'InvoiceMonth', columns = 'SalesCatI', values='Ammount')
dt18M=dataset18.groupby(['SalesCatI' , 'InvoiceMonth'])['Ammount'].sum().reset_index()
dt18Month=dt18M.pivot(index = 'InvoiceMonth', columns = 'SalesCatI', values='Ammount')

dt17Month.plot(figsize=(10,5))
plt.xlabel('Month')
plt.ylabel('Amount')
plt.title('Total Sales per Month per Sales Category for 2017', fontweight='bold')
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5), prop={'size':8})
```

```
dt18Month.plot(figsize=(10,5))
plt.xlabel('Month')
plt.ylabel('Amount')
plt.title('Total Sales per Month per Sales Category for 2018', fontweight='bold')
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5), prop={'size':8})
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

2. Clustering

```
#CLUSTERING

dtClust = dataset.groupby(['CustomerCode']).agg (
            {
                'Invoice': 'nunique',
                'Ammount': sum
            }
        )

dtClust.columns = ['Count of Invoices' , 'Total Sales']

dtClust['avgInvAmount'] = (dtClust['Total Sales'] / dtClust['Count of Invoices']).round(2)

#κανονικοποίηση των δεδομένων

rank_dt = dtClust.rank(method='first')
normalized_dtClust = (rank_dt - rank_dt.mean()) / rank_dt.std()
normalized_dtClust.head(5)

#Silhouette

for n_cluster in [4,5,6,7,8]:
    kmeans = KMeans(n_clusters=n_cluster).fit(
        normalized_dtClust[['Total Sales', 'Count of Invoices', 'avgInvAmount']])

    silhouette_avg = silhouette_score(
        normalized_dtClust[['Total Sales', 'Count of Invoices', 'avgInvAmount']],
        kmeans.labels_)

    print('Silhouette Score for %i Clusters: %0.4f' % (n_cluster, silhouette_avg))
```

```
#Elbow Method

sse = []
krange = list(range(2,11))
X = normalized_dtClust[['Total Sales', 'Count of Invoices', 'avgInvAmount']].values
for n in krange:
    model = cluster.KMeans(n_clusters=n, random_state=3)
    model.fit_predict(X)
    cluster_assignments = model.labels_
    centers = model.cluster_centers_
    sse.append(np.sum((X - centers[cluster_assignments]) ** 2))

plt.plot(krange, sse)
plt.title('SSE regarding number of Clusters')
plt.xlabel("$K$")
plt.ylabel("Sum of Squares")
plt.show()

kmeans = KMeans(n_clusters=4).fit(normalized_dtClust[['Total Sales', 'Count of Invoices', 'avgInvAmount']])

four_clusters = normalized_dtClust[['Total Sales', 'Count of Invoices', 'avgInvAmount']].copy(deep=True)
four_clusters['Cluster'] = kmeans.labels_

four_clusters.head(5)
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
cluster1_metrics = kmeans.cluster_centers_[0]
cluster2_metrics = kmeans.cluster_centers_[1]
cluster3_metrics = kmeans.cluster_centers_[2]
cluster4_metrics = kmeans.cluster_centers_[3]

data = [cluster1_metrics, cluster2_metrics, cluster3_metrics, cluster4_metrics]
cluster_center = pd.DataFrame(data)

cluster_center.columns = four_clusters.columns[0:3]
cluster_center
```

```
plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 0]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 0]['Total Sales'],
    c='blue')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 1]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 1]['Total Sales'],
    c='red')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 2]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 2]['Total Sales'],
    c='orange')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 3]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 3]['Total Sales'],
    c='green')

plt.title('Total Sales vs. Count of Invoices')
plt.xlabel('Count of Invoices')
plt.ylabel('Total Sales')

plt.grid()
plt.show()
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 0]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 0]['avgInvAmount'],
    c='blue')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 1]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 1]['avgInvAmount'],
    c='red')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 2]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 2]['avgInvAmount'],
    c='orange')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 3]['Count of Invoices'],
    four_clusters.loc[four_clusters['Cluster'] == 3]['avgInvAmount'],
    c='green')

plt.title('avg Invoice Amount vs. Count of Invoices')
plt.xlabel('Count of Invoices')
plt.ylabel('avg Invoice Amount')

plt.grid()
plt.show()
```

```
plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 0]['Total Sales'],
    four_clusters.loc[four_clusters['Cluster'] == 0]['avgInvAmount'],
    c='blue')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 1]['Total Sales'],
    four_clusters.loc[four_clusters['Cluster'] == 1]['avgInvAmount'],
    c='red')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 2]['Total Sales'],
    four_clusters.loc[four_clusters['Cluster'] == 2]['avgInvAmount'],
    c='orange')

plt.scatter(
    four_clusters.loc[four_clusters['Cluster'] == 3]['Total Sales'],
    four_clusters.loc[four_clusters['Cluster'] == 3]['avgInvAmount'],
    c='green')

plt.title('avg Invoice Amount vs. Total Sales Clusters')
plt.xlabel('Total Sales')
plt.ylabel('avg Invoice Amount')

plt.grid()
plt.show()
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
high_value = four_clusters.loc[four_clusters['Cluster'] == 3]

pd.DataFrame(dataset.loc[dataset['CustomerCode'].isin(high_value.index)].groupby(
    'ItemCode').count()['ItemNameHash'].sort_values(ascending=False).head())

mid_value = four_clusters.loc[four_clusters['Cluster'] == 1]

pd.DataFrame(dataset.loc[dataset['CustomerCode'].isin(mid_value.index)].groupby(
    'ItemCode').count()['ItemNameHash'].sort_values(ascending=False).head())

mid_value2 = four_clusters.loc[four_clusters['Cluster'] == 2]

pd.DataFrame(dataset.loc[dataset['CustomerCode'].isin(mid_valu2e.index)].groupby(
    'ItemCode').count()['ItemNameHash'].sort_values(ascending=False).head())

high_value.reset_index().CustomerCode.nunique()
mid_value.reset_index().CustomerCode.nunique()
mid_value2.shape
low_value.reset_index().CustomerCode.nunique()
```

```
#based on product
#για το clustering δεν χρειαζόμαστε Salesman, CustomerType, CustomerHashName, ItemHashName, Invoice, InvoiceType, InvoiceDate, InvoiceYear, ItemCode

Salesclust= dataset[['Country' , 'CustomerCode' , 'SalesCatI' , 'Quantity' , 'Ammount']]
Sclust=Salesclust.groupby('CustomerCode')['SalesCatI'].apply(lambda l: ' '.join([str(e) for e in list(l)]))

Sclust.shape

tfidf_vectorizer1 = TfidfVectorizer()
tfidf1 = tfidf_vectorizer1.fit_transform(Sclust)

kmeans = KMeans(n_clusters=4).fit_predict(tfidf1)
Sclust = pd.DataFrame(Sclust)
Sclust['cluster'] = kmeans
Sclust['cluster'].hist()
```

```
Cluster0 = Sclust.loc[Sclust['cluster'] == 0]
Cluster1 = Sclust.loc[Sclust['cluster'] == 1]
Cluster2 = Sclust.loc[Sclust['cluster'] == 2]
Cluster3 = Sclust.loc[Sclust['cluster'] == 3]

Clients0=pd.DataFrame(dataset.loc[dataset['CustomerCode'].isin(Cluster0.index)])
Clients1=pd.DataFrame(dataset.loc[dataset['CustomerCode'].isin(Cluster1.index)])
Clients2=pd.DataFrame(dataset.loc[dataset['CustomerCode'].isin(Cluster2.index)])
Clients3=pd.DataFrame(dataset.loc[dataset['CustomerCode'].isin(Cluster3.index)])
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

3. RFM ANALYSIS

```
#RFM ANALYSIS

dtCC=dataset.groupby(['Country'])['CustomerCode'].nunique().sort_values().reset_index().sort_values(by='CustomerCode' , ascending = False)

#percentage of unique count of customers per country
dtCC['pctCustomers'] = dtCC['CustomerCode']/dtCC['CustomerCode'].sum() * 100

#σχεδόν το 90% των πελατών βρίσκεται στην Ελλάδα, η ανάλυση μας επικεντρώνεται εκεί.

dataset.Country = dataset.Country.str.strip()
greece_data = dataset.loc[dataset.Country == 'ΕΛΛΑΔΑ']

#for RFM Analysis, columns: CustomerCode, Invoice, InvoiceDate, Quantity, Ammount

greece_data = greece_data[['CustomerCode' , 'Invoice' , 'InvoiceDate' , 'Quantity' , 'Ammount']]

greece_data['InvoiceDate'].max()
greece_data['InvoiceDate'].min()

greece_data['InvoiceDate'] = pd.to_datetime(greece_data['InvoiceDate'])

now = dt.datetime(2019, 9, 30)

rfm = greece_data.groupby(['CustomerCode']).agg(
       {
           'InvoiceDate':lambda date:(now-date.max()).days ,
           'Invoice': 'nunique' ,
           'Ammount': sum
        }
       )
rfm.columns
rfm.columns=['Regency' , 'Frequency' , 'Monetary']
rfm.Regency = rfm.Regency.astype(int)

#Computing Quantile of RFM Analysis
#1 highest/best value, 4 lowest/worst value

rfm['r_quantile'] = pd.qcut(rfm['Regency'] , 4 , [1 , 2 , 3 , 4])
rfm['f_quantile'] = pd.qcut(rfm['Frequency'], 4 , [4 , 3 , 2 , 1])
rfm['m_quantile'] = pd.qcut(rfm['Monetary'], 4 , [4 , 3 , 2 , 1])

rfm['RFM Score'] = rfm.r_quantile.astype(str) + rfm.f_quantile.astype(str) + rfm.m_quantile.astype(str)

rfm[rfm['RFM Score']=='111'].sort_values('Monetary', ascending=False).head()

print("Best Customers: ",len(rfm[rfm['RFM Score']=='111']))
print('Loyal Customers: ',len(rfm[rfm['f_quantile']==1]))
print("Big Spenders: ",len(rfm[rfm['m_quantile']==1]))
print('Almost Lost: ', len(rfm[rfm['r_quantile']== 3]))
print('Almost Lost Best Customers: ', len(rfm[rfm['RFM Score']=='311']))
print('Lost Customers: ',len(rfm[rfm['r_quantile']== 4]))
print('Lost Best Customers: ',len(rfm[rfm['RFM Score']=='411']))
print('Lost Cheap Customers: ',len(rfm[rfm['RFM Score']=='444']))

rfm.groupby(['RFM Score'])['Monetary'].mean()

rfm.loc[rfm['RFM Score'] == '311'].sort_values('Monetary', ascending=False)

rfm.loc[rfm['RFM Score'] == '411'].sort_values('Monetary', ascending=False)
```

```
sum(greece_data.Ammount)

rfm.sort_values('Monetary', ascending = False).head(10)
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ