ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

**SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY**

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

# Fast-moving consumer goods (FMCG) product similarities evaluation and visualization based on customer transaction sales

By

Frideriki D. Kostopoulou

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece
August 2021


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


**ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ**

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

# Αξιολόγηση και οπτικοποίηση της ομοιότητας καταναλωτικών αγαθών που καταναλώνονται γρήγορα FMCG με βάση τις συναλλαγές των πελατών

Φρειδερίκη Δ. Κωστοπούλου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα
Αύγουστος 2021


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Acknowledgements

I would like to express my sincere appreciation to each one of the professors of this Master's program for their important guidance and knowledge in my studies. Especially, I would like to express my sincere gratitude to my supervisor, professor Panagiotis Papastamoulis, for taking the time to help me, and provide me the right tools in order to successfully complete my dissertation. Also, I would like to thank Director-Innovation at IRI Vasileios Georgiou and Data Scientist - R&D Senior Manager at IRI Stavroula Poulopoulou for their insightful comments and their valuable help throughout my research.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Abstract

Frideriki Kostopoulou

## Fast-moving consumer goods (FMCG) product similarities evaluation and visualization based on customer transaction sales

August 2021

In the retail, market basket analysis is a set of statistical affinity calculations that facilitate shop owners more completely comprehend, and ultimately serve, their clients by highlighting purchasing patterns. The idea is that combinations of products most frequently occur together in orders, namely, customers who purchase a certain item (or group of items) are more likely to purchase another specific item (or group of items). The goal of this dissertation is to discover and visualize possible measures of similarity of FMCG products in the category of “juices” based on the annual transaction data of the customers. Our approach is based on examining pairs of goods and constructing models that we could evaluate them. After construction of models, in the second part, we classify both juices and consumers into reasonable groups. Finally, it is presented the overall conclusions of the dissertation.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Περίληψη

Φρειδερίκη Κωστοπούλου

## Αξιολόγηση και οπτικοποίηση της ομοιότητας καταναλωτικών αγαθών που καταναλώνονται γρήγορα FMCG με βάση τις συναλλαγές των πελατών

Αύγουστος 2021

Στο λιανεμπόριο, η ανάλυση καλαθιού αγοράς είναι ένα σύνολο στατιστικών υπολογισμών συνάφειας που διευκολύνουν τους ιδιοκτήτες καταστημάτων να κατανοήσουν πλήρως και τελικά να εξυπηρετήσουν τους πελάτες τους, επισημαίνοντας τα πρότυπα αγορών. Η ιδέα είναι ότι οι συνδυασμοί προϊόντων συμβαίνουν συχνότερα μαζί σε παραγγελίες, δηλαδή οι πελάτες που αγοράζουν ένα συγκεκριμένο είδος (ή ομάδα ειδών) είναι πιο πιθανό να αγοράσουν ένα άλλο συγκεκριμένο είδος (ή ομάδα ειδών). Ο στόχος αυτής της διπλωματικής εργασίας είναι να ανακαλύψει και να απεικονίσει πιθανά μέτρα ομοιότητας των προϊόντων FMCG στην κατηγορία «χυμοί» με βάση τα ετήσια δεδομένα συναλλαγών των πελατών. Η προσέγγισή μας βασίζεται στην εξέταση ζευγαριών αγαθών και στην κατασκευή μοντέλων που θα μπορούσαμε να τα αξιολογήσουμε. Μετά την κατασκευή μοντέλων, στο δεύτερο μέρος, κατατάσσουμε τόσο τους χυμούς όσο και τους καταναλωτές σε λογικές ομάδες. Τέλος, παρουσιάζονται τα συνολικά συμπεράσματα της διατριβής.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Table of Contents





ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# List of tables





ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# List of figures





ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Chapter 1

# Introduction

## 1.1 Introduction

This study aims to evaluate and visualize Fast-Moving Consumer Goods (FMCG) product similarities based on customer transaction sales. Fast Moving Consumer Goods are goods that are consumed at regular intervals by the average consumer and are changed over a period of days, weeks, months, and for one year.

Fast Moving Consumer Goods, also known as Consumer-Packaged Goods (CPG), are sold quickly at relatively low cost and do not require a lot of thought and time to purchase. These products generally are sold daily in large numbers and because of that, the cumulative profit on such products can be large. The Fast-Moving Consumer Goods Industry supply food and non-food everyday consumer goods to clients. Examples include fruits and vegetables, toilet paper, meat, dairy products, soft drinks etc. Purchasing of these consumables occurs at grocery stores, supermarkets, warehouse outlet etc. and are supported by the manufacturers, using advertising and promotion, most common on TV and in internet. Also, creativity and innovation are keys to generation of new ideas of improving goods for customer satisfaction.

As fast-moving consumer goods have a high turnover rate, the market is very competitive. Retailers need to focus their efforts on marketing to entice consumers to buy their products. They collect data about purchasing patterns, recording purchase data as item barcodes are scanned by point-of-sale systems. These purchasing patterns help in understanding the needs of their customers. Statistical models could look for co-occurrence in this data to determine which products are most likely to be purchased together. Finally, a retailer could adjust marketing and sales strategy to be beneficial in total. It is great to mention that the fast-moving consumer goods (FMCG) sector contributes a lot to the growth of India's GDP (Gross domestic product is a monetary measure of the market value of all the final goods and services produced in a specific period).



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## 1.2 Market Basket Analysis

Market Basket Analysis is one of the known key techniques that used by large retailers to discover associations between items. Its scope is to find combinations of items that both are purchased often in transactions.

Considering the field of retail, it is provided large amount of goods and the management is necessary to make decision, for instance what to put on sale, how to design discount vouchers, how to place merchandise on shelves to maximize their profit etc. A commonly used approach is to analyze transactions of past. Each customer has a basket data that is created from the bar-code of items that are purchased over a period, daily, monthly, yearly etc. There are several organizations that collect massive amounts of such data, like IRI, Market research company.

Mining association rules was first introduced by Agrawal (1993). He also discovered rules that had one product in the consequent and a union of a number of products in the antecedent. His research resulted in itemsets with their respective support count, meaning the % of transactions where these items were bought together.

Association rule mining's aim is to extract interesting correlations, frequent patterns, associations, or casual structures among sets of items in the transaction databases. Association rules are widely used in various areas like telecommunication networks, market, and risk management, in medical diagnosis etc.

The procedure of using association rules is often referred as "association rule mining" or "mining associations". In association rules we have two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data and a consequent is an item found in combination with the antecedent.

Searching data for regular patterns of if-then, it is generated association rules and criteria like support and confidence are used to check the validity of the association rules. Support is used for removing of no interest rule as low support rule occur just by chance and confidence provides the reliability of an association rule. Furthermore, a metric, called lift, can be used to compare confidence with expected confidence, in other words how many times an if-then statement is expected to be found true.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

An example of association rule (AR) is the following:

$$rusk, butter \rightarrow jam\ (0.01, 0.8)$$

This association rule leads us to the conclusion that if a customer buys rusk and butter, there is an 80% chance that he also buys jam at that time. Furthermore, 1% of all market-baskets contain all three items/products. In general, an association rule can be presented as following:

$$A, B \rightarrow C\ (support, confidence)$$

This means that those customers who purchased A and B will also purchase C with probability confidence and items A, B and C will be purchased all together with probability support.

In association rule mining, item sets that are made up of two or more items, are used. It is necessary to mention that more than 10.000 rules are generated in the industry application.

Many techniques have been proposed to mine frequent patterns with the most famous be Apriori Algorithm. This algorithm has been developed by Agarwal and Srikant in 1994. The principle of Apriori Algorithm is "If an item set is frequent, then all of its subsets must also be frequent". Apriori uses the support value to limit the number of candidate itemset. In essence, instead of examining all possible sets of itemsets, pruning procedure is done and thus smaller groups are created items that often appear together.

First of all, it is calculated the support value (the frequency of each itemset individually in the dataset) of all items in the database. A set of itemsets is created which contains the itemsets that have a support value equal to or greater than the minimum support value set. This set is defined as $L_k$, where $k = 1$ because each itemset contains an object.Then, from the itemsets of $L_k$, a set of itemsets are created that contain $k + 1$ items. This candidate set is denoted by $C_{k+1}$, which is created from the union of $L_k$, with itself. If there is even an infrequent set of objects $k$ within $C_{k+1}$, it is deleted. Furthermore, the support value is recalculated for all members of $C_{k+1}$ and if any of them are found have a lower support value than the set minimum value, then is deleted and the process for finding $L_{k+1}$, is completed. The above process is repeated, until $L_{k+1}$ is not empty.

Now, we will apply Apriori Algorithm in an example. Suppose we have the following dataset that has several transactions. From this dataset which consists of A, B, C, D, E items, we want to find the frequent itemsets and generate the association rules using the Apriori algorithm.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

| TID | Itemsets |
|---|---|
| T1 | A, B |
| T2 | B, D |
| T3 | B, C |
| T4 | A, B, D |
| T5 | A, C |
| T6 | B, C |
| T7 | A, C |
| T8 | A, B, C, E |
| T9 | A, B, C |

Given minimum support 2 and minimum confidence 50%

Step-1: Calculating $C_1$ and $L_1$:

In the first step, we will create a table that contains support count of each itemset in the given dataset. This table is called the Candidate set or $C_1$

| Itemset | Support count |
|---|---|
| A | 6 |
| B | 7 |
| C | 5 |
| D | 2 |
| E | 1 |

Now, we will take out all the itemsets that have the greater support count that the Minimum Support 2. It will give us the table for the frequent itemset $L_1$

| Itemset | Support count |
|---|---|
| A | 6 |
| B | 7 |
| C | 5 |
| D | 2 |

Step-2: Candidate Generation $C_2$, and $L_2$:

In this step, we will generate $C_2$with the help of $L_1$. In $C_2$, we will create the pair of the itemsets of $L_1$ in the form of subsets.

Creating the subsets, we will again find the support count from the main transaction table of datasets, i.e., how many times these pairs have occurred together in the given dataset. We will get the below table for $C_2$:



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

| Itemset | Support count |
|---|---|
| {A, B} | 4 |
| {A, C} | 4 |
| {A, D} | 1 |
| {B, C} | 4 |
| {B, D} | 2 |
| {C, D} | 0 |

Now, we need to compare the $C_2$ Support count with the minimum support count, and after comparing, the itemset with less support count will be eliminated from the table $C_2$. It will give us the below table for $L_2$.

| Itemset | Support count |
|---|---|
| {A, B} | 4 |
| {A, C} | 4 |
| {B, C} | 4 |
| {B, D} | 2 |

Step-3: Candidate Generation $C_3$and $L_3$:

We will repeat the same two processes, but now we will form the $C_3$, table with subsets of three itemsets together A, B, C, and will calculate the support count from the dataset. It will give the below table:

| Itemset | Support count |
|---|---|
| {A, B, C} | 2 |
| {B, C, D} | 1 |
| {A, C, D} | 0 |
| {A, B, D} | 0 |

As we can see from the above $C_3$ table, there is only one combination of itemset that has support count equal to the minimum support count. So, the $L_3$ will have only one combination, i.e., {A, B, C}

Step-4: Finding the association rules for the subsets:

We will create a new table with the possible rules from the occurred combination {A, B,C}. For all the rules, we will calculate the Confidence using formula sup(A^B)/A. After calculating the confidence value for all rules, we will exclude the rules that have less confidence than the minimum confidence, 50%. Using the following:

$$Confidence(A \rightarrow B) = \frac{Support\_count(A \cup B)}{Support\_count(A)}$$



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

| Rules | Support | Confidence |
|---|---|---|
| A, B → C | 2 | Sup{(A, B, C}/sup(A,B)= 2/4=0.5=50% |
| B, C → A | 2 | Sup{(B, C, A}/sup(B, C)= 2/4=0.5=50% |
| A, C → B | 2 | Sup{(A, C, B}/sup(A, C)= 2/4=0.5=50% |
| C→ A, B | 2 | Sup{(C, A, B)}/sup(C)= 2/5=0.4=40% |
| A→ B, C | 2 | Sup{(A, B, C)}/sup(A)= 2/6=0.33=33.33% |
| B→ B, C | 2 | Sup{(B, B, C)}/sup(B)= 2/7=0.28=28% |

Finally, as the given minimum confidence is 50%, the first three rules:

1. A, B → C
2. B, C → A
3. A, C → B

can be considered as the strong association rules for the given example.

## 1.3 Thesis outline

The purpose of this thesis is to find various possible measures of similarity (distance) of FMCG products that we could calculate them based on the transaction data of the customers. We apply these measures, and we show how they could be presented graphically.

In chapter 2, are introduced the descriptive analysis of the data set and the construction of model that can discover similarities-differences between products.

In chapter 3, are presented the concepts of clustering as in our FMCG products as in customers.

In chapter 4, are implemented the model construction that we abstracted in chapter 2. Clustering algorithms will be applied and evaluated to the prepared dataset to cluster the FMCG products and customers too in chapter 4.

Finally, the chapters end with chapter 5 by comparing and summarizing the results.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Chapter 2

## Description of the data and model construction

### 2.1 Description of the Data

Our research is based on data provided by IRI. The period under examination is from 3 March 2019 to 1 March 2020 on a daily basis, so we have sales data for one year.

Data contains the following variables:

1. PRODUCT_ID: Contains the id of the product
2. STORE: Indicates the id of the store
3. CUSTOMER_ID: Contains the id of the customer which is unique for each customer
4. DATE: Indicates the date of purchase
5. UNITS: Indicates the units of products that are purchased.

The data set consists of 8.669.400 transactions of 433.470 customers, sales of 20 products, that belong to product category “juices”. The 20 items are the following:



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

| PRODUCT ID | DESCRIPTION |
|---|---|
| 93 | Brand A Orange Juice 1L |
| 78 | Brand A Grapefruit Juice 1L |
| 83 | Brand A Apple Juice 1L |
| 33 | Brand A Clementine Juice 1L |
| 32 | Brand A Ananas Juice 1L |
| 4 | Brand A Concentrated Lemon Juice 1L |
| 113 | Brand B Orange Juice without pulp PET[1] packaging 90CL |
| 111 | Brand B Orange Juice with pulp PET packaging 90CL |
| 90 | Brand B Apple Juice PET packaging 90CL |
| 103 | Brand D Orange Juice 1L |
| 95 | Brand A Orange Juice BP[2] 1,50L |
| 50 | Brand C Mixed Fruit Juice 900ML |
| 21 | Brand C Orange/Carotte Juice 900ML |
| 47 | Brand C Bilberry/Black Currant/Cranberry Juice 900ML |
| 54 | Brand C Pear/Apple/Peach Juice 900ML |
| 101 | Brand C Orange Juice without pulp 900ML |
| 49 | Brand C Apple/Raspberry 900ML |
| 46 | Brand C Ananas/Passion Fruit Juice 900ML |
| 122 | Brand B Orange Juice without pulp 1 L |
| 123 | Brand B Orange Juice 1 L |

*Table 1: Products of our sample*

Firstly, "STORE" refers to the id of stores and this variable did not carry any meaningful information for this analysis and was removed. Each customer id, namely each customer, has a sequence of transactions of 20 products (variable "UNITS") and a basket is generated. The basket is necessary to be sorted by the variable "DATE" as we clearly discover customer behavior. Also, transactions whose number is less than 20 items were removed. Finally, our data contains 22.029 customers with their transactions of 20 items.

[1] Polyethylene terephthalate, often abbreviated PET, is the most common thermoplastic polymer resin of the polyester family and is used in fibres for clothing, packaging for liquids and foods. Pet bottles are shatterproof, recyclable and produces less product waste.

[2] Plastic Bottle



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The following barplot provides the purchase frequencies of our sample.

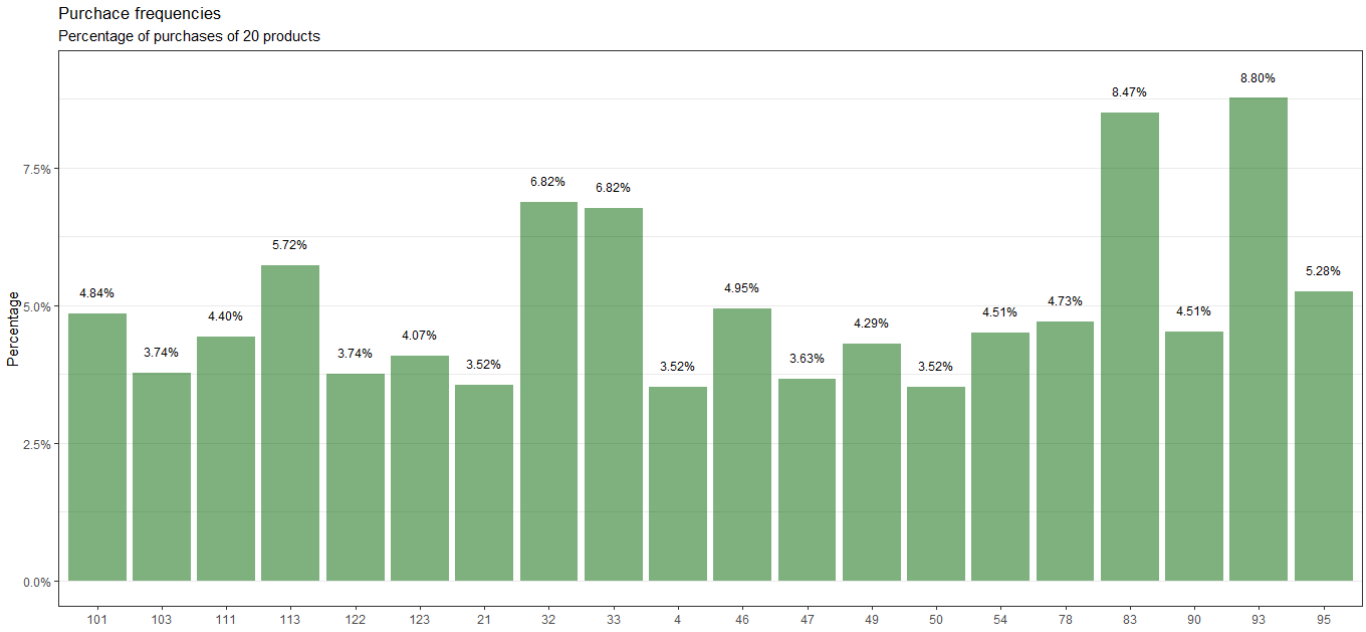



*Figure 1: Purchase frequencies*

As it is ascertained from the histogram above, the 8.8% of the purchases includes the product of Brand A Orange Juice 1L (*93*) and the 8.5% includes the product of Brand A Apple Juice (*83*). Besides, the sample was consisted of 3.5% of the purchases of product of Brand C Orange/Carotte Juice 900ML (*21*) and the product of Brand A concentrated Lemon Juice 1L (*4*).

Figure 2 illustrates how many units are purchased in each item (using the variable UNITS)

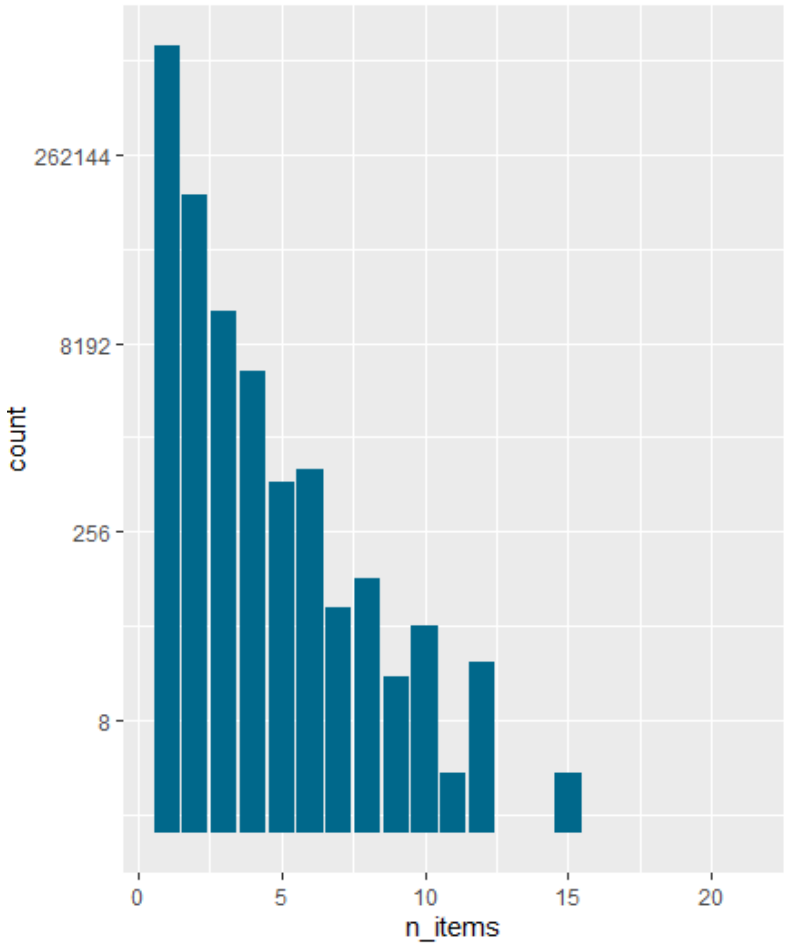



*Figure 2: Number of units*

Customers mostly purchased less than 2 units of each product.

Since our aim is to discover similarities between items in the category of juices, our analysis based on pairs of purchases from 20 products. In the following figure, it is presented the total number of customers who purchased each pair of items and the frequency of items in the diagonal.

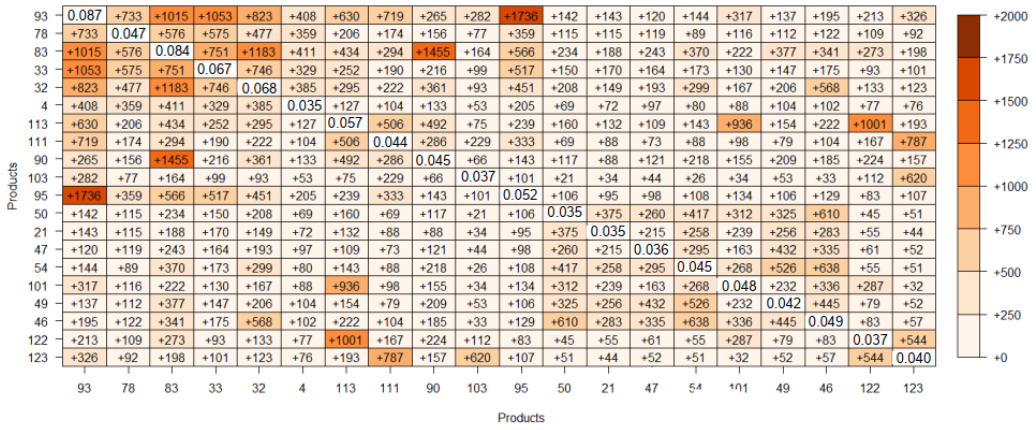



*Figure 3: Number of customers purchased each pair of items & frequencies of items*

It is clearly noticed that product of Brand A Orange Juice 1L (*93*) and product of Brand A Orange Juice BP 1,50L (*95*) are more purchased and then product of Brand A Apple Juice 1L (*83)* and Brand B Apple Juice PET packaging 90CL *(90)*, 1.736 and 1.455 customers bought them, respectively. At once, we could deduce that there is a preference in Brand A.

## 2.2 Model construction

We try to compare pairs of purchases from 20 products in the category of "juices". For instance (A, B), it is considered as "1" the purchase of product A and "0" the purchase of product B. Also, we consider $X_i$ transactions of product A and product B at the time $i = 1, \dots, n$, $n$ depends on $i$, like a stream of binary trials, with possible outcomes, 1 or 0. Suppose that there is $t = 2, 3, \dots, n-1$ when it is observed a switching in the shopping habit, with the preference in a specific product among A, B. This procedure is applied in every pair of 20 products of our analysis. Also, every customer has a different stream of purchases of the pairs, i.e., $n$ is not constant for each customer.

In the following figure, we present an example of stream of binary trial among products A, B. The total number of transactions of A, B is 30.

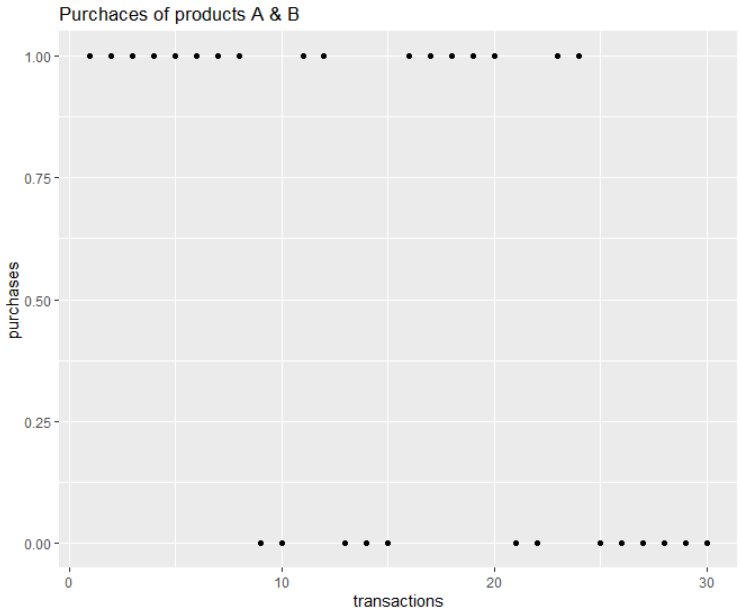



*Figure 4: Basket of purchases items A, B*

Looking the above figure, we notice that in the first 8 transactions is a preference for one item over other. In the next purchases, customer made switching from item A to B and opposite until the last purchases of one item.

At $i = 1, \dots, t$, suppose that $X_i$ is a random variable from the Bernoulli distribution with unknown parameter $p_1 \in [0,1]$. At $i = t+1, \dots, n$, suppose that $X_i$ is a random variable from the Bernoulli distribution with unknown parameter $p_2 \in [0,1]$. In this case, we imply that there is the switching model at a time $t$. We assume that $X_i$ are conditionally independent in the two situations. In other words, it is considered that in every t, namely in each transaction of a costumer, we have different models, the no switching and the switching models.

The probability mass function of this distribution is given by:

$$P(X = x) = p^x(1-p)^{(1-x)}, x \in [0,1] \tag{2.1}$$

The cumulative distribution function is 0 if $x < 0$, $1-p$ if $0 \leq x < 1$, and 1 if $x \geq 1$. The mean and the variance of the distribution are $p$ and $p(1-p)$, respectively.

The likelihood is:

$$\begin{aligned} L(p_1, p_2|X, t) &= \prod_{i=1}^{t} {p_1}^{x_i}(1-p_1)^{1-x_i} \prod_{i=t+1}^{n} {p_2}^{x_i}(1-p_2)^{1-x_i} \\ &= {p_1}^{\sum_{i\leq t} x_i}(1-p_1)^{t-\sum_{i\leq t} x_i} {p_2}^{\sum_{i>t} x_i}(1-p_2)^{(n-t-\sum_{i>t} x_i)} \end{aligned} \tag{2.2}$$

$$0 < p_1 < 1$$

$$0 < p_2 < 1$$

$$Given\ X, t$$

where $\boldsymbol{x_i} = \left(x_1, x_2, \dots, x_{n_i}\right) \in \{0,1\}^{n_i}$

The log-likelihood is:

$$\begin{aligned} l(p_1, p_2) &= logL(p_1, p_2|X, t) \\ &= \sum_{i\leq t} x_i log p_1 + (t - \sum_{i\leq t} x_i)\log(1-p_1) \\ &+ \sum_{i>t} x_i log p_2 + (n - t - \sum_{i>t} x_i)\log(1-p_2) \end{aligned} \tag{2.3}$$

We calculate the maximum likelihood estimation (MLE) of $p_1, p_2$.

$$\frac{\partial l(p_1, p_2)}{\partial p_1} = \frac{\sum_{i\leq t} x_i}{p_1} + \frac{t - \sum_{i>t} x_i}{1-p_1} \tag{2.4}$$

Setting $\frac{\partial l(p_1,p_2)}{\partial p_1} = 0$, we have:

$$\frac{\sum_{i\leq t} x_i}{p_1} + \frac{t - \sum_{i>t} x_i}{1-p_1} = 0$$

$$\sum_{i\leq t} x_i - p_1 \sum_{i\leq t} x_i = p_1(t - \sum_{i\leq t} x_i)$$

$$\hat{p}_1 = \frac{\sum_{i=1}^{t} x_i}{t} \tag{2.5}$$



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The second derivative is:

$$\frac{\partial^2 l(p_1, p_2)}{\partial {p_1}^2} = \frac{-\sum_{i \le t} x_i}{{p_1}^2} - \frac{t - \sum_{i>t} x_i}{(1-p_1)^2} \tag{2.6}$$

Since $p_1 \in [0,1]$ and $x_i \in \{0,1\}$, the second derivative is negative.

$$\frac{\partial l(p_1, p_2)}{\partial p_2} = \frac{\sum_{i>t} x_i}{p_2} + \frac{n - t - \sum_{i>t} x_i}{1-p_2} \tag{2.7}$$

Setting $\frac{\partial l(p_1, p_2)}{\partial p_2} = 0$, we have:

$$\frac{\sum_{i>t} x_i}{p_2} + \frac{n - t - \sum_{i>t} x_i}{1-p_2} = 0$$

$$\sum_{i>t} x_i - p_2 \sum_{i>t} x_i = p_2 n - p_2 t - p_2 \sum_{i>t} x_i$$

$$\hat{p}_2 = \frac{\sum_{i=t+1}^{n} x_i}{n-t}, \qquad \forall\, t = 2, \dots, n-1 \tag{2.8}$$

The second derivative is:

$$\frac{\partial^2 l(p_1, p_2)}{\partial {p_2}^2} = \frac{-\sum_{i>t} x_i}{{p_2}^2} - \frac{n - t - \sum_{i>t} x_i}{(1-p_2)^2} \tag{2.9}$$

Since $p_2 \in [0,1]$ and $x_i \in \{0,1\}$, the second derivative is negative.

Following the above procedure, we consider that there are several switching models depending on what we consider to be the point $t$. We evaluate all these models (no switching model, switching models) and we compare the models based on model selection criteria. This procedure is applied in each pair of 20 goods as we will have remarkable results. In the next section the model selection criteria will be discussed.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

### 2.2.1 Model selection

Model selection is the problem of choosing one from among a set of candidate models. As we must deal with several models, we need to discover which fits best our data by comparing them. For this purpose, few methods have been developed to ensure the best choice of model with the less cost. Here we will introduce two of them, the Akaike Information Criterion proposed by Akaike (1973), and the Bayesian Information Criterion proposed by Schwartz (1978).

The Akaike Information Criterion (AIC) is defined as:

$$AIC = D(\hat{\theta}) + 2d^* \tag{2.10}$$

where $d^*$ is the number of estimated parameters, $\hat{\theta}$ denotes the maximum likelihood estimator of parameters and finally the $D(\hat{\theta})$ is the estimate of the deviance at the estimated parameters.

The deviance generally is given by:

$$D(\theta) = -2\log f(x \mid \theta) \tag{2.11}$$

where $f(x \mid \theta)$ represents the likelihood function.

The Bayesian Information Criterion (BIC) is defined as:

$$BIC = D(\hat{\theta}) + d^* \log(N) \tag{2.12}$$

where $N$ denotes the number of observed variables.

We select the model giving smallest value of AIC and BIC over the set of models considered. Furthermore, this model can best predict a replicate dataset of the same structure as the observed and finally will give us more accurate results.

We simultaneously calculate the value of BIC in the no switching model (i.i.d. model). Also, we find the time of $t$ where it is observed the smallest value of BIC in the switching model. In this time, we assess that it is observed a switching in the shopping habit.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

### 2.2.2 Bayes factor & BIC approximation

Given a data set $x$, we need to compare two competing hypotheses $H_0$ and $H_1$, usually a null and an alternative. Suppose $f(x)$ be the marginal likelihood of the observed sample which is the probability of the observed data $x$, $P(H_i|x)$ be the posterior probability of a hypothesis being true, $P(H_i)$ be the prior probability, and $P(x|H_i)$ be the marginal likelihood of the sample under the assumption that hypothesis $H_i$ is true, with $i \in \{0,1\}$. We assume that $P(H_0) + P(H_1) = 1$. From Bayes' s theorem, we have the following:

$$P(H_0|x) = \frac{f(x|H_0)P(H_0)}{f(x)} \tag{2.13}$$

$$P(H_1|x) = \frac{f(x|H_1)P(H_1)}{f(x)} \tag{2.14}$$

Taking the ratio of the above equations we have:

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{f(x|H_0)P(H_0)}{f(x|H_1)P(H_1)} \tag{2.15}$$

The transformation is multiplication by

$$BF_{01} = \frac{P(x|H_0)}{P(x|H_1)} \tag{2.16}$$

which is the Bayes factor.

In other words,

$$Posterior\ odds\ =\ Bayes\ factor\ \times\ prior\ odds$$

The Bayes factor of $H_0$ versus $H_1$ is denoted $BF_{01}$, and is defined as the ratio of the marginal likelihoods of the observed data $x$ under the two different hypotheses. There is a different expression of Bayes Factor which is the ratio of posterior to prior hypotheses odds:

$$BF_{01} = \frac{\frac{P(H_0|x)}{P(H_1|x)}}{\frac{P(H_0)}{P(H_1)}} = \frac{P(H_0|x)P(H_1)}{P(H_1|x)P(H_0)} \tag{2.17}$$



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

If hypothesis $H_i$ assumes the existence of a parameter vector $\theta_i \in \Theta_i$ with a prior $p_i(\theta_i)$, then the marginal likelihood $P(x|H_i)$ is obtained by integration in the parameter space.

The marginal likelihood under model $H_i$ is given by:

$$P(x|H_i) = \int_{\Theta_i} f(x|\theta_i, H_i) p_i(\theta_i) d\theta_i \ , \text{i} \in \{0,1\} \tag{2.18}$$

Finally, Bayes factor is a measure of the strongness of the evidence provided by the data supports a hypothesis (or model) over another. Note that classical hypothesis testing gives one hypothesis (or model) preferred status (the "null hypothesis"), and only supposes evidence against it.

Jeffreys (1961) gave a scale for interpretation of Bayes factor as following:

| **Bayes Factor** | **Evidence against $H_0$** |
|---|---|
| $\boldsymbol{BF_{01} > 1}$ | Negative |
| $1 > \boldsymbol{BF_{01}} > \boldsymbol{10^{-0.5}}$ | Bare |
| $\boldsymbol{10^{-0.5} > BF_{01} > 10^{-1}}$ | Substantial |
| $\boldsymbol{10^{-1} > BF_{01} > 10^{-1.5}}$ | Strong |
| $\boldsymbol{10^{-1.5} > BF_{01} > 10^{-2}}$ | Very Strong |
| $\boldsymbol{10^{-2} > BF_{01}}$ | Decisive |

*Table 2: Jeffreys' grading of evidential strength given Bayes factors*

Kass and Raftery (1995) presented a modified version with the Bayes factor undergoing a $2 \times log_e$ transformation.

| $\boldsymbol{2 \times log_e(BF_{01})}$ | **Evidence against $H_0$** |
|---|---|
| $\boldsymbol{0 - 2}$ | Bare |
| $\boldsymbol{2 - 6}$ | Positive |
| $\boldsymbol{6 - 10}$ | Strong |
| $\boldsymbol{> 10}$ | Very strong |

*Table 3: Kass and Raftery' transformation of Jeffreys' scale*



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The Bayes factors could quantify the hypotheses and at the same time could easily and conveniently interpret the results. For instance, if the Bayes factor is $\mathrm{BF}_{01} = 3$, we would have the following interpretations:

- The observed data are three times as likely to occur under hypothesis $H_0$ than they are under $H_1$.
- The data support hypothesis $H_0$ three times as strongly than they do $H_1$.
- The odds of hypothesis $H_0$ versus $H_1$ after the experiment are three times what they were before it.

After all, we could consider that a Bayes factor in favor of a hypothesis $H_0$, can easily exclude the possibility of that hypothesis being doubtful and inadequate to describe the reality below an experiment. It only implies that $H_0$ is better compared to its alternative $H_1$.

In many applications, we have to deal with several models, and it is necessary to compare each of them with a baseline model, this can be a null model ($M_0$) with no independent variables or a saturated model ($M_S$) in which each data point is fit exactly. When we have to compare two models, $M_1$ , $M_2$ , we note the following, using the Bayes factor:

$$B_{12} = \frac{p(x|M_1)}{p(x|M_2)} = \frac{\frac{p(x|M_S)}{p(x|M_2)}}{\frac{p(x|M_S)}{p(x|M_1)}} = \frac{B_{S2}}{B_{S1}} \tag{2.19}$$

Given a data set $x$

Then we have:

$$2logB_{12} = 2logB_{S2} - 2logB_{S1} \approx BIC_2 - BIC_1 \tag{2.20}$$

Finally, the two models, the switching and the switching model, could be compared with the difference of their BIC values, as smaller the value is as better the data could be predicted by this model. Taking into concern all above, it is used the scale of Kass & Raftery in Bayes Factor since the logarithm of the marginal probability of the data leads to easier interpretation.

In our analysis, it is assumed as null the hypothesis that data are independent and identically distributed random variables (i.i.d) and as alternative the hypothesis that there is switching in the shopping habit in time $t^*$. We make comparisons considering all possible values of $t$. The $t^*$ is the purchase where we found the smallest value of BIC for all $t = 2, 3, \dots, n-1$ and then the time where we observe switching in shopping behavior.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## 2.3 Metric construction

The aim of our analysis is to discover possible similarities in 20 products of category of "juices" and moreover which product can replace someone else on the shelves of stores.

For each pair of 20 items, we take into account two hypothesis, the null $H_0$ in which there is no switching in the behavior of customer and the alternative $H_1$ in which there is switching. We find the number of transactions that we have strong evidence ($\mathrm{BF}_{01} \geq 10$) against the hypothesis $H_0$. Then, we create the following metric:

$$Q = \frac{Number\ of\ transactions\ with\ strong\ evidence\ in\ favor\ of\ H_1}{Total\ number\ of\ transactions\ of\ pair} \qquad (2.21)$$

Calculating this ratio we can discover the distance of pairs of juices, and it is necessary to consider the total number of transactions in each pair to have a logical result.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Chapter 3

# Clustering

Nowadays, the databases increase with exponential rate and data warehouses and other repositories store an enormous number of records. To deal with such a large amount of information, grouping them into meaningful categories is efficient.

Cluster analysis, as it is known, is applicable in many fields, such as Market or Customer Segmentation, where people who works in marketing, discover groups in their customer bases and then use this knowledge to develop targeted marketing campaigns. In this chapter, the clustering analysis is applied both to 22.029 customers and to 20 items.

## 3.1 Clustering the customers

Shortly to remind, each customer has a vector of 20 values, which represents the transactions of 20 items. It would be meaningful to group customers based on their shopping preferences and rank in a cluster those who have similar shopping habits. Converting each time the data, we could assume that they follow multivariate normal distribution (I Case) or multinomial distribution (II Case) too. In the two following sections, we will analyze the clustering of customers in the two assumptions of distributions.

### 3.1.1 Multivariate Normal Distribution

The multivariate normal distribution is considered as a multidimensional generalization of the one-dimensional normal distribution. This distribution represents the distribution of a multivariate random variable which consists of multiple random variables that could be correlated with each other.

In the same manner as the one-dimensional normal distribution, it is defined by two parameters, the mean vector $\mu$, i.e., the expected value of the distribution, the covariance matrix $\Sigma$, which measures how dependent two random variables are and



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

how they change together. This distribution is denoted as $N(\mu, \Sigma)$. The joint probability density of the multivariate normal with dimension $d$ is following:

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{2\pi^d|\Sigma|}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)} \quad (3.1)$$

where $\boldsymbol{x}$ is a random vector size d, $\mu$ is the mean vector and $\Sigma$ is the covariance matrix size $d \times d$.

The matrix, with the rows represents the customers and the column the quantities of the transactions, has a dimension of $22.029\ x\ 20$. In the following table, it is represented the relative frequencies of 4 customers:

| CUSTOMERS | 93 | 78 | 83 | 33 | 32 | 4 | 113 | 111 | 90 | 103 | 95 | 50 | 21 | 47 | 54 | 101 | 49 | 46 | 122 | 123 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.174 | 0 | 0.043 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.782 |
| 2 | 0.133 | 0.044 | 0 | 0.133 | 0.088 | 0 | 0.088 | 0 | 0 | 0 | 0 | 0.044 | 0 | 0.155 | 0.066 | 0 | 0.2 | 0 | 0.044 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0.604 | 0 | 0.046 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.348 | 0 |
| 4 | 0 | 0 | 0.032 | 0 | 0 | 0.032 | 0.322 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.064 | 0 | 0 | 0.548 | 0 |

*Table 4: Relative frequencies of customers*

As we could see from the above table, the 1st customer purchased the item Brand B Orange Juice without pulp PET packaging 90CL *(111)* with percentage of 17,4% and the item Brand D Orange Juice 1L *(103)* with percentage of 4,3%. Thereby, the 22.029 customers have percentages of purchases of juices.

### 3.1.2 Multinomial Distribution

The multinomial distribution is an extension of the binomial distribution. The binomial distribution models nominal data with two categories, the multinomial distribution is used to model nominal data whose outcome has more than two categories.

Consider a trial that results in exactly one of some fixed finite number $k$ of possible outcomes, with probabilities $q_1, q_2, \ldots, q_k$ with $q_i \geq 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^{k} q_i = 1$ and there are $n$ independent trials. Then let the random variables $X_i$ indicate the number of times outcome number $i$ was observed over the $n$ trials. Then $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ follows a multinomial distribution with parameters $n$ and $q$, where $q = (q_1, q_2, \ldots, q_k)$.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Our raw data is a matrix with 22.029 rows and 20 columns. Every row presents the customers with their transactions of 20 products. It is like an experiment you repeat in $n$ times with 20 possible results. We note how many successes we have in each category by committing to the total transactions made by each customer. Let $x_i$ , $i = 1, 2, \ldots, 22029$, denote the number of customers in $n$ trials-transactions of $k = 20$ categories-products. Then $x = (x_1, x_2, \ldots, x_{20})$ follows a multinomial distribution.

In the following table, it is represented the transactions of 4 customers:

| CUSTOMERS | 93 | 78 | 83 | 33 | 32 | 4 | 113 | 111 | 90 | 103 | 95 | 50 | 21 | 47 | 54 | 101 | 49 | 46 | 122 | 123 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 2 | 6 | 2 | 0 | 6 | 4 | 0 | 4 | 0 | 0 | 0 | 0 | 2 | 0 | 7 | 3 | 0 | 9 | 0 | 2 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 26 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 0 |
| 4 | 0 | 0 | 1 | 0 | 0 | 1 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 17 | 0 |

*Table 5: Transactions of customers*

According to the previous table, the 1st customer purchased in total 23 items of which there are item Brand B Orange Juice with pulp PET packaging 90CL *(111)* with 4 transactions, item Brand D Orange Juice 1L *(103)* with 1 transaction and item Brand B Orange Juice 1 L *(123)* with 18 transactions.

In the following section, we will present the model-based clustering that is used in the two assumptions of distributions.

### 3.1.2.1 Model-based clustering

One of the most known clustering procedures is the model-based cluster analysis in which the data is considered as coming from a mixture of basic probability distributions. It assumes that the data were generated by a model and tries to recover the original model from the data. The distribution of data could be Bernoulli, Gaussian or from another distribution family. Furthermore, the most common approach is the Gaussian Mixture Model, where the number of clusters is $g$, each observation is assumed to be distributed as one of $g$ multivariate-normal distributions.

$$P(X_i = x) = \sum_{g=1}^{G} \pi_g \, P(X_i = x | Z_i = g) \qquad (3.2)$$



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

where $Z_i \in \{1,2,\ldots,g\}$ is the latent variable representing the mixture component for $X_i$, $P(X_i|Z_i)$ is the mixture component and $\pi_g$ is the mixture proportion representing the probability that $X_i$ belongs to the $g-th$ mixture component.

Moreover, model-based clustering aims to identify unobserved heterogeneity in a population based on observed data. A commonly used criterion for estimating the model parameters is maximum likelihood (ML). The most used technique for estimating Gaussian Mixture Model based clustering is the Expectation Maximization algorithm (EM).

### 3.1.2.2 Expectation Maximization algorithm

Having in mind the general form of Gaussian Mixture Model we understand, that estimating its parameters, is a difficult task since its log likelihood form is rather complicated. To estimate the parameters $\Theta_g$ and the weights $\pi_g$ we should use the Expectation-Maximization Algorithm (EM). The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin. It is a technique that is suitable to estimate parameters in cases of missing data. Moreover, it is used in problems such as when dealing with grouped data, missing or truncated data etc. Suppose that we have a set of data $Y$ and we have a likelihood of the form $L(\theta|Y)$ that our aim is to maximize it to obtain the estimators for the parameters $\Theta$. In the case that we face mixture models, we add information with the form of missing data $Z$ to obtain a likelihood easier to maximize, of the form $L(\theta|Y,Z)$.

The steps of clustering by using Gaussian Mixture Model fitted by Expectation Maximization algorithm are two, the E-step or Estimation step and the M-step or Maximization step. In the first step we have to estimate the missing data $Z$, from the observed likelihood $L(\theta|Y)$. First, we define the following function given a current estimate $\theta^{(j)}$.

$$Q(\theta,\theta^{(j)}) = \int_Z \log L(\theta|Y) P(Z|\theta^{(j)},Y) dZ = E(log\, L(\theta|Y,Z)) \qquad (3.3)$$

where $P(Z|\theta^{(j)},Y)$ is the posterior distribution of $Z$. We finally calculate this $Q(\cdot)$ function. In the second step, we maximize the $L(\theta|Y,Z)$ likelihood with respect to $\theta$, using the estimation for $Z$ from the E-step, or in a formal way we maximize the function $Q(\theta,\theta^{(j)})$ with respect to $\theta$. In many cases the ML estimators needed for M step are already known, but there are also examples, where the estimators are computed numerically, using Newton Raphson for example. Starting with initial values, we repeat the above steps until a termination condition is satisfied. A widely used stopping



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

criterion, for instance, is the following:

$$\frac{|L^{(j+1)} - L^{(j)}|}{|L^{(j+1)}|} < e^* \tag{3.4}$$

Where $L^{(j)}$is the likelihood at the j-th iteration. The algorithm stops when the difference between two iterations is very small (the $e^*$ could be $10^{-6}$). It is proved that at each iteration the likelihood increases.

In clustering, we need to use EM to estimate the probabilities $\pi_g$ and the parameters $\Theta_g$ The log-likelihood of the data is:

$$l(\Theta, X) = \sum_{i=1}^{n} \log \left( \sum_{g=1}^{G} \pi_g \, p_g(x_i|\theta_g) \right) \tag{3.5}$$

The above likelihood is difficult to maximize and it necessary to use EM algorithm. This problem would be easier if we knew in which group every observation belongs to. The idea is to augment data with some unobserved variables $Z_{ig}$. We will suppose that these variables, can indicate in which cluster each observation belong to. If for example, $Z_{ig} = 1$ this means that the $i - th$ observations are from the $g - th$ group, and it would be zero everywhere else. The new log-likelihood with the augmented data is:

$$\begin{aligned} l(\Theta, X, Z) &= \sum_{i=1}^{n} log \left( \sum_{g=1}^{G} (\pi_g p_g(x_i|\theta_g))^{z_{ig}} \right) \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} (z_{ig} log \, \pi_g) + z_{ig} \, log \, (p_g(x_i|\theta_g)) \end{aligned} \tag{3.6}$$

When we estimate the $Z_{ig}$s, we know in which cluster each observation belongs to. We should also notice that $p_g$ can be any distribution and it isn't obligatory for all components to be from the same. The procedure that we follow is the same described previously.

The EM algorithm often fails to identify the true maximum and sometimes it gets trapped in local maxima. The procedure of choosing good initial values is a crucial part of the estimation. There are many strategies that one could follow for the initial values. First, if prior knowledge regarding the clusters of the data exists it would we useful to use it for initialization. Furthermore, one could use as initial values the centers of other clustering techniques such as the hierarchical clustering or the k-means algorithm. It is



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

important to be sure that we have covered the whole space of our data, so another strategy would be to run EM many times, using different initial values each time and choose the better solution.

In our two cases of distributions, we use different function in R to classify the customers into groups. In the assumption of the data follows a multivariate normal distribution, we use the function *Mclust()* in R and in the assumption of the data follows a multinomial distribution, we use the function *multmixEM()*, which belong to *mclust package* and *mixtools package* respectively. The criteria to discover the optimal number of clusters are the Bayesian Information Criterion (BIC), we analyzed this criterion in Chapter 2, and the Integrated Complete-data Likelihood (ICL), this criterion will be analyzed in the following section.

### 3.1.2.3 ICL Criterion

It is widely known that the integrated completed likelihood (ICL) (Biernacki, Celeux & Govaert 2000) criterion is an approach in model-based clustering which automatically chooses the number of clusters in a mixture model. This criterion is an alternative to BIC criterion, and it is characterized as a penalized likelihood criterion.

ICL is the criterion BIC penalized by the estimated mean entropy

$$E(K) = -\sum_{k=1}^{K}\sum_{i=1}^{n} t_{ik} log_{t_{ik}} \geq 0, with\ 1 \leq i \leq n, 1 \leq k \leq K \qquad (3.7)$$

$t_{ik}$ denoting the conditional probability that $x_i$ arises from the $k-th$ mixture component which penalty involves an "entropy" term. ICL was designed to select the model leading to the greatest evidence for clustering the data since it maximizes the integrated likelihood. Due to the mean entropy, ICL penalises clustering configurations exhibiting overlapping groups, this means that low- entropy solutions with well-separated groups are preferred to configurations that give the best match concerinng the distributional assumptions. While BIC allows a very efficient estimation of the number of components for the mixture model, ICL steadily and reliably estimates the number of clusters for real datasets and for simulated data sets from mixtures when the components are not overlapping (Baudry et al. 2010).



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## 3.2 Clustering 20 Products

With the process of clustering the products, we manage to organize the 20 items into "logical" groups, to discover similarities and differences between them, but also to draw useful conclusions about them. To accomplish this difficulty, it is used the hierarchical clustering. This procedure of clustering will be analyzed in the following section.

### 3.2.1 Hierarchical clustering

The aim of hierarchical clustering is to create a sequence of nested segments, which can be conveniently visualized via a tree or hierarchy of clusters, named as cluster dendrogram. The lowest level of the tree (the leaves) consists of all elements in each own cluster and the highest level (the root) consists of all elements in the cluster. There are two main algorithmic approaches to extract hierarchical clusters, agglomerative ("bottom-up") and divisive ("top-down").

To group data, we need a way to measure the elements and their distances relative to each other to decide which elements belong to a group. This can be a similarity, although on many occasions a dissimilarity measurement. There are several approaches to measure this metric between datasets.

- Single - linkage: the distance between two clusters is determined by a single pair of elements, those two elements (one in each cluster) that are closest to each other.
- Complete - linkage: the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other.
- Average - linkage: the average distance from any member of one cluster to any member of the other cluster.
- Centroid - linkage: the square of the Euclidean distance between the centroids of each cluster.
- Wards - linkage: is based on a sum of squared errors rationale that only works for Euclidean distance between observations. In addition, the sum of squared errors requires the consideration of the so-called centroid of each cluster



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

### 3.2.1.1 Agglomerative Clustering

This approach is also known as AGNES (Agglomerative Nesting). The steps of the approach are following:

1. Start by assigning each data point to a cluster and there are $n$ number of data points at total in the dataset, each containing just one data element.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now the cluster number decreases by one.
3. Continue the process until all data points are clustered into a single cluster, which includes all data elements in the dataset.

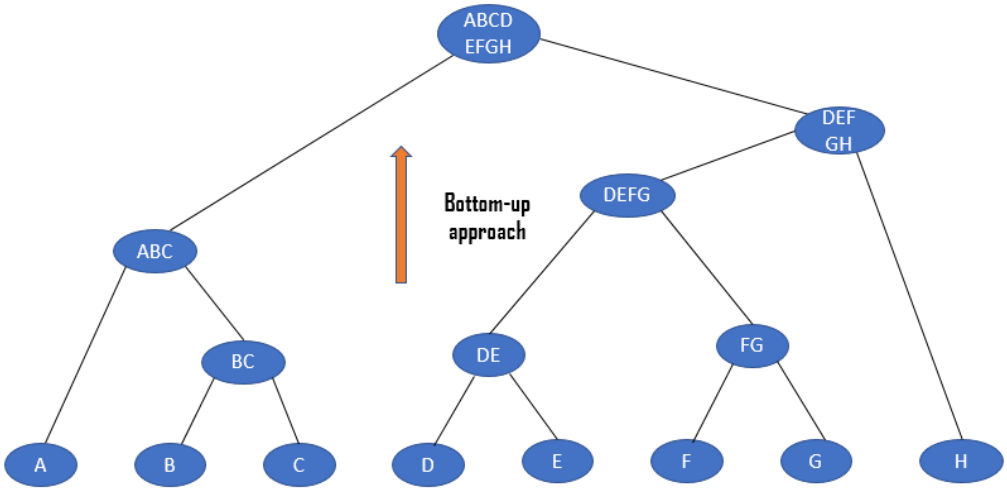



*Figure 5: Example of Agglomerative Hierarchical Clustering*

For instance, observing the above graph, firstly all data elements {A, B, C, D, E, F, G} are single clusters, the leaves of the dendrogram. Afterwards, data elements "B" and "C" merged into one cluster together with "D, E" and "F, G". After that, "cluster A" and "cluster B, C" become merged and at the same time "D, E" and "F, G" merged. And on the fourth step cluster "D, E, F, G" and cluster "H" merged and become one cluster as "D, E, F, G, H". At the last step, the cluster "A, B, C" and cluster "D, E, F, G, H" merged and with this step all data elements united and become only one cluster with members "A, B, C, D, E, F, G, H".

### 3.2.1.2 Divisive Hierarchical Clustering

This approach is also known as DIANA (DIvisive ANAlysis). The steps of the approach are following:

1. Start by assigning all data elements to one cluster and there are $n$ number of data points at total in one cluster.
2. Find the highest dissimilarity between data elements and split them into pair of clusters, so that now the cluster number increases by one.
3. Continue the process until all data elements are clustered into different clusters, and finally there are $n$ number of clusters, each containing just one data element.

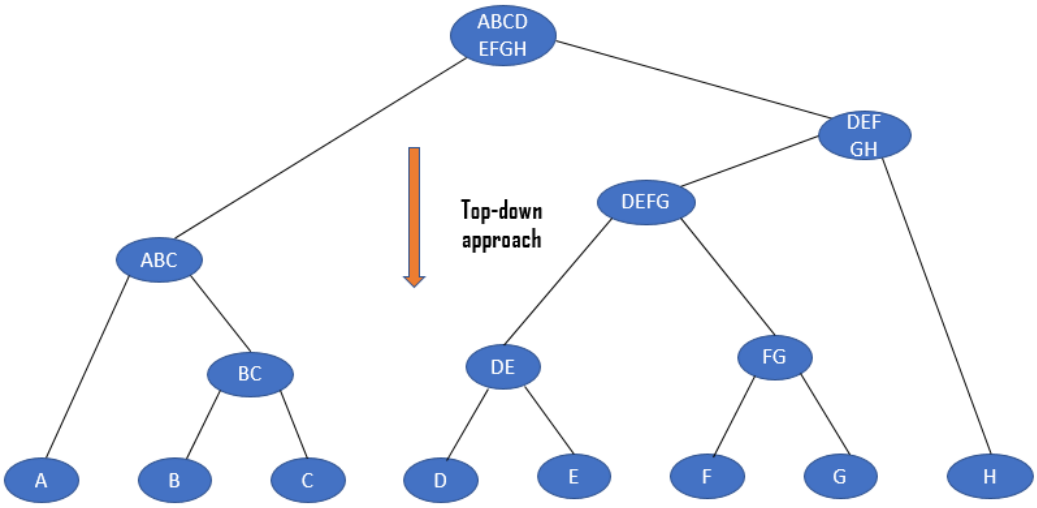



*Figure 6: Example of Divisive Hierarchical Clustering*

For instance, observing the above graph, firstly all data elements {A, B, C, D, E, F, G, H} is a single cluster, the root of the dendrogram. Afterward, there will be 2 clusters, cluster "A, B, C" and cluster "D, E, F, G, H". And then cluster "D, E, F, G, H" split into cluster "D, E, F, G" and cluster "H". And in the following step, cluster "A, B, C" split into cluster "B, C" and cluster "A" at the same time with separation of cluster "D, E, F, G" into "D, E" and "F, G". In the last step all data elements become separate clusters "A, B, C, D, E, F, G, H".

Afterward hierarchical clustering analysis of 20 item on the category of juices, a method that could be used to select the optimal number of clusters is Silhouette analysis which based on the Silhouette Coefficient. In the following section, Silhouette Analysis will be presented.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

### 3.2.2 Silhouette analysis

The Silhouette Coefficient for a data point $i$ is computed as following:

$$s(i) = \frac{b_i - a_i}{\max(a_i - b_i)} \tag{3.8}$$

where $a_i$ is the average distance of data point $i$ from members of its own cluster and $b_i$ is the average distance of data point $i$ from members of the nearest cluster. The silhouette ranges from −1 to +1.

The worst cases are when $s(i) < 0$ and values quite close to -1, because the negative value means that $a_i > b_i$, therefore the point $i$ is closer to the nearest cluster rather than to its own one. When $s(i)$ is near to zero, this value points out that the data point $i$ is in, or very close to, the neighboring area between the two adjacent clusters. If the value of the Silhouette Coefficient is closer to 1, this will indicate that the point is closest to its own cluster and far from the other clusters.

Depicting graphically the mean Silhouette Coefficient over all points is very useful, since is a measure of how tightly grouped all the points in the cluster are, for different values of $k$. After all we choose this $k$ with the highest value on the mean of the $s(i)$.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Chapter 4

# Implementation of methods

## 4.1 Implementation of model construction and metric construction

As we mentioned in section 2.2, it is assumed that in every $t$, each purchase, $t = 1, \ldots, n-1$, there is a different Bernoulli model with different unknown parameters p. We evaluate all these models using the Bayesian Information Criterion approximation., find all n-1 BIC values and find the $t^*$ where the BIC value is the smallest. Now we have two BIC values, the value of no switching model and the value of switching model in time $t^*$. Also, if the value of BIC of switching model is smaller than the value of no switching model, we will assess that there is switching on shopping behavior. Evidence that are strong against the hypothesis that data are independent and identically distributed random variables, namely there is not switching in the preference of customer, could be used for visualizing products similarities.

As it is analyzed above, our aim is to find similarities between the 20 products in the category of "juices". We potentially assess that if a customer purchases a product A in consecutive transactions and at once he purchases a product B, it will be considered a similarity between 2 these products. An example of these case is presenting below:

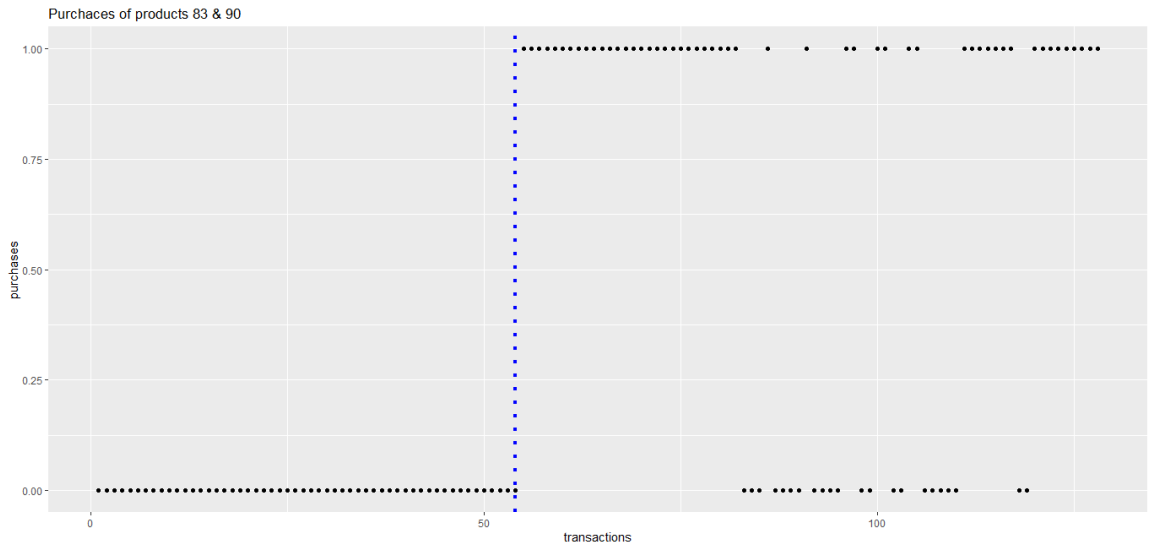



*Figure 7: Switching in 83 & 90 items*

The pair *(83,90)* is bought by 1.455 customers (look Figure 3). The shopping habit of a customer among of 1.455 is presented in the Figure 7. This customer has in total 128 transactions of Brand B Apple Juice PET packaging 90CL *(90)* and Brand A Apple Juice 1L *(83)*. Observing the figure, we could say that in total the customer had not a specific buying behavior and the customer vacillated between the two items. Suppose that there is $t = 2, \dots, 128$ when it is observed a switching in the shopping habit. Moreover, in each t we assume that the purchase is a random variable from the Bernoulli distribution with unknown parameter $p \in [0,1]$ and we calculate the BIC values. In the following figure, the BIC values are presenting:

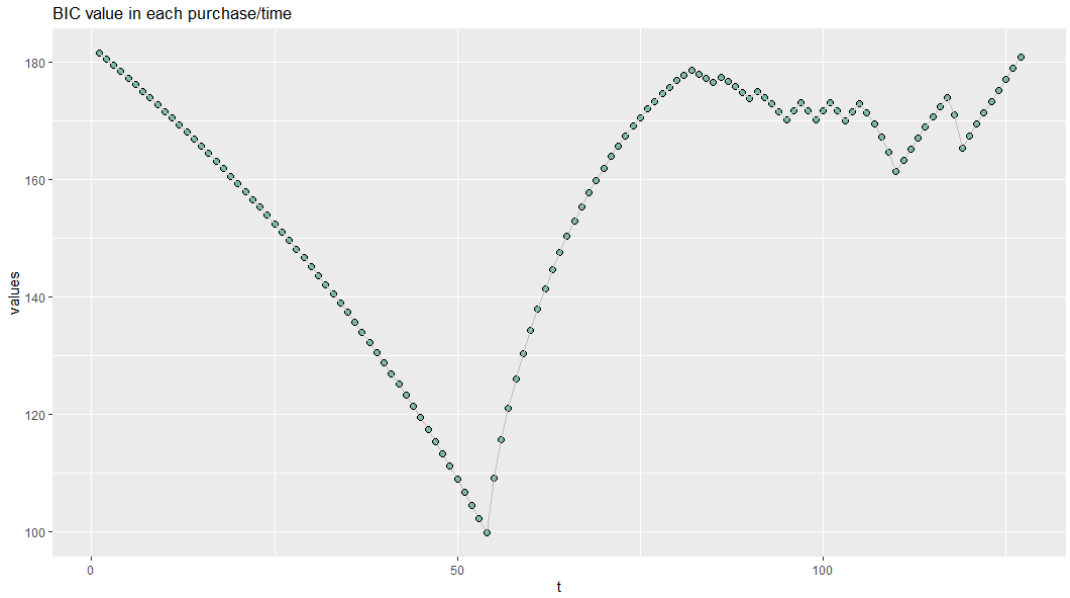



*Figure 8: BIC values in switching models in pair (83,90)*

Looking the above figure, we clearly infer that the minimum value is noticed in 54 purchase and the value is 99.77. Then, in the purchase 54 and afterward it is observed a preference in a product among two products. From the other side, the value of BIC in the no switching model is 177.77. The difference of two values of BIC gives us the Bayes Factor 78, strong values in favor of the hypothesis that there is switching after the $54^{th}$ transaction.

A distinct shopping habit in pair *(123,111)* is giving in the following figure.

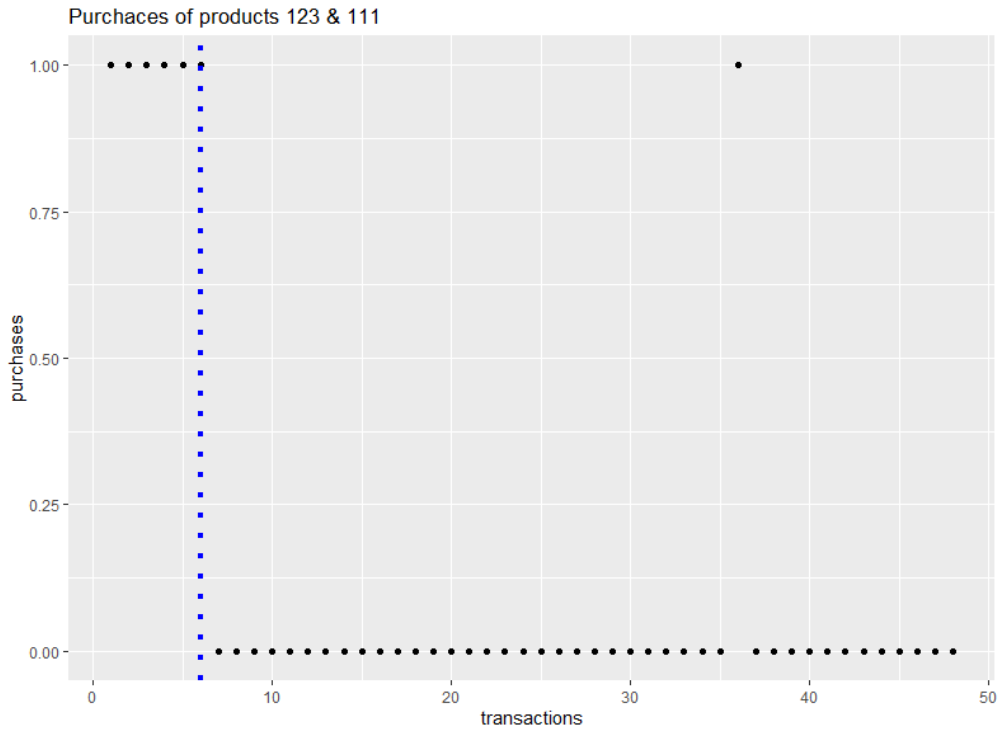



*Figure 9: Switching in 123 & 111 items*

It is obvious that among 48 purchases of customer there is a preference in a one item of pair *(123,111).* Corresponding to the previous example, we suppose that there is $t = 1, 2, \ldots, 47$ when it is observed a switching in the shopping habit, and we calculate the BIC values for each purchase from 47. It is created the following figure:

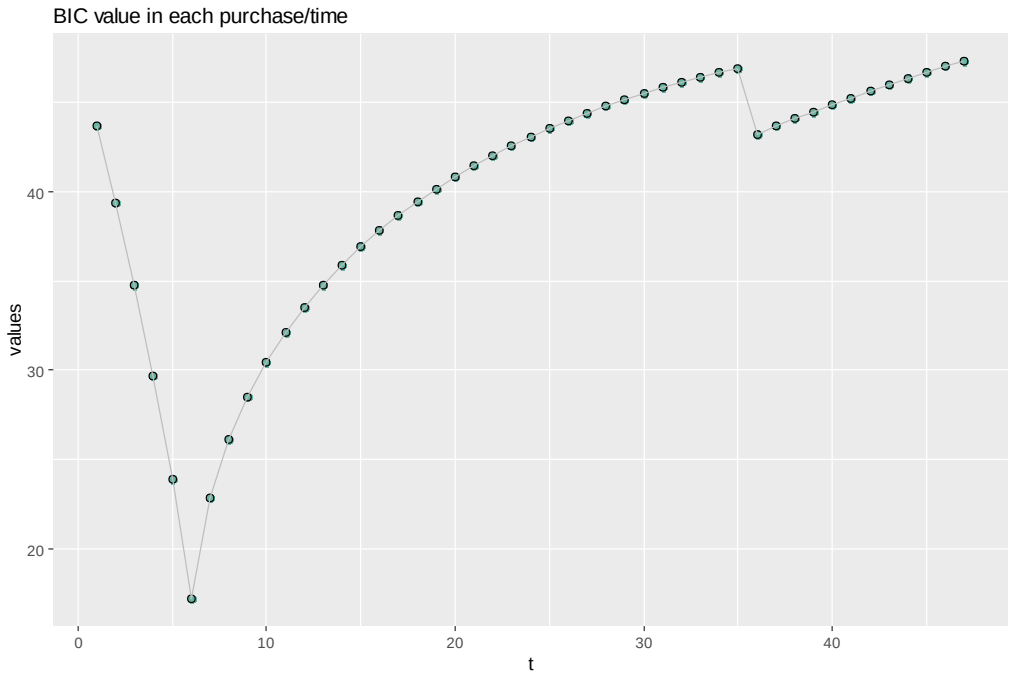



*Figure 10: BIC values in switching models in pair (123,111)*



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

We notice that the minimum value 17.19 is noticed in 6th purchase. Then, in the purchase 6th and afterward it is observed a preference in a product among two products. From the other side, the value of BIC in the no switching model is 43.75. The difference of two values of BIC provides us with the Bayes Factor value 26.56, strong evidence that there is switching after the 6th transaction.

Furthermore, we create a symmetric matrix which is made of the metric $Q$ for each pair. In the following figure, we could observe the high values of the metric:

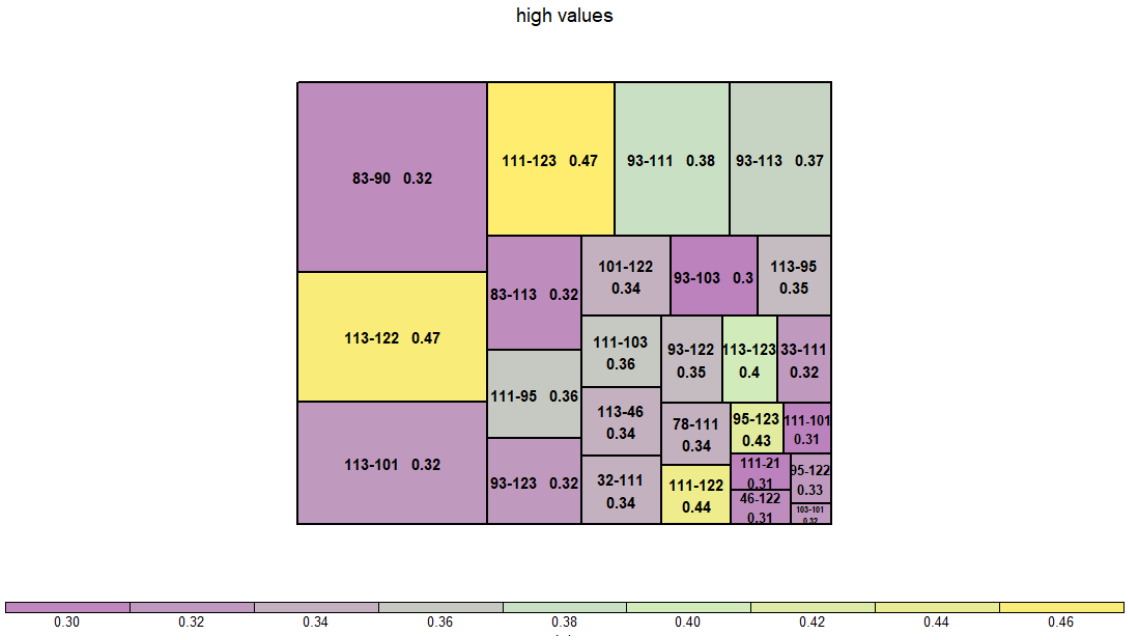



*Figure 11: High values of metric Q*

The scale of the colors shows us the ranges of the strong evidence such as the light yellow depicts that the metric ranges up to 0.44. From the other side, the area of the block represents how many transactions are observed in this metric. It is obvious that there is switching between item Brand B Orange Juice without pulp PET packaging 90CL *(113)* and item Brand B Orange Juice with SS pulp 1 L *(122)* with percentage of 47% among 1.001 customers. The percentage of 47% it is also observed in the pair of item Brand B Orange Juice with pulp PET packaging 90CL *(111)* and item Brand B Orange Juice 1 L *(123)* among 787 customers. On the contrary, it is observed a switching between items Brand A Orange Juice BP 1,50L *(95)* and Brand B Orange Juice 1 L *(123)* with percentage of 43% among 107 customers only.

## 4.2 Clustering Items

In this section, agglomerative hierarchical clustering algorithm with Ward method, where distances were calculated by using Ward distance, will be implemented, and evaluated. As mentioned before, this algorithm starts to assign all observations into different clusters as its initial step. And then distances will be calculated between all single data points as clusters and most similar observations (lower distance value) will be united in order to be member of same cluster. This procedure iterate until all observations will be member of one cluster.

In the attempt to group items and discover similarities, it is necessary to create the distance matrix $S$. Creating the similarity matrix $D$ which consists of all Q values, we calculate the following value in each pair.

$$S = max(Q) - D$$

This means that matrix S is created by finding the maximum value of the metric Q and subtracting the similarity matrix that we have already created. After implementation of this algorithm of grouping the items in the distance matrix $S$, a dendrogram is generated, which is shown below.

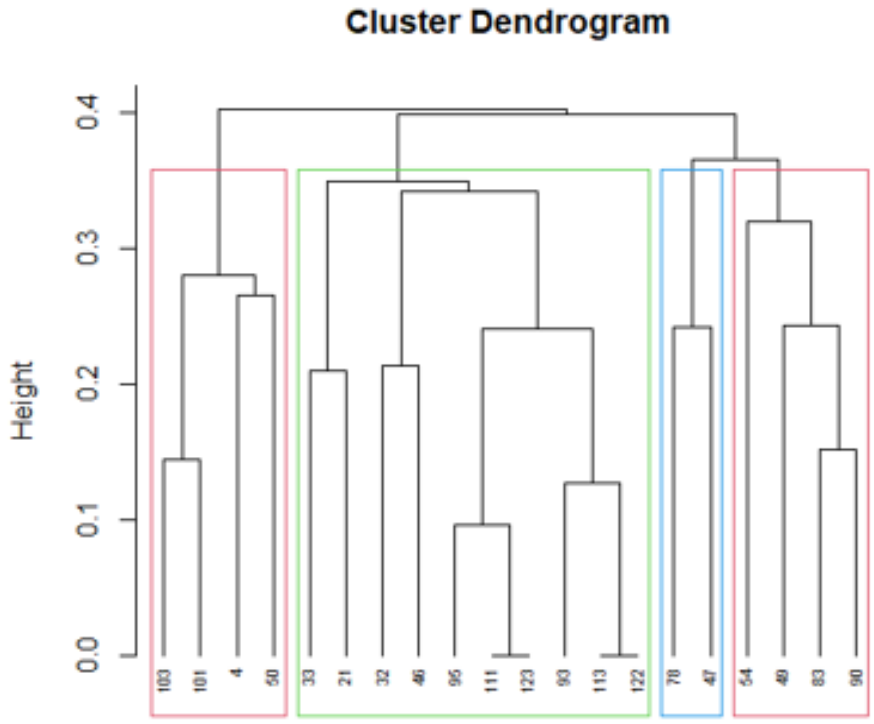



*Figure 12: Hierarchical Clustering Dendrogram in Similarity Matrix of 20 items*

From the dendrogram, a height point needs to be determined to cut the tree and split data into clusters. In the above dendrogram, it is expected to cluster data into 4, which can be obtained to cut the tree at height between 0.3 and 0.4.
In this step, Silhouette plots are used to determine the optimal number of clusters.

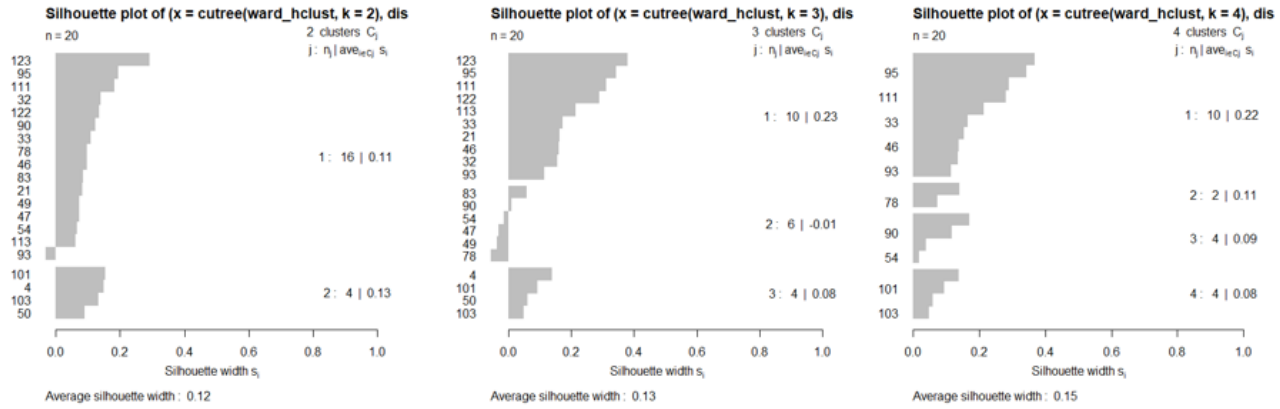



*Figure 13: Silhouette plots of different g=2, 3, 4*

From the graphs above, it can be said that "4" seems to be the optimal number. In details, we could easily observe that in the case of 4 clusters, any product has negative silhouette values and the overall average silhouette is 0.15.

## 4.3 Clustering Customers

### 4.3.1 Clustering in the Case I of Multivariate Normal Data

In the assumption that data follows multivariate normal distribution, In the following figure is presented the values of Bayesian Information Criterion and of Integrated Information Criterion for different number of clusters.

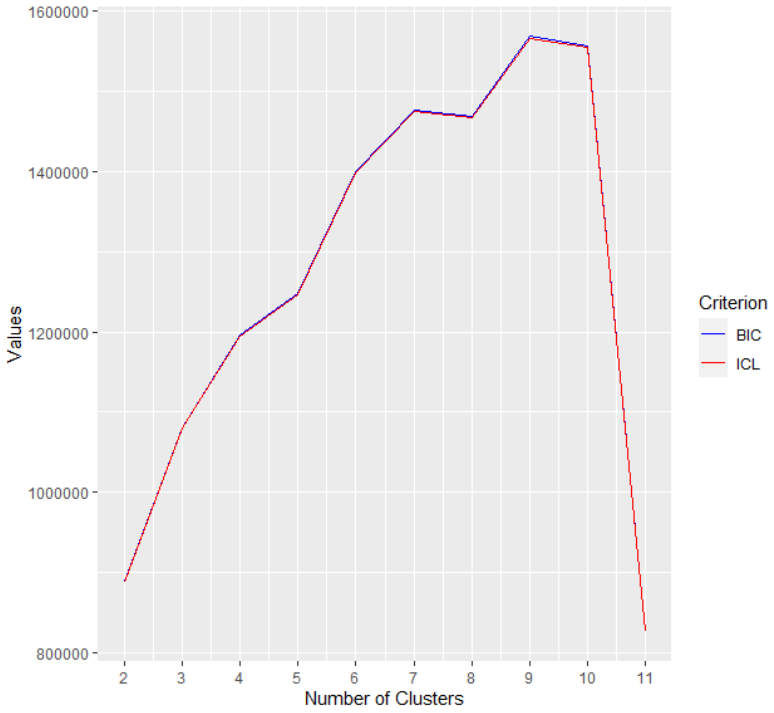



*Figure 14: Bayesian information criterion (BIC) values for g = 2, 3, …, 11 clusters in Case I*

Although, BIC is calculated as $(-2log(L) + dlog(N))$, where $(L)$ refers to the maximized value of the likelihood function of the model $N$ is number of observations, $d$ is the number of parameters, using Mclust function in R, BIC calculated as $\big(-2log(L) + dlog(N)\big)$, thus optimal model is selected according to the highest BIC score.

Moreover, BIC values are an approximation to integrated (not maximum) likelihood, and we seek the model with the greatest integrated likelihood (Bayes factor), and we select the model with the largest BIC, respectively with ICL with the two values being similar in each segmentation. Increasing the number of clusters, the value of BIC is becoming higher, and it is illogical in our size of data. We present the results of 7 clusters. Then after setting up 7 as cluster number, customers per cluster are distributed as it shown in the following histogram.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

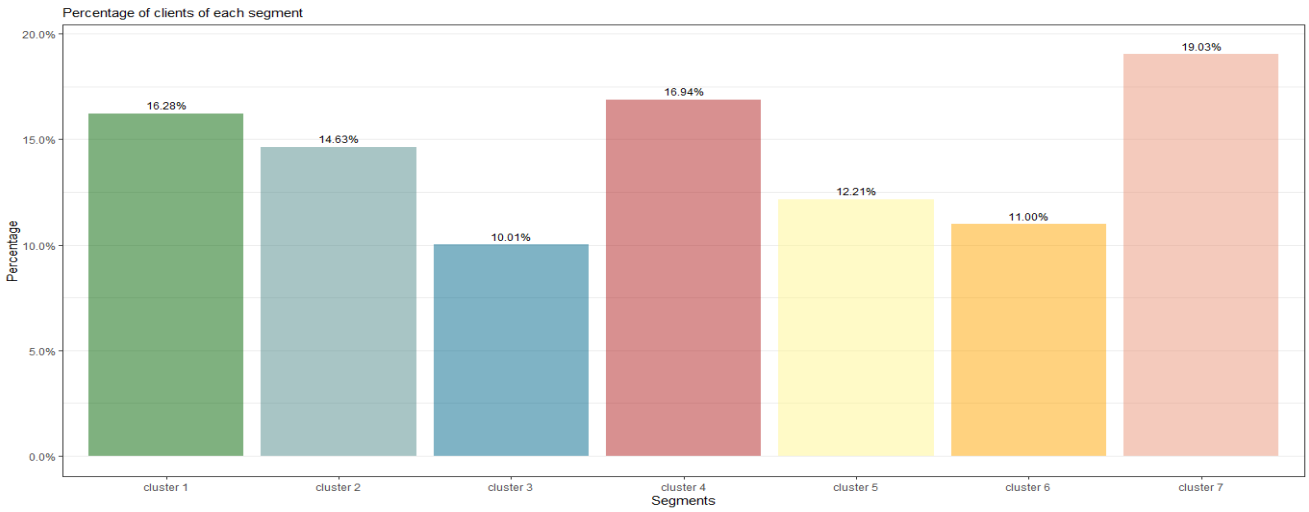



*Figure 15: Number of customers per segmentation in the assumption of Multivariate Normal Data*

At the above histogram, the largest number of customers is noticed in the Cluster 7 where it is assembled the 19,03% percentage of the customers. Moreover, it is assumed as special category the Cluster 3 since almost the 10% percentage of customers are in this segmentation.

Then after setting up 7 as cluster number, the metric, which measures the strongness of the evidence that there is switching between pairs, per cluster is created as it shown in the following figures.

## *Plots in the assumption of Multivariate Normal Data (I case)*

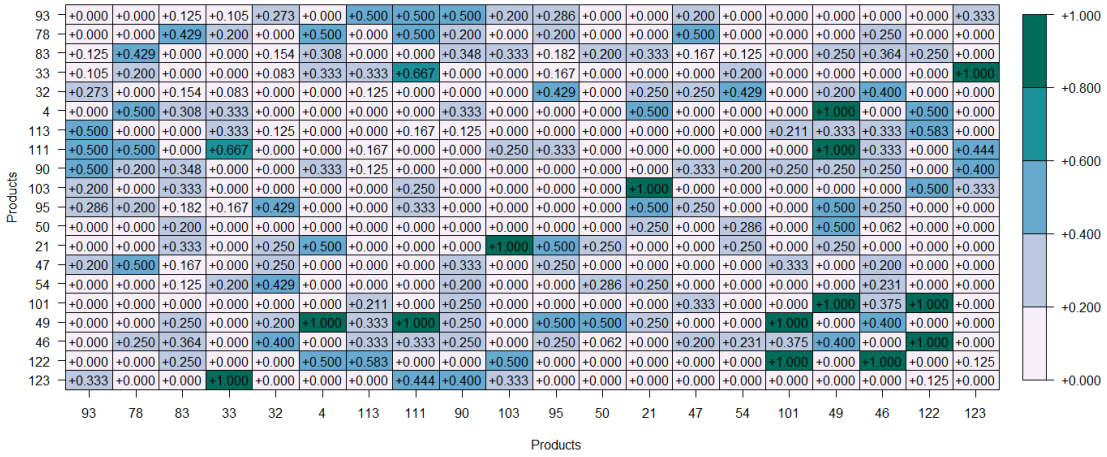



*Figure 16: Values of metric in each pair of 1st Cluster (I case)*

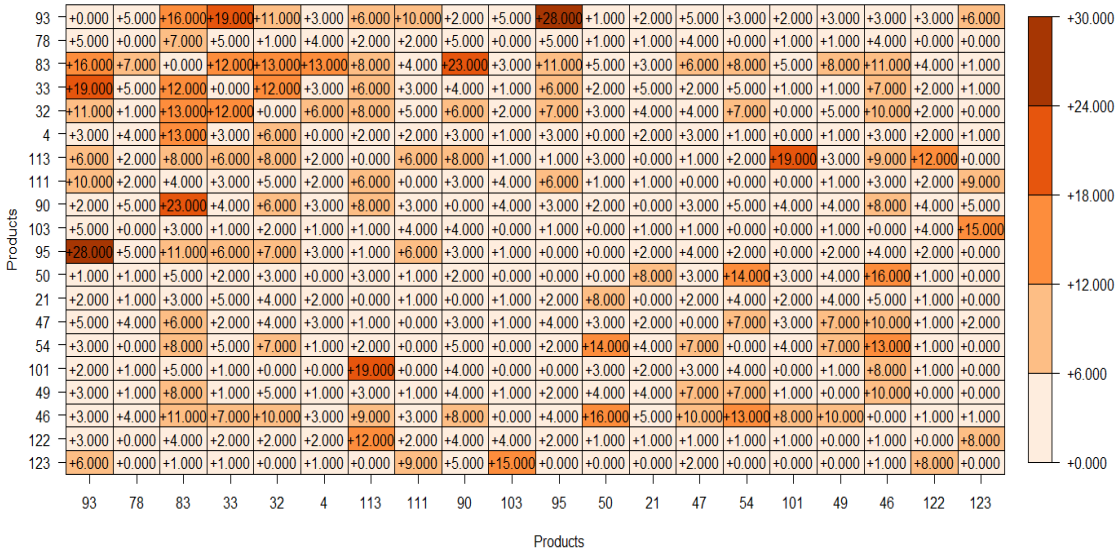



*Figure 17: Segmentation of transactions in each pair of 1st Cluster (I case)*

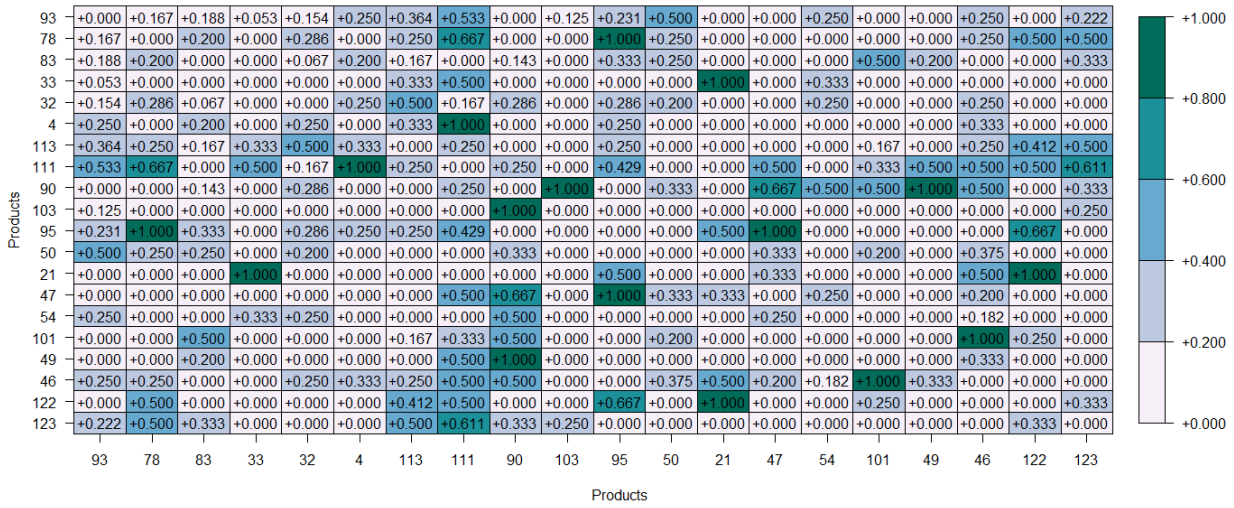



*Figure 19: Values of metric in each pair of 2nd Cluster (I case)*

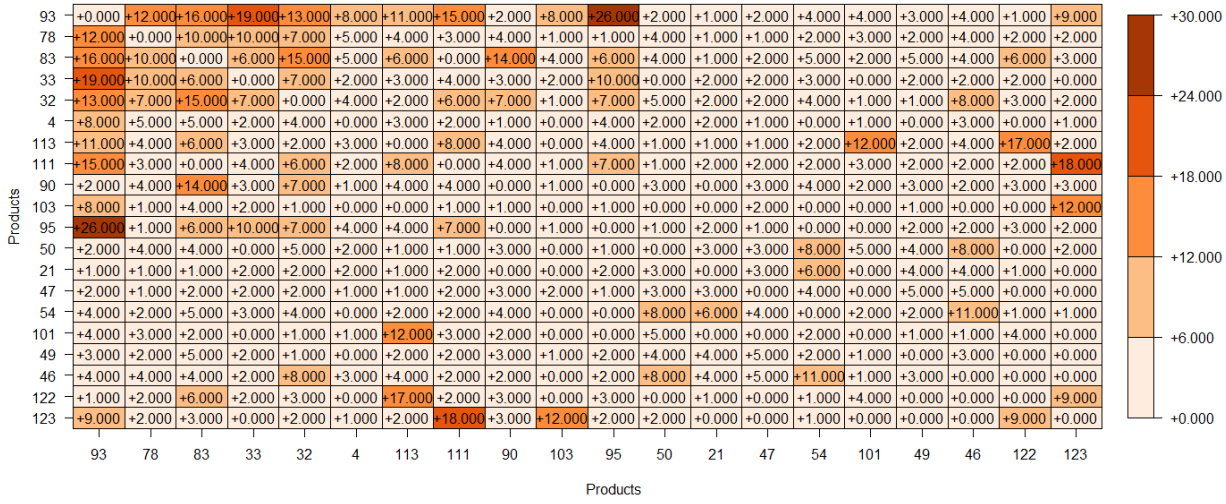



*Figure 20: Segmentation of transactions in each pair of 2nd Cluster (I case)*

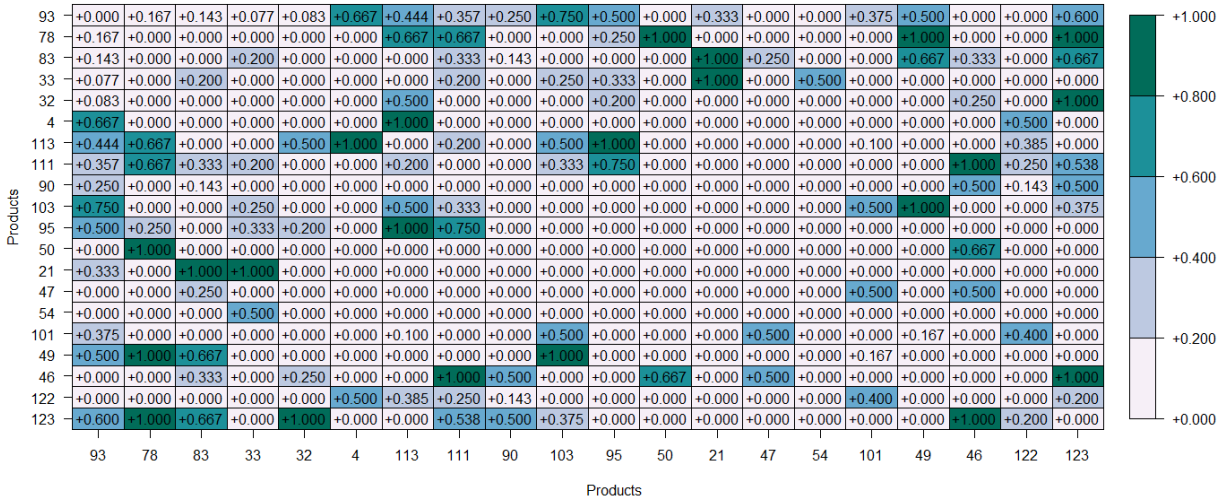



*Figure 21: Values of metric in each pair of 3rd Cluster (I case)*

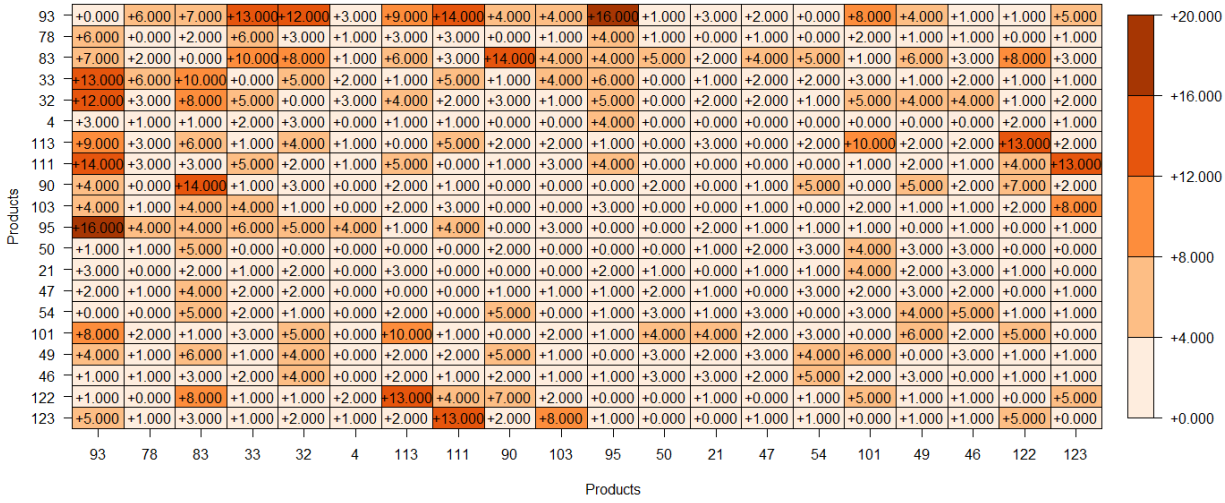



*Figure 22: Segmentation of transactions in each pair of 3rd Cluster (I case)*

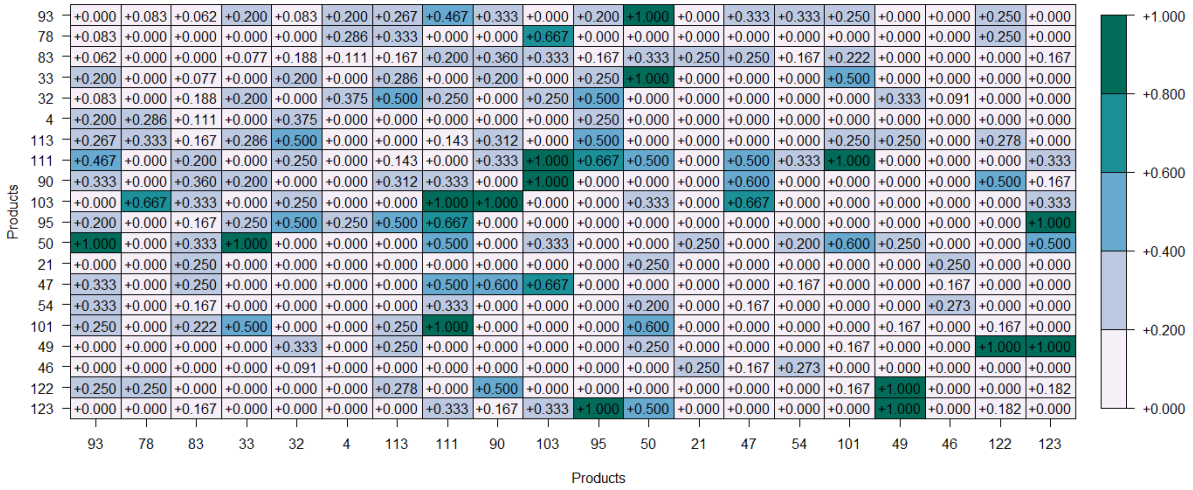



*Figure 23: Values of metric in each pair of 4th Cluster (I case)*

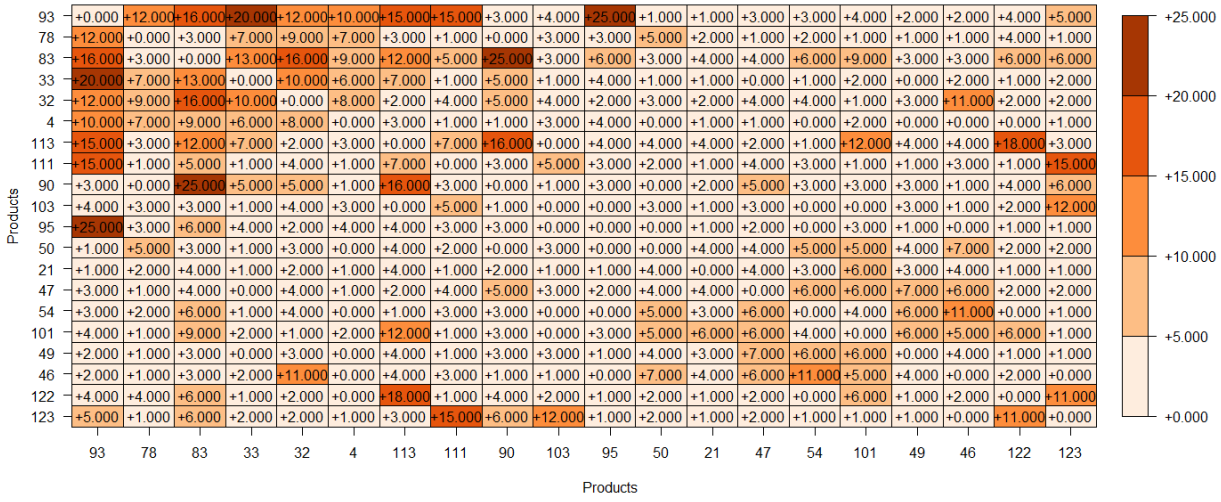



*Figure 24: Segmentation of transactions in each pair of 4th Cluster (I case)*

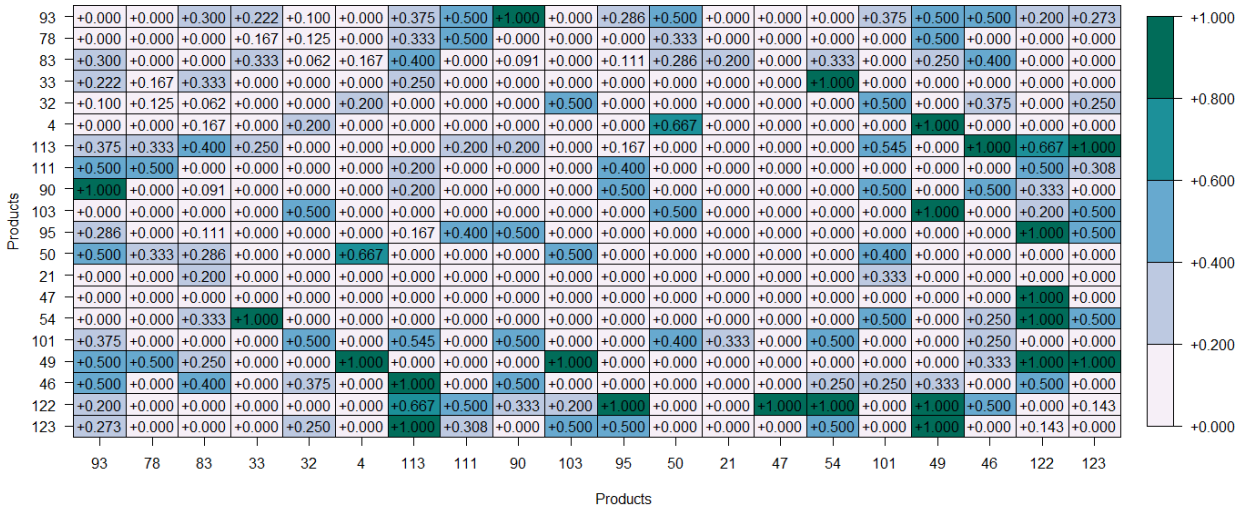



*Figure 25: Values of metric in each pair of 5th Cluster (I case)*

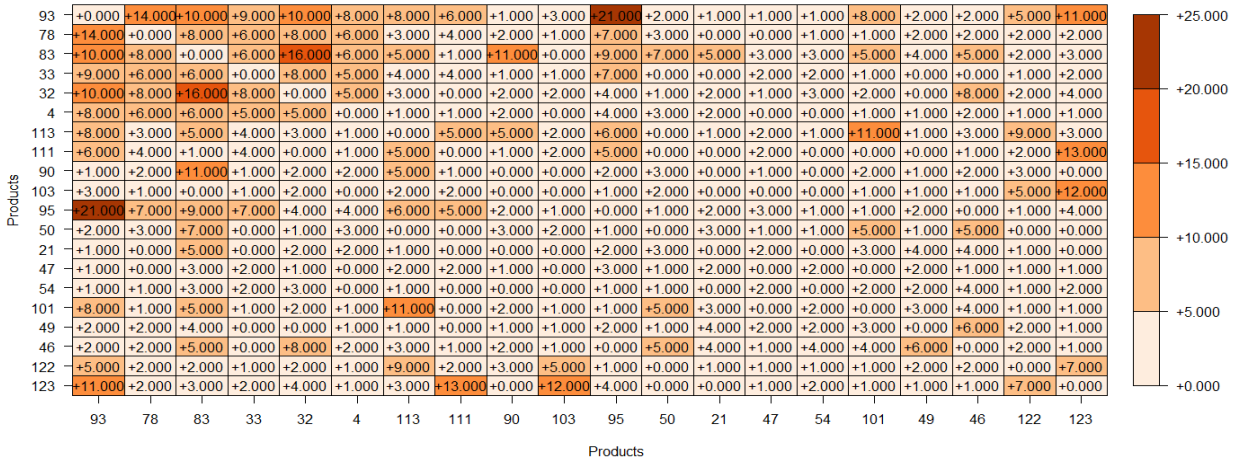



*Figure 26: Segmentation of transactions in each pair of 5th Cluster (I case)*

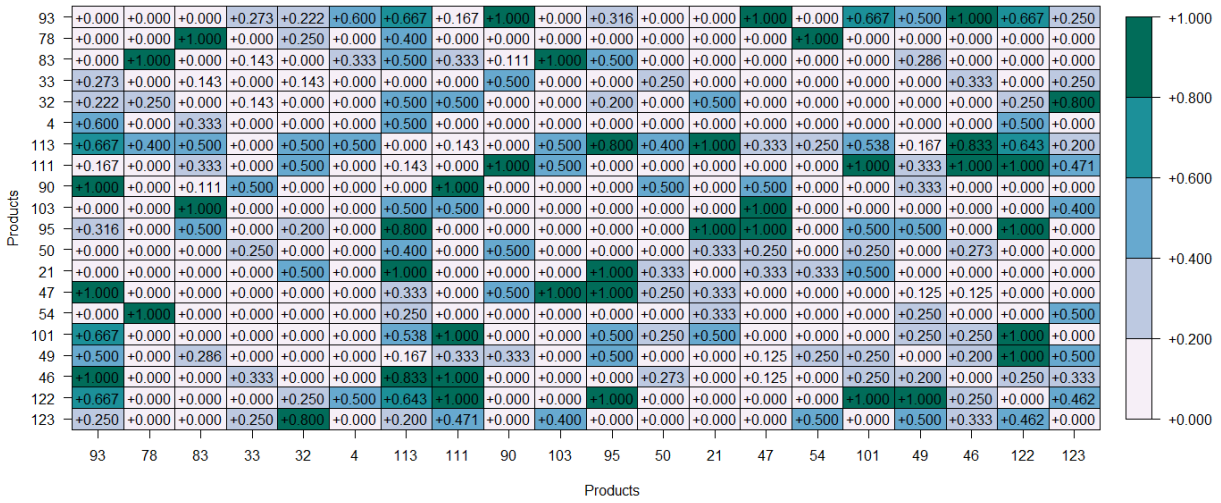



*Figure 27: Values of metric in each pair of 6th Cluster (I case)*

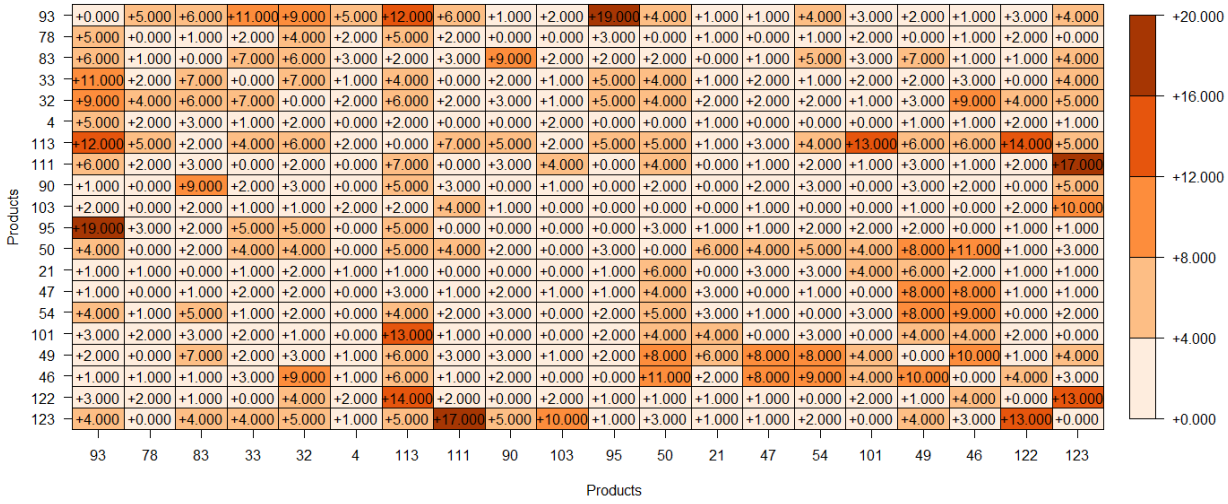



*Figure 28: Segmentation of transactions in each pair of 6th Cluster (I case)*



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

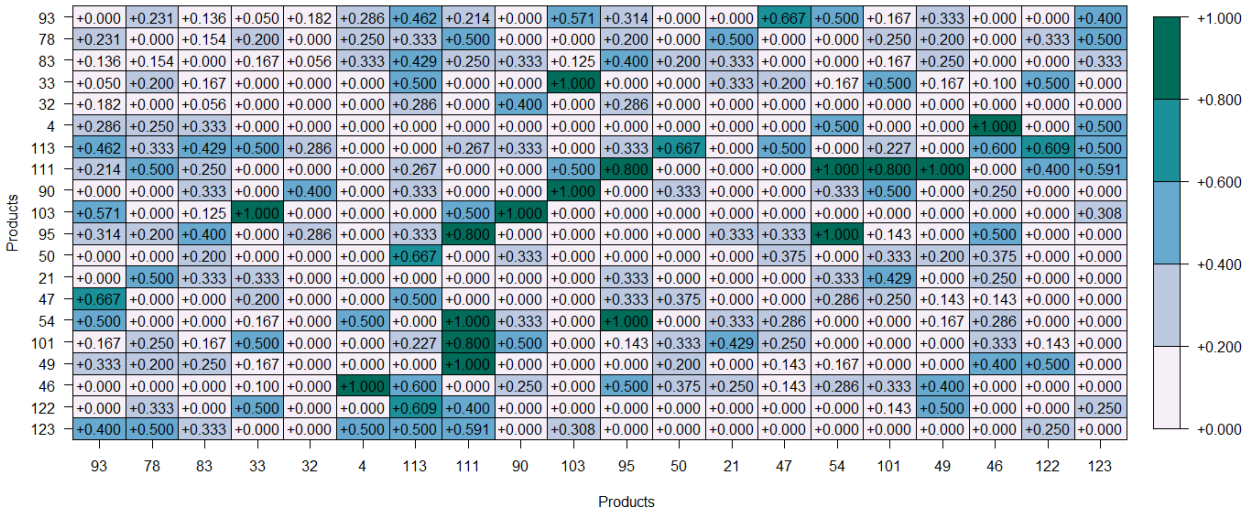



*Figure 29: Values of metric in each pair of 7th Cluster (I case)*

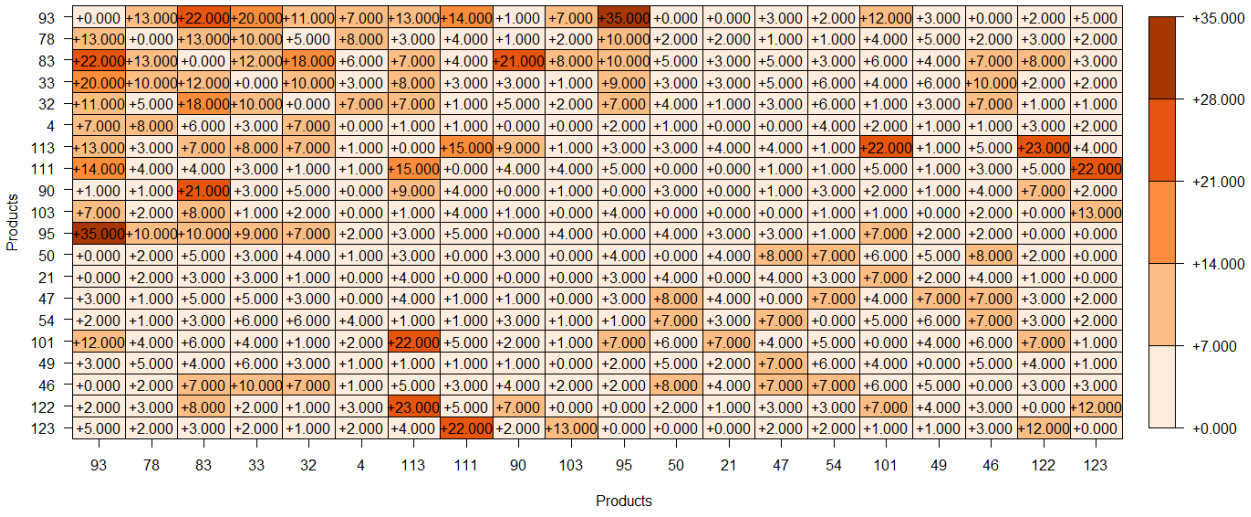



*Figure 30: Segmentation of transactions in each pair of 7th Cluster (I case)*

According to Figure 17, in Cluster 1 the largest number of transactions, 28, was observed in pair *(93, 95)* and there is percentage 28,6%, based on Figure 16, for switching between them. From Figure 18, it is illustrated that there was shift between item *(111, 123)* with percentage 61,1% and 16 transactions. Also, in Cluster 2 it is observed switching between items *(122, 113)* having 17 transactions and percentage 41,12%. Looking in Cluster 3 and in the shopping pair *(93, 95)* it is noticed switching percentage 50% and in shopping pair *(93, 111)* percentage 35,7%. Moreover, in Cluster 3, there were 13 transactions in pairs *(111, 123)* and *(113, 122)* with shift percentage 53,8%, 38,5% respectively. Exchange pairs that not observed in other 3 Segmentations was noticed in Cluster 4 such as *(93, 111)* with 15 purchases and percentage 46,7% and (83,90) with 25 purchases and percentage 36%. Comparing Cluster 1 with Clustering 5, in both segments there was percentage switching percentage 28,6% in transaction pair *(93, 95)* but in Cluster 5 21 observed transactions. From Figure 24, in segmentation 5 we inform that the largest percentage of switching, 66,7%, was in pair *(113, 122).* Looking all above figures, the significant percentage 83,3% is in Cluster 6, strong evidence for switching the pair (*46, 113)* but the transactions was only 6. Then, in Cluster 6 there was percentage 66,7% in pair (*93, 113*) with 12 purchases. Percentages above 50% were observed in shopping pairs *(111, 123)* and *(113, 122)* with 22 and 23 transactions respectively. Finally, the large number of transactions, 35, was noticed in purchases of items (93, 95) and having percentage of switching 31,4%.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

### 4.3.2 Clustering in the Case II of Multinomial Data

In the assumption that data follows multinomial distribution, the ICL score gives us the following plot with different number of clusters.

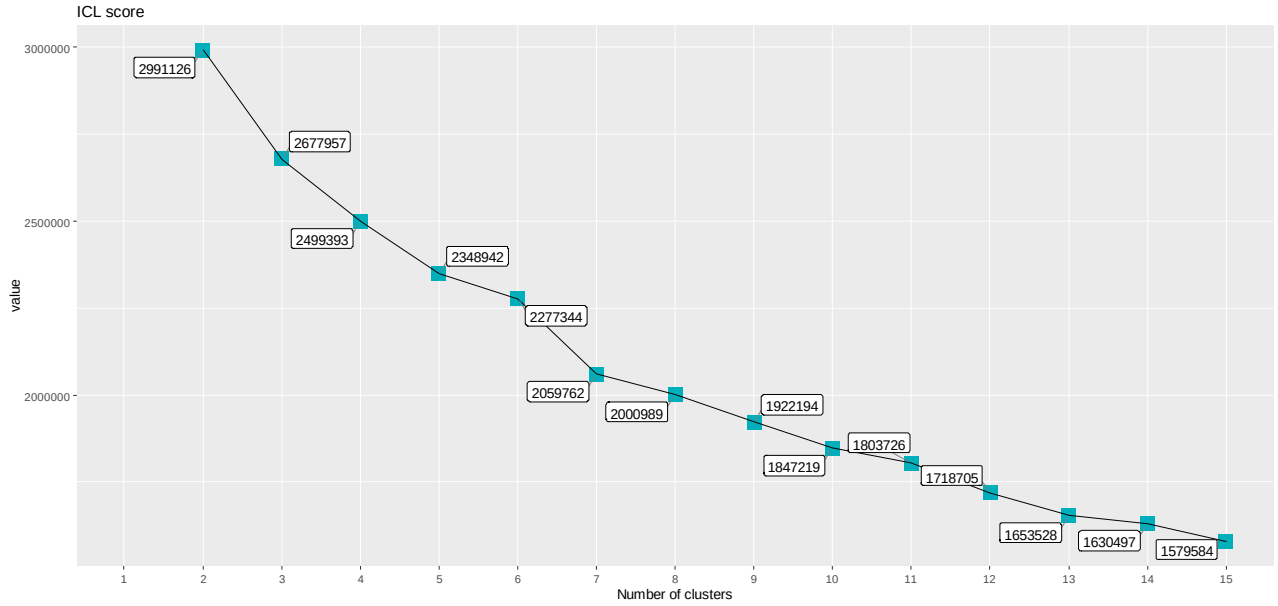



*Figure 31: ICL values in the assumption of multinomial data*

As seen in figure 31, when we changed the cluster value from 6 to 7, the ICL value reduced very sharply. This decrease in the ICL value reduces and eventually becomes constant as we increase the number of clusters further. Then, the cluster value where this decrease in ICL value becomes constant can be chosen as the right cluster value for our data. We can choose any number of clusters between 6 and 10. We can have 7, 8, or even 9 clusters and if we increase the number of clusters, the computation cost will also increase. Our choice is 7 with the ICL value be 2.059.762. After dividing the customers in 7 segments, the graph below is generated and shows us how the customers are distributed.

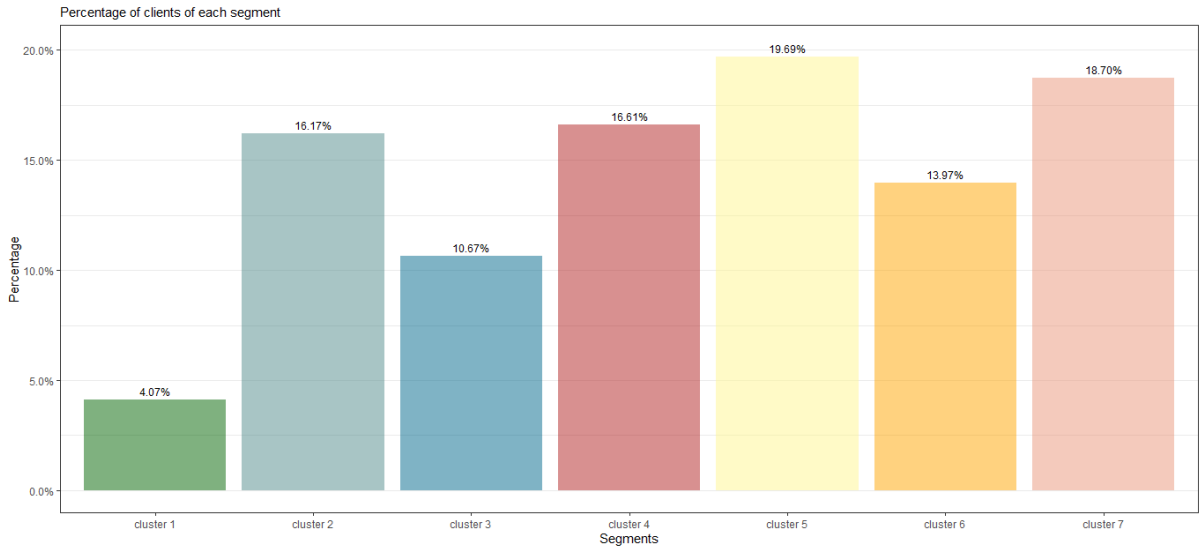



*Figure 32: Number of customers per segmentation in the assumption of Multinomial Data*

In a first view, in Cluster 1 only 4% percentage of total clients was consisted of, it is assumed a special category segmentation that leads us to different conclusions. On the other side, the higher percentage was detected in Cluster 5, 19,7%.

And in this case, it is calculated the ratio of the mean of every cluster to the mean of total customers and the below graphs depict how "far" is each cluster from the average customer.

After splitting data in 7 cluster number, the metric $Q$ per cluster is created as it shown in the following figures.

## *Plots in the assumption of Multinomial Data (II case)*

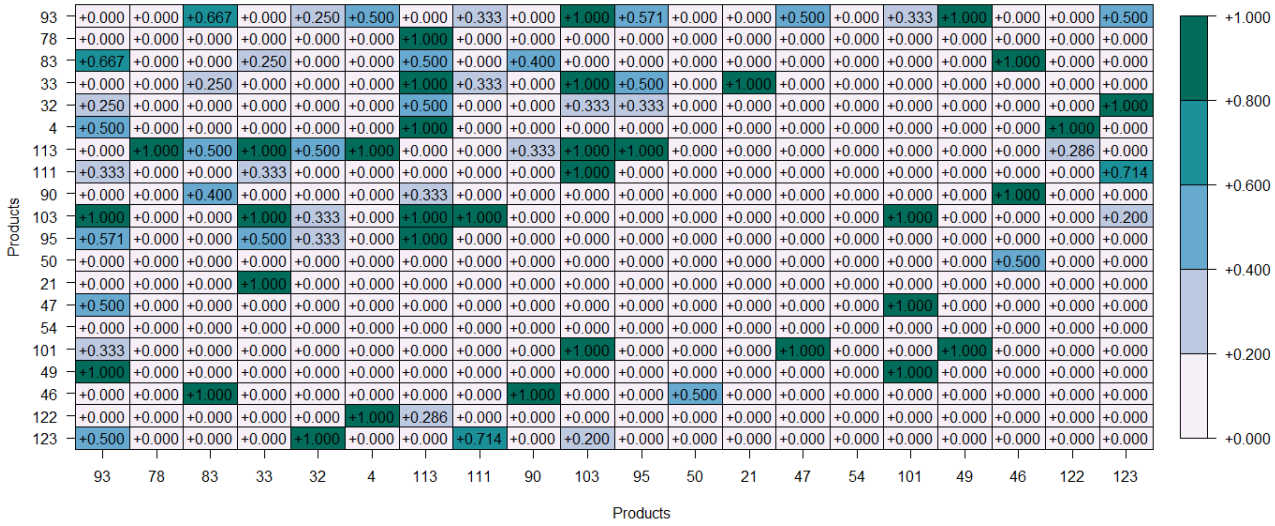



*Figure 33: Values of metric in each pair of 1st Cluster (II case)*

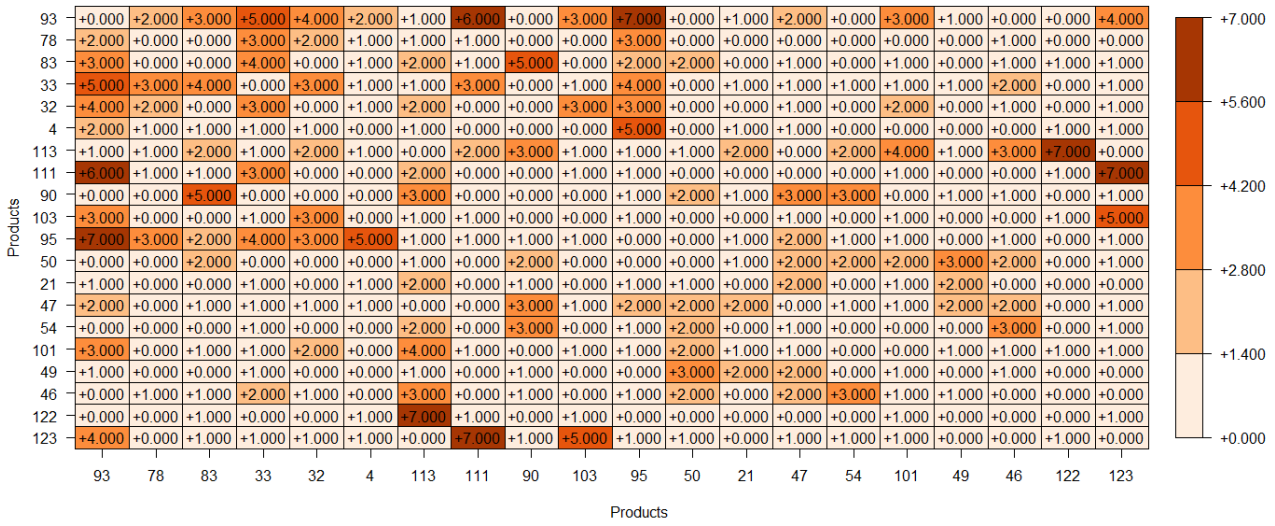



*Figure 34: Segmentation of transactions in each pair of 1st Cluster (II case)*

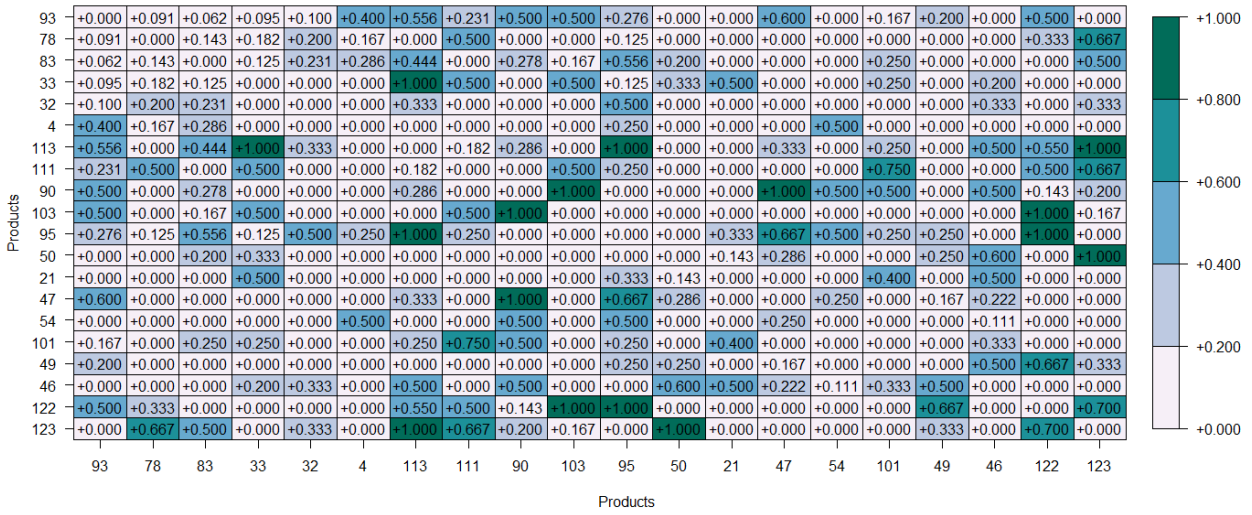



*Figure 35: Values of metric in each pair of 2nd Cluster (II case)*

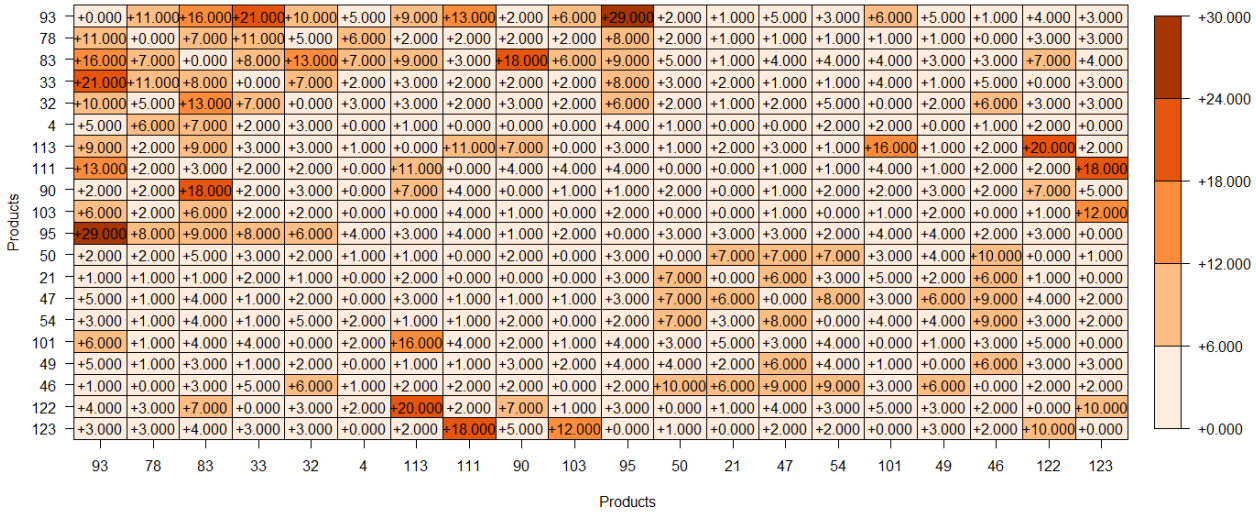



*Figure 36: Segmentation of transactions in each pair of 2nd Cluster (II case)*

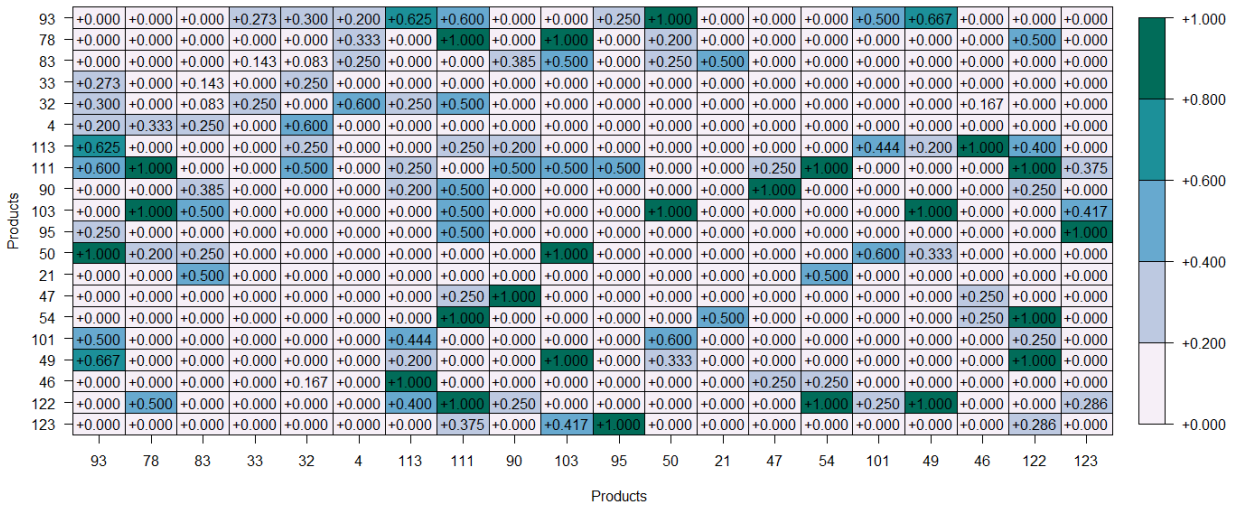



*Figure 37: Values of metric in each pair of 3rd Cluster (II case)*

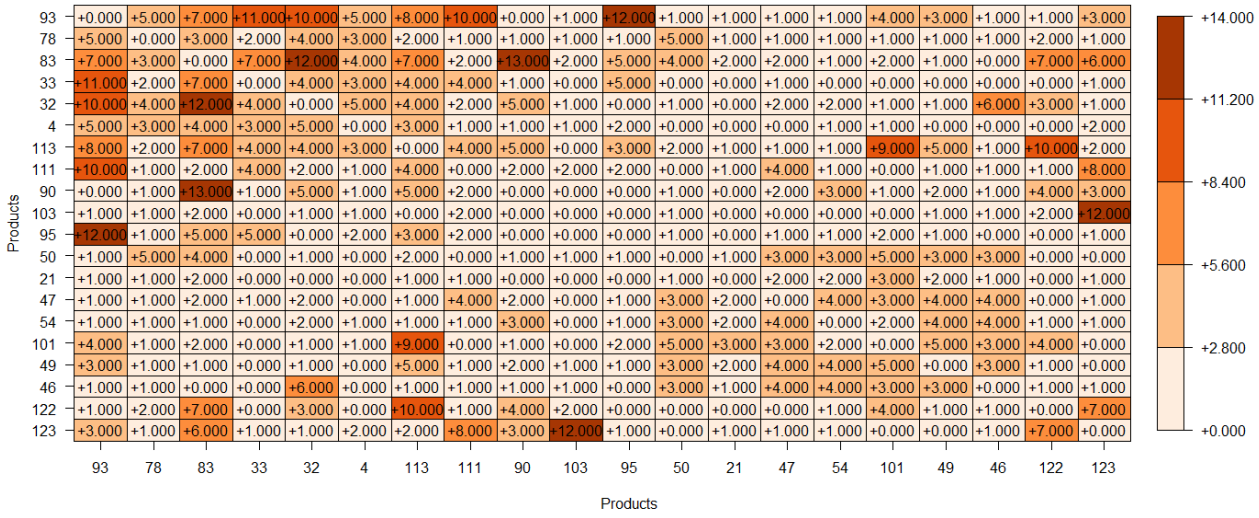



*Figure 38: Segmentation of transactions in each pair of 3rd Cluster (II case)*

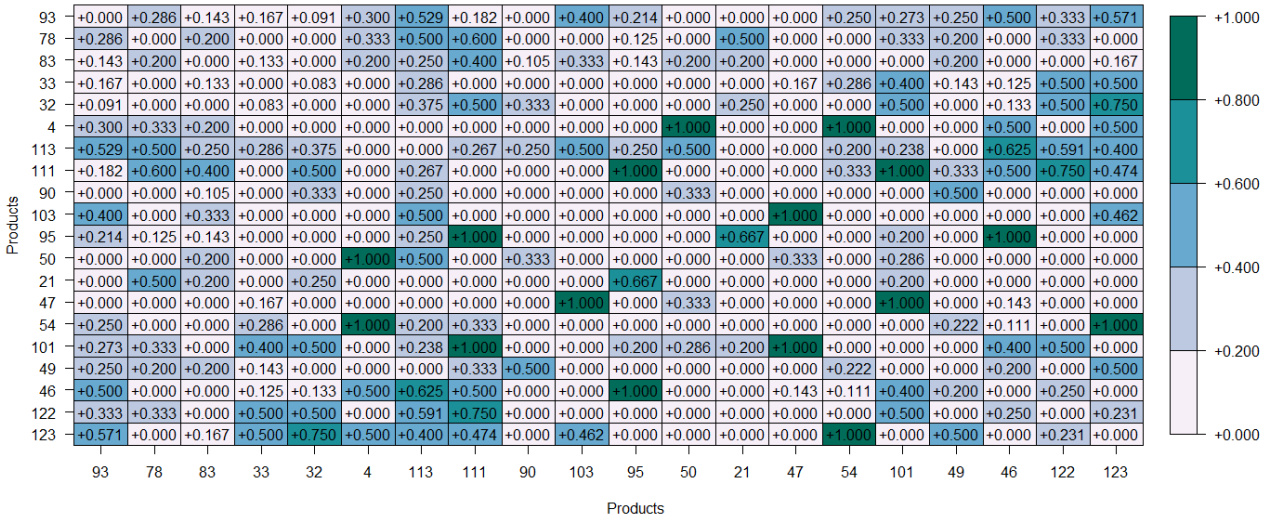



*Figure 39: Values of metric in each pair of 4th Cluster (II case)*

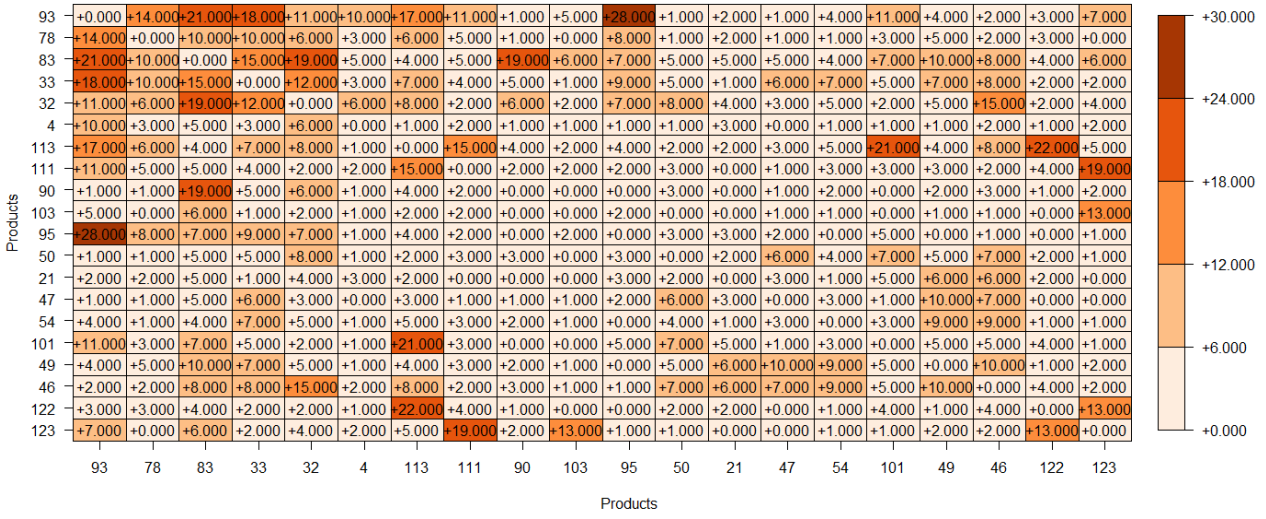



*Figure 40: Segmentation of transactions in each pair of 4th Cluster (II case)*

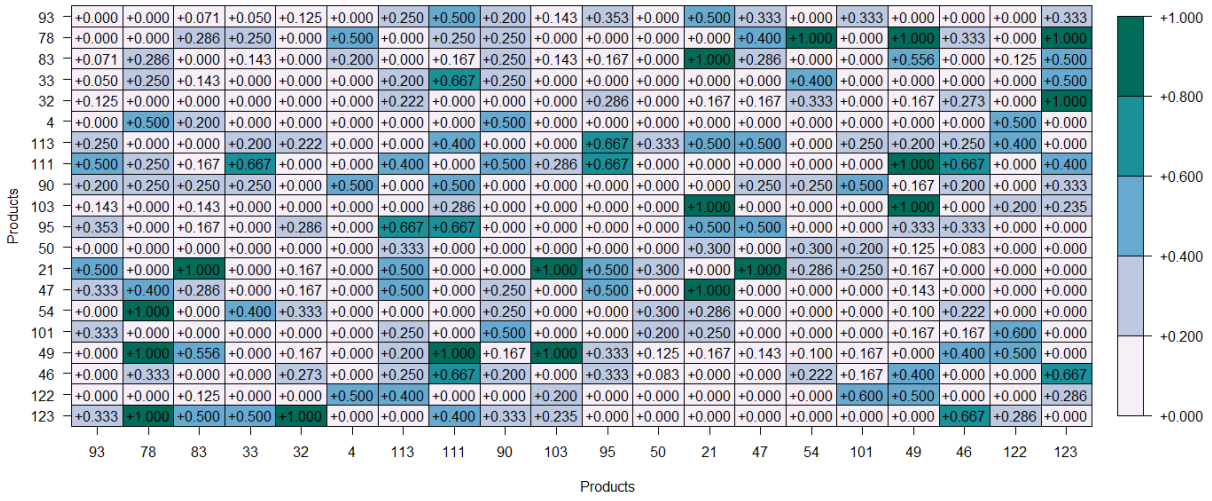



*Figure 41: Values of metric in each pair of 5th Cluster (II case)*

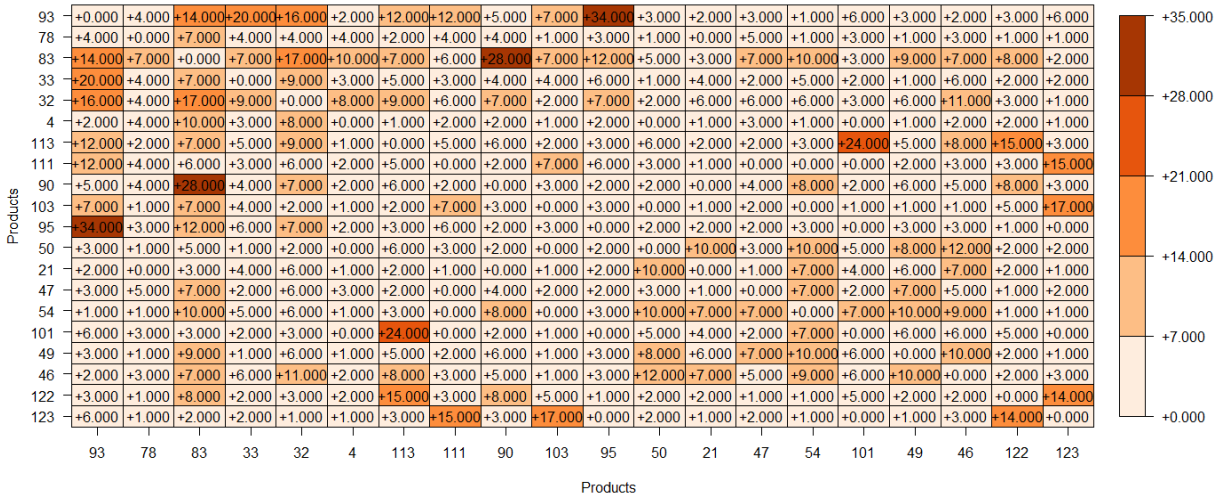



*Figure 42: Segmentation of transactions in each pair of 5th Cluster (II case)*

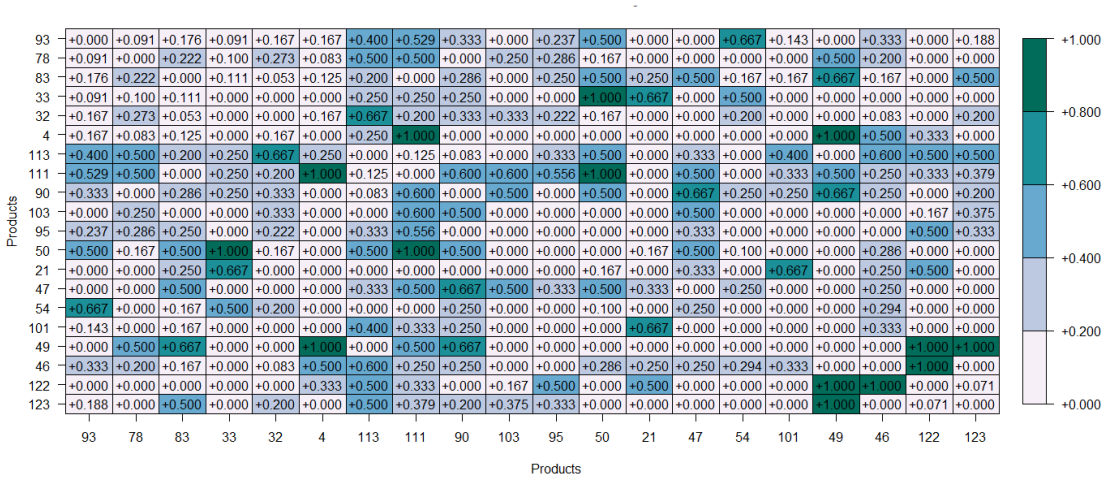



*Figure 43: Values of metric in each pair of 7th Cluster (II case)*

According to Figure 34, also in this case the pairs *(93, 95)* and *(111, 123)* had percentage of 57,1% with deficient transactions, only 7. Observing the Cluster 2, we could conclude that there was strong evidence of switching between pair *(93,95),* pair *(111,123),* pair *(122, 113)* which was verified with a percentage that varied 50%-65%. Furthermore, shopping pairs *(103, 123)*, *(90, 83)*, that had not observed in other clusters and in Case I with percentages 41,7% and 38,5%, were noticed in Cluster 3. In segmentation 4, although there were 28 transactions in pair *(93, 95)*, the percentage of shift is low 21,4%, the lowest in Case II. Also, in this segmentation and in pair (*93, 113)*, we had 17 transactions with percentage 52,9% and this similar percentage (62,5%) was observed in Cluster 3 with the cost of less transactions, 8. From Figure 43, 34 customers of our data purchased pair *(93, 95)* with a percentage of switching shopping behavior 35,3%. Also, 15 transactions were noticed in pairs *(111, 123)*, *(122, 113)* with strong evidence of switching in 40%. It's worth to mention that in Cluster 7, there is switching percentage 40% in pair *(93, 113)* and 52,9% in pair *(93,111).*

### 4.3.3 Compare profile of Clustering in 2 Cases

In order to have a clear figure of Customers segmentation, the optimization of clusters is a solution. For each segmentation, we discover the average profile of customers in 2 Cases of assumptions.

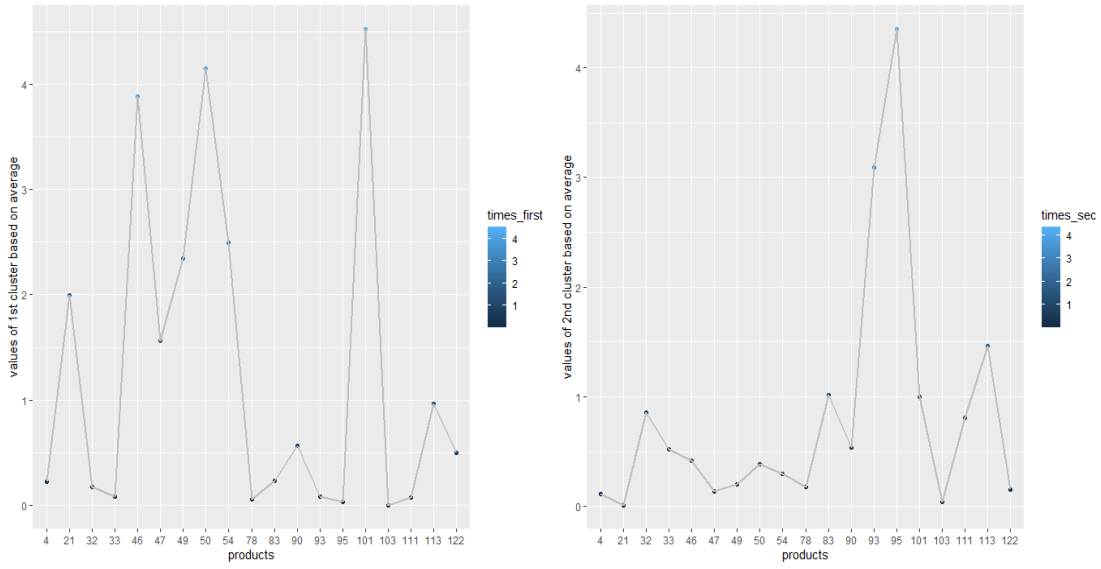
As we mentioned in section 3.1.1, in Case I, we use data that provide us with relative frequency of a customer purchasing the corresponding product. From the other Case, it is used the row data with the units of items are purchased by each customer. Then, it created a scale $R$ that shows the market share, i.e. percentage of the sales. The scale $R$ is the following:

$$R_{ig} = \frac{mean\ of\ frequency\ of\ product\ i\ in\ cluster\ g}{mean\ of\ frequency\ of\ product\ i} \qquad (4.1)$$

$$where\ i = 1, 2, \dots, 20,$$

$$g = 1, 2, \dots, 7$$

*Profile plots in the assumption of Multivariate Normal Data (I case)*



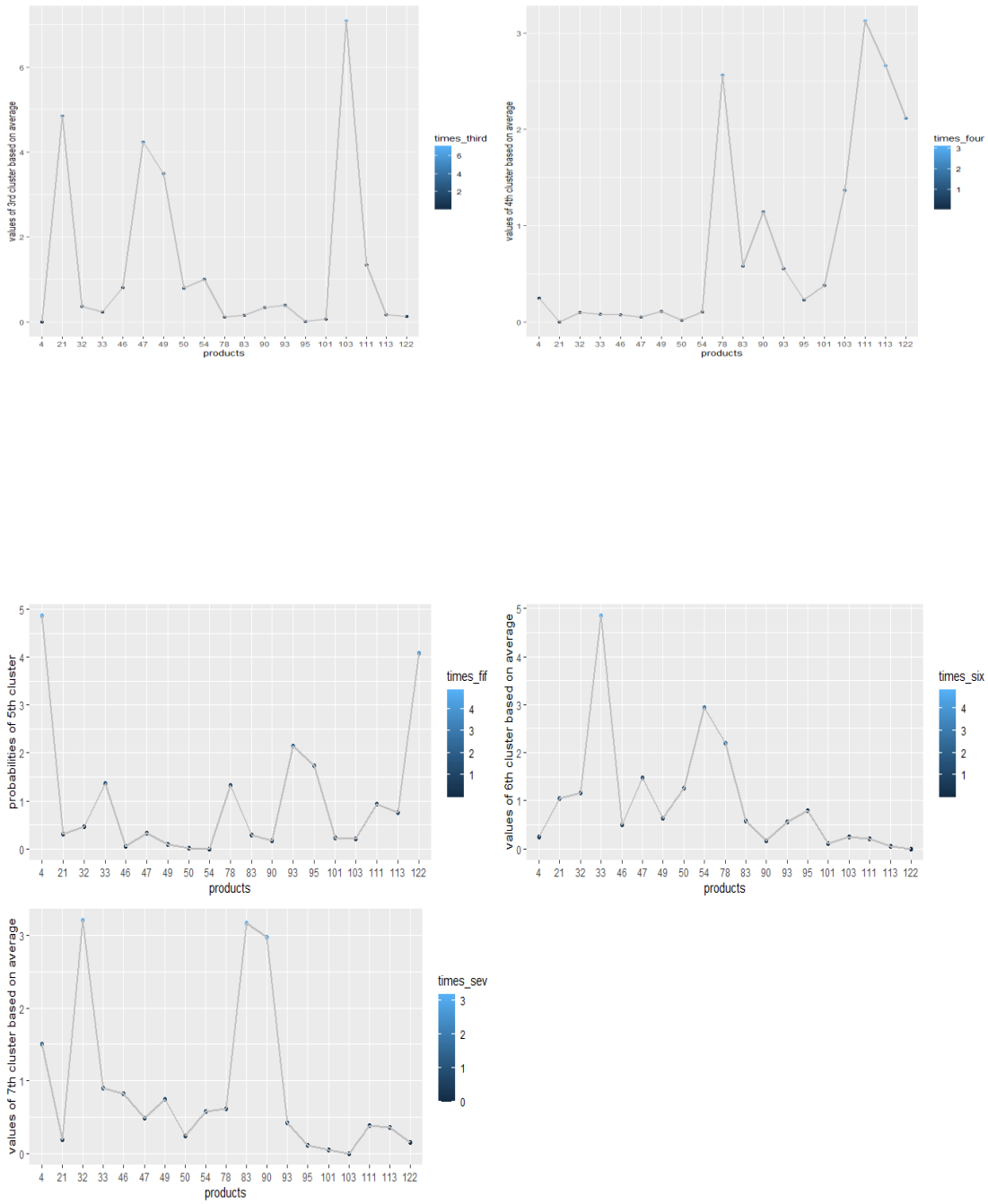
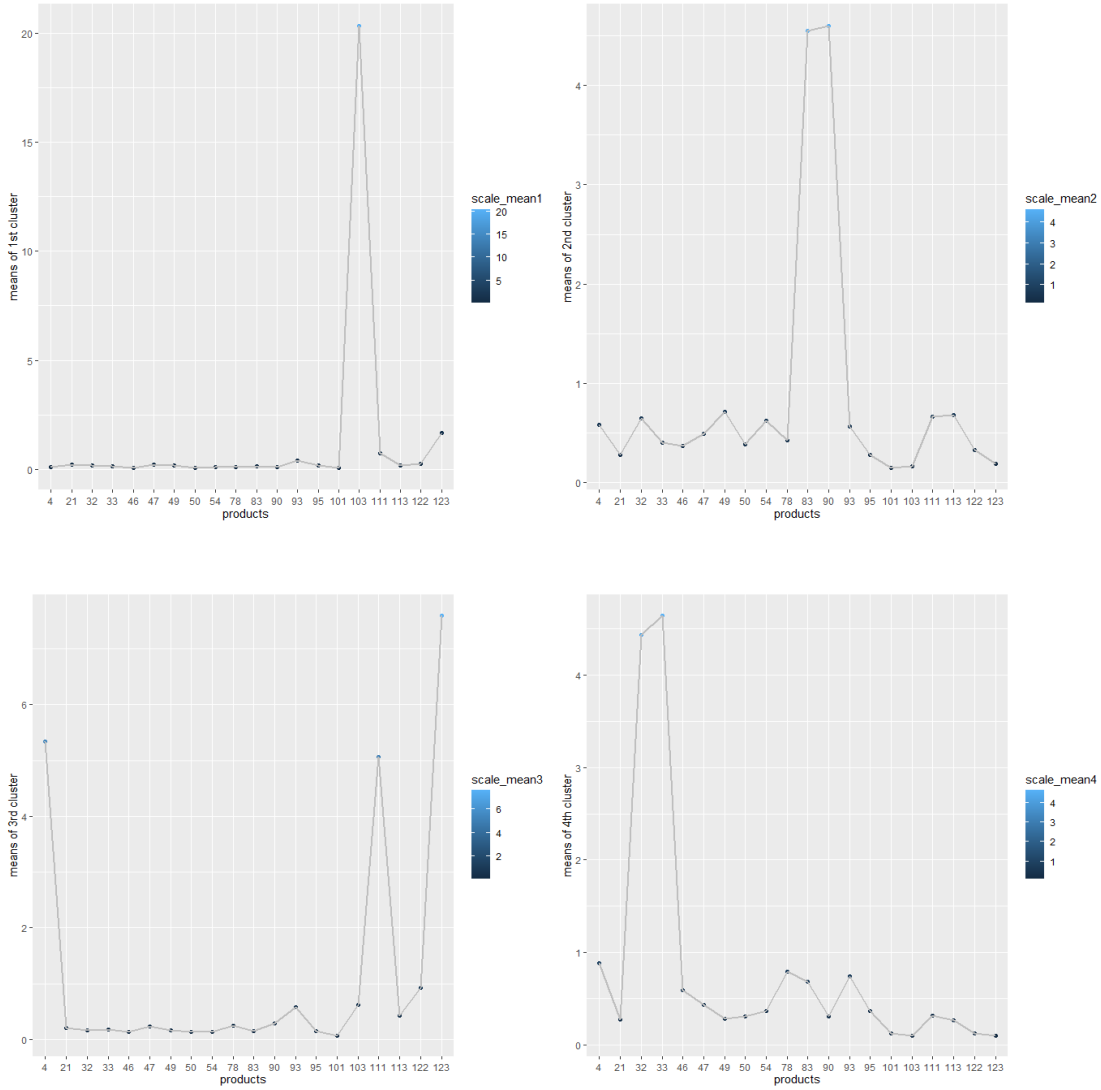



*Figure 44: Plots of Average Profile of 7 Clusters to the mean of Customers in the assumption of Multivariate Normal Data*

*Profile plots in the assumption of Multinomial Data (II case)*



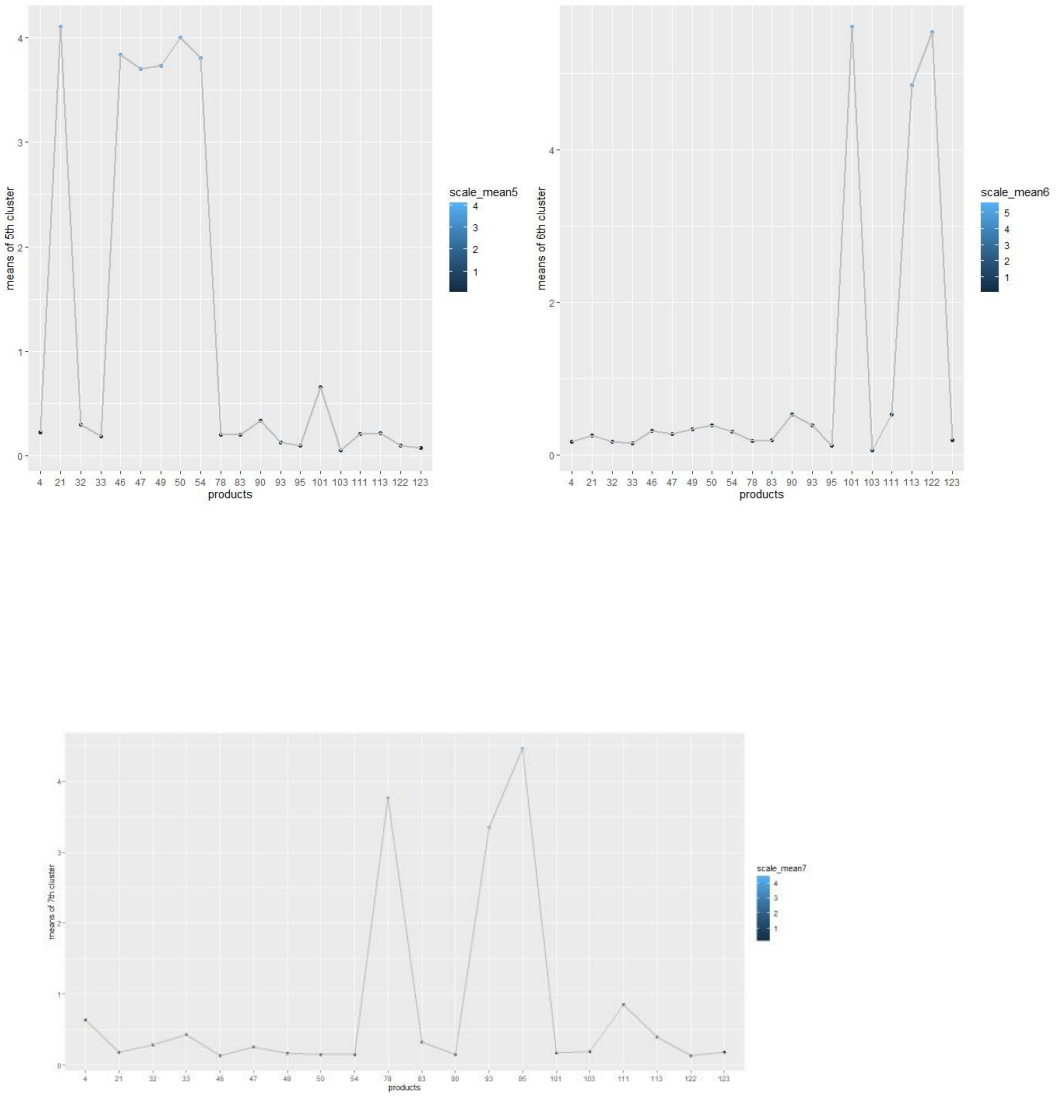



*Figure 45: Plots of Average Profile of 7 Clusters to the mean of Customers in the assumption of Multinomial Data*

As seen in Figures 43 and 44, in the 5th plot and 3rd plot respectively, item Brand A Concentrated Lemon Juice 1L *(4)* purchased almost 5 times more than the average of customers. Similar behavior we observe in the Cluster 2 (2nd figure), in Case I and in the Cluster 7 (7th figure), in Case II where item Brand A Orange Juice BP 1,50L *(95)* was bought 4 times more. It is worth to mention that in the 1st Figure and Case I, item Brand D Orange Juice 1L *(103)* was bought 20 times more than the average customer, it is a discrimination between the other clusters since Cluster 1 contains inadequate customers for remarkable conclusions. Moreover, clients of Cluster 6 in Case I and in Case II bought items Brand C Orange Juice without pulp 900ML *(101)* almost 5 times more than average clients.

## 4.4 Summary of the results of the implementations

First, we will present the results of model construction, observing the Figure 11 in section 4.1 which consists of high values of switching in pairs, the Figure 3 in section 2.1 with the number of transactions in each pair and the below figure which present all values of metric in each pair.

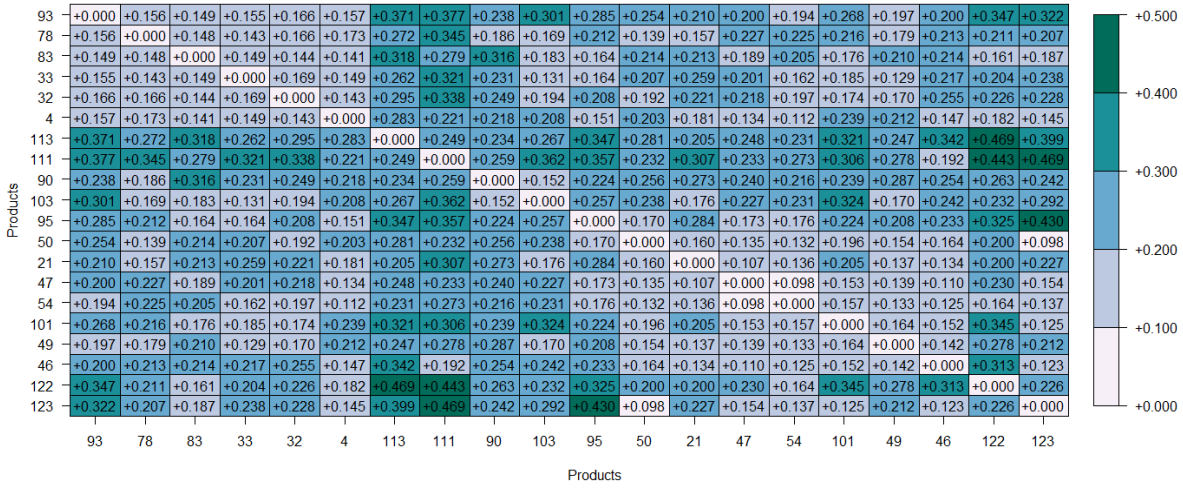



*Figure 46: Values of metric in each pair*

We could easily conclude that this process is able to detect similarities of our items with flavor of orange. Specifically, item Brand B Orange Juice without pulp PET packaging 90CL *(113)* and item Brand B Orange Juice without pulp 1 L *(122)* act similar in preference of our customers. Also, in this pair customers chose Brand B and flavor orange without pulp. Furthermore, item Brand B Orange Juice with pulp PET packaging 90CL (111) and item Brand B Orange Juice with pulp PET packaging 90CL (123) are similar and Brand B plays an important role regardless of packaging and existence of pulp. Similar behavior is observed in item Brand B Orange Juice without pulp 1 L *(122)* and Brand B Orange Juice with pulp PET packaging 90CL *(111)*.

On other side, customers was hesitated to choose item Brand A Apple Juice 1L *(83)* and item Brand B Apple Juice PET packaging 90CL *(90)*. This means that their choice was depended on taste, on brand and not on kind of packaging and capacity of juice.

As previously mentioned, we fit two different distributions in our data, the multinomial distribution and multivariate normal distribution. Although the assumption of multivariate normal distribution is incoherent due to zero elements, the results are similar Both in two assumptions, this procedure is able to discover similarities of products with flavor orange. Specifically, the pairs *(111, 123)*, *(122,113)*, *(93,95)*,

*(93,113)*, and *(93,111)* are detected in two cases of distributions and there is strong evidence that items in each pair behave similar in our data set.

Let us now demonstrate the results of hierarchical clustering in the 20 items of our category, using the figure 12. For this attempt it is created the following table with the description of each item in a cluster.

| **Cluster 1** | **Cluster 2** | **Cluster 3** | **Cluster 4** |
|---|---|---|---|
| 101: Brand C Orange Juice without pulp 900ML | 47: Brand C Bilberry/Black Currant/Cranberry Juice 900ML | 122: Brand B Orange Juice without pulp 1 L | 54: Brand C Pear/Apple/Peach Juice 900ML |
| 103: Brand D Orange Juice 1L | 78: Brand A Grapefruit Juice 1L | 113: Brand B Orange Juice without pulp PET packaging 90CL | 49: Brand C Apple/Raspberry 900ML |
| 4: Brand A Concentrated Lemon Juice 1L | | 93: Brand A Orange Juice 1L | 83: Brand A Apple Juice 1L |
| 50: Brand C Mixed Fruit Juice 900ML | | 123: Brand B Orange Juice 1 L | 90: Brand B Apple Juice PET packaging 90CL |
| | | 111: Brand B Orange Juice with pulp PET packaging 90CL | |
| | | 95: Brand A Orange Juice BP 1,50L | |
| | | 46: Brand C Ananas/Passion Fruit Juice 900ML | |
| | | 32: Brand A Ananas Juice 1L | |
| | | 21: Brand C Orange/Carotte Juice 900ML | |
| | | 33: Brand A Clementine Juice 1L | |

*Table 6: Hierarchical clustering of 20 items*

A notable finding is that in Cluster 3 it is concentrated the largest number of products and in particular the taste of orange apart from 3 items, *46, 32, 33*. Moreover, in Cluster 3 there are three items that consist of the fruit apple. As we can easily notice that in Cluster 2 is concentrated taste of juices that are not observed in other segmentations, bilberry, black currant, cranberry, and grapefruit.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Chapter 5

# Conclusion and discussion

Applying market analysis, shop owners can properly comprehend their target audience and the conditions of the market. Furthermore, they will also distinguish themself from great competition that is observed in this field. The shop owners must have a wide variety of foods and beverage to cover large part of the crowded market. In conclusion the results from this research on the category of "juices" recommend which items on the category of "juices" it is necessary to be provided on the store shelves. Products that have flavor orange there must be, since orange juice are the most popular juice consumed. Also results show that dividing the products into a few groups, juices that contain apples are gathered in a segmentation. The preferences of customers that we examined depend on the flavor of the juices and mainly on brand. The type of packaging, the capacity of the juice and the content of filtered juice do not play a role in our analysis. In conclusion, our method of finding distances of FMCG products, it can be used accordingly to collect data from some other product category.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Appendix

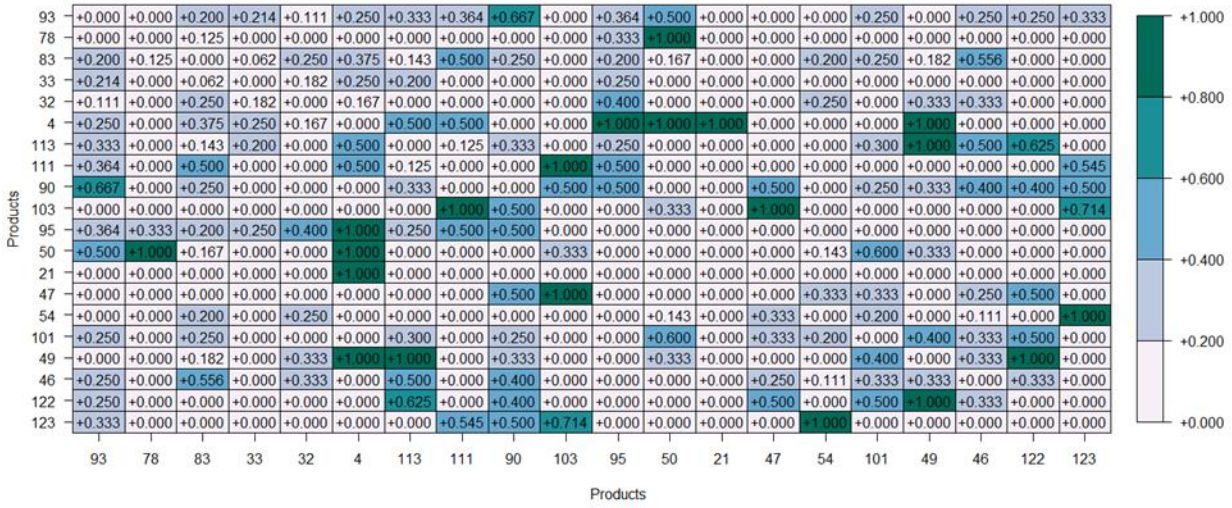



*Figure 47: Values of metric in each pair of 6th Cluster (II case)*

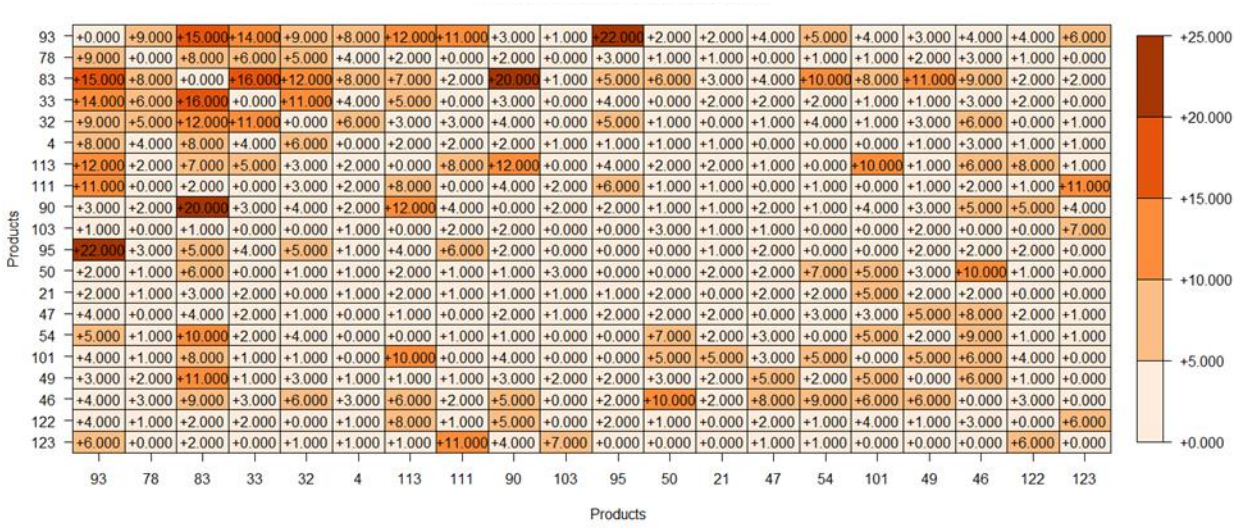



*Figure 48: Segmentation of transactions in each pair of 6th Cluster (II case)*

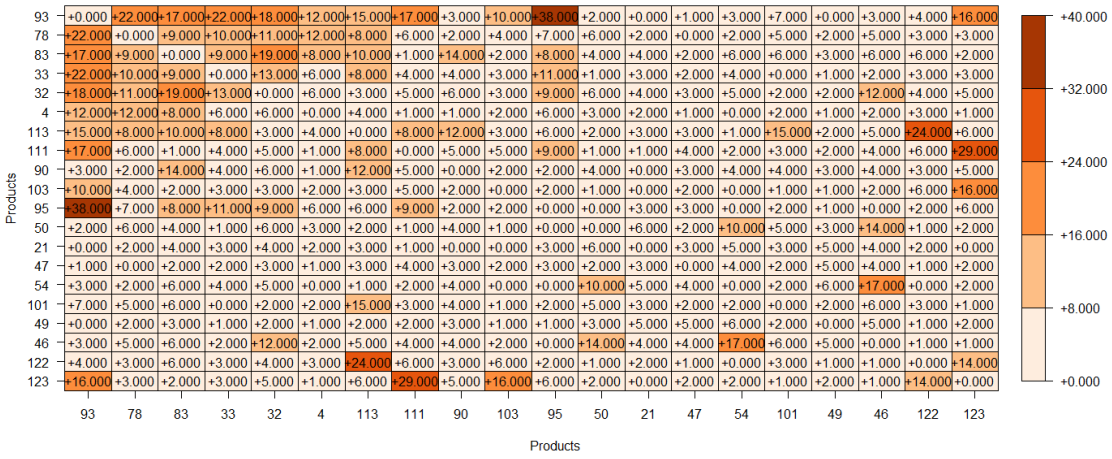



*Figure 49: Segmentation of transactions in each pair of 7th Cluster (II case)*

# References


**"A study on Consumer Buying Behaviour towards Selected FMCG Products"** International Journal of scientific research and management, 2014

**Agrawal, R., & Srikant, R. (1994).** Fast algorithms for mining association rules. In: Proceedings of the 20th VLDB Conference

**Agrawal, R., Imieliński, T., & Swami, A. (1993).** Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2), 207-216.

**Baudry, J. (2015).** Estimation and model selection for model-based clustering with the conditional classification likelihood. Electronic Journal Of Statistics, 9(1).

**Cubaynes, S., Lavergne, C., Marboutin, E., & Gimenez, O. (2012).** Assessing individual heterogeneity using model selection criteria: how many mixture components in capture-recapture models?. Methods In Ecology And Evolution, 3(3), 564-573.

**Dr.Vibhuti, Dr. Ajay Kumar Tyagi and Vivek Pandey (2014),** "A study on Consumer Buying Behaviour towards Selected FMCG Products" International Journal of scientific research and management (IJSRM) Volume 2, Issue 8,Pages 1168-1182

**Finucan, H. (1964).** The Mode of a Multinomial Distribution. Biometrika, 51(3/4), 513.

**Haberman, S. (1976).** Review: Y. M. M. Bishop, S. E. Fienberg, P. W. Holland, Discrete Multivariate Analysis: Theory and Practice. The Annals Of Statistics, 4(4).

**Kass, R. E., & Raftery, A. E. (1995).** Bayes factors. Journal of the American Statistical Association, 90(430), 773-795.

**Mun, E., von Eye, A., Bates, M., & Vaschillo, E. (2008).** Finding groups using model-based cluster analysis: Heterogeneous emotional self-regulatory processes and heavy alcohol use risk. Developmental Psychology, 44(2), 481-495.

**Murtagh, F. (2015).** "Hierarchical Clustering", Department of Computer Science, Royal Holloway, University of London, Egham TW20 0EX, England.

**Novak, T. (1993).** Log-Linear Trees: Models of Market Structure in Brand Switching Data. Journal Of Marketing Research, 30(3), 267.

**Pathak, K., Silakari, S., & Chaudhari, N. (2017).** Privacy Preserving Informative Association Rule Mining. International Journal Of Applied Information Systems, 12(8), 1-7.

**Peel, D., & McLachlan, G. (2000).** Journal search results Statistics And Computing, 10(4), 339-348.

**Polyethylene terephthalate - Wikipedia. (2021).** https://en.wikipedia.org/wiki/Polyethylene_terephthalate




ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

**Raftery, A. (1995).** Bayesian Model Selection in Social Research. Sociological Methodology, 25, 111.

**Roux, M. (2018),** "A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms", Journal of Classification, Springer Verlag.

**Scrucca L., Fop M., Murphy T. B. and Raftery A. E. (2016)** mclust 5: clustering, classification and density estimation using Gaussian finite mixture models The R Journal 8/1, pp. 289-317

**Sinz, Fabian; Gerwinn, Sebastian; Bethge, Matthias (2009).** "Characterization of the p-generalized normal distribution". Journal of Multivariate Analysis. 100 (5): 817–820

**Tatiana Benaglia, Didier Chauveau, David R. Hunter, Derek Young (2009).** mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29. URL http://www.jstatsoft.org/v32/i06/.

**Vijay Kotu, Bala Deshpande**, in Data Science (Second Edition), 2019

**Why Fast-Moving Consumer Goods Matter. (2021)**
https://www.investopedia.com/terms/f/fastmoving-consumer-goods-fmcg.asp