ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

**ΠΑΥΛΟΣ ΑΒΡΑΜΙΔΗΣ**

# Τεχνικές Εξόρυξης Δεδομένων στον Τομέα Ηλεκτρονικού Λιανικού Εμπορίου με Σκοπό την Ανάπτυξη Στρατηγικών Marketing

**Επιβλέπων :** Γεώργιος Λεκάκος – Αναπληρωτής Καθηγητής
**Κύριος συντελεστής :** Ανδρέας Ζάρας – Data Scientist

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

Αθήνα, Ιούλιος 2022


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

**ΠΑΥΛΟΣ ΑΒΡΑΜΙΔΗΣ**

# Τεχνικές Εξόρυξης Δεδομένων στον Τομέα Ηλεκτρονικού Λιανικού Εμπορίου με Σκοπό την Ανάπτυξη Στρατηγικών Marketing.

**Επιβλέπων :** Γεώργιος Λεκάκος – Αναπληρωτής Καθηγητής

**Κύριος συντελεστής :** Ανδρέας Ζάρας – Data Scientist

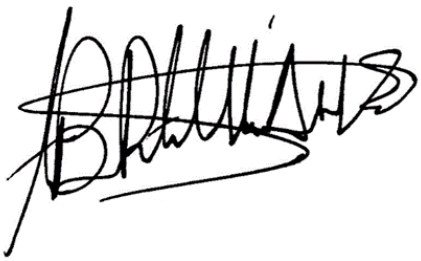
Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Ιούλιος 2022

**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)*



.....

**<ΑΒΡΑΜΙΔΗΣ ΠΑΥΛΟΣ>**

Φοιτητής MSc στη Διοικητική Επιστήμη και Τεχνολογία

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Ευχαριστίες

Με την εκπόνηση της παρούσας διπλωματικής εργασίας ολοκληρώνεται ο κύκλος των σπουδών μου στο Μεταπτυχιακό Πρόγραμμα της Διοικητικής Επιστήμης και Τεχνολογίας του Οικονομικού Πανεπιστημίου Αθηνών.

Ευχαριστώ θερμά τους Επιβλέποντες Καθηγητές μου, κύριο Ανδρέα Ζάρα και κύριο Γεώργιο Λεκάκο για την καθοδήγηση και τις πολύτιμες συμβουλές που μου έδωσαν από την αρχή ως το τέλος της προσπάθειάς μου.

Ευχαριστώ όλους τους καθηγητές μου για τις γνώσεις που μου παρείχαν, καθώς επίσης και την γραμματεία του προγράμματος κυρία Σοφία Γκίκα που ήταν πάντα διαθέσιμη να βοηθήσει.

Οφείλω να ευχαριστήσω όλους τους φίλους μου για την συμπαράσταση και την υπομονή που έδειξαν κατά τη διάρκεια των σπουδών μου.

Τέλος, ένα τεράστιο ευχαριστώ αξίζει στην οικογένειά μου, για την αμέριστη συμπαράσταση και υποστήριξη που πάντα μου προσφέρουν απλόχερα, καθώς είναι αυτοί που με οδήγησαν στην επίτευξη αυτού του πολύ σημαντικού στόχου της ζωής μου. Σε αυτούς οφείλω αυτό που είμαι και τους αφιερώνω την παρούσα διπλωματική εργασία.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Περίληψη

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η τμηματοποίηση των πελατών ενός ηλεκτρονικού καταστήματος σε ομάδες, η εύρεση συσχετίσεων ανάμεσα σε κατηγορίες προϊόντων που διαθέτει το κατάστημα αλλά και οι παρουσίαση ορισμένων προτάσεων marketing που θα βοηθήσουν την επιχείρηση να αυξήσει τα κέρδη της και να δημιουργήσει πελάτες με αφοσίωση. Το ηλεκτρονικό κατάστημα δραστηριοποιείται στο χώρο της ένδυσης υπόδησης και αξεσουάρ με την έδρα της επιχείρησης να βρίσκεται στην πόλη της Λάρισας.

Για την ανάλυση Clustering και την Market Basket Analysis χρησιμοποιήσαμε το λογισμικό SAS®, και πιο συγκεκριμένα το SAS Enterprise Guide 8.3 και το SAS Enterprise Miner Workstation 15.2. Το Guide χρησιμοποιήθηκε για την επεξεργασία, την μορφοποίηση και την διαδικασία προετοιμασίας των δεδομένων. Ενώ το Miner για την ανάλυση Clustering και την MBA. Ο αλγόριθμος που χρησιμοποιεί το λογισμικό για την κατηγοριοποίηση των πελατών σε clusters είναι ο K-means και για την διεξαγωγή των κανόνων συσχέτισης ο Apriori.

Αρχικά για την τμηματοποίηση των πελατών λάβαμε υπόψη την συχνότητα των αγορών, την χρονική περίοδο που πέρασε από την τελευταία αγορά και το χρηματικό ποσό που έχει ξοδέψει ο κάθε πελάτης συνολικά. Χρειάστηκε να εξαιρέσουμε ορισμένες ακραίες τιμές και να κάνουμε αρκετές δοκιμές με τα δεδομένα για να καταλήξουμε σε ένα επιθυμητό αποτέλεσμα. Έτσι η λύση που λάβαμε ήταν σχηματισμός 5 clusters πελατών, που εξυπηρέτησαν πλήρως τους σκοπούς την παρούσας διπλωματικής.

Στην συνέχεια για την εύρεση τον κανόνων συσχέτισης χρειάστηκε τα αφαιρέσουμε ένα μεγάλο μέρος τον δεδομένων μας, αφού χρειαζόμασταν αγορές με 2 ή περισσότερων προϊόντων και η πλειοψηφία των αγορών αφορούσε ένα μόνο προϊόν. Επίσης, χρησιμοποιήθηκαν οι κατηγορίες των προϊόντων καθώς, τα προϊόντα ήταν πολλά και με χαμηλή αγοραστική συχνότητα το καθένα και καταλήξαμε σε 43 κανόνες. Τέλος συνδυάζοντας τους κανόνες που βρήκαμε και τα clusters των πελατών που δημιουργήθηκαν, κάναμε ορισμένες προτάσεις marketing προς την επιχείρηση.

Τα αποτελέσματα της ανάλυσης βοηθούν την επιχείρηση στην καλύτερη κατανόηση του υπάρχοντος πελατολογίου της, στην διατήρησή του και ως αποτέλεσμα, στην βελτίωση ικανοποίησης αυξάνοντας έτσι τα κέρδη της. Η επιχείρηση θα μπορεί να εφαρμόσει αποτελεσματικά προωθητικές ενέργειες για τους πελάτες του ηλεκτρονικού της καταστήματος και να πετύχει έτσι την αύξηση του πελατολογίου της με την προσέλκυση νέων πελατών.

Μελλοντικά η συγκεκριμένη ανάλυση θα μπορούσε να πραγματοποιηθεί ξανά για ένα μεγαλύτερο χρονικό διάστημα με περισσότερα δημογραφικά και συμπεριφορά δεδομένα για τον κάθε πελάτη από της επιχείρησης. Έτσι, θα μπορούσε να γίνει σύγκριση των αποτελεσμάτων με αυτά της υπάρχουσας διπλωματικής για να διαπιστωθεί αν η επιχείρηση είχε κάποια βελτίωση στους στόχους που έχει θέσει και αν χρειάζεται να κάνει κάποιες αλλαγές στις προωθητικές της ενέργειες.

**Λέξεις-κλειδιά:** Εξόρυξη δεδομένων, Ανάλυση δεδομένων, Τμηματοποίηση πελατών, Clustering, Διαχείριση, RFM, MBA, Κανόνες συσχέτισης, Ηλεκτρονικό κατάστημα, Λιανικής, Προτάσεις marketing


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# Abstract

The purpose of this thesis is to segment the customers of an online store into groups, to find correlations between product categories that the store has and to present some marketing proposals that will help the company to increase its profits and create customers with loyalty. The online store is active in the field of clothing, footwear and accessories with the headquarters of the company located in the city of Larissa.

For Clustering and Market Basket Analysis we used SAS® software, more specifically SAS Enterprise Guide 8.3 and SAS Enterprise Miner Workstation 15.2. Guide was used for the data processing, formatting and preparation process. While Miner for Clustering analysis and MBA. The algorithm used by the software to categorize the customers into clusters is K-means and to carry out the correlation rules Apriori algorithm.

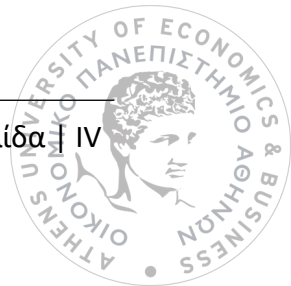
Initially, for customer segmentation we considered the frequency of purchases, the time that has passed since the last purchase and the amount of money that each customer has spent in total. We had to exclude some extreme values and do several tests with the data to arrive at a desired result. So, the solution we got was the formation of 5 customer clusters, which fully served the purposes of this thesis.

Then, to find the correlation rules, we had to remove a large part of our data since we needed purchases with 2 or more products and most purchases concerned a single product. The categories of the products were also used as there were many products and each with a low purchase frequency and so, we ended up with 43 rules. Finally, by combining the rules we found and the customer clusters I created, we made some marketing proposals for the company.

The results of the analysis help the company to better understand its existing clientele, to maintain it and as a result to improve satisfaction thus increasing its profits. The company will be able to effectively implement promotional actions for its customers of its online store and, thus, achieve an increase in its clientele by attracting new customers.

In the future this analysis could be performed again over a longer period with more demographic and behavioral data for each customer from the business. Thus, the results could be compared with those of the existing thesis to see if the company had any improvement in the goals, it has set and if it needs to make some changes in its promotional actions.

Keywords: Data Mining, Data Analysis, Customer Segmentation, Clustering, Management, RFM, MBA, Association Rules, Online Store, Retail, Marketing Proposals,


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# *Περιεχόμενα*

**1 Εισαγωγή ..... 1**
1.1 Το Ηλεκτρονικό Εμπόριο ..... 3
1.1.1 Η Ιστορία του Ηλεκτρονικού Εμπορίου ..... 3
1.1.2 Πόσο Αξίζει το Ηλεκτρονικό Εμπόριο ..... 4
1.2 Το Ευρωπαϊκό Ηλεκτρονικό Εμπόριο ..... 4
1.2.1 Επιχειρηματικά Μοντέλα ..... 5
1.3 Αντικείμενο Διπλωματικής ..... 7
1.3.1 Συνεισφορά ..... 8
1.4 Οργάνωση Κειμένου ..... 8
**2 Θεωρητικό Υπόβαθρο ..... 9**
2.1 Ανάλυση Συστάδων (Clustering) ..... 9
2.1.1 K-means Clustering ..... 13
2.1.2 Ανάλυση RFM ..... 14
2.1.3 DBSCAN ..... 16
2.1.4 Gauss ..... 18
2.2 Ανάλυση Καλαθιού Αγοράς (MBA) ..... 22
2.2.1 Τμηματοποίηση της Αγοράς ..... 25
2.2.2 Association Rules ..... 26
2.2.3 Apriori Algorithm ..... 27
**3 Σχετικές Εργασίες ..... 29**
3.1 Clustering - RFM ..... 29
3.2 Ανάλυση Καλαθιού Αγοράς (MBA) ..... 30
**4 Γνωριμία με τα δεδομένα ..... 36**
4.1 Η Επιχείρηση και τα Δεδομένα ..... 36
4.2 Προ-επεξεργασία Δεδομένων ..... 36
4.3 Κατανόηση Δεδομένων ..... 37
**5 Εξόρυξη Δεδομένων ..... 40**
5.1 Clustering ..... 40
5.1.1 RFM analysis ..... 40
5.2 Market Basket Analysis ..... 46
**6 Σχολιασμός αποτελεσμάτων ..... 49**
6.1 Customers Profile ..... 49
6.2 Προτιμήσεις στις Κατηγορίες Προϊόντων ..... 50



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

6.3 Προτάσεις Marketing ..... 51
**7 Αξιολόγηση ..... 58**
7.1 Αξιολόγηση των Clusters ..... 58
7.2 Αξιολόγηση των Association Rules ..... 60
**8 Τεχνικές Λεπτομέρειες ..... 61**
8.1 Clustering ..... 61
8.2 Market Basket Analysis ..... 66
***9 Επίλογος*..... 68**
9.1 Σύνοψη ..... 68
9.2 Συμπεράσματα ..... 69
9.3 Περιορισμοί ..... 69
9.4 Επεκτάσεις ..... 70
***10 ΒΙΒΛΙΟΓΡΑΦΙΑ* ..... 71**
**11 Παράρτημα I: Λίστα Πινάκων ..... 77**
**12 Παράρτημα II: Λίστα Γραφημάτων ..... 78**
**13 Παράρτημα III: Λίστα Εικόνων ..... 79**
**14 Παράρτημα IV: Appendix ..... 80**

# 1

# Εισαγωγή

Σε ένα ανταγωνιστικό περιβάλλον, οι επιχειρήσεις στοχεύουν στην ικανοποίηση των πελατών τους, διατηρώντας παράλληλα κερδοφόρες σχέσεις μακροπρόθεσμα. Ο τρόπος διατήρησης των ήδη υπαρχόντων πελατών μέσω πελατοκεντρικών πρακτικών είναι ιδιαίτερα σημαντικός για τις επιχειρήσεις, εξαιτίας κυρίως του ότι το να αποκτήσουν νέους πελάτες κοστίζει περισσότερο (Bodapati, 2008). Ο τρόπος διαχείρισης τώρα των πελατών αποτελεί στρατηγική των επιχειρήσεων με στόχο να ικανοποιούν τους πελάτες τους μακροπρόθεσμα, προσεγγίζοντας τους πελατοκεντρικά. Η στρατηγική αυτή επικεντρώνεται στις σχέσεις που έχουν δημιουργηθεί με τους πελάτες και απαιτεί την ανάλυση των δραστηριοτήτων των πελατών για την κατανόηση των αναγκών και των συμπεριφορών τους (Winer, 2001). Με αυτή την προοπτική, μια επιχείρηση πρέπει να παρακολουθεί τις ενέργειες των πελατών της, να διερευνά ό,τι θέλει, να αναλύει τα πρότυπα αγοράς τους και να διερευνά τις ανάγκες, τις προτιμήσεις και τις συμπεριφορές των πελατών της (Τσίπτσης & Χωριανόπουλος, 2009). Οι επιχειρήσεις λοιπόν τείνουν να χρησιμοποιούν τις συγκεκριμένες πληροφορίες για τη δημιουργία μακροχρόνιων επικερδών σχέσεων με τους πελάτες τους, με την προσφορά βέλτιστων υπηρεσιών (Sota et al., 2018). Τα τελευταία χρόνια, το ηλεκτρονικό εμπόριο άλλαξε εντελώς την αλληλεπίδραση ανάμεσα στους πελάτες και τις επιχειρήσεις. Απόρροια είναι ότι οι επιχειρήσεις έτειναν την προσοχή τους στις διαδικτυακές αγορές (Aksoy, 2008). Οι επιχειρήσεις μέσα από την πληροφορική μπορούν να λαμβάνουν, αποθηκεύουν και να βρίσκουν πιο εύκολα και γρήγορα τις πληροφορίες από τους πελάτες, παρέχοντας την δυνατότητα να εφαρμόζονται στρατηγικές αγοράς που βασίζονται στα δεδομένα. Ταυτόχρονα ενδιαφέρον παρουσιάζει η ανάγκη να εξάγονται χρήσιμες πληροφορίες μέσω μεγάλων συστάδων δεδομένων, οδηγώντας στο να εισάγονται νέες τεχνικές εξόρυξης δεδομένων (Han et al., 2012). Η τελευταία αποτελεί μια ημιαυτόματη διαδικασία, με στόχο να ανακαλύπτονται ενδιαφερόντα μοτίβα και στατιστικά από σημαντικές δομές στα σύνολα δεδομένων (Vahaplar & İnceoğlu, 2001). Με τα παραπάνω χρησιμοποιούνται τεχνικές για να εξάγονται ουσιαστικές και χρήσιμες πληροφορίες σε μεγάλα και μη δομημένα σύνολα δεδομένων (Dalal, 2012· Ching & Pong, 2002), διευκολύνοντας να εκτιμώνται οι μελλοντικές τάσεις της αγοράς, οι οποίες ασκούν επιρροή στους υπεύθυνους λήψης αποφάσεων (Hudairy, 2004).



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Η περιγραφή εξόρυξης δεδομένων συμπεριλαμβάνεται στη διαδικασία ανακάλυψης γνώσεων (Bramer, 2016:2), η οποία είναι υπεύθυνη για την καθαριότητα, ενσωμάτωση, επιλογή και μετασχηματισμό των δεδομένων πριν εφαρμοστεί ο αλγόριθμος εξόρυξης δεδομένων, ενώ ακολουθείται από την αξιολόγηση και παρουσίαση των μοτίβων (Han et al., 2012). Η εξόρυξη δεδομένων επιτρέπει να ανακαλύπτονται άγνωστα πρότυπα και σχέσεις μέσα σε μεγάλα σύνολα δεδομένων (Linoff & Berry, 2011; Marakas, 2003) με την γενική απασχόληση διεπιστημονικών μελετών διαφόρων τομέων (Savaş et al., 2012), όπως τα οικονομικά, το ηλεκτρονικό εμπόριο, η ιατρική, οι επιχειρήσεις και η εκπαίδευση. Κατά συνέπεια, η εξόρυξη δεδομένων διασταυρώνει την πληροφορική, τη μηχανική μάθηση, τη διαχείριση βάσεων δεδομένων, τα εφαρμοσμένα μαθηματικά και τη στατιστική (Emel & Taşkın, 2005). Εφαρμόζοντας την εξόρυξη εκτιμώνται οι μελλοντικές τάσεις που θα έχουν επίδραση στους υπεύθυνους λήψης αποφάσεων (Hudairy, 2004). Η εξόρυξη αποκάλυψε κάποια πρότυπα, βοηθώντας στην λήψη αποφάσεων των επιχειρηματιών ως προς το μάρκετινγκ, καθορίζοντας νέες ευκαιρίες προϊόντων και προβλέποντας τις συμπεριφορές των πελατών (Strauss & Frost, 2009). Επίσης, η τμηματοποίηση, ταξινόμηση και πρόβλεψη μπορούν να εφαρμοστούν σε διάφορα ζητήματα της επιχείρησης, ως προς τη μόχλευση δεδομένων, εφαρμόζοντας στρατηγικές που έχουν ως στόχο τους πελάτες. Ανάμεσα σε διάφορες επιχειρήσεις, το λιανικό εμπόριο αποτελεί έναν από τους κυριότερους κλάδους εφαρμογής εξόρυξης δεδομένων.

Οι Anderson et al. (2007), υποστήριξαν ότι το λιανικό εμπόριο που στοχεύει στην κάλυψη των αναγκών των πελατών, στρέφεται προς τις αποφάσεις ανταλλαγής δεδομένων. Η εξόρυξη δεδομένων προσαρμόζεται λοιπόν για ζητήματα του λιανικού εμπορίου, συμπεριλαμβάνοντας τη διατήρησης πελατών, τις διασταυρούμενες πωλήσεις, τη διαχείριση κινδύνου, την ανάλυση καλαθιού αγοράς, την απόκτηση πελατών και την ανίχνευση απάτης (Bala, 2008), τη τοποθέτηση ραφιών και τη διαχείριση αποθεμάτων (Hormozi & Giles, 2004). Τα δεδομένα που υπάρχουν στο καλάθι αποτελούν χρήσιμα στοιχεία για την εξόρυξη δεδομένων στο λιανικό εμπόριο και συχνά για την ανάλυση της τμηματοποίησης των πελατών και την εξαγωγή προτύπων αγοράς (Griva et al., 2018). Συγκεκριμένα, η ανάλυση καλαθιού αγοράς έχει συχνά αντιμετωπιστεί ως πρόβλημα εξόρυξης δεδομένων με στόχο την ανακάλυψη σχέσεων ή κανόνων σύνδεσης, οι οποίοι αντιπροσωπεύουν υποθέσεις σχετικά με τις προθέσεις αγοράς.

Όπως αναφέρουν οι Dippold & Hruschka, (2013a), η προηγούμενη έρευνα επικεντρώνεται συχνά σε αποφάσεις αγοράς μεταξύ των στοιχείων καλαθιού, όπου τα περισσότερα διερευνητικά μοντέλα αποκλείουν τα δημογραφικά στοιχεία και τα στοιχεία του μίγματος μάρκετινγκ. Παρόλα αυτά, υποστηρίζεται ότι μπορεί αν εφαρμόζεται η παραπάνω τεχνική σε διάφορα χαρακτηριστικά του προϊόντος στο λιανικό εμπόριο, όπως το εμπορικό σήμα (Kabasakal, 2020). Σε μελέτη των Kooti et al. (2016) αναλύθηκαν οι διαφορές στη λήψη απόφασης αγοράς πελατών, τονίζοντας ότι εάν καταμετρούνται εκτεταμένα τα χαρακτηριστικά, όπως η τοποθεσία και τα δημογραφικά χαρακτηριστικά, βοηθάει στο να προβλέπονται οι αποφάσεις στις αγορές μέσω διαδικτύου.

Επιπροσθέτως, σε μελέτη των Zhang και Pennacchiotti (2013) παρατηρήθηκε ότι τα δεδομένα από τα κοινωνικά μέσα των πελατών, το φύλο και η ηλικία, χρησιμεύουν, έτσι ώστε να προβλέπονται οι αποφάσεις αγοράς στο ηλεκτρονικό εμπόριο. Εξαιτίας του ότι επηρεάζονται οι συμπεριφορές αγορών, οι πελάτες και τα δημογραφικά τους χαρακτηριστικά χρησιμοποιούνται και για προτάσεις προϊόντων. Έχοντας αυτό ως κίνητρο, η συγκεκριμένη εργασία έχει ως στόχο να αναλύσει δεδομένα καλαθιού αγοράς και συστάδων μαζί. Συγκεκριμένα αποτελείται από μια περιπτωσιολογική μελέτη, στην οποία παρουσιάζονται κανόνες συσχέτισης, βοηθώντας στο να συσχετιστούν και οι προθέσεις αγοράς. Στις ακόλουθες ενότητες, παρέχεται εισαγωγή ανάλυσης καλαθιού αγοράς. Έπειτα, γίνεται η παρουσίαση της μελέτης περίπτωσης. Τα αποτελέσματα περιλαμβάνουν κανόνες που αρχικά είχαν επιλεγεί ανάλογα το ενδιαφέρον και στη συνέχεια έγινε η κατηγοριοποίηση τους.

Ο σκοπός της συγκεκριμένης εργασίας είναι να βοηθήσει ένα κατάστημα λιανικής πώλησης, με έδρα την πόλη της Λάρισας, να μπορέσει να γνωρίσει καλύτερα τις ανάγκες των πελατών του και την αγοραστική τους συμπεριφορά με αποτέλεσμα να μπορέσει να αναπτύξει στρατηγικές μάρκετινγκ, ώστε να αποκτήσει προβάδισμα έναντι του ανταγωνισμού, μέσα από την μελέτη της ανάλυσης συστάδων και του καλαθιού αγοράς στο ηλεκτρονικό εμπόριο.

# 1.1 Το Ηλεκτρονικό Εμπόριο

## 1.1.1 Η Ιστορία του Ηλεκτρονικού Εμπορίου

Η ιστορία του ηλεκτρονικού εμπορίου αναφέρεται συνήθως σε εμπορικά προϊόντα και υπηρεσίες μέσω οποιωνδήποτε ηλεκτρονικών μέσων όπως το Διαδίκτυο. Η έννοια "ηλεκτρονικό εμπόριο" εισήχθη για πρώτη φορά στα τέλη της δεκαετίας του 1970. Η διάδοση του βέβαια τη δεκαετία του 1990, οφείλεται στην εφεύρεση του παγκόσμιου ιστού από τον Tim Berners Lee. Κατά συνέπεια, το Διαδίκτυο δεν ανήκε πλέον στα στενά όριά της επιστήμης, αλλά σε ένα γενικότερο και παγκόσμιο επικοινωνιακό σύστημα. Από εκείνη τη στιγμή, το Διαδίκτυο μεταφέρθηκε στο εμπόριο και οι επιχειρήσεις πραγματοποίησαν την είσοδο τους σε αυτό. Αρχικά, το ηλεκτρονικό εμπόριο αφορούσε μόνο την ηλεκτρονική μεταφορά και την ανταλλαγή επιχειρηματικών πληροφοριών. Μέχρι το τέλος του 1994, αναπτύχθηκαν πρωτόκολλα ασφαλείας και η Pizza Hut είχε την πρώτη καταγεγραμμένη πώληση στο Διαδίκτυο. Στην Αμερική και την Ευρώπη από τις αρχές του 21$^{ου}$ αιώνα άρχισαν να προσφέρονται μέσω διαδικτύου προϊόντα από τις επιχειρήσεις. Με αυτό τον τρόπο, οι καταναλωτές άρχισαν να βιώνουν την εμπειρία του ηλεκτρονικού εμπορίου, έχοντας πλέον την δυνατότητα να αγοράζουν διάφορα προϊόντα από το διαδίκτυο, με τη χρήση υπηρεσιών ηλεκτρονικών συναλλαγών με ασφάλεια.

### 1.1.2 Πόσο Αξίζει το Ηλεκτρονικό Εμπόριο

Το διαδίκτυο μετέβαλλε σε σημαντικό βαθμό τον τρόπο με τον οποίο δραστηριοποιούνται οι επιχειρήσεις, δημιουργώντας μια ειδική επιχείρηση. Το ηλεκτρονικό εμπόριο άκμασε τα τελευταία χρόνια. Οι μεγαλύτερες επιχειρήσεις ανά τον κόσμο ξεκίνησαν να υπολογίζουν το διαδίκτυο ως βασικό μέρος της στρατηγικής μάρκετινγκ. Η Microsoft για παράδειγμα έχει να αποδώσει πολλές εμπειρίες από τις νέες καθαρές στρατηγικές (Turner, 2002). Ταυτόχρονα, φαίνεται ξεκάθαρα η ανάδυση βιώσιμων επιχειρηματικών μοντέλων διαδικτύου. Στα συγκεκριμένα επιχειρηματικά μοντέλα φαίνεται η ευελιξία αφού προχωρούν σε προσαρμογή των ιστοσελίδων τους για την υποστήριξη, ενίσχυση και βελτίωση όλων των τμημάτων της επιχείρησης.

Στα πλαίσια του ηλεκτρονικού εμπορίου, είναι και ο κίνδυνος που μπορεί να αντιμετωπίσει μια επιχείρηση που δραστηριοποιείται μέσω διαδικτύου. Διάφοροι οικονομολόγοι υποστηρίζουν ότι οι νεοσύστατες επιχειρήσεις μειώνονται ολοένα και περισσότερο εγείροντας ανησυχίες (Pepitone, 2011). Η φούσκα της dot-com το 1995-2000 αποτελεί ένα τέτοιο παράδειγμα. Οι υπέρογκες επενδυτικές αξίες για τις διαδικτυακές επιχειρήσεις θα μπορούσαν να προμηνύουν μια φούσκα διάφορους κινδύνους. Βέβαια υφίσταται το ότι οι σημερινές επιχειρήσεις υψηλής τεχνολογίας κατορθώνουν και διατηρούν συνεχή ανάπτυξη.

Με την κατάρρευση του ηλεκτρονικού εμπορίου, μια νέα εταιρεία δύσκολα μπορεί να μη κλείσει. Για μια αποτελεσματική χρήση του ηλεκτρονικού εμπορίου δεν πρέπει να επηρεάζονται από το μεγάλο περιβάλλον, αλλά να επικεντρωθούν στην υπεροχή της επιχειρηματικής λειτουργίας με βάση το Διαδίκτυο. Το ηλεκτρονικό εμπόριο δεν σχετίζεται με την πλήρη ψηφιοποίηση του λειτουργικού περιβάλλοντος των επιχειρήσεων. Πρόκειται περισσότερο για τη χρήση της ιδιαιτερότητας του Διαδικτύου για τη βελτίωση της δημιουργίας κερδών μέσω της ενίσχυσης της ανταγωνιστικότητας των επιχειρήσεων και της ανάπτυξης μιας ρεαλιστικής στάσης απέναντι στην επιχείρηση.

## 1.2 Το Ευρωπαϊκό Ηλεκτρονικό Εμπόριο

Η πρόσβαση στο διαδίκτυο ή η ανεπτυγμένη τεχνολογία επηρεάζει την ανάπτυξη του ηλεκτρονικού εμπορίου σε διάφορους τομείς λόγω διάφορων λόγων. Η τρέχουσα θέση των ευρωπαϊκών εξελίξεων ηλεκτρονικού εμπορίου ηλεκτρονικού εμπορίου ποικίλλει από κράτος σε κράτος, ακόμη και εντός της Ευρωπαϊκής Ένωσης. Τα βασικά κράτη με ενδιαφέρον ως προς το ηλεκτρονικό εμπόριο είναι η Αγγλία, Γαλλία, η Σουηδία, η Φινλανδία, η Νορβηγία και η Γερμανία. Τα κράτη αυτά έχουν επίσης υψηλό ποσοστό αγορών μέσω ηλεκτρονικού εμπορίου. Παρά την οικονομική κρίση, το 2011 το ηλεκτρονικό εμπόριο αναπτύχθηκε σημαντικά. Σύμφωνα με έρευνα της CRR που ανατέθηκε από την Kelkoo, η 9 συνολική ευρωπαϊκή αγορά ηλεκτρονικού εμπορίου άξιζε 200,52 δισεκατομμύρια ευρώ το 2011 (από 169,63 δισεκατομμύρια ευρώ πέρυσι, και οι διαδικτυακές πωλήσεις στην Ευρώπη ήταν 117,84 δισεκατομμύρια ευρώ (Kelkoo, 2012). Οι πωλητές μέσω διαδικτύου στα τρία ευρωπαϊκά κράτη



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

αντιπροσώπευαν το 71% των συνολικών διαδικτυακών αγορών στην Ευρώπη, με 48 δις ευρώ, 39,2 δις ευρώ και 25 δις ευρώ αντίστοιχα. Το πιο χαμηλό σύνολο διαδικτυακών πωλήσεων ανήκε στην Πολωνία με μόλις 3,4 δις ευρώ.

## 1.2.1 Επιχειρηματικά Μοντέλα

Για να μελετηθούν τα επιχειρηματικά μοντέλα του ηλεκτρονικού εμπορίου χρειάζεται να μελετηθεί ο ορισμός πρώτα των επιχειρηματικών μοντέλων γενικά. «*Η πολιτική, οι λειτουργίες, η τεχνολογία και η ιδεολογία της εταιρείας καθορίζουν το επιχειρηματικό της μοντέλο*» (Weill, 2005). Για αύξηση της αξίας μια καινούργια επιχείρηση μέσα από έναν νεωτερισμό, πρέπει να διαθέτει το κατάλληλο επιχειρηματικό μοντέλο. Αυτά τα μοντέλα τροποποιούν τη καινούργια τεχνολογία σε χρηματική αξία. Οι Chesbrough και Rosenbloom παρουσίασαν ένα βασικό πλαίσιο περιγραφής των στοιχείων ενός επιχειρηματικού μοντέλου. Το Διαδίκτυο έχει οδηγήσει σε διάφορα νέα είδη επιχειρηματικών μοντέλων. Υφίσταται μεγάλο μπέρδεμα ως προς τα επιχειρηματικά μοντέλα στο ηλεκτρονικό εμπόριο. Για την ακρίβεια, τα μοντέλα αυτά αποτελούν τις βασικές μεθόδους που εξηγούν πώς μια επιχείρηση ιστού πλουτίζει μέσω διαδικτύου. Ο συγγραφέας συνόψισε τις πιο κοινές επιχειρήσεις ηλεκτρονικού εμπορίου σε έξι βασικά επιχειρηματικά μοντέλα ηλεκτρονικού εμπορίου

- Μοντέλο διαμεσολάβησης

Το μοντέλο Brokerage είναι ένας ιστότοπος που φέρνει τους αγοραστές και τους πωλητές μαζί για να διεξάγουν επιχειρήσεις. Τα μοντέλα "Από επιχείρηση σε πελάτη" (B2C), "Από επιχείρηση σε επιχείρηση" (B2B) και "Πελάτης προς πελάτη" (C2C) συχνά το αντιπροσωπεύουν. Τα συγκεκριμένα 3 διαμεσολαβητικά μοντέλα ηλεκτρονικού εμπορίου, δηλαδή B2B, B2C, C2C μοιράστηκαν πάνω από το 90 % των συνολικών εσόδων ηλεκτρονικού εμπορίου το 2011 (Oracle, 2012). Το B2C αποτελεί τον πιο κοινό επιχειρηματικό τομέα. Για την ακρίβεια περιγράφει μια επιχείρηση με προϊόντα και υπηρεσίες απευθείας μέσω Διαδικτύου. Στις μέρες μας, οι πιο πολλές επιχειρήσεις ξεκίνησαν την εξυπηρέτηση της αγοράς B2C, ψάχνοντας καινούργιους πελάτες. Από την άλλη το B2B αναφέρεται σε δύο ή περισσότερες επιχειρήσεις με ηλεκτρονικές επιχειρηματικές σχέσεις, αποτελώντας τη μεγαλύτερη μορφή ηλεκτρονικού εμπορίου. Σε αυτήν περιλαμβάνονται εμπορικές συναλλαγές ανάμεσα σε διαφορετικές επιχειρήσεις, όπως ανάμεσα σε ένα εργοστάσιο και ένα κατάστημα, ή σε έναν κατάστημα χονδρικής πώλησης και έναν μεταπωλητή. Έτσι, τόσο οι καταναλωτές όσο και οι πωλητές αποτελούν επιχειρηματικά όντα και δεν αφορούν μεμονωμένους καταναλωτές. Δηλαδή μοιάζει σαν ένας παραγωγός να προμηθεύει τα προϊόντα του στον πωλητή. Τέλος το τελευταίο διάστημα, το τμήμα B2B αναπτύχθηκε ραγδαία, διότι όλο και πιο πολλοί επιχειρηματίες μετέφεραν την επιχείρησή τους στο Διαδίκτυο (Dvorak,2006). Εξαιτίας του μεγάλου πλήθους παραγγελιών στη συγκεκριμένη μορφή, δημιουργήθηκε υψηλότερη συναλλαγή χρηματοδότησης στο ηλεκτρονικό εμπόριο. Η C2C είναι οι συναλλαγές αγαθών ή υπηρεσιών που πραγματοποιούνται διαδικτυακά μεταξύ πελατών μέσω τρίτου. Για παράδειγμα, ένας πελάτης δημοσιεύει ένα προϊόν στο Amazon για δημοπρασία και άλλοι πελάτες

υποβάλλουν προσφορές για να το αγοράσουν. Ο ιστότοπος (στην περίπτωση αυτή το "Amazon") είναι απλά ένας μεσάζων, ο οποίος παρέχει μια πλατφόρμα για να επιτρέψει στους πελάτες να συναντηθούν μεταξύ τους.

- Μοντέλο εμπόρου

Το μοντέλο εμπόρου συνήθως χωρίζεται σε εικονικό έμπορο και έμπορο κλικ και κονιάματα. Το προηγούμενο τμήμα αναφέρεται στις εταιρείες που λειτουργούν αποκλειστικά μέσω διαδικτυακών επιχειρηματικών πρακτικών (Wang, 2007). Ενώ το τελευταίο σημαίνει μια εταιρεία που λειτουργεί τόσο με κατάστημα εκτός σύνδεσης όσο και με μια βιτρίνα ιστού. Μπορεί να είναι ένα τοπικό κατάστημα που κινείται στο διαδίκτυο ή το αντίστροφο. Περισσότερες εταιρείες δημιουργίας μετακινούν την επιχείρησή τους στο διαδίκτυο προσπαθώντας να επιτύχουν την πλειοψηφία της επιχείρησης 19 στο διαδίκτυο. Με αυτή την ενέργεια, βοηθά την εταιρεία να μειώσει το λειτουργικό κόστος και να μειώσει ταυτόχρονα τον επενδυτικό κίνδυνο.

- Μοντέλο διαφήμισης

Το μοντέλο διαφήμισης ηλεκτρονικού εμπορίου είναι η ψηφιακή επέκταση της παραδοσιακής επιχείρησης ραδιοτηλεοπτικών μεταδόσεων μέσων ενημέρωσης (Wang, 2007). Για τη συλλογή εσόδων, α) οι ιστοσελίδες μπορούν να παρέχουν χώρο για διαφημιστικό περιεχόμενο, όπως ημερήσιες εφημερίδες και περιοδικά. ή β) να πληρώνεται με κάθε κλικ μέσω του συνδέσμου στο περιεχόμενο. Για παράδειγμα, οι διαφημίσεις Google, η Google πληρώνεται με το ποσοστό κλικ μέσω.

- Μοντέλο θυγατρικών

Το online μοντέλο θυγατρικών βασίζεται στις πωλήσεις προμήθειας. Οι συνεργαζόμενοι συνεργάτες / ιδιοκτήτες ιστότοπων δεν χρειάζεται να αγοράσουν το εμπόρευμα για να μετακυληθούν, να παράσχουν μόνο πληροφορίες προϊόντων, δεν εξελίχθηκαν υπηρεσίες logistics. Ο χειρισμός και η αποστολή παρέχονται από τη μητρική εταιρεία στην οποία συνδέεται. Οι συνεργαζόμενοι συνεργάτες ανακατευθύνουν τον πελάτη από τη δική τους ιστοσελίδα στον ιστότοπο των μητρικών εταιρειών και κερδίζουν προμήθεια για τα έσοδα. Είναι ένα επιχειρηματικό μοντέλο που πληρώνεται για την απόδοση. Εάν μια συνδεδεμένη εταιρεία δεν αποφέρει έσοδα, η μητρική εταιρεία δεν έχει καμία ζημία. Έτσι, είναι ένα ηλεκτρονικό επιχειρηματικό μοντέλο, το οποίο είναι επίσης το δημοφιλές επιχειρηματικό μοντέλο και στο ηλεκτρονικό εμπόριο. Για παράδειγμα, η Amazon είναι ο πρώτος έμπορος που χρησιμοποιεί αυτή τη μέθοδο πώλησης, επιτρέποντας σε οποιονδήποτε να πουλήσει και προσφέροντας οικονομικά κίνητρα. Όλο και περισσότερες εταιρείες δημιούργησαν αυτό το μοντέλο στην επιχείρηση αφού είδαν το αποτέλεσμα έναντι χαμηλού κινδύνου να εξελίσσεται.

- Μοντέλο πληροφόρησης

Το μοντέλο της επιχείρησης της πληροφόρησης βασίζεται σε μεγάλες βάσεις δεδομένων εξειδικευμένων πληροφοριών για ένα συγκεκριμένο θέμα. Στο ηλεκτρονικό εμπόριο, η πληροφόρηση είναι το πιο πολύτιμο πράγμα, ειδικά όταν οι πληροφορίες αναλύονται προσεκτικά, ακαδημαϊκά. Μια

εταιρεία ηλεκτρονικού εμπορίου επιθυμεί να γνωρίζει πληροφορίες σχετικά με τους πελάτες και τις καταναλωτικές τους συνήθειες. Αυτοί οι ιστότοποι προσελκύουν μεγάλο αριθμό χρηστών που ενδιαφέρονται για το συγκεκριμένο πεδίο πληροφοριών τους και δημιουργούν έσοδα μέσω διαφημίσεων ή πληροφοριών πώλησης.

- Κοινοτικό μοντέλο

Ο ιστός ταιριάζει με την κοινοτική πρότυπη επιχείρηση που λειτουργεί εγγενώς. Αποτελεί επίσης το πιο παραγωγικό επιχειρηματικό μοντέλο στο ηλεκτρονικό εμπόριο, με τη ραγδαία του ανάπτυξη το τελευταίο διάστημα, όπως στα κοινοτικά δίκτυα. Συνδέεται έντονα με την αφοσίωση των χρηστών. Στο συγκεκριμένο μοντέλο η πώληση εισπράττει τα έσοδα μέσα από υπηρεσίες και προϊόντα ή έχουν ως βάση τις διαφημίσεις και συνδρομές για υπηρεσίες υψηλής ποιότητας.

## 1.3 Αντικείμενο Διπλωματικής

Αντικείμενο της παρούσας διπλωματικής εργασίας αποτελεί η ανάλυση και μελέτη των δεδομένων ενός ηλεκτρονικού καταστήματος με είδη ένδυσης, υπόδησης και αξεσουάρ. Το ηλεκτρονικό κατάστημα διαθέτει μια μεγάλη ποικιλία προϊόντων με πολλές κατηγορίες και απευθύνεται σε άντρες και γυναίκες όλων των ηλικιών. Το κατάστημα εξυπηρετεί όλη την Ελλάδα με την έδρα της επιχείρησης να βρίσκεται στην πόλη της Λάρισας. Η επιχείρηση έχει φυσική παρουσία στην Λάρισα με 3 καταστήματα.

Σκοπός της διπλωματικής είναι μέσα από την ανάλυση των δεδομένων του ηλεκτρονικού καταστήματος να μπορέσουμε να προτείνουμε αποτελεσματικές στρατηγικές marketing στην επιχείρηση. Ορισμένα βασικά ερωτήματα που κάνει κάθε επιχείρηση στον εαυτό της είναι: ποιοι είναι οι πιστοί πελάτες, ποιοι πελάτες πρόκειται να αποχωρήσουν, ποιοι είναι αυτοί που ξοδεύουν τα περισσότερα χρήματα ή ποια προϊόντα αγοράζονται περισσότερο μαζί. Απαντώντας στα ερωτήματα αυτά θα έχει περισσότερες πληροφορίες για τους πελάτες της οπότε θα μπορεί να εξυπηρετήσει καλύτερα τους πελάτες της και να αυξήσει το κέρδος της.

Για να μπορέσουμε να απαντήσουμε σε αυτά τα ερωτήματα θα χρησιμοποιήσουμε κάποιες τεχνικές όπως η Ομαδοποίηση (Clustering) σε συνδυασμό με τις τιμές RFM (Recency, Frequency, Monetary) και τον αλγόριθμο K-means καθώς και την Ανάλυση Καλαθιού Αγοράς (Market Basket Analysis) με την βοήθεια των Κανόνων Συσχέτισης (Association Rules). Το εργαλείο που θα μας βοηθήσει στην υλοποίηση είναι το λογισμικό SAS όπου η γνωριμία με τα τα δεδομένα και η ανάλυση RFM θα πραγματοποιηθούν στο SAS Enterprise Guide ενώ το Clustering και η Market Basket Analysis θα πραγματοποιηθούν στο SAS Enterprise Miner Workstation.

### 1.3.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

➔ Αντιμετωπίσαμε ορισμένες δυσκολίες στην υλοποίηση της ανάλυσης clustering και MBA καθώς είχαμε στην κατοχή μας ένα σύνολο δεδομένων με χαμηλή συχνότητα συναλλαγών ανά πελάτη, χαμηλή συχνότητα αγοράς του κάθε μοναδικού προϊόντος στο σύνολο των συναλλαγών και με χαμηλό πλήθος προϊόντων ανά συναλλαγή.

➔ Πραγματοποιήθηκε ανάλυση καλαθιού αγοράς για τις κατηγορίες των προϊόντων.

➔ Ενσωματώσαμε τον κώδικα αξιολόγησης της MBA στο SAS Enterprise Miner και αξιολογήσαμε την επίδοση των αποτελεσμάτων της MBA βρίσκοντας έτσι τους καταλληλότερους κανόνες συσχέτισης.

➔ Εισηγηθήκαμε προτάσεις μάρκετινγκ για το ηλεκτρονικό κατάστημα σύμφωνα με το προφίλ των καταναλωτών.

## 1.4 Οργάνωση Κειμένου

Στο κεφάλαιο 1 παρουσιάζεται η Εισαγωγή της παρούσας διπλωματικής αλλά και το αντικείμενο της διπλωματικής. Στο Κεφάλαιο 2 παρουσιάζεται το Θεωρητικό Υπόβαθρο της τεχνική Clustering όπως επίσης και της Market Basket Analysis. Στο Κεφάλαιο 3 γίνεται αναφορά σε Σχετικές Εργασίες με το αντικείμενο της παρούσας διπλωματικής εργασίας. Στο Κεφάλαιο 4 γίνεται η Γνωριμία με τα Δεδομένα για να μπορέσουμε να κατανοήσουμε πλήρως τους πελάτες του καταστήματος. Στο Κεφάλαιο 5γίνεται η Εξόρυξη των Δεδομένων με την Ανάλυση Clustering και την Market Basket Analysis. Στο Κεφάλαιο 6 γίνεται ο Σχολιασμός των αποτελεσμάτων που προέκυψαν και προτείνονται ορισμένες προτάσεις marketingστην επιχείρηση βάση αυτών. Στο Κεφάλαιο 7 γίνεται η Αξιολόγηση των αποτελεσμάτων ενώ στο Κεφάλαιο 8 παρουσιάζονται οι Τεχνικές Λεπτομέρειες της ανάλυσης μέσα από το λογισμικό SAS. Στην συνέχεια στο Κεφάλαιο 9 παρουσιάζεται ο Επίλογος της διπλωματικής εργασίας συμπεριλαμβανομένου των Συμπερασμάτων, των Μελλοντικών Επεκτάσεων αλλά και τους Περιορισμούς που αντιμετωπίσαμε. Τέλος το Κεφάλαιο 10 περιέχει την Βιβλιογραφία που χρησιμοποιήθηκε. Μετά το 10ο Κεφάλαιο παρουσιάζεται ο κώδικας που χρησιμοποιήθηκε για την αξιολόγηση των κανόνων συσχέτισης όπως επίσης και τα βήματα που ακολουθήθηκαν για την εισαγωγή των δεδομένων στο λογισμικό SAS. Στα Παρατήματα I, II, III, IV περιέχονται οι λίστες Πινάκων, Γραφημάτων, Εικόνων και Appendix αντίστοιχα.

# 2

# Θεωρητικό Υπόβαθρο

## 2.1 Ανάλυση Συστάδων (Clustering)

Η ανάλυση συστάδων ή αλλιώς clustering αποτελεί μία από τις κυριότερες εργασίες εξόρυξης δεδομένων. Γενικά η ανάλυση συστάδων σχετίζεται με τον τρόπο με τον οποίον εντάσσονται οι οντότητες σε ομοειδείς ομάδες. Η συγκεκριμένη δραστηριότητα χρησιμοποιείται και στους ανθρώπους και μάλιστα ξεκινάει αυθόρμητα από όταν είναι παιδιά. Οι άνθρωποι σε πρωτόγονες συνθήκες αλλά και όσοι έχουν μία σχετική εμπειρία μπορούν να κατανοήσουν αυθόρμητα τις ομάδες (Kruskal, 1977). Στην περίπτωση της επιστημονικής ανάλυσης συστάδων οι ομάδες θα εξαχθούν με βάση αλγόριθμους από τα δεδομένα. Στόχος τους είναι να διαχωρίσουν ένα σύνολο από παραδείγματα σε υποσύνολα. Τα τελευταία καλούνται ως συστάδες. Στον διαχωρισμό βασικό ρόλο διαδραματίζει η ομοιότητα, δηλαδή η συστάδα θα πρέπει να μοιάζει και τα παραδείγματα της να μη μοιάζουν με άλλες συστάδες. Συγγενές αντικείμενο των συστάδων είναι η κατηγοριοποίηση, η οποία έχει ως στόχο να προβλέπει την κατηγορία κάθε παρατήρησης. Οι διαφορές όμως ανάμεσα σε αυτήν και την ανάλυση συστάδων είναι πολλές. Στην κατηγοριοποίηση οι κατηγορίες γνωρίζονται από πριν. Τα δεδομένα υφίσταται ένα γνώρισμα, το γνώρισμα της κλάσης στο οποίο γίνεται η καταγραφή της κατηγορίας για την εκάστοτε παρατήρηση. Οι αλγόριθμοι θα μοντελοποιήσουν τις σχέσεις ανάμεσα στο γνώρισμα της κλάσης και στα υπόλοιπα. Στην περίπτωση της κατηγοριοποίησης υπάρχει εκπαίδευση μέσα από παραδείγματα. Αυτό καθιστά σαφές ότι η γνώση σε σχέση με τις κατηγορίες υπάρχει από πριν και ότι καθοδηγεί τη διαδικασία εκπαίδευσης, χαρακτηρίζοντας την κατηγοριοποίηση σαν επιβλεπόμενη μάθηση. Στην ανάλυση συστάδων δεν υφίσταται κάποιο γνώρισμα στο οποίο υπάρχει καταγραφή της κλάσης παραδειγμάτων και οι συστάδες δεν γίνονται από πριν γνωστές. Αντίθετα αυτό που ζητάται είναι ο εντοπισμός των συστάδων και η ένταξη των παραδειγμάτων στην κατάλληλη συστάδα. Οι συστάδες μπορούν να συγκροτηθούν στη βάση της ομοιότητας των μελών τους. Για αυτό το λόγο θεωρούνται ως μία μορφή εκπαίδευσης μέσα από την παρατήρηση. Ταυτόχρονα αυτό το γεγονός της μη ύπαρξης από πριν γνώσεων χαρακτηρίζει την ανάλυση συστάδων μη επιβλεπόμενη μάθηση. Ο βασικός στόχος της κατηγοριοποίησης είναι να διατυπώνονται προβλέψεις, ενώ βασικός στόχος της ανάλυσης συστάδων είναι η περιγραφή σε διαδραστικό επίπεδο. Η ανάλυση συστάδων θα αντιμετωπίσει όλα τα γνωρίσματα ισότιμα χρησιμοποιώντας στον υπολογισμό της την ομοιότητα στις παρατηρήσεις. Αντίθετα η



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

κατηγοριοποίηση θα χρησιμοποιήσει τα υπόλοιπα γνωρίσματα ως πρόβλεψη των τιμών του γνωρίσματος της κλάσης.

Στα πλαίσια εξόρυξης δεδομένων η ανάλυση συστάδων χρησιμοποιείται με πολλαπλό τρόπο, σαν αυτόνομη αναλυτική εργασία δίνει τη δυνατότητα στους αναλυτές να διαχωρίσουν τα δεδομένα σε ομοειδείς ομάδες παρατήρησης. Εν συνεχεία οι αναλυτές μπορούν να επικεντρωθούν στην κάθε ομάδα αναγνωρίζοντας τα κοινά τους χαρακτηριστικά και να εξάγουν τη γνώση για να λάβουν αποφάσεις. Η γνωστότερη εφαρμογή της ανάλυσης συστάδων στις επιχειρηματικές διαδικασίες είναι η διαφήμιση και ειδικά η τμηματοποίηση της αγοράς. Η τμηματοποίηση της αγοράς περιγράφει τον διαχωρισμό των καταναλωτών σε ομάδες που έχουν παρόμοια καταναλωτική συμπεριφορά και έχει κεφαλαιώδη σημασία για το μάρκετινγκ. Οι μαζικές διαφημίσεις έχουν μεγάλο κόστος και μικρό ποσοστό ανταπόκρισης. Εντοπίζοντας ομάδες που είναι όμοιες ως προς την κατανάλωση μπορεί να γίνει ο σχεδιασμός της διαφημιστικής εκστρατείας προσαρμοσμένη στα ιδιαίτερα χαρακτηριστικά. Ο στόχος στις συγκεκριμένες ομάδες θα κοστίσει λιγότερο και θα πετύχει σημαντικά πιο μεγάλα ποσοστά ανταπόκρισης στους καταναλωτές. Πέρα από αυτόνομα εργαλείο ανάλυσης, η ανάλυση συστάδων συνδυάζεται και με άλλες εργασίες εξόρυξης δεδομένων, αποτελώντας το στάδιο για την προεπεξεργασία. Εξαιτίας της ικανότητας των αλγορίθμων της στην ομαδοποίηση σύμφωνα με την ομοιότητα μπορεί να γίνει η χρήση της για να εντοπίζονται παρατηρήσεις ακραίων τιμών (Ng & Han, 1994; Shekhar & Chawla, 2003). Οι ακραίες αυτές παρατηρήσεις απομακρύνουν το σύνολο δεδομένων, έτσι ώστε να υπάρχει ένα βέλτιστο σύνολο εκπαίδευσης. Τέλος οι συστάδες οι οποίες προκύπτουν θεωρούνται και κατηγορίες. Σε επόμενο στάδιο μπορεί να γίνει η κατηγοριοποίηση και να αναπτυχθούν μοντέλα ικανά για την πρόβλεψη κατηγορίας. Ο συνδυασμός της ανάλυσης συστάδων και της κατηγοριοποίησης μπορεί να προκαλέσει υβριδικούς κατηγοριοποιητές.

Η ανάλυση συστάδων εφαρμόζεται πολύ συχνά στις σύγχρονες επιχειρήσεις. Το γνωστότερο πεδίο εφαρμογής της είναι το μάρκετινγκ. Οι εκστρατείες διαφημίσεων που απευθύνονται στο σύνολο του πληθυσμού έχουν μεγάλο κόστος και μικρό ποσοστό συμμετοχής ανταπόκρισης στους καταναλωτές. Με αυτόν τον τόπο γίνεται φανερό ότι κάθε προσφορά δεν μπορεί να χρησιμοποιηθεί για κάθε πελάτη και κάθε καταναλωτής δεν μπορεί να ανταποκριθεί με τον ίδιο τρόπο στο ίδιο διαφημιστικό μήνυμα. Οι διαφημιστές επιθυμούν να ξεχωρίσουν από τον πληθυσμό ομάδες με παρόμοια χαρακτηριστικά. Οι επιχειρήσεις διατηρούν στα μηχανογραφικά τους συστήματα πληροφορίες για τους καταναλωτές τους. Αυτά τα στοιχεία μπορούν να αναλυθούν, ώστε να υπάρξει εξαγωγή συμπερασμάτων. Θα χρησιμοποιηθούν για το διαφημιστικό σκοπό. Ένας καινούργιος όρος που μπορεί να περιγράψει τη συγκεκριμένη διαδικασία είναι το «data base marketing». Σε μία επιχείρηση που δεν διαθέτει τέτοια στοιχεία μπορεί να διεξαχθεί μία έρευνα αντιπροσωπευτικού δείγματος του πληθυσμού. Οι πελάτες μπορούν να γίνουν ομάδες σύμφωνα με διάφορα κριτήρια όπως την περιοχή που ζουν, τον τρόπο ζωής τους, το φύλο ή την ηλικία τους, το καταναλωτικό τους προφίλ. Ο διαχωρισμός του καταναλωτικού κοινού σε ομάδες με παρόμοια χαρακτηριστικά αναφέρεται ως τμηματοποίηση της αγοράς.

Οι διαφημιστές γνωρίζοντας τα τμήματα της αγοράς οργανώνουν στοχευμένες διαφημιστικές εκστρατείες για την κάθε ομάδα καταναλωτών. Οι συγκεκριμένες εκστρατείας κοστίζουν λιγότερο και έχουν καλύτερη ανταπόκριση. Ο ορθός καθορισμός της ομάδας αυξάνει περισσότερο και την ανταπόκριση των καταναλωτών. Η τμηματοποίηση της αγοράς και η βαθύτερη γνώση του καταναλωτικού προφίλ των ήδη πελατών αλλά και των εν δυνάμει πελατών δεν αφορούν μόνο τη διαφήμιση. Η εξυπηρέτηση τους έπειτα από την πώληση μπορεί να αλλάξει προς το καλύτερο με την ομαδοποίηση των απόψεών τους, των παραπόνων τους, των αναφορών για προβλήματα και τα λοιπά. Το μάρκετινγκ αποτελεί το πιο γνωστό παράδειγμα εφαρμογής της ανάλυσης συστάδων. Δεν αποτελεί όμως και το μοναδικό. Οι επιχειρηματικές δράσεις μπορούν να ορίζουν πρόσωπα, αντικείμενα και ενέργειες σε σχέση με τα γνωρίσματα τους και να αποδίδουν κέρδος ανακαλύπτοντας συστάδες. Η ανάλυση συστάδων γίνεται και για τη διαχείριση ανθρώπινων πόρων. Οι εργαζόμενοι ομαδοποιούνται σύμφωνα με την αξιολόγηση τους,ντην εκπαίδευση που έχουν λάβει, τη συμμετοχή τους σε ομάδες εργασίας, τις ειδικές τους δεξιότητες και τα λοιπά. Συγκεκριμένα γνωρίσματα χρησιμοποιούνται για να καθορίσουν το μισθό, την προαγωγή ή τη στελέχωση ομάδων εργασίας. Η ανάλυση συστάδων πραγματοποιείται και στην παραγωγή. Το κάθε προϊόν ανήκει σε κάποια κατηγορία ή εντάσσεται. Με το πέρασμα του χρόνου παρουσιάζονται προβλήματα όπως οι παλιές κατηγορίες ή η ένταξη προϊόντων σε λάθος κατηγορίες και τα λοιπά. Με την ανάλυση συστάδων επανακαθορίζονται οι κατηγορίες με πιο ουσιαστικό τρόπο και επανεντάσσονται προϊόντα στην κατάλληλη κατηγορία. Ταυτόχρονα γνωρίζοντας τα ειδικά χαρακτηριστικά των τμημάτων του καταναλωτικού κοινού αξιοποιείται αυτή η γνώση και στην παραγωγή μέσω του σχεδίου εξειδίκευσης προϊόντων για συγκεκριμένες ομάδες και με τη δημιουργία γραμμών παραγωγής. Τέλος ολόκληρες γραμμές παραγωγής και επιχειρήσεις μπορούν να ομαδοποιούνται σύμφωνα με την ταχύτητα, την ποιότητα και το κόστος τους.

Η ανάλυση συστάδων χρησιμοποιείται επίσης για να ελέγχεται και να εντοπίζεται η απάτη. Σε περίπτωση απάτης υφίσταται όριο. Μάλιστα τότε εντοπίζονται οι μικρές και απομονωμένες συστάδες οι οποίες καθορίζονται ως ύποπτες. Παράδειγμα τέτοιας ανάλυσης είναι η εργασία από τους Thiprungsri and Vasarhelyi (2011). Οι ερευνητές μελέτησαν το πρόβλημα ομαδικών ασφαλίσεων ζωής. Σε αυτές τις ασφάλειες οι πελάτες ήταν μία εταιρεία η οποία ασφάλισε το προσωπικό της ως μεμονωμένος ιδιώτης. Οι ασφαλιστικές εταιρείες που έδιναν τέτοια ασφαλιστικά συμβόλαια δεν τηρούσαν πληροφορίες ασφαλείας. Εντοπίστηκαν λοιπόν περιπτώσεις απάτης και εφαρμόστηκε η μέθοδος K-means με αποτέλεσμα οι ερευνητές να εντοπίσουν ολιγομελείς συστάδες οι οποίες απαιτούσαν ειδικό έλεγχο. Η ανάλυση συστάδων βρίσκεται και στη διαχείριση επιχειρηματικών διαδικασιών. Στις σημερινές επιχειρήσεις οργάνωσης διαδικασιών οι επιχειρηματικές διαδικασίες μπορούν να τροποποιηθούν και να αυτοματοποιηθούν σε εξειδικευμένα πληροφοριακά συστήματα. Σύμφωνα με τους Jung, Bae and Liu (2009) οι επιχειρηματικές διαδικασίες μπορούν να μετασχηματίσουν διανυσματικά μοντέλα και να ομαδοποιηθούν με την εφαρμογή της οριστικής ανάλυσης συστάδων. Τα αποτελέσματα που θα προκύψουν θα χρησιμοποιηθούν για τον καθορισμό καινούργιων ή τον ανασχεδιασμό των ήδη

υπαρχουσών επιχειρηματικών διαδικασιών. Επιπροσθέτως η ανάλυση συστάδων μπορεί να χρησιμοποιηθεί και στο στρατηγικό management. Στην περίπτωση των διοικητικών στελεχών γίνεται ανάλυση των στοιχείων του επιχειρηματικού κλάδου με την ομαδοποίηση των επιχειρήσεων αναλόγως με τα ιδιαίτερα χαρακτηριστικά τους, τις αδυναμίες και τα πλεονεκτήματα. Με αυτό τον τρόπο γίνεται καλύτερη κατανόηση των συνθηκών ανταγωνισμού και μπορεί να γίνει ο σχεδιασμός δράσεων που θα εξασφαλίσουν ένα πλεονέκτημα στον ανταγωνισμό. Παράδειγμα ανάλυσης τέτοιων στοιχείων είναι η εργασία των Mosleh, Nosratabadi και Bahrami (2015), οι οποίοι ανέλυσαν επιχειρηματικά μοντέλα εφαρμόζοντας τα σε τουριστικά γραφεία. Με την εφαρμογή ιεραρχικής ανάλυσης συστάδων εντόπισαν ότι εφαρμόζονταν τρία διαφορετικά επιχειρηματικά μοντέλα με πιο δημοφιλές το μοντέλο που έχει ως βάση τα χρηματοοικονομικά και το αμέσως επόμενο το μοντέλο που έχει ως βάση τους πελάτες και το μοντέλο που έχει ως βάση τις υπηρεσίες.

Ενδιαφέρον προκάλεσαν και τα πεδία εφαρμογής στα οποία χρησιμοποιείται η ανάλυση συστάδων σε σχέση με την ανάλυση χωρικής πληροφορίας. Για παράδειγμα μπορεί να πραγματοποιηθεί ομαδοποίηση κατοικιών ανάλογα με τη γεωγραφική τους θέση, τον τύπο και την αξία τους, καθορίζοντας ειδικές υπηρεσίες για συγκεκριμένες ομάδες. Ταυτόχρονα μπορεί να γίνει ο σχεδιασμός θεματικών χαρτών, οι οποίοι θα αναγνωρίσουν περιοχές που έχουν παρόμοια χρήση γης όπως τα αστικά κέντρα, ύπαιθρος, οι βιομηχανικές περιοχές. Αυτά τα στοιχεία χρησιμοποιούνται για κρατικούς φορείς και για να ασκηθεί πολιτική και να δημιουργηθούν υποδομές. Ιδιαίτερα σημαντικό αντικείμενο είναι ο εντοπισμός συστάδων επιχειρήσεων. Μία συστάδα επιχείρησης αποτελούν το σύνολο από επιχειρήσεις οι οποίες βρίσκονται στην ίδια περιοχή και ενώνονται με βάση κοινά στοιχεία λειτουργώντας ανταγωνιστικά ή συμπληρωματικά. Η πιο γνωστή περίπτωση επιχειρηματικού συνόλου είναι η Silicon Valley. Η γεωγραφική συγκέντρωση παρόμοιων και σχετιζόμενων επιχειρήσεων εξασφαλίζει μία σειρά από κέρδη, όπως να μειώνονται τα έξοδα μεταφοράς, να αναπτύσσονται οι υλικές υποδομές που είναι χρήσιμες για το συγκεκριμένο κλάδο, επιτυγχάνεται οικονομία κλίμακας, αυξάνεται η έκθεση στον ανταγωνισμό, ενισχύοντας τη βελτίωση της ποιότητας, αξιοποιείται από κοινού το τοπικό εργατικό δυναμικό, το οποίο σταδιακά εξειδικεύεται, δημιουργούνται βέλτιστες πρακτικές καινούργιας τεχνολογίας και επιτυγχάνεται το ανταγωνιστικό πλεονέκτημα. Ο εντοπισμός των επιχειρήσεων είναι ιδιαίτερα χρήσιμος, καθώς οι αρμόδιοι φορείς αφού λάβουν τα ειδικά μέτρα ενισχύουν το συγκεκριμένο φαινόμενο. Αυτά τα μέτρα περιλαμβάνουν την ενίσχυση των υποδομών, εκπαίδευση του προσωπικού, την εξασφάλιση επιδότησης, προσαρμογή φορολογικών πολιτικών και τη θέσπιση απαγορευτικών διατάξεων. Όπως έχει αποδειχθεί παραπάνω η ανάλυση συστάδων μπορεί να εφαρμοστεί σε πολλές επιχειρηματικές δραστηριότητες και προβλήματα. Σε κάθε περίπτωση οι αναλυτές οφείλουν να ξέρουν ότι στην ανάλυση συστάδων απαιτείται πειραματισμός με διάφορα γνωρίσματα και μεθόδους μέχρι την εξαγωγή συμπερασμάτων. Η ανάλυση συστάδων διαθέτει αρκετά έντονα περιγραφικό και διερευνητικό χαρακτήρα. Τα αποτελέσματα της μπορούν να συνδυαστούν με τη γνώμη από στελέχη τα οποία έχουν

ειδικές γνώσεις πάνω στο συγκεκριμένο θέμα. Αυτά τα στελέχη θα αξιολογήσουν τα αποτελέσματα των αλγορίθμων και θα αποφανθούν αν είναι λογικά και αν είναι χρήσιμα.

### 2.1.1 K-means Clustering

Η ομαδοποίηση k -means είναι μια μέθοδος κβαντοποίησης διανυσμάτων, αρχικά από την επεξεργασία σήματος, που στοχεύει να χωρίσει n παρατηρήσεις σε k συμπλέγματα στα οποία κάθε παρατήρηση ανήκει στο σύμπλεγμα με τον πλησιέστερο μέσο όρο (κέντρα συστάδων ή κέντρο συστάδων ), που χρησιμεύει ως πρωτότυπο το σύμπλεγμα. Αυτό έχει ως αποτέλεσμα την κατάτμηση του χώρου δεδομένων σε κελιά Voronoi. Το k -means Clustering ελαχιστοποιεί τις διακυμάνσεις εντός του συμπλέγματος ( τετράγωνες Ευκλείδειες αποστάσεις ), αλλά όχι τις κανονικές Ευκλείδειες αποστάσεις, που θα ήταν το πιο δύσκολο πρόβλημα Weber: ο μέσος όρος βελτιστοποιεί τα τετράγωνα σφάλματα, ενώ μόνο η γεωμετρική διάμεσος ελαχιστοποιεί τις Ευκλείδειες αποστάσεις. Για παράδειγμα, καλύτερες ευκλείδειες λύσεις μπορούν να βρεθούν χρησιμοποιώντας k -διάμεσους και k-medoids.

Το πρόβλημα είναι υπολογιστικά δύσκολο ( NP-hard ). Ωστόσο, αποτελεσματικοί ευρετικοί αλγόριθμοι συγκλίνουν γρήγορα σε ένα τοπικό βέλτιστο. Αυτά είναι συνήθως παρόμοια με τον αλγόριθμο προσδοκίας-μεγιστοποίησης για μείγματα γκαουσιανών κατανομών μέσω μιας επαναληπτικής προσέγγισης βελτίωσης που χρησιμοποιείται τόσο από την k- μέση όσο και από τη μοντελοποίηση μειγμάτων Gauss. Και οι δύο χρησιμοποιούν κέντρα συμπλέγματος για να μοντελοποιήσουν τα δεδομένα. Ωστόσο, η ομαδοποίηση k -means τείνει να βρίσκει συστάδες συγκρίσιμης χωρικής έκτασης, ενώ το μοντέλο μίξης Gauss επιτρέπει στις συστάδες να έχουν διαφορετικά σχήματα.

Ο μη εποπτευόμενος αλγόριθμος K-means έχει μια χαλαρή σχέση με τον ταξινομητή k -πλησιέστερου γείτονα, μια δημοφιλής εποπτευόμενη τεχνική μηχανικής μάθησης για ταξινόμηση που συχνά συγχέεται με το k -means λόγω του ονόματος. Η εφαρμογή του ταξινομητή 1-πλησιέστερου γείτονα στα κέντρα συμπλέγματος που λαμβάνονται από το k -means ταξινομεί νέα δεδομένα στις υπάρχουσες συστάδες. Αυτός είναι γνωστός ως ο πλησιέστερος κεντροειδής ταξινομητής ή αλγόριθμος Rocchio.

Δεδομένου ενός συνόλου παρατηρήσεων ( $x_1, x_2, ..., x_n$ ), όπου κάθε παρατήρηση είναι ένα d -διάστατο πραγματικό διάνυσμα, η ομαδοποίηση k -means στοχεύει να χωρίσει τις n παρατηρήσεις σε k (≤ n ) σύνολα S = { $S_1, S_2, ..., S_k$ } έτσι ώστε να ελαχιστοποιηθεί το άθροισμα τετραγώνων εντός του συμπλέγματος (WCSS) (δηλ. διακύμανση ). Επίσημα, ο στόχος είναι να βρεθεί:

$$E = \sum k\ i=1 \sum x \in Ci\ |xi - x\ |$$

Η ισοδυναμία μπορεί να συναχθεί από την ταυτότητα. Εφόσον η συνολική διακύμανση είναι σταθερή, αυτό ισοδυναμεί με τη μεγιστοποίηση του αθροίσματος των τετραγωνικών αποκλίσεων μεταξύ σημείων σε διαφορετικά συμπλέγματα (μεταξύ αθροίσματος τετραγώνων, BCSS). Αυτή η ντετερμινιστική σχέση σχετίζεται επίσης με τον νόμο της ολικής διακύμανσης στη θεωρία πιθανοτήτων (MacQueen, 1967).

## 2.1.2 Ανάλυση RFM

Η ανάλυση αυτή αποτελεί μια από τις πιο γνωστές τεχνικές εξόρυξης δεδομένων και ειδικά στο μάρκετινγκ. Το μοντέλο RFM είναι η πιο συχνά υιοθετημένη τεχνική τμηματοποίησης που περιλαμβάνει τρία μέτρα (επιείκεια, συχνότητα και νομισματική), τα οποία συνδυάζονται σε τριψήφιο κωδικό κυψελών RFM, που καλύπτει πέντε ίσα τμήματα (ομάδα 20%). Μεταξύ των τριών μέτρων RFM, η επιείκεια θεωρείται συχνά ως η σημαντικότερη. Ωστόσο, σύμφωνα με προηγούμενα πορίσματα, οι τιμές RFM τείνουν να είναι συγκεκριμένες και βασίζονται στη φύση των προϊόντων (Lumsden et al., 2008). Για παράδειγμα, η Fader et al. (2005) διαπίστωσε ότι, για χαμηλότερη διαφάνεια, οι πελάτες με μεγαλύτερη συχνότητα έτειναν να έχουν χαμηλότερο μελλοντικό αγοραστικό δυναμικό από εκείνους με χαμηλότερα ποσοστά στις τιμές προαγοράς. Οι Lumsden et al. (2008) έχουν παρόμοια ευρήματα ότι υπάρχουν σημαντικές διαφορές μεταξύ των ομάδων σε όλη την επιείκεια και τη συχνότητα. Η διαδικασία ποσοτικοποίησης της συμπεριφοράς των πελατών μέσω του μοντέλου RFM είναι η εξής. Πρώτα, ταξινομείται η βάση δεδομένων κατά κάθε διάσταση RFM και, στη συνέχεια, διαιρείται η λίστα πελατών σε πέντε ίσα τμήματα. Η μέθοδος είναι γνωστό ότι έχει ακριβώς ίσο μέγεθος. Τα διαφορετικά τμήματα RFM έχουν διαφορετικά ποσοστά απόκρισης. Για την επιείκεια, οι πελάτες ταξινομούνται κατά ημερομηνίες αγοράς. Η επιείκεια ορίζεται συνήθως από τον αριθμό των περιόδων από την τελευταία αγορά, η οποία μετρά το διάστημα μεταξύ του πιο πρόσφατου χρόνου συναλλαγής και του χρόνου ανάλυσης (ημέρες ή μήνες), δηλαδή, όσο χαμηλότερος είναι ο αριθμός των ημερών, τόσο υψηλότερη είναι η βαθμολογία της επιείκειας. Ένας πελάτης με υψηλή βαθμολογία επιείκειας υποδηλώνει ότι είναι πιο πιθανό να κάνει μια επαναλαμβανόμενη αγορά. Το κορυφαίο τμήμα 20% κωδικοποιείται ως 5, ενώ το επόμενο τμήμα 20% κωδικοποιείται ως 4 και ούτω καθεξής. Τέλος, η επιείκεια για κάθε πελάτη στη βάση δεδομένων υποδηλώνεται από έναν αριθμό από το 5 έως το 1 (Hughes, 1996).

Για τη συχνότητα, η βάση δεδομένων ταξινομείται κατά συχνότητα αγοράς (ο αριθμός των αγορών) που πραγματοποιούνται σε μια συγκεκριμένη χρονική περίοδο. Ο ορισμός της συχνότητας συχνά απλοποιείται για να εξεταστούν δύο μέρη, συμπεριλαμβανομένων των μεμονωμένων και επαναλαμβανόμενων αγορών. Στο επάνω μέρος αποδίδεται τιμή 5 και στα άλλα δίνονται οι τιμές των 4, 3, 2 και 1. Ωστόσο, η υψηλότερη βαθμολογία συχνότητας υποδεικνύει μεγαλύτερη αφοσίωση των πελατών. Ένας πελάτης με υψηλή βαθμολογία συχνότητας υποδηλώνει ότι έχει μεγάλη ζήτηση για το προϊόν και είναι πιο πιθανό να αγοράσει τα προϊόντα επανειλημμένα. Για νομισματικούς λόγους, οι

πελάτες κωδικοποιούνται με το συνολικό ποσό των χρημάτων που δαπανούν κατά τη διάρκεια μιας καθορισμένης χρονικής περιόδου. Ο ορισμός του νομισματικού καθορίζεται από την αξία χρημάτων που ξόδεψε ο πελάτης σε αυτή τη χρονική περίοδο ή από το μέσο ποσό ανά αγορά ή από όλες τις αγορές μέχρι σήμερα. Ο Marcus (1998) πρότεινε ότι είναι προτιμότερο να χρησιμοποιείται το μέσο ποσό αγοράς παρά το συνολικό συσσωρευμένο ποσό αγοράς, ώστε να μειωθεί η συν-γραμμικότητα της συχνότητας και της νομισματικής. Με βάση τις βαθμολογίες συμπεριφοράς RFM που έχουν εκχωρηθεί, οι πελάτες μπορούν να ομαδοποιηθούν σε τμήματα και η κερδοφορία τους μπορεί να αναλυθεί περαιτέρω (Chang et al., 2010).

Υπάρχουν διάφοροι λόγοι για τους οποίους το μοντέλο RFM είναι δημοφιλές στην τμηματοποίηση άμεσου μάρκετινγκ για δεκαετίες. Πρώτον, το RFM είναι οικονομικά αποδοτικό στην απόκτηση σημαντικής ανάλυσης συμπεριφοράς πελατών και είναι εύκολο να ποσοτικοποιηθεί η συμπεριφορά των πελατών (Kahan, 1998. Miglautsch, 2000), όπου οι πελάτες και τα δεδομένα συναλλαγών μπορούν να αποθηκευτούν σε προσβάσιμη ηλεκτρονική μορφή (Lumsden et al., 2008). Ως εκ τούτου, οι υπεύθυνοι λήψης αποφάσεων μπορούν εύκολα να κατανοήσουν την εφαρμογή του μοντέλου RFM (McCarty και Hastak, 2007). Δεύτερον, η RFM είναι πολύ πολύτιμη στην πρόβλεψη της ανταπόκρισης και μπορεί να ενισχύσει τα κέρδη μιας εταιρείας βραχυπρόθεσμα (Baecke και Van den Poel, 2009). Τρίτον, είναι πολύ αποτελεσματικό να διαμορφώνεται με μεταβλητές RFM, καθώς η συμπεριφορά αγοράς μπορεί να συνοψιστεί χρησιμοποιώντας έναν πολύ μικρό αριθμό μεταβλητών. Τέταρτον, οι μεταβλητές RFM συλλέγονται μέσω εσωτερικής βάσης δεδομένων που περιέχει πληροφορίες ειδικά για τον πελάτη σχετικά με το ιστορικό συναλλαγών και δεν λαμβάνονται μέσω των συγκεντρωτικών πληροφοριών επιπέδου στις δημογραφικές βάσεις δεδομένων. Ως εκ τούτου, η RFM είναι πιο σημαντική για τη στόχευση συγκεκριμένων πελατών (Kaymak, 2001). Πέμπτο, η RFM είναι μια πολύ γνωστή μέθοδος για τη μέτρηση της δύναμης της σχέσης με τους πελάτες, καθώς η RFM μπορεί να εντοπίσει αποτελεσματικά πολύτιμους πελάτες (Wang, 2010).

Αν και το μοντέλο RFM είναι ένα κρίσιμο εργαλείο για τις επιχειρήσεις να αναπτύξουν στρατηγικές μάρκετινγκ, έχει επίσης πολλά μειονεκτήματα. Πρώτον, δεδομένου ότι η RFM στοχεύει στον εντοπισμό πολύτιμων πελατών σε εταιρείες, επικεντρώνεται μόνο στους καλύτερους πελάτες. Παρέχει μικρή ουσιαστική βαθμολογία στην επιείκεια, τη συχνότητα και τη νομισματική όταν οι περισσότεροι πελάτες δεν αγοράζουν συχνά, ξοδεύουν λίγα και δεν έχουν αγοράσει πρόσφατα. Αυτό ισχύει ιδιαίτερα για τις περισσότερες πωλήσεις των επιχειρήσεων, το 80% των οποίων προέρχεται από το 20% των πελατών (Hughes, 1996· Wang, 2010) (κανόνας 80 έως 20). Κατά συνέπεια, αγνοεί την ανάλυση σχετικά με τις νέες επιχειρήσεις που ιδρύονται σε σύντομο χρονικό διάστημα και τους πελάτες που αγοράζουν μόνο μία φορά και τοποθετούν μικρές παραγγελίες. Η Miglautsch (2002) αναφέρθηκε σε αυτόν τον τύπο πελατών ως πελάτες 1-1-1 και ισχυρίστηκε ότι είναι το μεγαλύτερο τμήμα πελατών και μπορεί να έχει το μεγαλύτερο ανεκμετάλλευτο δυναμικό. Δεύτερον, το μοντέλο RFM μπορεί να χρησιμοποιήσει μόνο περιορισμένο αριθμό μεταβλητών επιλογής. Ωστόσο, τα περισσότερα οικιακά χαρακτηριστικά έχουν

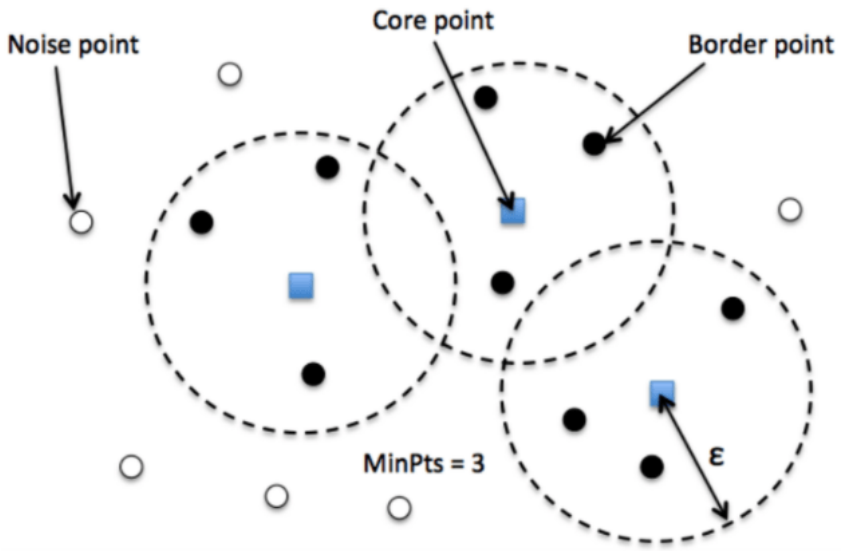
επίδραση στην πιθανότητα ανταπόκρισης των πελατών. Η απλότητα του μοντέλου RFM έχει υπερτονιστεί και η ικανότητά του να διαφοροποιείται δεν έχει πολλά να σκεφτεί. Προηγούμενη βιβλιογραφία έχει δείξει ότι είναι καλύτερο να λαμβάνονται υπόψη οι σχεσιακές πληροφορίες κατά τη χρήση μοντέλων RFM (McCarty και Hastak, 2007). Τρίτον, εκτός εάν οι μεταβλητές RFM είναι όλες αμοιβαία ανεξάρτητες, το μοντέλο RFM δεν μετρά διπλά (Bult και Wansbeek, 1995). Τέταρτον, η RFM επικεντρώνεται στους σημερινούς πελάτες μιας εταιρείας και δεν μπορεί να εφαρμοστεί στην αναζήτηση νέων πελατών, καθώς ένας έμπορος δεν έχει συναλλαγές για προοπτικές (McCarty και Hastak, 2007). Πέμπτον, η RFM εκτιμά ένα ενιαίο μοντέλο απόκρισης για όλους τους πελάτες στη βάση δεδομένων και, ως εκ τούτου, αναλαμβάνει την ομοιογενή βάση δεδομένων πελατών, η οποία συχνά αντιβαίνει στην πραγματική κατάσταση ότι οι πελάτες συχνά έχουν σημαντική ανομοιογένεια (Suh et al., 1999). Έκτο, η RFM δεν προβλέπεται ως ακριβές ποσοτικό μοντέλο και η σημασία κάθε μέτρου RFM είναι διαφορετική μεταξύ των βιομηχανιών (Yeh et al., 2009).

Συνοπτικά, το μοντέλο RFM έχει αδυναμίες σε ορισμένους τομείς που οδηγούν σε ορισμένες αποφάσεις σχετικά με τα μειονεκτήματα του μοντέλου RFM ή κάποια μικρή τροποποίηση ή επέκταση του μοντέλου RFM. Για παράδειγμα, ο Miglautsch (2002) πρότεινε ότι η υπο-τμηματοποίηση μπορεί να βοηθήσει στον εντοπισμό πελατών 1-1-1, που περιλαμβάνουν τρεις κατηγορίες μεταβλητών: εσωτερικές πληροφορίες αγοράς, γεω-δημογραφικές πληροφορίες που συνδέονται με ταχυδρομικό κώδικα και μεταβλητές πελατών. Η ανάλυση RFM συνδυάζεται με την clustering τεχνική έτσι ώστε να έχουμε αποτελέσματα RFM για κάθε ένα από τα clusters και έπειτα να δημιουργηθεί το προφίλ των πελατών ανάλογα με τις τιμές RFM που έχουν προκύψει.

### 2.1.3 DBSCAN

Μια από τις πιο δημοφιλής μεθόδους με βάση την πυκνότητα αποτελεί η «*χωρική ομαδοποίηση εφαρμογών με θόρυβο βάσει πυκνότητας*» ( DBSCAN ). Έχει ένα καλά καθορισμένο μοντέλο συμπλέγματος που ονομάζεται « *προσιτότητα πυκνότητας* ». η συγκεκριμένη τεχνική ομαδοποίησης προκαλεί την σύνδεση των σημείων δεδομένων, ικανοποιώντας συγκεκριμένα κριτήρια πυκνότητας (ελάχιστος αριθμός αντικειμένων σε μια ακτίνα). Με την ολοκλήρωση της ομαδοποίησης DBSCAN, υφίστανται 3 τύποι σημείων: πυρήνας, περίγραμμα και θόρυβος.

(https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html)




*Εικόνα 1: Τύποι σημείων DBSCAN*

Στο σχήμα, «*ο πυρήνας αποτελεί σημείο με μερικά (m) σημεία σε μια συγκεκριμένη (n) απόσταση από τον εαυτό του. Το περίγραμμα αποτελεί σημείο με τουλάχιστον ένα κεντρικό σημείο στην απόσταση n. Ο θόρυβος αποτελεί σημείο που δεν είναι ούτε σύνορο ούτε πυρήνας. Σημεία δεδομένων σε αραιές περιοχές που απαιτούνται για διαχωρισμό συστάδων θεωρούνται θόρυβος και ευρύτερα σημεία*». ([https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html](https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html)).

Το DBSCAN κάνει χρήση δύο παραμέτρων για τον προσδιορισμό των συμπλεγμάτων :

- «*minPts : για να θεωρηθεί μια περιοχή πυκνή, ο ελάχιστος αριθμός πόντων που απαιτείται είναι « minPts* ».
- «*eps : για τον εντοπισμό σημείων δεδομένων στη γειτονιά οποιωνδήποτε σημείων, το « eps (ε) » χρησιμοποιείται ως μέτρο απόστασης*».

Ο αλγόριθμος DBSCAN:

- *«Το DBSCAN ξεκινά με ένα τυχαίο σημείο δεδομένων (σημεία χωρίς επίσκεψη).*
- *Η γειτονιά αυτού του σημείου εξάγεται χρησιμοποιώντας ένα έψιλον απόστασης ε.*
- *Η διαδικασία ομαδοποίησης ξεκινάει εάν υπάρχουν επαρκή σημεία δεδομένων σε αυτήν την περιοχή και το τρέχον σημείο δεδομένων γίνεται το πρώτο σημείο στο νεότερο σύμπλεγμα, ή εάν το σημείο έχει επισημανθεί ως θόρυβος και έχει γίνει επίσκεψη.*
- *Το σημείο εντός της γειτονιάς απόστασης του έψιλον ε γίνεται επίσης μέρος του ίδιου συμπλέγματος για το πρώτο σημείο στο νέο σύμπλεγμα. Για όλα τα νέα σημεία δεδομένων που προστέθηκαν στο παραπάνω σύμπλεγμα, επαναλαμβάνεται η διαδικασία για να ανήκουν όλα τα σημεία δεδομένων στο ίδιο σύμπλεγμα.*

- *Τα παραπάνω δύο βήματα επαναλαμβάνονται μέχρι να καθοριστούν όλα τα σημεία του συμπλέγματος. Όλα τα σημεία εντός της γειτονιάς ε του συμπλέγματος έχουν επισκεφθεί και έχουν επισημανθεί. Μόλις τελειώσουμε με το τρέχον σύμπλεγμα, ανακτάται και υποβάλλεται σε επεξεργασία ένα νέο μη επισκέψιμο σημείο, οδηγώντας σε περαιτέρω ανακάλυψη του συμπλέγματος ή του θορύβου. Η διαδικασία επαναλαμβάνεται μέχρι να επισημανθούν όλα τα σημεία δεδομένων ως επισκέψιμα».*

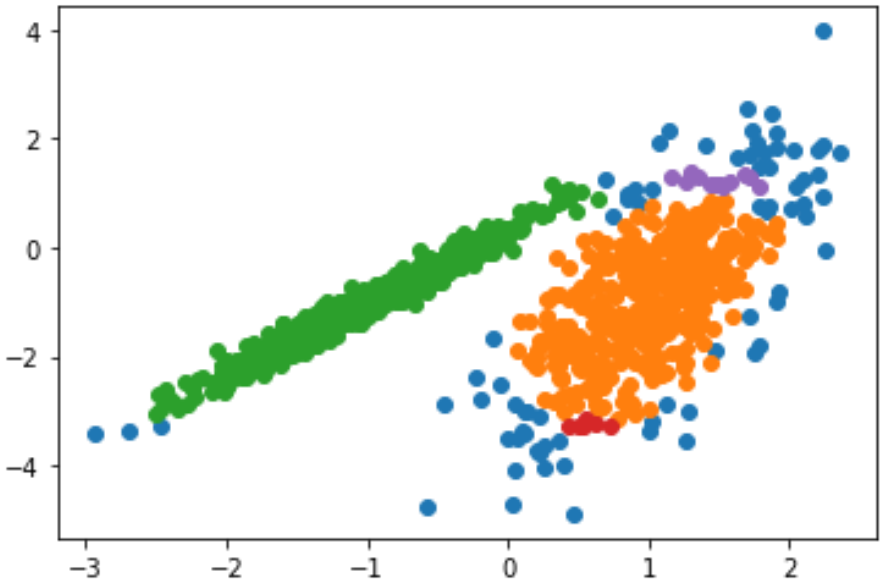
Ιδιότητα του DBSCAN αποτελεί επίσης η χαμηλή πολυπλοκότητά του. Απαιτεί έναν γραμμικό αριθμό ερωτήσεων εύρους στη βάση δεδομένων.

Το βασικό ζήτημα με το DBSCAN είναι ότι:

- Μπορεί να ρίξει την πυκνότητα για την ανίχνευση των ορίων των συστάδων. Το DBSCAN συνδέει περιοχές με υψηλή πυκνότητα. Τέλος ο αλγόριθμος είναι πιο καλός από το K-means στην περίπτωση δεδομένων με περίεργο σχήμα.

Επίσης δεν απαιτεί προκαθορισμένο αριθμό συμπλεγμάτων. Προσδιορίζει επίσης το θόρυβο και τα ακραία σημεία. Επιπλέον, τα συμπλέγματα αυθαίρετου μεγέθους και σχήματος βρίσκονται αρκετά καλά από τον αλγόριθμο.

(https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html)



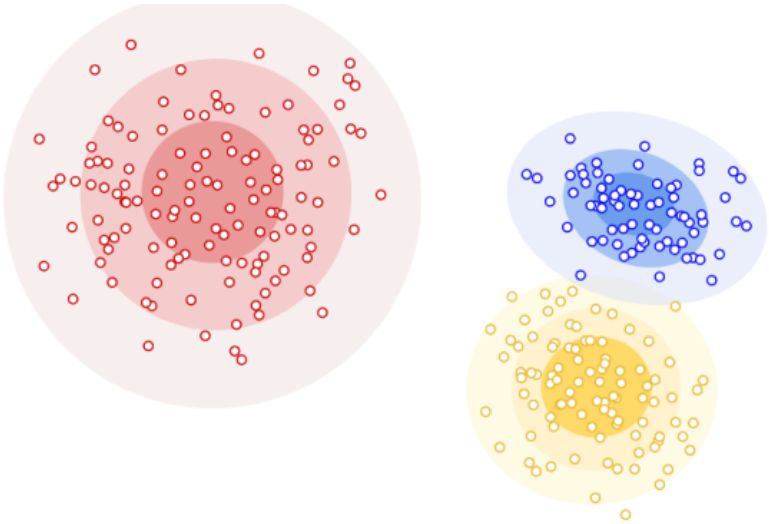

*Γράφημα 1: Συστάδες που λαμβάνονται από τον αλγόριθμο συμπλέγματος DBSCAN*

## 2.1.4 Gauss

Το μοντέλο μείγματος Gaussian αποτελεί έναν από τους τύπους ομαδοποίησης με βάση τη διανομή. Η ομαδοποίηση αυτή υποθέτει ότι τα δεδομένα περιλαμβάνουν κατανομές, όπως οι κατανομές Gauss. Παρακάτω, ο αλγόριθμος στο σχήμα έχει ως βάση του τη διανομή, ομαδοποιώντας τα δεδομένα

σε τρεις κατανομές Gauss. Με την αύξηση της απόστασης από την κατανομή, υπάρχει και μείωση της πιθανότητας το σημείο να ανήκει στην κατανομή.

Το μοντέλο χρησιμοποιείται επίσης για να βρίσκονται συμπλέγματα με τον ίδιο τρόπο όπως το K-means. Ένα σημείο μπορεί να ανήκει στο κέντρο της διανομής και να μειώνεται, όσο ταυτόχρονα αυξάνεται η απόσταση από το κέντρο διανομής. Παρακάτω φαίνεται η μείωση της πιθανότητας. Από αυτό το μοντέλο υπάρχει και η πιθανότητα εύρεσης εκχώρησης πιθανοτικών συστάδων. Όταν δεν είναι γνωστός ο τύπος κατανομής δεδομένων, θα πρέπει να χρησιμοποιείται διαφορετικός αλγόριθμος (https://developers.google.com/machine-learning/clustering/clustering-algorithms).

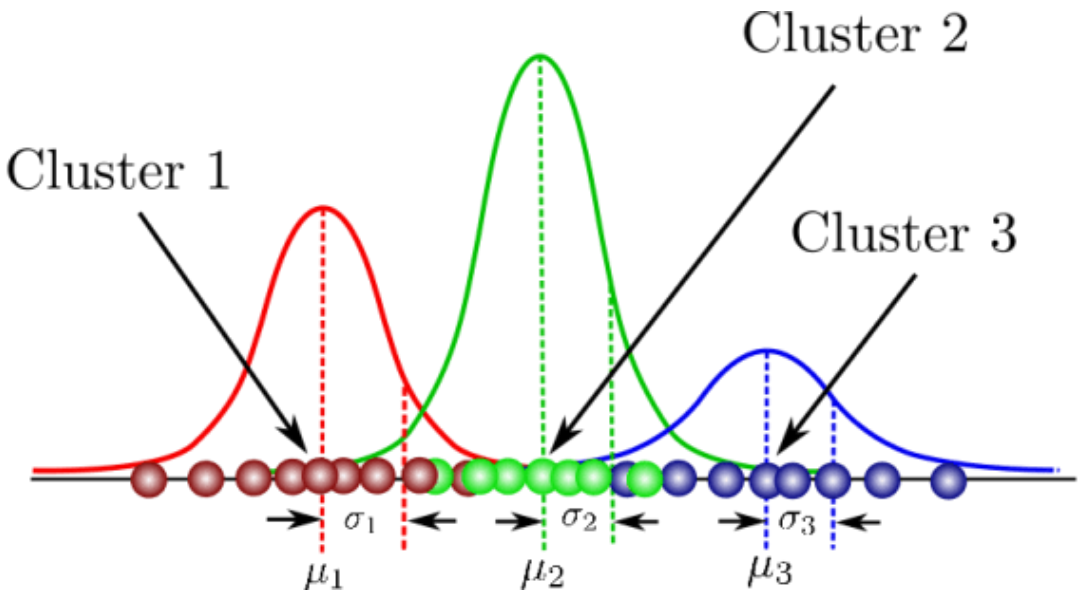


*Εικόνα 2: Παράδειγμα ομαδοποίησης βάσει διανομής*

Υπολογισμός πιθανοτήτων και εκχώρηση στο σημείο δεδομένων:

- «*Μια GM είναι μια συνάρτηση που αποτελείται από πολλά Gaussians, καθένα από τα οποία προσδιορίζεται με* $k \in \{1, \ldots, K\}$*, όπου K είναι ο αριθμός των συστάδων*».
- Κάθε Gaussian K στο μείγμα αποτελείται από τις ακόλουθες παραμέτρους:
    - Ένας μέσος όρος μ που ορίζει το κέντρο του.
    - Μια συνδιακύμανση Σ που ορίζει το πλάτος του.
    - Μια πιθανότητα ανάμειξης που καθορίζει πόσο μεγάλη ή μικρή ή μεγάλη θα είναι η Gaussian συνάρτηση.

Αυτές οι παράμετροι φαίνονται στην παρακάτω εικόνα:




*Εικόνα 3: Παράμετροι Μοντέλου μείγματος Gauss*

Για να βρει τη συνδιακύμανση, τον μέσο όρο, τη διακύμανση και τα βάρη των συστάδων, το GMM χρησιμοποιεί την τεχνική Μεγιστοποίησης προσδοκιών. Πρέπει να εκχωρήσουμε K αριθμό συστάδων, που σημαίνει K Gaussian κατανομές, με τις τιμές μέσου όρου και συνδιακύμανσης να είναι μ1, μ2, .. μk και Σ1, Σ2, .. Σk. Υπάρχει μια άλλη παράμετρος Πi που αντιπροσωπεύει την πυκνότητα κατανομής. Για να ορίσουμε την κατανομή Gauss πρέπει να βρούμε τις τιμές για αυτές τις παραμέτρους. Έχουμε ήδη αποφασίσει για τον αριθμό των συστάδων και έχουμε εκχωρήσει την τιμή για τη μέση τιμή, τη συνδιακύμανση και την πυκνότητα. Ακολουθούν το βήμα προσδοκίας και το βήμα μεγιστοποίησης.

Πλεονεκτήματα του GMM:

- Ένα από τα πλεονεκτήματα του GMM έναντι του K-means είναι ότι το K-means δεν λαμβάνει υπόψη τη διακύμανση (εδώ, η διακύμανση αναφέρεται στο πλάτος της καμπύλης σε σχήμα καμπάνας) και το GMM επιστρέφει την πιθανότητα τα σημεία δεδομένων να ανήκουν σε καθεμία από τις ομάδες K.
- Σε περίπτωση επικαλυπτόμενων συστάδων, όλοι οι παραπάνω αλγόριθμοι ομαδοποίησης αποτυγχάνουν να το αναγνωρίσουν ως ένα σύμπλεγμα.
- Το GMM χρησιμοποιεί μια πιθανολογική προσέγγιση και παρέχει πιθανότητα για κάθε σημείο δεδομένων που ανήκει στα συμπλέγματα.

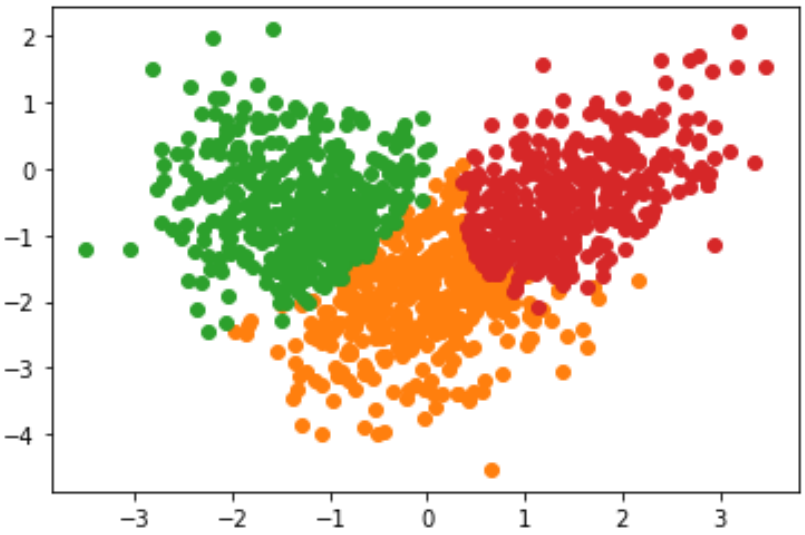
Μειονεκτήματα του GMM

- Τα μοντέλα μιγμάτων είναι υπολογιστικά ακριβά εάν ο αριθμός των διανομών είναι μεγάλος ή το σύνολο δεδομένων περιέχει λιγότερα παρατηρούμενα σημεία δεδομένων.

- Χρειάζεται μεγάλα σύνολα δεδομένων και είναι δύσκολο να εκτιμηθεί ο αριθμός των συμπλεγμάτων.

Ας δούμε τώρα πώς το GMM συγκεντρώνει δεδομένα. Ο παρακάτω κώδικας βοηθά:

- Στη δημιουργία δεδομένων,
- Προσαρμογή των δεδομένων στο μοντέλο «GaussianMixture»,
- Εύρεση τω σημείων δεδομένων που έχουν εκχωρηθεί σε ένα σύμπλεγμα,
- Απόκτηση των μοναδικών συμπλεγμάτων και
- Σχεδιασμό των συμπλεγμάτων (https://developers.google.com/machine-learning/clustering/clustering-algorithms).




*Γράφημα 2: Συστάδες που προέκυψαν από τον αλγόριθμο του Gaussian Mixture Model*

## 2.2 Ανάλυση Καλαθιού Αγοράς (MBA)

Η ανάλυση καλαθιού αγορών προϋποθέτει ότι υπάρχουν εξαρτήσεις μεταξύ των κατηγοριών προϊόντων, οι οποίες πρέπει να ενσωματωθούν στις προσεγγίσεις μοντέλων λόγω περιεχομένου και μεθοδολογικών πτυχών, προκειμένου να διασφαλιστεί η ενημερωτική αξία των εκτιμητών. Ο όρος «ανάλυση καλαθιού αγορών» ορίζει μοντέλα που προβλέπουν την πιθανότητα επιλογής δέσμης ή καλαθιού πολλών αντικειμένων κατά τη διάρκεια της συναλλαγής αγοράς (Russell και Petersen,2000). Όλα τα μοντέλα υποθέτουν ότι συνεπάγεται ένας σχεδιασμός μιας εξαρτώμενης από το καλάθι αποφάσεων επιλογής (Russell et al. 1997). Στην ανάλυση καλαθιού, επομένως, θα πρέπει να αποκαλυφθούν και να ποσοτικοποιηθούν σύνθετες σχέσεις εντός του καλαθιού αγορών και να εκτιμηθεί η πιθανότητα επιλογής ενός συγκεκριμένου καλαθιού αγορών.

Ο Russell et al. (1999) ορίζει την «επιλογή πολλαπλών κατηγοριών» ως διαδικασία λήψης αποφάσεων κατά την οποία η επιλογή ενός προϊόντος επηρεάζεται από την παρουσία ενός ή περισσότερων άλλων προϊόντων από άλλες κατηγορίες. Ως αποτέλεσμα, η λήψη αποφάσεων σε πολλές κατηγορίες συνεπάγεται διασταυρούμενη εξάρτηση. Οι Shocker et al. (2004) παρέχουν επίσης μια ολοκληρωμένη αιτιολόγηση της συνάφειας των "άλλων προϊόντων". Από νωρίς, οι Böcker και Merkle (1975, σ. 184) τόνισαν σαφώς το κίνητρο της κοινής ανάλυσης. Η Müller-Hagedorn (1978) επισημαίνει ότι η γνώση της ζήτησης των καταναλωτών έχει μεγάλη σημασία για τη χρήση των μέτρων πολιτικής πωλήσεων. Εάν, για παράδειγμα, μια εταιρεία δεν έλαβε υπόψη το γεγονός ότι οι αγοραστές εξαρτούν την αγορά ενός αγαθού από την αγορά άλλου αγαθού, δηλαδή εάν προσπαθούσε να επηρεάσει τη συμπεριφορά των πελατών όσον αφορά μεμονωμένα προϊόντα, αυτό θα εμπόδιζε τον βέλτιστο σχεδιασμό (Müller-Χάγκεντορν,1978, σ. 181). Οι λιανοπωλητές προσπαθούν να μεγιστοποιήσουν τον κύκλο εργασιών και τα κέρδη, τα οποία μπορούν να επιτευχθούν μόνο με τον συντονισμό στρατηγικών μάρκετινγκ μεταξύ κατηγοριών και μεταξύ προϊόντων μιας κατηγορίας (Song & Chintagunta, 2003). Για τη σύσταση στρατηγικών μάρκετινγκ, χρησιμοποιούνται πρότυπα συσχέτισης στις αγορές κατηγοριών (Russell και Petersen, 2000), έτσι ώστε οι στρατηγικές τιμολόγησης και προώθησης να μπορούν να δοκιμαστούν σε όλες τις κατηγορίες (Manchanda et al.,1999). Ως εκ τούτου, η ανάλυση καλαθιού αγορών είναι μια κατάλληλη μέθοδος για την αποκάλυψη των προτύπων αγορών των καταναλωτών (Chen et al., 2005).

Η συμπεριφορά των καταναλωτών είναι μια καταναλωτική δραστηριότητα στην απόφαση αγοράς, χρήσης και κατανάλωσης των αγορασμένων αγαθών και υπηρεσιών, συμπεριλαμβανομένων των παραγόντων των πελατών που μπορούν να οδηγήσουν στις αποφάσεις τους σχετικά με την αγορά και τη χρήση προϊόντων. Κάθε πελάτης έχει διαφορετικές ανάγκες και κλίσεις καθώς και έχει διαφορετικές συμπεριφορές στην εκπλήρωση αυτών των πραγμάτων. Ωστόσο, σε περίπτωση διαφορετικών συμπεριφορών για την κάλυψη των αναγκών τους, εξακολουθούν να μοιράζονται κάποιες ομοιότητες, μία από αυτές επιθυμεί να μεγιστοποιήσει την ικανοποίησή τους για την κατανάλωση ενός απαραίτητου



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

προϊόντος ή υπηρεσίας. Από αυτή τη δραστηριότητα κατανάλωσης, αυτό μπορεί να συναχθεί ως προς τη συμπεριφορά, το μοτίβο ή τη συνήθεια που κάνουν οι πελάτες στην ικανοποίηση των αναγκών και των επιθυμιών τους. Η καταγραφή απογειώνεται λόγω της απαιτούμενης τεκμηρίωσης και ταυτοποίησης δεδομένων ιστορικού κατά την εκτέλεση της συναλλακτικής δραστηριότητας.

Τα τελευταία χρόνια, τα δεδομένα συναλλαγών χρησιμοποιούνται συνήθως ως αντικείμενα έρευνας και ανάλυσης για ερευνητές. Για παράδειγμα, πληροφορίες ενός είδους του οποίου οι πωλήσεις είναι οι υψηλότερες. Εκτός αυτού, οι πληροφορίες μπορούν να χρησιμοποιηθούν σε σχέση με την προσθήκη αποθεμάτων αυτού του στοιχείου. Επιπλέον, από τα δεδομένα συναλλαγών μπορεί να χρησιμοποιηθεί ως προς τη σχέση κάθε αγορασμένου αντικειμένου μέσα στο καλάθι του πελάτη. Με αυτές τις πληροφορίες, μπορούμε να το χρησιμοποιήσουμε για αποτελεσματική προβολή/ποικιλία προϊόντων για να προσελκύσουμε το ενδιαφέρον των πελατών. Η κοινώς χρησιμοποιούμενη εφαρμογή για την ανάλυση του καλαθιού αγορών των πελατών δεδομένων συναλλαγών είναι η ανάλυση καλαθιού αγοράς. Η ανάλυση καλαθιού αγοράς είναι ένας από τους τρόπους από την τεχνική εξόρυξης δεδομένων που χρησιμοποιείται επικρατέστερα για την ανάλυση αντικειμένων/αγαθών σε ένα ή περισσότερα καλάθια αγορών που έχει ένας πελάτης σε μια συγκεκριμένη στιγμή (Kaur & Kang, 2016). Η εφαρμογή ανάλυσης καλαθιού αγοράς θα πρέπει να σχεδιάζεται και να εφαρμόζεται όχι μόνο λόγω της δυνατότητας να βοηθήσει το σχεδιασμό προώθησης των πωλήσεων, αλλά και να είναι σε θέση να γίνει ως αναφορά στην εκ νέου διαχείριση των εισερχόμενων και επερχόμενων αποθεμάτων αντικειμένων στην αποθήκη.

Η ανάλυση καλαθιού αγοράς είναι ένας τρόπος ανάλυσης που πραγματοποιείται στη συμπεριφορά των πελατών μέσω των μέσων αναγνώρισης της συσχέτισης και των συνδέσεων μεταξύ των διαφόρων αντικειμένων που τοποθετούνται από τους πελάτες στα καλάθια αγορών τους (Mansur& Kuncoro, 2012). Συγκεκριμένα, η Ανάλυση Καλαθιού Αγοράς στοχεύει στον ταυτόχρονα εντοπισμό των πιο συχνά αγορασμένων αντικειμένων από τους πελάτες. Χρησιμοποιώντας τη λειτουργία ανάλυσης καλαθιού αγοράς, μπορεί κάποιος να αποκτήσει γνώση των ειδών που αγοράζονται συχνά από τους πελάτες ταυτόχρονα και έχουν την ευκαιρία να προωθηθούν. Όσον αφορά τον στόχο του τρόπου ανάλυσης καλαθιού αγοράς για να αποφασιστεί ποια προϊόντα αγοράζουν ταυτόχρονα οι πελάτες, με το οποίο το όνομα αυτής της λειτουργίας λαμβάνεται από τη συμπεριφορά των πελατών στην τοποθέτηση προϊόντων αγορών στα καλάθια αγορών ή στη λίστα αγορών τους. Πέρα από τον προσδιορισμό του τρόπου καλαθιού αγορών ενός πελάτη θα είναι σημαντικά σε θέση να βοηθήσει μια εταιρεία στη χρήση αυτών των πληροφοριών όσον αφορά τις ανάγκες επιχειρηματικής στρατηγικής, ένα από αυτά τοποθετεί τα προϊόντα που αγοράζονται συχνότερα ταυτόχρονα σε έναν συγκεκριμένο τομέα. Ο κανόνας της Ένωσης Κανόνων Σύνδεσης σχετίζεται με τη δήλωση του "τι" με τι". Το θέμα αυτό μπορεί να έχει τη μορφή δήλωσης σχετικά με τη συναλλακτική δραστηριότητα που ασκείται από τους πελάτες. Από τη δήλωση αυτή, υπάρχει ισχυρή σχέση με τη μελέτη της βάσης δεδομένων συναλλαγών πελατών για τον προσδιορισμό της συνήθειας ενός αγορασμένου προϊόντος με ποιο

προϊόν, επομένως, ο κανόνας της σύνδεσης αναφέρεται συχνά ως ανάλυση καλαθιού αγοράς (Mansur& Kuncoro, 2012). Η σημασία ενός συνενωτικού κανόνα μπορεί να υποτεθεί με την παρουσία δύο παραμέτρων, δηλαδή της υποστήριξης και της εμπιστοσύνης. Η υποστήριξη (υποστηρικτική τιμή) είναι το ποσοστό των συνδυασμών στοιχείων προϊόντων στη βάση δεδομένων. Ενώ η εμπιστοσύνη (τιμή βεβαιότητας) είναι μια τιμή για τον προσδιορισμό της δύναμης των σχέσεων μεταξύ των στοιχείων στους κανόνες συσχέτισης.

$$S = \frac{\sum(Ta+Tc)}{\sum(T)} \quad (1)$$

$$Support\ (A) = \frac{The\ Number\ of\ Transaction\ that\ Contains\ A}{Total\ Transaction} \quad (2)$$

where:

S = Support

$\sum(Ta + Tc)$ = the number of transaction that contains *antecendent* and *consequent.*

$\sum(T)$ = the number of transaction.

$$C = \frac{\sum(Ta+Tc)}{\sum(Ta)} \quad (3)$$

$$Confidence\ =\ P(A \mid B)\ =\ \frac{The\ Number\ of\ Transaction\ that\ Contains\ A\ and\ B}{Total\ Transaction\ that\ Contains\ A} \quad (4)$$

where:

C = Confidence

$\sum(Ta + Tc)$ = the number of transaction that contains *antecendent* and *consequent.*

$\sum(Ta)$ = the number of transaction that contains *antecendent.*

Ο ορισμός της συμπεριφοράς των πελατών είναι μια δυναμική αλληλεπίδραση μεταξύ της γνώσης, της αγάπης, της συμπεριφοράς και του περιβάλλοντός της, με την οποία κάποιος εκτελεί δραστηριότητες ανταλλαγής στην τακτική του βάση (Su, 2007). Ενόψει αυτής της δήλωσης, υπάρχουν τρία σημαντικά ζητήματα που πρέπει να κατανοήσουμε, δηλαδή: 1. Η συμπεριφορά των πελατών φέρει δυναμικό χαρακτηριστικό, επομένως, δύσκολο να προβλεφθεί. 2. Περιλαμβάνει αλληλεπίδραση, όπως γνωστική λειτουργία, στοργή, συμπεριφορά και τα περιστατικά γύρω από τους πελάτες, 3. Περιλαμβάνει ανταλλαγή, όπως η ανταλλαγή αντικειμένων και χρημάτων από έμπορο σε πελάτη. Τέσσερις παράγοντες που θα μπορούσαν να προκαλέσουν την αγορά πελατών κατά τις αγορές (Armstrong, et al., 2014), είναι : 1. Πολιτιστικός Παράγοντας 2. Κοινωνικός παράγοντας 3. Προσωπικός παράγοντας 4. Ψυχολογικός παράγοντας. Υπήρχαν τρεις μεταβλητές που πρέπει να ληφθούν υπόψη στην κατανόηση της συμπεριφοράς των πελατών, δηλαδή η μεταβλητή ερεθισμάτων, η μεταβλητή απόκρισης και η παρεμβατική μεταβλητή (Sherman, Mathur, & Smith, 1997).

## 2.2.1 Τμηματοποίηση της Αγοράς

Η τμηματοποίηση της αγοράς εισήχθη για πρώτη φορά από τον Smith (1956), ο οποίος τον όρισε ως «Η τμηματοποίηση της αγοράς συνίσταται στην προβολή μιας ετερογενούς αγοράς ως μια σειρά μικρότερων ομοιογενών αγορών, ανταποκρινόμενη στις διαφορετικές προτιμήσεις, που αποδίδονται στις επιθυμίες των καταναλωτών για ακριβέστερη ικανοποίηση των διαφορετικών επιθυμιών τους». Ο κατακερματισμός της αγοράς έχει γίνει απαραίτητος για τη στρατηγική μάρκετινγκ, η οποία προσανατολίζει την εταιρεία να χρησιμοποιήσει τις περιορισμένες ικανότητές της για να εξυπηρετήσει τις απαιτήσεις των πελατών με τον πιο αποτελεσματικό τρόπο. Ο κατακερματισμός της αγοράς εξαρτάται σε μεγάλο βαθμό από τις βάσεις (μεταβλητές ή κριτήρια) και τις μεθόδους (Wedel & Kamakura 2000). Οι βάσεις επιλέγονται από το σκοπό της μελέτης και των ερωτήσεων μάρκετινγκ. Οι βάσεις ορίζουν και διακρίνουν την τμηματοποίηση. Η μέθοδος σχετίζεται στενά με τις βάσεις. Η μέθοδος τμηματοποίησης επιλέγεται επίσης από (1) σκοπό της μελέτης και (2) την καταλληλότητα με τις επιλεγμένες βάσεις. Υπάρχουν πολλά χαρακτηριστικά της αγοράς που μπορούν να επιλεγούν ως βάσεις για τον κατακερματισμό της αγοράς, γεγονός που οδηγεί σε πολλούς τρόπους για τον κατακερματισμό μιας αγοράς. Αλλά δεν είναι κάθε τρόπος αποτελεσματικός για να υποβληθεί αίτηση για στρατηγική μάρκετινγκ.

Οι Segal & Giacobbe (1994) έχουν δηλώσει δύο κυρίαρχα κριτήρια που εφαρμόζονται ειδικά για τα τμήματα της αγοράς λιανικής, τα οποία είναι: η συμπύκνωση και η ερμηνευτικότητα. «Τα τμήματα έπρεπε να είναι συμπαγή, διακριτά και έγκυρα για να έχει η έρευνα κάποια διαχειριστική αξία. Και η εξήγηση για τα τμήματα έπρεπε να είναι κατανοητή από τους διευθυντές, διαφορετικά η έρευνα δεν θα είχε ούτε πιστευτεί ούτε θα είχε γίνει αποδεκτή. Μαζί, τα κριτήρια αυτά συνέβαλαν στο κατά πόσον τα αποτελέσματα ήταν εφαρμόσιμα από διαχειριστική άποψη». Η Kotler (1988) έχει δημιουργήσει πέντε κριτήρια για την αξιολόγηση της αποτελεσματικότητας της τμηματοποίησης, τα οποία περιλαμβάνουν: μετρήσιμα, ουσιαστικά, προσβάσιμα, διαφορέα και εφαρμόσιμα. Αργότερα, η Wedel & Kamakura (2000) συνόψισε αρκετές επεκτάσεις του μοντέλου της Kotler, όπως τα έξι κριτήρια παρακάτω: - Αναγνωρίσιμη: είναι η επέκταση της μετρήσιμου, στην οποία οι πελάτες σε διαφορετική κατάτμηση είναι σε θέση να προσδιοριστούν από τις βάσεις της τμηματοποίησης και οι βάσεις (μεταβλητές) είναι εύκολο να μετρηθούν. - Ουσιαστικότητα: η τμηματοποίηση είναι αρκετά μεγάλη και δημιουργεί σημαντικό κέρδος για να εξυπηρετήσουν - Προσβασιμότητα: οι διαχειριστές μάρκετινγκ μπορούν να φτάσουν αποτελεσματικά στην τμηματοποίηση - Ανταπόκριση: είναι η επέκταση της διαφοροποίησης. Τα τμήματα είναι σε θέση να διακριθούν από τον ορισμό και ταυτόχρονα να έχουν διαφορετική ανταπόκριση από την προσπάθεια μάρκετινγκ. - Σταθερότητα: τα τμήματα είναι αρκετά σταθερά για να προσδιορίσουν και να εφαρμόσουν τη στρατηγική μάρκετινγκ. - Δυνατότητα δράσης: τα τμήματα πρέπει να έχουν νόημα για τον διευθυντή μάρκετινγκ όσον αφορά την καθοδήγηση για τη στρατηγική

μάρκετινγκ. Αυτά τα σύνολα κριτηρίων θα χρησιμοποιηθούν για την αξιολόγηση του κατακερματισμού της αγοράς αργότερα.

## 2.2.2 Association Rules

Η εξόρυξη κανόνων σύνδεσης έχει αναπτυχθεί για την ανάλυση δεδομένων καλαθιού σε ένα περιβάλλον μάρκετινγκ. Τα δεδομένα εισόδου αποτελούνται από συναλλαγές: κάθε συναλλαγή αποτελείται από αντικείμενα που αγοράζονται από έναν καταναλωτή κατά τη διάρκεια μιας μόνο επίσκεψης. Τα δεδομένα εξόδου αποτελούνται από κανόνες. Για παράδειγμα, ένας κανόνας μπορεί να είναι "το 90% των συναλλαγών που αφορούν την αγορά ψωμιού και βουτύρου περιλαμβάνουν επίσης γάλα" (Agrawal et al., 1993). Ακόμα και αν αυτή η μέθοδος έχει εισαχθεί στο πλαίσιο της Ανάλυσης Επιχειρήσεων αγοράς, έχει πολλές εφαρμογές σε άλλους τομείς, όπως η webmining ή η κειμενική. Μπορεί πραγματικά να χρησιμοποιηθεί για την αναζήτηση συχνών συνεννοήσεων σε κάθε μεγάλο σύνολο δεδομένων. Ένας κανόνας είναι ένας υπαινιγμός $A \rightarrow Γ$. Το αριστερό μέρος του κανόνα ονομάζεται προηγούμενο και το δεξί, το επακόλουθο. Τα σύνολα Α και Γ είναι αποσπασμένα, καθώς δεν μπορούμε να βρούμε το ίδιο στοιχείο τόσο στο προηγούμενο όσο και στο επακόλουθο. Ένας κανόνας έχει νόημα χάρη στην υποστήριξή του $s = \sup(A \rightarrow C) = P (A \cap C)$ και την αυτοπεποίθησή του $c = \text{conf}(A \rightarrow C) = P (C/A)$. Ο πρώτος αποτελεσματικός αλγόριθμος για τους κανόνες σύνδεσης εξορύξεων είναι Apriori (Agrawal και Srikant, 1994). Το πρώτο βήμα αυτού του αλγορίθμου είναι η έρευνα των συχνών itemsets. Ο χρήστης δίνει ένα ελάχιστο όριο για την υποστήριξη και ο αλγόριθμος αναζητά όλα τα itemsets που εμφανίζονται με υποστήριξη μεγαλύτερη από αυτό το όριο. Το δεύτερο βήμα είναι η δημιουργία κανόνων από itemsets που βρέθηκαν στο πρώτο βήμα. Ο αλγόριθμος υπολογίζει την εμπιστοσύνη κάθε κανόνα και διατηρεί μόνο εκείνους όπου η εμπιστοσύνη είναι μεγαλύτερη από ένα όριο που ορίζεται από τον χρήστη.

Ένα από τα κύρια προβλήματα είναι ο καθορισμός ορίων στήριξης και εμπιστοσύνης. Το Apriori βασίζεται στην ιδιότητα ότι κάθε υποσύνολο ενός συνόλου συχνών στοιχείων είναι επίσης συχνό. Τα υποψήφια k-itemsets δημιουργήσουν στην ανάγνωση kth του συνόλου δεδομένων και τα στηρίζουν τους υπολογίζονται στην k + 1η ανάγνωση. Εάν το K είναι το μεγαλύτερο μέγεθος συχνών στοιχείων, ο συνολικός αριθμός αναγνώσεων είναι K + 1. Άλλοι αλγόριθμοι έχουν προταθεί για τη μείωση του αριθμού των αναγνώσεων της βάσης δεδομένων και για τη βελτίωση της υπολογιστικής απόδοσης. Μεταξύ αυτών, μπορούμε να παραθέσουμε eclat (Zaki, 2000), Διαμέρισμα (Savasere et al., 1995), Δειγματοληψία (Toivonen, 1996), DIC (Brin et al., 1997a) ή FP-Ανάπτυξη (Han et al., 2000). Όλοι αυτοί οι αλγόριθμοι παρέχουν τα ίδια αποτελέσματα με την αναζήτηση κανόνων είναι ντετερμινιστική.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## 2.2.3 Apriori Algorithm

Ο αλγόριθμος Apriori αναφέρεται στον αλγόριθμο που χρησιμοποιείται για τον υπολογισμό των κανόνων συσχέτισης μεταξύ αντικειμένων. Σημαίνει πώς δύο ή περισσότερα αντικείμενα συνδέονται μεταξύ τους. Με άλλα λόγια, μπορούμε να πούμε ότι ο αλγόριθμος Apriori είναι ένας κανόνας συσχέτισης που αναλύει ότι οι άνθρωποι που αγόρασαν το προϊόν Α αγόρασαν επίσης το προϊόν Β. Ο πρωταρχικός στόχος του αλγορίθμου Apriori είναι να δημιουργήσει τον κανόνα συσχέτισης μεταξύ διαφορετικών αντικειμένων. Ο κανόνας συσχέτισης περιγράφει πώς σχετίζονται δύο ή περισσότερα αντικείμενα μεταξύ τους. Ο αλγόριθμος Apriori ονομάζεται επίσης συχνή εξόρυξη προτύπων. Γενικά, ο αλγόριθμος Apriori λειτουργεί σε μια βάση δεδομένων που αποτελείται από έναν τεράστιο αριθμό συναλλαγών. Ας κατανοήσουμε τον αλγόριθμο Apriori με τη βοήθεια ενός παραδείγματος. Ας υποθέσουμε ότι πηγαίνει κάποιος στο Big Bazar και αγοράζει διαφορετικά προϊόντα. Βοηθά τους πελάτες να αγοράσουν τα προϊόντα τους με ευκολία και αυξάνει την απόδοση πωλήσεων του Big Bazar.

Τα δεδομένα τρία στοιχεία αποτελούν τον αλγόριθμο Apriori.

1. Support
2. Confidence
3. Lift

Η υποστήριξη αναφέρεται στην προεπιλεγμένη δημοτικότητα οποιουδήποτε προϊόντος. Η υποστήριξη (support) υπολογίζεται ως το πηλίκο της διαίρεσης του αριθμού των συναλλαγών που περιλαμβάνει αυτό το προϊόν με τον συνολικό αριθμό συναλλαγών. Η εμπιστοσύνη (confidence) αναφέρεται στην πιθανότητα οι πελάτες να αγόρασαν μαζί μπισκότα και σοκολάτες. Επομένως, πρέπει να διαιρέσετε τον αριθμό των συναλλαγών που περιλαμβάνουν μπισκότα και σοκολάτες με τον συνολικό αριθμό συναλλαγών για να έχετε εμπιστοσύνη. Το lift αναφέρεται στην αύξηση της αναλογίας των πωλήσεων.

Η διαδικασία σύμφωνα με τους Agrawal και Srikant (1994) που ακολουθεί ο αλγόριθμος Apriori:

- πρώτα γίνεται ο υπολογισμός της τιμής support από όλα τα αντικείμενα στην βάση δεδομένων.
- έπειτα πραγματοποιείται η δημιουργία ενός συνόλου itemsets το οποίο περιέχει τα itemsets που είναι ισότιμα support ή μεγαλύτερα από την ελάχιστη τιμή support που ορίζεται. Το σύνολο αυτό ορίζεται ως Lk, όπου k=1 διότι κάθε itemset περιέχει ένα αντικείμενο.
- έπειτα τα itemsets του Lk, δημιουργούν ένα σύνολο itemsets που περιέχουν k+1 αντικείμενα. Το υποψήφιο αυτό σύνολο συμβολίζεται Ck+1, όπου για τη δημιουργία του συνενώνεται το Lk με τον εαυτό του.
- στην περίπτωση ενός έστω μη συχνού συνόλου αντικειμένων k μέσα στο Ck+1, γίνεται διαγραφή.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

- πραγματοποιείται εκ νέου ο υπολογισμός της τιμής support για όλα τα μέλη του Ck+1 και εάν κάποιο από αυτά είναι μικρότερο σε τιμή support από την ελάχιστη τιμή που ορίζεται, τότε γίνεται διαγραφή και ολοκλήρωση της διαδικασίας για να βρεθεί το Lk+1.
- τέλος γίνεται επανάληψη της διαδικασίας από το βήμα 3 εάν το Lk+1 δεν είναι κενό.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

# 3

# Σχετικές Εργασίες

## 3.1 Clustering - RFM

Πολλοί ερευνητές διαπιστώνεται μέσα από την βιβλιογραφία ότι ασχολούνται με το clustering των πελατών τόσο τη θεωρία όσο και στην ανάλυση των συστάδων. Το συγκεκριμένο θέμα έχει προσεγγιστεί από διάφορους, αλλά οι πιο πολλοί ασχολούνται με την ανάλυση RFM όπως περιεγράφηκε παραπάνω αλλά και με κάποιες μεταβολές καθώς και τον αλγόριθμο K-means συνδυασμένο και με άλλους αλγορίθμους. Σε έρευνα του Stone (1995) διαπιστώθηκε ότι οι τρεις μεταβλητές της ανάλυσης RFM διαφέρουν ως προς τη σημασία τους εξαιτίας των χαρακτηριστικών της αγοράς. Στις εποχιακές περιόδους κατά συνέπεια δημιουργούνται clusters ως βάση για ένα προσαρμοσμένο σταθμισμένο RFM μοντέλο, που ουσιαστικά αποτελεί ένα CLV μοντέλο.

Από την άλλη σε εργασία των Chen et al. (2009) αποδείχτηκε ότι οι πιο πολλοί από τους αλγόριθμους εξόρυξης διαδοχικών προτύπων μέχρι τότε επικεντρώνονταν στην συχνότητα ως σημαντικό κριτήριο για την επιλογή, σε αυτό το σημείο αξίζει να αναφερθεί ότι το ίδιο βασικό κριτήριο αποτελεί και η νομισματική αξία. Ανέλυσαν λοιπόν αυτά τα δύο κριτήρια (frequency, monetary) μαζί με τις πρόσφατες συναλλαγές (recency). Αλλάζοντας τον αλγόριθμο Apriori έφτιαξαν ένα πιο αποτελεσματικό αλγόριθμο, τον γνωστό RFM-Apriori. Βέβαια οι μεταβλητές recency, frequency και monetary εξετάστηκαν στο παρελθόν από πολλούς ερευνητές, οι συγκεκριμένοι όμως για πρώτη φορά εφάρμοσαν την ανάλυση RFM στο SPM (Sequential pattern mining). Αντί της κατηγοριοποίησης των πελατών ανάλογα με τις ανάγκες τους, για να ληφθούν αποφάσεις στο μάρκετινγκ και στην Διαχείριση Πελατειακών Σχέσεων (CRM), χρησιμοποιείται πλέον η κατηγοριοποίηση των πελατών ανάλογα με την Αξία Διάρκεια Ζωή του πελάτη (CLV) και υπολογίζεται η αξία του κάθε cluster.

Σε οδηγό των Khajvand & Tarokh (2011) κατέγραψαν στρατηγικές μάρκετινγκ για τη δημιουργία clusters διαφορετικών εποχιακών περιόδων. Επίσης σε εργασία των Abirami & Pattabiraman (2016) συνδυάστηκε η ανάλυση RFM με τις τεχνικές εξόρυξης δεδομένων για δεδομένα που είχαν σχέση με πελάτες λιανικού εμπορίου. Ειδικότερα κατασκεύασαν clusters με την ανάλυση RFM και τον αλγόριθμο K-means και στη συνέχεια χρησιμοποίησαν τους κανόνες συσχέτισης, για να βρεθούν μοτίβα από τα προϊόντα που υπήρχαν στις συναλλαγές των πελατών. Το clustering και οι συσχετίσεις συγκρίθηκαν με



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

σκοπό τα αποτελέσματα τους να συνδεθούν με τις κατάλληλες προσφορές. Οι Doğan et al. (2018), υποστηρίζουν ότι η βέλτιστη κατηγοριοποίηση των πελατών είναι ιδιαίτερα σημαντική για τις εταιρείες λιανικής, διότι η κατηγοριοποίηση πελατών που έχουν συναφείς ανάγκες, επιθυμίες και συμπεριφορές προσφέρουν την δυνατότητα στις εταιρείες να αντιληφθούν καλύτερα την αγορά-στόχο (target market). Η βελτιωμένη κατηγοριοποίηση συμπεριλαμβάνει τη σύγκριση μοντέλων, με σκοπό να δίνεται άριστη εικόνα για τους πελάτες. Με τη χρήση των δύο μοντέλων για την Ανάλυση RFM και η σύγκριση τους αλλά και η ήδη υπάρχουσα κατηγοριοποίηση της εταιρείας, αποδείχτηκε ότι τα αποτελέσματα της καλύτερης κατηγοριοποίησης που προκύπτουν δεν συμπίπτουν με αυτή που είχε κάνει ήδη η εταιρεία. Κατά συνέπεια η εταιρεία κατηγοριοποίησε τους πελάτες με τέτοιο τρόπο ώστε κάποιοι από αυτούς που είχαν εκδώσει golden loyalty card αποδείχθηκε ότι ήταν bronze καταναλωτές. Στο πρώτο μοντέλο εφαρμόστηκε η μέθοδος log-likelihood και το κριτήριο Schwarz's Bayesian, στο δεύτερο μοντέλο χρησιμοποιήθηκε ο αλγόριθμος K-means.

Οι Anitha & Patil (2019) συνδύασαν την ανάλυση RFM με τον αλγόριθμο K-means στο κλάδο του λιανικού εμπορίου, με σκοπό την εύρεση των clusters. Ο αλγόριθμος K-means εφαρμόστηκε 2 φορές για την ανάλυση του ποσού που πάρθηκε από τις πρόσφατες συναλλαγές και από τις συχνές συναλλαγές. Με αυτό τον τρόπο αρχικά έγινε η χρήση αλγόριθμου για τον διαχωρισμό των πελατών με βάση το ποσό που δημιουργήθηκε από τις πρόσφατες συναλλαγές και στην συνέχεια έγινε η χρήση του για την κατηγοριοποίηση των πελατών με βάση το ποσό που δημιουργήθηκε από τις συχνές συναλλαγές. Μετά πραγματοποιήθηκε ο υπολογισμός Silhouette Score για τη σύγκριση clusters που δημιουργήθηκαν από τις δύο εφαρμογές του αλγορίθμου K-means και επιλέχθηκαν τα clusters με το καλύτερο Silhouette Score. Στις διαδικτυακές αγορές οι ανάγκες και οι επιθυμίες των καταναλωτών μεταβάλλονται συνεχώς. Άρα κρίνεται επιτακτική η ανάγκη δημιουργίας εξαιρετικών στρατηγικών μάρκετινγκ ανάλογα με τους πελάτες, προβλέποντας τις διαδικτυακές συμπεριφορές με βάση την εξόρυξη δεδομένων. Σύμφωνα με τους Lin et.al (2020) που τμηματοποίησαν τους πελάτες, πρότειναν στρατηγικές μάρκετινγκ, με το συνδυασμό του μοντέλου RFM και της μεθόδου K-means σε ένα ηλεκτρονικό κατάστημα στη Κίνα για σχεδόν 2 χρόνια. Έκαναν χρήση της μεθόδου PCA, για να αποφασιστούν οι συντελεστές βαρύτητας των μεταβλητών RFM. Εφαρμόζοντας τις προτάσεις, προχώρησαν στην εξέταση των βασικών δεικτών απόδοσης της εταιρείας, διαπιστώνοντας τη προσθήκη 529 πελατών, με 279% αύξηση του όγκου των αγορών και 101.97% στο ποσό κατανάλωσης.

## 3.2 Ανάλυση Καλαθιού Αγοράς (MBA)

Η ανάλυση καλαθιού αγοράς (MBA) είναι μια τεχνική εξόρυξης δεδομένων που χρησιμοποιείται κυρίως στη βιομηχανία καταναλωτικών συσκευασμένων αγαθών (CPG) για να καθορίσει ποια προϊόντα επιλέγονται μαζί και, πώς η αγορά ενός στοιχείου επηρεάζει την πιθανότητα αγοράς ενός άλλου στοιχείου (Qualls, 2013). Το MBA είναι μόνο ένα από τα πολλά εργαλεία που υπάρχουν για να παρακολουθεί τις συνεχείς αλλαγές στη γεύση των πελατών, καθώς ποικίλλει γρήγορα

με την χρονική ταλάντευση (InfoCepts, 2004) υπάρχουν νέες συσκευές παρακολούθησης που εισάγονται ή γυαλιά που προσφέρονται σε εθελοντές αγοραστές και πολλές άλλες τεχνικές που είναι τα αποτελέσματα των τεχνολογικών καινοτομιών. Η δυνατότητα πρόβλεψης των συμπεριφορών αγοράς ενός πελάτη έχει γίνει τόσο ζωτικής σημασίας για τη λιανική επιχείρηση που εφαρμογές λογισμικού, (συστήματα), όπως salesForce, Microsoft Great Plains και λογιστικές εφαρμογές CRM όπως το Quickbooks, έχουν όλα προωθήσει το γεγονός ότι το προϊόν τους έχει τη δυνατότητα να κάνει επεξεργασία MBA (Gupta, 2013). Το θέμα αυτό είναι μάλλον πρακτικό και όχι ακαδημαϊκό, επομένως οι περισσότερες από τις μελέτες επί του θέματος έχουν ασκηθεί σε πραγματικά καταστήματα λιανικής πώλησης.

Η ανάλυση καλαθιού αγοράς προϊόντων ως παλιό πεδίο στην εξόρυξη δεδομένων είναι επίσης ένα από τα καλύτερα παραδείγματα κανόνων της ένωσης εξόρυξης (Gancheva, 2013). Οι Rakesh Agrawal και Usama Fayyad ως πρωτοπόροι στην εξόρυξη δεδομένων, την εξόρυξη κανόνων σύνδεσης (ARM) και τη ομαδοποίηση έχουν αναπτύξει διαφορετικούς αλγορίθμους για να βοηθήσουν τους χρήστες να επιτύχουν τους στόχους τους. Η ανάλυση καλαθιού αγοράς μπορεί να ταξινομηθεί σε δύο κατηγορίες, δηλαδή ποιοτικές και ποσοτικές, οι οποίες θα συζητηθούν εν συντομία. Οι ποσοτικοί κανόνες καθορίζονται από την άποψη του τύπου των χαρακτηριστικών που περιέχονται σε αυτούς τους σχεσινομικούς πίνακες. Τα χαρακτηριστικά μπορεί να είναι είτε ποσοτικά (μισθός, βάρος κ.λπ.) είτε κατηγορηματικά (τύπος προϊόντος, φύλο κ.λπ.). Τα χαρακτηριστικά Boolean είναι τέτοια χαρακτηριστικά που μπορούν να αναλάβουν μία από τις δύο επιλογές. Θεωρούνται ειδική περίπτωση κατηγορηματικών χαρακτηριστικών (Prakas – Parvati, 2011). Προτείνει μια ποιοτική προσέγγιση για τον κανόνα ποσοτικής συσχέτισης εξόρυξης και η προσέγγιση είναι ποιοτική, καθώς μετατρέπει αριθμητικά χαρακτηριστικά σε δυαδικά και ο τύπος που χρησιμοποιείται είναι κατηγορηματικός. Στις περισσότερες περιπτώσεις, οι κανόνες αυτοί εμφανίζονται ως δέντρα αποφάσεων, μοτίβα ή πίνακες εξάρτησης (Shapiro & Frawley, 1991). Οι Agrawal & Imielinski (1993) πρότειναν αυτόν τον αλγόριθμο εξόρυξης κανόνων που ανακαλύπτει ποιοτικούς κανόνες χωρίς περιορισμούς για τα χαρακτηριστικά Boolean. Τα δεδομένα που συλλέγονται από ένα μεγάλο κατάστημα λιανικής πώλησης χρησιμοποιούνται για τη δοκιμή της επίδρασης του αλγορίθμου και το χαρτί δείχνει τον εφαρμοσμένο αλγόριθμο. Η ανάλυση καλαθιού αγοράς έχει ταξινομηθεί σε δύο μοντέλα επεξηγηματικών και διερευνητικών σύμφωνα με τη βιβλιογραφία μάρκετινγκ.

Τα διερευνητικά μοντέλα χρησιμοποιούν δεδομένα POS (σημείο πώλησης) για τον προσδιορισμό των προτύπων αγορών, ωστόσο αυτό το μοντέλο δεν περιλαμβάνει πληροφορίες σχετικά με τα δημογραφικά στοιχεία των καταναλωτών ή τις μεταβλητές συνδυασμού μάρκετινγκ (Dipold - Khrushchev, 2010). Μέθοδοι όπως οι κανόνες συσχέτισης (Agrawal & Ramakrishnan, 1994) ή το συνεργατικό φιλτράρισμα (Mild & Reutterer, 2003) συνοψίζουν έναν τεράστιο όγκο δεδομένων σε λιγότερους σημαντικούς κανόνες ή μέτρα. Αυτές οι μέθοδοι θα μπορούσαν να είναι χρήσιμες για την ανακάλυψη ασυνήθιστων και απρόβλεπτων σχέσεων μεταξύ των θεμάτων των δεδομένων, επίσης είναι

εύκολο να εργαστούν με τη βοήθεια διαφορών λογισμικού. Αλλά το σημαντικό θέμα που πρέπει να λαμβάνεται υπόψη είναι ότι οι διερευνητικές προσεγγίσεις δεν είναι κατάλληλες για την πρόβλεψη και την εύρεση των αιτίων-ριζών σύνθετων προβλημάτων και χρησιμοποιούνται αποκλειστικά για την εύρεση αλληλεξαρτώμενων μεταξύ κατηγοριών με βάση τα πρότυπα των τύπων προϊόντων που αγοράζονται μαζί (Gancheva, 2013). Με απλά μέτρα συσχέτισης μια τυπική εφαρμογή αυτών των διερευνητικών προσεγγίσεων είναι ο προσδιορισμός των σχέσεων κατηγορίας προϊόντων. Η ίδια υπόθεση αφορά ενώσεις που χρησιμοποιούνται για τη σύγκριση οντοτήτων ανά ζεύγη και κρίνουν ποια οντότητα προτιμάται ή έχει μεγαλύτερο ποσό ποσοτικής περιουσίας, (Julander, 1992) συγκρίνει το ποσοστό των πελατών που αγοράζουν ένα προϊόν και το ποσοστό όλων των συνολικών πωλήσεων που παράγονται από το ίδιο προϊόν. Αυτές οι συγκρίσεις μπορούν να οδηγήσουν στην εξεύρεση των περαιτέρω προϊόντων και του μεριδίου τους στις πωλήσεις. Αυτές οι γνώσεις θα είναι χρήσιμες όχι μόνο στη στρατηγική, την προώθηση αλλά και την κυκλοφορία στο κατάστημα, καθώς είναι ζωτικής σημασίας να χρησιμοποιηθούν αυτές οι πληροφορίες για την τοποθέτηση άλλων συγκεκριμένων προϊόντων κοντά, επίσης το απόθεμα μπορεί να είναι πιο αποτελεσματικό τόσο από την άποψη των παραγγελιών όσο και από την άποψη του κόστους.

Το έγγραφο του Julander δείχνει πώς θα μπορούσε να χρησιμοποιηθεί συνδυαστική ανάλυση για τη μελέτη των προτύπων διασταυρούμενης αγοράς μεταξύ ορισμένων εμπορικών σημάτων ή τύπων προϊόντων. Ένας άλλος σημαντικός τομέας ενδιαφέροντος για την έρευνα, στον τομέα της διερευνητικής ανάλυσης είναι η διαδικασία δημιουργίας κανόνων σύνδεσης, ενώ έχουν εισαχθεί πολλά είδη αλγορίθμων για τη διάκριση των προτύπων εξόρυξης στην ανάλυση καλαθιού της αγοράς. Επίσης, εισάγεται άλλη χρήση των δεδομένων καλαθιού αγοράς στο πεπερασμένο μοντέλο μείγματος (Andrews & Currin, 2002), η ιδέα του οποίου είναι να εντοπιστούν τμήματα νοικοκυριών που έχουν πανομοιότυπη συμπεριφορά μεταξύ των επιλεγμένων προϊόντων και να απαντηθούν οι ερευνητικές ερωτήσεις. Το μοντέλο προτείνει τόσο μεταβλητές μάρκετινγκ όσο και δεδομένα πάνελ σαρωτή. Τα διερευνητικά μοντέλα δεν μπορούν να χρησιμοποιηθούν για να εξηγήσουν τη βασική αιτία των συν-περιστατικών μεταξύ των αντικειμένων, αλλά για να τα ανακαλύψουν αποκλειστικά, οπότε το κύριο καθήκον της διερευνητικής ανάλυσης καλαθιού αγοράς είναι να αποκαλύψει και να παρουσιάσει κρυφές σχέσεις μεταξύ των κατηγοριών προϊόντων. Ωστόσο, τα επεξηγηματικά μοντέλα αποσκοπούν στην εξήγηση αυτών των ευρημάτων. Τα επεξηγηματικά μοντέλα βοηθούν στον εντοπισμό και τον ποσοτικό προσδιορισμό των επιρροών μεταξύ κατηγοριών στις μεταβλητές μάρκετινγκ, όπως η τιμή, η προώθηση και άλλα χαρακτηριστικά μάρκετινγκ και κυρίως αυτά τα μοντέλα βασίζονται στην ανάλυση παλινδρόμησης και στο πολυεπίπεδο μοντέλο εφοδιαστικής (Mild &Reutterer, 2003).

Η χρήση δεδομένων συναλλαγών εξόρυξης και η συμμετοχή των πελατών βοηθά τους διαχειριστές σε ενέργειες προσαρμογής και στόχευσης μάρκετινγκ που είναι ζωτικής σημασίας στο ανταγωνιστικό στάδιο διαφόρων λιανοπωλητών που είναι διαθέσιμοι για τους ίδιους πελάτες. Η μελέτη και η ανάλυση προηγούμενων αλληλεπιδράσεων επιτρέπει στους διαχειριστές να κάνουν καλύτερη

τιμολόγηση με νοημοσύνη, ενώ εξυπηρετούν ετερογενείς καταναλωτές (Kumar & Rao, 2006). Οι πληροφορίες προσανατολισμένες στον καταναλωτή προέρχονται από καλάθια αγορών που χρησιμοποιούν οι λιανοπωλητές αντί να παρακολουθούν μεμονωμένα άρθρα, ενώ ταυτόχρονα οι λιανοπωλητές χρησιμοποιούν δεδομένα σαρωτή για να ανακαλύψουν τις επιπτώσεις των διαφημιστικών στρατηγικών στη συμπεριφορά των καταναλωτών τους. Δεδομένου ότι η συμπεριφορά αγοράς των καταναλωτών είναι ένας γνωστός τομέας στη βιβλιογραφία μάρκετινγκ μαζί με την ευαισθησία και την ελαστικότητα των τιμών είναι μέσω εφαρμοσμένων τεχνικών εξόρυξης δεδομένων, ενώ είναι δεδομένο ότι οι πελάτες χωρίζονται συνήθως σε αγοραστές μεγάλων καλαθιών και αγοραστές μικρών καλαθιών. Ενώ οι αγοραστές μεγάλων καλαθιών έχουν υψηλότερη αναμενόμενη ελκυστικότητα καλαθιού σε καθημερινά καταστήματα χαμηλών τιμών (όπως η Walmart), οι αγοραστές μικρών καλαθιών θα προτιμούσαν να πάνε για λιανοπωλητές υψηλής χαμηλής τιμής (Bell & Boztu, 2007). Για την πρόβλεψη της επιλογής μάρκας χρησιμοποιούνται κυρίως μοντέλα Logit και probit (Russell & Petersen, 2000) τα οποία μπορούν να βοηθήσουν σε καταστήματα λιανικής πώλησης που στοχεύουν σε οικογένειες της ανώτερης μεσαίας τάξης και στην πώληση επώνυμων ή πολυτελών αγαθών.

Η χρήση μοντέλων binominal logit επιτρέπει την εξέταση των επιπτώσεων αλληλεπίδρασης μεταξύ διαφόρων κατηγοριών προϊόντων με σχέσεις αλληλεξάρτησης που θα μπορούσαν να χρησιμοποιηθούν για την πρόβλεψη του σχεδιασμού στρατηγικής προώθησης, ενώ οι περισσότερες από τις άλλες προσεγγίσεις βρίσκουν αυτές τις συναντήσεις αποκλειστικές. Στο έγγραφο των Dippold & Hruschka, (2010) το πολυπολικό μοντέλο logit εισάγεται για τη μέτρηση των εξαρτήσεων και των προωθητικών αποτελεσμάτων σε διάφορους τύπους προϊόντων σε ένα κατάστημα λιανικής πώλησης και πώς αυτή η εξάρτηση μπορεί να χειριστεί τις πιθανότητες αγοράς. Η σχέση μεταξύ των δημογραφικών μεταβλητών των νοικοκυριών και της ευαισθησίας στις τιμές ήταν μετρήσιμη με τη βοήθεια δεδομένων σαρωτή και άλλων μεταβλητών ως μέρος της ευαισθησίας προς το μείγμα μάρκετινγκ και (Ainslie & Rossi, 1998) μετρά τη συγκατοίκηση της παρατηρούμενης και μη παρατηρούμενης ετερογένειας στην ευαισθησία του μίγματος μάρκετινγκ σε διάφορες κατηγορίες. Στην μελέτη των Russell & Petersen, (2000) εξετάζουν τη διαδικασία επιλογής εμπορικών σημάτων σε τέσσερις κατηγορίες αγαθών. Ενώ η επιλογή επωνυμίας μεταξύ άλλων κατηγοριών μπορεί να υπολογιστεί απλά με τη χρήση ενός τύπου πιθανότητας υπό όρους, από την άλλη πλευρά το επίπεδο πολυπλοκότητας μπορεί να αυξηθεί δραστικά με την προσθήκη κατηγοριών υπό μελέτη. Εξηγώντας τα πλεονεκτήματα της διερευνητικής ανάλυσης αυτή η μελέτη κινείται σε μειονεκτήματα και αδικήματα.

Παρά τη χρήση της ανακάλυψης αλληλεξαρτώμενων μεταξύ κατηγοριών, για διαφορετικά μέσα η διαχειριστική αξία αυτών των μοντέλων είναι μάλλον περιορισμένη. Δεδομένου ότι μπορεί να παράσχει μόνο περιορισμένο αριθμό ιδεών στη στρατηγική λήψης αποφάσεων και προώθησης και δεν υπάρχουν παραδοχές Apriori σχετικά με την «απάντηση» και το «αποτέλεσμα» και δεν ενσωματώνονται μεταβλητές μάρκετινγκ στην ανάλυση. Επίσης, η παράλειψη της ανομοιογένειας των καταναλωτών και του μίγματος μάρκετινγκ μπορεί να επηρεάσει τον σχεδιασμό και ακόμη και να

οδηγήσει σε προκαταλήψεις. Για να ξεπεραστούν οι περιορισμοί τόσο των επεξηγηματικών όσο και των διερευνητικών προσεγγίσεων (Boztug & Reutterer, 2006) προτείνεται ένα μοντέλο που συνδέει πρώτα τόσο τη συμπίεση δεδομένων όσο και την εκτίμηση των αποτελεσμάτων αγοράς μεταξύ κατηγοριών, έτσι ώστε η πολυπλοκότητα του μοντέλου να ελαχιστοποιείται επιλέγοντας μόνο ουσιαστικές κατηγορίες που σχετίζονται με ένα συγκεκριμένο τμήμα των νοικοκυριών. Για να επιλεγούν κατηγορίες που θα συμπεριληφθούν για την εκτίμηση των προϊόντων μεταξύ κατηγοριών, η διαδικασία Boztug & Reutterer που συνδυάζει τόσο διερευνητικά όσο και επεξηγηματικά μοντέλα μπορεί να χρησιμοποιηθεί ως κατευθυντήρια γραμμή. Με το συγκεχυμένο θέμα της ταξινόμησης των προϊόντων σε διάφορες κατηγορίες οι διάφοροι ερευνητές προσφέρουν διάφορους τύπους κατηγοριοποίησης, ενώ στη μελέτη των Berry & Linoff, (1997) προτείνεται μια απλή αλλά χρήσιμη ιδέα. Η προσέγγιση περιλαμβάνει όλα τα είδη αντικειμένων στις αντίστοιχες κατηγορίες τους. Τα συχνά στοιχεία δεν συγκεντρώνονται καθόλου, ενώ τα λιγότερο συχνά στοιχεία πρέπει να ωθηθούν σε υψηλότερο επίπεδο της ιεραρχίας. Ανάλογα με τις κατηγορίες τα ίδια τα αντικείμενα μπορούν να κινηθούν στην ταξινομία και όλα τα μεμονωμένα είδη ενός καταστήματος λιανικής πώλησης είναι στη συνέχεια το χαμηλότερο επίπεδο. Για παράδειγμα, υπάρχουν επτά διαφορετικοί τύποι iPhone. Όλα μπορούν να ομαδοποιηθούν σε μια κατηγορία «iPhone», η οποία είναι υποκατηγορία των «smartphones».

Ο όρος ταξινομία αναφέρεται στην ταξινόμηση διαφορετικών κατηγοριών με ιεραρχική μέθοδο. Στη συνήθη διαδικασία εξόρυξης κανόνων χρησιμοποιούνται τα δεδομένα σε επίπεδο συναλλαγής μιας ενιαίας αγοράς πελατών, ενώ, με τη σύμφωνη χρήση αυτού του στόχου, νέες έρευνες βρίσκονται σε συνολικό επίπεδο για τη διάκριση περισσότερων δεδομένων και αλληλεξαρτώμενων σε μεγαλύτερο πιο χρήσιμο επίπεδο που διεξάγονται σε περισσότερα δεδομένα καλαθιού της αγοράς. Τα στοιχεία αυτά συγκεντρώνονται κυρίως με τα μέτρα του χρόνου, έτσι ώστε να μην υπάρχει πλέον ατομική συναλλαγή, αλλά πηγαίνει τόσο μεγάλο όσο οι καθημερινές πωλήσεις ενός καταστήματος λιανικής πώλησης ή οποιωνδήποτε άλλων υπό μελέτη μονάδων. Είναι επίσης δυνατή η συλλογή δεδομένων σε επίπεδο συναλλαγών κατά περισσότερα από ένα χαρακτηριστικά που οδηγούν στο πρόβλημα της πολυδιάστασης που συζητείται λεπτομερώς στην μελέτη των Nestorov & Jukic, (2003) όπου υπάρχουν πληροφορίες για διάφορες διαστάσεις όπως το προϊόν, η τοποθεσία, ο πελάτης και το ημερολόγιο για κάθε συναλλαγή. Το συνηθισμένο ερώτημα μίας διάστασης για το ποια αντικείμενα αγοράζονται συχνά μαζί σε μια συναλλαγή, τώρα επεκτείνεται σε ποια προϊόντα αγοράζονται μαζί σε μια συγκεκριμένη περιοχή σε έναν συγκεκριμένο μήνα. Όταν εμπλέκονται πολλαπλές διαστάσεις, ορισμένες συσχετίσεις ενδέχεται να είναι κρυφές, οπότε προτείνεται ένα νέο μοντέλο που καταγράφει αυτές τις διαστάσεις. Τα πλεονεκτήματα της έννοιας της εκτεταμένης συσχέτισης όσον αφορά τους κανόνες που παράγονται είναι εύκολο να εξηγηθούν, παρέχοντας ακριβέστερες προβλέψεις για ορισμένες μεταβλητές και ο αριθμός των κανόνων που ανακαλύφθηκαν τείνει να είναι μικρότερος για την ίδια στήριξη. Διάφοροι συγγραφείς αφιέρωσαν τις εργασίες τους στη σύγκριση διαφορετικών

τεχνικών εξόρυξης. Ένα παράδειγμα θα ήταν η πρόσφατη έρευνα των Khan, Lee, Lee, & Khattak, (2010). Η έρευνα τους είναι μια συγκριτική ανάλυση δύο τεχνικών εξόρυξης δεδομένων: ARM (εξόρυξη κανόνα συσχέτισης) και Ομαδοποίηση. Χρήσιμες πληροφορίες εξάγονται από τα δεδομένα συναλλαγών ενός σούπερ-μάρκετ που ονομάζεται ημέρα πωλήσεων.

Ο αλγόριθμος Apriori χρησιμοποιείται για την εξόρυξη κανόνων συσχέτισης. Με την εύρεση συναφών προϊόντων και την αναδιάταξη των ραφιών προς όφελός τους, οι πιθανότητες αύξησης του όγκου πώλησης είναι υψηλότερες. Ωστόσο, στην ταξινόμηση, η ομαδοποίηση είναι πολλές φορές πιο προτιμώμενη για χρήση. Το έγγραφο Khan εφαρμόζει ομαδοποίηση K- που σημαίνει ομαδοποίηση για την ταξινόμηση διαφορετικών τύπων προϊόντων σε ένα καλάθι και τους πελάτες ανάλογα με τη συμπεριφορά και την αγοραστική τους δύναμη. Με τη σωστή χρήση της τεχνικής ομαδοποίησης οι λιανοπωλητές θα μπορούσαν να βελτιώσουν τις στρατηγικές προώθησης τους με αποτέλεσμα την ενίσχυση της πώλησης. Η τεχνική ομαδοποίησης εξάγει δεδομένα όπως το προφίλ των πελατών, όπως η ηλικία, η αγοραστική δύναμη, καθώς και η επισκεψιμότητα των πελατών.

# 4

# Γνωριμία με τα δεδομένα

Στο κεφάλαιο αυτό θα γνωρίσουμε καλύτερα τόσο το κατάστημα όσο και τα δεδομένα που μας έχει δώσει. Βασικό στάδιο πριν ξεκινήσουμε την ανάλυση είναι να έχουμε κατανοήσει πλήρως τα δεδομένα και να τα έχουμε επεξεργαστεί κατάλληλα ώστε να διευκολύνεται η διαδικασία.

## 4.1 Η Επιχείρηση και τα Δεδομένα

Η επιχείρηση από την οποία αντλήσαμε τα δεδομένα δραστηριοποιείται στον τομέα της λιανικής πώλησης ειδών ένδυσης και υπόδησης και αξεσουάρ σε όλη την Ελλάδα και πιο συγκεκριμένα στην πόλη της Λάρισας όπου είναι και η έδρα της. Διατηρεί φυσική παρουσία στην πόλη της Λάρισας με τρία καταστήματα και φυσικά εξυπηρετεί και όλη την Ελλάδα μέσω του ηλεκτρονικού της καταστήματος. Το καταναλωτικό κοινό στο οποίο απευθύνεται είναι άνδρες και γυναίκες όλων των ηλικιών.

Το αρχείο που μας δόθηκε από την επιχείρηση είχε όλα τα παραστατικά που εκδόθηκαν στο διάστημα που αναλύουμε αλλά και τα master data των ειδών του κάθε παραστατικού. Δηλαδή τον κωδικό παραγγελίας OrderID, τον αριθμό κάθε πελάτη CustomerID, το όνομα κάθε πελάτη Customer_Name, την ημερομηνία κατά την οποία πραγματοποιήθηκε η αγορά Purchase_Date, τον μοναδικό αριθμό του προϊόντος κάθε αγοράς ProductID, την περιγραφή του προϊόντος Product, την ποσότητα κάθε προϊόντος Quantity, την συνολική αξία κάθε προϊόντος Total, τους τομείς που ανήκει η κάθε κατηγορία Sector, την κατηγορία στην οποία ανήκει κάθε προϊόν Category, την μάρκα κάθε προϊόντος Brand, το φύλλο κάθε πελάτη Sex, όπως και τον τρόπο με τον οποίο πραγματοποιήθηκε η κάθε πληρωμή Payment.

## 4.2 Προ-επεξεργασία Δεδομένων

Πριν προχωρήσουμε στην ανάλυση είναι πολύ σημαντικό να γίνει η προ-επεξεργασία των δεδομένων. Η διαδικασία αυτή έχει ως στόχο τον εντοπισμό κενών και ακραίων τιμών ή ακόμη και τιμών οι οποίες μπορεί να αποτελέσουν εμπόδιο στην ανάλυση. Οι τιμές αυτές τροποποιούνται κατάλληλα ή ακόμη σε αρκετές περιπτώσεις εξαιρούνται.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

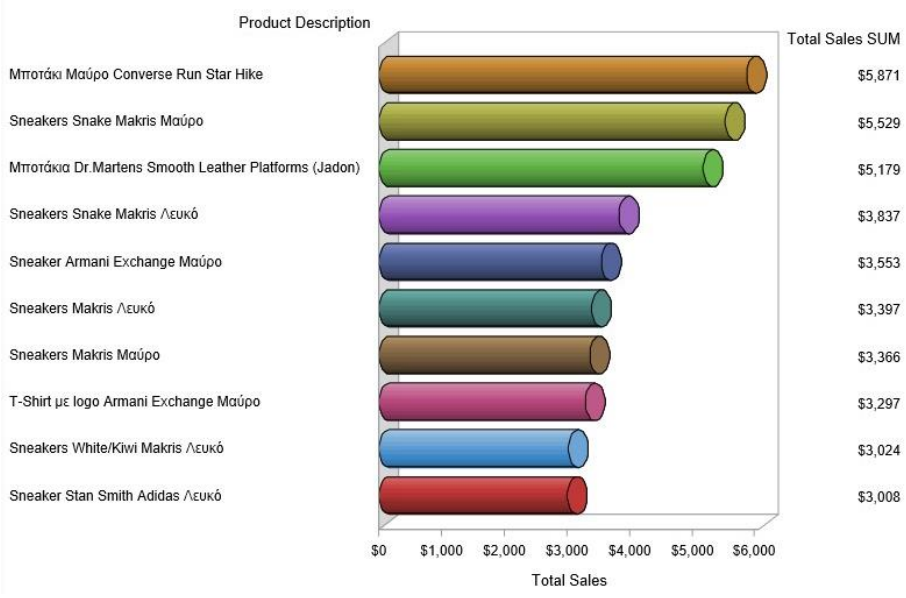
Τα δεδομένα που αντλήθηκαν από την βάση δεδομένων της επιχείρησης και χρειάστηκαν αρκετή επεξεργασία ώστε να μπορέσουμε να τα χρησιμοποιήσουμε. Παρατηρήθηκε έναν αριθμός πελατών που ήταν ανώνυμοι καθώς στο πεδίο Customer_Name η περιγραφή ήταν Πελάτης Λιανικής όπως επίσης και ένας ο αριθμός προϊόντων που στην την περιγραφή Product είχαν την ονομασία ΔΙΑΦΟΡΑ ΕΙΔΗ. Αν και οι τιμές αυτές θα μπορούσαν να εξαιρεθούν αποφασίσαμε να διατηρήσουμε τα δεδομένα αυτά στην ανάλυσή μας και στις ενότητες RFM και Market Basket Analysis αναφέρουμε κάθε φορά τον ακριβή αριθμό παρατηρήσεων που χρησιμοποιούμε.

## 4.3 Κατανόηση Δεδομένων

Πριν προχωρήσουμε στην ανάλυση των δεδομένων είναι πολύ σημαντικό να τα έχουμε κατανοήσει πλήρως.

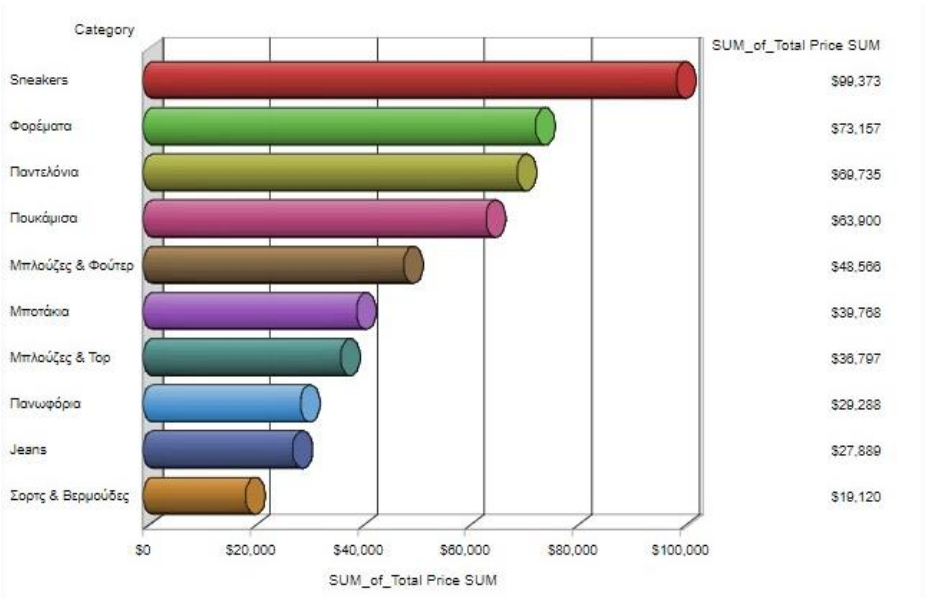
Τα δεδομένα που μας δόθηκαν ύστερα από την απαραίτητη προ-επεξεργασία αφορούν διάστημα 6 μηνών και πιο συγκεκριμένα από 1η Απριλίου του 2021 έως 31η Οκτωβρίου του έτους 2021. Το dataset αποτελείται από 13 στήλες και 10.552 σειρές. Πιο αναλυτικά για το διάστημα των 6 μηνών έχουμε 3.706 μοναδικούς πελάτες, 6.957 μοναδικές παραγγελίες, 2.883 μοναδικά προϊόντα και 50 κατηγορίες προϊόντων.

Στο Γράφημα 3 εμφανίζονται τα 15 προϊόντα με τις μεγαλύτερες πωλήσεις στην χρονική περίοδο των 7 μηνών που εξετάζουμε. Συγκεκριμένα, το «Μποτάκι Μαύρο Converse Run Star Hike», το «Sneakers Snake Makris Μαύρο» και το «Μποτάκια Dr.Martens Smooth Leather Platforms (Jadon)» είναι τα προϊόντα τα οποία έχουν συγκεντρώσει τα μεγαλύτερα χρηματικά ποσά. Πρέπει να αναφέρουμε ότι δεν συμπεριλάβαμε τα προϊόντα με την περιγραφή «Διάφορα Είδη» για να έχουμε μια καλύτερη εικόνα των δεδομένων.



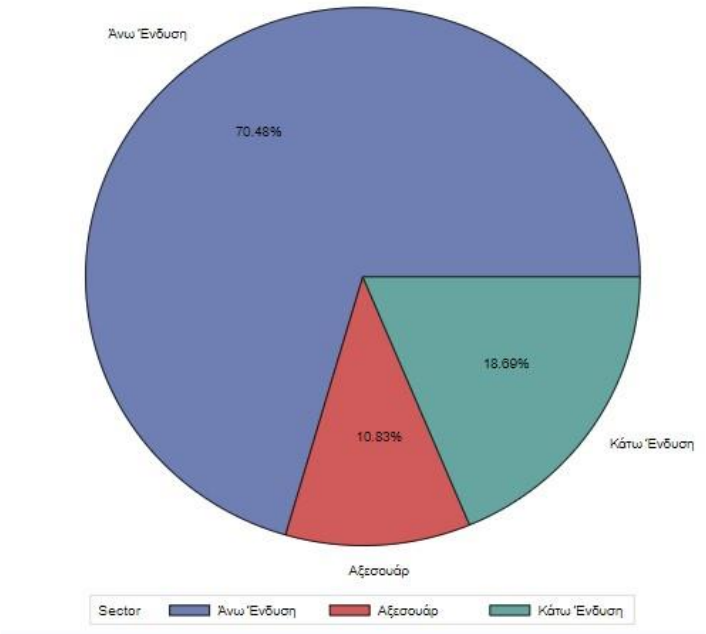

*Γράφημα 3: Τα 10 καλύτερα προϊόντα σε πωλήσεις*

Κοιτάζοντας το Γράφημα 4 και συγκρίνοντάς το με το παραπάνω επιβεβαιώνεται ότι τα παπούτσια συγκεντρώνουν τα περισσότερα χρηματικά ποσά για την επιχείρηση, αφού η κατηγορία «Sneakers» συγκεντρώνει συνολικά το μεγαλύτερο χρηματικό ποσό και στην συνέχεια την ακολουθούν οι κατηγορίες «Φορέματα», «Παντελόνια» και «Πουκάμισα».



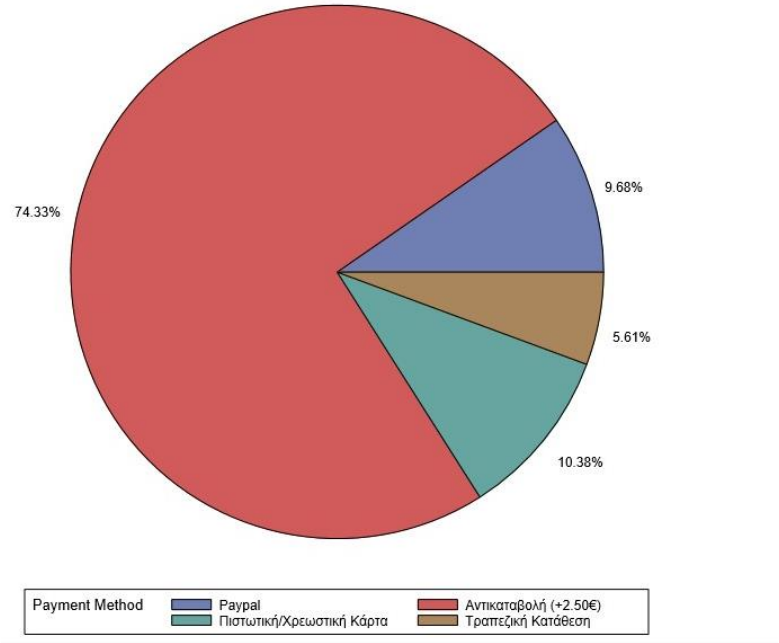

*Γράφημα 4: Οι 10 καλύτερες κατηγορίες σε πωλήσεις*

Παρακάτω στο Γράφημα 5 βλέπουμε ότι ο τομέας «Άνω Ένδυση» κυριαρχεί έναντι των άλλων δύο σε πωλήσεις προϊόντων. Ωστόσο όπως είδαμε και παραπάνω ορισμένες κατηγορίες που ανήκουν στην κάτω ένδυση αν και ο τομές έρχεται δεύτερος σε πωλήσεις καταφέρνουν και συγκεντρώνουν μεγάλα χρηματικά ποσά.




*Γράφημα 5: Πωλήσεις ανά τομέα*

Ο τρόπος πληρωμής που το μεγαλύτερο ποσοστό των πελατών επιλέγει (Γράφημα 6) είναι η πληρωμή με Αντικαταβολή (+2,50€), η οποία έχει επιπλέον χρέωση σε αυτή του προϊόντος κατά 2,50€. Αυτό πιθανότατα προκύπτει από το γεγονός ότι οι πελάτες νιώθουν έναν φόβο να δώσουν τις προσωπικές πληροφορίες της κάρτας τους ή μπορεί να μην αισθάνονται άνετα ακόμη με το κατάστημα, γιατί δεν έχουν αγοράσει αρκετές φορές. Το μικρότερο ποσοστό καταλαμβάνει η Τραπεζική κατάθεση ίσως επειδή απαιτεί περισσότερο χρόνο και προσοχή στην διαδικασία ή μπορεί και για τον ίδιο λόγο που επιλέγουν την Αντικαταβολή (+2,50€),.




*Γράφημα 6: Τρόποι πληρωμής των πελατών του ηλεκτρονικού καταστήματος*

# 5

# Εξόρυξη Δεδομένων

Έχοντας κατανοήσει τα δεδομένα του καταστήματος, στο κεφάλαιο αυτό παρουσιάζονται δύο τεχνικές εξόρυξη δεδομένων, η ανάλυση Clustering και η Market Basket Analysis. Θα δούμε τα βήματα της διαδικασίας που ακολουθήθηκαν όπως και τα προβλήματα που παρουσιαστήκαν σε αυτήν. Το κεφάλαιο χωρίζεται σε δυο ενότητες την 5.1 και 5.2 με την πρώτη να αφορά την ανάλυση clustering ενώ την δεύτερη την market basket analysis.

## 5.1 Clustering

Στην ενότητα αυτήν θα δημιουργήσουμε συστάδες των πελατών του καταστήματος. Βασικό στοιχείο της ανάλυσης clustering είναι ορισμένα χαρακτηριστικά που προκύπτουν από τα δεδομένα που έχουμε στην διάθεσή μας, μέσα από μια άλλη ανάλυση, την RFM. Στόχος είναι η δημιουργία συστάδων ή αλλιώς ομάδων πελατών οι οποίοι μοιράζονται περισσότερα κοινά χαρακτηριστικά μεταξύ τους παρά με πελάτες άλλων ομάδων.

### 5.1.1 RFM analysis

Κύριοι πυλώνες της RFM ανάλυσης είναι οι 3 μεταβλητές Recency, Frequency, Monetary. Για τον λόγο αυτό θα πρέπει να δημιουργήσουμε ένα νέο dataset με τις 3 αυτές μεταβλητές και τον μοναδικό κωδικό κάθε πελάτη (CustomerID).

Για την δημιουργία της μεταβλητής Recency χρειάστηκε να υπολογίσουμε το χρονικό διάστημα μεταξύ της ημερομηνίας της τελευταίας συναλλαγής που εμφανίζεται στο dataset (31 Οκτωβρίου 2021) και της ημερομηνίας που ο κάθε πελάτης πραγματοποίησε την τελευταία του συναλλαγή.

Υστέρα από αρκετές δοκιμές με τα δεδομένα διαπιστώθηκε προτιμότερο η συγκεκριμένη μεταβλητή να αντιπροσωπεύει τις εβδομάδες που έχουν περάσει από την τελευταία συναλλαγή του πελάτη. Φυσικά έγιναν δοκιμές τόσο σε ημέρες όσο και μήνες αλλά τα καλύτερα αποτελέσματα ήταν όταν υπολογίζαμε εβδομάδες. Σε αυτήν την απόφαση σημαντικό ρόλο είχε το μικρό χρονικό διάστημα των δεδομένων.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Η μεταβλητή Frequency αντιπροσωπεύει την συχνότητα με την οποία οι πελάτες του καταστήματος αγοράζουν από το σύνολο των δεδομένων. Για την δημιουργία της μεταβλητής αυτής χρησιμοποιήσαμε την στήλη CustomerID. Μετρήσαμε δηλαδή τον αριθμό διαφορετικών παραγγελιών που αντιστοιχούν σε κάθε πελάτη.

Τέλος δημιουργήσαμε την μεταβλητή Monetary με την βοήθεια της στήλης Total. Επί της ουσίας υπολογίσαμε την αξία κάθε αγοράς του κάθε πελάτη και έτσι καταλήξαμε στο χρηματικό ποσό που έχει δαπανήσει συνολικά.

Ωστόσο, βασικό στάδιο πριν προχωρήσουμε είναι να εξετάσουμε τα δεδομένα για να δούμε αν υπάρχουν ακραίες τιμές οι οποίες θα δημιουργήσουν πρόβλημα στην ανάλυση της ομαδοποίησης των πελατών. Οι μεταβλητές αυτές αφού εντοπιστούν θα πρέπει να εξαιρεθούν από τα δεδομένα.

Οι πίνακες που θα χρησιμοποιήσουμε είναι αυτοί των ποσοστιαίων σημείων και αυτοί των ακραίων τιμών. Επίσης θα δούμε και ορισμένα ιστογράμματα που θα μας βοηθήσουν περισσότερο να κατανοήσουμε πως είναι δομημένα τα δεδομένα.

Στον Πίνακα 1 φαίνεται μια μεγάλη απόκλιση ανάμεσα στην ελάχιστη και στην μέγιστη τιμή των μεταβλητών Frequency και Monetary. Ενώ στην μεταβλητή Recency το εύρος των τιμών είναι ικανοποιητικό.

| Variable | Mean | Minimum | Maximum |
|---|---|---|---|
| Recency | 2.6056110 | 0 | 6.0000000 |
| Frequency | 2.8467764 | 1.0000000 | 488.0000000 |
| Monetary | 189.9340680 | 0 | 32071.31 |

*Πίνακας 1: Summary Stat RFM αρχείο*

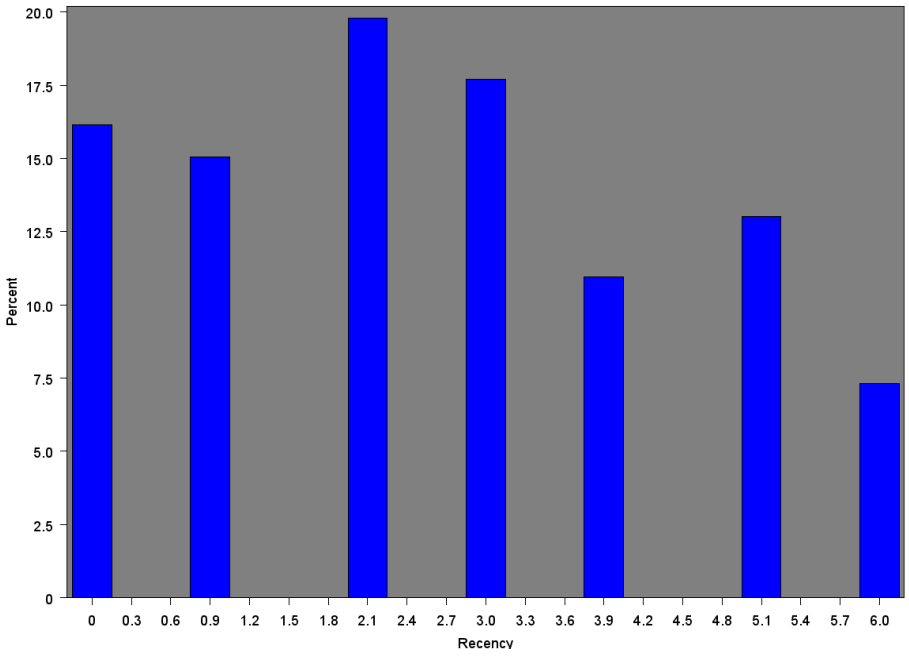
- **Recency**

Παρατηρώντας τον Πίνακα 2 με τα ποσοστιαία σημεία της μεταβλητής Recency η παραπάνω παρατήρησή μας, ότι επιβεβαιώνεται πως οι τιμές της μεταβλητής είναι καλά κατανεμημένες.

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 6 |
| 99% | 6 |
| 95% | 6 |
| 90% | 5 |
| 75% Q3 | 4 |
| 50% Median | 2 |
| 25% Q1 | 1 |
| 10% | 0 |
| 5% | 0 |
| 1% | 0 |
| 0% Min | 0 |

| Extreme Values | | | | | |
|---|---|---|---|---|---|
| Lowest | | | Highest | | |
| Order | Value | Freq | Order | Value | Freq |
| 1 | 0 | 599 | 3 | 2 | 734 |
| 2 | 1 | 558 | 4 | 3 | 656 |
| 3 | 2 | 734 | 5 | 4 | 406 |
| 4 | 3 | 656 | 6 | 5 | 483 |
| 5 | 4 | 406 | 7 | 6 | 271 |

*Πίνακας 2: Μεταβλητή Recency-Extreme Values & Quantiles*




*Γράφημα 7: Κατανομή μεταβλητής Recency*

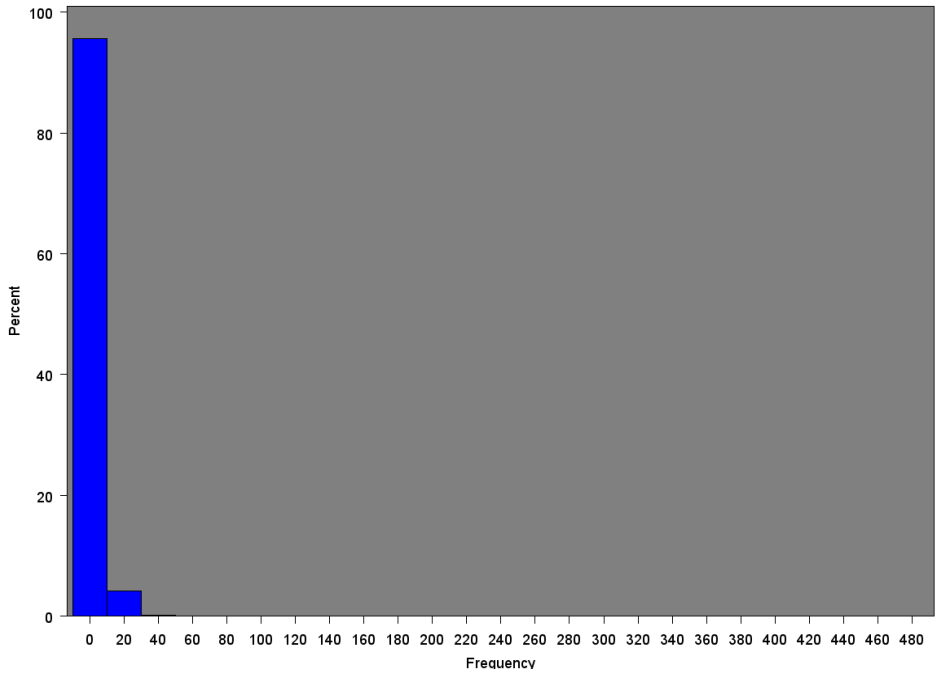
- **Frequency**

Όσον αφορά την μεταβλητή Frequency όπως φαίνεται και από τον Πίνακα 3 παρατηρούμαι ότι υπάρχουν ακραίες τιμές οι οποίες θα πρέπει να αφαιρεθούν ειδάλλως θα αποτελέσουν πρόβλημα στην μετέπειτα ανάλυση. Οι τιμές αυτές παρατηρούνται στο εκατοστό εκατοστημόριο και μπορούμε να δούμε ότι ένας πελάτης πραγματοποίησε 488 αγορές ενώ το 99% των πελατών πραγματοποίησε κάτω από 16 αγορές. Παρατηρείται επίσης ότι ένα μεγάλο ποσοστό πελατών (25%) έχει αγοράσει μια μόνο φορά.

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 488 |
| 99% | 16 |
| 95% | 9 |
| 90% | 6 |
| 75% Q3 | 3 |
| 50% Median | 2 |
| 25% Q1 | 1 |
| 10% | 1 |
| 5% | 1 |
| 1% | 1 |
| 0% Min | 1 |

| Extreme Values | | | | | |
|---|---|---|---|---|---|
| Lowest | | | Highest | | |
| Order | Value | Freq | Order | Value | Freq |
| 1 | 1 | 1661 | 26 | 28 | 1 |
| 2 | 2 | 971 | 27 | 33 | 3 |
| 3 | 3 | 350 | 28 | 37 | 1 |
| 4 | 4 | 226 | 29 | 44 | 2 |
| 5 | 5 | 101 | 30 | 488 | 1 |

*Πίνακας 3: Μεταβλητή Frequency-Extreme Values & Quantiles*




*Γράφημα 8: Κατανομή μεταβλητής Frequency*

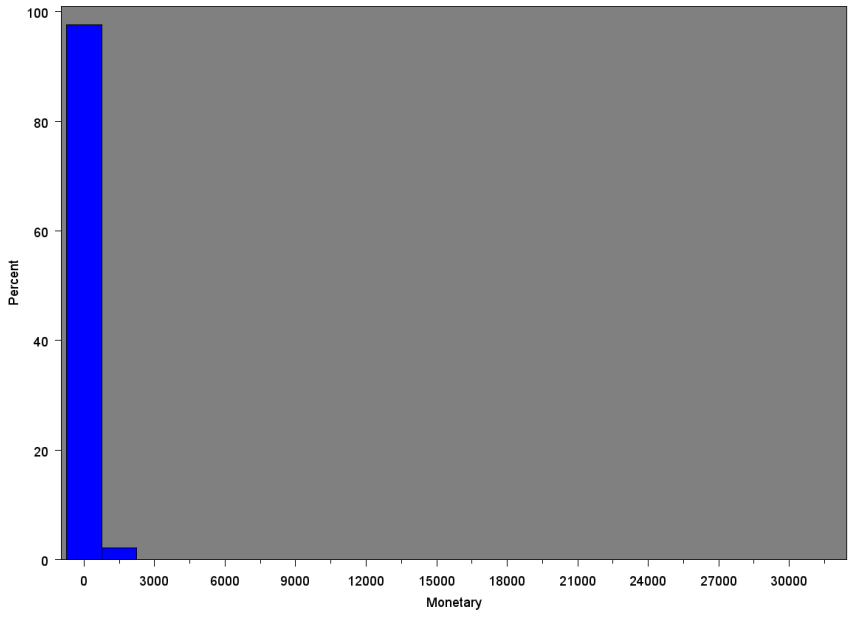
- **Monetary**

Όπως και παραπάνω έτσι και εδώ παρατηρούμε την μεταβλητή Monetary να έχει και αυτήν ακραίες τιμές που θα πρέπει να αφαιρέσουμε. Ενώ το 99% των παρατηρήσεων έχει καταναλώσει χρηματικό ποσό κάτω από 1020€ βλέπουμε ότι ένας ένα πελάτης έχει καταναλώσει 32.710€.

| Quantiles (Definition 5) | |
|---|---|
| Level | Quantile |
| 100% Max | 32071.31 |
| 99% | 1020.00 |
| 95% | 545.00 |
| 90% | 379.20 |
| 75% Q3 | 209.00 |
| 50% Median | 119.93 |
| 25% Q1 | 70.00 |
| 10% | 45.00 |
| 5% | 31.90 |
| 1% | 20.00 |
| 0% Min | 0.00 |

| Extreme Values | | | | | |
|---|---|---|---|---|---|
| Lowest | | | Highest | | |
| Order | Value | Freq | Order | Value | Freq |
| 1 | 0.0 | 2 | 1048 | 2010.00 | 1 |
| 2 | 5.0 | 6 | 1049 | 2233.00 | 1 |
| 3 | 10.0 | 1 | 1050 | 2603.51 | 1 |
| 4 | 12.0 | 1 | 1051 | 3443.00 | 1 |
| 5 | 12.5 | 1 | 1052 | 32071.31 | 1 |

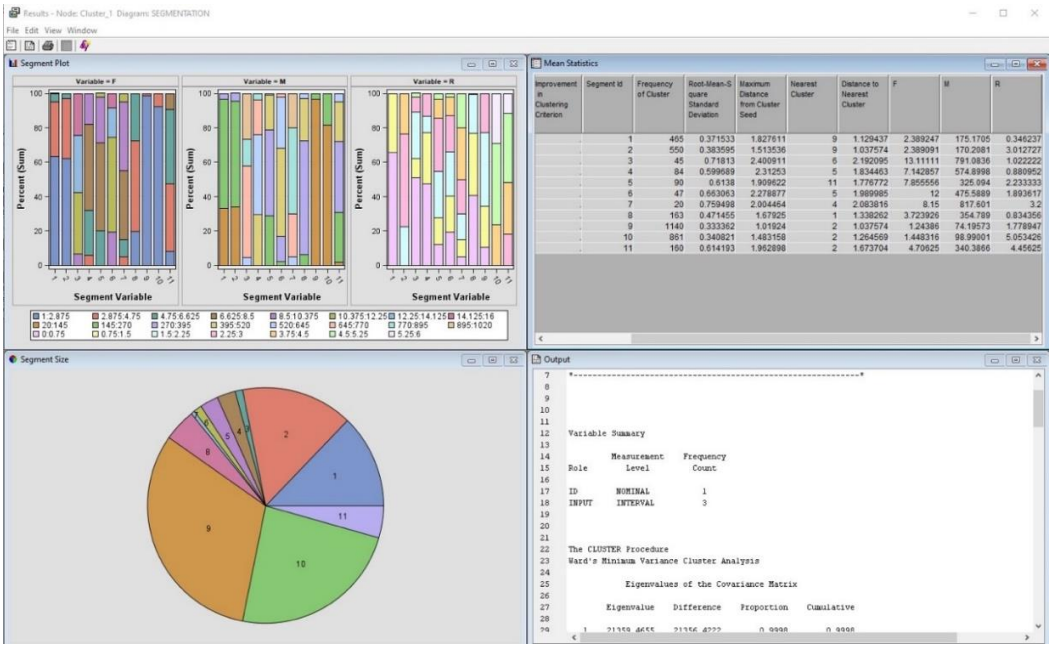
*Πίνακας 4: Μεταβλητή Monetary-Extreme Values & Quantiles*




*Γράφημα 9: Κατανομή μεταβλητής Monetary*

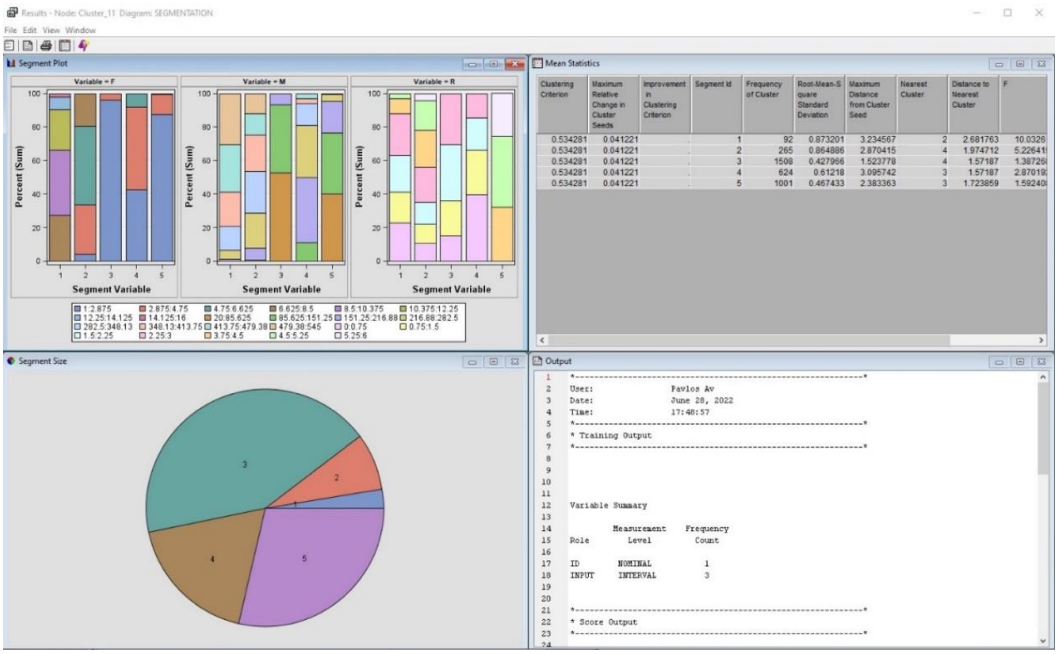
Οι ακραίες αυτές τιμές των μεταβλητών Frequency και Monetary που αποκλίνουν κατά πολύ από την πλειοψηφία των τιμών των δεδομένων μας θα πρέπει να αφαιρεθούν από την ανάλυση. Παρόλο που μιλάμε για πραγματικά δεδομένα εκπροσωπούν ένα πολύ μικρό ποσοστό του συνόλου και είναι προτιμότερο να αφαιρεθούν για να μπορέσουμε να έχουμε καλύτερα αποτελέσματα.

Έχοντας φιλτράρει τα δεδομένα και αφήνοντας τις προεπιλεγμένες παραμέτρους τα αποτελέσματα έδειξαν πως ο ιδανικός αριθμός των clusters είναι 11. Αυτός όμως ο αριθμός δεν είναι ικανοποιητικός για μπορέσουμε να περιγράψουμε τα διαφορετικά τμήματα πελατών στο σύνολο των δεδομένων.



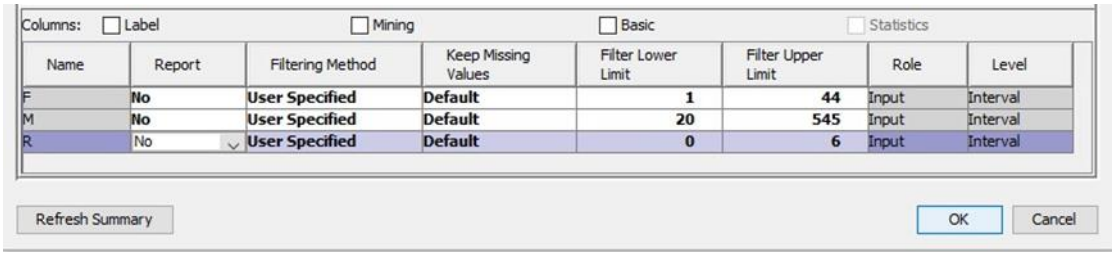

*Εικόνα 4: Clusters που προέκυψαν από την 1η προσπάθεια*

Έτσι έπειτα από πολλές δοκιμές ο αλγόριθμος κατέληξε σε 5 clusters πελατών. Το 5 είναι ένας ιδανικός αριθμός και ικανοποιεί πλήρως τις ανάγκες της ανάλυσης ώστε να παρέχουμε πολύτιμες πληροφορίες στην εταιρεία.




*Εικόνα 5: Τελικά Clusters*

Οι τιμές στις οποίες καταλήξαμε ήταν για το Recency να συμπεριλάβουμε όλες τις τιμές του, για το Frequency να εξαιρεθεί η τιμή 488 ενώ για το Monetary επιλέχθηκαν οι τιμές από ≤ 545 δηλαδή το 95,1% των παρατηρήσεων.




*Πίνακας 5: Εύρος τιμών μεταβλητών RFM μετά την αφαίρεση ακραίων τιμών*

Το αρχικό dataset είχε συνολικά 3.706 πελάτες και ύστερα από τις εξαιρέσεις το τελικό περιέχει 3.490 δηλαδή εξαιρέθηκαν 216 πελάτες. Στον Πίνακα 6 παρουσιάζονται τα στατιστικά για κάθε cluster και κάθε μεταβλητή.

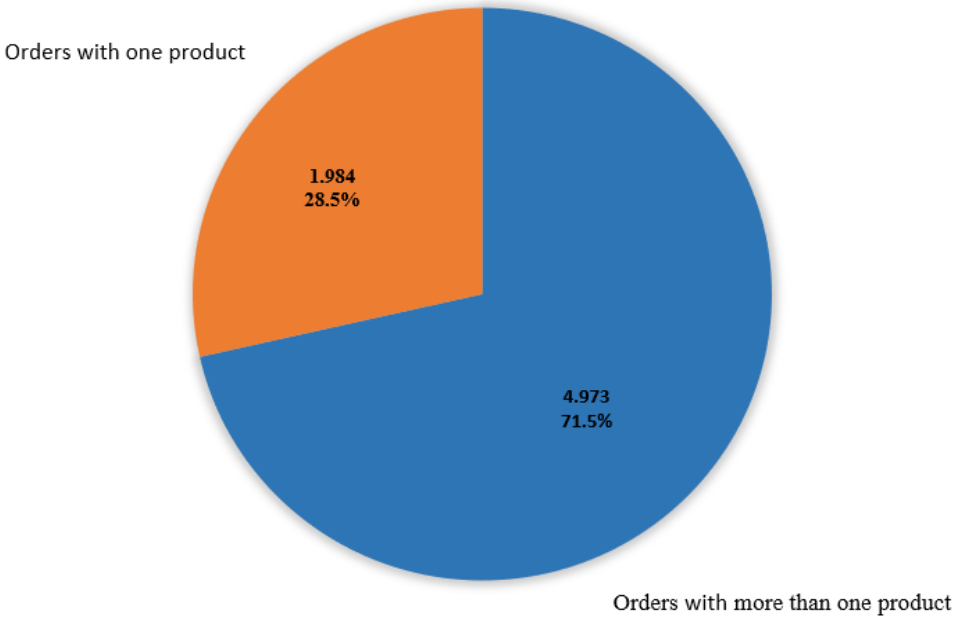
| Cluster | Αριθμός πελατών στο "Cluster" | % πελατών στο "Cluster" | Μέση τιμή "Recency" | Μέση τιμή "Frequency" | Μέση τιμή "Monetary" |
|---|---|---|---|---|---|
| 1 | 92 | 2.64% | 1,88 | 10,03 | 419,52 |
| 2 | 265 | 7.59% | 3,03 | 5,23 | 346,96 |
| 3 | 1508 | 43.21% | 1,80 | 1,39 | 86,75 |
| 4 | 624 | 17.88% | 1,09 | 2,87 | 226,77 |
| 5 | 1001 | 28.68% | 4,93 | 1,59 | 110,81 |
| Customer Base | 3.490 | 100% | 2,61 | 2,85 | 189,93 |

*Πίνακας 6: Η μέση τιμή των RFM ανά συστάδα σε σχέση με την συνολική μέση τιμή*

## 5.2 Market Basket Analysis

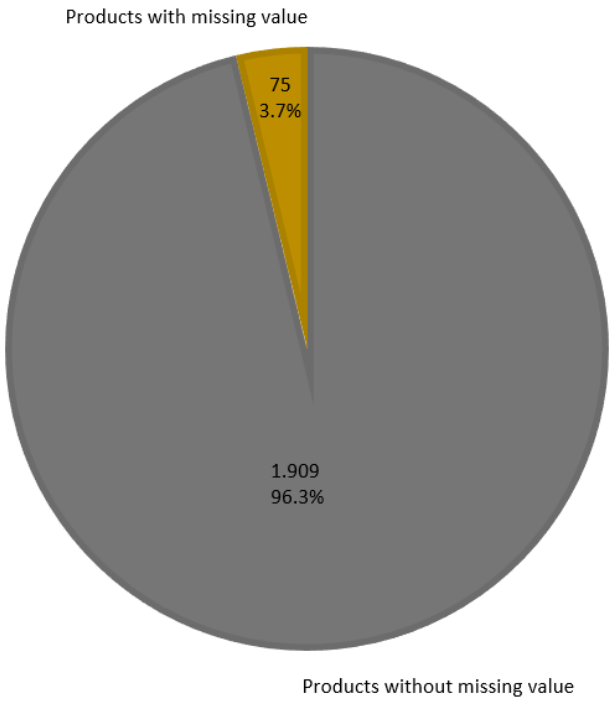
Σε αυτή την ενότητα γίνεται η ανάλυση του καλαθιού αγοράς των δεδομένων του καταστήματος. Η ανάλυση έχει ως στόχο να εντοπίσει κανόνες συσχέτισης μεταξύ των προϊόντων που αγοράστηκαν. Βασική προϋπόθεση είναι κάθε αγορά να περιέχει πάνω από ένα προϊόν, αφού οι συσχετίσεις που αναζητά έχουν στο επίκεντρο προϊόντα που αγοράστηκαν μαζί. Κάτι που θα βοηθούσε πολύ στην ανάλυσή είναι το σύνολο των δεδομένων να περιέχει προϊόντα που αγοράστηκαν πάνω από μια φορά. Βέβαια, τα δεδομένα μας υστερούν κατά πολύ σε αυτό το χαρακτηριστικό, αφού διαπιστώθηκε ότι περίπου το 1/3 του συνόλου των προϊόντων αγοράστηκαν μια μόνο φορά.

Στο Γράφημα 10 φαίνεται το ποσοστό των αγορών που περιέχουν μια ένα προϊόν είναι 71,45%. Αν και πρόκειται για ένα αρκετά μεγάλο ποσοστό, είμαστε αναγκασμένοι να τα αφαιρέσουμε, αφού δεν θα συμβάλουν κάπως στην ανάλυση.




*Γράφημα 10: Παραγγελίες που περιέχουν μόνο ένα προϊόν*

Από τις 1.984 αγορές που παρέμειναν και περιείχαν περισσότερα από 2 προϊόντα αφαιρέθηκαν ακόμη 75 αγορές οι οποίες περιείχαν 2 ακριβώς προϊόντα και το ένα είχε ονομασία NONE. Μπορούμε να δούμε και οπτικά (Γράφημα 11) αυτό τον μικρό αριθμό αγορών που χρειάστηκαν να εξαιρεθούν από την διαδικασία.




*Γράφημα 11: Προϊόντα με missing values*

Σημαντικό να σημειωθεί πως εξαιτίας της μεγάλης ποικιλίας των προϊόντων σε συνδυασμό με την μικρή εμφάνιση τους μέσα στο σύνολο των δεδομένων, για την ανάλυση χρησιμοποιήθηκε η κατηγορία των προϊόντων (38 στο σύνολο) και ο μοναδικός αριθμός κάθε παραγγελίας.

Οι κατηγορίες δίνονται αναλυτικά στα δεδομένα μας και έτσι γνωρίζουμε σε ποια κατηγορία ανήκει το καθένα. Το dataset που σχηματίστηκε περιέχει 5.423 γραμμές για 1.909 μοναδικές αγορές. Από την ανάλυση προέκυψαν 43 σχέσεις κατηγοριών.

| Relations | Expected Confidence (%) | Confidence (%) | Support(%) | Lift | Transaction Count | Rule | Left Hand Rule | Right Hand Rule |
|---|---|---|---|---|---|---|---|---|
| 2 | 7.81 | 88.89 | 2.93 | 11.39 | 56 | Κοστούμια ==> Αξεσουάρ | Κοστούμια | Αξεσουάρ |
| 2 | 3.3 | 37.58 | 2.93 | 11.39 | 56 | Αξεσουάρ ==> Κοστούμια | Αξεσουάρ | Κοστούμια |
| 2 | 7.81 | 52.38 | 2.3 | 6.71 | 44 | Σακάκια ==> Αξεσουάρ | Σακάκια | Αξεσουάρ |
| 2 | 4.4 | 29.53 | 2.3 | 6.71 | 44 | Αξεσουάρ ==> Σακάκια | Αξεσουάρ | Σακάκια |
| 2 | 23.15 | 74.19 | 4.82 | 3.2 | 92 | Φούστες ==> Μπλούζες & Top | Φούστες | Μπλούζες & Top |
| 2 | 6.5 | 20.81 | 4.82 | 3.2 | 92 | Μπλούζες & Top ==> Φούστες | Μπλούζες & Top | Φούστες |
| 2 | 31.27 | 71.43 | 2.36 | 2.28 | 45 | Κοστούμια ==> Πουκάμισα | Κοστούμια | Πουκάμισα |
| 2 | 35.83 | 70.24 | 3.09 | 1.96 | 59 | Σακάκια ==> Παντελόνια | Σακάκια | Παντελόνια |
| 2 | 7.81 | 14.07 | 4.4 | 1.8 | 84 | Πουκάμισα ==> Αξεσουάρ | Πουκάμισα | Αξεσουάρ |
| 2 | 31.27 | 56.38 | 4.4 | 1.8 | 84 | Αξεσουάρ ==> Πουκάμισα | Αξεσουάρ | Πουκάμισα |
| 2 | 31.27 | 53.57 | 2.36 | 1.71 | 45 | Σακάκια ==> Πουκάμισα | Σακάκια | Πουκάμισα |
| 2 | 24.52 | 36.96 | 1.78 | 1.51 | 34 | T-shirt ==> Μπλούζες & Φούτερ | T-shirt | Μπλούζες & Φούτερ |
| 2 | 35.83 | 52.09 | 16.29 | 1.45 | 311 | Πουκάμισα ==> Παντελόνια | Πουκάμισα | Παντελόνια |
| 2 | 31.27 | 45.47 | 16.29 | 1.45 | 311 | Παντελόνια ==> Πουκάμισα | Παντελόνια | Πουκάμισα |
| 2 | 31.27 | 45.37 | 2.57 | 1.45 | 49 | Ζώνες ==> Πουκάμισα | Ζώνες | Πουκάμισα |
| 2 | 24.52 | 33.73 | 2.99 | 1.38 | 57 | Σορτς & Βερμούδες ==> Μπλούζες & Φούτερ | Σορτς & Βερμούδες | Μπλούζες & Φούτερ |
| 2 | 8.85 | 12.18 | 2.99 | 1.38 | 57 | Μπλούζες & Φούτερ ==> Σορτς & Βερμούδες | Μπλούζες & Φούτερ | Σορτς & Βερμούδες |
| 2 | 11.31 | 13.68 | 3.35 | 1.21 | 64 | Μπλούζες & Φούτερ ==> Jeans | Μπλούζες & Φούτερ | Jeans |
| 2 | 24.52 | 29.63 | 3.35 | 1.21 | 64 | Jeans ==> Μπλούζες & Φούτερ | Jeans | Μπλούζες & Φούτερ |
| 2 | 11 | 13.25 | 3.25 | 1.2 | 62 | Μπλούζες & Φούτερ ==> Sneakers | Μπλούζες & Φούτερ | Sneakers |
| 2 | 24.52 | 29.52 | 3.25 | 1.2 | 62 | Sneakers ==> Μπλούζες & Φούτερ | Sneakers | Μπλούζες & Φούτερ |
| 2 | 23.15 | 26.39 | 2.99 | 1.14 | 57 | Jeans ==> Μπλούζες & Top | Jeans | Μπλούζες & Top |
| 2 | 11.31 | 12.9 | 2.99 | 1.14 | 57 | Μπλούζες & Top ==> Jeans | Μπλούζες & Top | Jeans |
| 2 | 23.15 | 25.44 | 2.25 | 1.1 | 43 | Σορτς & Βερμούδες ==> Μπλούζες & Top | Σορτς & Βερμούδες | Μπλούζες & Top |
| 2 | 11 | 12.06 | 3.77 | 1.1 | 72 | Πουκάμισα ==> Sneakers | Πουκάμισα | Sneakers |
| 2 | 31.27 | 34.29 | 3.77 | 1.1 | 72 | Sneakers ==> Πουκάμισα | Sneakers | Πουκάμισα |
| 2 | 35.83 | 39.08 | 1.78 | 1.09 | 34 | Πανωφόρια ==> Παντελόνια | Πανωφόρια | Παντελόνια |
| 2 | 31.27 | 33.8 | 3.82 | 1.08 | 73 | Jeans ==> Πουκάμισα | Jeans | Πουκάμισα |
| 2 | 11.31 | 12.23 | 3.82 | 1.08 | 73 | Πουκάμισα ==> Jeans | Πουκάμισα | Jeans |
| 2 | 23.15 | 22.46 | 2.2 | 0.97 | 42 | Φορέματα ==> Μπλούζες & Top | Φορέματα | Μπλούζες & Top |
| 2 | 35.83 | 33.33 | 1.89 | 0.93 | 36 | Ζώνες ==> Παντελόνια | Ζώνες | Παντελόνια |
| 2 | 35.83 | 31.54 | 2.46 | 0.88 | 47 | Αξεσουάρ ==> Παντελόνια | Αξεσουάρ | Παντελόνια |
| 2 | 24.52 | 21.44 | 6.71 | 0.87 | 128 | Πουκάμισα ==> Μπλούζες & Φούτερ | Πουκάμισα | Μπλούζες & Φούτερ |
| 2 | 31.27 | 27.35 | 6.71 | 0.87 | 128 | Μπλούζες & Φούτερ ==> Πουκάμισα | Μπλούζες & Φούτερ | Πουκάμισα |
| 2 | 35.83 | 30.95 | 3.4 | 0.86 | 65 | Sneakers ==> Παντελόνια | Sneakers | Παντελόνια |
| 2 | 31.27 | 25.44 | 2.25 | 0.81 | 43 | Σορτς & Βερμούδες ==> Πουκάμισα | Σορτς & Βερμούδες | Πουκάμισα |
| 2 | 24.52 | 19.44 | 6.97 | 0.79 | 133 | Παντελόνια ==> Μπλούζες & Φούτερ | Παντελόνια | Μπλούζες & Φούτερ |
| 2 | 35.83 | 28.42 | 6.97 | 0.79 | 133 | Μπλούζες & Φούτερ ==> Παντελόνια | Μπλούζες & Φούτερ | Παντελόνια |
| 2 | 23.15 | 16.08 | 5.76 | 0.69 | 110 | Παντελόνια ==> Μπλούζες & Top | Παντελόνια | Μπλούζες & Top |
| 2 | 35.83 | 24.89 | 5.76 | 0.69 | 110 | Μπλούζες & Top ==> Παντελόνια | Μπλούζες & Top | Παντελόνια |
| 2 | 35.83 | 24.54 | 2.78 | 0.68 | 53 | Jeans ==> Παντελόνια | Jeans | Παντελόνια |
| 2 | 24.52 | 13.12 | 3.04 | 0.54 | 58 | Μπλούζες & Top ==> Μπλούζες & Φούτερ | Μπλούζες & Top | Μπλούζες & Φούτερ |
| 2 | 23.15 | 12.39 | 3.04 | 0.54 | 58 | Μπλούζες & Φούτερ ==> Μπλούζες & Top | Μπλούζες & Φούτερ | Μπλούζες & Top |

*Πίνακας 7: Κανόνες Προϊόντων που προέκυψαν*

# 6

# Σχολιασμός αποτελεσμάτων

Σε αυτό το κεφάλαιο θα σχολιάσουμε τα αποτελέσματα που πήραμε από την ανάλυση Clustering και από την Market Basket Analysis. Στόχος μας είναι να ερμηνεύσουμε τα αποτελέσματα με τέτοιον τρόπο ώστε να είναι εύκολα κατανοητά από επιχειρησιακής πλευράς.

## 6.1 Customers Profile

Για να μπορέσει η επιχείρηση να κατανοήσει πλήρως τους πελάτες και τα χαρακτηριστικά τους πρέπει να προχωρήσουμε στο profiling των πελατών. Με την βοήθεια του Πίνακα 8 θα ονομάσουμε τους πελάτες του κάθε cluster, και θα ορίσουμε ένα βέλος σε κάθε RFM μεταβλητή. Τα βέλη δείχνουν αν ο μέσος όρος των τιμών Recency, Frequency και Monetary του κάθε cluster είναι καλύτερος (↑) ή χειρότερος (↓) από τον μέσο όρο του Recency, Frequency και Monetary του συνόλου των πελατών.

| Cluster | Recency | Frequency | Monetary | Segment Size |
|---|---|---|---|---|
| 1 | ↑ | ↑ | ↑ | 2,64% |
| 2 | ↓ | ↑ | ↑ | 7,59% |
| 3 | ↑ | ↓ | ↓ | 43,21% |
| 4 | ↑ | ↑ | ↑ | 17,88% |
| 5 | ↓ | ↓ | ↓ | 28,68% |

*Πίνακας 8: Η κατεύθυνση των τιμών RFM σε σχέση με το μέσο όρο του δείγματος*

Στο cluster 1 ανήκουν οι Champions και πρόκειται για το μικρότερο σε αριθμό cluster καθώς αποτελείται από 92 πελάτες. Αν και αντιπροσωπεύουν μόλις το 2,64% του πληθυσμού έχουν την μεγαλύτερη συχνότητα αγορών, έχουν ξοδέψει το μεγαλύτερο χρηματικό ποσό και έρχονται δεύτεροι, με μικρή διαφορά, σε πιο πρόσφατες επισκέψεις στο κατάστημα.

Στο cluster 2 ανήκουν οι Churners, οι οποίοι είναι πελάτες οι οποίοι έχουν επισκεφτεί αρκετές φορές το κατάστημα στο παρελθόν, έχουν ξοδέψει το δεύτερο μεγαλύτερο κατά μέσο όρο χρηματικό ποσό ενώ κατά μέσο όρο η τελευταία αγορά έγινε προσεγγιστικά πριν 3 εβδομάδες.

Στο cluster 3 βρίσκονται οι Recent πελάτες, πρόκειται για πελάτες που επισκέφτηκαν πρόσφατα το κατάστημα χωρίς να αφήσουν αρκετά χρήματα στο κατάστημα και συνολικά έχουν χαμηλή συχνότητα αγορών. Πρόκειται για το μεγαλύτερο cluster αφού αντιπροσωπεύουν το 43,21% του συνόλου των πελατών.

Στο cluster 4 βρίσκονται οι Valuable πελάτες, αντιπροσωπεύουν το 17,88% των πελατών και είναι το τρίτο μεγαλύτερο cluster. Έχουν επισκεφτεί πρόσφατα το κατάστημα έχουν μεγάλη επισκεψιμότητα και ξοδεύουν αρκετά χρήματα.

Στο cluster 5 ανήκουν οι Worst πελάτες, αυτοί δηλαδή που έχουν χαμηλές και τις τρεις τιμές RFM. Αντιπροσωπεύουν το 28,68% του συνόλου των πελατών και έρχεται στην δεύτερη θέση του μεγαλύτερου cluster. Πρόκειται για πελάτες που στο παρελθόν επισκέφτηκαν το κατάστημα μια φορά και ξόδεψα πολύ λίγα χρήματα.

| Cluster Name | Recency | Frequency | Monetary |
|---|---|---|---|
| Champions | ↑ | ↑ | ↑ |
| Churners | ↓ | ↑ | ↑ |
| Recent | ↑ | ↓ | ↓ |
| Valuable | ↑ | ↑ | ↑ |
| Worst | ↓ | ↓ | ↓ |

*Πίνακας 9: Χαρακτηρισμός των clusters σύμφωνα με τις τιμές των RFM*

# 6.2 Προτιμήσεις στις Κατηγορίες Προϊόντων

Όπως είδαμε σε προηγούμενο κεφάλαιο, προέκυψαν 64 συσχετίσεις ανάμεσα στις κατηγορίες των προϊόντων. Εξετάζοντας, τις τιμές confidence, support, και lift καταλήξαμε ότι 6 είναι οι πιο σημαντικές συσχετίσεις που ενδιαφέρουν την επιχείρηση.

1. Κοστούμια → Αξεσουάρ

Τα προϊόντα που ανήκουν στις κατηγορίες Κοστούμια και Αξεσουάρ με πιθανότητα 2,93% αγοράζονται μαζί. Εάν κάποιος πελάτης αγοράσει ένα προϊόν από την κατηγορία Κοστούμια με πιθανότητα 88,89% θα αγοράσει και ένα προϊόν από την κατηγορία Αξεσουάρ. Ένας πελάτης που έχει αγοράσει κάποιο προϊόν από την κατηγορία Κοστούμια είναι 11,39 φορές πιο πιθανό να αγοράσει και κάποιο προϊόν της κατηγορίας Αξεσουάρ συγκριτικά με έναν τυχαίο πελάτη.

2. Σακάκια → Αξεσουάρ

Τα προϊόντα που ανήκουν στις κατηγορίες Σακάκια και Αξεσουάρ με πιθανότητα 2,30% αγοράζονται μαζί. Εάν κάποιος πελάτης αγοράσει ένα προϊόν από την κατηγορία Σακάκια με πιθανότητα 52,38% θα αγοράσει και ένα προϊόν από την κατηγορία Αξεσουάρ. Ένας πελάτης που έχει

αγοράσει κάποιο προϊόν από την κατηγορία Σακάκια είναι 6,71 φορές πιο πιθανό να αγοράσει και κάποιο προϊόν της κατηγορίας Αξεσουάρ συγκριτικά με έναν τυχαίο πελάτη.

3. Φούστες → Μπλούζες & Top

Τα προϊόντα που ανήκουν στις κατηγορίες Φούστες και Μπλούζες & Top με πιθανότητα 4,82% αγοράζονται μαζί. Εάν κάποιος πελάτης αγοράσει ένα προϊόν από την κατηγορία Φούστες με πιθανότητα 74,19% θα αγοράσει και ένα προϊόν από την κατηγορία Μπλούζες & Top. Ένας πελάτης που έχει αγοράσει κάποιο προϊόν από την κατηγορία Φούστες είναι 3,20 φορές πιο πιθανό να αγοράσει και κάποιο προϊόν της κατηγορίας Μπλούζες & Top συγκριτικά με έναν τυχαίο πελάτη.

4. Σακάκια → Παντελόνια

Τα προϊόντα που ανήκουν στις κατηγορίες Σακάκια και Παντελόνια με πιθανότητα 3,09% αγοράζονται μαζί. Εάν κάποιος πελάτης αγοράσει ένα προϊόν από την κατηγορία Σακάκια με πιθανότητα 70,24% θα αγοράσει και ένα προϊόν από την κατηγορία Παντελόνια. Ένας πελάτης που έχει αγοράσει κάποιο προϊόν από την κατηγορία Σακάκια είναι 1,96 φορές πιο πιθανό να αγοράσει και κάποιο προϊόν της κατηγορίας Παντελόνια συγκριτικά με έναν τυχαίο πελάτη.

5. Αξεσουάρ → Πουκάμισα

Τα προϊόντα που ανήκουν στις κατηγορίες Αξεσουάρ και Κοστούμια με πιθανότητα 4,40% αγοράζονται μαζί. Εάν κάποιος πελάτης αγοράσει ένα προϊόν από την κατηγορία Αξεσουάρ με πιθανότητα 56,38% θα αγοράσει και ένα προϊόν από την κατηγορία Κοστούμια. Ένας πελάτης που έχει αγοράσει κάποιο προϊόν από την κατηγορία Αξεσουάρ είναι 1,80 φορές πιο πιθανό να αγοράσει και κάποιο προϊόν της κατηγορίας Κοστούμια συγκριτικά με έναν τυχαίο πελάτη.

6. Πουκάμισα → Παντελόνια

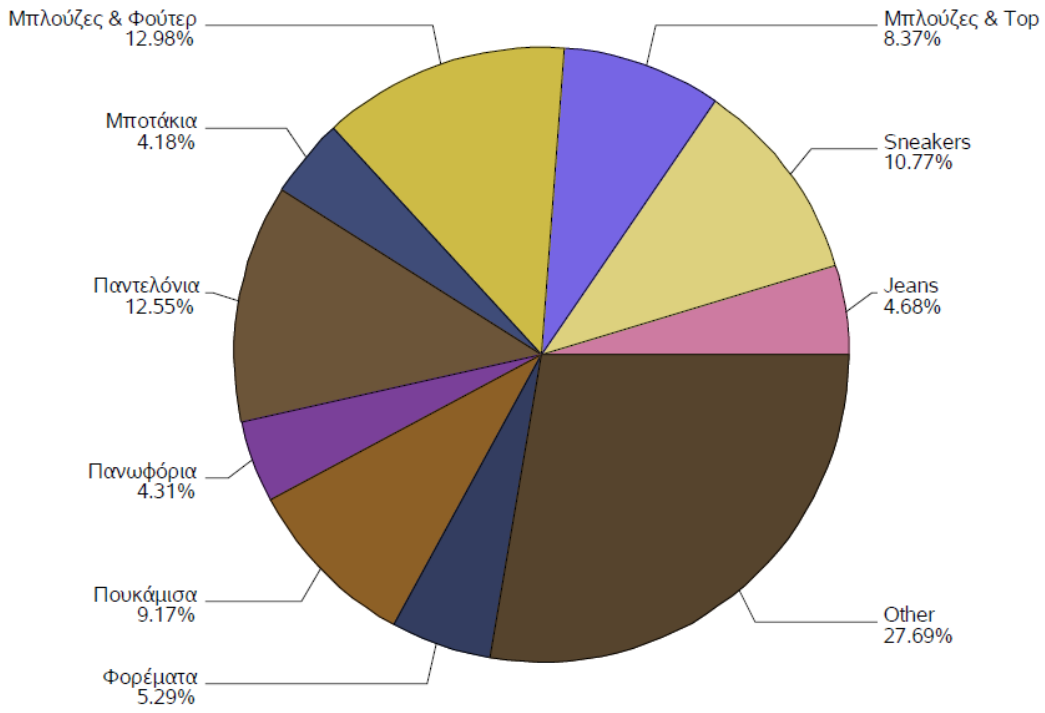
Τα προϊόντα που ανήκουν στις κατηγορίες Πουκάμισα και Παντελόνια με πιθανότητα 16,29% αγοράζονται μαζί. Εάν κάποιος πελάτης αγοράσει ένα προϊόν από την κατηγορία Πουκάμισα με πιθανότητα 52,09% θα αγοράσει και ένα προϊόν από την κατηγορία Παντελόνια. Ένας πελάτης που έχει αγοράσει κάποιο προϊόν από την κατηγορία Πουκάμισα είναι 1,45 φορές πιο πιθανό να αγοράσει και κάποιο προϊόν της κατηγορίας Παντελόνια συγκριτικά με έναν τυχαίο πελάτη.

## 6.3 Προτάσεις Marketing

Σε προηγούμενη ενότητα, πραγματοποιήσαμε ανάλυση RFM και τμηματοποιήσαμε τους πελάτες του ηλεκτρονικού καταστήματος σε clusters ανάλογα με τα κοινά χαρακτηριστικά τους. Ακόμη, μέσα από την Market Basket Analysis προέκυψαν ορισμένοι σημαντικοί κανόνες κατηγοριών προϊόντων οι οποίες αγοράζονται μαζί.

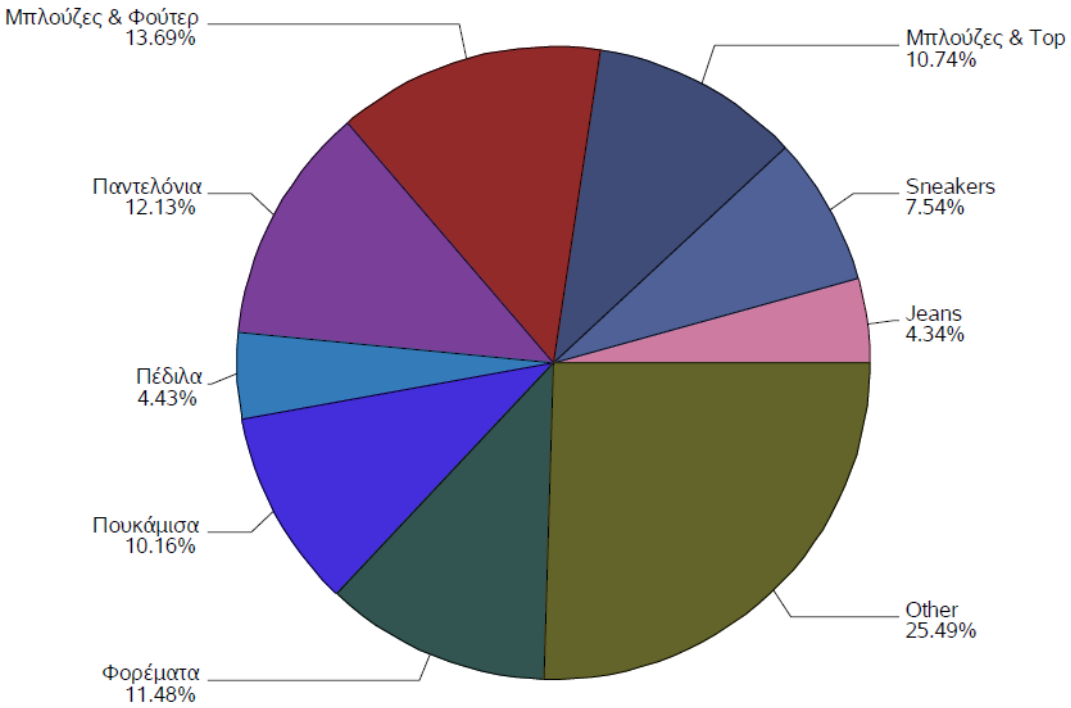
Σε αυτήν την ενότητα θα χρησιμοποιήσουμε την γνώση που αποκτήσαμε στις προηγούμενες ενότητες για να μπορέσουμε να εμβαθύνουμε σε κάθε cluster. Θα δούμε με γραφήματα τις κατηγορίες προϊόντων που κάθε cluster πελατών επιλέγει και στην συνέχεια με την βοήθεια των κανόνων συσχέτισης των κατηγοριών θα κάνουμε ορισμένες προτάσεις marketing που θα είναι επικεντρωμένες στο κάθε cluster.

Ξεκινώντας με το cluster των Champions βλέπουμε ότι πιο συχνή επιλογή τους είναι οι κατηγορίες Μπλούζες & Φούτερ και Παντελόνια. Το cluster αποτελείται από 92 πελάτες και είναι το μικρότερο ανάμεσα στα άλλα παρόλα αυτά επισκέπτονται συχνά το ηλεκτρονικό κατάστημα και ξοδεύουν αρκετά χρήματα. Αυτούς τους πελάτες δεν μας απασχολεί να τους δώσουμε εκπτώσεις αφού φαίνεται να δαπανούν αρκετά χρήματα αλλά θέλουμε να τους κάνουμε να καταλάβουν ότι είναι ξεχωριστοί για εμάς. Αν και διεκδικούν ένα μικρό ποσοστό του συνόλου η διατήρησή του είναι καίριας σημασίας καθώς η επιχείρηση συγκεντρώνοντας πληροφορίες από αυτούς μπορεί να μάθει πολλά για τον τρόπο λειτουργίας της και τι οδηγεί αυτούς τους πελάτες να αγοράζουν από αυτήν. Στους Champions μπορούμε κάθε φορά που ολοκληρώνουν μια αγορά να στέλνεται ένα email που θα τους ευχαριστεί για την προτίμησή τους και θα τους παροτρύνει να κάνουν μια ανάρτηση στα social media (Instagram, TikTok, Facebook, κλπ.) με hashtag το brand της επιχείρησης ή να σχολιάσουν θετικά για τα προϊόντα σε αυτά. Αυτό θα βοηθήσει στην προβολή των προϊόντων και στην προσέλκυση νέων πελατών. Ταυτόχρονα αξιοποιώντας τους κανόνες συσχέτισης που προέκυψαν και κάνοντας cross selling προτάσεις θα μπορέσουμε να τους οδηγήσουμε σε μεγαλύτερες αγορές άρα και μεγαλύτερα χρηματικά ποσά. Για παράδειγμα ένας συνδυασμός κατηγοριών που θα τους ενδιέφερε είναι Πουκάμισα → Παντελόνια αφού είναι 1,45 φορές πιο πιθανό αν ένας πελάτης που έχει αγοράσει κάποιο προϊόν από την μια κατηγορία να αγοράσει και κάποιο προϊόν από την άλλη, συγκριτικά με έναν τυχαίο πελάτη και με πιθανότητα 45,47% αν αγοράσουν από την κατηγορία Παντελόνια θα αγοράσουν και από την κατηγορία Πουκάμισα.



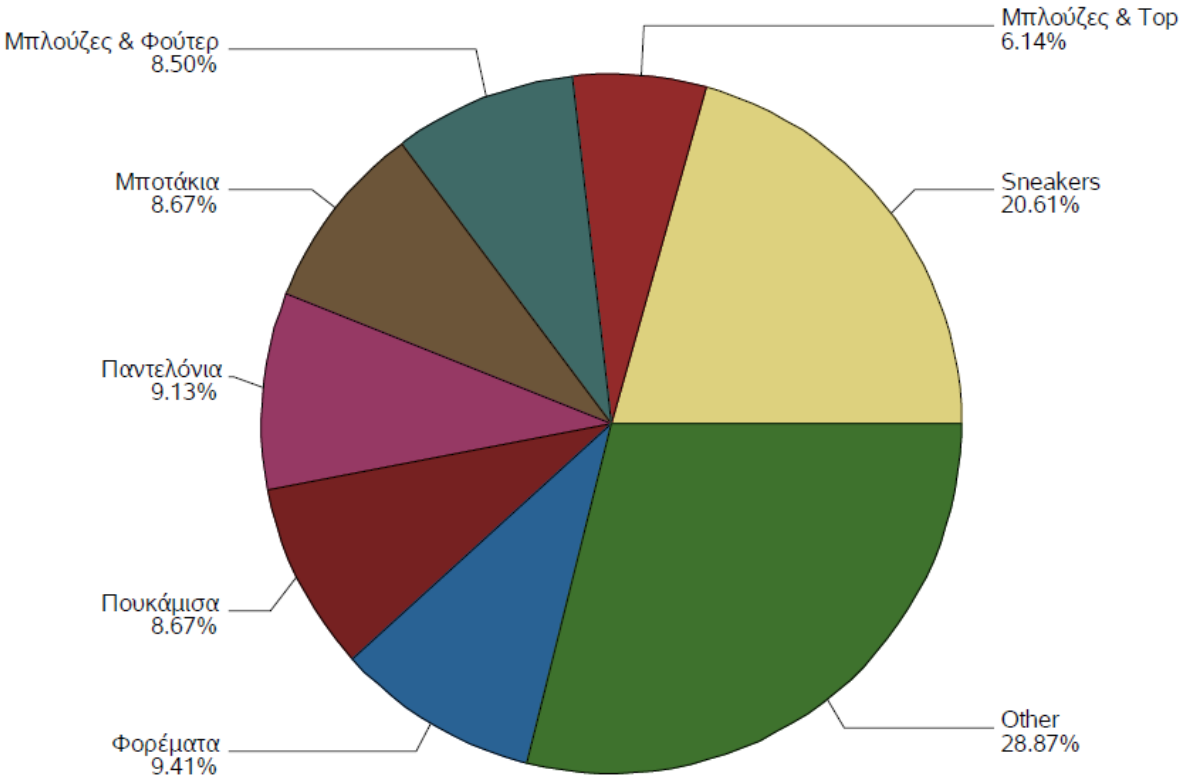

*Γράφημα 12: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster 'Champions'*

Το επόμενο cluster είναι οι Churners, και παρατηρώντας το Γράφημα 13 εντοπίζουμε αρκετές ομοιότητες στις αγοραστικές του προτιμήσεις με τους Champions πελάτες. Πρόκειται για ένα σημαντικό τμήμα πελατών διότι η συχνότητα αγορών τους ήταν υψηλή, όπως και το ποσό που δαπανούσαν σε σύγκριση με τον αντιπροσωπευτικό μέσο πελάτη της πελατειακής βάσης. Αυτοί οι πελάτες πριν γίνουν Churners θα μπορούσανε να είναι Champions. Στην περίπτωση αυτή είναι σημαντικό οι δράσεις να έχουν κύριο στόχο την επανένταξή τους ως ενεργούς πελάτες και την κατανόηση των αιτιών της αποχώρησής τους. Η ενημέρωση για προσφορές ώστε να δείξουμε πως δεν τους έχουμε ξεχάσει αλλά ενδιαφερόμαστε ακόμη για αυτούς, η υπενθύμιση για αγορά προϊόντων (customer reactivation program) και αίτημα για αξιολόγηση καταστήματος με στόχο την κατανόηση της πρόσφατης και «ξαφνικής» αποχώρησής τους είναι ορισμένες από αυτές τις δράσεις. Στην συνέχεια μέσω του feedback θα μπορέσουμε να δούμε πιθανά λάθη που έγιναν και αυτό θα βοηθήσει ώστε να μην επαναληφθούν και γίνουν αιτία αποχώρησης και άλλων πελατών. Φυσικά όσοι από τους πελάτες δεν ανταποκριθούν σε ένα χρονικό διάστημα ενός μήνα θα σημειωθούν και μελλοντικά δεν θα ασχοληθούμε ξανά μαζί τους.



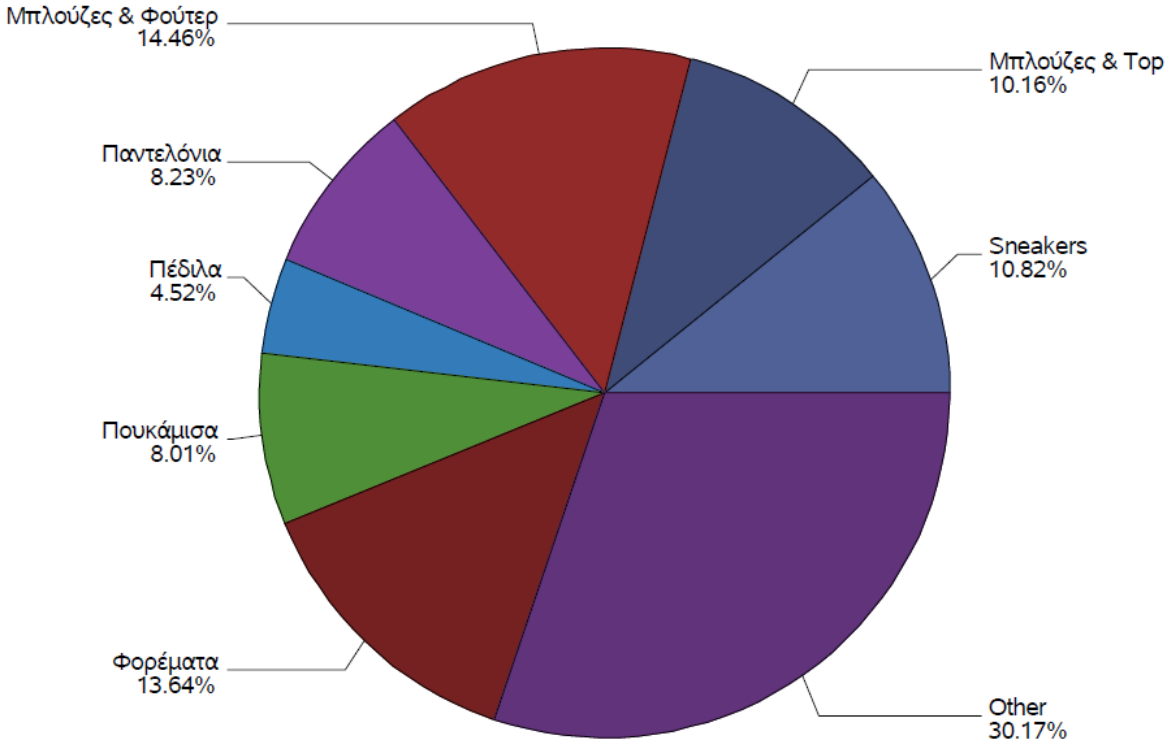

*Γράφημα 13: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster 'Churners'*

Όπως φαίνεται στο Γράφημα 14 οι Recent δείχνουν ιδιαίτερη προτίμηση στην κατηγορία Sneakers και το cluster αυτό είναι το μεγαλύτερο αφού αποτελείται από 1508 πελάτες δηλαδή το 43,21% όλων των πελατών. Πρόκειται για πελάτες που επισκέπτονται για μια φορά το κατάστημα και δεν ξοδεύουν μεγάλα χρηματικά ποσά. Η επιχείρηση παρατηρώντας τις κατηγορίες που προσελκύουν τους περισσότερους πελάτες θα μπορούσε να έχει συνεχώς στο ηλεκτρονικό της κατάστημα προτάσεις προϊόντων που θα οδηγήσουν σε μεγαλύτερες αγορές από την πρώτη κιόλας φορά. Για παράδειγμα ιδανικός συνδυασμός είναι οι κατηγορίες οι Sneakers → Μπλούζες & Φούτερ όπου είναι 3.25 φορές πιο πιθανό αν αγοραστεί η μια να αγοραστεί και η άλλη ή οι κατηγορίες Πουκάμισα → Παντελόνια όπου είναι 1.45 πιο πιθανό αν αγοραστεί η μια να αγοραστεί και η άλλη. Ακόμη σε αυτούς τους πελάτες το ηλεκτρονικό κατάστημα μπορεί μετά την πρώτη τους παραγγελία να τους αποστείλει κουπόνια για την πιο δημοφιλή κατηγορία, με έκπτωση ύψους 15% στην επόμενη αγορά τους και επιπλέον 5% αν το συνδυάσουν με ακόμη ένα προϊόν άλλης κατηγορίας, το οποίο θα έχει ισχύ 20 ημέρες. Έτσι θα τους παρακινήσει να επισκεφτούν ξανά το κατάστημα και σε σύντομο χρονικό διάστημα να πραγματοποιήσουν νέα παραγγελία και να δαπανήσουν περισσότερα χρήματα.



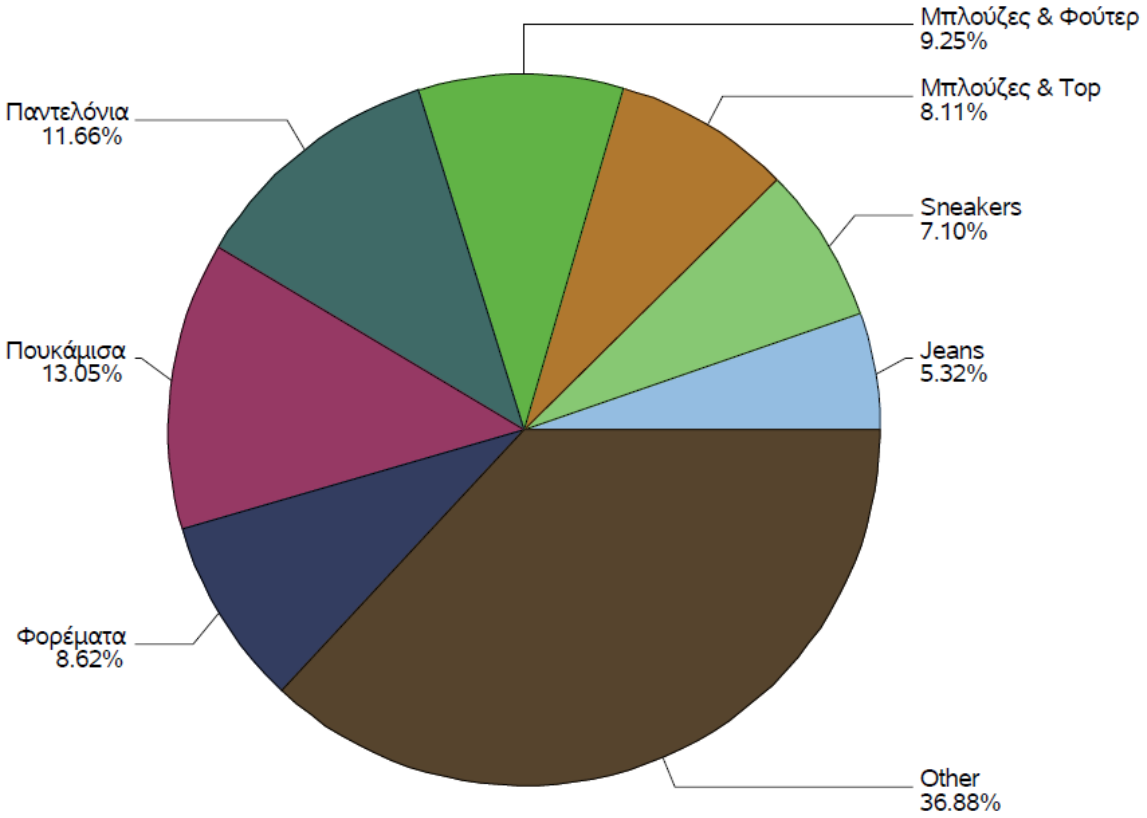

*Γράφημα 14: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster 'Recent'*

Οι Valuable πελάτες αποτελούν το 17,88% των πελατών. Έχουν επισκεφτεί πρόσφατα το ηλεκτρονικό κατάστημα και έχουν αφήσει ικανοποιητικό χρηματικό ποσό όσον αφορά όμως την επισκεψιμότητα τους βρίσκονται στο όριο με αυτό τους μέσου πελάτη. Αυτό μας κάνει να θέλουμε να διατηρήσουμε αυτούς τους πελάτες και να τους οδηγήσουμε να γίνουν οι Loyal πελάτες. Βλέπουμε μεγάλη προτίμηση σε Μπλούζες & Φούτερ όπως επίσης και σε Φορέματα. Σε αυτούς τους πελάτες το θα μπορούσαμε να στείλουμε μέσω ηλεκτρονικού ταχυδρομείου κάποια εκπτωτικό κουπόνι για την επόμενη αγορά τους με την προϋπόθεση ότι θα αγοράσουν μαζί δύο ή περισσότερες από τις δημοφιλείς κατηγορίες. Επίσης θα μπορούσαμε να ορίσουμε δυο ημέρες μέσα στον μήνα, κατά προτίμηση στην αρχή και στο τέλος, που οι αγορές τους θα έχουν τα μεταφορικά δωρεάν. Οι δράσεις αυτές έχουν στόχο να ενισχύσουν την αφοσίωση των πελατών και να τους κατευθύνει σε ένα σταθερό μοτίβο αγορών το οποίο είναι κάτι που λείπει από το ηλεκτρονικό κατάστημα και το έχει ανάγκη.




*Γράφημα 15: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster 'Valuable'*

Οι Worst είναι το δεύτερο μεγαλύτερο cluster και έχουν τις χειρότερες τιμές RFM. Επειδή όμως πρόκειται για ένα μεγάλο ποσοστό πελατών μπορούμε να στείλουμε αιτήματα για αξιολόγηση του καταστήματος (contact customers for feedback), για εντοπισμό ελλείψεων προϊόντων ή κακής παρουσίασης του καταστήματος (όπως δυσκολίες κατά την εξυπηρέτηση ή την εύρεση προϊόντων). Η δράση αποσκοπεί στην κατανόηση του λόγου αποχώρησης αυτών και μετέπειτα την αξιοποίηση της γνώσης αυτής για τη βελτίωση του ηλεκτρονικού καταστήματος και της εξυπηρέτησης των πελατών. Ακόμη θα μπορούσε να στείλει εκπτωτικά κουπόνια μικρής χρονικής διάρκειας, ώστε να τους παρακινήσει να κάνουν αγορές. Όσοι από αυτούς δεν ανταποκριθούν στις προωθητικές ενέργειες θα αποκλειστούν από την πελατειακή βάση και το ηλεκτρονικό κατάστημα δεν θα ξανά ενδιαφερθεί.




*Γράφημα 16: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster 'Worst'*

# 7

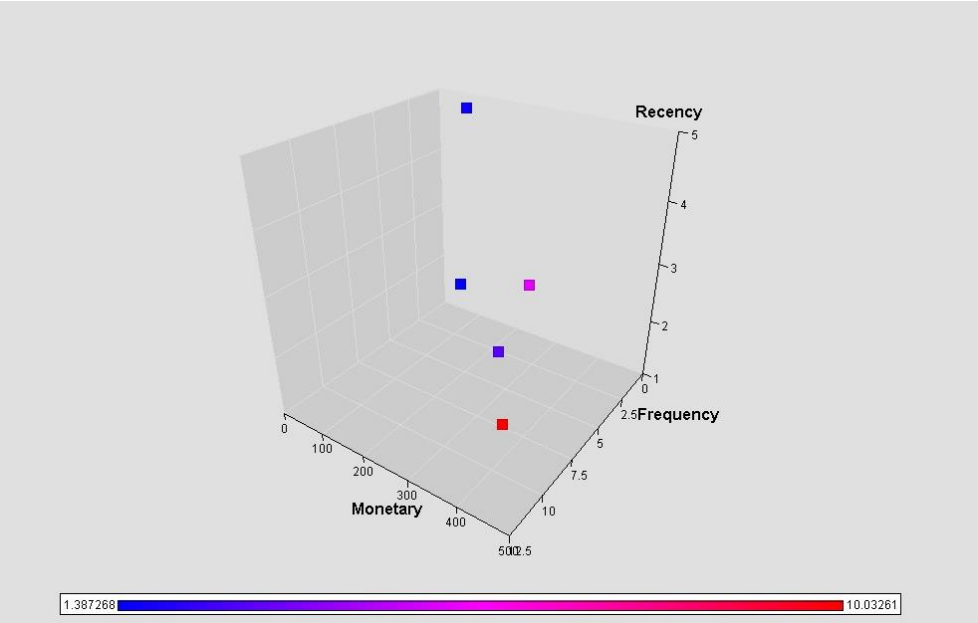
# Αξιολόγηση

Στο κεφάλαιο αυτό θα γίνει η αξιολόγηση του clustering και της market basket analysis.

## 7.1 Αξιολόγηση των Clusters

Το clustering έχει ως στόχο να εντοπίσει μέσα από τα δεδομένα πελάτες με κοινά χαρακτηριστικά και να δημιουργήσει ομάδες που στην συνέχεια μπορούν να βοηθήσουν από επιχειρησιακής πλευράς την εταιρεία. Ο ιδανικός αριθμός clusters δεν είναι ούτε πολύ μεγάλος ώστε να μπορεί να γίνει σωστή ανάλυση των αποτελεσμάτων αλλά ούτε πολύ μικρός. Στην δικιά μας περίπτωση δημιουργήθηκαν 5 ομάδες πελατών που είναι ένας εξαιρετικός αριθμός.

Για να μπορέσουμε να πούμε ότι έγινε μια καλή ομαδοποίηση θα πρέπει τα στοιχεία στο εσωτερικό κάθε cluster να έχουν κοινά χαρακτηριστικά και να υπάρχει μεγάλη απόσταση ανάμεσα στα clusters. Στο Γράφημα 17 μπορούμε και οπτικά να δούμε τα κέντρα του κάθε cluster σε σχέση με τις 3 μεταβλητές Recency, Frequency και Monetary.

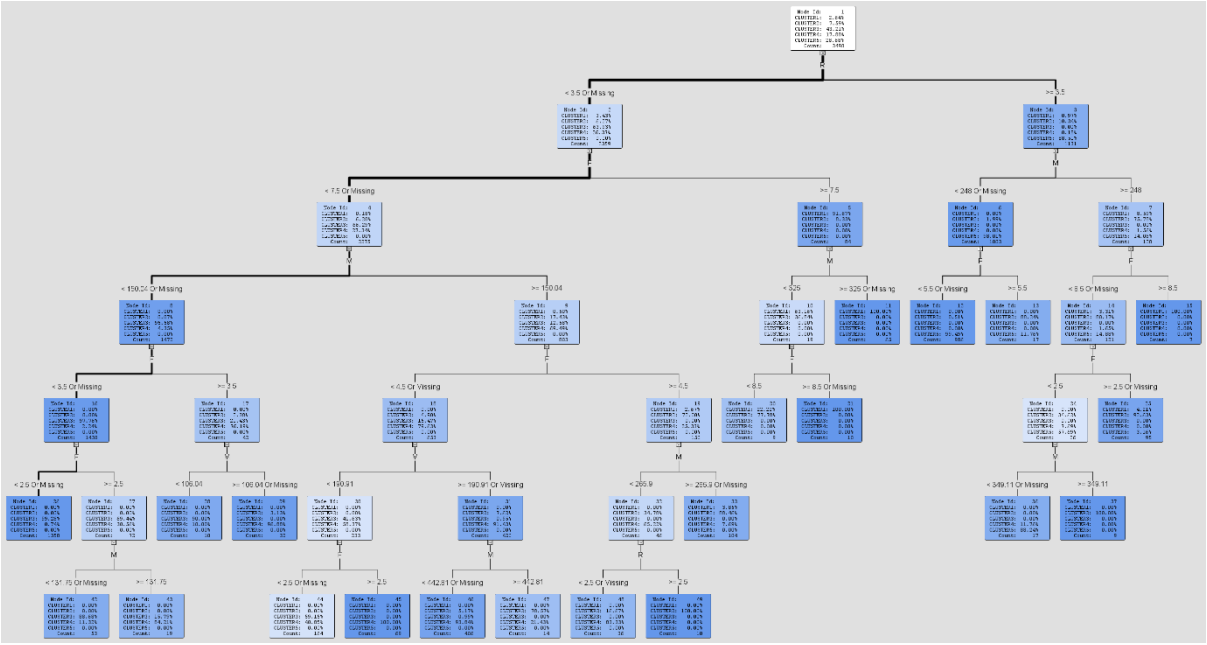



*Γράφημα 17: Scatterplot RFM ανά Cluster*

Φαίνεται ότι τα clusters που δημιουργήσαμε δεν έχουν μεγάλες αποστάσεις μεταξύ τους ενώ αυτό που θα θέλαμε ιδανικά ήταν το ακριβώς αντίθετο. Δεδομένου όμως του σχετικά μικρού διαστήματος το οποίο εξετάζουμε σε συνδυασμό με το μικρό εύρος τιμών των Recency και Frequency τα αποτελέσματα είναι πλήρως ικανοποιητικά.

Ένας τρόπος για να αναπαραστήσουμε την διαδικασία με την οποία δημιουργούνται τα clusters είναι το δέντρο αποφάσεων. Στο Γράφημα 18 φαίνεται το δέντρο το οποίο μας δείχνει με ποια λογική χωρίζονται τα clusters όπως επίσης και την σημαντικότητα των μεταβλητών στον διαχωρισμό αυτόν.

Παρατηρώντας τα κλαδιά του δένδρου αποφάσεων βλέπουμε για παράδειγμα ότι εάν ένα πελάτης έχει τιμή Recency ≥ 3,5, την Monetary < 248 και τιμή Frequency < 5,5 ή Missing τότε με πιθανότητα 99,49% θα ανήκει στο cluster 5.



*Γράφημα 18: Δένδρο αποφάσεων*

Από τον Πίνακα 10 Παρατηρούμε ότι η μεταβλητές Recency και Monetary έχουν βοηθήσει περισσότερο στον διαχωρισμό των clusters. Αν και η μεταβλητή Recency έχει χρησιμοποιηθεί τις λιγότερες φορές από τις τρεις σε διαχωρισμό κάποιου node η συνεισφορά της έχει το μεγαλύτερο ποσοστό σημαντικότητας.

| Variable Name | Label | Number of Splitting Rules | Importance |
|---|---|---|---|
| R | | 2 | 1.0000 |
| M | | 9 | 0.9355 |
| F | | 9 | 0.5994 |

*Πίνακας 10: Σημαντικότητα των μεταβλητών RFM στο διαχωρισμό των clusters*

# 7.2 Αξιολόγηση των Association Rules

Για τον προσδιορισμό των σημαντικών κανόνων συσχέτισης χρησιμοποιούμε τις τιμές support και confidence. Ορίζοντας κάθε φορά ένα επιθυμητό κατώτατο όριο μπορούμε να πούμε αν έναν κανόνας είναι σημαντικό ή όχι. Βέβαια μια ακόμη τιμή που είναι άκρως σημαντική στην απόφαση αυτήν είναι η τιμή lift. Έχει αποδειχθεί ότι παρά τα μέτρα που παίρνουμε, πολλοί από τους κανόνες που προκύπτουν, περίπου το 30%, είναι στατιστικά ασήμαντοι (Dorn et al., 2008). Έτσι ως ένα επιπλέον μέτρο αξιολόγησης της σημαντικότητας των κανόνων είναι η Chi-squared analysis.

Το Chi-squared ενός κανόνα μπορεί να υπολογιστεί απευθείας από τις τιμές της support, της frequency και της lift (ενδιαφέροντος) του εν λόγω κανόνα. Τα αποτελέσματά μας διευκολύνουν το κλάδεμα των συνόλων κανόνων και επιτρέπουν τον εντοπισμό στατιστικά σημαντικών κανόνων που μπορεί να έχουν παραλειφθεί από τον αλγόριθμο εξόρυξης (Alvarez, 2003).

Παρακάτω (Πίνακα 11) βλέπουμε τα αποτελέσματα της ανάλυσης Chi-squared για το δικό μας πρόβλημα. Παρατηρείται ότι υπάρχουν 12 κανόνες συσχέτισης που ενώ έχουν lift > 1 και καλές τιμές support και confidence, δεν είναι στατιστικά σημαντικοί. Αντίθετα υπάρχουν κανόνες που ενώ έχουν lift < 1 που φαίνεται να είναι στατιστικά σημαντικοί. Από τους κανόνες συσχέτισης που αναφέρθηκαν παραπάνω στατιστικά σημαντικοί θεωρούνται μόνο οι πρώτοι 6 από αυτούς.

| Obs | EXP_CONF | CONF | SUPPORT | LIFT | RULE | index | CHISQ | PVALUE |
|---|---|---|---|---|---|---|---|---|
| 1 | 7.81 | 88.89 | 2.93 | 11.39 | Κοστούμια ==> Αξεσουάρ | 1 | 595.245 | 0.00000 |
| 2 | 3.30 | 37.58 | 2.93 | 11.39 | Αξεσουάρ ==> Κοστούμια | 2 | 595.245 | 0.00000 |
| 3 | 7.81 | 52.38 | 2.30 | 6.71 | Σακάκια ==> Αξεσουάρ | 3 | 242.624 | 0.00000 |
| 4 | 4.40 | 29.53 | 2.30 | 6.71 | Αξεσουάρ ==> Σακάκια | 4 | 242.624 | 0.00000 |
| 5 | 23.15 | 74.19 | 4.82 | 3.20 | Φούστες ==> Μπλούζες & Top | 5 | 194.165 | 0.00000 |
| 6 | 6.50 | 20.81 | 4.82 | 3.20 | Μπλούζες & Top ==> Φούστες | 6 | 194.165 | 0.00000 |
| 7 | 31.27 | 71.43 | 2.36 | 2.28 | Κοστούμια ==> Πουκάμισα | 7 | 48.878 | 0.00000 |
| 8 | 35.83 | 70.24 | 3.09 | 1.96 | Σακάκια ==> Παντελόνια | 8 | 45.244 | 0.00000 |
| 9 | 7.81 | 14.07 | 4.40 | 1.80 | Πουκάμισα ==> Αξεσουάρ | 9 | 47.384 | 0.00000 |
| 10 | 31.27 | 56.38 | 4.40 | 1.80 | Αξεσουάρ ==> Πουκάμισα | 10 | 47.384 | 0.00000 |
| 11 | 31.27 | 53.57 | 2.36 | 1.71 | Σακάκια ==> Πουκάμισα | 11 | 20.327 | 0.00001 |
| 12 | 24.52 | 36.96 | 1.78 | 1.51 | T-shirt ==> Μπλούζες & Φούτερ | 12 | 8.085 | 0.00446 |
| 13 | 35.83 | 52.09 | 16.29 | 1.45 | Πουκάμισα ==> Παντελόνια | 13 | 99.930 | 0.00000 |
| 14 | 31.27 | 45.47 | 16.29 | 1.45 | Παντελόνια ==> Πουκάμισα | 14 | 99.930 | 0.00000 |
| 15 | 31.27 | 45.37 | 2.57 | 1.45 | Ζώνες ==> Πουκάμισα | 15 | 10.585 | 0.00114 |
| 16 | 24.52 | 33.73 | 2.99 | 1.38 | Σορτς & Βερμούδες ==> Μπλούζες & Φούτερ | 16 | 8.503 | 0.00355 |
| 17 | 8.85 | 12.18 | 2.99 | 1.38 | Μπλούζες & Φούτερ ==> Σορτς & Βερμούδες | 17 | 8.503 | 0.00355 |
| 18 | 11.31 | 13.68 | 3.35 | 1.21 | Μπλούζες & Φούτερ ==> Jeans | 18 | 3.442 | 0.06354 |
| 19 | 24.52 | 29.63 | 3.35 | 1.21 | Jeans ==> Μπλούζες & Φούτερ | 19 | 3.442 | 0.06354 |
| 20 | 11.00 | 13.25 | 3.25 | 1.20 | Μπλούζες & Φούτερ ==> Sneakers | 20 | 3.198 | 0.07371 |
| 21 | 24.52 | 29.52 | 3.25 | 1.20 | Sneakers ==> Μπλούζες & Φούτερ | 21 | 3.198 | 0.07371 |
| 22 | 23.15 | 26.39 | 2.99 | 1.14 | Jeans ==> Μπλούζες & Top | 22 | 1.433 | 0.23129 |
| 23 | 11.31 | 12.90 | 2.99 | 1.14 | Μπλούζες & Top ==> Jeans | 23 | 1.433 | 0.23129 |
| 24 | 23.15 | 25.44 | 2.25 | 1.10 | Σορτς & Βερμούδες ==> Μπλούζες & Top | 24 | 0.547 | 0.45970 |
| 25 | 11.00 | 12.06 | 3.77 | 1.10 | Πουκάμισα ==> Sneakers | 25 | 0.996 | 0.31816 |
| 26 | 31.27 | 34.29 | 3.77 | 1.10 | Sneakers ==> Πουκάμισα | 26 | 0.996 | 0.31816 |
| 27 | 35.83 | 39.08 | 1.78 | 1.09 | Πανωφόρια ==> Παντελόνια | 27 | 0.419 | 0.51753 |
| 28 | 31.27 | 33.80 | 3.82 | 1.08 | Jeans ==> Πουκάμισα | 28 | 0.722 | 0.39563 |
| 29 | 11.31 | 12.23 | 3.82 | 1.08 | Πουκάμισα ==> Jeans | 29 | 0.722 | 0.39563 |
| 30 | 23.15 | 22.46 | 2.20 | 0.97 | Φορέματα ==> Μπλούζες & Top | 30 | 0.056 | 0.81285 |
| 31 | 35.83 | 33.33 | 1.89 | 0.93 | Ζώνες ==> Παντελόνια | 31 | 0.310 | 0.57742 |
| 32 | 35.83 | 31.54 | 2.46 | 0.88 | Αξεσουάρ ==> Παντελόνια | 32 | 1.292 | 0.25575 |
| 33 | 24.52 | 21.44 | 6.71 | 0.87 | Πουκάμισα ==> Μπλούζες & Φούτερ | 33 | 4.438 | 0.03514 |
| 34 | 31.27 | 27.35 | 6.71 | 0.87 | Μπλούζες & Φούτερ ==> Πουκάμισα | 34 | 4.438 | 0.03514 |
| 35 | 35.83 | 30.95 | 3.40 | 0.86 | Sneakers ==> Παντελόνια | 35 | 2.442 | 0.11814 |
| 36 | 31.27 | 25.44 | 2.25 | 0.81 | Σορτς & Βερμούδες ==> Πουκάμισα | 36 | 2.931 | 0.08688 |
| 37 | 24.52 | 19.44 | 6.97 | 0.79 | Παντελόνια ==> Μπλούζες & Φούτερ | 37 | 14.812 | 0.00012 |
| 38 | 35.83 | 28.42 | 6.97 | 0.79 | Μπλούζες & Φούτερ ==> Παντελόνια | 38 | 14.812 | 0.00012 |
| 39 | 23.15 | 16.08 | 5.76 | 0.69 | Παντελόνια ==> Μπλούζες & Top | 39 | 29.959 | 0.00000 |
| 40 | 35.83 | 24.89 | 5.76 | 0.69 | Μπλούζες & Top ==> Παντελόνια | 40 | 29.959 | 0.00000 |
| 41 | 35.83 | 24.54 | 2.78 | 0.68 | Jeans ==> Παντελόνια | 41 | 13.510 | 0.00024 |
| 42 | 24.52 | 13.12 | 3.04 | 0.54 | Μπλούζες & Top ==> Μπλούζες & Φούτερ | 42 | 40.346 | 0.00000 |
| 43 | 23.15 | 12.39 | 3.04 | 0.54 | Μπλούζες & Φούτερ ==> Μπλούζες & Top | 43 | 40.346 | 0.00000 |

*Πίνακας 11: Αποτελέσματα chi-squared ανάλυσης*

# 8

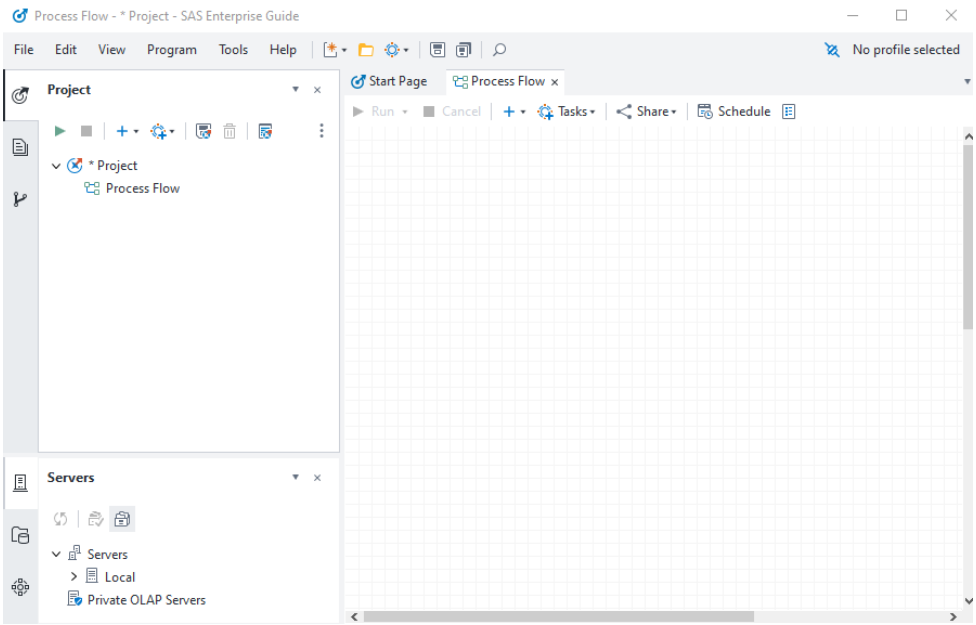
# Τεχνικές Λεπτομέρειες

Σε αυτό το κεφάλαιο παρουσιάζονται οι τεχνικές λεπτομέρειες που πραγματοποιήσαμε για την παρούσα διπλωματική. Το εργαλείο που χρησιμοποιήσαμε είναι το λογισμικό SAS και συγκεκριμένα το SAS Enterprise Guide 8.3 και το SAS Enterprise Miner Workstation 15.2 Το κεφάλαιο αναφέρεται στο Clustering analysis αλλά και στην MBA.

Το Enterprise Guide για την περιγραφική ανάλυση των δεδομένων μας και για την RFM ανάλυση. Ενώ το Enterprise Miner Workstation χρησιμοποιήθηκε για ανάλυση Clustering και την εξαγωγή των Association Rules.

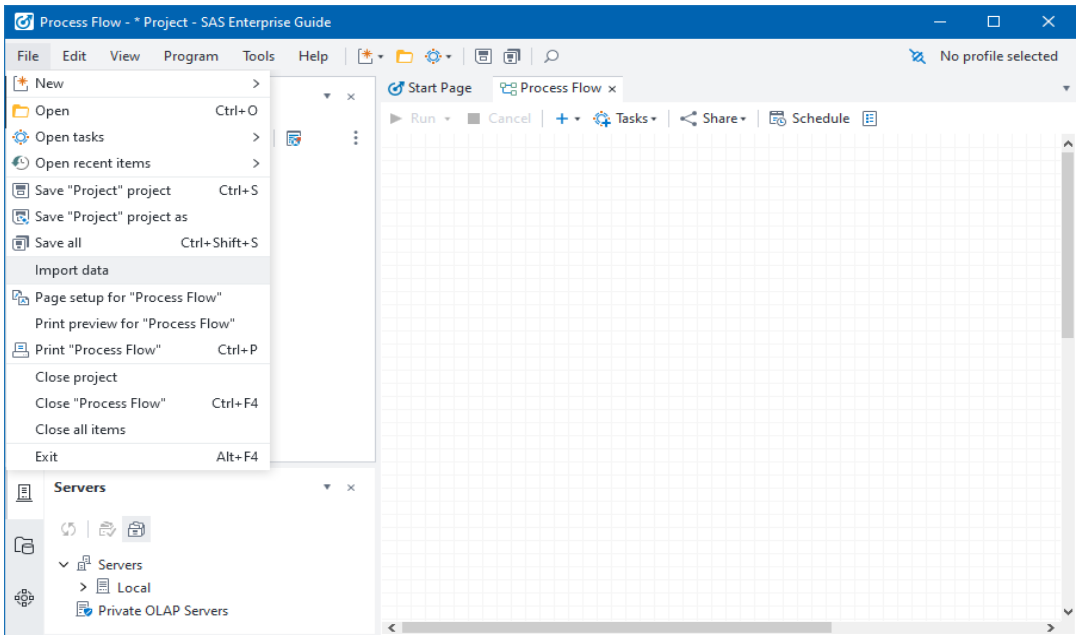
## 8.1 Clustering

Ανοίγοντας το παράθυρο του Enterprise Guide βλέπουμε στο πάνω μέρος μια γραμμή εργασιών από την οποία μπορούμε να δημιουργήσουμε να επεξεργαστούμε και να αποθηκεύσουμε project, να κάνουμε εισαγωγή δεδομένων, να δημιουργήσουμε βιβλιοθήκες και πολλές ακόμη ενέργειες.



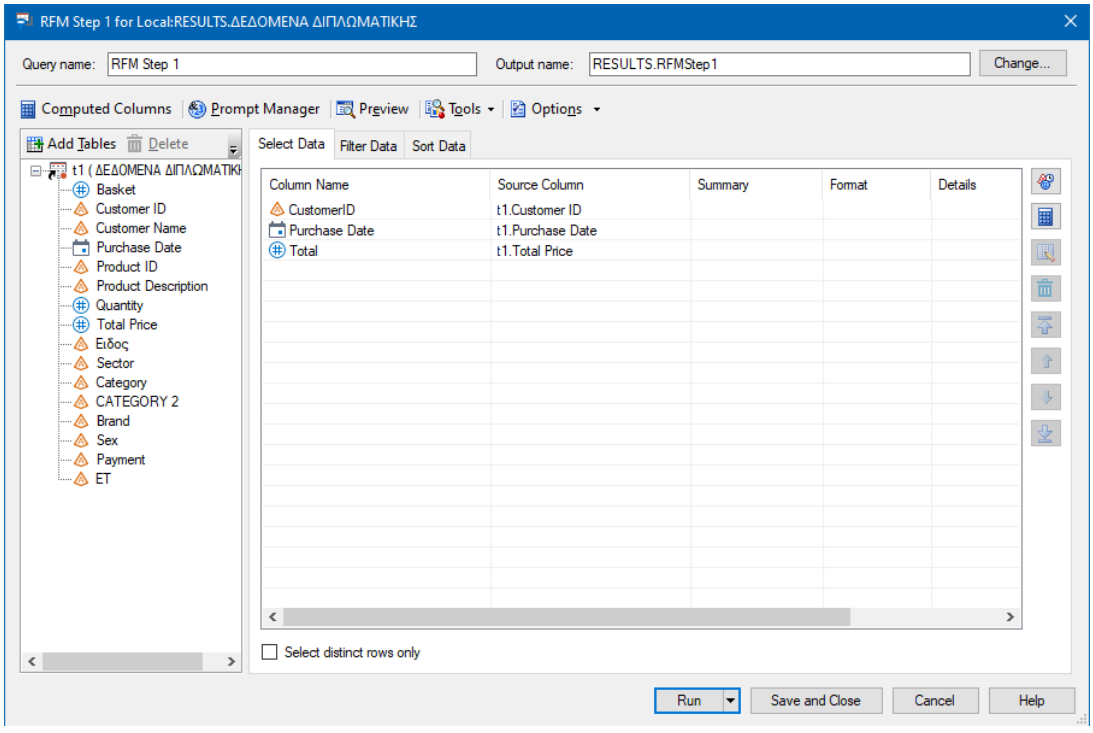

*Εικόνα 6: Η αρχική οθόνη του SAS Enterprise Guide*

Για να εισάγουμε το Excel με τα δεδομένα της εταιρείας, πρέπει να επιλέξουμε από την γραμμή εργασιών το File και από το dropdown menu που θα ανοίξει πατάμε την επιλογή import data.




*Εικόνα 7: Import αρχείου Excel στο Enterprise Guide*

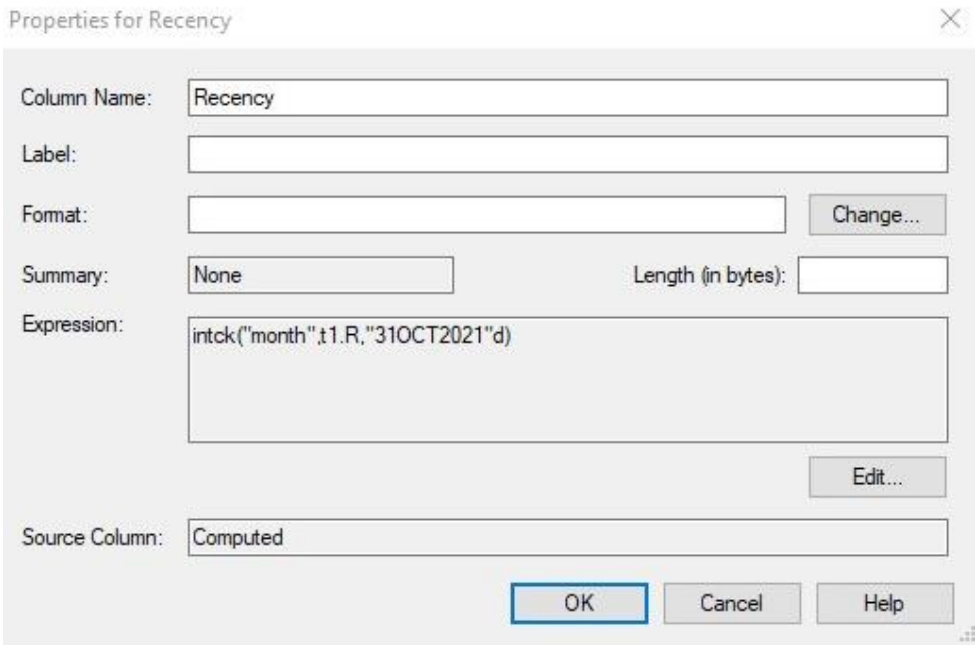
Για την δημιουργία του dataset RFM πρέπει να δημιουργήσουμε τις μεταβλητές Recency, Frequency, Monetary. Έτσι, το πρώτο βήμα είναι να δημιουργήσουμε ένα Query με τις μεταβλτές CustomerID, Purchase_Date και Total.




*Εικόνα 8: Query Builder για την δημιουργία του dataset RFM*

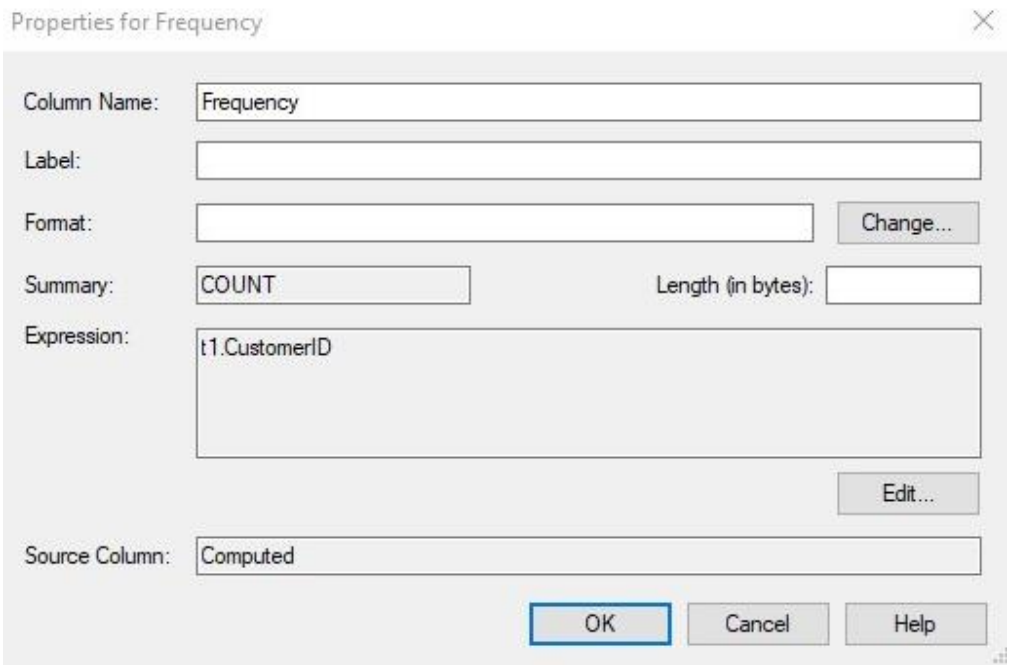
Για την δημιουργία της μεταβλητής Recency χρειάστηκε να υπολογίσουμε το χρονικό διάστημα μεταξύ της ημερομηνίας της τελευταίας συναλλαγής που εμφανίζεται στο dataset (31 Οκτωβρίου 2021) και της ημερομηνίας που ο κάθε πελάτης πραγματοποίησε την τελευταία του συναλλαγή.

Υστέρα από αρκετές δοκιμές με τα δεδομένα διαπιστώθηκε προτιμότερο η συγκεκριμένη μεταβλητή να αντιπροσωπεύει τις εβδομάδες που έχουν περάσει από την τελευταία συναλλαγή του πελάτη. Φυσικά έγιναν δοκιμές τόσο σε ημέρες όσο και μήνες αλλά τα καλύτερα αποτελέσματα ήταν όταν υπολογίζαμε εβδομάδες. Σε αυτήν την απόφαση σημαντικό ρόλο είχε το μικρό χρονικό διάστημα των δεδομένων.



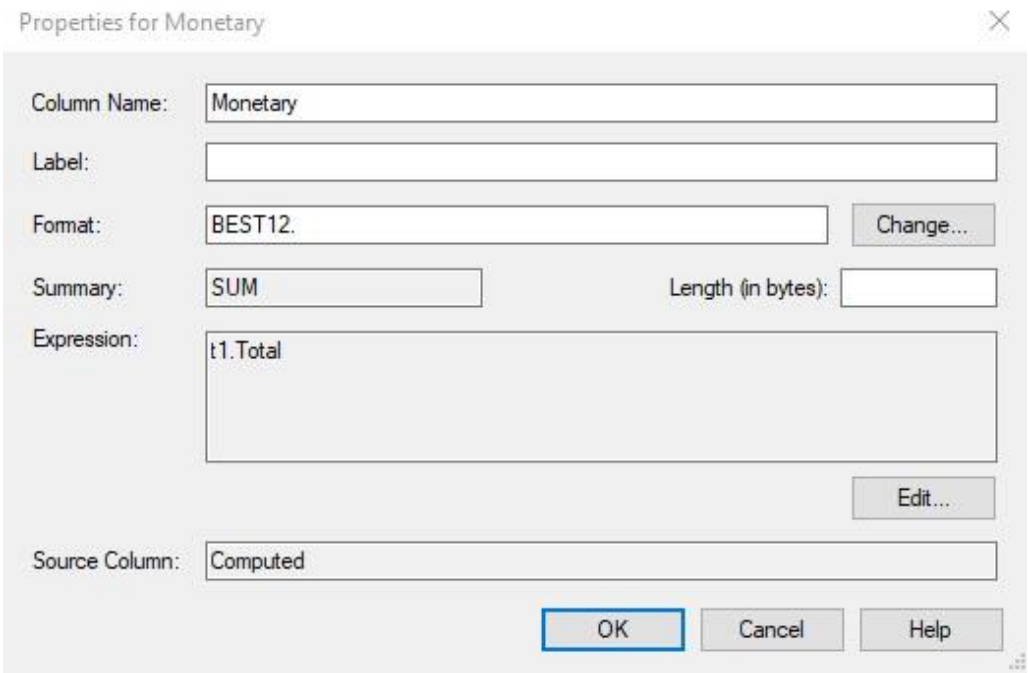

*Εικόνα 9: Σχηματισμός Recency-SAS Enterprise Guide*

Η μεταβλητή Frequency αντιπροσωπεύει την συχνότητα με την οποία οι πελάτες του καταστήματος αγοράζουν από το σύνολο των δεδομένων. Για την δημιουργία της μεταβλητής αυτής χρησιμοποιήσαμε την στήλη CustomerID. Μετρήσαμε δηλαδή τον αριθμό διαφορετικών παραγγελιών που αντιστοιχούν σε κάθε πελάτη.



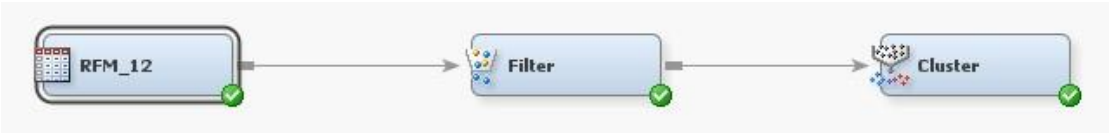

*Εικόνα 10: Σχηματισμός Frequency -SAS Enterprise Guide.*

Τέλος δημιουργήσαμε την μεταβλητή Monetary με την βοήθεια της στήλης Total. Επί της ουσίας υπολογίσαμε την αξία κάθε αγοράς του κάθε πελάτη και έτσι καταλήξαμε στο χρηματικό ποσό που έχει δαπανήσει συνολικά.




*Εικόνα 11: Σχηματισμός Monetary -SAS Enterprise Guide*

Εν συνεχεία το αρχείο εισήχθη στο Enterprise Miner και συνδέθηκε με το κόμβο Filter, με την βοήθεια του οποίου θα γίνει η εξαίρεση των ακραίων τιμών, και αυτός με την σειρά του με τον κόμβο Cluster, που είναι υπεύθυνος για το Clustering των πελατών.



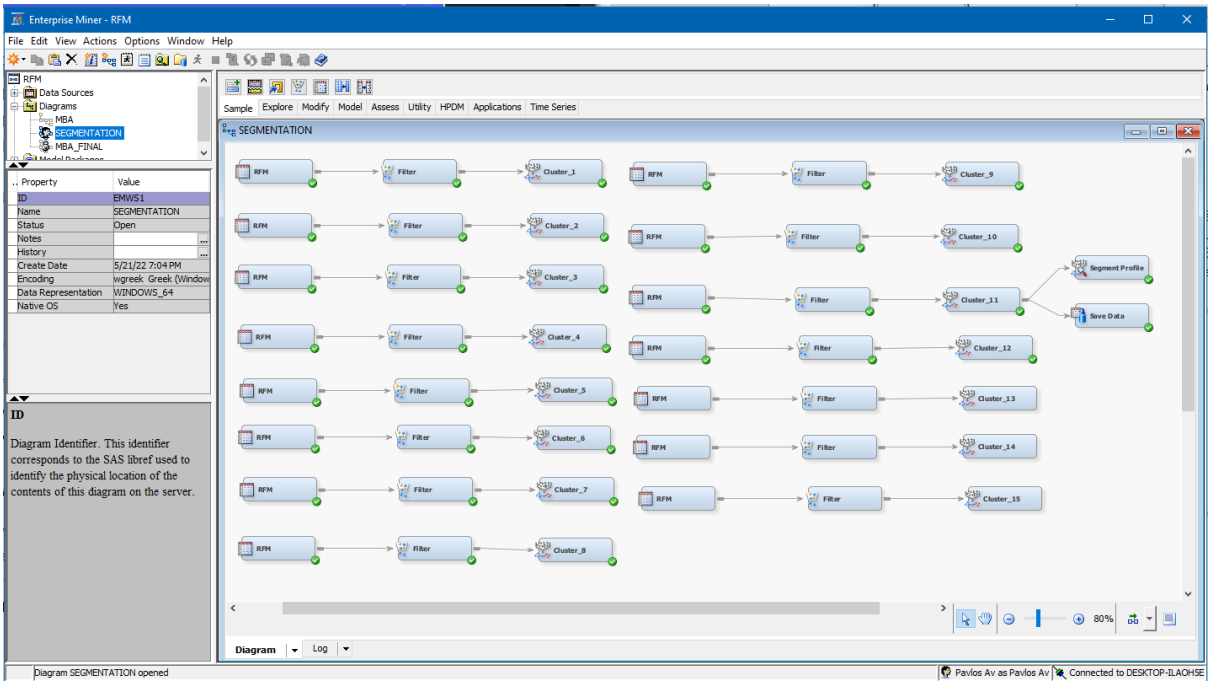

*Εικόνα 12 Σύνδεση RFM dataset με nodes Filter και Cluster*

Κατά την εισαγωγή του αρχείου στο Miner ορίσαμε την μεταβλητή CustomerID ως ID ενώ τις Recency, Frequency, Monetary ως εισακτέες για την ανάλυση μεταβλητές.

| Name | Role | Level |
|---|---|---|
| CustomerID | ID | Nominal |
| F | Input | Interval |
| M | Input | Interval |
| R | Input | Interval |

*Εικόνα 13: Χαρακτηρισμός μεταβλητών κατά την εισαγωγή τους στο Miner*

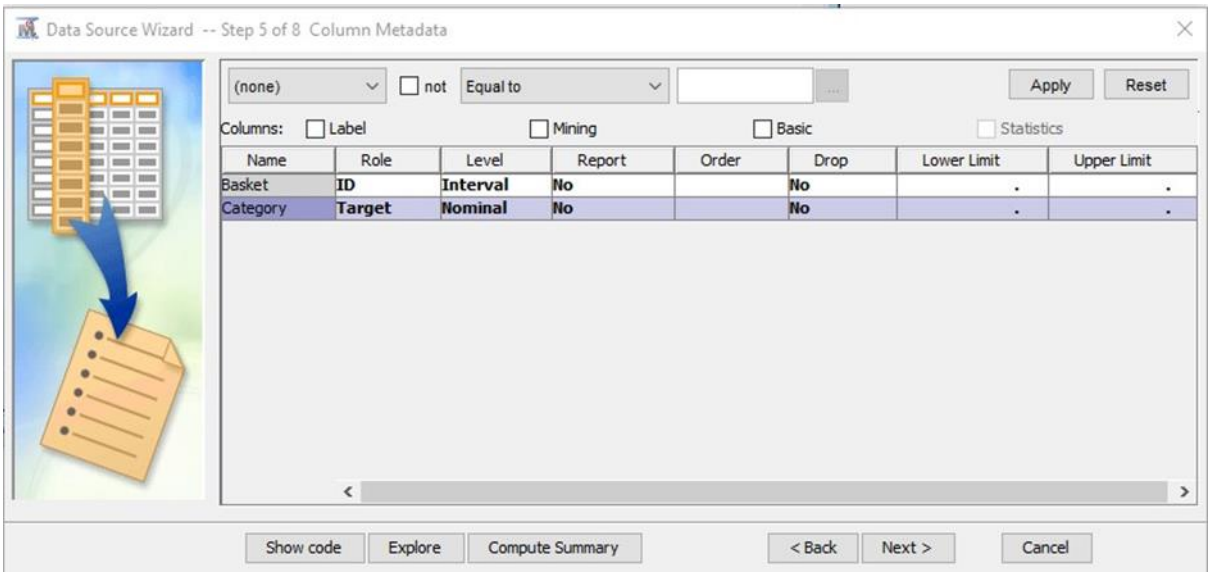
Όπως φαίνεται και στην Εικόνα 13 έγιναν αρκετές δοκιμές τόσο θέτοντας τον αλγόριθμό να τρέξει αυτόματα ώστε να καθορίσει τον κατάλληλο αριθμό clusters όσο και επεμβαίνοντας στον αλγόριθμο θέτοντας συγκεκριμένο επιθυμητό αριθμό clusters. Το τελικό αποτέλεσμα προέκυψε αφού ορίσαμε επιθυμητό αριθμό




*Εικόνα 14: Το Diagram Flow με τα dataset που δημιουργήθηκαν και εξετάστηκαν*

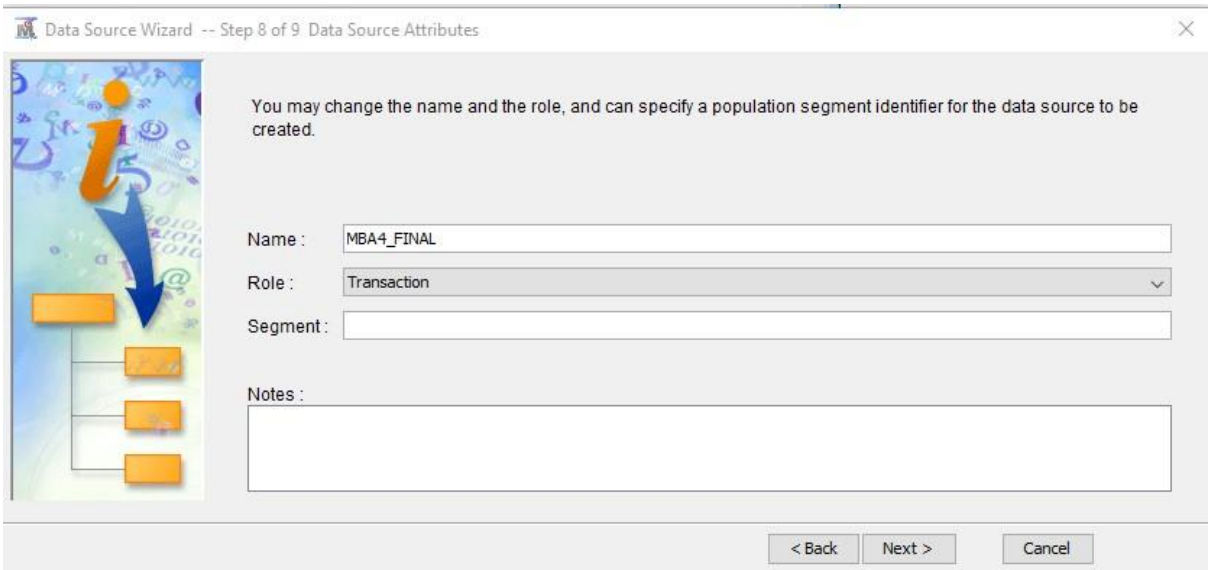
# 8.2 Market Basket Analysis

Για την διεξαγωγή της ανάλυσης καλαθιού αγοράς το πρώτο βήμα είναι να εισάγουμε τα δεδομένα στο Miner πατώντας File → New → Data Source. Στην συνέχεια επιλέγουμε το dataset από την βιβλιοθήκη και θέτουμε ρόλο για κάθε μεταβλητή του. Στην περίπτωσή μας ορίζουμε την μεταβλητή Basket ως ID και την Category ως Target.




*Εικόνα 15: Ορισμός ρόλων για τις μεταβλητές Basket και Category (Βήμα 5)*

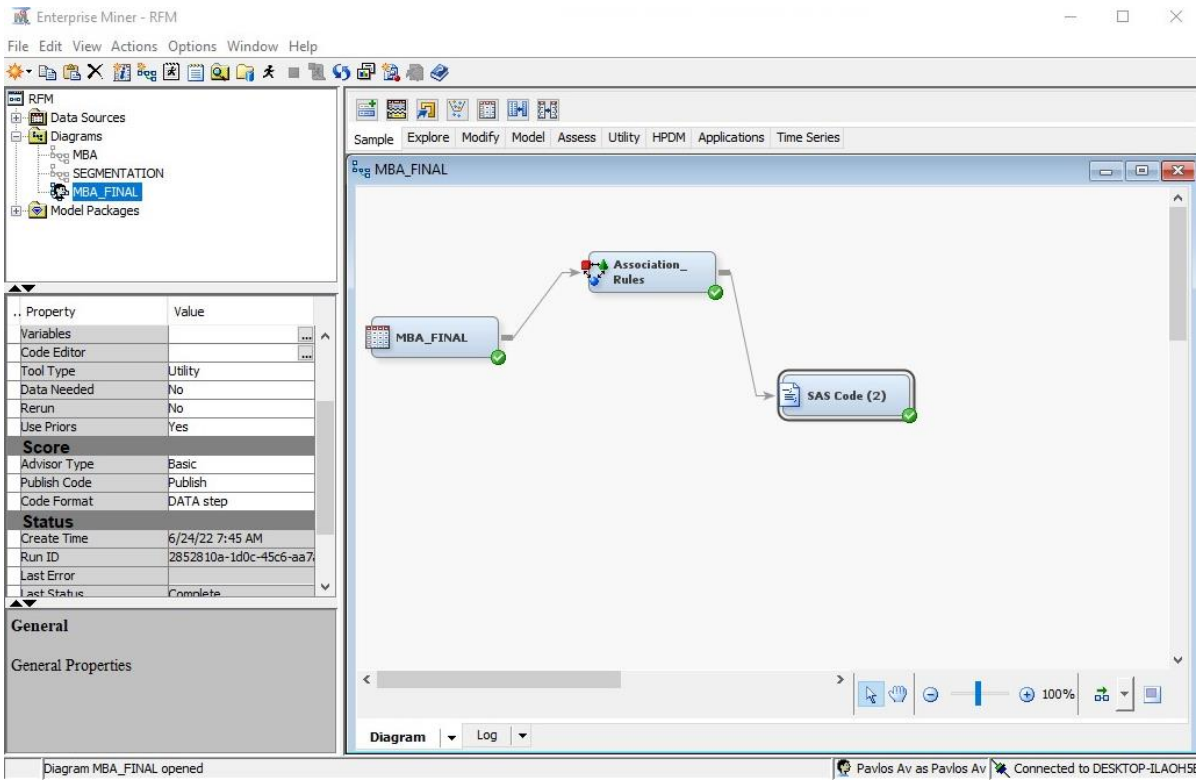
Το επόμενο σημαντικό βήμα είναι το 8 εκεί ορίζουμε ως ρόλο του dataset την επιλογή Transaction.




*Εικόνα 16: Ορισμός ρόλου για το dataset (Βήμα 8)*

Αφού ολοκληρωθεί η εισαγωγή του dataset το ενώνουμε με το Association node και τέλος με το SAS code node. Στο association node έχουμε αφήσει τις πρωταρχικές επιλογές που δίνονται δηλαδή: Maximum items= 2, Maximum Confidence= 5, Support Percentage= 5. Η επιλογή που πρέπει να

αλλάξουμε είναι η Export by Rule = ‘Yes’. Στο SAS code node πατώντας το στο Code Editor από τις ρυθμίσεις του node γράφουμε τον κώδικά για τον στατιστικό έλεγχο Chi-squared.




*Εικόνα 17: Ένωση του dataset με τα node Association Rules και SAS Code*

# 9

# *Επίλογος*

## 9.1 Σύνοψη

Ο όγκος των δεδομένων που συλλέγεται καθημερινά από ένα ηλεκτρονικό κατάστημα είναι χωρίς αμφιβολία τεράστιος. Βασικός στόχος κάθε επιχείρησης πρέπει να είναι η αξιοποίηση των δεδομένων αυτών με τρόπο ώστε να εξάγει συμπεράσματα για τους πελάτες της αλλά και τον τρόπο λειτουργίας της που θα την κάνουν να βελτιωθεί και να γίνει πιο ανταγωνιστική. Με την χρήση τεχνικών εξόρυξης η επιχείρηση μπορεί να τμηματοποιεί τους πελάτες της αλλά και να εξάγει συσχετίσεις που μπορεί να υπάρχουν ανάμεσα σε ορισμένα από τα προϊόντα της. Έτσι θα μπορέσει να δημιουργήσει εξατομικευμένες στρατηγικές προς τους πελάτες της.

Στην παρούσα διπλωματική δουλέψαμε με τα δεδομένα ενός ηλεκτρονικού καταστήματος με την έδρα της επιχείρησης στη Λάρισα. Για την εξόρυξη των δεδομένων έγινε χρήση του μοντέλου RFM (Recency, Frequency, Monetary) με σκοπό την τμηματοποίηση των πελατών σε ομάδες βάσει κοινών χαρακτηριστικών τους. Με αυτόν τον τρόπο η επιχείρηση θα μπορέσει να γνωρίσει καλυτέρα τους πελάτες της και την αγοραστική τους συμπεριφορά. Έτσι με μια πιο πελατοκεντρική προσέγγιση θα μπορέσει να δημιουργήσει στοχευμένες στρατηγικές και να εφαρμόσει σχέδια marketing τόσο σε πολύτιμους πελάτες όσο και σε πελάτες που έχουν καιρό να επισκεφτούν το ηλεκτρονικό κατάστημα. Έτσι θα μπορέσει να ενισχύσει τις σχέσεις της με τους πελάτες, να προσελκύσει κάποιους που έφυγαν αλλά και να δει σημεία στον τρόπο λειτουργίας της που πρέπει να αλλάξει. Στην συνέχεια της εξόρυξης των δεδομένων εφαρμόστηκε η Market Basket Analysis (MBA) για την εύρεση κανόνων συσχέτισης μεταξύ των προϊόντων, αν και στην δική μας περίπτωση χρησιμοποιήθηκαν οι κατηγορίες στην θέση των προϊόντων.

Αρχικά χρησιμοποιώντας της μεταβλητές CustomerID, Purchase_Date και Total από τα δεδομένα που μας δόθηκαν από την επιχείρηση για το ηλεκτρονικό της κατάστημα δημιουργήθηκαν 5 clusters πελατών: Champions, Churners, Recent, Valuable και Worst. Τα αποτελέσματα βοήθησαν την επιχείρηση να γνωρίσει καλύτερα την πελατεικαή της βάση και τον τρόπο υσμπεριφοράς τους.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Στην συνέχεια εφαρμόσαμε τον αλγόριθμο Apriori και με την βοήθεια των μεταβλητών Basket και Category καταλήξαμε σε ορισμένους κανόνες συσχέτισης μεταξύ των κατηγοριών που επιλέγουν οι πελάτες. Οι κανόνες αυτοί που θεωρήθηκαν στατιστικά σημαντικοί κρίνοντας τις υψηλές τιμές τους σε support, confidence και lift είναι οι Πουκάμισα → Παντελόνια, Φούστες → Μπλούζες & Top και Αξεσουάρ → Πουκάμισα.

## 9.2 Συμπεράσματα

Γνωρίζοντας καλύτερα τους πελάτες του ηλεκτρονικού καταστήματος είδαμε πως υπάρχει μια ομάδα πελατών και πιο συγκεκριμένα η Champions, που ενώ δεν αποτελείται από πολλά μέλη, για την ακρίβεια μόλις 97 πελάτες, έχει τις καλύτερες τιμές και στις τρεις μεταβλητές RFM. Από την άλλη πλευρά εντοπίσαμε ομάδες πελατών που αποτελούν τα μεγαλύτερα clusters όπως οι Worst που έχουν και τρεις μεταβλητές RFM είναι κάτω από το μέσο πελάτη ή τους Recent που ενώ επισκέφτηκαν πρόσφατα το ηλεκτρονικό κατάστημα δεν έχουν μεγάλη συχνότητα επισκέψεων ούτε αφήνουν μεγάλα χρηματικά ποσά. Αυτό που παρατηρήθηκε είναι ότι το ηλεκτρονικό κατάστημα έχει έλλειψη loyal πελατών. Αν και το cluster των Valuable πελατών τείνει να μοιάσει σε loyal πελάτες η χαμηλή τιμή, κοντά στον μέσο πελάτη, Frequency πρέπει να βελτιωθεί για να το επιτύχει. Για αυτόν ακριβώς τον λόγο οι προτάσεις marketing που διατυπώθηκαν απευθύνονται σε κάθε cluster ξεχωριστά και στόχο έχουν την αύξηση των εσόδων της επιχείρησης, την δημιουργία αφοσιωμένων πελατών αλλά και βελτίωση της λειτουργείας του ηλεκτρονικού καταστήματος. Ορισμένες από τις προτάσεις που έγιναν ήταν η αποστολή εκπτωτικών κουπονιών μέσω ηλεκτρονικού ταχυδρομείου, cross selling πωλήσεις προϊόντων και η απόκτηση feedback από τους λιγότερο καλούς πελάτες.

Αν και οι κανόνες που προέκυψαν για τις συσχετίσεις μεταξύ των κατηγοριών συνέβαλαν στην δημιουργία των προτάσεων marketing μπορούμε να πούμε ότι δεν βοήθησαν σημαντικά. Οι περισσότεροι από του πελάτες έχουν αγοράσει μια μόνο φορά και ένα μόνο προϊόν οπότε έπρεπε να αφαιρεθούν από την διαδικασία. Από τη ανάλυση προέκυψαν 43 κανόνες εκ των οποίων οι 13 δεν έχουν δυνατή συσχέτιση καθώς το lift < 1 ενώ από τους εναπομείναντες 3 είναι αυτοί που έδειξαν δυνατή συσχέτιση και είναι στατιστικά σημαντικοί.

## 9.3 Περιορισμοί

Στην παρούσα διπλωματική οι περιορισμοί και οι δυσκολίες που αντιμετωπίσαμε ήταν ο μεγάλος αριθμός παραγγελιών που περιείχαν μόνο ένα προϊόν, ο μεγάλος αριθμός προϊόντων που είχαν αγοραστεί μόνο μια φορά και ο επίσης μεγάλος αριθμός πελατών που είχε αγορές μια ή δυο φορές. Φυσικά, αυτό μπορεί να οφείλεται στο γεγονός ότι εξετάζουμε ένα διάστημα 7 μηνών και τα δεδομένα δεν ήταν αρκετά. Όλα τα παραπάνω χαρακτηριστικά δυσκόλεψαν τις τεχνικές εξόρυξης που χρησιμοποιήθηκαν.

## 9.4 Επεκτάσεις

Η παρούσα διπλωματική θα μπορούσε να επαναληφθεί εξετάζοντας ένα μεγαλύτερο χρονικό διάστημα με περισσότερα δημογραφικά και συμπεριφορά δεδομένα για τον κάθε πελάτη από την επιχείρηση. Ως αποτέλεσμα, θα μπορούσαν να δημιουργηθούν πιο λεπτομερή clusters έτσι ώστε η επιχείρηση να γνωρίσει καλύτερα τους πελάτες της. Φυσικά θα μπορούσε να γίνει σύγκριση ανάμεσα στα αποτελέσματα των δυο αυτών αναλύσεων, ελέγχοντας έτσι αν οι προτάσεις marketing που προτείναμε είναι σωστές ή αν χρειάζονται αλλαγές.

Κάτι ακόμη που θα μπορούσε να κάνει η επιχείρηση είναι forecasting στις πωλήσεις όλων των καταστημάτων τόσο του ηλεκτρονικού όσο και των φυσικών της. Βέβαια για επιτευχθεί κάτι τέτοιο απαιτούνται περισσότερα δεδομένα από την επιχείρηση καθώς αυτά που έχουμε στην διάθεσή μας δεν επαρκούν. Στόχος της πρόβλεψης της ζήτησης είναι η επιχείρηση να μπορέσει να κάνει αποτελεσματικότερες προωθητικές ενέργειες στο καταναλωτικό κοινό της, δεν θα μένει ποτέ out of stock χωρίς να μπορεί να εξυπηρετήσει την ζήτηση της αγορά αλλά ούτε θα της μένει μεγάλο απόθεμα αυξάνοντας έτσι τα κόστη για την αποθήκευση. Γνωρίζοντας τις ανάγκες της αγοράς η επιχείρηση θα μπορέσει να αυξήσει την κερδοφορία της και να γίνει πιο ανταγωνιστική. Για να επιτευχθεί όμως κάτι τέτοιο θα χρειαστεί η επιχείρηση να προσθέσει ένα πληροφοριακό σύστημα CRM το οποίο θα συγκεντρώνει δεδομένα από όλα τα καταστήματα. Στην συνέχεια τα δεδομένα αυτά θα μπορούσαν να γίνουν import στο SAS Forecast Server με την βοήθεια ενός module που θα κάνει όλη την δουλειά. Για να έχει αποτέλεσμα θα πρέπει να συγκεντρωθεί ένας μεγάλος όγκος δεδομένων και αυτό απαιτεί χρόνο.

# 10

# *ΒΙΒΛΙΟΓΡΑΦΙΑ*


Abirami, M., & Pattabiraman, V. (2016). Data Mining Approach for Intelligent Customer Behavior Analysis for a Retail Store. Smart Innovation, Systems and Technologies, 283–291. DOI:10.1007/978-3-319-30348-2_23

Agrawal, R., & Imielinski, T. (1993). Mining Association Rules between Sets of Items in Large Databases. Conference on Management Data, 207-216.

Agrawal, R., & Ramakrishnan, S. (1994). Fast Algorithms for Mining Association Rules. Santiago: Proceedings of the 20th VLDB Conference.

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

Agrawal, R., Imielinski, T., Swani, A., (1993). Mining association rules between sets of items in large databases. In: Proceedings of the ACM-SIGMOD Conference on Management of Data, Washington DC, USA.

Ainslie, A., & Rossi, P. E. (1998). Similarities in Choice Behavior Across Product Categories. Marketing Science, 17(2), 91-106.

Aksoy, R. (2008). İnternet Ortamında Pazarlama, Seçkin Yayıncılık, Ankara

Alvarez, A.S. (2003). Chi-squared computation for association rules: preliminary results. Technical Report BC-CS-2003-01.

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

Andrews, R. L., & Currin, I. S. (2002). Identifying segments with identical choice behaviors across product categories: An Intercategory Logit Mixture model. International Journal of Research in Marketing, 19, 65-79.

Anitha, P., & Patil, M. (2019). RFM model for customer purchase behavior using K-means algorithm. Journal of King Saud University - Computer and Information Sciences, In Press, Corrected Proof. https://doi.org/10.1016/j.jksuci.2019.12.011

Armstrong, G. Adam, S. Denize, S., Kotler P.,(2014). Principles of Marketing. Pearson Australia

Bala, P. K. (2008). "Exploring Various Forms of Purchase Dependency in Retail Sale", In Proceedings of the World Congress on Engineering and Computer Science 2008, San Francisco, USA, 1101-1104.

Bell, D. R., & Boztu, Y. (2007). The positive and negative effects of inventory on category purchase: An empirical analysis. Marketing Letters, 18, 1-14.





ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Berry, M., & Linoff, G. (1997). Data Mining Techniques for Marketing, Sales and Customer Support. John Wiley & Sons.

Böcker F, Merkle E (1975) Die Analyse des Sortimentsverbundes. In: Böcker F, Dichtl E (Hrsg) Erfolgskontrolle im Marketing. S 179–191

Bodapati, A.V. (2008). Recommendation Systems with Purchase Data, Journal of Marketing Research 45(1)

Boztug, Y., & Reutterer, T. (2006). A Combined Approach for Segment- Specific Analysis of Market Basket Data

Bramer, M. (2016). Principles of Data Mining, Third Edition, Springer.

Brin, S., Motwani, R., Ullman, J.D., Tsur, S., (1997a). Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the ACM-SIGMOD Conference on Management of Data, Tucson, Arizona, USA.

Chang EC, Huang SC, Wu HH (2010). Using K-means method and spectral clustering technique in an outfitter's value analysis. Qual Quant., 44(4): 807-815

Chen YL, Tang K, Shen RJ, Hu YH (2005) Market Basket Analysis in a Multiple Store Environment. Decis Support Syst 40(2):339–354

Chen, Y.-L., Kuo, M.-H., Wu, S.-Y., & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electronic Commerce Research and Applications, 8(5), 241–251. DOI:10.1016/j.elerap.2009.03.002

Ching, W. K., Pong, M. K. (2002). Advances in Data Mining and Modeling, World Scientific, 1st Edition, Hong Kong, China.

Dalal, M. K. (2012). "Automatic Text Classification of Sports Blog Data", Computing, Communications and Applications Conference, Hong Kong, China, 219-222.

Dippold, K., & Hruschka, H. (2010). Variable Selection for Market Basket Analysis. University of Regensburg Working Papers in Business,Economics and Management Information Systems

Dippold, K., & Hruschka, H. (2013a). "Variable selection for market basket analysis", Computational Statistics, 28(2), 519-539.

Dogan, O., Ayçin, E., & Bulut, E. (2018). Customer segmentation by using rfm model and clustering methods: A case study in retail industry. International Journal of Contemporary Economics and Administrative Sciences, 8(1), 1–19.

Dorn, M., Hou, W., Che, D., & Jiang, Z. (2008). An empirical study of qualities of association rules from a statistical view point. Journal of Information Processing Systems, 4 (1), 27-31.

Dvorak, J. (2006). E-commerce Strategy For Companies.

Emel, G. G., Taşkın, Ç. (2005). "Veri Madenciliğinde Karar Ağaçları ve Bir Satış Analizi Uygulaması", Eskişehir Osmangazi Üniversitesi Sosyal Bilimler Dergisi, 6, 221-239.

Fader PS, Hardie BGS, Lee KL (2005). RFM and CLV: Using Iso-Value curves for customer base analysis. J. Mark. Res. 415-430.

Gancheva, V. (2013). Market basket analysis of beauty products

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). "Retail business analytics: Customer visit segmentation using market basket data", Expert Systems with Applications, 100, 1-16

Gupta, R. (2013). Finding what to Sell Next to your Customer. from http://blogs.mydbsync.com/finding-what-to-sell-next-to-your-customer/

Han, J., Kamber, M., Pei, J. P. (2012). Data Mining Concepts and Techniques, Morgen Kaufmann Publishing, Third Edition, USA.

Han, J., Pei, J., Yin, Y., (2000). Mining frequent patterns without candidate generation. In: Proceedings of the ACM-SIGMOD Conference on Management of Data, Dallas, Texas, USA.

Hormozi, A. M., Giles, S. (2004). Data mining: A competitive weapon for banking and retail industries. Information Systems Management, 21(2), 62-71.

Hudairy, H. (2004). Data mining and decision making support in the governmental sector, Master Thesis, Faculty of Graduate School of the University of Louisville, Kentucky, USA.

Hughes AM (1996). Boosting reponse with RFM. Mark. Tools, 5: 4-10.

InfoCepts. (2004). Business Challenge. Mclean,USA.

Julander. (1992). Basket Analysis: A New Way of Analysing Scanner Data. International Journal of Retail and Distribution Management, 20(7), 10-18.

Jung, J. Y., Bae, J., & Liu, L. (2009). Hierarchical Clustering of Business Process Models. International Journal of Innovative Computing, Information and Control, 5(12), 1-10.

Kabasakal, İ. (2020). Understanding shopping behaviors with category and brand-level market basket analysis. In Tools and Techniques for Implementing International E-Trading Tactics for Competitive Advantage, Editor: Yurdagül Meral, IGI Global, 242-267.

Kaur M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining, Procedia Computer Science 85:78-85

Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. Procedia Computer Science, 3, 1327–1332. DOI:10.1016/j.procs.2011.01.011

Khan, A., Lee, S., Lee, Y.-K., & Khattak, A. (2010). Analyzing Association Rule Mining and Clustering on Sales Day Data with XLMiner and Weka. International Journal of Database Theory and Application, 3(1).

Kooti, F., Lerman, K., Aiello, L. M., Grbovic, M., Djuric, N., & Radosavljevic, V. (2016). Portrait of an online shopper: Understanding and predicting consumer behavior. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, USA, 205-214.

Kotler, P. (1988). Marketing Management: Analysis, Planning, Implementation, and Control. Englewood Cliffs, NJ: Prentice-Hall.

Kruskal, J. (1977). The Relationship between Multidimensional Scaling and Clustering. In J. Van Ryzin (Ed.), Classification and Clustering (pp 17-45). New York, NY: Academic Press Inc

Kumar, N., & Rao, R. (2006). Using Basket Composition Data for Intelligent Supermarket Pricing. Marketing Science, 25(2), 188-199

Lin, W.P, Tsai, S.B, Wu, J., Shi, L., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-means Algorithm.

Mathematical Problems in Engineering, 20. Article ID 8884227.
https://doi.org/10.1155/2020/8884227

Linoff, G. S., Berry, M. J. A. (2011). Data Mining Techniques for Marketing, Sales and Customer Relationship Management, Wiley, Third Edition, Canada.

Lumsden SA, Beldona S, Morison AM (2008). Customer value in an all-inclusive travel vacation club: An application of the RFM framework. J. Hosp. Leisure Mark., 16(3): 270-285

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1 (Univ. of Calif. Press, 1967), 281-297. https://projecteuclid.org/euclid.bsmsp/1200512992

Manchanda P, Ansari A, Gupta S (1999) The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions. Market Sci 18(2):95–114

Mansur, A., &Kuncoro, T., (2012). Product Inventory Predictions at Small Medium Enterprise Using Market Basket Analysis Approach-Neural Networks. Procedia Economics and Finance. International Conference on Small and Medium Enterprises Development with a Theme "Innovation and Sustainability in SME Development" (ICSMED), 4, 312-320

Marakas, G. M. (2003). Decision Support Systems in the 21st Century, Prentice Hall, Second Edition, USA.

Marcus C (1998). A practical yet meaningful apprach to customer segmentation. J. Consum. Mark., 15(5): 494-504.

Mild, A., & Reutterer, T. (2003). An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. Journal of Retailing and Consumer Services, 10, 123-133.

Mosleh, A., Nosratabadi, S., & Bahrami, P. (2015). Recognizing the Business Models types in Tourism Agencies: Utilizing the Cluster Analysis. International Business Research, 8(2), 173-180.

Müller-Hagedorn L (1978) Das Problem des Nachfrageverbundes in erweiterter Sicht. Z betriebswirt Forsch 3:181–193

Nestorov, S., & Jukic, N. (2003). Ad-Hoc Association-Rule Mining within the Data Warehouse. Hawaii: 36th Hawaii International Conference on System Sciences.

Ng, R. T., & Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. Proceedings of the 20th International conference on Very Large Data Bases, 144-155. San Francisco, CA: Morgan Kaufmann Publishers.

Oracle. (2012). 2012 B2B E-Commerce Survey: Results And Trends.

Pepitone, J. (2011). Groupon's latest accounting revision whacks sales in half. CNNMoney Tech.

Prakash, S., & Parvathi, R. (2011). A Novel Method of Mining Association Rule with Multilevel Concept Hierarchy. JCA Journal.

Qualls, B. (2013). Introduction to Market Basket Analysis. AA(07).

Russell GJ, Bell D, Bodapati A, Brown CL, Chiang J, Gaeth G, Gupta S, Manchanda P (1997) Perspectives on multiple category choice. Marketing Letters 8 (3) 297–305

Russell GJ, Ratneshwar S, Shocker AD, Bell D, Bodapati A, Degeratu A, Hildebrandt L, Kim N, Ramaswami S, h Shankar V (1999) Multiple– category decision–making: Review and synthesis. Marketing Letters 10 (3) 319–332

Russell, G. J., Petersen A. (20000. Analysis of cross category dependence in market basket selection. J. Retailing 76(3) 367–392.

Savaş, S., Topaloğlu, N., Yılmaz, M. (2012). "Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri", İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21, 1-23

Savasere, A., Omiecinski, E., Navathe, S., 1995. An efficient algorithm for mining association rules in large databases. In: Proceedings of the 21st VLDB Conference, Zurich, Swizerland.

Segal, M. N., & Giacobbe, R. W. (1994). Market Segmentation and Competitive Analysis for Supermarket Retailing, International Journal of Retail & Distribution Management, 22, 38- 48.

Shapiro, G., & Frawley, W. (1991). Knowledge Discovery in Databases. MIT Press

Shekhar, S., & Chawla, S. W. (2003). Spatial Databases: A Tour. Upper Saddle River, NJ: Prentice Hall.

Sherman, E., Mathur, A., Smith, R.B.(1997). "Store Environment and Consumer Purchase Behavior: Mediating Role of Consumer Emotions," Psychol. Mark., vol. 14, no. 4, pp. 361–378 Service Systems Design: Literature Review and Research Questions," Procedia CIRP, vol. 64, pp. 306–311, 2017.

Shocker, A. D., B. L. Bayus, and N. Kim. 2004. Product complements and substitutes in the real world: The relevance of other products. Journal of Marketing 68 (1): 28–40.

Smith, W. (1956) Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing, 21, 3-8.

Sota, S., Chaudhry, H., Chamaria, A., Chauhan, A. (2018). "Customer relationship management research from 2007 to 2016: An academic literature review", Journal of Relationship Marketing, 17(4), 277-291.

Stone, B.,& Jacobs, R. (1995).Successful Direct Marketing Methods. Lincolnwood, IL: NTC Business Books. 37–57.

Strauss, J., Frost, R. (2009). E-Marketing, Pearson Education, New Jersey

Su, X. (2007). "Intertemporal Pricing with Strategic Customer Behavior," Manage. Sci., vol. 53, no. 5, pp. 726–741]

Thiprungsri, S., & Vasarhelyi, M. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. The International Journal of Digital Accounting Research, 11, 69-84

Toivonen, H., 1996. Sampling large databases for association rules. In: Proceedings of the 22nd VLDB Conference, Mumbai (Bombay), India

Vahaplar, A., İnceoğlu, M. M. (2001). "Veri Madenciliği ve Elektronik Ticaret", VII. Türkiye'de İnternet Konferansı, İstanbul.

Wang P. (2007). Evaluate E-business Models.

Wedel, M., & Kamakura W. (2000). Market Segmentation: Conceptual andMethodological Foundations (2nd edition), Kluwer Academic Publishers, Boston

Weill, P and Malone, T.W. (2005) Do Some Business Models Perform Better Than Others? A study of the 1000 largest US firms.

Winer, R. S. (2001), "A framework for customer relationship management", California Management Review, Vol. 43 No. 4, pp. 89-105.

Zaki, M.J., (2000). Scalable algorithms for association mining. IEEE Trans. Knowledge Data Eng. 12 (3), 372–390.

Zhang, Y., Pennacchiotti, M. (2013). Predicting purchase behaviors from social media. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro Brazil, 1521-1532.

Τσίπτσης Κ. & Χωριανόπουλος, Α. (2009). Data Mining Techniques in CRM: Inside Customer Segmentation, ResearchGate

# 11

# Παράρτημα Ι: Λίστα Πινάκων

Πίνακας 1: Summary Stat RFM αρχείο ..... 41
Πίνακας 2: Μεταβλητή Recency-Extreme Values & Quantiles ..... 42
Πίνακας 3: Μεταβλητή Frequency-Extreme Values & Quantiles ..... 43
Πίνακας 4: Μεταβλητή Monetary-Extreme Values & Quantiles ..... 44
Πίνακας 5: Εύρος τιμών μεταβλητών RFM μετά την αφαίρεση ακραίων τιμών ..... 46
Πίνακας 6: Η μέση τιμή των RFM ανά συστάδα σε σχέση με την συνολική μέση τιμή ..... 46
Πίνακας 7: Κανόνες Προϊόντων που προέκυψαν ..... 48
Πίνακας 8: Η κατεύθυνση των τιμών RFM σε σχέση με το μέσο όρο του δείγματος ..... 49
Πίνακας 9: Χαρακτηρισμός των clusters σύμφωνα με τις τιμές των RFM ..... 50
Πίνακας 10: Σημαντικότητα των μεταβλητών RFM στο διαχωρισμό των clusters ..... 59
Πίνακας 11: Αποτελέσματα chi-squared ανάλυσης ..... 60

# 12

# Παράρτημα ΙΙ: Λίστα Γραφημάτων

Γράφημα 1: Συστάδες που λαμβάνονται από τον αλγόριθμο συμπλέγματος DBSCAN ..... 18
Γράφημα 2: Συστάδες που προέκυψαν από τον αλγόριθμο του Gaussian Mixture Model..... 21
Γράφημα 3: Τα 10 καλύτερα προϊόντα σε πωλήσεις ..... 37
Γράφημα 4: Οι 10 καλύτερες κατηγορίες σε πωλήσεις ..... 38
Γράφημα 5: Πωλήσεις ανά τομέα ..... 38
Γράφημα 6: Τρόποι πληρωμής των πελατών του ηλεκτρονικού καταστήματος ..... 39
Γράφημα 7: Κατανομή μεταβλητής Recency ..... 42
Γράφημα 8: Κατανομή μεταβλητής Frequency ..... 43
Γράφημα 9: Κατανομή μεταβλητής Monetary..... 44
Γράφημα 10: Παραγγελίες που περιέχουν μόνο ένα προϊόν..... 47
Γράφημα 11: Προϊόντα με missing values..... 47
Γράφημα 12: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster ‘Champions’ ..... 53
Γράφημα 13: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster ‘Churners’..... 54
Γράφημα 14: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster ‘Recent’ ..... 55
Γράφημα 15: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster ‘Valuable’..... 56
Γράφημα 16: Ποσοστά συχνότητας για τις κατηγορίες προϊόντων του cluster ‘Worst’..... 57
Γράφημα 17: Scatterplot RFM ανά Cluster .....**Σφάλμα! Δεν έχει οριστεί σελιδοδείκτης.**
Γράφημα 18: Δένδρο αποφάσεων..... 59

# 13

# Παράρτημα ΙΙΙ: Λίστα Εικόνων

Εικόνα 1: Τύποι σημείων DBSCAN ..... 17
Εικόνα 2: Παράδειγμα ομαδοποίησης βάσει διανομής ..... 19
Εικόνα 3: Παράμετροι Μοντέλου μείγματος Gauss ..... 20
Εικόνα 4: Clusters που προέκυψαν από την 1η προσπάθεια ..... 45
Εικόνα 5: Τελικά Clusters ..... 45
Εικόνα 6: Η αρχική οθόνη του SAS Enterprise Guide ..... 61
Εικόνα 7: Import αρχείου Excel στο Enterprise Guide ..... 62
Εικόνα 8: Query Builder για την δημιουργία του dataset RFM ..... 62
Εικόνα 9: Σχηματισμός Recency-SAS Enterprise Guide ..... 63
Εικόνα 10: Σχηματισμός Frequency -SAS Enterprise Guide ..... 64
Εικόνα 11: Σχηματισμός Monetary -SAS Enterprise Guide ..... 64
Εικόνα 12 Σύνδεση RFM dataset με nodes Filter και Cluster ..... 65
Εικόνα 13: Χαρακτηρισμός μεταβλητών κατά την εισαγωγή τους στο Miner ..... 65
Εικόνα 14: Το Diagram Flow με τα dataset που δημιουργήθηκαν και εξετάστηκαν ..... 65
Εικόνα 15: Ορισμός ρόλων για τις μεταβλητές Basket και Category (Βήμα 5) ..... 66
Εικόνα 16: Ορισμός ρόλου για το dataset (Βήμα 8) ..... 66
Εικόνα 17: Ένωση του dataset με τα node Association Rules και SAS Code ..... 67

# 14

# Παράρτημα IV: Appendix

```
*\\ Set macro variable transN to total number of market baskets (or
transactions);
proc sql noprint;
       select count(distinct(%EM_ID))
       into :transN
       from &EM_IMPORT_TRANSACTION
       ;
quit;

*\\ Create a dataset that will add Chi Square statistic to existing Rules;
data work.ChiSquare;
       *\\ Start with Association Node Rules output table;
       set &EM_IMPORT_RULES;

       *\\ add Chi Square statistic to each rule;
       CHISQ = .;  *missing in case of divide by zero scenario;
       PVALUE = .;
       if NOT (CONF=SUPPORT or LIFT=CONF) then do;
              CHISQ = (&transN * (LIFT-1)**2)*((SUPPORT/100) * (CONF/100))
                           /
                           ((CONF/100 - SUPPORT/100) * (LIFT - CONF/100));
              PVALUE = 1-Probchi(CHISQ,1);
       end;

       *\\ keep desired columns;
       keep index rule exp_conf conf support lift chisq pvalue;

run;

proc sort data=work.ChiSquare;
  by descending lift;
run;

proc print;
run;

*\\ The Chi Square table is currently in the temporary Work library;
*\\ The following code saves the table to YOUR PROJECT LIBRARY;
* data <yourPrjLibName>.ChiSquare;
*        set work.ChiSquare;
* run;
```