ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Εφαρμογή τεχνικών εξόρυξης γνώσης για την αναγνώριση των αγοραστικών αποστολών των καταναλωτών»

Αναστασία Γρίβα

Αριθμός Μητρώου:
M312011

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2014

ΜΕΤΑΠΤΥΧΙΑΚΟ ΔΙΠΛΩΜΑ ΕΙΔΙΚΕΥΣΗΣ (MSc)
στα ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Εφαρμογή τεχνικών εξόρυξης γνώσης για την αναγνώριση των αγοραστικών αποστολών των καταναλωτών»

Αναστασία Γρίβα

Αριθμός Μητρώου:
M312011

Επιβλέπων Καθηγητής:
Παναγιώτης Μηλιώτης

Εξωτερικοί Κριτές:
Κατερίνα Πραματάρη
Κλεοπάτρα Μπαρδάκη

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΑΘΗΝΑ, ΦΕΒΡΟΥΑΡΙΟΣ 2014

Athens University of Economics and Business

Department of Informatics
MSc in Information Systems

Thesis Title:

"A Data Mining-based Framework to Identify Shoppers' Missions"

by

Anastasia Griva

Thesis Advisor:

Panagiotis Miliotis

Committee Members:

Katerina Pramatari
Cleopatra Bardaki

Athens, February 2014

ΕΠΙΤΕΛΙΚΗ ΣΥΝΟΨΗ

Στις μέρες μας η συμπεριφορά των καταναλωτών έχει αλλάξει, οι πελάτες έχουν γίνει πλέον πιο απαιτητικοί. Οι επιχειρήσεις έχοντας αντιληφθεί ότι η ικανοποίηση του πελάτη σχετίζεται άμεσα με την κερδοφορία τους, δεδομένου του ολοένα και αυξανόμενου ανταγωνισμού, έχουν στραφεί σε μια περισσότερο πελατοκεντρική φιλοσοφία (Bull, 2003; Phan & Vogel, 2010). Μάλιστα, προκειμένου να ανταποκριθούν στις ολοένα και αυξανόμενες απαιτήσεις των καταναλωτών, προσπαθούν να αναπτύξουν καινοτόμες μεθόδους διαχείρισης του πελάτη (Anderson, Jolly, & Fairhurst, 2007). Το παραπάνω έρχεται να υποβοηθήσει η χρήση και η ανάπτυξη των νέων τεχνολογιών, όπως Big Data, Επιχειρηματική Ευφυΐα (Business Intelligence - BI), Εξόρυξη Γνώσης (Data Mining - DM) (Provost & Fawcett, 2013). Οι νέες αυτές τάσεις της τεχνολογίας προσφέρουν τη δυνατότητα επεξεργασίας μεγάλου όγκου δεδομένων και την άντληση πολύτιμων πληροφοριών. Το παραπάνω υποβοηθά σημαντικά τη λήψη αποφάσεων. Όπως κάθε επιχείρηση, έτσι και οι λιανέμποροι, έχουν αντιληφθεί την προστιθέμενη αξία που προκύπτει από τη διερεύνηση μεγάλου όγκου δεδομένων με τη χρήση των νέων τεχνολογιών, ως σημαντικό στοιχείο για την ικανοποίηση του πελάτη (Bertino, 2011). Παρόλα αυτά, δεν έχει γίνει αρκετή έρευνα γύρω από διερεύνηση μεγάλου όγκου δεδομένων, ώστε να ανακαλυφθούν τα αγοραστικά μοτίβα των καταναλωτών όταν επισκέπτονται ένα κατάστημα (Wang & Zhou, 2013).

Έχοντας ως κίνητρο το παραπάνω ερευνητικό κενό, αυτή η διπλωματική εργασία προτείνει μια μεθοδολογία, η οποία με τη χρήση τεχνικών DM στα δεδομένα πωλήσεων από το καταστήματα ενός λιανέμπορου, απαντά στο ερώτημα: «Ποιές είναι οι αγοραστικές αποστολές των καταναλωτών όταν επισκέπτονται το συγκεκριμένο κατάστημα;». Με τον όρο αγοραστικές αποστολές εννοούμε το λόγο ή τους λόγους για τους οποίους ένας καταναλωτής επισκέφτηκε το κατάστημα αυτό, για παράδειγμα για να αγοράσει προϊόντα για το πρωινό του, ή για να αγοράσει είδη υγιεινής. Με τη χρήση DM θα αναγνωριστούν οι κατηγορίες των προϊόντων οι οποίες αγοράζονται μαζί. Έπειτα αυτά τα μοτίβα των κατηγοριών, θα βοηθήσουν στην αναγνώριση των αγοραστικών αποστολών.

Στηριζόμενη στη "Design science" μεθοδολογία (Hevner, March, Park, & Ram, 2004), η διπλωματική αυτή ξεκινά με τη διερεύνηση του παραπάνω επιχειρηματικού

προβλήματος, και την ανασκόπηση της σχετικής με αυτό βιβλιογραφίας. Στη συνέχεια, αναπτύσσεται μια μεθοδολογία βασισμένη σε DM τεχνικές και αλγορίθμους, η υλοποίηση της οποίας συμβάλλει στη λύση του επιχειρηματικού προβλήματος. Έπειτα, η μεθοδολογία αυτή τίθεται σε εφαρμογή, ώστε να αξιολογηθεί και να διαπιστωθεί αν πραγματικά ανταποκρίνεται αποτελεσματικά στη λύση του προβλήματος. Τέλος, αναλύεται τόσο η θεωρητική, όσο και η πρακτική συμβολή αυτής της διπλωματικής, ενώ παράλληλα υπογραμμίζονται οι περιορισμοί που προκύπτουν από τη μεθοδολογία, καθώς και προτάσεις για περεταίρω έρευνα.

Πιο συγκεκριμένα, αρχικά αναλύεται λεπτομερώς το ερευνητικό πρόβλημα. Μιας και η ανάγκη για την ικανοποίηση του πελάτη ανήκει στο ευρύτερο πλαίσιο του Customer Relationship Management (CRM), παραθέτονται τα πλεονεκτήματα από τη χρήση αυτού. Το CRM αποτελείται από τέσσερις συνιστώσες, οι οποίες είναι η αναγνώριση, η προσέλκυση, η διατήρηση και η ανάπτυξη του πελάτη (Hosseini, Maleki, & Gholamian, 2010; Ngai, Xiu, & Chau, 2009). Κάθε μια από αυτές, μπορεί να υποβοηθηθεί με τη χρήση διαφόρων DM μοντέλων και τεχνικών. Οι τεχνικές αυτές μπορούν να συμβάλλουν στο CRM βοηθώντας στη βαθύτερη ανάλυση και κατανόηση της συμπεριφοράς του πελάτη. Τα DM μοντέλα που χρησιμοποιούνται για να υποβοηθήσουν το CRM διακρίνονται σε επτά μεγάλες κατηγορίες, αυτές είναι: Association, Classification, Clustering, Sequence discovery, Regression, Forecasting, Visualization (Ngai et al., 2009).

Όπως και σε κάθε άλλο κλάδο, έτσι και στο λιανεμπόριο υπάρχουν αρκετές εφαρμογές βασισμένες στο DM. Το CRM, συνδυασμένο με τεχνικές εξόρυξης γνώσης, μπορεί να βοηθήσει τους λιανέμπορους να αποκτήσουν μια πιο ολοκληρωμένη αντίληψη για τους πελάτες τους, και να ανταποκριθούν στις ολοένα και αυξανόμενες απαιτήσεις τους (Anderson et al., 2007). Ένα γνωστό παράδειγμα είναι αυτό της Tesco, όπου με τη χρήση τεχνικών DM σε δεδομένα που λάμβανε από τις κάρτες πιστότητας, κατάφερε να επαναπροσδιορίσει τις σχέσεις της με τους πελάτες (Humby, Hunt, & Phillips, 2003). Εκτός από τα παραδοσιακά παραδείγματα BI και DM εφαρμογών, στη βιβλιογραφία δεν υπάρχουν πολλές δημοσιεύσεις σχετικές με την ιδέα των αγοραστικών αποστολών. Υπάρχουν ορισμένες δημοσιεύσεις οι οποίες προσπαθούν, με τη χρήση δεδομένων πωλήσεων από σουπερμάρκετ, να αναγνωρίσουν τις συσχετίσεις μεταξύ των προϊοντικών κατηγοριών. Η αναγνώριση αυτή στηρίζεται σε μοντέλα βασισμένα σε "association"

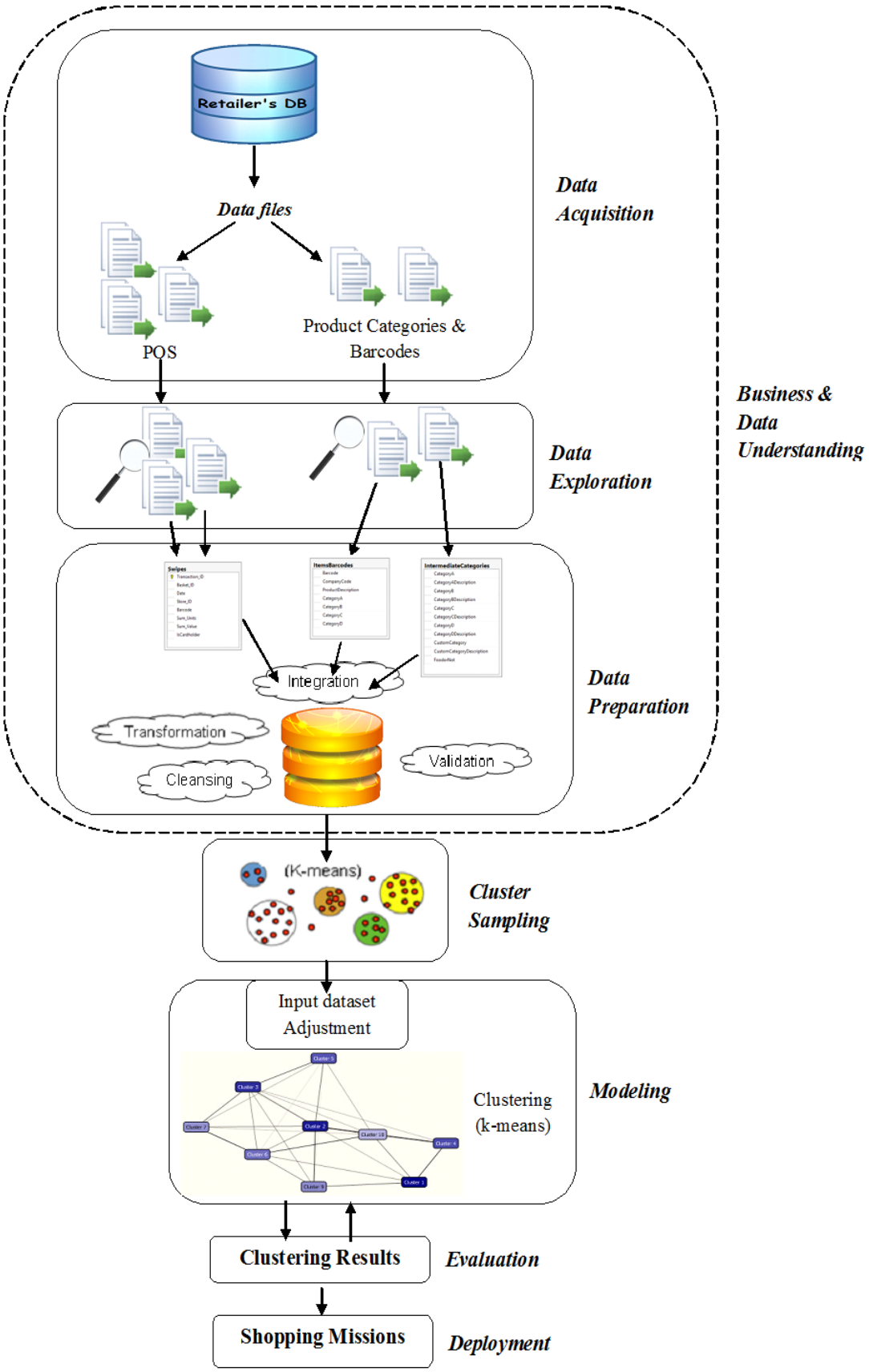
τεχνικές και "apriori" αλγορίθμους ή αλγορίθμους «κοντινότερου γείτονα» (Ahn, 2012; Borges & Babin, 2010; Cil, 2012; Raorane, Kulkarni, & Jitkar, 2012; Shrivastava & Sahu, 2007). Παρόλα αυτά, δεν υπάρχει καμία μεθοδολογία που να δίνει συγκεκριμένες οδηγίες για το πως μπορούν να αναγνωριστούν οι αγοραστικές αποστολές των καταναλωτών, με τη χρήση τεχνικών βασισμένων σε "clustering" και "k-means" αλγορίθμων.

Η προτεινόμενη μεθοδολογία στηρίζεται στην CRISP-DM (Chapman et al., 2000), μια μεθοδολογία για την ανάπτυξη έργων Εξόρυξης Γνώσης. Οι φάσεις της μεθοδολογίας απεικονίζονται στην Εικόνα 1. Η αρχική φάση επικεντρώνεται στην απόκτηση του απαραίτητου συνόλου δεδομένων (Data Acquisition), έχοντας στο μυαλό μας τον επιχειρηματικό μας στόχο, ο οποίος είναι να αναγνωρίσουμε τις αγοραστικές αποστολές. Τα απαραίτητα δεδομένα που πρέπει να αποκτηθούν είναι: (Α) δεδομένα πωλήσεων, όπως προέρχονται από τις αγορές στα ταμεία των σουπερμάρκετ (Point Of Sales - POS), (Β) δεδομένα που αφορούν τα προϊόντα και την ιεραρχία κατηγοριών στις οποίες οι λιανέμποροι τα κατηγοριοποιούν. Οι λιανέμποροι συνηθίζεται να στέλνουν περισσότερα από τα δεδομένα που χρειάζονται για την ανάλυση, καθώς στην εταιρική τους βάση τα δεδομένα είναι κάπως άναρχα. Για το λόγο αυτό θα πρέπει να εξεταστούν αυτά τα δεδομένα (Data Exploration), και να αποφασίσουμε με ποιά από αυτά θα συνεχίσουμε στην ανάλυση. Σημαντικό βήμα κατά τη διερεύνηση των δεδομένων, είναι η δημιουργία νέων προϊοντικών κατηγοριών οι οποίες θα έχουν νόημα για την μετέπειτα ανάλυση. Για παράδειγμα, για την ανάλυση δεν έχει νόημα το προϊόν πορτοκαλάδα μάρκας «χ», αλλά ότι αυτό το προϊόν είναι αναψυκτικό.

Τα αρχικά δεδομένα, όπως είναι λογικό, δε θα είναι σε μορφή έτοιμη για επεξεργασία, ούτε θα είναι «καθαρά», χωρίς λάθη και ασυνέπειες. Επομένως, πρέπει να προετοιμαστούν (Data Preparation) κατάλληλα, ώστε να μπορούν να αναλυθούν. Η προετοιμασία αυτή περιλαμβάνει τη συγχώνευση όλων των δεδομένων σε μια βάση, τον καθαρισμό αυτών από λάθη και ασυνέπειες, και τον μετασχηματισμό τους, ώστε να μπορέσουν να εισαχθούν σε κάποιο DM εργαλείο. Εδώ πρέπει να σημειωθεί ότι είναι σημαντικός ο μετασχηματισμός των δεδομένων από επίπεδο συναλλαγής, σε επίπεδο απόδειξης-καλαθιού. Τέλος, ο έλεγχος ότι οι παραπάνω διεργασίες πραγματοποιήθηκαν ορθά είναι βασικό κομμάτι της προετοιμασίας των δεδομένων. Μετά τα παραπάνω, θα πρέπει να αφαιρεθούν από τα δεδομένα οι ακραίες τιμές. Ο

αποκλεισμός των ακραίων τιμών θα πραγματοποιηθεί με τη χρήση "Cluster Sampling". Ακραίες τιμές είναι καλάθια τα οποία έχουν πολύ μικρό ή πολύ μεγάλο αριθμό προϊόντων, από τα οποία δεν μπορούν να εξαχθούν αγοραστικές αποστολές.

Πριν τη δημιουργία του τελικού μοντέλου, το οποίο θα χρησιμοποιηθεί για να την αναγνώριση των αγοραστικών αποστολών, θα πρέπει να συγχωνευτούν όλα τα δεδομένα που θα χρησιμοποιηθούν στην ανάλυση σε ένα αρχείο ή πίνακα βάσης. Το παραπάνω είναι υποχρεωτικό έτσι ώστε να διασφαλιστεί η διαλειτουργικότητα των δεδομένων, δηλαδή η δυνατότητα ανάλυσής τους από πολλά DM εργαλεία. Μετά αυτός ο πίνακας-αρχείο θα εισαχθεί στο DM εργαλείο, ώστε να δημιουργηθεί το μοντέλο (Modeling Phase). Όπως προαναφέρθηκε, το μοντέλο που θα δημιουργηθεί θα στηρίζεται στην DM τεχνική "clustering", ενώ ο αλγόριθμος που θα χρησιμοποιηθεί είναι ο "k-means". Τα αποτελέσματα του μοντέλου θα είναι ομάδες-clusters από κατηγορίες προϊόντων. Κάθε κατηγορία θα καταλαμβάνει ένα ποσοστό ανάλογα με τη συμβολή της σε κάθε cluster. Οι επικρατούσες σε κάθε cluster κατηγορίες βοηθούν στο χαρακτηρισμό αυτού, ως μια αγοραστική αποστολή. Μετά από ένα πρώτο, γρήγορο χαρακτηρισμό των αγοραστικών αποστολών, θα πρέπει τα αποτελέσματα να παρουσιαστούν στους ανθρώπους που γνωρίζουν το πεδίο, ώστε να διαπιστωθούν τυχόν παραλήψεις-ασυνέπειες. Ύστερα από τυχόν διορθώσεις στο μοντέλο, θα πρέπει να ξαναεκτελεστεί η διαδικασία, τα τελικά αποτελέσματα της οποίας θα πρέπει να αναλυθούν σε βάθος, ώστε να εξαχθούν οι τελικές αγοραστικές αποστολές. Την παραπάνω ανάλυση θα βοηθήσει ο υπολογισμός του μέσου όρου των προϊόντων, αλλά και των κατηγοριών που περιέχονται στα καλάθια που αποτελούν κάθε ένα από τα clusters. Μερικά από τα clusters που προκύπτουν μπορεί να μην γίνεται να χαρακτηριστούν, καθώς μπορεί να περιέχουν καλάθια που αποτελούνται από πολλά προϊόντα, από διάφορες κατηγορίες. Σε αυτά τα clusters μπορεί να υλοποιηθεί εκ νέου clustering, ώστε να διερευνηθεί εάν περιέχουν παραπάνω από μια αγοραστικές αποστολές. Τέλος, τα αόριστα και δύσκολα να χαρακτηριστούν clusters παρατηρούνται σε καταστήματα που περιέχουν πολλές κατηγορίες προϊόντων, και οι πελάτες συνήθως πραγματοποιούν σε αυτά μαζικές, χωρίς συγκεκριμένο στόχο, αγορές. Το παραπάνω αποτελεί ένα σημαντικό περιορισμό της προτεινόμενης μεθοδολογίας.



Εικόνα 1 Προτεινόμενη Μεθοδολογία

Στη συνέχεια, προκειμένου να διαπιστωθεί εάν η μεθοδολογία που προτείνεται ανταποκρίνεται στην πραγματικότητα, χρησιμοποιήθηκαν δεδομένα από οκτώ καταστήματα ενός έλληνα λιανέμπορου. Τα καταστήματα αυτά είχαν ανά δυο κοινά χαρακτηριστικά, χωρίζονταν σε μικρά καταστήματα για γρήγορες αγορές (convenience stores), σε σουπερμάρκετ, σε μικρά υπερκαταστήματα, και σε μεγαλύτερα υπερκαταστήματα. Τα δεδομένα που δόθηκαν αφορούσαν συναλλαγές που είχαν πραγματοποιηθεί από τον Ιανουάριο του 2012, έως τον Μάιο του 2013. Για να εξαχθούν οι αγοραστικές αποστολές, υλοποιήθηκε κάθε ένα από τα βήματα όπως περιγράφηκαν πιο πάνω. Οι αγοραστικές αποστολές εξήχθησαν ανά κατάστημα, όμως παρατηρήθηκε ότι τα αποτελέσματα για κάθε τύπο καταστήματος έμοιαζαν αρκετά. Όπως επισημαίνεται στη μεθοδολογία, τα clusters που προέκυψαν και αφορούσαν τα δυο μεγάλα υπερκαταστήματα, ήταν δύσκολο να χαρακτηριστούν, ακόμα και με την εκ νέου υλοποίηση clustering σε κάθε cluster. Συνεπώς, στα μεγάλα υπερκαταστήματα πολλά clusters έμειναν αχαρακτήριστα, ενώ οι αγοραστικές αποστολές που εξήχθησαν ήταν αρκετά γενικές. Το παραπάνω επιβεβαιώνει τον περιορισμό που τέθηκε από την μεθοδολογία. Τέλος, τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση των ανωτέρω είναι: Microsoft SQL Server 2012 από το SQL Server Management Studio, SQL Server Data Tools (SSDT) του Visual Studio, και το Data Mining add-in του Excel.

Μετά την πρακτική εφαρμογή της μεθοδολογίας, αναλύεται η θεωρητική συμβολή αυτής της διπλωματικής. Όπως αναφέρθηκε, είναι η πρώτη που προτείνει μια συγκεκριμένη μεθοδολογία για την εξαγωγή των αγοραστικών αποστολών με τη χρήση clustering. Επιπρόσθετα η διατριβή αυτή συμβάλει στη θεωρία, προτείνοντας έναν νέο τρόπο κατηγοριοποίησης των καταναλωτών, σύμφωνα με τα αγοραστικά τους ταξίδια. Παράλληλα, η προτεινόμενη μεθοδολογία μπορεί να χρησιμοποιηθεί ως εργαλείο για τη στήριξη της διαδικασίας λήψης των αποφάσεων, έχοντας διαφορετικές εφαρμογές στο χώρο του λιανεμπορίου. Για παράδειγμα, μπορεί να χρησιμοποιηθεί από τους managers για το μετασχηματισμό ενός σουπερμάρκετ με βάση τις αγοραστικές αποστολές των πελατών τους. Μπορεί επίσης να χρησιμοποιηθεί από τους marketers για το σχεδιασμό cross-coupon προγραμμάτων. Κλείνοντας, συστήνονται προτάσεις για περεταίρω έρευνα, όπως χρήση διαφόρων τεχνικών DM και σύγκριση των αποτελεσμάτων με αυτά που προκύπτουν από το clustering. Προτείνεται, επίσης, διερεύνηση εναλλακτικών τρόπων αναγνώρισης των

αγοραστικών αποστολών, όπως χρήση RFID στα καλάθια των καταναλωτών και καταγραφή της πορείας του καταναλωτή μέσα στο κατάστημα. Τέλος, προτείνεται η περεταίρω ανάλυση όχι μόνο των κατηγοριών που αγοράζονται μαζί, αλλά και αυτών που δεν εμφανίζονται σχεδόν ποτέ στην ίδια αγοραστική αποστολή.

ABSTRACT

Consumers' behavior and expectations for service have changed dramatically in recent years, as they have become more demanding. Many organizations have identified the need to become more customer centric, facing increased global competition (Bull, 2003; Phan & Vogel, 2010). Therefore, in order to respond to the ever increasing demands of consumers, they are trying to develop innovative methods for managing their customers (Anderson et al., 2007). At the same time computers have become far more powerful, and new technological trends have been developed, such as Big Data, Business Intelligence (BI), Data Mining (DM) (Provost & Fawcett, 2013). These new trends give us the opportunity to process large volumes of data and extract valuable information. Like any other business, so do retailers have realized the importance of applying these new technological trends to support decision making and satisfy their customers (Bertino, 2011). However, there is only sparse research in the context of retailing in order to discover patterns in customers' behavior, to empower decision making, and to satisfy the demanding consumers (Wang & Zhou, 2013).

Motivated by the above, this thesis presents an effort to fill this research gap. It introduces a DM-based framework, which could be used to identify consumers' shopping missions in a supermarket. A shopping mission identifies the purpose or purposes a customer went to the supermarket. For instance, to buy products for his breakfast or meal, or to buy detergents etc. Via applying "clustering" as DM technique to consumers' daily purchases, this research identifies the product categories that are purchased together.

Based on the "Design Science" approach (Hevner et al., 2004), this thesis explores the above research problem, and overviews relevant researches. Then, an artifact is developed. This artifact is a DM-based framework, that draws on CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology (Chapman et al., 2000). It provides a certain manner to handle the appropriate data, in order to extract the shopping missions. The phases of this framework are five, (A) Business and Data Understanding, sub-steps of which are: Data Acquisition, Data Exploration and Data Preparation, (B) Cluster Sampling, (C) Modeling, (D) Evaluation, and (E) Deployment. Then, the above artifact is been evaluated, by implementing its phases in practice, and confirming if it can solve the original problem. The proposed framework

had been evaluated by applying it to a real case. This case concerned data of seventeen consecutive months of eight stores of a Greek retailer.

Additionally, the proposed framework has a few limitations, but it could be useful for both academia and retail industry. It's the first framework that indicates how to extract shopping missions from retail data, by identifying correlations in product categories, using clustering. Furthermore, it provides a new way to study customers in groups. Retails, can utilize it as a decision-making tool for different usages. For example, the framework's results could be used to modify and re-design the supermarket's layout, based on the customers' desires. Moreover, marketers could use it to develop marketing campaigns and promotions. Last but not least, further research could be conducted in order to apply other DM techniques, and to examine whether there are alternative ways to identify the shopping missions, such as RFID in customers' shopping carts.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. Research Motivation

Consumers' behavior and expectations for service have changed dramatically in recent years, as they have become more demanding. Many organizations have identified the need to become more customer centric, facing increased global competition (Bull, 2003). With growing competition from enterprises, keeping customers satisfied, increasing potential sales, and maintaining customer loyalty, become strategically important to business success (Phan & Vogel, 2010). For that reason, businesses try to apply new innovative methods to gain customers insights, support their decision making, and improve customer relationships.

At the same time, computers have become far more powerful, networking is ubiquitous, and Data Mining algorithms have been developed that can connect datasets to enable broader and deeper analysis than previously possible (Provost & Fawcett, 2013). These technological improvements allow us to analyze large volumes of data in order to improve the profitability and the success of many enterprises. As a consequence, data-driven decision making is now recognized broadly, and there is growing enthusiasm for the notion of "Big Data". Decisions that previously were based on purely intuition, can now be made based on data itself. Big Data analytics now drives near every aspect of our modern society, including retail industry, financial services etc (Bertino, 2011). Big Data research looks at how to analyze data in different domains with such characteristics and in a way that generate deeper knowledge and adds value to the decision making process in businesses. (Sharda, Asamoah, & Ponna, 2013).

With vast amounts of data now available, companies in almost every industry are focused on exploiting data for competitive advantage. The volume and variety of data have far outstripped the capacity of manual analysis (Provost & Fawcett, 2013). As a consequence, Business Intelligence (BI) tools have developed to assist decision making. In this context, retailers are also trying to use BI and DM in order to exploit the hidden knowledge, derived from the vast amount of data they have. However, not a lot of research has been done in retailing in order to discover patterns in customers'

behaviors in order to empower decision making and satisfy the demanding consumers (Wang & Zhou, 2013).

Motivated by the above, this thesis presents an effort to fill the research gap. It introduces a framework, which could be used to identify consumers shopping patterns in a supermarket. The results of this framework could be exploited by decision makers to satisfy the demanding customers.

1.2. Research Objective & Questions

The purpose of the research is to design a methodology-framework that answers to the question: "Which are the *shopping missions* of the consumers when they visit a supermarket?". *A shopping mission identifies the purpose or the purposes a customer went to the supermarket*. For instance, to buy products for his breakfast or meal, or to buy detergents etc.

The key input of this framework is the consumers' daily purchases collected by the point of sales/ cashiers in the retail stores. The shopping missions will be formulated applying Data Mining Techniques (DMT) to the given dataset. *I will identify the product categories that are purchased together* via Data Mining (DM) techniques. Product category is a way to describe the main characteristics of a product, for example coca-cola can is a beverage. Then these category association patterns will indicate the shopping missions.

1.3. Research Methodology

The research draws on a conceptual framework for understanding, executing and evaluating IS (Information System) research combining behavioral and design science. Design science is foundational to the IS research. In design science, the researcher creates and evaluates IT (Information Technology) artifacts intended to solve identified organizational problems. Behavioral science addresses research through the development and justification of theories that explain or predict phenomena related to the identified business need (Hevner et al., 2004).

According to the above framework the research methodology of this thesis which I followed to answer the key question is shown in Figure 1.1. As it has been mentioned above *this thesis aims to solve a business problem-need in the retailing environment*,

that is to identify the shopping missions of customers in a supermarket. Hence, I developed a technology-based solution that is relevant to the above research problem. In this research *the developed artifact is a framework*, providing a certain manner to handle the appropriate data with purpose to extract the shopping missions. In order to build this framework I used a *theoretical foundation*. Databases, Data Mining Techniques (DMT) such as clustering, Data Mining (DM) algorithms such as k-means, and CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology (Chapman et al., 2000) were used as the basic knowledge inputs in this framework.

Then, there was a need to put the framework in practice and realize if it can solve the original problem. The proposed framework had been *evaluated* by applying it to a real case, concerning the purchases of consumers in eight stores of a Greek retailer. At the end of this thesis, the theory contribution and the practical implications of the proposed framework are detailed.

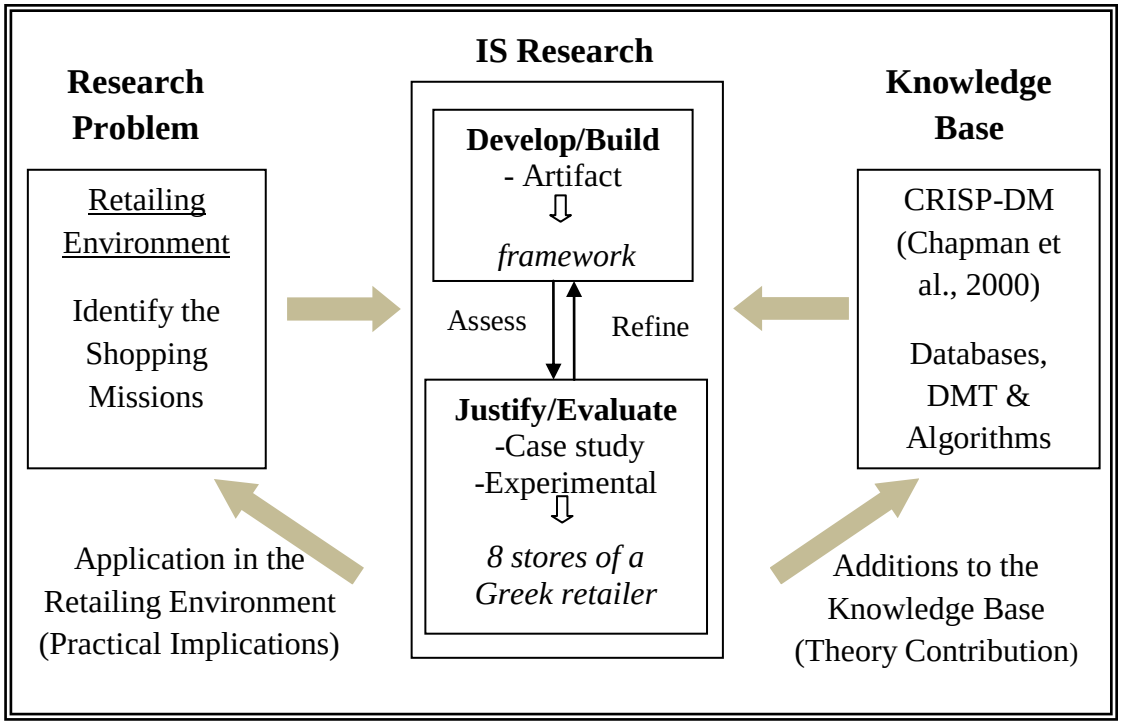


Figure 1.1 Research's Methodology Framework

1.4. Thesis Outline

This thesis consists of five (5) chapters. Each chapter implements a phase of the research's methodology framework, as described above. Figure 1.2 presents the thesis chapters, according to the research methodology.

The content of each chapter is summarized as follows:

- *Chapter 1 (Introduction)*

The purpose of this chapter is to introduce the reader to the main concept of this thesis, which is the identification of shopping missions. At first, trying to introduce the research area, research motivation and objectives are analyzed. Then, in order to guide the reader into the subsequent chapters, research's methodology and thesis structure are presented.

- *Chapter 2 (Research Problem)*

It analyzes the research area, surveys the literature, and points out the research gap. Firstly, it highlights how important is Customer Relationship Management (CRM) for enterprises, and therefore retailers. Then, it emphasizes at the emerging trend of DM applications in CRM, and presents common DM models used in CRM. Last but not least, after examining relevant studies, it highlights how this study puts an effort to fill the identified research.

- *Chapter 3 (A Framework that Identifies Shopping Missions)*

It proposes a framework that supports the identification of shopping missions in a supermarket. In this chapter the basic steps of each framework's phase are analyzed in order to provide a guide of how to extract shopping missions using as basic input Point Of Sales (POS) data.

- *Chapter 4 (Framework Evaluation)*

It presents the application of the proposed framework in a real case. The purpose of this chapter is to evaluate the framework, and realize if it accomplishes to solve the original problem.

- *Chapter 5 (Conclusions & Discussion)*

The last chapter overviews the main outcomes of this thesis. It also presents the theoretical contribution and the practical implications of the proposed framework. It outlines the framework's limitations and illuminates the area that could require further research.

Last but not least, this thesis includes an Appendix that complements the above chapters. The following figure presents thesis' structure according to the research methodology, as described at the previous sub-section.

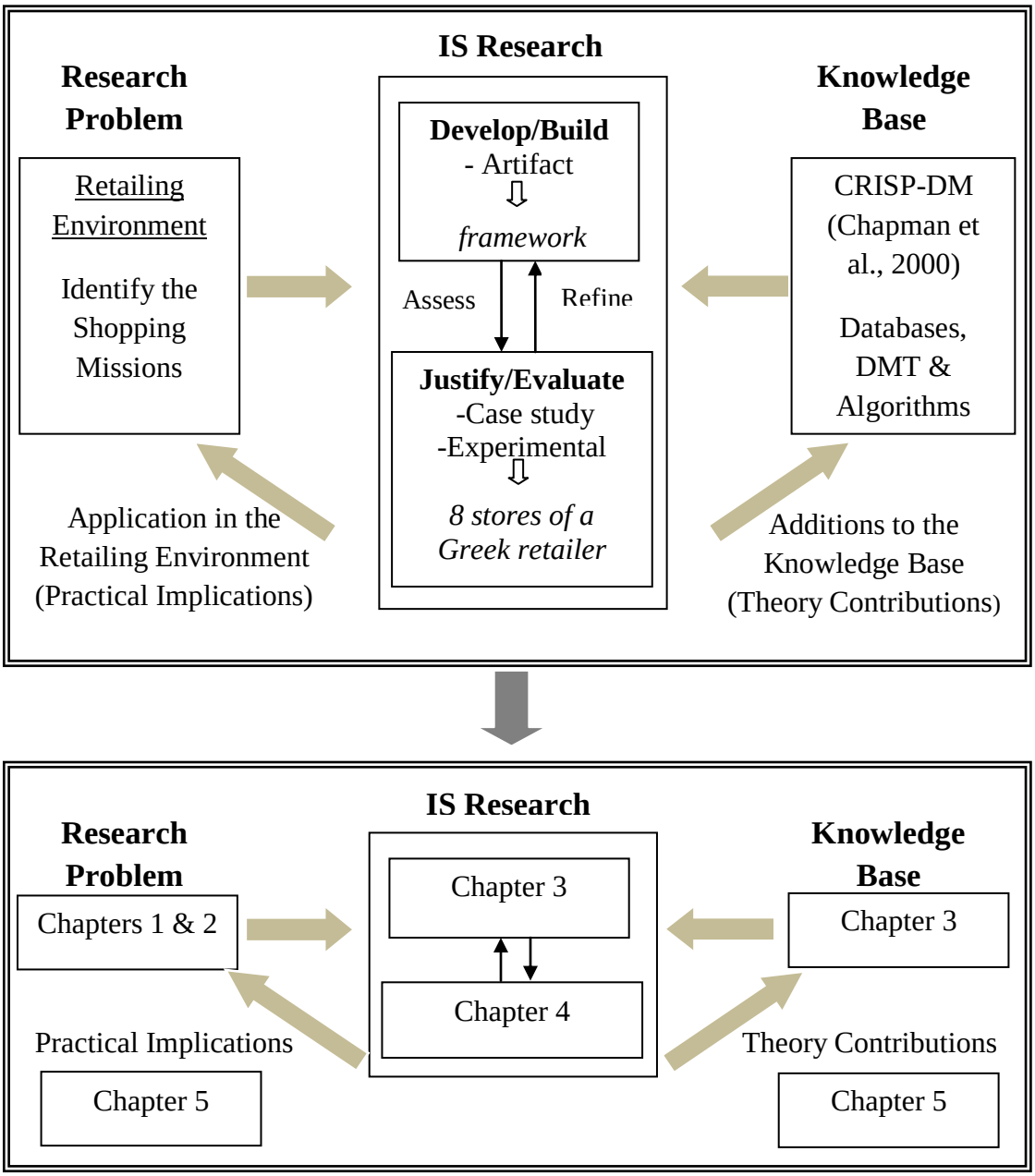


Figure 1.2 Thesis Structure (First framework: Research Methodology, Second: Thesis Chapters for each Research Phase)

2. RESEARCH PROBLEM

2.1. Introduction

In this chapter the research problem will be pointed out. At first the need of customer satisfaction and effective customer management will be indicated. Since the above, belong to the context of Customer Relationship Management (CRM), CRM's importance will be highlighted. Then, the advantages of using Data Mining (DM) in CRM and common DM models, used in CRM, will be presented in detail. At the end, relevant studies are examined in order to highlight how this study puts an effort to fill the identified research gap.

2.2. Customer Relationship Management (CRM)

Consumer behavior and expectations for service have changed dramatically in recent years. Nowadays, consumers have become demanding; thus, this affects how retailers act. In order to face this dynamic environment, and retain their market shares retailers should be alerted to satisfy customers' desires. As customer satisfaction has an effect on the profitability of nearly every business, retailers need to embrace a customer-centric focus and find out smart and innovative ways to manage their customers (Anderson et al., 2007).

Many researchers have pointed out how important is for enterprises to understand and satisfy their customers. Every forward-looking company is moving toward the goal of understanding each customer individually and using that to make customers to do business with them and not with the competitors (Linoff & Berry, 2011). As Grewal, Levy, & Kumar (2009) mentioned, "the key to the retailing success is to understand one's customers". Moreover, Jeevananda (2011) pointed that customer satisfaction is an asset that should be monitored and managed just like any other physical asset, and it is the key factor in knowing the success of any retail store or business. As a consequence, CRM has risen to the agenda of many organizational strategies (Bull, 2003). As retailers attempt to satisfy customers and improve their relationships, they have also try to develop innovative CRM strategies. However, there is a lack of innovative CRM applications in retail industry. As well not many research had been done concerning the concept of shopping mission.

CRM itself is not a new concept, it builds on the principles of relationship marketing, the formal study of which goes back 30 years (Berry, 1983). Trying to give a definition of CRM, we could tell that is a comprehensive business and marketing strategy which used to built long term and profitable relations with specific customers (Ling & Yen, 2001). CRM consists of four dimensions (Hosseini et al., 2010; Ngai et al., 2009).:

- *Customer Identification:* also referred as customer acquisition- is the initial phase of CRM, which involves targeting and analyzing the potential customers, the customers that will offer the most profits to the company, and the old customers that the competitors have gain. Target customer analysis and customer segmentation are the basic elements of this phase.
- *Customer Attraction:* follows customer identification, this is the phase in which organizations can attract specific customer segments, as been identified by the previous phase. An element of this phase is direct marketing such as direct mail or coupon.
- *Customer Retention:* is the central phase of CRM. Customer satisfaction is the basic condition to retain customers. The basic elements of this phase is one-to-one marketing, loyalty programs and complaints management.
- *Customer Development*: involves the empowerment of each customer's transaction intensity and value, which will expand the profitability of the firm. Basic elements are market basket analysis and up-cross selling.

CRM has evolved from advances in information technology and organizational changes in customer-centric processes. Both academic and industry people have highlighted the importance of CRM. As Chen & Popovich (2003) say companies that successfully implement CRM will benefit in customers' loyalty and long run profitability. Moreover, consulting firms assert that the CRM concept enables customer loyalty-building and profitable segmentation (Minami & Dawson, 2008).

2.3. Applying Data Mining (DM) Techniques for CRM purposes

The use of IT has created new ways for firms to exploit vast potentials of CRM. Business intelligence (BI) tools are used to assist CRM systems to focus on decision support, market research, target marketing, customer service, and customer collaboration in products and services (Phan & Vogel, 2010). Many companies have collected and stored data, however, they are unable to transform these data into valuable knowledge. For that reason the application of Data Mining (DM) tools in CRM is an emerging trend in global economy (Ngai et al., 2009).

DM applied to CRM could be used to create a deeper understanding of customers, and to maximize customer's value. It enables in-depth analysis of datasets, and allows extracting hidden customer characteristics, and behaviors of large volumes of data (Liao, Chu, & Hsiao, 2012; Ngai et al., 2009). The core of DM process is to build a model that will be used to mine knowledge from a dataset and solve a problem.

DM models are classified into two big categories (Gorunescu, 2011):

- *Predictive:* that are using a part of variables to predict one or more other variables.
- *Descriptive:* used to identify patterns that describe data and can be easily understood by the user.

According to Ngai et al. (2009) each DM model applied for CRM purposes can perform one or more of the following types of data modeling.:

- *Association:* is used to make simple correlations between two or more items in order to identify hidden relations and patterns among data. An association rule has the form $X \rightarrow Y$, that means that customers who purchase product X tend to buy also product Y (Ahn, 2012). Market basket analysis and cross-selling programs are typical examples for which association modeling is usually adopted. Common DM Techniques (DMT) are association rules and Markov chains. Common tools are Apriori and tree-based algorithms etc (Ahn, 2012; Ngai et al., 2009). A famous example is the association rule of beer and diapers.
- *Classification:* is the most common learning model in DM. It aims to build a model that will predict customer behaviors based on predefined characteristics. It works as follows, it has a predefined set of groups or models based on what values you

want to predict. Then the rest data will be classified in the predefined classes according to the relation of their characteristics with those that have been predefined (Ngai et al., 2009; Yang & Ma, 2013). For example[1], you can easily classify cars into different types (sedan, 4x4) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, you can classify it into an existing class comparing the attributes it has. You can apply the same principles to customers, for example by classifying them by age and social group. Common DMT are neural-networks and decision trees (Ngai et al., 2009).

- *Clustering:* is the task of segmenting the objects into groups called clusters, so that the object within a cluster are "similar" to another and "dissimilar" to objects in other clusters. Similarity is defined of how close the objects are in space, based on a distance function (Phan & Vogel, 2010). It differs from classification as there are no predefined clusters. Common DMT are neural-networks, k-means etc (Ngai et al., 2009).
- *Sequence discovery:* focuses on analyzing and identifying associations or patterns or trends between events over time (Ngai et al., 2009; Yang & Ma, 2013). For example in a store often people buy tomato sauce after pasta.
- *Regression:* is a statistical estimation technique which uses known data formats like linear or logistic in order to map each data object to a real value provide prediction value (Ngai et al., 2009).
- *Forecasting:* it can be used to estimate future values by analyzing past events and instances. A well-known example is demand forecast. Common DMT used is neural networks (Ngai et al., 2009).
- *Visualization:* refers to the presentation of data in a way so that users can view complex patterns and extract valuable information. It is used in conjunction with other DM models. Common tools are 3D graphs and Multidimensional scaling (MDS) (Ngai et al., 2009).

In Figure 2.1 below is shown the classification of the above DM model types into "Predictive" and "Descriptive.

[1] http://www.ibm.com/developerworks/opensource/library/ba-data-mining-techniques/index.html

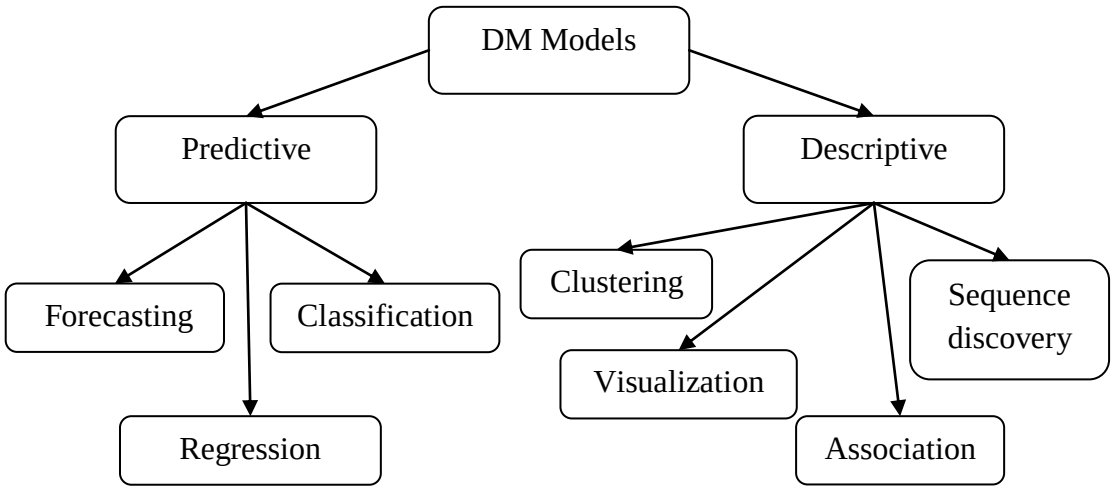


Figure 2.1 Classification of DM models

Each CRM dimension, as described in the previous section, could be supported by the different types of DM models. There are various applications of DM in CRM, an important example is the case of Fingerhut Inc. - a large catalogue mail order company in the U.S.- that by using BI and DM manage to predict consumers shopping patterns and credit behavior. So by providing one-to-one CRM-marketing, we gain customers' satisfaction and trust (Phan & Vogel, 2010).

2.4. Data Mining (DM) Applications in Retail Industry

DM can also consolidate retail data, and could be used to analyze them in order to create "one view" of the customer. By using DM, retailers could adopt a customer-centric approach, and stay competitive. That happens because the data gathered, could be used to gain customer and company insight, and support CRM (Anderson et al., 2007). CRM based on DM could help retailers cope with the ever changing demands of customers, so customers' satisfaction will be increased. DM help retailers patronage customers' behavior, gain insights, and retain customer's that really add value to the business (Min, 2006). By discovering patterns in customers' behaviors, enterprise's decision making could also be empowered (Wang & Zhou, 2013). A well-known example is Tesco, that by using Loyalty Card reinvent its relation with its customer and used it as a core element of its marketing strategy (Humby et al., 2003).

Except for the traditional BI and DM applications in retailing, in literature there are few papers that are relevant to the concept of shopping mission. The most recent relevant research is that of Cil (2012). The main purpose of this research is to identify the associations among categories in a supermarket, that will be used to change store's layout. In order to implement the above he identifies the categories' clusters or else the "consumption universes". Association rules and Apriori algorithm, used as DMT and as DM tool to analyze POS data of a Turkish supermarket. Then to identify the product categories clusters he used the extracted association rules as input in a Multidimensional scaling (MDS) tool. However, he was not the fist who introduces the idea of consumption universes. The above research builds on the research of Borges (2003). Borges uses the same techniques as above to identify the associations patterns among product categories in three French supermarkets, without making changes in supermarkets' layout.

Also, few scholars try to understand customer's behavior in supermarkets via market-basket analysis (Ahn, 2012; Raorane et al., 2012; Shrivastava & Sahu, 2007). In these cases Apriori and nearest neighbor algorithms are used to identify the hidden associations between product categories, and to gain customer insights. However, the extracted association rules didn't used to identify shopping universes, missions or trips.

Another interesting research is that of Larson, Bradlow, & Fader (2005). They used RFID (Radio Frequency Identification) in shopping carts to record customers' shopping paths in a grocery store. Then by implementing clustering they classify the shopping paths and examine common travel behaviors. However, to the best of my knowledge there is no other framework - methodology which gives specific steps and guidelines of how to extract shopping missions from retail data, by identifying correlations in product categories. Moreover, in contrast to the relevant works, that are using association rule mining, apriori and nearest neighbor algorithms to identify the correlations between products, this is the first research that introduces clustering, as DM model and k-means, as DMT to implement the above.

3. A FRAMEWORK THAT IDENTIFIES SHOPPING MISSIONS

3.1. Methodology

The proposed framework (Figure 3.2) draws on CRISP-DM methodology. CRISP-DM (Cross-Industry Standard Process for Data Mining) (Chapman et al., 2000) is a Knowledge Discovery and Data Mining process model (KDDM) created by a group of organizations from different industries with great experience in this type of knowledge discovery. A KDDM concerns the entire knowledge extraction process, including how the data is stored and accessed, how to develop efficient and scalable algorithms that can be used to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine (Kurgan & Musilek, 2006).

The proposed framework that supports the identification of the shopping missions consists of five (5) phases:

1. Business and Data Understanding, sub-steps of which are:
   1.1. Data Acquisition
   1.2. Data Exploration
   1.3. Data Preparation
2. Cluster Sampling
3. Modeling
4. Evaluation
5. Deployment

In Table 3.1 below are shown the mappings between the steps of CRISP-DM and the proposed approach.

CRISP-DM	Proposed Framework
1. Business Understanding	**1. Business & Data Understanding** **1.1. Data Acquisition** **1.2. Data Exploration**
2. Data Understanding	
3. Data Preparation	**1.3. Data Preparation** ✓ Data Integration ✓ Data Validation ✓ Data Cleansing ✓ Data Transformation
	2. Cluster Sampling ✓ Meaningful baskets Selection ✓ Outliers' Subtraction
4. Modeling	**3. Modeling** ✓ Clustering ✓ DMT: K-means ✓ Pre-condition: input dataset adjustment
5. Evaluation	**4. Evaluation** ✓ Evaluation of DM results ✓ First identification of Shopping Missions ✓ Model's re-execution
6. Deployment	**5. Deployment** ✓ In depth clusters' analysis ✓ Final identification of Shopping Missions

Table 3.1 Side-by-side comparison of the two approaches

The initial phase focuses on obtaining, exploring and preparing the dataset, having in mind the business objective that is to extract shopping missions. For that purpose, the necessary dataset is the following:

- Data that concern the consumers' daily purchases collected from the Point Of Sales (POS) /cashiers in the retail stores.
- Data that concern the product categories hierarchy and the barcodes the retailer has. These data are required in order to identify the product categories hierarchy of a purchased product at a POS.

For example, from the POS data you are given the information that a customer has bought the product with barcode "x". But you, also, need to know that this product belongs to the product category "rye toast bread", which belongs to the product category "packed", which belongs to the parent category "bakery products" etc. The above is required because exploring these category taxonomies you have to decide new and more specific category names. These customized names would be meaningful in the analysis and the identification of shopping missions. Examining the previous example, the new category name you will create could be "toast breads", including also products of the category "white toast breads" (Figure 3.1). *The processing of the categories taxonomies is a main step of Data Exploration phase.* Moreover, in this phase as getting familiar with the data, you will also make a *Data Selection*. As it is common the retailer gives you more than the required data. So data selection is needed to omit the irrelevant data to the DM goal. At the subsection 3.1.1. there is detailed list of the data you have to choose.

It is obvious that the given dataset is "dirty", and not in an appropriate format ready for analysis. For that reason it is need to be prepared. *Data Preparation* includes:

- *Data Integration*: involves merging the given POS data in a database table, and dropping the other selected data in tables that belong to the same database. For example, you may receive the POS data in batches concerning months or quarters etc. These batches of data need to be integrated into a single table.
- *Data Cleansing*: involves deleting duplicate values, or corrupted records, or deleting records that are not associated with the given data. You may also need to delete generally wrong values, such as negative values in a field that concern products' units etc.
- *Data Transformation*: involves creating new data from the existing. For example, a necessary need is to group POS data and transform them from transaction to basket level.
- *Data Validation:* happens after each of the above steps to ensure that you will not have wrong inputs in the next. The main tasks of this step are: cross-check the data with the initial raw files, ensure that there are not missing data after each step's execution, search for mistakes in data types that caused corrupted records etc.

After preparing the dataset, you have to implement *Cluster Sampling with k-means in order to extract the outliers*. Outliers are baskets with too little number of products, from which you can't extract a shopping mission, and baskets with too many products; that includes many abstract shopping missions. *At the Modeling phase you need to use Clustering as modeling type and k-means as DMT.* A DM pre-condition is to integrate all the tables you have in your database into one table or file. This will be used as the input dataset of the model. The above is a compulsory step to ensure the interoperability of your dataset with other DM tools. After this, you have to *Evaluate the DM results with the assistance of industry's people*. *You have to decide whether the results meets the reality, or you have to re-execute the previous phase*. In order to evaluate the results, you need to make a first identification of the shopping missions. Specifically, *the results of the modeling phase would be groups-clusters of product categories, each resulting cluster will represent a shopping mission. By identifying the prevailing categories of a cluster, you can characterize the shopping missions*. Finally, *at the Deployment phase the final-verified modeling results, will be explained and analyzed in depth.* In this phase the final identification of the shopping missions will be implemented. These missions could be used by the retailers to support CRM strategies.

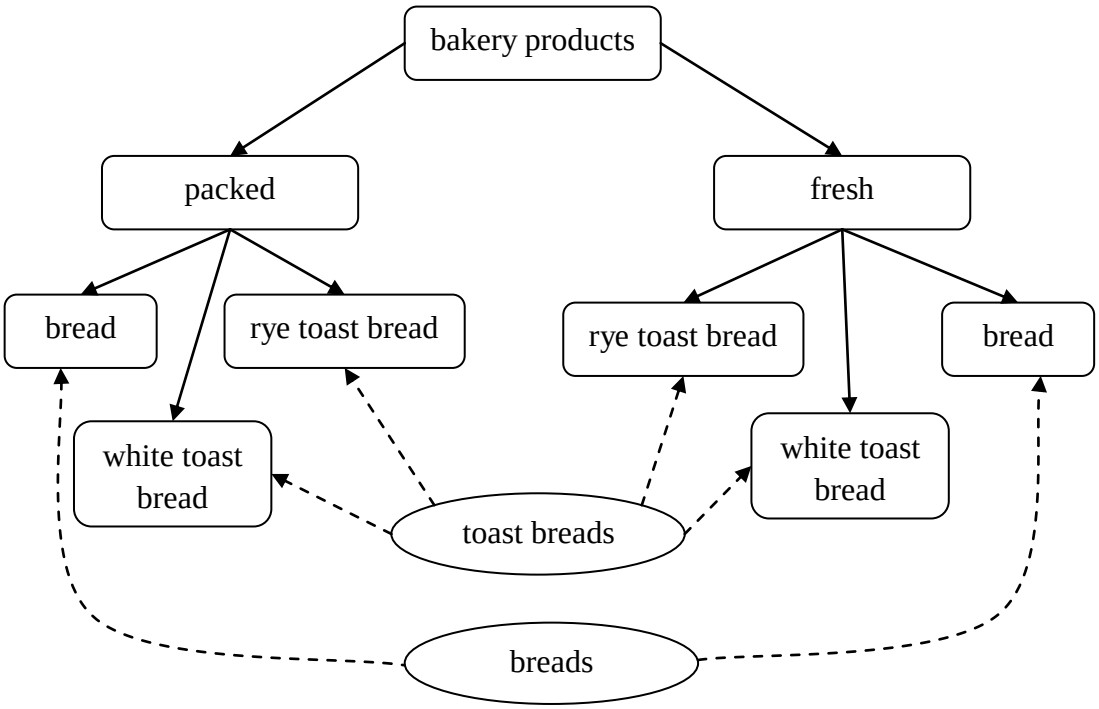


Figure 3.1 Creation of new customized categories

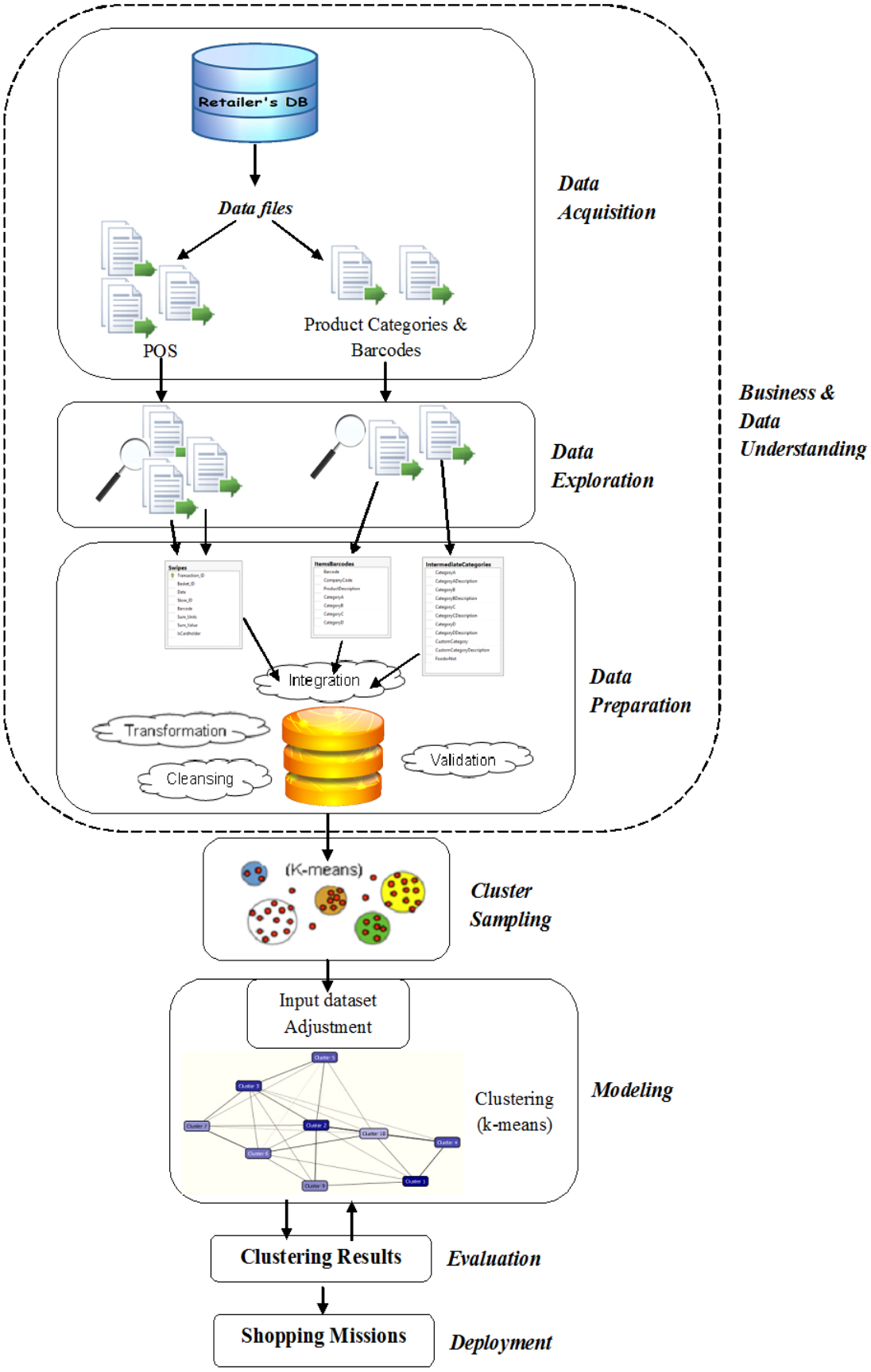


Figure 3.2 Proposed Framework

3.2. Business & Data Understanding

3.2.1. Data Acquisition

The necessary dataset required by the retailer includes two type of data, (A) POS data, and (B) barcodes data, with their product categories. Regarding the POS data, you need to know all the transactions-product scans happened in the POS, with additional information the receipt-basket they belong to. For example, if a customer bought five products, you need to have five transactions associated with the same basket-receipt number. That is important because in the next steps, you need to have the ability to group these transactions in basket level, in order to proceed with the sample clustering based on basket size.

Moreover, you could enrich the framework by asking the retailer for additional data, such as customers' loyalty cards data. These data may concern demographic characteristics about the retailer's loyal customers and their purchases. Having the above data, you can associate each transaction happening in the POS with a cardholder. Then, you can extract additional information about the demographic characteristics of each shopping mission.

It is common that the retailer provides you with more data than the required, as the received files may contain additional un-useful fields for your analysis. For that reason, a data exploration phase is necessary: (A) to evaluate whether the data acquired satisfy the relevant requirements, (B) to select the appropriate data, and (C) to process the hierarchy of categories you will use (Chapman et al., 2000).

3.2.2. Data Exploration

Data Exploration phase is the beginning phase of the analysis. It is a high level of data understanding, and it is needed to get familiar with the dataset (Shahbaba, 2012). During this phase you have to examine whether the given data are those you asked for, or you need to ask for more. A description of the acquired data, including data variety (flat files, excel files, csv (Comma Separated Values) files, database tables etc), dataset volume (number of records, number of files received etc), and the data quality, would be useful in this step.

Furthermore, in this phase you have to make a data selection, as you have to choose which data files and fields will be useful for your analysis. In Table 3.2 there is a list of the necessary fields you have to obtain. Last but not least, in this phase you need to examine the product hierarchy categories (category taxonomies), and determine new customized category names that will be used to identify the shopping missions (Figure 3.1). During the Evaluation phase you may reconsider these category names, and merge the categories in a different way. The above could happen if the first DM results are not satisfactory. At the description of the Evaluation phase there is a detailed example.

POS Data	
Necessary fields	**Short Description**
Transaction id	The id of each transaction-product scan in the POS (If the given data do not contain this field, you have to generate it during Data Integration)
Barcode	The barcode of the bought product
Basket-receipt id	The identification of the basket each transaction belongs to
Store id	The store each transaction took place (if you receive data for more than one store)
Number of units	The number of units of the bough product (i.e. 6 coca-cola cans)
Product Value	The total value of the above product units
Date	The date the transaction happened
Barcodes Data	
Product Barcodes	The barcode of the bought product
Description	Product's description
Parent Category id	The id of the Parent Category the product belongs to
Sub-Category A id	The id of the sub-Category A the product belongs to
Sub-Category B id	The id of the sub-Category B the product belongs to
etc	Data about all the other category taxonomies
Categories Data	
Parent Category id	The id of the Parent Category the product belongs to
Description of Parent Category	A short description of the parent category, i.e. "Bread"
Sub-Category A id	The id of the sub-Category A the product belongs
Description of Sub-Category A	A short description of the sub-category A, i.e. "packed breads"
Sub-Category B id	The id of the sub-Category B the product belongs
Description of Sub-Category B	A short description of the sub-category B, i.e. "rye toast bread"

etc	Data about all the other category levels
New Category Name	The new-customized category name you created, i.e. in this case it could be "toast bread"
New Category id	A generated unique identification of the above

Table 3.2 List of necessary data fields

3.2.3. Data Preparation

Data Preparation phase covers all activities to construct the final dataset, before applying cluster sampling, and before proceeding with the modeling. The basic tasks in this framework are four: (A) Data Integration, (B) Data Cleansing, (C) Data Transformation and (D) Data Validation. Data validation is a repetitive task which is implemented at the end of each other task. It ensures that the correct data will be used as input for the next task.

Data Integration

Data Integration involves combining data residing in different sources (Lenzerini, 2002). The retailer will send you the POS data in batches, concerns months, or quarters, etc. A basic task is to integrate all these data in a table that will concern all the POS data. Moreover, you have to integrate the data you have selected in the previous phase, in tables that belong to the same database. For that reason, at the beginning you need to create in your database three kind of tables, one for the POS data, one for barcodes, and another for the processed categories' hierarchy data. At this step, if you, also, have asked for loyal customers' data, you need to create more tables. During this task, you may also need to reformat some data, such as change the data type of date field from "datetime" to "integer", in order to store less space in server.

The dataset you will receive from the retailer will be heterogeneous. For that reason, in this framework it is recommended not to use foreign keys between the database tables, as due to this heterogeneity, integrity problems will occur. In the modeling phase it is proposed a specific way to handle these non-integrity tables.

Data Validation

Data Validation happens after each task to ensure that you will not have wrong inputs in the next one, so at the end the DM will not have any wrong inputs. The main validation tasks, that should happen after data integration, via querying the uploaded dataset, are:

- Cross-check the data with the initial raw files
- Check for missing data, such as not uploaded fields
- Search for mistakes in data types that caused corrupted records. For example, a string field has unrecognized value due to wrong unicode, or decimal values may be transformed by mistake into integer values, or they could have missing digits.

Data validation commonly leads to the re-execution of the data integration phase, till data integration outputs are validated.

Data Cleansing

Frequently, the collected data will not be in a format ready for analysis. It is obvious that the given dataset contain "dirty data", so data cleansing is required. Data Cleansing involves detecting and removing errors or inconsistencies of data, in order to improve data quality (Rahm & Do, 2000). Common cleansing tasks, that need to be done at these kind of datasets, have to do with:

- Deleting the corrupted records you found out with data validation. This should happen after ensuring that these records are not a result of your fault, and you could not correct them.
- Deleting duplicate values from all the tables.
- Recognizing and correcting the units of products that are not packed, such as fresh rice. In order to proceed with Cluster Sampling you have to count these kind of products as one. For example, in the above case the "2,5" kilos of fresh rice, should be converted into "1".
- Deleting records with unexpected negative values, such as negative numbers in "units" and "value" fields. Since many retailers count the returned products with negative values, before deleting these records you have to ensure that they are not

returns. You can ensure it by checking whether there is another record with positive values with the same characteristics.

- Deleting data from tables that do not have a matching with the rest, such as:
  - Delete records from barcode's table that do not have a matching with the given product categories.
  - Delete records from barcodes that haven't got a matching with any barcode in POS data. It is common the retailer sent you data about barcodes and product categories he does not use anymore. So, you do not need to carry these data on the next steps.
  - Delete records from POS data, such as barcodes that were not contained in barcode files.

After the cleansing phase, it would be useful to check the impact of cleansing in comparison with the initial dataset, such as number of records deleted. A validation phase is also needed to check, mainly via querying, whether corrections and deleting tasks happened as expected.

Data Transformation

Some of your data need to be transformed and represented in a different way, more suitable for your analysis. In this step you need to create new tables from the existing ones. A common need is to transform the information you have from the level of transaction in POS data, to basket level. You can implement this by aggregating the transactions that belong to the same basket. You need this information because in the next steps it will be used to implement Cluster Sampling, based on basket size. In Figure 3.3 it can be shown an example of the transformation from the transaction to basket level.

Moreover, you have to create another table that matches each barcode with the customized categories created in Data Exploration phase. This could happen by combining two other tables you have in your database: (A) The table that matches the hierarchy of categories with the new customized categories, and (B) the table concerning barcodes. This task is needed in order to omit the un-useful data fields, and keep only that which really matters. Last but not least, you have to ensure via querying that the transformed data are as expected to be.

Transaction id	Basket-receipt id	Barcode	Store id	Number of units	Products' Value	Date
1000	100	850037172	3	2	4,5	3/1/2013
1001	100	909249274	3	3	5	3/1/2013
1002	101	239570183	2	1	1,5	9/1/2014
1003	102	234017429	1	1	2,4	10/1/2014

Basket-receipt id	Store id	Basket Size	Products' Value	Date
100	3	5	9,5	3/1/2013
101	2	1	1,4	9/1/2014
102	1	1	2,4	10/1/2014

Figure 3.3 Transformation from Transaction to Basket level

3.3. Cluster Sampling

Having considered the purpose of the research there is no meaning in applying DM on all the dataset. In the dataset derived from the previous phases there are outliers. You have two types of outliers:

- Baskets having too little number of products, that will not indicate you a shopping mission. For instance, if you have a basket with two products, such as milk and detergent this basket does not indicate a specific shopping mission.
- Baskets having too many products, that contain many and abstract shopping missions. For example, if somebody bought a basket with more than a hundred products, this basket would not have a specific shopping mission.

In order to proceed with the modeling you have to identify the baskets that really matter. Basket size will help you find out these meaningful baskets, as it will be used as the key to identify the prevailing baskets of each store, and extract outliers. The sampling technique you have to use in this phase is Cluster Sampling, with equal sampling weights. Cluster Sampling is a technique where the finite population is grouped into subpopulations-groups called clusters, then a subset of these clusters is selected (Särndal, Swensson, & Wretman, 2003).

The proposed algorithm to be used is k-means. K-means is one of the well-known algorithms for clustering, searching for a nearly optimal partition with a fixed number

of clusters. First, a starting point with the chosen number of clusters is built, and then the partition is improved iteratively (Phan & Vogel, 2010). Before executing the model you need to generate a procedure to test the model's quality, validity and accuracy. Therefore, you have to split the dataset into training and testing sets. The training set will be used to implement the clustering, and the testing set to estimate its quality and accuracy (Chapman et al., 2000). The proposed percentage of training and testing set is sixty and forty percent respectively.

Concerning the algorithm's parameters, you have to set as data source the table "basket", and as input the column that must be analyzed, in this case the "basket size". Regarding the number of clusters, it is proposed not to indicate the actual number, but leave the algorithm execute the optimum case. After cluster sampling results, you have to calculate the actual number of baskets that belong to each cluster. You need the above in order to determine the meaningful and prevailing baskets. Moreover, you could calculate the revenues derived from each cluster. Then, taken into consideration the above two factors you can decide which clusters (basket size range) are those you will use to proceed with the modeling.

3.4. Modeling

According to CRISP-DM the purpose of this phase is to create the model you will use to apply DM in the dataset. The basic tasks in this phase are: (A) select the DM model type and technique you will use to build the model, (B) set the model's parameters, and (C) generate a procedure or mechanism to test model's quality (Chapman et al., 2000). The proposed framework specifies the above tasks and adds one more. This task has to do with the need to adjust the dataset in order to ensure interoperability with other tools.

3.4.1. Input Dataset Adjustment

As mentioned above, a DM pre-condition is to present the data in a way that would be analyzable by many tools. You need this in order to assure the interoperability of your dataset with other clustering tools. In your database you have plain tables which have no foreign keys. In this stage you need to merge the data from all these tables into one table or file. To achieve it, it is necessary to implement a second phase of data integration. This new file-table will include all the information about your database

tables in basket level. Each row will represent a basket, having as columns all the attributes that basket table has, and the customized categories you had created in previous steps. If you had asked your retailer about additional data, such as loyal customers' data, in this fact table you would also have more columns concerning the extra data. In Figure 3.4 the columns of this table are shown. All columns that concern the categories will be filled with true or false, according to the condition: "products of this category are contained or not, in this basket?".

The creation of this table could be implemented either by writing code in a preferred language, or by writing SQL scripts. Then, by converting this table-file into a csv (Comma Separated Values) file, it can be processed by almost every clustering tool. At this phase, you can also make a second phase of data selection. You can extract fields from your final dataset that finally you will not use for your analysis, but you had kept from the first phase of data selection.

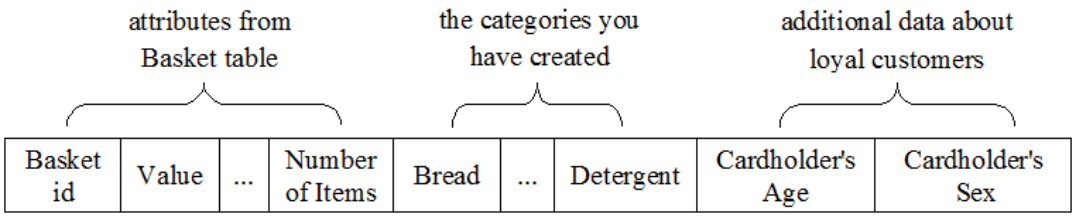


Figure 3.4 Fact Table Columns

3.4.2. Model Implementation

In this framework, Clustering is proposed as the type of DM model, and k-means as DMT. Before executing the model you need to generate a procedure or mechanism to test the model's quality, validity and accuracy. As described before, in this phase you also have to set the training and testing set, in order to test model's accuracy. The proposed percentage of training and testing set is sixty and forty percent respectively.

Regarding the algorithm parameters, you have to set as data source the above created fact table and, as input and predict values, all the customized categories you have. Moreover, it is proposed not to indicate the actual number of clusters, but leave the algorithm execute the optimum case. With that way the algorithm will process all the categories, and will calculate the categories' clusters. After model's implementation, you have to assess the model, by checking its accuracy. This assessment could be

based on the results extracted from the mechanism made before, to evaluate the model's quality and validity.

3.5. Evaluation

At this section you have to examine the modeling results, and check whether these results make sense, or you have to return at modeling phase and re-execute the model with changes in input dataset. This evaluation will be done by examining the product categories each cluster contains, and the contribution of each category to a cluster, in order to make a first identification of the shopping missions. Specifically, the results of the modeling phase would be clusters. Each resulting cluster will contain associated categories. These categories will be appeared in percentages in each cluster. The percentages indicate the contribution of a category to each cluster. The prevailing categories of each one will indicate you a shopping mission, and will help you characterize this cluster. For example, if the prevailing categories of a cluster are: milk, cereals, coffee and sugar, this might be a shopping mission for "breakfast".

As referred above, in this phase you have to check the associated categories the clustering extracted, and assess whether these associations make sense. For this reason, DM results should be shown and verified by the people of the industry, that know the domain. These people will help you assess whether these results meet the reality, or you have to make changes in the input dataset and re-execute the model.

These changes have to do with deleting and merging some of the customized categories you had created in data exploration phase. For example, the people who know the domain may expect a shopping mission which concerning "breakfast". But in DM results there is not a cluster with associated categories that concern "breakfast". Hence, you have to check whether there are relative categories in low percentages, in a cluster, that if you merged them they would appear in higher percentages. For example, it may exist in a cluster the following categories: low-fat milk, high-fat milk, chocolate milk, low calories cereal and high calories cereal, appeared in a low percentage. If you merge the cereal related categories into one category named "cereals", and the milk related categories into another category named "milk", and re-execute the model, these two new categories will be appeared in a higher percentage, so the hidden related to "breakfast" shopping mission will be

appeared. Furthermore, you have to check whether there are categories that appeared in a too little average percentage in all the clusters. You could delete these categories from the input dataset, in order not to affect the results, and then you can re-execute the model. For example, if the category named "camping" appeared in average percentage less than 2% within all the cluster, you can delete it.

3.6. Deployment

In this phase, the final and verified modeling results will be explained and analyzed in depth. In the Deployment phase will be implemented the final identification of the shopping missions, the conclusions of which could be used in practice by the retailers to support CRM strategies. Thus, in this and in the previous phase, it would be useful to choose a tool to achieve a greater visualization of your results. This tool will be used as the bridge between the knowledge extracted from the DM, and the managers.

In order to proceed with the identification of the shopping missions, as in the evaluation phase, you have to examine the product categories, and their contribution in each cluster. As each basket is categorized by the algorithm in only one cluster, in order to analyze them in depth, it could be helpful to calculate for each one of the resulting clusters the following: average basket size, average number of unique product categories of the baskets contained in each cluster. These two factors will help you analyze and characterize the final shopping missions. Trying to identify the shopping missions, you may observe some resulting clusters with high average basket size and product categories. These clusters can't be easily characterized, as they contain too general and abstract shopping missions. Drill-down in these clusters will be helpful to characterize the hidden shopping missions. With the drill-down you will perform clustering in a single cluster, so maybe more shopping missions will occur. For example, you may have a cluster with many products and product categories concerning both food and non-food products, such as vegetables, shampoos, toothpastes, pasta, fish etc. If you implement drill-down clustering in this cluster the tool may split this cluster into two sub-clusters. For instance, as shown in Figure 3.5 these could be, a cluster with "oral and body hygiene" products, and another cluster with "meal preparation" products. Therefore, these two hidden shopping missions will be appeared.

It is common the resulting sub-clusters to be the same as those you have already found and characterized. If this happens, or if the sub-clusters are still too general and abstract, then you can just leave this cluster uncharacterized. Moreover, during the characterization of the clusters, you may observe a few small clusters that also can't be characterized. In these clusters, since they contain baskets with a few products and categories, there is no meaning in implementing drill-down, so you can just leave them un-characterized.

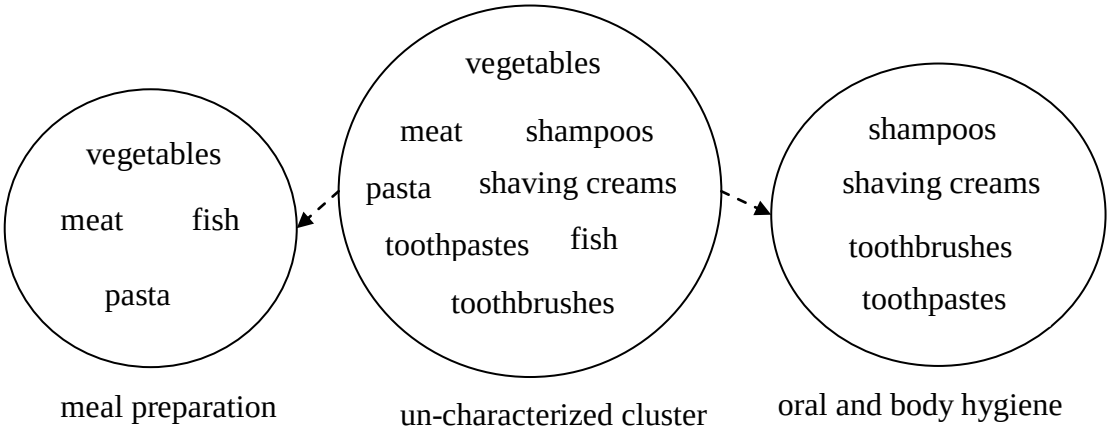


Figure 3.5 Drill-Down in a cluster

Abstract and too general shopping missions are more common to happen when you are trying to analyze the shopping patterns in a big store. For instance, a hypermarket with a wide range of food and non-food products. This happens because customers commonly visit this type of stores to make bulk purchases. Hence, their baskets contain a lot of irrelevant products, without a certain shopping purpose. In this kind of datasets it is too difficult to extract specific shopping missions, even by applying drill-down.

4. FRAMEWORK EVALUATION

In this section the proposed framework will be put in practice in order to realize if it manage to solve the original problem, that is to identify the shopping missions. One of the biggest retail chains in Greece in terms of both turnover and number of stores provided us with retail data, in order to extract the shopping missions. The next paragraphs describe the above framework's evaluation by applying it to this real case.

The tools used in this case in order to upload, prepare and analyze the data are the following:

- Microsoft SQL Server 2012 of SQL Server Management Studio (SSMS)
- Microsoft SQL Server Integrated Services (SSIS) of SQL Server Data Tools (SSDT) of Visual Studio
- Microsoft SQL Server Analysis Services (SSAS) of SQL Server Data Tools (SSDT) of Visual Studio
- Data Mining add-in for Excel

4.1. Business & Data Understanding

4.1.1. Data Acquisition

For the research's purposes the collaborator retailer extracted from his corporate database product categories, barcodes and POS data from January 2012 to May 2013, from eight of the stores he has. Except for these necessary data, the retailer was willing to give customers' loyalty cards data. These data contained demographic characteristics about the cardholders he has in the 8 stores. With this kind of data the framework will be enriched, as it could be extracted more information about the profile of loyal customers of each shopping mission.

4.1.2. Data Exploration

At this phase the given dataset had been examined. After a first data exploration, the given dataset was considered to be sufficient for the forthcoming analysis, so there was not a need to ask for more data. As the framework indicates, in this phase, there is a need to select the files and fields that will be used in the analysis. In Table 4.1 there is a detailed description of the received dataset and the fields will be used. The other

basic task of this phase was to decide the new customized categories. After exploring the category hierarchies-taxonomies, it has been decided to categorize all the products in 104 new categories. In Figure 4.1 there is an example of the new customized categories.

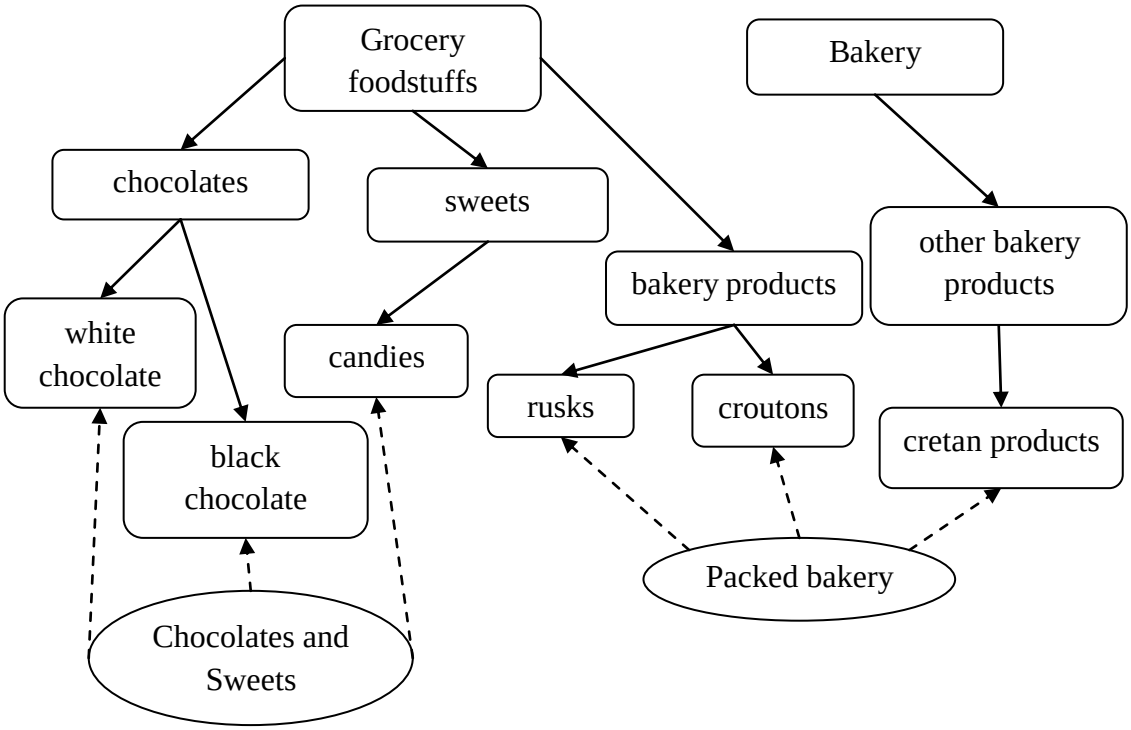


Figure 4.1 Example of the new customized categories

4.1.3. Data Preparation

In the context of this research the data were loaded and processed in Microsoft SQL Server 2012 of SQL Server Management Studio (SSMS), using SQL Server Integration Services (SSIS) of SQL Server Data Tools (SSDT) concerning Business Intelligence for Visual Studio.

Data Integration

The purpose of this task was to load the data from the different given data files in a database in SSMS. The initialization of the database was a compulsory task, as the empty database tables should be created to load the data. Five tables without foreign keys had been created. One table in order to load and integrate the POS data, that as expected they had been sent in batches concerning half months. Another table, to match each loyal customer with the baskets he had bought. Two more tables for the

product categories and the barcodes the retailer has, and another one to keep the information about loyal customers.

Then, the selected data had been loaded creating an SSIS (SQL Server Integration Services) project in Visual Studio, in order to indicate the data flow tasks from the initial files to the empty database tables. During this task, there was a need to execute some sub-tasks, such as reformatting the data type of some fields. The details of the Data Integration task are shown below in Table 4.2.

Data Validation

It has been checked whether the data were the same as those from the initial raw files, or whether there were missing data. Some execution tasks were revised and re-executed multiple times, as mistakes in data types were found. Other common errors arising concerned wrong unicode and lost digits, mainly at the numeric values. The validation has been done via writing queries in SSMS and cross-checking the resulting data with the initial.

Data Kind	Description	Time Period	Number of stores	Store types	Number of Transactions	Number of files received	Data Selction
POS	Data about each swipe-product scan happened in the cashier, such as: the barcode of the bought product, the number of the items of that product, the sum value of the product items, the basket the transaction belongs, the day and the store the transaction happened, the unique identification of the cardholder (if exists)	17 consecutive months, January 2012 to May 2013 for all store types. Except mini-hyper stores for which they sent data only for April and May 2013	8 different stores of Attica region.	• convenience •supermarket • mini-hyper market • flag-hyper market 2 stores of each type, with same size, from regions with common characteristics	36.797.639	68 flatfiles	All the given files and fields have been used
	Description	**Number of categories**		**Number of barcodes**		**Number of files received**	**Data Selction**
Product Categories and Barcodes	Data about the categories the retailer has, and the hierarchy of them (name - unique identification)	7 big parent categories, 5 sub-categories concerning the hierarchy of the categories (1st sub-category 51 unique product categories, 2nd-469, 3rd-1943, 4th &5th undefined number or unique product categories)				1 excel file (6 sheets)	Only fields till sub-category 3 have been selected
	Data about all the barcodes of the products, in all stores the retailer has. The fields contained: product description, company code, association with the categories the barcode belongs (till sub-category 3)			840.017		1 excel file (1 sheet)	Extra given fields, such as company codes, have been extracted

	Description	Time period of segmentation	Number of loyalty customers	Number of files received	Data Selction
Customers' Loyalty Cards	Data about demographic characteristics of loyalty customers (age, sex, marital status, number of householder, number of children, postal code), segmentation of them in food or non-food buyers, and categorization of customers according to their purchases (bronze, silver, gold, diamond, newcomers). This segmentation was a result of the RFM[2] analysis	Fourth quarter of 2012 and first quarter of 2013	191.129	1 excel file	All the given fields were used. At input dataset adjustment phase more fields will be eliminated

Table 4.1 Dataset Description

[2] Recency of last purchase, Frequency of the purchases, Monetary value of the purchase in a particular period

Data Kind	Number of tables created	Number of data flow tasks created	Uploading Details	Extra Tasks	Common arising errors during the execution of data flows
POS	2 tables One to upload the common data the 68 files had. Another to split the extra information about loyal customers (their id and the baskets they bought)	3 flow tasks -1 for non-cardholders data (32 files executed with this flow task) -1 for cardholders data (32 files executed with this flow task) -1 to split the data concerning cardholders (re- execution of the above 32 files)	-32 files concerning non-loyalty customers uploaded "as-is" -32 files concerning loyalty customers uploaded "as-is" without the extra attribute	Derived Columns: add an extra column to identify if the shopper is loyal or not Data Reformation: Conversion of attribute date from data type "Date" to data type "Int" (integer) in order to store less space in server Split and drop the extra column of cardholders in another table	False data type declaration during the creation of the tables in SSMS. Data types which do not correspond to those of the data from the flat file records. Wrong mappings between source and destination data columns i.e. drop a column of the flat file in a wrong column of a table in the OLE DB[3] destination
Product Categories and Barcodes	1 table	1 flow task	The 1 sheet of the given excel file uploaded "as-is" after adding manual the new customized categories		The same as above
	1 table	1 flow task	At first the file table uploaded "as is"		
Customers' Loyalty Cards	1 table	1 flow task	The file uploaded "as-is"		The same as above

Table 4.2 Data Integration Details

[3] Object Linking and Embedding Database

Data Cleansing

After having validated the dataset, ensuring that the inputs are correct and as it has been expected, there was a need to clean the "dirty records". This cleansing was made either via writing, deleting or updating queries in the SSMS, or via SSDT. In SSDT, the SSIS package was used to parse the "dirty data", clean and drop them in "clean tables".

As it has been expected the retailer sent more data than those that really matters. For instance, he sent information about all the barcodes he had, and not those which concerned the transactions from January 2012 to May 2013. For that reason, there was a need to delete the extra and non-useful data. Moreover, the units' field of packed products also needed to be cleaned and corrected. More details of Data Cleansing tasks are shown below in Table 4.3.

Data Transformation

After validating the cleaned data via repeating the tasks as described above, there was a need to transform and represent the dataset in a different way, more suitable for the analysis. At this phase the basic tasks are two:

- Transform the information from the level of transaction to basket level, by aggregating the transactions that belong to the same basket. In order to execute this task another data flow task has been created in the SSIS. With this task 3.973.215 baskets entered in a table in SSMS.
- In order to keep information only about the barcodes and the 104 categories created in previous phase, there was a need to transform the existing dataset. "Look-up" and "derived column" functions of SSIS were used to combine product categories and barcode tables, in order to materialize this task.

Data Kind	Problem Located	Checking Tasks	Solution	Tool used	Number of records Deleted
POS	Negative values in columns concerning the number of units and the sum value of a transaction	Examining these data I come to the conclusion that are either mistakes, (as some of them didn't had a matching with an existing barcode), or returns. It appeared important to examine if there were returns. For that reason, I searched for transactions with the same characteristics, but with positive values within the same basket, that essentially cancel the negative records. Coming to the conclusion that is not true, negative transactions have been deleted	Delete all negative value records	An SSIS flow task used to delete these wrong records simultaneously	126.834 records
	There were remaining positive records that didn't had a matching with an existing barcode	The number of barcodes involved in these records where at about 258. Examining these barcodes I found out that were wrong values which had been extracted from the corporate database with less digits than a common barcode.	Delete records with no matching with existing barcodes		426.103 records
	With the variable "number of units" the retailer used to count not only the units of each item bought, but also the kilos of the weighable products. For this analysis it was meaningful to count the weighable products as one product. For instance, "2,5" kilos of meat been set as 1	Checking via querying the description of the barcodes referred to these data.	A query that checks "units" column in the Transactions table, and when it founds a positive value having comma, it converts it to 1	An updated query in SSMS used	
	Duplicate records resulted during splitting cardholders data in another table		Keep only distinct records	Another SSIS flow task used to extract duplicate records	

Product Categories	Duplicate records existed in the table concerning barcodes		Keep only distinct records	Another SSIS flow task used to extract duplicate records	
	The retailer gave data about all the barcodes he had even for old barcodes that were not used anymore. But, in POS data there were used only 126.402 barcodes		Keep data only about barcodes that exist in POS data	Deleting query in SSMS	713.615 records
Customers' Loyalty Cards	Data contained a lot of null and zero values, as customers didn't filled their personal data. But, zero values in columns, such as age and household size need to be converted into null. Otherwise, the results of the forthcoming analysis would be influenced i.e. average results		Updated queries, to convert zero values into null	Two updated queries in SSMS used	

Table 4.3 Data Cleansing Details

Data Preparation Impact in Dataset Volume

In Table 4.4 you can observe the impact in the final dataset, after the implementation of data preparation tasks, and the data exclusions. The percentage of product categories is low because, as mentioned above, the retailer sent data of all the barcodes he had, but in the POS data existed only the 15% of them.

Data Kind	Number of Initial Records		Number of deleted Records	Percentage of dataset kept
POS	Transaction:	36.797.639	552.937	98,50%
	Baskets:	3.973.215	72.581	98,17%
Product Categories	Barcodes:	840.017	713.805	15,02%

Table 4.4 Impact in dataset volume

4.2. Cluster Sampling

In order to split the baskets in clusters according to their size and extract the outliers, SQL Server Data Tools (Visual Studio) was used again. This time the features of SQL Server Analysis Services (SSAS) package were exploited, to proceed with the cluster sampling.

As mentioned before, the retailer sent data about 8 stores, which had common characteristics in pairs. So, it has been decided to study them and extract the outliers in pairs. For that reason, four different views of the data in the table concerning the aggregated baskets, had been created. Each view represented each one of the four store types. After querying the appropriate table and creating the views, clustering details had to be set. The algorithm works as follows: the input data will be randomly split into two sets, a training set and a testing set, based on the selected percentage of the testing data. Training set is used to create the sampling model. Testing set is used to check model accuracy. Testing set was set 40%, so the training set was 60%. After having set clustering algorithms parameters, non-scalable K-means algorithm has been used as clustering method. According to the framework, the number of clusters hasn't been indicated, in order the algorithm to execute the optimum case.

After clustering results k-means indicated, there was a need to check the generated accuracy charts. After ensuring that the model's accuracy was approximately to the optimum model I ran queries in SSMS in order to specify the actual number of baskets that belong to each cluster, and the value-revenue derived from them. In the charts below are shown the segmentations resulting from the clustering, and their impact in both number of baskets and revenues-values. There are two different charts for each one of the store types. The first one concerns information about number of baskets, and the second one information about revenues. Each bar of the chart depicts a cluster. Regarding the first chart of each store type, the amount at the top of each bar indicates the actual number of baskets that belong to the cluster. The other amount indicates the percentage of baskets that belong to this cluster, related to the total number of baskets of each store type. The second one, indicates the revenue that concerns each cluster and the percentage compared with the total revenues of each store type. Information about revenues was helpful to decide the clusters that will be analyzed. Examining the charts below, the blue color clusters have been decided to be analyzed, as according to the above two factors, these were the most meaningful to mining the shopping missions.

View no. 1: Convenience stores

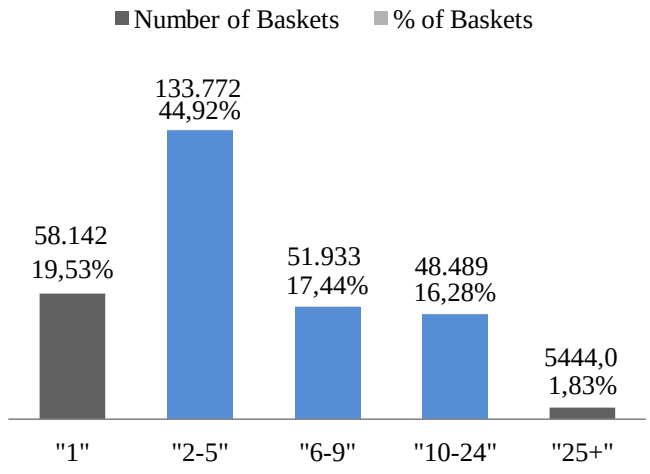


Figure 4.2 Convenience-Number of Baskets

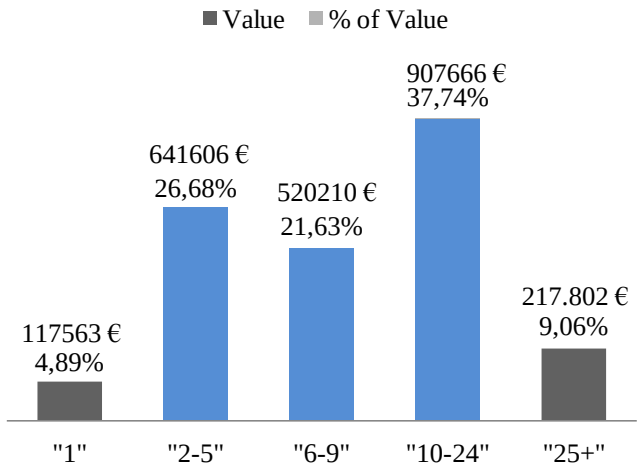


Figure 4.3 Convenience-Revenues

View no. 2 : Supermarkets

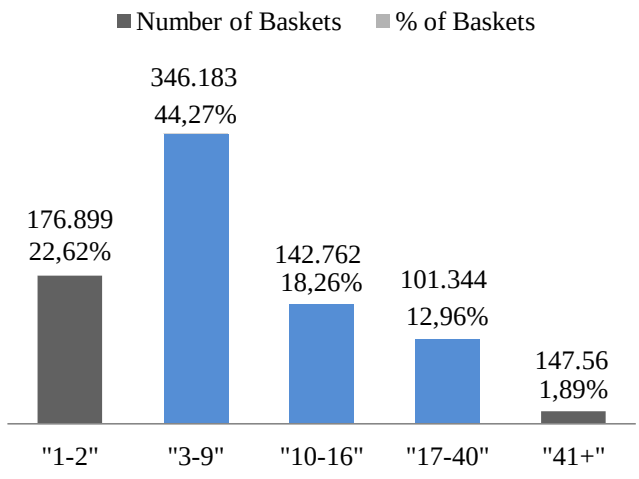


Figure 4.4 Supermarket- Number of Baskets

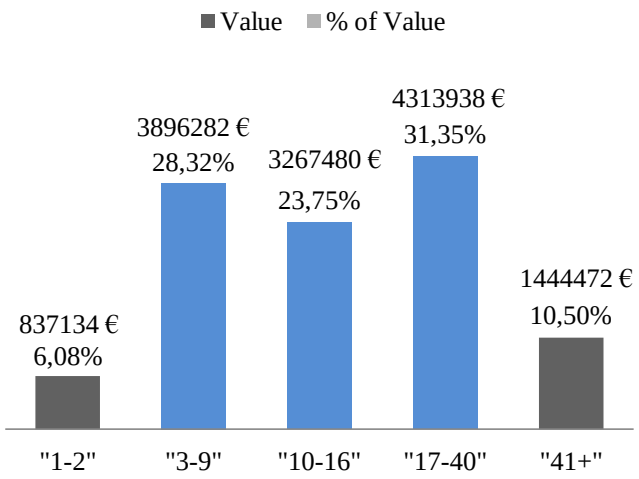


Figure 4.5 Supermarket-Revenues

View no. 3: Mini-hyper stores

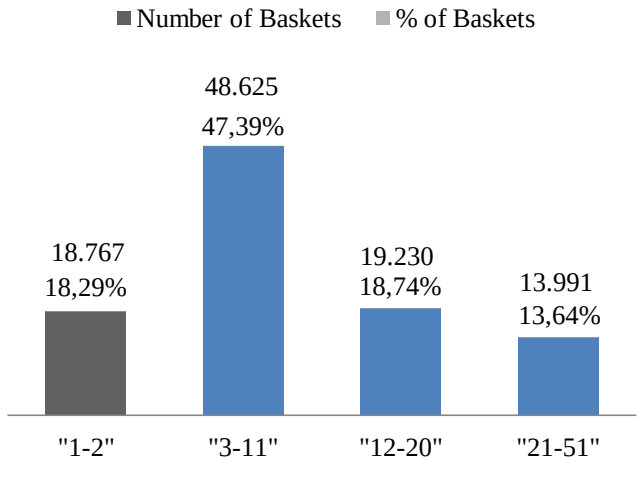


Figure 4.6 Mini-hyper - Number of Baskets

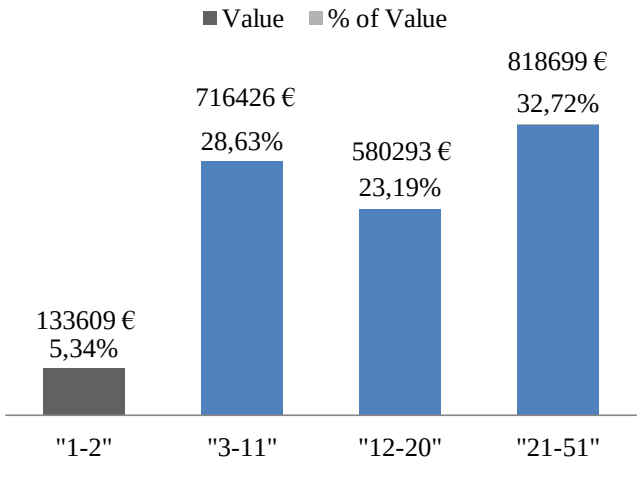


Figure 4.7 Mini-hyper- Revenues

View no. 4: Flag-hyper stores

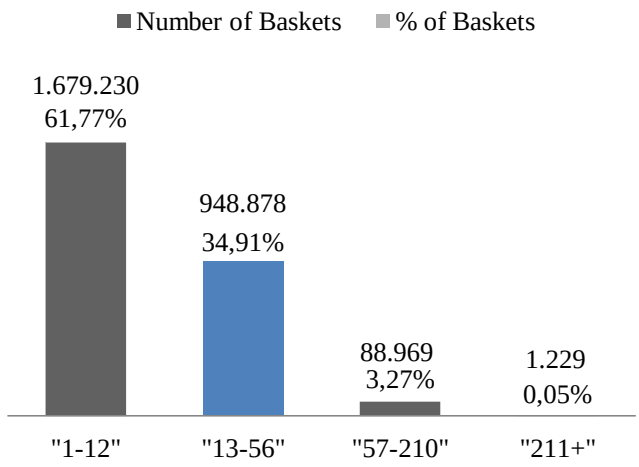


Figure 4.8 Flag-hyper- Number of Baskets

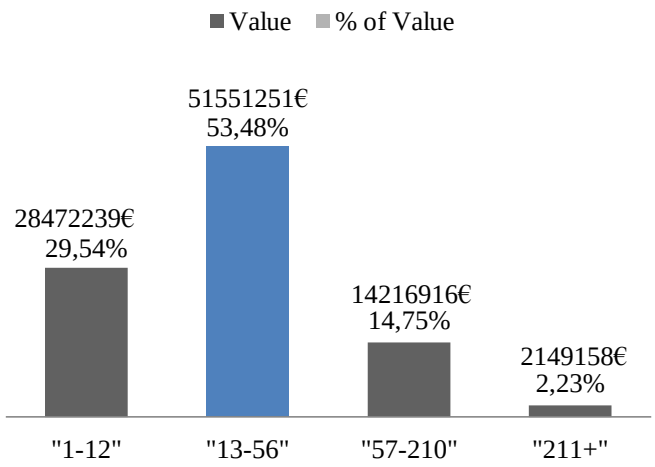


Figure 4.9 Flag-hyper- Revenues

In Table 4.5 there is a summary of the selections that have been made. Second row represents basket size range of each store type, that will be used in the forthcoming analysis. In the two last columns are shown the percentage of baskets which will be kept for the next step of the process, and the percentage of the revenue these baskets cause.

Store Type	Basket Size Range Sample	Total baskets	Percentage of the total baskets kept	Percentage of revenue kept
Convenience	2-24	231.194	78,64%	86,05%
Supermarket	3-40	590.289	75,49%	83,42%
Mini-Hyper	3-51	81.846	79,75%	84,54%
Flag-Hyper	13-56	948.878	34,91%	53,48%

Table 4.5 Summarized results of Cluster Sampling

4.3. Modeling

The purpose of the DM phase is to identify the category association patterns that will indicate the shopping missions in each separate store. For that needs, SQL Server Analysis Services (SSAS) package of SQL Server Data Tools (SSDT), will be used again as DM tool.

4.3.1. Input Dataset Adjustment

In order to ensure the interoperability of the dataset, there was a need to merge all the data of the database tables, into csv (Comma Separated Values) files. At this point, although there were stores having common characteristics, it has been decided to analyze each store separately. This happened because probably in same store types, different shopping missions may result. For that reason, Java code was used in order to create 8 csv files, one for each store. After making a second phase of data selection, fields concerning loyal customers were extracted from the final dataset. The csv files had the following columns:

- Basket identification
- Basket Value
- Number of items
- Date
- Store identification
- The 104 customized Categories
- Whether the Baskets was bought by a cardholder (True or False)
- Cardholder identification
- Age
- Marital status
- Household Size

After the csv files creation, the 8 above files were loaded in SQL Server, and were used as input fact tables in SSAS, in order to proceed with the modeling.

In this case it was not necessary to integrate the data into csv files. The tool that was used, could process and analyze the dataset in the existing format, although in order to implement the above, the tables need to be connected with foreign keys. In this phase

this task could happen, as with data preparation it has been ensured the homogeneity of the dataset, so previous integrity problems have been eliminated. With this alternative way the tables that should be linked were those concerning POS, loyal customers, and barcodes data. There was no need to use the basket table that concerned the aggregated information of POS data. The above task was quicker than the creation of the csv files, although csv files have been preferred. The above happened because the interoperability of the dataset was more important than receiving results in lower time.

4.3.2. Model Implementation

As mentioned, SSAS was used as DM tool, Microsoft Clustering as the tool to develop the DM model, and Microsoft non-scalable K-means as DMT.

Microsoft Clustering

SSDT provides Microsoft (MS) Clustering[4]. This is a technique used to place data elements into related groups-segments. Microsoft Clustering allows for more flexibility in input types and grouping methodologies. It uses iterative techniques to group cases in a dataset into clusters that contain similar characteristics. These groupings are useful for exploring data, identifying anomalies in the data, and creating predictions. MS Clustering first identifies relationships in a dataset and generates a series of clusters based on those relationships. After first defining the clusters, it calculates how well the clusters represent groupings of the points, and then tries to redefine the groupings to create clusters that better represent the data. It iterates through this process until it cannot improve the results more by redefining the clusters.

Microsoft K-means

K-means[5] clustering is a well-known method of assigning cluster membership by minimizing the differences among items in a cluster while maximizing the distance between clusters. Microsoft k-means algorithm provides two methods of sampling the data set: non-scalable K-means, which loads the entire data set and makes one clustering pass, and scalable k-means, where the algorithm uses the first 50.000 cases

[4] http://technet.microsoft.com/en-us/library/ms174879.aspx
[5] http://technet.microsoft.com/en-us/library/cc280445.aspx

and reads more cases only if it needs more data to achieve a good fit of model to data. For this research non-scalable k-means was used.

Eight models have been created and executed, one for each table that had the information about each store. For each store, the model was executed maintaining the same algorithm parameters with the others. According to the framework, in all the created models I didn't indicate the actual number of clusters, but left the algorithm execute the optimum case. Moreover, regarding the algorithm's parameters, all the product categories were used as input and predict values. Last but not least, it was necessary to generate a mechanism to test model's accuracy. For this reason, the testing set was set 40%, therefore the training set was 60%.

According to the above, the tool produced DM accuracy charts[6] for each one of the predicted values (categories) of each store. In Figure 4.10 is depicted an accuracy chart of the attribute "cereals" at a convenience store. The chart show how the model performs for all states of the predictable attribute "cereals" i.e. exists or not in a basket. Specifically, the y-axis represents the percentage of predictions that are correct, and the x-axis represents the percentage of the dataset that is used to compare the predictions. The blue line shows the ideal results for the training dataset, if you create a model that always predicts perfectly. The green line shows the actual lift for the model. The closer these two lines are, the more accurate the model is. For instance, at the grey line (60% of x-axis), the ideal model of predicting cereals shows that the 60% of the population is needed in order the model to correctly predict the 60% of the cases, which is the maximum that can be expected. Moreover, at the 60% of the population, the model that was built, predicts correctly the 57,83% of the cases. The rest accuracy charts of all the categories in all stores follows the same trend, as they have a successful rate which ranges from 54% to 59%, at the 60% of the population.

[6] http://technet.microsoft.com/en-us/library/ms175428.aspx

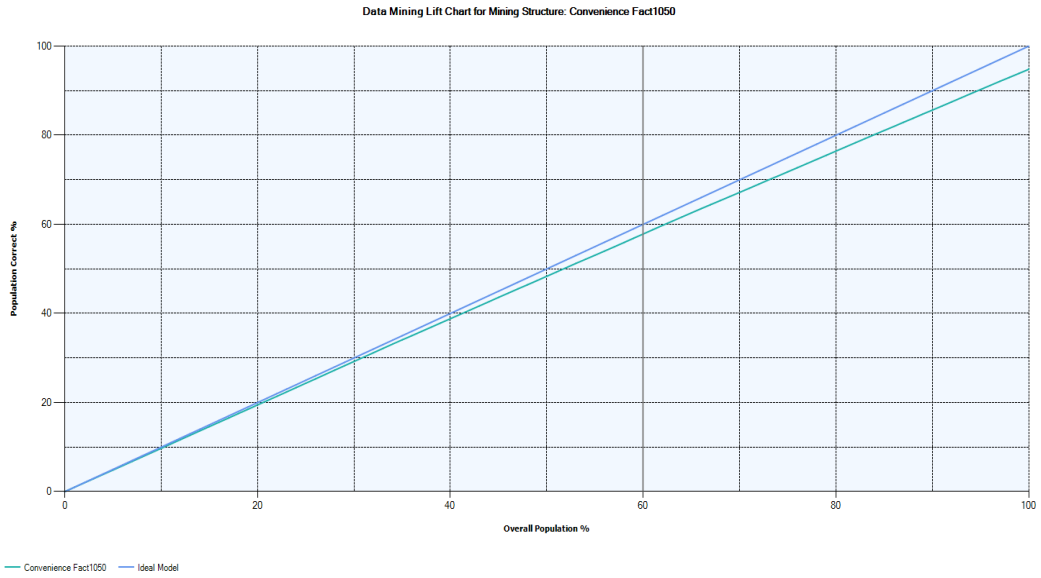


Figure 4.10 Accuracy chart of attribute "Cereals" at a convenience store

4.4. Evaluation

After receiving the first clustering results, and characterizing for the first time the shopping missions, there was a need to show them at the people of the industry, who know the domain. According to these results industry's people realize an important omission in the resulting clusters. In any cluster didn't exist categories related to meat. But, there were categories such as beef, burgers, pork, frozen meat, lambs etc appeared in low percentages. If these categories had been merged in a new higher level category, such as "red meat", they would have appeared in a higher percentage in a cluster. So, according to this guidance, these categories had been merged into one. Furthermore, categories related to beauty products and cosmetics, such as perfumes, beauty accessories, makeup etc had also been merged into one. Finally, about 15 categories were deleted, as they appeared in an average percentage less than 2% in all clusters.

The same changes were applied in all the stores by writing queries to modify the 8 tables that had been created above. Finally, from the 104 categories, the clustering process continued with the 75 of them. After the evaluation step, the DM model re-executed 8 times with the same parameters as described before, one for each one of the stores. After the re-execution of the DM model and the final approval of the industry people, the final clustering results were produced. Both in this and the next

phase, excel data mining add-in had been used as the bridge between the knowledge that was extracted from the DM, and the managers, as it provides a great visualization of the DM results.

4.5. Deployment

In Figure 4.1, there is the final cluster diagram for a supermarket. The more densely populated clusters have darker color. The shading of the line that connects one cluster to another represents the strength of the similarity of the clusters. If the shading is light or nonexistent, the clusters are not very similar. As the line becomes darker, the similarity of the links becomes stronger[7].

After the clustering results, in order to help us with the cluster identification, there was a need to calculate the average basket size (number of products), and the average unique categories each cluster contained. The cluster diagram of a supermarket can be seen in Figure 4.11. For each resulting cluster the two above factors have been calculated. According to the clustering results of the supermarket the cluster number 2 contains in high percentages categories related to fresh vegetables, red meat, chicken, white cheese, pasta, eggs, bread, oil, vinegar etc. So, according to the contribution of the percentages that these categories have, this cluster could be characterized as "Main Course Preparation". Cluster number 3 is named "Breakfast", as it contains categories related to milk, yogurt, cereals, coffee, tea, juice, fruit, chocolate, sweets etc. Cluster number 4 contains categories related to biscuits, sweets, candies, ice cream, chips, snacks, juice, beverages etc. So this cluster is called "Snack". Cluster number 5 contains product categories related to household cleaners, soft tissue paper products, disinfectants, detergents, shampoos, oral hygiene etc. This non-food cluster is named "Detergents and Hygiene". Cluster number 6 concerns categories related to packed cheese and cold cuts, bakery products etc, so it is a shopping mission for "Toast with packed products". Moreover, the same product categories as above are contained in cluster 8, with a difference that this time they were not packed, but fresh cutting products. So, cluster 8 is named "Toast with serviced products". Cluster number 7, is named "Light meal" as it contains product categories such as pasta, rice, pulses and canned food.

[7] http://technet.microsoft.com/en-us/library/ms174801.aspx

Clusters number 9 and 10 were more abstract, the first one concerned food products, and the other non-food products. Drill-down was performed in both the above clusters, but the results were not significant, as the occurring sub-clusters either contained the same shopping missions as those had been characterized before, or they didn't indicate a certain shopping purpose. Last but not least, cluster number 1 could not be characterized since it contains baskets with a few products and product categories.

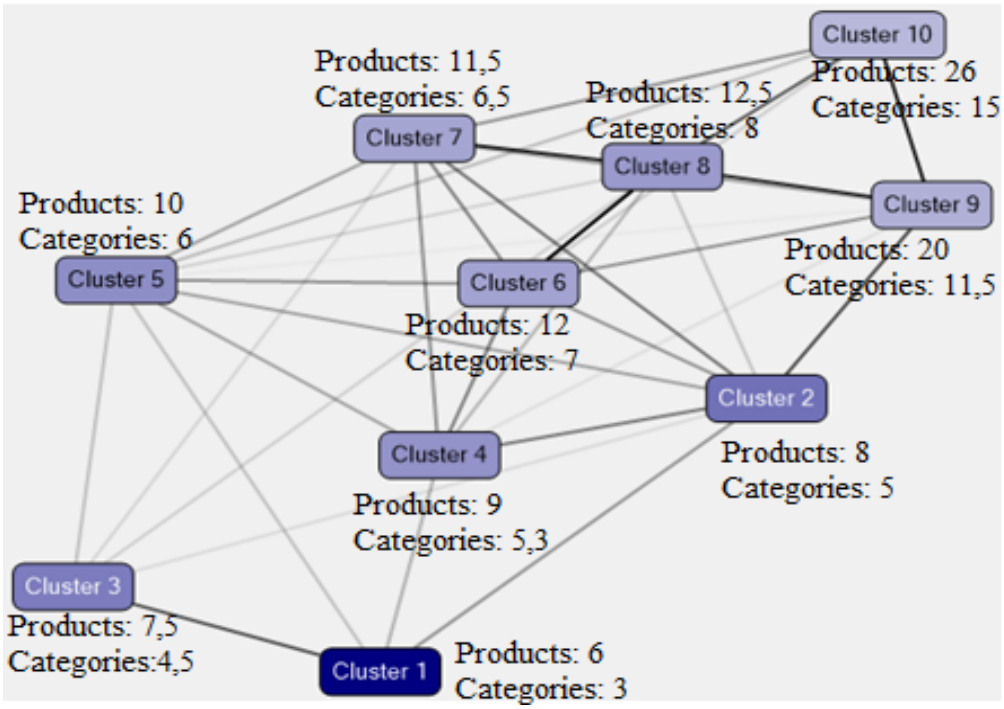


Figure 4.11 Cluster Diagram for a Supermarket

In the same context as above, for the other stores, the next shopping missions have occurred. The numbers in the brackets, next to each shopping mission represent the average basket size and the average number of unique categories each cluster contained. The rest Cluster Diagrams of these stores are cited at the Appendix (Appendices A-G).

For the other supermarket the resulting clusters are 10. The resulting shopping missions are the following:

- Cluster 2: Main Course Preparation - (Products: 7 , Categories: 4)
- Cluster 3: Breakfast - (Products: 7 , Categories: 4)
- Cluster 4: Breakfast and Vegetables - (Products: 11 , Categories: 6)

- Cluster 5: Detergents and Hygiene - (Products: 9 , Categories: 5)
- Cluster 6: Toast - (Products: 9 , Categories: 6)
- Cluster 7: Snack (with Beverages) - (Products: 8 , Categories: 5)
- Cluster 9: Coffee Break - (Products: 9 , Categories: 5)
- Cluster 10: Snack (around food) - (Products: 13 , Categories: 6)
- Cluster 1: Un-characterized food based cluster - (Products:5 , Categories: 3)
- Cluster 8: Un-characterized mixed food and non-food based cluster - (Products:25 , Categories: 14)

According to the clustering results of these two supermarkets, it is observed that, the resulting shopping missions resemble, but are not the same. In both supermarkets the smallest and the biggest clusters could not be characterized. Concerning the two small clusters, they contain too little product categories to indicate a shopping purpose. Regarding the three big clusters, they also could not be characterized, as according to the drill-down, they contained shopping missions that have already been found. The revised shopping missions in all these three clusters are: Detergents and Hygiene, and Light meal.

For the first convenience store the resulting clusters are 6. The resulting shopping missions are the following:

- Cluster 1: Snack (with beverages) - (Products: 5 , Categories: 3)
- Cluster 2: Breakfast - (Products: 5,5 , Categories: 3,5)
- Cluster 3: Snack (around food) - (Products: 7 , Categories: 4)
- Cluster 4: Toast - (Products: 10 , Categories: 7)
- Cluster 5: Detergents and Hygiene - (Products: 7,5 , Categories: 5)
- Cluster 6: Light meal - (Products: 9 , Categories: 6)

For the other convenience store the resulting clusters are 7. The resulting shopping missions are the following:

- Cluster 1: Snack (with beverages) - (Products: 4,5 , Categories: 2,5)
- Cluster 2: Breakfast - (Products: 5,5 , Categories: 3,5)
- Cluster 3: Snack (around food) - (Products: 5 , Categories: 4)
- Cluster 5: Toast - (Products: 8 , Categories: 5)
- Cluster 6: Light meal - (Products: 9 , Categories: 5,5)

- Cluster 4: Un-characterized mixed food and non-food based cluster - (Products: 6 , Categories: 3)
- Cluster 7: Un-characterized mixed food and non-food based cluster - (Products: 15, Categories: 10)

The resulting shopping missions of the convenience stores, are almost the same. The difference is that in the second store, there is not a shopping mission related to detergent and hygiene products. In the resulting clusters these product categories appeared in low percentages in almost every cluster, but they didn't appeared as prevailing categories in any cluster, or sub-cluster. Moreover, for the second convenience store the results of the drill-down didn't indicate a new hidden shopping mission, but they confirmed shopping missions that had already been found. Last but not least, the difference between this type of stores and the previous, is that customers in convenience stores buy less products than they buy in supermarkets, and their shopping missions basically concern food products.

For the first mini-hyper store the resulting clusters are 9. The resulting shopping missions are the following:

- Cluster 2: Semi-prepared food - (Products: 8 , Categories: 4)
- Cluster 3: Biological products - (Products: 11, Categories: 4)
- Cluster 4: Detergents and Hygiene - (Products: 11 , Categories: 4)
- Cluster 5: Main Course Preparation (meat based) - (Products: 13 , Categories: 5)
- Cluster 6: Main Course Preparation (fish based) - (Products: 12 , Categories: 6)
- Cluster 7: Main Course Preparation with Dessert - (Products: 29 , Categories: 13)
- Cluster 8: Detergents and Hygiene - (Products: 27 , Categories: 13)
- Cluster 9: Snacks and Animal Feed - (Products: 18 , Categories: 9)
- Cluster 1: Un-characterized food based cluster - (Products: 7 , Categories: 2)

For the other mini-hyper store the resulting clusters are 9. The resulting shopping missions are the following:

- Cluster 2: Main Course Preparation (with dessert) - (Products: 9 , Categories: 4)
- Cluster 3: Main Course Preparation (with dessert) - (Products: 19 , Categories: 9)
- Cluster 4: Detergents and Hygiene - (Products: 13 , Categories: 6)
- Cluster 5: Semi-Prepared food - (Products: 10 , Categories: 4)

- Cluster 7: Desserts and Sweet Preparation - (Products: 12 , Categories: 5)
- Cluster 9: Snacks and Animal Feed - (Products: 21 , Categories: 10)
- Cluster 1: Un-characterized food based cluster - (Products:7, Categories: 2)
- Cluster 6: Un-characterized mixed food and non-food based cluster - (Products: 11 , Categories: 5)
- Cluster 8: Un-characterized mixed food and non-food based cluster - (Products: 33 , Categories: 16)

For these two mini-hyper stores the resulting clusters and the derived shopping missions resemble a lot. For both stores there is a resulting cluster with too little product categories (only two), that can't be characterized. As far as concerns the second store, there are also two other uncharacterized resulting clusters. These clusters are: cluster 8, that contains a lot of products and product categories, and cluster 6 that contains only five product categories. Drill-down was applied in both these clusters, but the results were not meaningful, as the resulting sub-clusters were still abstract.

For the first flag-hyper store the resulting clusters are 10. The resulting shopping missions are the following:

- Cluster 1: Main Course Preparation (with dessert) - (Products: 20 , Categories: 9)
- Cluster 2: Semi-Prepared food - (Products: 19 , Categories: 4)
- Cluster 3: Semi-Prepared food and Desserts- (Products: 20 , Categories: 7)
- Cluster 4: Main Course Preparation (with dessert) - (Products: 22 , Categories: 9)
- Cluster 9: Main course Preparation with Dessert (meat based) - (Products: 34 , Categories: 15)
- Cluster 10: Detergents and Hygiene - (Products: 24 , Categories: 10)
- Cluster 5: Un-characterized mixed food and non-food based cluster - (Products: 25, Categories: 11)
- Cluster 6: Un-characterized mixed food and non-food based cluster - (Products: 23, Categories: 9)
- Cluster 7: Un-characterized mixed food and non-food based cluster - (Products: 26, Categories: 11)
- Cluster 8: Un-characterized mixed food and non-food based cluster - (Products: 40, Categories: 18)

For the other flag-hyper store the resulting clusters are 10. The resulting shopping missions are the following:

- Cluster 1: Main Course Preparation (with dessert) - (Products: 20 , Categories: 7)
- Cluster 4: Non-refrigerated long-term food supply - (Products: 25 , Categories: 10)
- Cluster 5: Detergents and Hygiene - (Products: 25 , Categories: 9)
- Cluster 7: Course Preparation meat based (with dessert) - (Products: 26 , Categories: 10)
- Cluster 8: Semi-prepared food - (Products: 21 , Categories: 8)
- Cluster 9: Main Course Preparation meat based (with dessert) - (Products: 27 , Categories: 12)
- Cluster 3: Un-characterized mixed food cluster - (Products: 23 , Categories: 8)
- Cluster 2: Un-characterized mixed food and non-food based cluster - (Products: 33, Categories: 14)
- Cluster 6: Un-characterized mixed food and non-food based cluster - (Products: 41, Categories: 18)
- Cluster 10: Un-characterized mixed food and non-food based cluster - (Products: 32 , Categories: 14)

These flag-hyper stores the Greek retailer gave data for, have a wide range of food and non-food products. So, as it has been expected customers had bought from these stores a lot of products (and product categories) without having a certain shopping purpose. This is the reason why the shopping missions in this type of stores were too general and abstract. Even by applying drill-down in the un-characterized clusters, still abstract and non-meaningful shopping missions derived.

At this point, it should be mentioned that, after all, the given loyalty data were not used. The initial purpose was to use the loyalty data in order to extract more valuable information about each resulting cluster-shopping mission. But, while exploring these data, it has been found out two significant things. Firstly, it has been noticed that many customers didn't filled their personal data such as age, sex, household size etc, so data quality was poor. Secondly, the percentage of purchases that loyal customers had made in the period covered by the given POS data, was significantly low.

5. CONCLUSIONS & DISCUSSION

5.1. Overview

Nowadays, computers have become far more powerful and new technological trends have been developed, such as Big Data, Business Intelligence (BI), Data Mining (DM). At the same time consumers' behavior and expectations for service have changed dramatically. Many enterprises have realized the importance of applying these new technological trends to support decision making and satisfy their customers. Motivated by the above, this thesis tried to assist retailers to satisfy their demanding customers, via exploiting these technological improvements. Firstly, by exploring this business problem, it has been found out a significant research gap, as there is only sparse research in the context of retailing in order to discover patterns in customers' behavior, to empower decision making, and to satisfy the demanding consumers. Trying to fill this gap this thesis presented a framework to identify consumers' shopping missions in a supermarket. The above was implemented using clustering as DM technique to identify the product categories that are purchased is the same shopping mission. This framework was successfully applied to a real case. This case concerned eight stores of a Greek retailer, these stores had common characteristics in pairs, as the retailer provided data for two convenience stores, two supermarkets, two mini-hyper markets and two hyper-markets. For each one of the eight stores the shopping missions were identified, but it had been noticed that the results of each store type resemble a lot. It had been also noticed that the resulting shopping missions of the two hyper stores were too general and abstract. This is related with framework limitations, as described below. The other resulting shopping missions were significant, and could be used in practice by the retailers to support CRM strategies; these practical implications are also described below.

5.2. Theory Contribution

To the best of my knowledge, there is no other framework - methodology which indicates how to extract shopping missions from retail data, by identifying correlations in product categories, using clustering. This proposed approach differs from the well-known "consumption universes" and market basket analysis in several aspects. First of all, other researches (Ahn, 2012; Borges, 2003; Chen, Chen, & Tung,

2006; Cil, 2012; Raorane et al., 2012; Shrivastava & Sahu, 2007) make category correlations in retail stores data by using association rule mining, apriori and nearest neighbor algorithms. This research introduces Clustering and k-means as DM model and technique. Furthermore, market baskets analysis is mainly devoted to a market-level analysis (Chen et al., 2006). The proposed framework is devoted to a store-level analysis, as it provides a specific way of how to extract customers' shopping missions within a store.

Moreover, it is the first study that takes into consideration and identifies the outliers (Cluster Sampling) that exist in the dataset i.e. too large and too small basket sizes. Other researchers, such as Cil (2012), extract from the dataset only baskets containing one product. Last but not least, by enriching this framework with loyal customers data, you can segment the customers in groups according to their purchases. Therefore, it could derive a new way to study and analyze customers in groups, except from the well-know RFM (Recency, Frequency, Monetary) analysis.

5.3. Practical Implications

The proposed framework could be used as a decision-making tool for different usages in the retailing domain. Specifically, it could be used by the managers as a complementary tool in order to modify and re-design the supermarket's layout, based on the customers' desires, as emerged from the DM. This allows retailers to cluster products around customers buying habits and appeal to busy customers. The above could happen by placing products that belong to the same shopping mission, in nearby supermarket's aisles and shelves. As mentioned above, since it is more convenient for consumers to find what they need by spending the least amount of time in the supermarket, customer satisfaction increases, but retailers profit as well. Moreover, managers could use the knowledge extracted from the framework to support commercial decision making in retailing. They could develop marketing campaigns and promotions for products that belong to the same shopping mission.

An important implication is that the framework's results could be also used by the marketers to increase customer loyalty. Marketers could design cross-coupon programs from products that belong to the same shopping mission. Additionally, they could use the resulting shopping missions to determine the display layout of the

products in the supermarket's on-line catalog, in order to increase the basket size of the online shoppers. Furthermore, another innovative implication could be to use the extracted knowledge to create a recommendation system for real time purchases in a supermarket. This system could propose to a customer the products that maybe forgot to buy, according to the shopping mission (or missions) that belong the products he already has in his basket. Last but not least, via this system, real-time cross-coupons could be offered.

5.4. Limitations & Further Research

An important factor-limitation that should be taken into consideration is the store type. If the store is too big, with a wide range of food and non-food products, it may be too difficult to extract shopping missions, even by applying drill-down. For instance, it could happen in a hypermarket. This happens because customers commonly visit this type of stores to make bulk purchases of several products. Hence, their basket contains a lot of products and product categories that do not have relevance, as their shopping purpose is not certain.

Further research could be conducted in order to apply another DM technique, such as association rules, and compare the resulting shopping missions with those that had been derived from clustering. Moreover, it would be of great interest to study the loyal customers' purchases, to identify their shopping missions, and then to compare these shopping missions, with those of non-loyal customers. Additionally, other research could be performed to examine whether there are alternative ways to identify the shopping missions. For instance, new research could be carried out to identify the shopping missions via using RFID (Radio Frequency Identification) in shopping carts to record customers' shopping paths in a store, or via conducting consumers' survey for their purchases while they are in a store. Then it could be notable to compare and contrast the resulting shopping missions of these different approaches.

In relation to the proposed framework, there are some issues that could be further studied. First of all, further research could take into consideration not only the product categories that are purchased in the same shopping mission, but also those categories that have been never purchased together. By analyzing the relation between the "non-purchased together" product categories, managers could use this information as an

additional factor in changing the supermarket's layout, or marketers could provide incentives to the customers in order to purchase these irrelevant categories simultaneously.

Last but not least, it could be examined whether is meaningful to implement cluster sampling in the number of categories each basket has, rather than in the number of products. This may be useful, since the whole analysis and the identification of shopping missions were based on products categories, and not on each unique product. In the proposed approach, there have been extracted baskets with many products, without taking into consideration if these products belong to fewer product categories. For example, there have been extracted baskets containing more than twenty-five products, without examining if these baskets include products from a small number of product categories i.e. five, that could be useful to identify a shopping mission.

REFERENCES

Ahn, K.-II. 2012. Effective product assignment based on association rule mining in retail. *Expert Systems with Applications*, 39(16): 12551-12556.

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.

Berry, L. L. 1983. *Emerging perspectives on services marketing*: Chicago, Ill.

Bertino, E., Bernstein, P., Agrawal, D., Davidson, S. 2011. 'Challenges and Opportunities with Big Data', A community white paper developed by leading researchers across the United States.

Borges, A. 2003. *Toward a new supermarket layout: from industrial categories to one stop shopping organization through a data mining approach*. Paper presented at the Proceedings of the 2003 Society for Marketing Advances Annual Symposium on Retail Patronage and Strategy, Montreal.

Borges, A., & Babin, B. 2010. KDD: Applying in Marketing Practice Using Point of Sale Information. In J. Casillas, & F. Martínez-López (Eds.), *Marketing Intelligent Systems Using Soft Computing*, Vol. 258: 35-41: Springer Berlin Heidelberg.

Bull, C. 2003. Strategic issues in customer relationship management (CRM) implementation. *Business Process Management Journal*, 9(Emerald): 592-602.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., & Shearer, C. 2000. CRISP-DM 1.0 Step-by-step data mining guide, *CRISP-DM Consortium*.

Chen, I. J., & Popovich, K. 2003. Understanding customer relationship management (CRM): People, process and technology. *Business Process Management Journal*, 9(5): 672 - 688.

Chen, Y.-L., Chen, J.-M., & Tung, C.-W. 2006. A data mining approach for retail knowledge discovery with consideration of the effect of shelf-space adjacency on sales. *Decision Support Systems*, 42(3): 1503-1520.

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

Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques* Springer-Verlag Berlin Heidelberg.

Grewal, D., Levy, M., & Kumar, V. 2009. Customer Experience Management in Retailing: An Organizing Framework. *Journal of Retailing*, 85(1): 1-14.

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

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. 2010. Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, 37(7): 5259-5264.

Humby, C., Hunt, T., & Phillips, T. 2003. *Scoring Points: How Tesco is winning customer loyalty*: Kogan Page, London Hardback.

Jeevananda, S. 2011. Study on Customer Satisfaction Level at Hypermarkets in Indian Retail Industry. *The International Journal's – Research Journal of Social Science and Management*, 1(3): 2.

Kurgan, L. A., & Musilek, P. 2006. A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, 21(01): 1-24.

Larson, J. S., Bradlow, E. T., & Fader, P. S. 2005. An exploratory look at supermarket shopping paths. *International Journal of Research in Marketing*, 22(4): 395-414.

Lenzerini, M. 2002. Data integration: a theoretical perspective, *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*: 233-246. Madison, Wisconsin: ACM.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. 2012. Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications*, 39(12): 11303-11311.

Ling, R., & Yen, D. 2001. Customer relationship management: An analysis framework and implementation strategies. *The Journal of Computer Information Systems*(41): 82-97.

Linoff, G. S., & Berry, M. J. 2011. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3rd ed.): Wiley.

Min, H. 2006. Developing the profiles of supermarket customers through data mining. *The Service Industries Journal*, 26(7): 747-763.

Minami, C., & Dawson, J. 2008. The CRM process in retail and service sector firms in Japan: Loyalty development and financial return. *Journal of Retailing and Consumer Services*, 15(5): 375-385.

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

Phan, D. D., & Vogel, D. R. 2010. A model of customer relationship management and business intelligence systems for catalogue and online retailers. *Information & Management*, 47(2): 69-77.

Provost , F., & Fawcett, T. 2013. Data Science and its Relationship to Big Data and Data-Driven Decision Making *Big Data*, 1(1): 51-59.

Rahm, E., & Do, H.-H. 2000. Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, 23(4): 3-13.

Raorane, A., Kulkarni, R., & Jitkar, B. 2012. Association Rule–Extracting Knowledge Using Market Basket Analysis. *Research Journal of Recent Sciences*, 1: 19-27.

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

Shahbaba, B. 2012. Data Exploration, *Biostatistics with R*: 17-59: Springer New York.

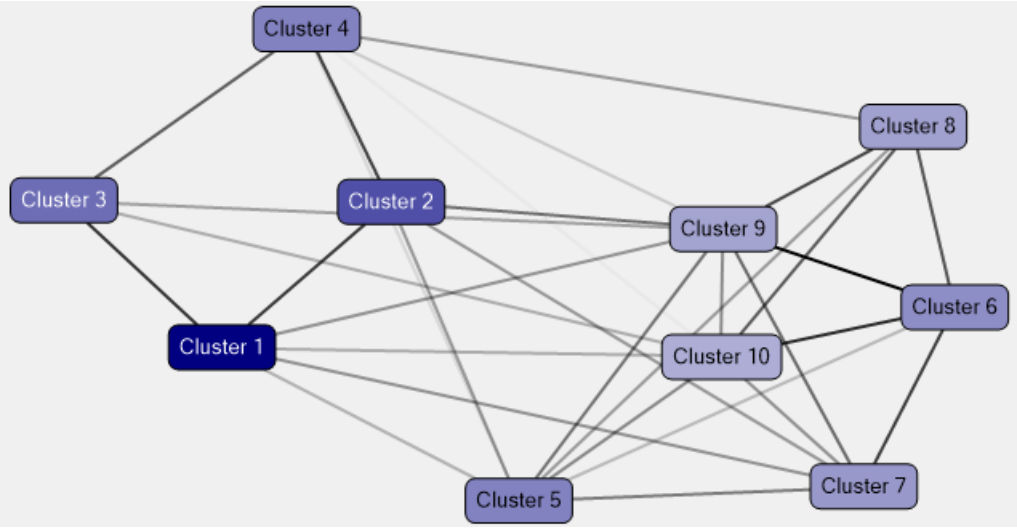
Sharda, R., Asamoah, D. A., & Ponna, N. 2013. Business analytics: Research and teaching perspectives. In V. Luzar-Stiffler, & I. Jarec (Eds.), *Proceedings of the ITI 2013 35th International Conference on Information Technology Interfaces*: 19-27. Cavtat / Dubrovnik, Croatia: IEEE.

Shrivastava, A., & Sahu, R. 2007. Efficient Association Rule Mining for Market Basket Analysis. *Global Journal of e-Business & Knowledge Management*, 3(1): 21-25.

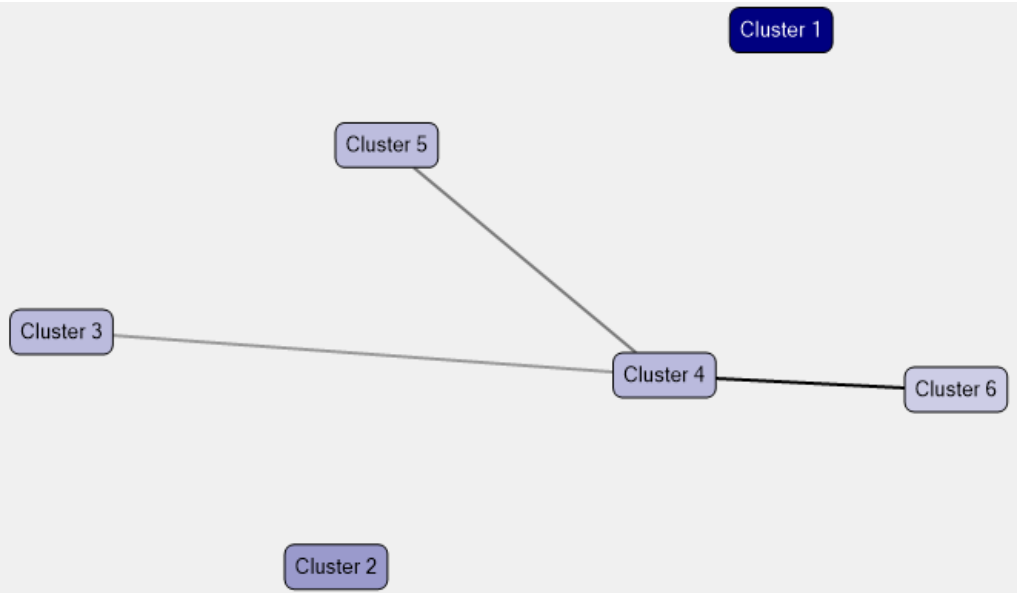
Wang, Y., & Zhou, T. 2013. Research of Data Mining in Customer Relationship Management. In Y. Yang, & M. Ma (Eds.), *Proceedings of the 2nd International Conference on Green Communications and Networks 2012 (GCN 2012): Volume 2*, Vol. 224: 163-170: Springer Berlin Heidelberg.

Yang, Y., & Ma, M. 2013. *Proceedings of the 2nd International Conference on Green Communications and Networks 2012 (GCN 2012): Volume 5*: Springer.

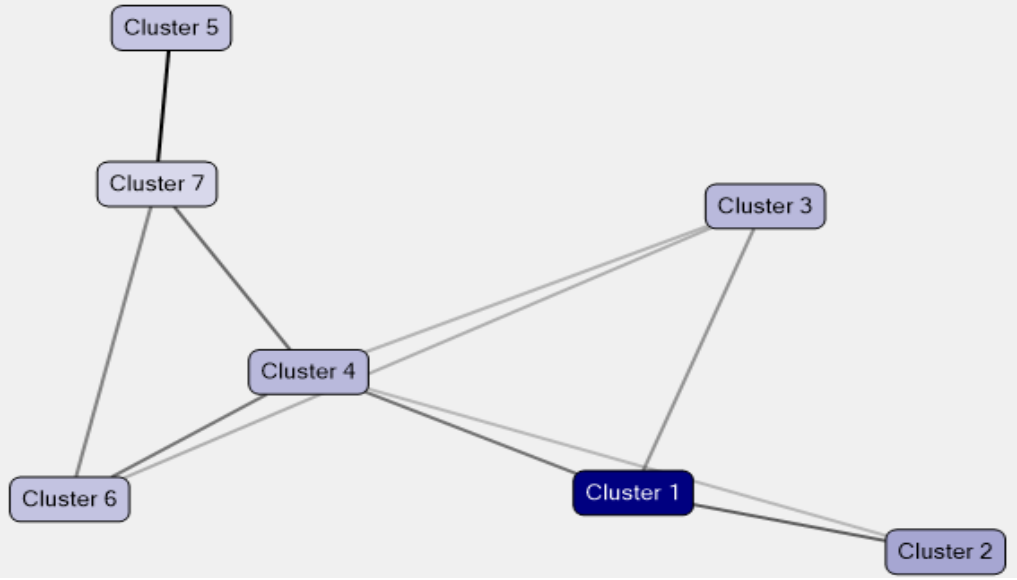
APPENDIX



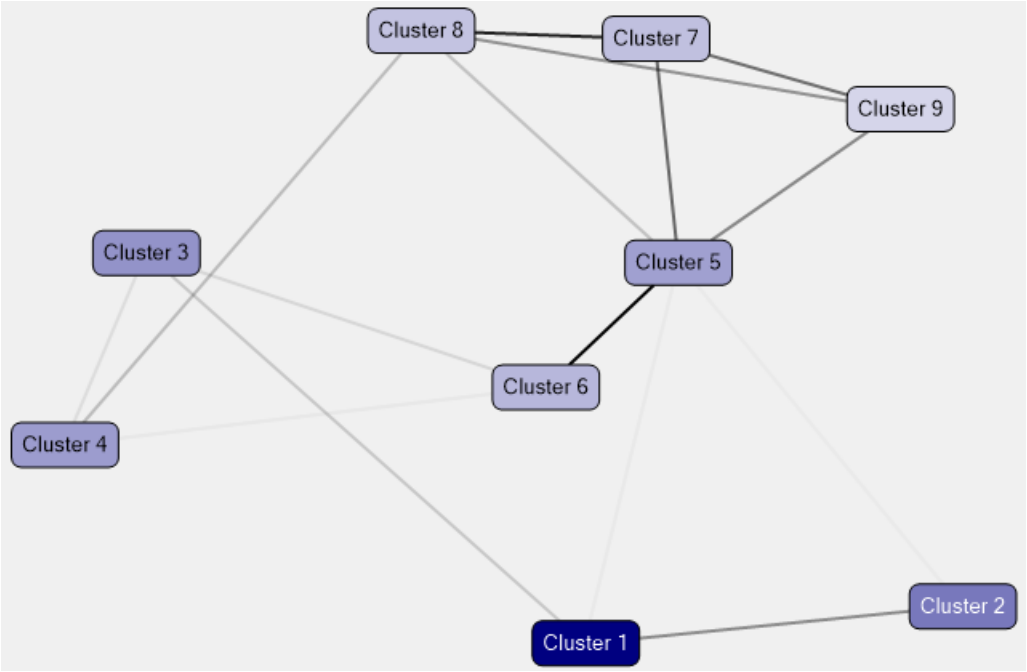
Appendix - Figure 1 Supermarket Number 2 -Cluster Diagram



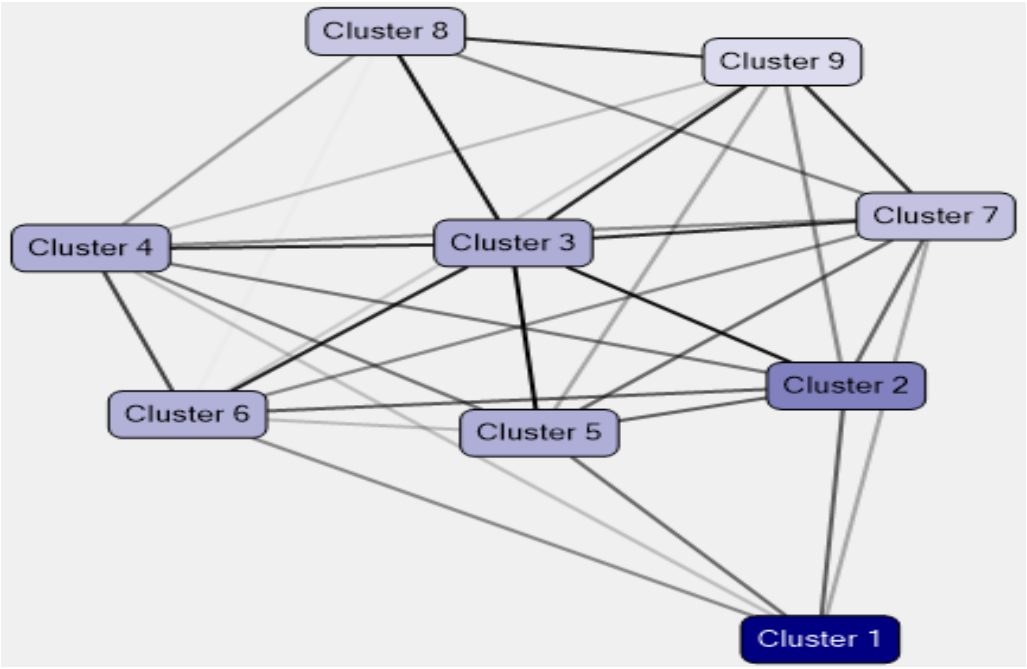
Appendix - Figure 2 Convenience Store Number 1 -Cluster Diagram



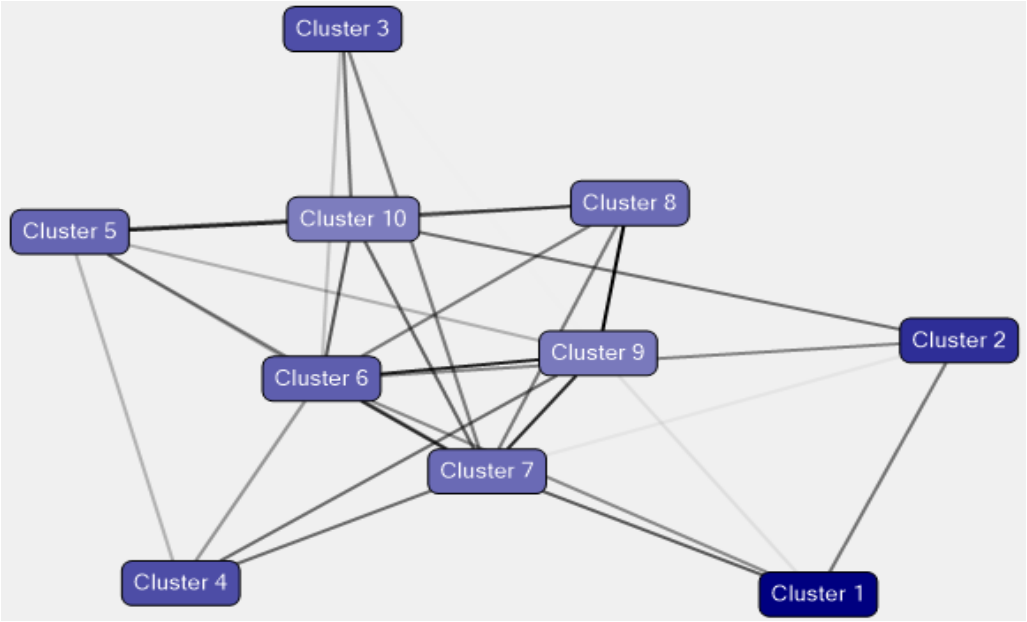
Appendix - Figure 3 Convenience Store Number 2 -Cluster Diagram



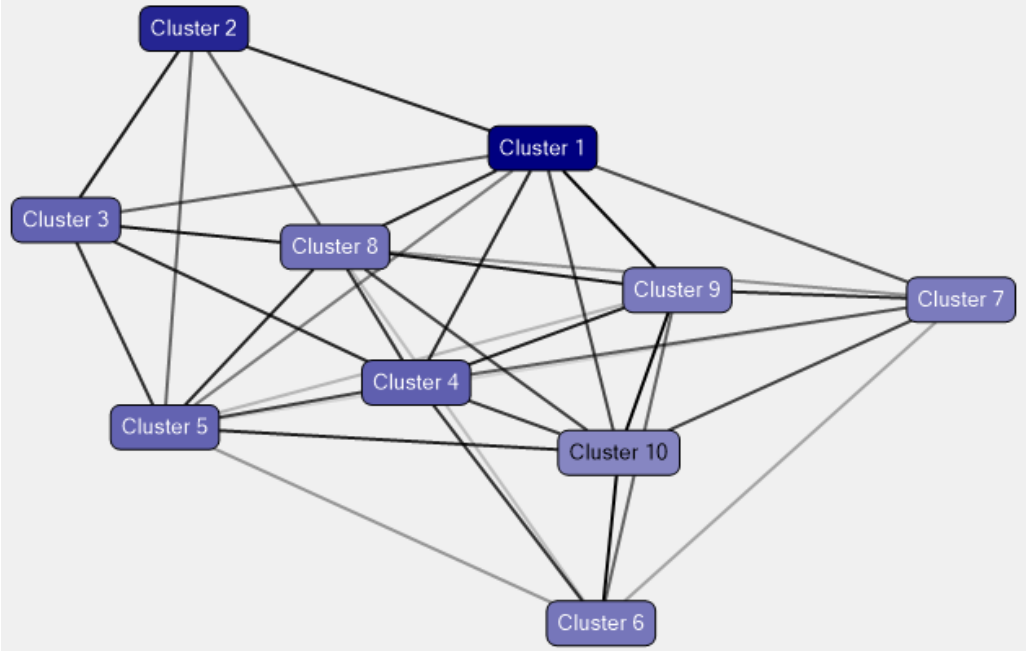
Appendix - Figure 4 Mini-Hyper Store Number 1 -Cluster Diagram



Appendix - Figure 5 Mini-Hyper Store Number 2 -Cluster Diagram



Appendix - Figure 6 Flag-Hyper Store Number 1 -Cluster Diagram



Appendix - Figure 7 Flag-Hyper Store Number 2 -Cluster Diagram