Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τμήμα Πληροφορικής

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Ομαδοποίηση Πελατών Με Μεθόδους Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών

Βασίλειος Λιούτας

Επιστημονικός Υπεύθυνος:
Καθ. Ιωάννης Βλαχάβας

Ιούνιος 2016

Aristotle University of Thessaloniki
Computer Science Department

THESIS

Customer Segmentation Using Methods Of Multiple Correspondence Analysis

Vasileios Lioutas

Thesis Supervisor:

Prof. Ioannis Vlahavas

June 2016

To Vasia, my partner in crime.

Περίληψη

Η σύγχρονη επιχείρηση για να επιβιώσει, χρειάζεται να παίρνει τις σωστές αποφάσεις την κατάλληλη στιγμή. Η διαδικασία λήψης απόφασης της επιχείρησης είναι μια σύνθετη διαδικασία και οι απαιτήσεις της αγοράς απαιτούν από αυτή να είναι συνεπής και ευέλικτη. Αυτό που κάνει την επιχείρηση να υφίσταται είναι οι πελάτες της. Θα πρέπει να είναι σε θέση να μαθαίνει τις ανάγκες τους αλλά ταυτόχρονα να μπορεί να καθιερώνει και τις δικές της τάσεις στην αγορά. Η Επιχειρηματική Ευφυΐα και η Μηχανική Μάθηση σήμερα μπορούν να προσφέρουν πολλά εργαλεία, που θα μπορούσαν να βοηθήσουν την επιχείρηση να γνωρίσει καλύτερα το πελατειακό της κοινό, μέσα από την οργάνωσή τους σε ομάδες με βάση κάποια κοινά χαρακτηριστικά. Ένα πολύ σημαντικό στοιχείο για την επιχείρηση και ιδιαίτερα για το τμήμα Marketing που διαθέτει, είναι να γνωρίσει και να ξεχωρίσει ομάδες πελατών με βάση την αγοραστική τους συμπεριφορά, δηλαδή τα προϊόντα ή τις κατηγορίες προϊόντων που αγοράζουν. Σκοπός της πτυχιακής εργασίας είναι η ανάπτυξη μιας διαδικτυακής εφαρμογής επιχειρηματικής ευφυίας με το πακέτο Shiny της στατιστικής γλώσσας R, για την ομαδοποίηση πελατών με βάση την αγοραστική τους συμπεριφορά, εφαρμόζοντας τη μέθοδο της Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών (MCA) και του Ιεραρχικού Αλγορίθμου, δίνοντας όλα τα απαραίτητα εργαλεία στο χρήστη ώστε να γνωρίσει τα δεδομένα του, να εμβαθύνει και να βελτιστοποιήσει την ανάλυση της ομαδοποίησης.

Λέξεις Κλειδιά: «Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών, Ιεραρχικός Αλγόριθμος, Επιχειρηματική Ευφυΐα, Μηχανική Μάθηση, Ανάλυση Καλαθιού Αγοράς, Shiny, R»

Abstract

In order for a modern business to survive, the right decisions must be made at the right time. The process of making a decision for a business is a complex one, and the market demands from it to be consistent and flexible. What defines a business is its customers. Not only does a business have to know its customers' needs, but it also has to be able to set its own trends on the market. Today, Business Intelligence and Machine Learning, can offer various tools that would allow a business to learn more about its customer basis by organizing them into clusters based on specific similar characteristics. It is vital for a business, and particularly for its Marketing division, to know and differentiate between its customers based on their purchase behavior, or in other words the products or groups of products customers tend to buy. The goal of this thesis is the development of an online business intelligence application, using the Shiny package of the R statistical language, to group customers based on their purchase behavior, implementing the Multiple Correspondence Analysis (MCA) method and the method of the Hierarchical Algorithm, and giving the user all the tools necessary to learn more about their data, delve deeper and improve their clustering analysis.

Keywords: «Multiple Correspondence Analysis, Hierarchical Algorithm, Business Intelligence, Machine Learning, Market Basket Analysis, Shiny, R»

Ευχαριστίες

Αισθάνομαι την ανάγκη, πριν την παρουσίαση των αποτελεσμάτων της παρούσας εργασίας, να εκφράσω την απέραντη ευγνωμοσύνη μου σε κάποιους ανθρώπους που άμεσα ή έμμεσα βοήθησαν καθοριστικά στην εκπόνηση της.

Πρώτα απ' όλα θα ήθελα να ευχαριστήσω τον καθηγητή και επιβλέποντα μου κ. Ιωάννη Βλαχάβα για την ανάθεση του θέματος, την πολύτιμη βοήθειά του, το ενδιαφέρον του αλλά και το χρόνο που διέθεσε για την διεκπεραίωση της πτυχιακής εργασίας.

Ιδιαίτερα θα ήθελα να ευχαριστήσω, τον συν-επιβλέποντα της διατριβής μου, υποψήφιο διδάκτωρ και μέντορα μου Ανέστη Φαχαντίδη που με ανέχτηκε όλο αυτό το διάστημα. Υπήρξε καταλυτικός αρωγός σε όλη τη προσπάθεια. Πάντα φιλικός, προσιτός και συνεργάσιμος, μου έδωσε όλα εκείνα τα απαραίτητα εφόδια για να αντεπεξέλθω στις όποιες δυσκολίες εμφανίστηκαν κατά τη διάρκεια της εργασίας. Η ευκαιρία να συνεργαστώ μαζί του αποτέλεσε κίνητρο στο να αποκτήσω μια πρώτη εμπειρία στο χώρο της έρευνας και να αγαπήσω ακόμα περισσότερο το αντικείμενό μας.

Θα ήθελα ακόμα να ευχαριστήσω όλους τους διδάσκοντες του τμήματος, για τις γνώσεις που μου προσέφεραν αδιάκοπα τα τελευταία τέσσερα χρόνια, με τις οποίες έγινε εφικτή η πραγματοποίηση της παρούσας εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω το οικογενειακό και φιλικό μου περιβάλλον για τη συνεχή υποστήριξη τους, η οποία έπαιξε σημαντικό ρόλο στην πραγματοποίηση και ολοκλήρωση των σπουδών μου.

Βασιλειος Λιοτταs
Θεσσαλονίκη
Ιούνιος 2016

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1

Εισαγωγή

1.1 Αντικείμενο της Διπλωματικής

Η επιβίωση μιας επιχείρησης σήμερα σχετίζεται με μια καθημερινή προσπάθεια συνεχούς βελτίωσης. Ο ανταγωνισμός είναι μεγάλος και απαιτείται από αυτή να μπορέσει να ανταποκριθεί σε αυτό. Χρειάζεται συνεχώς να μπορεί να βρίσκει λύσεις ώστε να είναι ευέλικτη στις αδιάκοπες αλλαγές της αγοράς. Η πληροφορική και η επιστήμη των δεδομένων έχουν αλλάξει τον τρόπο που δρα μια σύγχρονη επιχείρηση. Έχει δώσει σε αυτή εργαλεία με τα οποία μπορεί πλέον να μαθαίνει οργανωμένα και μεθοδευμένα από το παρελθόν της αλλά και να προβλέπει με αρκετή ακρίβεια την πορεία που πρέπει να ακολουθήσει. Αυτό που κάνει την επιχείρηση βιώσιμη, είναι η ικανότητα της να μπορεί ταυτόχρονα να ακολουθεί και να γνωρίζει τους πελάτες της αλλά συγχρόνως να είναι σε θέση να προτείνει νέες κατευθύνσεις στην αγορά. Γι΄ αυτό είναι σημαντικό και αναγκαίο για εκείνη να γνωρίζει της επιθυμίες των πελατών της όπως επίσης και τα προφίλ των αγορών τους. Πρέπει να γνωρίζει την πραγματική αγοραστική αξία των προϊόντων και των υπηρεσιών που προσφέρει ώστε να βελτιώνεται συνεχώς.

Η ευρύτερη περιοχή της Μηχανικής Μάθησης έχει εργαλεία που κάθε επιχείρηση σήμερα θα πρέπει να αξιοποιήσει αν επιθυμεί να είναι ανταγωνιστική. Μπορεί να βοηθήσει καταλυτικά στο πρόβλημα της γνώσης της αγοράς. Οι μάνατζερ και οι αναλυτές της επιχείρησης μπορούν να χρησιμοποιήσουν αυτά τα εργαλεία ώστε να καταφέρουν να αποκτήσουν μια εικόνα και μια διαίσθηση των πελατών τους και τι αυτοί προτιμούν. Έτσι μπορούν να μάθουν τι είναι αυτό που κάνει την επιχείρηση τους ανταγωνιστική και να το προωθήσουν κατάλληλα.

1.2 Σκοπός της Διπλωματικής

Μία από τις πιο γνωστές τεχνικές από τη βιβλιογραφία αναφέρει στο διαχωρισμό και την ομαδοποίηση των πελατών με βάση κάποια κοινά χαρακτηριστικά. Στην περίπτωση των προϊόντων ή των κατηγοριών αυτών σαν κοινά χαρακτηριστικά, το πρόβλημα που εγείρεται αφορά στο ότι οι πιο γνωστές μέθοδοι

ομαδοποίησης δεν επιδέχονται ποιοτικές μεταβλητές παρά μόνο συνεχείς. Η μέθοδος που έρχεται να λύσει αυτό το πρόβλημα αναφέρεται ως Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών (Multiple Correspondence Analysis - MCA) και είναι μια μέθοδος της οικογένειας μεθόδων Πολυμεταβλητής Ανάλυσης (Multivariate Analysis).

Η παρούσα πτυχιακή εργασία αναπτύσσει μια διαδικτυακή εφαρμογή επιχειρηματικής ευφυίας με τη χρήση της βιβλιοθήκης Shiny στη γλώσσα R που εφαρμόζει τη μέθοδο της Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών σε συνδυασμό με τον Ιεραρχικό Αλγόριθμο Ομαδοποίησης σε ένα οποιοδήποτε σύνολο δεδομένων από πελάτες και κατηγορίες προϊόντων, με σκοπό τη δημιουργία ομάδων που ακολουθούν κάποια κοινά αγοραστικά χαρακτηριστικά.

Η MCA στοχεύει στην προσπάθεια μετασχηματισμού των ποιοτικών δεδομένων σε ένα νέο χώρο συνεχούς τιμών και στη συνέχεια την εφαρμογή του Ιεραρχικού Αλγορίθμου για την υλοποίηση της ομαδοποίησης στις καινούργιες πλέον συνεχείς μεταβλητές.

Μέσο αυτής της εφαρμογής ο αναλυτής είναι σε θέση να πάρει δει ποια προϊόντα αγοράζονται συχνά μαζί και να δώσει τα αποτελέσματα στο τμήμα Marketing της επιχείρησης για να τα αξιοποιήσει κατάλληλα.

1.3 Δομή της Διπλωματικής

Μέχρι το σημείο αυτό, έγινε μία σύντομη εισαγωγή που αναφέρει επιγραμματικά τις επιστημονικές περιοχές και τις μεθόδους που χρησιμοποιούνται στην παρούσα εργασία.

Το **2ο Κεφάλαιο** εισάγει την έννοια της Επιχειρηματικής Ευφυίας και των Συστημάτων αυτής, στο οποίο περιγράφεται το θεωρητικό υπόβαθρο, όπως επίσης αναφέρονται λεπτομερώς όλες οι μεθοδολογίες της Επιχειρηματικής Ευφυίας.

Στο **3ο Κεφάλαιο** περιγράφονται οι βασικές έννοιες μιας Αποθήκης Δεδομένων καθώς αναλύεται και η μέθοδος της Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών και ο Ιεραρχικός Αλγόριθμος.

Το **4ο Κεφάλαιο** κάνει αναφορά στην γλώσσα R και στο πακέτο Shiny και περιγράφει την υλοποίηση της διαδικτυακής εφαρμογής, μέσω της οποίας εφαρμόζεται η μέθοδος της Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών για την παραγωγή ομάδων.

2

Επιχειρηματική Ευφυΐα

2.1 Εισαγωγή

Επιχειρηματική Ευφυΐα (BI) είναι η διαδικασία ανάλυσης, εξόρυξης και παρουσίασης αξιοποιήσιμης πληροφορίας, με στόχο να βοηθήσει κατά βάση τις επιχειρήσεις και τα άτομα που τις στελεχώνουν, να πάρουν περισσότερο έγκυρες και έγκαιρες αποφάσεις, να κατανοήσουν την αγοραστική συμπεριφορά και να αξιοποιήσουν ευκαιρίες. Αποτελείται από μια πληθώρα τεχνικών και εργαλείων που επιτρέπουν 1) την συλλογή και την οργάνωση (ακατέργαστων) δεδομένων τόσο από την ίδια την επιχείρηση όσο και από εξωτερικές πηγές, 2) την (προ)επεξεργασία τους ώστε να είναι έτοιμα για ανάλυση, 3) την ανάπτυξη στατιστικών μοντέλων και μοντέλων πρόβλεψης και 4) τη δημιουργία αναφορών (analytics), πινάκων και απεικονίσεων των δεδομένων και την διάθεσή τους στα στελέχη της επιχείρισης.

Σύμφωνα με την βιβλιογραφία υπάρχουν αρκετοί ορισμοί της επιχειρηματικής ευφυΐας, όπου κάθε ένας περιγράφει το αντικείμενο από μια διαφορετική οπτική γωνία. Μερικοί από αυτούς είναι:

- Η Επιχειρηματική Ευφυΐα ορίζεται ως η ικανότητα που έχει μια επιχείρηση να συλλάβει τις αλληλεξαρτήσεις που παρουσιάζονται από τα γεγονότα με τέτοιο τρόπο ώστε να καθοδηγήσει τις ενέργειες της προς την κατεύθυνση ενός επιθυμητού στόχου.

  — *Hans Peter Luhn, 1958*

- Η Επιχειρηματική Ευφυΐα δεν είναι ούτε ένα προϊόν ούτε κάποιο σύστημα. Είναι μια αρχιτεκτονική και μια συλλογή από ολοκληρωμένες λύσεις, όπως επίσης εφαρμογών υποστήριξης λήψεων αποφάσεων και βάσεων δεδομένων που παρέχει στις επιχειρήσεις εύκολη πρόσβαση στα δεδομένα τους.

  — *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications - Larissa T. Moss, Shaku Atre*

- Οι διαδικασίες, οι τεχνολογίες και τα εργαλεία που απαιτούνται για να μετατρέψουν τα δεδομένα σε πληροφορίες, τις πληροφορίες σε γνώση και τη γνώση σε αποφάσεις που θα οδηγήσουν την επιχείρηση σε κέρδη. Η Επιχειρηματική Ευφυΐα περιλαμβάνει αποθήκες δεδομένων, αναφορές επιχειρήσεων και διαχείριση γνώσης.

  — *The Data Warehouse Institute, Q4/2002*

- Η Επιχειρηματική Ευφυΐα ορίζεται ως «η γνώση που αποκτήθηκε για μία επιχείρηση μέσω της χρήσης διαφόρων τεχνολογιών, υλικού και λογισμικού, τα οποία επιτρέπουν στους οργανισμούς να μετατρέψουν τα δεδομένα σε πληροφορίες».

  — *Data Management Review*

- Η Επιχειρηματική Ευφυΐα επιφέρει μεγαλύτερη αντίληψη της εικόνας των επιχειρήσεων και επιτρέπει σε αυτές να γίνουν πιο αποτελεσματικές, ευέλικτες και ανταγωνιστικές.

  — *Oracle*

- Η Επιχειρηματική Ευφυΐα απλοποιεί την ανακάλυψη γνώσης και τις αποδοτικές αναλύσεις, δίνοντας τη δυνατότητα στην λήψη αποφάσεων όλων των επιπέδων να είναι γίνεται πιο εύκολα προσβάσιμη, κατανοητή, συνεργάσιμη και πιο αναλυτική.

  — *Microsoft*

2.2 Εισαγωγή στα Συστήματα Επιχειρηματικής Ευφυΐας

Ένα σύγχρονο σύστημα επιχειρηματικής ευφυΐας, περιλαμβάνει μια σειρά από μεθοδολογίες και εργαλεία που βοηθούν την επιχείρηση να γίνει πιο αποτελεσματική και κερδοφόρα. Με τη χρήση ενός τέτοιου συστήματος τα στελέχη μίας επιχείρησης μπορούν να αντλήσουν όλη τη δύναμη των δεδομένων της και να οργανώσουν προσεγμένες στρατηγικές όπου θα οδηγήσουν την επιχείρηση ένα επίπεδο πιο ψηλά από τον ανταγωνισμό.

Μερικά από τα πλεονεκτήματα των συστημάτων επιχειρηματικής ευφυΐας περιλαμβάνουν ταχύτερη, ορθότερη και ακριβέστερη λήψη αποφάσεων, βοηθώντας στην βελτίωση της συνολικής λειτουργίας και αποδοτικότητας της επιχείρησης. Τα συστήματα αυτά μπορούν να περιέχουν τόσο ιστορικά όσο και νεότερα δεδομένα από διάφορες πηγές, επιτρέποντας την ανάλυση επιχειρηματικής ευφυΐας να υποστηρίζει στρατηγικές και τακτικές απόφασης. Συνδυάζουν ένα ευρύ σύνολο από εργαλεία και εφαρμογές όπως άμεση επεξεργασία δεδομένων, ανάλυση πραγματικού χρόνου, mobile επιχειρηματική ευφυΐα κ.ά. αλλά

και τεχνικές οπτικοποίησης των δεδομένων με διαγράμματα και γραφήματα όπως και σε συγκεντρωτικούς πίνακες αναφοράς.

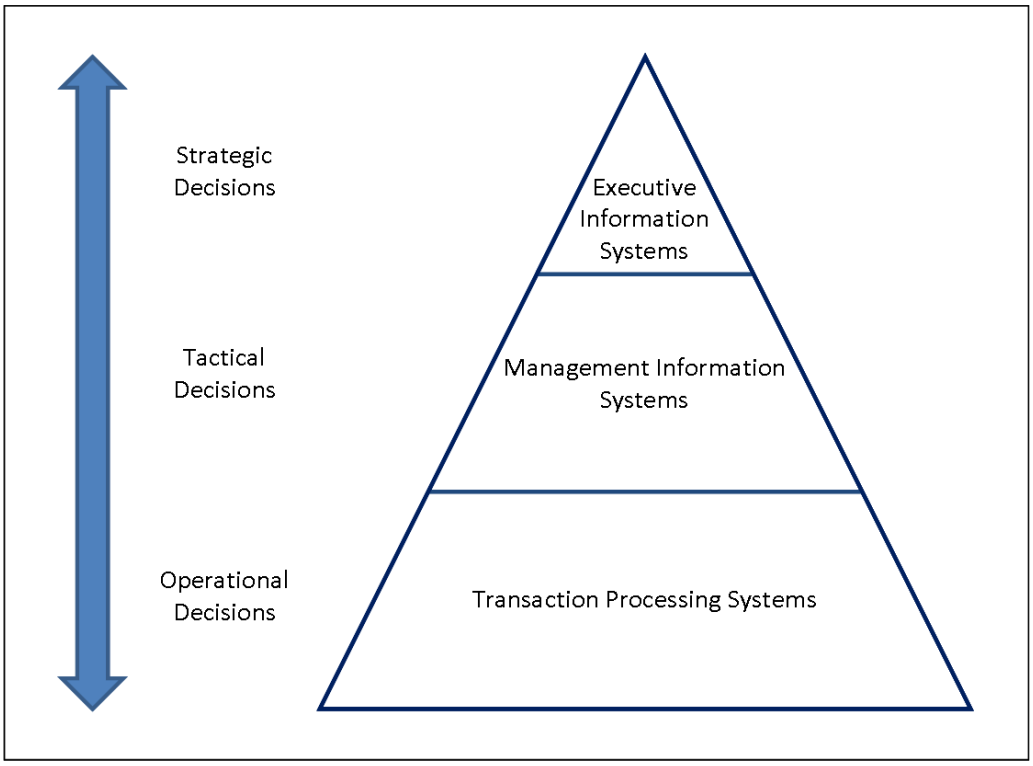
2.2.1 Λήψη Αποφάσεων και Συστήματα Επιχειρηματικής Ευφυίας

Σε μια επιχείρηση, οι αποφάσεις της είναι ζωτικής σημασίας και καθοριστικές για την πορεία και την επιτυχία της. Αυτές οι ακριβείς και καίριες αποφάσεις μπορούν να βοηθήσουν την επιχείρηση να ξεπεράσει τυχόν προβλήματα που μπορεί να προκύψουν, και να να ξεπεράσει τον ανταγωνισμό. Γενικά, οι αποφάσεις που λαμβάνουν οι ομάδες χωρίζονται σε τρία επίπεδά [1], [2] (**ΣΧΗΜΑ 2.1**): Στρατηγικές Αποφάσεις, Αποφάσεις Τακτικής και Λειτουργικές Αποφάσεις.

Στρατηγικές Αποφάσεις: Σε αυτό το επίπεδό τα στελέχη της επιχείρησης είναι υπεύθυνα για τις αποφάσεις και ενδιαφέρονται κυρίως να ελέγξουν της μεταβολές μεταξύ διαφόρων αποδόσεων των μεταβλητών ή των δεικτών σε σχετικά μεγάλα χρονικά διαστήματα (τετράμηνα, εξάμηνα κλπ.) και ανάλογα να αποφασίσουν. Δηλαδή, δεν τους ενδιαφέρει η αιτία της μεταβολής αλλά μόνο το αποτέλεσμα που έχει στην επιχείρηση. Χρησιμοποιούν συνοπτικές εκθέσεις από το προηγούμενο ιεραρχικά επίπεδο όπου περιέχουν τα δεδομένα αφαιρετικά και με ανάλυση της κλίσης αυτών των διαστάσεων, καταλήγουν σε αποφάσεις για την γενική και λογιστική λειτουργία της επιχείρησης.

Αποφάσεις Τακτικής: Αυτό ο τύπος απόφασης αποτελεί το δεύτερο επίπεδο στην πυραμίδα τύπου αποφάσεων. Εδώ τα στελέχη που το απαρτίζουν είναι συνήθως διευθυντικά στελέχη και έχουν την υποχρέωση να αναλύουν δεδομένα τμημάτων και να παίρνουν αποφάσεις για την καλύτερη και αποδοτικότερη λειτουργία τους. Όπως και στο επίπεδο των στρατηγικών αποφάσεων, έτσι και σε αυτό, τα δεδομένα που παίρνουν από τα υφιστάμενα τμήματα είναι συνοπτικά και συγκεντρωτικά αλλά με μεγαλύτερη πληθώρα πληροφοριών, που θα τους βοηθήσουν να εντοπίσουν τα αίτια αναποτελεσματικότητας των τμημάτων. Οι αποφάσεις αυτές συνήθως παίρνονται ανά μήνα και είναι ιδιαίτερα σημαντικό να έχουν σαν στόχο την γρήγορη προσαρμοστικότητα για την αποδοτικότερη και έγκαιρη λειτουργία της επιχείρησης.

Λειτουργικές Αποφάσεις: Τέλος, στο επίπεδο αυτό, τα στελέχη παίρνουν αποφάσεις που βασίζονται σε όλη τη διαθέσιμη λεπτομέρεια που μπορούν να προσφέρουν τα δεδομένα. Συνήθως, τα δεδομένα αυτά είναι μεγάλα σε μέγεθος και οι αποφάσεις που πρέπει να παρθούν αφορούν διαστήματα μικρής διάρκειας, από μία έως εφτά ημέρες. Οι αναλυτές αυτού του επιπέδου είναι υπεύθυνοι να παίρνουν αποφάσεις για τα υποτμήματα που διευθύνουν και το κύριο χαρακτηριστικό τους είναι η ταχύτητα με την οποία παίρνουν τις αποφάσεις τους.

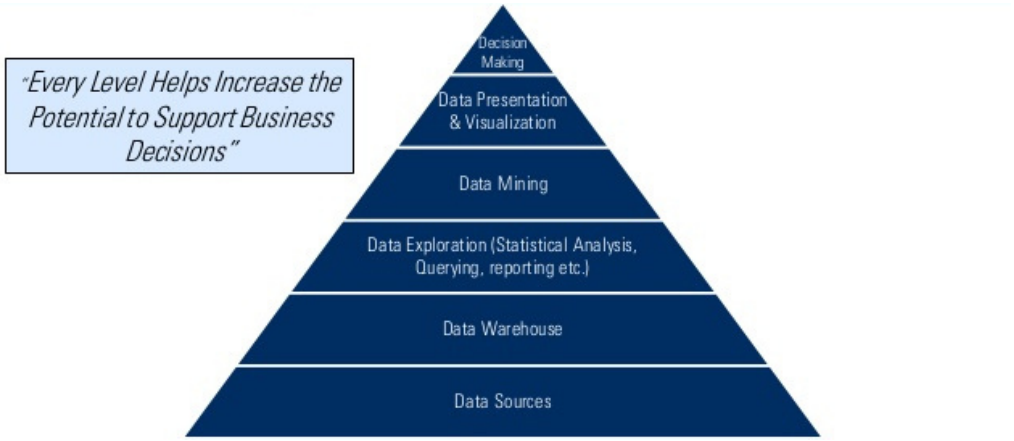


Σχημα 2.1: *Πυραμίδα βασισμένη στον τύπο απόφασης που παίρνει κάθε ομάδα διαχείρισης πληροφοριακών συστημάτων της επιχείρησης.*

Γενικά, η αρχή είναι ότι τα στελέχη που είναι υπεύθυνα να παίρνουν αποφάσεις και βρίσκονται ψηλότερα στην πυραμίδα, ασχολούνται περισσότερο με την συνολική εικόνα της επιχείρησης και θέτουν μακροπρόθεσμους στόχους. Όσο κατεβαίνουμε πιο κάτω τα στελέχη παίρνουν περισσότερο λεπτομερείς και συγκεκριμένες αποφάσεις για το τμήμα ή τα τμήματα που διαχειρίζονται. Όλα όμως τα επίπεδα έχουν την ιδιότητα να θέτουν στόχους, να αξιολογούν το ποσοστό επιτυχίας τους και ανάλογα να αναπροσαρμόζονται. Τα στελέχη που είναι υπεύθυνα να πάρουν αυτές της αποφάσεις πρέπει να διαθέτουν κάποια ισχυρά εργαλεία για να οπτικοποιούν, να αναλύουν και να μοντελοποιούν τα δεδομένα σε χρηστική πληροφορία, ικανή να τους οδηγήσει σε σωστές αποφάσεις. Η επιχειρηματική ευφυΐα είναι το πιο ολοκληρωμένο και σύγχρονο εργαλείο όπου προσφέρει πολυάριθμες δυνατότητες.

2.2.2 Επίπεδα Λειτουργίας Συστήματος Επιχειρηματικής Ευφυίας

Τα συστήματα επιχειρηματικής ευφυίας αποτελούν ολοκληρωμένα συστήματα μεγάλης κλίμακας, και περιέχουν αρκετά επίπεδα εργαλείων και μεθοδολογιών. Ωστόσο, όλα τα συστήματα στηρίζονται σε κάποια βασικά επίπεδα λειτουργίας όπως φαίνεται στο **ΣΧΗΜΑ 2.2**. Αναλυτικότερα τα επίπεδα είναι: [2]



Σχημα 2.2: *Τα διάφορα στάδια της διαδικασία ενός συστήματος επιχειρηματικής ευφυίας.*

Πηγές Δεδομένων (Data Sources): Το επίπεδο αυτό αποτελεί τη βάση των συστημάτων επιχειρηματικής ευφυίας. Αφορά τη συλλογή οποιασδήποτε ακατέργαστης μορφής δεδομένων και την μετατροπή τους σε δεδομένα με νόημα, έτοιμα για περαιτέρω ανάλυση. Οι αναλυτές σε αυτό το στάδιο μπορούν να χρησιμοποιήσουν διάφορα εργαλεία όπως γραφήματα και διαγράμματα για να αποκτήσουν μια πρώτη εικόνα της χρηστικότητας των δεδομένων και ποιες ερωτήσεις μπορούν να απαντήσουν με αυτά. Η χρήση αυτής της στρατηγική, επιτρέπει τις επιχειρήσεις να πάρουν αποτελεσματικότερες και πιο πραγματολογικές αποφάσεις, παρέχοντας έτσι μια πιο ολοκληρωμένη εικόνα των αναγκών τόσο της ίδιας της επιχείρησης όσο και της αγοράς γενικότερα.

Αποθήκες Δεδομένων (Data Warehouses): Οι αποθήκες δεδομένων επιτρέπουν στα στελέχη της επιχείρησης, να εξετάσουν τα δεδομένα λεπτομερώς μέσα από υποσύνολα και αλληλεξαρτήσεις. Οι αποθήκες αυτές μπορούν να χρησιμοποιηθούν για εξαγωγή χρήσιμων στατιστικών στοιχείων της λειτουργίας της επιχείρισης και πως αυτά συνδέονται το ένα με το άλλο όπου συνήθως χρησιμοποιούν μεθόδους πολυδιάστατης ανάλυσης. Για παράδειγμα, οι ιδιοκτήτες μια επιχείρησης μπορούν να συγκρίνουν χρόνους αποστολής σε σχέση με τις εισπράξεις σε διάφορα τμήματα πωλήσεων και να ελέγξουν, ποια διαδικασία είναι πιο αποδοτική. Οι αποθήκες δεδομένων αφορούν την αποθήκευση τεράστιου όγκου δεδομένων[1], με τέτοιο τρόπο ώστε να είναι ταυτόχρονα αποδοτικά και χρηστικά. Αφορούν κάθε είδους δεδομένα, από αδόμητα έως πλήρως δομημένα.

Διερεύνηση Δεδομένων (Data Exploration): Η Διερεύνηση Δεδομένων πρόκειται για την διαδικασία ανάλυσης και εξαγωγής συμπερασμάτων από τα δεδομένα της επιχείρησης. Συνήθως η διαδικασία αυτή

[1]Σήμερα, λόγο της τεράστιας έκρηξης των δεδομένων, η αναφορά σε μεγέθη όπως TB (Terabyte) και PB (Petabyte) ιδιαίτερα για μεσαίες και μεγάλες επιχειρήσεις είναι συχνό έως και καθημερινό φαινόμενο.

γίνεται με την εφαρμογή συμπερασματικής στατιστικής ανάλυσης και ερωτήσεων στα δεδομένα. Στόχος αυτού του επιπέδου είναι να δώσει στα στελέχη της επιχείρησης τη δυνατότητα να κατανοήσουν συγκεντρωτικά όλη τη πληροφορία που περιέχουν τα δεδομένα δίνοντας την κατάλληλη ερμηνεία στα συμπεράσματα που προκύπτουν. Επιτυγχάνεται κάνοντας διάφορους ελέγχους υπόθεσης και οπτικοποιώντας τα αποτελέσματα σε γραφήματα κατάλληλα για την εξαγωγή συμπεράσματος.

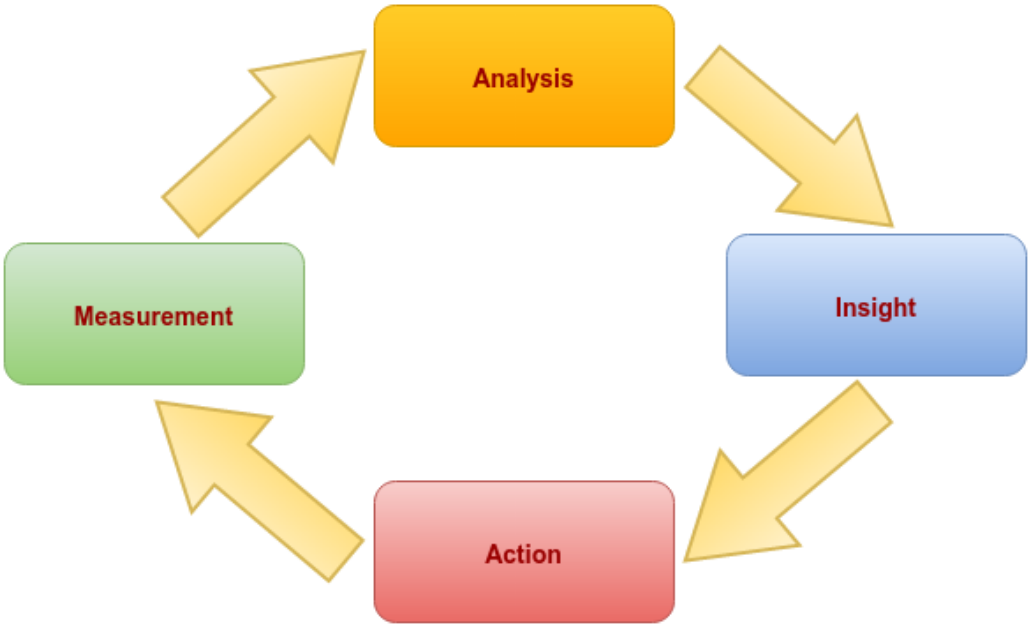
Εξόρυξη Δεδομένων (Data Mining): Στο επίπεδο της Εξόρυξης Δεδομένων, οι αναλυτές μπορούν να κατανοήσουν την φιλοσοφία των δεδομένων τους και να ανακαλύψουν πρότυπα που θα τους βοηθήσουν να προβλέψουν μελλοντικές συνήθειες πελατών. Είναι ο πυρήνας κάθε σύγχρονου συστήματος επιχειρηματικής ευφυίας. Στη πράξη αυτό που συμβαίνει είναι ο αναλυτής να εκπαιδεύει κάποια μοντέλα πρόβλεψης βασισμένα σε κάποιο αντιπροσωπευτικό σύνολο των δεδομένων και να εφαρμόζει άγνωστα δεδομένα πάνω σε αυτό. Βασίζεται κυρίως σε μεθόδους μηχανικής μάθησης και ουσιαστικά διευρύνει τα όρια της διερεύνησης δεδομένων.

Παρουσίαση Δεδομένων (Data Presentation): Η Παρουσίαση Δεδομένων αποτελεί το τελευταίο επίπεδο ενός συστήματος επιχειρηματικής ευφυίας πριν την διαδικασία λήψης απόφασης. Είναι υπεύθυνο να παρουσιάσει τα αποτελέσματα όλων των προηγουμένων επιπέδων και κυρίως αυτό επιτυγχάνεται με τη δημιουργία αποδοτικών αναφορών και εκθέσεων. Επίσης μπορεί περιέχει και συγκεντρωτικούς πίνακες (Dashboards) που θα αναφέρουν κάποια στατιστικά στοιχεία για τα δεδομένα αυτά και θα παρουσιάζουν εξαρτήσεις και ομάδες που μπορεί να προκύψουν. Τέλος στηρίζεται ιδιαίτερα στη χρήση γραφημάτων και οπτικοποιήσεων για να παρουσιάσει συνοπτικά, αρκετές πληροφορίες των δεδομένων με στόχο να είναι άμεσα κατανοητές από τα στελέχη της επιχειρήσεις.

Όσο ανεβαίνουμε τη πυραμίδα της διαδικασία ενός συστήματος επιχειρηματικής διαδικασίας, οδηγούμαστε στην μετατροπή των δεδομένων σε πληροφορία και από πληροφορία σε γνώση ικανή να υποστηρίξει τις αποφάσεις των στελεχών της επιχείρησης. Για τη διαδικασία αυτή το σύστημα το σύστημα απαιτεί σε κάθε επίπεδο στελέχη, εμπειρογνώμονα στον τομέα τους όπως διαχειριστές βάσεων δεδομένων, στατιστικολόγους, επιστήμονες πληροφορικής, ανθρώπους υπεύθυνους να κατανοήσουν τα αποτελέσματα της διαδικασίας και να πάρουν αποφάσεις κ.ά.

2.2.3 Ο Κύκλος της Επιχειρηματικής Ευφυίας

Ένα σύστημα επιχειρηματικής ευφυίας δεν είναι μόνο ένα σύνολο εργαλείων για την ανάλυση δεδομένων και την υποστήριξη αποφάσεων και στρατηγικών. Πρόκειται επίσης για ένα πλαίσιο, που προσφέρει και δίνει τη δυνατότητα στα στελέχη της επιχείρισης να κατανοήσουν τι να αναζητούν σε μεγάλους όγκους δεδομένων της επιχείρησης. Παρόλο που κάθε επιχείρηση έχει το δικό της σύστημα επιχειρηματικής ευφυίας προσαρμοσμένο για αυτήν και ακολουθεί τη



Σχημα 2.3: *Ο κύκλος της επιχειρηματικής ευφυίας.*

δική του φιλοσοφία, είναι δυνατόν να καθοριστεί μία κοινή βάση για το πως αυτά κινούνται. Ουσιαστικά πρόκειται για ένα συνεχή κύκλο ανάλυσης, επίγνωσης, δράσης και αποτίμησης [2] (βλ. **Σχήμα 2.3**). Πιο συγκεκριμένα:

Ανάλυση (Analysis): Το στάδιο της Ανάλυσης, βασίζεται ουσιαστικά στη γνώση που έχει η επιχείρηση για τι είναι σημαντικό για εκείνη, ενώ φιλτράροντας παράλληλα τις πτυχές της επιχείρησης που δεν θεωρούνται κρίσιμες ή επιβλαβείς για την ανάπτυξη της. Η απόφαση τι είναι σημαντικό για εκείνη βασίζεται στην κατανόηση και σε υποθέσεις που κάνει για το τι είναι σημαντικό για τους πελάτες, τις προμήθειες, τους ανταγωνιστές και τους εργαζομένους της. Όλη αυτή η γνώση είναι μοναδική για μία επιχείρηση και αποτελεί κύριο πόρο για τη διαμόρφωση της επιχειρηματικής στρατηγικής που θα ακολουθήσει. Τα στελέχη της επιχείρησης που είναι υπεύθυνα να πάρουν κάποιες αποφάσεις, με αυτόν τον τρόπο μπορούν να κάνουν πολλές ερωτήσεις προς ένα σύστημα και να πάρουν γρήγορες διαδραστικές απαντήσεις.

Επίγνωση (Insight): Η Επίγνωση έρχεται σε πολλές μορφές. Υπάρχει για παράδειγμα η *Λειτουργικής Επίγνωση* και η *Στρατηγική Επίγνωση*. Πρόκειται για τη μετατροπή της πληροφορίας που αντλούμε από την ανάλυση, σε γνώση. Τα στελέχη μπορούν να καταφέρουν την εξαγωγή γνώσης που είναι είτε βασισμένη σε προηγούμενη εμπειρία είτε σε υποθέσεις. Το πρόβλημα που πρέπει να αντιμετωπίσει ο υπεύθυνος σχετικά με την επίγνωση, είναι να καταφέρει να πείσει τους υπόλοιπους ότι η γνώση που παρήγαγε είναι χρήσιμη, εφαρμόσιμη και ρεαλιστική. Όπως στη ζωή, οτιδήποτε νέο ή διαφορετικό θέλει το χρόνο του για να γίνει καθολικά αποδεχτό. Ένα καλά οργανωμένο σύστημα επιχειρηματικής ευφυίας που υποστηρίζει την επίγνωση, παρέχει σαφή και ξεκάθαρα τα

δεδομένα, τα πρότυπα, τη λογική, την παρουσίαση (π.χ. γραφικές παραστάσεις, εκθέσεις) και τους υπολογισμούς που γίνονται για να κατανοήσει πιο γρήγορα η επιχείρηση την νέα αντίληψη.

Δράση (Action): Όταν η ανάλυση έχει γίνει και η επίγνωση έχει εξαχθεί και έχει γίνει αποδεχτή, η επόμενη διαδικασία στον κύκλο επιχειρηματικής ευφυίας είναι να εκτελεστεί η Δράση ή η λήψη αποφάσεων. Οι καλά μελετημένες αποφάσεις υποστηρίζονται από την καλή ανάλυση και η επίγνωση δίνει αυτοπεποίθηση στην εφαρμογή της προτεινόμενης δράσης. Σε αντίθετη περίπτωση, οι αποφάσεις που δεν υποστηρίζονται από ποιότητα στην ανάλυση γίνονται κάτω από αυστηρά μέτρα ασφαλείας ή με μικρότερη αφοσίωση και δέσμευση από τους εμπλεκόμενους φορείς. Επιπλέον, η ποιότητα των αποτελεσμάτων της επιχειρηματικής ευφυΐας γρήγορα βελτιώνει την ταχύτητα της δράσης. Οι επιχειρήσεις σήμερα πρέπει να αντιδρούν με μεγαλύτερη ταχύτητα, να αναπτύσσουν νέες προσεγγίσεις γρηγορότερα, να είναι πιο ευέλικτες στη διεξαγωγή έρευνας και ανάπτυξης και να παράγουν προϊόντα και υπηρεσίες στην αγορά πιο γρήγορα από τον ανταγωνισμό. Η λήψη αποφάσεων βασισμένη σε συστήματα επιχειρηματικής ευφυίας, γίνεται με ταχύτερη πρόσβαση στις πληροφορίες και την ανατροφοδότηση που παίρνουμε από αυτά και παρέχει περισσότερες δυνατότητες για ταχύτερη κατασκευή πρωτοτύπων και δοκιμών.

Αποτίμηση (Measurement): Το όφελος των αναφορών που προκύπτουν από τα συστήματα επιχειρηματικής ευφυίας, συχνά δημιουργούν την αναγκαιότητα να τεκμηριωθούν τα αποτελέσματα αυτών σε σχέση με κάποια ποσοτικά πρότυπα. Η επιχειρηματική ευφυΐα επιτρέπει τη θέσπιση κάποιων προτύπων και σημείων αναφοράς για την παρακολούθηση της απόδοσης και της παροχής ανατροφοδοτήσεων σε κάθε τμήμα της επιχείρησης, χρησιμοποιώντας μετρήσεις που επεκτείνονται πολύ πιο πέρα από τις παραδοσιακές οικονομικές μετρήσεις.

Ένας καλά καθορισμένος κύκλος επιχειρηματικής ευφυίας, βοηθά τις επιχειρήσεις να θέτουν στόχους, να αναλύουν την πρόοδο τους, να αποκτούν γνώσεις, να αναλαμβάνουν δράσεις και να αξιολογούν τα αποτελέσματα τους. Η επιχειρηματική ευφυΐα αποτελεί ένα πλαίσιο διαχείρισης της απόδοσης της επιχείρησης και θα πρέπει συνεχώς να εξελίσσεται καθώς η οργάνωση ωριμάζει και προσπαθεί να αποκτήσει ανταγωνιστικό πλεονέκτημα.

2.2.4 Σχεδιασμός και Ανάλυση Συστημάτων Επιχειρηματικής Ευφυίας

Ένα σύστημα επιχειρηματικής ευφυίας, αφορά τη διαδικασία λήψης αποφάσεων μέσα από διάφορα στάδια επεξεργασίας των διαθέσιμων δεδομένων της επιχείρησης. Αρχικά, τα δεδομένα αυτά συλλέγονται με την βοήθεια των αποθηκών δεδομένων, μετατρέπονται σε χρηστική πληροφορία μέσω ad-hoc ερωτημάτων και της άμεσης επεξεργασίας δεδομένων, εξάγεται γνώση μέσα από αυτές τις πληροφορίες μέσω της εξόρυξης δεδομένων και της παρουσίασης

των αποτελεσμάτων και με την γνώση αυτή τα στελέχη της επιχείρησης είναι σε θέση να πάρουν σημαντικές και ανταγωνιστικές αποφάσεις.

Ο χρόνος και η προσπάθεια που απαιτείται για να σχεδιαστεί ένα σύστημα επιχειρηματικής ευφυΐας διαφέρει από επιχείρηση σε επιχείρηση. Αυτή η προσπάθεια εξαρτάται κυρίως από την πολυπλοκότητα του συστήματος και το επίπεδο της λειτουργίας του. Ωστόσο, στην πλειονότητα των περιπτώσεων, η δημιουργία ενός προσαρμοσμένου συστήματος επιχειρηματικής ευφυίας απαιτεί πολύ χρόνο. Ο χρόνος που απαιτείται, δαπανάται όχι μόνο για το σχεδιασμό των επιμέρους επιπέδων, αλλά και για να διασφαλίσουμε ότι το σύνολο του συστήματος είναι λογικό και συνεπές. Ένα άλλο σημαντικό στάδιο του σχεδιασμού ενός συστήματος επιχειρηματικής ευφυίας περιλαμβάνει την κατασκευή μιας αποθήκης δεδομένων που θα εκτελεί δύο λειτουργίες: ενός αποθετηρίου για περαιτέρω αναλύσεις, και τη βάση για το σύστημα επιχειρηματικής ευφυίας. Η διαδικασία γενικά πρέπει να ακολουθεί κάποιους κανόνες όπως: [2]–[4]

- Τον καθορισμό ενός πλαισίου για το ποια από τα δεδομένα της επιχείρησης που είναι αποθηκευμένα στην αποθήκη δεδομένων είναι σημαντικά για εκείνη.
- Τον καθορισμό των διασυνδέσεων μεταξύ των δεδομένων που μπορούν να βρεθούν σε διάφορα συστήματα και έχουν την ίδια αξία. Ως αποτέλεσμα αυτών των ενεργειών, θα πρέπει να δημιουργηθεί ένα σύνολο δεδομένων όπου θα επιτρέψουν το σχεδιασμό μιας βάσης δεδομένων που τα δεδομένα από την πηγή θα στέλνονται.
- Η δημιουργία ενός σχεδίου αποθήκης δεδομένων, η οποία εξυπηρετεί ως βάση για τη λειτουργία ενός συστήματος επιχειρηματικής ευφυίας. Ένα τέτοιο σχέδιο πρέπει να δημιουργηθεί για να παρέχει εύκολη διαμόρφωση της βάσης δεδομένων σχετικά με εύκολες αναφορές και ερωτήσεων προς αυτή. Το σχέδιο που γενικά προτείνεται είναι ενός σχήματος τύπου «αστέρα» ή «νιφάδας» που απλοποιεί περαιτέρω την εφαρμογή, στην αποθήκη δεδομένων, με τη χρήση τεχνικών όπως OLAP και εξόρυξης δεδομένων.

Μια αποθήκη δεδομένων πρέπει να ενημερώνεται συστηματικά για να συμπεριλαμβάνει και στοιχεία που έρχονται σε τακτά χρονικά διαστήματα, όπως δεδομένα συναλλαγών. Οι μηχανισμοί που θα χρησιμοποιηθούν είναι απαραίτητο να επιτρέπουν στα στελέχη της επιχείρησης την εισαγωγή όλων των δεδομένων αλλά και να εκτελούν σταδιακές εισαγωγές, οι οποίες απαιτούν την επεξεργασία των δεδομένων που φτάνουν εκείνη τη στιγμή. Οι σταδιακές εισαγωγές, δεν θα πρέπει να επιβαρύνουν το σύστημα και των μηχανισμό επεξεργασίας των δεδομένων. Οι μηχανισμοί μεταφοράς δεδομένων ταυτόχρονα πρέπει να εκτελούν μια λειτουργία ελέγχου που έχει την ευθύνη της συνοχής των δεδομένων. Σε πολλές περιπτώσεις, οι μηχανισμοί αυτοί επιτρέπουν την εύρεση ασυνεπειών και σφαλμάτων στο στάδιο της υλοποίησης. Οι διαδικασίες εισαγωγής δημιουργήθηκαν ώστε να μπορούν να καταγράφουν εσφαλμένα δεδομένα που θα πρέπει να μεταφερθούν στην αποθήκη δεδομένων όπου εκεί

δύνεται η δυνατότητα διόρθωσης τους.

Η αποθήκη δεδομένων που δημιουργήθηκε, θα αποτελέσει τη βάση για την κατασκευή των επιπέδων, σχετικά με τις αναφορές και τις εκθέσεις. Είναι αναγκαίο να παρέχονται τουλάχιστον δύο ομάδες αναφορών. Θα πρέπει να δημιουργηθούν προκαθορισμένες εκθέσεις που να ενημερώνονται συστηματικά από τα στελέχη της επιχείρησης. Επίσης θα πρέπει να επιτρέπει στους χρήστες τη δυνατότητα να καθορίσουν οι ίδιοι δικές του αναφορές σύμφωνα με τις ανάγκες τους.

Η ιστορία έχει δείξει ότι - αργά ή γρήγορα - οι επιχειρήσεις θα πρέπει να χρησιμοποιήσουν πολυδιάστατες και πολυμεταβλητές αναλύσεις, γεγονός που σημαίνει ότι ένα σύστημα επιχειρηματικής ευφυίας πρέπει να στοχεύει σε ενέργειες OLAP και εξόρυξης δεδομένων. Ανάλογα με τις ιδιαιτερότητες και τις ανάγκες των επιχειρήσεων, είναι δυνατόν η εξέταση του ενδεχόμενου διεξαγωγής αναλύσεων που είναι υπεύθυνες για την μεταξύ άλλων:

- Εξέταση του κέρδους και της αξία των πελατών,
- Επαλήθευση τις σημασίας των διαφόρων παραμέτρων,
- Κατάτμηση και ομαδοποίηση των πελατών,
- Παρακολούθηση της αφοσίωσης των πελατών,
- Έλεγχο διατήρησης των πελατών,
- Εύρεση ομοιοτήτων,
- Μελέτη για απάτες,
- Μελέτη για εκστρατείες marketing,
- Διενέργεια μεθόδων Cross-selling και Up-selling.

Τέλος, ένα επίπεδο παρουσίασης των δεδομένων είναι ιδιαίτερα σημαντικό στοιχείο ενός συστήματος επιχειρηματικής ευφυίας. Αναφορικά με το προηγούμενο, είναι απαραίτητο να δοθεί ιδιαίτερη προσοχή στο σχεδιασμό της διεπαφής.

2.3 Εμπορικά Συστήματα Επιχειρηματικής Ευφυίας

Σήμερα οι επιχειρήσεις, μπορούν να βρουν χιλιάδες λύσεις επιχειρηματικής ευφυίας σε διάφορα εργαλεία υποστηριζόμενα από μεγάλες και αξιόπιστες εταιρίες. Μερικές αξιόλογες από αυτές μπορούν να θεωρηθούν οι εξής:

Oracle BI:

ORACLE
BUSINESS INTELLIGENCE

Το σύστημα επιχειρηματικής ευφυίας της Oracle αποτελείται από μία συμπαγής συλλογή λειτουργιών, που καλύπτουν μια μεγάλη γκάμα από δυνατότητες επιχειρηματικής ευφυίας, συμπεριλαμβανομένου συγκεντρωτικούς πινάκες, πλήρη ad-hoc ερωτήματα, προληπτικές προειδοποιήσεις και πολλά άλλα. Τυπικά, οι οργανισμοί μπορούν να αποθηκεύουν και να παρακολουθούν μεγάλο όγκο δεδομένων σχετικά με προϊόντα, πελάτες, συναλλαγές, επαφές, υπαλλήλους και πολλές ακόμα κατηγορίες. Τα δεδομένα αυτά συνήθως διαδίδονται μεταξύ πολλών διαφορετικών βάσεων δεδομένων σε διαφορετικές τοποθεσίες και σε διαφορετικές εκδόσεις αυτών.

IBM Congos:

IBM COGNOS

Η πληροφορία είναι σκορπισμένη σε διαφορετικά σιλό δεδομένων, πολλαπλές πλατφόρμες και η υπερβολική εξάρτηση από λογιστικές αναφορές μπορούν να παρεμποδίζουν την διαδικασία της ανάλυσης των δεδομένων της επιχείρησης. Με το σύστημα επιχειρηματικής ευφυίας της IBM, Congos, οι επιχειρήσεις μπορούν να αναλύσουν και να παρακολουθήσουν τα δεδομένα από διάφορες οπτικές γωνίες και σκοπιές και να τα συγκρίνει με δεδομένα πραγματικού χρόνου και τάσεων για πιο εκτεταμένη ανάλυση των αναγκών της επιχείρησης.

SAP:

SAP

Οι προγνωστικές αναλύσεις, επιτρέπουν στα στελέχη της επιχείρησης να εξάγουν γνώση ικανή να τους βοηθήσει να προβλέψουν νέες εξελίξεις, να εκμεταλλευτούν μελλοντικές τάσεις της αγοράς και να ανταποκριθούν στις προκλήσεις, πριν αυτές συμβούν. Το σύστημα SAP, αποτελεί ένα

συνδυασμό επιχειρηματικής ευφυίας πραγματικού χρόνου και προβλεπτικής ανάλυσης επιτρέποντας στις επιχειρήσεις να εξάγουν μελλοντικές αποφάσεις από μεγάλα σύνολα δεδομένων (Big Data), εκμεταλλεύοντας την δύναμη της γλώσσας R και δημιουργώντας οπτικοποιήσεις των δεδομένων με μια μεγάλη γκάμα εργαλείων.

SAS:

Ένα από τα πιο σημαντικά συστήματα επιχειρηματικής ευφυίας, το SAS, αποτελεί μία ολοκληρωμένη πλατφόρμα που προτιμάτε από πάρα πολλές επιχειρήσεις, λόγο του μεγάλου αριθμού από δυνατότητες που προσφέρει. Καταφέρνει να συνδυάζει δυνατότητες όπως επεκτασιμότητα, πολύ καλή ενσωμάτωση δεδομένων, υποστήριξη πολλαπλών γλωσσών ερωτημάτων, APIs με αρκετές παραμετροποιήσεις, προηγμένα εργαλεία ανάλυσης και αναφορών κ.ά. Το σύστημα SAS κατέχει ένα μεγάλο ποσοστό της αγοράς συστημάτων επιχειρηματικής ευφυίας.

Tableau:

Η οικογένεια υπηρεσιών Tableau παρέχει στις επιχειρήσεις εργαλεία διαισθητικής επιχειρηματικής ευφυΐας για να βοηθήσει και να βελτιώσει την ανακάλυψη από τα δεδομένα και την κατανόηση τους για όλους τους τύπους των επιχειρήσεων και από όλα τα στελέχη της. Με απλές λειτουργίες drag-and-drop, οι χρήστες είναι σε θέση να έχουν εύκολη πρόσβαση και να αναλύουν βασικά δεδομένα, να δημιουργούν καινοτόμες εκθέσεις και απεικονίσεις και να μοιράζονται κρίσιμες ιδέες σε ολόκληρη την επιχείρηση.

Teradata:

Η πλατφόρμα επιχειρηματικής ευφυίας Teradata με τη χρήση σύγχρονων αποθηκών δεδομένων ικανών για υποστήριξη Big Data, αυξάνει την ταχύτητα των επιχειρηματικών διαδικασιών, δίνοντας στα στελέχη την δυνατότητα να παρακολουθήσουν την συνολική εικόνα της επιχείρησης

και να κατανοήσουν βαθύτερα τις ανάγκες των πελατών τους και της αγοράς.

Cloudera:

Η Cloudera είναι μια υπηρεσία βασισμένη στο Hadoop αλλά με υποστήριξη και πιο μοντέρνων τεχνικών όπως το Spark. Διαθέτει στις επιχειρήσεις ένα σύστημα κατανεμημένης αποθήκευσης των δεδομένων (Data Hub), το οποίο αυξάνει την ταχύτητα για αποθήκευση δεδομένων και εκτέλεση μεγάλου αριθμού φόρτου εργασιών, που κυμαίνονται από μαζική επεξεργασία και επιχειρησιακή έρευνα μέχρι διαδραστικά ερωτήματα SQL και προηγμένες αναλύσεις.

Eclipse BIRT Project:

Το πρότζεκτ BIRT είναι ένα ευέλικτο, ανοιχτού κώδικα και αμιγώς φτιαγμένο σε Java εργαλείο, για τη δημιουργία και την εξαγωγή αναφορών από πηγές δεδομένων που κυμαίνονται από τυπικές σχεσιακές βάσεις δεδομένων, σε πηγές δεδομένων XML αλλά και σε αντικείμενα Java στη μνήμη. Το BIRT αναπτύχθηκε ως top-level πρότζεκτ του Eclipse Foundation και αξιοποιεί τις πλούσιες δυνατότητες που προσφέρει η πλατφόρμα Eclipse όπως και την πολύ ενεργή ανοιχτή κοινότητα των χρηστών που την απαρτίζουν. Χρησιμοποιώντας το BIRT, οι προγραμματιστές όλων των επιπέδων μπορούν να ενσωματώσουν ισχυρές αναφορές σε εφαρμογές Java, J2EE και έργων βασισμένων σε Eclipse περιβάλλον.

MicroStrategy:

Από τοπικά ήμι-δομημένα δεδομένα μέχρι εταιρικά συστήματα δεδομένων και δεδομένα βασισμένα στο cloud, το MicroStrategy προσφέρει εύκολη πρόσβαση σε όλα τα δεδομένα από ένα σημείο. Χρησιμοποιεί συνδέσεις μεταξύ δεδομένων, βελτιστοποιημένες για κάθε πηγή και επιτρέποντας τα ερωτήματα να φτάνουν στην καλύτερη δυνατή απόδοση. Συνδέεται με μία ή πολλές πηγές, χωριστά ή σε συνδυασμό. Στοχεύει σε υψηλής ταχύτητας αναλύσεις και σε συνδυασμό με την υπηρεσία MapR μπορεί

εύκολα να κλιμακωθεί σε μεγάλα σύνολα δεδομένων με την βοήθεια εργαλείων όπως το Spark.

Αξίζει να σημειωθεί εδώ ότι πολλές επιχειρήσεις σήμερα, ιδιαίτερα μεσαίου προς μεγάλου μεγέθους, στρέφονται σε λύσεις επιχειρηματικής ευφυίας προσαρμοσμένες για τις δικές τους ανάγκες και σε εργαλεία που έχουν δημιουργηθεί για αυτές. Η μέθοδος αυτή μεταξύ άλλων, προσφέρει στις επιχειρήσεις τα εξής πλεονεκτήματα:

- Μικρό χρόνο εκμάθησης του συστήματος διότι η ίδια η επιχείρηση ενσωματώνει την κουλτούρα της μέσα σε αυτό.
- Το σύστημα προσφέρει όλες της δυνατότητες που χρειάζεται η επιχείρηση χωρίς συμβιβασμό.
- Τα εργαλεία που θα χρησιμοποιεί η επιχείρηση θα είναι μοναδικά και ανταγωνιστικά σε σχέση με αυτά της αγοράς.
- Η εγκατάσταση και η ρύθμιση του συστήματος μπορεί να γίνει σταδιακά και προσαρμοστικά.
- Συχνά τα συστήματα αυτά είναι πολύ περισσότερο προσιτά από ήδη έτοιμες λύσεις.

2.4 Μεθοδολογίες Επιχειρηματικής Ευφυίας

Η επιχειρηματική ευφυΐα αποτελεί ένα μεγάλο σύνολο από πολλές και διάφορες μεθόδους, που η κάθε μία είναι υπεύθυνη να δώσει στον αναλυτή τη δυνατότητα να εξερευνήσει και να κατανοήσει τα δεδομένα της επιχείρησης από πολλές πλευρές. Αντλεί αυτές τις δυνατότητες από αρκετές επιστήμες όπως μαθηματικά, στατιστική, πληροφορική, οικονομικά και ειδικότητες αυτών.

2.4.1 Διερευνητική Ανάλυση Δεδομένων

Η Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis ή EDA) είναι κρίσιμο πρώτο βήμα στην ανάλυση δεδομένων. Ουσιαστικά αποτελεί μία μέθοδο σύνοψης των κύριων χαρακτηριστικών από τα δεδομένα συνήθως με οπτικές μεθόδους. Κυρίως η EDA δίνει στον αναλυτή την δυνατότητα να κατανοήσει πριν από μοντέλα πρόβλεψης και ελέγχους υποθέσεων, ποια πληροφορία μπορεί να του δώσουν τα δεδομένα και τι απαντήσεις μπορεί να πάρει από αυτά. Ο πατέρας της μεθόδου αυτής είναι ο John Tukey[2]. Μερικοί από τους κύριους λόγους όπου η EDA θεωρείται ιδιαίτερα σημαντική είναι:

[2]Ο John Wilder Tukey ήταν Αμερικάνος μαθηματικός, γνωστός κυρίως για την συμβολή του στην ανάπτυξη του αλγορίθμου FFT και των Box Plots. Το όνομά του έχουν πάρει επίσης μερικές πολύ σημαντικές συνεισφορές του όπως η Tuckey λάμδα κατανομή, το τεστ Tuckey κ.ά.

- Η εύρεση σφαλμάτων.
- Ο έλεγχος υποθέσεων.
- Μια πρώτη επιλογή των κατάλληλων μοντέλων.
- Ο προσδιορισμός των σχέσεων μεταξύ των ανεξάρτητων μεταβλητών.
- Η διερεύνηση της κατεύθυνσης που κινούνται οι αλληλεξαρτήσεις μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών.

Γενικά μιλώντας, όποια μέθοδος χρησιμοποιείται για τη διερεύνηση των δεδομένων που δεν περιλαμβάνει τυπικά στατιστικά μοντέλα και μοντέλα πρόβλεψης θεωρείται μέθοδος της διερευνητικής ανάλυσης δεδομένων.

Τα Είδη Διερευνητικής Ανάλυσης Δεδομένων

Τα δεδομένα μιας επιχειρήσεις τις περισσότερες φορές είναι δομημένα σε βάσεις δεδομένων και έχουν την μορφή πολυδιάστατων πινάκων. Κάθε διάσταση περιέχει είτε κατηγορικές είτε διακριτές είτε συνεχείς τιμές και μπορεί να χαρακτηριστεί είτε σαν εξαρτημένη μεταβλητή όταν θέλουμε να προβλέψουμε αυτήν την διάσταση είτε σαν ανεξάρτητη μεταβλητή όταν η διάσταση αυτή βοηθά στην πρόβλεψη της εξαρτημένης μεταβλητής.

Ό άνθρωπος γενικά δεν μπορεί εύκολα να βγάλει συμπεράσματα κοιτώντας απλά και μόνο πίνακες με αριθμούς και να διαπιστώσει μοτίβα στα δεδομένα. Με την διερευνητική ανάλυση των δεδομένων όμως ο αναλυτής μπορεί να ξεπεράσει τέτοιες δυσκολίες. Οι περισσότερες από τις τεχνικές που χρησιμοποιεί η EDA λειτουργούν κρύβοντας συγκεκριμένες πτυχές των δεδομένων κάνοντας παράλληλα άλλες πτυχές πιο διακριτές.

Η διερευνητική ανάλυση δεδομένων γενικά ταξινομείται με δύο τρόπους. Ο ένας είναι κάθε μέθοδος που χρησιμοποιείται μπορεί να αναπαρασταθεί γραφικά ή όχι και ο δεύτερος είναι αν οι μέθοδοι αφορούν μία μόνο μεταβλητή ή περισσότερες. [5], [6]

Οι μη-γραφικές μέθοδοι κυρίως αφορούν υπολογισμούς συνοπτικών στατιστικών ενώ οι γραφικές μέθοδοι συνοψίζουν τα δεδομένα σε γραφικές παραστάσεις και γραφήματα. Οι μέθοδοι που αφορούν μία μεταβλητή, ελέγχουν μία μεταβλητή (διάσταση) τη φορά, ενώ οι πολυμεταβλητές μέθοδοι εξετάζουν δύο ή περισσότερες μεταβλητές ταυτόχρονα για να ανακαλύψουν σχέσεις μεταξύ τους. Τις περισσότερες φορές χρησιμοποιούνται δύο μεταβλητές τη φορά αλλά είναι φρόνιμο ο αναλυτής πριν εξετάσει τις ενδιαφερόμενες μεταβλητές μαζί να τις ελέγξει κάθε μία ξεχωριστά.

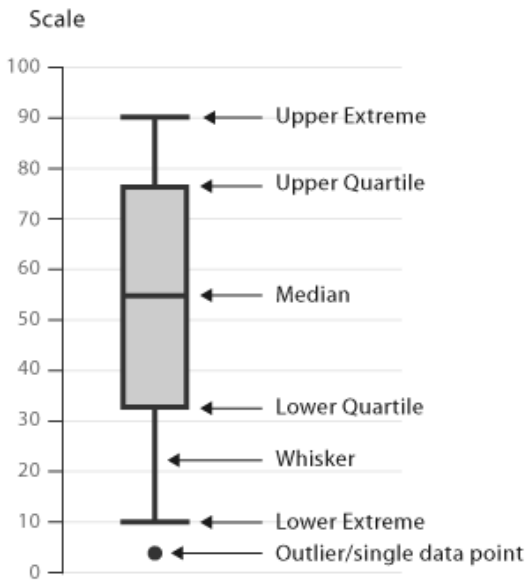
Εκτός από αυτές τις δύο κατατάξεις, κάθε μία κατηγορία μπορεί να διαιρεθεί περαιτέρω σε σχέση με το ρόλο (συμπερασματική ή επεξηγηματική) και το είδος (κατηγορική ή ποσοτική) των μεταβλητών που εξετάζονται.

Παρόλο που υπάρχει μια γενική μεθοδολογία ποιας τεχνικής διερευνητικής ανάλυσης δεδομένων είναι η πιο χρηστική σε κάθε περίπτωση, η επιλογή της καταλληλότερης συχνά είναι θέμα της εμπειρίας και του ταλέντου του αναλυτή. Η αυτοπεποίθηση και η ικανότητα επιλογής έρχεται με την εξάσκηση και την μελέτη της δουλειάς άλλων αναλυτών. Επίσης, οι τεχνικές που χρησιμοποιούνται δεν είναι ντε φάκτο και μερικές φορές μπορεί να χρειαστούν καινούργιοι τρόποι εξέτασης των δεδομένων.

Τεχνικές Διερευνητικής Ανάλυσης Δεδομένων

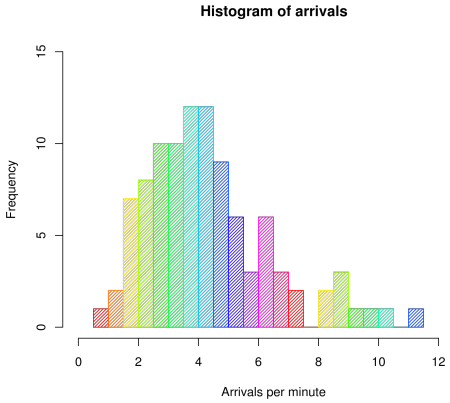
Μερικές από της πιο σημαντικές τεχνικές γραφικής απεικόνισης της EDA είναι οι εξής:

Box Plots:



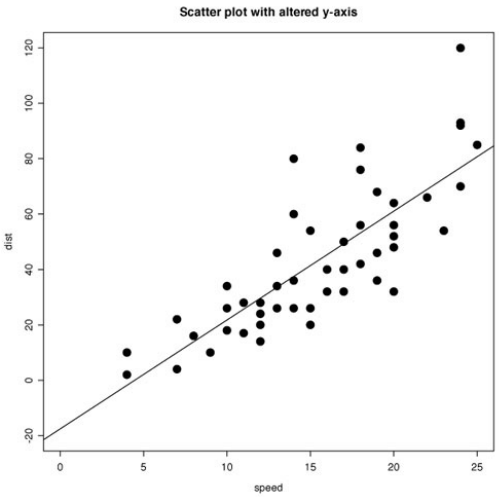
Ένα διάγραμμα τύπου Box Plot είναι ένας βολικός τρόπος αναπαράστασης ομάδων από αριθμητικά δεδομένα μέσω των τεταρτημορίων τους. Τα γραφήματα αυτά μπορούν επίσης να έχουν γραμμές που επεκτείνονται κάθετα από το "κουτί" δείχνοντας τη μεταβλητότητα του άνου και του κάτου τεταρτημορίου. Επίσης, οι ακραίες τιμές μπορούν να αναπαρίσταται σαν ξεχωριστά σημεία. Τα διαγράμματα αυτά είναι χρήσιμα διότι προσφέρουν συνοπτικά όλες σχεδόν τις πληροφορίες κάθε μεταβλητής. Σήμερα, κυκλοφορούν και διάφορες αξιόλογες παραλλαγές των Box Plots όπως τα Violin Plots και τα Bean Plots.

Histograms:



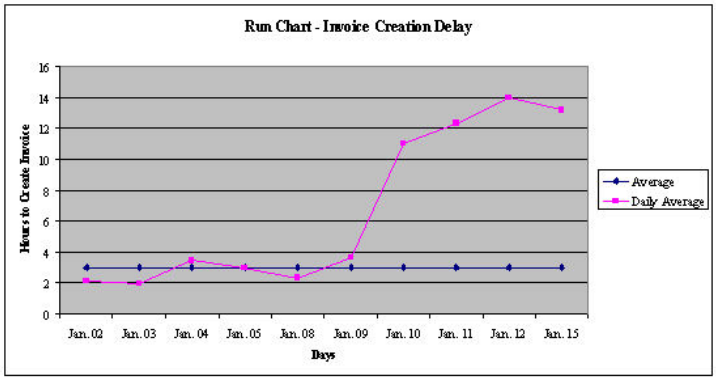
Ένα ιστόγραμμα είναι μια γραφική αναπαράσταση της κατανομής πιθανότητας συνεχών αριθμητικών μεταβλητών. Ουσιαστικά δείχνει τη συχνότητα εμφάνισης μεταξύ ενός εύρους τιμών. Είναι το πρώτο και συνηθέστερο διάγραμμα.

Scatter Plot:



Ένα Scatter Plot είναι ένας τύπος μαθηματικού διαγράμματος που χρησιμοποιεί το Καρτεσιανό σύστημα συντεταγμένων για να απεικονίσει τις τιμές δύο (ή περισσοτέρων μεταβλητών). Τα δεδομένα προβάλλονται σαν σημεία στους κάθετους άξονες με κάθε τιμή της μεταβλητής να καθορίζει τη θέση τους στο σχήμα. Η απεικόνιση περισσότερο από δύο μεταβλητών μπορεί να επιτευχθεί με τη χρήση διαφορετικού χρώματος και σχήματος στο κάθε σημείο.

Run Chart:



Το Run Chart είναι ένα τύπος διαγράμματος που απεικονίζει τα δεδομένα σε χρονολογική σειρά. Συχνά, τα διαγράμματα αυτά χρησιμοποιούνται για να απεικονίσουν κάποια πτυχή της διαδικασίας παραγωγής ή τις επιδόσεις ενός τμήματος στο χρόνο ή άλλες επιχειρηματικές δραστηριότητες.

Μερικές τεχνικές που αφορούν συνοπτικά στατιστικά συμπεράσματα της EDA είναι τα εξής:

Median Polish: Η τεχνική Median Polish βρίσκει ένα αθροιστικό προσαρμοσμένο μοντέλο στα δεδομένα σε έναν αμφίδρομο πίνακα με τιμές:

παρατηρούμενη τιμή = συνολικό μέσο όρο (ή συνολικό διάμεσο) + όρο γραμμής + όρο στήλης + υπόλοιπο

Ordination: Το Ordination είναι μία μέθοδος της πολυμεταβλητής ανάλυσης και ουσιαστικά συμπληρώνει την μέθοδο του Clustering. Η τεχνική αυτή οργανώνει τα δεδομένα σε αντικείμενα που χαρακτηρίζονται από τιμές σε πολλαπλές μεταβλητές, έτσι ώστε όμοια αντικείμενα να είναι κοντά το ένα με το άλλο και τα ανόμοια το αντίθετο. Μερικές από τις τεχνικές ordination είναι η Principal Components Analysis (PCA), η Correspondence Analysis (CA) κ.ά.

Trimean: Ο όρος Trimean μετράει το κέντρο της κατανομής πιθανότητας και υπολογίζεται ως μέσος σταθμισμένος όρος της διαμέσου της κατανομής και των δύο τεταρτημορίων.

$$TM = \frac{1}{2}(Q_2 + \frac{Q_1 + Q_3}{2}) \tag{2.1}$$

2.4.2 Συμπερασματική Στατιστική

Η Συμπερασματική Στατιστική αναφέρεται ως η διαδικασία εξαγωγής συμπερασμάτων για ένα πληθυσμό από ένα μικρότερο θορυβώδες δείγμα. Ουσιαστικά με αυτόν τον τρόπο προσπαθούμε να εξάγουμε γνώση με (απλά) μοντέλα για να εξηγήσουμε πολύπλοκα φαινόμενα. Τα πιθανοκρατικά μοντέλα που θα κατασκευάσουμε θα χρησιμοποιηθούν για να περιγράψουμε τον κόσμο και τη χρήση αυτών των πιθανοκρατικών μοντέλων ως σύνδεση μεταξύ των δεδομένων και του πληθυσμού που αντιπροσωπεύει τον πιο αποτελεσματικό τρόπο για εξαγωγή συμπεράσματος.

Η εξαγωγή γενικών συμπερασμάτων για έναν πληθυσμό χρησιμοποιόντας μόνο ένα δείγμα των δεδομένων δεν είναι εύκολη υπόθεση. Ο αναλυτής θα πρέπει μεταξύ άλλων να προσέξει και τα εξής:

- Είναι το δείγμα αντιπροσωπευτικό του πληθυσμού που θέλουμε να εξάγουμε συμπέρασμα;
- Υπάρχουν γνωστές και παρατηρήσιμες, γνωστές και μη παρατηρήσιμες ή άγνωστες και μη παρατηρήσιμες μεταβλητές που θορυβίζουν τα συμπερασματά μας;
- Υπάρχει συστηματική μεροληψία που δημιουργείται από ελλιπής τιμές ή το σχεδιασμό και την εξαγωγή της μελέτης;
- Τι ποσοστό τυχαιότητας υπάρχει στα δεδομένα και πως μπορούμε να το ρυθμίσουμε; Η τυχαιότητα αυτή μπορεί να οφείλεται ή από τυχαία επιλογή ή από τυχαία δειγματοληψία ή έμμεσα από το άθροισμα πολλών άγνωστων και πολύπλοκων παραγόντων.
- Μήπως προσπαθούμε να εκτιμήσουμε ένα υποκείμενο μοντέλο φαινομένων υπό μελέτη;

Η Συμπερασματική Στατιστική απαιτεί την έρευνα σε ένα σύνολο παραδοχών και εργαλείων και τη συνεχή σκέψη σχετικά με το πώς να εξαχθούν συμπεράσματα από τα δεδομένα.

Ο Στόχος Της Συμπερασματικής Στατιστικής

Ο αναλυτής θα πρέπει να είναι σε θέση να αναγνωρίζει πότε πρέπει να χρησιμοποιήσει συμπερασματική στατιστική. Μερικοί από τους στόχους αυτής της μεθόδου είναι:

- Η εκτίμηση και η ποσοτικοποίηση της αβεβαιότητας μιας εκτίμησης του πληθυσμού (π.χ. Η αναλογία των ανθρώπων που θα ψηφίσει για έναν υποψήφιο).
- Ο προσδιορισμός αν η ποσότητα του πληθυσμού είναι ικανή τιμή αναφοράς (π.χ. Είναι η θεραπεία αποτελεσματική;).

- Να συμπεράνει μια καθολική σχέση, όταν οι ποσότητες που μετρούνται έχουν θόρυβο.
- Ο προσδιορισμός των επιπτώσεων διαφορετικών επιλογών πολιτικής (π.χ. Αν μειώσουμε τα επίπεδα ρύπανσης, τα ποσοστά άσθματος θα μειωθούν;).
- Να εξηγήσει την πιθανότητα κάτι να συμβεί.

Εργαλεία Της Συμπερασματικής Στατιστικής

Η συμπερασματική στατιστική αποτελείται από μια πληθώρα εργαλείων. Μερικά από τα πιο σημαντικά είναι τα παρακάτω: [7]

1. *Τυχαιοποίηση*: ασχολείται με την εξισορρόπηση απαραίτητων μεταβλητών που μπορεί να επηρεάσουν τα συμπεράσματα.
2. *Τυχαία Δειγματοληψία*: ασχολείται με την απόκτηση των δεδομένων που είναι αντιπροσωπευτικά του ενδιαφερόμενου πληθυσμού.
3. *Μοντέλα Δειγματοληψίας*: ασχολούνται με τη δημιουργία ενός μοντέλου για τη διαδικασία δειγματοληψίας.
4. *Έλεγχοι Υποθέσεων*: ασχολούνται με τη λήψη αποφάσεων υπό την παρουσία αβεβαιότητας.
5. *Διαστήματα Εμπιστοσύνης*: ασχολούνται με την ποσοτικοποίηση της αβεβαιότητας στις εκτιμήσεις.
6. *Πιθανοκρατικά Μοντέλα*: μια τυπική σύνδεση μεταξύ των δεδομένων και του πληθυσμού που μας ενδιαφέρει. Συχνά, τα πιθανοκρατικά μοντέλα είναι υποτιθέμενα ή προσεγγίζονται.
7. *Σχεδιασμός Μελέτης*: η διαδικασία σχεδιασμού ενός πειράματος για να ελαχιστοποιήσουμε τη μεροληψία και τις διακυμάνσεις.
8. *Μέθοδος Bootstrapping*: η διαδικασία της χρήσης των δεδομένων με επανατοποθέτηση και με τις ελάχιστες παραδοχές στα πιθανοκρατικά μοντέλα για την εξαγωγή συμπερασμάτων.
9. *Μεταθέσεις, Τυχαιοποίηση και Εναλλαγή Δοκιμών*: η διαδικασία χρήσης μεταθέσεων στα δεδομένα για την εξαγωγή συμπερασμάτων.

Είδη Συμπερασματικής Στατιστικής

Υπάρχουν αρκετά διαφορετικά στυλ συμπερασματικής στατιστικής. Τα σημαντικότερα από αυτά είναι δύο: [7]

- *Frequentist Συμπερασματική Στατιστική*: χρησιμοποιεί ερμηνείες των Πιθανοτήτων Συχνοτήτων[3] για τον έλεγχο του ποσοστού σφάλματος.

[3] *Πιθανότητες Συχνοτήτων* είναι η μακροπρόθεσμη αναλογία των εμφανίσεων ενός γεγονότος σε ανεξάρτητες, ομοιόμορφα κατανεμημένες επαναλήψεις.

Απαντάει σε ερωτήσεις όπως «Τι πρέπει να αποφασίσω από τα δεδομένα μου με ελεγχόμενο μακροπρόθεσμο και αποδεκτό ποσοστό σφάλματος».

- *Bayesian Συμπερασματική Στατιστική*: χρησιμοποιεί αναπαραστάσεις Bayesian Πιθανοτήτων[4] για την εξαγωγή συμπερασμάτων. Απαντάει σε ερωτήσεις όπως «Δεδομένου των υποκειμενικών μου πεποιθήσεων και της αντικειμενικής πληροφορίας από τα δεδομένα, τι θα πρέπει να αποφασίσω τώρα».

2.4.3 Πολυμεταβλητή Ανάλυση

Η Πολυμεταβλητή Ανάλυση όπως φαίνεται και από την ονομασία της αναφέρεται σε μεθόδους όπου με τη χρήση πολλών μεταβλητών προσπαθούμε να εξάγουμε συμπεράσματα. Οι μέθοδοι αυτοί για χρόνια δεν χρησιμοποιούνταν λόγο της αυξημένης πολυπλοκότητας τους και της δύσκολη εφαρμογή τους στην πράξη. Σήμερα όμως με την ανάπτυξη της τεχνολογίας αυτές οι δυσκολίες περιορίζονται συνεχώς. Στην πράξη, τα δεδομένα που έχει μια επιχείρηση αλλά και γενικότερα, αποτελούνται από πολλές μεταβλητές και σχεδόν πάντα αυτές σχετίζονται και επηρεάζουν η μία την άλλη. Σκοπός της πολυμεταβλητής ανάλυσης είναι να ανακαλύψει συσχετισμούς των μεταβλητών και να χρησιμοποιήσει όλη τη διαθέσιμη πληροφορία.

Τα Πολυμεταβλητά δεδομένα προκύπτουν όταν οι ερευνητές καταγράφουν τις τιμές αρκετών τυχαίων μεταβλητών σε ένα αριθμό θεμάτων ή αντικειμένων ή ίσως μιας πληθώρας από άλλα πράγματα, η οποίες μας ενδιαφέρουν, οδηγώντας στην αναπαράσταση ενός διανύσματος ή μιας πολυδιάστατης παρατήρησης για το καθένα. Τα δεδομένα αυτά συλλέγονται σε ένα ευρύ φάσμα επιστημονικών κλάδων, και πράγματι είναι μάλλον λογικό να ισχυριστεί ότι η πλειοψηφία του συνόλου δεδομένων που συναντώνται στην πράξη είναι πολυμεταβλητά. Σε ορισμένες μελέτες, οι μεταβλητές επιλέγονται στο σχεδιασμό, επειδή είναι γνωστό ότι είναι ουσιώδης περιγραφές του συστήματος υπό έρευνα. Σε άλλες μελέτες, ιδιαίτερα εκείνων που είναι δύσκολες ή δαπανηρές στην οργάνωση, πολλές μεταβλητές μπορεί να μετρηθούν απλώς με τη συγκέντρωση όσο το δυνατόν περισσότερων πληροφοριών με στόχο την ευκολία και την οικονομία της διαδικασίας. [8]

Μερικοί λόγοι για τους οποίους η πολυμεταβλητή ανάλυση είναι ιδιαίτερα χρήσιμη είναι οι εξής:

- Την ανακάλυψη και τη δημιουργία ομάδων, σύμφωνα με κάποια χαρακτηριστικά, μέσα από μεταβλητές
- Την εύρεση και ερμηνεία συσχετίσεων μεταξύ των μεταβλητών.
- Τη μείωση των διαστάσεων του προβλήματος, δηλαδή τη συρρίκνωση της πληροφορίας πολλών μεταβλητών σε λιγότερες.

[4]Οι *Bayesian Πιθανότητες* αφορά τον υπολογισμό εξαρτημένων πιθανοτήτων, δεδομένου ότι οι εξαρτήσεις ακολουθούν ορισμένους κανόνες.

- Την πρόβλεψη νέων ή ελλιπών τιμών.
- Την αναγωγή του προβλήματος σε πολλές διαστάσεις που μπορεί να μας βοηθήσει να ερμηνεύσουμε πολλές μεταβλητές σε σχέση με άλλες.
- Τη ποσοτικοποίηση μεταβλητών που δεν είναι άμεσα παρατηρήσιμες.

Αξίζει να σημειωθεί βέβαια ότι η πολυμεταβλητή ανάλυση μπορεί να οδηγήσει σε ασαφή αποτελέσματα διότι οργανώνει πολύπλοκες συσχετίσεις όπως επίσης μπορεί να επιφέρει αχρείαστη πολυπλοκότητα διότι δεν χρειάζεται πάντα να γίνει ανάλυση των συσχετίσεων σε μεγάλο βάθος.

Μέθοδοι Πολυμεταβλητής Ανάλυσης

Υπάρχει μια μεγάλη γκάμα από μεθόδους και αλγορίθμους πολυμεταβλητής ανάλυσης. Μερικές από τις πιο σημαντικές είναι οι παρακάτω:

Ανάλυση Κύριων Συνιστωσών (Principal Compoment Analysis): Η μέθοδος της Ανάλυσης Κύριων Συνιστωσών είναι η πιο γνωστή μέθοδος MVA διότι είναι σχετικά εύκολη στην εφαρμογή και τα αποτελέσματα που παράγει δεν έχουν μεγάλη πολυπλοκότητα στην ερμηνεία τους. Αποτελεί ένα μαθηματικό μετασχησμό των δεδομένων και ουσιαστικά ο αλγόριθμος PCA βρίσκει γραμμικούς συνδυασμούς από τα δεδομένα με στόχο οι νέες μεταβλητές που θα προκύψουν να είναι ασυσχέτιστες μεταξύ τους αλλά να περιέχουν όσο το δυνατό όλη την πληροφορία από τα αρχικά δεδομένα. Βρίσκει δηλαδή τις συσχετίσεις μεταξύ των μεταβλητών και οργανώνει αυτές τις μεταβλητές και μειώνει την διάσταση των δεδομένων επιτρέποντας την ανακάλυψη απλούστερων ερμηνεύσιμων σχέσεων μεταξύ των μεταβλητών.

Ομαδοποίηση (Cluster Analysis): Η Ομαδοποίηση πρόκειται για τη δημιουργία ομάδων από παρατηρήσεις με τη χρήση μεταβλητών για τις οποίες τα δεδομένα δείχνουν πως έχουν παρόμοια ή κοινά χαρακτηριστικά. Υπάρχουν αρκετοί αλγόριθμοι και μεθοδολογίες για να επιτευχθεί αυτό και κάποιοι από αυτούς είναι εμπειρικοί ενώ άλλοι βασίζονται σε μοντέλα (π.χ. k-Means), οι οποίοι έχουν ένα σημαντικό θεωρητικό υπόβαθρο και μπορούν εύκολα να συγκριθούν και να αξιολογηθούν για την ορθότητα τους. Οι περισσότεροι από αυτούς βασίζονται στην έννοια της απόστασης.

Παραγοντική Ανάλυση (Factor Analysis): Η Παραγοντική Ανάλυση αφορά στην εύρεση και στην ερμηνεία παραγόντων που δεν είναι άμεσα μετρήσιμοι αλλά όμως υπάρχουν και προκαλούν συσχετίσεις μεταξύ των παρατηρούμενων μεταβλητών. Σε αντίθεση με την μέθοδο PCA, επιτρέπει τη στατιστική εξέταση διάφορων υποθέσεων σχετικά με τα φαινόμενα που μελετούνται και είναι μια μέθοδος που χρησιμοποιείται αρκετά σε κοινωνικές επιστήμες και αλλά και σε αναλύσεις marketing όπως διαχείριση προϊόντων κ.ά.

Παραγοντική Ανάλυση Αντιστοιχιών (Correspondence Analysis): Η Παραγοντική Ανάλυση Αντιστοιχιών είναι παρόμοια μέθοδος με την μέθοδος Ανάλυση Κύριων Συνιστωσών. Αντίθετα με την μέθοδο PCA η Παραγοντική Ανάλυση Αντιστοιχιών ασχολείται με κατηγορικά δεδομένα. Ουσιαστικά δημιουργεί άξονες πάνω στους οποίους προβάλλονται οι μεταβλητές και έτσι δίνεται η δυνατότητα να ανακαλυφθούν ερμηνεύσιμες συσχετίσεις μεταξύ τους.

Διαχωριστική Ανάλυση (Discriminant Analysis): Η Διαχωριστική Ανάλυση πρόκειται για την δημιουργία κανόνων από τα δεδομένα με στόχο να μπορεί να γίνει πρόβλεψη άγνωστων δεδομένων στις υπάρχουσες ομάδες. Χωρίζεται σε τεχνικές που χρησιμοποιούν υποθετικές παραμετροποιήσεις και σε τεχνικές χωρίς τη χρήση παραμετρικών τεχνικών. Η πιο γνωστή μέθοδος είναι η Γραμμική Διαχωριστική Ανάλυση (Linear discriminant analysis - LDA) και είναι στενά συνδεδεμένη με την PCA και την FA.

Πολυδιάστατη Κλιμάκωση (Multidimensional Scaling): Η Πολυδιάστατη Κλιμάκωση αποτελεί τη μέθοδο όπου ο αναλυτής προβάλλει τις διαστάσεις των δεδομένων συνήθως στο χώρο των δύο ή περισσότερων διαστάσεων και έχει σκοπό ουσιαστικά να διευκολύνει την ερμηνεία των αποτελεσμάτων που θα προκύψουν και να παρουσιαστούν σε διαγράμματα λιγότερων διαστάσεων.

2.4.4 Εξόρυξη Δεδομένων

Η Εξόρυξη Δεδομένων αποτελεί τον πυρήνα κάθε συστήματος επιχειρηματικής ευφυίας και ουσιαστικά αποτελείται από όλα τα εργαλεία επιχειρηματικής ευφυίας που αναφέρθηκαν προηγουμένως αλλά και πολλών ακόμα. Ουσιαστικά είναι ένα συνονθύλευμα από Συστήματα Βάσεων Δεδομένων, τεχνικών Ομαδοποίησης και Κατηγοριοποίησης (2.4.3), Συμπερασματικής Στατιστικής (2.4.2), Τεχνητής Νοημοσύνης, Αναγνώρισης Προτύπων και Μηχανικής Μάθησης. Σκοπός της εξόρυξης δεδομένων είναι η ανακάλυψη γνώσης από (μεγάλα) σύνολα δεδομένων και η εξόρυξη προτύπων συνήθως με αυτόματο τρόπο όπου ο αναλυτής και τα στελέχη μιας επιχείρησης θα χρησιμοποιήσουν για να πάρουν αποφάσεις. Όλες οι μεγάλες επιχειρήσεις σήμερα βασίζουν σε μεγάλο βαθμό τις αποφάσεις στους σε συστήματα που χρησιμοποιούν εξόρυξη δεδομένων. Κυρίως, η εξόρυξη δεδομένων παράγει προβλεπτικά μοντέλα που βασίζονται σε πρότυπα από τα δεδομένα της βάσης δεδομένων και δοκιμάζονται σε άγνωστα δεδομένα του πραγματικού κόσμου.

Μέθοδοι Εξόρυξης Δεδομένων

Γενικά η εξόρυξη δεδομένων βασίζεται σε δύο τύπους μεθοδολογιών: *Περιγραφική Εξόρυξη Δεδομένων* όπου περιγράφουν τις γενικές ιδιότητες των δεδομένων και *Προβλεπτική Εξόρυξη Δεδομένων* όπου επιχειρεί να προβλέψει βάση συμπερασμάτων από τα διαθέσιμα δεδομένα.

Οι κυριότερες μεθοδολογίες εξόρυξης δεδομένων είναι οι εξής: [9], [10]

Κατηγοριοποίηση (Classification): Η Κατηγοριοποίηση είναι η οργάνωση των δεδομένων σε κλάσεις. Πρόκειται για τεχνική επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιεί δοσμένες κλάσεις για να κατηγοριοποιεί αντικείμενα. Το μοντέλο της κατηγοριοποίησης αρχικά εκπαιδεύεται σε ένα σύνολο δεδομένων που έχουν από πριν χωριστεί σε σταθερές κλάσεις και αφού έχει εκπαιδευτεί χτίζεται το μοντέλο και είναι σε θέση να χρησιμοποιηθεί για να κατηγοριοποιήσει άγνωστα δεδομένα.

Παλινδρόμηση (Regression): Η Παλινδρόμηση έχει ιδιαίτερο ενδιαφέρον για τις επιχειρήσεις λόγω των πιθανών θετικών επιπτώσεων από τις επιτυχείς προβλέψεις. Υπάρχουν δύο κύριοι τύποι παλινδρόμησης: η πρόβλεψη κάποιας ελλιπής τιμής στα δεδομένα ή τάσεις που τείνουν να έχουν αυτά και η πρόβλεψη μιας κατηγορίας για ορισμένα δεδομένα. Η τελευταία συνδέεται με την κατηγοριοποίηση. Η παλινδρόμηση ωστόσο, πιο συχνά αναφέρεται στην πρόβλεψη των ελλιπών αριθμητικών τιμών ή την αύξηση και μείωση τάσεων σε χρονικά δεδομένα. Η κύρια ιδέα είναι η χρησιμοποίηση ενός μεγάλου αριθμού παρατηρούμενων τιμών για να εξετάσουν πιθανές μελλοντικές τιμές.

Ομαδοποίηση (Clustering): Παρόμοια με την κατηγοριοποίηση, η Ομαδοποίηση είναι η οργάνωση των δεδομένων σε ομάδες. Ωστόσο, σε αντίθεση με την κατηγοριοποίηση, στην ομαδοποίηση, οι ομάδες είναι άγνωστες και εξαρτάται από τον αλγόριθμο για να ανακαλύψει αποδεκτές ομάδες. Η ομαδοποίηση είναι μία κατηγορία μη επιβλεπόμενης μηχανικής μάθησης, διότι οι ομάδες δεν είναι γνωστές από την αρχή. Υπάρχουν αρκετοί αλγόριθμοι ομαδοποίησης αλλά όλοι έχουν σαν αρχή την μεγιστοποίηση της ομοιότητας μεταξύ των δεδομένων σε μία ομάδα και ελαχιστοποιώντας της ομοιότητας μεταξύ των δεδομένων διαφορετικών ομάδων.

Ανάλυση Κανόνων Συσχετισμού (Association Rules Analysis): Η Ανάλυση Κανόνων Συσχετισμού μελετά τη συχνότητα των στοιχείων που εμφανίζονται μέσα σε ένα σύστημα βάσεων δεδομένων, βασισμένη σε ένα κατώτατο όριο που ονομάζεται Υποστήριξη, όπου προσδιορίζει τα συχνότερα σύνολα δεδομένων. Ένα άλλο όριο, η Εμπιστοσύνη, η οποία είναι η δεσμευμένη πιθανότητα από ένα στοιχείο που εμφανίζεται σε μια ενέργεια, όταν εμφανίζεται ένα άλλο στοιχείο και χρησιμοποιείται για να εντοπίσει κανόνες συσχέτισης. Το πρώτο μέρος ενός κανόνα συσχετισμού ονομάζεται Υπόθεση και το τελευταίο ονομάζεται Συμπέρασμα. Η ανάλυση κανόνων συσχετισμού συνήθως χρησιμοποιείται για την ανάλυση του καλαθιού αγοράς.

Χαρακτηρισμός (Characterization:): Ο Χαρακτηρισμός των δεδομένων είναι μια περιληπτική παρουσίαση των γενικών χαρακτηριστικών των δεδομένων σε μία συγκεκριμένη κλάση όπου εξάγει χαρακτηριστικούς κανόνες. Τα δεδομένα που ανήκουν σε μία κλάση που έχει ορίσει ο χρήστης συνήθως προέρχονται από ένα ερώτημα βάσης δεδομένων και εκτελείται μία λειτουργία εξαγωγής της "περίληψης" σε αυτά. Αξίζει να σημειωθεί

ότι σε πολυδιάστατα δεδομένα τα οποία αποτελούνται από μία περιληπτική παρουσίαση των δεδομένων, οι απλές εργασίες OLAP ταιριάζουν στο σκοπό του χαρακτηρισμού των δεδομένων.

Διακριτοποίηση (Discrimination): Η Διακριτοποίηση των δεδομένων παράγει κανόνες διάκρισης και είναι ουσιαστικά η σύγκριση των γενικών χαρακτηριστικών από αντικείμενα μεταξύ δύο κλάσεων που ονομάζονται κλάση στόχος και αντίθετη κλάση. Οι τεχνικές που χρησιμοποιούνται για τη διακριτοποίηση των δεδομένων είναι παρόμοιες με τις τεχνικές που χρησιμοποιούνται για τον χαρακτηρισμό των δεδομένων με τη διαφορά ότι τα αποτελέσματα της διακριτοποίησης περιλαμβάνουν συγκρίσιμα μέτρα.

Ανάλυση Ακραίων Τιμών (Outlier Analysis): Οι ακραίες τιμές είναι τα δεδομένα τα οποία δεν μπορούν να ομαδοποιηθούν σε μια ομάδα ή κλάση και είναι συχνά πολύ σημαντικές να προσδιοριστούν. Ενώ οι ακραίες τιμές μπορεί να θεωρούνται θόρυβος και να απορρίπτονται σε μερικές εφαρμογές, μπορεί να αποκαλύψουν σημαντική γνώση σε άλλες εφαρμογές και έτσι να είναι πολύτιμη η ανάλυσή τους.

2.4.5 Αναφορές Τελικού Χρήστη

Ένα συστήματα επιχειρηματικής ευφυίας οφείλει να περιέχει αποδοτικές μεθόδους παρουσίασης των αποτελεσμάτων των μεθόδους που περιγράψαμε πιο πάνω, όπως επίσης και να επιτρέπει στο χρήστη να επεξεργάζεται σε πραγματικό χρόνο τα δεδομένα και να θέτει ερωτήματα προς αυτά.

Μερικές μέθοδοι όπου ένα σύστημα επιχειρηματικής ευφυίας μπορεί να περιέχει για τη δημιουργία αναφορών προς στο χρήστη είναι: [11], [12]

Άμεση Επεξεργασία Συναλλαγών (OLTP): Τα συστήματα OLTP χαρακτηρίζονται από ένα μεγάλο αριθμό σύντομων συναλλαγών με την βάση. Το κύριο χαρακτηριστικό των OLTP είναι η έμφαση που δίνουν στην γρήγορη επεξεργασία ερωτημάτων, διατηρώντας την ακεραιότητα των δεδομένων σε περιβάλλοντα με πολλαπλές προσβάσεις και με αποτελεσματικότητα που μετριέται από των αριθμό συναλλαγών ανά δευτερόλεπτο. Σε μία βάση δεδομένων OLTP υπάρχουν λεπτομερείς και πραγματικού χρόνου δεδομένα και το σχήμα της βάσης που συνήθως χρησιμοποιείται είναι το μοντέλο οντοτήτων - συσχετίσεων. Ένα από τα κύρια αρνητικά αυτού του συστήματος είναι η αδυναμία του να αντεπεξέλθει σε συναθροιστικές ερωτήσεις που απαιτούν πολύπλοκους μαθηματικούς υπολογισμούς, λόγο του μεγάλου χρόνου που θα δεσμεύσει την βάση και με αποτέλεσμα να μην είναι προσπελάσιμη όλη εκείνη τη στιγμή.

Άμεση Επεξεργασία Δεδομένων (OLAP): Σε αυτό το επίπεδο, τα συστήματα επιχειρηματικής ευφυίας παρέχουν δυνατότητες προβλεπτικής και προγνωστικής ανάλυσης (Predictive/Forecasting Analysis), ανάλυσης τάσεων (Trend Analysis), οικονομικές αναφορές και εκθέσεις

προϋπολογισμού για περαιτέρω εξόρυξη γνώσης. Τα ερωτήματα στα συστήματα OLAP, είναι συχνά αρκετά πολύπλοκα και περιέχουν συναθροίσεις και πολύπλοκους μαθηματικούς υπολογισμούς. Για ένα τέτοιο σύστημα οι απαντήσεις βρίσκονται σε πραγματικό χρόνο και η ταχύτητα να απαντηθούν είναι το μέτρο αποτελεσματικότητας τους. Η μέθοδος OLAP αποτελείται από τρεις βασικές τεχνικές ανάλυσης: Συσσωμάτωση (Roll-Up), Drill-Down και Slice & Dice. Η τεχνικής της συσσωμάτωσης, αφορά την ενοποίηση και συσσώρευση δεδομένων σε διάφορες διαστάσεις. Αντίθετα, η τεχνική Drill-Down επιτρέπει στα στελέχη να αναλύσουν τα δεδομένα με μεγαλύτερη λεπτομέρεια. Τέλος, με την τεχνική Slice & Dice οι χρήστες μπορούν να χωρίσουν τα δεδομένα σε διάφορα υποσύνολα με βάση κάποιες διαστάσεις και να τα συγκρίνουν από διαφορετικές οπτικές γωνίες. Οι εφαρμογές των συστημάτων OLAP χρησιμοποιούν ευρέως τεχνικές εξόρυξης δεδομένων. Η βάση δεδομένων που χρησιμοποιεί περιέχει συναθροιστικά και ιστορικά δεδομένα, αποθηκευμένα σε πολυδιάστατα σχήματα βάσης.

Αναφορές (Reports): Οι αναφορές έχουν ως στόχο την δημιουργία ανά τακτά χρονικά διαστήματα αποδοτικών και φιλικών ως προς τη χρήση εκθέσεων. Η μορφή των αναφορών είναι προκαθορισμένη από την αρχή και παράγονται είτε αυτόματα σε περιοδικό χρόνο είτε έπειτα από απαίτηση του χρήστη, και σκοπός τους είναι η ομαδοποίηση δεδομένων από διάφορες, αξιόπιστες, πηγές.

Αναλύσεις (Analytics): Ένα από τα βασικά επίπεδα ενός συστήματος επιχειρηματικής ευφυίας είναι οι αναλύσεις. Πρόκειται για αυτοματοποιημένες ευφυείς αναλύσεις δεδομένων, βασισμένες σε αρκετά εκλεπτυσμένες τεχνικές όπως η ασαφής λογική και η λογική ασαφούς νευρώνων[5]. Ένα από τα κύρια χαρακτηριστικά αυτού του επιπέδου, είναι ότι έχει δημιουργηθεί με τέτοιο τρόπο ώστε να είναι φιλικό προς το χρήστη και κρύβοντας όλη την πολυπλοκότητα των τεχνικών που χρησιμοποιεί από πίσω. Παράγει χρήσιμες συσχετίσεις για τα δεδομένα και βασίζεται σε ανάλυση σεναρίων WHAT-IF και προηγμένων προσομοιώσεων που βοηθούν στην καλύτερη λήψη αποφάσεων.

Συγκεντρωτικούς Πίνακες - Γραφήματα (Dashboards): Περιέχουν υψηλού επιπέδου συγκεντρωτικές παρουσιάσεις δεδομένων της επιχείρησης, κυρίως γραφήματα, διαγράμματα και δείκτες απόδοσης, τόσο στατικού όσο και δυναμικού περιεχομένου. Επιτρέπουν επίσης, κάποιες βασικές αλληλεπιδράσεις αλλά και διάφορα επίπεδα εμβάθυνσης δίνοντας την δυνατότητα εξαγωγής χρήσιμης γνώσης. Να σημειωθεί βέβαια, ότι η επεξηγηματική δύναμη αυτών των παρουσιάσεων βασίζεται κυρίως στην ερμηνεία που δίνει ο χρήστης στη γνώση.

[5]Η λογική ασαφούς νευρώνων (Neuro-fuzzy) είναι μια μέθοδος τεχνητής νοημοσύνης που βασίζεται σε ένα συνδυασμό τεχνιτών νευρωνικών δικτύων και ασαφής λογικής. Σκοπός τους είναι η προσομοίωση ενός συνδυαστικού υβριδικού συστήματος που βασίζεται στην λογική του ανθρώπινου εγκεφάλου και πως αυτός μαθαίνει και αντιλαμβάνεται την ασάφεια.

3

Ανάλυση Δεδομένων Συναλλαγών Καλαθιού Αγοράς

3.1 Εισαγωγή

Η ανάλυση δεδομένων από συναλλαγές πελατών αποτελεί ένα από τα πιο σημαντικά ζητήματα στο τομέα του marketing και της λήψης αποφάσεων σε μια επιχείρηση. Τα τελευταία 30 χρόνια έχει γίνει μεγάλη έρευνα και έχουν προταθεί διάφορες μέθοδοι με πολύ καλά αποτελέσματα. Μια επιχείρηση για να μπορέσει να είναι ανταγωνιστική και να επιβιώσει σήμερα είναι απαραίτητο να συμπεριλάβει στο ρεπερτόριο των αναλύσεων Επιχειρηματικής Ευφυίας τέτοιες αναλύσεις. Τυπικά, η ανάλυση συναλλαγών καλαθιού αγοράς μπορεί να οριστεί ως οι ενέργειες που πρέπει να γίνουν ώστε να ανακαλυφθεί χρήσιμη και εφαρμόσιμη γνώση από μία βάση συναλλαγών ή με άλλα λόγια σύμφωνα με τον Julander[13] αποτελεί την μελέτη προϊόντων που έχουν αγοραστεί μαζί από ένα πελάτη. Επίσης οι Russell και Petersen[14] αναφέρουν πως η ανάλυση αυτή επικεντρώνεται στη διαδικασία λήψης αποφάσεων στην οποία ο καταναλωτής επιλέγει προϊόντα από ένα δεδομένο σύνολο των κατηγοριών στην ίδια αγορά. Τα δεδομένα καλαθιού ενός καταστήματος μπορούν να αναλυθούν από διάφορες οπτικές γωνίες όπως από την πλευρά της ημερομηνίας της συναλλαγής (πόσα και ποια προϊόντα αγοράστηκαν μαζί σε ένα χρονικό πλαίσιο), τα χαρακτηριστικά των προϊόντων (επωνυμία, προώθηση, εύρος τιμής) αλλά και στα προϊόντα κάθε αυτά (αφορά κάθε επίπεδο αφαίρεσης των κατηγοριών). Επίσης, η ανάλυση μπορεί να έχει σαν στόχο τα προϊόντα ή τους πελάτες.

Στη βιβλιογραφία όπως έχει αναφέρθηκε και προηγουμένως, υπάρχουν αρκετοί τρόποι ανάλυσης τέτοιων δεδομένων όπου κάθε ένας προσπαθεί να ερμηνεύσει τα δεδομένα από διαφορετική σκοπιά. Μία από τις πιο σημαντικές μεθόδους ανάλυσης είναι η εξόρυξη συχνών στοιχειοσύνολων και κανόνων συσχέτισης όπως προτάθηκε από τον Agrawal[15]. Ακόμα, ο αναλυτής μπορεί να δει τα δεδομένα αυτά με την σκέψη ότι θα αναζητήσει ομάδες πελατών ή συναλλαγών. Τα αποτελέσματα της ανάλυσης δεδομένων από συναλλαγές είναι

ποικίλα και μερικά από αυτά στοχεύουν στο να δώσει στην επιχείρηση τη δυνατότητα να κατανοήσει την συμπεριφορά των πελατών της και να διαμορφώσει ανάλογα τις αποφάσεις που θα πρέπει να πάρει ώστε να μπορέσει να κρατήσει τους πελάτες της, να προσελκύσει νέους, να μεγιστοποιήσει τα κέρδη της αλλά και να δημιουργήσει μία σχέση εμπιστοσύνης μεταξύ των πελατών της. Πιο συγκεκριμένα, η επιχείρηση, μέσω τέτοιων αναλύσεων μπορεί να καταλάβει την αγοραστική συμπεριφορά των πελατών λαμβάνοντας υπόψη διάφορα κριτήρια (δημογραφικά στοιχεία, περίοδος πωλήσεων, επωνυμία, τιμή κ.α), να προβλέψει ποια προϊόντα θα αγοράσει στο άμεσο μέλλον ο πελάτης και κατάλληλα να τα προωθήσει, όπως ακόμα να βρει και τις συσχετίσεις προϊόντων.

Όταν το ενδιαφέρων της επιχείρησης έχει σαν στόχο την έρευνα από την πλευρά των προϊόντων, η ανάλυση επικεντρώνεται σε δεδομένα ταυτόχρονων συναλλαγών όπου επιτρέπουν στην ανακάλυψη της σχέσης μεταξύ των απαιτήσεων των προϊόντων (ανεξαρτησία, υποκατάσταση, συμπληρωματικότητα) και συγχρόνως, εξαλείφει απλές συμπτώσεις (π.χ. η ανακάλυψη σχέσεων από την ταυτόχρονη αγορά προϊόντων λόγω της αγοραστικής συχνότητάς τους ή τη διαδρομή που ακολούθησε κατά τη διάρκεια της πορείας μιας αγοράς). Όταν το ενδιαφέρων της επιχείρησης έχει σαν στόχο την έρευνα από την πλευρά των πελατών, η ερώτηση που μπορεί να διατυπωθεί είναι να καθοριστούν προσωπικές προτιμήσεις πελατών ή ομάδων πελατών για τα εμπορεύματα, τις εμπορικές επωνυμίες ή τα χαρακτηριστικά των προϊόντων και των προωθητικών ενεργειών (μέσα ενημέρωσης κ.α.) σε μια βάση δεδομένων δυαδικών μεταβλητών. Μπορεί επίσης να χρησιμοποιηθεί για να προβλέψει την πιθανότητα για μια αγορά ενός προϊόντος με την εργασία αυτή να εστιάζει σε αυτό το σημείο.

3.2 Βασικές έννοιες μιας Αποθήκης Δεδομένων

> *«Η Αποθήκη Δεδομένων πρόκειται για μία συλλογή δεδομένων που χρησιμοποιείται κυρίως για την λήψη αποφάσεων σε ένα οργανισμό, και είναι θεματικά προσανατολισμένη, έχοντας «ολοκληρωμένα» και τελικά δεδομένα, τα οποία κρατούνται σε βάθος χρόνου χωρίς να διαγράφονται»*
>
> – Bill Inmon, *Building the Data Warehouse (1992)*[16]

Σύμφωνα με τον Inmon, αποθήκη δεδομένων ονομάζεται η βάση δεδομένων που σχεδιάστηκε για ανάλυση και επεξεργασία δεδομένων. Αποτελεί μια αποθήκη από ολοκληρωμένες, συναθροιστικές και τελικές πληροφορίες οι οποίες είναι διαρκώς διαθέσιμες για ερωτήσεις και αναλύσεις από τα στελέχη της επιχείρησης με στόχο, να διευκολύνουν την αναλυτική επεξεργασία του οργανισμού και την υποστήριξη λήψης αποφάσεων. Διατηρεί δεδομένα από όπου αντλεί από βάσεις δεδομένων των πληροφοριακών συστημάτων της επιχείρησης όπως επίσης και διάφορες άλλες πηγές δεδομένων (αρχεία καταγραφών του οργανισμού, δεδομένα από εξωτερικές πηγές κ.α.). Τα δεδομένα αυτά οργανώνονται και δομούνται στην αποθήκη δεδομένων σε δομές, κατάλληλες και

έτοιμες να ικανοποιήσουν αποδοτικά τις απαιτήσεις των υπεύθυνων στελεχών για τη λήψη και τη στήριξη αποφάσεων. Μια αποθήκη δεδομένων θα μπορούσε να χαρακτηριστεί ως μια μεγάλη Βάση Δεδομένων με συγκεκριμένες λειτουργίες, όπου ο χρήστης έχει πρόσβαση μόνο για ανάγνωση και του δίνεται η δυνατότητα να εξάγει, να φιλτράρει και να ολοκληρώνει τη σχετική πληροφορία εκ των προτέρων ενός ερωτήματος.

Οι Αποθήκες Δεδομένων έχουν προσελκύσει ευρύτατο ενδιαφέρον αφού μπορούν να αποτελέσουν λύση για μια πολύ μεγάλη ομάδα προβλημάτων, όπως την ολοκλήρωση, την ενοποίηση και τον καθαρισμό δεδομένων από ετερογενές πηγές συνήθως προερχόμενες από αναχρονισμένα συστήματα βάσεων δεδομένων. Συγχρόνως, υπάρχει τεράστια ποικιλία από προϊόντα και στρατηγικές από προμηθευτές, τεράστια γκάμα εφαρμογών και απαιτήσεων (ακόμη και οι προμηθευτές παραδέχονται ότι κάθε αποθήκη δεδομένων χρειάζεται ειδική κατασκευή και βελτιστοποίηση) και έλλειψη σαφούς και τυπικής κατανόησης απαιτήσεων.

3.2.1 Λειτουργία Της Αποθήκης Δεδομένων

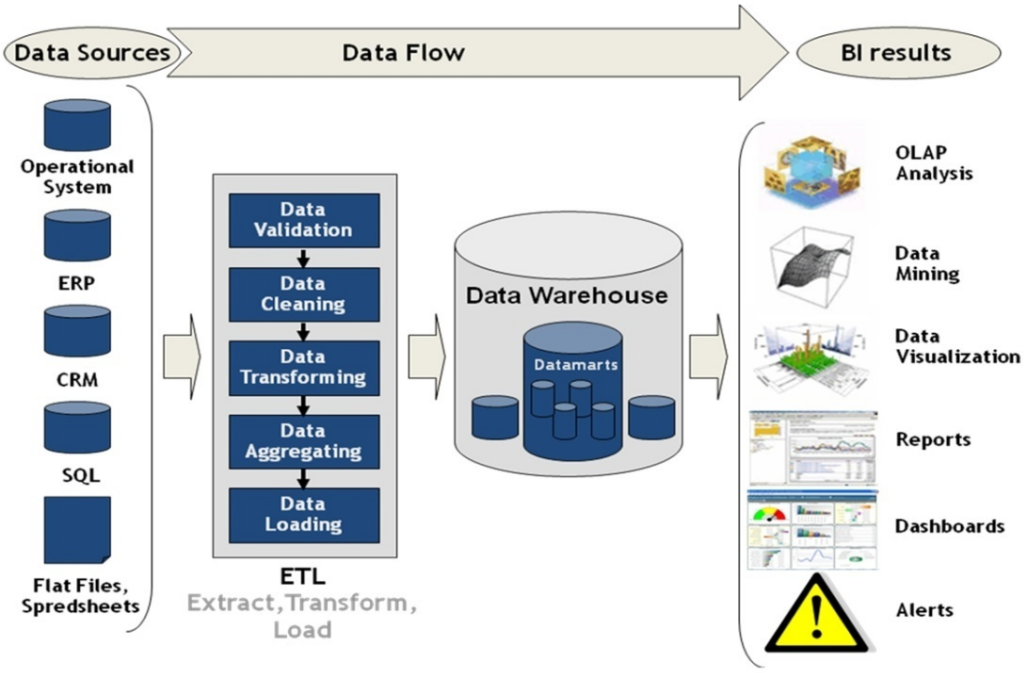
Μερικές από τις βασικές λειτουργίες-ιδιότητες μιας αποθήκης δεδομένων είναι:

- Η χρησιμοποίηση για συλλογή πληροφοριών (ιστορικών και χρονικών κυρίως).
- Η χρησιμοποίηση των συλλογών δεδομένων για την ανάλυσή τους με στόχο τη μεγαλύτερη κατανόηση της επιχείρησης και την εξέλιξή της.
- Η χρησιμοποίηση ως βάση σε Συστήματα Στήριξης και Λήψης Αποφάσεων (DSS)

Στις αποθήκες δεδομένων συνήθως κρατάμε όλη τη διαθέσιμη πληροφορία χωρίς να διαγράφεται οτιδήποτε. Αυτό έχει ως αποτέλεσμα να συγκεντρώνεται ένας μεγάλος όγκος δεδομένων, όπου αποτελούν ιστορικά στοιχεία της επιχείρησης που διατηρεί την αποθήκη δεδομένων.

Επίσης οι αποθήκες δεδομένων:

- Συνδυάζουν όλες τις πληροφορίες που μπορούν να συλλέξουν από διάφορες πηγές όπως OLTP, Spreadsheets, Web, κείμενα κ.α., τα καθαρίζει και τα ελέγχει για ακρίβεια τους και τα οργανώνει έτσι ώστε οι χρήστες να είναι σε θέση να αναζητούν εύκολα αυτό που επιθυμούν.
- Συχνά χωρίζονται σε άλλα μικρότερα τμήματα εξειδικευμένων αποθηκών δεδομένων που ονομάζονται Datamarts. Αυτά τα τμήματα είναι συχνά περισσότερο προτιμητέα καθώς έχουν πιο μικρό κόστος υλοποίησης και χρειάζονται λιγότερο χρόνο για την κατασκευή τους.



Σχημα 3.1: *Η δομή μια τυπικής αποθήκης δεδομένων.*

Η αξία μιας αποθήκης δεδομένων φαίνεται τελικά στην δυνατότητα ανάλυσης που παρέχει στα στελέχη της επιχείρησης. Τα δεδομένα προσπελάζονται και αναλύονται χρησιμοποιώντας διάφορα εργαλεία, όπως ερωτήματα OLAP, μηχανισμούς εξόρυξης δεδομένων και μηχανικής μάθησης και εργαλεία οπτικοποίησης και απεικόνισης πληροφορίας όπως και στατιστικές αναλύσεις.

Σκοπός των αποθηκών δεδομένων είναι η άμεση, αποδοτική και ακριβής απάντηση οποιαδήποτε ερώτησης σχετικά με τη λειτουργία της επιχείρησης στην οποία εφαρμόζεται. Συνεπώς, η σχεδίαση των αποθηκών δεδομένων έχει σαν στόχο την αποδοτική απάντηση πολύπλοκων ερωτήσεων που θέτονται κατά την αναλυτική επεξεργασία δεδομένων από εφαρμογές στρατηγικού σχεδιασμού και λήψης αποφάσεων. Στο παρακάτω σχήμα (3.1) φαίνεται η τυπική δομή μίας αποθήκης δεδομένων.

3.2.2 Χαρακτηριστικά της Αποθήκης Δεδομένων

Τα χαρακτηριστικά μιας αποθήκης δεδομένων όπως προκύπτουν από τον πατέρα του όρου William Inmon είναι:

Προσανατολισμένη σε ένα θέμα (Subject Oriented): Οι αποθήκες δεδομένων είναι σχεδιασμένες ώστε να βοηθήσουν στην ανάλυση των δεδομένων. Για παράδειγμα, για να μάθουμε περισσότερα για τους πελάτες της επιχείρησής μας, μπορούμε να δημιουργήσουμε μια αποθήκη δεδομένων που επικεντρώνετε στις συναλλαγές των πελατών μας και χρησιμοποιώντας την αυτή, μπορούμε να απαντήσουμε σε ερωτήσεις όπως: "Ποιες είναι οι ομάδες των πελατών μας και πως μπορούμε να τους προ-

σφέρουμε τα προϊόντα που χρειάζονται την στιγμή που χρειάζεται;". Αυτή η ικανότητα της αποθήκης δεδομένων την κάνει προσανατολισμένη στο θέμα.

Ολοκληρωμένη (Integrated): Η ολοκλήρωση συνδέεται στενά με τον προσανατολισμό που έχει η αποθήκη δεδομένων. Οι αποθήκες πρέπει να τοποθετήσουν δεδομένα από διαφορετικές πηγές σε ένα συνεχές ανομοιογενές σχήμα. Πρέπει να επιλύσουν προβλήματα όπως η ασυνέπεια μεταξύ των μονάδων αυτών. Όταν το επιτυγχάνουν αυτό τότε λέμε ότι έχουν ολοκληρωθεί.

Ευμετάβλητη (Non-volatile): Ευμετάβλητη σημαίνει ότι από τη στιγμή που τα δεδομένα εισήχθησαν μια φορά στην αποθήκη δεδομένων δεν πρέπει να αλλάξουν. Αυτός είναι και ο σκοπός μιας αποθήκης δεδομένων άλλωστε, δηλαδή να μην έχουμε συχνά την διαγραφή εγγραφών, παρά μόνο προσθήκη νέων εγγραφών ώστε να επιτραπεί η ανάλυση.

Ιστορικά Δεδομένα (Time-variant): Προκειμένου να ανακαλυφθούν οι τάσεις στην επιχείρηση, οι αναλυτές χρειάζονται μεγάλο όγκο δεδομένων. Αυτό είναι αρκετά σημαντικό και έρχεται σε αντίθεση με τη λειτουργία του OLTP, όπου για λόγους απόδοσης απαιτούν να κινηθούν τα ιστορικά δεδομένα σε ένα αρχείο. Αντίθετα μια αποθήκη δεδομένων εστιάζει στην αλλαγή κατά τη διάρκεια του χρόνου.

3.2.3 Πλεονεκτήματα των Αποθηκών Δεδομένων

Ο σωστός σχεδιασμός και υλοποίηση των αποθηκών δεδομένων είναι σε θέση να προσφέρουν στην επιχείρηση σημαντικά πλεονεκτήματα όπως:

Ενισχυμένη Επιχειρηματική Ευφυΐα: Η γνώση θα αποκτηθεί μέσω μιας βελτιωμένης πρόσβασης στις πληροφορίες. Οι διευθυντές και τα στελέχη της επιχείρησης θα είναι σε θέση να παίρνουν αποφάσεις που δεν θα βασίζονται σε περιορισμένα δεδομένα και τη δική τους διαίσθηση. Οι αποφάσεις που επηρεάζουν τη στρατηγική και τη λειτουργία των επιχειρήσεων θα βασίζονται σε αξιόπιστα στοιχεία και θα υποστηρίζονται από αποδείξεις και πραγματικά δεδομένα της επιχείρησης. Επίσης, οι υπεύθυνοι για την λήψη αποφάσεων θα είναι καλύτερα ενημερωμένοι, καθώς θα είναι σε θέση να διερευνούν τα πραγματικά στοιχεία με στόχο να ανακτήσουν πληροφορίες βασισμένες σε προσωπικές τους ανάγκες. Επιπλέον, οι αποθήκες δεδομένων μπορούν να χρησιμοποιηθούν άμεσα στις επιχειρηματικές διαδικασίες, συμπεριλαμβανομένων της κατάτμησης και της ομαδοποίησης της αγοράς, τη διαχείριση αποθεμάτων, τη οικονομική διαχείριση και τις πωλήσεις.

Αυξημένη απόδοση των ερωτημάτων: Οι αποθήκες δεδομένων έχουν σκόπιμα σχεδιαστεί και κατασκευαστεί με έμφαση στην ταχύτητα ανάκτησης και ανάλυσης των δεδομένων μιας επιχείρησης. Επίσης, μια

αποθήκη δεδομένων έχει σχεδιαστεί για την αποθήκευση μεγάλων όγκων δεδομένων αλλά και να είναι σε θέση να διερευνώνται γρήγορα τα δεδομένα. Αυτά τα συστήματα ανάλυσης είναι κατασκευασμένα διαφορετικά από ότι τα συστήματα λειτουργικότητας που επικεντρώνονται στην δημιουργία και τροποποίηση των δεδομένων. Σε αντίθεση, η αποθήκη δεδομένων είναι χτισμένη για την ανάλυση και την ανάκτηση των δεδομένων και όχι αποτελεσματική συντήρηση των επιμέρους εγγραφών (π.χ. συναλλαγές). Περαιτέρω, οι αποθήκες δεδομένων επιτρέπουν την αποφυγή μιας μεγάλης επιβάρυνσης του συστήματος λειτουργικότητας της επιχείρησης και διανέμει αποτελεσματικά το φόρτο του συστήματος σε όλη την τεχνολογική υποδομή του οργανισμού.

Επιχειρηματική ευφυΐα από πολλαπλές πηγές: Για πολλές επιχειρήσεις, τα πληροφοριακά συστήματα που χρησιμοποιούν, αποτελούνται από πολλά υποσυστήματα, διαχωρισμένα και κατασκευασμένα σε διαφορετικές πλατφόρμες. Επίσης, η συγχώνευση των δεδομένων από πολλαπλές διαφορετικές πηγές δεδομένων είναι μια κοινή ανάγκη κατά την εφαρμογή της επιχειρηματικής ευφυΐας. Για να λύσει αυτό το πρόβλημα, η αποθήκη δεδομένων εκτελεί μια συγχώνευση των υφιστάμενων ανόμοιων πηγών δεδομένων και τις καθιστά προσβάσιμες από ένα μέρος. Η ενοποίηση των δεδομένων της επιχείρησης σε ένα ενιαίο αποθετήριο δεδομένων ελαφρύνει το βάρος της επαναληπτικής προσπάθειας συλλογής των ίδιων δεδομένων, και επιτρέπει την εξαγωγή των πληροφοριών που διαφορετικά θα ήταν αδύνατο. Επιπλέον, η αποθήκη δεδομένων θα θεωρείται ως η κοινή πηγή για την εξαγωγή της γνώσης για την επιχείρηση παρά από τις πολλαπλές γνώσεις που μπορεί να προκύψουν από την υποβολή εκθέσεων σε σχέση με τα επιμέρους υποσυστήματα.

Έγκαιρη πρόσβαση στα δεδομένα: Η αποθήκη δεδομένων επιτρέπει στα στελέχη των επιχειρήσεων και των υπεύθυνων λήψης αποφάσεων να έχουν πρόσβαση σε δεδομένα από πολλές διαφορετικές πηγές, διότι θα χρειαστούν να έχουν πρόσβαση στα δεδομένα. Οι χρήστες των επιχειρήσεων θα περάσουν λίγο χρόνο στη διαδικασία ανάκτησης των δεδομένων. Οι προγραμματισμένες ρουτίνες συγχώνευσης των δεδομένων, γνωστές και ως ETL (Extract, Transform, Load), είναι κομμάτι μέσα στο περιβάλλον αποθήκης δεδομένων. Αυτές οι ρουτίνες εδραιώνουν τα δεδομένα από πολλαπλά συστήματα πηγών και τα μετατρέπουν σε χρήσιμη μορφή. Στη συνέχεια, τα στελέχη μπορούν να έχουν πρόσβαση στα δεδομένα από ένα ενιαίο περιβάλλον. Επιπλέον, οι χρήστες των δεδομένων θα είναι σε θέση να αναζητούν δεδομένα απευθείας με λιγότερη υποστήριξη από ειδικευμένους χρήστες πληροφορικής. Ο χρόνος αναμονής για τους επαγγελματίες πληροφορικής να αναπτύξουν τις εκθέσεις και τα ερωτήματα μειώνεται σημαντικά καθώς στους χρήστες των επιχειρήσεων δίνεται η δυνατότητα να δημιουργήσουν αναφορές και ερωτήσεις αναφορικά με τις δικές τους ανάγκες. Η χρήση των ερωτημάτων και των εργαλείων ανάλυσης όπως επίσης η χρήση μιας συνεπής και εμπεριστατωμένης αποθήκης δεδομένων επιτρέπει στους χρήστες των επιχειρήσεων

να περνούν περισσότερο χρόνο στην εκτέλεση της ανάλυση δεδομένων και λιγότερο χρόνο στη συλλογή τους.

Βελτιωμένη ποιότητα και συνέπεια των δεδομένων: Η υλοποίηση μιας αποθήκης δεδομένων συνήθως περιλαμβάνει τη μετατροπή των δεδομένων από πολλαπλά συστήματα πηγών, αρχείων και ετερόκλητων στοιχείων σε κοινή μορφή. Τα δεδομένα από τις διαφορετικές επιχειρηματικές μονάδες και τμήματα είναι τυποποιημένες και η ασυνεπής φύση των δεδομένων από τα διαφορετικά συστήματα αφαιρείται. Επιπλέον, επιμέρους επιχειρηματικές μονάδες και υπηρεσίες, συμπεριλαμβανομένων των πωλήσεων, του μάρκετινγκ, του οικονομικού, και του λειτουργικού, θα αρχίσουν να χρησιμοποιούν το ίδιο αποθετήριο δεδομένων, όπως το σύστημα πηγής για τα μεμονωμένα ερωτήματα και τις εκθέσεις τους. Έτσι, κάθε μία από αυτές τις επιμέρους επιχειρηματικές μονάδες θα παράγουν αποτελέσματα που θα είναι συνεπείς σε σχέση με τις υπόλοιπες επιχειρηματικές μονάδες εντός του οργανισμού. Έτσι, η συνολική εμπιστοσύνη στα δεδομένα της επιχείρησης αυξάνεται σημαντικά.

Ευφυΐα στα ιστορικά δεδομένα: Οι αποθήκες δεδομένων συνήθως περιέχουν δεδομένα αξίας πολλών χρόνων που δεν μπορούν ούτε να αποθηκευτούν ούτε να αναλυθούν μέσα από ένα σύστημα συναλλαγών. Συνήθως τα συστήματα συναλλαγών ικανοποιούν τις περισσότερες απαιτήσεις υποβολής λειτουργικών εκθέσεων για ένα δεδομένο χρονικό διάστημα, αλλά χωρίς τη συμπερίληψη των ιστορικών δεδομένων. Σε αντίθεση, οι αποθήκες δεδομένων αποθηκεύουν μεγάλες ποσότητες ιστορικών δεδομένων και μπορεί να επιτρέψει την προηγμένη επιχειρηματική ευφυΐα συμπεριλαμβανομένης της ανάλυσης χρονικών περιόδων, της ανάλυση τάσεων και της πρόβλεψη τάσεων. Το πλεονέκτημα των αποθηκών δεδομένων είναι ότι επιτρέπει την προηγμένη αναφορά και ανάλυση από πολλαπλές χρονικές περιόδους.

Υψηλή απόδοση της επένδυσης (ROI): Η απόδοση της επένδυσης πρόκειται για την ποσότητα της αύξησης των εσόδων ή της μείωσης των εξόδων όπου μια επιχείρηση θα είναι σε θέση να εισπράξει από κάθε έργο ή επένδυση κεφαλαίου. Έτσι, η υλοποίηση των αποθηκών δεδομένων και των συμπληρωματικών συστημάτων επιχειρηματικής ευφυίας επέτρεψαν τις επιχειρήσεις να παράγουν υψηλότερα ποσά εσόδων και σημαντική εξοικονόμηση κόστους. Σύμφωνα με μια μελέτη του 2002 International Data Corporation (IDC) «*The Financial Impact of Business Analytics*»[17], τα πρότζεκτ αναλύσεων επέφεραν σημαντική θετική επίδραση στην οικονομική κατάσταση μιας επιχείρηση. Επιπλέον, η μελέτη έδειξε ότι οι εφαρμογές επιχειρηματικής ευφυίας έχουν επιφέρει ένα μέσο επιστροφής σε πέντε χρόνια από την επένδυση της τάξης του 112% με μέση αποπληρωμή τα 1.6 χρόνια. Από τις επιχειρήσεις που περιλαμβάνονταν στη μελέτη, το 54% είχαν μια απόδοση της επένδυσης της τάξης του 101% ή και περισσότερο.

3.3 Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών

Η Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών (Multiple Correspondence Analysis - MCA) είναι μια ειδική εφαρμογή της παραγοντικής ανάλυσης αντιστοιχιών για πίνακες με μια σειρά από κατηγορικές μεταβλητές. Ειδικότερα η MCA εφαρμόζεται σε πίνακες με ανεξάρτητες παρατηρήσεις στις γραμμές και κατηγορικές μεταβλητές στις στήλες.

Η ανάλυση μπορεί να γίνει από τη πλευρά των ανεξάρτητων παρατηρήσεων και έχει σαν στόχο την κατανόηση των ομοιοτήτων μεταξύ των παρατηρήσεων σε σχέση με το σύνολο των μεταβλητών, δημιουργώντας μια τυπολογία ή έναν χάρτη με τις ανεξάρτητες παρατηρήσεις όπου ξεχωρίζουν ποιες είναι οι πιο όμοιες μεταξύ τους και ποιες όχι. Μπορούμε επίσης να ανακαλύψουμε αν υπάρχουν ομάδες παρατηρήσεων οι οποίες να είναι συνεπείς όσον αφορά τις ομοιότητες τους.

Η MCA μπορεί να παρουσιαστεί με πολλούς τρόπους. Στη Γαλλία, μετά το έργο του Jean-Paul Benzécri[18], όπου παρουσίασε την πιο κοινή μέθοδο, την Παραγοντική Ανάλυση Αντιστοιχιών, μια μέθοδος σχεδιασμένη για να μελετήσει τη σχέση μεταξύ δύο ποιοτικών μεταβλητών. Με την προοπτική της ταυτόχρονης επεξεργασίας ποσοτικών και ποιοτικών μεταβλητών για τις ίδιες παρατηρήσεις, που είναι ένα από τα δυνατά σημεία της ανάλυσης πολλαπλών παραγόντων (MFA), είναι σημαντικό να εστιάσουμε στις ομοιότητες μεταξύ της Ανάλυσης Κύριων Συνιστωσών (PCA) και MCA. Έτσι λοιπόν όταν η ανάλυση εστιάζεται από τη πλευρά των ανεξάρτητων μεταβλητών, όπως και με την PCA, ο στόχος είναι να συνοψίσουμε τις σχέσεις μεταξύ τους. Ψάχνουμε λοιπόν για συνθετικές μεταβλητές που συνοψίζουν τις πληροφορίες που περιέχονται σε ένα αριθμό από μεταβλητές (με μέγιστο αριθμό τη διάσταση του πίνακα).

Σύμφωνα και με τον Husson F.[19] στην MCA, κυρίως επικεντρωνόμαστε στην μελέτη των κατηγοριών, όπου ως κατηγορίες αντιπροσωπεύονται και από τις μεταβλητές και από τις παρατηρήσεις που ανήκουν στην κατηγορία. Με άλλα λόγια είναι οι ποιοτικές τιμές που εμφανίζονται σε κάθε ζεύγος παρατήρησης και μεταβλητής.

3.3.1 Ορισμός Αποστάσεων

Στο σημείο αυτό και σύμφωνα με το [19] μπορούμε να αναφέρουμε τον τρόπο με τον οποίο υπολογίζονται η αποστάσεις και από την πλευρά των παρατηρήσεων και από των κατηγοριών.

Απόσταση μεταξύ των παρατηρήσεων:
Η απόσταση μεταξύ δύο παρατηρήσεων ορίζεται ως:

$$d^2_{i,i'} = C \sum_{k=1}^{K} \frac{(x_{ik} - x_{i'k})^2}{I_k} \tag{3.1}$$

όπου x_{ik} είναι η κατηγορία k της παρατήρησης i, I_k ο αριθμός των παρατηρήσεων που περιλαμβάνουν την κατηγορία k και C μια σταθερά.

Απόσταση μεταξύ των κατηγοριών:
Η απόσταση μεταξύ δύο κατηγοριών ορίζεται ως:

$$d^2_{k,k'} = C' \sum_{i=1}^{I} (\frac{x_{ik}}{I_k} - \frac{x_{ik'}}{I_{k'}})^2 \tag{3.2}$$

όπου x_{ik} είναι η κατηγορία k της παρατήρησης i, I_k ο αριθμός των παρατηρήσεων που περιλαμβάνουν την κατηγορία k και C' μια σταθερά.

3.3.2 Indicator Matrix

Η πιο συνηθισμένη μέθοδος που χρησιμοποιεί η Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών είναι ο Indicator Matrix, όπου με τη βοήθεια του απεικονίζονται γραφικά τα δεδομένα. Ο πίνακας αυτός συμβολίζεται με X και είναι διάστασης $N \times C$ όπου N το πλήθος των παρατηρήσεων και C ο αριθμός όλων των κατηγοριών.

Ουσιαστικά οι μεταβλητές στον πίνακα αυτό είναι ψευδομεταβλητές που δημιουργήθηκαν με βάση τις ποιοτικές μεταβλητές και τις κατηγορίες αυτών. Δηλαδή αν είχαμε 5 μεταβλητές με 4 δυνατές κατηγορίες η κάθε μία θα είχαμε στο σύνολο 20 ψευδομεταβλητές. Οι τιμές που παίρνει κάθε κελί του πίνακα αποτελείται από 1 ή 0 δηλώντας αν μία παρατήρηση διαθέτει ή όχι την αντίστοιχη κατηγορία μιας μεταβλητής. Αν M είναι το σύνολο των μεταβλητών τότε μια άλλη αναπαράσταση του X είναι η $X = [X_1, X_2, ..., X_m, ..., X_M]$ όπου για μία παρατήρηση i ο υποπίνακας X_m περιέχει την τιμή 0 $k_m - 1$ φορές, όπου k ο αριθμός των κατηγοριών της μεταβλητής m, και την τιμή 1 μία φορά στην κατηγορία που διαθέτει η παρατήρηση. Από αυτήν την αναπαράσταση μπορεί εύκολα να παρατηρηθεί ο Indicator Matrix είναι αρκετά αραιός (sparse) και αποτελείται από πολλά μηδενικά. Επίσης, επειδή ο πίνακας είναι δυαδικός μπορούμε να εφαρμόσουμε και μεθόδους ομαδοποίησης μεταξύ των παρατηρήσεων με μέτρα απόστασης για δυαδικές μεταβλητές. Στο Σχήμα 3.2 φαίνεται η δομή του Indicator Matrix.

Σύμφωνα με τον Husson F.[19] αποδεικνύεται λοιπόν ότι η Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών δεν είναι τίποτα άλλο από μια εφαρμογή της Απλής Παραγοντικής Ανάλυσης Αντιστοιχιών στον Indicator Matrix, διότι αν αναλύσουμε τον πίνακα σαν να ήταν ένας two-way frequency table, το αποτέλεσμα της ανάλυσης θα επέστρεφε τις συντεταγμένες των στηλών που θα

$$X = \begin{bmatrix} \overbrace{0\ 1\ 0}^{X_1} & \overbrace{1\ 0}^{X_2} & \cdots & \overbrace{1\ 0\ 0\ 0\ 0}^{X_M} \\ 1\ 0\ 0 & 0\ 1 & \cdots & 0\ 0\ 1\ 0\ 0 \\ & & \vdots & \\ 0\ 1\ 0 & 1\ 0 & \cdots & 0\ 0\ 0\ 0\ 1 \\ 0\ 0\ 1 & 0\ 1 & \cdots & 0\ 1\ 0\ 0\ 0 \end{bmatrix} \Bigg\} N$$
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{C}$$

Σχημα 3.2: *Ο Indicator Matrix διάστασης* $N \times C$.

μας επέτρεπαν να συσχετίσουμε τις κατηγορίες μεταξύ τους. Τα αποτελέσματα αυτά βασίζονται στις αποστάσεις μεταξύ των παρατηρήσεων ή των κατηγοριών χρησιμοποιώντας τις παρακάτω μετρικές αποστάσεις χ^2 αντίστοιχα:

- $d^2_{\chi^2}(row\ profile\ i,\ row\ profile\ i') = \sum_{k=1}^{K} \frac{1}{f_{\bullet k}} (\frac{f_{ik}}{f_{i\bullet}} - \frac{f_{i'k}}{f_{i'\bullet}})^2$
- $d^2_{\chi^2}(column\ profile\ k,\ column\ profile\ k') = \sum_{i=1}^{I} \frac{1}{f_{i\bullet}} (\frac{f_{ik}}{f_{\bullet k}} - \frac{f_{ik'}}{f_{\bullet k'}})^2$

όπου έχουμε $C = \frac{I}{J}$ (Εξίσωση 3.1) και

- $f_{ik} = \frac{x_i k}{IJ}$
- $f_{\bullet k} = \sum_{i=1}^{I} \frac{x_i k}{IJ} = \frac{I_k}{IJ}$
- $f_{i\bullet} = \sum_{k=1}^{K} \frac{x_i k}{IJ} = \frac{1}{I}$

Αξίζει να σημειωθεί ότι η χρήση του Indicator Matrix με την Απλή Παραγοντική Ανάλυση Αντιστοιχιών είναι ακριβή διαδικασία αν δεν χρησιμοποιηθούν κατάλληλες δομές που να αξιοποιούν τις ιδιαιτερότητες αυτού του πίνακα.

3.3.3 Πίνακας Burt

Μία ακόμα μέθοδος που χρησιμοποιεί η Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών είναι η εφαρμογή της Απλής Παραγοντικής Ανάλυσης Αντιστοιχιών στον πίνακα Burt. Ο πίνακας αυτός δεν είναι τίποτα άλλο από το εσωτερικό γινόμενο του Indicator Matrix με τον εαυτό του, δηλαδή $X^\mathsf{T}X$. Έχει διάσταση $C \times C$ όπου C το πλήθος όλων των κατηγοριών και στο Σχήμα 3.3 φαίνεται η δομή του.

Κάθε τιμή στο πίνακα Burt αποτελεί τη συχνότητα εμφάνισης μεταξύ δύο κατηγοριών. Με άλλα λόγια δείχνει το πόσες φορές εμφανίζονται ταυτόχρονα δύο κατηγορίες. Η εφαρμογή της Απλής Παραγοντικής Ανάλυσης Αντιστοιχιών σε αυτόν τον πίνακα χρησιμοποιείται για να αναπαραστήσει τις κατηγορίες. Δεδομένου ότι ο πίνακας αυτός είναι συμμετρικός, η αναπαράσταση του χάρτη των γραμμών είναι ταυτόσημη με εκείνη του χάρτη των στηλών. Ουσιαστικά με αυτόν τον τρόπο διαφαίνεται η γραμμικότητα των κύριων συνιστωσών της

$$
\begin{aligned}
B &= X^\top X \\
&= \begin{bmatrix} X_1^\top X_1 & X_1^\top X_2 & \cdots & X_1^\top X_M \\ X_2^\top X_1 & X_2^\top X_2 & \cdots & X_2^\top X_M \\ \vdots & \vdots & \ddots & \vdots \\ X_M^\top X_1 & X_M^\top X_2 & \cdots & X_M^\top X_M \end{bmatrix} \\
&= \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1C} \\ B_{21} & B_{22} & \cdots & B_{2C} \\ \vdots & \vdots & \ddots & \vdots \\ B_{C1} & B_{C2} & \cdots & B_{CC} \end{bmatrix}
\end{aligned}
\tag{3.3}
$$

Σχημα 3.3: *Ο πίνακας Burt διάστασης $C \times C$.*

ίδιας τάξης. Η αναπαράσταση με χρήση του πίνακα Burt ενδείκνυται σε μεγάλο βαθμό όταν ο αριθμός των παρατηρήσεων είναι αρκετά μεγάλος, διότι το μέγεθος αποθήκευσης του πίνακα είναι αρκετά μικρότερο σε σχέση με εκείνο του Indicator Matrix και με τον πίνακα αυτόν μπορούμε να αποθηκεύσουμε την ίδια πληροφορία σχετικά με την συσχέτιση από την πλευρά των κατηγοριών.

3.3.4 Αδράνεια

Ο όρος Αδράνεια (Inertia) προέρχεται από τη Μηχανική, συμβολίζεται συνήθως με I ή με ϕ^2. Είναι ένα συνολικό μέτρο ανομοιογένειας μεταξύ των προφίλ/συνιστωσών. Δείχνει με άλλα λόγια το ποσό της διαφοράς μεταξύ των συνιστωσών, μετρώντας τη διαφορά ανάμεσα σε κάθε ζευγάρι σημείων. Η αδράνεια μπορεί επίσης να θεωρηθεί και ως τη Διακύμανση (Variance) που προσφέρει κάθε προφίλ. Αυτοί οι δύο όροι πολλές φορές χρησιμοποιούνται ταυτόσημα στη βιβλιογραφία. Καθώς η αδράνεια αυξάνεται γίνεται πιο έντονη και η διαφοροποίηση μεταξύ των γραμμών. Η αδράνεια δίνεται από τον παρακάτω τύπο:

$$I = \phi^2 = \frac{\chi^2}{n} \tag{3.4}$$

όπου χ^2 είναι η απόσταση chi-square και δίνεται από τον τύπο:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} \tag{3.5}$$

Αδράνεια στον Indicator Matrix:

Η συνολική αδράνεια σε έναν Indicator Matrix έχει μια ιδιαίτερα απλή μορφή. Βασίζεται μόνο στον αριθμό των μεταβλητών και των κατηγοριών και όχι στα ίδια τα δεδομένα. Ας υποθέσουμε ότι έχουμε M μεταβλητές και κάθε μία μεταβλητή m έχει C_m κατηγορίες, όπου C είναι ο συνολικός αριθμός

των κατηγοριών. Αν συμβολίσουμε τον Indicator Matrix με X και με C τον αριθμό στηλών όπου αποτελείται από X_M υποπίνακες στοιβαγμένους τον ένα δίπλα στον άλλο τότε ορίζουμε σαν συνολική αδράνεια αυτού του πίνακα την μέση αδράνεια των υποπινάκων. Κάθε υποπίνακας X_m αποτελείται από έναν άσσο σε κάθε γραμμή και σε κάθε άλλη περίπτωση μηδέν. Αποδεικνύεται οι αδράνειες είναι ίσες με 1 σε κάθε κύρια συνιστώσα του υποπίνακα και ότι η συνολική αδράνεια του υποπίνακα X_m είναι ίση με τη διάσταση του η οποία είναι $C_m - 1$. Συνεπάγεται λοιπόν ότι η αδράνεια του πίνακα X ισούται με:

$$I_X = \frac{1}{M} \sum_m I_{X_m} = \frac{1}{M} \sum_m (C_m - 1) = \frac{C - M}{M} \tag{3.6}$$

Καθώς ο αριθμός $C - M$ είναι η διαστασιμότητα του πίνακα X, η μέση αδράνεια κάθε διάστασης είναι $\frac{1}{M}$. Αυτή η τιμή εξυπηρετεί σαν ένα κάτω όριο για την απόφαση ποια συνιστώσα αξίζει να ερμηνευτεί στην MCA.

Αδράνεια στον πίνακα Burt:

Οι υποπίνακες του πίνακα Burt έχουν τα ίδια περιθώρια γραμμής σε κάθε ένα σύνολο των οριζόντιων πινάκων και τα ίδια περιθώρια στήλης σε κάθε ένα σύνολο των κάθετων πινάκων. Συνεπώς η αδράνεια υπολογίζεται το ίδιο. Δηλαδή η αδράνεια του πίνακα B ορίζεται ως η μέση αδράνεια των πινάκων $B_{M_i M_j}$. Η αδράνεια ενός πίνακα με τον εαυτό του υπολογίζεται με τον ίδιο ορισμό όπως και στην Εξίσωση 3.6 του indicator matrix - πρόκειται για πίνακες $C_m \times C_m$ διάστασης $C_m - 1$ με τέλεια συσχέτιση μεταξύ των γραμμών-στηλών και έτσι έχει μέγιστη αδράνεια ίση με τον αριθμό των διαστάσεων. Οι τιμές της αδράνειας των πινάκων μεταξύ τους είναι αρκετά μεγαλύτερες από τις υπόλοιπες τιμές. Για τον λόγο αυτό η συνολική αδράνεια του πίνακα Burt είναι αρκετά μεγάλη ενώ ταυτόχρονα οι επιμέρους αδράνειες μεταξύ των συνιστωσών έχουν χαμηλά ποσοστά.

3.3.5 Μέθοδοι Αποσύνθεσης

Για να απεικονίσει τη σύνδεση μεταξύ των γραμμών και των στηλών ενός Indicator Matrix, κάθε σειρά και στήλη πρέπει να ποσοτικοποιηθεί με κάποια βαθμολογία. Οι βαθμολογίες των σειρών συχνά αναφέρονται και ως component scores, ενώ οι βαθμολογίες των στηλών ονομάζονται component loadings. Η ορολογία αυτή είναι δανεισμένη από την Ανάλυση Κύριων Συνιστωσών όπου εκεί προαναφέρθηκαν[20].

Για να αποκτηθεί βέλτιστη ποσοτικοποίηση των κατηγοριών ή με άλλα λόγια να βρεθεί η μέγιστη συσχέτιση χρησιμοποιώντας όσο το δυνατόν λιγότερες διαστάσεις μπορούν να χρησιμοποιηθούν οι δύο παρακάτω τρόποι:

1. Η Αποσύνθεση Ιδιαζουσών Τιμών (Singular Value Decomposition - SVD) στον indicator matrix X διάστασης $N \times C$.
2. Η Ιδιοαποσύνθεση (Eigendecomposition) του πίνακα Burt B διάστασης $C \times C$ καθώς ο πίνακας είναι υπερ-διαγώνιος.

Αποσύνθεση Ιδιαζουσών Τιμών (Singular Value Decomposition - SVD)

Στην πρώτη μέθοδο για την Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών, τα δεδομένα συνοψίζονται με τη χρήση του Indicator Matrix[21]. Δεδομένου ότι εφαρμόζουμε την μέθοδο στον Indicator Matrix, οι γραμμές αντιπροσωπεύουν στις παρατηρήσεις και οι στήλες στις κατηγορίες των διάφορων μεταβλητών. Η γραφική απεικόνιση των σημείων γραμμής και στήλης σε έναν χαμηλής διάστασης χώρο λαμβάνει υπόψιν διαφορετικά συστήματα βάρους για τις παρατηρήσεις και τις μεταβλητές. Συνήθως, τα βάρη των παρατηρήσεων είναι ομοιόμορφα ενώ τα βάρη των μεταβλητών εξαρτώνται από τις οριακές συχνότητες των στηλών του πίνακα X. Δεδομένου ότι ο πίνακας X είναι διάστασης $N \times C$, ο υποχώρος θα αποτελείται το πολύ από $M^* = \min(N, C)$ διαστάσεις. Σύμφωνα λοιπόν με τον Greenacre θα εφαρμόσουμε SVD στον πίνακα:

$$\frac{1}{p\sqrt{N}} X D^{-1/2} \tag{3.7}$$

δηλαδή:

$$SVD\left(\frac{1}{p\sqrt{N}} X D^{-1/2}\right) = U S_X V^{\mathsf{T}} \tag{3.8}$$

όπου $S_X = diag(\lambda_m^X)$, για $m = 1, 2, \ldots, \min(N, C - p)$, είναι ο πίνακας με τις ιδιοτιμές σε φθίνουσα διάταξη. Οι πίνακες U και V περιέχουν τα αριστερά και δεξιά ιδιοδιανύσματα αντίστοιχα και υπακούν στους παρακάτω περιορισμούς:

$$U^{\mathsf{T}} U = I_{M^*} \tag{3.9}$$

και

$$(V^{\mathsf{T}} D^{-1/2}) D (D^{-1/2}) V = V^{*\mathsf{T}} D V^* = I_{M^*} \tag{3.10}$$

Ιδιοαποσύνθεση

Στην πρώτη μέθοδο για την Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών, τα δεδομένα συνοψίζονται στη μορφή του πίνακα Burt[22] B. Ο πίνακας κατασκευάζεται έτσι ώστε οι γραμμές και οι στήλες του να αποτελούνται από όλες τις κατηγορίες των μεταβλητών. Συνεπώς, θα αποτελεί έναν υπερ-διαγώνιο πίνακας όπου οι διαγώνιοι υποπίνακες είναι διαγώνιοι και περιέχουν το μέγιστο ποσοστό συμμετοχής κάθε κατηγορίας για κάθε μεταβλητή. Ομοίως, οι μη διαγώνιοι υποπίνακες θα αποτελούν τον αμφίδρομο πίνακα συσχέτισης (ή τον ανάστροφο του) για κάθε ζεύγος κατηγοριών (δες Παράγραφο 3.3.3). Ως εκ τούτου, ο πίνακας Burt μπορεί να υπολογιστεί ως:

$$B = X^{\mathsf{T}} X$$

Κάθε μη διαγώνιος υποπίνακας $X_m^{\mathsf{T}} X'_m$, για $m \neq m'$ είναι ένας αμφίδρομος πίνακας συσχέτισης μεταξύ των μεταβλητών m και m'. Ο χαμηλής διάστασης χώρος που χρησιμοποιείται για τη γραφική απεικόνιση της συσχέτισης μεταξύ

των κατηγοριών, χρησιμοποιώντας τον πίνακα Burt αποτελείται το μέγιστο από $\max(C, C) = C$ διαστάσεις. Επομένως, η εφαρμογή του MCA σε έναν πολυεπίπεδο πίνακας συσχετίσεων χρησιμοποιώντας τον πίνακα Burt μπορεί να γίνει με την εκτέλεση μιας Ιδιοαποσύνθεσης στο μετασχηματισμένο πίνακα Burt:

$$\frac{1}{p^2 N} D^{-1} B \tag{3.11}$$

έτσι ώστε

$$ED\left(\frac{1}{p^2 N} D^{-1} B\right) = Q \Lambda_B Q^{\mathsf{T}} \tag{3.12}$$

όπου Λ_B είναι ο διαγώνιος πίνακας με τις ιδιοτιμές λ_m^B όπου $m = 1, 2, \ldots, C-p$, έτσι ώστε $\Lambda_B = diag(\lambda_m^B)$. Επιπλέον, ο πίνακα Q περιέχει τα ιδιοδυανύσματα του πίνακα (3.11).

Ο Greenacre και ο Lebart απέδειξαν ότι η σχέση μεταξύ της ιδιοτιμής m του Indicator Matrix - που ορίζεται έτσι λ_m^X - και η ιδιοτιμή m του πίνακα Burt - που ορίζεται έτσι λ_m^B - δίνεται από τη σχέση:

$$\lambda_m^B = (\lambda_m^X)^2 \tag{3.13}$$

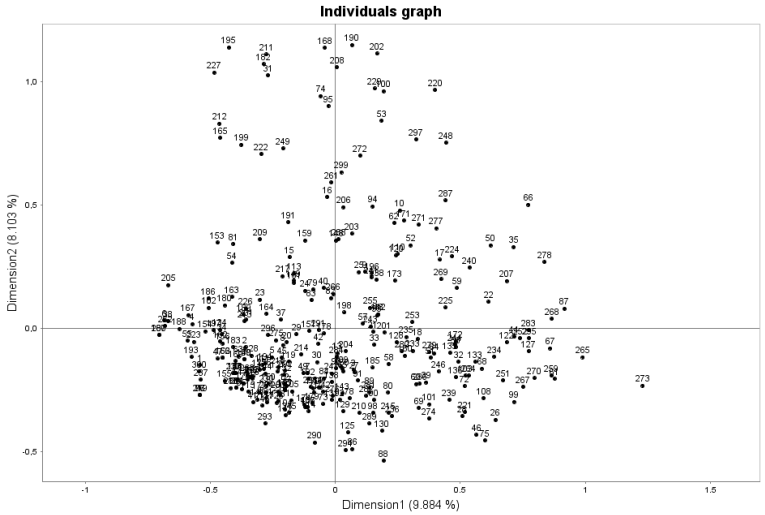
Συνεπώς, η Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών στον πίνακα Burt δίνει ιδιοτιμές οι οποίες είναι ίσες με το τετράγωνο της τιμής των ιδιοτιμών της ανάλυσης στον Indicator Matrix.

3.3.6 Απεικόνιση Αποτελεσμάτων

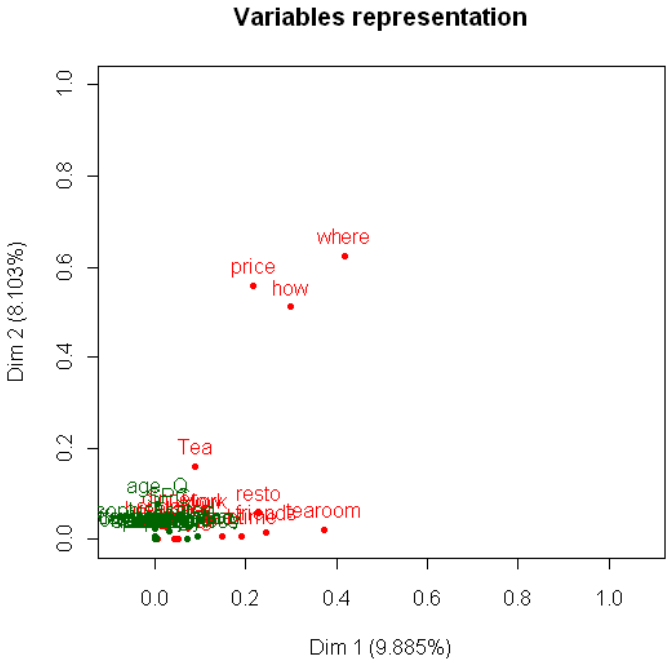
Για την απεικόνιση των αποτελεσμάτων της ανάλυσης MCA χρησιμοποιούνται οι τρεις βασικοί τύποι γραφημάτων συνήθως στις δύο πρώτες διαστάσεις που περιγράφουν την περισσότερη πληροφορία. Σύμφωνα λοιπόν και με το [19] μπορούν να απεικονισθούν με:

Χάρτης των παρατηρήσεων:
Ο χάρτης των παρατηρήσεων αναπαρίσταται χρησιμοποιώντας τις κύριες συνιστώσες (όπως στην PCA) και μεγιστοποιώντας την αδράνεια χάρτη των κατηγοριών προβάλλοντας τις παρατηρήσεις σε μια σειρά από ορθογώνιους άξονες. Αυτή η διερευνητική προσέγγιση μπορεί να είναι κουραστική λόγω του μεγάλου αριθμού των παρατηρήσεων, και γενικεύεται με τη μελέτη των κατηγοριών μέσω των παρατηρήσεων που εκπροσωπούν.

Χάρτης των μεταβλητών:
Σε ένα χάρτη μεταβλητών, οι μεταβλητές μπορούν να αναπαριστώνται από τον υπολογισμό των συσχετίσεων μεταξύ των συντεταγμένων των παρατηρήσεων ενός αντικειμένου και κάθε ποιοτικής μεταβλητής. Αν η συσχέτιση μεταξύ της μεταβλητής j και του αντικειμένου s είναι κοντά στο 1, οι παρατηρήσεις που φέρουν την ίδια κατηγορία (για αυτή την



Σχημα 3.4: *Παράδειγμα χάρτη παρατηρήσεων.*

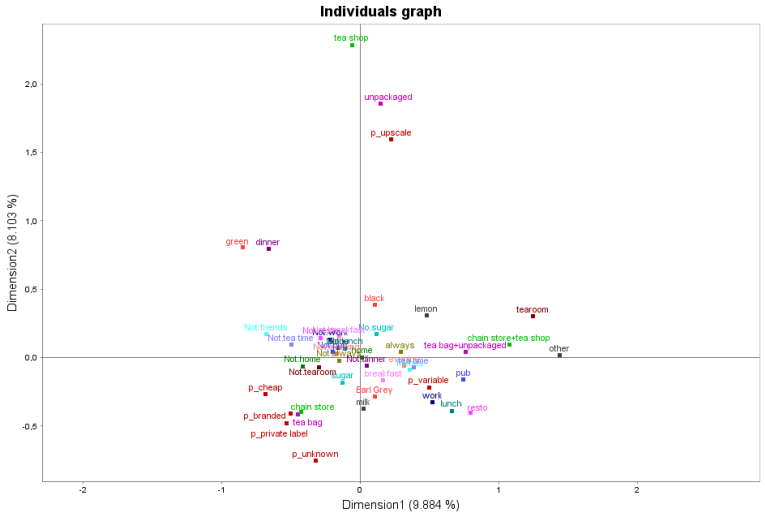


Σχημα 3.5: *Παράδειγμα χάρτη μεταβλητών.*

ποιοτική μεταβλητή) έχουν παρόμοιες συντεταγμένες για το αντικείμενο s.

Χάρτης των κατηγοριών:

Σε ένα χάρτη κατηγοριών, οι κατηγορίες μπορούν να αναπαριστώνται στο κέντρο βάρους των παρατηρήσεων που έχουν αυτές τις κατηγορίες. Αυτή αναπαράσταση είναι βέλτιστη σε σχέση με τις υπόλοιπες καθώς αντιστοιχεί στην αναπαράσταση που λαμβάνεται με τη μεγιστοποίηση της αδράνειας του χάρτη κατηγοριών σε μια ακολουθία από ορθογώνιους άξονες με τη χρήση του μετασχηματισμού της ομοιοθεσίας.



Σχημα 3.6: *Παράδειγμα χάρτη κατηγοριών.*

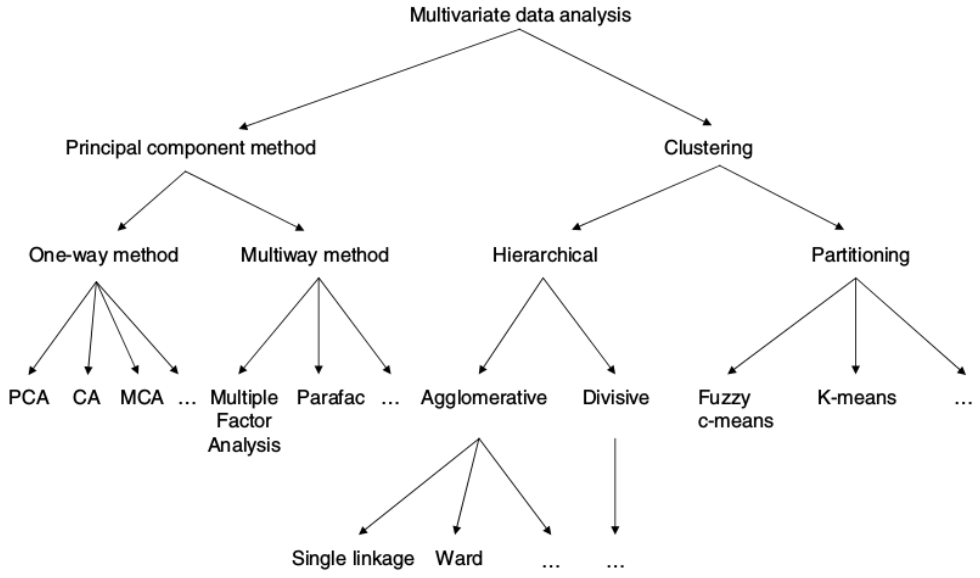
3.4 Ιεραρχική Ομαδοποίηση Στην MCA

3.4.1 Εισαγωγή

Η ομαδοποίηση ή αλλιώς ο διαχωρισμός των δεδομένων σε τμήματα έχει ένα πλήθος στόχων. Όλοι αφορούν την ομαδοποίηση ή κατάτμηση μιας συλλογής από αντικείμενα σε υποσύνολα ή «ομάδες», όπου εκείνα τα αντικείμενα μέσα σε κάθε ομάδα είναι πιο στενά συνδεδεμένα ή με άλλα λόγια μοιράζονται περισσότερα κοινά χαρακτηριστικά σε σχέση με τα αντικείμενα που βρίσκονται στις υπόλοιπες ομάδες. Ένα αντικείμενο μπορεί να περιγράφεται από ένα σύνολο μετρικών ή με τη σχέση του με άλλα αντικείμενα. Επιπλέον, ο στόχος πολλές φορές μπορεί να είναι η ομαδοποίηση των αντικειμένων με βάση μια φυσική ιεραρχία. Αυτό περιλαμβάνει την διαδοχική ομαδοποίηση των ίδιων των ομάδων, έτσι ώστε σε κάθε επίπεδο της ιεραρχίας, οι ομάδες μέσα στην ίδια ομάδα να είναι πιο όμοιες μεταξύ τους από ότι εκείνες στις διαφορετικές ομάδες.

Η ομαδοποίηση χρησιμοποιείται επίσης για να σχηματίσουν περιγραφικά στατιστικά ώστε να εξακριβωθεί αν τα δεδομένα αποτελούνται από ένα σύνολο διακριτών υποομάδων, κάθε ομάδα που αντιπροσωπεύεται από αντικείμενα με σημαντικά διαφορετικές ιδιότητες. Αυτός ο στόχος απαιτεί την εκτίμηση του βαθμού της διαφοράς μεταξύ των αντικειμένων που υπάρχουν στις αντίστοιχες ομάδες.

Κεντρικό ρόλο σε όλους τους στόχους της ομαδοποίησης είναι η έννοια του βαθμού ομοιότητα (ή ανομοιότητας) μεταξύ των επιμέρους αντικειμένων που ομαδοποιούνται. Μια μέθοδος ομαδοποίησης επιχειρεί να ομαδοποιήσει τα αντικείμενα με βάση τον ορισμό ομοιότητας που παρέχεται σε αυτό. Αυτό μπορεί να προέλθει μόνο την εκτίμηση που κάνει ο αναλυτής σχετικά με τις ομάδες που περιμένει να βρει.



Σχημα 3.7*: Ιεραρχικό δέντρο που δηλώνει τις σχέσεις μεταξύ των μεθόδων ανάλυσης κύριων συνιστωσών και ομαδοποίησης[19].*

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο, οι μέθοδοι ανάλυσης σε κύριες συνιστώσες όπως η MCA, αποδίδουν αναπαραστάσεις στον Ευκλείδειο χώρο. Επίσης μία μέθοδος αναπαράστασης των δεδομένων είναι με την μορφή ενός ιεραρχικού δέντρου. Αυτό μπορεί να επιτευχθεί αν υιοθετήσουμε την προσέγγιση που χρησιμοποιείται στις μεθόδους ανάλυσης σε κύριες συνιστώσες, και πιο συγκεκριμένα, στην ανάλυση ενός πίνακα δεδομένων, χωρίς να έχει παρθεί κάποια προηγούμενη απόφαση παρά μόνο των παραδοχών που έγιναν για τη κατασκευή του πίνακα για την επιλογή των παρατηρήσεων και των μεταβλητών. Ο στόχος είναι η κατασκευή ενός ιεραρχικού δέντρου και όχι ένας χάρτης κυρίων συνιστωσών, για να απεικονίσει τις σχέσεις μεταξύ των αντικειμένων. Πρόκειται δηλαδή για ένα εργαλείο μελέτης της μεταβλητότητας εντός του πίνακα. Η διαφορά με τις μεθόδους ανάλυσης σε κύριες συνιστώσες ουσιαστικά είναι ο τρόπος που αναπαριστώνται.

Ο αλγόριθμος που χρησιμοποιείται για την κατασκευή αυτών των δέντρων ονομάζεται ιεραρχική ομαδοποίηση. Υπάρχουν δύο βασικές στρατηγικές που μπορεί να κινηθεί αυτός ο αλγόριθμος οι οποίες είναι:

Συσσωρευτικός (Agglomerative): Αρχίζει με τα σημεία ως ξεχωριστές ομάδες και σε κάθε βήμα, συγχωνεύει το πιο κοντινό ζευγάρι ομάδων μέχρι να μείνει μόνο μία ομάδα.

Διαιρετικός (Divisive): Αρχίζει με μία ομάδα που περιέχει όλα τα σημεία και σε κάθε βήμα, διαχωρίζει μία ομάδα, έως ότου κάθε ομάδα να περιέχει μόνο ένα σημείο.

Η συσσωρευτική ιεραρχική ομαδοποίηση είναι αυτή που χρησιμοποιείται πιο συχνά και κυρίως μέσω του αλγορίθμου Ward. Στο σχήμα 3.7 φαίνεται η σχέση μεταξύ των μεθόδων ανάλυσης κύριων συνιστωσών και ομαδοποίησης.

3.4.2 Η MCA σαν προ-επεξεργασία της ομαδοποίησης

Ο κύριος στόχος της Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών αλλά και όλων των μεθόδων ανάλυσης σε πρωτεύοντες συνιστώσες είναι να περιγράψουν έναν πίνακα δεδομένων X με N παρατηρήσεις και M μεταβλητές, χρησιμοποιώντας έναν μικρότερο αριθμό S, όπου $S < M$, από ασυσχέτιστες μεταξύ τους μεταβλητές διατηρώντας παράλληλα όση περισσότερη πληροφορία γίνεται. Αυτό επιτυγχάνεται μετατρέποντας τα κατηγορικά δεδομένα μας σε συνεχείς μεταβλητές ή αλλιώς πρωτεύοντες συνιστώσες.

Η ομαδοποίηση σε κατηγορικές μεταβλητές είναι ένα θέμα που έχει απασχολήσει αρκετά την επιστημονική κοινότητα. Στη βιβλιογραφία υπάρχουν πολλά μέτρα ομοιότητας για τέτοιες περιπτώσεις όπως ο δείκτης Jaccard, ο συντελεστή Sørensen–Dice, ο συντελεστής απλής αντιστοίχισης κ.α. Το πρόβλημα με αυτές τις μετρικές είναι ότι είναι σχεδιασμένες για να δουλεύουν αρκετά καλά με δυαδικά δεδομένα. Όταν όμως οι κατηγορίες των μεταβλητών είναι περισσότερες από δύο τότε είναι σύνηθες να χρησιμοποιείται η απόσταση χ^2. Χρησιμοποιώντας αυτήν την μετρική για ομαδοποίηση είναι ισάξιο με την εφαρμογή της ομαδοποίησης σε όλες τις κύριες συνιστώσες που παράγονται από την MCA. Συνεπώς, η εφαρμογή της MCA στα δεδομένα μπορεί να θεωρηθεί ως ένας τρόπος μετασχηματισμού των κατηγορικών μεταβλητών σε συνεχείς. Μία συνήθης πρακτική είναι να εφαρμόζεται η ομαδοποίηση στις πρώτες κύριες συνιστώσες της MCA.

3.4.3 Αδράνεια Ανάμεσα και Μέσα στις ομάδες

Η αδράνεια ανάμεσα στις ομάδες ονομάζεται η αδράνεια η οποία μετράει τη διακύμανση μεταξύ των διαφορετικών ομάδων γραμμής σε έναν πίνακα δεδομένων. Η διαφορά ανάμεσα στην ολική αδράνεια και στην αδράνεια ανάμεσα στις ομάδες ονομάζεται αδράνεια μέσα στις ομάδες και ουσιαστικά μετράει τη διακύμανση μεταξύ των διαφορετικών ομάδων η οποία χάθηκε κατά την συγχώνευση των γραμμών σε ομάδες. Αυτή η αποσύνθεση της αδράνεια είναι ένα κλασσικό αποτέλεσμα της Ανάλυση Διακύμανσης (ANOVA), συνήθως χρησιμοποιούμενη για μία μεταβλητή, αλλά εξίσου εφαρμόσιμη και για πολυμεταβλητά δεδομένα. Η εξίσωση της αποσύνθεσης είναι η εξής:

$$Total\ inertia = Between\ inertia + Within\ inertia$$

$$\sum_{c=1}^{C}\sum_{q=1}^{Q}\sum_{n=1}^{N_q}(x_{nqc} - \bar{x}_c)^2 = \sum_{c=1}^{C}\sum_{q=1}^{Q} N_q(\bar{x}_{qc} - \bar{x}_c)^2 + \sum_{c=1}^{C}\sum_{q=1}^{Q}\sum_{n=1}^{N_q}(x_{nqc} - \bar{x}_{qc})^2$$

Γενικά η ποιότητα μιας ομάδας διακρίνεται όταν:

- Οι παρατηρήσεις μιας ομάδα διαφέρουν αισθητά από τις παρατηρήσεις μιας άλλης ομάδας (δηλ. έχει μεγάλη τιμή αδράνειας ανάμεσα στις ομάδες).
- Οι παρατηρήσεις μιας ομάδα είναι ομοιογενώς κατανεμημένες (δηλ. έχει μικρή τιμή αδράνειας μέσα στην ομάδα).

Η επιλογή βελτιστοποίησης μίας εκ των δύο τιμών (αδράνεια ανάμεσα ή μέσα στις ομάδες) για την εξασφάλιση και την μέτρηση ποιότητας των ομάδων δεν είναι πάντα εύκολη ή αυτονόητη επιλογή. Μία λύση είναι να χρησιμοποιηθεί το ποσοστό:

$$\frac{Between - groups\ inertia}{Total\ inertia} * 100\% \tag{3.14}$$

όπου δηλώνει το πόσο αυτή η ομάδα συμμετέχει σε σχέση με την συνολική μεταβλητότητα των παρατηρήσεων. Ο αριθμός των ομάδων παίζει καθοριστικό ρόλο διότι όσο περισσότερες ομάδες επιλέγονται τόσο το ποσοστό προσεγγίζει το 100%. Παρόλα αυτά τέτοιες ομάδες δεν έχουν κάποια πραγματική αξία. Χρειάζεται να υπάρχει σωστή ισορροπία ανάμεσα σε αυτά τα δύο.

3.4.4 Μέθοδος Ward

Η μέθοδος του Ward πρόκειται για ένα κριτήριο που εφαρμόζεται στον Ιεραρχικό Αλγόριθμο ομαδοποίησης. Πρωτοπαρουσιάστηκε από τον Joe H. Ward, Jr[23] το 1993 και ουσιαστικά πρότεινε μία γενική μέθοδο συσσωρευτικής ιεραρχικής ομαδοποίησης, όπου το κριτήριο επιλογής των ομάδων από ζευγάρια για συγχώνευση σε κάθε βήμα βασίζεται σε στην βέλτιστη τιμή μίας αντικειμενικής συνάρτησης. Στην περίπτωσή μας, ο ιεραρχικός αλγόριθμος εφαρμόζεται στις κύριες συνιστώσες και τον ρόλο της αντικειμενικής συνάρτησης παίζει η αδράνεια μέσα στην ομάδα, όπου αναζητούμαι την ελάχιστη τιμής σε κάθε βήμα του αλγορίθμου. Με άλλα λόγια αναζητούμε την ελαχιστοποίηση της μείωσης της αδράνειας μεταξύ των ομάδων. Η αδράνεια μέσα σε μια ομάδα πρακτικά δηλώνει την ομοιογένεια των ομάδων. Η κατάταξη ιεραρχικά αναπαριστάται με ένα δενδρόγραμμα το οποίο βαθμονομείται από το κέρδος της αδράνεια μέσα στις ομάδες.

Σύμφωνα λοιπόν, όπως αναφέρει ο Husson στο [19], αν υποθέσουμε μία ομάδα p με κέντρο βάρους g_p και μέγεθος δείγματος N_p και την ομάδα q με κέντρο βάρους g_q και μέγεθος δείγματος N_q τότε η αύξηση $\Delta(p,q)$ στην αδράνεια μέσα στην ομάδα που προκλήθηκε από τη συγχώνευση των δύο ομάδων p και q μπορεί να εκφραστεί ως:

$$\Delta(p,q) = \frac{N_p N_q}{N_p + N_q} d^2(g_p, g_q) \tag{3.15}$$

Με άλλα λόγια, η επιλογή δύο ομάδων p και q που ελαχιστοποιούν το $\Delta(p,q)$ σημαίνει, η επιλογή ομάδων όπου τα κέντρα τους είναι κοντά (δηλ. η απόσταση $d^2(g_p, g_q)$ είναι μικρή) ή ομάδων όπου το μέγεθός τους είναι μικρό (δηλ. η τιμή $\frac{N_p N_q}{N_p + N_q}$ είναι μικρή).

Η επιλογή του αριθμού των ομάδων αποτελεί βασικό θέμα κάθε αλγορίθμου ομαδοποίησης και πάνω σε αυτό έχουν προταθεί διάφορες προσεγγίσεις. Μερικές από αυτές βασίζονται καθαρά στο δενδρόγραμμα ή ιεραρχικό δέντρο και αυτό διότι ένα ιεραρχικό δέντρο μπορεί να θεωρηθεί ως μια ακολουθία από

εμφωλευμένες ομάδες, δηλαδή από το σημείο όπου κάθε παρατήρηση είναι μια ομάδα από μόνη της μέχρι το σημείο όπου όλες οι παρατηρήσεις είναι μια ομάδα μαζί. Ο αριθμός των ομάδων μπορεί στη συνέχεια να επιλέγεται κοιτάζοντας τη συνολική απεικόνιση του δέντρου, το ιστόγραμμα του κέρδους της αδράνειας μέσα στις ομάδες, κλπ. Αυτοί οι κανόνες εφαρμόζονται συνήθως έμμεσα ή άμεσα στην αύξηση της αδράνειας. Συνοπτικά οι κανόνες επιλογής που πρέπει να ληφθούν υπόψιν συνοψίζονται παρακάτω[19]:

1. Η όψη του ιεραρχικού δέντρου.
2. Ο αριθμός των ομάδων να μην είναι πολύ μεγάλος αλλιώς χάνεται η αξία της ομαδοποίησης.
3. Ο αριθμός των παρατηρήσεων που υπάρχουν σε κάθε διαχωρισμό. Αν η μείωση είναι αισθητή σε έναν διαχωρισμό τότε είναι καλό να υπάρξει και ένας επόμενος. Συνήθως αναπαριστάται με ένα ιστόγραμμα των διαχωρισμών σε κάθε βήμα.
4. Η ερμηνεία των ομάδων. Αν η ομάδες που θα παραχθούν έχουν αρκετά καλή αδράνεια μεταξύ τους αλλά παρόλα αυτά δεν μπορούν να ερμηνευτούν και τότε δεν έχει κάποια αξία η ανάλυση. Είναι προτιμητέο να γίνει θυσιαστεί κάποια από την αδράνεια και να δημιουργηθούν περισσότερες πιο κατανοητές ομάδες.

Σύμφωνα λοιπόν με τον κανόνα 3. μπορεί να εφαρμοστεί ένα κριτήριο για την αυτόματη επιλογή του αριθμού των ομάδων. Δηλαδή, η διαίρεση σε Q ομάδες όταν η αύξηση της αδράνειας μεταξύ των ομάδων $Q-1$ και Q είναι αρκετά μεγαλύτερη από αυτήν των ομάδων Q και $Q+1$. Ένα εμπειρικό κριτήριο για να επιτευχθεί αυτό είναι το παρακάτω:

$$\min_{Q_{\min} \leq Q \leq Q_{\max}} \frac{\Delta(Q)}{\Delta(Q+1)}$$

όπου $\Delta(Q)$ είναι η διαφορά της αδράνεια μεταξύ των ομάδων $Q-1$ και Q όπως επίσης $Q_{\min}$ και $Q_{\max}$ είναι τα όρια του αριθμού των ομάδων που επιθυμεί ο χρήστης. Ο αριθμός Q όπου ελαχιστοποιεί το κριτηρίου επιστρέφεται.

3.4.5 K-means πριν και μετά την Ιεραρχική Ομαδοποίηση

Υπάρχουν διαθέσιμες αρκετές τεχνικές ομαδοποίησης. Ίσως η πιο απλή από αυτές αποτελεί η επιλογή Q ομάδων όπως αυτές ορίζονται από το ιεραρχικό δέντρο που αναφέραμε στην προηγούμενη παράγραφο. Μία δεύτερη στρατηγική που θα μπορούσε κάποιος να ακολουθήσει είναι να εφαρμόσει τον αλγόριθμο K-means με τον αριθμό των ομάδων προκαθορισμό σε Q. Ακόμα μία στρατηγική θα μπορούσε να συνδυάζει τις δύο αυτές τεχνικές, δηλαδή οι ομάδες που απομονώθηκαν από την τομή του δέντρου να εισαχθούν στον αλγόριθμο K-means σαν αρχική κατάσταση και να συνεχιστούν πολλές επαναλήψεις σε αυτόν. Η τμηματοποίηση που θα προκύψει είναι η τελική που θα κρατηθεί.

Συνήθως η τμηματοποίηση που ορίζεται σαν αρχική κατάσταση στον K-means δεν είναι ποτέ τελείως διαφορετική από την τελική αλλά είναι μια βελτίωση της.

Στην περίπτωση μας η τεχνική που μας ενδιαφέρει σύμφωνα με τον Husson ανάγεται στη συγχώνευση των δύο τεχνικών με την αντίστροφη σειρά από ότι περιγράφηκε πιο πάνω. Δηλαδή:

Φάση $1^η$**:** Ομαδοποιούμε με μεγάλο αριθμό ομάδων χρησιμοποιώντας τον αλγόριθμο K-means. Οι ομάδες που θα εξαχθούν συνήθως δεν μπορούν να ερμηνευτούν αμέσως διότι οι ομάδες είναι πάρα πολλές και πολλές από αυτές θα βρίσκονται αρκετά κοντά μεταξύ τους. Παρόλα αυτά έχουν μικρή τιμή αδράνειας μέσα στις ομάδες και οι παρατηρήσεις που ομαδοποιήθηκαν μαζί είναι σίγουρο ότι δεν θα έπρεπε να διαχωριστούν.

Φάση $2^η$**:** Στη συνέχεια εφαρμόζουμε τον ιεραρχικό αλγόριθμο όπως αυτός παρουσιάστηκε στην προηγούμενη παράγραφο χρησιμοποιώντας τις ομάδες του προηγούμενου βήματος σαν παρατηρήσεις που πρέπει να ομαδοποιηθούν. Έτσι καταφέρνουμε να ομαδοποιήσουμε ιεραρχικά τις παρατηρήσεις κατά προσέγγιση όπως θα ομαδοποιούνταν με άμεση εφαρμογή του αλγορίθμου στις παρατηρήσεις.

Η παραπάνω τεχνική είναι ιδιαίτερα χρήσιμη όταν οι παρατηρήσεις είναι πάρα πολλές και δεν είναι υπολογιστικά δυνατό να χρησιμοποιηθεί ο ιεραρχικός αλγόριθμος απευθείας σε αυτές λόγο της μεγάλης πολυπλοκότητας του αλγορίθμου ($\mathcal{O}(n^3)$).

Αλγόριθμος K-means

Τέλος, όπως περιγράφει και ο Husson στο [19] ο αλγόριθμος του K-means πρόκειται για έναν επαναληπτικό αλγόριθμο όπου συγκλίνει στις ομάδες με βάση κάποια συνθήκη (ένα κάτω όριο). Ο χρήστης πρέπει να ορίσει από την αρχή τον αριθμό των ομάδων που θα συγκλίνουν.

Έστω P_n μία ομαδοποίηση των παρατηρήσεων στο βήμα n και ρ_n το πηλίκο της σχέσης (3.14) αυτής της ομαδοποίησης. Αρχικά:

1. Έστω μια αρχική ομαδοποίηση P_0 με ρ_0.

   Στην επανάληψη n του αλγορίθμου:

2. Υπολογίζουμε το κέντρο $g_n(q)$ κάθε ομάδας q της ομαδοποίησης P_n.

3. Αναθέτουμε ξανά κάθε παρατήρηση στην ομάδα q, η οποία παρατήρηση βρίσκεται πιο κοντά στο κέντρο $g_n(q)$ της q σε σχέση με όλες τις άλλες ομάδες και δημιουργούμε την καινούργια ομαδοποίηση P_{n+1} με πηλίκο ρ_{n+1}.

4. Η διαδικασία επαναλαμβάνεται έως ότου ισχύει η συνθήκη $\rho_{n+1} - \rho_n > threshold$, αν δηλαδή η ομαδοποίηση P_{n+1} είναι καλύτερη από την P_n. Σε αντίθετη περίπτωση ο αλγόριθμος επιστρέφει στο βήμα 1.

Στη πράξη αλγόριθμος συγκλίνει πολύ γρήγορα, δηλαδή συνήθως μετά από λίγες επαναλήψεις. Επειδή το ρ_n μετά από κάθε επανάληψη μικραίνει, η σύγκλιση είναι εγγυημένη. Παρόλα αυτά ο K-means δεν εγγυάται ότι θα συγκλίνει σε ολικό βέλτιστο. Συνήθως αυτό που γίνεται, είναι ο αλγόριθμος να συγκλίνει σε ένα τοπικό βέλτιστο το οποίο όμως να προσεγγίζει σε ανεκτό βαθμό το ολικό. Μεγάλο ρόλο σε αυτό παίζει η αρχική τυχαία κατάσταση του αλγορίθμου που θα του δοθεί, καθώς είναι πολύ πιθανόν ο αλγόριθμος να μην βρίσκει την ίδια ομαδοποίηση μετά από κάθε εκτέλεση και να βρίσκεται σε διαφορετικό τοπικό βέλτιστο κάθε φορά.

4

Ανάπτυξη Εφαρμογής Επιχειρηματικής Ευφυίας Με Χρήση Της R

4.1 Εισαγωγή

Όπως έχει περιγραφεί και στα προηγούμενα κεφάλαια, μια επιχείρηση είναι ζωτικής σημασίας να γνωρίζει τους πελάτες της και τι είναι αυτό που τους κάνει να μένουν ή να φεύγουν από αυτή. Είναι σημαντικό για εκείνη να μπορεί να δημιουργεί ομάδες πελατών που ακολουθούν συγκεκριμένες αγοραστικές συμπεριφορές. Με στόχο αυτό και με τις γνώσεις της Πολυμεταβλητής Παραγοντικής Ανάλυσης που περιγράφτηκαν στο προηγούμενο κεφάλαιο, στο κεφάλαιο αυτό θα αναλυθεί η δημιουργία μια διαδικτυακής εφαρμογής επιχειρηματικής ευφυίας (BI Web Application) όπου ένας καταρτισμένος αναλυτής της επιχείρησης θα μπορούσε να χρησιμοποιήσει για να διαχωρίσει τους πελάτες της σε ομάδες με βάση τα προϊόντα ή της κατηγορίες προϊόντων που επιλέγουν σε ένα χρονικό διάστημα και έτσι να αποκτήσουν μια εικόνα στο ποια προϊόντα πωλούνται συνήθως μαζί και ενδεχομένως για παράδειγμα να μπορούσε να εκμεταλλευτεί αυτή τη πληροφορία το τμήμα Marketing.

4.2 Εργαλεία ανάπτυξης εφαρμογής

4.2.1 Η γλώσσα R

Η γλώσσα R αποτελεί μια ολοκληρωμένη σουίτα από εργαλεία για το χειρισμό των δεδομένων, τον υπολογισμό, την ανάλυση και τη γραφική απεικόνιση τους. Μεταξύ των άλλων έχει τη δυνατότητα να:

- παρέχει έναν αποδοτικό τρόπο χειρισμού και αποθήκευσης δεδομένων,
- παρέχει επίσης μία πληθώρα από μεθόδους και εργαλεία για υπολογισμούς σε πίνακες, και συγκεκριμένα πίνακες γραμμικής άλγεβρας,
- παρέχει επίσης μια μεγάλη ολοκληρωμένη συλλογή από εργαλεία για την ανάλυση δεδομένων, την γραφική απεικόνιση της ανάλυσης των δεδομένων είτε απευθείας στον υπολογιστή ή και σε έντυπη μορφή και
- παρέχει τέλος μια καλά αναπτυγμένη, απλή και συναρτησιακή γλώσσα προγραμματισμού, η οποία περιλαμβάνει ελέγχους, βρόχους, ορισμένες από το χρήστη αναδρομικές συναρτήσεις και βιβλιοθήκες εισόδου και εξόδου.

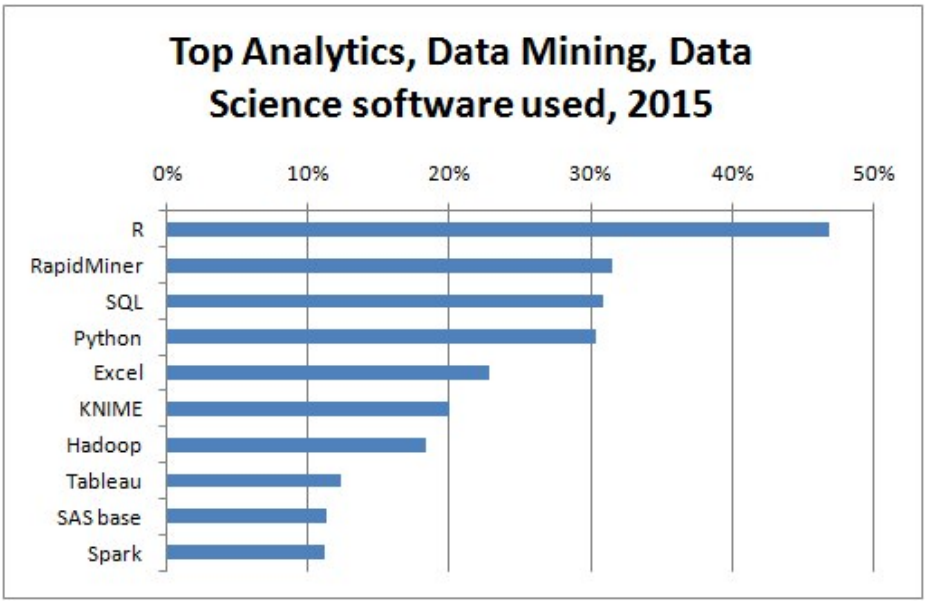
Η γλώσσα R είναι μια διερμηνευμένη γλώσσα δηλαδή είναι μια γλώσσα προγραμματισμού όπου οι εντολές που διαβάζονται εκτελούνται αμέσως. Συνήθως η επικοινωνία της γλώσσας με το χρήστη γίνεται μέσα από μια κονσόλα της R όπου οι εντολές δίνονται γραμμή γραμμή και τα αποτελέσματα εμφανίζονται αμέσως.

Η R πρόκειται για μια υλοποίηση της γλώσσας προγραμματισμού S σε συνδυασμό με την σημασιολογία που βασίζεται σε Lexical scope εμπνευσμένη από τη γλώσσα Scheme (όπως και η γλώσσα Lisp). Η S δημιουργήθηκε από τον John Chambers καθώς δούλευε στα γραφεία της Bell Labs. Υπάρχουν αρκετές σημαντικές διαφορές, αλλά ένα μεγάλο μέρος του κώδικα που γράφεται για την S μπορεί να τρέξει αναλλοίωτο.

Η R ξεκίνησε σαν project από τους Ross Ihaka και Robert Gentleman στο πανεπιστήμιο του Auckland, της Νέας Ζηλανδίας, και σήμερα υποστηρίζεται από την R Development Core Team όπου ο John Chambers είναι επίσης μέλος. Η R πήρε το όνομά της από το πρώτο γράμμα του ονόματος των δύο κατασκευαστών της.

Η R είναι ένα έργο ανοιχτού κώδικα. Ο πηγαίος κώδικας για το περιβάλλον λογισμικού της R είναι γραμμένο κυρίως σε C, Fortran και R και υποστηρίζεται επίσης από την άδεια GNU General Public License. Ενώ η R έχει μια διεπαφή γραμμής εντολών, υπάρχουν πολλά ολοκληρωμένα γραφικά περιβάλλοντα ανάπτυξης κώδικα (IDE) διαθέσιμα με κυρίαρχο και πιο δημοφιλή το RStudio, ένα ανοιχτού κώδικα project όπου δημιουργήθηκε από τον JJ Allaire το 2010.

Η R θεωρείται μία από τις πιο σημαντικές λύσεις για μια σύγχρονη επιχείρηση στο τομέα των αναλύσεων και της εξαγωγής αναφορών. Προσφέρει όλα τα απαραίτητα εργαλεία που θα χρειαστεί μια επιχείρηση για να καταφέρει να αναπτύξει μία εφαρμογή επιχειρηματικής ευφυίας. Υποστηρίζεται από μια μεγάλη πληθώρα βιβλιοθηκών και επεκτάσεων μέσω του συστήματος CRAN (Comprehensive R Archive Network) όπου ο αναλυτής μπορεί να βρει και να αναζητήσει οποιοδήποτε εργαλείο ανάλυσης επιθυμεί. Σε μία δημοσκόπηση που πραγματοποιείται κάθε χρόνο και συγκεκριμένα αυτή του 2015 που



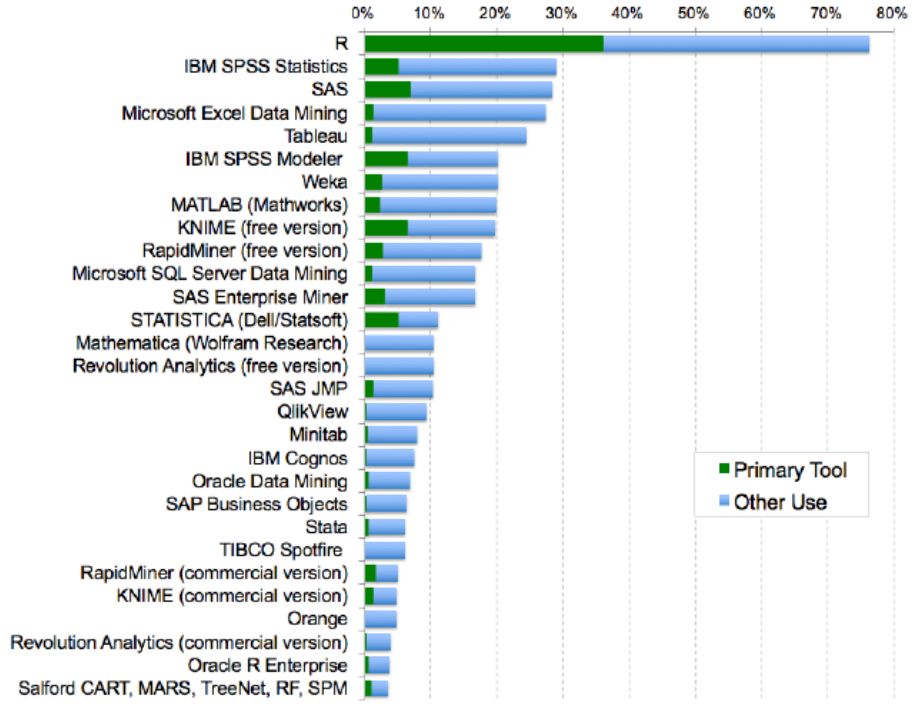
Σχημα 4.1: *Δημοσκόπηση που πραγματοποίησε η KDnuggets το 2015 σχετικά με τα πιο δημοφιλή εργαλεία ανάλυσης δεδομένων.*

πραγματοποίησε μία από τις πιο έγκριτες ιστοσελίδες που ασχολείται σχετικά με εξορύξεις δεδομένων, αναλύσεις και γενικότερα με την επιστήμη των δεδομένων η KDnuggets ρώτησε την κοινότητα των αναλυτών δεδομένων σχετικά με το πιο εργαλείο κυρίως χρησιμοποιούν στην εργασία τους. Τα αποτελέσματα φαίνονται στο σχήμα 4.1. Στην έρευνα πήραν μέρος περίπου 2.800 επιστήμονες δεδομένων, όπου στο σύνολό τους διάλεξαν συνολικά 93 διαφορετικά εργαλεία. Η συμμετοχή ανά περιοχή ήταν: US/Καναδά (41.5%), Ευρώπη (38.4%), Ασία (8.2%), Λατινική Αμερική (6.3%), Αυστραλία/Νέα Ζηλανδία (3.1%) και Αφρική/Μέση Ανατολή (2.5%).

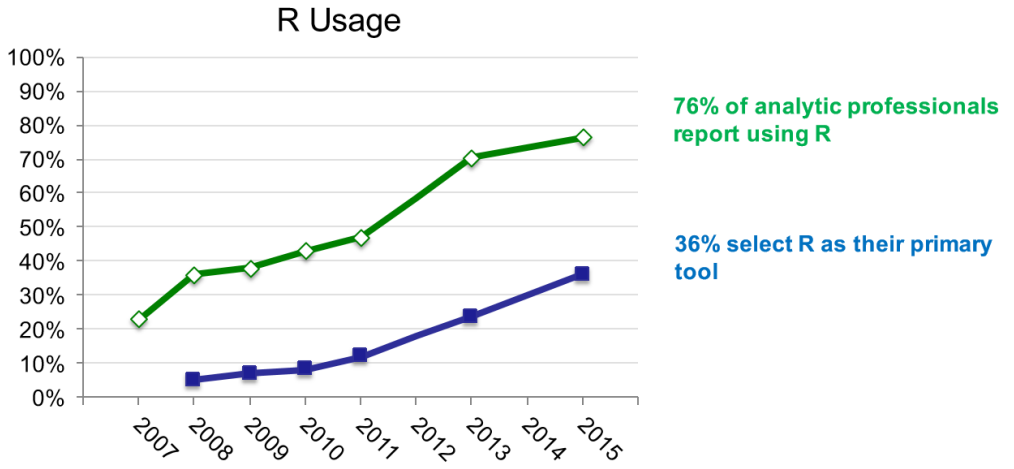
Επίσης η Rexer Analytics διεξάγει μια παρόμοια έρευνα κάθε χρόνο, με ερωτήσεις σε ένα ευρύ φάσμα θεμάτων που αφορούν την επιστήμη των δεδομένων. Στο σχήμα 4.2 φαίνονται τα εργαλεία όπου οι ερωτηθέντες ανέφεραν τη χρήση τους το 2015.

Στην ίδια έρευνα δημοσιεύτηκε ένα ερώτημα σχετικά με τη χρήση της R για την εξαγωγή επαγγελματικών αναφορών όπως επίσης και για την χρήση της ως επαγγελματικό εργαλείο σε βάθος χρόνου 9 ετών. Τα αποτελέσματα φαίνονται στο σχήμα 4.3.

Επιπλέον η RJmetrics διεξήγαγε το 2015 μία μεγάλη έρευνα με το όνομα The State of Data Science με δεξιότητες που ζητούν οι επιχειρήσεις σε θέσεις αναλυτή δεδομένων αλλά και δεξιότητες που έχουν κάποιοι αναγνωρισμένοι επιστήμονες στο χώρο. Φαίνεται ξεκάθαρα από τα σχήματα και ότι η R είναι αδιαμφισβήτητα ένα από τα απαραίτητα εργαλεία που ένα αναλυτής πρέπει να έχει στη σύγχρονη επιχείρηση.

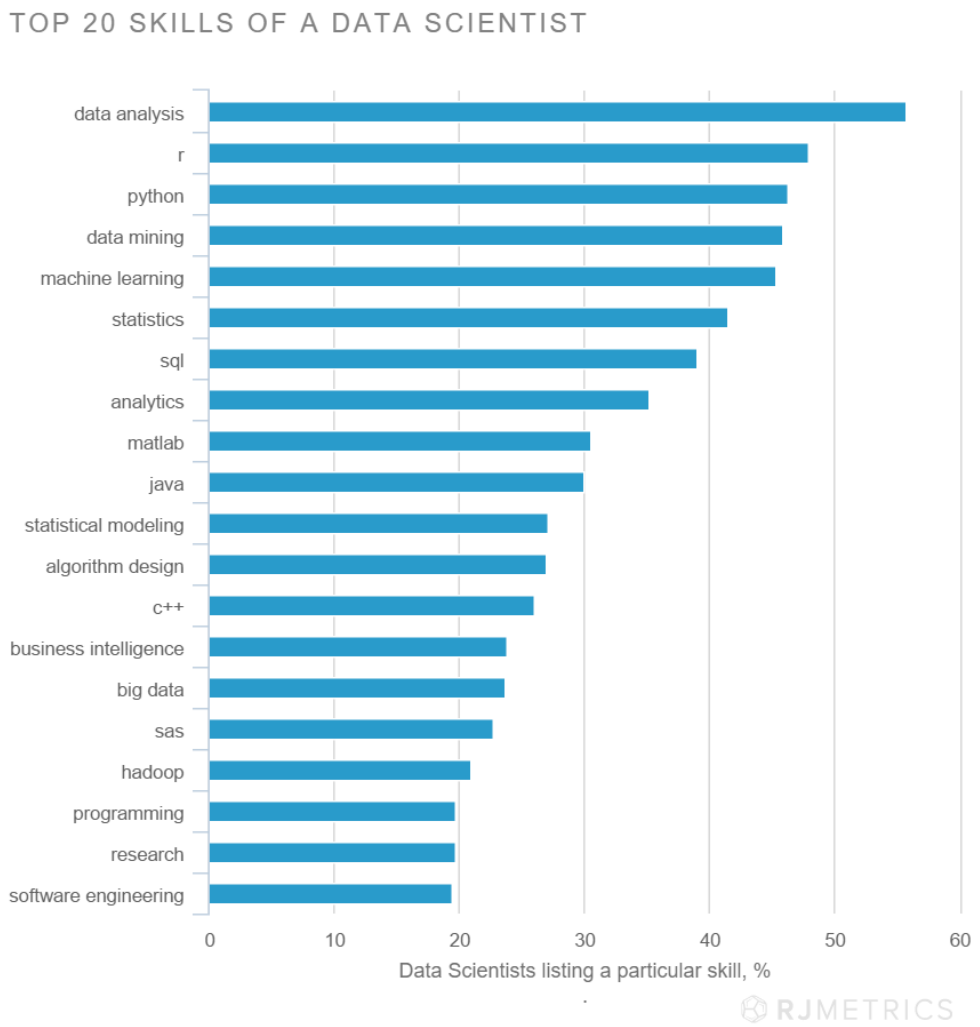


Σχημα 4.2: *Δημοσκόπηση που πραγματοποίησε η Rexer Analytics το 2015 σχετικά με τα πιο δημοφιλή εργαλεία ανάλυσης δεδομένων.*



Σχημα 4.3: *Δημοσκόπηση που πραγματοποίησε η Rexer Analytics το 2015 σχετικά με τη χρήση της R.*

Είναι ξεκάθαρη η επιρροή που έχει η R στην εφαρμογή επιχειρηματικής ευφυίας σε μια σύγχρονη επιχείρηση. Ένας ενδιαφερόμενος αναλυτής δεδομένων σε μια επιχείρηση είναι απαραίτητο να συμπεριλάβει στις δεξιότητες του τη γνώση της R. Είναι ένα εργαλείο που όπως φαίνεται και στο σχήμα 4.5 είναι κυρίαρχο στην εφαρμογή εξαγωγής αναφορών και κυρίως γραφημάτων καθώς διαθέτει μία από τις πιο δυνατές μηχανές εξαγωγής γραφημάτων με τη δική του γλώσσα γραφημάτων. Δεν είναι τυχαίο άλλωστε που η IBM ανακοίνωσε στις 7 Ιουνίου 2016 πως θα στηρίξει και θα γίνει πλατινένιος συνεργάτης στην

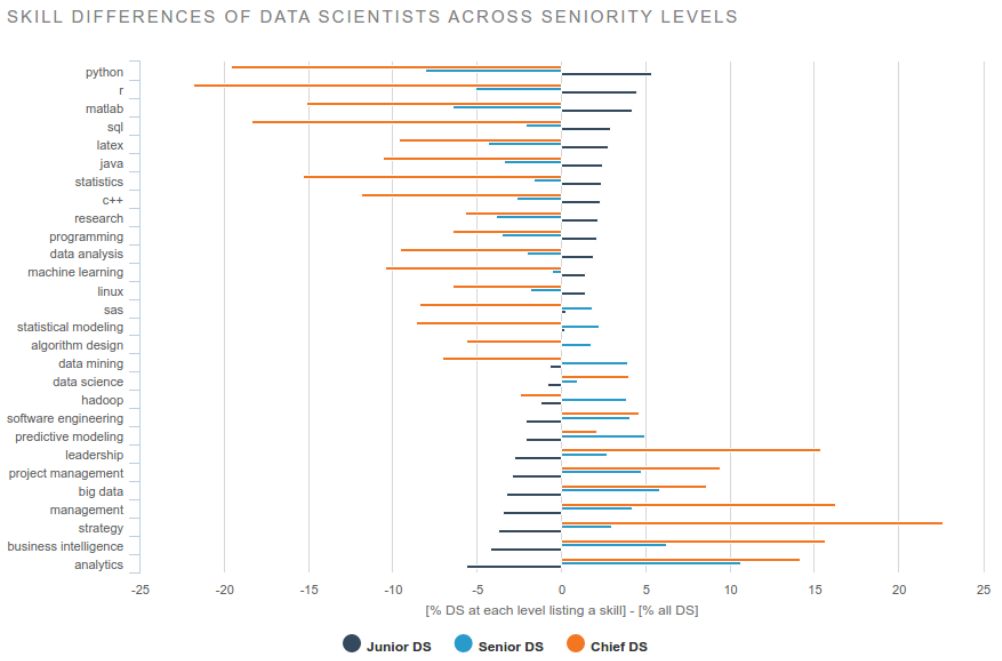


Σχημα 4.4: *Δημοσκόπηση που πραγματοποίησε η RJmetrics το 2015 σχετικά με τις πρώτες 20 πιο διαδεδομένες δεξιότητες ενός επιστήμονα δεδομένων.*

ανοιχτή κοινότητα της R με την κοινότητα R Consortium με στόχο να γίνει η κύρια γλώσσα της επιστήμης δεδομένων (Data Science).

4.2.2 Το πακέτο FactoMineR της R

Στο κεφάλαιο αυτό παρουσιάζεται το FactoMineR ένα πακέτο της R που ασχολείται με την πολυμεταβλητή ανάλυση δεδομένων. Τα κύρια χαρακτηριστικά αυτού του πακέτου είναι η δυνατότητα λήψης υπόψη τα διαφορετικά είδη μεταβλητών (ποσοτικές ή ποιοτικές), διάφορους τύπους δομής δεδομένων (ομάδες ή ιεραρχίες μεταβλητών και ομάδες παρατηρήσεων) όπως επίσης και συμπληρωματικές πληροφορίες (συμπληρωματικές παρατηρήσεις και μεταβλητές).



Σχημα 4.5: *Δημοσκόπηση που πραγματοποίησε η RJmetrics το 2015 σχετικά με τη διαφορά επιλογής δεξιοτήτων μεταξύ διάφορων επιπέδων επιστημόνων δεδομένων.*

Επιπλέον, οι διαστάσεις που εξάγονται από τις διάφορες αναλύσεις διερευνητικών δεδομένων μπορούν αυτόματα να περιγραφούν από τις ποσοτικές ή/και κατηγορικές μεταβλητές. Πολυάριθμες γραφικές παραστάσεις είναι επίσης διαθέσιμες με μια πληθώρα επιλογών.

Εγκατάσταση του πακέτου FactoMineR

Η εγκατάσταση του πακέτου είναι πολύ εύκολη. Απλά ο χρήστης αρκεί να πληκτρολογήσει την παρακάτω εντολή

```
1 install.packages("FactoMineR")
```

στην κονσόλα της R και στη συνέχεια για να φορτώσουμε το πακέτο αρκεί να χρησιμοποιήσουμε μία από τις παρακάτω εντολές

```
1 library(FactoMineR)
2 #or
3 require(FactoMineR)
```

Από εδώ και πέρα μπορεί να προχωρήσει η ανάπτυξη της εφαρμογής.

Χρήση του πακέτου FactoMineR

Στη παρούσα εργασία το πακέτο FactoMineR χρησιμοποιήθηκε ως το βασικό εργαλείο για την εκτέλεση της Πολυμεταβλητής Παραγοντικής Ανάλυσης

Αντιστοιχιών. Η συναρτήσεις του πακέτου που χρησιμοποιήθηκαν παρουσιάζονται παρακάτω:

- Συνάρτηση MCA Πρόκειται για τον πυρήνα της ανάλυσης. Πραγματοποιεί την Πολυμεταβλητή Παραγοντική Ανάλυση Αντιστοιχιών σύμφωνα με τη θεωρία που περιγράφτηκε στην Ενότητα 3.3. Δέχεται δεδομένα σε δομή τύπου data.frame. Τα τιμές των μεταβλητών που θα συμμετέχουν ενεργά στην ανάλυση πρέπει υποχρεωτικά να είναι τύπου factor. Ένα πλεονέκτημα της Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών με τη χρήση αυτού του πακέτου έναντι άλλων πακέτων της R όπως του πακέτου MASS ή του πακέτου ca είναι ότι μεταξύ των άλλων επιτρέπει τη χρήση συμπληρωματικών (Supplementary) παρατηρήσεων όπως και ποιοτικών και αριθμητικών μεταβλητών, δυνατότητα όπου λείπει από τα υπόλοιπα πακέτα. Οι συμπληρωματικές μεταβλητές και παρατηρήσεις δεν χρησιμοποιούνται για τον καθορισμό των κύριων συνιστωσών. Αντίθετα, οι συντεταγμένες τους στους χάρτες των κατηγοριών και των μεταβλητών καθορίζεται χρησιμοποιώντας μόνο την πληροφορία που παράγεται από την ανάλυση στις ενεργές μεταβλητές/παρατηρήσεις.

Η χρήση της συνάρτησης είναι η εξής:

```
1 MCA(X, ncp = 5, ind.sup = NULL, quanti.sup = NULL, quali.sup = NULL,
  ↪  graph = TRUE, level.ventil = 0, axes = c(1,2), row.w = NULL,
  ↪  method="Indicator", na.method="NA", tab.disj=NULL)
```

Τα ορίσματα που δέχεται είναι αναλυτικά:

`X`	ένα data.frame με n γραμμές (παρατηρήσεις) και p στήλες (ποιοτικές μεταβλητές)
`ncp`	ο αριθμός των διαστάσεων που θα επισταφεί στο αποτέλεσμα (προεπιλεγμένο 5)
`ind.sup`	ένα διάνυσμα που περιέχει τους δείκτες των συμπληρωματικών παρατηρήσεων
`quanti.sup`	ένα διάνυσμα που περιέχει τους δείκτες των συμπληρωματικών αριθμητικών μεταβλητών
`quali.sup`	ένα διάνυσμα που περιέχει τους δείκτες των συμπληρωματικών ποιοτικών μεταβλητών
`graph`	αν η τιμή είναι `True` τότε εμφανίζονται τα επιπλέον σχετικά γραφήματα
`level.ventil`	ένα ποσοστό που αντιστοιχεί στο επίπεδο με βάση το οποίο μια κατηγορία δεν θα λάβει μέρος (από προεπιλογή, η τιμή είναι 0 και δεν γίνεται κάποιος διαχωρισμός)
`axes`	ένα διάνυσμα μεγέθους 2 που δηλώνει τις κύριες συνιστώσες που θα εμφανιστούν στο γράφημα
`row.w`	προαιρετικά ένα διάνυσμα με βάρη συμμετοχής των παρατηρήσεων (από προεπιλογή, τα βάρη είναι όλα 1)
`method`	μια συμβολοσειρά που αντιστοιχεί στο όνομα της μεθόδου που θα χρησιμοποιηθεί: ''Indicator'' (από προεπιλογή) είναι η CA στον Indicator matrix, ''Burt'' είναι η CA στον πίνακα Burt
`na.method`	μια συμβολοσειρά που αντιστοιχεί στο όνομα της μεθόδου που θα χρησιμοποιείται αν υπάρχουν τιμές που λείπουν διαθέσιμες μέθοδοι είναι ''NA'' ή ''Average'' (από προεπιλογή, ''NA'')
`tab.disj`	ένα προαιρετικό data.frame που αντιστοιχεί στο διαζευκτικό πίνακα που χρησιμοποιείται για την ανάλυση αντιστοιχεί σε έναν διαζευκτικό πίνακα που λαμβάνεται από τη μέθοδο εκτίμησης (βλ. πακέτο missMDA).

- Συνάρτηση HCPC Η συνάρτηση αυτή είναι υπεύθυνη για να πραγματοποιήσει την (agglomerative) ιεραρχική ομαδοποίηση στους πελάτες/παρατηρήσεις όπως προκύπτουν από την ανάλυση MCA.

Η χρήση της συνάρτησης είναι η εξής:

```
1 HCPC(res, nb.clust=0, consol=TRUE, iter.max=10, min=3, max=NULL,
  ↪ metric="euclidean", method="ward", order=TRUE,
  ↪ graph.scale="inertia", nb.par=5, graph=TRUE, proba=0.05,
  ↪ cluster.CA="rows",kk=Inf,...)
```

Τα ορίσματα που δέχεται είναι αναλυτικά:

`res`	η έξοδος της ανάλυσης MCA.
`nb.clust`	ο αριθμός των ομάδων. Αν είναι -1 τότε το δέντρο 'κόβεται' αυτόματα στο προτεινόμενο ύψος.
`consol`	αν είναι `TRUE` τότε μία ομαδοποίηση k-Means πραγματοποιείται μετά τον ιεραρχικό αλγόριθμο για να δώσει πιο συμπαγείς ομάδες.
`iter.max`	ο μέγιστος αριθμός επαναλήψεων του consolidation.
`min`	ο ελάχιστος αριθμός ομάδων που θα προταθούν από την αυτόματη αποκοπή του δέντρου.
`max`	ο μέγιστος αριθμός ομάδων που θα προταθούν από την αυτόματη αποκοπή του δέντρου.
`metric`	η μετρική για την κατασκευή του δέντρου (`euclidean` ή `manhattan`).
`method`	η μέθοδος για την κατασκευή του δέντρου (`average`, `single`, `complete`, `ward`, `weighted` ή `flexible`).
`order`	αν είναι `TRUE` τότε οι ομάδες ακολουθούν τις συντεταγμένες του κέντρου του πρώτου άξονα.
`graph.scale`	δέχεται συμβολοσειρές. Προεπιλογή είναι η λέξη 'inertia' και το ύψος του δέντρου δηλώνει το κέρδος της αδράνειας, αν η λέξη είναι 'sqrt-inertia' δηλώνει τη τετραγωνική ρίζα του κέρδους της αδράνειας.
`nb.par`	ο αριθμός των επεξεργασμένων υποδειγμάτων παρατηρήσεων.
`graph`	αν η τιμή είναι `TRUE` τότε εμφανίζονται τα επιπλέον σχετικά γραφήματα
`proba`	η πιθανότητα επιλογής αξόνων και μεταβλητών για την περιγραφή των ομάδων.
`cluster.CA`	μια συμβολοσειρά που δηλώνει τις γραμμές και τις στήλες για την ομαδοποίηση του πίνακα Παραγοντικής Ανάλυσης.
`kk`	ο αριθμός των ομάδων για την εφαρμογή του k-means πριν την ιεραρχική ομαδοποίηση.

4.2.3 Το πακέτο Shiny της R

Στο κεφάλαιο αυτό παρουσιάζεται το πακέτο Shiny της R, ένα πακέτο ανοιχτού κώδικα που παρέχει ένα κομψό και ισχυρό διαδικτυακό πλαίσιο για τη δημιουργία διαδικτυακών εφαρμογών χρησιμοποιώντας την R. Το Shiny είναι σε θέση να μετατρέψει τις αναλύσεις σε διαδραστικές διαδικτυακές εφαρμογές

χωρίς να απαιτείται HTML, CSS, ή γνώση JavaScript. Βασίζεται κυρίως στο περιβάλλον του Bootstrap και δίνει πολλές έτοιμες (out-of-the-box) δυνατότες.

Εγκατάσταση του πακέτου Shiny

Η εγκατάσταση του πακέτου είναι πολύ εύκολη. Απλά ο χρήστης αρκεί να πληκτρολογήσει την παρακάτω εντολή

```
1 install.packages("shiny")
```

στην κονσόλα της R και στη συνέχεια για να φορτώσουμε το πακέτο αρκεί να χρησιμοποιήσουμε μία από τις παρακάτω εντολές

```
1 library(shiny)
2 #or
3 require(shiny)
```

Από εδώ και πέρα μπορεί να προχωρήσει η ανάπτυξη της εφαρμογής.

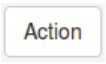
Χρήση του πακέτου Shiny

Ένα project γραμμένο με τη χρήση του Shiny αποτελείται συνήθως από δύο αρχεία: το αρχείο ui.R και το αρχείο server.R. Επίσης συνήθη πρακτική είναι η χρήση και κάποιον εξωτερικών αρχείων με κώδικα που θα χρειαστεί για τις διάφορες αναλύσεις.

Βασικά Widgets

Μερικά από τα πιο βασικά widgets που μπορούν να χρησιμοποιηθούν απευθείας από το περιβάλλον του Shiny είναι τα εξής:

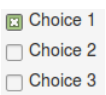
- Action Button



Button όπου η λειτουργία του ορίζεται στο αρχείο server.R.

```
1 actionButton("action", label = "Action")
```

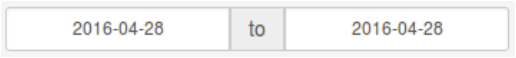
- (Grouped) Checkbox



Κουμπί δύο καταστάσεων (True ή False) που μπορεί να οριστεί και σαν ομάδα.

```
checkboxInput("checkbox", label = "Choice A", value = TRUE)
#or
checkboxGroupInput("checkGroup", label = h3("Checkbox group"),
↪  choices = list("Choice 1" = 1, "Choice 2" = 2, "Choice 3" =
↪  3), selected = 1)
```

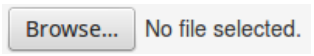
- **Date selector**



Κουμπί που επιτρέπει την εύκολη επιλογή ημερομηνίας.

```
dateRangeInput("dates", label = h3("Date range"))
#or
dateInput("date", label = h3("Date input"), value =
↪  "2014-01-01")
```

- **File Uploader**



Κουμπί που επιτρέπει την εύκολη επιλογή αρχείου για ανέβασμα στην εφαρμογή.

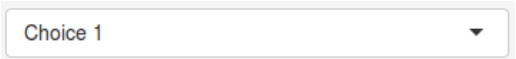
```
fileInput("file", label = h3("File input"))
```

- **Numeric Field**

Πλαίσιο εισαγωγής αριθμητικών τιμών.

```
numericInput("num", label = h3("Numeric input"), value = 1)
```

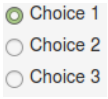
- **Select Box**



Πλαίσιο για την επιλογή ενός ή περισσοτέρων τιμών από μία λίστα.

```
selectInput("select", label = h3("Select box"), choices =
↪  list("Choice 1" = 1, "Choice 2" = 2, "Choice 3" = 3),
↪  selected = 1)
```

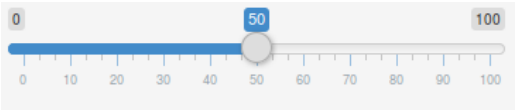
- Radio Buttons



Μια ομάδα κουμπιών όπου επιτρέπεται η επιλογή ενός μόνο.

```
radioButtons("radio", label = h3("Radio buttons"), choices =
↪  list("Choice 1" = 1, "Choice 2" = 2, "Choice 3" = 3),
↪  selected = 1)
```

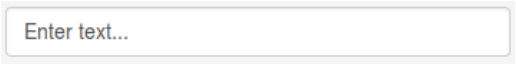
- Slider



Ένας slider αριθμητικών τιμών όπου δέχεται τιμές σε ένα περιορισμένο εύρος.

```
sliderInput("slider1", label = h3("Slider"), min = 0, max = 100,
↪  value = 50)
```

- Text Field



Πλαίσιο εισαγωγής σύντομου κειμένου.

```
textInput("text", label = h3("Text input"), value = "Enter
↪  text...")
```

Το αρχείο ui.R

Στο αρχείο αυτό περιέχεται εκείνο το κομμάτι του προθεστ όπου περιγράφει την εμφάνιση της εφαρμογής. Αν τα widgets που θα δηλωθούν και θα χρησιμοποιηθούν είναι στατικά τότε δεν μπορούν να αλλάξουν παρά μόνο τις ιδιότητες τους κατά τη διάρκεια της εκτέλεσης. Δηλαδή σε αυτό το αρχείο δηλώνεται η στατική εμφάνιση της εφαρμογής. Εκτός από τα βασικά widgets μπορούν επίσης να τοποθετηθούν στο αρχείο, συναρτήσεις εξόδου προς το αρχείο server.R που από εκεί θα οριστούν δυναμικά αποτελέσματα αναλύσεων ή δυναμικά widgets μέσω των συναρτήσεων render. Οι συναρτήσεις εξόδου και οι αντίστοιχες συναρτήσεις render είναι οι εξής:

Output function	Render function	Usage
uiOutput	renderUI	Raw HTML
htmlOutput	renderUI	Raw HTML
plotOutput	renderPlot	Plot
tableOutput	renderTable	Table
dataTableOutput	renderDataTable	Advanced Table
textOutput	renderText	Text
verbatimTextOutput	renderText	Text
imageOutput	renderImage	Image

Πίνακας 4.1: Οι συναρτήσεις εξόδου με τις αντίστοιχες render συναρτήσεις και τη χρήση τους.

Το αρχείο server.R

Το αρχείο αυτό είναι υπεύθυνο για τη δυναμική λειτουργία της εφαρμογής. Επιτρέπει στο χρήστη να διενεργεί με την εφαρμογή μέσω δύο παραμέτρων: της **input** και της **output**. Η πρώτη είναι υπεύθυνη για να μεταφέρει τις τιμές των widget που έχει ρυθμίσει ο χρήστης και η output είναι υπεύθυνη για να σχεδιάσει τις συναρτήσεις εξόδου μέσω των συναρτήσεων render. Η πιο συμαντική λειτουργία που προσφέρει το Shiny είναι οι εκφράσεις **reactive**. Πρόκειται για ένα δοκιμή στοιχείο που δίνει τη δυνατότητα για διαδραστική δημιουργία και ενημέρωση των ωιδγετς αλλά και την διαδραστική εκτέλεση λειτουργιών στο παρασκήνιο. Η συνταξή της είναι η εξής:

```
1  plotToPrint <- reactive({
2          plot(data, input$x, input$y)
3  })
4
5  output$plot1 <- renderPlot({
6          plotToPrint()
7  })$
```

4.3 Scienica for Business - Εφαρμογή Επιχειρηματικές Ευφυίας

Σε αυτό το κεφάλαιο περιγράφεται η εφαρμογή Επιχειρηματικής Ευφυίας που αναπτύχθηκε στο πλαίσιο της παρούσας εργασίας. Στόχος είναι να δημιουργηθεί μια πλατφόρμα προσιτή για τον αναλυτή μιας επιχείρησης που θέλει

να εφαρμόσει τις τεχνικές Πολυμεταβλητής Παραγοντικής Ανάλυσης που αναλύθηκαν στο προηγούμενο κεφάλαιο, για να δημιουργήσει ομάδες πελατών σε σχέση με την αγοραστική τους συμπεριφορά. Η κατασκευή έγινε με τη χρήση του πακέτου Shiny της στατιστικής γλώσσας R για ερευνητικούς σκοπούς.

4.3.1 Ο σκοπός της εφαρμογής

Η διαδικτυακή εφαρμογή επιχειρηματικής ευφυίας που αναπτύχθηκε, αποτελεί το κύριο κομμάτι αυτής της εργασίας και σαν στόχο έχει να δημιουργηθεί ένα περιβάλλον όπου ένας αναλυτής με κάποιες γνώσεις πάνω στη Πολυμεταβλητή Παραγοντική Ανάλυση θα τη χρησιμοποιούσε για να εφαρμόσει με αυτές τις τεχνικές μια τμηματοποίηση των πελατών της επιχείρησης. Επειδή η φιλοσοφία της εφαρμογής μπορεί να χρησιμοποιηθεί και για οποιοδήποτε σύνολο δεδομένων το οποίο πλήρη της προδιαγραφές όπως παρουσιάζονται στο Κεφάλαιο 4.4, αποφασίστηκε ο χαρακτήρας της εφαρμογής να διατηρηθεί ουδέτερος. Σε όλες τις δοκιμές χρησιμοποιήθηκε ένα σύνολο δεδομένων από πελάτες ενός καταστήματος τροφίμων που βρέθηκε στο πακέτο arules[24] της R. Η εφαρμογή αποτελείται από τρία βασικά μέρη: **Οπτικοποίηση Δεδομένων** όπου απαρτίζεται κυρίως από εργαλεία οπτικοποίησης του συνόλου δεδομένων με μεγάλη γκάμα από διαγράμματα, **Ανάλυση MCA** όπου πραγματοποιείται η κύρια ανάλυση MCA με διάφορα γραφήματα και ποιο αναλυτικές περιγραφές για την διασφάλιση της καλής ποιότητας της ανάλυσης από τον αναλυτή και τέλος **Ανάλυση HCPC** όπου εδώ πραγματοποιείται και ο στόχος της εφαρμογής, δηλαδή η τμηματοποίηση και η ομαδοποίηση των παρατηρήσεων μέσω του ιεραρχικού αλγορίθμου και του αποτελέσματος της MCA. Παρακάτω περιγράφονται πιο αναλυτικά.

4.3.2 Τα βασικά μέρη της εφαρμογής

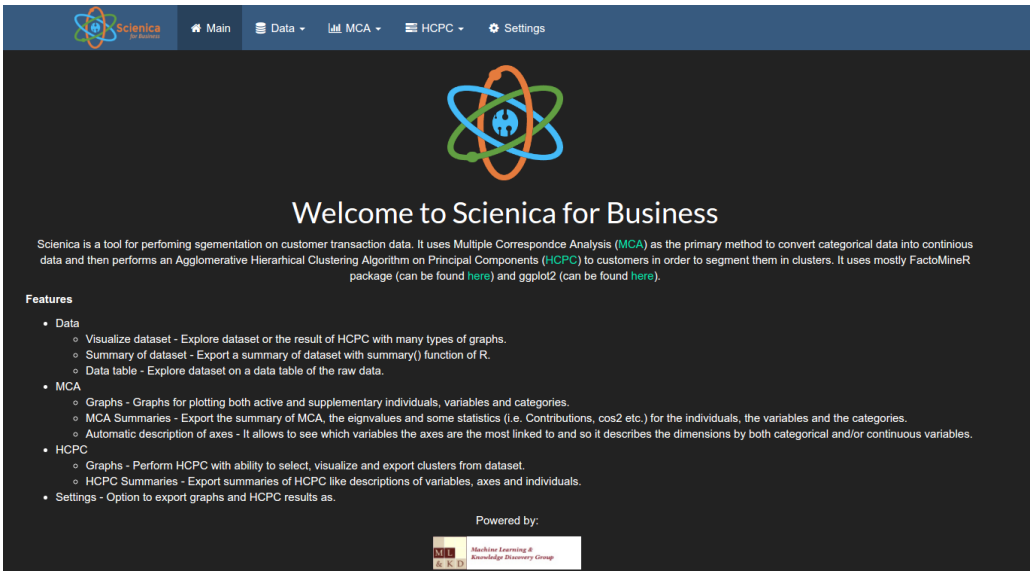
Η εφαρμογή αποτελείται από τρία μέρη, το μενού πλοήγησης, το πλαϊνό μενού και το κυρίως πάνελ. Στην έναρξη της εφαρμογής το μενού πλοήγησης και το κυρίως πάνελ είναι ορατά ενώ κατά τις υπόλοιπες λειτουργίες της εφαρμογής εμφανίζεται όπου χρειάζεται και το πλαϊνό μενού.

4.3.2.1 Αρχική Οθόνη (Home)

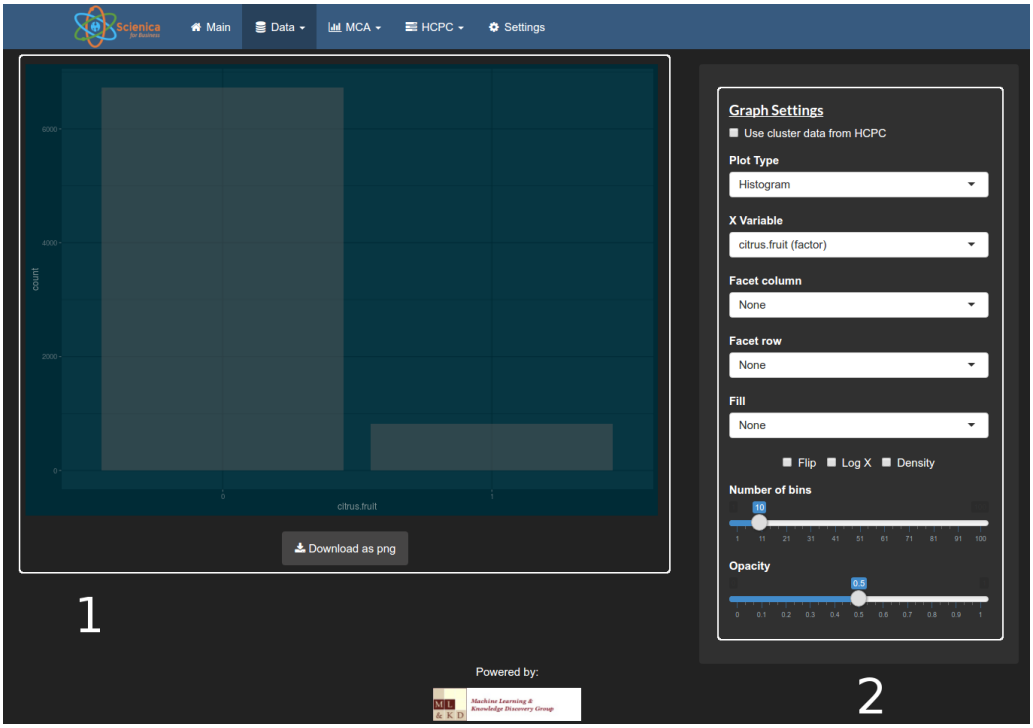
Στην εικόνα 4.6 φαίνεται η όψη της αρχικής οθόνης που αντικρίζει ο χρήστης κατά την εισαγωγή του στην εφαρμογή. Σκοπός αυτής της οθόνης είναι να ενημερώσει το χρήστη σύντομα για τις λειτουργίες της εφαρμογής.

4.3.2.2 Οπτικοποίηση Δεδομένων (Visualize dataset)

Η εικόνα 4.7 δείχνει οθόνη οπτικοποίησης του συνόλου δεδομένων που ο αναλυτής θα χρησιμοποιήσει για την ανάλυση του περαιτέρω. Σκοπός αυτής της λειτουργίας είναι να δώσει στο χρήστη ένα εργαλείο ώστε να καταλάβει καλύτερα και σε βάθος τα δεδομένα του και να εντοπίσει τυχών μοτίβα που



Σχημα 4.6: *Η αρχική οθόνη της εφαρμογής.*



Σχημα 4.7: *Η οθόνη οπτικοποίησης των δεδομένων.*

μπορεί να υπάρχουν στα δεδομένα ή μεταβλητές οι οποίες προσθέτουν θόρυβο στην ανάλυση.

Η οθόνη χωρίζεται σε δύο μέρη. Στη κύρια οθόνη και στο πλαϊνό μενού. Αναλυτικά:

1. **Κύρια Οθόνη:** είναι υπεύθυνη για να απεικονίζει τα διάφορα γραφήματα που προκύπτουν από τις διάφορες επιλογές του χρήστη. Επίσης δίνει την δυνατότητα στο χρήστη να κατεβάσει το γράφημα.

2. **Πλαϊνό Μενού:** οι λειτουργίες που προσφέρει φαίνονται στον πίνακα 4.2.

Η συνάρτηση που δημιουργήθηκε γι΄ αυτό το σκοπό βρίσκεται στο Παράρτημα Πηγαίου Κώδικα και συγκεκριμένα στο πίνακα Α΄.1.

Function Name	**Values**	**Usage**
Use cluster data from HCPC	True/False	Either use the original data or the data from HCPC
Plot Type	Histogram; Density; Scatter; Line; Bar; Box Plot; Violin	Choose between different types of plots
X Variable	Variables from dataset	Choose the variable to be plotted on x axes
Y Variable	Variables from dataset	Choose the variable to be plotted on y axes (available in all types of plot except Histogram and Density)
Facet column	Variables from dataset	Choose the variable to be plotted on columns facet (adds an extra dimension on plot)
Facet row	Variables from dataset	Choose the variable to be plotted on rows facet (adds an extra dimension on plot)
Fill/Color	Variables from dataset	Choose the variable to be plotted as an extra dimension of colors
Flip	True/False	Flipped cartesian coordinates so that horizontal becomes vertical, and vertical, horizontal
LogX/LogY	True/False	Log scale continuous position (X or Y)
Density	True/False	Plot density line
Line	True/False	Plot linear regression line on scatter plot
Loess	True/False	Plot loess regression line on scatter plot
Jitter	True/False	Plot jitter points
Sort	True/False	Sort available values on bar plot
Number of bins	1-100	Number of bins on histogram
Smooth	0.1-3	Smooth of density plot
Opacity	0-1	Opacity of plots

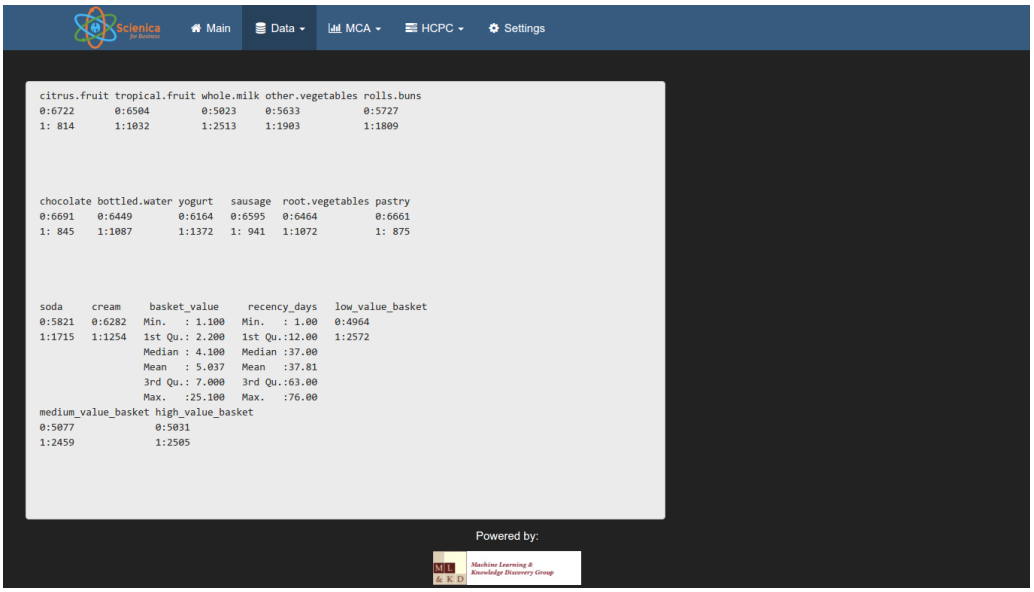
Πίνακας 4.2: Οι λειτουργίες της οπτικοποίησης των δεδομένων.

4.3.2.3 Σύνοψη των δεδομένων (Summary of dataset)

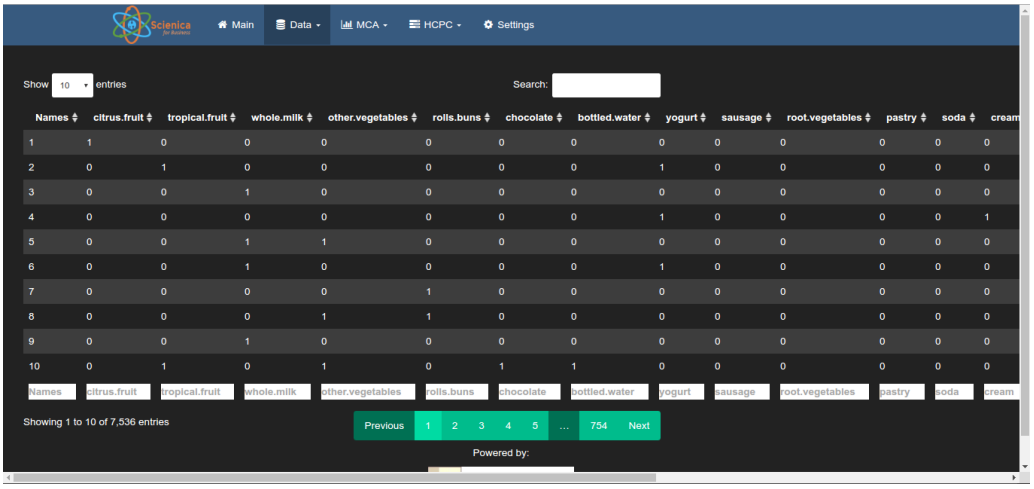
Η εικόνα 4.8 δείχνει οθόνη σύνοψης του συνόλου δεδομένων. Εδώ μπορεί ο αναλυτής να πάρει κάποιες πολύ σύντομες πληροφορίες για τα δεδομένα του. Χρησιμοποιεί την built-in συνάρτηση της R `summary(data)`.

4.3.2.4 Πίνακας Δεδομένων (Data Table)

Στην εικόνα 4.9 φαίνεται η οθόνη του πίνακα δεδομένων. Εδώ ο αναλυτής έχει τη δυνατότητα να κοιτάξει απευθείας στα δεδομένα και να αναζητήσει συγκεκριμένες εγγραφές.



Σχημα 4.8: *Η οθόνη σύνοψης των δεδομένων.*



Σχημα 4.9: *Η οθόνη του πίνακα δεδομένων.*

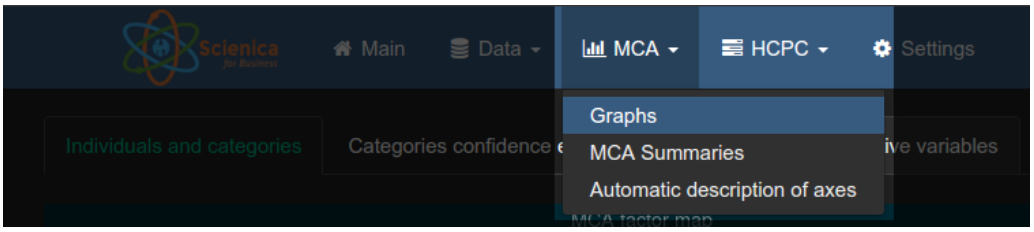
4.3.2.5 Ανάλυση MCA

Η χρήση της μεθόδου ανάλυσης MCA μπορεί εύκολα να πραγματοποιηθεί όταν ο χρήστης επιλέξει την επιλογή MCA από το μενού πλοήγησης και επιλέγοντας μία από τις επιλογές: **Graphs (Γραφήματα Ανάλυσης MCA)**, **MCA Summaries (Σύνοψη της ανάλυσης MCA)** ή **Automatic description of axes (Αυτόματη περιγραφή των αξόνων)** όπως φαίνεται και στο σχήμα 4.10.

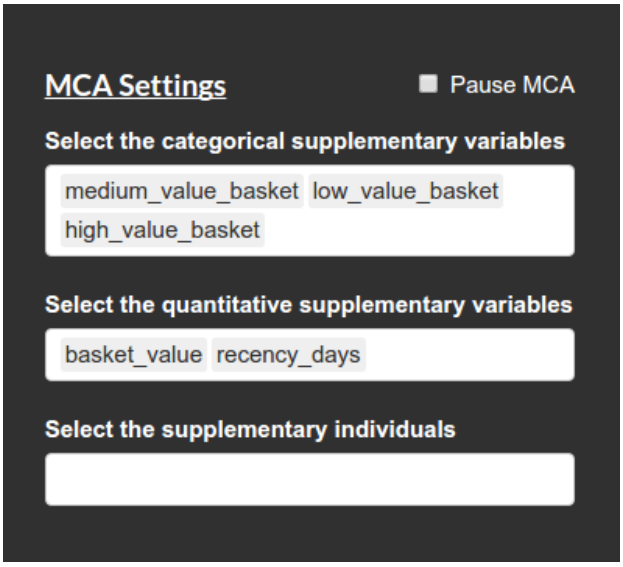
Σε όλες τις παραπάνω επιλογές δίνεται η δυνατότητα άμεσης αλλαγής των ρυθμίσεων της ανάλυσης MCA όπως φαίνεται στο σχήμα 4.11.

Οι επιλογές είναι αναλυτικά:

Select the categorical supplementary variables: Εδώ δίνεται η δυνατότητα στο χρήστη να διαλέξει τις κατηγορικές συμπληρωματικές μετα-



Σχημα 4.10: *Το κεντρικό μενού της ανάλυσης MCA.*



Σχημα 4.11: *Οι ρυθμίσεις της ανάλυσης MCA.*

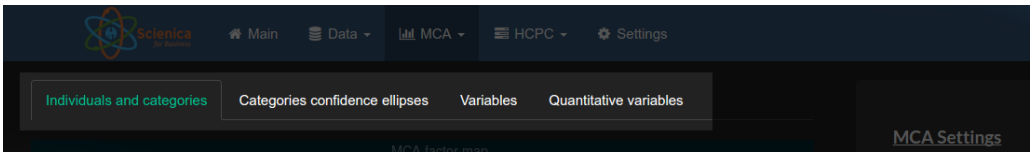
βλητές ανάμεσα στις κατηγορικές μεταβλητές του συνόλου δεδομένων με στόχο να μην λάβουν μέρος στην ανάλυση παρά μόνο στο αποτέλεσμα.

Select the quantitative supplementary variables: Εδώ δίνεται η δυνατότητα στο χρήστη να διαλέξει τις ποσοτικές συμπληρωματικές μεταβλητές ανάμεσα στις ποσοτικές μεταβλητές του συνόλου δεδομένων εφόσον υπάρχουν και με στόχο να μην λάβουν μέρος στην ανάλυση παρά μόνο στο αποτέλεσμα.

Select the supplementary individuals: Εδώ δίνεται η δυνατότητα στο χρήστη να διαλέξει τις συμπληρωματικές παρατηρήσεις του συνόλου δεδομένων ώστε να μην λάβουν μέρος στην ανάλυση παρά μόνο στο αποτέλεσμα.

Pause MCA: Ο χρήστης έχοντας αυτή την επιλογή ενεργοποιημένη, κάθε αλλαγή στις ρυθμίσεις δεν θα έχει άμεσο υπολογισμό της ανάλυσης αλλά θα γίνεται όταν απενεργοποιηθεί.

Στη συνέχεια παρουσιάζονται αναλυτικά κάθε λειτουργία και παράθυρο που προσφέρει η ανάλυση MCA.



Σχημα 4.12: *Οι επιλογές γραφημάτων της ανάλυσης MCA.*

4.3.2.5.1 Γραφήματα Ανάλυσης MCA (MCA Graphs)

Τα διαθέσιμα γραφήματα της ανάλυσης MCA είναι τέσσερα και μπορούν να πλοηγηθούν από το μενού του σχήματος 4.12. Οι σημασία κάθε γραφήματος φαίνεται συνοπτικά στον πίνακα 4.3.

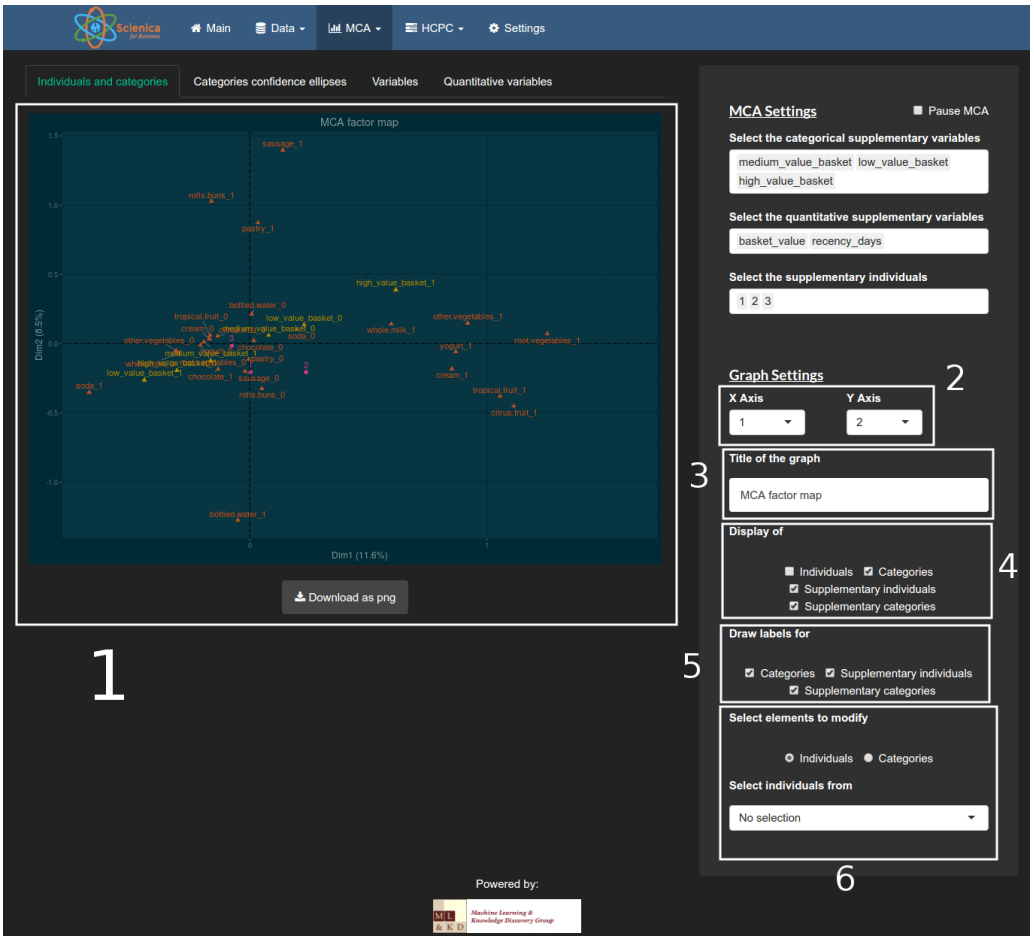
Όνομα Γραφήματος	**Σημασία**
Individuals and categories	Στο γράφημα αυτό ο χρήστης μπορεί να διακρίνει τις συσχετίσεις μεταξύ των (συμπληρωματικών) κατηγοριών αλλά και μεταξύ των (συμπληρωματικών) παρατηρήσεων.
Categories confidence ellipses	Στο γράφημα αυτό ο χρήστης μπορεί να δει για κάθε ποιοτική μεταβλητή την εμπιστοσύνη κάθε κατηγορίας της μεταβλητής με κατάλληλα χρωματισμένες της αντίστοιχες παρατηρήσεις και με σχεδιασμένους κύκλους εμπιστοσύνης για κάθε κατηγορία.
Variables	Το γράφημα αυτό δείχνει τη συσχέτιση μεταξύ όλων των μεταβλητών.
Quantitative variables	Στο γράφημα φαίνεται η συσχέτιση μεταξύ των αριθμητικών μεταβλητών και των κύριων συνιστωσών που ονομάζονται loadings. Οι μεταβλητές απεικονίζονται ως σημεία στο χώρο των κύριων συνιστωσών χρησιμοποιώντας τα loadings για συντεταγμένες.

Πίνακας 4.3: Τα γραφήματα και η σημασία τους της ανάλυσης MCA.

Παρακάτω παρουσιάζονται κάθε γράφημα με τις λειτουργίες του ξεχωριστά.

Individuals and categories Σε αυτήν την επιλογή γραφήματος ο χρήστης έχει τη δυνατότητα να παράγει δύο από τα γραφήματα που περιγράφτηκαν στην Ενότητα 3.3.6. Συγκεκριμένα του δύνεται η δυνατότητα να κατασκευάσει τον **Χάρτη των παρατηρήσεων** και τον **Χάρτη των κατηγοριών** όπως και συνδυασμό αυτών όπως φαίνεται στο σχήμα 4.14. Ο χρήστης μέσω αυτού του γραφήματος μπορεί να διακρίνει τις σχέσεις μεταξύ των κατηγοριών όπου πλέον έχουν αποκτήσει συνεχής και αριθμητικές τιμές. Είναι σε θέση επίσης να συγκρίνει την θέση των παρατηρήσεων στο χάρτη και να διακρίνει έτσι μια πρώτης μορφή ομαδοποίηση.

Στο σχήμα 4.13 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:



Σχημα 4.13: *Η οθόνη του γραφήματος Individuals and categories της ανάλυσης MCA.*

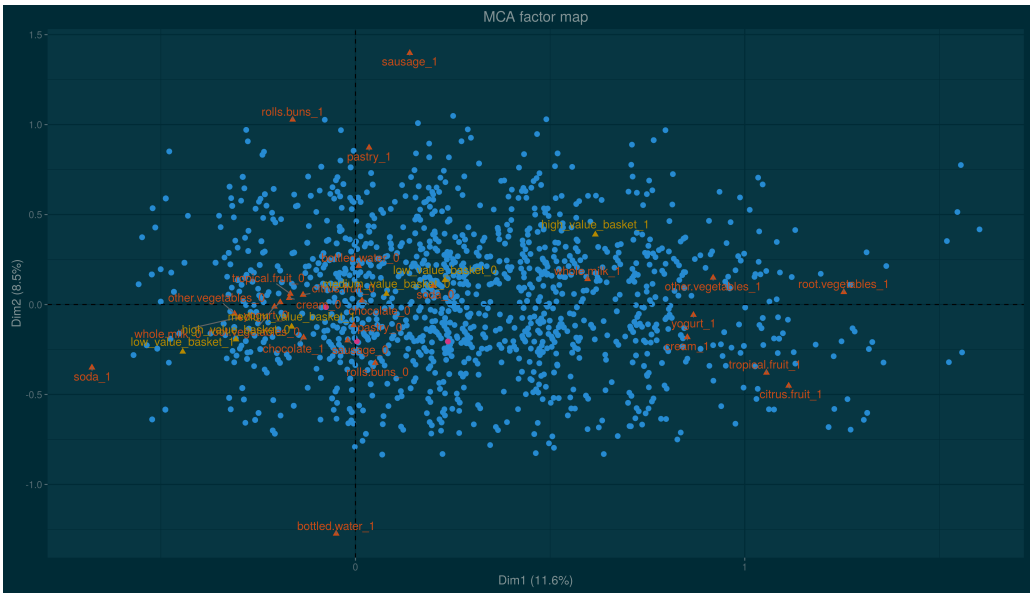
1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Εδώ ο χρήστης μπορεί να αλλάξει ποια κύρια συνιστώσα θα σχεδιαστεί στο γράφημα για τον άξονα x και τον άξονα y.

3. Ο χρήστης σε αυτήν την επιλογή έχει τη δυνατότητα να αλλάξει τον τίτλο του γραφήματος.

4. Ο χρήστης εδώ μπορεί να διαλέξει αν θα σχεδιαστούν στο γράφημα οι **Παρατηρήσεις (Individuals)**, οι **Κατηγορίες (Categories)**, οι **Επιπρόσθετες Παρατηρήσεις (Supplementary Individuals)** και οι **Επιπρόσθετες Κατηγορίες (Supplementary Categories)**.

5. Ο χρήστης εδώ μπορεί επίσης να διαλέξει αν θα σχεδιαστούν στο γράφημα οι ετικέτες των **Παρατηρήσεων (Individuals)**, των **Κατηγοριών (Categories)**, των **Επιπρόσθετων Παρατηρήσεων (Supplementary**



Σχημα 4.14: *Το γράφημα Individuals and categories.*

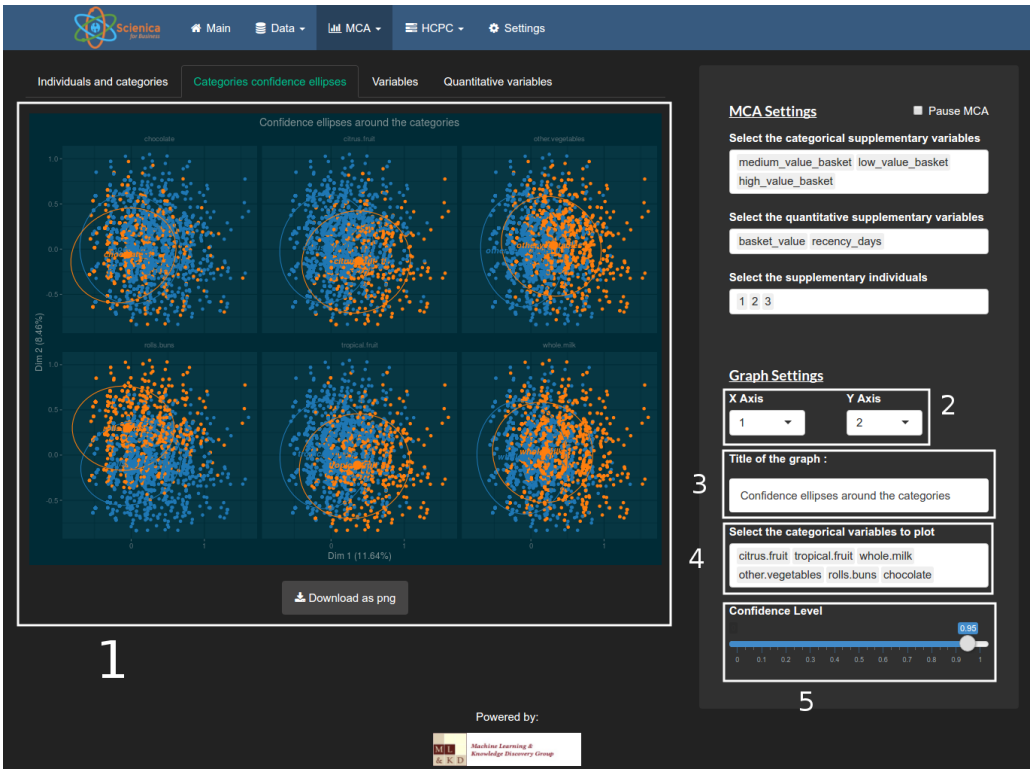
Individuals) και των **Επιπρόσθετων Κατηγόριων (Supplementary Categories)**.

6. Εδώ ο χρήστης μπορεί να φιλτράρει τις παρατηρήσεις και τις κατηγορίες με βάση τρεις επιλογές:

   - **Χειροκίνητα (Manual):** όπου ο χρήστης διαλέγει χειροκίνητα ποιες παρατηρήσεις και ποιες κατηγορίες θα σχεδιαστούν στο γράφημα.
   - **cos²:** όπου ο χρήστης διαλέγει ένα κατώφλι της τιμής cos^2 [0-1] όπου παρατηρήσεις ή και κατηγορίες που ξεπερνούν αυτό το κατώφλι θα σχεδιάζονται στο γράφημα. Το τετραγωνικό συνημίτονο βοηθάει στο να εντοπίζεται η σημαντικότητα των παρατηρήσεων και των κατηγοριών.
   - **Συνεισφορά (Contribution):** ο χρήστης μπορεί να διαλέξει να σχεδιάσει με βάση την συνεισφορά των παρατηρήσεων και των κατηγοριών ώστε να εμφανίζονται αυτές που δεν ξεπερνούν αυτό το όριο.

Categories confidence ellipses Σε αυτό το γράφημα ο χρήστης έχει τη δυνατότητα να παράγει το διάγραμμα εμπιστοσύνης των ποιοτικών μεταβλητών όπως φαίνεται στο σχήμα 4.16. Ο χρήστης με αυτό το γράφημα μπορεί να δει ποιες ποιοτικές μεταβλητές παρουσιάζουν τη μεγαλύτερη διακύμανση μεταξύ των διαφορετικών κατηγοριών επιλέγοντας ένα επίπεδο εμπιστοσύνης του ανάλογου κύκλου εμπιστοσύνης.

Στο σχήμα 4.15 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:



Σχημα 4.15: *Η οθόνη του γραφήματος Categories confidence ellipses της ανάλυσης MCA.*

1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Εδώ ο χρήστης μπορεί να αλλάξει ποια κύρια συνιστώσα θα σχεδιαστεί στο γράφημα για τον άξονα x και τον άξονα y.

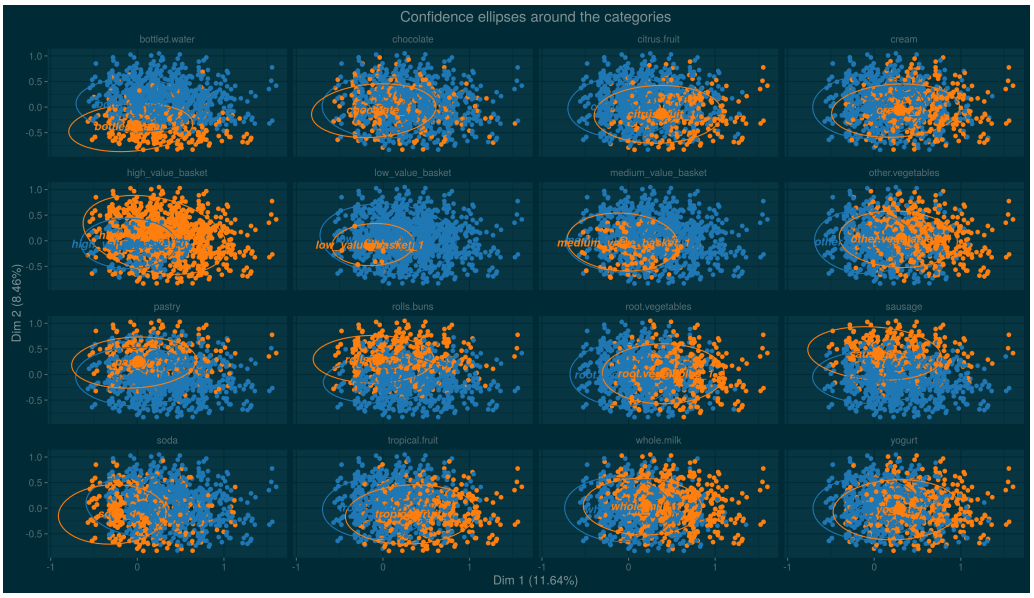
3. Ο χρήστης σε αυτήν την επιλογή έχει τη δυνατότητα να αλλάξει τον τίτλο του γραφήματος.

4. Ο χρήστης εδώ μπορεί να διαλέξει ποιες ποιοτικές μεταβλητές να σχεδιαστούν στο γράφημα όπως φαίνεται και στην εικόνα 4.15. Αν δεν επιλεγεί καμία τότε εμφανίζονται όλες.

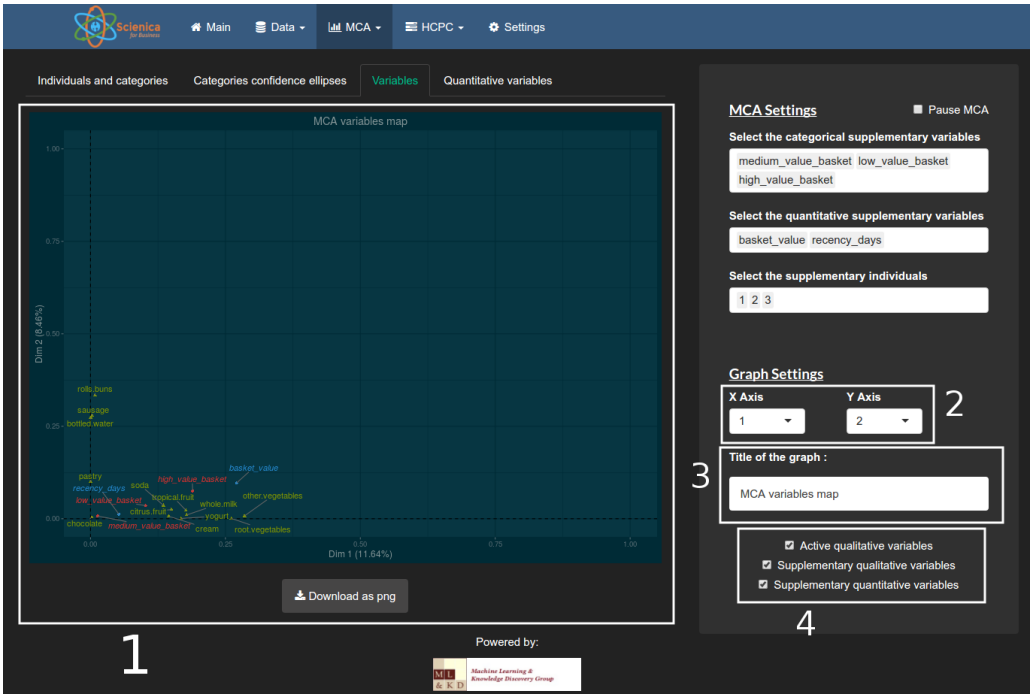
5. Εδώ ο χρήστης μπορεί να διαλέξει το επίπεδο εμπιστοσύνης των κύκλων εμπιστοσύνης στο γράφημα. Οι επιτρεπτές τιμές είναι [0-1].

Η συνάρτηση που δημιουργήθηκε γι΄ αυτό το σκοπό βρίσκεται στο Παράρτημα Πηγαίου Κώδικα και συγκεκριμένα στο πίνακα Α΄.2.

Variables Με αυτό το γράφημα ο χρήστης μπορεί να παράξει το **Χάρτη των μεταβλητών** που περιγράφτηκε στην Ενότητα 3.3.6 όπως φαίνεται στο σχήμα 4.18. Ο χρήστης με αυτό το γράφημα μπορεί να διακρίνει και να συγκρίνει τη συσχέτιση μεταξύ όλων των μεταβλητών.



Σχημα 4.16: *Το γράφημα Categories confidence ellipses.*

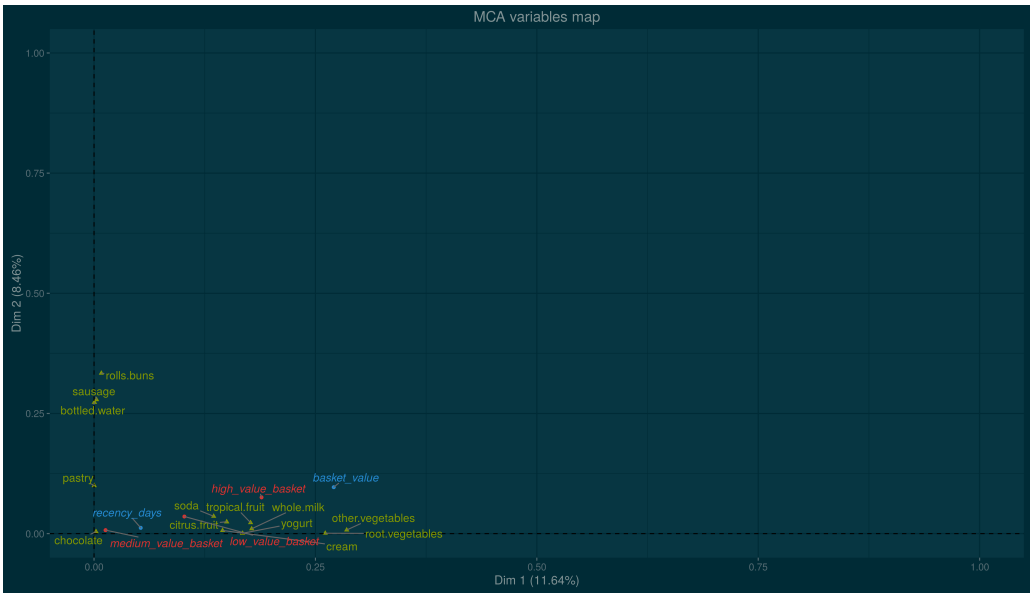


Σχημα 4.17: *Η οθόνη του γραφήματος Variables της ανάλυσης MCA.*

Στο σχήμα 4.17 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:

1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Εδώ ο χρήστης μπορεί να αλλάξει ποια κύρια συνιστώσα θα σχεδιαστεί



Σχημα 4.18: *Το γράφημα Variables.*

στο γράφημα για τον άξονα x και τον άξονα y.

3. Ο χρήστης σε αυτήν την επιλογή έχει τη δυνατότητα να αλλάξει τον τίτλο του γραφήματος.

4. Ο χρήστης εδώ μπορεί να διαλέξει ποιες μεταβλητές να σχεδιαστούν στο γράφημα. Οι επιλογές είναι τρεις: **Active qualitative variables**, **Supplementary qualitative variables** και **Supplementary quantitative variables**.

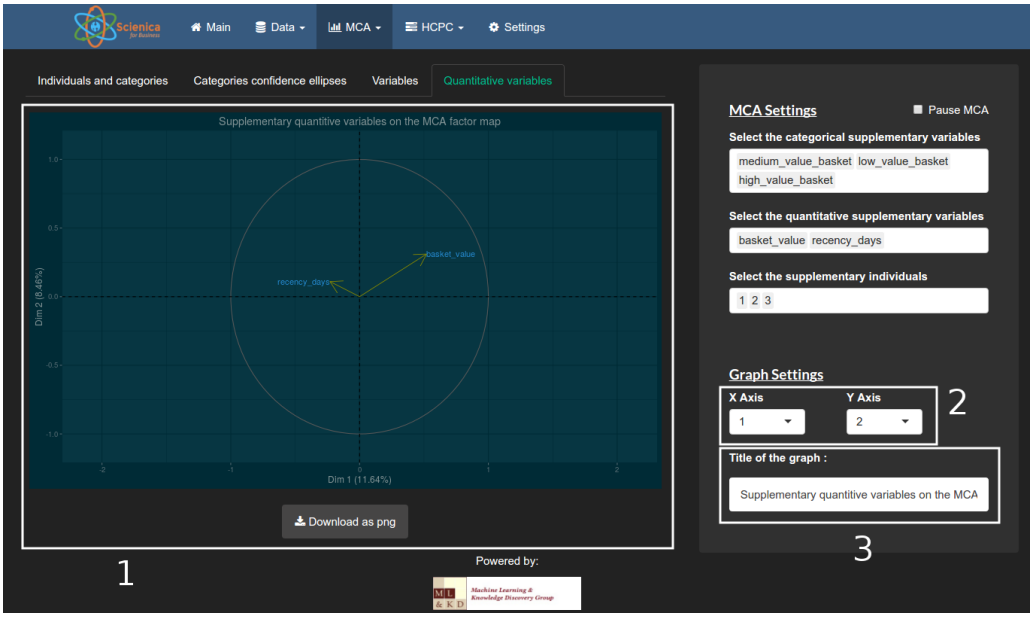
Η συνάρτηση που δημιουργήθηκε γι΄ αυτό το σκοπό βρίσκεται στο Παράρτημα Πηγαίου Κώδικα και συγκεκριμένα στο πίνακα Α΄.3.

Quantitative variables Σε αυτό το γράφημα ο χρήστης έχει τη δυνατότητα να παράγει το διάγραμμα της συσχέτιση μεταξύ των αριθμητικών μεταβλητών και των κύριων συνιστωσών τους που ονομάζονται loadings όπως φαίνεται στο σχήμα 4.18. Όταν ολοκληρωθεί η ανάλυση MCA γίνεται επίσης και μια ανάλυση PCA (Principal Component Analysis) ώστε οι αριθμητικές μεταβλητές να μεταφερθούν στο χώρο των συνιστωσών της MCA και να μπορούν να είναι συγκρίσιμες.

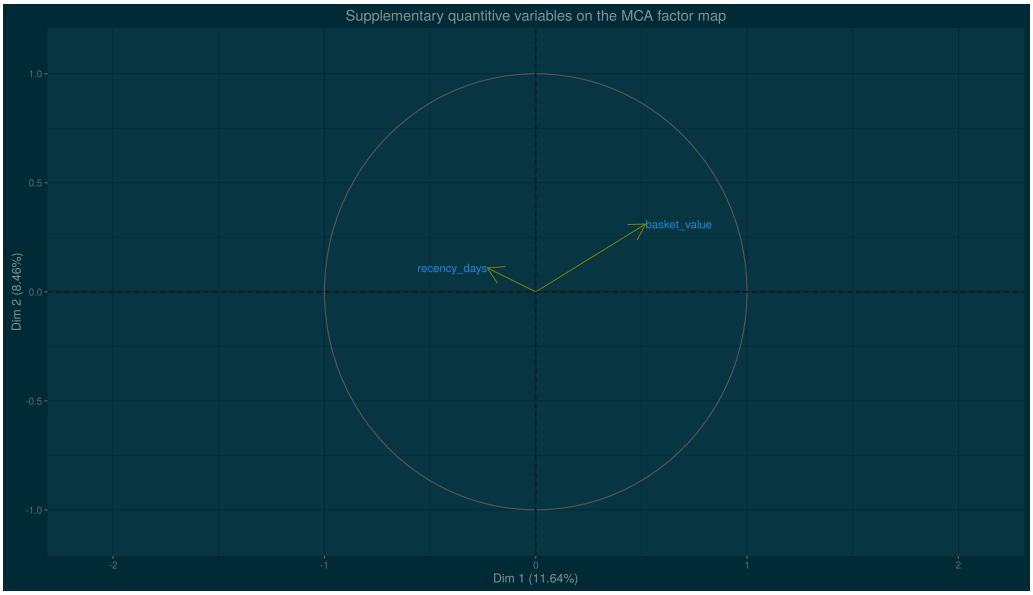
Στο σχήμα 4.17 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:

1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Εδώ ο χρήστης μπορεί να αλλάξει ποια κύρια συνιστώσα θα σχεδιαστεί στο γράφημα για τον άξονα x και τον άξονα y.



Σχημα 4.19: *Η οθόνη του γραφήματος Quantitative variables της ανάλυσης MCA.*



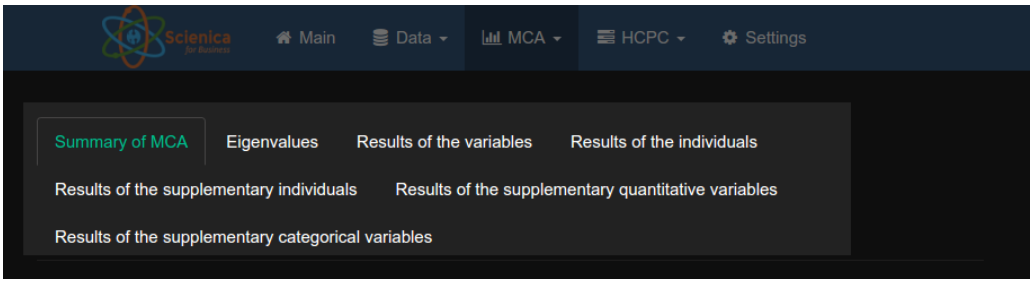
Σχημα 4.20: *Το γράφημα Quantitative variables.*

3. Ο χρήστης σε αυτήν την επιλογή έχει τη δυνατότητα να αλλάξει τον τίτλο του γραφήματος.

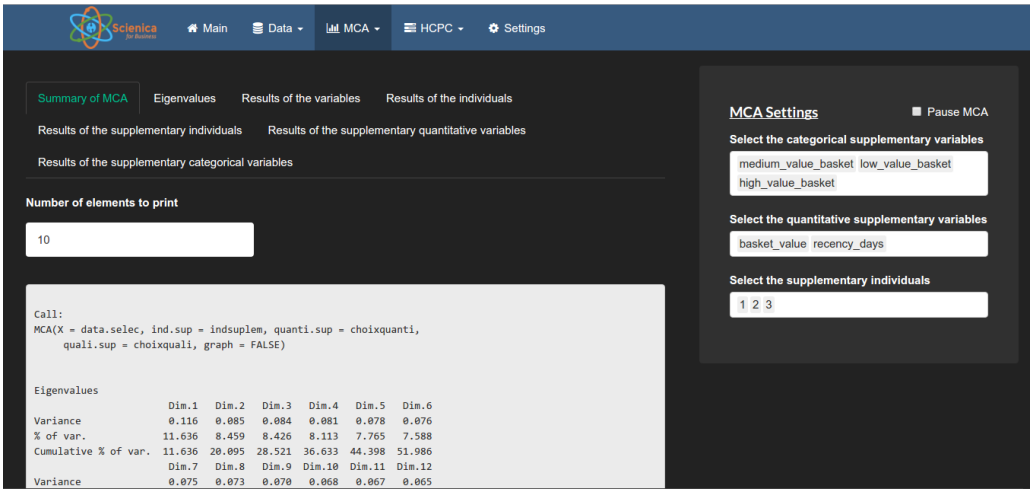
Η συνάρτηση που δημιουργήθηκε γι΄ αυτό το σκοπό βρίσκεται στο Παράρτημα Πηγαίου Κώδικα και συγκεκριμένα στο πίνακα Α΄.4.

4.3.2.5.2 Σύνοψη της ανάλυσης MCA (MCA Summaries)

Σε αυτήν την επιλογή ο χρήστης μπορεί να δει κάποιες πληροφορίες για την ανάλυση MCA όπως να διερευνήσει τις ιδιοτιμές των πρωτευόντων πα-



Σχημα 4.21: *Το μενού με τις επιλογές της σύνοψης MCA.*



Σχημα 4.22: *Η οθόνη της σύνοψης MCA.*

ραγόντων. Στο σχήμα 4.21 φαίνεται πως μπορεί να γίνει η πρόσβαση στην καρτέλα.

Summary of MCA

Σε αυτήν την επιλογή ο χρηστής μπορεί να δει μια σύνοψη των πληροφοριών όπως παράγεται από την εντολή `summary(res.mca)`. Εδώ ο χρήστης μπορεί πολύ γρήγορα να πληροφορηθεί για τις ιδιοτιμές των αξόνων, κάποιες πληροφορίες για τις 10 πρώτες (συμπληρωματικές) παρατηρήσεις, (συμπληρωματικές) μεταβλητές και κατηγορίες. Δίνεται επίσης η επιλογή να αλλάξει ο αριθμός των στοιχείων που εμφανίζονται (αρχικά 10) αλλά και να αποθηκεύσει ο χρήστης σε αρχείο κειμένου το αποτέλεσμα. Στο σχήμα 4.22 φαίνεται ένα παράδειγμα του παράθυρου διαλόγου της εφαρμογής και παρακάτω το πλήρες αποτέλεσμα που εμφανίζεται.

```
Call:
MCA(X = data.selec, ind.sup = indsuplem, quanti.sup = choixquanti,
     quali.sup = choixquali, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
                       ↪  Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13
```

```
Variance               0.116   0.085   0.084   0.081   0.078   0.076   0.075
↪  0.073   0.070   0.068   0.067   0.065   0.061
% of var.             11.636   8.459   8.426   8.113   7.765   7.588   7.529
↪  7.259   6.976   6.844   6.745   6.513   6.148
Cumulative % of var.  11.636  20.095  28.521  36.633  44.398  51.986  59.515
↪  66.774  73.750  80.594  87.339  93.852 100.000

Individuals (the 10 first)
                       Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
                       ↪  cos2
4                    |  0.192  0.004  0.040 | -0.147  0.003  0.023 | -0.080  0.001
↪  0.007 |
5                    |  0.201  0.005  0.075 |  0.038  0.000  0.003 | -0.385  0.023
↪  0.275 |
6                    |  0.163  0.003  0.040 | -0.033  0.000  0.002 | -0.157  0.004
↪  0.037 |
7                    | -0.326  0.012  0.240 |  0.287  0.013  0.187 |  0.033  0.000
↪  0.002 |
8                    | -0.048  0.000  0.004 |  0.340  0.018  0.180 |  0.008  0.000
↪  0.000 |
9                    | -0.076  0.001  0.017 | -0.014  0.000  0.001 | -0.359  0.020
↪  0.383 |
10                   |  0.228  0.006  0.027 | -0.582  0.053  0.175 |  0.237  0.009
↪  0.029 |
11                   |  0.708  0.057  0.227 | -0.677  0.072  0.207 |  0.657  0.068
↪  0.195 |
12                   | -0.523  0.031  0.403 |  0.167  0.004  0.041 |  0.250  0.010
↪  0.092 |
13                   | -0.002  0.000  0.000 | -0.187  0.005  0.050 |  0.089  0.001
↪  0.011 |

Supplementary individuals
                       Dim.1   cos2    Dim.2   cos2    Dim.3   cos2
1                    |  0.004  0.000 | -0.204  0.049 |  0.045  0.002 |
2                    |  0.238  0.055 | -0.205  0.041 |  0.291  0.083 |
3                    | -0.076  0.017 | -0.014  0.001 | -0.359  0.383 |

Categories (the 10 first)
                        Dim.1     ctr    cos2  v.test     Dim.2     ctr    cos2
                        ↪  v.test     Dim.3     ctr    cos2  v.test
citrus.fruit_0        |  -0.135   1.069   0.150 -33.591 |   0.055   0.242   0.025
↪  13.617 |  -0.089   0.647   0.066 -22.246 |
citrus.fruit_1        |   1.113   8.835   0.150  33.591 |  -0.451   1.997   0.025
↪  -13.617 |   0.737   5.351   0.066  22.246 |
tropical.fruit_0      |  -0.167   1.599   0.177 -36.486 |   0.060   0.283   0.023
↪  13.091 |  -0.136   1.455   0.116 -29.617 |
tropical.fruit_1      |   1.056  10.084   0.177  36.486 |  -0.379   1.786   0.023
↪  -13.091 |   0.857   9.177   0.116  29.617 |
whole.milk_0          |  -0.298   3.924   0.178 -36.615 |  -0.071   0.305   0.010
↪  -8.700 |   0.233   3.314   0.109  28.634 |
whole.milk_1          |   0.596   7.843   0.178  36.615 |   0.142   0.609   0.010
↪  8.700 |  -0.466   6.624   0.109 -28.634 |
other.vegetables_0    |  -0.311   4.764   0.285 -46.356 |  -0.050   0.173   0.008
↪  -7.536 |   0.024   0.039   0.002   3.577 |
other.vegetables_1    |   0.919  14.096   0.285  46.356 |   0.149   0.512   0.008
↪  7.536 |  -0.071   0.116   0.002  -3.577 |
rolls.buns_0          |   0.051   0.131   0.008   7.876 |  -0.325   7.288   0.334
↪  -50.136 |  -0.188   2.444   0.111 -28.974 |
rolls.buns_1          |  -0.161   0.414   0.008  -7.876 |   1.028  23.061   0.334
↪  50.136 |   0.594   7.732   0.111  28.974 |

Categorical variables (eta2)
                       Dim.1 Dim.2 Dim.3
citrus.fruit         | 0.150 0.025 0.066 |
tropical.fruit       | 0.177 0.023 0.116 |
whole.milk           | 0.178 0.010 0.109 |
other.vegetables     | 0.285 0.008 0.002 |
rolls.buns           | 0.008 0.334 0.111 |
chocolate            | 0.002 0.004 0.036 |
bottled.water        | 0.000 0.273 0.194 |
```

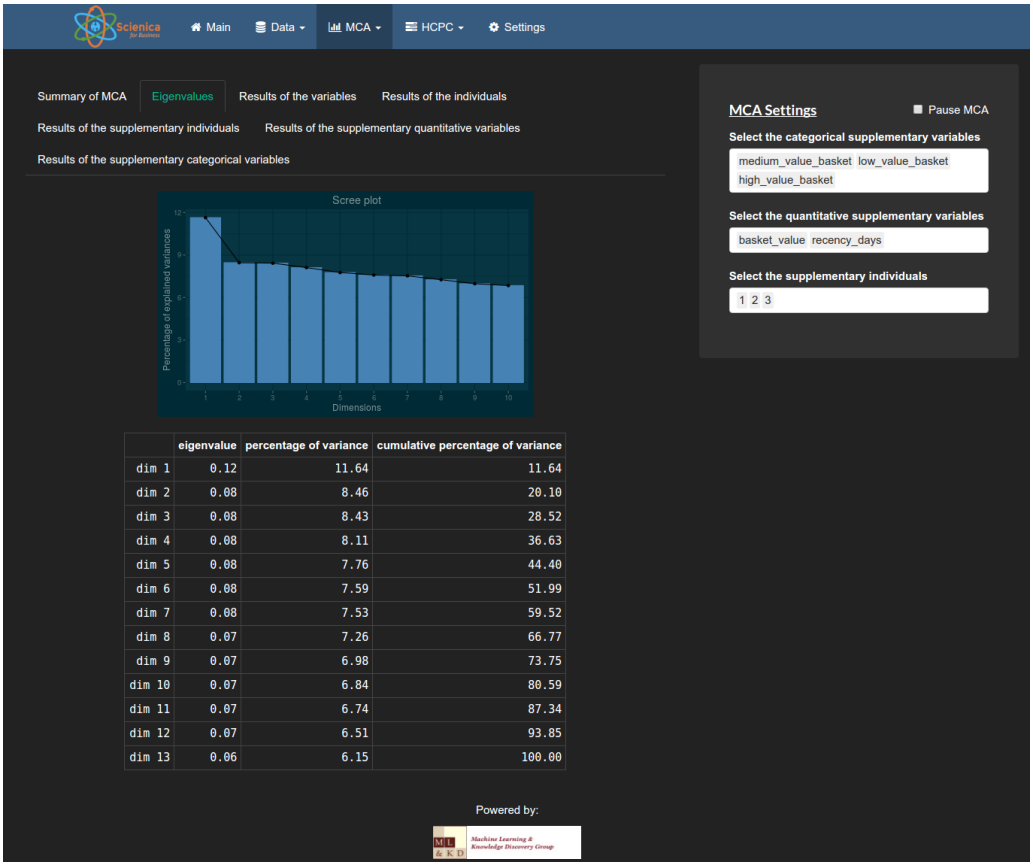
```
yogurt                | 0.168 0.001 0.087 |
sausage               | 0.003 0.279 0.222 |
root.vegetables       | 0.261 0.001 0.000 |

Supplementary categories
                         Dim.1    cos2  v.test     Dim.2    cos2  v.test     Dim.3
                         ↪   cos2  v.test
medium_value_basket_0 |   0.079   0.013   9.902 |   0.059   0.007   7.406 |   0.060
↪   0.007   7.504 |
medium_value_basket_1 |  -0.164   0.013  -9.902 |  -0.123   0.007  -7.406 |  -0.124
↪   0.007  -7.504 |
low_value_basket_0    |   0.230   0.102  27.717 |   0.136   0.036  16.377 |   0.129
↪   0.032  15.608 |
low_value_basket_1    |  -0.444   0.102 -27.717 |  -0.262   0.036 -16.377 |  -0.250
↪   0.032 -15.608 |
high_value_basket_0   |  -0.307   0.189 -37.746 |  -0.194   0.076 -23.851 |  -0.188
↪   0.071 -23.175 |
high_value_basket_1   |   0.616   0.189  37.746 |   0.389   0.076  23.851 |   0.378
↪   0.071  23.175 |

Supplementary categorical variables (eta2)
                       Dim.1 Dim.2 Dim.3
medium_value_basket   | 0.013 0.007 0.007 |
low_value_basket      | 0.102 0.036 0.032 |
high_value_basket     | 0.189 0.076 0.071 |

Supplementary continuous variables
                         Dim.1    Dim.2    Dim.3
basket_value          |  0.520 |  0.311 |  0.327 |
recency_days          | -0.230 |  0.110 | -0.069 |
```

Eigenvalues

Εδώ ο χρηστής μπορεί να δει όλες τις πληροφορίες σχετικά με τις ιδιοτιμές των κύριων συνιστωσών που προκύπτουν από την ανάλυση MCA όπως η τιμή της ιδιοτιμής, το ποσοστό της διακύμανσης αλλά και το αθροιστικό ποσοστό της διακύμανσης για πιο εύκολη σύγκριση. Επίσης μπορεί να δει το πως διαμορφώνεται το ποσοστό της διακύμανσης σε ένα bar γράφημα. Στο σχήμα 4.23 φαίνεται ένα παράδειγμα του παράθυρου διαλόγου της εφαρμογής.

Results of the variables

Σε αυτήν την επιλογή μπορεί ο χρήστης να δει όλες τις πληροφορίες σχετικά με τις κατηγορίες των ποιοτικών μεταβλητών που έγινε η ανάλυση MCA. Ο χρήστης μπορεί να πάρει πληροφορίες σχετικά για κάθε κατηγορία σε κάθε κύρια συνιστώσα και πιο συγκεκριμένα αφορούν τις τιμές για τις **Συντεταγμένες (Coordinates)** όπου ο χρήστης μπορεί να συγκρίνει ποιοι άξονες έχουν μεγαλύτερη διακύμανση, τις **Συνεισφορές (Contributions)** όπου φαίνεται η συνεισφορά κάθε κατηγορίας, τις τιμές **cos²** που ο χρήστης μπορεί μέσω του τετραγωνικού συνημίτονου να εντοπίζει τη σημαντικότητα των κατηγοριών, τις τιμές **v.test** όπου αφορά ένα κριτήριο με μια κανονική κατανομή και τέλος οι τιμές **eta²** όπου αφορούν τη τετραγωνική συσχέτιση μεταξύ των μεταβλητών και των αξόνων. Οι πληροφορίες δίνονται σε διαγράμματα για να μπορούν να διακριθούν εύκολα οι τιμές εκείνες που είναι σημαντικές. Ο χρήστης αν επιθυμεί μπορεί να προβάλει και τις αριθμητικές τιμές αν θέλει στο



	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.12	11.64	11.64
dim 2	0.08	8.46	20.10
dim 3	0.08	8.43	28.52
dim 4	0.08	8.11	36.63
dim 5	0.08	7.76	44.40
dim 6	0.08	7.59	51.99
dim 7	0.08	7.53	59.52
dim 8	0.07	7.26	66.77
dim 9	0.07	6.98	73.75
dim 10	0.07	6.84	80.59
dim 11	0.07	6.74	87.34
dim 12	0.07	6.51	93.85
dim 13	0.06	6.15	100.00

Σχημα 4.23: *Η οθόνη της σύνοψης των ιδιοτιμών.*

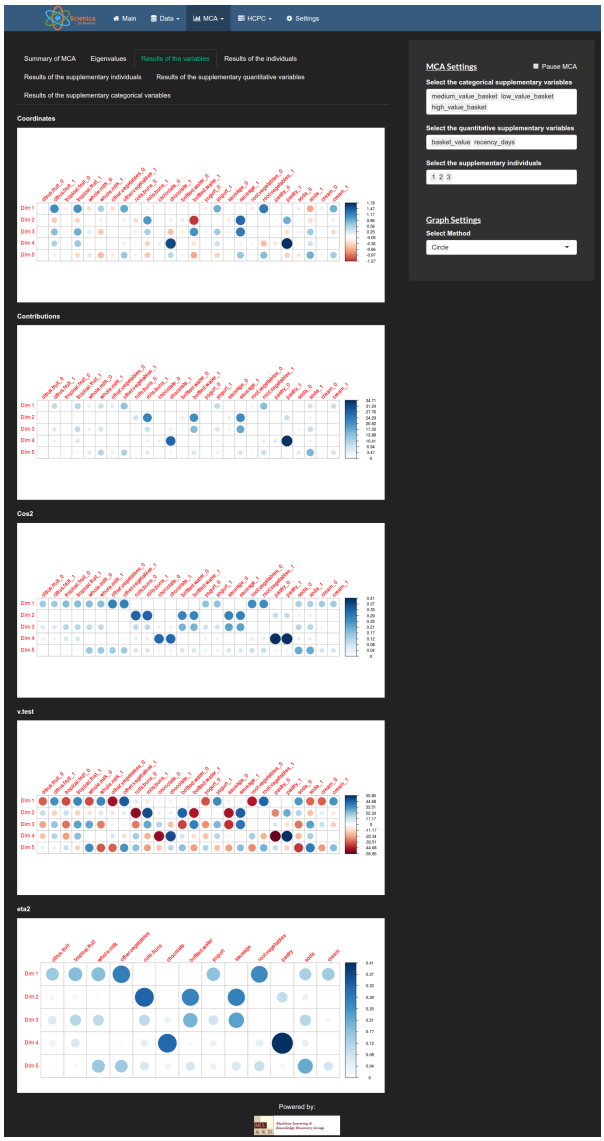
γράφημα. Στο σχήμα 4.24 φαίνεται ένα παράδειγμα του παράθυρου διαλόγου της εφαρμογής.

Results of the individuals

Εδώ ο χρήστης είναι να πάρει πληροφορίες σχετικά με τις ενεργές παρατηρήσεις του συνόλου δεδομένων. Δεδομένου ότι οι παρατηρήσεις σε τέτοια σύνολα είναι πολλές επιλέχθηκε η παρουσίαση των αποτελεσμάτων σε πίνακες δεδομένων. Ο χρήστης μπορεί να δει πληροφορίες για τις **Συντεταγμένες (Coordinates)**, τις **Συνεισφορές (Contributions)** και τις τιμές $\mathbf{cos^2}$. Στο σχήμα 4.25 φαίνεται ένα παράδειγμα του παράθυρου διαλόγου της εφαρμογής.

Results of the supplementary individuals

Ο χρήστης εδώ είναι να πάρει επίσης πληροφορίες σχετικά με τις επιπρόσθετες παρατηρήσεις του συνόλου δεδομένων και συγκεκριμένα μπορεί να δει πληροφορίες για τις **Συντεταγμένες (Coordinates)** και τις τιμές $\mathbf{cos^2}$. Στο σχήμα 4.26 φαίνεται ένα παράδειγμα του παράθυρου διαλόγου της εφαρμογής.



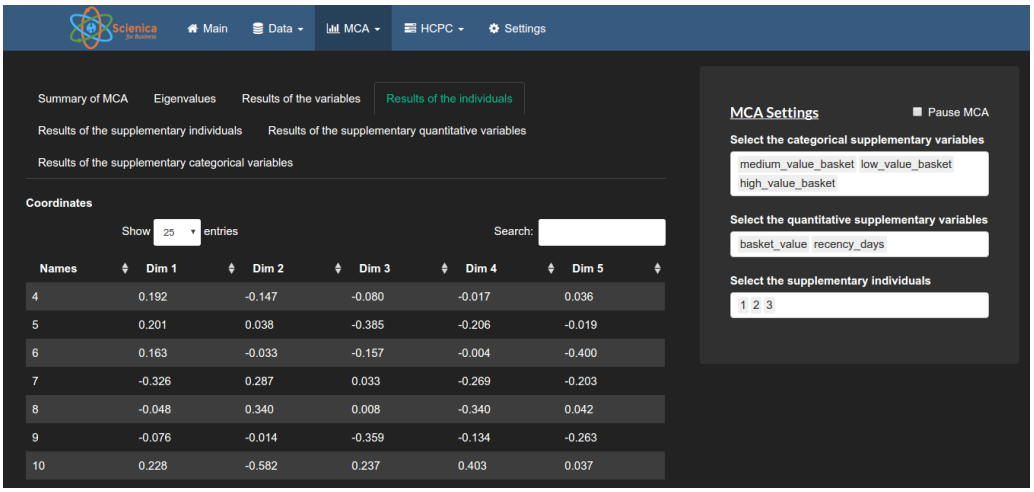
Σχημα 4.24: *Η οθόνη της σύνοψης Results of the variables.*

Results of the supplementary quantitative variables

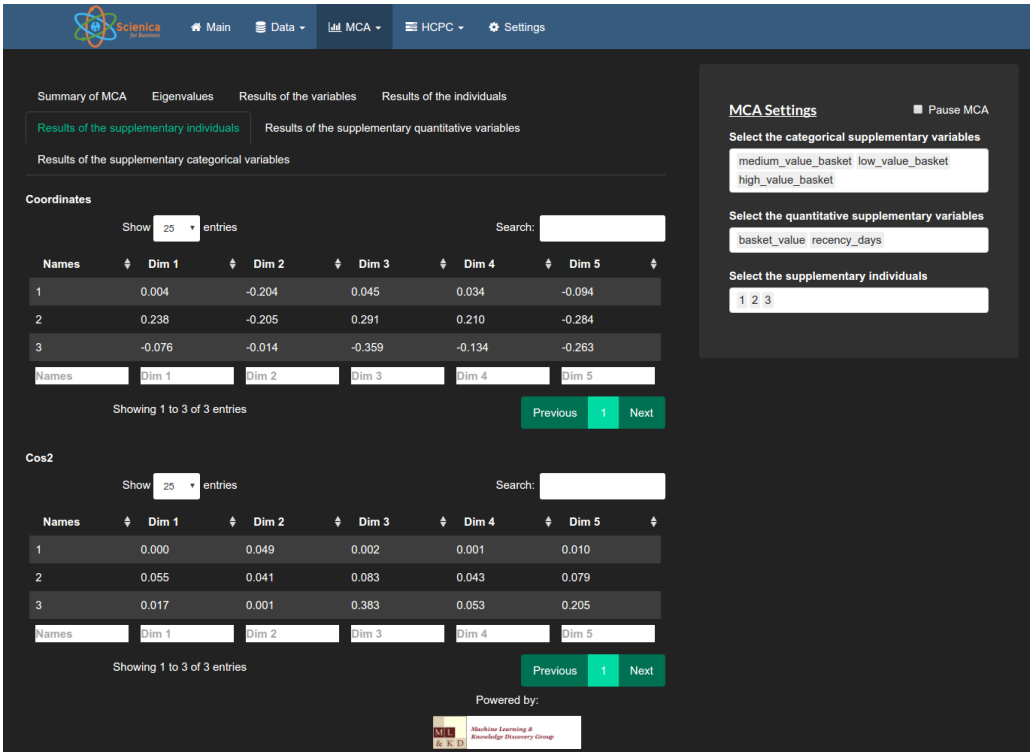
Ο χρήστης και εδώ μπορεί να πάρει πληροφορίες σχετικά με τις επιπρόσθετες αριθμητικές μεταβλητές από το σύνολο των δεδομένων και συγκεκριμένα μπορεί να δει πληροφορίες για τις **Συντεταγμένες (Coordinates)**. Στο σχήμα 4.27 φαίνεται ένα παράδειγμα του παράθυρου διαλόγου της εφαρμογής.

Results of the supplementary categorical variables

Τέλος ο χρήστης εδώ μπορεί να πάρει πληροφορίες σχετικά με τις επιπρόσθετες ποιοτικές μεταβλητές από το σύνολο των δεδομένων και συγκεκριμένα μπορεί να δει πληροφορίες για τις **Συντεταγμένες (Coordinates)**, τις τιμές **cos²**, τις τιμές **v.test** και τις τιμές **eta²**. Στο σχήμα 4.28 φαίνεται ένα παράδειγμα του παράθυρου διαλόγου της εφαρμογής.



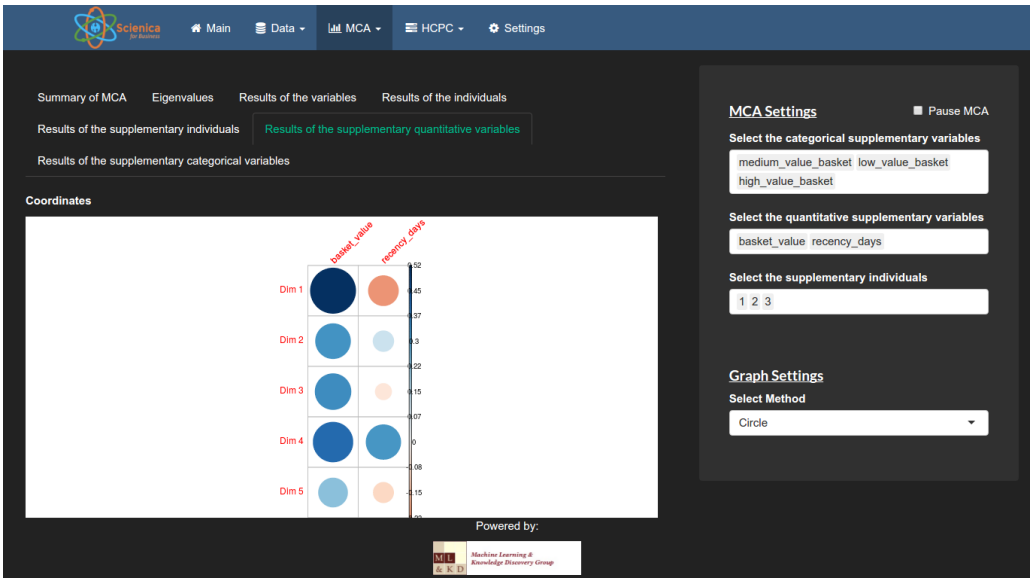
Σχημα 4.25: *Η οθόνη της σύνοψης Results of the individuals.*



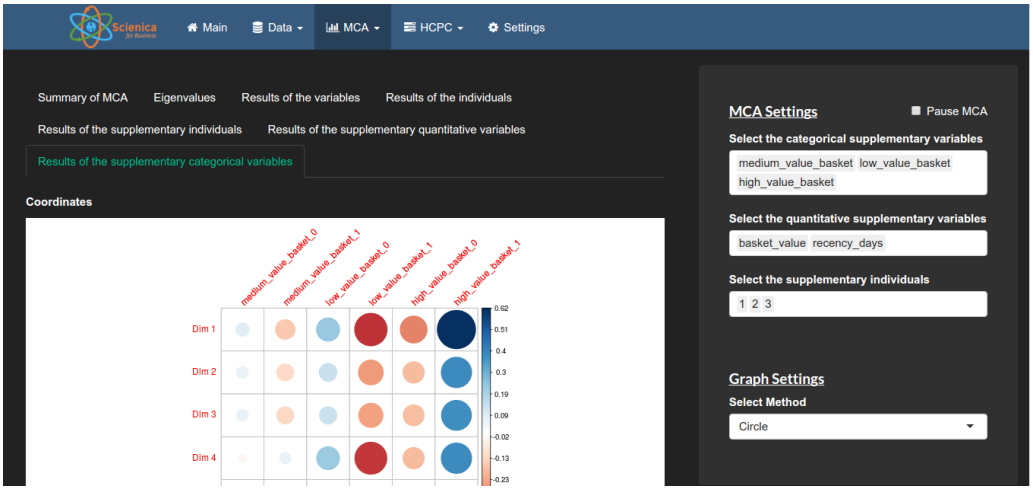
Σχημα 4.26: *Η οθόνη της σύνοψης Results of the supplementary individuals.*

4.3.2.5.3 Αυτόματη περιγραφή αξόνων (Automatic description of axes)

Σε αυτήν την επιλογή ο χρήστης μπορεί να πάρει μια πρώτη περιγραφή των αξόνων που παράχθηκαν από την ανάλυση. Η επιλογή αυτή είναι ιδιαίτερα χρήσιμη όταν υπάρχουν πολλές μεταβλητές στην ανάλυση. Επιτρέπει στο χρήστη να δει με ποιες μεταβλητές είναι οι άξονες περισσότερο συσχετισμένοι. Ποιο συγκεκριμένα, ποιες ποσοτικές μεταβλητές έχουν μεγαλύτερη συσχέτιση με κάθε άξονα και ποιες ποιοτικές μεταβλητές και κατηγορίες περιγράφουν κα-



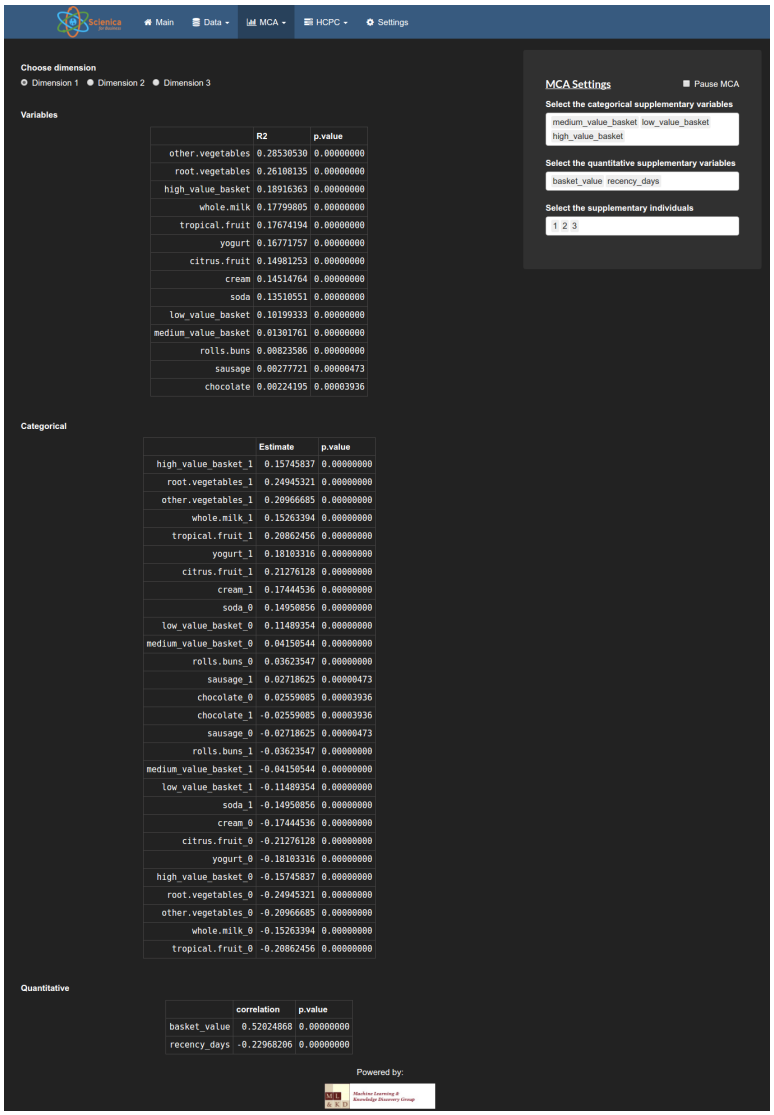
Σχημα 4.27: *Η οθόνη της σύνοψης Results of the supplementary quantitative variables.*



Σχημα 4.28: *Η οθόνη της σύνοψης Results of the supplementary categorical variables.*

λύτερα κάθε άξονα. Σε όλες τις περιπτώσεις δύνεται η p-value του αντίστοιχου test. Στο σχήμα 4.29 φαίνεται πως μπορεί να γίνει η πρόσβαση στην καρτέλα.

Για τις ποιοτικές μεταβλητές εφαρμόζεται το μοντέλο **One-Factor ANOVA**[25] για κάθε διάσταση με τις συντεταγμένες των παρατηρήσεων να εξηγούνται από τη ποιοτική μεταβλητή. Ένα **F-test**[25] εφαρμόζεται για να αποφανθεί αν η μεταβλητή επηρεάζει τη διάσταση και επιπλέον εφαρμόζονται κάποια **T-tests**[25] ανά κατηγορία. Μπορούμε να δούμε αν οι συντεταγμένες των παρατηρήσεων ενός μέρους του πληθυσμού που ορίζεται από μια κατηγορία είναι σημαντικά διαφορετικές από τη συνολική εικόνα. Οι μεταβλητές και οι κατηγορίες κατατάσσονται σύμφωνα με **p-value** και μόνο οι σημαντικές από αυτές διατηρούνται.



Σχημα 4.29: *Η οθόνη της αυτόματης περιγραφής αξόνων.*

Η συνάρτηση του πακέτου FactoMineR που είναι υπεύθυνη για να γίνει αυτή αυτή η ενέργεια είναι η παρακάτω:

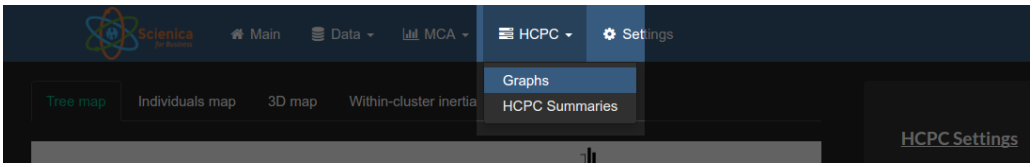
```
1 dimdesc(res, axes = 1:3, proba = 0.05)
```

4.3.2.6 Ανάλυση HCPC

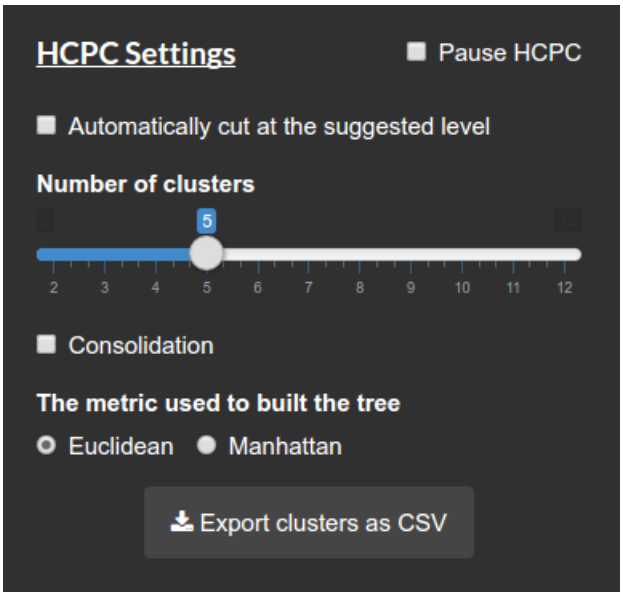
Με την επιλογή HCPC από το μενού πλοήγησης ο χρήστης μπορεί εύκολα να πραγματοποιήσει την ανάλυση επιλέγοντας μία από τις επιλογές: **Graphs (Γραφήματα Ανάλυσης HCPC)**, **HCPC Summaries (Σύνοψη της ανάλυσης HCPC)** όπως φαίνεται και στο σχήμα 4.30.

Σε όλες τις παραπάνω επιλογές δίνεται η δυνατότητα άμεσης αλλαγής των ρυθμίσεων της ανάλυσης HCPC όπως φαίνεται στο σχήμα 4.31.

Οι επιλογές είναι αναλυτικά:



Σχημα 4.30: *Το κεντρικό μενού της ανάλυσης HCPC.*



Σχημα 4.31: *Οι ρυθμίσεις της ανάλυσης HCPC.*

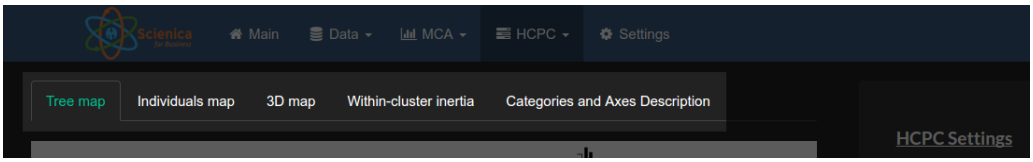
Automatically cut at the suggested level: Ο χρήστης εδώ μπορεί να επιλέξει είτε σε αυτόματη επιλογή του αριθμού των ομάδων ανάλογα με την αδράνεια ή να επιλέξει αυτός χειροκίνητα τον αριθμό.

Consolidation: Εδώ δίνεται η δυνατότητα στο χρήστη να 'εδραιώσει' consolidate τα αποτελέσματα του Ιεραρχικού αλγορίθμου εφαρμόζοντας έναν k-Means μετά από αυτόν ώστε να ανατεθούν σε περισσότερο συμπαγείς ομάδες οι πελάτες.

The metric used to built the tree: Εδώ μπορεί ο χρήστης να επιλέξει τη μετρική που θα χρησιμοποιηθεί για τον Ιεραρχικό Αλγόριθμο Ομαδοποίησης ανάμεσα σε δύο επιλογές: Ευκλείδεια απόσταση και απόσταση Manhattan.

Export clusters as CSV: Ο χρήστης με αυτήν την επιλογή μπορεί να κατεβάσει ένα αρχείο με τα αποτελέσματα της ομαδοποίησης στη μορφή που έχει επιλέξει στις ρυθμίσεις. Πρόκειται για δεδομένα που εισήγαγε με την προσθήκη μιας ακόμα στήλης που δηλώνει την ομάδα που ανήκει κάθε πελάτης/παρατήρηση.

Pause HCPC: Ο χρήστης έχοντας αυτή την επιλογή ενεργοποιημένη, κάθε αλλαγή στις ρυθμίσεις δεν θα έχει άμεσο υπολογισμό της ανάλυσης αλλά θα γίνεται όταν απενεργοποιηθεί.



Σχημα 4.32: *Οι επιλογές των γραφημάτων HCPC.*

Σχημα 4.33: *Η οθόνη του γραφήματος Tree Map.*

Στη συνέχεια παρουσιάζονται αναλυτικά κάθε λειτουργία και παράθυρο που προσφέρει η ανάλυση HCPC.

4.3.2.6.1 Γραφήματα Ανάλυσης HCPC (HCPC Graphs)

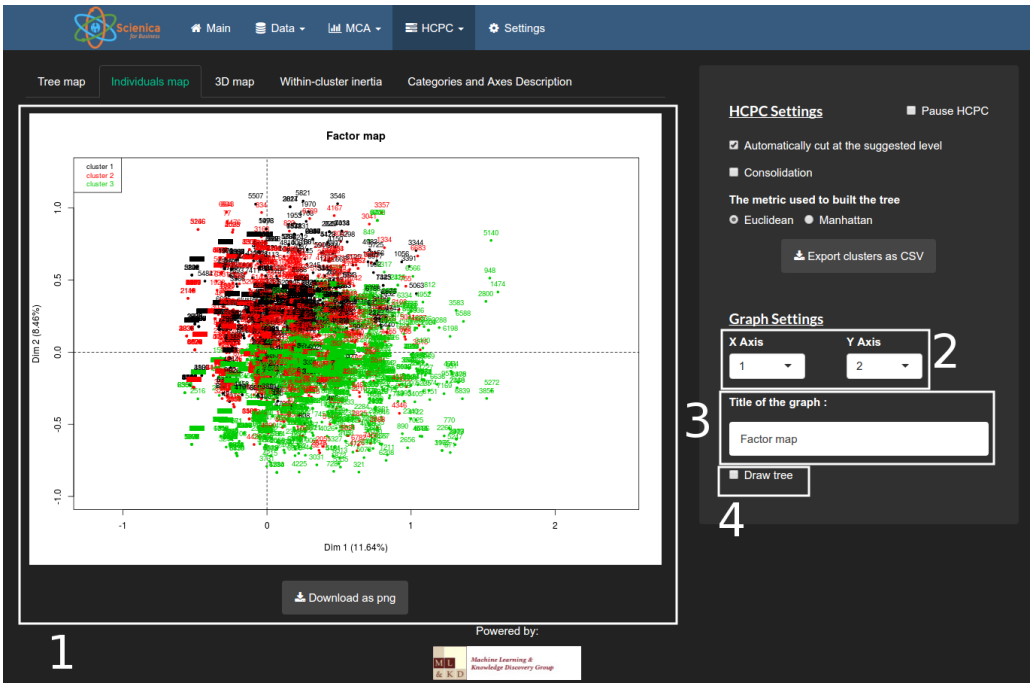
Τα διαθέσιμα γραφήματα της ανάλυσης HCPC είναι πέντε και μπορούν να πλοηγηθούν από το μενού του σχήματος 4.32.

Παρακάτω παρουσιάζεται κάθε γράφημα με τις λειτουργίες του ξεχωριστά.

Tree Map Σε αυτήν την επιλογή γραφήματος ο χρήστης έχει τη δυνατότητα να δει το γράφημα του δέντρου του ιεραρχικού αλγορίθμου μαζί με το προτεινόμενο ύψος για να 'κοπεί' το δέντρο. Επιπλέον ο χρήστης μπορεί να δει το η αδράνεια που χρησιμοποιήθηκε για να κοπεί το δέντρο.

Στο σχήμα 4.33 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:

1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)



Σχημα 4.34: *Η οθόνη του γραφήματος Individuals map.*

2. Ο χρήστης σε αυτήν την επιλογή έχει τη δυνατότητα να αλλάξει τον τίτλο του γραφήματος.

Individuals map Εδώ ο χρήστης μπορεί να δει ένα γράφημα των παρατηρήσεων όπου ομαδοποιούνται χρωματικά ανάλογα την ομάδα που ανήκουν.

Στο σχήμα 4.34 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:

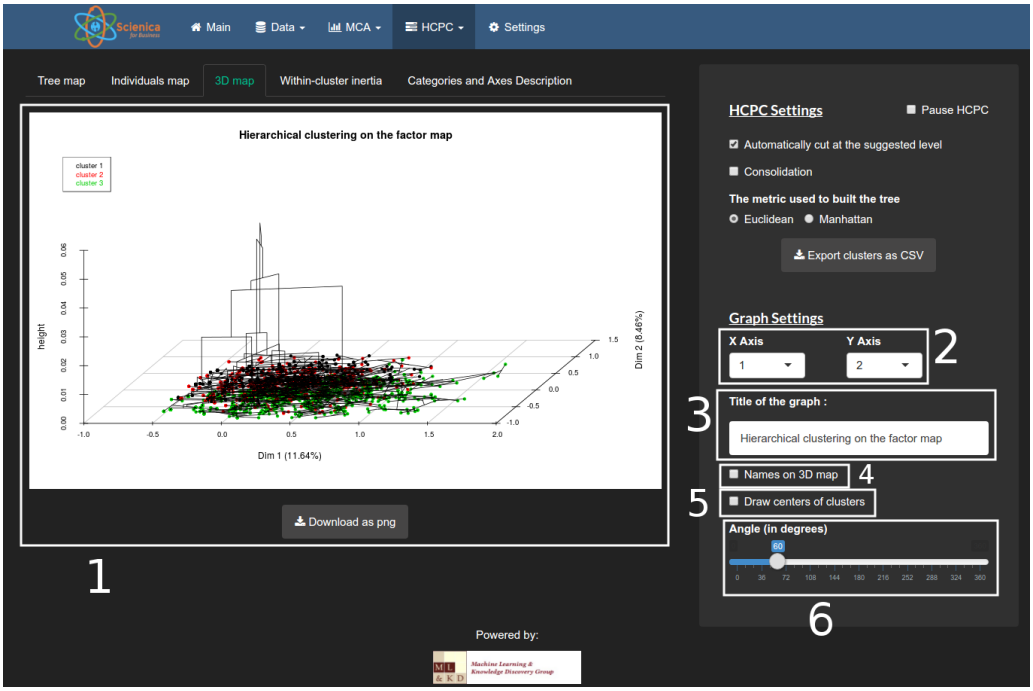
1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Εδώ ο χρήστης μπορεί να αλλάξει ποια κύρια συνιστώσα θα σχεδιαστεί στο γράφημα για τον άξονα x και τον άξονα y.

3. Ο χρήστης σε αυτήν την επιλογή έχει τη δυνατότητα να αλλάξει τον τίτλο του γραφήματος.

4. Με αυτήν την επιλογή ο χρήστης μπορεί να σχεδιάσει το δέντρο του ιεραρχικού αλγορίθμου πάνω από τις παρατηρήσεις.

3D map Με αυτό το γράφημα μπορεί ο χρήστης να δει πως ομαδοποιούνται οι παρατηρήσεις στο χώρο. Προβάλλονται τρισδιάστατα χρωματισμένες ανάλογα με την ομάδα που ανήκουν.



Σχημα 4.35: *Η οθόνη του γραφήματος 3D map.*

Στο σχήμα 4.35 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:

1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Εδώ ο χρήστης μπορεί να αλλάξει ποια κύρια συνιστώσα θα σχεδιαστεί στο γράφημα για τον άξονα x και τον άξονα y.

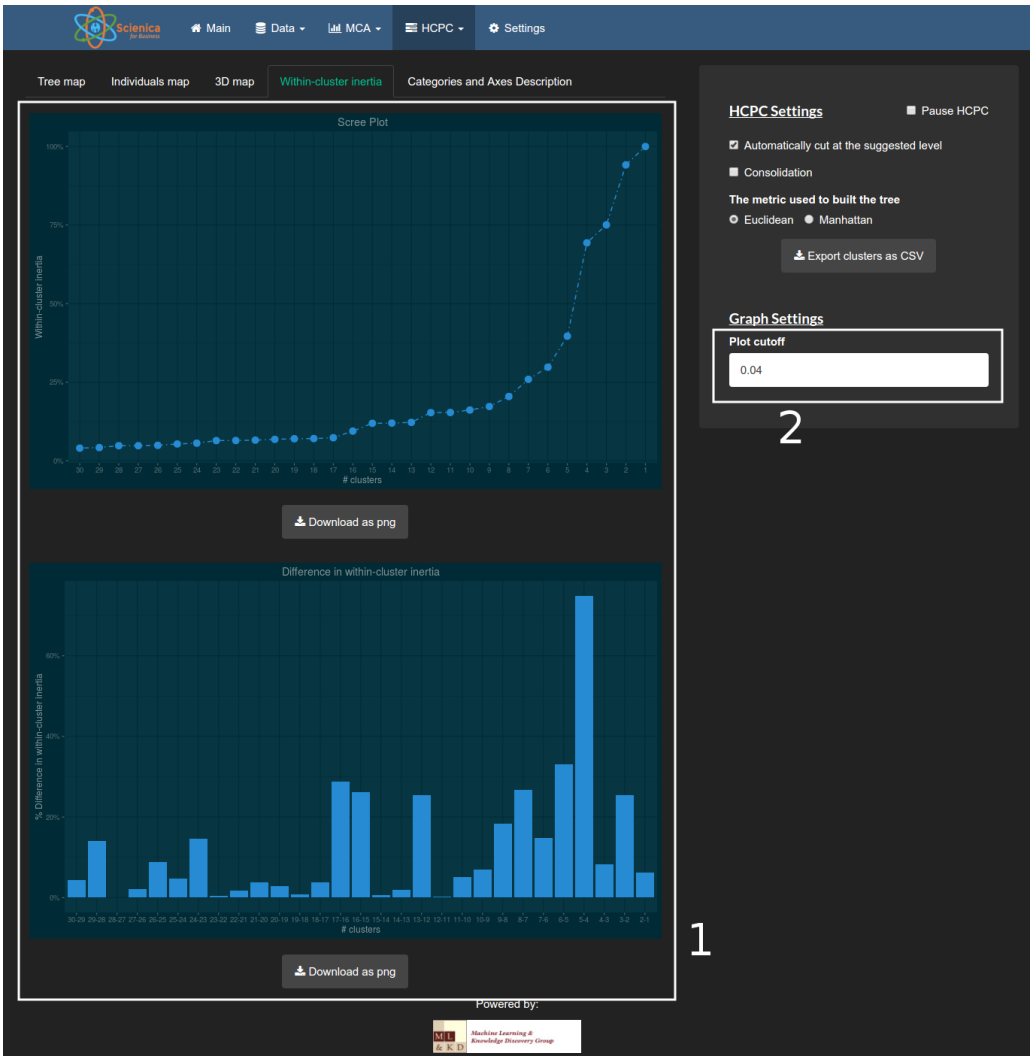
3. Ο χρήστης σε αυτήν την επιλογή έχει τη δυνατότητα να αλλάξει τον τίτλο του γραφήματος.

4. Ο χρήστης με αυτήν την επιλογή μπορεί να σχεδιάσει τα ονόματα των παρατηρήσεων.

5. Εδώ ο χρήστης μπορεί να σχεδιάσει ξεχωριστά τα κέντρα κάθε κλάσης.

6. Με αυτήν την επιλογή ο χρήστης μπορεί να επιλέξει τη γωνία όπου σχεδιάζεται το τρισδιάστατο γράφημα.

Within-cluster inertia Με αυτά τα δύο γραφήματα ο χρήστης είναι σε θέση να δει αρχικά πως εξελίσσεται η ενδο-αδράνεια μεταξύ των ομάδων εξελίσσεται ανάλογα τον αριθμό των ομάδων και στο δεύτερο φαίνεται η διαφορά της αδράνειας μεταξύ των αριθμών των ομάδων που επιλέγονται στον ιεραρχικό αλγόριθμο.



Σχημα 4.36: *Η οθόνη του γραφήματος Within-cluster inertia.*

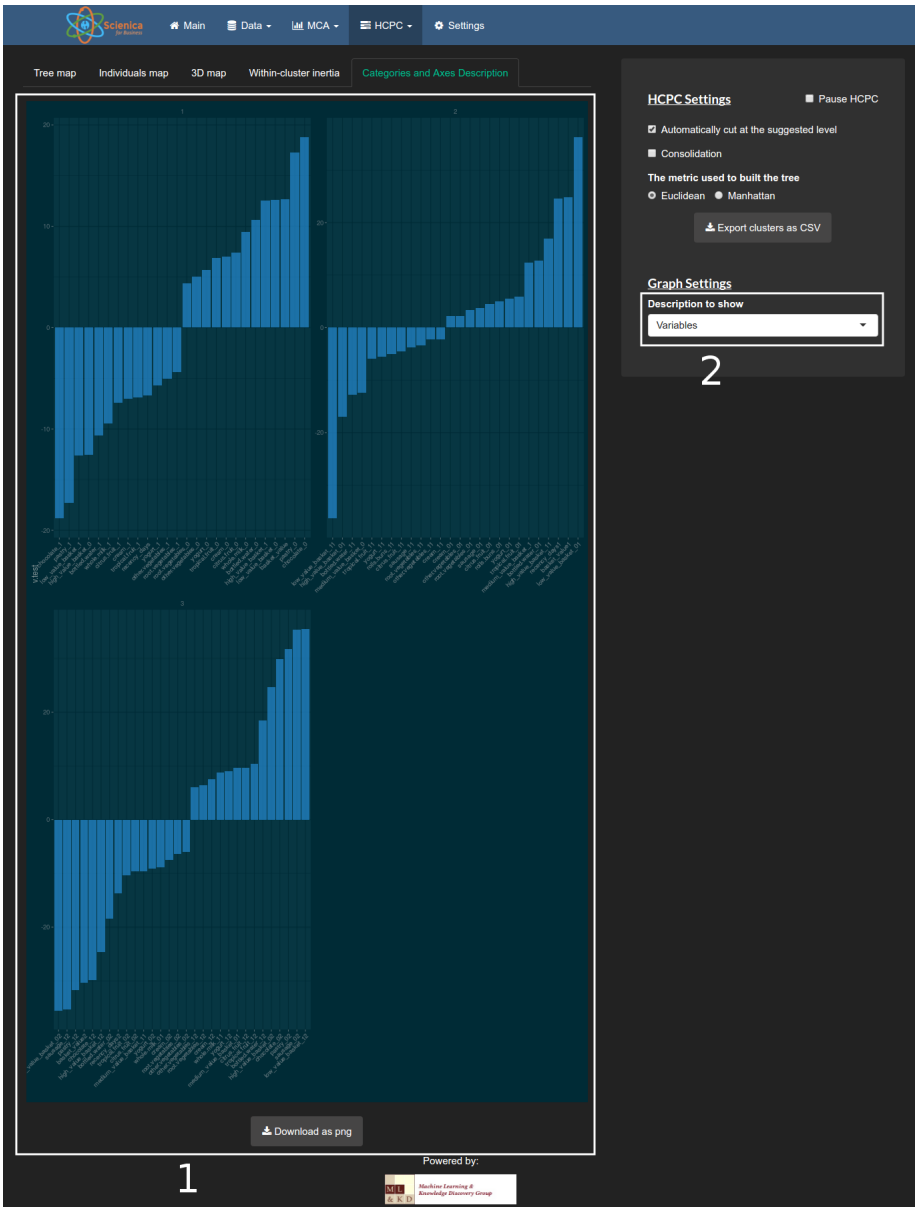
Στο σχήμα 4.36 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:

1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Ο χρήστης με αυτήν την επιλογή μπορεί το κάτω όριο της αδράνειας επί τις εκατό που να σταματήσει να υπολογίζει ο αλγόριθμος.

Η συνάρτηση που δημιουργήθηκε γι΄ αυτό το σκοπό βρίσκεται στο Παράρτημα Πηγαίου Κώδικα και συγκεκριμένα στο πίνακα Α΄.5.

Categories and Axes Description Στο γράφημα αυτό ο χρήστης μπορεί να δει κατά πόσο κάθε κατηγορία ή άξονας περιγράφει θετικά ή αρνητικά κάθε κατηγορία.



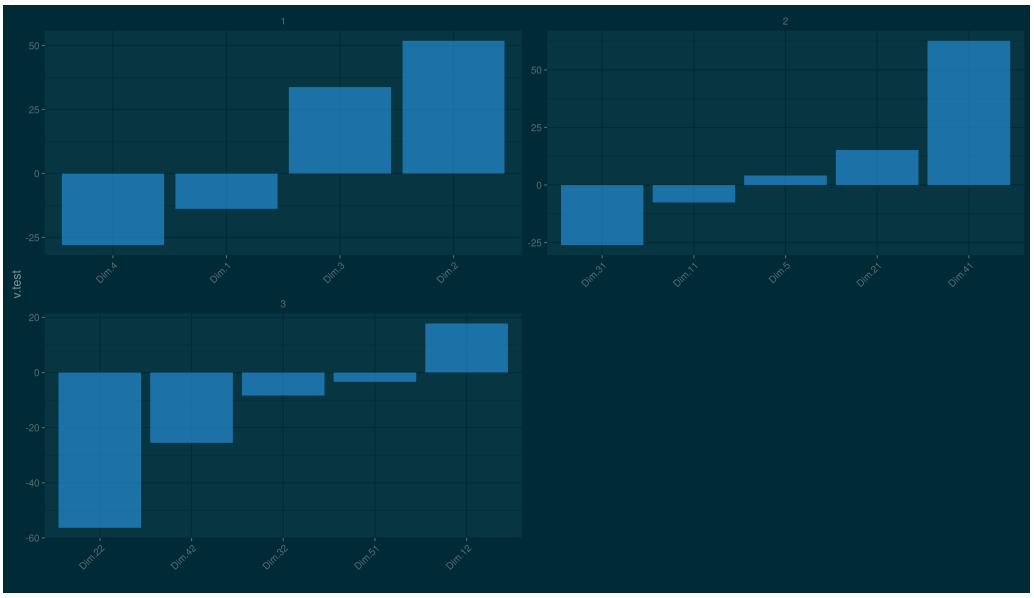
Σχημα 4.37: *Η οθόνη του γραφήματος Categories and Axes Description.*

Στο σχήμα 4.37 φαίνεται το παράθυρο του γραφήματος. Παρακάτω εξηγούνται όλες οι λειτουργίες του:

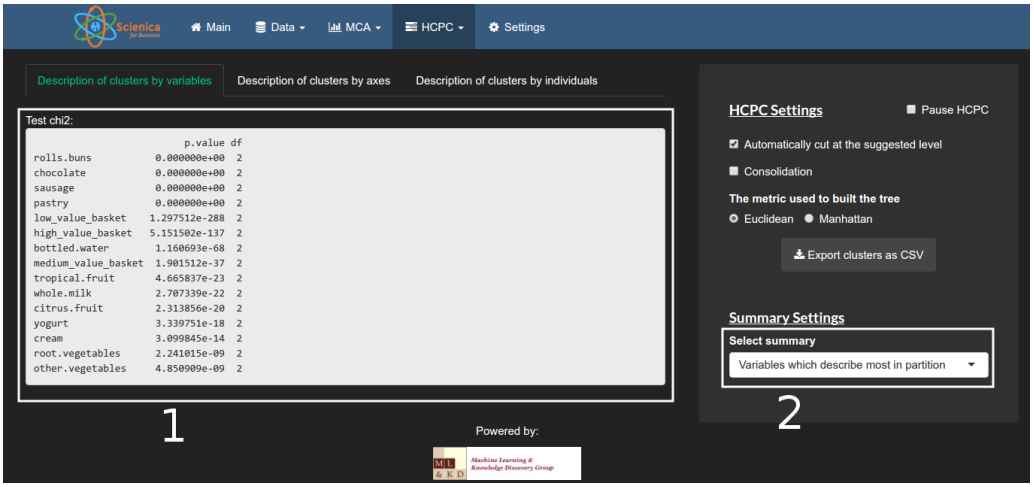
1. Το κύριο panel του γραφήματος όπου φαίνεται το γράφημα και δίνεται η δυνατότητα στο χρήστη να αποθηκεύσει το γράφημα στον υπολογιστή του σε μία από τις τρεις μορφές που έχει επιλέξει στις ρυθμίσεις. (Παράγραφος 4.3.2.7)

2. Με αυτήν την επιλογή ο χρήστης μπορεί να επιλέξει στο να εμφανίσει την σύνοψη σχετικά με τις κατηγορίες ή τους άξονες που περιγράφουν κάθε ομάδα.

Ένα γράφημα που δείχνει την περιγραφή κάθε ομάδας σχετικά με κάθε άξονα φαίνεται στο σχήμα 4.38.



Σχημα 4.38: *Το γράφημα Categories and Axes Description.*



Σχημα 4.39: *Η οθόνη της σύνοψης HCPC.*

Η συνάρτηση που δημιουργήθηκε γι΄ αυτό το σκοπό βρίσκεται στο Παράρτημα Πηγαίου Κώδικα και συγκεκριμένα στο πίνακα Α΄.6.

4.3.2.6.2 Σύνοψη της ανάλυσης HCPC (HCPC Summaries)

Σε αυτήν την επιλογή ο χρήστης μπορεί να δει κάποια αποτελέσματα από την ανάλυση HCPC όπως να διερευνήσει ποιες μεταβλητές ή κατηγορίες εξηγούν περισσότερο κάθε ομάδα. Στο σχήμα 4.39 φαίνεται πως είναι το παράθυρο της σύνοψης HCPC. Αποτελείται από τρεις καρτέλες: **Description of clusters by variables**, **Description of clusters by axes** και **Description of clusters by individuals**.

Παρακάτω εξηγούνται οι λειτουργίες:

1. Το κύριο panel της σύνοψης. Εδώ εμφανίζονται όλες οι λεπτομέρειες

της ανάλυσης.

2. Εδώ ο χρήστης μπορεί να αλλάξει ποιον τύπο σύνοψης θέλει να δει για κάθε καρτέλα.

Παρακάτω περιγράφονται η σημασία κάθε καρτέλας μαζί με κάθε τύπο σύνοψης.

Description of clusters by variables

Σε αυτήν την ενότητα ο χρήστης έχει τη δυνατότητα να δει μια περιγραφή των ομάδων από τη σκοπιά των μεταβλητών. Έχει στη διάθεσή του να δει σε αύξουσα σειρά σε σχέση με την p-value όλες τις παρακάτω πληροφορίες που εξηγούνται στο πίνακα 4.4.

Τύπος Περιγραφής	**Σημασία**
Variables which describe most in partition	Σε αυτόν τον τύπο σύνοψης ο χρήστης μπορεί να δει ποιες μεταβλητές επεξηγούν καλύτερα το σύνολο των ομάδων.
Description of each category by cluster	Εδώ ο χρήστης μπορεί να δει ποιες κατηγορίες των μεταβλητών επεξηγούν καλύτερα κάθε ομάδα.
Global description of quantitative variables	Με αυτήν τη σύνοψη ο χρήστης είναι σε θέση να δει ποιες ποσοτικές μεταβλητές περιγράφουν καλύτερα το σύνολο των ομάδων.
Description of quantitative variables by cluster	Εδώ ο χρήστης μπορεί αναλυτικά ποιες ποσοτικές μεταβλητές περιγράφουν καλύτερα κάθε ομάδα.

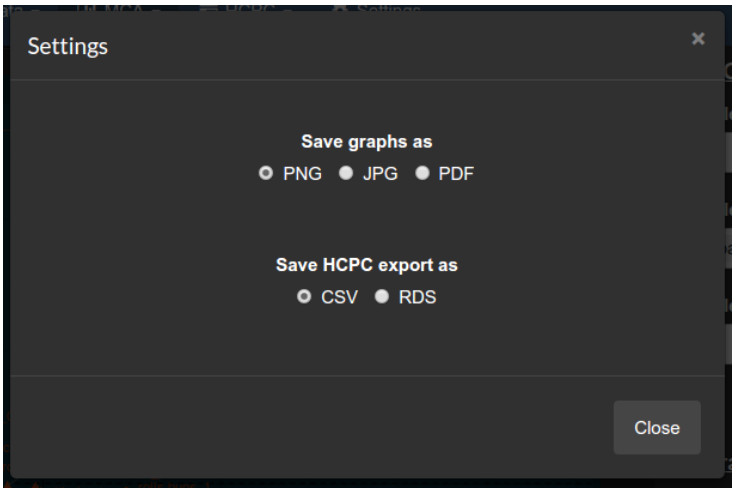
Πίνακας 4.4: Οι τύποι περιγραφής των ομάδων σε σχέση με τις μεταβλητές.

Description of clusters by axes

Σε αυτήν την ενότητα ο χρήστης έχει τη δυνατότητα να δει μια περιγραφή των ομάδων από τη σκοπιά των αξόνων και πως κάθε ένα περιγράφει κάθε ομάδα. Και εδώ μπορεί να δει του όλες τις πληροφορίες που εξηγούνται στο πίνακα 4.5 σε αύξουσα σειρά σε σχέση με την p-value.

Τύπος Περιγραφής	**Σημασία**
Global description by axes	Ο χρήστης με αυτήν την επιλογή έχει τη δυνατότητα να δει μια περιγραφή όλων των ομάδων με βάση τους άξονες (κύριες συνιστώσες).
Description of clusters by axes	Ο χρήστης με αυτήν την επιλογή έχει τη δυνατότητα να δει μια περιγραφή κάθε ομάδας με βάση τους άξονες (κύριες συνιστώσες) τις εξηγούν καλύτερα.

Πίνακας 4.5: Οι τύποι περιγραφής των ομάδων σε σχέση με τους άξονες.



Σχημα 4.40: *Η οθόνη των ρυθμίσεων της εφαρμογής.*

Description of clusters by individuals

Τέλος σε αυτήν την ενότητα ο χρήστης έχει τη δυνατότητα να δει μια περιγραφή των ομάδων από τη σκοπιά των παρατηρήσεων. Έχει στη διάθεσή του δύο τύπους περιγραφής που εξηγούνται στο πίνακα 4.6.

Τύπος Περιγραφής	**Σημασία**
Paragons (Closest to cluster center)	Σε αυτόν τον τύπο ο χρήστης μπορεί να δει τις 5 πρώτες παρατηρήσεις κάθε ομάδας που βρίσκονται πιο κοντά στο κέντρο της ομάδας.
Dist (Farthest from other clusters)	Σε αυτόν τον τύπο ο χρήστης μπορεί να δει τις 5 πρώτες παρατηρήσεις κάθε ομάδας που βρίσκονται πιο μακριά από κάθε άλλη ομάδα.

Πίνακας 4.6: Οι τύποι περιγραφής των ομάδων σε σχέση με τις παρατηρήσεις.

4.3.2.7 Ρυθμίσεις (Settings)

Στην εικόνα 4.40 φαίνεται η οθόνη ρυθμίσεων. Ο χρηστής από εδώ έχει τη δυνατότητα να αλλάξει τον τύπο αποθήκευσης των γραφημάτων ανάμεσα σε τρεις επιλογές: **PNG**, **JPG** και **PDF**. Επίσης έχει την επιλογή να διαλέξει τον τύπο εξαγωγής των δεδομένων από την HCPC ανάλυση ανάμεσα σε δύο επιλογές: **CSV** και **RDS** (σειριακό αντικείμενο της R).

4.4 Αναπαράσταση δεδομένων από συναλλαγές πελατών

Σε αυτό το κεφάλαιο θα αναλύσουμε την αναπαράσταση των δεδομένων όπως χρησιμοποιούνται από την εφαρμογή καθώς επίσης και ένα δείγμα από

τα δεδομένα που χρησιμοποιήθηκαν για τις δοκιμές της εφαρμογής.

Η αναπαράσταση των πελατών σε σχέση με την αγοραστική της συμπεριφορά γίνεται με τον δυαδικό πίνακα προφίλ πελατών (Binary Profile Data Table). Ένας τέτοιος πίνακας χρησιμοποιείται για να αποτυπώσει τις συναλλαγές όπου κάθε γραμμή αποτελεί έναν πελάτη, κάθε στήλη ένα προϊόν και κάθε κελί μία λογική τιμή 0 ή 1 που δηλώνει αν ο πελάτης έχει αγοράσει έστω μία φορά ή όχι το προϊόν. Ο πίνακας αυτός είναι ιδιαίτερα γνωστός στο χώρο της ανάλυσης καλαθιού πελατών (Market Basket Analysis) και συγκεκριμένα με τη χρήση της τεχνικής των κανόνων συσχέτισης. Στην ανάλυση με τη χρήση της μεθόδου MCA, προτείνεται η χρήση ενός ανώτερου πιο αφαιρετικού επιπέδου στην κατηγορία προϊόντων έναντι του τελευταίο εξαντλητικού επιπέδου που ουσιαστικά είναι το ίδιο προϊόν. Η ανάλυση αυτή έχει περισσότερο νόημα για να καταφέρουν να δουν τα στελέχη της επιχείρησης την πιο γενική εικόνα του πως κινείται η αγορά και να μπορέσουν να βρουν πιο εύκολα κοινές συμπεριφορές μη εξαρτημένες από την μεροληψία της επωνυμίας του προϊόντος.

Ένας τυπικός πίνακας με προφίλ πελατών αποτελούμενος από 1000 κατηγορίες προϊόντων θα έδειχνε όπως φαίνεται στο πίνακα 4.7.

Item / Client	I_1	I_2	I_3	I_4	I_5	...	I_{1000}
C_1	1	0	1	0	0	...	1
C_2	0	1	1	1	0	...	0
C_3	1	0	1	1	1	...	0
C_4	0	0	0	0	1	...	1
C_5	0	1	1	0	1	...	1
...	...	...	...	...	...	...	...
C_{100}	1	0	0	0	1	...	1
...	...	...	...	...	...	...	...

Πίνακας 4.7: Τυπικός πίνακας με προφίλ πελατών.

Για να χτιστεί και να δοκιμαστεί η εφαρμογή, χρησιμοποιήθηκε ένα σύνολο δεδομένων που αποτελείται από 7536 προφίλ πελατών και 13 κατηγορίες προϊόντων. Επιπλέον υπήρχαν και στο σύνολο ακόμα 5 μεταβλητές που δήλωναν την αξία του καλαθιού (Basket Value) και την προσφατότητα (Recency), δηλαδή τον αριθμό των ημερών που έχουν περάσει από την τελευταία φορά που ένας πελάτης έκανε μια αγορά. Οι μεταβλητές αυτές χρησιμοποιήθηκαν ως συμπληρωματικές μεταβλητές στην ανάλυση. Οι δύο από αυτές ήταν συνεχείς μεταβλητές και οι υπόλοιπες τρεις χώριζαν την μεταβλητή Basket Value σε 3 κλάσεις με τη μέθοδο της διακριτοποίησης. Ένα παράδειγμα του συνόλου δεδομένων φαίνεται στο πίνακα 4.8.

	citrus.fruit	tropical.fruit	whole.milk	other.vegetables	rolls.buns	chocolate	bottled.water
1	1	0	0	0	0	0	0
2	0	1	0	0	0	0	0
3	0	0	1	0	0	0	0
4	0	0	0	0	0	0	0
5	0	1	1	0	0	0	0
6	0	0	1	0	0	0	0
7	0	0	0	0	1	0	0
8	0	0	0	1	1	0	0
9	0	0	1	0	0	0	0
10	0	1	0	1	0	1	1

	yogurt sausage	root.vegetables	pastry	soda	cream	basket_value	recency_days
1	0	0	0	0	0	1.1	2
2	1	0	0	0	0	3.6	31
3	0	0	0	0	0	1.2	7
4	1	0	0	0	1	6.6	7
5	0	0	0	0	0	2.5	66
6	1	0	0	0	0	2.7	61
7	0	0	0	0	0	2.2	67
8	0	0	0	0	0	3.5	47
9	0	0	0	0	0	1.2	7
10	0	0	0	0	0	7.9	17

	low_value_basket	medium_value_basket	high_value_basket
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	1	0	0
6	1	0	0
7	1	0	0
8	0	1	0
9	1	0	0
10	0	0	1

Πίνακας 4.8: Οι 10 πρώτες παρατηρήσεις του πίνακα προφίλ πελατών που χρησιμοποιήθηκε για δοκιμή.

5

Συμπέρασμα - Μελλοντική Εργασία

5.1 Συμπέρασμα

Η εργασία αυτή είχε σκοπό την ανάδειξη της μεθόδου Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών και την παρουσίαση μιας διαδικτυακής εφαρμογής που θα χρησιμοποιεί τη μέθοδο, για να δημιουργήσει ομάδες πελατών με βάση τα προϊόντα ή τις κατηγορίες προϊόντων που αυτοί αγοράζουν μέσω του Ιεραρχικού Αλγορίθμου.

Η μέθοδος της Πολυμεταβλητής Παραγοντικής Ανάλυσης Αντιστοιχιών δεν είναι ευρέως γνωστή στο χώρο της Επιχειρηματικής Ευφυίας και μέσα από την εργασία παρουσιάστηκαν οι δυνατότητες που προσφέρει και για το πως θα μπορούσε να χρησιμοποιηθεί για να βοηθήσει την επιχείρηση να εξελιχθεί και να ανταποκριθεί ακόμα καλύτερα στην αγορά.

Με το Scienica for Business, τη διαδικτυακή εφαρμογή επιχειρηματικής ευφυΐας που δημιουργήθηκε στο πλαίσιο της διατριβής, αναδείχθηκαν οι ανάγκες αλλά και τρόποι ώστε αυτές να καλυφθούν, που εγείρεται από την επιχείρηση και ιδιαίτερα από το τμήμα Marketing. Με την ανακάλυψη ομάδων από πελάτες, οι αναλυτές είναι σε θέση να δουν ποια προϊόντα αγοράζονται μαζί, αλλά ακόμα πιο σημαντικό είναι σε θέση να εξάγουν έναν μέσο/πρότυπο πελάτη κάθε ομάδας. Αυτός ο πελάτης θα έχει όλα τα χαρακτηριστικά που κάνουν την ομάδα. Με βάση αυτόν, το τμήμα Marketing θα είναι σε θέση να μελετήσει γιατί οι υπόλοιποι πελάτες της ομάδας δεν συμπεριφέρονται σαν τον πρότυπο πελάτη και να εφαρμόσουν στοχευμένες προωθητικές ενέργειες σε συγκεκριμένους πελάτες για να ακολουθήσουν κάποια κοινά αγοραστικά χαρακτηριστικά.

Η εφαρμογή χρησιμοποιεί αρκετά γραφήματα που βοηθούν τον αναλυτή να κατανοήσει καλύτερα το σύνολο δεδομένων και τα αποτελέσματα της ανάλυσης. Μέσω αυτών μπορεί να πάρει πιο γρήγορα αποφάσεις ώστε να βελτιώσει τις παραμέτρους της μεθόδου και να εξάγει τα καλύτερα αποτελέσματα από

αυτή. Παρόλα αυτά ο αναλυτής έχει πρόσβαση και σε πιο λεπτομερή αποτελέσματα της ανάλυσης και μπορεί να φτάσει σε όποιο βάθος επιθυμεί για να βελτιώσει ακόμα περισσότερο τα αποτελέσματα. Δημιουργήθηκε εξολοκλήρου με εργαλεία ανοιχτού κώδικα, πράγμα που καθιστά την εφαρμογή προσβάσιμη από όλους και εύκολα αναβαθμίσιμη.

5.2 Μελλοντική Εργασία

Σε μελλοντική υλοποίηση της εφαρμογής, ο χρήστης αρχικά θα μπορούσε να συνδέει τη Βάση Δεδομένων με την εφαρμογή για να μπορεί να γίνει απευθείας η ανάλυση στα δεδομένα της επιχείρησης.

Επιπλέον, πολλές λειτουργίες και παραμέτρους της ανάλυσης θα μπορούσε να αυτοματοποιούνται με σκοπό η εφαρμογή να είναι εύκολα χρησιμοποιήσιμη από μη καταρτισμένο προσωπικό της επιχείρησης όπως αναλυτές και επιστήμονες δεδομένων.

Επιπροσθέτως, ένα πολύ σημαντικό βήμα της ανάλυσης είναι η βελτίωση της αναπαράστασης των δεδομένων από δυαδική μορφή σε πίνακα περισσοτέρων κατηγοριών που θα δώσουν νόημα σε όλες τις κατηγορίες. Μια πρώτη σκέψη είναι η χρησιμοποίηση των συχνοτήτων αγοράς κάθε προϊόντος από κάθε πελάτη και τη δημιουργία αυτών σε ομάδες/κατηγορίες.

Τέλος η εφαρμογή θα μπορούσε να αποτελέσει μέρος ενός μεγαλύτερου πρότζεκτ εφαρμογής επιχειρηματικής ευφυίας όπως το πρόσφατα δημοσιευμένο Segmento[26] που βασίζεται και αυτό στη γλώσσα R και τη βιβλιοθήκη Shiny.

Παράρτημα Α′

Πηγαίος Κώδικας

Πηγαίος Κώδικας Α′.1: Συνάρτηση Παραγωγής Διαγράμματος "Visualize dataset".

```
viz_data <- function(dataset, xvar,
                     yvar = "",
                     type = "hist",
                     facet_row = "None",
                     facet_col = "None",
                     color = "None",
                     fill = "None",
                     bins = 10,
                     smooth = 1,
                     fun = c("mean","median"),
                     check = "",
                     axes = "",
                     alpha = .5) {
  vars <- c(xvar)

  if (!type %in% c("scatter","line"))
    color <- "None"
  if (!type %in% c("bar","hist","density"))
    fill <- "None"
  if (type != "scatter") {
    check %<>% sub("line","",.) %>% sub("loess","",.)
  }

  byvar <- NULL

  if (length(yvar) == 0 || identical(yvar, "")) {
    if (!type %in% c("hist","density")) {
      return()
    }
  } else {
    if (type %in% c("hist","density")) {
      yvar <- ""
    } else {
      vars <- c(vars, yvar)
    }
  }

  if (color != "None") {
    vars <- c(vars, color)
    if (type == "line")
      byvar <- color
  }
  if (facet_row != "None") {
    vars <- c(vars, facet_row)
    byvar <-
```

```
      if (is.null(byvar))
        facet_row
    else
      unique(c(byvar, facet_row))
  }
  if (facet_col != "None") {
    vars <- c(vars, facet_col)
    byvar <-
      if (is.null(byvar))
        facet_col
    else
      unique(c(byvar, facet_col))
  }
  if (fill != "None") {
    vars <- c(vars, fill)
    if (type == "bar")
      byvar <- if (is.null(byvar))
        fill
    else
      unique(c(byvar, fill))
  }

  dat <- select_(dataset, .dots = vars)

  dc <- getclass(dat)

  isChar <- dc == "character"
  if (sum(isChar) > 0) {
    if (type == "density") {
      dat[,isChar] <-
        select(dat, which(isChar)) %>% mutate_each(funs(as_numeric))
    } else {
      dat[,isChar] <-
        select(dat, which(isChar)) %>% mutate_each(funs(as.factor))
    }
    dc <- getclass(dat)
  }

  if (type == "bar") {
    isFctY <- "factor" == dc & names(dc) %in% yvar
    if (sum(isFctY)) {
      dat[,isFctY] <-
        select(dat, which(isFctY)) %>% mutate_each(funs(as.integer(. ==
        ↪  levels(.)[1])))
      dc[isFctY] <- "integer"
    }
  }

  log_trans <- function(x)
    ifelse(x > 0, log(x), NA)

  if ("log_x" %in% axes) {
    if (any(!dc[xvar] %in% c("integer","numeric")))
      return("Log X is only meaningful for X variables of type integer or numeric")
    to_log <- (dc[xvar] %in% c("integer","numeric")) %>% xvar[.]
    dat[, to_log] <-
      select_(dat, .dots = to_log) %>% mutate_each(funs(log_trans))
  }

  if ("log_y" %in% axes) {
    if (any(!dc[yvar] %in% c("integer","numeric")))
      return("Log Y is only meaningful for Y variables of type integer or numeric")
    to_log <- (dc[yvar] %in% c("integer","numeric")) %>% yvar[.]
    dat[, to_log] <-
      select_(dat, .dots = to_log) %>% mutate_each(funs(log_trans))
  }

  plot_var <-  list()
  if (type == "hist") {
    hist_par <- list(alpha = alpha, position = "dodge")
```

```
    plot_var <- ggplot(dat, aes_string(x = xvar))
    if ("density" %in% axes && !"factor" %in% dc[xvar]) {
      hist_par <-
        list(aes(y = ..density..), alpha = alpha, position = "dodge")
      plot_var <-
        plot_var + geom_density(color = "#d9c263", size = .5)
    }
    if ("factor" %in% dc[xvar]) {
      plot_fun <- get("geom_bar")
      if ("log_x" %in% axes)
        axes <- sub("log_x","",axes)
    } else {
      plot_fun <- get("geom_histogram")
      hist_par[["binwidth"]] <-
        select_(dat,xvar) %>% range %>% {
          diff(.) / bins
        }
    }


    plot_var <- plot_var + do.call(plot_fun, hist_par)
    if ("log_x" %in% axes)
      plot_var <- plot_var + xlab(paste("log", xvar))

  }
  else if (type == "density") {
    plot_var <- ggplot(dat, aes_string(x = xvar)) +
      if (fill == "None")
        geom_density(
          adjust = smooth, color = "#268bd2", fill = "#268bd2", alpha = alpha, size
          ↪  = 1
        )
      else
        geom_density(adjust = smooth, alpha = alpha, size = 1)

    if ("log_x" %in% axes)
      plot_var <- plot_var + xlab(paste("log", xvar))
  } else if (type == "scatter") {
    if ("jitter" %in% check) {
      gs <-
        geom_jitter(alpha = alpha, position = position_jitter(width = 0.4, height =
        ↪  0.1))
      check <- sub("jitter","", check)
    } else {
      gs <- geom_point(alpha = alpha)
    }

    if ("log_x" %in% axes &&
        dc[xvar] == "factor")
      axes <- sub("log_x","",axes)


    plot_var <- ggplot(dat, aes_string(x = xvar, y = yvar)) + gs

    if ("log_x" %in% axes)
      plot_var <- plot_var + xlab(paste("log", xvar))
    if ("log_y" %in% axes)
      plot_var <- plot_var + ylab(paste("log", yvar))

    if (dc[xvar] == "factor") {
      suppressWarnings(ymax <- max(dat[[yvar]]) %>% {if (. < 0) 0 else .})
      suppressWarnings(ymin <- min(dat[[yvar]]) %>% {if (. > 0) 0 else .})

      plot_var <- plot_var + ylim(ymin,ymax)

      meanf <- function(y) {
          y <- mean(y)
          data.frame(ymin = y, ymax = y, y = y)
      }
      medianf <- function(y) {
```

```
        y <- median(y)
        data.frame(ymin = y, ymax = y, y = y)
      }
      plot_var <- plot_var + ylab(paste(plot_var$labels$y, "(mean<blue> |
      ↪  median<red>)")) +
        stat_summary(fun.data = meanf, geom = "crossbar", color = "blue") +
        stat_summary(fun.data = medianf, geom = "crossbar", color = "red")
    }

  } else if (type == "line") {
    if (color == 'None') {
      if (dc[xvar] %in% c("factor","date")) {
        tbv <- if (is.null(byvar)) xvar else c(xvar, byvar)
        tmp <- dat %>% group_by_(.dots = tbv) %>% select_(yvar) %>%
        ↪  summarise_each(funs(mean(.)))
        plot_var <- ggplot(tmp, aes_string(x = xvar, y = yvar)) +
        ↪  geom_point(aes(color = "blue")) +
          geom_line(aes(color ="blue", group = 1))
      } else {
        plot_var <- ggplot(dat, aes_string(x = xvar, y = yvar)) +
        ↪  geom_line(aes(color = "blue"))
      }
    } else {
      if (dc[xvar] %in% c("factor","date")) {
        tbv <- if (is.null(byvar)) xvar else c(xvar, byvar)
        tmp <- dat %>% group_by_(.dots = tbv) %>% select_(yvar, color) %>%
        ↪  summarise_each(funs(mean(.)))
        plot_var <- ggplot(tmp, aes_string(x = xvar, y = yvar, color = color, group
        ↪  = color)) +
          geom_point() + geom_line()
      } else {
        plot_var <- ggplot(dat, aes_string(x = xvar, y = yvar, color = color, group
        ↪  = color)) + geom_line()
      }
    }
    if ("log_x" %in% axes)
      plot_var <- plot_var + xlab(paste("log", xvar))
    if ("log_y" %in% axes)
      plot_var <- plot_var + ylab(paste("log", yvar))
    if (dc[xvar] %in% c("factor","date"))
      plot_var$labels$y %<>% paste0(., " (mean)")
  } else if (type == "bar") {
    if ("log_x" %in% axes)
      axes <- sub("log_x","",axes)

    tbv <- if (is.null(byvar)) xvar else c(xvar, byvar)
    tmp <- dat %>% group_by_(.dots = tbv) %>% select_(yvar) %>%
    ↪  summarise_each(funs(mean(.)))

    if ("sort" %in% axes && facet_row == "None" && facet_col == "None") {
      tmp <- arrange_(ungroup(tmp), yvar)
      tmp[[xvar]] %<>% factor(., levels = unique(.))
    }

    plot_var <- ggplot(tmp, aes_string(x = xvar, y = yvar)) +
                geom_bar(stat = "identity", position = "dodge", alpha = alpha)

    if ("log_y" %in% axes)
      plot_var <- plot_var + ylab(paste("log", yvar))

    plot_var$labels$y %<>% paste0(., " (mean)")

  } else if (type == "box") {
    plot_var <- ggplot(dat, aes_string(x = xvar, y = yvar, fill = xvar)) +
      geom_boxplot(alpha = alpha,aes_string(colour = xvar)) +
      theme(legend.position = "none")

    if ("log_y" %in% axes)
      plot_var <- plot_var + ylab(paste("log", yvar))
  } else if (type == "violin") {
```

```
    plot_var <- ggplot(dat, aes_string(x = xvar, y = yvar, fill = xvar)) +
      geom_violin(alpha = alpha,aes_string(colour = xvar)) +
      theme(legend.position = "none")

    if ("log_y" %in% axes)
      plot_var <- plot_var + ylab(paste("log", yvar))
  }

  if (facet_row != "None" || facet_col != "None") {
    facets <- {
      if (facet_row == "None")
        paste("~", facet_col)
      else if (facet_col == "None")
        paste(facet_row, '~', ".")
      else
        paste(facet_row, '~', facet_col)
    }
    scl <- if ("scale_y" %in% axes) "free_y" else "fixed"
    if (facet_row == "None") plot_var <- plot_var +  facet_wrap(as.formula(facets),
                                       scales = scl, labeller = label_both)
    else plot_var <- plot_var +  facet_grid(as.formula(facets), scales = scl,
    ↪  labeller = label_both)
  }

  if (color != 'None') {
    plot_var <- plot_var + aes_string(color = color)
  }

  if (fill != 'None') {
    plot_var <- plot_var + aes_string(fill = fill)
  }

  if ("jitter" %in% check) {
    plot_var <- plot_var + geom_jitter(alpha = alpha, position =
    ↪  position_jitter(width = 0.4, height = 0.1))
  }

  if ("line" %in% check) {
    plot_var <- plot_var + invisible(geom_smooth(method = "lm", alpha = .1, size =
    ↪  .75,linetype = "dashed"))
  }

  if ("loess" %in% check) {
    plot_var <- plot_var + invisible(geom_smooth(span = smooth, method = "loess",
                size = .75, linetype = "dotdash", aes(group = 1)))
  }

  if ("flip" %in% axes) {
    plot_var <- plot_var + coord_flip()
  }

  if ((type == "scatter" ||
       type == "line") &&
      color != "None" &&
      getclass(select_(dat, .dots = c(color))) %in% c("integer","numeric")) {
    return(plot_var + theme_solarized_2(light = F))
  }
  return(plot_var + theme_solarized_2(light = F))
}
```

Πηγαίος Κώδικας Α΄.2: Συνάρτηση Παραγωγής Διαγράμματος "Categories confidence ellipses".

```
plotellipses <-
  function (model, keepvar = "all", axes = c(1, 2), means = TRUE,
            level = 0.95, magnify = 2, cex = 1, pch = 20, pch.means = 15,
            type = c("g", "p"), keepnames = TRUE, namescat = NULL, xlim =
              NULL, ylim = NULL, lwd = 1, label = "none",
            autoLab = c("auto","yes","no"), title = "Confidence ellipses around the
            ↪  categories",...)
  {
    monpanel <- function(x, y, level, means, nommod, magnify = magnify,
                         pchmeans = pchmeans, ...) {
      lattice::panel.xyplot(x, y, ...)
      panel.superpose2(
        x, y, level = level, means = means,
        nommod = nommod, pchmeans = pchmeans, magnify = magnify,
        panel.groups = "monpanel.ellipse", ...
      )
    }
    monpanel.ellipse <- function(x, y, col.line, lty, lwd, subscripts,
                                 pch, cex, font, font.family, col, col.symbol,
                                 ↪  fill, alpha,
                                 type, group.number, level, means, nommod, magnify,
                                 ↪  pchmeans,
                                 ...) {
      if (length(x) > 1) {
        matrice <- matrix(c(x, y), ncol = 2)
        cdg <- colMeans(matrice)
        if (means) {
          variance <- var(matrice) / length(x)
        }
        else {
          variance <- var(matrice)
        }
        coord.ellipse <- ellipse::ellipse(variance, centre = cdg,
                                          level = level)
        lattice::llines(
          coord.ellipse[, 1], coord.ellipse[, 2], col = col.line,
          lty = lty, lwd = lwd
        )
      }
      else
        cdg <- c(x, y)
      lattice::ltext(
        cdg[1], cdg[2], unique(nommod), cex = cex * magnify,
        col = col.line, pos = 3, offset = 0.4 * cex * magnify
      )
      lattice::lpoints(cdg[1], cdg[2], pch = pchmeans[group.number], col =
      ↪  col.line,
                       cex = cex * magnify)
    }
    panel.superpose2 <- function(x, y = NULL, subscripts, groups,
                                 panel.groups = "panel.xyplot", col = NA, col.line
                                 ↪  = superpose.line$col,
                                 col.symbol = superpose.symbol$col, pch =
                                 ↪  superpose.symbol$pch,
                                 cex = superpose.symbol$cex, fill =
                                 ↪  superpose.symbol$fill,
                                 font = superpose.symbol$font, fontface =
                                 ↪  superpose.symbol$fontface,
                                 fontfamily = superpose.symbol$fontfamily, lty =
                                 ↪  superpose.line$lty,
                                 lwd = superpose.line$lwd, alpha =
                                 ↪  superpose.symbol$alpha,
                                 type = "p", nommod = nommod, ..., distribute.type
                                 ↪  = FALSE) {
      if (distribute.type) {
```

```
        type <- as.list(type)
      }
      else {
        type <- unique(type)
        wg <- match("g", type, nomatch = NA)
        if (!is.na(wg)) {
          lattice::panel.grid(h = -1, v = -1)
          type <- type[-wg]
        }
        type <- list(type)
      }
      x <- as.numeric(x)
      if (!is.null(y))
        y <- as.numeric(y)
      if (length(x) > 0) {
        if (!missing(col)) {
          if (missing(col.line))
            col.line <- col
          if (missing(col.symbol))
            col.symbol <- col
        }
        superpose.symbol <-
          lattice::trellis.par.get("superpose.symbol")
        superpose.line <- lattice::trellis.par.get("superpose.line")
        vals <- if (is.factor(groups))
          levels(groups)
        else
          sort(unique(groups))
        nvals <- length(vals)
        col <- rep(col, length = nvals)
        col.line <- rep(col.line, length = nvals)
        col.symbol <- rep(col.symbol, length = nvals)
        pch <- rep(pch, length = nvals)
        fill <- rep(fill, length = nvals)
        lty <- rep(lty, length = nvals)
        lwd <- rep(lwd, length = nvals)
        alpha <- rep(alpha, length = nvals)
        cex <- rep(cex, length = nvals)
        font <- rep(font, length = nvals)
        if (!is.null(fontface))
          fontface <- rep(fontface, length = nvals)
        if (!is.null(fontfamily))
          fontfamily <- rep(fontfamily, length = nvals)
        type <- rep(type, length = nvals)
        panel.groups <- if (is.function(panel.groups))
          panel.groups
        else if (is.character(panel.groups))
          get(panel.groups)
        else
          eval(panel.groups)
        subg <- groups[subscripts]
        ok <- !is.na(subg)
        for (i in seq_along(vals)) {
          id <- ok & (subg == vals[i])
          if (any(id)) {
            args <- list(
              x = x[id], subscripts = subscripts[id],
              pch = pch[i], cex = cex[i], font = font[i],
              fontface = fontface[i], fontfamily = fontfamily[i],
              col = col[i], col.line = col.line[i], col.symbol = col.symbol[i],
              fill = fill[i], lty = lty[i], lwd = lwd[i],
              alpha = alpha[i], type = type[[i]], group.number = i,
              nommod = (nommod[subscripts])[id], ...
            )
            if (!is.null(y))
              args$y <- y[id]
            do.call(panel.groups, args)
          }
        }
      }
```

```
    }
    autoLab <- match.arg(autoLab,c("auto","yes","no"))
    if (autoLab == "yes")
      autoLab = TRUE
    if (autoLab == "no")
      autoLab = FALSE
    nomtot <- names(model$call$X)
    nbevartot <- ncol(model$call$X)
    eliminer <- NULL
    if (class(model)[1] == "PCA") {
      if (all(names(model$call) != "quali.sup"))
        return(NULL)
      else {
        nomtot <- names(model$call$X)[model$call$quali.sup$numero]
        eliminer <- (1:nbevartot)[-model$call$quali.sup$numero]
        if (is.character(keepvar)) {
          if (length(keepvar) == 1) {
            possibilites <- c("all", "quali", "quali.sup")
            choix <- match(keepvar, possibilites)
            if (is.na(choix)) {
              if (!any(keepvar == nomtot))
                return(NULL)
              else
                eliminer <- unique(c(eliminer, (1:nbevartot)[!(keepvar ==
                ↪  nomtot)]))
            }
            else {
              if (choix == 2)
                return(NULL)
            }
          }
          else
            eliminer <- unique(c(eliminer, (1:nbevartot)[!(nomtot %in% keepvar)]))
        }
        if (is.numeric(keepvar))
          eliminer <- unique(c(eliminer, (1:nbevartot)[-keepvar]))
        if (is.logical(keepvar))
          eliminer <- unique(c(eliminer, (1:nbevartot)[!keepvar]))
      }
      nomtot <- names(model$call$X)
      if (all((1:nbevartot) %in% eliminer))
        return(NULL)
    }
    else if (class(model)[1] == "MCA") {
      if (any(names(model$call) == "quanti.sup"))
        eliminer <- model$call$quanti.sup
      if (is.character(keepvar)) {
        if (length(keepvar) == 1) {
          possibilites <- c("all", "quali", "quali.sup")
          choix <- match(keepvar, possibilites)
          if (is.na(choix)) {
            if (!any(keepvar == nomtot))
              return(NULL)
            else
              eliminer <-
                unique(c(eliminer, (1:nbevartot)[!(keepvar == nomtot)]))
          }
          else {
            if (choix == 2)
              eliminer <- c(eliminer, model$call$quali.sup)
            if (choix == 3)
              eliminer <- (1:nbevartot)[-model$call$quali.sup]
            if (choix == 1) {
              if (all(!(names(model$call) %in% c("quali", "quali.sup"))))
                return(NULL)
            }
          }
        }
        else
          eliminer <-
```

```
            unique(c(eliminer, (1:nbevartot)[!(nomtot %in% keepvar)]))
      }
      if (is.numeric(keepvar))
        eliminer <- unique(c(eliminer, (1:nbevartot)[-keepvar]))
      if (is.logical(keepvar))
        eliminer <- unique(c(eliminer, (1:nbevartot)[!keepvar]))
    }
    else if (class(model)[1] == "MFA") {
      eliminer <- which(unlist(lapply(model$call$X,is.numeric)))
      if (is.character(keepvar)) {
        if (length(keepvar) == 1) {
          possibilites <- c("all", "quali", "quali.sup")
          choix <- match(keepvar, possibilites)
          if (is.na(choix)) {
            if (!any(keepvar == nomtot))
              return(NULL)
            else
              eliminer <- unique(c(eliminer, (1:nbevartot)[!(keepvar == nomtot)]))
          }
          else {
            if (choix == 2)
              eliminer <- c(eliminer, which(model$call$nature.var == "quali.sup"))
            if (choix == 3)
              eliminer <- c(eliminer, which(model$call$nature.var == "quali"))
            if (choix == 1) {
              if (all(!(names(model$call) %in% c("quali", "quali.sup"))))
                return(NULL)
            }
          }
        }
        else
          eliminer <- unique(c(eliminer, (1:nbevartot)[!(nomtot %in% keepvar)]))
      }
      if (is.numeric(keepvar))
        eliminer <- unique(c(eliminer, (1:nbevartot)[-keepvar]))
      if (is.logical(keepvar))
        eliminer <- unique(c(eliminer, (1:nbevartot)[!keepvar]))
    }
    if (!is.null(eliminer))
      nomvargardees <- nomtot[-eliminer]
    else
      nomvargardees <- nomtot
    if (!is.logical(keepnames)) {
      if (is.numeric(keepnames))
        nomvartrimmees <- nomtot[-unique(c(eliminer, keepnames))]
      else
        nomvartrimmees <- nomvargardees[!(nomvargardees %in% keepnames)]
    }
    else {
      if (length(keepnames) == 1) {
        if (keepnames)
          nomvartrimmees <- NULL
        else
          nomvartrimmees <- nomvargardees
      }
      else {
        if (length(keepnames) == length(nomtot))
          nomvartrimmees <-
            nomtot[(!keepnames) & ((1:nbevartot) != eliminer)]
        else
          return(NULL)
      }
    }
    if (is.null(model$call$ind.sup)) {
      if (!is.null(eliminer))
        donnees <- model$call$X[,-eliminer, drop = FALSE]
      else
        donnees <- model$call$X
    } else {
      if (!is.null(eliminer))
```

```
        donnees <- model$call$X[-model$call$ind.sup,-eliminer, drop = FALSE]
      else
        donnees <- model$call$X[-model$call$ind.sup,, drop = FALSE]
    }
    nbevar <- ncol(donnees)
    don <-
      apply(
        model$ind$coord[, axes], 2, FUN = function(x, k)
          rep(x, k), k = nbevar
      )
    nindiv <- nrow(donnees)
    rownames(don) <- NULL
    colnames(don) <- c("x", "y")
    nomvar <- rep(nomvargardees, each = nindiv)
    modalite2 <- as.vector(apply(data.matrix(donnees), 2, unlist))
    if (is.null(namescat))
      nommod <- as.vector(apply(donnees, 2, unlist))
    else
      nommod <- namescat
    if (!is.null(nomvartrimmees) & (is.null(namescat))) {
      kept <- (1:nbevar)[nomvargardees %in% nomvartrimmees]
      selecti <-
        as.vector(mapply(seq, (kept - 1) * nindiv + 1, (kept) * nindiv))
      nommod[selecti] <-
        substr(nommod[selecti], nchar(nomvar[selecti]) + 2, nchar(nommod[selecti]))
    }
    modalite = factor(modalite2)
    don <-  cbind.data.frame(don, var = nomvar, modalite = factor(modalite2))
    if (length(pch.means) < max(modalite2))
      pch.means <- rep(pch.means,length = max(modalite2))
    if (is.null(xlim) & is.null(ylim)) {
      don$class <- nommod
      centroids <- aggregate(cbind(x,y) ~ var + modalite + class,don,mean)
      g <- ggplot(don, aes(x,y,color = modalite)) + geom_point() +
      ↪  stat_ellipse(level = level) + geom_point(data = centroids,size = 5) +
      ↪  geom_text(data = centroids, aes(label = class, fontface =
      ↪  "bold.italic")) + facet_wrap( ~ var) + labs( title = title, x =
      ↪  paste("Dim ", axes[1], " (", round(model$eig[axes[1],2], 2), "%)", sep =
      ↪  ""), y = paste("Dim ", axes[2], " (", round(model$eig[axes[2], 2], 2),
      ↪  "%)", sep = ""))
      g <-
        g + theme_solarized_2(light = FALSE) + scale_colour_tableau() +
        ↪  theme(legend.position = "none")
      g
    }else {
      if (is.null(xlim))
        xlim <- ylim
      else {
        if (is.null(ylim))
          ylim <- xlim
      }
      don$class <- nommod
      centroids <- aggregate(cbind(x,y) ~ var + modalite + class,don,mean)
      g <- ggplot(don, aes(x,y,color = modalite)) + geom_point() +
      ↪  stat_ellipse(level = level) + geom_point(data = centroids,size = 5) +
      ↪  geom_text(data = centroids, aes(label = class, fontface =
      ↪  "bold.italic")) + facet_wrap( ~ var) + labs(title = title, x =
      ↪  paste("Dim ", axes[1], " (", round(model$eig[axes[1],2], 2), "%)", sep =
      ↪  ""), y = paste("Dim ", axes[2], " (", round(model$eig[axes[2], 2], 2),
      ↪  "%)", sep = "")) + xlim(c(xlim)) + ylim(c(ylim))

      g <-
        g + theme_solarized_2(light = FALSE) + scale_colour_tableau() +
        ↪  theme(legend.position = "none")
      g
    }
}
```

Πηγαίος Κώδικας Α΄.3: Συνάρτηση Παραγωγής Διαγράμματος "Variables".

```
mapVars <-
  function(x, axes = c(1, 2), choix = c("ind","var","quanti.sup"),
           xlim = NULL, ylim = NULL, invisible = c("none","ind", "var", "ind.sup",
           ↪  "quali.sup", "quanti.sup"),
           col.ind = "blue", col.var = "red", col.quali.sup = "darkgreen",
           col.ind.sup = "darkblue", col.quanti.sup = "blue",
           label = c("all","none","ind", "var", "ind.sup", "quali.sup",
           ↪  "quanti.sup"), title = NULL, habillage = "none", palette =
             NULL,
           autoLab = c("auto","yes","no"),new.plot = FALSE,select =
             NULL,selectMod = NULL, unselect = 0.7, shadowtext = FALSE,...) {
    label <-
      match.arg(
        label,c(
          "all","none","ind", "var", "ind.sup", "quali.sup", "quanti.sup"
        ),several.ok = TRUE
      )
    choix <- match.arg(choix,c("ind","var","quanti.sup"))
    autoLab <- match.arg(autoLab,c("auto","yes","no"))
    if (autoLab == "yes")
      autoLab = TRUE
    if (autoLab == "no")
      autoLab = FALSE
    autoLab = FALSE
    invisible <-
      match.arg(
        invisible,c("none","ind", "var", "ind.sup", "quali.sup",
        ↪  "quanti.sup"),several.ok =
          TRUE
      )
    if ("none" %in% invisible)
      invisible = NULL

    res.mca <- x
    if (!inherits(res.mca, "MCA"))
      stop("non convenient data")
    if (is.numeric(unselect))
      if ((unselect > 1) |
          (unselect < 0))
        stop("unselect should be betwwen 0 and 1")

    lab.x <-
      paste("Dim ",axes[1]," (",format(res.mca$eig[axes[1],2],nsmall = 2,digits =
                                         2),"%)",sep = "")
    lab.y <-
      paste("Dim ",axes[2]," (",format(res.mca$eig[axes[2],2],nsmall = 2,digits =
                                         2),"%)",sep = "")

    lab.var <- lab.quali.sup <- lab.quanti.sup <- FALSE
    if (length(label) == 1 &&
        label == "all")
      lab.var <- lab.quali.sup <- lab.quanti.sup <- TRUE
    if ("var" %in% label)
      lab.var <- TRUE
    if ("quali.sup" %in% label)
      lab.quali.sup <- TRUE
    if ("quanti.sup" %in% label)
      lab.quanti.sup <- TRUE

    test.invisible <- vector(length = 3)
    if (!is.null(invisible)) {
      test.invisible[1] <- match("var", invisible)
      test.invisible[2] <- match("quali.sup", invisible)
      test.invisible[3] <- match("quanti.sup", invisible)
    }
    else
```

```
    test.invisible <- rep(NA, 3)

  if ((new.plot) &
      !nzchar(Sys.getenv("RSTUDIO_USER_IDENTITY")))
    dev.new()
  if (is.null(palette))
    palette(
      c("black", "red", "green3", "blue", "cyan", "magenta", "darkgray",
      ↪  "darkgoldenrod", "darkgreen", "violet", "turquoise", "orange",
      ↪  "lightpink", "lavender", "yellow", "lightgreen", "lightgrey",
      ↪  "lightblue", "darkkhaki", "darkmagenta", "darkolivegreen",
      ↪  "lightcyan", "darkorange", "darkorchid", "darkred", "darksalmon",
      ↪  "darkseagreen", "darkslateblue", "darkslategray", "darkslategrey",
      ↪  "darkturquoise", "darkviolet", "lightgray", "lightsalmon",
      ↪  "lightyellow", "maroon"
      )
    )
  if (is.null(xlim))
    xlim <- c(0,1)
  if (is.null(ylim))
    ylim <- c(0,1)
  if (is.null(title))
    title <- "Variables representation"
  coo <- labe <- coll <- ipch <- fonte <- NULL
  coord.actif <- res.mca$var$eta2[, axes,drop = FALSE]
  if (!is.null(res.mca$quali.sup$eta2))
    coord.illu <- res.mca$quali.sup$eta2[,axes,drop = FALSE]
  if (!is.null(res.mca$quanti.sup$coord))
    coord.illuq <- res.mca$quanti.sup$coord[,axes,drop = FALSE] ^ 2
  if (is.na(test.invisible[1])) {
    coo <- rbind(coo,coord.actif)
    if (lab.var) {
      labe <- c(labe,rownames(coord.actif))
    } else
      labe <- c(labe,rep("",nrow(coord.actif)))
    coll <- c(coll,rep(col.var,nrow(coord.actif)))
    ipch <- c(ipch,rep(20,nrow(coord.actif)))
    fonte <- c(fonte,rep(1,nrow(coord.actif)))
  }
  if ((!is.null(res.mca$quali.sup$eta2)) &&
      (is.na(test.invisible[2]))) {
    coo <- rbind(coo,coord.illu)
    if (lab.quali.sup) {
      labe <- c(labe,rownames(coord.illu))
    } else
      labe <- c(labe,rep("",nrow(coord.illu)))
    coll <- c(coll,rep(col.quali.sup,nrow(coord.illu)))
    ipch <- c(ipch,rep(1,nrow(coord.illu)))
    fonte <- c(fonte,rep(3,nrow(coord.illu)))
  }
  if ((!is.null(res.mca$quanti.sup$coord)) &&
      (is.na(test.invisible[3]))) {
    coo <- rbind(coo,coord.illuq)
    if (lab.quanti.sup) {
      labe <- c(labe,rownames(coord.illuq))
    } else
      labe <- c(labe,rep("",nrow(coord.illuq)))
    coll <- c(coll,rep(col.quanti.sup,nrow(coord.illuq)))
    ipch <- c(ipch,rep(1,nrow(coord.illuq)))
    fonte <- c(fonte,rep(3,nrow(coord.illuq)))
  }
  g <- ggplot(as.data.frame(coo), aes(coo[, 1], coo[, 2], label = labe, colour =
  ↪  coll,shape = as.factor(ipch))) + labs(title = title,x = lab.x, y = lab.y)
  ↪  + geom_point() + geom_hline(yintercept = 0, linetype = "dashed") +
  ↪  geom_vline(xintercept = 0, linetype = "dashed") + xlim(xlim) + ylim(ylim)
  ↪  + geom_text_repel(aes(colour = coll, fontface = fonte)) +
  ↪  scale_colour_solarized()
  g
}
```

Πηγαίος Κώδικας Α΄.4: Συνάρτηση Παραγωγής Διαγράμματος "Quantitative variables".

```
mapQuanti <-
  function(x, axes = c(1, 2), choix = c("ind","var","quanti.sup"),
           xlim = NULL, ylim = NULL, invisible = c("none","ind", "var", "ind.sup",
           ↪  "quali.sup", "quanti.sup"),
           col.ind = "blue", col.var = "red", col.quali.sup = "darkgreen",
           col.ind.sup = "darkblue", col.quanti.sup = "blue",
           label = c("all","none","ind", "var", "ind.sup", "quali.sup",
           ↪  "quanti.sup"), title = NULL, habillage = "none", palette =
             NULL,
           autoLab = c("auto","yes","no"),new.plot = FALSE,select =
             NULL,selectMod = NULL, unselect = 0.7, shadowtext = FALSE,...) {
    label <- match.arg(label,c("all","none","ind", "var", "ind.sup", "quali.sup",
    ↪  "quanti.sup"),several.ok = TRUE)
    choix <- match.arg(choix,c("ind","var","quanti.sup"))
    autoLab <- match.arg(autoLab,c("auto","yes","no"))
    if (autoLab == "yes")
      autoLab = TRUE
    if (autoLab == "no")
      autoLab = FALSE
    autoLab = FALSE
    invisible <-
      match.arg(invisible,c("none","ind", "var", "ind.sup", "quali.sup",
      ↪  "quanti.sup"),several.ok = TRUE)
    if ("none" %in% invisible)
      invisible = NULL

    res.mca <- x
    if (!inherits(res.mca, "MCA"))
      stop("non convenient data")
    if (is.numeric(unselect))
      if ((unselect > 1) |
          (unselect < 0))
        stop("unselect should be betwwen 0 and 1")

    lab.x <-
      paste("Dim ",axes[1]," (",format(res.mca$eig[axes[1],2],nsmall = 2,digits =
                                       2),"%)",sep = "")
    lab.y <-
      paste("Dim ",axes[2]," (",format(res.mca$eig[axes[2],2],nsmall = 2,digits =
                                       2),"%)",sep = "")

    if (!is.null(res.mca$quanti.sup)) {
      if ((new.plot) &
          !nzchar(Sys.getenv("RSTUDIO_USER_IDENTITY")))
        dev.new()
      if (is.null(palette))
        palette(
          c("black", "red", "green3", "blue", "cyan", "magenta", "darkgray",
          ↪  "darkgoldenrod", "darkgreen", "violet", "turquoise", "orange",
          ↪  "lightpink", "lavender", "yellow", "lightgreen", "lightgrey",
          ↪  "lightblue", "darkkhaki", "darkmagenta", "darkolivegreen",
          ↪  "lightcyan", "darkorange", "darkorchid", "darkred", "darksalmon",
          ↪  "darkseagreen", "darkslateblue", "darkslategray", "darkslategrey",
          ↪  "darkturquoise", "darkviolet", "lightgray", "lightsalmon",
          ↪  "lightyellow", "maroon"
          )
        )
      if (is.null(title))
        title <- "Supplementary variables on the MCA factor map"
      x.cercle <- seq(-1, 1, by = 0.01)
      y.cercle <- sqrt(1 - x.cercle ^ 2)
```

```
    g <- ggplot(circleFun(c(0,0),2,npoints = 100), aes(x,y)) + geom_path(color =
    ↪  "#595959") + labs(title = title,x = lab.x, y = lab.y) +
    ↪  geom_rect(aes(xmin = -1,ymin = -1,xmax = 1,ymax = 1), fill = NA) +
    ↪  geom_hline(yintercept = 0, linetype = "dashed") + geom_vline(xintercept
    ↪  = 0, linetype = "dashed") + xlim(c(-2.1,2.1)) + ylim(c(-1.1,1.1))

    if (!is.null(select)) {
      if (mode(select) == "numeric")
        selection <- select
      else {
        if (sum(rownames(res.mca$quanti.sup$coord) %in% select) != 0)
          selection <- which(rownames(res.mca$quanti.sup$coord) %in% select)
        else {
          if (grepl("coord",select))
            selection <-
              (rev(order(
                apply(res.mca$quanti.sup$coord[,axes] ^ 2,1,sum)
              )))[1:min(nrow(res.mca$quanti.sup$coord),sum(as.integer(unlist(
                strsplit(select,"coord")
              )),na.rm = T))]
          if (is.integer(select))
            selection <- select
        }
      }
      res.mca$quanti.sup$coord = res.mca$quanti.sup$coord[selection,,drop =
      ↪  FALSE]
    }
    df <- data.frame(
      x = numeric(),
      y = numeric(),
      xend = numeric(),
      yend = numeric(),
      pos1 = numeric(),
      pos2 = numeric(),
      labels = character(),stringsAsFactors = FALSE
    )
    for (v in 1:nrow(res.mca$quanti.sup$coord)) {
      df[v,1] <- 0
      df[v,2] <- 0
      df[v,3] <- res.mca$quanti.sup$coord[v, axes[1]]
      df[v,4] <- res.mca$quanti.sup$coord[v, axes[2]]
      pos1 <- 0.5
      pos2 <- 0.5
      if (abs(res.mca$quanti.sup$coord[v,axes[1]]) >
      ↪  abs(res.mca$quanti.sup$coord[v,axes[2]])) {
        if (res.mca$quanti.sup$coord[v,axes[1]] >= 0)
          pos1 <- 0
        else
          pos1 <- 1
      }
      else {
        if (res.mca$quanti.sup$coord[v,axes[2]] >= 0)
          pos2 <- 0
        else
          pos2 <- 1
      }
      df[v,5] <- pos1
      df[v,6] <- pos2
      df[v,7] <- rownames(res.mca$quanti.sup$coord)[v]
    }
    g <- g + geom_segment(data = df, mapping = aes(x = x, y = y, xend = xend,
    ↪  yend = yend), arrow = arrow(), size = 0.4, color = col.quanti.sup)
    g <- g + geom_text(data = df,aes(x = xend,y = yend,label = labels,hjust =
    ↪  pos1,vjust = pos2,colour = col.quanti.sup))
    g + scale_colour_solarized()
  }
}
```

Πηγαίος Κώδικας Α΄.5: Συνάρτηση Παραγωγής Διαγράμματος "Within-cluster inertia".

```
clust_plots <- function(x,
                        type = "scree",
                        cutoff = 0.05,
                        title = "") {
  object <- x
  rm(x)

  object <- object / max(object)

  plot_var <- list()
  if (type == "scree") {
    object <-
      data.frame(height = object[object > cutoff], nr_clus = length(object[object >
      ↪  cutoff]):1)
    return(
      ggplot(object,aes(
        x = factor(nr_clus,levels = nr_clus), y = height, group = 1
      )) +
        geom_line(
          colour = "#268bd2", linetype = 'dotdash', size = .7
        ) +
        geom_point(colour = "#268bd2", size = 4) +
        scale_y_continuous(labels = scales::percent) +
        labs(
          list(title = "Scree Plot", x = "# clusters",
               y = "Within-cluster inertia")
        ) + theme_solarized_2(light = F)
    )
  }
  else {
    object <- object[object > cutoff]
    object <- (object - lag(object)) / lag(object)
    object <-
      na.omit(data.frame(bump = object, nr_clus = paste0((
        length(object) + 1
      ):2, "-", length(object):1)))
    return(
      ggplot(object, aes(
        x = factor(nr_clus, levels = nr_clus), y = bump
      )) +
        geom_bar(
          stat = "identity", alpha = 1, fill = "#268bd2"
        ) +
        scale_y_continuous(labels = scales::percent) +
        labs(
          list(title = "Difference in within-cluster inertia",
               x = "# clusters", y = "% Difference in within-cluster inertia")
        ) + theme_solarized_2(light = F)
    )
  }

}
```

Πηγαίος Κώδικας Α΄.6: Συνάρτηση Παραγωγής Διαγράμματος "Categories and Axes Description".

```
plot.catdes <- function (x,col="deepskyblue",show="all",numchar=10,...){
  l=max(length(x$quanti),length(x$category))
  long=rep(0,l)
  list.catdes=list(long)
  minimum=0
  maximum=0
  count=0
  for (i in 1:l){
    if (!is.null(x$quanti[[i]])){
      quanti=as.data.frame(x$quanti[[i]])
      quanti.catdes=as.vector(quanti[,1])
      names(quanti.catdes)=rownames(quanti)
    } else quanti.catdes=NULL
    if (!is.null(x$category[[i]])){
      category=as.data.frame(x$category[[i]])
      category.catdes=as.vector(category[,5])
      names(category.catdes)=rownames(category)
    } else category.catdes=NULL
    if (show=="all") catdes.aux=c(quanti.catdes,category.catdes)
    if (show=="quanti") catdes.aux=quanti.catdes
    if (show=="quali") catdes.aux=category.catdes

    if (!is.null(catdes.aux)) {
      count=count+1
      long[i]=length(catdes.aux)
      mn = c(catdes.aux,minimum)
      mx = c(catdes.aux,maximum)
      minimum=min(mn[is.finite(mn)])
      maximum=max(mx[is.finite(mx)])
    } else long[i]=0
    list.catdes[[i]]=catdes.aux
  }
  if(count!=0){
    if (count<=4){
      numc=count
      numr=1
    } else{
      numc=4
      numr=round(count/4)+1
    }
    catdes.aux=sort(list.catdes[[1]],decreasing=FALSE)
    df <- data.frame(y=catdes.aux, clust=1)
    for(i in 2:l){
      catdes.aux=sort(list.catdes[[i]],decreasing=FALSE)
      df1 <- data.frame(y=catdes.aux, clust=i)
      df <- rbind(df, df1)
    }
    t <- data.frame(strsplit(rownames(df), "="))
    t <- unlist(t[dim(t)[1],])
    df$x <- t
    df$x <- factor(df$x, levels = df$x)
    df <- df[is.finite(df$y), ]
  }
  g <- ggplot(df, aes(x, y)) + geom_bar(fill = "#268bd2", alpha=0.7, stat =
  ↪  "identity", position = "identity") + xlab("") + ylab("v.test") +
    facet_wrap(~ clust, scales = "free", ncol = 2) + theme_solarized_2(light = F) +
    ↪  theme(axis.text.x = element_text(angle = 45, hjust = 1))
  g
}
```

Βιβλιογραφία

[1] J. Taylor, *Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics*, English, ser. IBM Press. Pearson Education, 2011, ISBN: 9780132884440. [Online]. Available: https://books.google.gr/books?id=X2pJuopB2JsC.

[2] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*, English. Wiley, 2011, ISBN: 9781119965473. [Online]. Available: https://books.google.gr/books?id=Yl_yAn2bhZ0C.

[3] A. Bara, I. Botha, V. Diaconita, I. Lungu, A. Velicanu, and M. Velicanu, "A model for business intelligence systems' development", English, *Informatica Economica*, vol. 13, no. 4, pp. 99–108, 2009. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.492.7906&rep=rep1&type=pdf (visited on 11/24/2015).

[4] Z. Martin, "Methodology framework for analysis and design of business intelligence systems", English, *Applied Mathematical Sciences*, vol. 7, no. 4, pp. 1523 –1528, 2013. [Online]. Available: http://www.m-hikari.com/ams/ams-2013/ams-29-32-2013/zavodnyAMS29-32-2013.pdf.

[5] D. Cook, A. Buja, D. Lang, D. Swayne, H. Hofmann, H. Wickham, and M. Lawrence, *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi*, English, ser. Use R! Springer New York, 2007, ISBN: 9780387717623. [Online]. Available: https://books.google.gr/books?id=sC0Ij2r_KLcC.

[6] D. Hoaglin, F. Mosteller, and J. Tukey, *Fundamentals of Exploratory Analysis of Variance*, English, ser. Wiley Series in Probability and Statistics. Wiley, 2009, ISBN: 9780470317662. [Online]. Available: https://books.google.gr/books?id=2KC9uig0_04C.

[7] M. DeGroot and M. Schervish, *Probability and Statistics*, English, ser. Addison-Wesley series in statistics. Addison-Wesley, 2002, ISBN: 9780201524888. [Online]. Available: https://books.google.gr/books?id=iH4ZAQAAIAAJ.

[8] B. Everitt and T. Hothorn, *An Introduction to Applied Multivariate Analysis with R*, English, ser. Use R! Springer New York, 2011, ISBN: 9781441996503. [Online]. Available: https://books.google.gr/books?id=BB51TWe94EwC.

[9] J. Ledolter, *Data Mining and Business Analytics with R*, English. Wiley, 2013, ISBN: 9781118572153. [Online]. Available: https://books.google.gr/books?id=gLnuxUmk5eEC.

[10] F. Provost and T. Fawcett, *Data Science for Business*, English. O'Reilly, 2013, ISBN: 9781449361327. [Online]. Available: https://books.google.gr/books?id=_1b4nAEACAAJ.

[11] D. Bălăceanu, "Components of a business intelligence software solution", English, *Informatica Economică*, vol. 2, p. 42, 2007. [Online]. Available: http://revistaie.ase.ro/content/42/balaceanu.pdf (visited on 11/24/2015).

[12] M. Alexander, J. Robert, and S. Edgar, *Better decision making with proper business intelligence*, English. [Online]. Available: https://www.atkearney.com/documents/10192/247903/Better_Decision_Making_with_Proper_Business_Intelligence.pdf.

[13] C. Julander, "Basket analysis: a new way of analysing scanner data", English, *International Journal of Retail & Distribution Management*, vol. 20, no. 7, 1992. DOI: 10.1108/09590559210022362. eprint: http://dx.doi.org/10.1108/09590559210022362. [Online]. Available: http://dx.doi.org/10.1108/09590559210022362.

[14] G. J. Russell and A. Petersen, "Analysis of cross category dependence in market basket selection", English, *Journal of Retailing*, vol. 76, no. 3, pp. 367–392, 2000.

[15] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases", English, *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993, ISSN: 0163-5808. DOI: 10.1145/170036.170072. [Online]. Available: http://doi.acm.org/10.1145/170036.170072.

[16] W. H. Inmon, *Building the Data Warehouse*, English. New York, NY, USA: John Wiley & Sons, Inc., 1992, ISBN: 0471569607.

[17] P. A. K. M. R. B. D. V. N. M. M. C. Henry Morris Stephen Graham, *The financial impact of business analytics*, English, IBM, 2002. [Online]. Available: https://www-07.ibm.com/services/hk/pdf/finimpact.pdf.

[18] J. Benzécri, *L'analyse des données: L'analyse des correspondances*, English, ser. L'analyse des données: leçons sur l'analyse factorielle et la reconnaissance des formes et travaux. Dunod, 1973, ISBN: 9782040072254. [Online]. Available: https://books.google.gr/books?id=sDTwAAAAMAAJ.

[19] F. Husson, S. Lê, and J. Pagès, *Exploratory multivariate analysis by example using R*, English, ser. Chapman & Hall/CRC computer science and data analysis. Boca Raton: CRC Press, 2011, 228 pp., ISBN: 978-1-4398-3580-7.

[20] H. Hotelling, *Analysis of a Complex of Statistical Variables Into Principal Components*, English. Warwick & York, 1933. [Online]. Available: https://books.google.gr/books?id=qJfXAAAAMAAJ.

[21] M. Greenacre, *Theory and Applications of Correspondence Analysis*, English. Academic Press, 1984, ISBN: 9780122990502. [Online]. Available: https://books.google.gr/books?id=LsPaAAAAMAAJ.

[22] L. Lebart, A. Morineau, and K. M. Warwick, *Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices*, English, ser. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1984, ISBN: 9780471867432. [Online]. Available: https://books.google.gr/books?id=iJwQAQAAIAAJ.

[23] J. H. Ward, "Hierarchical grouping to optimize an objective function", English, *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963. DOI: 10.1080/01621459.1963.10500845. eprint: http://www.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845.

[24] M. Hahsler, C. Buchta, B. Gruen, and K. Hornik, *Arules: mining association rules and frequent itemsets*, English, R package version 1.4-1, 2016. [Online]. Available: http://CRAN.R-project.org/package=arules.

[25] G. Snedecor and W. Cochran, *Statistical methods*, English. Iowa State University Press, 1980, ISBN: 9780813815602. [Online]. Available: https://books.google.gr/books?id=x4EuAAAAIAAJ.

[26] A. Fachantidis, A. Tsiaras, G. Tsoumakas, and I. Vlahavas, "Segmento: an r-based visualization-rich system for customer segmentation and targeting", English, in *Proceedings of the 9th Hellenic Conference on Artificial Intelligence*, ser. SETN '16, Thessaloniki, Greece: ACM, 2016, 23:1–23:4, ISBN: 978-1-4503-3734-2. DOI: 10.1145/2903220.2903245. [Online]. Available: http://doi.acm.org/10.1145/2903220.2903245.