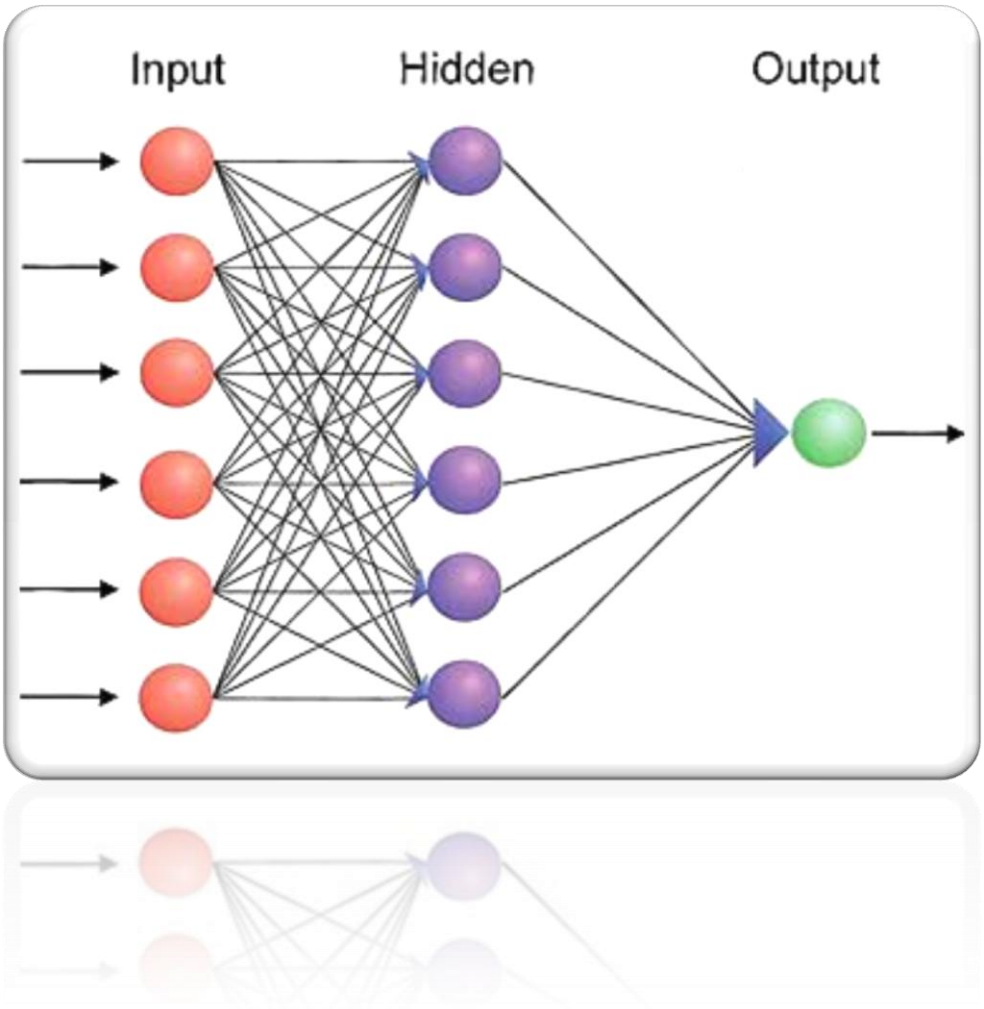
ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΠΡΩΗΝ ΤΜΗΜΑ ΕΦΑΡΜΟΓΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΔΙΟΙΚΗΣΗ ΚΑΙ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ (ΠΑΤΡΑ)

« Τεχνικές Οπτικοποίησης στην Εξόρυξη Δεδομένων »



Πτυχιακή Εργασία των:

Ζαχαρίας Ξενάκης Α.Μ: 12780, Παναγιώτα Μουτοπούλου Α.Μ: 12682

Επιβλέπων καθηγητής:

Γεράσιμος Αντζουλάτος

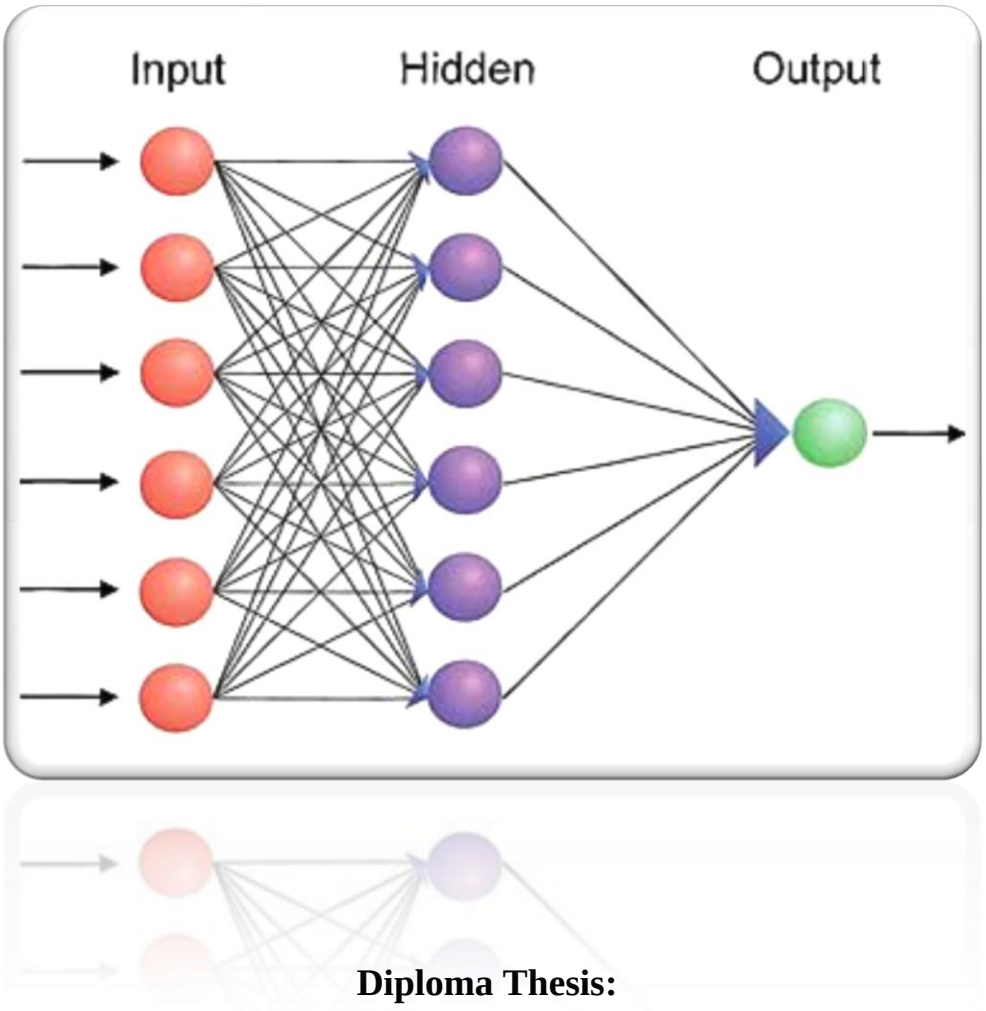
Πάτρα 2/09/2015

TECHNOLOGICAL EDUCATIONAL INSTITUTION OF WESTERN GREECE

FORMER TECHNOLOGICAL INSTITUTE OF MANAGEMENT AND ECONOMICS DEPARTMENT OF APPLIED COMPUTING ON MANAGEMENT AND THE ECONOMY

DEPARTMENT OF BUSINESS ADMINISTRATION (PATRA)

«Visual Data Mining Techniques and Tools for Data Visualization and Mining»



Diploma Thesis:

Zacharias Xenakis A.M.: 12780, Panagiota Moutopoulou A.M.: 12682

Supervised Professor:

Gerasimos Antzoulatos

Patra 2/09/2015

Ευχαριστίες

Θα θέλαμε αρχικά να ευχαριστήσουμε τον επιβλέποντα καθηγητή μας, κ. Γεράσιμο Αντζουλάτο για την καθοδήγηση και υποστήριξη που μας παρείχε αλλά και τις γνώσεις που μας μετέδωσε καθ΄ όλη την διάρκεια εκπόνησης αυτής της πτυχιακής εργασίας.

Επίσης θα θέλαμε να ευχαριστήσουμε τις οικογένειές μας για την πολύπλευρη υποστήριξή τους σε όλη την διάρκεια των σπουδών μας.

Περίληψη

Η οπτικοποίηση των δεδομένων είναι μια απεικόνιση της πληροφορίας κυρίως σε μια γραφική ή πινακοειδή μορφή. Η επιτυχής οπτικοποίηση απαιτεί τα δεδομένα και οι πληροφορίες να αναπαρίστανται με τρόπο εύγλωτο και ευκολοκατανόητο σε μια οπτική μορφή. Το κυριότερο κίνητρο για τη χρήση της οπτικοποίησης είναι πως οι άνθρωποι αντιλαμβάνονται εύκολα και είναι σε θέση να απορροφήσουν μεγάλες ποσότητες οπτικής πληροφορίας και να εντοπίσουν υποδείγματα, τάσεις και πρότυπα σε αυτά. Στο πεδίο της Εξόρυξης Δεδομένων οι τεχνικές οπτικοποίησης μπορούν να συνεισφέρουν τόσο στην διερεύνηση των δεδομένων στη φάση της προεπεξεργασίας τους, όσο και στην ερμηνεία των αποτελεσμάτων της εφαρμογής των μεθοδολογιών της Εξόρυξης Δεδομένων. Η γνώση που εξάγεται θα πρέπει να εκφράζεται σε οπτικές αναπαραστάσεις ώστε να αποτυπώνεται η πληροφορία εύκολα, να είναι κατανοητή στον άνθρωπο έτσι ώστε να είναι άμεση η χρησιμοποίησή της. Στην παρούσα εργασία γίνεται μια προσπάθεια κατηγοριοποίησης των τεχνικών οπτικοποίησης σε τεχνικές βασισμένη στο πλήθος και στον τύπο των εμπλεκόμενων χαρακτηριστικών. Αρχικά, παρουσιάζονται τεχνικές οπτικοποίησης από την Διερευνητική Ανάλυση Δεδομένων, όπως είναι τεχνικές για μονομεταβλητή απεικόνιση (ιστογράμματα, γραφήματα γραμμών, στηλών, θηκογράμματα κ.α.), τεχνικές διμεταβλητής απεικόνισης δεδομένων, όπως είναι τα γραφήματα διασποράς. Εν συνεχεία, περιγράφονται τεχνικές οπτικοποίησης πολυμεταβλητών δεδομένων και ερμηνείας των αποτελεσμάτων της Εξόρυξης Δεδομένων, όπως είναι η παλινδρόμηση, τα δέντρα απόφασης, τα δενδρογράμματα, οι κύβοι OLAP. Επιπλέον, οι προαναφερθείσες τεχνικές εφαρμόστηκαν στο σύνολο δεδομένων που δημιουργήθηκε από τις απαντήσεις για την αποτύπωση της ψυχολογικής κατάστασης φοιτητών. Για την καταγραφή, αποτύπωση και αξιολόγηση της ψυχολογικής κατάστασης χρησιμοποιήθηκε η σταθμισμένη κλίμακα ψυχοπαθολογίας Symptom Checklist-90 (SCL-90), η οποία εξετάζει ένα ευρύ φάσμα ψυχολογικών προβλημάτων και συμπτωμάτων ψυχοπαθολογίας. Για την υλοποίηση της εφαρμογής των τεχνικών οπτικοποίησης χρησιμοποιήθηκε το προγραμματιστικό πακέτο λογισμικού R.

Abstract

The visualization of data is an information portrayal in a graphic or tabular format. The successful visualization demands the data and the information to be described with an understandable way in a visual format. The significant initiative for the visualization use is the manner with which people easily understand and are capable enough to receive much visual information and find small samples, inclinations and symbols on them. As far as concerned the data mining, the techniques of visualization can help not only the data to be broadened in the preparation face but also to be interpreted. The knowledge should be expressed with specific ways so as to be understandable and useful. In this work, visualization techniques classified in the number and the type of characteristics. At the beginning, visualization techniques are presented by the exploratory data analysis. These techniques concern univariate portrayal (line charts, histograms, pie charts and boxplots e.t.c), bivariate data such as scatter plots. Moreover, for the illustration of multivariate data in order the data mining to be analyzed, were used regression, dendrograms, decision trees and OLAP cubes. Furthermore, the quoted techniques applied in the answers for the description of psychological situation which had high school students. For the recording, description and evaluation of this situation, Symptom Checklist-90 (SCL-90) was used to. The (SCL-90) concerns a wide range of psychological problems and psychopathic symptoms. Last but not least, all these visualizuation techniques were materialized by using the software programming package R.

Περιεχόμενα

Πίνακας Εικόνων

Εισαγωγή

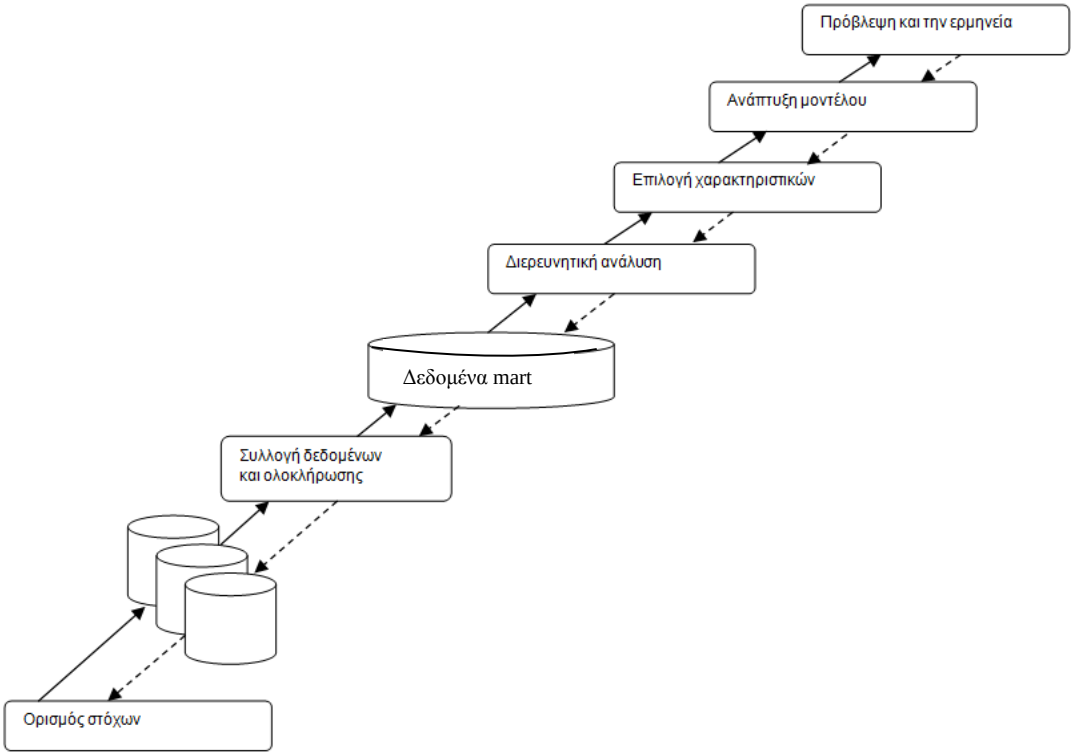
Η εξόρυξη δεδομένων είναι ένας τρόπος μάθησης από το παρελθόν ούτως ώστε να λαμβάνουν καλύτερες αποφάσεις στο μέλλον και έχει σχεδιαστεί για την επίλυση προβλημάτων των επιχειρήσεων.

Η διαδικασία εξόρυξης δεδομένων βασίζεται στην επαγωγική μάθηση μεθόδων, των οποίων βασικός σκοπός είναι να αντλήσει γενικούς κανόνες, ξεκινώντας από ένα σύνολο παραδειγμάτων, αποτελούμενο από παλαιότερες παρατηρήσεις που καταγράφονται σε μία ή περισσότερες βάσεις δεδομένων. Με άλλα λόγια, ο σκοπός μιας ανάλυσης εξόρυξης δεδομένων είναι να συνάγει ορισμένα συμπεράσματα που αρχίζει από ένα δείγμα των παρατηρήσεων του παρελθόντος και γενικεύουν τα συμπεράσματα σε σχέση με το σύνολο του πληθυσμού, με τέτοιο τρόπο ώστε να είναι όσο το δυνατόν ακριβέστερες. Τα μοντέλα και τα σχέδια που προσδιορίζονται με αυτόν τον τρόπο μπορεί να πάρουν διάφορες μορφές, οι οποίες θα περιγραφούν στις επόμενες ενότητες, όπως γραμμικών εξισώσεων, συμπλέγματα, γραφήματα και δέντρα απόφασης (Diane, 1999) (Ning Tan, Steinbach, & Kumar, 2010).

Η οπτικοποίησης είναι μια μεθοδολογία ανάλυσης των στοιχείων εξόρυξης. Δίνει την δυνατότητα στην διαδικασία της εξόρυξης δεδομένων να παρέχει μια απλή και συνοπτική αντιπροσώπευση των πληροφοριών που αποθηκεύονται σε ένα μεγάλο σύνολο δεδομένων. Αρχικά είναι δύσκολο να επιτευχθεί μια ουσιαστική οπτικοποίηση των δεδομένων ωστόσο με έναν καλό σχεδιασμό μπορεί να αντιπροσωπευθεί και να αναλυθεί μέσω ενός γραφήματος (Diane, 1999).

Κεφάλαιο 1ο Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων είναι η διαδικασία της αυτόματης ανακάλυψης χρήσιμων πληροφοριών μέσα από μεγάλες δεξαμενές δεδομένων. Οι τεχνικές εξόρυξης δεδομένων εφαρμόζονται για να ερευνήσουν σε βάθος μεγάλες βάσεις δεδομένων με σκοπό να βρεθούν νέα και χρήσιμα πρότυπα, τα οποία σε διαφορετική περίπτωση θα παρέμεναν άγνωστα. Επίσης παρέχουν δυνατότητες πρόβλεψης του αποτελέσματος μιας μελλοντικής παρατήρησης. Η εξόρυξη δεδομένων είναι αναπόσπαστο κομμάτι της Ανεύρεσης Γνώσης από τις βάσεις δεδομένων η οποία αποτελεί την συνολική διεργασία της μετατροπής ακατέργαστων δεδομένων σε σημαντικές πληροφορίες (Ning Tan, Steinbach, & Kumar, 2010). Στην (Εικόνα 1. Βήματα της διαδικασίας Ανεύρεσης Γνώσης από Βάσεις Δεδομένων) παρουσιάζονται οι κύριες φάσεις της διαδικασίας Ανεύρεσης Γνώσης από Βάσεις Δεδομένων (Knowledge Discovery in Databases - KDD) (Carlo, 2009) (Han & Kameber, 2006).



Εικόνα 1. Βήματα της διαδικασίας Ανεύρεσης Γνώσης από Βάσεις Δεδομένων

Ορισμός στόχων. Η ανάλυση της εξόρυξης δεδομένων πραγματοποιείται σε συγκεκριμένους τομείς εφαρμογής και προορίζετε για φορείς λήψης αποφάσεων για την παροχή χρήσιμων γνώσεων. Προκειμένου να διατυπωθούν εύλογοι και καλά καθορισμένοι στόχοι της έρευνας οι ειδικοί πρέπει να έχουν ιδιαίτερα καλές γνώσεις

του προβλήματος διαφορετικά μπορεί να εμφανιστούν σφάλματα σε μελλοντικές προσπάθειες για την εξόρυξη δεδομένων για το συγκεκριμένο πρόβλημα.

Συλλογή δεδομένων και την ολοκλήρωση. Μόλις προσδιοριστούν οι στόχοι της έρευνας, η συγκέντρωση των δεδομένων αρχίζει. Τα δεδομένα μπορούν να προέρχονται από διαφορετικές πηγές και επομένως μπορεί να απαιτηθεί χρόνος για την ολοκλήρωσης της. Οι πηγές των δεδομένων μπορεί να είναι εσωτερικές, εξωτερικές ή ένας συνδυασμός των δύο.

Δεδομένων Mart. Η έξοδος που θα δώσει ένα martδεδομένοθα είναι ένας συγκεντρωτικός πίνακας με δεδομένα που θα απεικονίζουν αυτό ακριβώς που έθεσε ως ερώτημα ο αναλυτής. Αυτοί οι πίνακες μπορούν να αποθηκευθούν σε ένα σύστημα διαχείρισης σχεσιακής βάσης δεδομένων (RDBMS), Oracle, Microsoft SQL Server και IBM.

Διερευνητική ανάλυση. Μια προκαταρκτική ανάλυση των δεδομένων πραγματοποιείται με σκοπό να πάρει δεδομένα εξοικειωμένα με τις διαθέσιμες πληροφορίες και τη διεξαγωγή της εκκαθάρισης. Τα δεδομένα που αποθηκεύονται σε αποθήκες δεδομένων επεξεργάζονται με το χρόνο κατά τρόπον ώστε να αφαιρέσουν της συντακτικές ανακολουθίες φόρτωσης.

Χαρακτηριστικό επιλογή. Η καταλληλότητα για τα διαφορετικά χαρακτηριστικά αξιολογείται σε σχέση με τους στόχους της ανάλυσης. Αφαιρούνται τα χαρακτηριστικά που δεν έχουν νόημα, προκειμένου να καθαρίσει τις πλεονάζουσες πληροφορίες από το σύνολο δεδομένων.

Μοντέλο ανάπτυξης. Συνήθως η εκπαίδευση των μοντέλων γίνεται χρησιμοποιώντας ένα δείγμα των καταγραφών που προέρχονται από το αρχικό σύνολο δεδομένων, η διαγνωστική ακρίβεια κάθε μοντέλου που δημιουργείται μπορεί να εκτιμηθεί χρησιμοποιώντας το υπόλοιπο των δεδομένων. Το διαθέσιμο σύνολο δεδομένων είναι χωρισμένο σε δύο υποσύνολα. Συνήθως το μέγεθος του δείγματος από την εκπαίδευση που θα επιλεγεί να είναι σχετικά μικρό, αν και σημαντικό, από στατιστική άποψη.Το δεύτερο υποσύνολο είναι το σύνολο δοκιμής και χρησιμοποιείται για να αξιολογήσει την ακρίβεια των εναλλακτικών μοντέλων που παράγονται κατά τη φάση της κατάρτισης, για να μπορέσει να προσδιοριστεί το καλύτερο μοντέλο πραγματικών μελλοντικών προβλέψεων.

Πρόβλεψη και την ερμηνεία. Μόλις ολοκληρωθεί η διαδικασία εξόρυξης δεδομένων, το μοντέλο που επιλέγετε μεταξύ εκείνων που έχουν δημιουργηθεί κατά τη διάρκεια της φάσης ανάπτυξης, πρέπει να εφαρμοστεί και να χρησιμοποιηθεί για να επιτευχθούν οι στόχοι που είχαν αρχικά προσδιοριστεί και θα πρέπει να ενσωματωθεί στις διαδικασίες υποστήριξης λήψης αποφάσεων.

Η διαδικασία εξόρυξης δεδομένων περιλαμβάνει κύκλους ανατροφοδότησης, που εκπροσωπούνται από το διάστικτο βέλος στο σχήμα (Εικόνα 1. Βήματα της διαδικασίας Ανεύρεσης Γνώσης από Βάσεις Δεδομένων), που μπορεί να υποδεικνύει μια επιστροφή σε κάποια προηγούμενη φάση, ανάλογα με το αποτέλεσμα των μετέπειτα φάσεων. Η συμμετοχής και της αλληλεπίδρασης από διάφορους επαγγελματικούς ρόλους προκειμένου η διαδικασία εξόρυξης δεδομένων να εμφανίσει αληθή αποτελέσματα:

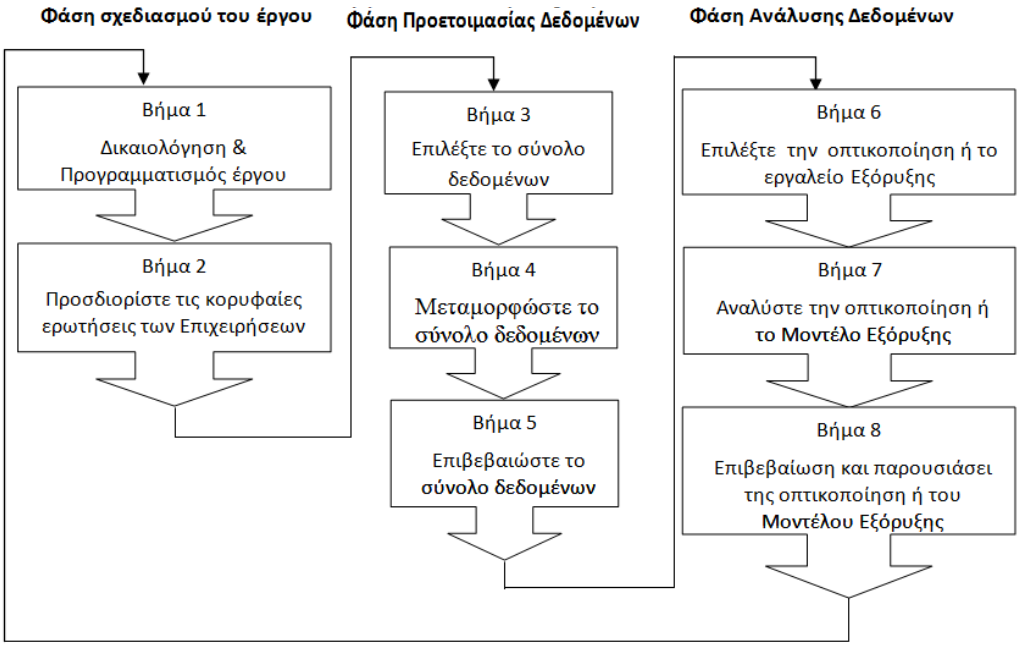
- εμπειρογνώμονας στον τομέα της εφαρμογής, αναμένεται να καθορίσει τους αρχικούς στόχους της ανάλυσης
- ειδικός στα πληροφοριακά συστήματα της εταιρείας, αναμένεται να επιβλέπουν την πρόσβαση στις πηγές πληροφόρησης.
- ειδικός στη μαθηματική θεωρία της μάθησης και στατιστικές, για την ανάλυση των διερευνητικών δεδομένων και για την παραγωγή των μοντέλων πρόβλεψης.

1.1 Τεχνικές Οπτικοποίησης Εξόρυξης Δεδομένων & Εργαλεία

Εφαρμόζοντας απεικονίσεις και τεχνικές εξόρυξης δεδομένων, οι επιχειρήσεις μπορούν να αξιοποιήσουν πλήρως επιχειρησιακά στοιχεία, να ανακαλύψουν άγνωστες τάσεις και συμπεριφορές των καταναλωτών τους χρησιμοποιώντας εργαλεία και τεχνικές οπτικοποίησης δεδομένων, επιτυγχάνοντας τη δημιουργία εικόνων δύο και τριών διαστάσεων οι οποίες εύκολα μπορούν να ερμηνευτούν. Η απεικόνιση θεωρείται κλειδί για την βοήθεια που παρέχει στις επιχειρήσεις και πιο συγκεκριμένα στους αναλυτές καθώς είναι μια δοκιμασμένη μέθοδος για λήψης αποφάσεων (Carlo, 2009).

1.1.1 Τα οκτώ βήματα οπτικοποίησης & η μεθοδολογία εξόρυξης δεδομένων (VDM)

Οι (Soukup&Davidson, 2002) πρότειναν τα οκτώ (8) βήματα οπτικοποίησης που φαίνονται στην Εικόνα 2. Τα οκτώ βήματα οπτικοποίησης.



Εικόνα 2. Τα οκτώ βήματα οπτικοποίησης

Φάση Σχεδιασμού του έργου

Πρώτο βήμα: Η φάση αυτή χωρίζεται σε δύο μεγάλες κατηγορίες οπτικοποίησης:

1) Εργαλεία οπτικοποίησης δεδομένων και τεχνικών, οι οποίες βοηθούν στην δημιουργία δύο και τριών διαστάσεων εικόνων επιχειρηματικών δεδομένων, οι ερμηνείες των οποίων είναι απλές για την απόκτηση γνώσεων και ιδεών για τον επιχειρηματία. Με δύο ή τριών διαστάσεων εικόνων δίνεται η δυνατότητα να αναγνωριστούν οι πληροφορίες που μας ενδιαφέρουν ή να σχεδιαστούν τα σύνολα των δεδομένων της επιχείρησης,

2) Οπτική των δεδομένων εξόρυξης (εργαλεία και τεχνικές), που βοηθούν για την απεικόνιση προτύπων εξόρυξης δεδομένων, να αποκτηθούν γνώσεις και διορατικότητες στα πρότυπα που εντόπισε ο αλγόριθμος εξόρυξης δεδομένων για την λήψη σωστών αποφάσεων και προβλέψεων νέων επιχειρηματικών ευκαιριών.

Και οι δύο κατηγορίες οπτικοποίησης βοηθούν τους ανθρώπους στην ανακάλυψη νέων προτύπων και τάσεων.

Δεύτερο βήμα: Ο προσδιορισμός και η βελτίωση των κορυφαίων επιχειρηματικών ερωτήσεων για να μπορέσει να βρεθεί μέσα από την οπτικοποίηση των δεδομένων και την οπτική της εξόρυξης δεδομένων. Με αυτόν τον τρόπο καθοδηγούμαστε μέσω της χαρτογράφησης στην επιλογή της κορυφαίας επιχειρηματικής ερώτησης για το VDM έργο σε οπτικοποίησης δεδομένων και τον ορισμό του προβλήματος εξόρυξης δεδομένων.

Φάση Προετοιμασίας Δεδομένων

Τρίτο βήμα: Το συγκεκριμένο βήμα ασχολείται με τον τρόπο επιλογής τον δεδομένων που σχετίζονται με την οπτικοποίηση των δεδομένων. Σε αυτό το βήμα χρησιμοποιούνται οι αποθήκες δεδομένων για την δημιουργία και την διατήρηση του συνόλου των επιχειρηματικών δεδομένων για την κάλυψη των επιχειρηματικών ερωτήσεων που βρίσκονται υπό διερεύνηση. Οι διερευνητικές αποθήκες δεδομένων χρησιμοποιούνται στην συνέχεια για εξαγωγή, φόρτωση, και συγχώνευση των πρώτων επιχειρήσεων παραγωγής.

Τέταρτο βήμα: Στο βήμα αυτό υλοποιούνται διάφοροι μετασχηματισμοί στο σύνολο των επιχειρηματικών δεδομένων που αποθηκεύονται στις αποθήκες δεδομένων. Αυτοί οι μετασχηματισμοί γίνονται για να βοηθήσουν στην επέκταση των επιχειρηματικών δεδομένων ώστε να αποκτηθούν περισσότερες δυνατότητες αντίληψης του προβλήματος της επιχειρησιακής έρευνας.

Πέμπτο βήμα: Στο συγκεκριμένο βήμα επαληθεύονται τα σύνολα των δεδομένων, χρησιμοποιούνται τα αναμενόμενα στοιχεία και όλα τα βήματα των προηγούμενων βημάτων πρέπει να έχουν εφαρμοστεί σωστά, ώστε να μην υπάρχουν σφάλματα και ελλιπής δεδομένα της επιχείρησης.

Φάση της Ανάλυσης Δεδομένων

Έκτο βήμα: Η φάση αυτή χωρίζεται σε δύο κατηγορίες οπτικοποίησης:

1) Την απεικόνιση των δεδομένων η οποία είναι η χρήση εικόνων με τις οποίες είναι πιο εύκολα κατανοητές από ότι με ένα σύνολο από αριθμούς ή κανόνων, από έναν άμορφο αλγόριθμο εξόρυξης δεδομένων. Με την απεικόνιση αυτή υπάρχει και η δυνατότητα να επικεντρωθούν στο σύνολο των δεδομένων που χρειάζονται με την επιλογή του κατάλληλου τύπου γραφημάτων, την σωστή επιλογή του χρώματος και χρησιμοποιώντας τις κατάλληλες οντότητες γραφικών.

2) Οι στόχοι των εργαλείων και των τεχνικών οπτικοποίησης δεδομένων είναι να βοηθήσει στη δημιουργία δύο ή τριών διαστάσεων εικόνων από τα έτοιμα σύνολα δεδομένων της επιχείρησης που μας ενδιαφέρει για να μπορέσουν να αναλυθούν και να αποκτηθούν γνώσεις και απόψεις. Επιπλέον μπορούν να χρησιμοποιήθουν τα αποτελέσματα που θα εμφανιστούν στην οπτικοποίηση δεδομένων όπως γραφήματα στηλών, γραφήματα πιτών ώστε να εμφανιστούν τα δεδομένα και οι αναλύσεις των επιχειρήσεων και των φορέων λήψης αποφάσεων. Το ανθρώπινο μυαλό είναι ένα πολύ εξελιγμένο μηχάνημα αναγνώρισης και επεξεργασίας μοτίβων.

Η επιλογή του κατάλληλου εργαλείου οπτικοποίησης δεδομένων ή της τεχνικής εξαρτάται από τη φύση των δεδομένων που αφορούν τις επιχειρήσεις και το υπόβαθρο της δομής της.

Έβδομο βήμα: Το βήμα αυτό ασχολείται με τη χρήση της απεικόνισης των δεδομένων και των μοντέλων εξόρυξης δεδομένων για την απόκτηση επιχειρηματικών ιδεών για να απαντηθούν τα επιχειρηματικά ζητήματα. Επιπλέον θα αξιολογηθούν και θα συγκριθούν η δύναμη του κάθε μοντέλου στα προγνωστικά αποκτώντας έτσι την δυνατότητα να επιλεχθεί το καλύτερο μοντέλο που απευθύνεται στις ερωτήσεις της επιχείρησης.

Όγδοο βήμα: Το τελευταίο βήμα ασχολείται με το μοντέλο οπτικοποίησης και εξόρυξης δεδομένων ώστε να ικανοποιούνται οι επιχειρηματικοί στόχοι. Επίσης παρουσιάζεται η οπτικοποίηση και τα δεδομένα εξόρυξης για τους φορείς των λήψεων αποφάσεων και αναπτύσσονται οι απεικονίσεις και τα μοντέλα εξόρυξης σε περιβάλλον παραγωγής.

1.2 Εργασίες Εξόρυξης Δεδομένων

Οι δραστηριότητες της εξόρυξης δεδομένων μπορούν να υποδιαιρεθούν σε επτά βασικές κατηγορίες, αναλόγως τα καθήκοντα και τους στόχους της ανάλυσης, οι οποίες είναι οι ακόλουθες (Carlo, 2009) (Ning Tan, Steinbach, & Kumar, 2010) (Γκίζα, 2007):

1. Ο χαρακτηρισμός και η διακριτοποίηση (discrimination) προκύπτει πριν την ανάπτυξη του προτύπου ταξινόμησης και γίνεται μια διερευνητική ανάλυση ώστε

να συγκριθούν οι τιμές αν ανήκουν στην ίδια κατηγορία με τις ιδιότητες ή αν υπάρχουν διαφορές μεταξύ τους.

2. Η ταξινόμηση ή κατηγοριοποίηση προσπαθεί να φτιάξει μοντέλα τα οποία περιγράφουν την μεταβλητή στόχο την οποία καλούμε κλάση με κάποια ανεξάρτητα χαρακτηριστικά του συνόλου δεδομένων.

3. Η παλινδρόμιση βάση ορισμού είναι η εργασία εκμάθησης μιας στοχευμένης συνάρτησης f, η οποία απεικονίζει κάθε χαρακτηριστικό X σε μια έξοδο συνεχών τιμών Y, είναι δηλαδή μια τεχνική προγνωστικής μοντελοποίησης, όπου η στοχευμένη μεταβλητή που πρέπει να εκτιμηθεί είναι συνεχής.

4. Η ανάλυση χρονοσειρών σκοπό έχει την πρόβλεψη της τιμής της μεταβλητής προορισμού για μια ή περισσότερες μελλοντικές περιόδους.

5. Οι κανόνες συσχέτισης χρησιμοποιούνται για να εντοπίσουν αν υπάρχει ισχυρή σχέση μεταξύ των ομάδων του συνόλου δεδομένων ώστε να αναπτυχθούν καλύτεροι τρόποι κατανόησης και αλληλεπίδρασης.

6. Η ομαδοποίηση ή συσταδοποίηση αναλύει τα δεδομένα για τα οποία δεν έχουν εκ των προτέρων γνώση για την κατηγορία στην οποία ανήκουν. Επιπλέον η ομαδοποίηση αποτελεί μια μέθοδο η οποία ασχολείται με τεχνικές πολυμεταβλητότητας.

7. Η περιγραφή και η οπτικοποίηση είναι ο τρόπος με τον οποίο παρουσιάζονται τα δεδομένα, οι γραφικές παραστάσεις, οι εικόνες για να μπορέσει να γίνει εύκολα κατανοητό στο κοινό τα δεδομένα τα οποία λαμβάνει.

1.3 Τύποι Δεδομένων

Όσο αναπτύσσεται και ωριμάζει το πεδίο εξόρυξης δεδομένων τόσο μεγαλώνουν και τα σύνολα δεδομένων που μπορούν να αναλυθούν. Υπάρχουν τρεις κύριοι τύποι συνόλων δεδομένων οι οποίοι είναι τα δεδομένα εγγράφων, τα δεδομένα γράφων και τα διατεταγμένα δεδομένα. Πριν όμως περιγράψουμε τις λεπτομέρειες για τα είδη των συνόλων δεδομένων, εξετάζουμε τρία χαρακτηριστικά τα οποία και εφαρμόζονται στα περισσότερα σύνολα δεδομένων και έχουν σημαντική επίδραση για τις τεχνικές εξόρυξης δεδομένων, τα οποία είναι η διάσταση, η σποραδικότητα και η ανάλυση (Ning Tan, Steinbach, & Kumar, 2010).

Διάσταση: Διάσταση συνόλου δεδομένων είναι το πλήθος των χαρακτηριστικών που περιέχουν τα αντικείμενα του συνόλου. Τα δεδομένα που έχουν μικρό αριθμό διαστάσεων είναι ποσοτικά διαφορετικά από τις μεσαίες ή υψηλές πλήθους διαστάσεων δεδομένων.

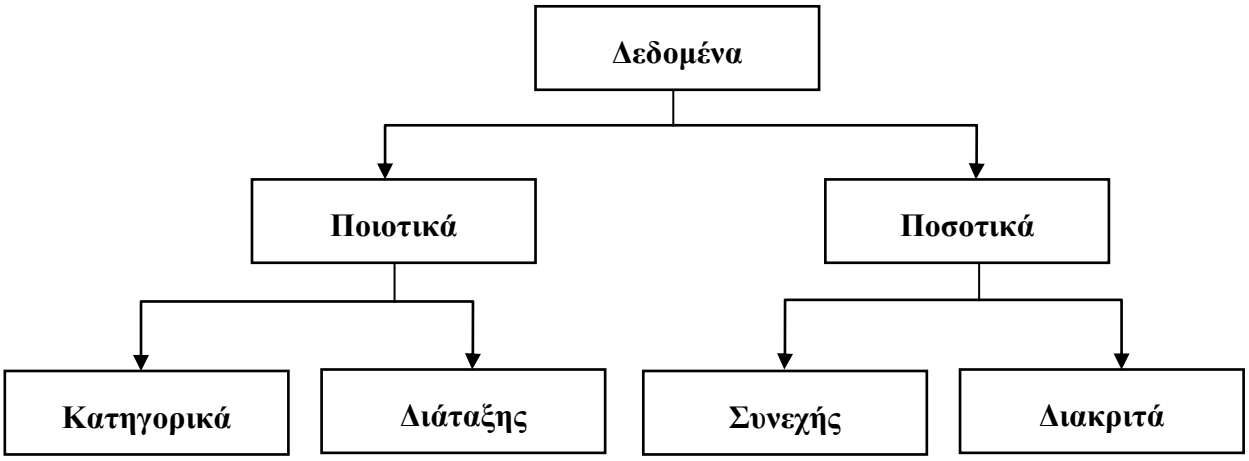
Σποραδικότητα: Η σποραδικότητα έχει ένα πλεονέκτημα γιατί μόνο οι μη μηδενικές τιμές αποθηκεύονται και διαχειρίζονται. Μερικοί αλγόριθμοι εξόρυξης δεδομένων λειτουργούν καλά μόνο για σποραδικά δεδομένα.

Ανάλυση: Η ανάλυση δεδομένων ασχολείται με τα διαφορετικά επίπεδα αναλύσεων, με τις ιδιότητες των δεδομένων που είναι σε διαφορετικά επίπεδα ανάλυσης, και με το αν η ανάλυση είναι υψηλή με αποτέλεσμα το υπόδειγμα να μην είναι ορατό ή και να χαθεί από το θόρυβο, ή αν η ανάλυση είναι χονδροειδής που έχει ως αποτέλεσμα την εξαφάνιση του υποδείγματος.

Δεδομένα εγγραφών: Είναι ένα σταθερό σύνολο πεδίων δεδομένων το οποίο είναι μια συλλογή από εγγραφές και για το οποίο δεν υπάρχει μια σαφής σχέση ανάμεσα στις εγγραφές ή τα πεδία δεδομένων και κάθε εγγραφή έχει το ίδιο σύνολο χαρακτηριστικών. Τα δεδομένα εγγραφών αποθηκεύονται συνήθως σε τυποποιημένα αρχεία ή σχεσιακές βάσεις δεδομένων. Μια βάση δεδομένων χρησιμεύει ως ένας βολικός χώρος εύρεσης εγγραφών. Τα δεδομένα εγγραφών χωρίζονται σε τρεις υποκατηγορίες τα δεδομένα συναλλαγών ή καλαθιού αγοράς, τη μήτρα δεδομένων και τη μήτρα σποραδικών δεδομένων. Τα δεδομένα συναλλαγών ή καλαθιού αγοράς είναι ειδικός τύπος δεδομένων εγγραφών, όπου στην κάθε εγγραφή υπάρχει ένα σύνολο στοιχείων τα οποία μπορούν να εξετάζονται και σαν ασύμμετρα χαρακτηριστικά τα οποία συνήθως είναι σε δυαδική μορφή. Η μήτρα Δεδομένων είναι μια παραλλαγή των δεδομένων εγγραφών αλλά λόγο του ότι αποτελείται από αριθμητικά χαρακτηρισικά μπορούν να εφαρμοστούν λειτουργίες πινάκων για το μετασχηματικσμό και τη διαχείριση των δεδομένων. Η μήτρα Σποραδικών Δεδομένων είναι μια ειδική περίπτωση μήτρας δεδομένων στην οποία τα χαρακτηριστικά είναι ίδιου τύπου και ασύμμετρα. Αν παραληφθεί μια σειρά λέξεων σε ένα έγγραφο τότε το έγγραφο μπορεί να αναπαρασταθεί ως ένα διάνυσμα όρων, όπου κάθε όρος είναι ένα στοιχείο του διανύσματος και η τιμή κάθε στοιχείου είναι ο αριθμός των φορών εμφάνισης του όρου μέσα στο έγγραφο (Ning Tan, Steinbach, & Kumar, 2010).

Δεδομένα Γράφων: Υπάρχουν δύο συγκεκριμένες περιπτώσεις γράφων που αποτελεί ένα βολικό και ισχυρό τρόπο αναπαράστασης δεδομένων, ο γράφος καταγραφής της σχέσης μεταξύ των αντικειμένων δεδομένων και τα ίδια τα αντικείμενα δεδομένων να αναπαρίστανται ως γράφοι. Τα δεδομένα γράφων έχει δύο υποκατηγορίες τα δεδομένα με σχέσεις μεταξύ αντικειμένων και τα δεδομένα αντικειμένων που είναι γράφοι. Τα Δεδομένα με Σχέσεις μεταξύ Αντικειμένων συνήθως περιέχουν πληροφορίες που συχνά αναπαρίστανται ως γράφοι. Τα αντικείμενα δεδομένων απεικονίζονται σε κόμβους του γράφου και οι σχέσεις μεταξύ των αντικειμένων δείχνουν τους συνδέσμους μεταξύ τους και τις ιδιότητες των συνδέσμων όπως είναι η κατεύθυνση ή το βάρος. Τα Δεδομένα Αντικειμένων που είναι Γράφοι έχουν μια δομή, τα αντικείμενα περιέχουν υπο-αντικείμενα τα οποία σχετίζονται και αυτού του είδους τα αντικείμενα αναπαρίστανται συνήθως ως γράφοι.

Διατεταγμένα Δεδομένα: Τα διατεταγμένα δεδομένα είναι η διάταξη στο χρόνο ή στο χώρο για ορισμένους τύπους δεδομένων και τα διατεταγμένα δεδομένα έχουν διάφορους τύπους όπως τα ακολουθιακά δεδομένα, τα δεδομένα ακολουθίας, τα δεδομένα χρονικών σειρών, τα χωρικά δεδομένα και τη διαχείριση δεδομένων μη εγγραφών. Τα Ακολουθιακά Δεδομένα μπορούν να θεωρηθούν και ως μια επέκταση των δεδομένων εγγραφών στην οποία κάθε εγγραφή συσχετίζεται με ένα χρόνο. Τα Δεδομένα Ακολουθίας είναι μια ακολουθία ατομικών οντοτήτων π.χ. μια ακολουθία λέξεων ή γραμμάτων. Είναι αρκετά παρόμοια με τα ακολουθιακά δεδομένα αλλά διαφέρουν στο ότι υπάρχουν θέσεις σε μια διατεταγμένη ακολουθία και δεν υπάρχει η έννοια του χρόνου. Τα Δεδομένα Χρονικών Σειρών είναι ένας ειδικός τύπος ακολουθιακών δεδομένων στα οποία κάθε εγγραφή είναι μια χρονική σειρά. Χωρικά Δεδομένα είναι κάποια αντικείμενα που έχουν ιδιαίτερα χαρακτηριστικά όπως οι θέσεις, οι περιοχές και άλλοι τύποι χαρακτηριστικών. Ένα χαρακτηριστικό παράδειγμα χωρικών δεδομένων είναι τα καιρικά δεδομένα (βροχόπτωση, θερμοκρασία). Διαχείριση Δεδομένων μη Εγγραφών είναι οι εγγραφοστρεφείς τεχνικές που εφαρμόζονται σε δεδομένα μη εγγραφών. Οι κυριότεροι τύποι των μεταβλητών διακρίνονται σε ποιοτικά και ποσοτικά δεδομένα.



Εικόνα 3. Τύποι δεδομένων

Ποιοτικά: Όταν τα δεδομένα είναι ποιοτικά δεν μπορούν να πραγματοποιηθούν μαθηματικές πράξεις. Όμως μπορούν να καταμετρηθούν οι συχνότητες εμφάνισης κάθε κατηγορίας. Οι τιμές ενός ονομαστικού χαρακτηριστικού είναι απλός διαφορετικά ονόματα. Οι λειτουργίες που μπορούν να γίνουν για τα ποιοτικά δεδομένα είναι επικρατούσα τιμή, εντροπία, συσχέτιση ενδεχομένων και έλεγχος του x^2. Επιπλέον όταν τα δεδομένα είναι σε διάταξη παρουσιάζουν διάταξη και μπορούν να υπολογίσουν τις ακόλουθες λειτουργίες όπως είναι η διάμεσος, τα εκατοστημόρια, η συσχέτιση κατάταξης και οι έλεγχοι εκτέλεσης και προσίμου.

Ποσοτικά: Στα ποσοτικά δεδομένα εκτός από γραφήματα μπορούν να εφαρμοστούν αριθμητικοί μέθοδοι για την παρουσίαση ενός δείγματος. Οι συνηθέστερη γραφική μέθοδος στις ποσοτικές μεταβλητές είναι το ιστόγραμμα. Οι συνεχής μεταβλητές λαμβάνουν ως τιμές πραγματικούς αριθμούς. Όμως μπορούν να πάρουν και άλλες λειτουργίες όπως είναι ο μέσος, η τυπική απόκλιση και η συσχέτιση. Στην περίπτωση που τα δεδομένα είναι διακριτά μπορεί να υπολογιστή ο γεωμετρικός μέσος και η ποσοστιαία μεταβολή (Ning Tan, Steinbach, & Kumar, 2010).

Κεφάλαιο 2ο Διερευνητική Ανάλυση Δεδομένων

Ο πρωταρχικός σκοπός της διερευνητικής ανάλυσης δεδομένων είναι να τονίσει τις σχετικές ιδιότητες του κάθε χαρακτηριστικού που προέρχεται από ένα σύνολο δεδομένων. Επιπλέον χρησιμοποιεί γραφικές μεθόδους και προσδιορίζει στατιστικά στοιχεία. Η διερευνητική ανάλυση δεδομένων περιλαμβάνει τρεις κύριες φάσεις:

1. Μονοπαραγοντική ανάλυση στην οποία οι ιδιότητες του κάθε μεμονωμένου χαρακτηριστικού είναι ένα σύνολο δεδομένων οι οποίες ερευνούνται.
2. Διμεταβλητή ανάλυση στην οποία οι ιδιότητες θεωρούνται ζεύγη χαρακτηριστικών για την μέτριση της έντασης της σχέσης που υπάρχει μεταξύ τους. Για τα μοντέλα εποπτευόμενης μάθησης είναι ιδιαίτερα σημαντική η ανάλυση των σχέσεων μεταξύ των επεξηγηματικών χαρακτηριστικών και της μεταβλητής στόχου.
3. Ανάλυση πολλαπλών μεταβλητών στην οποία οι σχέσεις κατέχουν κάποιο χαρακτηριστικό εντός του υποσυνόλου που διερευνώνται (Carlo, 2009).

2.1 Γραφική ανάλυση των κατηγορηματικών χαρακτηριστικών

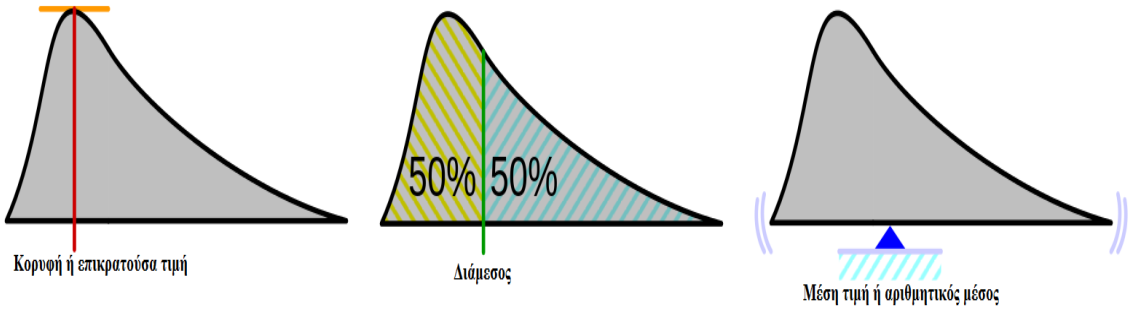
Ένα κατηγορικό χαρακτηριστικό αναλύεται γραφικά σε διάφορες παραστάσεις για εμπειρική κατανομή των παρατηρήσεων. Για τα κατηγορικά δεδομένα ισχύει, κάθε τιμή είναι και ένα τμήμα. Αυτό έχει σαν αποτέλεσμα την ύπαρξη μεγάλου πλήθους τιμών και οι τιμές αυτές συνδιάζονται μεταξύ τους. Ωστόσο για συνεχή χαρακτηριστικά, το εύρος των τιμών χωρίζεται σε τμήματα αλλά όχι αναγκαστικά ίδιου μήκους. Μόλις ολοκληρωθούν οι μετρήσεις κάθε τμήματος, σχεδιάζεται ένα γράφημα ράβδων έτσι ώστε κάθε τμήμα να αναπαρίσταται από μια ράβδο και η περιοχή κάθε ράβδου να είναι ανάλογη του πλήθους των τιμών. Θεωρείται πολύ χρήσιμο για την καταγραφή διακριτών δεδομένων.

Ανάλυση εμπειρικής πυκνότητας

Η σχετική εμπειρική συχνότητα είναι ένα πολύ χρήσιμο εργαλείο για την γραφική ανάλυση τόσο σε κατηγορικά χαρακτηριστικά όσο και σε αριθμητικά και ο τρόπος υπολογισμού της δίνεται από τον μαθηματικό τύπο $f(X_i) = \frac{f_i}{\sum f_i}$. Προκειμένου να επιτευχθεί ο στόχος θα πρέπει να μελετηθούν η ασυμμετρία των καμπύλων

πυκνότητας, η κύρτωση των καμπύλων πυκνότητας και τα εμπειρικά ιστογράμματα πυκνότητας (Vercellis, 2009).

2.2 Γραφική ανάλυση αριθμητικών χαρακτηριστικών



Εικόνα 4. Μέτρα θέσης

Μέσος:

Το πιο γνωστό που χρησιμοποιείται για να περιγραφεί μια αριθμητική ιδιότητα είναι η αριθμητική μέση τιμή του δείγματος η οποία δέχεται ποσοτικά δεδομένα. Προσδιορίζει ένα κεντρικό σημείο γύρω από το οποίο τείνουν να συγκεντρώνονται τα δεδομένα. Επιπλέον χαρακτηριστικό του ότι γίνεται εύκολα κατανοητός και υπολογίζεται σχετικά εύκολα. Αντίθετα, επηρεάζεται πάρα πολύ από ακραίες τιμές και δεν μπορεί να υπολογίσει ποιοτικά δεδομένα. Η μέση τιμή ενός πληθυσμού συμβολίζεται με το μ και η μέση τιμή του δείγματος με $\bar{\mu}$.

$$\bar{\mu} = \frac{\chi_1 + \chi_2 + \cdots + \chi_m}{m} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Διάμεσος:

Η διάμεσος ενός δείγματος πολλών παρατηρήσεων διατεταγμένες σε μη φθίνουσα σειρά μπορεί να χαρακτηριστεί ως η κεντρική τιμή. Επιπλέον η διάμεσος επηρεάζεται από τον αριθμό των στοιχείων που βρίσκονται στην σειρά αλλά όχι με τις ακραίες τιμές.

$$\delta = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

Επικρατούσα Τιμή:

Το μέτρο της κεντρικής τάσης είναι η επικρατούσα τιμή , που ορίζεται ως η τιμή που αντιστοιχεί στην κορυφή της καμπύλης εμπειρικής πυκνότητας, η οποία έχει υπολογιστεί σε ένα χρονικό διάστημα και παρουσιάζεται μέσω γραφικών μεθόδων

κάθε τιμή του διαστήματος της που αντιστοιχίζεται στην μέγιστη εμπειρική συχνότητα.

Μεσαίος Δείκτης:

Ο μεσαίος δείκτης θέσης, ορίζεται ως το μεσαίο σημείο στο διάστημα μεταξύ της ελάχιστης και μέγιστης τιμής.

Γεωμετρικός Μέσος:

Ο γεωμετρικός μέσος πολλών παρατηρήσεων αντιπροσωπεύεται από το γεωμετρικό μέσο που ορίζεται από την μαθηματική ρίζα (Vercellis, 2009).

2.3 Μέτρα της διασποράς για αριθμητικά χαρακτηριστικά

Με τον όρο μέτρα διασποράς αντιπροσωπεύονται τα επίπεδα μεταβλητότητας που εκφράζονται από τις παρατηρήσεις σε σχέση με την κεντρική τιμή του δείγματος. Τα μέτρα διασποράς είναι: Εύρος, Μέση Απόλυτη Διασπορά (Mean Absolute Variance), Διασπορά Συντελεστής Μεταβλητότητας (Coefficient of Variance) (Ning Tan, Steinbach, & Kumar, 2010).

Εύρος τιμών:

Το εύρος τιμών ορίζεται ως η διαφορά της μεγαλύτερης από την μικρότερη παρατήρηση του δείγματος. Είναι πολύ απλό στον υπολογισμό όμως δεν θεωρείται ως αξιόπιστο μέτρο διασποράς διότι βασίζεται μόνο στην μικρότερη και την μεγαλύτερη παρατήριση του δείγματος

$$R = x_{max} - x_{min}$$

Τεταρτημόρια:

Τα τεταρτημόρια είναι γενικευμένη μορφή της διαμέσου και μπορούν να δώσουν την ένδειξη του κέντρου και το σχήμα τη κατανομής.

Q_1 το πρώτο τεταρτημόριο: Το σημείο κάτω από το οποίο βρίσκεται το 25% των διατεταγμένων τιμών του δείγματος $Q_1 = L_i + \frac{\delta_i}{f_i}\left(\frac{n}{4} - F_{i-1}\right)$

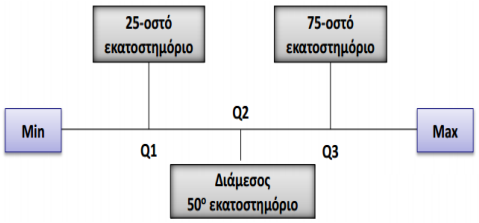
Q_2 το δεύτερο τεταρτημόριο: Συμπίπτει με την διάμεσο

Q_3 το τρίτο τετατρημόριο: Είναι το σημείο πάνω από το οποίο βρίσκεται το 25% των διατεταγμένων τιμών του δείγματος $Q_3 = L_i + \frac{\delta_i}{f_i}\left(\frac{3n}{4} - F_{i-1}\right)$

Εκατοστημόρια:

Τα εκατοστημόρια αποτελούν μια πιο γενική έννοια της διαμέσου. Τα πιο συχνά εκατοστιαία σημεία είναι:

- 25% πρώτο τεταρτημόριο
- 50% δεύτερο τεταρτημόριο ή διάμεσος
- 75% τρίτο τεταρτημόριο



Εικόνα 5. Απεικόνιση Εκατοστημορίου

Θέση εκατοστημορίου p%: $L_p = (n+1)\frac{p}{100}$

Γενικευμένος τύπος περίπτωσης κατανομών συχνοτήτων:

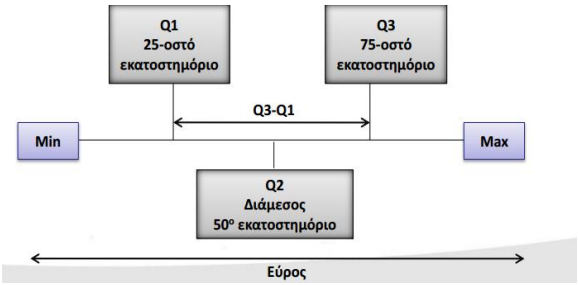
$$X_p = L + \frac{c}{f_i}(pn - F_{i-1})$$

όπου L το κάτω όριο της τάξης που περιέχει το p-οστό εκατοστιαίο σημείο, c το εύρος της τάξης, f_i η συχνότητων, n το πλήθος των δεδομένων και F_{i-1} η αθροιστική συχνότητα της προηγούμενης τάξης.

Ενδοτεταρτημοριακό εύρος:

Τα τεταρτημόρια βοηθούν στον ορισμό ενός νέου δείκτη μεταβλητότητας το οποίο είναι το ενδοτεταρτημοριακό εύρος.

$$IQR = Q_3 - Q_1$$



Εικόνα 6. Απεικόνιση Ενδοτεταρτημορίου

Συντελεστής μεταβλητότητας:

Όταν ένα δείγμα εξεταζόμενο ως προς μια ποσοτική μεταβλητή παρουσιάζει μέση τιμή και τυπική απόκλιση τότε ο συντελεστής μεταβλητότητας ονομάζεται το πηλίκο της τυπικής απόκλισης προς την μέση τιμή επί της 100%. Η μαθηματική έκφραση του συντελεστή μεταβλητότητας δίνεται από τον παρακάτω τύπο:

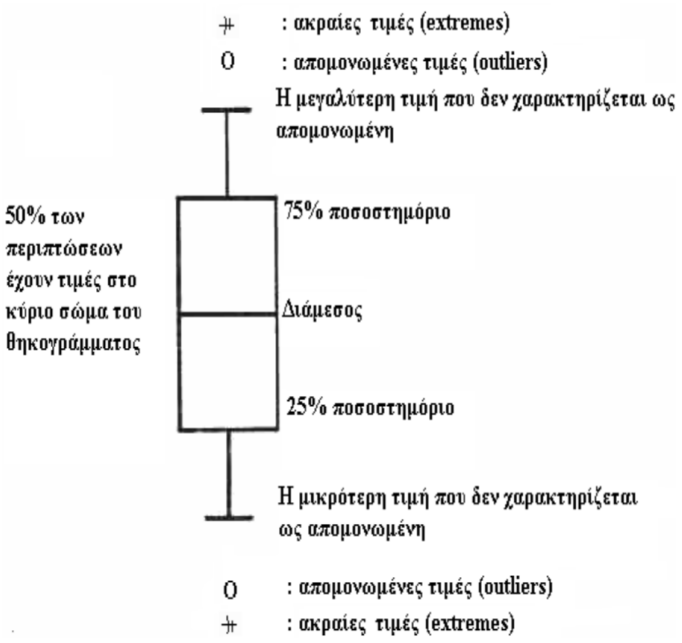
$$CV = \frac{s}{\bar{x}}100\%$$

Ακραίες τιμές:

Οι ακραίες τιμές είναι είτε αντικείμενα δεδομένων τα οποία κατά μια έννοια περιέχουν χαρακτηριστικά τα οποία είναι διαφορετικά από τα περισσότερα από τα άλλα αντικείμενα του συνόλου. Επίσης οι τιμές ενός χαρακτηριστικού οι οποίες είναι ασυνηθίστες σε σχέση με τις τυπικές τιμές του συγκεκριμένου χαρακτηριστικού. Επιπλέον καλούνται και ως ανώμαλες τιμές του συνόλου. Οι ακραίες τιμές διαφέρουν από την έννοια του θορύβου καθώς παρουσιάζουν ενδιαφέρον στην εύρεση αντικειμένων μέσα από μεγάλο αριθμό δεδομένων (Ning Tan, Steinbach, & Kumar, 2010).

Θηκόγραμμα:

Είναι ένας δημοφιλής τρόπος απεικόνισης ο οποίος περιλαμβάνει ένα οριζόντιο ορθογώνιο κουτί του οποίου αρχικά στα άκρα του βρίσκονται τα τεταρτημόρια ενώ όλο το μήκος του πλαισίου είναι το διατεταρτημοριακό φάσμα. Επιπλέον η διάμεσος εντοπίζεται με μια οριζόντια γραμμή μέσα στο κουτί. Επιπρόσθετα έξω από το κουτί εμφανίζονται δύο γραμμές στο πάνω και κάτω του μέρος αντίστοιχα οι οποίες προσδιορίζουν τα όρια των μέγιστων και κατώτερων παρατηρήσεων (Vercellis, 2009).



Εικόνα 7. Θηκόγραμμα

Αν η διάμεσος βρίσκεται στο κέντρο του θηκογράμματος πρόκειται για μια συμμετρική κατανομή, ενώ αν εάν πλησιάζει στο άνω ή στο κάτω άκρο της θήκης, τότε έχουμε αντίστοιχα αρνητική ή θετική ασυμμετρία (Carlo, 2009).

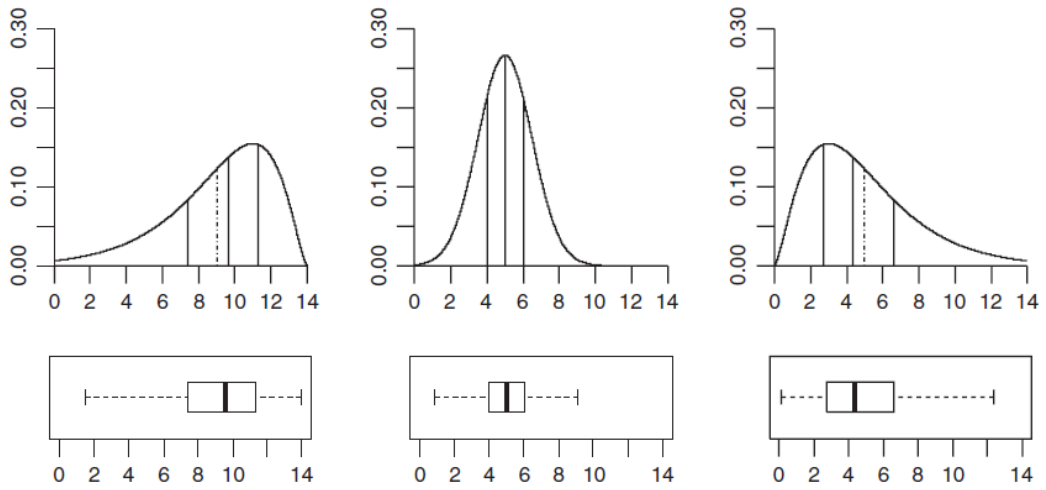
2.3.1 Μέτρα Ασυμμετρίας και Κύρτωσης

Μια κατανομή συχνοτήτων καλείται συμμετρική όταν οι τιμές της τοποθετούνται γύρω από την μέση αριθμητική τιμή. Στην περίπτωση που δεν συμβαίνει, η κατανομή καλείται ασυμμετρική. Οι μαθηματικοί τύποι των μέτρων ασυμμετρίας είναι:

$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, όπου $\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^k$ είναι η κεντρική ροπή τάξης κ, $\beta_1 > 0$ κατανομή με θετική ασυμμετρία, $\beta_1 < 0$ κατανομή με αρνητική ασυμμετρία, $\beta_1 = 0$ συμμετρική κατανομή.

Η κύρτωση μιας κατανομής είναι ο βαθμός συγκέντρωσης των τιμών μιας μεταβλητής γύρω από την μέση αριθμητική του τιμή. Η κύρτωση μετράει την αιχμηρότητα ή την πλάτυνση μιας κατανομής συχνοτήτων. Οι μαθηματικοί τύποι των μέτρων κύρτωσης είναι:

$\beta_2 = \frac{\mu_4}{\mu_2^2}$, όπου $\beta_2 > 3$ κατανομή λεπτόκυρτη, $\beta_1 < 3$ κατανομή πλατύκυρτη, $\beta_3 = 0$ κατανομή μεσόκυρτη (Κιτικίδου).



Εικόνα 8. Εμπειρικές καμπύλες πυκνότητας: Ασύμμετρη αριστερά, Συμμετρική , Ασύμμετρη δεξιά (Carlo, 2009)

2.4 Η μέση απόλυτη απόκλιση

Μέση τυπική απόκλιση:

Το μέτρο αυτό θεωρεί την μεταβλητότητα ως το βαθμό στον οποίο οι τιμές ενός συνόλου δεδομένων αποκλίνουν από το μέσο τους. Ως μέση απόλυτη απόκλιση ορίζεται ο μέσος των απόλυτων τιμών των αποκλίσεων των παρατηρήσεων από το μέσο των παρατηρήσεων αυτών.

Διακύμανση:

Η διακύμανση ορίζεται από κάποιο πληθυσμός N τιμών με μέση τιμή μ την μέση τετραγωνική απόκλιση των n μετρήσεων από τη μέση τιμή μ του πληθυσμού. Η διακύμανση πλεονεκτεί ως μέτρο μεταβλητότητας έναντι της μέσης απόλυτης απόκλισης καθώς θεωρείται πιο εύχρηστη σε μαθηματικές πράξεις. Δηλώνη πόσο μακριά από τη μέση τιμή απέχουν οι παρατηρίσεις. Επιπλέον όταν οι τιμές απέχουν πολύ από την μέση τιμή η διασπορά είναι μεγάλη αντίθετα όταν οι τιμές δεν διαφέρουν πολύ η διασπορά είναι μικρή.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Τυπική απόκλιση:

Ως τυπική απόκλιση ορίζεται η θετική τετραγωνική ρίζα της διακύμανσης

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

2.5 Διμεταβλητή Ανάλυση

Υπάρχουν διάφοροι τύποι των γραφικών αναπαραστάσεων που επιτρέπουν την σχέση μεταξύ δύο χαρακτηριστικών. Για κάθε τύπο γραφήματος θα αναφερθούν οι κατηγορίες των ιδιοτήτων με τις οποίες μπορεί να εφαρμοστεί.

2.5.1 Γραφήματα Διασποράς (Scatter Plots)

Ένα γράφημα διασποράς συχνά αναφέρεται ως διάγραμμα διασποράς και εκφράζει διαισθητικά την γραφική αναπαράσταση μεταξύ δύο αριθμητικών χαρακτηριστικών. Αυτό αναπαριστάται σε ένα Καρτεσιανό διάγραμμα το οποίο λαμβάνεται με την τοποθέτηση του πρώτου γνωρίσματος στον οριζόντιο άξονα και του δεύτερου χαρακτηριστικού στον κάθετο άξονα. Παρέχει την δυνατότητα να σχεδιάζει σημεία χωρίς την προσθήκη γραμμών σύνδεσης (Soukup & Davidson, 2002). Το γράφημα διασποράς είναι πολύ χρήσιμο για στην οπτικοποίηση της εξόρυξης δεδομένων διότι διακρίνονται πολύ εύκολα οι συστάδες των σημείων και των ακραίων τιμών των παρατηρήσεων (Berry & Linoff, 2004) (Carlo, 2009).

2.5.2 Γραφήματα Διασποράς Τριών Διαστάσεων (3D Scatter Plots)

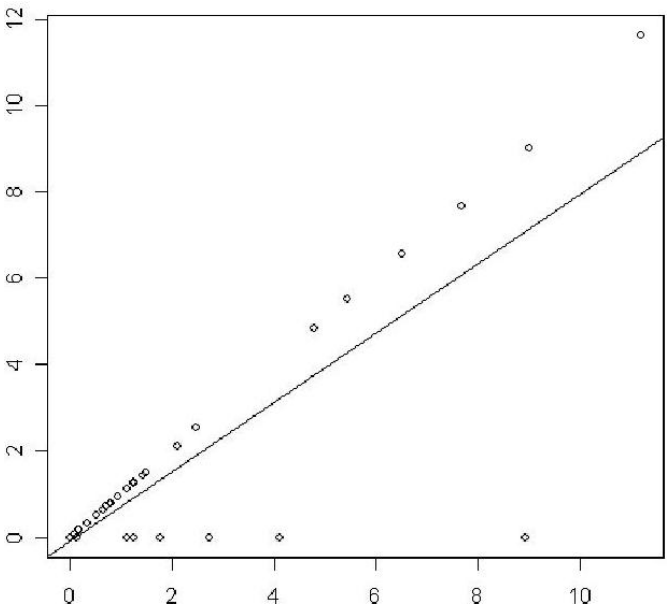
Ένα γράφημα διασποράς τριών διαστάσεων χρησιμοποιείται για να διερευνήσει την σχέση μεταξύ τριών μεταβλητών. Κάθε γραμμή στον πίνακα δεδομένων αναπαρίσταται από έναν δείκτη του οποίου η θέση εξαρτάται από τις τιμές, στις στήλες των τριών αξόνων. Επιπλέον παρέχει την δυνατότητα μιας τέταρτης μεταβλητής η οποία θα μπορεί να ρυθμίζει το μέγεθος των δεικτών και τα χρώματά τους αντίστοιχα. Η σχέση μεταξύ των διάφορων μεταβλητών ονομάζεται συσχέτιση. Στην περίπτωση που οι δείκτες τείνουν να φτιάξουν μια ευθεία γραμμή σε οποιαδήποτε κατεύθυνση του τρισδιάστατου χώρου η συσχέτιση μεταξύ αυτών των μεταβλητών καλείται υψηλή. Σε αντίθετη περίπτωση αν οι δείκτες είναι ισομερώς κατανεμημένοι τότε η συσχέτηση καλείται χαμηλή ή μηδενική. **(http://stn.spotfire.com/spotfire_client_help/3d_scat/3d_scat_what_is_a_3d_scatter_plot.htm)**

Κεφάλαιο 3ο Πολυμεταβλητή Ανάλυση

3.1 Παλινδρόμηση

Ο όρος παλινδρόμηση σημαίνει την δημιουργία μιας συνάρτησης από ανεξάρτητες μεταβλητές (γνωστές και ως προγνωστικοί δείκτες) για την πρόβλεψη μιας εξαρτημένης μεταβλητής (το οποίο είναι η απάντηση). Υπάρχουν διάφορα είδη παλινδρόμησης όπως η γραμμική παλινδρόμηση, γενικευμένη γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση, μη γραμμική παλινδρόμηση. Όταν το μοντέλο είναι σε μορφή μη γραμμικής παλινδρόμησης θεωρούμε συνήθως ότι η μεταβλητή X δείχνει το οριζόντιο άξονα και έχει το ρόλο της ερμηνευτικής μεταβλητής και η μεταβλητή Y δείχνει τον κάθετο άξονα και έχει το ρόλο του μεταβλητού αποτελέσματος (Yanchang, 2012) (Maindonald&Braun, 2003).

Η γραμμική παλινδρόμηση αντιπροσωπεύει την μεγαλύτερη ομάδα των μοντέλων παλινδρόμησης και βασίζεται σε μια κατηγορία υποθέσεων που αποτελείται από γραμμικές συναρτήσεις. Ο μαθηματικός τύπος της γραμμικής είναι ο ακόλουθος $y = \alpha + \beta x$ όπου ο συντελεστής α είναι η τιμή του y και το β είναι η κλίση (slope) της ευθείας και αναπαραστάτε γραφήκα όπως στο ακόλουθο γράφημα. Η αξιοπιστία της γραμμικής παλινδρόμησης δίνεται από την τιμή του συντελεστή προσδιορισμού r^2 και από την τιμή του κριτηρίου F. Δίνεται από την μαθηματική σχέση $r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$.



Η Γενικευμένη Γραμμική Παλινδρόμηση είναι ένα γενικευμένο γραμμικό μοντέλο (GLM) το οποίο περιγράφει το κανονικό σφάλμα γραμμικού μοντέλου παλινδρόμησης, μη γραμμικού, λογιστικά και μοντέλα παλινδρόμησης Poisson και η δομή του συντάσσεται από τα εξής τρία στοιχεία:

1. Το στοιχείο τυχαιότητας (random component) το οποίο καθορίζει την υποθετική κατανομή της μεταβλητής απόκρισης Y_i δοθέντων των εκτιμητών. Οι $Y_{1,\dots,}Y_n$ είναι ανεξάρτητες αποκρίσεις που ακολουθούν μια κατανομή που ανήκει στην εκθετική οικογένεια με αναμενόμενη τιμή $E\{Y_i\}=\mu_i$.
2. Μια γραμμική συνάρτηση των συντελεστών παλινδρόμησης, από την οποία εξαρτάται η μέση τιμή του Y_i, που καλείται γραμμικός εκτιμητής (linear predictor) και βασίζεται στις $\chi_{i1},\dots\dots,x_{i,p-1}$ (μεταβλητές πρόβλεψης). Θα δηλωθεί με $X'_i\beta$ όπου $X'_i\beta=\beta_0+\beta_1X_{i1}+\dots\dots+\beta_{p-1}X_{i,p-1}$.
3. Μια αντιστρέψιμη συνάρτηση σύνδεσης (link function) g η οποία μετασχηματίζει τη μέση τιμή της απόκρισης στην γραμμική εκτίμηση, δηλαδή $X'_i\beta=g(\mu_i)$. Η αντίστροφη συνάρτηση της συνάρτησης σύνδεσης λέγεται και συνάρτηση μέσης τιμής (mean function) και προφανώς θα ισχύει $g^{-1}(X'_i\beta)=\mu_i$.

Η λογιστική παλινδρόμηση χρησιμοποιείται για να προβλέψει ποια είναι η πιθανότητα να εμφανιστεί ένα γεγονός προσαρμόζοντας τα δεδομένα πάνω σε μια λογιστική καμπύλη. Οι τεχνικές λογιστικής παλινδρόμησης μπορεί να ευθύνονται για τις συνδυασμένες επιπτώσεις της αλληλεπίδρασης μεταξύ όλων των μεταβλητών πρόβλεψης βάσει της μη γραμμικής συνάρτησης που ορίζει την εξαρτώμενη μεταβλητή y. Ο μαθηματικός τύπος της λογιστικής παλινδρόμησης είναι ο ακόλουθος: $P(y=0|x)=\frac{1}{1+e^{w'x}}$,

$$P(y=1|x)=\frac{e^{w'x}}{1+e^{w'x}}$$

Παλινδρόμηση Poisson είναι όταν ένα μη γραμμικό μοντέλο παλινδρόμησης όπου τα αποτελέσματα της Υ είναι διακριτά, χρησιμοποιεί μετρήσιμα αποτελέσματα για να βρει λύση και να εξετάσει ποιοι είναι οι παράγοντες και σε πιο μέγεθος επιδρούν για τα αποτελέσματα που βρέθηκαν, σύμφωνα με τον επόμενο μαθηματικό τύπο:

$f(Y)=\frac{\mu^{!}exp(-\mu)}{Y!}$, *όπου Y=0,1,2,3,.. .και το* $f(Y)$*δηλώνει την πιθανότητα το αποτέλεσμα να είναι Y και Y!= Y(Y-1)....3*2*1.*

Μη γραμμική παλινδρόμηση είναι συνδεδεμένη με το πλήθος των ανεξάρτητων μεταβλητών του μοντέλου. Σύμφωνα με τον μαθηματικό τύπο $Y_i = f(X_i, \gamma) + \varepsilon_i$ η μη γραμμική παλινδρόμηση υπάρχει όταν δεν είναι υποχρεωτικό το πλήθος των παραμέτρων παλινδρόμησης που στον τύπο είναι η μεταβλητή γ (Τόλιας, 2008) (Carlo, 2009)(Yanchang, 2012).

3.2 Άμεση Αναλυτική Επεξεργασία (On-Line Analytical Processing – OLAP)

Οι αναλύσεις OLAP βασίζονται στις γνώσεις, την υποβολή εκθέσεων, τη διαίσθηση και τα κριτήρια της οπτικοποίησης των εργαζομένων. Η ροή ανάλυσης της OLAP έχει δομή από πάνω προς τα κάτω η οποία μερικές φορές παίζει σημαντικό ρόλο στην επιτυχία ή την αποτυχία του μοντέλου (Carlo, 2009).

3.2.1 Τι είναι η Άμεση Αναλυτική Επεξεργασία (OLAP) και πως λειτουργεί

Η OLAP (**O**n-**L**ine **A**nalytical **P**rocessing) είναι μια σημαντική βελτίωση σε σχέση με τα συστήματα adhoc ερωτημάτων, επειδή το σύστημα OLAP σχεδιάζει την δομή των δεδομένων σύμφωνα με τις απαιτήσεις των χρηστών του. Τα εργαλεία OLAP παρέχουν πρακτική ανάλυση λειτουργίας που είναι δύσκολο ή αδύνατο να εκφραστούν σε SQL. Τα OLAP εργαλεία έχουν ένα μειονέκτημα το οποίο είναι ότι οι επαγγελματικοί χρήστες αρχίζουν να επικεντρώνονται μόνο στις διαστάσεις των δεδομένων τα οποία αναπαριστώνται από το εργαλείο. Η εξόρυξη δεδομένων, από την άλλη πλευρά, είναι ιδιαίτερα πολύτιμη για τη δημιουργική σκέψη. Στόχος της OLAP είναι να δίνει στον χρήστη τη δυνατότητα να «βλέπει» τα λειτουργικά δεδομένα της επιχείρησης σαν σύνολο χωρίς να μας ενδιαφέρει που καταγράφηκε, σε διαφορετικά επίπεδα ανάλυσης, από διάφορες οπτικές γωνίες, με ανθρωποκεντρική μεθοδολογία χωρίς τεχνικά θέματα.

OLAP είναι μια ισχυρή αναβάθμιση σε παλαιότερες μεθόδους αναφοράς. Η δύναμή του στηρίζεται σε δύο βασικά χαρακτηριστικά:

- Πρώτον, ένα καλά σχεδιασμένο σύστημα OLAP έχει ένα σύνολο σχετικών μερών όπως τη γεωγραφία, το προϊόν, και χρονικά κατανοητά για τους επαγγελματίες χρήστες. Οι διαστάσεις αυτές αποδεικνύονται σημαντικοί για τους σκοπούς της εξόρυξη δεδομένων.

- Δεύτερον, ένα καλά σχεδιασμένο σύστημα OLAP έχει μια σειρά από χρήσιμα μέτρα που σχετίζονται με την επιχείρηση.

Οι τρεις κύριοι τύποι εξυπηρετητών της OLAP είναι:

1) η σχεσιακή OLAP (ROLAP) που τα δεδομένα είναι αποθηκευμένα σε σχεσιακές βάσεις δεδομένων,

2) η πολυδιάστατη OLAP (MOLAP) που τα πολυδιάστατα δεδομένα τους αποθηκεύονται άμεσα σε ειδικές δομές δεδομένων και υλοποιούν τις λειτουργίες της OLAP πάνω σε αυτές τις δομές και οι συναθροίσεις αποθηκεύονται σε ειδική πολυδιάστατη δομή και

3) η υβριδική OLAP (HOLAP) είναι ο συνδιασμός των πολυδιάστατων και των σχεσιακών δομών για να αποθηκευτούν τα πολυδιάστατα δεδομένα δηλαδή επεξεργάζεται τα δεδομένα όπως ROLAP και τις συναθροίσεις όπως η MOLAP.

Οι γρήγοροι χρόνοι απόκρισης είναι σημαντικοί για να πάρει την αποδοχή των χρηστών στα συστήματα αναφοράς. Όταν οι χρήστες πρέπει να περιμένουν, μπορεί να ξεχνούν το ερώτημα που ζήτησαν. Οι μεγαλύτεροι χρόνοι απόκριση, όπως τη βιώνουν οι τελικούς χρήστες, πρέπει να είναι της τάξης των 3-5 δευτερόλεπτα.

Οι δυνατότητες αυτές είναι συμπληρωματικές προς την εξόρυξη δεδομένων, αλλά δεν αποτελεί υποκατάστατο για αυτό. Ωστόσο, OLAP είναι ένα πολύ σημαντικό (ίσως ακόμη και το σημαντικότερο) μέρος των δεδομένων αποθήκης αρχιτεκτονικής, επειδή έχει το μεγαλύτερο αριθμό χρηστών.

Τα γεγονότα των κύβων είναι πολύ ισχυρά. Η χρήση τους είναι περιορισμένη, επειδή γίνονται γρήγορα πολύ μεγάλοι πίνακες των βάσεων δεδομένων που εκπροσωπούν, μπορεί να έχει εκατομμύρια, εκατοντάδες εκατομμύρια, ή ακόμη και δισεκατομμύρια γραμμές. Ακόμη και με τη δύναμη της OLAP και παράλληλων υπολογιστών, όπως κύβους απαιτείται ένα κομμάτι του χρόνου επεξεργασίας για ένα συνηθισμένο ερώτημα. Παρόλα αυτά, τα γεγονότα των κύβων είναι ιδιαίτερα πολύτιμα, επειδή καθιστούν δυνατόν την "περαιτέρω διερεύνηση" από άλλους κύβους ώστε να βρεθεί το ακριβές σύνολο των γεγονότων που χρησιμοποιείται για τον υπολογισμό μιας συγκεκριμένης αξίας. Τα γεγονότα είναι τα μέτρα σε κάθε υπό κύβο. Τα πιο χρήσιμα στοιχεία είναι αθροιστικά, έτσι ώστε να συνδυαστούν μαζί με πολλούς άλλους διαφορετικούς υπό κύβους, ώστε να δοθούν απαντήσεις σε αυθαίρετων επιπέδων ερωτημάτων με μια περιληπτική παρουσίαση της πληροφορίας. Κάθε πρόσθετο

γεγονός καθιστά δυνατή τη σύνοψη δεδομένων κατά μήκος της κάθε διάστασης ή κατά μήκος αρκετών διαστάσεων ταυτόχρονα ο οποίος είναι ο σκοπός του κύβου. Ακόμη και για την υποβολή εκθέσεων adhoc, η πρόσβαση σε δομή κύβου μπορεί να αποδειχθεί πολύ πιο εύκολη από την πρόσβαση σε μια κανονικοποιημένη σχεσιακή βάση δεδομένων (Michael & Gordon, 2004) (Γκίζα, 2007)**.**

3.3 Η αποθήκη δεδομένων και πολυδιάστατη ανάλυση

Αποθήκη δεδομένων είναι ένα πρώτο μέρος για την τοποθέτηση των διαθέσιμων δεδομένων για την δημιουργία επιχειρηματικών πληροφοριών και αρχιτεκτονικής και τη δημιουργία συστημάτων υποστήριξης των αποφάσεων, δηλαδή είναι το σύνολο των δραστηριοτήτων που αλληλοσυνδέονται για τον σχεδιασμό και την υλοποίηση χρησιμοποιόντας μια αποθήκη δεδομένων. Υπάρχουν τρεις (3) βασικές κατηγορίες δεδομένων που γεμίζουν μια αποθήκη δεδομένων: τα εσωτερικά δεδομένα, τα εξωτερικά δεδομένα και τα προσωπικά δεδομένα (Carlo, 2009).

Εσωτερικά δεδομένα: Κατά κύριο λόγο εσωτερικά δεδομένα είναι τα δεδομένα που αποθηκεύονται σε βάσεις δεδομένων που ζητούνται από τα συστήματα συναλλαγών ή τα λειτουργικά συστήματα που είναι ο κύριος κορμός ενός συστήματος επιχειρηματικών πληροφοριών.

Εξωτερικά δεδομένα: Τα εξωτερικά δεδομένα είναι όλες οι πληροφορίες που έρχονται από εξωτερικές πηγές π.χ. ερωτηματολόγια, από γεωγραφικά συστήματα πληροφοριών τα οποία είναι ένα σύνολο από εφαρμογές που παρουσιάζουν τα εδαφικά στοιχεία με αποτέλεσμα να ληφθεί γνώση σε κάθε περιοχή τι προβλήματα υπάρχουν και οι λύσεις τους να μπορούν να απεικονιστούν γραφικά.

Προσωπικά δεδομένα: Προσωπικά δεδομένα είναι οι προσωπικές εκτιμήσεις που υπάρχουν στα φύλλα εργασίας ή σε βάσεις δεδομένων της επιχείρησης και η εισχώρηση τους με σωστά και δομημένα δεδομένα από τις πηγές είναι ένας από τους στόχους των συστημάτων διαχείρισης γνώσης.

Οι αποθήκες δεδομένων και τα marts δεδομένα έχουν βασιστεί σε ένα παράδειγμα για να αντιπροσωπεύσουν στοιχεία που περιέχουν δύο τουλάχιστον σημαντικά πλεονεκτήματα τα οποία είναι: στην λειτουργική πλευρά όπου μπορεί να εγγυηθεί για τους γρήγορους χρόνους ανταπόκρισης ακόμα και σε σύνθετες ερωτήσεις και στη λογική πλευρά όπου οι διαστάσεις ταιριάζουν φυσικά με τα κριτήρια από τους

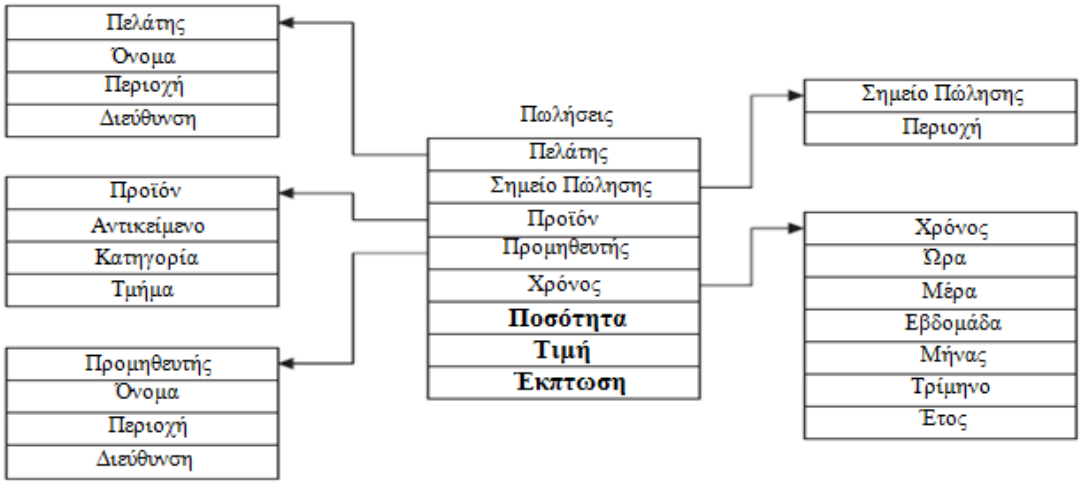
γνώστες στην εργασία για να μπορεί να εκτελεί τις αναλύσεις τους. Αυτή η πολυδιάστατη αντιπροσώπευση βασίζεται στο σχήμα αστέρι το οποίο περιέχει δύο τύπους στοιχείων πινάκων τους πίνακες διάστασης και τους πίνακες γεγονότος.

Πίνακες διάστασης: Γενικά οι διαστάσεις συνδέονται με τις οντότητες γύρω από τις οποίες περιστρέφονται οι διαδικασίες μιας οργάνωσης. Οι πίνακες διάστασης αντιστοιχούν στις αρχικές οντότητες που βρίσκονται σε αποθήκες δεδομένων και στις περισσότερες περιπτώσεις βρίσκονται άμεσα από τους κύριους πίνακες που αποθηκεύονται σε συστήματα OLTP και συνήθως η δομή τους είναι εσωτερική σύμφωνα με τις ιεραχικές σχέσεις.

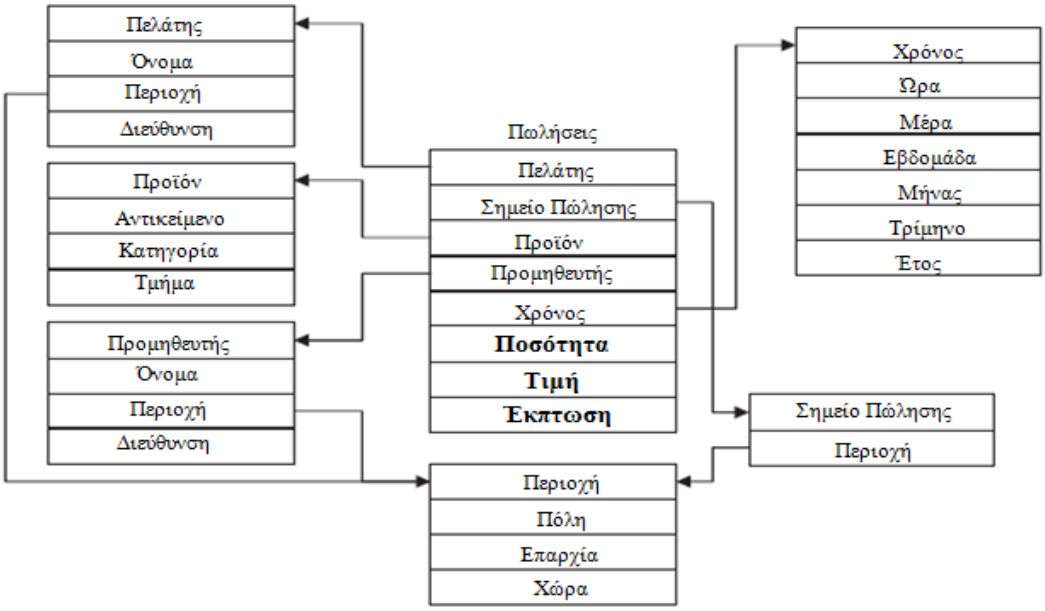
Πίνακες γεγονότος: Οι πίνακες αυτοί αναφέρονται κατά κύριο λόγο σε συναλλαγές και περιέχουν δύο (2) τύπους στοιχείων:

- Σύνδεση με τους πίνακες διάστασης, οι οποίοι είναι απαραίτητοι για να μας οδηγήσουν στις κατάλληλες πληροφορίες που βρίσκονται σε πίνακες γεγονότων,
- Σε αριθμητικές τιμές ιδιοτήτων που χαρακτηρίζουν την αντιστοιχία συναλλαγών που αντιπροσωπεύουν τον πραγματικό στόχο της OLAP ανάλυσης.

Στην συνέχεια παρουσιάζεται το σχήμα αστέρι σε ένα παράδειγμα συναλλαγών πωλήσεων. Ο πίνακας γεγονός βρίσκεται στην μέση του σχήματος και συνδέεται με τους υπόλοιπους πίνακες χρησιμοποιόντας τις κατάλληλες αναφορές και τα μέτρα στον πίνακα γεγονός εμφανίζονται με έντονη γραφή (Carlo, 2009).

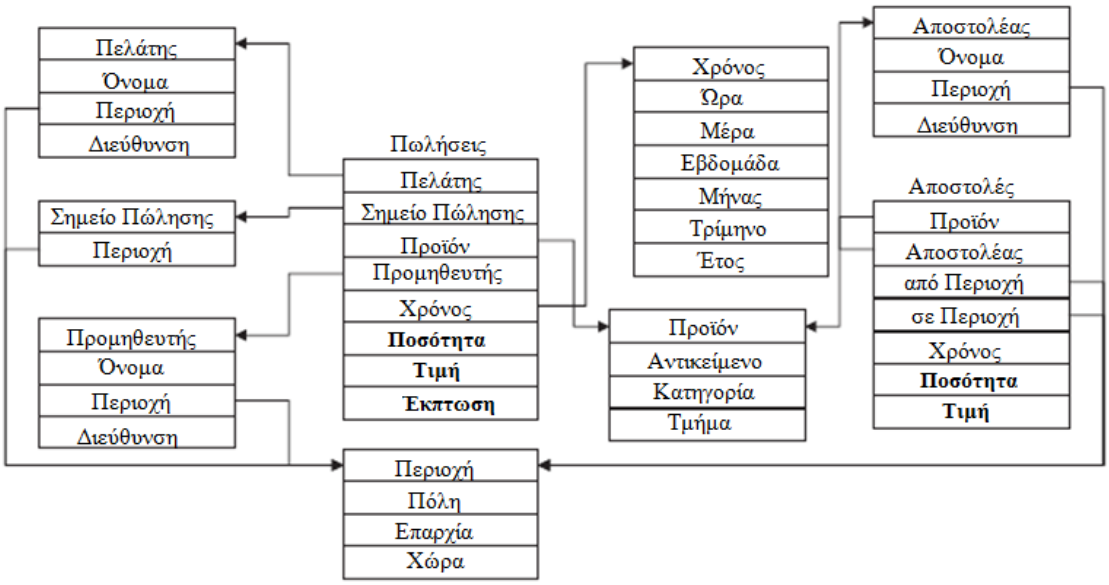


Εικόνα 9. Σχήμα αστέρι



Εικόνα 10. Σχήμα νιφάδας

Άλλη μια τροποποίηση του σχήματος αστεριού είναι η νιφάδα χιονιού που βλέπουμε στο πιο πάνω σχήμα. Σε αυτή την περίπτωση οι πίνακες διάστασης συνδέονται με άλλους πίνακες διάστασης για να δημιουργήσουνμια διαδικασία μερικής τυποποίησης στοιχείων δημιουργώντας έτσι μείωση στη χρήση μνήμηςπου χρειάζεται (Carlo, 2009).

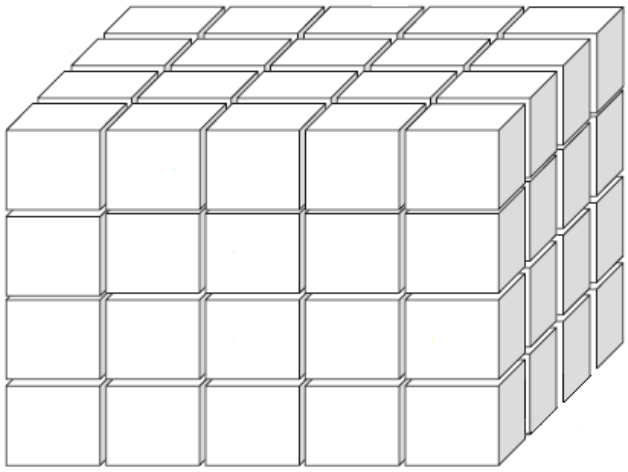


Εικόνα 11. Σχήμα Γαλαξία

Το σχήμα γαλαξία είναι αντιπροσωπεύει πίνακες γεγονότων που συνδέονται με πίνακες διαστάσεων οι οποίοι ενώνονται με άλλους πίνακες διαστάσεων και οι ενώσεις αυτές επιστρέφουν το αποτέλεσμα που ζητήθηκε (Carlo, 2009).

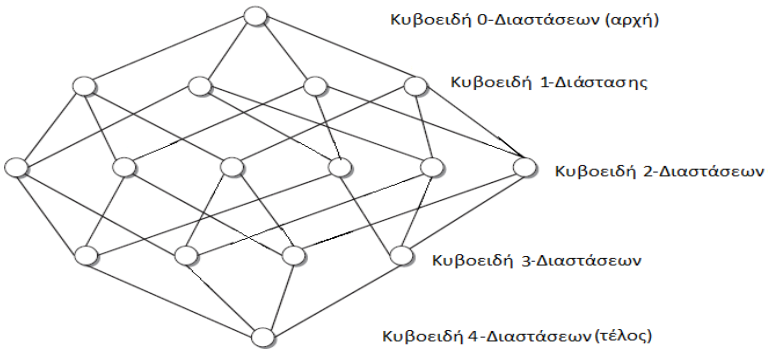
3.3.1 Οπτική αναπαράσταση πολυδιάστατων δεδομένων και ιεράρχιση εννοιών και λειτουργιών της OLAP

Ένας πίνακας γεγονότος με πίνακες πολλών διαστάσεων n μπορεί να εμφανιστεί με την βοήθεια n διαστάσεων δεδομένων κύβου στον οποίο κάθε άξονας αντιπροσωπεύει και μια διάσταση. Ο τριδιάσταστος κύβος είναι η συνέχεια των γνωστών δισδιάστατων λογιστικών φύλλων που μπορούν να εξηγηθούν και ως δισδιάστατοι κύβοι και η μορφή απεικόνισεις του είναι η ακόλουθη (Carlo, 2009).



Εικόνα 12. Τριδιάστατος κύβος

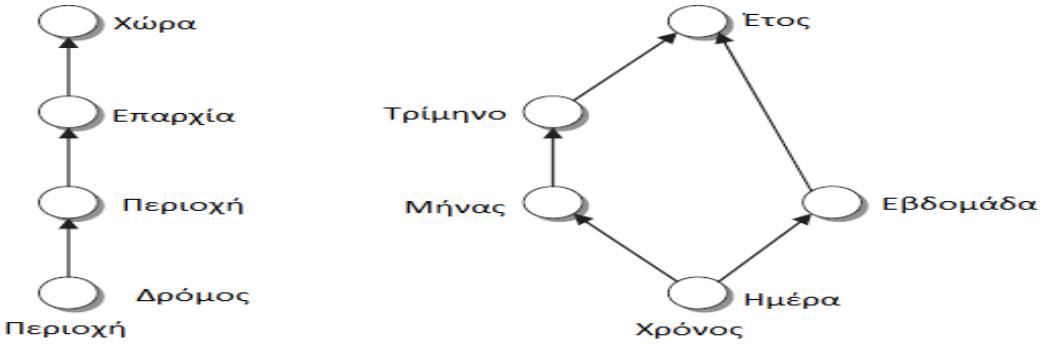
Αντίθετα όταν υπάρχουν τεσσάρων διαστάσεων δεδομένα δεν μπορούμενα το αναπαραστήσουμε με τον τριδιάστατο κύβο αλλά μπορούν να πάρουν τέσσερις λογικές όψεις που προέρχονται από τον τριδιάστατο κύβο, οι οποίες ονομάζονται κυβοειδή, και να τις βάλουμε στον τεσσάρων διαστάσεων κύβο καθορίζοντας πρώτα τις τιμές της μιας διάστασης. Μία απεικόνιση που μπορούμε να δημιουργήσουμε είναι η ακόλουθη (Carlo, 2009).



Εικόνα 13. Τρόπος απεικόνισεις τεσσάρων διαστάσεων και πάνω δεδομένα

Τις περισσότερες φορές οι αναλύσεις OLAP βασίζονται στην ιεράρχιση των εννοιών για να συγκεντρώσουν τα δεδομένα και να δημιουργήσουν λογικές απόψεις έχοντας

τις διαστάσεις της αποθήκης δεδομένων. Ως ιεραχία μπορεί να ορίσει ένα σύνολο χαρτών από ένα χαμηλό επίπεδο εννοιών έως ένα υψηλό επίπεδο. Όπως βλέπουμε (Εικόνα 14. Τρόποι εκτέλεσης ιεράρχισης) μια διατεταγμένη ιεραχία μπορεί να αναπτυχθεί κατά μήκος ή μερικώς διασκορπισμένη. Οι συγκεκριμένες ιεραρχίες δεν χρειάζεται οι αναλυτές να καθορίσουν τις σχέσεις μεταξύ εννοιών όπως είναι απαραίτητο σε άλλες ιεραχίες. Επίσης οι ιεραρχίες μπορούν να εκτελέσουν και διαδικασίες απεικόνισης που ασχολούνται με κύβους δεδομένων μέσα σε μια αποθήκη δεδομένων (Carlo, 2009).



Εικόνα 14. Τρόποι εκτέλεσης ιεράρχισης

Κύλιση προς τα πάνω: Η λειτουργία κύλιση προς τα πάνω (roll-up) ή διείσδυση προς τα πάνω (drill-up) είναι μια ομαδοποίηση δεδομένων στον κύβο τα οποία μπορούν εμφανιστούν και με δύο άλλους τρόπους (Carlo, 2009):

- Ανεβαίνοντας σε υψηλότερο επίπεδο μιας διάστασης που έχει οριστεί στην διαδικασία της ιεράρχισης εννοιών,
- Μειώνοντας απο μια διάσταση, αφαιρώντας δηλαδή μια διάσταση η οποία μας πηγαίνει σε ενοποιημένα μέτρα βάση των αθροισμάτων που γινόνταν σε όλη τη διάρκεια των χρονικών περιόδων στα δεδομένα του κύβου.

Κύλιση προς τα κάτω: Η λειτουργία κύλιση προς τα κάτω (roll-down) ή διείσδυση προς τα κάτω (drill-down) κάνει την αντίθετη λειτουργία με την κύλιση προς τα πάνω, δηλαδή επιτρέπει να δούμε μέσα σε ένα κύβο δεδομένων τις ενοποιημένες και ομαδοποιημένες πληροφορίες για να έχουμε πιο λεπτομερείς πληροφόριση. Υπάρχουν δύο (2) τρόποι να πραγματοποιηθεί η συγκεκριμένη λειτουργια (Carlo, 2009):

- Πηγαίνοντας σε χαμηλότερο επίπεδο κατά μήκος μιας μόνο ιεραχικής διάστασης,
- Προσθέτοντας μια διάσταση.

Τεμαχισμός σε κύβο (Slice and dice) : Με αυτόν τον τρόπο επιλέγετε η αξίας μιας ιδιότητας και καθορίζεται το μήκος μιας διάστασης. Όταν υπάρχει σταθερό μήκος μιας διάστασης επιλέγετε ένας κύβος ο οποίος γίνεται ένας υπο-κύβος μέσα σε ένα υποχώρο επιλέγοντας ταυτόχρονα τις διαστάσεις που επιθυμούμε.

Περιστροφή (pivot): Η λειτρουγία αυτή που ονομάζεται περιστροφή δημιουργεί μια περιστροφή γύρω από τον άξονα του αλλάζοντας κάποιες διαστάσεις για να παρατηρίσουμε μια διαφορετική άποψη των δεδομένων του κύβου (Carlo, 2009).

3.4 Ο κύβος και τα τρία είδη του κύβου

Κύβος είναι ο σχεδιασμός της δομής των δεδομένων σύμφωνα με τις απαιτήσεις των χρηστών η οποία είναι ιδανική για τον τεμαχισμό των δεδομένων σε κύβους. Ο ίδιος κύβος αποθηκεύεται είτε σε μία σχεσιακή βάση δεδομένων, συνήθως χρησιμοποιώντας ένα σχήμα αστεριού, ή σε μια ειδική βάση δεδομένων πολυδιάστατη που βελτιστοποιεί λειτουργίες OLAP.

Για να φτιαχτεί ο κύβος απαιτείται να αναλυθούν τα δεδομένα και οι ανάγκες των τελικών χρηστών, η οποία γίνεται γενικά από ειδικούς που είναι εξοικειωμένοι με τα δεδομένα και το εργαλείο, μέσω μιας διαδικασίας που ονομάζεται τρισδιάστατη μοντελοποίηση. Αν και ο σχεδιασμός και τοποθέτηση ενός συστήματος OLAP απαιτεί μια αρχική επένδυση, το αποτέλεσμα παρέχει κατατοπιστική και γρήγορη πρόσβαση στους τελικούς χρήστες, γενικά πολύ πιο χρήσιμο από ότι τα αποτελέσματα από ένα εργαλείο ερωτήματος γενιάς. Ο χρόνος απόκρισης του κύβου που έχει δημιουργηθεί σχεδόν πάντα χρειάζεται λίγα δευτερόλεπτα, επιτρέποντας έτσι στους χρήστες να εξερευνήσουν γρήγορα τα δεδομένα.

Τα είδη των κύβων είναι:

1) Ο κύβος των συνοπτικών δεδομένων. Το συγκεκριμένο είδος κύβου μαζί με κάποια δεδομένα τα οποία εμφανίζονται σε ένα σχήμα τριών διαστάσεων μας βοηθούν να εμφανίσουν εύκολα και σαφή συμπεράσματα για το ερώτημα το οποίο έχουμε θέσει.
2) Κύβος αντιπροσώπευσης ατομικών γνωρισμάτων. Αυτοί οι κύβοι περιέχουν πιο λεπτομερή δεδομένα που σχετίζονται με τις αλληλεπιδράσεις των πελατών, όπως οι κλήσεις εξυπηρέτησης πελατών, πληρωμές, επιμέρους λογαριασμοί. Οι περιλήψεις είναι κατασκευασμένες για την άθροιση των γεγονότων πέρα από τον

κύβο. Μια τέτοια εκδήλωση κύβου έχει συνήθως μια διάσταση του πελάτη ή κάτι παρόμοιο, όπως έναν λογαριασμό, Web cookies, ή οικιακή χρήση, που δένει το συμβάν πίσω στον πελάτη. Ένας μικρός αριθμός των διαστάσεων, όπως το αναγνωριστικό πελάτη, ημερομηνία και τύπος συμβάντος είναι συχνά επαρκής για τον εντοπισμό του κάθε υπό-κύβου.

3) Ο τρίτος τύπος του κύβου είναι μια παραλλαγή στον κύβο συμβάν. Σκοπός αυτού του είδους του κύβου είναι να αντιπροσωπεύσει τα αποδεικτικά στοιχεία για κάτι που έχει συμβεί. (Michael & Gordon, 2004)

3.5 Ομαδοποίηση - Συσταδοποίηση

Οι μέθοδοι ομαδοποίησης (cluster analysis) είναι τεχνικές πολυμεταβλητής στατιστικής οι οποίες έχουν στόχο στην δημιουργία ομοιογενών ομάδων έτσι ώστε τα στοιχεία (παρατηρήσεις) που βρίσκονται στην ίδια ομάδα να παρουσιάζουν παρόμοια συμπεριφορά από άποψη κατανομής ενώ τα στοιχεία διαφορετικών ομάδων να αντιστοιχούν σε απομακρυσμένες κατανομές (Ζαφειροπούλου, 2007).

3.5.1Διαφορετικοί τύποι συσταδοποίησης.

Η συσταδοποίηση αποτελείται από τους παρακάτω πέντε τύπους (Ning Tan, Steinbach, & Kumar, 2010):

1. Αποκλειστικές (exclusive). Καλούνται οι συσταδοποιήσεις που αποδίδουν κάθε αντικείμενο σε μια μόνο συστάδα.
2. Επικαλυπτόμενη (over lapping) ή Μη αποκλειστικές (non-exclusive). Καλούνται οι συσταδοποιήσεις που επιτρέπουν σε ένα αντικείμενο να ανήκει ταυτόχρονα σε περισσότερες από μια ομάδες.
3. Ασαφής (fuzzy). Στην ασαφή συσταδοποίηση ένα σημείο ανήκει σε κάθε συστάδα με μια στάθμιση ιδιότητας μέλους (κάποιου βάρους) μεταξύ του 0 το οποίο εντελώς δεν ανήκει και του 1 το οποίο ανήκει.
4. Πλήρης (complete). Καλούνται οι συσταδοποιήσεις που αποδίδουν κάθε αντικείμενο σε μια συστάδα (cluster).
5. Μερική (partial). Καλούνται οι συσταδοποιήσεις που δεν αποδίδουν όλα τα αντικείμενα σε συστάδες. Ωστόσο σε μερικές περιπτώσεις ομαδοποιούνται μόνο κάποια από τα δεδομένα καθώς τα υπόλοιπα μπορεί να βρίσκονται εντός των ομάδων και δεν αποτελούν θόρυβο ή απομονωμένες τιμές (outliers).

Διαφορετικοί τύποι συστάδων (Ning Tan, Steinbach, & Kumar, 2010):

1. Καλά διαχωρισμένες:

Μια συστάδα είναι ένα σύνολο από σημεία τέτοια ώστε κάθε σημείο μιας συστάδας είναι κοντινότερο ή πιο όμοιο με όλα τα άλλα σημεία της συστάδας από ότι σε οποιαδήποτε άλλο σημείο που δεν ανήκει στην συστάδα

2. Βασισμένες σε πρότυπο ή κέντρο:

Μια συστάδα είναι ένα σύνολο από αντικείμενα τέτοια ώστε ένα αντικείμενο στην συστάδα είναι κοντινότερο με το κέντρο ή πρότυπο της συστάδας από ότι το κέντρο οποιαδήποτε άλλης συστάδας.

Το κέντρο της ομάδας είναι συχνά:

Α) centroid, ο μέσος όρος των σημείων της συστάδας.

Β) medoid, το πιο αντιπροσωπευτικό σημείο της συστάδα.

3. Βασισμένες σε γράφο:

Α) Όταν τα δεδομένα αναπαρίστανται με τη μορφή γράφου, όπου οι κόμβοι αποτελούν τα αντικείμενα και οι σύνδεσμοι αντιπροσωπεύουν τις συνδέσεις μεταξύ των αντικειμένων, τότε μια συστάδα ορίζεται ως συνδεδεμένη συνιστώσα (connected component).

Β) Συστάδες βασισμένες στην γειτνίαση.

Μια συστάδα είναι ένα σύνολο σημείων τέτοιο ώστε κάθε σημείο να είναι πιο κοντά σε ένα ή περισσότερα σημεία της συστάδας από ό,τι σε οποιοδήποτε άλλο σημείο εκτός συστάδας.

Γ) Κλίκα ονομάζεται το σύνολο των κόμβων σε ένα γράφο οι οποίοι είναι πλήρως συνδεδεμένοι μεταξύ τους.

4. Βασισμένη στην πυκνότητα:

Α) Μια συστάδα είναι μια πυκνή περιοχή από σημεία, η οποία χωρίζεται από άλλες περιοχές μεγάλης πυκνότητας με περιοχές χαμηλής πυκνότητας.

Β) Χρησιμοποιούνται συχνά σε περιπτώσεις συστάδων με μη κανονικό σχήμα ή με αλληλοπλεκόμενα σχήματα ή όταν υπάρχει θόρυβος ή απομονωμένες τιμές.

5. Εννοιολογικές συστάδες:

Σύμφωνα με τον ορισμό της συστάδας ως ένα σύνολο αντικειμένων, μπορούν να μοιράζονται κάποια ιδιότητα.

Έχουν αναπτυχθεί διάφοροι αλγόριθμοι συσταδοποίησης μερικοί από τους οποίους παρουσιάζουν ιδιαίτερο ενδιαφέρον μιας και έχουν ευρεία αποδοχή. Ένας τέτοιος αλγόριθμος είναι και ο K-Means (Ning Tan, Steinbach, & Kumar, 2010), ο οποίος χαρακτηρίζεται για την απλότητα της εφαρμογής. Ο εν λόγω αλγόριθμος χρησιμοποιεί κάποια βήματα ώστε να «τεμαχίζει» τα δεδομένα σε ξεχωριστές, μη-επικαλυπτόμενες συστάδες (clusters).

- Επιλέγει τον αριθμό k των clusters που θα σχηματιστούν.
- Επιλέγει στυγμιότυπα, με τυχαίο τρόπο, ως τα πρώτα κέντρα των clusters.
- Χρησιμοποιεί την γνωστή Ευκλείδια απόσταση, προκειμένου να κατατάξει τα υπόλοιπα στιγμιότυπα στα clusters με τρόπο τέτοιο ώστε η απόστασή τους με τα κέντρα των clusters να είναι η μικρότερη.
- Χρησιμοποιεί τα στιγμιότυπα του κάθε cluster για να υπολογίσει την μέση τιμή αυτών.
- Η μέση τιμή των στιγμιότυπων σε κάθε cluster προσδιορίζει την νέα τιμή του κέντρου του cluster.

Ορισμένα γενικά συμπεράσματα για τον αλγόριθμο K-means:

1. Ο συγκεκριμένος αλγόριθμος εφαρμόζεται μόνο στην περίπτωση αριθμητικών δεδομένων. Στην περίπτωση την οποία υπάρξουν δεδομένα κατηγορικά θα πρέπει να μετατραπούν οι τιμές σε αριθμητικές προκειμένου να χρησιμοποιηθούν.
2. Ο αλγόριθμος επιτρέπει στον χρήστη να καθορίσει τον αριθμό των clusters προκειμένου να οδηγήσει στην βέλτιστη συσταδοποίηση, διότι ο αλγόριθμος δεν έχει την δυνατότητα να καθορίσει τον αριθμό των clusters από μόνος του. Για το λόγο αυτό ενδέχεται η εφαρμογή του αλγορίθμου να γίνει επαναληπτικά για διαφορετικό αριθμό clusters ώστε να σχηματιστεί ένα αποτελεσματικό μοντέλο.
3. Ο εν λόγω αλγόριθμος βρίσκει την βέλτιστη λύση όταν τα clusters που σχηματίζονται έχουν προσεγγιστικά το ίδιο μέγεθος. Στην περίπτωση που ο K-means καταλήγει σε μια λύση η οποία αποτελείται από clusters διαφορετικών μεγεθών τότε ο αλγόριθμος δεν είναι σε θέση να αποτυπώσει την βέλτιστη λύση.
4. Δεν είναι εφικτό να καθοριστούν ποια κατηγορικά δεδομένα είναι σημαντικά για τον σχηματισμό των clusters. Για το λόγο αυτό διάφορα κατηγορικά δεδομένα τα οποία είναι ασυσχέτιστα μεταξύ τους θα οδηγήσει σε μια όχι καλή λύση.

Όλοι ανωτέρω περιορισμοί του αλγορίθμου K-means δεν σταματούν να τον καθιστούν ένα δημοφιλές εργαλείο το οποίο παρέχει αξιόπιστα αποτελέσματα (Ning Tan, Steinbach, & Kumar, 2010).

3.5.2 Ιεραρχική ομαδοποίηση

Στην ιεραρχική ομαδοποίηση, ο αριθμός των ομάδων δεν απαιτείται από πριν. Οι μέθοδοι λειτουργούν ιεραρχικά με την έννοια ότι ξεκινούν χρησιμοποιώντας κάθε παρατήρηση μιας ομάδας και σε κάθε βήμα ενώνουν ομάδες ή παρατηρήσεις που βρίσκονται πιο κοντά. Επιπλέον χρησιμοποιούν έναν πίνακα αποστάσεων. Ο πίνακας αυτός λειτουργεί με έναν αλγόριθμο τεσσάρων βημάτων. Σε κάποιες ειδικές περιπτώσεις όπου ο αλγόριθμος ξεκινάει με πολλές ομάδες, ενώνει κάθε φορά ανά δύο τις παρατηρήσεις μειώνοντας κατά ένα, ώσπου όλες οι παρατηρήσεις να είναι σε μια ομάδα (Ning Tan, Steinbach, & Kumar, 2010).

Τα δενδρογράμματα είναι ένας τρόπος απεικόνισης της ιεραρχικής συσταδοποίησης και δημιουργείται καθώς εκτελεστούν τα παρακάτω τέσσερα βήματα. Αποτελείται από ένα σύνολο εμφωλευμένων συστάδων όπου επιτρέπεται σε κάθε μια να έχει υπο-συστάδες οργανωμένες σαν ένα ιεραρχικό δένδρο, το οποίο καλείται δενδρόγραμμα. Κάθε κόμβος (συστάδα) στο δένδρο είναι μια ένωση των παιδιών τους (υπό-συστάδες) και η ρίζα του δένδρου αποτελεί την συστάδα που περιέχει όλα τα αντικείμενα. Οι κόμβοι – φύλλα του δένδρου δεν έχουν υπό-συστάδες (Ning Tan, Steinbach, & Kumar, 2010).

3.6 Ιεραρχικοί Αλγόριθμοι Συσταδοποίησης

Οι ιεραρχικοί Αλγόριθμοι Συσταδοποίησης διαχωρίζονται σε δύο κατηγορίες σύμφωνα με την μέθοδο παραγωγής συστάδων (Carlo, 2009):

1. Συσσωρευτικοί Ιεραρχικοί Μέθοδοι (Agglomerative Hierarchical Methods).

Ξεκινώντας ο αλγόριθμος θεωρεί κάθε στοιχείο ως μια ξεχωριστή συστάδα. Σε κάθε βήμα του αλγορίθμου συγχωνεύεται το ζεύγος των συστάδων με την μεγαλύτερη ομοιότητα ή το ζεύγος συστάδων με την κοντινότερη απόσταση. Ο αλγόριθμος σταματάει όταν μόνο ένα σύμπλεγμα συμπεριλαμβανομένων όλων των παρατηρήσεων έχει επιτευχθεί. Η διαδικασία μπορεί να αναπαρασταθεί γραφικά με ένα δενδρόγραμμα του οποίο ο ένας άξονας θα αναγράφει την αξία της ελάχιστης

απόστασης που αντιστοιχούν σε κάθε συγχώνευση και στον άλλον άξονα θα αναγράφονται οι παρατηρήσεις.

Ο αλγόριθμος αποτελείται από τα ακόλουθα τέσσερα βήματα:

- Στην φάση της προετοιμασίας κάθε παρατήρηση αποτελεί μια συστάδα. Επομένως η απόσταση μεταξύ των συστάδων καταγράφεται σε έναν πίνακα D ο οποίος περιέχει όλες τις αποστάσεις μεταξύ των ζευγών των παρατηρήσεων.
- Στην συνέχεια υπολογίζεται η ελάχιστη απόσταση μεταξύ των συστάδων και συγχωνεύονται οι δύο συστάδες X και Cf με την ελάχιστη απόσταση dist(X,Cf) η οποία καταγράφεται.
- Η απόσταση μεταξύ της νέας ομαδοποίησης Ce που προέκυψε μετά την συγχώνευση μεταξύ της Ch και της Cf υπολογίζεται.
- Τέλος, εάν όλες οι παρατηρήσεις περιλαμβάνονται σε μια ενιαία συστάδα η διαδικασία σταματά, αλλιώς επαναλαμβάνεται από το βήμα Β.

2. Διαιρετικοί Ιεραρχικοί Μέθοδοι (Divisive Hierarchical Methods).

   Οι διαιρετικοί αλγόριθμοι λειτουργούν αντίθετα από τους συσσωρευτικούς, καθώς βασίζονται σε μία τεχνική από πάνω προς τα κάτω η οποία τοποθετεί αρχικά το σύνολο όλων των παρατηρήσεων σε μια μεγάλη συστάδα. Στην συνέχεια υποδιαιρεί μια συστάδα σε μικρότερα μεγέθη με στόχο να ελαχιστοποιεί τις αποστάσεις μεταξύ των υποομάδων που δημιουργούνται. Η συγκεκριμένη διαδικασία επαναλαμβάνεται έως ότου οι συστάδες να περιέχουν μια ενιαία παρατήρηση ή πληρούνται οι ανάλογες προυποθέσεις για διακοπή.

Συνοψίζοντας οι συγκεκριμένες μεθόδοι παρουσιάζουν ένα πολύ χαρακτηριστηκό πλεονεκτήμα καθώς υοθετούν διαφορετικές μετρικές ομοιότητες και κριτήρια συγχώνευσης συστάδων με αποτέλεσμα να προσδίδουν μεγαλύτερη ευελιξία στον ερευνητή (Ning Tan, Steinbach, & Kumar, 2010).

3.6.1 Κανόνες σύνδεσης

Προκειμένου να εκτιμηθούν οι αποστάσεις μεταξύ δύο συστάδων, οι αλγόριθμοι επιλέγουν ένα από τα κριτήρια σύνδεσης, τα οποία διαφέρουν μεταξύ τους για τον τρόπο με τον οποίο υπολογίζουν τις αποστάσεις. Για το λόγο αυτό κάθε φορά που επιλέγεται διαφορετικό κριτήριο σύνδεσης προκύπτει και διαφορετικό αποτέλεσμα συσταδοποίησης. Οι κανόνες σύνδεσης αποτελούνται από τα ακόλουθα κριτήρια:

1) Κριτήριο κοντινότερου γείτονα (single linkage).

   Το κριτήριο σύνδεσης χρησιμοποιείται από τον συσσωρευτικό αλγόριθμο και βασίζεται κυρίως στην ελάχιστη απόσταση των παρατηρήσεων αλλά και στα δύο πιο όμοια σημεία στις διαφορετικές συστάδες. Αυτό γραφικά αναπαρίσταται με μια ακμή.

2) Κριτήριο πλήρης σύνδεσης (complete linkage).

   Το κριτήριο της πλήρης σύνδεσης χρησιμοποιείται από τον συσσωρευτικό ιεραρχικό αλγόριθμο, ο οποίος εξετάζει τα δύο λιγότερα όμοια και πιο απομακρυσμένα σημεία στις διαφορετικές συστάδες.

3) Κριτήριο μέσης απόστασης (average linkage).

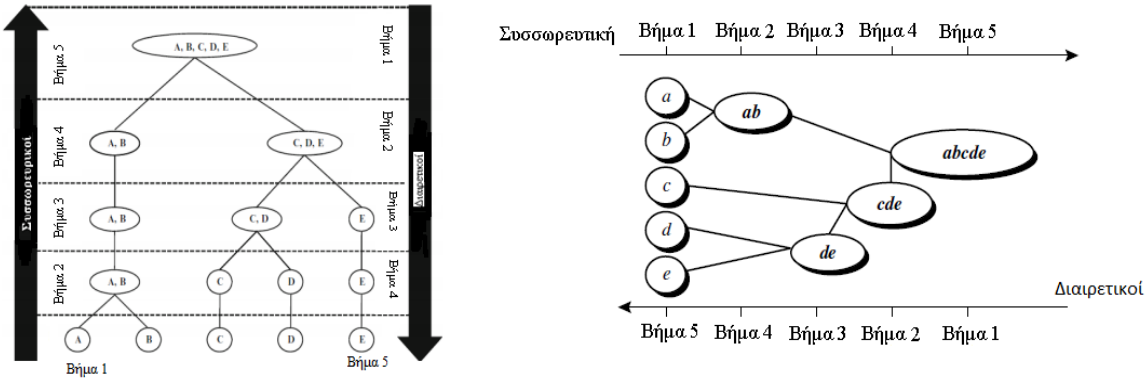
   Α. Η μέση απόσταση εκφράζει την ανομοιότητα μεταξύ δύο συστάδων μέσω του μέσου όρου των αποστάσεων μεταξύ όλων των ζευγαριών των παρατηρήσεων που ανήκουν στις δύο συστάδες.

   Κριτήριο σταθμισμένης μέσης απόστασης.

   Β. Η απόσταση του μέσου όρου των σημείων των συστάδων καθορίζει την ανομοιότητα μεταξύ των δύο συστάδων και χρησιμοποιείται όταν οι κλάσεις έχουν πολύ διαφορετικό μέγεθος.

4) Κριτήριο ελαχίστων διακυμάνσεων Ward.

Σύμφωνα με τον Ward, 1963 στόχος του ήταν να ελαχιστοποιήσει την απώλεια πληροφορίας μετά από κάθε ομαδοποίηση. Για το λόγο αυτό πρότεινε μια διαδικασία η οποία πρώτα θα υπολογίζει το άθροισμα των τετραγώνων των αποστάσεων μεταξύ όλων των ζεύγων των παρατηρήσεων που ανήκουν σε μια συστάδα. Στην συνέχεια, όλα τα ζεύγη των συστάδων που συγχωνεύθηκαν στην τρέχουσα επανάληψη εξετάζονται, και για κάθε ζεύγος υπολογίζεται η συνολική διακύμανση ως το άθροισμα των δύο διακυμάνσεων των αποστάσεων μέσα από κάθε συστάδα, η οποία αξιολογήθηκε στο πρώτο βήμα. Τέλος το ζεύγος των συστάδων που σχετίζονται με την ελάχιστη συνολική διακύμανση συγχωνεύεται. Με αυτόν τον τρόπο παράγεται ένας μεγάλος αριθμός ομάδων που η κάθε μια περιέχει μερικές παρατηρήσεις (Carlo, 2009).



Εικόνα 15. Συσσωρευτική και Διαιρετικοί Ιεραρχικοί Μέθοδοι

3.7 Κατηγοριοποίηση - Ταξινόμηση

Η Κατηγοριοποίηση - Ταξινόμηση (Classification) είναι ίσως η πιο γνωστή και πιο δημοφιλής τεχνική της εξόρυξης δεδομένων της μηχανικής μάθησης και της αναγνώρισης προτύπων. Είναι προγνωστικού τύπου και χρησιμοποιείται για διακριτές μεταβλητές στόχους. Σύμφωνα με μελέτες τα δέντρα απόφασης θεωρούνται από τα πιο δημοφιλή για την αναπαράσταση των κατηγοριοποιητών και ταξινομεί τα δεδομένα με προσέγγιση από πάνω προς τα κάτω (Maimon&Rokach, 2010).

3.7.1 Κατηγοριοποίηση βασισμένη σε δέντρα απόφασης

Ένα δέντρο απόφασης (Decision tree) ή δέντρο κατηγοριοποίησης ακόλουθεί τα παρακάτω βήματα επαγωγής:

1. Αρχικά ξεκινάει με έναν κόμβο που περιέχει όλες τις εγγραφές.
2. Στη δεύτερη φάση διασπά τον κόμβο, δηλαδή διαμοιράζει τις εγγραφές με βάση μιας συνθήκης διαχωρισμού σε κάποιο από τα γνωρίσματα. Μετά από αυτήν την διαδικασία επιλέγεται το καλύτερο γνώρισμα διαχωρισμού.
3. Στην συνέχεια υλοποιείται αναδρομική κλήση του δεύτερου βήματος.
4. Η διαδικασία σταματάει όταν ικανοποιηθεί κάποιο κριτήριο τερματισμού ή μέχρι να φτάσει σε ένα φύλλο το οποίο θα δίνει την προβλεπόμενη έξοδο.
5. Τέλος, εκτελεί κλάδεμα για την βελτίωση της επίδοσής του.

3.7.2 Αλγόριθμοι κατασκευής δέντρου απόφασης.

Από τους πρώτους αλγόριθμους είναι ο Hunto οποίος κατασκεύαζε τα δέντρα απόφασης αναδρομικά καταχωρώντας αρχικά όλες τις εγγραφές σε έναν κόμβο (ρίζα) μέχρι το σύνολο των εγγραφών να εκπαιδευτεί και να καταλήξουν σε ένα φύλλο.

Ο αλγόριθμος ID3 αποτελεί τον θεμέλιο λίθο σχετικά με τα δέντρα αποφάσεων καθώς ακολουθεί μια προσέγγιση από την κορυφή προς τα κάτω (top-down) χωρίς οπισθοδρομήσεις. Επιπλέον ο αλγόριθμος CART βασίζεται σε αριθμητικά δεδομένα καιέχει την ικανότητα να παράγει δέντρα παλινδρόμησης και να αναζητά διασπάσεις οι οποίες θα ελαχιστοποιούν την πρόβλεψη τετραγωνικού σφάλματος.Ο αλγόριθμος C4.5 είναι ο διάδοχος του ID3, καθώς χρησιμοποιεί το κριτήριο της αναλογίας κέρδους (GainRation),το οποίο εξασφαλίζει μεγαλύτερο από το μέσο όρο κέρδος πληροφορίας. Τέλος, διασπά τις πληροφορίες όπου το μέγεθος του υποσυνόλου είναι κοντά στο αρχικό (Maimon & Rokach, 2010) (Berry & Linoff, 2004).

3.7.3 Βήματα Αλγορίθμου ID3

Η ανάλυση των βημάτων του αλγορίθμου ID3 είναι η ακόλουθη (Κουμπάρου, 2010):

1) Επιλογή ενός πεδίου ως ρίζα του δέντρου με βάσει την μικρότερη εντροπία και σχηματισμός διακλάδωσης για κάθε διαφορετική τιμή ή διάστημα του πεδίου αυτού.
2) Χρησιμοποίηση του συνόλου εκπαίδευσης του μέχρι τώρα κατασκευασμένου δέντρο απόφασης. Στην συνέχεια οι εγγραφές της ίδιας τάξης ταξινομούνται σε ένα συγκεκριμένο φύλλο με την ονομασία της τάξης και αν όλα τα φύλλα έχουν ονομασθεί σε κάποια τάξη ο αλγόριθμος τελειώνει.
3) Αντίθετα για τα φύλλα που δεν έχουν ονομασθεί με κάποια τάξη, επιλέγεται ένα πεδίο που δεν έχει επιλεγεί στο μονοπάτι από το φύλλο έως τη ρίζα, βάση της μικρότερης εντροπίας. Ο κόμβος παίρνει την ονομασία αυτού του διαστήματος και σχηματίζεται διακλάδωση μεταξύ του κόμβου και του φύλλου για κάθε διαφορετική τιμή αυτού του πεδίου.
4) Επανάληψη του δεύτερου βήματος.

Το κριτήριο διαχωρισμού των γνωρισμάτων για την επιλογή του καλύτερου κόμβου από τα γνωρίσματα γίνεται χρησιμοποιώντας ένα στατιστικό μέτρο την πληροφορία του κέρδους (information gain) (Αποστόλου, 2008).

Κεφάλαιο 4ο Εφαρμογή Τεχνικών Οπτικοποίησης

4.1 Εισαγωγή

Στην παρούσα βάση δεδομένων εφαρμόζονται μέθοδοι Οπτικοποίησης και Εξόρυξης Δεδομένων για την αποτύπωση της ψυχολογικής κατάστασης φοιτητών ηλικιακής ομάδας 18 έως 26 χρόνων. Για την καταγραφή, αποτύπωση και αξιολόγηση της ψυχολογικής κατάστασης χρησιμοποιήθηκε η σταθμισμένη κλίμακα ψυχοπαθολογίας Symptom Check List-90 (SCL-90), η οποία εξετάζει ένα ευρύ φάσμα ψυχολογικών προβλημάτων και συμπτωμάτων ψυχοπαθολογίας. Το SCL-90 αποτελείται από 90 ερωτήσεις και εντοπίζει 9 κλινικά σημεία (υποκλίμακες) και 3 σύνολα απαντήσεων που περικλείουν τη βαθμολογία των επιμέρους κλιμάκων. Οι υποκλίμακες ορίζονται ως εξής: σωματοποίηση (somatization), ψυχαναγκαστικότητα – καταναγκαστικότητα (obsessive compulsive), διαπροσωπική ευαισθησία (interpersonal sensitivity), κατάθλιψη (depression), άγχος (anxiety), θυμός-επιθετικότητα (anger - hostility), φοβικό άγχος (phobic anxiety), παρανοειδής ιδεασμός (paranoid ideation), ψυχωτισμός (psychotism).

Πιο αναλυτικά,

- Η σωματοποίηση αντανακλά τη δυσφορία (άγχος, ανησυχία) που προέρχονται από την αντίληψη της σωματικής δυσλειτουργίας. Τα συμπτώματα επικεντρώθηκαν στο καρδιαγγειακό, γαστρεντερικό, αναπνευστικό και άλλα συστήματα στα οποία το μέσο του αυτόνομου νευρικού συστήματος είναι πολύ σημαντικό.
- Η ψυχαναγκαστικότητα - καταναγκαστικότητα αντικατοπτρίζουν συμπεριφορές που είναι στενά όμοιες με κλινικό σύνδρομο που περιγράφεται ως ιδεοψυχαναγκαστική διαταραχή. Η πλειοψηφία των στοιχείων της κλίμακας αναφέρεται στις εμμονές ή τις βασανιστικές σκέψεις.
- Η διαπροσωπική ευαισθησία η οποία εστιάζει σε αισθήματα προσωπικής ανεπάρκειας και κατωτερότητας όπως είναι τα συναισθήματα του άγχους και η ανησυχία κατά την διάρκεια διαπροσωπικών σχέσεων. Οι πληροφορίες αυτές για την συμπεριφορά κρίνονται πολύ σημαντικές καθώς εμφανίζουν διάφορες ψυχολογικές καταστάσεις όπως είναι η κατάθλιψη, η σχιζοφρένεια και η κοινωνική φοβία.

- Κλίμακα της κατάθλιψης η οποία αντανακλά μια φαινομενική περιοχή των δεδομένων οι οποίες προκύπτουν από τις κλινικές κατθλιπτικού συνδρόμου. Παρατηρήθηκαν συμπτόματα δυσφορίας συναισθημάτων και διάθεση για τη ζωή καθώς δεν υπήρχαν κίνητρα και ζωτική ενέργεια.
- Κλίμακα του άγχους η οποία περιλαμβάνει μια ποικιλία συμπτωμάτων και εμπειριών που συνδέεται συνήθως με τα υψηλά επίπεδα ανησυχίας.
- Κλίμακα του θυμού - επιθετικότητας η οποία αντικατοπτριζεί τρεις κατηγορίες της επιθετικής συμπεριφοράς: σκέψεις, συναισθήματα και δράσεις.
- Το φοβικό άγχος παρουσιάζει συμπτώματα με υψηλή συχνότητα συνθηκών φοβικών διαταραχών ή αγοροφοβίας. Επιπρόσθετα περιλαμβάνονται διάφορες κλίμακες φοβικής συμπεριφοράς.
- Ο παρανοειδής ιδεασμός προέρχεται από την αντίληψη ότι η παρανοϊκή συμπεριφορά θεωρείται περισσότερο ως ένα σύνδρομο ο οποίος περιλαμβάνει θυμό, βία, καχυποψία και παραισθήσεις.
- Ο ψυχωτισμός ο οποίος περιγράφει μια πλήρη συνέχεια της ψυχωτικής συμπεριφοράς. Τέσσερα στοιχεία απεικονίζουν τη γραμμή συμπτωμάτων (Schneider) για την σχιζοφρένια. Αυτά είναι οι ακουστικές παραισθήσεις, μετάδοση σκέψεων, εξωτερίκευση σκέψεων και εξωτερική παρέμβαση της σκέψης.

4.2 Τι είναι και πως εκτελείται το πρόγραμμα R/RStudio

Η ανάλυση των αποτελεσμάτων έγινε με χρήση του λογισμικού πακέτου R (http://www.r-project.org/) το οποίο προσφέρει ένα ολοκληρωμένο σύνολο υπηρεσιών λογισμικού για ανάλυση δεδομένων, υπολογισμών και γραφικών αναπαραστάσεων και είναι ο διάδοχος της γλώσσα προγραμματισμού S.

Στην συνέχεια ακολουθεί η εφαρμογή της οπτικοποίησης βασισμένη στην βάση δεδομένων excelSCL90 που προέκυψε από επεξεργασία των ερωτήσεων του testSCL-90.

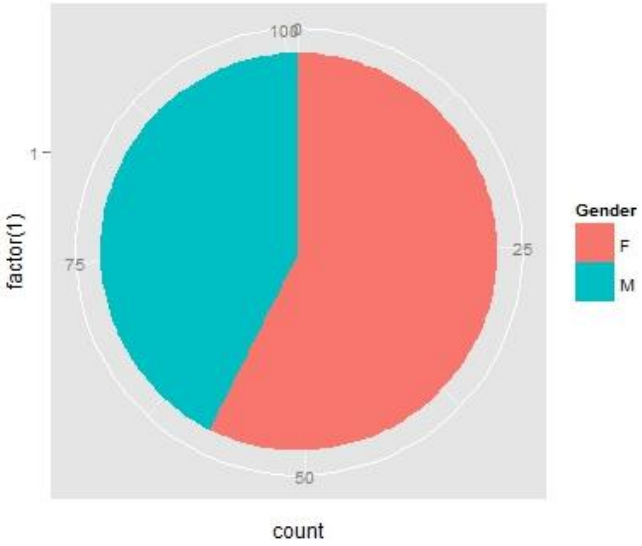
Για την εφαρμογή της οπτικοποίησης χρησιμοποιήθηκαν οι ακόλουθες βιβλιοθήκες:

- ggplot2: Τυπικά χρησιμοποιείται για την κατασκευή ενός οικοπέδου χρησιμοποιόντας τον τελεστή + για την προσθήκη περισσότερων επιπέδων. Αυτό είναι το πλεονέκτημα του κώδικα καθώς γίνονται σαφής τα επιπρόσθετα επίπεδα και η σειρά με την οποία προστίθενται. Συνηστάται για πολύπλοκα γραφικά με πολλά επίπεδα.
- ape: Παρέχει λειτουργίες για την αναγνώριση και τον χειρισμό των φυλογενετικών δέντρων και ακολουθείες σαν το DNA, υπολογίζοντας τις αποστάσεις. Πιο συγκεκριμένα η μέθοδός του βασίζεται στην απόσταση και χρησιμοποιεί μια σειρά από μεθόδους συγκριτικών αναλύσεων και αναλύσεων της διαφοροποίησης.
- plotrix: Το πακέτο plotrix έχει ως στόχο να επιτρέπει στους χρήστες να οπτικοποιήσουν πολλά είδη εξειδικευμένων οικοπέδων γρήγορα, επιτρέποντας τους εύκολη προσαρμογή χώρις μεγάλη εξειδίκευση στην εκμάθηση της σύνταξης του κώδικα.
- scatterplot3d: Το πακέτο scatterplot3d οπτικοποιεί ένα γράφημα διασποράς τριών διαστάσεων για πολυμεταβλητά δεδομένα.
- rgl: Η rgl είναι ένα σύστημα τριών διαστάσεων το οποίο οπτικοποιεί τα πολυμεταβλητά δεδομένα και δίνει την δυνατότητα στον χρήστη να περιστρέψει το γράφημα τρακόσιες εξήντα μοίρες σε ένα καινούργιο παράθυρο το οποίο ανοίγει αυτόματα με την βοήθεια του υποπρογράμματος R Commander.
- Rcmdr: Η Rcmdr είναι μια πλατφόρμα ανεξάρτητη για το R η οποία οπτικοποιεί πολυμεταβλητά δεδομένα και δίνει την δυνατότητα στον χρήστη να περιστρέψει το γράφημα τρακόσιες εξήντα μοίρες σε ένα καινούργιο παράθυρο το οποίο ανοίγει αυτόματα με την βοήθεια του υποπρογράμματος R Commander.
- rpart: Το πακέτο rpart δίνει την δυνατότητα οπτικοποίησης δέντρων απόφασης.
- dendextend: Το πακέτο dendextend οπτικοποιεί δεντδρογράμματα τα οποία έχουν διάφορες ιδιότητες.
- e1071: Είναι η βιβλιοθήκη η οποία εμφανίζει τον αριθμό κύρτωσης και ασυμμετρίας του δείγματος.

- shiny: Είναι εξαιρετικά εύκολη η οικοδόμιση διαδραστικών διαδυκτιακών εφαρμογών για το πρόγραμμα R. Παρουσιάζει μεγάλο ενδιαφέρον η αυτοματοποιημένη δέσμευση μεταξύ των μεταβλητών που εισάγονται και των αποτελεσμάτων χτίζει όμορφες, υπεύθυνες και ισχυρές εφαρμογές με ελάχιστη προσπάθεια.
- shinyApp: Είναι η συνάρτηση η οποία δημιουργεί την εφαρμογή Shiny είτε από ένα ζεύγος UI είτε μεταβιβάζει την διαδρομή του καταλόγου που περιέχει η εφαρμογή Shiny.
- shinydashboard: Είναι η βιβλιοθήκη η οποία δημιουργεί το πάνε πάνω στο οποίο θα οπτικοποιούνται τα δεδομένα. Επιπλέον στο πλάι του πάνελ καταχωρούνται οι λίστες εντολών και των δεδομένων.
- shinyfiles: Η συγκεκριμένη βιβλιοθήκη έχει την ιδιότητα να δημιουργεί ένα κουμπί το οποίο έχει τη λειτουργία εμφάνισης ή απόκρυψης της πλαϊνής μαύρης λίστας.
- rpivotTable: Χρησιμοποιεί συγκεντρωτικούς πίνακες στην R με τη δυνατότητα μίας html εφαρμογής. Για να εγκατασταθεί απαιτείται εγκατάσταση της παραμέτρου devtools με την ακόλουθη γραμμή εντολής devtools::install_github(c("ramnathv/htmlwidgets","smartinsightsfromdata/rpivotTable")) .
- dygraphs: Διασυνδέει γραφήματα σχεδίασης χρονοσειρών στην R μέσο της βιβλιοθήκης JavaScript.

4.3 Μονομεταβλητή Ανάλυση Δεδομένων

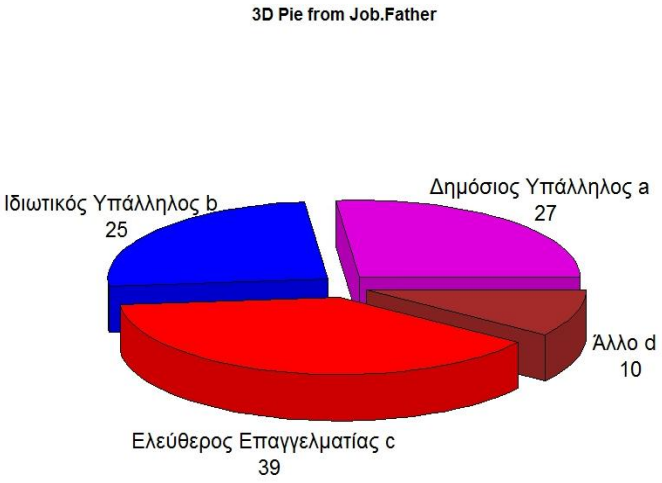
Πίτες:



Στην εικόνα απεικονίζεται ένα μονοδιάστατο κυκλικό διάγραμμα το οποίο εμφανίζει την συχνότητατης μεταβλητής φύλο από τα δεδομένα και δείχνει πως έχουν κατανεμηθεί.

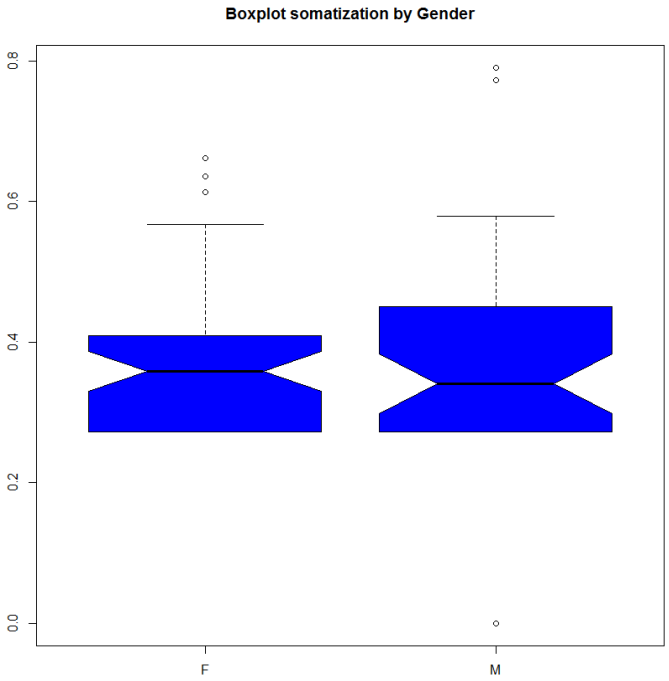
```
Κώδικας: library(ggplot2)
ds_gender<- cbind( 'Gender'=df.SCL90$Gender, ds )
pie<- ggplot(ds_gender, aes(x=factor(1), fill = Gender)) + geom_bar(width = 1)
pie + coord_polar(theta = "y")
```

Στηνεικόνα απεικονίζεται ένα τρισδιάστατο κυκλικό διάγραμμα το οποίο εμφανίζει την συχνότητα των μεταβλητών σύμφωνα με το επάγγελμα του πατέρακαι οι μεταβλητές είναι ποιοτικές. Απαιτείται εγκατάσταση του πακέτου plotrix.




```
Κώδικας : library (plotrix)
counts<- table(df.SCL90$Job.Father)
abcd <-c("Δημόσιος Υπάλληλος", "Ιδιωτικός Υπάλληλος", "Ελεύθερος Επαγγελματίας", "Άλλο")
lbls<- paste(abcd,names(counts), "\n", counts)
pie3D(counts, labels = lbls, explode=0.1, main="3D Pie from Job.Father\n ", col=c("#dd00dd","blue","red","brown"))
```

Θηκόγραμμα:



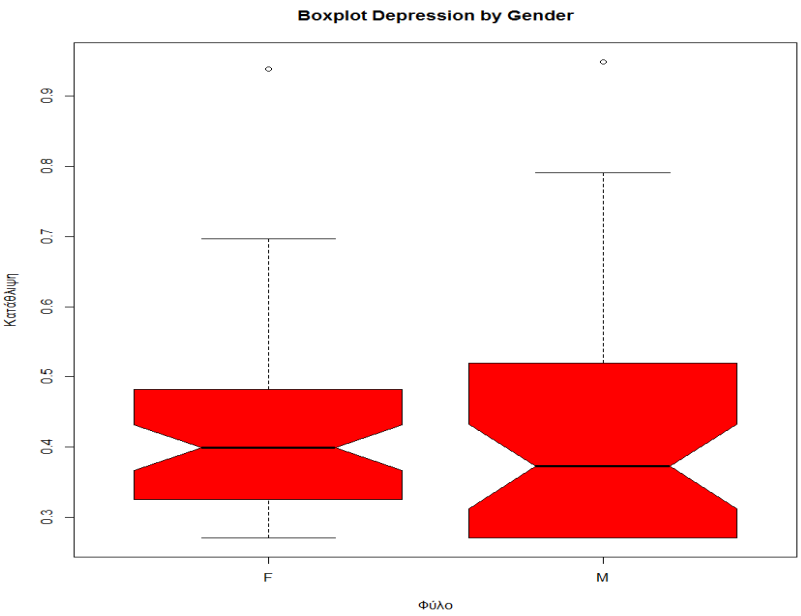
Στην εικόνα αυτή παρουσιάζονται δύο διαγνωστικά γραφήματα παίρνοντας ως δεδομένα από την βάση τα δύο φύλα, δηλαδή γυναίκες και άντρες αντίστοιχα και αναπαριστάται η αρνητική ασυμμετρία για τις γυναίκες σύμφωνα με την μεταβλητή σωματοποίησης ενώ για τους άνδρες φαίνεται η θετική ασυμμετρία σε σχέση με την σωματοποίηση. Επιπλέον οι κάθετες γραμμές δείχνουν τις μέγιστες τιμές που δεν χαρακτηρίζονται ως απομονωμένες σε αντίθεση με τα κλυκλικά σημεία που καλούνται απομονωμένες τιμές. συμμετρικότητα

Κώδικας:ds<- f.SCL90[,12:20]

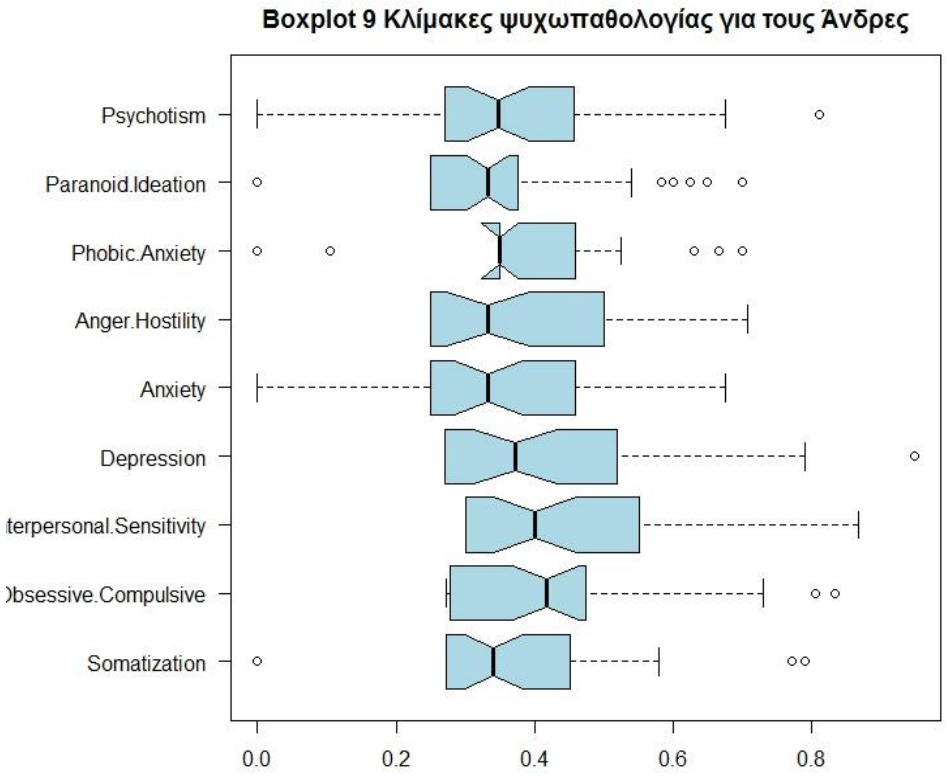
boxplot(x=ds, horizontal = FALSE, notch = TRUE)

box_som_gender<- boxplot(Somatization~Gender, data = df.SCL90, horizontal = FALSE, notch = TRUE, col = "blue", main="Boxplot Somatization by Gender")

Στο συγκεκριμένο γράφημα αναπαριστάται ποια είναι η συμμετρικότητα των δύο φύλων, σε σχέση με την κατάθλιψη.Για τις γυναίκες υπάρχει συμμετρική κατανομή ενώ για τους άνδρες υπάρχει θετική ασυμμετρία.



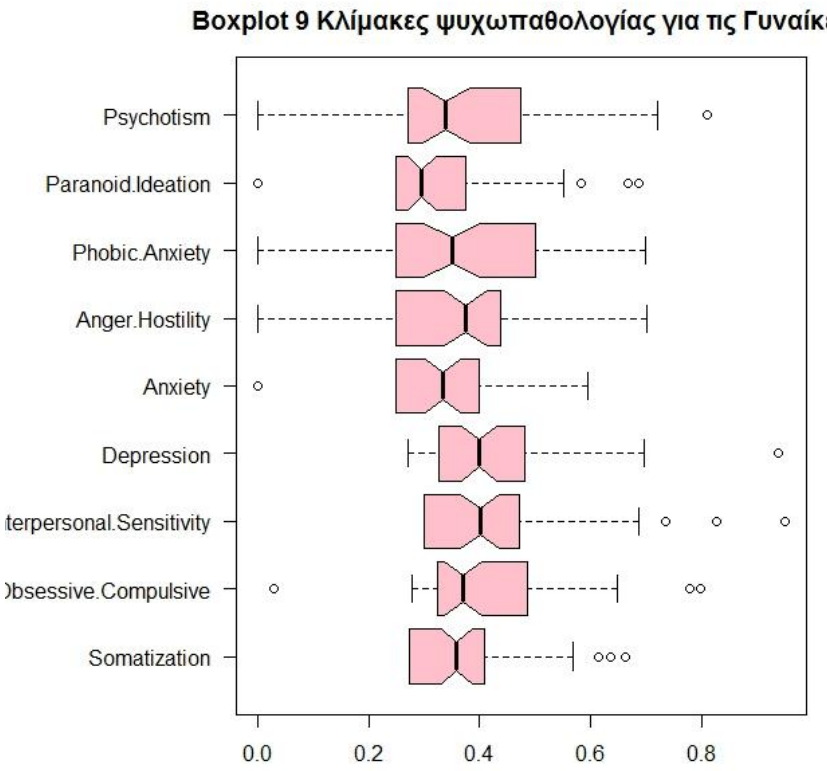
Κώδικας: thikogrammabox_som_gender<- boxplot (Depression~Gender, data = df.SCL90, horizontal = FALSE, notch = TRUE, col = "red", xlab="Φύλο", ylab="Κατάθλιψη", main= "Boxplot Depression by Gender")



Η παραπάνω εικόνα οπτικοποιεί ένα θηκόγραμμα εννέα μεταβλητών με εξαρτημένη μεταβλητή τους άνδρες. Οι μεταβλητές ψυχωτισμός, παρανοειδής ιδεασμός, φοβικό άγχος, κατάθλιψη, ψυχαναγκαστικότητα – καταναγκαστικότητα και η σωματοποίηση παρουσιάζουν απομονωμένες τιμές.

Κώδικας: ds.M <- df.SCL90[df.SCL90$Gender=='M', 12:20]

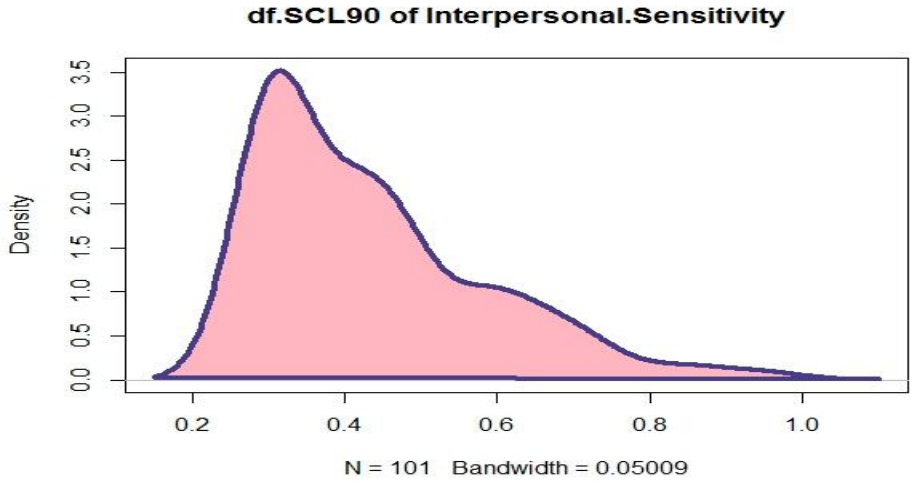
```
par(mar=c(3,9,3,7))
boxplot(x=ds.M, horizontal = TRUE, notch = TRUE, las=1,
main="Boxplot 9 Κλίμακες ψυχωπαθολογίας για τους Άνδρες", col = "light blue")
```



Η παραπάνω εικόνα οπτικοποιεί ένα θηκόγραμμα εννέα μεταβλητών με εξαρτημένη μεταβλητή τις γυναίκες. Οι μεταβλητές ψυχωτισμός, παρανοειδής ιδεασμός, άγχος, κατάθλιψη, διαπροσωπική ευαισθησία, ψυχαναγκαστικότητα – καταναγκαστικότητα και η σωματοποίηση παρουσιάζουν απομονωμένες τιμές.

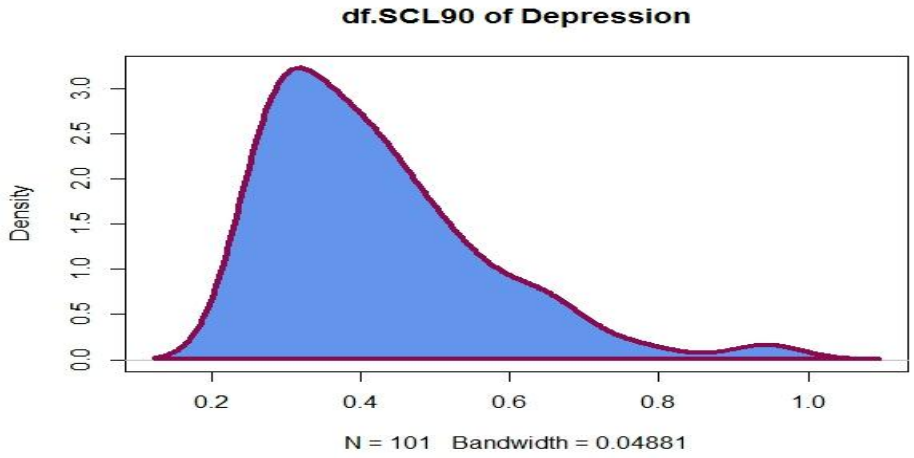
Κώδικας: ds.F<- df.SCL90[df.SCL90$Gender=='F', 12:20]

```
par(mar=c(3,9,3,7))
boxplot(x=ds.F, horizontal = TRUE, notch = TRUE, las=1,
main="Boxplot 9 Κλίμακες ψυχωπαθολογίας για τις Γυναίκες",col = "pink")
```



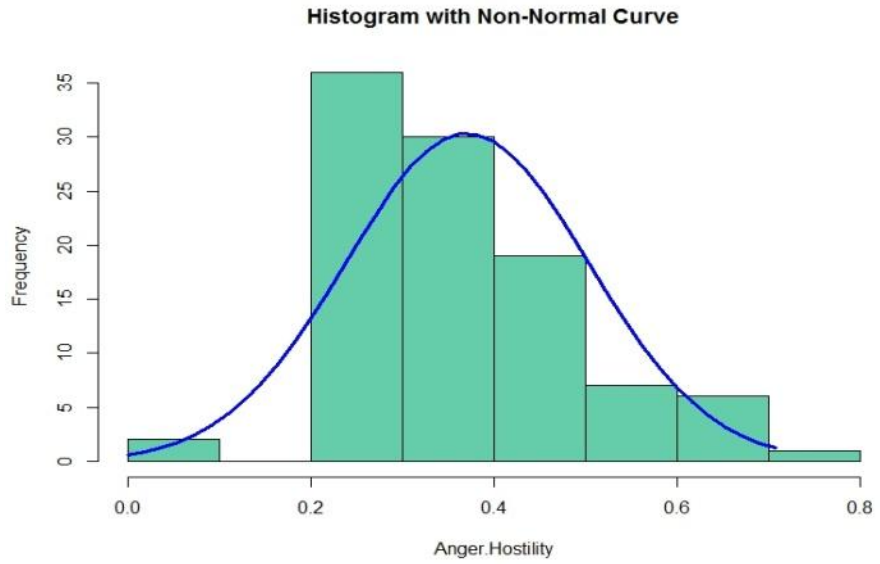
Η εικόνα παρουσιάζει ένα γράφημα ασύμμετρης θετικής κατανομής σύμφωνα με την μεταβλητή διαπροσωπική ευαισθησία. Φαίνεται το εύρος εκτίμησης πυκνότητας για το πλήθος των δεδομένων το οποίο ανέρχεται στο σύνολο των εκατών ένα δεδομένων το οποίο είναι 0,05009.

```
Κώδικας: d <- density(df.SCL90$Interpersonal.Sensitivity)
plot(d, main="df.SCL90 of Interpersonal.Sensitivity")
polygon(d, col="lightpink", border="darkslateblue",lwd=4)
```



Η εικόνα παρουσιάζει ένα γράφημα ασύμμετρης θετικής κατανομής σύμφωνα με την μεταβλητή κατάθλιψη. Φαίνεται πως η διάμεση τιμή είναι μεγαλύτερη από το μέσο όρο της καμπύλης και έτσι ονομάζεται ασύμμετρη προς τα αριστερά. Εμαφανίζεται το εύρος εκτίμησης πυκνότητας για το πλήθος των δεδομένων το οποίο ανέρχεται στο σύνολο των εκατών ένα δεδομένων το οποίο είναι 0,04881.

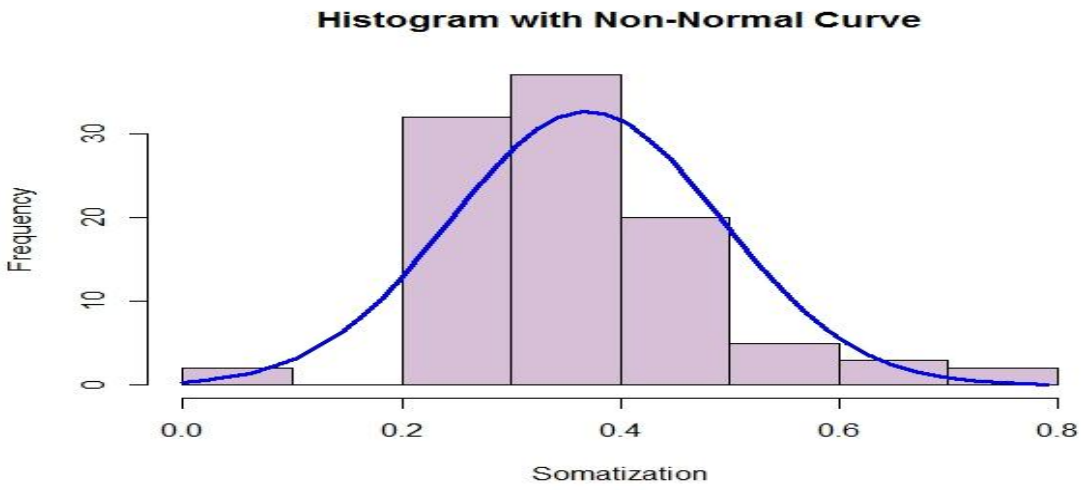
```
Κώδικας: d <- density(df.SCL90$Depression)
plot(d, main="df.SCL90 of Depression")
polygon(d, col="cornflowerblue", border="deeppink4",lwd=4)
```



Το ακόλουθο γράφημα παρουσιάζει ένα ιστόγραμμα μη κανονικής κατανομή σύμφωνα με την μεταβλητή φοβικού άγχους. Δηλαδή είναι μία εμπειρική πυκνότητα και η μπλέ γραμμή προσεγκίζει την απόκλιση από την κανονική πυκνότητα. Επιπλέον το μέτρο της ασυμμετρίας εμφάνισε τον αριθμό 0.3839127 ο οποίος είναι μεγαλύτερος του μηδέν και η κατανομή καλείται θετική ασύμμετρη. Η κύρτωση εμφάνισε τον αριθμό 0.4725694 ο οποίος είναι μικρότερος του τρία και η κατανομή ονομάζεται πλατύκυρτη.

Κώδικας:

```
x <- df.SCL90$Anger.Hostility
h<-hist(x, breaks=10, col="red", xlab="Anger.Hostility",
main="Histogram with Non-Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit<- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=3)
library(e1071)
skewness(x)
kurtosis(x)
```

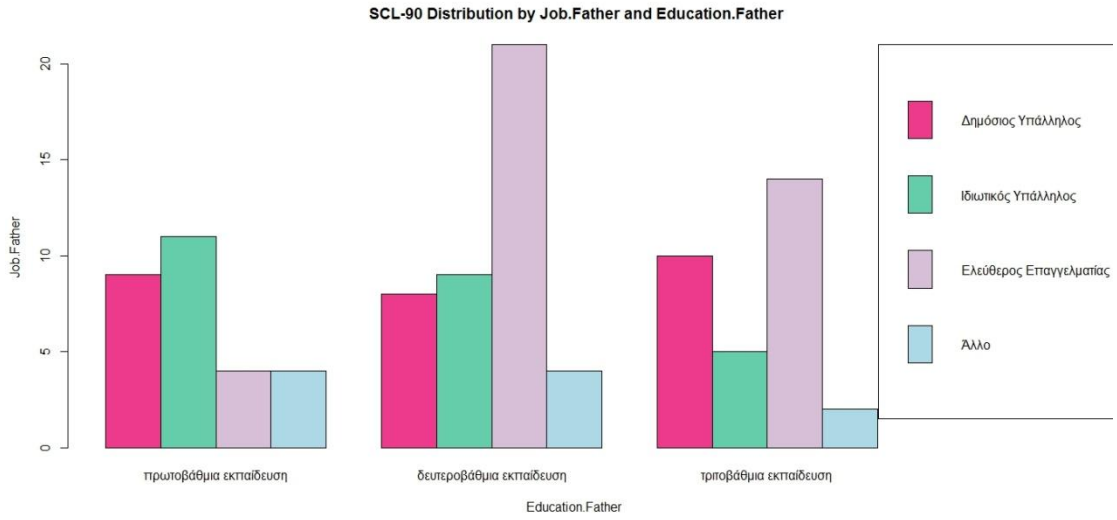


Το ακόλουθο γράφημα παρουσιάζει ένα ιστόγραμμα μη κανονικής κατανομή σύμφωνα με την μεταβλητή σωματοποίηση. Δηλαδή είναι μία εμπειρική πυκνότητα και η μπλέ γραμμή προσεγκίζει την απόκλιση από την κανονική πυκνότητα. Επιπλέον το μέτρο της ασυμμετρίας εμφάνισε τον αριθμό 0.6600879 ο οποίος είναι μεγαλύτερος του μηδέν και η κατανομή καλείται θετική ασύμμετρη. Η κύρτωση εμφάνισε τον αριθμό 2.241598 ο οποίος είναι μικρότερος του τρία και η κατανομή ονομάζεται πλατύκυρτη.

Κώδικας:

```
x <- df.SCL90$Somatization
h<-hist(x, breaks=10, col="yellow", xlab="Somatization",
main="Histogram with Non-Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit<- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=3)
library(e1071)
skewness(x)
kurtosis(x)
```

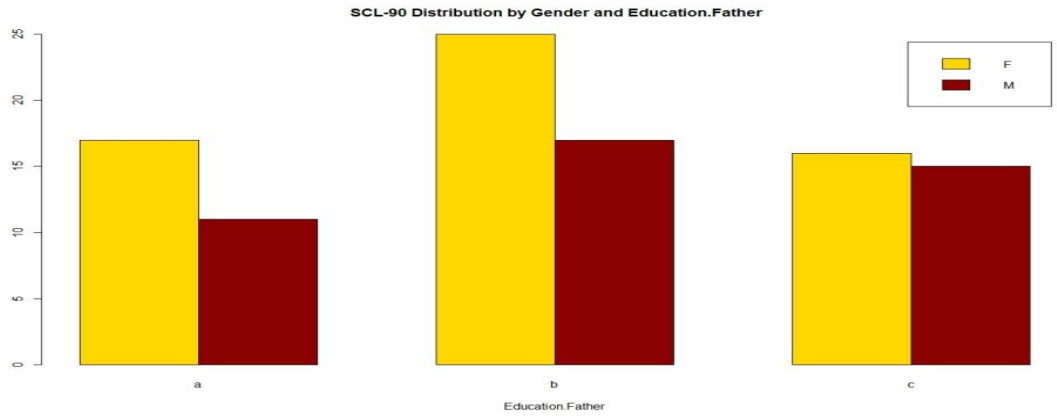
4.4 Διμεταβλητή Ανάλυση Δεδομένων



Η εικόνα παρουσιάζει ένα ραβδόγραμμα συχνοτήτων. Στον άξονα των Υ είναι η εξαρτημένη μεταβλητή επάγγελμα πατέρα η οποία χωρίζεται σε δημόσιο υπάλληλο, ιδιωτικό υπάλληλο, ελεύθερο επαγγελματία και άλλο. Στον άξονα των Χ είναι η ανεξάρτητη μεταβλητή εκπαίδευση πατέρα η οποία χωρίζεται σε πρωτοβάθμια, δευτεροβάθμια και τριτοβάθμια εκπαίδευση.

Κώδικας: counts<- table(df.SCL90$Job.Father, df.SCL90$Education.Father)

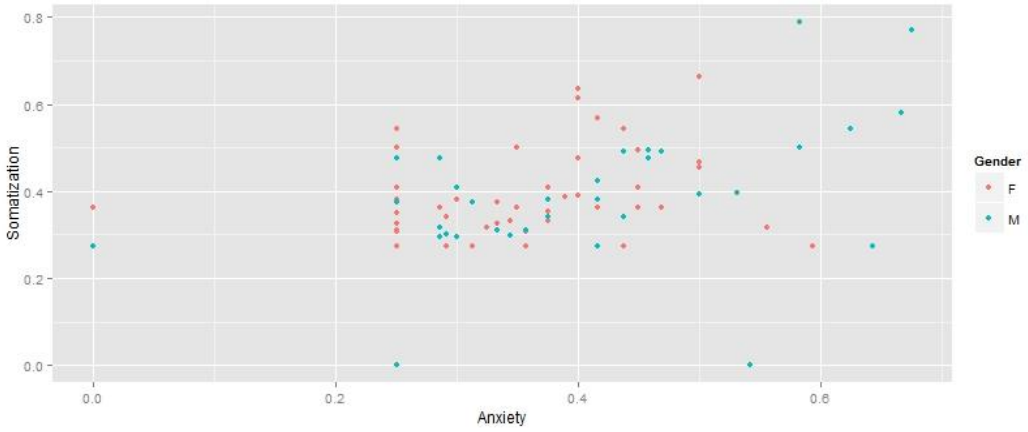
```
barplot(counts, main=" SCL-90 Distribution by Job.Father and Education.Father",
    xlim = c(1,18),
    xlab="Education.Father",
    ylab="Job.Father",
    col=c("violetred2","aquamarine3","thistle", "lightblue"),
    legend =c("Δημόσιος Υπάλληλος", "Ιδιωτικός Υπάλληλος", "Ελεύθερος Επαγγελματίας", "Άλλο"),
    args.legend = list(x ="topleft",inset=c(0.8,0)),
    names.arg = c("πρωτοβάθμια εκπαίδευση", "δευτεροβάθμια εκπαίδευση", "τριτοβάθμια εκπαίδευση"),
axisnames=TRUE, beside=TRUE )
```



Η εικόνα παρουσιάζει ένα ραβδόγραμμα συχνοτήτων κατά πόσο επηρεάζονται τα δύο φύλα σχετικά με την εκπαίδευση του πάτερα.

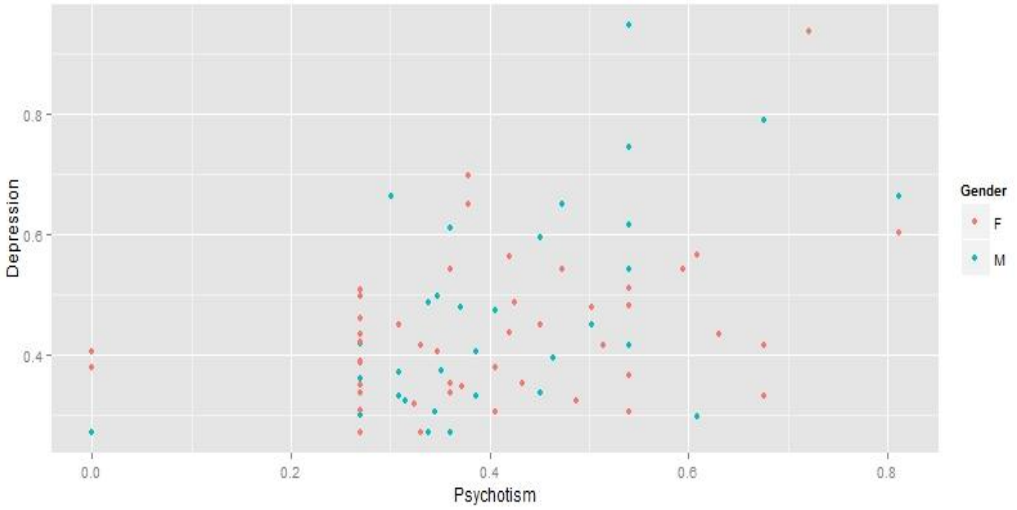
Κώδικας:

```
counts<- table(df.SCL90$Gender, df.SCL90$Education.Father)
barplot(counts, main=" SCL-90 Distribution by Gender and Education.Father",
xlab=" Education.Father ", col=c("gold","darkred"),
legend = rownames(counts), beside=TRUE)
```



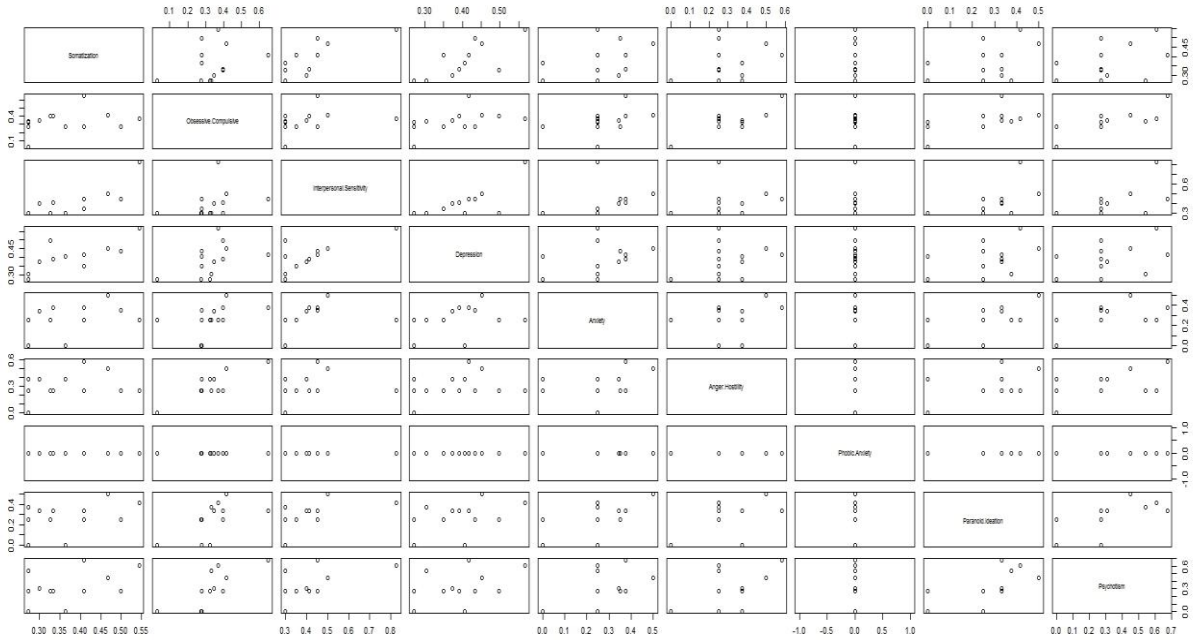
Το γράφημα παρουσιάζει τη διασπορά (scatterplot) του άγχους σε σχέση με την σωματοποίηση ανά φύλο..Για την συγκεκριμένη διεργασία εγκαταστήθηκε το πακέτο ggplot2.

```
Κώδικας:library(ggplot2)
ds_gender<-cbind( 'Gender' = df.SCL90$Gender, ds )
ggplot(data=ds_gender, aes(x=Anxiety, y=Somatization, color=Gender)) + geom_point()
```



Το γράφημα παρουσιάζει τη διασπορά (scatterplot) του φύλου (άντρες-γυναίκες) σε σχέση με την κατάθλιψη.

```
Κώδικας:library(ggplot2)
ds_gender<- cbind( 'Gender'=df.SCL90$Gender, ds )
ggplot(data=ds_gender,aes(x=Psychotism, y=Depression, color=Gender)) + geom_point()
```

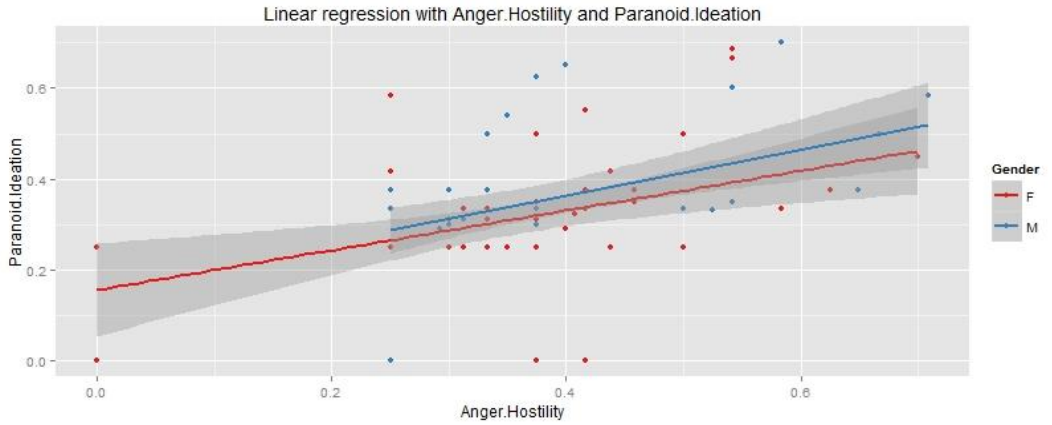


Το διάγραμμα διασποράς δείχνει το κάθε ζεύγος χαρακτηριστικών του συνόλου δεδομένων των εννέα υποκλιμάκων ψυχωπαθολογίας. Η διευθέτιση των γραφημάτων διασποράς ζευγών χαρακτηριστικών σε πινακοειδή μορφή είναι γνωστή και ως μήτρα διαγραμμάτων διασποράς, και παρέχει ένα οργανωμένο τρόπο ώστε να εξετάζεται ταυτόχρονα ένα μεγάλο πλήθος από διαγράμματα διασποράς.

Κώδικας: library(ggplot2)

```
df.SCL90.same.city.y<- subset(df.SCL90, df.SCL90$Same.City=='yes')
df.SCL90.same.city.n<- subset(df.SCL90, df.SCL90$Same.City=='no')
df.SCL90.same.city.y.urban<- df.SCL90.same.city.y[df.SCL90.same.city.y$City.of.Birth=='urban',]
df.SCL90.same.city.y.urban.Phob0<-df.SCL90.same.city.y.urban[df.SCL90.same.city.y.urban$Phobic.Anxiety==0,]
pairs(df.SCL90.same.city.y.urban.Phob0[12:20])
table(df.SCL90.same.city.y$City.of.Birth)
```

Γραμμική παλινδρόμηση:



Η εικόνα παρουσιάζει ένα γράφημα διασποράς με δύο γραμμές παλινδρόμησης. Η εξαρτημένη μεταβλητή είναι το φύλο το οποίο καλείται και ως απάντηση. Οι ανεξάρτητες μεταβλητές είναι ο παρανοειδής ιδεασμός στον άξονα των Y και η επιθετικότητα η οποία είναι ανεξάρτητη μεταβλητή στον άξονα των X. Οι ανεξάρτητες μεταβλητές ονομάζονται και ως προγνωστικοί δείκτες.Και για αυτήν την εφαρμογή απαιτείται εγκατάσταση του πακέτου ggplot2 (Ning Tan, Steinbach, & Kumar, 2010).

Κώδικας: library(ggplot2)

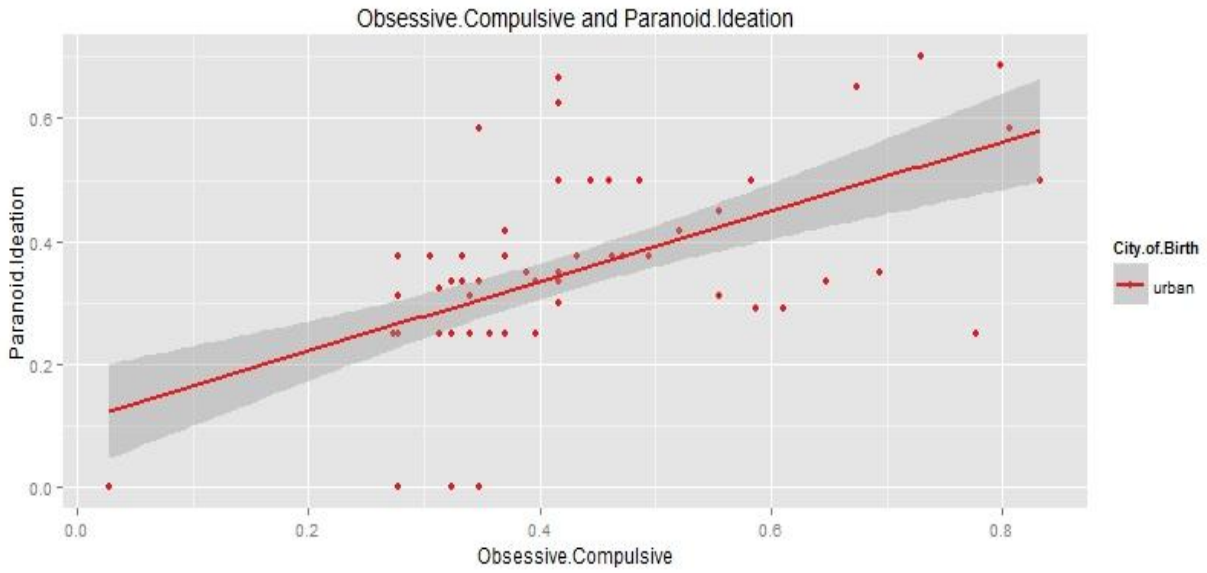
p<- ggplot(df.SCL90, aes_string(x='Anger.Hostility', y='Paranoid.Ideation'))

p <- p + aes_string(color='Gender') + geom_point()

p <- p + geom_smooth(method = "lm", size = 1)

p <- p + scale_color_brewer(type="qual",palette='Set1') + scale_fill_brewer() + ggtitle("Linear regression with Anger.Hostility and Paranoid.Ideation")

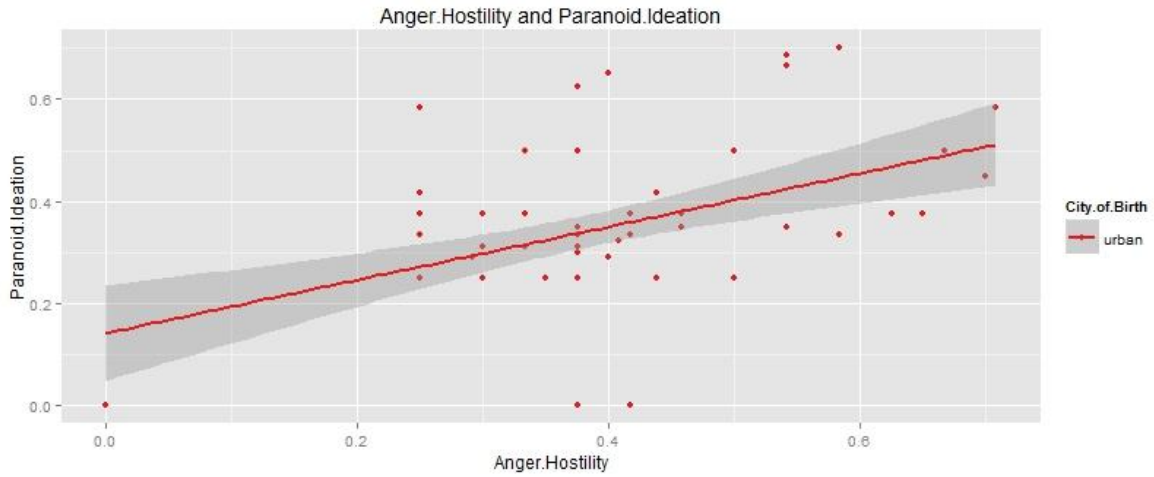
print(p)



Στην παραπάνω είκονα παρουσιάζεται μια γραμμική παλινδρόμηση με δεδομένα την εξαρτημένη μεταβλητή παρανοειδής ιδεασμός στον άξονα των Y και την ανεξάρτητη μεταβλητή ψυχαναγκαστικότητα – καταναγκαστικότητα στον άξονα των X. Επιπλέον παρατηρούμε την απόσταση που έχουν οι μεταβλητές από την κύρια ομάδα (γραμμή) και στην σκούρα γκρι περιοχή φαίνεται το διάστημα εμπιστοσύνης των παρατηρήσεων. Για αυτή την εφαρμογή απαιτήται εγκατάσταση του πακέτου ggplot2.

Κώδικας:

```
p1 <- ggplot(df.SCL90.same.city.y.urban, aes_string(x='Obsessive.Compulsive', y='Paranoid.Ideation'))
p1 <- p1 + aes_string(color='City.of.Birth') + geom_point()
p1 <- p1 + geom_smooth(method = "lm", size = 1)
p1 <- p1 + scale_color_brewer(type="qual",palette='Set1') + scale_fill_brewer()
print(p1)
```



Στην παραπάνω είκονα παρουσιάζεται μια γραμμική παλινδρόμηση με δεδομένα την εξαρτημένη μεταβλητή παρανοειδής ιδεασμός στον άξονα των Υ και την ανεξάρτητη μεταβλητή επιθετικότητα στον άξονα των Χ. Επιπλέον παρατηρούμε την απόσταση που έχουν οι μεταβλητές από την κύρια ομάδα (γραμμή) και στην σκούρα γκρι περιοχή φαίνεται το διάστημα εμπιστοσύνης των παρατηρήσεων. Για αυτή την εφαρμογή απαιτήται εγκατάσταση του πακέτου ggplot2.

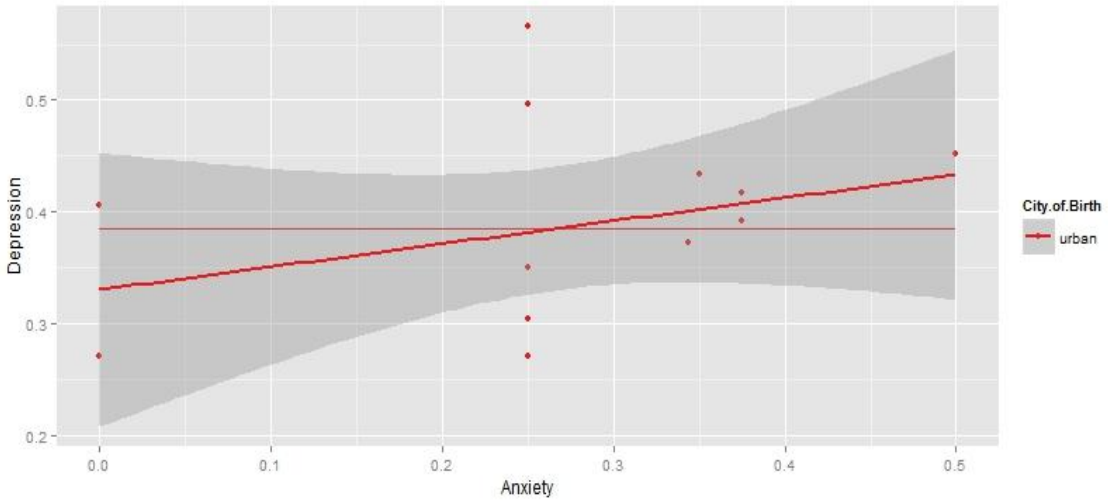
Κώδικας: p1 <- ggplot(df.SCL90.same.city.y.urban, aes_string(x='Anger.Hostility', y='Paranoid.Ideation'))

p1 <- p1 + aes_string(color='City.of.Birth') + geom_point()

p1 <- p1 + geom_smooth(method = "lm", size = 1)

p1 <- p1 + scale_color_brewer(type="qual",palette='Set1') + scale_fill_brewer()+ ggtitle("Anger.Hostility and Paranoid.Ideation")

print(p1)



Στην παραπάνω είκονα παρουσιάζεται μια γραμμική παλινδρόμηση με δεδομένα την εξαρτημένη μεταβλητή κατάθλιψη και την ανεξάρτητη μεταβλητή άγχος. Παρατηρούμε την απόσταση που έχουν οι μεταβλητές από την κύρια ομάδα (γραμμή) και στην σκούρα γκρι περιοχή φαίνεται το διάστημα εμπιστοσύνης των παρατηρήσεων. Και στην συγκεκριμένη εφαρμογή απαιτήται εγκατάσταση του πακέτου ggplot2.

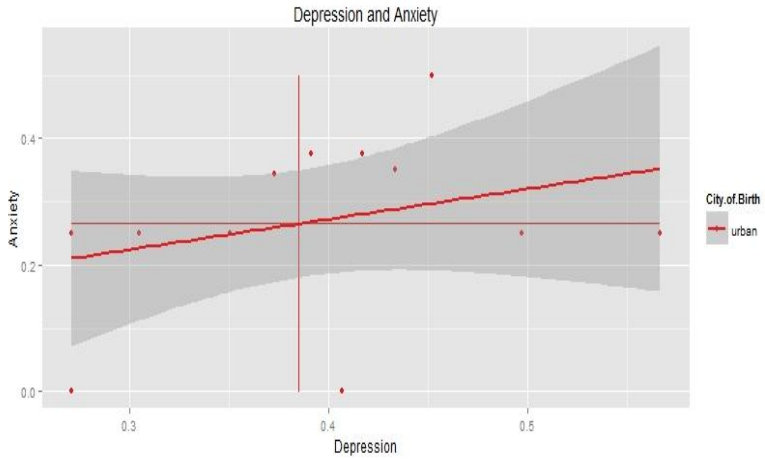
Κώδικας: p2 <- ggplot(df.SCL90.same.city.y.urban.Phob0, aes_string(x='Anxiety', y='Depression'))

p2 <- p2 + aes_string(color='City.of.Birth') + geom_point() + geom_line(stat = "hline", yintercept="mean")

p2 <- p2 + geom_smooth(method = "lm", size = 1)

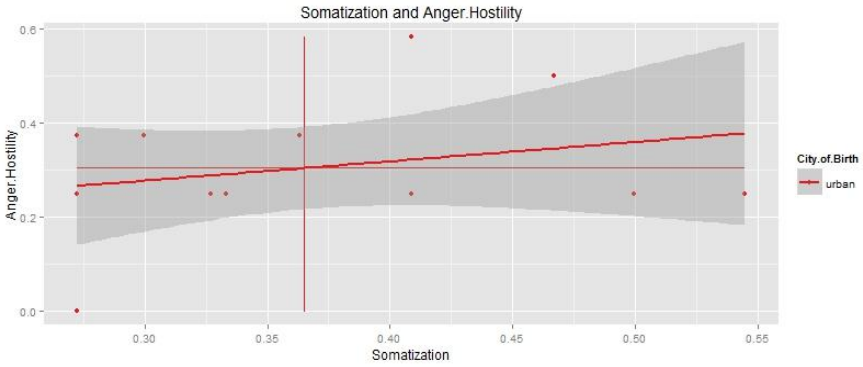
p2 <- p2 + scale_color_brewer(type="qual",palette='Set1') + scale_fill_brewer()

print(p2)



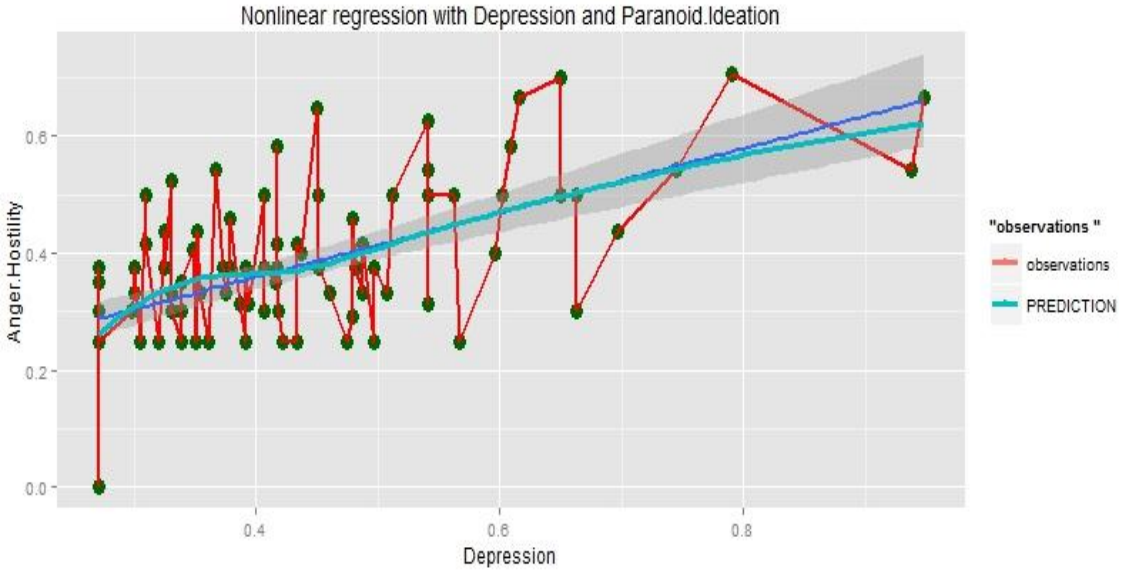
Στην παραπάνω είκονα παρουσιάζεται μια γραμμική παλινδρόμηση με δεδομένα το άγχος και την κατάθλιψη με τα οποία παρατηρούμε την απόσταση που έχουν οι μεταβλητές από την κύρια ομάδα (γραμμή) και στην σκούρα γκρι περιοχή φαίνεται το διάστημα εμπιστοσύνης των παρατηρήσεων. Η κατακόρυφη και οριζόντια κόκκινη γραμμή δείχνουν το μέσο των σημείο των δύο μεταβλητών. Και στην συγκεκριμένη εφαρμογή απαιτήται εγκατάσταση του πακέτου ggplot2.

```
Κώδικας: p1 <- ggplot(df.SCL90.same.city.y.urban.Phob0, aes_string(x='Depression', y='Anxiety'))
p1 <- p1 + aes_string(color='City.of.Birth') + geom_point() + geom_line( stat = "vline", xintercept="mean")
p1 <- p1 + geom_line( stat = "hline", yintercept="mean")
p1 <- p1 + geom_smooth(method = "lm", size = 1)
p1 <- p1 + scale_color_brewer(type="qual",palette='Set1') + scale_fill_brewer()
print(p1)
```



Στην παραπάνω είκονα παρουσιάζεται μια γραμμική παλινδρόμηση με δεδομένα θυμός-επιθετικότητα και σωματοποίηση με τα οποία παρατηρούμε την απόσταση που έχουν οι μεταβλητές από την κύρια ομάδα (γραμμή) και στην σκούρα γκρι περιοχή φαίνεται το διάστημα εμπιστοσύνης των παρατηρήσεων. Η κατακόρυφη και οριζόντια κόκκινη γραμμή δείχνουν το μέσο των σημείο των δύο μεταβλητών. Και στην συγκεκριμένη εφαρμογή απαιτήται εγκατάσταση του πακέτου ggplot2.

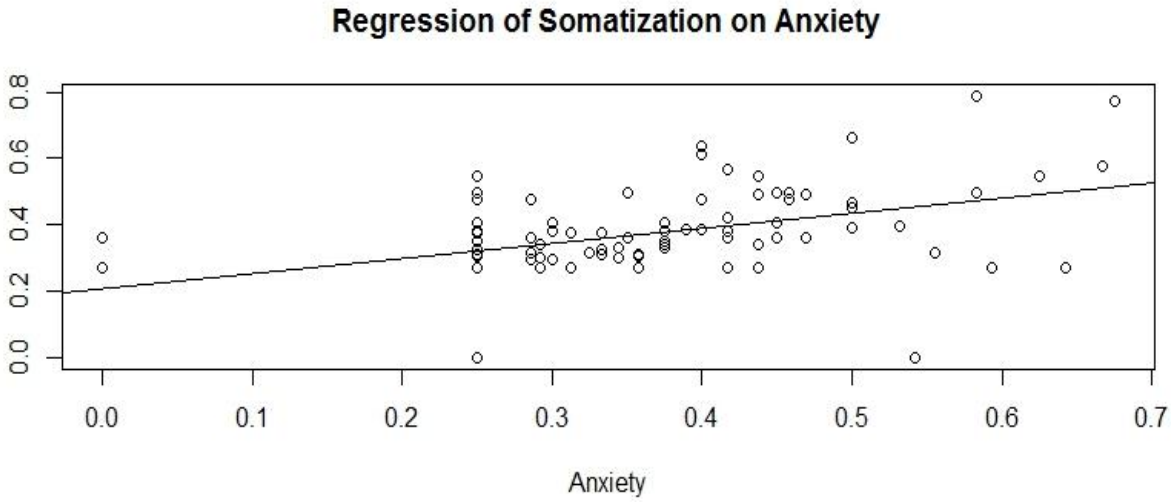
```
Κώδικας: p1 <- ggplot(df.SCL90.same.city.y.urban.Phob0, aes_string(x='Somatization', y='Anger.Hostility'))
p1 <- p1 + aes_string(color='City.of.Birth') + geom_point() + geom_line( stat = "vline", xintercept= "mean")
p1 <- p1 + geom_line( stat = "hline", yintercept="mean")
p1 <- p1 + geom_smooth(method = "lm", size = 1)
p1 <- p1 + scale_color_brewer(type="qual",palette='Set1') + scale_fill_brewer()
print(p1)
```



Η παραπάνω εικόνα απεικονίζει μια μη γραμμική παλινδρόμηση με δεδομένα τον θυμό-επιθετικότητα στον κάθετο άξονα και την κατάθλιψη στον οριζόντιο. Στην εικόνα φαίνονται οι παρατηρήσεις με πρασινο χρώμα και η κόκκινη γραμμή που ενώνει τα σημεία είναι οι παρατηρήσεις ενώ η γαλάζια φανερώνει την προσέγγιση της μεταβλητής Y με μη γραμμικό τρόπο σε σχέση με την X. Και για την συγκεκριμένη απεικόνιση χρησιμοποιείται η βιβλιοθήκη ggplot2.

Κώδικας: library(ggplot2)

```
ggplot ( df.SCL90, aes_string(x= "Depression","Anger.Hostility"))+
geom_point(colour = "blue",size=1 )+
geom_point(aes(color = "observations "))+
geom_point(colour = "darkgreen",size= 4)+
geom_line(colour = "red",size=0.9)+
geom_smooth(method="lm", size=1)+
geom_smooth(aes(colour = "PREDICTION" ),se= F, size = 1.1)+ ggtitle("Nonlinear regression with Depression and Paranoid.Ideation")
```



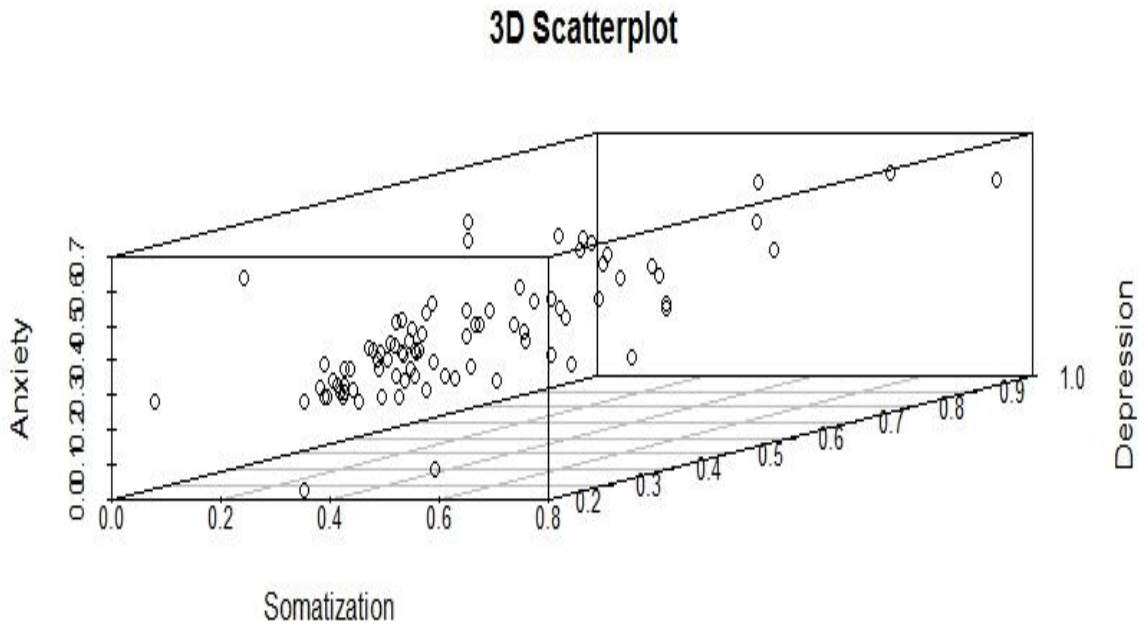
Το παραπάνω γράφημα απεικονίζει μια γραμμική παλινδρόμηση συσχετίζοντας τις αριθμητικές μεταβλητές σωματοποίησης και άγχος.

```
Κώδικας:attach(df.SCL90)
plot(Anxiety, Somatization)
abline(lm(Somatization~Anxiety))
title("Regression of Somatization on Anxiety")
```

4.5 Πολυμεταβλητή Ανάλυση Δεδομένων

4.5.1 Διαγράμματα Διασποράς

Τρισδιάστατη απεικόνιση διαγράμματος διασποράς:

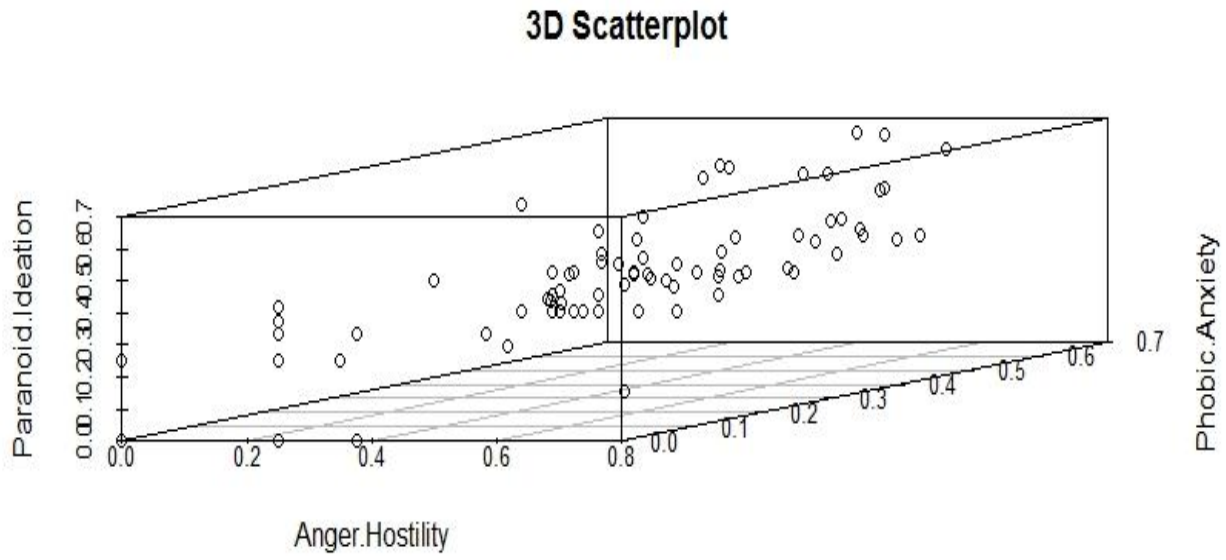


Στην εικόνα παρουσιάζεται ένας τρισδιάστατος κύβος ο οποίος έχει λάβει τις μεταβλητές άγχος, σωματοποίηση και κατάθλιψη. Με την βοήθεια των μεταβλητών αυτών παρουσιάζεται μέσα στον κύβο οι τιμές του δείγματος. Για την συγκεκριμένη εφαρμογή χρειάζεται η εγκατάσταση του πακέτουscatterplot3d.

Κώδικας:library(scatterplot3d)

attach(df.SCL90)

scatterplot3d(Somatization,Depression,Anxiety, main="3D Scatterplot")

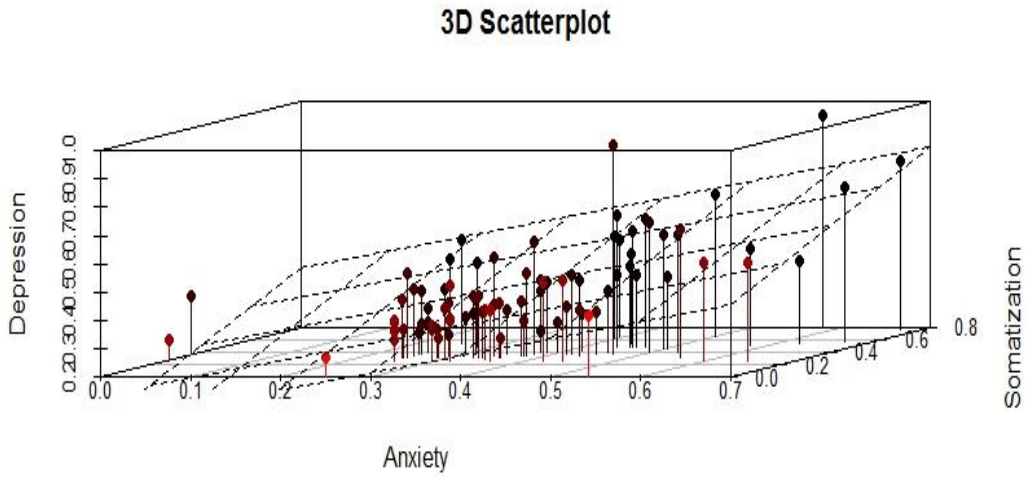


Στην εικόνα παρουσιάζεται ένας τρισδιάστατος κύβος ο οποίος έχει λάβει τις μεταβλητές παρανοειδής ιδεασμός, θυμός-επιθετικότητα και φοβικό άγχος. Με την βοήθεια των μεταβλητών αυτών παρουσιάζεται μέσα στον κύβο οι τιμές του δείγματος. Για την συγκεκριμένη εφαρμογή χρειάζεται η εγκατάσταση του πακέτουscatterplot3d.

Κώδικας:library(scatterplot3d)

attach(df.SCL90)

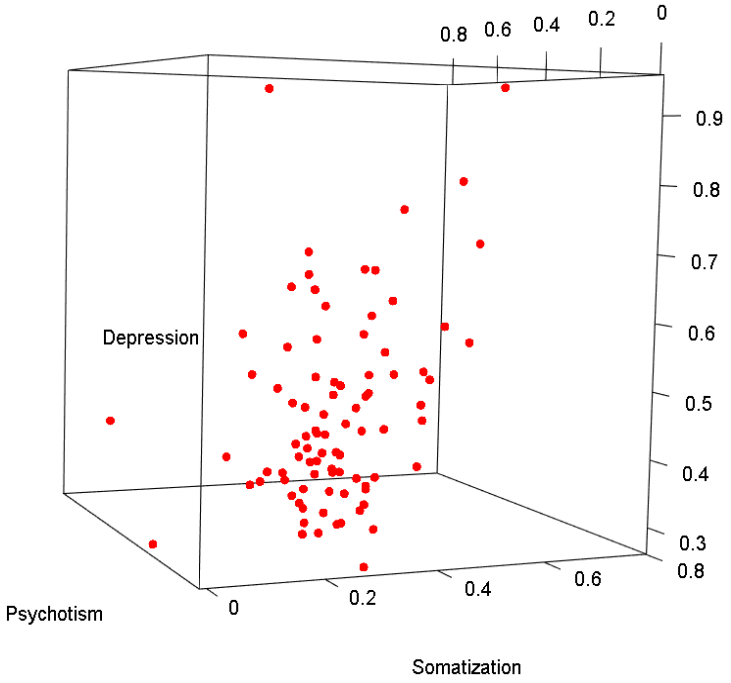
scatterplot3d(Anger.Hostility,Phobic.Anxiety,Paranoid.Ideation, main="3D Scatterplot")



Στην εικόνα παρουσιάζεται ένας τρισδιάστατος κύβος ο οποίος έχει λάβει τις μεταβλητές κατάθλιψη, άγχος και σωματοποίηση. Με την βοήθεια των μεταβλητών αυτών παρουσιάζεται μέσα στον κύβο με πιο μεγάλη λεπτομέρια οι τιμές που έχουν δοθεί στο δείγμα. Για την συγκεκριμένη εφαρμογή χρειάζεται η εγκατάσταση του πακέτουscatterplot3d.

Κώδικας:library(scatterplot3d)

```
attach(df.SCL90)
s3d <-scatterplot3d(Anxiety,Somatization,Depression, pch=16, highlight.3d=TRUE,
                    type="h", main="3D Scatterplot")
fit<- lm(Depression ~ Anxiety+Somatization)
s3d$plane3d(fit)
```

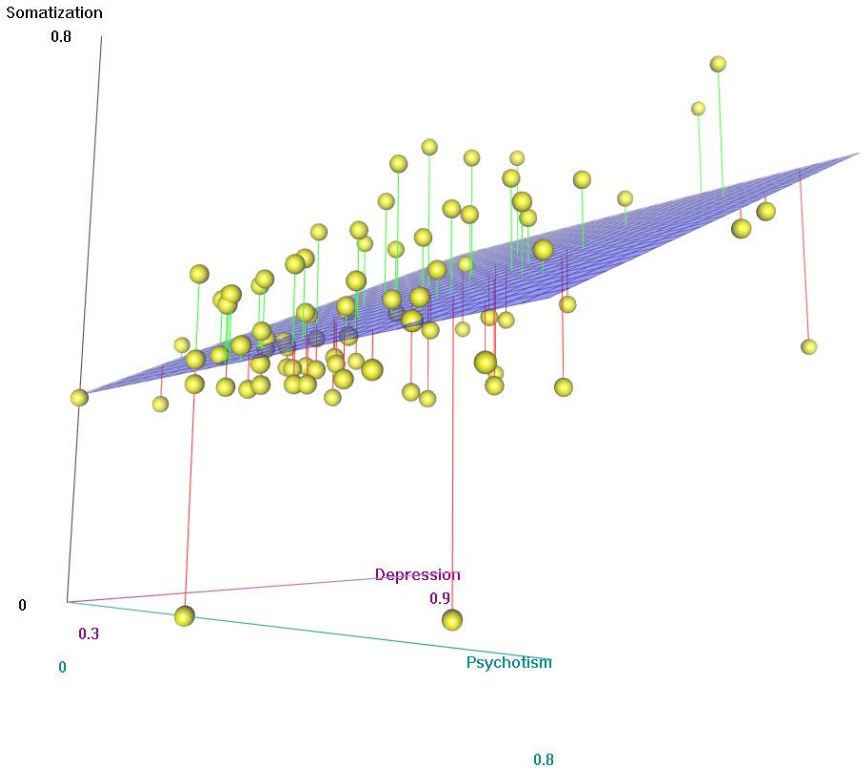


Στην εικόνα παρουσιάζεται ένας τρισδιάστατος κύβος ο οποίος έχει λάβει τις μεταβλητές ψυχωτισμός, κατάθλιψη και σωματοποίηση. Με την βοήθεια των μεταβλητών αυτών παρουσιάζεται μέσα στον κύβο με μεγαλύτερη λεπτομέρια οι τιμές που έχουν δοθεί στο δείγμα. Για την συγκεκριμένη εφαρμογή χρειάζεται η εγκατάσταση του πακέτουrgl και με την οποία μπορούμε να περιστρέψουμε τον κύβο για να έχουμε μια καλύτερη είκονα των τιμών που υπάρχουν μέσα στον κύβο.

Κώδικας:library(rgl)

attach(df.SCL90)

plot3d(Depression,Somatization,Psychotism, col="red", size=9)



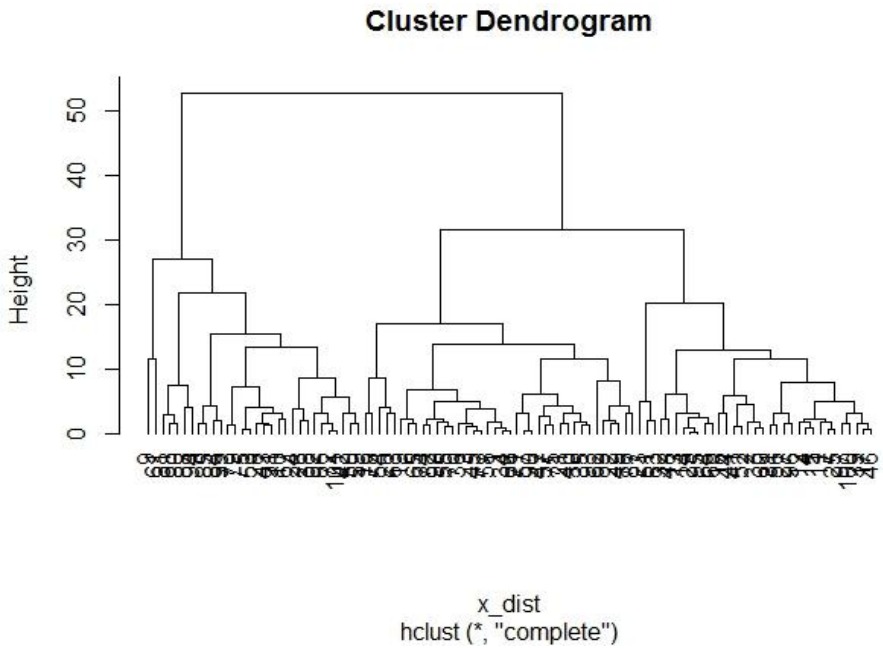
Στην εικόνα παρουσιάζεται ένα τρισδιάστατο περιστροφικό διάγραμματο οποίο έχει λάβει τις μεταβλητές ψυχωτισμός, κατάθλιψη και σωματοποίηση. Με την βοήθεια των μεταβλητών αυτών παρουσιάζεται μέσα στο περιστροφικό διάγραμμα με μεγάληλεπτομέρια οι τιμές που έχουν δοθεί στο δείγμα. Για την συγκεκριμένη εφαρμογή χρειάζεται η εγκατάσταση του πακέτουRcmdrκαι με την οποία μπορούμε να περιστρέψουμε το διάγραμμα για να έχουμε μια καλύτερη είκονα των τιμών που υπάρχουν μέσα στον κύβο.

Κώδικας: library(Rcmdr)

attach(df.SCL90)

scatter3d(Depression,Somatization,Psychotism)

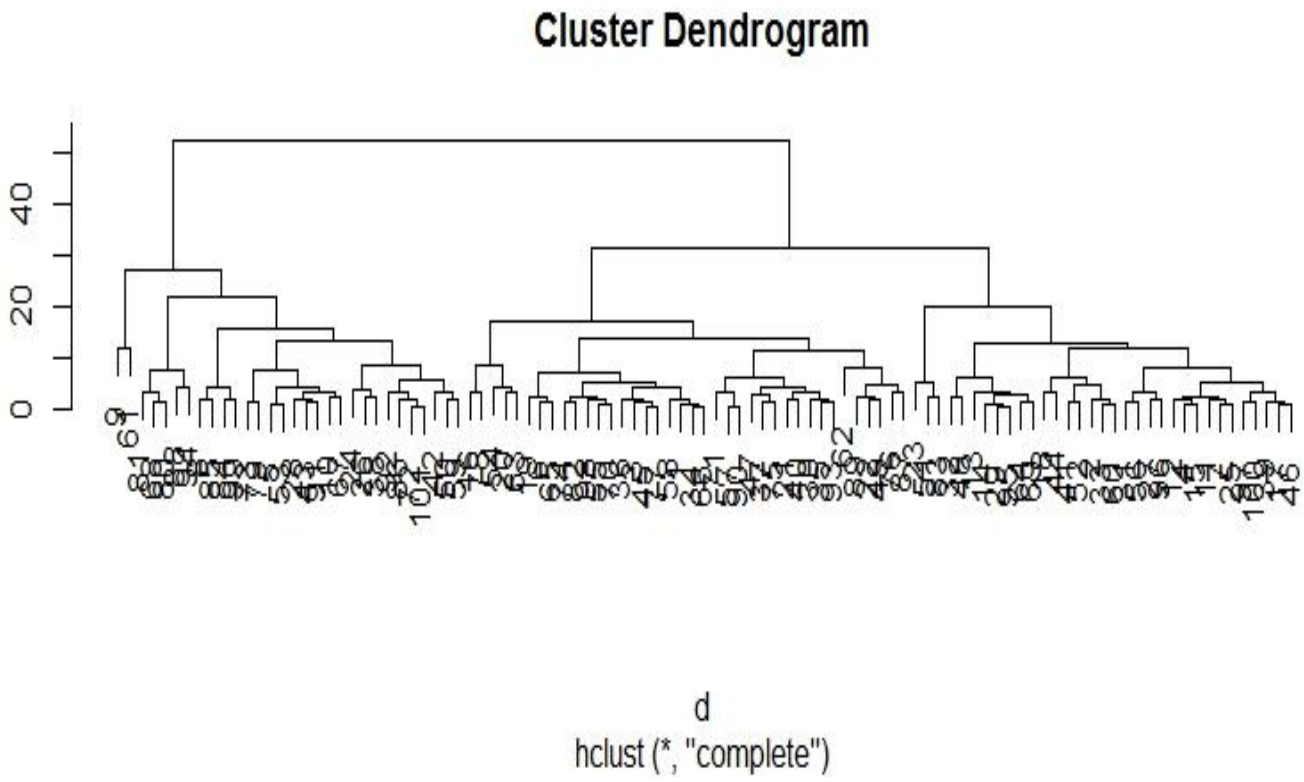
4.5.2 Διαγράμματα Ομαδοποίησης - Συσταδοποίησης



Το δεντρόγραμμα που ακολουθεί απεικονίζει όλες τις μεταβλητές της ψυχοπαθολογίας. Ο τύπος του δεντρογράμματος καλείται πλήρης ή ολοκληρωμένος. Είναι μια από τις πολλές μεθόδους ιεραρχικής ομαδοποίησης. Κατά την έναρξη της διαδικασίας, κάθε στοιχείο είναι ένα σύμπλεγμα και στην συνέχεια οι συστάδες διαδοχικά συνδιάζονται σε μεγαλύτερες συστάδες μέχρις ότου όλα τα στοιχεία στο τέλος να ανήκουν σε μία ομάδα. Για την οπτικοποίηση της παραπάνω εικόνας απαιτείται η εγκατάσταση του πακέτου dendextend.

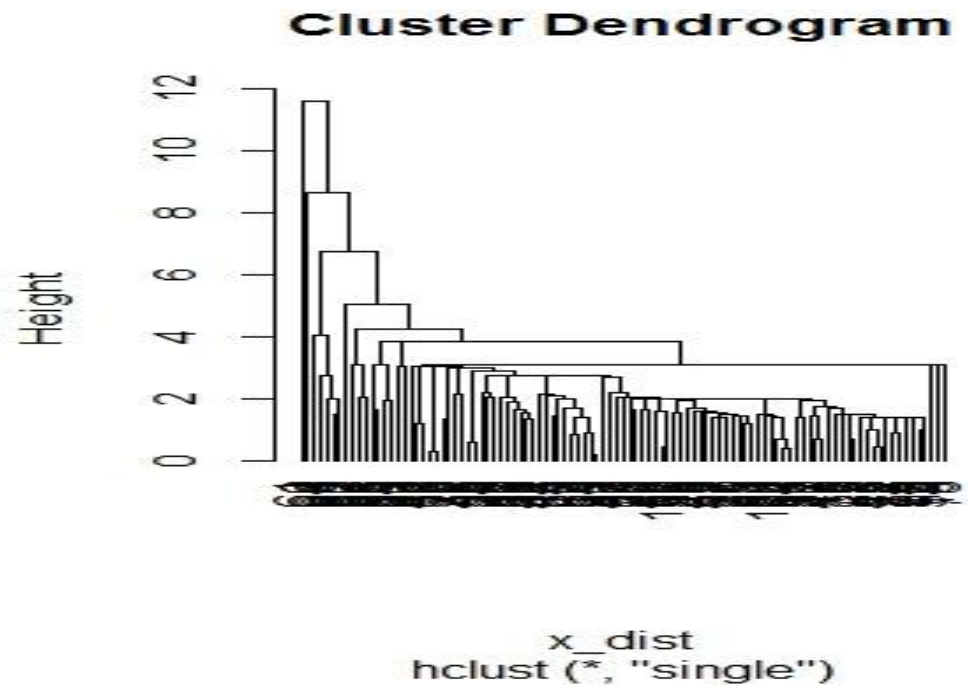
Κώδικας: library(dendextend)

```
library(dendextendRcpp)
data(df.SCL90)
df.SCL90_data <-(data=df.SCL90)
df.SCL90_data <- df.SCL90(c["Anxienty", "Somatization", "Psychotism", "Obsessive.Compulsive", "Interpersonal.Sensitivity","Depression", "Anger.Hostility", "Phobic.Anxiety", "Paranoid.Ideation" ])
x_dist<- dist(df.SCL90, diag = TRUE)
hc1 <- hclust(x_dist, method = "complete" )
plot(hc1,hang = -1)
dend1 <- as.dendrogram(hc1)
str(hc1)
str(dend1)
str(unclass(dend1))
```



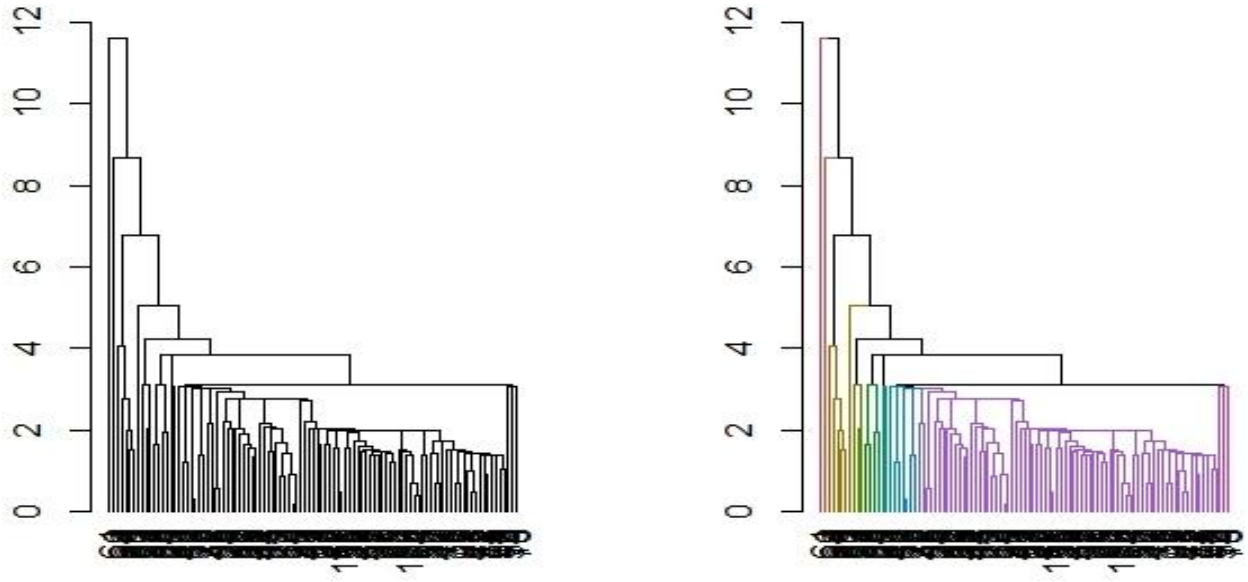
Έναςάλλος διαφορετικός κώδικας για την οπτικοποίηση ενός πλήρης δεντρογράμματος είναι ο ακόλουθος.

```
Κώδικας:d <- dist(as.matrix(df.SCL90))
hc<- hclust(d)
plot(hc)
```



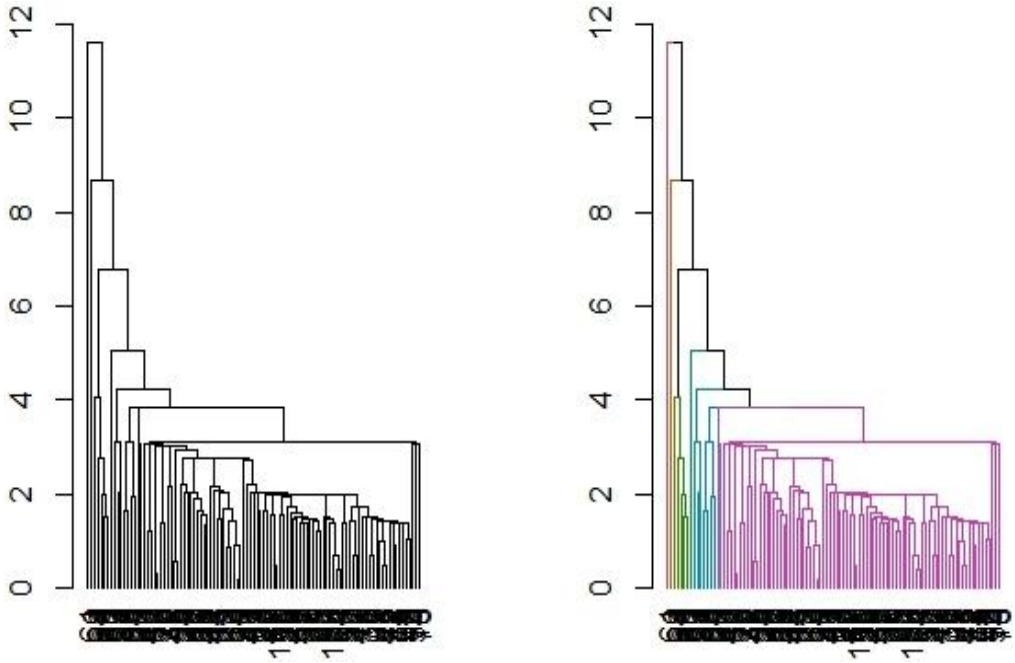
Η παραπάνω εικόνα απεικονίζει ένα δεντρόγραμμα απλής σύνδεσης (singlelinkage). Το κριτήριο σύνδεσης που χρησιμοποιείται από τον συσσωρευτικό αλγόριθμο βασίζεται κυρίως στην ελάχιστη απόσταση των παρατηρήσεων αλλά και στα δύο πιο όμοια σημεία στις διαφορετικές συστάδες. Αυτό γραφικά αναπαρίσταται με μια ακμή. Για την συγκεκριμένη απεικόνιση απαιτείται εγκατάσταση του πακέτου dendextend.

```
Κώδικας:library(dendextend)
library(dendextendRcpp)
data(df.SCL90)
df.SCL90_data <- (data=df.SCL90)
df.SCL90_data <- df.SCL90(c["Anxienty", "Somatization", "Psychotism", "Obsessive.Compulsive", "Interpersonal.Sensitivity", "Depression" , "Anger.Hostility", "Phobic.Anxiety", "Paranoid.Ideation" ])
x_dist<- dist(df.SCL90, diag = TRUE)
hc1 <- hclust(x_dist, method = "single" )
plot(hc1,hang = -1)
dend1 <- as.dendrogram(hc1)
str(hc1)
str(dend1)
str(unclass(dend1))
```



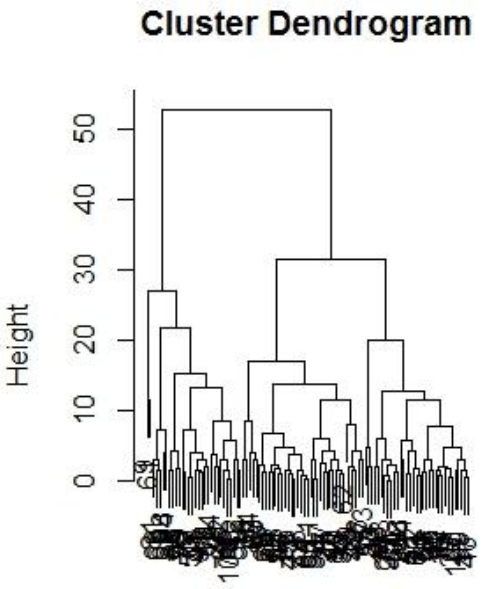
Στην παραπάνω εικόνα απεικονίζονται δύο δεντρογράμματα μονού τύπου τα οποία έχουν μέγεθος είκοσι συστάδων. Στην δεξιά εικόνα φαίνονται όλα τα δεδομένα ενώ στην αριστερί διακρίνονται οι ομαδοποιήσεις με τους διαφορετικούς τύπους χρωμάτων. Για την παραπάνω οπτικοποίησηαπαιτήται εγκατάσταση του πακέτου dendextend.

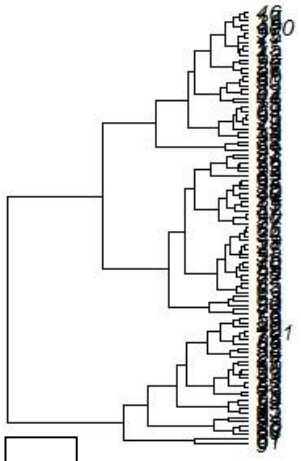
```
Κώδικας: require(dendextend)
dend1_mod_01 <- dend1
dend1_mod_01 <- color_branches(dend1_mod_01,k = 20)
col_for_labels<- c ("purple","red","green","blue","yellow")
dend1_mod_01 <- col_for_labels(dend1_mod_01,col = col_for_labels )
par(mfrow = c (1,2))
plot(dend1)
plot(dend1_mod_01)
```



Στην παραπάνω εικόνα απεικονίζονται δύο δεντρογράμματα μονού τύπου τα οποία έχουν μέγεθος εννέα συστάδων. Στην δεξιά εικόνα φαίνονται όλα τα δεδομένα ενώ στην αριστερί διακρίνονται οι ομαδοποιήσεις με τους διαφορετικούς τύπους χρωμάτων. Για την παραπάνω οπτικοποίησηαπαιτήται εγκατάσταση του πακέτου dendextend.

```
Κώδικας: require(dendextend)
dend1_mod_01 <- dend1
dend1_mod_01 <- color_branches(dend1_mod_01,k = 9)
col_for_labels<- c ("purple","red","green","blue","yellow")
dend1_mod_01 <- col_for_labels(dend1_mod_01,col = col_for_labels )
par(mfrow = c (1,2))
plot(dend1)
plot(dend1_mod_01)
```





2

1

dist(df.SCL90)
hclust (*, "complete")

Κώδικας:library(ape)

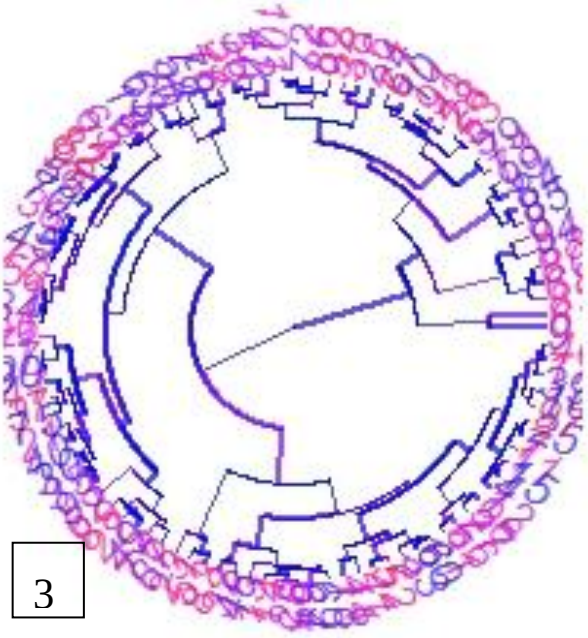
hc = hclust(dist(df.SCL90))

plot(hc)

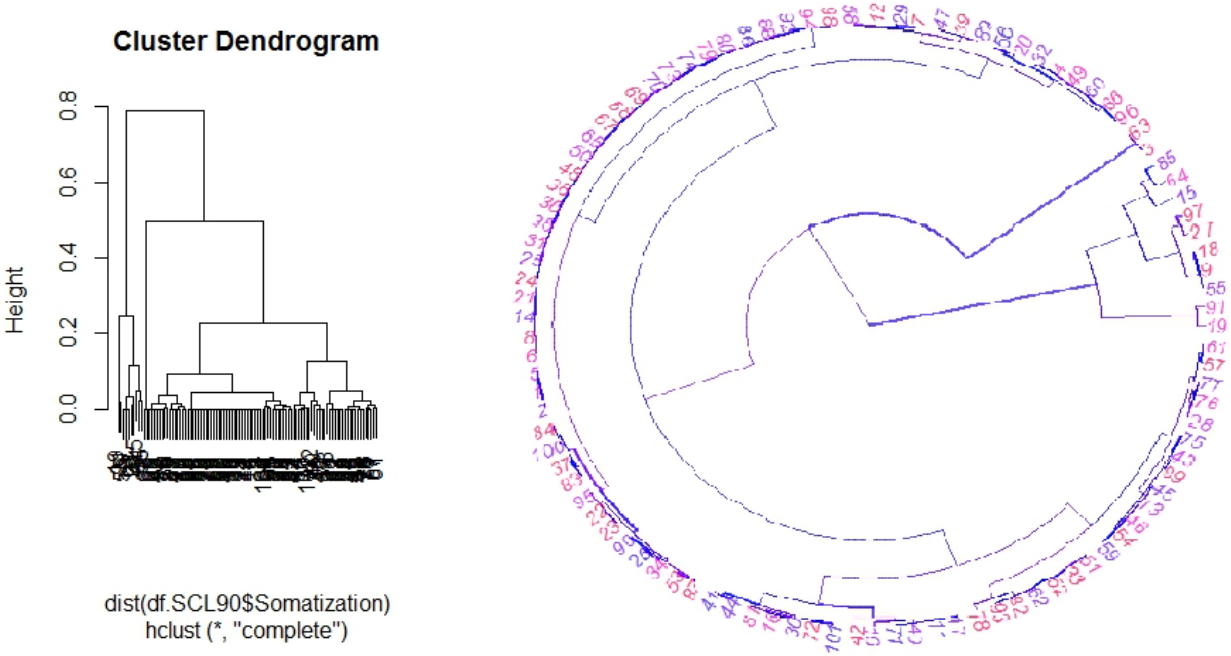
plot(as.phylo(hc), cex = 0.9, label.offset = 1)

plot(as.phylo(hc), type = "fan", tip.color = hsv(runif(15, 0.65,0.95), 1, 1, 0.7), edge.color = hsv(runif(10, 0.65, 0.75), 1, 1, 0.7), edge.width = runif(20,0.5, 3),use.edge.length = TRUE, col = "gray80")

Ο παραπάνω κώδικας οπτικοποιεί τρεις διαφορετικές εικόνες (1,2,3). Στην πρώτη εικόνα παρουσιάζεται ένα πολύ μεγάλο πλήρης κάθετο ιεραρχικό δεντρόγραμμα το οποίο περιέχει όλα τα δεδομένα της βάσης. Στην δεύτερη εικόνα το δεντρόγραμμα στρέφεται οριζόντια ενώ στην τελευταία εικόνα παρουσιάζεται ένα κυκλικό δεντρόγραμμα. Για την συγκεκριμένη οπτικοποίηση απαιτείται εγκατάσταση του πακέτου ape.



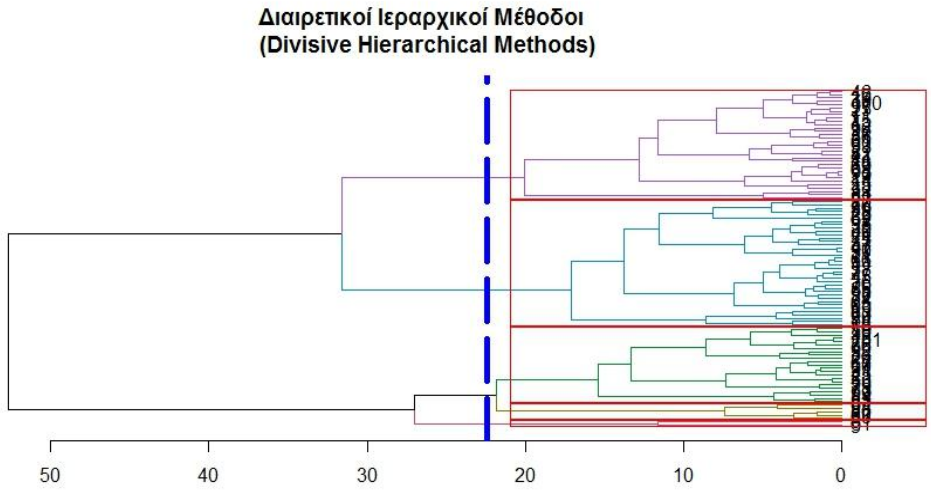
3



Κώδικας: library(ape)

```
hc = hclust(dist(df.SCL90$Somatization,df.SCL90$Psychotism,df.SCL90$Depression))
plot(hc)
plot(as.phylo(hc), type = "fan", tip.color = hsv(runif(15, 0.65,0.95), 1, 1, 0.7), edge.color = hsv(runif(10, 0.65, 0.75), 1, 1, 0.7), edge.width = runif(20,0.5, 3),use.edge.length = TRUE, col = "gray80"
```

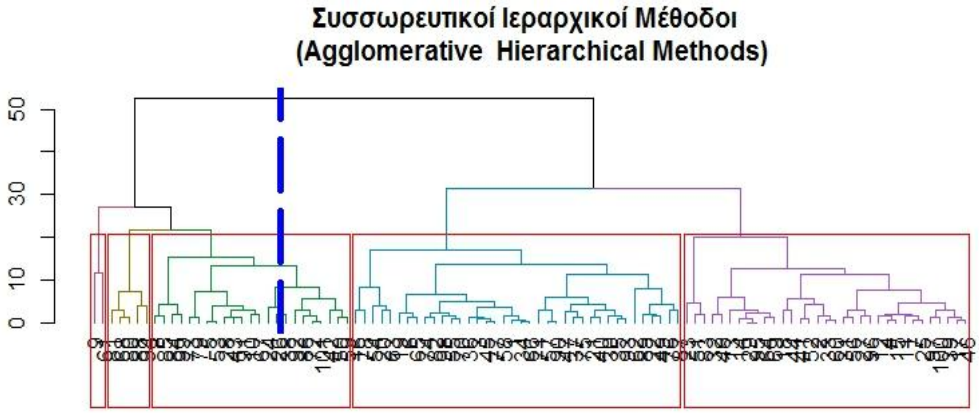
Στην παραπάνω εικόνα παρουσιάζονται δυόδεντρογράμματα σύμφωνα με τις μεταβλητές σωματοποίησης, ψυχωτισμός και κατάθλιψη. Η δεξιά εικόνα οπτικοποιεί ένα πλήρης κάθετο δεντρόγραμμα ενώ η αριστερή εικόνα παρουσιάζει τα δεδομένα σε ένα κυκλικό δεντρόγραμμα ή αλλιώς φυλογενετικό δέντρο.



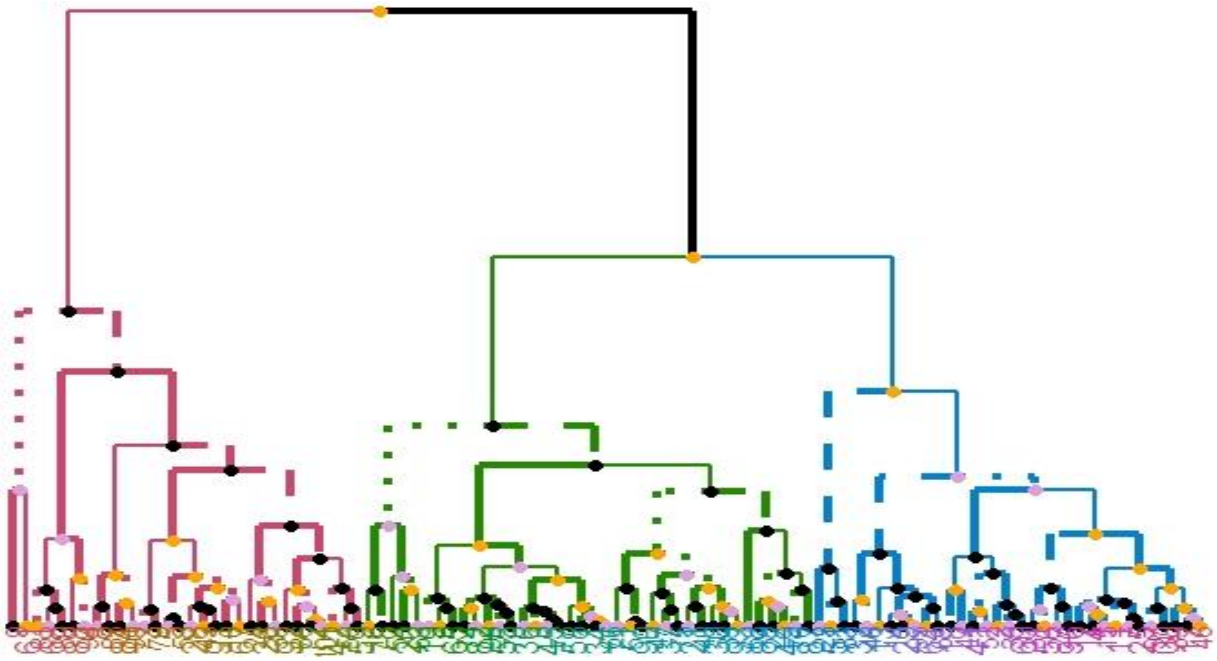
Η εικόνα απεικονίζει τα δεδομένα σε ένα οριζόντιο δεντρόγραμμα το οποίο έχει οριστεί με πέντε συστάδες. Για την συγκεκριμένη οπτικοποίηση χρειάζεται εγκατάσταση των πακέτων dendextend και dendextendRcpp.

```
Κώδικας:ds <- df.SCL90[,1:20]

library(dendextend)

library(dendextendRcpp)

dend<- df.SCL90[,1:20] %>% dist %>% hclust %>% as.dendrogram

dend %>% color_branches(k=5) %>% plot(horiz=TRUE, main = "Διαιρετικοί Ιεραρχικοί Μέθοδοι \n (Divisive Hierarchical Methods)")

dend %>% rect.dendrogram(k=5,horiz=TRUE)

abline(v = heights_per_k.dendrogram(dend)["5"] + .6, lwd = 5, lty = 5, col = "blue")
```

Η εικόνα απεικονίζει τα δεδομένα σε ένα κάθετο δεντρόγραμμα το οποίο έχει οριστεί με πέντε συστάδες.




```
Κώδικας:dend<- df.SCL90[,1:20] %>% dist %>% hclust %>% as.dendrogram

dend %>% color_branches(k=5) %>% plot(horiz=FALSE, main = "ΣυσσωρευτικοίΙεραρχικοί Μέθοδοι \n (Agglomerative Hierarchical Methods)")

dend %>% rect.dendrogram(k=5,horiz=FALSE)

abline(v = heights_per_k.dendrogram(dend)["5"] + .6, lwd = 5, lty = 5, col = "blue")
```



Ηεικόναπαρουσιάζει μια διαφορετική οπτικοποίηση ενός κάθετου ολοκληρωμένου δεντρογράμματος με τρεις συστάδες χρωματισμένες με ροζ, πράσινο και μπλέ αντίστοιχα.. Το συγκεκριμένο γράφημα καλείται και ως δεντρόγραμμα χαρτογράφησης.

Κώδικας: dend<- df.SCL90[1:101,-5] %>% dist %>% hclust %>% as.dendrogram %>%

set("branches_k_color", k=3) %>% set("branches_lwd", c(1.5,1,1.5)) %>%

set("branches_lty", c(1,1,3,1,1,2)) %>%

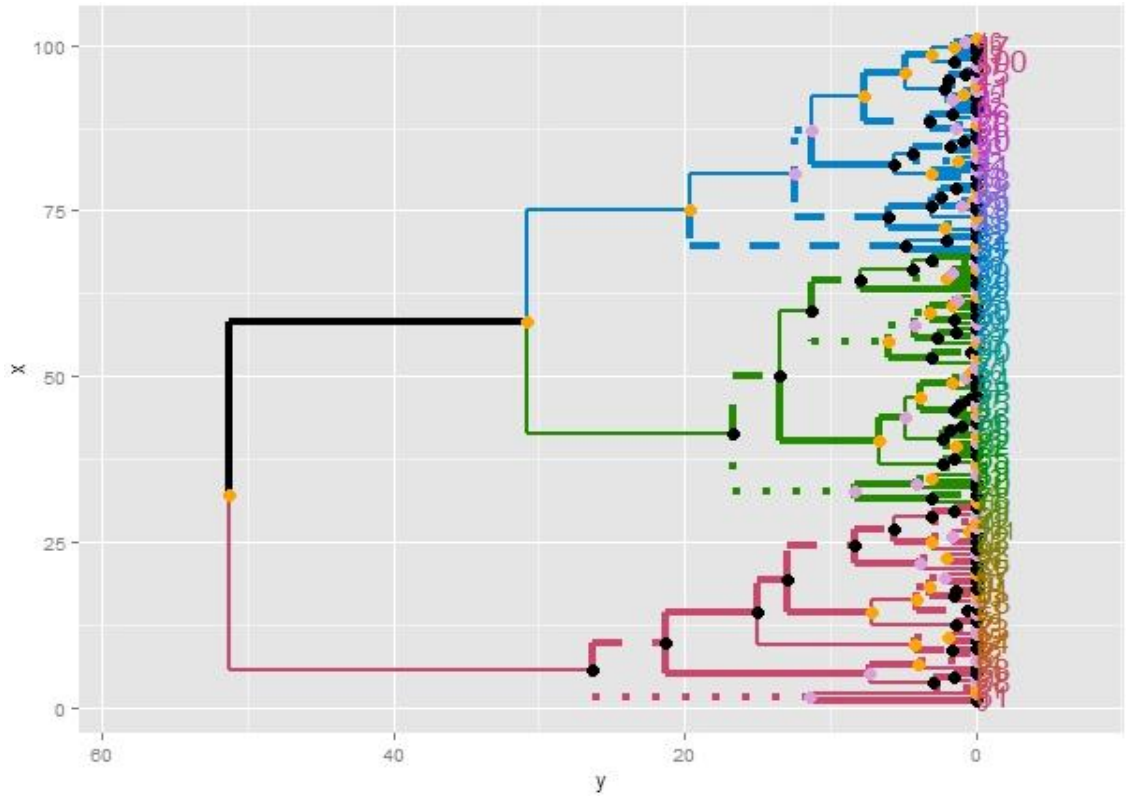
set("labels_colors") %>% set("labels_cex", c(.9,1.2)) %>%

set("nodes_pch", 19) %>% set("nodes_col", c("orange", "black", "plum", NA))

ggd1 <- as.ggdend(dend)

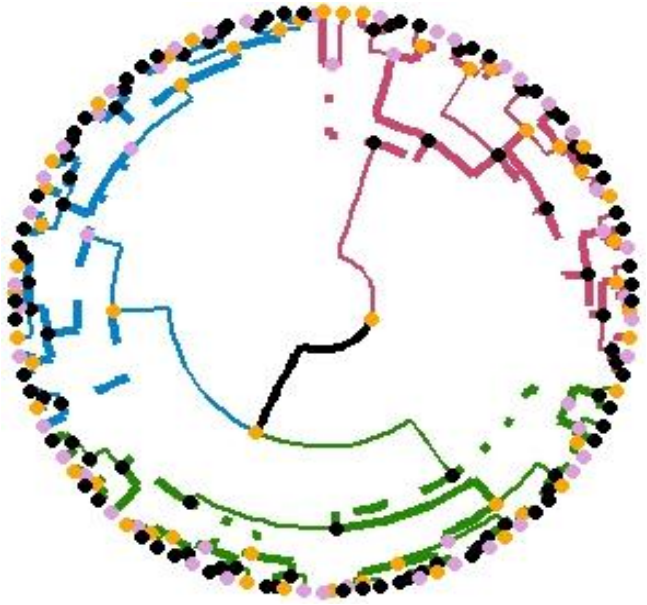
library(ggplot2)

ggplot(ggd1)



Ηεικόναπαρουσιάζει μια διαφορετική οπτικοποίηση ενός οριζόντιου ολοκληρωμένου δεντρογράμματος με τρεις συστάδες χρωματισμένες με ροζ, πράσινο και μπλέ αντίστοιχα.

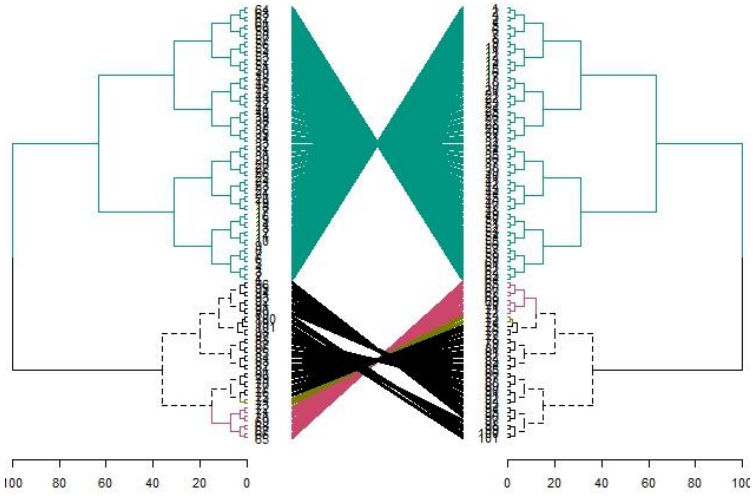
```
Κώδικας:dend<- df.SCL90[1:101,-5] %>% dist %>% hclust %>% as.dendrogram %>%
set("branches_k_color", k=3) %>% set("branches_lwd", c(1.5,1,1.5)) %>%
set("branches_lty", c(1,1,3,1,1,2)) %>%
set("labels_colors") %>% set("labels_cex", c(.9,1.2)) %>%
set("nodes_pch", 19) %>% set("nodes_col", c("orange", "black", "plum", NA))
ggd1 <- as.ggdend(dend)
library(ggplot2)
ggplot(ggd1, horiz = TRUE, theme = NULL)
```



Η εικόνα απεικονίζει ένα κυκλικό δενδρόγραμμα με όλα τα δεδομένα από την βάση και είναι χωρισμένο σε τρεις συστάδες.

Κώδικας:

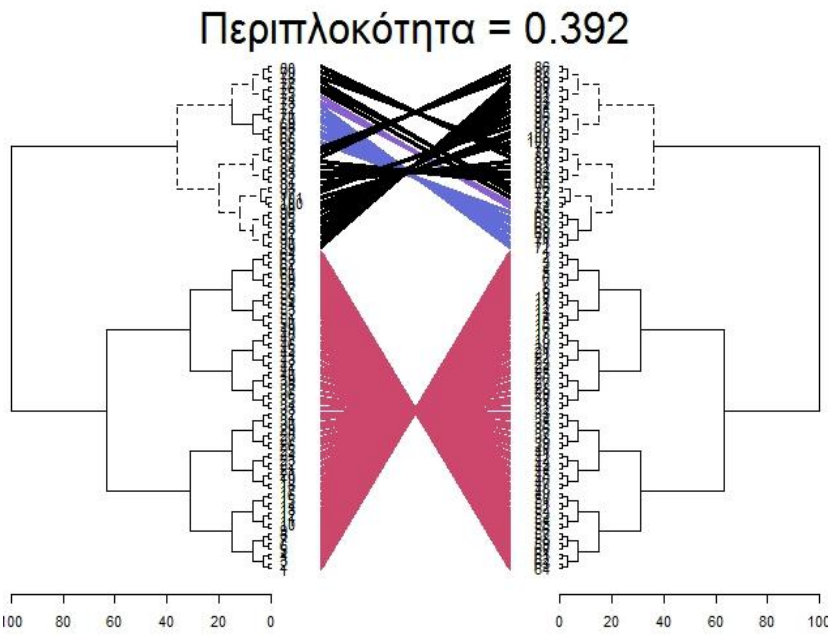
```
dend<- df.SCL90[1:101,-5] %>% dist %>% hclust %>% as.dendrogram %>%
set("branches_k_color", k=3) %>% set("branches_lwd", c(1.5,1,1.5)) %>%
set("branches_lty", c(1,1,3,1,1,2)) %>%
set("labels_colors") %>% set("labels_cex", c(.9,1.2)) %>%
set("nodes_pch", 19) %>% set("nodes_col", c("orange", "black", "plum", NA))
ggd1 <- as.ggdend(dend)
library(ggplot2)
ggplot(ggd1, labels = FALSE) + scale_y_reverse(expand = c(0.2, 0)) + coord_polar(theta="x")
```



Ο συγκεκριμένος κώδικα εμφανίζει δύο εικόνες. Στην πρώτη εικόνα απεικονίζεται ένα ολοκληρωμένο ιεραρχικό δεντρόγραμμα το οποίο στη συνέχεια αντιστρέφεται και ενώνει τις παρατηρήσεις που είναι ίδιες μεταξύ τους με διαφορετικά χρώματα.

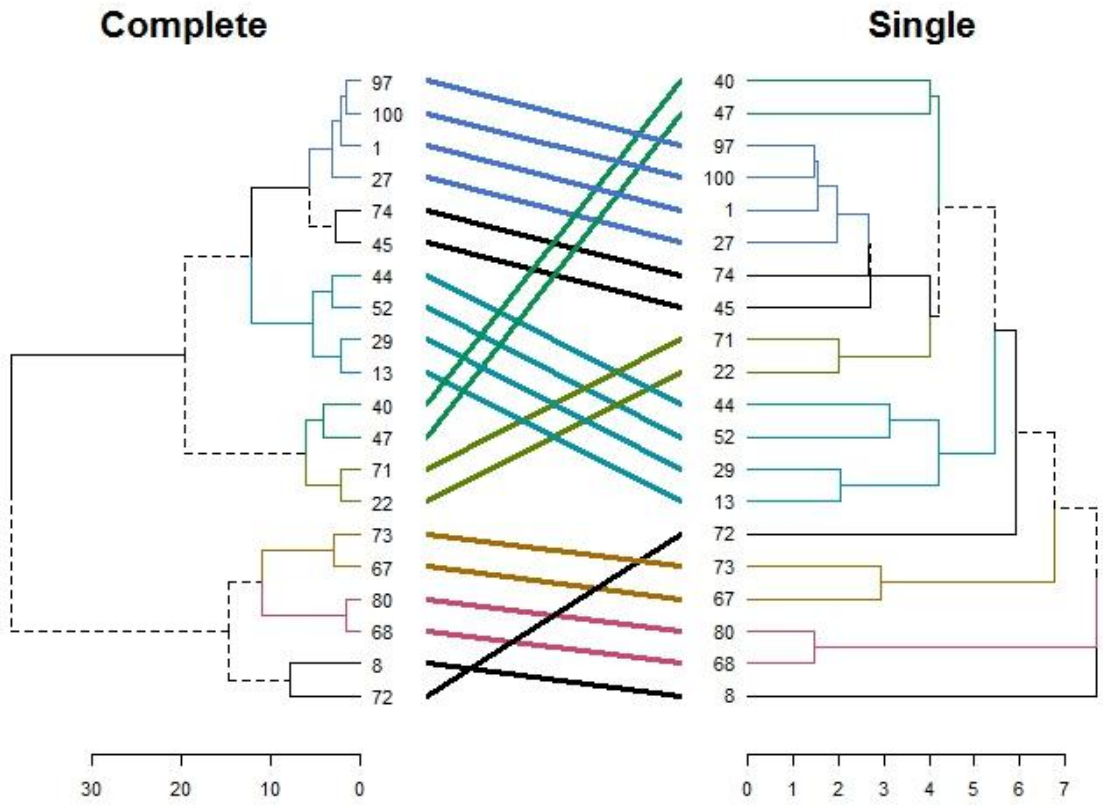
```
Κώδικας:dend15 <- c(1:101) %>% dist %>% hclust(method = "complete") %>% as.dendrogram
dend15 <- dend15 %>% set("labels_to_char")
dend51 <- dend15 %>% set("labels", as.character(101:1)) %>% match_order_by_labels(dend15)
dends_15_51 <- dendlist(dend15, dend51)
tanglegram(dends_15_51, common_subtrees_color_branches = TRUE)
x <- dends_15_51 %>% untangle(method = "ladderize")
x %>% plot(main = paste("Περιπλοκότητα =", round(entanglement(x), 3)))
```

Σε αυτήν την εικόνα είναι φανερό πως το δεντρόγραμμα αναποδογύρισε και εμφανίζεται η περιπλοκότητά δηλαδή η συνάρτηση εμπλοκής της ποιότητας διάταξης η οποία ισούται με 0,392 και είναι πολύ χαμηλή λόγο του μεγάλου αριθμού των παρατηρήσεων.

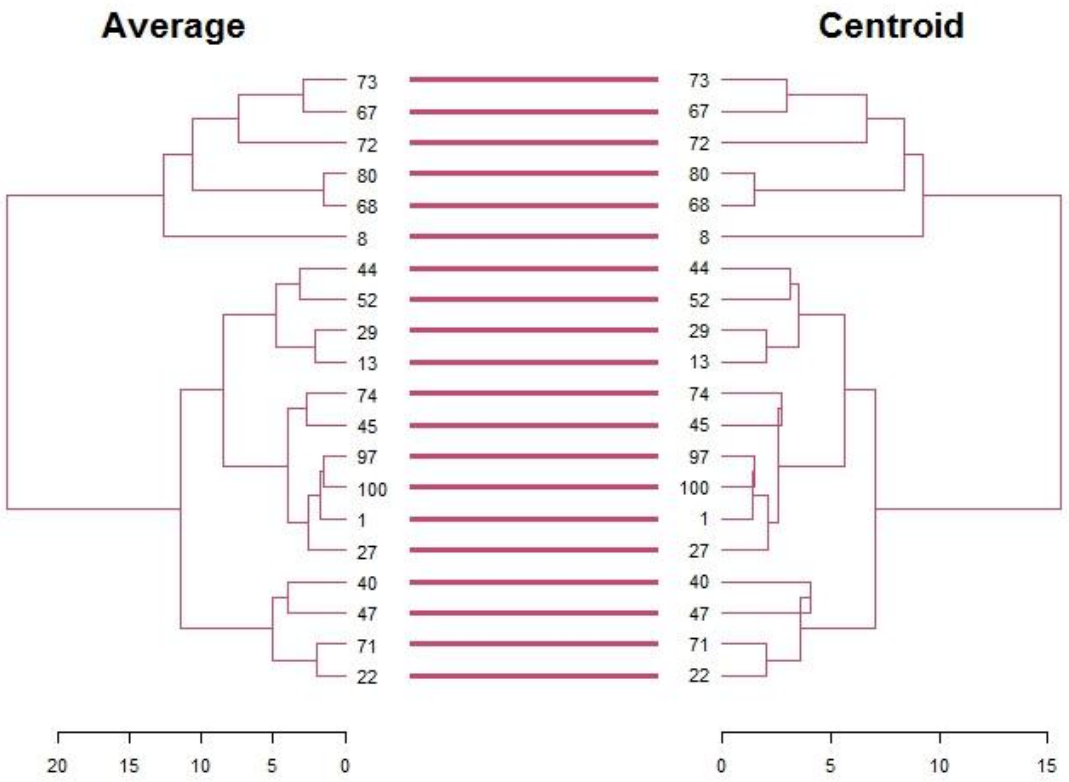


Στον ακόλουθο κώδικα παρουσιάζονται 6 διαφορετικές οπτικοποιήσεις δύο σύνδεδεμένων δεντρογραμμάτων:

```
ss<- sample(1:101,20 )
dend1 <- df.SCL90[ss,-5] %>% dist %>% hclust("com") %>% as.dendrogram
dend2 <- df.SCL90[ss,-5] %>% dist %>% hclust("single") %>% as.dendrogram
dend3 <- df.SCL90[ss,-5] %>% dist %>% hclust("ave") %>% as.dendrogram
dend4 <- df.SCL90[ss,-5] %>% dist %>% hclust("centroid") %>% as.dendrogram
dend1234 <-dendlist("Complete" = dend1, "Single" = dend2, "Average" = dend3, "Centroid" = dend4)
dend1234 %>% tanglegram(which = c(1,2), common_subtrees_color_branches = TRUE)  (1)
dend1234 %>% tanglegram(which = c(3,4), common_subtrees_color_branches = TRUE)  (2)
dend1234 %>% tanglegram(which = c(1,4), common_subtrees_color_branches = TRUE)  (3)
dend1234 %>% tanglegram(which = c(2,4),common_subtrees_color_branches = TRUE)  (4)
dend1234 %>% tanglegram(which = c(2,3), common_subtrees_color_branches = TRUE) (5)
dend1234 %>% tanglegram(which = c(3,1), common_subtrees_color_branches = TRUE)  (6)
```



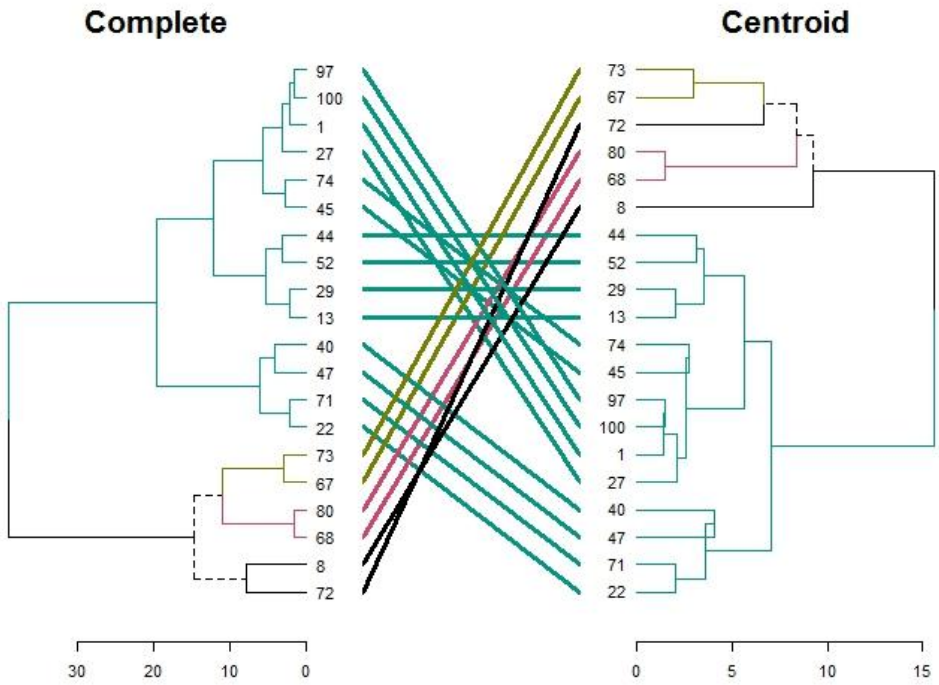
Στην παραπάνω εικόνα απεικονίζεται ένα δέντρο με μέγεθος 101 παρατηρήσεων το οποίο έχει 20 συστάδες. Στην αριστερή μεριά παρουσιάζεται ένα ολοκληρωμένο δενδρόγραμμα ενώ στα δεξιά ένα μονό δενδρόγραμμα.



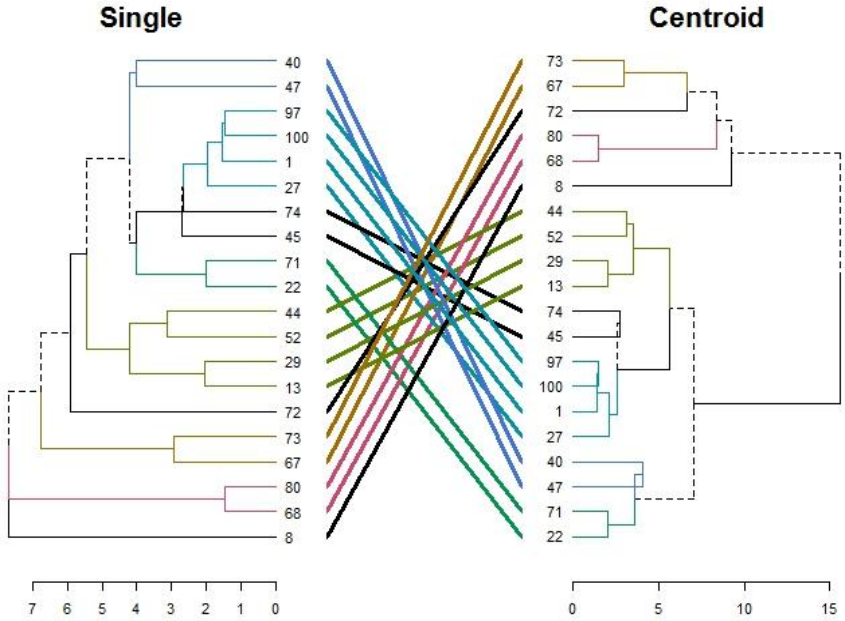
1 dend1234 %>% tanglegram(which = c(3,4), common_subtrees_color_branches = TRUE)

Στην παραπάνω εικόνα απεικονίζεται ένα δεντρόγραμμα με μέγεθος 101 παρατηρήσεων το οποίο έχει 20 συστάδες. Στην αριστερή μεριά παρουσιάζεται ένα δενδρόγραμμα τύπου μέσου όρου ενώ στα δεξιά ένα δενδρόγραμμα κεντροειδούς τύπου.

Στην εικόνα απεικονίζεται ένα δεντρόγραμμα με μέγεθος 101 παρατηρήσεων το οποίο έχει 20 συστάδες.Στην αριστερή μεριά παρουσιάζεται ένα δενδρόγραμμα τύπου ολοκληρωμένου ενώ στα δεξιά ένα δενδρόγραμμα κεντροειδούς τύπου.



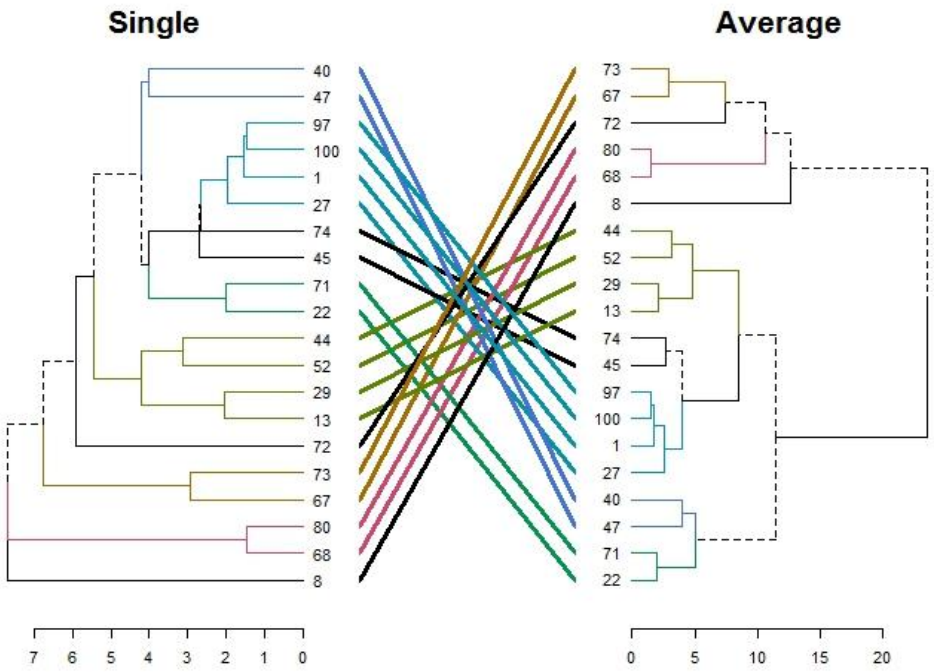
2 dend1234 %>% tanglegram(which = c(1,4), common_subtrees_color_branches = TRUE)



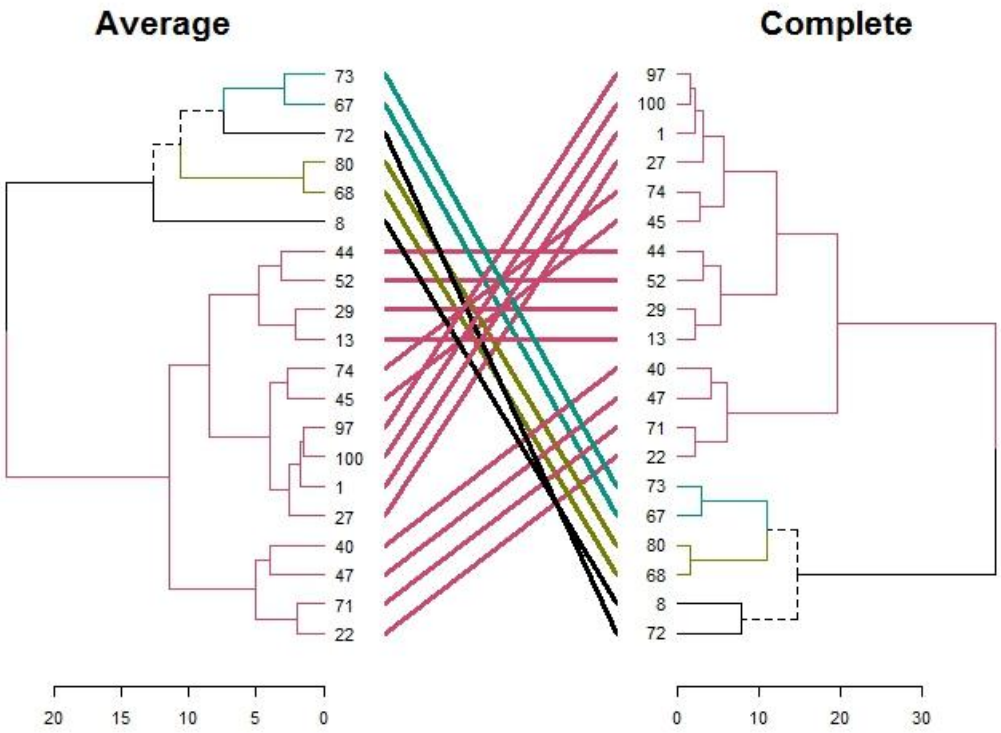
Στην εικόνα απεικονίζεται ένα δεντρόγραμμα με μέγεθος 101 παρατηρήσεων το οποίο έχει 20 συστάδες. Στην αριστερή μεριά παρουσιάζεταιένα δενδρόγραμμα μονού τύπου ενώ στα δεξιά ένα δενδρόγραμμα κεντροειδούς τύπου.

3 dend1234 %>% tanglegram(which = c(2,4), common_subtrees_color_branches = TRUE)

Στην εικόνα απεικονίζεται ένα δεντρόγραμμα με μέγεθος 101 παρατηρήσεων το οποίο έχει 20 συστάδες. Στην αριστερή μεριά παρουσιάζεται ένα δενδρόγραμμα μονού τύπου ενώ στα δεξιά ένα δενδρόγραμμα τύπου μέσου όρου.



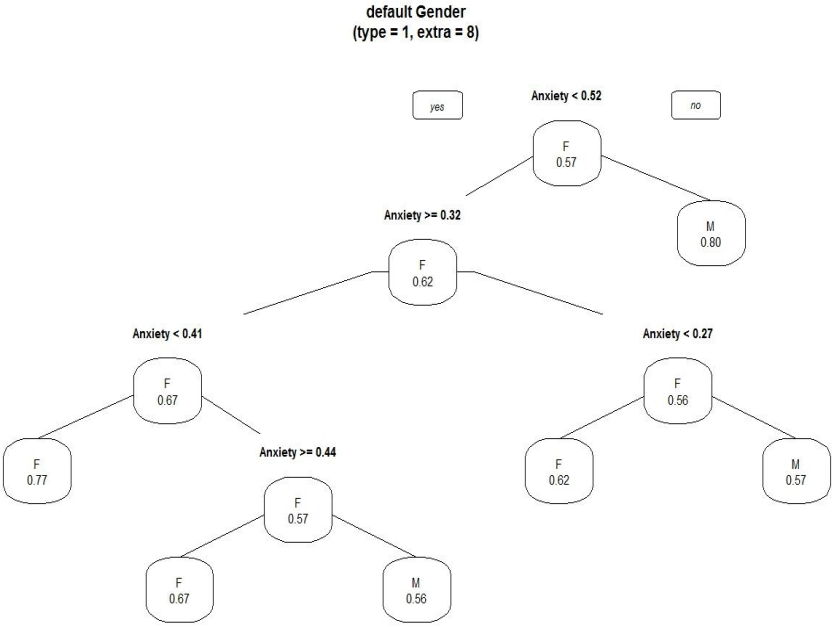

```
4 dend1234 %>% tanglegram(which = c(2,3), common_subtrees_color_branches = TRUE)
```




```
5 dend1234 %>% tanglegram(which = c(3,1), common_subtrees_color_branches = TRUE)
```

Στην εικόνα απεικονίζεται ένα δεντρόγραμμα με μέγεθος 101 παρατηρήσεων το οποίο έχει 20 συστάδες. Στην αριστερή μεριά παρουσιάζεται ένα δενδρόγραμμα τύπου μέσου όρου ενώ στα δεξιά ένα δενδρόγραμμα ολοκληρωμένου τύπου.

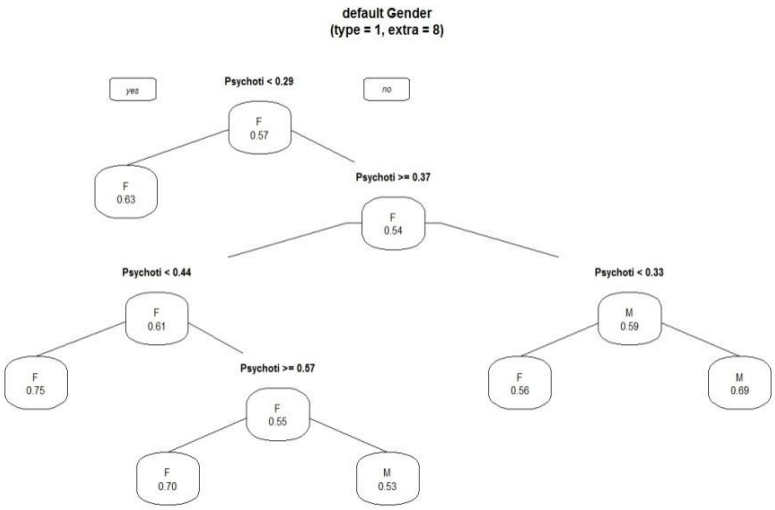
4.5.3 Διαγράμματα Δέντρων Αποφάσεων



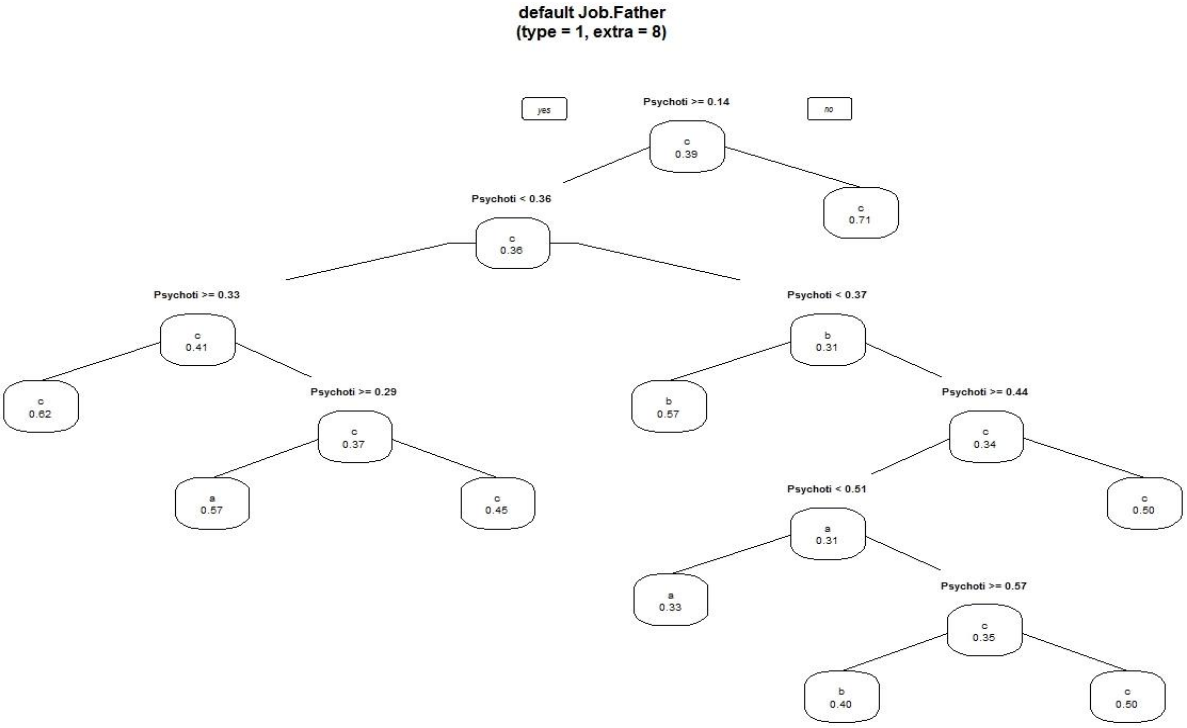
Στην εικόνα παρουσιάζεται ένα δέντρο απόφασης, το οποίο έχει διαχωρίσει τις μεταβλητές σύμφωνα με τα δεδομένα της βάσης φύλο (άντρες – γυναίκες) και το άγχος. Για την συγκεκριμένη απεικόνιση απαιτείται εγκατάσταση του πακέτου rpart.

```
Κώδικας: library(rpart)
library(rpart.plot)
data(df.SCL90)
tree<- rpart(Gender ~ Anxiety, data=df.SCL90, cp= 0.01)
prp(tree, main="default Gender\n(type = 1, extra = 8)",type = 1, extra = 8)
```

Στην εικόνα παρουσιάζεται ένα δέντρο απόφασης, το οποίο έχει διαχωρίσει τις μεταβλητές σύμφωνα με τα δεδομένα της βάσης φύλο (άντρες – γυναίκες) και τον ψυχωτισμό. Για την συγκεκριμένη απεικόνιση απαιτείται εγκατάσταση του πακέτου rpart.




```
Κώδικας: library(rpart)
library(rpart.plot)
data(df.SCL90)
tree<- rpart(Gender ~ Psychotism, data=df.SCL90, cp= .01)
prp(tree, main="default Gender\n(type = 1, extra = 8)",type = 1, extra = 8)
```

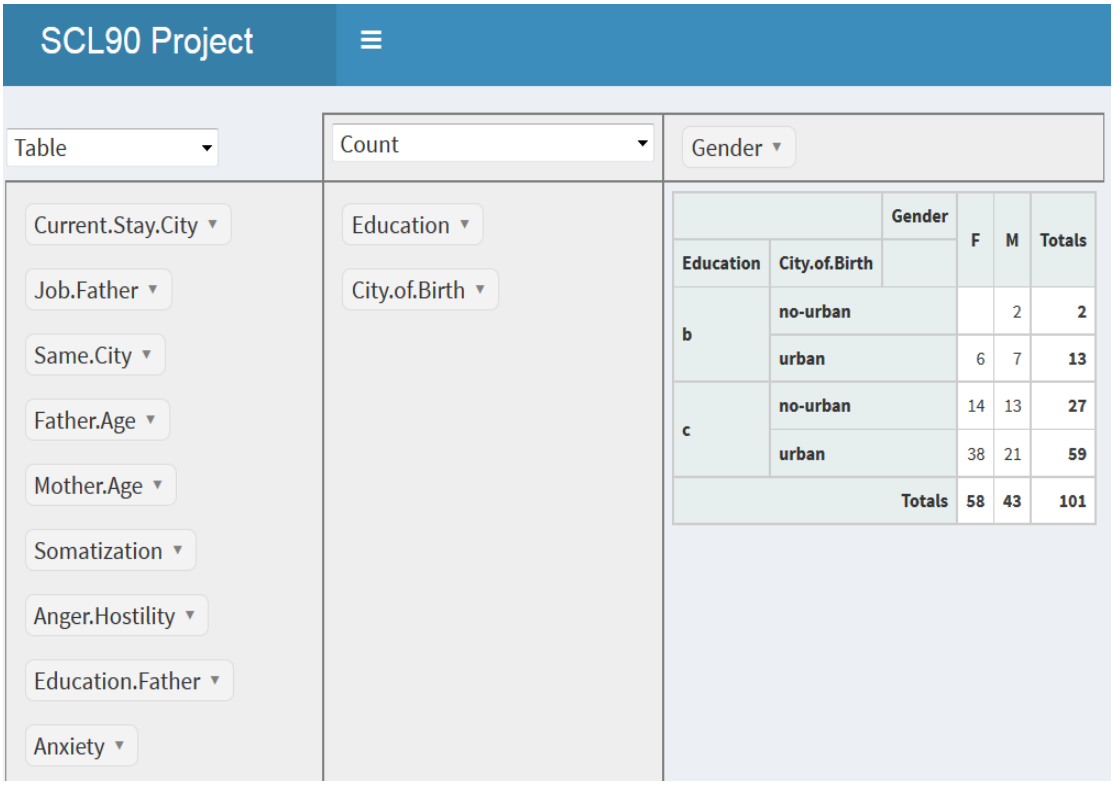


Στην παραπάνω εικόνα παρουσιάζεται ένα δέντρο απόφασης, το οποίο έχει διαχωρίσει τις μεταβλητές σύμφωνα με τα δεδομένα της βάσης το επάγγελμα του πατέρα και τον ψυχωτισμό. Για την συγκεκριμένη απεικόνιση απαιτείται εγκατάσταση του πακέτου rpart.

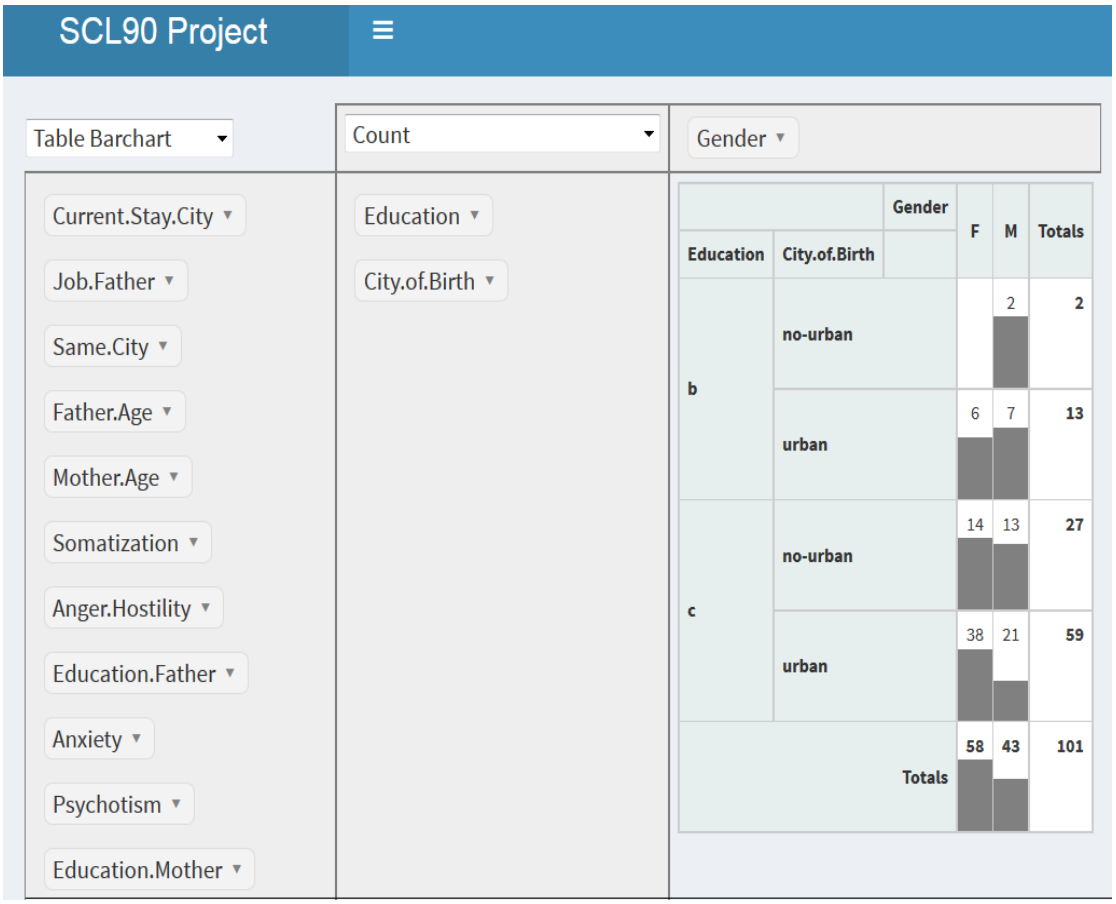
Κώδικας: library(rpart)

```
library(rpart.plot)
data(df.SCL90)
tree <- rpart(Job.Father ~ Psychotism, data=df.SCL90, cp= 0.01)
prp(tree, main="default Job.Father\n(type = 1, extra = 8)",type = 1, extra = 8)
```

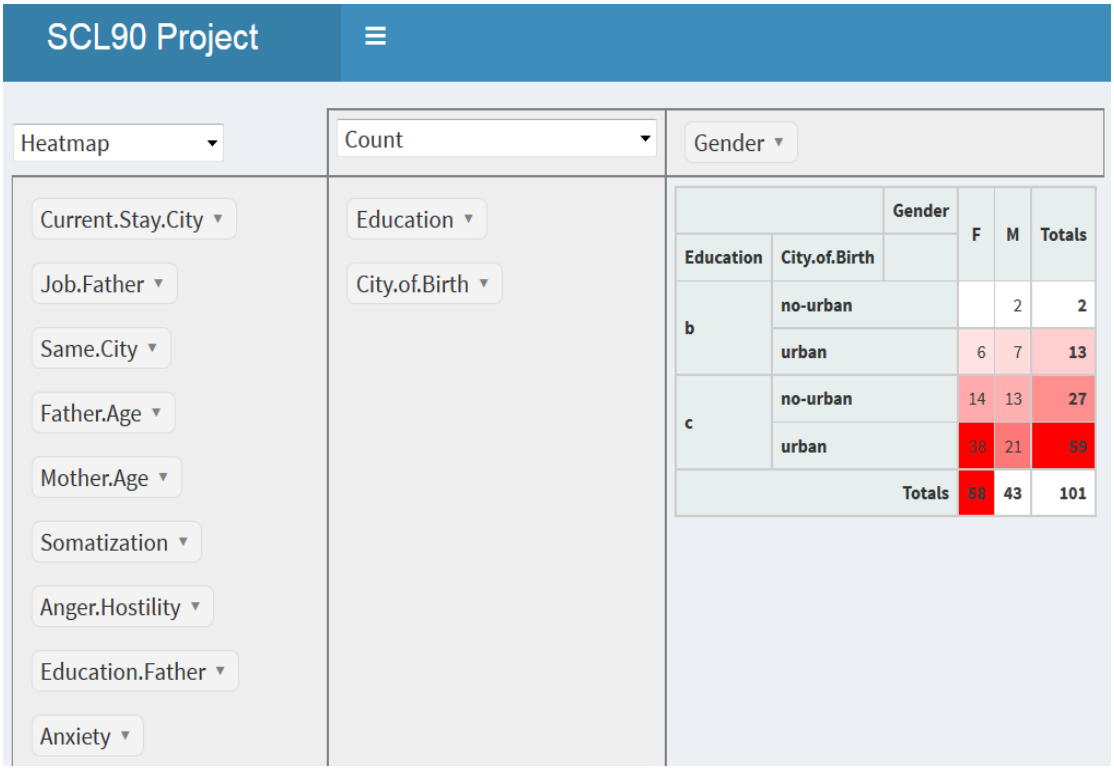
4.5.4 OLAP με Shiny



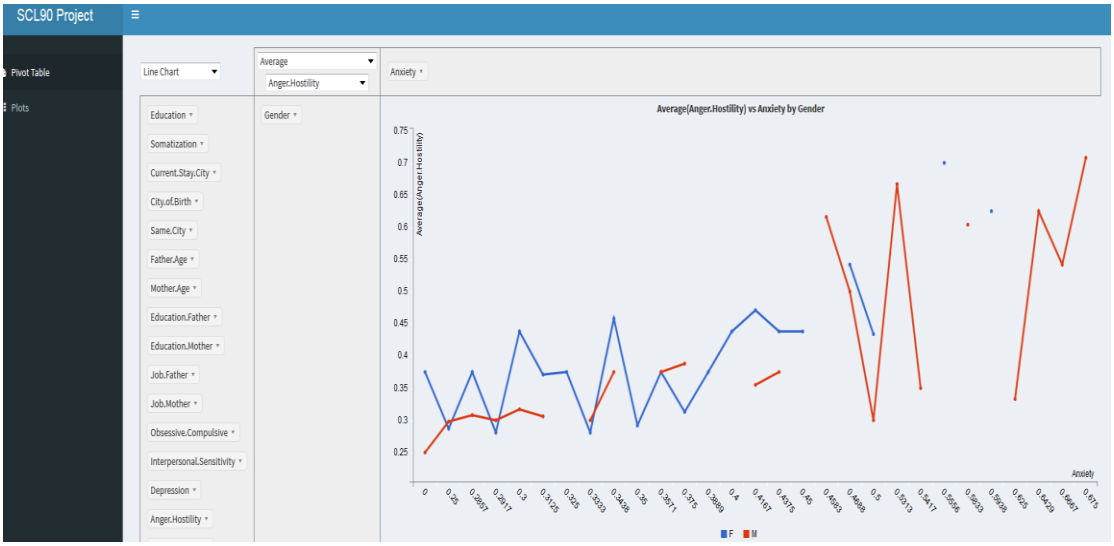
Η παραπάνω εικόνα απεικονίζει έναν συγκεντωτικό πίνακα δεδομένων στον διαδυκτιακό χώρο του Shiny. Ο συγκεκριμένος πίνακας έλαβε τις κατηγορικές μεταβλητές εκπαίδευση και πόλη στην οποία γεννήθηκαν οι ερωτηθέντες του δείγματος, σύμφωνα με την μεταβλητή φύλο.



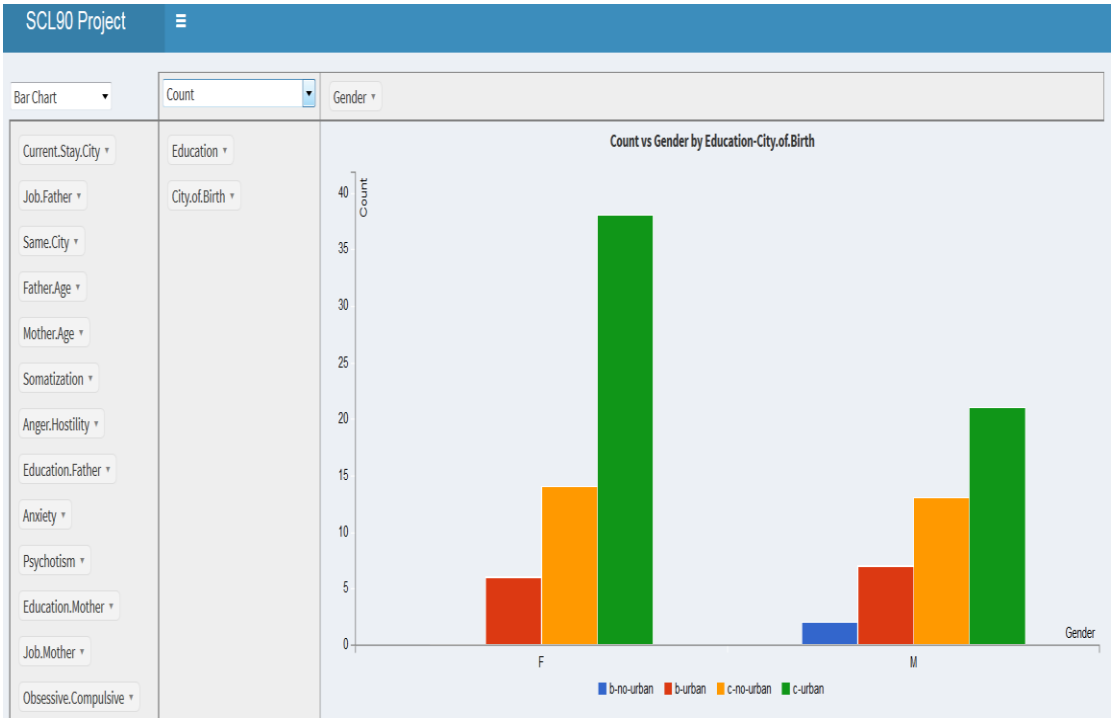
Η παραπάνω εικόνα απεικονίζει έναν συγκεντωτικό πίνακα δεδομένων στον διαδυκτιακό χώρο του Shiny. Ο συγκεκριμένος πίνακας έλαβε τις κατηγορικές μεταβλητές εκπαίδευση και πόλη στην οποία γεννήθηκαν οι ερωτηθέντες του δείγματος, σύμφωνα με την μεταβλητή φύλο. Επιπλέον η εφαρμογή δίνει την δυνατότητα μέσω της εντολής Table Barchar να εμφανίζει ραβδογράμματα συγκέντρωσης των δεδομένων εντός του πίνακα.



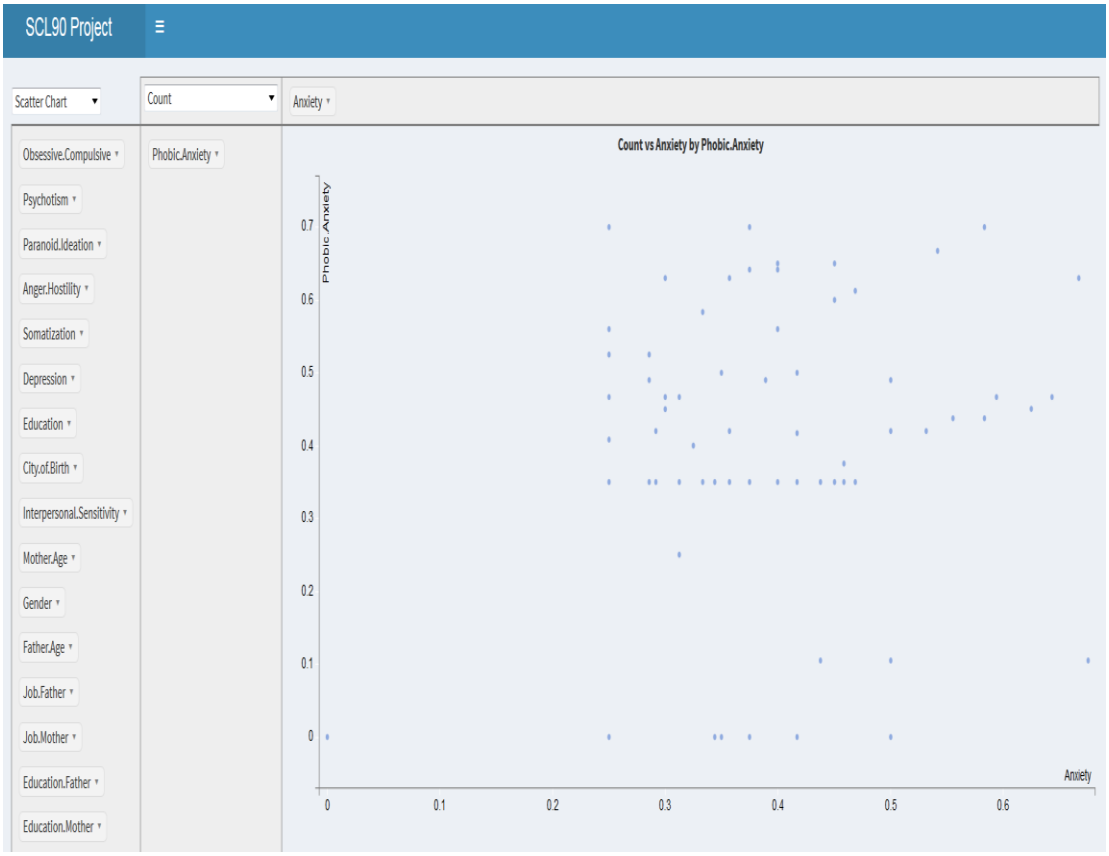
Η παραπάνω εικόνα απεικονίζει έναν συγκεντωτικό πίνακα δεδομένων στον διαδυκτιακό χώρο του Shiny. Ο συγκεκριμένος πίνακας έλαβε τις κατηγορικές μεταβλητές εκπαίδευση και πόλη στην οποία γεννήθηκαν οι ερωτηθέντες του δείγματος, σύμφωνα με την μεταβλητή φύλο. Επιπλέον η εφαρμογή δίνει την δυνατότητα μέσω της εντολής Heatmap να εμφανίζει τις περιοχές οι οποίες έχουν μεγαλύτερη συγκέντρωση δεδομένων εντός του πίνακα με πιο έντονα χρώματα.



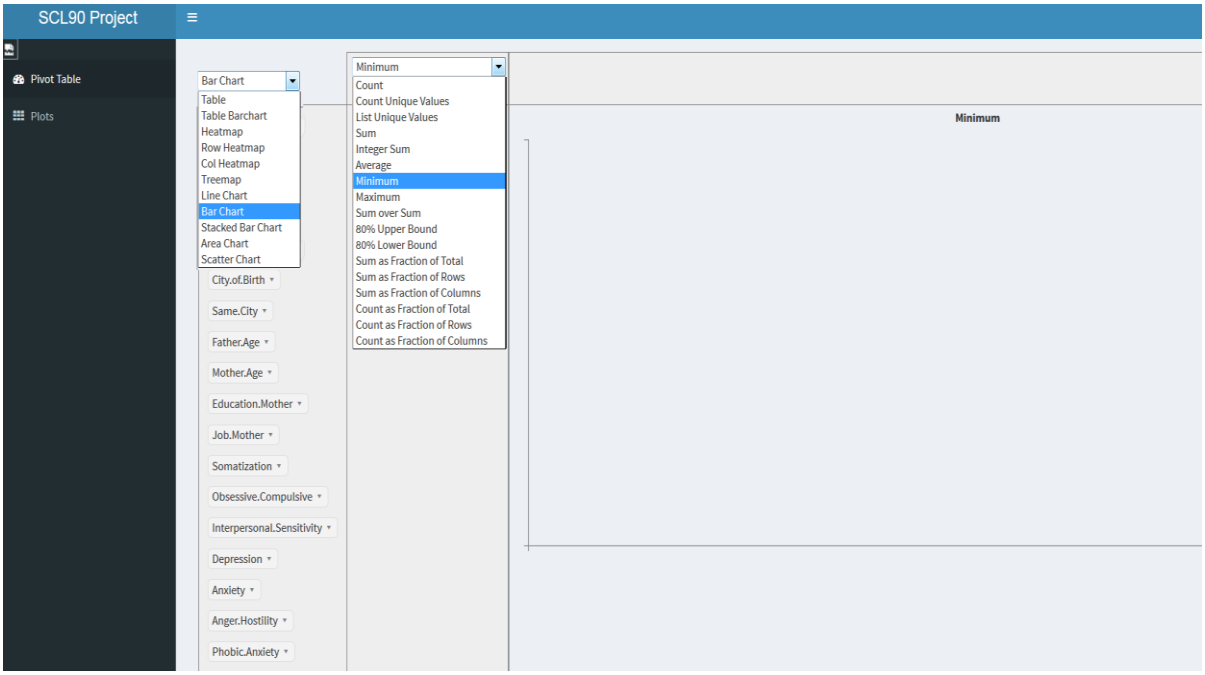
Η παραπάνω εικόνα απεικονίζει έναν γράφημα γραμμών. Το συγκεκριμένο γράφημα γραμμών έλαβε την κατηγορική μεταβλητή φύλο και την ποσοτική μεταβλητή άγχος. Διαλέγοντας από την λίστα τον τύπο μέσο όρο του βάση τησ μεταβλητής θυμός-επιθετικότητα σχεδιάστικε το παραπάνω γράφημα. Η εντολή Line Chart δίνει την δυνατότητα να εμφανίζει γραμμές συγκέντρωσης δεδομένων.



Η παραπάνω εικόνα απεικονίζει έναν συγκεντωτικό πίνακα δεδομένων στον διαδυκτιακό χώρο του Shiny. Ο συγκεκριμένος πίνακας έλαβε τις κατηγορικές μεταβλητές εκπαίδευση και πόλη στην οποία γεννήθηκαν οι ερωτηθέντες του δείγματος, σύμφωνα με την μεταβλητή φύλο. Η εντολή Bar Chart δίνει την δυνατότητα να εμφανίζει ραβδογράμματα με διαφορετικά χρώματα για το κάθε δεδομένο.



Η παραπάνω εικόνα απεικονίζει ένα γράφημα διασποράς στον διαδυκτιακό χώρο του Shiny λαμβάνοντας τις αριθμητικές μεταβλητές άγχος και φοβικό άγχος των ερωτηθέντων του δείγματος. Η εντολή Scatter Chart δίνει την δυνατότητα να εμφανίζει τα σημεία συγκέντρωσης των δεδομένων.



Σύμφωνα με τον κώδικα που δημιουργήθηκε (Παράρτημα Β) η εφαρμογή Shiny δίνει την δυνατότητα στο χρήστη να επεξεργαστεί τα δεδομένα από τις λίστες εντολών και να οπτικοποιήσει τις μεταβλητές που χρειάζεται, με τις παραμέτρους που του εμφανίζει η λίστα στο δεξί μέρος.

Συμπεράσματα

Το αντικείμενο της εξόρυξης δεδομένων, είναι σχετικά νέο σε σχέση με την στατιστική. Η εξόρυξη δεδομένων είναι αναπόσπαστο κομμάτι της ανακάλυψης γνώσης από τις βάσεις δεδομένων, και μετασχηματίζει τα ακατέργαστα δεδομένα σε σημαντικές πληροφορίες ώστε να γίνουν κατανοητές από τον άνθρωπο (Ning Tan, Steinbach, & Kumar, 2010). Μια από τις μεθοδολογίες ανάλυσης των αποτελεσμάτων της εξόρυξης είναι και η διαδικασία της οπτικοποίησης. Η οπτικοποίηση δίνει την δυνατότητα στην διαδικασία της εξόρυξης δεδομένων να παρέχει μια απλή και συνοπτική απεικόνιση των πληροφοριών που αποθηκεύονται σε ένα μεγάλο σύνολο δεδομένων. Οι πιο συνηθισμένες τεχνικές οπτικοποίησης είναι τα γραφήματα, οι πίνακες, τα διαγράμματα, τα δέντρα αποφάσεων, τα δενδρογράμματα και η OLAP. Ωστόσο με τις τεχνικές οπτικοποίησης όπως για παράδειγμα τα δεντρογράμματα και τα δέντρα αποφάσεων καθιστάτε δυνατή η απεικόνιση και η οπτική ερμηνεία των αποτελεσμάτων των διαδικασιών της εξόρυξης δεδομένων, δηλαδή της συσταδοποίησης και της κατηγοριοποίησης.

Τέλος στην παρούσα εργασία εφαρμόστηκε η στατιστική γλώσσα προγραμματισμού R, η οποία ανέλυσε δεδομένα και τεχνικές οπτικοποίησης δεδομένων πολλών διαστάσεων, με αποτέλεσμα να προσφέρει εντυπωσιακές γραφικές αναπαραστάσεις οι οποίες γίνονται κατανοητές όχι μόνο στον ερευνητή αλλά και σε κάθε άνθρωπο.

Βιβλιογραφία

1.Berry, M., & Linoff, G. (2004). *Data Mining Techiques For Marketing, Sales, and Customer Relationship Management* (2η Έκδοση εκδ.). Indianapolis, Indiana: Wiley Publishing, Inc.

Carlo, V. (2009). *Business intelligence: Data Mining and Optimization for Decision Making* (1η Έκδοση εκδ.). (S. Ltd, Επιμ.) Milano, Italy: John Wiley.

Diane, C. (1999). *Data Preparation for Data Mining.* (E. Wade, Επιμ.) San Francisco, 340 Pine Street, Sixth Floor, ΗΠΑ: Morgan Kaufmann.

Han, J., & Kameber, M. (2006). *Data Mining: Concepts and Techniques, Second Edition* (2η Έκδοση εκδ.). Amsterdam, Boston, London: Diane Cerra.

John, M., & John, B. (2003). *Data Analysis and Graphics Using R* (3η Έκδοση εκδ.). New York, America: Cambridge University Press.

Maimon, O., & Rokach, L. (2010). *Data Mining And Knowledge Discovery Handbook* (2η Έκδοση εκδ.). Ramat, Israel: Springer.

Michael, B., & Gordon, L. (2004). *Data Mining Techiques For Marketing, Sales, and Customer Relationship Management* (2η Έκδοση εκδ.). Indianapolis, Indiana: Wiley Publishing, Inc.

Ning Tan, P., Steinbach, M., & Kumar, V. (2010). *Εισαγωγή στην εξόρυξη δεδομένων.* (Β. Βερύκιος, Επιμ., & Σ. Σουραβλάς, Μεταφρ.) Πανεπιστημίου 39, Αθήνα, Ελλάδα: Τζίολα.

Soukup, T., & Davidson, I. (2002). *Visual Data MIning: Techniques and Tools for Data Visualization and Mining.* (R. Elliott, Επιμ.) New York, Third Avenue, America: Wiley & Sons, John.

Tom, S., & Ian, D. (2002). *Visual Data MIning: Techniques and Tools for Data Visualization and Mining.* (R. Elliott, Επιμ.) New York, Third Avenue, America: Wiley & Sons, John.

Usama, F. (1996). *Advances in Knowledge Discovery and Data Mining.* United Kingdom, Duchess Street, London: Gregory Piatetsky-Shapiro.

Vercellis, C. (2009). *Business intelligence: Data Mining and Optimization for Decision Making* (1η Έκδοση εκδ.). (S. Ltd, Επιμ.) Milano, Italy: John Wiley.

Wickham, H. (2009). *ggplot2 Elegant Graphics for Data Analysis.* Houston, USA: Springer.

Winston, W. (2011). *Microsoft® Excel® 2010:Data Analysis and.* (R. Caperton, Επιμ.) Washington, Redmond, ΗΠΑ: Microsoft Press.

Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies.* Elsevier: Academic Press.

Αποστόλου, Ι. (2008). *ΕΠΕΞΕΡΓΑΣΙΑ ΣΕΙΣΜΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ.* Θεσσαλονίκη.

Γκίζα, Ε. (2007). *Χρήση της OLAP Τεχνικής στην Οπτικοποίηση Κανόνων Data Mining.* Πάτρα.

Ζαφειροπούλου, Φ. (2007). *Μέθοδοι Εύρεσης Βέλτιστου Πλήθους Ομάδων για Πολυδιάστατα Δεδομένα.* Πειραιάς.

Κιτικίδου, Κ. *Εισαγωγή στην Παλινδρόμηση "http://www.fmenr.duth.gr/labwebpages/dasikiviometria/regression.pdf".*

Κουμπάρου, Ν. (2010). *ΑΞΙΟΛΟΓΗΣΗ ΚΑΝΟΝΩΝ ΒΑΣΕΙ ΠΟΛΛΑΠΛΩΝ ΜΕΤΡΩΝ.* Κύπρος.

Τόλιας, Γ. (2008). *Μη γραμμική παλινδρόμηση.* Πάτρα.

Παράρτημα Α

```
library(readxl)

pathname = paste(getwd(), "SCL90.xlsx", sep="/")

df.SCL90 <- read_excel(pathname, sheet = 1 ,col_names = TRUE, na='na')

View(df.SCL90)

ds <- df.SCL90[,12:20]

ds <- df.SCL90[,1:20]

library(ggplot2)

ds_gender<- cbind( 'Gender'=df.SCL90$Gender, ds )

pie<- ggplot(ds_gender, aes(x=factor(1), fill = Gender)) + geom_bar(width = 1)

pie + coord_polar(theta = "y")

library (plotrix)

counts<- table(df.SCL90$Job.Father)

abcd <-c("Δημόσιος Υπάλληλος", "Ιδιωτικός Υπάλληλος", "Ελεύθερος Επαγγελματίας", "Άλλο")

lbls<- paste(abcd,names(counts), "\n", counts)

pie3D(counts, labels = lbls, explode=0.1, main="3D Pie from Job.Father\n ", col=c("#dd00dd","blue","red","brown"))

ds<- f.SCL90[,12:20]

boxplot(x=ds, horizontal = FALSE, notch = TRUE)

box_som_gender<- boxplot(Somatization~Gender, data = df.SCL90, horizontal = FALSE, notch = TRUE, col = "blue",  main="Boxplot Somatization by Gender")

thikogrammabox_som_gender<- boxplot (Depression~Gender, data = df.SCL90, horizontal = FALSE, notch = TRUE, col = "red", xlab="Φύλο", ylab="Κατάθλιψη", main= "Boxplot Depression by Gender")

ds.M <- df.SCL90[ df.SCL90$Gender=='M', 12:20]
```

```
par(mar=c(3,9,3,7))
boxplot(x=ds.M, horizontal = TRUE, notch = TRUE, las=1,
main="Boxplot 9 Κλίμακες ψυχωπαθολογίας για τους Άνδρες", col = "light blue")
ds.F<- df.SCL90[ df.SCL90$Gender=='F', 12:20]
par(mar=c(3,9,3,7))
boxplot(x=ds.F, horizontal = TRUE, notch = TRUE, las=1,
    main="Boxplot 9 Κλίμακες ψυχωπαθολογίας για τις Γυναίκες",col = "pink")
counts<- table( df.SCL90$Job.Father, df.SCL90$Education.Father)
barplot(counts, main=" SCL-90 Distribution by Job.Father and Education.Father",
    xlim = c(1,18),
    xlab="Education.Father",
    ylab="Job.Father",
    col=c("violetred2","aquamarine3","thistle", "lightblue"),
legend =c("Δημόσιος Υπάλληλος", "Ιδιωτικός Υπάλληλος", "Ελεύθερος Επαγγελματίας", "Άλλο"),
    args.legend = list(x ="topleft",inset=c(0.8,0)),
    names.arg = c("πρωτοβάθμια εκπαίδευση", "δευτεροβάθμια εκπαίδευση", "τριτοβάθμια εκπαίδευση"), axisnames=TRUE, beside=TRUE )
counts<- table(df.SCL90$Gender, df.SCL90$Education.Father)
barplot(counts, main=" SCL-90 Distribution by Gender and Education.Father",
    xlab=" Education.Father ", col=c("gold","darkred"),
    legend = rownames(counts), beside=TRUE)
d <- density(df.SCL90$Interpersonal.Sensitivity)
plot(d, main="df.SCL90 of Interpersonal.Sensitivity")
polygon(d, col="lightpink", border="darkslateblue",lwd=4)
d <- density(df.SCL90$Depression)
plot(d, main="df.SCL90 of Depression")
```

```
polygon(d, col="cornflowerblue", border="deeppink4",lwd=4)
x <- df.SCL90$Anger.Hostility
h<-hist(x, breaks=10, col="red", xlab="Anger.Hostility",
      main="Histogram with Non-Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit<- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=3)
library(e1071)
skewness(x)
kurtosis(x)
x <- df.SCL90$Somatization
h<-hist(x, breaks=10, col="yellow", xlab="Somatization",
      main="Histogram with Non-Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit<- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=3)
library(e1071)
skewness(x)
kurtosis(x)
library(ggplot2)
ds_gender<-cbind( 'Gender' = df.SCL90$Gender, ds )
ggplot(data=ds_gender, aes(x=Anxiety, y=Somatization, color=Gender)) +
geom_point()
library(ggplot2)
```

```
ds_gender<- cbind( 'Gender'=df.SCL90$Gender, ds )
ggplot(data=ds_gender,aes(x=Psychotism, y=Depression, color=Gender)) + geom_point()
counts<- table( df.SCL90$Job.Father, df.SCL90$Education.Father)
barplot(counts, main=" SCL-90 Distribution by Job.Father and Education.Father",
    xlim = c(1,18),
    xlab="Education.Father",
    ylab="Job.Father",
    col=c("violetred2","aquamarine3","thistle", "lightblue"),
    legend =c("Δημόσιος Υπάλληλος", "Ιδιωτικός Υπάλληλος", "Ελεύθερος Επαγγελματίας", "Άλλο"),
    args.legend = list(x ="topleft",inset=c(0.8,0)),
    names.arg = c("πρωτοβάθμια εκπαίδευση", "δευτεροβάθμια εκπαίδευση", "τριτοβάθμια εκπαίδευση"), axisnames=TRUE, beside=TRUE )
counts<- table(df.SCL90$Gender, df.SCL90$Education.Father)
barplot(counts, main=" SCL-90 Distribution by Gender and Education.Father",
    xlab=" Education.Father ", col=c("gold","darkred"),
    legend = rownames(counts), beside=TRUE)
d <- density(df.SCL90$Interpersonal.Sensitivity)
plot(d, main="df.SCL90 of Interpersonal.Sensitivity")
polygon(d, col="lightpink", border="darkslateblue",lwd=4)
d <- density(df.SCL90$Depression)
plot(d, main="df.SCL90 of Depression")
polygon(d, col="cornflowerblue", border="deeppink4",lwd=4)
x <- df.SCL90$Anger.Hostility
h<-hist(x, breaks=10, col="red", xlab="Anger.Hostility",
    main="Histogram with Non-Normal Curve")
```

```
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit<- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=3)
library(e1071)
skewness(x)
kurtosis(x)
x <- df.SCL90$Somatization
h<-hist(x, breaks=10, col="yellow", xlab="Somatization",
      main="Histogram with Non-Normal Curve")
xfit<-seq(min(x),max(x),length=40)
yfit<-dnorm(xfit,mean=mean(x),sd=sd(x))
yfit<- yfit*diff(h$mids[1:2])*length(x)
lines(xfit, yfit, col="blue", lwd=3)
library(e1071)
skewness(x)
kurtosis(x)
library(ggplot2)
ds_gender<-cbind( 'Gender' = df.SCL90$Gender, ds )
ggplot(data=ds_gender, aes(x=Anxiety, y=Somatization, color=Gender)) +
geom_point()
library(ggplot2)
ds_gender<- cbind( 'Gender'=df.SCL90$Gender, ds )
ggplot(data=ds_gender,aes(x=Psychotism, y=Depression, color=Gender)) +
geom_point()
library(scatterplot3d)
attach(df.SCL90)
```

```
scatterplot3d(Somatization,Depression,Anxiety, main="3D Scatterplot")

library(scatterplot3d)

attach(df.SCL90)

scatterplot3d(Anger.Hostility,Phobic.Anxiety,Paranoid.Ideation, main="3D
Scatterplot")

library(scatterplot3d)

attach(df.SCL90)

s3d <-scatterplot3d(Anxiety,Somatization,Depression, pch=16, highlight.3d=TRUE,

        type="h", main="3D Scatterplot")

fit<- lm(Depression ~ Anxiety+Somatization)

s3d$plane3d(fit)

library(rgl)

attach(df.SCL90)

plot3d(Depression,Somatization,Psychotism, col="red", size=9)

library(Rcmdr)

attach(df.SCL90)

scatter3d(Depression,Somatization,Psychotism)

library(dendextend)

library(dendextendRcpp)

data(df.SCL90)

df.SCL90_data <-(data=df.SCL90)

df.SCL90_data <- df.SCL90(c["Anxienty", "Somatization", "Psychotism",
"Obsessive.Compulsive", "Interpersonal.Sensitivity","Depression", "Anger.Hostility",
"Phobic.Anxiety", "Paranoid.Ideation" ])

x_dist<- dist(df.SCL90, diag = TRUE)

hc1 <- hclust(x_dist, method = "complete" )

plot(hc1,hang = -1)
```

```
dend1 <- as.dendrogram(hc1)

str(hc1)

str(dend1)

str(unclass(dend1))

d <- dist(as.matrix(df.SCL90))

hc<- hclust(d)

plot(hc)

library(dendextend)

library(dendextendRcpp)

data(df.SCL90)

df.SCL90_data <- (data=df.SCL90)

df.SCL90_data <- df.SCL90(c["Anxienty", "Somatization", "Psychotism", "Obsessive.Compulsive", "Interpersonal.Sensitivity", "Depression" , "Anger.Hostility", "Phobic.Anxiety", "Paranoid.Ideation" ])

x_dist<- dist(df.SCL90, diag = TRUE)

hc1 <- hclust(x_dist, method = "single" )

plot(hc1,hang = -1)

dend1 <- as.dendrogram(hc1)

str(hc1)

str(dend1)

str(unclass(dend1))

require(dendextend)

dend1_mod_01 <- dend1

dend1_mod_01 <- color_branches(dend1_mod_01,k = 20)

col_for_labels<- c ("purple","red","green","blue","yellow")

dend1_mod_01 <- col_for_labels(dend1_mod_01,col = col_for_labels )

par(mfrow = c (1,2))
```

```
plot(dend1)

plot(dend1_mod_01)

require(dendextend)

dend1_mod_01 <- dend1

dend1_mod_01 <- color_branches(dend1_mod_01,k = 9)

col_for_labels<- c ("purple","red","green","blue","yellow")

dend1_mod_01 <- col_for_labels(dend1_mod_01,col = col_for_labels )

par(mfrow = c (1,2))

plot(dend1)

plot(dend1_mod_01)

library(ape)

hc = hclust(dist(df.SCL90))

plot(hc)

plot(as.phylo(hc), cex = 0.9, label.offset = 1)

plot(as.phylo(hc), type = "fan", tip.color = hsv(runif(15, 0.65,0.95), 1, 1, 0.7), edge.color = hsv(runif(10, 0.65, 0.75), 1, 1, 0.7), edge.width = runif(20,0.5, 3),use.edge.length = TRUE, col = "gray80")

library(ape)

hc=hclust(dist(df.SCL90$Somatization,df.SCL90$Psychotism,df.SCL90$Depression))

plot(hc)

plot(as.phylo(hc), type = "fan", tip.color = hsv(runif(15, 0.65,0.95), 1, 1, 0.7), edge.color = hsv(runif(10, 0.65, 0.75), 1, 1, 0.7), edge.width = runif(20,0.5,3),

    use.edge.length = TRUE, col = "gray80"

    ds <- df.SCL90[,1:20]

    library(dendextend)

    library(dendextendRcpp)

    dend<- df.SCL90[,1:20] %>% dist %>% hclust %>% as.dendrogram
```

```
dend %>% color_branches(k=5) %>% plot(horiz=TRUE, main = "Διαιρετικοί Ιεραρχικοί Μέθοδοι \n (Divisive Hierarchical Methods)")
dend %>% rect.dendrogram(k=5,horiz=TRUE)
abline(v = heights_per_k.dendrogram(dend)["5"] + .6, lwd = 5, lty = 5, col = "blue")
dend<- df.SCL90[,1:20] %>% dist %>% hclust %>% as.dendrogram
dend %>% color_branches(k=5) %>% plot(horiz=FALSE, main = "ΣυσσωρευτικοίΙεραρχικοί Μέθοδοι \n (Agglomerative  Hierarchical Methods)")
dend %>% rect.dendrogram(k=5,horiz=FALSE)
abline(v = heights_per_k.dendrogram(dend)["5"] + .6, lwd = 5, lty = 5, col = "blue")
dend<- df.SCL90[1:101,-5] %>% dist %>% hclust %>% as.dendrogram %>%
  set("branches_k_color", k=3) %>% set("branches_lwd", c(1.5,1,1.5)) %>%
  set("branches_lty", c(1,1,3,1,1,2)) %>%
  set("labels_colors") %>% set("labels_cex", c(.9,1.2)) %>%
  set("nodes_pch", 19) %>% set("nodes_col", c("orange", "black", "plum", NA))
ggd1 <- as.ggdend(dend)
library(ggplot2)
ggplot(ggd1)
dend<- df.SCL90[1:101,-5] %>% dist %>% hclust %>% as.dendrogram %>%
  set("branches_k_color", k=3) %>% set("branches_lwd", c(1.5,1,1.5)) %>%
  set("branches_lty", c(1,1,3,1,1,2)) %>%
  set("labels_colors") %>% set("labels_cex", c(.9,1.2)) %>%
  set("nodes_pch", 19) %>% set("nodes_col", c("orange", "black", "plum", NA))
ggd1 <- as.ggdend(dend)
library(ggplot2)
ggplot(ggd1, horiz = TRUE, theme = NULL)
dend<- df.SCL90[1:101,-5] %>% dist %>% hclust %>% as.dendrogram %>%
  set("branches_k_color", k=3) %>% set("branches_lwd", c(1.5,1,1.5)) %>%
```

```
    set("branches_lty", c(1,1,3,1,1,2)) %>%
    set("labels_colors") %>% set("labels_cex", c(.9,1.2)) %>%
    set("nodes_pch", 19) %>% set("nodes_col", c("orange", "black", "plum", NA))
  ggd1 <- as.ggdend(dend)
  library(ggplot2)
  ggplot(ggd1, labels = FALSE) + scale_y_reverse(expand = c(0.2, 0)) + coord_polar(theta="x")
  dend15 <- c(1:101) %>% dist %>% hclust(method = "complete") %>% as.dendrogram
  dend15 <- dend15 %>% set("labels_to_char")
  dend51 <- dend15 %>% set("labels", as.character(101:1)) %>% match_order_by_labels(dend15)
  dends_15_51 <- dendlist(dend15, dend51)
  tanglegram(dends_15_51, common_subtrees_color_branches = TRUE)
  x <- dends_15_51 %>% untangle(method = "ladderize")
  x %>% plot(main = paste("Περιπλοκότητα =", round(entanglement(x), 3)))
  ss<- sample(1:101,20 )
  dend1 <- df.SCL90[ss,-5] %>% dist %>% hclust("com") %>% as.dendrogram
  dend2 <- df.SCL90[ss,-5] %>% dist %>% hclust("single") %>% as.dendrogram
  dend3 <- df.SCL90[ss,-5] %>% dist %>% hclust("ave") %>% as.dendrogram
  dend4 <- df.SCL90[ss,-5] %>% dist %>% hclust("centroid") %>% as.dendrogram
  dend1234 <-dendlist("Complete" = dend1, "Single" = dend2, "Average" = dend3, "Centroid" = dend4)
dend1234 %>% tanglegram(which = c(1,2), common_subtrees_color_branches = TRUE)
dend1234 %>% tanglegram(which = c(3,4), common_subtrees_color_branches = TRUE)
dend1234 %>% tanglegram(which = c(1,4), common_subtrees_color_branches = TRUE)
dend1234 %>% tanglegram(which = c(2,4),common_subtrees_color_branches = TRUE)
```

```
dend1234 %>% tanglegram(which = c(2,3), common_subtrees_color_branches = TRUE)
dend1234 %>% tanglegram(which = c(3,1), common_subtrees_color_branches = TRUE)
    library(rpart)
    library(rpart.plot)
    data(df.SCL90)
    tree<- rpart(Gender ~ Anxiety, data=df.SCL90, cp= 0.01)
    prp(tree, main="default Gender\n(type = 1, extra = 8)",type = 1, extra = 8)
    library(rpart)
    library(rpart.plot)
    data(df.SCL90)
    tree<- rpart(Gender ~ Psychotism, data=df.SCL90, cp= .01)
    prp(tree, main="default Gender\n(type = 1, extra = 8)",type = 1, extra = 8)
    library(rpart)
    library(rpart.plot)
    data(df.SCL90)
    tree <- rpart(Job.Father ~ Psychotism, data=df.SCL90, cp= 0.01)
    prp(tree, main="default Job.Father\n(type = 1, extra = 8)",type = 1, extra = 8)



    devtools::install_github('rstudio/shinyapps')
    devtools::install_github(c("ramnathv/htmlwidgets",
"smartinsightsfromdata/rpivotTable"))
```

Παράρτημα Β

 # Shiny App for SCL90 project(σε νέα καρτέλα)

```
library(shiny)
library(shinydashboard)
library(shinyFiles)
source('helpers.R')
source('dashboardHeader.R')
source('dashboardSidebar.R')
source('dashboardBody.R')
dashboardPage(header,sidebar,body)
```

 # ML app for Scoring Startups... (σε νέα καρτέλα)

```
library(shiny)
shinyServer(function(input, output, session) {
  output$pivTable = renderRpivotTable({
    rpivotTable(df.SCL90) })
  })
```

helpers.R (σε νέα καρτέλα)

```
library(ggplot2)
library(dygraphs)
library(rpivotTable)
library(readxl)
pathname = paste(getwd(), "SCL90.xlsx", sep="/")
df.SCL90 <- read_excel(pathname, sheet = 1 ,col_names = TRUE, na='na')
```

dashboardSidebar.R(σε νέα καρτέλα)

```
sidebar <- dashboardSidebar(
  img(src='logo.jpg',class ='img-responsive')
,sidebarMenu(
    menuItem("Pivot Table", tabName = "pivotTable", icon = icon("dashboard"))  ,
    menuItem("Plots", icon = icon("th"), tabName = "plots")
  )
)
```

dashboardHeader.R(σε νέα καρτέλα)

```
header <- dashboardHeader(title = "SCL90 Project")
```

dashboardBody.R(σε νέα καρτέλα)

```
body <- dashboardBody(
  tabItems(
    tabItem(tabName = 'pivotTable',
              fluidPage(
                 rpivotTableOutput('pivTable')
              ) # end fluid page
           ) # end tabItem
    ,tabItem(tabName = 'plots', h2("Create Plots for Analysis....")
           ) # end tabItem
  )#end tabItems
)
#            , fluidRow(
#              column(width = 9,
#              box(title ='Bass Diffusion Model'
#              ,width = NULL
```

```
#             ,solidHeader = TRUE
#             ,collapsible = TRUE
#             ,dygraphOutput("dygraph", height = 500)
#              ) # end box
#             ) # end column
#             , column(width = 3,
#                   box( title = 'Criteria'
#                       ,status = "warning"
#                       ,width = NULL
#                       ,solidHeader = TRUE
#                       ,collapsible = TRUE
#                        ,sliderInput('marketSize','Market Size:', min = 500,max = 500000,
value = 10000, step = 500, sep = '.')
#                        ,sliderInput('pSize',' P-Innovation:', min = 0.01,max = 1, value =
0.03, step = 0.01)
#                        ,sliderInput('qSize','Q-imitation:', min = 0.2,max = 0.8, value =
0.38, step = 0.01)
#                       ,sliderInput('Years','Years:'
#                                   ,min = as.integer(format(Sys.Date(), format="%Y")) +1
#                                   ,max = as.integer(format(Sys.Date(), format="%Y"))+30
#                                                   ,value = c(as.integer(format(Sys.Date(),
format="%Y"))+1,as.integer(format(Sys.Date(), format="%Y"))+30), step = 1, sep =
")
#                        ,checkboxInput("showgrid", label = "Show Grid", value = TRUE))
# end box
#                    ) # end column
#              )# end fluidRow
```