**ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ**

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

****

**ΤΕΛΙΚΗ ΕΡΓΑΣΙΑ**

ΜΑΘΗΜΑ:

**«****[0531] ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ ΓΝΩΣΗΣ»**

ΟΝΟΜΑΤΕΠΩΝΥΜΑ:

**Κωνσταντίνος Π. Γιαννακόπουλος**

**Αριστείδης Δ. Καλούδης**

ΑΕΜ:

**16051**

**15730**

ΘΕΣΣΑΛΟΝΙΚΗ, ΦΕΒΡΟΥΑΡΙΟΣ 2018

**ΠΡΟΛΟΓΟΣ**

Η παρούσα εργασία αποτελεί την τελική εργασία εξέτασης στο υποχρεωτικό κατ’επιλογήν μάθημα του 5ου Εξαμήνου του ΠΠΣ του Τμήματος Μαθηματικών ΑΠΘ «[0531] ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ ΓΝΩΣΗΣ» για το Ακαδημαϊκό Έτος 2017-2018 και παρουσιάστηκε τον Φεβρουάριο του 2018. Διδάσκοντες καθηγητές του μαθήματος ήταν ο Καθηγητής κ. Ιωάννης Αντωνίου και ο Δρ. Χαράλαμπος Μπράτσας.

Στη συγκεκριμένη εργασία μελετήθηκε ένα σύνολο δεδομένων της εταιρείας Movehub, εταιρείας ειδικών στην παροχή συμβουλών σε επιχειρήσεις και φυσικά πρόσωπα που επιθυμούν να μεταφερθούν ή μεταναστεύσουν οπουδήποτε στον πλανήτη. Το σύνολο περιείχε παρατηρήσεις από 216 πόλεις ανά τον κόσμο, με πληροφορίες όπως το Μέσο Εισόδημα μετά φόρων, το μέσο ενοίκιο ενός διαμερίσματος κ.α.

Στα δεδομένα αυτά διενεργήθηκε περιγραφική στατιστική ώστε να δημιουργηθεί μια αρχική γενική εικόνα για το συνόλο και στη συνέχεια εφαρμόστηκαν διάφοροι στατιστικοί έλεγχοι υποθέσεων καθώς και η κατάλληλη οπτικοποίηση των δεδομένων για να εξαχθούν κάποια βασικά συμπεράσματα για τις σχέσεις μεταξύ των παρατηρήσεων ή των μεταβλητών. Τέλος, τον πυρήνα αυτής της εργασίας αποτελεί η εφαρμογή των μεθόδων στατιστικής μάθησης που έχουν διδαχθεί στο προαναφερθέν μάθημα, κάποιες από χρόνια παγιωμένες (λ.χ. γραμμική παλινδρόμηση) και κάποιες πιο πρόσφατες όπως οι Μηχανές Διανυσμάτων Υποστήριξης(SVMs), και τα Δένδρα Ταξινόμησης και Παλινδρόμησης.

Όλα τα παραπάνω εκπονήθηκαν χρησιμοποιώντας την γλώσσα R και το λογισμικό R Studio.

Table of Contents

**Datasets4**

Γενικές Πληροφορίες4

**Περιγραφική Στατιστική7**

Ραβδογράμματα7

Διαγράμματα Πίτας9

Περιγραφική Στατιστική για Συνεχείς Μεταβλητές11

Scatterplot Matrix13

Ιστογράμματα16

ANOVA19

**Παλινδρόμηση23**

Γραμμική Παλινδρόμηση23

Λογιστική Παλινδρόμηση27

Μέθοδοι Αναδειγματοληψίας29

**Δένδρα33**

Δένδρα Ταξινόμησης33

Δένδρα Παλινδρόμησης37

**Support Vector Machines (SVMs)40**

**Παράρτημα47**

**Βιβλιογραφία48**

**Datasets**

Για την εργασία χρησιμοποιήθηκαν 3 datasets από το Kaggle:

1. **movehubqualityoflife.csv**

Περιέχει 216 πόλεις από 79 διαφορετικές χώρες με τα εξής χαρακτηριστικά για την καθεμία:

**Movehub Rating**: ένας συνδυασμός όλων των σκορ, για ένα τελικό rating για κάθε πόλη.

**Purchase Power**: Συγκρίνει το μέσο κόστος ζωής με το μέσο τοπικό μισθό.

**Health Care**: Συντίθεται από το πώς νιώθουν οι πολίτες σχετικά με την πρόσβαση τους

στο σύστημα υγείας και η ποιότητα του.

**Pollution**: Όσο χαμηλότερη είναι η τιμή τόσο το καλύτερο. Ένας δείκτης για το πόσο μολυσμένη βρίσκουν την πόλη οι πολίτες, συμπεριλαμβανομένου του αέρα, του νερού και της ηχορύπανσης

**Quality of Life**: Ένας συνυπολογισμός των Health Care, Pollution, Purchase Power, Crime Rate που δίνει ένα συνολικό σκορ για την ποιότητα ζωής.

**Crime Rating**: Όσο χαμηλότερη είναι η τιμή τόσο το καλύτερο. Όσο χαμηλότερος είναι τόσο πιο ασφαλείς νιώθουν οι άνθρωποι σε αυτή την πόλη.

1. **movehubcostofliving.csv**

Σε αυτό το dataset περιέχονται οι παρακάτω πληροφορίες:

Χρηματική μονάδα: GBP(£)

**City**: 216 ονόματα πόλεων ανά τον κόσμο.

**Cappuccino**: Η μέση τιμή μιας κούπας Cappuccino

**Cinema**: Η μέση τιμή του εισιτηρίου για τον κινηματογράφο.

**Wine**: Η μέση τιμή κρασιού.

**Gasoline**: Η τιμή της βενζίνης.

**Avg. Rent**: Το μέσο ενοίκιο.

**Avg Disposable Income**: (DPI) Το χρηματικό ποσό που είναι διαθέσιμο για να ξοδευτεί ή να αποταμιευτεί, μετά φόρου, ανά άτομο.

**3. cities.csv**

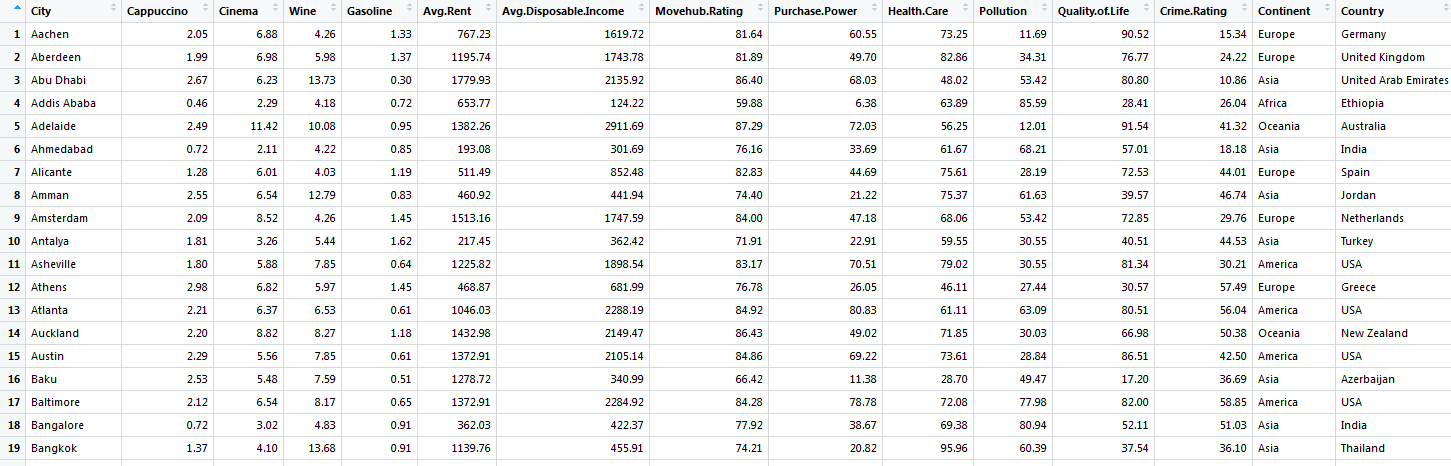
Μια λίστα από πόλεις που είχαν πληθυσμό πάνω από 100.000, για το 2006.

* Στη συνέχεια, τα τρία .csv αρχεία ενώθηκαν σε ένα και προστέθηκαν δύο νέες στήλες-μεταβλητές με τα ονόματα:

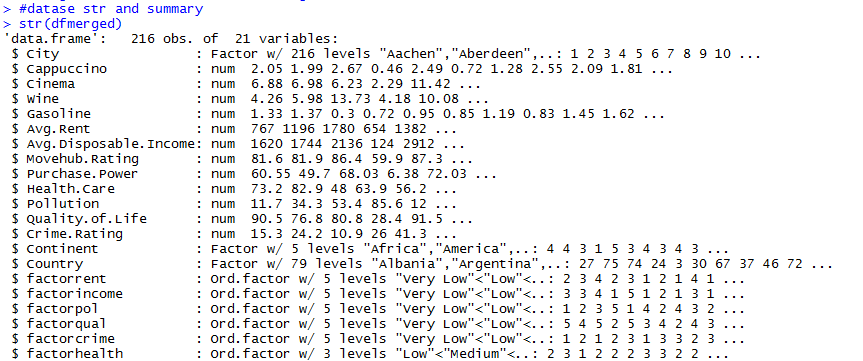
**Continent**: δείχνει την ήπειρο στην οποία ανήκει η κάθε πόλη - παρατήρηση.

**Country**: δείχνει τη χώρα στην οποία ανήκει η κάθε πόλη - παρατήρηση.

Στη συνέχεια, το νέο πλέον ενωμένο αρχείο .csv εισήχθει στην R έχοντας την εξής μορφή.

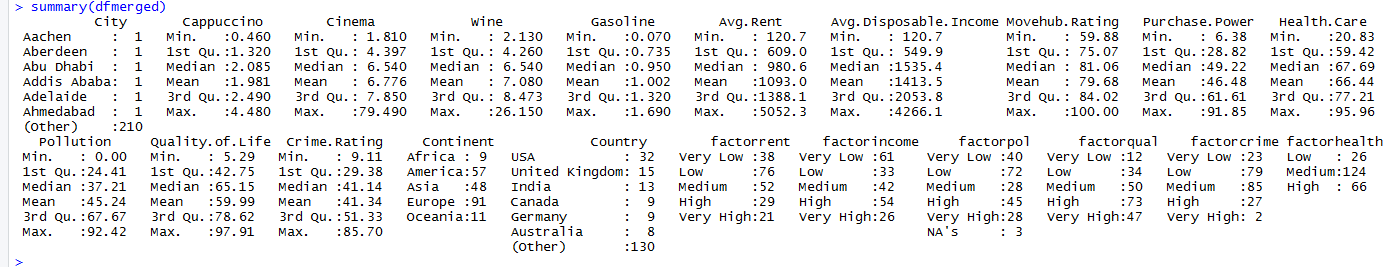


* Το dataset δεν έχει missing values. Ωστόσο, παρατηρούμε ότι δεν έχει και ποιοτικές μεταβλητές εκτός από τις City, Continent και Country. Κρίνεται απαραίτητο, τόσο για την περιγραφική στατιστική όσο και για κάποια tests αλλά και μοντέλα πρόβλεψης, να δημιουργηθούν κάποιες ποιοτικές μεταβλητές. Επομένως, ορίζονται οι factorrent, factorincome, factorhealth, factorpol, factorqual, factorcrime.

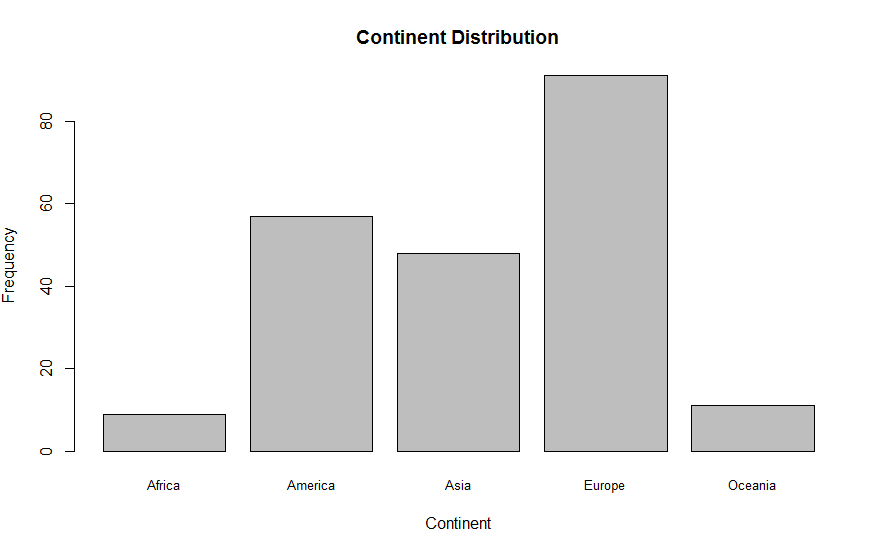
****

**Περιγραφική Στατιστική**

Αρχικά, βλέπουμε το summary του dataset όπου υπολογίζονται μερικά στατιστικά μέτρα με τα οποία όμως θα ασχοληθούμε και λίγο αργότερα.



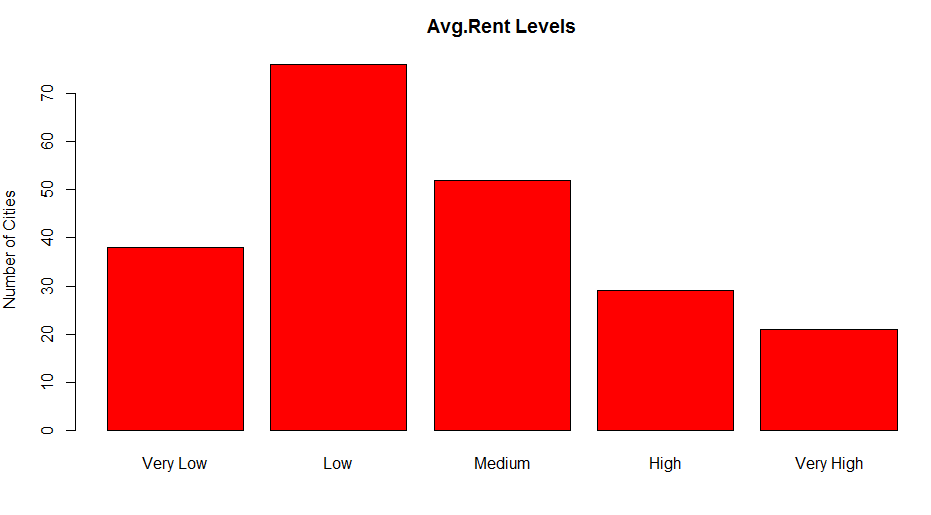
**Διαγράμματα Barplot για τις μεταβλητές Continent, factorincome, factorrent**



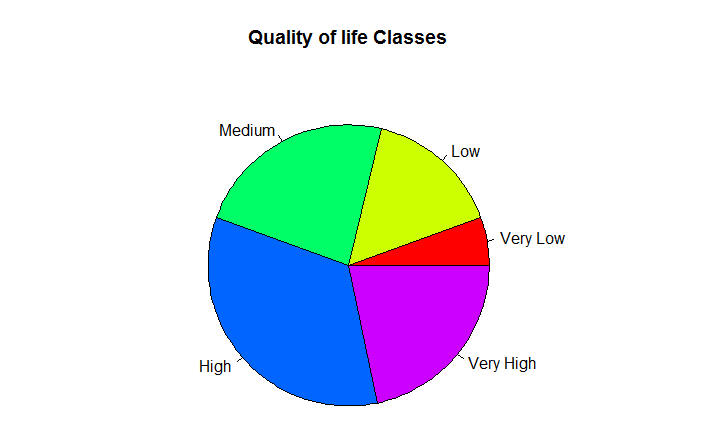
Στο διάγραμμα αυτό παρατηρούμε ότι μεγάλο μέρος των παρατηρήσεων συγκεντρώνεται στην Ευρώπη, ενώ είναι λίγες οι πόλεις στην Αφρική και στην Ωκεανία.



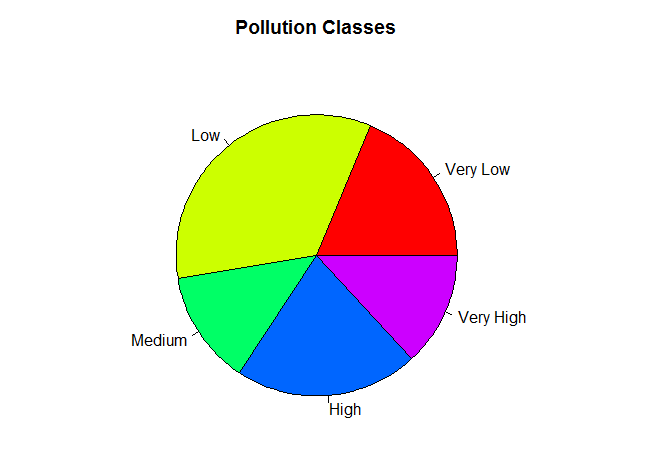
Στο Barplot για το Avg. Disposable Income, παρατηρούμε ότι οι ακραίες κατηγορίες δεν είναι συμμετρικές, δηλαδή στην κατηγορια "Very Low" υπάρχουν περίπου οι διπλάσιες παρατηρήσεις σε σχέση με την κατηγορία "Very High".



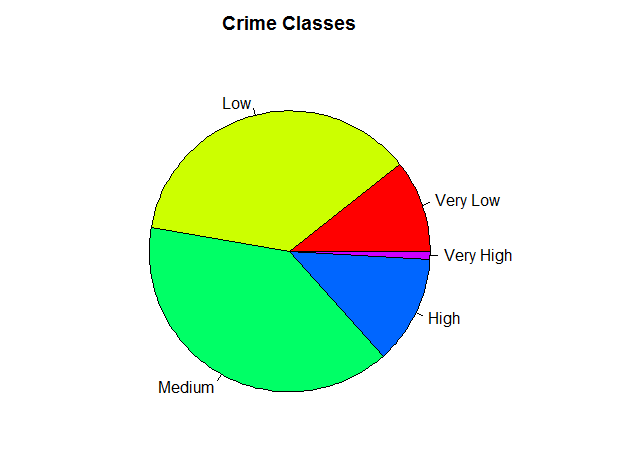
Αντίστοιχα για τις κλάσεις της μεταβλητής Avg. Rent, η περισσότερες παρατηρήσεις συγκεντρώνονται στην κατηγορία "Low", ενώ στις κατηγορίες "High" και "Very High" υπάρχουν λίγες παρατηρήσεις.

**Διαγράμματα Πίτας για τις μεταβλητές fqual, fpol, fcrime**

Στο παραπάνω διάγραμμα, φαίνεται ότι περισσότερες από τις μισές παρατηρήσεις βρίσκονται στις κατηγορίες "High" και "Very High", ενώ λιγότερες από το ένα τέταρτο αυτών βρίσκονται στις κατηγορίες "Low" και "Very Low".



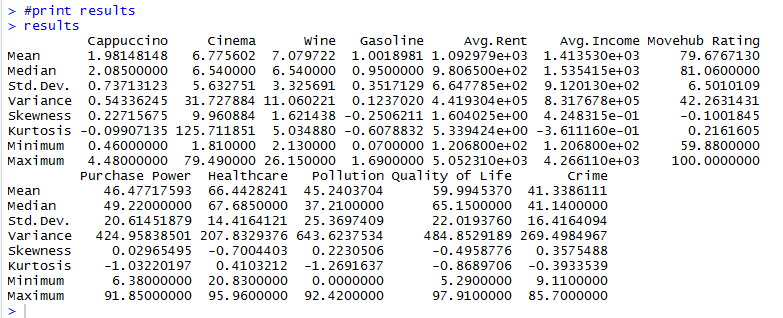
Όσον αφορά το Pollution, οι παρατηρήσεις μοιράζονται σχετικά ομοιόμορφα μεταξύ των κατηγοριών "Very Low", "Medium", "High" και "Very High", ενώ είναι αρκετά περισσότερες στην κατηγορία 'Low".



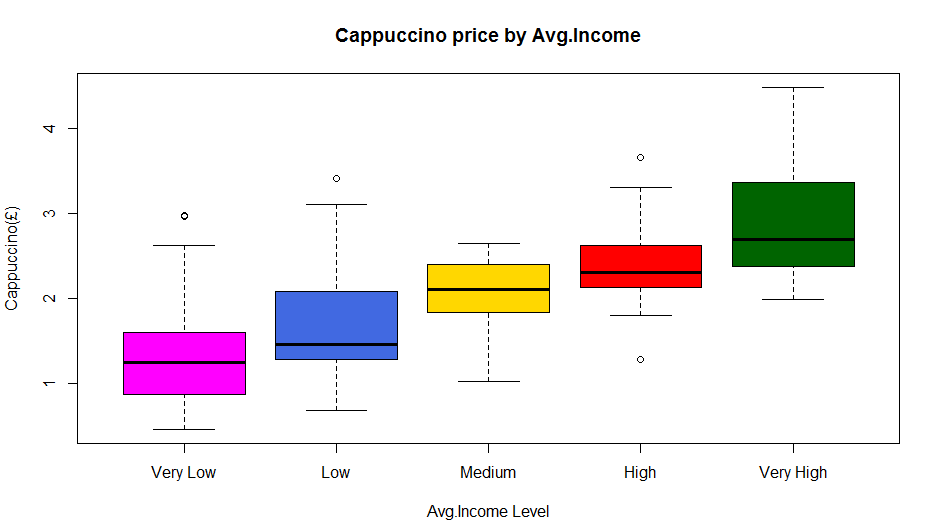
Στην μεταβλητή Crime, τα τρία τέταρτα των παρατηρήσεων μοιράζονται περίπου ισόποσα στις κατηγορίες "Low" και "Medium", ενώ η κατηγορία "Very High" έχει ελάχιστες παρατηρήσεις.

**Περιγραφική στατιστική για τις συνεχείς μεταβλητές**

Αρχικά, για τις ποσοτικές μεταβλητές του dataset υπολογίστηκαν κάποια βασικά στατιστικά μέτρα: η μέση τιμή, η διάμμεσος, η διακύμανση, η τυπική απόκλιση, η λοξότητα, η κύρτωση, η ελάχιστη και η μέγιστη τιμή.

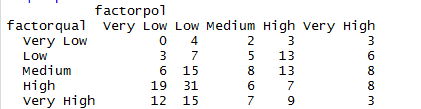


Στη συνέχεια έχουμε ένα Boxplot του Avg. Disposable Income συναρτήσει της τιμής το Cappuccino:



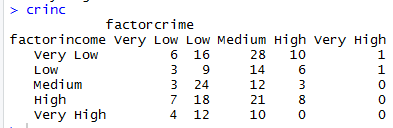
Παρατηρούμε ότι η τιμή του Cappuccino έχει σχέση με το Avg. Disposable Income, με ένα μεγάλο άλμα να γίνεται μεταξύ των κατηγοριών "Low" και "Medium" αλλά και των "High" και "Very High".

**Πίνακας Crosstabs για τις μεταβλητές Quality of Life και Pollution**

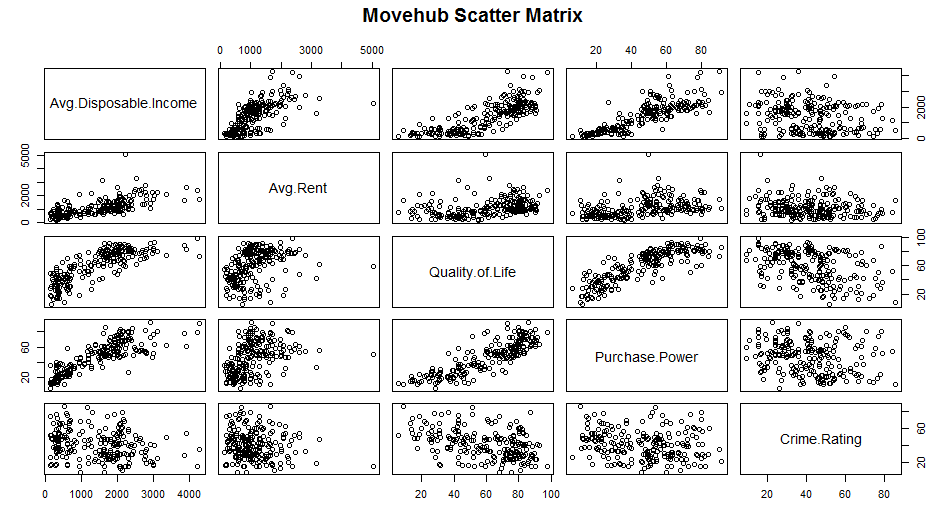
****

Εδώ τα αποτελέσματα είναι σχετικά αναμενόμενα, καθώς στις πόλεις με χαμηλότερη ρύπανση η ποιότητα ζωής είναι χαμηλότερη, ενώ όπου είναι υψηλότερη είναι και υψηλότερος ο δείκτης ποιότητας ζωής. Ενδιαφέρον έχει η κατηγορία "Medium" στη μεταβλητή Quality of Life, καθώς 21 παρατηρήσεις βρίσκονται στις κατηγορίες "Very Low" και "Low", 8 στην κατηγορία "Medium" και 21 στις κατηγορίες "High" και "Very High", πράγμα που δείχνει ότι η μέτρια ποιότητα ζωής μπορεί να συνυπάρχει και με Χαμηλή και με Υψηλή ρύπανση.

**Πίνακας Crosstabs για τις μεταβλητές Crime και Avg. Disposable Income**



Παρατηρούμε ότι στις πόλεις με υψηλό εισόδημα δεν έχουμε υψηλή εγκληματικότητα.

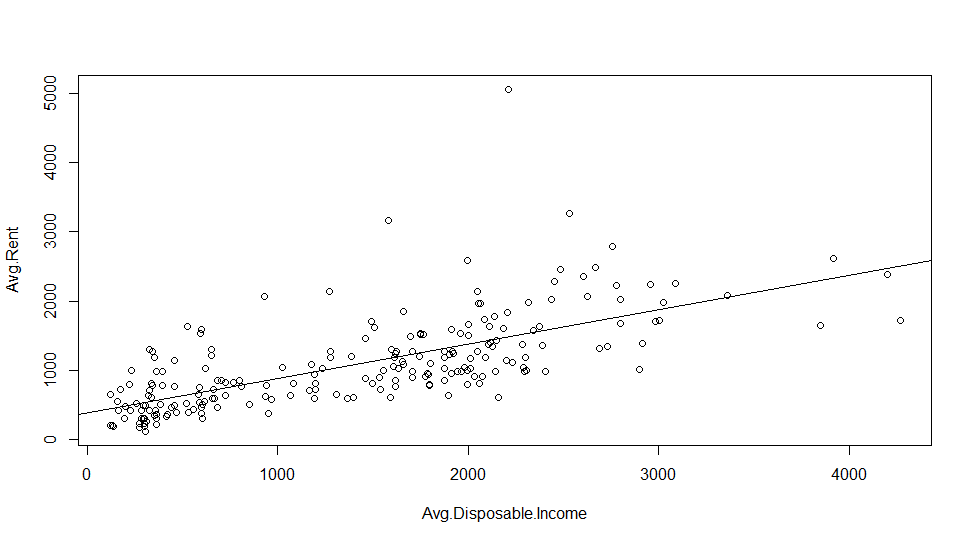
**Scatterplot Matrix**

O παραπάνω πίνακας γραφημάτων μας δείχνει ότι δεν υπάρχει γραμμική σχέση μεταξύ κάποιων μεταβλητών, όπως για παράδειγμα το Crime Rating με όλες τις υπόλοιπες ή το Purchase Power - Avg. Rent, ενώ υπάρχει κάποια ίσως γραμμική σχέση μεταξύ των Quality of Life - Purchase Power, Avg. Disposable Income - Purchase Power, Avg. Disposable Income - Avg. Rent και Quality of Life - Avg. Disposable Income.

Μπορούμε να υπολογίσουμε τον συντελεστή συσχέτισης Pearson για τις μεταβλητές Avg. Disposable Income - Avg. Rent.

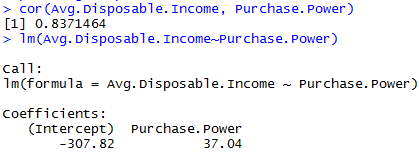


Μπορούμε επίσης να δημιουργήσουμε και ένα Scatterplot για τις μεταβλητές μας, που θα περιέχει και την ευθεία παλινδρόμησης.

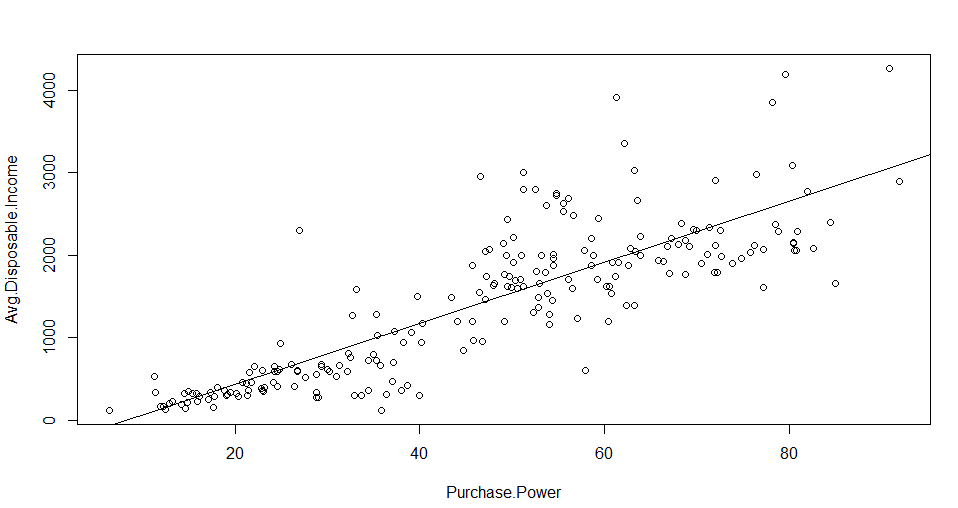


Τόσο ο συντελεστής Pearson (~0.68), όσο και το παραπάνω γράφημα μας δείχνουν πως δεν υπάρχει τόσο καλή γραμμική σχέση μεταξύ τους.

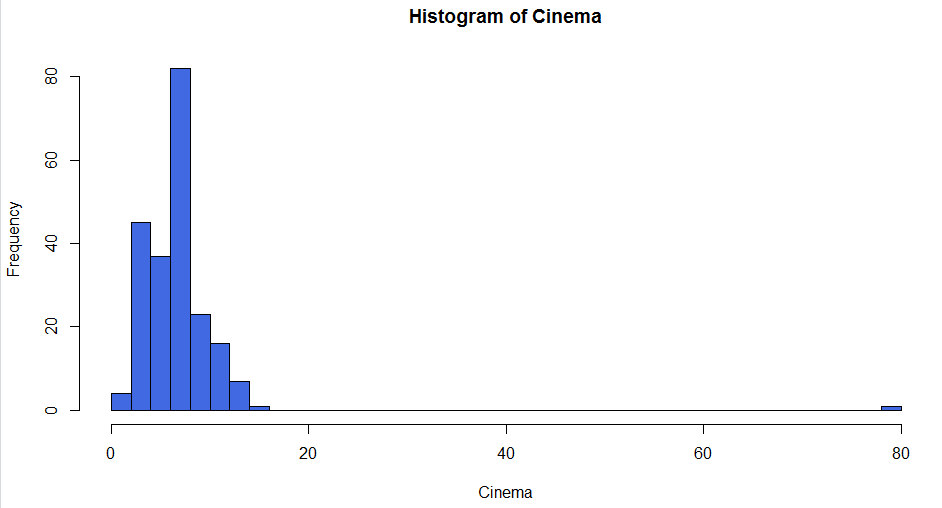
Ωστόσο, για τις μεταβλητές Avg. Disposable Income και Purchase Power έχουμε καλό συντελεστή Pearson.



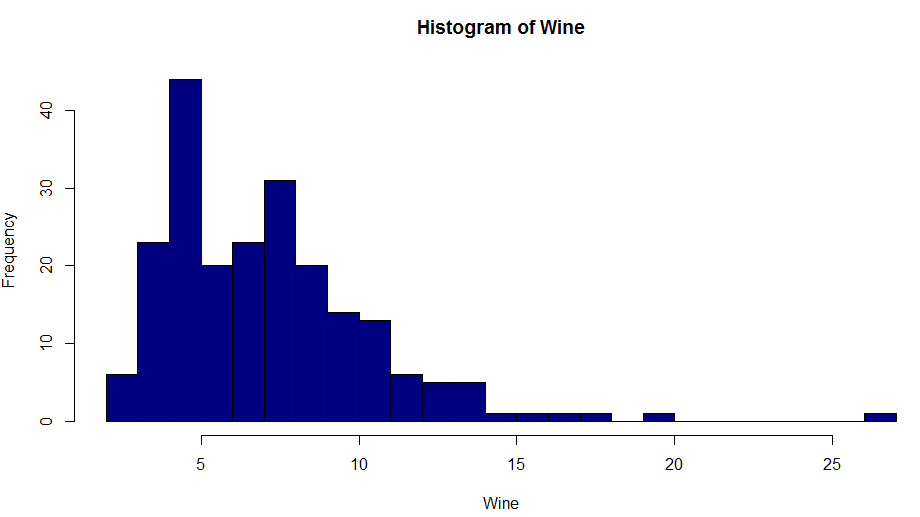
Η καλή γραμμική συσχέτιση αποτυπώνεται και στο scatterplot, κυρίως στις χαμηλότερες τιμές.



**Ιστογράμματα**

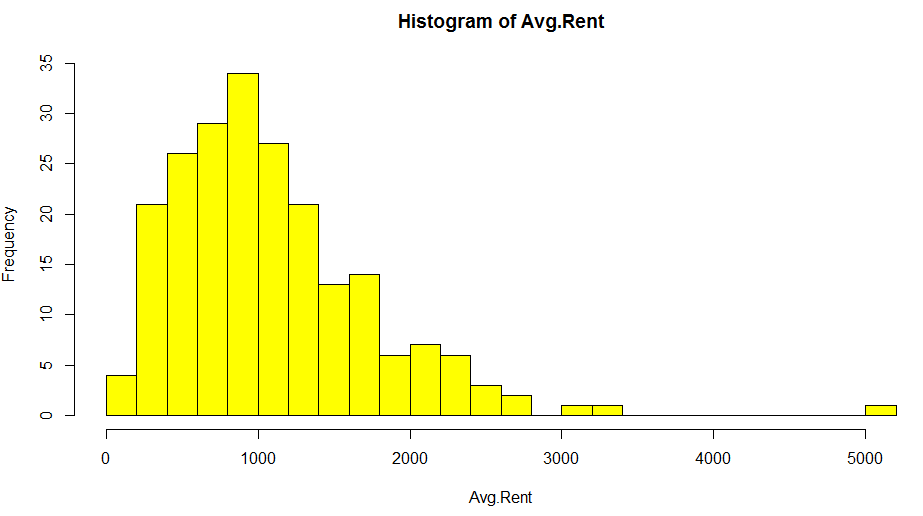


Η ακραία τιμή παρατηρείται στη Σαουδική Αραβία, καθώς το σινεμά είχε απογορευτεί για 35 χρόνια και η άρση της απαγόρευσης έχει προγραμματιστεί για το Μάρτιο του 2018.



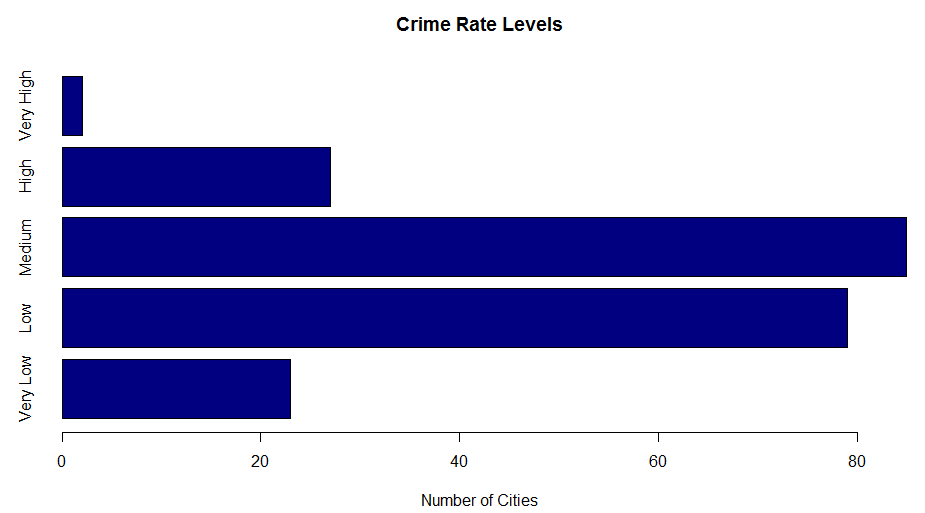
Αντίστοιχα, έχουμε το ιστόγραμμα για τη μεταβλητή Wine και Avg. Rent.

Για τη μεταβλητή Wine, οι 5 υψηλότερες τιμές της παρατηρούνται σε χώρες στις οποίες η επικρατέστερη θρησκεία απαγορεύει το αλκοόλ.

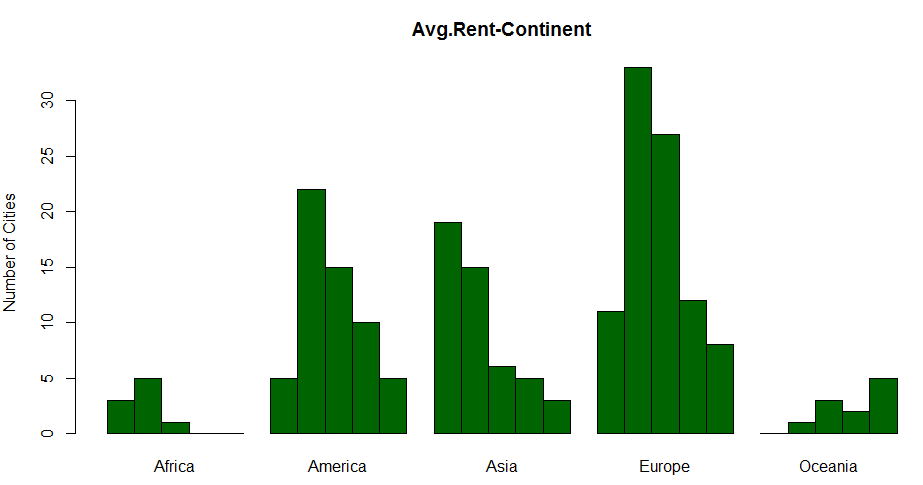


**Plots**

Τέλος, έχουμε δύο διαφορετικά barplots. Στο παρακάτω παρατηρούμε ότι υπάρχουν ελάχιστες παρατηρήσεις στην κατηγορία "Very High"



Στο παρακάτω διάγραμμα παρατηρούμε τις διαφορετικές κατανομές του Avg. Rent σε κάθε ήπειρο.

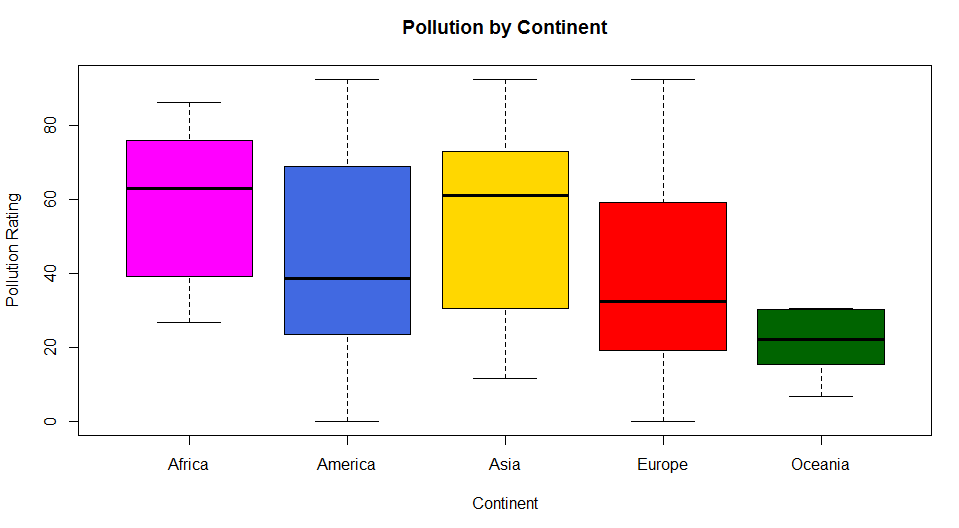
****

**ΑNOVA - Ανάλυση Διασποράς**

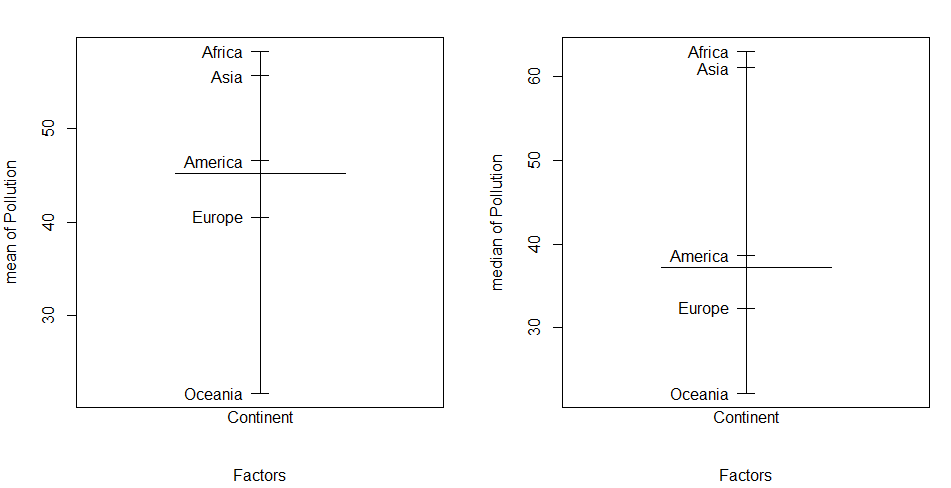
Θα χρησιμοποιήσουμε τη μέθοδο ANOVA, ένα στατιστικό τεστ για έλεγχο των στατιστικών μέσων αναλύοντας την μεταβλητότητα του δείγματός μας.

Αρχικά, θα ελέγξουμε αν υπάρχει σημαντική διαφορά της μέσης τιμής της ρύπανσης ανά ήπειρο.

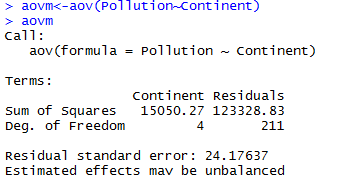
Εποπτικά, το γράφημα Boxplot δείχνει ότι θα υπάρχει διαφορα στις μέσες τιμές.



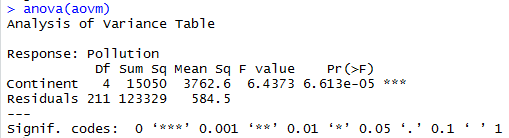
Το ίδιο ισχύει και για το επόμενο γράφημα στο οποίο φαίνεται η διαφορά των μέσων τιμών αλλά και των διαμέσων.



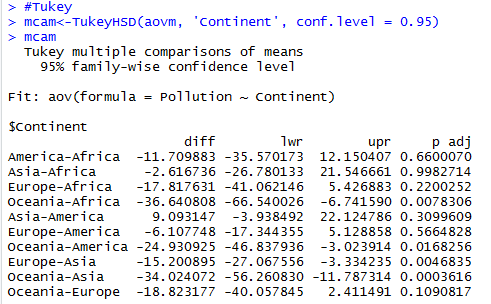
Στη συνέχεια πραγματοποιούμε την ανάλυση διασποράς:

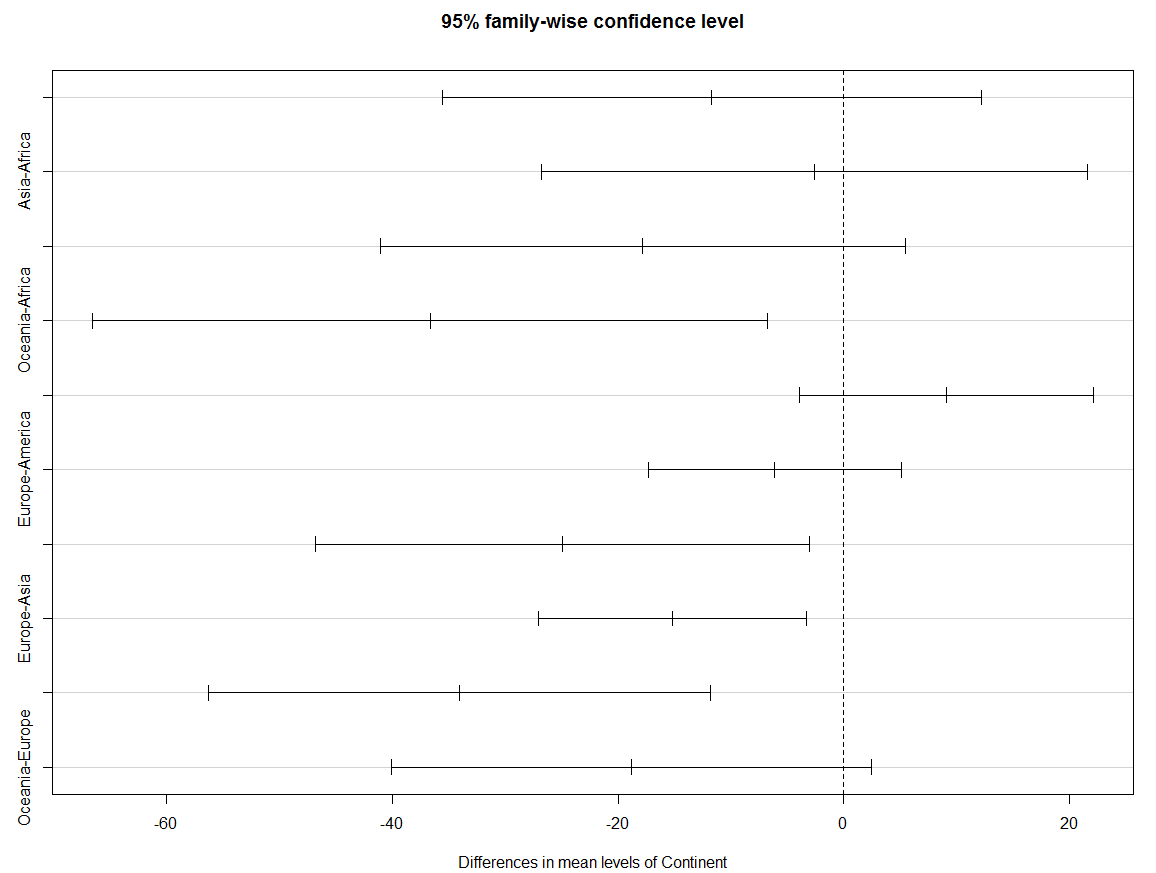


Παρατηρούμε ότι η p-value είναι μικρότερη από 0.05. Επομένως, οι μέσες τιμές διαφέρουν στατιστικά.

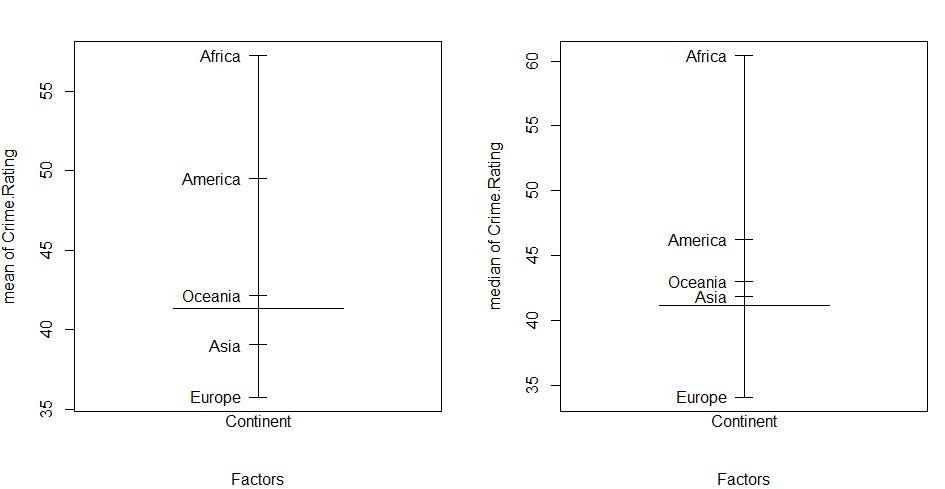


**Tukey Test**

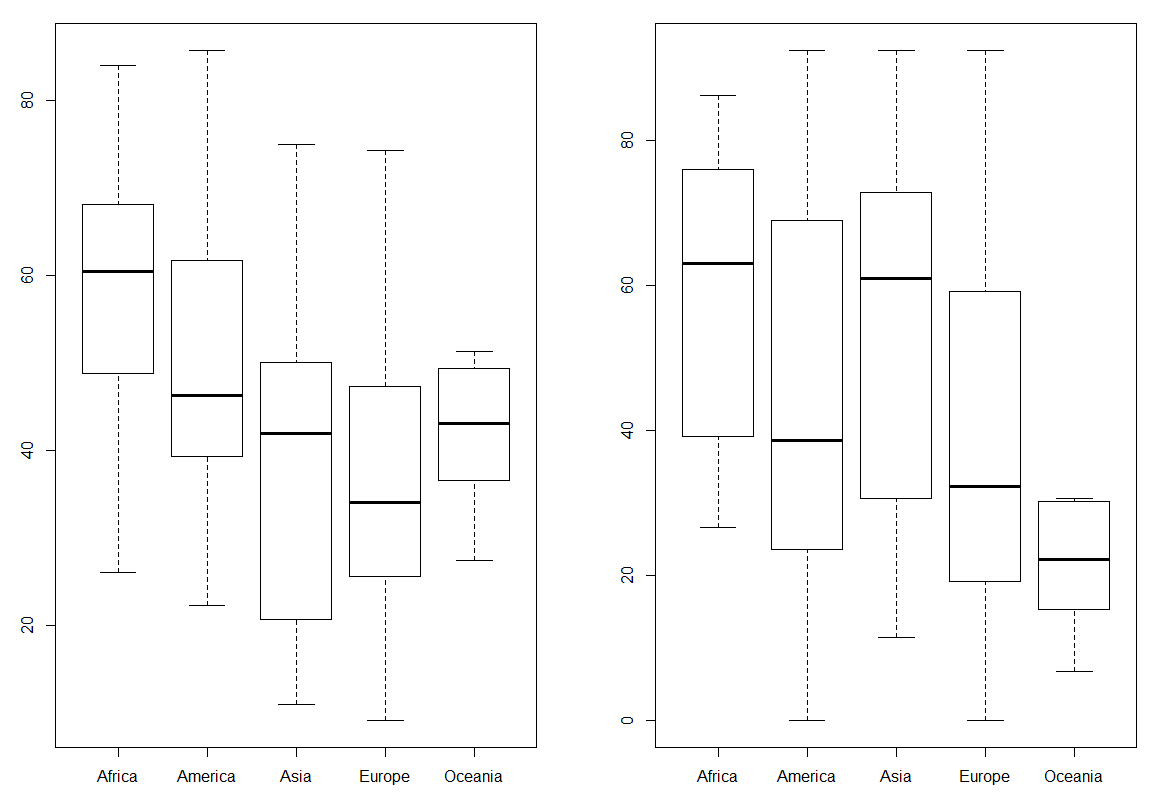
Στη συνέχεια προχωρούμε σε ένα Tukey test, έτσι ώστε να συγκρίνουμε τις μέσες τιμές σε ζεύγη ανά δύο, δημιουργώντας διαστήματα εμπιστοσύνης 95%.

Όπως είναι εμφανές, τόσο από τον παραπάνω πίνακα όσο και από το παρακάτω διάγραμμα, τα Δ.Ε. των μέσων τιμών της ρύπανσης των ζευγών Ωκεανία-Αφρική, Ωκεανία-Αμερική, Ευρώπη-Ασία και Ωκεανία-Ασία δεν περιέχουν το 0 και κατά συνέπεια διαφέρουν σημαντικά οι μέσες τιμές.

Στη συνέχεια, εντελώς εποπτικά βλέποντας τα παρακάτω διαγράμματα μπορούμε να υποθέσουμε ότι και οι μέσες τιμές της ρύπανσης ανά ήπειρο παρουσιάζουν σημαντικές διαφορές.

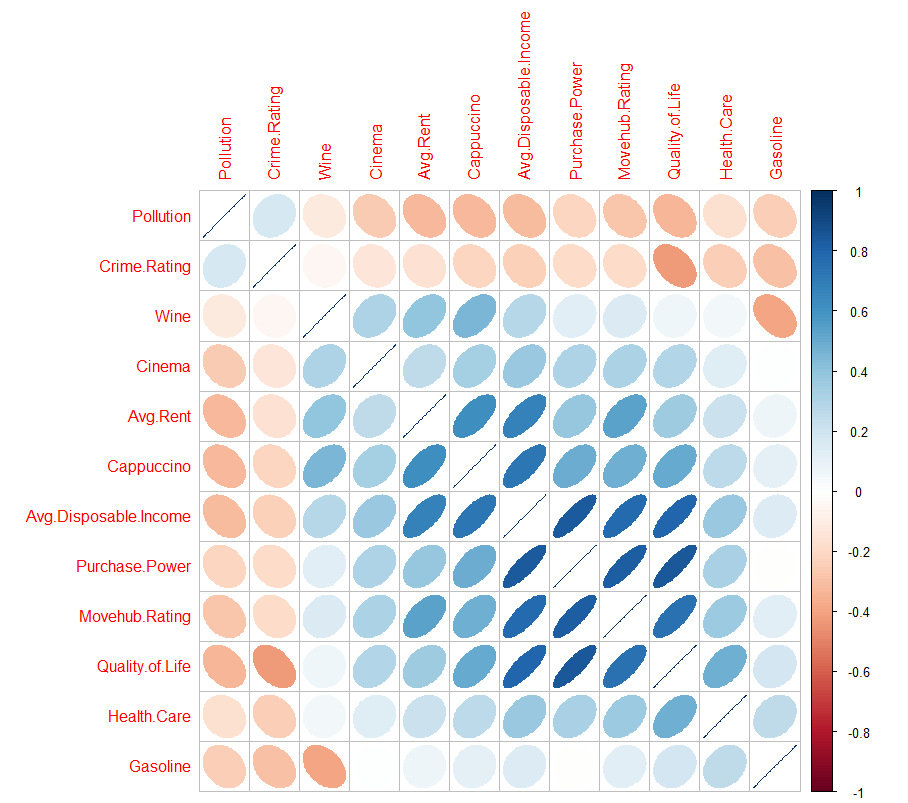


Παρακάτω μπορούμε να συγκρίνουμε τις διαφορές των μέσων τιμών της εγκληματικότητας (αριστερά) και της ρύπανσης (δεξιά).

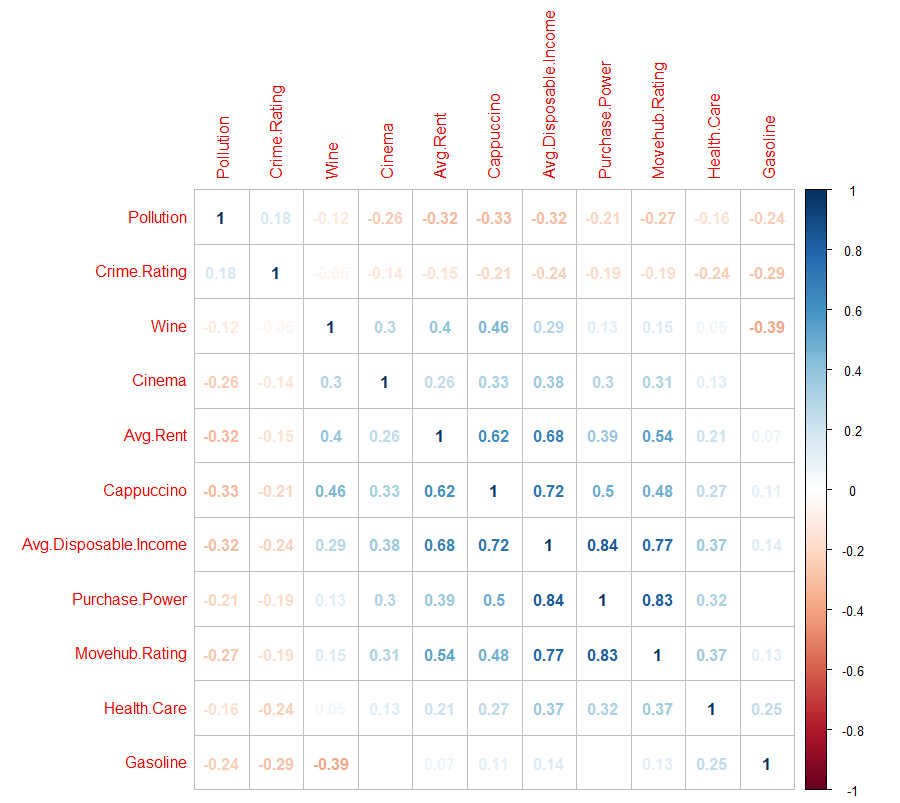


**Γραμμική Παλινδρόμηση**

Στη συνέχεια, θα δημιουργήσουμε ένα μοντέλο γραμμικής παλινδρόμησης. Προκειμένου του μοντέλο να έχει σχετικά καλά αποτελέσματα, θα πρέπει αν ελέγξουμε το συντελεστή συσχέτισης των μεταβλητών. Ένας πολύ καλός και οργανωμένος τρόπος είναι οι δύο παρακάτω πίνακες.



Το μπλε χρώμα δείχνει θετική συσχέτιση, ενώ το κόκκινο αρνητική. Το πάχος της έλλειψης είναι αντιστρόφως ανάλογο με την απόλυτη τιμή του συντελεστή συσχέτισης, άρα όταν οι τιμές πλησιάζουν το +1 και -1 γίνεται λεπτότερη, ενώ όταν προσεγγίζουν το 0 γίνεται πλατύτερη.



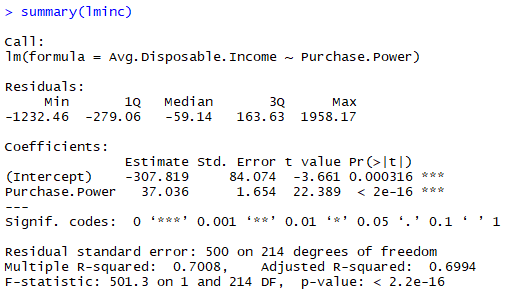
Σε αυτόν τον πίνακα δίνονται αριθμητικά και συγκεντρωτικά οι συντελεστές συσχέτισης όλων των μεταβλητών.

Όπως είναι εμφανές, τη μεγαλύτερη γραμμική συσχέτιση έχουν οι μεταβλητές:

1. Purchase Power~ Avg. Disposable Income (p=0.84)
2. Movehub Rating~ Purchase Power (p=0.83)
3. Movehub Rating~ Avg. Disposable Income (p=0.77)
4. Avg. Disposable Income~ Cappuccino (p=0.72)

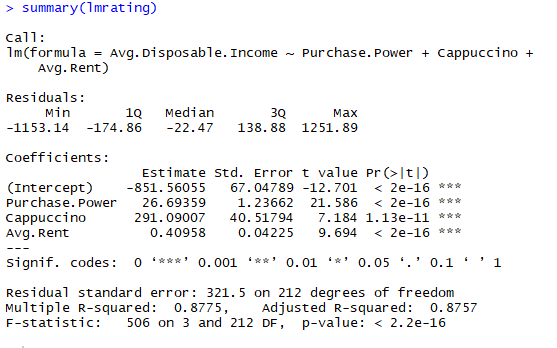
Δημιουργούμε λοιπόν το μοντέλο παλινδρόμησης:

**Avg. Disposable Income ~ Purchase Power**

****

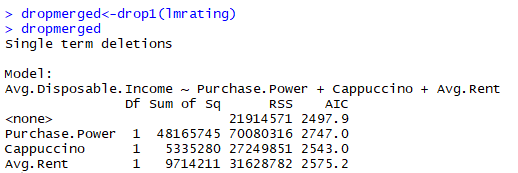
Το R-squared είναι ~70% που σημαίνει ότι το 70% περίπου της διακύμανσης του Avg. Disposable Income εξαρτάται από τη μεταβλητή Purchase Power.

Δεν είναι τόσο ικανοποιητικό το αποτέλεσμα, γι' αυτό θα δημιουργήσουμε ένα νέο μοντέλο, προσθέτοντας άλλους δύο όρους:

**Avg. Disposable Income~ Purchase Power + Cappuccino + Avg. Rent**

Η διαφορά στο R-squared είναι σημαντική αφού περίπου το 87.5% της διακύμανσης του Avg. Disposable Income εξαρτάται από τις μεταβλητές Purchase Power, Cappuccino, Avg. Rent.

Στη συνέχεια, θα εξετάσουμε αν το μοντέλο θα μας έδινε καλύτερα αποτελέσματα αφαιρώντας κάποιον όρο. Αυτό θα γίνει χρησιμοποιώντας το κριτήριο πληροφορίας Akaike (AIC).



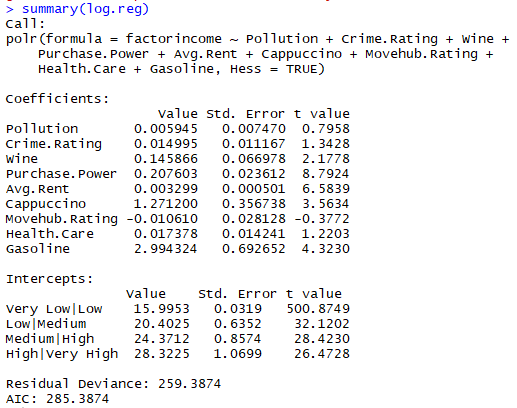
Παρατηρούμε ότι το μικρότερο AIC το έχει το μοντέλο χωρίς να αφαιρέσουμε κάποιον όρο, AIC=2497.9.

**Λογιστική Παλινδρόμηση**

Για τη λογιστική παλινδρόμηση χρησιμοποιήσαμε την εντολή polr από το πακέτο MASS, προκειμένου να γίνει για μια ordered factor μεταβλητή.

Για τη λογιστική παλινδρόμηση δημιουργήσαμε το παρακάτω μοντέλο:

**factorincome~ Pollution + Crime.Rating + Wine + Purchase Power + Avg.Rent + Cappuccino + Movehub.Rating + Health.Care + Gasoline**

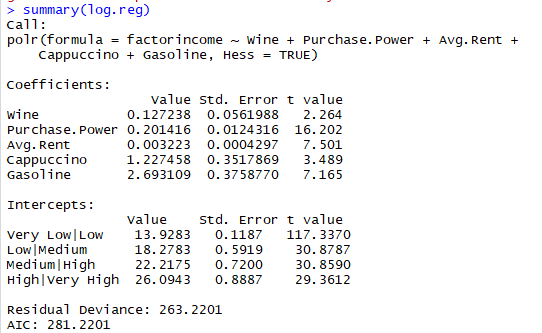
Η μεταβλητή factorincome έχει οριστεί ως ordered factor με 5 levels: Very Low < Low < Medium < High < Very High.

To AIC του μοντέλου είναι περίπου 285.

Φαίνεται ότι οι μεταβλητές που είναι πιο σημαντικές για το μοντέλο είναι οι: Wine, Purchase Power, Avg. Rent, Cappuccino, Gasoline.

Δημιουργούμε λοιπόν ένα νέο μοντέλο λογιστικής παλινδρόμησης με τις παραπάνω μεταβλητές.

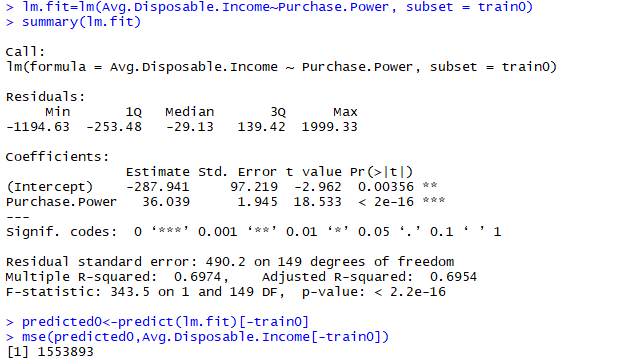
**factorincome~ Wine + Purchase Power + Avg.Rent + Cappuccino + Gasoline**

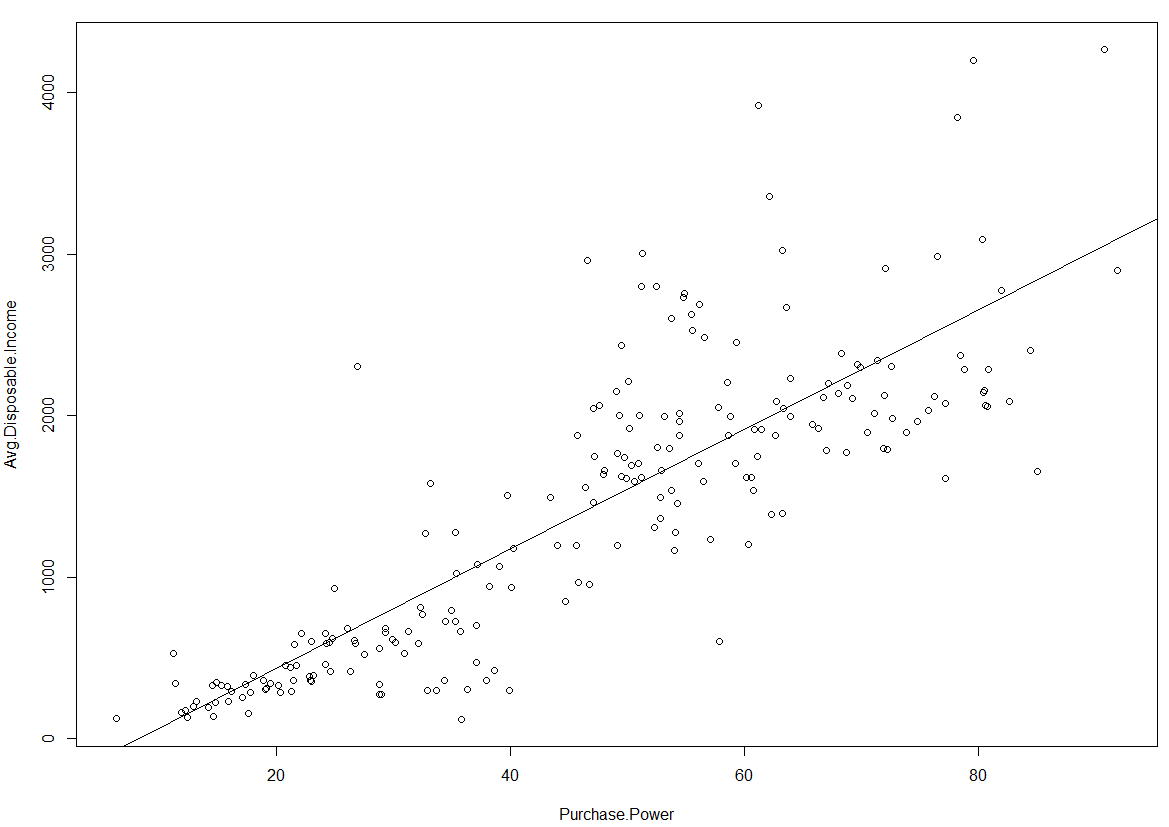


Το AIC του μοντέλου έχει μειωθεί και είναι περίπου 281, μειώθηκε και επομένως το μοντέλο μας είναι καλύτερο.

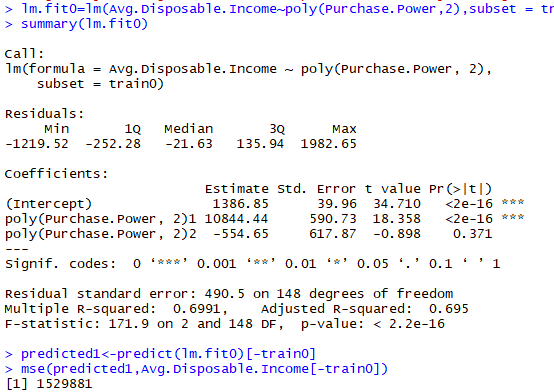
**Μέθοδοι αναδειγματοληψίας**

**Μέθοδος HOLD OUT**

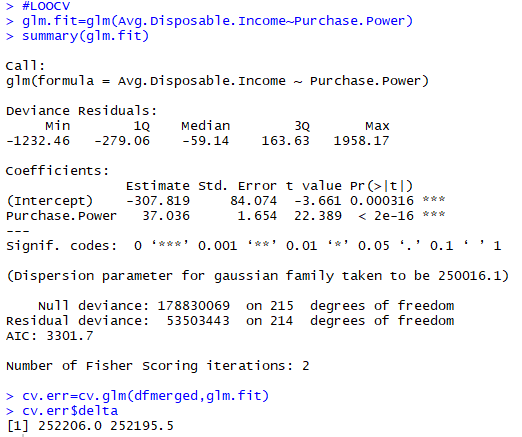
Χρησιμοποιώντας τη μέθοδο HOLD OUT θα χωρίσουμε τις παρατηρήσεις μας σε δύο υποσύνολα: 70% των παρατηρήσεων ως training set και το 30% ως test set.



Στη συνέχεια θα εφαρμόσουμε μια γραμμική παλινδρόμηση και μια γραμμική πολυωνυμική παλινδρόμηση.

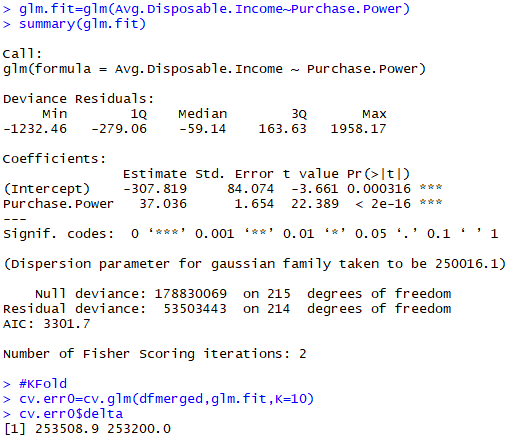


Το μέσο τετραγωνικό σφάλμα μειώθηκε στη δεύτερη περίπτωση. Στο σημείο αυτό θα πρέπει να λάβουμε υπόψιν και τη διακύμανση της μεταβλητής Avg. Disposable Income που είναι 831767.8

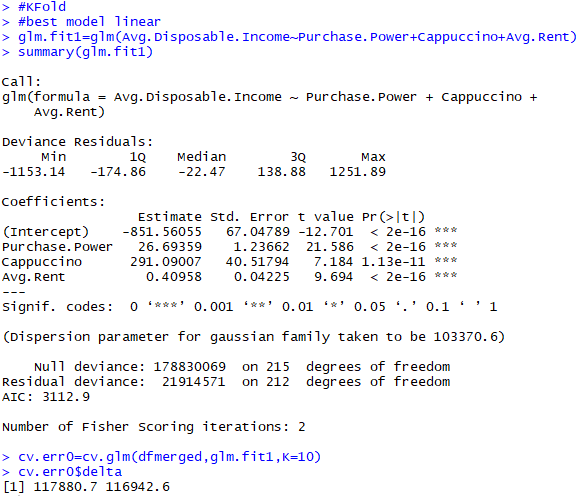
**Μέθοδος Leave-one-out Cross Validation**

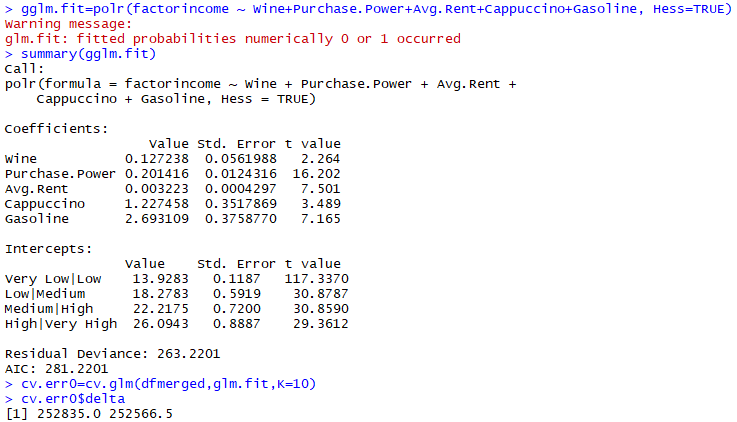
Παρατηρούμε ότι η εκτίμηση μας για το σφάλμα είναι 252,206.

**Μέθοδος 10-fold Cross Validation**

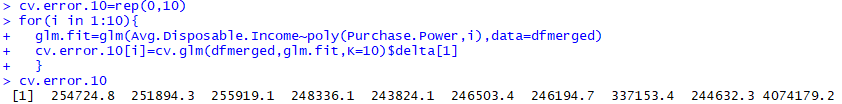


Θα χρησιμοποιήσουμε την ίδια μέθοδο για το βέλτιστο γραμμικό μοντέλο που είχαμε βρει νωρίτερα. Παρατηρούμε ότι το σφάλμα μειώθηκε κατά 56%.



Θα χρησιμοποιήσουμε πάλι την ίδια μέθοδο, αυτή τη φορά για ένα μοντέλο λογιστικής παλινδρόμησης.

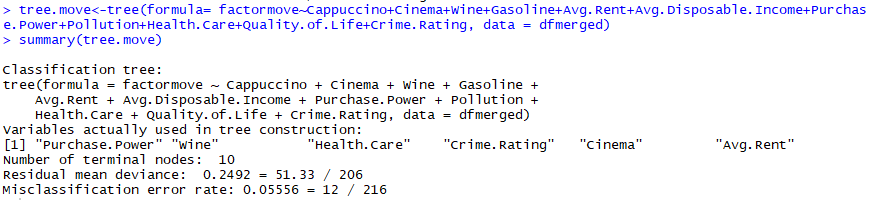
Τέλος, θα βρούμε το καλύτερο μοντέλο γραμμικής πολυωνυμικής παλινδρόμησης σύμφωνα με το σφάλμα που προκύπτει.



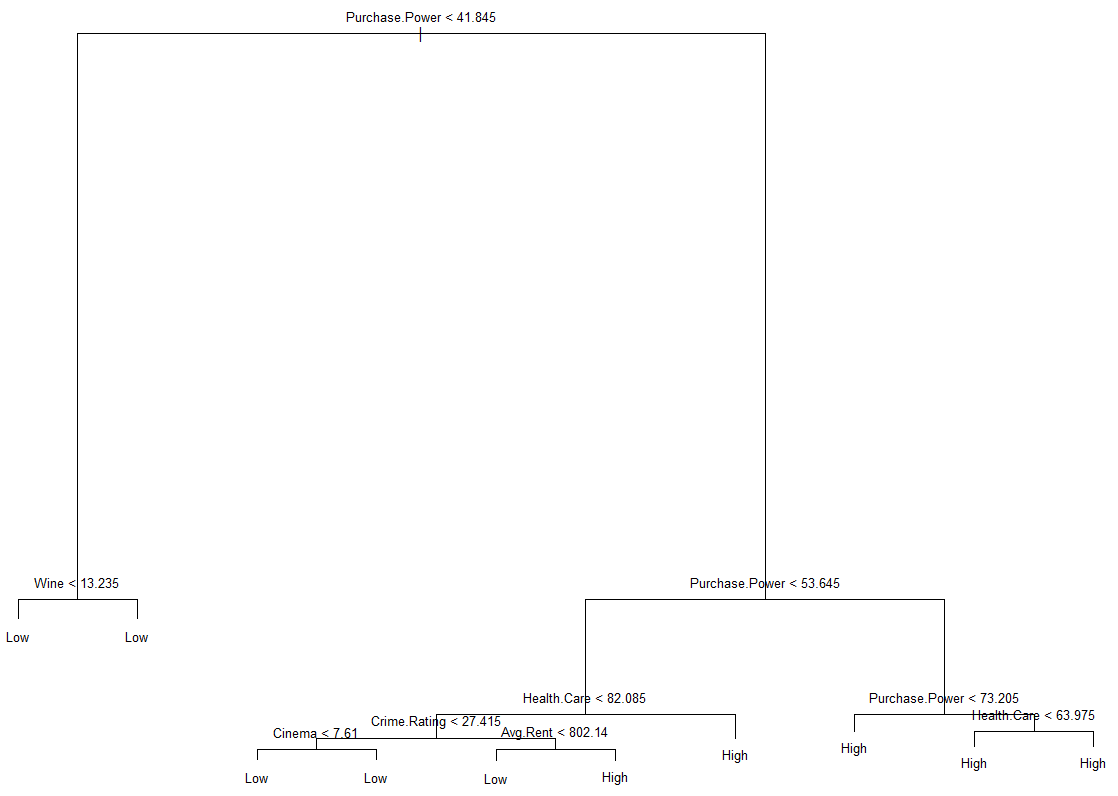
Έχουμε λοιπόν ότι η πολυωνυμική τάξης 5 δίνει το μικρότερο σφάλμα.

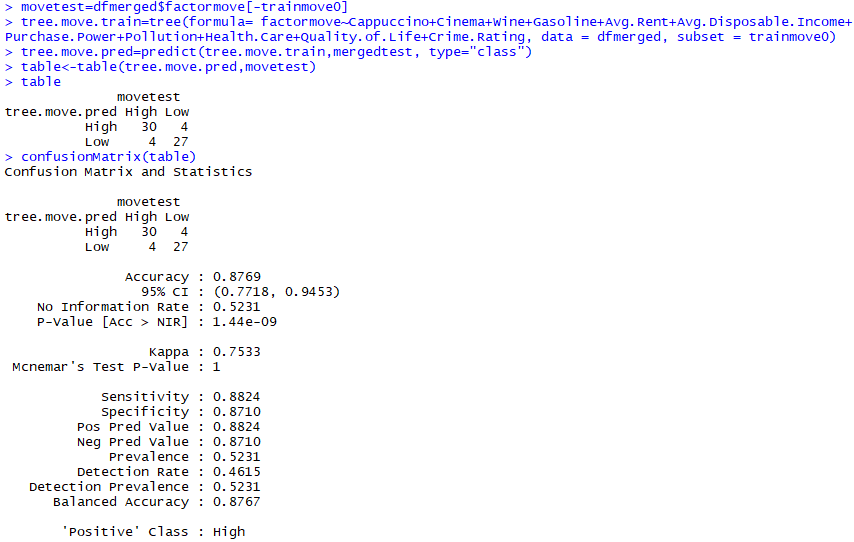
**Δέντρα Ταξινόμησης**

Δημιουργούμε και εισάγουμε στο dataset μια νέα ordered factor μεταβλητή την factormove, με δύο επίπεδα Low και High, ανάλογα με το αν η παρατήρηση είναι μικρότερη ή μεγαλύτερη από τη διάμεσο της μεταβλητής Movehub Rating.

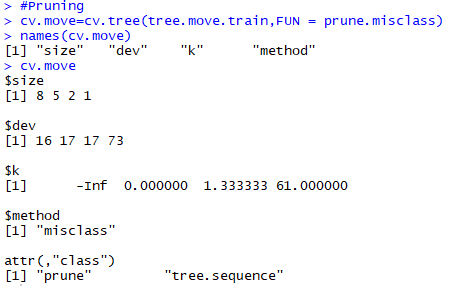
Στη συνέχεια, δημιουργούμε ένα δέντρο ταξινόμησης για να προβλέψουμε τη μεταβλητή factormove χρησιμοποιώντας όλες τις άλλες μεταβλητές του dataset.

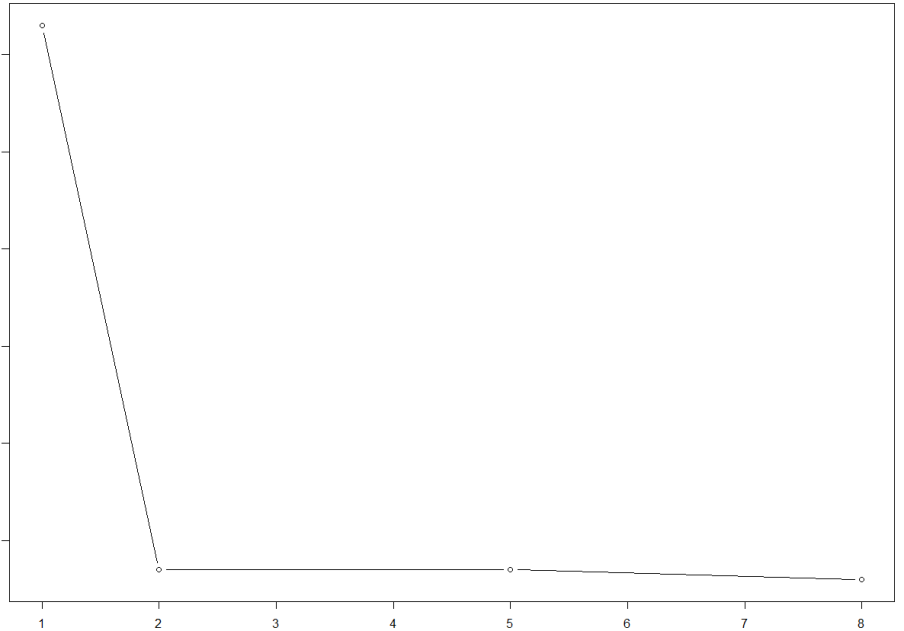
Παρατηρούμε ότι το δέντρο θα έχει 10 κόμβους και το σφάλμα εκπαίδευσης του είναι 12/216 δηλαδή περίπου 5.6%. Γραφικά φαίνεται παρακάτω:



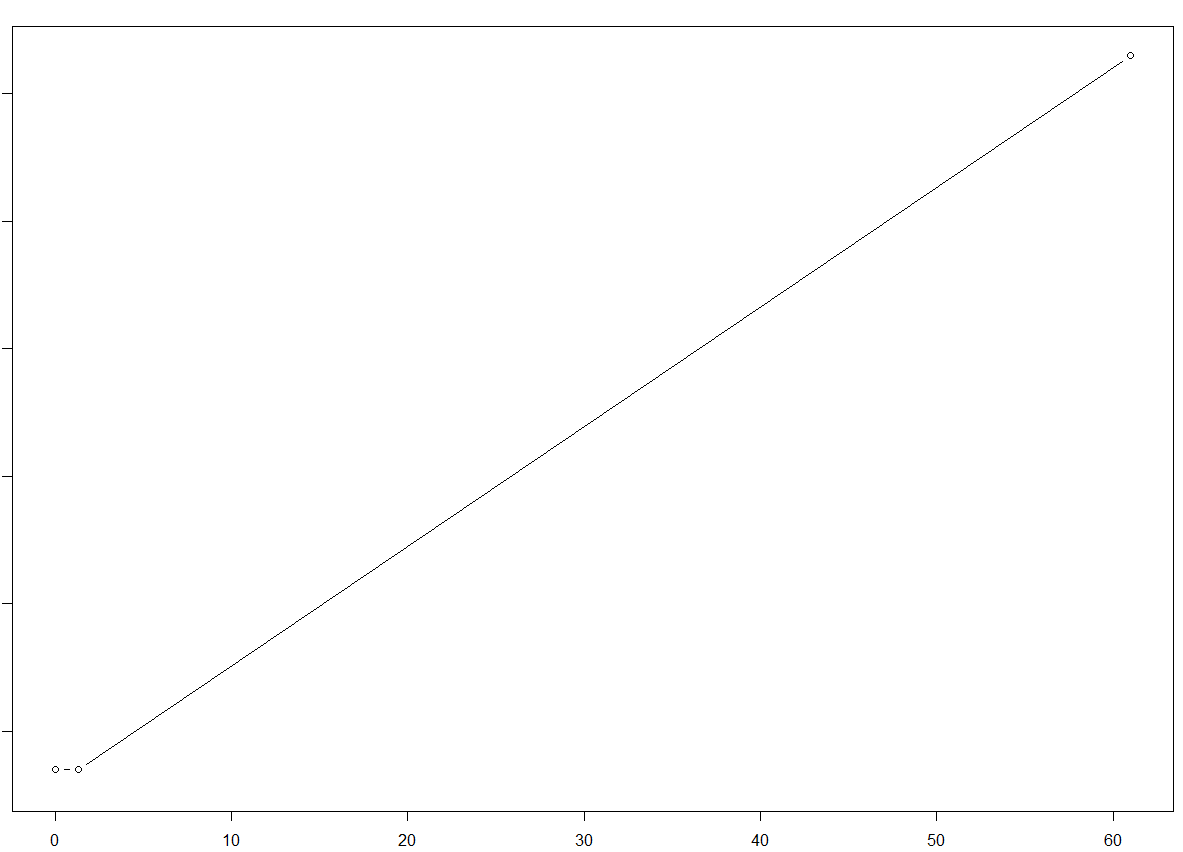
Στη συνέχεια, θα πρέπει να ορίσουμε ένα training set και ένα test set, έτσι ώστε να υπολογίσουμε και το test error του δέντρου.

To accuracy του δέντρου ταξινόμησης είναι 87,69%, δηλαδή ταξινόμησε στη σωστή κατηγορία το 87,69% των παρατηρήσεων του test set, 57/65.

Θα δούμε κατά πόσο μπορούμε να βελτιστοποιήσουμε το μοντέλο μας, δηλαδή να "κλαδέψουμε το δέντρο".

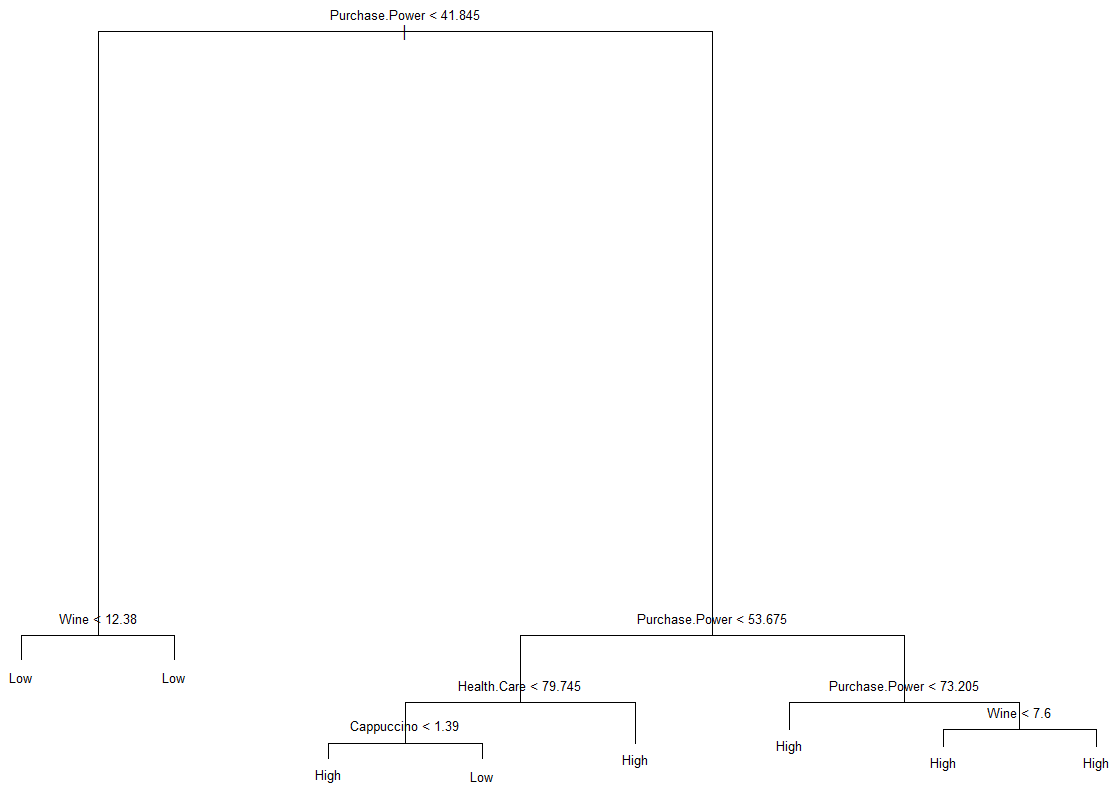


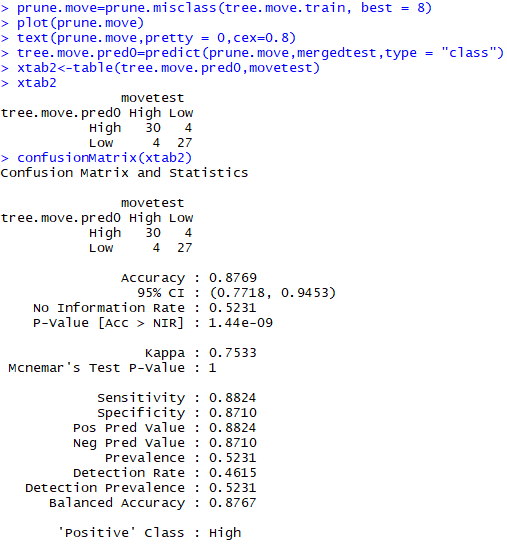
**Γραφική παράσταση ποσοστού σφάλματος - μεγέθους δέντρου.**



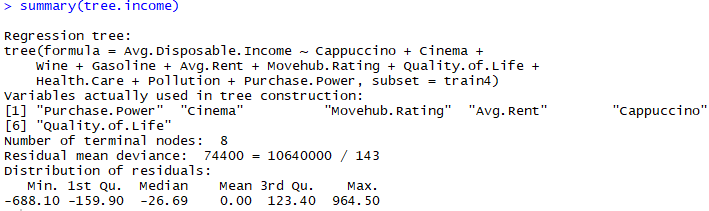
**Γραφική παράσταση κόστους πολυπλοκότητας - μεγέθους δέντρου.**

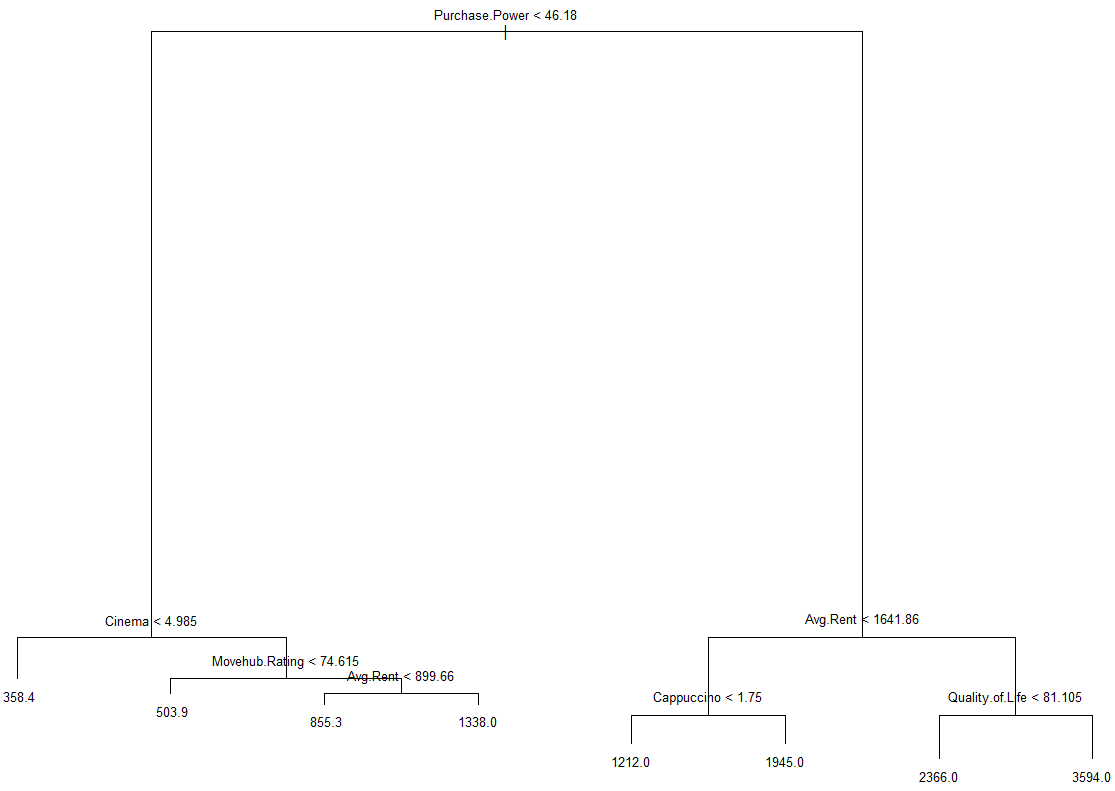
Όπως είναι εμφανές, το δέντρο με τους 8 κόμβους έχει το μικρότερο σφάλμα.

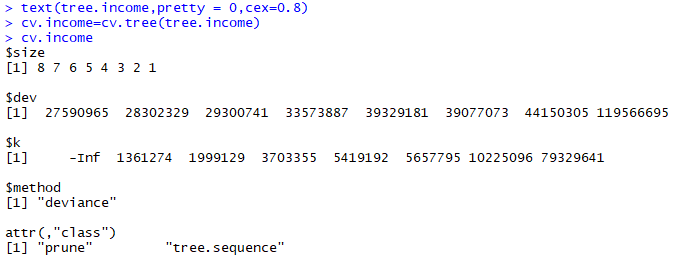
Επομένως, "κλαδεύουμε" το δέντρο στους 8 κόμβους και έχουμε το παρακάτω αποτέλεσμα.

Μετά το κλάδεμα το αποτέλεσμα του accuracy δεν άλλαξε.

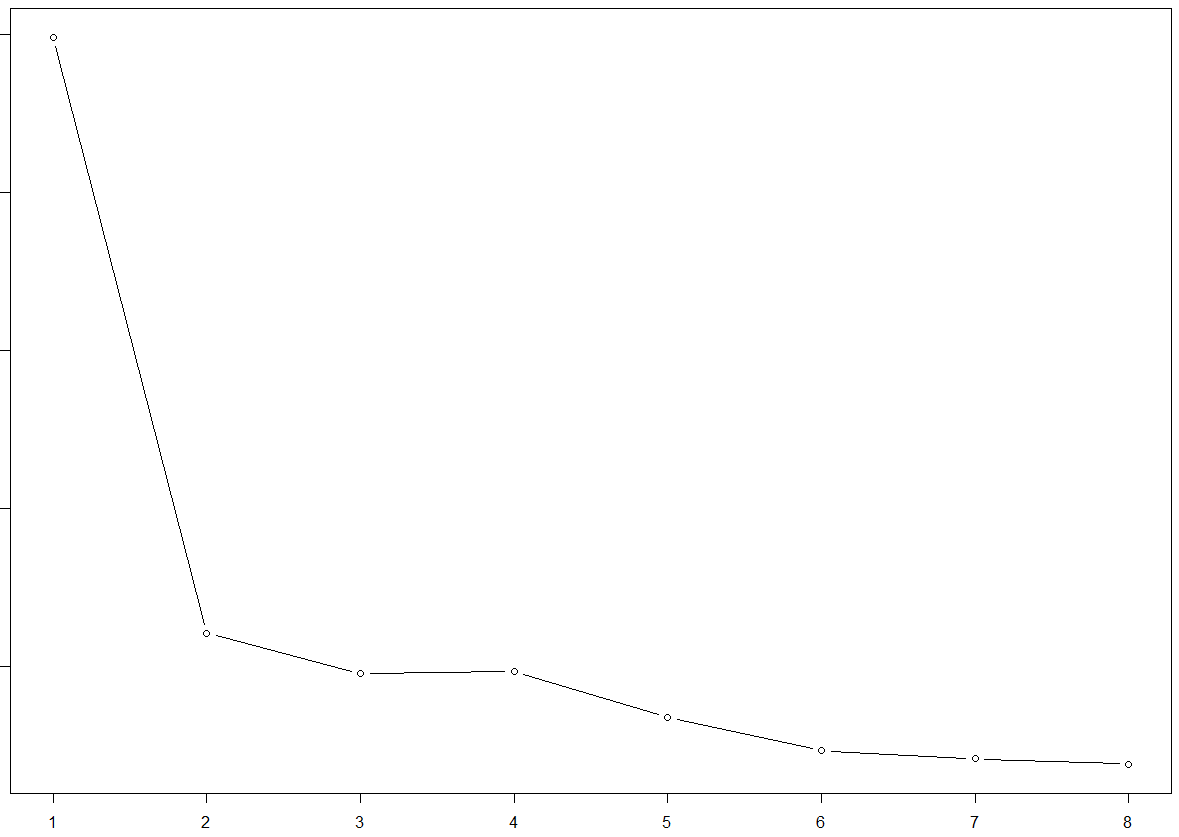
**Δέντρα Παλινδρόμησης**

Αρχικά, δημιουργούμε το δέντρο παλινδρόμησης το οποίο θα προβλέπει τη μεταβλητή Avg. Disposable Income χρησιμοποιώντας τις υπόλοιπες μεταβλητές του dataset.

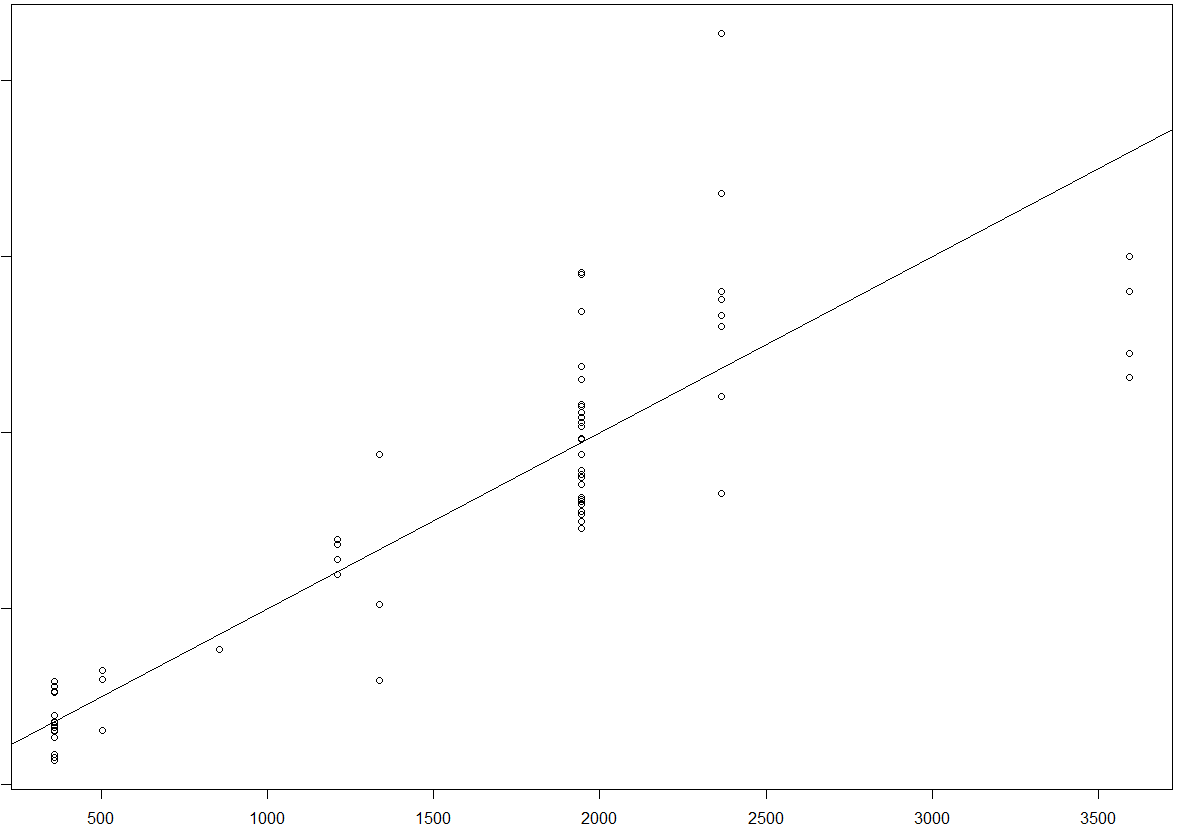
Το δέντρο έχει 8 κόμβους και γραφικά φαίνεται παρακάτω:

Στη συνέχεια, θα εξετάσουμε αν μπορούμε να βελτιστοποιήσουμε το μοντέλο μας.

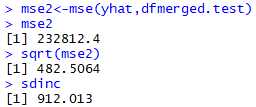
Το μικρότερο σφάλμα προκύπτει για τους 8 κόμβους, που σημαίνει ότι το δέντρο δεν μπορεί να κλαδευτεί για να δώσει καλύτερα αποτελέσματα.



**Γραφική παράσταση απόκλισης - μεγέθους δέντρου.**



**Γραφική παράσταση των πραγματικών τιμών ως προς τις τιμές που προέβλεψε το δέντρο**

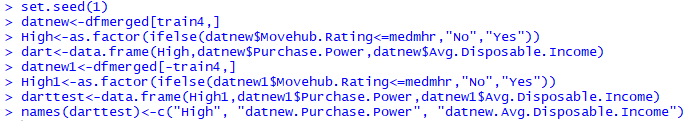
Τέλος, θα υπολογίσουμε το μέσο τετραγωνικό σφάλμα για τις προβλέψεις.

To μ.τ.σ. είναι 232,812.4. Η ρίζα του μ.τ.σ. είναι περίπου 482.5 και συγκρίνοντας το με την τυπική απόκλιση της μεταβλητής βλέπουμε πως είναι περίπου στο 50% της τυπικής απόκλισης, που σημαίνει ότι οι προβλέψεις του μοντέλου είναι αρκετά ικανοποιητικές

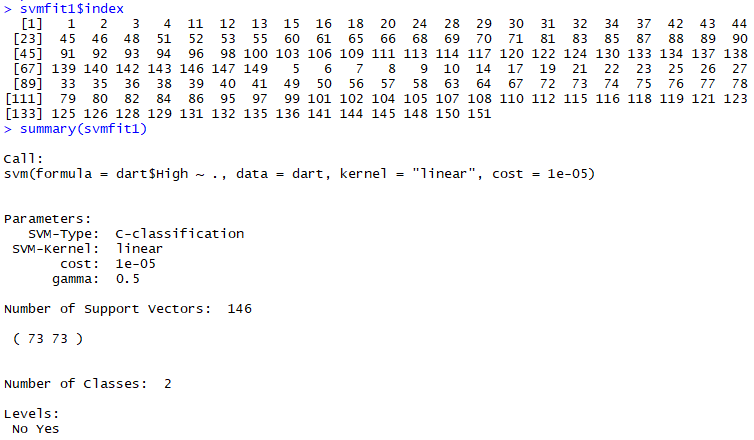
**Support Vector Machines**

Οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ, SVMs) είναι μια μέθοδος μηχανικής μάθησης. Χρησιμοποιούνται σε προβλήματα ταξινόμησης και στην προσέγγιση της μορφής της συνάρτηση σε προβλήματα παλινδρόμησης.

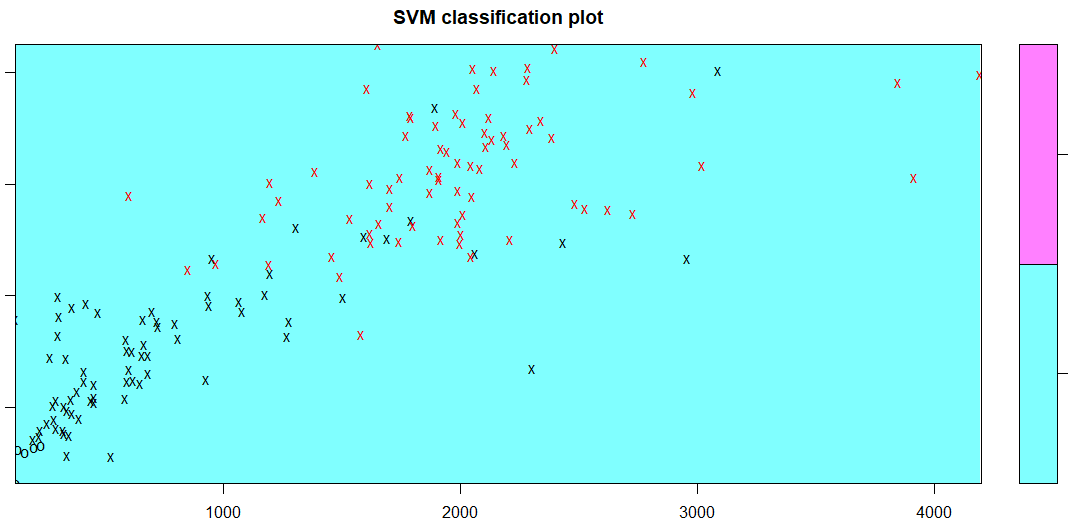
Αρχικά δημιουργήσαμε δύο νέα data frames, ένα με 151 παρατηρήσεις (70%) και θα το χρησιμοποιήσουμε ως training set και ένα με 65 παρατηρήσεις (30%) που θα χρησιμοποιηθεί ως test set. Επίσης, προσθέσαμε και μια νέα μεταβλητή High, που είναι factor με δύο levels: No, Yes, ανάλογα με το αν το Movehub Rating της εκάστοτε παρατήρησης είναι μεγαλύτερο ή όχι από τη διάμεσο.



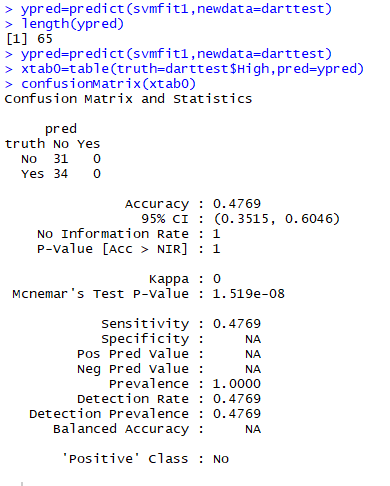
Στη συνέχεια, δημιουργήσαμε το πρώτο SVM, το οποίο θα προβλέπει τη μεταβλητή High χρησιμοποιώντας τις μεταβλητές Purchase Power και Avg. Disposable Income. Ο πυρήνας του μοντέλου είναι γραμμικός και το κόστος είναι 0.0001.

Τα διανύσματα υποστήριξης που προκύπτουν είναι 146, από 73 σε κάθε κατηγορία

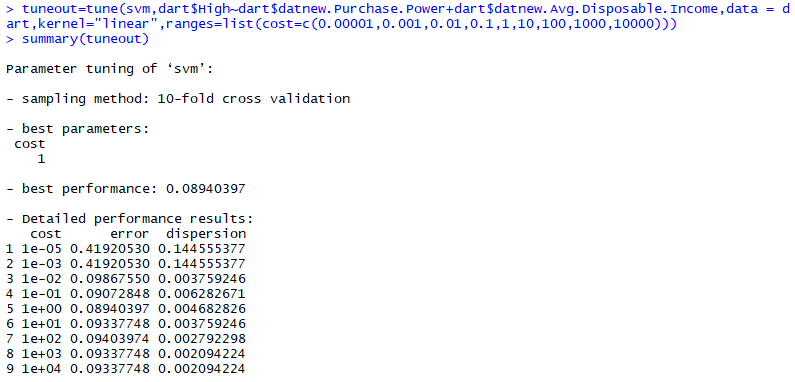
Η γραφική παράσταση του SVM:



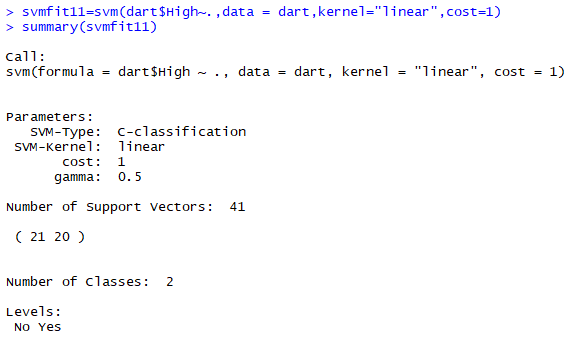
Είναι εμφανές ότι το μοντέλο δεν θα έχει πολύ καλό accuracy. Όπως φαίνεται και στον παρακάτω πίνακα, το accuracy είναι 47.69%.



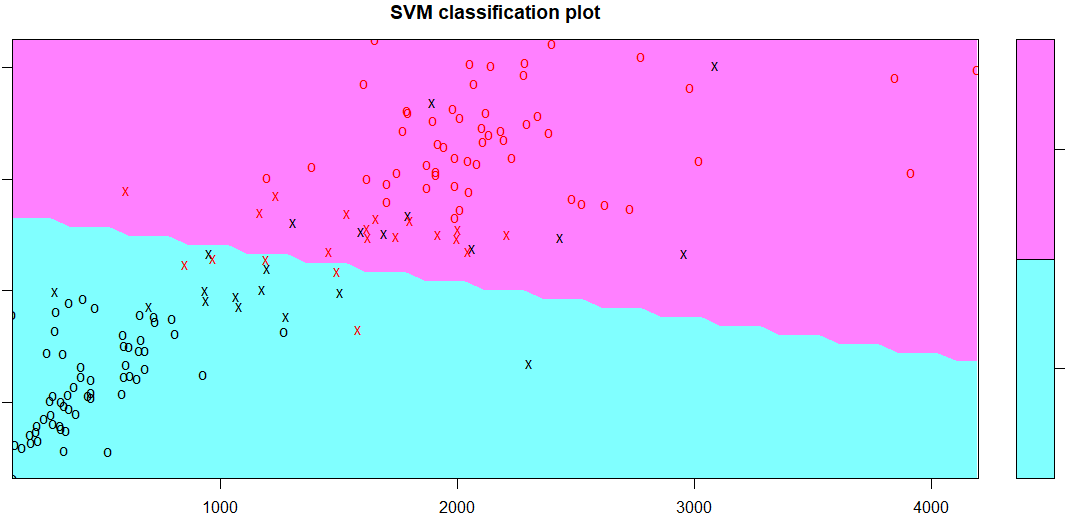
Θα προσπαθήσουμε να βελτιστοποιήσουμε το μοντέλο μας με την εντολή tune, η οποία πραγματοποιεί 10-fold validation για διάφορες τιμές κόστους.

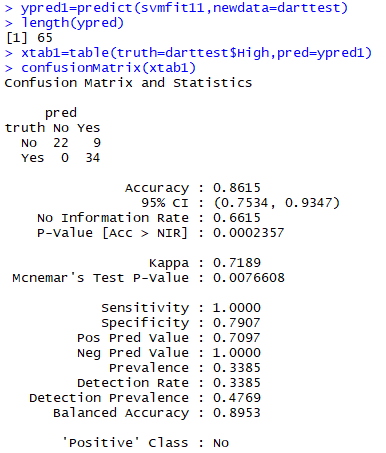


Το βέλτιστο μοντέλο, δηλαδή αυτό με το μικρότερο σφάλμα, είναι αυτό με κόστος 1.



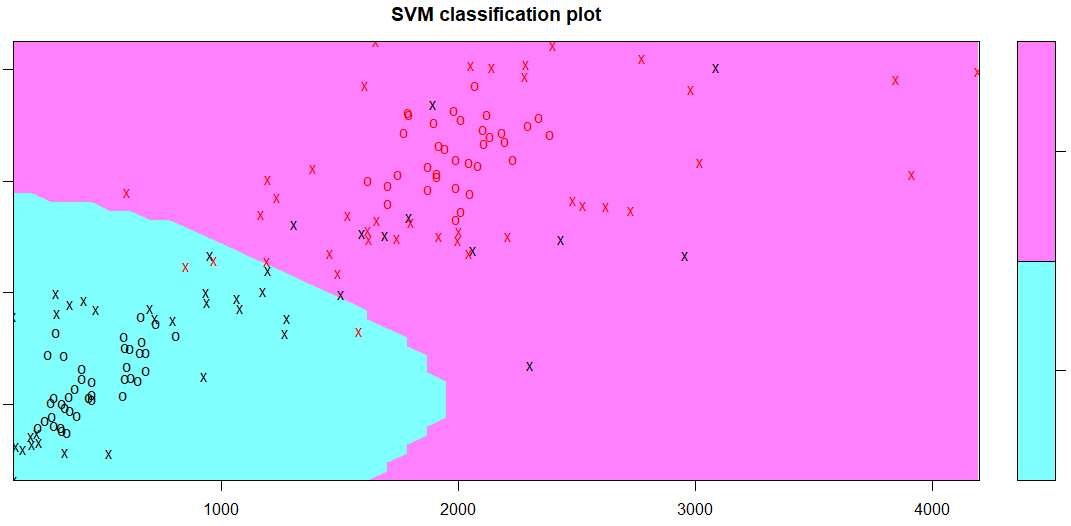
Και η γραφική του παράσταση:



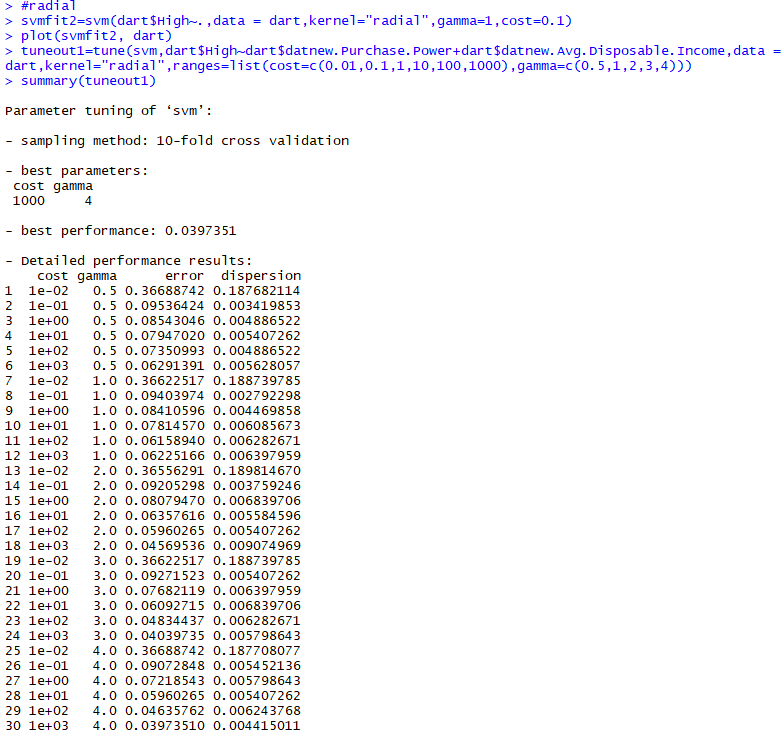
Εποπτικά, αναμένουμε σημαντική βελτίωση.

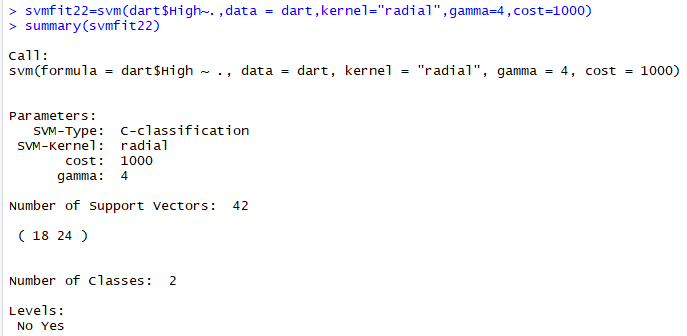
Και όπως αναμέναμε, το accuracy αυτού του μοντέλου είναι 86.15%, δηλαδή ταξινόμησε σωστά τις 56/65 παρατηρήσεις.

**Μη Γραμμικός πυρήνας**

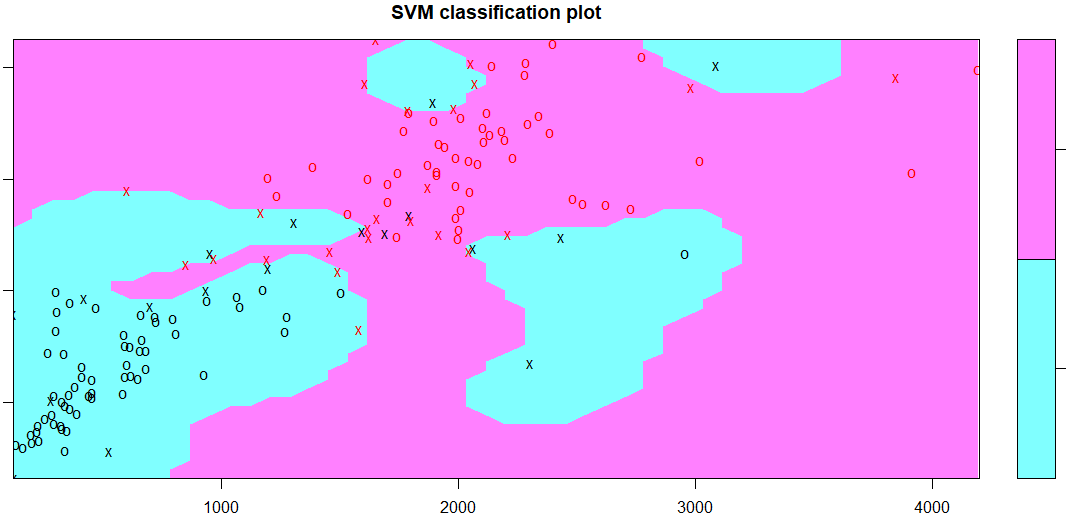
Ορίζουμε αρχικά ένα SVM με πυρήνα radial, gamma=1 και κόστος 0. 1

Και πάλι προσπαθούμε να βελτιστοποιήσουμε το μοντέλο μας και παρατηρούμε ότι το μικρότερο σφάλμα το έχουμε για κόστος 1000 και gamma=4.



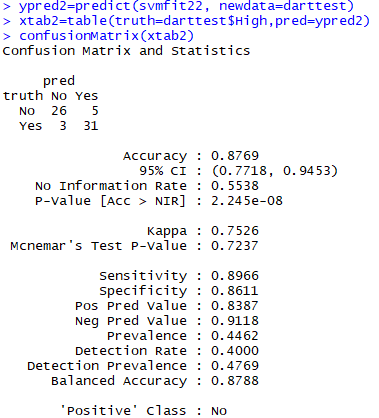


Η γραφική παράσταση του μοντέλου είναι η παρακάτω:



Ο διαχωρισμός φαίνεται καλύτερος, μένει να δούμε το accuracy.

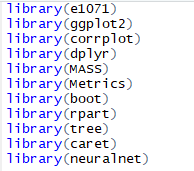
Το accuracy του βέλτιστου μοντέλου είναι 87.69%, δηλαδή ταξινομήθηκαν σωστά 57/65 παρατηρήσεις.



Τέλος, αν συγκρίνουμε το μοντέλο με το γραμμικό πυρήνα και το μοντέλο με τον ακτινικό, βλέπουμε ότι το τελευταίο δίνει λίγο καλύτερα αποτελέσματα στην ταξινόμηση.

**Παράρτημα**

* **Libraries**



* **Δείκτης Τιμών Καταναλωτή**

Στην [Οικονομική Επιστήμη](https://el.wikipedia.org/wiki/%CE%9F%CE%B9%CE%BA%CE%BF%CE%BD%CE%BF%CE%BC%CE%B9%CE%BA%CE%AE_%CE%95%CF%80%CE%B9%CF%83%CF%84%CE%AE%CE%BC%CE%B7) ο **Δείκτης Τιμών Καταναλωτή** (ΔΤΚ) ή **τιμάριθμος** είναι ο δείκτης μέτρησης του κόστους ζωής και διαβίωσης που βασίζεται στις μεταβολές των λιανικών τιμών των περισσότερων αγαθών ή υπηρεσιών. Μετρά τη διακύμανση των τιμών στα [αγαθά](https://el.wikipedia.org/wiki/%CE%91%CE%B3%CE%B1%CE%B8%CF%8C_(%CE%BF%CE%B9%CE%BA%CE%BF%CE%BD%CE%BF%CE%BC%CE%AF%CE%B1)) και τις [υπηρεσίες](https://el.wikipedia.org/wiki/%CE%A0%CE%B1%CF%81%CE%BF%CF%87%CE%AE_%CF%85%CF%80%CE%B7%CF%81%CE%B5%CF%83%CE%B9%CF%8E%CE%BD) που περιλαμβάνονται στο "[καλάθι του καταναλωτή](https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CE%BB%CE%AC%CE%B8%CE%B9_%CF%84%CE%BF%CF%85_%CE%9A%CE%B1%CF%84%CE%B1%CE%BD%CE%B1%CE%BB%CF%89%CF%84%CE%AE)".

* **Disposable Income**

**Disposable income**, επίσης γνωστό ως disposable personal income (DPI), είναι το ποσό των χρημάτων που έχει ένα νοικοκυριό για να ξοδέψει και να αποταμιέψει μετά φόρων. Αυτός ο δείκτης συχνά παρατηρείται ως ένας από τους κύριους οικονομικούς δείκτες που χρησιμοποιύνται για να μετρηθεί η συνολική κατάσταση της οικονομίας.

* Τα δεδομένα συλλέχθηκαν από το Numbeo, τη μεγαλύτερη βάση δεδομένων, παγκοσμίως, που αποτελείται από δεδομένα χρηστών που αφορούν τα κόστη και τις τιμές ανά τον κόσμο

**Βιβλίογραφία**

[1] Φ. Κολυβά- Μαχαίρα Ε. Μπόρα- Σέντα, Στατιστική, Εκδόσεις Ζήτη, 2014

[2] Torsten Hothorn, Brian S. Everitt, A Handbook of Statistical Analyses Using R, second edition, Chapman and Hall, 2009

[3] <https://www.investopedia.com/terms/d/disposableincome.asp#ixzz54AIYkgW1>

[4] https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

[5] http://blog.minitab.com/blog/statistics-and-quality-data-analysis/what-are-t- values-and-p-values-in-statistics

[6] https://www.movehub.com/

[7] http://www.bankofgreece.gr/Pages/el/Statistics/prices.aspx