ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ
ΠΜΣ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ
ΕΙΔΙΚΕΥΣΗ: ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ & ΑΝΑΛΥΤΙΚΗ

ΑΠΑΛΛΑΚΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΑΘΗΜΑ: ΠΡΟΒΛΕΠΤΙΚΗ ΑΝΑΛΥΤΙΚΗ

Διδάσκων: Φιλιππάκης Μιχαήλ

Εκπόνηση: Κέζιου Παναγιώτα

Ράπτη Χαρίκλεια

ΠΕΙΡΑΙΑΣ, ΙΟΥΝΙΟΣ 2020

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

1. ΕΚΦΩΝΗΣΗ

1.1. ΜΕΡΟΣ ΠΡΩΤΟ

1.1.1. ΑΣΚΗΣΗ 1

Υλοποιείστε μία απλή εκδοχή του K –means για δισδιάστατα δεδομένα στο [0,10] που θα χρησιμοποιεί i) την ευκλείδεια απόσταση και ii) την Manhattan απόσταση και θα επιλέγει ως κεντρικό σημείο κάθε συστάδα το μεσαίο σημείο (όχι το αριθμητικό μέσο όπως στην ευκλείδεια απόσταση). Γράψτε ένα πρόγραμμα που θα παράγει N δισδιάστατα σημεία που να ανήκουν σε m κύκλους με την ίδια ακτίνα. Ο αριθμός των σημείων N, ο αριθμός m και η ακτίνα των κύκλων θα είναι είσοδος στο πρόγραμμα σας.

i) Δημιουργήσετε 250 τυχαία σημεία και τρέξτε τον αλγόριθμο με K=10
ii) Δημιουργείστε 250 σημεία που να ανήκουν σε 5 κύκλους με την ίδια ακτίνα και ξένους μεταξύ τους. Αναθέστε ίδιο αριθμό σημείων σε κάθε κύκλο και τρέξτε τον αλγόριθμο με k=3, k=6, k=12. Δώστε μία επιλογή αρχικών σημείων που να οδηγεί σε καλά αποτελέσματα και μία επιλογή που δεν οδηγεί σε καλά αποτελέσματα.
iii) Τρέξτε τον αλγόριθμο DBSCAN για το ερώτημα (ii) θεωρώντας την ευκλείδεια απόσταση. Πειραματιστείτε με την επιλογή του eps και MinPts. Δώστε 2 διαφορετικές συσταδοποιήσεις για διαφορετικές επιλογές αυτών των τιμών.

1.1.2. ΑΣΚΗΣΗ 2

In this assignment, you are to implement the simple Logistic Regression algorithm for a single independent variable (IV) and the Perceptron algorithm for the general case of multiple independent variables. You may assume that data is provided to you in 2-column (n-column) format, where the first column(s) is(are) the IV(s) and the second(last) one is the dependent variable (DV). The first row contains the number of independent variables and the second row contains the name(s) of the variables. Each subsequent row corresponds to a single observation. Hence, you may need to aggregate the values of the DV (i.e., counts) for each value of the IV before doing the regression.

1. Implement Logistic Regression for a single IV.

- Your program must:
  - Output the values of the intercept (w0) and the slope (w1) of the model.
  - Output the value of R2 from the probabilities (i.e., compute the SSE and TSS on probabilities rather than on odds).

2. Use your algorithm on this (modified) Coronary Heart Disease.txt (CHD) problem.

- First, use simple linear regression (no logit function) to build a linear model of the data
  - Report the parameters of the model (w0, w1 and R2).
  - Plot the original (probability) data and graph the linear model your program produced.
  - What does the model predict for the probability of someone 41 years old suffering of CHD?
- Next, use logistic regression to build a nonlinear model of the data.
  - Report the parameters of the model (w0, w1 and R2).
  - Plot the original (probability) data and graph the nonlinear model your program produced.
  - What does the model predict for the probability of someone 41 years old suffering of CHD?

3. Generalize your program to handle multiple independent variables by implementing the Perceptron Algorithm

- Your program must:
  - Output the values of the perceptron weights (w0, w1, ..., wn).
  - Output the value of R2 (i.e., compute the SSE and TSS for the output of the perceptron)
- Use your algorithm on the CHD problem.
  - Report the parameters of the model (weights and R2).
  - Plot the SSE over time for 100 epochs of training for a learning rate that is too large (no convergence).
  - Plot the SSE over time for 100 epochs of training for a learning rate that is small enough (asymptotic error).
  - Plot the original (probability) data and graph the nonlinear model your program produced.

- How do the three models you produced for the CHD problem compare?

- Use your algorithm on this (modified) Iris problem.
  - Report the parameters of the model (weights and R2).
  - Plot the SSE over time for 1000 epochs of training for a learning rate that is too large (no convergence).
  - Plot the SSE over time for 1000 epochs of training for a learning rate that is small enough (asymptotic error).
  - Report the predictions your model makes for the following data (iris1.txt)

1.1.3. ΑΣΚΗΣΗ 3

Θέλουμε να μελετήσουμε τους παράγοντες που προκαλούν το διαβήτη. Δημιουργείστε το κατάλληλο μοντέλο πρόβλεψης και κάντε ανάλυση προς κάθε κατεύθυνση, ποιες μεταβλητές είναι σημαντικές?
Ερμηνεύστε τα αποτελέσματα
Να προβλέψετε τα αποτελέσματα για τα επόμενα 4 άτομα

Πίνακας 1. Δεδομένα άσκησης 3

	Npreg	***Glu***	***BP***	***BMI***	***SKIN***	***PED***	***Age***
1	*5*	*140*	*76*	*70*	*20*	*0.6*	*40*
2	*1*	*80*	*70*	*25*	*45*	*0.56*	*25*
3	*8*	*120*	*60*	*27*	*30*	*0.5*	*44*
4	*2*	*91*	*50*	*68*	*23*	*0.7*	*34*

1.1.4. ΑΣΚΗΣΗ 4

Τα στοιχεία του πίνακα παρουσιάζουν τις πωλήσεις μπίρας μίας βιομηχανίας (σε εκατομμύρια μπουκάλια) ανά περιοχή από το 2016μέχρι το 2019

Πίνακας 2. Δεδομένα άσκησης 4

Ετος	***Χειμώνας***	***Άνοιξη***	***Καλοκαίρι***	***Φθινόπωρο***
2016	*1*	*3*	*6*	*4*
2017	*2*	*2*	*7*	*5*
2018	*2*	*4*	*8*	*5*
2019	*1*	*3*	*8*	*6*

i) Να γίνει εξάλειψη της εποχικότητας και να υπολογιστούν οι εποχιακές δείκτες αυτών των δεδομένων
ii) Να κατασκευάσετε τη γραμμή τάσης
iii) Να προσδιορίσετε την κυκλική μεταβολή με τη μέθοδο των σχετικών κυκλικών καταλοίπων

1.1.5. ΑΣΚΗΣΗ 5

Στον επόμενο πίνακα δίνονται οι εισπράξεις από τις ετήσιες πωλήσεις (σε χιλ. ευρώ) μία βιομηχανίας.

Πίνακας 3. Δεδομένα άσκησης 5

Έτος	*Έσοδα*
2011	*37,44*
2012	*44,14*
2013	*46,25*
2014	*43,99*
2015	*51,84*
2016	*49,10*
2017	*58,56*
2018	*58,02*
2019	*70,28*

i) Να γίνει εξάλειψη της εποχικότητας και να υπολογιστούν οι εποχιακές δείκτες αυτών των δεδομένων
ii) Να κατασκευάσετε τη γραμμή τάσης
iii) Να προσδιορίσετε την κυκλική μεταβολή με τη μέθοδο των σχετικών κυκλικών καταλοίπων

1.1.6. ΑΣΚΗΣΗ 6

www.bankofcanada.ca

The daily USD/CAD exchange rates for the last 2 years

I. Plot the time series
II. Βρείτε την ευθεία της τάσης με γραμμική παλινδρόμηση και να κάντε το κοινό γράφημα.
III. Βρείτε quadratic trend και κάντε πάλι το γράφημα

IV. Υπολογίστε τις πρώτες και δεύτερες διαφορές και κάντε τα αντίστοιχα γραφήματα

V. Κάντε μία πρόβλεψη με ένα από τα παραπάνω μοντέλα (4) για τις επόμενες 22 μέρες.

VI. Προσθέστε μία εποχική συνιστώσα (π.χ. την επίδραση του Ιανουαρίου) στην γραμμική τάση (φτιάξτε μία binary μεταβλητή που θα δείχνει αν το exchange rate είναι από το Γενάρη ή όχι) Κάντε fit το μοντέλο με την επίδραση του Γενάρη.
(hint : χρησιμοποιείστε τη συνάρτηση format to convert the dates to strings)

VII. Κάνετε fit a quadratic trend θεωρώντας και την επίδραση του Γενάρη

VIII. Κανετε το γράφημα διασποράς μεταξύ των τιμών exchange rates and the lagged values

IX. Υπολογίστε την αυτοσυσχέτιση r_1 υστέρησης 1. Τι παρατηρείτε?

X. Φτιάξτε μια συνάρτηση (ονομάστε την cal AC(y,k) που θα υπολογίζει την αυτοσυσχέτιση υστέρισης k γενικά για μία χρονοσειρά

XI. Στα παραπάνω δεδομένα φτιάξτε το AR(1) και MA(1) μοντέλο

AR(1)=ARIMA(1,0,0)

MA(1)=ARIMA(0,0,1)

XII. Συγκρίνετε τα παραπάνω δύο μοντέλα με το ARIMA(2,1,2).
Υπολογίστε τα MAE (MEAN ABSOLUTE ERROR), mse, mape

1.1.7. ΑΣΚΗΣΗ 7

Δεδομένα iris

Αναλύστε τα δεδομένα κάντε PCA με πλήρη ανάλυση προς κάθε κατεύθυνση με τον πίνακα συνδιακυμάνσεων

1.2. ΜΕΡΟΣ ΔΕΥΤΕΡΟ

1.2.1. ΑΣΚΗΣΗ 1

Εισάγετε το πακέτο rpart. Φορτώστε τα δεδομένα «kyphosis». Περιγράψτε το dataset. Κάντε το θηκόγραμμα για την μεταβλητή number και μετά να βρείτε τα outliers (τις τιμές τους). Σε ποιες γραμμές αντιστοιχούν τα συγκεκριμένα δεδομένα (**%in%-which**). Επαναλάβετε το τελευταίο με τη συνάρτηση (**Identify**).

Υπόδειξη: Η συνάρτηση **identify** παίρνει τις τιμές των τετμημένων και τεταγμένων από το scatterplot ως ορίσματα.

1.2.2. ΑΣΚΗΣΗ 2

Θεωρείστε το dataset capital.csv.

i. Να γίνουν οι γραφικές παραστάσεις της Balance σε σχέση με την Gender (πίνακας σχετικών συχνοτήτων-ραβδόγραμμα-πίττα)
ii. Να γίνει το θηκόγραμμα των δεδομένων μας και τα θηκογράμματα σε σχέση με το gender
iii. Να υπολογιστούν τα μέτρα κεντρικής τάσης και απόκλισης
iv. Εξετάστε αν τα δεδομένα μας προέρχονται από κανονική κατανομή (πχ. Κάντε Q-Q-plot)

1.2.3. ΑΣΚΗΣΗ 3

Θεωρείστε τα δεδομένα του αρχείου mtcars που ακολουθούν κανονική κατανομή. Να βρεθεί το διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης 0.95 για τη διαφορά των μέσων που αντιστοιχούν στις μεταβλητές κατανάλωσης καυσίμου για το μηχανικό και αυτόματο αυτοκίνητο. (εφαρμόστε την συνάρτηση **t.test()**)

1.2.4. ΑΣΚΗΣΗ 4

Θεωρείστε το dataset OctopusF.txt. Διαβάστε τα δεδομένα, υπολογίστε τα περιγραφικά μέτρα του δείγματος (μέση τιμή, τυπική απόκλιση) . Κατασκευάστε το ιστόγραμμα. Ελέγξτε τη κανονικότητα των δεδομένων και κατασκευάστε το διάστημα εμπιστοσύνης.

1.2.5. ΑΣΚΗΣΗ 5

Φορτώστε τη βιβλιοθήκη MASS και στο αρχείο δεδομένων survey. Η στήλη smoke δείχνει τον βαθμό καπνίσματος των φοιτητών, ενώ η στήλη Exer σημειώνεται το επίπεδο σωματικής τους άσκησης. Τα επίπεδα καπνίσματος είναι «Heavy", "regul",

“Occas” “Never”. Για τη μεταβλητή Exer αυτά είναι «Freq”, “Some”, “None”. Να εξετάσετε πόσο ο βαθμός καπνίσματος επηρεάζει τη σωματική άσκηση

Υπόδειξη: Πίνακας συνάφειας και chi^2 test

1.2.6. ΑΣΚΗΣΗ 6

Τα δεδομένα μας είναι αποθηκευμένα ως Concrete_Data.xls. και αναφέρονται σε μεταβλητές που επηρεάζουν την ανθεκτικότητα του τσιμέντου. Η ανθεκτικότητα του τσιμέντου είναι μη γραμμική συνάρτηση των μεταβλητών ηλικίας και διαφόρων συστατικών όπως, blast furnace slag, fly ash, water, super-plasticizer, coarse aggregate. Οι πρώτες 8 είναι ανεξάρτητες ποσοτικές ενώ η Concrete compressive strength είναι η εξαρτημένη. Χρησιμοποιήστε κάποια πακέτα ώστε να εκπαιδεύσετε το νευρωνικό δίκτυο π.χ. το neuralnet, nnet, RSNNS.

Κάνετε ανάγνωση των δεδομένων. Στη συνέχεια κάνετε τυποποίηση των δεδομένων σας. Μετά δημιουργείστε τα σύνολα εκπαίδευσης και ελέγχου, Εκπαιδεύστε το μοντέλο σας, κάνετε τη γραφική αναπαράσταση του νευρωνικού και αξιολογήστε το. (χρησιμοποιήστε τη συνάρτηση compute() και δείτε αν λειτουργεί διαφορετικά και γιατί από τη συνάρτηση predict(). Δείτε τι κάνει η συνάρτηση cor(). Βελτιώστε το μοντέλο σας αν γίνετε και δείτε πως επηρεάζεται η συμπεριφορά του μοντέλου σας αν αυξηθεί ο αριθμός των κρυφών κόμβων.

1.2.7. ΑΣΚΗΣΗ 7

Θεωρείστε το σύνολο δεδομένων faithful και εκτιμήστε την επόμενη έκρηξη αν ο αναμενόμενος χρόνος από την τελευταία έκρηξη είναι 80 λεπτά. Να βρεθεί ο συντελεστής προσδιορισμού και η ευθεία παλινδρόμησης. Ελέγξτε αν υπάρχει στατιστικά σημαντική σχέση μεταξύ των δύο μεταβλητών στο μοντέλο της ευθείας παλινδρόμησης για τις δύο μεταβλητές των δεδομένων σας με επίπεδο σημαντικότητας α=0.05. Κατασκευάστε ένα 95% Δ.Ε. για την μεταβλητή eruption duration για δεδομένο χρόνο αναμονής 80 min. Κατασκευάστε ένα 95% διάστημα πρόβλεψης της μεταβλητής eruption duration δεδομένο χρόνο αναμονής 80 min. Επιπλέον να γίνει το διάγραμμα υπολοίπων της γραμμικής παλινδρόμησης έναντι της μεταβλητής waiting. Να γίνει το διάγραμμα της κανονικότητας για τα τυποποιημένα υπόλοιπα της γραμμικής παλινδρόμησης.

1.2.8. ΑΣΚΗΣΗ 8

Θεωρείστε τα δεδομένα stackloss και εκτιμήστε την τιμή της stack loss αν η τιμή της airflow είναι 72 και η τιμή της water temperature είναι 20 και της air concentration

είναι 85. Να βρεθεί ο συντελεστής προσδιορισμού για το πολλαπλό γραμμικό μοντέλο. Να εξετάσετε την σημαντικότητα του μοντέλου στα δεδομένα μας με επίπεδο σημαντικότητας α=0.05. Να γίνει το 95% Δ.Ε. για το μέσο της εξαρτημένης μεταβλητής όταν έχουμε τιμές για τις μεταβλητής air flow ίσο με 72, water temperature ίσο με 20 και acid concentration ίσο με 85, επίσης να γίνει το 95% Δ.Ε. πρόβλεψης για την εξαρτημένη μεταβλητή όταν έχουμε τιμές για τις μεταβλητής air flow ίσο με 72, water temperature ίσο με 20 και acid concentration ίσο με 85.

1.2.9. ΑΣΚΗΣΗ 9

Θέλουμε να προβλέψουμε ένα μοντέλο πρόβλεψης της οργανικής περιεκτικότητας σε άνθρακα στο έδαφος από φασματοσκοπικές μετρήσεις. Τα δεδομένα μας αποτελούνται από 177 χρώματα (άτομα) και την οργανική τους περιεκτικότητα σε άνθρακα (OC μεταβλητή απόκρισης), τα φάσματα που λαμβάνουμε στην ορατή και στην εγγύς υπέρυθρη σειρά (400nm-2500nm) η οποία δίνει 2101 εξηρτημένες μεταβλητές. Να εφαρμοστεί η μέθοδος μερικών ελαχίστων τετραγώνων με παλινδρόμηση και να γίνει ανάλυση υπολοίπων. Spectrum_Breizh.txt

1.2.10. ΑΣΚΗΣΗ 10

Ο αντιπρόσωπος πωλήσεων ενός προϊόντος δεν είναι ικανοποιημένος από τις πωλήσεις του προϊόντος στη ζώνη ευθύνης του. Διαπιστώνει ότι τα προϊόντα που πουλήθηκαν διαφέρουν από κατάστημα σε κατάστημα σε ένα εύρος τιμών από 921 έως 2604 τεμάχια με μέση τιμή 1846,8. Θέλει να γνωρίζει που οφείλεται αυτή η διαφορά και για αυτό επιλέγει τυχαία 37 καταστήματα ίδιου μεγέθους παρατηρώντας τις πωλήσεις του προϊόντος, την τιμή του προϊόντος, τα έξοδα πωλήσεων και το πλήθος των αφίξεων στα εν λόγω καταστήματα για μία ορισμένη χρονική περίοδο. Η παρακάτω έρευνα πρέπει να απαντήσει στο ερώτημα αν οι πωλήσεις σετ-τεμάχια του προϊόντος επηρεάζονται από τα υπόλοιπα μεγέθη που επέλεξε.

Εξετάστε αν:

i. Το μοντέλο σας είναι στατιστικά σημαντικό
ii. Αν τα μεγέθη είναι στατιστικά σημαντικά
iii. Αν υπάρχει αλληλεξάρτηση μεταξύ των μεγεθών

Να υπολογιστούν τα τυποποιημένα υπόλοιπα και μη και να βρεθεί το μικρότερο και μεγαλύτερο υπόλοιπο

Να εξεταστεί η κατανομή των υπολοίπων

i. Να γραφεί η σχέση πρόβλεψης για τις πωλήσεις. Που γίνονται οι καλύτερες προβλέψεις.
ii. Γράψετε τη θεωρία που αφορά την ομοιοσκεδαστικότητα και εξετάστε αν υπάρχει τέτοια στο μοντέλο σας

1.2.11. ΑΣΚΗΣΗ 11

Θεωρούμε τα δεδομένα insurance.csv που περιέχει 1338 δεδομένα με τα χαρακτηριστικά

i. Age:ηλικία του ασφαλισμένου
ii. Sex
iii. Bmi: Δείκτης μάζας του σώματος (BMI= βάρος σε κιλά του ασφαλισμένου προς το τετράγωνο του ύψους), ένας ιδεατός δείκτης είναι μεταξύ 18,5 και 24,9
iv. Children: αριθμός παιδιών που συμμετέχουν στο πρόγραμμα του κυρίως ασφαλισμένου
v. Smoker
vi. Region: Η περιοχή διαμονής του ασφαλισμένου: northeast, southeast, southwest or northwest

Σκοπός μας είναι να μελετηθεί πως αυτές οι μεταβλητές επηρεάζουν τα έξοδα.

Κάντε

i. Ανάγνωση των δεδομένων
ii. Έλεγχο των μεταβλητών (χρησιμοποιήστε το πακέτο “psych”)
iii. Δημιουργείστε το κατάλληλο μοντέλο ώστε να απαντήσετε το παραπάνω ερώτημα

1.2.12. ΑΣΚΗΣΗ12

Η MF είναι μία εταιρεία ηλεκτρονικών ειδών. Αν η εταιρεία έχει καλό σύστημα ποιοτικής διασφάλισης των προϊόντων της υπάρχουν περιπτώσεις επιστροφής. Στη διάρκεια λειτουργίας της δημιούργησε μία βάση δεδομένων. Στην πρώτη στήλη υπάρχουν οι αποζημιώσεις για κάθε επιστροφή, στη δεύτερη οι βάρδιες κατασκευής των εμπορευμάτων στην τρίτη τα είδη των παραπόνων και στην τέταρτη το μέρος παραγωγής. Ο υπεύθυνος ποιότητας της εταιρίας μελέτησε ένα τυχαίο δείγμα 110 επιστροφών (mf.xls)

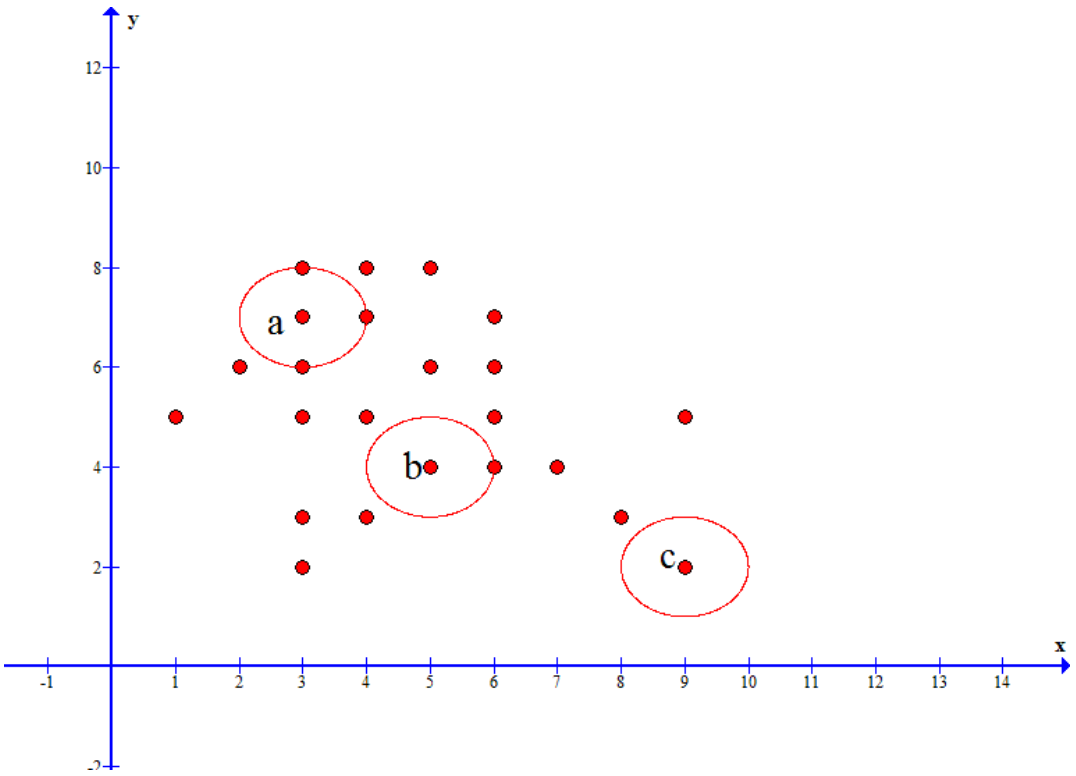
i. Παρουσιάστε τα δεδομένα με πίνακες συνάφειας εξετάζοντας τα μεγέθη ανά δύο (αποζημιώσεις έναντι των υπολοίπων καθώς και τύπος παραπόνων έναντι μίας βάρδιας ή τόπος παραγωγής). Σχολιάσατε τα δεδομένα.

ii. Ο υπεύθυνος θέλει να δώσει μία απάντηση στην υποψία του ότι υπάρχει σχέση μεταξύ των τύπων παραπόνων και του τόπου παραγωγής. Υπολογίσετε τις αναμενόμενες τιμές σε κάθε κελί. Υποδείξτε ένα τρόπο ώστε σε κάθε κελί να υπάρχει αναμενόμενη τιμή 5. Ισχύει η υποψία του υπευθύνου σε επίπεδο σημαντικότητας 0.01?

iii. Ο υπεύθυνος θέλει να γνωρίζει αν υπάρχει διαφορά στο ύψος των αποζημιώσεων μεταξύ των τόπων παραγωγής (Boise kai Salt lake city). Ελέγξτε το σε επίπεδο σημαντικότητας α=0,02.

1.2.13. ΑΣΚΗΣΗ 13

Εφαρμόστε τον αλγόριθμό DBSCAN με ε=1 και MinPts=3. Σε ποια κατηγορία ανήκουν τα σημεία a,b,c? Εφαρμόστε τον K-means και βρείτε το καλύτερο δυνατό πλήθος συστάδων k.



1.2.14. ΑΣΚΗΣΗ 14

Δίνεται ο πίνακας :

	X	Y
1	0,4005	0,5306
2	0,2148	0,3854
3	0,3457	0,3156
4	0,2652	0,1875
5	0,0789	0,4139
6	0,4548	0,3022

Βρείτε το δενδρόγραμμα και τις ακριβείς τιμές των αποστάσεων, στις οποίες δημιουργούνται οι αντίστοιχες συστάδες κατά την συσσωρευτική ιεραρχική συσταδοποίηση, όπως αυτά προκύπτουν από την εφαρμογή της Ευκλείδειας απόστασης σε συνδυασμό με την τεχνική του: α) απλού συνδέσμου, και β) πλήρους συνδέσμου

1.3. ΜΕΡΟΣ ΤΡΙΤΟ

Task 1:

Υλοποιήστε το SVM μοντέλο με χρήση της ‘fitcsvm’ μεθόδου

Task 2:

Αξιολογήστε το μοντέλο που προπονήσατε στο προηγούμενο βήμα χρησιμοποιώντας τη 'predict' μέθοδο και υπολογίζοντας την απόκλιση από τις πραγματικές ετικέτες με όποιο τρόπο επιθυμείτε.

Task 3:

Επαναλάβετε τα προηγούμενα βήματα για διαφορετικά KerneIs.

Task 4:

Δημιουργείστε μια δίκης αρχιτεκτονική για ένα CNN μοντέλο με τη βοήθεια του Deep Learning TooIbox του MATLAB και προσδιορίστε οποιαδήποτε άλλο χαρακτηριστικό του (π.χ. optimizer) χρειάζεται. Θυμηθείτε ότι ένα νευρωνικό με πολλά layers μπορεί να είναι καλύτερος classifier αλλά έχει μεγαλύτερες υπολογιστικές ανάγκες. Επομένως καλείσθε να βρείτε τη χρυσή τομή για τη λύση αυτού του προβλήματος.

2. ΕΠΙΛΥΣΗ

2.1. ΜΕΡΟΣ 1ο

2.1.1. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 1

Αρχικά, ορίσαμε τη συνάρτηση K-means που χρησιμοποιεί την ευκλείδεια απόσταση και τη συνάρτηση K-means που χρησιμοποιεί την Manhattan απόσταση. Η συνάρτηση K-means δέχεται ως ορίσματα τον αριθμό των συστάδων και το σύνολο δεδομένων. Ο αλγόριθμος K-means είναι ένας επαναλαμβανόμενος αλγόριθμος που προσπαθεί να χωρίσει το σύνολο δεδομένων σε Κ προκαθορισμένες, διακριτές και μη επικαλυπτόμενες συστάδες όπου κάθε σημείο από τα δεδομένα ανήκει σε μία μόνο συστάδα. Βασικός σκοπός είναι τα σημεία εντός μίας συστάδας να είναι όσο το δυνατόν πιο παρόμοια, διατηρώντας ταυτόχρονα τις συστάδες όσο το δυνατόν πιο διαφορετικές. Εκχωρεί σημεία σε μία συστάδα έτσι ώστε το άθροισμα της τετραγωνικής απόστασης μεταξύ των σημείων και του κέντρου της συστάδας να είναι στο ελάχιστο. Όσο λιγότερη μεταβολή έχουμε ανάμεσα στις συστάδες τόσο πιο ομοιογενή είναι τα σημεία που βρίσκονται στην ίδια συστάδα.

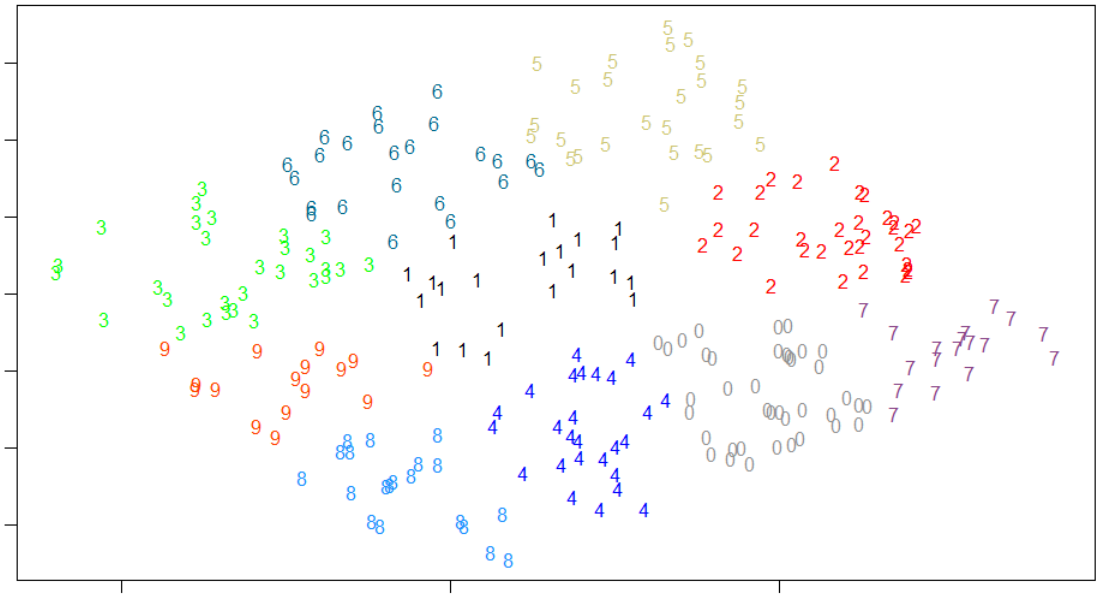
Ο αλγόριθμος K-means ακολουθεί τα εξής βήματα:

- Καθορισμός του αριθμού των συστάδων Κ.
- Αρχικοποίηση των κέντρων ανακατεύοντας πρώτα το σύνολο δεδομένων και στη συνέχεια επιλέγοντας τυχαία Κ σημεία ως κέντρα χωρίς αντικατάσταση.
- Συνέχεια των επαναλήψεων μέχρι να μην υπάρχει αλλαγή στα κέντρα.
- Υπολογισμός του αθροίσματος της τετραγωνικής απόστασης μεταξύ των σημείων και όλων των κέντρων.
- Εκχώρηση κάθε σημείου στην πλησιέστερη συστάδα.
- Υπολογισμός των κέντρων των συστάδων από το μέσο όλων των σημείων που ανήκουν σε κάθε συστάδα.

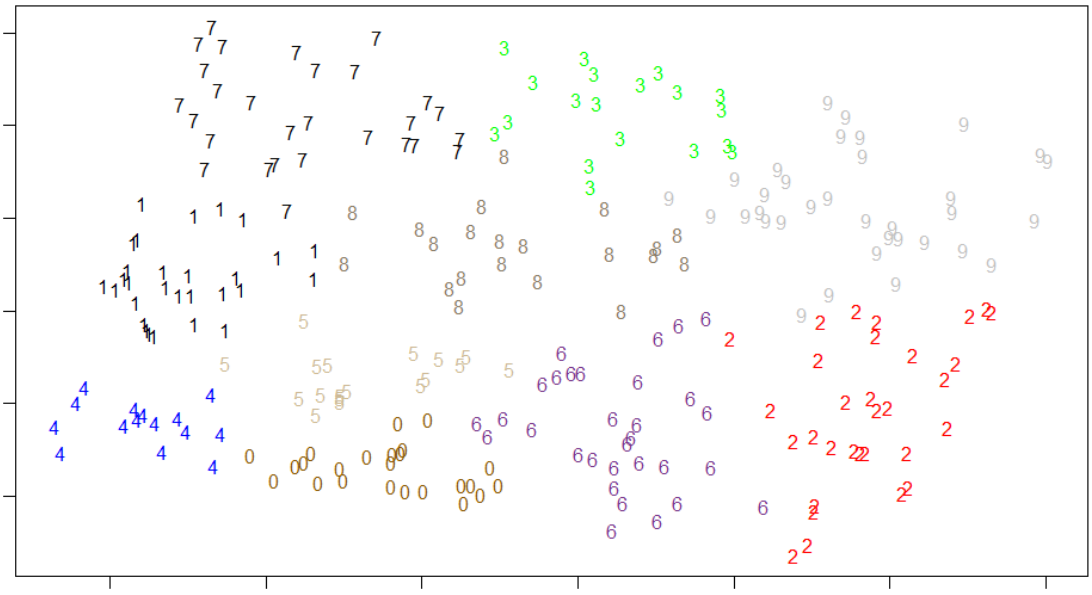
Ερώτημα i

Αφού ορίστηκαν οι δύο συναρτήσεις δημιουργήσαμε 250 τυχαία σημεία στα οποία εφαρμόσαμε τις δύο συναρτήσεις με αριθμό συστάδων K=10.

K-means με ευκλείδεια απόσταση

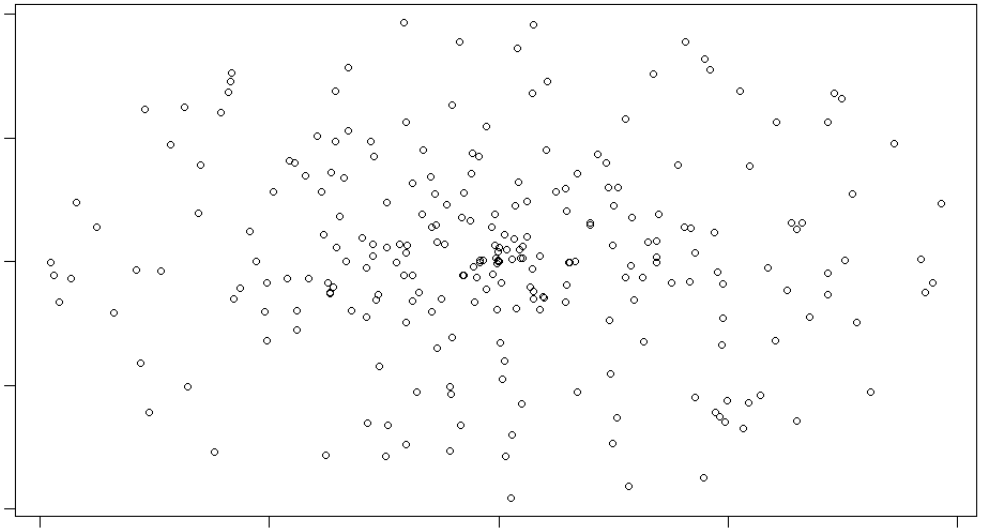


K-means με Manhattan απόσταση

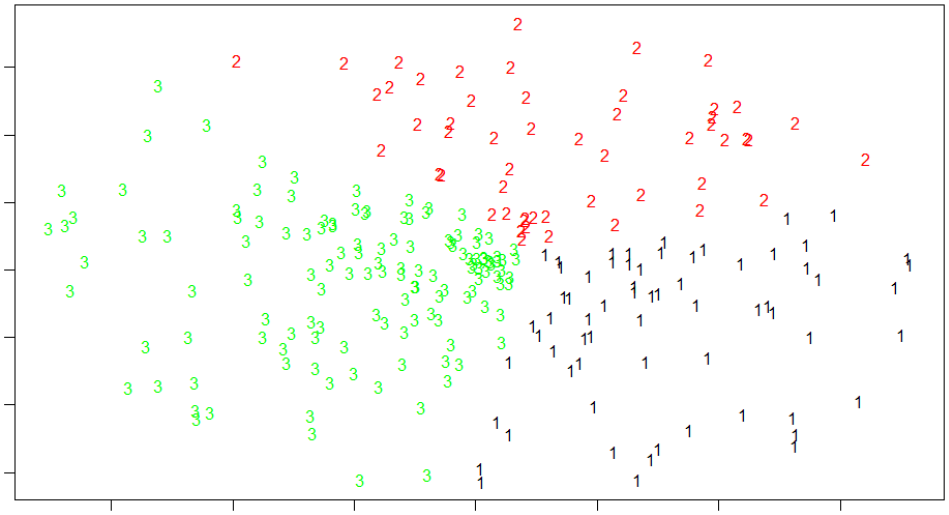


<u>Ερώτημα ii</u>

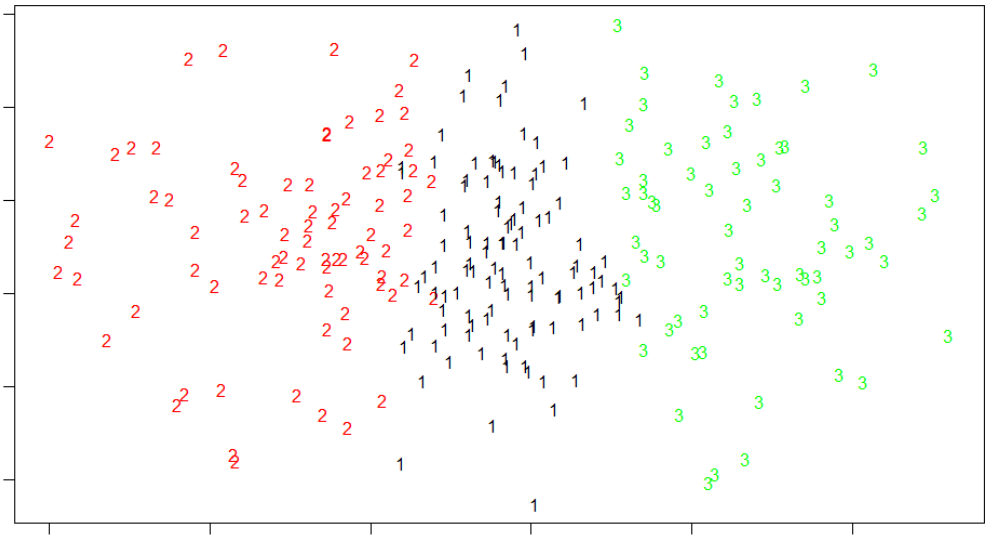
Εφόσον πρέπει να δημιουργήσουμε 250 σημείων που να ανήκουν σε 5 κύκλους με την ίδια ακτίνα όπου όλοι οι κύκλοι να έχουν τον ίδιο αριθμό σημείων, κάθε κύκλος θα έχει 50 σημεία. Τα σημεία που δημιουργήθηκαν φαίνονται στο παρακάτω σχήμα.



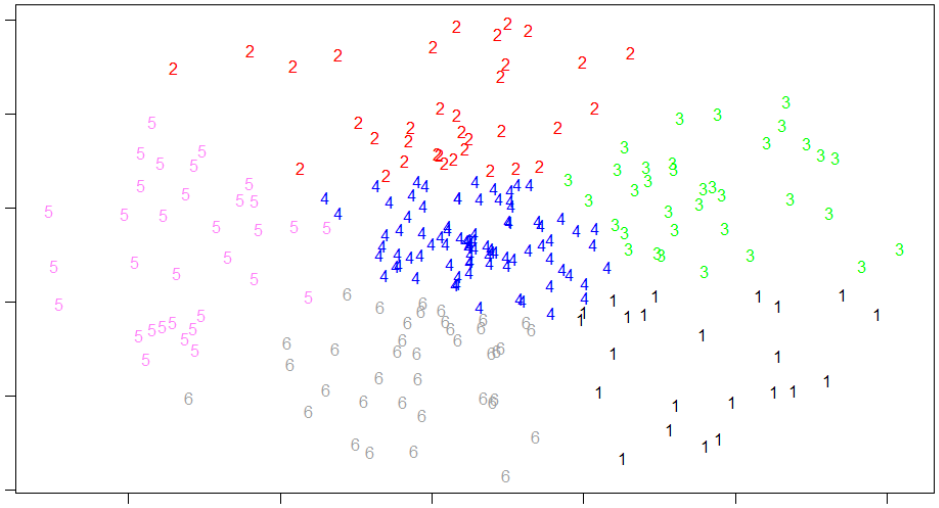
Ο αλγόριθμος K-means με την ευκλείδεια απόσταση με αριθμό συστάδων K=3



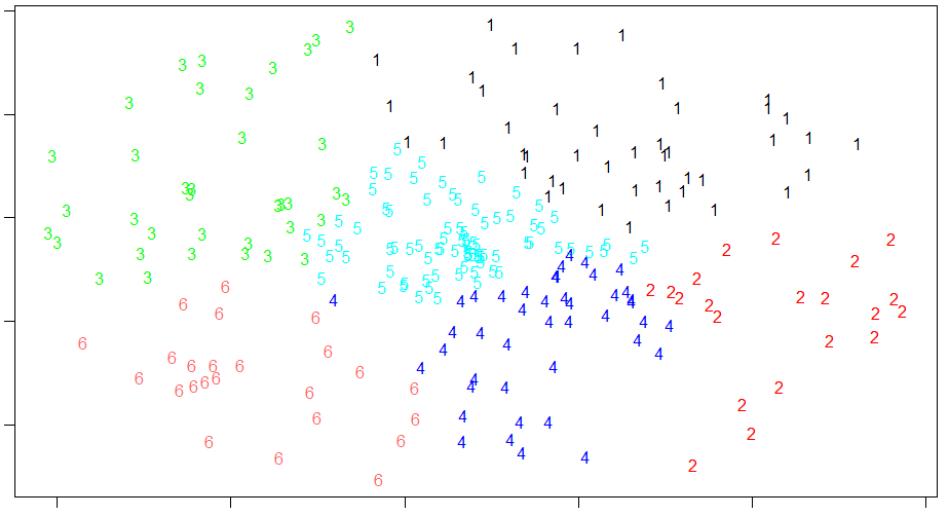
Ο αλγόριθμος K-means με την Manhattan απόσταση με αριθμό συστάδων K=3



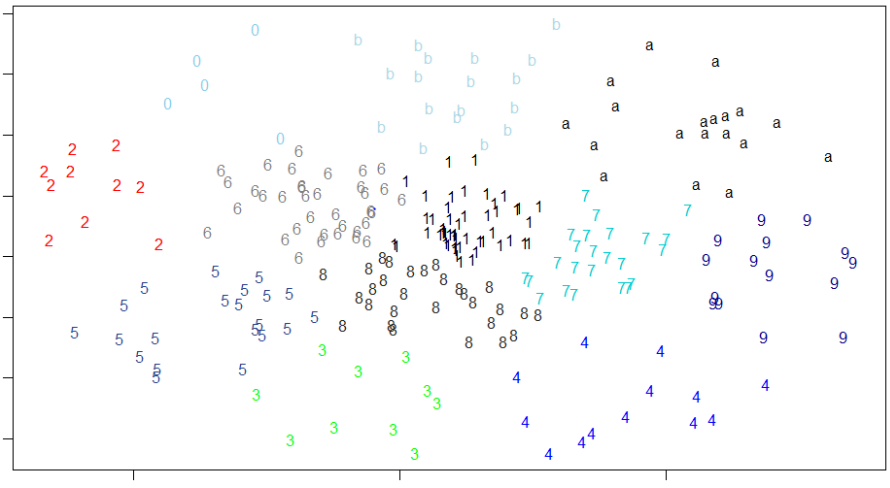
Ο αλγόριθμος K-means με την ευκλείδεια απόσταση με αριθμό συστάδων K=6



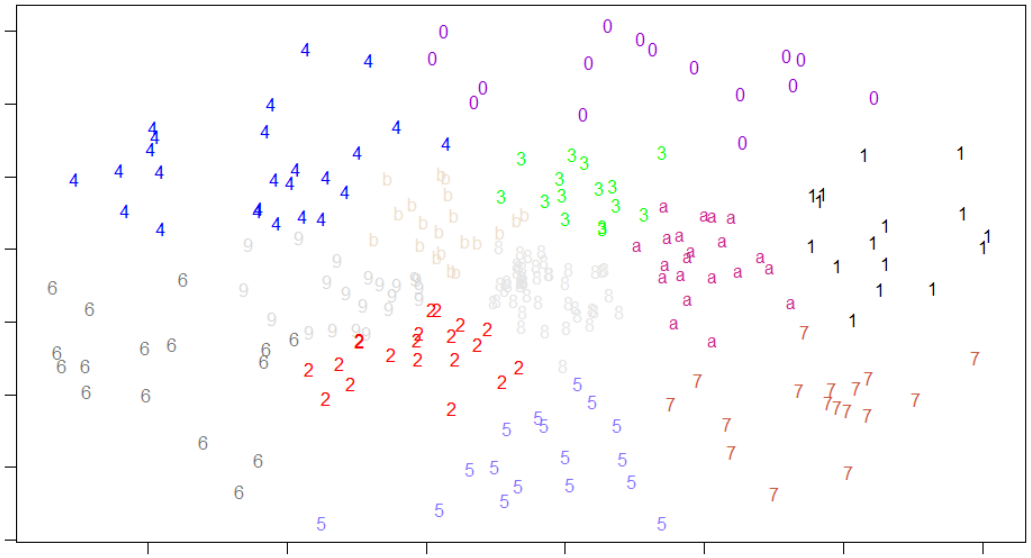
Ο αλγόριθμος K-means με την Manhattan απόσταση με αριθμό συστάδων K=6



Ο αλγόριθμος K-means με την ευκλείδια απόσταση με αριθμό συστάδων K=12



Ο αλγόριθμος K-means με την Manhattan απόσταση με αριθμό συστάδων K=12



Μία επιλογή αρχικών σημείων που οδηγεί σε καλά αποτελέσματα είναι ο αλγόριθμος K-means με αριθμό συστάδων K=3. Ενώ μία επιλογή αρχικών σημείων που δεν οδηγεί σε καλά αποτελέσματα είναι ο αλγόριθμος K-means με αριθμό συστάδων K=6.

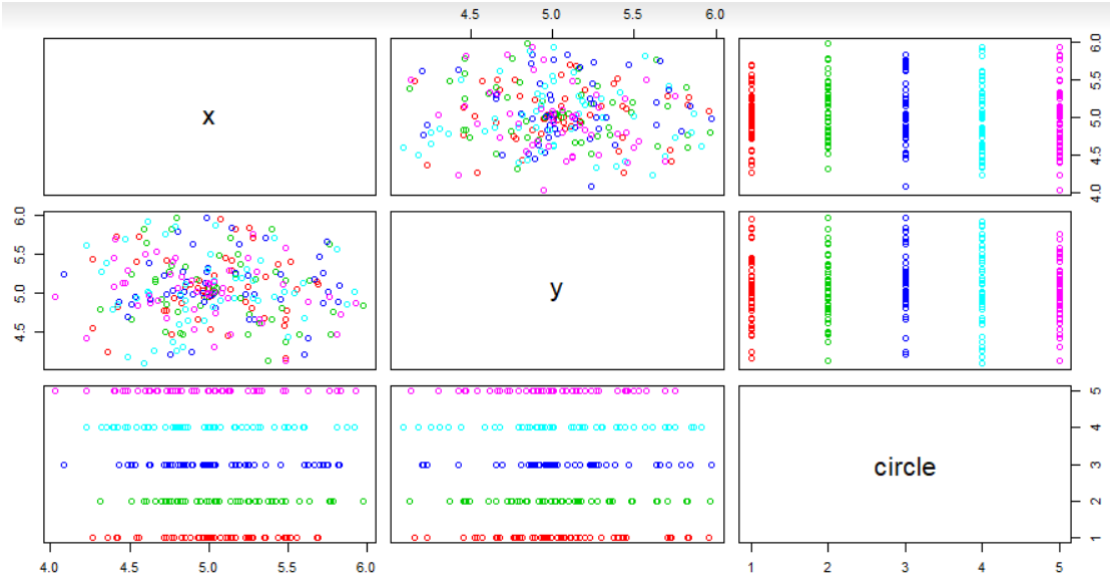
Ερώτημα iii

Η συσταδοποίηση με βάση την πυκνότητα αναφέρεται σε μη επιβλεπόμενες μεθόδους μάθησης που προσδιορίζουν διακριτές συστάδες στα δεδομένα. Βασίζεται στην ιδέα ότι μία συστάδα στον χώρο των δεδομένων είναι μία συνεχής περιοχής με υψηλή πυκνότητα μεταξύ των σημείων, διαχωρισμένη από άλλες τέτοιες συστάδες με συνεχείς περιοχές χαμηλής πυκνότητας μεταξύ των σημείων. Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) είναι ένας βασικός αλγόριθμος για συσταδοποίηση με βάση την πυκνότητα. Μπορεί να ανακαλύψει συστάδες διαφορετικών σχημάτων και μεγεθών από μία μεγάλη ποσότητα δεδομένων με θόρυβο και και ακραίες τιμές.

Ο αλγόριθμος DBSCAN δέχεται δύο ορίσματα:

minPts: Ο ελάχιστος αριθμός σημείων (ένα κατώφλι) που θα συσταδοποιηθούν μαζί ώστε να θεωρηθεί μία περιοχή πυκνή.

eps (ε): Ένα μέτρο απόστασης που θα χρησιμοποιηθεί για τον εντοπισμό των σημείων στην περιοχή οποιουδήποτε σημείου.



2.1.2. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 2

Ερώτημα 1ο:

Η λογιστική παλινδρόμηση προβλέπει την πιθανότητα ενός αποτελέσματος που μπορεί να πάρει μόνο δυο τιμές. Η πρόβλεψη γίνεται με τη χρήση ενός ή περισσοτέρων εκτιμητών.

Η γραμμική παλινδρόμηση δεν είναι κατάλληλη για την πρόβλεψη τιμών δυαδικού τύπου, καθώς προβλέπει τιμές πέραν του αποδεκτού διαστήματος [0,1] και τα υπόλοιπα που θα προκύψουν από το μοντέλο δεν θα ακολουθούν την κανονική κατανομή εξαιτίας της φύσης του προβλήματος. Αντίθετα, η λογιστική παλινδρόμηση παράγει μια λογιστική καμπύλη, η οποία προβλέπει τιμές στο διάστημα [0,1].

$$p = \frac{1}{1 + \mathrm{e}^{-(w0+w1x)}}$$

Ο σταθερός όρος w_0 του μοντέλου εκφράζει τη μετατόπιση του μοντέλου, ενώ ο όρος w_1 εκφράζει πόσο απότομη είναι η καμπύλη και το μέγεθος συνεισφοράς της αντίστοιχης μεταβλητής. Θετική τιμή του συντελεστή δηλώνει ότι η επεξηγηματική μεταβλητή αυξάνει την πιθανότητα της επιτυχημένης έκβασης (να συμβεί δηλαδή το γεγονός), αρνητική τιμή σημαίνει ότι η μεταβλητή μειώνει την πιθανότητα αυτής της έκβασης. Υψηλή τιμή του συντελεστή σημαίνει ότι η ανεξάρτητη μεταβλητή επηρεάζει πολύ ισχυρά την πιθανότητα να συμβεί το γεγονός ή μη, ενώ χαμηλή τιμή

δηλώνει μικρή επίδραση της ανεξάρτητης μεταβλητής στην πιθανότητα εμφάνισης της ανάλογης έκβασης.

Υπάρχουν αρκετές αναλογίες μεταξύ της γραμμικής και της λογιστικής παλινδρόμησης. Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην μεν πρώτη μπορεί να είναι κατηγορική, τακτική ή ονομαστική, στη δε δεύτερη αποκλειστικά ποσοτική. Οι συντελεστές της γραμμικής παλινδρόμησης που οδηγούν στην βέλτιστη ευθεία προκύπτουν με την Μέθοδο των Ελαχίστων Τετραγώνων, ενώ στη λογιστική παλινδρόμηση οι συντελεστές που θα οδηγήσουν στην πρόβλεψη της κλάσης από τη μέθοδο του λόγου πιθανοφάνειας.

Οι συντελεστές πολλαπλού προσδιορισμού τύπου R^2, ή αλλιώς ψευδο-συντελεστές R^2, αποτελούν προσέγγιση εκτίμησης της καλής προσαρμογής του μοντέλου παρέχοντας ερμηνεία παρόμοια με εκείνη της γραμμικής παλινδρόμησης, δηλαδή εκφράζουν το ποσοστό της διακύμανσης που επεξηγείται από τις ανεξάρτητες μεταβλητές. Δημοφιλέστερος θεωρείται ο συντελεστής R^2 του McFadden.

Ο συντελεστής τύπου R_2 του McFadden γνωστός και ως δείκτης του λόγου πιθανοφανειών (Likelihood-ratio index) συγκρίνει το μοντέλο με k ανεξάρτητες μεταβλητές με το μοντέλο εκείνο στο οποίο απουσιάζουν οι συγκεκριμένες μεταβλητές.

$$R^2 = 1 - \frac{LL_{full\_model}}{LL_{intercept}}$$

όπου $LL_{intercept}$ η εκτίμηση πιθανοφάνειας στο μοντέλο χωρίς την ένταξη των ανεξάρτητων μεταβλητών και LL_{full_model} η εκτίμηση στο πλήρες μοντέλο.

Υψηλές τιμές του δείκτη R^2 δηλώνουν ένδειξη καλής προσαρμογής του μοντέλου.

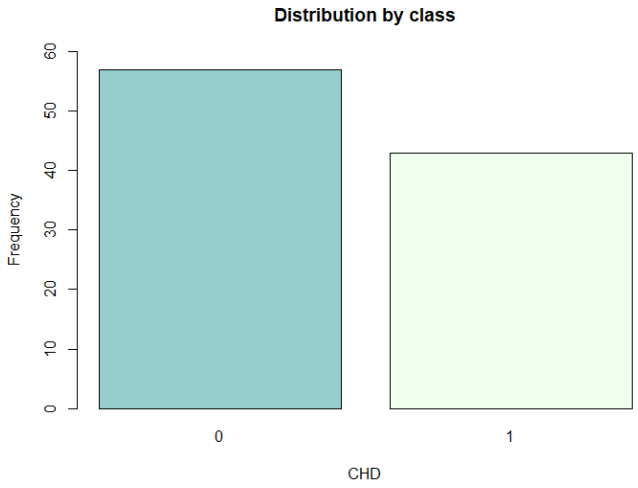
Ένας ακόμα συντελεστής R^2 είναι ο Count που αντιστοιχεί στο μέτρο καλής προσαρμογής στης ακρίβειας για προβλήματα classification.

$$R^2 = 1 - \frac{Correct\ Predicted}{Total\ Count}$$

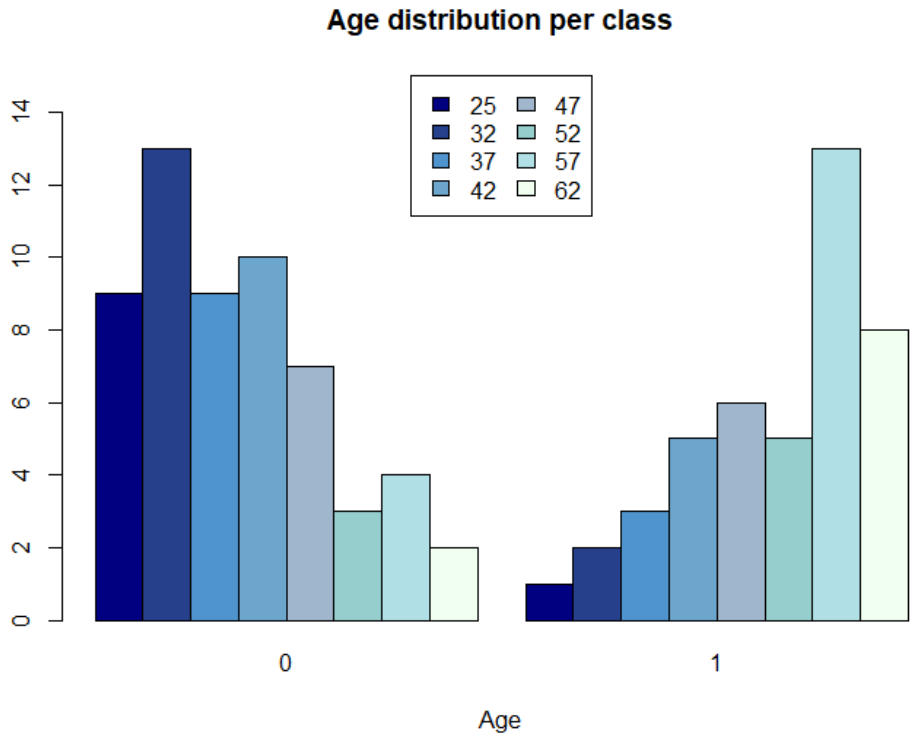
Ερώτημα 2ο:

Το σύνολο δεδομένων Coronary Heart Disease αποτελείται από μια μεταβλητή, που αντιστοιχεί στην ηλικία των ασθενών (Age) και την κλάση (0 - Υγιής ή 1 - Ασθενής).

Παρατηρώντας το παρακάτω γράφημα προκύπτει ότι στην κλάση 0 αντιστοιχούν 57 άτομα και στη κλάση 1, 43. Με αποτέλεσμα να υπάρχει μια μικρή ανισορροπία την οποία πρόκειται να αγνοήσουμε.



Μελετώντας την κατανομή των δεδομένων ανά κλάση προκύπτει ότι είναι συχνότερη η εμφάνιση CHD στις μεγαλύτερες ηλικιακές ομάδες από ότι στις νεότερες.

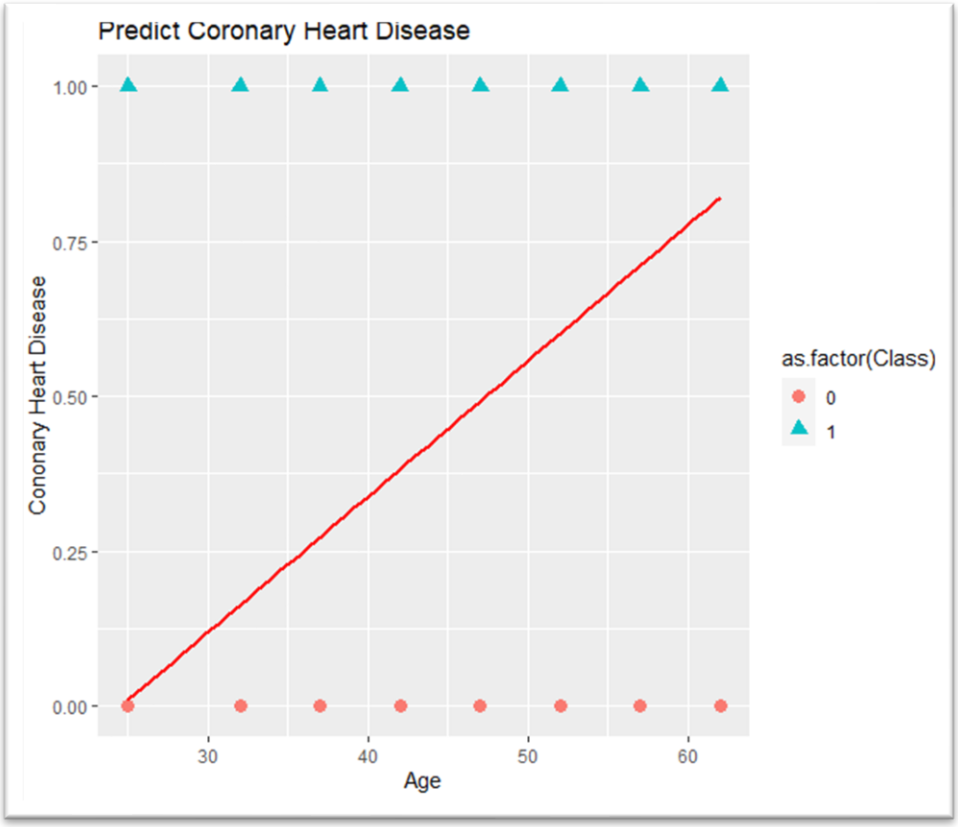


1. Εφαρμόζοντας τη συνάρτηση lm για γραμμική παλινδρόμηση προκύπτουν τα εξής αποτελέσματα:
   - Οι συντελεστές της γραμμικής παλινδρόμησης είναι στατιστικά σημαντικοί και ίσοι με $w_0 = - 0.54 \pm 0.17$ και $w_1 = 0.02 \pm 0.003$

- Ο συντελεστής R^2 ισούται με 0.26. Η τιμή αυτή είναι ιδιαίτερα χαμηλή που σημαίνει ότι το γραμμικό μοντέλο δεν προσαρμόζεται καλά στα δεδομένα, γεγονός αναμενόμενο εξαιτίας του δυαδικού χαρακτήρα του προβλήματος.

```
Call:
lm(formula = Class ~ Age, data = CHD)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82020 -0.38177 -0.00911  0.28940  0.99089

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.538933   0.171510  -3.142  0.00222 **
Age          0.021922   0.003756   5.837 6.91e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4308 on 98 degrees of freedom
Multiple R-squared:  0.2579,    Adjusted R-squared:  0.2504
F-statistic: 34.07 on 1 and 98 DF,  p-value: 6.905e-08
```

Η γραφική αναπαράσταση των δεδομένων και το γραμμικό μοντέλο που προσαρμόζεται βέλτιστα στα δεδομένα παρουσιάζονται στο παρακάτω διάγραμμα.

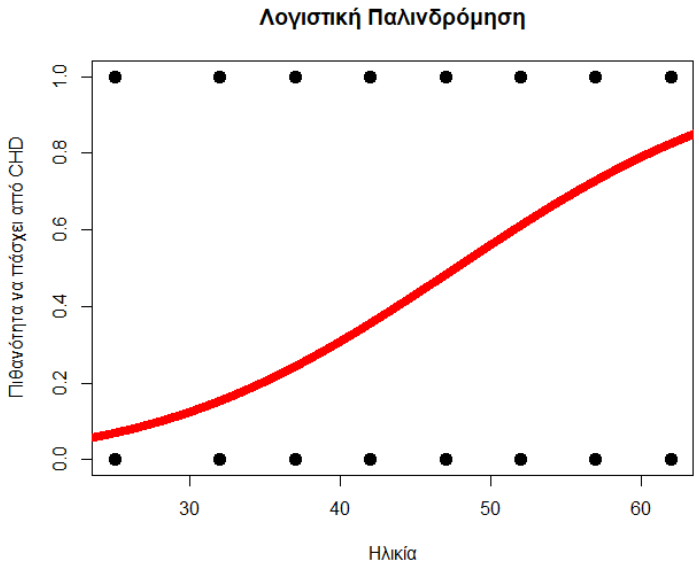


Για κάποιον άνθρωπο ηλικίας 41 ετών, το γραμμικό μοντέλο προβλέπει πιθανότητα 36% να πάσχει από CHD.

2. Εφαρμόζοντας τη συνάρτηση glm για λογιστική παλινδρόμηση προκύπτουν τα παρακάτω αποτελέσματα:
   - Οι συντελεστές της γραμμικής παλινδρόμησης είναι στατιστικά σημαντικοί και ίσοι με $w_0 = -3.11 \pm 0.62$ και $w_1 = 0.07 \pm 0.01$
   - Από τη συνάρτηση pR2, προκύπτει συντελεστής τύπου R2 τύπου McFadden ίσος με 0.21 και συντελεστής R2 τύπου Count ίσος 0.74. Από τα παραπάνω προκύπτει ότι το μοντέλο δεν προσαρμόζεται ικανοποιητικά στα δεδομένα, ενώ καταφέρνει να ταξινομήσει σωστά το 74% των παραδειγμάτων εκπαίδευσης.

```
Call:
glm(formula = Class ~ Age, family = binomial(link = "probit"),
    data = CHD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8676  -0.9376  -0.3795   0.7956   2.3094

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.11132    0.62139  -5.007 5.53e-07 ***
Age          0.06527    0.01332   4.901 9.53e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 108.36  on 98  degrees of freedom
AIC: 112.36

Number of Fisher Scoring iterations: 4
```

Η γραφική αναπαράσταση των δεδομένων και το γραμμικό μοντέλο που προσαρμόζεται βέλτιστα στα δεδομένα παρουσιάζονται στο παρακάτω διάγραμμα.



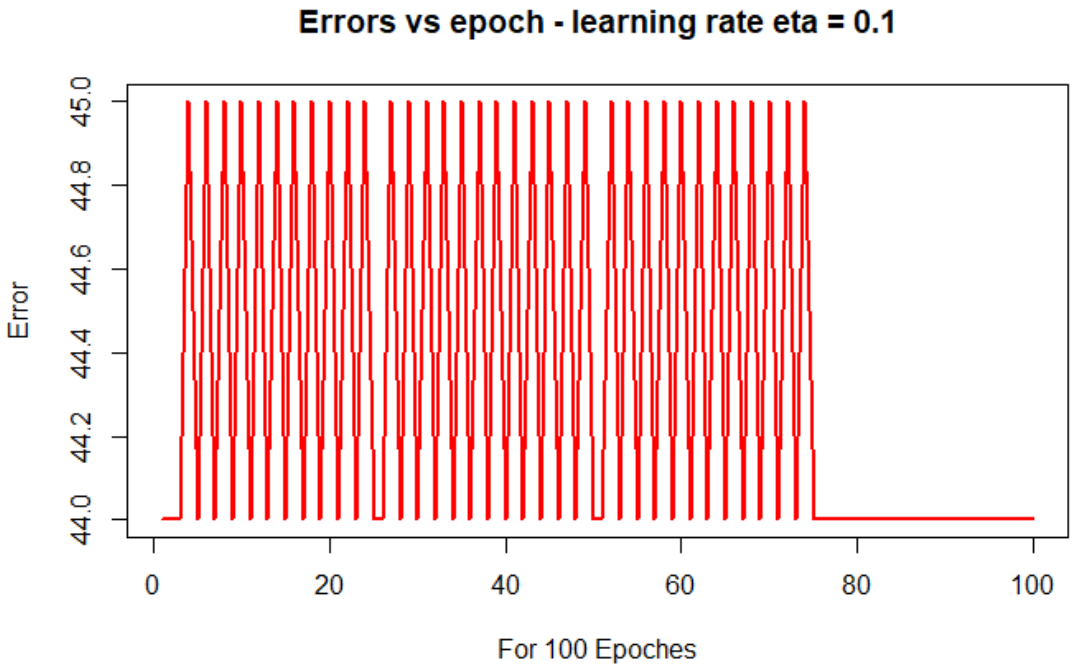
Για κάποιον άνθρωπο ηλικίας 41 ετών, το μοντέλο λογιστικής παλινδρόμησης προβλέπει πιθανότητα 33% να πάσχει από CHD.

Ερώτημα 3ο:

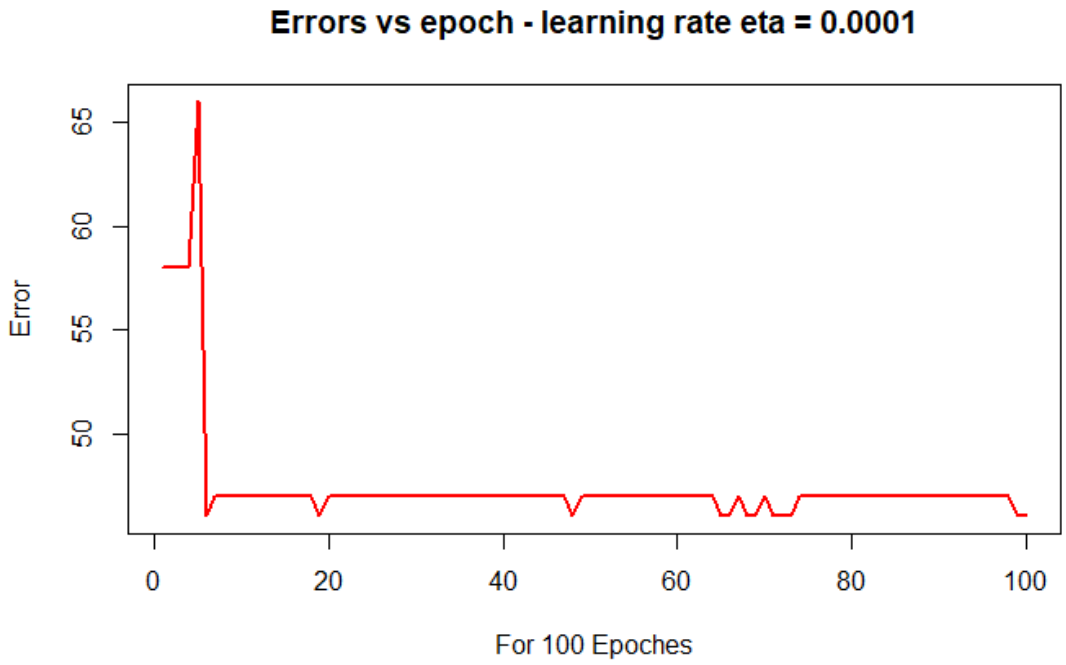
Ο αλγόριθμος Perceptron χρησιμοποιείται για την ταξινόμηση δυο κλάσεων ω_1 και ω_2 γραμμικά διαχωρίσιμων και έχει ως σκοπό να προσδιορίσει τις παραμέτρους του υπερεπιπέδου που διαχωρίζει πλήρως τις δυο κλάσεις. Οι παράμετροι του υπερεπιπέδου προσδιορίζονται μέσω της ελαχιστοποίησης της συνάρτησης κόστους του Perceptron, μέσω μια επαναληπτικής διαδικασίας. Η συνάρτηση κόστους μηδενίζεται όταν όλες οι παρατηρήσεις ταξινομούνται σωστά.

Ο αλγόριθμος Perceptron που υλοποιήθηκε παίρνει ως όρισμα ένα data.frame με τις επεξηγηματικές μεταβλητές, μια λίστα με τις τιμές των κλάσεων, το ρυθμό μάθησης και τον αριθμό των επαναλήψεων. Η επαναληπτική διαδικασία ξεκινάει για μοναδιαία αρχικά βάρη και γίνεται πρόβλεψης της κλάση για κάθε σημείο και ανανέωση των βαρών με ένα ρυθμό εκμάθησης. Η διαδικασία ολοκληρώνεται όταν συμπληρωθεί ο αριθμός των επαναλήψεων. Η ακρίβεια του μοντέλου υπολογίζεται μέσω της μετρικής accuracy και το σφάλμα ως το πλήθος των λάθος ταξινομημένων παρατηρήσεων.

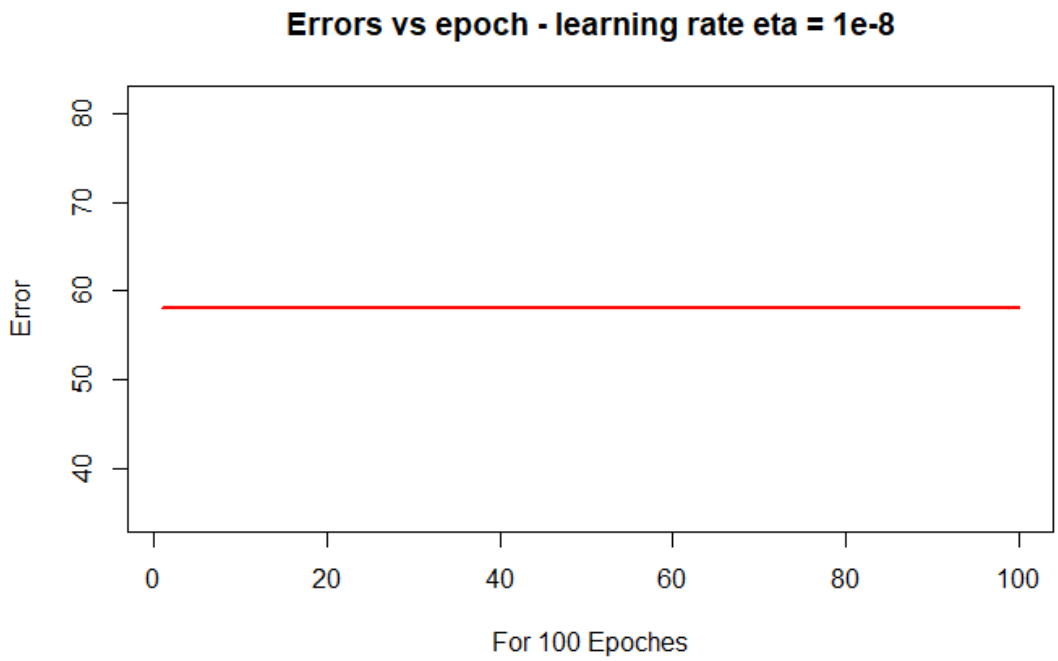
Για το σύνολο δεδομένων CHD για αριθμό εποχών ίσο με 100 και ρυθμό εκμάθησης μεγάλο και ίσο με 0.1, ο αλγόριθμος συγκλίνει έπειτα από τις 77 επαναλήψεις ταξινομώντας εσφαλμένα το 45% των παρατηρήσεων.



Για ίδιο αριθμό επαναλήψεων και ρυθμό εκμάθησης ίσο με 0.0001, προκύπτει ότι η σύγκλιση είναι πιο αργή και το βέλτιστο σφάλμα που προκύπτει είναι 46 εσφαλμένες προβλέψεις σε σύνολο των 100.



Για πολύ μικρό ρυθμό εκμάθησης το σφάλμα, όπως, φαίνεται και στο διάγραμμα που ακολουθεί, δεν συγκλίνει.

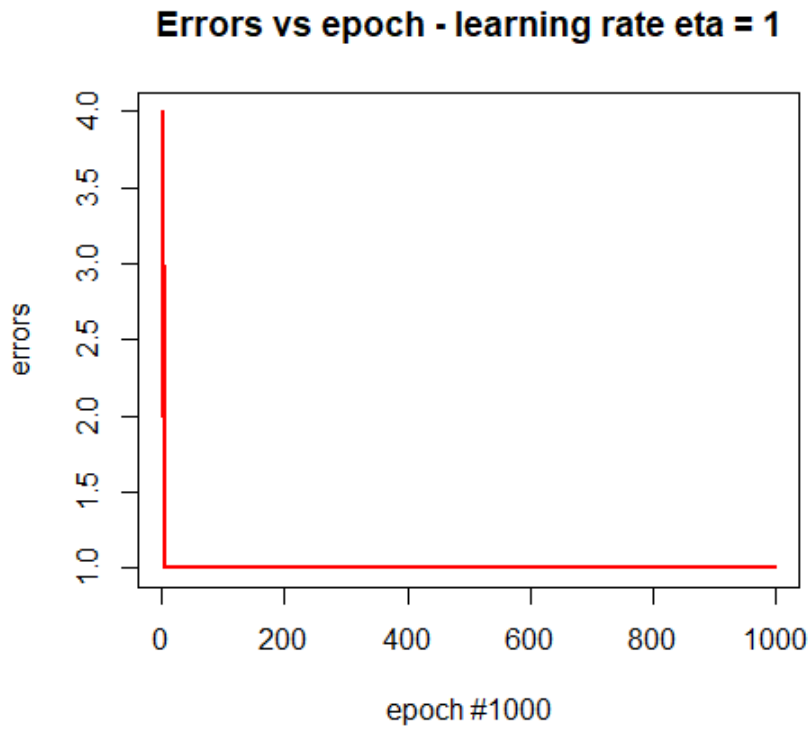


Τα αποτελέσματα και των τριών αλγορίθμων για το σύνολο δεδομένων CHD παρουσιάζονται στον πίνακα που ακολουθεί. Η λογιστική παλινδρόμηση καταφέρνει να κατηγοριοποίησει τις δυο κλάσεις με καλύτερη ακρίβεια.

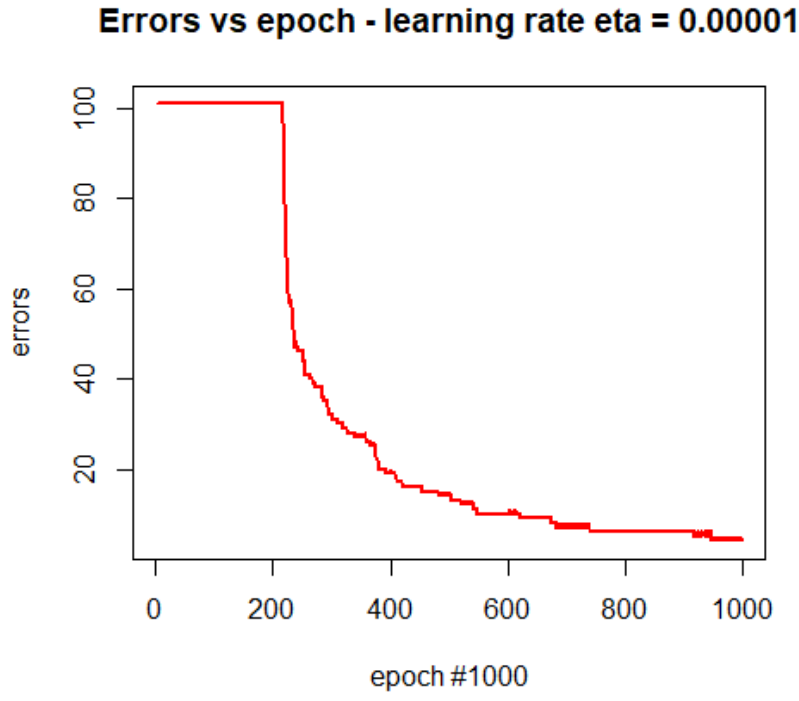
Αλγόριθμος	***R2 – pseudo R2 Count***
Γραμμική Παλινδρόμηση	*0.26*
Λογιστική Παλινδρόμηση	*0.74*
Perceptron	*0.62*

Εφαρμόζοντας τον ίδιο αλγόριθμο για το σύνολο δεδομένων Iris, με σκοπό την πρόβλεψη της κατηγορίας λουλουδιού ***setosa***, προκύπτουν τα εξής:

- Για 1000 επαναλήψεις και μεγάλο ρυθμό μάθηση, η σύγκλιση του αλγορίθμου πραγματοποιείται στην 6 επανάληψη. Ο αλγόριθμος με αυτές τις παραμέτρους εκπαίδευσης επιτυγχάνει 100 % ακρίβεια στο σύνολο εκπαίδευσης.



- Για 1000 επαναλήψεις και μικρό ρυθμό μάθηση, η σύγκλιση του αλγορίθμου είναι αργή με τελικό πλήθος σφαλμάτων 4/150. Αύξηση των επαναλήψεων αναμένεται να βελτιώσει τα αποτελέσματα.



Η απόδοση του αλγορίθμου πρόκειται να αξιολογηθεί και σε ένα καινούριο σύνολο δεδομένων που δεν έχει ξαναδεί ο αλγόριθμος, το iris1.txt. Η απόδοση του μοντέλου είναι ικανοποιητική και σε αυτό το σύνολο δεδομένων καθώς το σύνολο των παρατηρήσεων ταξινομούνται σωστά.

2.1.3. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 3

Το σύνολο δεδομένων στη συγκεκριμένη άσκηση είναι το Pima.te από το MASS package. Τα δεδομένα περιέχουν στοιχεία από έναν πληθυσμό γυναικών τουλάχιστον 21 ετών από την Ινδία που ελέγχθηκαν για διαβήτη σύμφωνα με τα κριτήρια του Παγκόσμιου Οργανισμού Υγείας. Το σύνολο δεδομένων περιέχει 8 στήλες όπου στην τελευταία αναγράφεται αν ο ασθενής έχει διαβήτη (‘Yes’) ή όχι (‘No’) και 332 εγγραφές.

Αρχικά χωρίσαμε το σύνολο δεδομένων σε σύνολο εκπαίδευση με το 80% των εγγραφών και σε σύνολο ελέγχου με το 20% των εγγραφών.

Στη συνέχεια εφαρμόσαμε ένα μοντέλο λογιστικής παλινδρόμησης. Η λογιστική παλινδρόμηση είναι κατάλληλη όταν η εξαρτημένη μεταβλητή είναι binary.

Εφαρμόσαμε το μοντέλο αρκετές φορές ώστε να βρούμε το καταλληλότερο μοντέλο και ποιες μεταβλητές είναι πιο σημαντικές για τη διάγνωση του διαβήτη. Αρχικά, εφαρμόσαμε το μοντέλο με όλες τις ανεξάρτητες μεταβλητές, στη συνέχεια με όλες τις μεταβλητές εκτός της μεταβλητής bp, έπειτα χωρίς τις bp και skin και τέλος χωρίς τις bp, skin και age. Το καλύτερο μοντέλο αποδείχτηκε το μοντέλο χωρίς τις μεταβλητές bp και skin.

```
Logistic Regression Model

lrm(formula = type ~ npreg + glu + bmi + ped + age, data = train)

                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes           Indexes
Obs            265    LR chi2     114.13    R2       0.492    C        0.874
 No            182    d.f.             5    g        2.069    Dxy      0.748
 Yes            83    Pr(> chi2) <0.0001    gr       7.920    gamma    0.748
max |deriv| 3e-07                          gp       0.319    tau-a    0.323
                                           Brier    0.128

          Coef     S.E.   Wald Z Pr(>|Z|)
Intercept -10.0611 1.3315 -7.56  <0.0001
npreg       0.1847 0.0670  2.76  0.0058
glu         0.0415 0.0064  6.46  <0.0001
bmi         0.0801 0.0262  3.06  0.0022
ped         0.7373 0.5126  1.44  0.1503
age         0.0079 0.0188  0.42  0.6726
```

Συνεπώς, γίνεται χρήση του συγκεκριμένου μοντέλου για την πρόβλεψη της πιθανότητας να έχει κάποιος διαβήτη.

Παρακάτω φαίνεται ο πίνακας ταξινόμησης του μοντέλου:

```
           Predictedvalue
ActualValue FALSE TRUE
        No     34    7
        Yes    10   16
```

Η ακρίβεια του συγκεκριμένου μοντέλου είναι 0.746.

Στη συνέχεια έγινε πρόβλεψη για τα επόμενα 4 άτομα.

	Npreg	Glu	BP	BMI	SKIN	PED	age
1	5	140	76	70	20	0.6	40
2	1	80	70	25	45	0.56	25
3	8	120	60	27	30	0.5	44
4	2	91	50	68	23	0.7	34

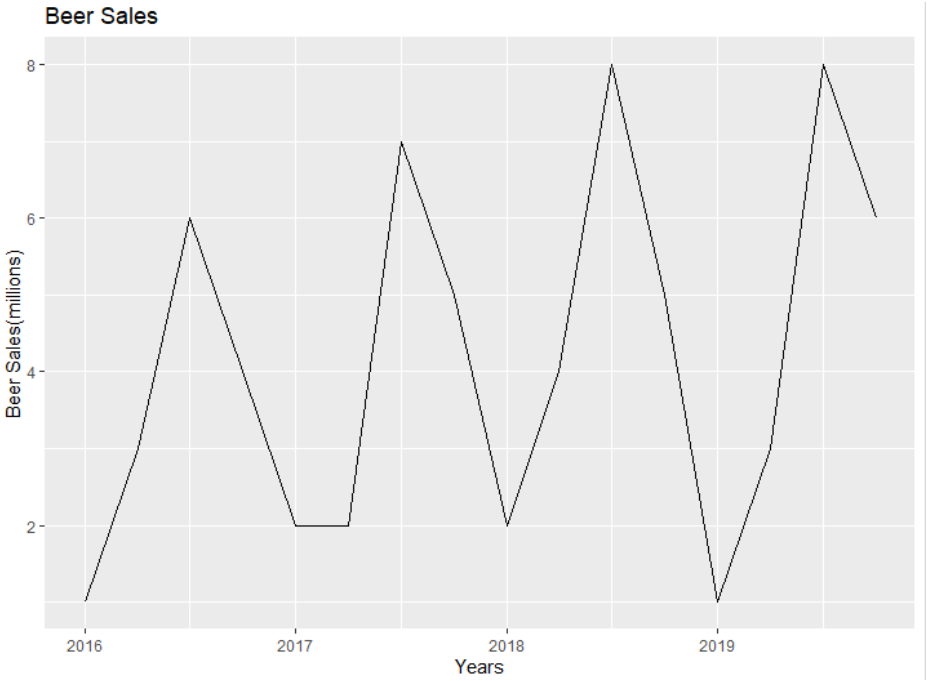
Προέκυψαν τα εξής αποτελέσματα:

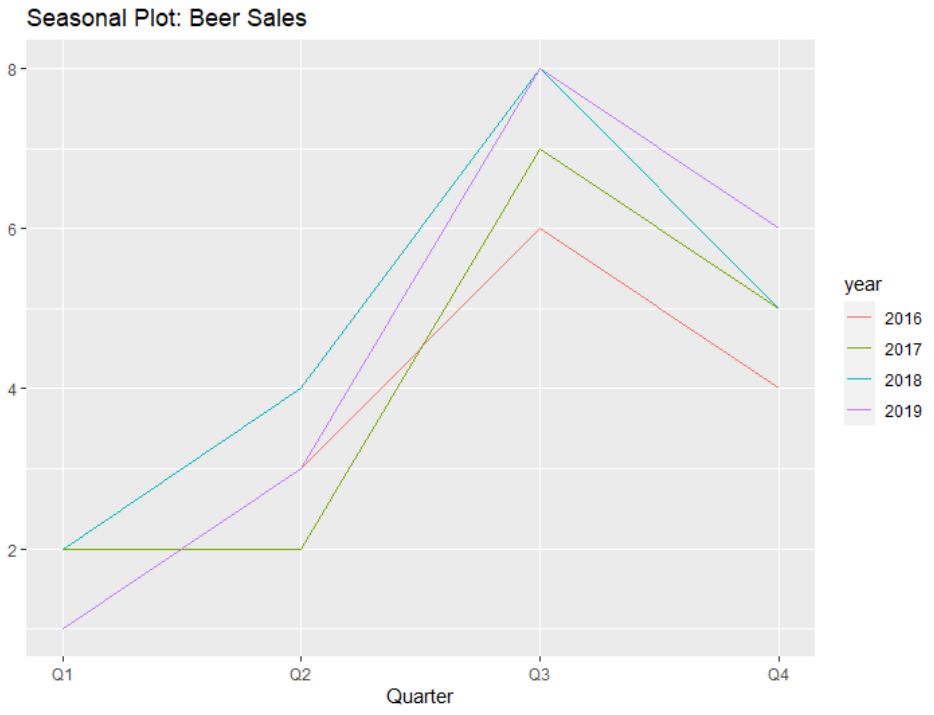
```
        1         2         3         4
0.9542685 0.0190288 0.3268434 0.5783647
```

Συνέπως, σύμφωνα με το συγκεκριμένο μοντέλο λογιστικής παλινδρόμησης οι ασθενείς 1 και 4 έχουν διαβήτη ενώ οι 2 και 3 όχι.

2.1.4. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 4

Στο παρακάτω διάγραμμα παρουσιάζονται οι πωλήσεις μπύρας μια βιομηχανίας από το 2016 έως το 2019 σε εκατομμύρια μπουκάλια. Στα δεδομένα παρατηρείται εποχιακή μεταβλητότητα (seasonality) με φανερή αύξηση της κατανάλωσης (μέγιστο) το τρίτο τρίμηνο κάθε έτους. Επίσης, ορατή είναι μια ελαφριά ανοδική μακροχρόνια τάση.

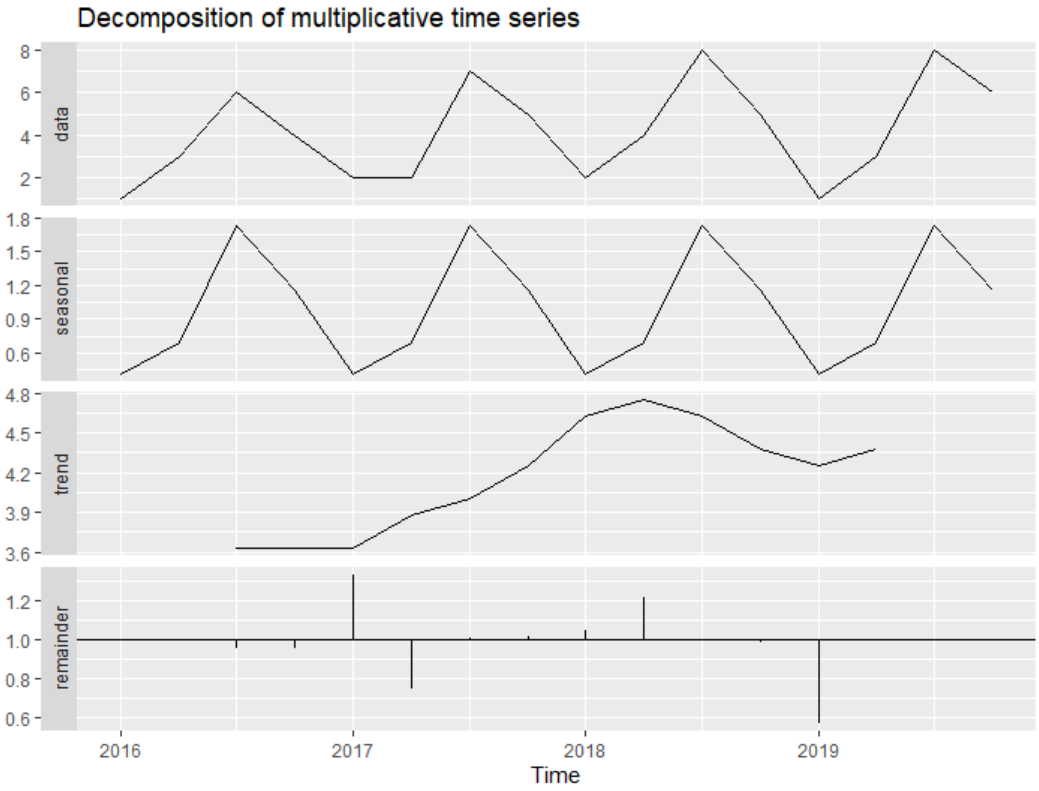




1. Για την εξάλειψη της εποχικότητας από τα δεδομένα μας, πρόκειται να υπολογιστούν οι εποχιακοί δείκτες των δεδομένων. Το πρώτο βήμα γίνεται με τον υπολογισμό των κινητών μέσων, με αποτέλεσμα την εξομάλυνση της χρονοσειράς. Η γραμμή των κινητών μέσων αποτυπώνει την κύρια κίνηση της χρονοσειράς χωρίς τις αμελητέες διακυμάνσεις.

Μεσώ των κινητών μέσων είναι δυνατός ο υπολογισμός των εποχιακών δεικτών, που ποσοτικοποιούν την επίδραση της εποχικότητας στη χρονοσειρά. Διαιρώντας ή αφαιρώντας τις πραγματικές τιμές της χρονοσειράς με το δείκτη εποχικότητας της αντίστοιχης περιόδου προκύπτει η χρονοσειρά απαλλαγμένη από την εποχιακή μεταβλητότητα.

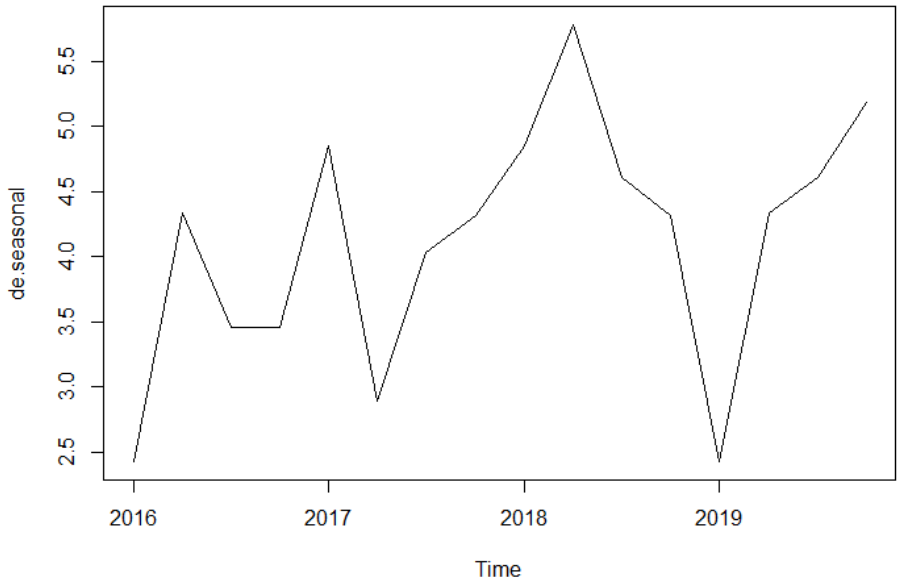
Στην R η διαδικασία του decomposition, πραγματοποιείται με τη συνάρτηση decompose κάνοντας χρήση του πολλαπλασιαστικού μοντέλου.



Οι δείκτες εποχικότητας που χρησιμοποιήθηκαν είναι οι εξής:

Qtr1	***Qtr2***	***Qtr3***	***Qtr4***
0.413	0.692	1.738	1.158

Κάποιες φορές όταν έχουμε καθορίσει τους εποχιακούς δείκτες, μπορούμε να εξαλείψουμε την επίδραση της εποχικότητας από τις χρονοσειρές. Αυτή η διόρθωση επιτρέπει τον υπολογισμό των κυκλικών μεταβολών που συμβαίνουν κάθε χρόνο. Όταν εξαλείψουμε την επιρροή των εποχιακών μεταβολών τότε έχουμε μη εποχιακές χρονολογικές σειρές.



2. Στη συνέχεια πρόκειται να υπολογίσουμε την τάση της χρονοσειράς, όντας απαλλαγμένη από την εποχιακή μεταβλητότητα, με χρήση της γραμμικής παλινδρόμησης.

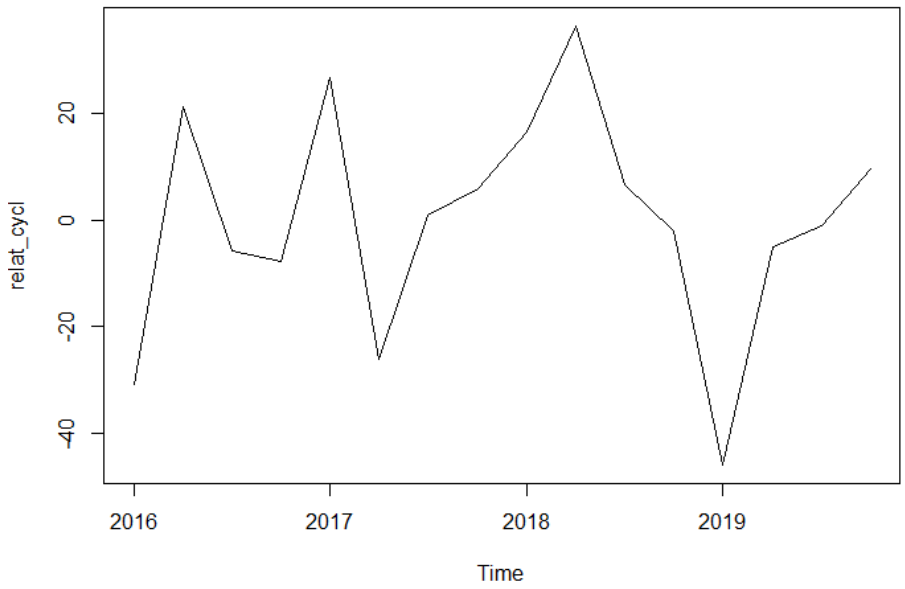
   Η εξίσωση της τάσης που προκύπτει είναι:

$$\bar{Y}_i = -659.720 + 0.329\, t_i$$

3. Η κυκλικότητα προκύπτει όταν τα δεδομένα παρουσιάζουν διακυμάνσεις, χωρίς σταθερή συχνότητα. Για την μελέτη της κυκλικής μεταβολής υπολογίζεται η σχετική τάση.

$$\frac{\bar{Y}_i - Y_i}{Y_i} * 100$$

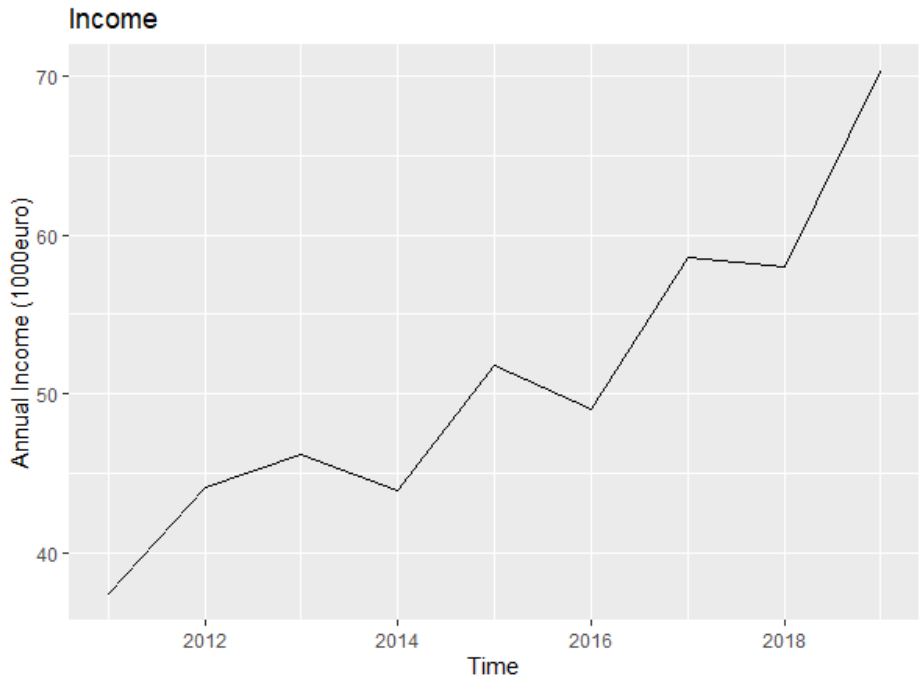
   Στο παρακάτω διάγραμμα η ύπαρξη κυκλικότητας γίνεται φανερή και οφείλεται πιθανώς στον κύκλο ζωής της επιχείρησης.



2.1.5. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 5

Σε μία χρονολογική σειρά με ετήσια δεδομένα μόνο η μακροχρόνια τάση, η κυκλική μεταβολή και η ακανόνιστη μεταβολή λαμβάνονται υπόψη. Η εποχιακή μεταβολή κάνει ένα ολόκληρο κύκλο στη διάρκεια ενός έτους και έτσι δεν επηρεάζει ένα χρόνο περισσότερο από ένα άλλο.

Στη συγκεκριμένη άσκηση δίνεται πίνακας με τις ετήσιες εισπράξεις μίας βιομηχανίας (σε χιλ. ευρώ). Για την κατασκευή της γραμμής τάσης της χρονοσειράς θα γίνει χρήση της γραμμικής παλινδρόμησης.



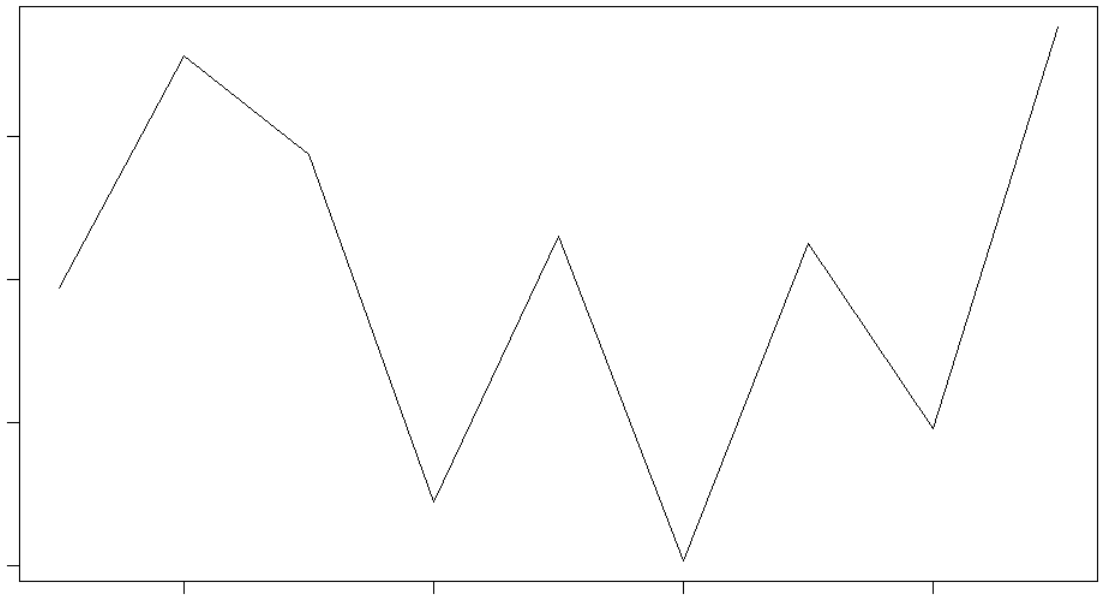
Η εξίσωση της τάσης που προκύπτει είναι:

$$\bar{Y}_i = 34.174722 + 3.378833\, t_i$$

Η κυκλικότητα προκύπτει όταν τα δεδομένα παρουσιάζουν διακυμάνσεις, χωρίς σταθερή συχνότητα. Για την μελέτη της κυκλικής μεταβολής υπολογίζεται η σχετική τάση.

$$\frac{\bar{Y}_i - Y_i}{Y_i} * 100$$

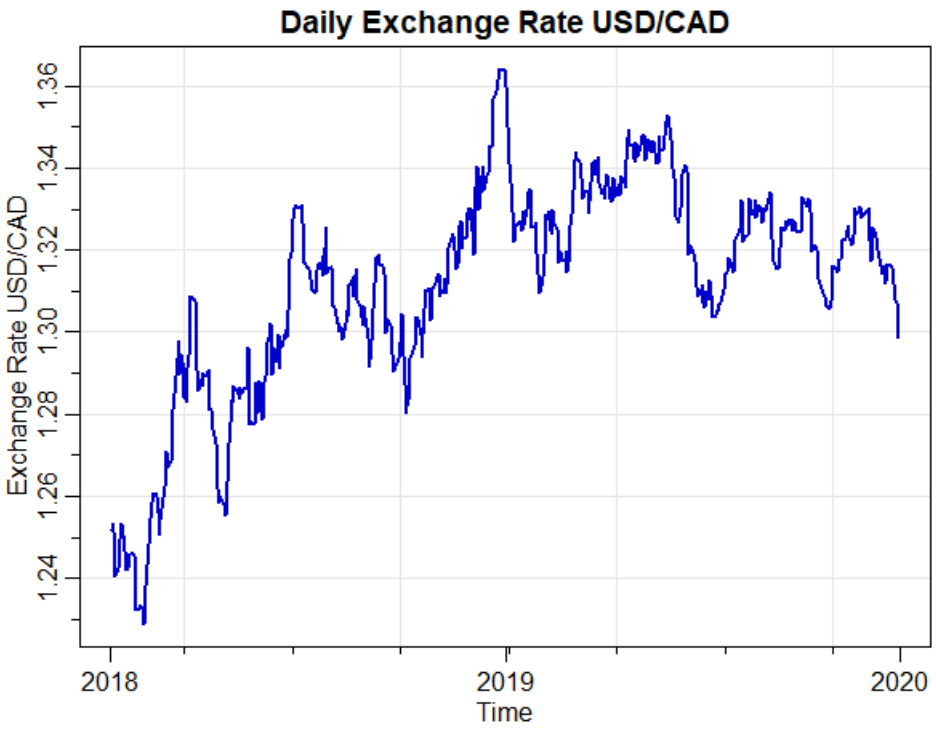
Στο παρακάτω διάγραμμα η ύπαρξη κυκλικότητας γίνεται φανερή και οφείλεται πιθανώς στον κύκλο ζωής της επιχείρησης.



2.1.6. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 6

I. Τα δεδομένα που χρησιμοποιήθηκαν για τη συγκεκριμένη άσκηση είναι η ισοτιμία συναλλάγματος USD/CAD τα τελευταία 2 χρόνια. Ειδικότερα τα δεδομένα περιλαμβάνουν την ισοτιμία συναλλάγματος για από τις 2018-01-02 έως τις 2019-12-31, σύνολο 729 ημερών.

   Τα δεδομένα αυτά δεν περιλαμβάνουν δεδομένα για τα Σαββατοκύριακα και τις αργίες, για αυτό το λόγο τα κενά συμπληρώνονται με την τιμή συναλλάγματος της προηγούμενης μη κενής ημέρας.

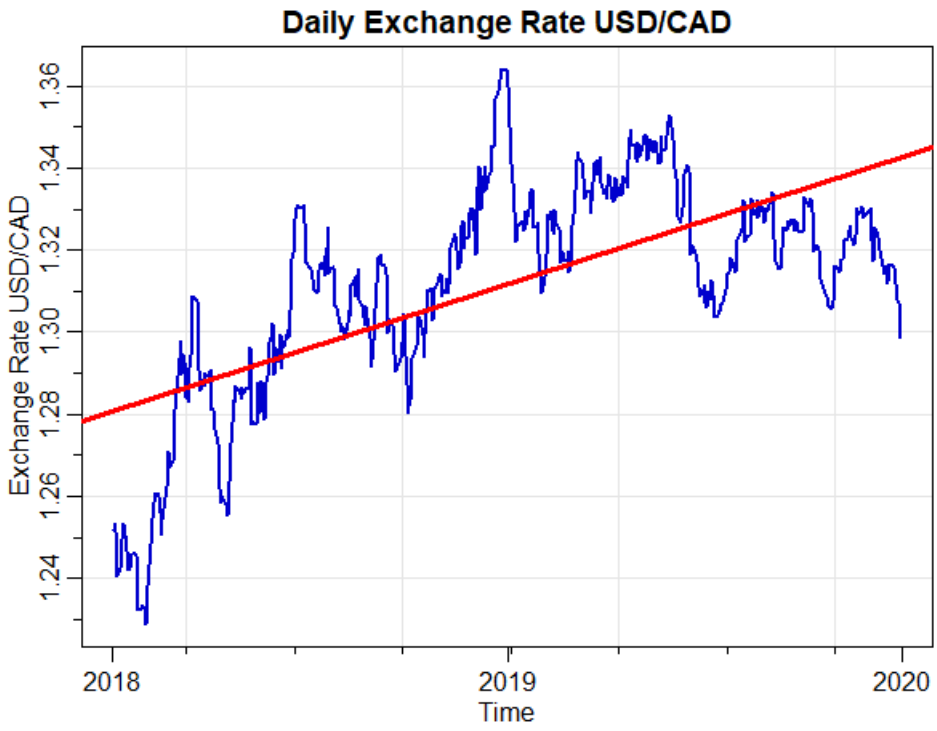


II. Για τον υπολογισμό της γραμμής της τάσης χρησιμοποιήθηκε η μέθοδος lm της R. Η εξίσωση που προέκυψε είναι η εξής:

$$\bar{Y}_i = -0.2004 + 0.000084\, t_i$$

```
Call:
lm(formula = myts ~ time(myts))

Residuals:
      Min        1Q    Median        3Q       Max
-0.054637 -0.013525  0.000769  0.014440  0.052867

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.004e-01  6.168e-02  -3.249  0.00121 **
time(myts)   8.449e-05  3.446e-06  24.515  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01958 on 727 degrees of freedom
Multiple R-squared:  0.4526,    Adjusted R-squared:  0.4518
F-statistic:   601 on 1 and 727 DF,  p-value: < 2.2e-16
```

Η σχέση μεταξύ των μεταβλητών προκύπτει στατιστικά σημαντική. Ο συντελεστής προσδιορισμού του μοντέλου R^2 προκύπτει ίσως με 0.45, γεγονός που οδηγεί στο συμπέρασμα ότι η γραμμή της τάσης δεν προσαρμόζεται καλά στα δεδομένα.

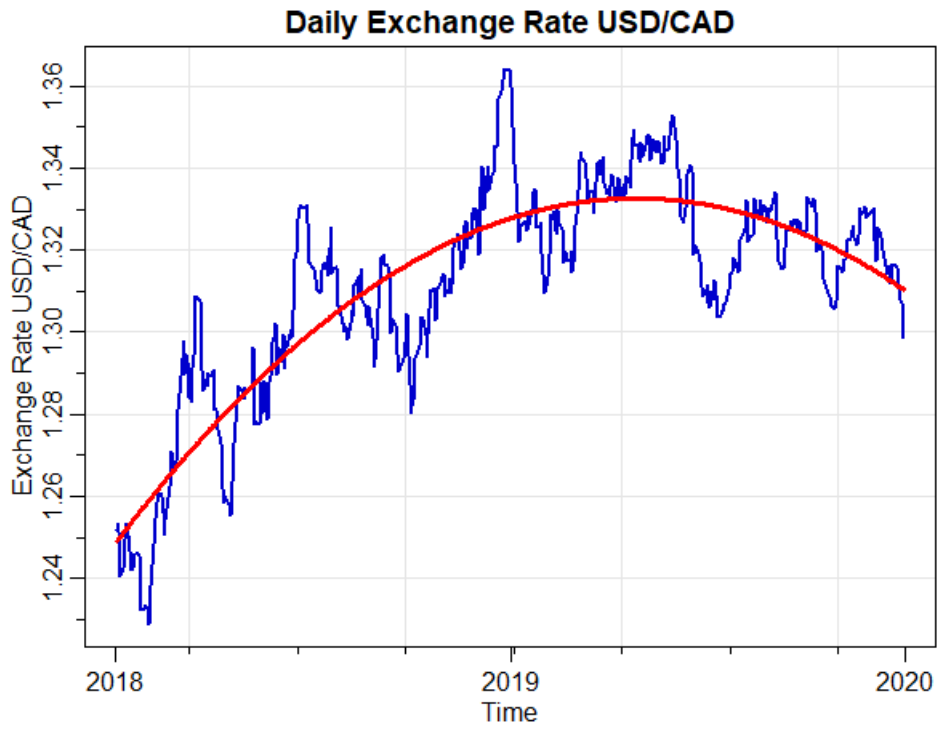


III. Εκφράζοντας τη γραμμή της τάσης ως πολυώνυμο δευτέρου βαθμού αναμένουμε καλύτερη προσαρμογή στα δεδομένα.

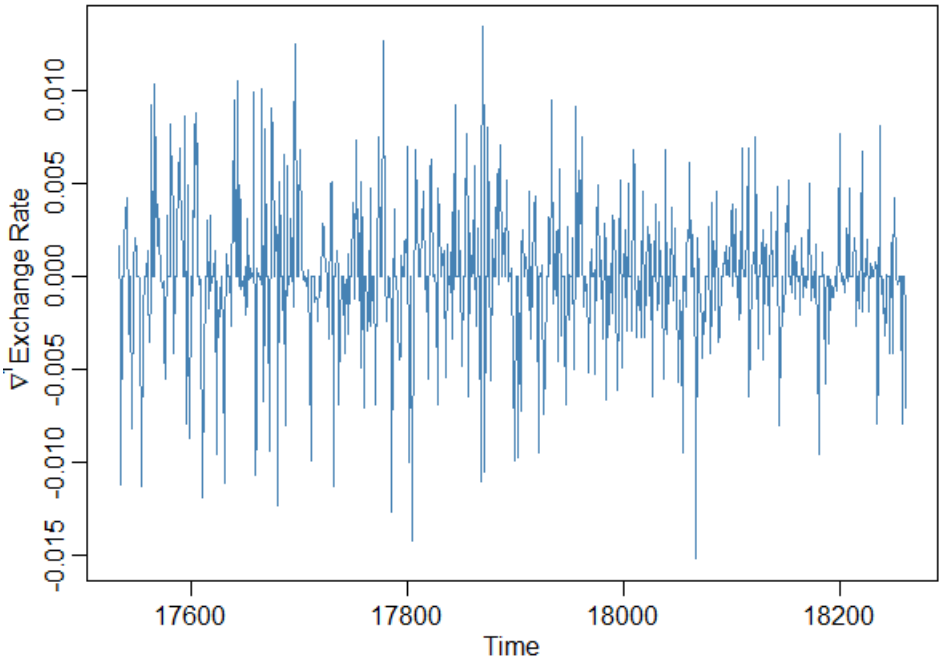
$$\bar{Y}_i = -1.16 * 10^6 + 1.308 * 10^{-2}\, t_i + -3.630 * 10^{-7} * t_i^2$$

```
Call:
lm(formula = myts ~ Time + Time2)

Residuals:
      Min        1Q    Median        3Q       Max
-0.036584 -0.008860  0.000340  0.008706  0.036802

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.164e+02  3.978e+00  -29.27   <2e-16 ***
Time         1.308e-02  4.445e-04   29.41   <2e-16 ***
Time2       -3.630e-07  1.242e-08  -29.23   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01328 on 726 degrees of freedom
Multiple R-squared:  0.7485,    Adjusted R-squared:  0.7478
F-statistic:  1080 on 2 and 726 DF,  p-value: < 2.2e-16
```

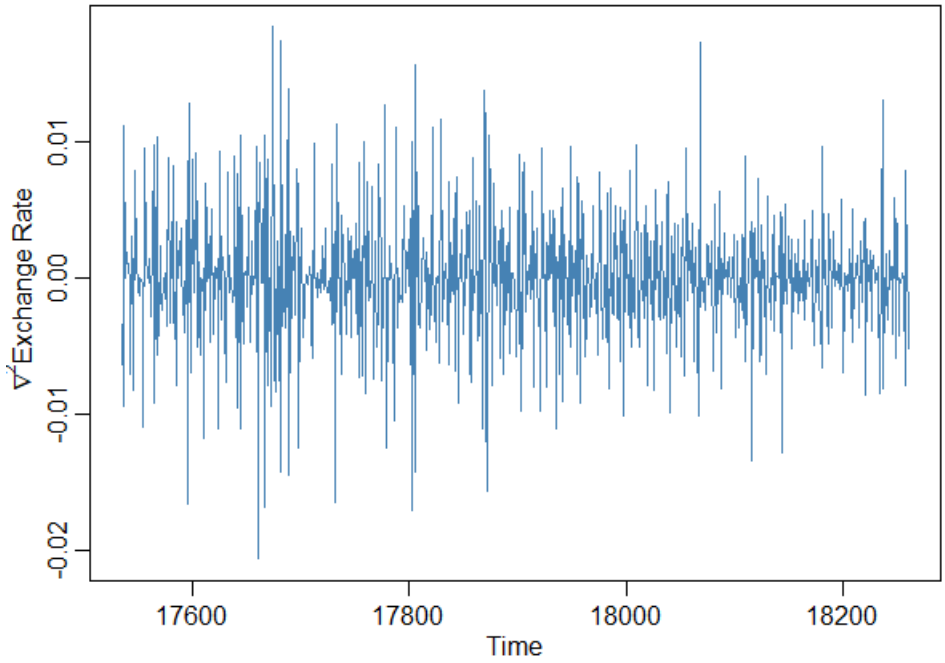
Η σχέση μεταξύ των μεταβλητών προκύπτει στατιστικά σημαντική. Ο συντελεστής προσδιορισμού του μοντέλου R^2 προκύπτει ίσως με 0.75, γεγονός που οδηγεί στο συμπέρασμα ότι το πολυώνυμο μοντέλο δευτέρου βαθμού προσαρμόζεται καλύτερα στα δεδομένα από ότι μοντέλο της απλής γραμμικής παλινδρόμησης, περιγράφοντας μεγαλύτερο ποσοστό της διακύμανσης των δεδομένων.



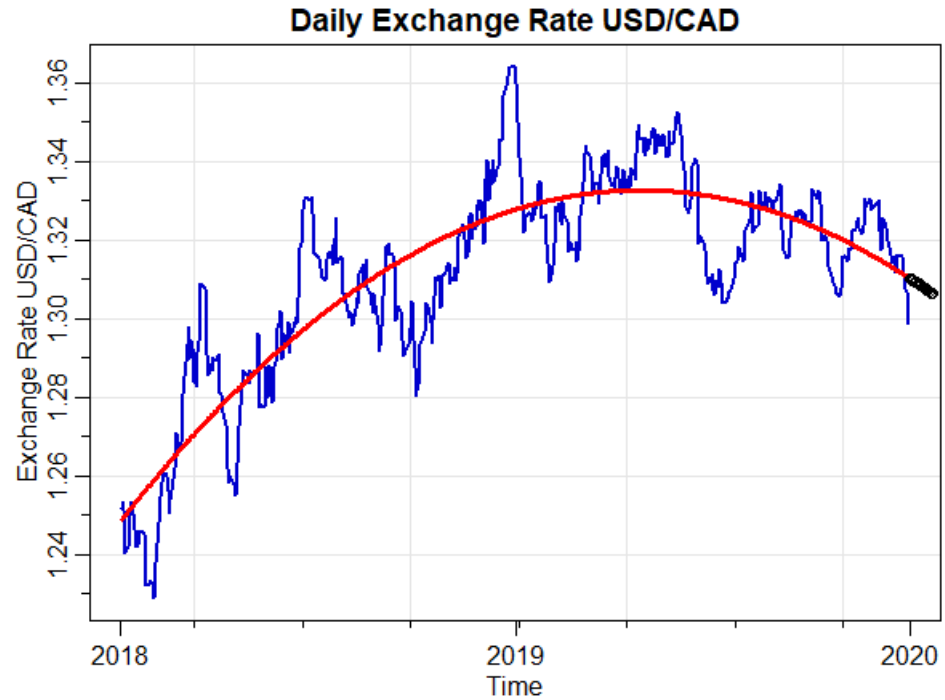
IV. Η απομάκρυνση της τάσης από τα δεδομένα είναι δυνατή με τη χρήση του διαφορών (differencing).



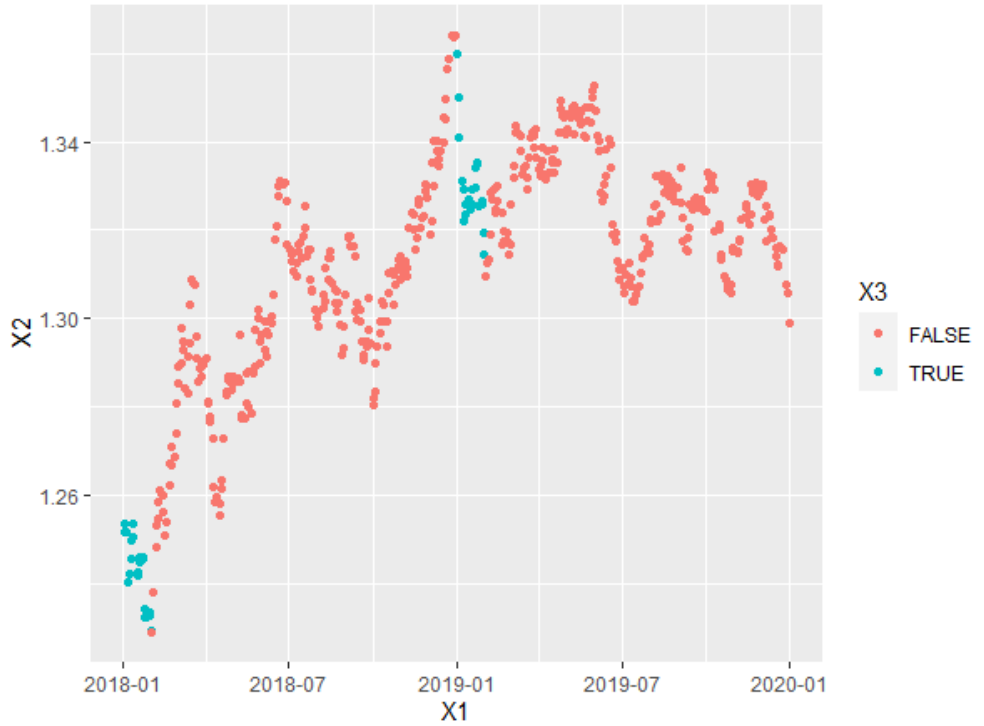
Με τη χρήση των πρώτων διαφορών απομακρύνεται η γραμμική τάση της χρονοσειράς, ενώ με τη δεύτερη η τάση που περιγράφεται από πολυώνυμο δευτέρου βαθμού. Από το διάγραμμα των πρώτων διαφορών φαίνεται ότι έχει αφαιρεθεί η τάση. Με τη χρήση των δεύτερων διαφορών φαίνεται να μην έχει εξαφανιστεί η εποχικότητα, καθώς τα δεδομένα είναι ημερήσια.



V. Κάνοντας χρήση του μοντέλου του quatradic trend, πρόκειται να προβλέψουμε την ισοτιμία συναλλάγματος για τις επόμενες 22 ημέρες. Τα αποτελέσματα που προκύπτουν υποδηλώνουν μια πτωτική τάση, ωστόσο η προσαρμογή του μοντέλου στα δεδομένα δεν φαίνεται ικανοποιητική.

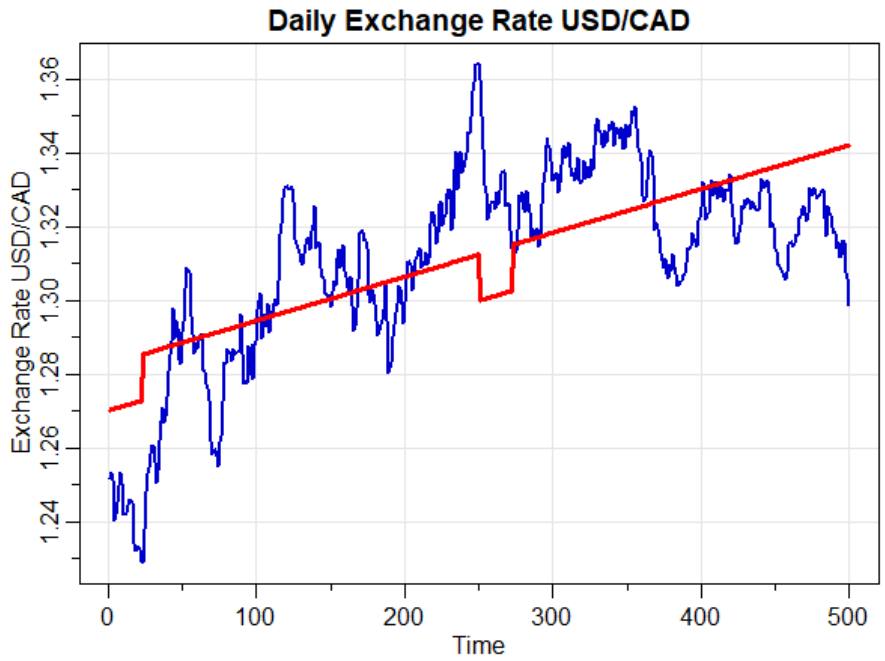


VI. Για την επίδραση του Γενάρη, προσθέτουμε μια Boolean μεταβλητή στη χρονοσειρά. Η τιμή TRUE αντιστοιχεί στο μήνα Γενάρη και η τιμή False σε όλους τους άλλους μήνες.



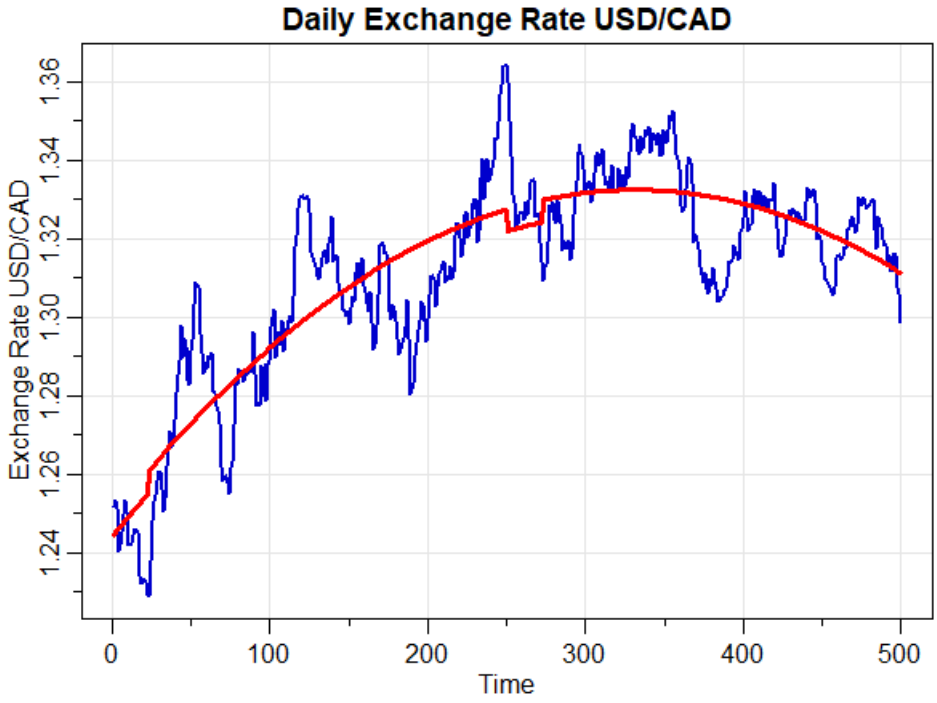
Εφαρμόζοντας το μοντέλο της γραμμικής παλινδρόμησης για τον προσδιορισμό της τάση με την επίδραση του Γενάρη, προκύπτει το εξής report:

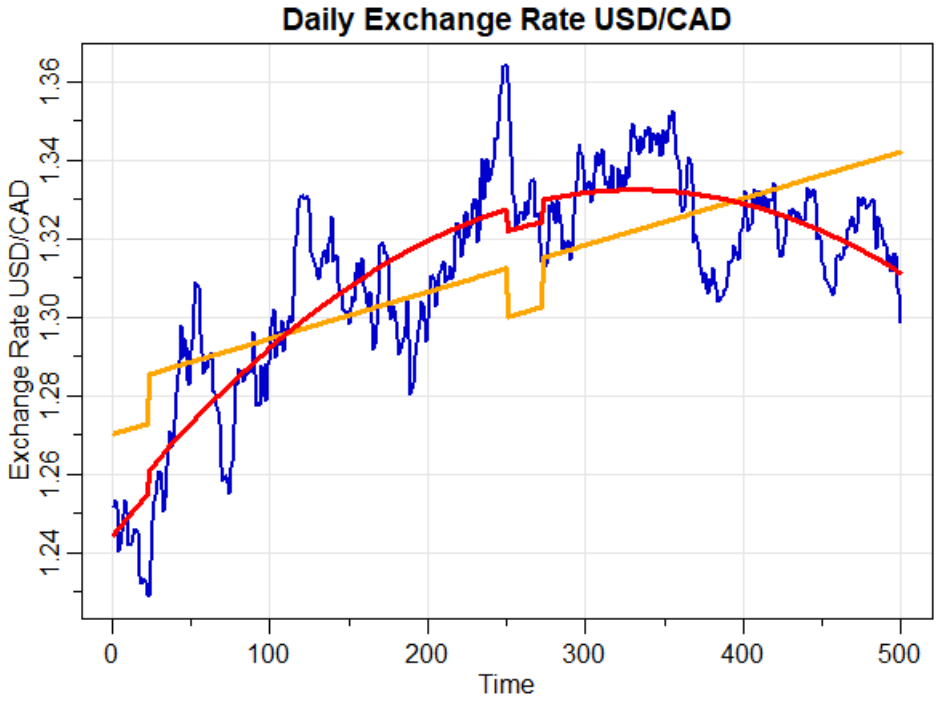
```
> lm.xreg<-tslm(x ~ trend+y)
> summary(lm.xreg)

Call:
tslm(formula = x ~ trend + y)

Residuals:
      Min        1Q    Median        3Q       Max
-0.056555 -0.014046 -0.000354  0.014891  0.059957

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.283e+00  1.836e-03 698.499  < 2e-16 ***
trend        1.190e-04  6.126e-06  19.422  < 2e-16 ***
yTRUE       -1.244e-02  3.121e-03  -3.985 7.75e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01917 on 497 degrees of freedom
Multiple R-squared:   0.48,     Adjusted R-squared:  0.4779
F-statistic: 229.3 on 2 and 497 DF,  p-value: < 2.2e-16
```



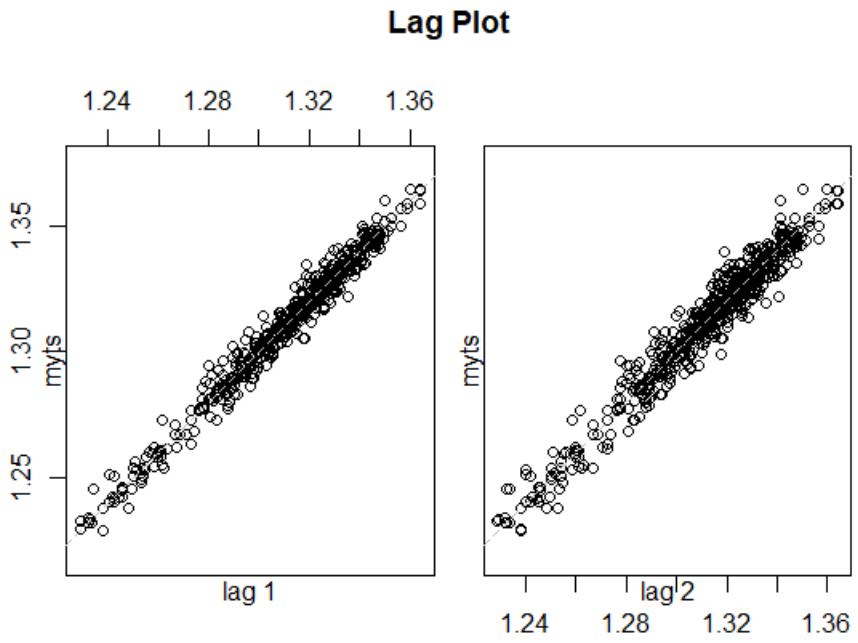
VII. Εφαρμόζοντας το μοντέλο του quadradic trend με την επίδραση του Γενάρη, προκύπτει το εξής report:

```
> lm.poly.xreg<-tslm(x ~ trend+I(trend^2)+y)
> summary(lm.poly.xreg)

Call:
tslm(formula = x ~ trend + I(trend^2) + y)

Residuals:
      Min        1Q    Median        3Q       Max
-0.036794 -0.008631  0.000551  0.008786  0.038073

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.250e+00  1.889e-03 661.439   <2e-16 ***
trend        4.982e-04  1.672e-05  29.800   <2e-16 ***
I(trend^2)  -7.504e-07  3.201e-08 -23.444   <2e-16 ***
yTRUE       -5.586e-03  2.172e-03  -2.572   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01321 on 496 degrees of freedom
Multiple R-squared:  0.7533,    Adjusted R-squared:  0.7518
F-statistic: 504.9 on 3 and 496 DF,  p-value: < 2.2e-16
```





VIII. Από τα διαγράμματα διασποράς πρώτων και δεύτερων διαφορών προκύπτει ότι τα δεδομένα έχουν γραμμικό πρότυπο, γεγονός που υποδηλώνει την ύπαρξη αυτοσυσχέτισης (autocorrelation). Το γεγονός ότι τα δεδομένα βερίσκονται με μεγάλη πυκνότητα πάνω στην διαγώνιο για υστέρηση ίση με ένα (lag=1) σε σχέση το διάγραμμα lag = 2, αποτελεί ένδειξη ισχυρότερης αυτοσυσχέτισης. Δεν παρατηρείται η ύπαρξη κάποιας ακραίας τιμής ούτε εποχικότητας.



IX. Από τον υπολογισμό της αυτοσυσχέτισης για υστέρηση ίση με ένα, προκύπτει η τιμή 0.986. Για υστέρηση ίση με 2, η αυτοσυσχέτιση είναι ελαφρώς ασθενέστερη με τιμή 0.972.

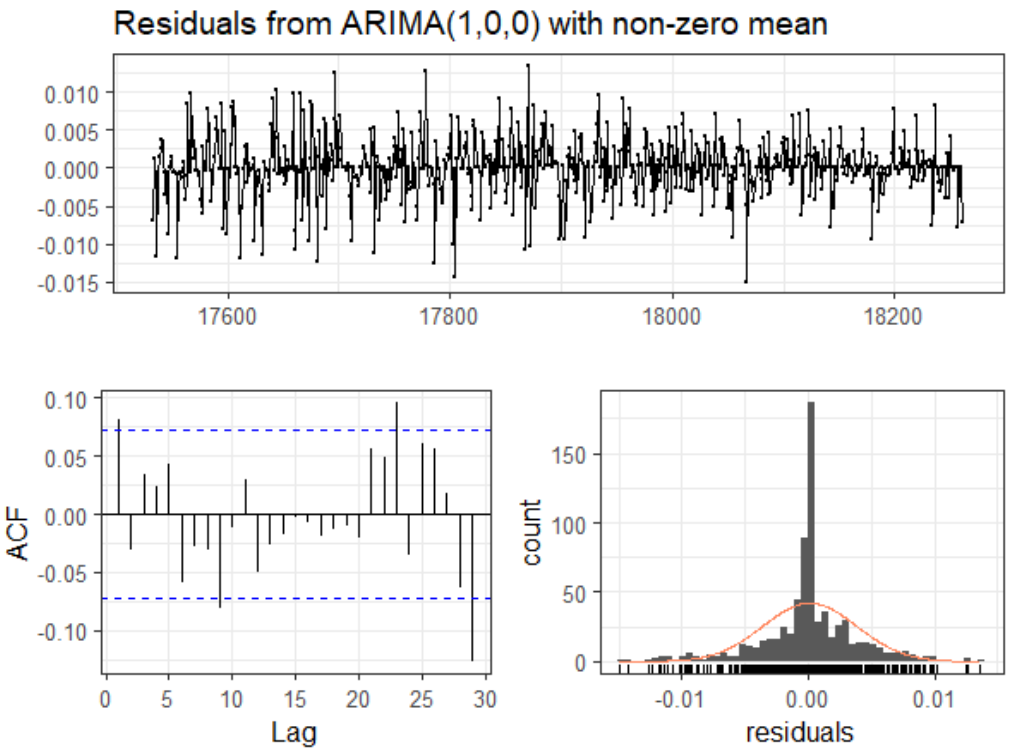
X. Στο ερώτημα αυτό δημιουργούμε μια συνάρτηση για τον υπολογισμό της αυτοσυσχέτισης μεταξύ δυο χρονοσειρών από την αρχή. Το αποτέλεσμα που προκύπτει είναι ίδιο, 0.986.

```
#Autocorrelation function
AC <- function(y,k) {

  X <- ts.union(yt = y, yt2 = stats::lag(x = y, k = k))
  c0 <- var(y)
  m <- mean(y)
  n <- length(y)
  ct <- sum((X[, 1] - m) * (X[, 2] - m), na.rm = TRUE) / (n - 1)
  rt <- ct / c0

  return(rt)
}

AC(myts,1)
```

XI. Το μοντέλο AR(1) ή αλλιώς ARIMA(1,0,0) έχει τον εξής τύπο:

$$Y_t = c + \varphi Y_{t-1} + e_t$$

Όπου Y_{t-1} αποτελεί το προηγούμενο χρονικό σημείο και e_t αποτελεί το λευκό θόρυβο.

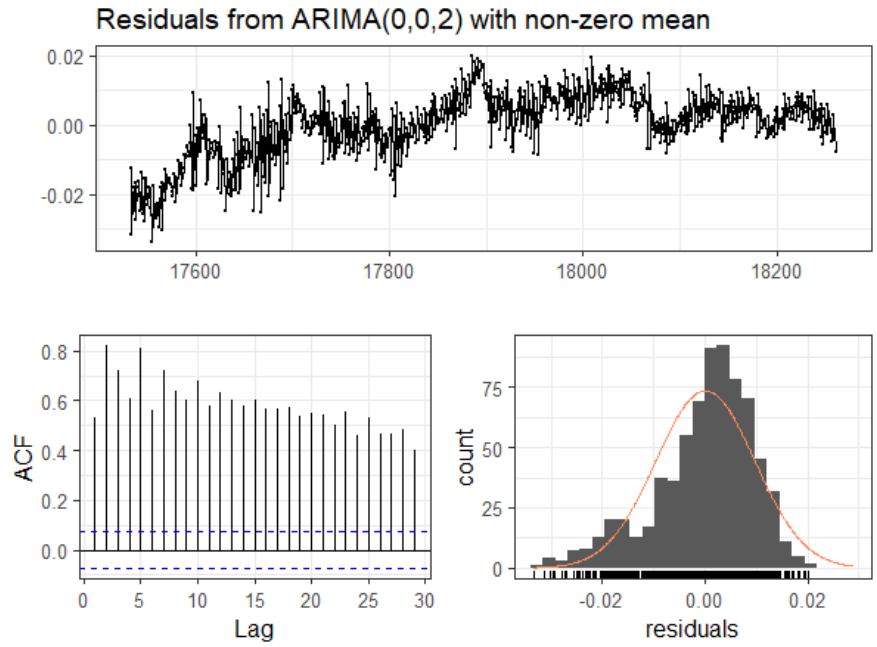



```
> ar.model = arima(myts, order = c(1,0,0)); ar.model

Call:
arima(x = myts, order = c(1, 0, 0))

Coefficients:
          ar1  intercept
      0.9918     1.3058
s.e.  0.0047     0.0145

sigma^2 estimated as 1.38e-05:  log likelihood = 3042.65,  aic = -6079.3
```

Το μοντέλο MA(1) ή αλλιώς ARIMA(0,0,1) έχει τον εξής τύπο:

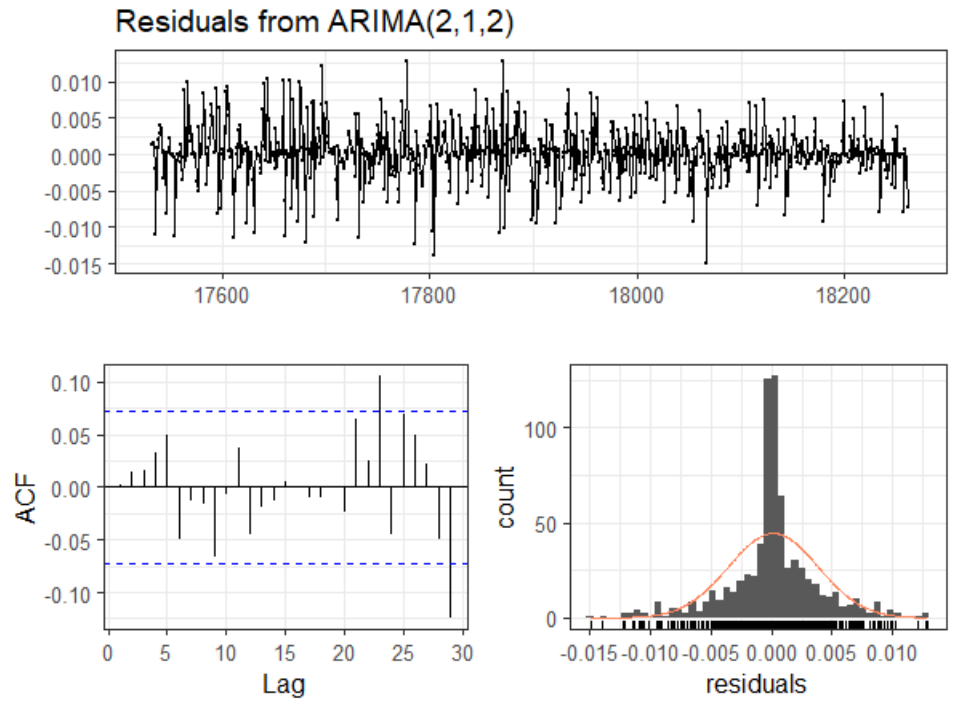
$$Y_t = c + \varphi e_{t-1} + e_t$$

Όπου e_{t-1} αποτελεί το σφάλματα από την προηγούμενη πρόβλεψη και e_t αποτελεί το λευκό θόρυβο. Κάθε προβλεπόμενη τιμή μπορεί να θεωρηθεί ως ένας σταθμισμένος μέσος των προηγούμενων σφαλμάτων πρόβλεψης.




```
> ma.model = arima(myts, order = c(0,0,2)); ma.model

Call:
arima(x = myts, order = c(0, 0, 2))

Coefficients:
          ma1     ma2  intercept
      1.4460  0.7406     1.3116
s.e.  0.0264  0.0187     0.0011

sigma^2 estimated as 9.261e-05:  log likelihood = 2349.36,  aic = -4690.72
```

XII. Το μοντέλο ARIMA(2,1,2) εφαρμόζει την πρώτη διαφορά για φτάσει σε στάσιμη χρονοσειρά και στη συνέχεια προσαρμόζει το μοντέλο ARMA(p=2, q=2).



Από την μελέτη των υπολοίπων προκύπτει ότι τα υπόλοιπα δεν είναι συσχετισμένα και συνεπώς δεν περιλαμβάνουν πληροφορία που θα μπορούσε να αυξήσει την ακρίβεια της πρόβλεψης.

```
> arima.model = arima(myts, order = c(2,1,2)); arima.model

Call:
arima(x = myts, order = c(2, 1, 2))

Coefficients:
         ar1     ar2      ma1      ma2
      0.5355  0.4220  -0.4538  -0.5219
s.e.  0.4942  0.4858   0.4743   0.4693

sigma^2 estimated as 1.367e-05:  log likelihood = 3043.92,  aic = -6077.85
```

Για την αξιολόγηση της ποιότητα της πρόβλεψης θα χρησιμοποιήσουμε τις εξής μετρικές:

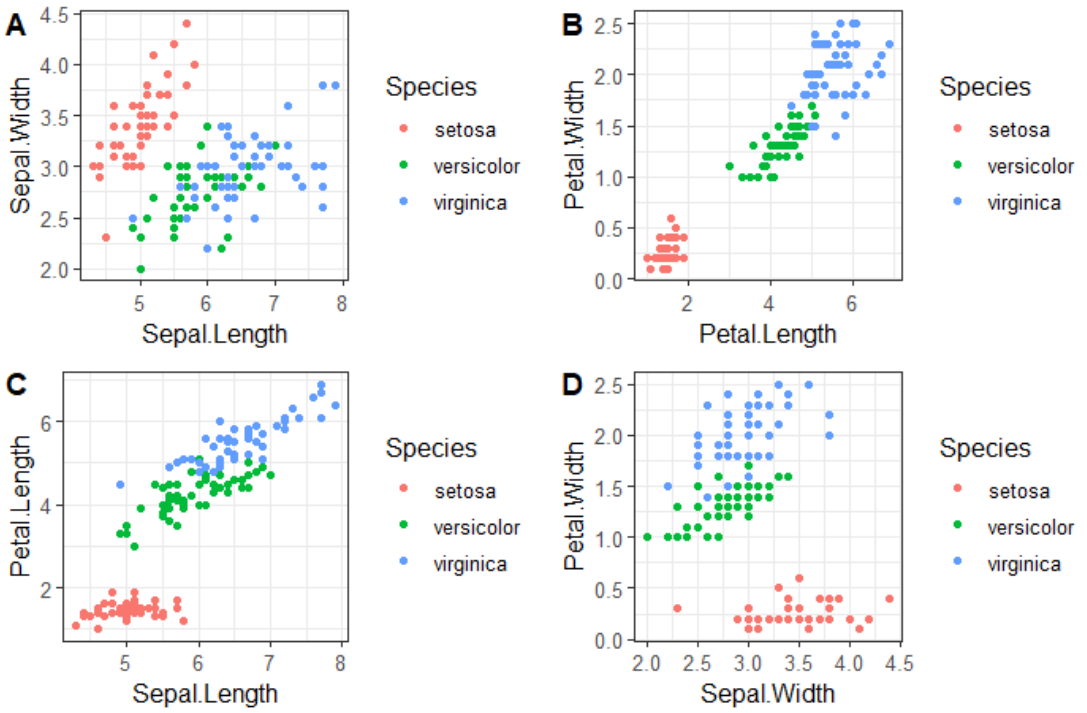
- *MAE – Mean Absolute Error:* που αντιστοιχεί στη μέση τιμή της απόλυτης τιμής των διαφορών μεταξύ της πραγματικής τιμής και της προβλεπόμενης τιμής
- *RMSE – Root Mean Squared Error*: η τυπική απόκλιση των υπολοίπων
- *MAPE – Mean Absolute Percentage Error:* αποτελεί το μέσο όρο της διαφοράς των σφαλμάτων πρόβλεψης, διαιρώντας τη με την πραγματική τιμή. Δεν επιτρέπει την μηδενικής τιμές και δίνει μεγάλο βάρος σε ακραίες τιμές και θετικά σφάλματα. Το πλεονέκτημα της μετρικής είναι ότι ανεξάρτητη κλίμακας.

Συγκρίνοντας τα σφάλματα των διαφορετικών μοντέλων προκύπτει ότι η απόδοση του μοντέλο AR(1) και ARIMA(2,1,2) είναι όμοια και βέλτιστη του μοντέλου MA(1).

	AR(1)	MA(1)	ARIMA(2,1,2)
MAE	0.002	0.007	0.002
RMSE	0.004	0.010	0.004
MAPE	0.183	0.572	0.184

2.1.7. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 7

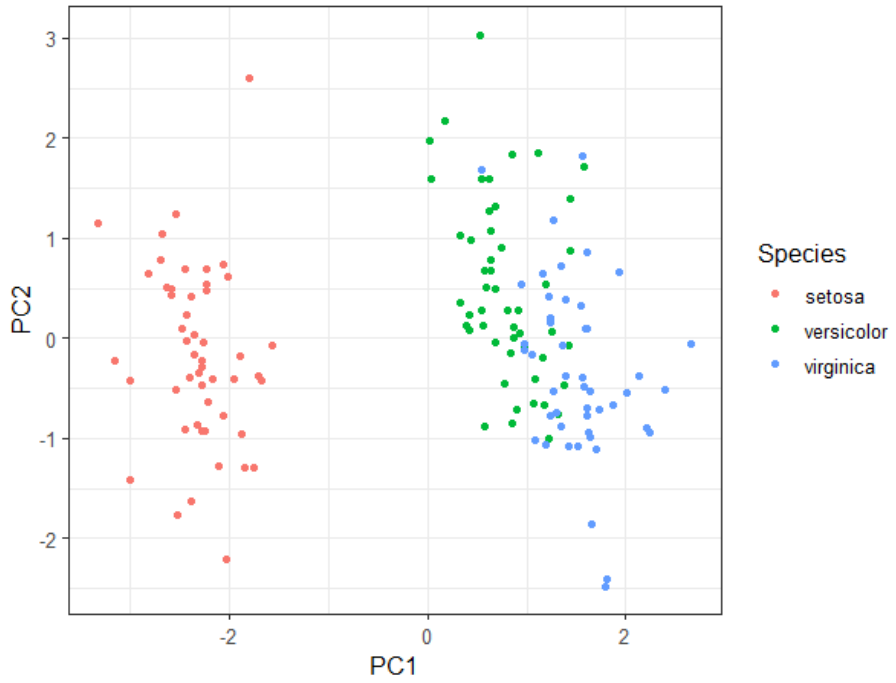
Το σύνολο δεδομένων που χρησιμοποιείται είναι το iris, το οποίο περιλαμβάνει πληροφορίες για τα χαρακτηριστικά διαφορετικών ειδών λουλουδιών. Από τα διαγράμματα διασποράς, προκύπτει ότι με βάση τα χαρακτηριστικά Petal.Length και Petal.Width οι κλάσεις είναι καλύτερα διαχωρίσιμες. Η ανάλυση κυρίων συνιστωσών επιτρέπει την αναγνώριση προτύπων στα δεδομένα.



Τα βασικά βήματα της ανάλυσης κυριών συνιστωσών είναι η προετοιμασία των δεδομένων ώστε να περιλαμβάνουν μόνο αριθμητικές στήλες και ο μετασχηματισμός των δεδομένων ώστε να έχουν μέση τιμή μηδέν και μοναδιαία διακύμανση.

Εφαρμόζοντας την ανάλυση των κυριών συνιστωσών, προκύπτει η τυπική απόκλιση κάθε συνιστώσας, ο rotation matrix, καθώς και οι νέες συνιστώσες που αποτελούν γραμμικό συνδυασμό των αρχικών.

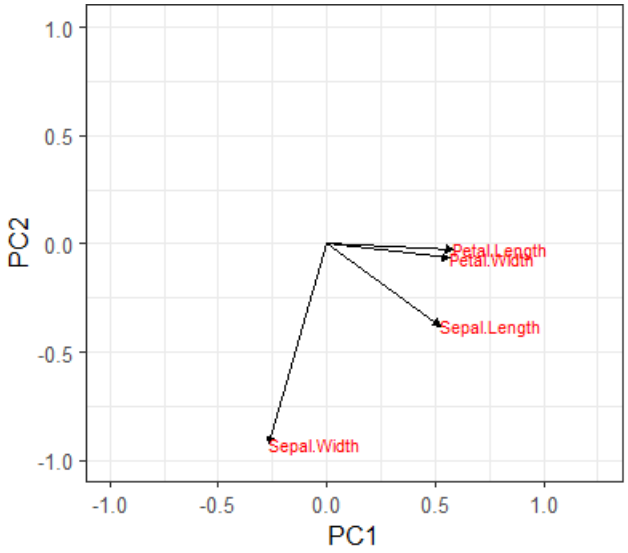
Απεικονίζοντας τα δεδομένα με τις δυο πρώτες συνιστώσες, παρατηρείται ότι οι κλάσεις είναι διακριτά διαχωρίσιμες μεταξύ τους.



Από το notation matrix μπορούμε να δούμε σε τι ποσοστό συμμετέχουν τα δεδομένα στις κύριες συνιστώσες. Παρατηρείται ότι η μεταβλητή Petal.Length συμμετέχει περισσότερο στη PC1, ενώ η Sepal.Width στη PC2

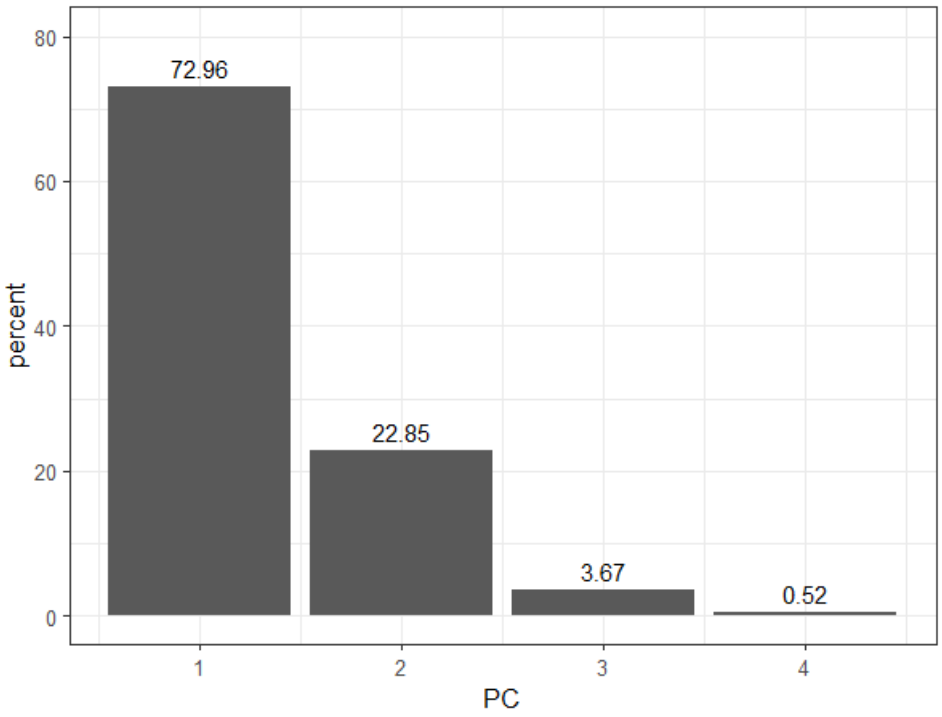
```
> pca$rotation
                    PC1         PC2        PC3        PC4
Sepal.Length  0.5210659 -0.37741762  0.7195664  0.2612863
Sepal.Width  -0.2693474 -0.92329566 -0.2443818 -0.1235096
Petal.Length  0.5804131 -0.02449161 -0.1421264 -0.8014492
Petal.Width   0.5648565 -0.06694199 -0.6342727  0.5235971
```

Από το διάγραμμα που ακολουθεί φαίνεται ξεκάθαρα ότι η μεταβλητές Petal.Length, Petal.Width και Sepal.Length συμμετέχουν κυρίως στη συνιστώσα PC1, ενώ η μεταβλητή Sepal.Width κυριαρχεί στη PC2.



Τέλος είναι δυνατό να ελέγξουμε το ποσοστό της διακύμανσης των αρχικών δεδομένων που εξηγείται από κάθε συνιστώσα. Συμπεραίνεται ότι η πρώτη συνιστώσα εξηγεί το 73% της συνολικής διακύμανσης των δεδομένων, η δεύτερη το 23%, η τρίτη το 4% και η τελευταία το 0.5%.

```
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```



Ο πίνακας συνδιακυμάνσεων αποτελεί τον πυρήνα της PCA. Τα ιδιοδιανύσματα αποτελούν τις κύριες συνιστώσες και καθορίζουν τις συντεταγμένες των σημείων στον νέο χώρο που προβάλλονται και οι ιδιοτιμές την διακύμανση των νέων δεδομένων.

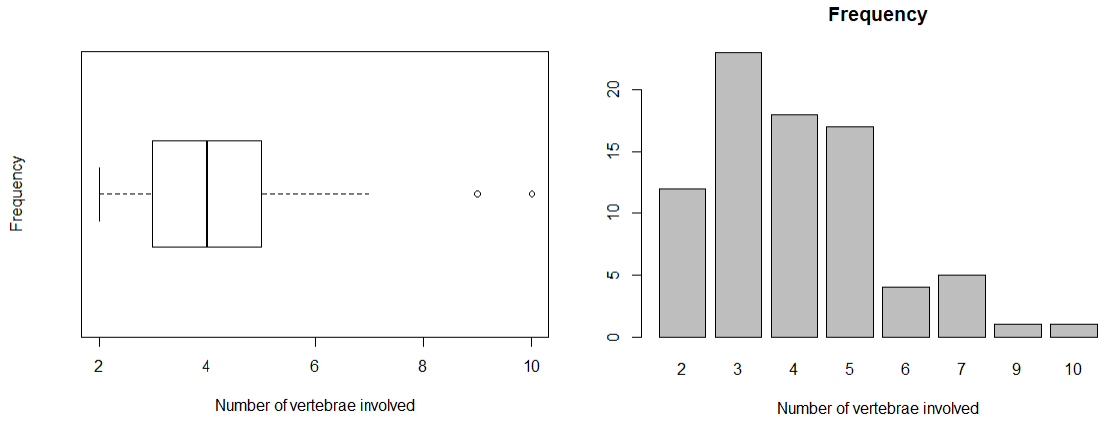
2.2. ΜΕΡΟΣ 2ο

2.2.1. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 1

Στη συγκεκριμένη άσκηση γίνεται χρήση του συνόλου δεδομένων kyphosis. Το σύνολο αυτό δεδομένων περιλαμβάνει 4 μεταβλητές και 81 παρατηρήσεις, ενώ πραγματεύεται την παρουσία ή απουσία κύφωσης σε παιδιά. Τα περιγραφικά στατιστικά του συνόλου παρουσιάζονται στην εικόνα που ακολουθεί.

```
> summary(kyphosis)
    Kyphosis        Age              Number          Start
 absent :64   Min.   :  1.00   Min.   : 2.000   Min.   : 1.00
 present:17   1st Qu.: 26.00   1st Qu.: 3.000   1st Qu.: 9.00
              Median : 87.00   Median : 4.000   Median :13.00
              Mean   : 83.65   Mean   : 4.049   Mean   :11.49
              3rd Qu.:130.00   3rd Qu.: 5.000   3rd Qu.:16.00
              Max.   :206.00   Max.   :10.000   Max.   :18.00
```

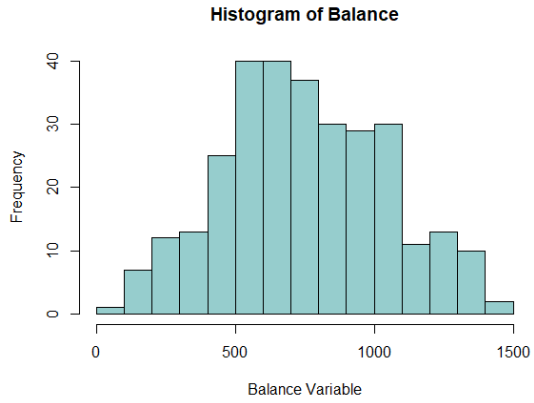
Για την εξέταση της μεταβλητής 'Number', που αντιστοιχεί στον αριθμό των σπονδύλων που έχουν επηρεαστεί από την πάθηση, πραγματοποιείται απεικόνιση της με ιστόγραμμα και θηκόγραμμα, οδηγώντας στην ανάδειξη δυο ακραίων τιμών (outliers).

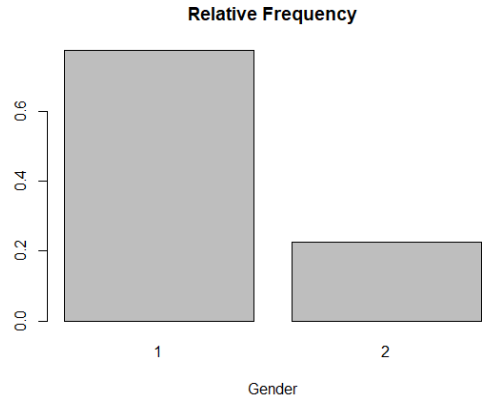


Με χρήση της εντολής which ή identify, είναι δυνατή η ανάκτηση του index αυτών των ακραίων παρατηρήσεων και η απομάκρυνση τους από τα δεδομένα. Τα index των outliers είναι 43 και 53.

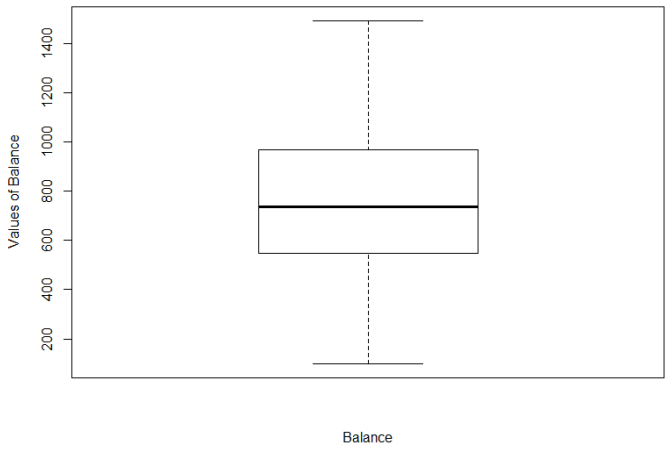
2.2.2. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 2

Για το σύνολο δεδομένων capital.csv, ελέγχουμε αρχικά τον τύπο των μεταβλητών που περιλαμβάνει και μετατρέπουμε τη μεταβλητή Gender σε κατηγορική. Στη συνέχεια παρουσιάζονται τα ιστογράμματα σχετικής συχνότητας των μεταβλητών Balance και Gender, αντίστοιχα.



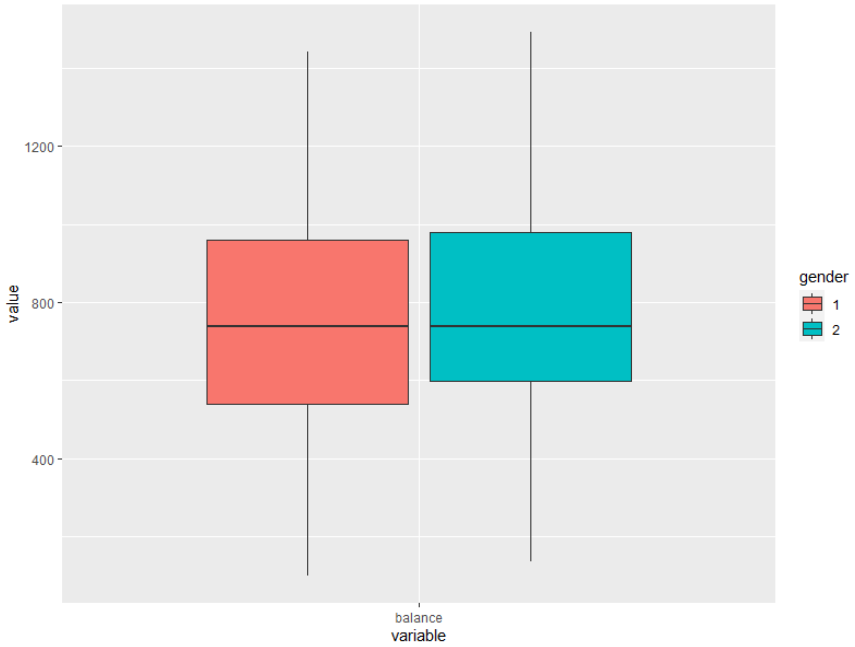


Το θηκόγραμμα αποτελεί έναν τρόπο απεικόνισης της κατανομής των δεδομένων βασισμένο σε 5 περιγραφικά μέτρα, το ελάχιστο, το μέγιστο, τη διάμεσο και το πρώτο και τρίτο τεταρτημόριο. Το γεγονός ότι το θηκόγραμμα βρίσκεται στο κέντρο του διαγράμματος, όπως συμβαίνει και για τη μεταβλητή Balance, υποδηλώνει ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Σε άλλες περιπτώσεις, μπορεί να συνεισφέρει στην ανάδειξη ακραίων τιμών.



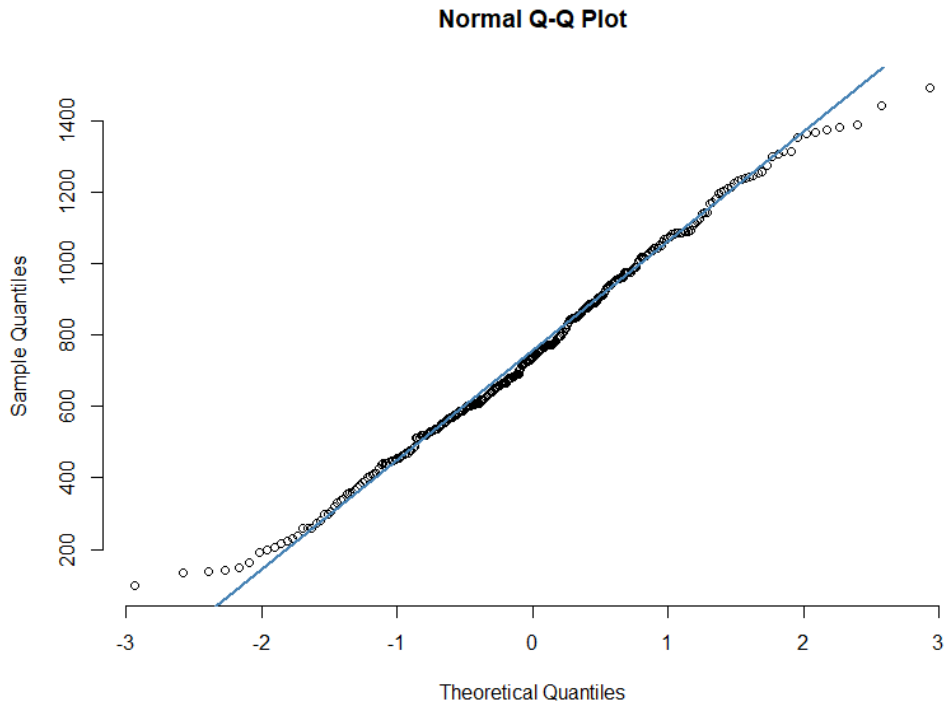
Το θηκόγραμμα χρησιμοποιείται και για την ανάλυση της σχέσης ανάμεσα σε μια κατηγορική και μια συνεχή μεταβλητή. Από το διάγραμμα που ακολουθεί δεν προκύπτει κάποια σημαντική διαφορά για το Balance ανδρών και γυναικών, καθώς η διάμεσος είναι παρόμοια. Θα μπορούσε να θεωρηθεί ότι το Balance των γυναικών είναι ελαφρώς πιο διασκορπισμένο, ωστόσο δεν παρατηρούνται ακραίες τιμές.

```
 Descriptive statistics by group
group: 1
        vars   n   mean     sd median trimmed    mad min  max range skew kurtosis    se
balance    1 232 746.51 294.52  738.5  743.88 312.83  99 1443  1344 0.08    -0.64 19.34
gender*    2 232   1.00   0.00    1.0    1.00   0.00   1    1     0  NaN      NaN  0.00
------------------------------------------------------------------------------------------
group: 2
        vars  n   mean     sd median trimmed    mad min  max range skew kurtosis   se
balance    1 68 778.13 295.22    737  773.38 287.62 135 1493  1358 0.21    -0.35 35.8
gender*    2 68   2.00   0.00      2    2.00   0.00   2    2     0  NaN      NaN  0.0
```



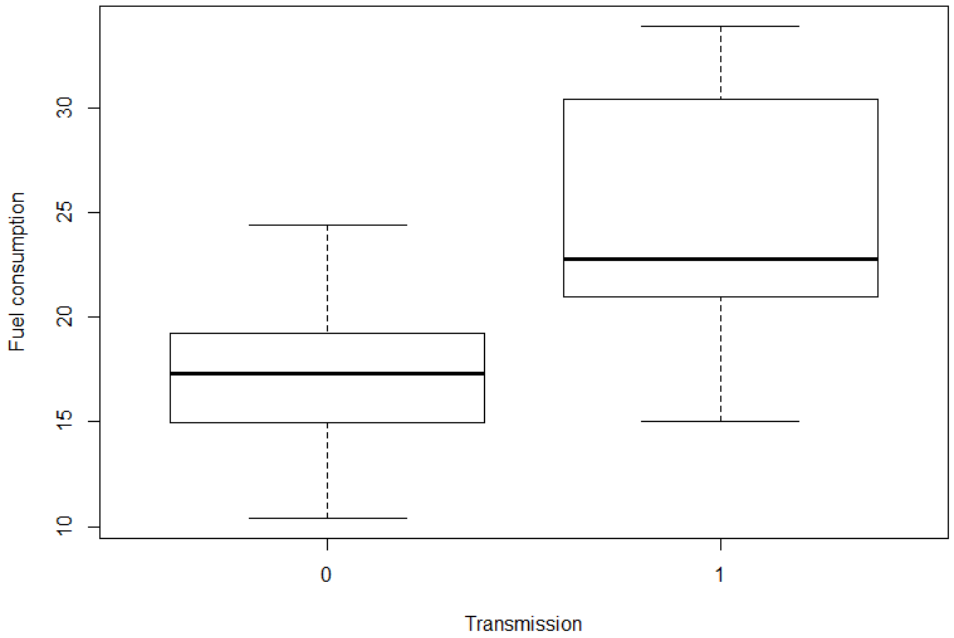

```
> summary(capital)
    balance        gender
 Min.   :  99.0   1:232
 1st Qu.: 550.8   2: 68
 Median : 737.0
 Mean   : 753.7
 3rd Qu.: 964.2
 Max.   :1493.0
```

Την κανονικότητα της συνεχούς μεταβλητής Balance έρχεται να επιβεβαιώσει και το Q-Q Plot που ακολουθεί.



2.2.3. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 3

Απεικονίζοντας τα δεδομένα σε θηκόγραμμα σε σχέση με την κατηγορική μεταβλητή Transmission, που αφορά το μηχανικό και αυτόματο αυτοκίνητο, προκύπτει ότι η κατανάλωση καυσίμου διαφέρει για τις δυο κατηγορίες. Με το μηχανικό αυτοκίνητο να παρουσιάζει μεγαλύτερη κατανάλωση.



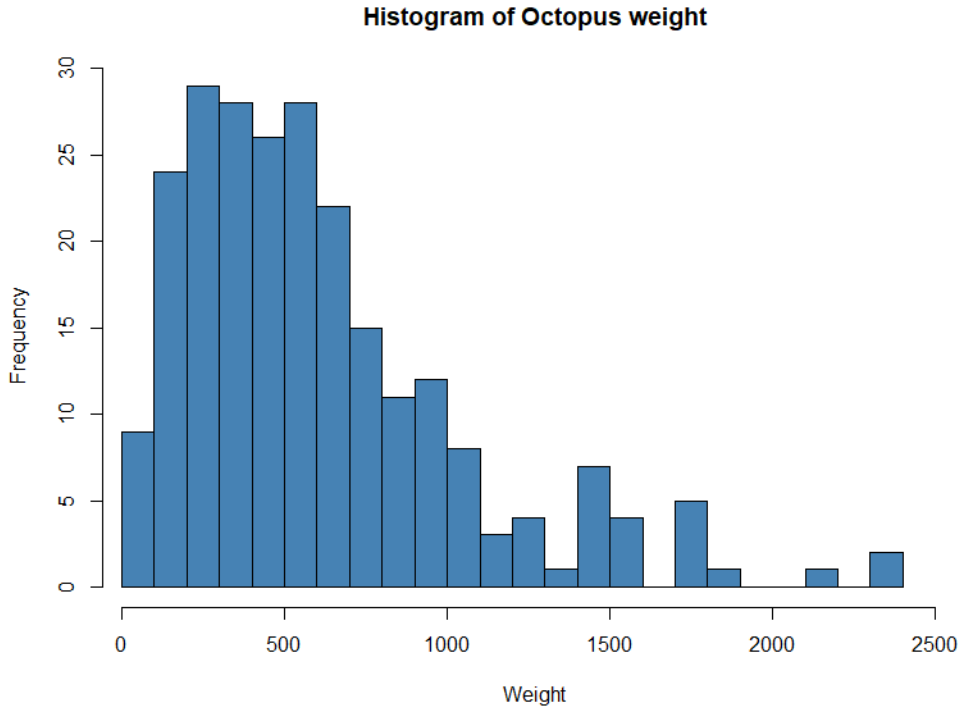
Θεωρώντας ότι το σύνολο δεδομένων mtcars ακολουθεί την κανονική κατανομή, πρόκειται να αποδείξουμε αν τα δείγματα κατανάλωσης καυσίμου για μηχανικό και αυτόματο αυτοκίνητο είναι ανεξάρτητα. Ειδικότερα ότι προέρχονται από ασυσχέτιστους πληθυσμούς και τα δείγματα δεν επηρεάζονται το ένα από το άλλο.

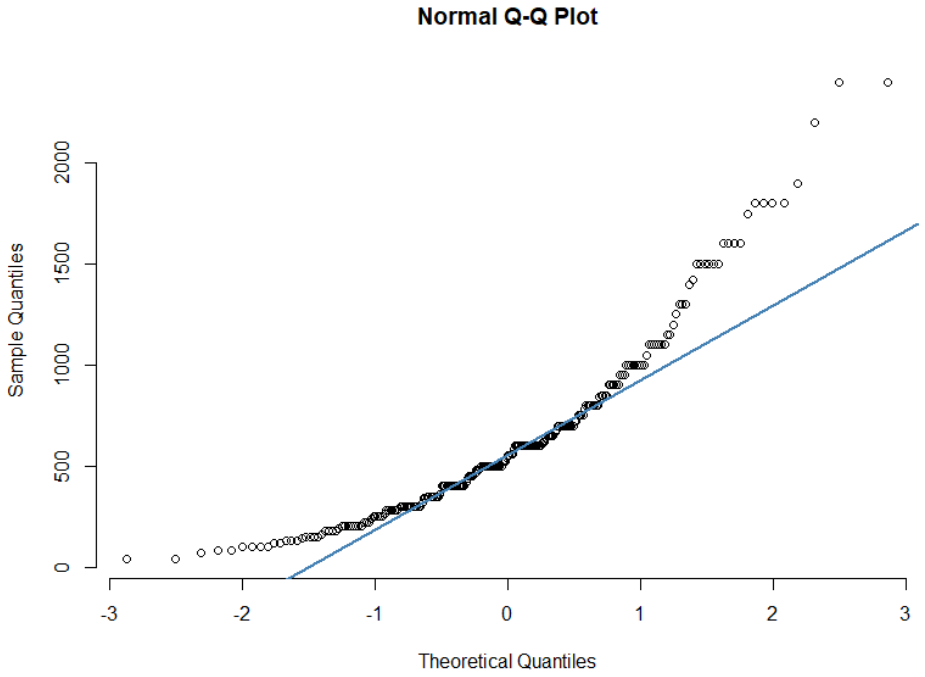
Πρόκειται να εφαρμόσουμε τη συνάρτηση t.test για τον υπολογισμό της διαφοράς των μέσων των δυο δειγμάτων. Για το σύνολο δεδομένων, η μέση κατανάλωση για αυτόματο αυτοκίνητο είναι 17.147mpg και για μηχανικό 24.392mpg. Με 95% συντελεστή εμπιστοσύνης, η διαφορά της μέσης κατανάλωσης είναι μεταξύ 3.210mpg και 11.280mpg.

```
        Welch Two Sample t-test

data:  mtcars$mpg by mtcars$am
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.280194  -3.209684
sample estimates:
mean in group 0 mean in group 1
       17.14737        24.39231
```

2.2.4. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 4

Για το σύνολο δεδομένων OctopusF, υπολογίζονται τα περιγραφικά μέτρα και προκύπτει μέση τιμή ίση με 639.6 kg και τυπική απόκλιση 445.9. Τα δεδομένα παρουσιάζουν skewness με θετική ασυμμετρία, όπως παρατηρείται και στο ιστόγραμμα που ακολουθεί. Από το Q-Q Plot, δίνεται ακόμα μια ένδειξη ότι η μεταβλητή δεν ακολουθεί την κανονική κατανομή.





Από το Shapiro-Wilk τεστ κανονικότητας, προκύπτει ότι η μηδενική υπόθεση του τεστ ότι τα δεδομένα ακολουθούν την κανονική κατανομή δεν μπορεί να επιβεβαιωθεί, καθώς το p-value έχει τιμή 1.863e-12 κατά πολύ μικρότερη από το 0.05.

2.2.5. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 5

Στο σύνολο δεδομένων survey, η στήλη Smoke καταγράφει το βαθμό καπνίσματος μαθητών και τη στήλη Exer το βαθμό σωματικής άσκησης. για την εύρεση αν οι δυο μεταβλητές είναι ανεξάρτητες αρχικά υπολογίζεται ο πίνακας συνάφειας.

		Exercise		
Smoking		***Freq***	***None***	***Some***
	Heavy	7	1	3
	Never	87	18	84
	Occas	12	3	4
	Regul	9	1	7

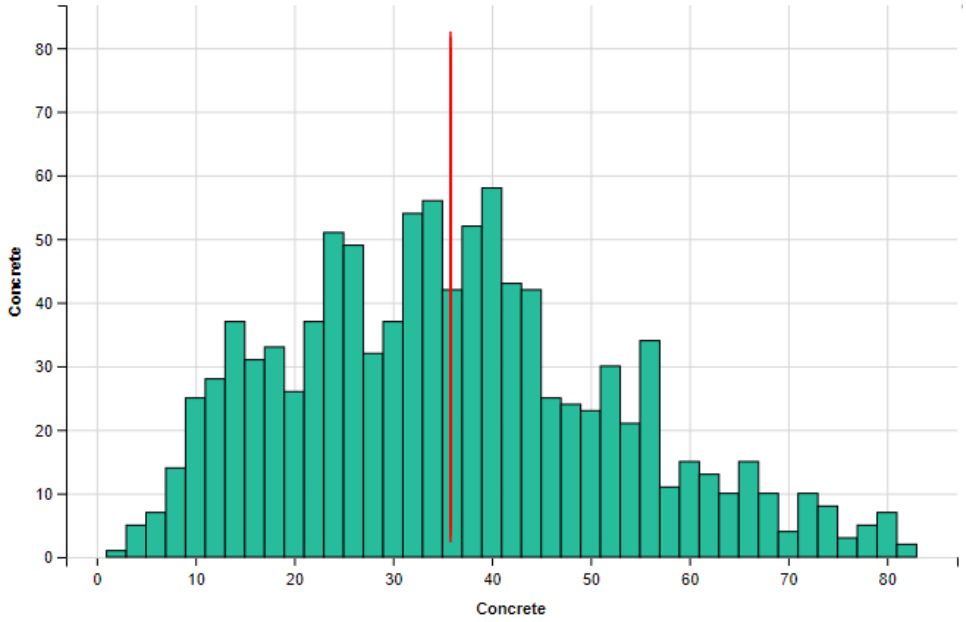
Η υπόθεση που ελέγχεται είναι αν η συνήθεια καπνίσματος των μαθητών είναι ανεξάρτητη από το επίπεδο σωματικής άσκησης με συντελεστή σημαντικότητας 0.05. Εφαρμόζοντας το x2-test στον πίνακα συνάφειας προκύπτει ότι η τιμής της p-value είναι 0.4828 μεγαλύτερη από τον συντελεστή σημαντικότητας 0.05, με αποτέλεσμα να μην μπορεί να απορριφθεί η μηδενική υπόθεση ότι η συνήθεια του καπνίσματος είναι ανεξάρτητη από το επίπεδο της σωματικής άσκησης.

```
        Pearson's Chi-squared test

data:  tab
X-squared = 5.4885, df = 6, p-value = 0.4828
```

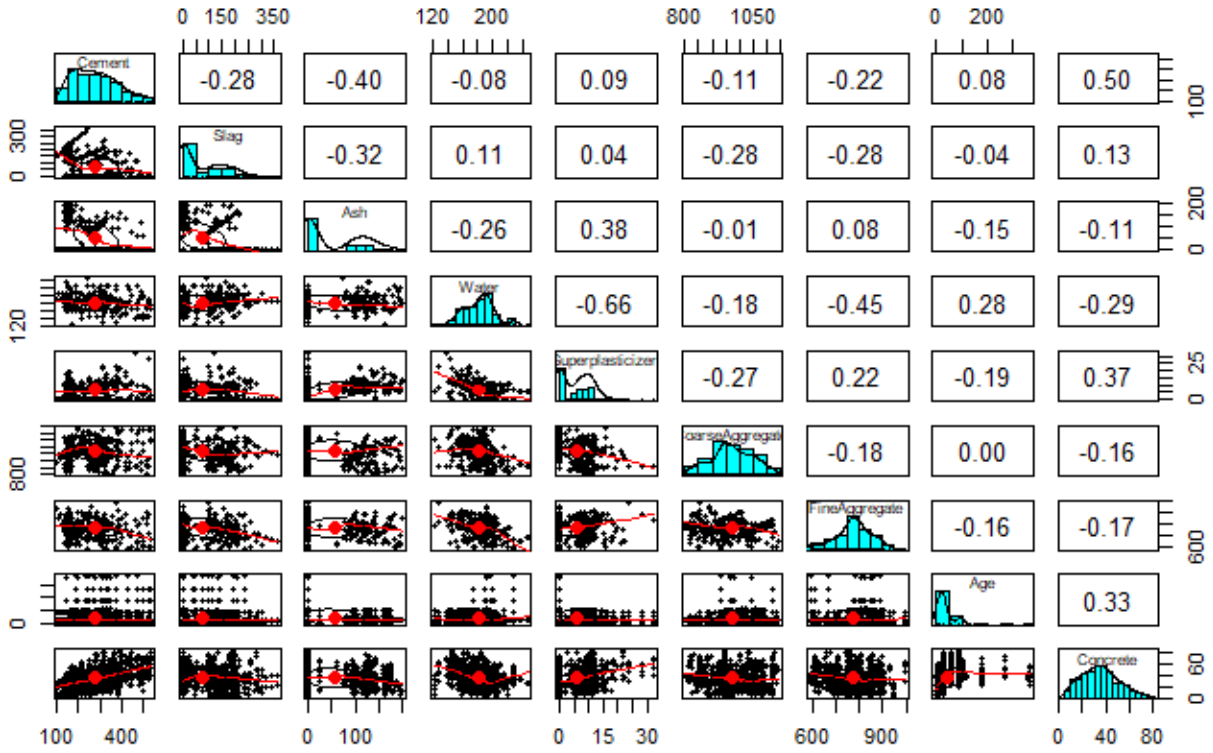
2.2.6. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 6

Στη συγκεκριμένη άσκηση γίνεται χρήση του συνόλου δεδομένων Concrete_Data.xls, που περιλαμβάνει μεταβλητές που επηρεάζουν την ανθεκτικότητα του σκυροδέματος. Από το ιστόγραμμα που ακολουθεί προκύπτει ότι η μεταβλητή της ανθεκτικότητας του σκυροδέματος παρουσιάζει ελαφριά θετική ασυμμετρία, ωστόσο η πλειονότητα των τιμών της μεταβλητής εντοπίζονται κοντά στη μέση τιμή 35.82, χωρίς πολλές ακραίες τιμές.

	vars	n	mean	sd	median	trimmed	mad	min	max	range
Cement	1	1030	281.17	104.51	272.90	273.47	117.72	102.00	540.0	438.00
Slag	2	1030	73.90	86.28	22.00	62.43	32.62	0.00	359.4	359.40
Ash	3	1030	54.19	64.00	0.00	46.85	0.00	0.00	200.1	200.10
Water	4	1030	181.57	21.36	185.00	181.19	19.27	121.75	247.0	125.25
Superplasticizer	5	1030	6.20	5.97	6.35	5.56	7.87	0.00	32.2	32.20
CoarseAggregate	6	1030	972.92	77.75	968.00	973.49	68.64	801.00	1145.0	344.00
FineAggregate	7	1030	773.58	80.18	779.51	776.41	67.44	594.00	992.6	398.60
Age	8	1030	45.66	63.17	28.00	32.53	31.13	1.00	365.0	364.00
Concrete	9	1030	35.82	16.71	34.44	34.96	16.20	2.33	82.6	80.27



Μελετώντας τη συσχέτιση μεταξύ των μεταβλητών προκύπτει ότι η πιοισχυρή συσχέτιση παρουσιάζεται μεταξύ σκυροδέματος ανθεκτικότητας και του Cement και είναι ίση με 0.5.

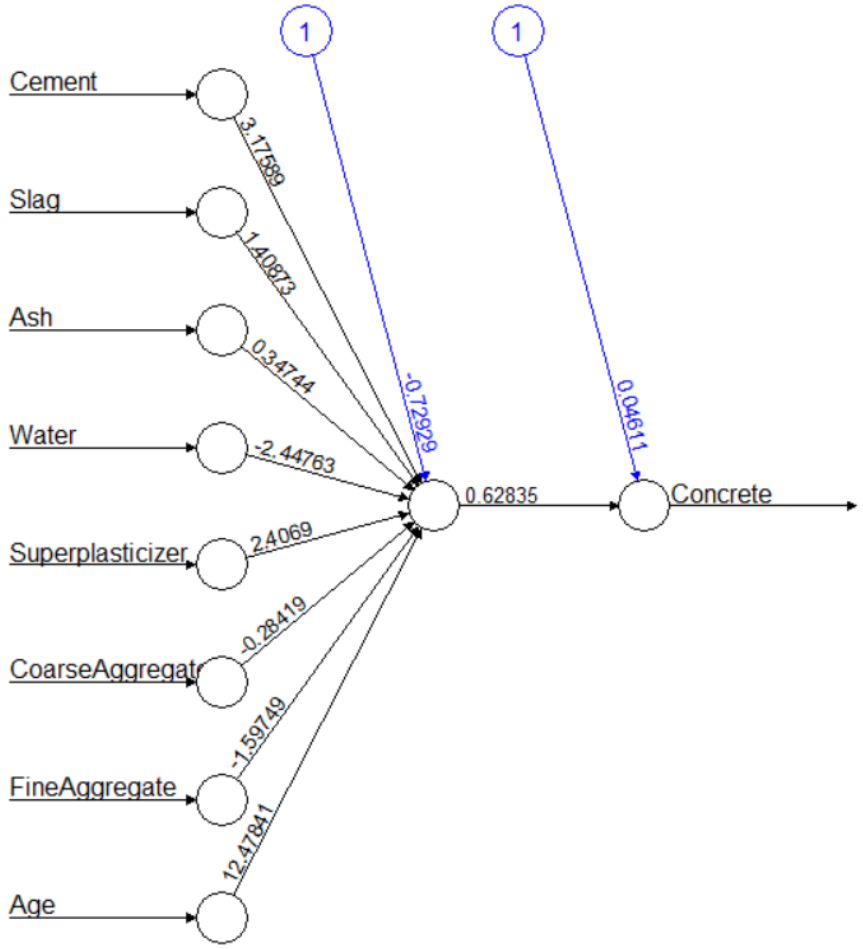


Τα δεδομένα όπως προκύπτει από τα περιγραφικά στα στατιστικά δεν είναι ίδια κλίμακας, με αποτέλεσμα να απαιτείται κανονικοποίηση, με στόχο τη βελτίωση των αποτελεσμάτων του νευρωνικού δικτύου. Η κανονικοποίηση των μεταβλητών γίνεται στο διάστημα [0,1].

Για την εκπαίδευση του νευρωνικού, το σύνολο δεδομένων χωρίζεται σε train και test set με ποσοστά 75% και 25% αντίστοιχα.

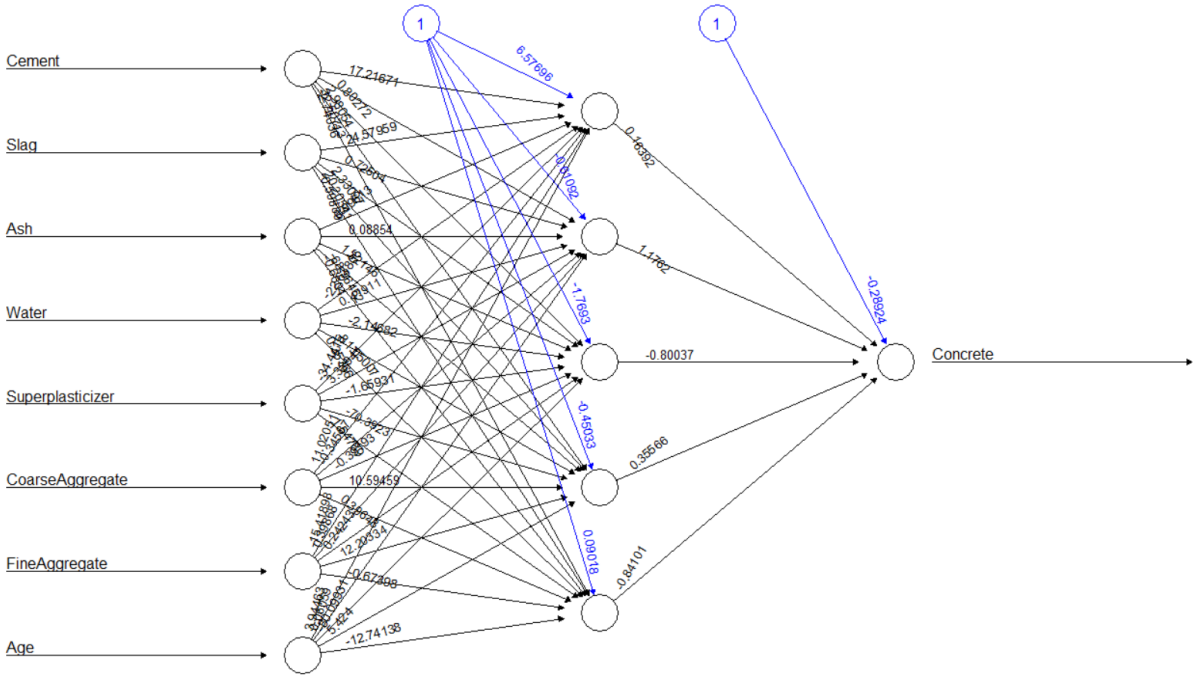
Για την εκπαίδευση του μοντέλου θα χρησιμοποιηθεί το πακέτο neuralnet. Η αρχιτεκτονική που υλοποιήθηκε αρχικά περιλαμβάνει το επίπεδο εισόδου, το επίπεδο εξόδου και ένα κρυφό επίπεδο με ένα νευρώνα. Το μοντέλο έχει MSE ίσο με 5.67. Για την αξιολόγηση της απόδοσης του μοντέλου πραγματοποιείται πρόβλεψη της ανθεκτικότητας του σκυροδέματος για test set. Για την πρόβλεψη γίνεται χρήση της εντολής compute, η οποία αποθηκεύει τα αποτελέσματα της πρόβλεψης και τα συναπτικά βάρη των νευρώνων, έτσι όπως αυτά προέκυψαν από την εκπαίδευση.

Εφόσον το πρόβλημα μας περιλαμβάνει συνεχείς μεταβλητές, μπορεί να υπολογιστή η συσχέτιση μεταξύ των προβλεπόμενων τιμών και των πραγματικών. Με αυτό τον τρόπο, γίνεται γνωστή η ισχυρότητα της γραμμικής συσχέτισης μεταξύ τους. Όσο πιο ισχυρή είναι τόσο καλύτερη η πρόβλεψη. Για τη συγκεκριμένη αρχιτεκτονική, προκύπτει συσχέτιση ίση με 0.72.



Error: 5.668196 Steps: 2618

Με στόχο την βελτίωση της απόδοσης του μοντέλου, αυξάνουμε τον πλήθος των νευρώνων του κρυφού επιπέδου από έναν σε πέντε. Το MSE μειώνεται σημαντικά από 5.67 σε 1.61. Η συσχέτιση μεταξύ της εξαρτημένης μεταβλητής του test set της προβλεπόμενης από το βελτιωμένο δίκτυο είναι ίση με 1.

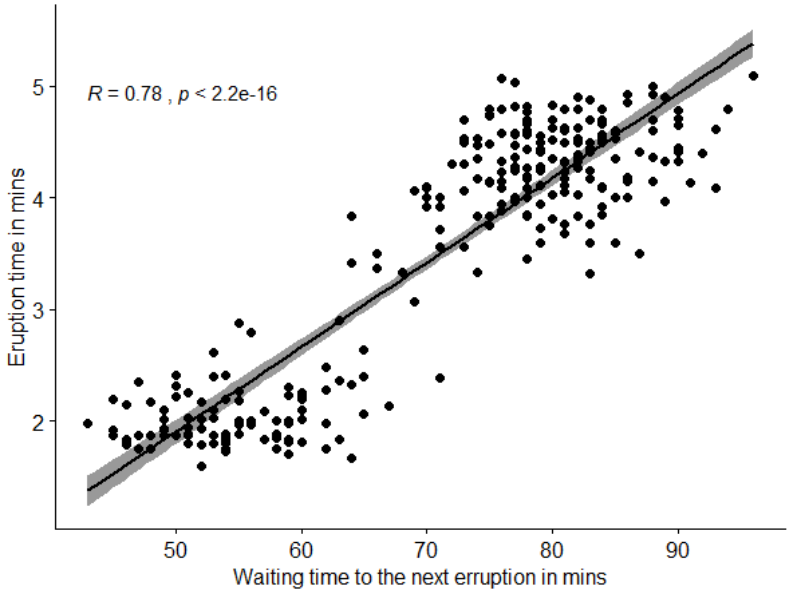


2.2.7. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 7

Το σύνολο δεδομένων faithfull περιλαμβάνει δυο μεταβλητές, η μεταβλητή eruptions αντιστοιχεί στη διάρκεια μιας έκρηξης και η μεταβλητή waiting στο χρονικό διάστημα μεταξύ δυο διαδοχικών εκρήξεων.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
eruptions	1	272	3.49	1.14	4	3.53	0.95	1.6	5.1	3.5	-0.41	-1.51	0.07
waiting	2	272	70.90	13.59	76	71.50	11.86	43.0	96.0	53.0	-0.41	-1.16	0.82

Ο συντελεστής συσχέτισης μεταξύ των δυο μεταβλητών είναι ίσος με 0.78 και από το Correlation test προκύπτει ότι δεν μπορούμε να αποδεχθούμε την μηδενική υπόθεση ότι ο συντελεστής συσχέτισης είναι μη σημαντικά διαφορετικός του μηδενός.

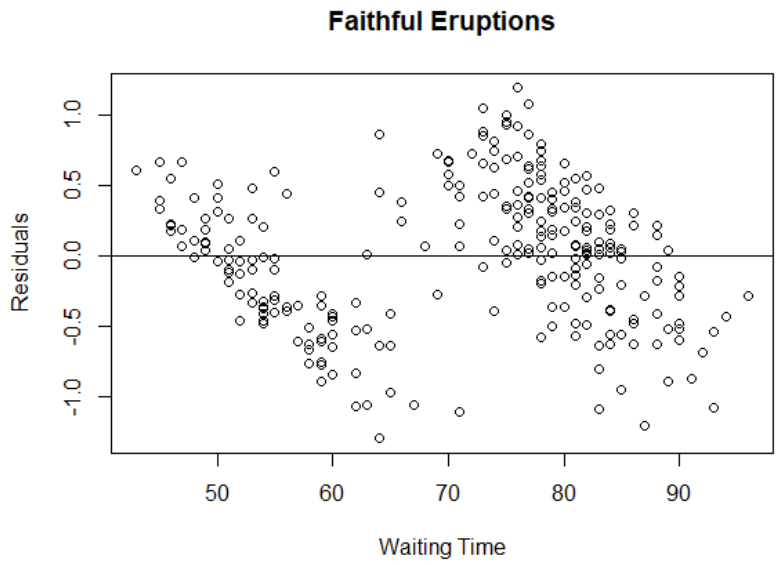


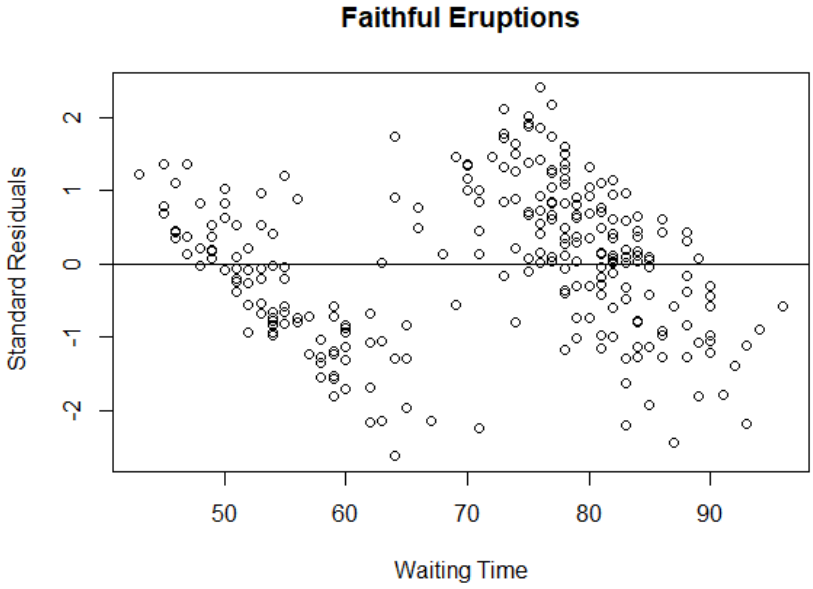
Εφαρμόζοντας γραμμική παλινδρόμηση μεταξύ των μεταβλητών προκύπτει ότι η σχέση μεταξύ των παραμέτρων στη γραμμική παλινδρόμηση είναι στατιστικά σημαντική, καθώς το p-value είναι μικρότερο του συντελεστή εμπιστοσύνης 0.05 και απορρίπτεται η μηδενική υπόθεση ότι ο συντελεστής β=0.

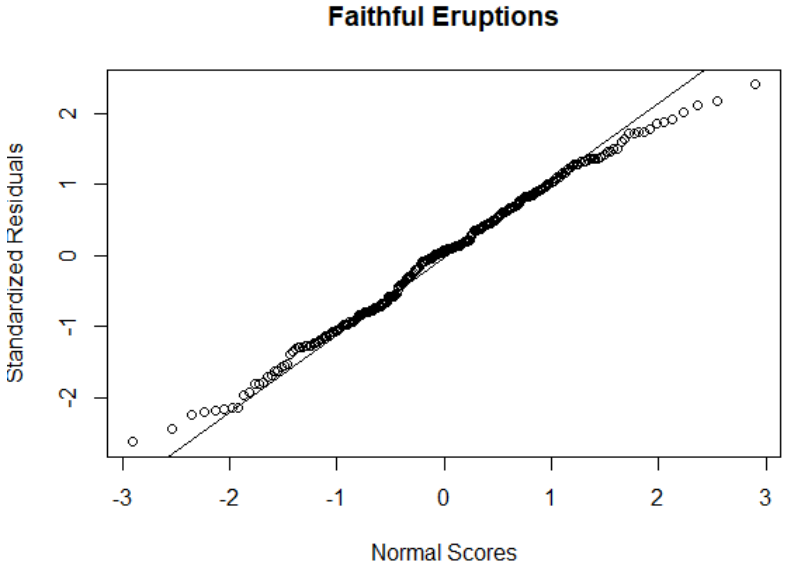
Ο συντελεστής προσδιορισμού είναι ίσος με 0.81, γεγονός που σημαίνει ότι το 81% της διακύμανσης των δεδομένων εξηγείται από το μοντέλο.

```
Call:
lm(formula = eruptions ~ waiting, data = faithfull)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
waiting      0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

Για την μελέτη της κατανομής των υπολοίπων παρουσιάζονται τα διαγράμματα των υπολοίπων και τυποποιημένων υπολοίπων. Από τα scatterplot, το Q-Q Plot και το Shapiro-Wilk normality test, μπορούμε να υποθέσουμε ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή.


Faithful Eruptions
Residuals
1.0
0.5
0.0
-0.5
-1.0
50
60
70
80
90
Waiting Time


Faithful Eruptions
Standard Residuals
2
1
0
-1
-2
50
60
70
80
90
Waiting Time


Faithful Eruptions
Standardized Residuals
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
Normal Scores

Η πρόβλεψη της μεταβλητής eruption για τιμή waiting ίση με 80 λεπτά προκύπτει ως εξής:

$$duration = -1{,}87 + 0.08 * (waiting = 80) = 4.18 \text{ min}$$

Δεδομένου ότι τα υπόλοιπα είναι ανεξάρτητα της μεταβλητής x, για δοσμένη τιμή του x, μπορεί να εκτιμηθεί το διάστημα της μέσης τιμή του εκτιμώμενου y, που ονομάζεται confidence interval. Το 95% διάστημα εμπιστοσύνης για τη μεταβλητή eruption duration για χρόνο αναμονής 80 min είναι μεταξύ των 4.10 και 4.25 min.

```
predict(model, newdata, interval="confidence")
    fit      lwr      upr
4.17622 4.104848 4.247592
```

Για δοσμένη τιμή x, το εκτιμώμενο διάστημα της εξαρτημένης μεταβλητής y ονομάζεται prediction interval. Το 95% διάστημα πρόβλεψης για τη μεταβλητή eruption duration για χρόνο αναμονής 80 min είναι μεταξύ των 3.20 και 5.16 min.

```
predict(model, newdata, interval="predict")
    fit      lwr      upr
4.17622 3.196089 5.156351
```

2.2.8. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 8

Το σύνολο δεδομένων stack loss αποτελείται από 21 παρατηρήσεις και 4 μεταβλητές, τα περιγραφικά στατιστικά των οποίων παρουσιάζονται στην εικόνα που ακολουθεί.

```
> describe(stackloss)
           vars  n  mean    sd median trimmed  mad min max range
Air.Flow      1 21 60.43  9.17     58   59.35 5.93  50  80    30
Water.Temp    2 21 21.10  3.16     20   20.82 2.97  17  27    10
Acid.Conc.    3 21 86.29  5.36     87   86.76 4.45  72  93    21
stack.loss    4 21 17.52 10.17     15   16.12 5.93   7  42    35
```

Εφαρμόζοντας γραμμική παλινδρόμηση μεταξύ των μεταβλητών προκύπτει ότι η σχέση μεταξύ της μεταβλητής Stack.loss και των ανεξάρτητων μεταβλητών Air.Flow και Water.Temp είναι στατιστικά σημαντική, καθώς το p-value είναι μικρότερο του συντελεστή εμπιστοσύνης 0.05. Η σχέση ωστόσο μεταξύ της μεταβλητής Acid.Conc και της Stack.loss δεν είναι στατιστικά σημαντική και δεν μπορούμε να υποθέσουμε ότι ο συντελεστής β_i είναι διάφορος του μηδενός για διάστημα εμπιστοσύνης 0.05.

Ο συντελεστής προσδιορισμού είναι ίσος με 0.91, γεγονός που σημαίνει ότι το 91% της διακύμανσης των δεδομένων εξηγείται από το μοντέλο.

```
> summary(stackloss.lm)

Call:
lm(formula = stack.loss ~ ., data = stackloss)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2377 -1.7117 -0.4551  2.3614  5.6978

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -39.9197    11.8960  -3.356  0.00375 **
Air.Flow      0.7156     0.1349   5.307  5.8e-05 ***
Water.Temp    1.2953     0.3680   3.520  0.00263 **
Acid.Conc.   -0.1521     0.1563  -0.973  0.34405
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.243 on 17 degrees of freedom
Multiple R-squared:  0.9136,    Adjusted R-squared:  0.8983
F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

Η πρόβλεψη της μεταβλητής Stack.loss για τιμή η τιμή της airflow είναι 72, της water temperature είναι 20 και της air concentration είναι 85 προκύπτει ως εξής:

Stack.loss = -39.92 + 0.71 (Air.Flow =72) + 1.30 (Water.Temp=20) - 0,15 (Acid.Conc) = 24.58

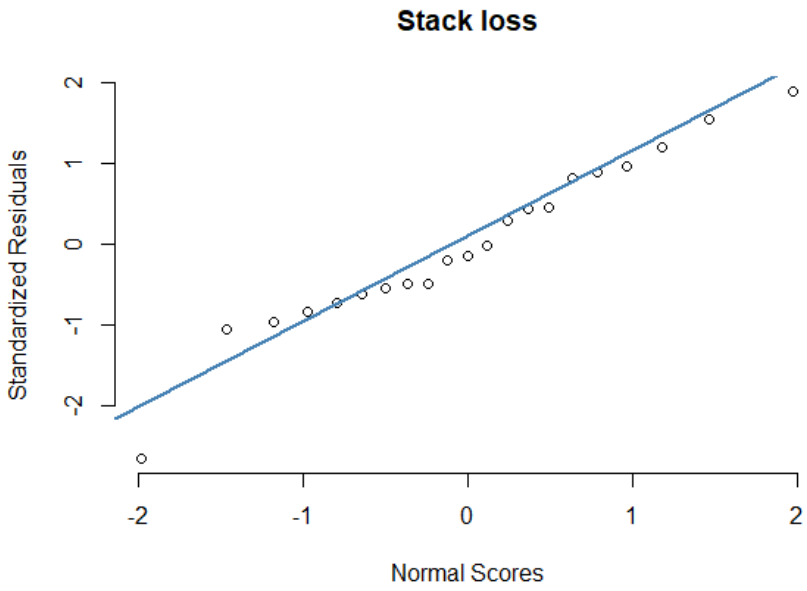
Δεδομένου ότι τα υπόλοιπα είναι ανεξάρτητα της μεταβλητής x, για δοσμένη τιμή του x, μπορεί να εκτιμηθεί το διάστημα της μέσης τιμή του εκτιμόμενου y, που ονομάζεται confidence interval. Το 95% διάστημα εμπιστοσύνης της μέσης τιμής της μεταβλητή Stackloss είναι μεταξύ των 20.22 και 28.95.

```
> predict(stackloss.lm, newdata, interval="confidence")
       fit      lwr    upr
1 24.58173 20.21846 28.945
```

Για δοσμένη τιμή x, το εκτιμώμενο διάστημα της εξαρτημένης μεταβλητής y ονομάζεται prediction interval. Το 95% διάστημα πρόβλεψης της μεταβλητής Stackloss είναι μεταξύ των 16.47 και 32.70.

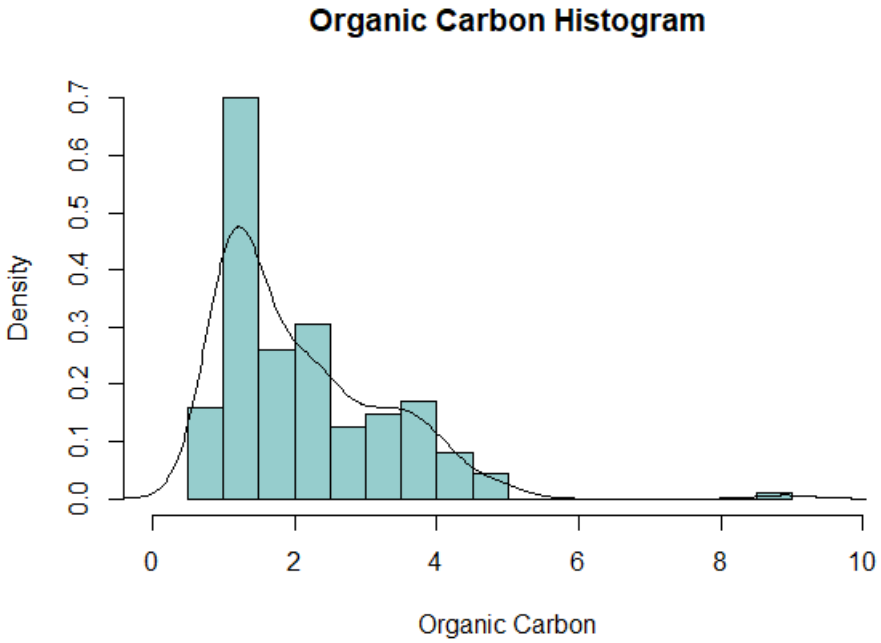
```
> predict(stackloss.lm, newdata, interval="predict")
       fit     lwr      upr
1 24.58173 16.4661 32.69736
```

Για την μελέτη της κατανομής των τυποποιημένων υπολοίπων παρουσιάζεται το διάγραμμα Q-Q Plot σύμφωνα με το οποίο μπορούμε να υποθέσουμε ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή.

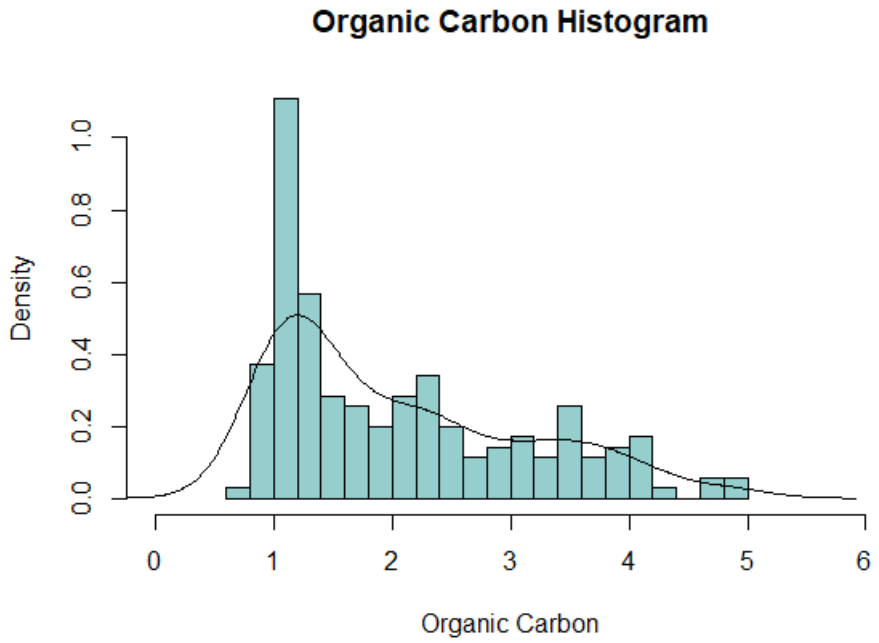


2.2.9. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 9

Το σύνολο δεδομένων Spectrum_Breizh.txt περιέχει μεταβλητές από φασματοσκοπικές μετρήσεις με σκοπό την πρόβλεψη της περιεκτικότητας του εδάφους σε άνθρακα. Μελετώντας την κατανομή της εξαρτημένης μεταβλητής παρατηρούνται κάποια outliers.



Με την αφαίρεση των ακραίων τιμών, παρατηρείται ότι η μεταβλητή OC δεν ακολουθεί την κανονική κατανομή.



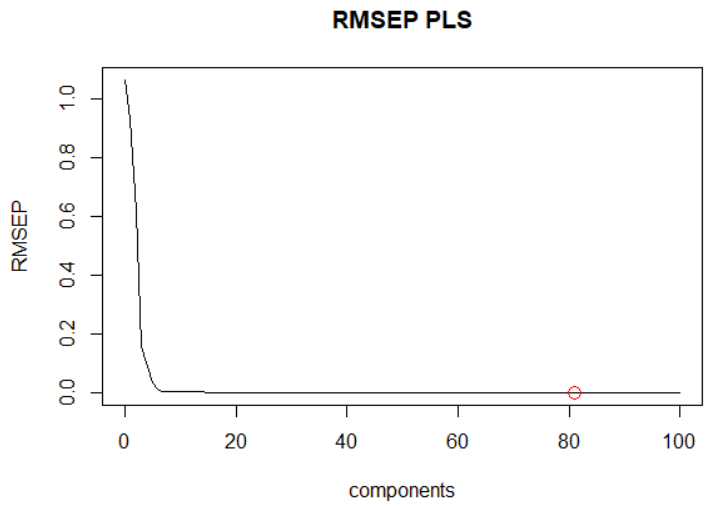
Πρώτο βήμα της μοντελοποίησης αποτελεί η κανονικοποίηση των ανεξάρτητων μεταβλητών στο διάστημα [0,1]

Στη συνέχεια, πρόκειται να εφαρμόσουμε τη Μέθοδο των Μερικών Ελαχίστων Τετραγώνων. Η παραδοσιακή παλινδρόμηση κρίνεται ανεπαρκής εξαιτίας των πολλών συνιστωσών του συνόλου δεδομένων, του μικρού αριθμού παρατηρήσεων και την ύπαρξη αυτό-συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών.

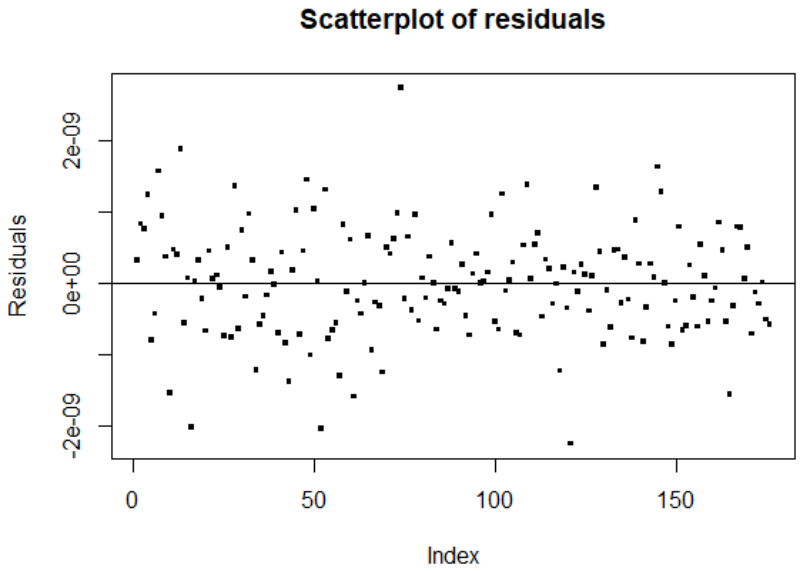
Η μέθοδος των μερικών ελαχίστων τετραγώνων βασίζεται σε μεθόδους μείωσης των διαστάσεων των δεδομένων, όπως η Principal Component Regression (PCR). Η PCR βασίζεται στη μέθοδο PCA, η οποία προβάλλει ένα σύνολο από διανύσματα υψηλής διάστασης σε ένα χώρο χαμηλότερης διάστασης, ενώ οι μετρικές ανάμεσα τους διατηρούνται. Τα νέα διανύσματα που προκύπτουν είναι ασυσχέτιστα μεταξύ τους και διατηρούν σε ένα μεγάλο βαθμό τη διακύμανση των αρχικών δεδομένων. Αυτός ο μετασχηματισμός ορίζεται με τέτοιο τρόπο, ώστε η πρώτη συνιστώσα να έχει τη μεγαλύτερη δυνατή διακύμανση και κάθε επόμενη συνιστώσα έχει με τη σειρά της την υψηλότερη δυνατή διακύμανση υπό το περιορισμό ότι είναι κάθετη ως προς τις υπόλοιπες συνιστώσες. Οι κύριες συνιστώσες που προκύπτουν χρησιμοποιούνται για τη κατασκευή του μοντέλου της γραμμικής παλινδρόμησης. Ένα μειονέκτημα της μεθόδου PCR είναι ότι δεν υπάρχει βεβαιότητα ότι οι συνιστώσες που θα επιλεχθούν για την κατασκευή του μοντέλου συσχετίζονται με την εξαρτημένη μεταβλητή του μοντέλου, την οποία θέλουμε και να προβλέψουμε.

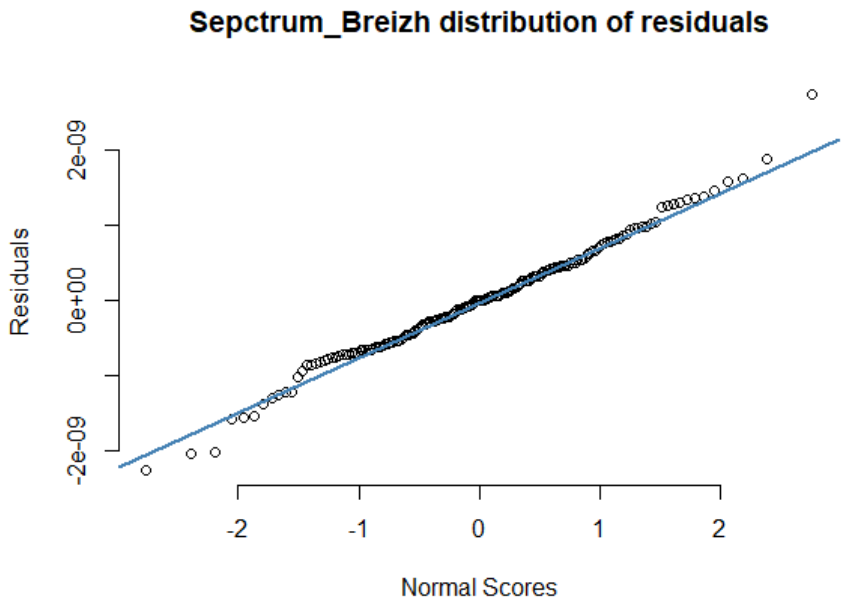
Η Partial Least Squares (PLS) regression, αντίθετα, οδηγεί στη δημιουργία νέων συνιστωσών από τα αρχικά δεδομένα προβάλλοντας τα σε έναν άλλο χώρο ενώ παράλληλα εγγυάται τη συσχέτισης τους με την εξαρτημένη μεταβλητή. Οι νέες αυτές μεταβλητές χρησιμοποιούνται για τη πρόβλεψη του μοντέλου.

Αρχικά, εξετάζουμε το βέλτιστο πλήθος από συνιστώσες που πρέπει να συμμετάσχουν στο μοντέλο. Από το διάγραμμα προκύπτει ότι η ελαχιστοποίηση του σφάλματος πρόβλεψης προκύπτει για πλήθος συνιστωσών ίσο με 81.



Για την ανάλυση των υπολοίπων εκπαιδεύεται το μοντέλο για πλήθος συνιστωσών ίσο με 81. Από το διάγραμμα διασποράς των υπολοίπων δεν προκύπτει κάποιο pattern και σε συνδυασμό με το Q-Q Plot μπορούμε να υποθέσουμε την κανονική κατανομή των υπολοίπων.



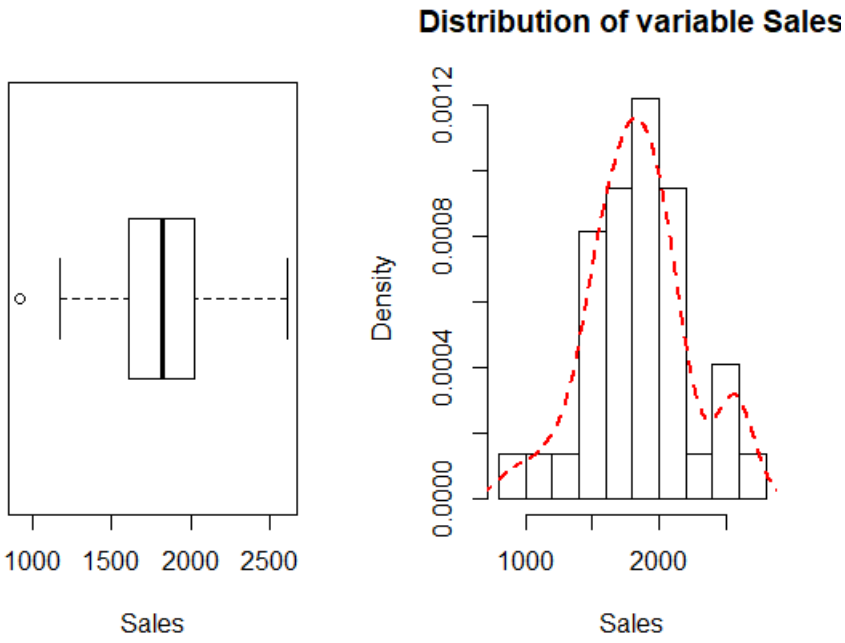


2.2.10. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 10

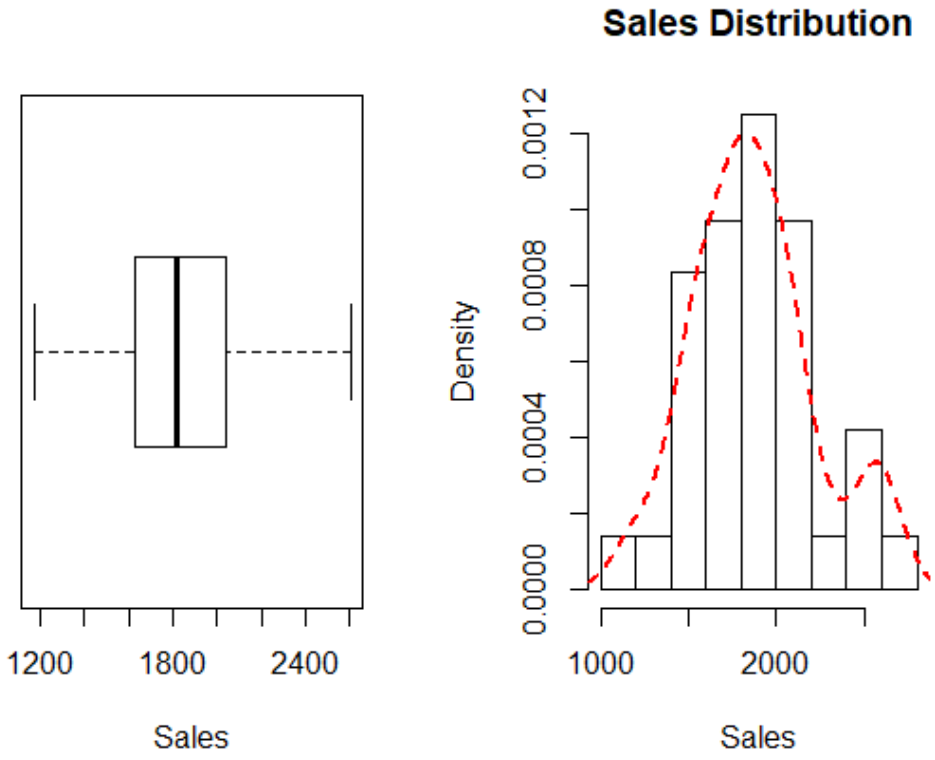
Το σύνολο δεδομένων market περιλαμβάνεις μεταβλητές για τις πωλήσεις, την τιμή, το κόστος ενός προϊόντος και την αφίξεις σε 37 διαφορετικά καταστήματα. Στόχος είναι η εύρεση της σχέσης των πωλήσεων με της μεταβλητές.

```
> summary(market)
     Sales           Preis           Costs         Arrivals
 Min.   : 921   Min.   : 7.00   Min.   :   0   Min.   : 60.00
 1st Qu.:1612   1st Qu.: 9.00   1st Qu.: 800   1st Qu.: 79.00
 Median :1819   Median :10.00   Median :1300   Median : 89.00
 Mean   :1847   Mean   :10.43   Mean   :1240   Mean   : 90.27
 3rd Qu.:2026   3rd Qu.:12.00   3rd Qu.:1600   3rd Qu.:103.00
 Max.   :2604   Max.   :13.00   Max.   :2000   Max.   :125.00
```

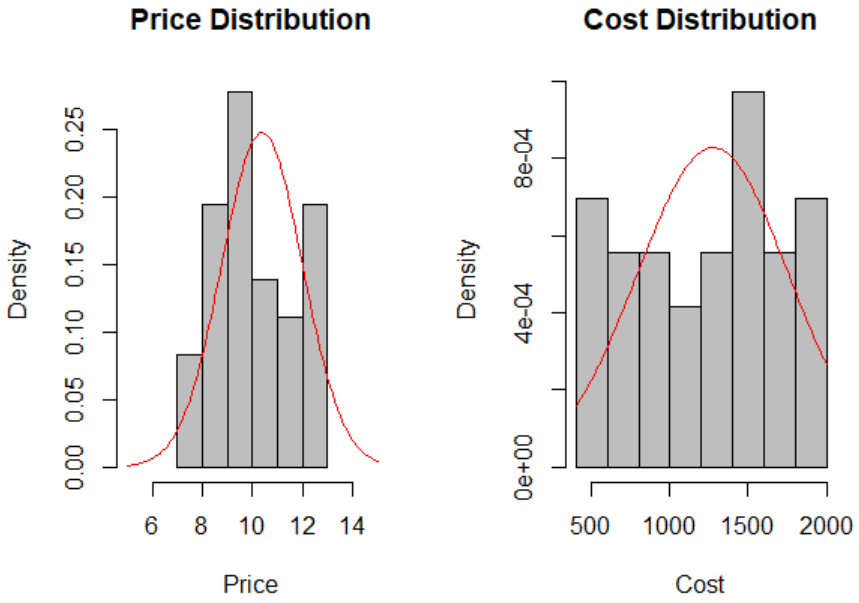
Μελετώντας τη μεταβλητή Sales, αναδεικνύεται η ύπαρξη ακραίας τιμής.

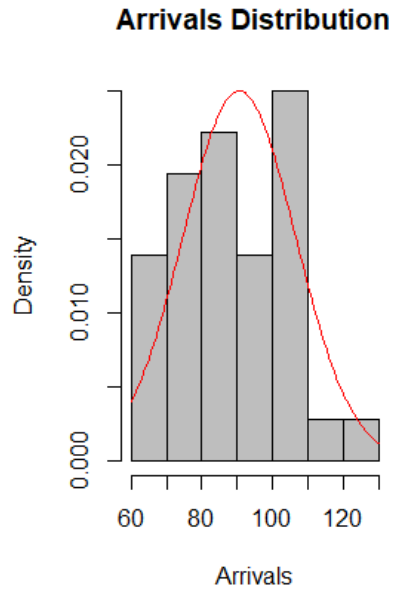


Αφαιρώντας την, παρατηρείται ότι η μεταβλητή ακολουθεί την κανονική κατανομή με το μεγαλύτερο πλήθος των παραμέτρων να παρατηρείται γύρω από τη μέση τιμή και διάμεσο.

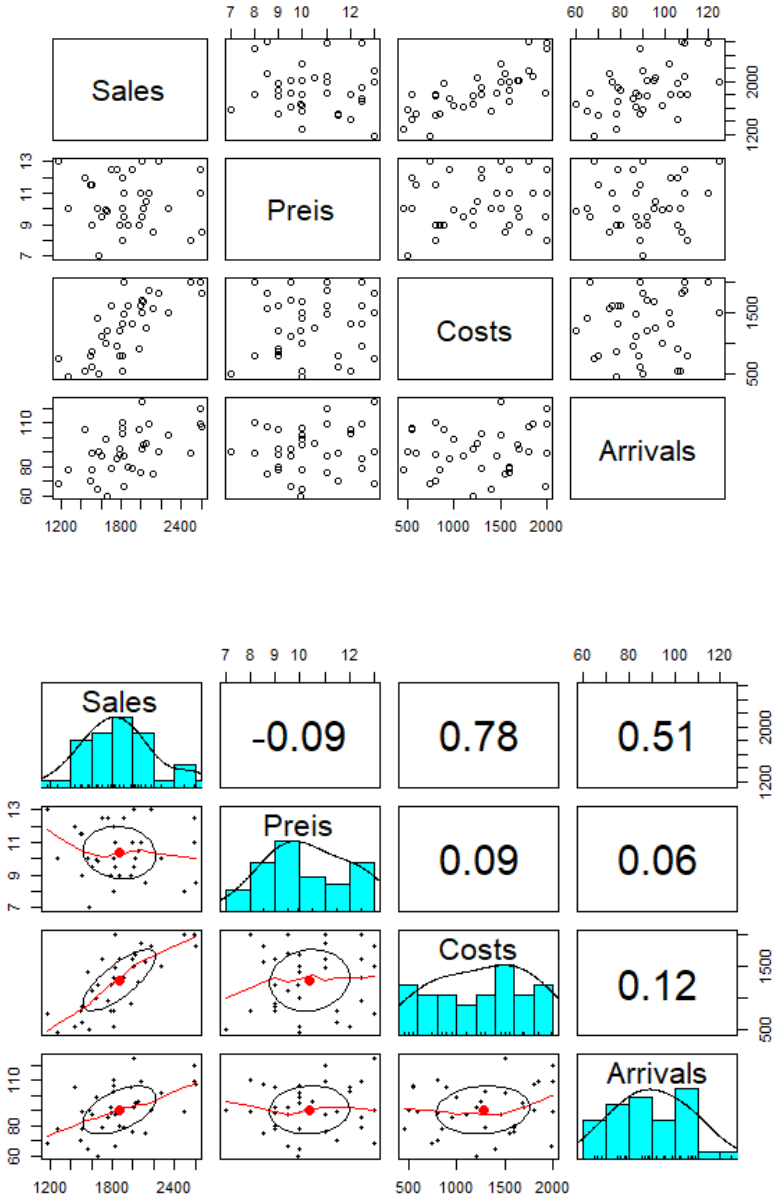


Η μελέτη της κατανομής πραγματοποιείται και για τις υπόλοιπες μεταβλητές, οι οποίες σύμφωνα με το Shapiro-Wilk normality test φαίνεται να ακολουθούν ην κανονική κατανομή.





Εξαιτίας της διαφορετικής κλίμακας των δεδομένων επιλέχθηκε η κανονικοποίηση των ανεξάρτητων μεταβλητών στο διάστημα [0,1]. Η συσχέτιση μεταξύ της μεταβλητής Sales με την Arrivals και Costs παρατηρείται υψηλή, ίση με 0.51 και 0.78 αντίστοιχα. Παράλληλα, η συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών παίρνει χαμηλές τιμές.



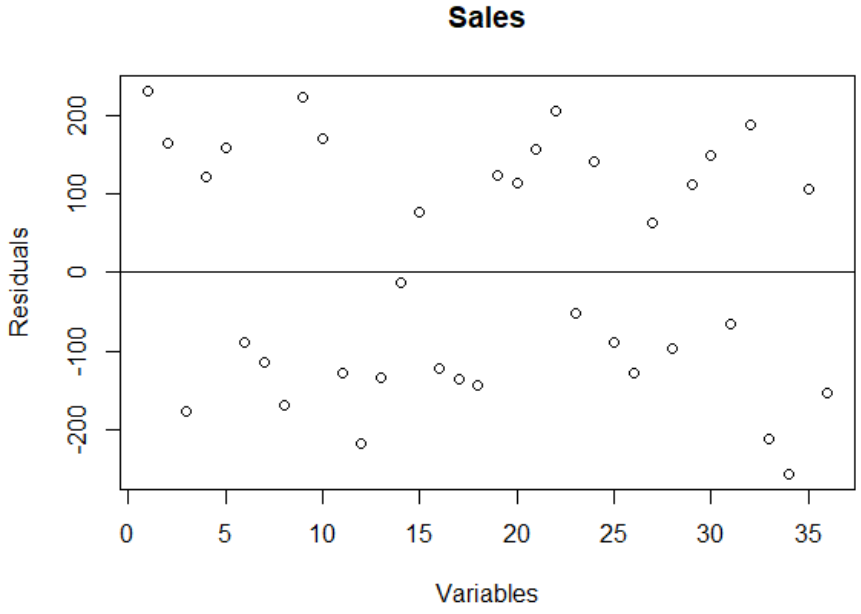
Το report του γραμμικού μοντέλου παρουσιάζεται στην εικόνα που ακολουθεί και προκύπτει ότι η σχέση μεταξύ των μεταβλητών είναι στατιστικά σημαντική, καθώς το p-value προκύπτει μικρότερο του 0.05 και μπορούμε να απορρίψουμε τη μηδενική υπόθεση ότι οι συντελεστές της γραμμικής παλινδρόμησης είναι ίσοι με το μηδέν. Από το F-statistic test, προκύπτει ότι το σύνολο του μοντέλου είναι στατιστικά σημαντικό.

Ο συντελεστής προσδιορισμού του μοντέλου είναι ικανοποιητικά υψηλός με τιμή ίση με 0.82. Η σχέση πρόβλεψης που προκύπτει είναι η εξής:

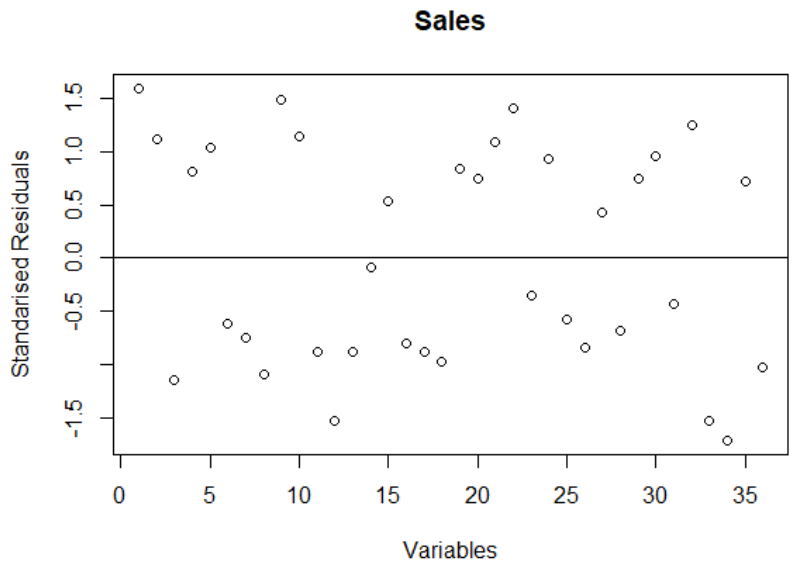
Sales = 734.84 - 40.17 Price + 0.54 Costs +9.60 Arrivals

```
Call:
lm(formula = Sales ~ ., data = market)

Residuals:
    Min      1Q  Median      3Q     Max
-255.67 -129.26  -32.28  142.75  229.84

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 734.8450   227.3280   3.233  0.00284 **
Preis       -40.1698    16.6341  -2.415  0.02163 *
Costs         0.5383     0.0559   9.629 5.68e-11 ***
Arrivals      9.5953     1.6843   5.697 2.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 157.6 on 32 degrees of freedom
Multiple R-squared:  0.815,     Adjusted R-squared:  0.7977
F-statistic:    47 on 3 and 32 DF,  p-value: 7.876e-12
```

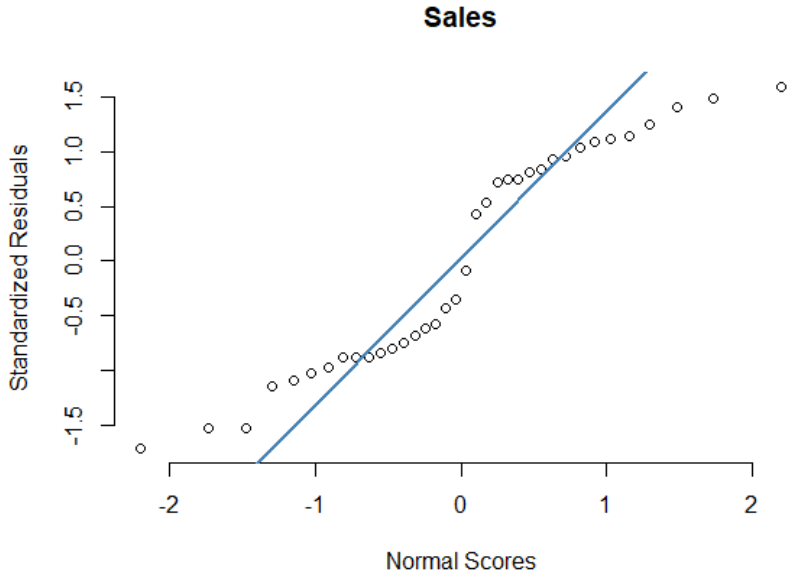
Από τα scatterplot των υπολοίπων, παρατηρείται ένα pattern γεγονός που αποτελεί ένδειξη για την μη κανονικότητα των υπολοίπων. Η ένδειξη αυτή επιβεβαιώνεται από το Shapiro-Wilk normality test, καθώς η μηδενική υπόθεση για κανονικότητα των δεδομένων δεν γίνεται αποδεκτή για p-value 0.005 μικρότερο από το 0.05.




```
> describe(stdres)
   vars  n mean   sd median trimmed  mad   min  max range skew kurtosis   se
X1    1 36    0 1.02  -0.22    0.01 1.38 -1.71 1.59   3.3    0    -1.57 0.17
```




```
> describe(res)
   vars  n mean     sd median trimmed    mad     min    max  range skew kurtosis    se
X1    1 36    0 150.67 -32.28    0.96 208.71 -255.67 229.84 485.51 0.01     -1.6 25.11
```




```
> shapiro.test(stdres)

        Shapiro-Wilk normality test

data:  stdres
W = 0.90731, p-value = 0.005441
```

Ομοσκεδαστικότητα:

Για την ανάλυση της γραμμικής παλινδρόμησης μια από της υποθέσεις που κάνουμε είναι ότι το σφάλμα της παλινδρόμησης έχει μέση τιμή μηδέν για κάθε τιμή της ανεξάρτητης μεταβλητής X και η διασπορά του είναι σταθερή και δεν εξαρτάται από τη μεταβλητή X. Η ιδιότητα αυτή ονομάζεται ομοσκεδαστικότητα. Στην αντίθετη

περίπτωση έχουμε ετεροσκεδαστικότητα, όταν δηλαδή η διασπορά της Υ μεταβάλλεται με την τιμής της Χ.

Γενικά για να εκτιμήσουμε τις παραμέτρους της γραμμικής παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων, δεν είναι απαραίτητο να υποθέσουμε κάποια συγκεκριμένη δεσμευμένη κατανομή της Υ ως προς τη Χ. Αν θέλουμε όμως να υπολογίσουμε παραμετρικά διαστήματα εμπιστοσύνης για τις παραμέτρους ή να κάνουμε παραμετρικούς στατιστικούς ελέγχους χρειάζεται να υποθέσουμε κανονική δεσμευμένη κατανομή για τη Υ. Επίσης οι υποθέσεις για γραμμική σχέση και σταθερή διασπορά αποτελούν χαρακτηριστικά πληθυσμών με κανονική κατανομή. Συνήθως λοιπόν σε προβλήματα γραμμικής παλινδρόμησης υποθέτουμε ότι η δεσμευμένη κατανομή της Υ είναι κανονική.

Αναφορικά με την ομοσκεδαστικότητα στο συγκεκριμένο παράδειγμα, από την ανάλυσης των υπολοίπων καταλήγουμε στο συμπέρασμα ότι η διασπορά της Υ μεταβάλλεται με την τιμής της Χ, καθώς τα υπόλοιπα παρουσιάζουν κάποιο πρότυπο και συνεπώς υφίσταται ετεροσκεδαστικότητα.

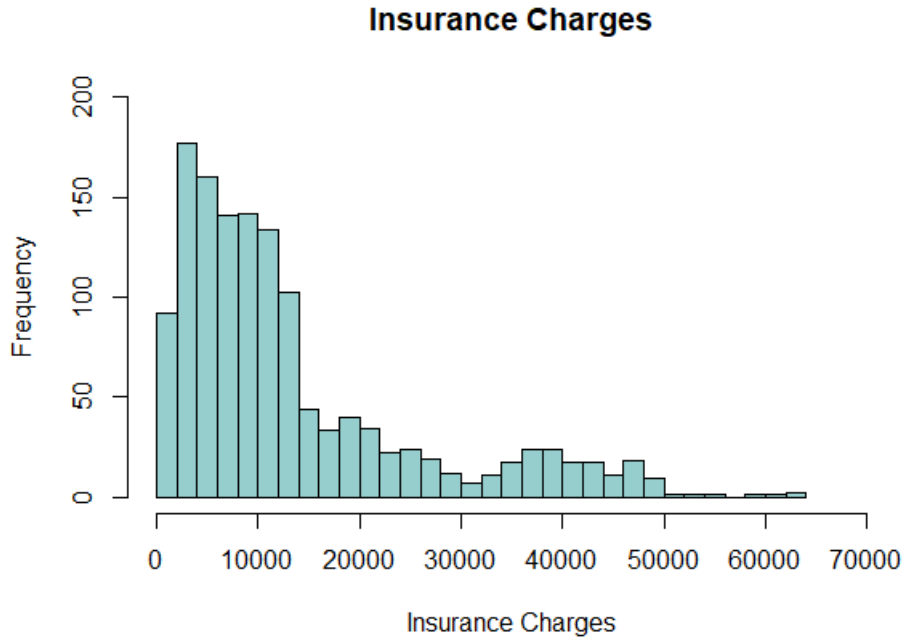
2.2.11. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 11

Ερώτημα 1 και 2:

Το σύνολο δεδομένων insurance.csv περιέχει 1338 παρατηρήσεις και 7 στήλες. Οι τέσσερεις εκ των όποιων είναι αριθμητικές ((age, bmi, children and expenses) και τρεις κατηγορικές (sex, smoker and region). Τα περιγραφικά στατιστικά των δεδομένων παρουσιάζονται στην εικόνα που ακολουθεί.

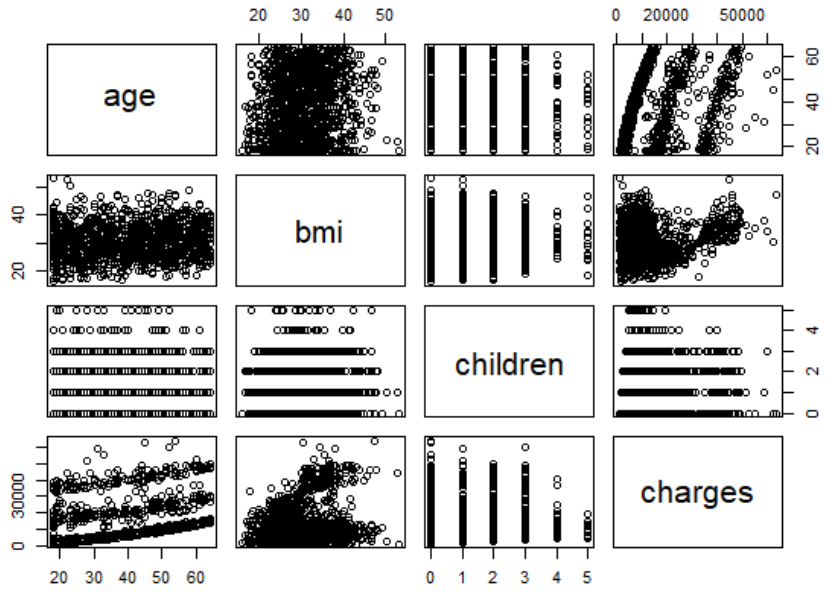
```
> #Descriptive Statistics
> summary(insurance)
      age            sex           bmi           children     smoker          region       charges
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064   northeast:324   Min.   : 1122
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325   1st Qu.: 4740
 Median :39.00                Median :30.40   Median :1.000              southeast:364   Median : 9382
 Mean   :39.21                Mean   :30.66   Mean   :1.095              southwest:325   Mean   :13270
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000                              3rd Qu.:16640
 Max.   :64.00                Max.   :53.13   Max.   :5.000                              Max.   :63770
> describe(insurance)
         vars    n     mean       sd  median  trimmed     mad     min      max    range  skew kurtosis     se
age         1 1338    39.21    14.05   39.00    39.01   17.79   18.00    64.00    46.00  0.06    -1.25   0.38
sex*        2 1338     1.51     0.50    2.00     1.51    0.00    1.00     2.00     1.00 -0.02    -2.00   0.01
bmi         3 1338    30.66     6.10   30.40    30.50    6.20   15.96    53.13    37.17  0.28    -0.06   0.17
children    4 1338     1.09     1.21    1.00     0.94    1.48    0.00     5.00     5.00  0.94     0.19   0.03
smoker*     5 1338     1.20     0.40    1.00     1.13    0.00    1.00     2.00     1.00  1.46     0.14   0.01
region*     6 1338     2.52     1.10    3.00     2.52    1.48    1.00     4.00     3.00 -0.04    -1.33   0.03
charges     7 1338 13270.42 12110.01 9382.03 11076.02 7440.81 1121.87 63770.43 62648.55  1.51     1.59 331.07
```

Από το ιστόγραμμα της μεταβλητής των εξόδων, παρατηρείται θετική ασυμμετρία στα δεδομένα.

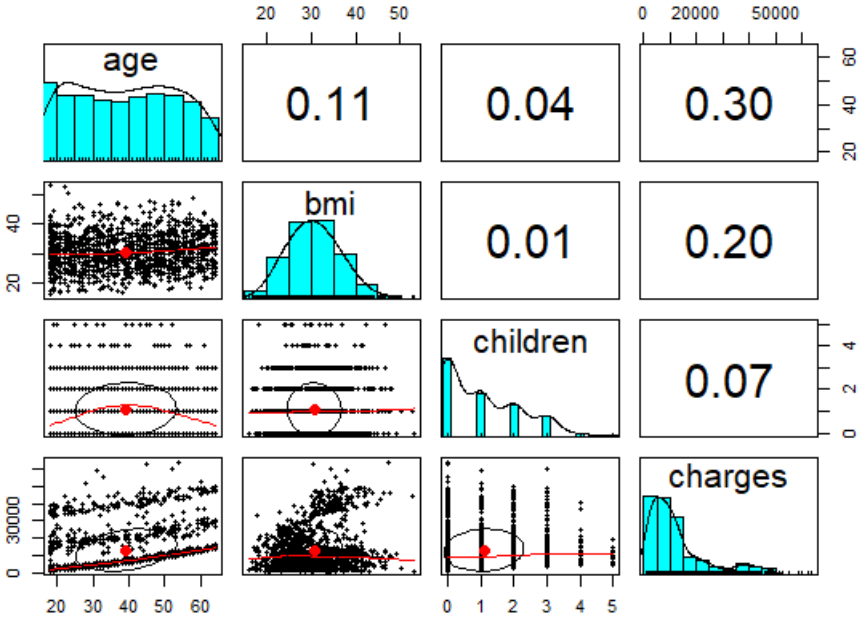


Για τη μελέτη της σχέσης μεταξύ των αριθμητικών μεταβλητών είναι η χρήση του πίνακα συσχέτισης. Η συσχέτιση παίρνει τιμές που κυμαίνονται μεταξύ -1 και 1. Όταν ο δείκτης έχει αρνητικό πρόσημο υποδηλώνει αντίστροφη συσχέτιση και το θετικό θετική συσχέτιση.

Το υψηλότερο ζεύγος συσχέτισης παρατηρείται για τη μεταβλητή της ηλικίας και τα έξοδα με τιμή συντελεστή ίση με 0.3 (θετική συσχέτιση), γεγονός που υποδηλώνει ότι ένα ηλικιωμένο άτομο ξοδεύει περισσότερα στην περίθαλψη.



Στο διάγραμμα που ακολουθεί παρουσιάζονται τρία είδη πληροφορίας. Στη κάτω διαγώνιο του διαγράμματος φαίνονται τα scatterplot με τη πρόσθετη πληροφορία του μέσου των δυο μεταβλητών και την έλλειψη της συσχέτισης που δείχνει την ισχυρότατα αυτής. Στη διαγώνιο του πίνακα, παρουσιάζονται τα ιστογράμματα των του κάθε χαρακτηριστικού και η καμπύλη της κατανομής. Δεξιά της διαγώνιου παρουσιάζονται οι τιμές συσχέτισης μεταξύ των μεταβλητών.



Ερώτημα 3:

Στη συνέχεια , εφαρμόζουμε το μοντέλο της γραμμικής παλινδρόμησης, θέτοντας ως εξαρτημένη μεταβλητή τα charges και ως ανεξάρτητης τις υπόλοιπες μεταβλητές. Από το report του μοντέλου προκύπτει ότι η σχέση μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων είναι στατιστικά σημαντική για τις περισσότερες μεταβλητές με εξαίρεση της κατηγορικές dummies sexmale και regionnorthwest. Ο συντελεστής προσδιορισμού είναι ίσος με 0.75, γεγονός που σημαίνει ότι το μοντέλο εξηγεί το 75% της συνολικής διακύμανσης της μεταβλητής των ιατρικών εξόδων.

```
> summary(ins_model)

Call:
lm(formula = charges ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Η απόδοση του μοντέλου αυξάνεται θεωρώντας ότι η σχέση μεταξύ της μεταβλητής των εξόδων και της ηλικίας δεν είναι γραμμική, αλλά πολυωνιμική δευτέρου βαθμού. Επίσης βασιζόμενοι στη μεταβλητή του δείκτη μάζας σώματος δημιουργούμε μια νέα μεταβλητή την ύπαρξη ή όχι παχυσαρκίας. Παρατηρώντας ότι υπάρχει θετική συσχέτιση μεταξύ των εξόδων και τη μεταβλητή του καπνίσματος και του δείκτη bmi, είναι λογικό να υποθέσουμε ότι ένας παχύσαρκος καπνιστής ξοδεύει περισσότερα χρήματα σε ιατρικές δαπάνες από ότι άτομα με παχυσαρκία που δεν είναι καπνιστές ή άτομα που είναι μόνο καπνιστές.

Σύμφωνα με τα παραπάνω το νέο μοντέλο που προκύπτει έχει μεγαλύτερο συντελεστή προσδιορισμού ίσο με 0.87. επίσης παρατηρείται ότι οι νέοι όροι προστίθενται (age2, bmi30 and bmi30*smokeyes) είναι όλοι στατιστικά σημαντικοί, που σημαίνει ότι έχει νόημα να τους συμπεριλάβουμε στο μοντέλο.

```
> summary(ins_model2)

Call:
lm(formula = charges ~ age + age2 + children + bmi + sex + bmi30 *
    smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-17296.4  -1656.0  -1263.3   -722.1  24160.2

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       134.2509  1362.7511   0.099 0.921539
age               -32.6851    59.8242  -0.546 0.584915
age2                3.7316     0.7463   5.000 6.50e-07 ***
children          678.5612   105.8831   6.409 2.04e-10 ***
bmi               120.0196    34.2660   3.503 0.000476 ***
sexmale          -496.8245   244.3659  -2.033 0.042240 *
bmi30           -1000.1403   422.8402  -2.365 0.018159 *
smokeryes       13404.6866   439.9491  30.469  < 2e-16 ***
regionnorthwest  -279.2038   349.2746  -0.799 0.424212
regionsoutheast  -828.5467   351.6352  -2.356 0.018604 *
regionsouthwest -1222.6437   350.5285  -3.488 0.000503 ***
bmi30:smokeryes 19810.7533   604.6567  32.764  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4445 on 1326 degrees of freedom
Multiple R-squared:  0.8664,    Adjusted R-squared:  0.8653
F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

2.2.12. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 12

Ερώτημα 1:

Το σύνολο δεδομένων mf.xls αποτελείται από 110 παρατηρήσεις και 4 μεταβλητές:

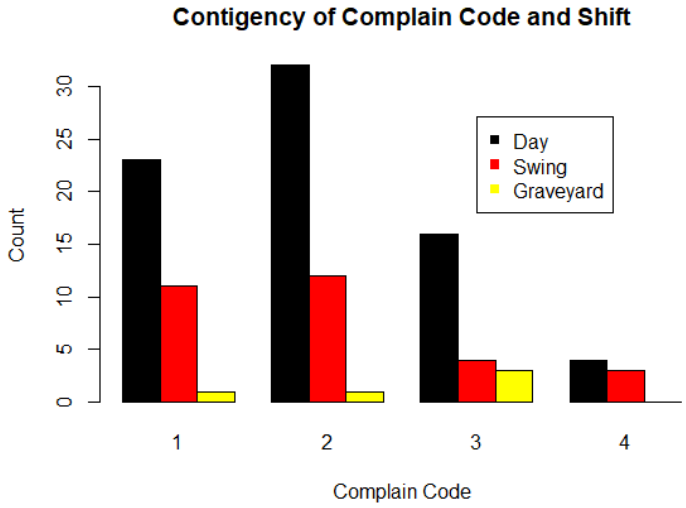
- Της αποζημιώσεις από κάθε επιστροφή
- Τις βάρδιες κατασκευής των εμπορευμάτων
- Τα είδη παραπόνων
- Το μέρος παραγωγής

Κάποια από τα περιγραφικά μέτρα των μεταβλητών παρουσιάζονται στην εικόνα που ακολουθεί.

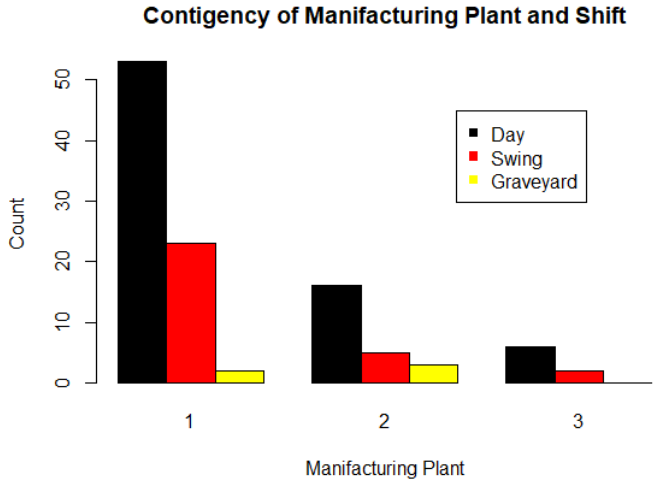
```
> summary(mf)
 Dollar Claim Amount Shift  Complaint Code Manufacturing Plant
 Min.   :14600       1:75   1:35           1:78
 1st Qu.:23650       2:30   2:45           2:24
 Median :27000       3: 5   3:23           3: 8
 Mean   :27245              4: 7
 3rd Qu.:30975
 Max.   :39300
```

Από τον πίνακα συνάφειας μεταξύ της μεταβλητής Complain Code και Shift, παρατηρείται ότι το πλήθος των παραπόνων είναι μεγαλύτερο για τη βάρδια με

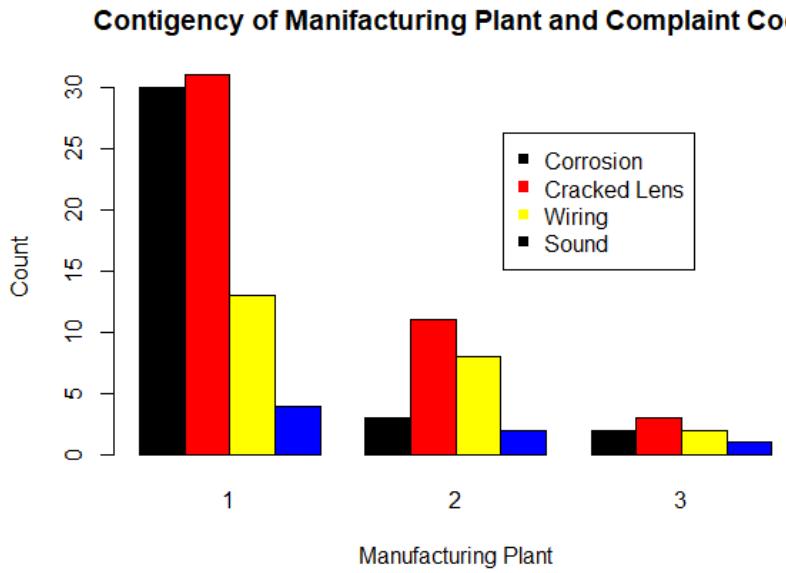
κωδικό ‘Day’, ενώ αμέσως επόμενη με αρκετή διαφορά ακολουθεί η βάρδια ‘Swing’. Στη βάρδια ‘Day’ το παράπονο με ‘Cracked Lens’ αντιστοιχεί στο πιο συχνό παράπονο στο σύνολο των παραπόνων.



Από τον πίνακα συνάφειας μεταξύ της μεταβλητής Manufacturing Plant και Shift, παρατηρείται ότι η περιοχή ‘Boise’ παρουσιάζει το μεγαλύτερο πλήθος παραπόνων κατά τη βάρδια ‘Day’ και ακολουθεί η περιοχή ‘Salt Lake City’.



Τέλος, από τον πίνακα συνάφειας μεταξύ της μεταβλητής Manufacturing Plant και Complain Code, παρατηρείται στην περιοχή ‘Boise’ μεγαλύτερα σε πλήθος είναι τα παράπονα ‘Collision’ και ‘Cracked Lens’, ενώ στις υπόλοιπες περιοχές τα παράπονα ‘Cracked Lens’ και ‘Wiring’.



Ερώτημα 2:

Η ανεξαρτησία των μεταβλητών του Complain Code και Manufacturing Plant ελέγχεται με χρήση του chisq.test. επειδή ο έλεγχος x^2 είναι ευαίσθητος στις μικρές συχνότητες εξασφαλίζουμε ελάχιστο αριθμό παρατηρήσεων ίσο με 3 συμπτύσσοντας τις κλάσεις 'Salt Lake City' και 'Toronto'.

```
> chisq.test(ctbl)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 6.7592, df = 3, p-value = 0.07998
```

Για επίπεδο σημαντικότητας 0.01 και 3 βαθμούς ελευθερίας, προκύπτει από τον πίνακα τη κατανομής x^2 τιμή 11.344 που είναι μεγαλύτερη της τιμής 6.7592 που προέκυψε από το τεστ, με αποτέλεσμα να μην μπορούμε να απορρίψουμε ότι οι δυο μεταβλητές είναι ανεξάρτητες.

Ερώτημα 3:

Για να ελέγξουμε αν υπάρχει στατιστικά σημαντική διαφορά στο ύψος των αποζημιώσεων μεταξύ των τόπων παραγωγής Boise και Salt Lake City για επίπεδο σημαντικότητας 0.02, εφαρμόζουμε t.test στα δυο δείγματα αποζημιώσεων. Με p-value μεγαλύτερη του επιπέδου σημαντικότητας 0.02, δεν μπορεί να απορριφθεί ή μηδενική υπόθεση ότι τα δυο δείγματα ανήκουν στον ίδιο πληθυσμό.

```
> t.test(maninf1$`Dollar Claim Amount`,maninf2$`Dollar Claim Amount`, conf.level = 0.98)

        Welch Two Sample t-test

data:  maninf1$`Dollar Claim Amount` and maninf2$`Dollar Claim Amount`
t = -0.63437, df = 32.887, p-value = 0.5302
alternative hypothesis: true difference in means is not equal to 0
98 percent confidence interval:
 -4319.299  2539.812
sample estimates:
mean of x mean of y
 26843.59  27733.33
```

2.2.13. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 13

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) είναι ένας βασικός αλγόριθμος για συσταδοποίηση με βάση την πυκνότητα. Μπορεί να ανακαλύψει συστάδες διαφορετικών σχημάτων και μεγεθών από μία μεγάλη ποσότητα δεδομένων με θόρυβο και και ακραίες τιμές. Είναι κατάλληλος για δεδομένα με υψηλή πυκνότητα σημείων που είναι διαχωρισμένες από άλλες περιοχές χαμηλής πυκνότητας, όπως θόρυβος. Προϋποθέτει την ύπαρξη ομάδων παρόμοιας συχνότητας χωρίς μεγάλες διακυμάνσεις. Η πυκνότητα για ένα σημείο στον αλγόριθμο ορίζεται ως ο αριθμός σημείων (MinPts) μέσα σε μια προκαθορισμένη ακτίνα (Eps) από αυτό.

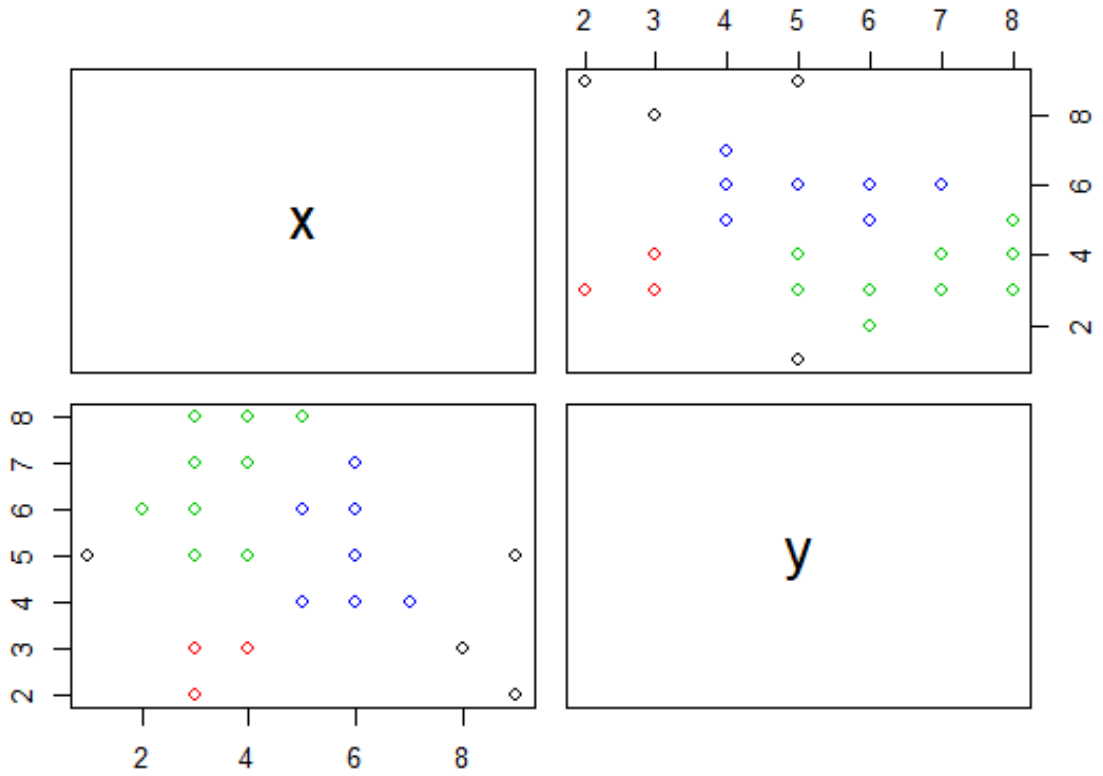
Για την ρύθμιση του αλγορίθμου συνεπώς απαιτείται ο καθορισμός δυο παραμέτρων της μέγιστης ακτίνας της γειτονιάς, Eps και του ελάχιστου αριθμού σημείων στην Eps-γειτονιά ενός σημείου, MinPts.

Τα σημεία κατά τη συσταδοποίηση διαχωρίζονται σε:

- *Βασικά (core) –σημεία πυρήνα*: τα σημεία για τα οποία στην γειτονία τους υπάρχουν περισσότερα σημεία από ένα προκαθορισμένο ελάχιστο αριθμό (MinPts). Τα σημεία αυτά εντοπίζονται στο εσωτερικό μιας συστάδας (ομάδας πυκνών σημείων)
- *Οριακά (border)–σημεία ορίου:* τα σημεία για τα οποία υπάρχουν λιγότερα από ένα προκαθορισμένο αριθμό (MinPts) σημεία σε ακτίνα Eps, αλλά είναι στη γειτονιά (τουλάχιστον) ενός βασικού σημείου(απόσταση μικρότερη ή ίση από Eps)
- *Θορύβου (noise):* τα σημεία που δεν είναι ούτε σημεία πυρήνα ούτε σημεία ορίου

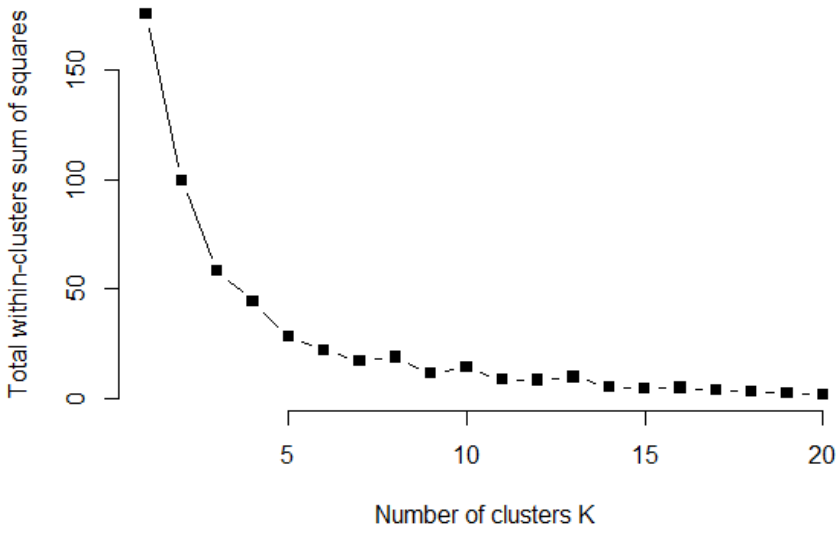
Για τα σημεία που δίνονται στην εκφώνηση, πραγματοποιείται ο αλγόριθμος dbscan για MinPts = 3 και Eps = 1 και οι συστάδες που προκύπτουν παρουσιάζονται στο διάγραμμα που ακολουθεί. Από τον αλγόριθμο προκύπτουν 3 συστάδες και μια που αντιστοιχεί σε θόρυβο.

```
> dbsc <- dbscan(data, eps=1, minPts=3);dbsc
DBSCAN clustering for 23 objects.
Parameters: eps = 1, minPts = 3
The clustering contains 3 cluster(s) and 4 noise points.

0 1 2 3
4 3 9 7

Available fields: cluster, eps, minPts
```

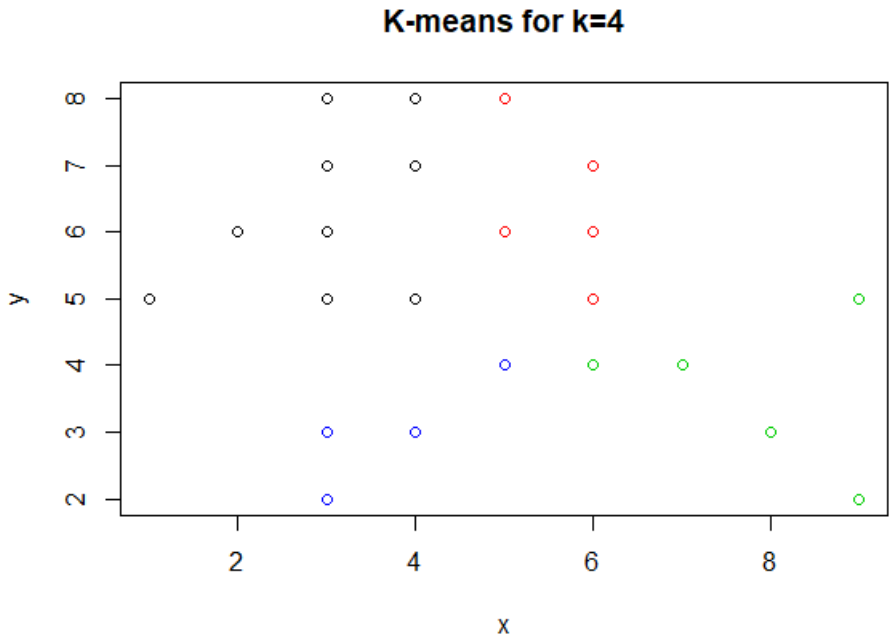


Τέλος τα νέα σημεία A(3,7) και B(4,5) ανατίθενται στο cluster 2, ενώ το σημείο C(2,9) σε θόρυβο.

Ο αλγόριθμος k-means απαιτεί την εκ των προτέρων γνώση του αριθμού των συστάδων που θέλουμε να γνωρίσουμε τα δεδομένα μας. Στόχος του k-means είναι να χωρίσει τα δεδομένα σε k συστάδων, ώστε να ελαχιστοποίηση τις αποστάσεις μεταξύ των σημείων της ίδιας συστάδας και να μεγιστοποιήσει την απόσταση μεταξύ διαφορετικών συστάδων. Ένα διάγραμμα του αθροίσματος των τετραγώνων των αποστάσεων εντός συστάδας για διαφορετικό αριθμό συστάδων αποτελεί μια λύση. Μια τέτοια μέθοδος είναι η μέθοδος του αγκώνα. Σύμφωνα με την οποία ο βέλτιστος αριθμός συστάδων εντοπίζεται στο σημείο απότομης αλλαγής κλίσης της καμπύλης.



Από την μέθοδο του αγκώνα, όπως, αυτή παρουσιάζεται στο παραπάνω διάγραμμα προκύπτει βέλτιστος αριθμός συστάδων ίσος με 4. Τα αποτελέσματα του k-means είναι κοντινά με του αλγορίθμου dbscan.

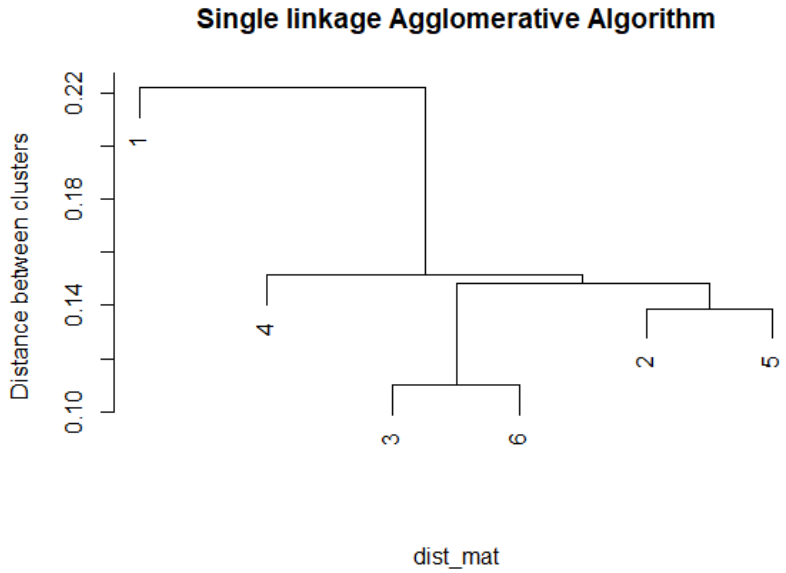


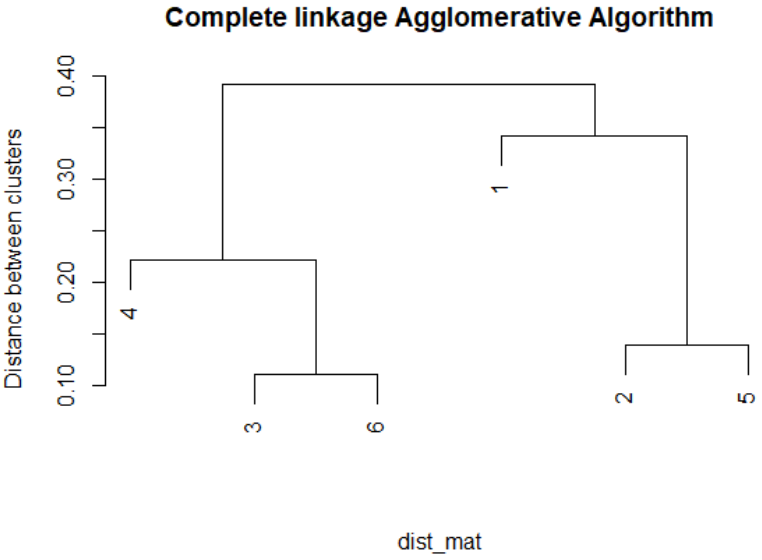
2.2.14. ΕΠΙΛΥΣΗ ΑΣΚΗΣΗΣ 14

Η ιεραρχική συσταδοποίηση παράγει ένα σύνολο από εμφωλευμένες συστάδες οργανωμένες σε ένα ιεραρχικό δέντρο. Μπορεί να αναπαρασταθεί με τη βοήθεια ενός δενδρογράμματος, το οποίο καταγράφει τις ακολουθίες συγχωνεύσεων και διαχωρισμών. Η συσσωρευτική ιεραρχική συσταδοποίηση έχει την ιδιαιτερότητα ότι όλα τα σημεία αρχικά αποτελούν μια συστάδα. Σε κάθε βήμα, συγχωνεύονται οι δύο πιο κοντινές συστάδες, δηλαδή το πλήθος των συστάδων μειώνεται κατά ένα. Αυτή η διαδικασία επαναλαμβάνεται, μέχρι ο αλγόριθμος να καταλήξει σε μια μοναδική συστάδα, η οποία θα εμπεριέχει όλα τα *n* δείγματα. Για την υλοποίηση του αλγορίθμου απαιτείται ορισμός της εγγύτητας (proximity) μεταξύ των συστάδων . για τις ανάγκες της άσκηση θα αναφερθούμε στις τεχνικής του απλού συνδέσμου και του σύνθετου συνδέσμου.

- Απλός Σύνδεσμο (single linkage): Η απόσταση μεταξύ δυο συστάδων ορίζεται ως η ελάχιστη απόσταση από τις αποστάσεις μεταξύ κάθε σημείου της κάθε συστάδας.
- Σύνθετου Συνδέσμου (complete linkage): Η απόσταση μεταξύ δυο συστάδων ορίζεται ως η μέγιστη από τις αποστάσεις μεταξύ κάθε σημείου της κάθε συστάδας.

Τα αποτελέσματα της συσσωρευτικής συσταδοποίησης με τη χρήση των δυο τεχνικών παρουσιάζονται στη συνέχεια. Οι αποστάσεις μεταξύ των συστάδων προκύπτουν από τον άξονα y.


Single linkage Agglomerative Algorithm
Distance between clusters
0.10
0.14
0.18
0.22
1
4
3
6
2
5
dist_mat


Complete linkage Agglomerative Algorithm
Distance between clusters
0.10
0.20
0.30
0.40
4
3
6
1
2
5
dist_mat

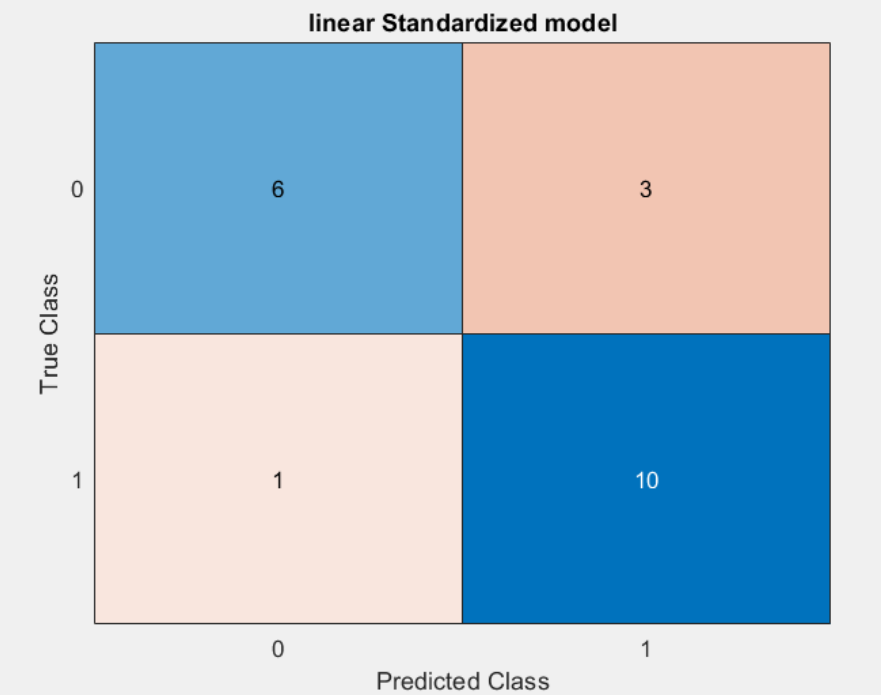
2.3. ΜΕΡΟΣ 3ο

Taskσ 1,2 και 3

Ο αλγόριθμος SVM έχει ως στόχο να βρει το βέλτιστο υπερεπίπεδο που διαχωρίζει δυο κλάσεις. Το βέλτιστο αυτό υπερεπίπεδο προκύπτει μεγιστοποιώντας την απόσταση του υπερεπιπέδου από τα διανύσματα στήριξης. Στην περίπτωση που οι δυο κλάσεις είναι γραμμικά διαχωρίσιμες ο αλγόριθμος linear SVM είναι ικανοποιητικός. Για πιο σύνθετα σύνορα απόφασης απαιτείται μετασχηματισμός των δεδομένων μέσω πολυωνυμικών ή gaussian συναρτήσεων κελύφους. Στη συγκεκριμένη εργασία, θα χρησιμοποιηθεί linear, polynomial και rbf μετασχηματισμός. Η κανονικοποίηση των δεδομένων θεωρείται απαραίτητη για την αύξηση της ακρίβειας την πρόβλεψης.

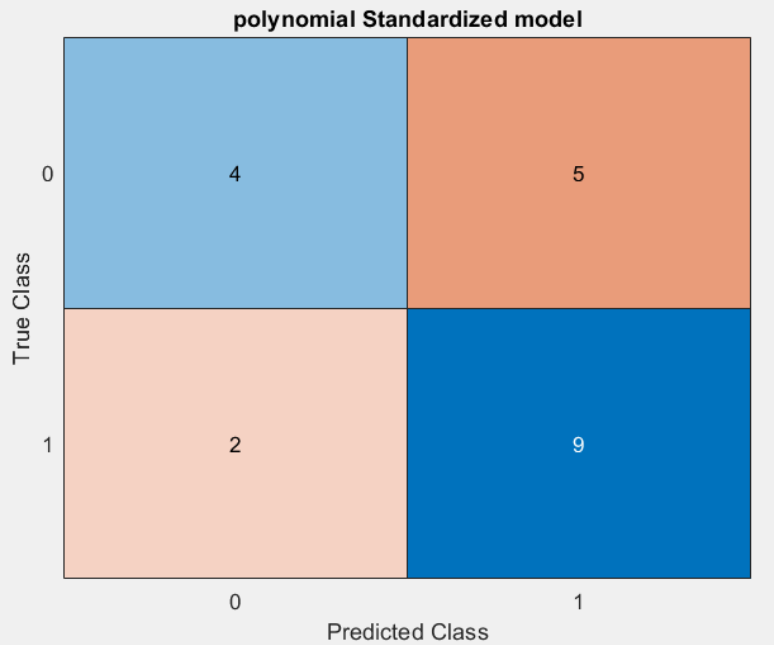
- **Linear Kernel Transformation:**

  Το μοντέλο SVM με χρήση γραμμικού μετασχηματισμού των δεδομένων επιτυγχάνει ακρίβεια ίση με 80% και σφάλμα 0.183030.



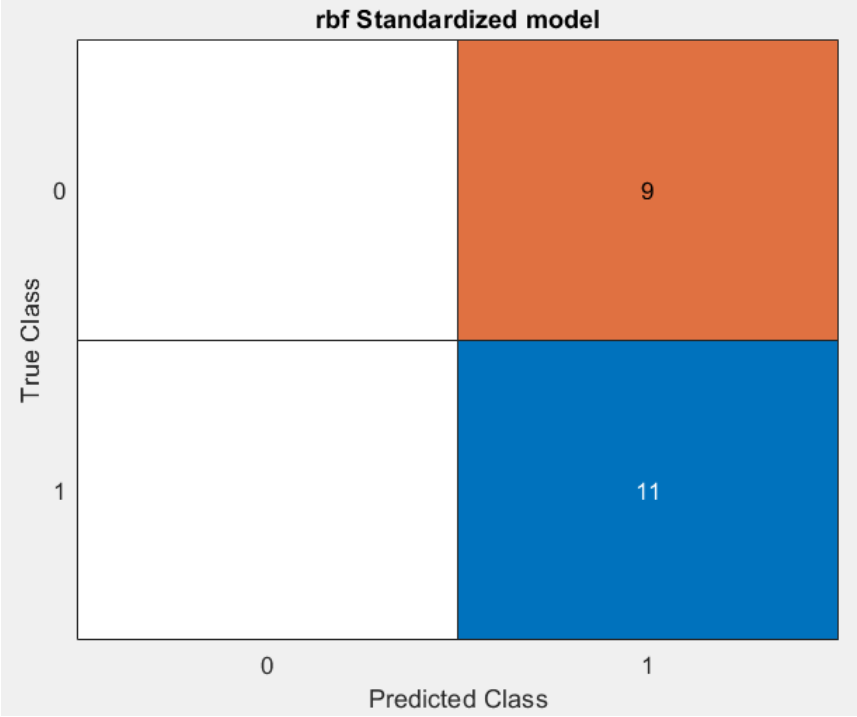
- **Polynomial Kernel Transformation:**

  Το μοντέλο SVM με χρήση πολυωνυμικού μετασχηματισμού των δεδομένων επιτυγχάνει ακρίβεια ίση με 65% και σφάλμα 0.323838.



- **RBF Kernel Transformation:**

  Το μοντέλο SVM με χρήση rbf μετασχηματισμού των δεδομένων επιτυγχάνει ακρίβεια ίση με 55% και σφάλμα 0.380000. Κατατάσσει σωστά όλες τις εικόνες της δεύτερης κατηγορίας και καμία της πρώτης. Μπορούμε να συμπεράνουμε ότι το μοντέλο έχει κάνει overfitting στη μια κατηγορία.



Βέλτιστη είναι η απόδοση του SVM με γραμμικό μετασχηματισμό καθώς επιτυγχάνει βέλτιστο accuracy και loss. Για το συγκεκριμένο πρόβλημα μπορούμε να συμπεράνουμε ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα.

Task 4,5 και 6:

Η αρχιτεκτονική νευρωνικού δικτύου που χρησιμοποιήθηκε αποτελείται από 2 συνελικτικά blocks, ενός επιπέδου συνέλιξης που ακολουθείται από τη συνάρτηση ενεργοποίησης ReLu και ένα επίπεδο pooling με χρήση της τεχνικής max pooling. Για την αποφυγή overfitting χρησιμοποιήθηκε η τεχνική dropout, που μηδενίζει κάποιο ποσοστό των κόμβων του νευρωνικού.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 1. Define a CNN architecture and training options:
layers = [
    imageInputLayer([256 256 1])

    convolution2dLayer(3,32, 'Padding', 'Same')
    reluLayer
    maxPooling2dLayer(2)
    dropoutLayer(0.2)

    convolution2dLayer(3,16, 'Padding', 'Same')
    reluLayer
    maxPooling2dLayer(2)
    dropoutLayer(0.2)

    fullyConnectedLayer(128)
    reluLayer

    fullyConnectedLayer(64)
    reluLayer

    dropoutLayer(0.2)
    fullyConnectedLayer(32)
    reluLayer

    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer];
```

Για την εκπαίδευση επιλέχθηκε η μέθοδος βελτιστοποίησης ‘sgdm’ και ρυθμός εκμάθησης 0.001. Το πλήθος των επαναλήψεων κατά την εκπαίδευση επιλέχθηκε ίσο με 10.

Η ακρίβεια του νευρωνικού δικτύου που επιτεύχθηκε είναι ίση με 0.70.

```
Training on single CPU.
Initializing input data normalization.
|=========================================================================================|
|  Epoch  |  Iteration  |  Time Elapsed  |  Mini-batch  |  Mini-batch  |  Base Learning  |
|         |             |   (hh:mm:ss)   |   Accuracy   |     Loss     |      Rate       |
|=========================================================================================|
|       1 |           1 |       00:00:02 |       60.00% |       0.6810 |          0.0010 |
|       5 |          50 |       00:02:27 |       80.00% |       0.5335 |          0.0010 |
|      10 |         100 |       00:04:48 |       90.00% |       0.3283 |          0.0010 |
|=========================================================================================|

accuracy =

    0.7000
```

Για πλήθος των επαναλήψεων ίσο με 10, η ακρίβεια του νευρωνικού δικτύου που επιτεύχθηκε είναι ίση με 0.75

```
Training on single CPU.
Initializing input data normalization.
|=========================================================================================|
|  Epoch  |  Iteration  |  Time Elapsed  |  Mini-batch  |  Mini-batch  |  Base Learning  |
|         |             |   (hh:mm:ss)   |   Accuracy   |     Loss     |      Rate       |
|=========================================================================================|
|       1 |           1 |       00:00:03 |       60.00% |       0.6899 |          0.0010 |
|       5 |          50 |       00:02:24 |       90.00% |       0.2464 |          0.0010 |
|      10 |         100 |       00:04:43 |      100.00% |       0.0790 |          0.0010 |
|      15 |         150 |       00:07:03 |      100.00% |       0.0105 |          0.0010 |
|      20 |         200 |       00:09:23 |      100.00% |       0.0014 |          0.0010 |
|=========================================================================================|

accuracy =

    0.7500
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Φιλιππάκης Μ. (2020), Προβλεπτική Αναλυτική, Πανεπιστημιακές σημειώσεις, ΠΜΣ Μεγάλα Δεδομένα και Αναλυτική, Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς
2. https://repository.kallipos.gr/bitstream/11419/2128/1/04_chapter03.pdf
3. https://saedsayad.com/logistic_regression.htm
4. https://www.rdocumentation.org/packages/amap/versions/0.8-18/topics/Kmeans?fbclid=IwAR0YfTFn1cXMmUfnjffmzaykXo5SdDBqQomhAQL_FVRmGFPhWmztdrwh-q8
5. https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf
6. https://www.mathworks.com/help/deeplearning/ug/create-simple-deep-learning-network-for-classification.html
7. https://www.mathworks.com/help/deeplearning/examples.html?category=deep-learning-with-images