Data Analytics eLearning Academy Final Project

Συγγραφέας: Σιδηροπούλου Νικολέτα

Φορέας: College Link

23/10/2021

# Εισαγωγή

Η ανάλυση δεδομένων αποτελεί έναν ευρέως διαδεδομένο και ανερχόμενο επιστημονικό κλάδο της εποχής μας. Με τη χρήση και συνεχή ανάπτυξη των πληροφοριακών συστημάτων και γλωσσών προγραμματισμού αποτελεί πεδίο μελέτης και διαχείρισης δεδομένων της στατιστικής. Μία από τις γνωστότερες γλώσσες προγραμματισμού που χρησιμοποιούνται διεθνώς για την επίλυση στατιστικών προβλημάτων και αναλύσεων αποτελεί η R.

Η στατιστική ανάλυση που θα εφαρμοστεί σε αυτή την αναφορά, βασίζεται σε δεδομένα ενός τυχαίου δείγματος 117 περιπτώσεων από τους φακέλους πωλήσεων σπιτιών μιας μεγάλης μεσιτικής εταιρείας στις ΗΠΑ από τις 15 Φεβρουαρίου έως τις 30 Απριλίου 1993. Οι διαθέσιμες μεταβλητές με τις μονάδες μέτρησης που χρησιμοποιούνται αποτελούνται από τη PRICE (τιμή πώλησης σε χιλιάδες USD), την SQFT (τετραγωνικά πόδια κατοικίας), την AGE (ηλικία κατοικίας σε χρόνια), τη FEATS(αριθμός επιπλέων χαρακτηριστικών από σύνολο 11), την NE (NorthEast, δίτιμη μεταβλητή που δηλώνει εάν η κατοικία βρίσκεται στη βορειοδυτική πλευρά της πόλης), τη COR (δίτιμη μεταβλητή που δηλώνει εάν η κατοικία είναι γωνιακή) και τέλος τη TAX (ετήσια φορολογία σε USD).

Στις σελίδες που ακολουθούν παρουσιάζεται η ανάλυση δομημένη σε τρία βασικά μέρη. Αρχικά πραγματοποιείται ο καθαρισμός και επεξεργασία των δεδομένων για την ανάλυση τους. Στη συνέχεια εφαρμόζεται περιγραφική στατιστική και οπτικοποίηση των επιμέρους μεταβλητών με κατάλληλα διαγράμματα για την κατανόηση των στατιστικών σχέσεων. Τέλος, διεξάγονται τεχνικές εξόρυξης δεδομένων μέσω στατιστικών μοντέλων για την εξαγωγή των τελικών συμπερασμάτων.

# Περίληψη

Η παρούσα εργασία αποτελείται από τρεις ενότητες. Στη πρώτη ενότητα παρουσιάζεται η εισαγωγή του αρχείου δεδομένων στη γλώσσα προγραμματισμού R, καθώς και η προετοιμασία και ο κατάλληλος καθαρισμός τους (Data Cleaning). Στην επόμενηενότητα εφαρμόζεται η περιγραφική στατιστική και η οπτικοποίηση των δεδομένων μέσω ελέγχων υποθέσεων, καθώς και η παρουσίαση και επεξήγηση πινάκων και διαγραμμάτων (Descriptive statistics and visualization). Στην τελευταία ενότητα, διεξάγονται διάφορα μοντέλα παλινδρόμησης και η κατηγοριοποίηση μέσω αλγορίθμων προκειμένου να εξορυστεί αποτελεσματικά νέα χρήσιμη πληροφορία από τα δεδομένα (Data mining) και την οδήγηση στα συνολικά συμπεράσματα της ανάλυσης.

# Μέρος πρώτο: Data Cleaning

Αρχικά, εισάγεται το αρχείο δεδομένων(House Prices), μορφής( .txt), στη γλώσσα προγραμματισμού R, αποθηκευμένο με το όνομα (Housedata.txt) και ελέγχεται η μορφή εισαγωγής του.

Έπειτα, γίνεται προσδιορισμός των στηλών που περιέχουν ελλείπουσες τιμές, αναπαριστώμενες με (“\*”) ώστε να αντικατασταθούν με NA τιμές. Η εφαρμογή αυτή γίνεται μέσω της μετατροπής των στηλών αυτών, AGE και TAX, από διανύσματα χαρακτήρων σε αριθμητικές μεταβλητές, οδηγώντας στη δημιουργία κατά αυτόν τον τρόπο των NA που ζητούνται. Εν συνεχεία, ελέγχεται ο τύπος των μεταβλητών του δείγματος με τις PRICE, SQFT, FEATS, NE και COR να αποτελούν ακέραιες μεταβλητές, ενώ οι υπόλοιπες να διατηρούν το τελικό τύπο της προηγούμενης διαδικασίας.

Στο επόμενο βήμα, παρέχεται η μετατροπή των δίτιμων αριθμητικών μεταβλητών, NE και COR σε κατηγορικές, μέσω της αντικατάστασης για τη κάθε μεταβλητή σε αντιστοιχία τις τιμές 0 και 1 με τους χαρακτήρες “No” και “Yes” δοσμένες πλέον στη μορφή που ζητείται.

Στη συνέχεια, πραγματοποιείται η αφαίρεση από τα δεδομένα των γραμμών αυτών που περιέχουν τουλάχιστον δύο NA τιμές. Με την αποθήκευση τους σε ένα νέο διάνυσμα, del, έπειτα από τον κατάλληλο λογικό έλεγχο, αφαιρείται το διάνυσμα τιμών από το Housedata και το και το αποτέλεσμα αποθηκεύεται εξαρχής, στο αρχικό δείγμα.

Ακολουθεί η μετατροπή της μεταβλητής SQFT (τετραγωνικά πόδια κατοικίας) σε SQM (τετραγωνικά μέτρα κατοικίας). Με την ισχύουσα ισοδυναμία τους, 1m2=10.746ft2, διαιρείται το διάνυσμα τιμών της SQFT με το 10.764, και αποθηκεύονται οι νέες τιμές στην υπάρχουσα μεταβλητή στα δεδομένα. Επίσης, τη μετονομάζεται από SQFT σε SQM.

Το τελευταίο μας βήμα σε αυτή τη διαδικασία αφορά την αντικατάσταση όλων των ελλειπουσών τιμών του δείγματος με τις νέες, στρογγυλοποιημένες στις εκάστοτε τιμές, αναμενόμενες. Οι τελευταίες, θα προβλεφθούν με τη χρήση κάθε φορά ενός εφαρμοσμένου μοντέλου γραμμικής παλινδρόμησης για τις μεταβλητές που περιέχουν NA, AGE και TAX. Πρωτίστως, αναγκαίος είναι ο ορισμός του δείγματος δεδομένων των γραμμών του Housedata που δεν περιέχουν κανένα NA και έπειτα η αποθήκευση τους ως subfull ώστε να χρησιμοποιηθούν στα μοντέλα παλινδρόμησης. Κατόπιν, εφαρμόζονται τα δύο αντίστοιχα γραμμικά μοντέλα ANCOVA. Το lmfullA, με εξαρτημένη μεταβλητή την AGE και επεξηγηματικές όλες τις υπόλοιπες του subfull και το lmfullT με εξαρτημένη μεταβλητή τη TAX και ανεξάρτητες όλες τις υπόλοιπες μεταβλητές αντιστοίχως. Για την εκτίμηση των αναμενόμενων τιμών, αποθηκεύονται σε μορφή data frame στα αντίστοιχα newdataA και newdataT, οι τιμές των επιμέρους μεταβλητών που αντιστοιχούν στις παρατηρήσεις που περιέχουν ΝΑ στην εξαρτημένη για τη πρώτη περίπτωση και ακολούθως για τα δεύτερα κατά παρόμοιο τρόπο. Έπειτα, προβλέπονται οι αναμενόμενες τιμές των AGE και TAX, με την εφαρμογή αρχικά του newdataA στο lmfullA, τη στρογγυλοποίηση των εξαγόμενων αποτελεσμάτων και την αποθήκευση τους στο διάνυσμα predA. Τέλος, γίνεται η αντικατάσταση του διανύσματος αυτού στις θέσεις των NA στο Housedata. Με παρόμοια διαδικασία, εκτιμώνται και αποθηκεύονται οι αναμενόμενες τιμές μέσω του lmfullT στο predT ώστε να αντικατασταθούν στις θέσεις της TAX στο Housedata.

# Μέρος δεύτερο: Descriptive Statistics & Visualization

Αυτή η ενότητα ξεκινάει με τη παροχή μερικών περιγραφικών μέτρων. Πιο συγκεκριμένα τη μέση τιμή(mean), τη διάμεσο(median), τη τυπική απόκλιση(sd), τη μικρότερη τιμή(minimum) και τέλος τη μέγιστη τιμή(maximum) για όλες τις αριθμητικές μεταβλητές PRICE, SQM, AGE, FEATS και TAX.

Η πρώτη μεταβλητή είναι η PRICE, με τις εξής τιμές για τη κατανομή της: mean=1076.11, sd=383.01, median=975, min=540 και max=2150. Η μέση τιμή αποτελεί μέτρο θέσης και είναι πιο ευαίσθητη στον υπολογισμό της σε ακραίες παρατηρήσεις της κατανομής. Ωστόσο η διάμεσος, εξ’ ορισμού αποτελεί την παρατήρηση της κατανομής που βρίσκεται στο 50% των διατεταγμένων τιμών του εύρους αυτής. Το γεγονός ότι η μέση τιμή είναι μεγαλύτερη από τη διάμεσο προμηνύει την ύπαρξη θετικά ασύμμετρης κατανομής. Παρατηρείται μία συγκέντρωση τιμών γύρω από τη πρώτη, η οποία λόγω των ακραίων παρατηρήσεων έλκεται προς αυτές στο εύρος τιμών. Η τυπική απόκλιση αποτελεί την τετραγωνική ρίζα της διακύμανσης της κατανομής και μας δίνει την εικόνα διασποράς των παρατηρήσεων γύρω από τη μέση τιμή. Όσο αφορά την ελάχιστη και μέγιστη τιμή, επρόκειτο για ένα αρκετά μεγάλο εύρος τιμών στη χρηματική αξία των σπιτιών του δείγματος.

Για τη δεύτερη μεταβλητή SQM, παρέχοντα οι τιμές: mean=154.78, sd=49.01, median=145.39, min= 77.76, max=348.38. Με τη μέση τιμή ελαφρώς μεγαλύτερη από τη διάμεσο, παραπέμπει σε μια πιθανών ελαφρώς θετικά ασύμμετρη κατανομή για την SQM. Το εύρος τιμών είναι αισθητά μικρότερο από την προηγούμενη μεταβλητή το οποίο εξαρτάται και από τη μονάδα μέτρησης του. Το διάστημα αυτό αντιπροσωπεύει τη διασπορά των παρατηρήσεων στην κατανομή των τετραγωνικών μέτρων κάθε κατοικίας δηλαδή τη τυπική απόκλιση.

Επακόλουθα, τα περιγραφικά μέτρα της AGE είναι: mean=18.1, sd=11.54, median=17, min=1, max=53. Ομοίως με πριν δημιουργείται η υποψία άλλης μία ελάχιστα θετικά ασύμμετρης κατανομής. Η αρκετά μικρή τιμή για την τυπική απόκλιση είναι ανάλογη του εύρους των τιμών της μεταβλητής και προκύπτει όπως εξηγήσαμε παραπάνω.

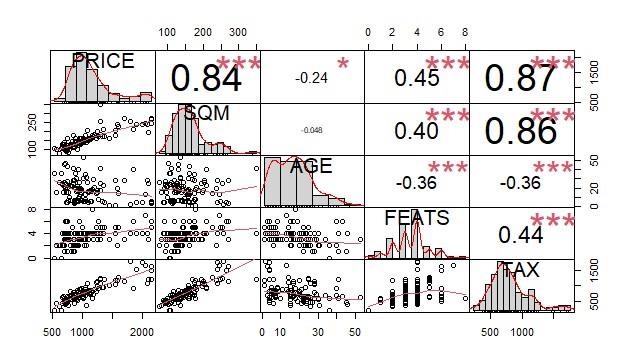
Εν συνεχεία οι τιμές για την μεταβλητή FEATS είναι οι εξής: mean=3.53, sd=1.4, median=4, min=0, max=8. Η μέση τιμή είναι ελαφρώς μικρότερη από τη διάμεσο της κατανομής. Ωστόσο, επειδή πρόκειται για διακριτή μεταβλητή με μικρό εύρος τιμών η σύγκριση των αποτελεσμάτων ίσως έδινε ένα παραπλανητικό αποτέλεσμα. Όσο αφορά το δείγμα, πληροφορεί για την ύπαρξη ενός μικρού αριθμού κατοικιών με ελάχιστες ή και καθόλου επιπλέον παροχές όπως αναφέρονται για τη μεταβλητή. Επίσης, παρατηρείται μία αναμενόμενη σχετικά μικρή τιμή και για την τυπική απόκλιση.

Τέλος για τη TAX, παρέχονται τα επόμενα περιγραφικά μέτρα: mean=793.75, sd=305.83, median=731, min=223, max=1765. Παρατηρείται άλλη μία περίπτωση αισθητά θετικά ασύμμετρης κατανομής. Επίσης υπάρχει ένα αρκετά μεγάλο εύρος τιμών και αρκετά μεγάλη διασπορά παρατηρήσεων γύρω από τη μέση τιμή όπως εύκολα συμπεραίνεται από τη τιμή της τυπικής απόκλισης. Η κατανομή μοιάζει αρκετά με αυτή της PRICE, κάτι που προκύπτει από τη μελέτη της στατιστικής συσχέτισης στον παρακάτω πίνακα.

Στη συνέχεια υπολογίζεται το πλήθος των παρατηρήσεων για το κάθε επίπεδο των κατηγορικών μεταβλητών. Πιο συγκεκριμένα όσο αφορά την NE, παρέχεται ένας πίνακας συχνοτήτων με 37 έναντι 72 παρατηρήσεων για την τιμή “No” και “Yes” αντίστοιχα. Επίσης ο πίνακας συχνοτήτων της COR επιστρέφει 87 παρατηρήσεις για την τιμή “No” και μόλις 22 για την τιμή “Yes”. Για την τελευταία η διαφορά στις συχνότητες εμφάνισης των τιμών είναι αρκετά μεγαλύτερη από την πρώτη.

Για τη διακριτή μεταβλητή FEATS, υπολογίζεται ο πίνακας σχετικών συχνοτήτων δηλαδή το διάνυσμα τιμών των συχνοτήτων εμφάνισης της κάθε παρατήρησης προς το μήκος των παρατηρήσεων συνολικά. Το τελευταίο επιστρέφει τη συχνότητα εμφάνισης της κάθε παρατήρησης συνολικά στον πληθυσμό μας. Συμπεραίνεται η αρκετά μικρή συχνότητα για τις μεγαλύτερες τιμές κυρίως από την 7 και μετά.

Έπειτα, απεικονίζεται ένα συγκεντρωτικό σημειόγραμμα για όλες τις αριθμητικές μεταβλητές. Το γράφημα που ζητάτε είναι το εξής:



## Εικόνα 1: Correlation Matrix of Housedata

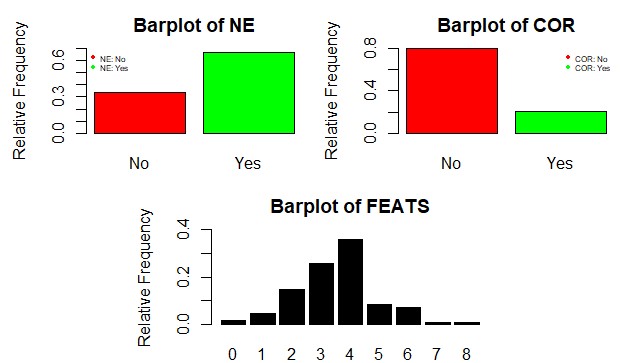
Αρχικά, η μέθοδος που ακολουθείται είναι ο υπολογισμός του συντελεστή συσχέτισης

Pearson που δίνει την ένταση της γραμμικής σχέσης των αριθμητικών μεταβλητών του γραφήματος. Η τιμή του δίνεται στο πάνω δεξιό μέρος και αντιστοιχεί κάθε φορά στις μεταβλητές που βρίσκονται επί του καθέτου και οριζοντίου άξονα. Τα πρόσημα του συντελεστή δείχνουν την, κατά μέσο όρο, είτε θετική είτε αρνητική συσχέτιση, ενώ οι αστερίσκοι δίνουν πληροφορία για το επίπεδο σημαντικότητας που θα απορρίπτονταν σε κάθε περίπτωση η μηδενική υπόθεση για τη στατιστική ισότητα των παραμέτρων με το 0. Ενδεικτικά, πιθανών εντοπίζεται μια ισχυρή θετική γραμμική συσχέτιση για τα ζεύγη μεταβλητών PRICE και SQM, καθώς επίσης για τη PRICE με TAX και της SQM με TAX, με συντελεστές μεγαλύτερους του 0.8. Ακολουθούν και για τις υπόλοιπες μεταβλητές ωστόσο με μικρότερους δείκτες που δείχνουν μια ασθενέστερη στατιστική σχέση, όπως για την αρνητική γραμμική σχέση των AGE και FEATS, παρομοίως και για την AGE με τη TAX.

Στη κύρια διαγώνιο απεικονίζονται τα ιστογράμματα αρχικά των συνεχών αριθμητικών μεταβλητών αλλά και της μοναδικής διακριτής FEATS. Η μη παραμετρική εκτίμηση της πυκνότητας που αναπαρίσταται από τη κόκκινη γραμμή αποτυπώνει το ύψος της καμπύλης της κατανομής. Επίσης, είναι φανερές οι ασυμμετρίες των κατανομών κατά μέσο όρο, για τις οποίες έγινε αναφορά προηγουμένως.

Οπτικά οι σχέσεις αυτές δίνονται και από τα σημειογράμματα στο κάτω αριστερό μέρος του γραφήματος. Για κάθε διάγραμμα βάση της συγκέντρωσης των παρατηρήσεων γύρω από τη γραμμή τάσης εφικτή καθίσταται η κρίση του βαθμού συσχέτισης και φυσικά το είδος αυτής που απεικονίζεται από την κλίση της γραμμής. Για τα ζεύγη μεταβλητών με θετική γραμμική συσχέτιση μεγάλες τιμές για την μία μεταβλητή συνεπάγονται, κατά μέσο όρο, μεγάλες τιμές για την άλλη ενώ για τα αντίστοιχα με αρνητική συσχέτιση μεγάλες τιμές για την μία μεταβλητή θα συνεπάγονται μικρές αναμενόμενες τιμές για την άλλη.

Στο επόμενο βήμα, δίνονται τα ραβδογράμματα για τις κατηγορικές μεταβλητές COR και NE και την διακριτή αριθμητική μεταβλητή FEATS:



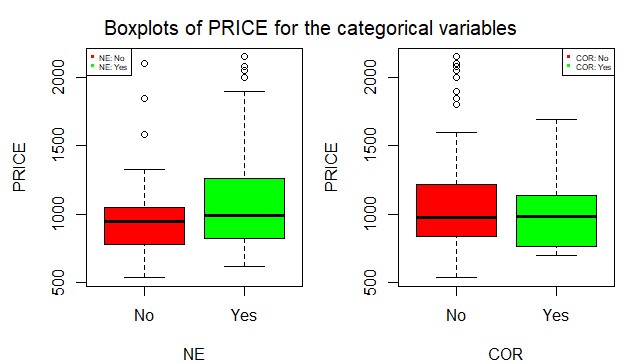
## Εικόνα 2: Barplots of NE, COR & FEATS

## 

Στα πάνω σχήματα αναπαρίστανται, αριστερά το ραβδόγραμμα των σχετικών συχνοτήτων της NE μεταβλητής και δεξιά το αντίστοιχο ραβδόγραμμα της COR μεταβλητής.

Για το πρώτο, είναι φανερή η υπερίσχυση της σχετικής συχνότητας εμφάνισης της τιμής “Yes” έναντι της τιμής “No”. Αυτό εξηγεί ότι οι περισσότερες κατοικίες ανήκουν στην βορειοδυτική γεωγραφική θέση σε σχέση με τις υπόλοιπες στο σύνολο. Συνεχίζοντας με το γράφημα για τις σχετικές συχνότητες της COR, η εμφάνιση της τιμής “No” είναι σχεδόν τετραπλάσια από αυτή της τιμής “Yes”. Με άλλα λόγια, τα μη γωνιακά σπίτια αποτελούν σχεδόν μόλις το ¼ των συνολικών κατοικιών του δείγματος. Αναφορικά με το κάτω γράφημα, απεικονίζεται το ραβδόγραμμα της διακριτής μεταβλητής FEATS. Η επικρατούσα τιμή αντιστοιχεί στο 4 και είναι αρκετά κοντά στη μέση τιμή της κατανομής. Επομένως, παρατηρείται μια συγκέντρωση τιμών γύρω από τη μέση τιμή, και ένα σχήμα κατανομής με πιθανή θετική ασυμμετρία.

Παρακάτω, αναπαρίστανται τα θηκογράμματα της μεταβλητής PRICE ανά κατηγορία για τα διαφορετικά επίπεδα της μεταβλητής NE, και έπειτα της COR:



## Εικόνα 3: Boxplots of PRICE for the categorical variables

## 

Με κόκκινο χρώμα απεικονίζονται οι κατανομές της PRICE για τη τιμή “No” τόσο για τη κατηγορική μεταβλητή NE όσο και για τη COR, ενώ με πράσινο οι αντίστοιχες κατανομές για τη τιμή “Yes”.

Ξεκινώντας από το αριστερό γράφημα δίνονται οι θήκες για τις κατανομές της PRICE για τα δύο προαναφερθέντα επίπεδα της NE. Αρχικά, είναι αρκετά εμφανές το γεγονός ότι η κατανομή της “Yes” είναι ελαφρώς πιο ψηλά από αυτή της “No”, με άλλα λόγια η βορειοδυτική θέση δείχνει, κατά μέσο όρο, να επιδρά θετικά στη τιμή των κατοικιών.

Επίσης, διακρίνεται η διασπορά για την κάθε κατανομή, πιο συγκεκριμένα μέσω της θέσης της διαμέσου που αποτελεί το 50% των μεσαίων παρατηρήσεων σε σχέση πάντα με το μέγεθος των θηκών. Οι θέσεις των διαμέσων τους με τη μαύρη οριζόντια γραμμή εντός τους, βρίσκονται αρκετά κοντά σε σχέση με τις τιμές της PRICE, άρα κινούνται σε παρόμοιο εύρος τιμών. Όσο αφορά τις απολήξεις τους τόσο στο πάνω όσο και το κάτω μέρος των θηκών, η κατανομή για τη τιμή “Yes” φαίνεται να παρουσιάζει θετική ασυμμετρία ενώ και οι δύο έχουν κάποιες θετικές ακραίες τιμές.

Αρκετά κοντά βρίσκονται και οι κατανομές της PRICE για τα διαφορετικά επίπεδα της COR όπως και οι διάμεσοι τους. Αυτό εξηγεί και την πιθανή μη στατιστική επίδραση της κατηγορικής μεταβλητής στην αριθμητική αλλά σε κάθε περίπτωση αυτό μεταβάλλεται σύμφωνα με τα μεγέθη των επιμέρους κατανομών, με θετικές ακραίες τιμές για τη τιμή “No”, και πιθανή θετική ασυμμετρία για τη τιμή ‘Yes”.

Το επόμενο βήμα στην ανάλυση αποτελεί η εύρεση των μέσων τιμών της μεταβλητής PRICE για τα διαφορετικά επίπεδα της μεταβλητής NE και αντίστοιχα της COR. Ως αποτέλεσμα δίνονται οι τιμές 992.43 και 1119.11 για τη πρώτη, ομοίως 1095.28 και1000.32 για τη δεύτερη αντίστοιχα στα επίπεδα “No” και “Yes”. Ωστόσο, για να ελεγχθεί κατά πόσο παρατηρείται στατιστική διαφορά μεταξύ των μέσων, εφαρμόζεται αρχικά ένα Shapiro test για την κατανομή της PRICE ως προϋπόθεση της περαιτέρω εφαρμογής του πρώτου ελέγχου. Με εναλλακτική να μην ακολουθεί κανονική κατανομή υπάρχουν ισχυρές ενδείξεις για την απόρριψη σε επίπεδο σημαντικότητας α=1% της μηδενική υπόθεση που αναφέρεται στην ύπαρξη κανονικότητας. Ως απόρροια αυτού, επιλέγεται η εφαρμογή ενός μη παραμετρικού ελέγχου, του Wilcoxon test, για τη στατιστική ισότητα των διαμέσων. Τα αποτελέσματα για τις κατανομές της PRICE στα επίπεδα και των δύο κατηγορικών οδηγούν στην μη απόρριψη της μηδενικής υπόθεσης που αναφέρεται στην στατιστική ισότητα των επιμέρους διαμέσων για επίπεδο σημαντικότητας α=10%, έναντι της εναλλακτικής που ισχυρίζεται τη στατιστική διαφορά των διαμέσων.

Στη συνέχεια, ελέγχονται και επιλέγονται οι μεταβλητές με υψηλή θετική ασυμμετρία ώστε να εφαρμοστούν σε αυτές ένας έλεγχος κανονικότητας. Οι μεταβλητές είναι οι PRICE, SQM και TAX, με τιμή, skewness>1. Έπειτα, μέσω ενός Shapiro test, επιβεβαιώνεται το αρχικό συμπέρασμα και για τις τρεις, αφού απορρίπτεται για ένα επίπεδο σημαντικότητας α=1% η μηδενική υπόθεση που αναφέρεται στην ύπαρξη κανονικότητας.

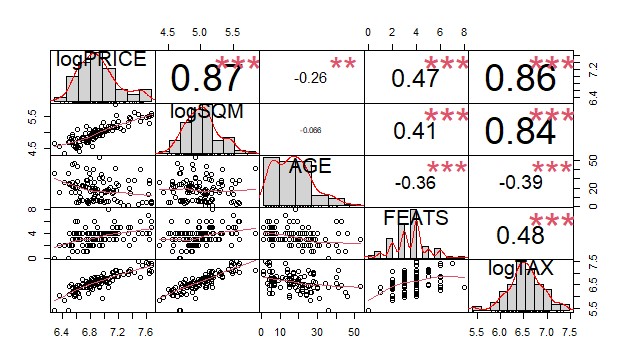
Έπειτα υπολογίζεται ο λογάριθμος τους και εφαρμόζεται εκ νέου το Shapiro test για κάθε μία από τις κατανομές. Πρωτίστως, στον έλεγχο της PRICE κατανομής, επιστρέφεται μία εξαιρετικά μικρή p-value που επιτρέπει την απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας α=1%. Για ένα περαιτέρω έλεγχο αυτού, εφαρμόζεται και ένας επιπλέον έλεγχος, το Lillie test, όπου το επιστρεφόμενο αποτέλεσμα δεν επιτρέπει την απόρριψη της μηδενικής υπόθεσης για επίπεδο σημαντικότητας α=1% που αναφέρεται στη κανονικότητα της κατανομής έναντι της εναλλακτικής που ισχυρίζεται το αντίθετο.

Η ερμηνεία των, ελαφρώς, αυτών διαφορετικών αποτελεσμάτων, οφείλεται στον τρόπο υπολογισμού των επιμέρους στατιστικών των ελέγχων. Πιο συγκεκριμένα, το Shapiro test είναι πιο ‘’ευαίσθητο’’ στην ύπαρξη ακραίων παρατηρήσεων στην κατανομή σε αντίθεση με το Lillie test που δεν ακολουθεί αυτό τον τρόπο υπολογισμού.

Για τις μεταβλητές SQM και TAX, μέσω των αποτελεσμάτων εφαρμογής των Shapiro test γίνεται δεκτή η ύπαρξη κανονικότητας σε ένα επίπεδο σημαντικότητας α= 10%. Κατά αυτό τον τρόπο προκύπτει ο ισχυρισμός της διόρθωσης του προβλήματος της έντονης ασυμμετρίας για τις αριθμητικές μεταβλητές και της προσέγγισης στη κανονική κατανομή. Έχοντας τις νέες τιμές μετά το μετασχηματισμό τους με τον υπολογισμό των λογαρίθμων για τη PRICE, SQM και TAX, εφαρμόζεται η αντικατάσταση στις προηγούμενες ανά στήλη της κάθε μεταβλητής στο Housedata.

Συνακολούθως, δίνονται και τα νέα αντίστοιχα ονόματα για κάθε στήλη.

Παρακάτω επαναλαμβάνεται ένα συγκεντρωτικό γράφημα για τις νέες αριθμητικές τιμές μετά τους κατάλληλους μετασχηματισμούς με λογαρίθμους:



## Εικόνα 4: New Correlation Matrix of Housedata

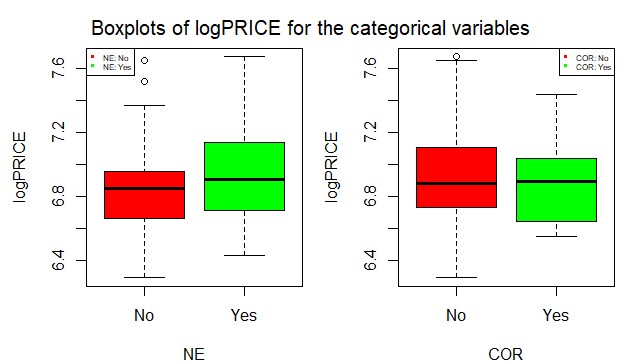
## 

Από το παραπάνω γράφημα μπορούν εύκολα να διακριθούν αναλυτικότερα οι κατανομές των αριθμητικών μεταβλητών μετά το μετασχηματισμό και τη συγκέντρωση των παρατηρήσεων στο κάθε σχήμα ξεχωριστά, για τα ιστογράμματα. Με το τελευταίο, γίνονται αντιληπτές οπτικά οι υποθέσεις που κάναμε προηγουμένως για τις logPRICE, logSQM και logTAX.

Κατά μέσο όρο υπάρχουν και ισχυρότερες συσχετίσεις για τις μεταβλητές ανά ζεύγη όπως ερμηνεύεται από τους συντελεστές συσχέτισης του Pearson, οι οποίοι είναι κατά απόλυτη τιμή μεγαλύτεροι από τους αντίστοιχους προηγουμένως. Οι αστερίσκοι οδηγούν επικουρικά στο ίδιο συμπέρασμα.

Τέλος, παρατηρούνται ισχυρότερες γραμμικές συσχετίσεις, κατά μέσο όρο, κυρίως για τα ζεύγη με τους αντίστοιχους μεγαλύτερους συντελεστές συσχέτισης, όπως η logPRICE με τη logSQM, ή τη logTAX, ακολουθώντας και όλα τα υπόλοιπα. Εντονότερη γραμμικότητα ίσως εμφανίζουν και ζεύγη μεταβλητών με αρνητική συσχέτιση, όπως η AGE με τη logTAX.

Για την ερμηνεία του τελευταίου γραφήματος για το δεύτερο μέρος της ανάλυσης, δίνονται δύο θηκογράμματα που αφορούν τις κατανομές των δειγμάτων της logPRICE ως προς τα επίπεδα των κατηγορικών μεταβλητών NE και COR αντιστοίχως:



## Εικόνα 5: Boxplots of logPRICE for the categorical variables

## 

Με τη σύγκριση του γραφήματος με το προηγούμενο, παρατηρείται ότι για τις τιμές ‘’No’’ τόσο στη μία κατηγορική όσο και στην άλλη, οι απολήξεις προς τα πάνω και κάτω, δηλαδή δεξιά και αριστερά των κατανομών είναι μεγαλύτερες. Αυτό σημαίνει ότι τα δεδομένα εκτείνονται περισσότερο και ίσως πιο συμμετρικά στο εύρος τιμών. Οι θέσεις των διαμέσων κινούνται σε παρόμοια επίπεδα μεταξύ των κατηγοριών.

Όσο για τις κατανομές ανά γράφημα φαίνονται να κινούνται στο ίδιο εύρος και πάλι, με άλλα λόγια δεν φαίνεται η logPRICE να επηρεάζεται ιδιαίτερα από τα επίπεδα των κατηγορικών.

# Μέρος τρίτο: Data Mining

Στο τρίτο και τελευταίο μέρος το πρώτο βήμα είναι η εφαρμογή ενός μοντέλου ANCOVA, με εξαρτημένη μεταβλητή τη logPRICE και ανεξάρτητες όλες τις υπόλοιπες του Housedata.

Ακολουθούν αναλυτικά οι ερμηνείες των βi παραμέτρων του μοντέλου.

Λόγω της ύπαρξης τόσο αριθμητικών όσο και κατηγορικών μεταβλητών στο μοντέλο μας, αρχικά ερμηνεύεται η σταθερά(intercept), ως το επίπεδο αναφοράς για τις κατηγορικές NE και COR. Με άλλα λόγια αποτελεί το μέσο αναμενόμενο λογάριθμο της αξίας μίας κατοικίας, εάν αυτή δε βρίσκεται στη βορειοδυτική πλευρά της πόλης και δεν είναι γωνιακή.

Σχετικά με την ερμηνεία της για τις αριθμητικές μεταβλητές αποτελεί την αναμενόμενη τιμή της logPRICE, για μηδενικές τιμές στις πρώτες, κάτι που δεν ερμηνεύεται πρακτικά πάντα όπως συμβαίνει και με τα δεδομένα που εξετάζονται. Ένα παράδειγμα αυτού είναι μία κατοικία με μηδενική τιμή SQM που ουσιαστικά δεν νοείται.

Η ερμηνεία της εκτιμώμενης παραμέτρου για την κάθε αριθμητική μεταβλητή αντίστοιχα ισούται με την αναμενόμενη τιμή της εξαρτημένης στη μονάδα μέτρησης της. Λόγω της εμφάνισης λογαρίθμων στις μεταβλητές είναι σημαντική η ορθή ερμηνεία τους, με τη χρήση αντιστρόφου μετασχηματισμού αλλά και το είδος της συσχέτισης μεταξύ των επεξηγηματικών με την εξαρτημένη που μπορεί να μετατραπεί σε εκθετική.

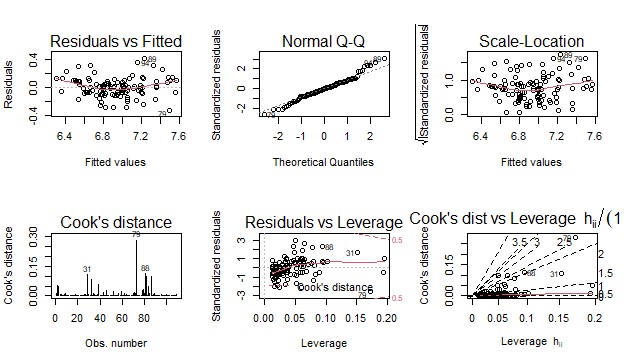
Όσο αφορά τις παραμέτρους των κατηγορικών επεξηγηματικών, εάν γινόταν αναφορά σε μία παρατήρηση με τιμή ‘’No’’ για τη COR, με αφαίρεση της τιμής της εκτιμώμενης παραμέτρου, για το επίπεδο της COR, από τη τιμή της σταθεράς, θα εξάγονταν η αναμενόμενη διαφορά στη logPRICE για την παρατήρηση με αυτά τα χαρακτηριστικά. Αντίστοιχα δίνεται και η ερμηνεία της logPRICE με αλλαγή στο επίπεδο της NE.

Στη συνέχεια, εφαρμόζεται ένας βηματικός έλεγχος μέσω αλγορίθμου(stepwise) ώστε να επιλεγχθεί το τελικό μοντέλο, mfinal, με εξαρτημένη τη logPRICE. Οι μεταβλητές που επεξηγούν το μοντέλο μας, είναι η logSQM, AGE, CORYes και τέλος η logTAX.

Όσο αφορά τη σταθερά του μοντέλου αποτελεί το σημείο αναφοράς της μοναδικής κατηγορικής COR, επομένως αναφέρεται σε μια κατοικία με χαρακτηριστικά την τιμή “No” όταν φυσικά η παράμετρος της CORYes όπως και οι υπόλοιπες αριθμητικές είναι ίσες με 0 και δίνει τη μέση αναμενόμενη τιμή της logPRICE.

Αξίζει να αναφερθεί ότι κάποιες εκ των εκτιμώμενων παραμέτρων φαίνεται να έχουν θετική στατιστική συσχέτιση με την εξαρτημένη μεταβλητή, όπως η logSQM, η AGE και η logTAX, ενώ οι υπόλοιπες έχουν αρνητικό πρόσημο στην εκτίμηση τους. Βέβαια, ιδιαίτερη προσοχή στη περαιτέρω ανάλυση των σχέσεων θα απαιτούνταν για τη περίπτωση ύπαρξης πολυσυγγραμικότητας ανάμεσα στις μεταβλητές. Επίσης, στην πλειονότητα επρόκειτο για στατιστικά σημαντικές μεταβλητές κάτι που αναμένεται με αρκετές ενδείξεις από τις τιμές p-value για τις εκτιμήσεις τους. Τέλος, δίνεται και μία αρκετά καλή τιμή για το R2adj=0.81 που αποτελεί το ποσοστό μεταβλητότητας που μπορεί να ερμηνευθεί από το μοντέλο, σταθμισμένο ως προς τον αριθμό των επεξηγηματικών μεταβλητών.

Κατόπιν, ελέγχονται και οι υποθέσεις του μοντέλου, συνοδευτικά με τα επακόλουθα διαγνωστικά γραφήματα:



## Εικόνα 6: Diagnostic Plots of mfinal model

## 

Οι υποθέσεις που ελέγχονται είναι η κανονικότητα, ομοσκεδαστικότητα και ανεξαρτησία των εκτιμώμενων καταλοίπων καθώς και η γραμμικότητα στη στατιστική συσχέτιση της εξαρτημένης με τις ανεξάρτητες μεταβλητές. Τέλος γίνεται η υπόθεση και της μη ύπαρξης πολυσυγγραμικότητας δηλαδή ανεξαρτησία μεταξύ των επεξηγηματικών μεταβλητών.

Από το γράφημα Residuals vs Fitted, παρατηρείται ένα νέφος το οποίο δεν προϊδεάζει για την ύπαρξη κάποιας δομής στα κατάλοιπα. Ωστόσο, ίσως διακρίνεται υψηλότερη διακύμανση καταλοίπων προς τα άκρα της κατανομής. Ενδεχομένως να υπάρχει παραβίαση της γραμμικότητας στη σχέση της εξαρτημένης και επεξηγηματικής μεταβλητής. Περαιτέρω θα διαπιστωθεί από την εφαρμογή και κάποιων ελέγχων υποθέσεων, όπως επίσης και πιθανοί τρόποι αντιμετώπισης.

Αναφορικά με τον έλεγχο της κανονικότητας για τα κατάλοιπα, παρέχεται μία πρώιμη πληροφορία από το Normal Q-Q Plot, όπου γίνεται η υπόθεση κατά πόσο η κατανομή των καταλοίπων συμφωνεί με την κανονική κατανομή, σύμφωνα με τα θεωρητικά και δειγματικά ποσοστημόρια. Από το γράφημα δεν φαίνονται κάποιες σημαντικές αποκλείσεις εκτός από κάποιες παρατηρήσεις στα άκρα που ωστόσο δε είναι ανησυχητικές.

Από το γράφημα Scale-Location δίνεται η εικόνα για πιθανή ύπαρξη εταιροσκεδαστικότητας στο δείγμα, δηλαδή μη σταθερής διακύμανσης των καταλοίπων. Πιθανόν να υπάρχει αύξηση διακύμανσης στα άκρα της κατανομής, το οποίο θα ελεγχθεί και αργότερα.

Τέλος από τα υπόλοιπα τρία γραφήματα, ο δείκτης Cook’s distance αποτελεί ένα μέτρο για την επίδραση ενός σημείου που απέχει αισθητά από τη γραμμή παλινδρόμησης με μεγάλες τιμές για την επεξηγηματική μεταβλητή(influence point). Ενδεικτικά οι παρατηρήσεις 31,79 και 88 ανήκουν σε αυτά. Οι τελευταίες αποτελούν και σημεία(Leverage) δηλαδή με μεγάλη τιμή της επεξηγηματικής. Από τα γραφήματα δεν φαίνεται κάποιο ιδιαίτερο πρόβλημα.

Για την διαπίστωση της ευστάθειας των παραπάνω, εφαρμόζονται και κάποιοι διαγνωστικοί έλεγχοι υποθέσεων. Αρχικά εφαρμόζεται ένας Shapiro έλεγχος για την κατανομή των καταλοίπων, όπου σύμφωνα με τη δοσμένη p-value και την εναλλακτική υπόθεση που χρησιμοποιήθηκε και σε προηγούμενους ελέγχους, δεν καθίσταται εφικτή α απόρριψη σε ένα επίπεδο σημαντικότητας α=10% της μηδενικής υπόθεσης για ύπαρξη κανονικότητας.

Στη συνέχεια ελέγχεται εάν υπάρχει παραβίαση της ομοσκεδαστικότητας. Από το αποτέλεσμα του Bartlett test βάση ισχυρών ενδείξεων δεν απορρίπτεται η μηδενική υπόθεση σε επίπεδο σημαντικότητας α=10% για σταθερή διακύμανση στο δείγμα των καταλοίπων στα διαστήματα διαχωρισμού των αναμενόμενων τιμών, έναντι της εναλλακτικής που υποθέτει τουλάχιστον μία εκ των τριών διαστημάτων να διαφέρει στατιστικά από τις υπόλοιπες.

Έπειτα πραγματοποιείται έλεγχος για την ανεξαρτησία των καταλοίπων. Επιλέγεται αρχικά το Durbin-Watson test με μηδενική υπόθεση τη μη ύπαρξη αυτοσυσχέτισης στα δεδομένα και εναλλακτική να είναι διάφορη του μηδενός. Η τιμή p-value που επιστρέφεται είναι μικρότερη του 0.05, δηλαδή δεν γίνεται απόρριψη της μηδενικής υπόθεσης σε επίπεδο εμπιστοσύνης α=10%.

Περαιτέρω μπορεί να διερευνηθεί με κάποιο έλεγχο όπως το Runs test που δίνει ισχυρές ενδείξεις για τη μη απόρριψη της υπόθεσης για μη ύπαρξη δομής στα κατάλοιπα σε επίπεδο εμπιστοσύνης α= 10%.

Τελευταία υπόθεση που ελέγχεται είναι η γραμμικότητα μέσω του residualPlots ελέγχου. Λόγω του επιστρεφόμενου αποτελέσματος του Tukey test, απορρίπτεται η μηδενική υπόθεση για μη ύπαρξη στατιστικά σημαντικού δευτεροβάθμιου όρου που δεν ελήφθη υπόψη και συγκλίνει προς την εναλλακτική υπόθεση που ισχυρίζεται την ύπαρξη αυτού, για επίπεδο εμπιστοσύνης α=1%. Η μικρή τιμή του p-value προκύπτει από μικρές τιμές των επιμέρους μεταβλητών για τον ίδιο έλεγχο. Ένας τρόπος αντιμετώπισης αυτού είναι η πρόσθεση του δευτεροβάθμιου όρου στην αντίστοιχη μεταβλητή του μοντέλου και η εφαρμογή ενός νέου βηματικού ελέγχου. Το νέο επιστρεφόμενο μοντέλο είναι το mfinal1 με στατιστικά σημαντικές επεξηγηματικές μεταβλητές σύμφωνα με τις pvalue τιμές κατά πλειοψηφία, και ελαφρώς καλύτερη τιμή για το R2adj περίπου ίση με 0.82 που ερμηνεύεται όπως εξηγήθηκε παραπάνω.

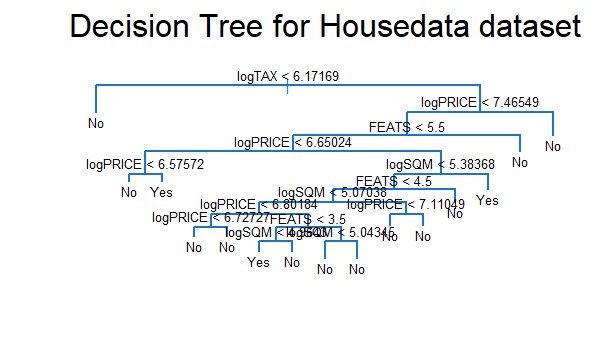
Έπεται δημιουργείται η νέα κατηγορική μεταβλητή catFEATS, η οποία διαβαθμίζει τις τιμές της FEATS στα τρία επίπεδα “Low”,”Moderate” και “High”, βάση του αριθμού των επιπλέων χαρακτηριστικών της κάθε κατοικίας. Στη συνέχεια εφαρμόζεται ένα μοντέλο ANOVA για τη logPRICE στα επίπεδα της catFEATS. Με άλλα λόγια ελέγχεται η στατιστική σημαντικότητα άρα η επίδραση της κατηγορικής στην εξαρτημένη μεταβλητή. Με μηδενική υπόθεση τη μη ύπαρξη στατιστικής διαφοράς στις μέσες τιμές, βάση της επιστρεφόμενης τιμής p-value για το μοντέλο υπάρχουν ισχυρές ενδείξεις για απόρριψη της υπόθεσης για επίπεδο σημαντικότητας α=1% και η σύγκλιση στην εναλλακτική που ισχυρίζεται ότι τουλάχιστον μία εκ των μέσων διαφέρει στατιστικά από τις υπόλοιπες.

Με την πρόσθεση της catFEATS στο Housedata πραγματοποιείται η επιλογή ενός βέλτιστου μοντέλου λογιστικής παλινδρόμησης, finalN, μέσω ενός βηματικού ελέγχου. Το μοντέλο αυτό έχει ως εξαρτημένη μεταβλητή την NE και επεξηγηματικές όλες τις υπόλοιπες των δεδομένων εκτός της κατηγορικής που εισήχθη προηγουμένως.

Πλέον με τη χρήση του finalN γίνεται η εκτίμηση της πιθανότητας ένα σπίτι να βρίσκεται στη βορειοδυτική πλευρά με καθορισμένες τις υπόλοιπες τιμές, SQM=180, AGE=15, FEATS=5 και

TAX=1000, με τη χρήση των λογαρίθμων σε όσες απαιτείται στο μοντέλο. Με την εφαρμογή των νέων αυτών δεδομένων και την ύψωση του finalN στον εκθέτη υπολογίζεται η πιθανότητα επιτυχίας που ζητάτε η οποία ισούται περίπου με 0.88.

Το τελευταίο μας βήμα είναι η εφαρμογή και οπτικοποίηση ενός δέντρου αποφάσεων για τη μεταβλητή COR, όπως επίσης και η παροχή του ποσοστού των σωστά εκτιμώμενων τιμών .Με το ομώνυμο πακέτο, αρχικά γίνονται οι απαραίτητες μετατροπές των μεταβλητών σε διατάξιμες και εφαρμόζεται το δέντρο αποφάσεων για τη μεταβλητή fitCOR για τα δεδομένα χωρίς τη χρήση της catFEATS με τις μεταβλητές που χρησιμοποιούνται από τον αλγόριθμο να είναι οι logTAX, logPRICE, FEATS και logSQM. Ακολούθως, παρέχεται ο πίνακας συχνοτήτων της COR όπου δίνονται οι σωστές και λανθασμένες εκτιμήσεις για τα δύο επίπεδα, “Yes” και “No”. Με τη διαίρεση του αθροίσματος των τιμών της κύριας διαγώνιου προς το σύνολο των παρατηρήσεων δίνεται ως πηλίκο το ποσοστό των σωστά εκτιμώμενων τιμών.



## Εικόνα 7: Decision Tree for Housedata dataset

Η ανάλυση του παραπάνω γραφήματος ξεκινάει από τον πρώτο κόμβο και προχωρώντας προς τα κάτω, κάθε φορά εφαρμόζεται ο λογικός έλεγχος που παρέχεται ως συνθήκη για τη τιμή της επιμέρους μεταβλητής. Εάν αυτή η συνθήκη είναι αληθής επιλέγεται η αριστερή κατεύθυνση, σε αντίθετη περίπτωση η δεξιά, μέχρι τη διαπίστωση της τελικής τιμής για την εκτίμηση της COR.

Δοθέντων κάποιων τιμών για τις μεταβλητές που περιέχονται στο δέντρο αποφάσεων ζητείται η εκτίμηση εάν μία κατοικία είναι γωνιακή ή όχι. Με τη χρήση και πάλι λογαρίθμων όπως απαιτούνται για τις PRICE=1000, SQM=150, FEATS=4 και TAX=800 και την εφαρμογή της διαδικασίας που εξηγήθηκε παραπάνω, προκύπτει το συμπέρασμα πως μία κατοικία των ανωτέρω χαρακτηριστικών δεν είναι γωνιακή.

# Συμπεράσματα

Προχωρώντας από τη πρώτη ενότητα προς την δεύτερη, δίνονται τα κατάλληλα γραφήματα και έπειτα οι έλεγχοι για τις στατιστικές συσχετίσεις των μεταβλητών. Παρατηρούνται θετικά ασύμμετρες κατανομές σχεδόν για κάθε αριθμητική μεταβλητή που εύλογα οδηγεί στη διαπίστωση ύπαρξης ακραίων παρατηρήσεων. Οι προβλέψεις των σχέσεων αυτών εξακριβώνονται μέσω στατιστικών ελέγχων. Περαιτέρω ελέγχεται η επίδραση των κατηγορικών μεταβλητών στην αριθμητική εξαρτημένη μεταβλητή συμπεραίνοντας τη μη στατιστική επίδραση τους πάνω στη δεύτερη. Στη τελική ανάλυση, εφαρμόζονται διάφορα μοντέλα γραμμικής παλινδρόμησης με αρκετά καλή αξιολόγηση της προσαρμογής τους καθώς συμπεραίνεται και η στατιστική επίδραση της διατάξιμης στην εξαρτημένη αριθμητική μεταβλητή. Τέλος, παρουσιάζονται και γίνεται χρήση κάποιων χρήσιμων εφαρμογών για κατηγοριοποίηση των δεδομένων.