ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

"Περιγραφική στατιστική, στατιστικοί έλεγχοι και δημιουργία μοντέλων πρόβλεψης"

ΚΕΛΕΠΙΡΗ ΖΩΗ, 78
ΚΟΥΡΟΣ ΧΡΗΣΤΟΣ, 158

ΘΕΣΣΑΛΟΝΙΚΗ 2023

Περιεχόμενα

Εισαγωγή

Στόχος της παρούσας εργασίας είναι να δημιουργήσουμε ένα μοντέλο πρόβλεψης τιμών πώλησης ακινήτων ενώ ταυτόχρονα να εξετάσουμε τις σχέσεις μεταξύ των χαρακτηριστικών αυτών από τα δεδομένα που έχουμε στην διάθεση μας. Με σκοπό την καλύτερη κατανόηση των δεδομένων, θα εξετάσουμε προσεκτικά κάθε μεταβλητή ξεχωριστά και θα πραγματοποιήσουμε στατιστικούς ελέγχους για να αποφασίσουμε αν θα τις αποδεχτούμε ή θα τις απορρίψουμε.

Θεωρητική επισκόπηση - Μεθοδολογία

Σε αυτό το σημείο αναλύουμε τα δεδομένα μας. Αρχικά χωρίζουμε τις μεταβλητές μας σε δύο κατηγορίες: ποσοτικές και ποιοτικές. Στην συνέχεια, ελέγχουμε για τυχόν ελλείπουσες τιμές και αφαιρούμε σχετικές παρατηρήσεις από το σύνολο δεδομένων, εφόσον αυτές δεν αποτελούν σημαντικό ποσοστό των γνωστών τιμών. Η εξαρτημένη μας μεταβλητή είναι η **sales price** η οποία είναι ποσοτική (διακριτή).

Περιγραφική Στατιστική

Στόχος της περιγραφικής είναι η συνοπτική και αποτελεσματική παρουσίαση των χαρακτηριστικών του συνόλου δεδομένων μας. Χωρίζεται σε δυο μέρη:

Για τις **ποιοτικές μεταβλητές** θα υπολογίσουμε:

- Απόλυτες συχνότητες (frequencies)
- Σχετικές συχνότητες (relative frequencies)
- Αθροιστικές συχνότητες (cumulative frequencies)

Έπειτα για κάθε μεταβλητή θα φτιάξουμε και θα εμφανίσουμε ραβδόγραμμα.

Οι ποιοτικές μεταβλητές που διαθέτει το dataset παρουσιάζονται στον παρακάτω πίνακα. (πίνακας 1)

Variable	Details	Type
MSZoning	Ζώνη κατοικίας	ποιοτική (ονομαστική)
Street	Είδος του μπροστινού δρόμου	ποιοτική (ονομαστική)
Utilities	Παροχές	ποιοτική (ονομαστική)
OverallQual	Αξιολόγηση όλων των υλικών του εξωτερικού του ακινήτου	ποιοτική (διατάξιμη)
OverallCond	Αξιολόγηση της συνολικής κατάστασης του ακινήτου	ποιοτική (διατάξιμη)
ExterQual	Αξιολόγηση του υλικού του εξωτερικού	ποιοτική (διατάξιμη)
BsmtCond	Αξιολόγηση κατάστασης υπογείου	ποιοτική (διατάξιμη)
HeatingQC	Ποιότητα και κατάσταση θέρμανσης	ποιοτική (διατάξιμη)
CentralAir	Ύπαρξη κεντρικού κλιματισμού	ποιοτική (διατάξιμη)

Πίνακας 1: Ποιοτικές Μεταβλητές

Για τις **ποσοτικές μεταβλητές** θα υπολογίσουμε:

- Μέτρα κεντρικής τάσης
    - Μέση τιμή (Mean)
    - Διάμεσος (Median)
    - Επικρατούσα τιμή (Mode)
- Μέτρα μεταβλητότητας
    - Διασπορά (Variance)
    - Τυπική απόκλιση (Standard Deviation)
    - Εύρος (Range)
    - Ενδοτεταρτημοριακό εύρος (Interquartile Range)
- Μέτρα σχήματος κατανομής

Έπειτα για κάθε μεταβλητή θα φτιάξουμε και θα παρουσιάσουμε ιστόγραμμα κατανομής συχνοτήτων.

Οι ποσοτικές μεταβλητές που διαθέτει το dataset παρουσιάζονται στον παρακάτω πίνακα. (πίνακας 2)

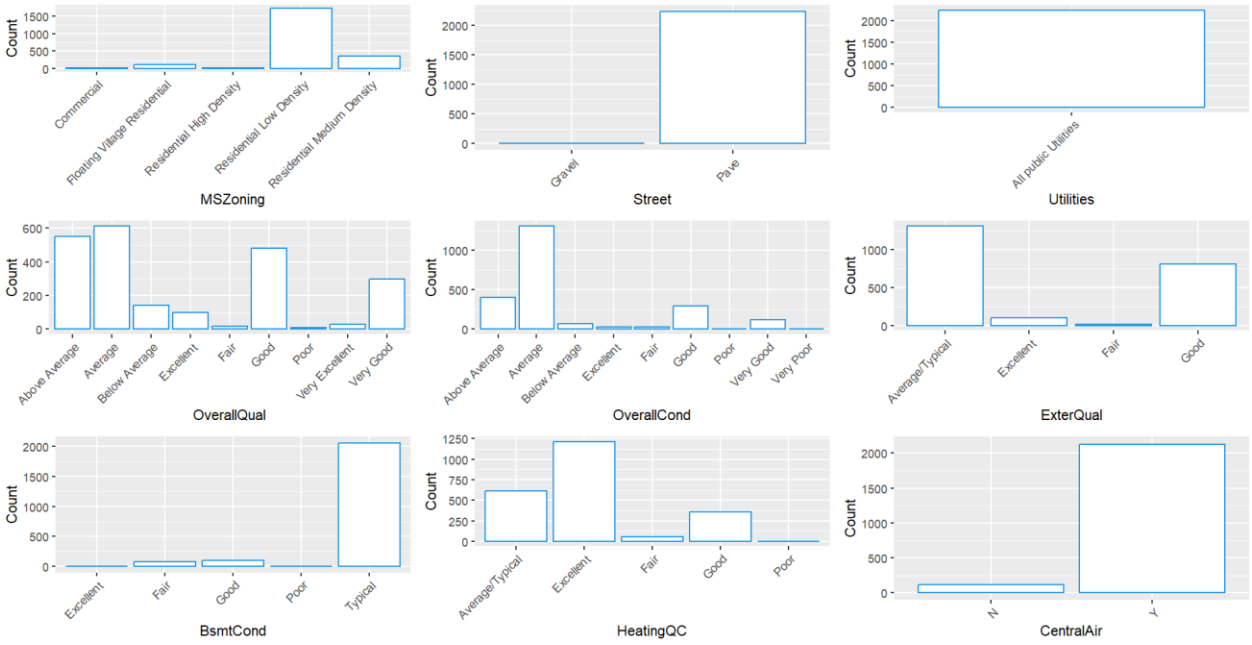
Variable	Details	Type
Order	Αύξων αριθμός οικίας	ποσοτική (διακριτή)
Lot Frontage	Έκταση δρόμου που συνδέεται με το ακίνητο	ποσοτική (διακριτή)
LotArea	Τετραγωνική επιφάνεια οικοπέδου	ποσοτική (διακριτή)
YearBuilt	Έτος κατασκευής ακινήτου	ποσοτική (διακριτή)
YearRemodAdd	Έτος ανακαίνισης	ποσοτική (διακριτή)
TotalBsmtSF	Τετραγωνική έκταση υπογείου	ποσοτική (διακριτή)
@1stFlrSF	Τετραγωνική έκταση 1ου ορόφου	ποσοτική (διακριτή)
@2ndFlrSF	Τετραγωνική έκταση 2ου ορόφου	ποσοτική (διακριτή)
BedroomAbvGr	Αριθμός υπνοδωματίων (εκτός υπογείου)	ποσοτική (διακριτή)
KitchenAbvGr	Αριθμός κουζινών	ποσοτική (διακριτή)
TotRmsAbvGrd	Αριθμός όλων των δωματίων	ποσοτική (διακριτή)
Fireplaces	Πλήθος τζακιών	ποσοτική (διακριτή)
GarageYrBlt	Έτος κατασκευής γκαράζ	ποσοτική (διακριτή)
PoolArea	Τετραγωνική έκταση πισίνας (σε ft2)	ποσοτική (διακριτή)
YrSold	Έτος πώλησης ακινήτου	ποσοτική (διακριτή)

Πίνακας 2: Ποσοτικές Μεταβλητές

Περιγραφική Στατιστική - Ποιοτικές Μεταβλητές

Για κάθε μεταβλητή εξετάζουμε όλες τις συχνότητες των χαρακτηριστικών και προσπαθούμε να εντοπίσουμε ανισορροπίες.

Παρακάτω απεικονίζονται τα γραφήματα που υλοποιήθηκαν στην εικόνα 1.



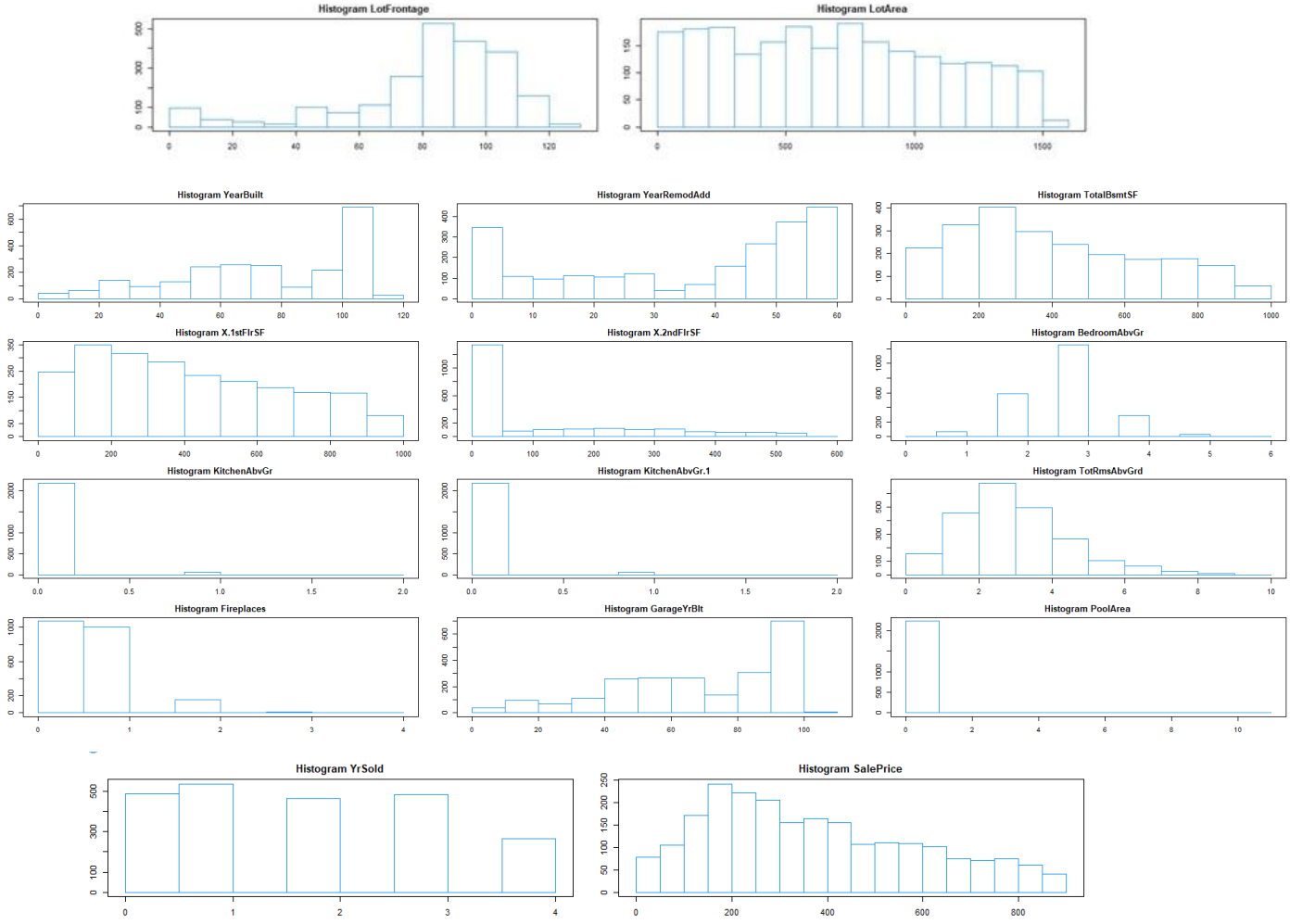
Εικόνα 1: Ραβδόγραμμα ποιοτικών μεταβλητών

Παρατηρούμε πως οι μεταβλητές “Street” και “Utilities” έχουν πολύ μεγάλη κλίση προς μια μόνο τιμή. Το ποσοστό είναι 99.65% και για τις δυο περιπτώσεις και επομένως τις κάνουμε drop.

Περιγραφική Στατιστική - Ποσοτικές Μεταβλητές

Τα παρακάνω ιστογράμματα επιβεβαιώνουν το γεγονός ότι δεν ακολουθείται για καμία μεταβλητή κανονική κατανομή και επομένως θα πρέπει να μετασχηματιστεί η εξαρτημένη μεταβλητή του dataset sales price, με την χρήση του log. Επίσης αξίζει να αναφερθεί ότι έγινε αφαίρεση των ελλείπουσων τιμών από το dataset.

Για κάθε μεταβλητή ισχύουν τα παρακάτω όπως απεικονίζονται στην εικόνα 2.



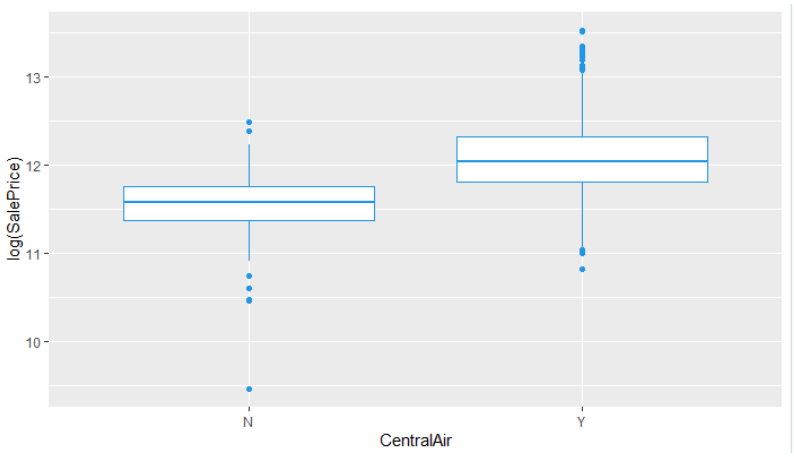
Εικόνα 2: Ιστόγραμμα για ποσοτικές μεταβλητές

Καθώς δεν ακολουθείται η κανονική κατανομή, από τα διαγράμματα μπορούμε να εξάγουμε συμπεράσματα για την λοξότητα και την κύρτωση της κάθε μεταβλητής. Επιπλέον, επιλέγουμε να υπολογίσουμε από τα μέτρα κεντρικής τάσης την διάμεσο ενώ για τα μέτρα μεταβλητότητας το ενδοτεταρτομοριακό εύρος. Τα αποτελέσματα αυτά συνοψίζονται στον παρακάτω πίνακα (πίνακας 3).

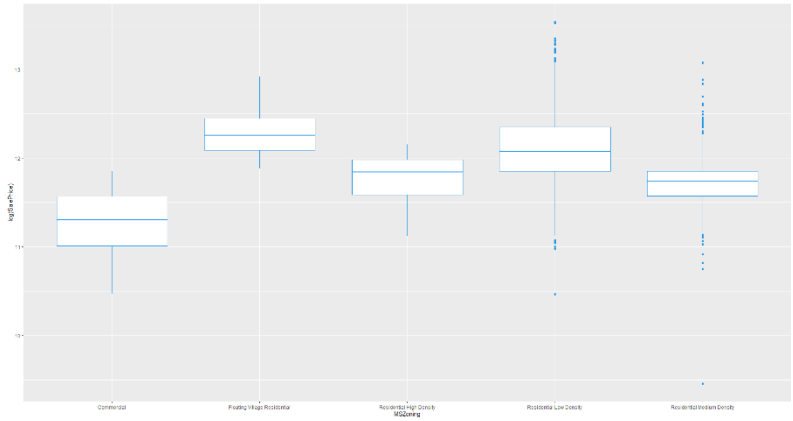
Variables	Normality	Skewness	Kurtosis	Median	Q1-Q3
Lot Frontage	Όχι	Αρνητική	Λεπτόκυρτη	69	21
LotArea	Όχι	Θετική	Πλατύκυρτη	9360	3937
YearBuilt	Όχι	Αρνητική	Πλατύκυρτη	1974	50
YearRemodAdd	Όχι	Αρνητική	Πλατύκυρτη	1994	40
TotalBsmtSF	Όχι	Θετική	Πλατύκυρτη	995	530.5
@1stFlrSF	Όχι	Θετική	Πλατύκυρτη	1089	517.5
@2ndFlrSF	Όχι	Θετική	Πλατύκυρτη	0	700.5
BedroomAbvGr	Όχι	Θετική	Λεπτόκυρτη	3	1
KitchenAbvGr	Όχι	Θετική	Λεπτόκυρτη	1	0
TotRmsAbvGrd	Όχι	Θετική	Λεπτόκυρτη	6	2
Fireplaces	Όχι	Θετική	Λεπτόκυρτη	1	1
GarageYrBlt	Όχι	Αρνητική	Πλατύκυρτη	1980	44
PoolArea	Όχι	Θετική	Λεπτόκυρτη	0	0
YrSold	Όχι	Θετική	Πλατύκυρτη	2008	2
SalesPrice	Όχι	Θετική	Πλατύκυρτη	162500	86918.5

Πίνακας 3: Περιγραφική Στατιστική ποσοτικών μεταβλητών

Διερευνητική Ανάλυση



Εικόνα 3: Boxplot για την ποιοτική μεταβλητή CentralAir



Εικόνα 4: Boxplot για την μεταβλητή MSZoning

Στα θηκογράμματα παραπάνω (εικόνα 3 και εικόνα 4) μπορούμε να καταλάβουμε την συμπεριφορά των ποιοτικών μεταβλητών μας. Για παράδειγμα στην εικόνα 3, η διάμεσος για την κατηγορία Yes είναι υψηλότερη σε σχέση με την κατηγορία No. Το ενδοτεταρτομοριακό εύρος για την κατηγορία Yes είναι πολύ μεγαλύτερο δηλαδή, υπάρχει μεγαλύτερη μεταβλητότητα στην κατανομή log(price) για τα σπίτια προς πώληση. Επίσης είναι εμφανής η ύπαρξη ακραίων σημείων και για τις δύο τιμές. Όμοια στην εικόνα 4, συμπεραίνουμε ότι η διάμεσος είναι διαφορετική για την MSZoning ενώ αντίστοιχα η κατηγορία Commercial έχει σαφώς μεγαλύτερη μεταβλητότητα από την Residential Medium Density. Τέλος, για τη κατηγορία Residential High Density η διάμεσος βρίσκεται στο άνω μέρος του ενδοτεταρτομοριακού εύρους υποδεικνύουν την ύπαρξη λοξότητας στην κατανομή, συσσώρευση στις υψηλές τιμές της κατανομής (αρνητική λοξότητα).

Ερευνητικά ερωτήματα

Η εξαρτημένη μεταβλητή sales price δεν ακολουθεί κανονική κατανομή και επομένως είναι απαραίτητος ο μετασχηματισμός της με την χρήση του λογαρίθμου log. Ο μετασχηματισμός εξασφάλισε ότι η sales price ακολουθεί κανονική κατανομή και επομένως μπορεί να γίνει χρήση χρήση του t-test για 2 ανεξάρτητους πληθυσμούς και ANOVA για περισσότερους από δυο πληθυσμούς. Παρακάτω έγινε υλοποίηση των test που αναφέρθηκαν για όλες τις ποιοτικές μας μεταβλητές.

Ανάλυση Διασποράς

1ο Ερώτημα: Υπάρχουν ενδείξεις για στατιστικά σημαντική διαφορά στις πληθυσμιακές μέσες τιμές των sales price ανάλογα με την ύπαρξη air-condition ή όχι.

Για την απάντηση του συγκεκριμένου ερωτήματος θα γίνει έλεγχος για την ισότητα των διασπορών των 2 πληθυσμών. Επομένως είναι απαραίτητος ο ορισμός της μηδενικής και της εναλλακτικής υπόθεσης. Δηλαδή,

Ho: Ισότητα των διασπορών

H1: Ανισότητα των διασπορών.

Για τον έλεγχο της μηδενικής υπόθεσης έγινε αρχικά χρήση του Levene's test για την ισότητα των διακυμάνσεων μεταξύ δύο ομάδων. Αυτό το τεστ είναι σημαντικό όταν εφαρμόζουμε στατιστικές μεθόδους που βασίζονται στην υπόθεση ότι οι διακυμάνσεις των διαφορετικών ομάδων είναι ίδιες.

Η διεξαγωγή του Levene's Test έδειξε ότι $p = 0.2236 > 0.05$ και επομένως δεν θα πρέπει να διεξαχθεί το t-test για ανεξάρτητους πληθυσμούς κάτω υπό την υπόθεση άνισων διασπορών. Άρα η μηδενική υπόθεση είναι αληθής, οπότε υπάρχει ισότητα ανάμεσα στις διασπορών των σπιτιών ανάλογα με την ύπαρξη air-condition ή όχι.

Δηλαδή, "Η μέση τιμή της τιμής του sales price που διαθέτουν air-condition δεν διαφέρει από εκείνη που δεν διαθέτουν air-condition."

Οι μεταβλητές MSZoning, OverallQual, OverallCond καθώς και οι Utilities, ExterQual, BsmtCond, HeatingQC επειδή έχουν πληθυσμούς περισσότερους από δύο θα γίνει χρήση της μεθόδου ANOVA.

Η διαδικασία αυτή πραγματοποιείται σε δυο στάδια:

- F-ratio: Ελέγχει την μηδενική υπόθεση ότι όλοι οι πληθυσμοί εμφανίζουν ίσες μέσες τιμές.
- Ζευγαρωτοί έλεγχοι (post-hoc analysis): Συγκρίνει τις μέσες τιμές των πληθυσμών κατά ζεύγη

2ο Ερώτημα: Ο τύπος του MSZoining (ζώνη κατοικίας) έχει επίδραση στις πληθυσμιακές μέσες τιμές της sales price. Άρα:

Ho: Ισότητα των μέσων τιμών των διαφόρων τιμών του MSZoning

H1: Υπάρχει τουλάχιστον ένα ζευγάρι πληθυσμών που εμφανίζουν διαφορετικές μέσες τιμές

Τα αποτελέσματα F-ratio υπέδειξαν ότι η $p < 0.001$ και άρα υπάρχει τουλάχιστον ένα ζεύγος μέσων τιμών που εμφανίζουν στατιστικά σημαντική διαφορά, οπότε είναι σημαντική η διενέργεια ζευγαρωτών συγκρίσεων. Στη συνέχεια η εφαρμογή του Levene's test επιβεβαίωσε ότι $p < 0.001$ και συμπερασματικά απορρίπτεται η μηδενική υπόθεση, επομένως θα πρέπει να επιλεχθεί post-hoc έλεγχος κάτω υπό την υπόθεση άνισων διασπορών (Dunnett T3). Τα αποτελέσματα από το Dunnett T3 υπέδειξαν τα ζεύγη εκείνα που εμφανίζουν στατιστικά σημαντικές διαφορές.

Αποτελέσματα post-hoc ανάλυσης:

- Floating Village Residential - Commercial ($p < 0.05$)
- Residential Low Density - Commercial ($p < 0.05$)
- Residential High Density – Commercial ($p < 0.05$)
- Residential Medium Density – Commercial ($p < 0.05$)

Τελική Αναφορά αποτελεσμάτων ANOVA

Η ανάλυση διασποράς (ANOVA) φανέρωσε στατιστικά σημαντική επίδραση του παράγοντα MSZoning, $F(4, 2234) = 120$, $p < 0.001$. Η post hoc ανάλυση με τον έλεγχο Dunnett T3 φανέρωσε την ύπαρξη στατιστικά σημαντικής διαφοράς μεταξύ των Floating Village Residential - Commercial , Residential Low Density – Commercial, Residential High Density – Commercial , Residential Medium Density – Commercial ζευγαριών.

Παρόμοια εφαρμόστηκε ο ίδιος έλεγχος και για τις υπόλοιπες ποιοτικές μεταβλητές. Για ευσύνοπτους λόγους στις επόμενες μεταβλητές παρουσιάζεται η τελική αναφορά αποτελεσμάτων ANOVA.

3ο Ερώτημα: Η αξιολόγηση όλων των υλικών του εξωτερικού ακινήτου OverallQual έχει επίδραση στις πληθυσμιακές μέσες τιμές της sales price. Άρα:
Ho: Ισότητα των μέσων τιμών των διαφόρων τιμών του OverallQual
H1: Υπάρχει τουλάχιστον ένα ζευγάρι πληθυσμών που εμφανίζουν διαφορετικές μέσες τιμές

Από αποτελέσματα γίνεται αποδοχή της εναλλακτικής υπόθεσης δηλαδή:
"Η αξιολόγηση όλων των υλικών του εξωτερικού ακινήτου OverallQual έχει επίδραση στις πληθυσμιακές μέσες τιμές του sales price." Τα αποτελέσματα από το Dunnett T3 υπέδειξαν τα ζεύγη εκείνα που εμφανίζουν στατιστικά σημαντικές διαφορές.

Αποτελέσματα post-hoc ανάλυσης:

- Average – Above Average (p <0.001)
- Below Average – Above Average (p <0.001)
- Excellent – Above Average (p <0.001)
- Fair – Above Average (p <0.001)
- Good – Above Average (p <0.001)
- Poor – Above Average (p <0.001)
- Very Excellent – Above Average (p <0.001)
- Very Good – Above Average (p <0.001)

Τελική Αναφορά αποτελεσμάτων ANOVA

Η ανάλυση διασποράς (ANOVA) φανέρωσε στατιστικά σημαντική επίδραση του παράγοντα OverallQual, $F(8, 2230) = 660.5$ $p< 0.001$. Η post hoc ανάλυση με τον έλεγχο Dunnett T3 φανέρωσε την ύπαρξη στατιστικά σημαντικής διαφοράς μεταξύ όλων των υπό εξέταση ζευγαριών.

4ο Ερώτημα: Η αξιολόγηση της συνολικής κατάστασης του ακινήτου έχει επίδραση στις πληθυσμιακές μέσες τιμές της sales price. Άρα:

Ho: Ισότητα των μέσων τιμών των διαφόρων τιμών του OverallCond

H1: Υπάρχει τουλάχιστον ένα ζευγάρι πληθυσμών που εμφανίζουν διαφορετικές μέσες τιμές

Από αποτελέσματα γίνεται αποδοχή της εναλλακτικής υπόθεσης δηλαδή:

"Η αξιολόγηση όλων των υλικών του εξωτερικού ακινήτου OverallCond έχει επίδραση στις πληθυσμιακές μέσες τιμές του sales price." Τα αποτελέσματα από το Dunnett T3 υπέδειξαν τα ζεύγη εκείνα που εμφανίζουν στατιστικά σημαντικές διαφορές.

Αποτελέσματα post-hoc ανάλυσης:

- Average – Above Average (p <0.001)
- Below Average – Above Average (p <0.001)
- Excellent – Above Average (p <0.001)
- Fair – Above Average (p <0.001)
- Poor – Above Average (p <0.001)
- Very Poor – Above Average (p <0.001)

Τελική Αναφορά αποτελεσμάτων ANOVA

Η ανάλυση διασποράς (ANOVA) φανέρωσε στατιστικά σημαντική επίδραση του παράγοντα OverallCond, $F(8, 2230) = 70.89$ $p< 0.001$. Η post hoc ανάλυση με τον έλεγχο Dunnett T3 φανέρωσε την ύπαρξη στατιστικά σημαντικής διαφοράς μεταξύ όλων των υπό εξέταση ζευγαριών.

5ο Ερώτημα: Η αξιολόγηση του υλικού του εξωτερικού του ακινήτου έχει επίδραση στις πληθυσμιακές μέσες τιμές της sales price. Άρα:

Ho: Ισότητα των μέσων τιμών των διαφόρων τιμών του ExterQual

H1: Υπάρχει τουλάχιστον ένα ζευγάρι πληθυσμών που εμφανίζουν διαφορετικές μέσες τιμές

Από αποτελέσματα γίνεται αποδοχή της εναλλακτικής υπόθεσης δηλαδή:

"Η αξιολόγηση όλων των υλικών του εξωτερικού ακινήτου ExterQual έχει επίδραση στις πληθυσμιακές μέσες τιμές του sales price." Τα αποτελέσματα από το Dunnett T3 υπέδειξαν τα ζεύγη εκείνα που εμφανίζουν στατιστικά σημαντικές διαφορές.

Αποτελέσματα post-hoc ανάλυσης:

- Excellent – Average/Typical (p <0.001)
- Fair – Average/Typical (p <0.001)
- Good – Average/Typical (p <0.001)

Τελική Αναφορά αποτελεσμάτων ANOVA

Η ανάλυση διασποράς (ANOVA) φανέρωσε στατιστικά σημαντική επίδραση του παράγοντα ExterQual, $F(3, 2235) = 760.5$ $p< 0.001$. Η post hoc ανάλυση με τον έλεγχο Dunnett T3 φανέρωσε την ύπαρξη στατιστικά σημαντικής διαφοράς μεταξύ των Excellent – Average/Typical, Fair – Average/Typical , Good – Average/Typical ζευγαριών.

6ο Ερώτημα: Η αξιολόγηση κατασκευής του υπογείου του ακινήτου έχει επίδραση στις πληθυσμιακές μέσες τιμές της sales price. Άρα:
Ho: Ισότητα των μέσων τιμών των διαφόρων τιμών του BsmtCond
H1: Υπάρχει τουλάχιστον ένα ζευγάρι πληθυσμών που εμφανίζουν διαφορετικές μέσες τιμές

Τα αποτελέσματα F-ratio φανερώνουν ότι υπάρχει σημαντική στατιστική διαφορά μεταξύ των μέσων τιμών των πληθυσμών, ωστόσο η διεξαγωγή του Levene's test υπέδειξε ότι δεν υπάρχει στατιστικά σημαντική διαφορά διότι η πιθανότητα δεν ήταν μικρότερη από 0.05. Θεωρήσαμε λοιπόν ότι αυτή η διαφορά υπάρχει διότι το σύνολο των δεδομένων είναι μεγάλο, οι στατιστικές δοκιμές έχουν την τάση να είναι πολύ ακριβείς και μπορεί να εντοπίζουν μικρές, ασήμαντες διαφορές που δεν θα είχαν νόημα.

Από τα αποτελέσματα λοιπόν γίνεται αποδοχή της μηδενικής υπόθεσης δηλαδή:
"Η αξιολόγηση όλων των υλικών του εξωτερικού ακινήτου BsmtCond δεν έχει επίδραση στις πληθυσμιακές μέσες τιμές του sales price."

7ο Ερώτημα: Η ποιότητα και κατάσταση θέρμανσης του ακινήτου έχει επίδραση στις πληθυσμιακές μέσες τιμές της sales price. Άρα:

Ho: Ισότητα των μέσων τιμών των διαφόρων τιμών του HeatingQC

H1: Υπάρχει τουλάχιστον ένα ζευγάρι πληθυσμών που εμφανίζουν διαφορετικές μέσες τιμές

Από τα αποτελέσματα γίνεται αποδοχή της εναλλακτικής υπόθεσης δηλαδή:

"Η αξιολόγηση όλων των υλικών του εξωτερικού ακινήτου HeatingQC έχει επίδραση στις πληθυσμιακές μέσες τιμές του sales price." Τα αποτελέσματα από το Dunnett T3 υπέδειξαν τα ζεύγη εκείνα που εμφανίζουν στατιστικά σημαντικές διαφορές.

Αποτελέσματα post-hoc ανάλυσης:

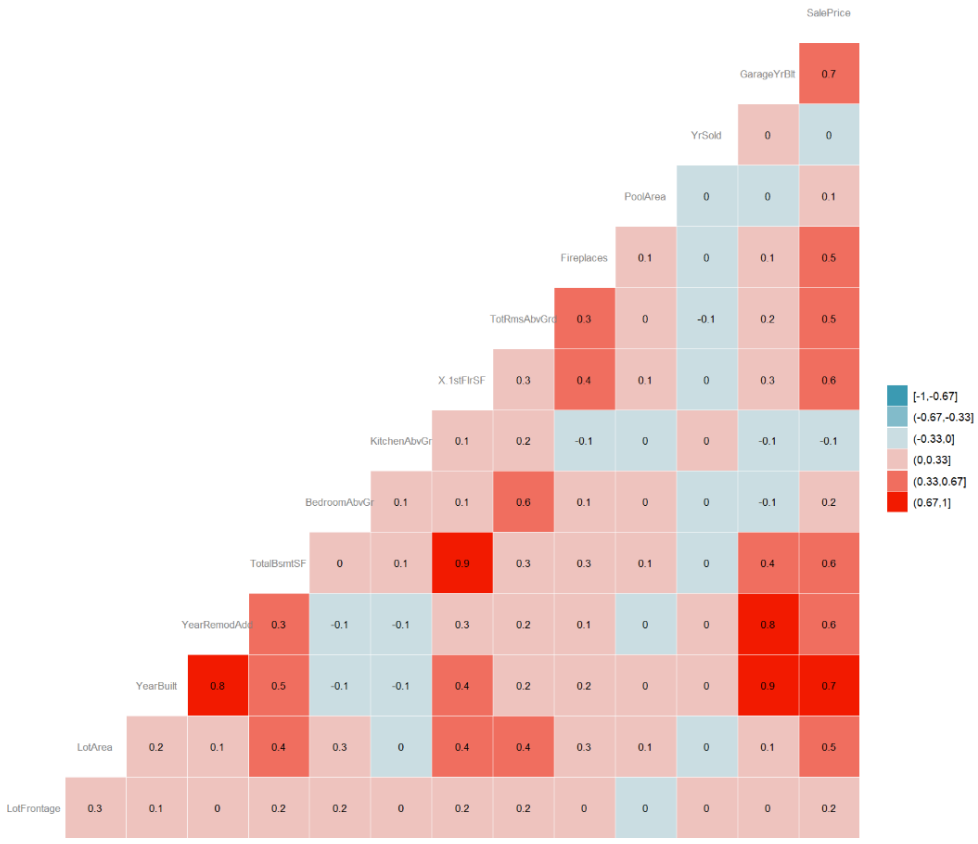
- Excellent – Average/Typical ($p < 0.001$)
- Good – Average/Typical ($p < 0.001$)

Τελική Αναφορά αποτελεσμάτων ANOVA

Η ανάλυση διασποράς (ANOVA) φανέρωσε στατιστικά σημαντική επίδραση του παράγοντα HeatingQC, $F(4, 2234) = 205.5$ $p < 0.001$. Η post hoc ανάλυση με τον έλεγχο Dunnett T3 φανέρωσε την ύπαρξη στατιστικά σημαντικής διαφοράς μεταξύ των Excellent – Average/Typical , Good – Average/Typical ζευγαριών.

Ανάλυση Συσχέτισης

Στόχος της συσχέτισης είναι η μελέτη της φύσης και της ισχύς ανάμεσα σε δύο ποσοτικές μεταβλητές. Η μελέτη συσχέτισης ανάμεσα στις μεταβλητές μπορεί να γίνει είτε με διαγράμματα διασποράς (scatter plot) είτε με την εύρεση του συντελεστή συσχέτισης r. Ανάλογα με το εάν οι μεταβλητές ακολουθούν κανονική ή όχι κατανομή, το είδος της σχέσης των μεταβλητών και το είδος τους θα γίνει και η επιλογή του κατάλληλο συντελεστή συσχέτισης καθώς υπάρχει ο Pearson και ο Spearman. Όπως αναφέρθηκε και παραπάνω οι ποσοτικές μεταβλητές μας δεν ακολουθούν κανονική κατανομή και άρα η ανάλυση συσχέτισης έγινε με την μέθοδο spearman. Παρακάτω η εικόνα 5 απεικονίζει την εφαρμογή συσχέτισης για τις ποσοτικές μεταβλητές ενώ παρουσιάζεται και ο αντίστοιχος πίνακας (πίνακας 4) από την ανάλυση καθώς και τα συμπεράσματα που προκύπτουν.



Εικόνα 5: Ανάλυση Συσχέτισης μεταξύ των μεταβλητών

Variables	Sales Price (Correlation)	Φύση	Ισχύς
YearBuilt	0,696	Θετική	Υψηλή
GarageYrBlt	0,660	Θετική	Υψηλή
YearRemodAdd	0,636	Θετική	Υψηλή
1stFlrSF	0,615	Θετική	Υψηλή
TotalBsmtSF	0,611	Θετική	Υψηλή
TotRmsAbvGrd	0,534	Θετική	Μέτρια
Fireplaces	0,497	Θετική	Μέτρια
LotArea	0,453	Θετική	Μέτρια
BedroomAbvGr	0,189	Θετική	Πολύ χαμηλή
Lot Frontage	0,182	Θετική	Πολύ χαμηλή
PoolArea	0,06	Θετική	Πολύ χαμηλή
YrSold	-0,016	-	Καμία συσχέτιση
KitchenAbvGr	-0,101	-	Καμία συσχέτιση

Πίνακας 4: Ανάλυση Συσχέτισης

Μοντέλο πρόβλεψης

Το τελικό μοντέλο πρόβλεψης που δημιουργήθηκε περιέχει τις παρακάτω μεταβλητές YearRemodAdd , YearBuilt, 1stFlrSF, GarageYrBlt, TotRmsAbvGrd, MSZoning, OverallQual, OverallCond και θεωρείται το πιο κατάλληλο για την πρόβλεψη της sales price.

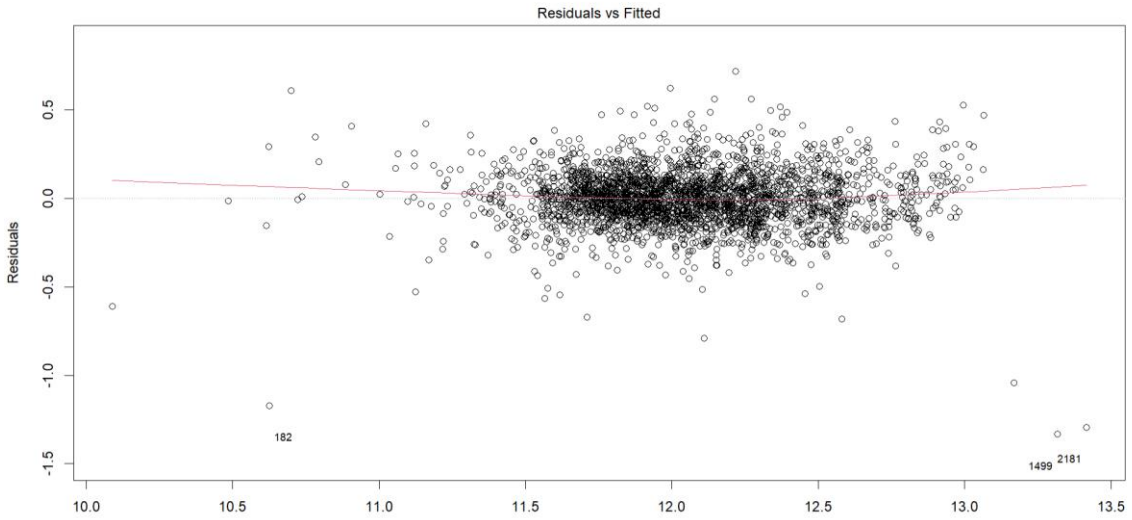
Παρακάνω απεικονίζονται οι έλεγχοι που έγιναν προκειμένου να καταλήξουμε στο τελικό μοντέλο πρόβλεψης. Παρατηρούμε ότι έγινε προσθήκη μίας-μίας μεταβλητής και συγκρίθηκε το εκάστοτε μοντέλο με το προηγούμενο, με την χρήσης της ANOVA. Η προσθήκη όλων των μεταβλητών εξηγούν όλο και περισσότερο το ποσοστό μεταβλητότητας της εξαρτημένης μας μεταβλητής. Για παράδειγμα στον πίνακα 5, που παρουσιάζεται παρακάτω, το μοντέλο 5 εξηγεί το 65,65% της μεταβλητότητας της εξαρτημένης μεταβλητής log(price), δηλαδή αύξηση σε σχέση με το μοντέλο 4, το οποίο εξηγεί το 65,45% της μεταβλητότητας. Ωστόσο, η προσθήκη του BsmtCond αύξησε το ποσοστό R-squared αλλά συγκριτικά με το προηγούμενο μοντέλο οι συντελεστές του δεν ήταν στατιστικά σημαντικοί και έτσι καταλήγουμε στο model 9 ύστερα από την σύγκριση των δύο τελευταίων μοντέλων .

Number	Model	Multiple R - squared	Adjusted R - squared	ANOVA
1	log(YearBuilt)	0,3882	0,3879	
2	+ log(GarageYrBlt)	0,4076	0,407	p < 0,001
3	+ log(YearRemodAdd)	0,4615	0,4608	p < 0,001
4	+ log(`@1stFlrSF`)	0,6546	0,654	p < 0,001
5	+ log(TotalBsmtSF)	0,6565	0,6557	p < 0,001
6	+ log(TotRmsAbvGrd)	0,7324	0,7317	p < 0,001
7	+ MSZoning	0,7389	0,7377	p < 0,001
8	+ OverallQual	0,8304	0,829	p < 0,001
9	+ OverallCond	0,8432	0,8414	p < 0,001
10	+ BsmtCond	0,8435	0,8414	p > 0,05

Πίνακας 5: Έλεγχος Μοντέλων Πρόβλεψης

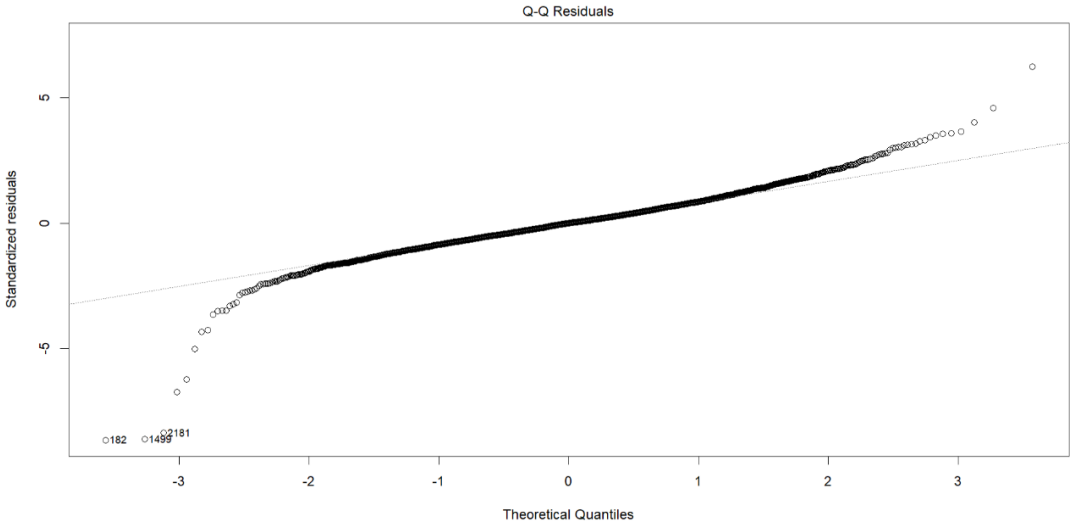
Στη συνέχεια παρουσιάζονται διαγνωστικά γραφήματα για τον έλεγχο γραμμικότητας, κανονικότητας και ομοσκεδαστικότητας.

Υπόθεση Γραμμικότητας: Διάγραμμα καταλοίπων vs. προσαρμοσμένων τιμών (residuals vs. fitted values) (εικόνα 6). Είναι φανερό ότι η οριζόντια γραμμή χωρίς διακριτά πρότυπα είναι ένδειξη της ύπαρξης γραμμικής σχέσης.



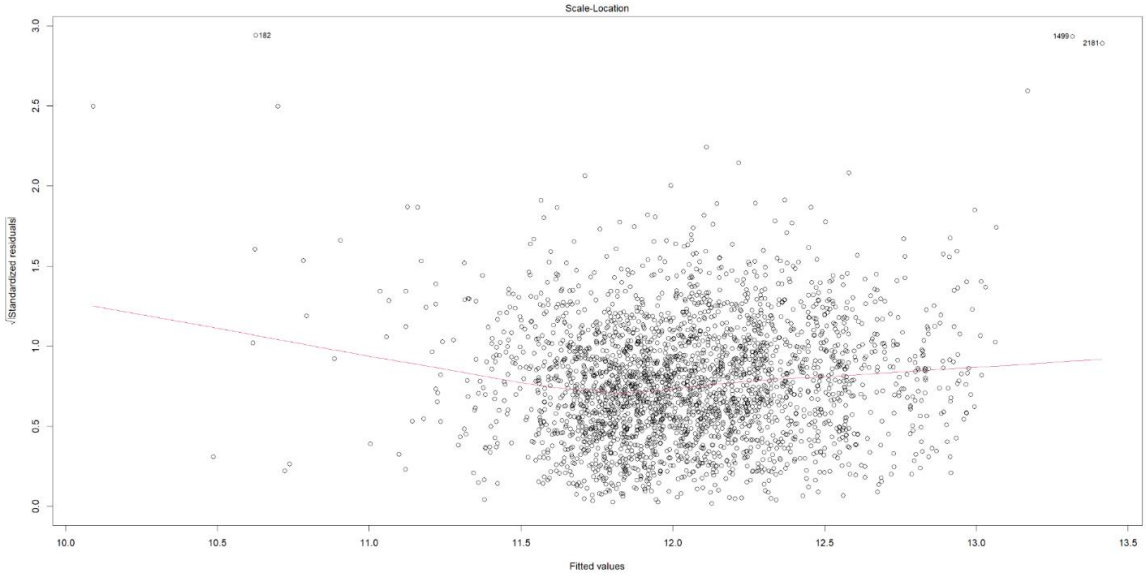
Εικόνα 6: Υπόθεση γραμμικότητας

Υπόθεση Κανονικής Κατανομής Καταλοίπων: Ελέγχεται η υπόθεση κανονικής κατανομής για τα κατάλοιπα (εικόνα 7), η οποία ικανοποιείται αφού τα περισσότερα σημεία βρίσκονται κοντά στην ευθεία.



Εικόνα 7: Έλεγχος κανονικής κατανομής

Υπόθεση Ομοσκεδαστικότητας: Διάγραμμα scale location, ελέγχεται η υπόθεση της ομοσκεδαστικότητας των καταλοίπων, (εικόνα 8) τα οποία διασπείρονται με παρόμοιο τρόπο σε όλο το εύρος τιμών.



Εικόνα 8: Έλεγχος ομοσκεδαστικότητας

Απόδοση πρόβλεψης

Χρησιμοποιήσαμε το τελικό μοντέλο πρόβλεψης για να δημιουργήσουμε το LR το οποίο συγκρίναμε με το KNN (για K = 20). Η επιλογή του K = 20 έγινε ύστερα από δοκιμές με μικρότερα καθώς και με μεγαλύτερα K. Ωστόσο για K = 20 παρατηρήθηκε ότι ήταν το πιο βέλτιστο. Το σύνολο δεδομένων χωρίστηκε σε δύο υποσύνολα. Σύνολο εκπαίδευσης (2/3 σε μέγεθος)για προσαρμογή μοντέλου και σύνολο ελέγχου (1/3 σε μέγεθος) για επικύρωση της προβλεπτικής ικανότητας του μοντέλου. Στον πίνακα 6 παρουσιάζονται τα αποτελέσματα από αυτή τη σύγκριση.

Loss Function	**LR**	**KNN**
Mean Error (ME)	3216.995	6209.986
Median Error (MdE)	890.2915	4000
Mean Absolute Error (MAE)	22204.52	40327.99
Median Absolute Error (MdAE)	14335.28	25000
Mean Magnitude of Relative Error (MMRE)	0.117%	0.22%
Median Magnitude of Relative Error (MdMRE)	0.088%	0.15%
Mean Magnitude of Relative Error to the Estimate (MMER)	0.116%	0.223%
Median Magnitude of Relative Error to the Estimate (MdMER)	0.086%	0.15%

Πίνακας 6: Σύγκριση Μοντέλων Πρόβλεψης

Αυτές οι μετρήσεις συλλογικά υποδηλώνουν ότι, κατά μέσο όρο, το μοντέλο Γραμμικής παλινδρόμησης έχει μικρότερο μέγεθος σφαλμάτων και σχετικών σφαλμάτων σε σύγκριση με το μοντέλο K-Nearest Neighbors στα δεδομένα δοκιμής, οπότε προτιμάτε το μοντέλο γραμμικής παλινδρόμησης.

Σύγκριση Μοντέλων

Παρακάτω γίνεται διεξαγωγή wilcox-test για τα δυο μοντέλα προκειμένου να διαπιστωθεί αν υπάρχει διαφορά στις κατανομές. Από τα παρακάτω αποτελέσματα παρατηρούμε ότι εκτός από την κατανομή του Error σε όλες τις υπόλοιπες κατανομές Absolute Error, Magnitude of Relative Error και Magnitude of Relative Error to the Estimate παρατηρείται σημαντική στατιστική διαφορά στις κατανομές. ($p < 0.001$) (πίνακας 7).

Wilcox Test	**Results**
Error	$p > 0.05$
Absolute Error	$p < 0.001$
Magnitude of Relative Error	$p < 0.001$
Magnitude of Relative Error to the Estimate	$p < 0.001$

Πίνακας 7: Αποτελέσματα Wilcox Test

Συμπεράσματα

Από την παραπάνω ανάλυση μας κατανοήσαμε την συμπεριφορά τόσο των ποιοτικών όσο και των ποσοτικών μεταβλητών. Στηριζόμενοι σε αυτά τα αποτελέσματα εφαρμόσαμε για τις ποιοτικές μεταβλητές τα κατάλληλα ερευνητικά ερωτήματα είτε με την χρήση t-test είτε με Anova. Στην συνέχεια για τις ποσοτικές μεταβλητές έγινε ανάλυση συσχέτισης με την εξαρτημένη μεταβλητή, γεγονός που μας βοήθησε στην δημιουργία του τελικού μοντέλου πρόβλεψης, στο οποίο έγινε πρώτα προεπεξεργασία των δεδομένων (αφαίρεση ελλείπουσων τιμών). Χωρίσαμε τα δεδομένα μας σε Train και Test με σκοπό την εκπαίδευση των μοντέλων LR, KNN και τέλος συγκρίναμε τα αποτελέσματα τους. Από τα αποτελέσματα αυτά συμπεράναμε ότι το μοντέλο LR είναι καλύτερο και το πιο βέλτιστο.