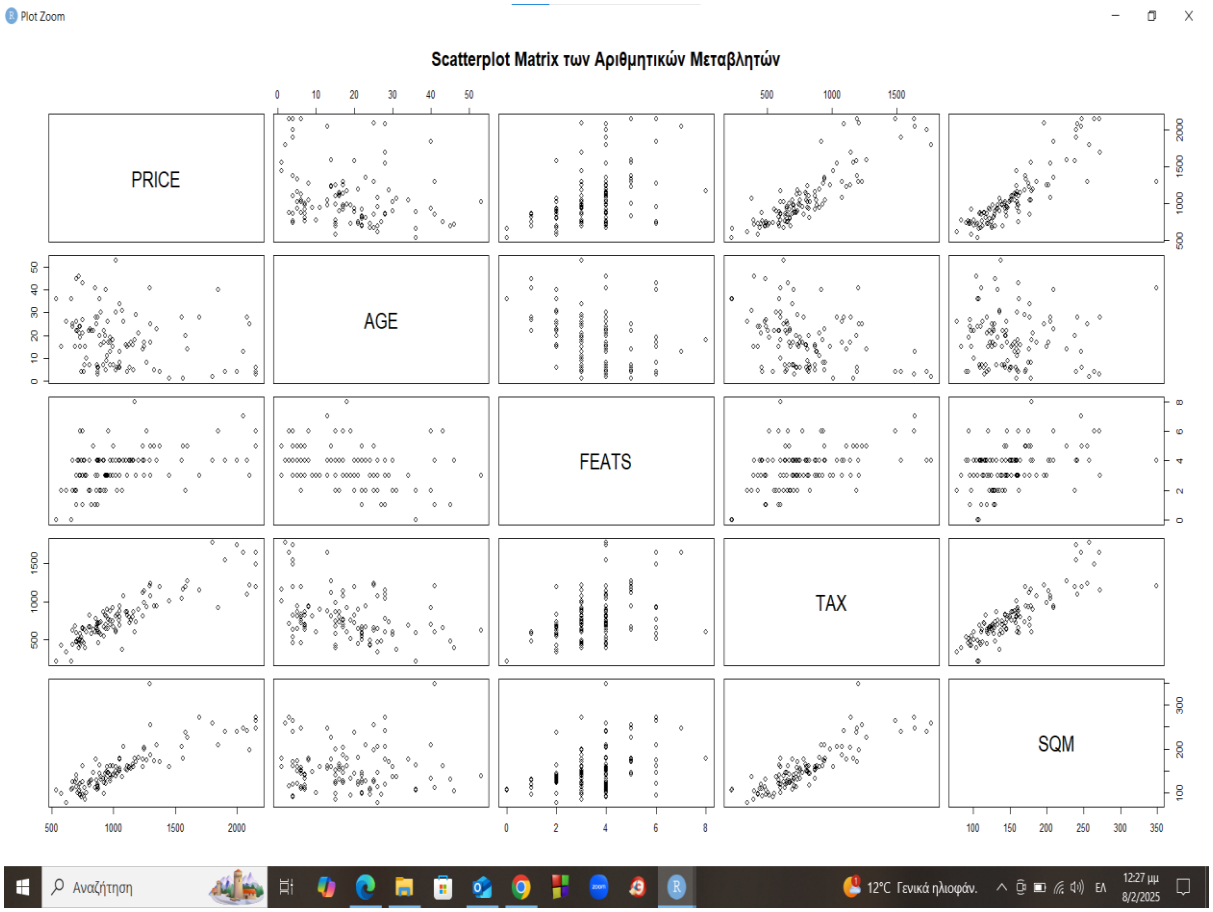
Task 2: Descriptive statistics and visualization

b)



Μέσα στο Scatterplot Matrix , κάθε ζεύγος αριθμητικών μεταβλητών συγκρίνεται μεταξύ τους. Κάποια σημεία παρατήρησης :

1. **PRICE vs. SQM** :

Αναμένουμε θετική συσχέτιση -> Όσο αυξάνεται το εμβαδόν(SQM), αυξάνεται και η τιμή του σπιτιού (PRICE) .

2. **AGE vs. PRICE**:

Πιθανώς αρνητική συσχέτιση -> Παλαιότερα σπίτια μπορεί να έχουν χαμηλότερη αξία, όσο αναφορά τη τιμή (PRICE)

3. **TAX vs. PRICE**:

Πιθανή θετική συσχέτιση -> Υψηλότερη Φορολογία μπορεί να σχετίζεται με πιο ακριβά σπίτια.

Συμπέρασμα

- Οι ισχυρότερες σχέσεις είναι **PRICE ~ SQM** και **PRICE ~ TAX**, που είναι θετικές.
- Η AGE έχει αρνητική συσχέτιση με την τιμή PRICE
- Οι υπόλοιπες μεταβλητές φαίνεται να μην έχουν τόσο ισχυρή γραμμική σχέση

Πίνακας Συσχέτισης

```
> cor(numeric_vars, use = "complete.obs")
           PRICE          AGE      FEATS        TAX         SQM
PRICE  1.0000000 -0.23598240  0.4453661  0.8746898  0.84300862
AGE   -0.2359824  1.00000000 -0.3550884 -0.3643916 -0.04818656
FEATS  0.4453661 -0.35508839  1.0000000  0.4446787  0.39850888
TAX    0.8746898 -0.36439163  0.4446787  1.0000000  0.85883201
SQM    0.8430086 -0.04818656  0.3985089  0.8588320  1.00000000
```

Ανάλυση των αποτελεσμάτων

1. **PRICE & SQM (0.84300862) -> Ισχυρή θετική συσχέτιση**

   Όσο αυξάνονται τα τετραγωνικά μέτρα, αυξάνεται και η τιμή.

2. **PRICE & TAX (0.8746898) -> Ισχυρή θετική συσχέτιση**

   Οι φόροι είναι πιο υψηλοί στα ακριβότερα ακίνητα.

3. **TAX & SQM (0.85883201) -> Ισχυρή θετική συσχέτιση**
   Όσο αυξάνονται τα τετραγωνικά μέτρα, αυξάνεται και η φορολογία.

4. **PRICE & AGE (-0.23598240) -> Αρνητική συσχέτιση**

Τα παλαιότερα σπίτια τείνουν να έχουν χαμηλότερη τιμή (αλλά η συσχέτιση είναι μέτρια).

5. **AGE & FEATS (-0.3550884) -> Αρνητική συσχέτιση**

Τα νεότερα σπίτια τείνουν να έχουν περισσότερα χαρακτηριστικά.

6. **AGE & TAX (-0.3643916) -> Αρνητική συσχέτιση**

Τα παλαιότερα σπίτια τείνουν να έχουν χαμηλότερη φορολογία.

7. **Αδύναμες ή μη σημαντικές συσχετίσεις:**

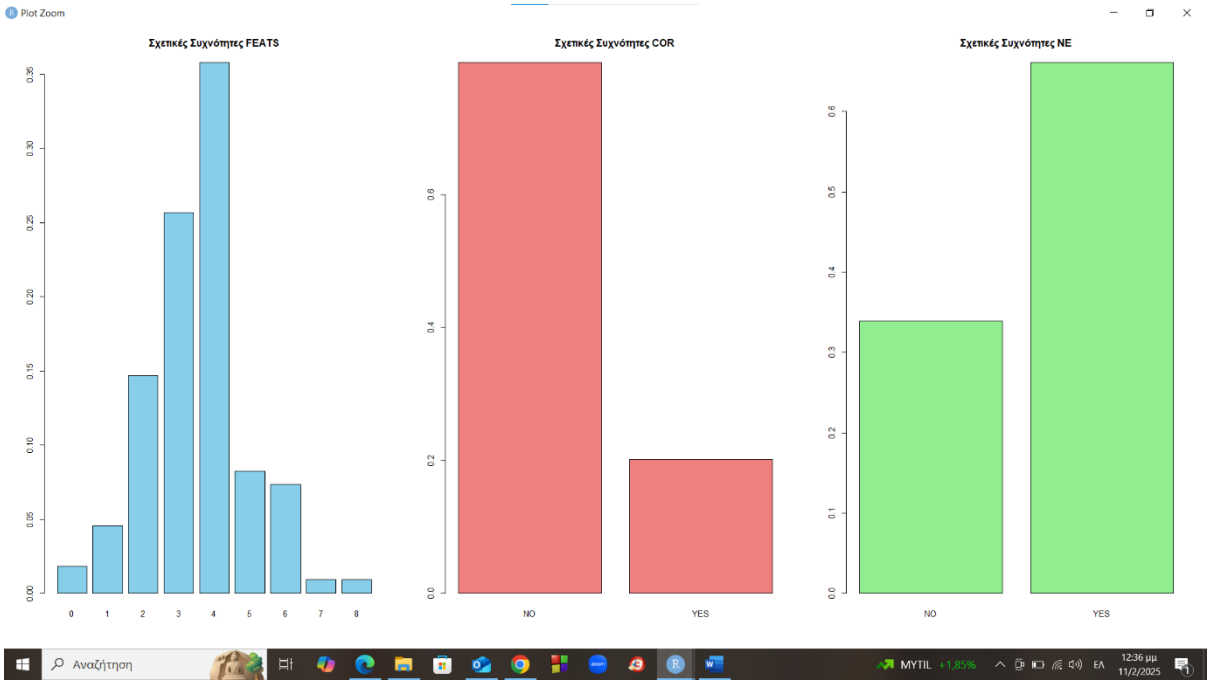
AGE & SQM (-0.04818656) -> Αδύναμη συσχέτιση

Πολύ χαμηλή συσχέτιση, που σημαίνει ότι η ηλικία του σπιτιού δεν επηρεάζει πολύ τα τετραγωνικά.

Συμπέρασμα

- **Η τιμή ενός σπιτιού εξαρτάται κυρίως από τα τετραγωνικά (SQM) και τη φορολογία (TAX).**
- **Τα νεότερα σπίτια έχουν περισσότερα χαρακτηριστικά (FEATS) και τείνουν να είναι πιο ακριβά.**
- **Η φορολογία επηρεάζεται έντονα από το μέγεθος και την τιμή του ακινήτου.**
- **Η ηλικία του σπιτιού έχει σχετικά μικρή επίδραση στην τιμή και τα τετραγωνικά.**

c)



Ανάλυση του πολλαπλού διαγράμματος (barplot σχετικών συχνοτήτων)

Το διάγραμμα περιλαμβάνει τρεις κατανομές σχετικών συχνοτήτων για διαφορετικές μεταβλητές: **FEATS, COR, NE**.

1. Αριστερό γράφημα (Γαλάζιο) - Σχετικές Συχνότητες FEATS

- Η μεταβλητή **FEATS** φαίνεται να παίρνει διακριτές τιμές (0, 1, 2, ..., 8).
- Παρατηρούμε ότι οι τιμές **3 και 4 έχουν τη μεγαλύτερη συχνότητα**, κάτι που σημαίνει ότι τα περισσότερα δείγματα έχουν 3-4 χαρακτηριστικά.
- Οι τιμές 0, 7, 8 εμφανίζονται σπάνια.

2. Κεντρικό γράφημα (Κόκκινο) - Σχετικές Συχνότητες COR

- Η μεταβλητή **COR** είναι κατηγορική, με δύο τιμές: **NO & YES**.
- Το **NO (όχι)** έχει πολύ μεγαλύτερη σχετική συχνότητα από το **YES (ναι)**, δηλαδή η πλειοψηφία των παρατηρήσεων ανήκει στην κατηγορία **NO**.

3. Δεξιό γράφημα (Πράσινο) - Σχετικές Συχνότητες NE

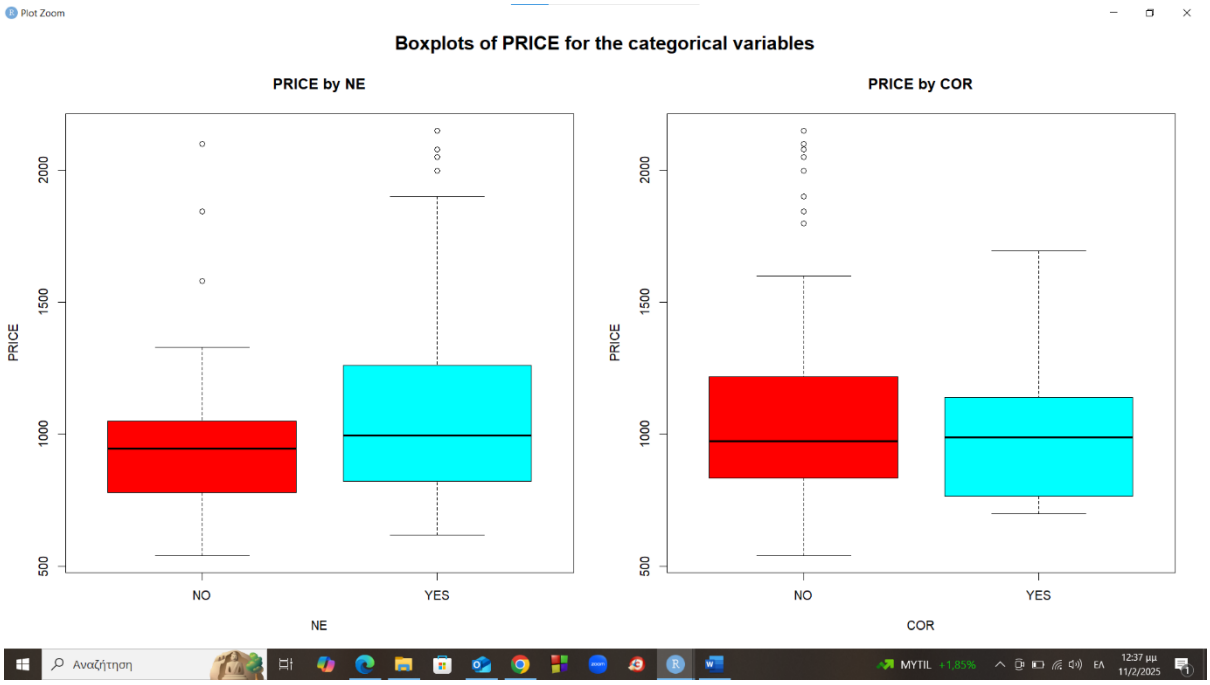
- Όπως και η μεταβλητή **COR**, έτσι και η **NE** έχει δύο κατηγορίες: **NO & YES**.
- Σε αντίθεση με το **COR**, εδώ η κατηγορία **YES (ναι)** εμφανίζεται συχνότερα από το **NO (όχι)**.

- Δηλαδή, η πλειοψηφία των περιπτώσεων κατατάσσεται στη θετική κατηγορία.

Συμπεράσματα

FEATS: Τα περισσότερα δείγματα έχουν 3-4 χαρακτηριστικά.
COR: Οι περισσότερες περιπτώσεις ανήκουν στην κατηγορία **NO**.
NE: Οι περισσότερες περιπτώσεις ανήκουν στην κατηγορία **YES**.

d)



Ανάλυση των Boxplots της μεταβλητής PRICE ανά κατηγορία

Το διάγραμμα δείχνει δύο boxplots για την **PRICE (Τιμή)**, ομαδοποιημένη με βάση δύο κατηγορικές μεταβλητές: **NE και COR**.

---

1. Boxplot αριστερά (PRICE by NE)

- **NE ("NO" vs. "YES")**:
    - Οι ιδιοκτησίες που ανήκουν στην κατηγορία **"YES" (γαλάζιο boxplot)** φαίνεται να έχουν **υψηλότερη διάμεσο (median)** σε σχέση με τα **"NO" (κόκκινο boxplot)**.
    - Το εύρος των τιμών είναι μεγαλύτερο για την κατηγορία **"YES"**, με ορισμένες πολύ υψηλές τιμές (outliers).
    - Οι τιμές της **"NO"** κατηγορίας είναι πιο συγκεντρωμένες γύρω από τη διάμεσο και έχουν μικρότερο spread.

Συμπέρασμα: Οι κατοικίες που ανήκουν στην κατηγορία **"YES"** της μεταβλητής **NE** φαίνεται να έχουν γενικά υψηλότερες τιμές, αλλά και μεγαλύτερη διακύμανση.

---

2. Boxplot δεξιά (PRICE by COR)

- **COR ("NO" vs. "YES")**:
    - Οι ιδιοκτησίες με **"NO" (κόκκινο boxplot)** και **"YES" (γαλάζιο boxplot)** έχουν παρόμοιες κατανομές στην τιμή, με ελαφρώς υψηλότερη διάμεσο στην κατηγορία **"NO"**.
    - Υπάρχουν **αρκετά outliers** και στις δύο περιπτώσεις, δείχνοντας ότι ορισμένες ιδιοκτησίες είναι πολύ πιο ακριβές από τον μέσο όρο.
    - Το spread (εύρος τιμών) φαίνεται παρόμοιο, αλλά η κατηγορία **"NO"** φαίνεται να έχει ελαφρώς περισσότερη μεταβλητότητα.

Συμπέρασμα: Η μεταβλητή **COR** δε φαίνεται να έχει σημαντική επίδραση στις τιμές των κατοικιών, καθώς οι κατανομές είναι παρόμοιες.

Γενική Ερμηνεία

Η μεταβλητή NE φαίνεται να σχετίζεται περισσότερο με την τιμή των κατοικιών, σε σύγκριση με τη μεταβλητή COR.
Οι κατοικίες που ανήκουν στην κατηγορία "YES" της NE μεταβλητής τείνουν να είναι ακριβότερες.
Υπάρχουν αρκετά outliers, κάτι που δείχνει ότι κάποιες ιδιοκτησίες είναι πολύ πιο ακριβές από τις υπόλοιπες.

e)

Η έξοδος για το t.test(PRICE ~ NE, data = dataproject) είναι p-value=0.07754 που είναι μεγαλύτερο από 0.05, αυτό σημαίνει ότι **δεν υπάρχει στατιστικά σημαντική διαφορά**, στη μέση τιμή της PRICE μεταξύ των κατηγοριών NE=YES και NE=NO, με επίπεδο σημαντικότητας 0.05. Το p-value = 0.07754 είναι μεγαλύτερο από 0.05, άρα δεν απορρίπτουμε την υπόθεση ότι οι μέσες τιμές είναι ίδιες. Αυτό σημαίνει ότι η γεωγραφική τοποθεσία (NE) δεν φαίνεται να επηρεάζει σημαντικά την τιμή του σπιτιού.

Η έξοδος για το t.test(PRICE ~ COR, data = dataproject) είναι p-value=0.1815 που είναι μεγαλύτερο από 0.05, αυτό σημαίνει ότι **δεν υπάρχει στατιστικά σημαντική διαφορά,** στις μέσες τιμές της PRICE μεταξύ των σπιτιών που βρίσκονται σε γωνιακά (COR=YES) και μη γωνιακά (COR=NO) οικόπεδα. Το p-value=0.1815 είναι πολύ μεγαλύτερο από 0.05, άρα δεν μπορούμε να απορρίψουμε την υπόθεση ότι η μέση τιμή της PRICE είναι ίδια για τα γωνιακά και μη γωνιακά σπίτια. Αυτό σημαίνει ότι η γωνιακή τοποθεσία δεν φαίνεται να επηρεάζει σημαντικά την τιμή του.

Σύγκριση με το NE

- Και τα δύο p-values (NE:0.07754, COR:0.1815) είναι μη στατιστικά σημαντικά στο επίπεδο 0.05.
- Το NE είχε μία οριακή τάση (p=0.07754), που σημαίνει ότι **ίσως** υπάρχει μία μικρή επίδραση της τοποθεσίας, αλλά όχι αρκετά ισχυρή.
- Το COR είναι ακόμα πιο ασθενές (p=0.1815), άρα μάλλον δεν παίζει ρόλο στη τιμή.

Συμπέρασμα:

1. Η τοποθεσία (NE) μπορεί να έχει μικρή επίδραση στην τιμή, αλλά όχι αρκετά ισχυρή για να θεωρηθεί στατιστικά σημαντική.
2. Η γωνιακή τοποθεσία (COR) δεν επηρεάζει σημαντικά την τιμή του σπιτιού.
3. Πιθανόν άλλες μεταβλητές (όπως SQM και TAX) να είναι οι βασικοί καθοριστικοί παράγοντες της τιμής.

f)

1. Βρίσκουμε τις μεταβλητές που έχουν υψηλή ασυμμετρία (skewness>1) .
2. Ελέγχουμε αν ακολουθούν κανονική κατανομή με Shapiro-Wilk test.

3. Αν η κατανομή δεν είναι κανονική, εφαρμόζουμε λογαριθμικό μετασχηματισμό (log()).
4. Ξανακάνουμε το τεστ κανονικότητας για να δούμε αν ο μετασχηματισμός βελτίωσε την κατανομή.

Αντικαθιστούμε τις μεταβλητές στο dataset με τις λογαριθμικές τιμές, προσθέτοντας το πρόθεμα “log” στο όνομά τους.

Συμπέρασμα:

1. Βρήκαμε ποιες μεταβλητές είχαν skewness>1
2. Κάναμε Shapiro-Wilk test για να δούμε αν ήταν κανονικές.
3. Εφαρμόσαμε λογαριθμικό μετασχηματισμό όπου χρειάστηκε.
4. Ξανακάναμε Shapiro-Wilk test για να δούμε αν η κατανομή βελτιώθηκε.
5. Αντικαταστήσαμε τις αρχικές μεταβλητές με τις λογαριθμικές.

Ερμηνεία των p-values

Μεταβλητή	p-value	Συμπέρασμα
PRICE	7.206128e-09 (≈ **0.0000000072**)	**Απορρίπτουμε την κανονικότητα** (η κατανομή δεν είναι κανονική)
TAX	2.087056e-05 (≈ **0.00002**)	**Απορρίπτουμε την κανονικότητα** (η κατανομή δεν είναι κανονική)
SQM	3.257518e-06 (≈ **0.0000032**)	**Απορρίπτουμε την κανονικότητα** (η κατανομή δεν είναι κανονική)

Όλα τα p-values είναι < 0.05, επομένως οι μεταβλητές **δεν ακολουθούν κανονική κατανομή** και χρειάζονται **μετασχηματισμό**.

Συμπέρασμα

Πριν τον μετασχηματισμό, οι PRICE, TAX, SQM δεν ήταν κανονικές.

Μετά τον λογαριθμικό μετασχηματισμό, πρέπει να ελέγξουμε αν οι νέες logPRICE, logTAX, logSQM έχουν γίνει κανονικές.

Αν το Shapiro-Wilk test επιστρέψει p-value > 0.05, σημαίνει ότι ο μετασχηματισμός βελτίωσε την κατανομή!

Τα νέα p-values από το **Shapiro-Wilk test** μετά τον λογαριθμικό μετασχηματισμό είναι:

Μεταβλητή	**p-value**	**Συμπέρασμα**
logPRICE	0.000589878 (**< 0.05**)	**Εξακολουθεί να μην είναι κανονική**
logTAX	0.127297935 (**> 0.05**)	**Τώρα είναι κανονική**
logSQM	0.187687054 (**> 0.05**)	Τώρα είναι κανονική

Τελική Ανάλυση

- **Η logTAX και η logSQM βελτιώθηκαν και τώρα ακολουθούν κανονική κατανομή .**
- **Η logPRICE εξακολουθεί να μην είναι κανονική (p < 0.05), αλλά η κατανομή της ίσως έχει βελτιωθεί.**

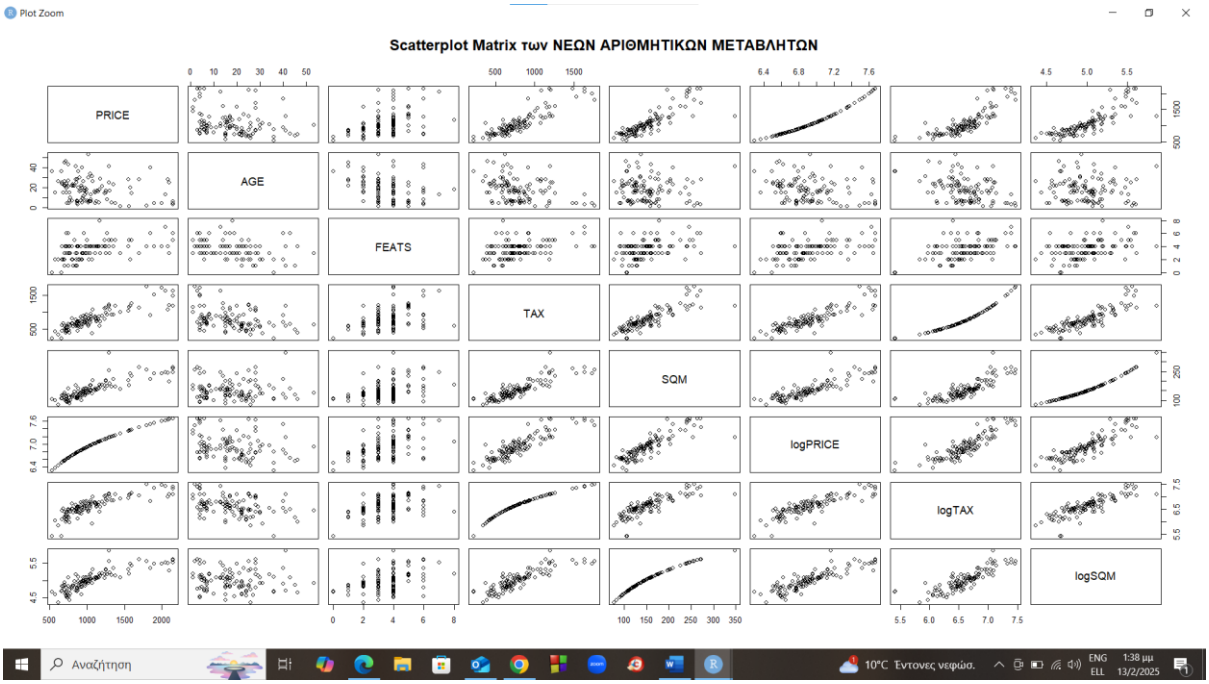
Συμπέρασμα

Ο μετασχηματισμός βοήθησε τις TAX και SQM, αλλά όχι απόλυτα την PRICE.

Οι logTAX και logSQM είναι πλέον κανονικές, άρα μπορούμε να τις χρησιμοποιήσουμε σε στατιστικές αναλύσεις που απαιτούν κανονικότητα.

Η logPRICE βελτιώθηκε, αλλά όχι τέλεια – ίσως να χρειάζεται άλλος μετασχηματισμός.

g)



Γενικές Παρατηρήσεις:

Συσχέτιση μεταξύ μεταβλητών:

- Βλέπουμε διάφορα **μοτίβα σχέσεων**, από γραμμικές έως πιο καμπυλωτές.
- Οι μεταβλητές **TAX, SQM και log-μετασχηματισμένες μεταβλητές** φαίνεται να έχουν **ισχυρές θετικές σχέσεις** με την PRICE.
- Η **μεταβλητή AGE (ηλικία ακινήτου)** δείχνει **αρνητική συσχέτιση** με την PRICE, όσο μεγαλύτερο είναι το ακίνητο, τόσο χαμηλότερη τείνει να είναι η τιμή του.

Log-μετασχηματισμένες μεταβλητές (logPRICE, logTAX, logSQM):

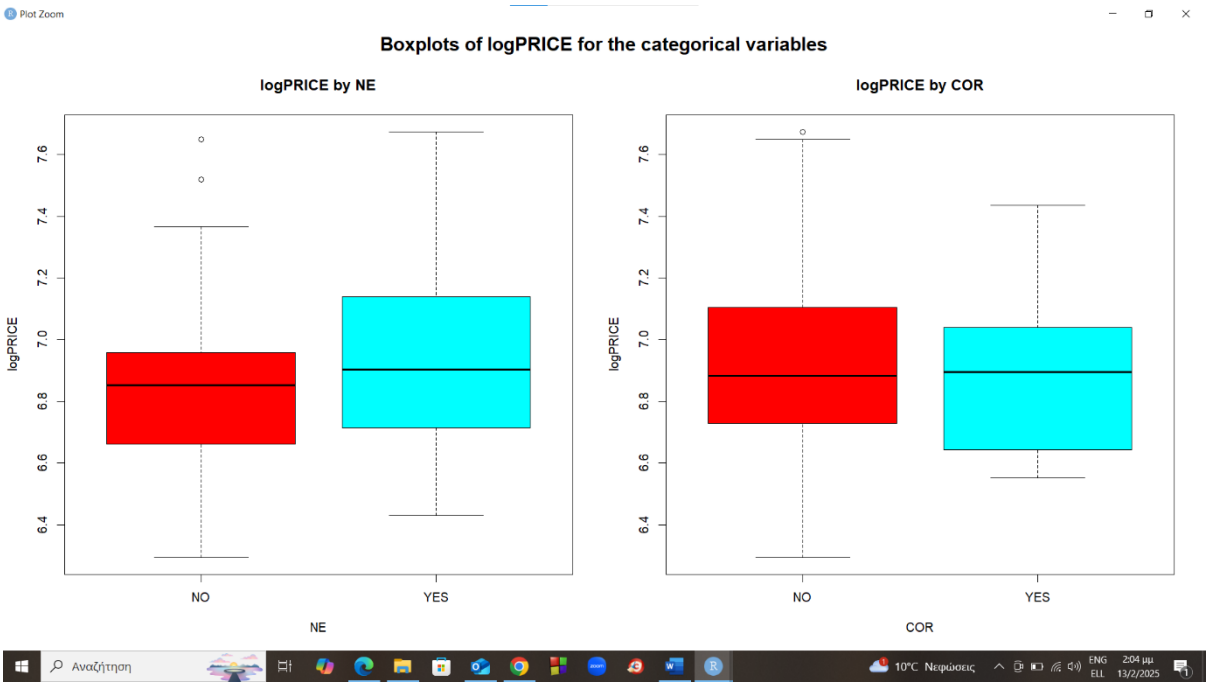
- Ο λογαριθμικός μετασχηματισμός των μεταβλητών φαίνεται να έχει **ομαλοποιήσει τις σχέσεις** και να έχει κάνει τη συσχέτιση **πιο γραμμική**.
- Η σχέση μεταξύ **logPRICE και logSQM** είναι **ιδιαίτερα έντονη και γραμμική**, πράγμα που δείχνει ότι ο log-μετασχηματισμός βελτίωσε τη σχέση μεταξύ αυτών των δύο μεταβλητών.

Παρατηρούμε Outliers

- Σε κάποιες μεταβλητές, όπως **PRICE και TAX**, υπάρχουν **μεμονωμένες τιμές μακριά από το κύριο σύνολο των δεδομένων**, που πιθανώς είναι outliers.

Συμπέρασμα

Το scatterplot matrix μας δείχνει:
Δυνατές θετικές σχέσεις μεταξύ μεταβλητών όπως **PRICE, TAX, SQM**.
Αρνητική σχέση μεταξύ **PRICE και AGE**.
Η log-μετατροπή βελτιώνει τη γραμμικότητα και αποτρέπει πιθανή ετεροσκεδαστικότητα.



1.logPRICE by NE (αριστερό γράφημα)

Ερμηνεία:

- Δείχνει τη διακύμανση της **logPRICE** για τις κατηγορίες **NE = NO** (κόκκινο) και **NE = YES** (γαλάζιο).
- Η **διάμεσος** (μαύρη γραμμή μέσα στο κουτί) είναι **χαμηλότερη** για την κατηγορία **NO**, ενώ η κατηγορία **YES** έχει λίγο υψηλότερες τιμές.
- Το **εύρος των δεδομένων** (interquartile range - IQR) φαίνεται ελαφρώς μεγαλύτερο στην κατηγορία **YES**, αλλά οι διαφορές δεν είναι πολύ έντονες.
- **Υπάρχουν outliers** (σημεία πάνω από το ανώτερο όριο).

Συμπέρασμα:

- Τα ακίνητα που ανήκουν στην κατηγορία **NE = YES** τείνουν να έχουν **ελαφρώς υψηλότερη logPRICE** από αυτά στην κατηγορία **NE = NO**.

---

2. logPRICE by COR (δεξί γράφημα)

Ερμηνεία:

- Παρουσιάζει τη διανομή της **logPRICE** για τις κατηγορίες **COR = NO** (κόκκινο) και **COR = YES** (γαλάζιο).
- Οι διάμεσοι φαίνεται να είναι **πολύ κοντινές** μεταξύ των δύο κατηγοριών.
- Το **εύρος των δεδομένων (IQR)** είναι παρόμοιο μεταξύ των δύο ομάδων.
- Υπάρχουν **μερικά outliers** και στις δύο κατηγορίες.

Συμπέρασμα:

- Φαίνεται ότι η μεταβλητή **COR** δεν επηρεάζει σημαντικά την **logPRICE**, καθώς η κατανομή είναι σχεδόν ίδια και στις δύο κατηγορίες.

Γενικό Συμπέρασμα

Η μεταβλητή **NE** φαίνεται να επηρεάζει κάπως τη logPRICE, αλλά η διαφορά είναι μικρή.
Η μεταβλητή **COR** δεν φαίνεται να έχει σημαντική επίδραση.
Υπάρχουν μερικά **outliers**, που μπορεί να επηρεάζουν τα αποτελέσματα.

Task 3: Data mining

a)

```
> summary(model)

Call:
lm(formula = logPRICE ~ . - PRICE, data = dataproject)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31314 -0.07991  0.00508  0.06264  0.38802

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7573777  1.0006857   3.755 0.000292 ***
AGE         -0.0008739  0.0016979  -0.515 0.607899
FEATS        0.0162007  0.0115638   1.401 0.164313
NEYES       -0.0065245  0.0311933  -0.209 0.834746
CORYES      -0.0410623  0.0336638  -1.220 0.225418
TAX          0.0005049  0.0001996   2.530 0.012956 *
SQM         -0.0007622  0.0016852  -0.452 0.652038
logTAX      -0.0295739  0.1482359  -0.200 0.842272
logSQM       0.6105193  0.2778883   2.197 0.030330 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1354 on 100 degrees of freedom
Multiple R-squared:  0.8312,    Adjusted R-squared:  0.8177
F-statistic: 61.57 on 8 and 100 DF,  p-value: < 2.2e-16

>
```

Ανάλυση του νέου summary(model) μετά την αφαίρεση της PRICE

Τώρα έχουμε ένα **σωστότερο μοντέλο**, αφού η PRICE δεν εμφανίζεται πλέον ως ανεξάρτητη μεταβλητή:

---

1. Υπόλοιπα (Residuals)

Residuals:

Τα υπόλοιπα (residuals) φαίνονται πιο ισορροπημένα.
Όμως το Min και Max (-0.31, 0.38) δείχνουν ότι υπάρχουν **μερικά ακραία υπόλοιπα**.

---

2. Ανάλυση Συντελεστών (βi)

Ανάλυση σημαντικών μεταβλητών (p < 0.05)

Μεταβλητή	Συντελεστής (β)	p-value	Συμπέρασμα
TAX	0.0005049	0.0129	**Σημαντική**
logSQM	0.6105	0.0303	**Σημαντική**
Intercept	3.7574	0.0003	Σταθερός όρος

Μεταβλητές που ΔΕΝ είναι σημαντικές (p > 0.05)

- AGE, FEATS, NEYES, CORYES, SQM, logTAX **δεν είναι στατιστικά σημαντικές**.
- logTAX είχε σημαντικότητα στο προηγούμενο μοντέλο, αλλά εδώ **έγινε ασήμαντο (p = 0.8423)**.

---

3. Ανάλυση R^2 και ποιότητας μοντέλου

Multiple R-squared: 0.8312, Adjusted R-squared: 0.8177

F-statistic: 61.57 on 8 and 100 DF, p-value: < 2.2e-16

Το μοντέλο εξηγεί το 83.12% της διακύμανσης της logPRICE, που είναι αρκετά καλό!
Το p-value του μοντέλου είναι πολύ μικρό (< 2.2e-16), άρα το μοντέλο συνολικά είναι σημαντικό.
Όμως, η Adjusted R^2 (81.77%) είναι χαμηλότερη, που σημαίνει ότι κάποιες μεταβλητές ίσως είναι περιττές.

Τελικό Συμπέρασμα

Οι πιο σημαντικοί παράγοντες στην πρόβλεψη της logPRICE είναι:

- **TAX**: Όταν αυξάνεται ο φόρος, η logPRICE αυξάνεται.
- **logSQM**: Όταν αυξάνονται τα τετραγωνικά μέτρα, η logPRICE αυξάνεται.

Οι μεταβλητές AGE, FEATS, NEYES, CORYES, SQM, logTAX δεν είναι στατιστικά σημαντικές και θα μπορούσαν να αφαιρεθούν από το μοντέλο.

b)

```
> summary(stepwise_model)

Call:
lm(formula = logPRICE ~ FEATS + TAX + logSQM, data = dataproject)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35660 -0.07785  0.01118  0.05894  0.39364

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.233e+00  3.743e-01  11.307  < 2e-16 ***
FEATS       1.792e-02  1.030e-02   1.740   0.0848 .
TAX         5.100e-04  8.393e-05   6.076 2.00e-08 ***
logSQM      4.452e-01  8.578e-02   5.190 1.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1338 on 105 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8218
F-statistic:   167 on 3 and 105 DF,  p-value: < 2.2e-16
```

Ανάλυση του summary(stepwise_model)

Το τελικό μοντέλο μετά το **Stepwise Selection** περιλαμβάνει τις μεταβλητές **FEATS, TAX και logSQM**, που σημαίνει ότι αυτές είναι οι πιο σημαντικές για την πρόβλεψη της logPRICE.

---

1. Υπόλοιπα (Residuals)

Residuals:

Τα υπόλοιπα φαίνονται ισορροπημένα, αλλά υπάρχουν κάποια ακραία σημεία (-0.36, 0.39).

---

2. Ανάλυση Συντελεστών (βi)

Ανάλυση σημαντικών μεταβλητών (p < 0.05)

Μεταβλητή	**Συντελεστής (β)**	**p-value**	**Ερμηνεία**
TAX	0.00051	2.00e-08 (**< 0.001**)	**Σημαντική**. Όταν ο φόρος αυξάνεται κατά **1 USD**, η logPRICE αυξάνεται κατά **0.00051 μονάδες**.
logSQM	0.4452	1.03e-06 (**< 0.001**)	**Σημαντική**. Αν τα τετραγωνικά αυξηθούν **κατά 1%**, η PRICE αυξάνεται περίπου **0.445%**.

Μεταβλητές που ΔΕΝ είναι στατιστικά σημαντικές (p > 0.05)

- Η μεταβλητή **FEATS (επιπλέον χαρακτηριστικά)** έχει **p = 0.0848**, που σημαίνει ότι **δεν είναι σημαντική στο επίπεδο 0.05**.
- Παρόλο που το p-value είναι κοντά στο 0.05, δεν είναι αρκετά χαμηλό για να τη θεωρήσουμε σίγουρα σημαντική.

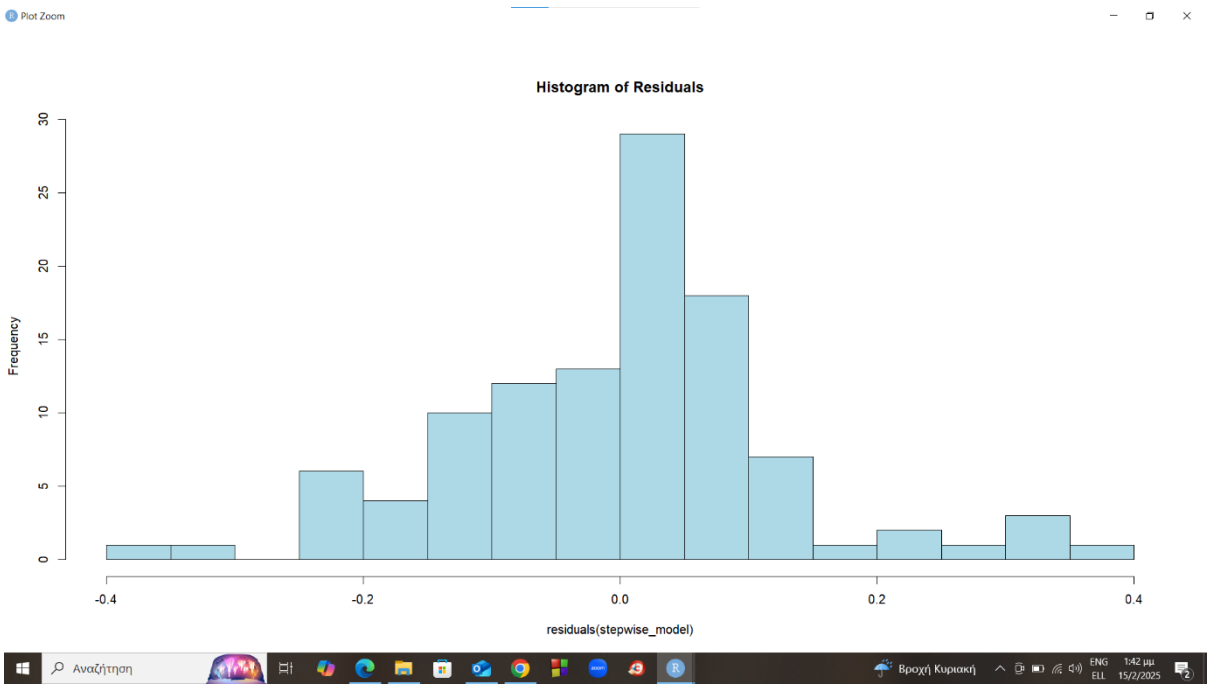
3. Αξιολόγηση R^2 και Ποιότητας του Μοντέλου

Multiple R-squared: 0.8268, Adjusted R-squared: 0.8218

F-statistic: 167 on 3 and 105 DF, p-value: < 2.2e-16

Το μοντέλο εξηγεί το 82.68% της διακύμανσης της logPRICE (πολύ καλό).
Το p-value του μοντέλου είναι πολύ μικρό (< 2.2e-16), άρα συνολικά το μοντέλο είναι στατιστικά σημαντικό.
Η Adjusted R^2 (82.18%) είναι κοντά στο R^2, που σημαίνει ότι το μοντέλο δεν έχει πολλές περιττές μεταβλητές.

Αν αφαιρέσουμε τη FEATS, μπορούμε να δούμε αν το Adjusted R^2 αλλάζει σημαντικά.



Ανάλυση Ιστογράμματος Υπολειμμάτων (Residuals Histogram)

Το γράφημα, είναι **ιστόγραμμα των υπολειμμάτων (residuals)** από το **μοντέλο stepwise regression** :

---

1. **Οριζόντιος άξονας (X-axis):**
    - Δείχνει τις τιμές των υπολειμμάτων (**residuals**), δηλαδή τη διαφορά μεταξύ των **πραγματικών τιμών** και των **προβλεπόμενων τιμών** από το μοντέλο.
    - Οι τιμές κυμαίνονται περίπου από **-0.4** έως **+0.4**.
2. **Κάθετος άξονας (Y-axis):**
    - Δείχνει τη **συχνότητα** εμφάνισης των υπολειμμάτων σε κάθε διάστημα τιμών.
3. **Κατανομή των υπολειμμάτων:**
    - Παρατηρούμε ότι τα περισσότερα υπολείμματα συγκεντρώνονται γύρω από **το 0**, κάτι που είναι **θετικό σημάδι**.

- Υπάρχουν περισσότερες τιμές προς τα **αριστερά του 0** (ελαφρώς ασύμμετρο προς τα αρνητικά).
- Ορισμένα υπολείμματα είναι αρκετά μεγάλα (στα άκρα), αλλά δεν φαίνεται να έχουμε **ακραίες αποκλίσεις (extreme outliers)**.

Κανονικότητα των Υπολειμμάτων:

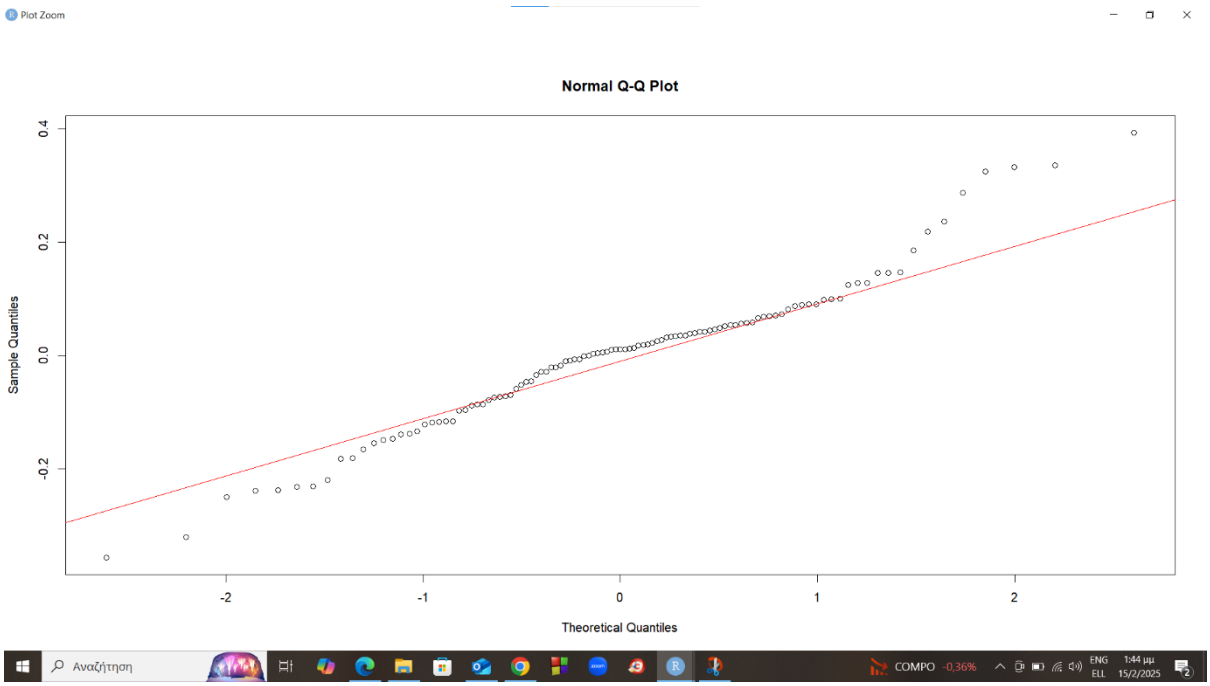
- Για να ισχύουν οι υποθέσεις της **γραμμικής παλινδρόμησης**, τα υπολείμματα πρέπει να **ακολουθούν κανονική κατανομή** (δηλαδή συμμετρικό σχήμα γύρω από το 0).
- Το ιστόγραμμα **δεν είναι τέλεια κανονικό**, καθώς φαίνεται να έχει μια ελαφριά ασυμμετρία προς τα αριστερά.

Θετικά σημεία:

- Τα υπολείμματα συγκεντρώνονται κοντά στο 0, πράγμα που δείχνει ότι το μοντέλο δεν έχει μεγάλα σφάλματα.
- Δεν παρατηρούμε μεγάλα outliers που να δείχνουν σοβαρό πρόβλημα.

Συμπέρασμα

Τα υπολείμματα φαίνονται αρκετά καλά κατανεμημένα, αλλά υπάρχει μια μικρή ασυμμετρία προς τα αριστερά.
Το μοντέλο φαίνεται να λειτουργεί ικανοποιητικά, αλλά μπορεί να υπάρχει μικρή υποεκτίμηση σε κάποιες τιμές.
Για να είμαστε σίγουροι, μπορούμε να κάνουμε επιπλέον στατιστικούς ελέγχους για την κανονικότητα των υπολειμμάτων.



Ανάλυση του Q-Q Plot για τα υπολείμματα

Το **Q-Q Plot (Quantile-Quantile Plot)** μας βοηθά να ελέγξουμε αν τα υπολείμματα ακολουθούν **κανονική κατανομή (Normality)**.

1. **Κεντρικό Τμήμα** (γύρω από 0):
    - Οι περισσότερες κουκκίδες βρίσκονται **κοντά στη διαγώνιο κόκκινη γραμμή**, κάτι που δείχνει **καλή προσέγγιση στην κανονικότητα** στο κέντρο της κατανομής.
2. **Ακραίες Τιμές (Outliers) στις Άκρες:**
    - Στις **δεξιές και αριστερές άκρες** του γραφήματος, **οι κουκκίδες αρχίζουν να αποκλίνουν από τη διαγώνιο**.
    - Αυτό σημαίνει ότι υπάρχουν **λίγα outliers (ακραίες τιμές)**, ειδικά στις μεγάλες θετικές και αρνητικές τιμές.
    - Οι τιμές αυτές είναι **μεγαλύτερες ή μικρότερες από τις αναμενόμενες στην κανονική κατανομή**.

Συμπέρασμα

Γενικά τα υπολείμματα ακολουθούν την κανονική κατανομή, αλλά υπάρχει **ελαφρά απόκλιση στα άκρα**.
Πιθανά outliers υπάρχουν στις **ακραίες θετικές και αρνητικές τιμές**.

```
> bptest(stepwise_model)

        studentized Breusch-Pagan test

data:  stepwise_model
BP = 14.942, df = 3, p-value = 0.001867

> library(car)
Loading required package: carData
Warning messages:
1: package 'car' was built under R version 4.4.1
2: package 'carData' was built under R version 4.4.1
> vif(stepwise_model)
   FEATS      TAX   logSQM
1.250133 3.971708 3.817651
```

Το Breusch-Pagan test ελέγχει την παρουσία ετεροσκεδαστικότητας στα υπόλοιπα του μοντέλου.

Ερμηνεία των αποτελεσμάτων:

- Η **μηδενική υπόθεση** (H0) του τεστ είναι ότι δεν υπάρχει ετεροσκεδαστικότητα (δηλαδή τα υπόλοιπα έχουν σταθερή διακύμανση).
- Η **εναλλακτική υπόθεση** (H1) είναι ότι υπάρχει ετεροσκεδαστικότητα (η διακύμανση των υπολοίπων δεν είναι σταθερή).
- Το **p-value = 0.001867** είναι πολύ μικρότερο από 0.05, επομένως απορρίπτουμε τη μηδενική υπόθεση και **συμπεραίνουμε ότι υπάρχει ετεροσκεδαστικότητα** στο μοντέλο.

Τι σημαίνει αυτό για το μοντέλο:

Η παρουσία ετεροσκεδαστικότητας μπορεί να επηρεάσει τις εκτιμήσεις των σταθερών σφαλμάτων και να κάνει τα p-values των μεταβλητών λιγότερο αξιόπιστα.

Το **Variance Inflation Factor (VIF)** μετρά την πολυπραγμία (multicollinearity) μεταξύ των ανεξάρτητων μεταβλητών.

Ερμηνεία των τιμών VIF:

- **FEATS = 1.25** → Πολύ χαμηλό VIF, δεν υπάρχει πολυπραγμία.
- **TAX = 3.97** → Είναι κάτω από 5, άρα δεν υπάρχει σημαντικό πρόβλημα πολυπραγμίας.
- **logSQM = 3.82** → Επίσης κάτω από 5, άρα δεν υπάρχει σοβαρό πρόβλημα πολυπραγμίας.

Συμπέρασμα:

Οι τιμές VIF είναι όλες κάτω από 5, που σημαίνει ότι **το μοντέλο δεν αντιμετωπίζει σοβαρό πρόβλημα πολυπραγμίας**. Αν κάποια τιμή ήταν πάνω από 5 ή 10, θα έπρεπε να εξετάσουμε την αφαίρεση ή τον συνδυασμό μεταβλητών.

Το μόνο ζήτημα που υπάρχει είναι η **ετεροσκεδαστικότητα (Breusch-Pagan test)**.

c)

```
> summary(anova_model)
             Df Sum Sq Mean Sq F value   Pr(>F)
catFEATS      2  1.350  0.6749   7.523 0.000881 ***
Residuals   106  9.509  0.0897
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(anova_model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = logPRICE ~ catFEATS, data = dataproject)

$catFEATS
                    diff         lwr       upr     p adj
Moderate-Low  0.22235970  0.08512638 0.3595930 0.0005857
High-Low      0.25543668 -0.46349360 0.9743670 0.6762246
High-Moderate 0.03307698 -0.68512527 0.7512792 0.9934145
```

Ερμηνεία των αποτελεσμάτων του ANOVA :summary(anova_model)

Το μοντέλο ANOVA εξετάζει αν υπάρχουν στατιστικά σημαντικές διαφορές στις μέσες τιμές της **logPRICE** μεταξύ των κατηγοριών της **catFEATS** ("Low", "Moderate", "High").

- **p-value = 0.000881** → Πολύ μικρότερο από 0.05, άρα απορρίπτουμε την υπόθεση ότι οι μέσες τιμές είναι ίσες.
- **F-value = 7.523** → Υψηλή τιμή F δείχνει ότι υπάρχει διαφορά μεταξύ των ομάδων.
- **Συμπέρασμα:** Υπάρχει στατιστικά σημαντική διαφορά μεταξύ τουλάχιστον δύο από τις κατηγορίες **catFEATS**.

Ερμηνεία των αποτελεσμάτων του :TukeyHSD(anova_model)

Τα αποτελέσματα του **Tukey HSD test** παρέχουν συγκρίσεις μεταξύ των επιπέδων της κατηγορικής μεταβλητής **catFEATS** όσον αφορά τη μέση τιμή της **logPRICE**. Ας αναλύσουμε τα αποτελέσματα:

Σύγκριση	**Διαφορά Μέσων (diff)**	**Κατώτερο Όριο (lwr)**	**Ανώτερο Όριο (upr)**	**p-value**
Moderate - Low	0.2224	0.0851	0.3596	0.0005857
High - Low	0.2554	-0.4635	0.9744	0.6762246
High - Moderate	0.0331	-0.6851	0.7513	0.9934145

Ερμηνεία:

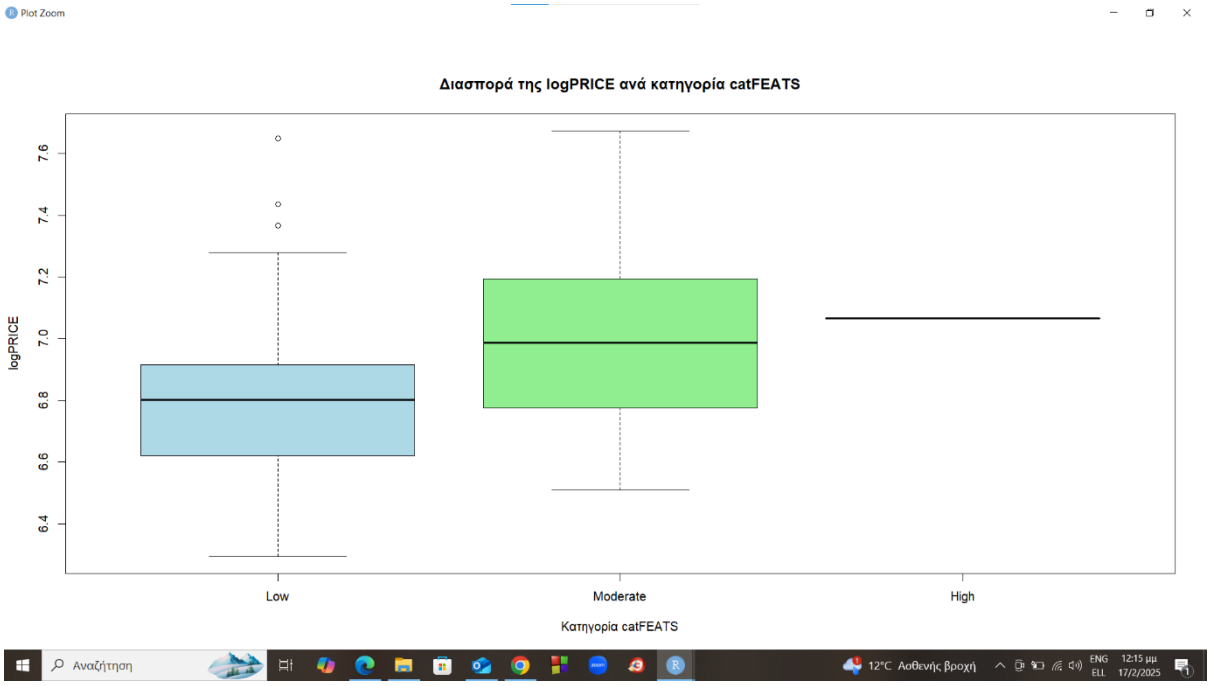
- **Moderate - Low:** Η διαφορά στις μέσες τιμές της logPRICE μεταξύ των κατηγοριών "Moderate" και "Low" είναι **0.2224**. Το 95% διάστημα εμπιστοσύνης για αυτή τη διαφορά είναι από **0.0851** έως **0.3596**, και το p-value είναι **0.0005857**. Αυτό υποδηλώνει ότι η διαφορά είναι στατιστικά σημαντική, με την κατηγορία "Moderate" να έχει υψηλότερη μέση τιμή logPRICE σε σχέση με την "Low".
- **High - Low:** Η διαφορά στις μέσες τιμές μεταξύ "High" και "Low" είναι **0.2554**. Ωστόσο, το διάστημα εμπιστοσύνης περιλαμβάνει το μηδέν (-0.4635 έως 0.9744), και το p-value είναι **0.6762246**, υποδεικνύοντας ότι αυτή η διαφορά δεν είναι στατιστικά σημαντική.

- **High - Moderate:** Η διαφορά μεταξύ "High" και "Moderate" είναι **0.0331**, με διάστημα εμπιστοσύνης από -0.6851 έως 0.7513, και p-value **0.9934145**. Επίσης, αυτή η διαφορά δεν είναι στατιστικά σημαντική.

Συμπέρασμα:

Παρόλο που το αρχικό ANOVA έδειξε ότι υπάρχουν συνολικά διαφορές μεταξύ των ομάδων, το **Tukey HSD test** αποκαλύπτει ότι η στατιστικά σημαντική διαφορά εντοπίζεται μόνο μεταξύ των κατηγοριών "Moderate" και "Low". Οι άλλες συγκρίσεις δεν παρουσιάζουν στατιστικά σημαντικές διαφορές.

Αυτό σημαίνει ότι οι ιδιοκτησίες με "Moderate" αριθμό χαρακτηριστικών (FEATS) έχουν σημαντικά υψηλότερη μέση τιμή logPRICE σε σύγκριση με εκείνες με "Low" αριθμό χαρακτηριστικών, ενώ οι κατηγορίες "High" δεν διαφέρουν σημαντικά από τις άλλες δύο.



Το **boxplot** δείχνει τη διασπορά των τιμών **logPRICE** για κάθε επίπεδο της μεταβλητής **catFEATS** ("Low", "Moderate", "High"):

1. **Low (Γαλάζιο κουτί)**
    - Η διάμεσος (μαύρη γραμμή στο κουτί) φαίνεται χαμηλότερη από την κατηγορία "Moderate".
    - Οι τιμές κυμαίνονται σε ένα πιο στενό εύρος.

- Υπάρχουν κάποια outliers (κυκλάκια πάνω από το κουτί).

2. **Moderate (Πράσινο κουτί)**
    - Η διάμεσος είναι υψηλότερη από την "Low", κάτι που συμφωνεί με το στατιστικά σημαντικό αποτέλεσμα του Tukey HSD.
    - Υπάρχει μεγαλύτερη διασπορά των δεδομένων.
3. **High (Μαύρη γραμμή)**
    - Δεν εμφανίζεται κουτί, κάτι που σημαίνει ότι υπάρχουν λίγα δεδομένα σε αυτή την κατηγορία, μόνο μία τιμή.
    - Αυτό μπορεί να εξηγεί γιατί οι διαφορές μεταξύ "High" και των άλλων κατηγοριών δεν είναι στατιστικά σημαντικές.

Συμπέρασμα:

- **Η κατηγορία "Moderate" έχει σημαντικά υψηλότερη μέση τιμή logPRICE από την "Low".**
- **Η κατηγορία "High" φαίνεται να έχει πολύ λίγα δεδομένα, κάτι που μπορεί να επηρεάζει τα στατιστικά αποτελέσματα.**
- **Οπτικά, το αποτέλεσμα του ANOVA και του Tukey HSD επιβεβαιώνεται.**

d)

```
> summary(log_model)

Call:
glm(formula = NE ~ PRICE + SQM + AGE + FEATS + COR + TAX, family = binomial,
    data = dataproject)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.869e+00  1.361e+00  -3.578 0.000346 ***
PRICE        2.271e-05  1.575e-03   0.014 0.988492
SQM         -4.284e-02  1.431e-02  -2.994 0.002757 **
AGE          1.684e-01  4.089e-02   4.118 3.83e-05 ***
FEATS        8.115e-01  2.654e-01   3.057 0.002235 **
CORYES      -3.772e-01  5.795e-01  -0.651 0.515117
TAX          8.405e-03  2.625e-03   3.202 0.001367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 139.67  on 108  degrees of freedom
Residual deviance: 107.28  on 102  degrees of freedom
AIC: 121.28

Number of Fisher Scoring iterations: 5
```

Ας αναλύσουμε τα αποτελέσματα και να προχωρήσουμε στην επιλογή του **βέλτιστου μοντέλου**:

---

1. Ανάλυση των αποτελεσμάτων του αρχικού μοντέλου

Σημαντικά ευρήματα:

- **Σημαντικές μεταβλητές (p < 0.05)**:
  - **SQM** (β = -0.0428, p = 0.0028) → Αρνητική επίδραση
  - **AGE** (β = 0.1684, p = 3.83e-05) → Θετική επίδραση
  - **FEATS** (β = 0.8115, p = 0.0022) → Θετική επίδραση
  - **TAX** (β = 0.0084, p = 0.0014) → Θετική επίδραση
- **Μη σημαντικές μεταβλητές (p > 0.05)**:
  - **PRICE** (p = 0.9885) → Καμία επίδραση
  - **COR** (p = 0.5151) → Καμία επίδραση

Συμπέρασμα:
Οι μεταβλητές **PRICE** και **COR** δεν είναι σημαντικές και μπορούν να αφαιρεθούν για να απλοποιηθεί το μοντέλο.

```
> summary(optimized_log_model)

Call:
glm(formula = NE ~ SQM + AGE + FEATS + TAX, family = binomial,
    data = dataproject)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.890218   1.329261  -3.679 0.000234 ***
SQM         -0.043560   0.013349  -3.263 0.001101 **
AGE          0.167923   0.040518   4.144 3.41e-05 ***
FEATS        0.801713   0.259763   3.086 0.002026 **
TAX          0.008547   0.002495   3.425 0.000614 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 139.67  on 108  degrees of freedom
Residual deviance: 107.72  on 104  degrees of freedom
AIC: 117.72

Number of Fisher Scoring iterations: 5
```

Το νέο βελτιστοποιημένο μοντέλο έχει **μικρότερο AIC (117.72 αντί για 121.28)**, που σημαίνει ότι είναι πιο αποδοτικό!

Σύγκριση:

Όλες οι μεταβλητές είναι στατιστικά σημαντικές (p < 0.05).
Το AIC μειώθηκε, άρα το νέο μοντέλο έχει καλύτερη προσαρμογή με λιγότερες περιττές μεταβλητές.
Η μεταβλητή PRICE και η μεταβλητή COR αφαιρέθηκαν, καθώς δεν πρόσφεραν πληροφορία.

e)

accuracy = 0.8073394 : **Ανάλυση του Accuracy (Ακρίβεια Μοντέλου)**

Το Decision Tree έχει ακρίβεια **80.73%**, που σημαίνει ότι το μοντέλο **προβλέπει σωστά αν ένα σπίτι είναι γωνιακό (COR = YES/NO) σε περίπου 81% των περιπτώσεων**.

Τι σημαίνει αυτό;

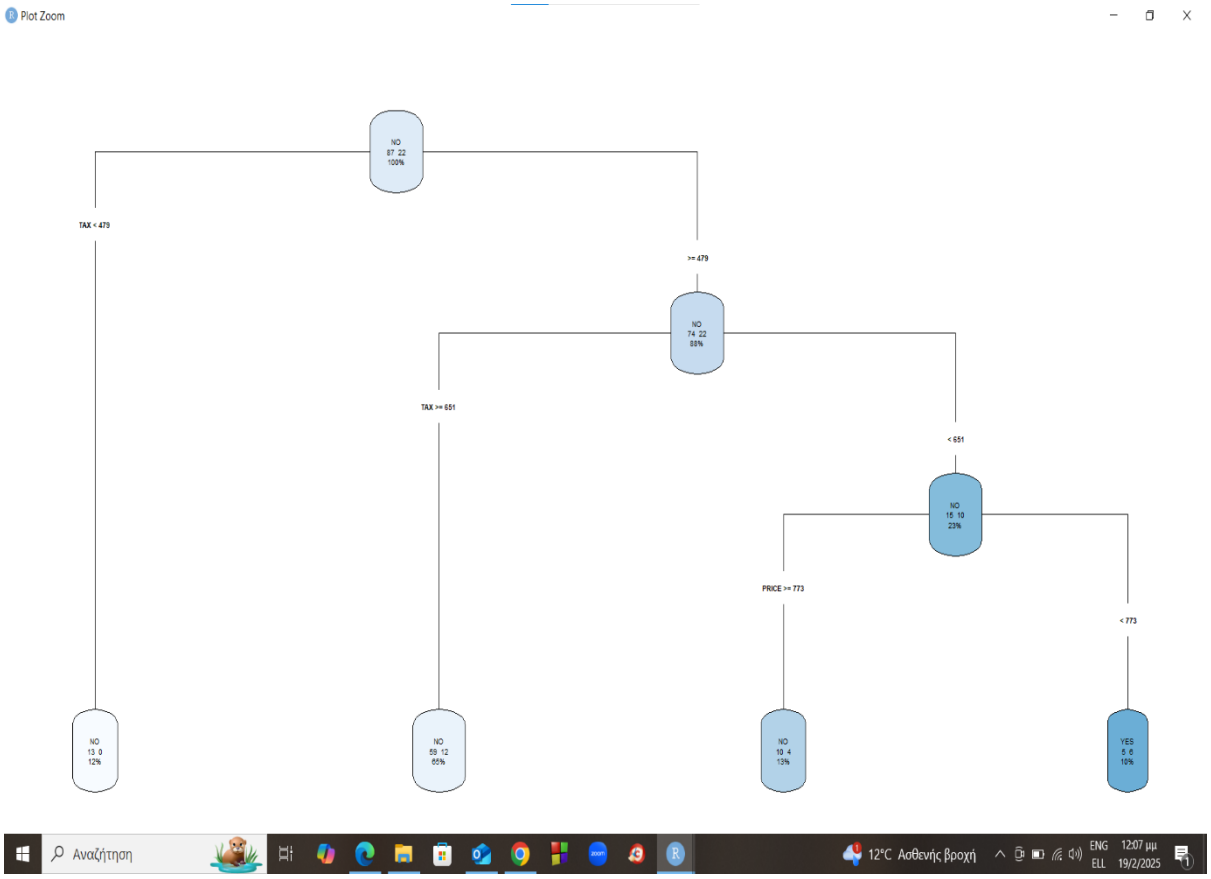
- Ένα accuracy **πάνω από 80% θεωρείται καλό** για ένα μοντέλο ταξινόμησης (classification).
- Το μοντέλο μπορεί να προβλέψει με σχετικά μεγάλη ακρίβεια αν ένα σπίτι είναι γωνιακό ή όχι, με βάση τις μεταβλητές **PRICE, SQM, AGE, FEATS, NE, TAX**.

Ανάλυση της Πρόβλεψης για το Νέο Σπίτι

Το Decision Tree προβλέπει **"NO"**, που σημαίνει ότι **το σπίτι ΔΕΝ είναι γωνιακό (COR = NO)**, με βάση τις τιμές :

Μεταβλητή	Τιμή
PRICE	1000
SQM	150
AGE	17
FEATS	4
NE	YES
TAX	800

Συμπέρασμα:
Με βάση το Decision Tree, **ένα σπίτι με αυτά τα χαρακτηριστικά είναι πιο πιθανό να ΜΗΝ είναι γωνιακό (COR = NO)**.



1. **Ριζικός κόμβος (Root Node)**:
    - Το αρχικό split γίνεται με βάση το **TAX (φορολογία)** στο **479**.
    - Αν **TAX < 479**, τότε το σπίτι προβλέπεται ως **"NO" (όχι γωνιακό)**.
    - Αν **TAX >= 479**, τότε συνεχίζουμε σε περαιτέρω διαχωρισμούς.
2. **Επόμενα splits:**
    - Για **TAX >= 479**, γίνεται νέο split στο **651**.
    - Αν **TAX >= 651**, η πρόβλεψη παραμένει **"NO"** (όχι γωνιακό).
    - Αν **TAX < 651**, τότε λαμβάνεται υπόψη και η τιμή **PRICE (τιμή σπιτιού)** στο **773**.
    - Αν **PRICE < 773**, υπάρχει πιθανότητα να είναι **"YES" (γωνιακό σπίτι)**.

Ερμηνεία των αριθμών στους κόμβους:

Κάθε κόμβος περιέχει τρεις πληροφορίες:

1. **Κατηγορία (YES/NO)** → Η προβλεπόμενη κλάση.
2. **Πλήθος (π.χ., 74 22)** → Δηλώνει το πλήθος των περιπτώσεων που έφτασαν σε αυτόν τον κόμβο.
3. **Ποσοστό (%)** → Δείχνει το ποσοστό των συνολικών δεδομένων που βρίσκονται στον κόμβο.

Τι σημαίνει για την πρόβλεψη του νέου σπιτιού (PRICE=1000, SQM=150, AGE=17, FEATS=4, NE=YES, TAX=800);

- **TAX = 800** → Είναι **πάνω από 651**, άρα πάει στο αριστερό μονοπάτι.
- **PRICE = 1000** → Είναι **πάνω από 773**, άρα η πρόβλεψη είναι **"NO" (όχι γωνιακό)**.