Statistical Analysis with R in Saratoga Dataset

Εισαγωγή

Σκοπός της παρούσας μελέτης είναι εξαγωγή συμπερασμάτων για τα δεδομένα των σπιτιών της Saratoga. Το dataset που θα επεξεργαστούμε έχει τις ακόλουθες μεταβλητές: Price: η τιμή (σε 1000s $) (εξαρτημένη μεταβλητή) lotSize: μέγεθος οικοπέδου (square feet) age: παλαιότητα του σπιτιού (years) landValue: αξία οικοπέδου (1000s of US dollars) livingArea: εμβαδό σπιτιού (square feet) pctCollege: ποσοστό γειτόνων που είναι απόφοιτη κολεγίου bedrooms: αριθμός υπνοδωματίων firplaces: αριθμός τζακιών bathrooms: αριθμός μπάνιων rooms: αριθμός δωματίων heating type: τύπος θέρμανσης fuel: καύσιμο θέρμανσης sewer type: τύπος υπονόμου waterfront: ύπαρξη προκυμαίας newConstruction: αν είναι νεόδμητο centralAir: αν έχει κεντρικό αερισμό

Ερυνητικά ερωτήματα

Παρακάτω θα μελετήσουμε τη σχέση όλων των μεταβλητών με την τιμή του σπιτιού και θα προσπαθήσουμε να διερευνήσουμε το βαθμό της συσχέτισης τους με αυτήν κατασκευάζοντας κατάλληλα μοντέλα πρόβλεψης. Ακολούθως θα διερευνήσουμε κατά πόσο η διαισθητικά εύλογη συσχετίση του αν είναι νεόδμητο το σπίτι με το ποσοστό αποφοίτων κολεγίου της γειτονιάς, τον αριθμό δωματίων, την τιμή και (την εξ ορισμού) ηλικία της κατασκευής, επιβεβαιώνονται από τα δεδομένα μας:

```
system("defaults write org.R-project.R force.LANG el_GR.UTF-8")
```

```
## Warning in system("defaults write org.R-project.R force.LANG el_GR.UTF-8"):
## 'defaults' not found
```

```
## [1] 127
```

```
SH<-read.csv("D:/Desktop/SaratogaHouses.csv", header=T)
SH<-subset(SH, select=-c(X))
```

Η παρακάτω εντολή σημαίνει ότι η βάση δεδομένων αναζητείται από την R κατά την αξιολόγηση μιας μεταβλητής, έτσι ώστε τα αντικείμενα στη βάση δεδομένων να είναι προσβάσιμα προς ανάλυση δίνοντας απλώς τα ονόματά τους.

```
attach(SH)
```

Με τηνεντολή names() πραγματοποιείται εκτύπωση όλων των μεταβλητών στο πλαίσιο δεδομένων.

```
names(SH)
```

```
##  [1] "price"           "lotSize"         "age"             "landValue"
##  [5] "livingArea"      "pctCollege"      "bedrooms"        "fireplaces"
##  [9] "bathrooms"       "rooms"           "heating"         "fuel"
## [13] "sewer"           "waterfront"      "newConstruction" "centralAir"
```

Πραγματοποιείται εκτύπωση των πρώτων 6 εγγραφών για κάθε στήλη του dataframe

```
head(SH)
```

```
##    price lotSize age landValue livingArea pctCollege bedrooms fireplaces
## 1 132500    0.09  42     50000        906         35        2          1
## 2 181115    0.92   0     22300       1953         51        3          0
## 3 109000    0.19 133      7300       1944         51        4          1
## 4 155000    0.41  13     18700       1944         51        3          1
## 5  86060    0.11   0     15000        840         51        2          0
## 6 120000    0.68  31     14000       1152         22        4          1
##   bathrooms rooms         heating     fuel             sewer waterfront
## 1       1.0     5        electric electric            septic         No
## 2       2.5     6 hot water/steam      gas            septic         No
## 3       1.0     8 hot water/steam      gas public/commercial         No
## 4       1.5     5         hot air      gas            septic         No
## 5       1.0     3         hot air      gas public/commercial         No
## 6       1.0     8         hot air      gas            septic         No
##   newConstruction centralAir
## 1              No         No
## 2              No         No
## 3              No         No
## 4              No         No
## 5             Yes        Yes
## 6              No         No
```

Με την εντολή summary() βρίσκονται όλα τα σημαντικά στατιστικά στοιχεία της εξαρτημένης και των ανεξάρτητων μεταβλητών(Διάμεσος , μέση τιμή, 1ο & 3ο τεταρτημόριο)

```
summary(SH)
```

```
##      price            lotSize              age           landValue
##  Min.   :  5000   Min.   : 0.0000   Min.   :  0.00   Min.   :   200
##  1st Qu.:145000   1st Qu.: 0.1700   1st Qu.: 13.00   1st Qu.: 15100
##  Median :189900   Median : 0.3700   Median : 19.00   Median : 25000
##  Mean   :211967   Mean   : 0.5002   Mean   : 27.92   Mean   : 34557
##  3rd Qu.:259000   3rd Qu.: 0.5400   3rd Qu.: 34.00   3rd Qu.: 40200
##  Max.   :775000   Max.   :12.2000   Max.   :225.00   Max.   :412600
##    livingArea     pctCollege       bedrooms       fireplaces       bathrooms
##  Min.   : 616   Min.   :20.00   Min.   :1.000   Min.   :0.0000   Min.   :0.0
##  1st Qu.:1300   1st Qu.:52.00   1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:1.5
##  Median :1634   Median :57.00   Median :3.000   Median :1.0000   Median :2.0
##  Mean   :1755   Mean   :55.57   Mean   :3.155   Mean   :0.6019   Mean   :1.9
##  3rd Qu.:2138   3rd Qu.:64.00   3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:2.5
##  Max.   :5228   Max.   :82.00   Max.   :7.000   Max.   :4.0000   Max.   :4.5
##      rooms          heating              fuel              sewer
##  Min.   : 2.000   Length:1728        Length:1728        Length:1728
```

```
##  1st Qu.: 5.000   Class :character   Class :character   Class :character
##  Median : 7.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 7.042
##  3rd Qu.: 8.250
##  Max.   :12.000
##   waterfront        newConstruction     centralAir
##  Length:1728        Length:1728        Length:1728
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Η μεταβλητή lotSize δίνεται σε εκτάρια και θα πρέπει να μετατραπεί σε square feet

```
lotSize=lotSize*43560
```

```
SH$fireplaces <- as.factor(SH$fireplaces)
SH$bedrooms <- as.factor(SH$bedrooms)
SH$bathrooms <- as.factor(SH$bathrooms)
SH$rooms <- as.factor(SH$rooms)
```

Εύρεση της δειγματικής διαμέσου για κάθε αριθμητική μεταβλητή του dataframe

```
sapply(Filter(is.numeric, SH), FUN = mean, na.rm = TRUE)
```

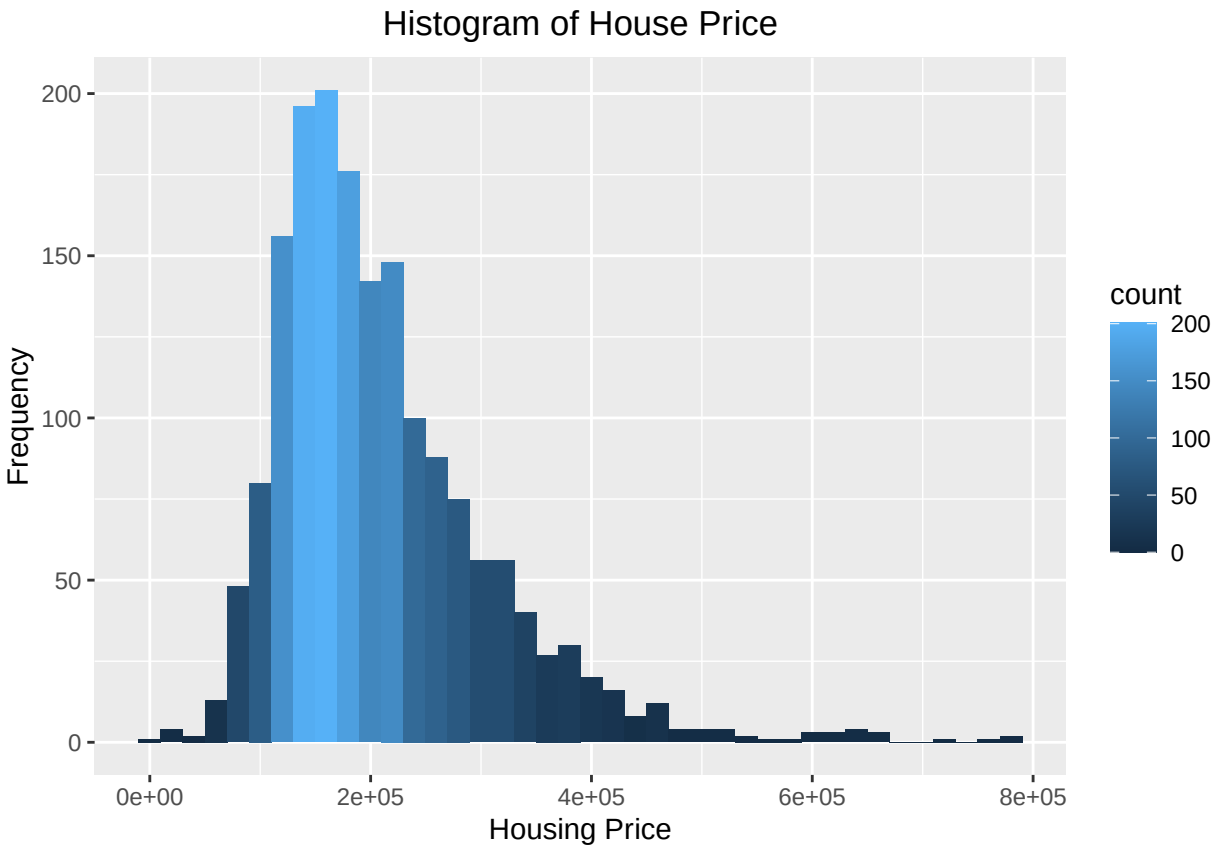
```
##        price      lotSize          age    landValue   livingArea   pctCollege
## 2.119667e+05 5.002141e-01 2.791609e+01 3.455719e+04 1.754976e+03 5.556771e+01
```

Τα ακόλουθα γραφήματα που δημιουργήθηκαν με τη βιβλιοθήκη ggplot() απεικονίζουν γραφικά τις μεταβλητές και τις μεταξύ τους σχέσεις.

```
library(ggplot2)
```

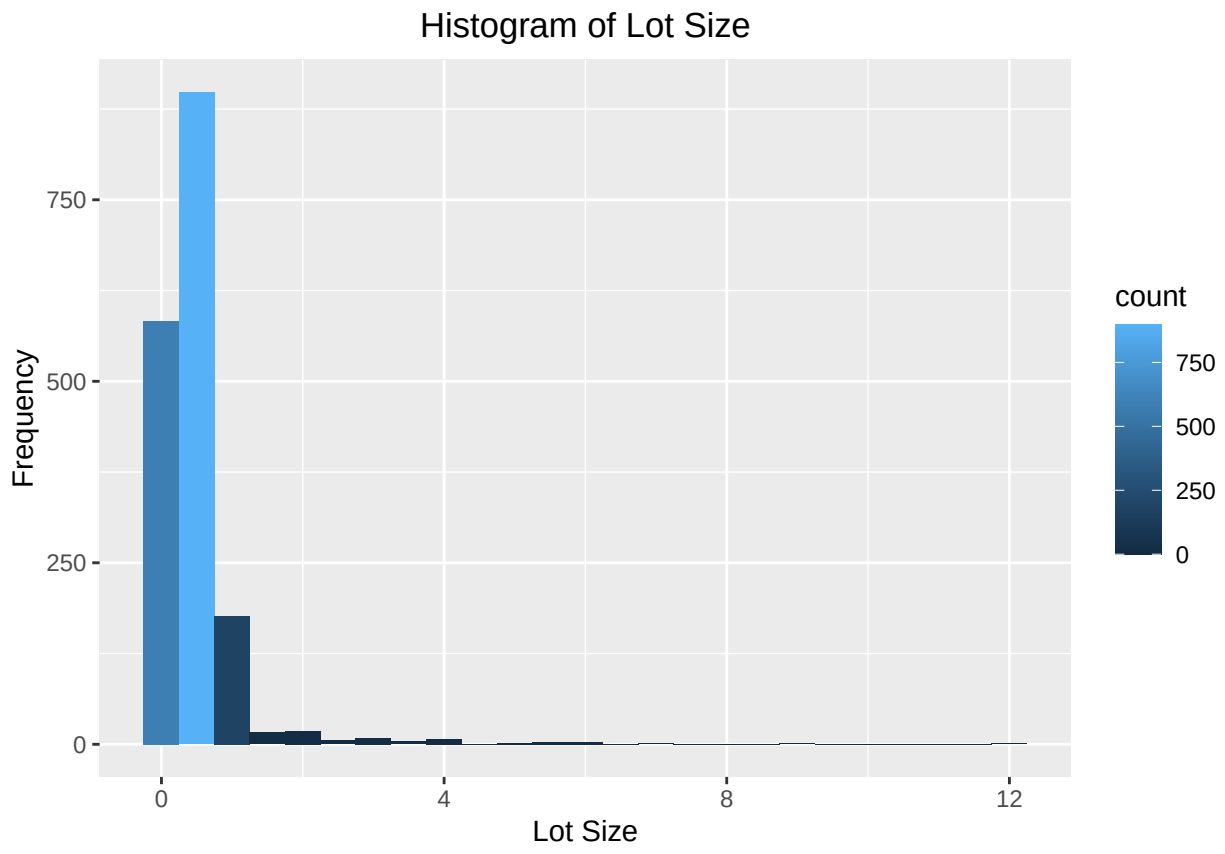
```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
ggplot(SH, aes(x = price, fill = ..count..)) +
  geom_histogram(binwidth = 20000) +
  ggtitle("Histogram of House Price") +
  ylab("Frequency") +
  xlab("Housing Price") +
  theme(plot.title = element_text(hjust = 0.5))
```

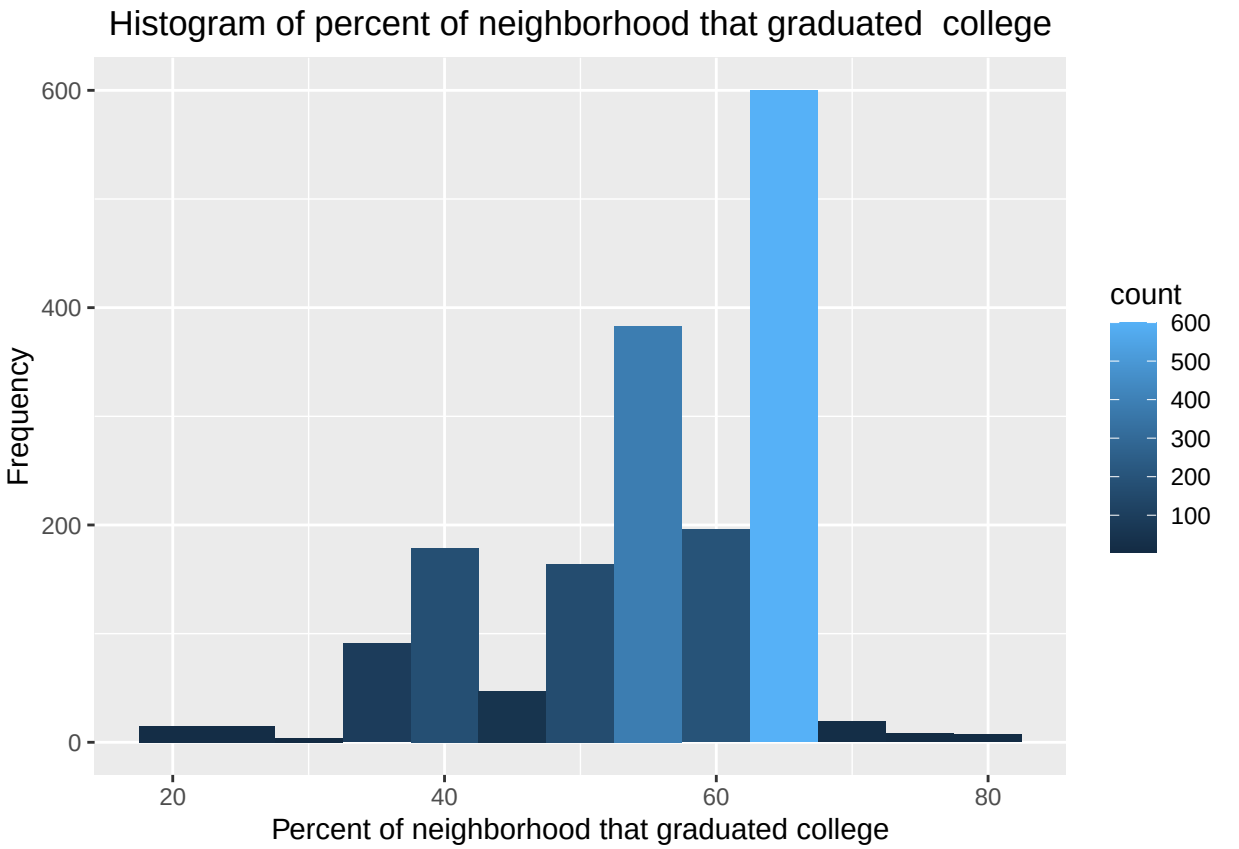


Στο πρώτο ιστόγραμμα φαίνεται η συχνότητα των σπιτιών συγκριτικά με την τιμή. όπως βλέπουμε και απο το plot τα δεδομένα ακολουθούν Skewed Distribution - Η λοξή κατανομή η οποία είναι ασύμμετρη επειδή ένα φυσικό όριο αποτρέπει τα αποτελέσματα στη μία πλευρά.Ο συγκεκριμένος τύπος κατανομής μας δείχνει θετική ασσυμετρία στο δείγμα. Στην συγκεκριμένη περίπτωση φαίνεται οτι υπάρχει δεξιά λοξότητα καθώς το μεγαλύτερο πλήθος των σπιτιών έχουν αγοραστική αξία από 0 - 400.000, ενώ αρκετά λιγότερα έχουν αξία >400.000. Η κορυφή της διανομής είναι εκτός κέντρου προς το όριο και μια ουρά απλώνεται μακριά από αυτό.

```
ggplot(SH, aes(x = lotSize, fill = ..count..)) +
  geom_histogram(binwidth = 0.5) +
  ggtitle("Histogram of Lot Size") +
  ylab("Frequency") + xlab("Lot Size") +
  theme(plot.title = element_text(hjust = 0.5))
```

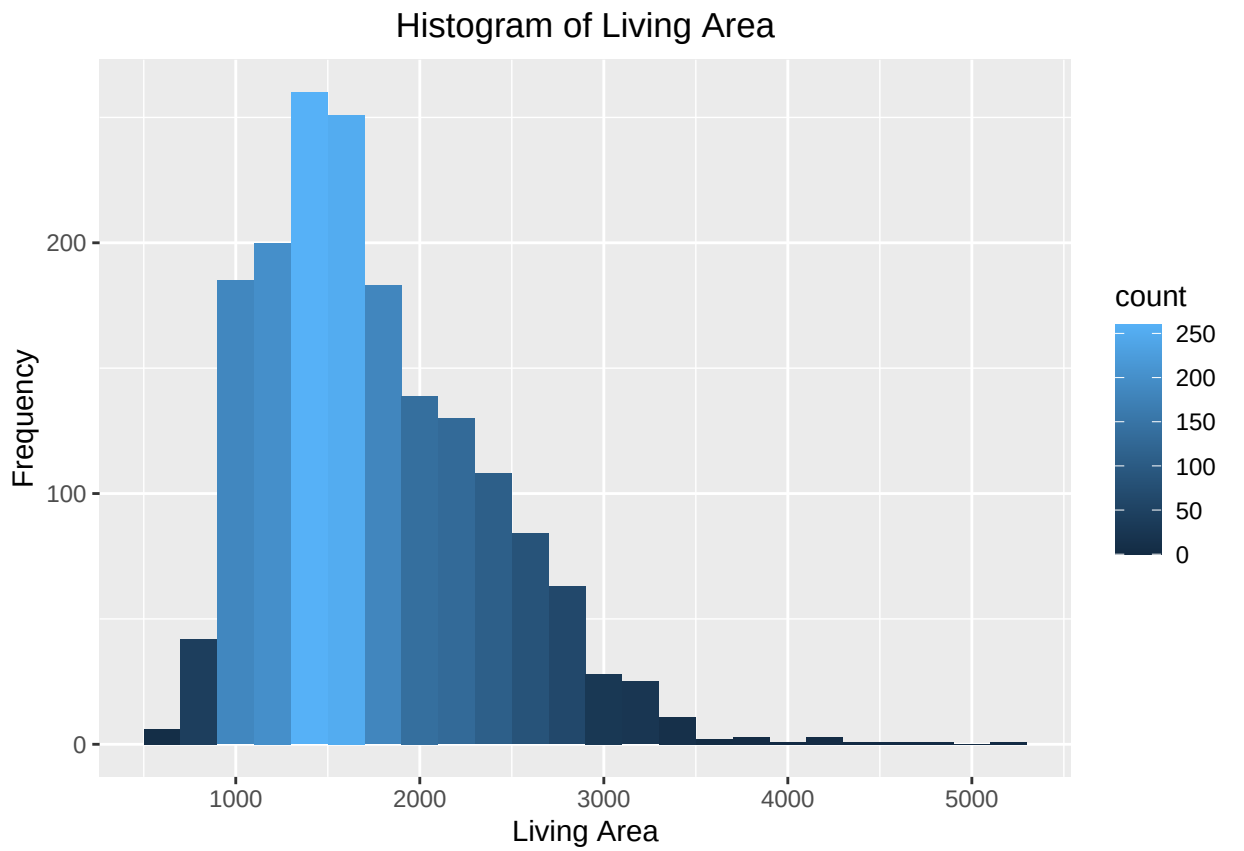



```
ggplot(SH, aes(x = pctCollege, fill = ..count..)) +
  geom_histogram(binwidth = 5) +
  ggtitle("Histogram of percent of neighborhood that graduated  college") +
  ylab("Frequency") +
  xlab("Percent of neighborhood that graduated college") +
  theme(plot.title = element_text(hjust = 0.5))
```



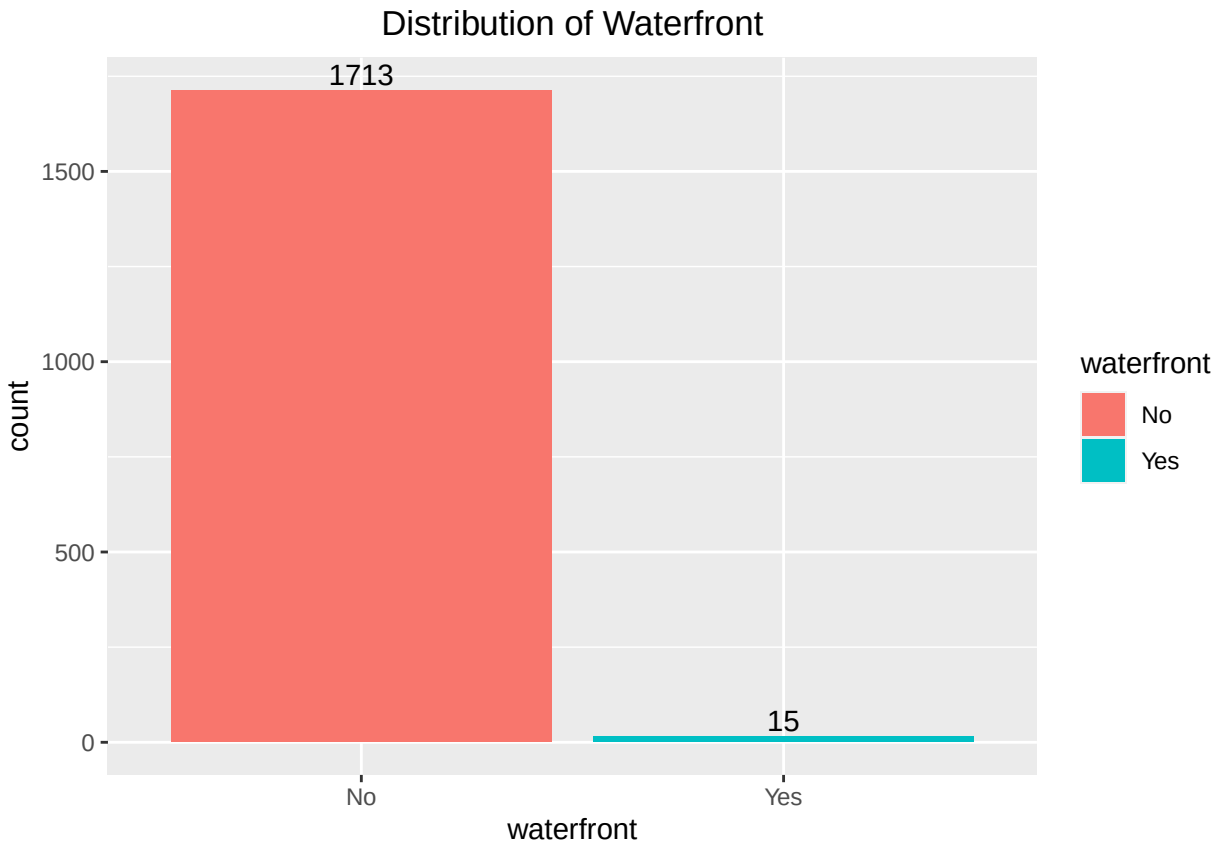
Το διάγραμμα για την ανεξάρτητη μεταβλητή pctCollege δείχνει το ποσοστό αποφοίτων κολεγίου της γειτονιάς. Το δείγμα ακολουθεί λοξή κατανομή απο την δεξιά πλευρά ή αρνητικά λοξή κατανομή, έχοντας επίσης αρνητική ασσυμετρία. Αυτή η συνθήκη προκύπτει καθώς οι μικρότερες τιμές του δείγματος έχουν λιγότερες πιθανότητες να εμφανιστούν.Πιο συγκεκριμένα είναι πολυ μικρό το ποσοστό (<30%) των γειτόνων που δεν έχουν αποφοιτήσει.

```
ggplot(SH, aes(x = livingArea, fill = ..count..)) + geom_histogram(binwidth = 200) +ggtitle("Histogram
```



Το τελευταίο ιστόγραμμα υποδεικνύει την σχέση της LivingArea με την συχνότητα ακολουθώντας λοξή κατανομή επίσης με θετική ασσυμετρία.

```
## `geom_smooth()` using formula 'y ~ x'
ggplot(SH, aes(x = waterfront, fill = waterfront )) +
geom_bar()+ ggtitle("Distribution of Waterfront")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



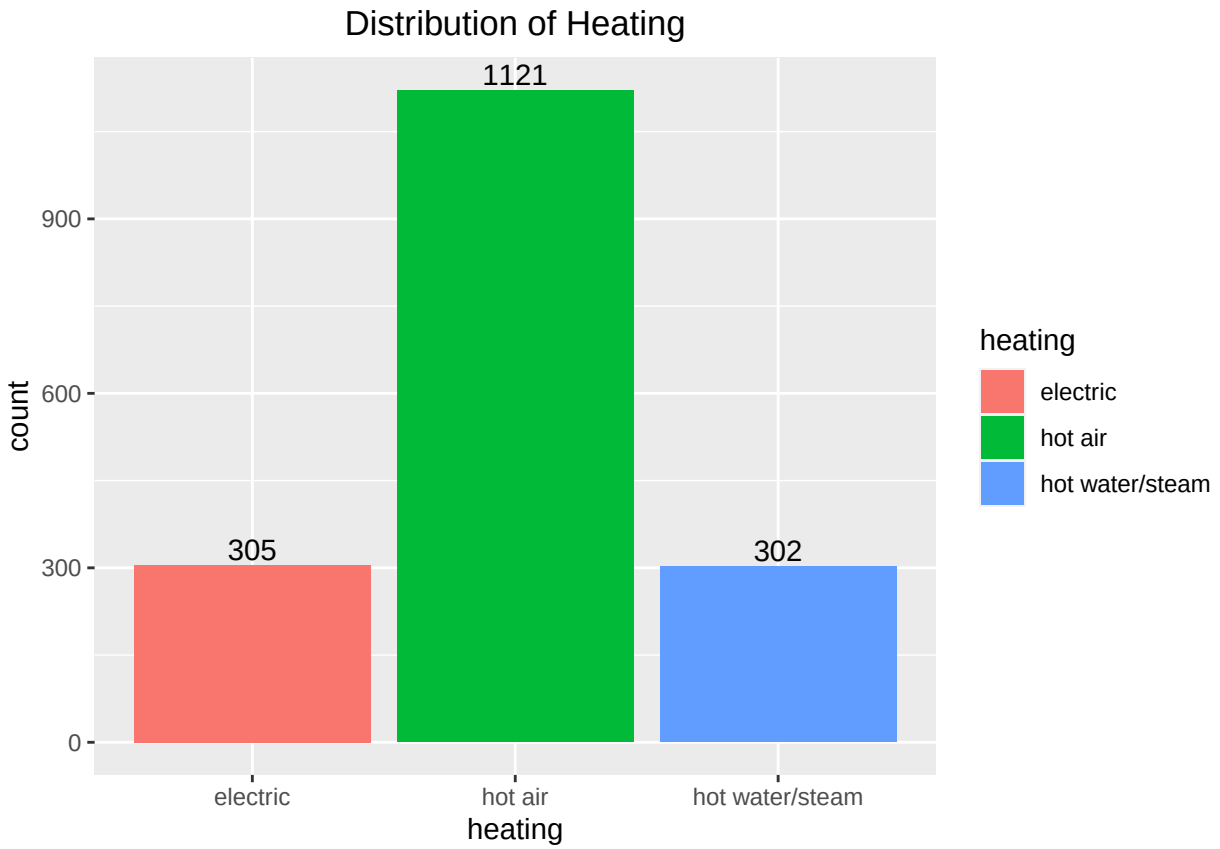
Το παραπάνω ραβδόγραμμα περιλαμβανει την κατονομή και δείχνει πόσα σπίτια στην Saratoga διαθέτουν προκυμαία - waterfront. Παρατηρείται οτι η πλειονότητα των σπιτιών δεν διαθέτει προκυμαία στο σπίτι του.

Ο πίνακας συχνοτήτων και σχετικών συχνότητων σχετικά με το αν το σπίτι διαθέτει προκυμαία ή όχι

~	Waterfront	Συχνότητα	Σχετική συχνότητα
1	Yes	15	0.9913
2	no	1713	0.008

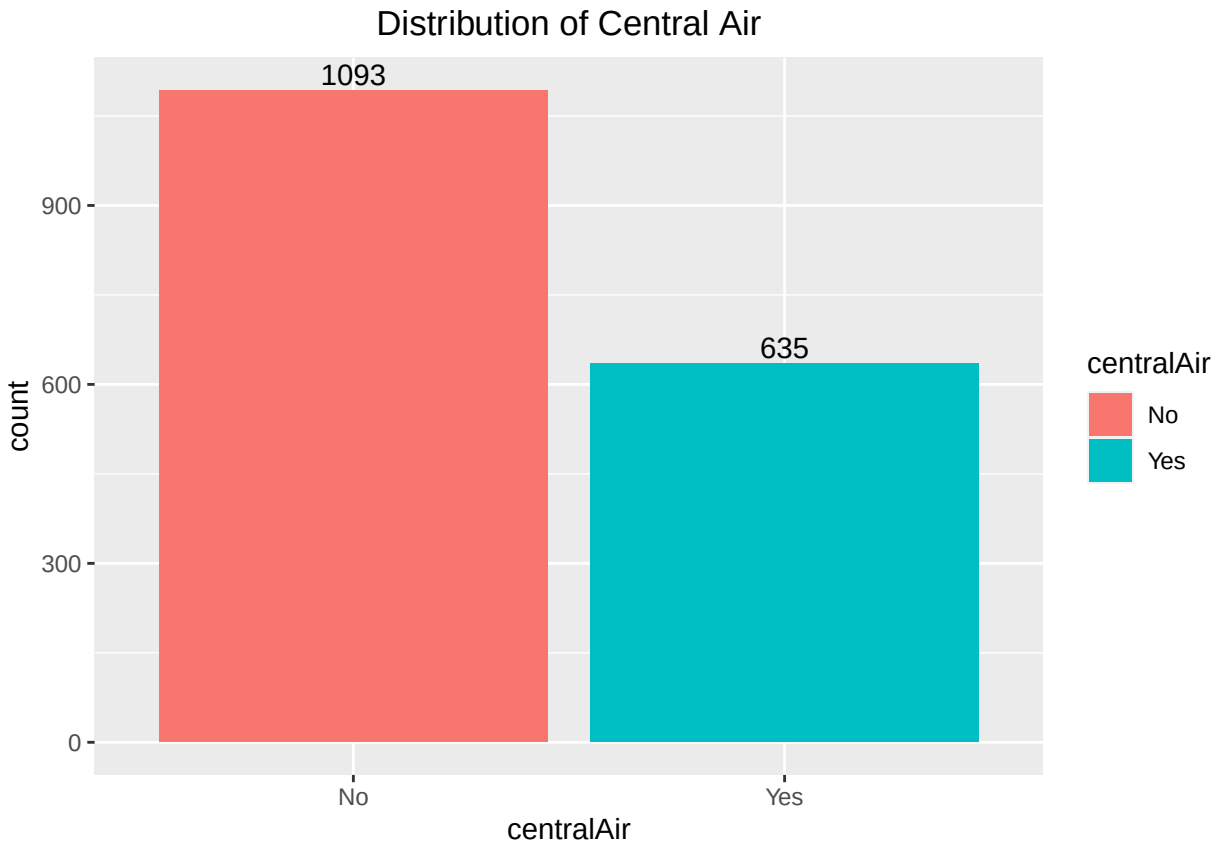
Η κατανομή της μεταβλητής heating, φαίνεται παρακάτω.

```
ggplot(SH, aes(x = heating, fill = heating )) +
geom_bar()+ ggtitle("Distribution of Heating")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



Η κατανομή της μεταβλητής centralAir, φαίνεται παρακάτω. Παρατηρούμε οτι η κατανομή αυτη ειναι περισσότερο κανονικοποιημένη καθώς οι τιμές του 'Εχω κεντρικη θέρμανση' και του 'δεν εχω' δεν έχουν τόσο μεγάλη απόκλιση όσο προηγουμένως.

```
ggplot(SH, aes(x = centralAir, fill = centralAir )) +
geom_bar()+ ggtitle("Distribution of Central Air")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```

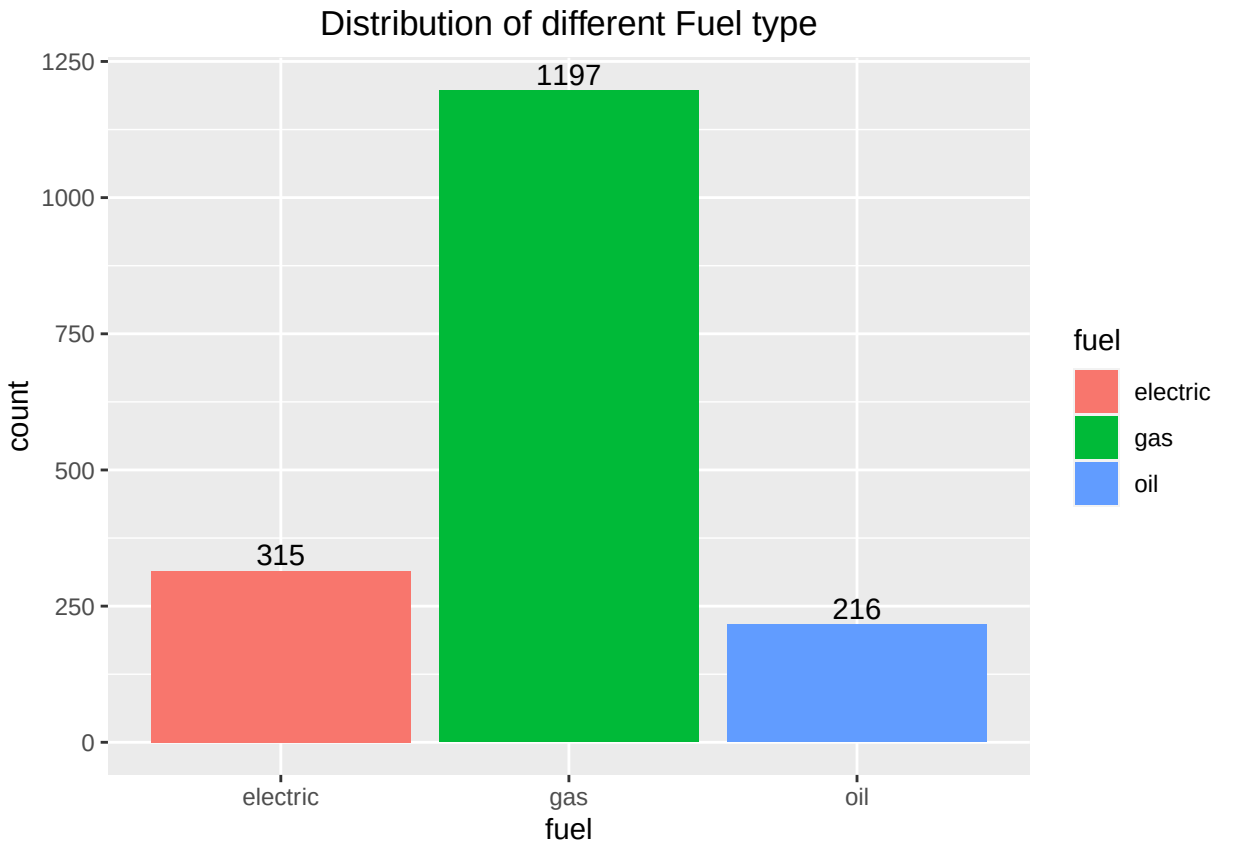


Ο πίνακας συχνοτήτων και σχετικών συχνότητων σχετικά με το αν το σπίτι διαθέτει κεντρική θέρμανση

~	CentralAir	Συχνότητα	Σχετική συχνότητα
1	Yes	635	0.632
2	no	1093	0.368

Η κατανομή για τα διάφορα είδη καυσίμου.

```
ggplot(SH, aes(x = fuel, fill = fuel )) +
geom_bar()+ ggtitle("Distribution of different Fuel type")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



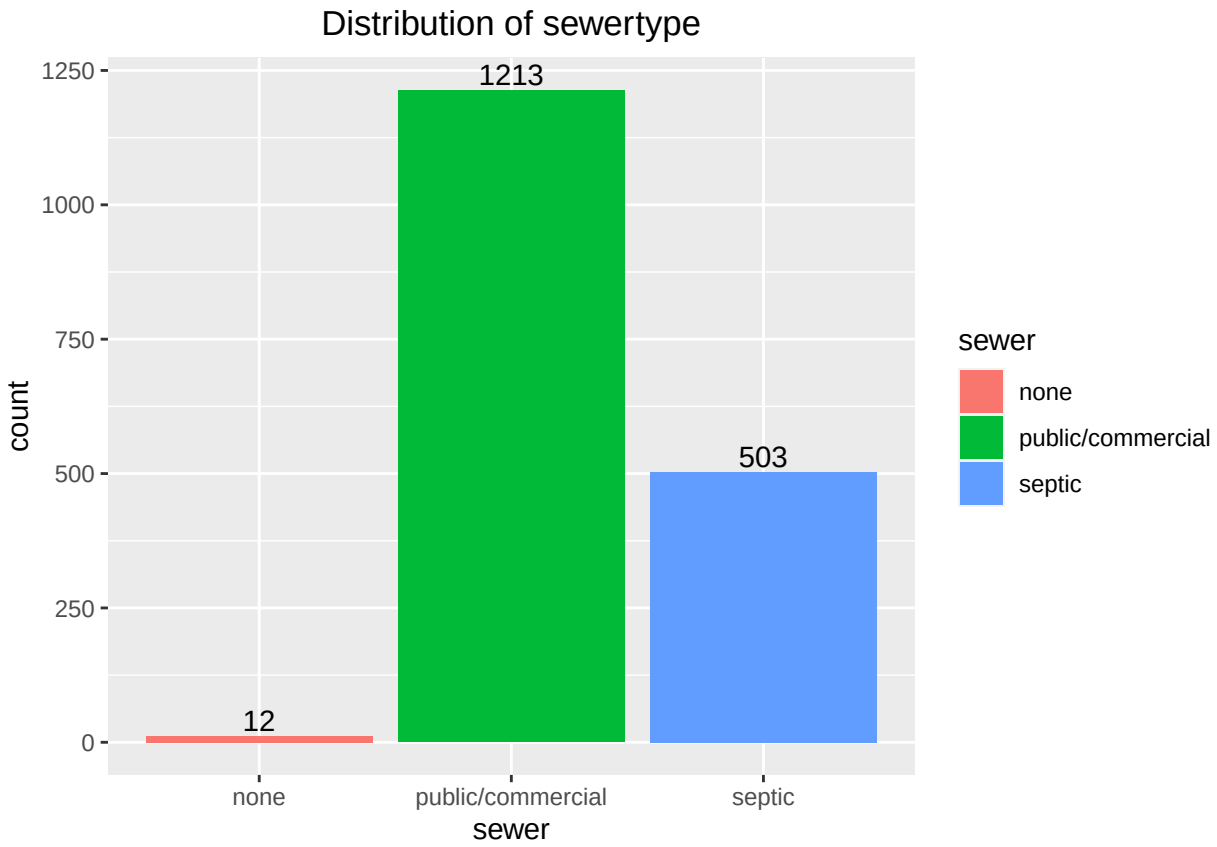
Η κατανομή για τα διάφορα είδη καυσίμου. Παρατηρούμε οτι τα περισσότερα παραδείγματα βρίσκονται στο feature:oil.

Ο πίνακας συχνοτήτων και σχετικών συχνότητων σχετικά με τις διαφορετικό τύπο θέρμανσης

~	FuelType	Συχνότητα	Σχετική συχνότητα
1	electric	315	0.1822
2	gas	1197	0.6927
3	oil	216	0.125

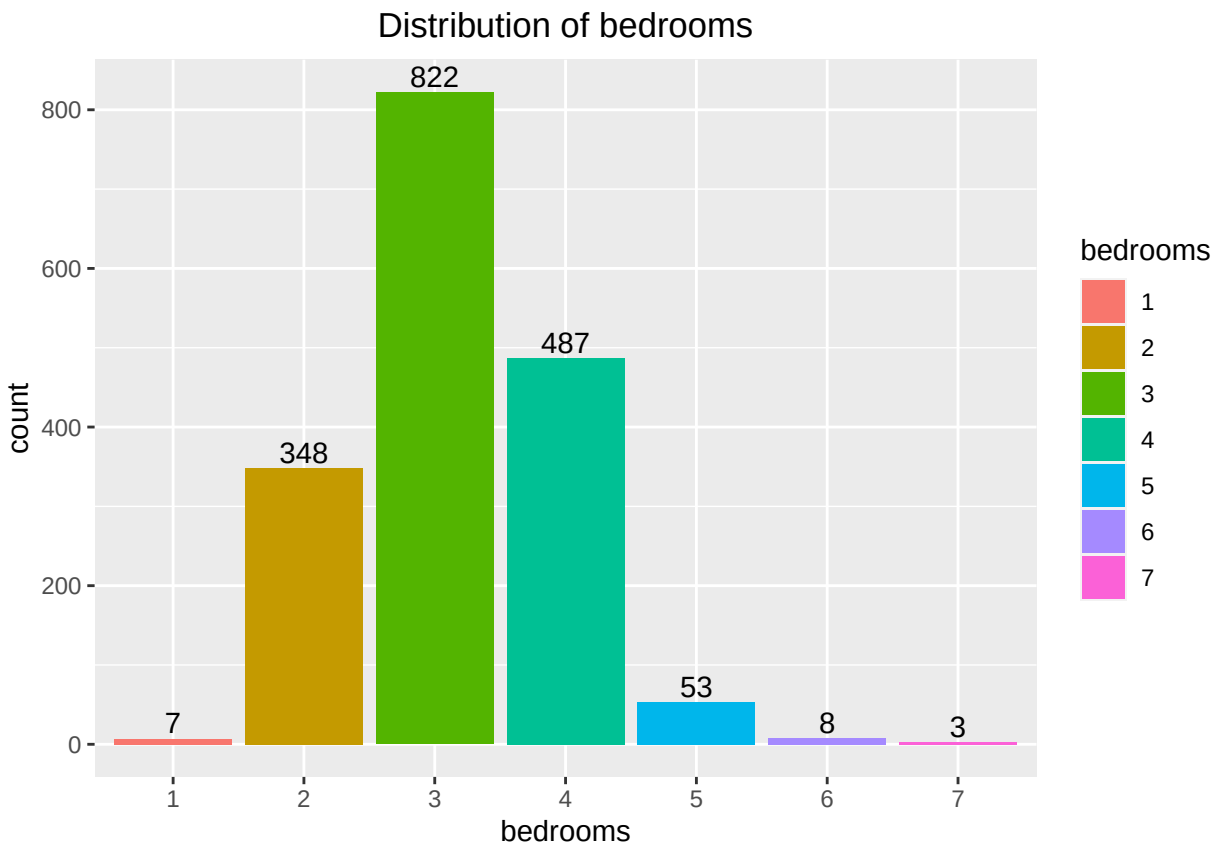
Η παρακάτω κατανομή υποδυκνύει αν σπίτια διαθέτουν δημόισιο, ιδιωτικό υπόνομο. Πολυ μικρός ο αριθμός των αγνωστων παρατηρήσεων σε αυτή την μεταβλητή.

```
ggplot(SH, aes(x = sewer, fill = sewer )) +
geom_bar()+ ggtitle("Distribution of sewertype")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



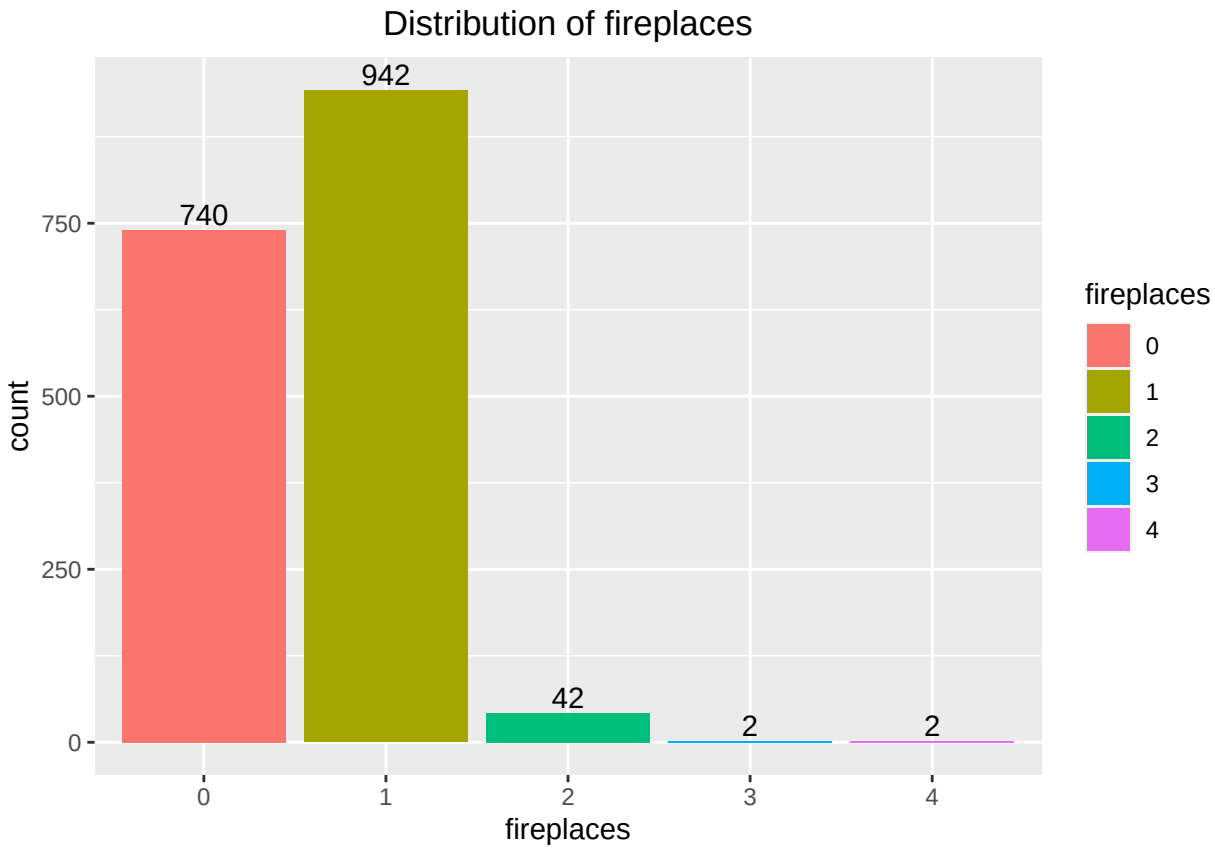
Η μεταβλητή bedrooms δεν μπορούμε να απορρίψουμε οτι ακολουθεί κανονική κατανομή

```
ggplot(SH, aes(x = bedrooms, fill = bedrooms )) +
geom_bar()+ ggtitle("Distribution of bedrooms")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



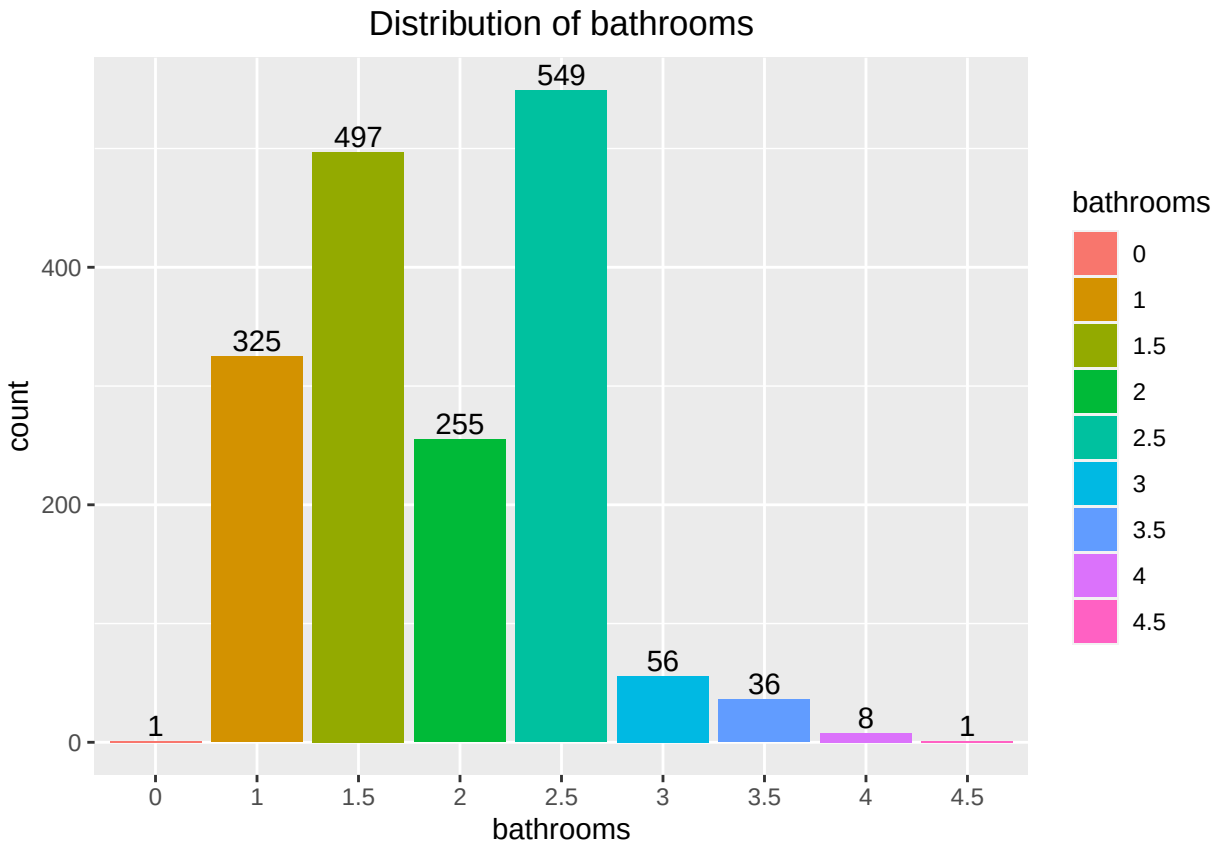
Ραβδόγραμμα για τον αριθμό των τζακιών

```
ggplot(SH, aes(x = fireplaces, fill = fireplaces )) +
geom_bar()+ ggtitle("Distribution of fireplaces")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



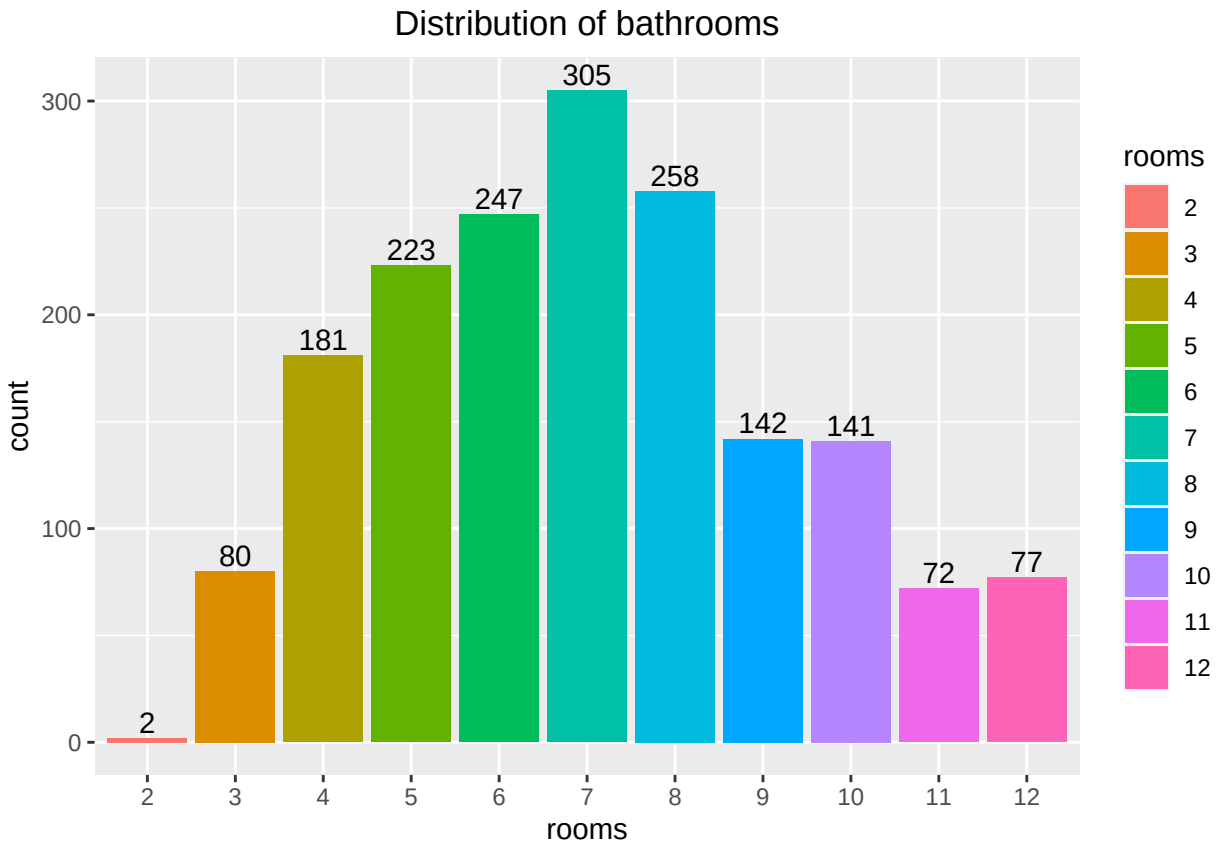
Ραβδόγραμμα για τον αριθμό των bathrooms

```
ggplot(SH, aes(x = bathrooms, fill = bathrooms )) +
geom_bar()+ ggtitle("Distribution of bathrooms")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



Ραβδόγραμμα για τον αριθμό των δωματίων.

```
ggplot(SH, aes(x = rooms, fill = rooms )) +
geom_bar()+ ggtitle("Distribution of bathrooms")+
theme(plot.title = element_text(hjust = 0.5))+
  geom_text(stat='count',aes(label=..count..),vjust=-0.25)
```



#Συσχέτιση μεταβλητών μεταξύ τους

```
library("ggcorrplot")
```

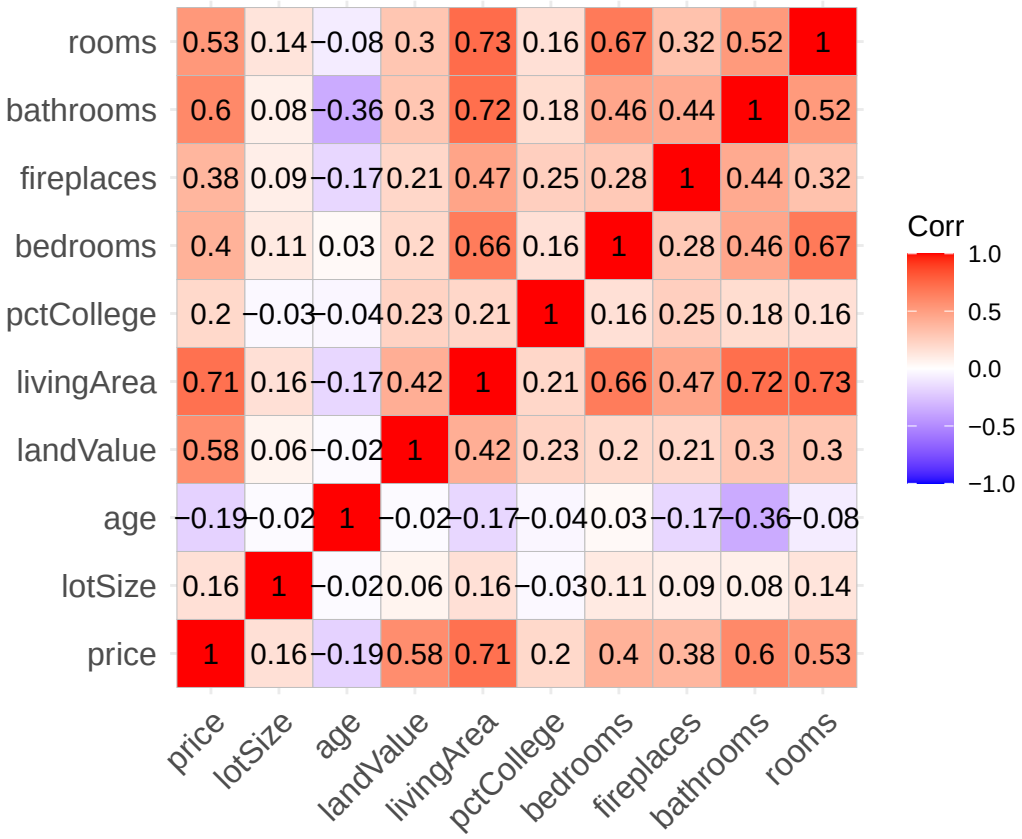
```
## Warning: package 'ggcorrplot' was built under R version 4.1.2
```

```
SHcor = data.frame(price, lotSize, age, landValue, livingArea, pctCollege, bedrooms, fireplaces, bathroo
correl = cor(SHcor)
round(correl, 2)
```

```
##             price lotSize   age landValue livingArea pctCollege bedrooms
## price        1.00    0.16 -0.19      0.58       0.71       0.20     0.40
## lotSize      0.16    1.00 -0.02      0.06       0.16      -0.03     0.11
## age         -0.19   -0.02  1.00     -0.02      -0.17      -0.04     0.03
## landValue    0.58    0.06 -0.02      1.00       0.42       0.23     0.20
## livingArea   0.71    0.16 -0.17      0.42       1.00       0.21     0.66
## pctCollege   0.20   -0.03 -0.04      0.23       0.21       1.00     0.16
## bedrooms     0.40    0.11  0.03      0.20       0.66       0.16     1.00
## fireplaces   0.38    0.09 -0.17      0.21       0.47       0.25     0.28
## bathrooms    0.60    0.08 -0.36      0.30       0.72       0.18     0.46
## rooms        0.53    0.14 -0.08      0.30       0.73       0.16     0.67
##             fireplaces bathrooms rooms
## price             0.38      0.60  0.53
## lotSize           0.09      0.08  0.14
## age              -0.17     -0.36 -0.08
```

```
## landValue        0.21      0.30  0.30
## livingArea       0.47      0.72  0.73
## pctCollege       0.25      0.18  0.16
## bedrooms         0.28      0.46  0.67
## fireplaces       1.00      0.44  0.32
## bathrooms        0.44      1.00  0.52
## rooms            0.32      0.52  1.00
```

```
ggcorrplot(correl,lab=TRUE)
```



#Συσχέτιση ανεξάρτητων μεταβλητών μεταξύ τους

Καύσιμο και τύπος θέρμανσης (heating,fuel)

Υποθέτουμε ότι θα υπάρχει κάποια σχέση ανάμεσα στον τύπο του καυσίμου και στον τύπο θέρμανσης.

```
prop.table(table(heating,fuel),2)
```

```
##                  fuel
## heating              electric         gas         oil
##   electric        0.946031746 0.005012531 0.004629630
##   hot air         0.050793651 0.802840434 0.666666667
##   hot water/steam 0.003174603 0.192147034 0.328703704
```

Φαίνεται πως υφίσταται συσχέτιση του fuel electric με τον τύπο θέρμανσης electric, του fuel gas με τον τύπο θέρμανσης ζεστού αέρα και του fuel oil με τον τύπο θέρμανσης ζεστού αέρα.

κεντρικός αερισμός και παραλία (centralAir,waterfront)

θα εξετάσουμε αν υπάρχει συσχέτιση του centralAir με το waterfront.

```
prop.table(table(centralAir,waterfront),2)
```

```
##           waterfront
## centralAir        No       Yes
##        No  0.6316404 0.7333333
##        Yes 0.3683596 0.2666667
```

```
x=as.matrix(c(64,36,73,27))
dim(x)=c(2,2)
fisher.test(x)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  x
## p-value = 0.2232
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.3438828 1.2516078
## sample estimates:
## odds ratio
##  0.6589328
```

Από το Fisher's test προκύπτει ότι p= 0.22 > 0.05 άρα η μηδενική υπόθεση δεν απορρίπτεται, δηλαδή οι μεταβλητές «centralAir» και «waterfront» είναι ανεξάρτητες (δεν έχουν σχέση).

Λοιπές αριθμητικές

Από τον πίνακα συσχέτισης φαίνεται ότι ισχυρή συσχέτιση μεταξύ τους έχουν οι ακόλουθες ανεξάρτητες αριθμητικές μεταβλητές: 1)οι rooms, bedrooms, bathrooms, fireplaces, livingArea (κάτι το οποίο είναι εύλογο αφού όλες έχουν σχέση με το μέγεθος του σπιτιού) 2)οι landValue με την livingArea

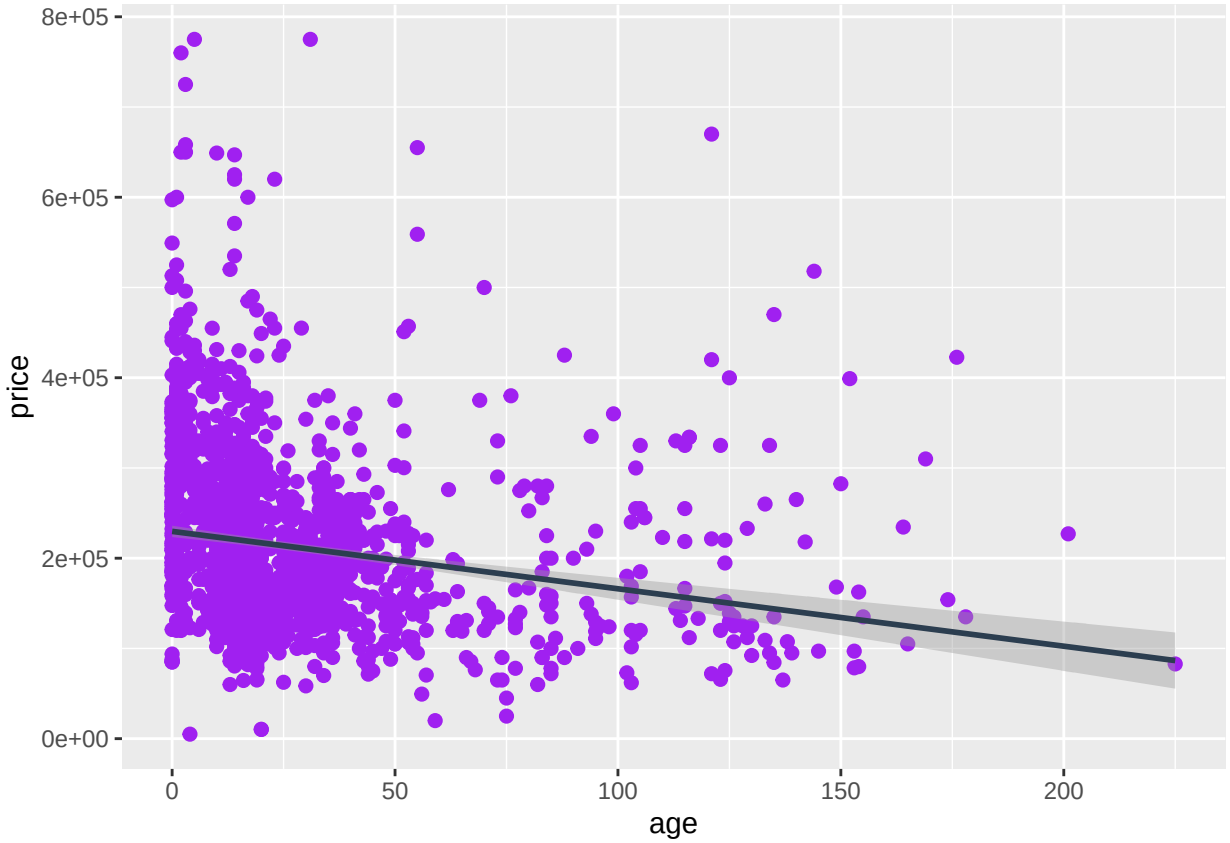
#Συσχέτιση ανεξάρτητων μεταβλητών με την εξαρτημένη

Δημιουργία scatterplots -μέσω της βιβλιοθήκης ggplot2 με σκοπό την οπτικοποίηση των σχεσεων της μεταβλητής price και των ανεξάρτητων μεταβλητών.

Price ~ Age

```
ggplot(SH, aes(x=age, y=price)) +
  geom_point(color='purple', size = 2) +
  geom_smooth(method=lm, color='#2C3E50')
```

```
## `geom_smooth()` using formula 'y ~ x'
```

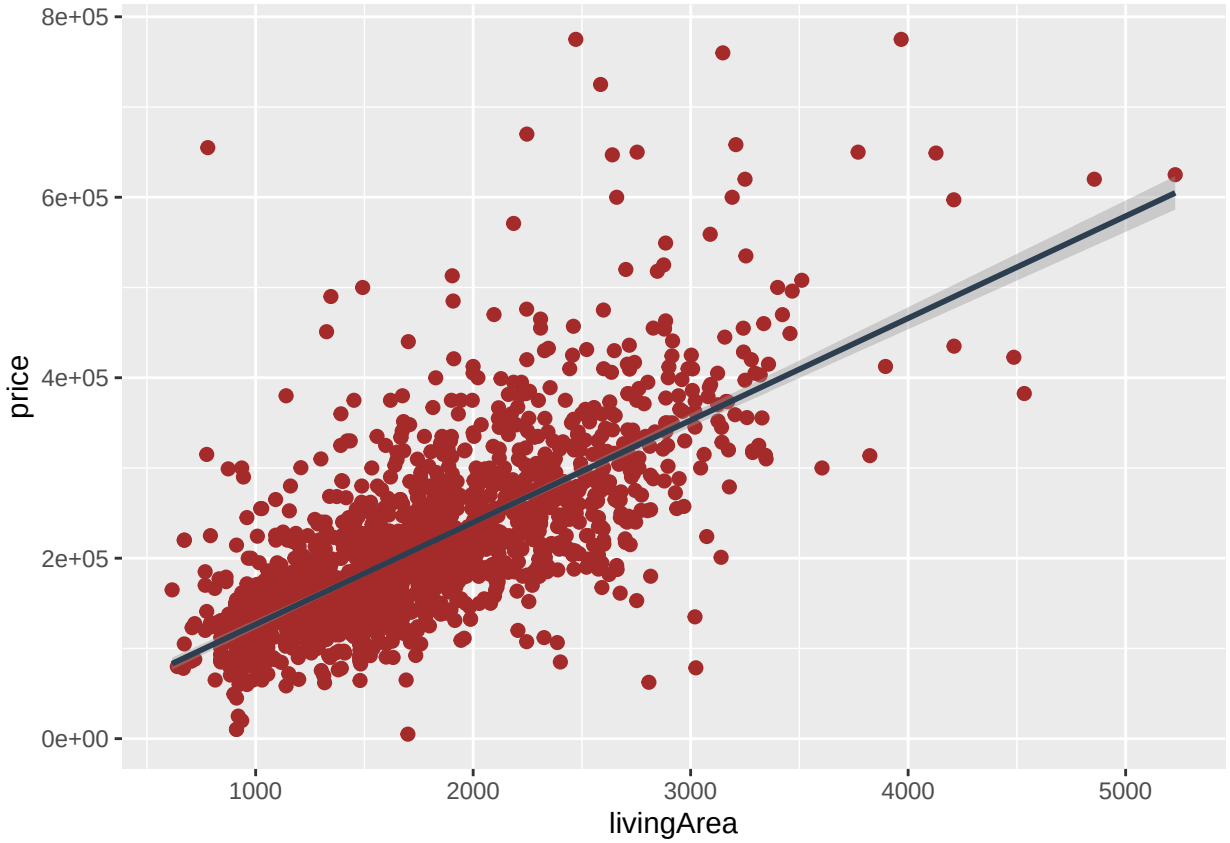


Η συσχέτιση των δυο μεταβήτών είναι αρκετά ισχυρή, καθώς υπάρχει αναλογική σταθερή μείωση των τιμών.Αυτό ειναι λογικό, άμα λάβουμε υπόψιν οτι όσο πιο παλιό είναι το σπίτι τόσο και χαμηλότερη αγοραστική αξία θα έχει

Living Area ~ Price

```
ggplot(SH, aes(x=livingArea, y=price)) +
  geom_point(color='brown', size = 2) +
  geom_smooth(method=lm, color='#2C3E50')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



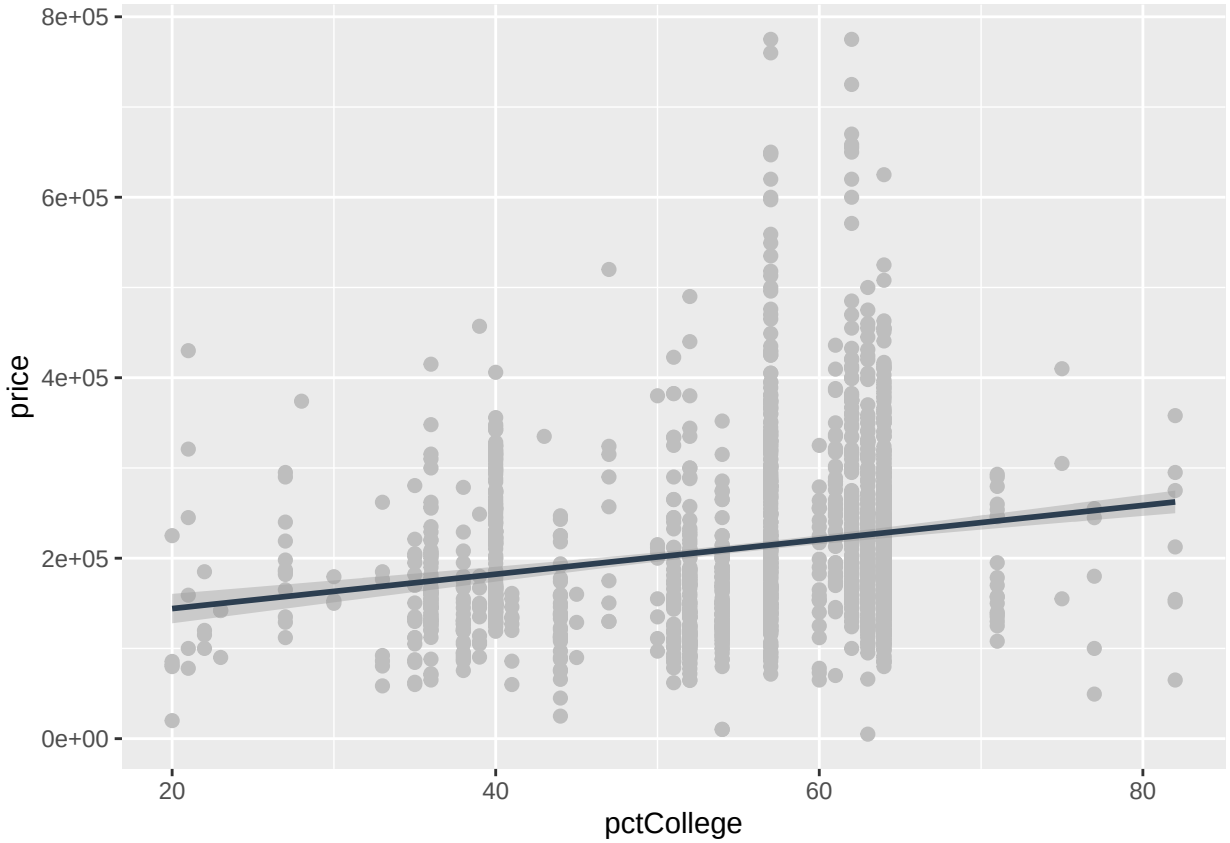
Η living Area εχει τη μεγαλύτερη θετική συσχέτιση απο κάθε άλλη μεταβλητή, καθώς ο ρυθμός με τον οποίο αυξάνονται οι τιμές των δειγμάτων ειναι παραπλήσιος.

---

pctCollege - price

```
ggplot(SH, aes(x=pctCollege, y=price)) +
  geom_point(color='grey', size = 2) +
  geom_smooth(method=lm, color='#2C3E50')
```

```
## `geom_smooth()` using formula 'y ~ x'
```

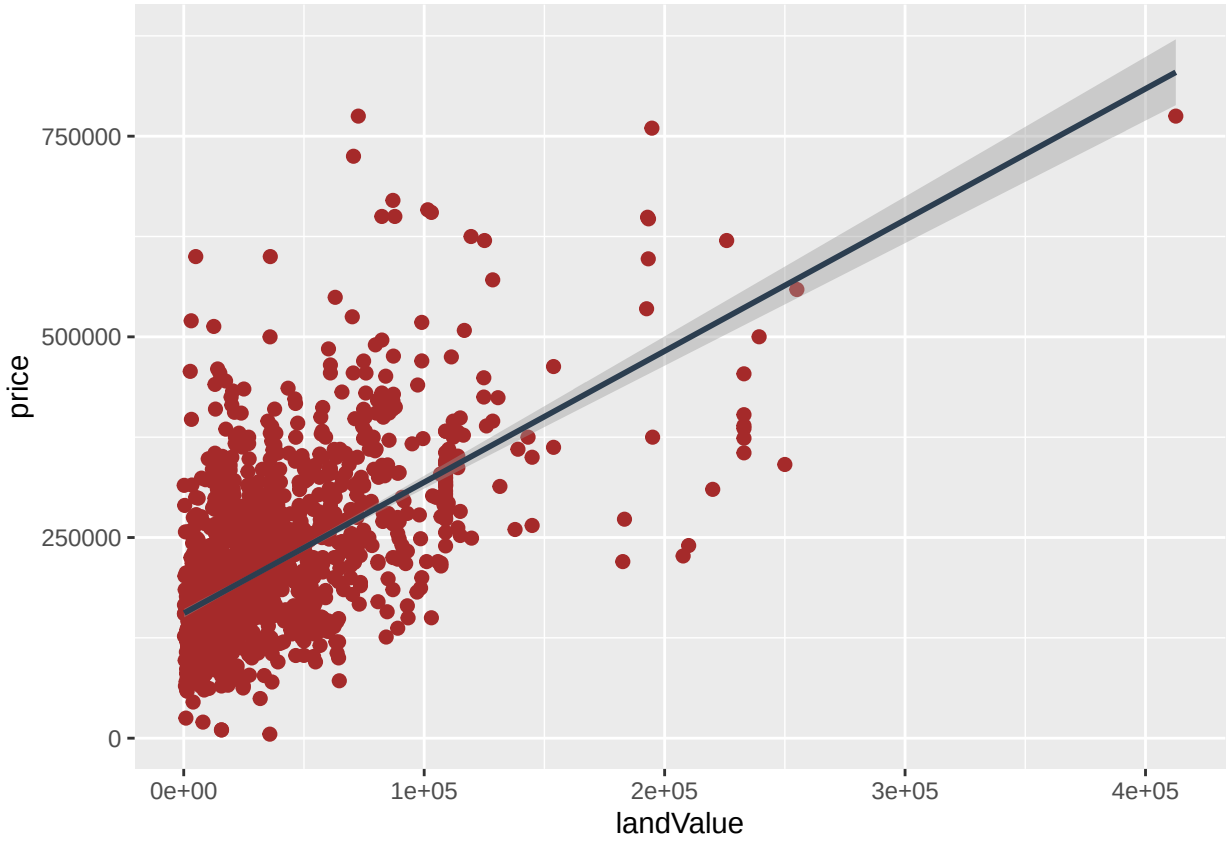


Οι παρατηρήσεις του δείγματος εχουν μεγάλη διασπορά, το οποίο μας δείχνει οτι οι δεν μεταβλητές δεν έχουν ισχυρή συσχέτιση.

landValue - price

```
ggplot(SH, aes(x=landValue, y=price)) +
  geom_point(color='brown', size = 2) +
  geom_smooth(method=lm, color='#2C3E50')
```

```
## `geom_smooth()` using formula 'y ~ x'
```

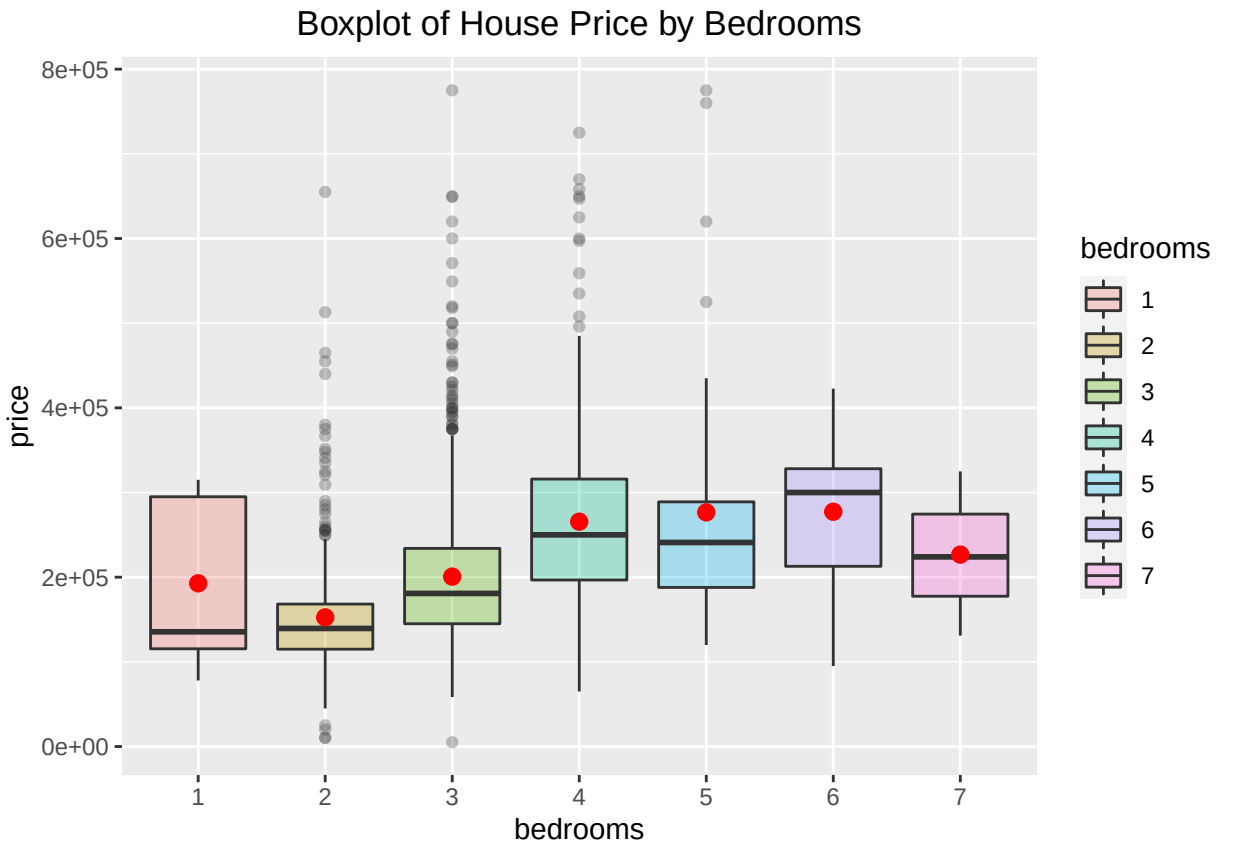


οι δύο μεταβλητές φαίνεται οτι έχουν συσχέτιση καθώς όσο αυξανεται το landValue, τείναι να αυξάνεται και η τιμή του σπιτιού.Οι περισσότερες παρατηρήσεις βρίσκονται στο διάστημα 0 - 100.000.

Θηκόγραμμα μεταξύ των μεταβλητών price-bedrooms

```
ggplot(SH, aes(x=bedrooms, y=price, fill=bedrooms)) +
  geom_boxplot(alpha=0.3) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=4, color="red", fill="red")+
  ggtitle("Boxplot of House Price by Bedrooms")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



Για τον αριθμό υπνοδωματίων:1 φαίνεται οτι υπάρχει θετική ασυμμετρία όπου η διάμεσος(το 50%-ποσοστιμόριο) πλησιάζει περισσότερο το πρώτο τεταρτημόιρο(25%-ποσοστιμόριο) δηλαδή οι περισότερες τιμές της μεταβλητής κατανέμονται ανάμεσα στην διάμεσο και το τρίτο τεταρτημόριο.

Για τον αριθμό υπνοδωματίων:6 φαίνεται οτι υπάρχει αρνητική ασυμμετρία όπου η διάμεσος(το 50%-ποσοστιμόριο) πλησιάζει περισσότερο το τρίτο τεταρτημόιρο(75%-ποσοστιμόριο) δηλαδή οι περισότερες τιμές της μεταβλητής κατανέμονται ανάμεσα στην διάμεσο και το πρώτο τεταρτημόριο.

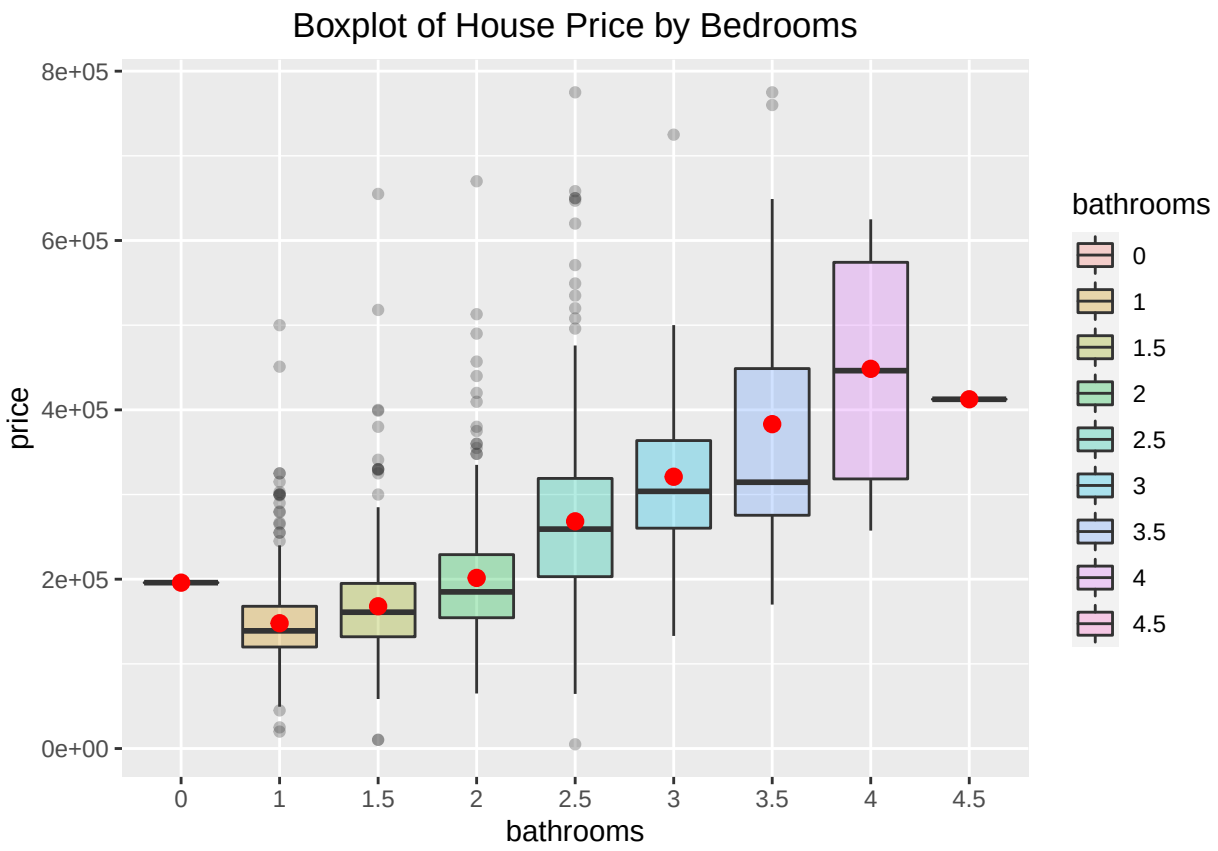
Στις υπόλοιπες περιπτώσεις το δείγμα ακολουθεί κανονική κατανομή

Για τις περιπτώσεις των υπνοδωματίων 2,3,4,5,7 υπάρχουν αρκετές εξωτερικές τιμές(παράτυπα σημεία), γεγονός που φανερώνει οτι δεν έχει τόσο σωστή δειγματοληψία ωστε το δείγμα να ειναι ανταπροσωπευτικό για τον πληθυσμό.

Θηκόγραμμα μεταξύ των μεταβλητών price-bathrooms

```
ggplot(SH, aes(x=bathrooms, y=price, fill=bathrooms)) +
  geom_boxplot(alpha=0.3) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=4, color="red", fill="red")+
  ggtitle("Boxplot of House Price by Bedrooms")+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```



Για τον αριθμό μπάνιων:3.5 φαίνεται οτι υπάρχει θετική ασυμμετρία όπου η διάμεσος(το 50%-ποσοστιμόριο) πλησιάζει περισσότερο το πρώτο τεταρτημόιρο(25%-ποσοστιμόριο) δηλαδή οι περισσότερες τιμές της μεταβλητής κατανέμονται ανάμεσα στην διάμεσο και το τρίτο τεταρτημόριο.

Επιπλέον, παρατηρούμε οτι όσο περισσότερα μπάνια διαθέτει ένα σπίτι, τόσο μεγαλύτερη είναι και η τιμή του σπιτιού.

Για τις περιπτώσεις των σπιτιών που έχουν τόσα μπάνια 1 - 3.5 υπάρχουν αρκετές εξωτερικές τιμές(παράτυπα σημεία), γεγονός που φανερώνει οτι δεν έχει τόσο σωστή δειγματοληψία ωστε το δείγμα να ειναι ανταπροσωπευτικό για τον πληθυσμό.

Μοντέλα με εξαρτημένη μεταβλητή την τιμή

Παρατηρούμε ότι υπάρχει ισχυρή θετική συσχέτιση της τιμής με τον αριθμό δωματίων, υπνοδωματίων μπάνιων, εμβαδού σπιτιού, αξίας οικοπέδου. Οι παρατηρήσεις αυτές θα είναι πολύ χρήσιμες για την ακόλουθη σχεδίαση των μοντέλων.

Θα σχεδιάσουμε 3 μοντέλα με εξαρτημένη μεταβλητή την τιμή: ένα με όλες τις ανεξάρτητες αριθμητικές, ένα με μόνο κατηγορικές και ένα και με αριθμητικές και με κατηγορικές.

#1ο Μοντέλα με εξαρτημένη μεταβλητή την price (all numeric) Στο πρώτο μοντέλο με μόνο numerical μεταβλητές θα επιλέξουμε τις 3 με το μεγαλύτερο correlation με την τιμή. Θα ξεκινήσουμε με έλεγχο των κυρίων επιδράσεων, όλων των αλληλεπιδράσεων 2ης τάξης και την αλληλεπίδραση 3ης τάξης.

```
model1=lm(price ~ livingArea*landValue*bathrooms)
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ livingArea * landValue * bathrooms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217304  -35694   -6357   27271  437889
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     6.669e+04  1.752e+04   3.806 0.000146 ***
## livingArea                      3.664e+01  1.098e+01   3.336 0.000867 ***
## landValue                       7.071e-01  3.407e-01   2.076 0.038063 *
## bathrooms                      -1.379e+04  9.450e+03  -1.459 0.144760
## livingArea:landValue            3.411e-05  1.731e-04   0.197 0.843843
## livingArea:bathrooms            2.022e+01  4.737e+00   4.269 2.07e-05 ***
## landValue:bathrooms             3.807e-01  1.678e-01   2.269 0.023405 *
## livingArea:landValue:bathrooms -1.288e-04  6.513e-05  -1.978 0.048046 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60410 on 1720 degrees of freedom
## Multiple R-squared:  0.625,  Adjusted R-squared:  0.6234
## F-statistic: 409.5 on 7 and 1720 DF,  p-value: < 2.2e-16
```

Παρατηρούμε ότι p<0.001<0.05 και επομένως το μοντέλο συνεισφέρει σημαντικά στην ερμηνεία της εξαρτημένης μεταβλητής. Επίσης εξηγεί το (R-squared) 62,5% της μεταβλητότητας της εξαρτημένης μεταβλητής(price). Οι συντελεστές livingArea, landValue, η αλληλεπίδραση των δύο με την bathrooms καθώς και η τριπλή αλληλεπίδραση τους είναι στατιστικά σημαντικοί αφού έχουν p<0.001<0.05 και επομένως συνεισφέρουν στην ερμηνεία της εξαρτημένης μεταβλητής. Αντίθετα η bathrooms και η αλληλεπίδραση των livingArea και landValue δεν είναι στατιστικά σημαντικές (έχουν p>0.05) και επομένως μπορουν να αφαιρεθούν από το μοντέλο.

```
model2=update(model1, ~. -bathrooms -livingArea:landValue)
anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ livingArea * landValue * bathrooms
## Model 2: price ~ livingArea + landValue + livingArea:bathrooms + landValue:bathrooms +
##     livingArea:landValue:bathrooms
##   Res.Df        RSS Df   Sum of Sq      F Pr(>F)
## 1   1720 6.2766e+12
## 2   1722 6.2853e+12 -2 -8713385517 1.1939 0.3033
```

```
model3=update(model2, ~. -landValue:bathrooms)
summary(model3)
```

```
##
## Call:
## lm(formula = price ~ livingArea + landValue + livingArea:bathrooms +
##     livingArea:landValue:bathrooms)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -217260  -35683   -5975   27709  437600
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     4.978e+04  6.315e+03   7.882 5.68e-15 ***
## livingArea                      3.824e+01  6.111e+00   6.259 4.89e-10 ***
## landValue                       1.203e+00  8.406e-02  14.312  < 2e-16 ***
## livingArea:bathrooms            1.691e+01  1.739e+00   9.721  < 2e-16 ***
## livingArea:landValue:bathrooms -4.901e-05  1.285e-05  -3.814 0.000141 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60470 on 1723 degrees of freedom
## Multiple R-squared:  0.6235, Adjusted R-squared:  0.6226
## F-statistic: 713.4 on 4 and 1723 DF,  p-value: < 2.2e-16
```

```
anova(model2,model3)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ livingArea + landValue + livingArea:bathrooms + landValue:bathrooms +
##     livingArea:landValue:bathrooms
## Model 2: price ~ livingArea + landValue + livingArea:bathrooms + livingArea:landValue:bathrooms
##   Res.Df        RSS Df   Sum of Sq      F  Pr(>F)
## 1   1722 6.2853e+12
## 2   1723 6.3008e+12 -1 -1.5531e+10 4.2551 0.03928 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model4=update(model3, ~. -livingArea:landValue:bathrooms)
anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ livingArea * landValue * bathrooms
## Model 2: price ~ livingArea + landValue + livingArea:bathrooms + landValue:bathrooms +
##     livingArea:landValue:bathrooms
##   Res.Df        RSS Df   Sum of Sq      F Pr(>F)
## 1   1720 6.2766e+12
## 2   1722 6.2853e+12 -2 -8713385517 1.1939 0.3033
```

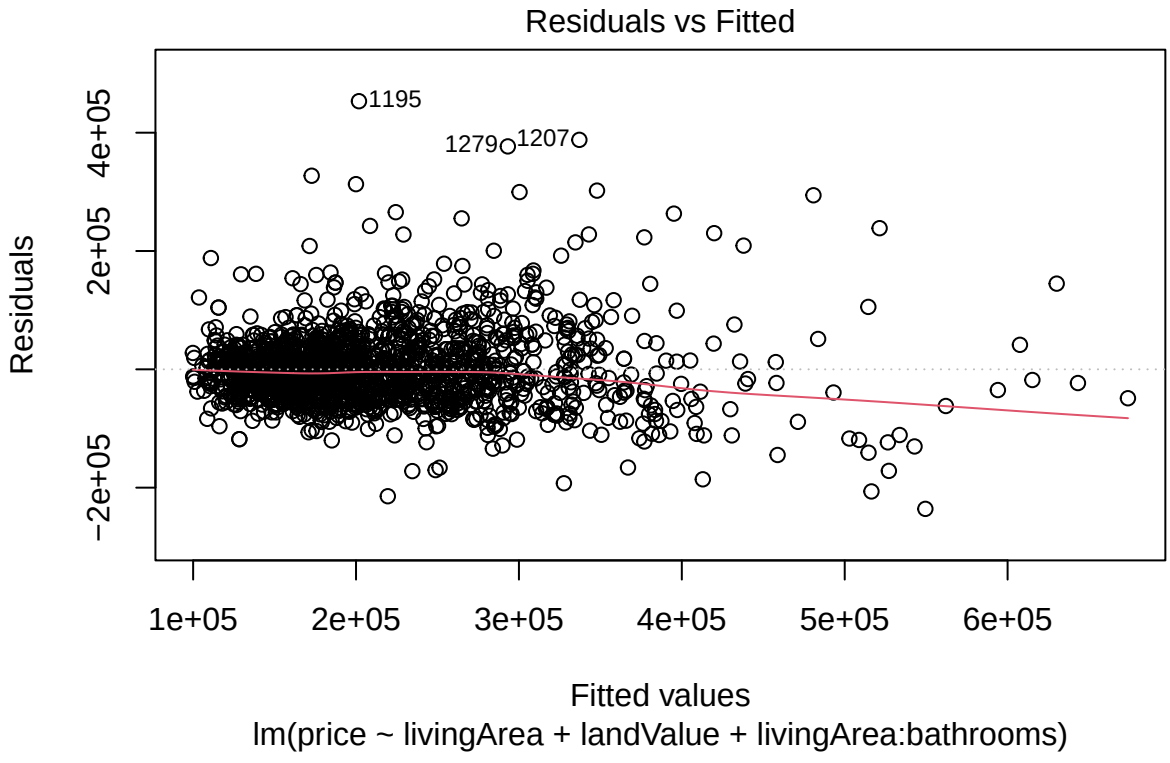
```
summary(model4)
```

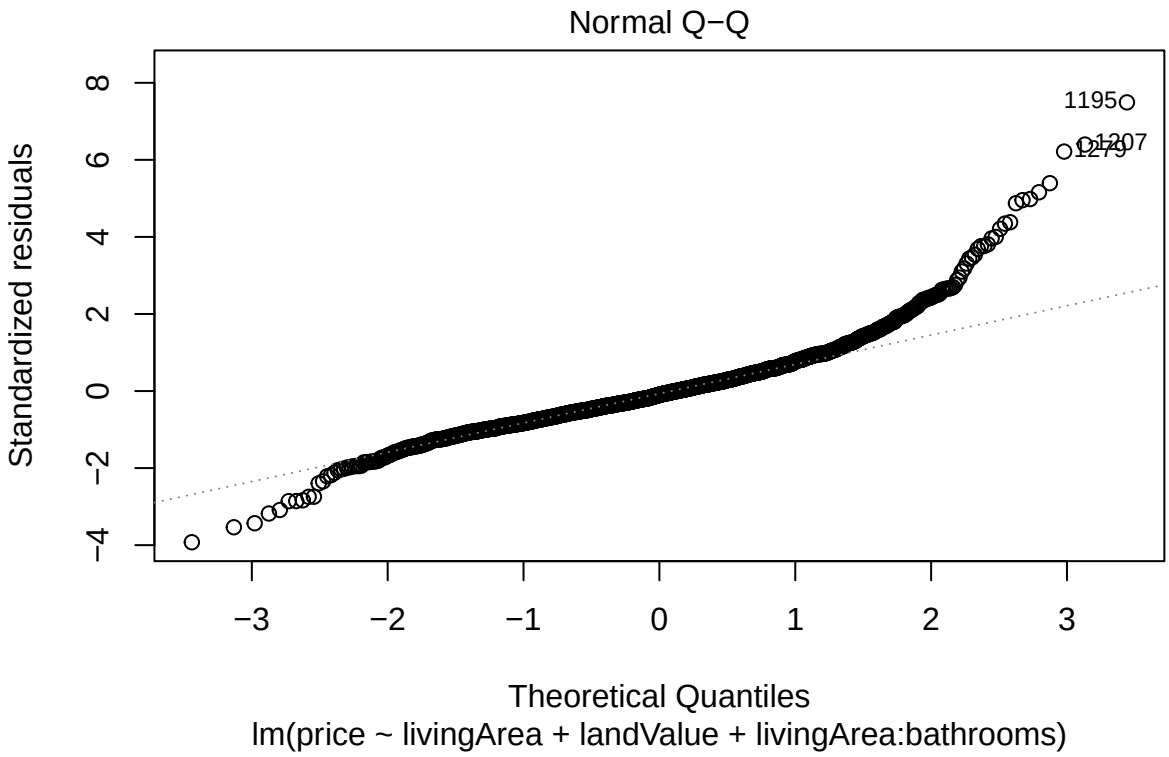
```
##
## Call:
## lm(formula = price ~ livingArea + landValue + livingArea:bathrooms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

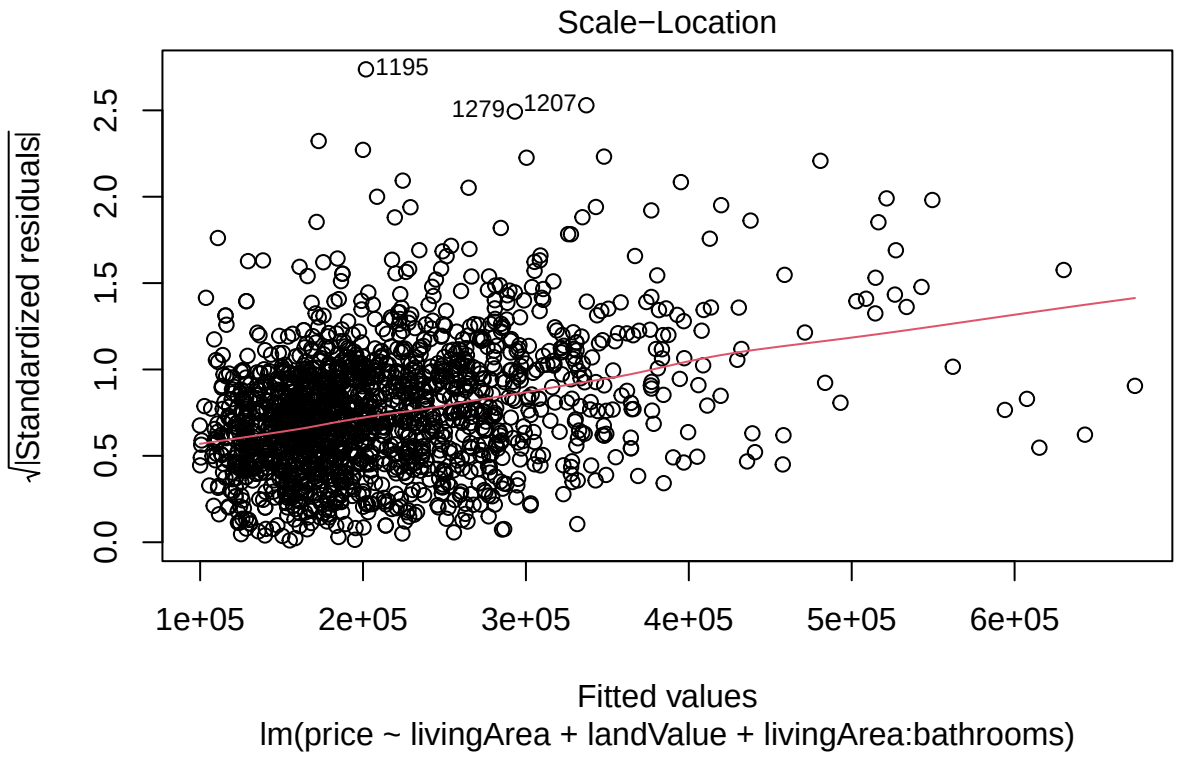
```
## -235866  -35095   -5845   26977  453197
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.730e+04  6.023e+03   9.513  < 2e-16 ***
## livingArea              4.102e+01  6.091e+00   6.734 2.25e-11 ***
## landValue               9.346e-01  4.612e-02  20.265  < 2e-16 ***
## livingArea:bathrooms 1.389e+01  1.555e+00   8.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60710 on 1724 degrees of freedom
## Multiple R-squared:  0.6203, Adjusted R-squared:  0.6197
## F-statistic: 938.9 on 3 and 1724 DF,  p-value: < 2.2e-16
```

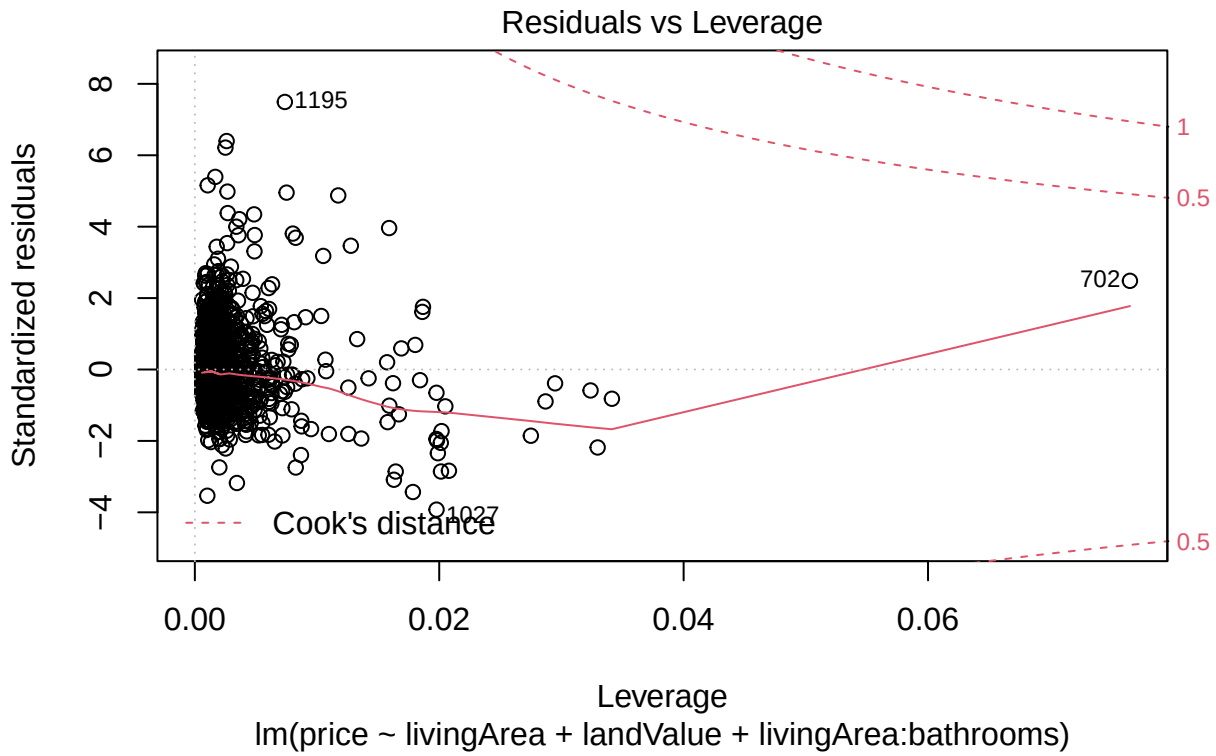
Παρατηρούμε ότι τα μοντέλα 1,2,3 δεν έχουν στατιστικά σημαντική διαφορά μεταξύ τους ενώ το μοντέλο 3 με το 4 έχουν στατιστικά σημαντική διαφορά (p<0,001). Ωστόσο το μοντέλο 4 ίσως να είναι προτιμητέο καθώς είναι λιγότερο περίπλοκο (αφού έχει αφαιρεθεί και η τριπλή αλληλεπίδραση παρά το ότι είναι στατιστικά σημαντική) και εξακολουθεί να συνεισφέρει σημαντικά στην ερμηνεία της εξαρτημένης μεταβλητής και να εξηγεί το (R-squared) σε παραπλήσιο βαθμό (62%) την μεταβλητότητας της εξαρτημένης μεταβλητής(price).

```
plot(model4)
```




Normal Q−Q
Standardized residuals
8
6
4
2
0
−2
−4
1195
1207
1279
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(price ~ livingArea + landValue + livingArea:bathrooms)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
1195
1279
1207
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
Fitted values
lm(price ~ livingArea + landValue + livingArea:bathrooms)



Παρατηρούμε ότι υφίσταται κάποιο pattern στα υπόλοιπα (ύπαρξη ετεροσκεδαστικότητας).Ακόμα η κατανομή τους ξεφεύγει από την κανονική (QQplot). Θα πειραματιστούμε λοιπόν με το λογαριθμικό μετασχηματισμό της εξαρτημένης μεταβλητής και των livingArea και landValue.

```
model5=lm(log(price) ~ log(livingArea) + log(landValue) + livingArea:bathrooms)
```

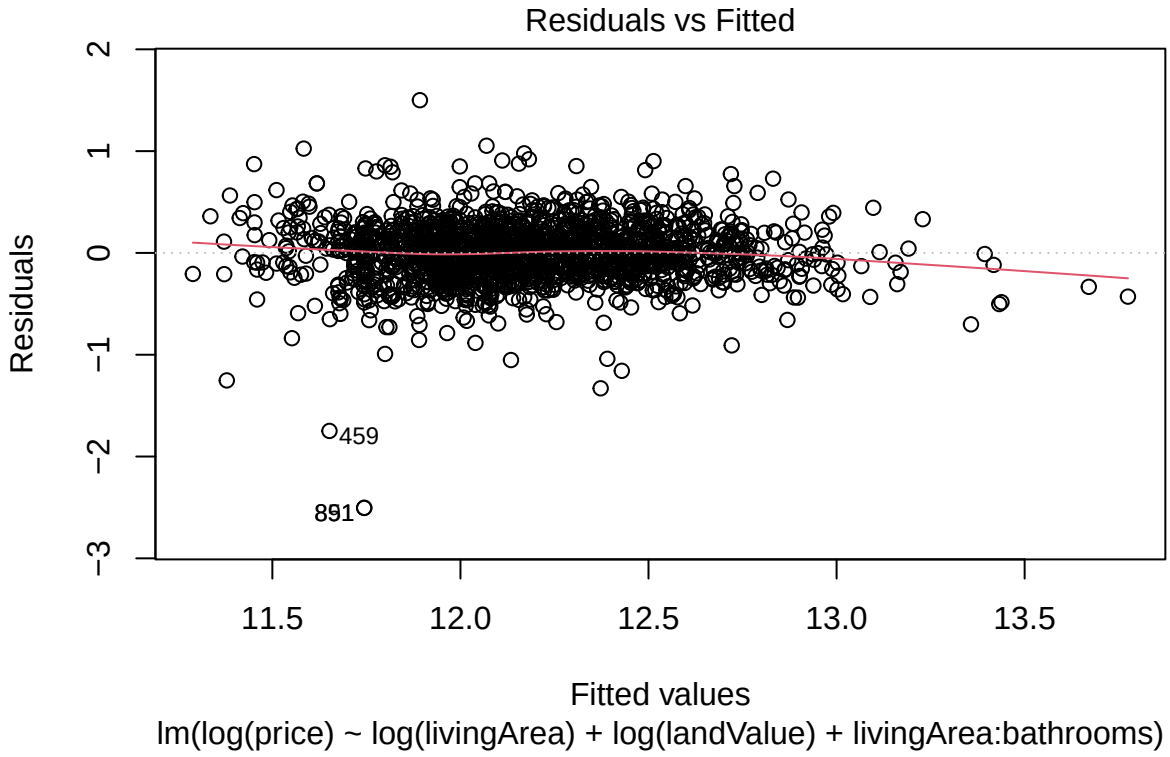
Το πρόβλημα της ετεροσκεδαστικότητας έχει αντιμετωπιστεί και η κατανομή των καταλοίπων είναι σημαντικά βελτιωμένη, προσεγγίζοντας την κανονική. Σε όλα τα σχήματα φαίνεται πως η παρατήρηση #1011 έχει άσχημη επίδραση στην παλινδρόμηση και επομένως στο τελικό μοντέλο θα αφαιρεθεί.

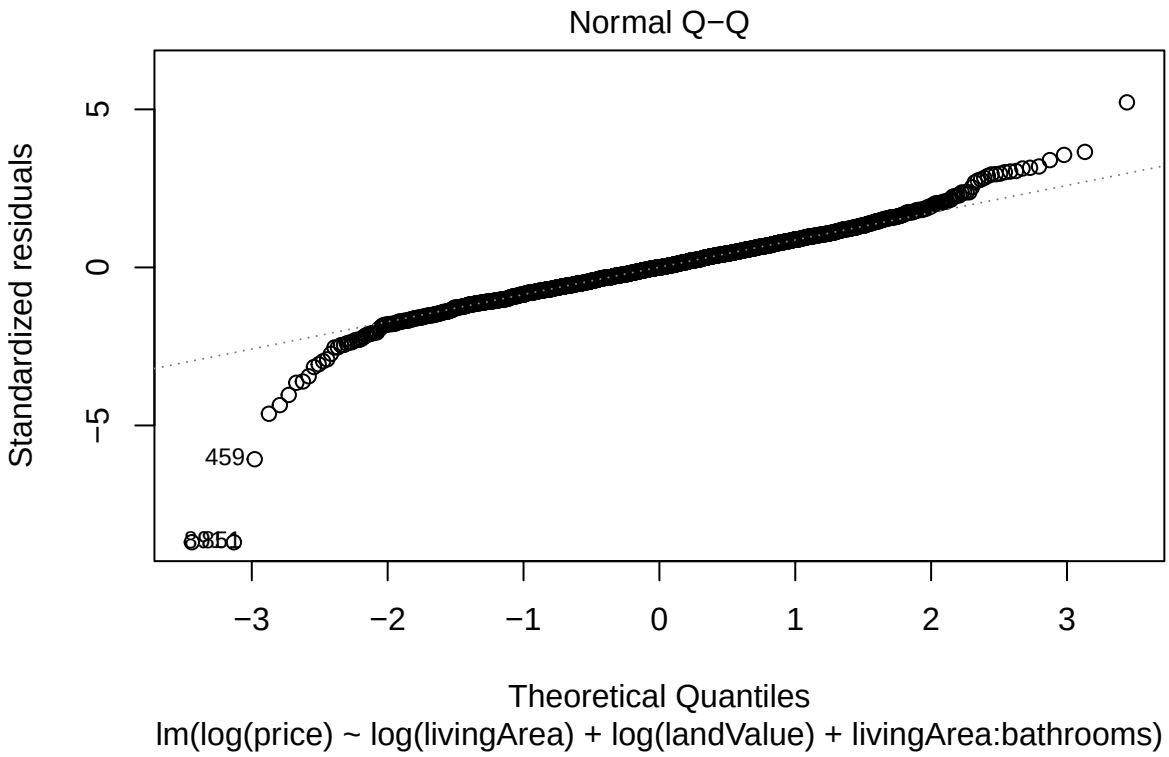
```
model6=lm(log(price) ~ log(livingArea) + log(landValue) + livingArea:bathrooms, subset=(1:length(price)
summary(model6)
```

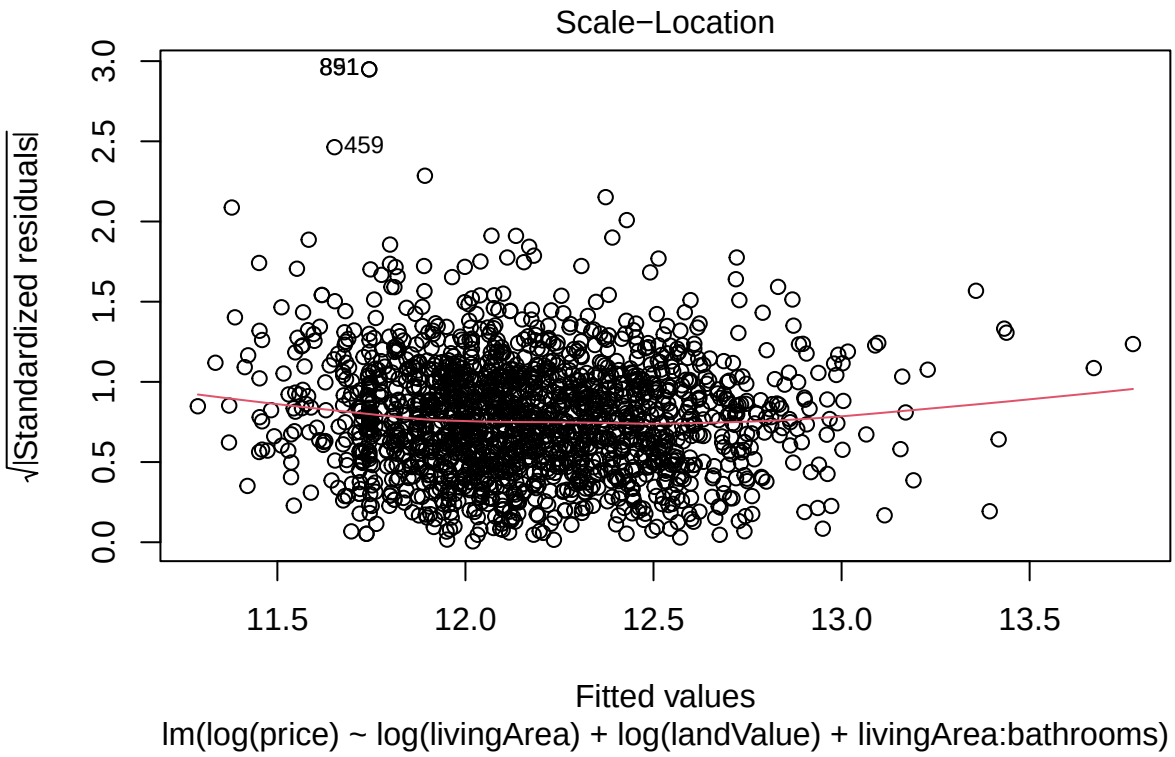
```
##
## Call:
## lm(formula = log(price) ~ log(livingArea) + log(landValue) +
##     livingArea:bathrooms, subset = (1:length(price) != 1011))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50429 -0.16588  0.00001  0.16933  1.50013
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.477e+00  2.847e-01  26.262   <2e-16 ***
```

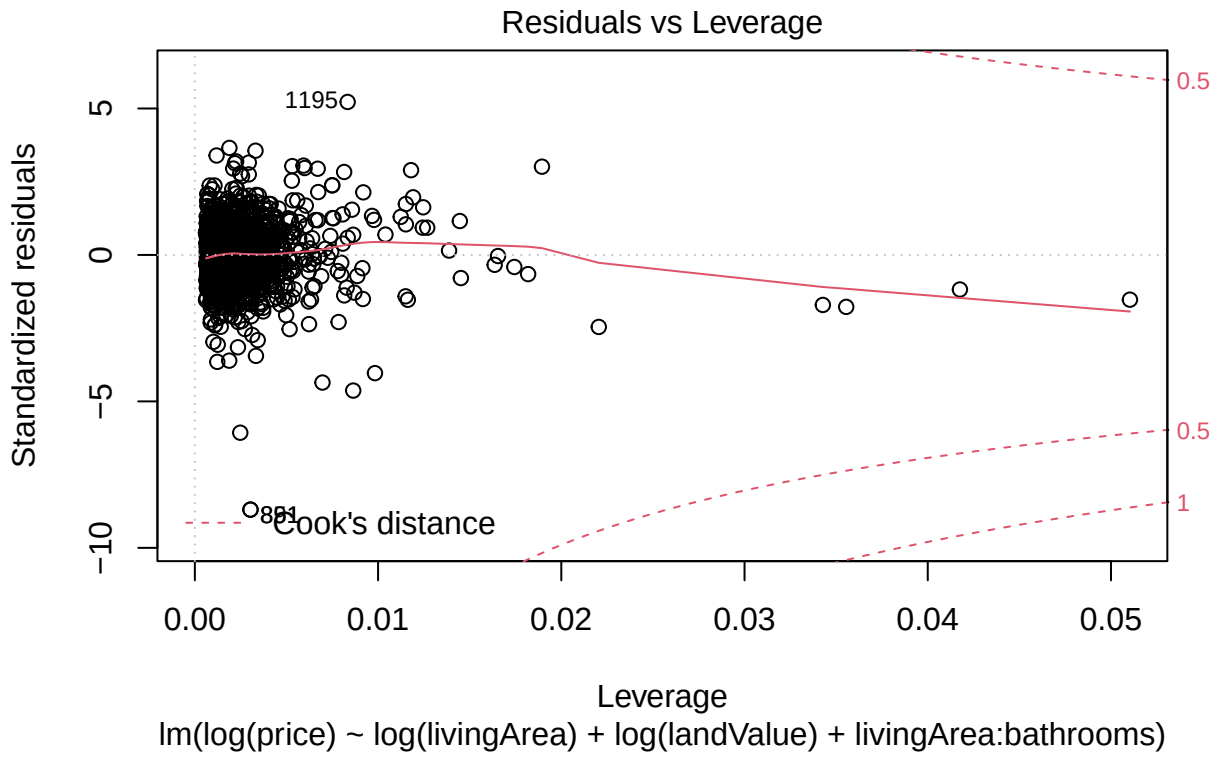
```
## log(livingArea)      4.440e-01  4.101e-02  10.827   <2e-16 ***
## log(landValue)       1.211e-01  7.385e-03  16.393   <2e-16 ***
## livingArea:bathrooms 5.165e-05  5.881e-06   8.781   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2885 on 1723 degrees of freedom
## Multiple R-squared:  0.5793, Adjusted R-squared:  0.5785
## F-statistic: 790.8 on 3 and 1723 DF,  p-value: < 2.2e-16
```

```
plot(model6)
```




Normal Q−Q
Standardized residuals
5
0
−5
459
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(price) ~ log(livingArea) + log(landValue) + livingArea:bathrooms)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
891
459
11.5
12.0
12.5
13.0
13.5
Fitted values
lm(log(price) ~ log(livingArea) + log(landValue) + livingArea:bathrooms)



Παρατηρούμε ότι το τελικό μοντέλο έχει στατιστικά σημαντική ερμηνευτική ισχύ της τιμής των σπιτιών (εξαρτημένη μεταβλητή), και καλή εξήγηση της μεταβλητότητας της (~58%). Και οι τρεις όροι είναι στατιστικά σημαντικοί και οι θετικοί συντελεστές στην livingArea, στην landValue και στην αλληλεπίδραση των livingArea και bathrooms, σημαίνει ότι όταν έχουμε αύξηση στους 3 όρους κατά μια μονάδα ασκείται ανάλογη με το συντελεστή αυξητική συνιστώσα στην τιμή.

#2ο Μοντέλα με εξαρτημένη μεταβλητή την price (all categorical)

Διερευνήσαμε την ανάλυση διακύμανσης για όλες τις κατηγορικές μεταβλητές μας (heating, fuel, sewer, waterfront, newConstruction, centralAir) αλλά και για τις numeric που μπορούν να εκληφθούν ώς ordinal (rooms, bedrooms, fireplace, bathrooms) και βρέθηκε στατιστικά σημαντική επίδραση όλων των παραγόντων στην τιμή.

Ενδεικτικά:

Δωμάτια - rooms

```
library("DescTools")
```

```
## Warning: package 'DescTools' was built under R version 4.1.2
```

```
rooms3=as.factor(rooms)
model30=aov(price~rooms3)
summary(model30)
```

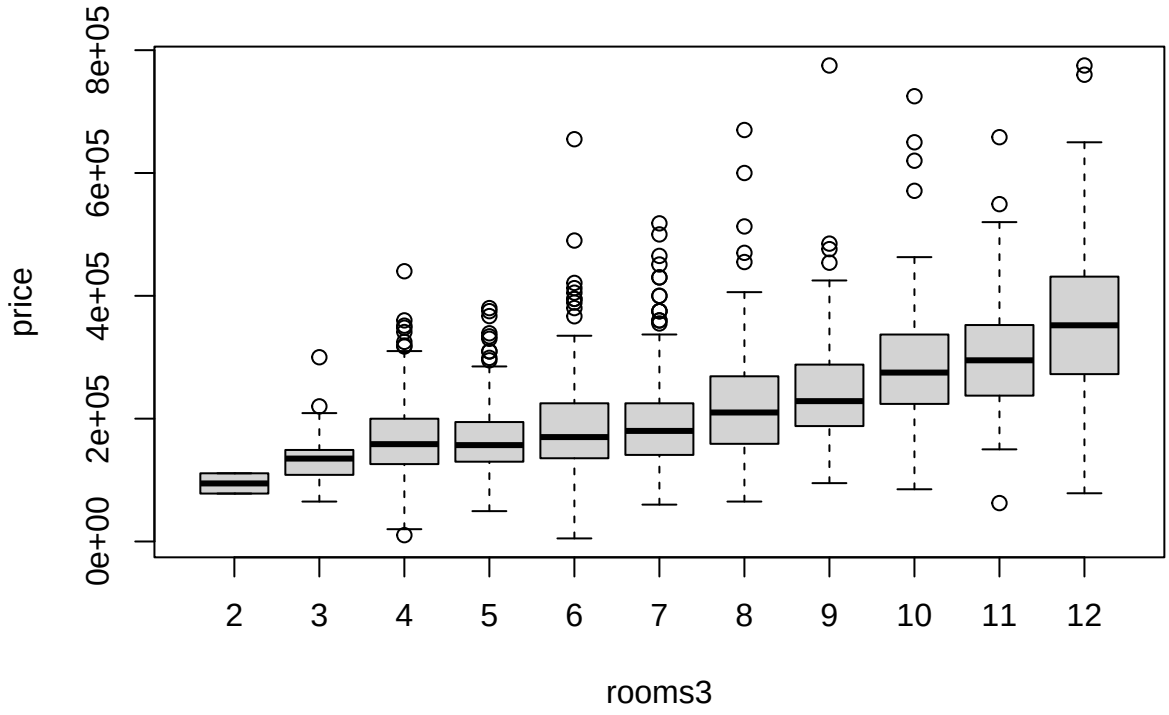
```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## rooms3        10 5.233e+12 5.233e+11   78.12 <2e-16 ***
## Residuals   1717 1.150e+13 6.699e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PostHocTest(model30)
```

```
##
##   Posthoc multiple comparisons of means : Tukey HSD
##     95% family-wise confidence level
##
## $rooms3
##             diff       lwr.ci    upr.ci    pval
## 3-2    39656.225 -149196.8398 228509.29 0.99986
## 4-2    74417.724 -113145.7908 261981.24 0.97236
## 5-2    72797.610 -114572.7677 260167.99 0.97621
## 6-2    90813.660  -96475.7845 278103.10 0.89760
## 7-2    97328.980  -89817.3745 284475.34 0.84776
## 8-2   126096.814  -61160.5586 313354.19 0.52646
## 9-2   150778.796  -37066.0047 338623.60 0.25619
## 10-2  194067.206    6213.1538 381921.26 0.03590 *
## 11-2  211413.736   22304.9460 400522.53 0.01433 *
## 12-2  278718.818   89776.0441 467661.59 0.00011 ***
## 4-3    34761.499    -655.6075  70178.61 0.05977 .
## 5-3    33141.385   -1238.2288  67521.00 0.07045 .
## 6-3    51157.435   17221.6805  85093.19 6.9e-05 ***
## 7-3    57672.755   24535.8037  90809.71 1.3e-06 ***
## 8-3    86440.589   52682.2860 120198.89 3.9e-12 ***
## 9-3   111122.571   74244.8201 148000.32 3.9e-12 ***
## 10-3  154410.981  117486.1350 191335.83 3.8e-12 ***
## 11-3  171757.511  128903.8773 214611.14 3.8e-12 ***
## 12-3  239062.593  196947.6157 281177.57 3.8e-12 ***
## 5-4    -1620.114  -28012.3426  24772.11 1.00000
## 6-4    16395.936   -9415.4437  42207.32 0.61541
## 7-4    22911.257   -1840.5003  47663.01 0.09936 .
## 8-4    51679.090   26101.4648  77256.72 5.3e-09 ***
## 9-4    76361.072   46788.1047 105934.04 3.9e-12 ***
## 10-4  119649.482   90017.8074 149281.16 3.8e-12 ***
## 11-4  136996.012  100239.7090 173752.32 3.8e-12 ***
## 12-4  204301.094  168408.7104 240193.48 3.8e-12 ***
## 6-5    18016.050   -6352.2438  42384.34 0.37824
## 7-5    24531.370    1288.3956  47774.35 0.02845 *
## 8-5    53299.204   29178.6458  77419.76 9.0e-11 ***
## 9-5    77981.186   49658.9895 106303.38 3.9e-12 ***
## 10-5  121269.596   92886.1051 149653.09 3.8e-12 ***
## 11-5  138616.126  102858.4383 174373.81 3.8e-12 ***
## 12-5  205921.208  171052.1728 240790.24 3.8e-12 ***
## 7-6     6515.320  -16065.9434  29096.58 0.99769
## 8-6    35283.154   11799.5655  58766.74 7.4e-05 ***
## 9-6    59965.136   32183.4079  87746.86 2.8e-10 ***
## 10-6  103253.546   75409.3337 131097.76 3.8e-12 ***
## 11-6  120600.076   85268.9306 155931.22 3.8e-12 ***
## 12-6  187905.158  153473.6721 222336.64 3.8e-12 ***
```

```
## 8-7    28767.834    6454.1369  51081.53 0.00170 **
## 9-7    53449.815   26649.6934  80249.94 9.2e-09 ***
## 10-7   96738.225   69873.3360 123603.11 3.8e-12 ***
## 11-7  114084.756   79520.1499 148649.36 3.9e-12 ***
## 12-7  181389.838  147745.3826 215034.29 3.8e-12 ***
## 9-8    24681.982   -2882.7056  52246.67 0.12825
## 10-8   67970.392   40342.7293  95598.05 4.1e-12 ***
## 11-8   85316.922   50156.1851 120477.66 4.4e-12 ***
## 12-8  152622.004  118365.4016 186878.61 3.8e-12 ***
## 10-9   43288.410   11925.4826  74651.34 0.00047 ***
## 11-9   60634.940   22469.2108  98800.67 1.8e-05 ***
## 12-9  127940.022   90605.5838 165274.46 3.8e-12 ***
## 11-10  17346.530  -20864.7068  55557.77 0.93159
## 12-10  84651.613   47270.6543 122032.57 2.8e-11 ***
## 12-11  67305.082   24057.8192 110552.34 3.1e-05 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(price~rooms3)
```

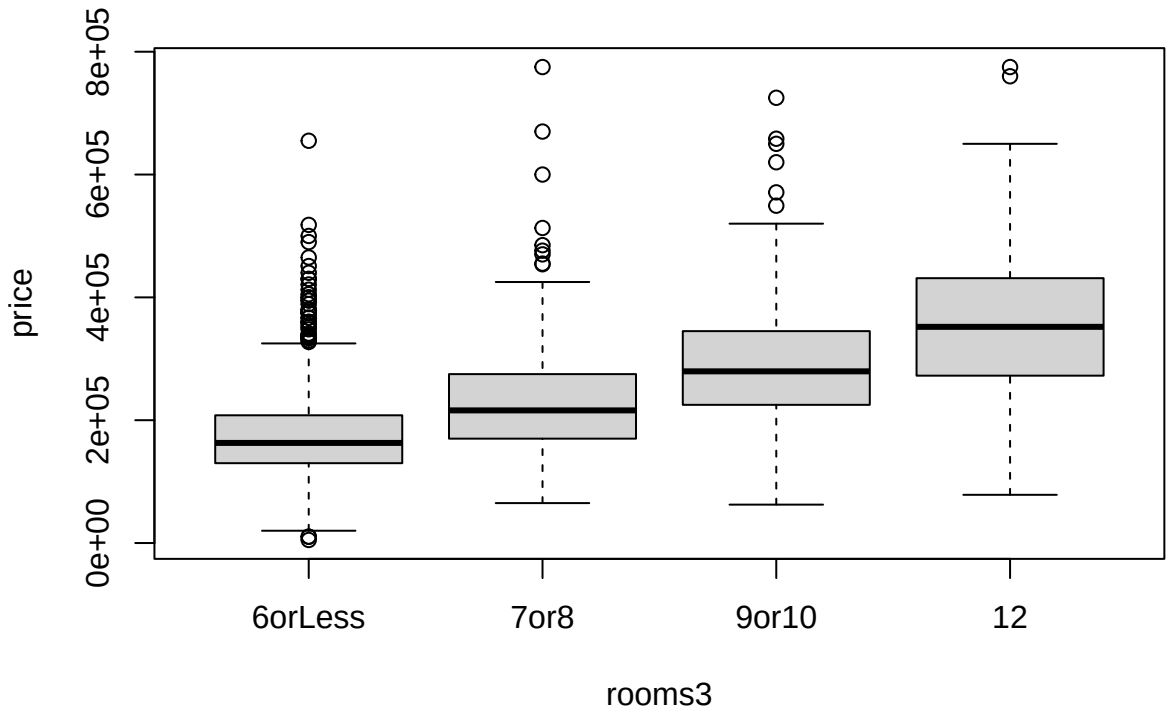


Η ανάλυση διακύμανσης δείχνει στατιστικά σημαντική επίδραση του παράγοντα rooms στην τιμή (p<0,001). Ο Post hoc έλεγχος Tukey HSD δείχνει ότι δεν υπάρχει διαφορά ως προς την τιμή ανάμεσα στον αριθμό δωματίων 1 έως 6, 7 με 8, 9 με 10. Επομένως μπούμε να χωρίσουμε τον παράγοντα στις κατηγορίες: μεχρι6, 7ή8, 9ή10, 11. Το παραπάνω συμπέρασμα γίνεται και οπτικά κατανοητό από το σχετικό boxplot.

```
levels(rooms3)[c(1:6)]= "6orLess"
levels(rooms3)[c(2:3)]= "7or8"
levels(rooms3)[c(3:4)]= "9or10"
model30=aov(price~rooms3)
summary.aov(model30)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## rooms3         3 4.886e+12 1.629e+12     237 <2e-16 ***
## Residuals   1724 1.185e+13 6.873e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(price~rooms3)
```




```
PostHocTest(model30)
```

```
##
##   Posthoc multiple comparisons of means : Tukey HSD
##     95% family-wise confidence level
##
## $rooms3
##                    diff    lwr.ci    upr.ci    pval
## 7or8-6orLess    52978.1  40431.40  65524.80 4.3e-12 ***
## 9or10-6orLess 118050.0 102013.16 134086.85 4.3e-12 ***
```

```
## 12-6orLess    196838.0 171657.05 222018.96 4.3e-12 ***
## 9or10-7or8     65071.9  46988.12  83155.68 4.3e-12 ***
## 12-7or8       143859.9 117328.35 170391.45 4.3e-12 ***
## 12-9or10       78788.0  50438.68 107137.32 1.2e-11 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.lm(model30)
```

```
##
## Call:
## aov(formula = price ~ rooms3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -294719  -52364  -13531   38612  545641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   176381       2573   68.54   <2e-16 ***
## rooms37or8     52978       4879   10.86   <2e-16 ***
## rooms39or10   118050       6236   18.93   <2e-16 ***
## rooms312      196838       9792   20.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82910 on 1724 degrees of freedom
## Multiple R-squared:  0.292,  Adjusted R-squared:  0.2907
## F-statistic:   237 on 3 and 1724 DF,  p-value: < 2.2e-16
```

Η ανάλυση διακύμανσης του απλοποιημένου και βελτιωμένου μοντέλου εξακολουθεί δείχνει στατιστικά σημαντική επίδραση του παράγοντα rooms στην τιμή (p<0,001). Από το PostHocTest βλέπουμε ότι όλα τα επίπεδα έχουν στατιστικά σημαντική διαφορά μεταξύ τους καθώς και την διαφορά που παρατηρείται στη μέση τιμή του σπιτιού όταν τα δωμάτια αυξάνονται από το κάθε επίπεδο στο άλλο (ενδεικτικά όταν τα δωμάτια αυξάνονται από το επίπεδο των '6orLess' στα '11' η μέση τιμή των σπιτιών αυξάνεται κατά 196.838).

Τζάκια - fireplaces

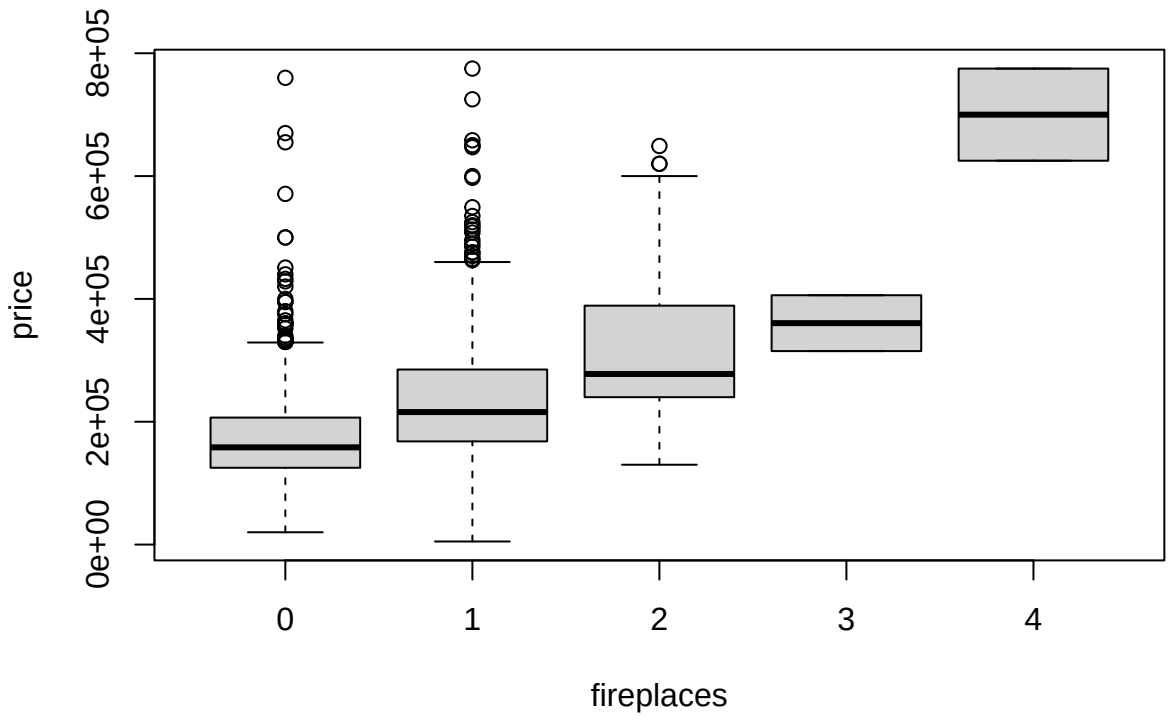
```
fireplaces=factor(fireplaces)
model35=aov(price~fireplaces)
summary(model35)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## fireplaces     4 2.537e+12 6.343e+11   76.97 <2e-16 ***
## Residuals   1723 1.420e+13 8.241e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
PostHocTest(model35)
```

```
##
##   Posthoc multiple comparisons of means : Tukey HSD
##     95% family-wise confidence level
##
## $fireplaces
##          diff     lwr.ci    upr.ci    pval
## 1-0  60509.59   48333.12  72686.07 3.3e-12 ***
## 2-0 144168.01  104848.07 183487.94 3.4e-12 ***
## 3-0 185846.65   10328.90 361364.40  0.0317 *
## 4-0 525346.65  349828.90 700864.40 3.4e-12 ***
## 2-1  83658.42   44565.56 122751.27 6.1e-08 ***
## 3-1 125337.06  -50129.96 300804.07  0.2911
## 4-1 464837.06  289370.04 640304.07 1.0e-11 ***
## 3-2  41678.64 -137727.22 221084.51  0.9695
## 4-2 381178.64  201772.78 560584.51 7.8e-08 ***
## 4-3 339500.00   91615.18 587384.82  0.0018 **
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(price~fireplaces)
```



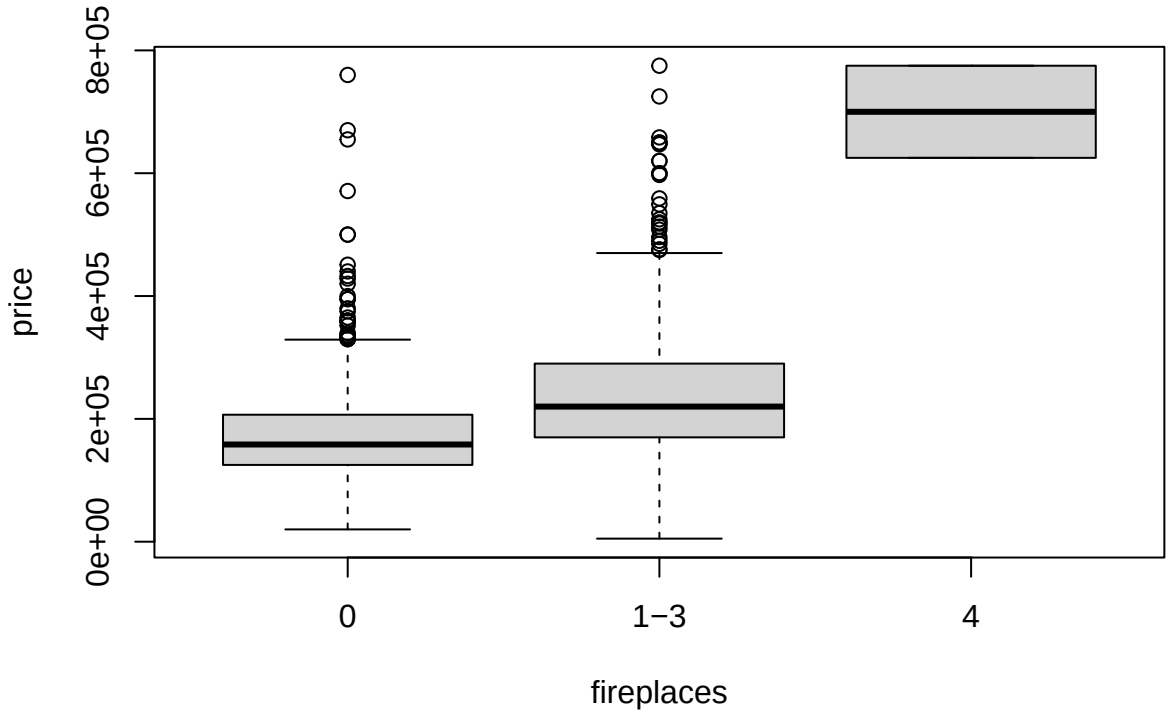
Η ανάλυση διακύμανσης δείχνει στατιστικά σημαντική επίδραση του παράγοντα fireplaces στην τιμή (p<0,001). Ο Post hoc έλεγχος Tukey HSD δείχνει ότι δεν υπάρχει διαφορά ως προς την τιμή ανάμεσα

στον αριθμό τζακιών από 1 έως 3. Επομένως μπούμε να χωρίσουμε τον παράγοντα στις κατηγορίες: 0, 1-3, 4. Το παραπάνω συμπέρασμα γίνεται και οπτικά κατανοητό από το σχετικό boxplot.

```
levels(fireplaces)[c(2:4)]= "1-3"
model35=aov(price~fireplaces)
summary.aov(model35)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## fireplaces     2 2.226e+12 1.113e+12   132.3 <2e-16 ***
## Residuals   1725 1.451e+13 8.411e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
boxplot(price~fireplaces)
```




```
PostHocTest(model35)
```

```
##
##   Posthoc multiple comparisons of means : Tukey HSD
##     95% family-wise confidence level
##
## $fireplaces
##            diff    lwr.ci   upr.ci    pval
## 1-3-0  64327.37  53863.84  74790.9 5.2e-12 ***
```

```
## 4-0   525346.65 373017.45 677675.8 5.3e-12 ***
## 4-1-3 461019.28 308741.31 613297.3 1.1e-11 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary.lm(model35)
```

```
##
## Call:
## aov(formula = price ~ fireplaces)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -233981  -58981  -18653   42329  585347
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     174653       3372   51.80  < 2e-16 ***
## fireplaces1-3    64327       4461   14.42  < 2e-16 ***
## fireplaces4     525347      64939    8.09 1.11e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91710 on 1725 degrees of freedom
## Multiple R-squared:  0.133,  Adjusted R-squared:  0.132
## F-statistic: 132.3 on 2 and 1725 DF,  p-value: < 2.2e-16
```

Η ανάλυση διακύμανσης του απλοποιημένου και βελτιωμένου μοντέλου εξακολουθεί δείχνει στατιστικά σημαντική επίδραση του παράγοντα fireplaces στην τιμή (p<0,001). Από το PostHocTest βλέπουμε ότι όλα τα επίπεδα έχουν στατιστικά σημαντική διαφορά μεταξύ τους καθώς και την διαφορά που παρατηρείται στη μέση τιμή του σπιτιού όταν τα τζάκια αυξάνονται από το κάθε επίπεδο στο άλλο (ενδεικτικά όταν τα τζάκια αυξάνονται από το επίπεδο των '0' στα '4' η μέση τιμή των σπιτιών αυξάνεται κατά 525.347).

Τύπος θέρμανσης - heating

Παρατηρούμε ότι η μεταβλητή heating έχει στατιστικά σημαντική επίδραση στην τιμή (p<0,001) όμως χαμηλή ερμηνευτική ισχύ της μεταβλητότητας της τιμής(R-squared= 5,97%). Από το PostHocTest βλέπουμε ότι όλα τα επίπεδα της μεταβλητής έχουν στατιστικά σημαντική διαφορά μεταξύ τους ως προς την τιμή. Από την ανάλυση του μοντέλου βλέπουμε την διαφορά που παρατηρείται στη μέση τιμή του σπιτιού όταν αλλάζει ο τύπος της θέρμανσης (ενδεικτικά όταν ο τύπος της θέρμανσης γίνεται από 'electric' 'hot air' η μέση τιμή των σπιτιών αυξάνεται κατά 64.467). Τα παραπάνω φαίνονται καθαρά και στο boxplot (όπου φαίνεται και η ύπαρξη πολλών outliers).

Νεόδμητο - newConstruction

```
model41=aov(price~newConstruction)
summary.lm(model41)
```

```
##
## Call:
## aov(formula = price ~ newConstruction)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -203507  -64632  -21007   41493  566493
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          208507       2396  87.034  < 2e-16 ***
## newConstructionYes    73800      11065   6.669 3.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97230 on 1726 degrees of freedom
## Multiple R-squared:  0.02512,    Adjusted R-squared:  0.02456
## F-statistic: 44.48 on 1 and 1726 DF,  p-value: 3.446e-11
```

```
# plot(newConstruction, price)
```

Παρατηρούμε ότι η μεταβλητή newConstruction έχει στατιστικά σημαντική επίδραση στην τιμή (p<0,001) όμως πολύ χαμηλή ερμηνευτική ισχύ της μεταβλητότητας της τιμής(R-squared= 2,5%). Δηλαδή η μέση τιμή των νέων σπιτιών είναι στατιστικά σημαντικά υψηλότερη αλλά η διακυμάνσεις στην τιμή πολύ λίγο οφείλονται στο αν είναι νέο το σπίτι. Συγκεκριμένα βλέπουμε ότι η διαφορά που παρατηρείται στη μέση τιμή του σπιτιού όταν αυτό είναι νεόδμητο είναι 73.800. Τα παραπάνω φαίνονται καθαρά και στο boxplot (όπου φαίνεται και η ύπαρξη πολλών outliers στα παλαιά σπίτια και καθόλου στα νέα).

Ωστόσο όπως δείχθηκε παραπάνω πολλές από τις μεταβλητές αυτές έχουν συσχέτιση μεταξύ τους (collinearity) και επομένως δε θα χρειαστούν όλες στο τελικό μοντέλο.

#3ο Μοντέλα με εξαρτημένη μεταβλητή την price (numerical & categorical) Για το επόμενο μοντέλο θα λάβουμε υπόψη τη συμμεταβλητότητα των παραγόντων και θα επιλέξουμε να συμπεριλάβουμε αυτές με χαμηλό collinearity, επιλέγοντας μία συνεχή numerical (την landValue) μια διακριτή numerical που μπορεί να μετατραπεί σε Ordinal (την fireplaces) μια categorical (την heating) και μία binary (την centralAir)
Θα ακολουθήσουμε διαδικασία ANCOVA σχηματίζοντας το μοντέλο με κύριες επιδράσεις και αλληλεπιδράσεις 2ης τάξης.

```
model16= aov(price~ landValue +fireplaces +heating +centralAir +landValue:fireplaces+heating:centralAir
summary.lm(model16)
```

```
##
## Call:
## aov(formula = price ~ landValue + fireplaces + heating + centralAir +
##     landValue:fireplaces + heating:centralAir + landValue:heating +
##     centralAir:fireplaces + landValue:centralAir + fireplaces:heating)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -273177  -47378   -9078   34409  428709
##
## Coefficients: (3 not defined because of singularities)
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           1.082e+05  7.711e+03  14.028  < 2e-16 ***
```

```
## landValue                              2.320e+00  3.310e-01   7.010 3.43e-12 ***
## fireplaces1-3                          2.173e+04  9.627e+03   2.257 0.024130 *
## fireplaces4                            8.492e+05  2.197e+05   3.865 0.000115 ***
## heatinghot air                         2.463e+04  8.679e+03   2.838 0.004600 **
## heatinghot water/steam                 2.928e+04  1.093e+04   2.679 0.007454 **
## centralAirYes                         -9.141e+03  1.310e+04  -0.698 0.485367
## landValue:fireplaces1-3               -1.381e-01  1.217e-01  -1.134 0.256749
## landValue:fireplaces4                 -4.773e+00  2.224e+00  -2.146 0.031974 *
## heatinghot air:centralAirYes           3.434e+04  1.276e+04   2.692 0.007171 **
## heatinghot water/steam:centralAirYes   3.550e+04  1.922e+04   1.848 0.064838 .
## landValue:heatinghot air              -9.331e-01  3.220e-01  -2.898 0.003805 **
## landValue:heatinghot water/steam      -1.126e+00  3.400e-01  -3.313 0.000942 ***
## fireplaces1-3:centralAirYes           -4.934e+02  8.859e+03  -0.056 0.955592
## fireplaces4:centralAirYes                    NA         NA      NA       NA
## landValue:centralAirYes                1.883e-01  1.254e-01   1.502 0.133344
## fireplaces1-3:heatinghot air           2.248e+04  1.074e+04   2.093 0.036533 *
## fireplaces4:heatinghot air                   NA         NA      NA       NA
## fireplaces1-3:heatinghot water/steam   2.647e+04  1.296e+04   2.043 0.041178 *
## fireplaces4:heatinghot water/steam           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73640 on 1711 degrees of freedom
## Multiple R-squared:  0.4456, Adjusted R-squared:  0.4404
## F-statistic: 85.96 on 16 and 1711 DF,  p-value: < 2.2e-16
```

```
summary(model16)
```

```
##                          Df    Sum Sq   Mean Sq  F value   Pr(>F)
## landValue                 1 5.655e+12 5.655e+12 1042.772  < 2e-16 ***
## fireplaces                2 1.181e+12 5.906e+11  108.914  < 2e-16 ***
## heating                   2 1.923e+11 9.617e+10   17.735 2.38e-08 ***
## centralAir                1 2.564e+11 2.564e+11   47.288 8.56e-12 ***
## landValue:fireplaces      2 2.329e+10 1.164e+10    2.147  0.11711
## heating:centralAir        2 4.501e+10 2.251e+10    4.150  0.01592 *
## landValue:heating         2 6.577e+10 3.288e+10    6.064  0.00237 **
## fireplaces:centralAir     1 6.843e+07 6.843e+07    0.013  0.91057
## landValue:centralAir      1 1.053e+10 1.053e+10    1.942  0.16367
## fireplaces:heating        2 2.862e+10 1.431e+10    2.639  0.07173 .
## Residuals              1711 9.278e+12 5.423e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

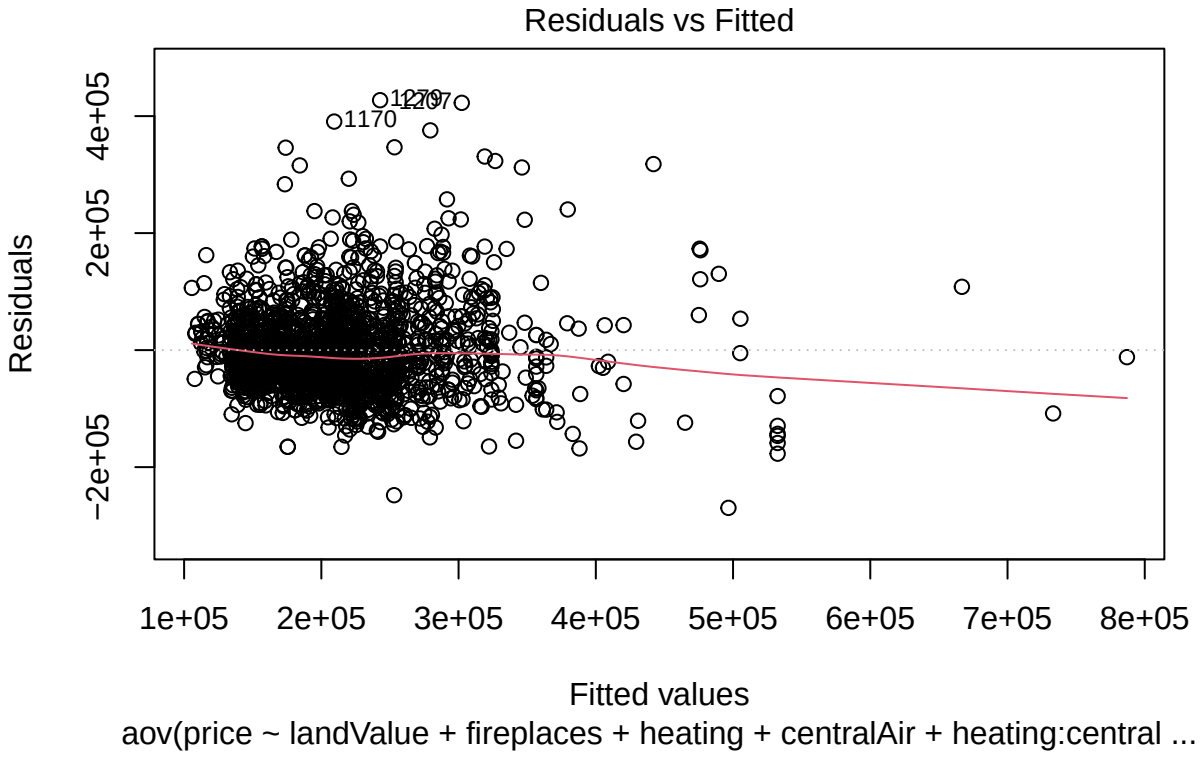
Παρατηρούμε ότι p<0.001<0.05 και επομένως το μοντέλο συνεισφέρει σημαντικά στην ερμηνεία της εξαρτημένης μεταβλητής. Επίσης εξηγεί το (R-squared) 45% της μεταβλητότητας της εξαρτημένης μεταβλητής(price). Οι συντελεστές landValue, fireP, heating, centralAir,όπως και η αλληλεπίδραση των landValue με την heating είναι στατιστικά σημαντικοί αφού έχουν p<0.001<0.05 και επομένως συνεισφέρουν στην ερμηνεία της εξαρτημένης μεταβλητής. Επίσης σημαντική είναι η αλληλεπίδραση των heating και centralAir (p=0,015<0.05). Αντίθετα οι υπόλοιπες αλληλεπιδράσεις δεν είναι στατιστικά σημαντικές (έχουν p>0.05) και επομένως μπορούν να αφαιρεθούν από το μοντέλο.

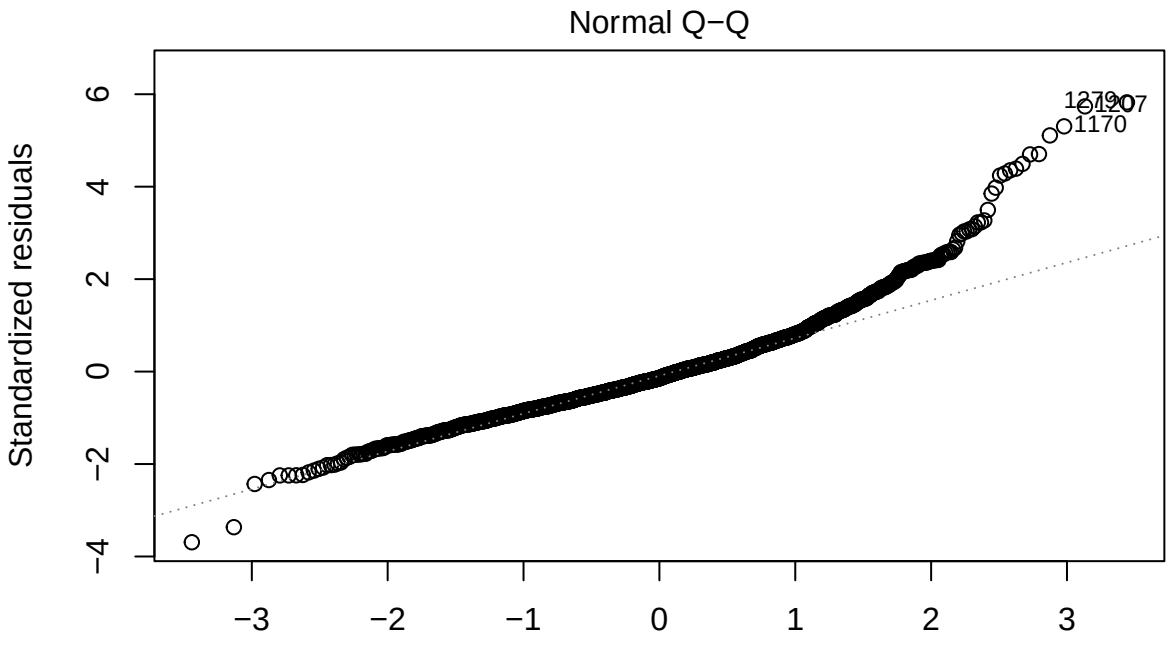
```
model17=update(model16, ~. -landValue:fireplaces -fireplaces:centralAir -landValue:centralAir -fireplac
anova(model16,model17)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ landValue + fireplaces + heating + centralAir + landValue:fireplaces +
##     heating:centralAir + landValue:heating + centralAir:fireplaces +
##     landValue:centralAir + fireplaces:heating
## Model 2: price ~ landValue + fireplaces + heating + centralAir + heating:centralAir +
##     landValue:heating
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1   1711 9.2781e+12
## 2   1717 9.3418e+12 -6 -6.371e+10 1.9582 0.06845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

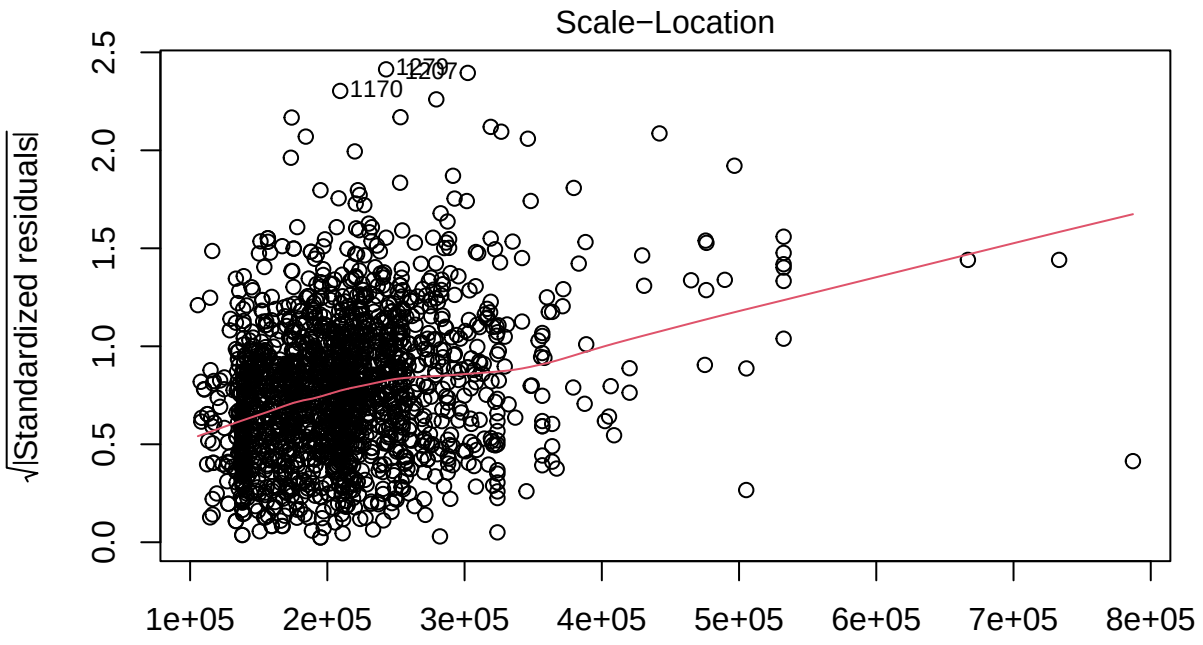
Παρατηρούμε ότι το μοντέλο από το οποίο αφαιρέθηκαν οι μη στατιστικά σημαντικοί όροι δεν έχει στατιστικά σημαντική διαφορά από το αρχικό αφού p=0.068>0.05.

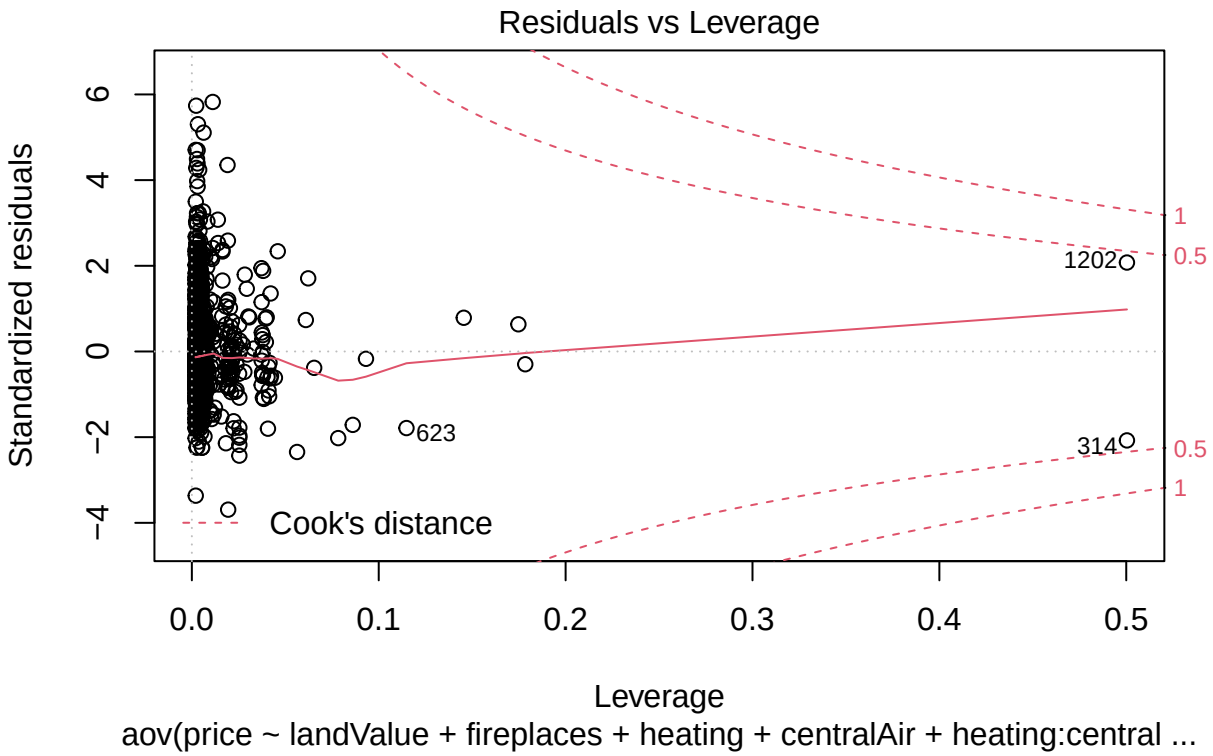
```
plot(model17)
```





aov(price ~ landValue + fireplaces + heating + centralAir + heating:central ...

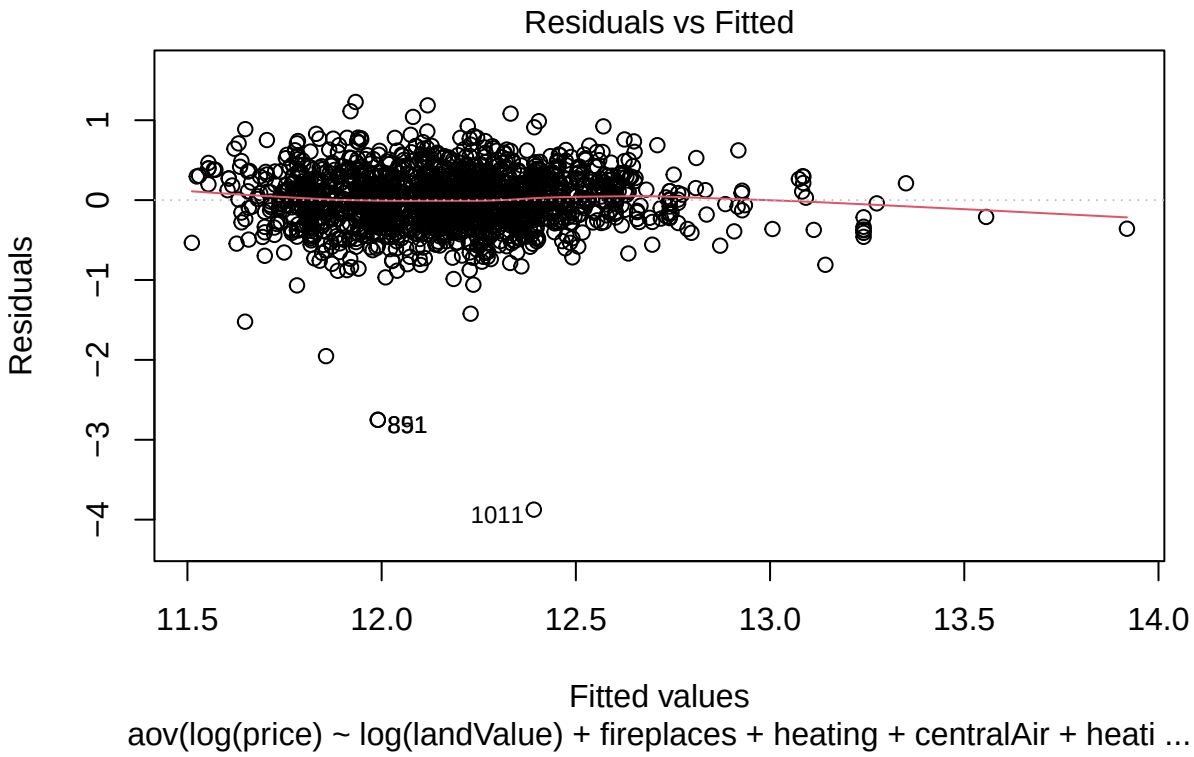

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
1170
1207
1279
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
8e+05
Fitted values
aov(price ~ landValue + fireplaces + heating + centralAir + heating:central ...

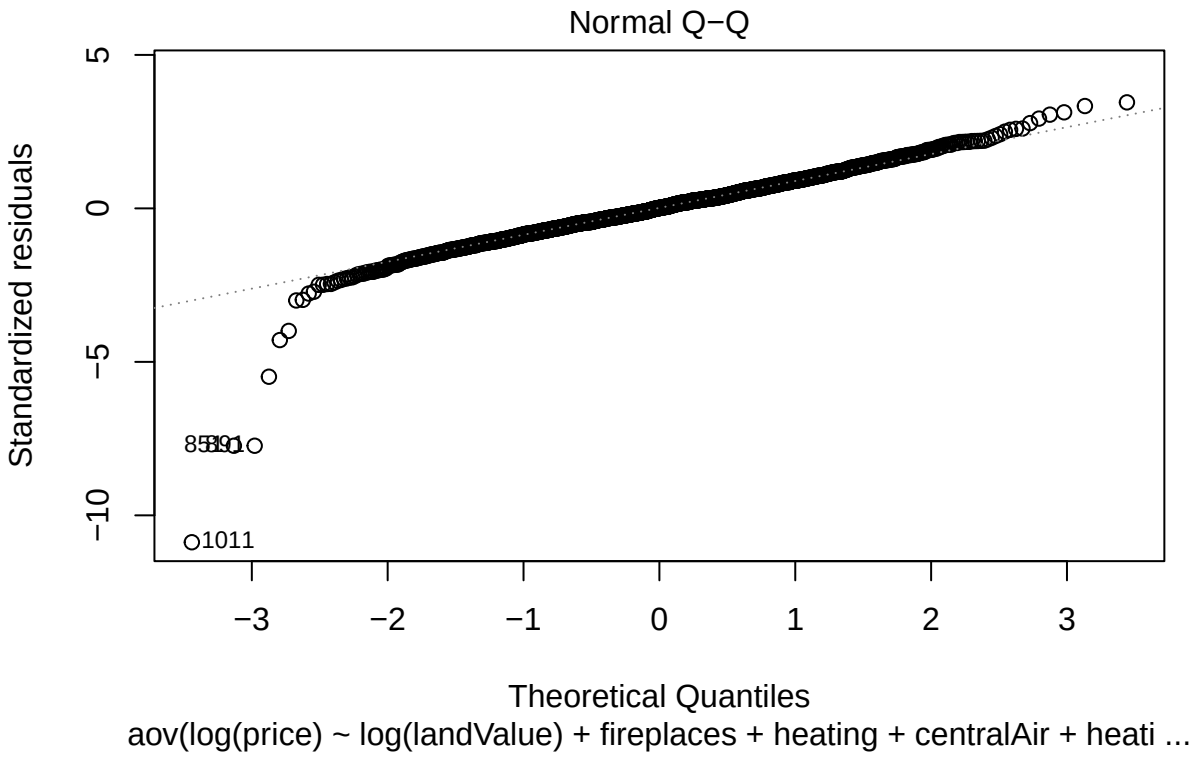


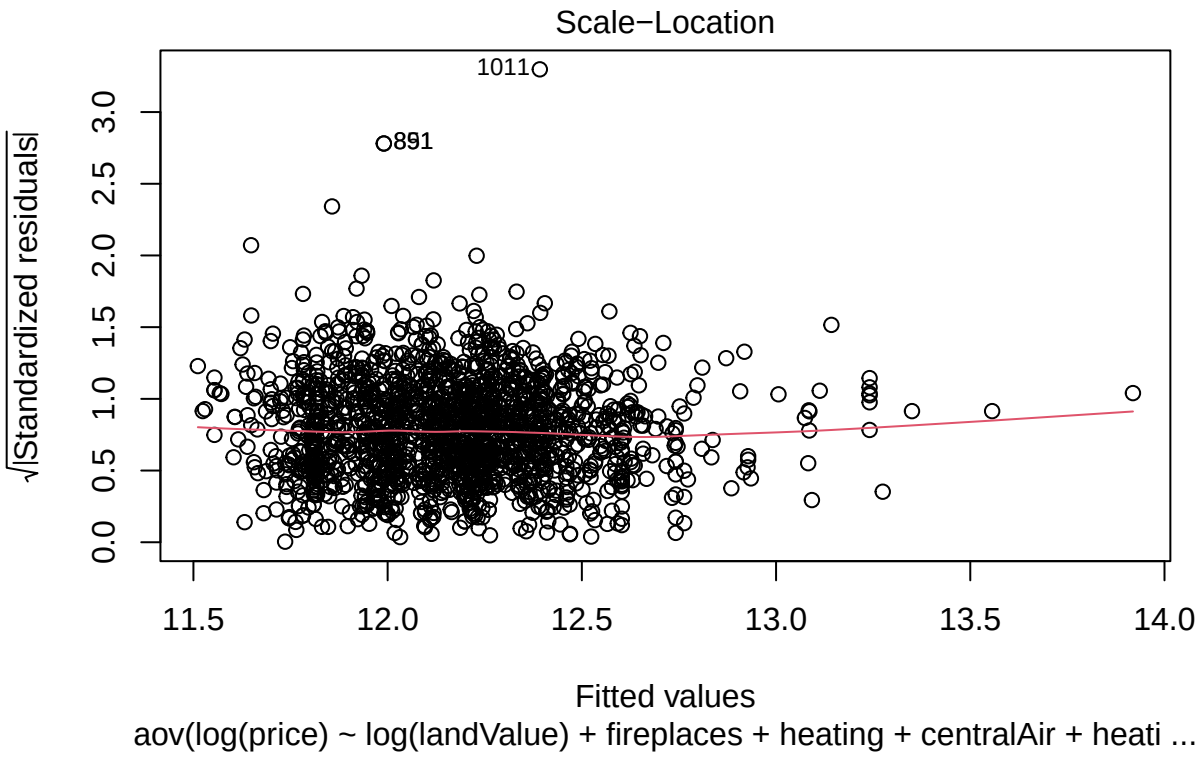
aov(price ~ landValue + fireplaces + heating + centralAir + heating:central ...

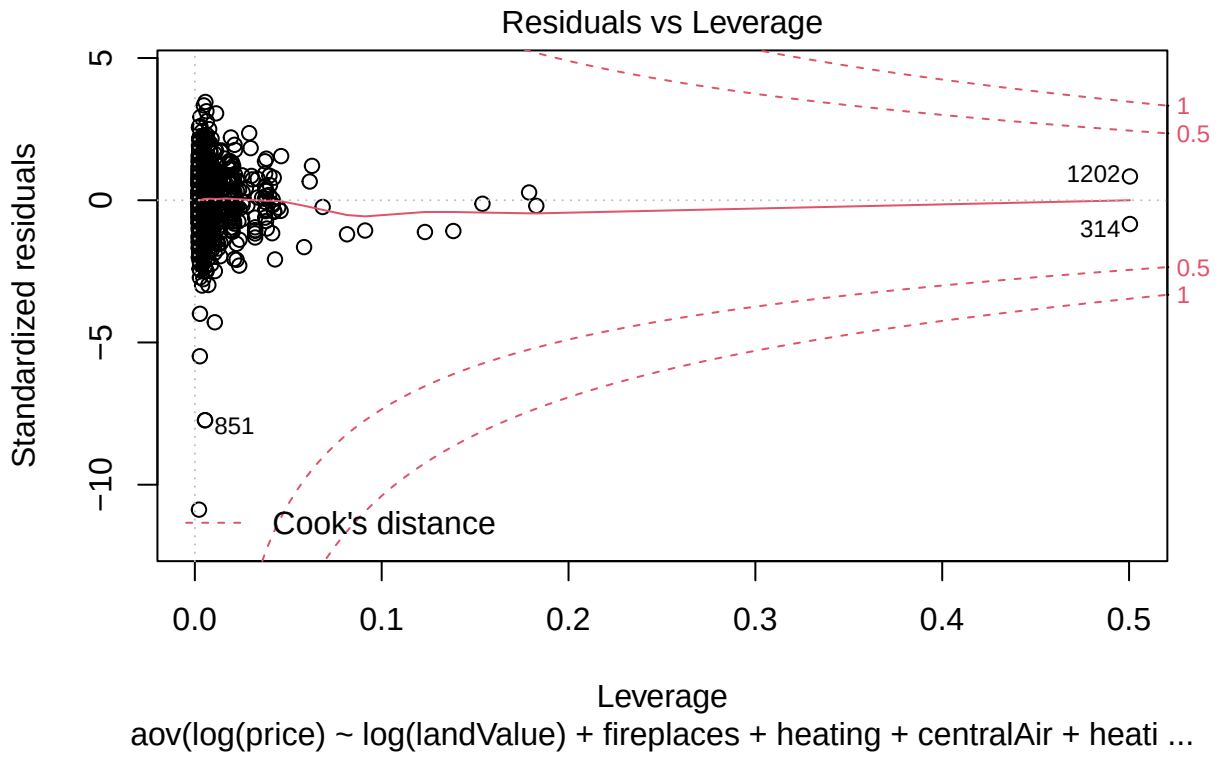
Παρατηρούμε η κατανομή των υπολοίπων ξεφεύγει από την κανονική (QQplot). Θα πειραματιστούμε λοιπόν με το λογαριθμικό μετασχηματισμό της εξαρτημένης μεταβλητής και της landValue.

```
model18= aov(log(price)~ log(landValue) +fireplaces +heating +centralAir +heating:centralAir +landValue
plot(model18)
```


Residuals vs Fitted
Residuals
1
0
−1
−2
−3
−4
891
1011
11.5
12.0
12.5
13.0
13.5
14.0
Fitted values
aov(log(price) ~ log(landValue) + fireplaces + heating + centralAir + heati ...


Normal Q−Q
Standardized residuals
5
0
−5
−10
1011
−3
−2
−1
0
1
2
3
Theoretical Quantiles
aov(log(price) ~ log(landValue) + fireplaces + heating + centralAir + heati ...


Scale−Location
1011
891
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
11.5
12.0
12.5
13.0
13.5
14.0
Fitted values
aov(log(price) ~ log(landValue) + fireplaces + heating + centralAir + heati ...

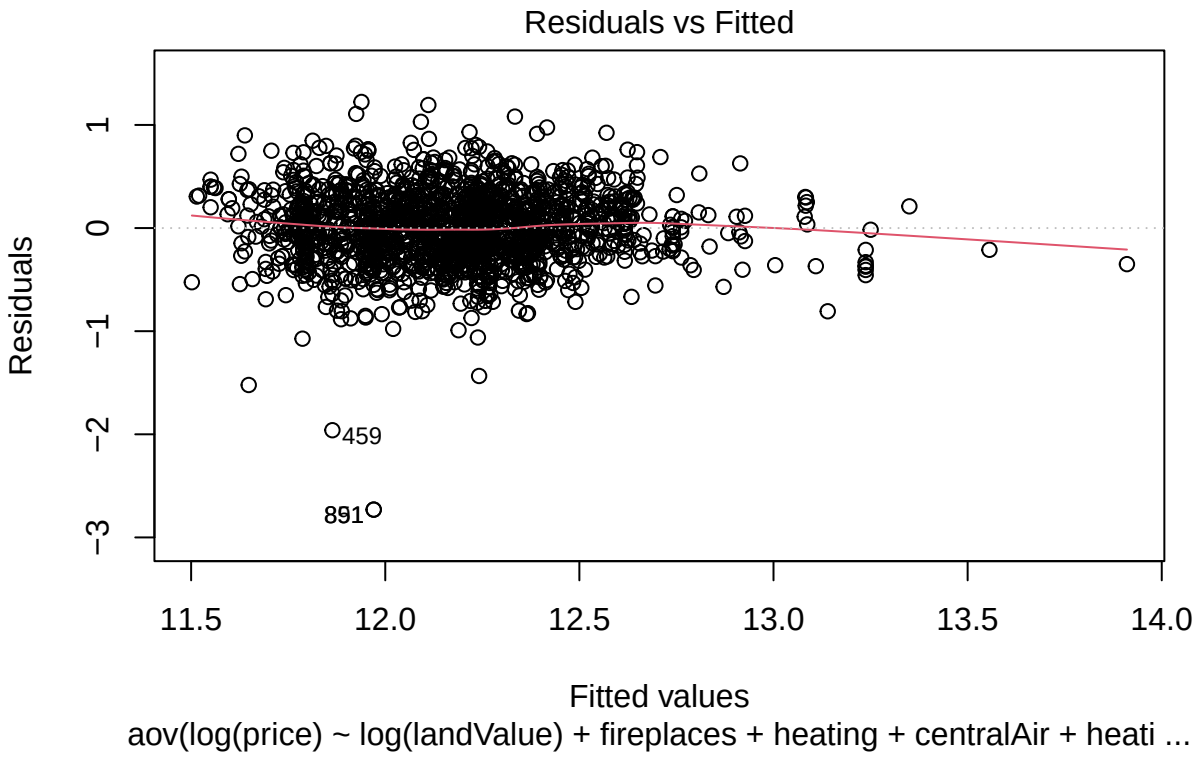


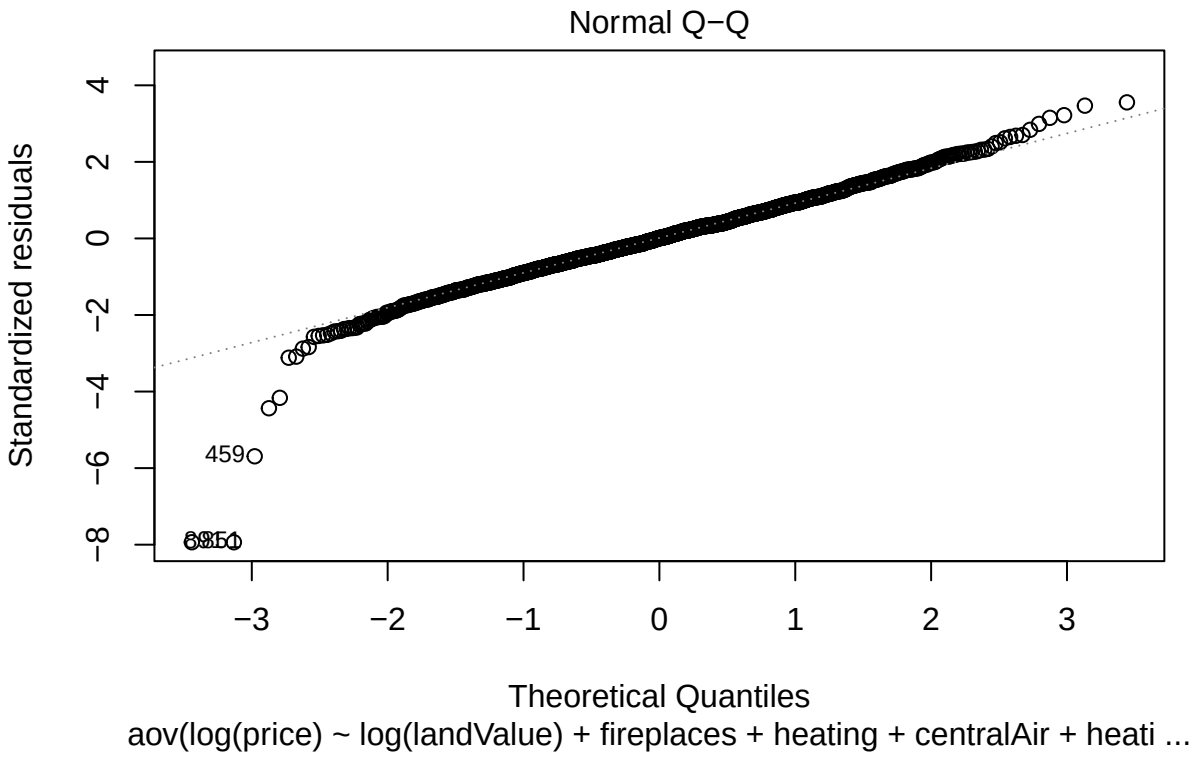

```
summary(model18)
```

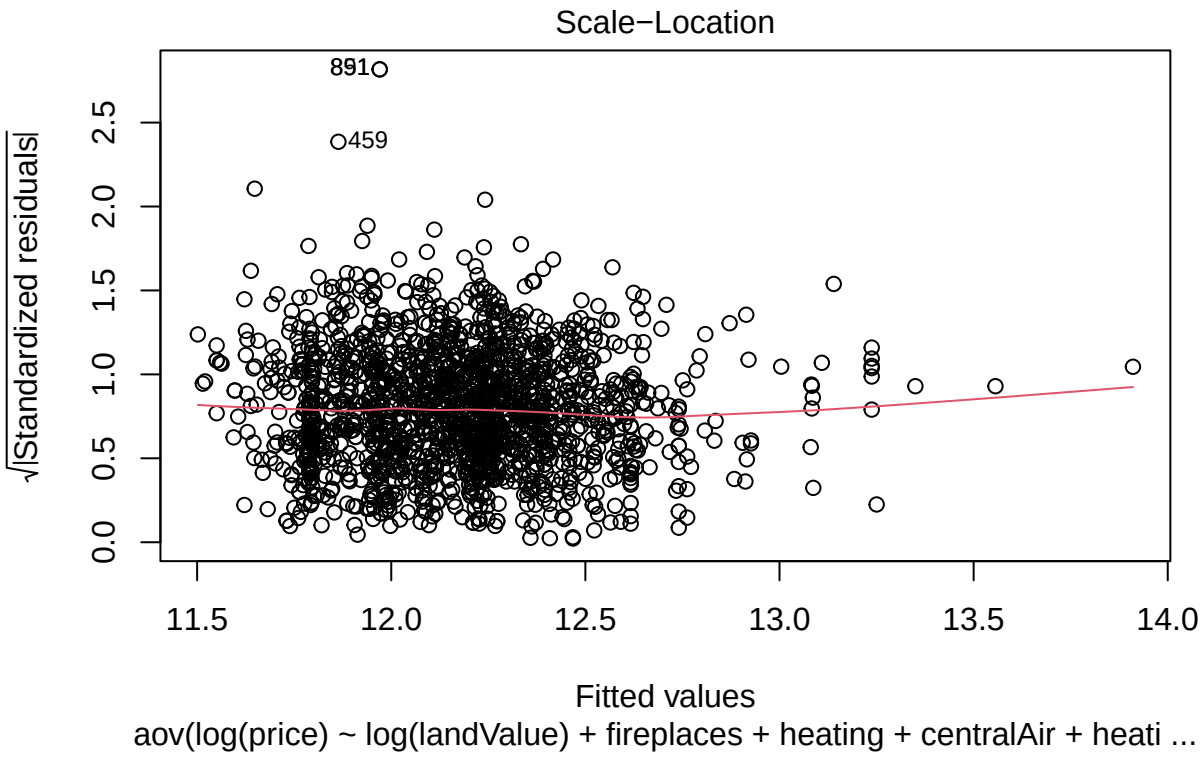
```
##                       Df Sum Sq Mean Sq F value   Pr(>F)
## log(landValue)         1  90.01   90.01  707.70  < 2e-16 ***
## fireplaces             2  20.19   10.10   79.38  < 2e-16 ***
## heating                2   9.09    4.55   35.75 6.14e-16 ***
## centralAir             1   5.53    5.53   43.50 5.62e-11 ***
## heating:centralAir     2   0.61    0.30    2.38   0.0929 .
## heating:landValue      3  10.36    3.45   27.15  < 2e-16 ***
## Residuals           1716 218.26    0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

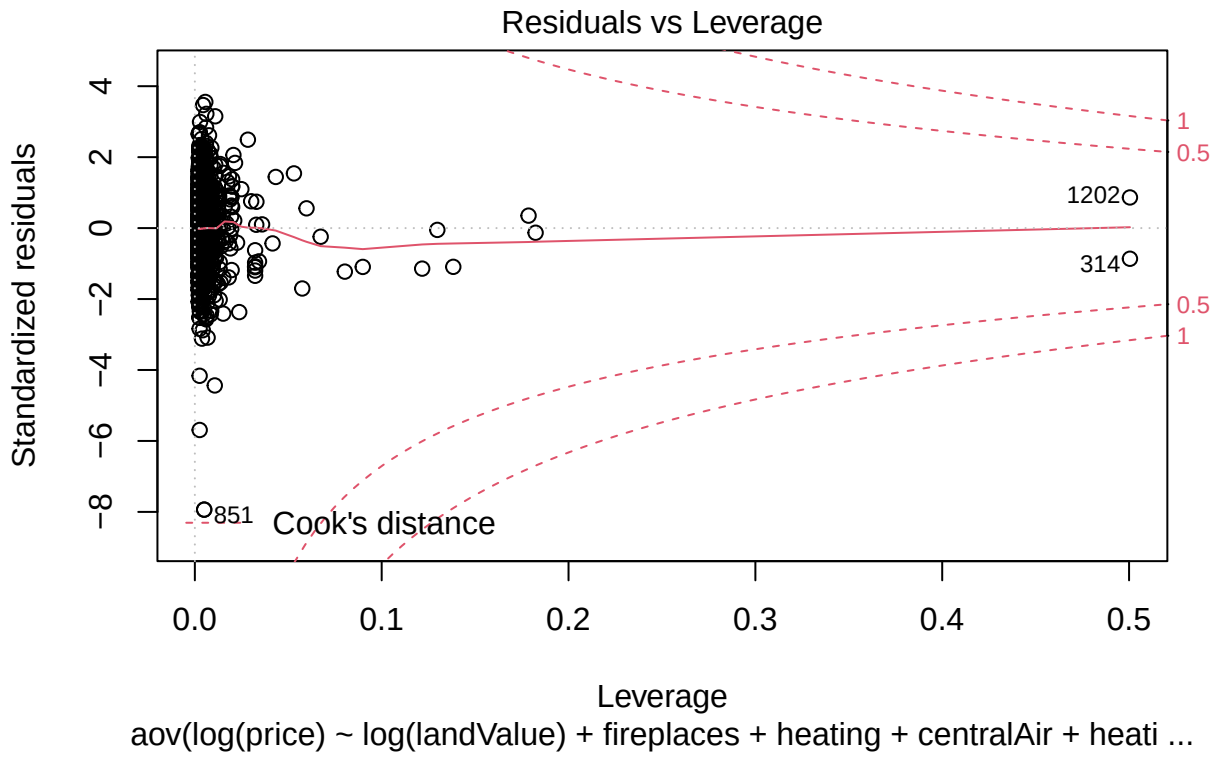
Η κατανομή των καταλοίπων είναι σημαντικά βελτιωμένη, προσεγγίζοντας την κανονική. Στο νέο μοντέλο η αλληλεπίδραση των heating και centralAir δεν είναι στατιστικά σημαντική (p=0,09>0.05) και επομένως μπορεί να αφαιρεθούν από το μοντέλο. Ακόμα φαίνεται πως η παρατήρηση #1011 έχει άσχημη επίδραση στην παλινδρόμηση και επομένως στο τελικό μοντέλο θα αφαιρεθεί.

```
model19=update(model18, ~. -heating:centralAir, subset=(1:length(price)!=1011))
plot(model19)
```


Residuals vs Fitted
Residuals
1
0
−1
−2
−3
459
891
11.5
12.0
12.5
13.0
13.5
14.0
Fitted values
aov(log(price) ~ log(landValue) + fireplaces + heating + centralAir + heati ...


Normal Q−Q
Standardized residuals
4
2
0
−2
−4
−6
−8
459
−3
−2
−1
0
1
2
3
Theoretical Quantiles
aov(log(price) ~ log(landValue) + fireplaces + heating + centralAir + heati ...


Scale−Location
891
459
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
11.5
12.0
12.5
13.0
13.5
14.0
Fitted values
aov(log(price) ~ log(landValue) + fireplaces + heating + centralAir + heati ...




```
summary.lm(model19)
```

```
##
## Call:
## aov(formula = log(price) ~ log(landValue) + fireplaces + heating +
##     centralAir + heating:landValue, subset = (1:length(price) !=
##     1011))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.73037 -0.20714  0.00328  0.21584  1.22286
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.086e+01  1.247e-01  87.071  < 2e-16 ***
## log(landValue)                    8.739e-02  1.336e-02   6.538 8.17e-11 ***
## fireplaces1-3                     1.747e-01  1.806e-02   9.669  < 2e-16 ***
## fireplaces4                       9.454e-01  2.452e-01   3.856 0.000119 ***
## heatinghot air                    1.901e-01  3.813e-02   4.985 6.81e-07 ***
## heatinghot water/steam            2.082e-01  4.439e-02   4.691 2.93e-06 ***
## centralAirYes                     1.239e-01  1.963e-02   6.311 3.51e-10 ***
## heatingelectric:landValue         5.772e-06  1.451e-06   3.979 7.21e-05 ***
## heatinghot air:landValue          3.468e-06  4.084e-07   8.493  < 2e-16 ***
## heatinghot water/steam:landValue 3.117e-06  6.067e-07   5.138 3.10e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3449 on 1717 degrees of freedom
## Multiple R-squared:  0.4006, Adjusted R-squared:  0.3974
## F-statistic: 127.5 on 9 and 1717 DF,  p-value: < 2.2e-16
```

Παρατηρούμε ότι το τελικό μοντέλο έχει στατιστικά σημαντική ερμηνευτική ισχύ της τιμής των σπιτιών (p-value: < 2.2e-16), και σχετικά καλή εξήγηση της μεταβλητότητας της (~40%). Όλοι οι όροι είναι στατιστικά σημαντικοί.

Μοντέλα με εξαρτημένη μεταβλητή τη newConstruction

Θα προσπαθήσουμε να κατασκευάσουμε κατάλληλο μοντέλο πρόβλεψης για το αν είναι νέα κατασκευή και το βαθμό της συσχέτισης με αυτό των μεταβλητών: ποσοστό αποφοίτων κολεγίου της γειτονιάς, αριθμό δωματίων, τιμής και (εξ ορισμού) ηλικίας της κατασκευής.

```
table(newConstruction)
```

```
## newConstruction
##   No  Yes
## 1647   81
```

```
tapply(price,newConstruction,mean)
```

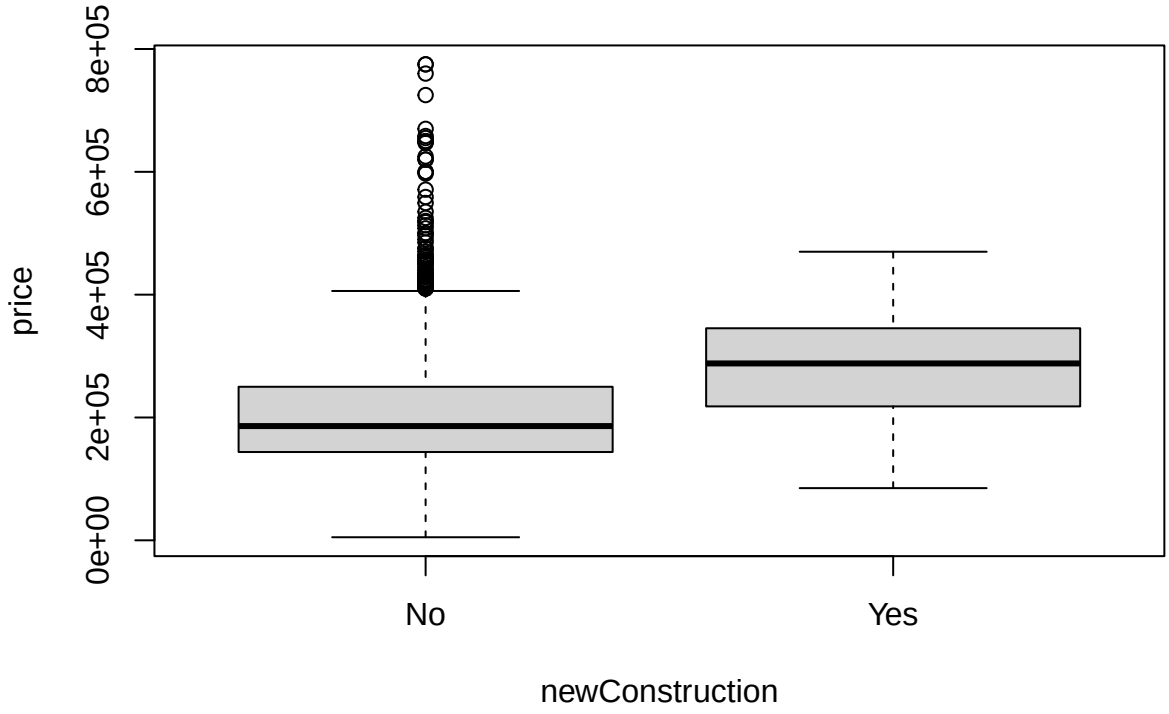
```
##       No      Yes
## 208507.4 282306.8
```

Παρατηρούμε ότι οι παρατηρήσεις (σπίτια) δεν είναι ισομερώς κατανεμημένες στις δύο κατηγορίες.

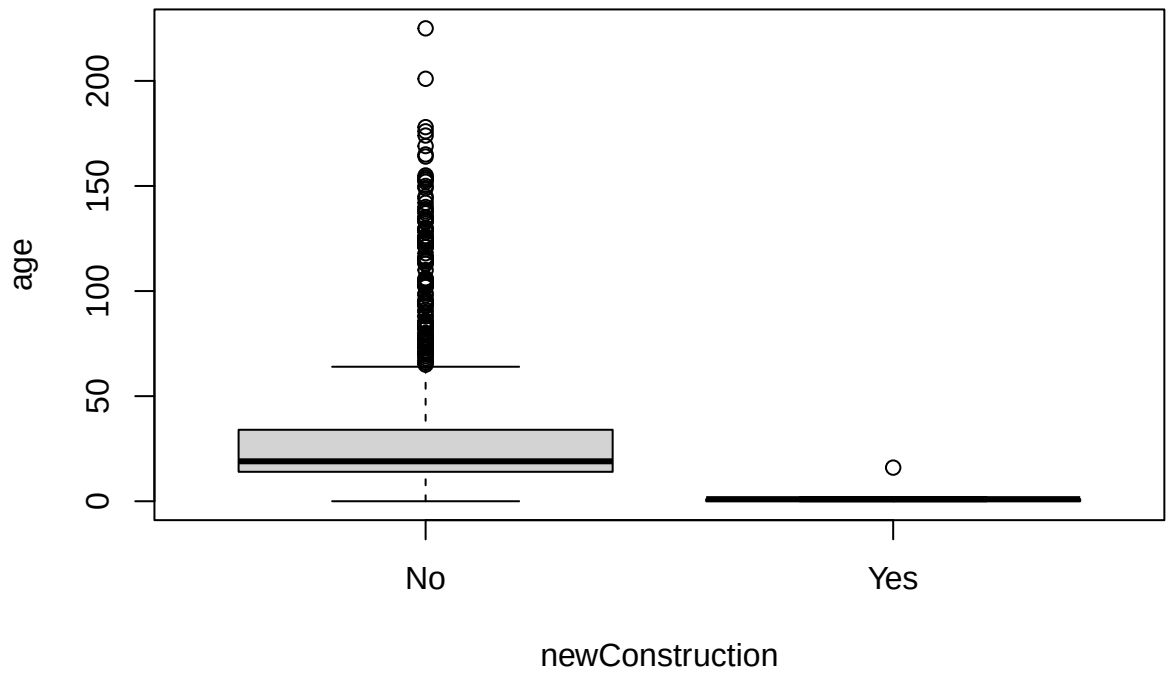
```
newCon = as.factor(newConstruction)
levels(newCon)[1]= 0
levels(newCon)[2]= 1
newCon=ifelse(newCon==1,0,1)
```

```
boxplot(pctCollege~newConstruction)
```

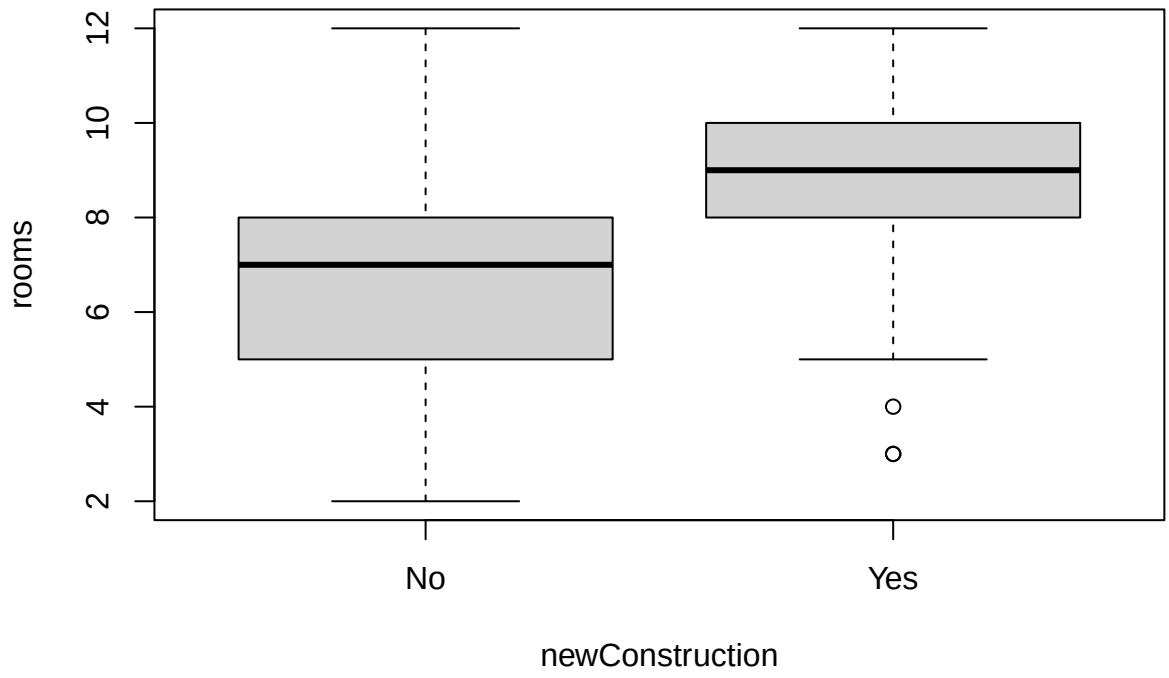

```
boxplot(price~newConstruction)
```




```
boxplot(age~newConstruction)
```




```
boxplot(rooms~newConstruction)
```



Παρατηρούμε σε όλα τα boxplots πως τα επίπεδα της εξαρτημένης μεταβλητής που εξετάζουμε (newConstruction) φαίνεται να διαφέρουν για όλες τις εξαρτημένες που επιλέξαμε. Ακολούθως θα το διερευνήσουμε και με τον συντελεστή συσχέτισης του Kendall.

```
cor.test(pctCollege,newCon,method="kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  pctCollege and newCon
## z = 3.1814, p-value = 0.001466
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.06582315
```

```
cor.test(price,newCon,method="kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  price and newCon
## z = -7.6348, p-value = 2.262e-14
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
```

```
##        tau
## -0.1502536
```

```
cor.test(age,newCon,method="kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  age and newCon
## z = 14.248, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2833876
```

```
cor.test(rooms,newCon,method="kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  rooms and newCon
## z = -8.0407, p-value = 8.936e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.1670935
```

Παρατηρούμε ότι υπάρχει στατιστικά σημαντική θετική συσχέτιση του αν είναι νεόδμητο με την τιμή και με τον αριθμό των δωματίων. Επίσης υπάρχει στατιστικά σημαντική αρνητική συσχέτιση του αν είναι νεόδμητο με το ποσοστό αποφοίτων κολεγίου της γειτονιάς και της ηλικίας της κατασκευής(εξ ορισμού).

Στην περίπτωσή μας έχουμε binary μεταβλητή απόκρισης. Ενδείκνυται λοιπόν η χρήση των γενικευμένων γραμμικών μοντέλων. Θα σχηματίσουμε το μοντέλο με τις κύριες επιδράσεις και τις αλληλεπιδράσεις 2ης τάξης.

```
model80 = glm(newCon~pctCollege+rooms+age+price ,binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model80)
```

```
##
## Call:
## glm(formula = newCon ~ pctCollege + rooms + age + price, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0486   0.0000   0.0019   0.0131   2.0288
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.022e+00  9.036e-01  -1.131  0.25788
## pctCollege   4.962e-02  1.512e-02   3.281  0.00103 **
## rooms       -2.673e-01  8.703e-02  -3.072  0.00213 **
## age          6.317e-01  1.044e-01   6.050 1.45e-09 ***
## price        2.245e-06  1.742e-06   1.289  0.19756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 653.91  on 1727  degrees of freedom
## Residual deviance: 260.43  on 1723  degrees of freedom
## AIC: 270.43
##
## Number of Fisher Scoring iterations: 12
```

Όλοι όροι που φαίνονται σημαντικοί στο μοντέλο εκτός της τιμής μπορεί να αφαιρεθεί.

```
model81 = glm(newCon~pctCollege+rooms+age,binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model81)
```

```
##
## Call:
## glm(formula = newCon ~ pctCollege + rooms + age, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1703   0.0000   0.0015   0.0111   1.9831
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.11209    0.90040  -1.235 0.216791
## pctCollege   0.05337    0.01490   3.582 0.000341 ***
## rooms       -0.20861    0.07318  -2.851 0.004365 **
## age          0.65330    0.10777   6.062 1.35e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 653.91  on 1727  degrees of freedom
## Residual deviance: 262.16  on 1724  degrees of freedom
## AIC: 270.16
##
## Number of Fisher Scoring iterations: 12
```

```
anova(model80,model81,test="Chi")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: newCon ~ pctCollege + rooms + age + price
## Model 2: newCon ~ pctCollege + rooms + age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1723     260.43
## 2      1724     262.17 -1  -1.7388   0.1873
```

Τα δύο μοντέλα δε φαίνεται να διαφέρουν σημαντικά. Θα προτιμήσουμε το δεύτερο και απλούστερο μοντέλο που ουσιαστικά δεν χάνει κάτι αφαιρώντας τον όρο τιμή που δεν είναι στατιστικά σημαντικός. Όπως είχαμε διαπιστώσει και παραπάνω ο αριθμός των δωματίων έχει θετική συσχέτιση με το αν είναι νεόδμητο ενώ αρνητική σχέση έχει το ποσοστό αποφοίτων κολεγίου της γειτονιάς και ασφαλώς η ηλικία του.