BUSINESS ANALYTICS
Master of Science

STATISTICS FOR BUSINESS ANALYTICS I
ASSIGNMENT II

ELENI RALLI (f2822312)

21/12/2023

Contents

The data for this assignment are a random sample of 63 cases from the files of a big real estate agency in USA concerning house sales from February 15 to April 30, 1993. The data was collected from many cities (and corresponding local real estate agencies) and is used as a basis for the whole company. The variables in this datasets are:

1. PRICE = Selling prices (in hundreds$)

2. SQFT = Square Feet of living space

3. AGE = Age of home (in years)

4. FEATS = Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access

5. NE = Located in northeast sector of city (1) or not (0)

6.COR = Corner location (1) or not (0).

Taks:

(I) Read the “usdata” dataset and use str() to understand its structure.

The data are in txt form . The command to read them is read.table( ) and

we will see the structure with the str( ) command.

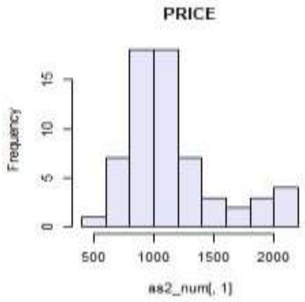
(II) Convert the variables PRICE, SQFT, AGE, FEATS to be numeric variables and NE, COR to be factors.

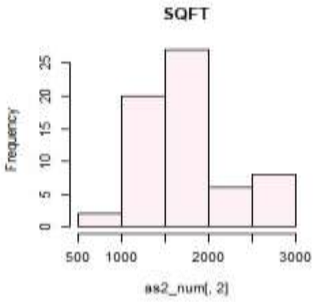
We will convert the variables PRICE, SQFT, AGE, FEATS to be numeric with the command as.numeric( ) .

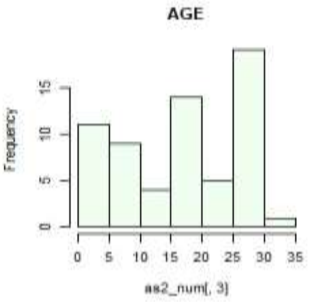
We will convert the variables NE, COR to be factors with the command as.factor( ) .

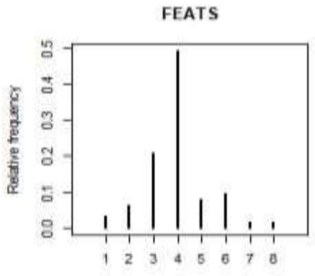
(III) Perform descriptive analysis and visualization for each variable to get an initial insight of what the data looks like. Comment on your findings.

Numeric Variables	PRICE Selling prices (in hundreds$)	SQFT Square Feet of living space	AGE Age of home (in years)	FEATS Number out of 11 features
Mean	1158	1730	17.5	4
Median	1049	1680	20	4
sd	393	507	9	1
Minimum	580	970	2	1
Maximum	2150	2931	31	8
Range	580 : 2150	970 : 2931	2 : 31	1 : 8
Skewness	1.2	0.8	-0.2	0.5
Kurtosis	3.6	2.9	1.6	4.2
1st Quartile	910	1400	7	3
3ed Quartile	1250	1920	27.5	4









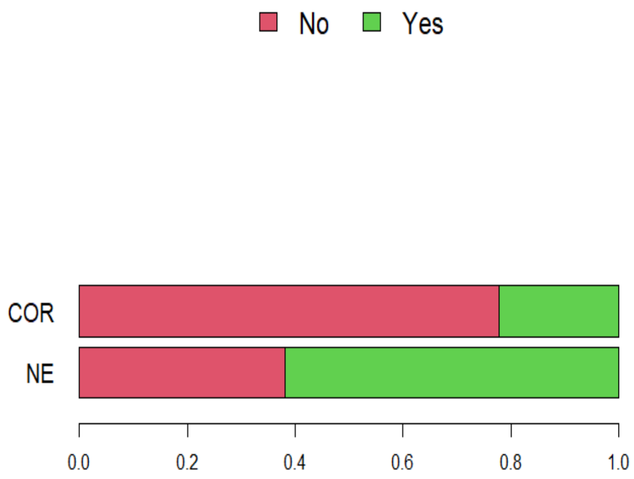
The mean **price** of houses is 1158 hundread dollars .The median of prices is 1049 hundread dollars and is lower than the mean, showing right skewness (1.2). The standard deviation is 393 hundread dollars showing variability in prices. The range of prices is 580 to 2150 hundread dollars which is a wide range.The Kurtosis is 3.6 a little more peaked than a normal distribution.The 50% of prices are between 910 hunderad dollars and 1250 hundread dollars.

The average size of living space in houses is 1730 square feet. The median is at 1680 square feet, showing a a little right skewness (0.8). The standard deviation is 507 square feet, indicating a wide variety in house sizes, ranging from 970 to 2931 square feet. The kurtosis is 2.9 , very close to normal distibution .

Homes are on average 17.5 years old .The median is 20 years. The age range from 2 to 31 years. The slight left skewness(-0.2) shows that we have more new homes (but not a big tend), and the distribution is flatter than normal, showing ages are spread out evenly. Half of the houses are between 7 and 27.5 years old.

The average number of features in a house is 4, with most homes having between 3 and 4 features. This shows a consistent level of features across properties. The range is 1 to 8 features, with a small skewness (0.5)showing that more houses have more feautures . The distribution is a bit more peaked than normal(kurtsosis 4.2) , showing that most houses are close to the average number of features.

Categorical Variables	NE	COR
Level 0 (no)	Doesn't located in northeast sector of city 24 houses Proportion : 0.38	No Corner location 49 houses Proportion : 0.78
Level 1 (yes)	Located in northeast sector of city 39 houses Proportion : 0.62	Corner location 14 houses Proportion : 0.22



In our sample , 39 houses (62%) are located in the northeast sector of the city, while 24 houses (38%) are not in the northeast.

Corner locations, only 14 houses (22%) are situated on a corner, but the majority, 49 houses (78%), are not located on a corner. So most houses are positioned away from corners and are more likely to be in the northeast sector of the city.

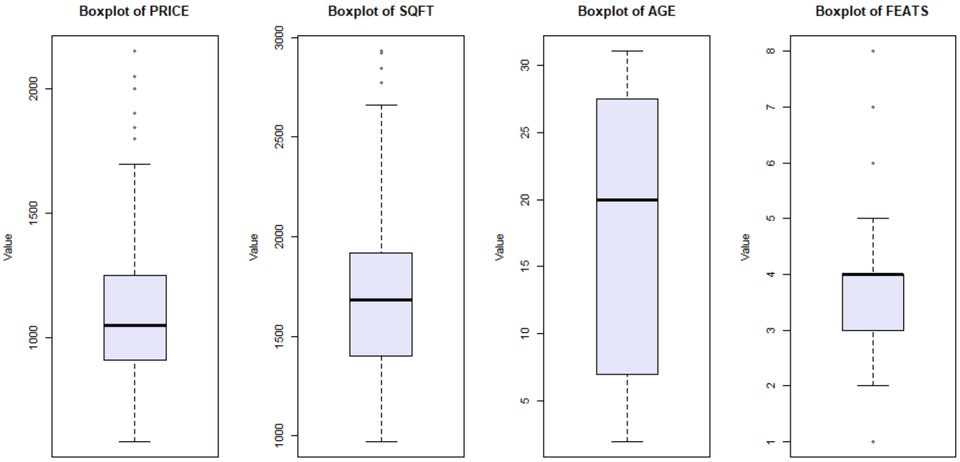
(IV) Conduct pairwise comparisons between the variables in the dataset to investigate if there are any associations implied by the dataset.(Hint: Plot variables against one another and use correlation plots and measures for the numerical variables.). Comment on your findings. Is there a linear relationship between PRICE and any of the variables in the dataset?

For the pairwise comparison we will examine the relationships between numerical and numerical variables, numerical and categorical, categorical and categorical but first we will check each variable separatelly .

We will do shapiro test and lillie test for each numerical variable .

Numeric Variables	PRICE Selling prices (in hundreds$)	SQFT Square Feet of living space	AGE Age of home (in years)	FEATS Number out of 11 features
Shapiro-Wilk test	p-value = 4.944e-06	p-value = 0.001711	p-value = 7.74e-05	p-value = 5.719e-05
Lilliefors test	p-value = 1.006e-05	p-value = 0.002785	p-value = 0.008282	p-value = 2.433e-13

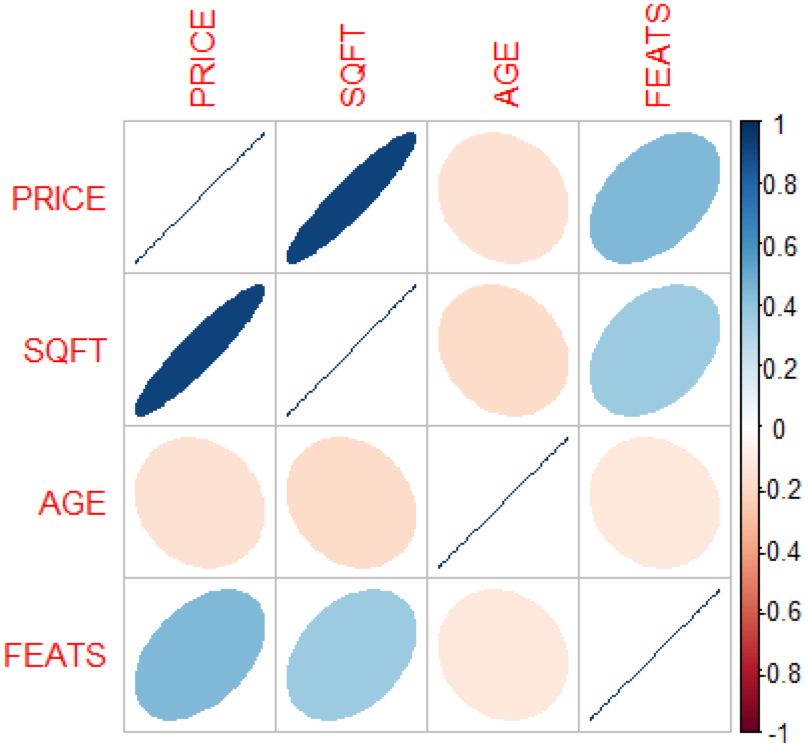
All the p-values<a=0.05 → Ho of normality rejected . The variables can be assume that follow the normal distribution . As you can see and from the boxplots.



Relationship between Price and all numeric variables :

As we can see in the plot appears to be a strong positive(cor=0.95) correlation between PRICE and SQFT, indicating that as the SQFT of a house increases, so does its price. The relationship (cor=-0.15) between PRICE and AGE is not as clear, show that other factors may influence the price with the age. FEATS shows some level of positive correlation with PRICE (cor=0.44), so more features may contribute to higher prices. The plot do not show any obvious patterns between AGE and SQFT or FEATS, so we will not have mutlicolinearity problems probably .

We also did some correlation test however we use the non parametric method spearman because our variables are not follow normal distribution . The results gave as significant relationships between 'PRICE' and the other variables (all p-values of the test <a) , so there is monotonic relationships, (not necessary linearity , rho= 0.93,-0.27,0.40 respectively between price and sqft, age, feats) .



Relationship between Price and all the categorical variables NE and COR :

For houses not in the northeast sector (NE = 0), the p-value was 0.0002725, leading to the rejection of the null hypothesis, showing that PRICE is not normally distributed in this subgroup.For houses in the northeast sector (NE = 1), the p-value was 0.002699, also rejection of the null hypothesis, so a non-normal distribution of 'PRICE' in this subgroup too.

For houses not on a corner (COR = 0), the p-value was 7.805e-05, rejection of the null hypothesis, non-normal distribution of PRICE. For houses on a corner (COR = 1), the p-value was 0.05699 ,the null hypothesis not rejected, PRICE could be normally distributed for this subgroup.

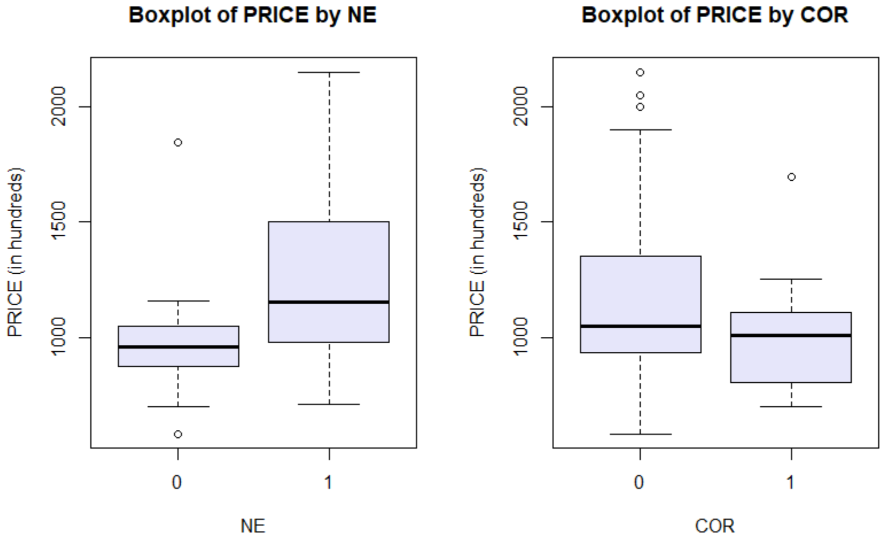
For the NE the p-value was 0.002179, rejection of the null hypothesis. This shows a significant difference in the median prices between houses in and not in the northeast sector.

For the COR the p-value was 0.1389, we don't reject the null hypothesis, no significant difference in the median prices between houses based on corner location.

the PRICE variable within each level of each categorical variable	NE	COR

Shapiro-Wilk test (H0) is the PRICE variable within each level of the categorical variable is normally distributed.	p-value = 0.0002725 (for NE = 0)
	p-value = 0.002699 (for NE = 1)
	the Ho rejected for the 2 groups
		p-value = 7.805e-05 (for COR=0)
		p-value = 0.05699 (for COR=1)
		the Ho rejected only for the COR = 0 group

Wilcoxon test (H0) is that the medians of the two groups are equal	p-value = 0.002179 the Ho rejected	p-value = 0.1389 the Ho doesn't rejected



Relationship between SQFT /AGE /FEATS and the categorical variables :

Between the SQFT and the location in the NE, the Shapiro-Wilk normality doesn't reject the normality (p-value > 0.05),. However the variance test show that variances are not equal across NE categories (p-value < 0.05), reject the null hypothesis of equal variances. Next the t-test shows a significant difference in the mean SQFT between the NE categories (p-value < 0.05), show that the NE has a statistically significant effect on the size of the house. The plotmeans graph show this(in the appendix)   .

Between SQFT and COR, the Shapiro-Wilk test reject the normality (p-value < 0.05), the SQFT is not normally distributed across COR categories. Both sample sizes are less than 50, so we will use the median than the mean. The Wilcoxon test does not show any significant difference in median SQFT between the COR categories (p-value > 0.05), so being on a corner lot does not significantly affect the size of the house . The boxplots show this(in the appendix)   .

Between AGE and NE, the Shapiro-Wilk test reject the normality (p-value < 0.05), the age distribution differs from the normal distribution within NE categories. The sample is small (less than 50 in each group), we will do non-parametric Wilcoxon test, which show significant difference in the median ages of homes between the NE categories (p-value < 0.05). So we can suppose that the location within the northeast sector influences the median age of homes. The boxplots can show this (in the appendix) .

Between AGE and COR, the Shapiro-Wilk test reject the normality for age within COR categories (p-value < 0.05). With small sample sizes, we apply the non-parametric Wilcoxon test, which shows no significant difference in the median ages of homes between the COR categories (p-value > 0.05).So, we have no evidence to suggest that the corner lot status affects the median age of homes. The boxplots can show this (in the appendix) .

Between FEATS and NE, the Shapiro-Wilk reject the normality (p-value < 0.05). The size of the groups are small , we do non-parametric Wilcoxon test and show significant differences in the median number of features between homes in different NE categories (p-value < 0.05). This suggests that homes in different NE locations tend to have different numbers of features. The boxplots can show this (in the appendix) .

Between FEATS and COR, the Shapiro-Wilk reject the normality (p-value < 0.05). The size of the groups are small, less than 50. We do Wilcoxon test to see if there are differences in the median number of features. There is a significant difference in the median number of features between homes with different COR (p-value < 0.05), so we can assume that COR may be associated with the number of features a home. The boxplots can show this (in the appendix) .

Relationship between the categorical variables NE and COR :

Between NE and COR, the Pearsons chi-squared test (p-value = 0.9172>a), we don't reject the Ho, show that probably there is no significant association between the variables . The barplot can show this(in the appendix) .

(V) Construct a model for the expected selling prices (PRICE) according to the remaining features.(hint: Conduct multiple regression having PRICE as a response and all the other variables as predictors). Does this linear model fit well to the data? (Hint: Comment on R^2 adj).

In order to construct a model for the expected selling price(PRICE) according to the remaining features we will first include all independent variables in our dataset (SQFT, AGE , FEATS, NE, COR).

- **Adjusted R^2**: Show the proportion of variance explained by the model, adjusted for the number of variables. A value > 0.7 show good model fit and tell us that the model makes good predictions. In our case , we have Adj- R^2= 0.864 >0.7 , so its is very good.

However from the summary of the model we must examine also the followings:

- **F Statistic and p-value:** The first thing that we are checking is the F-statistic (before R^2 adj), which tells if the model is different from the null model(model with no predictors , only the intercept (the intercept here is the average PRICE in the dataset.)
  With a p-value less than 2.2e-16, we reject the null hypothesis (H0), that means that our model is statistically significant and different from the null model.
- **Coefficient Analysis**: In order to include a independent-explanatory variable in our model we must reject the null hypothesis H0: $\beta i = 0$ (we will use level of significance a= 0.05) ,
  that mean that this i explanatory variable is statistically significant for our dependent variable (PRICE) and contributes meaningfully to the model (so we must included in our model) .
  Also, we must see if whether these variables have a positive or negative effect and the extent of their influence.

	Intercept	SQFT	AGE	FEATS	NE	COR
Estimate	-193.34926	0.67662	2.22907	34.36573	30.00446	-53.07940
Interpretation	-193.34926 is the expected value of PRICE when all predictors are zero and we are not in the northest sector in the city without corner location	for each additional square feet, the PRICE increases by approximately 0.67662 hundread$. if all the other explanatory variables are remain constant aand we are not in the northest sector in the city without corner location	for each additional year in age, the PRICE increases by about 2.22907 hundread$ **(BUT THIS IS NOT RELIABLE due to the lack of statistical significance.)** if all the other explanatory variables are remain constant and we are not in the northest sector in the city without corner location	For each additional feature , the PRICE increases approximately 34.36573 hundread$ (units) if all the other explanatory variables are remain constant and we are not in the northest sector in the city without corner location	being in NE1(northest sector of the city)) increases the PRICE by about 30.00446 units. **(BUT THIS IS NOT RELIABLE due to the lack of statistical significance.)** if all the other explanatory variables are remain constant and we are without corner location	the presence of the COR1 characteristic (corner location)decreases PRICE by about 53.07940 units.(hundread$) if all the other explanatory variables are remain constant and we are not in the northest sector in the city .
Significance	p-value: 0.0454 <0.05 → Ho reject → the intercept is statistically significant	p-value<2e-16 → Ho reject → the SQFT is highly statistically significant	p-value: 0.3337>0.05→Ho doesn't reject→ the age isn't statistically significant	p-value: 0.0391<0.05→ Ho reject → the FEATS is statistically significant	p-value: 0.5339>0.05→Ho doesn't reject→ the NE isn't statistically significant	p-value: 0.2550>0.05→Ho doesn't reject→ the COR isn't statistically significant
Influence	-	Positive relationship : as SQFT increases, so does PRICE.	This relationship is not reliable due to the lack of statistical significance.	Positive relationship : as FEATS increases, so does PRICE.	This relationship is not reliable due to the lack of statistical significance.	This relationship is not reliable due to the lack of statistical significance.

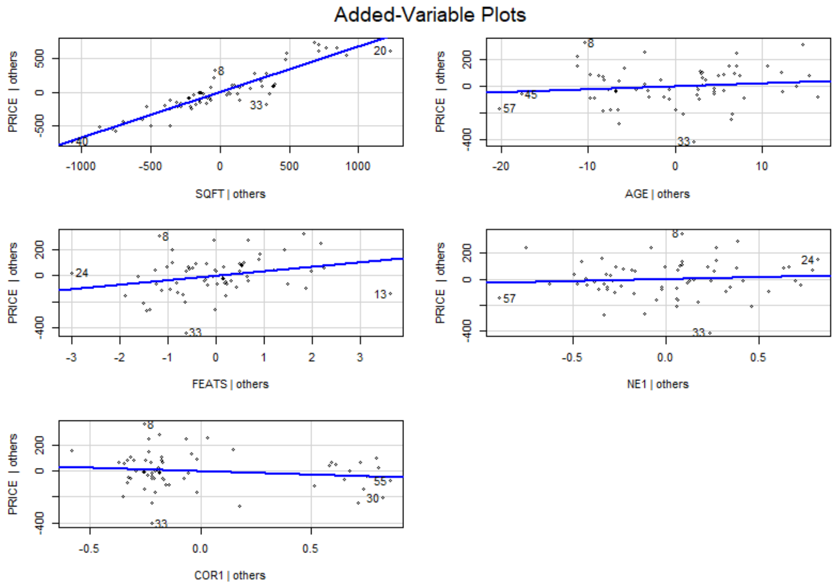
In short :

Significant explanatory variables : **SQFT** and **FEATS** are statistically significant and have a positive influence on **PRICE**(we need do include them in our model , the other explanatory variables we will not include them , but we will decide after the stepwise methods) .

- **Standard Error of Coefficients**: The 'std.error' in the coefficients shows the standard error of the estimates (the precision of the coefficient estimates).

- **Residual Standard Error**: 144.8. This measures the precision of the model's predictions and suggests that about 95% of the results will fall within a range of ± 2 * 144.8 thousand dollars. This range gives an idea of the variability around the predicted values of price .

The following plot support the foundings above :



Specifically in this plot (which is a visual representation of the individual relationship between the dependent variable PRICE and each of the explanatory variables in the context of a multiple regression model)

- If the line is upward (positive slope), there's a positive association between the explanatory variable and PRICE.
- If the line is downward (negative slope) we have a negative association.
- If the line is flat we have no association.

- The closer the points are to the fitted line, the stronger the relationship is between the explanatory and PRICE.
- If they're spread out, the relationship is weaker.

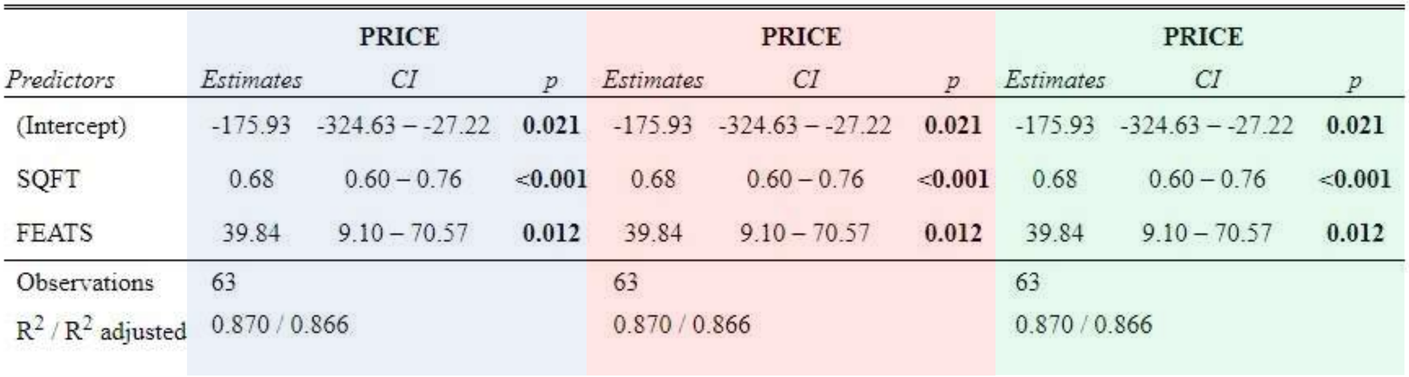
We can see that significant explanatory variables are SQFT and FEATS and have a positive influence on PRICE and the points are not very close to the fitted line but as we see from the summary of our model SQFT is highly statistically significant (the points are more close to the line than the points in FEATS with PRICE) .Also we can see the positive slops (positive association) .

(VI) Find the best model for predicting the selling prices (PRICE). Select the appropriate features using stepwise methods. (Hint: Use Forward, Backward or Stepwise procedure according to AIC or BIC to choose which variables appear to be more significant for predicting selling PRICES).

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are both used for model selection in statistical analysis (choosing between multiple competing models) .They help decide which model is the best fit for the data.

Between AIC and BIC for predicting is better the AIC and for interpretation is better the BIC . So , we will select to use AIC here .

I will use all the three step methods and I will see if we take the same results from each .



	PRICE			PRICE			PRICE		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	-175.93	-324.63 – -27.22	**0.021**	-175.93	-324.63 – -27.22	**0.021**	-175.93	-324.63 – -27.22	**0.021**
SQFT	0.68	0.60 – 0.76	**<0.001**	0.68	0.60 – 0.76	**<0.001**	0.68	0.60 – 0.76	**<0.001**
FEATS	39.84	9.10 – 70.57	**0.012**	39.84	9.10 – 70.57	**0.012**	39.84	9.10 – 70.57	**0.012**
Observations	63			63			63		
R^2 / R^2 adjusted	0.870 / 0.866			0.870 / 0.866			0.870 / 0.866		

As we can see here :

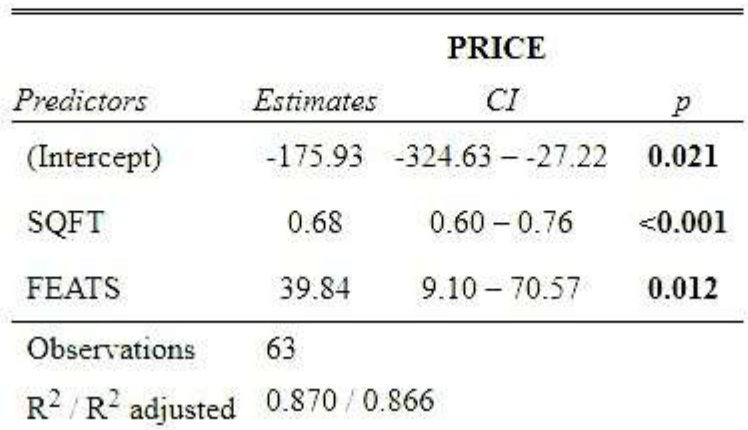
First we did a stepwise method with direction both (blue color result), that is means that we create the full model (all explanatory variables of the dataset included) and then will add and remove variables to find the best model based on AIC (the algoritm do this)

Then we did a stepwise method with direction back(red color result) , that is means that we started with the full model and sequentially the algoritm removes the least significant explanatory variables, one at a time, until the best model according to AIC is found.

And then we did a stepwise method with direction forward (green color result) , that is means that we started with the null model (only the intercepts has in the model) and sequentially the algoritm adds the most significant predictor variables, one at a time, until the best model according to AIC is found.

- ✓ **All the three step methods gave us the same result , which is : We must include in our model the significant explanatory variables SQFT and FEATS and the interecept** .(The estimates and the pvalues and why is statistically signifficant these specific variables are explained in the task (vii))

(VII) Get the summary of your final model, (the model that you ended up having after conducting the stepwise procedure) and comment on the output. Interpret the coefficients. Comment on the significance of each coefficient and write down the mathematical formulation of the model (e.g PRICES = Intercept + coef1*Variable1 + coef2*Variable2 +…. + ε , where ε ~ N(0, ...)). Should the intercept be excluded from our model?



	PRICE		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-175.93	-324.63 – -27.22	**0.021**
SQFT	0.68	0.60 – 0.76	**<0.001**
FEATS	39.84	9.10 – 70.57	**0.012**
Observations	63		
R^2 / R^2 adjusted	0.870 / 0.866		

	Intercept	SQFT	FEATS
Estimate	-175.93	0.68	39.84
Interpretation	-175.93 is the expected value of PRICE when SQFT and FEATS are zero . (this is not meaningfull because a house canot have) SQFT)	for each additional square feet, the PRICE increases by approximately 0.68hundread$ (units) if the other explanatory variable (FEATS) is remain constant	For each additional feature , the PRICE increases approximately 39.84 hundread$ (units) . if the other explanatory variable (SQFT) is remain constant Alternative : if we compare two houses with the same SQFT and a one-unit difference in FEATS, the house with the greater number of features is expected to have a higher PRICE by 39.84 hundred dollars according to the model
Significance	p-value: 0.021<0.05 → Ho reject → the intercept is statistically significant	p-value<0.001 <0.05 → Ho reject → the SQFT is highly statistically significant	p-value: 0.012<0.05→ Ho reject → the FEATS is statistically significant
Influence	-	Positive relationship : as SQFT increases, so does PRICE.	Positive relationship : as FEATS increases, so does PRICE.

From the p-values we can see that SQFT is higher statistically significant than the FEATS.

Also , from the summary of our model we see that :

- **Adjusted R^2**: Show the proportion of variance explained by the model, adjusted for the number of variables. A value > 0.7 show good model fit and tell us that the model makes good predictions. In our case , we have Adj- R^2= 0.866 >0.7 , so its is very good.
- **Standard Error of Coefficients**: The 'std.error' in the coefficients shows the standard error of the estimates (the precision of the coefficient estimates)
  - The standard error **of the intercept** is 18.10 (1.810e+01). This value indicates the precision of the estimate of the baseline 'PRICE' when all other variables (SQFT and FEATS) are held at zero. (smaller standard error → more reliable estimate of the intercept)

- The standard error **of the 'SQFT'** coefficient is 0.03868 (3.868e-02). This show a high level of precision in the estimate of the 'SQFT' effect on 'PRICE' (we can be more “sure” about the influence of square feet on the price from this dataset-sample)
- The standard error **for the 'FEATS'** coefficient is 15.37 (1.537e+01). This doesn’t show a high level of precision in the estimate of of how the number of features -FEATS affects. (the exact size of the effect is less certain than for SQFT)

➢ **Residual Standard Error**: 143.7. This measures the precision of the model's predictions and suggests that about 95% of the results will fall within a range of ± 2 * 143.7 thousand dollars. This range gives an idea of the variability around the predicted values of price .

PRICE = −175.93 + 0.68*SQFT + 39.84 * FEATS + ε , ε ~ N(0, 143.7^2)

Why to include the intercept in our model :

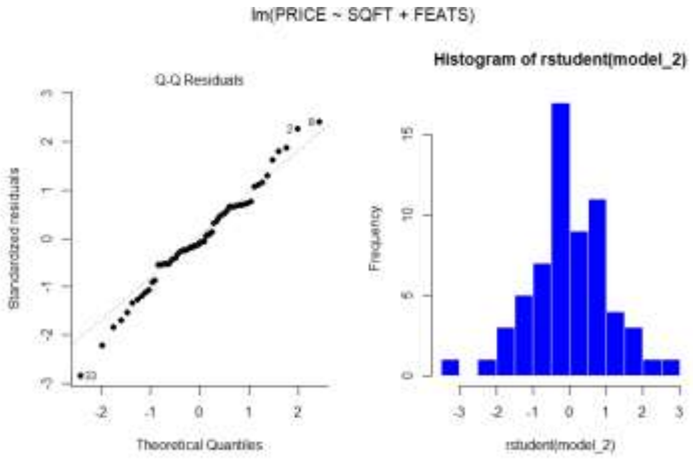
It is the expected value of PRICE when all the explanatory variables is 0. Here is not meanigful (in the price of a house is not make sense to interpret a situation where 'SQFT' and 'FEATS' are both zero, also there is no a house that someone pays us to buy it), however here the p-value: 0.021<0.05 → Ho reject → the intercept is statistically significant .

(VIII) Check the assumptions of your final model. Are the assumptions satisfied? If not, what is the impact of the violation of the assumption not satisfied in terms of inference? What could someone do about it?

Assumptions Analysis for Regression Residuals:

1. **Normality of Errors:**

- I will use studentized residuals to check normality.
- The normality will be visually inspected with Q-Q plots (we can do and a histogram) and tested with normality tests such as the Shapiro-Wilk and lillie test (kolmogorov-Smirnof) test applied to the studentized residuals.



From the qq-plot of the studendized residuals we see some deviation from the line at the ends, (heavy tails) which show the residuals may not be normally distributed.

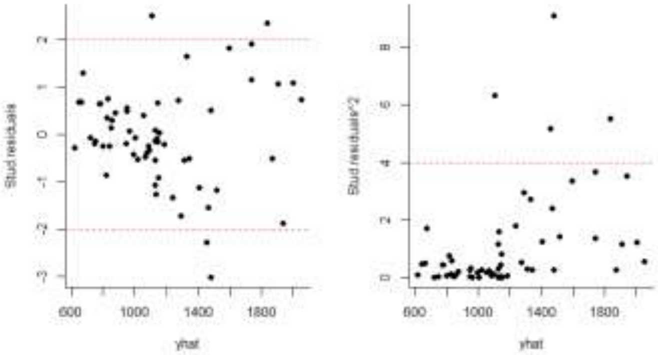
From the histogram of the studendized residuals we see that does have a bell shape, but it's not perfectly symmetric, showing some skewness

Also we run a Shapiro-Wilk test (p-value = 0.4591 >a =0.05 → H0: normality doesn't rejected) and a Lilliefors test (p-value = 0.04856<a =0.05 → H0: normality rejected) for the studendized residuals .

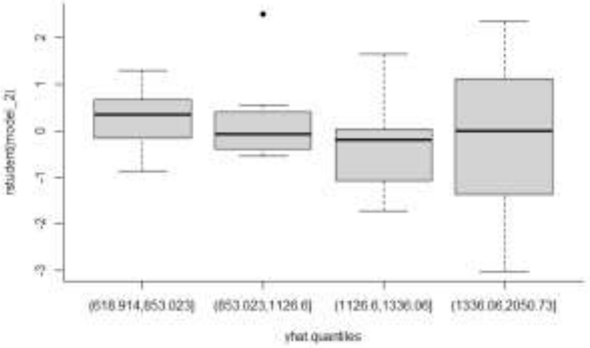
In total , the assumption of normalilty is violated.

2. **Homoscedasticity of Errors:**

    - To check for constant variance of the residuals, I will plot the studentized residuals against the fitted values. If the variance appears to be constant across the range of fitted values, this assumption is satisfied.
    - Also , I wil do Levene's Test to check the variances of the studentized residuals are equal across all levels of the independent variables (H0) (p-value=2.249e-05 <a → Ho rejected) .



From the plots of the studendized residuals against the predicted values (yhat) there appears to be some spread that increases with yhat, which might show potential heteroscedasticity (non-constant variance of residuals) (we see a lot points out of the red lines)



The boxplots show potential heteroscedasticity.

Varying spread of residuals across the quartiles of fitted values.

Differences in spread between the outer and middle quartiles.

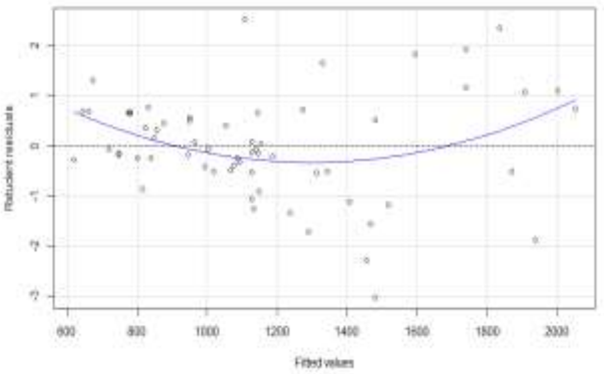
The assumption of Homoscedasticity of Errors is violated .

3. **Independence of Errors:**

    - This assumption is typically checked when data involves a time sequence. As our data does not have a time sequence, we'll proceed under the assumption that the residuals are independent.

4. **Linearity:**

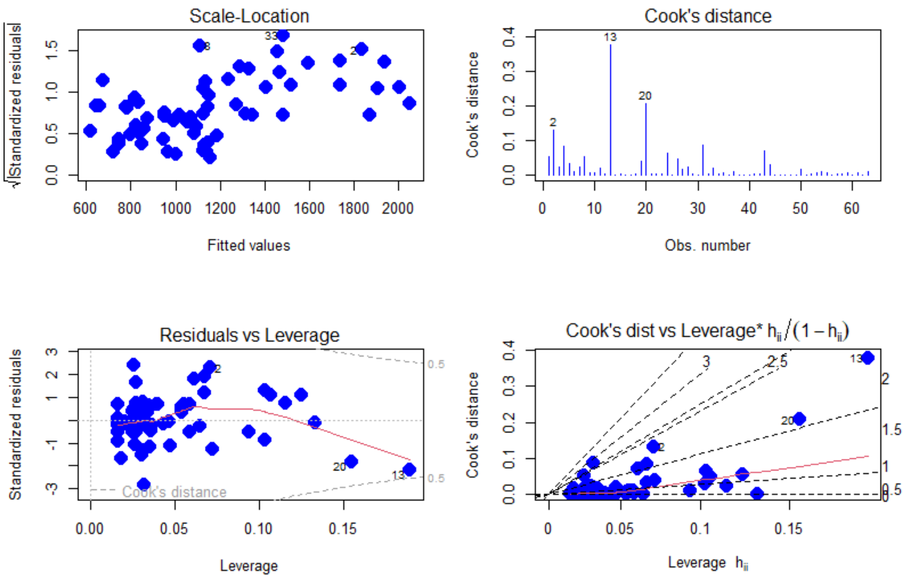
- The linearity assumption will be checked by plotting the residuals(student residuals) versus the fitted values.
- I will conduct Tukey's test (H0: is that the model is correctly specified as linear p-value=0.009317 <a →Ho rejected)



The curve of the blue line show a non-linear relationship . For a linear model, we want the residuals to be randomly scattered around the horizontal line at zero, without any specific pattern.

<u>The assumption of lienearity between the depentent variables and the dependent variable is violated</u>

Furthermore, I will also assess for outliers using **leverage plots and Cook's distance**, although these are not formal assumptions of linear regression. They are important to identify influential data points .



From the plots I see that there may be some issues with homoscedasticity and that a few data points could be do influence on the model. Observation 13 and 20 shows that have high Cook's distances, so may they are influential points and could be affecting the stability and interpretation of the model..

<u>Multicollinearity Check</u>

First, I look at the Pearson correlations and histograms to see how the variables relate to each other one-on-one and to understand their patterns (see task (IV)) .

But this doesn't show how all the variables might work together. Next, I use a command VIF to make sure my variables aren't overlapping too much (multicollinearity, it means that some of the independent variables in the model are explaining much of the same variance in the dependent variable as other independent variables in the model.).

If the VIF number is under 10 we don’t have multicollinearity problem . Here we take 1.15 <10 so we will not have this problem .

What can someone do about it :

The assumptions for linear regression may not be 100% satisfied in practice, but large deviations from these assumptions should be avoided to ensure the stability of the model.

If violations are detected, corrective actions such as transformations may be applied to the model .

Also , if after the transoformation we don’t have again satisfied assumptions.
We can do tests for leverage points and remove specific obserbvation of our dataset .
However the outliers and levege points , a lot of times are important to see valuable information about the behavior of the population being studied. Removing them might lead to an oversimplified or biased model that doesn't accurately reflect reality.
Here we can write the model :

log(PRICE) = 5.959+ 0.0005402*SQFT + 0.0285 * FEATS + ε , ε ~ N(0, 0.1079^2)

and all Assumptions Analysis for Regression Residuals will be fixed .

(IX) Conduct LASSO as a variable selection technique and compare the variables that you end up having using LASSO to the variables that you ended up having using stepwise methods in (VI). Are you getting the same results? Comment

LASSO prevent overfitting and perform variable selection.

As the log lambda increases:

Some coefficients quickly shrink towards zero, indicating that LASSO is believe that them less important in the presence of regularization. Other coefficients remain non-zero even as the penalty increases, that says that these explanatory variables are more important in predicting the response variable.

Stepwise regression methods, such as forward, backward, or both, add or remove ot both variables based on some criterion (e.g. AIC) , e.g. SQFT has higly statistically signifficant and the FEATS was statistically signifficant , so these methods choose them .

The lasso suggest to include in the model the explanatory variables FEATS and SQFT (as the stepwise suggest too) and also the intercept . So , lasso and stepwice methods gave us the same result .

Appendix

Rcode

```
install.packages("sjPlot")
library(sjPlot)
library(foreign)
library(psych)     # for the describe function
library(DescTools)  # to calculate the mode
library(corrplot)   #for the corrplot
library(nortest) # for lillie.test
library(lawstat) # symmetry.test
library(moments)  #agostino.test , test for skewness ( δεν υπαρχει πια η normtest)
library(Hmisc)    # cut function
library(car)      # for levene test
library(gplots)   # for error bars
library(gmodels)  #for cross table
```

```
# Tasks:
# (I)  Read the “usdata” dataset and use str() to understand its structure.
file.choose() # μου έβγαλε το path αφου επέλεξα το αρχειο txt που κατέβασα
as2 <- read.table("C:\\Users\\eleni\\Downloads\\usdata.txt", header=TRUE)
str(as2) # για να δούμε λίγο την μορφή του
head(as2)
```

```
# (II) Convert the variables PRICE, SQFT, AGE, FEATS to be numeric variables and NE, COR to be  factors.
as2$NE <- as.factor(as2$NE)   #the int is numeric already however lets do it for the question
as2$COR <- as.factor(as2$COR)    #the int is numeric already however lets do it for the question
as2$PRICE <- as.numeric(as2$PRICE)   #the int is numeric already however lets do it for the question
as2$SQFT <- as.numeric(as2$SQFT)       #the int is numeric already however lets do it for the question
```

```
as2$AGE <- as.numeric(as2$AGE)
as2$FEATS <- as.numeric(as2$FEATS)
str(as2)   # ok
```

```
# (III) Perform descriptive analysis and visualization for each variable to get an initial insight of what  the data looks like. Comment on your findings.
summary(as2)  #min, Quartiles, max or frequencies if categorical
describe(as2)
#keep the numeric-variables :
index <- sapply(as2, is.numeric)
as2_num <- as2[index]
head(as2_num)

#mode (η επικρατουσα τιμή)
for (i in 1:4) {
  print(Mode(as2_num[, i]))
}
for (i in 1:4) {
  mode_value <- Mode(as2_num[, i])
  mode_freq <- sum(as2_num[, i] == mode_value)
  cat("The mode for", names(as2_num)[i], "is", mode_value, "with frequency of", mode_freq, "\n")
}

#Range
for (i in 1:4) {
  print( range(as2_num[,i]) )
}

#i will create this function to check quickly all these
mean_median_skew_kurt <- function(data_vector) {
  if (!is.numeric(data_vector)) {
    stop("Data vector is not numeric.")
  }

  results <- list(
    Mean = mean(data_vector, na.rm = TRUE),
    Median = median(data_vector, na.rm = TRUE),
    Skewness = skewness(data_vector, na.rm = TRUE),
    Kurtosis = kurtosis(data_vector, na.rm = TRUE)
  )

  print(results)

}

mean_median_skew_kurt(as2$PRICE)
mean_median_skew_kurt(as2$SQFT)
mean_median_skew_kurt(as2$AGE)
mean_median_skew_kurt(as2$FEATS)

#standard deviation

for (i in 1:4) {
  print( sd(as2_num[,i]) )
}


# visualizations :
#Visual Analysis for numerical variables   (για τις συνεχεις θα κάνω histogram  για τις διακριτές θα κανω barplot)
par(mfrow=c(2,3)); n <- nrow(as2_num)
hist(as2_num[,1], main=names(as2_num)[1], col= "lavender" )
hist(as2_num[,2], main=names(as2_num)[2], col="lavenderblush")

hist(as2_num[,3], main=names(as2_num)[3] ,col = "honeydew")

plot(table(as2_num[,4])/n, type='h', xlim=range(as2_num[,4])+c(-1,1), main=names(as2_num)[4], ylab='Relative frequency')

#PRICE hist:  κατανομή με μία κορυφή- δεν κατανέμονται κανονικά,βλέπω αρκετα μεγάλη διασπορά κατανομής με μερικές πώλησεις στα 2000 εκατ δολ.

#           η πλειοψηφία των σπιτιών πωλήθηκε σε τιμή γύρω στα 800-1000 εκατοντάδες δολάρια.

# SQFT hist:   κατανομή με δύο κορυφές, μία γύρω στα 1000 τετραγωνικά πόδια και μία άλλη γύρω στα 2000 τετραγωνικά πόδια, οχι καννική κατ.

#           opote βλέπω ότι οι περισσότερες πωλήσεις συγκεντρώνονται σε αυτά τα δύο μεγέθη.

# AGE hist :    κατανομή με περισσότερες τιμές στις μικρότερες ηλικίες σπιτιών ,αλλα υπάρχει  σημαντική συχνότητα σπιτιών  στα 25-30 ετών.

# FEATS hist : διακριτή μεταβλητή αρα διακριτή  κατανομή με μεγαλύτερη συχνότητα τιμής = 4


#Visual Analysis for factors   # μπορείς και πίνακες για συχνότητες

as2_fac <- as2[,!index]                   # keep the factor variables

par(mfrow=c(1,1),oma=c(1,4,1,1))

barplot(sapply(as2_fac,table)/n, horiz=T, las=1, col=2:3, ylim=c(0,8), cex.names=1.3)

legend('top', fil=2:3, legend=c('No','Yes'), ncol=2, bty='n',cex=1.5)

sjt.xtab(as2$NE, as2$COR) # πινακας για συχνότητες , ουσιαστικά το ίδιο με table(as2$NE, as2$COR)

tab_xtab(as2$NE, as2$COR, show.n = FALSE, show.row.prc = TRUE, show.col.prc = TRUE) #  το ιδιο με prop.table(table(as2$NE, as2$COR), 1) και prop.table(table(as2$NE, as2$COR), 2)

#  NE = Located in northeast sector of city (1) or not (0) , so we have more houses in  norteset secton in this dataset ( 39 vs 24)

#  COR = Corner location (1) or not (0).  , so we have more houses that are not in corner location  (49 vs 14)
```

```
# (IV) Conduct pairwise comparisons between the variables in the dataset to investigate if there are

#     any associations implied by the dataset.(Hint: Plot variables against one another and use correlation

#     plots and measures for the numerical variables.). Comment on your findings.

#     Is there a linear relationship between PRICE and any of the variables in the dataset?

str(as2)
```

```
#ok now we will examining and analyzing the relationships between each pair of variables in the dataset.

#but first we must do analysis fir each variable separately

# we Performed descriptive analysis and visualization for each variable in the question (III)

# but now i want to check and with test if the numeric variables follow normal distribution ( tests for normality)


#Ho:The sample is from a normally distributed population.

#H1:The sample  is not from a normally distributed population.


for(column in names(as2_num)) {

    print(paste("Lilliefors test for", column))

    print(lillie.test(as2[[column]]))


    print(paste("Shapiro-Wilk test for", column))

    print(shapiro.test(as2[[column]]))

  }

str(as2_num) #είναι n>50 , οποτε πάμε να δούμε αν ο μέσος είναι κατάλληλο μέτρο περογραφης της κεντρικής θέσης( στη κανονική ο mean και ο median συμπεφτουν )

par(mfrow = c(1, 4))

for (i in names(as2_num)) {

  boxplot(as2_num[[i]],col = c( "lavender","lavenderblush1","honeydew","blue" ) ,main = paste("Boxplot of", i), ylab = "Value")

}


## οκ ας πουμε οτι δεν είναι καταλληλος ο μέσος και ας πάμε με

# Βασικα αππέριψα την κανονικότητα αρα θα πάμε με spearman για να δούμε αυξουσα και και φθινουσα σχέση( μη παραμτερικα)

# και επιπλεον για να κανω ta cor.test θα βάλω μεθοδο spearman



#Test for normality for each group γαι τις κατηγορικές ( τα δειγματα ειναι μικρα κυριως το shapiro θα κοιταω )

as2_1 <- data.frame(PRICE=as2$PRICE,NE=as2$NE)

#by(as2_1$PRICE, as2_1$NE, lillie.test) # αποριψη κανονικότηατς kai sta 2 gkroyp
```

```
by(as2_1$PRICE, as2_1$NE, shapiro.test) # αποριψη κανονικότηατς kai sta 2 group

table(as2_1$NE) # NO large samples ΠΑΩ ΜΗ ΠΑΡΑΜΕΤΡΙΚΑ

wilcox.test(as2_1$PRICE~ as2_1$NE) # ΑΠΟΡΡΙΨΗ ΟΙ ΔΙΑΜΕΜΕΣΟΙ ΔΕΝ ΜΠΟΡΟΥΝ ΝΑ ΘΕΩΡΗΘΟΎΝΗ ΙΣΕΣ ΣΤΑ 2 ΓΚΡΟΥΠ , ΚΑΙ ΒΛΕΠΩ ΚΑΙ ΤΑ BOXPLOT


as2_2<-data.frame(PRICE=as2$PRICE,COR=as2$COR)

by(as2_2$PRICE, as2_2$COR, shapiro.test) # αποριψη κανονικότηατς  στο γκρουπ με COR=O στο αλλο οχι

table(as2_2$COR) # NO large samples ΠΑΩ ΜΗ ΠΑΡΑΜΕΤΡΙΚΑ

wilcox.test(as2_2$PRICE~ as2_2$COR) # DEN ΑΠΠΡΟΡΙΠΤΩ εχω ενδειξη οτι ΟΙ ΔΙΑΜΕΜΕΣΟΙ  ΜΠΟΡΟΥΝ ΝΑ ΘΕΩΡΗΘΟΎΝΗ ΙΣΕΣ ΣΤΑ 2 ΓΚΡΟΥΠ , ΚΑΙ ΒΛΕΠΩ ΚΑΙ ΤΑ BOXPLOT


par(mfrow=c(1,2))

# Boxplot for PRICE across NE levels

boxplot(PRICE ~ NE, data = as2, main = "Boxplot of PRICE by NE", xlab = "NE", ylab = "PRICE (in hundreds)", col="lavender")


# Boxplot for PRICE across COR levels

boxplot(PRICE ~ COR, data = as2, main = "Boxplot of PRICE by COR", xlab = "COR", ylab = "PRICE (in hundreds)", col="lavender")



#first we will do a visualization of bivariate associations ( cor plots ).

#-->Pairs of numerical variables:

pairs(as2_num) # για να δω πως μοιαζει η σχέση των numeric μετβλατητών

        # from the scatterplots i can see prrice and sqft έχουν μια σχέση που μοιαζει κάπως γραμμική

        #για την feats τα scatterplots είναι έτσι γιατι είναι διακριτή

corrplot(cor(as2_num),method= "ellipse") #για να δω κατα πόσο η σχέση των numeric μεταβλητών είναι γραμμική

cor(as2_num)

par(mfrow=c(1,3)) # 4 numeric εκ των οποιοων η μια ειναι η εξαρτημένη μου οποτε θα χρειαστώω 3 θέσεις

for(j in 2:4){

  plot(as2_num[,j], as2_num[,1], xlab=names(as2_num)[j], ylab='Price',cex.lab=1.5)

  abline(lm(as2_num[,1]~as2_num[,j])) }
```

```
#θα κάνω και μερικά boxplot  να το δω και ετσι ( αν η numeric δεν είναι διακριτή δεν έχει νόημα )
#par(mfrow=c(1,3))
#for(j in 2:4){
#  boxplot(as2_num[,1]~as2_num[,j], xlab=names(as2_num)[j], ylab='Price',cex.lab=1.5)
#  abline(lm(as2_num[,1]~as2_num[,j]),col=2)}
 par(mfrow=c(1,1))
boxplot(as2_num[,1]~as2_num[,4], xlab=names(as2_num)[4], ylab='Price',cex.lab=1.5)
abline(lm(as2_num[,1]~as2_num[,4]),col=2)

#_______________________________________________________________________
par(mfrow=c(1,2)) # 2 κατηγορικές έχω
for(j in 1:2){
  boxplot(as2_num[,1]~as2_fac[,j], xlab=names(as2_fac)[j], ylab='Price',cex.lab=1.5)
}
#_________________________________________________________________________________________________________

par(mfrow = c(1,1))
corrplot(cor(as2_num), method = "number") # δεν νομίζω οτι θα έχω θεματα πολλυσυγγραμικότητας

cor_matrix <- cor(as2_num, method = "pearson") # βεβαια για να δουμε και το
cor_matrix <- cor(as2_num, method = "spearman") # βεβαια για να δουμε και το
cor.test(as2_num$PRICE,as2_num$SQFT,method = "spearman")
cor.test(as2_num$PRICE,as2_num$AGE,method = "spearman")
cor.test(as2_num$PRICE,as2_num$FEATS,method = "spearman")


#Relationship between SQFT ,AGE ,FEATS  and the categorical variables:

by(as2$SQFT, as2$NE, shapiro.test) #not rejected
var.test(as2$SQFT~ as2$NE) #reject Ho
t.test(as2$SQFT~as2$NE, var.equal=F) #reject Ho
```

```
library(gplots)

plotmeans(as2$SQFT~as2$NE)


by(as2$AGE, as2$NE, shapiro.test) #normality rejected

wilcox.test(as2$AGE~as2$NE)

boxplot(as2$AGE~as2$NE)


by(as2$FEATS, as2$NE, shapiro.test) #normality rejected

wilcox.test(as2$FEATS~as2$NE) #rejected Ho

boxplot(as2$FEATS, as2$NE)


#_________________________________

by(as2$SQFT, as2$COR, shapiro.test)

wilcox.test(as2$SQFT~as2$COR)

boxplot(as2$SQFT, as2$COR)


by(as2$AGE, as2$COR, shapiro.test)

wilcox.test(as2$AGE~as2$COR)


by(as2$FEATS, as2$COR, shapiro.test)

wilcox.test(as2$FEATS~as2$COR)


par(mfrow=c(1,2)) # 2 κατηγορικές έχω


for(j in 1:2){

  boxplot(as2_num[,2]~as2_fac[,j], xlab=names(as2_fac)[j], ylab='SQFT',cex.lab=1.5)

}


par(mfrow=c(1,2)) # 2 κατηγορικές έχω

for(j in 1:2){

  boxplot(as2_num[,3]~as2_fac[,j], xlab=names(as2_fac)[j], ylab='AGE',cex.lab=1.5)
}

par(mfrow=c(1,2)) # 2 κατηγορικές έχω
for(j in 1:2){
  boxplot(as2_num[,4]~as2_fac[,j], xlab=names(as2_fac)[j], ylab='FEATS ',cex.lab=1.5)
}

par(mfrow=c(1,1))

#Relationship between the categorical variables NE and COR:
tab <- table(as2$NE, as2$COR)
p <-prop.table(tab)
chisq.test(tab) #not reject
barplot(prop.table(tab), main="NE by COR", beside=T,ylim=c(0,0.6),legend=levels(as2$NE))
```

```
# (V) Construct a model for the expected selling prices (PRICE) according to the remaining
#     features.(hint: Conduct multiple regression having PRICE as a response and all the other variables
#     as predictors). Does this linear model fit well to the data? (Hint: Comment on R^2 adj ).

str(as2)

model_full <- lm(PRICE ~., data = as2)          # τρέχει το φουλ μοντέλο , τρέχει όλες τις μεταβλητές
summary(model_full)
# εδω βλέπουμε τα εξης :
# πρωτο πραγμα που κοιτάω είναι το F statistic  , που μου λέει αν το μοντέλο διαφέρει απο το σταθερό ,p-value: < 2.2e-16 απορiπτω HO οποτε οκ για αρχη
# οσον αφορα τους συντελεστες:
# θέλουμε να απορρίπτεται ο έλεγχος Ho: βi=0 , ώστε να έχει νόημα να το βάλω στο μοντέλο να ειναι στατιστικά σημαντική η i επεξηγηματική ( signifficant effect)
```

# για τη Υ (να συνεισφέρουν σημαντικά στην επεξήγηση της Υ)

# επισης κοιτάω π΄ροσημο , αν επηρεάζουν θετικα ή αρντητικά και το κατα πόσο επηρεάζουν

# βλέπω οτι δεν έχει νόημα να βάλω την AGE , NE και THN COR γιατι σύμφωνα με τα p.value δεν απορ η H0

# to std.error στα coefficients είναι το standard error of the estimates

#βλεπω residual std error =144.8 , αυτο measures the precision of the model predictions και ουσιαστικα και σου λεει το 95 % των αποτελεσμάτων θα είναι μέσα σε ένα εύρος

# +-2 * 144.8 χιλιάδες δολάρια

#βλεπώ και το R^2 adj = 0.864 που είναι το ποσοστο διακυμανσης που εξηγεί το μοντέλο διορθωμνεο για τον αριθμο των μεταβλητών

#>0.7 καλες προβλέψεις goof fit .

#(ας δουμε διάγραμμα προστιθέμενης μεταβλητής, ΚΑΛΑ ΕΔΩ οχι σε μεγαλυτερα dataset

#το οποίο είναι εξτρα γράφημα στην πολλαπλή παλιονδρόμση και μου επιτρέπει να παρω μια 1η γεύση

#του τι να περιμένω με την σχέση της εξαρτημένης μου μεταβλητής με τις υποψήφιες επεξηγηματικές)

# ας δούμε και ένα διαγραμμα προστιθέμενης μεταβήτης

```
library(car)

avPlots(model_full)
```

# οπου βλέπω μια ευθεία οριζόντια γραμμή , λογικό να μην είναι στατ σημαν κα να μην συνεισφέρει

# (VI) Find the best model for predicting the selling prices (PRICE). Select the appropriate features

# using stepwise methods. (Hint: Use Forward, Backward or Stepwise procedure according to AIC or BIC to choose which variables

# appear to be more significant for predicting selling PRICES).

# για predictions είναι καλύτερο το AIC , για επεξήγηση είναι καλυτερο το BIC

```
#stepwise με αρχικό μοντέλο το full & AIC ,BIC
##stepwice procedure - προσθαφαιρώ μεταβλητες , θα ξεκινήσω απο το πληρες
model_full <- lm(PRICE ~., data = as2)          # τρέχει το φουλ μοντέλο , τρέχει όλες τις μεταβλητές
step1<-step(model_full,direction ="both")
step1
summary(step(model_full,direction ="both"))     # το περιμενα οτι θα κρατησει μόνο αυτες

#backward procedure -ξενινάω απο το πλήρες μοντέλο και αφαιρω μεταβλητες
model_full <- lm(PRICE ~., data = as2)
step2<-step(model_full,direction ="back")
step2
summary(step(model_full,direction ="back"))

#forward procedure - ξεκινάω έχοντας μόνο την σταθερά μου
model_0<-lm(PRICE ~1, data = as2)
step3<-step(model_0,scope=list(lower=model_0, upper=model_full),direction='forward')
step3
summary(step(model_0,scope=list(lower=model_0, upper=model_full),direction='forward'))

library("sjmisc")
tab_model(step1,step2,step3)

#το μοντέλο που επιλέχθηκε με αυτή την διαδικάσια και κριτήριο το AIC
library(sjPlot)
tab_model(step(model_full,direction ="both")) # είναι το πρώτο που βλέπεις στο επάνω κουτάκι
```

```
# (VII) Get the summary of your final model, (the model that you ended up having after conducting
#      the stepwise procedure) and comment on the output. Interpret the coefficients. Comment on the
#     significance of each coefficient and write down the mathematical formulation of the model (e.g

#    PRICES = Intercept + coef1*Variable1 + coef2*Variable2 +…. + ε , where ε ~ N(0, ...) ). Should the intercept be excluded from our model?


summary(step(model_full,direction ="both"))

tab_model(step(model_full,direction ="both"))


model_2<-lm(PRICE ~ SQFT + FEATS , data = as2)

summary(model_2)


install.packages("stargazer")

library(stargazer)

stargazer(model_2, type = "text") # to have it in a table
```

```
# (VIII) Check the assumptions of your final model. Are the assumptions satisfied? If not, what is the

#      impact of the violation of the assumption not satisfied in terms of inference? What could someone do about it?


#ΓΕΝΙΚΑ παίρνω μια πρώτη εικόνα απο τους συντελ συσχετ. κατα pearson και τα ιστογράμματα αλλά έτσι δεν λαμβάνω την επιρροή

#που θα έχω με πολλές μεταβλητές μαζι- μεταξύ τους( πάνω απο ανα 2 που κοιτάει ο pearson)

library(car)

vif(model_2) #δεν αντιμετωπίζω πρόβλημα πολλυσυγραμμικότητας vif<10 , καλα το είχα δει και απο το corrplot οτι δεν παιζει μαλλον θεμα


#προυποθέσεις(ανάλυση καταλοποιπών): (δεν θα πληρούνται συνήθως 100% -αρκεί να μην έχω μεγάλες αποκλίσεις απο αυτές)

#(αν υπάρχουν παραβιάσεις γυρίζω πίσω και διορθώνω το μοντέλο μου μέσω μετασχημ. κλπ)(εξασφαλίζουν την "σταθερότητα" του μοντέλο μου )

# 1)normality of errors ( i will use studendized residuals )--> qqplots + normality tests for the studendized residuals

# 2)homoscedacity of errors( check the assumption of constant variance)--> plot of studendized residuals vs the fitted values
# 3)check the independence of errors ( only if a time sequence applies) here we dont have time sequence
# 4)linearity( plot for X vs Y και residuals vs fitted value using the residual plot + apply Tukeys test for quadratic terms , οκ εδω δεν εχω τετοια  )
# και θα τσεκάρω αν και δεν ειναι μέσα στις προυποθέσεις και τα outliers με plot leverage και cook distance


#1)
Stud.residuals <- rstudent(model_2)
plot(model_2, which = 2) #Normality of the residuals( ελέγχο την διαφορα της εκτίμησης απο το actual) # κακο
par(pch=16, bty='l')
hist(rstudent(model_2), breaks=10, border='white', col="blue")  #ιστόγραμμα των τυποποιημένων καταλοίπων ( καλα δεν χρειάζεται τοσο )


#συνοδεύω τους παραπάνω οπτικούς ελέγχους με στατιστικούς ελέγχους
#θα εφαρμόσω και τα 2 τεστ αφου έχω περισσότερες απο 50 παρατηρήσεις
shapiro.test( rstudent(model_2) ) #των  καταλοίπων p-value>0.05=> HO ΔΕΝ απορρίπτεται
lillie.test( rstudent(model_2) )#των καταλοίπων  p-value <0.05=> HO  απορρίπτεται


#2)
library(car)
# ncvTest(model_2)    # ( υποθέτει κανονικότητα καταλοίπων ) για να δούμε ομοσλεδατσικότητα , αποριψη της H0 , η διακυμανση των καταλοίπων δεν είναι σταθερη
#the Breusch-Pagan test to assess the homoscedasticity (constant variance) of the residuals from a linear regression mode
#pvalue<a--> Ho απορρίπτεται
yhat.quantiles<-cut(yhat, breaks=quantile(yhat, probs=seq(0,1,0.25)), dig.lab=6)  # μποεςι; να σκαπσεις τα καταλοιπα να δεις πως πάει η διακύμναση σε κάθε γκρούπ
table(yhat.quantiles)
leveneTest(rstudent(model_2)~yhat.quantiles)  #p.value= 2.249e-05 <α --> Ho απορ , ΔΕΝ Μπορω να θεωρησω ίσες διακυμανσεις
par(mfrow=c(1,1))
boxplot(rstudent(model_2)~yhat.quantiles)      # παρατηρω η διακυμναση τους δεν ειναι το ίδιο το βλέπεις και απο το p value του levene
```

```
yhat <- fitted(model_2)  # κταταω τις εκτιμήσεις
par(mfrow=c(1,2))
plot(yhat, Stud.residuals)
abline(h=c(-2,2), col=2, lty=2)
plot(yhat, Stud.residuals^2)
abline(h=4, col=2, lty=2)  # έχουμε θεμα , πολλά καταλοιπα εκτος των οριών που θα θέλαμε


#4)linearity check
par(mfrow=c(1,1))
library(car)
residualPlot(model_2, type='rstudent')   #θες αυτα να είναι ευθέια και να απλβωνονται σωστα γυρω απο την ευθέια που εδω δεν γινεται
residualPlots(model_2, plot=F, type = "rstudent")       # απλως του λες μην μου εμφανίσεις το Plot, το tukey αποριπτει


#η γραμμικότητα ειναι η πιο σημαντική apo τις σχέσεις γιατι εάν έχουμε έστω λίγο γραμμικότητα κάτι μπορεί να γίνει με το μοντέλο
# ας κάνω και έναν έλεγχο για outliers
#οποτε τώρα θα πάω να τις αφαιρέσω και να κάνω όλο το μοντελο μου χωρίς αυτή
#( προσπαθώ να πηράζω τα πραγματικά μου δεδομένα όσο λιγότερο γίνεται, για αυτο και θα κάνω ελέγχους να δω άμα μου επιτρέπεται να αφαιρέσω αυτη την τιμή )


#ελεγχοι outliers κλπ για ακραιές παρατηρησεις   για να δω αν όντως προκύπτει ακραία
#και δεν έχω κάποια γνωση απο πηγή ότι πρέπει να παραμείνει μέσα , θα την αφαιρέσω


par(mfrow=c(2,2))
plot( model_2 , pch=16, cex=2, col='blue', add.smooth=F, which=3)
plot( model_2 , pch=16, cex=2, col='blue', which=4)
abline(h=4/5, col='red', lty=2, lwd=2)
plot( model_2 , pch=16, cex=2, col='blue', which=5)
plot( model_2, pch=16, cex=2, col='blue', which=6)
# λίγο η 20 και η 13 παρατήρηση δημιουργουν θέμα αλλα θα πάω να κάνω έναν μετασχηματισμό και αν δεν φτιάξει θα τις αφαιρέσω
```

```
#γενικά
#τα σημεία μοχλευσης =δεν επηρεάζουν ιδιαιτερα τους συντελ. παλλινδρόμησης αλλα μπορεί να επιδράσουν
#στο συντελεστή προσδιορισμού R^2 και στα τυπικά εκτιμώμενα σφάλματα των συντελεστών παλινδρόμησης αν παραμείνουν μέσα στο μοντελο.
#τα σημεία επιρροής = επιδρούν σημαντικά στους συντελεσες παλινδρόμησης και τραβούν την γραμμή παλινδρομ.
#προς την κατεύθυνση στην οποία βρίσκονται


# οποτε πάμε να δούμε με μετασχηματισμό , γενικά
model_2<-lm(PRICE ~ SQFT + FEATS , data = as2)
summary(model_2)
par(mfrow=c(1,2))
plot(lm(PRICE ~ SQFT + FEATS , data = as2_c.num),2, main='Price')
plot(lm(log(PRICE)~SQFT + FEATS,data = as2_c.num),2, main='log of price')


logmodel_2<-lm(log(PRICE)~SQFT + FEATS,data = as2)
par(mfrow=c(2,2))
plot( logmodel_2, 2 )
plot( logmodel_2, 3 )
residualPlot(logmodel_2, type='rstudent')
plot(rstudent(logmodel_2), type='l')


summary( logmodel_2)


# ΠΑΜΕ ΠΑΛΙ ΕΛΕΓΧΟ ΥΠΟΘΕΣΕΩΝ
#1)
Stud.residuals.log <- rstudent(logmodel_2)
plot(logmodel_2, which = 2) #Normality of the residuals( ελέγχο την διαφορα της εκτίμησης απο το actual) # κακο
par(pch=16, bty='l')
hist(rstudent(logmodel_2), breaks=10, border='white', col="blue")  #ιστόγραμμα των τυποποιημένων καταλοίπων ( καλα δεν χρειάζεται τοσο )
yhatlog <- fitted(logmodel_2)  # κταταω τις εκτιμήσεις
```

```
par(mfrow=c(1,2))

plot(yhat, Stud.residuals.log)

abline(h=c(-2,2), col=2, lty=2)

plot(yhat, Stud.residuals.log^2)

abline(h=4, col=2, lty=2)  # ισως έχουμε θεμα , πολλά καταλοιπα εκτος των οριών που θα θέλαμε

#συνοδεύω τους παραπάνω οπτικούς ελέγχους με στατιστικούς ελέγχους

#θα εφαρμόσω και τα 2 τεστ αφου έχω περισσότερες απο 50 παρατηρήσεις

shapiro.test( rstudent(logmodel_2) ) #των  καταλοίπων p-value>0.05=> HO ΔΕΝ απορρίπτεται

lillie.test( rstudent(logmodel_2) )#των καταλοίπων  p-value >0.05=> HO ΔΕΝ  απορρίπτεται

# οκ με κανονικότητα


#2)

library(car)

ncvTest(logmodel_2)    # για να δούμε ομοσλεδατσικότητα , DEN απορρίπτω την H0 , η διακυμανση των καταλοίπων θεωρειται σταθερη

#pvalue<a--> Ho απορρίπτεται

yhatlog.quantiles<-cut(yhatlog, breaks=quantile(yhatlog, probs=seq(0,1,0.25)), dig.lab=6)  # μποεςι; να σκαπσεις τα καταλοιπα να δεις πως πάει η διακύμναση σε κάθε γκρούπ

table(yhatlog.quantiles)

leveneTest(rstudent(logmodel_2)~yhatlog.quantiles)  #p.value= 0.22 >α --> Ho  ΔΕΝ απορ ,  Μπορω να θεωρησω ίσες διακυμανσεις

par(mfrow=c(1,1))

boxplot(rstudent(logmodel_2)~yhatlog.quantiles)      # παρατηρω η διακυμναση τους πολύ καλύτερα


#4)linearity check

library(car)

residualPlot(logmodel_2, type='rstudent') #θες αυτα να είναι ευθέια και να απλβωνονται σωστα γυρω απο την ευθέια , πολυ καλύτερα .

residualPlots(logmodel_2, plot=F, type = "rstudent")       # απλως του λες μην μου εμφανίσεις το Plot, ολα οκ
```

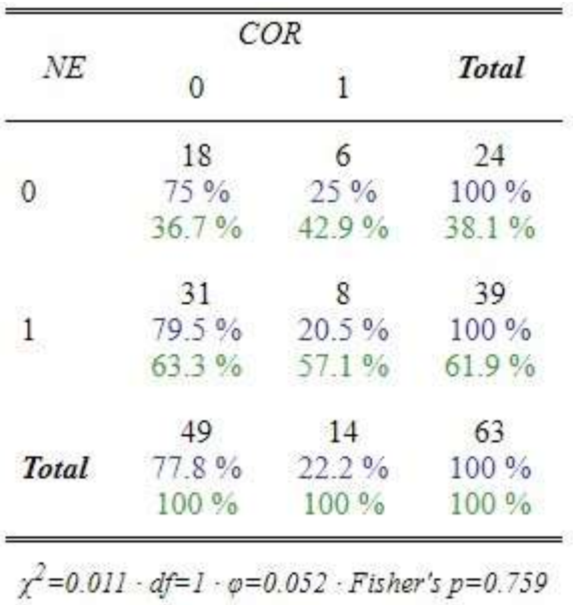
```
# ola ok πλεον
# πάμε και έναν έλεγχογια outliers ( εδω δεν χρειάζεται καθως με τον λογάριθμο διορθώθηκε )
#▪ Apply the Bonferroni test for the largest
#absolute residual value:
outlierTest(logmodel_2)
#H0: There are no outliers in the data. --> all data points (or residuals) fit well to the model, and any large
#residuals are attributed to random variation rather than being οντως outlier
#bonferonii p value=0.17<a => H0 δεν εχω ενδείξεις να την απορρίψω



#▪ Leverage points:
#A leverage point in the context of statistical analysis,
#particularly in regression models, is a data point that exerts a strong influence on the parameters of the model
#A leverage point can significantly affect the slope and position of the regression line.
#Are often(not always) extreme or outlier values .
#Some times distort the results of a regression analysis.
#not all leverage points negatively impact a model.
#Some are valid observations that are essential for understanding the underlying relationship in the data.
#So i will not always put them out of my model , only if they create a big problem in the regression analysis
lev <- hatvalues(logmodel_2)
plot(lev, ylim=range(c(0,3.5*2/10, lev)), main="as2" )
abline( h=c(2,3)*2/10, col=2:3, lty=2:3 )                        # all ok

#▪ Influence measures via Cook’s distance:
# identifying outliers or influential observations that could skew the results of the regression analysis.
#Cook's distance measures the effect of deleting a single observation on the fitted regression model.
#It assesses the influence of each data point by looking at
#how much the predicted values for the dependent variable change when the model is re-estimated without that point.
par(mfrow=c(1,1))
cooks1<- cooks.distance(logmodel_2)
critical.value <- 4/(nrow(as2_c.num)- length(logmodel_2$coef))
```

```
plot(cooks1, ylab="Cook s distance", ylim=range(c(0,cooks1, critical.value )) )

abline(h=critical.value,col=2,lty=2)
```

```
# (IX) Conduct LASSO as a variable selection technique and compare the variables that you end up

#	having using LASSO to the variables that you ended up having using stepwise methods in (VI). Are you getting the same results? Comment.


model_full <- lm(PRICE ~., data = as2)


par(mfrow=c(1,1))

require(glmnet)

X <- model.matrix(model_full)[,-1]        # διώχνω την σταθερα αρχικά , θελω να κρατησω μονο τις μεταβλητες , αυτες που πέφτουν πιο γρηγορα στο μηδεν ειναι αυτες που πεταω πιο γρηγορα αν δεν φαινεται καλα θα δω τα coefficients

lasso <- glmnet(X, as2$PRICE)

plot(lasso, xvar = "lambda", label = T)


# i see that the variables that first doesnt want to include in the model is

# SQFT , AGE , NE ,COR , FEATS

#PRICE ~ SQFT + FEATS  with step method

#Use cross validation to find a reasonable value for lambda

lasso1 <- cv.glmnet(X,as2$PRICE , alpha = 1)                    #Πως επιλέγω το πεναλτι , καλα αυτο γινεται αυτοματα  , το δεφαθλτ μαλλον ειναι 10

lasso1$lambda

lasso1$lambda.min

lasso1$lambda.1se   # οποτε διαλέγουμε αυτο γιατι πεταει πιο γρηγορα , αν δεν γινεται το αλλο

plot(lasso1)

coef(lasso1, s = "lambda.min")  # π.χ. εδω πεναλτι το μινιμομ λάμδα , με βαση αυτο οι μεταβλητλες που θα φύγουν ειναι αυτες που εχουν την τελίτσα

coef(lasso1, s = "lambda.1se")
```

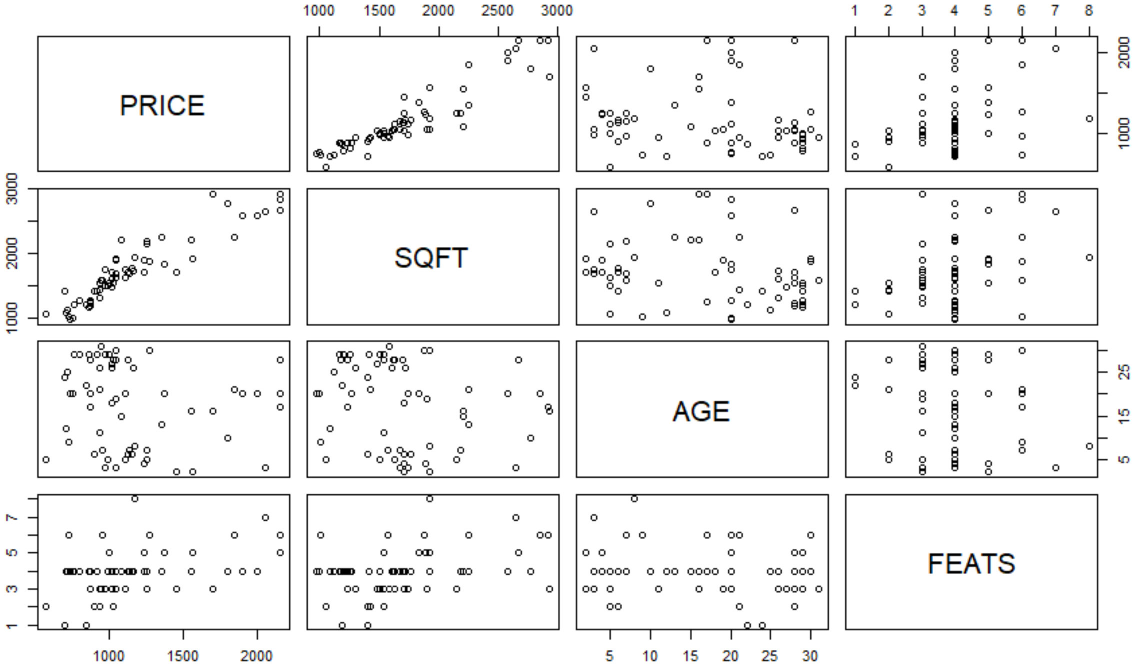
```
plot(lasso1$glmnet.fit, xvar = "lambda")   #επιλέγω το λαμδα οπου με ταδε λαμδα βλεπεις ποσες  μεταβλητές εχουν μεινει

abline(v=log(c(lasso1$lambda.min, lasso1$lambda.1se)), lty =2)
```

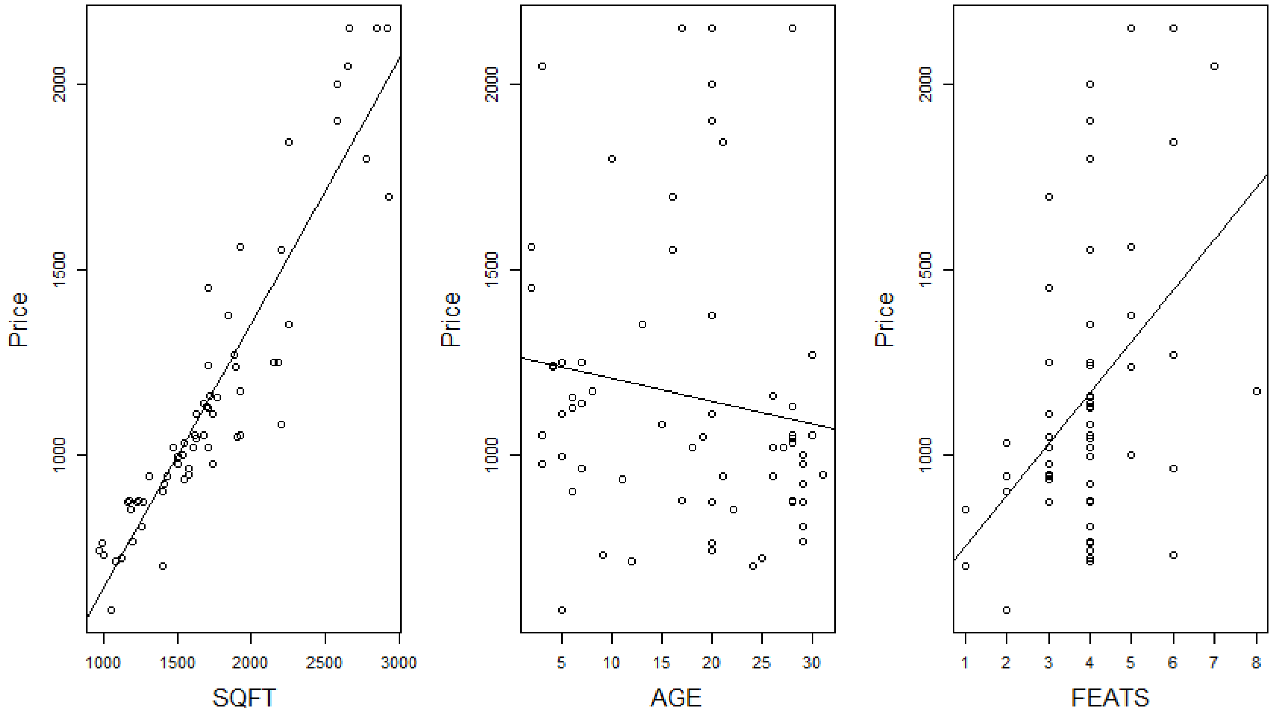
Plots

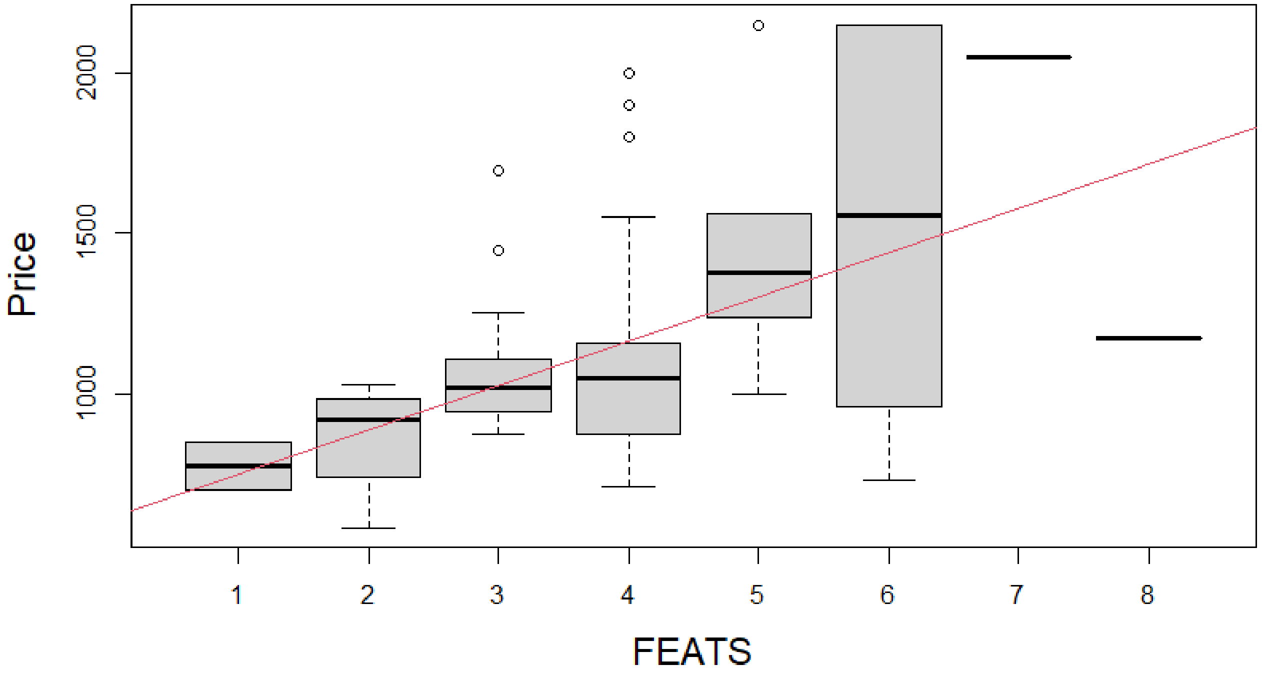


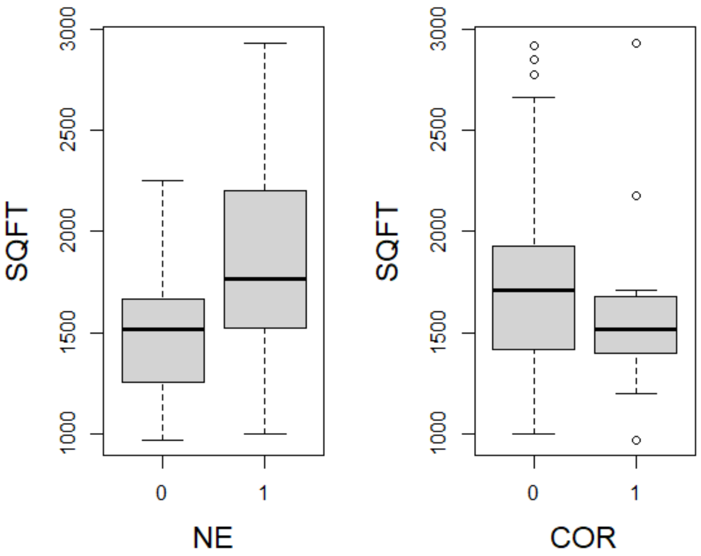
NE	*COR*		***Total***
	0	1	
0	18 75 % 36.7 %	6 25 % 42.9 %	24 100 % 38.1 %
1	31 79.5 % 63.3 %	8 20.5 % 57.1 %	39 100 % 61.9 %
Total	49 77.8 % 100 %	14 22.2 % 100 %	63 100 % 100 %

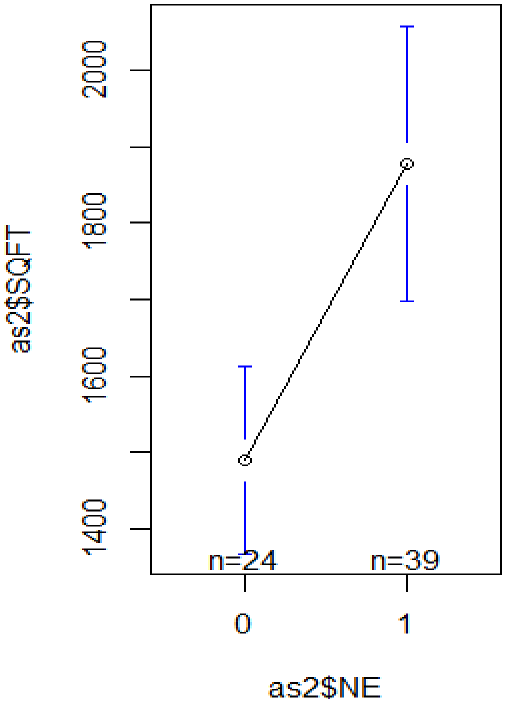
χ^2=0.011 · df=1 · φ=0.052 · Fisher's p=0.759

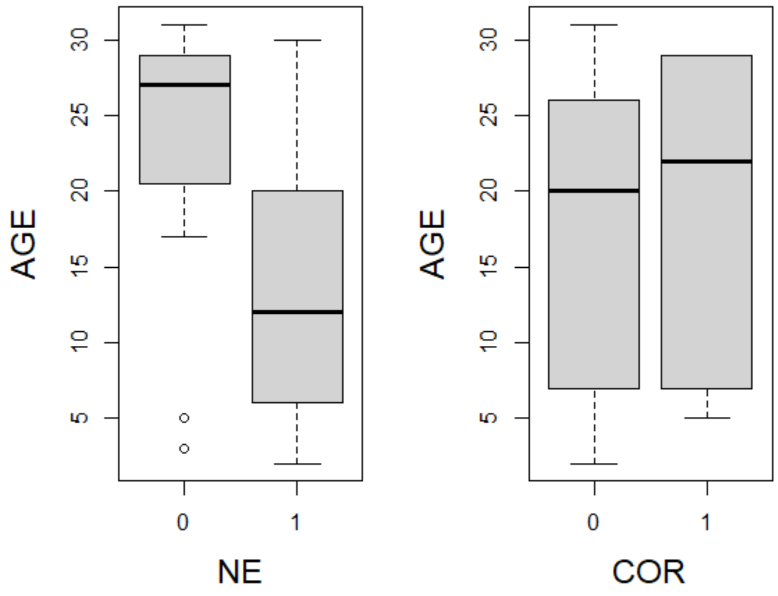


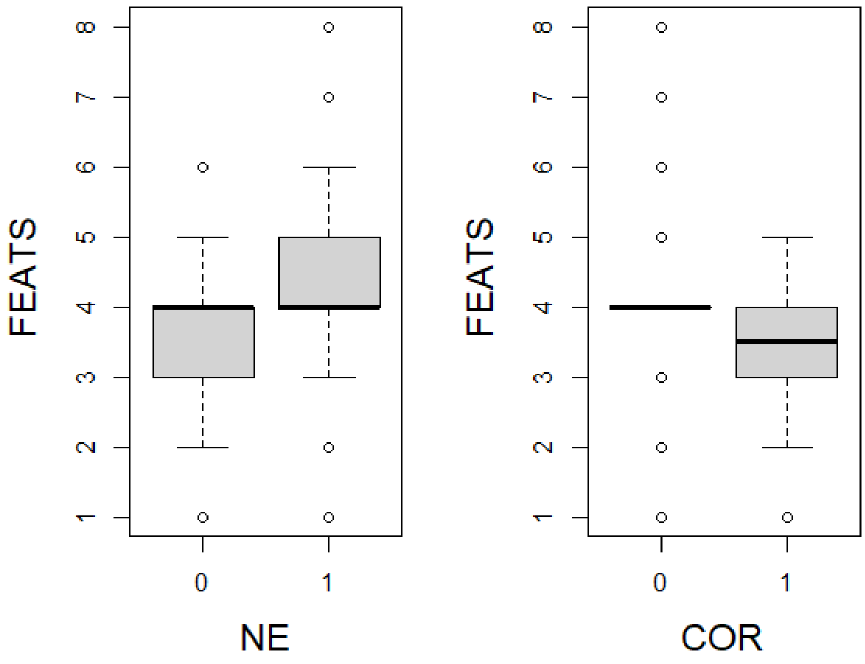

Price
1000
1500
2000
1000
1500
2000
2500
3000
SQFT
Price
1000
1500
2000
5
10
15
20
25
30
AGE
Price
1000
1500
2000
1
2
3
4
5
6
7
8
FEATS

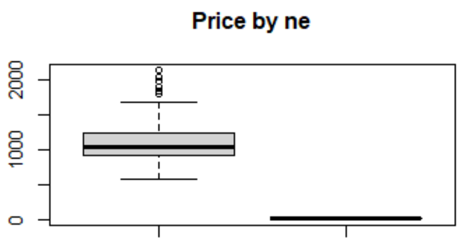

Price
1000
1500
2000
1
2
3
4
5
6
7
8
FEATS

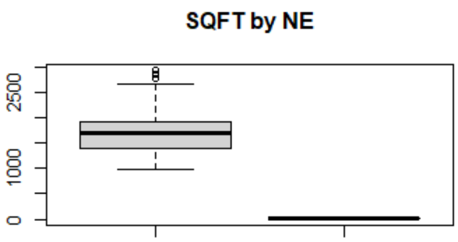

SQFT
1000
1500
2000
2500
3000
0
1
NE
SQFT
1000
1500
2000
2500
3000
0
1
COR


as2$SQFT
1400
1600
1800
2000
n=24
n=39
0
1
as2$NE

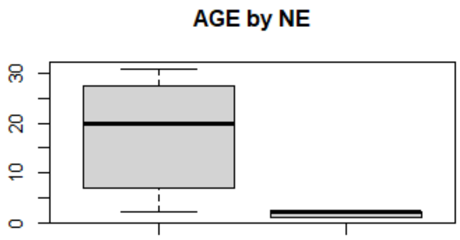

AGE
30
25
20
15
10
5
0
1
NE
AGE
30
25
20
15
10
5
0
1
COR

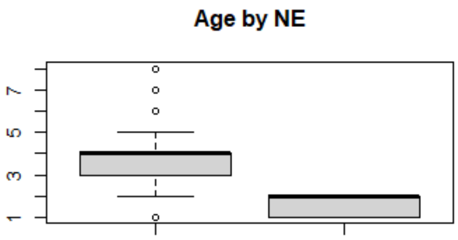

FEATS
8
7
6
5
4
3
2
1
0
1
NE
FEATS
8
7
6
5
4
3
2
1
0
1
COR

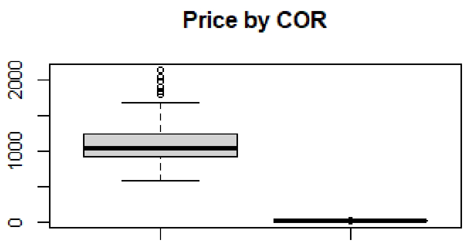

Price by ne

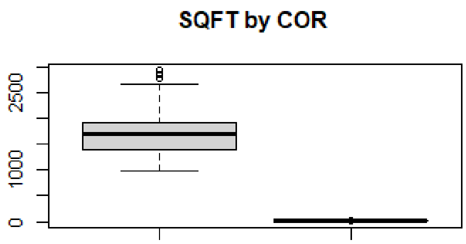

SQFT by NE

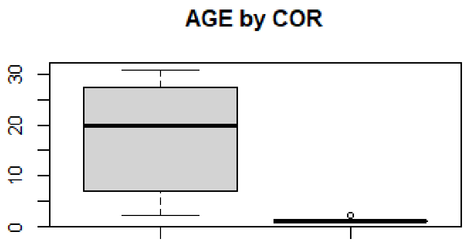
NE

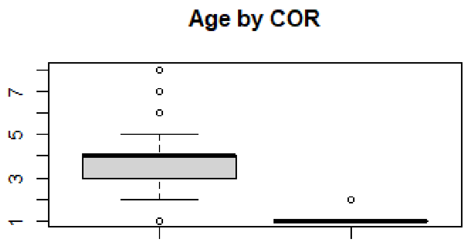

AGE by NE

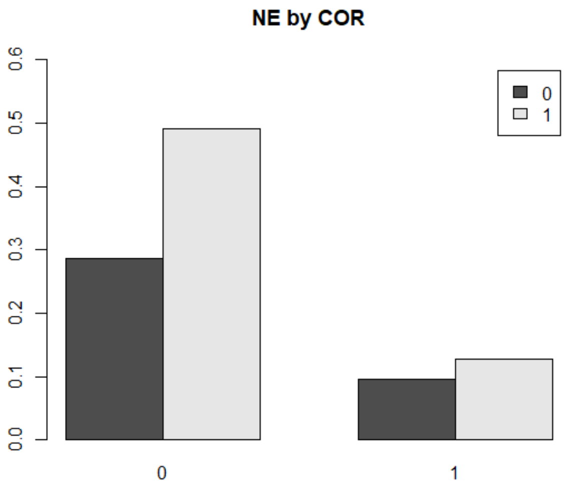

Age by NE

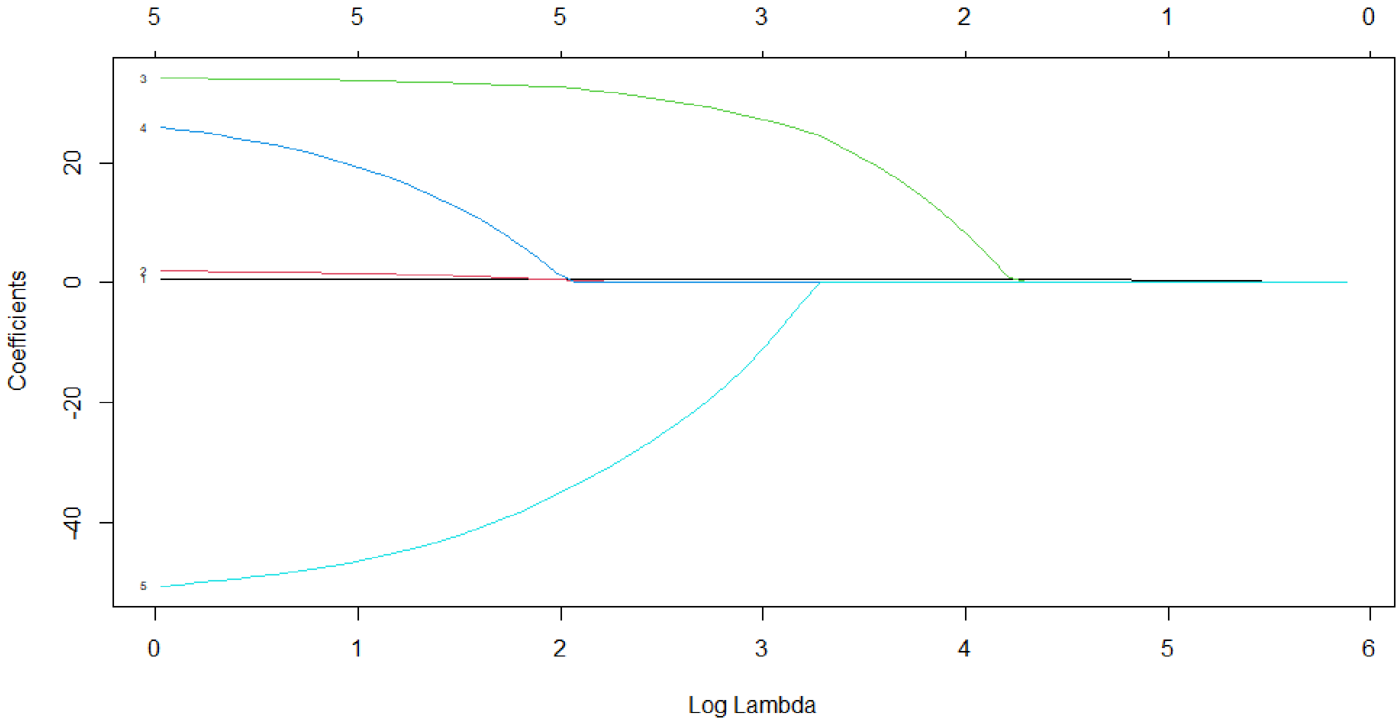

Price by COR

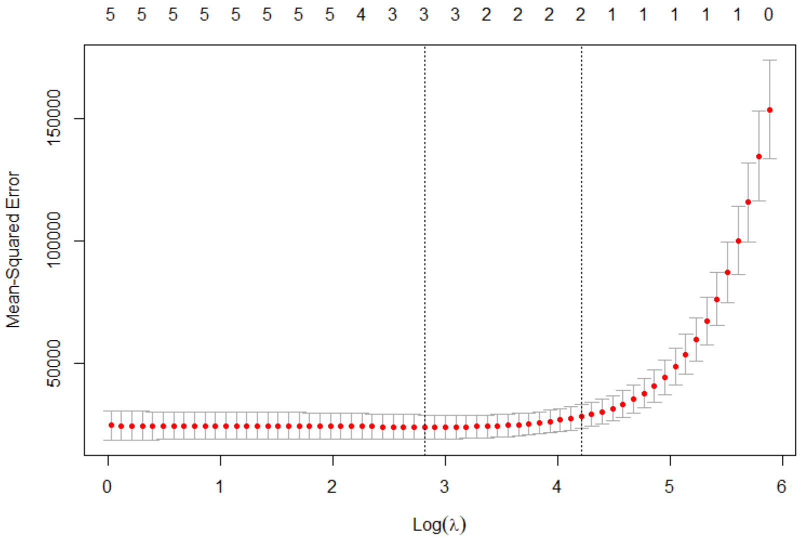

SQFT by COR


AGE by COR


Age by COR


NE by COR
0.0
0.1
0.2
0.3
0.4
0.5
0.6
■ 0
□ 1
0
1


5
5
5
3
2
1
0
3
4
2
1
5
20
0
-20
-40
Coefficients
0
1
2
3
4
5
6
Log Lambda


5 5 5 5 5 5 5 5 4 3 3 3 2 2 2 2 1 1 1 1 1 0
Mean-Squared Error
50000
100000
150000
0
1
2
3
4
5
6
Log(λ)