MSc in
Business Analytics

# STATISTICS FOR BUSINESS ANALYTICS I

## LAB ASSIGNMENT II

Georgia Vlassi – p2822001

Professors: Mr. Ntzoufras, Mr. Leriou

MSc in Business Analytics | Part Time 2020

Department of Management Science & Technology

Athens, Greece
December 23rd, 2020

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.
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).

## I. Read the 'usdata' dataset and use str() to understand its structure.

To import and read the dataset, we set the working directory and use read.table() function:

| | PRICE | SQFT | AGE | FEATS | NE | COR |
|---|---|---|---|---|---|---|
| 1 | 2050 | 2650 | 3 | 7 | 1 | 0 |
| 2 | 2150 | 2664 | 28 | 5 | 1 | 0 |
| 3 | 2150 | 2921 | 17 | 6 | 1 | 0 |
| 4 | 1999 | 2580 | 20 | 4 | 1 | 0 |
| 5 | 1900 | 2580 | 20 | 4 | 1 | 0 |
| 6 | 1800 | 2774 | 10 | 4 | 1 | 0 |
| 7 | 1560 | 1920 | 2 | 5 | 1 | 0 |
| 8 | 1449 | 1710 | 2 | 3 | 1 | 0 |
| 9 | 1375 | 1837 | 20 | 5 | 1 | 0 |
| 10 | 1270 | 1880 | 30 | 6 | 1 | 0 |

We use the function str() to display the structure of the data frame:

```
'data.frame':   63 obs. of  6 variables:
 $ PRICE: int  2050 2150 2150 1999 1900 1800 1560 1449 1375 1270 ...
 $ SQFT : int  2650 2664 2921 2580 2580 2774 1920 1710 1837 1880 ...
 $ AGE  : int  3 28 17 20 20 10 2 2 20 30 ...
 $ FEATS: int  7 5 6 4 4 4 5 3 5 6 ...
 $ NE   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ COR  : int  0 0 0 0 0 0 0 0 0 0 ...
```

File 'usdata' contains 63 observations (objects) with 6 variables each. All variables are integers.

If we open the file with an editor, e.g. Notepad++ , we notice that variables NE and COR are strings and not integers. As a result, the variables are not defined correct.

```
1  "PRICE" "SQFT" "AGE" "FEATS" "NE" "COR"
2  "1" 2050 2650 3 7 "1" "0"
3  "2" 2150 2664 28 5 "1" "0"
4  "3" 2150 2921 17 6 "1" "0"
5  "4" 1999 2580 20 4 "1" "0"
6  "5" 1900 2580 20 4 "1" "0"
7  "6" 1800 2774 10 4 "1" "0"
8  "7" 1560 1920 2 5 "1" "0"
9  "8" 1449 1710 2 3 "1" "0"
10 "9" 1375 1837 20 5 "1" "0"
```

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

By the use of as.numeric() function, we convert variables PRICE, SQFT, AGE and FEATS to be numeric. For the rest variables, NE and COR, we use as.factor() function, to convert them to factors with two levels each.

```
'data.frame':   63 obs. of  6 variables:
 $ PRICE: num  2050 2150 2150 1999 1900 ...
 $ SQFT : num  2650 2664 2921 2580 2580 ...
 $ AGE  : num  3 28 17 20 20 10 2 2 20 30 ...
 $ FEATS: num  7 5 6 4 4 4 5 3 5 6 ...
 $ NE   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ COR  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

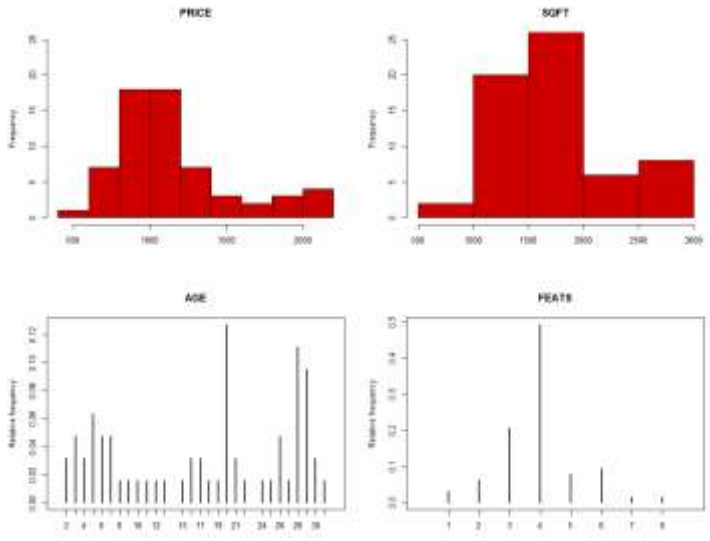
## III. Perform descriptive analysis and visualization for each variable to get an initial insight of what the data looks like. Comment on your findings

In the beginning, we will analyze the numerical variables. We use describe() function, to get the basic Descriptive Statistics of these variables:

| | PRICE | SQFT | AGE | FEATS |
|---|---|---|---|---|
| vars | 1.00 | 2.00 | 3.00 | 4.00 |
| n | 63.00 | 63.00 | 63.00 | 63.00 |
| mean | 1158.41 | 1729.54 | 17.46 | 3.95 |
| sd | 392.71 | 506.70 | 9.60 | 1.28 |
| median | 1049.00 | 1680.00 | 20.00 | 4.00 |
| trimmed | 1105.96 | 1685.18 | 17.75 | 3.92 |
| mad | 262.42 | 392.89 | 11.86 | 1.48 |
| min | 580.00 | 970.00 | 2.00 | 1.00 |
| max | 2150.00 | 2931.00 | 31.00 | 8.00 |
| range | 1570.00 | 1961.00 | 29.00 | 7.00 |
| skew | 1.18 | 0.74 | -0.21 | 0.45 |
| kurtosis | 0.54 | -0.16 | -1.47 | 1.12 |
| se | 49.48 | 63.84 | 1.21 | 0.16 |

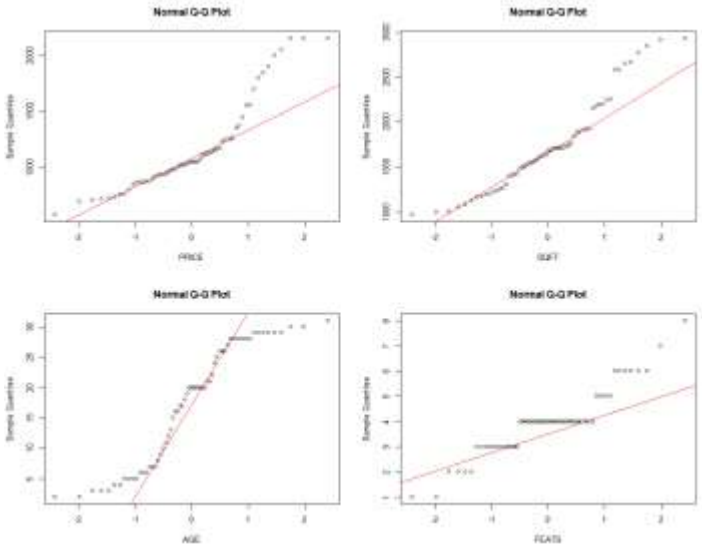
For each numeric variable we have the number of objects, the mean, the standard deviation (sd), the median, the adjusted mean(trimmed), the mean absolute deviation (mad), the min value, the max value, the range, the skewness, the kurtosis and the standard error.

Histograms



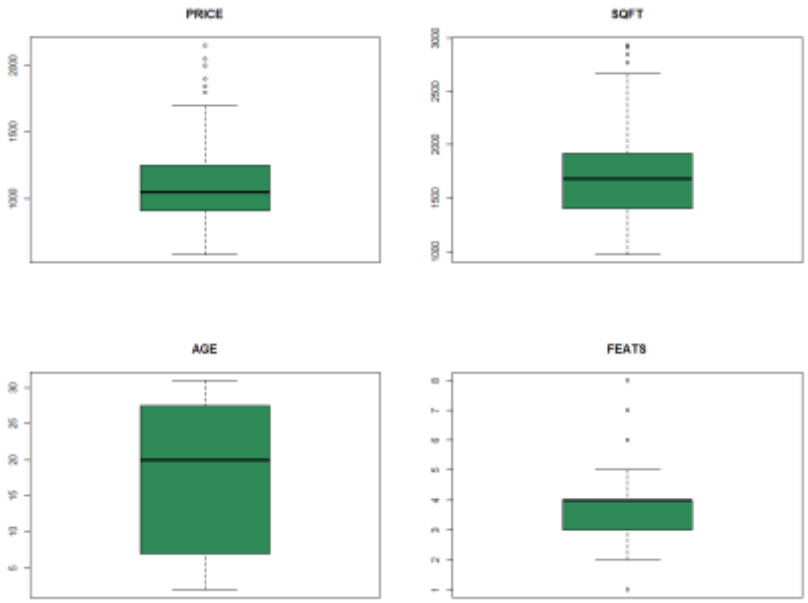
None of the above histograms is bell shaped. PRICE and SQFT are right-skewed, which indicates that there are several data points, perhaps **outliers,** that are greater than the mode. AGE is undefined bimodal, which means that there are intervals equally representing the maximum frequency of the distribution. We use relative frequency for AGE and FEATS, as they have discrete values.

## QQ Plots



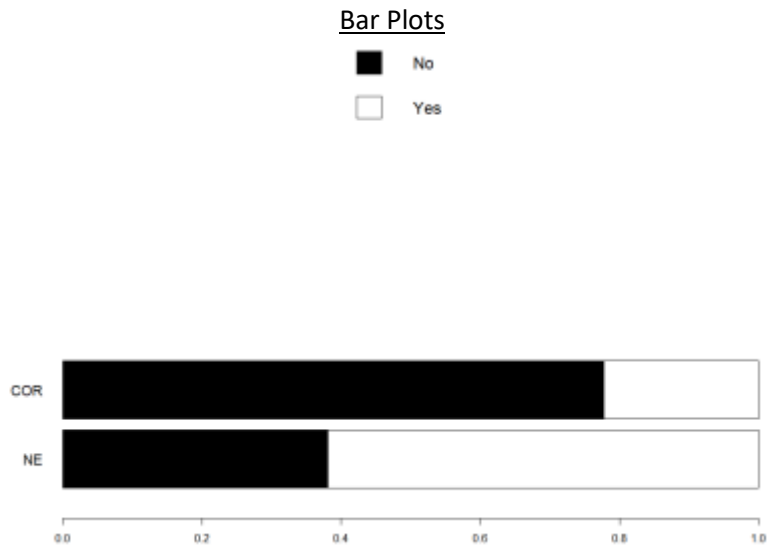

We use QQ plot, which is more effective way to view the distribution of a variable. QQ plot can identify from how the values in some section of the plot differ locally from an overall linear trend by seeing whether the values are more or less concentrated than the theoretical distribution would suppose in that section of a plot.

## Box Plots



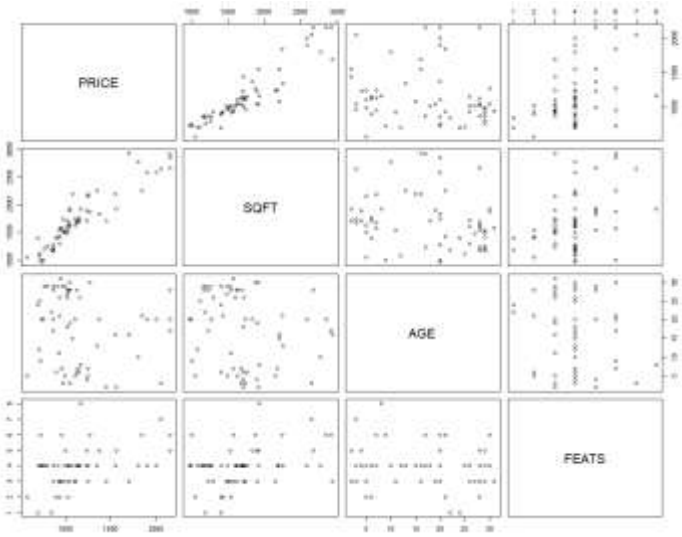
Finally, by using Boxplot, we conclude that the mean measure is greater than the middle value for all numeric variables, apart from PRICE, where it is lower. In FEATS the mean is in the third quartile.

Below, there is a visualization for the two factor variables COR and NE:



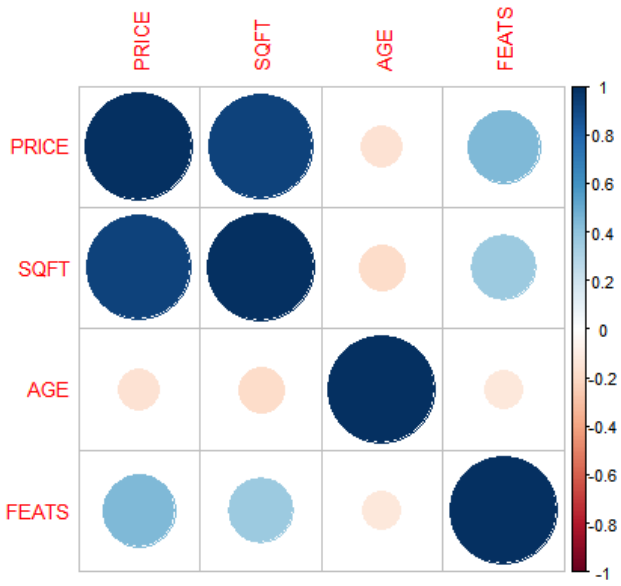

**IV. Conduct pairwise comparisons between the variables in the dataset to investigate if there are any associations implied by the dataset.(Hint: Plot variables against one another and use correlation plots and measures for the numerical variables.). Comment on your findings.**
**Is there a linear relationship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

We conduct pairwise comparisons between the numeric variables, by using pairs() function.
From the result below, we conclude that there is association between PRICE and SQFT.



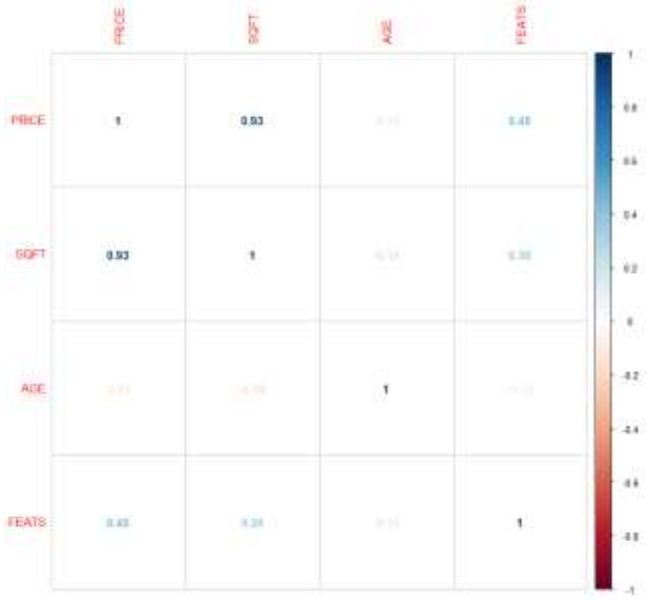
We, also, draw corrplot() to have an image of associations between variables.

Correlation Plot




Positive correlations are displayed in dark blue, like PRICE ~ SQFT and negative correlations in red color. There is zero to negative correlation between PRICE ~ AGE. Color intensity and the size of the circle are proportional to the correlation coefficients. In the right side of the correlogram, the legend color shows the correlation coefficients and the corresponding colors.

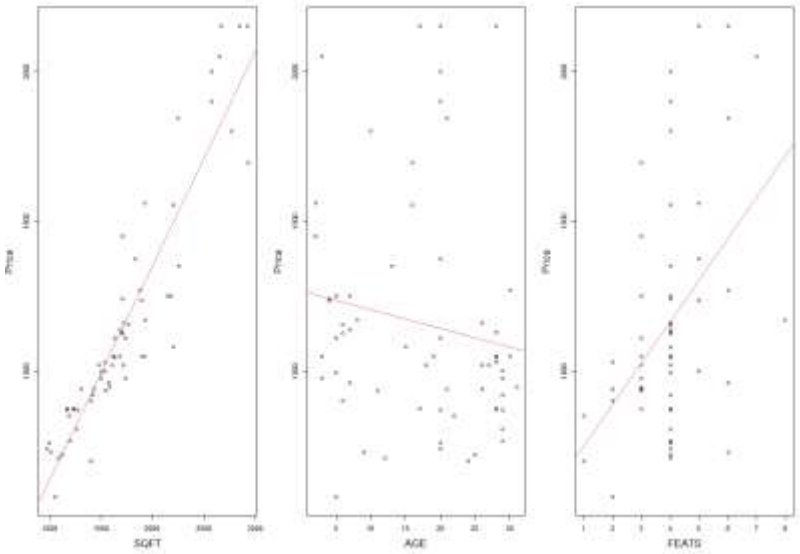
Correlations



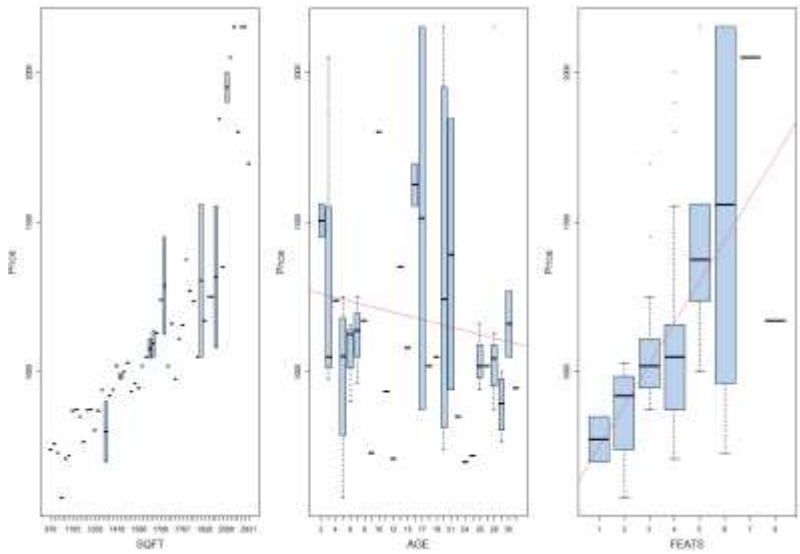
The numbers that are closer to 1, indicate that these two variables are strongly correlated.

In order to find if there is a linear relationship between PRICE and the other variables, we draw the following Scatter Plots and Box Plots, by using lm() function.

Scatter Plots



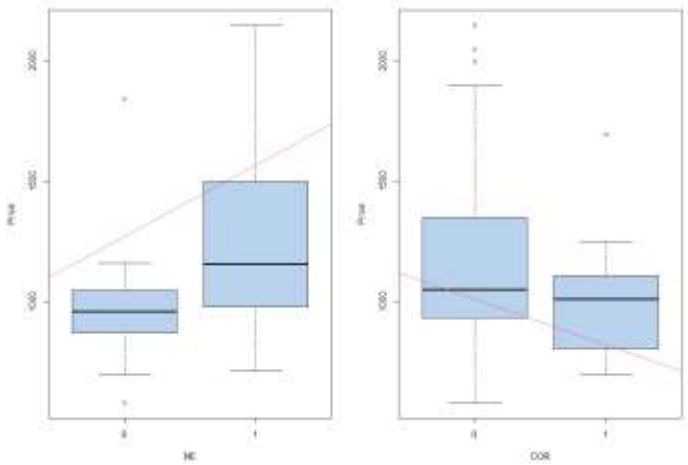

Box Plots




We conclude that there is a negative linear relationship between PRICE ~ AGE.
On the other hand, there is a strongly positive linear relationship between PRICE ~ SQFT and PRICE ~ FEATS.

To find if there is a linear relationship between PRICE and factor variables COR and NE, we also draw the following box plot:

Box Plots



From the above graph, we conclude that there is a negative linear relationship between PRICE ~ COR and a positive linear relationship between PRICE ~ 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 ).**

The model below is achieved by using the lm() function and the output is called using the summary() function on the model.

**Full model**

```
Call:
lm(formula = PRICE ~ ., data = usdata)

Coefficients:
(Intercept)         SQFT          AGE        FEATS          NE1         COR1
  -193.3493       0.6766       2.2291      34.3657      30.0045     -53.0794
```

**Summary of Full model**

```
Call:
lm(formula = PRICE ~ ., data = usdata)

Residuals:
    Min      1Q  Median      3Q     Max
-416.11  -71.03  -15.26   83.02  347.77

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -193.34926   94.52382  -2.046   0.0454 *
SQFT           0.67662    0.04098  16.509   <2e-16 ***
AGE            2.22907    2.28626   0.975   0.3337
FEATS         34.36573   16.27114   2.112   0.0391 *
NE1           30.00446   47.93940   0.626   0.5339
COR1         -53.07940   46.15653  -1.150   0.2550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 144.8 on 57 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.864
F-statistic: 79.76 on 5 and 57 DF,  p-value: < 2.2e-16
```

The most important statistics from linear regression are Pr(>t) and Adjusted R-squared.
The Pr relates to the probability of observing any value equal or larger than t. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope

indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between PRICE and SQFT, as well as between PRICE and FEATS.
The R-squared (R2) statistic provides a measure of how well the model is fitting the actual data. It takes the form of a proportion of variance.
The adjusted R2 is a measure of the linear relationship between our predictor variable (PRICE) and our response / target variables (SQFT, AGE, FEATS, NE, COR). It takes values between 0 and 1, where a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1(0,86 in our example) does explain the observed variance in the response variable.
We can conclude that the 86% of the variability is explained by this model.

**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).**

Our purpose is to find the best model for predicting the selling prices. We will rely on the AIC, which provides a means for model selection.
We will use the Stepwise procedure according to AIC, where in each step we choose which variable to include, by adding or removing covariates. In the beginning, we are going to construct the full model, by using *lm(PRICE ~., data = usdata),* including all covariates of the dataset. Our scope is to select the covariates with the minimum AIC, minimum BIC, or minimum p-value.

**Stepwise procedure**

```
Start:  AIC=632.62
PRICE ~ SQFT + AGE + FEATS + NE + COR

       Df Sum of Sq     RSS    AIC
- NE    1      8218 1203977 631.05
- AGE   1     19942 1215701 631.66
- COR   1     27743 1223502 632.07
<none>              1195759 632.62
- FEATS 1     93580 1289339 635.37
- SQFT  1   5717835 6913594 741.17

Step:  AIC=631.05
PRICE ~ SQFT + AGE + FEATS + COR

       Df Sum of Sq     RSS    AIC
- AGE   1     12171 1216147 629.69
- COR   1     23099 1227076 630.35
<none>              1203977 631.05
+ NE    1      8218 1195759 632.62
- FEATS 1    106953 1310930 634.42
- SQFT  1   6288869 7492846 744.24

Step:  AIC=629.69
PRICE ~ SQFT + FEATS + COR

       Df Sum of Sq     RSS    AIC
- COR   1     22454 1238602 628.84
<none>              1216147 629.69
+ AGE   1     12171 1203977 631.05
+ NE    1       447 1215701 631.66
- FEATS 1    104259 1320407 632.87
- SQFT  1   6352036 7568184 742.87

Step:  AIC=628.84
PRICE ~ SQFT + FEATS

       Df Sum of Sq     RSS    AIC
<none>              1238602 628.84
+ COR   1     22454 1216147 629.69
+ AGE   1      9526 1229076 630.35
+ NE    1       218 1238384 630.83
- FEATS 1    138761 1377363 633.53
- SQFT  1   6389899 7628501 741.37

Call:
lm(formula = PRICE ~ SQFT + FEATS, data = data)

Coefficients:
(Intercept)         SQFT        FEATS
   -175.9278       0.8803      39.8389
```

From the results above, we conclude that we have minimum AIC (= 628.84) when we predict the selling prices with formula PRICE ~ SQFT + FEATS. So, SQFT and FEATS are more significant to predict the PRICE.

**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 of Final model**

```
Call:
lm(formula = PRICE ~ SQFT + FEATS, data = usdata)

Residuals:
    Min      1Q  Median      3Q     Max
-400.44  -71.70  -11.21   93.12  341.82

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -175.92760   74.34207  -2.366   0.0212 *
SQFT           0.68046    0.03868  17.594   <2e-16 ***
FEATS         39.83687   15.36531   2.593   0.0119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 143.7 on 60 degrees of freedom
Multiple R-squared:  0.8705,    Adjusted R-squared:  0.8661
F-statistic: 201.6 on 2 and 60 DF,  p-value: < 2.2e-16
```

Observing the coefficients of our final model, we can conclude:

- The coefficient Estimate contains three rows. The first one is the intercept. The intercept, in our example, has negative value, which means that the expected value PRICE will be less than 0 when SQFT and FEATS are set to 0.
- The 86,61% of the variability is explained by this model, as indicates the Adjusted R-squared. We take into concern the Adjusted R-squared , as we have one predictor variable and two target variables.
- The Pr relates to the probability of observing any value equal or larger than t. A small p-value for the intercept and the slope indicates that we can reject the null hypothesis, that the β0, β1, β2 and so on can be zero, which allows us to conclude that there is a relationship between PRICE ~ SQFT and FEATS.
- The RSE (ε) estimate gives a measure of error of prediction. The lower the RSE, the more accurate the model.

The mathematical formulation of the model is:

PRICES = -175.926027 + 0.68046*SQFT + 39.83687*FEATS + ε where ε ~ N(0, 143.7)

Below, we can see the summary of our model without the intercept.

**Final model without Intercept**

```
Call:
lm(formula = PRICE ~ SQFT + FEATS - 1, data = usdata)

Residuals:
    Min      1Q  Median      3Q     Max
-384.11  -80.82  -31.34   49.69  373.64

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
SQFT   0.62538    0.03203  19.524   <2e-16 ***
FEATS 22.06792   13.90199   1.587    0.118
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 149 on 61 degrees of freedom
Multiple R-squared:  0.9856,    Adjusted R-squared:  0.9851
F-statistic:  2089 on 2 and 61 DF,  p-value: < 2.2e-16
```

We observe that when we exclude Intercept from our model, the updated statistic Adjusted R-squared has been increased from 0,8661 to 0,9851. Although, this value is closer to 1, which means that 98% of the variability is explained by this model, this value is iconic. The real Adjusted R-squared value of our final model without Intercept is 0.858371, which means that it has been decreased. Respectively we reach the outcome that we should not exclude Intercept.

**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?**

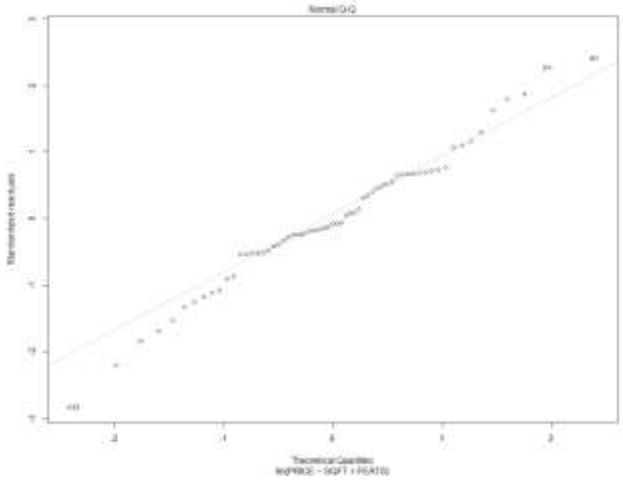
For our final model, we should check the below assumptions:

I. Normality of residuals. The residual errors are assumed to be normally distributed.
II. Homogeneity of residuals variance. The residuals are assumed to have a constant variance (homoscedasticity)
III. Linearity of the data. The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
IV. Independence of residuals error terms.

In the beginning, we have to implement Shapiro-Wilcoxon test to test normality of residuals:

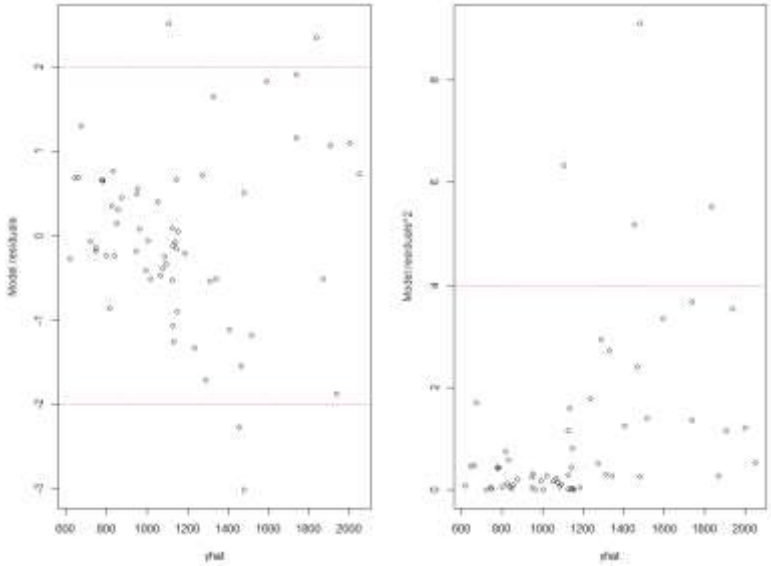
```
        Shapiro-Wilk normality test

data:  Model.residuals
W = 0.98145, p-value = 0.4591
```

As the p-value is greater than 0,05, which is significant evidence of Null Hypothesis, we have to accept normality. As our dataset is greater than 50 observations we should check for the normality of errors, by using QQ plot:



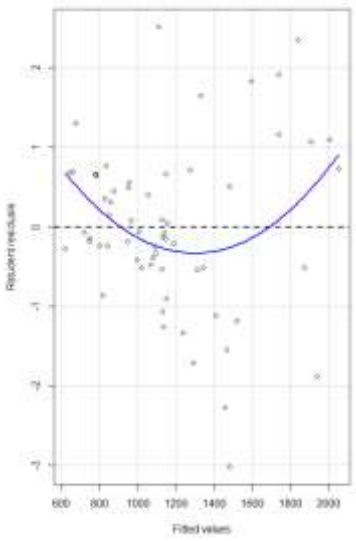
The fitted (or predicted) values are the y-values that we expect for the given x-values according to the built regression model (or visually, the best-fitting straight regression line).From the above graph, we observe that there are a outliers.

Subsequently, we should test the homoscedasticity of residuals:

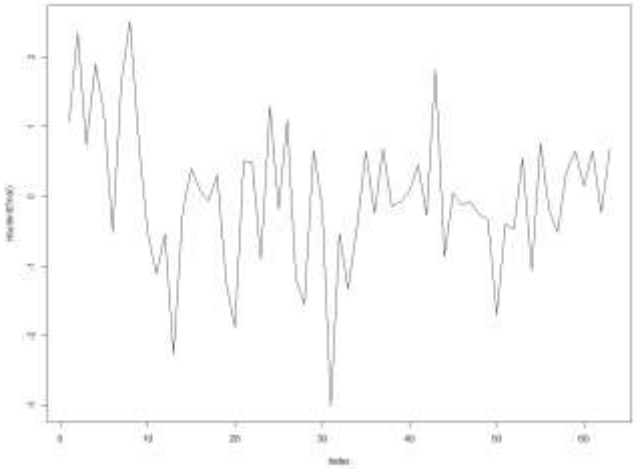


We observe that there is Heteroscedasticity. As a result, the constant Variance assumption violated because there are extreme values outside the red lines.

Moreover, we check for Linearity of our data:




The linearity is also violated as we can see at the outcome - predictor relationships.



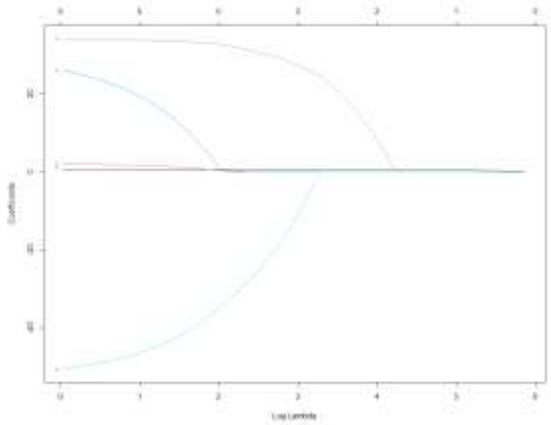
```
        Durbin-Watson test

data:  final
DW = 1.5734, p-value = 0.03571
alternative hypothesis: true autocorrelation is greater than 0
```

Finally, for Independence of residuals error terms, we draw the relative graph and we also run a Durbin-Watson test. The Hypotheses for the Durbin Watson test are: H0 = no first order autocorrelation. H1 = first order correlation exists. From the results, we conclude that we have serial correlation.

There are many ways to solve the violation regarding Normality, Heteroscedasticity and Non-Linearity. Indicative, we could run again all of the above graphs and tests with logarithms. If the issue is still present, we could use the polynomials of the predictors.

**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.**

In the beginning, we implement Lasso with the coefficients against the log-lambda value and with each curve labeled.



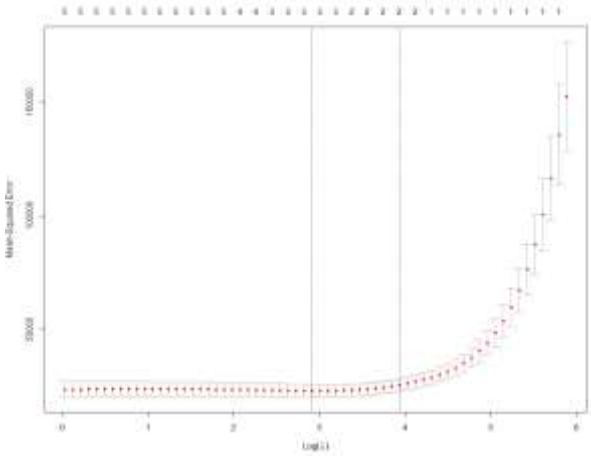
We use cross validation to find a reasonable value for lambda. To show explicitly the selected optimal values of λ,

- the λ at which the minimal MSE is achieved:
  lambda.min = 12.65537

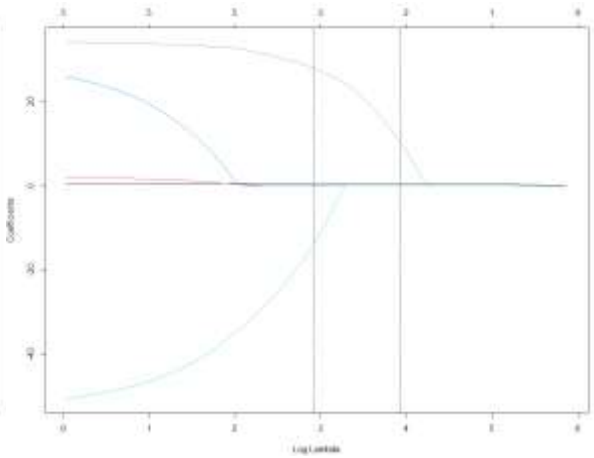
- The most regularized model whose mean squared error is within one standard error of the minimal:
  lambda.1se = 51.08993

As we calculated the outcomes of lambda.min and lambda.1se, we construct the below graphs of MSE against log-lambda value and coefficients against the log-lambda value, respectively.

**MSE ~ Log(lambda)**



**Coefficients ~ Log(lambda)**



**Coefficients with lambda.min**

```
6 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -99.7933996
SQFT          0.6612236
AGE           .
FEATS        30.3653110
NE1           .
COR1        -24.3970328
```

**Coefficients with lambda.1se**

```
6 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 47.9819799
SQFT         0.6126077
AGE          .
FEATS       12.8786649
NE1          .
COR1         .
```

From Lasso we choose the coefficients produced from lambda.1se, as it has fewer variables while it does not differ significantly from the model derived from lambda.min in terms of Mean Square Error.

By the use of Lasso, we observe that we end up having the same variables as using Stepwise method, with different values at coefficients.

MSc in
Business Analytics

# APPENDIX: R source code

```
Question 1
#Import data
setwd('C:\\Users\\tzina\\OneDrive\\Documents\\LabAssignment')
usdata <- read.csv("usdata", header = TRUE, sep = "")

#Understanding the structure
str(usdata)
View(usdata)

Question 2
#Convert the variables PRICE, SQFT, AGE, FEATS to be numeric variables and
NE, COR to be factors
usdata$PRICE <- as.numeric(usdata$PRICE)
usdata$SQFT  <- as.numeric(usdata$SQFT)
usdata$AGE   <- as.numeric(usdata$AGE)
usdata$FEATS <- as.numeric(usdata$FEATS)
usdata$NE    <- as.factor(usdata$NE)
usdata$COR   <- as.factor(usdata$COR)

Question 3
#Perform descriptive analysis and visualization for each variable to get
an initial insight of what the data looks like. Comment on your findings
summary(usdata)
#Descriptive analysis and visualization for numerical variables
require(psych)
index <- sapply(usdata, class) == "numeric"
usdata_num <- usdata[,index]
round(t(describe(usdata_num)),2)

install.packages("ggplot2")
install.packages("cowplot")
library(ggplot2)
library(cowplot)

#Histograms
par(mfrow=c(2,2))
n <- nrow(usdata_num)
hist(usdata_num$PRICE,ylim=c(0,25),main="PRICE",xlab="",ylab="Frequency",c
ol="red3",cex.axis=1.0,cex.lab=1.0)
hist(usdata_num$SQFT,ylim=c(0,25),main="SQFT",xlab="",ylab="Frequency",col
="red3",cex.axis=1.0,cex.lab=1.0)
plot(table(usdata_num[,3])/n, type='h', xlim=range(usdata_num[,3])+c(-
1,1), main=names(usdata_num)[3], ylab='Relative frequency')
plot(table(usdata_num[,4])/n, type='h', xlim=range(usdata_num[,4])+c(-
1,1), main=names(usdata_num)[4], ylab='Relative frequency')

#QQ Plots
library(psych)
par(mfrow= c(2,2))
qqnorm(usdata_num$PRICE, xlab=deparse(substitute(PRICE))) +
qqline(usdata_num$PRICE, col = 'red')
```

```
qqnorm(usdata_num$SQFT, xlab=deparse(substitute(SQFT)))    +
qqline(usdata_num$SQFT, col = 'red')
qqnorm(usdata_num$AGE, xlab=deparse(substitute(AGE)))      +
qqline(usdata_num$AGE, col = 'red')
qqnorm(usdata_num$FEATS, xlab=deparse(substitute(FEATS))) +
qqline(usdata_num$FEATS, col = 'red')

#Box Plots
par(mfrow= c(2,2))
boxplot(usdata_num$PRICE,data=usdata_num, main="PRICE", col= "seagreen")
boxplot(usdata_num$SQFT,data=usdata_num, main="SQFT", col= "seagreen")
boxplot(usdata_num$AGE,data=usdata_num, main="AGE", col= "seagreen")
boxplot(usdata_num$FEATS,data=usdata_num, main="FEATS", col= "seagreen")

#Visual Analysis for factors
usdata_fac <- usdata[,!index]
n <- nrow(data_num)
par(mfrow=c(1,1))
barplot(sapply(usdata_fac,table)/n, horiz=T, las=1, col=1:0, ylim=c(0,8),
cex.names=1.3)
legend('top', fil=1:0, legend=c('No','Yes'), ncol=1, bty='n',cex=1.5)

Question 4
#Pairs of numerical variables
pairs(usdata_num)
require(corrplot)

#Use corrplot
corrplot(cor(usdata_num))

#Price regarding other variables (PLOTS)
par(mfrow=c(1,3))
for(j in 2:4){
  plot(usdata_num[,j], usdata_num[,1], xlab=names(usdata_num)[j],
ylab='Price',cex.lab=1.5)
  abline(lm(usdata_num[,1]~usdata_num[,j]),col=2 )
}

#Price regarding other variables (BOXPLOTS)
par(mfrow=c(1,3))
for(j in 2:4){
  boxplot(usdata_num[,1]~usdata_num[,j], xlab=names(usdata_num)[j],
ylab='Price',col ="slategray2",cex.lab=1.5)
  abline(lm(usdata_num[,1]~usdata_num[,j]),col=2)
}

#Price regarding other variables (CORRPLOTS)
#Correlations valid on the numerical variables
round(cor(usdata_num), 2)

#Diagonal is always one since every variables is perfectly correlated with
itself!
par(mfrow = c(1,1))
corrplot(cor(usdata_num), method = "number")
```

```
cor(usdata_num$PRICE, usdata_num$SQFT, method = "pearson")
cor(usdata_num$PRICE, usdata_num$AGE, method = "pearson")
cor(usdata_num$PRICE, usdata_num$FEATS, method = "pearson")

#Price regarding factor variables
par(mfrow=c(1,2))
for(j in 1:2){
  boxplot(usdata_num[,1]~usdata_fac[,j], xlab=names(usdata_fac)[j],
ylab='Price',cex.lab=1.0,col="slategray2")
  abline(lm(usdata_num[,1]~usdata_fac[,j]),col=2)
}

Question 5
#Construct the full model
mfull <- lm(PRICE ~., data = usdata)
mfull

#Interpret the coefficients
summary(mfull)

Question 6
#Construct the model

#Stepwise
step(mfull, direction='both')
mfull <- lm(PRICE ~., data = usdata)

mnull <- lm(PRICE~1,data=usdata)
step(mnull, scope=list(lower=mnull,upper=mfull), direction='both')


Question 7
#Final Model
final <- lm(PRICE ~ SQFT + FEATS, data=usdata)
summary(final)$coefficient
summary(final)

#Remove intercept from model
no_intercept <- lm(PRICE ~ SQFT + FEATS-1,data=usdata)
summary(no_intercept)

#Misleading result without intercept
#Calculate Adjusted R^2 for the above model
true.r2 <- 1-sum(no_intercept$res^2)/((n-1)*var(usdata$PRICE))
true.r2

#Summary of the model selected by stepwise
summary(step(mfull, direction='both'))

Question 8
#Checking assumptions

#Step 1:Checking Normality of the residuals
#QQ Plot - Checking for the normality of errors
```

```
plot(final, which = 2) #there are outliers

#Find residuals
Model.residuals <- rstudent(final)

#Shapiro Wilk for residuals sample > 50
shapiro.test(Model.residuals)

#Step 2: Check Constant variance
yhat <- fitted(final)
par(mfrow=c(1,2))
plot(yhat, Model.residuals)
abline(h=c(-2,2), col=2, lty=2)
plot(yhat, Model.residuals^2)
abline(h=4, col=2, lty=2)

#Step 3: Check for non linearity
library(car)
residualPlot(final, type='rstudent')
residualPlots(final, plot=F, type = "rstudent")

#Step 4: Check Independence of errors
plot(rstudent(final), type='l')
library(randtests);
runs.test(final$res)
library(lmtest);
dwtest(final)
library(car);
durbinWatsonTest(final)

#model 1 only with SQFT
par(mfrow=c(1,1))
plot(lm(PRICE~SQFT,data=usdata),2, main='Price')
plot(lm(log(PRICE)~SQFT,data=usdata),2, main='log of price')

logmodel<-lm(log(PRICE)~SQFT,data=usdata)
par(mfrow=c(2,2))
plot( logmodel, 2 )
plot( logmodel, 3 )
residualPlot(logmodel, type='rstudent')
plot(rstudent(logmodel), type='l')

ncvTest(logmodel)
residualPlots(logmodel, plot=F)

# model 1 only with SQFT
par(mfrow=c(1,2))
plot(lm(PRICE~SQFT,data=data),2, main='Price')
plot(lm(log(PRICE)~SQFT,data=data),2, main='log of price')

Question 9
#Create Lasso graph
require(glmnet)
X <- model.matrix(mfull)[,-1]
```

```
lasso <- glmnet(X, usdata$PRICE)
plot(lasso, xvar = "lambda", label = T)

#Use cross validation to find a reasonable value for lambda
lasso1 <- cv.glmnet(X, usdata$PRICE, alpha = 1)
lasso1$lambda
lasso1$lambda.min
lasso1$lambda.1se
plot(lasso1)
coef(lasso1, s = "lambda.min")
coef(lasso1, s = "lambda.1se")
plot(lasso1$glmnet.fit, xvar = "lambda")
abline(v=log(c(lasso1$lambda.min, lasso1$lambda.1se)), lty =2)
```