STATISTICS FOR BA I – ASSIGNMENT 4

Question 1 – Read file 'usdata' and interpret its data

Observing the contents of 'usdata' file it is noticed that the sample is comprised of 63 observations and 6 variables. Specifically, the variables and their corresponding data types included in the sample are the following:

- **PRICE:** Selling price of the house (in hundred $) - Integer
- **SQFT:** Square feet of living **-** Integer
- **AGE:** Age of home in years – Integer
- **FEATS:** Number of features (0-11) – Integer
- **NE:** Index that determines whether the house is located in Northeast Coast – Integer
- **COR:** Corner Location – Integer

```
> str(House_sales)
'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 ...
```

1. Raw data of the sample

Question 2 – Update the data type for each variable

First, data cleaning must be performed to the sample. Initially, all variables are examined to ensure that there are no missing values in the sample. Afterwards, the data types of the variables are updated. The variables PRICE, SQFT, AGE and FEATS are assigned a numeric data type and the variables NE and COR are assigned a factor data type with labels 'No' and 'Yes' for their values '0' and '1' respectively.

```
> sum(is.na(House_sales))
[1] 0
```

2. Check for NAs

```
> str(House_sales)
'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 "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ COR  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

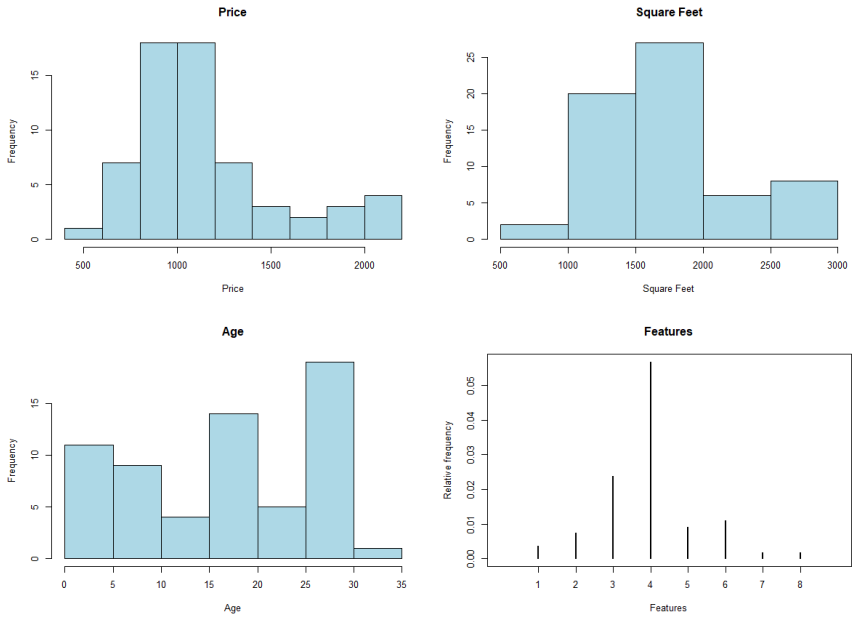
3. Clean data

Question 3 – Perform descriptive analysis for each variable

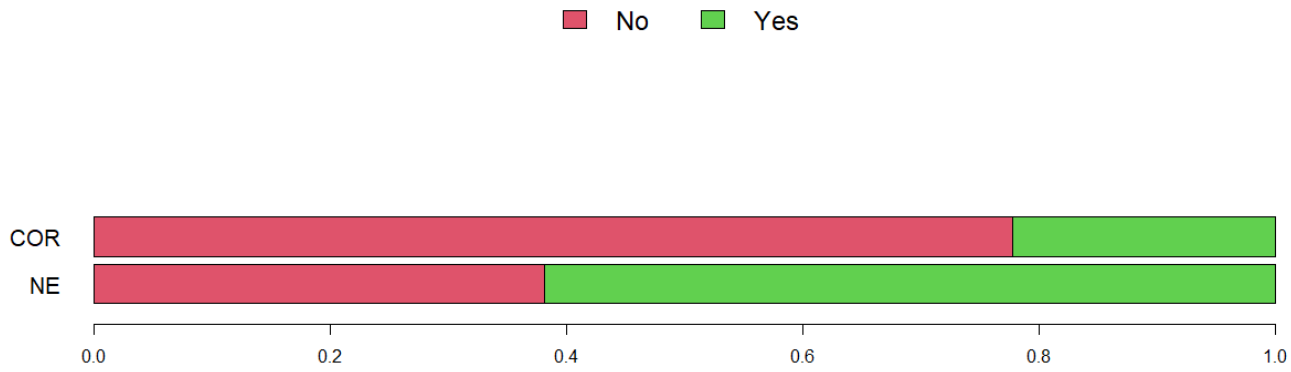
Observing the summarized data, the following deductions can be drawn about the sample's variables. First, it can be observed that most houses are sold with prices close to 110.000$, while the number of houses sold with smaller or higher values is much smaller. The cheapest and the most expensive houses that were recorded in the sample were sold with a price of 58.000$ and 215.000$ respectively. The average square feet area of the houses that are recorded is approximately equal to 1.700 sqft., while only a small portion of the houses has an area smaller than 1000 or greater than 2000 sqft. Both the price and the sqft. variables have big standard deviations. In other words, both variables have a lot of values that are spread out and not close to their corresponding mean. The average age of the houses recorded is approximately equal to 17 years, and most of the houses (mode) included in the sample have an age between 25-30 years old. Regarding the number of features of the houses, it can be deducted that most houses sold include only 4 features, and the second most frequent houses sold included only 3 features. Also, it can be observed that none of the houses have no features, or more features than 8. Regarding the factor variables, first it can be observed that most houses are placed in the northeast cost and are not in corner locations.

```
> summary(House_sales)
     PRICE           SQFT           AGE            FEATS          NE      COR
 Min.   : 580   Min.   : 970   Min.   : 2.00   Min.   :1.000   No :24   No :49
 1st Qu.: 910   1st Qu.:1400   1st Qu.: 7.00   1st Qu.:3.000   Yes:39   Yes:14
 Median :1049   Median :1680   Median :20.00   Median :4.000
 Mean   :1158   Mean   :1730   Mean   :17.46   Mean   :3.952
 3rd Qu.:1250   3rd Qu.:1920   3rd Qu.:27.50   3rd Qu.:4.000
 Max.   :2150   Max.   :2931   Max.   :31.00   Max.   :8.000
> describe(House_num)
      vars  n    mean     sd median trimmed    mad min  max range  skew kurtosis    se
PRICE    1 63 1158.41 392.71   1049 1105.96 262.42 580 2150  1570  1.18     0.54 49.48
SQFT     2 63 1729.54 506.70   1680 1685.18 392.89 970 2931  1961  0.74    -0.16 63.84
AGE      3 63   17.46   9.60     20   17.75  11.86   2   31    29 -0.21    -1.47  1.21
FEATS    4 63    3.95   1.28      4    3.92   1.48   1    8     7  0.45     1.12  0.16
```

4. Summary of clean data



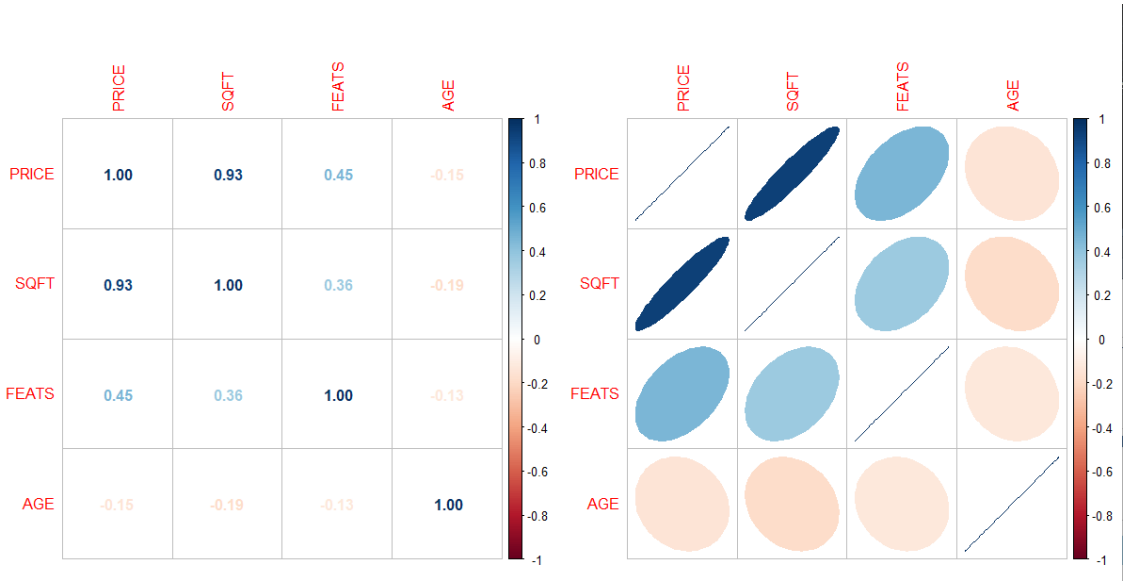
5. Histograms of numeric variables



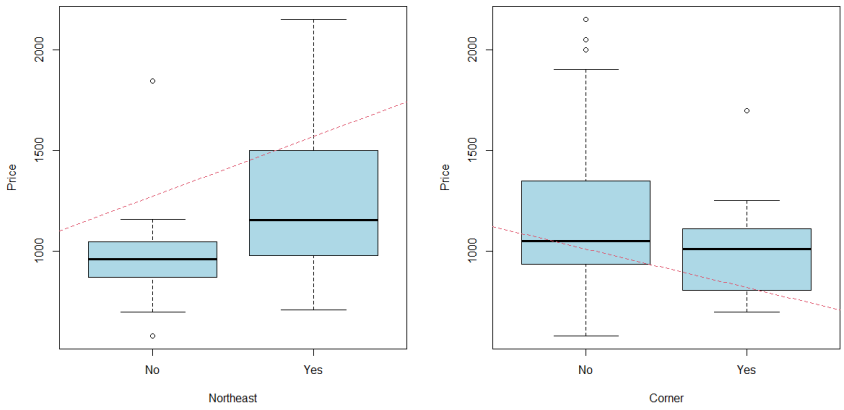
6. Barplots for factor variables

Question 4 – Conduct pairwise comparisons and interpret the results

Initially, it is noticed that there is a strong positive linear relationship between the price and the square feet area of the houses. In other words, while the square feet area of the house is rising, the price tends to increase too. Afterwards, it is noticed that the number of the features has a medium positive linear relationship with the price and the square feet of the houses. Specifically, while the price and the square feet area of the house are rising, the number of features is increasing at a lower rate. Also, it is noticed that while the age of the houses rises, the price, the square feet area and the number of features tend to decrease. In other words, older houses tend to be cheaper and to have smaller square feet areas and less features than the new houses based on the data of the sample. Furthermore, it can be remarked that houses that are placed in the northeast cost tend have higher prices, while houses that are placed in corner locations tend to have lower prices.



7. Pairwise comparisons for numeric variables



8. Boxplots of factor variables vs Price

Question 5 – Construct a linear model predicting the price of the houses – comment on its fitness

To evaluate the fitness of the model we must examine the value of the Adjusted R-Squared value, since it's a multiple linear regression model. In the full model we observe that the Adjusted R-Squared value is equal to 0.864 (or 86.4%), which means that the current explanatory variables are a good fit for the prediction of the final price. In other words, 86.4% of the variability of the houses' price can be interpreted from all the explanatory variables of the sample. Finally, a constant model was created (a model that has no variables) and an anova test was executed. Specifically, the anova test compares the two models, and examines if the full model is predicting the value of the houses better than the constant model. Observing that the p-value of the test is (<0.05) we reject the null hypothesis that the constant model predicts the price of the houses in a better manner.

```
> summary(full_model)

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

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 *
NEYes         30.00446   47.93940   0.626   0.5339
CORYes       -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
```

9. Full model

```
> summary(constant_model)

Call:
lm(formula = PRICE ~ 1, data = House_sales)

Residuals:
    Min      1Q  Median      3Q     Max
-578.41 -248.41 -109.41   91.59  991.59

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1158.41      49.48   23.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 392.7 on 62 degrees of freedom
```

10. Constant model

```
> anova(full_model, constant_model)
Analysis of Variance Table

Model 1: PRICE ~ SQFT + AGE + FEATS + NE + COR
Model 2: PRICE ~ 1
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     57 1195759
2     62 9561651 -5  -8365892 79.758 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

11. Comparisons of the null and the constant models

Question 6 – Use Stepwise methods to find the best models

Since the objective is to produce a predictive model, the AIC method will be used to construct the final model. In most trials, with one exception, the results of the final model were the same. The final model and the exception model can be noticed below.

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

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
```

12. Best model

```
Call:
lm(formula = PRICE ~ 1, data = House_sales)

Residuals:
    Min      1Q  Median      3Q     Max
-578.41 -248.41 -109.41   91.59  991.59

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1158.41      49.48   23.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 392.7 on 62 degrees of freedom
```

13. Exception model

Regarding the exception model, we can observe that it's the constant model, for which has already been established that is worse than the full model, so it won't be selected as the best model in the current case. On the other hand, observing the best model that was selected from most of the stepwise methods executed it can be noticed that the remaining variables are the square feet area and the number of features in the house.

Question 7 – Interpret the results of the best model

First, it can be observed that the constant of the model is negative (-175.92), which does not make any sense. Interpreting that value as it is, it would mean that a house with 0 sqft and no features would cost -175.92 monetary units. To fix that issue, the model's values are centered. Then, it can be remarked that a house that has average characteristics, that is 1730 square feet area and 4 features number of features. In the next step, to make the interpretation of the model easier, the square feet area is transformed to square meter area using the appropriate formula. So, it can be implied that for each additional square meter of the house, the price rises for 7.32 monetary units and that for each additional feature the price rises for almost 40 monetary units.

Furthermore, it can be observed that the intercept and the remaining variables are all statistically significant for a confidence level of (a = 0.05) since each of their p-values is lower than 'a'. Finally, it can be remarked that residual standard error for the best model is slightly lower than that of the full model. As a result, the predictions that will be conducted through the best model will be slightly better than those of the full model. The mathematical formulation of the final model can be described from the following formula: $PRICE = 1158.41 + 7.32 \times SQM + 39.84 \times FEATS + \varepsilon, \quad \varepsilon \sim N(0, 143.7^2)$

Also, it can be observed that the Adjusted R-squared value (86.61%) is approximately the same with the full model that was initially constructed. To put it another way, the 2 variables that were maintained in the model describe 86.61% of the houses' price variability and can produce good predictions for it.

```
#transform square feet to square meters
House_num$SQM <- House_num$SQFT/10.764
```

14. Transformation of square feet to square meters

```
#center the variables
House_num2 <- as.data.frame(scale(House_num, center = TRUE, scale = F))
House_num2$PRICE<-House_num$PRICE
```

15. Centering the variables

```
Call:
lm(formula = PRICE ~ . - AGE - SQFT, data = House_num2)

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) 1158.4127    18.1017  63.995   <2e-16 ***
FEATS         39.8369    15.3653   2.593   0.0119 *
SQM            7.3245     0.4163  17.594   <2e-16 ***
---
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
```

16. Best model – Centered

Below, the centered model without its intercept can be observed. In that model it can be noticed that the adjusted R-squared dropped to almost 6%, from 86.61% that was assigned to the model that included the intercept. The value dropped vastly because a statistically significant value was removed. As a result, it can be deducted that the intercept cannot be removed from the model because it will reduce greatly the predictive abilities of the model.

```
Call:
lm(formula = PRICE ~ . - 1 - AGE - SQFT, data = House_num2)

Residuals:
   Min     1Q Median     3Q    Max
   758   1087   1147   1252   1500

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
FEATS   39.837    126.817   0.314   0.7545
SQM      7.325      3.436   2.132   0.0371 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1186 on 61 degrees of freedom
Multiple R-squared:  0.08845,   Adjusted R-squared:  0.05856
F-statistic: 2.959 on 2 and 61 DF,  p-value: 0.05934
```

17. Centered model without the intercept

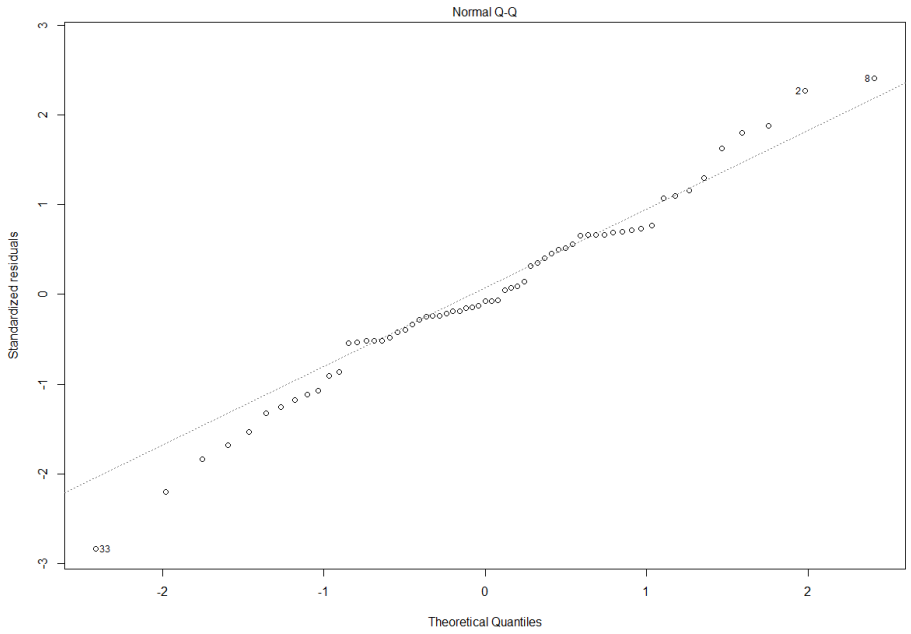
Question 8 – Check the assumptions of the final model

The first assumption that is examined for the final model, is the multi-collinearity of its explanatory variables. Since all the explanatory variables of the final model are numeric, the vif test is used to examine that assumption. It is noticed that there is no multi-collinearity between the variables because none of the variables have a vif value greater than 10.

```
> vif(model11)
   FEATS      SQM
1.153477 1.153477
```

18. Multi-collinearity

The second assumption that is examined, is the normality of the model's residuals. It is not comprehensible whether the residuals of the final model are following a normal distribution from the visual interpretation of the residuals. Although, it is clear from the normality tests that were conducted that the residuals indeed follow a normal distribution.

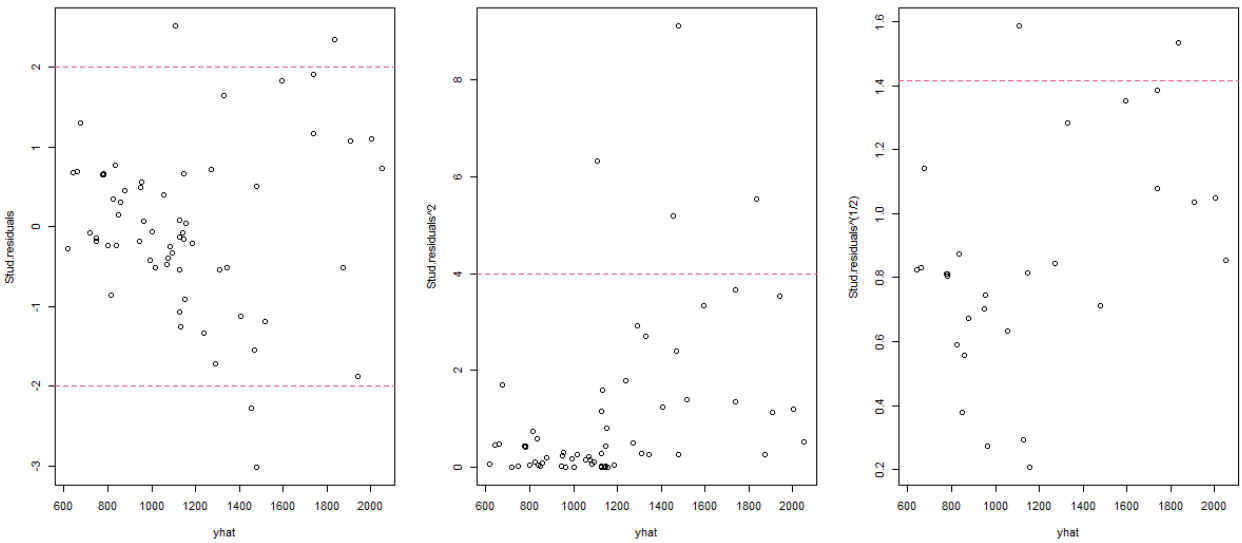


19. QQ-plot, examines normality of the residuals

```
> shapiro.test(model11$residuals)

        Shapiro-Wilk normality test

data:  model11$residuals
W = 0.98483, p-value = 0.6303

> ks.test(model11$residuals, 'pnorm', mean(model11$residuals), sd(model11$residuals))

        One-sample Kolmogorov-Smirnov test

data:  model11$residuals
D = 0.10234, p-value = 0.4924
alternative hypothesis: two-sided
```

20. Normality tests

The third assumption that was examined is the homoscedasticity of residuals. In other words, it is examined whether the residuals' variance differ for different values of the dependent variable. Observing the visual interpretation of the data, it is noticed (especially in the Fitted values vs Residuals graph) that the variance of the residuals is greater for higher values of the dependent variable (the price of the house).

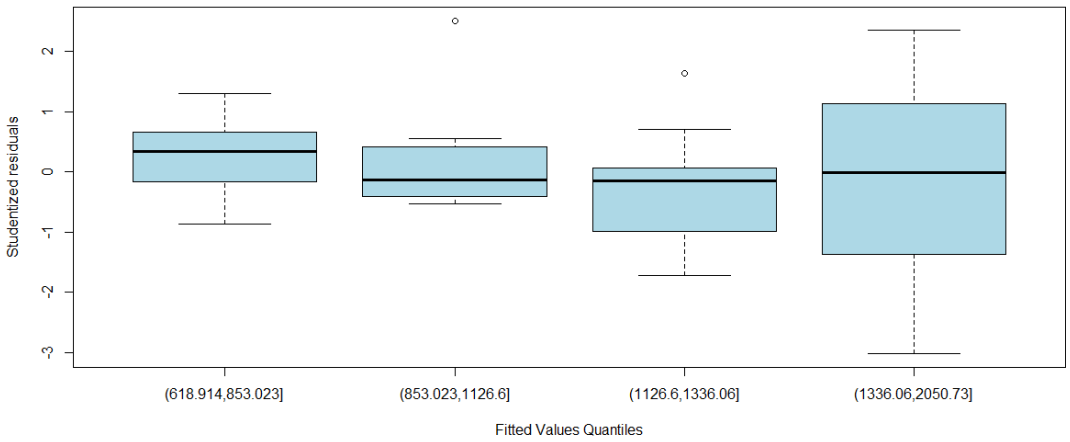


21. Fitted values vs residuals, squared residuals & squared root of residuals

To confirm the above assumption, a Levene test is conducted, which examines the equality of the variance in the 4 quantiles of the fitted values. Since the p-value of the test is <0.05, and with a 5% confidence level, the null hypothesis that the variance is equal for all quantiles of the fitted values is rejected. Also, a nonconstant variance test (ncv in short) is conducted to examine whether there is a linear effect on the variance of the fitted values. It is observed that the p-value of the ncv test is <0.05, so with a confidence level of 5%, the null hypothesis that there is no linear effect on the variance of the fitted values is rejected.

```
> ncvTest(model11)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 14.99402, Df = 1, p = 0.00010785
> leveneTest(rstudent(model11)~yhat.quantiles)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  3  9.4383 3.583e-05 ***
      58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

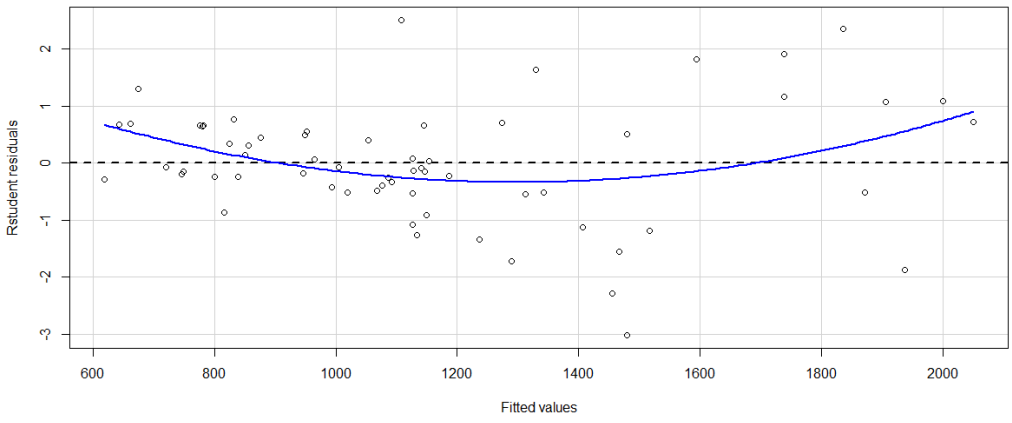
22. Homoscedasticity tests



23. Boxplots of variance per quantile of fitted values

It can be concluded that there is no homoscedasticity between the residuals of the fitted values. As a result, the error variance estimator is not estimated correctly, the standard errors are not estimated appropriately, and the performance of the hypothesis tests and the confidence intervals is affected negatively. Some methods that may resolve the homoscedasticity issue of the model, are the use of weighted least squares regression models, or the use of a general linear model with more complex covariates instead of the current model.

The final assumption that was examined is the existence of a linear relationship between the residuals and the fitted values of the model. From the graphical presentation of the models fitted values vs their covariates we cannot draw any certain assumption on whether there is a linear relationship between the explanatory and the dependent variables.



24. Plot of studentized residuals vs Fitted values

Observing the results of the Tukey's test, since that there are p-values that are less than 0.05, with a confidence level of 5% the null hypothesis that there is no linear relationship between the residuals and the fitted values is rejected. The departure of linearity may result to the appearance of the error variance as non-constant, even if it is constant due to the model misspecification. It can also result to an inadequate

model in terms of performing the predictions it was produced for. This issue can be resolved by transforming the response or the covariates, by using non-linear models to predict the price of the houses or by using polynomial regression models.

```
> residualPlots(model11, plot=F, type = "rstudent")
           Test stat Pr(>|Test stat|)
FEATS        -0.2876         0.774643
SQM           2.0388         0.045959 *
Tukey test    2.6002         0.009317 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

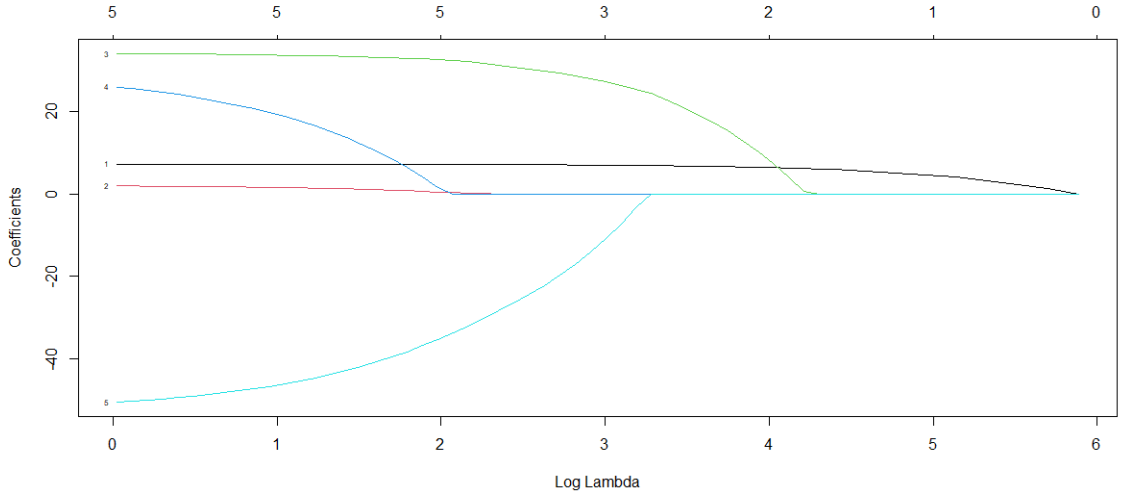
25. Non-linearity test

Since the model's variables have no meaning in terms of chronological sequence, the variables' independence will not be examined.

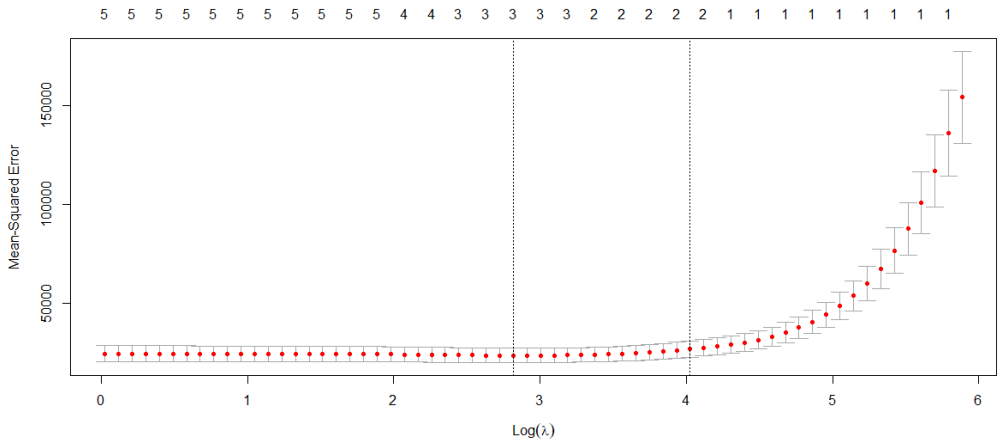
Question 9 – Conduct lasso and compare the results with the results of the stepwise method

To select the best variables for a model, the lasso method can be conducted instead of the stepwise methods. The lasso method shrinks variables and removes them from the final model it produces based on a tuning parameter λ. Very small values of that parameter can set many coefficients equal to 0, while a small value of that parameter can lead to over-fitted models (models with unnecessary explanatory variables).

To produce a model comparable to the model produced from the pairwise method, first, the square feet area in the is transformed to square meter area in the full model. The model that is selected as the best from the lasso method is the model with the largest lambda value, that has an error within one standard error of the minimum.



26. Shrinkage of variables based on the value of λ



27. Min MSE value and largest value of lambda such that error is within 1 se of the minimum

It is noticed that the stepwise and the lasso methods selected the same explanatory variables. Although, it should be mentioned that the coefficients that were produced from the lasso method are smaller than those of the stepwise method. These results occur due to the fact that the lasso method is attempting to set the coefficients equal to zero.

```
> coef(lasso1, s = "lambda.1se")
6 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 93.772282
SQM          6.444738
AGE          .
FEATS        7.365617
NEYes        .
CORYes       .
```

28. Best model with lasso method

R-Code

```
setwd("C:\\Users\\ninas\\R\\RPackages")


.libPaths('C:\\Users\\ninas\\R\\RPackages')

library('psych')

library('corrplot')

library(nortest)

library('randtests')

library('lmtest')

library(car)

require('glmnet')


#Question 1 - Read file and interpret descriptives

House_sales <- read.csv("C:/Users/ninas/OneDrive/Desktop/MSc Business Analytics/1st Quarter/Statistics for BA 1/Statistics_Lab_Assignment4/usdata", sep = "")

str(House_sales)

summary(House_sales)



#Question 2 - Update data types for variables

sum(is.na(House_sales))

House_sales$PRICE <- as.numeric(House_sales$PRICE)

House_sales$SQFT <- as.numeric(House_sales$SQFT)

House_sales$AGE <- as.numeric(House_sales$AGE)

House_sales$FEATS <- as.numeric(House_sales$FEATS)

House_sales$NE <- factor(House_sales$NE, levels = c(0,1), labels = c('No','Yes'))

House_sales$COR <- factor(House_sales$COR, levels = c(0,1), labels = c('No','Yes'))

str(House_sales)
```

```
#Question 3 - Perform descriptive analysis for each variable
index <- sapply(House_sales, class) == 'numeric'
House_num <- House_sales[,index]
#functions to show the variables' descriptives
summary(House_sales)
describe(House_num)
n <- nrow(House_sales)

#visualize numeric variables with histograms
numnames <- c('Price', 'Square Feet', 'Age', 'Features')
par(mfrow = c(2,2))
for (i in 1:3){
  hist(House_num[,i], main = numnames[i], xlab = numnames[i]
       ,col = 'lightblue')
}
plot(table(House_num[,4])/n, type='h', xlim=range(House_num[,4])+c(-1,1)
     , main='Features', ylab='Relative frequency', xlab = 'Features')

#visualize factor variables
factnames <- c('Northeast', 'Corner')
par(mfrow = c(1,1))
House_factors <- House_sales[,c(5,6)]
barplot(sapply(House_factors,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)


#Question 4 - Conduct pairwise comparisons and interpret the results
cor_table <- cor(House_num)
```

```
cor_table <- round(cor_table,2)
index <- c(1, 2, 4, 3)
cor_table <- cor_table[index, index]
#corrplots for numeric variables
par(mfrow = c(1,2))
corrplot(cor_table, method = 'number')
corrplot(cor_table, method = 'ellipse')

#boxplots for factors vs prices
for (j in 1:2){
  boxplot(House_sales$PRICE ~ House_sales[,j+4], col = 'lightblue', ylab = 'Price', xlab = factnames[j])
  abline(lm(House_sales$PRICE ~ House_sales[,j+4]), col = 2, lty = 2)
}


#Question 5 - Construct a full model
full_model <- lm(PRICE~., data = House_sales)
constant_model <- lm(PRICE~1, data = House_sales)
summary(full_model)
summary(constant_model)
anova(full_model, constant_model)
#adjusted R-squared =  86,4% - good fit for the model


#Question 6 - Find the best model using stepwise methods
#AIC Methods
model1 <- step(full_model, direction = 'both')
summary(model1)
model2 <- step(full_model, direction = 'backward')
```

```
summary(model2)

#model 3 ends up in a different model - the constant model

model3 <- step(constant_model, direction = 'forward')

summary(model3)

model4 <- step(constant_model, scope=list(lower=constant_model,upper=full_model), direction =
'forward')

summary(model4)

model5 <- step(constant_model, scope=list(lower=constant_model,upper=full_model), direction =
'both')

summary(model5)



#Question 7 - Interpret the best model

#transform square feet to square meters

House_num$SQM <- House_num$SQFT/10.764

House_num <- House_num[,-2]

#center the variables

House_num2 <- as.data.frame(scale(House_num, center = TRUE, scale = F))

House_num2$PRICE<-House_num$PRICE


#recreate the best model with centered covariates

model11 <- lm(PRICE~.-AGE, data=House_num2)

summary(model11)

model11$coefficients


#trying to remove the intercept

model12 <- lm(PRICE~.-1-AGE, data=House_num2)

summary(model12)

true.r2 <- 1-sum(model11$res^2)/((n-1)*var(House_num2$PRICE)); true.r2
```

```
#Question 8 - Check the assumptions of the model
#multi-collinearity of the x variables
vif(model11) #no multi collinearity

#normality of the residuals
par(mfrow = c(1,1))
plot(model11, which = 2)
shapiro.test(model11$residuals)
ks.test(model11$residuals, 'pnorm', mean(model11$residuals), sd(model11$residuals))

#Costant variance
Stud.residuals <- rstudent(model11)
yhat <- fitted(model11)
par(mfrow=c(1,3))
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)
plot(yhat, Stud.residuals^(1/2))
abline(h=sqrt(2), col=2, lty=2)

ncvTest(model11)

yhat.quantiles<-cut(yhat, breaks=quantile(yhat, probs=seq(0,1,0.25)), dig.lab=6)
table(yhat.quantiles)
leveneTest(rstudent(model11)~yhat.quantiles)
par(mfrow = c(1,1))
```

```
boxplot(rstudent(model11)~yhat.quantiles, col = 'lightblue', ylab = 'Studentized residuals'
      ,xlab = 'Fitted Values Quantiles')

#Non-linearity
residualPlot(model11, type='rstudent')
residualPlots(model11, plot=F, type = "rstudent")

#Question 9 - Conduct Lasso and compare the results with the stepwise method
House_sales$SQFT <- House_sales$SQFT/10.764
names(House_sales)[2] <- 'SQM'
full_model1 <- lm(PRICE~., data = House_sales)
summary(full_model1)

full_matrix <- model.matrix(full_model1)[,-1]
lasso <- glmnet(full_matrix, House_sales$PRICE)
par(mfrow = c(1,1))
plot(lasso, xvar = "lambda", label = T)
#Use cross validation to find a reasonable value for lambda
lasso1 <- cv.glmnet(full_matrix, House_num2$PRICE, alpha = 1)
lasso1$lambda
lasso1$lambda.min
lasso1$lambda.1se
plot(lasso1)
coef(lasso1, s = "lambda.min")
coef(lasso1, s = "lambda.1se")
```