Assignment_2

Burcin_Sarac_FT18

Q1

I began with reading the data in R and check the structure of it;

```
usdata <- read.table("usdata")
str(usdata)
```

```
## '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 ...
```

```
summary(usdata)
```

```
##      PRICE            SQFT           AGE            FEATS
##  Min.   : 580   Min.   : 970   Min.   : 2.00   Min.   :1.000
##  1st Qu.: 910   1st Qu.:1400   1st Qu.: 7.00   1st Qu.:3.000
##  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
##        NE             COR
##  Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.000   1st Qu.:0.0000
##  Median :1.000   Median :0.0000
##  Mean   :0.619   Mean   :0.2222
##  3rd Qu.:1.000   3rd Qu.:0.0000
##  Max.   :1.000   Max.   :1.0000
```

The dataset includes 63 observations and 6 variables. It seems from str() command that, all variables are integers. And from summary of data, all variables seems to variate in a normal distance, mean and median values are closely located and there is not seen any extreme values from min and max values.

Q2

```
numvar <- c("PRICE", "SQFT", "AGE", "FEATS")
boolvar <- c("NE", "COR")
usdata[,numvar] <- apply(usdata[,numvar], 2, function(x) as.numeric(x))
usdata[,boolvar] <- apply(usdata[,boolvar], 2, function(x) as.factor(x))
usnum <- usdata[numvar]
```

```r
usfac <- usdata[boolvar]
str(usdata)
```

```
## '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   : chr  "1" "1" "1" "1" ...
##  $ COR  : chr  "0" "0" "0" "0" ...
```

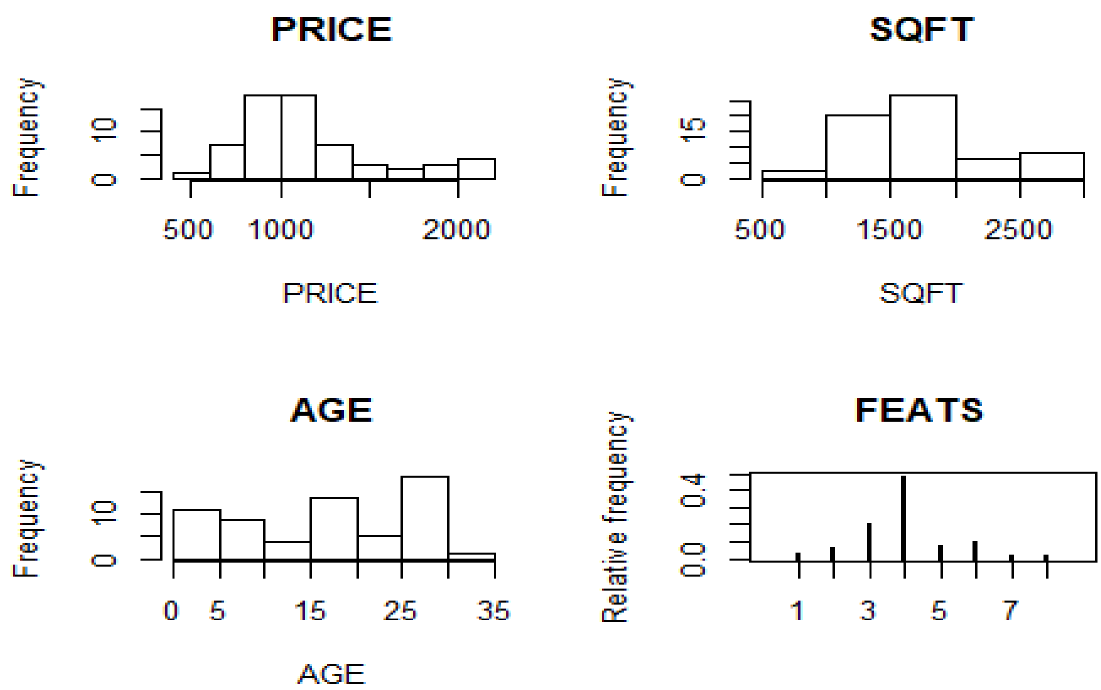
Q3

For analyzing each variable, firstly I started with numerical variables.

```r
round(t(describe(usdata[1:4])),2)
```

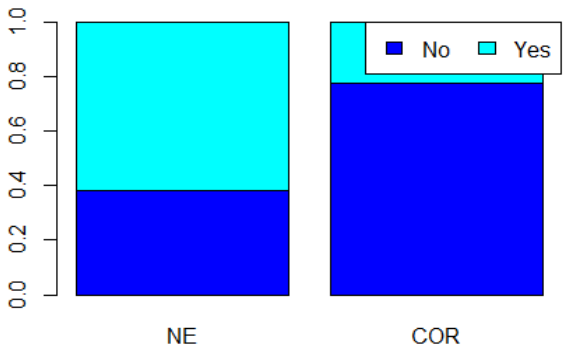
```
##              PRICE    SQFT   AGE FEATS
## vars          1.00    2.00  3.00  4.00
## n            63.00   63.00 63.00 63.00
## mean       1158.41 1729.54 17.46  3.95
## sd          392.71  506.70  9.60  1.28
## median     1049.00 1680.00 20.00  4.00
## trimmed    1105.96 1685.18 17.75  3.92
## 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
```

```r
par(mfrow=c(2,2))
for(i in 1:3){
  hist(usdata[,i], main=names(usdata)[i], xlab = names(usdata)[i])
}
n <- nrow(usdata)
plot(table(usdata[,4])/n, type='h', xlim=range(usdata[,4])+c(-1,1), main=name
s(usdata)[4], ylab='Relative frequency')
```



From generated codes and graphs it could be seen that, the numeric variables are not symetrically distributed. According to mean and median values and histograms we may roughly say that PRICE, SQFT have right skewed distribution and AGE has left skewed distribution. FEATS variable's distribution cannot understood from graph clearly, however it can determined by its "skew" value that it has right skewed distribution too. After that, I checked below the boolian variables(which were determined as factors in our dataset);

```
for(i in 5:6){
  tbl= table(usdata[i])
  tbl = cbind(tbl,round(prop.table(tbl),2))
  colnames(tbl) <- c(names(usdata)[i], "prob")
  print(tbl)
}

##   NE prob
## 0 24 0.38
## 1 39 0.62
##   COR prob
## 0  49 0.78
## 1  14 0.22

par(mfrow=c(1,1))
barplot(sapply(usfac,table)/n, col=4:8)
legend('topright', fil=4:8, legend=c('No','Yes'), ncol=2)
```

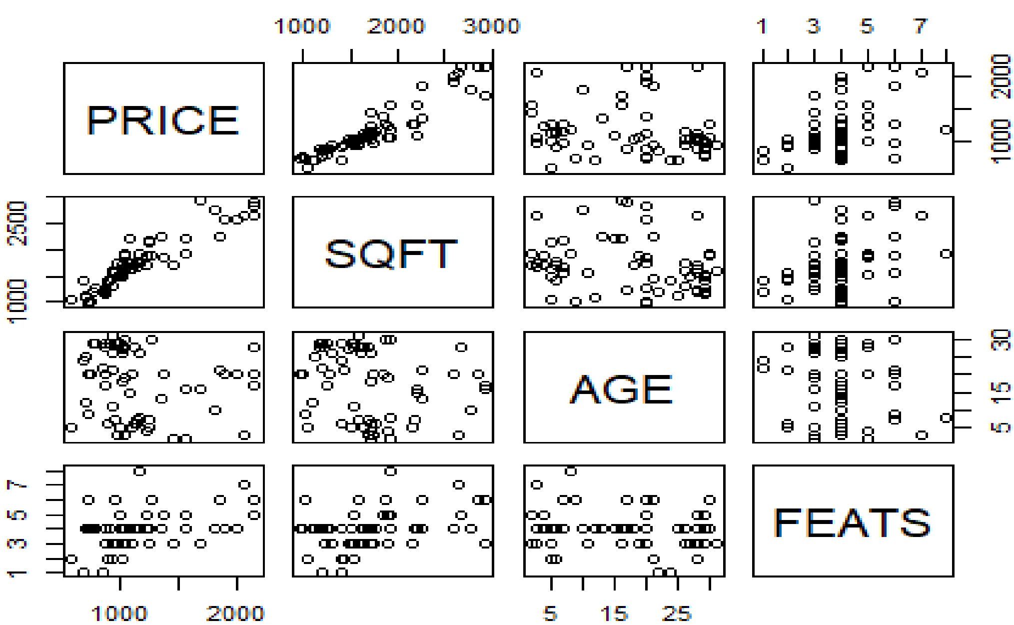


As it can be seen from tables that, in FEATS data about 70 percent of its variables are generated from 3-4. And around 62% of houses located in North East sector of the city and 22% of 63 houses are located in corner of a street.

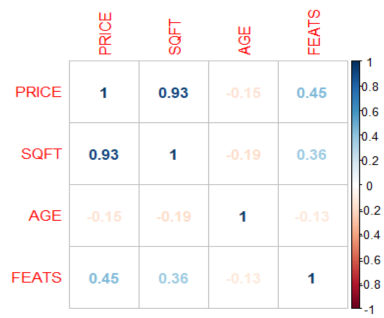
Q4

In this question, I tried to analyze numeric data in terms of correlations and visualization of bivariate associations;

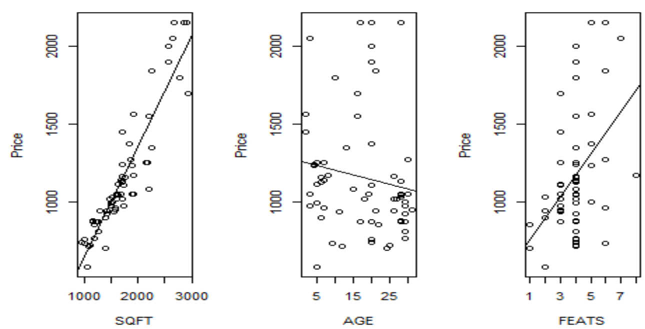
```
pairs(usnum)
```



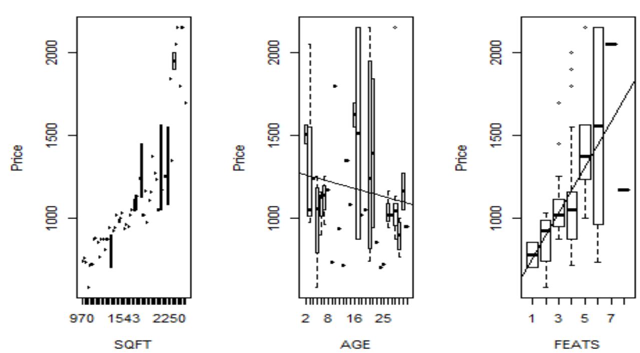
```
par(mfrow=c(1,1))
corrplot(cor(usnum), method = "number")
```



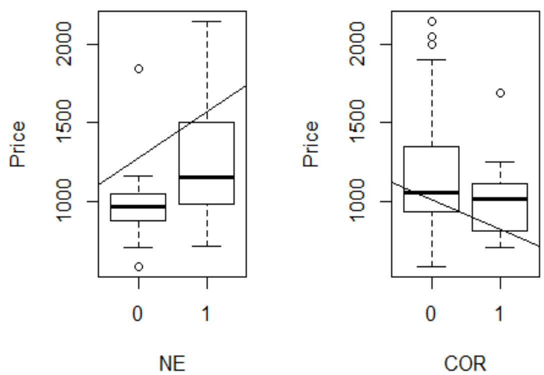

```
par(mfrow=c(1,3))
for(i in 2:4){
  plot(usnum[,i], usnum[,1], xlab=names(usnum)[i], ylab='Price')
  abline(lm(usnum[,1]~usnum[,i]))
}
```




```
par(mfrow=c(1,3))
for(i in 2:4){
  boxplot(usnum[,1]~ usnum[,i], xlab=names(usnum)[i], ylab='Price')
  abline(lm(usnum[,1]~usnum[,i]))
}
```



```
par(mfrow=c(1,2))
for(i in 1:2){
  boxplot(usnum[,1]~usfac[,i], xlab=names(usfac)[i], ylab='Price')
  abline(lm(usnum[,1]~usfac[,i]))
}
```



Firstly, I visualize the association between numeric variables and response variable. It seems that, PRICE has clear and strong positive linear relationship with SQFT and weaker but still positive linear relationship with FEATS variables. However, it is difficult to say something about relationship between PRICE and AGE from graphs. Therefore, correlation plot also supports the relationships just mentioned. Additionally it can be said that PRICE has weak negative linear relationship with AGE variable. And PRICE has close to perfect positive linear relationship with SQFT with 0.93. Moreover boxplots also supports the descriptions above about relations with ablines. Addition to this, the boolian variables can be assessed from boxplots. The NE variable has positive linear relation with PRICE but COR has negative. By the way for explaining, if there is a positive linear relationship determined between variables, like PRICE vs SQFT, this means if SQFT increases it causes to increase in PRICE as well.

Q5

```
model <- lm(PRICE~.,usdata)
summary(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
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: PRICE
##           Df  Sum Sq Mean Sq  F value  Pr(>F)
## SQFT       1 8184288 8184288 390.1325 < 2e-16 ***
## AGE        1    5501    5501   0.2622 0.61058
## FEATS      1  142786  142786   6.8064 0.01158 *
## NE         1    5574    5574   0.2657 0.60822
## COR        1   27743   27743   1.3225 0.25495
## Residuals 57 1195759   20978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After fitting a multiple linear regression model which includes all variables as predictors except PRICE as response variable, the fitted model has 86.4% adjusted R^2. If I only take into account this value, it can be said that the predictors would expain PRICE level with 86% accuracy. However, only taking into account R^2 may causes wrong interpretations, it cannot be trustworthy alone to explain model. Moreover, it can be seen from summary that, only SQFT,FEATS variables and constant variable seem significant in 95% confidence level. And the analysis of variance(anova) also supports our decision with testing the significance of each covariate using F-Tests. The p-values of predictors should be less than the selected confidence level for significance(<0.05).

Q6

```
mnull <- lm(PRICE~1,usdata)
step(mnull, scope=list(lower=mnull,upper=model), direction='forward')
```

```
## Start:  AIC=753.6
## PRICE ~ 1
##
##         Df Sum of Sq     RSS    AIC
## + SQFT   1   8184288 1377363 633.53
## + FEATS  1   1933150 7628501 741.37
## + NE     1   1304205 8257446 746.36
## + COR    1    393427 9168225 752.95
## <none>               9561651 753.60
```

```
## + AGE    1    218683 9342968 754.14
##
## Step:  AIC=633.53
## PRICE ~ SQFT
##
##         Df Sum of Sq     RSS    AIC
## + FEATS  1    138761 1238602 628.84
## + COR    1     56956 1320407 632.87
## <none>               1377363 633.53
## + NE     1      5874 1371489 635.26
## + AGE    1      5501 1371862 635.28
##
## Step:  AIC=628.84
## PRICE ~ SQFT + FEATS
##
##        Df Sum of Sq     RSS    AIC
## <none>              1238602 628.84
## + COR   1   22454.3 1216147 629.69
## + AGE   1    9525.5 1229076 630.35
## + NE    1     217.6 1238384 630.83

##
## Call:
## lm(formula = PRICE ~ SQFT + FEATS, data = usdata)
##
## Coefficients:
## (Intercept)         SQFT        FEATS
##   -175.9276       0.6805      39.8369
```

```r
step(model, direction='back')
```

```
## 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     25099 1229076 630.35
## <none>               1203977 631.05
## - 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
## - 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
## - FEATS  1    138761 1377363 633.53
## - SQFT   1   6389899 7628501 741.37

##
## Call:
## lm(formula = PRICE ~ SQFT + FEATS, data = usdata)
##
## Coefficients:
## (Intercept)         SQFT        FEATS
##   -175.9276       0.6805      39.8369
```

In this command, I tried to find best model for predicting PRICE with both backward and forward procedure. I firstly created a null model only included the constant variable. And I defined the scope of implemented method from null model to full, which included all variables. With the helps of this I aimed to test the model with all variables without any missing predictor. It resulted with a model only used SQRT and FEATS varibles as predictors and it determined from min AIC variable as 628.84.

Q7

```
modelf <- lm(PRICE~SQFT+FEATS, usdata)
summary(modelf)
```

```
##
## 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
```

After determining my model with only variables SQFT and FEATS, I created modelf for final model. And from summary of my final model, the adjusted R^2 equals to 0.8661 and all my coefficients and constant variable seems significant from their p-values, which are lower than 0.05 significance level. With this result my model become;

PRICE = -175.93 + 0.68xSQFT + 39.84xFEATS + E

E~N(0,143.7^2)

The intercept variable may be thought as fixed value added to the Price of a house. However, according to the constant with the value -175.93, I should assume if there is a house without any FEATS and SQFT, its PRICE should be -175.93. In other words the constant variable is not meaningful as a negative number and should be removed.

```r
modelf2 <- lm(PRICE~SQFT+FEATS-1, usdata)
summary(modelf2)
```

```
##
## 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
```

After removing the constant the adjusted R^2 become 0.9851 and still coefficients seem significant. So the last model should be like this;

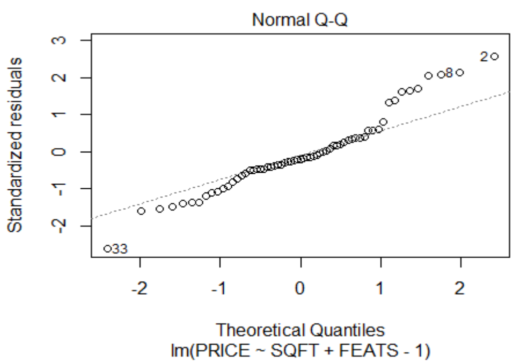
PRICE = 0.68xSQFT + 39.84xFEATS + E

E~N(0,149^2)

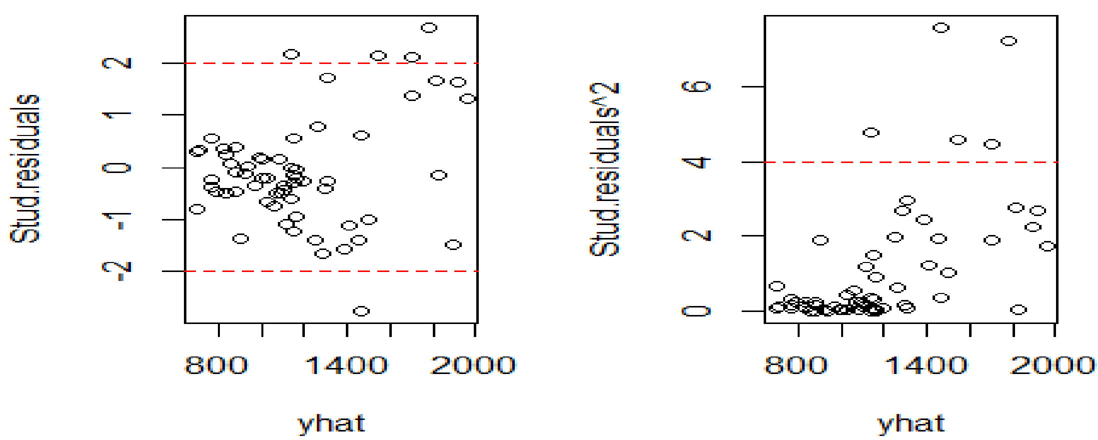
Q8

If model assumptions are not met, the model used for regression analysis may not be valid and it may cause to draw wrong conclusions and to make uneffective decisions.

```
par(mfrow=c(1,1))
plot(modelf2, which = 2)
```

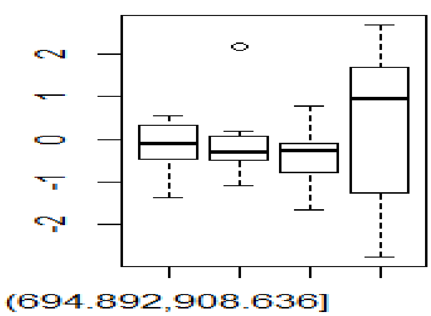


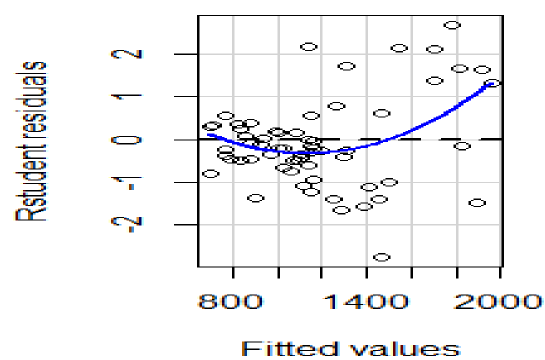

```
Stud.residuals <- rstudent(modelf2)
yhat <- fitted(modelf2)
par(mfrow=c(1,2))
plot(yhat, Stud.residuals)
abline(h=c(-2,2), col=2, lty=2)
plot(yhat, Stud.residuals^2)
abline(h=4, col=2, lty=2)
```



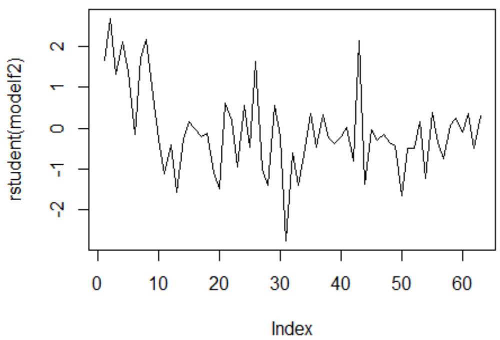

```
ncvTest(modelf2)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 23.79561, Df = 1, p = 1.0713e-06
```

```
yhat.quantiles<-cut(yhat, breaks=quantile(yhat, probs=seq(0,1,0.25)), dig.lab
=6)
table(yhat.quantiles)

## yhat.quantiles
## (694.892,908.636] (908.636,1135.61]  (1135.61,1309.6]  (1309.6,1959.15]
##                15                16                15                16

leveneTest(rstudent(modelf2)~yhat.quantiles)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value    Pr(>F)
## group  3  10.203 1.714e-05 ***
##       58
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

boxplot(rstudent(modelf2)~yhat.quantiles)
residualPlot(modelf2, type='rstudent')
```






```
residualPlots(modelf2, plot=F, type = "rstudent")

##            Test stat Pr(>|Test stat|)
## SQFT          2.9030         0.005164 **
## FEATS         1.6575         0.102639
## Tukey test    2.9836         0.002849 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(1,1))
plot(rstudent(modelf2), type='l')
```




```r
dwtest(modelf2)
```

```
##
##  Durbin-Watson test
##
## data:  modelf2
## DW = 1.415, p-value = 0.008131
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
durbinWatsonTest(modelf2)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2708188      1.415005   0.014
##  Alternative hypothesis: rho != 0
```

However, as it seems from graphs, both normality and homoscedasticity assumptions do not met. Because residuals in QQPlot don't seem normally distributed and they distributed like a funnel starts with low variance and getting higher. And also p-value in Non-constant variance score test and Levene's Test for Homogenity were above my confidence level(0.05). It ensures me that my model has also heteroscedasticity problem. And from the graph I can easily say that there is a linearity problem as well. However, it seems from DW test results that, the predictors are independent.

For fixing this problem, one mostly used solution is to transform response variable and also predictors. And then it is needed to check the assumptions again.

After spending too much time and many tries, I found a model mentioned below, which ensures all assumptions are met and meaningful. However, because of it is out of scope of this question, I would not prefer to add detailed test results and analysis in this report.

Last Model = PRICE ~ SQFT + I(FEATS^(7/8)) + E

E = N(0, 148.6^2)

```r
round(vif(model),1)
```

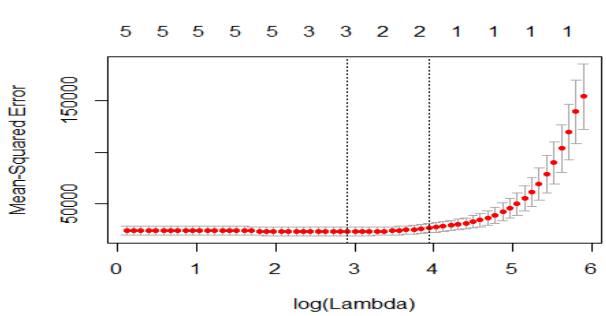
```
##   SQFT    AGE  FEATS     NE    COR
##    1.3    1.4    1.3    1.6    1.1
```

```r
round(vif(modelf2),1)

## Warning in vif.default(modelf2): No intercept: vifs may not be sensible.

##  SQFT FEATS
##   9.4   9.4
```

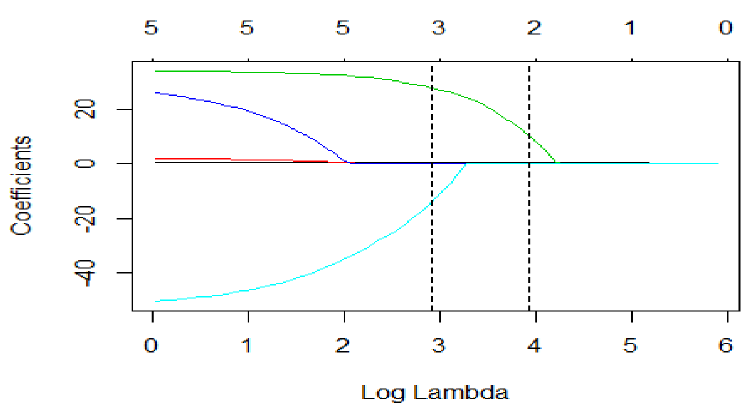
At last I checked multicollinearity both in my very beginning model with all variables and also my last model. From the queries, it can be seen that there was not any multicollinearity problem in any coefficient. I checked it with this formula and the VIF value in my last model is; VIF = $(1-R^{2})^{-1}$ = 66.67 And according to my VIF value, none of the coefficients got higher values than VIF value.

Q9

```r
X <- model.matrix(model)[,-1]
lasso <- glmnet(X, usdata$PRICE)
lasso1 <- cv.glmnet(X, usdata$PRICE, alpha = 1)
lasso1$lambda

##  [1] 360.429381 328.409829 299.234805 272.651609 248.429992 226.360156
##  [7] 206.250944 187.928178 171.233157 156.021275 142.160775 129.531604
## [13] 118.024373 107.539413  97.985908  89.281110  81.349622  74.122746
## [19]  67.537886  61.538006  56.071139  51.089934  46.551245  42.415760
## [25]  38.647661  35.214310  32.085967  29.235538  26.638334  24.271858
## [31]  22.115613  20.150923  18.360770  16.729650  15.243434  13.889249
## [37]  12.655367  11.531099  10.506708   9.573321   8.722853   7.947939
## [43]   7.241866   6.598519   6.012324   5.478206   4.991537   4.548103
## [49]   4.144062   3.775915   3.440473   3.134831   2.856341   2.602592
## [55]   2.371385   2.160717   1.968765   1.793866   1.634503   1.489299
## [61]   1.356993   1.236442   1.126600

lasso1$lambda.min

## [1] 18.36077

lasso1$lambda.1se

## [1] 51.08993

plot(lasso1)
```



```
coef(lasso1, s = "lambda.min")

## 6 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) -78.8223332
## SQFT          0.6532394
## AGE           .
## FEATS        27.9711239
## NE1           .
## COR1        -14.0444633

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

## 6 x 1 sparse Matrix of class "dgCMatrix"
##                      1
## (Intercept) 69.8128877
## SQFT         0.6059918
## AGE          .
## FEATS       10.2502734
## NE1          .
## COR1         .

plot(lasso1$glmnet.fit, xvar = "lambda")
abline(v=log(c(lasso1$lambda.min, lasso1$lambda.1se)), lty =2)
```



The Lasso technique also ended up with same variables(SQRT,FEATS) as determined in the stepwise method, according to determined lambda(lambda.1se). Lambda.1se is calculated with; 1 standard error + min lambda and I prefer to use this lambda value because of decreasing error and avoid overfitting. And in this lambda level the “coef(lasso1, s =”lambda.1se“)” query shows me the most important last 2 predictors as SQFT and FEATS.