ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
BUSINESS ANALYTICS Master of Science

Statistics for Business Analytics I (Part Time)

Lab Assignment 2: Real Estate Agency

Despotis Spyridon: p282211

2021

1. Read the "usdata" dataset and use str() to understand its structure

```
> any(is.na(df))
[1] FALSE
>
> str(df)
'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 ...
```

The dataset "usdata" contains **63** rows and **6** columns (variables) with the same data type: integer. We have to convert NE & COR columns to factor type and the rest columns to numerical. Last but not least, our dataset does not contain missing values.

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

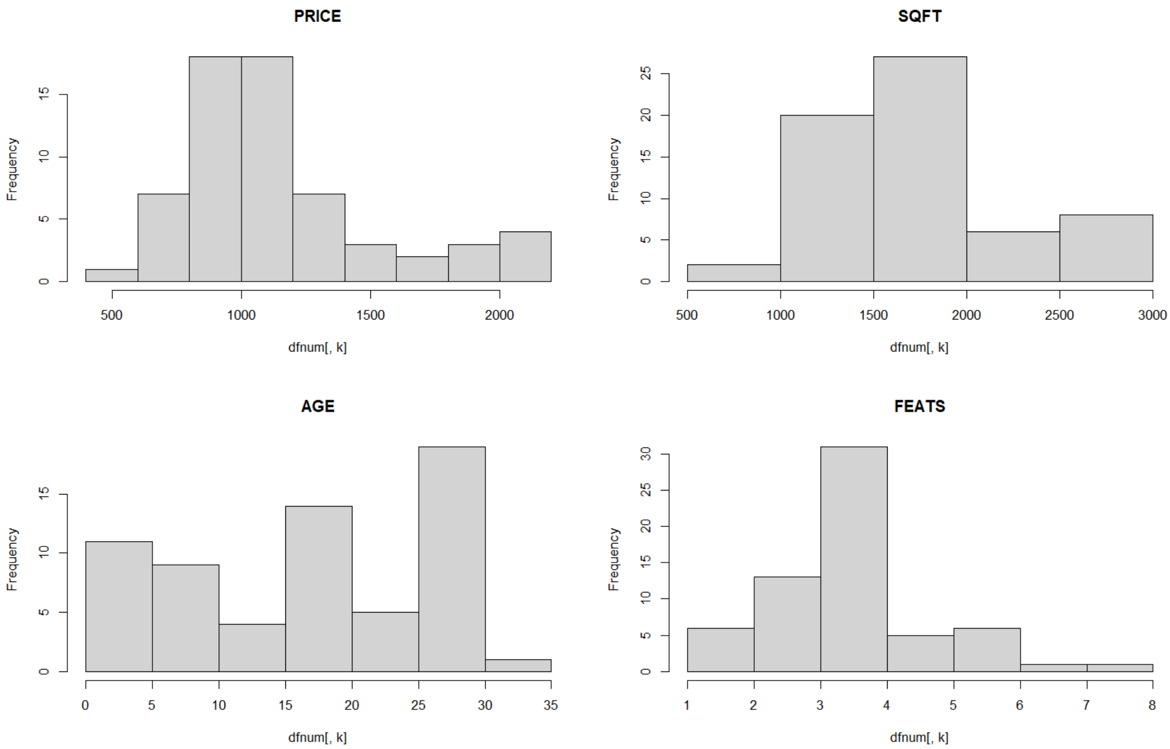
We proceed with converting each column to the appropriate data type. Specifically we converted the variables PRICE, SQFT, AGE, FEATS to **numeric** and the variables NE, COR to be **factors,** because these are indicator variables.

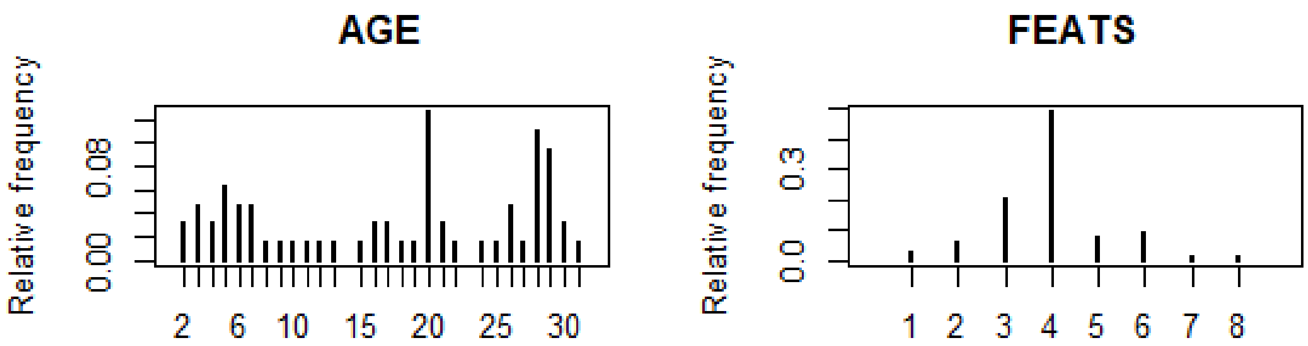
```
> str(df)
'data.frame':	63 obs. of  6 variables:
 $ PRICE: num  2050 2150 2150 1999 1900 ...
 $ SQFT : num  2650 2664 2921 2580 2580 ...
 $ AGE  : num  3 28 17 20 20 10 2 2 20 30 ...
 $ FEATS: num  7 5 6 4 4 4 5 3 5 6 ...
 $ NE   : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
 $ COR  : Factor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

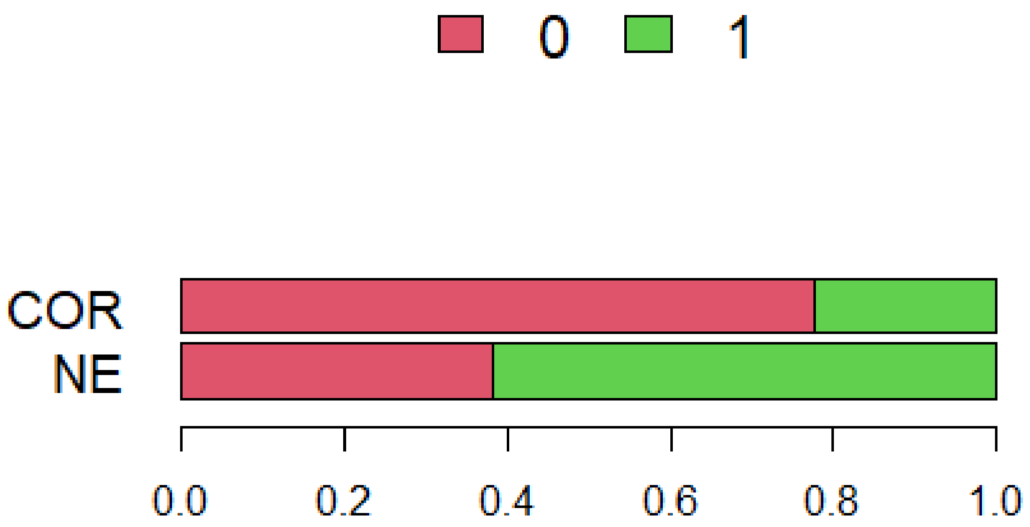
3. Perform descriptive analysis and visualization for each variable to get an initial insight of what the data looks like. Comment on your findings.

The variables do not appear to be symmetrically distributed, and thus to follow a normal distribution. The values for all the numeric variables seem plausible. Also there is evidence of some extreme values in the histograms, which could possibly indicate the existence of outliers. Finally, there seems to be moderate imbalance in the factor variable COR.

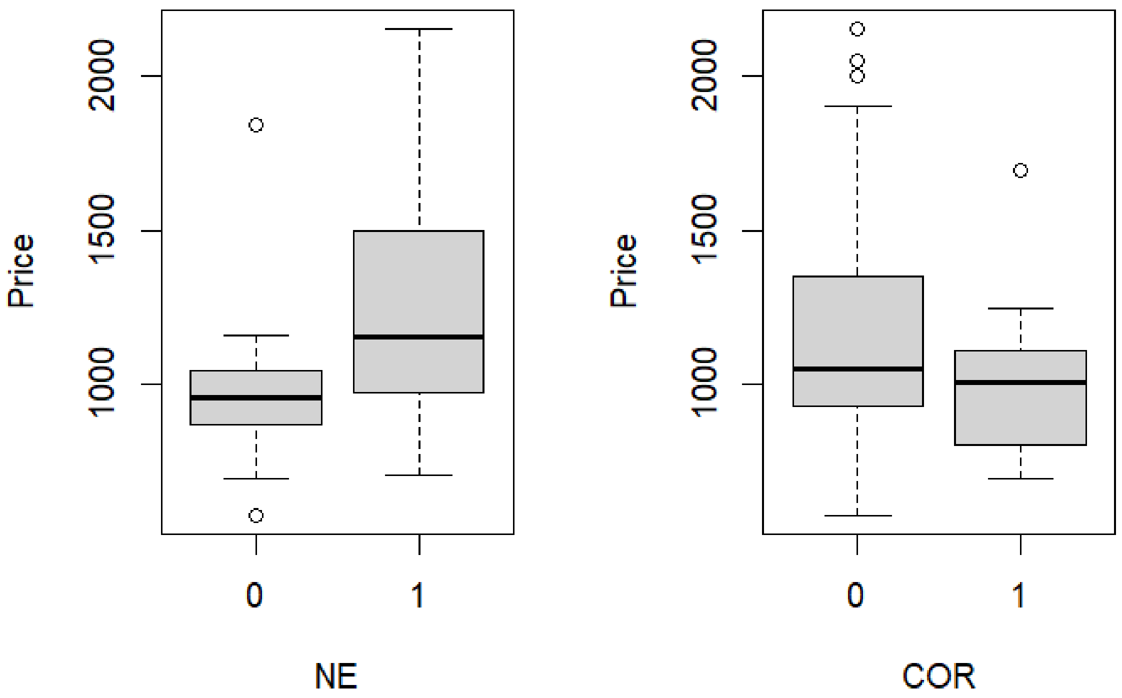
```
> summary(df)
     PRICE            SQFT             AGE              FEATS          NE     COR
 Min.   : 580   Min.   : 970   Min.   : 2.00   Min.   :1.000   0:24   0:49
 1st Qu.: 910   1st Qu.:1400   1st Qu.: 7.00   1st Qu.:3.000   1:39   1:14
 Median :1049   Median :1680   Median :20.00   Median :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(dfnum)
      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
```




AGE
Relative frequency
0.00
0.08
2
6
10
15
20
25
30
FEATS
Relative frequency
0.0
0.3
1
2
3
4
5
6
7
8


0
1
COR
NE
0.0
0.2
0.4
0.6
0.8
1.0

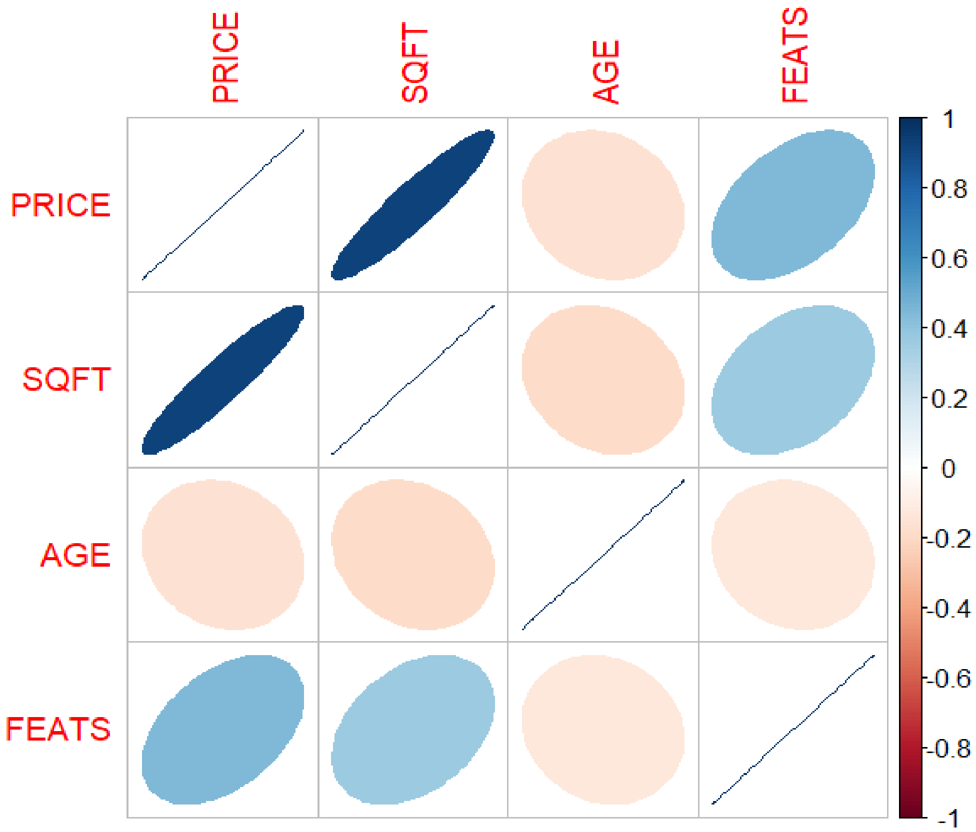
4. **Conduct pairwise comparisons between the variables in the dataset to investigate if there are any associations implied by the dataset.(Hint: Plot variables against one another and use correlation plots and measures for the numerical variables.). Comment on your findings. Is there a linear relationship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



In the above graphs we try to examine if there is any association between the price variable and the factor variables.

From the first boxplot we can see that there is a difference in price if the house is located in northeast sector of city. Specifically if the house is locate in northeast, the price is slightly higher, so we may have a positive effect of NE on the target variable.

In the second boxplot we examine if there is any association between the PRICE and the COR variable. As we can see, there is a relatively small difference in the median price corresponding to the two levels of COR.
So there is little evidence of association between COR and price.

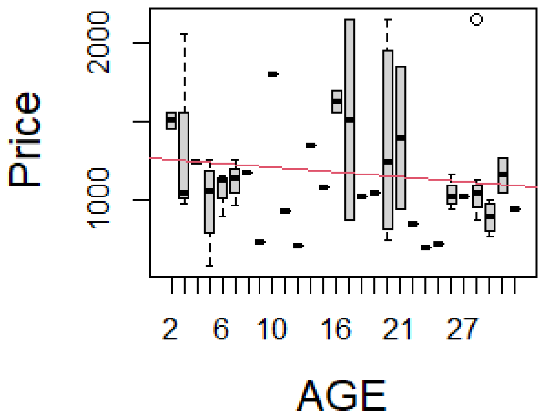


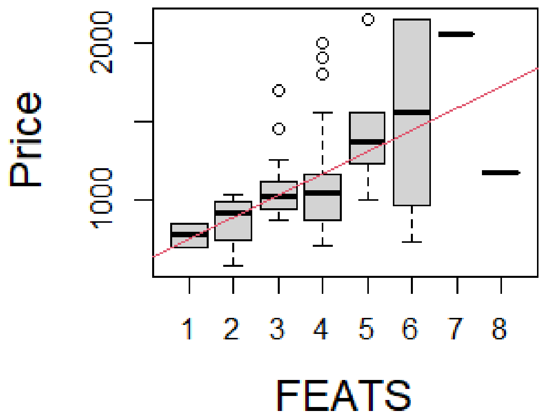
From the above correlation plot we can see two possible linear relationships of the PRICE variable especially with the SQFT variable and with the FEATS variable. The first relationship seems to be stronger. The next table show the correlation coefficients among all pairs of numeric variables:

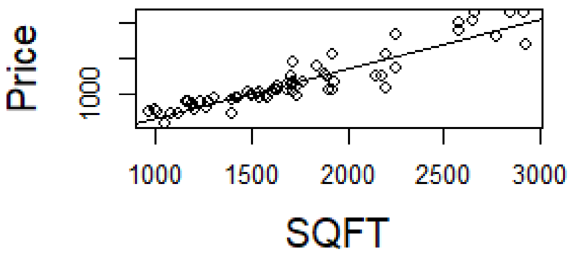
	PRICE	SQFT	AGE	FEATS
PRICE	1.00	0.93	-0.15	0.45
SQFT	0.93	1.00	-0.19	0.36
AGE	-0.15	-0.19	1.00	-0.13
FEATS	0.45	0.36	-0.13	1.00

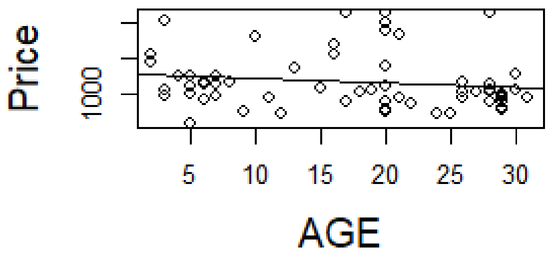
From the output we can conclude that the SQFT and the PRICE have a correlation coefficient of 0.93, while for FEATS with the PRICE r = 0.45. So, first pair appears to have a strong linear relationship, while for the second pair the relationship is moderate.

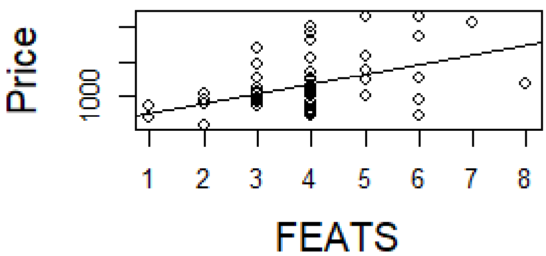
In the next graphs, the relationship of PRICE with the numeric predictors can be visually investigated.


Price
1000
2000
2
6
10
16
21
27
AGE


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


Price
1000
1000
1500
2000
2500
3000
SQFT


Price
1000
5
10
15
20
25
30
AGE


Price
1000
1
2
3
4
5
6
7
8
FEATS

5. **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).**

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

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
```

From the above output we conclude that our linear model seems to fit well to our data, since the Adjusted R-squared is 0.864 and therefore it is close to the 1. We continue with checking the F-statistic with the following hypothesis:

- *Ho: The model has no predictive capability (all of the regression coefficients are equal to zero)*
- *H1: : The model has predictive capability*

The corresponding p-value is <2.2*10-16 in test, so it is lower than the significance level of alpha = 0.05. So we reject our null hypothesis that our model has no predictive capability.

Lastly, we check the significance of the regressors. As we can see, only the SQFT and FEATS, and the Intercept, are statistically significant.

6. **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).**

```
> step(model1, direction='both')
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
+ 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=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 = df)

Coefficients:
(Intercept)         SQFT        FEATS
  -175.9276       0.6805      39.8369
```

```
> step(model1, direction='both', k=log(63)
Start:  AIC=645.48
PRICE ~ SQFT + AGE + FEATS + NE + COR

        Df Sum of Sq     RSS    AIC
- NE     1      8218 1203977 641.77
- AGE    1     19942 1215701 642.38
- COR    1     27743 1223502 642.78
<none>               1195759 645.48
- FEATS  1     93580 1289339 646.09
- SQFT   1   5717835 6913594 751.89

Step:  AIC=641.77
PRICE ~ SQFT + AGE + FEATS + COR

        Df Sum of Sq     RSS    AIC
- AGE    1     12171 1216147 638.26
- COR    1     25099 1229076 638.93
<none>               1203977 641.77
- FEATS  1    106953 1310930 642.99
+ NE     1      8218 1195759 645.48
- SQFT   1   6288869 7492846 752.81

Step:  AIC=638.26
PRICE ~ SQFT + FEATS + COR

        Df Sum of Sq     RSS    AIC
- COR    1     22454 1238602 635.27
<none>               1216147 638.26
- FEATS  1    104259 1320407 639.30
+ AGE    1     12171 1203977 641.77
+ NE     1       447 1215701 642.38
- SQFT   1   6352036 7568184 749.30

Step:  AIC=635.27
PRICE ~ SQFT + FEATS

        Df Sum of Sq     RSS    AIC
<none>               1238602 635.27
- FEATS  1    138761 1377363 637.82
+ COR    1     22454 1216147 638.26
+ AGE    1      9526 1229076 638.93
+ NE     1       218 1238384 639.40
- SQFT   1   6389899 7628501 745.66

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

Coefficients:
(Intercept)         SQFT        FEATS
  -175.9276       0.6805      39.8369
```

We select the stepwise procedure (direction=" both") as most appropriate because of double checking. The procedure is applied twice using the AIC and the BIC respectively. The results with AIC and BIC are the same and showed that the best subset of variables is SQFT and FEATS.

7. **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?**

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

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
```

From the above output we conclude that all variables are significant.

Parameter interpretation: When the house has zero Square Feet, no Number out of 11 features, then the expected value is equal to 175.93$

- ✓ This interpretation is not sensible
- ✓ We may consider them as fixed costs e.g. buying the land of the house without the house itself (which is frequent in Economics)

<u>Mathematical model is:</u>

$$\textbf{Price} = -175.93 + 0.68 \times \textbf{SQFT} + 39.84 \times \textbf{FEATS} + \varepsilon$$

$$\varepsilon \sim N(0,\ 143.72^{\wedge}2)$$

- If we compare two houses with the same characteristics which differ only by 1 sq.ft, then the expected difference in the price will be 0.68$ in favor of the larger house
- If we compare two houses with the same characteristics which differ only by 1 feature out of 11, then the expected difference in the price will be 39.84$ in favor of the house with the feature

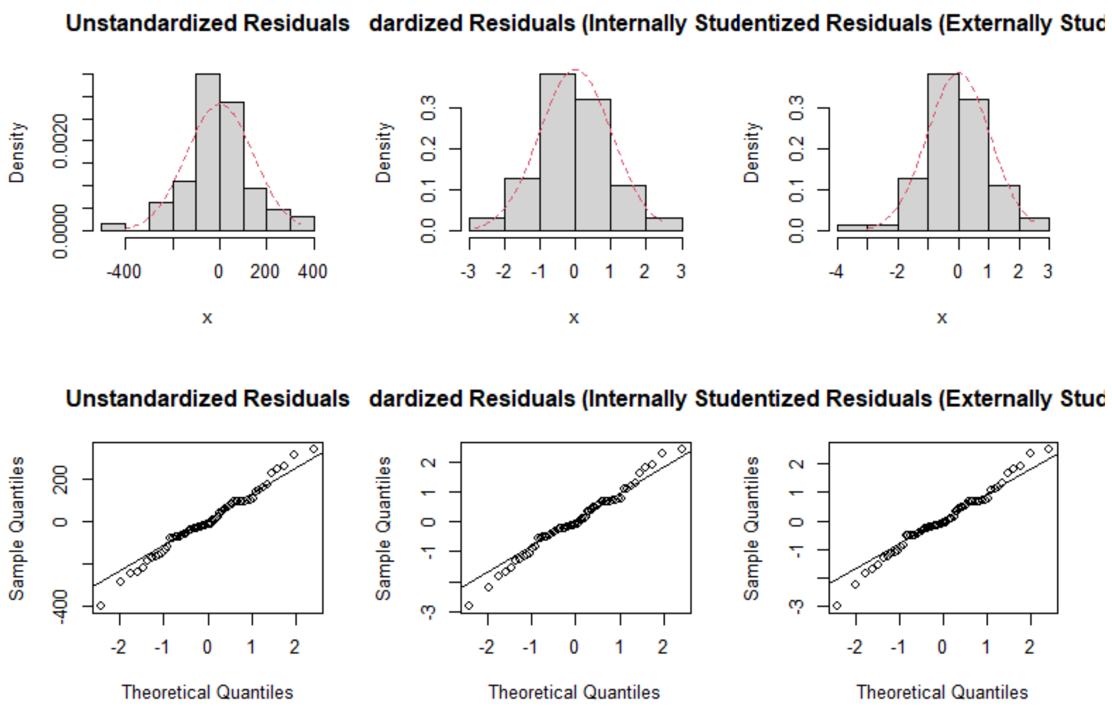
One way to improve the interpretability of the coefficients, especially the intercept, would be to rescale the predictors to mean. This way, the intercept would correspond to PRICE when all predictors are equal to the respective means. The remaining coefficients would be interpreted as the expected change in price, resulting from increasing each predictor by 1 in relation to its mean.

8. 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 to be checked:

- ➢ Independence of errors [not relevant since we do not have time-series]
- ➢ **<u>Normality of errors</u>**

  Comparisons for different residuals using QQplots and histograms.



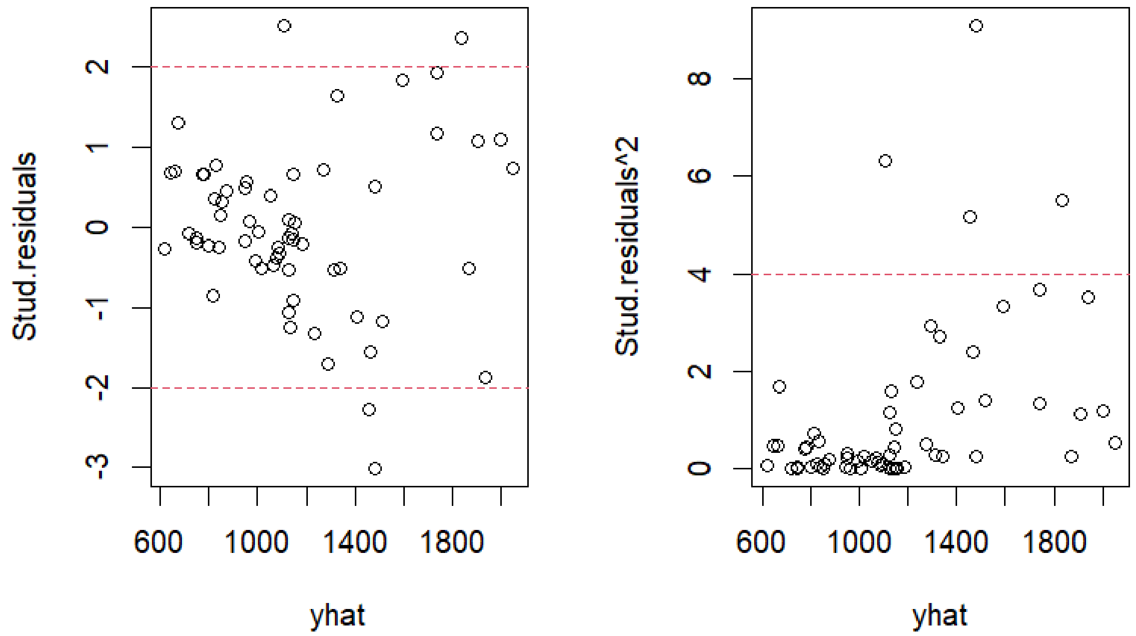
The residuals distribution seem to follow a normal distribution but though the graphs we can see that the left tails differ. We proceed with normality tests:

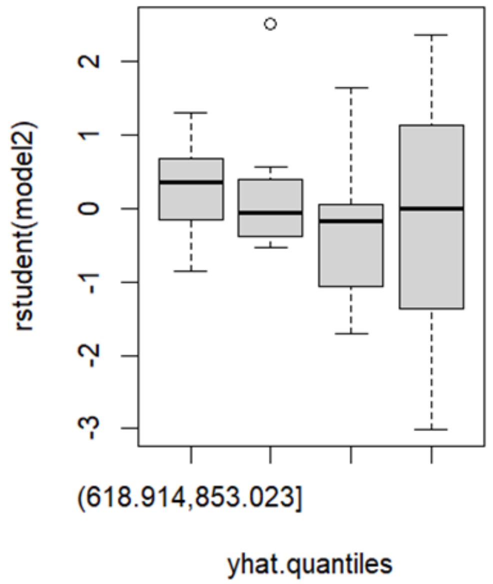
```
> normality.pvalues
                        Lillie KS        SW
Unstandardized     0.09853762 0.6303383
Standardized       0.07427354 0.6051777
Ext. Studentized 0.04855771 0.4590853
```

From the above outputs we can conclude that we don’t have strong evidence to reject the null hypothesis of normality in most cases.

- **Homogeneity of the variance of the residuals**

The variance of the residuals tends to shift for different yhats , which means we may not have homogeneity in the variance of our residuals.





We proceed with ncvTest:

- *Ho: The variance of the residuals is constant*
- *H1: : The variance of the residuals is not constant*

```
> ncvTest(model2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 14.99402, Df = 1, p = 0.00010785
```

The corresponding p-value is 0.00010785, so it is lower than the significance level of alpha = 0.05. So we reject our null hypothesis that *The variance of the residuals is constant.*

Similar conclusion can be obtained with the Levene test.

```
> leveneTest(rstudent(model2)~yhat.quantiles)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  3  9.9191 2.249e-05 ***
      58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Consequences of departures from homoscedasticity:

• Estimators of coefficients are still unbiased

• The error variance estimator is not estimated correctly

• Standard errors are not estimated appropriately

• This affects the performance of the hypothesis tests and confidence intervals.

How to cure the problem:

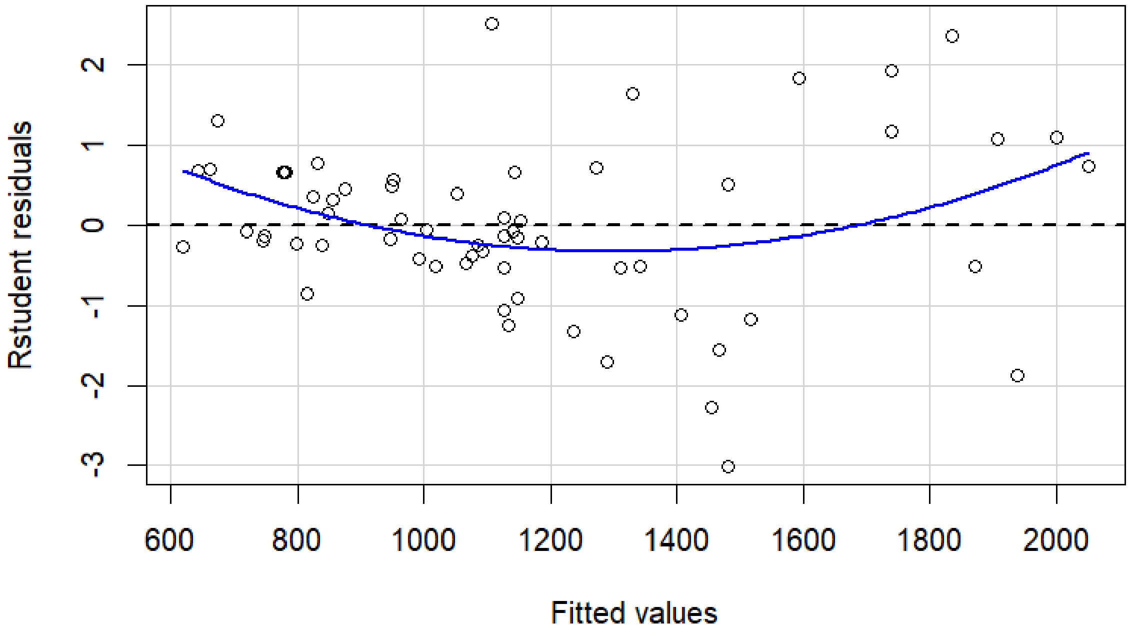
• Use weighted least squares regression models

• Use transformed response

• Use GLMs with more complicated distributions

• Use GAMLSS to use covariates in the variance componenets 117

➢ **Linearity**



From the above graph, we can see that the relation between PRICE and SQFT,FEATS is not Linear due to the fact that the blue loess line shows a non linear pattern between the residuals and the fitted values.

Consequences of departures from linearity:

• The error variance will appear as non-constant even if it is constant due to the model misspecification

• the model is inadequate, especially for prediction.

How to cure the problem:

• Transform the response

• Transform the covariates

• Use polynomial regression or non-parametric regression models

• Use non-linear models

- **Multi-collinearity**

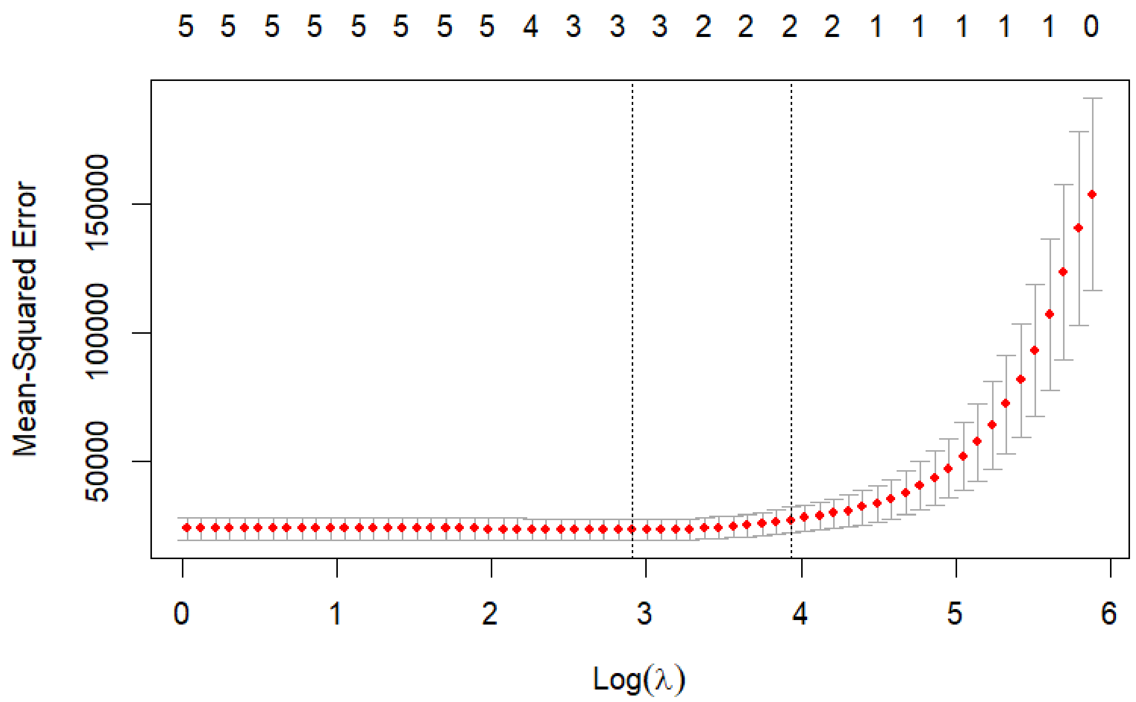
Multicollinearity occurs when independent variables in a regression model are correlated. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results. For a given predictor (p), multicollinearity can assessed by computing a score called the variance inflation factor (or VIF), which measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

The following results show that multicollinearity is not an issue in the final model.

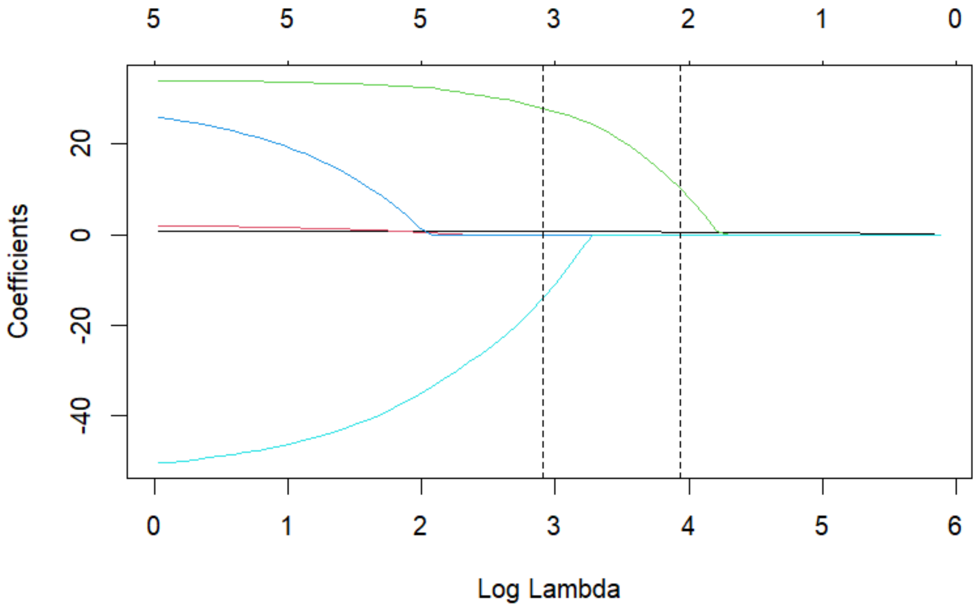
```
> vif(model2)
    SQFT    FEATS
1.153477 1.153477
```

9. 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.

An important aspect in Lasso is to select the optimal λ. We can draw a graph of log(λ) and MSE (mean squared error). The first candidate is the λ at which the minimal MSE is achieved (lambda.min) but it is likely that this model have many variables. The second is the largest λ at which the MSE is within one standard error of the minimal MSE (lambda.1se).



The following figure shows the change of estimated coefficients with respect to the change of the penalty parameter log(λ) which is the shrinkage path. The vertical lines are drawn at lambda.min and lambda.1se



The following result reports the estimated coefficients under the MSE minimized 1se λ. Thus, we end up with the same set of predictors (SQFT, FEATS) selected also with the stepwise procedure above.

```
> coef(lasso1, s = "lambda.1se")
6 x 1 sparse Matrix of class "dgCMatrix"
                    s1
(Intercept) 69.8128873
SQFT         0.6059918
AGE          .
FEATS       10.2502735
NE1          .
COR1         .
```

APPENDIX

```
##Question 1 ---

df <- read.table("usdata")
any(is.na(df))
str(df)

##Question 2 ---

df[, 5:6] <- lapply(df[, 5:6], as.factor)
df[, 1:4] <- lapply(df[, 1:4], as.numeric)
str(df)

##Question 3 ---

summary(df)
describe(dfnum)


install.packages('psych')
library('psych')
index <- sapply(df, class) == "numeric";
dfnum <- df[,index];
dffactor <- df[,!index]
round(t(describe(dfnum)),2)

library(purrr)
library(tidyr)
library(ggplot2)
dfnum %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()

par(mfrow=c(2,2));
for(k in 1:4){
  hist(dfnum[,k], main=names(dfnum)[k])
}


library('psych')
par(mfrow=c(1,1))
barplot(sapply(dffactor,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)
```

```r
plot(table(dfnum[,3])/n, type='h', xlim=range(dfnum[,3])+c(-1,1),
main=names(dfnum)[3], ylab='Relative frequency')
plot(table(dfnum[,4])/n, type='h', xlim=range(dfnum[,4])+c(-1,1),
main=names(dfnum)[4], ylab='Relative frequency')


##Question 4 ---

install.packages('corrplot')
library('corrplot')
corrplot(cor(dfnum), method = "ellipse")


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

round(cor(dfnum),2)

pairs(dfnum)

par(mfrow=c(2,2))
for(j in 2:4){
  plot(dfnum[,j], dfnum[,1], xlab=names(dfnum)[j],
ylab='Price',cex.lab=1.5)
  abline(lm(dfnum[,1]~dfnum[,j]))
}

par(mfrow=c(1,2))
for(j in 3:4){
  boxplot(dfnum[,1]~dfnum[,j], xlab=names(dfnum)[j],
ylab='Price',cex.lab=1.5)
  abline(lm(dfnum[,1]~dfnum[,j]),col=2)
}

#Price (our response) on factor variables
par(mfrow=c(1,2))
for(j in 1:2){
  boxplot(dfnum[,1]~dffactor[,j], xlab=names(dffactor)[j],
ylab='Price',cex.lab=1.0)
}



##Question 5 ---

model1 <- lm(PRICE ~., data = df)
summary(model1)
```

```
##Question 6

step(model1, direction='both')

step(model1, direction='both', k=log(63))


##Question 7

model2 <- lm(formula = PRICE ~ SQFT + FEATS, data = df)
summary(model2)


##Question 8

model2 <- lm(PRICE ~.-AGE-NE-COR, data = df)

#Normality of the residuals
plot(model2, which = 2)


102

par( mfcol=c(2,3) )
allres <- list(); allres[[1]] <- model2$res
allres[[2]] <- rstandard(model2); allres[[3]] <- rstudent(model2)
mt<-c(); mt[1] <- 'Unstandardized Residuals'
mt[2] <- 'Standardized Residuals (Internally Studentized)'
mt[3] <- 'Studentized Residuals (Externally Studentized)'
for (i in 1:3){
  x<-allres[[i]]
  hist(x, probability=T, main=mt[i])
  x0<-seq( min(x), max(x), length.out=100)
  y0<-dnorm( x0, mean(x), sd(x) )
  lines(x0,y0, col=2, lty=2)
  qqnorm(x, main=mt[i])
  qqline(x)
}


normality.pvalues <- matrix( nrow=3,ncol=2)
row.names(normality.pvalues) <- c( 'Unstandardized',
                                   'Standardized', 'Ext.
Studentized' )
colnames(normality.pvalues) <- c( 'Lillie KS', 'SW' )
library(nortest)
allres <- list()
allres[[1]] <- model2$res; allres[[2]] <- rstandard(model2);
allres[[3]] <- rstudent(model2)
for (i in 1:3){
  res <- allres[[i]]
```

```
  normality.pvalues[i,1]<-lillie.test(res)$p.value
  normality.pvalues[i,2]<-shapiro.test(res)$p.value
}
normality.pvalues


#Costant variance
Stud.residuals <- rstudent(model2)
yhat <- fitted(model2)
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)



library(car)
ncvTest(model2)
# ------------------
yhat.quantiles<-cut(yhat, breaks=quantile(yhat,
probs=seq(0,1,0.25)), dig.lab=6)
table(yhat.quantiles)
leveneTest(rstudent(model2)~yhat.quantiles)
boxplot(rstudent(model2)~yhat.quantiles)


##Multi Collinearity

#Using VIF
require(car)

vif(model2)

##Question 9

require(glmnet)
mfull <- lm(PRICE~.,data=df)
X <- model.matrix(mfull)[,-1]
lasso <- glmnet(X, df$PRICE)
plot(lasso, xvar = "lambda", label = T)

#Use cross validation to find a reasonable value for lambda
lasso1 <- cv.glmnet(X, df$PRICE, alpha = 1)
plot(lasso1)

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