Harshad Kumar Elangovan

19200349 || Predictive Analytics Assignment 1

The provided data file “House.csv” contains imformation on the sale of 76 single family homes in Dublin during 2005.Using R, the data is read and analysis is carried out on Price.

```
housedata=read.csv("C:\\house.csv",header=TRUE)
head(housedata)
```

```
##   Ã¯..Price  Size Lot Bath Bed Year Garage School
## 1    388.0 2.180   4    3   4 1940      0   High
## 2    450.0 2.054   5    3   4 1957      2   High
## 3    386.0 2.112   5    2   4 1955      2   High
## 4    350.0 1.442   6    1   2 1956      1   Alex
## 5    155.5 1.800   1    2   4 1994      1   Alex
## 6    220.0 1.965   5    2   3 1940      1   Alex
```

```
names(housedata)[names(housedata)=='Ã¯..Price']<-'Price'
head(housedata)
```

```
##   Price  Size Lot Bath Bed Year Garage School
## 1 388.0 2.180   4    3   4 1940      0   High
## 2 450.0 2.054   5    3   4 1957      2   High
## 3 386.0 2.112   5    2   4 1955      2   High
## 4 350.0 1.442   6    1   2 1956      1   Alex
## 5 155.5 1.800   1    2   4 1994      1   Alex
## 6 220.0 1.965   5    2   3 1940      1   Alex
```

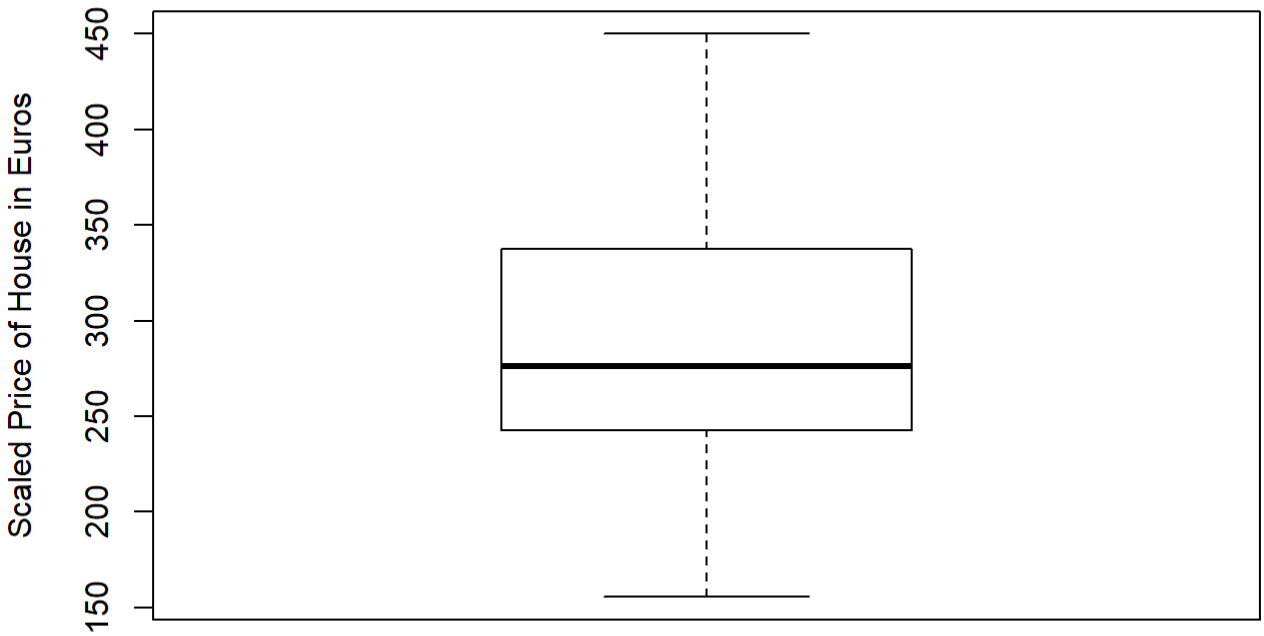
```
str(housedata)
```

```
## 'data.frame':    76 obs. of  8 variables:
##  $ Price : num  388 450 386 350 156 ...
##  $ Size  : num  2.18 2.05 2.11 1.44 1.8 ...
##  $ Lot   : int  4 5 5 6 1 5 4 4 4 5 ...
##  $ Bath  : num  3 3 2 1 2 2 1.1 2 2.1 2.1 ...
##  $ Bed   : int  4 4 4 2 4 3 4 4 4 3 ...
##  $ Year  : int  1940 1957 1955 1956 1994 1940 1958 1961 1965 1968 ...
##  $ Garage: int  0 2 2 1 1 1 1 2 2 2 ...
##  $ School: Factor w/ 6 levels "Alex","High",..: 2 2 2 1 1 1 4 4 4 4 ...
```

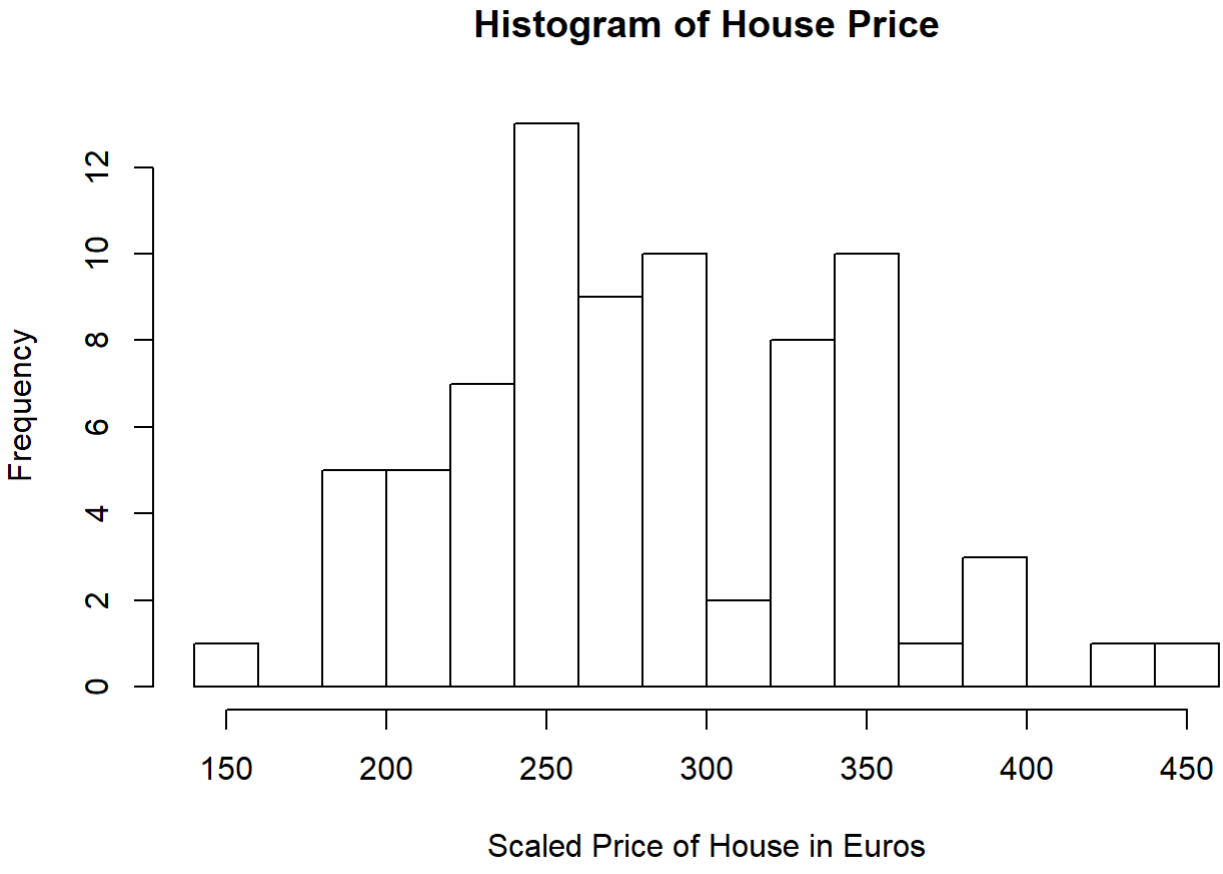
Exploratory Data Analysis

1. Using a boxplot, histogram and summary. Describe the distribution of the sales price of the houses.

```
boxplot(housedata$Price,ylab="Scaled Price of House in Euros")
```




```
hist(housedata$Price,,xlab="Scaled Price of House in Euros",breaks = 20,main = "Histogram of House Pric
e")
```




```
summary(housedata$Price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   242.8   276.0   285.8   336.8   450.0
```

The graphs and summary of the data shows that the average price of the house is 285.8 (scaled by 1000 Euros) and the price range of most of the house is between 242.8 to 336.8 thousand Euros(1st Quartile and 3rd Quartile).The minimum sale price of the house is 155.5 and the max price is 450 thousand Euros. The box plot suggest that there are no outliers in the given price range.

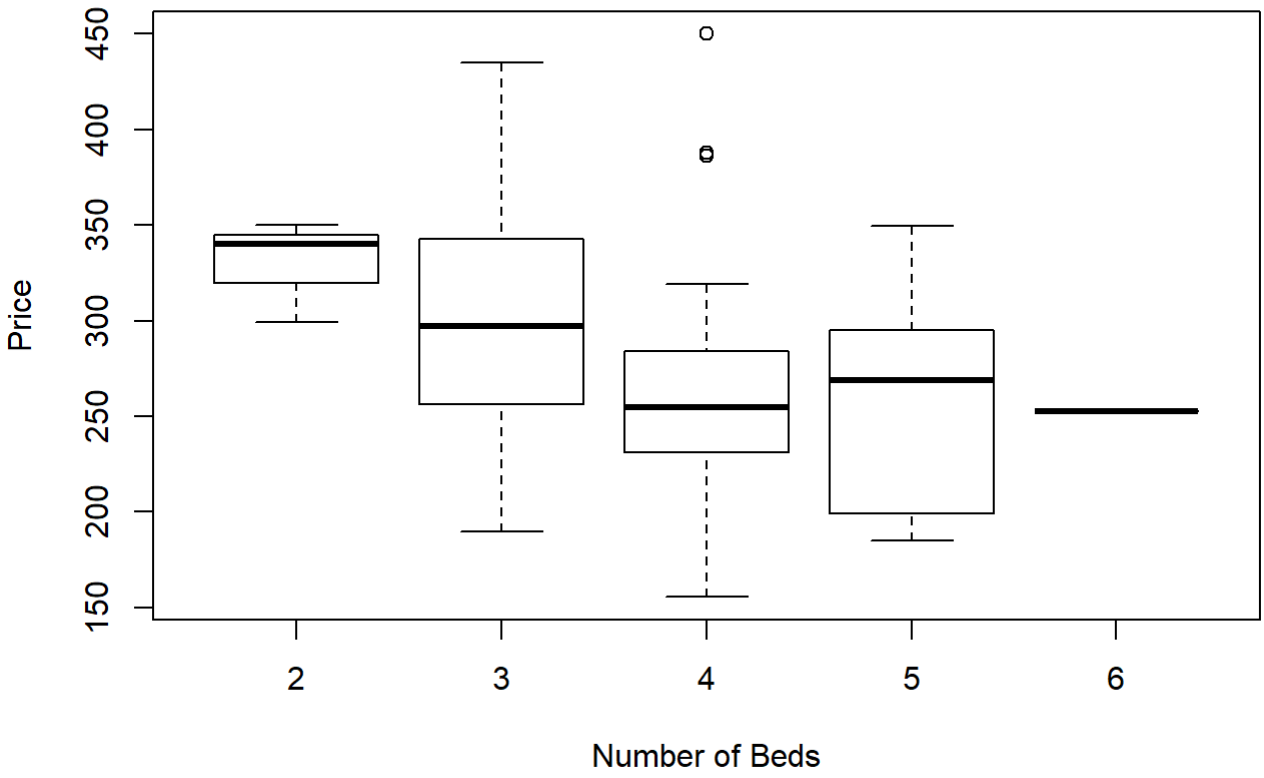
2. Convert all the categorical variables to factors. Using the summary and a boxplot describe how sales prices vary with respect to the number of bedrooms, bathrooms, garage size and school.

```
housedata$Bed<-factor(housedata$Bed)
housedata$Bath<-factor(housedata$Bath)
housedata$Garage<-factor(housedata$Garage)
#School is already a factor
str(housedata)
```

```
## 'data.frame':    76 obs. of  8 variables:
##  $ Price : num  388 450 386 350 156 ...
##  $ Size  : num  2.18 2.05 2.11 1.44 1.8 ...
##  $ Lot   : int  4 5 5 6 1 5 4 4 4 5 ...
##  $ Bath  : Factor w/ 6 levels "1","1.1","2",..: 5 5 3 1 3 3 2 3 4 4 ...
##  $ Bed   : Factor w/ 5 levels "2","3","4","5",..: 3 3 3 1 3 2 3 3 3 2 ...
##  $ Year  : int  1940 1957 1955 1956 1994 1940 1958 1961 1965 1968 ...
##  $ Garage: Factor w/ 4 levels "0","1","2","3": 1 3 3 2 2 2 2 3 3 3 ...
##  $ School: Factor w/ 6 levels "Alex","High",..: 2 2 2 1 1 1 4 4 4 4 ...
```

Number of Bedrooms

```
boxplot(housedata[,c("Price")]~housedata[,c("Bed")],ylab = "Price",xlab = "Number of Beds")
library(psych)
```



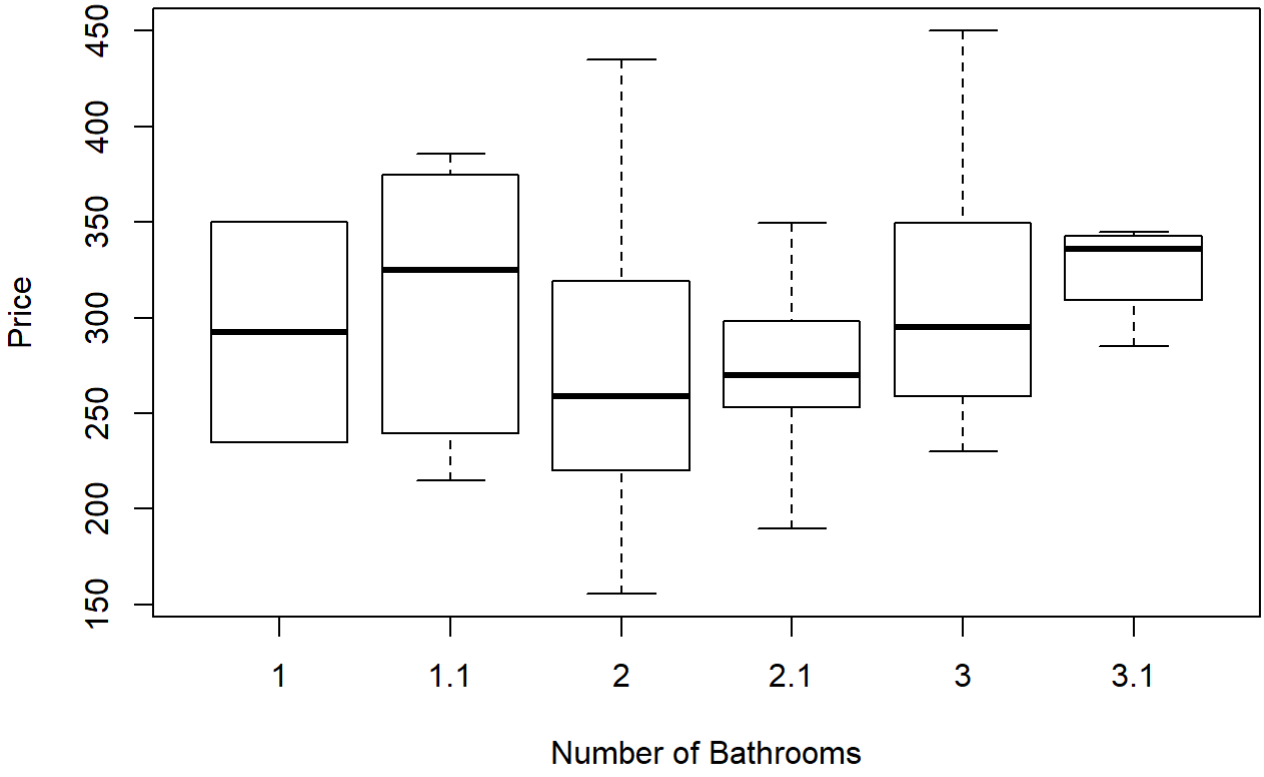

```
describeBy(housedata[,c("Price")],housedata[,c("Bed")])
```

```
##
##  Descriptive statistics by group
## group: 2
##    vars n   mean    sd median trimmed   mad min max range  skew kurtosis
## X1    1 3 329.63 27.01  339.9  329.63 14.97 299 350    51 -0.33    -2.33
##       se
## X1 15.59
## ------------------------------------------------------------
## group: 3
##    vars  n   mean    sd median trimmed   mad   min max range skew kurtosis
## X1    1 43 297.28 55.66    297  297.34 62.27 189.5 435 245.5 0.08    -0.69
##      se
## X1 8.49
## ------------------------------------------------------------
## group: 4
##    vars  n  mean    sd median trimmed   mad   min max range skew kurtosis
## X1    1 24 266.6 65.25 254.45  260.39 39.29 155.5 450 294.5 1.12     1.06
##       se
## X1 13.32
## ------------------------------------------------------------
## group: 5
##    vars n  mean   sd median trimmed    mad min   max range skew kurtosis
## X1    1 5 259.5 68.3    269   259.5 103.78 185 349.5 164.5 0.09    -1.98
##       se
## X1 30.55
## ------------------------------------------------------------
## group: 6
##    vars n  mean sd median trimmed mad   min   max range skew kurtosis se
## X1    1 1 252.5 NA  252.5   252.5   0 252.5 252.5     0   NA       NA NA
```

The summary and plots shows that houses of 3 bedrooms were sold the most(43 houses) and its maximum price reached 435 thousand Euros.The price of houses of bedrooms above three were less than the price of 3 bedrooms.This suggest that the price of houses decreases as the number of beds increases. The plot also shows that two houses of four bedrooms has unusually high price.

Number of Bathrooms:

```
boxplot(housedata[,c("Price")]~housedata[,c("Bath")],ylab = "Price",xlab = "Number of Bathrooms")
```



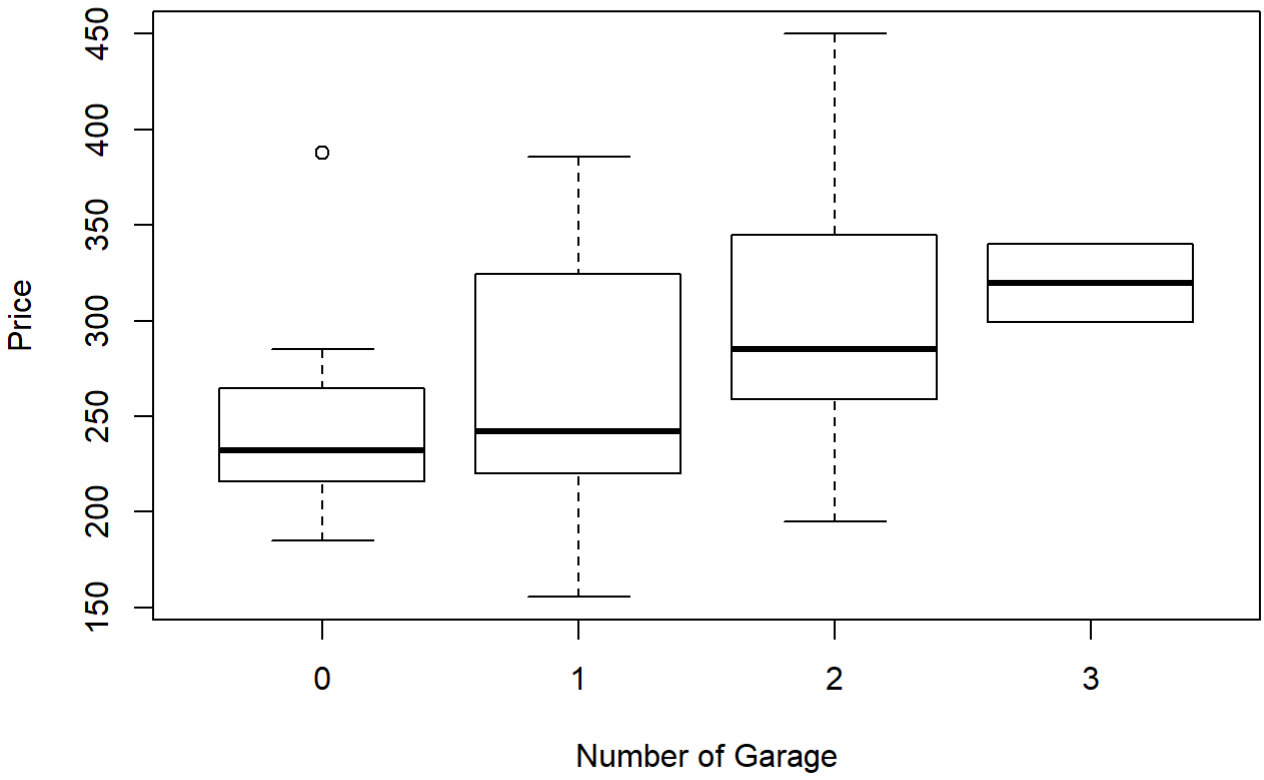

```
#library(psych)
describeBy(housedata[,c("Price")],housedata[,c("Bath")])
```

```
##
##  Descriptive statistics by group
## group: 1
##    vars n  mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 2 292.5 81.32  292.5   292.5 85.25 235 350   115    0    -2.75
##      se
## X1 57.5
## ------------------------------------------------------
## group: 1.1
##    vars n  mean    sd median trimmed  mad min   max range  skew kurtosis
## X1    1 5 307.9 77.55    325   307.9 89.7 215 385.5 170.5 -0.15    -2.16
##       se
## X1 34.68
## ------------------------------------------------------
## group: 2
##    vars  n  mean    sd median trimmed   mad   min max range skew kurtosis
## X1    1 33 270.7 64.89    259  267.28 62.27 155.5 435 279.5 0.52    -0.43
##      se
## X1 11.3
## ------------------------------------------------------
## group: 2.1
##    vars  n   mean    sd median trimmed   mad   min   max range  skew
## X1    1 16 274.52 38.54 269.95  275.24 34.92 189.5 349.5   160 -0.06
##    kurtosis   se
## X1    -0.21 9.63
## ------------------------------------------------------
## group: 3
##    vars  n   mean    sd median trimmed  mad min max range skew kurtosis
## X1    1 13 307.76 65.93    295   301.9 77.1 230 450   220 0.63    -0.76
##       se
## X1 18.29
## ------------------------------------------------------
## group: 3.1
##    vars n   mean    sd median trimmed   mad min max range  skew kurtosis
## X1    1 7 324.25 27.14    336  324.25 13.34 285 345    60 -0.69    -1.63
##       se
## X1 10.26
```

There is not much information about houses with 1 and 1.1 bathrooms since the count is low(1 and 5 respectively).Apart from that, the plot trend suggests that the price of house increases with increase in bathrooms.

Garage count:

```
boxplot(housedata[,c("Price")]~housedata[,c("Garage")],ylab = "Price",xlab = "Number of Garage")
```



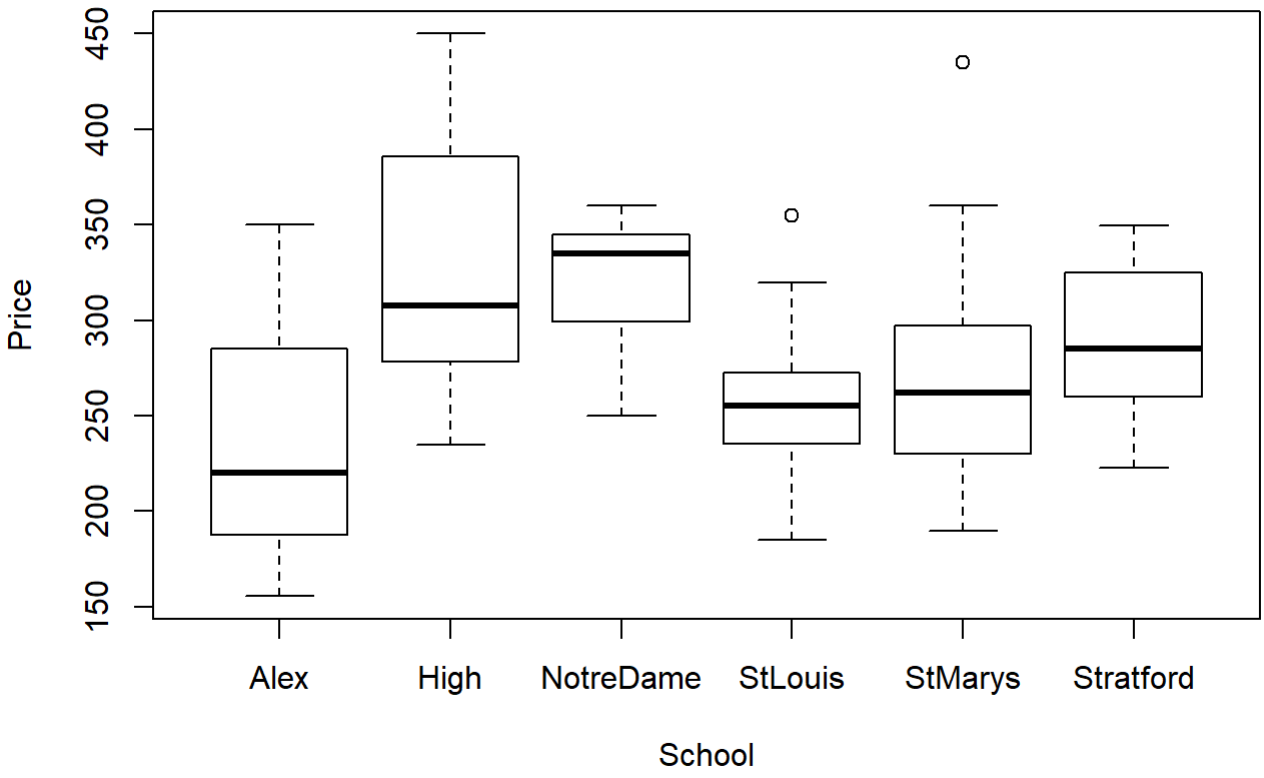

```
#Library(psych)
describeBy(housedata[,c("Price")],housedata[,c("Garage")])
```

```
##
##  Descriptive statistics by group
## group: 0
##    vars  n   mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 11 246.85 56.48    232  238.04 41.36 185 388   203  1.2     0.81
##       se
## X1 17.03
## ------------------------------------------------------------
## group: 1
##    vars  n  mean    sd median trimmed   mad   min   max range skew
## X1    1 13 260.6 67.45    242   258.8 63.75 155.5 385.5   230 0.38
##    kurtosis    se
## X1    -1.13 18.71
## ------------------------------------------------------------
## group: 2
##    vars  n   mean    sd median trimmed   mad min max range skew kurtosis
## X1    1 50 299.57 55.14    285  297.24 55.82 195 450   255 0.49    -0.07
##     se
## X1 7.8
## ------------------------------------------------------------
## group: 3
##    vars n   mean    sd median trimmed   mad min   max range skew kurtosis
## X1    1 2 319.45 28.92 319.45  319.45 30.32 299 339.9  40.9    0    -2.75
##       se
## X1 20.45
```

From the boxplot, we can see that there is a house with unsually high price with no garage.Other than that, the trend suggests increase in house price with increase in garage count.

School:

```
boxplot(housedata[,c("Price")]~housedata[,c("School")],ylab = "Price",xlab = "School")
```




```
#Library(psych)
describeBy(housedata[,c("Price")],housedata[,c("School")])
```

```
##
##  Descriptive statistics by group
## group: Alex
##    vars n   mean    sd median trimmed   mad   min max range skew kurtosis
## X1    1 3 241.83 99.07    220  241.83 95.63 155.5 350 194.5 0.21    -2.33
##      se
## X1 57.2
## ------------------------------------------------------
## group: High
##    vars  n  mean   sd median trimmed   mad min max range skew kurtosis
## X1    1 12 327.1 67.8  307.5  324.02 90.44 235 450   215 0.27    -1.45
##       se
## X1 19.57
## ------------------------------------------------------
## group: NotreDame
##    vars  n  mean    sd median trimmed   mad   min   max range  skew
## X1    1 14 319.1 37.75 334.88  321.47 22.05 249.9 359.9   110 -0.82
##    kurtosis    se
## X1       -1 10.09
## ------------------------------------------------------
## group: StLouis
##    vars  n   mean    sd median trimmed  mad min max range skew kurtosis
## X1    1 15 257.45 43.84    255  255.52 29.5 185 355   170  0.5    -0.35
##       se
## X1 11.32
## ------------------------------------------------------
## group: StMarys
##    vars  n   mean    sd median trimmed   mad   min max range skew kurtosis
## X1    1 26 269.76 58.84    262   265.2 50.41 189.5 435 245.5 0.82     0.36
##       se
## X1 11.54
## ------------------------------------------------------
## group: Stratford
##    vars n   mean    sd median trimmed   mad   min   max range  skew
## X1    1 6 287.82 45.27    285  287.82 48.26 222.5 349.5   127 -0.03
##    kurtosis    se
## X1     -1.6 18.48
```

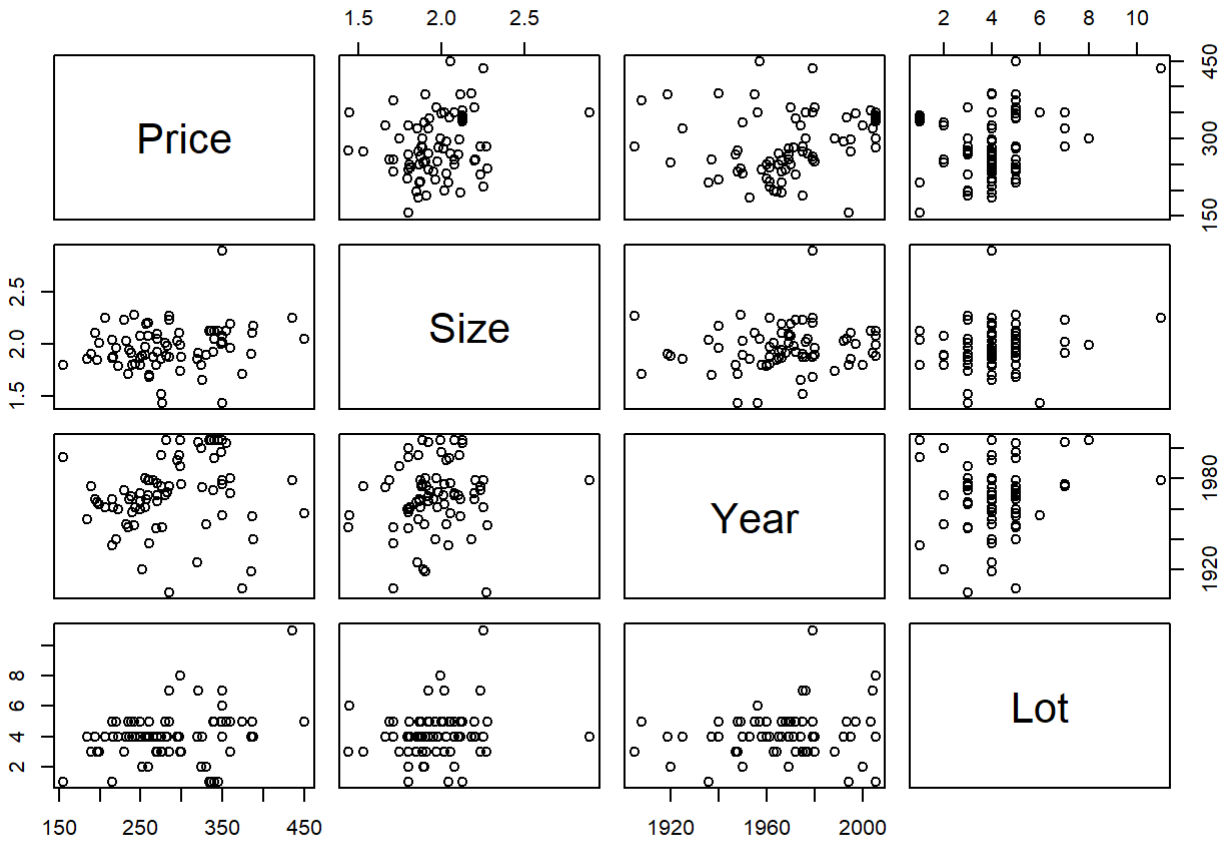
From the plot, we see that the Price of houses near High School and NotreDame school are comparatively higher than other four schools. Houses near Alex school have low price compared with other schools.

3. Using the summary, correlation and the pairs plots discuss the relationship between the response sales price and each of the numeric predictor variables.

```
cor(housedata[,c("Price","Size","Year","Lot")])
```

```
##           Price       Size        Year         Lot
## Price 1.0000000 0.20143783  0.15412476  0.24423228
## Size  0.2014378 1.00000000  0.17656934  0.04079199
## Year  0.1541248 0.17656934  1.00000000 -0.03933975
## Lot   0.2442323 0.04079199 -0.03933975  1.00000000
```

```
#data1<-c("Price","Size","Lot","Year")
pairs(housedata[,c("Price","Size","Year","Lot")])
```




```
#lines(housedata$Price,fitted.values(data1),col="red")
```

The pair plots suggests that there is a linear relationship between the response sales price and the numeric predictor variables Size,Year and Lot and there is a positive correlation among each variables with Price. The correlation values are 0.201,0.154 and 0.244 for variables Size,year and Lot respectively. So, as the correlation increases, the price of house also increases.

Regression Model:

1. Fit a multiple linear regression model to the data with sales price as the response and size, lot, bath, bed, year, garage and school as the predictor variables. Write down the equation for this model.

```
Rmodel<-lm(Price~Size+Lot+Year+Bath+Bed+Garage+School,data = housedata)
summary(Rmodel)
```

```
##
## Call:
## lm(formula = Price ~ Size + Lot + Year + Bath + Bed + Garage +
##     School, data = housedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -87.601 -21.429   0.173  24.248  72.581
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -884.3531   661.7589  -1.336  0.18693
## Size              59.4503    28.9813   2.051  0.04501 *
## Lot               11.7701     3.7842   3.110  0.00296 **
## Year               0.5567     0.3384   1.645  0.10565
## Bath1.1          135.8983    49.1990   2.762  0.00779 **
## Bath2             73.9317    47.8636   1.545  0.12817
## Bath2.1           76.9433    48.1208   1.599  0.11556
## Bath3             98.0694    50.4663   1.943  0.05711 .
## Bath3.1           85.8037    54.3074   1.580  0.11985
## Bed3            -228.1052    70.6732  -3.228  0.00211 **
## Bed4            -238.2609    72.4883  -3.287  0.00177 **
## Bed5            -237.6155    76.4733  -3.107  0.00299 **
## Bed6            -255.0211    88.0955  -2.895  0.00543 **
## Garage1          -10.9191    22.4871  -0.486  0.62920
## Garage2           18.2435    18.2212   1.001  0.32111
## Garage3         -209.9038    80.7191  -2.600  0.01193 *
## SchoolHigh       113.2774    36.9154   3.069  0.00334 **
## SchoolNotreDame   80.9317    35.6893   2.268  0.02730 *
## SchoolStLouis      9.0367    37.3439   0.242  0.80969
## SchoolStMarys     27.3408    35.8760   0.762  0.44926
## SchoolStratford   31.9254    40.9171   0.780  0.43859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.13 on 55 degrees of freedom
## Multiple R-squared:  0.6425, Adjusted R-squared:  0.5125
## F-statistic: 4.942 on 20 and 55 DF,  p-value: 1.265e-06
```

The intercept value is negative which is not interpretable. So, the numeric predictor variables can be stardardized by subtracting its mean value and the intercept can be calculated again with the new values.

```
housedata$Size<-housedata$Size-mean(housedata$Size)
housedata$Lot<-housedata$Lot-mean(housedata$Lot)
housedata$Year<-housedata$Year-mean(housedata$Year)
Rmodel<-lm(Price~Size+Lot+Year+Bath+Bed+Garage+School,data = housedata)
summary(Rmodel)
```

```
##
## Call:
## lm(formula = Price ~ Size + Lot + Year + Bath + Bed + Garage +
##     School, data = housedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -87.601 -21.429   0.173  24.248  72.581
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      376.1016    51.7258   7.271 1.36e-09 ***
## Size              59.4503    28.9813   2.051  0.04501 *
## Lot               11.7701     3.7842   3.110  0.00296 **
## Year               0.5567     0.3384   1.645  0.10565
## Bath1.1          135.8983    49.1990   2.762  0.00779 **
## Bath2             73.9317    47.8636   1.545  0.12817
## Bath2.1           76.9433    48.1208   1.599  0.11556
## Bath3             98.0694    50.4663   1.943  0.05711 .
## Bath3.1           85.8037    54.3074   1.580  0.11985
## Bed3            -228.1052    70.6732  -3.228  0.00211 **
## Bed4            -238.2609    72.4883  -3.287  0.00177 **
## Bed5            -237.6155    76.4733  -3.107  0.00299 **
## Bed6            -255.0211    88.0955  -2.895  0.00543 **
## Garage1          -10.9191    22.4871  -0.486  0.62920
## Garage2           18.2435    18.2212   1.001  0.32111
## Garage3         -209.9038    80.7191  -2.600  0.01193 *
## SchoolHigh       113.2774    36.9154   3.069  0.00334 **
## SchoolNotreDame   80.9317    35.6893   2.268  0.02730 *
## SchoolStLouis      9.0367    37.3439   0.242  0.80969
## SchoolStMarys     27.3408    35.8760   0.762  0.44926
## SchoolStratford   31.9254    40.9171   0.780  0.43859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.13 on 55 degrees of freedom
## Multiple R-squared:  0.6425, Adjusted R-squared:  0.5125
## F-statistic: 4.942 on 20 and 55 DF,  p-value: 1.265e-06
```

The equation of this model is,

Price <-beta0 + beta1.Size + beta2.Lot + beta3.Year + beta4.Bath1.1 + beta5.Bath2 + beta6.Bath2.1 + beta7.Bath3 + beta8.Bath3.1 + beta9.Bed3 + beta10.Bed4 + beta11.Bed5 + beta12.Bed6 + beta13.Garage1 + beta14.Garage2 + beta15.Garage3 + beta16.SchoolHigh + beta17.SchoolNotreDame + beta18.SchoolStLouis + beta19.SchoolStMarys + beta20.SchoolStratford +error

(i.e) Price<- 376.1016 + 59.4503(Size) + 11.7701(Lot) + 0.5567(Year) + 135.8983(Bath1.1) + 73.9317(Bath2) + 76.9433(Bath2.1) + 98.0694(Bath3) + 85.8037(Bath3.1) + (-228.1052)(Bed3) + (-238.2609)(Bed4) + (-237.6155)(Bed5) + (-255.0211)(Bed6) + (-10.9191)(Garage1) + 18.2435(Garage2) + (-209.9038)(Garage3) + 113.2774(SchoolHigh) + 80.9317(SchoolNotreDame) + 9.0367(SchoolStLouis) + 27.3408(SchoolStMarys) + 31.9254(SchoolStratford)

2. Interpret the estimate of the intercept term Î²0.

The estimated intercept value Î²0 i.e average Price is 376.1016 taking the average values of Size,Lot,Year and the house with 1 bathroom,2 bedrooms,0 Garage which is near to Alex School. This estimate is significant with price as the p-value is less than 0.05 and the 95% confidence interval for this intercept is in the range (376.1016 (+,-)(1.96(51.7258))) ie [274.7190,477.4841].

3. Interpret the estimate of Î²size the parameter associated with ï¬□oor size (Size).

The estimate of Î²size is 59.4503 ie. for a given values of predictor variables lot, bath, bed, year, garage and school, a 1 unit increase in new size(after subtracting mean(size)) will increase the house price by 59.4503.This estimate is significant and the 95% confidence interval for this variable is in the range (59.4503 (+,-)(1.96(28.9813))) ie [2.6469,116.253648].

4. Interpret the estimate of Î²Bath1.1 the parameter associated with one and a half bathrooms.

The estimate of Î²Bath1.1 is 135.8983 ie. for a given values of predictor variables lot, size, bed, year, garage and school, a 1 unit increase in bath1.1 will increase the house price by 135.8983.This estimate is significant and the 95% confidence interval for this variable is in the range (135.8983 (+,-)(1.96(49.1990))) ie [39.4683,232.3284].

5. Discuss and interpret the eï¬□ect the predictor variable bed on the expected value of the house prices.

The predictor variable Bed is seperated into four co-efficients as Bed3,Bed4,Bed5 and Bed6.The interpretition of each of these is given below:

-The estimate of Bed3 is -228.1052 ie. for a given values of predictor variables lot, size, bath, year, garage and school, a 1 unit increase in bed3 will decrease the house price by 228.1052.This estimate is significant with price and the 95% confidence interval for this variable is in the range (-228.1052 (+,-)(1.96(70.6732))) ie [-366.6247,-89.5857].

-The estimate of Bed4 is -238.2609 ie. for a given values of predictor variables lot, size, bath, year, garage and school, a 1 unit increase in bed4 will decrease the house price by 238.2609.This estimate is significant with price and the 95% confidence interval for this variable is in the range (-238.2609 (+,-)(1.96(72.4883))) ie [-380.3379,-96.1838].

-The estimate of Bed5 is -237.6155 ie. for a given values of predictor variables lot, size, bath, year, garage and school, a 1 unit increase in bed5 will decrease the house price by 237.6155.This estimate is significant with price and the 95% confidence interval for this variable is in the range (-237.6155 (+,-)(1.96(76.4733))) ie [-387.5032,-87.7278].

-The estimate of Bed6 is -255.0211 ie. for a given values of predictor variables lot, size, bath, year, garage and school, a 1 unit increase in bed6 will decrease the house price by 255.0211.This estimate is significant with price and the 95% confidence interval for this variable is in the range (-255.0211 (+,-)(1.96(88.0955))) ie [-427.6883,-82.3539].

6. List the predictor variables that are signiï¬□cantly contributing to the expected value of the house prices

From the linear model data, we can see that all the predictor variables are significantly contributing to the expected value of the house price. These variables are listed as,
-Bath
-School
-Size
-Lot
-Garage
-Bed

7. For each predictor variable what is the value that will lead to the largest expected value of the house prices.

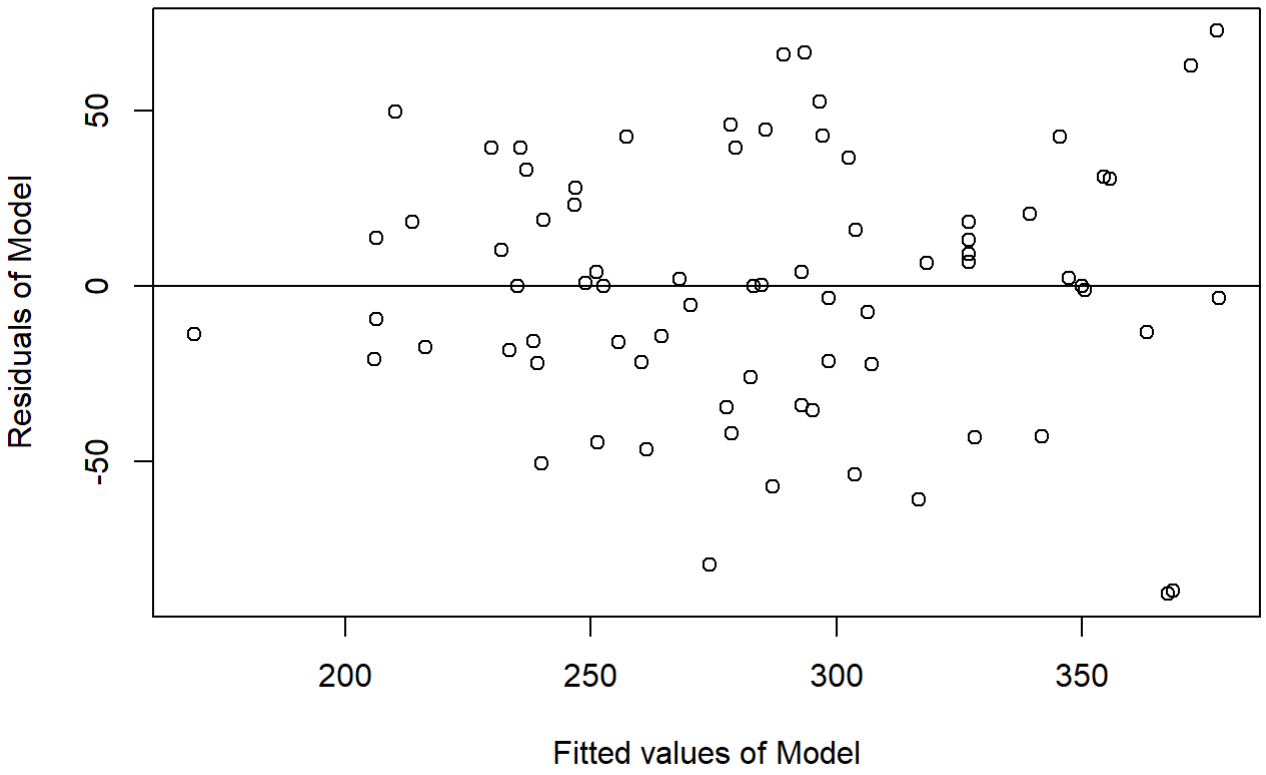
Based on the estimated values, the following values for each predictor variable gives the largest expected value,
Size - 2.896
Lot - 11
Bath - 1.1
Bed - 2
Garage - 2
School - High

8. For each predictor variable what is the value that will lead to the lowest expected value of the house prices.

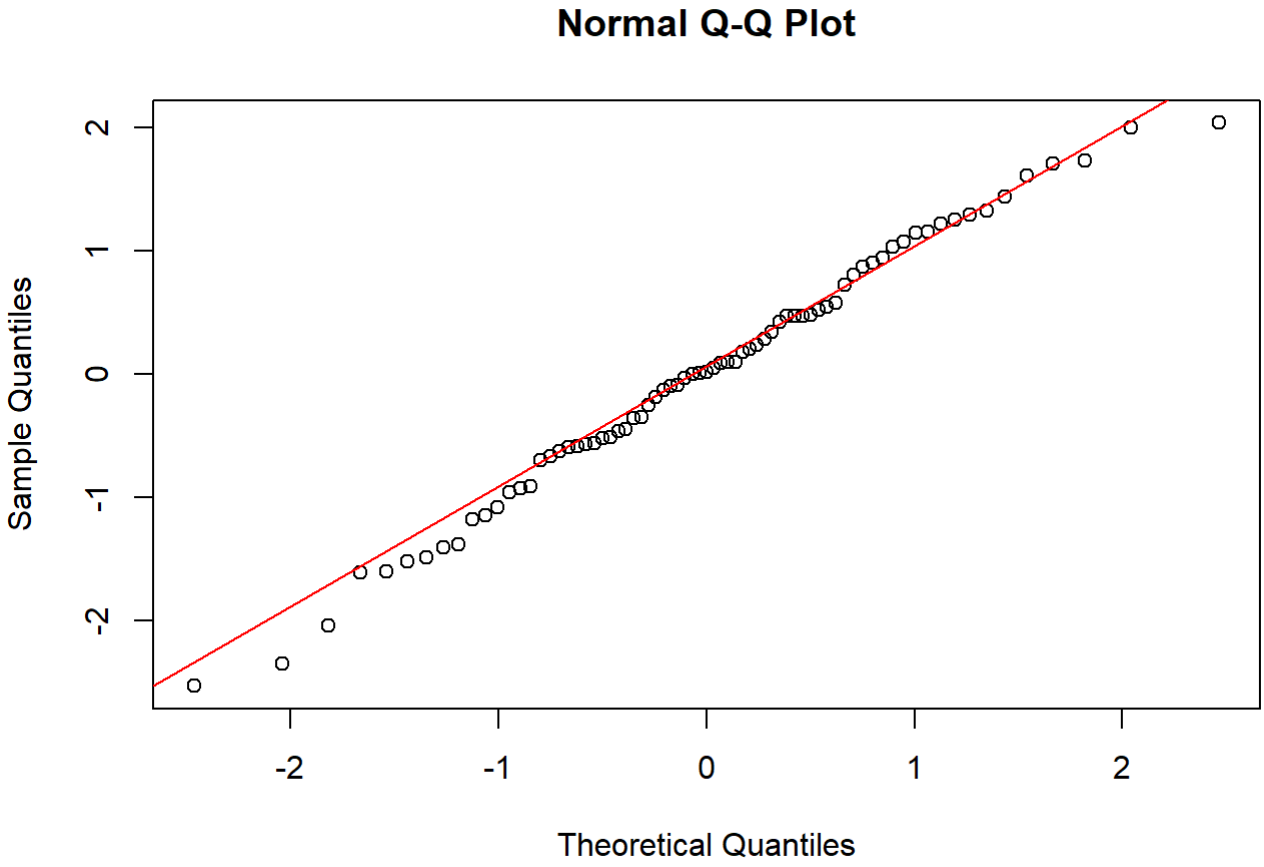
Based on the estimated values, the following values for each predictor variable gives the largest expected value,
Size - 1.44
Lot - 1
Bath - 1
Bed - 6
Garage - 3
School - Alex

9. By looking at the information about the residuals in the summary and by plotting the residuals do you think this is a good model of the expected value of the house prices.

```
plot(fitted(Rmodel),residuals(Rmodel),xlab = "Fitted values of Model",ylab = "Residuals of Model")
abline(h=0)
```




```
qqnorm(rstudent(Rmodel))
qqline(rstudent(Rmodel),col=2)
```




```
summary(Rmodel)
```

```
##
## Call:
## lm(formula = Price ~ Size + Lot + Year + Bath + Bed + Garage +
##     School, data = housedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -87.601 -21.429   0.173  24.248  72.581
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      376.1016    51.7258   7.271 1.36e-09 ***
## Size              59.4503    28.9813   2.051  0.04501 *
## Lot               11.7701     3.7842   3.110  0.00296 **
## Year               0.5567     0.3384   1.645  0.10565
## Bath1.1          135.8983    49.1990   2.762  0.00779 **
## Bath2             73.9317    47.8636   1.545  0.12817
## Bath2.1           76.9433    48.1208   1.599  0.11556
## Bath3             98.0694    50.4663   1.943  0.05711 .
## Bath3.1           85.8037    54.3074   1.580  0.11985
## Bed3            -228.1052    70.6732  -3.228  0.00211 **
## Bed4            -238.2609    72.4883  -3.287  0.00177 **
## Bed5            -237.6155    76.4733  -3.107  0.00299 **
## Bed6            -255.0211    88.0955  -2.895  0.00543 **
## Garage1          -10.9191    22.4871  -0.486  0.62920
## Garage2           18.2435    18.2212   1.001  0.32111
## Garage3         -209.9038    80.7191  -2.600  0.01193 *
## SchoolHigh       113.2774    36.9154   3.069  0.00334 **
## SchoolNotreDame   80.9317    35.6893   2.268  0.02730 *
## SchoolStLouis      9.0367    37.3439   0.242  0.80969
## SchoolStMarys     27.3408    35.8760   0.762  0.44926
## SchoolStratford   31.9254    40.9171   0.780  0.43859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.13 on 55 degrees of freedom
## Multiple R-squared:  0.6425, Adjusted R-squared:  0.5125
## F-statistic: 4.942 on 20 and 55 DF,  p-value: 1.265e-06
```

From the plot between Residuals and Fitted values of the model, it suggests that the values are symmetrically distributed and in a constant band.The qqplot also shows that the errors are normally arranged. Also, from the summary of residuals, the median is 0.173 which is close to zero and the 1st and 3rd quantiles are -21.429 & 24.248. So, this model is a good model for predicting the expected price of the house.

10. Interpret the Adjusted R-squared value.

The Adjusted R-squared value indicates how well the data points fits a curve or line. If addition of useful variables in the model improves the model, it will increase the R-squared value and decreases with addition of useless variables to the model. In this model, the Adjusted R-squared value is 0.5125. So,approximately half of the variations is explained by the model's input.

11. Interpret the F-statistic in the output in the summary of the regression model. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

The hypothesis for this model can be

Null Hypothesis,H0: All the coefficients are zero (ie) beta1=beta2=…=beta20=0
Alternate Hypothesis,Ha: Atleast one coefficient is not zero Beta[Size||Year||Lot||Bath||Bed||Garage||School]!=0

F-Statistics checks tests whether any of the independent variables in a model is significant in the prediction. In this model, the p-value is 1.265e-06 which is less than 0.05,hence we reject NUll Hypothesis. Thus this test suggests that any of the independent variables significantly improves the model.

ANOVA:

1. Compute the type 1 anova table. Interpret the output. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

Type 1 Analysis of Variance determines whether there is any statistically significant difference in the model by introducing new variables one by one to the model.

```
anova(Rmodel)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Size       1  11078 11077.7  6.2426  0.015489 *
## Lot        1  15232 15232.5  8.5839  0.004929 **
## Year       1   4741  4740.6  2.6715  0.107872
## Bath       5  37939  7587.9  4.2760  0.002345 **
## Bed        4  20200  5049.9  2.8458  0.032393 *
## Garage     3  16101  5367.1  3.0245  0.037179 *
## School     5  70112 14022.4  7.9020 1.153e-05 ***
## Residuals 55  97599  1774.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The initial hypothesis will be adding Size to the model.

Null Hypothesis H0:Beta[Size]=0;
Alternate Hypothesis Ha: Beta[Size]!=0;
The anova table shows that there is a significance by adding Size to the model. So it can be added to the model. Now, the second variable Lot will be added to the existing model and the table is compared to check if the new variable fits in the model. If the variable fits, then the next variable untill last variable is added to the existing model and compared. So, it works in a sequential way with one variable at a time.

From the anova table, it can be seen that Year predictor variable doesnot contribute to the model.

2. Which predictor variable does the type 1 anova table suggest you should remove the regression analysis.

From the anova table, it can be seen that Year predictor variable doesnot contribute to the model.It can be removed from the model.

3. Compute a type 2 anova table comparing the full model with all predictor variables to the the reduced model with the suggested predictor variable identiï¬□ed in the previous question removed. Hint: State the hypothesis being tested, the test statistic and p-value and the conclusion in the context of the problem.

```
Rmodelnew<- lm(Price~Size+Lot+Bath+Bed+Garage+School,data = housedata)
anova(Rmodel,Rmodelnew)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Size + Lot + Year + Bath + Bed + Garage + School
## Model 2: Price ~ Size + Lot + Bath + Bed + Garage + School
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     55  97599
## 2     56 102402 -1   -4802.6 2.7064 0.1057
```

Null Hypothesis H0:Beta[Year]=0;
Alternate Hypothesis Ha: Beta[Year]!=0;
The test statistic is the F-Value. From the table above, we see that the p-value is 0.1057 which is greater than 0.05.Thus we fail to reject the Null Hypothesis. This proves that predictor variable “Year” doesnot contribute to the model and it can be removed (ie) beta[Year]=0.

Diagnostics:

1. Check the linearity assumption by interpreting the added variable plots and component-plus-residual plots. What eï¬□ect would non-linearity have on the regression model and how might you correct or improve the model in the presence of non-linearity?

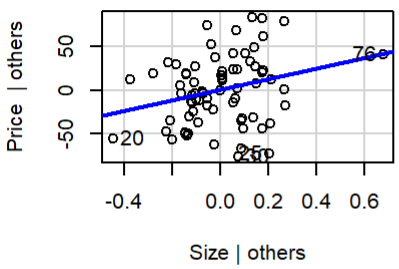
```
library(car)
```

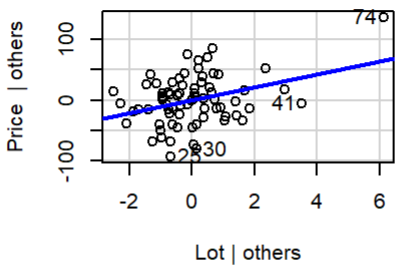
```
## Loading required package: carData
```

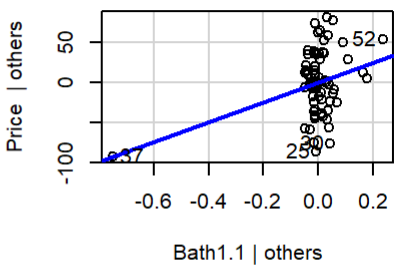
```
##
## Attaching package: 'car'
```

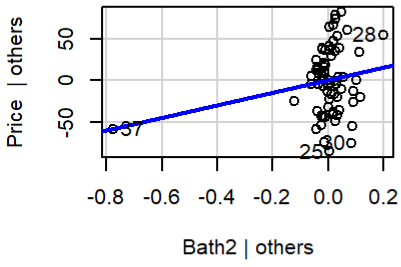
```
## The following object is masked from 'package:psych':
##
##     logit
```

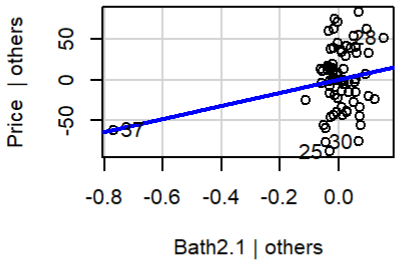
```
avPlots(Rmodelnew)
```

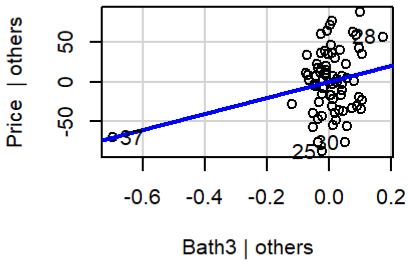


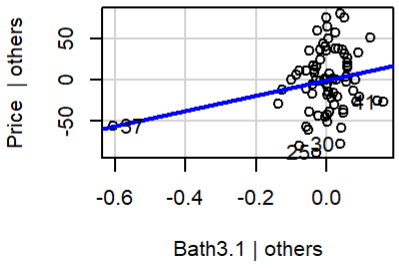


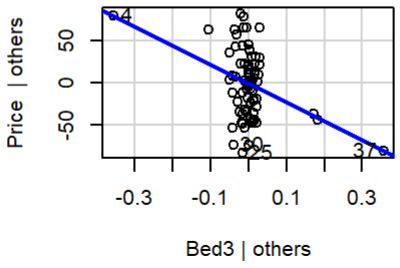


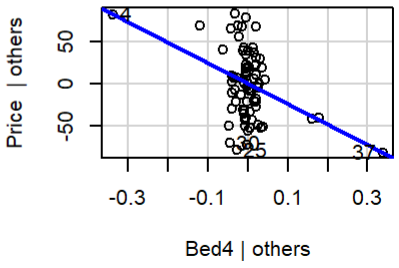


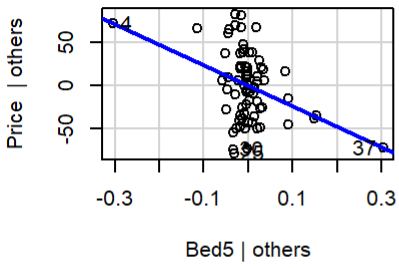


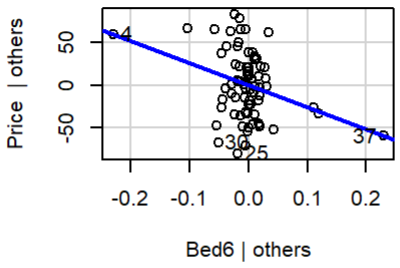


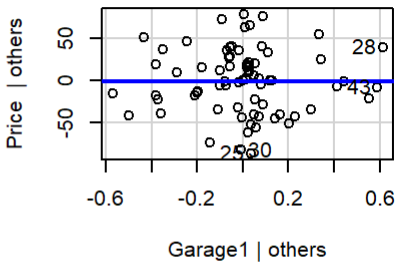


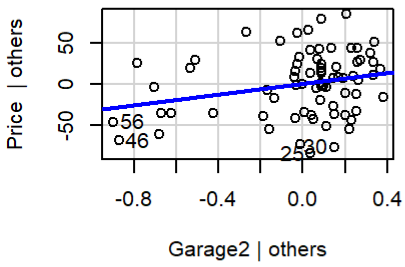


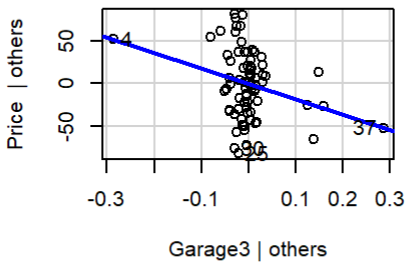


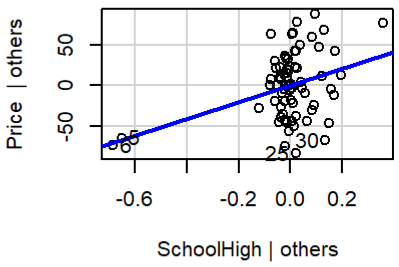


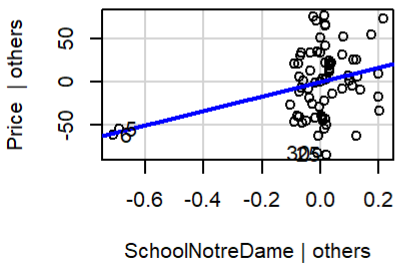


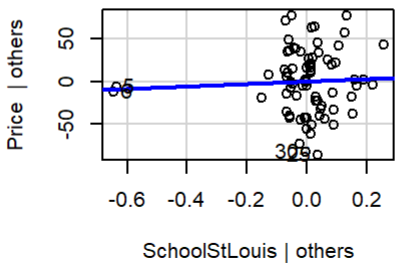


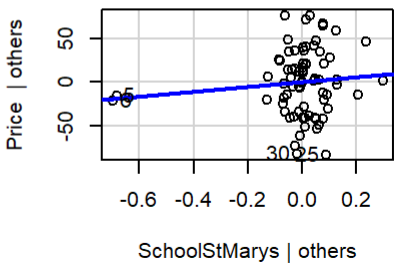




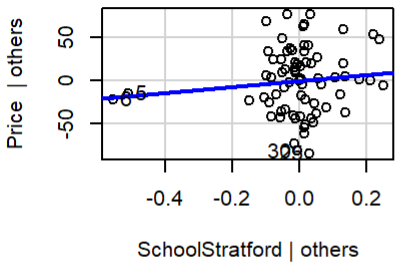




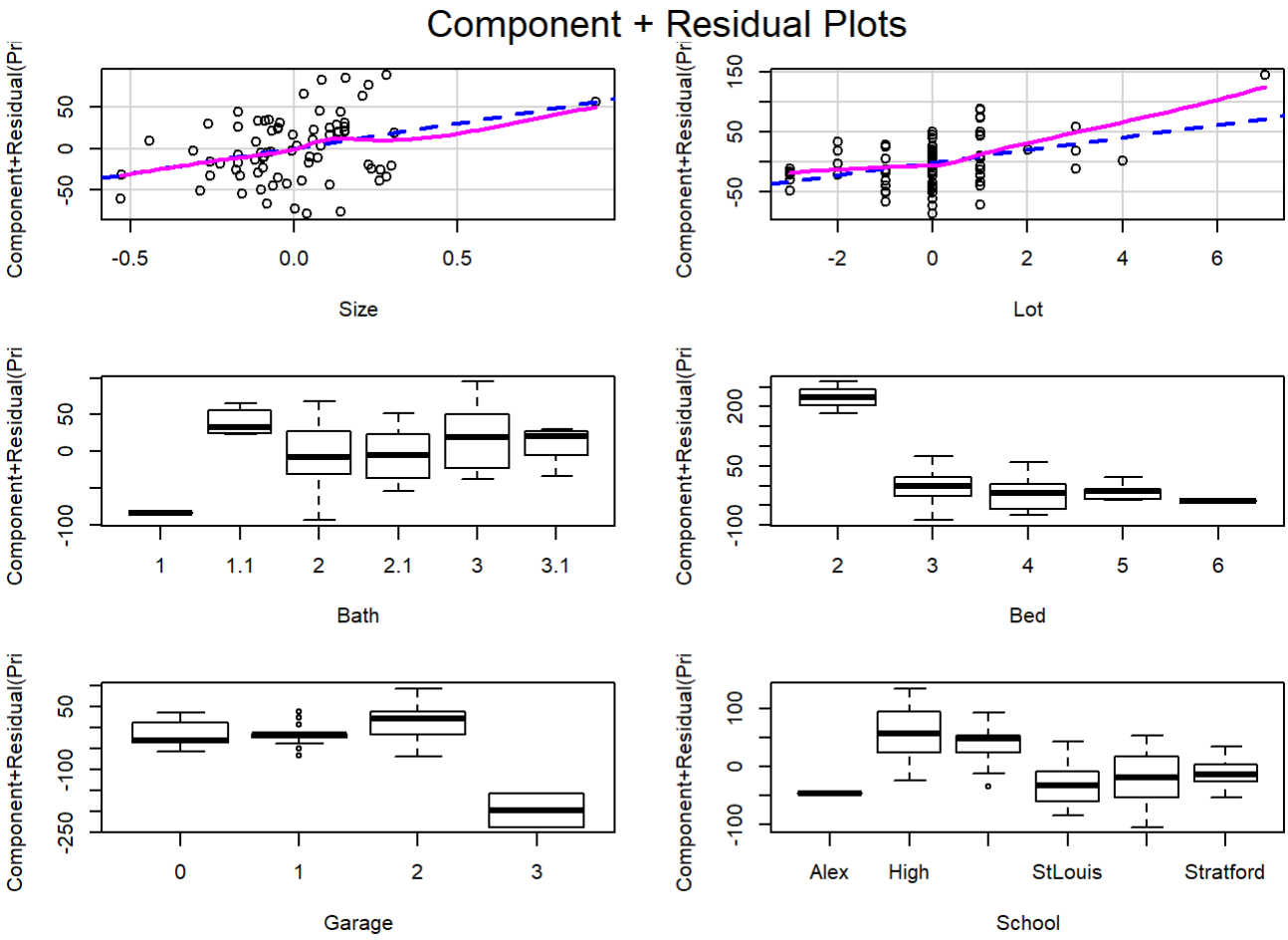




Added-Variable Plots



```
crPlots(Rmodelnew)
```



The Added variable plot(avPlots) depicts a pictorial format of price with different coefficient variables seperately. Each plot suggest a linear relationship between the response variable Price and each of the predictor variables. All the predictor variabes plots have a positive linear regression except for Bed coefficients and Garbage3 coefficients.Bed and Garbage3 coefficients has a negative linear relationship with the variable Price.

The Component+Residuals plots describes the Linear regression and the smoothness curve for the relationship between the response variable Sale price of the house and each of the predictor variable.
For numeric variable, there is no difference between residual(pink line) and component(blue dashed) as both the lines coincides with each other.
For categorical variables, the boxplot shows the significant values for each variable such as bath1.1 is highly significant for Bath, High school is highly significant for school. Hence, the linearity assumption holds the model.

The effect of non-linearity on the regression model would lead to inconsistence or biased esitmates.In the presence of non-linearity we can improve the model by trying to fit the data using various polynomials or splines. Depending on the data, these two methods may provide similar fits.

2. Check the random/i.i.d. sample assumption by carefully reading the data description and computing the Durbin Watson test (state the hypothesis of the test, the test statistic and p-value and the conclusion in the context of the problem). What are the two common violations of the random/i.i.d. sample assumption? What eï¬□ect would dependant samples have on the regression model and how might you correct or improve the model in the presence of dependant samples?

```
dwt(Rmodelnew)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2316982      1.511734   0.012
##  Alternative hypothesis: rho != 0
```

The Hypothesis for this test is, Null Hypothesis: No autocorrelation rho=0
Alternate Hypothesis: rho!=0
From the Durbin Watson Test, the D-W Statistic value is 1.51 and its p-value is 0.006. This p-value is less than 0.005 which results in rejecting the hypothesis of No autocorrelation and confirms that observations cannot be classified as independent.
Dwt result shows 0.231 as autocorrelation.So, rejection of Null Hypothesis can be due to outliers. So, the outliers have to be removed for a better model and hypothesis should be checked again.

The two common violations of the random/i.i.d. sample assumption can be
-Observations are separated to various groups
-Repeated Measures

The effect of dependant samples can be Heteroskedasticity,Outlier from different group can lead to biased result.
The model can be improved using Time series Analysis,Mixed Effect Models.

3. Check the collinearity assumption by interpreting the correlation and variance inï¬□ation factors. What eï¬□ect would multicollinearity have on the regression model and how might you correct or improve the model in the presence of multicollinearity.

```
vif(Rmodelnew)
```

```
##             GVIF Df GVIF^(1/(2*Df))
## Size    1.599498  1        1.264713
## Lot     1.586836  1        1.259697
## Bath    8.302874  5        1.235728
## Bed    17.585637  4        1.431017
## Garage 15.567849  3        1.580173
## School  5.342986  5        1.182438
```

The Variance Inflation Factor tests the multicollinearity in the model.its has the values that are approximately 1.So, there is no correlation between the predictor variables ie no Multicollinearity in the model.

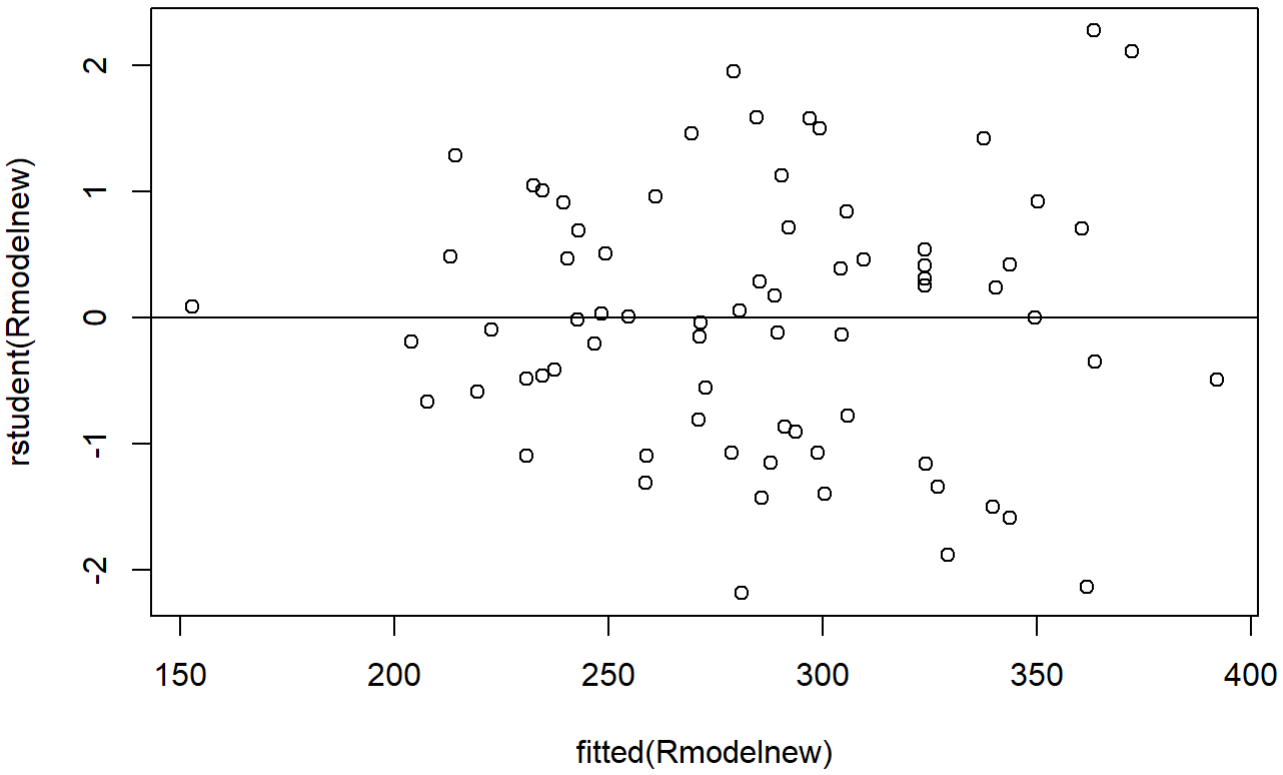
Presence of Multicollinerity can result in coefficients estimates to be unstable and difficult to interpret the model.

The model with multicollinerity can be improved by
-Using Partial Least Squares Regression(PLS),Principal Components Analysis,regression methods that seperates the number of predictors to small set of uncorrelated components.
-Removing highly correlated predictors from the model.

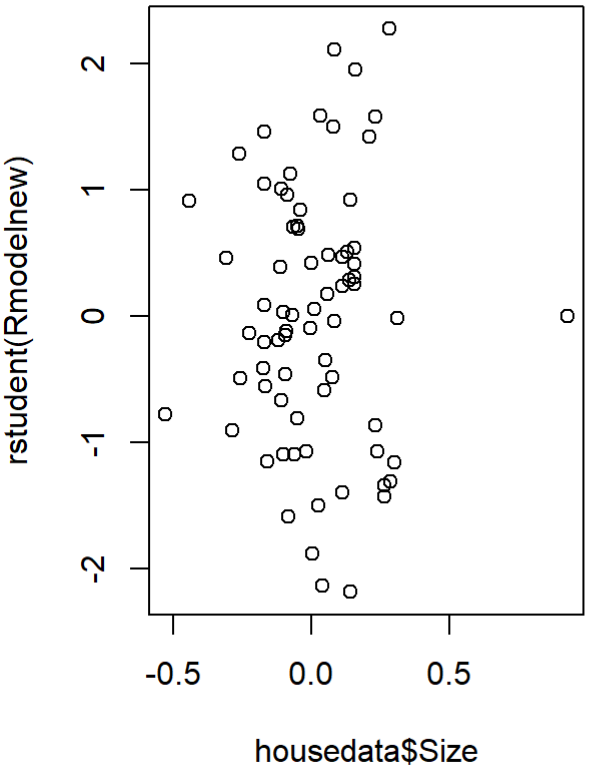
4. Check the zero conditional mean and homoscedasticity assumption by interpreting the studentized residuals vrs ï¬□tted values plots and the studentized residuals vrs predictor

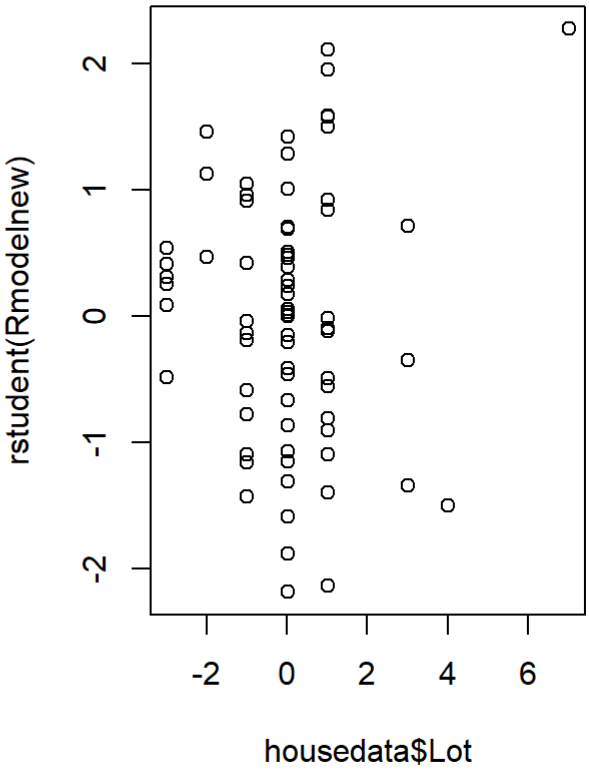
variable plots. What eï¬□ect would heteroscedasticity have on the regression model and how might you correct or improve the model in the presence of heteroscedasticity.

```
plot(fitted(Rmodelnew),rstudent(Rmodelnew))
abline(h=0)
```




```
par(mfrow=c(1,2))
plot(housedata$Size,rstudent(Rmodelnew))
plot(housedata$Lot,rstudent(Rmodelnew))
```





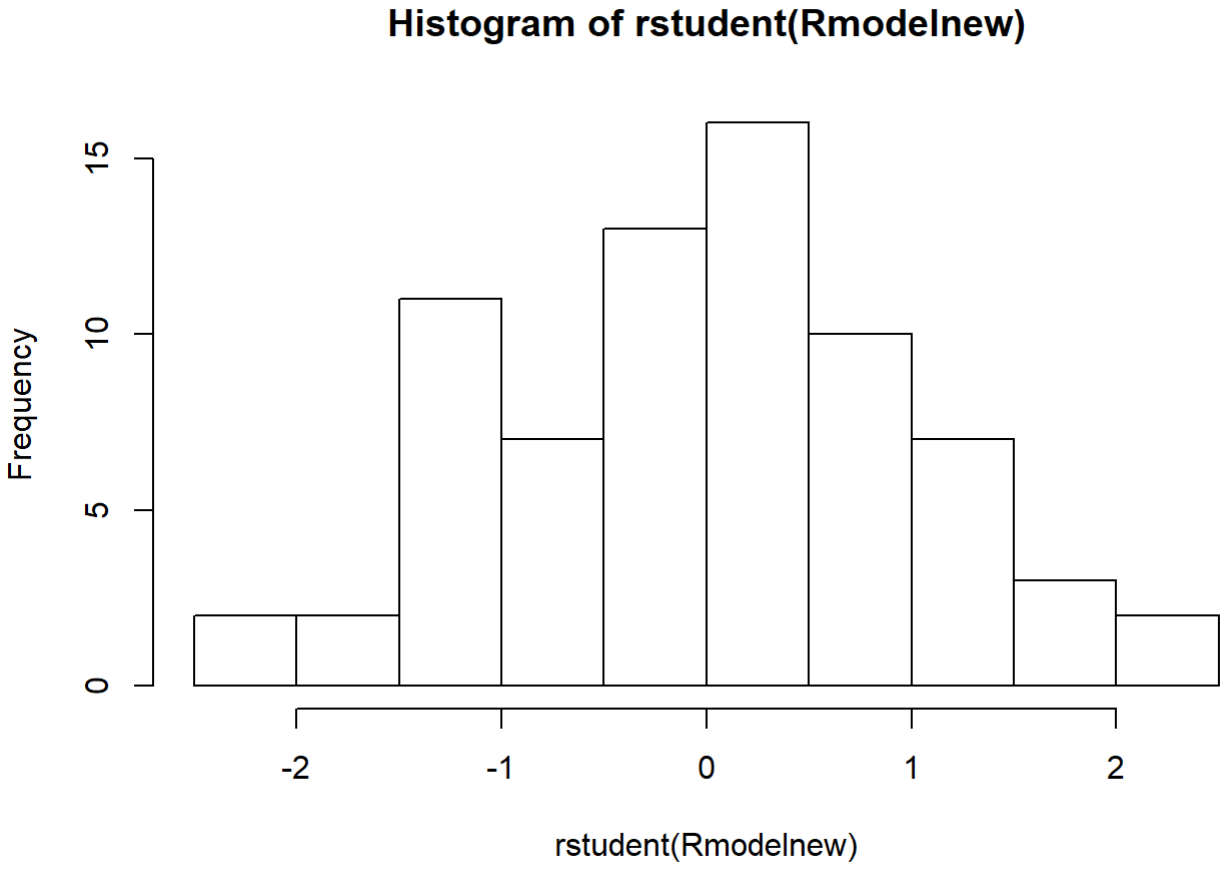
The plot of fitted model values and studentized residuals suggests that the values are uncorrelated as they should be in a homoscedastic linear model. The second plot shows that there is no funnel shapped pattern with the values. So, the model holds the assumptions on zero conditional mean and homoscedasticity.

Effects of heteroscedasticity are
-Predictions are no longer efficient
-The T-test and F-test are not valid due to inconsistency in coefficients.

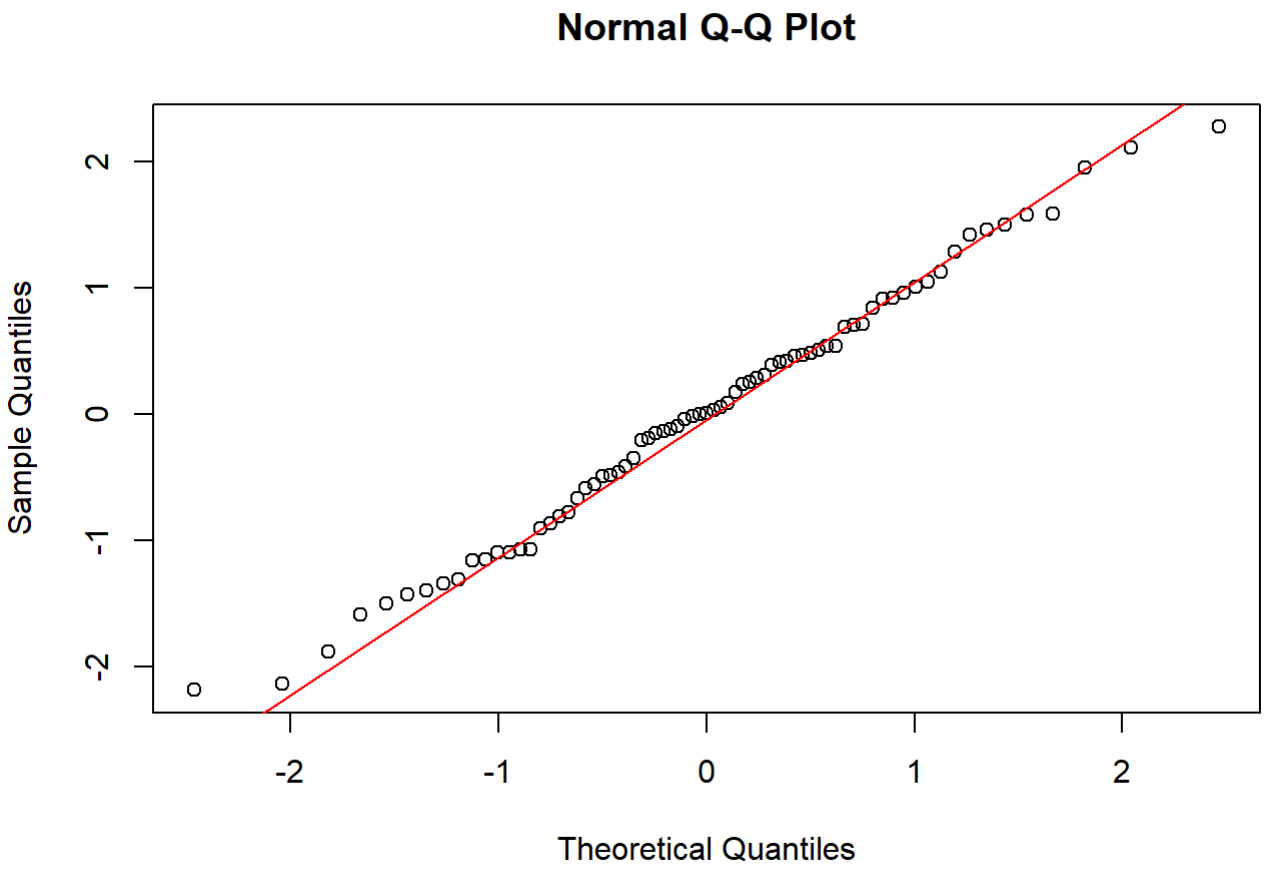
Heteroscedasticity can be corrected using Weighted Least Squares method.

5. Check the Normality assumption by interpreting the histogram and quantilequantile plot of the studentized residuals. What eï¬□ect would non-normality have on the regression model and how might you correct or improve the model in the presence of non-normality.

```
hist(rstudent(Rmodelnew),breaks = 10)
```




```
qqnorm(rstudent(Rmodelnew))
qqline(rstudent(Rmodelnew),col=2)
```



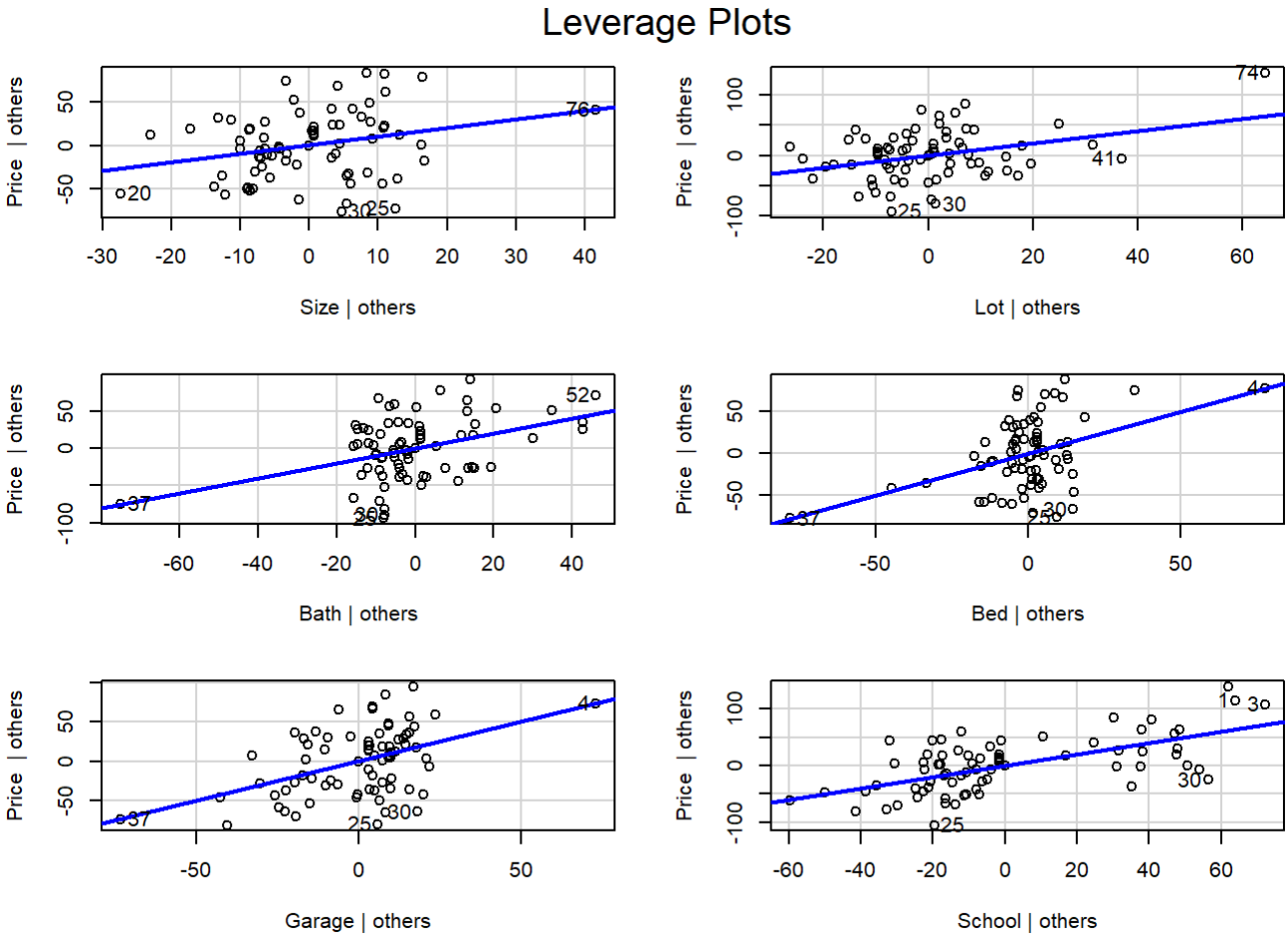
The plots suggests that the values are normally distributed from the approximately straight line. Hence the model holds Normality assumption. The effect of Non-normality is that the critical values of F-test and T-test may be misleading.To correct Non-Normality, the data can be transformed and tested again.

Leverage, Inï¬□uence and Outliers:

1. What is a leverage point? What eï¬□ect would a leverage point have on the regression model? Use the leverage values and the leverage plots to see if there is any leverage points.

The Leverage point is a measure of how far an observation(X-value) on the predictor variable from the mean of the predictor variable.This value can affect the summary of the model such as the R-squared value but has less impact on the estimates of the coefficients.The higher the leverage point of an observation, the more potential it has to impact the fitted model.

```
leveragepoints<-as.numeric(which(hatvalues(Rmodelnew)>((2*20)/length(housedata$Price))))
leveragePlots(Rmodelnew)
```




```
leveragepoints
```

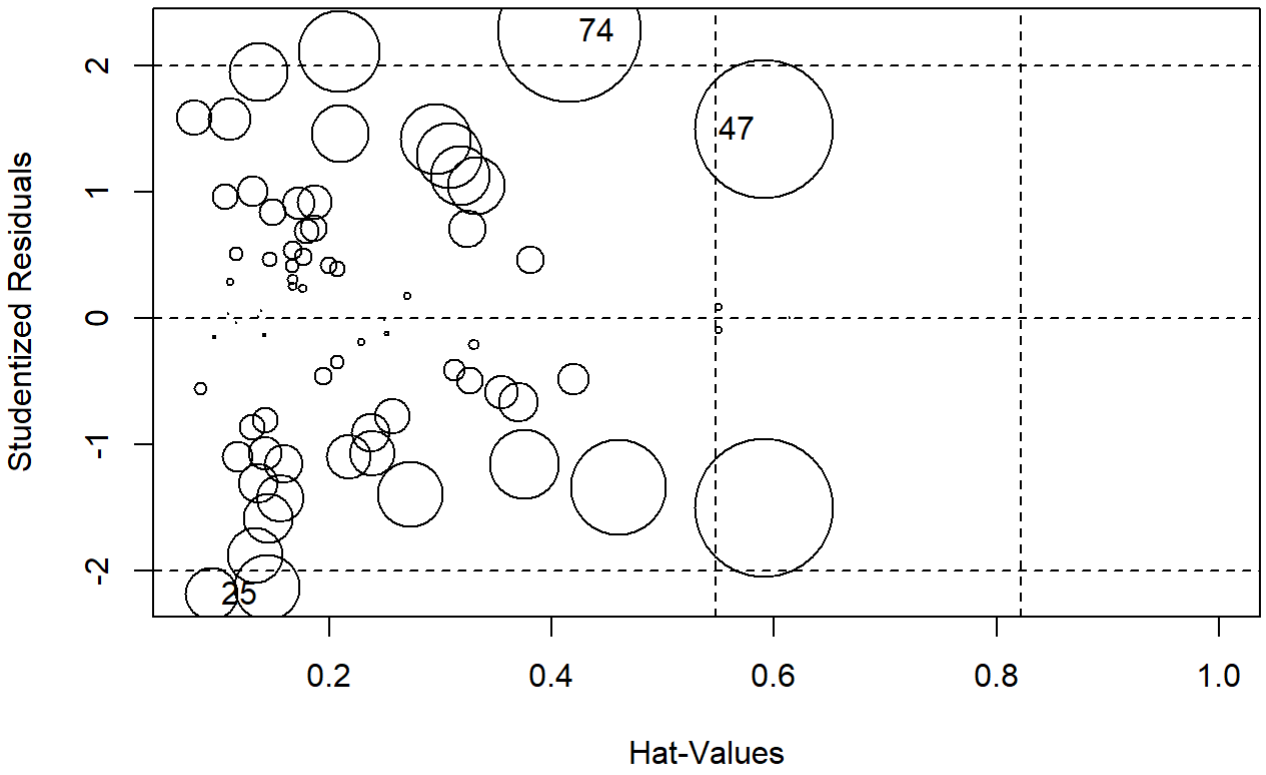
```
## [1]  4  5  6 21 35 37 47 76
```

The leverage points mentioned in the plots are,
Size: 20,25,30,36
Lot: 25,30,41,76
Bath: 25,30,37,52
Bed: 4,25,30,37
Garage: 4,25,30,37
School:1,3,25,30

2. What is an inï¬□uential point? What eï¬□ect would an inï¬□uential point have on the regression model? Use the inï¬□uence plot to see if there is any inï¬□uence points.

An influential point is an outlier(Y-value) that greatly impacts the slope of the regression line.

```
influencePlot(Rmodelnew)
```




```
##      StudRes        Hat      CookD
## 4        NaN 1.00000000        NaN
## 25 -2.183451 0.09446984 0.02330078
## 35       NaN 1.00000000        NaN
## 47  1.499325 0.59085455 0.15877875
## 74  2.276139 0.41584296 0.17159204
```

The influential points for this model are 4,25,35,47,74.

3. What is an outlier? What eï¬□ect would an outlier have on the regression model? How would you correct for outliers? Use the outlier test and outlier and leverage diagnostics plot to see if there is any outliers. Deal with the outliers if any are identiï¬□ed.

An outlier in regression model are the observations that fall far from the values of a fitted model.Outliers have a strong effect on the least squares line.This will lead to poor fitting of the regression model. Best method of correcting the outliers is to remove the obervations and refit the model again.

```
outlierTest(Rmodelnew)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 74 2.276139           0.026753           NA
```

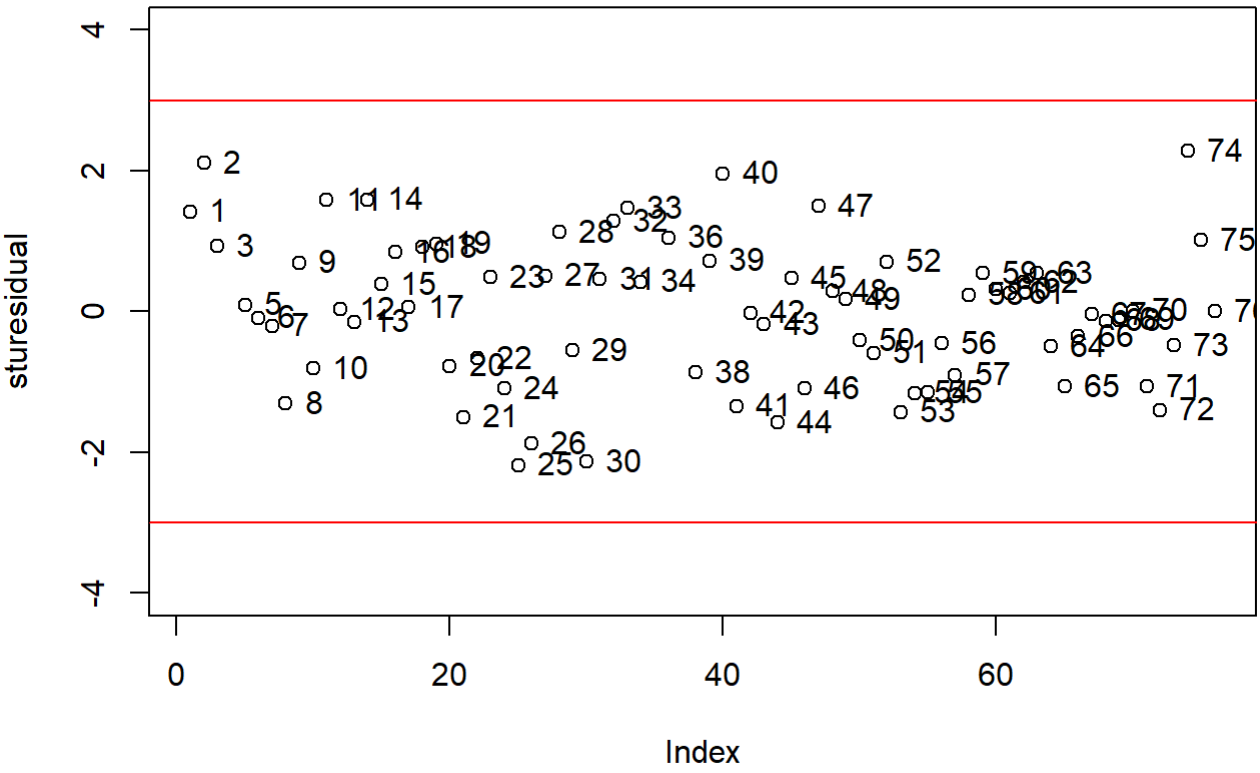
```
housedata[74,]
```

```
##    Price      Size      Lot Bath Bed     Year Garage  School
## 74   435 0.2826053 7.013158    2   3 9.592105      2 StMarys
```

```
sturesidual=rstudent(Rmodelnew)
sturesidual[74]
```

```
##       74
## 2.276139
```

```
plot(sturesidual,ylim = c(-4,4))
abline(h=3,col="red")
abline(h=-3,col="red")
text(sturesidual,labels = names(sturesidual),pos=4)
```

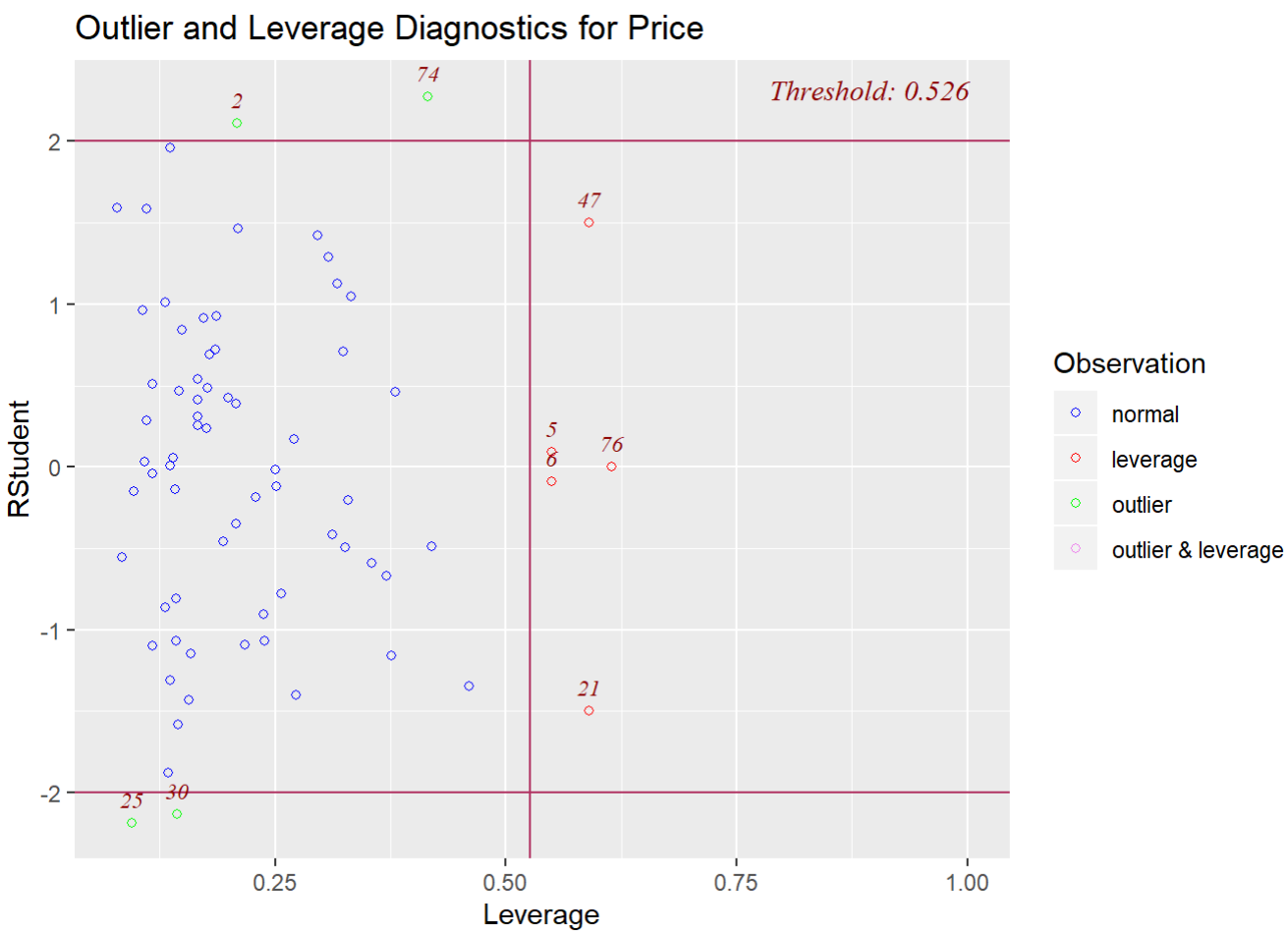



```
library(olsrr)
```

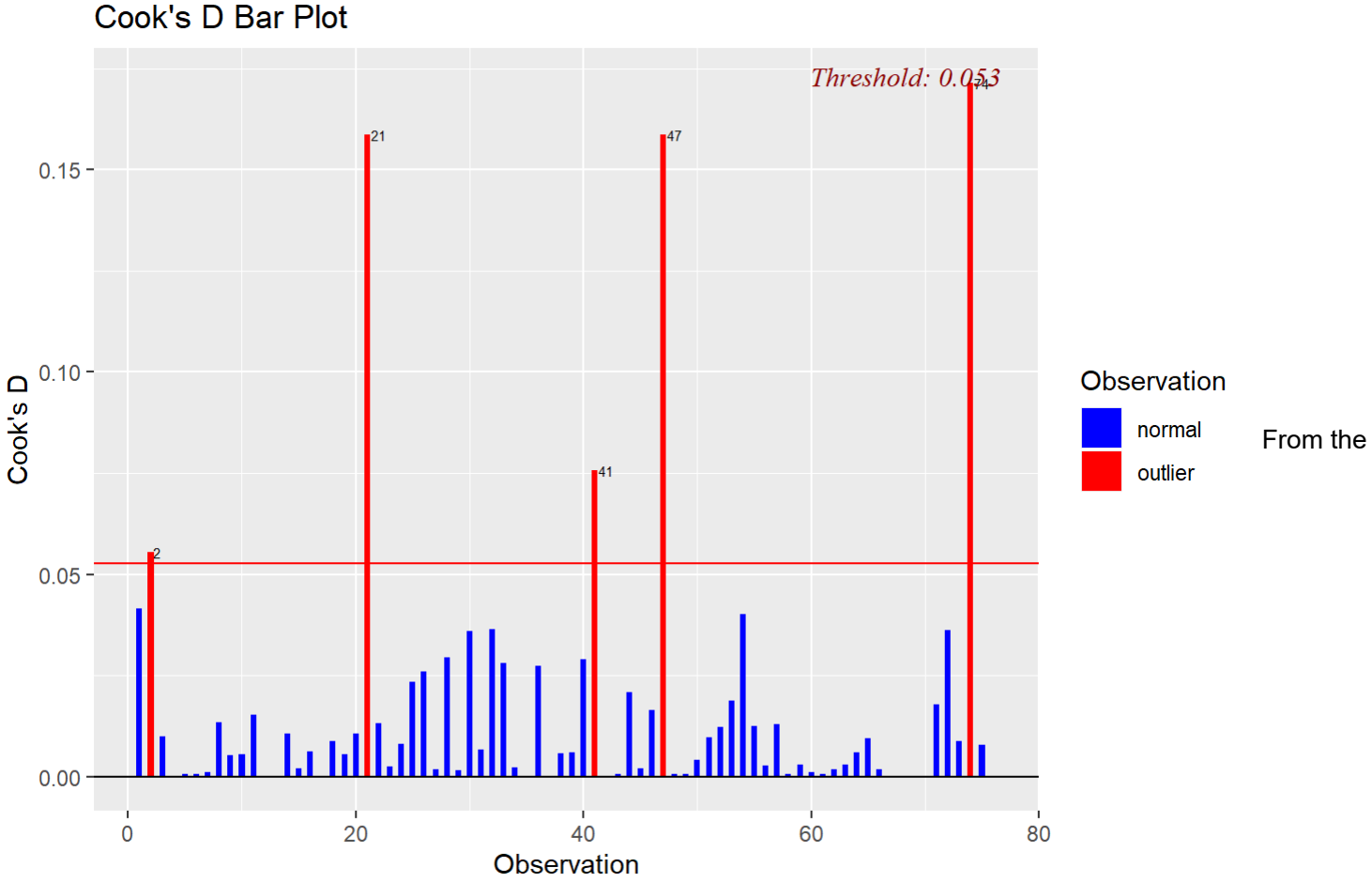
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
ols_plot_resid_lev(Rmodelnew)
```




```
ols_plot_cooksd_bar(Rmodelnew)
```



From the outlier test, there are four outliers in the data. They are 2,25,30,74. These values are removed and model is refitted.

```
housedatanew<-housedata[-c(2,25,30,74),]
Rmodelnew1<- lm(Price~Size+Lot+Bath+Bed+Garage+School,data=housedatanew)
summary(Rmodelnew1)
```

```
##
## Call:
## lm(formula = Price ~ Size + Lot + Bath + Bed + Garage + School,
##     data = housedatanew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.538 -24.489   0.602  19.567  70.107
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      373.762     45.368   8.238 5.36e-11 ***
## Size              67.082     26.470   2.534 0.014318 *
## Lot                4.724      4.036   1.170 0.247177
## Bath1.1          116.868     43.505   2.686 0.009677 **
## Bath2             76.213     42.867   1.778 0.081269 .
## Bath2.1           76.338     43.066   1.773 0.082155 .
## Bath3             90.863     45.322   2.005 0.050200 .
## Bath3.1           79.021     48.452   1.631 0.108950
## Bed3            -241.156     63.060  -3.824 0.000353 ***
## Bed4            -266.523     64.678  -4.121 0.000136 ***
## Bed5            -258.138     68.243  -3.783 0.000402 ***
## Bed6            -295.330     79.100  -3.734 0.000469 ***
## Garage1            2.173     19.251   0.113 0.910575
## Garage2           35.400     14.651   2.416 0.019230 *
## Garage3         -183.283     70.369  -2.605 0.011965 *
## SchoolHigh       112.768     32.884   3.429 0.001192 **
## SchoolNotreDame   87.015     31.666   2.748 0.008224 **
## SchoolStLouis     25.248     33.214   0.760 0.450597
## SchoolStMarys     35.082     32.023   1.096 0.278326
## SchoolStratford   49.856     36.604   1.362 0.179056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.39 on 52 degrees of freedom
## Multiple R-squared:  0.6616, Adjusted R-squared:  0.5379
## F-statistic:  5.35 on 19 and 52 DF,  p-value: 7.479e-07
```

After removing the outliers, the Adjusted R-squared value increased from 0.51 to 0.54 ie to approximately 54%. So this is the best model and model summary is increased.

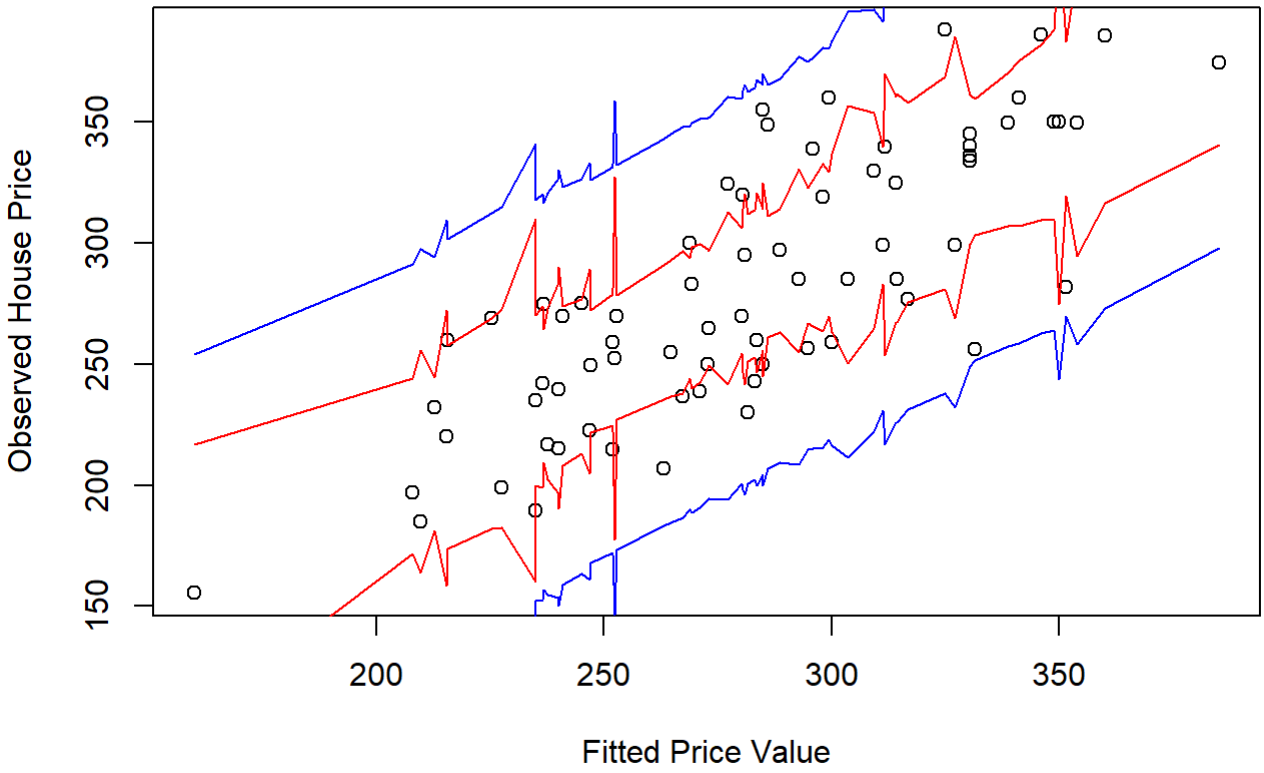
Expected Value, CI and PI:

1. Plot the observed house prices, their expected vale (ï¬□tted value), conï¬□dence intervals (in red) and prediction intervals (in blue). Looking at this plot is this model providing a good estimate of the house prices.

```
pred<-predict(Rmodelnew1,interval="predict")
```

```
## Warning in predict.lm(Rmodelnew1, interval = "predict"): predictions on current data refer to _futur
e_ responses
```

```
colnames(pred)<-c('p_fit','p_lwr','p_upr')
conf<-predict(Rmodelnew1,interval="confidence")
colnames(conf)<-c('c_fit','c_lwr','c_upr')
new_data<-cbind(housedatanew,pred,conf)
new_data<-new_data[order(new_data$p_fit),]
plot(c(new_data$p_fit),c(new_data$Price),xlab = "Fitted Price Value", ylab = "Observed House Price")
lines(c(new_data$p_fit),c(new_data$p_lwr),col="blue")
lines(c(new_data$p_fit),c(new_data$p_upr),col="blue")
lines(c(new_data$p_fit),c(new_data$c_lwr),col="red")
lines(c(new_data$p_fit),c(new_data$c_upr),col="red")
```



The generated plot proves that the predicted model provides a good estimate of the response variable Price.