PAA

Shubhang Periwal 19201104

11/13/2019

R Markdown

```r
library(car)
```

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

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(olsrr)
```

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

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

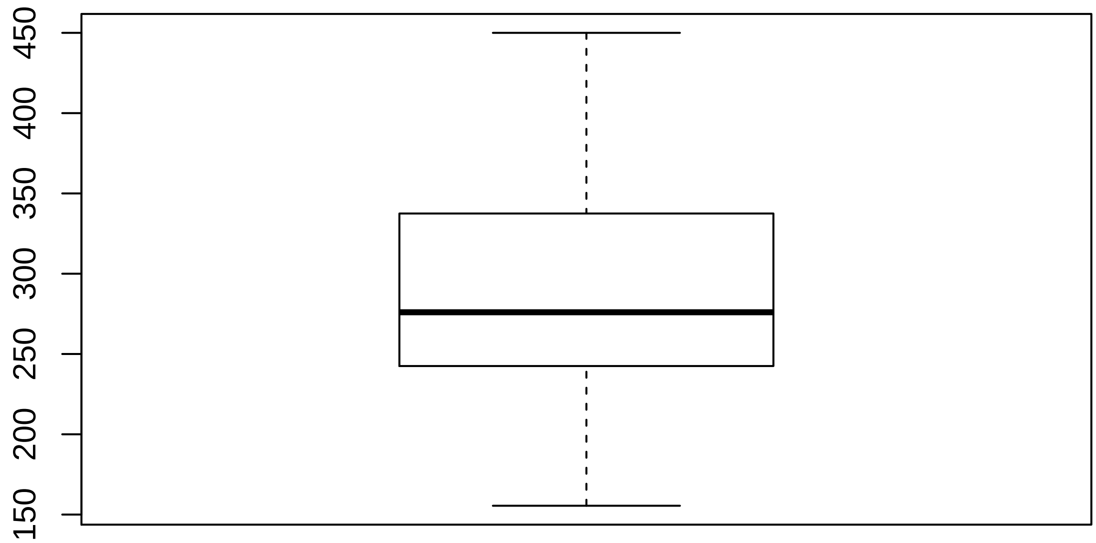
Libraries required to plot and do other things required in the assignment

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

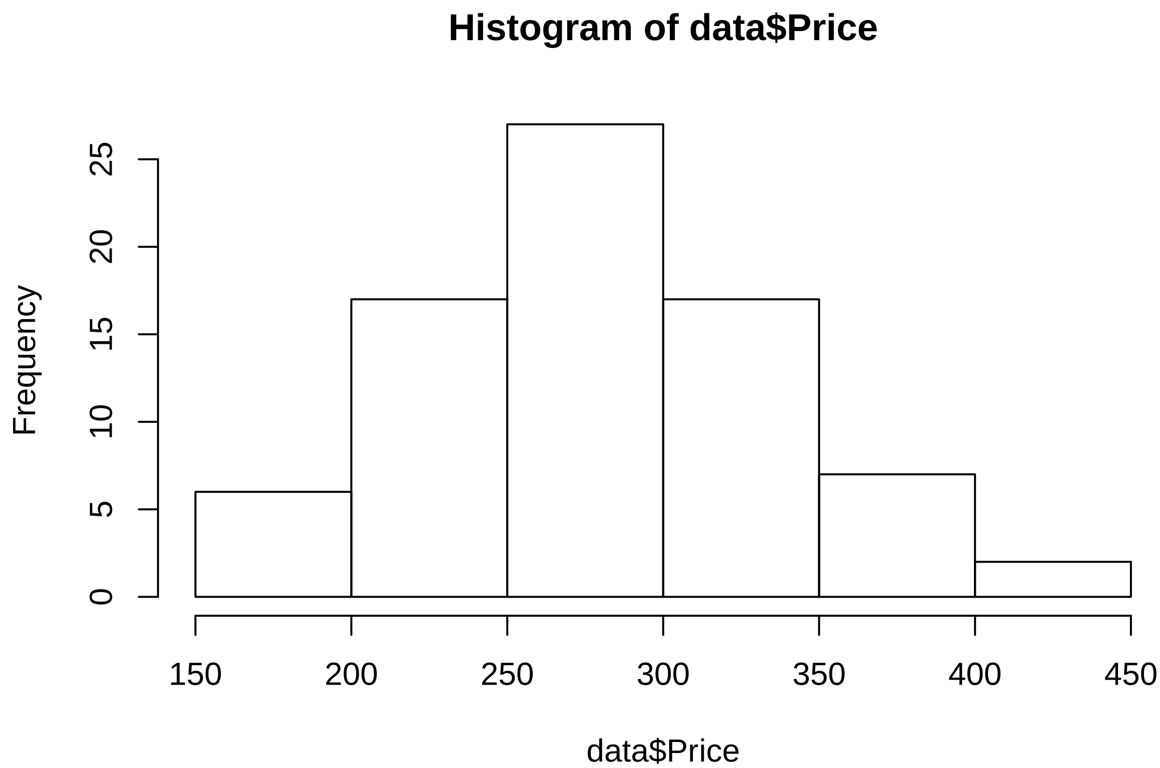
```r
#Exploratory Data Analysis q1
data = read.csv("House.csv",header = TRUE)
summary(data$Price)
```

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

```r
boxplot(data$Price)
```




```r
hist(data$Price)
```



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

From the boxplot we can see the distribution of prices with the frequency of the numbet of houses. We can see that 50% of the houses are distributed around 250k and 340k with median close to 280. The cost of houses start at aroun 150k and is there till 450k.

From the histogram we can see that the data is normally distributed with most number of houses priced between 250k and 300k.

From the summary we can see that the price minimum price of the house is 155.5k. 25% of the houses cost under 242.8k(1st quartile). 50% of the houses cost under 276k. 75%of the houses cost under 336.8k and the maximum cost of a house is 450k. Summary also specifies the mean which is the average cost of all houses with the value of 285.8k

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.

```
#Exploratory Data Analysis q2
data$School=factor(data$School)
data$Bath = factor(data$Bath,levels = c(1,1.1,2,2.1,3,3.1),labels=c(1,1.1,2,2.1,3,3.1))
data$Bed = factor(data$Bed, levels = c(2,3,4,5,6),labels=c(2,3,4,5,6))
data$Lot = factor(data$Lot)
by(data$Price,data$Bath,summary)
```

```
## data$Bath: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   235.0   263.8   292.5   292.5   321.2   350.0
```

```
## ----------------------------------------------------------
## data$Bath: 1.1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   215.0   239.5   325.0   307.9   374.5   385.5
## ----------------------------------------------------------
## data$Bath: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   220.0   259.0   270.7   319.0   435.0
## ----------------------------------------------------------
## data$Bath: 2.1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   189.5   254.8   269.9   274.5   297.7   349.5
## ----------------------------------------------------------
## data$Bath: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   230.0   259.0   295.0   307.8   349.5   450.0
## ----------------------------------------------------------
## data$Bath: 3.1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   285.0   309.4   336.0   324.2   342.5   345.0
```

```r
by(data$Price,data$Bed,summary)
```

```
## data$Bed: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   299.0   319.4   339.9   329.6   344.9   350.0
## ----------------------------------------------------------
## data$Bed: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   189.5   256.2   297.0   297.3   342.5   435.0
## ----------------------------------------------------------
## data$Bed: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   231.5   254.4   266.6   283.5   450.0
## ----------------------------------------------------------
## data$Bed: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   185.0   199.0   269.0   259.5   295.0   349.5
## ----------------------------------------------------------
## data$Bed: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   252.5   252.5   252.5   252.5   252.5   252.5
```

```r
by(data$Price,data$Garage,summary)
```

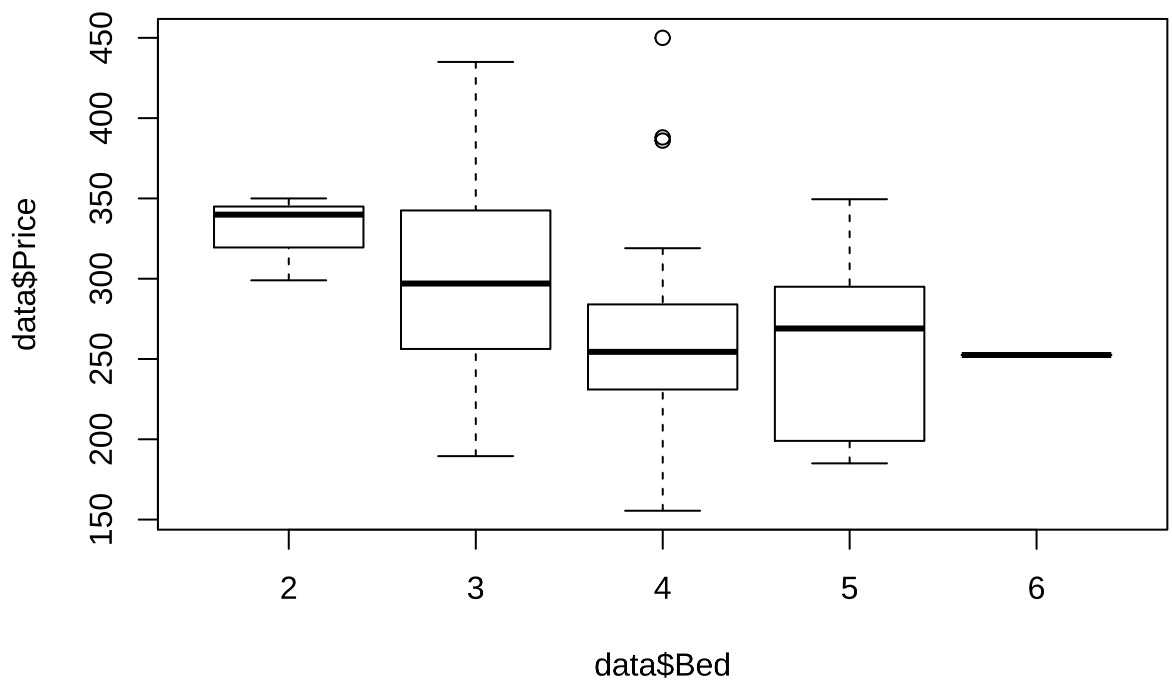
```
## data$Garage: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   185.0   216.0   232.0   246.9   264.4   388.0
## ----------------------------------------------------------
## data$Garage: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   220.0   242.0   260.6   324.5   385.5
## ----------------------------------------------------------
```

```
## data$Garage: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   195.0   259.0   285.0   299.6   343.8   450.0
## --------------------------------------------------------
## data$Garage: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   299.0   309.2   319.4   319.4   329.7   339.9
```

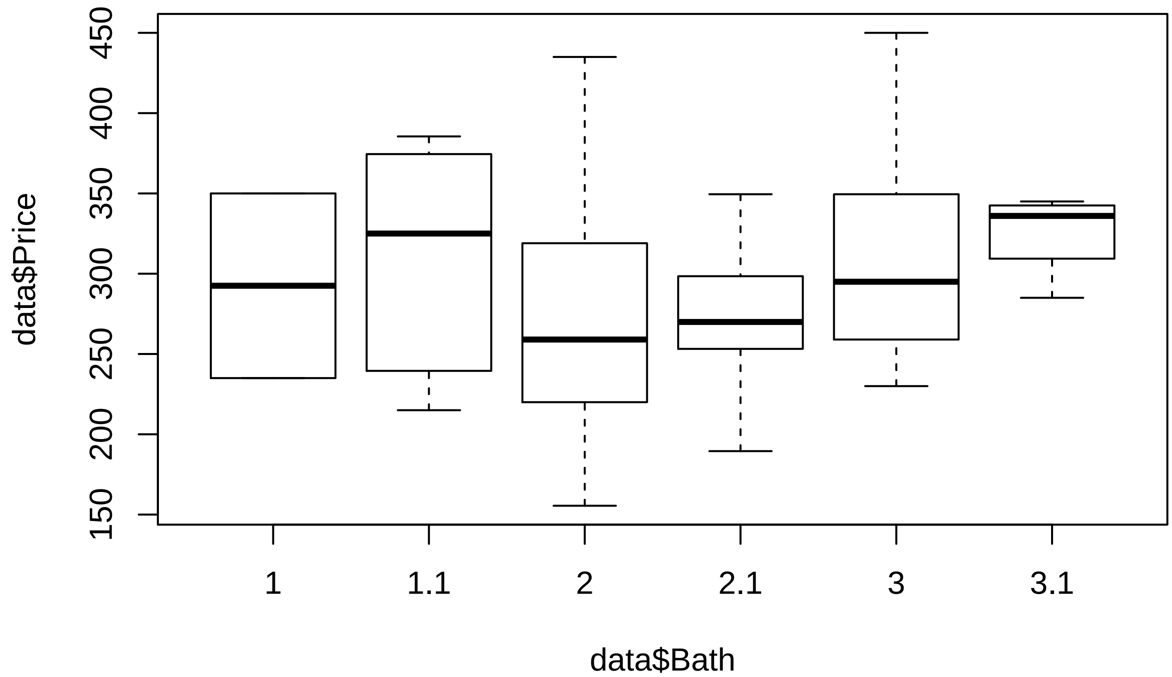
```
by(data$Price,data$School,summary)
```

```
## data$School: Alex
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   187.8   220.0   241.8   285.0   350.0
## --------------------------------------------------------
## data$School: High
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   235.0   279.2   307.5   327.1   385.6   450.0
## --------------------------------------------------------
## data$School: NotreDame
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   249.9   304.0   334.9   319.1   345.0   359.9
## --------------------------------------------------------
## data$School: StLouis
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   185.0   235.4   255.0   257.4   272.4   355.0
## --------------------------------------------------------
## data$School: StMarys
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   189.5   231.6   262.0   269.8   296.5   435.0
## --------------------------------------------------------
## data$School: Stratford
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   222.5   266.2   285.0   287.8   315.0   349.5
```

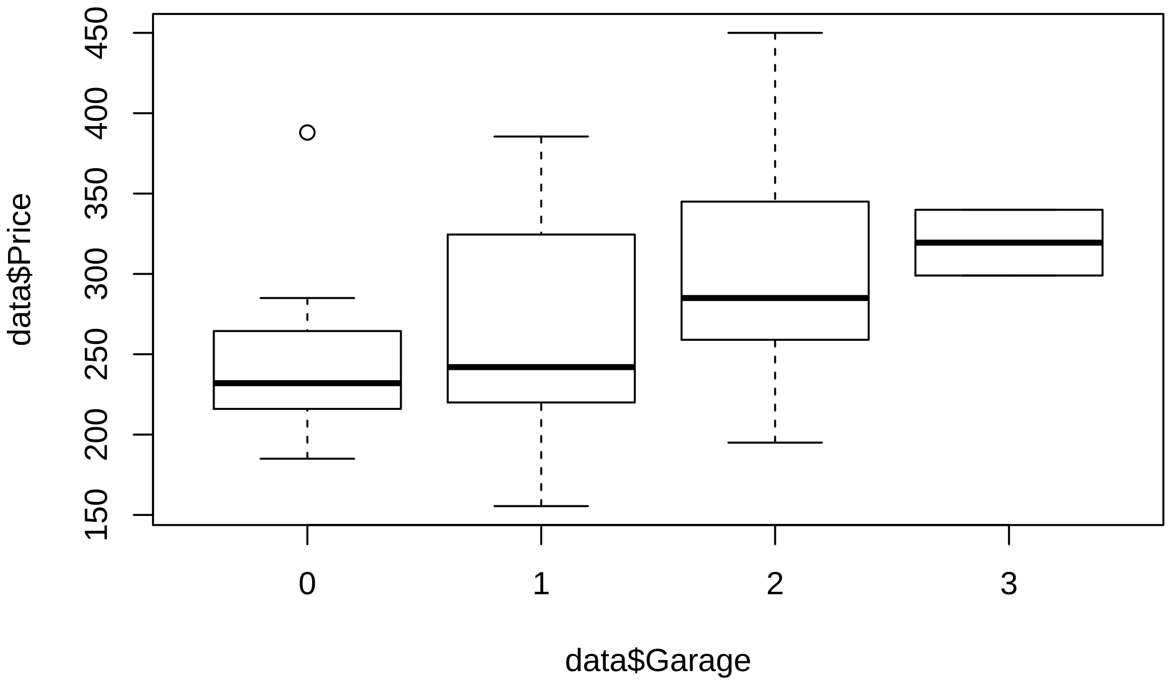
```
boxplot(data$Price~data$Bed)
```



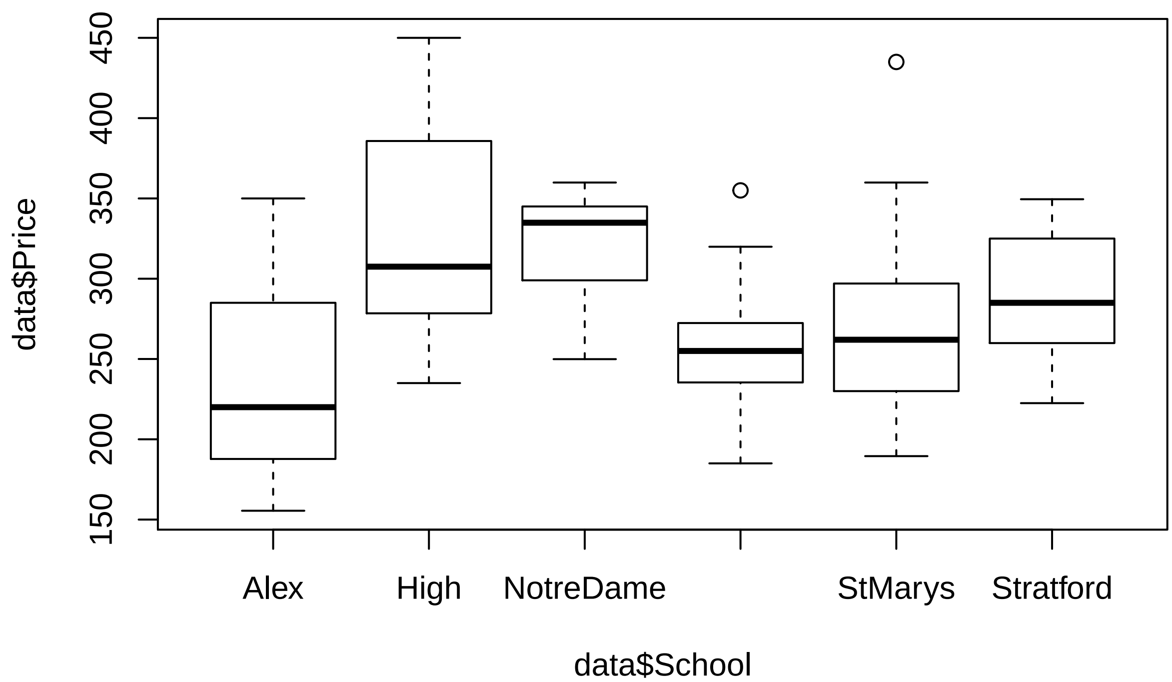

```r
boxplot(data$Price~data$Bath)
```




```r
boxplot(data$Price~data$Garage)
```




```r
boxplot(data$Price~data$School)
```



Number of bedrooms:

When the number of bedrooms is 2, the range is just of 51k. The price starts at 299k upto 350k. This means that the prices with 2 bedrooms do not have a high variance when compared to prices of house having 3,4 and 5 bedrooms which have a high range.

3 bed: range: 245.5k with prices starting from 189.5k upto 435k, Mean is 297 and median is 297.3 which are very close values. This shows that the data might be uniformly distributed.

4 bed: range: 294.5k with prices from 155.5k and 450k.Mean is 266.6 and median is 254.4 which are close values. This shows that the data might be uniformly distributed.

5 bed: range: 164.5k with prices from 185k to 349k. Mean is 259.5k and median is 269k which are close values. This shows that the data might be uniformly distributed.

6 bed: range is 0. There is just once input everything is the same. Number of Bathrooms

1 Bath: Range: 115k with prices from 235k to 350k . The mean is 292.5 and the median is also the same. Which shows that the data could be uniformly distributed. We can also check this by comparing the differences between each of the quartiles which is around 30k in all cases.

1.1 Bath:Range:170.5k with prices from 215k to 385.5k. The mean is 307 and the median is 325 which are not close. This shows that the data is not uniformly distributed. The difference between each quartile is also not close to each other.

2 Bath: Range: 279.5k with prices varying from 155.5k to 435k. The mean is 270k and median is 259k. The data is more weighted on the lower side and the data is not uniformly distributed.

2.1 Bath:Range:160k. The mean is 274.5k and the median is 269.9k which shows that the data might be uniformly distributed as the difference between each quartile is also approximately 20.

3 Bath:Range:220k. The mean is 307.8 and the median is 295k.

3.1 Bath:Range: 60k. The range is very less this shows that the prices are concentrated within a small region between 285 and 345k Garage size: 0: Range: 208k. The median and mean are far apart which shows that the data is not uniformly distributed. The gap between median and third quartile is much more than the difference between 1st and second. 1:Range:230k. Similar observations as above. 2:Range:205k. Similar but little less gap than above observations. 3:Range:40k. . The range is very less this shows that the prices are concentrated within a small region School : The prices of houses near Alexandera school are in general lower when compared to other schools. Cost of house near High school is much more higher. This shows that school does affect the cost of a house other things being constant.

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

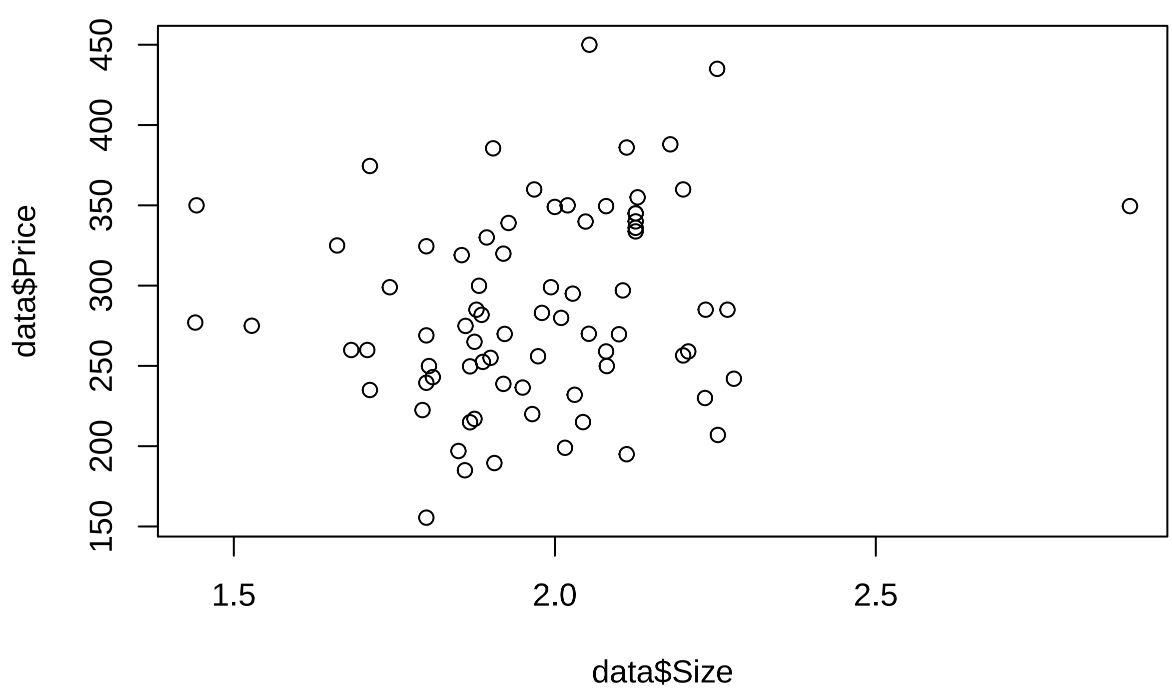
```
cor(data$Price,data$Size)
```

```
## [1] 0.2014378
```

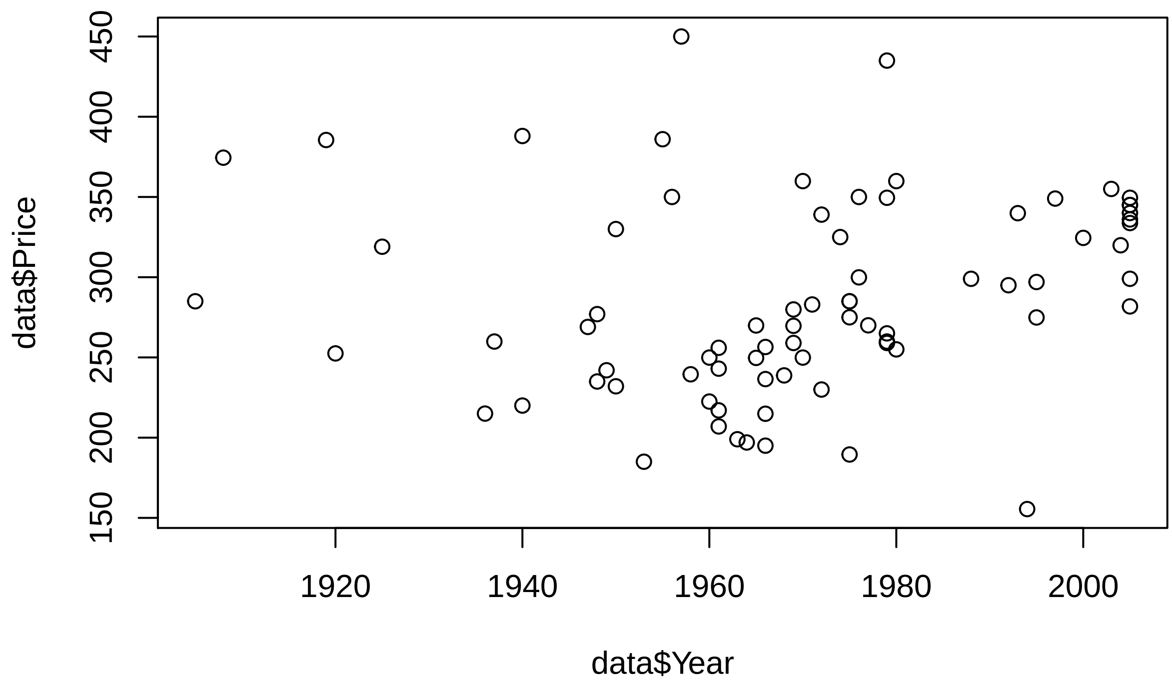
```
cor(data$Price,data$Year)
```

```
## [1] 0.1541248
```

```
plot(data$Size,data$Price)
```



```r
plot(data$Year,data$Price)
```




```r
by(data$Price,data$Size,summary)
```

```
## data$Size: 1.44
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     277     277     277     277     277     277
## ------------------------------------------------------------
## data$Size: 1.442
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     350     350     350     350     350     350
## ------------------------------------------------------------
## data$Size: 1.528
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     275     275     275     275     275     275
## ------------------------------------------------------------
## data$Size: 1.661
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     325     325     325     325     325     325
## ------------------------------------------------------------
## data$Size: 1.683
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   259.9   259.9   259.9   259.9   259.9   259.9
## ------------------------------------------------------------
## data$Size: 1.708
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   259.9   259.9   259.9   259.9   259.9   259.9
## --------------------------------------------------------
## data$Size: 1.712
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   235.0   269.9   304.8   304.8   339.6   374.5
## --------------------------------------------------------
## data$Size: 1.743
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     299     299     299     299     299     299
## --------------------------------------------------------
## data$Size: 1.794
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   222.5   222.5   222.5   222.5   222.5   222.5
## --------------------------------------------------------
## data$Size: 1.8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   218.5   254.2   247.1   282.9   324.5
## --------------------------------------------------------
## data$Size: 1.804
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   249.9   249.9   249.9   249.9   249.9   249.9
## --------------------------------------------------------
## data$Size: 1.81
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     243     243     243     243     243     243
## --------------------------------------------------------
## data$Size: 1.85
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     197     197     197     197     197     197
## --------------------------------------------------------
## data$Size: 1.855
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     319     319     319     319     319     319
## --------------------------------------------------------
## data$Size: 1.86
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     185     185     185     185     185     185
## --------------------------------------------------------
## data$Size: 1.861
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   274.9   274.9   274.9   274.9   274.9   274.9
## --------------------------------------------------------
## data$Size: 1.868
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   214.9   223.6   232.3   232.3   241.0   249.7
## --------------------------------------------------------
## data$Size: 1.875
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     217     229     241     241     253     265
## --------------------------------------------------------
## data$Size: 1.878
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     285     285     285     285     285     285
```

```
## ------------------------------------------------------------
## data$Size: 1.882
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   299.9   299.9   299.9   299.9   299.9   299.9 
## ------------------------------------------------------------
## data$Size: 1.886
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   281.8   281.8   281.8   281.8   281.8   281.8 
## ------------------------------------------------------------
## data$Size: 1.888
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   252.5   252.5   252.5   252.5   252.5   252.5 
## ------------------------------------------------------------
## data$Size: 1.894
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     330     330     330     330     330     330 
## ------------------------------------------------------------
## data$Size: 1.9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     255     255     255     255     255     255 
## ------------------------------------------------------------
## data$Size: 1.904
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   385.5   385.5   385.5   385.5   385.5   385.5 
## ------------------------------------------------------------
## data$Size: 1.906
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   189.5   189.5   189.5   189.5   189.5   189.5 
## ------------------------------------------------------------
## data$Size: 1.92
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   238.8   259.1   279.4   279.4   299.6   319.9 
## ------------------------------------------------------------
## data$Size: 1.922
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   269.9   269.9   269.9   269.9   269.9   269.9 
## ------------------------------------------------------------
## data$Size: 1.928
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     339     339     339     339     339     339 
## ------------------------------------------------------------
## data$Size: 1.95
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   236.5   236.5   236.5   236.5   236.5   236.5 
## ------------------------------------------------------------
## data$Size: 1.965
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     220     220     220     220     220     220 
## ------------------------------------------------------------
## data$Size: 1.968
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   359.9   359.9   359.9   359.9   359.9   359.9 
## ------------------------------------------------------------
## data$Size: 1.974
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     256     256     256     256     256     256
## --------------------------------------------------------
## data$Size: 1.98
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     283     283     283     283     283     283
## --------------------------------------------------------
## data$Size: 1.994
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     299     299     299     299     299     299
## --------------------------------------------------------
## data$Size: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     349     349     349     349     349     349
## --------------------------------------------------------
## data$Size: 2.01
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   279.9   279.9   279.9   279.9   279.9   279.9
## --------------------------------------------------------
## data$Size: 2.016
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     199     199     199     199     199     199
## --------------------------------------------------------
## data$Size: 2.02
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     350     350     350     350     350     350
## --------------------------------------------------------
## data$Size: 2.028
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     295     295     295     295     295     295
## --------------------------------------------------------
## data$Size: 2.031
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     232     232     232     232     232     232
## --------------------------------------------------------
## data$Size: 2.044
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     215     215     215     215     215     215
## --------------------------------------------------------
## data$Size: 2.048
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   339.9   339.9   339.9   339.9   339.9   339.9
## --------------------------------------------------------
## data$Size: 2.053
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     270     270     270     270     270     270
## --------------------------------------------------------
## data$Size: 2.054
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     450     450     450     450     450     450
## --------------------------------------------------------
## data$Size: 2.08
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   259.0   281.6   304.2   304.2   326.9   349.5
```

```
## --------------------------------------------------------
## data$Size: 2.081
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   249.9   249.9   249.9   249.9   249.9   249.9
## --------------------------------------------------------
## data$Size: 2.1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   269.7   269.7   269.7   269.7   269.7   269.7
## --------------------------------------------------------
## data$Size: 2.106
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     297     297     297     297     297     297
## --------------------------------------------------------
## data$Size: 2.112
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   195.0   242.8   290.5   290.5   338.2   386.0
## --------------------------------------------------------
## data$Size: 2.126
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   333.8   336.0   340.0   339.9   345.0   345.0
## --------------------------------------------------------
## data$Size: 2.129
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     355     355     355     355     355     355
## --------------------------------------------------------
## data$Size: 2.18
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     388     388     388     388     388     388
## --------------------------------------------------------
## data$Size: 2.2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   256.5   282.4   308.2   308.2   334.1   359.9
## --------------------------------------------------------
## data$Size: 2.208
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     259     259     259     259     259     259
## --------------------------------------------------------
## data$Size: 2.234
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     230     230     230     230     230     230
## --------------------------------------------------------
## data$Size: 2.235
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     285     285     285     285     285     285
## --------------------------------------------------------
## data$Size: 2.253
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     435     435     435     435     435     435
## --------------------------------------------------------
## data$Size: 2.254
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     207     207     207     207     207     207
## --------------------------------------------------------
## data$Size: 2.269
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     285     285     285     285     285     285
## --------------------------------------------------------
## data$Size: 2.279
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     242     242     242     242     242     242
## --------------------------------------------------------
## data$Size: 2.896
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   349.5   349.5   349.5   349.5   349.5   349.5
```

```r
by(data$Price,data$Year,summary)
```

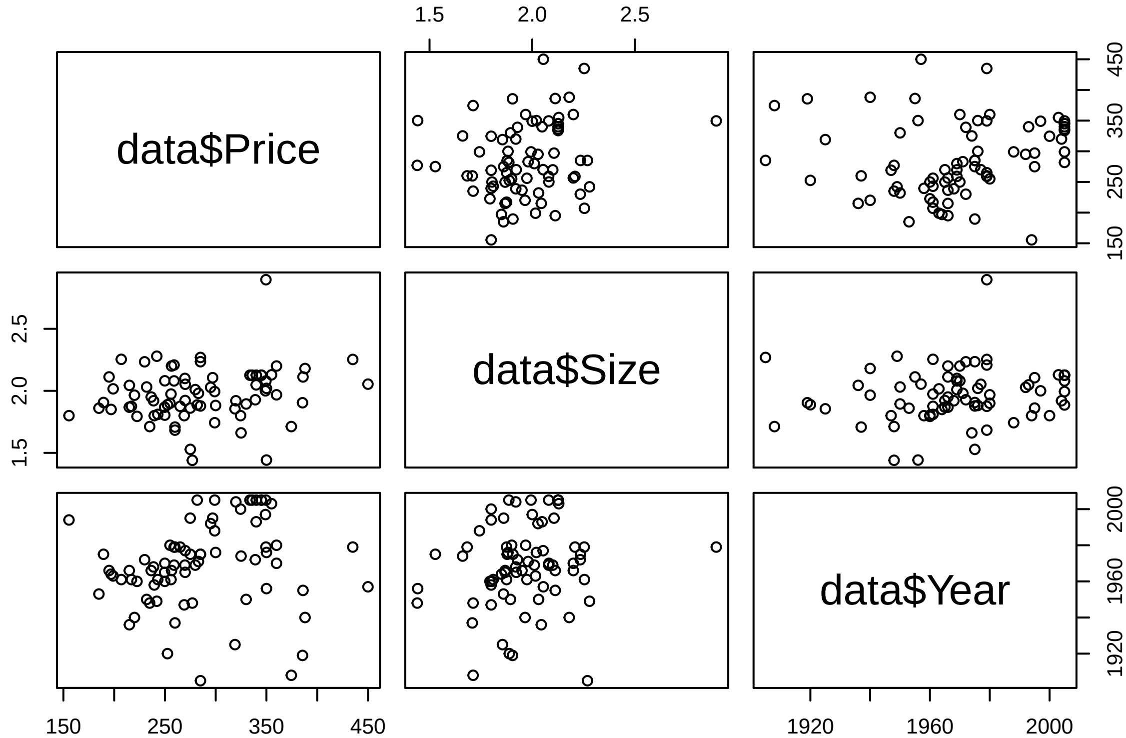
```
## data$Year: 1905
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     285     285     285     285     285     285
## --------------------------------------------------------
## data$Year: 1908
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   374.5   374.5   374.5   374.5   374.5   374.5
## --------------------------------------------------------
## data$Year: 1919
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   385.5   385.5   385.5   385.5   385.5   385.5
## --------------------------------------------------------
## data$Year: 1920
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   252.5   252.5   252.5   252.5   252.5   252.5
## --------------------------------------------------------
## data$Year: 1925
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     319     319     319     319     319     319
## --------------------------------------------------------
## data$Year: 1936
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     215     215     215     215     215     215
## --------------------------------------------------------
## data$Year: 1937
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   259.9   259.9   259.9   259.9   259.9   259.9
## --------------------------------------------------------
## data$Year: 1940
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     220     262     304     304     346     388
## --------------------------------------------------------
## data$Year: 1947
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     269     269     269     269     269     269
## --------------------------------------------------------
## data$Year: 1948
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   235.0   245.5   256.0   256.0   266.5   277.0
## --------------------------------------------------------
## data$Year: 1949
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     242     242     242     242     242     242
## ------------------------------------------------------------
## data$Year: 1950
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   232.0   256.5   281.0   281.0   305.5   330.0
## ------------------------------------------------------------
## data$Year: 1953
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     185     185     185     185     185     185
## ------------------------------------------------------------
## data$Year: 1955
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     386     386     386     386     386     386
## ------------------------------------------------------------
## data$Year: 1956
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     350     350     350     350     350     350
## ------------------------------------------------------------
## data$Year: 1957
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     450     450     450     450     450     450
## ------------------------------------------------------------
## data$Year: 1958
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   239.5   239.5   239.5   239.5   239.5   239.5
## ------------------------------------------------------------
## data$Year: 1960
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   222.5   229.3   236.2   236.2   243.1   249.9
## ------------------------------------------------------------
## data$Year: 1961
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   207.0   214.5   230.0   230.8   246.2   256.0
## ------------------------------------------------------------
## data$Year: 1963
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     199     199     199     199     199     199
## ------------------------------------------------------------
## data$Year: 1964
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     197     197     197     197     197     197
## ------------------------------------------------------------
## data$Year: 1965
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   249.7   254.8   259.8   259.8   264.9   269.9
## ------------------------------------------------------------
## data$Year: 1966
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   195.0   209.9   225.7   225.7   241.5   256.5
## ------------------------------------------------------------
## data$Year: 1968
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   238.8   238.8   238.8   238.8   238.8   238.8
```

```
## ----------------------------------------------------------
## data$Year: 1969
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   259.0   264.4   269.7   269.5   274.8   279.9 
## ----------------------------------------------------------
## data$Year: 1970
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   249.9   277.4   304.9   304.9   332.4   359.9 
## ----------------------------------------------------------
## data$Year: 1971
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     283     283     283     283     283     283 
## ----------------------------------------------------------
## data$Year: 1972
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   230.0   257.2   284.5   284.5   311.8   339.0 
## ----------------------------------------------------------
## data$Year: 1974
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     325     325     325     325     325     325 
## ----------------------------------------------------------
## data$Year: 1975
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   189.5   253.6   280.0   258.6   285.0   285.0 
## ----------------------------------------------------------
## data$Year: 1976
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   299.9   312.4   324.9   324.9   337.5   350.0 
## ----------------------------------------------------------
## data$Year: 1977
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     270     270     270     270     270     270 
## ----------------------------------------------------------
## data$Year: 1979
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   259.0   259.9   265.0   313.7   349.5   435.0 
## ----------------------------------------------------------
## data$Year: 1980
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   255.0   281.2   307.4   307.4   333.7   359.9 
## ----------------------------------------------------------
## data$Year: 1988
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     299     299     299     299     299     299 
## ----------------------------------------------------------
## data$Year: 1992
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     295     295     295     295     295     295 
## ----------------------------------------------------------
## data$Year: 1993
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   339.9   339.9   339.9   339.9   339.9   339.9 
## ----------------------------------------------------------
## data$Year: 1994
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   155.5   155.5   155.5   155.5   155.5
## --------------------------------------------------------
## data$Year: 1995
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   274.9   280.4   285.9   285.9   291.5   297.0
## --------------------------------------------------------
## data$Year: 1997
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     349     349     349     349     349     349
## --------------------------------------------------------
## data$Year: 2000
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   324.5   324.5   324.5   324.5   324.5   324.5
## --------------------------------------------------------
## data$Year: 2003
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     355     355     355     355     355     355
## --------------------------------------------------------
## data$Year: 2004
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   319.9   319.9   319.9   319.9   319.9   319.9
## --------------------------------------------------------
## data$Year: 2005
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   281.8   325.1   338.0   328.8   345.0   349.5
```

```r
pairs(data$Price~data$Size+data$Year)
```



Price vs Year As we move forward in time from 1900's the number of data points increase with maximum number of data points between 1960 and 1980. The houses are priced between 150k to 450k. The range of the cost of houses increase as the number of data points increase with the maximum range in around 1950-1970. During the years 1950-1980 the prices are more scattered when compared to prices of the house afterwards which are not very scattered and lie between 270k and 350k in the years 1990 to end of the data. During late 1950s and 1960s the prices of a house varied from as low as 175 to ass high as 450. The correlation between price and year is positive and non zero which shows that the prices tend to increase over time.

Price vs Size We can see that the prices and size have a positive correlation of 0.2, which shows that as size increases there is an average increase in the cost of a house. The maximum number of houses have the size in the range between 1.75k and 2.3k. the houses outside this range can be considered as outliers as there are very less number of data points outside of this range.

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.

```
model=lm(Price~Size+Lot+Bath+Bed+Year+Garage+School,data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ Size + Lot + Bath + Bed + Year + Garage +
##     School, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

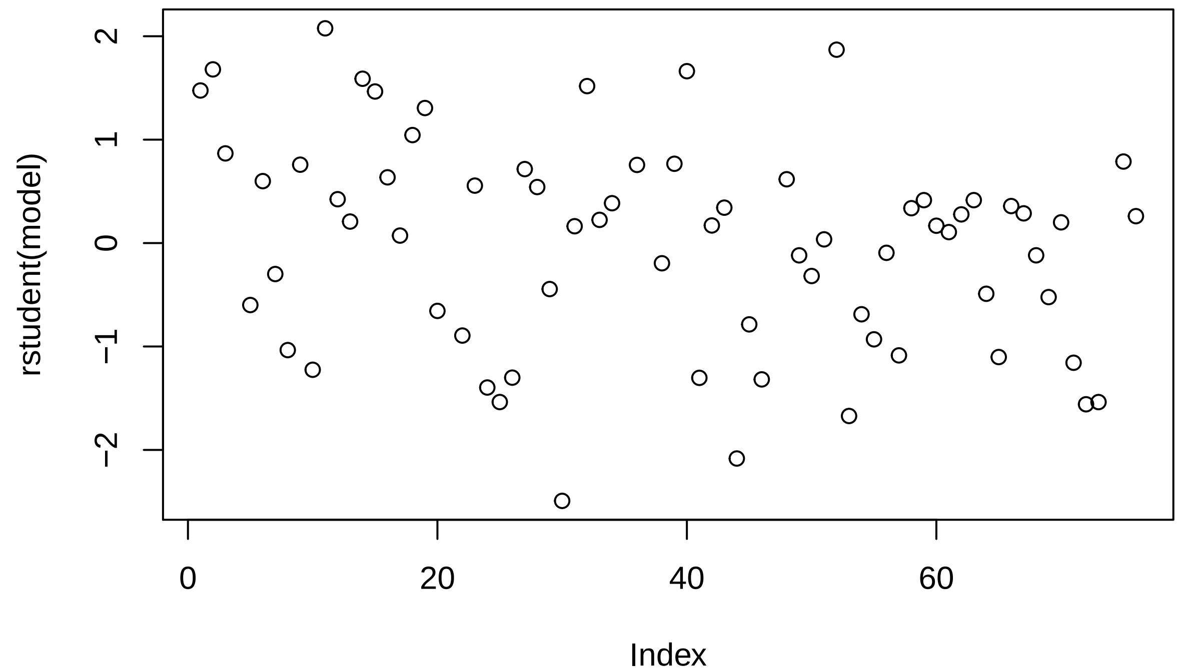
```
## -83.381 -18.517   0.181  20.956  74.167
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -704.9030   719.8736  -0.979 0.332194
## Size                37.6596    29.3930   1.281 0.206019
## Lot2                26.7554    39.0779   0.685 0.496714
## Lot3                 0.7350    33.1003   0.022 0.982372
## Lot4                 6.9079    31.8178   0.217 0.829007
## Lot5                31.7736    33.2014   0.957 0.343173
## Lot6               250.3842    84.2767   2.971 0.004553 **
## Lot7                37.2430    34.7310   1.072 0.288722
## Lot8               -53.0001    69.8926  -0.758 0.451827
## Lot11              175.6594    53.3634   3.292 0.001831 **
## Bath1.1            142.5320    47.8965   2.976 0.004493 **
## Bath2               94.7010    44.8319   2.112 0.039672 *
## Bath2.1             99.2583    46.2510   2.146 0.036744 *
## Bath3              137.1779    47.1175   2.911 0.005362 **
## Bath3.1            110.6906    52.3422   2.115 0.039458 *
## Bed3               -36.7818    44.1971  -0.832 0.409242
## Bed4               -49.5788    45.6583  -1.086 0.282746
## Bed5               -46.7850    49.9368  -0.937 0.353322
## Bed6               -90.6022    68.9303  -1.314 0.194708
## Year                 0.3762     0.3590   1.048 0.299725
## Garage              14.4581     8.7289   1.656 0.103915
## SchoolHigh         133.2335    37.0731   3.594 0.000744 ***
## SchoolNotreDame     99.3080    35.0883   2.830 0.006681 **
## SchoolStLouis       47.0708    35.1413   1.339 0.186475
## SchoolStMarys       48.1541    34.9070   1.379 0.173880
## SchoolStratford     70.7054    39.6900   1.781 0.080915 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.78 on 50 degrees of freedom
## Multiple R-squared:  0.6955, Adjusted R-squared:  0.5432
## F-statistic: 4.568 on 25 and 50 DF,  p-value: 2.475e-06
```

```r
#removing garage 3 as it is giving na values
data$Garage = factor(data$Garage, levels=c(0,1,2,3),labels=c(0,1,2,2))
model=lm(Price~Size+Lot+Bath+Bed+Year+Garage+School,data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Price ~ Size + Lot + Bath + Bed + Year + Garage +
##     School, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.634 -18.850   3.274  16.949  72.932
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -884.2593   718.3257  -1.231 0.224199
```

```
## Size              39.3738    28.8314   1.366 0.178283
## Lot2              56.9196    42.0487   1.354 0.182056
## Lot3              16.2949    33.6585   0.484 0.630454
## Lot4              20.0807    32.0972   0.626 0.534467
## Lot5              50.5869    34.2969   1.475 0.146617
## Lot6             224.5395    83.9993   2.673 0.010179 *
## Lot7              48.2408    34.6292   1.393 0.169888
## Lot8             -31.1897    69.6543  -0.448 0.656287
## Lot11            192.0355    53.1530   3.613 0.000712 ***
## Bath1.1          131.5562    47.3757   2.777 0.007754 **
## Bath2             65.1129    47.1245   1.382 0.173325
## Bath2.1           74.2050    47.5720   1.560 0.125232
## Bath3            104.2021    49.9269   2.087 0.042103 *
## Bath3.1           84.5954    53.4588   1.582 0.119983
## Bed3             -47.4253    43.3741  -1.093 0.279564
## Bed4             -54.4886    44.9627  -1.212 0.231376
## Bed5             -48.2645    48.5054  -0.995 0.324608
## Bed6            -118.6555    67.5833  -1.756 0.085391 .
## Year               0.4996     0.3590   1.392 0.170263
## Garage1          -24.3447    23.8857  -1.019 0.313107
## Garage2           18.8511    18.0653   1.043 0.301838
## SchoolHigh        99.7440    41.1252   2.425 0.019023 *
## SchoolNotreDame   72.6369    37.6585   1.929 0.059555 .
## SchoolStLouis     11.1148    40.1721   0.277 0.783190
## SchoolStMarys     13.1969    39.6816   0.333 0.740878
## SchoolStratford   29.6125    45.5153   0.651 0.518341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.97 on 49 degrees of freedom
## Multiple R-squared:  0.7132, Adjusted R-squared:  0.561
## F-statistic: 4.687 on 26 and 49 DF,  p-value: 1.663e-06
```

```r
plot(rstudent(model))
```



```r
summary(rstudent(model))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -2.49266 -0.76169  0.16909 -0.01179  0.61262  2.07634        6
```

1) y = -884.2593 + 39.3738*Size + Lot2*56.9196 + Lot3*16.2949 + Lot4*20.0807 + Lot5*50.5869 + Lot6*224.5395

- Lot7*48.2408 + Lot8+(-31.1897) + Lot11*192.0355 + Bath1.1*131.5562
- Bath2*65.1129 + Bath2.1*74.2050 + Bath3*104.2021 + Bath3.1*84.5954 + Bed3*(-47.4253) + Bed4*(-54.4886)
- Bed5*(-48.2645) + Bed6*(-118.6555) + 0.4996*Year + Garage1*(-24.3447) + Garage2*(18.8511) + School-High*99.7440 + SchoolNotreDame*72.6369 + SchoolStLouis*11.1148 + SchoolStMarys*13.1969 + School-Stratford*29.6125

2)

0 : -884.2593

3) size : 39.3738

4) Bath1.1 : 131.5562

5) Bed2 is the default so there would be no change in the price of the house given that other factors remain constant. incase of 3 beds, the cost would decrease by 47000 approximately. Similarly the price would change based on the slope as given in the above table. The price would decrease by 54488 in case of 4 beds, it would decrease by 48264 in case of 5 beds and by 118655 in case of six beds given that none of the other factors are changed.

6) The predictor variables that change the price significantly are Lot11
Bath1.1

7) Size : 1 Unit increase leads to an increase of 39.3738k in increase of price (Maximum : 2.896) Lot : Lot6 should be 1 Bath : Bath1.1 should be 1 Bed : House with 2 beds would have maximum value Year : 1 Unit of increase leads to 0.4996k increase in price (Year : 2005) Garage : Garage2 should be selected School : High School

8) Size : 1 Unit decrease leads to a decrease of 39.3738k in increase of price (Minimum : 1.44) Lot : Lot8 should be 1 Bath : Bath1 should be 1 i.e. 0 Bed : House with 6 beds would have minimum value Year : 1 Unit of decrease leads to 0.4996k decrease in price (Year 1905) Garage : Garage2 should be selected School : Alexandra School

9) The residuals are scattered around the 0 value with some variance. Some values are outside the range of -2 to 2 which can be considered as outliers.The gap between first quartile, median and second quartile is approximately the same which shows that the data is uniformly distributed (can be inferred from the summary) with the range of around 4.5. The residual standard error is 39.97%, which shows that the model is not adequate for getting the prices.

10) Adjusted R-squared value: .561

11) F-statistic: 4.687 on 26 and 49 Degrees of freedom, p-value: 1.663e-06 The hypothesis being tested is that all beta values are 0. As p-value is really close to 0 and is not significant when compared to the F-statistic value we can safely reject the NULL hypothesis.

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Size       1  11078 11077.7  6.9327 0.0112903 *
## Lot        8  45378  5672.2  3.5498 0.0025749 **
## Bath       5  41999  8399.8  5.2568 0.0006084 ***
## Bed        4  23601  5900.2  3.6925 0.0104985 *
## Year       1   2057  2056.9  1.2872 0.2620764
## Garage     2  10786  5393.2  3.3752 0.0423383 *
## School     5  59808 11961.6  7.4859 2.766e-05 ***
## Residuals 49  78296  1597.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1)

Analysis of Variance (ANOVA) is a statistical analysis to test NULL hypothesis for H~0 and non zero beta for H~1.

Size: 0.0112903 We can reject the NULL hypothesis Lot : 0.0025749 We can reject the NULL hypothesis Bath : 0.0006084 We can reject the NULL hypothesis Bed : 0.0104985 We can reject the NULL hypothesis year

: 0.2620764 We can't reject the NULL hypothesis as P value is significant, we can put year as 0. Garage : 0.0423383 We can reject the NULL hypothesis School : 2.766e-05 We can reject the NULL hypothesis Conclusion : Year has a high value of NULL hypothesis, so we can consider removing year.

```
model=lm(Price~Size+Lot+Bath+Bed+Year+Garage+School,data = data)
Anova(model,type=2)
```

```
## Anova Table (Type II tests)
##
## Response: Price
##           Sum Sq Df F value    Pr(>F)
## Size        2980  1  1.8650 0.1782835
## Lot        52190  8  4.0827 0.0008723 ***
## Bath       27430  5  3.4333 0.0097351 **
## Bed         5181  4  0.8105 0.5245257
## Year        3095  1  1.9372 0.1702629
## Garage      9399  2  2.9412 0.0621867 .
## School     59808  5  7.4859 2.766e-05 ***
## Residuals  78296 49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#removing variables with high null hypothesis
model1=lm(Price~Lot+Bath+Garage+School,data = data)
Anova(model1,type=2)
```

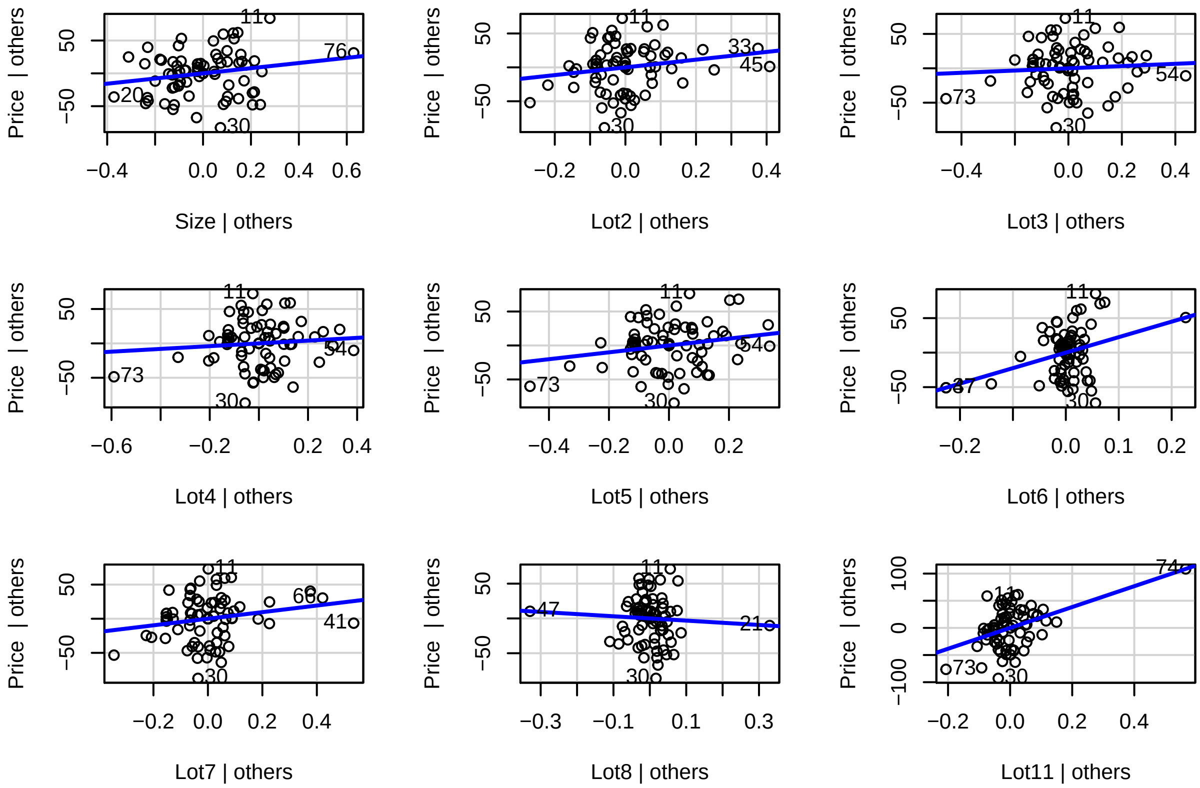
```
## Anova Table (Type II tests)
##
## Response: Price
##           Sum Sq Df F value    Pr(>F)
## Lot        63199  8  4.6396 0.0002271 ***
## Bath       26864  5  3.1555 0.0141673 *
## Garage     19639  2  5.7670 0.0053242 **
## School     55622  5  6.5334 7.781e-05 ***
## Residuals  93648 55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

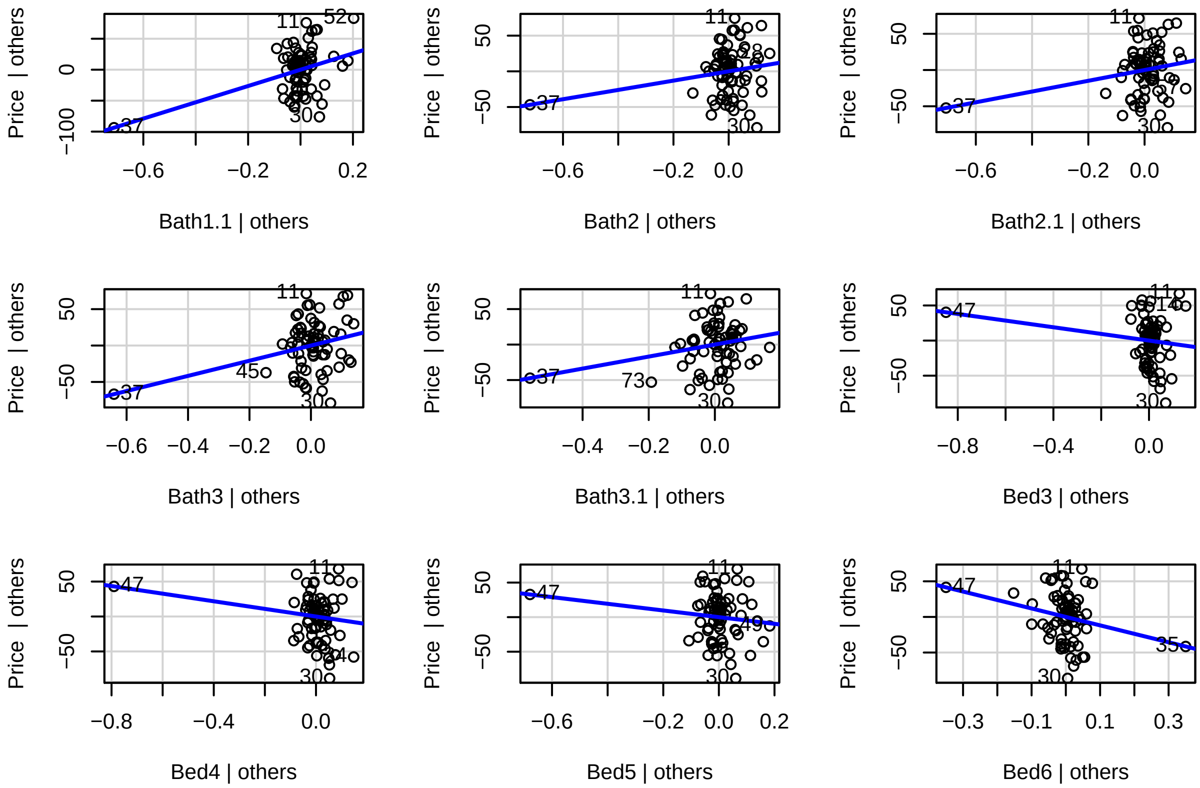
Size: 0.1782835 We can't reject the NULL hypothesis as P value is significant. Lot : 0.0008723 We can reject the NULL hypothesis Bath : 0.0097351 We can reject the NULL hypothesis Bed : 0.5245257 We can't reject the NULL hypothesis as P value is significant. year : 0.1702629 We can't reject the NULL hypothesis as P value is significant. Garage : 0.0621867 We can reject the NULL hypothesis School : 2.766e-05 We can reject the NULL hypothesis

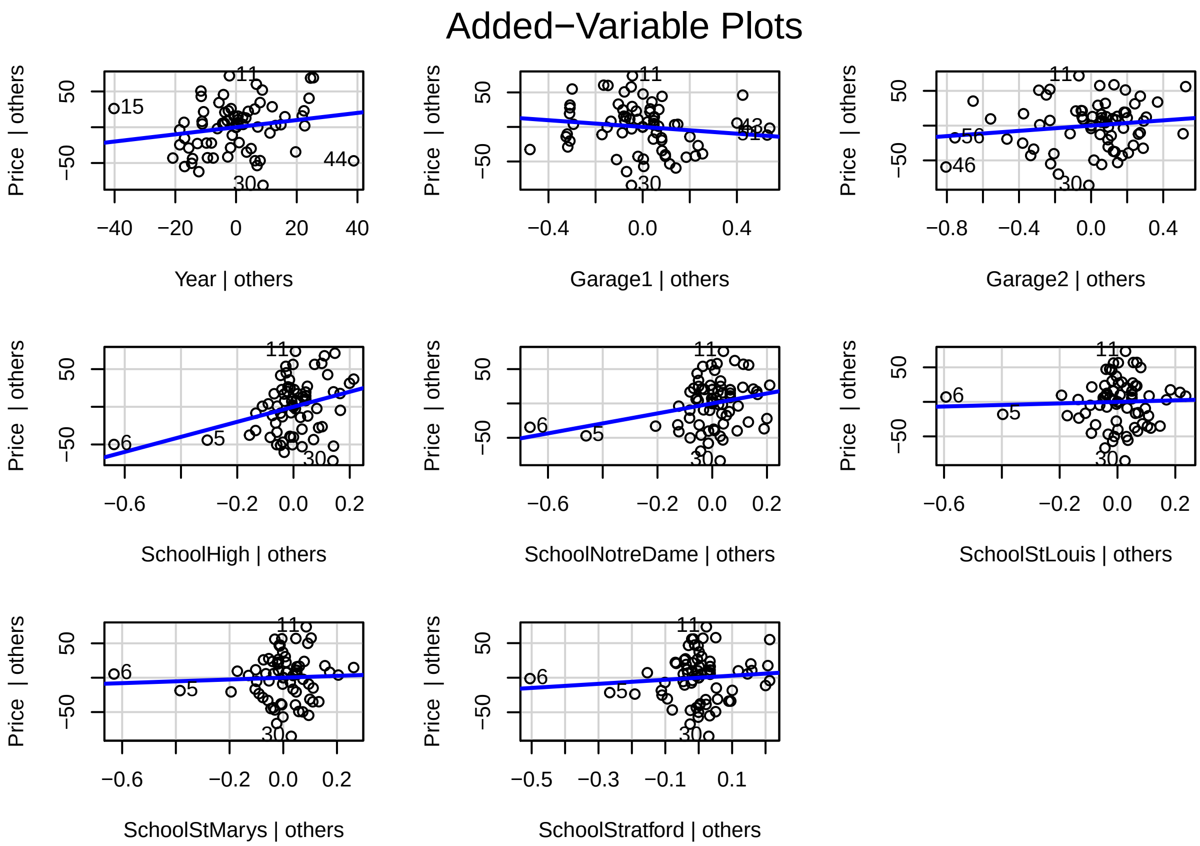
After removing variables with low significance Lot : 0.0002271 We can reject the NULL hypothesis Bath : 0.0141673 We can reject the NULL hypothesis Garage :0.0053242 We can reject the NULL hypothesis School : 7.781e-05 We can reject the NULL hypothesis

As we decrease the number of variables, the value of other p-values tend towards zero showing that they are even more significant now.

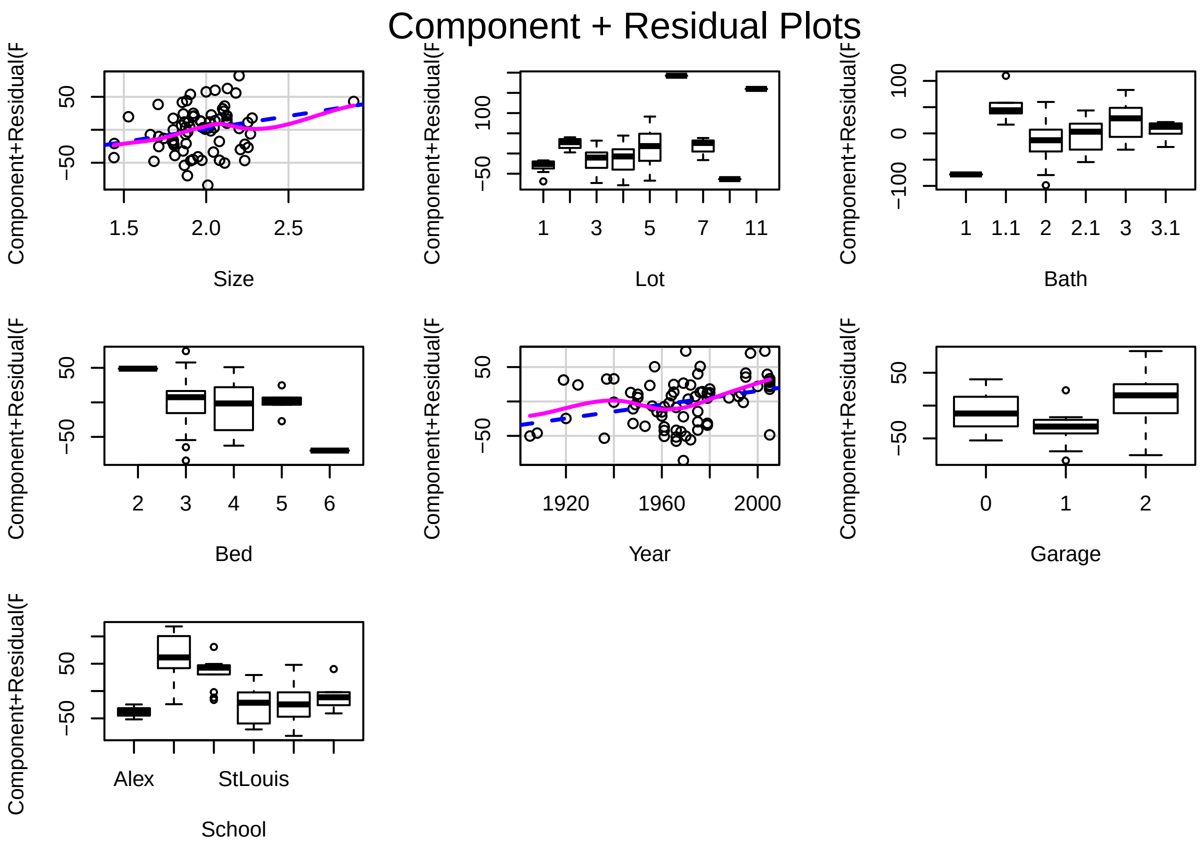
```
#q1
avPlots(model)
```


Price | others
Size | others
Lot2 | others
Lot3 | others
Lot4 | others
Lot5 | others
Lot6 | others
Lot7 | others
Lot8 | others
Lot11 | others


Price | others
Bath1.1 | others
Bath2 | others
Bath2.1 | others
Bath3 | others
Bath3.1 | others
Bed3 | others
Bed4 | others
Bed5 | others
Bed6 | others




```
crPlots(model)
```



From the above AV plots we can summarize that even though we are trying to model the data in a linear plot, the data is highly scattered along the regression line which shows high amount of variance. The plot is accurate for a few variables which have less amount of variance such as for Alexandra school, the box plot is highly concentrated, same is true for garage1. We can improve this by using splines, polynomials(more than 1 degree) or transformations. From CV plots we can see that linear model and smooth model are close.

Size,Lot2,lot5,lot6,lot7,lot11,Bath1.1,Bath2,Bath2.1,Bath3,Bath3.1 and price have a postive correlation, as the line is upward sloping whereas lot8,Bed3,Bed4,Bed5 and bed6 have a negative correlation .

The copmponent + residual plot shows whether the relationship is linear or not. In the fist plot we can see that the dashed line and the pink line are very close which shows that this model can be represented linearly.

We can improve this by using polynomials (greater than 1 degree) or perhaps transformations, or even dimensionality reduction etc.

```
#q2
dwt(model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1668749      1.636786    0.05
##  Alternative hypothesis: rho != 0
```

We can improve the given model, as there is a significant amount of autocorrelation. We can reject the NULL hypothesis as p value is just .044.

2 common violations are:

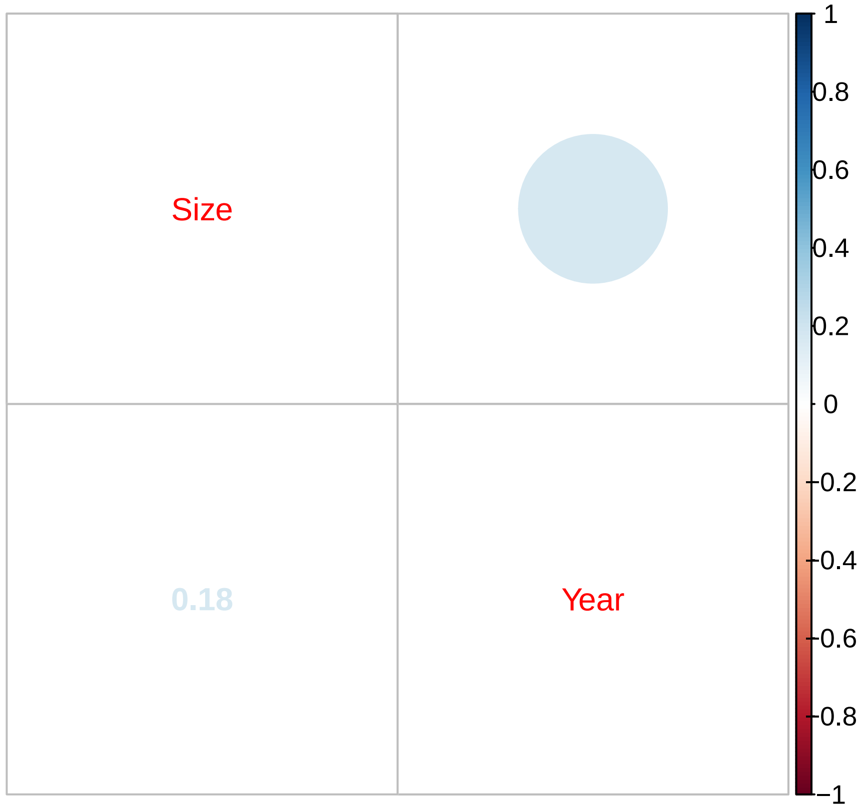
- Outliers in the model lead to non constant variance and biased and inefficient result.

- Violation of random/i.i.d sample assumption results in heteroscedasticity.

The dependent samples would lead to a biased result which could push a model in one direction leading to misleading output.

We could correct the outliers by using filters after creating the first model of data, or we could use time series analysis (ARIMA modelling) or we could use PCA to remove non significant variables.

```
#q3
corr = corrplot.mixed(cor(data[,c(2,6)]))
```




```
corr
```

```
##           Size      Year
## Size 1.0000000 0.1765693
## Year 0.1765693 1.0000000
```

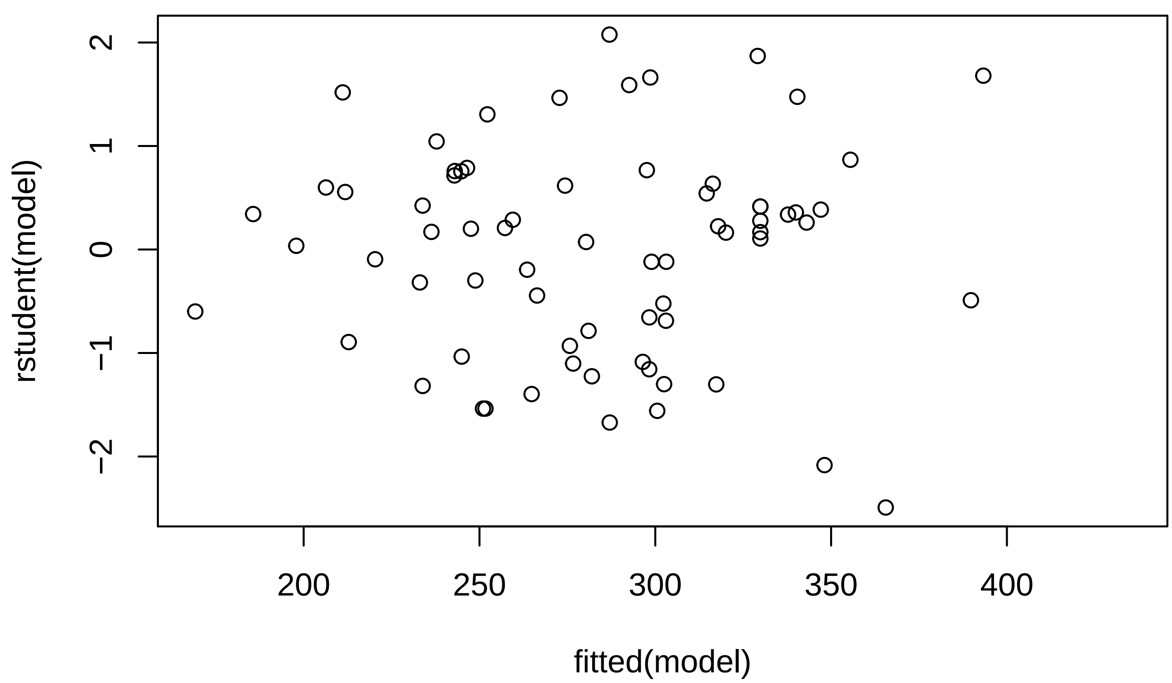
```
vif(model)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## Size     1.760512  1        1.326843
## Lot    113.958044  8        1.344456
## Bath    28.162186  5        1.396261
## Bed     11.744272  4        1.360593
## Year     3.338228  1        1.827082
## Garage   4.227787  2        1.433931
## School  12.945853  5        1.291853
```

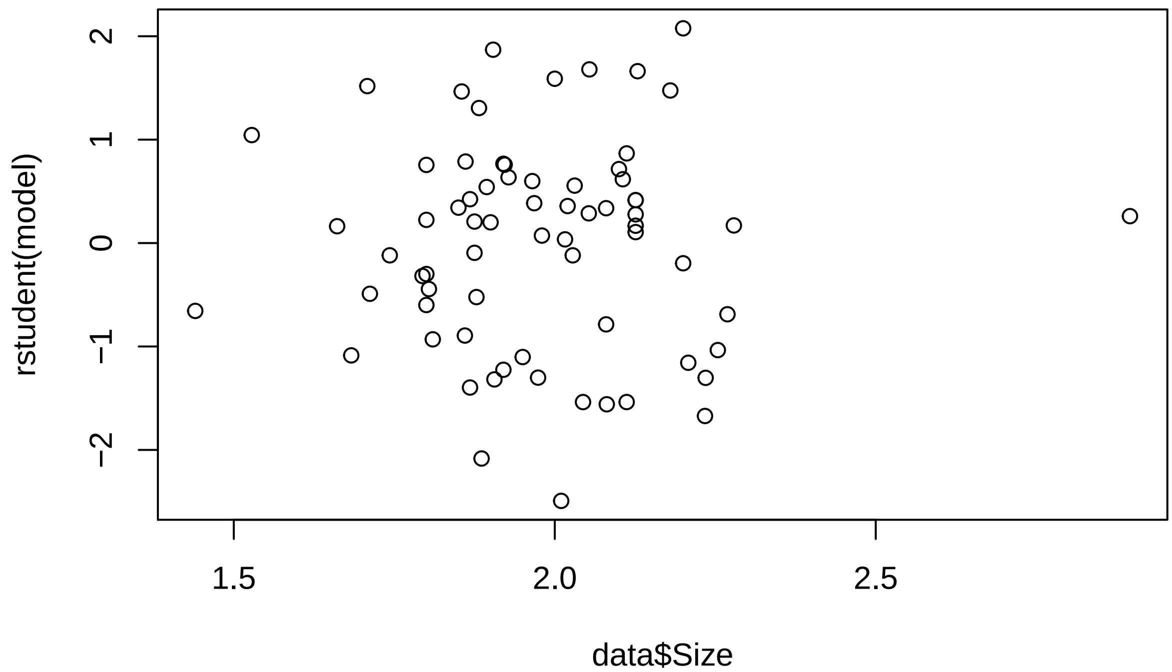
3) We can create corelation plot of only numerical variables (The value must be numeric). The year and size are correlated with a factor of 0.17 which means that we should not have any problem regarding the same.

Multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. This could be an issue as one variable could be derived by a linear combination of others making it redundant, as it could safely be removed by modifying the other coefficients. This makes the model unnecessarily complex, adding parameters that we don't actually need. We can remove variables which are highly dependant on others, and modify our linear model accordingly. We can do a dimension reduction using PCA or any other techniques.

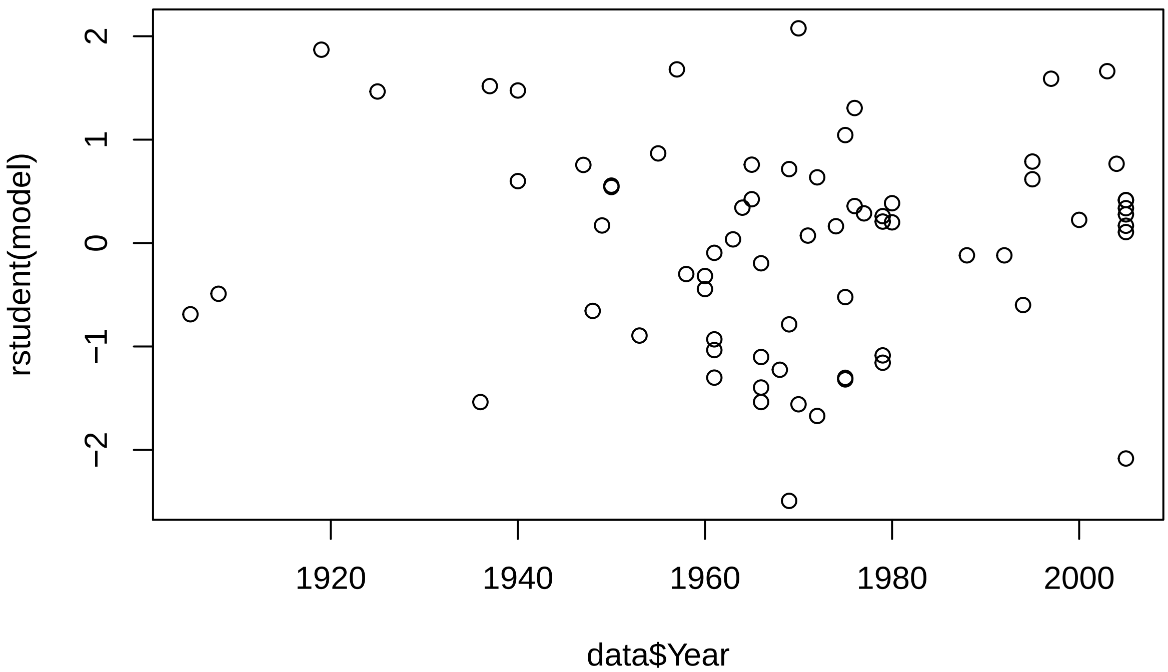
```r
#q4
plot(fitted(model),rstudent(model))
```




```r
plot(data$Size,rstudent(model))
```

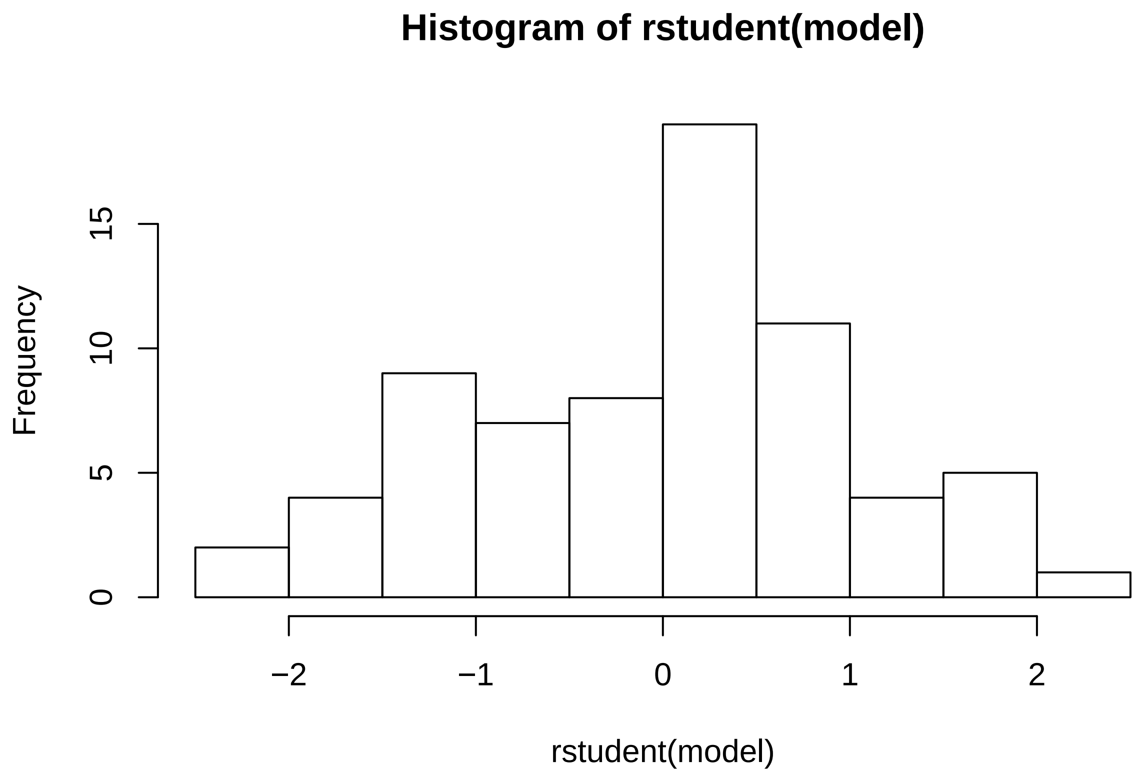



```
plot(data$Year,rstudent(model))
```

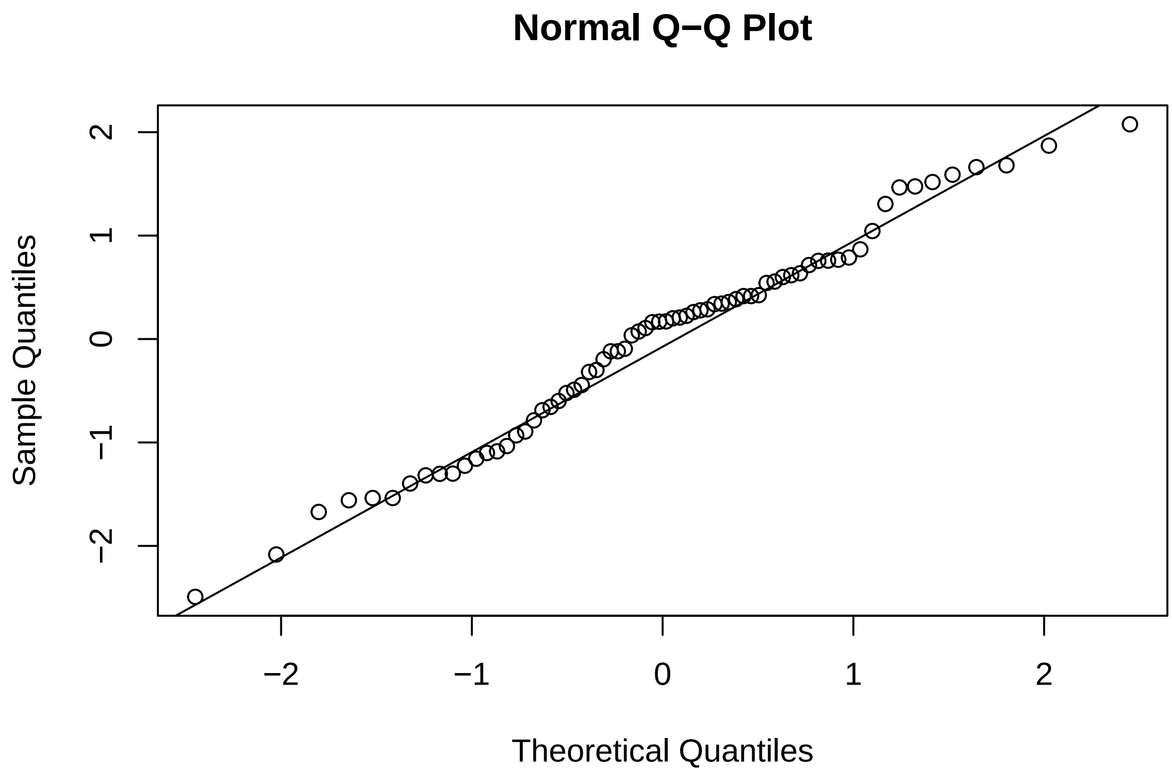


4) the points are randomly scattered around the origin, they have no as such defined pattern and the variance seems to be constant. There is no heteroscedasticity as no pattern can be seen (errors do not seem biased). If the errors are biased or there is heteroscedasticity then that means the model is not well represented as we are missing out some pattern present in the output. heteroscedasticity can be removed by taking out by using weighted least squares technique.

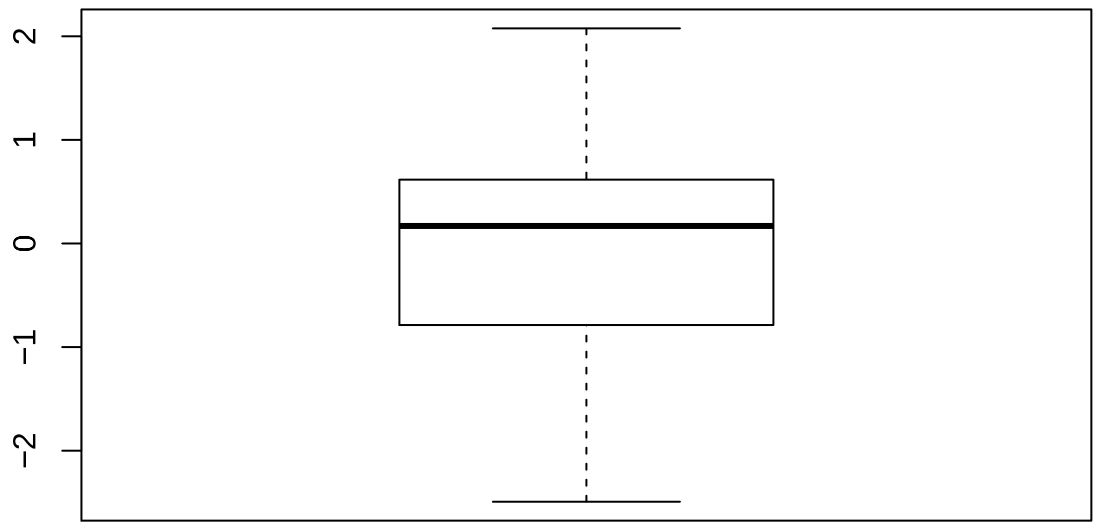
```
#q5
hist(rstudent(model))
```




```r
qqnorm(rstudent(model))
qqline(rstudent(model))
```




```
boxplot(rstudent(model))
```




```r
summary(rstudent(model))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
## -2.49266 -0.76169  0.16909 -0.01179  0.61262  2.07634        6
```

The data seems to be normally distributed. No outliers were found in the box plot.The data seems to be a little rightward shifted. This can be seen in both histogram and box plot. The density is higher on the right side.

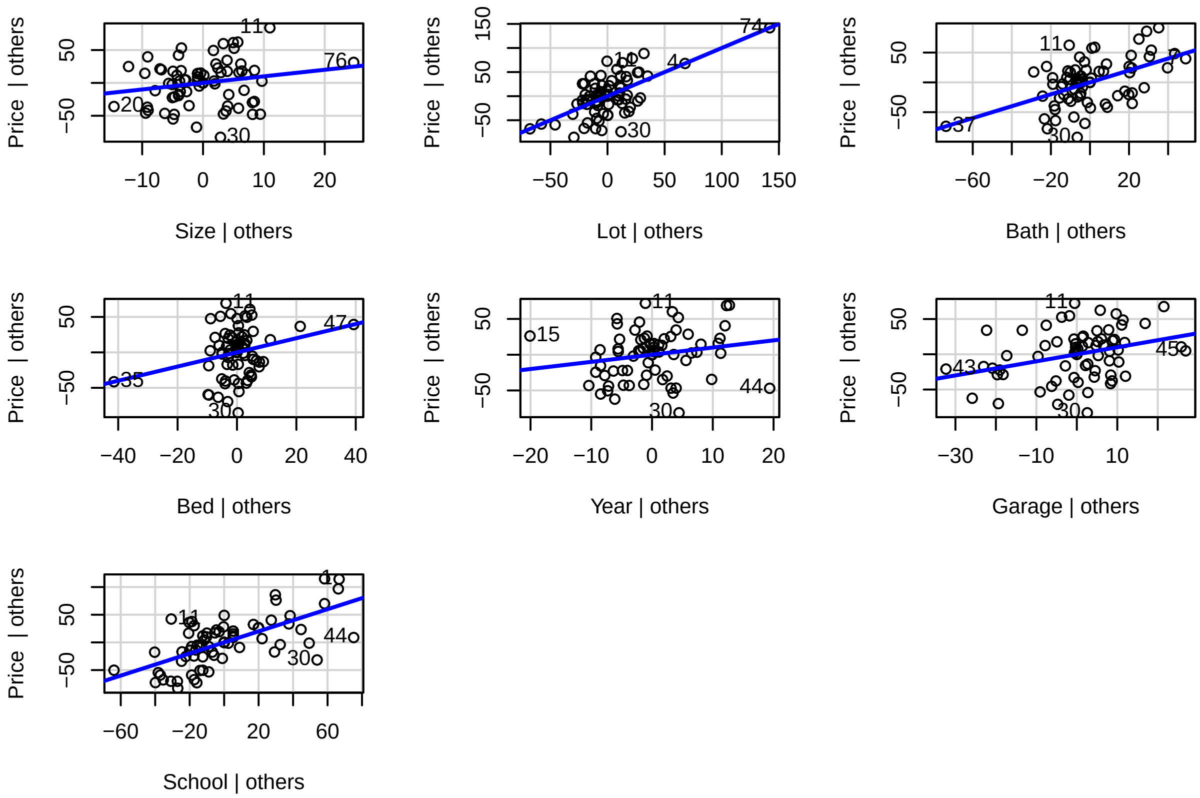
From the qq plot we can see that most of the points lie on the straight line which means that the residuals are normally distributed.

Effect of non-normality have on the regression model the amount of error in our model is not consistent across the full range of your observed data. The values of T-test and F-test are affected

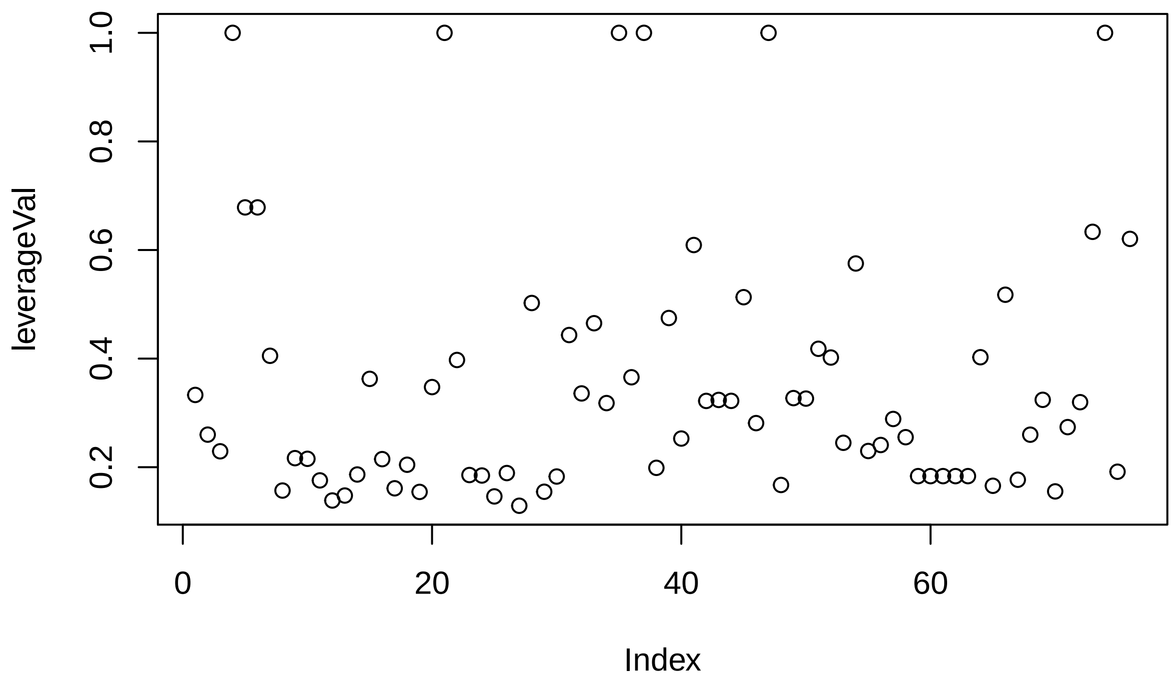
The non-normality can be corrected by using a different model, interactions and transformations.

```r
#q1
leveragePlots(model)
```

Leverage Plots




```r
leverageVal = hat(model.matrix(model))
plot(leverageVal)
```




```
data[leverageVal>0.2,]
```

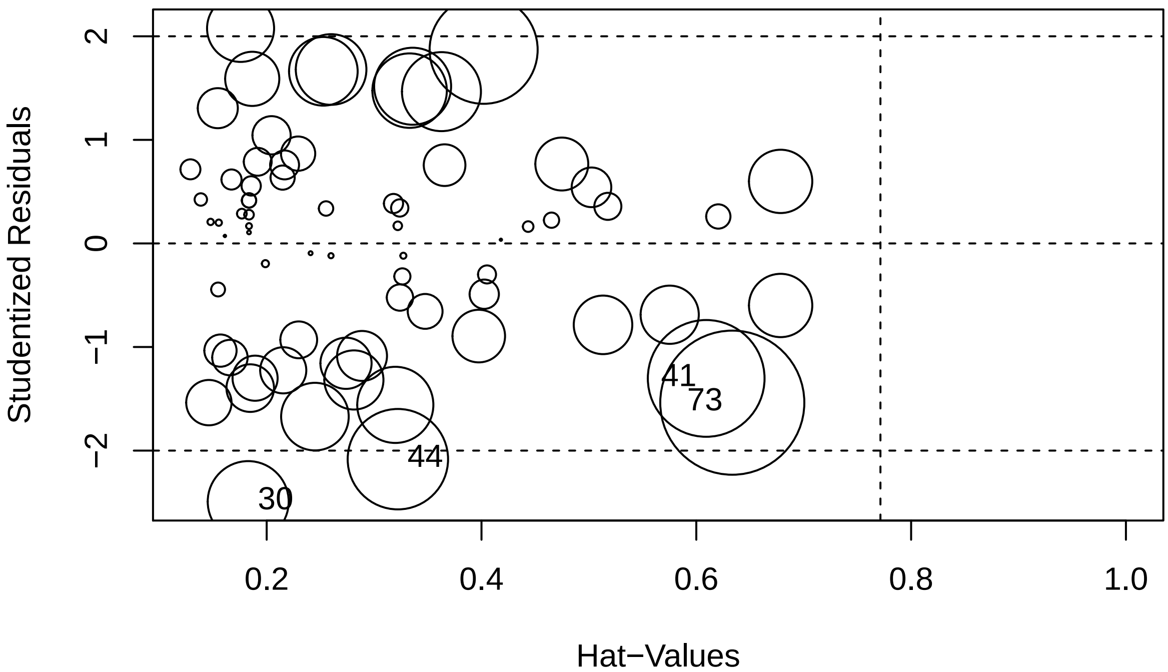
```
##    Price  Size Lot Bath Bed Year Garage    School
## 1  388.0 2.180   4    3   4 1940      0      High
## 2  450.0 2.054   5    3   4 1957      2      High
## 3  386.0 2.112   5    2   4 1955      2      High
## 4  350.0 1.442   6    1   2 1956      1      Alex
## 5  155.5 1.800   1    2   4 1994      1      Alex
## 6  220.0 1.965   5    2   3 1940      1      Alex
## 7  239.5 1.800   4  1.1   4 1958      1   StLouis
## 9  269.9 1.922   4  2.1   4 1965      2   StLouis
## 10 238.8 1.920   5  2.1   3 1968      2   StLouis
## 15 319.0 1.855   4    2   4 1925      2 NotreDame
## 16 339.0 1.928   5    3   3 1972      2   StMarys
## 18 275.0 1.528   3  2.1   3 1975      2   StMarys
## 20 277.0 1.440   3    2   3 1948      2      High
## 21 299.0 1.994   8    3   2 2005      2   StLouis
## 22 185.0 1.860   4    2   5 1953      0   StLouis
## 28 330.0 1.894   2    2   3 1950      1      High
## 31 325.0 1.661   4  1.1   3 1974      2 Stratford
## 32 259.9 1.708   4    2   4 1937      0 Stratford
## 33 324.5 1.800   2  2.1   3 2000      1 NotreDame
## 34 359.9 1.968   3    3   3 1980      2 NotreDame
## 35 252.5 1.888   2    2   6 1920      0      High
## 36 269.0 1.800   3    3   5 1947      0   StMarys
```

```
## 37 235.0 1.712   5    1    3 1948       1      High
## 39 319.9 1.920   7  2.1    3 2004       2   StLouis
## 40 355.0 2.129   5    2    3 2003       2   StLouis
## 41 285.0 2.235   7  3.1    4 1975       2 Stratford
## 42 242.0 2.279   5  2.1    4 1949       1   StLouis
## 43 197.0 1.850   3    2    4 1964       1   StMarys
## 44 281.8 1.886   4    2    3 2005       2      High
## 45 259.0 2.080   2    2    4 1969       2   StMarys
## 46 189.5 1.906   3  2.1    3 1975       0   StMarys
## 47 339.9 2.048   5    2    2 1993       2   StMarys
## 49 295.0 2.028   4    3    5 1992       2   StMarys
## 50 222.5 1.794   4    2    3 1960       0 Stratford
## 51 199.0 2.016   3    2    5 1963       1   StMarys
## 52 385.5 1.904   4  1.1    3 1919       1      High
## 53 230.0 2.234   3    3    4 1972       2   StMarys
## 54 285.0 2.269   3  3.1    4 1905       0      High
## 55 243.0 1.810   4    3    3 1961       2   StMarys
## 56 217.0 1.875   4    2    3 1961       0   StMarys
## 57 259.9 1.683   5  2.1    3 1979       1 NotreDame
## 58 349.5 2.080   4  2.1    3 2005       2 NotreDame
## 64 374.5 1.712   5  1.1    3 1908       2      High
## 66 350.0 2.020   7    2    3 1976       2 NotreDame
## 68 299.0 1.743   3    2    3 1988       2 NotreDame
## 69 285.0 1.878   5  2.1    3 1975       2 Stratford
## 71 259.0 2.208   4    3    3 1979       2   StLouis
## 72 249.9 2.081   5  2.1    4 1970       1 NotreDame
## 73 215.0 2.044   1  1.1    4 1936       0   StLouis
## 74 435.0 2.253  11    2    3 1979       2   StMarys
## 76 349.5 2.896   4    3    5 1979       2 Stratford
```

```r
leverage_points=as.numeric(which(hatvalues(model)>((2*25)/length(data$Price))))
leverage_points
```

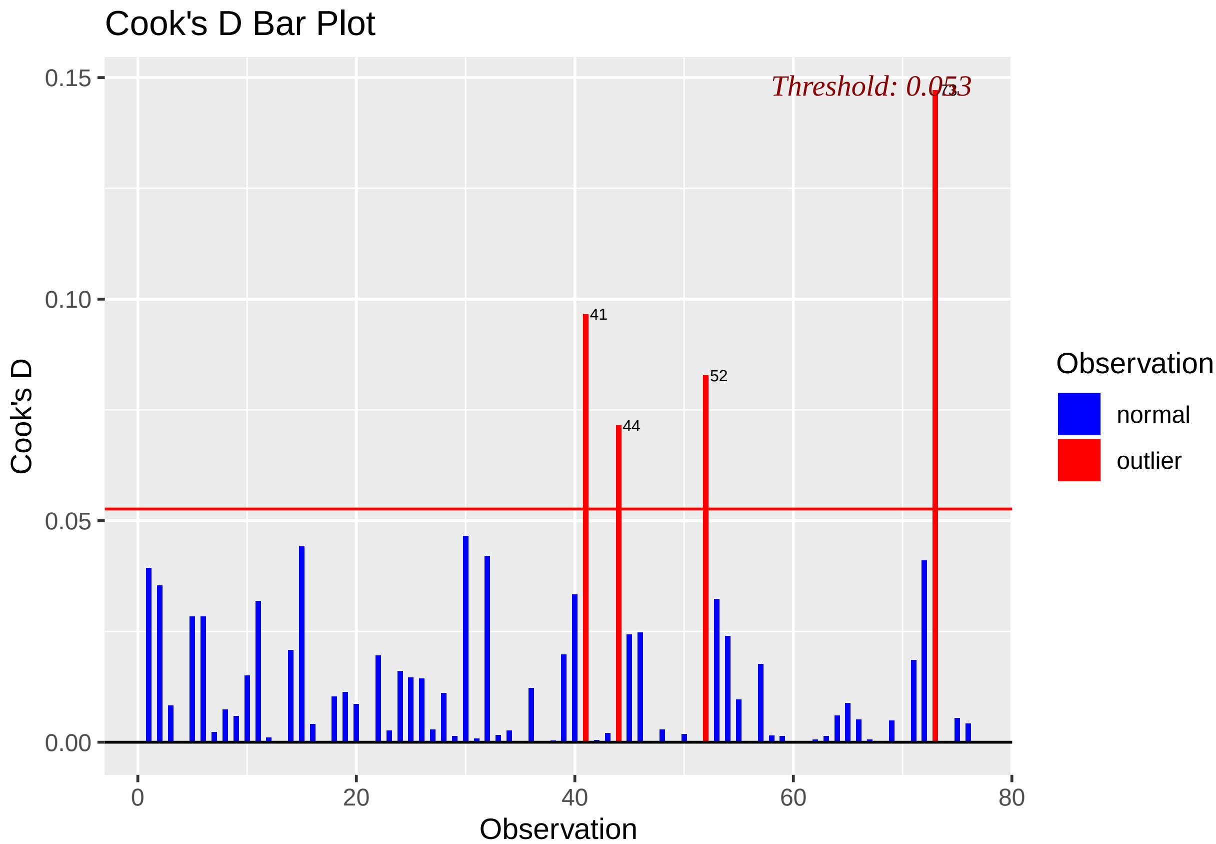
```
## [1]  4  5  6 21 35 37 47 74
```

```r
#q2
influencePlot(model)
```

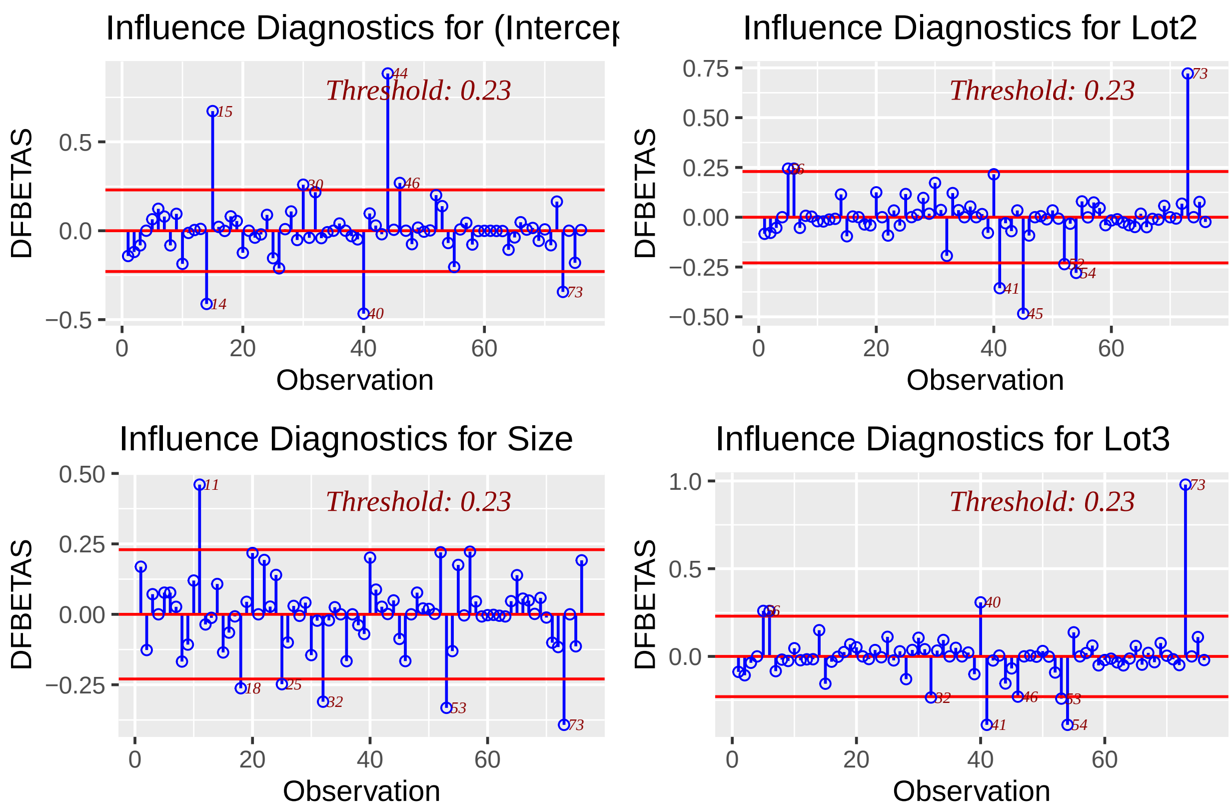



```
##        StudRes       Hat       CookD
## 4          NaN 1.0000000         NaN
## 21         NaN 1.0000000         NaN
## 30 -2.492660 0.1827863 0.04652220
## 41 -1.303051 0.6091925 0.09665145
## 44 -2.083053 0.3221395 0.07150069
## 73 -1.537421 0.6335028 0.14722402
```

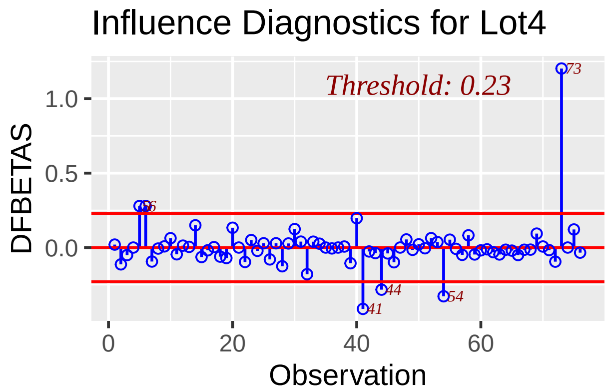
```
ols_plot_cooksd_bar(model)
```

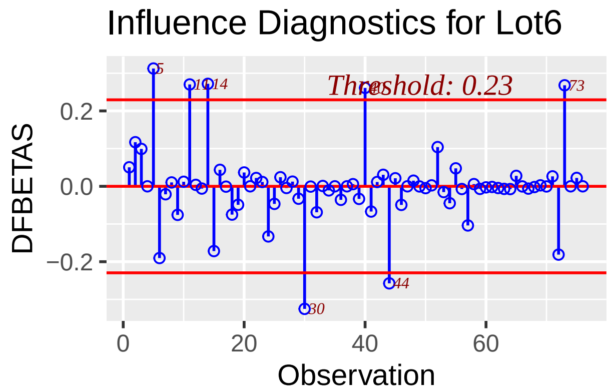


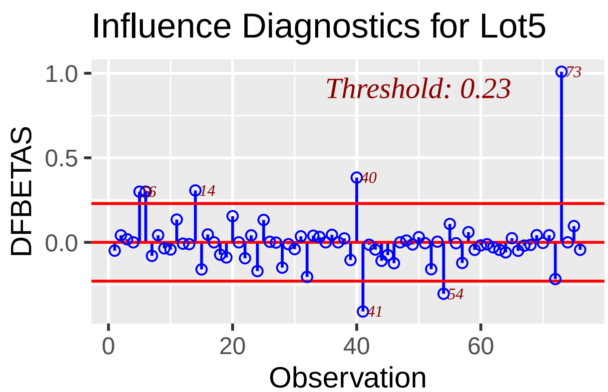

```
ols_plot_dfbetas(model)
```

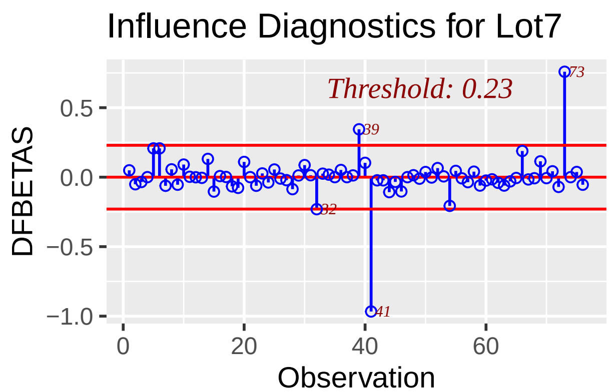


Influence Diagnostics for (Intercep
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for Lot2
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for Size
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for Lot3
Threshold: 0.23
DFBETAS
Observation

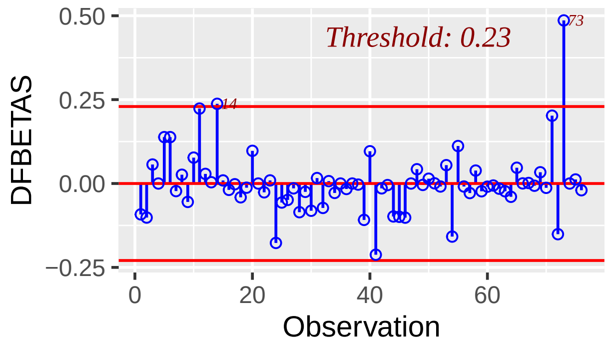

Influence Diagnostics for Lot4
Threshold: 0.23
DFBETAS
Observation
73
6
41
44
54


Influence Diagnostics for Lot6
Threshold: 0.23
DFBETAS
Observation
5
14
73
30
44

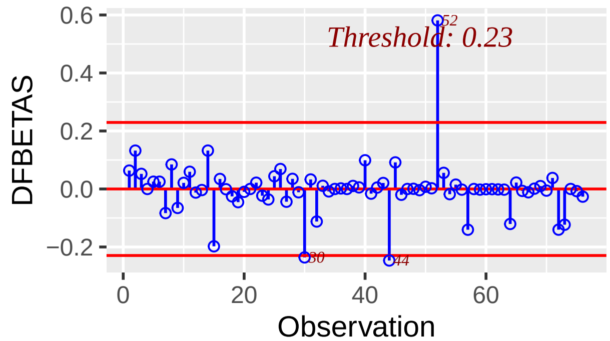

Influence Diagnostics for Lot5
Threshold: 0.23
DFBETAS
Observation
73
6
14
40
41
54


Influence Diagnostics for Lot7
Threshold: 0.23
DFBETAS
Observation
73
39
32
41

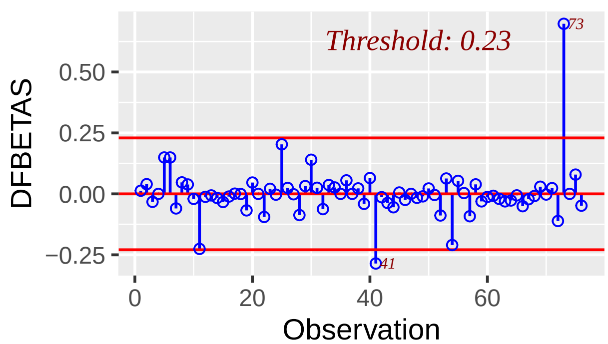
Influence Diagnostics for Lot8



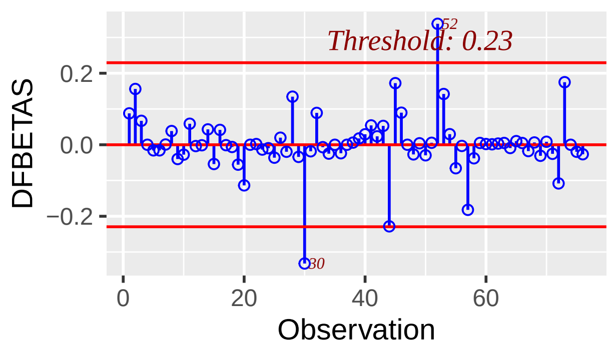
Influence Diagnostics for Bath1.1

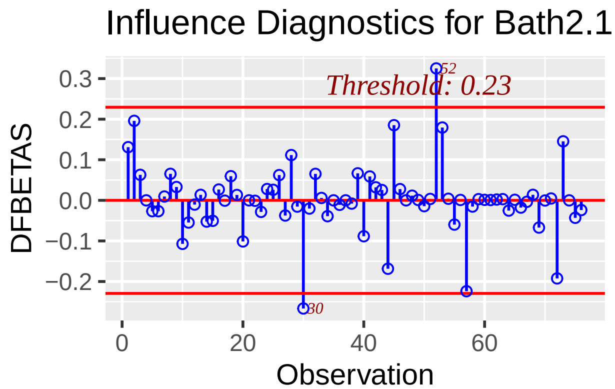


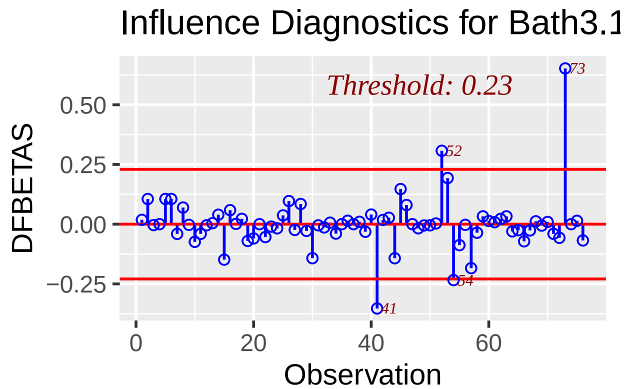
Influence Diagnostics for Lot11

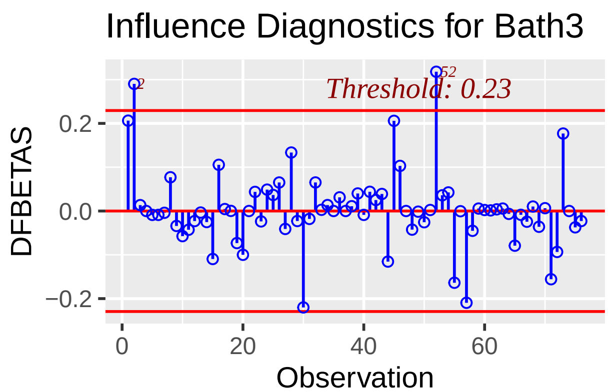


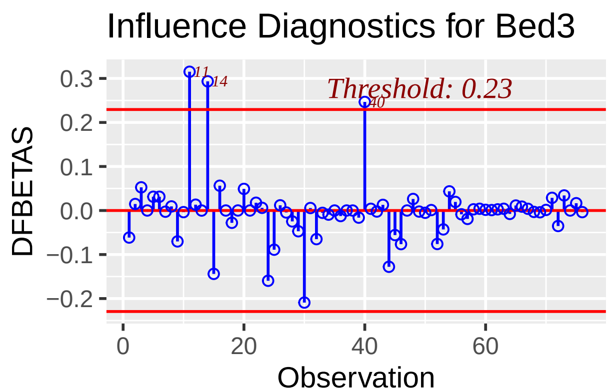
Influence Diagnostics for Bath2

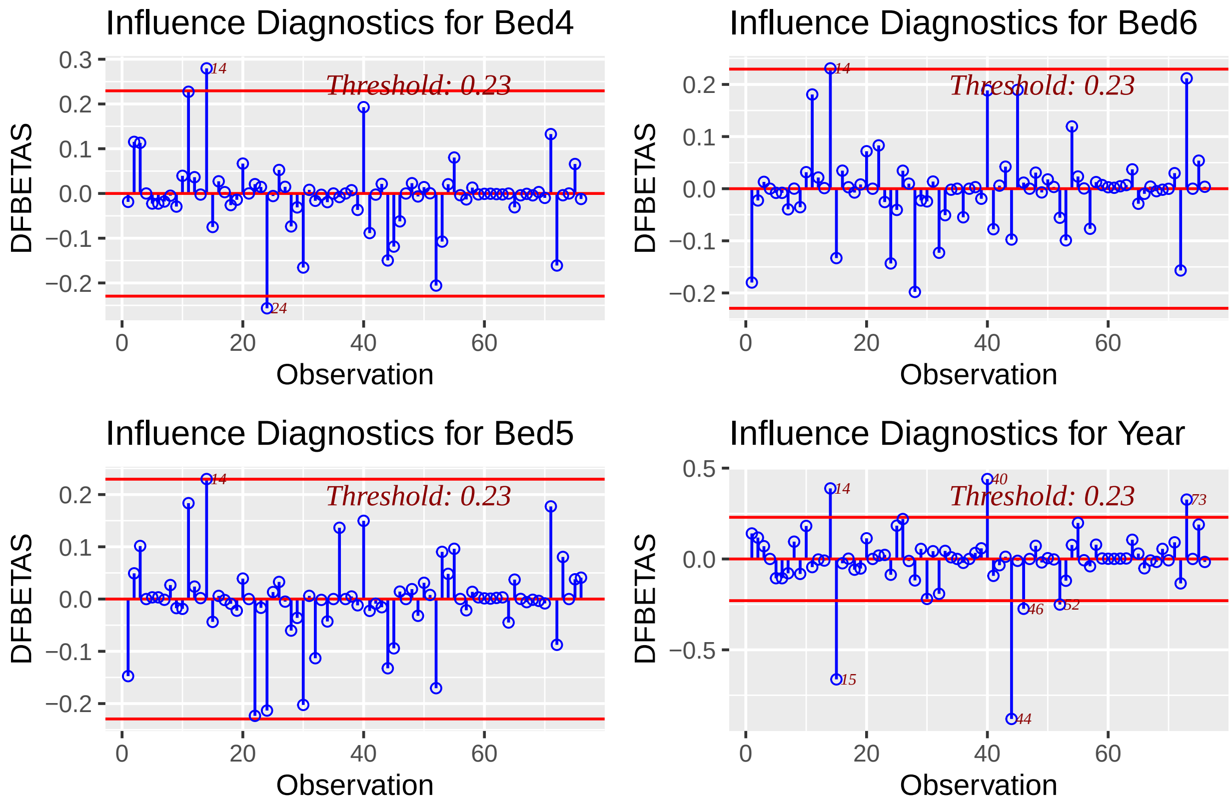



Influence Diagnostics for Bath2.1
Threshold: 0.23
DFBETAS
Observation
0.3
0.2
0.1
0.0
−0.1
−0.2
0
20
40
60
52
30

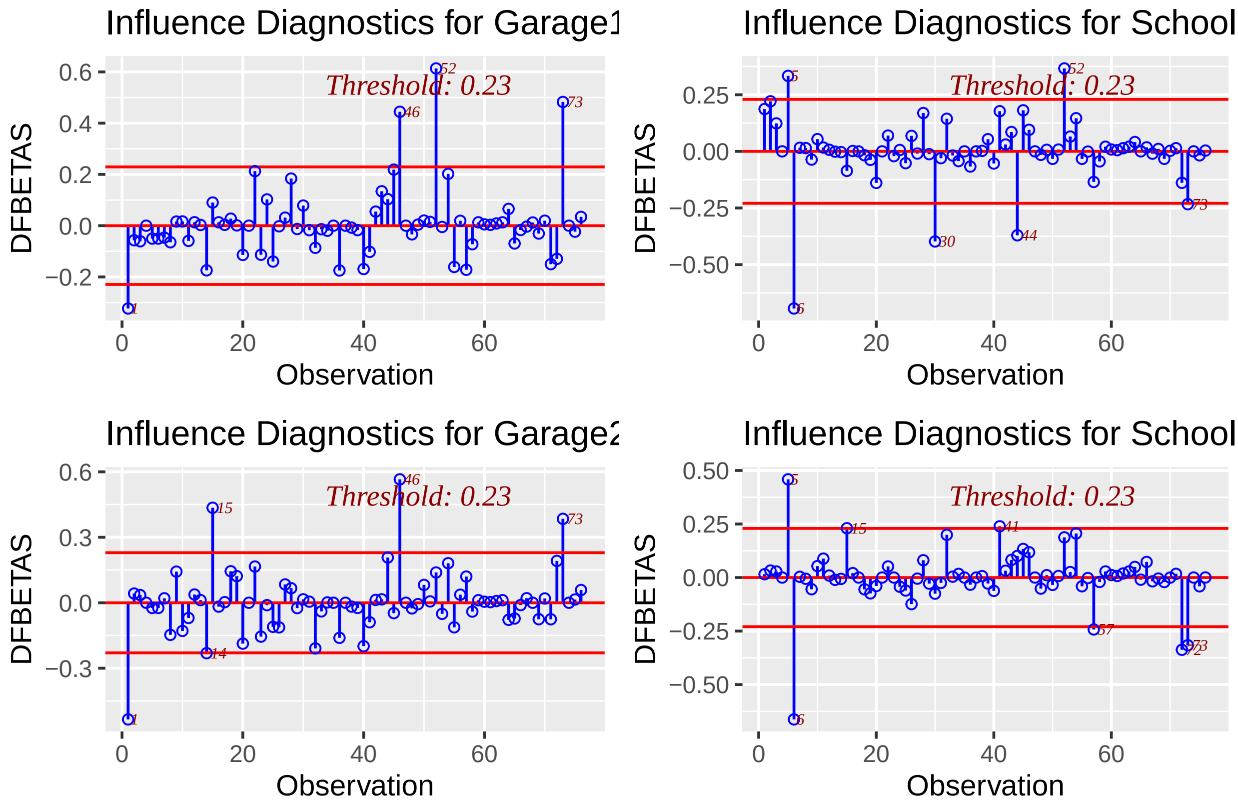

Influence Diagnostics for Bath3.1
Threshold: 0.23
DFBETAS
Observation
0.50
0.25
0.00
−0.25
0
20
40
60
73
52
54
41


Influence Diagnostics for Bath3
Threshold: 0.23
DFBETAS
Observation
0.2
0.0
−0.2
0
20
40
60
2
52

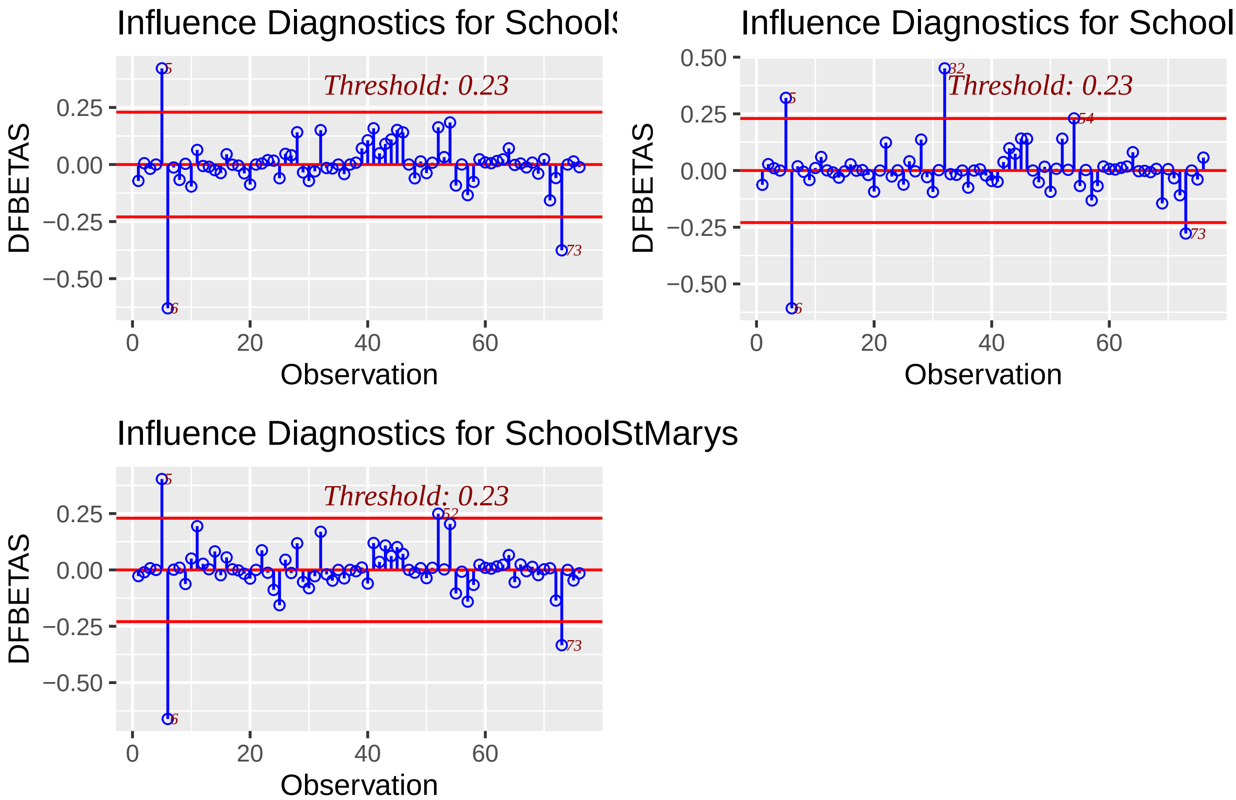

Influence Diagnostics for Bed3
Threshold: 0.23
DFBETAS
Observation
0.3
0.2
0.1
0.0
−0.1
−0.2
0
20
40
60
11
14
40


Influence Diagnostics for Bed4
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for Bed6
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for Bed5
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for Year
Threshold: 0.23
DFBETAS
Observation

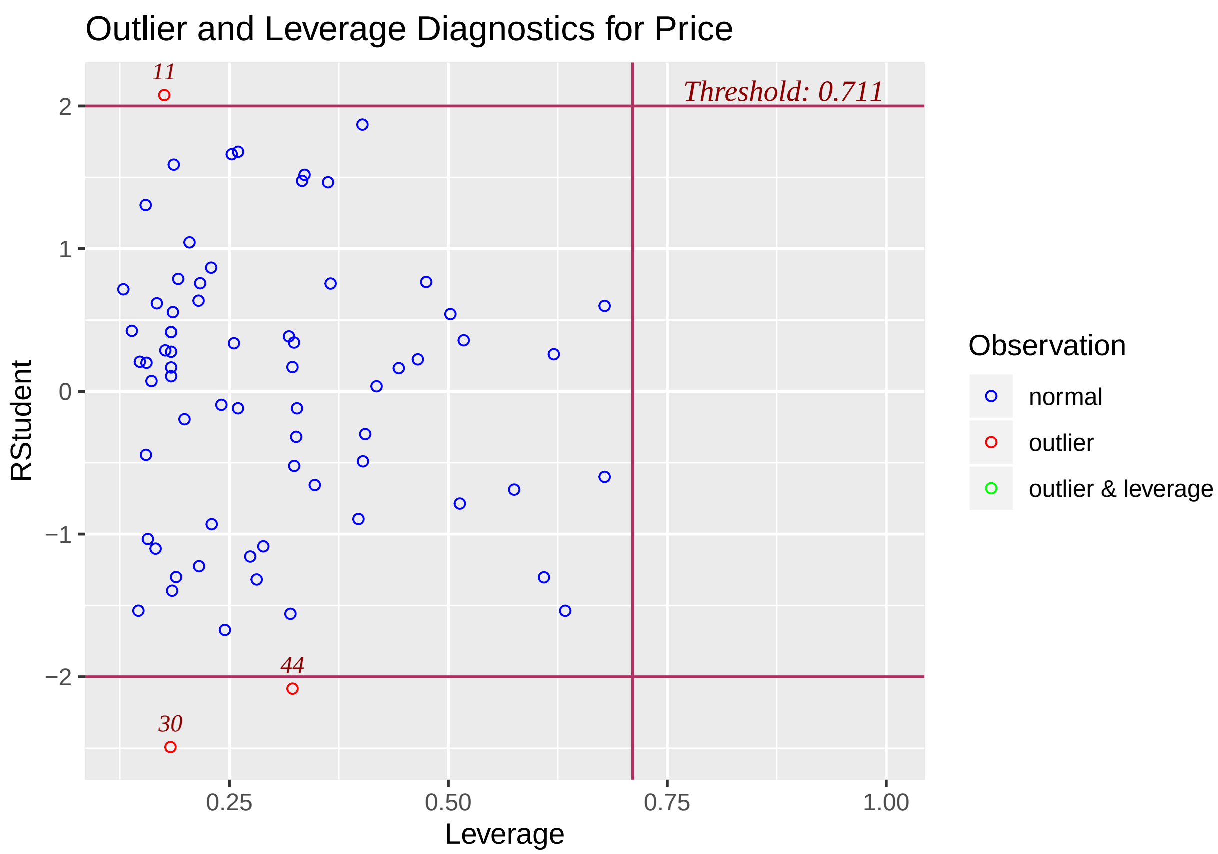
page 6 of 7


Influence Diagnostics for Garage1
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for SchoolI
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for Garage2
Threshold: 0.23
DFBETAS
Observation
Influence Diagnostics for SchoolI
Threshold: 0.23
DFBETAS
Observation

page 7 of 7




```
ols_plot_resid_lev(model)
```




```
#q3
outlierTest(model)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 30 -2.49266           0.016182           NA
```

1) leverage point:A leverage point is one with an unusual X-value. It affects the modelsummary statistics (e.g.R2,SSE, etc…) but has little effect on the estimates of the regression coefficients. High leverage points have the potential to affect the fit of the model.

- This point does not affect the estimates of the regression coefficients.
- It affects the model summary statistics e.g., 2 R , standard errors of regression coefficients etc.

Leverage Points : 4 5 6 21 35 37 47 74

2) An influential point has an usual Y -value also. It has a noticeable impact on the model coefficients: it 'pulls' the regression model in its direction.

- It has a noticeable impact on the model coefficients.
- It pulls the regression model in its direction.

Influential Points : 30,44,41,73

3) An outlier is an extreme observation. Typically points further than, say, three or four standard deviations from the mean are considered as “outliers”. An outlier is an observation where the response does not correspond to the model fitted to the bulk of the data.

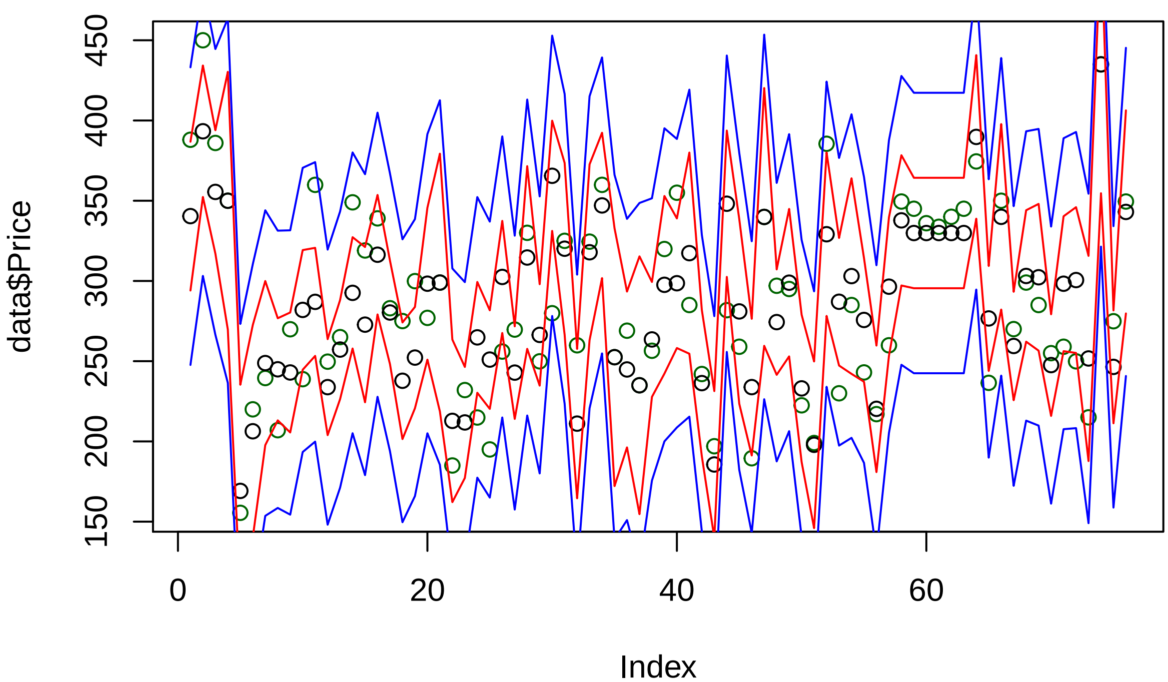
We can see 3 outliers in the graph. We can neglect them as they are very close to the actual data. We can also drop these points if the distance increases.

Check data entry.Investigate whether the context provide an explanation. Some scientific discoveries come from noticing unexpected irregularities. Exclude the outlier, see its influence. Perhaps present analysis with and without the outlier. When the model is changed, try to reintroduce the outlier.

```
ci=predict(model,level = 0.95,interval = "confidence")
pi=predict(model,level = 0.95,interval = "predict")
```

```
## Warning in predict.lm(model, level = 0.95, interval = "predict"): predictions on current data refer t
```

```
plot(data$Price,col="dark green")
points(fitted(model),col="black")
lines(ci[,2],col="red")
lines(ci[,3],col="red")
lines(pi[,2],col="blue")
lines(pi[,3],col="blue")
```



The above plot is a good estimate for calculating house prices. All the predicted values are within the confidence interval.