Predictive Analysis Assignment

Kshitij Baranwal and 19200078

15/11/2019

R Markdown

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------
------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------
------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(car)
```

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

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

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

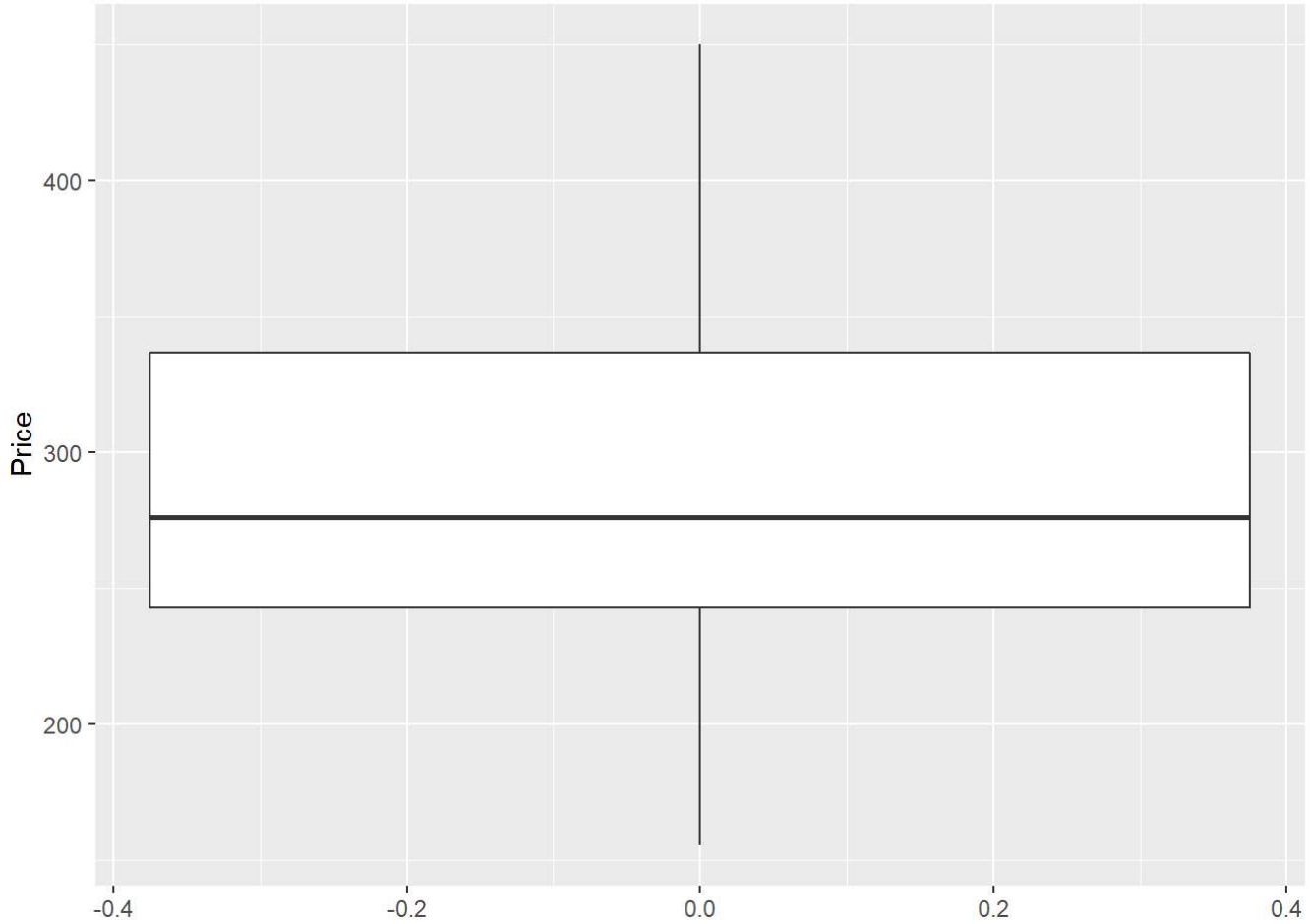
```
library(dplyr)
data=read.csv('House.csv')
names(data)[1] <- "Price"
head(data)
```

```
##   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
```

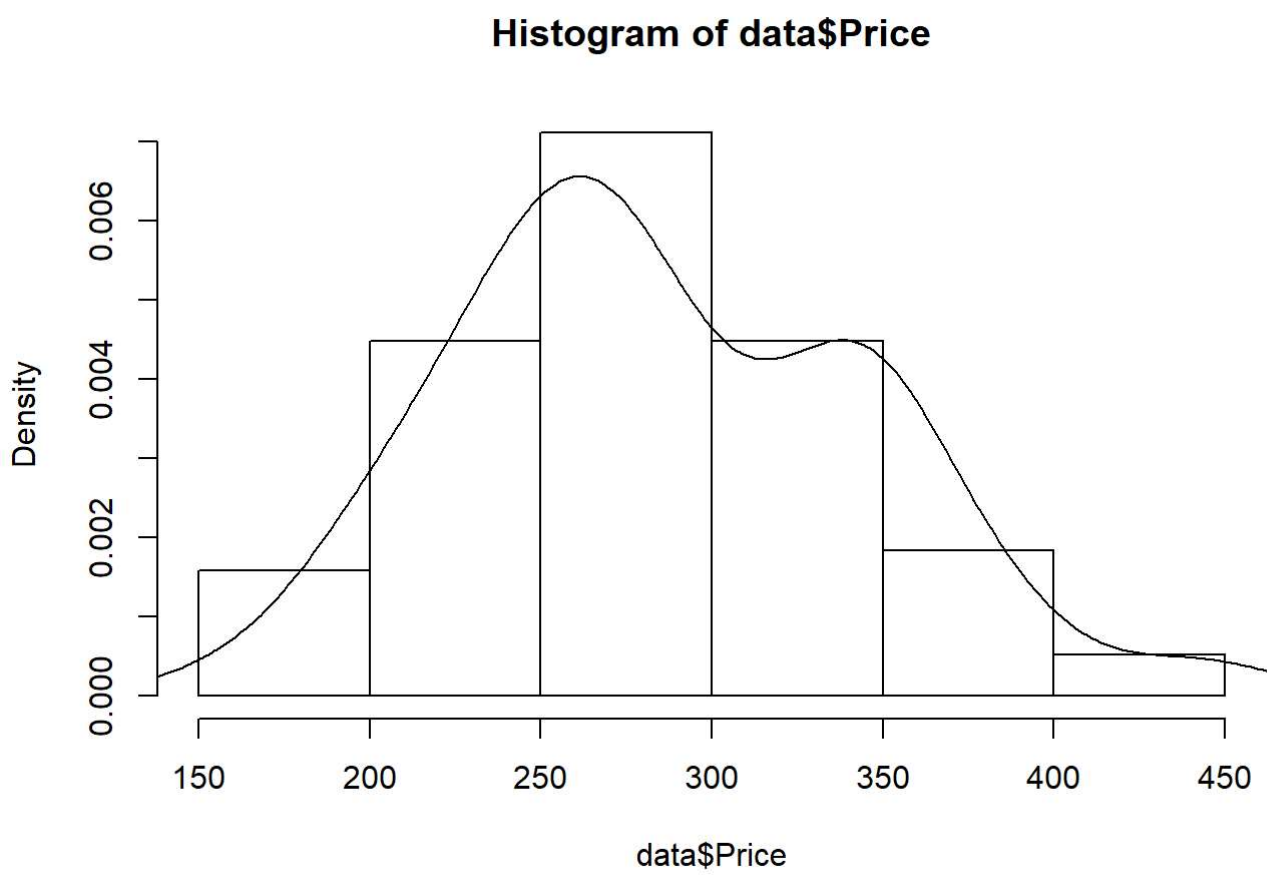
Exploratory Data Analysis:

Q1)

```
ggplot(data, aes(y=Price)) + geom_boxplot()
```




```
hist(data$Price,freq=FALSE)
lines(density(data$Price))
```

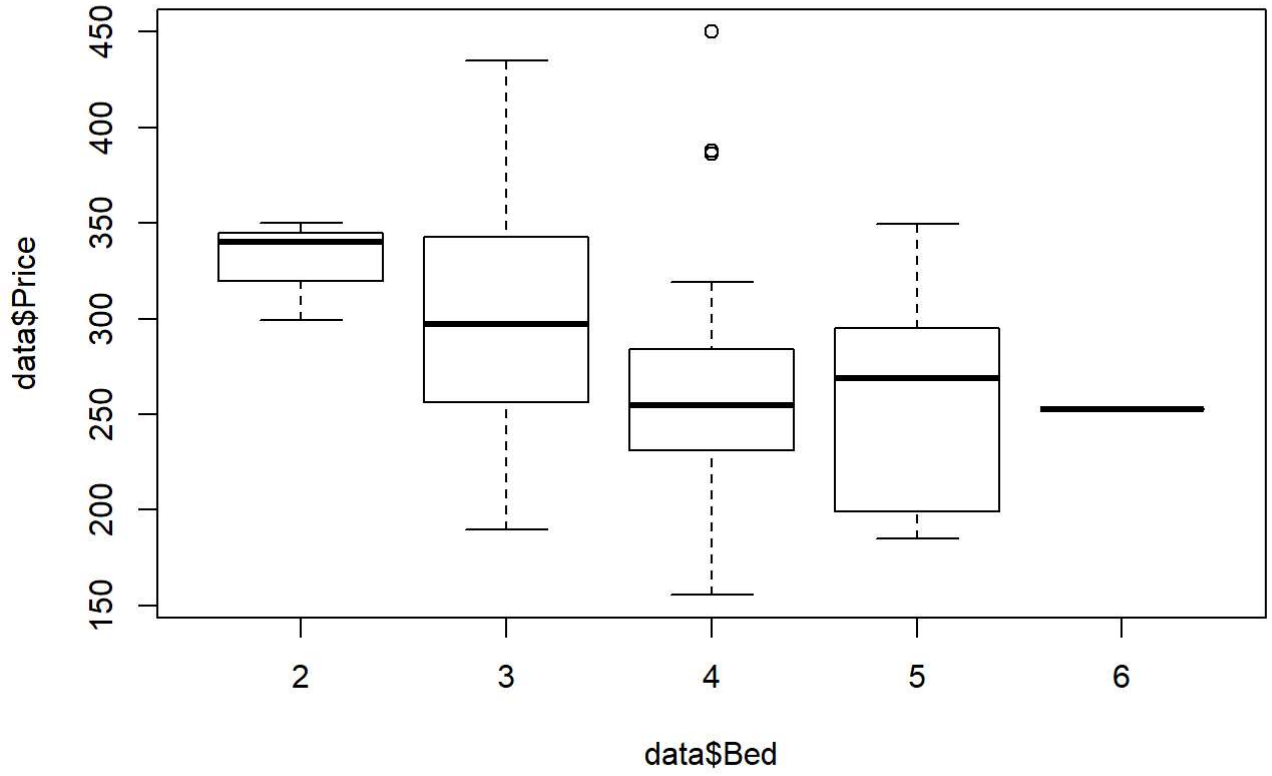



```
summary(data$Price)
```

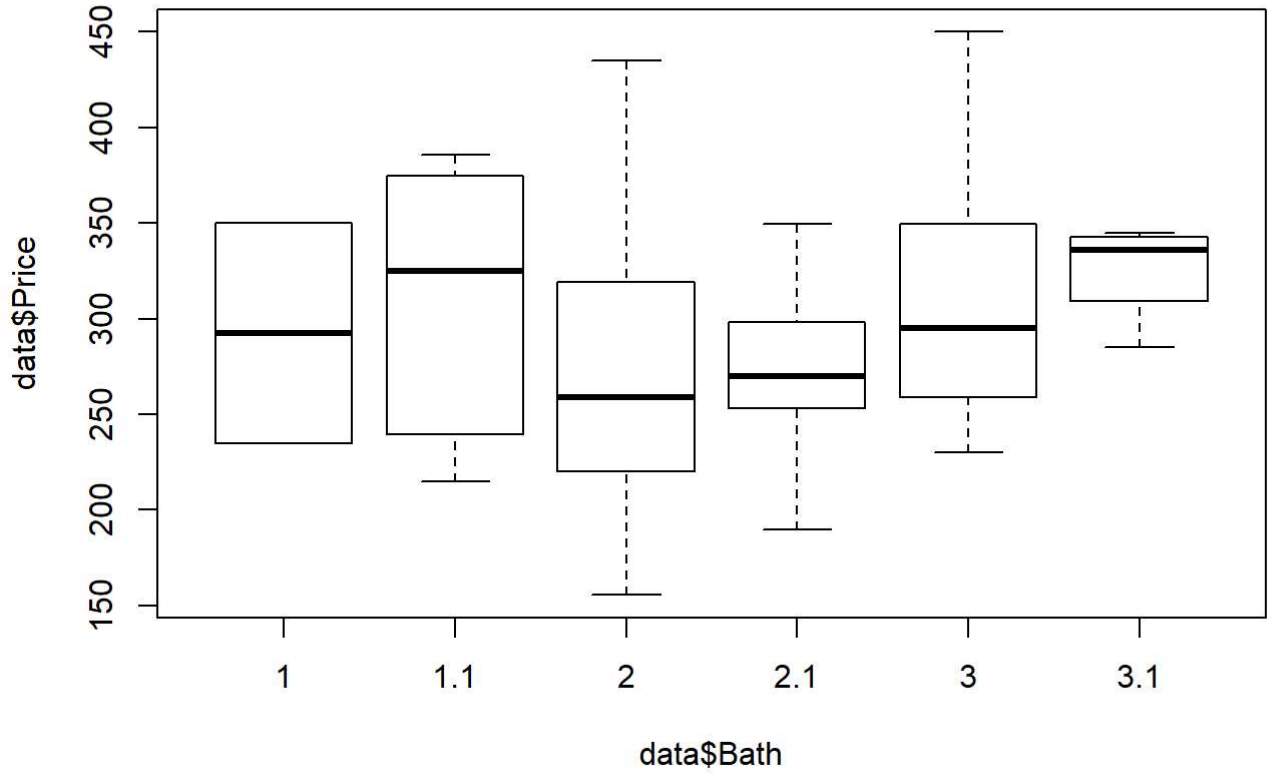
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   242.8   276.0   285.8   336.8   450.0
```

Q2)

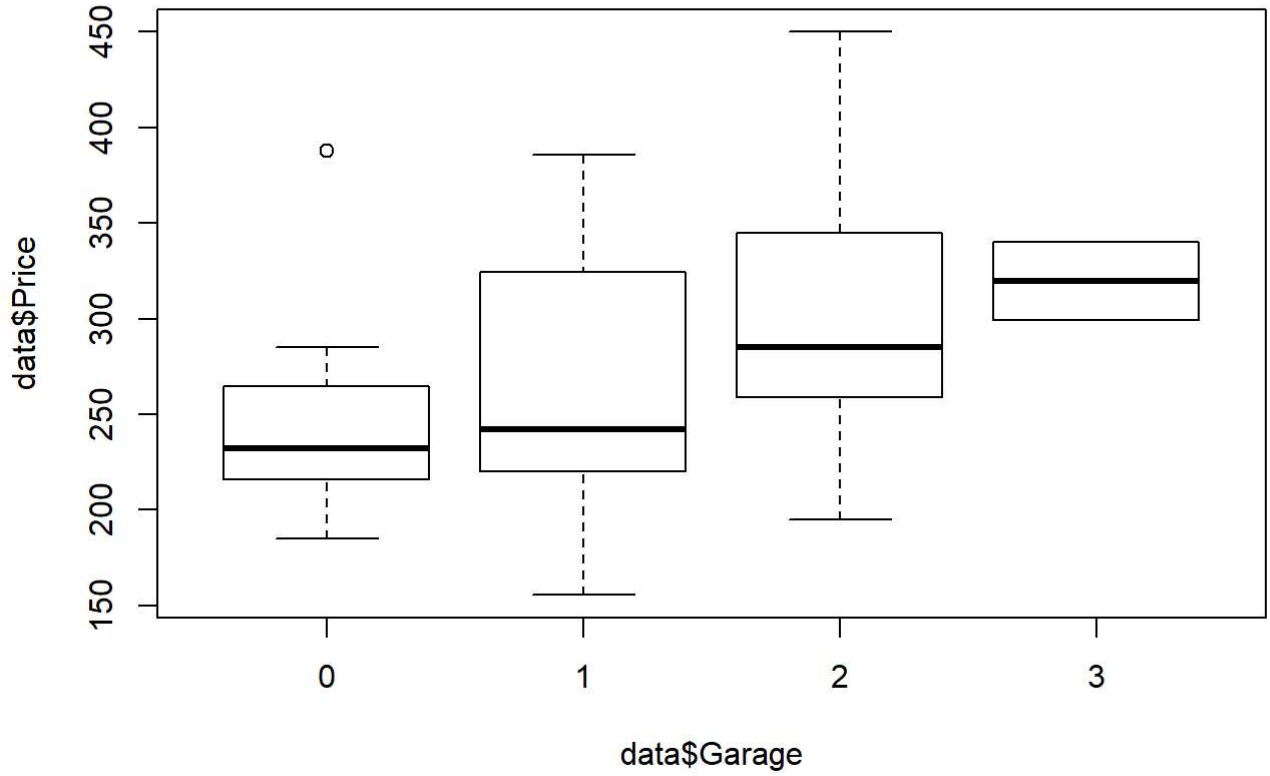
```
data$Bed <- factor(data$Bed)
data$Bath <- factor(data$Bath)
data$Garage <- factor(data$Garage)
data$School <- factor(data$School)
boxplot(data$Price~data$Bed)
```



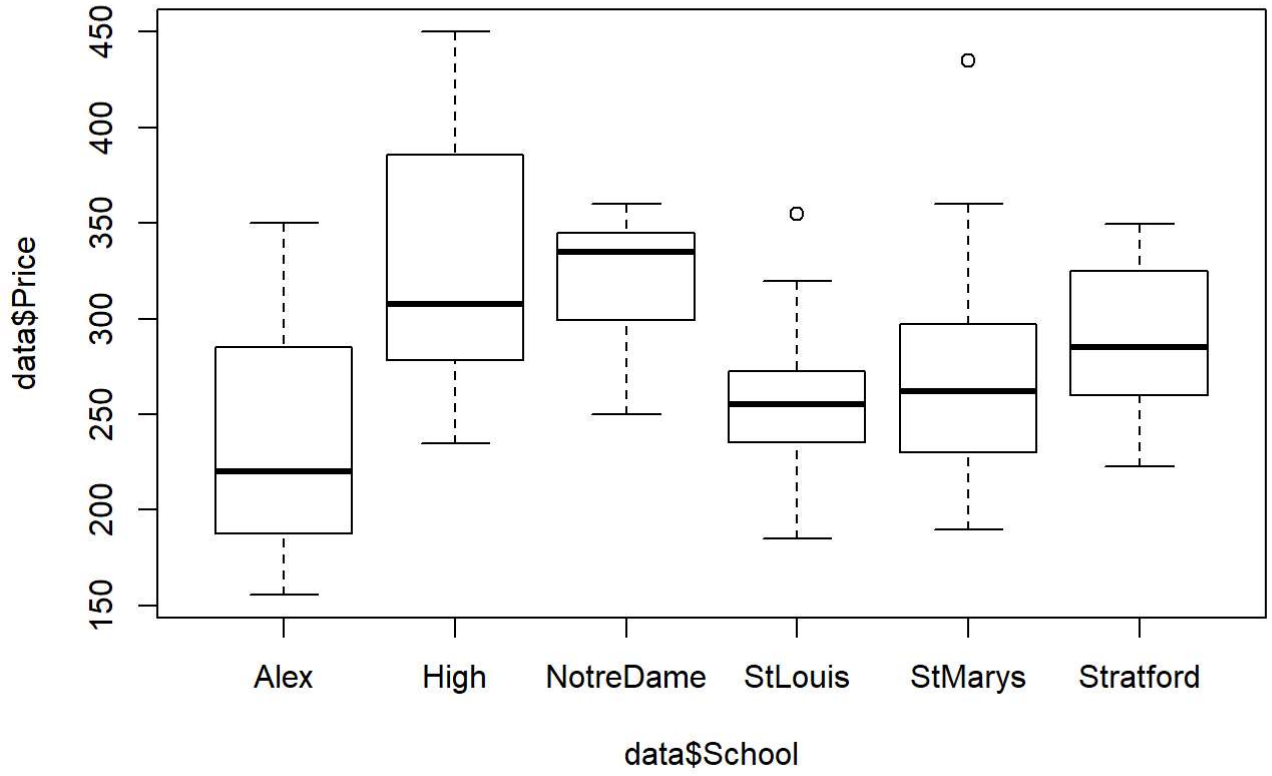

```
boxplot(data$Price~data$Bath)
```




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




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



Q3)

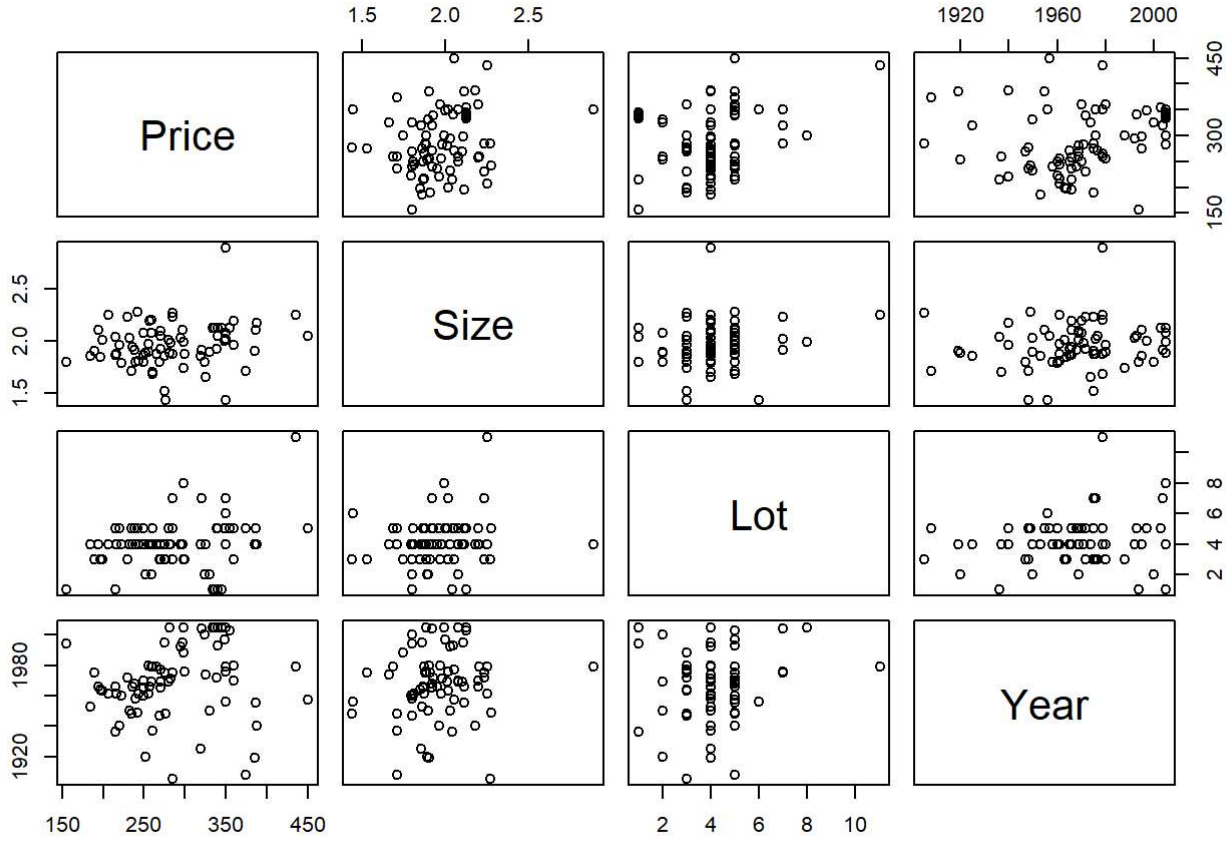
```
summary(data[,c(1,2,3,6)])
```

```
##      Price            Size            Lot              Year
##  Min.   :155.5   Min.   :1.440   Min.   : 1.000   Min.   :1905
##  1st Qu.:242.8   1st Qu.:1.861   1st Qu.: 3.000   1st Qu.:1958
##  Median :276.0   Median :1.966   Median : 4.000   Median :1970
##  Mean   :285.8   Mean   :1.970   Mean   : 3.987   Mean   :1969
##  3rd Qu.:336.8   3rd Qu.:2.107   3rd Qu.: 5.000   3rd Qu.:1980
##  Max.   :450.0   Max.   :2.896   Max.   :11.000   Max.   :2005
```

```
cor(data[,c(1,2,3,6)])
```

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

```
pairs(Price~Size+Lot+Year,data=data)
```



Regression Model:

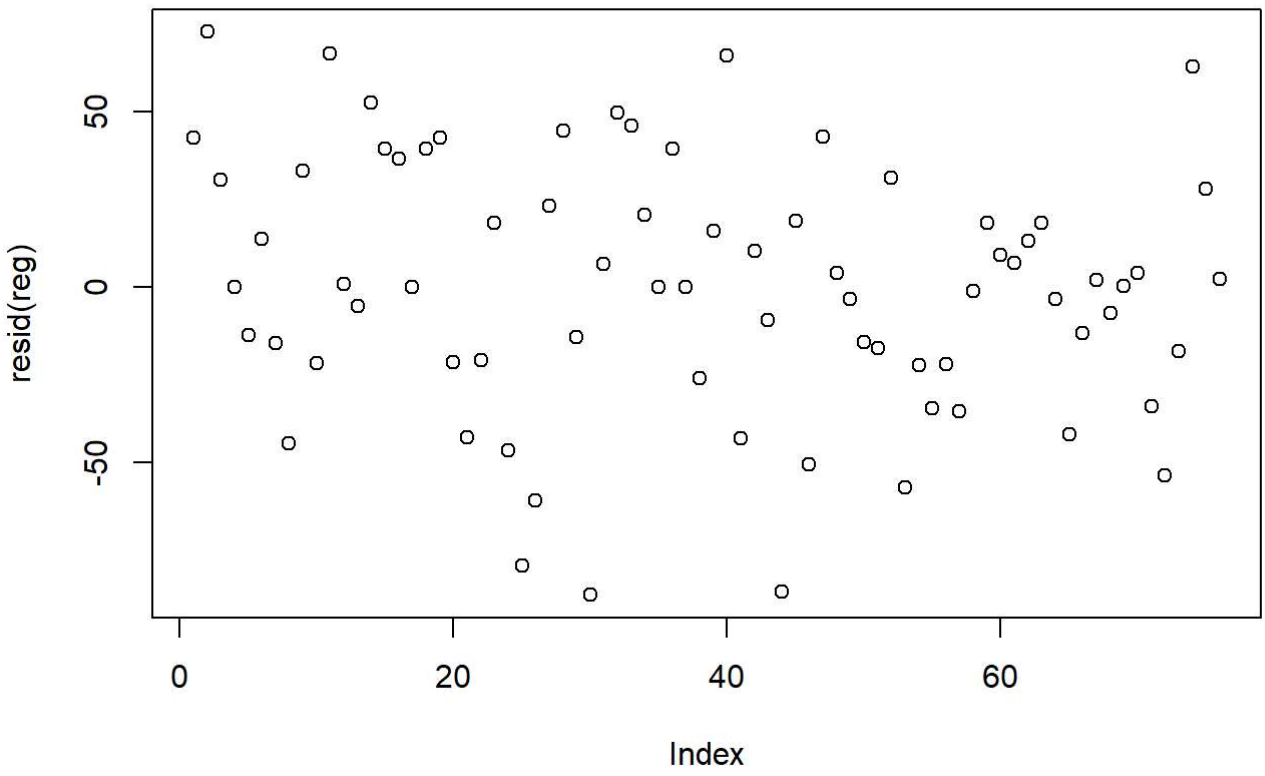
```
data$Size=data$Size-mean(data$Size)
data$Lot=data$Lot-mean(data$Lot)
data$Year=data$Year-mean(data$Year)
summary(data[,c(2,3,6)])
```

```
##       Size                Lot                Year          
##  Min.   :-0.530395   Min.   :-2.98684   Min.   :-64.40789  
##  1st Qu.:-0.109645   1st Qu.:-0.98684   1st Qu.:-11.65789  
##  Median :-0.003895   Median : 0.01316   Median :  0.09211  
##  Mean   : 0.000000   Mean   : 0.00000   Mean   :  0.00000  
##  3rd Qu.: 0.137105   3rd Qu.: 1.01316   3rd Qu.: 10.59211  
##  Max.   : 0.925605   Max.   : 7.01316   Max.   : 35.59211
```

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

```
##
## Call:
## lm(formula = Price ~ School + Bed + Bath + Garage + Lot + Size +
##     Year, data = data)
##
## 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 ***
## 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
## 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 **
## 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
## 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 *
## Lot               11.7701     3.7842   3.110  0.00296 **
## Size              59.4503    28.9813   2.051  0.04501 *
## Year               0.5567     0.3384   1.645  0.10565
## ---
## 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
```

```
plot(resid(reg))
```



ANOVA:

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## School     5  60570 12113.9  6.8265 5.119e-05 ***
## Bed        4  30180  7545.1  4.2519 0.0045381 **
## Bath       5  16741  3348.2  1.8868 0.1115268
## Garage     3  35965 11988.4  6.7558 0.0005831 ***
## Lot        1  19206 19206.4 10.8233 0.0017525 **
## Size       1   7938  7937.9  4.4732 0.0389754 *
## Year       1   4803  4802.6  2.7064 0.1056506
## Residuals 55  97599  1774.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
reg2=lm(Price~School+Bed+Bath+Garage+Lot+Size,data=data)
anova(reg,reg2)
```

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

Diagnostics:

Q1)

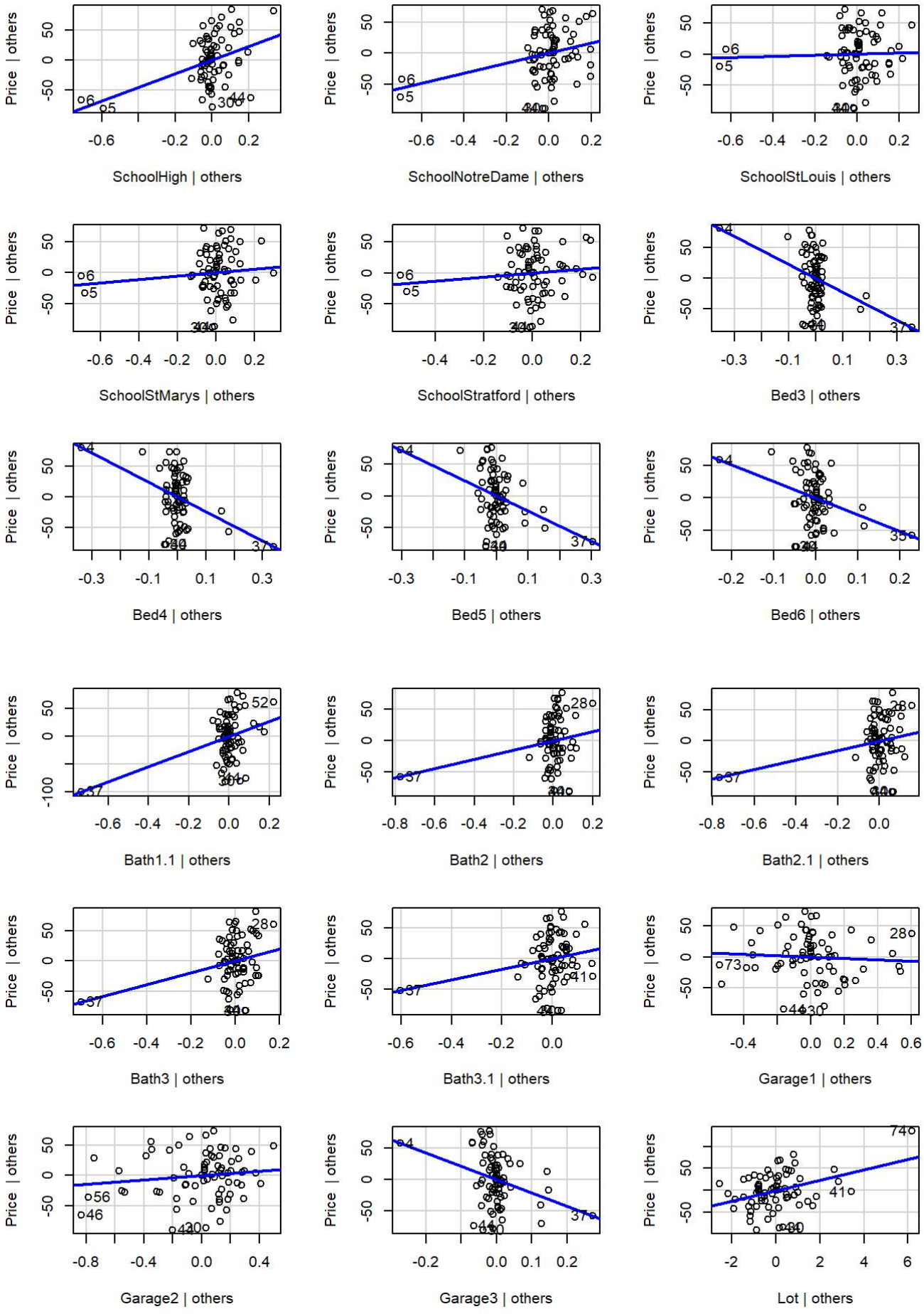
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

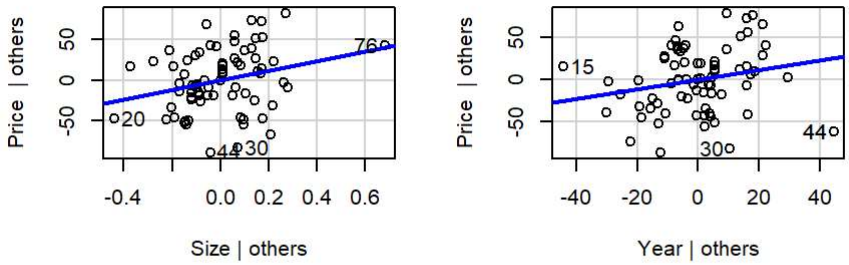
```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

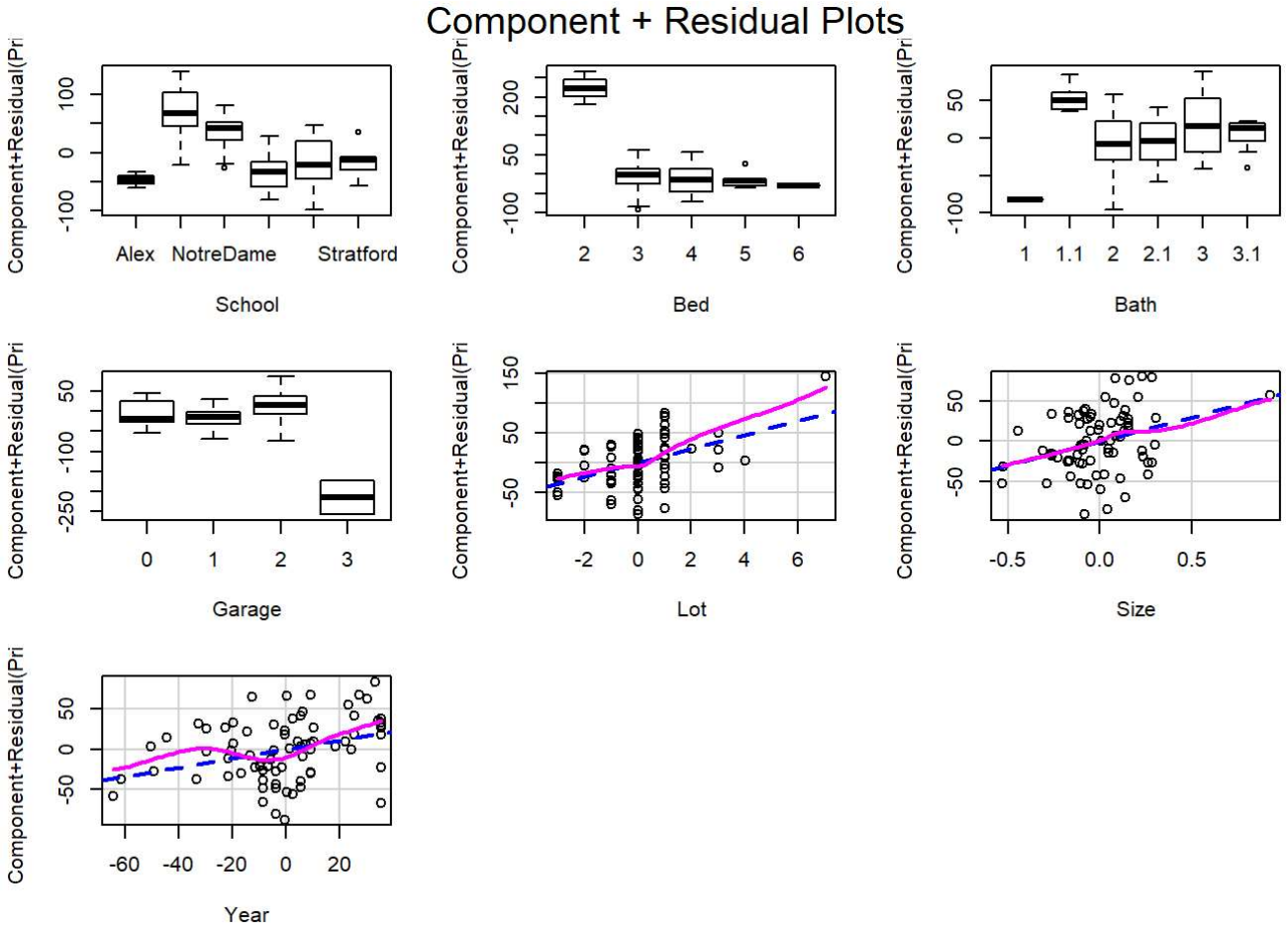
```
avPlots(reg)
```



Added-Variable Plots


Price | others
50
0
-50
76
20
44
30
-0.4
0.0
0.2
0.4
0.6
Size | others
Price | others
50
0
-50
15
44
30
-40
-20
0
20
40
Year | others

```
crPlots(reg)
```



Q2)

```
dwt(reg)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1836122      1.614157   0.038
##  Alternative hypothesis: rho != 0
```

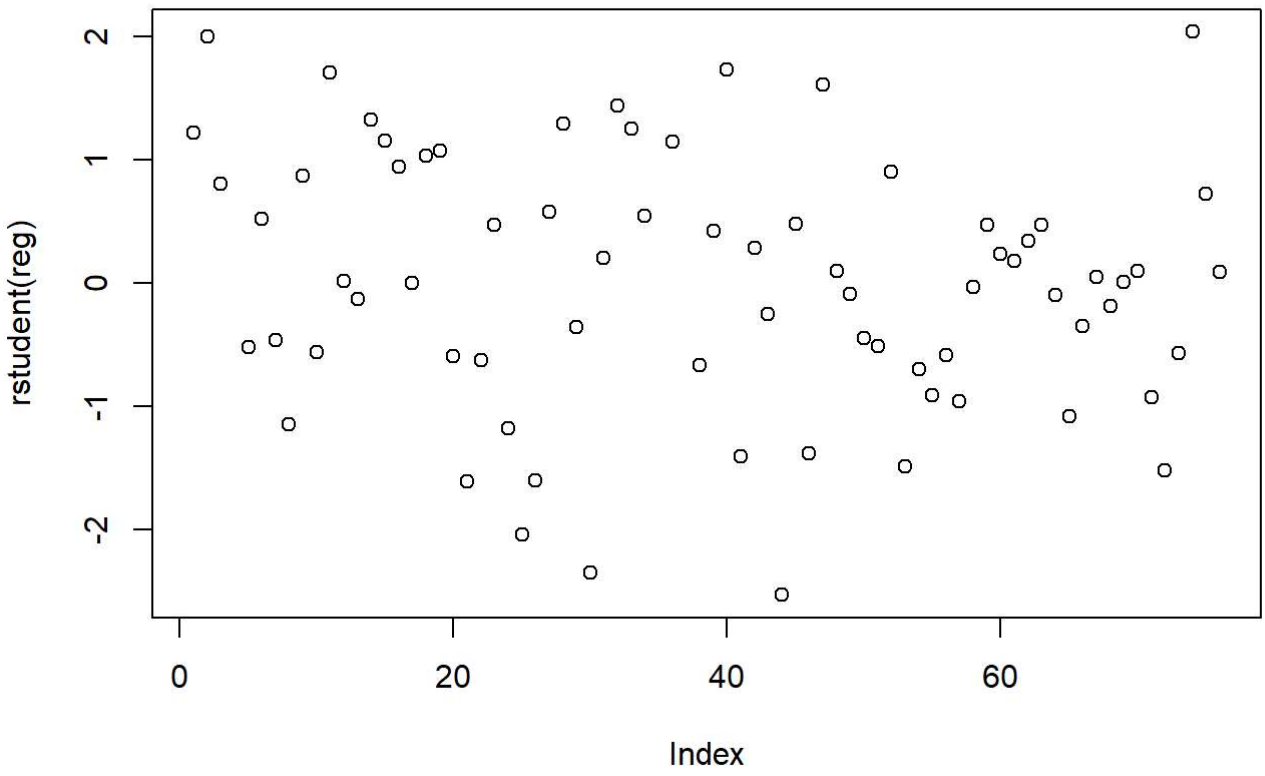
Q3)

```
vif(reg)
```

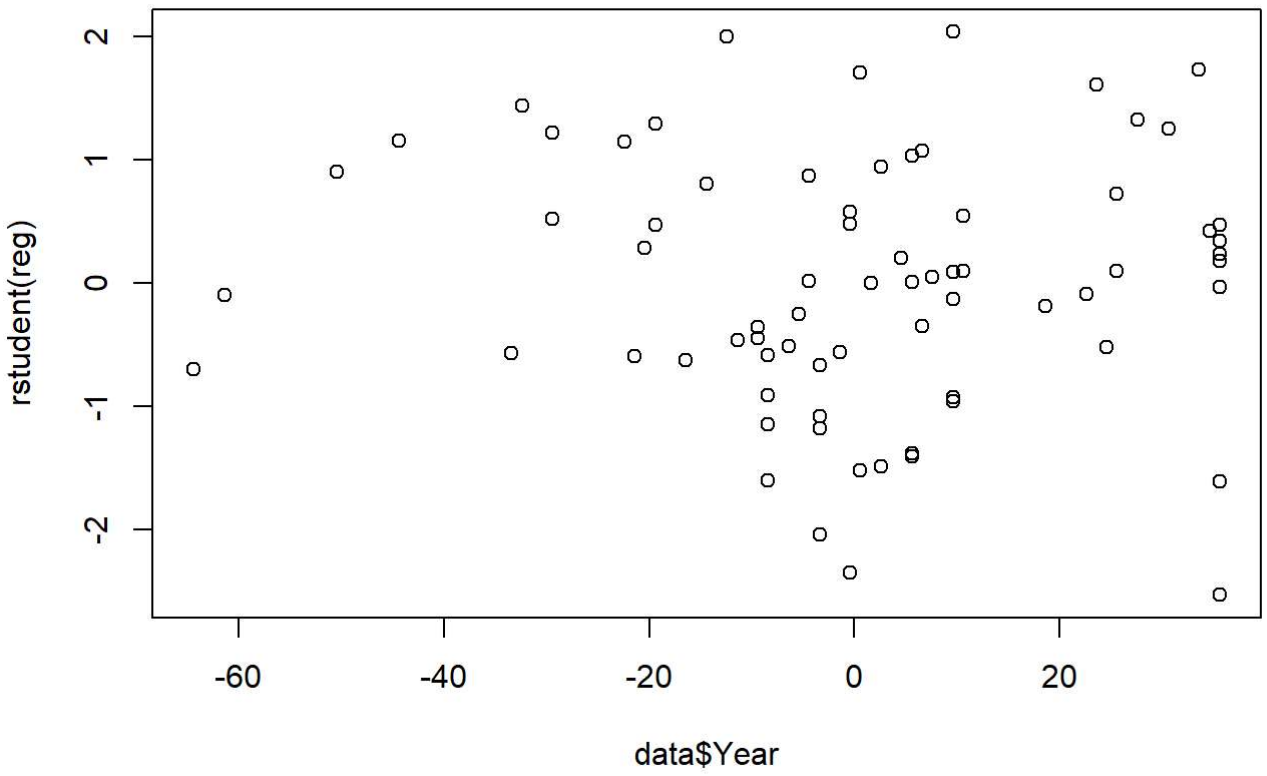
```
##             GVIF Df GVIF^(1/(2*Df))
## School  6.768538  5        1.210736
## Bed    20.215797  4        1.456168
## Bath    9.757455  5        1.255838
## Garage 19.811449  3        1.644950
## Lot     1.654167  1        1.286144
## Size    1.601785  1        1.265616
## Year    2.671175  1        1.634373
```

Q4)

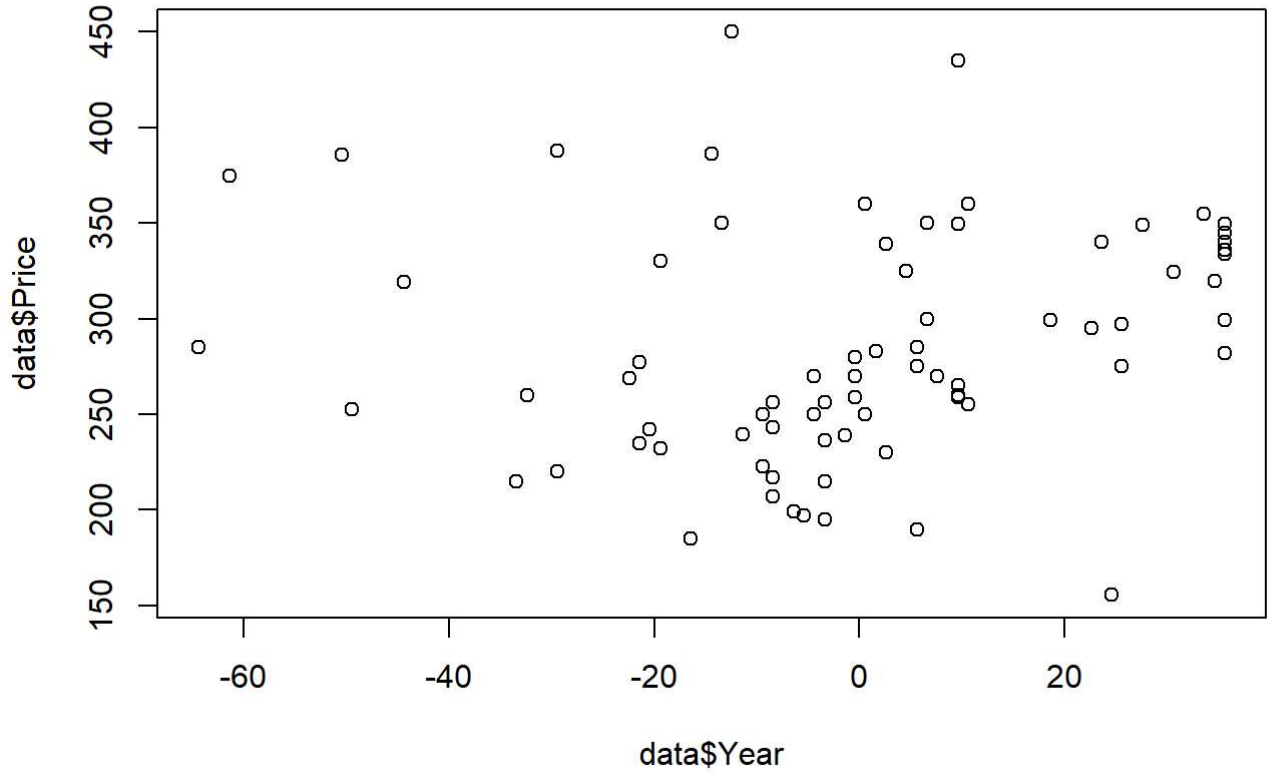
```
plot(rstudent(reg))
```



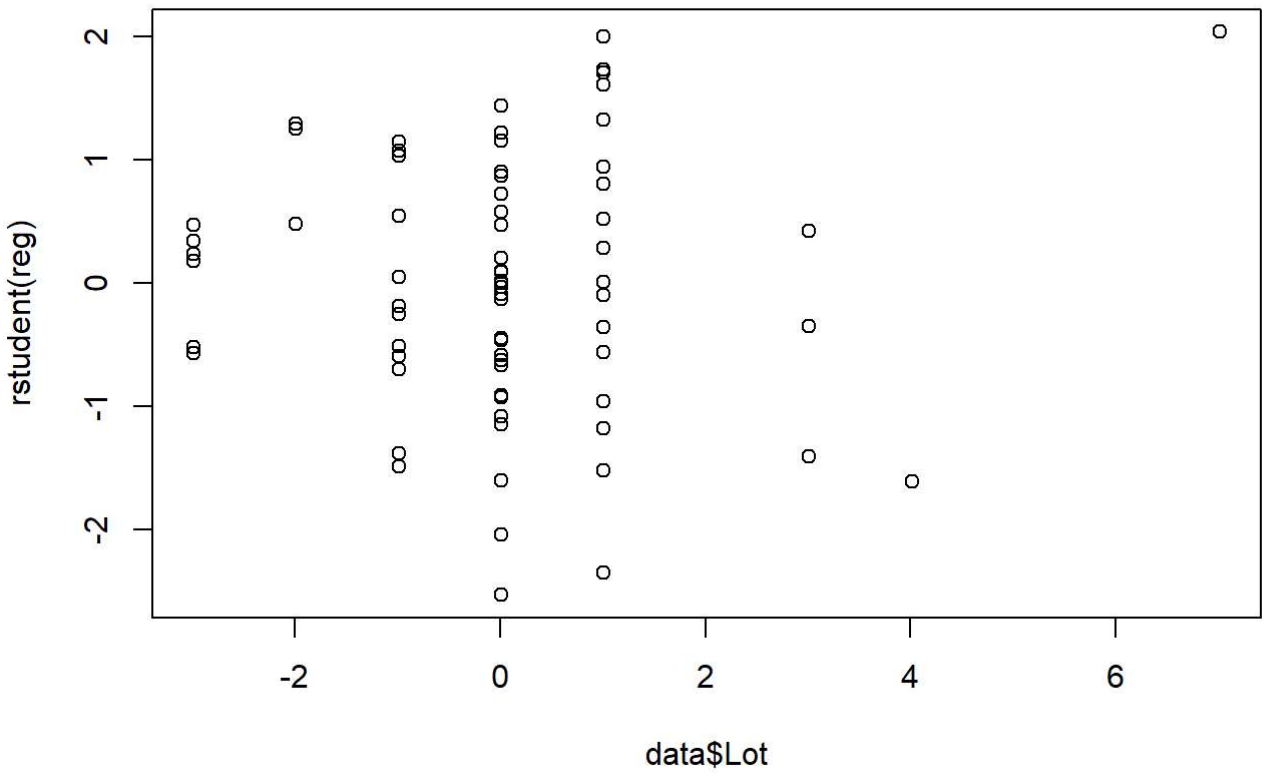

```
plot(data$Year,rstudent(reg))
```



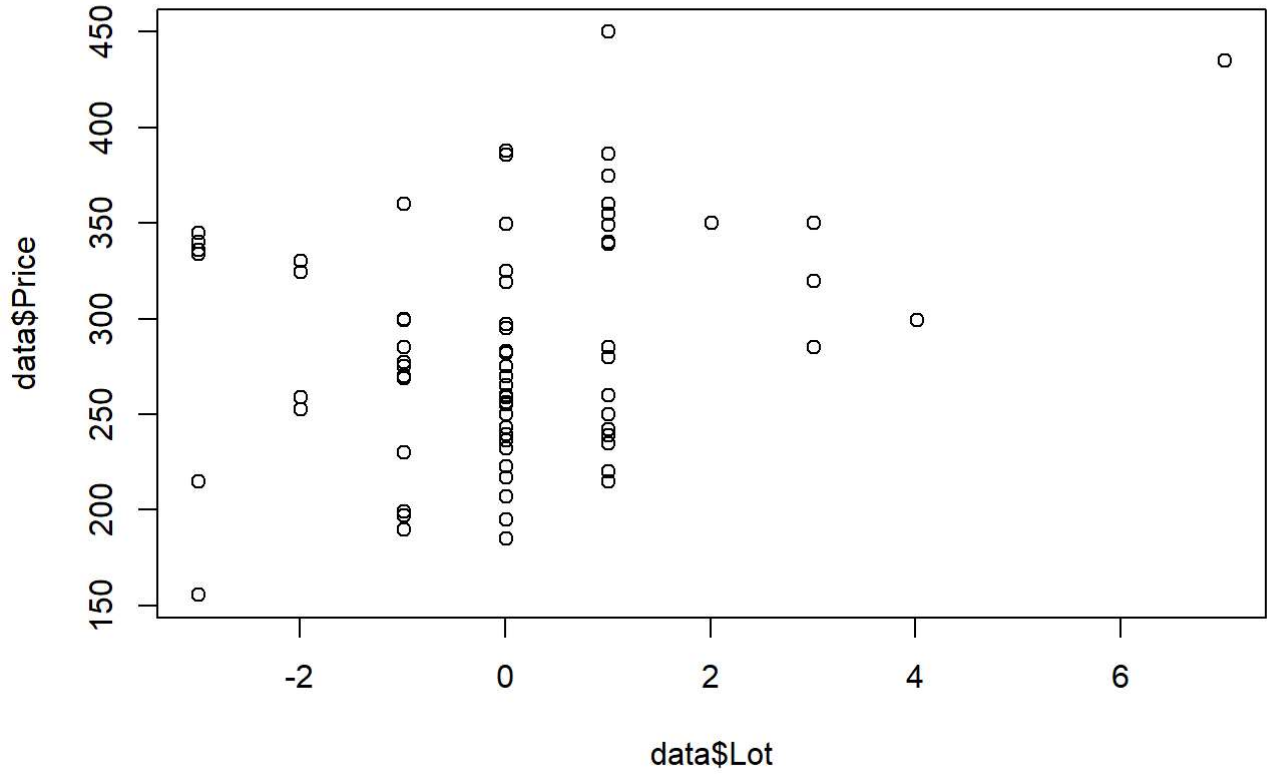

```
plot(data$Year,data$Price)
```



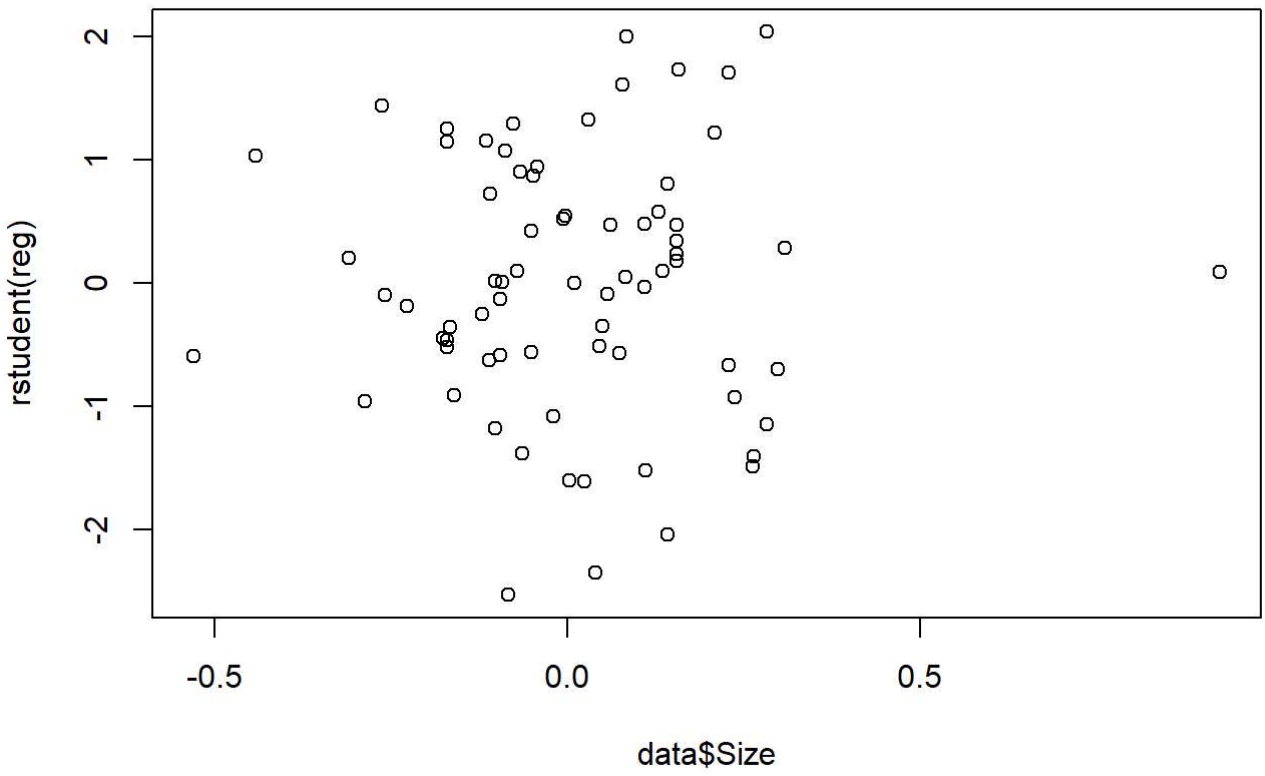

```
plot(data$Lot,rstudent(reg))
```



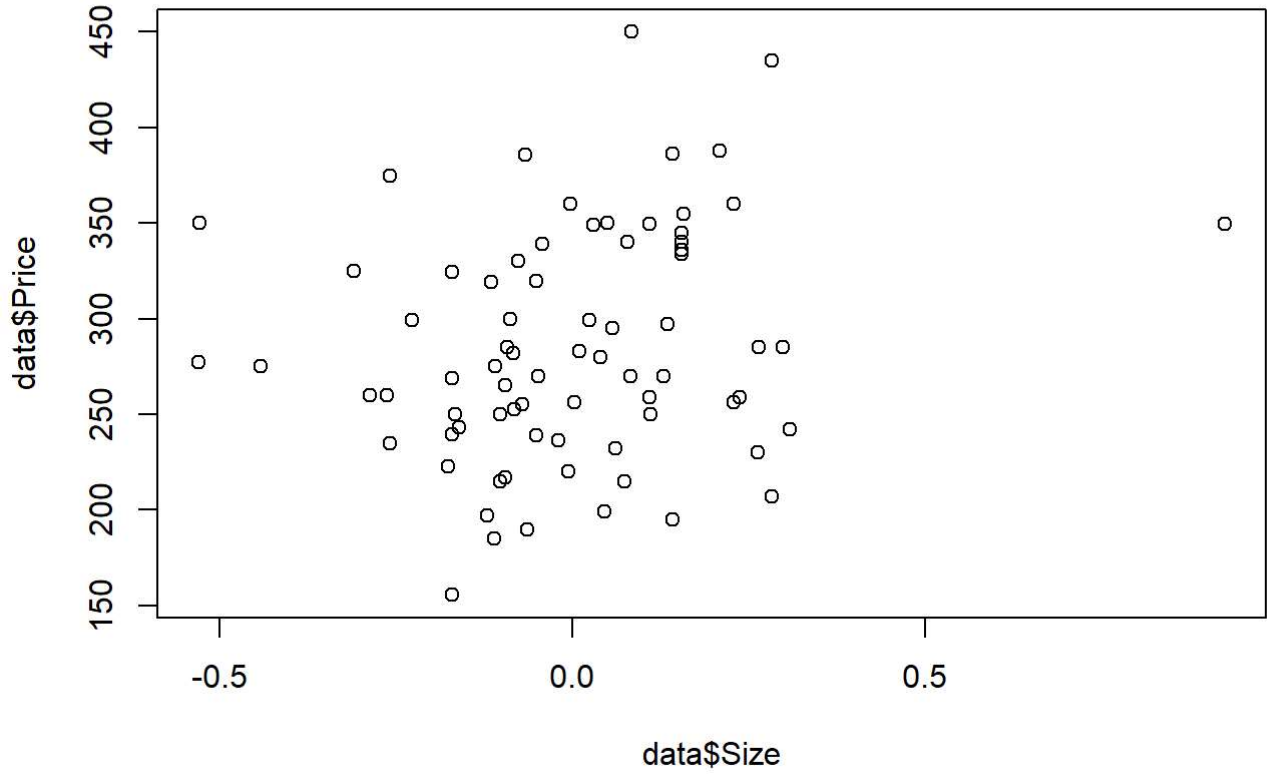

```
plot(data$Lot,data$Price)
```



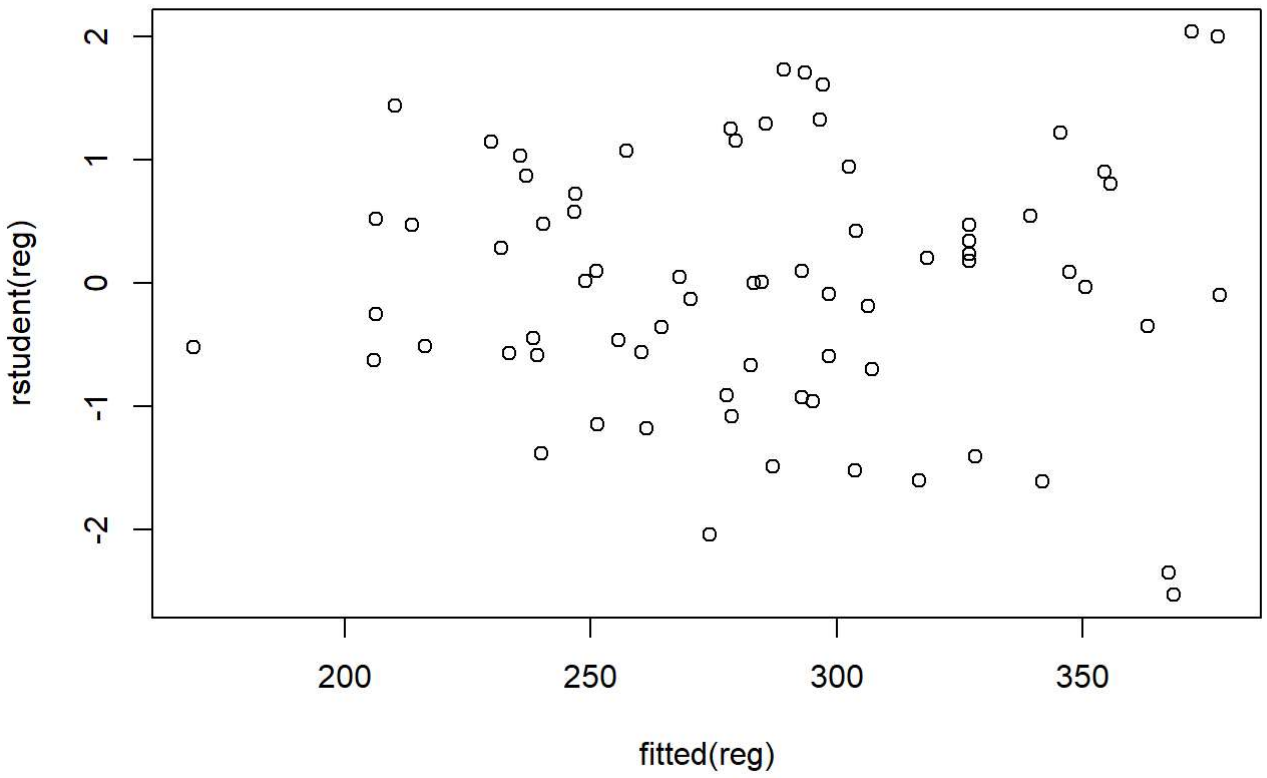

```
plot(data$Size,rstudent(reg))
```




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

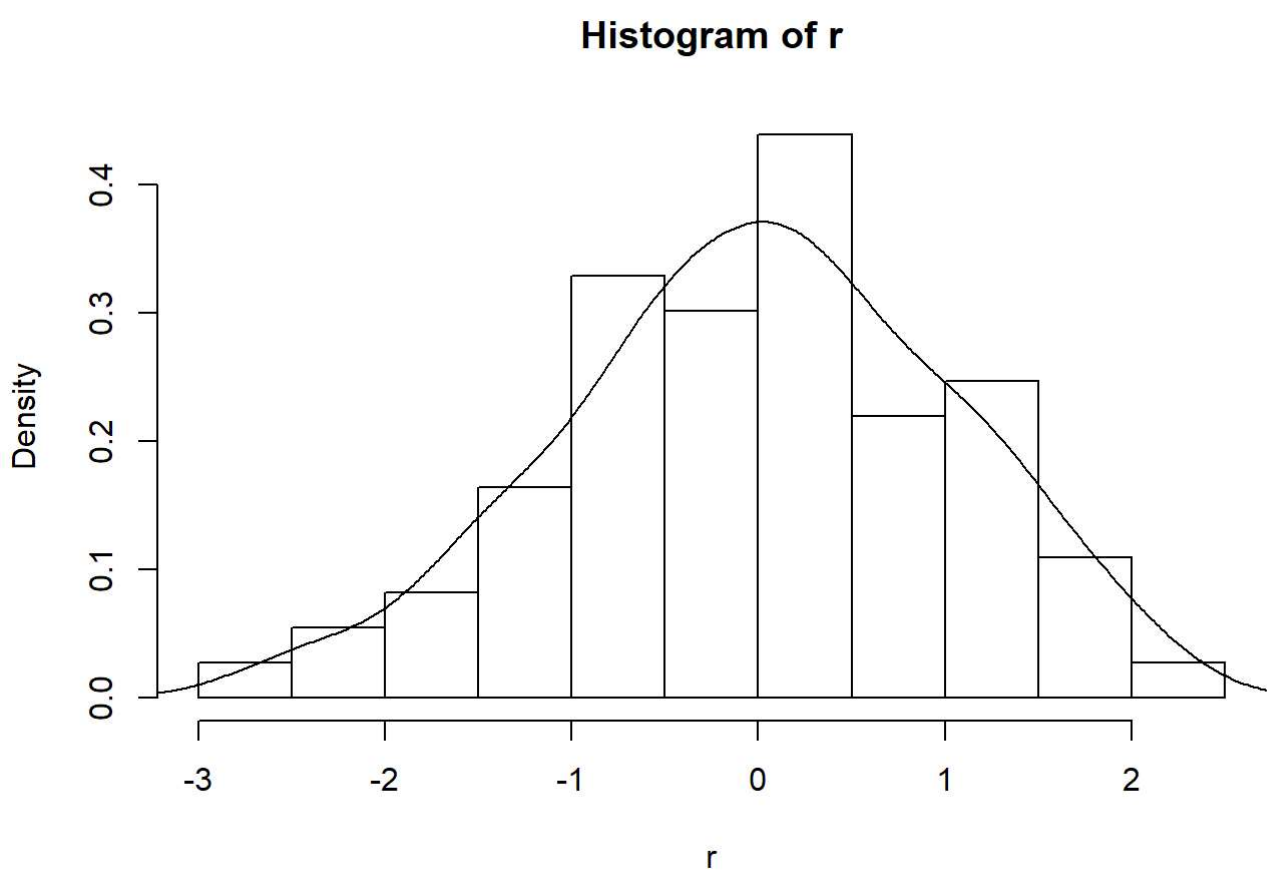



```
plot(fitted(reg),rstudent(reg))
```

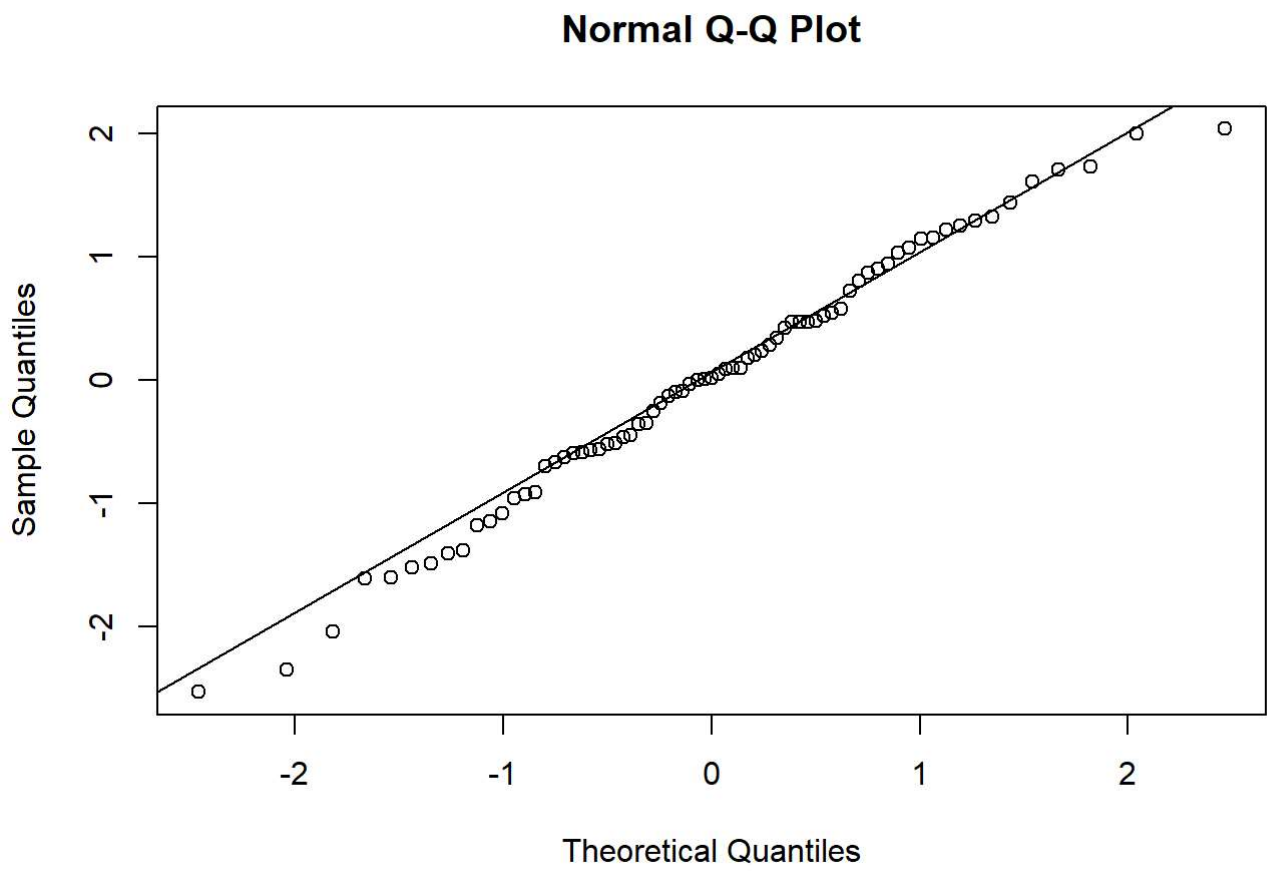


Q5)

```
r= rstudent(reg)
hist(r,freq=FALSE)
lines(density(r, na.rm = T))
```



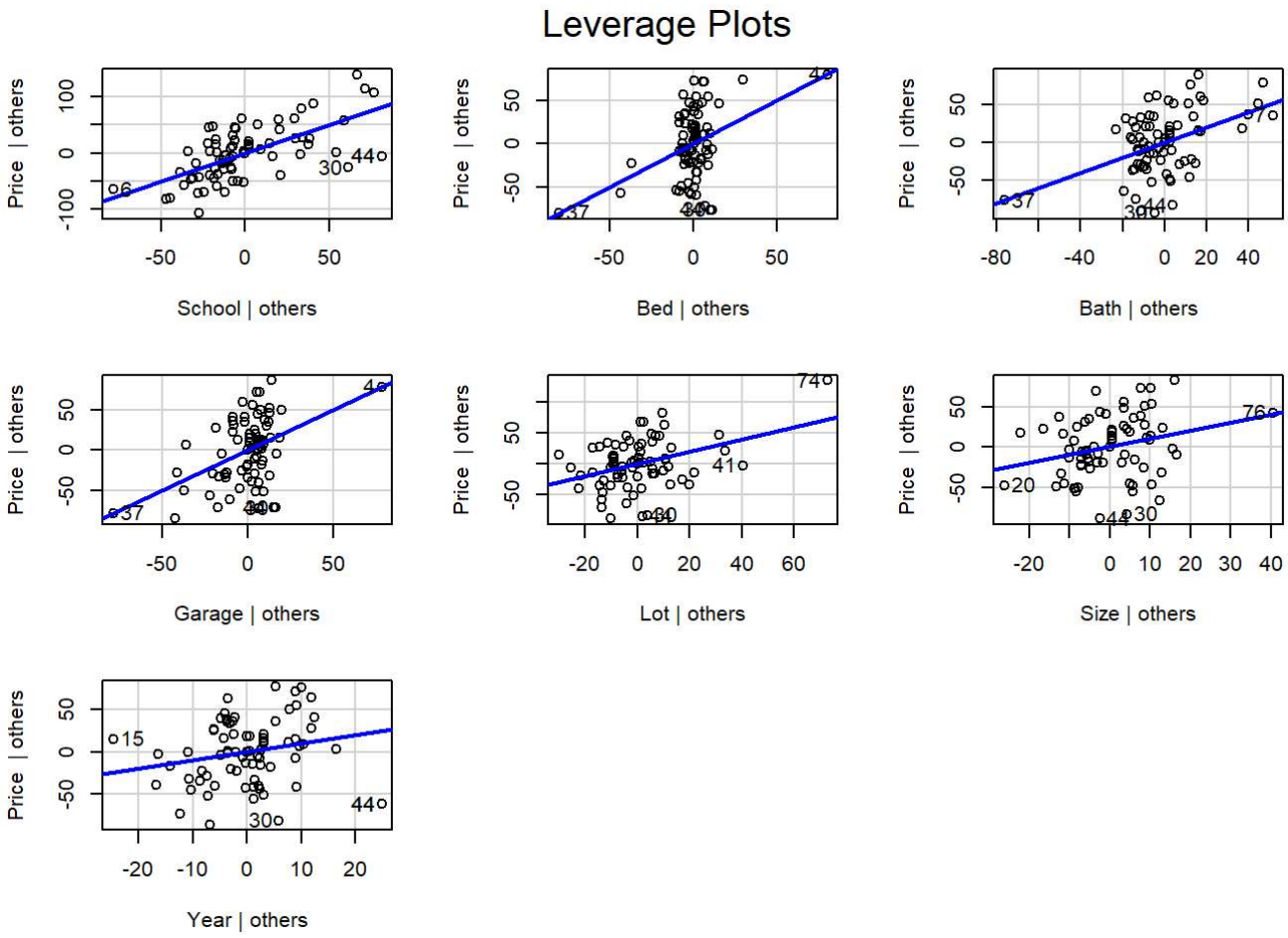

```
qqnorm(r)
qqline(r)
```



Leverage, Influence and Outliers:

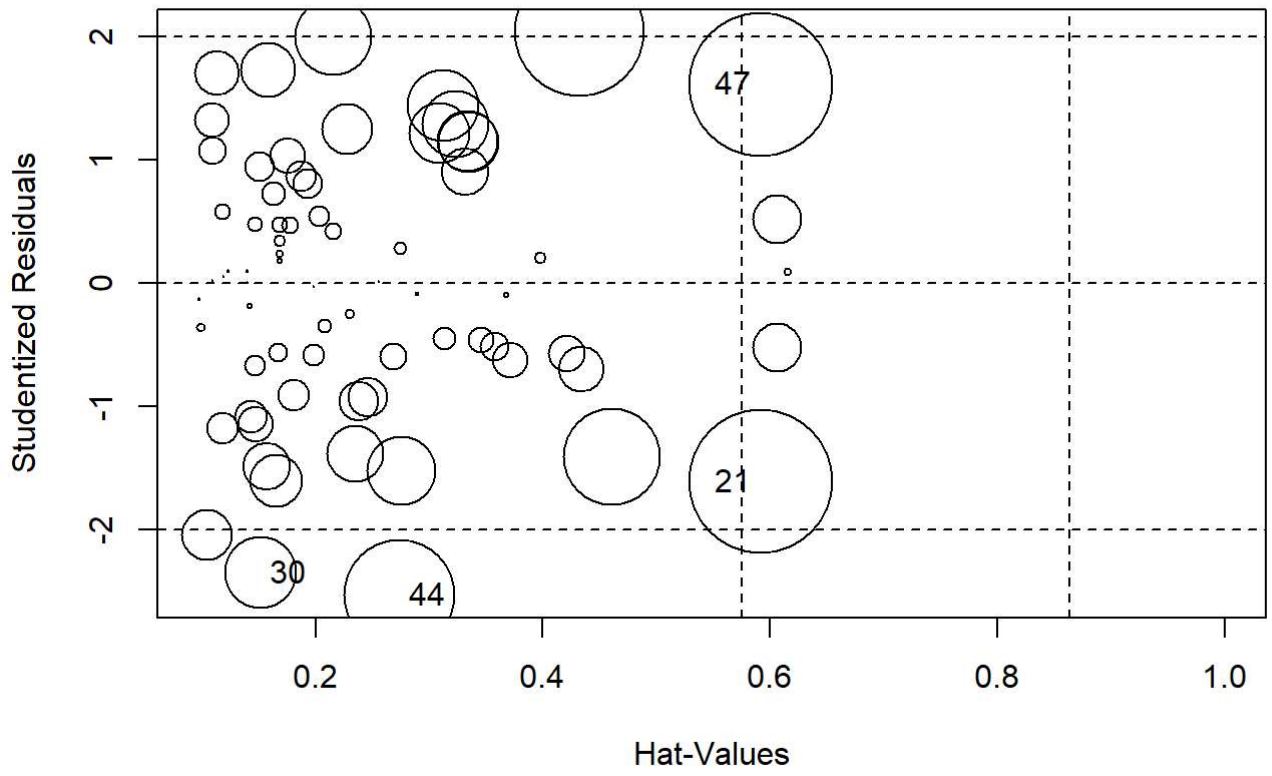
Q1)

```
lev=as.numeric(which(hatvalues(reg)>((2*7)/length(data$Price))))
leveragePlots(reg)
```



Q2)

```
influencePlot(reg)
```




```
##      StudRes       Hat      CookD
## 4        NaN 1.0000000        NaN
## 21 -1.611675 0.5918587 0.17430443
## 30 -2.348239 0.1513825 0.04328835
## 35       NaN 1.0000000        NaN
## 44 -2.527660 0.2736926 0.10441550
## 47  1.611675 0.5918587 0.17430443
```

Q3)

```
library(olsrr)
```

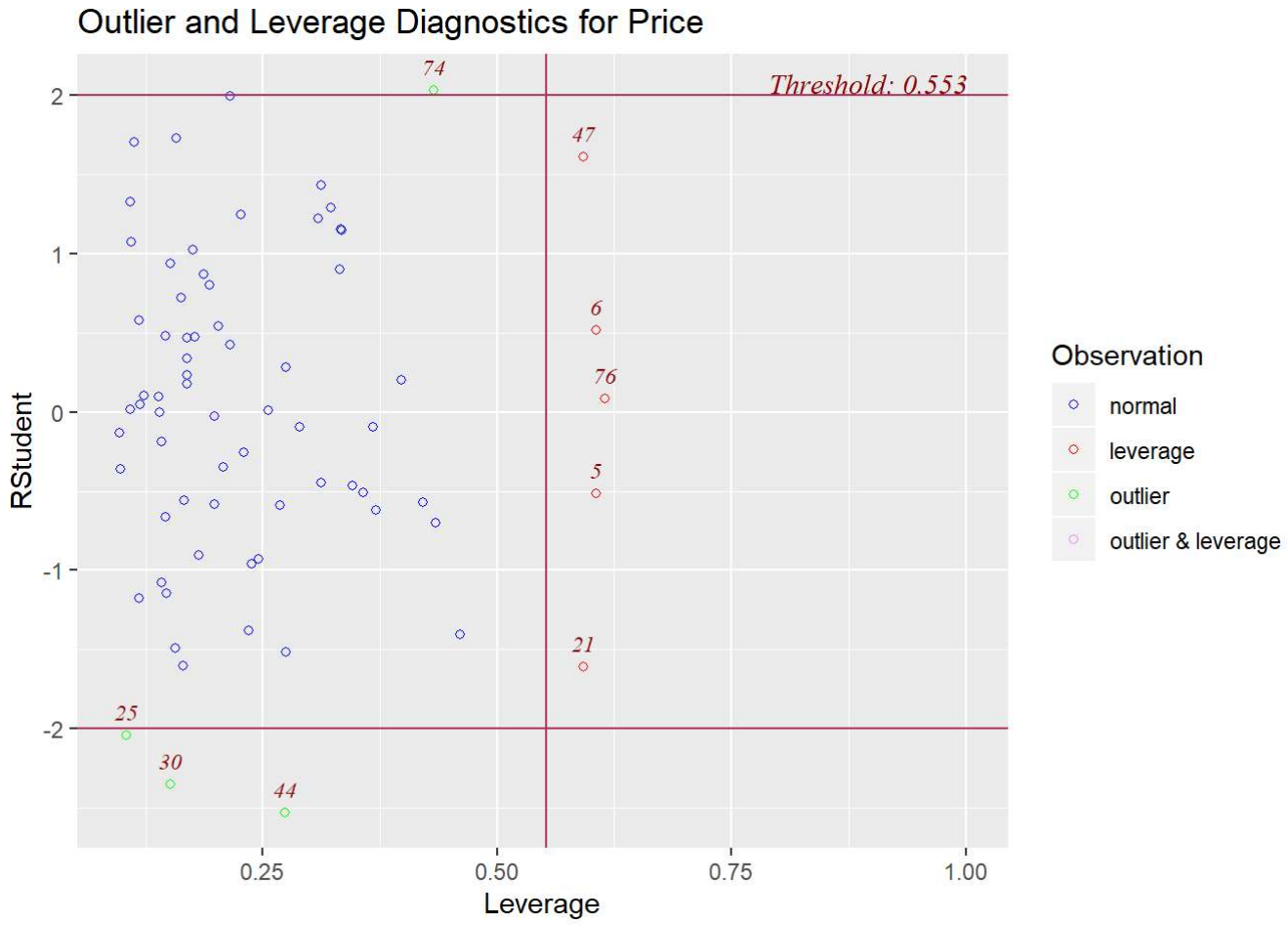
```
##
## Attaching package: 'olsrr'
```

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

```
outlierTest(reg)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 44 -2.52766           0.014441           NA
```

```
ols_plot_resid_lev(reg)
```

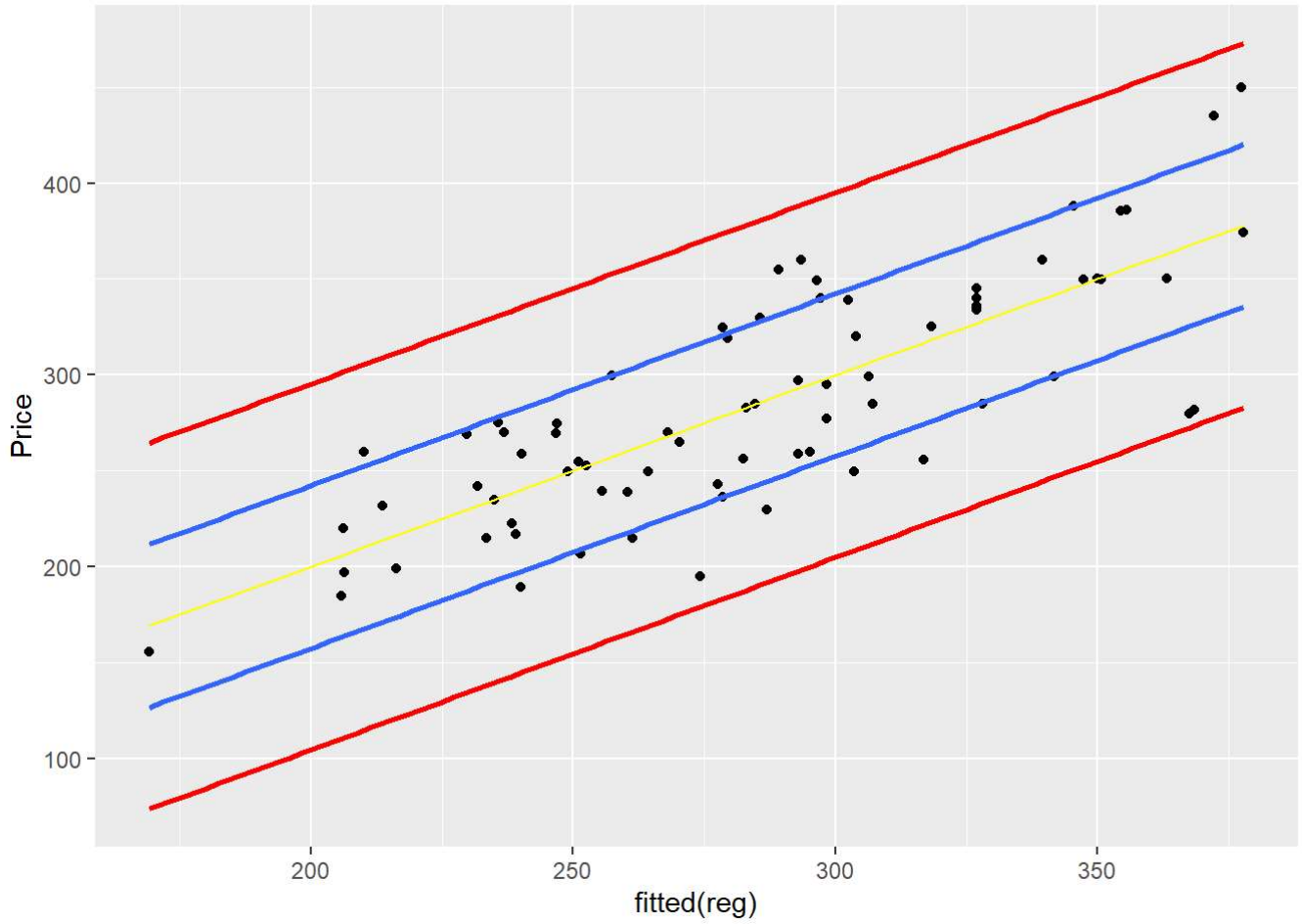


Expected Value, CI and PI:

```
conf=predict(reg,level=0.95,interval='confidence')
pred=predict(reg,level=0.95,interval='prediction')
```

```
## Warning in predict.lm(reg, level = 0.95, interval = "prediction"): predictions on current dat
a refer to _future_ responses
```

```
ggplot(data,aes(y=Price,x=fitted(reg)))+geom_point()+
  stat_smooth(aes(y=conf[,'upr']),method=lm, se=F)+
  stat_smooth(aes(y=conf[,'lwr']),method=lm, se=F)+
  stat_smooth(aes(y=pred[,'upr']),method=lm, se=F,col='red')+
  stat_smooth(aes(y=pred[,'lwr']),method=lm, se=F,col='red')+
  geom_line(aes(y=conf[,'fit']),col='yellow')
```


Price
400
300
200
100
200
250
300
350
fitted(reg)