Predicitve Assignment

Srajan Rai- 19200436

11/16/2019

Import the House dataset and load few packages

```
House<-read.csv("/Users/apple/Documents/UCD/Predictive analytics/assignments/House.csv")
library(GGally)
```

```
## Loading required package: ggplot2
```

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

```
library(car)
```

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

```
library(olsrr)
```

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

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

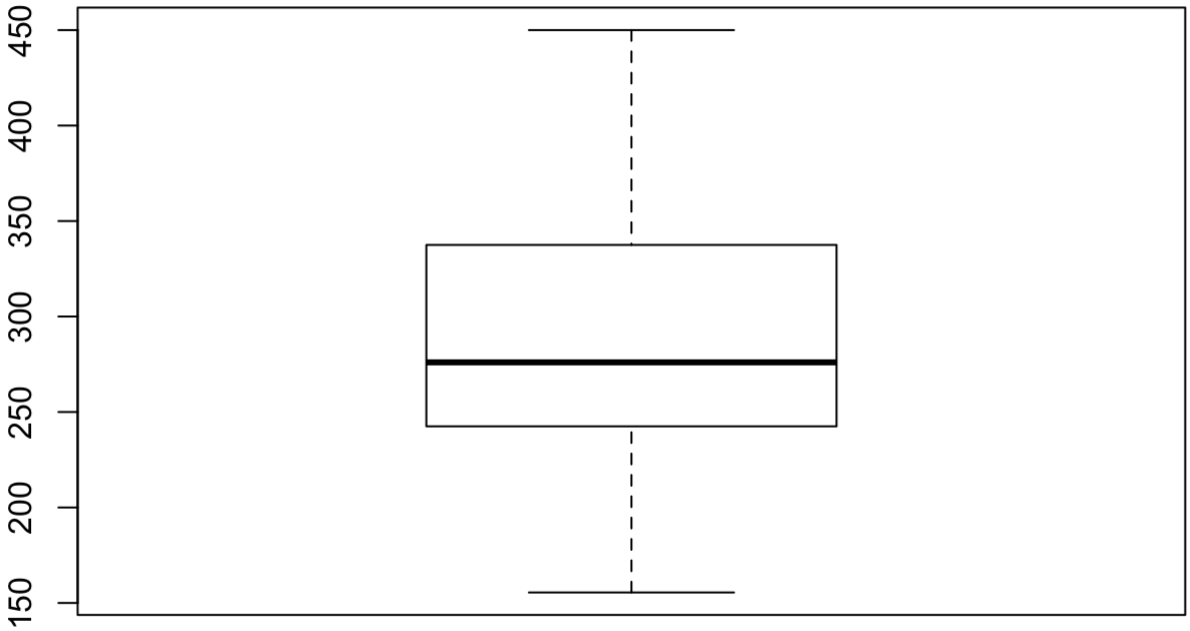
EDA

BoxPlot,Histogram, Summary of House Sale Price

```
summary(House$Price)
```

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

```
boxplot(House$Price)
```



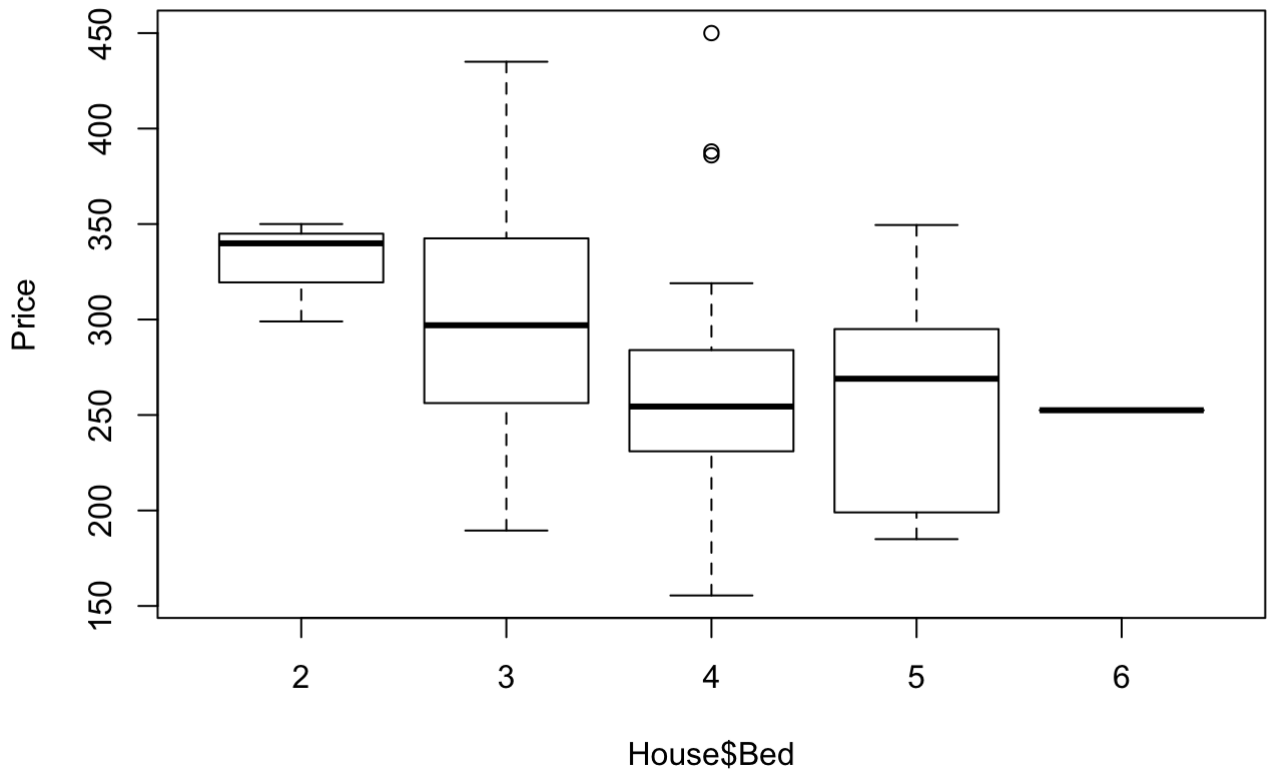

```
hist(House$Price)
```

2- categorical variables to factors

```
House$Bath<-factor(House$Bath)
House$Bed<-factor(House$Bed)
House$Garage<-factor(House$Garage)
House$School<-factor(House$School)
```

2- Summary and Boxplot describing how sales price varies with respect to the number of bedrooms, bathrooms, garage size and school.

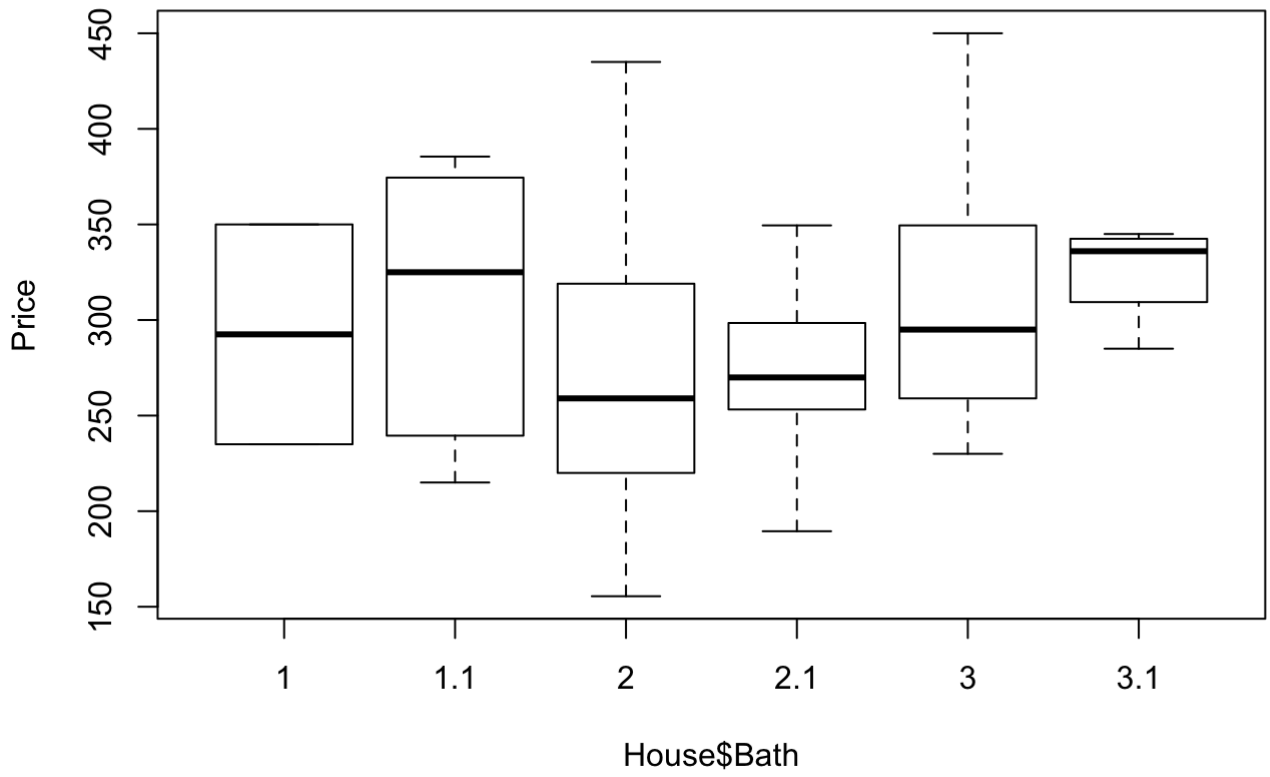
```
attach(House)
boxplot(Price~House$Bed)
```




```
by(Price,Bed,summary)
```

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

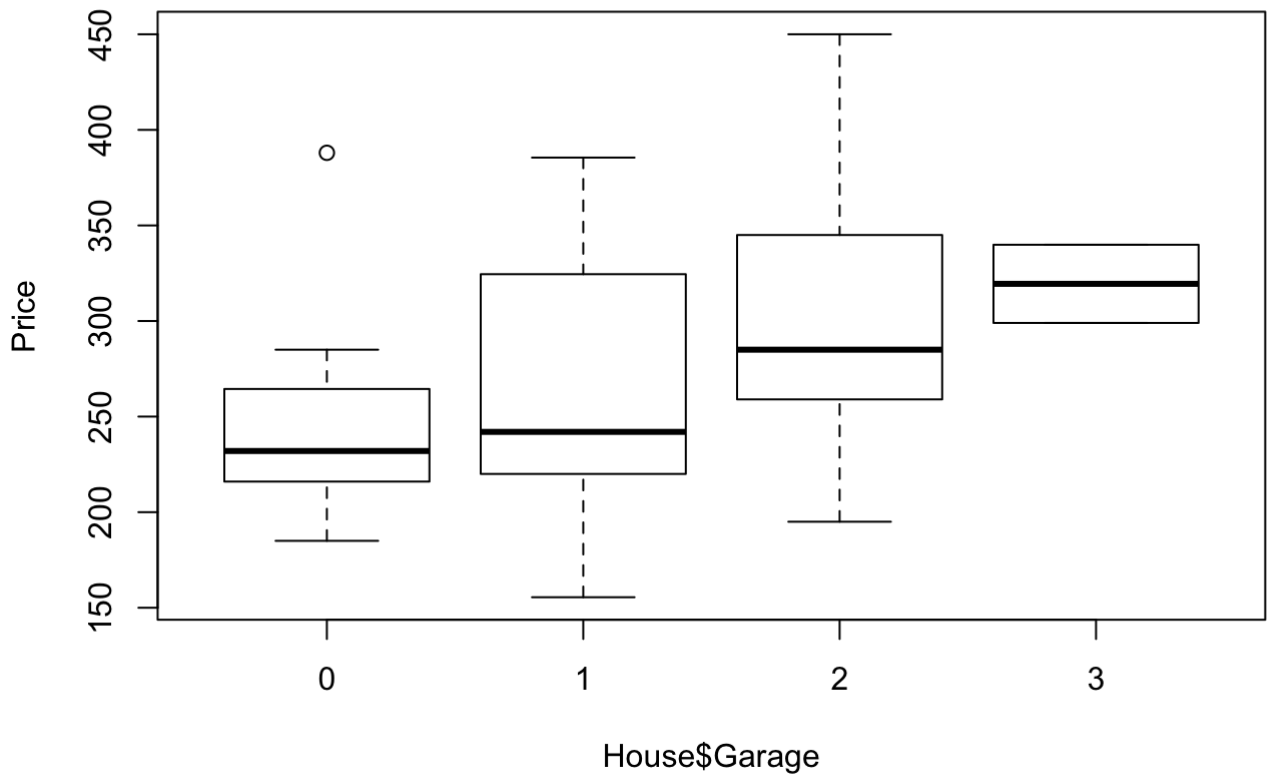
```
boxplot(Price~House$Bath)
```




```
by(Price,Bath,summary)
```

```
## Bath: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   235.0   263.8   292.5   292.5   321.2   350.0
## ------------------------------------------------------------
## Bath: 1.1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   215.0   239.5   325.0   307.9   374.5   385.5
## ------------------------------------------------------------
## Bath: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   220.0   259.0   270.7   319.0   435.0
## ------------------------------------------------------------
## Bath: 2.1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   189.5   254.8   269.9   274.5   297.7   349.5
## ------------------------------------------------------------
## Bath: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   230.0   259.0   295.0   307.8   349.5   450.0
## ------------------------------------------------------------
## Bath: 3.1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   285.0   309.4   336.0   324.2   342.5   345.0
```

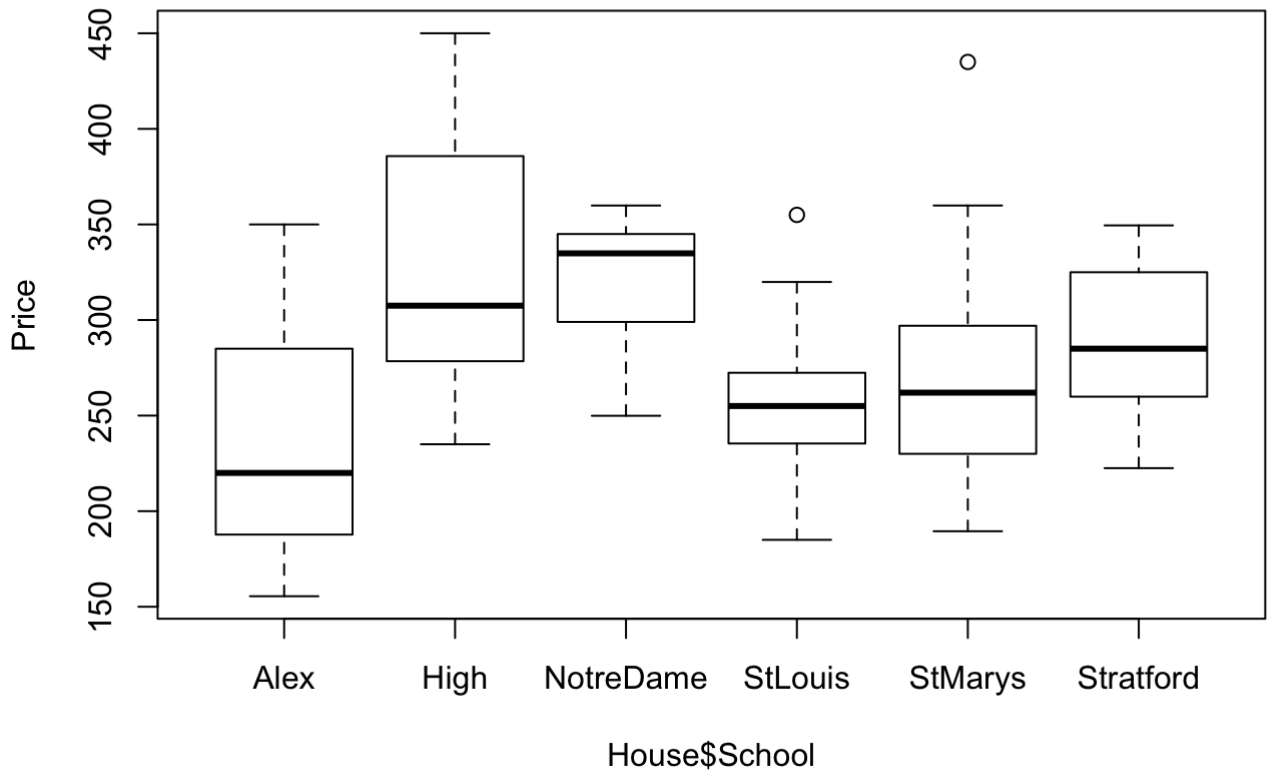
```
boxplot(Price~House$Garage)
```




```
by(Price,Garage,summary)
```

```
## Garage: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   185.0   216.0   232.0   246.9   264.4   388.0
## ------------------------------------------------------------
## Garage: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.5   220.0   242.0   260.6   324.5   385.5
## ------------------------------------------------------------
## Garage: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   195.0   259.0   285.0   299.6   343.8   450.0
## ------------------------------------------------------------
## Garage: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   299.0   309.2   319.4   319.4   329.7   339.9
```

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

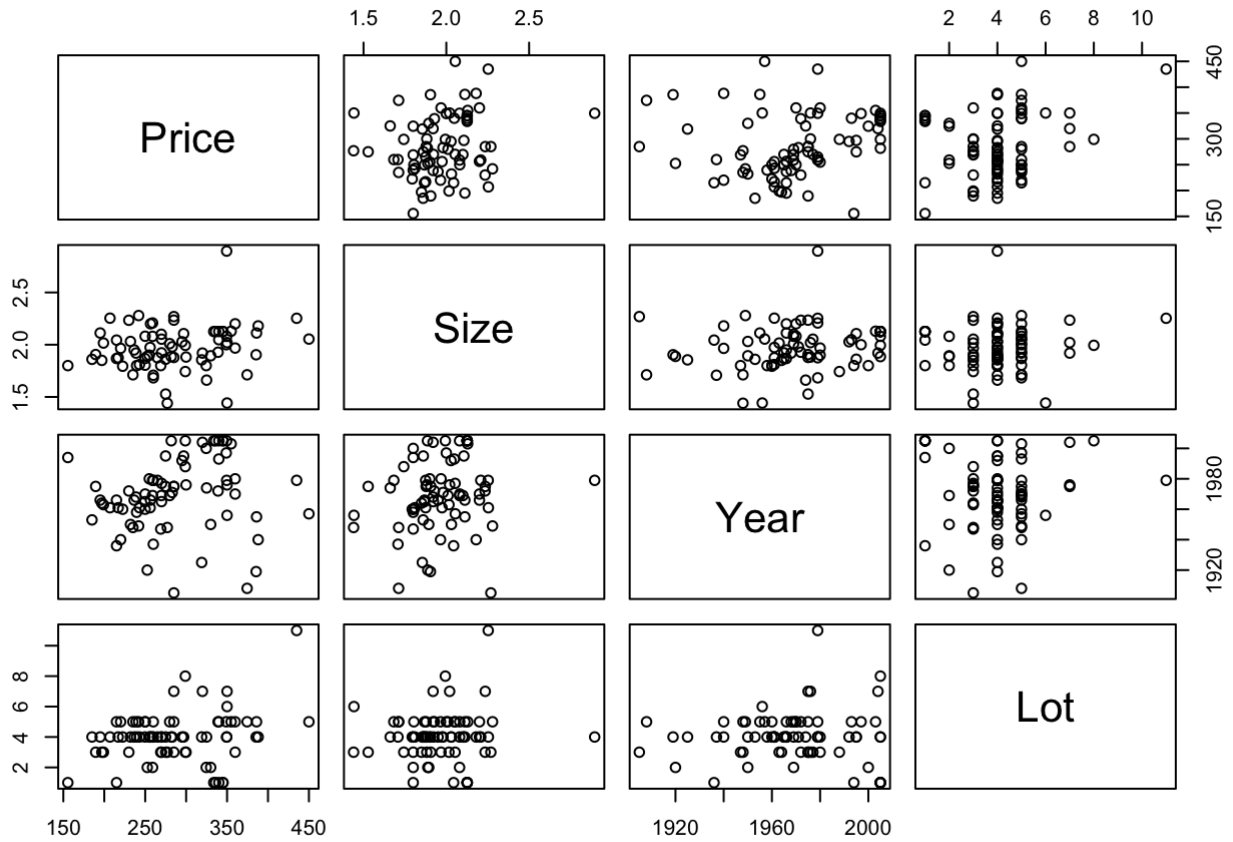



```
by(Price,School,summary)
```

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

3-Correlation, Pairs plot

```
pairs(Price~Size+Year+Lot)
```



```
cor(House[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
```

```
House$Lot<-House$Lot-mean(House$Lot)
House$Year<-House$Year-mean(House$Year)
House$Size<-House$Size-mean(House$Size)
```

Regression model

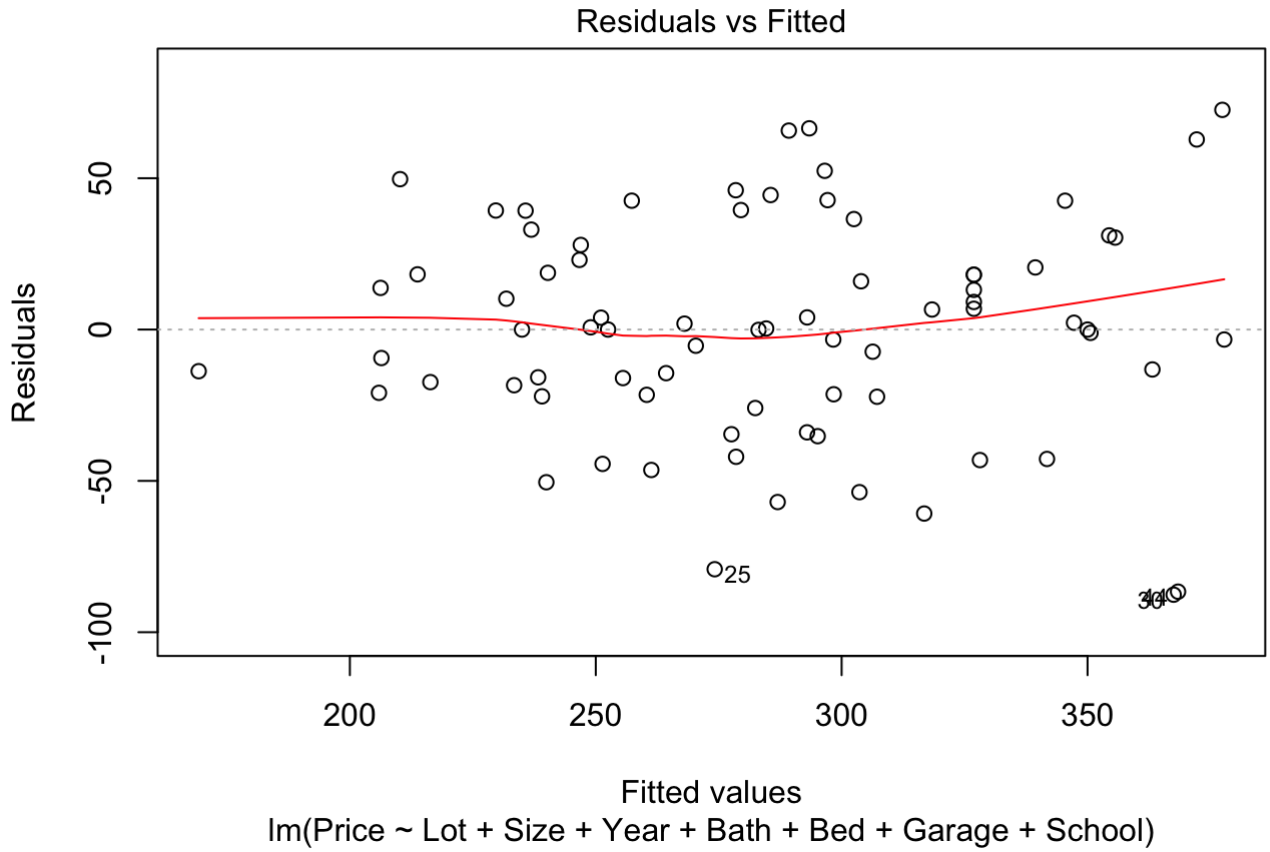
Multiple Linear regression model

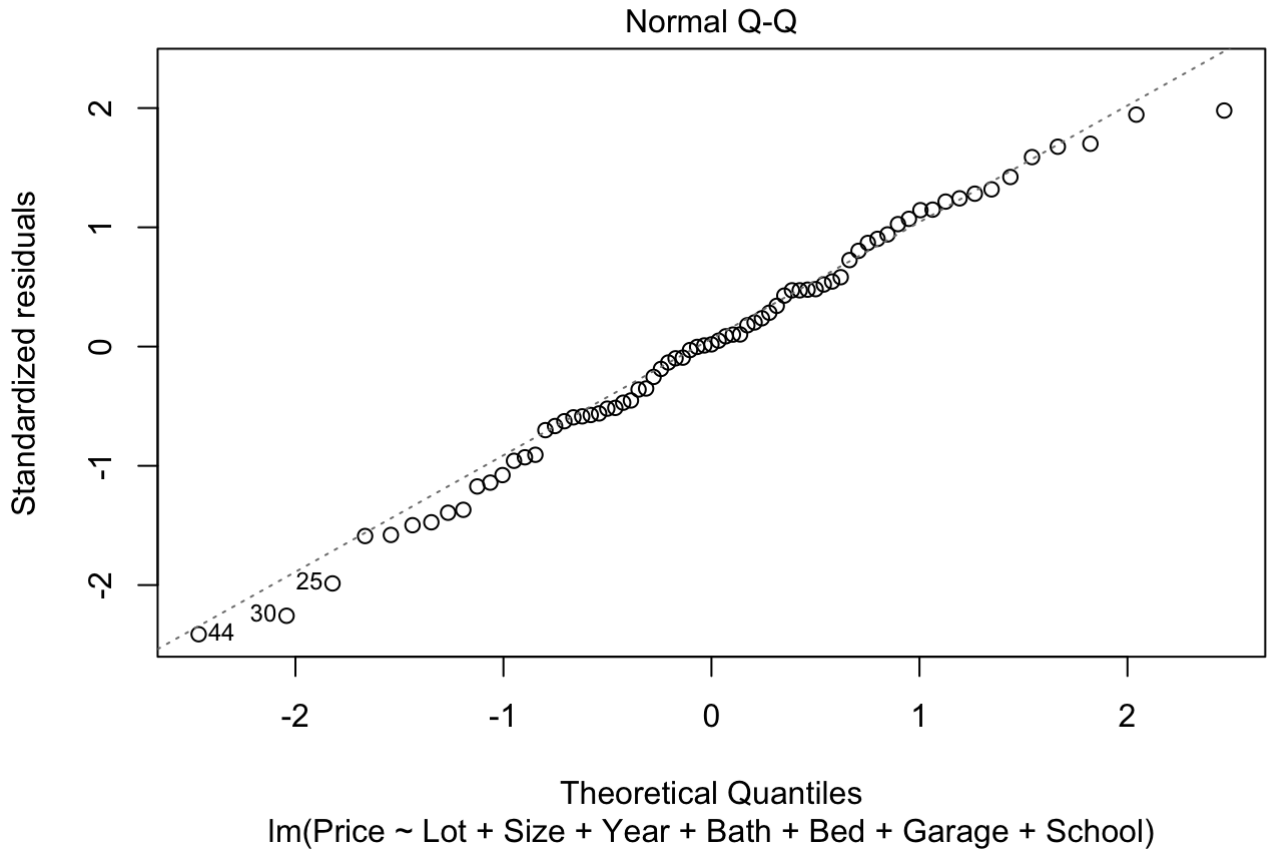
```
mod<-lm(Price~Lot+Size+Year+Bath+Bed+Garage+School,data=House)
summary(mod)
```

```
##
## Call:
## lm(formula = Price ~ Lot + Size + Year + Bath + Bed + Garage +
##     School, data = House)
##
## 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 ***
## 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
## 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
```

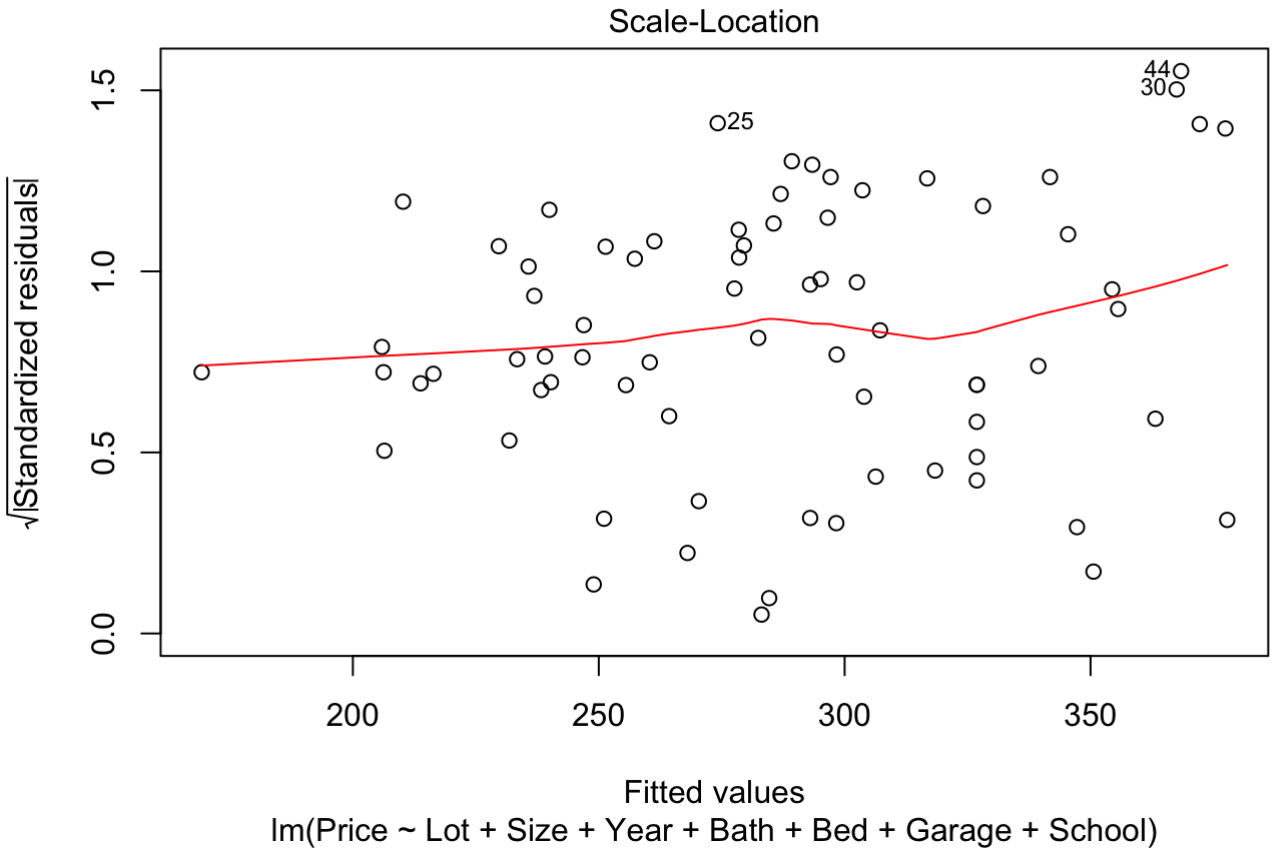
```
plot(mod)
```

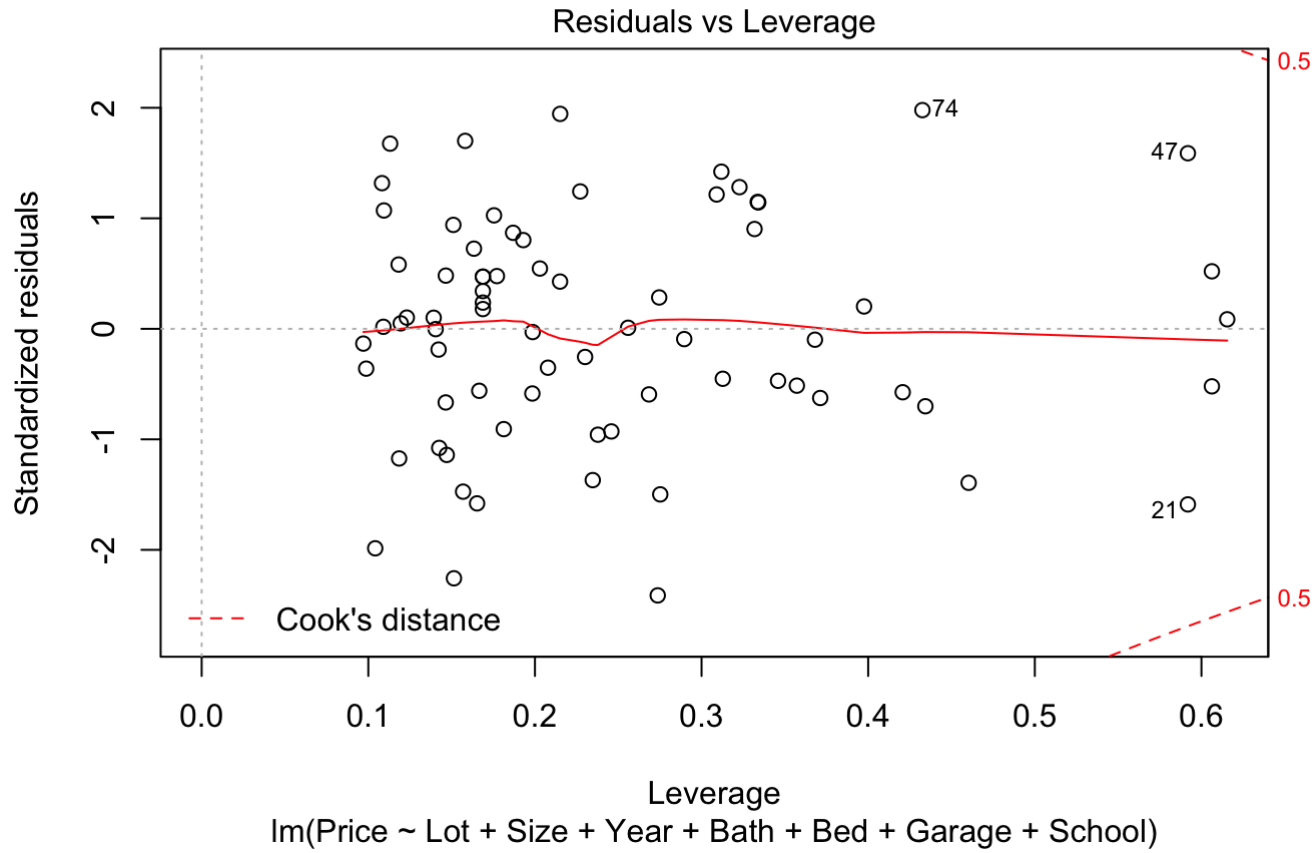
```
## Warning: not plotting observations with leverage one:
##   4, 35, 37
```






```
## Warning: not plotting observations with leverage one:
##   4, 35, 37
```





Anova ### Type 1 anova and non-significant predictor variable (Year) is removed

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Lot        1  16284 16284.4  9.1767  0.003729 **
## Size       1  10026 10025.7  5.6498  0.020964 *
## 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
```

```
mod2<-lm(Price~Lot+Size+Bath+Bed+Garage+School,data = House)
```

Type 2 anova

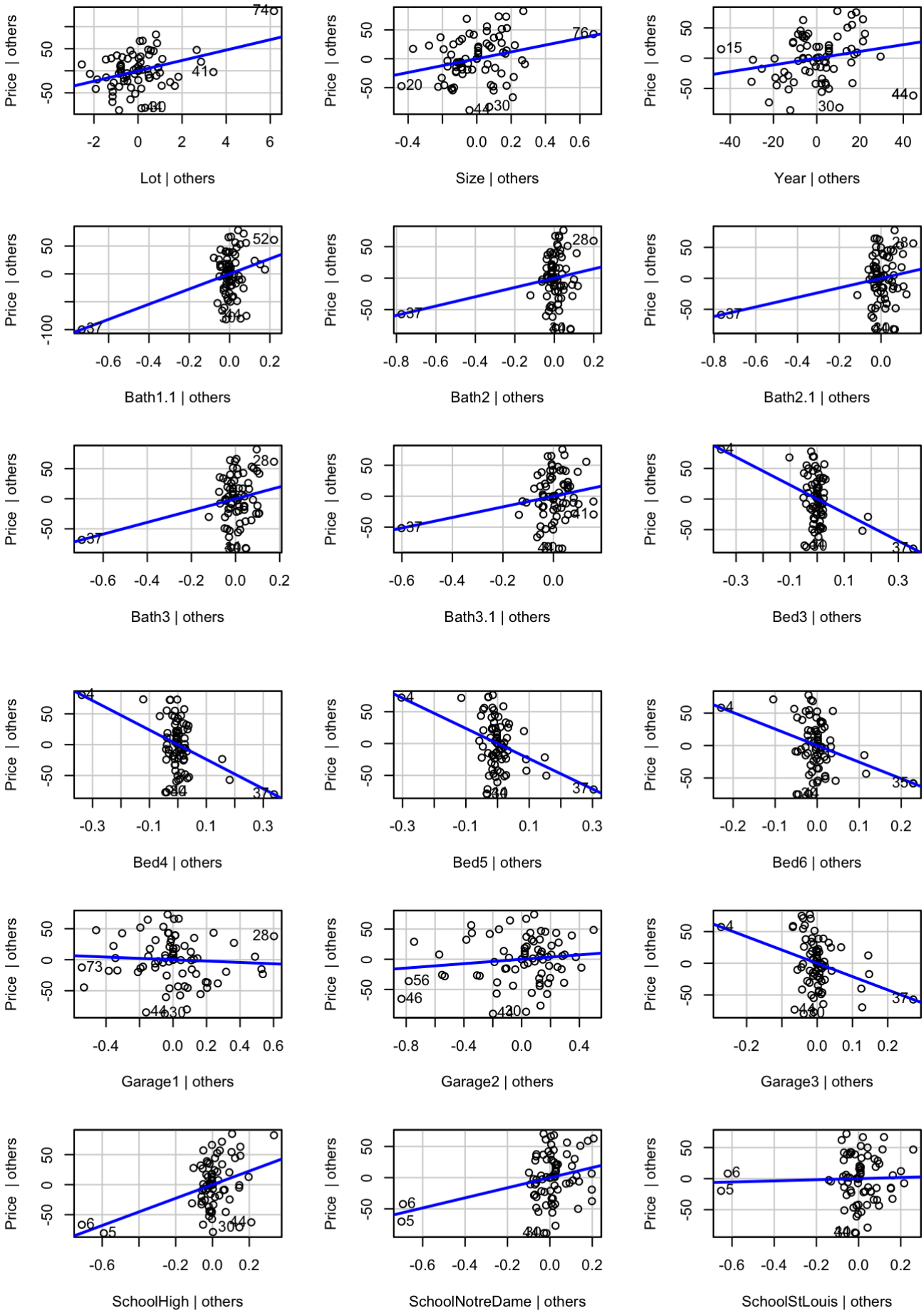
```
Anova(mod,mod2)
```

```
## Anova Table (Type II tests)
##
## Response: Price
##           Sum Sq Df F value   Pr(>F)
## Lot        17168  1  9.3883 0.003355 **
## Size        7467  1  4.0835 0.048094 *
## Year        4803  1  2.6264 0.110720
## Bath       23324  5  2.5511 0.037771 *
## Bed        19278  4  2.6356 0.043432 *
## Garage     25373  3  4.6252 0.005841 **
## School     70112  5  7.6683 1.51e-05 ***
## Residuals 102402 56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

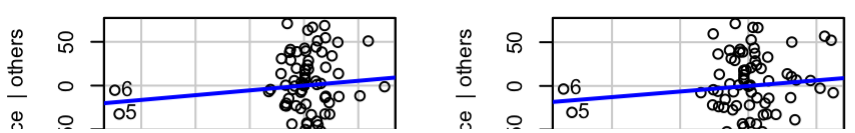
Diagnostics:

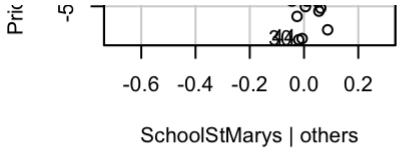
1-Added variable plot and component plus residual plot

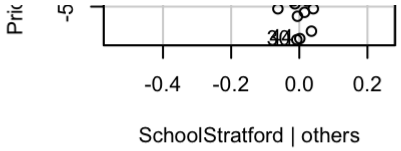
```
avPlots(mod)
```



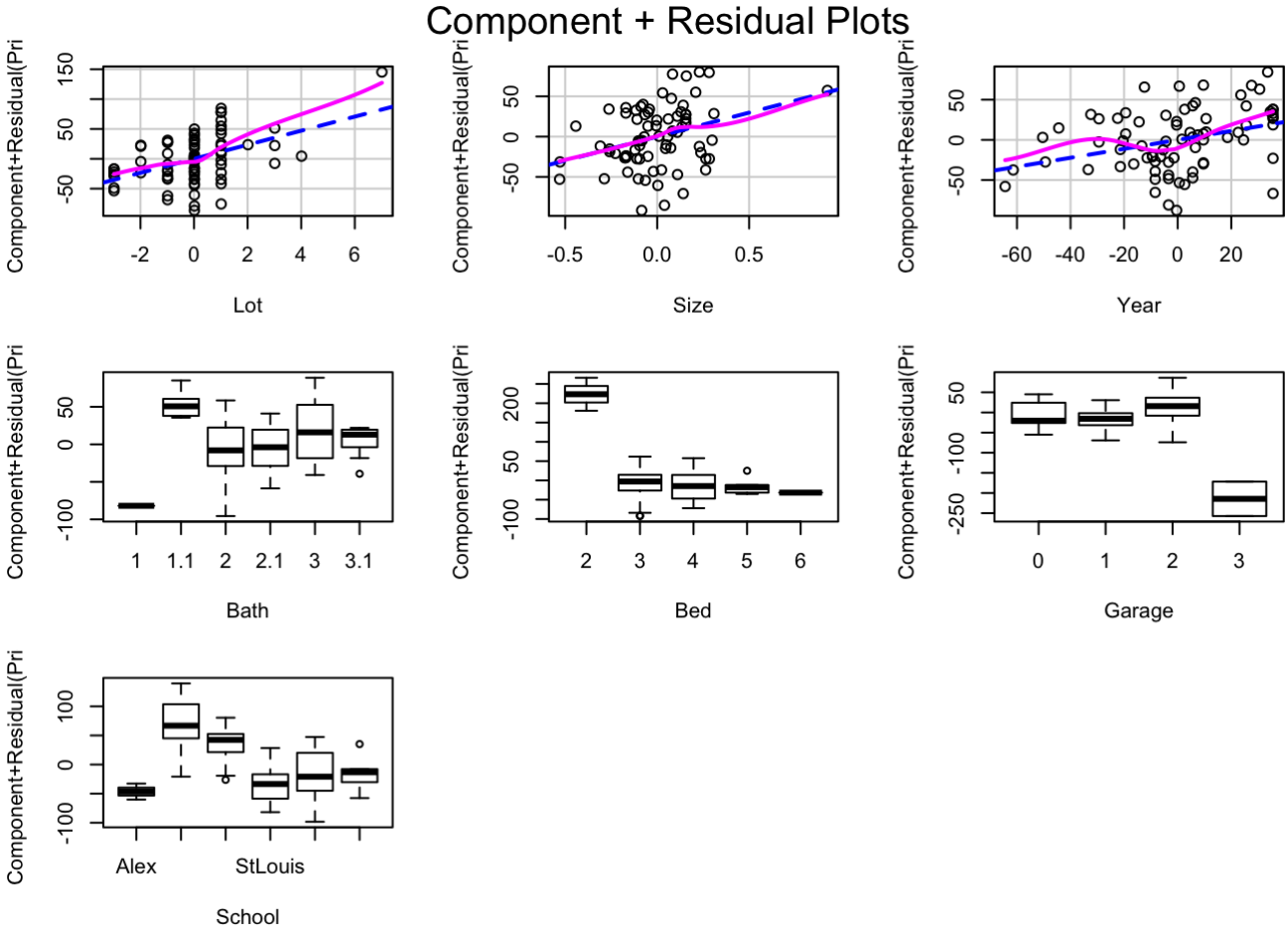
Added-Variable Plots








```
crPlots(mod)
```



2-Durbin-Watson test

```
dwt(mod)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1836122      1.614157    0.03
##  Alternative hypothesis: rho != 0
```

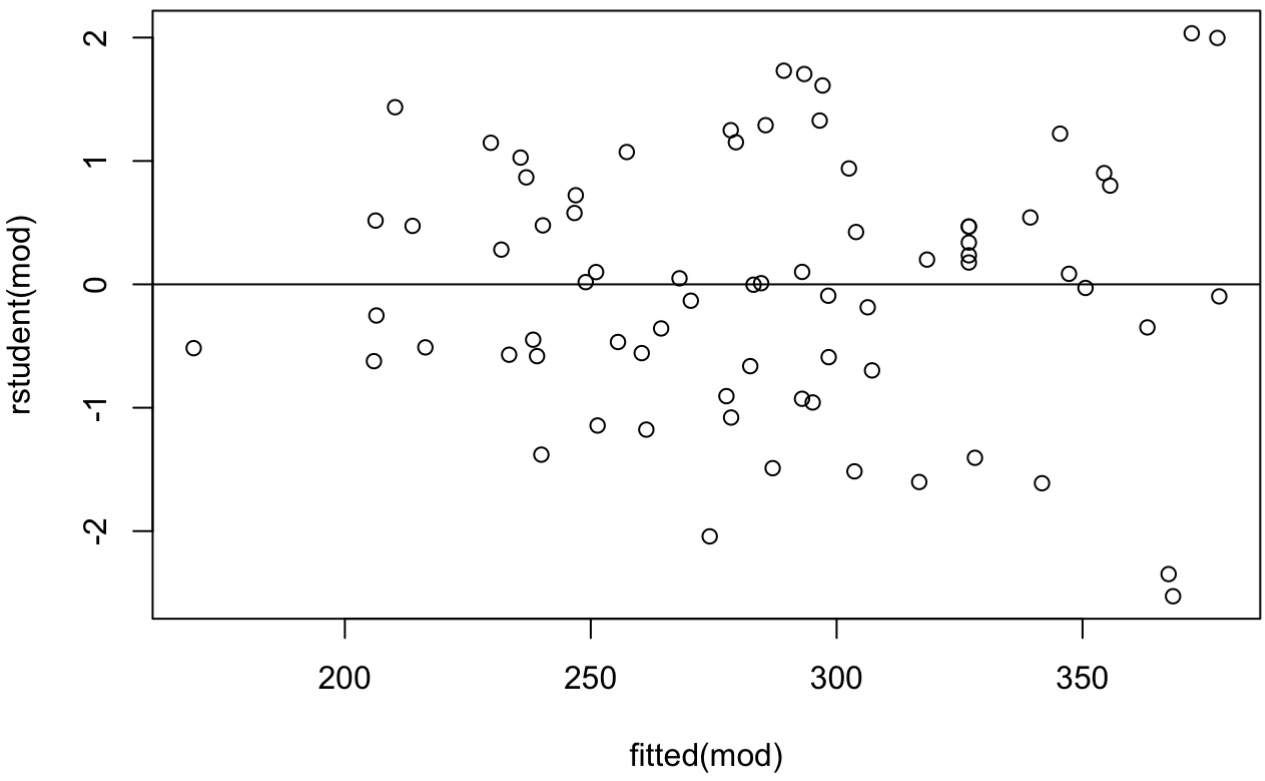
3-Collinearity Check-variation inflation factor

```
vif(mod)
```

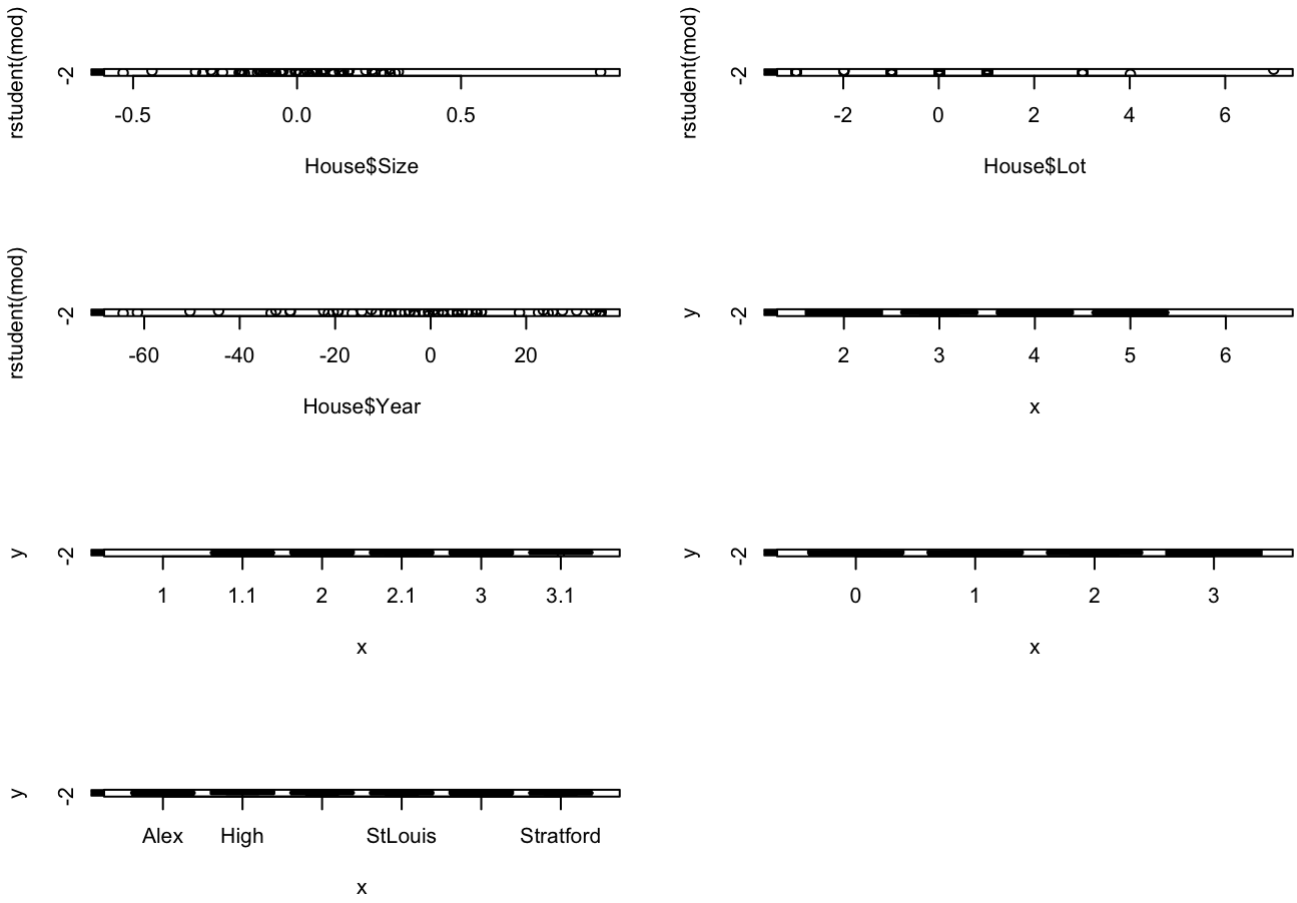
```
##             GVIF Df GVIF^(1/(2*Df))
## Lot     1.654167  1        1.286144
## Size    1.601785  1        1.265616
## Year    2.671175  1        1.634373
## Bath    9.757455  5        1.255838
## Bed    20.215797  4        1.456168
## Garage 19.811449  3        1.644950
## School  6.768538  5 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

4-Zero conditional mean and homoscedasticity

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

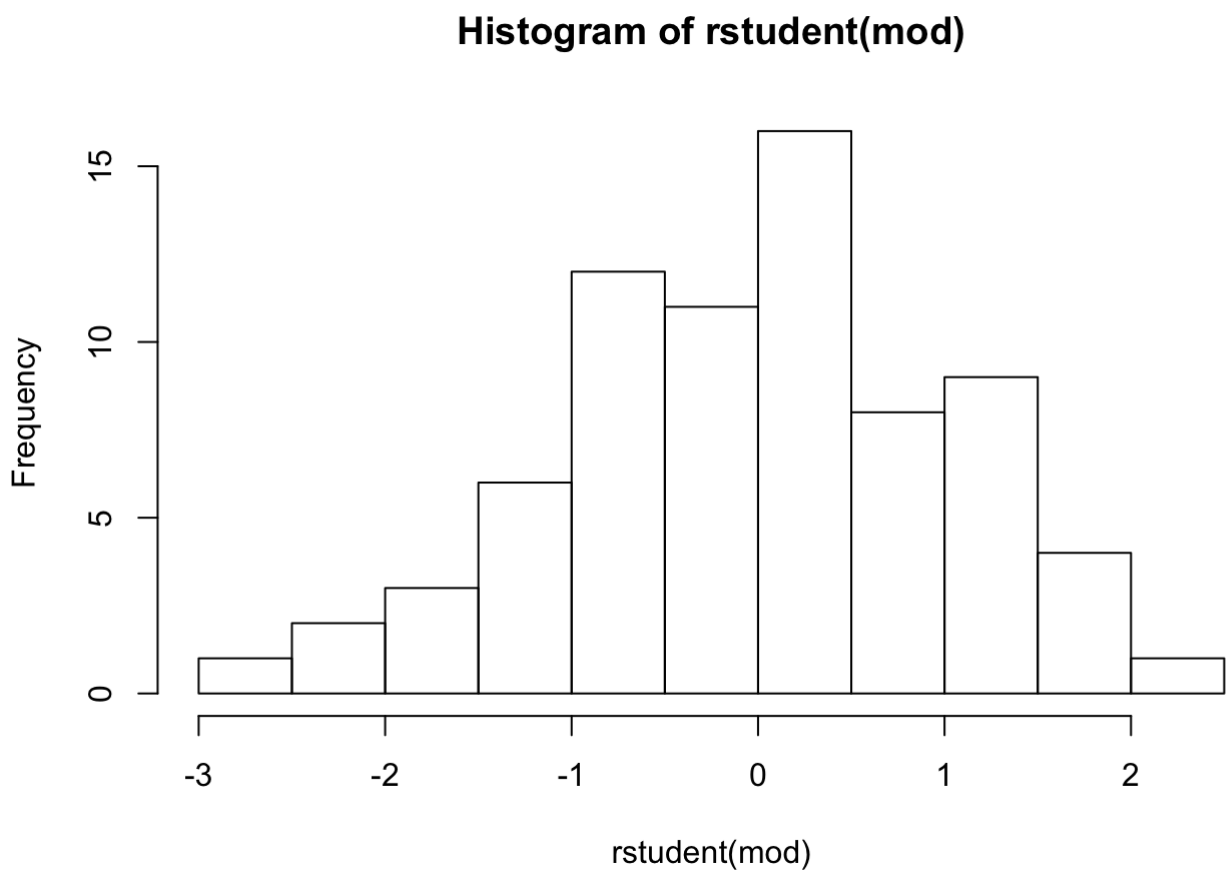


```
par(mfrow=c(4,2))
plot(House$Size,rstudent(mod))
plot(House$Lot,rstudent(mod))
plot(House$Year,rstudent(mod))
plot(House$Bed,rstudent(mod))
plot(House$Bath,rstudent(mod))
plot(House$Garage,rstudent(mod))
plot(House$School,rstudent(mod))
```

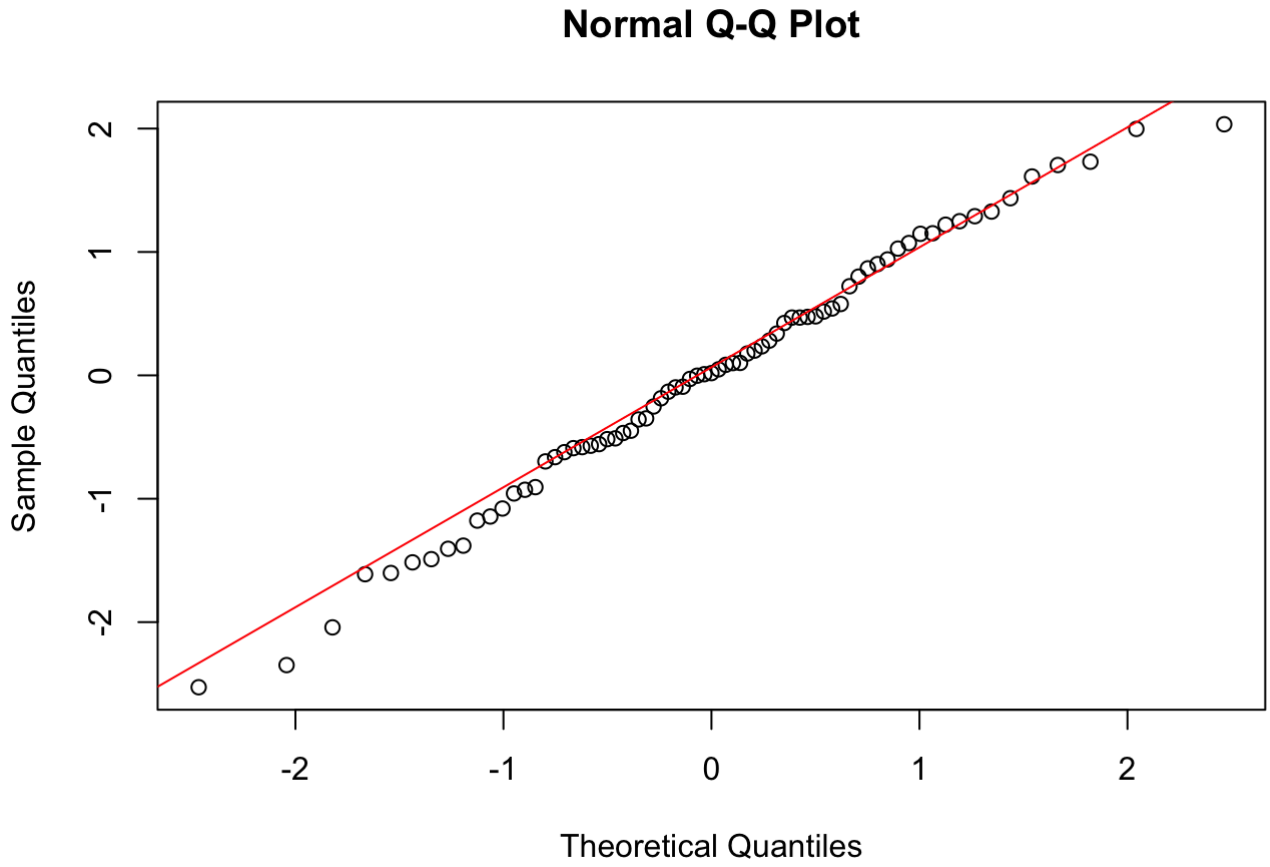


5-Normality Assumption

```
hist(rstudent(mod))
```




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



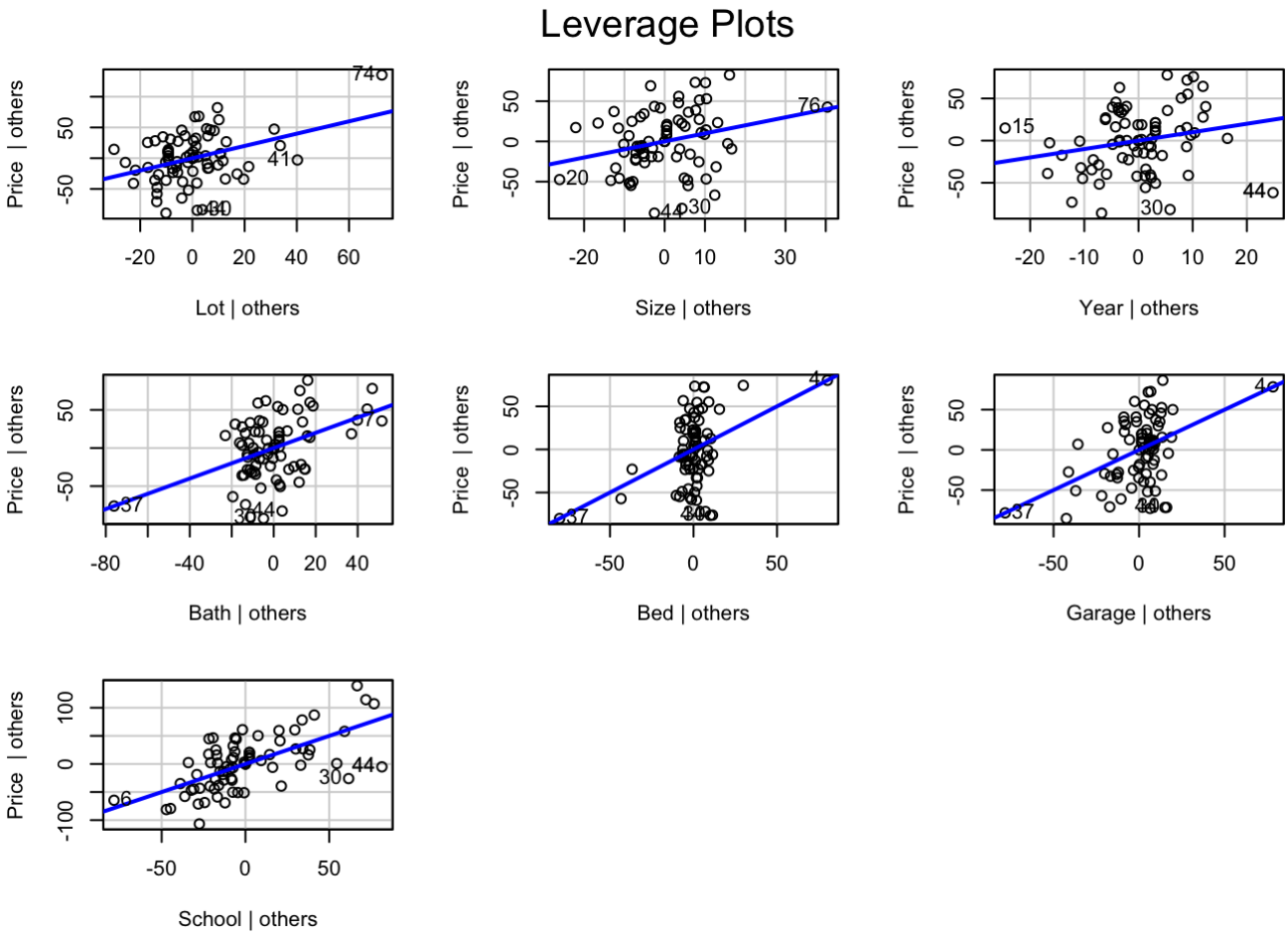
Leverage, Influence and Outliers:

Leverage values and Leverage plots

```
lev_point<-as.numeric(which(hatvalues(mod)>((2*7)/76)))
lev_point
```

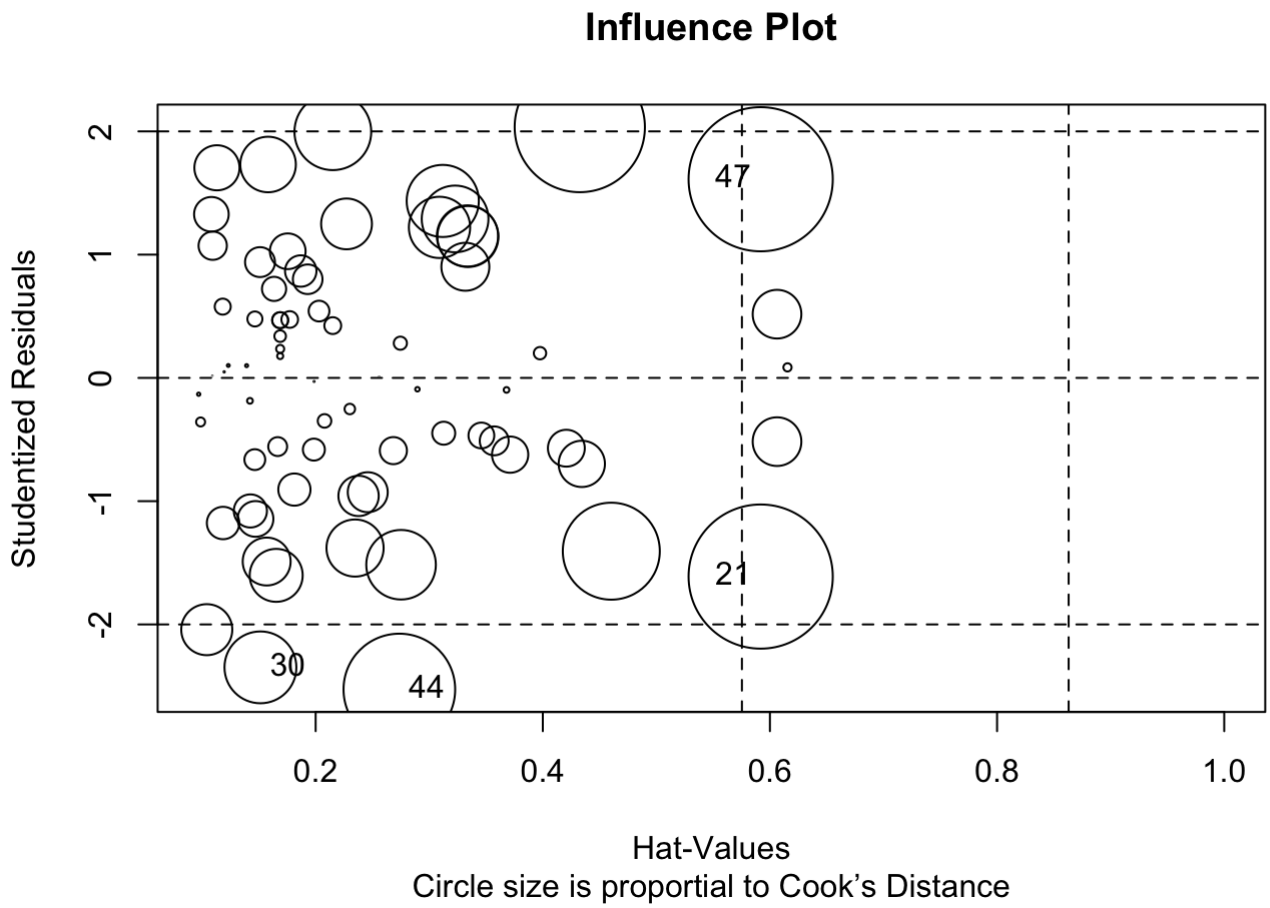
```
##  [1]  1  2  3  4  5  6  7  9 15 20 21 22 28 31 32 33 34 35 36 37 39 41 42
## [24] 43 44 46 47 49 50 51 52 54 56 57 58 64 66 69 71 72 73 74 76
```

```
leveragePlots(mod)
```



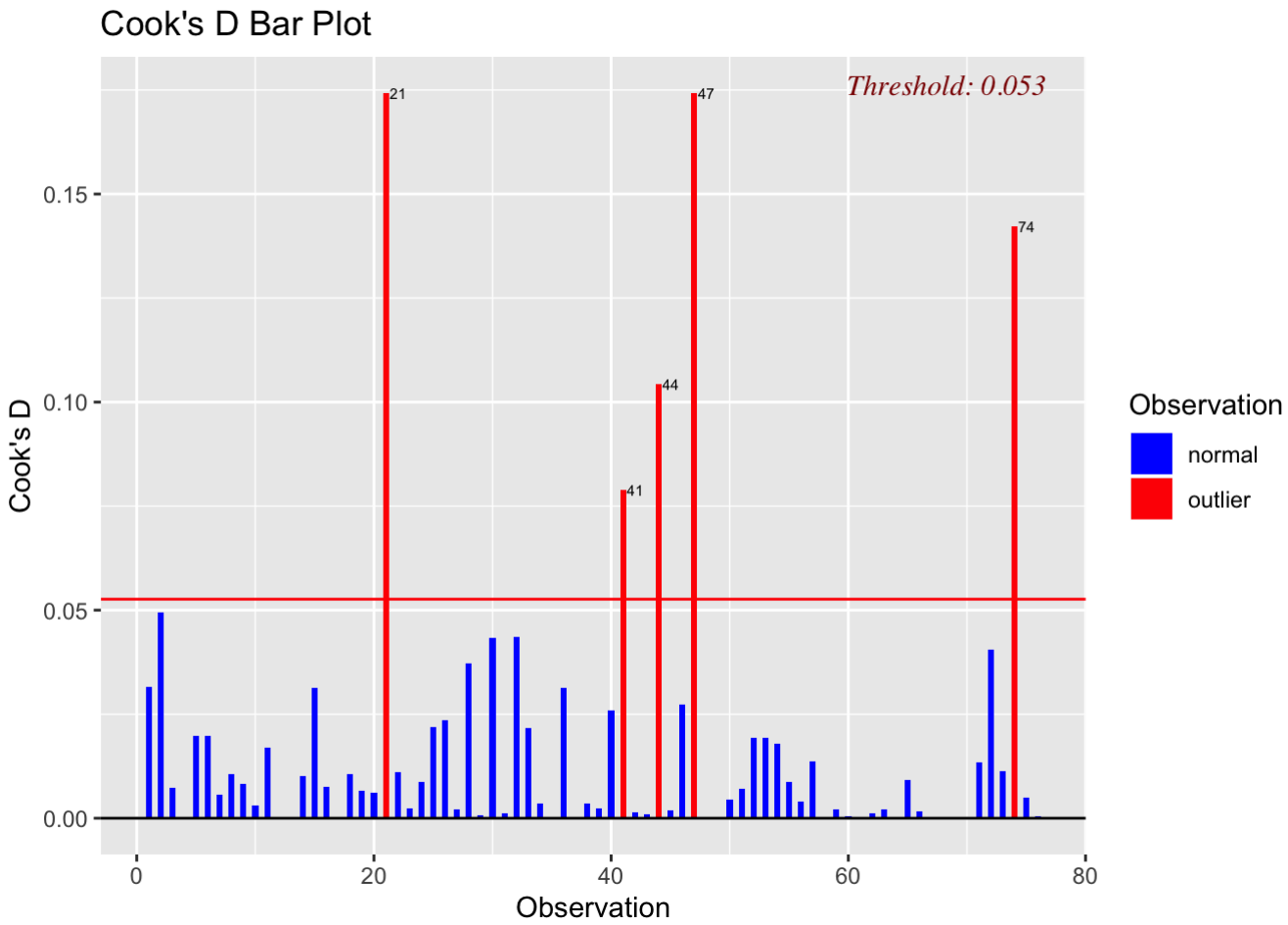
2- influencial Plot

```
influencePlot(mod, main="Influence Plot",sub="Circle size is proportial to Cook's Distance")
```




```
##      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
```

```
ols_plot_cooksd_bar(mod)
```

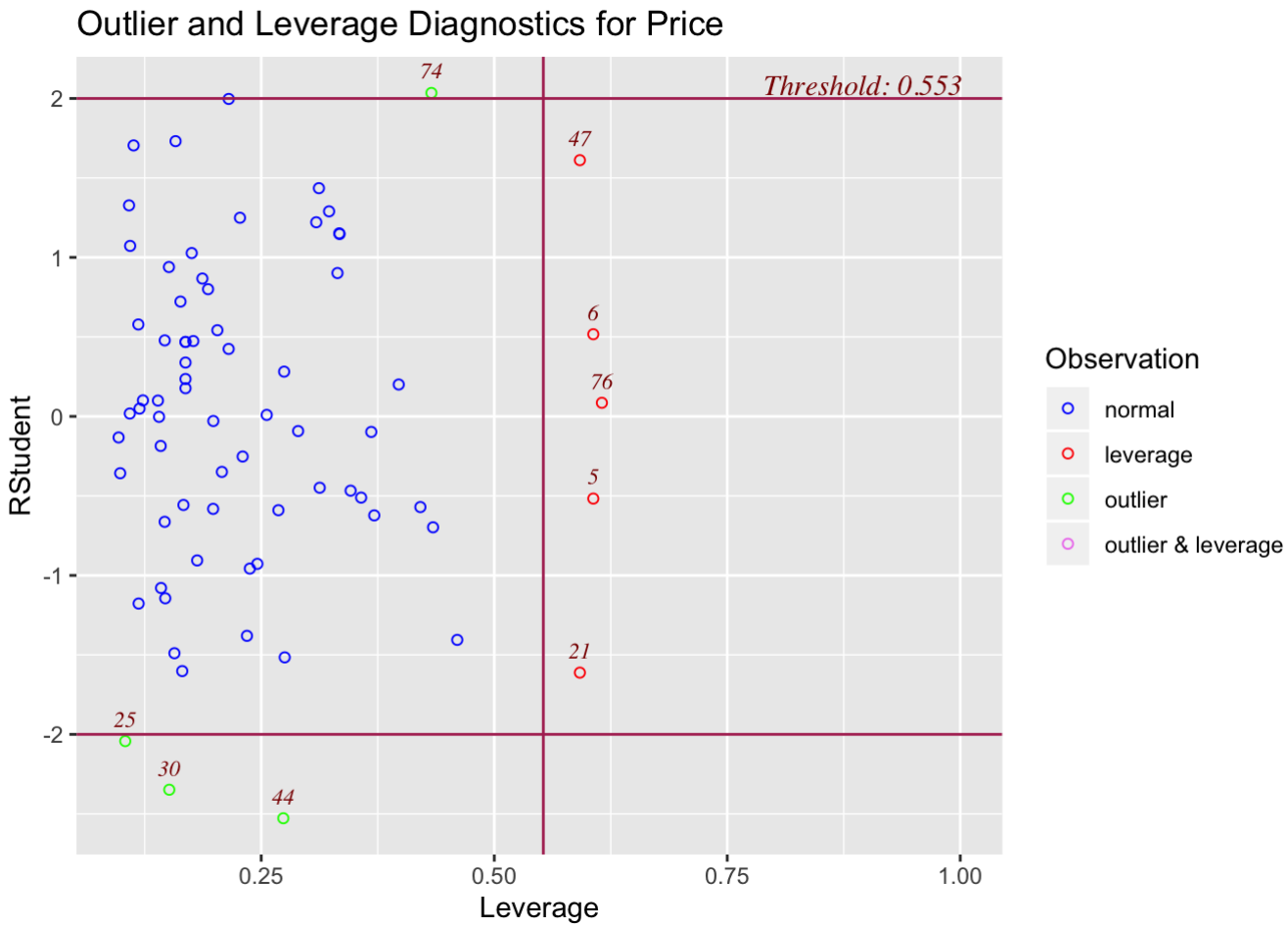


Outlier

```
outlierTest(mod)
```

```
## 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(mod)
```

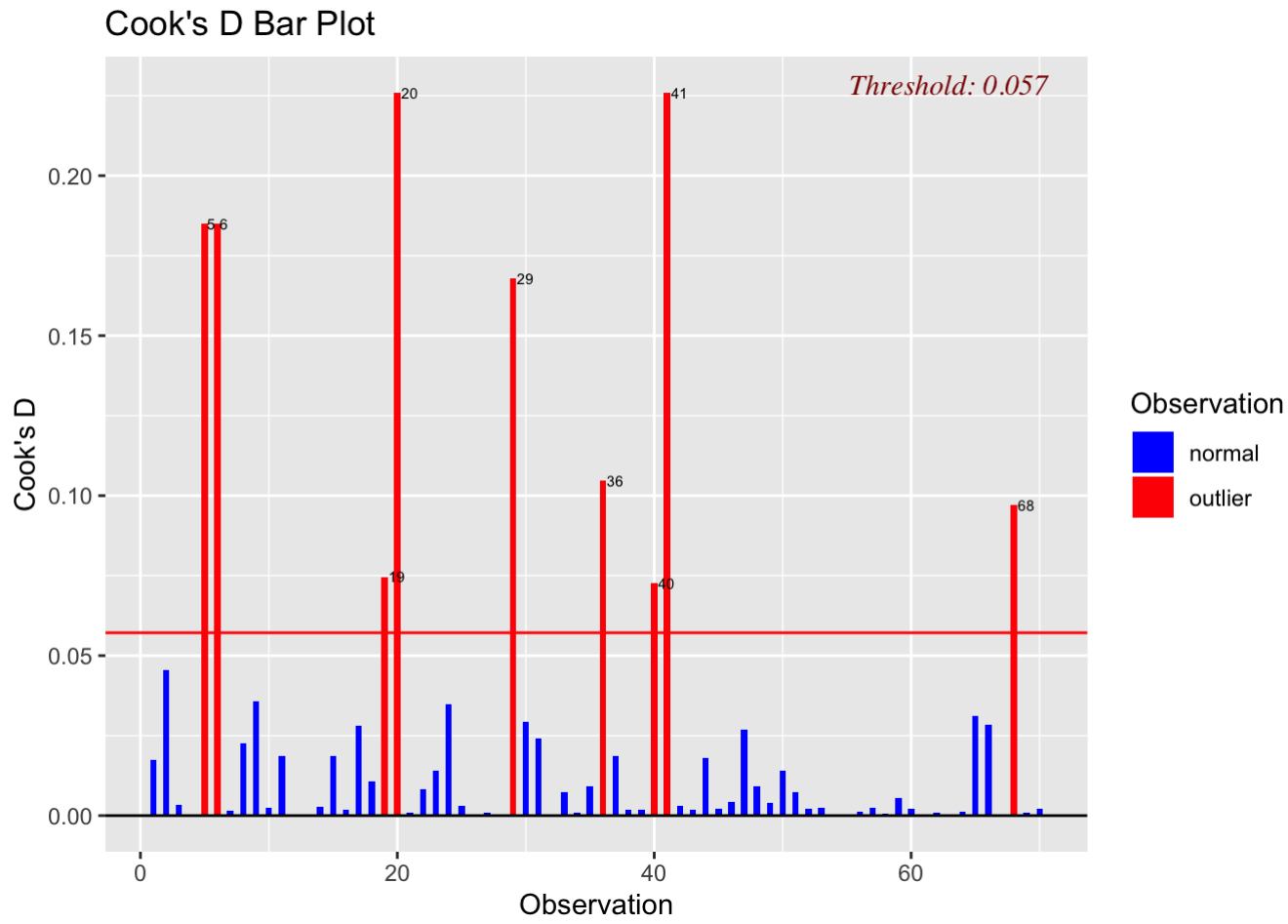


Outliers treated and model built

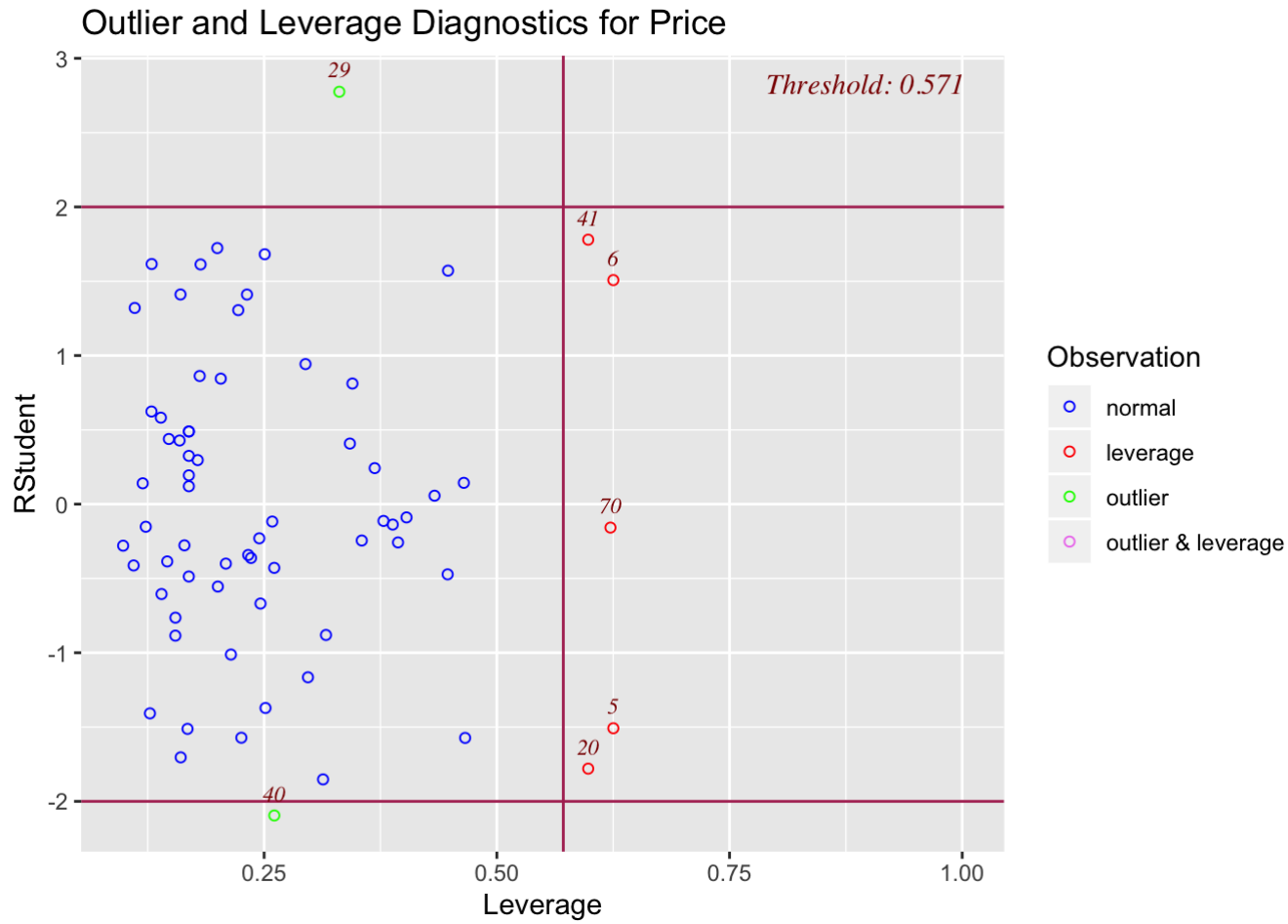
```
n1<-House
n1<-n1[-c(44),]
n1<-n1[-c(30),]
n1<-n1[-c(25),]
n1<-n1[-c(15),]
n1<-n1[-c(32),]
n1<-n1[-c(32),]

modd<-lm(Price~Lot+Size+Year+Bed+Bath+Garage+School,data = n1)
outlierTest(modd)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 32 2.775184          0.0077886      0.52963
```

```
ols_plot_cooksd_bar(modd)
```




```
ols_plot_resid_lev(modd)
```



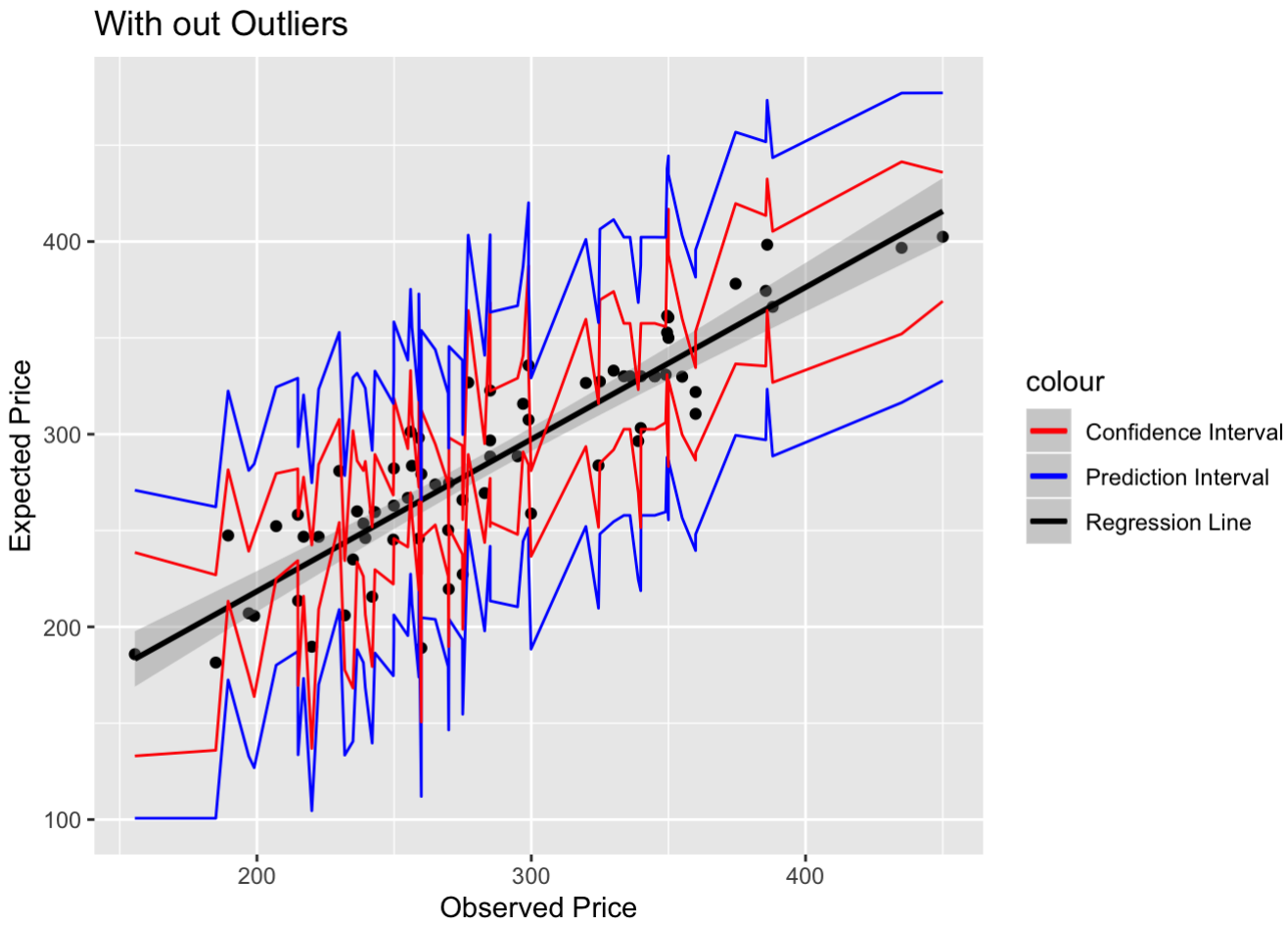
```
ci=predict(mod,level=0.95,interval='confidence')

pi=predict(mod,level=0.95,interval='prediction')
```

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

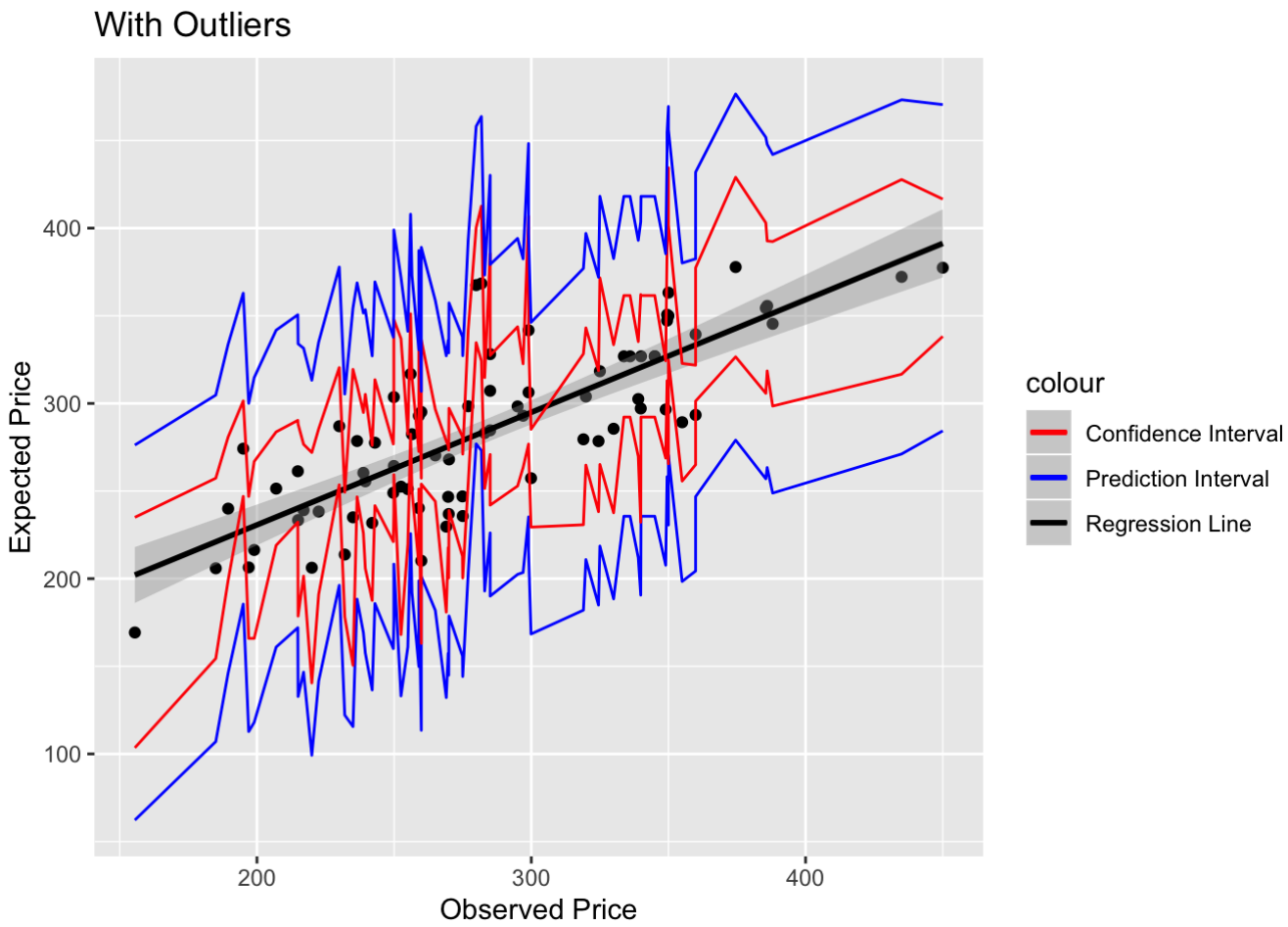
```
cipiplot = ggplot(House, aes(House$Price,pi[,1])) + geom_point() + geom_smooth(method
=lm,aes(color="Regression Line")) + geom_line(aes(y=pi[,2], color="Prediction Interva
l")) +geom_line(aes(y=ci[,2], color="Confidence Interval"))+geom_line(aes(y=ci[,3], c
olor="Confidence Interval")) + geom_line(aes(y=pi[,3], color="Prediction Interval"))
 + labs(x="Observed Price", y="Expected Price")+scale_color_manual(values = c("red",
"blue","black"))+ggtitle("With Outliers")


cinew=predict(modd,level=0.95,interval='confidence')

pinew=predict(modd,level=0.95,interval='prediction')
```

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

```
cipiplotnew = ggplot(n1, aes(n1$Price,pinew[,1])) + geom_point() + geom_smooth(method
=lm,aes(color="Regression Line")) + geom_line(aes(y=pinew[,2], color="Prediction Inte
rval")) +geom_line(aes(y=cinew[,2], color="Confidence Interval"))+geom_line(aes(y=cin
ew[,3], color="Confidence Interval")) + geom_line(aes(y=pinew[,3], color="Prediction
 Interval")) + labs(x="Observed Price", y="Expected Price")+scale_color_manual(values
= c("red","blue","black"))+ggtitle("With out Outliers")
cipiplotnew
```


With out Outliers
Expected Price
Observed Price
100
200
300
400
200
300
400
colour
Confidence Interval
Prediction Interval
Regression Line

```
cipiplot
```


With Outliers
Expected Price
Observed Price
100
200
300
400
200
300
400
colour
Confidence Interval
Prediction Interval
Regression Line