Analyzing The Diamond Price Based on 4 C's

STAT 40001-002/ STAT 59800-001 Statistical Computing

Submitted by

Shahrukh Alam Khan

Rishitha Sunduri

Objective

The objective of our project is to analyze and model the price of the diamond based on carat, colour, clarity and certification.

- Data Source
  - Pricing the C's of Diamond Stones – (https://vincentarelbundock.github.io/Rdatasets/datasets.html)
  - Journal of Statistics Education's data archive : http://www.amstat.org/publications/jse/jse_data_archive.htm.
  - Chu, Singfat (2001) "Pricing the C's of Diamond Stones", Journal of Statistics Education, 9(2).

  This data set consists of weight of diamond stones in carat unit, a factor with levels and certification body.

Attributes

It contains 5 variables with 4 regressor and 1 response.

Regressor Variable

- Carat (Carat Unit)
- Color (D,E,F,G,H,I)
- Clarity (IF,VVS1,VVS2,VS1,VS2)
- Certification (GIA,IGI,HRD)

Response Variable

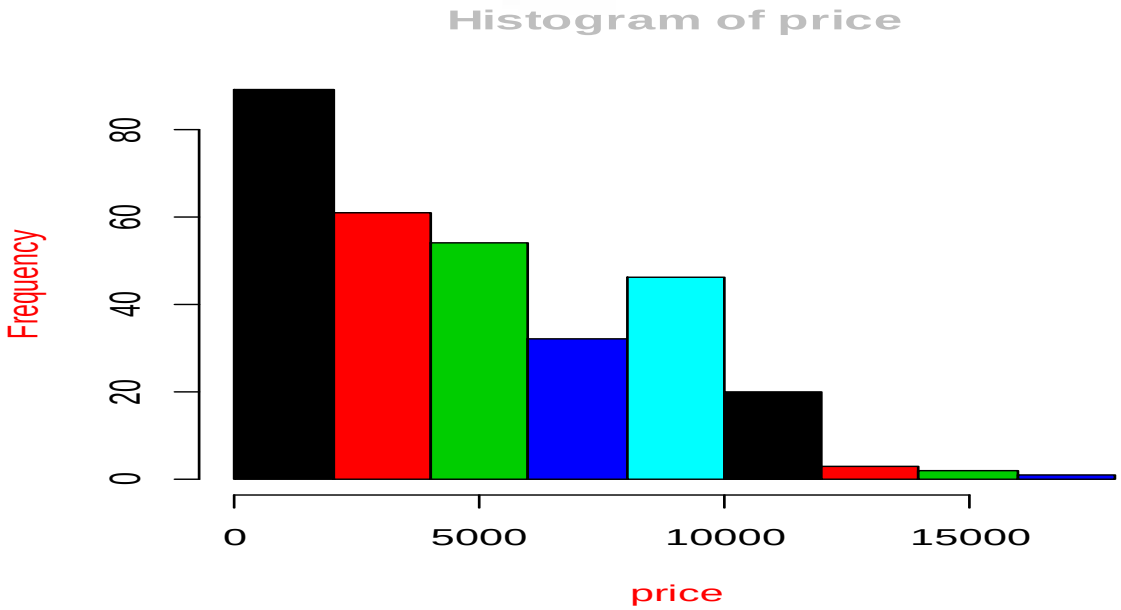
- Price ($)

Data Set

	Carat	Colour	Clarity	Certification	Price
1	0.30	D	VS2	GIA	1302
2	0.30	E	VS1	GIA	1510
3	0.30	G	VVS1	GIA	1510
4	0.30	G	VS1	GIA	1260
5	0.31	D	VS1	GIA	1641
6	0.31	E	VS1	GIA	1555

Histogram of Price

- Calculating the frequency of the price.



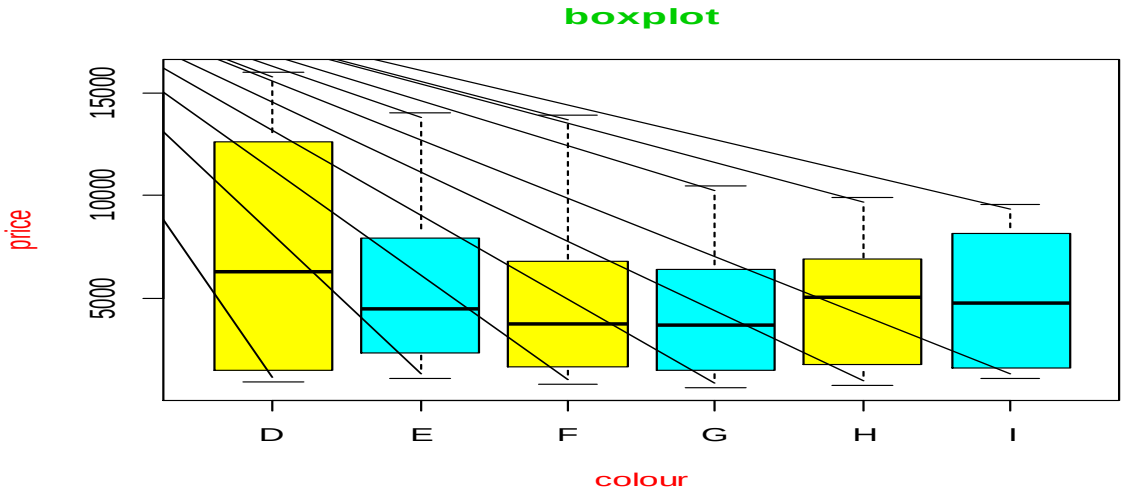
- From the histogram we can see that most of diamond prices are below 10000.

Boxplot

- Price Vs Colour: The variation of price depending on the colour.

Colour sequence from best to worst

D
E
F
G
H
I



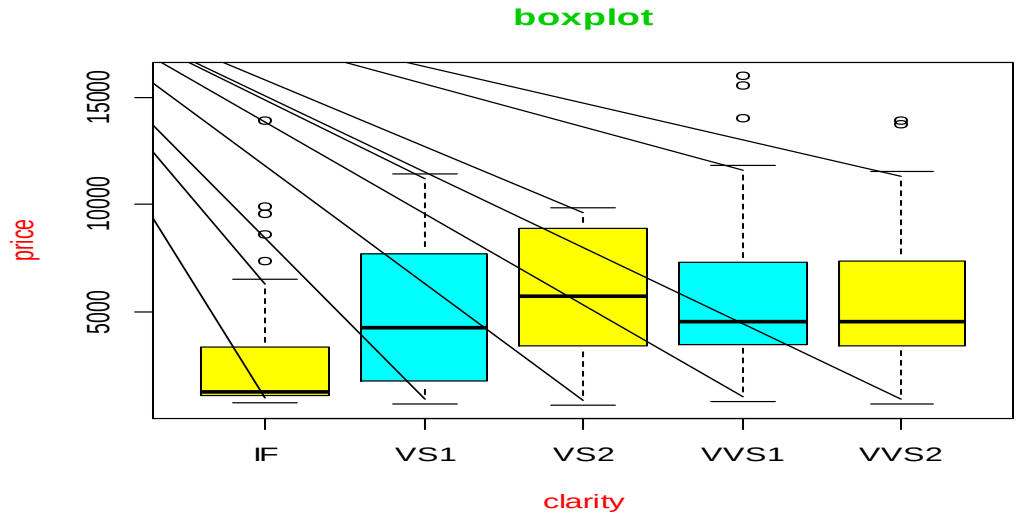
- From the above boxplot, we can conclude that prices of Diamonds are highly depend on the colour of the Diamond.

Boxplot - Contd

- Price Vs Clarity: The variation of price depending on the clarity.

Clarity sequence from best to worst

- IF
- VS1
- VS2
- VVS1
- VVS2



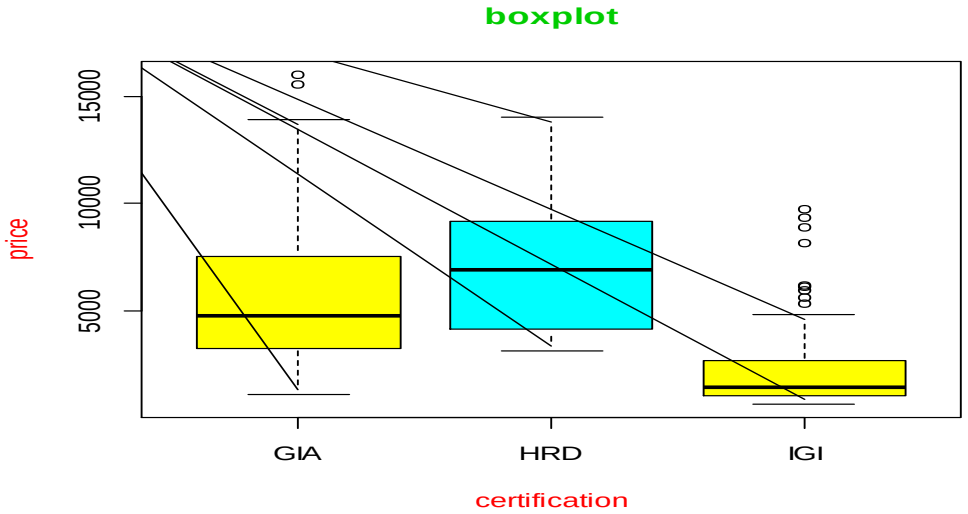
- From the above boxplot, we can conclude that price does not depend on the clarity.

Boxplot - Contd

- Price Vs Certification: The variation of price depending on the certification.

Certification sequence from best to worst

- GIA
- HRD
- IGI

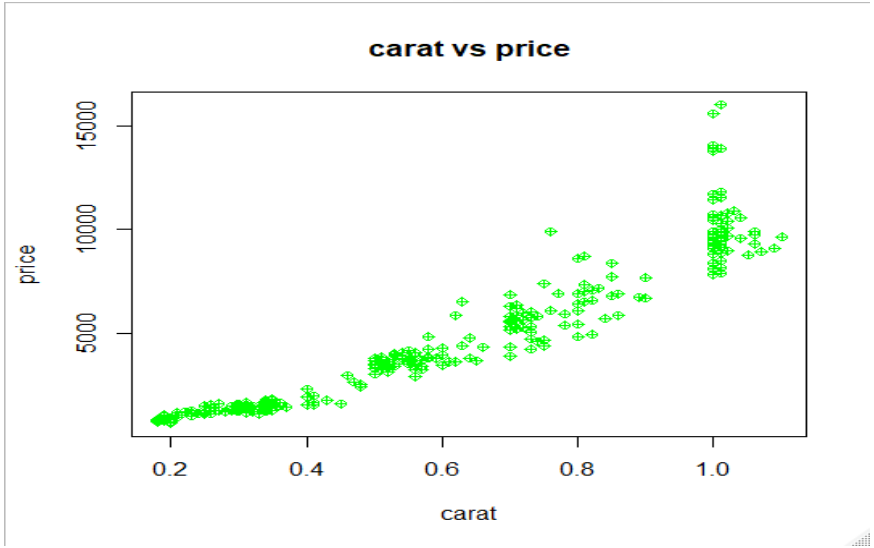


- Certification:

A Diamond Grading Report or Diamond Dossier® is a report created by a team of gemologists.

Scatter Plot

- Price Vs Carat: The variation of price depending on the carat.



- From the above boxplot, we can conclude that price of diaomonds is highly depend on carat.

T.Test

- Testing whether price is dependent on the colour

```
> D=subset(data,colour=="D")
> I=subset(data,colour=="I")
>t.test(D$price,I$price)
Welch Two Sample t-test
data:  D$price and I$price
t = 1.2982, df = 18.704, p-value = 0.21
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1181.727  5031.427
sample estimates:
mean of x mean of y
7099.875  5175.025
```

- From above test, the p value is higher than 0.05 so we failed to reject the null hypothesis and conclude that we don't have enough evidence to tell that price is dependent on colour.

T.Test - Contd

- Testing whether price is dependent on the clarity using high and low clarity levels.

```
> IF=subset(data,clarity=="IF")
> VS2=subset(data,clarity=="VS2")
>t.test(IF$price,alt="greater",VS2$price)
 Welch Two Sample t-test
data:  IF$price and VS2$price
t = -5.1172, df = 92.098, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
-4190.556       Inf
sample estimates:
mean of x mean of y
2694.773  5858.170
```

- From above test, the p value is higher than 0.05 so we failed to reject the null hypothesis and conclude that we don't have enough evidence to tell that price is dependent on clarity.

T.Test - Contd

- We will test whether the average price is 6000.

```
> t.test(price,alt="greater",mu=6000)

        One Sample t-test

data:  price
t = -5.0565, df = 307, p-value = 1
alternative hypothesis: true mean is greater than 6000
95 percent confidence interval:
 4699.564      Inf
sample estimates:
mean of x
 5019.484
```

- From above test, since p value is greater than 0.05 so we fail to reject the null hypothesis and we conclude that we don't have enough evidence to say that average price is greater than 6000.

T.Test - Contd

- Constructing the 95% confidence interval for the price.

```
> t.test(price,conf.level=0.95)
One Sample t-test
data: price
t = 25.886, df = 307, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
4637.922 5401.046
sample estimates:
mean of x
5019.484
```

- From above test, the 95% confidence interval is (4637.922, 5401.046).

Model Analysis

- We are using the model to predict the price depending on the colour, clarity, certification and carat.
- Here we are using multiple linear regression model.
- Because we have multiple variables and price is a continiuos variable.

Model Analysis - Contd

PURDUE UNIVERSITY NORTHWEST

Model Building:

We are building the model based on Price , Clarity, Carat, Colour and Certification.

```
> model=lm(price~carat+factor(colour)+factor(clarity)+factor(certification))
```

SUMMARY

```
> summary(model)

Call:
lm(formula = price ~ carat + factor(colour) + factor(clarity) +
    factor(certification))

Residuals:
     Min       1Q   Median       3Q      Max
-1740.0  -428.8  -128.3   314.3  3634.1

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                169.18     255.02   0.663    0.508
carat                    12766.40     190.02  67.183   < 2e-16 ***
factor(colour)E          -1439.09     207.98  -6.919 2.83e-11 ***
factor(colour)F          -1841.69     195.23  -9.433   < 2e-16 ***
factor(colour)G          -2176.67     200.39 -10.862   < 2e-16 ***
factor(colour)H          -2747.15     202.91 -13.538   < 2e-16 ***
factor(colour)I          -3313.10     212.71 -15.575   < 2e-16 ***
factor(clarity)VS1       -1474.57     159.67  -9.235   < 2e-16 ***
factor(clarity)VS2       -1792.01     171.19 -10.468   < 2e-16 ***
factor(clarity)VVS1       -689.29     159.93  -4.310 2.23e-05 ***
factor(clarity)VVS2      -1191.16     148.76  -8.007 2.73e-14 ***
factor(certification)HRD    15.23     107.25   0.142    0.887
factor(certification)IGI   141.26     128.26   1.101    0.272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 710.4 on 295 degrees of freedom
Multiple R-squared:  0.9581,    Adjusted R-squared:  0.9564
F-statistic: 562.5 on 12 and 295 DF,  p-value: < 2.2e-16
```

Model Analysis - Contd

Model Building:

ANOVA

```
> anova(model)
Analysis of Variance Table

Response: price
                      Df     Sum Sq    Mean Sq   F value Pr(>F)
carat                  1 3173248722 3173248722 6287.9809 <2e-16 ***
factor(colour)         5  142360234   28472047   56.4191 <2e-16 ***
factor(clarity)        4   90327430   22581857   44.7473 <2e-16 ***
factor(certification)  2     618313     309156    0.6126 0.5426
Residuals            295  148872648     504653
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Analysis - Continued

- From the summary of the model we got R square values as

R-squared: 0.9581, and Adjusted R-squared: 0.9564.

```
>summary(model)$r.squared
[1] 0.9581281
```

- We build the Anova of the model and we conclude that certification is not a significant variable.

Model Improvement:

We ran stepAIC for our model to check for the insignificant variable.

STEP AIC

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> stepAIC=stepAIC(model,direction="both")
Start:  AIC=4057.26
price ~ carat + factor(colour) + factor(clarity) + factor(certification)

                        Df  Sum of Sq        RSS     AIC
- factor(certification)  2      618313  149490961 4054.5
<none>                                  148872648 4057.3
- factor(clarity)        4    68244838  217117487 4165.5
- factor(colour)         5   169058942  317931590 4281.0
- carat                  1  2277773876 2426646525 4914.9

Step:  AIC=4054.54
price ~ carat + factor(colour) + factor(clarity)

                        Df  Sum of Sq        RSS     AIC
<none>                                  149490961 4054.5
+ factor(certification)  2      618313  148872648 4057.3
- factor(clarity)        4    90327430  239818391 4192.1
- factor(colour)         5   170206583  319697545 4278.7
- carat                  1  3001644888 3151135849 4991.4
```

Insignificant variable is found to be Certification

Model Improvement:

We are building the model of Price with Clarity, Carat, and Colour.

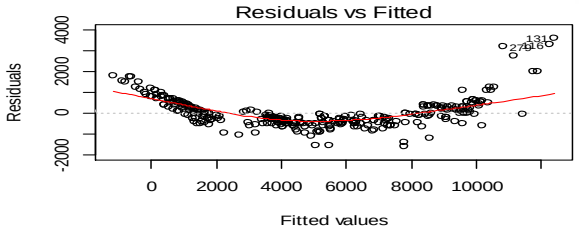
```
> model=lm(price~carat+factor(colour)+factor(clarity))
```

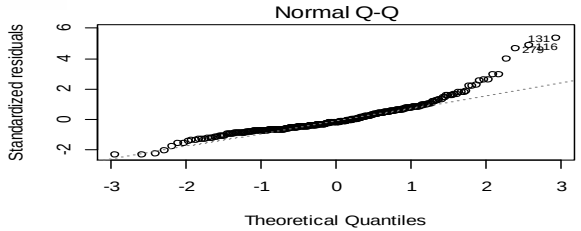
```
> summary(model)

Call:
lm(formula = price ~ carat + factor(colour) + factor(clarity))

Residuals:
    Min      1Q  Median      3Q     Max
-1613.5  -455.2  -148.8   328.3  3614.5

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)            316.8      216.6   1.463    0.145
carat                12683.8      164.2  77.224   < 2e-16 ***
factor(colour)E      -1447.5      207.3  -6.984 1.89e-11 ***
factor(colour)F      -1843.9      194.6  -9.473   < 2e-16 ***
factor(colour)G      -2178.8      199.6 -10.915   < 2e-16 ***
factor(colour)H      -2763.2      201.3 -13.726   < 2e-16 ***
factor(colour)I      -3315.9      212.4 -15.610   < 2e-16 ***
factor(clarity)VS1   -1548.8      143.8 -10.773   < 2e-16 ***
factor(clarity)VS2   -1860.7      158.9 -11.712   < 2e-16 ***
factor(clarity)VVS1   -733.9      153.8  -4.771 2.88e-06 ***
factor(clarity)VVS2  -1235.4      143.1  -8.633 3.72e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 709.5 on 297 degrees of freedom
Multiple R-squared:  0.958,     Adjusted R-squared:  0.9565
F-statistic: 676.7 on 10 and 297 DF,  p-value: < 2.2e-16
```

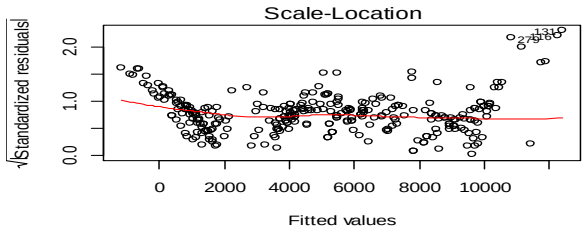
The R square value is same as previous model R square value (R-squared: 0.958, and Adjusted R-squared: 0.9565). So we have to find better model.

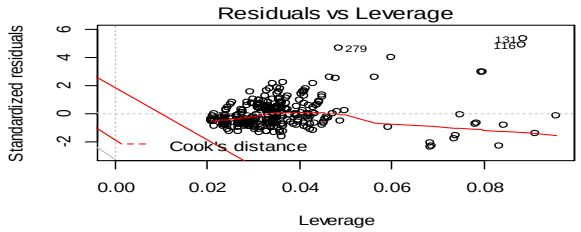
Residual Analysis

The fitted model is,





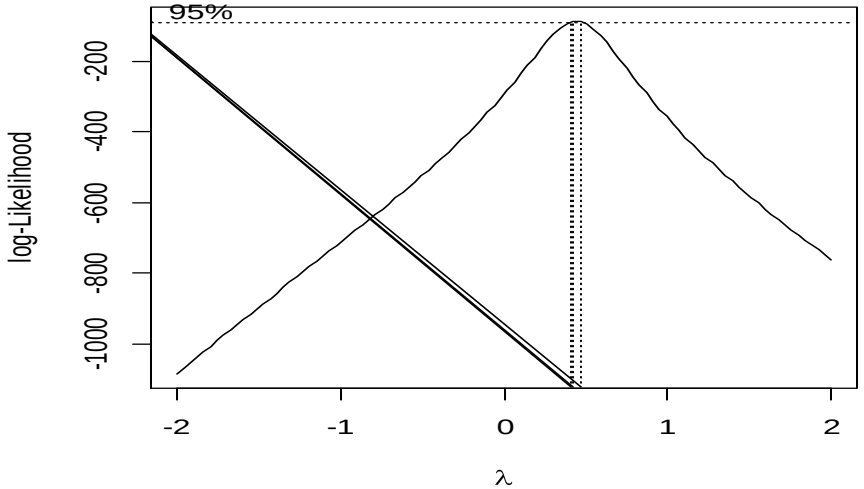




Box – Cox Transformation

- The residual plot has structured pattern that the data do not follow the underlying assumptions of the model.
- To improve the model we are using Box – Cox transformation.

```
>b=boxcox(model)
```



- It appears that lambda should be approximately 0.5.

Model Analysis

Model Improvement:

> new_price=price^0.5

> model3=lm(new_price ~ carat + factor(colour) + factor(clarity))

```
> summary(model3)

Call:
lm(formula = new_price ~ carat + factor(colour) + factor(clarity))

Residuals:
    Min      1Q  Median      3Q     Max
-5.6890 -1.4619 -0.0677  1.2392 11.3275

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          23.6727     0.7388  32.044  < 2e-16 ***
carat                92.6820     0.5602 165.440  < 2e-16 ***
factor(colour)E      -5.3900     0.7069  -7.624 3.33e-13 ***
factor(colour)F      -8.1199     0.6639 -12.231  < 2e-16 ***
factor(colour)G     -10.6551     0.6809 -15.649  < 2e-16 ***
factor(colour)H     -13.8782     0.6866 -20.212  < 2e-16 ***
factor(colour)I     -18.1974     0.7245 -25.117  < 2e-16 ***
factor(clarity)VS1   -7.5305     0.4903 -15.358  < 2e-16 ***
factor(clarity)VS2  -10.0981     0.5419 -18.634  < 2e-16 ***
factor(clarity)VVS1  -2.0863     0.5247  -3.976 8.80e-05 ***
factor(clarity)VVS2  -5.3791     0.4881 -11.020  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.42 on 297 degrees of freedom
Multiple R-squared:  0.9907,    Adjusted R-squared:  0.9904
F-statistic:  3166 on 10 and 297 DF,  p-value: < 2.2e-16
```

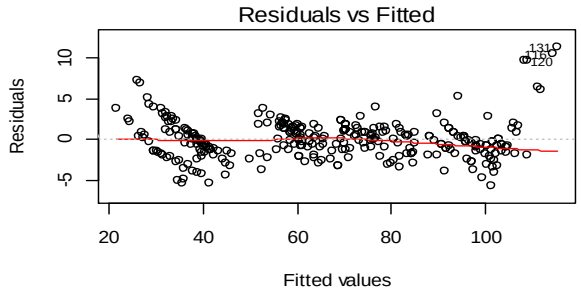
The R squared value has now increased to 0.9907 (R-squared: 0.9907, Adjusted R-squared: 0.9904).

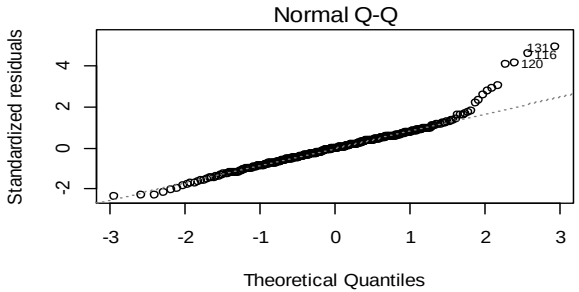
From

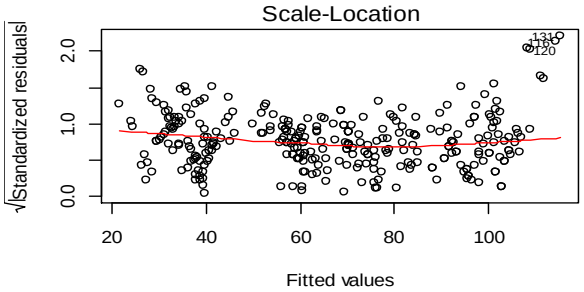
(R-squared: 0.958, and Adjusted R-squared: 0.9565).

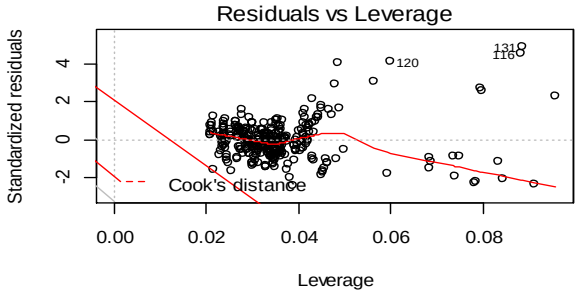
Residual Analysis of the Updated Model

The fitted model is,









Residual Analysis

- We can observe that the residual plots after the transformation is structure less and the co-efficient of determination after the transformation is increased compared to before the transformation.
- Thus we can conclude that transformation helped to improve our model.

Prediction of data using model

- We are predicting the price of a diamond by using carat = 0.75, colour = I and clarity = VS1.

```
> predict(model3,data.frame(carat=0.75,colour="I",clarity="VS1"),type="resp")
1
67.45631
>predict(model3,data.frame(carat=0.75,colour="I",clarity="VS1"),type="resp",interval="pred",level=0.95)
      fit       lwr      upr
1 67.45631 62.60827 72.30435
>predict(model3,data.frame(carat=0.75,colour="I",clarity="VS1"),type="resp",interval="conf",level=0.95)
      fit       lwr      upr
1 67.45631 66.54805 68.36457
```

- The predicted diamond value is equal to (67.45631)^2 = 4550, the original diamond value in our data set is 4335.
- So our model is predicting approximately the real diamond value.

Conclusion

- In this project, we analyzed and modeled the price of a diamond based on carat, colour, clarity and certification.
- Also we did the several T test to conclude which factor effecting the price of the diamond.
- Using our model, we determined the price of diamond.

PURDUE UNIVERSITY NORTHWEST

Thank You