**Statistical Analysis Report**

**Author:** Aurora

**Background**

Diamond is very precious and expensive, and the price of diamond is related to many factors. So it is necessary to find out which kind of factor could impact the diamond price. This analysis will seek to discover some variables to and build the model to explain the price of the diamond. It is helpful to the market to predict the price of diamond.

**Data Source**

Data is about the price of diamond. There are 7 variables, price (price the diamond sold for), carat (size of diamond in carats), clarity (a numerical measure of clarity associated with standard measures in diamonds), color (a numerical measure of colour, also using standard diamond evaluations), cut (a numeric measure of quality of cut), source (the diamond manufacturing who mined, graded and cut the diamond), year (the year the diamond was first cut)

**Data Transformation and Cleaning (Description)**

**Price**

Price was transformed from integer to numeric.

**Clarity**

Clarity was transformed from integer to numeric.

**Color**

Color was transformed from integer to numeric.

**Cut**

Cut was transformed from integer to numeric.

**Source**

The data identifying the diamond manufacturing who mined, graded and cut the diamond was transformed to four dummy variables.

**Year**

Year was transformed from integer to numeric.

**Descriptive Data Analysis**

Price Carat Clarity Color

Min. : 1000 Min. :0.3000 Min. :1.000 Min. : 4.000

1st Qu.: 1801 1st Qu.:0.6000 1st Qu.:2.000 1st Qu.: 5.000

Median : 3604 Median :0.9000 Median :3.000 Median : 7.000

Mean : 3971 Mean :0.8701 Mean :3.235 Mean : 6.997

3rd Qu.: 5544 3rd Qu.:1.0600 3rd Qu.:5.000 3rd Qu.: 9.000

Max. :10000 Max. :2.0200 Max. :7.000 Max. :11.000

Cut Source Year SourceAlrosa SourceDeBeers

Min. :1.000 Alrosa :657 Min. :1963 Min. :0.0000 Min. :0.000

1st Qu.:2.000 DeBeers :651 1st Qu.:1976 1st Qu.:0.0000 1st Qu.:0.000

Median :2.000 Debswana:706 Median :1989 Median :0.0000 Median :0.000

Mean :2.449 RioTinto:676 Mean :1990 Mean :0.2442 Mean :0.242

3rd Qu.:3.000 3rd Qu.:2003 3rd Qu.:0.0000 3rd Qu.:0.000

Max. :4.000 Max. :2017 Max. :1.0000 Max. :1.000

SourceDebswana SourceRioTinto

Min. :0.0000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000

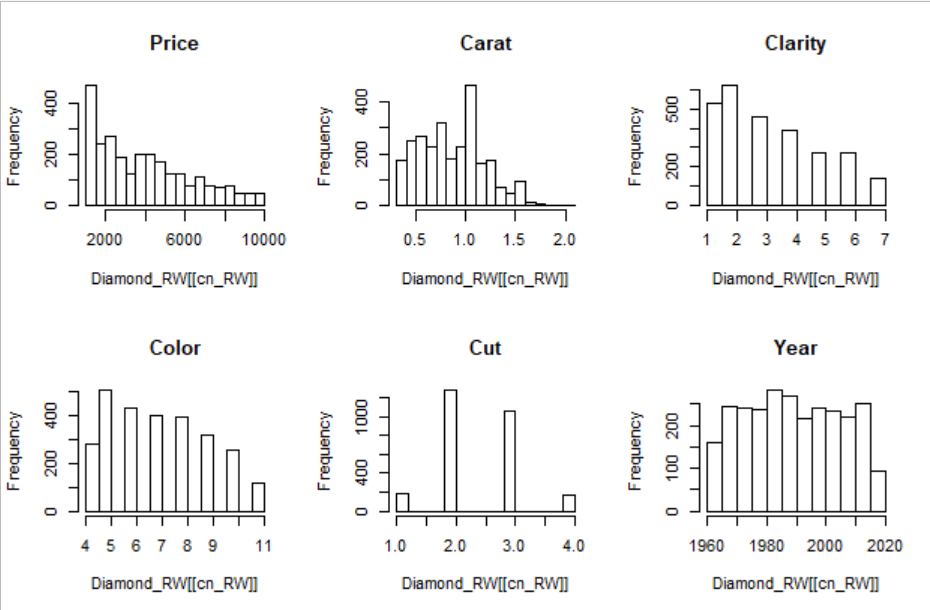
Median :0.0000 Median :0.0000

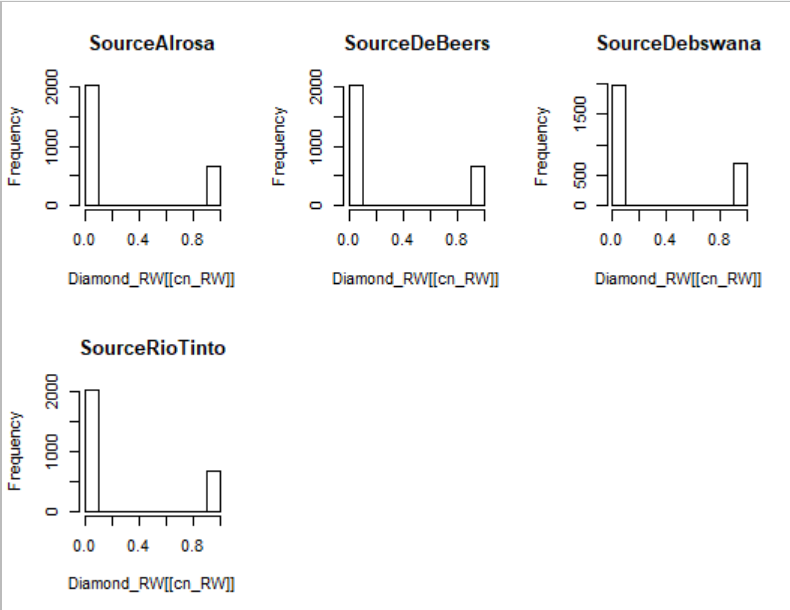
Mean :0.2625 Mean :0.2513

3rd Qu.:1.0000 3rd Qu.:1.0000

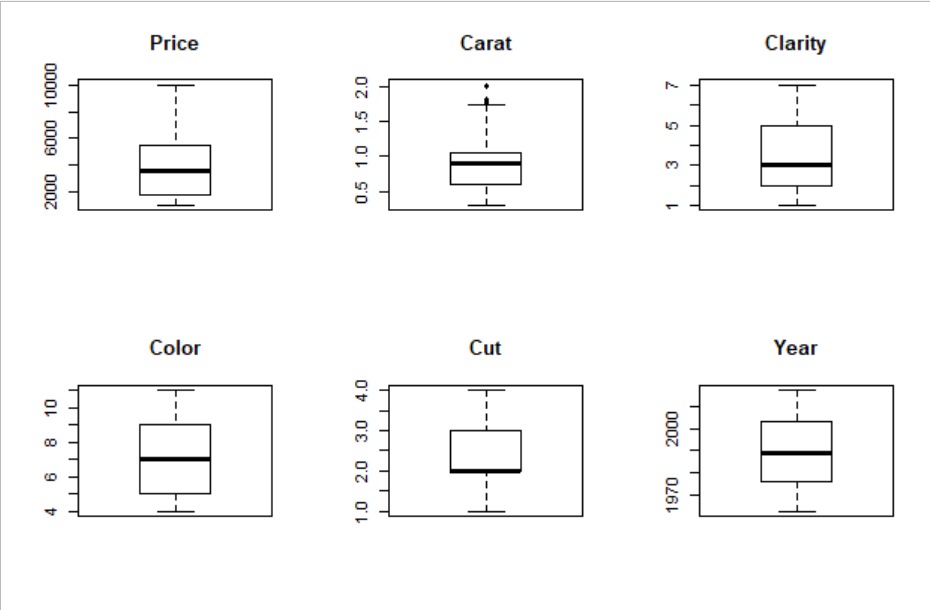
Max. :1.0000 Max. :1.0000

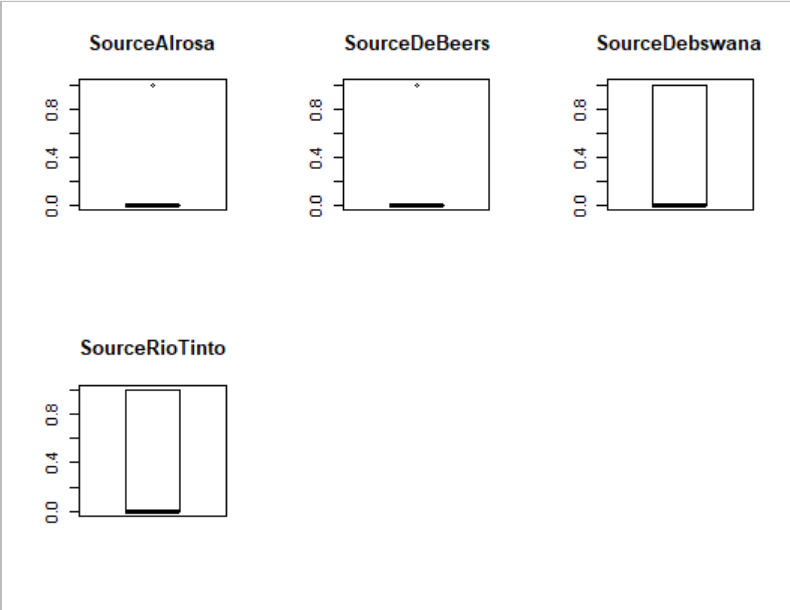
From the summary statistics we conclude that the transformation of data worked properly. And all of the data looks reasonable. There are no values that seems to be necessarily wrong. About the carat, the maximum seems to be a little higher. Maybe there are some extreme values.





**Outlier**





There seem to be outliers in Carat. And just leave it here, to decide what we could do about it later.

**Exploratory Data Analysis**

statistic p.value

Price 0.9213888 2.416619e-35

Carat 0.9720116 1.572993e-22

Clarity 0.9075301 1.236322e-37

Color 0.9395519 8.533522e-32

Cut 0.8328177 1.253891e-46

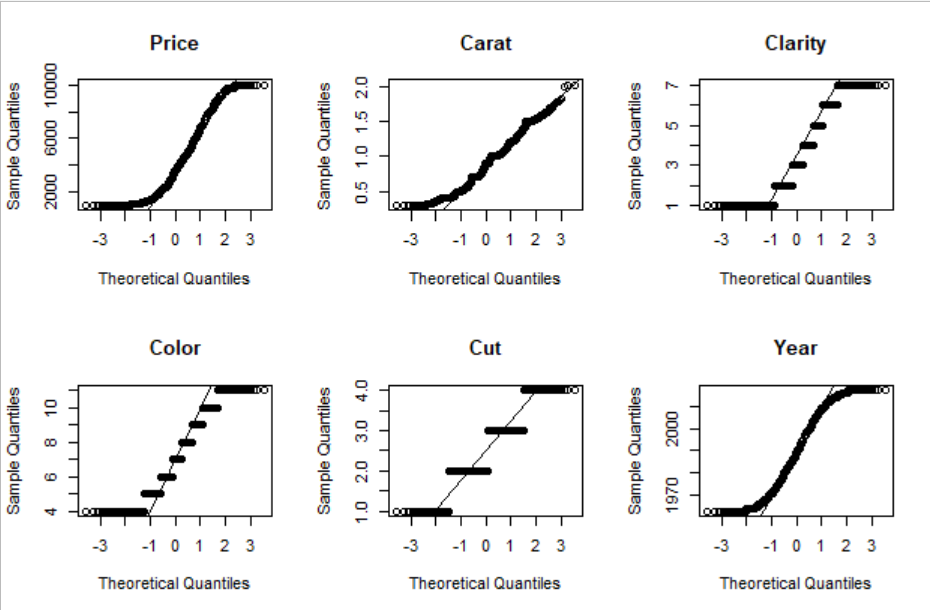
Year 0.9553 5.80826e-28

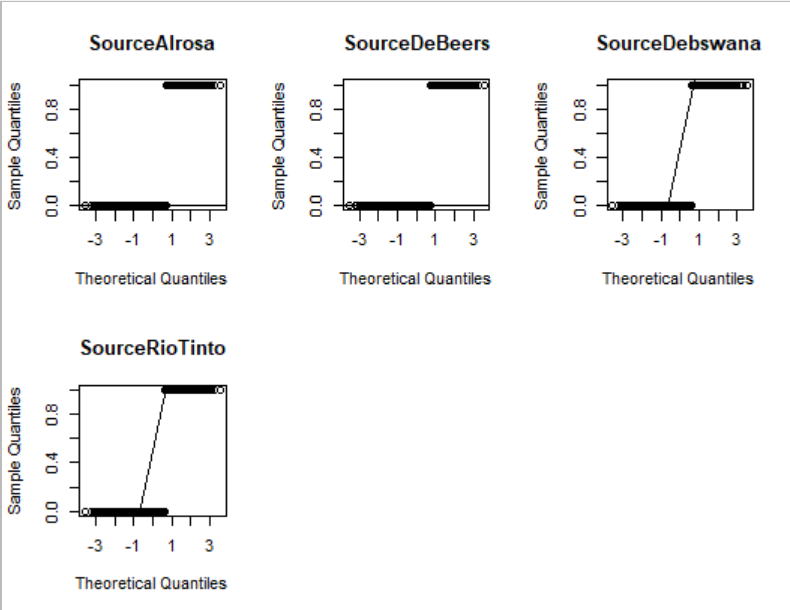
SourceAlrosa 0.5335691 1.392702e-64

SourceDeBeers 0.531601 1.158704e-64

SourceDebswana 0.5487832 5.905137e-64

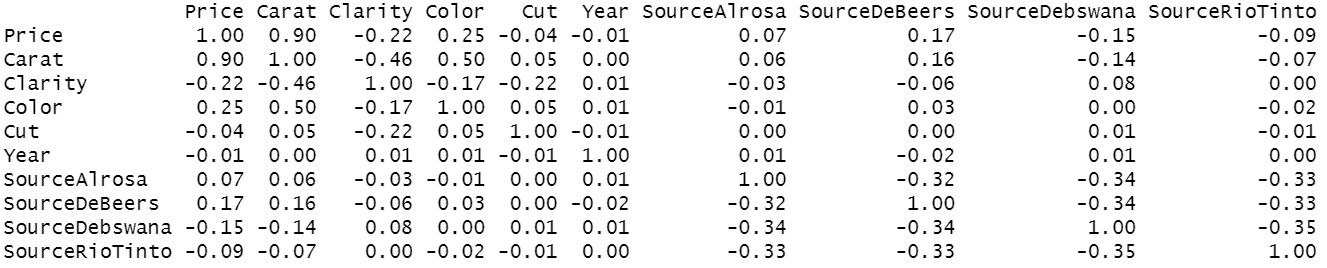
SourceRioTinto 0.5396479 2.468368e-64

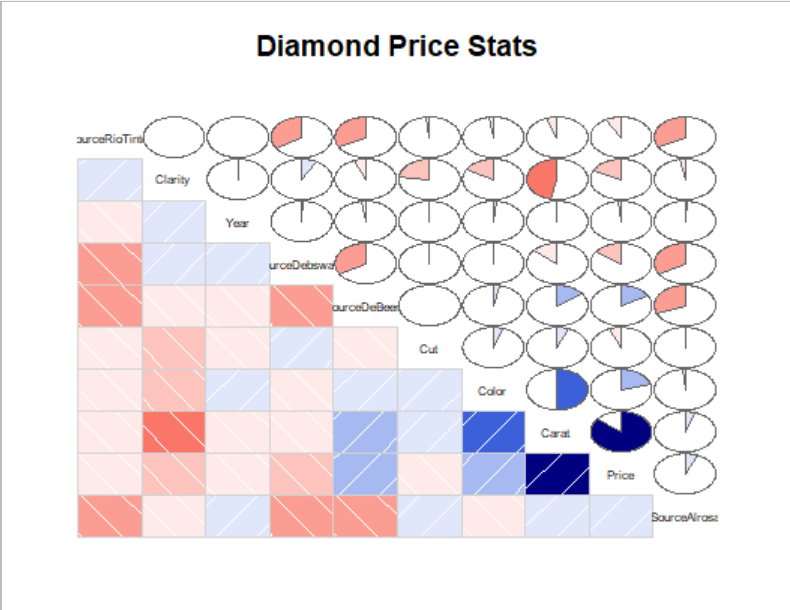




It seems that no variable is normally distributed because of the p value.

Correlations





Price seems to be very strongly positively correlated with Carat. And the positive correlation between Color and Carat is also strong. And there are also positive correlation between:

1 Color and Price.

2 SourceDebeers and Carat.

3 SourceDebeers and Price.

Also, clarity and carat has strongly negative correlation. There are also positive correlation between:

1 Clarity and Cut.

2 Clarity and Color.

3 Clarity and Price.

4 SourceDebswana and Price.

5 SourceDebswana and Carat.

**Models**

**Model 1: All Variables included**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 89.5% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Six variables look significant (p-values of t-test <0.05). Carat, Clarity, Color, Cut, SourceAlrosa, SourceDeBeers
5. Variable Clarity is positively correlated with price instead of negatively.

Call:

lm(formula = Price ~ Carat + Clarity + Color + Cut + Year + SourceAlrosa +

SourceDeBeers + SourceDebswana, data = Diamond\_RW, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-2907.8 -474.9 -82.1 333.2 3536.8

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 28.2840 1895.1384 0.015 0.988094

Carat 8783.3676 62.0966 141.447 < 2e-16 \*\*\*

Clarity 403.4900 9.7556 41.360 < 2e-16 \*\*\*

Color -403.0785 8.8292 -45.653 < 2e-16 \*\*\*

Cut -176.5198 21.8572 -8.076 1e-15 \*\*\*

Year -0.9194 0.9517 -0.966 0.334078

SourceAlrosa 139.9556 43.1149 3.246 0.001184 \*\*

SourceDeBeers 151.2230 43.5422 3.473 0.000523 \*\*\*

SourceDebswana 23.9222 42.3178 0.565 0.571917

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 784.4 on 2681 degrees of freedom

Multiple R-squared: 0.8953, Adjusted R-squared: 0.895

F-statistic: 2865 on 8 and 2681 DF, p-value: < 2.2e-16

**Model 2: Backward Selection**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 89.5% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. All six variables (and the intercept) look significant (p-values of t-test <0.001).
5. Variable Clarity is still positively correlated with price instead of negatively.

Call:

lm(formula = Price ~ Carat + Clarity + Color + Cut + SourceAlrosa +

SourceDeBeers, data = Diamond\_RW, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-2923.0 -471.6 -86.2 333.0 3524.4

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1789.396 97.894 -18.279 < 2e-16 \*\*\*

Carat 8782.467 62.040 141.562 < 2e-16 \*\*\*

Clarity 403.637 9.749 41.405 < 2e-16 \*\*\*

Color -403.073 8.821 -45.695 < 2e-16 \*\*\*

Cut -176.183 21.848 -8.064 1.1e-15 \*\*\*

SourceAlrosa 127.666 37.447 3.409 0.000661 \*\*\*

SourceDeBeers 139.850 38.014 3.679 0.000239 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 784.3 on 2683 degrees of freedom

Multiple R-squared: 0.8952, Adjusted R-squared: 0.895

F-statistic: 3820 on 6 and 2683 DF, p-value: < 2.2e-16

**Model 3: Forward Selection**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 89.5% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Six variables (and the intercept) look significant (p-values of t-test <0.01).
5. Variable Clarity is still positively correlated with price instead of negatively.

Call:

lm(formula = Price ~ Carat + Color + Clarity + Cut + SourceDeBeers +

SourceAlrosa, data = Diamond\_RW, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-2923.0 -471.6 -86.2 333.0 3524.4

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1789.396 97.894 -18.279 < 2e-16 \*\*\*

Carat 8782.467 62.040 141.562 < 2e-16 \*\*\*

Color -403.073 8.821 -45.695 < 2e-16 \*\*\*

Clarity 403.637 9.749 41.405 < 2e-16 \*\*\*

Cut -176.183 21.848 -8.064 1.1e-15 \*\*\*

SourceDeBeers 139.850 38.014 3.679 0.000239 \*\*\*

SourceAlrosa 127.666 37.447 3.409 0.000661 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 784.3 on 2683 degrees of freedom

Multiple R-squared: 0.8952, Adjusted R-squared: 0.895

F-statistic: 3820 on 6 and 2683 DF, p-value: < 2.2e-16

**Model 4: Criteria Selection**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 89.5% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Six variables (and the intercept) look significant (p-values of t-test <0.01).
5. Variable Clarity is still positively correlated with price instead of negatively.

Call:

lm(formula = Price ~ Carat + Clarity + Color + Cut + SourceAlrosa +

SourceDeBeers, data = Diamond\_RW, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-2923.0 -471.6 -86.2 333.0 3524.4

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1789.396 97.894 -18.279 < 2e-16 \*\*\*

Carat 8782.467 62.040 141.562 < 2e-16 \*\*\*

Clarity 403.637 9.749 41.405 < 2e-16 \*\*\*

Color -403.073 8.821 -45.695 < 2e-16 \*\*\*

Cut -176.183 21.848 -8.064 1.1e-15 \*\*\*

SourceAlrosa 127.666 37.447 3.409 0.000661 \*\*\*

SourceDeBeers 139.850 38.014 3.679 0.000239 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 784.3 on 2683 degrees of freedom

Multiple R-squared: 0.8952, Adjusted R-squared: 0.895

F-statistic: 3820 on 6 and 2683 DF, p-value: < 2.2e-16

**Model Evaluation**

**Verifying Assumptions**

1. **Independence of Predictors**

The Spearman rho value for Carat, Clarity, Color, Cut, SourceAlrosa, AourceDeBeers are all very low except for Carat and Color which is 0.5. That means there maybe some kind of relation between Carat and Color. The others are independent.

1. **Distribution of Error Terms**

The p value is all very small. So the error terms seem to be not normally distributed of all models.

Shapiro-Wilk normality test

data: DiaRes\_RW

W = 0.93805, p-value < 2.2e-16

Shapiro-Wilk normality test

data: BckDiaRes\_RW

W = 0.9382, p-value < 2.2e-16

Shapiro-Wilk normality test

data: FwdDiaRes\_RW

W = 0.9382, p-value < 2.2e-16

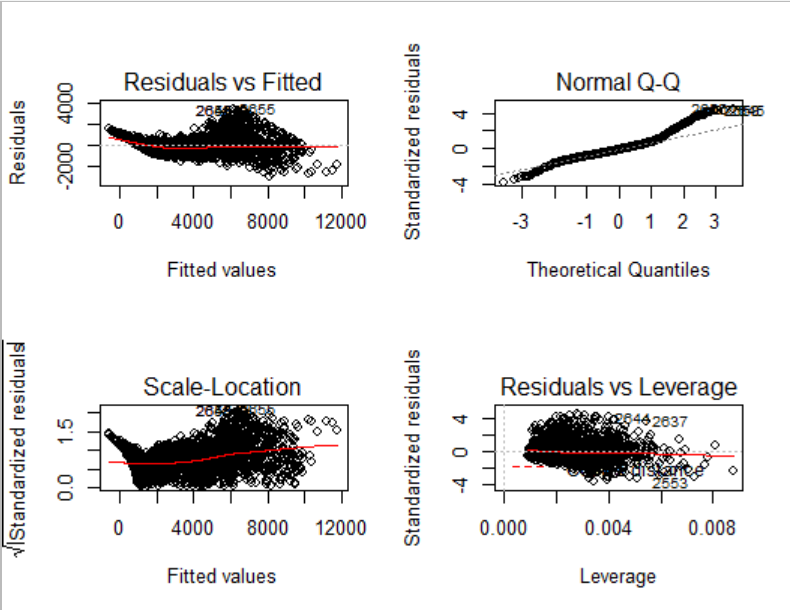
Shapiro-Wilk normality test

data: StpDiaRes\_RW

W = 0.9382, p-value < 2.2e-16

1. **Non-AutoCorrelation and Homoscedasticity**

Based on Residuals vs. Fitted and Scale-Location, the fitted values between 0 and 2000 are all between 0 and 4000. It seems that these data have a pattern. Based on Residuals vs. Leverage and Cook’s Distance, there is no data point exerting undue influence or leverage on the model.



**Final Model, Recommendation and Interpretation**

Based on the above, the results of all four models are similar. But the first model has more variables and the number of variables in other models are the same. So we could pick anyone of these three model. I recommend the following model (developed with Backward selection):

Price =

8782.467\*Carat + 403.637\*Clarity + (-403.073)\*Color + (-176.183)\*Cut +

127.666\*SourceAlrosa + (139.850)\* SourceDeBeers