Multiple Linear Regression

Author: Bharat Thakur

Background

In the first century AD, the Roman naturalist Pliny said, “Diamond is the most valuable, not only of precious stones, but of all things in this world.” But what determines the value of diamond? In this analysis we will attempt to answer this question.

Data Source

Data used in this analysis contains information about diamonds from several of world’s largest mining companies. The various parameters analyzed include:

- Size
- Clarity
- Color
- Quality of cut (Excellent, Good, etc.
- Source
- Insurance Value
- The year the diamond was first cut.

Data Transformation and Cleaning (Description)

The data in the file does not need much processing except for the column counsellor.

Source

The data identifying the diamond source was transformed to four dummy variables.

The code used for transformations is included in the Code Report.

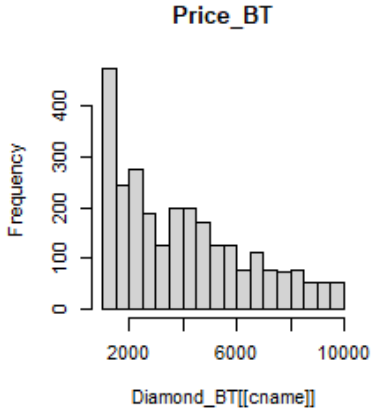
Descriptive Data Analysis

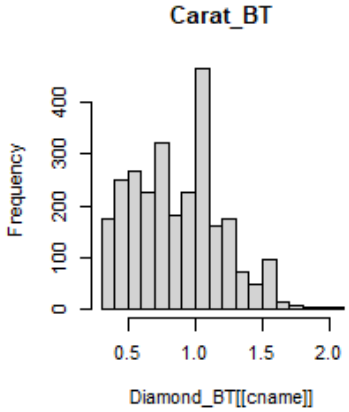
```
   Price_BT        Carat_BT         Year_BT       Clarity_BT
Min.   : 1000   Min.   :0.3000   Min.   :1963   Min.   :-0.1948
1st Qu.: 1801   1st Qu.:0.6000   1st Qu.:1976   1st Qu.: 1.6892
Median : 3604   Median :0.9000   Median :1989   Median : 2.9441
Mean   : 3971   Mean   :0.8701   Mean   :1990   Mean   : 3.2389
3rd Qu.: 5544   3rd Qu.:1.0600   3rd Qu.:2003   3rd Qu.: 4.6163
Max.   :10000   Max.   :2.0200   Max.   :2017   Max.   : 8.4893
   Color_BT          Cut_BT            Val_BT         Alrosa_BT
Min.   : 2.733   Min.   :-0.2387   Min.   : 1.514   Min.   :0.0000
1st Qu.: 5.289   1st Qu.: 1.8816   1st Qu.: 3.393   1st Qu.:0.0000
Median : 6.862   Median : 2.4107   Median : 4.672   Median :0.0000
Mean   : 6.999   Mean   : 2.4506   Mean   : 4.679   Mean   :0.2442
3rd Qu.: 8.605   3rd Qu.: 3.0228   3rd Qu.: 5.770   3rd Qu.:0.0000
```

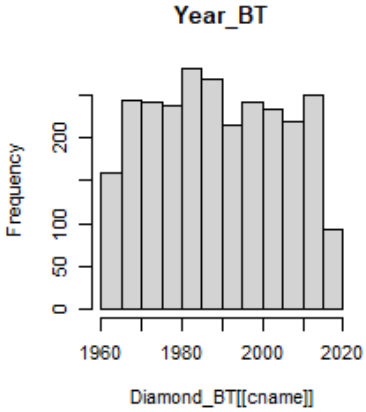
```
Max.   :11.952   Max.   : 5.1077   Max.   :10.170   Max.   :1.0000
  DeBeers_BT       Debswana_BT        Petra_BT        RioTinto_BT
Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000
1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000
Median :0.00000   Median :0.0000   Median :0.00000   Median :1.000
Mean   :0.01152   Mean   :0.0342   Mean   :0.03532   Mean   :0.655
3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:1.000
Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.000
 Rockwell_BT
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.0197
3rd Qu.:0.0000
Max.   :1.0000
```

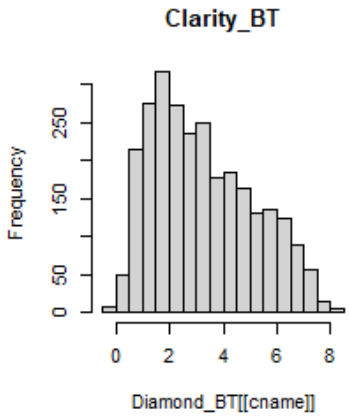
The statistical summary shows that Price, Clarity tend to skew to the right whereas cut appears to be normally distributed. Other than all the data looks reasonable and apparently there are no major anomalies.

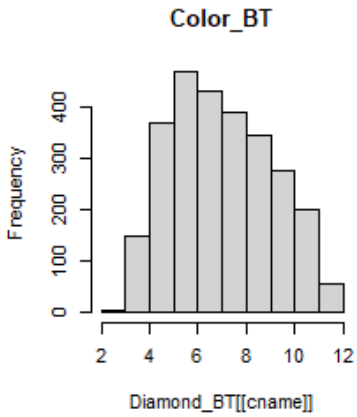
The histograms confirm the above mentioned observations. Furthermore, it can be observed that there are extreme values in Carat, Cut and Val which will be influence the outcomes. Additionally, it is interesting to note that Year is uniformly distributed suggesting that on an average every year equal number of diamonds are cut. Dummy variables are excluded from histograms since they are not expected to give meaningful plots because of the transformation, instead Source is analyzed separately.

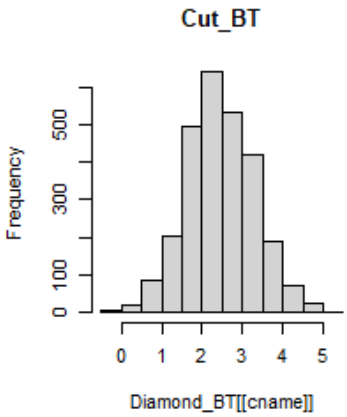


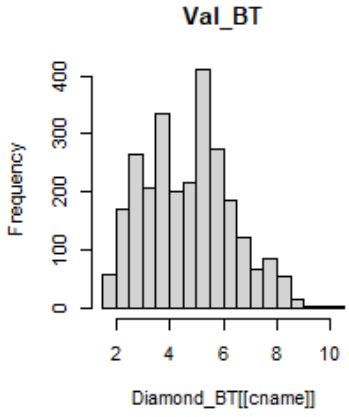




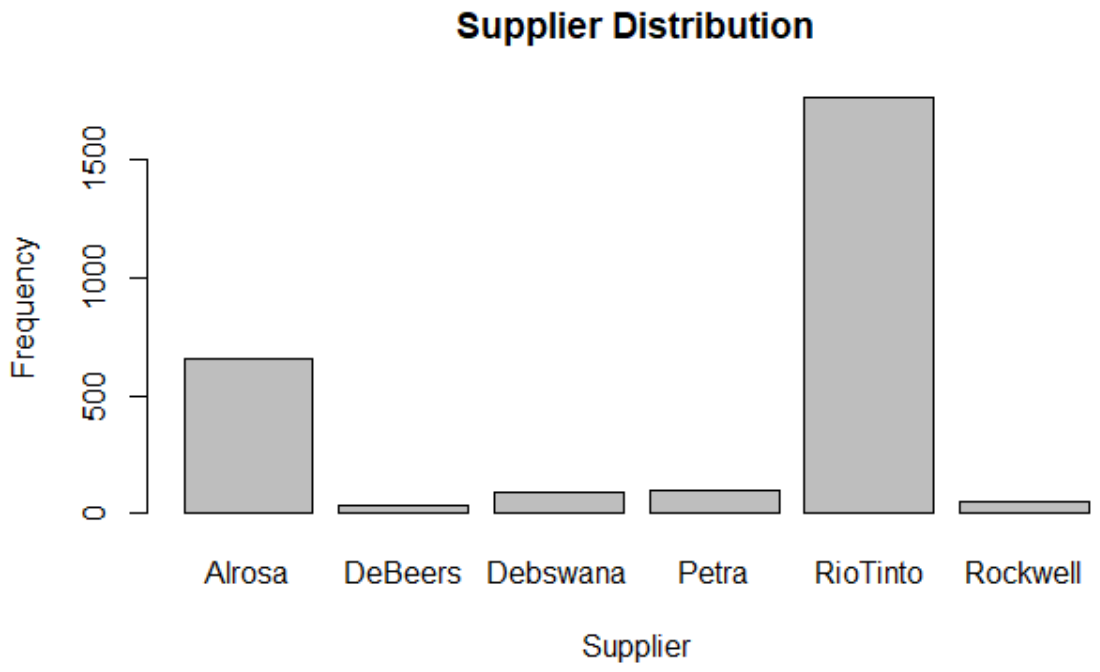




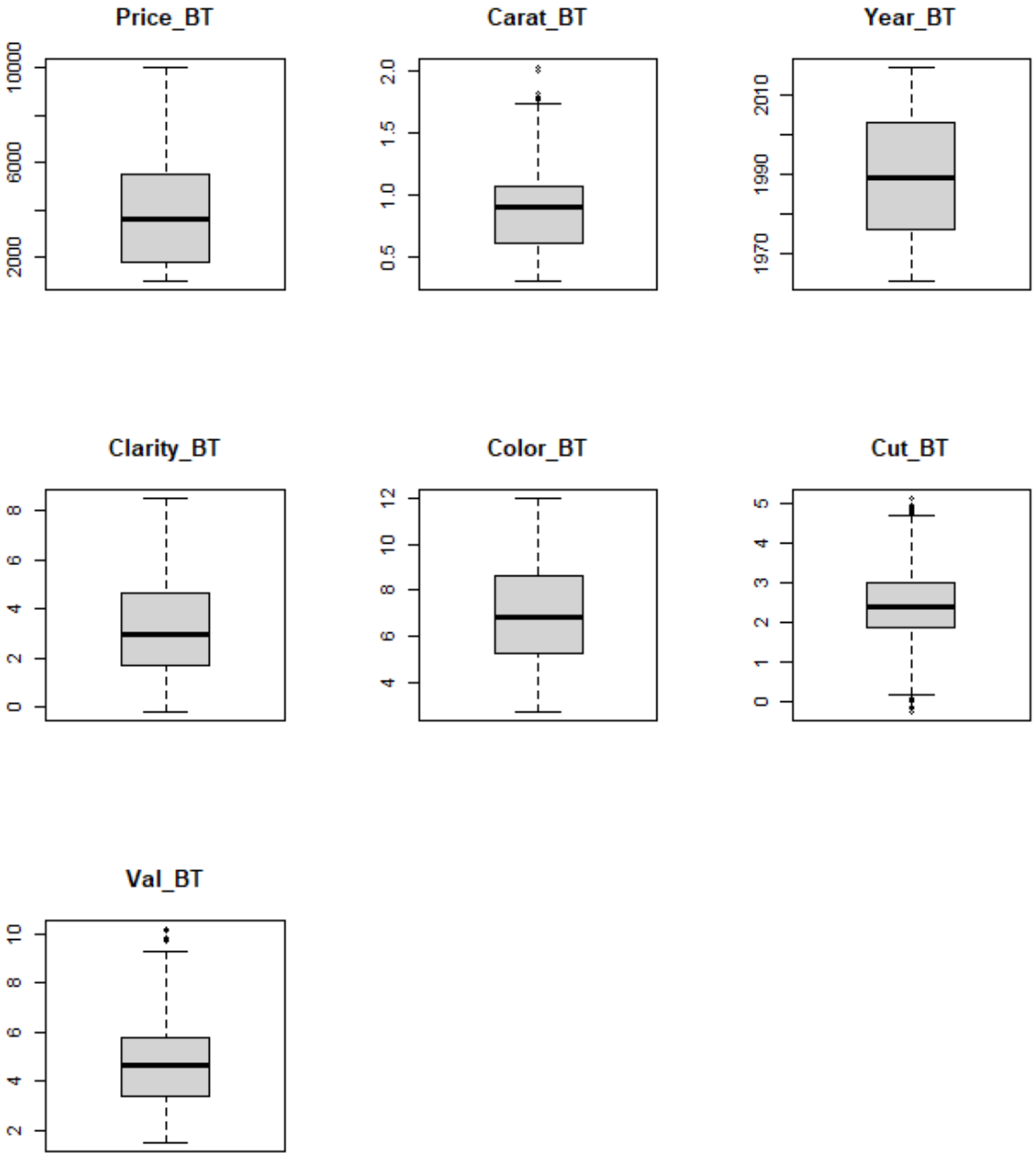




Following plot shows that in the given sample dataset most of the diamonds are sourced from RioTinto and Alrosa, RioTinto being the biggest supplier.

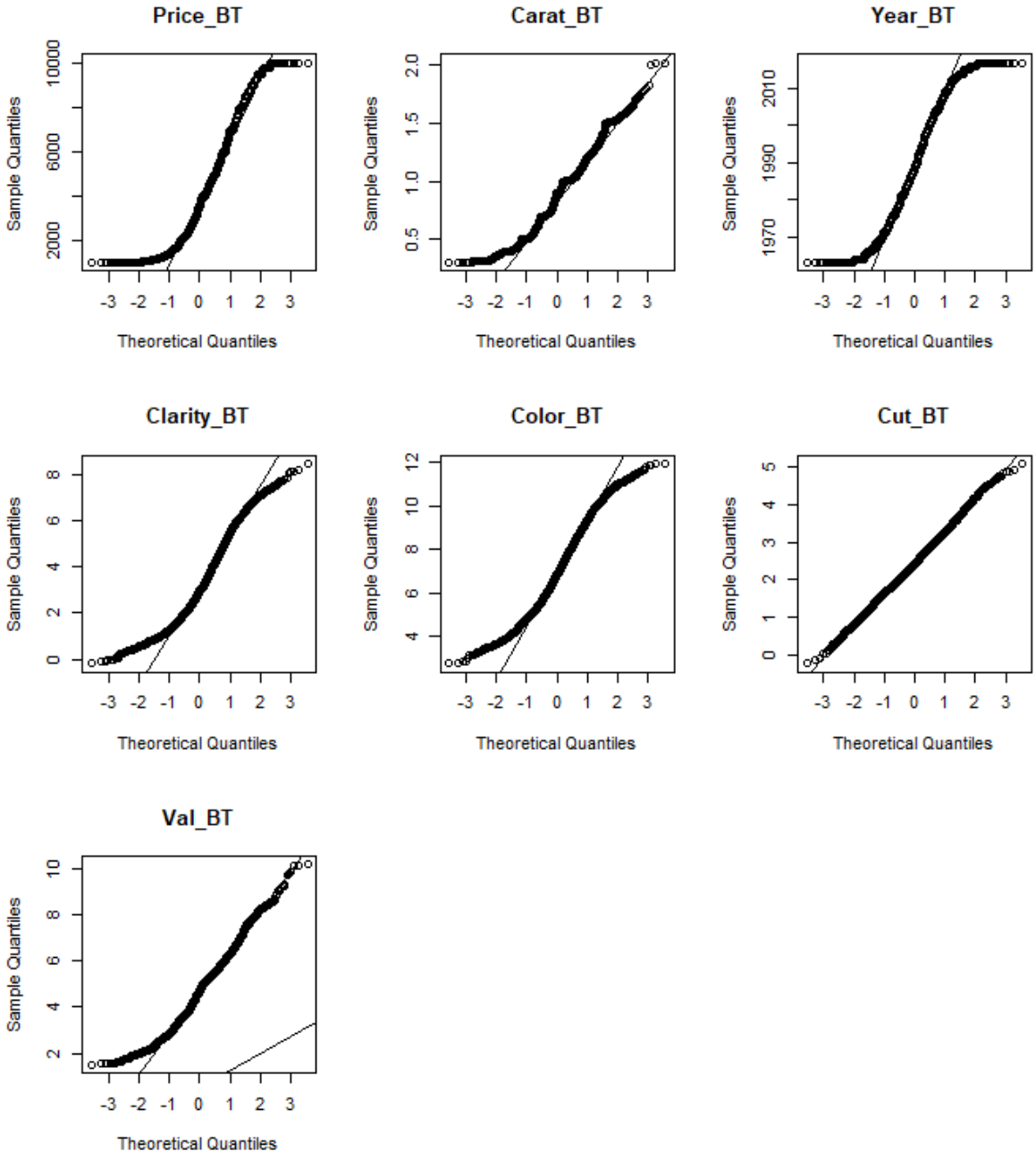


Outlier



As we noted earlier, there are some outliers in Carat, Cut and Val. Since, the outliers are reasonably close to the maximum and minimum value we are not removing them from analysis as of now.

Exploratory Data Analysis



	statistic	p.value
Price_BT	0.9213888	2.416619e-35
Carat_BT	0.9720116	1.572993e-22
Year_BT	0.9553	5.80826e-28
Clarity_BT	0.9560629	9.437154e-28
Color_BT	0.9742246	1.224655e-21
Cut_BT	0.9985379	0.01757149
Val_BT	0.9810517	1.767513e-18
Alrosa_BT	0.5335691	1.392702e-64
DeBeers_BT	0.07982884	3.954161e-78
Debswana_BT	0.1739726	6.874573e-76
Petra_BT	0.1778008	8.568645e-76
RioTinto_BT	0.6006981	1.138869e-61
Rockwell_BT	0.1182952	3.071532e-77

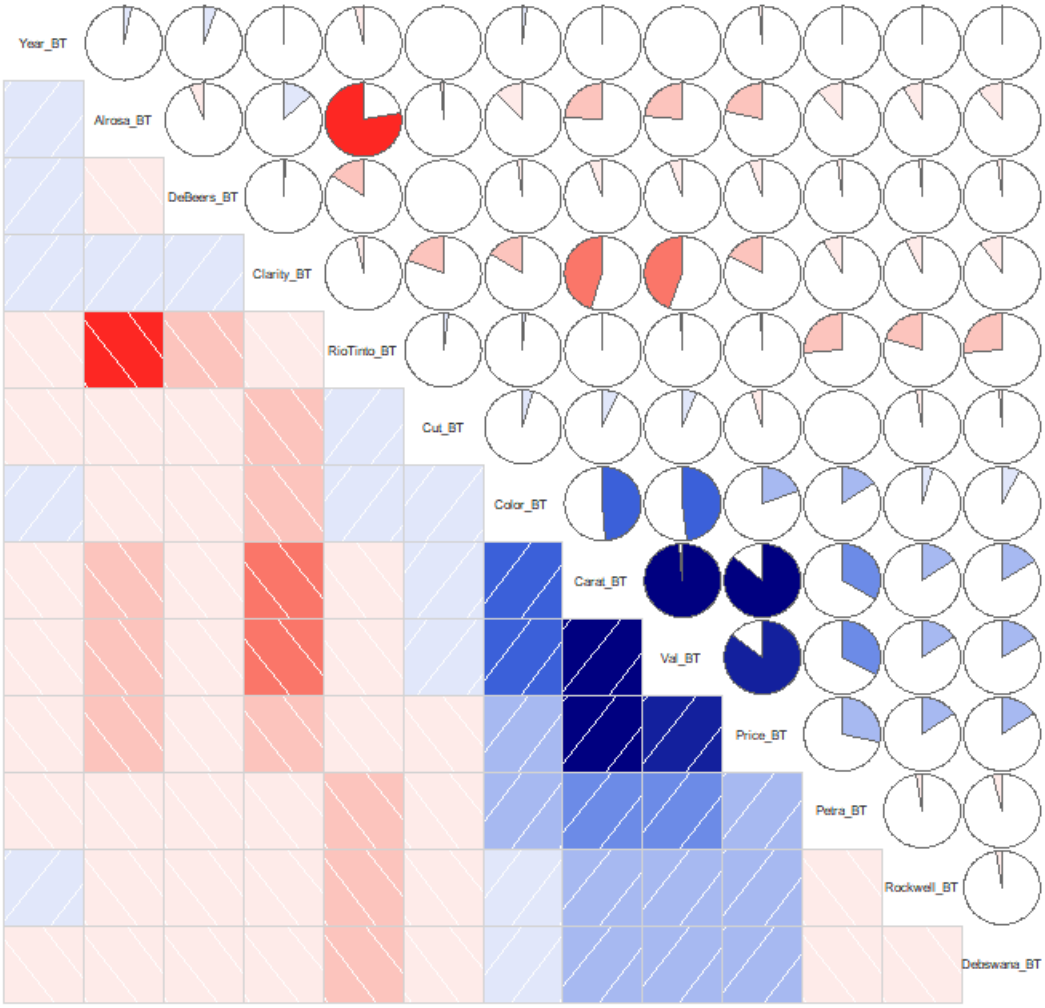
Based on the QQNorm plots and numeric test only Cut bears some resemblance to that of a normal distribution. All the other parameters are not normally distributed.

Correlations

	Price_BT	Carat_BT	Year_BT	Clarity_BT	Color_BT	Cut_BT	Val_BT
Price_BT	1.00	0.90	-0.01	-0.21	0.24	-0.02	0.89
Carat_BT	0.90	1.00	0.00	-0.45	0.49	0.05	0.98
Year_BT	-0.01	0.00	1.00	0.00	0.01	0.00	0.00
Clarity_BT	-0.21	-0.45	0.00	1.00	-0.16	-0.18	-0.44
Color_BT	0.24	0.49	0.01	-0.16	1.00	0.03	0.48
Cut_BT	-0.02	0.05	0.00	-0.18	0.03	1.00	0.06
Val_BT	0.89	0.98	0.00	-0.44	0.48	0.06	1.00
Alrosa_BT	-0.24	-0.25	0.03	0.13	-0.12	-0.01	-0.24
DeBeers_BT	-0.05	-0.06	0.06	0.01	-0.02	0.00	-0.06
Debswana_BT	0.17	0.17	-0.01	-0.10	0.08	-0.02	0.17
Petra_BT	0.25	0.29	-0.01	-0.08	0.15	0.00	0.29
RioTinto_BT	0.02	0.01	-0.04	-0.03	0.01	0.03	0.01
Rockwell_BT	0.14	0.17	0.00	-0.07	0.05	-0.02	0.16

	Alrosa_BT	DeBeers_BT	Debswana_BT	Petra_BT	RioTinto_BT	Rockwell_BT
Price_BT	-0.24	-0.05	0.17	0.25	0.02	0.14
Carat_BT	-0.25	-0.06	0.17	0.29	0.01	0.17
Year_BT	0.03	0.06	-0.01	-0.01	-0.04	0.00
Clarity_BT	0.13	0.01	-0.10	-0.08	-0.03	-0.07
Color_BT	-0.12	-0.02	0.08	0.15	0.01	0.05
Cut_BT	-0.01	0.00	-0.02	0.00	0.03	-0.02
Val_BT	-0.24	-0.06	0.17	0.29	0.01	0.16
Alrosa_BT	1.00	-0.06	-0.11	-0.11	-0.78	-0.08
DeBeers_BT	-0.06	1.00	-0.02	-0.02	-0.15	-0.02
Debswana_BT	-0.11	-0.02	1.00	-0.04	-0.26	-0.03
Petra_BT	-0.11	-0.02	-0.04	1.00	-0.26	-0.03
RioTinto_BT	-0.78	-0.15	-0.26	-0.26	1.00	-0.20
Rockwell_BT	-0.08	-0.02	-0.03	-0.03	-0.20	1.00

Diamond Price Stats



Price seems to be strongly positively correlated with Carat, Val and Petra and moderately positively correlated with Color, Rockwell, and Debswana. Whereas Clarity and Alrosa appear to be negatively correlated with Price. It is obvious that all else being equal, higher the carat value i.e., bigger the diamond higher the price. Also, it makes sense the colored diamonds (higher color index) are more expensive than the white diamonds(lower color index) and clear diamonds (lower clarity index) are more expensive than the diamond with impurities(higher clarity index).

I believe the variable Val, that is insurance value placed on the diamond, has high correlation with Price because of collinearity. Since only fewer diamonds are sourced from Alrosa, Rockwell and Debswana I don't expect that they will play a major role in predicting price of a random diamond.

The other correlations worth noticing are:

- RioTinto and Alrosa

- Carat and Clarity
- Carat and Color
- Val and Color
- Val and Clarity

Some of these (like RioTinto and Alrosa) may have confounding effects on the model.

Models

Model 1: All Variables included

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 88.49% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Five variables (but not the intercept) look significant (p-values of t-test <0.05). Variable Cut had very low correlation with Price, but in the model looks significant.
5. Variable Clarity is positively correlated with Price instead of negatively, whereas Color and Petra are negatively correlated with Price instead of positively.

```
Call:
lm(formula = Price_BT ~ Carat_BT + Year_BT + Clarity_BT + Color_BT +
    Cut_BT + Val_BT + Alrosa_BT + DeBeers_BT + Debswana_BT +
    Petra_BT + RioTinto_BT, data = Diamond_BT, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2825.4  -495.1   -79.7   352.2  3737.4

Coefficients:
             Estimate Std. Error t value          Pr(>|t|)
(Intercept) -220.6261  1990.7692  -0.111             0.912
Carat_BT    8290.7879   281.2594  29.477           < 2e-16 ***
Year_BT       -0.7585     0.9982  -0.760             0.447
Clarity_BT   380.8679     9.7541  39.047           < 2e-16 ***
Color_BT    -377.3389     8.8770 -42.508           < 2e-16 ***
Cut_BT      -141.2622    19.3370  -7.305 0.000000000000363 ***
Val_BT        81.0522    54.3516   1.491             0.136
Alrosa_BT   -176.8127   120.4974  -1.467             0.142
DeBeers_BT  -151.5498   188.1658  -0.805             0.421
Debswana_BT  103.4761   141.7192   0.730             0.465
Petra_BT    -559.4419   141.4947  -3.954 0.000078911501479 ***
RioTinto_BT -118.2734   116.3087  -1.017             0.309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 821.1 on 2678 degrees of freedom
```

```
Multiple R-squared:  0.8854, Adjusted R-squared:  0.8849
F-statistic:  1881 on 11 and 2678 DF,  p-value: < 2.2e-16
```

Model 2: Forward Selection

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 88.5% 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 <0.05). Variable Cut had very low correlation with Price, but in the model looks significant.
5. Variable Clarity is positively correlated with Price instead of negatively, whereas Color and Petra are negatively correlated with Price instead of positively.

```
Call:
lm(formula = Price_BT ~ Carat_BT + Color_BT + Clarity_BT + Cut_BT +
    Petra_BT + Debswana_BT + Alrosa_BT + Val_BT, data = Diamond_BT,
    na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2849.1  -495.5   -79.2   349.4  3751.1

Coefficients:
             Estimate Std. Error t value          Pr(>|t|)
(Intercept) -1850.506     98.233 -18.838           < 2e-16 ***
Carat_BT     8298.319    281.124  29.518           < 2e-16 ***
Color_BT     -377.794      8.868 -42.603           < 2e-16 ***
Clarity_BT    380.898      9.751  39.064           < 2e-16 ***
Cut_BT       -141.950     19.320  -7.347 0.000000000000267 ***
Petra_BT     -449.570     91.903  -4.892 0.000001058236470 ***
Debswana_BT   216.273     89.008   2.430            0.0152 *
Alrosa_BT     -60.880     38.126  -1.597            0.1104
Val_BT         81.877     54.324   1.507            0.1319
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 820.9 on 2681 degrees of freedom
Multiple R-squared:  0.8853, Adjusted R-squared:  0.885
F-statistic:  2587 on 8 and 2681 DF,  p-value: < 2.2e-16
```

Model 3: Stepwise Selection

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 88.49% of variation is explained by the model.
3. The residuals look approximately symmetrical.
4. Seven variables (and intercept) look significant (p-values of t-test <0.05). Variables Cut and RioTinto had very low correlation with Price, but in the model look significant.

5. Variable Clarity is positively correlated with Price instead of negatively, whereas Color and Petra are negatively correlated with Price instead of positively.

```
Call:
lm(formula = Price_BT ~ Carat_BT + Clarity_BT + Color_BT + Cut_BT +
    Val_BT + Alrosa_BT + Petra_BT + RioTinto_BT, data = Diamond_BT,
    na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2846.7  -494.8   -79.0   346.8  3754.0

Coefficients:
             Estimate Std. Error t value          Pr(>|t|)
(Intercept) -1712.847    118.296 -14.479           < 2e-16 ***
Carat_BT     8302.830    281.164  29.530           < 2e-16 ***
Clarity_BT    380.854      9.753  39.050           < 2e-16 ***
Color_BT     -377.520      8.871 -42.555           < 2e-16 ***
Cut_BT       -141.641     19.332  -7.327 0.000000000000311 ***
Val_BT         80.849     54.340   1.488           0.13692
Alrosa_BT    -200.107     72.527  -2.759           0.00584 **
Petra_BT     -588.982    106.519  -5.529 0.000000035245099 ***
RioTinto_BT  -141.932     66.148  -2.146           0.03199 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 821.1 on 2681 degrees of freedom
Multiple R-squared:  0.8853, Adjusted R-squared:  0.8849
F-statistic:  2585 on 8 and 2681 DF,  p-value: < 2.2e-16
```

Model Evaluation

Model 1: All Variables included

Verifying Assumptions

1. **Independence of Predictors**
   The Spearman rho value for Carat, Clarity, Color, and Petra are relatively high (-0.45, -0.49, .29) suggesting that the predictors are not independent.

2. **Distribution of Error Terms**
   The error terms do not seem to be normally distributed.

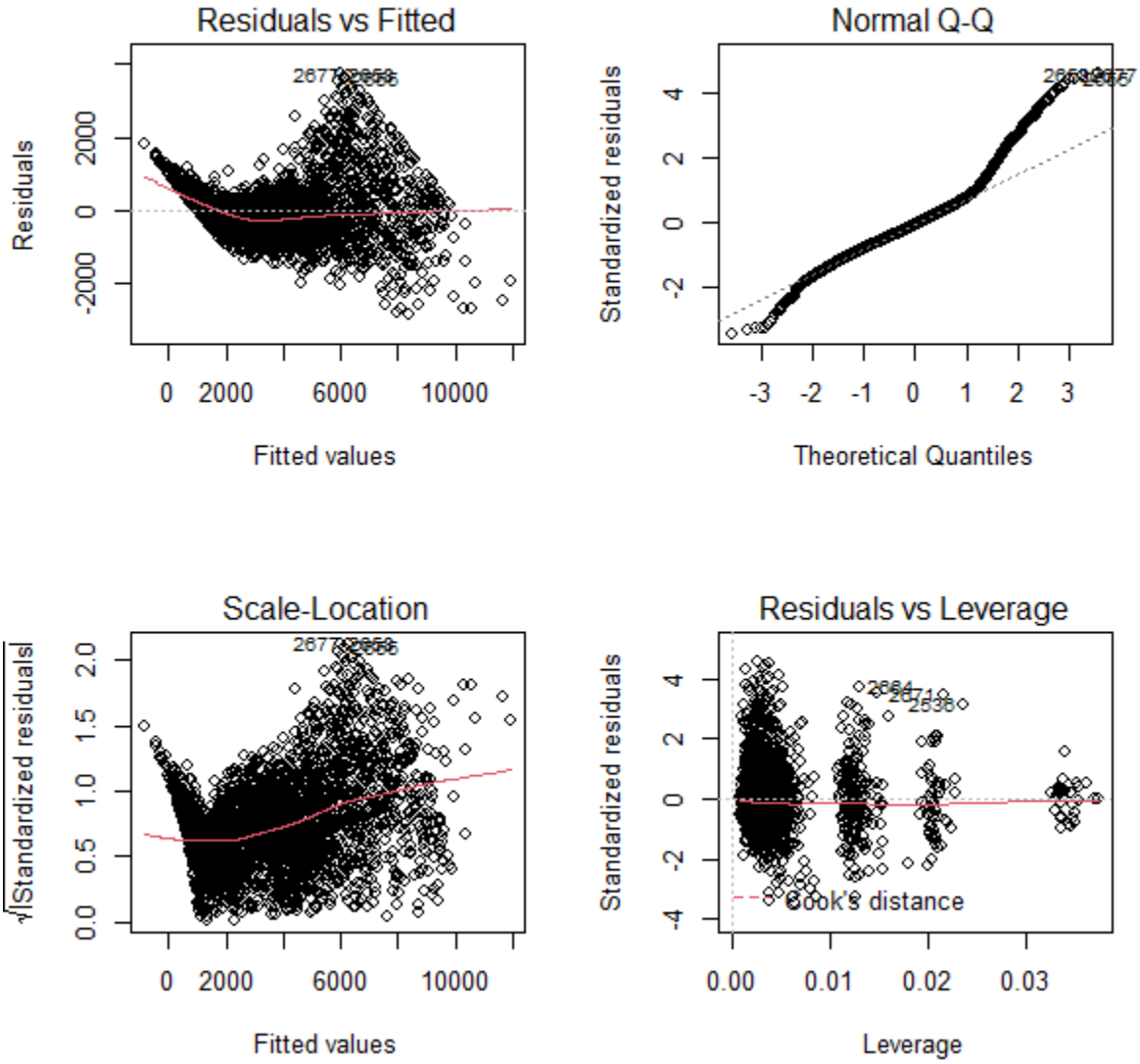
```
Shapiro-Wilk normality test

data:  DiaRes_BT
W = 0.94203, p-value < 2.2e-16
```

3. **Non-AutoCorrelation and Homoscedasticity**

   Based on Residuals vs. Fitted and Scale-Location, there appears to be no explicit pattern to the residuals. Therefore, no there is no appearance of autocorrelation.

   Based on Residuals vs. Leverage and Cook's Distance, there is no data point exerting undue influence or leverage on the model.



Model 2: Forward Selection

Verifying Assumptions

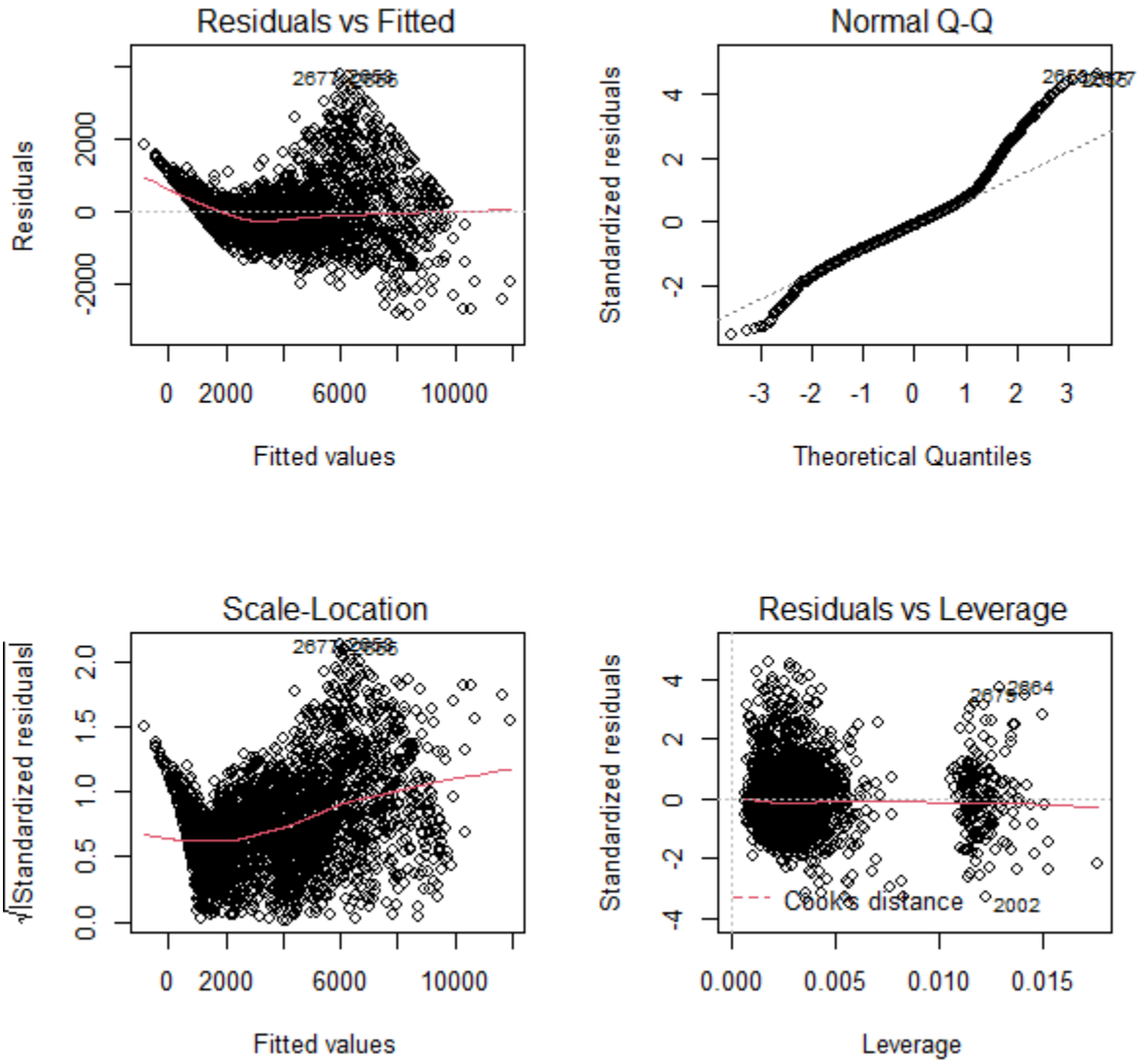
1. **Independence of Predictors**
   The Spearman rho value for Carat, Clarity, Color, and Petra are relatively high (-0.45, -0.49, .29) suggesting that the predictors are not independent. The Spearman rho value for Carat and Debswana is nominal (.17) so they are relatively independent.

2. **Distribution of Error Terms**
   The error terms do not seem to be normally distributed.

```
Shapiro-Wilk normality test

data:  FwdDiaRes_BT
W = 0.94222, p-value < 2.2e-16
```

3. **Non-AutoCorrelation and Homoscedasticity**
   Based on Residuals vs. Fitted and Scale-Location, there appears to be no explicit pattern to the residuals. Therefore, no there is no appearance of autocorrelation.
   Based on Residuals vs. Leverage and Cook's Distance, there is no data point exerting undue influence or leverage on the model.



Model 3: Stepwise Selection

Verifying Assumptions

1. **Independence of Predictors**
   The Spearman rho value for Carat, Clarity, Color, and Petra are relatively high (-0.45, -0.49, .29) suggesting that the predictors are not independent. The Spearman rho value for Carat, Debswana, and RioTinto are nominal (.17, .01) so they are relatively independent.

2. **Distribution of Error Terms**
   The error terms do not seem to be normally distributed.

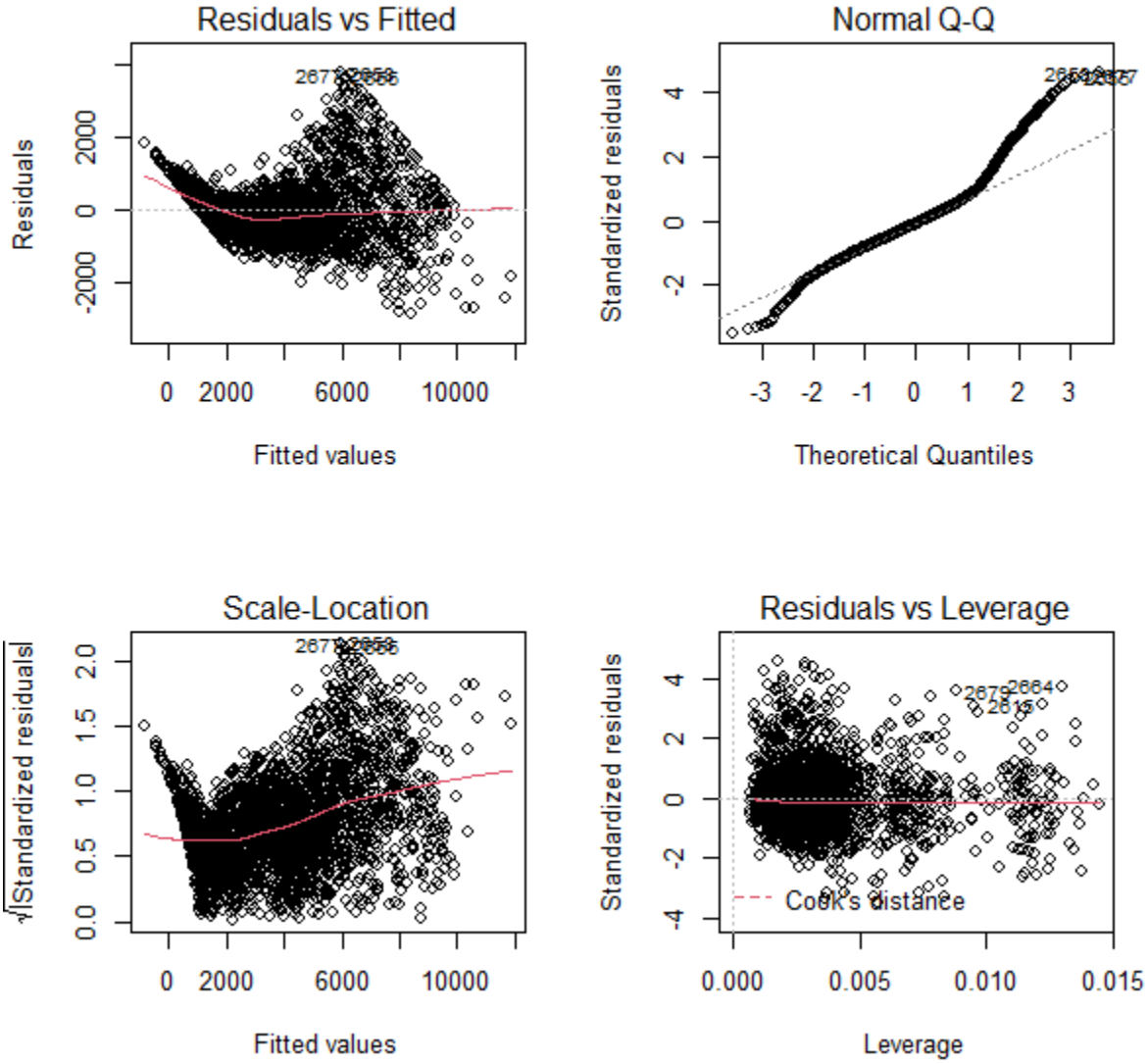
```
Shapiro-Wilk normality test
```

```
data:  StpDiaRes_BT
W = 0.94209, p-value < 2.2e-16
```

3. **Non-AutoCorrelation and Homoscedasticity**

   Based on Residuals vs. Fitted and Scale-Location, there appears to be no explicit pattern to the residuals. Therefore, no there is no appearance of autocorrelation.

   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

All the models developed above have reasonably similar statistics, but I recommend the model developed with forward selection since it has a slightly better F-statistic value :

```
Price_BT = -1850.506
 (8298.319) * Carat_BT +
 (-377.794) * Color_BT +
 (380.898) * Clarity_BT +
 (-141.950) * Cut_BT +
 (-449.570) * Petra_BT +
 (216.273) * Debswana_BT +
 (-60.880) * Alrosa_BT +
 ( 81.877) * Val_BT
```