Multivariate Linear Regression

Background

The goal of this exercise is to determine which factors help to best predict the price of a diamond using various multi-variate regression techniques.

Data Source

The dataset is a flat text file, obtained from Marsh, 2020. It is a compilation of various diamond characteristics, manufacturing sources, and prices. These can be seen in more detail in the provided data dictionary in Appendix 1. There are negative values seen in the dataset for the Cut and Clar variables. It is assumed these are part of the standard measures used during classification and are not mis-entered input.

Data Transformation and Cleaning

Price Variable

The Price variable was converted to the numeric data type, which is a requirement for participation in regression analysis. This is very important since Price is the dependent variable in this analysis.

Year Variable

The Year variable was also converted to the numeric data type in preparation for regression analysis.

Source Variable

The data identifying the source of the diamonds was categorical. It was transformed into six dummy variables to participate in the regression analysis. Each dummy variable represents one of the categories originally identified in the Source variable, coded with ones and zeros to represent their existence or non-existence respectively for each record. Only one dummy variable can exist with a value of 1 for any given record. The original Source column was then removed since it was no longer needed for regression analysis. However, its original description can still be seen in Appendix 1.

Other Notes

All variables were abbreviated to a max of three characters for ease in summarization during the various tasks below. They were also appended with the '_TG' suffix, since they were all altered in some way, some more minor than others. The updates can be seen in seen in Appendix 2, to complement the Data Dictionary. All transformation source code can be seen in Appendix 3.

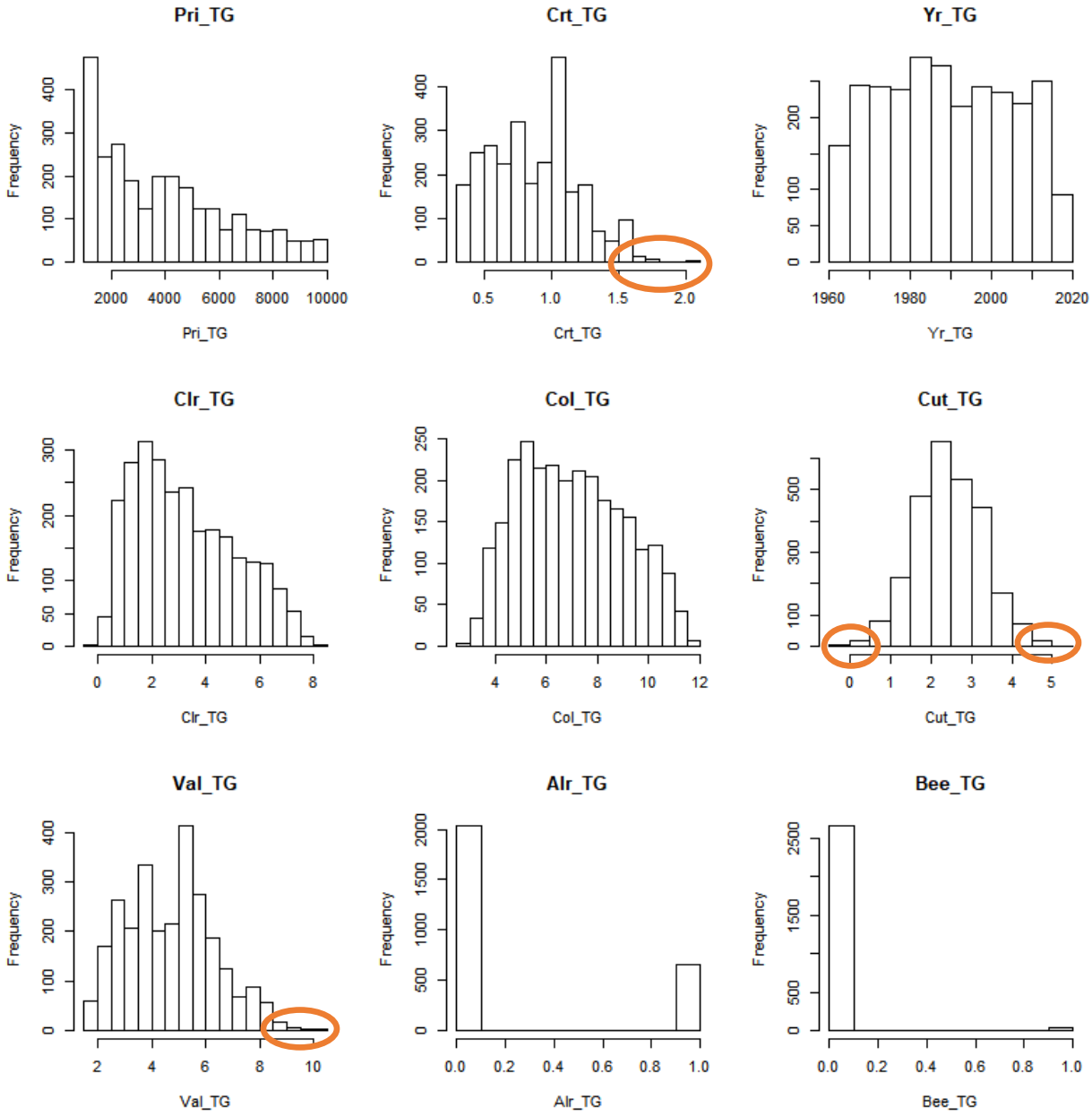
Descriptive Data Analysis

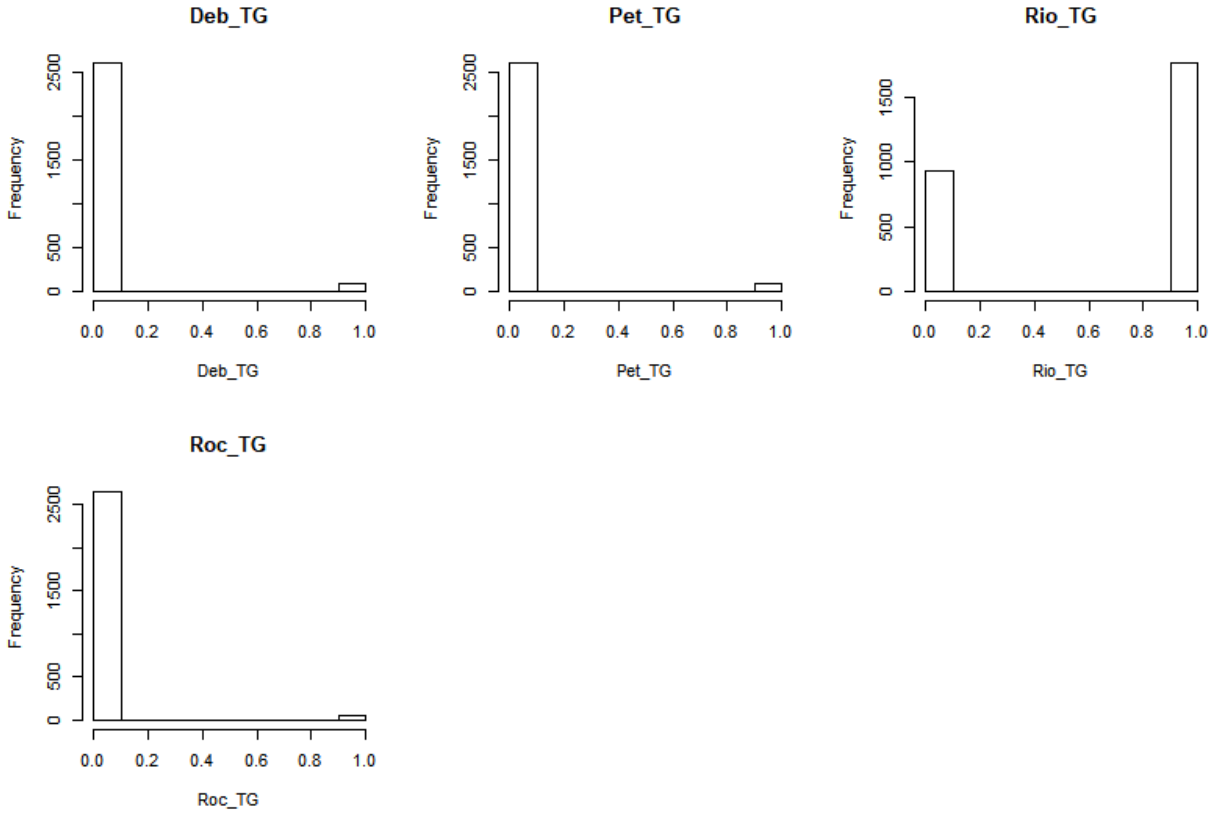
A summary table was then generated from the dataset to identify any unusual or interesting behaviour. The source code can be seen in Appendix 4.

```
     Pri_TG          Crt_TG          Yr_TG          Clr_TG          Col_TG
 Min.   : 1000   Min.   :0.30   Min.   :1963   Min.   :-0.19   Min.   : 2.77
 1st Qu.: 1801   1st Qu.:0.60   1st Qu.:1976   1st Qu.: 1.71   1st Qu.: 5.29
 Median : 3604   Median :0.90   Median :1989   Median : 2.94   Median : 6.88
 Mean   : 3972   Mean   :0.87   Mean   :1990   Mean   : 3.24   Mean   : 7.00
 3rd Qu.: 5544   3rd Qu.:1.06   3rd Qu.:2003   3rd Qu.: 4.62   3rd Qu.: 8.60
 Max.   :10000   Max.   :2.02   Max.   :2017   Max.   : 8.41   Max.   :11.77
     Cut_TG          Val_TG          Alr_TG          Bee_TG          Deb_TG
 Min.   :-0.05   Min.   : 1.51   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 1.91   1st Qu.: 3.39   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 2.42   Median : 4.67   Median :0.000   Median :0.0000   Median :0.0000
 Mean   : 2.45   Mean   : 4.68   Mean   :0.245   Mean   :0.0115   Mean   :0.0342
 3rd Qu.: 3.03   3rd Qu.: 5.77   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   : 5.03   Max.   :10.17   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
     Pet_TG          Rio_TG          Roc_TG
 Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
 Median :0.0000   Median :1.000   Median :0.0000
 Mean   :0.0353   Mean   :0.655   Mean   :0.0193
 3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.000   Max.   :1.0000
```

The Max values for Crt_TG, Cut_TG, and Val_TG are all greater than 1.5 of their respective interquartile ranges (IQR), suggesting the presence of outliers. The Min value for Cut_TG is lower than 1.5 of its IQR, suggesting the presence of outliers on the lower end as well.

A series of histograms were also generated to identify any patterns. Although the dummy variables just show values of 0 and 1, they were also included. The source code to generate the histograms can be seen in Appendix 4.

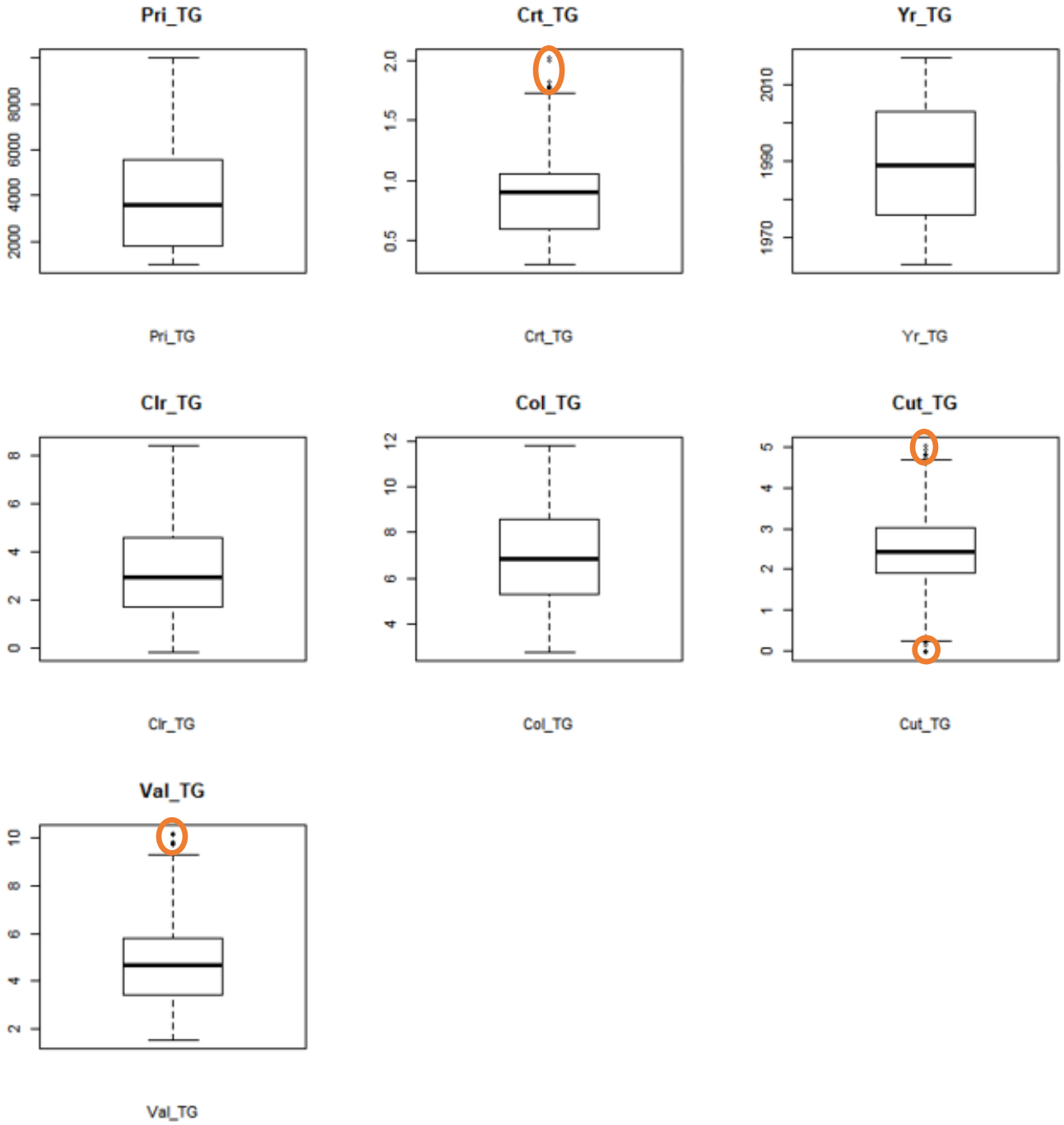




With the exceptions of the dummy variables, Yr_TG and Cut_TG, the histograms mostly show a range of positively skewed distributions. Pri_TG shows a trend of decreasing frequency with increasing price, with a very high frequency for the cheapest price. Crt_TG and Val_TG seem to have some extremely low values on the right tail that may influence the results. Although Cut_TG seems to display a more normal distribution, the extreme low values of both tails may also influence the results. Alr_TG and Rio_TG show a higher frequency of 1 than the other dummy variables. Yr_TG seems to display more of a regular distribution, with an interesting drop in the 2015-2020 bin.

Outliers

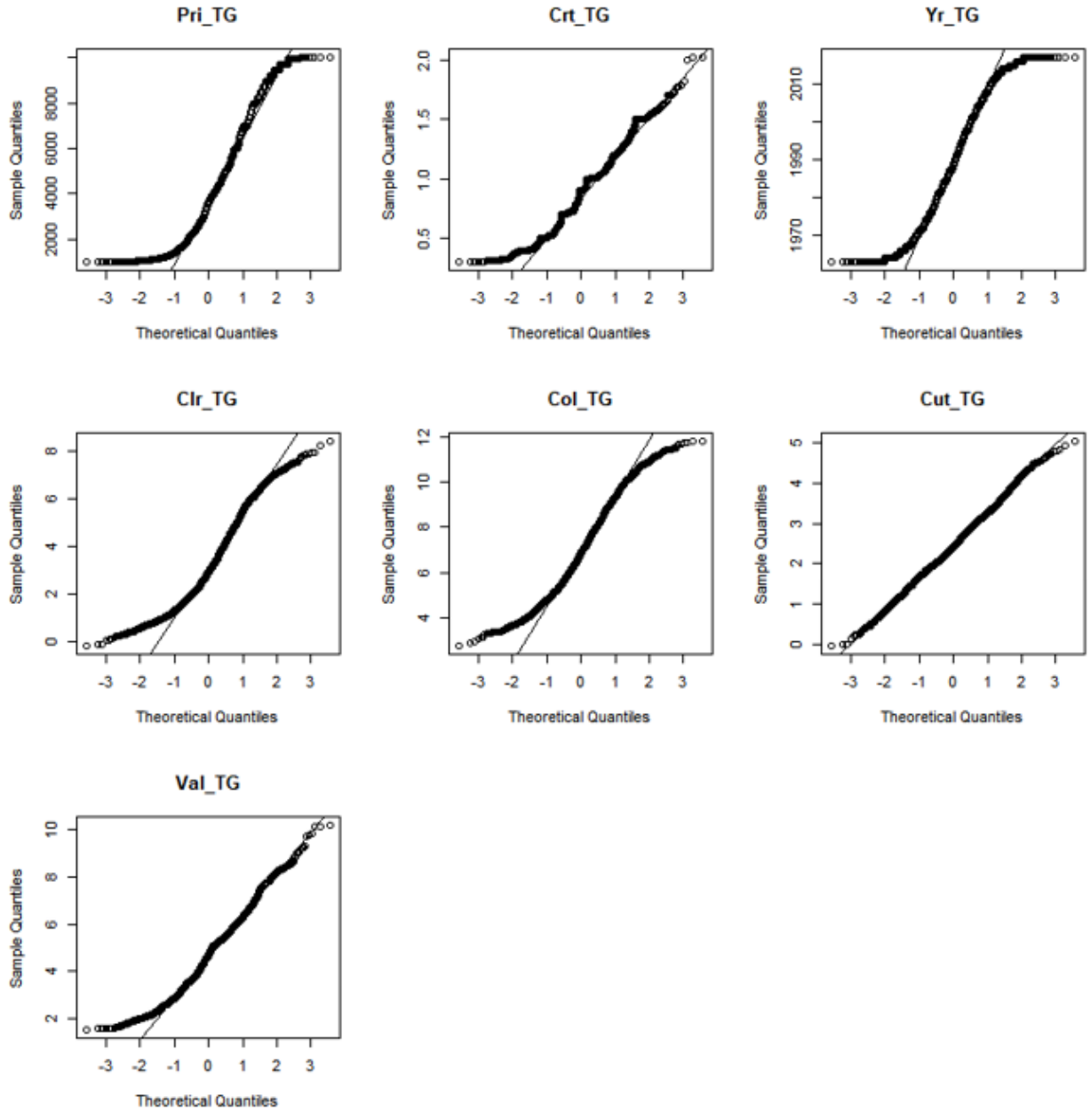
A series of boxplots were generated to identify the presence of outliers. The display excludes the dummy variables, due to their binary data of 1s and 0s. The source code to generate all boxplots can be seen in Appendix 5.



The boxplots reflect the distributions noticed earlier and confirms the presence of outliers in Crt_TG, Cut_TG and Val_TG.

Exploratory Data Analysis

QQ Norm Plots were generated to test for normality. The displays for the dummy variables were excluded due to the binary nature of their data, having a non-normal fit by default. The source code to generate all QQ Norm Plots can be seen in Appendix 6.



The only variable that visually comes close to a normal distribution fit is the Cut_TG variable, with relatively smaller deviations around the 0 and 5 edge sample quartiles.

Shapiro Wilks tests were performed on the variables to test for normality. All source code can be seen in Appendix 6.

	statistic	p.value
Pri_TG	0.92139	2.4166e-35
Crt_TG	0.97201	1.573e-22
Yr_TG	0.9553	5.8083e-28
Clr_TG	0.95469	3.9565e-28
Col_TG	0.97328	5.0267e-22
Cut_TG	0.99819	0.0039476
Val_TG	0.98105	1.7675e-18
Alr_TG	0.53422	1.4803e-64
Bee_TG	0.079829	3.9542e-78
Deb_TG	0.17397	6.8746e-76
Pet_TG	0.1778	8.5686e-76
Rio_TG	0.60088	1.1606e-61
Roc_TG	0.11669	2.815e-77

All the variables were confirmed to have significant deviations from a normal distribution fit, where all p-values are less than 0.05. The Cut_TG variable, although still non-normal, had the closest p-value of approximately 0.00394 to the 0.05 mark. This was visually reflected in the corresponding QQ Norm Plot above.

A Spearman correlation matrix was then run, since all the data are non-normal. This was done to identify any linear relationships between the predictor variables and Pri_TG, noted in the top half of the matrix, and any linear relationships amongst the predictor variables themselves, noted in the bottom half of the matrix. All source code can be seen in Appendix 6.

	Pri_TG	Crt_TG	Yr_TG	Clr_TG	Col_TG	Cut_TG	Val_TG	Alr_TG	Bee_TG	Deb_TG	Pet_TG	Rio_TG	Roc_TG
Pri_TG	1	0.9	-0.01	-0.21	0.24	-0.02	0.89	-0.23	-0.05	0.17	0.25	0.02	0.14
Crt_TG	0.9	1	0	-0.45	0.49	0.05	0.98	-0.24	-0.06	0.17	0.29	0.01	0.17
Yr_TG	-0.01	0	1	0	0.01	-0.01	0	0.04	0.06	-0.01	-0.01	-0.04	0
Clr_TG	-0.21	-0.45	0	1	-0.16	-0.19	-0.44	0.13	0.01	-0.1	-0.08	-0.03	-0.07
Col_TG	0.24	0.49	0.01	-0.16	1	0.03	0.48	-0.12	-0.02	0.08	0.15	0.01	0.06
Cut_TG	-0.02	0.05	-0.01	-0.19	0.03	1	0.06	-0.01	0	-0.02	0	0.02	-0.02
Val_TG	0.89	0.98	0	-0.44	0.48	0.06	1	-0.23	-0.06	0.17	0.29	0	0.16
Alr_TG	-0.23	-0.24	0.04	0.13	-0.12	-0.01	-0.23	1	-0.06	-0.11	-0.11	-0.78	-0.08
Bee_TG	-0.05	-0.06	0.06	0.01	-0.02	0	-0.06	-0.06	1	-0.02	-0.02	-0.15	-0.02
Deb_TG	0.17	0.17	-0.01	-0.1	0.08	-0.02	0.17	-0.11	-0.02	1	-0.04	-0.26	-0.03
Pet_TG	0.25	0.29	-0.01	-0.08	0.15	0	0.29	-0.11	-0.02	-0.04	1	-0.26	-0.03
Rio_TG	0.02	0.01	-0.04	-0.03	0.01	0.02	0	-0.78	-0.15	-0.26	-0.26	1	-0.19
Roc_TG	0.14	0.17	0	-0.07	0.06	-0.02	0.16	-0.08	-0.02	-0.03	-0.03	-0.19	1

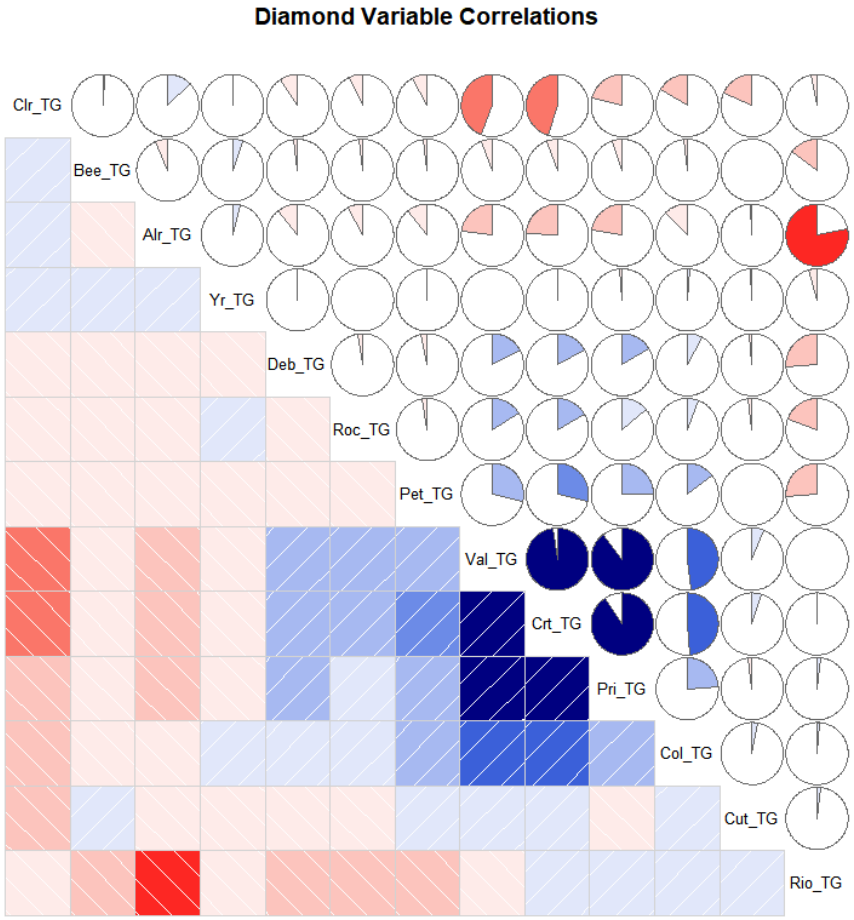
Several correlations were found between Pri_TG and other variables. There is a very strong positive linear correlation between Pri_TG and Crt_TG of 90% and between Pri_TG and Val_TG of 89%. There is a weak positive linear relationship between Pri_TG and Pet_TG of 25%. Col_TG has very near-

weak positive relationship to Pri_TG as well of 24%. Others worth mentioning are Clr_TG and Alr_TG, both having a near-weak negative relationship with Pri_TG of -21% and -23%. All other variables showed almost no linear correlation to Pri_TG.

Several predictor variables were found to have correlations amongst themselves. Val_TG has a very strong positive relationship with Crt_TG of 98%. Other variables also had relationships to Crt_TG, ranging from near-moderate (Clr_TG with -45% and Col_TG with 49%) to weak (Pet_TG with 29%) and near-weak (Alr_TG with -24%). Val_TG was also weakly correlated with Clr_TG and Col_TG of -44% and 48% respectively. Val_TG can be seen to have a near-weak relationship with Alr_TG of -23% and a weak relationship with Pet_TG of 29%.

Relationships between dummy variables were mostly ignored, given their inherent nature from the original Source variable. However, it was interesting to note the strong negative correlation of -78% between Rio_TG and Alr_TG. It is expected that these two variables are not both in the final model, and will be considered during model development, when dropping one of the dummy variables.

A graphical representation of the correlation matrix to show the highlights. All source code can be seen in Appendix 6.



Diamond Prices (Pri_TG) seem to be more strongly positively correlated with carat size (Crt_TG) and insurance value (Val_TG). As a result, there is an expectation to have at least one of these predictor variables in the final model. There is also a very strong, near-perfect, positive correlation between carat size and insurance value, suggesting collinearity. Hence, it is expected that both are not present in the final model, as they may have confounding effects.

Other correlations to notice are:

1. Insurance value (Val_TG) and Color (Col_TG)
2. Carat size (Crt_TG) and Color (Col_TG)
3. Insurance value (Val_TG) and Color (Col_TG)
4. Insurance value (Val_TG) and Clarity (Clr_TG)
5. Carat size (Crt_TG) and Clarity (Clr_TG)
6. Petra manufacturer (Pet_TG) and Insurance value (Val_TG)
7. Petra manufacturer (Pet_TG) and Carat size (Crt_TG)

Model Development

Multi-variate models were generated to predict diamond prices using various techniques below. The source code to generate the models can be seen in Appendix 7.

Model 1: All Variables included

1. The model is significant since the F-statistic p-value is less than 0.05.
2. 88.7% of the variation can be explained by the model, as seen with the Adjusted R-squared value.
3. The residuals are approximately symmetrical, with close differences between Q1 and the Median and Q3 and the Median of 406 and 416, respectively.
4. Six variables seem significant: Crt_TG, Clr_TG, Col_TG, Cut_TG, Pet_TG, with p-values of the t-test less than 0.001, and Deb_TG with p-value less than 0.05.
5. Clr_TG, Col_TG and Pet_TG are behaving inversely in the model when compared to the observations seen in the correlation matrix above.

```
Call:
lm(formula = Pri_TG ~ Crt_TG + Yr_TG + Clr_TG + Col_TG + Cut_TG +
    Val_TG + Alr_TG + Bee_TG + Deb_TG + Pet_TG + Roc_TG, data = diamond_data_TG,
    na.action = na.omit)

Residuals:
   Min     1Q Median     3Q    Max
 -2958   -483    -77    339   3752

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -347.837   1971.122   -0.18    0.860
Crt_TG      8326.886    279.081   29.84  < 2e-16 ***
Yr_TG         -0.746      0.990   -0.75    0.452
Clr_TG       384.826      9.716   39.61  < 2e-16 ***
Col_TG      -382.838      8.864  -43.19  < 2e-16 ***
Cut_TG      -149.070     19.547   -7.63  3.3e-14 ***
Val_TG        80.038     53.917    1.48    0.138
Alr_TG       -51.677     37.933   -1.36    0.173
Bee_TG       -25.420    148.101   -0.17    0.864
Deb_TG       207.252     88.524    2.34    0.019 *
Pet_TG      -444.263     91.595   -4.85  1.3e-06 ***
Roc_TG       115.661    116.357    0.99    0.320
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 814 on 2678 degrees of freedom
Multiple R-squared:  0.887,    Adjusted R-squared:  0.887
F-statistic: 1.92e+03 on 11 and 2678 DF,  p-value: <2e-16
```

Model 2: Forward Selection

1. The model is significant since the F-statistic p-value is less than 0.05.
2. 88.7% of the variation can be explained by the model, as seen with the Adjusted R-squared value.
3. The residuals are symmetrical, with differences between Q1 and the Median and Q3 and the Median 414.

4. Six variables and the intercept seem significant: Crt_TG, Col_TG, Clr_TG, Cut_TG, Pet_TG, and intercept, with p-values of the t-test less than 0.001, and Deb_TG with p-value less than 0.05.
5. Clr_TG, Col_TG and Pet_TG are behaving inversely in the model when compared to the observations seen in the correlation matrix above.

```
Call:
lm(formula = Pri_TG ~ Crt_TG + Col_TG + Clr_TG + Cut_TG + Pet_TG +
    Deb_TG + Val_TG + Alr_TG, data = diamond_data_TG, na.action = na.omit)

Residuals:
   Min     1Q Median     3Q    Max
 -2982   -487    -73    341   3732

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1833.68      98.07  -18.70  < 2e-16 ***
Crt_TG       8333.49     278.95   29.87  < 2e-16 ***
Col_TG       -383.22       8.86  -43.27  < 2e-16 ***
Clr_TG        384.86       9.71   39.62  < 2e-16 ***
Cut_TG       -149.73      19.53   -7.67  2.4e-14 ***
Pet_TG       -452.29      91.17   -4.96  7.5e-07 ***
Deb_TG        201.94      88.29    2.29    0.022 *
Val_TG         80.86      53.89    1.50    0.134
Alr_TG        -54.33      37.72   -1.44    0.150
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 814 on 2681 degrees of freedom
Multiple R-squared:  0.887,    Adjusted R-squared:  0.887
F-statistic: 2.63e+03 on 8 and 2681 DF,  p-value: <2e-16
```

Model 3: Stepwise Selection

1. The model is significant since the F-statistic p-value is less than 0.05.
2. 88.7% of the variation can be explained by the model, as seen with the Adjusted R-squared value.
3. The residuals are symmetrical, with differences between Q1 and the Median and Q3 and the Median of 414.
4. Six variables and the intercept seem significant: Crt_TG, Clr_TG, Col_TG, Cut_TG, Pet_TG, and intercept, with p-values of the t-test less than 0.001, and Deb_TG with p-value less than 0.05.
5. Clr_TG, Col_TG and Pet_TG are behaving inversely in the model when compared to the observations seen in the correlation matrix above.

```
Call:
lm(formula = Pri_TG ~ Crt_TG + Clr_TG + Col_TG + Cut_TG + Val_TG +
    Alr_TG + Deb_TG + Pet_TG, data = diamond_data_TG, na.action = na.omit)

Residuals:
   Min     1Q Median     3Q    Max
 -2982   -487    -73    341   3732

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1833.68      98.07  -18.70  < 2e-16 ***
Crt_TG       8333.49     278.95   29.87  < 2e-16 ***
Clr_TG        384.86       9.71   39.62  < 2e-16 ***
Col_TG       -383.22       8.86  -43.27  < 2e-16 ***
Cut_TG       -149.73      19.53   -7.67  2.4e-14 ***
Val_TG         80.86      53.89    1.50    0.134
Alr_TG        -54.33      37.72   -1.44    0.150
Deb_TG        201.94      88.29    2.29    0.022 *
Pet_TG       -452.29      91.17   -4.96  7.5e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 814 on 2681 degrees of freedom
Multiple R-squared:  0.887,    Adjusted R-squared:  0.887
F-statistic: 2.63e+03 on 8 and 2681 DF,  p-value: <2e-16
```

Model Evaluation

A brief study into the three models validated was conducted to verify all the assumptions associated with multi-variate regression analysis. The source code can be seen in Appendix 8.

Verifying Assumptions of Model 1 (All variables included)

1. **Independence of Predictors**
   The Spearman rho value seen in the Exploratory Analysis mostly show some weak to no correlations amongst the predictor values. However, the high correlation between the Crt_TG and Val_TG of 98% suggests there is some dependency amongst the predictor variables.

2. **Distribution of Error Terms**
   The error terms seem to deviate from a normal distribution significantly, as seen by the p-value below that is less than 0.05. They are not non-normal.

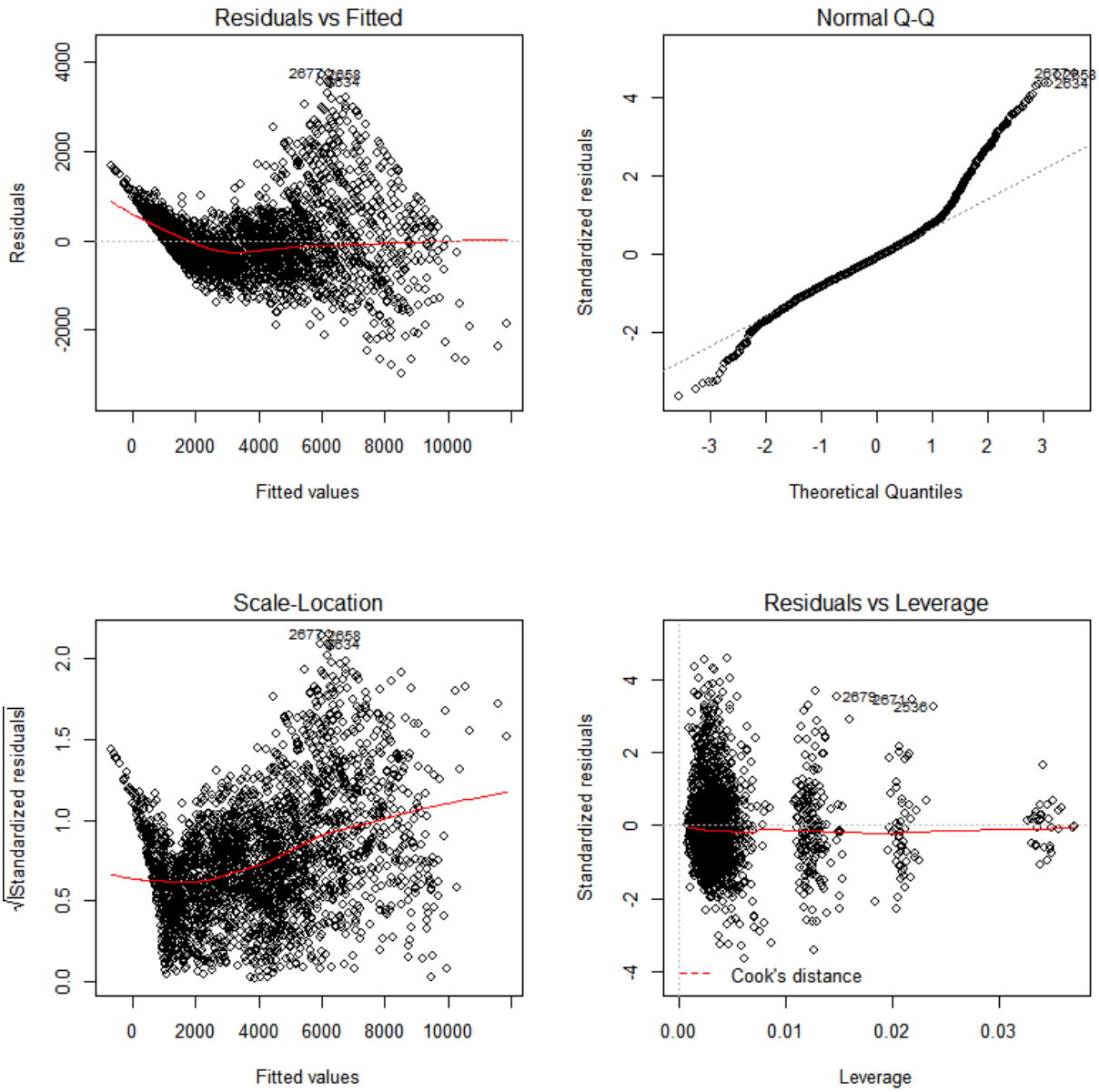
```
        Shapiro-Wilk normality test

data:  diamond_all_res_TG
W = 0.943, p-value <2e-16
```

3. **Non-Autocorrelation and Homoscedasticity**
   Based on Residuals vs. Fitted graph below, there seems to be a noticeable pattern of the points between 0 and 2000 Fitted values. The point cluster tends to follow the negative line before

becoming more dispersed after the 2000 Fitted Values, suggesting evidence of a non-linear relationship or auto-correlation.

The Scale-Location also shows a similar, yet more dispersed version of this behaviour. Therefore, there seems to be some evidence of non-homoscedasticity.

Based on Residuals vs. Leverage and Cook's Distance, there seems to be no data points influencing the model. The dashed line for Cook's Distance cannot even be seen.



Verifying Assumptions of Model 2 (Forward Selection)

1. **Independence of Predictors**
   Similarly, the Spearman rho value shows a high correlation between the Crt_TG and Val_TG of 98%, suggesting dependency exists amongst the predictor variables.

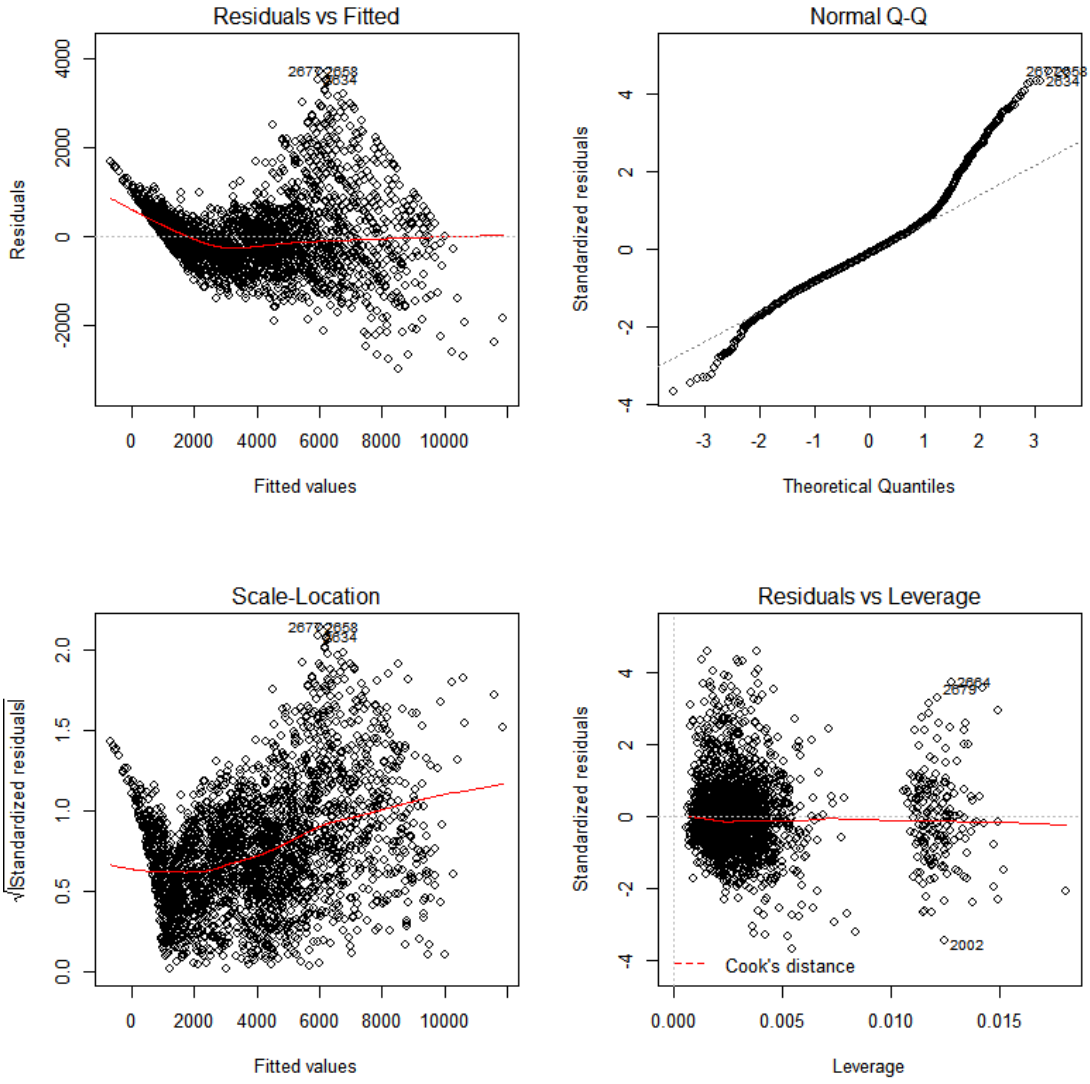
2. **Distribution of Error Terms**
   The error terms seem to deviate from a normal distribution significantly, as seen by the p-value below that is less than 0.05. They are not non-normal.

```
        Shapiro-Wilk normality test

data:  diamond_all_res_TG
W = 0.943, p-value <2e-16
```

3. **Non-Autocorrelation and Homoscedasticity**
   Based on Residuals vs. Fitted graph below, there seems to be a noticeable pattern of the points between 0 and 2000 Fitted values. The point cluster tends to follow the negative line before becoming more dispersed after the 2000 Fitted Values. There seems to be some evidence of auto correlation

   The Scale-Location also shows a similar, yet more dispersed version of this behaviour. Therefore, there seems to be some evidence of non-homoscedasticity. There is slightly more dispersion than Model 1.

   Based on Residuals vs. Leverage and Cook's Distance, there seems to be no data points influencing the model. The dashed line for Cook's Distance cannot even be seen. There are two distinct clustering of data points, as opposed to the four seen in Model 1.



Verifying Assumptions of Model 3 (Stepwise Selection)

1. **Independence of Predictors**
   Again, the Spearman rho value shows a high correlation between the Crt_TG and Val_TG of 98%, suggesting dependency exists amongst the predictor variables.

2. **Distribution of Error Terms**
   The error terms seem to deviate from a normal distribution significantly, as seen by the p-value below that is less than 0.05. They are not non-normal and are the same for the first two models.

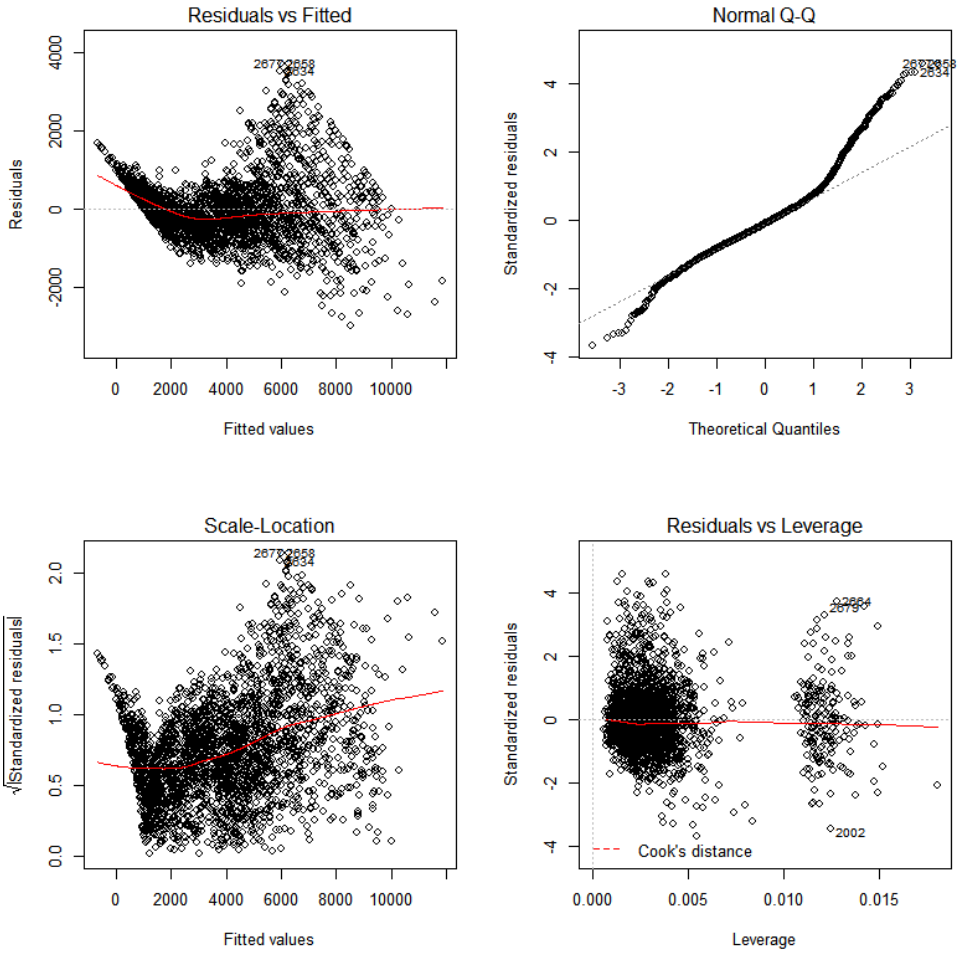
```
        Shapiro-Wilk normality test

data:  diamond_all_res_TG
W = 0.943, p-value <2e-16
```

3. **Non-Autocorrelation and Homoscedasticity**

Based on Residuals vs. Fitted graph below, there seems to be a noticeable pattern of the points between 0 and 2000 Fitted values. The point cluster tends to follow the negative line before becoming more dispersed after the 2000 Fitted Values. There is some evidence of autocorrelation.

The Scale-Location also shows a similar, yet more dispersed version of this behaviour. Therefore, there seems to be some evidence of non-homoscedasticity also within the 0 and 2000 Fitted Values.

Based on Residuals vs. Leverage and Cook's Distance, there seems to be no data points influencing the model. The dashed line for Cook's Distance cannot even be seen. There are two distinct clustering of data points, as opposed to the four seen in Model 1.



Final Recommendation

Based on the regression analysis above, none of the models meets the assumptions properly. All three models produced very similar results. To remove further effects of collinearity, the models should be run a few more times, using some judgmental analysis to manually remove variables, instead of only depending on the models doing that on their own. A further look into the variables, such as removing outliers, investigating collinearity more in depth or trying another sample would also be recommended. However, a recommendation would be for Model 3, using stepwise selection, since after adding or removing a variable, it checks the significance of all other variables first (Marsh, 2020 b).

Diamond Price =

(8333.49) * Carat size +

(384.86) * Clarity +

(-383.22) * Color +

(-149.73) * Cut +

(80.86) * Insurance value +

(-54.33) * Alrosa manufacturing +

(201.93) * Debswana manufacturing +

(-452.29) * Petra manufacturing +

(-18333.68)

References

Marsh, D. (2020). *Assignment 4 - Regression*. eConestoga. Retrieved July 10, 2020 from https://conestoga.desire2learn.com/d2l/lms/dropbox/user/folder_submit_files.d2l?db=349649&grpid=0&isprv=0&bp=0&ou=354666

Marsh, D. (2020) b. *PROG8430 – Data Analysis, Modeling and Algorithms*. eConestoga. Retrieved July 10, 2020 from https://conestoga.desire2learn.com/d2l/le/content/354666/viewContent/7408176/View

APPENDIX 1: Data Dictionary (Marsh, 2020)

Variable	Description
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 (Excellent, Good, etc.
Source	The diamond manufacturing who mined, graded and cut the diamond.
Val	Insurance Value placed on the diamond
Year	The year the diamond was first cut.

APPENDIX 2: Updated Variable Names

Dictionary Name	Dataset Name	New Name	Notes
Price	Price	Pri_TG	Converted to Numeric data type.
Carat	Carat.Size	Crt_TG	Inconsistent name
Clarity	Clar	Clr_TG	Shortened for ease
Color	Col	Col_TG	Involved in rename transformation (minor).
Cut	Cut	Cut_TG	Involved in rename transformation (minor).
Val	Val	Val_TG	Involved in rename transformation (minor).
Year	Year	Yr_TG	Converted to Numeric data type.
Source	Source	N/A	Removed from dataset.
		Alr_TG	Dummy coded from Source variable, representing Alrosa.
		Bee_TG	Dummy coded from Source variable, representing DeBeers.
		Deb_TG	Dummy coded from Source variable, representing Debswana.
		Pet_TG	Dummy coded from Source variable, representing Petra.
		Rio_TG	Dummy coded from Source variable, representing RioTinto.
		Roc_TG	Dummy coded from Source variable, representing Rockwell.

APPENDIX 3: Transformation and Clean-up Source Code

Transforming Price to Numeric

```
diamond_data_TG$Price <- as.numeric(diamond_data_TG$Price)
```

Transforming Year to Numeric

```
diamond_data_TG$Year <- as.numeric(diamond_data_TG$Year)
```

Transforming Source to Dummy Variables

```
# convert factor Source to index dummy variables
source_dummies_TG <- model.matrix(~ Source -1, data = diamond_data_TG)

# combine the datasets
diamond_data_TG <- cbind(diamond_data_TG, source_dummies_TG)
```

Clean-up by Dropping Source

```
diamond_data_TG <- diamond_data_TG[-c(3)]
```

Clean-up by Renaming Variables

```
names(diamond_data_TG) <- c("Pri_TG", "Crt_TG", "Yr_TG", "Clr_TG", "Col_TG", "Cut_TG",
                            "Val_TG", "Alr_TG", "Bee_TG", "Deb_TG", "Pet_TG", "Rio_TG",
                            "Roc_TG")
```

Result

```
> # confirm variable data types
> str(diamond_data_TG)
'data.frame':    2690 obs. of  13 variables:
 $ Pri_TG: num  1000 1000 1000 1000 1000 ...
 $ Crt_TG: num  0.3 0.44 0.31 0.66 0.47 0.4 0.36 0.52 0.53 0.43 ...
 $ Yr_TG : num  1979 2001 1982 2004 2015 ...
 $ Clr_TG: num  6.63 2.73 5.66 2.22 2.87 4.04 2.65 1.27 0.56 3.47 ...
 $ Col_TG: num  5.06 4.99 5.08 10.74 8.16 ...
 $ Cut_TG: num  3.13 2.04 2.39 2.09 3.04 2.03 2.07 2.97 3.31 2.22 ...
 $ Val_TG: num  1.88 2.21 2.08 3.86 2.35 ...
 $ Alr_TG: num  0 1 0 1 1 1 0 0 1 1 ...
 $ Bee_TG: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Deb_TG: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Pet_TG: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Rio_TG: num  1 0 1 0 0 0 1 1 0 0 ...
 $ Roc_TG: num  0 0 0 0 0 0 0 0 0 0 ...
```

APPENDIX 4: Descriptive Analysis Source Code

Generating Summary Statistics

```
summary(diamond_data_TG)
```

Generating Histograms

```
# generate 3 x 3 grid for graphs
par(mfrow = c(3,3))

# generate histograms for all numeric variables
# loop over column *names* instead of actual columns
sapply(names(diamond_data_TG), function(cname){
  # plot only the numeric columns
  if (is.numeric(diamond_data_TG[[cname]]))
    # set column name as plot title with 'main' param
    print(hist(diamond_data_TG[[cname]], main = cname, xlab = cname))
})
```

APPENDIX 5: Outliers Source Code

Generating Boxplots

```
# generate 3 x 3 grid for graphs
par(mfrow = c(3,3))

# generate box plots for all numeric variables
# loop over column *names* instead of actual columns
sapply(names(diamond_data_TG), function(cname){
  # plot only the numeric columns
  if (is.numeric(diamond_data_TG[[cname]]))
    # set column name as plot title with 'main' param
    print(boxplot(diamond_data_TG[[cname]], main = cname, xlab = cname))
})
```

APPENDIX 6: Exploratory Analysis Source Code

Generating Boxplots

```
# generate 3 x 3 grid for graphs
par(mfrow = c(3,3))

# generate QQ Norm plot for normality visual
# loop over column *names* instead of actual columns
sapply(names(diamond_data_TG), function(cname){
  # plot only the numeric columns
  if (is.numeric(diamond_data_TG[[cname]]))
    # set column name as plot title with 'main' param
    qqnorm(diamond_data_TG[[cname]], main = cname)
  qqline(diamond_data_TG[[cname]])
})
```

Running Shapiro Wilks Tests

```
# run Shapiro Wilks tests for normality
diamond_norm_TG <- lapply(diamond_data_TG, shapiro.test)

# group all the tests together
diamond_ngroup_TG <- sapply(diamond_norm_TG, `[`, c("statistic","p.value"))

# transpose the group for an easier read
diamond_ngroupt_TG <- t(diamond_ngroup_TG)
diamond_ngroupt_TG
```

Generating a Spearman Correlation Matrices

```
# run spearman correlation table since all data is not normally distributed
diamond_corr_TG <- cor(diamond_data_TG, method="spearman")
round(diamond_corr_TG, 2)

# generate visual correlation representation
corrgram(diamond_data_TG, order=TRUE, lower.panel=panel.shade,
         upper.panel=panel.pie, text.panel=panel.txt,
         main = "Diamond Variable Correlations", cor.method="spearman")
```

APPENDIX 7: Model Development Source Code

Generating Model with All Variables

```
diamond_all_lm_TG = lm(Pri_TG ~ Crt_TG + Yr_TG + Clr_TG + Col_TG +
                       Cut_TG + Val_TG + Alr_TG + Bee_TG + Deb_TG + Pet_TG +
                       Rio_TG, data=diamond_data_TG, na.action=na.omit)
diamond_all_lm_TG
summary(diamond_all_lm_TG)
```

Generating Model with Forward Selection

```
diamond_min_lm_TG <- lm(Pri_TG ~ 1, data=diamond_data_TG, na.action=na.omit)
diamond_fwd_lm_TG = step(diamond_min_lm_TG, direction="forward", scope =(
                        ~ Crt_TG + Yr_TG + Clr_TG + Col_TG +
                        Cut_TG + Val_TG + Alr_TG + Bee_TG + Deb_TG +
                        Pet_TG + Rio_TG), details=TRUE)

diamond_fwd_lm_TG
summary(diamond_fwd_lm_TG)
```

Generating Model with Stepwise Selection

```
diamond_all_lm_TG = lm(Pri_TG ~ Crt_TG + Yr_TG + Clr_TG + Col_TG +
                         Cut_TG + Val_TG + Alr_TG + Bee_TG + Deb_TG + Pet_TG +
                         Roc_TG, data=diamond_data_TG, na.action=na.omit)
diamond_step_lm_TG <- step(diamond_all_lm_TG)
diamond_step_lm_TG
summary(diamond_step_lm_TG)
```

APPENDIX 8: Model Evaluation Source Code

Verifying Assumptions for each model

```
# Create model and residual vectors
diamond_all_fit_TG <- predict(diamond_all_lm_TG)
diamond_all_res_TG <- residuals(diamond_all_lm_TG)

diamond_fwd_fit_TG <- predict(diamond_fwd_lm_TG)
diamond_fwd_res_TG <- residuals(diamond_fwd_lm_TG)

diamond_step_fit_TG <- predict(diamond_step_lm_TG)
diamond_step_res_TG <- residuals(diamond_step_lm_TG)

# Test normality of residuals - none has normal fit
shapiro.test(diamond_all_res_TG)
shapiro.test(diamond_fwd_res_TG)
shapiro.test(diamond_step_res_TG)

# Run diagnositcs and plot graphs on models
par(mfrow = c(2, 2))
plot(diamond_all_lm_TG)
par(mfrow = c(1, 1))

par(mfrow = c(2, 2))
plot(diamond_fwd_lm_TG)
par(mfrow = c(1, 1))

par(mfrow = c(2, 2))
plot(diamond_step_lm_TG)
par(mfrow = c(1, 1))
```