NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

MH3511 Data Analysis with Computer
Group Project

Relationship between “Dimensional Properties of a Diamond” with “Price”

Name	Matriculation Number
Yalamanchili Sanjana	U2323273H
Goh Jie Rong, Sean	U2322220J
Thatam Setty Venkata Nanda Sai	U2340694D
Sutheerth Gopalakrishnan	U2340025E
Srinivasan Shreyas	U2340729C

1. Introduction

Amongst the multitude of options that proliferates the luxury industry, the reliability and allure of Diamonds withstands the test of time. Seen as the epitome of luxurial status in a USD56.5 billion industry, diamonds remain highly sought after especially by jewelry and luxury watch brands. Diamonds also see use within the construction industry, where diamonds of different calibers and grades are leveraged for mining. As such, knowing the worth of diamonds and what variables affect its price plays an instrumental role for those involved with it's usage

In this project, a dataset containing details of 54,000 diamonds is used. Variables/Characteristics of the diamonds include its Price, Carat, Cut, Color, Clarity…
Based on this dataset, the following questions can be answered, hence gaining useful insights and practical use:

1. Does the Price of a Diamond depend on its Carat?
2. Does the Price of a Diamond depend on its Cut?
3. Which variable of a Diamond contributes most to its price fluctuations?
4. ……

With the use of R language, statistical analysis and inference were employed to gain insights into our research question.

2. Data description

The dataset, titled "diamonds", is obtained from the online data science community kaggle.com. The original data consists of a single data file named "diamonds.csv". The dataset was originally posted on the official website of the Nature Publishing Group, the official website responsible for publishing academic journals, magazines, online databases, and services in science and medicine, including the flagship journal Nature, and it is open to the public for study and research.

Before, proceeding to data analysis, appropriate data clearing needed to be done:

- ➔ Majority of the rows were already clean, and there were no null values present.
- ➔ Ultimately due to the large size of the dataset, we decided to choose only those data points, whose "table" values are in the range 64 to 95.
- ➔ ……..

After, the data preparation, now we had 590 diamonds, with 10 variables for analysis:

1. **price:** price in US dollars (\$326--\$18,823)

2. **carat:** weight of the diamond (0.2--5.01)
3. **cut:** quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4. **color:** diamond colour, from J (worst) to D (best)
5. **clarity:** a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
6. **x:** length in mm (0--10.74)
7. **y:** width in mm (0--58.9)
8. **z:** depth in mm (0--31.8)
9. **depth:** total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
10. **table:** width of top of diamond relative to widest point **(64--95)**

3. **Statistical Description**

In this section, we shall look into the data in more detail. Each variable is investigated individually to look for possible outliers, and/or to perform a transformation to avoid highly skewed data.

3.1 Summary statistics of the main variable of interest, "Price"

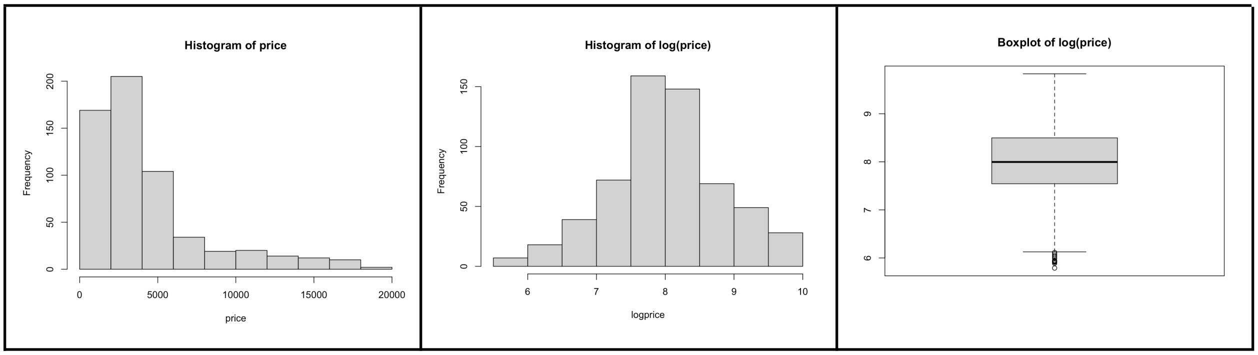
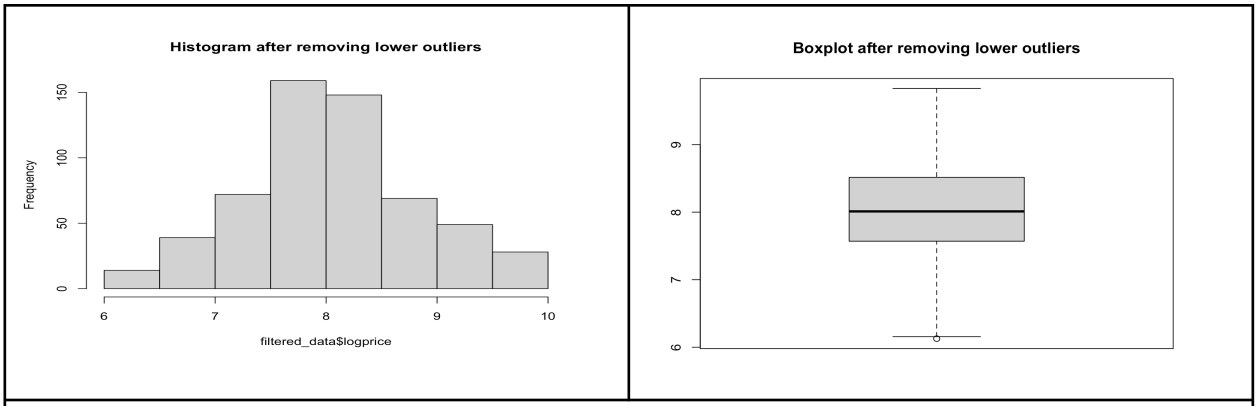
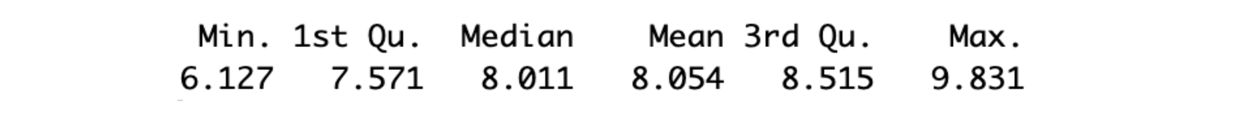


Fig 1. Summary Statistics of Price.

It appears that the variable: "Price" is highly skewed, hence we apply a log-transformation to the variable. The log-transformed data appears to have some outlying values at the left tail. Upon further investigation, we notice that some diamonds are of extremely poor quality due to which they have lower prices, compared to the other diamonds.Therefore, we remove those diamonds, whose price is below $389.11, which is approximately 3.9% of the data.

The histogram and the boxplot of the log-transformed variable, with the outliers removed are shown below with summary statistics/ the dataset is now more symmetric and has only one outlier.





```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.127   7.571   8.011   8.054   8.515   9.831 
```

Fig 2. Summary Statistics of "Price"(After Log-Transformation).

We shall proceed to the next section, with the trimmed dataset.

The "histogram", "boxplot", "transformation applied" and "outliers removed" from each variable are tabulated in the following subsections.

3.2.1 Carat of the diamonds, "Carat"

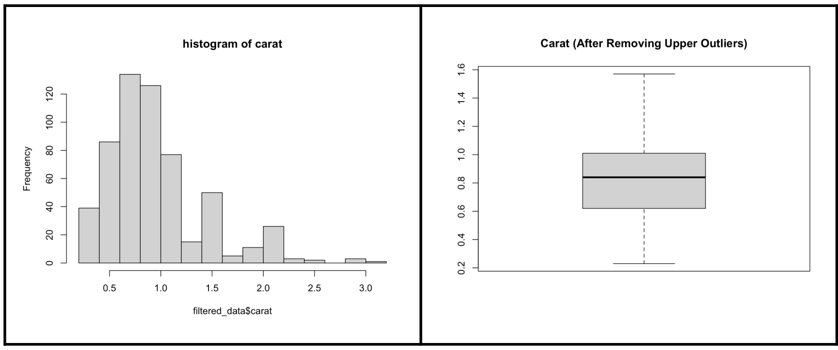
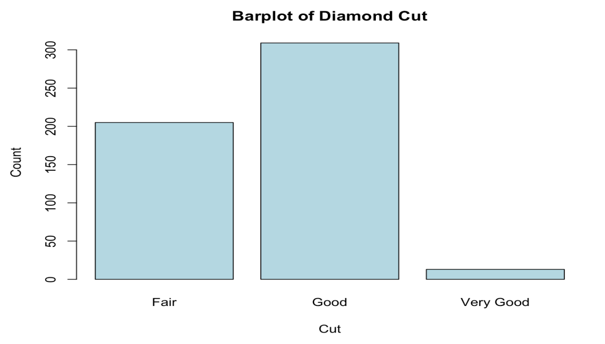
	-The outliers were removed with the help of the upper and lower bound. -No outliers remain after the transformation.

Fig 3. Statistical Visualization of "Carat" after cleaning.

3.2.2 Cut of the diamonds, "Cut"

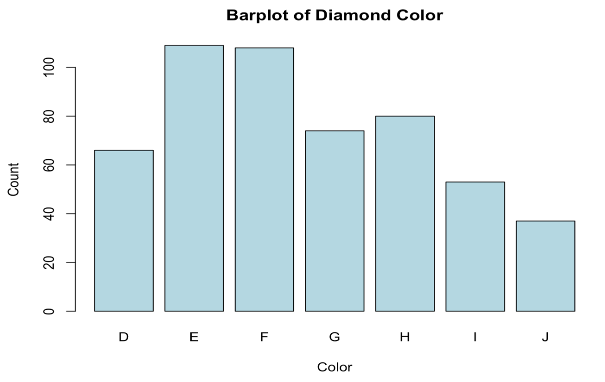


-There is a huge difference in the number of diamonds having a very good cut, and the other two categories, namely: “good cut" and “fair cut”

-No outliers were present.

Fig 4. Statistical Visualization of “Cut” after cleaning.

3.2.3 Color of the diamond, “Color”

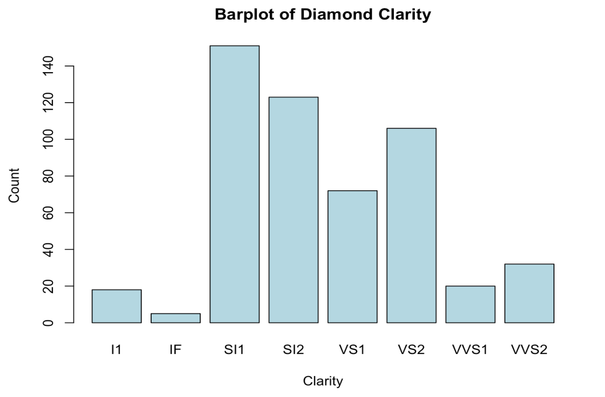


-It is very clearly seen that E and F color diamonds enjoy more popularity than the other colors.

-But, there are no huge differences between the colors of the diamonds.

Fig 5. Statistical Visualization of “Color” after cleaning

3.2.4 Clarity of the diamond, “Clarity”

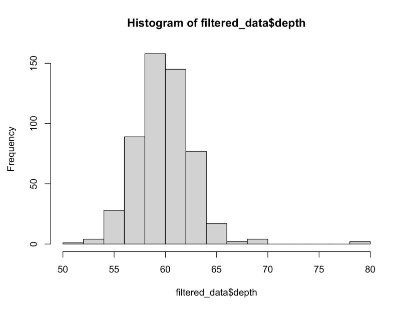


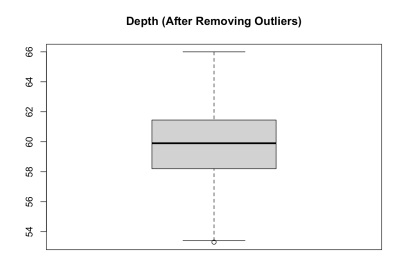
-There is a huge difference amongst the clarity of the diamonds present.

-There are currently no outlying data points present, so no values are removed.

Fig 6. Statistical Visualization of “Clarity” after cleaning.

3.2.5 Depth of the diamond, “Depth”

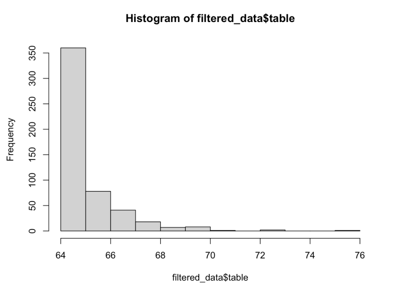


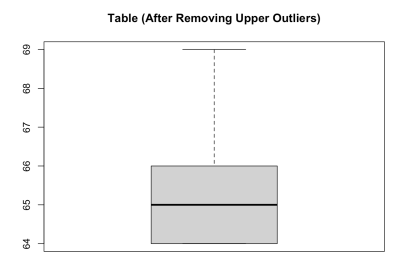


-There were a huge number of outliers in this case for the depth of the diamond.
-They were removed with the help of the IQR method.

Fig 7. Statistical Visualization of “Depth” after cleaning.

3.2.6 Table of the diamond, “Table” (Table → top flat surface of a Diamond)

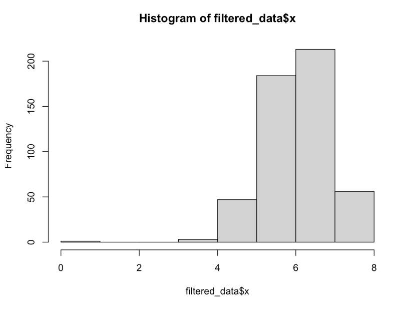


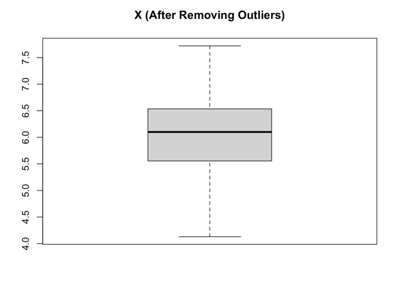


-The existing outliers were removed with the help of the IQR method.
-Since we already set a lower bound for the table, no lower outlier exists.
-Only 4 outliers were present, which were removed.

Fig 8. Statistical Visualization of “Table” after cleaning.

3.2.7 X-value of the diamond, “X”





-The outliers were removed with the help of IQR.
-The remaining x values of the diamonds stay.

Fig 9. Statistical Visualization of “X-Value” after cleaning.

3.2.8 Y-value of the diamond, “Y”

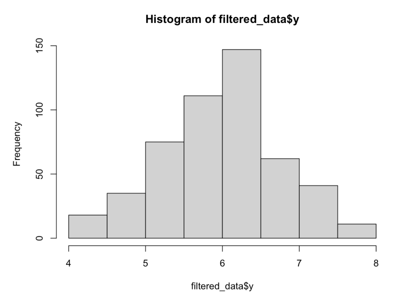
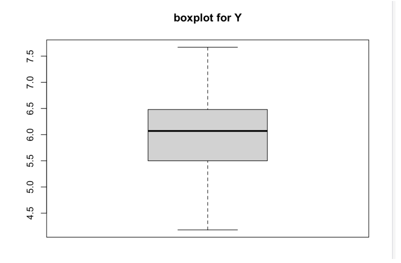
		-There were no outliers in the case of y.

Fig 10. Statistical Visualization of "Y-Value" after cleaning.

3.2.9 Z-value of the diamond, "Z"

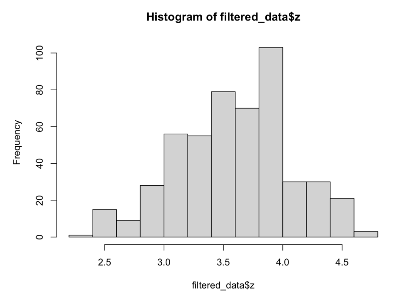
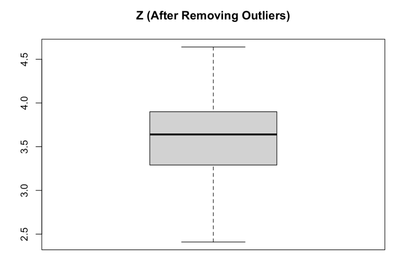
		-The outliers present in z, have been removed with the help of IQR

Fig 11. Statistical Visualization of "Z-Value" after cleaning.

3.3 Final dataset for analysis

Based on the above analysis the dataset is further reduced to 499 observations, and with 11 variables, because this time log(price) has been appended to the list of variables. Therefore, log(price) will be compared with all the other variables.

4. Statistical Analysis

4.1 Correlations between log (price) and Continuous Variables

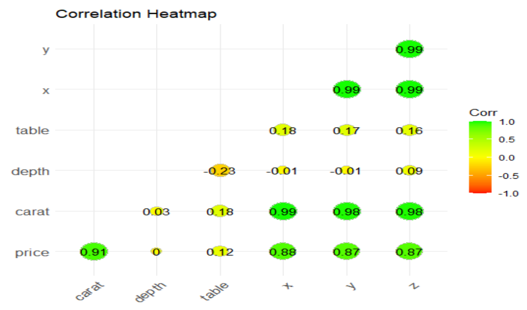


Fig 12. Correlation Heatmap comparing "log(Price)" and all Continuous Variables

The correlation coefficient can help us to analyse the relationships between two variables. With a positive coefficient, it can suggest a positive relationship between two variables, where an increase in one variable means there is an increase in another. The opposite is true for negative coefficients, whereas a zero would mean there is no relationship between the variables. On top of this the p values can show whether these observations are significant or not.

The below variables have an extremely small p value, which indicates that the observations are significant, and the following are their respective relationships with log(price):

- Carat has a strong positive relationship with log(price) (r = 0.94)
- Table has a weak positive relationship with log(price) (r = 0.15)
- Length, width and depth (x,y,z) all have equal strong positive relationship (r=0.96)

Depth has both a high p value of 0.896 and a very low r value of 0.01, which shows that it is insignificant and has virtually no correlation with log(price)

Also to take note is that x, y and z are highly correlated with carat, with the values of 0.99, 0.98 and 0.96 respectively. This is because x,y and z make up carats.

First part is q test - test for outliers, a lot of outliers (nanda)
Categorical - anova for each - all p value - all significant (nanda)

4.2 Statistical Tests

4.2.1 Visualization of the outliers in price

In order to visualise the outliers in price, we conduct a q-test with our price. The q-Test calculates the standardized scores using the t-test method. The scores help us ascertain how far the data points are away from the mean, the t-scores are then presented as a box plot as found below.

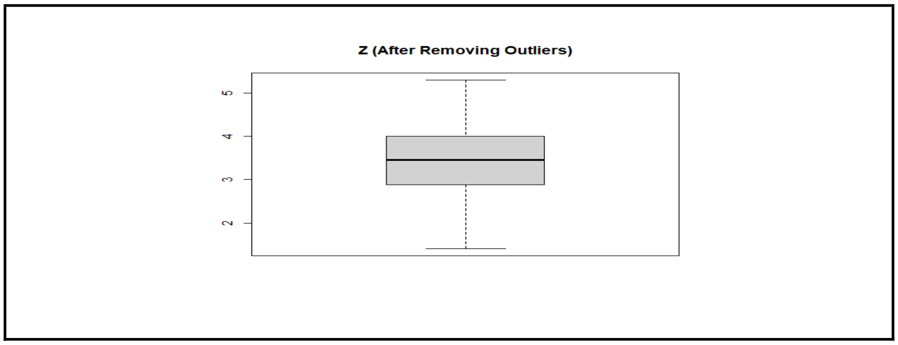
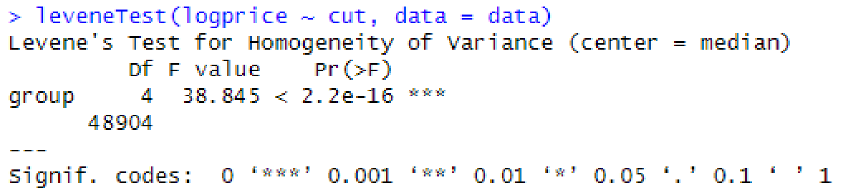


Fig 13. Box plot of all "t-scores" generated

From the graph, we can observe that the graph is symmetric about 0, implying that the price values are not skewed. Outliers are present below the box, this means that some values of price are significantly lower than expected with t-scores < -2.

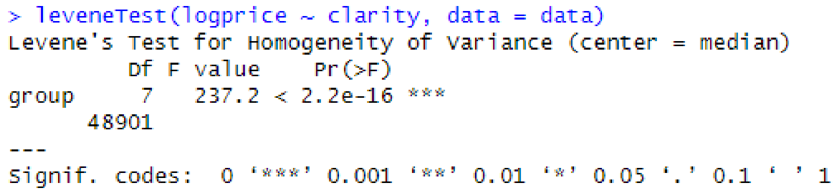
In the following section, we conduct more statistical tests to ascertain the relationships between our diamond attributes against Price. With our dataset containing both categorical (cut, clarity and color) and numerical attributes, (“x”, “y”, “z”, carat, depth and table). We begin with ascertaining the relationships between our categorical variables and Price of the diamonds.

To achieve this, we use the “Analysis of Variance” (ANOVA) test to determine whether the mean of Price depends significantly on cut, clarity or color. However, an assumption of utilising ANOVA is the homogeneity of variance - that the variance must be approximately the same across the levels of a factor. To ascertain this, we first conducted a Levene's Test of Homogeneity on the 3 factors: cut, clarity and color. The results of the test are found below



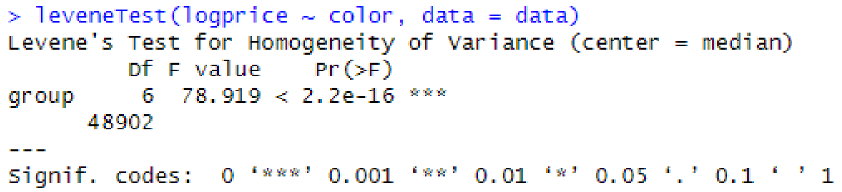
```
> leveneTest(logprice ~ cut, data = data)
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)    
group     4  38.845 < 2.2e-16 ***
      48904                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 14. Levene's Test of Homogeneity on “Cut”



```
> leveneTest(logprice ~ clarity, data = data)
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)    
group     7   237.2 < 2.2e-16 ***
      48901                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 15. Levene's Test of Homogeneity on “Clarity”



```
> leveneTest(logprice ~ color, data = data)
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)    
group     6  78.919 < 2.2e-16 ***
      48902                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig 16. Levene's Test of Homogeneity on “Color”

H_0 : Variances are equal across groups

H_1 : Variances are different for at least one group

As we can see from the figures on top, all the 3 p-values for cut,clarity and color are <2.2e-16, which is less than the level of significance of 0.05. Hence we reject Null Hypothesis and conclude that the variances are different for the groups of cut, clarity and color.

Since the variance is not homogeneous, the criteria for conducting an ANOVA fails. With some research, we identified a variation of ANOVA that can be utilised in non-homogenous variances, Welch's ANOVA.

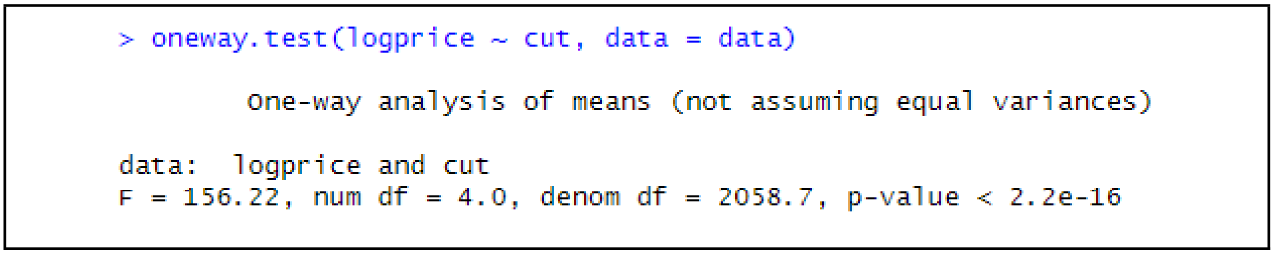
4.2.2 Relationship between cut and price

In the ANOVA to explore whether different cut categories (Fair, Good and Very Good) have significantly different prices.

H_0 : The mean price is the same across all cut categories Fair, Good and Very Good.

$\mu_{\text{Fair}} = \mu_{\text{Good}} = \mu_{\text{Very Good}}$

H_1 : The mean price is different between at least one of the cut categories



```
> oneway.test(logprice ~ cut, data = data)

        One-way analysis of means (not assuming equal variances)

data:  logprice and cut
F = 156.22, num df = 4.0, denom df = 2058.7, p-value < 2.2e-16
```

Fig 17. One-way ANOVA on "Cut"

Test returns a p-value of 6.739091e-80, which is lesser than our level of significance of 0.05. This indicates that the means of the different cut categories are significantly different. As such, there is sufficient evidence to reject the Null Hypothesis and conclude that the mean price of at least one cut category is significantly different from the rest.

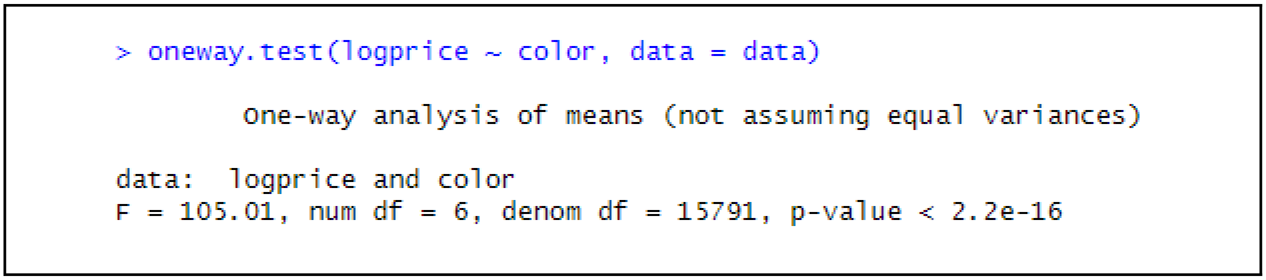
4.2.3 Relationship between color and logprice

We will conduct an ANOVA between color and logprice to determine whether there are differences in color categories (D, E, F, G, H, I, J) present.

H_0 : The mean price is the same across all color categories.

$\mu_D = \mu_E = \mu_E = \mu_F = \mu_G = \mu_H = \mu_I = \mu_J$

H_1 = The mean price is different between at least one of the color categories



```
> oneway.test(logprice ~ color, data = data)

        One-way analysis of means (not assuming equal variances)

data:  logprice and color
F = 105.01, num df = 6, denom df = 15791, p-value < 2.2e-16
```

Fig 18. One-way ANOVA on "Color"

Test returns a p-value of 1.436858e-136, which is lesser than our level of significance of 0.05. This indicates that the means of at least one of the different color categories is significantly different. As such, there is sufficient evidence to reject the Null Hypothesis and conclude that the mean price of at least one color category is significantly different from the rest and a posthoc test will be done to see whether the difference is between only a few specific groups or differences between all the color categories..

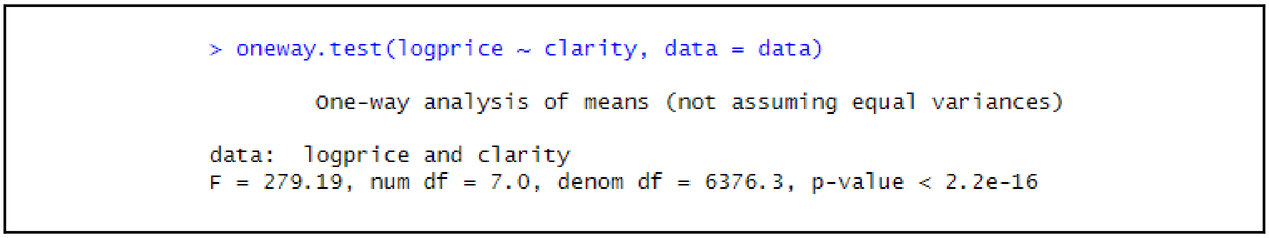
4.2.4 Relationship between clarity and logprice

Similar to what we have done with logprice and color, we will now run an ANOVA between logprice and clarity to see whether there are differences in mean price present between the different clarities(I1, IF, SI1, SI2, VS1,VS2,VVS1,VVS2).

H_0 : The mean price is the same across all clarity categories.

$\mu_{I1} = \mu_{IF} = \mu_{SI1} = \mu_{SI2} = \mu_{VS1} = \mu_{VS2} = \mu_{VVS1} = \mu_{VVS2}$

H_1 = The mean price is different between at least one of the clarity categories



```
> oneway.test(logprice ~ clarity, data = data)

        One-way analysis of means (not assuming equal variances)

data:  logprice and clarity
F = 279.19, num df = 7.0, denom df = 6376.3, p-value < 2.2e-16
```

Fig 19. One-way ANOVA on "Clarity"

Test returns a p-value of 2.3339987e-120, which is lesser than our level of significance of 0.05. This indicates that the means of at least one of the different clarity categories is significantly different. As such, there is sufficient evidence to reject the Null Hypothesis and conclude that the mean price of at least one clarity category is significantly different from the rest.

Similar to color, a post-hoc analysis is needed to ascertain which are the specific clarity types that have these significantly different mean prices as compared to the rest.

4.3 Understanding the interdependence of the categorical variables

The objective of this analysis is to investigate the relationships between three key categorical variables in the dataset: **Cut**, **Color**, and **Clarity**. These variables represent essential attributes of diamonds, and understanding their interdependence is crucial for uncovering underlying patterns in the data.

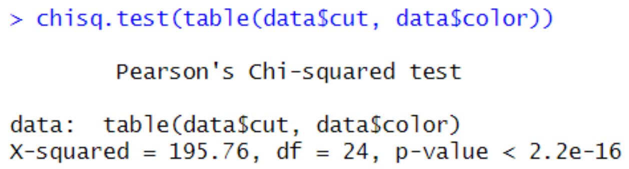
Specifically, we want to determine whether the distribution of one categorical variable (e.g., Clarity) is influenced by another (e.g., Cut). If two variables are **not independent**, it may indicate that one variable could potentially be used to predict or explain variations in the other.

To assess the relationships between the three categorical variables—**Cut**, **Color**, and **Clarity**—we conducted pairwise Pearson's Chi-Square tests of independence using R's chisq.test() function.

Null Hypothesis (H_0): The two variables are independent.
Alternative Hypothesis (H_1): The two variables are not independent.

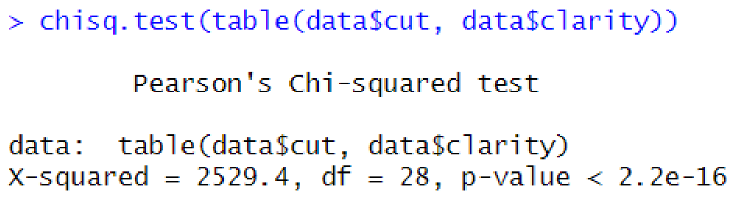
The output includes:

X-squared: The Chi-Square test statistic, measuring the discrepancy between observed and expected counts under independence. **df (degrees of freedom)**: Calculated as (number of rows −1)×(number of columns −1) , where r and c are the number of categories in each variable. **p-value**: Indicates the statistical significance of the result. A p-value < 0.05 suggests strong evidence to reject the null hypothesis.



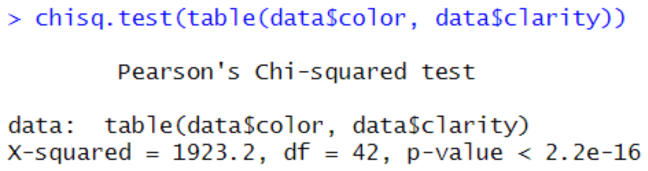
```
> chisq.test(table(data$cut, data$color))

	Pearson's Chi-squared test

data:  table(data$cut, data$color)
X-squared = 195.76, df = 24, p-value < 2.2e-16
```

Fig 20. Chi-Square Test between "Cut" & "Color"



```
> chisq.test(table(data$cut, data$clarity))

	Pearson's Chi-squared test

data:  table(data$cut, data$clarity)
X-squared = 2529.4, df = 28, p-value < 2.2e-16
```

Fig 21. Chi-Square Test between "Cut" & "Clarity"



```
> chisq.test(table(data$color, data$clarity))

	Pearson's Chi-squared test

data:  table(data$color, data$clarity)
X-squared = 1923.2, df = 42, p-value < 2.2e-16
```

Fig 22. Chi-Square Test between "Color" & "Clarity"

All three tests returned extremely small p-values (< 2.2e-16), providing strong evidence against the null hypothesis in each case. Therefore, we conclude that there are statistically significant associations between:
(Cut and Color) , (Cut and Clarity) , and (Color and Clarity)

This suggests that none of the three categorical variables are independent of each other in this dataset.

4.4 Games-Howell Post Hoc Analysis

We run a post Hoc test in R using the function called posthocTGH() from the userfriendlyscience package. This test checks whether there is a statistically significant difference among the means of several groups.
SInce the ANOVA detected at least one significant difference among the three categories, a post hoc test is performed to figure out which specific pairs of groups differ from each other.

The output contains the following terms,

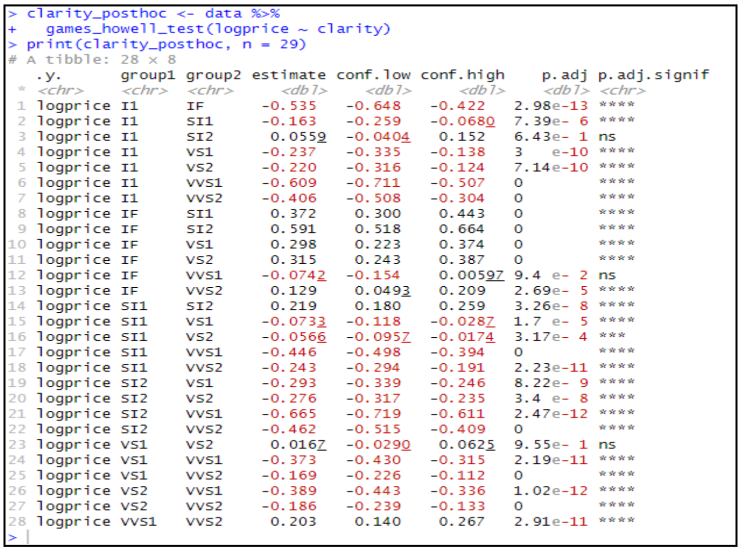
Estimate:
The difference in mean price between the two groups in a variable. A positive value means the first group listed has a higher mean price; a negative value means it has a lower mean price.

Conf.low and conf.high :
The lower and upper bounds of the *95% confidence interval* for that difference. If this interval does not contain zero, that difference is typically considered statistically significant. If zero is within the interval, the test did not detect a statistically significant difference between the two means (at the chosen family-wise confidence level).
p adj:
The p-value, adjusted for multiple comparisons using the Tukey HSD procedure. If p adj is less than 0.05, it indicates a statistically significant difference in mean price between those two clarity categories. This adjustment helps reduce the likelihood of incorrectly claiming statistical significance when making many pairwise comparisons.

Null Hypothesis (H_0): There is no significant difference between the two groups
Alternative Hypothesis (H_1): There is significant difference between the two groups

4.4.1 Games-Howell Post Hoc for clarity:

Shows the pairwise comparisons between each level of clarity (IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1).



```
> clarity_posthoc <- data %>%
+   games_howell_test(logprice ~ clarity)
> print(clarity_posthoc, n = 29)
# A tibble: 28 × 8
   .y.      group1 group2 estimate conf.low conf.high    p.adj p.adj.signif
 * <chr>    <chr>  <chr>     <dbl>    <dbl>     <dbl>    <dbl> <chr>       
 1 logprice I1     IF      -0.535   -0.648   -0.422   2.98e-13 ****        
 2 logprice I1     SI1     -0.163   -0.259   -0.0680  7.39e- 6 ****        
 3 logprice I1     SI2      0.0559  -0.0404   0.152   6.43e- 1 ns          
 4 logprice I1     VS1     -0.237   -0.335   -0.138   3   e-10 ****        
 5 logprice I1     VS2     -0.220   -0.316   -0.124   7.14e-10 ****        
 6 logprice I1     VVS1    -0.609   -0.711   -0.507   0        ****        
 7 logprice I1     VVS2    -0.406   -0.508   -0.304   0        ****        
 8 logprice IF     SI1      0.372    0.300    0.443   0        ****        
 9 logprice IF     SI2      0.591    0.518    0.664   0        ****        
10 logprice IF     VS1      0.298    0.223    0.374   0        ****        
11 logprice IF     VS2      0.315    0.243    0.387   0        ****        
12 logprice IF     VVS1    -0.0742  -0.154    0.00597 9.4 e- 2 ns          
13 logprice IF     VVS2     0.129    0.0493   0.209   2.69e- 5 ****        
14 logprice SI1    SI2      0.219    0.180    0.259   3.26e- 8 ****        
15 logprice SI1    VS1     -0.0733  -0.118   -0.0287  1.7 e- 5 ****        
16 logprice SI1    VS2     -0.0566  -0.0957  -0.0174  3.17e- 4 ***         
17 logprice SI1    VVS1    -0.446   -0.498   -0.394   0        ****        
18 logprice SI1    VVS2    -0.243   -0.294   -0.191   2.23e-11 ****        
19 logprice SI2    VS1     -0.293   -0.339   -0.246   8.22e- 9 ****        
20 logprice SI2    VS2     -0.276   -0.317   -0.235   3.4 e- 8 ****        
21 logprice SI2    VVS1    -0.665   -0.719   -0.611   2.47e-12 ****        
22 logprice SI2    VVS2    -0.462   -0.515   -0.409   0        ****        
23 logprice VS1    VS2      0.0167  -0.0290   0.0625  9.55e- 1 ns          
24 logprice VS1    VVS1    -0.373   -0.430   -0.315   2.19e-11 ****        
25 logprice VS1    VVS2    -0.169   -0.226   -0.112   0        ****        
26 logprice VS2    VVS1    -0.389   -0.443   -0.336   1.02e-12 ****        
27 logprice VS2    VVS2    -0.186   -0.239   -0.133   0        ****        
28 logprice VVS1   VVS2     0.203    0.140    0.267   2.91e-11 ****        
> 
```

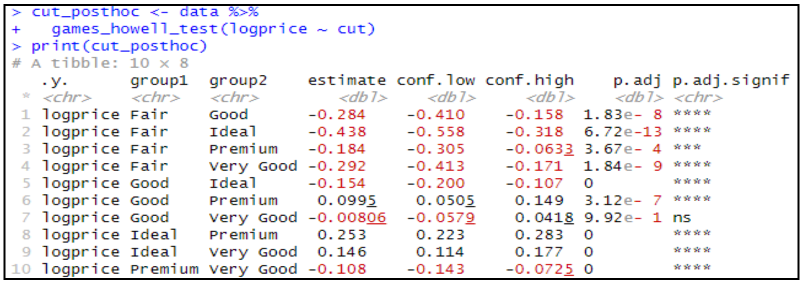
Fig 23. Games-Howell Post Hoc for "Clarity"

Rows with p-value<0.05 are statistically significant and indicate a difference in mean price of the two categories. For example IF and I1 are significantly different with I1 being more expensive on average.

We can infer from the table that SI2 has the highest mean value followed by SI1, I1. Just behind them come VS1 and VS2 with not much significant difference between them, followed by VVS2. VVS1 and IF come last with the lowest mean price.

4.4.2 Games-Howell Post Hoc for cut:

Shows the pairwise comparisons between each level of cut (Premium, Very Good, Good, Ideal, Fair).



```
> cut_posthoc <- data %>%
+   games_howell_test(logprice ~ cut)
> print(cut_posthoc)
# A tibble: 10 × 8
   .y.      group1  group2    estimate conf.low conf.high    p.adj p.adj.signif
 * <chr>    <chr>   <chr>        <dbl>    <dbl>     <dbl>    <dbl> <chr>       
 1 logprice Fair    Good      -0.284    -0.410    -0.158  1.83e- 8 ****        
 2 logprice Fair    Ideal     -0.438    -0.558    -0.318  6.72e-13 ****        
 3 logprice Fair    Premium   -0.184    -0.305    -0.0633 3.67e- 4 ***         
 4 logprice Fair    Very Good -0.292    -0.413    -0.171  1.84e- 9 ****        
 5 logprice Good    Ideal     -0.154    -0.200    -0.107  0        ****        
 6 logprice Good    Premium    0.0995    0.0505    0.149  3.12e- 7 ****        
 7 logprice Good    Very Good -0.00806  -0.0579    0.0418 9.92e- 1 ns          
 8 logprice Ideal   Premium    0.253     0.223     0.283  0        ****        
 9 logprice Ideal   Very Good  0.146     0.114     0.177  0        ****        
10 logprice Premium Very Good -0.108    -0.143    -0.0725 0        ****        
```

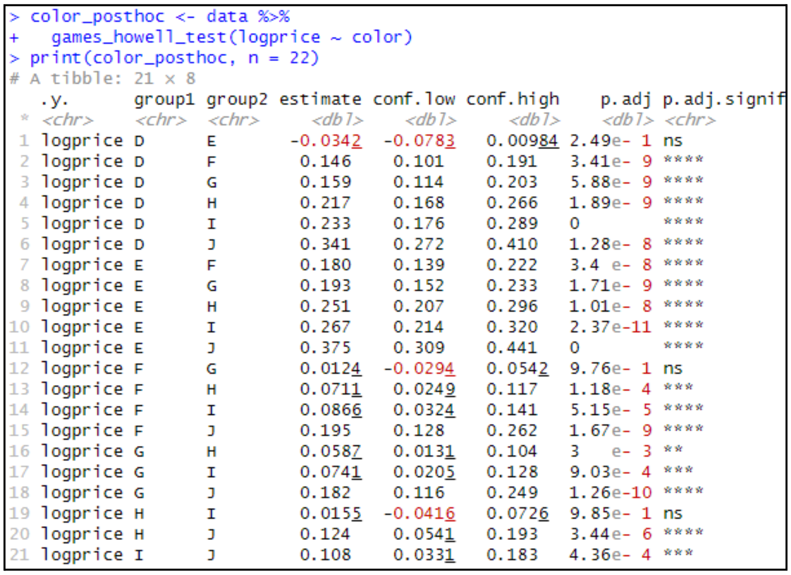
Fig 24. Games-Howell Post Hoc for “Cut”

Same as before, rows with p-value<0.05 are statistically significant and it gives us enough evidence to reject the null hypothesis.

We can infer from the results that Fair has the highest mean price followed closely by Premium. Very Good and Good are not significantly different. Ideal takes last place

4.4.3 Games-Howell Post Hoc for colour:

Shows the pairwise comparisons between each level of colour (D,E,F,G,H,I,J)



```
> color_posthoc <- data %>%
+   games_howell_test(logprice ~ color)
> print(color_posthoc, n = 22)
# A tibble: 21 x 8
   .y.      group1 group2 estimate conf.low conf.high    p.adj p.adj.signif
 * <chr>    <chr>  <chr>     <dbl>    <dbl>     <dbl>    <dbl> <chr>
 1 logprice D      E       -0.0342  -0.0783   0.00984 2.49e- 1 ns
 2 logprice D      F        0.146    0.101    0.191   3.41e- 9 ****
 3 logprice D      G        0.159    0.114    0.203   5.88e- 9 ****
 4 logprice D      H        0.217    0.168    0.266   1.89e- 9 ****
 5 logprice D      I        0.233    0.176    0.289   0        ****
 6 logprice D      J        0.341    0.272    0.410   1.28e- 8 ****
 7 logprice E      F        0.180    0.139    0.222   3.4 e- 8 ****
 8 logprice E      G        0.193    0.152    0.233   1.71e- 9 ****
 9 logprice E      H        0.251    0.207    0.296   1.01e- 8 ****
10 logprice E      I        0.267    0.214    0.320   2.37e-11 ****
11 logprice E      J        0.375    0.309    0.441   0        ****
12 logprice F      G        0.0124  -0.0294   0.0542  9.76e- 1 ns
13 logprice F      H        0.0711   0.0249   0.117   1.18e- 4 ***
14 logprice F      I        0.0866   0.0324   0.141   5.15e- 5 ****
15 logprice F      J        0.195    0.128    0.262   1.67e- 9 ****
16 logprice G      H        0.0587   0.0131   0.104   3    e- 3 **
17 logprice G      I        0.0741   0.0205   0.128   9.03e- 4 ***
18 logprice G      J        0.182    0.116    0.249   1.26e-10 ****
19 logprice H      I        0.0155  -0.0416   0.0726  9.85e- 1 ns
20 logprice H      J        0.124    0.0541   0.193   3.44e- 6 ****
21 logprice I      J        0.108    0.0331   0.183   4.36e- 4 ***
```

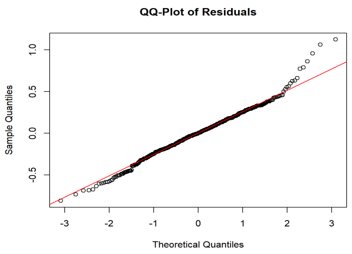
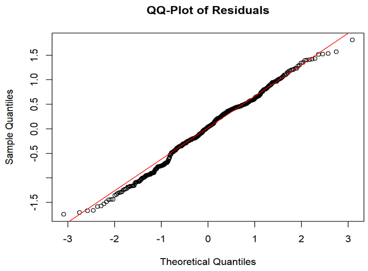
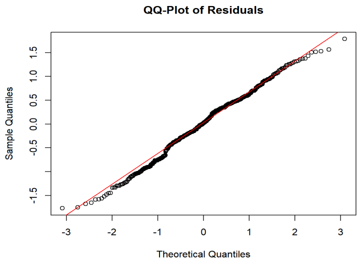
Fig 25. Games-Howell Post Hoc for “Color”

Once again we analyse the rows with p-value less than 0.05 as they suggest significant difference in the mean price.
We can see that J has the highest mean price. Best of the rest is taken by H and I. Then comes G, F with D and E sharing last.

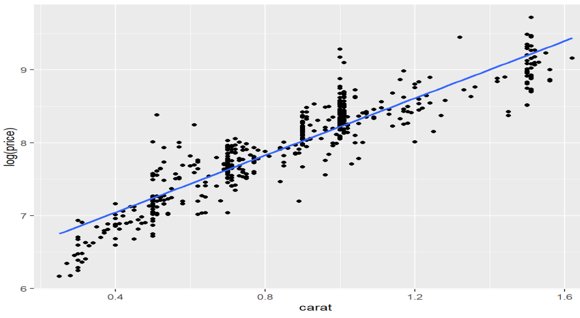
4.5 Relationship between Carat, Depth, Table and logprice

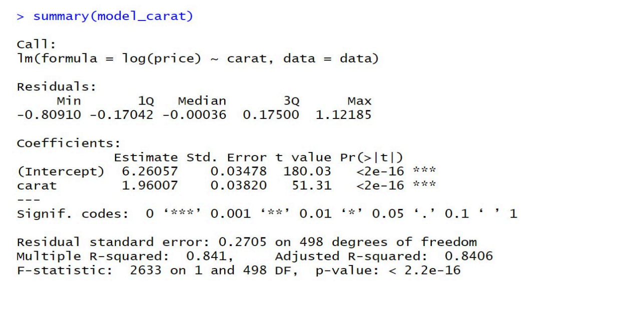
In this section, we determine whether the log(price) of a diamond depends on its carat, depth or table. Simple linear regression analysis was employed for each numerical variable, that is: **(Logprice and Carat) , (Logprice and Depth) ,** and **(Logprice and Table)**

Comparing metrics from the summary of each linear regression like “R-Square” value and QQ-Plot of residuals, ”Carat”was the single most important variable in affecting log(Price).

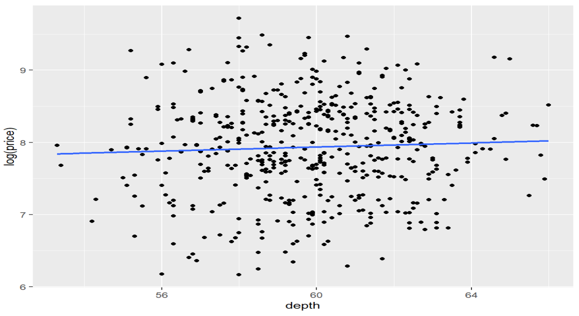
Variable	**P-value**	**R-squared**	**QQ-plot of Residuals**
Carat	< 2e-16	0.8406	
Depth	0.281	0.0003275	
Table	0.664	-0.001629	

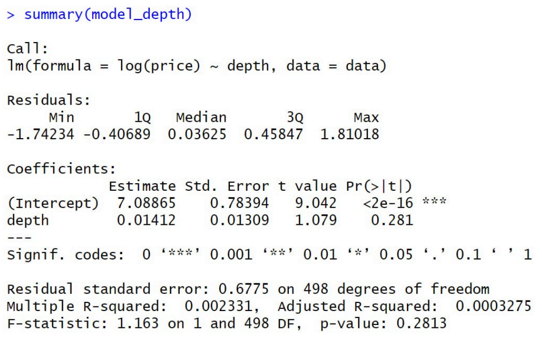
The linear regression model was plotted along with the summary statistic for each linear model



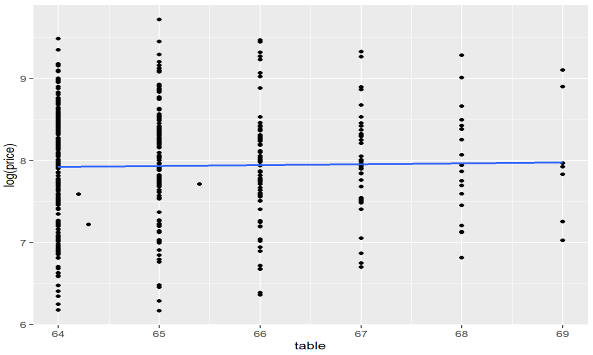


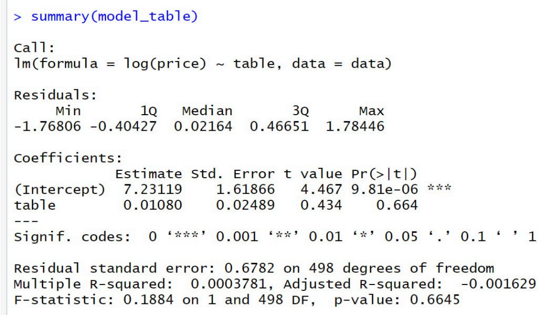
```
> summary(model_carat)

Call:
lm(formula = log(price) ~ carat, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.80910 -0.17042 -0.00036  0.17500  1.12185 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.26057    0.03478  180.03   <2e-16 ***
carat        1.96007    0.03820   51.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2705 on 498 degrees of freedom
Multiple R-squared:  0.841,	Adjusted R-squared:  0.8406 
F-statistic:  2633 on 1 and 498 DF,  p-value: < 2.2e-16
```





```
> summary(model_depth)

Call:
lm(formula = log(price) ~ depth, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.74234 -0.40689  0.03625  0.45847  1.81018 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.08865    0.78394   9.042   <2e-16 ***
depth        0.01412    0.01309   1.079    0.281    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6775 on 498 degrees of freedom
Multiple R-squared:  0.002331,	Adjusted R-squared:  0.0003275 
F-statistic: 1.163 on 1 and 498 DF,  p-value: 0.2813
```



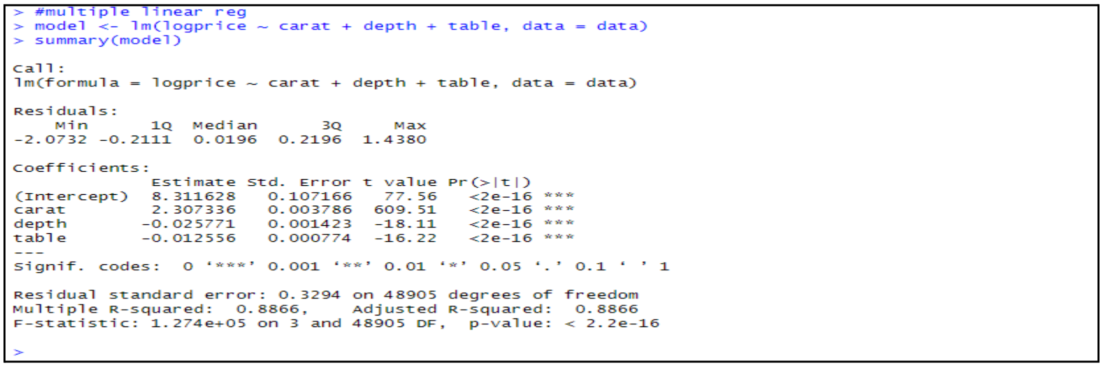


```
> summary(model_table)

Call:
lm(formula = log(price) ~ table, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.76806 -0.40427  0.02164  0.46651  1.78446 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.23119    1.61866   4.467 9.81e-06 ***
table        0.01080    0.02489   0.434    0.664    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6782 on 498 degrees of freedom
Multiple R-squared:  0.0003781,	Adjusted R-squared:  -0.001629 
F-statistic: 0.1884 on 1 and 498 DF,  p-value: 0.6645
```

4.6 Multiple Regression Model

We attempt to build a multiple regression model for log(price) based on the carat, depth and table. X, Y and Z are not included as they are highly correlated with carat and thus will lead to multicollinearity when building the model. All three variables are significant as indicated by their low p values as seen in the R code below. Using the information we obtained, the fitted model will be:

log(price) = 8.3116 + 2.3073*carat – 0.0258*depth – 0.0126*table



```
> #multiple linear reg
> model <- lm(logprice ~ carat + depth + table, data = data)
> summary(model)

Call:
lm(formula = logprice ~ carat + depth + table, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0732 -0.2111  0.0196  0.2196  1.4380 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.311628   0.107166   77.56   <2e-16 ***
carat        2.307336   0.003786  609.51   <2e-16 ***
depth       -0.025771   0.001423  -18.11   <2e-16 ***
table       -0.012556   0.000774  -16.22   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3294 on 48905 degrees of freedom
Multiple R-squared:  0.8866,	Adjusted R-squared:  0.8866 
F-statistic: 1.274e+05 on 3 and 48905 DF,  p-value: < 2.2e-16

>
```

Fig 27. Summary Statistic for Multivariate Linear Regression Model

5. Conclusion & Discussion

Diamonds are a luxury jewelry that people often look to when purchasing for special occasions such as weddings and engagement rings, and purchasing them can often consist of a significant portion of an individual's savings and income. Therefore having the information of what factors into the prices of diamonds can help an individual form a more informed decision for their purchases. We attempt to answer some of these questions with our analysis done in this report.

From our results, we can conclude the following:

- Carat of the diamond, which also includes length, width and height within it, will increase price significantly with a corresponding increase.
- Depth of the diamond does not affect its price.
- The categorical variables are not independent of each other.
- Clarity affects price, with SI2 generally corresponding to a higher priced diamond and IF to a lower priced one
- Cut affects the price, with Fair cuts corresponding to higher priced diamonds and decreases with each cut level with Ideal being the least
- Colour affects price, with better colors being higher priced

We are also able to model the price via a linear model with the factors of carat, depth and table, although the depth and table does not affect the price significantly as carat does. The two factors also decreases price with an increase in their values, while the opposite is true for the carat which is also shown in the linear model equation.

These conclusions are however only made with the one dataset that we have used for the report, therefore a full analysis of the whole market may change the values and significance of the factors. Another thing to take note is that the value of diamonds are tied to both its supply and also its perception as a jewelry. Decreases in the world's supply of diamond will eventually also affect prices in the long run. Furthermore the beauty of diamonds is subjective and it can be possible for the demand to shift preferences to other factors for example the table of the diamond instead of carat following fashion trends, hence these results that we obtain now may nor always hold true.

6)Appendix

Refer to the one in Zip file

7)References

https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction/notebook