STAT 1910 Project Final Submission - Diamonds Dataset

Zeyad A., Shriram S., Emmanuel O., Amanda Y. and Oluwademilade T.

Last updated on 2022-12-11

Contents

---

1. Introduction

Several factors determine the price of diamonds, such as, weight and cut quality. The goal of this project is to better understand the influence of the weight and the cut quality of the diamonds on their price.

We have decided to use this dataset which contains the prices and other attributes of 53,940 diamonds. The dataset was originally downloaded from Kaggle and modified to include the *id* column which represents a unique diamond identification number. Each case in the dataset represents a unique diamond number.

Here is a snapshot of 5 randomly chosen rows of the data set we'll use:

```
## # A tibble: 5 x 4
##      id price carat cut
##   <dbl> <dbl> <dbl> <fct>
## 1 34020   849  0.39 Ideal
## 2  8826  4478  1.12 Very Good
## 3 46208  1750  0.51 Very Good
## 4 47128  1829  0.52 Very Good
## 5 16740   612  0.28 Very Good
```

---

2. Exploratory data analysis

As seen in Table 1, our sample size (number of observations) is 53,940. The mean price of diamonds was the greatest for the Premium cut ($n = 13791$, $\bar{x} = 4584.3$, $sd = 4349.2$), Fair cut had the second-largest mean price ($n = 1610$, $\bar{x} = 4358.8$, $sd = 3560.4$). Next, the Very Good cut had the third-highest mean price ($n = 12082$, $\bar{x} = 3981.8$, $sd = 3935.9$). The Good cut had the second-lowest price ($n = 4906$, $\bar{x} = 3928.9$, $sd = 3681.6$). Finally, the Ideal cut had the lowest mean price for diamonds in this sample ($n = 1810$, $\bar{x} = 3457.5$, $sd = 3800.5$).

Table 1. Summary statistics of diamonds' prices across five cut qualities.

```
## # A tibble: 5 x 7
##   cut           n  mean median    sd   min   max
##   <fct>     <int> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Fair       1610 4359.  3282  3560.   337 18574
## 2 Good       4906 3929.  3050. 3682.   327 18788
## 3 Very Good 12082 3982.  2648  3936.   336 18818
## 4 Premium   13791 4584.  3185  4349.   326 18823
## 5 Ideal     21551 3458.  1810  3808.   326 18806
```

As depicted in Figure 2, the distribution of diamond prices is right-skewed - there are more observations for lower prices than mid-to-high prices. Hence, The median and the IQR would be used as summary statistics for this distribution because they are robust to outliers.

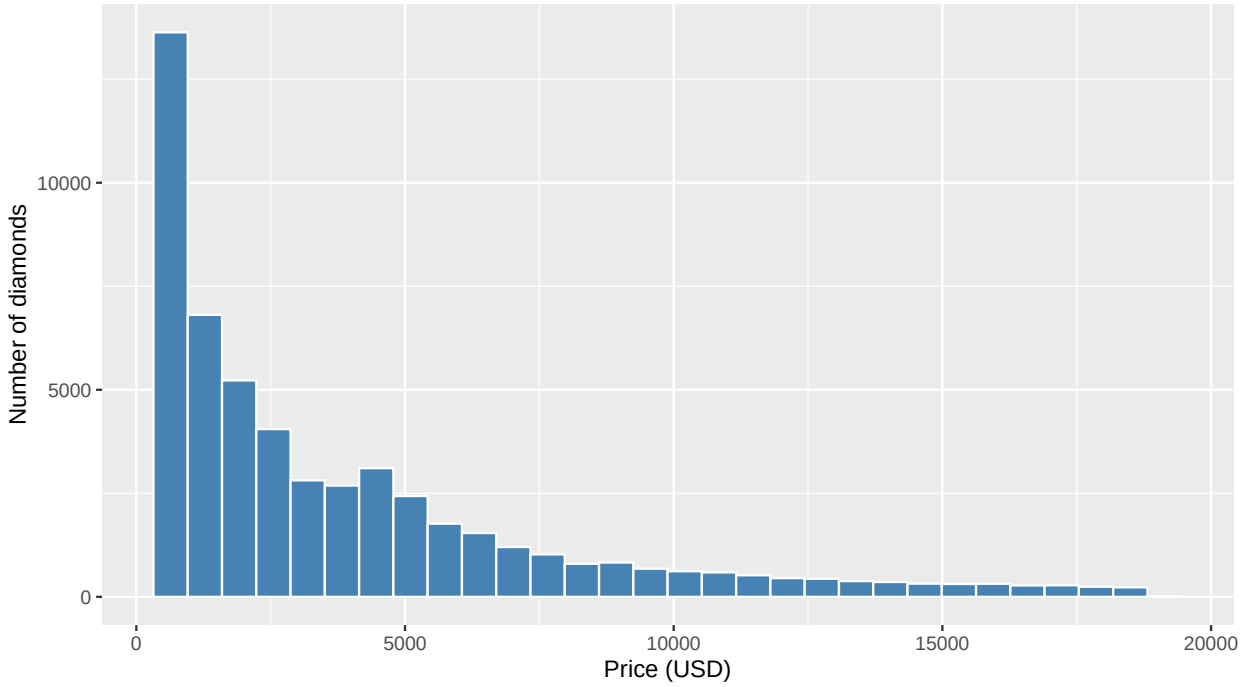

Figure 1: Figure 1. The distribution of diamonds' prices in US dollars

There seems to be a strong positive correlation between the weight of the diamonds and the price - seen in Figure 2.1. As the weight increases, there is an associated increase in the price. However, there appear to be some outliers on the right-hand-side of the graph.

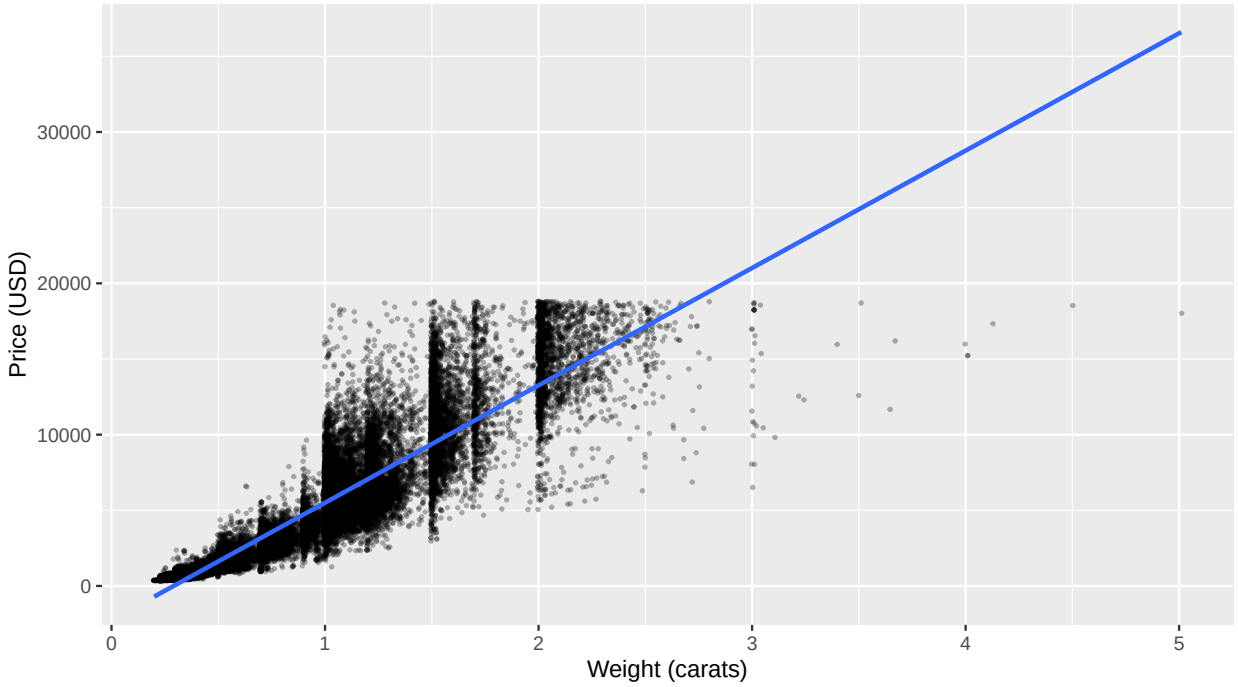

Figure 2: Figure 2.1. Relationship between the weight in carats of diamonds and price

We decided to visualize the log-log relationship, too, between the variables due to outliers. From this logarithmic-scaled plot in Figure 2.2, we can see that there is a stronger, positive correlation between price and weight. The linear model of this scaled data fits well. This suggests that the equation of the line may be similar to $log_{10}(\widehat{price}) = \beta_0 + \beta_1 \times log_{10}(carat)$ therefore, $\widehat{price} = 10^{\beta_0} \times carat^{\beta_1}$ which is also equivalent to $\widehat{price} = A \times carat^{\beta_1}$

Looking at Figure 3.1, all the cuts have outliers,i.e., there are some extreme prices for each cut. For most, the median is closer to the bottom of the box and the whisker is shorter on the lower end of the box which confirms that the distribution is right-skewed. Furthermore, the location of all cuts is about the same. However, their spread is not the same. The Premium cut has the largest spread, whereas the Fair cut has the lowest spread.

Again, logarithmic scale reduces the number of outliers and roughly reduces skewness as shown in Figure 3.2. Moreover, diamonds prices look to be the greatest for fair and premium cut quality, and the lowest for the ideal, however, the differences do not seem to be extreme.

As depicted by Figure 4.1, the positive relationship between the weight of the diamond and the price still holds true for each cut quality. However, the slope of the fair cut seems to be less steeper than the other cuts, i.e., for every unit increase in the weight of a fair-cut diamond, there is a relatively-lower associated increase in the price with respect to other, better, cuts.

Looking at Figure 4.2, the regression lines for the cuts tells us that diamonds with Fair cuts, on average, have less value for their weight compared to all other cuts.

---

3. Multiple linear regression

3.1 Methods

The components of our multiple linear regression model are the following:

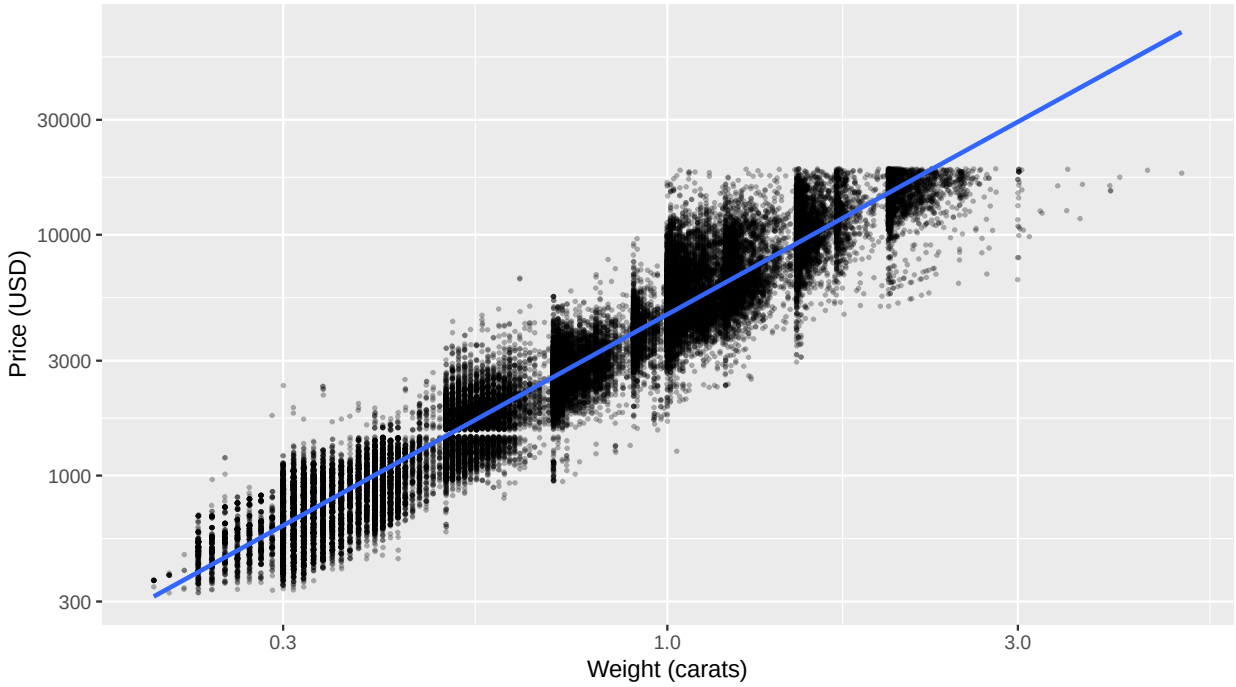

Figure 3: Figure 2.2. Relationship between the weight of diamonds and price - Logarithmic scale

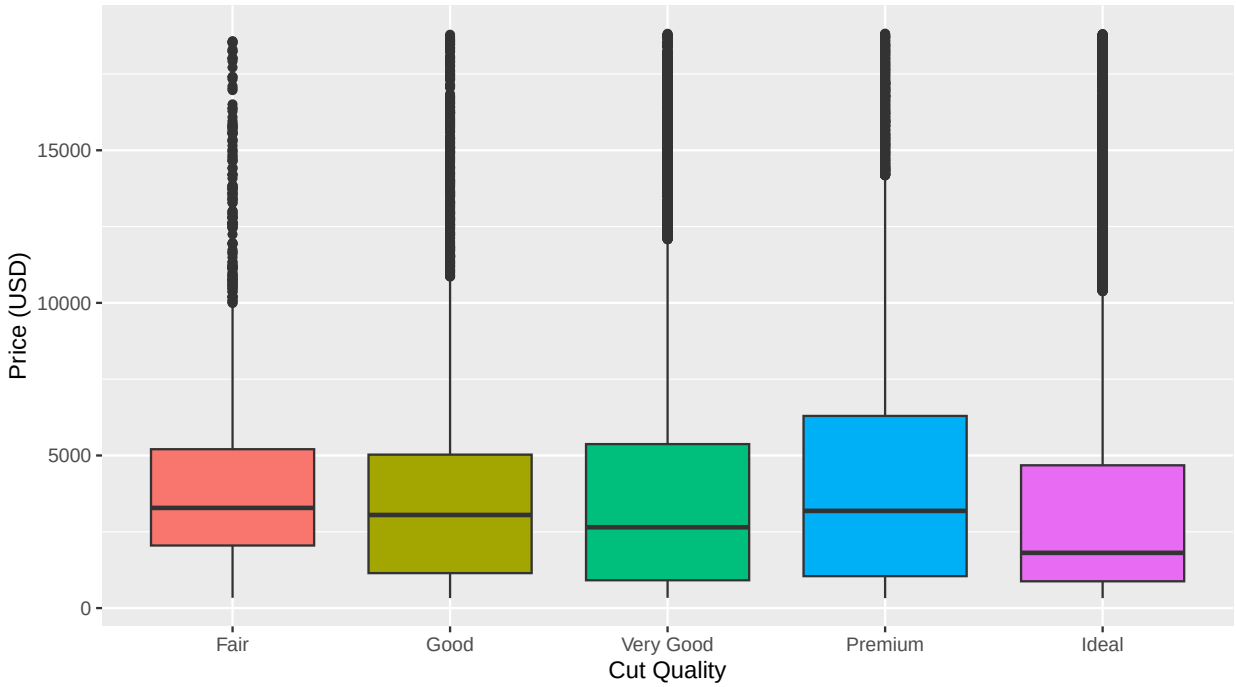

Figure 4: Figure 3.1. Relationship between the cut and the price of diamonds

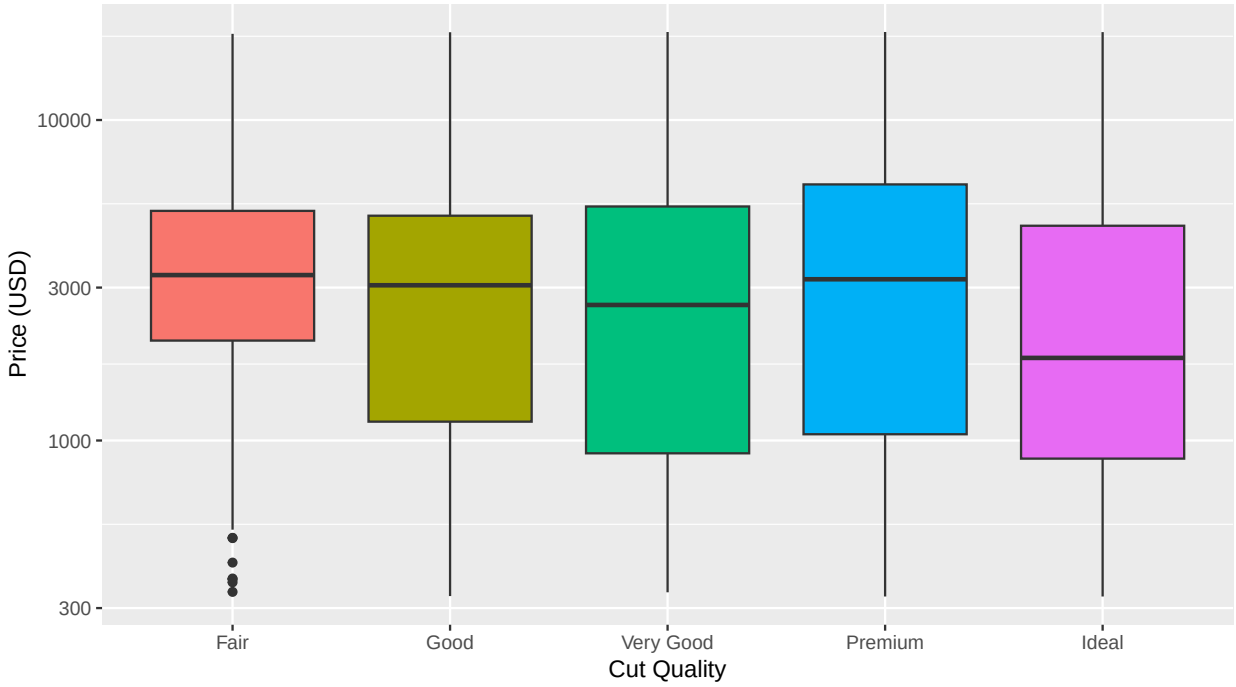

Figure 5: Figure 3.2. Relationship between the cut and the price of diamonds - Logarithmic scale

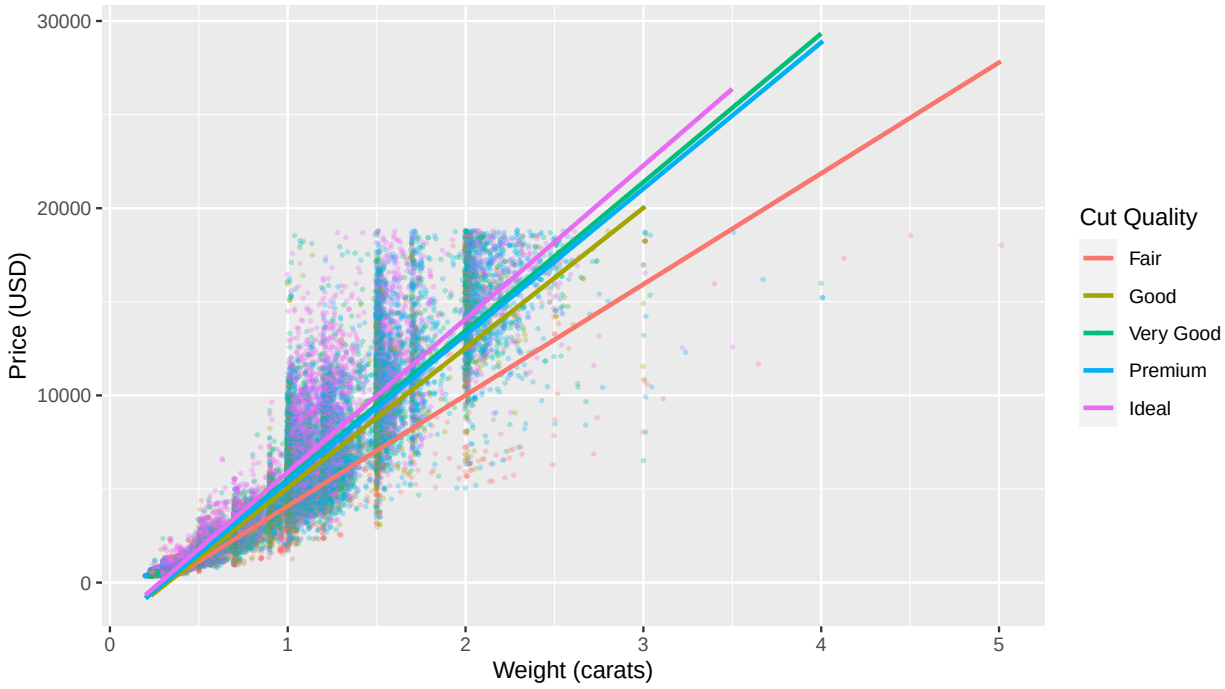

Figure 6: Figure 4.1. Relationship between the weight of the diamonds in carats, the price, and the cut

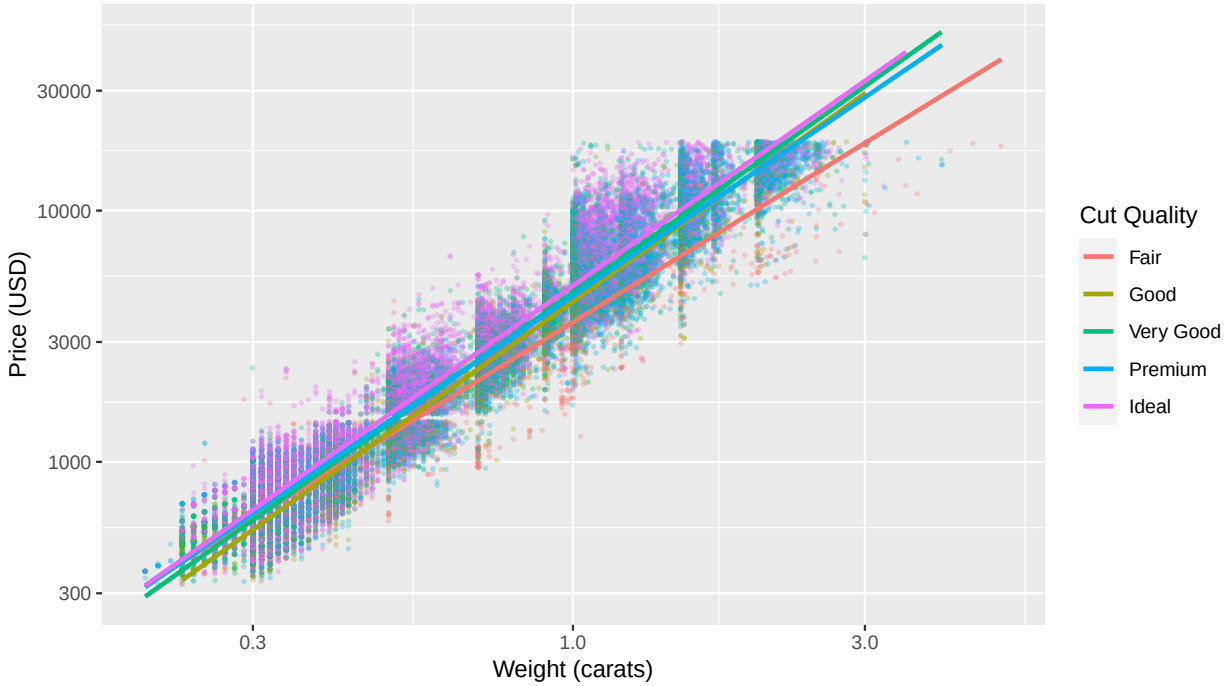

Figure 7: Figure 4.2. Relationship between the weight of the diamonds in carats, the price, and the cut - Logarithmic scale

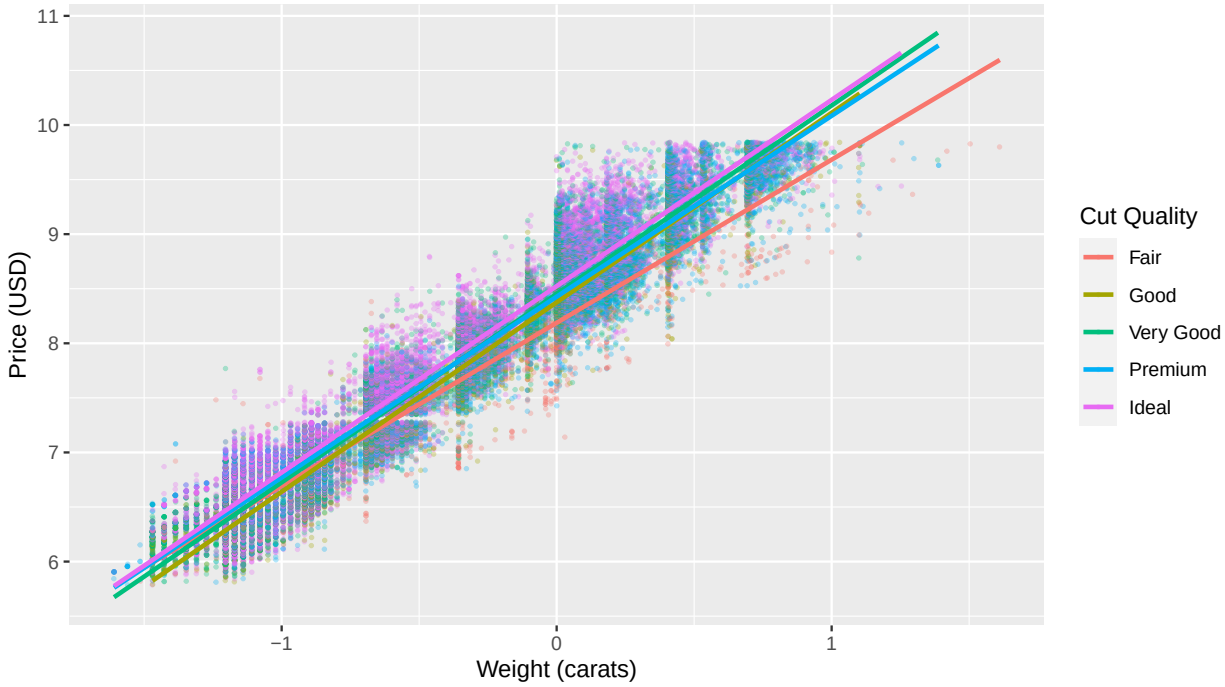

Figure 8: Figure 4.2. Relationship between the weight of the diamonds in carats, the price, and the cut - Logarithmic scale

- Outcome variable *price* = Price of diamonds in USD.
- Numerical explanatory variable *carat*= Weight, in carats, of the diamonds.
- Categorical explanatory variable *cut* = The quality of the cut of diamonds (fair, good, very good, premium, and ideal).

3.2 Model Results

Table 2.1. Regression table of diamonds' prices as a function of weight and cut quality with the baseline being the Fair cut quality.

```
## # A tibble: 6 x 7
##   term           estimate std_error statistic p_value lower_ci upper_ci
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept        -3875.      40.4     -95.9       0   -3955.   -3796.
## 2 carat             7871.      14.0     563.        0    7844.    7898.
## 3 cut: Good         1120.      43.5      25.8       0    1035.    1206.
## 4 cut: Very Good    1510.      40.2      37.5       0    1431.    1589.
## 5 cut: Premium      1439.      39.9      36.1       0    1361.    1517.
## 6 cut: Ideal        1801.      39.3      45.8       0    1724.    1878.
```

Table 2.2. Regression table of diamonds' log prices as a function of log weight and cut quality with the baseline being the Fair cut quality.

```
## # A tibble: 6 x 7
##   term           estimate std_error statistic p_value lower_ci upper_ci
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept         8.2       0.006    1293.        0    8.19     8.21
## 2 log(carat)        1.70      0.002     888.        0    1.69     1.7
## 3 cut: Good         0.163     0.007      22.3       0    0.149    0.178
## 4 cut: Very Good    0.241     0.007      35.5       0    0.227    0.254
## 5 cut: Premium      0.238     0.007      35.5       0    0.225    0.251
## 6 cut: Ideal        0.317     0.007      47.8       0    0.304    0.33
```

3.3 Interpreting the regression table

The regression equation for the price of diamonds is the following:

$$\begin{aligned}\widehat{price} =& b_0 + b_{carat} \cdot carat + b_{Good} \cdot 1_{is\ Good}(cut) + b_{Very\ Good} \cdot 1_{is\ Very\ Good}(cut) \\ &+ b_{Premium} \cdot 1_{is\ Premium}(cut) + b_{Ideal} \cdot 1_{is\ Ideal}(cut) \\ =& -3875.470 + 7871.082 \cdot carat + 1120.332 \cdot 1_{is\ Good}(cut) + 1510.135 \cdot 1_{is\ Very\ Good}(cut) \\ &+ 1439.077 \cdot 1_{is\ Premium}(cut) + 1800.924 \cdot 1_{is\ Ideal}(cut)\end{aligned}$$

- The intercept ($b_0 = -3875.470$) represents the price of a diamond when the cut quality is fair and the weight is zero (Table 2).
- The estimate for the slope for the weight of a diamond ($b_{carat} = 7871.082$) is the associated change in price depending on the weight of that diamond. Based on this estimate, for every one point increase in the weight, in carats, of a diamond, there was an associated increase in price of the diamond by 7871.082 USD on average.
- The estimate for Good Cut quality ($b_{Good} = 1120.332$).
- Very Good Cut quality ($b_{Very\ Good} = 1510.135$).
- Premium Cut quality ($b_{Premium} = 1439.077$).

- Ideal Cut quality ($b_{Ideal} = 1800.924$).

These slopes are the offsets in intercept relative to the baseline cut quality, Fair Cut, which is the intercept (Table 2). In simple terms, on average, the price of a Good Cut is 1120.332 USD higher than the price of a Fair Cut diamond, all else being equal, while the price of a Very Good Cut diamond is 1510.135 higher, all else being equal. Also, the higher tier diamonds, like Premium Cut and Ideal Cut diamonds, are 1439.077 and 1800.924 more than the price of a Fair Cut diamond respectively, all else being equal.

Thus, the five regression lines would have the equations:

$$\begin{aligned}
\text{Fair Cut (in red)} &: \widehat{price} = -3875.470 + 7871.082 \cdot carat \\
\text{Good Cut (in brown)} &: \widehat{price} = -2755.138 + 7871.082 \cdot carat \\
\text{Very Good Cut (in green)} &: \widehat{price} = -2365.335 + 7871.082 \cdot carat \\
\text{Premium Cut (in cyan)} &: \widehat{price} = -2436.393 + 7871.082 \cdot carat \\
\text{Ideal Cut (in pink)} &: \widehat{price} = -2074.546 + 7871.082 \cdot carat
\end{aligned}$$

3.4 Inference for multiple regression

Using the output of our regression table we will test two different null hypotheses. The first null hypothesis is that there is no relationship between the weight of diamonds and the price at the population level.

$$\begin{aligned}
H_0 &: \beta_{carat} = 0 \\
H_A &: \beta_{carat} \neq 0
\end{aligned}$$

We can see a positive relationship between the weight and price of the diamond ($\beta_{carat} = 7871.082$) in Table 2.1. Furthermore, this appears to be a meaningful relationship since in Table 2.1 we can see:

- The 95% confidence interval for the slope β_{carat} is $(7843.682, 7898.482)$ which is positive and does not contain zero.
- The p-value is extremely small that it is rounded to 0, so we reject the null hypothesis H_0 that $\beta_{carat} = 0$ in favor of the alternative H_A that β_{carat} is not 0 and positive.

Taking potential sampling variability into account (collecting diamond prices from a different seller for instance) the relationship appears to be positive.

In the second set of null hypotheses, we test whether all the differences in intercept for the non-baseline groups (good, very good, premium and ideal cuts) are zero.

$$\begin{aligned}
H_0 &: \beta_{Good} = 0 \\
H_A &: \beta_{Good} \neq 0
\end{aligned}$$

$$\begin{aligned}
H_0 &: \beta_{VeryGood} = 0 \\
H_A &: \beta_{VeryGood} \neq 0
\end{aligned}$$

$$\begin{aligned}
H_0 &: \beta_{Premium} = 0 \\
H_A &: \beta_{Premium} \neq 0
\end{aligned}$$

$$\begin{aligned}
H_0 &: \beta_{Ideal} = 0 \\
H_A &: \beta_{Ideal} \neq 0
\end{aligned}$$

In other words, we check if the intercept of the baseline cut, Fair-cut-diamond, is equal or not equal to the other cuts in non-baseline groups (Good, Very Good, Ideal and Premium).

From table 2.1, we can see that the observed differences in intercepts of the cuts of the diamond are all positive ($\beta_{Good} = 1120.332$, $\beta_{VeryGood} = 1510.135$, $\beta_{Premium} = 1439.077$ and $\beta_{Ideal} = 1800.924$) which is meaningful since in Table 2.1 we can see that:

- the 95% confidence intervals for the difference in intercepts β_{Good}, $\beta_{VeryGood}$, $\beta_{Premium}$ and β_{Ideal} do not include zero: (1035.073, 1205.591), (1431.265, 1589.006), (1360.941, 1517.214) and (1723.809, 1878.039) respectively. Thus, the difference of intercepts is not zero, and all cuts are not the same when compared to the Fair cut.
- The p-values of the differences are all extremely close to 0, so we reject all the null hypotheses that they are 0.

Therefore, the differences in intercepts are different from 0, and all the intercepts are not equal. This might not seem obvious in our observations from the visualization of the five regression lines in Figure 4.1 because the scale is very large due to the presence of outliers. However, statistically, these intercepts are not the same.

3.5 Residual Analysis

We conducted a residual analysis to see if there was any systematic pattern of residuals for the statistical model we ran. Because if there are systematic patterns, then we cannot fully trust our confidence intervals and p-values above.

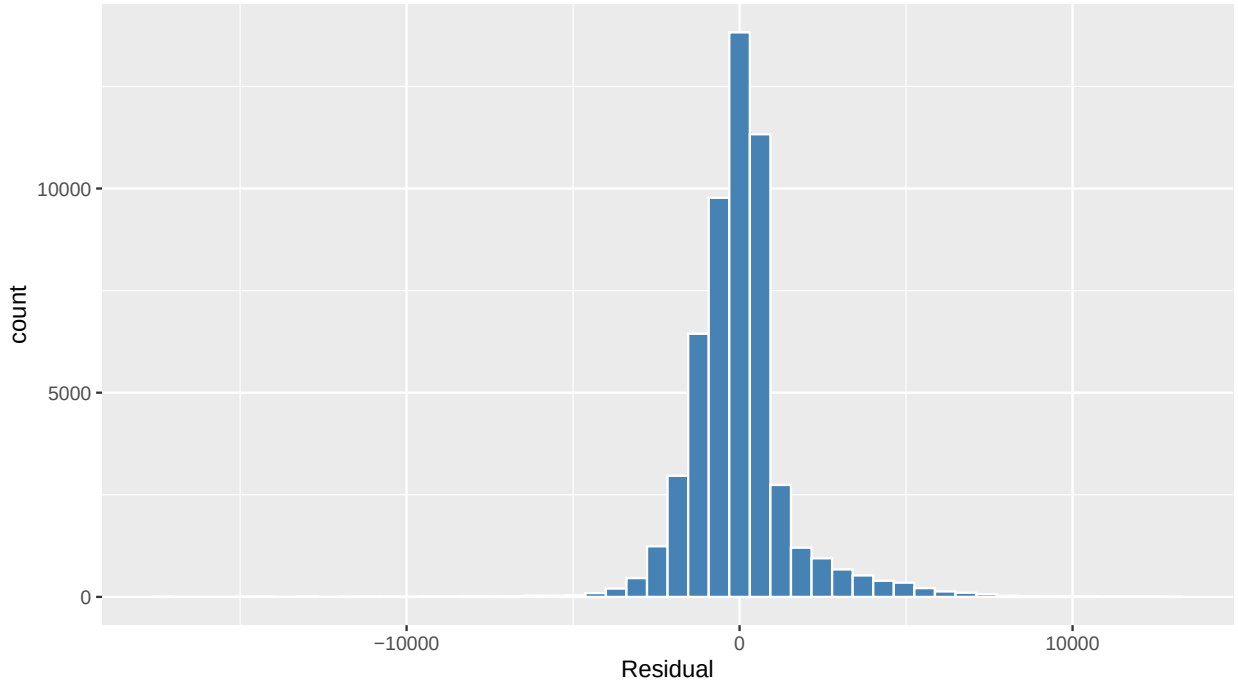

Figure 9: Figure 5.1. Histogram of residuals for statistical model

The model residuals were roughly normally distributed, though there were several outliers (Fig. 5.1). There appears to be a decreasing pattern of the residuals as the fitted or the weight values increase (Figs 6.1 & 7.1). There are numerous outliers on the bottom right side of the plots.

However, using the log-log model, the residuals were basically normally distributed with potentially less outliers (Fig. 5.2). In addition, there is less systematic pattern (decreasing trend) - roughly no pattern to either of the scatterplots (Figs. 6.2 & 7.2) and less outliers due to the log transformation as discussed previously.

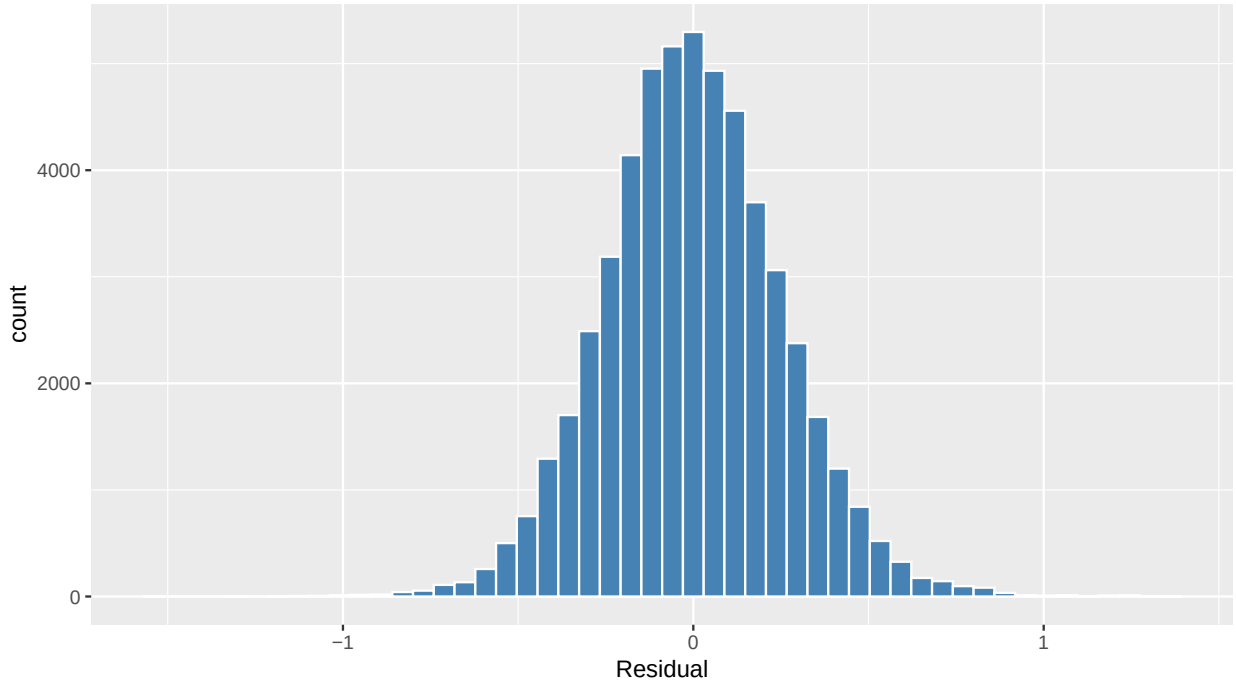

Figure 10: Figure 5.2. Histogram of residuals for statistical log model

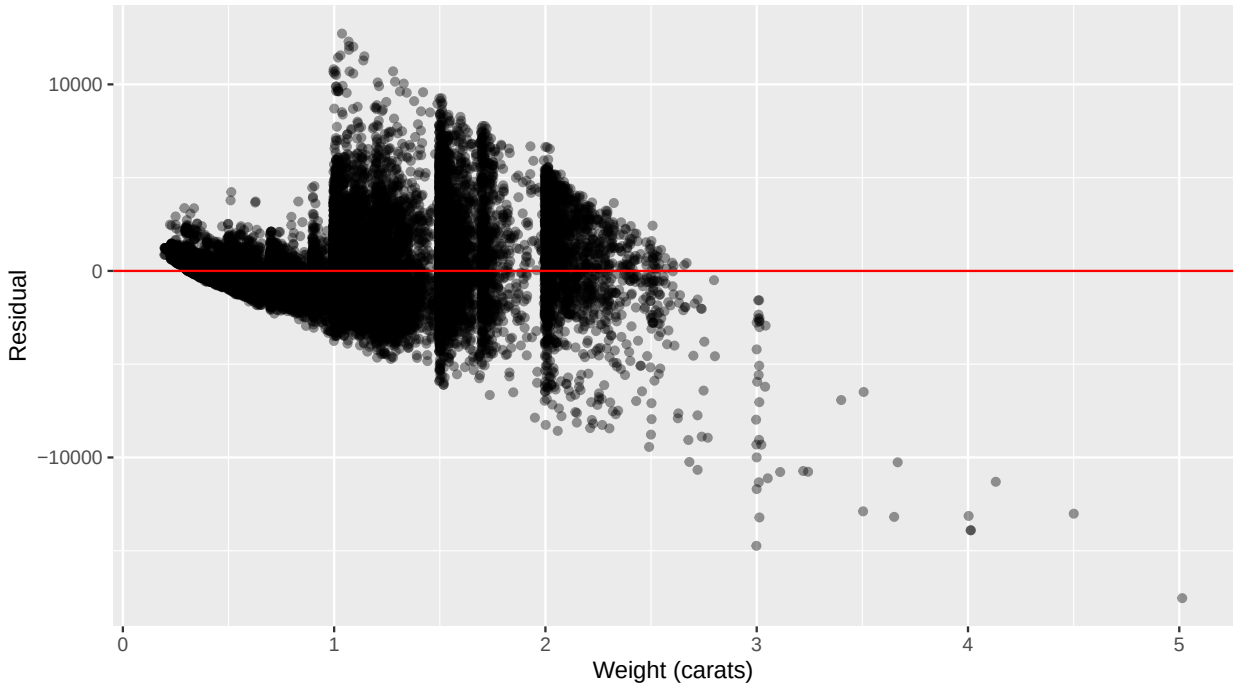

Figure 11: Figure 6.1. Scatterplots of residuals against the numeric explanatory variable.

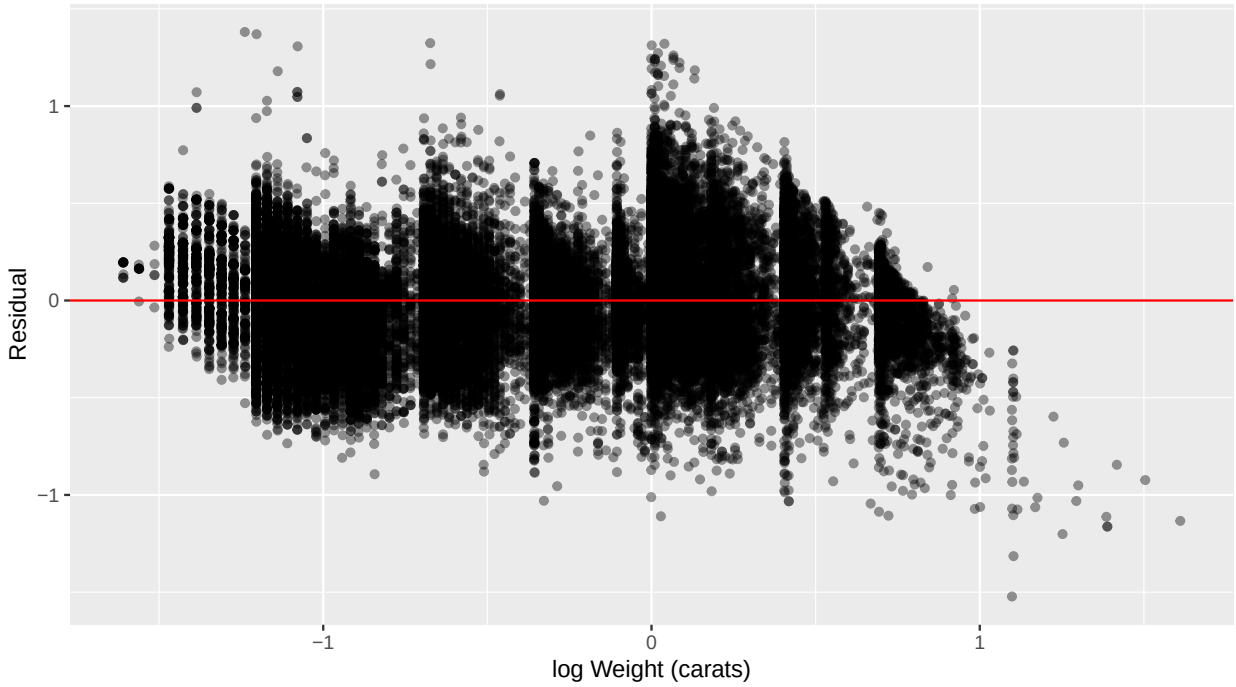

Figure 12: Figure 6.2. Scatterplots of residuals against the log numeric explanatory variable.

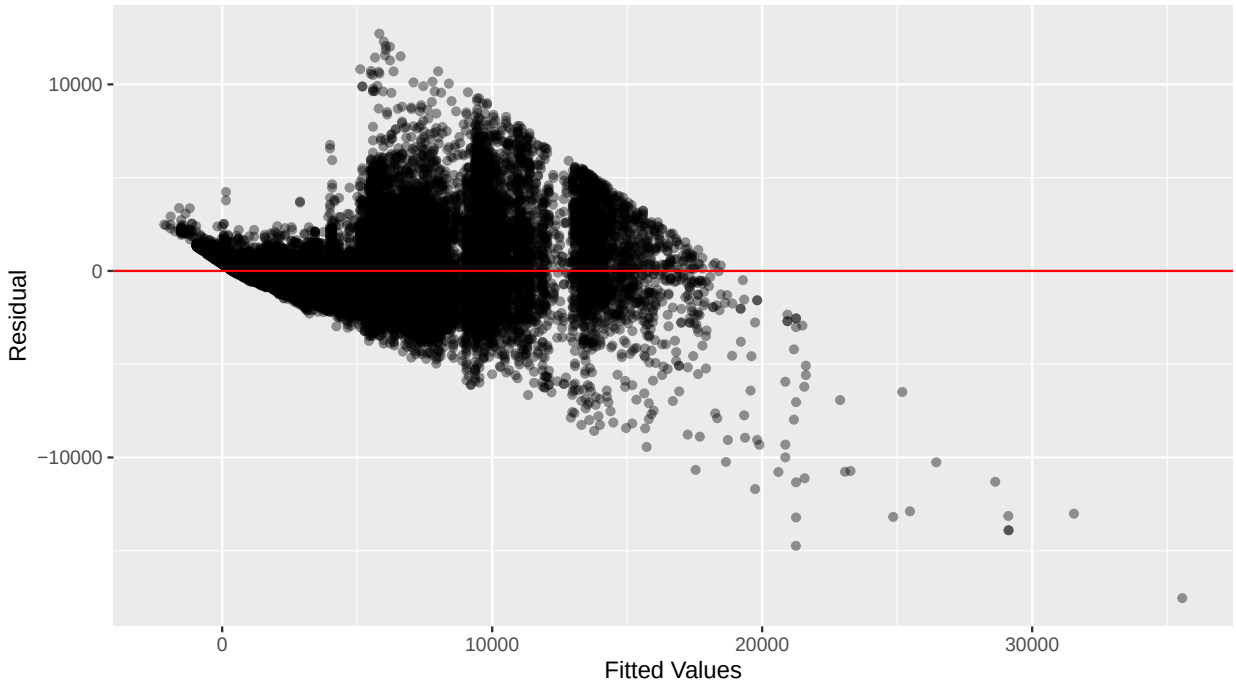

Figure 13: Figure 7.1. Scatterplots of residuals against fitted values.

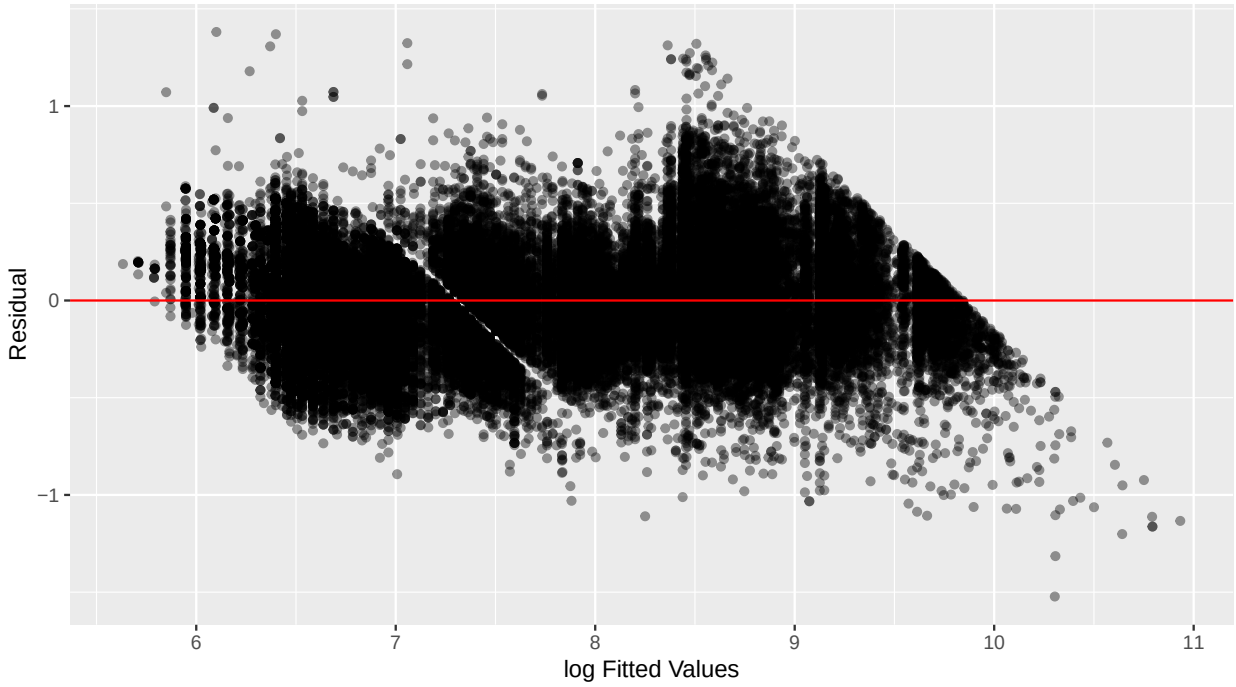

Figure 14: Figure 7.2. Scatterplots of residuals against log fitted values.

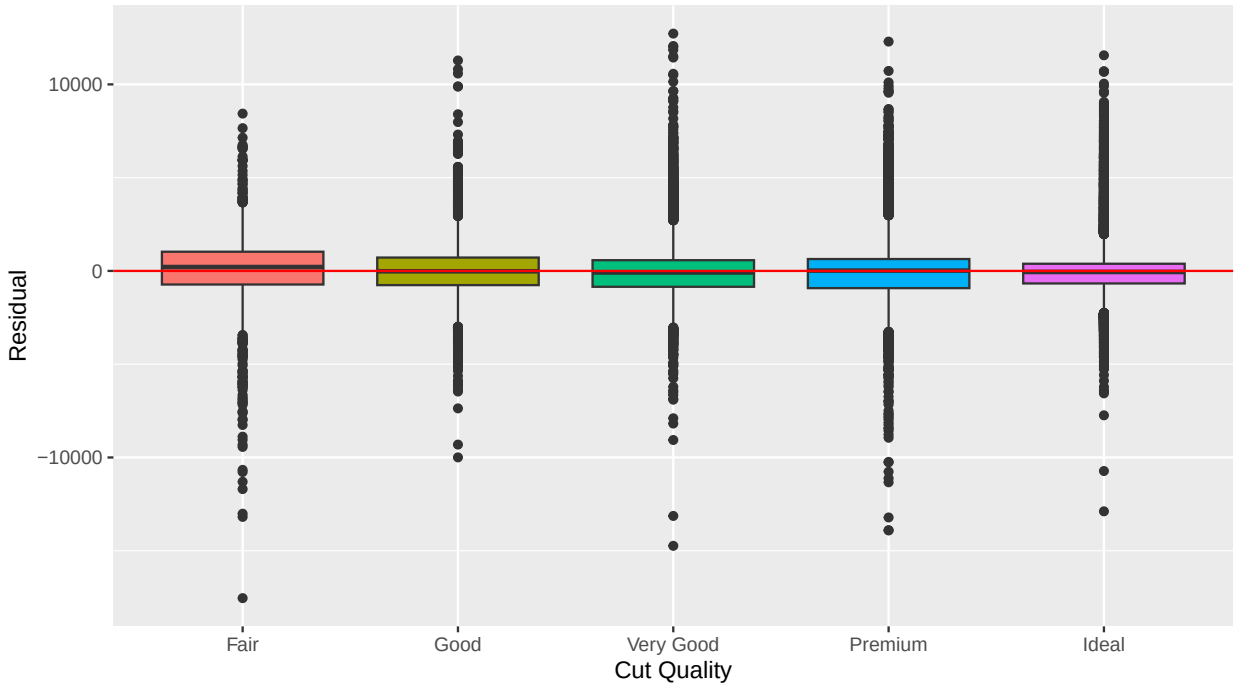

Figure 15: Figure 8.1. Boxplot of residuals for each cut quality.

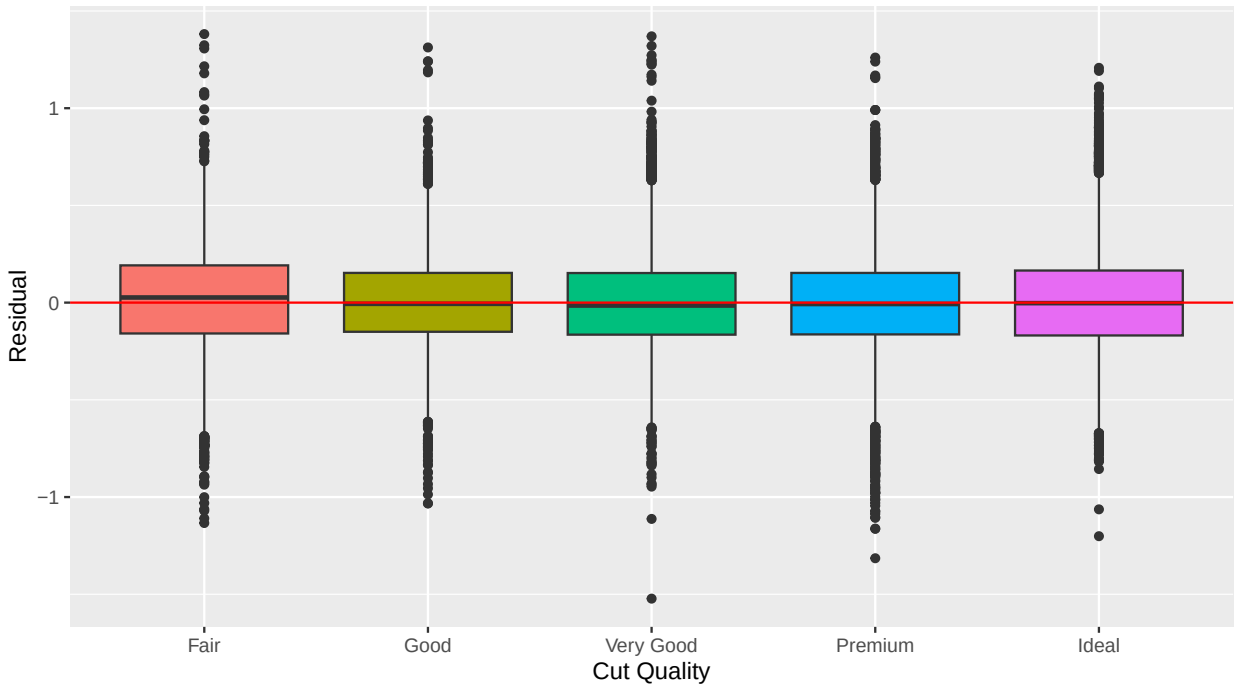

Figure 16: Figure 8.2. Boxplot of residuals for each cut quality, log-log model.

The boxplots show an even spread of residuals at each cut quality, and roughly similar values across the different cuts - with the Fair cut having slightly more positive residuals on average. However, there are several outliers in all cuts (Figs. 8.1 & 8.2).

We conclude that the assumptions for inference in multiple linear regression are not well met for the first model, mainly due to the violation of constant variance assumption and the linearity assumption as depicted by the patterns in Figs. 6.1 & 7.1.

However, we might consider the option of using the log-log model as there are no major violations to the linear regression conditions. Nevertheless, there are extreme outliers that require further investigation to see if they affect the conclusions.

---

4. Discussion

4.1 Conclusions

We see from the data that there is a notable difference in the price of different diamonds for every cut. As the cut quality gets better, the price is expected, on average, to rise. However, Premium cut did not quite follow this trend, we might want to investigate why this was the case.

Moreover, we can see that as the Weights (carat) of diamonds increase, the prices (USD) increase significantly. It is expected that for every one unit increase in the weight (carats) of a diamond, on average, the price (USD) of a diamond increases by $7871. This tells us that there is a positive relationship between the weights and the prices of the diamonds.

It is worth mentioning that these correlations do not necessarily imply causation, because this data was collected from a retrospective observational study.

For the most part, these results tell us that the price and cut quality of a diamond determine its worth and

value. Our findings are uniform with initial discoveries that the size or carat weight of a diamond and the cut quality are two of the 4 C's that affect their prices.

Diamonds are as expensive as they are because of their popularity as a gemstone. Due to the rareness of this gemstone, the weight of the diamond has a significant impact on its price. Additionally, the cut quality, as its balance and brilliance, play a major role in determining a diamond's price.

4.2 Limitations

A limitation to this data set is that the number of observations for high-tier cuts were extremely larger than the others (Table 1). The diamonds with an Ideal cut, Premium cut and Very good cut all have more observations compared to those with Fair cut and Good cut. That might have resulted in calculations that are heavily more dependent on the better cuts, β_{carat} in particular. Additionally, we can observe outliers in several plots which all have influenced our findings.

Given that, we might want to preform this study using a log-log model which will give estimates of how much percent would the price change given one percent increase in weight, for instance.

4.3 Further questions

To get a better understanding on what influences the price of diamonds we would like to work with a dataset that includes the other two variables from the 4 C's(Clarity and Color).

According to preceding studies, diamond clarity is the assessment of small imperfections on the surface and within the stone. Diamonds with the fewest and smallest inclusions receive the highest clarity grades, which results in high prices of diamond. Likewise, the color of diamonds has a huge impact on their value. A diamond's color affects how rare it is, so affects the price that you can sell the diamond for. We could also conduct a study to determine which of the 4 C's influence diamonds prices the most.

Apart from the 4 C's we can also test whether the prices of diamonds are different in the US and Canada if we have the country of each sold diamond.

---

5. Citations and References

- Shivam Agrawal. Diamonds data. Kaggle, 2016. https://www.kaggle.com/datasets/shivam2503/diamonds
- Blue Nile. The 4Cs Of Diamonds. https://www.bluenile.com/ca/education/diamonds?gclid=Cj0KCQiAnNacBhDvARIsABnDa6-rOgrV8C7l0Z8yJIdWF65ac8uT_mAVD_QK6HWOHBNh3N8p9RmAcX4aAgjXEA wcB&click_id=715353937&utm_source=google&utm_medium=text&utm_campaign=Google_%7C_P1_%7C_CA_%7C_English_%7C_Text_%7C_Non-Brand_%7C_ENG_%7C_Desktop_%7C_Engagement_%7C_Core_&utm_content=Diamonds_Education&utm_term=4cs
- Chicago Diamond Buyer. Diamond Colors Explained: Does the Color of My Diamond Affect Its Value?. February 2020. https://chicagodiamondbuyer.net/diamond-colors-does-the-color-of-my-diamond-affect-its-value/
- Blue Nile. Diamond Clarity Chart. https://www.bluenile.com/ca/education/diamonds/clarity?gclid=Cj0KCQiAnNacBhDvARIsABnDa6_7BezeJLsHj3aWWelN7iIDIDskZ6kk83MXG93VLcAo02h0lqZvntIaAkhXEALw_wcB&click_id=171182911&utm_source=google&utm_medium=text&utm_campaign=Google_%7C_P1_%7C_CA_%7C_English_%7C_Text_%7C_Non-Brand_%7C_ENG_%7C_Desktop_%7C_Engagement_%7C_Core_&utm_content=Diamond_Clarity&utm_term=diamond+clarity