Analysing Influence of 4 Cs in Diamond Price

For centuries, humans have been captivated by the majesty of diamonds, which are frequently referred to as "nature's marvels" (May, 2022). However, the valuation of these gemstones is what truly captivates gemmologists, sellers, and purchasers. The value of a diamond is not solely determined by its radiance, it is influenced by a variety of complex characteristics. This study aims to explore the relationship between a diamond's price and its features such as **carat (weight), clarity, colour, and cut** known as the "Four Cs.". Using a linear regression model, we will quantitatively assess how each of these variables influences the overall market value of a diamond. By treating carat as a continuous variable and clarity, colour, and cut as categorical factors, this statistical approach allows us to isolate and evaluate the impact of each attribute on price. Gemmologists, sellers, and consumers will all benefit from an understanding of their pricing mechanisms through data analysis.

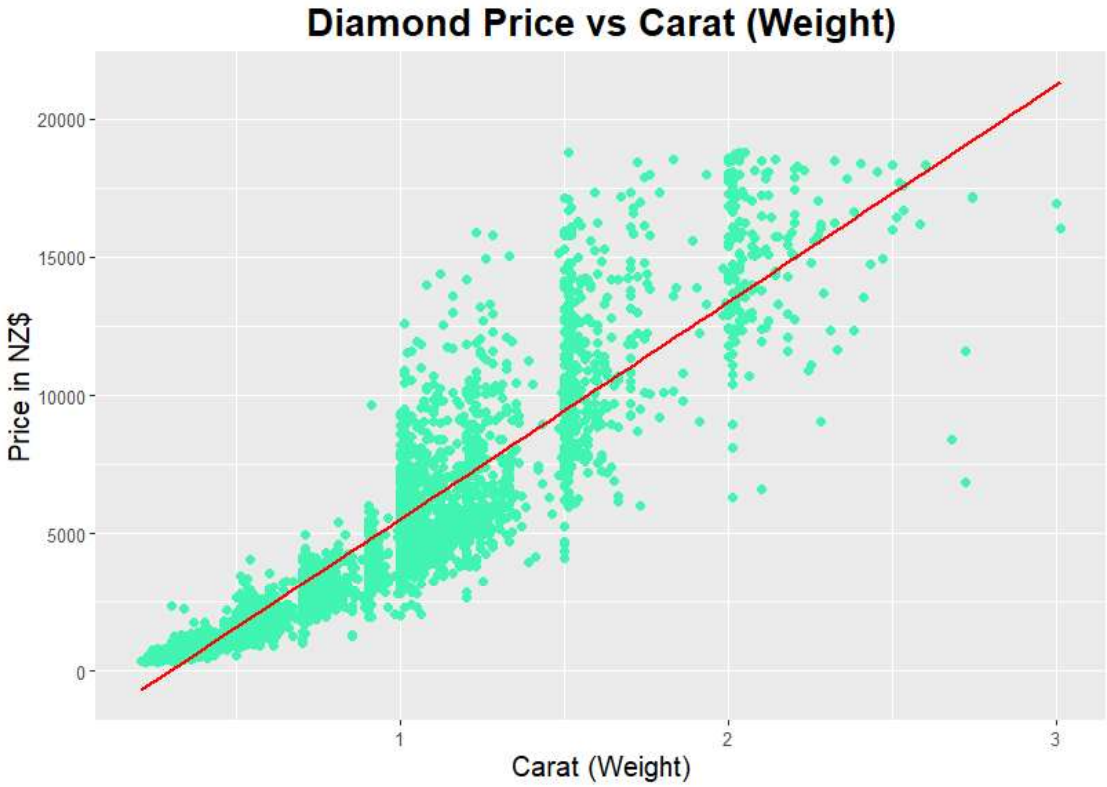
a) Correlation between the diamond **Price** and **Weight (Carat)**

In this case, we will analyse how much the weight of a diamond influences its price. Naturally, larger diamonds are more valuable, but how strong is this connection? We will find it out using correlation to measure how two variables move together.

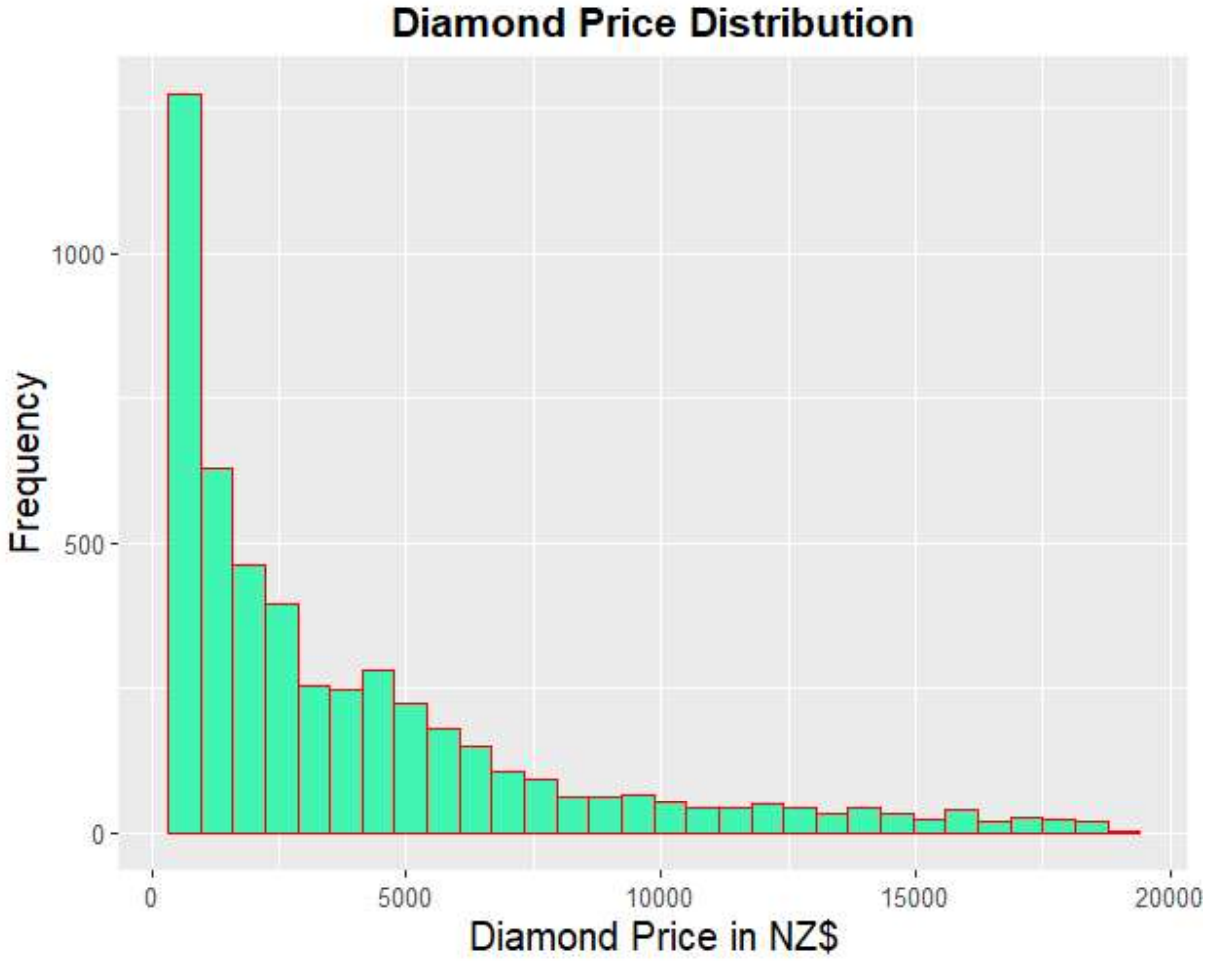
In our case, the correlation between diamond weight and price is 0.928, which is very close to +1. To determine whether the correlation we see is meaningful, we use a p-value. In this case, the p-value is less than 2.2e-16, which tells us that the correlation between price and carat is statistically significant.

Additionally, the confidence interval for the correlation is between 0.924 and 0.932, meaning we are 95% confident that the true correlation lies within this narrow range. This further solidifies the strength of the relationship between carat and price.

This strong correlation provides a powerful predictor for diamond pricing. It's the factor that explains most of the price variation, so any pricing decisions should give it significant weight.



b) Creating a histogram of the prices of the diamonds



We can also create a histogram to visualize the values in the dataset. We will explain how to explore, summarize, and visualize the diamonds dataset in R. (Bobbitt, 2022)

The histogram titled "Diamond Price Distribution" provides a clear visualization of the frequency distribution of diamond prices in NZ dollars. The x-axis represents the price of diamonds in NZ$, while the y-axis shows the frequency.

Skewness of the Distribution

The histogram is right-skewed (positively skewed). This means that most diamonds are priced at the lower end of the spectrum, while fewer diamonds have higher prices.

Low-Priced Diamonds

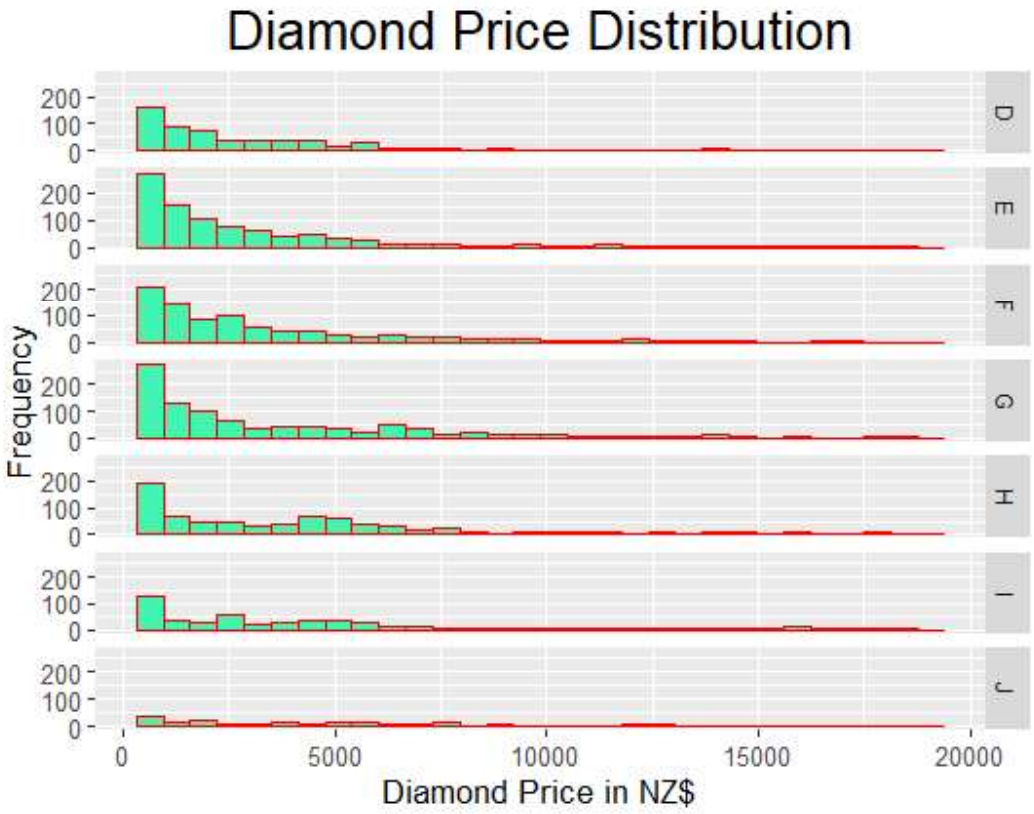
There is a high frequency of diamonds priced below NZ$ 2,000, which is evident from the tall bars on the left side of the histogram. This indicates that lower-priced diamonds dominate the market. These lower-priced diamonds are likely smaller in size or have lower quality in terms of cut, clarity, or colour, making them more affordable to consumers.

High-Priced Diamonds

As prices increase, the frequency drops significantly. This reflects the fact that larger, high-quality diamonds are much rarer. The distribution tail extends to prices nearing NZ$ 20,000, but these high-priced diamonds are quite rare.

Overall, the distribution suggests that most diamonds are more affordable, with a small percentage of premium-priced diamonds. The skewness in the data highlights that the market for diamonds is heavily concentrated at lower price points, likely due to the relative scarcity of large, high-quality diamonds.

v) Creating a facet chart based on the colour of the diamonds



The above facet chart offers a detailed look at of diamond price distribution, segmented by colour grades ranging from D (colourless) to J (noticeable tint). Each horizontal bar represents a specific colour grade. The x-axis indicates the price of diamonds in NZ$, while the y-axis represents the frequency.

Right-Skewed Distribution

In every colour category (D to J), the price distribution is right-skewed, meaning that most diamonds fall within the lower price range (below NZ$ 5,000). The frequency suggests that lower-priced diamonds dominate the market, while higher-priced diamonds are rarer.

Highest Quality (D and E Grades)

The long tail of diamonds priced above NZ$ 10,000 is most prominent in D and E colour grades, indicating that colourlessness is highly valued in diamond pricing. The colour is a major factor for customers seeking the best quality.

Mid-Range Colour Quality (F, G, H Grades)

F, G and H diamonds fall mainly in the mid-range price. H diamonds show a notable long tail, indicating that some still reach higher prices, potentially due to factors like size or cut.

Lower Quality

I and J colour diamonds are the least expensive, with most diamonds concentrated in the NZ$ 0 to NZ$ 2,000 price range. However, I diamond also exhibit a long tail, with a few reaching higher prices. For some buyers, other factors like carat weight or exceptional cuts can still drive-up prices, even for lower colour grades.

The visualization clearly shows how colour influences diamond pricing. The right-skewed nature across all categories highlights the predominance of lower-priced diamonds in the market. D and E diamonds, being the most colourless, as a premium, while F-H diamonds occupy the mid-range. H and I diamonds, although they have lower in colour quality, can still reach higher prices, but less frequently, due to factors like size and cut.

(i) How the linear regression model uses categorical data in model creation

In the study of diamond pricing, understanding how linear regression models incorporate categorical data is crucial and this study aims to explore this is utilized in linear regression and the implications for price predictions.

To incorporate categorical variables like clarity, colour, and cut, they are first converted into dummy variables.

lm(formula = price ~ carat + clarity + colour + cut, data = Dia)

Results and Interpretation

R-squared = 0.9209, this means the model explains approximately 92% of the variance in diamond prices, meaning the model captures most of the factors that affect diamond prices This is a high value, the model is good at predicting price based on carat, clarity, colour, and cut.

Adjusted R-squared = 0.9206: The adjusted R-squared is nearly the same as the R-squared, meaning that additional variables (such as the different levels of clarity, colour, and cut) indeed contributing useful information and has not led to overfitting.

The p-value for all coefficients is highly significant the p-value < 2.2e-16, meaning that all variables (carat, clarity, colour, and cut) have a significant effect on the price of diamonds, no chance that the result occurred randomly. The F-statistic is also extremely high, indicating that the overall model is statistically significant.

Coefficient:

a) Carat: For each additional carat in weight, the price of a diamond increases by approximately $8,932. This high value shows that carat has a significant impact on price.
b) Clarity: Different clarity levels have a varying impact on price. For instance, diamonds with clarityIF add approximately $4,767 to the price.
c) Colour: Diamonds with colourE decrease in price by approximately $145 compared to the baseline colour. And colourJ reduce the price by $2,387. As a result, colourE diamonds are NZ$ 2,241.53 more expensive than colourJ diamonds
d) Cut: Diamonds with Ideal cuts increase in price by about $926 which indicates that a better cut increases the price significantly.

In this linear regression, categorical variables are transformed into dummy variables, allowing the model to assess how each category influences the dependent variable (price). It indicates the strong fit of the model and the model can effectively handle non-numeric data, making it a powerful tool for predicting diamond prices.

ii) Effects on price excluding Carat

```
Call:
lm(formula = price ~ carat + clarity + colour + cut, data = dia)

Residuals:
    Min      1Q  Median      3Q     Max
-8894.8  -678.1  -191.9   454.4  7071.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6968.43     169.60 -41.088  < 2e-16 ***
carat         8932.20      38.70 230.831  < 2e-16 ***
clarityIF     4767.44     167.01  28.546  < 2e-16 ***
claritySI1    3167.56     143.91  22.011  < 2e-16 ***
claritySI2    2218.76     144.89  15.313  < 2e-16 ***
clarityVS1    4155.91     146.83  28.305  < 2e-16 ***
clarityVS2    3811.46     144.82  26.319  < 2e-16 ***
clarityVVS1   4667.13     154.68  30.172  < 2e-16 ***
clarityVVS2   4430.33     151.27  29.288  < 2e-16 ***
colourE       -145.61      59.33  -2.454   0.0141 *
colourF       -307.58      60.02  -5.125 3.09e-07 ***
colourG       -442.11      59.29  -7.457 1.04e-13 ***
colourH       -935.05      62.28 -15.014  < 2e-16 ***
colourI      -1341.03      68.98 -19.439  < 2e-16 ***
colourJ      -2387.33      92.49 -25.812  < 2e-16 ***
cutGood        611.84     108.58   5.635 1.85e-08 ***
cutIdeal       926.02      99.27   9.329  < 2e-16 ***
cutPremium     786.05     100.05   7.856 4.81e-15 ***
cutVery Good   826.51     101.07   8.178 3.63e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1133 on 4981 degrees of freedom
Multiple R-squared:  0.9209,    Adjusted R-squared:  0.9206
F-statistic:  3221 on 18 and 4981 DF,  p-value: < 2.2e-16

>
```

Carat: For every 1 unit increase in carat, the diamond's price increases by $8932.20.

The model shows that for every one-unit increase in carat, there is a **significant jump in price**. This aligns with what we know that diamonds are priced exponentially by weight, but size isn’t everything, and here’s where other factors come into play.

Greatest Effect on Price

Moving beyond carat, you may find that clarity, the number of imperfections in a diamond is the next big driver of price. Diamonds with fewer imperfections, like those graded as "IF" (internally flawless), command much higher prices than those with visible inclusions. Clarity has the largest coefficients, especially for ClarityIF (4967.44), ClarityVVS1 (3515.96), and ClarityVS1 (2825.21). Additionally, all the p-values for clarity categories are highly significant ($p < 0.001$). This suggests that clarity has the greatest effect on price after carat. Specifically, the model reveals that a top-grade clarity like ClarityIF adds nearly $5000 to the price of a diamond. That’s a significant leap compared to diamonds with lower clarity grades.

Least Effect on Price

Color also plays a role in determining diamond price, though to a lesser degree than carat, clarity, or cut. Diamonds that are closer to colourless tend to be more valuable. For instance, ColourJ has a negative coefficient (-1371.63), are expected to decrease the price by about $1371.63. Other colours (such as ColourI with -781.05 or ColourH with -495.47) also show negative coefficients, but the value of ColorJ is the largest in terms of reducing price.

The Bigger Picture

These numbers show that diamond pricing is a careful balance of multiple factors. While Carat weight is the headline feature, clarity and colour play crucial supporting roles.

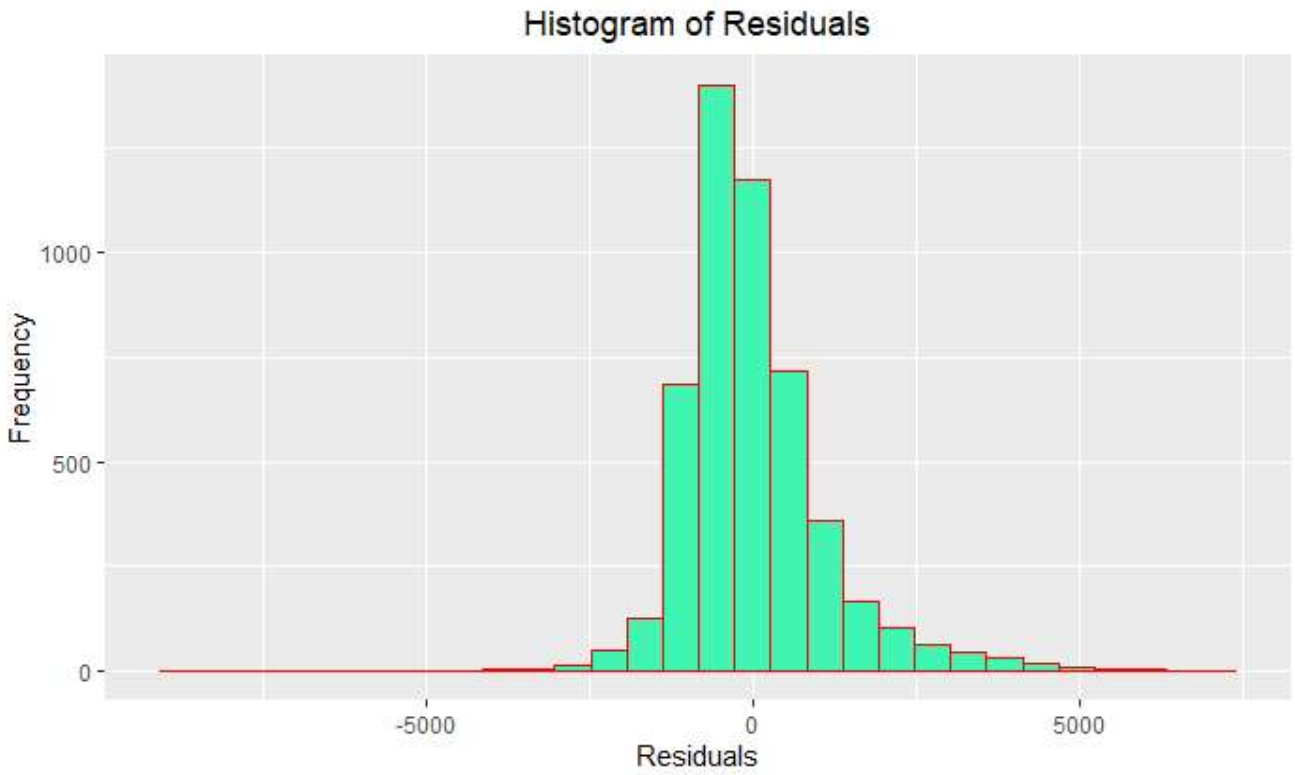
iii) Residual Analysis

We have established relationship between diamond prices and their characteristics by using linear regression model. We need to verify if the underlying assumptions hold true, or if there are areas where the model may fall short.

The Shapiro-Wilk Test

We began by conducting the Shapiro-Wilk normality test to determine the difference between the actual diamond prices and the prices predicted by our model.

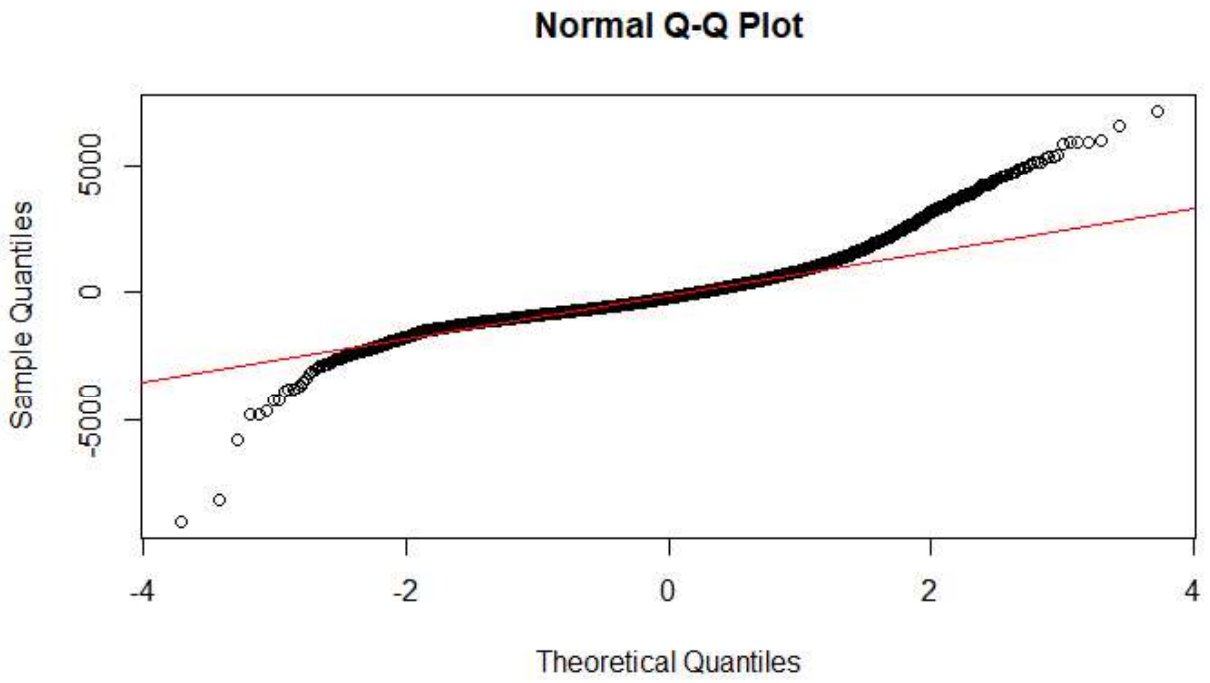
We obtained a W-statistic of 0.90098, indicating some degree of normality, but still far from the ideal value of 1. More importantly, the p-value was less than 2.2e-16, an incredibly small value. This p-value allowed us to reject the null hypothesis that our residuals were normally distributed. We will explore further and visualize these deviations.



The Story Told by the Histogram

Next, we created a histogram to visualize the distribution of the residuals. The histogram show peak near zero, which is a good sign in regression model. The distribution appears to be symmetrical and there are some longer tails, particularly on the left and right ends of the histogram these indicate that there are certain predictions where the model has either over- or under-predicted the diamond prices by a substantial margin.

Q-Q Plot: The Final Confirmation



To confirm these findings, we turned to a **Q-Q plot (Quantile-Quantile plot)**.
For most of the plot, the points follow the red line quite well, particularly in the middle range of the data. This suggests the model is performing well but in the left tail, we see that the points start to dip below the line, indicating that the model underestimates prices for some diamonds. On the other side, in the right tail, the points curve above the line, meaning the model overestimates prices for higher-end diamonds.

Overall, the model performs well for the majority of diamonds but struggles with extreme cases. The residuals deviate from normality, which affects the model's reliability in these extremes.

References:

Bobbitt, Z. (2022, April 13). *A Complete Guide to the diamonds Dataset in R*. Statology. https://www.statology.org/diamonds-dataset-r/

Andrew May. (2022, January 18). Diamonds: Formation, grading and other facts Livescience.com. https://www.livescience.com/diamonds-facts