Diamond Price Analysis: A Comprehensive Study Using Multiple Linear Regression

Anju Sambasivan

1. Introduction

This project is about analyzing the price and features of diamonds using data from the `dia.csv` file, which contains information on 5000 diamonds. The data includes details like weight (carat), cut quality, color, clarity, and price. The main goal is to understand how these features affect the price of a diamond.

The project involves:

1. Checking how the weight (carat) and price are related.
2. Creating visual charts, like a histogram and facet chart, to show the distribution of diamond prices and differences based on color.
3. Building a regression model to see how carat, clarity, color, and cut together influence the price.

We'll also identify which of the features (besides carat) has the biggest and smallest impact on price. Lastly, we'll check if the regression model works well by analyzing its residuals and creating visualizations.

```
# Load required library

library(ggplot2)
library(knitr)
```

```
# Load the dataset

dia <- read.csv("dia.csv")
```

2. Handling Categorical Variables in Regression Analysis

Categorical variables are variables that group data into categories instead of using numbers. In the diamond dataset, the variables `cut`, `colour`, and `clarity` are examples of categorical variables. These variables describe the quality of a diamond in terms of its cut, color grade, and clarity, which are essential for understanding what influences a diamond's price.

- **Cut**: Groups diamonds based on how well they are cut. Categories include `Fair`, `Good`, `Very Good`, `Premium`, and `Ideal`.
- **Colour**: Grades diamonds on how colorless they are, ranging from `D` (best, completely colorless) to `J` (worst, noticeable tint).
- **Clarity**: Ranks diamonds based on how many inclusions (imperfections) they have, from `FL` (Flawless) to `I3` (the most included).

While these categories are important, regression models cannot use text or categories directly because they require numbers to perform calculations. Therefore, we need to convert these categorical variables into a numeric format before including them in the model. This is done using a method called **dummy encoding**.

Dummy Encoding

Dummy encoding is a way to convert categories into numbers by creating new variables for each category.

1. **Assign Binary Values**: Each category of a variable is turned into a binary variable that takes the value `1` if the observation belongs to that category and `0` otherwise.
2. **Choose a Reference Level**: One of the categories is chosen as the baseline (called the **reference level**) and is not explicitly given a variable. Instead, the model compares all other categories to this reference level.

For example, for the `cut` variable:

- Categories: `Fair`, `Good`, `Very Good`, `Premium`, `Ideal`
- If `Fair` is chosen as the reference level, the model creates four new variables:

  `CutGood`: 1 if the diamond's cut is `Good`, 0 otherwise.

  `CutVeryGood`: 1 if the diamond's cut is `Very Good`, 0 otherwise.

  `CutPremium`: 1 if the diamond's cut is `Premium`, 0 otherwise.

  CutIdeal: `1 if the diamond's cut is`Ideal‘, 0 otherwise.

If a diamond has a `Good` cut, `CutGood` will be 1, and the other variables (`CutVeryGood`, `CutPremium`, `CutIdeal`) will be 0.

The regression model uses these dummy variables to calculate how much the price changes when the category changes. For example, if the model shows that the coefficient for `Good` is 200, it means diamonds with a `Good` cut are $200 more expensive than those with a `Fair` cut. Similarly, coefficients for other categories show how much more (or less) those categories affect the price compared to the reference level.

Some categorical variables, like `cut`, `colour`, and `clarity`, have a natural order. These are called **ordinal variables** because their categories follow a hierarchy. For example, in `cut`, `Fair` is the lowest quality, followed by `Good`, `Very Good`, `Premium`, and `Ideal`. In R, we use the `factor()` function to tell the model this order, ensuring the analysis respects the progression from lower to higher quality.

By properly encoding these variables and including them in the regression model, we can understand how much each quality level affects the price of a diamond. For example, higher-quality levels like `Ideal` cut, better colour grades like `D`, and clearer diamonds like `FL` are expected to increase the price significantly compared to lower-quality levels. This process ensures that the regression analysis provides accurate, meaningful results and helps us identify the most important factors driving a diamond's value.

```r
# Step 1: Correct Factor Levels for Categorical Variables

dia$cut <- factor(dia$cut, levels = c("Fair", "Good", "Very Good",
                                      "Premium", "Ideal"))
dia$colour <- factor(dia$colour, levels = c("D", "E", "F", "G",
                                            "H", "I", "J"))
dia$clarity <- factor(dia$clarity, levels = c("FL", "IF", "VVS1", "VVS2",
                                              "VS1", "VS2", "SI1", "SI2",
                                              "I1", "I2", "I3"))
```

```r
# Structure of data
str(dia)
```

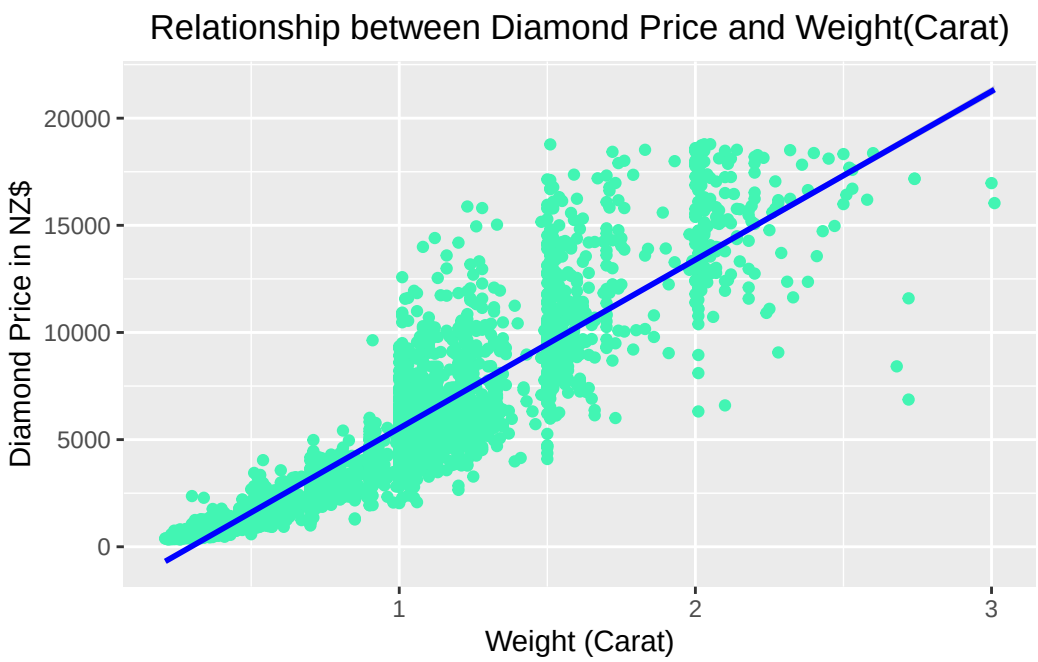
```
'data.frame':   5000 obs. of  11 variables:
 $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ carat  : num  0.59 0.3 0.42 0.95 0.32 0.52 1.04 0.5 0.72 0.24 ...
 $ cut    : Factor w/ 5 levels "Fair","Good",..: 3 2 4 5 4 4 5 4 5 2 ...
 $ colour : Factor w/ 7 levels "D","E","F","G",..: 5 6 3 5 1 2 5 2 3 3 ...
 $ clarity: Factor w/ 11 levels "FL","IF","VVS1",..: 4 5 2 7 3 6 7 6 7 3 ...
 $ depth  : num  61.1 63.3 62.2 61.9 62 60.7 62.3 62.1 62 64.8 ...
 $ table  : num  57 59 56 56 60 58 57 62 55 57 ...
```

```
 $ price   : int  1771 473 1389 4958 973 1689 5102 1559 2737 492 ...
 $ x       : num  5.39 4.2 4.85 6.31 4.4 5.17 6.45 5.1 5.76 3.9 ...
 $ y       : num  5.48 4.23 4.8 6.35 4.37 5.21 6.48 5.08 5.79 3.94 ...
 $ z       : num  3.32 2.67 3 3.92 2.72 3.15 4.03 3.16 3.58 2.54 ...
```

3. Relationship between Diamond Price and Weight (Carat)

```r
# Create the scatter plot

ggplot(dia, aes(x = carat, y = price)) +
  geom_point(color = "#42f5b3") +
  labs(title = "Relationship between Diamond Price and Weight(Carat)",
       x = "Weight (Carat)",
       y = "Diamond Price in NZ$") +
  # Add linear regression line
  geom_smooth(method = "lm", color = "blue") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
  ggsave("RelationshipBtnPrice_Carat.png", width = 11, height = 5)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

This graph shows how the weight of a diamond (measured in carats) relates to its price (in NZ dollars). Each black dot represents a single diamond, with the horizontal position showing its weight and the vertical position showing its price. The blue line in the middle, called a trendline, highlights the general pattern: as the weight of a diamond increases, its price also increases. This means that heavier diamonds are generally more expensive.

However, the graph also shows that diamonds of the same weight can have very different prices. For example, at 1 carat, some diamonds are priced much higher or lower than others. This spread suggests that factors other than weight, like the diamond's clarity, cut, or color, also play an important role in determining the price. These qualities can make some diamonds more valuable even if they weigh the same as others.

Another interesting observation is that the price increase is not linear. The blue trendline becomes steeper for heavier diamonds, meaning that the price rises faster as the diamond's weight increases. Larger diamonds are much rarer, so their value grows disproportionately compared to smaller ones.

Finally, the graph shows that most diamonds in the dataset are smaller (under 1 carat), as indicated by the dense cluster of points in the lower-left corner. Diamonds weighing more than 2 carats are much less common, reflecting their rarity. In summary, the graph highlights that while weight has a strong influence on price, other characteristics of the diamond also contribute to its value.

4. Correlation Between Price and Carat

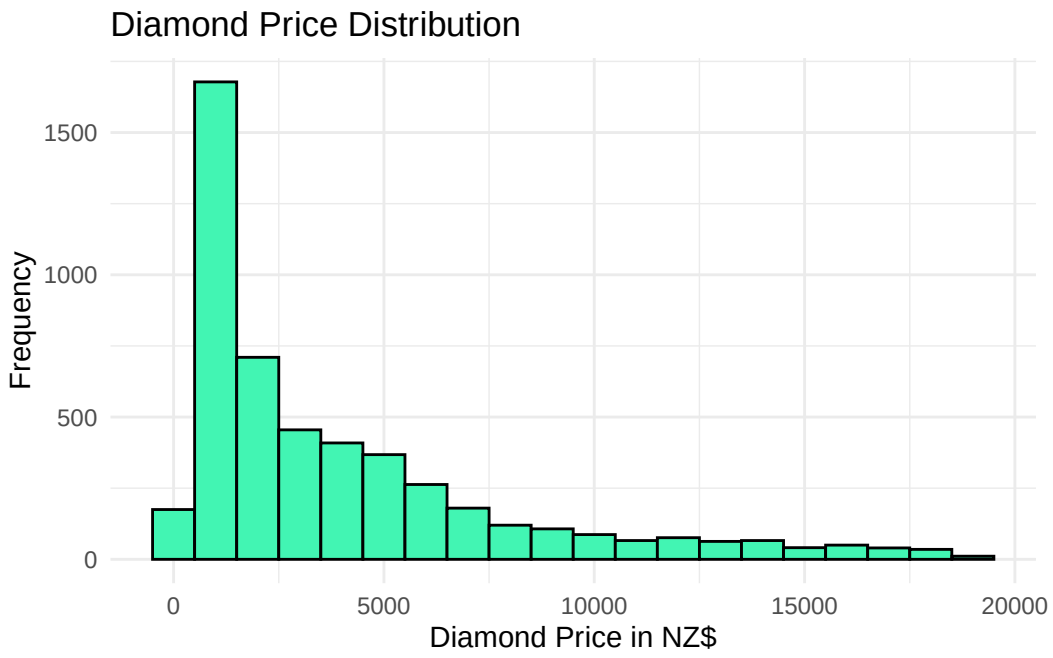
```
# Step 2: Correlation Between Price and Carat

correlation <- cor(dia$price, dia$carat)
cat("Correlation between price and carat:", correlation, "\n")
```

```
Correlation between price and carat: 0.9281015
```

The correlation between price and carat is **0.93**, which indicates a **very strong positive relationship** between these two variables. This means that as the weight of a diamond (measured in carats) increases, its price also increases significantly. The high value (close to 1) shows that carat weight is one of the most important factors influencing diamond price, with heavier diamonds being much more expensive on average.

5. Distribution of Diamond Prices

```
# Step 3: Histogram of Diamond Prices
ggplot(dia, aes(x = price)) +
  geom_histogram(fill = "#42f5b3", binwidth = 1000, colour = "black") +
  labs(
    title = "Diamond Price Distribution",
    x = "Diamond Price in NZ$",
    y = "Frequency"
  ) +
  theme_minimal()
```



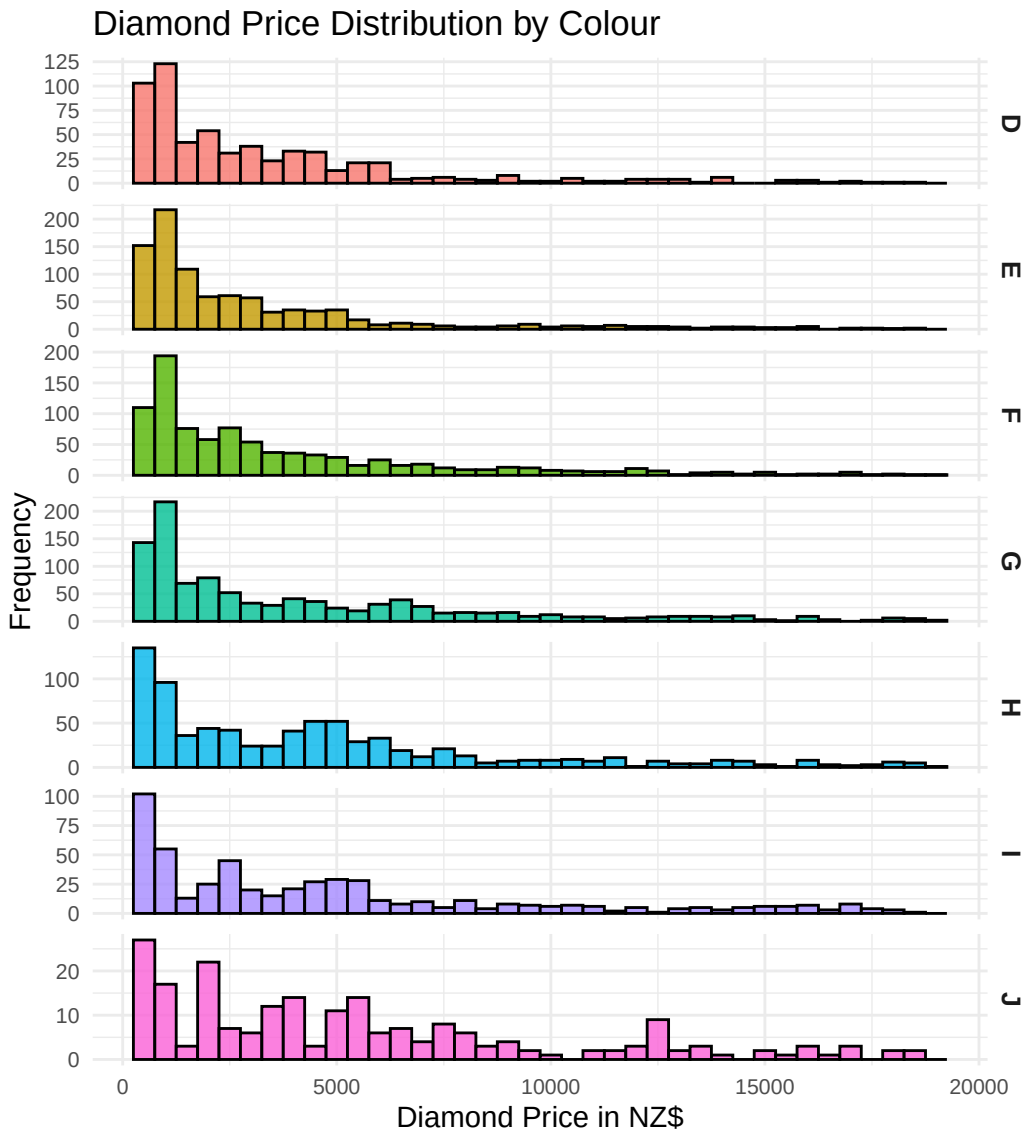
This graph shows the distribution of diamond prices in NZ dollars. The x-axis represents the price range of diamonds, while the y-axis shows the number of diamonds (frequency) within each price range. Each bar covers a price range of $1,000, and the height of the bar indicates how many diamonds fall within that range.

The graph reveals that most diamonds are in the lower price ranges, with the tallest bar representing diamonds priced between $0 and $1,000. As the price increases, the number of diamonds decreases steadily, indicating that higher-priced diamonds are less common. Beyond $10,000, the bars are much shorter, showing that very few diamonds are priced in this range. The graph also has a long tail on the right side, extending to nearly $20,000, which reflects the presence of a few extremely expensive diamonds.

In summary, the graph highlights that diamonds are most commonly priced under $5,000, with very few being exceptionally expensive. This pattern suggests that more affordable diamonds dominate the market, while high-priced diamonds are rare and likely targeted at a smaller, niche market.

6. Distribution of Diamond Prices by Colour

```
# Step 4: Create the facet chart

ggplot(dia, aes(x = price, fill = colour)) +
  geom_histogram(binwidth = 500, color = "black", alpha = 0.8) +
  facet_grid(colour ~ ., scales = "free") +
  labs(
    title = "Diamond Price Distribution by Colour",
    x = "Diamond Price in NZ$",
    y = "Frequency"
  ) +
  theme_minimal() +
  theme(
    strip.text.y = element_text(size = 10, face = "bold"),
    axis.text.x = element_text(size = 8),
    axis.text.y = element_text(size = 8),
    legend.position = "none"
  )
```



This faceted graph shows the **distribution of diamond prices (in NZ dollars)** grouped by their **color grade** (D to J). Each panel represents a specific color grade, and the bars show the number of diamonds (frequency) in different price ranges. The x-axis represents the price in NZ$, and the y-axis represents the frequency of diamonds in each price range.

For every color grade, the majority of diamonds are priced below $5,000. The tallest bars are always on the left side of each panel, showing that most diamonds in the dataset are affordable. As the price increases (above $10,000), the frequency of diamonds decreases sharply for all color grades. This trend is consistent across the different grades, indicating that high-priced diamonds are rare regardless of color. Diamonds with higher color grades (like **D**, **E**, and **F**)

have more diamonds in the higher price ranges (above $10,000) compared to lower grades like **I** or **J**.

For lower-quality grades (**I** and **J**), diamonds tend to cluster in lower price ranges, with very few reaching prices above $10,000. Some color grades, like **E** and **G**, have a higher overall frequency of diamonds compared to others like **D** or **J**. This could indicate that diamonds of certain color grades are more commonly found in the dataset or are more popular.

7. Regression Model

```r
# Step 5: Linear Regression Model

model <- lm(price ~ carat + clarity + colour + cut, data = dia)
summary(model)
```

```
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)  -2200.99     139.87 -15.736  < 2e-16 ***
carat         8932.20      38.70 230.831  < 2e-16 ***
clarityVVS1   -100.31     104.88  -0.956 0.338907
clarityVVS2   -337.12     101.30  -3.328 0.000882 ***
clarityVS1    -611.53      96.41  -6.343 2.45e-10 ***
clarityVS2    -955.99      94.05 -10.165  < 2e-16 ***
claritySI1   -1599.88      94.48 -16.933  < 2e-16 ***
claritySI2   -2548.68      98.75 -25.811  < 2e-16 ***
clarityI1    -4767.44     167.01 -28.546  < 2e-16 ***
colourE       -145.61      59.33  -2.454 0.014150 *
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 ***
```

```
cutVery Good   826.51     101.07   8.178 3.63e-16 ***
cutPremium     786.05     100.05   7.856 4.81e-15 ***
cutIdeal       926.02      99.27   9.329  < 2e-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
```

The regression model output explains how different factors influence diamond prices. The model predicts the price of diamonds based on four key attributes: **carat (weight)**, **clarity**, **color**, and **cut**. Each factor's impact on price is represented by coefficients, which indicate how much the price changes when that factor changes, while keeping all other factors constant.

Intercept:

The intercept value (-2200.99) is the baseline price of a diamond when all factors are at their reference levels:

- **Cut**: Fair (lowest quality).
- **Color**: D (highest color grade).
- **Clarity**: FL (Flawless).
- **Carat**: 0 (though diamonds can't actually have zero weight). While the intercept itself doesn't represent a realistic diamond, it serves as the starting point to which other effects (e.g., weight, cut, color) are added or subtracted.

Carat (Weight):

Carat is the most significant factor affecting diamond price. For every additional carat, the price increases by approximately **$8932.20**. This relationship is highly significant, as indicated by its p-value ($<$ 2e-16). This means carat has a very strong and reliable impact on the price, and larger diamonds are much more expensive.

Clarity:

Clarity measures the presence of internal flaws or inclusions in a diamond. The model uses the highest clarity grade (FL) as the reference. Diamonds with lower clarity grades (e.g., SI2, I1) tend to have lower prices: For example, diamonds graded **SI2** are, on average, **$2548.68** less expensive than FL diamonds.

Most clarity levels are statistically significant (p-value < 0.05), meaning they have a clear impact on price. However, some grades, such as **VVS1** ($p = 0.34$), are not significant, suggesting they don't strongly influence price in this dataset.

Color:

Color reflects how colorless a diamond is. The reference level is the best color grade, **D**. Diamonds with lower grades (e.g., J) tend to be less expensive: For example, a diamond with color grade **J** is, on average, **$2387.33** cheaper than a diamond with color grade **D**. The p-values for most color grades are highly significant, meaning color is an important factor in determining price.

Cut:

Cut refers to the quality of a diamond's shape and symmetry. The reference level is the lowest cut grade, **Fair**. Diamonds with higher cut grades, such as **Ideal**, have higher prices: For instance, Diamonds with an Ideal cut are **NZ$926.02 more expensive** than Fair-cut diamonds. All cut levels are statistically significant, showing that cut quality is a critical factor in pricing.

Model Performance:

The model performs very well:

1. **R-squared (0.9209)**: This means that 92.09% of the variation in diamond prices is explained by the factors in the model. This is an excellent result, indicating the model is highly reliable.
2. **Residual Standard Error (1133)**: On average, the model's predictions deviate by about **$1133** from the actual prices. While there is some error, it's relatively small compared to the scale of diamond prices.
3. **F-statistic (3221, p $<$ 2e-16)**: The model as a whole is statistically significant, meaning that the combination of factors (carat, clarity, color, and cut) strongly influences diamond prices.

Clarity has the strongest effect on diamond price after carat weight. Compared to the best clarity grade, **FL (Flawless)**, lower clarity grades cause large price drops. For example, diamonds with clarity **SI2** are about **$2548.68 cheaper**, and those with clarity **I1** are **$4767.44 cheaper** than flawless diamonds. This shows that consumers value diamonds with fewer flaws, and a lower clarity grade significantly reduces the price. Clarity levels are also highly significant statistically ($p < 0.001$ for most grades), confirming their strong impact on price.

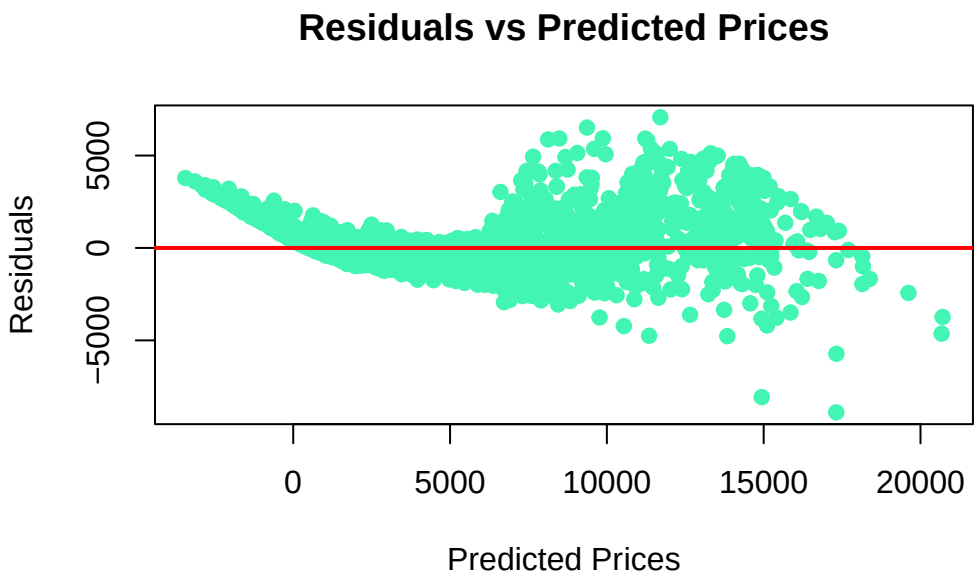
On the other hand, cut has the smallest effect on price. Diamonds with a better cut, like **Ideal** or **Good**, are slightly more expensive compared to the baseline **Fair** cut. For instance, an **Ideal cut** increases the price by about **$926.02**, and a **Good cut** adds around **$611.84**. These price increases are much smaller compared to the price drops caused by lower clarity grades. Even though cut is statistically significant, it has a smaller influence on price compared to clarity and color.

Color also affects price but less than clarity. For example, diamonds with color **J** cost about **$2387.33 less** than those with the best color grade, **D**. While this is significant, it still has a smaller effect than the price drop caused by lower clarity grades like **I1** or **SI2**.

In summary, clarity has the greatest impact on price, with lower grades causing big price drops, such as **$4767.44 less** for clarity **I1**. Cut has the least effect, with only small price increases, like **$926.02 more** for an **Ideal cut**. These findings are based on the size of the effects and their statistical importance in the model.

8. Residual Analysis

```
# Extract the residuals and fitted (predicted) values

residuals <- residuals(model)
fitted_values <- fitted(model)

# Step 6: Create the residual plot

plot(fitted_values, residuals,
     xlab = "Predicted Prices",
     ylab = "Residuals",
     main = "Residuals vs Predicted Prices",
     col = "#42f5b3",
     pch = 19)

# Add a horizontal reference line at y = 0

abline(h = 0, col = "red", lwd = 2)
```



This is a **Residuals vs. Predicted Prices** plot, which is a common tool for evaluating the performance of a regression model and checking if its assumptions are satisfied. The **x-axis** represents the predicted diamond prices generated by the regression model, while the **y-axis** shows the residuals, which are the differences between the actual diamond prices and the predicted prices. Residuals are calculated as:

Residual = Actual Price − Predicted Price.

Each **blue dot** represents a single diamond. The position of each dot shows how far the predicted price is from the actual price. The **red horizontal line** indicates a residual of zero. If a prediction perfectly matches the actual price, the residual will be zero, and the dot will lie on this line.

Observations from the Graph:

The residuals show a **curved pattern**, suggesting that the model does not fully capture the relationship between the predictors (e.g., carat, clarity, color, cut) and diamond prices. For diamonds with **lower predicted prices**, the residuals are mostly negative, meaning the model tends to **overestimate prices** in this range. For **mid-range predicted prices**, the residuals cluster around zero, indicating that the model performs reasonably well here. For **higher predicted prices**, the residuals are mostly positive, showing that the model tends to **underestimate prices** for expensive diamonds.

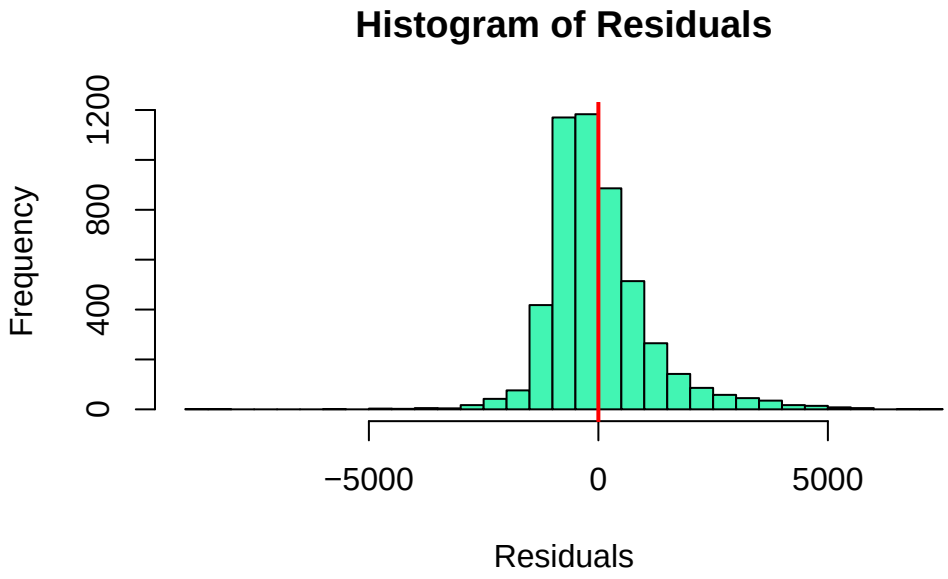
The graph also shows **heteroscedasticity**, which means that the spread of residuals increases as the predicted prices increase. This suggests that the model's errors are not consistent across price ranges. Specifically, the model struggles to predict expensive diamonds accurately, as seen by the larger spread of residuals at higher predicted prices. Additionally, there are a few **outliers**, where the residuals are far from the red line. These outliers indicate that the model significantly over- or under-predicts prices for certain diamonds.

What This Means for the Model:

The **curved pattern** of residuals indicates potential issues with the model's fit. This suggests that the relationship between the predictors and diamond prices might be **non-linear** or that important interaction terms between variables are missing. The **heteroscedasticity** (increasing spread of residuals at higher prices) highlights that the model's predictions are less reliable for higher-priced diamonds, which violates a key assumption of linear regression (constant variance of residuals).

Histogram of Residuals

```
# Create the histogram

hist(residuals,
     breaks = 50,
     col = "#42f5b3",
     border = "black",
     main = "Histogram of Residuals",
     xlab = "Residuals",
     ylab = "Frequency")

# Add a vertical line at zero for reference

abline(v = 0, col = "red", lwd = 2)
```



The histogram of residuals provides insights into the performance of the regression model for predicting diamond prices. Most of the residuals are centered around zero, indicating that the model's predictions are generally accurate on average. However, the shape of the histogram reveals several important issues.

The distribution is not perfectly normal, as shown by a sharp peak at zero and heavy tails on both sides. This suggests that the model does not fully meet the assumption of normally distributed residuals, which is critical for accurate hypothesis testing and confidence intervals. Additionally, outliers are observed in the tails, with some residuals exceeding -5000 and 5000. These outliers indicate that the model significantly over- or under-predicts prices for certain diamonds, possibly due to unusual data points, errors, or unaccounted patterns.

The presence of heavy tails also hints at potential heteroscedasticity, where the variance of residuals may not be consistent across the range of predicted prices. This violates another key assumption of regression and can reduce the reliability of the model's results.

9. Conclusion

This project analyzed the factors that influence diamond prices using a dataset of 5000 diamonds. The key attributes studied were carat weight, clarity, color, and cut. The analysis showed that carat weight has the strongest impact on price, with larger diamonds being significantly more expensive. Clarity also plays a major role, as diamonds with fewer flaws (higher clarity) command much higher prices. Color influences price to a lesser extent, with diamonds

closer to colorless being more valuable. Cut has the smallest effect, though better cuts still increase the price slightly.

The regression model used to predict diamond prices performed well, explaining over 92% of the variation in prices. This indicates that the chosen features are strong predictors of price. However, the residual analysis revealed some challenges. The model struggled to accurately predict prices for very low- and high-priced diamonds, showing non-linearity and inconsistent error variance (heteroscedasticity). These findings suggest that additional factors or more advanced modeling techniques could further improve accuracy.

Overall, the project provided meaningful insights into how diamond characteristics affect price. It highlighted the importance of carat weight and clarity as the most influential factors, while also identifying areas where the model could be refined for better performance.