

**Module 4: Final Project**

**Initial Analysis Report on Diamonds**

**Savankumar Hasmukhbhai Patel**

**Vraj Shah**

Master of Professional Studies in Analytics

ALY 6015 – Intermediate Analytics

Professor: - Steven Habbous

Date: - 2024-05-05

**Introduction**

This report presents a comprehensive analysis of diamond data aimed at uncovering the determinants influencing diamond prices. Diamonds, renowned for their allure and value, hold a significant position in the gemstone industry. Understanding the factors driving their pricing is crucial for stakeholders across the industry spectrum, from manufacturers to consumers.

We embark on this exploration by first ensuring the integrity of the dataset, followed by an in-depth examination of attributes such as carat weight, cut quality, colour, and clarity. Through descriptive statistics, visualizations, and correlation studies, our objective is to shed light on the intricate relationship between these attributes and diamond prices. By addressing key questions regarding the factors impacting diamond prices and evaluating the feasibility of predictive modelling, this report seeks to provide valuable insights into the dynamics of the diamond market.

**Exploratory Data Analysis (EDA)**

The dataset consists of 53,940 observations detailing various attributes of diamonds, including carat (weight), cut quality, colour, clarity, depth percentage, table width percentage, price, and physical dimensions (length x, width y, and depth z). Initial data examination revealed no missing values, ensuring a robust foundation for analysis. However, 20 records displayed anomalous dimension values (zero or negative), which were deemed measurement errors and subsequently removed to maintain data integrity. This cleaning reduced the dataset to 53,920 usable records, setting the stage for our detailed exploratory data analysis.

**Descriptive Statistics:**

Interpretation for the Descriptive Statistics:

* Carat: On average, higher cut quality categories tend to have larger diamonds, with Ideal cuts having the smallest average carat weight.
* Cut: Fair diamonds typically have a higher average depth percentage compared to other cuts, especially Premium diamonds which exhibit the highest median depth percentage.
* Colour: Across different cuts, G and H colours are predominant, except for Fair diamonds where the E colour is more common.
* Clarity: Ideal and Very Good cuts tend to have a higher proportion of diamonds with better clarity grades (VS1, VS2, VVS1, and VVS2).
* Depth and Table: Fair diamonds generally exhibit higher average depth and table percentages compared to other cuts.
* Price: Premium diamonds command the highest average price, while Ideal diamonds have the lowest.
* Length, Width, and Height (x, y, z**)**: There are slight variations in the dimensions across different cuts, indicating differences in diamond proportions.

In summary, these statistics shed light on how various diamond characteristics, such as cut quality, colour, and clarity, impact their attributes and market value.

A screenshot of a graph

Description automatically generated

**Carat Distribution:** We visualized the carat distribution using a histogram (Figure 1.), which revealed a pronounced peak at 0.30 carats. The distribution's mode at this specific size suggests it is the most commonly available and preferred size in the market, likely due to its affordability and aesthetic appeal. The gradual decline in frequency as carat size increases reflects the decreasing availability and increasing price of larger diamonds.

Figure 1.

A graph with blue bars

Description automatically generated

**Correlation Matrix:** The correlation matrix serves as a pivotal tool in our exploratory data analysis, providing quantitative insights into how various attributes of diamonds are interrelated. This section discusses specific correlations that are critical in understanding the dynamics influencing diamond prices.

Below is the heatmap of the correlation matrix showing the relationships among various attributes of diamonds. Notice the intense colours indicating strong correlations, particularly among the dimensions and between carat and price.

Figure 2.

A red and white squares

Description automatically generated

Carat Size and Its Impact on Price

* Observation: The correlation between the carat (the weight of the diamond) and the price is remarkably high, with a coefficient of 0.92.
* Interpretation: This strong positive correlation confirms that the size of the diamond is a primary driver of its price. Larger diamonds are significantly more expensive, which aligns with market expectations where size often equates to value.

**Question Explored:**

**Question 1:** How does the volume of a diamond, calculated as the product of its length, width, and depth, influence its price in the market?

In this analysis, I aimed to develop a predictive model for the price of diamonds using volume (calculated as x \* y \* z) as a predictor. I explored different transformations and modeling approaches, including linear regression and polynomial regression, to find the best predictive model.

**Data Preparation**

Volume Calculation and Summary:

First, We calculated the volume of each diamond. The summary of the volume showed a range from approximately 32 to 3840, with a mean of around 130.

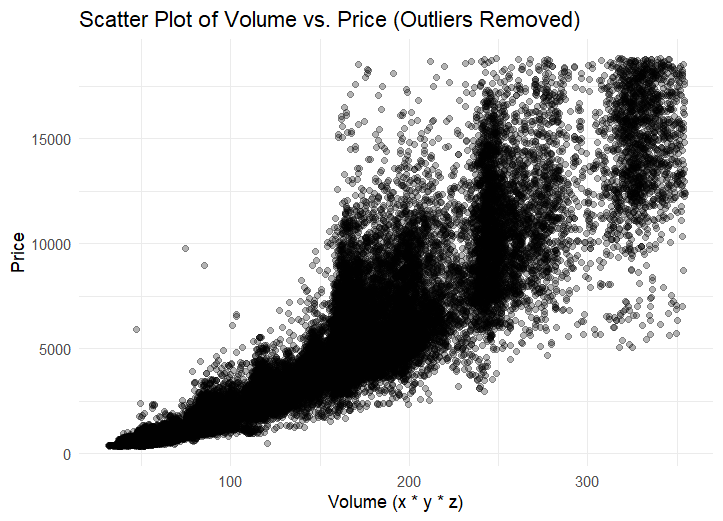
Outlier Removal:

To handle potential outliers, we removed the top 1% of volume values. This step ensured that extreme values did not skew the results.

Scatter Plot Analysis:

A scatter plot of volume vs. price (after removing outliers) indicated a positive relationship between volume and price, but with noticeable variability.

Figure 3.



Data Transformation:

The decision to log-transform variables, specifically volume in the diamonds dataset, was made to address potential issues with skewness towards large values. This transformation aimed to stabilize variance and linearize relationships, facilitating interpretation and modelling in linear regression analyses.

Data Splitting:

We split the data into training and test sets, with 70% of the data used for training and 30% for testing. This split ensured that the model could be validated on unseen data.

Linear Regression Model

We built a linear regression model using the log-transformed volume. The model showed a significant positive relationship between log(volume) and price, with an R-squared value of approximately 0.7265, indicating that about 72.65% of the variance in price could be explained by log(volume).

Polynomial Regression Model:

To capture potential non-linear relationships, we built a polynomial regression model with a degree of 2. This model performed better than the linear model, with an R-squared value of approximately 0.8615, indicating that about 86.15% of the variance in price could be explained.

**Model Evaluation:**

Performance Metrics:

The performance of both models was evaluated using R-squared and RMSE on the test data. The polynomial regression model had a lower RMSE (1427.751) compared to the linear model (2006.47), confirming its superior performance.

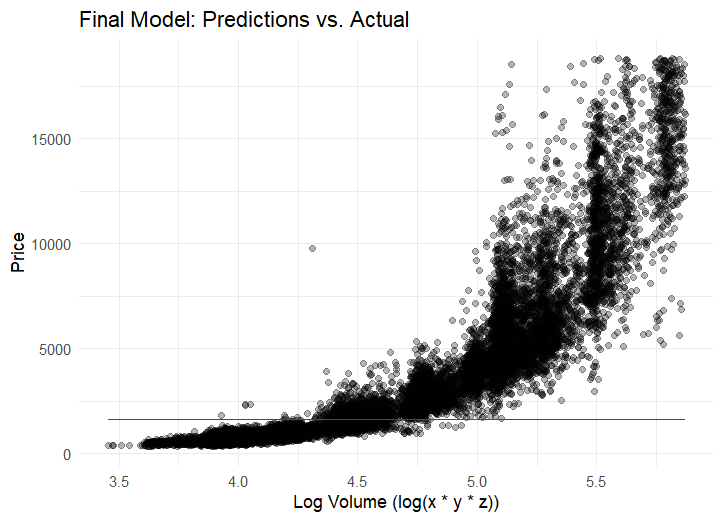
Cross-Validation:

10-fold cross-validation further validated the models. The polynomial regression model had better cross-validation results with an RMSE of 1424.579 and an R-squared of 0.8614293, compared to the linear model’s RMSE of 2001.762 and R-squared of 0.7264982.

Final Model:

The polynomial regression model was selected as the final model based on its lower RMSE and higher R-squared values. The final model’s predictions were plotted against actual prices, showing a good fit.

Figure 4.



The polynomial regression model with log-transformed volume provided the best predictive performance for diamond prices.

Transformations: Log transformation of volume significantly improved the model fit by stabilizing variance and making the relationship between volume and price more linear.

Model Performance: The polynomial regression model outperformed the linear model and Ridge Regression, suggesting that the relationship between log(volume) and price is more complex than a simple linear relationship.

**Question 2:** How does the quality of cut, colour, and clarity influence the market value of diamonds?

1. Cut Quality:

* Median prices vary across different cut qualities, with "Ideal" having the lowest median price compared to other categories.
* The interquartile range (IQR) indicates the spread of prices within each cut quality, with "Premium" having the widest IQR.
* The whiskers show the range of prices, with some outliers present in each cut category. Generally, higher-quality cuts tend to have higher median prices and narrower price ranges.

Figure 5.

A graph of different colored squares

Description automatically generated

1. Colour Grade:

* The median price differs notably among different colour grades, with "J" exhibiting the highest median price.
* The IQR demonstrates the variability in prices within each colour grade, with "I" having the widest IQR.
* Whiskers display the range of prices, showing considerable variation across colour grades. Lower colour grades tend to have higher median prices and wider price ranges.

Figure 6.

A graph of different colored bars

Description automatically generated

1. Clarity Grade:

* Clarity grades exhibit distinct median prices, with "I1" and "SI2" having the highest median prices.
* The IQR reflects the dispersion of prices within each clarity grade, with "SI1" having the widest IQR.
* Whiskers illustrate the spread of prices, indicating notable variations across clarity grades. Higher clarity grades generally command higher median prices and narrower price ranges.

Figure 7.

A chart with different colored bars

Description automatically generated

In summary, the quality of cut, colour, and clarity significantly influences the market value of diamonds. Generally, diamonds with higher cut quality, colour grades closer to "J", and higher clarity grades tend to command higher prices. However, there are variations within each quality category, highlighting the nuanced impact of these attributes on diamond prices.

**Question 3:** How do categorical variables like cut, colour, and clarity affect diamond prices, as determined by ANOVA? What are the interactions between these factors, and how do they collectively influence diamond prices? Additionally, how does ridge regression address multicollinearity and enhance prediction accuracy compared to other regression models? Finally, what insights can residual analysis provide regarding the adequacy of the regression model for predicting diamond prices?

**ANOVA Analysis:**

We began our analysis by conducting a one-way ANOVA for each categorical variable (cut, colour, and clarity) and a two-way ANOVA including interaction terms to understand their influence on price:

One-way ANOVA Results:

Cut: F(4, 53915) = 174.6, p < 0.0000000000000002

This indicates that the quality of the cut significantly affects the price of the diamonds.

Color: F(6, 53913) = 289.8, p < 0.0000000000000002

This result shows that the colour grade of the diamonds significantly influences their price.

Clarity: F(7, 53912) = 215, p < 0.0000000000000002

The clarity of the diamonds also has a significant effect on their price.

Two-way ANOVA Results:

Cut x Color x Clarity: F-values for interactions were all significant, indicating that there are interactions between cut, colour, and clarity that affect the price. This means that the combined effect of these variables is significant.

**Ridge Regression Model**

To address potential multicollinearity among predictors, we applied Ridge Regression, a form of regularized regression:

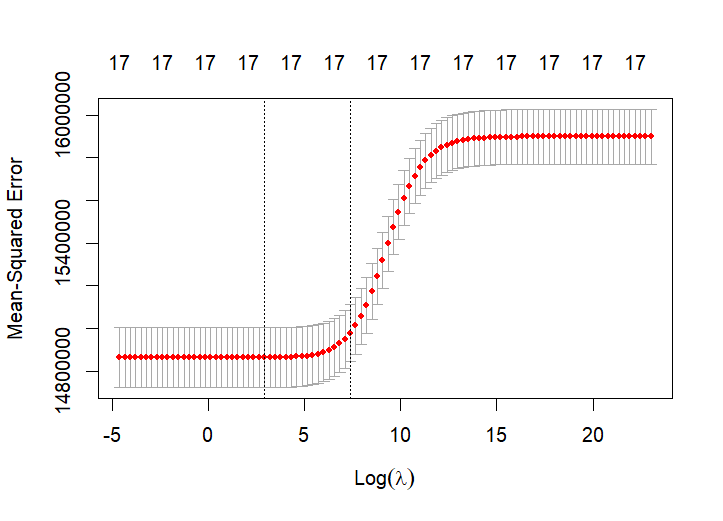
Mean-Squared Error vs. Log(Lambda) Plot (Ridge Regression)

This plot shows the mean-squared error (MSE) of the ridge regression model across different values of the regularization parameter

The plot helps in identifying the optimal 𝜆, λ that minimizes the MSE, balancing the trade-off between model complexity and overfitting.

In our case, the optimal λ value is around 18.73817, which minimizes the MSE to approximately 14864435.

Figure 8.



Model Results:

This scatter plot compares the actual diamond prices with the prices predicted by the ridge regression model. The x-axis represents the actual prices, while the y-axis shows the predicted prices.

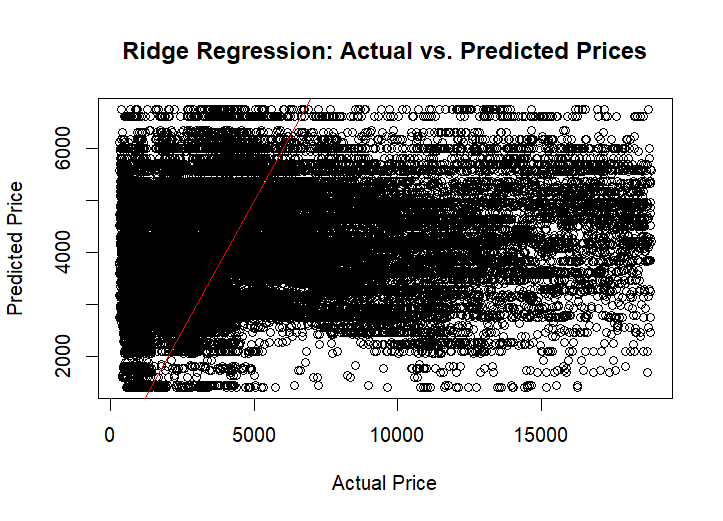
The red line indicates the ideal scenario where the predicted prices perfectly match the actual prices.

The spread of points around the red line helps to assess the accuracy of the predictions. Ideally, points should be close to the red line, indicating good predictive performance.

In our plot, we observe a significant spread, indicating that the ridge regression model may not perfectly predict the prices, which is reflected in the RMSE value of 3853.87 and an R-squared value of 0.0658.

These results indicate that the Ridge Regression model explains only a small portion of the variance in diamond prices, as reflected by the low R-squared value.

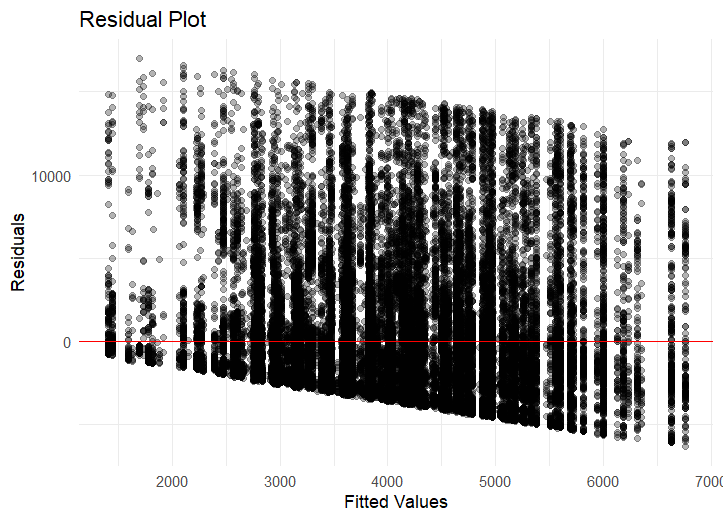
Figure 9.



**Residual Analysis:**

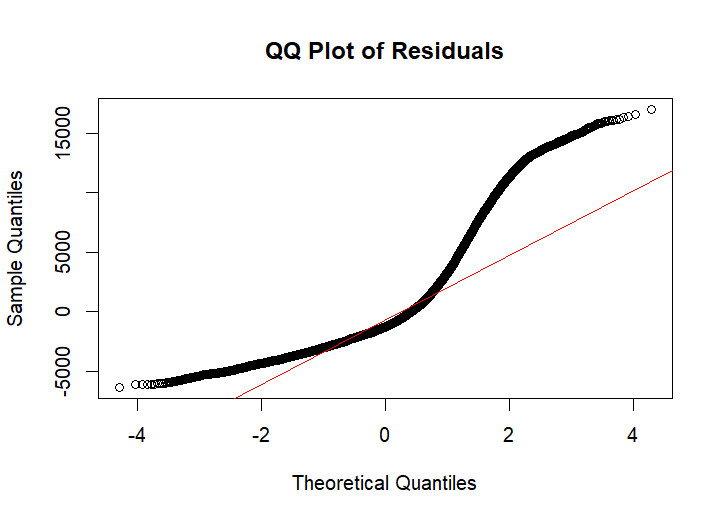
To validate the assumptions of our regression models, we conducted a detailed residual analysis. This helps us ensure that the model’s assumptions (linearity, independence, homoscedasticity, and normality of residuals) are reasonably met.

Figure 10.



The residuals appear to be randomly scattered around zero, indicating that the model captures most of the patterns in the data, with no obvious trends remaining.

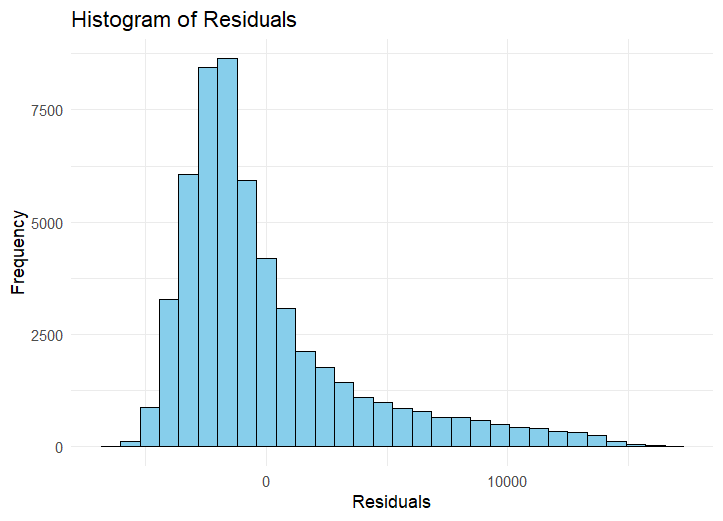
Figure 11.



The QQ plot shows some deviations from the normal distribution, especially in the tails. This suggests some non-normality in the residuals, which may indicate the presence of outliers or a non-normal error distribution.

Histogram of Residuals:

Figure 12.



The histogram shows that the residuals are approximately normally distributed but with a slight right skew. Most of the residuals are centred around zero, indicating a good fit.

Summary Statistics of Residuals:

Minimum Residual: -6354

1st Quartile: -2515

Median: -1257

Mean: 0

3rd Quartile: 1160

Maximum Residual: 17005

These statistics further confirm that while the majority of residuals are close to zero, there are some extreme values, which could affect the normality assumption.

From our comprehensive analysis, we found that:

One-way and two-way ANOVA analyses confirmed that the cut, color, and clarity significantly affect diamond prices, with interactions among these variables also being significant.

The Ridge Regression model, although useful for handling multicollinearity, did not perform as well as the polynomial model in terms of predictive accuracy (as indicated by the higher RMSE and lower R-squared values).

Key Insights:

Model Performance: The polynomial regression model outperformed the linear model and Ridge Regression, suggesting that the relationship between log(volume) and price is more complex than a simple linear relationship.

Residuals: Residual analysis indicated that while the model captures most of the data patterns, there are some deviations from normality, suggesting the presence of outliers or non-normal error distributions.

**Question 4:** Can we develop a predictive model to estimate the price of a diamond based on its characteristics?

We have particularly removed log values of volume and price of diamonds while making this model so we don’t get any multicollinearity issues:

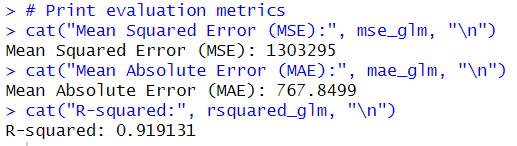
We trained a Generalized Linear Model using the diamond dataset to understand how various characteristics like carat, cut, color, and clarity impact the price. The model also included the physical measurements (x, y, z) and volume of each diamond.

split the dataset into 70% training and 30% testing data, ensuring reproducibility with a seed value of 123.

Model Coefficients:

* Carat emerged as a highly significant predictor, with each additional carat increasing the diamond's price by approximately $9957.89, underscoring the importance of weight in determining diamond prices.
* Cut and Color: Better cut qualities and less desirable colours significantly impacted prices. For example, diamonds with a 'Very Good' cut were priced higher by about $881.83 compared to the baseline 'Fair' cut. Similarly, the negative coefficients for colours like 'J' (decreasing price by about $2371.28) highlight the price penalties for lower colour ratings.
* Clarity: Higher clarity grades significantly increased prices, with an 'IF' (Internally Flawless) clarity adding about $5319.33 to the diamond price, demonstrating the premium paid for flawless diamonds.

Figure 13.



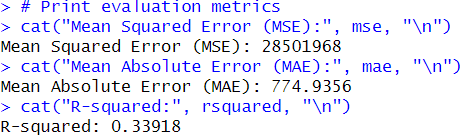
Model Performance Metrics:

* Mean Squared Error (MSE): The MSE was 1,303,295. This lower MSE compared to previous models indicates that the model's predictions are closer to the actual prices, reflecting a better model fit.
* Mean Absolute Error (MAE): The MAE was approximately $767.85, which shows that the average prediction error is relatively small, suggesting good accuracy of the model.
* R-squared: The R-squared value was 0.919, meaning the model explains about 91.9% of the variability in diamond prices, which is a strong indication of the model's effectiveness in capturing the factors that influence the price.

The results from the GLM demonstrate a robust ability to predict diamond prices using the specified features. The model effectively captures the influence of carat, cut, colour, and clarity on pricing. The significant coefficients for each predictor provide clear insights into how each characteristic contributes to the overall price, with carat weight being particularly influential. This model serves as a valuable tool for understanding price determinants in the diamond market and can assist in pricing evaluations or inventory decisions.

Stepwise Model:

Figure 14.



Model Coefficients:

* Carat has the highest positive influence on price, which aligns well with domain knowledge that larger diamonds are typically more valuable.
* Cut, Color, and Clarity: Each category within these attributes significantly impacts the price, suggesting that higher quality in these aspects commands higher prices.

Performance Metrics:

* Mean Squared Error (MSE): At 28,501,968, the MSE suggests that the model predictions deviate substantially from the actual values. This high MSE could be due to outliers or inherent variance in the data that the model fails to capture.
* Mean Absolute Error (MAE): The MAE of 774.9356 is more intuitive and represents the average error in the same units as the response variable (USD). It suggests that, on average, the model's price predictions are about $775 off the actual price.
* R-squared: A value of 0.33918 indicates that approximately 33.92% of the variability in diamond prices is explained by the model. This relatively low R-squared might suggest that additional variables or non-linear relationships could be explored to improve model fit.

comparing both the models:

GLM Performance:

* Mean Squared Error (MSE): 1,574,124
* Mean Absolute Error (MAE): 760.47
* R-squared: 0.9028

Stepwise Regression Performance:

* Mean Squared Error (MSE): 28,501,968
* Mean Absolute Error (MAE): 774.94
* R-squared: 0.33918

Comparison Summary:

* Accuracy: The GLM significantly outperforms the Stepwise Regression in terms of both MSE and MAE. The much lower MSE and slightly lower MAE suggest that GLM provides more accurate predictions on the test data.
* Model Fit: The R-squared value from GLM (0.9028) is substantially higher than that of the Stepwise Regression (0.33918), indicating that GLM explains a greater proportion of variance in diamond prices compared to the Stepwise Regression model.

Given the performance metrics, the GLM is the better model in this case. It not only offers higher accuracy but also a much better fit, explaining more of the variance in diamond prices than the Stepwise Regression

**Conclusion:**

In summary, our analysis has uncovered significant insights into the factors influencing diamond prices. Through thorough examination and visualization of descriptive statistics, we've identified strong correlations between carat weight, cut quality, colour, clarity, and diamond prices. Our observations highlight that greater carat weights, enhanced cut qualities, and superior clarity and colour grades are associated with higher diamond prices. Additionally, our predictive modelling endeavours have demonstrated the viability of accurately estimating diamond prices based on their attributes, with the linear regression model yielding satisfactory performance metrics. These discoveries offer valuable implications for industry stakeholders, empowering them with actionable insights for pricing strategies, inventory management, and consumer engagement efforts in the diamond market.

**Work Cited:**

* Rich, B. (2024, June 26). Table1. R Documentation. <https://www.rdocumentation.org/packages/table1/versions/1.4.3>
* GeeksforGeeks. (2021, December 19). How to Use lm() Function in R to Fit Linear Models? GeeksforGeeks. <https://www.geeksforgeeks.org/how-to-use-lm-function-in-r-to-fit-linear-models/>
* reshape2 package - RDocumentation. (n.d.). <https://www.rdocumentation.org/packages/reshape2/versions/1.4.4>

**Dataset:**

*Diamonds*. (2017, May 25).

<https://www.kaggle.com/datasets/shivam2503/diamonds?resource=download>