**Module 4: Final Project**

**Initial Analysis Report on Diamonds**

**Savankumar Hasmukhbhai Patel**

**Vraj Shah**

Master of Professional Studies in Analytics

Student ID Number: - 002806850

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, color, 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 modeling, 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, color, 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.
* Color: Across different cuts, G and H colors are predominant, except for Fair diamonds where E color 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, color, 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 colors 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 carat (the weight of the diamond) and 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:** What factors have the most significant impact on the price of a diamond?

According to the linear regression model results, the factors exerting the most pronounced influence on diamond price are:

1. Carat Weight: The weight of the diamond, measured in carats, emerges as the most impactful determinant. As carat weight increases, diamond prices exhibit a significant rise.
2. Clarity: The clarity grades assigned to diamonds (including IF, SI1, SI2, VS1, VS2, VVS1, VVS2) also wield substantial influence over pricing. Diamonds with higher clarity grades typically command higher prices.
3. Cut Quality: Varied cut qualities significantly affect diamond pricing. Notably, diamond cuts classified as Ideal, Premium, and Very Good are associated with higher price points, indicative of their superior quality.
4. Color Grading: The color grading of diamonds (ranging from E to J) plays a pivotal role in determining prices. Lower color grades, such as J, are typically correlated with lower prices compared to their higher-grade counterparts.

A screenshot of a computer

Description automatically generated

A close-up of a number

Description automatically generated

**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.

A graph of different colored squares

Description automatically generated

1. Color Grade:

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

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.

A chart with different colored bars

Description automatically generated

In summary, the quality of cut, color, and clarity significantly influences the market value of diamonds. Generally, diamonds with higher cut quality, color 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:** Can we develop a predictive model to estimate the price of a diamond based on its characteristics?

Yes, we can indeed develop a predictive model to estimate the price of a diamond based on its characteristics.

After training a linear regression model on 70% of the dataset and evaluating it on the remaining 30%, we obtained the following performance metrics:

* The Mean Squared Error (MSE) was 1,292,521.75.
* The Mean Absolute Error (MAE) amounted to 738.20.
* The R-squared value stood at 0.919.

These metrics indicate that the model performs adequately in predicting diamond prices using their characteristics. The relatively low MSE and MAE suggest that the model's predictions are close to the actual prices on average. Moreover, the high R-squared value of 0.919 suggests that the model explains approximately 91.9% of the variance in diamond prices, indicating a good fit to the data.

Thus, based on these evaluation metrics, we can affirm that a linear regression model is effective in estimating the price of a diamond based on its characteristics.

A screenshot of a computer code

Description automatically generated

**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, color, clarity, and diamond prices. Our observations highlight that greater carat weights, enhanced cut qualities, and superior clarity and color grades are associated with higher diamond prices. Additionally, our predictive modeling endeavors 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>