**Module 4: Final Project – Initial Analysis Report on Diamonds**

Name: - 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**

In this project, we conducted an exploratory data analysis of a dataset comprising various attributes of diamonds. The goal was to identify factors that significantly influence diamond prices, which could later be used to build a predictive model. Our analysis includes descriptive statistics, distribution analysis, correlation studies, and bivariate analysis focusing on the relationship between price and other attributes.

**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:**

In our analysis, we compiled comprehensive descriptive statistics for each attribute of the diamonds segmented by the quality of the cut. This data is crucial as it provides a foundational understanding of the distribution and range of each characteristic across different categories.

The table presents the mean, standard deviation, and range (minimum to maximum values) for key attributes such as depth percentage, table width, price, and dimensions (x, y, z) of the diamonds. These metrics vary significantly by cut, highlighting how each cut category differs in terms of average characteristics and variability.

**Depth and Table**: The average depth percentage ranges narrowly from 61.7% to 62.4% across all cuts, suggesting a standard cutting practice in terms of depth. However, the variability (standard deviation) and the range of depth highlight differences in cut precision.

**Price Variability:** The average prices show a broad range, from $2,400 for the least expensive cuts to $4,580 for the most expensive on average. Notably, the range of prices within each cut category is extensive, demonstrating the significant influence of factors beyond cut quality, such as carat size and clarity.

**Dimensions (x, y, z):** The dimensions of the diamonds also vary with cut quality, with higher-quality cuts tending to have larger average dimensions. This correlation between size and cut quality suggests that larger diamonds are often cut to a higher standard.

These descriptive statistics are pivotal as they not only reflect the physical and aesthetic attributes of the diamonds but also underscore the economic factors at play within different segments of the diamond market. For instance, the wide range of prices within each cut category can be attributed to the interplay between cut quality and other factors like clarity and color, which are not uniformly distributed.

**Univariate Analysis**

**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

**Bivariate Analysis**

In our analysis of how diamond prices are influenced by carat size across different cuts, we utilized scatter plots to visualize the pricing trends for each category of diamond cut. These visual representations demonstrated that as the carat size of a diamond increases, its price also tends to rise. This trend was particularly noticeable in diamonds with higher-quality cuts, such as those labeled as 'Ideal.' (Refer to Figure 2 below)

**Trend in Cuts:** For diamonds with an 'Ideal' cut, the increase in price relative to the increase in carat was more pronounced compared to other cuts. This suggests that not only are larger diamonds more valuable, but their value escalates more steeply when they are also of superior cut quality.

Figure 2.

A graph showing different colored lines

Description automatically generated

**Visual Trends:** The scatter plots for each cut category showed a clustering of data points that generally moved upward as carat size increased. This pattern was especially marked in the 'Ideal' and 'Premium' cut categories, indicating a strong linkage between size and price among high-quality cuts.

Next, in our exploratory analysis, we used boxplots to examine the distribution of prices for diamonds based on their cut quality. Boxplots are a useful visual tool because they show not only the typical range of prices within each cut category but also highlight the outliers—prices that are significantly higher or lower than most of the other prices.

Figure 3.

A chart with different colored squares

Description automatically generated

**Range of Prices:** Our boxplots revealed a wide range of prices within each cut category. For instance, in the 'Premium' cut category, most diamonds are priced between $3,630 and $10,340. However, we also noticed that there are some diamonds priced well above or below this range.

**Outliers in the Data:** Each boxplot showed several outliers, which are prices that stand out because they are much higher or lower than most of the diamonds in that category. These outliers might represent unique diamonds that are either exceptionally large, perfectly clear, or maybe colored, which can significantly affect their value.

**Why We're Keeping Outliers in the Analysis (For Now):**

**Potential Impact on Modeling**: Outliers can sometimes distort statistical models, making them less accurate. However, they can also provide valuable insights into special cases or exceptional values in the dataset. At this stage, we're choosing not to remove these outliers because we aren't yet sure how they will affect our future predictive modeling efforts.

**Future Considerations:** As we move forward, especially when we start building models to predict diamond prices, we will closely examine these outliers to determine their impact. If they skew our model predictions too much, we might reconsider whether to exclude them or adjust our model to accommodate their influence.

**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 4.

A red and white squares

Description automatically generatedCarat 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.

**Relationships Among Physical Dimensions of Diamonds:**

* **Observation:** The dimensions of diamonds, represented by x (length), y (width), and z (depth), show extremely high correlations among each other, all exceeding 0.97.
* **Interpretation:** These high correlations indicate that the dimensions are almost perfectly proportional. This proportionality suggests that as any one-dimension increases, the others do as well, maintaining a consistent shape and form.
* **Practical Implication:** For predictive modeling, this insight allows us to consider potential multicollinearity issues if all three dimensions are used as predictors. Simplifying the model by using only one dimension or a derived feature like volume (calculated as x\*y\*z) might be more effective.

**Conclusion from EDA**

The correlation analysis has unequivocally highlighted that carat size, physical dimensions, and the perceived quality of the cut are pivotal in determining diamond pricing. These insights guide our focus towards leveraging these variables for robust predictive modeling. Furthermore, the analysis indicates that despite the importance of cut quality, its impact on price is not uniform across the board, suggesting nuanced interactions with other attributes like carat and dimensions.

**Overall Conclusion**

Our comprehensive exploratory analysis has not only deepened our understanding of the factors that influence diamond prices but also equipped us with crucial empirical insights. These findings form the foundation for the next phase of our project, where we will apply advanced regression techniques to predict diamond prices based on the key predictors identified. This approach will help refine our predictive capabilities and offer valuable models for stakeholders in the diamond industry.

**Work Cited:**

* Zach, B. (2021, April 1). *How to Create a Confusion Matrix in R (Step-by-Step)*. STATOLOGY. <https://www.statology.org/confusion-matrix-in-r/>
* Zach, B. (2021, March 23). *How to Use the predict function with glm in R (With Examples)*. STATOLOGY. <https://www.statology.org/r-glm-predict/>
* Zach, B. (2021, November 15). *How to Interpret glm Output in R (With Example)*. STATOLOGY. <https://www.statology.org/interpret-glm-output-in-r/>
* (2024, January 25). *AUC ROC Curve in Machine Learning*. GeeksForGeeks. <https://www.geeksforgeeks.org/auc-roc-curve/>
* Zach, B. (2020, October 29). *How to Plot a ROC Curve Using ggplot2 (With Examples)*. STATOLOGY. <https://www.statology.org/roc-curve-ggplot2/>