Diamonds Market Analysis Report

Introduction

The diamond market is well known for its complexity, with prices determined by multiple factors such as carat, cut, clarity, and other physical attributes. This project focuses on a simplified analysis of a sample of 500 diamonds in order to explore how certain measurable features—particularly carat, depth, and table—relate to price. The main goal is to apply basic statistical methods and data visualization to better understand the structure of the dataset and to observe how a few key variables drive val...

Methodology

The dataset Diamonds_sample.csv was randomly selected from a larger collection to reduce computational burden and make exploratory analysis more manageable. All analysis was conducted in R. The workflow consisted of four main steps:

1. Data Sampling and Summary

- Extracted 500 random rows from the full dataset.

- Descriptive statistics were generated to provide an overview of price, carat, depth, and table values.

2. Exploratory Data Analysis (EDA)

- The distribution of prices was examined with histograms and boxplots.

- Carat values were grouped into half-carat intervals to compare average prices across weight categories.

3. Correlation and Visualization

- Correlation coefficients were calculated between price and the three main attributes (carat, depth, and table).

- Scatter plots were produced to visualize the relationships between price and carat, as well as price and depth.

4. Distribution Checks

- Normality was tested for depth and table using the Shapiro-Wilk test.

- Q-Q plots were created to evaluate whether these variables followed normal distributions.

Analysis and Results

The histogram of diamond prices revealed a right-skewed distribution, with most diamonds priced at the lower end of the scale but a few extreme outliers that significantly increased the range. The boxplot confirmed the presence of these outliers.

Grouping diamonds by carat weight showed that average price rose sharply as carat increased, particularly beyond the 1-carat threshold. This observation is consistent with industry knowledge that carat is one of the strongest predictors of diamond value.

Correlation analysis further supported these findings. The correlation between price and carat was approximately 0.9, indicating a very strong positive relationship. By contrast, depth and table had little to no correlation with price, suggesting that while they are part of the standard grading system, they may not directly drive market value in the same way.

The Shapiro-Wilk tests indicated that depth values were approximately normal, while table values were close to normal but not perfect. These findings were also visible in the Q-Q plots. Price, on the other hand, did not follow a normal distribution and instead showed

characteristics closer to an exponential distribution.

Discussion

The results confirm that carat weight is the single most influential factor in determining diamond price within this dataset. While other factors such as cut and clarity were not included in this analysis, the findings illustrate the importance of quantitative methods in validating assumptions about market behavior.

It is worth noting that the right-skewed price distribution mirrors the nature of many real-world markets, where a small number of high-value items push the average higher. This pattern reinforces the need to look beyond mean values and consider median and distributional shape when analyzing price data.

Conclusion

This project demonstrates how a relatively small sample of data can be used to uncover meaningful insights about the diamond market. By applying descriptive statistics, correlation analysis, and visualization techniques in R, it is possible to identify the dominant role of carat weight in pricing and to gain a clearer understanding of market distribution.

The project serves as an introduction to practical data analysis and highlights skills that are broadly transferable to other industries, including exploratory data analysis, correlation testing, and data visualization.

Skills Demonstrated

- Data handling and sampling in R

- Statistical analysis and correlation testing

- Visualization with histograms, boxplots, and scatter plots

- Writing structured, reproducible analysis