# Diamond Attribute Insights: Unveiling Value and Quality

Final Report  
  
Introduction

## Problem Statement

The diamond industry is a multi-billion-dollar global market where pricing is influenced by a variety of factors such as carat weight, cut, color, clarity, and other physical attributes. However, customers and sellers often struggle with inconsistent pricing due to the lack of transparency and a clear understanding of how these attributes impact diamond value. The task involves analyzing a dataset containing detailed characteristics of diamonds to uncover trends and relationships that can help improve pricing accuracy and transparency in the marketplace.

## Business Impact

This analysis can enable both buyers and sellers to make more informed decisions. Sellers could optimize their pricing strategies, ensuring competitiveness in the market, while buyers could better assess the fairness of the prices they are offered. Over time, this transparency could improve trust in the diamond marketplace and potentially attract a wider range of customers.

## Data

**File Name:** Diamonds.csv 🡪 **Diamonds price Dataset**

**Description:** This Dataset represents a range of diamonds with different features such as size, and clarity, and they have prices

**Dataset Details**: 53,941 Rows & 10 Columns

**Size**: 2.43 MB

**Source**: Kaggle [Diamonds Price Dataset](https://www.kaggle.com/datasets/amirhosseinmirzaie/diamonds-price-dataset)

## Data Analysis & Computation

# The dataset used in this project was sourced from Kaggle, a popular platform for open datasets. It included attributes such as cut, color, clarity, carat weight, and price, along with additional features like depth, table, and dimensions.



Figure : Carat Weight

A diagram of different types of diamonds

AI-generated content may be incorrect.

Figure : Diamond Clarity

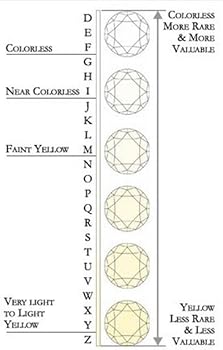


Figure : Diamond Color

### Data Cleaning Process:

# **Outlier Detection:** We used boxplots to detect and handle outliers that could skew the analysis.

# **Feature Engineering:** We created new features, such as price per carat, to enhance the depth of our analysis.

# The cleaning process was justified by the need to ensure data quality and reliability, which are critical for meaningful insights.

## Exploratory Data Analysis (EDA)

# EDA was conducted to uncover patterns, trends, and relationships within the data. Key visualizations and insights include:

# **Price Distribution:** A histogram of diamond prices revealed a right-skewed distribution, indicating that most diamonds are priced at the lower end of the spectrum.

# **Scatterplots:** Scatterplots of carat weight vs. price highlighted a nonlinear relationship, suggesting that carat weight is a significant driver of price.

# These visualizations helped us understand the data structure and provided valuable insights into the factors influencing diamond prices.

### Statistical Analysis

# While predictive modeling was not the focus of this project, I conducted statistical analysis to explore relationships within the data:

# **Descriptive Statistics:** calculated summary statistics (e.g., mean, median, standard deviation) for numerical features to understand their distribution.

# **Correlation Analysis:** examined correlations between numerical features to identify potential multicollinearity and understand their impact on price.

# These analyses provided a solid foundation for understanding the dataset and highlighted areas for future exploration.

### Challenges and Solutions

1. **Challenge:** Limitations of the dataset, such as the lack of real-time market data and additional attributes.  
   **Solution:** While I worked with the available data, I acknowledged its limitations and proposed future improvements, such as incorporating real-time data and additional features.
2. **Challenge:** Outliers and missing values in the dataset.  
   **Solution:** I used robust data cleaning techniques to address these issues and ensure the reliability of our analysis.
3. **Challenge:** Nonlinear relationships between features and price.  
   **Solution:** I used visualizations like scatterplots and boxplots to explore these relationships and gain insights.

### Description of Dashboard

The dashboard developed for this project provides an interactive platform for users to explore diamond data and gain insights into pricing trends. Key features include:

* **Interactive Visualizations:** Users can interact with charts (e.g., scatterplots, histograms) to explore relationships between diamond attributes and price.
* **Filtering Options:** Users can filter data by cut, color, clarity, and carat weight to focus on specific subsets of the dataset.
* **Summary Statistics:** The dashboard displays key statistics (average price per carat, average carat weight) based on user-selected filters.

**Primary Use Cases:**

* Retailers can use the dashboard to identify pricing trends and make informed decisions.
* Consumers can use it to understand how different attributes influence diamond prices.

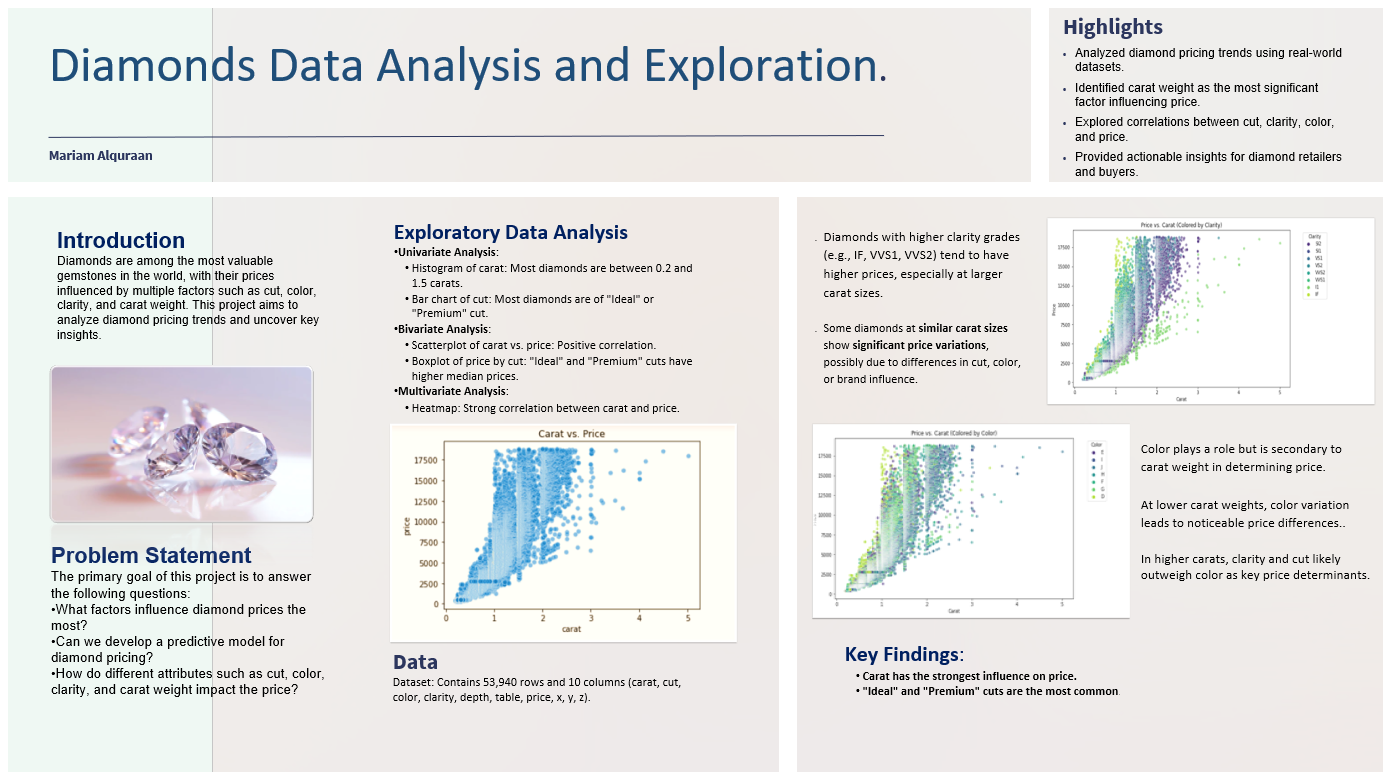


Figure : Datafolio

A screenshot of a computer

AI-generated content may be incorrect.

Figure : Dashboard

* Price Distribution

Purpose: Shows the overall distribution of diamond prices in the dataset.

Why: Helps identify price concentration, outliers, and trends in diamond pricing.

* Carat Distribution

Purpose: Displays how diamond carat weights are distributed.

Why: Highlights the most common carat sizes and reveals any patterns in weight distribution.

* Average Price by Cut

Purpose: Shows how the average price varies by diamond cut quality.

Why: Provides insights into which cut grades (Ideal, Premium, etc.) command higher prices.

* Price by Color and Clarity

Purpose: Visualizes the relationship between diamond color and clarity on pricing.

Why: Highlights price variations based on quality attributes, helping understand their impact on valuation.

* Carat vs. Price | Colored by …

Purpose: Shows the relationship between carat weight and price, with clarity, color or cut levels as color-coded categories.

Why: Helps analyze how carat influences price while factoring in clarity, color or cut variations.

* Carat Avg & Avg Price per Carat

Purpose: Provides a quick summary of the dataset's average carat weight and average price per carat.

Why: Useful for understanding overall dataset characteristics and benchmarking individual diamonds.

## Conclusion & Future Work

This project successfully demonstrated the potential of data-driven approaches to uncover insights into diamond pricing. While predictive modeling was not the focus, our analysis provided a deeper understanding of the factors influencing diamond prices and highlighted the limitations of the current dataset. Key conclusions include:

* Carat weight, cut, color, and clarity are significant drivers of diamond prices.
* Discovered that **carat weight** has the strongest impact on diamond prices, with prices increasing significantly as carat increases.
* Found that **higher-quality cuts** (e.g., Ideal, Premium) and **better clarity grades** (e.g., IF, VVS) lead to higher prices.
* Identified that **diamonds with better color grades** (e.g., D, E) are priced higher than those with lower grades in some cases.
* Observed that **smaller diamonds (lower carat)** tend to have better cuts (e.g., Ideal, Premium) and are often associated with higher prices, reflecting manufacturing or market trends.
* The dataset has limitations, such as the lack of real-time data and additional attributes, which could be addressed in future work.

**Future Work:**

1. **Incorporate Real-Time Data:** Using real-time market data would improve the accuracy and relevance of the analysis.
2. **Add More Attributes:** Including additional features such as fluorescence, polish, and symmetry could provide a more comprehensive understanding of diamond pricing.
3. **Expand Dataset:** A larger and more diverse dataset would enhance the robustness of the analysis.
4. **Predictive Modeling:** Future work could focus on building predictive models to estimate diamond prices based on the insights gained from this analysis.

## Reference

* **Data Source:** Kaggle ([Diamonds Price Dataset](https://www.kaggle.com/datasets/amirhosseinmirzaie/diamonds-price-dataset))
* **Tools Used:** Python, Pandas, Matplotlib, Seaborn, Tableau (for dashboard)