# Overview

The wholesale diamond dataset spans a period of ten years and serves as a comprehensive repository of diamond sales data for the company. It comprises 407,280 diamonds sold between 2012 and 2023, each represented by 11 attributes: carat, cut, color, depth, table, length (mm), width (mm), height (mm), cost (dollars), clarity, and year of sale.

The goal of the exploratory data analysis is to leverage this dataset to uncover patterns, trends, and correlations that can enhance our understanding of the diamond market. The insights gained from this exploratory analysis will guide our efforts in building predictive models for future diamond prices. By combining numerical and categorical data, we aim to identify factors influencing diamond prices and provide actionable recommendations for strategic decision-making.

# Data Quality

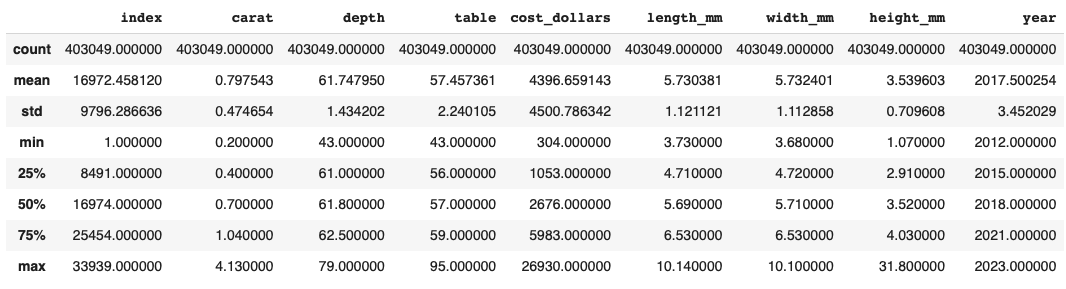
Data quality is a critical aspect that underpins the reliability and effectiveness of any analysis or modeling endeavor. In the context of our wholesale diamond dataset, ensuring data quality involved a thorough examination and cleansing process to address potential issues and inaccuracies, which included:

* 2048 blank entries for carat,
* 2037 negative entries for cost/diamond prices,
* 3 zero cost/diamond price,
* Several zero entries for length, width and height of diamonds,
* Column heading nomenclature, among others.
* Finally, after data cleaning, we ended with 403049 entries

# Findings

### Descriptive statistics

A fundamental aspect of our analysis involves summarizing the distribution of numerical attributes such as carat weight, diamond dimensions (length, width, height), and price. Calculating summary statistics, including measures such as mean, median, standard deviation, and range, will enable the department to gain insights into the central tendencies and variability of these critical variables. Figure 1 below provides the summary statistics of the cleaned diamond dataset.

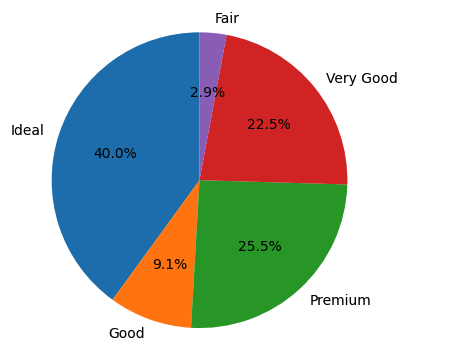


### Diamonds description.

The diamonds in the dataset can be classified according to three categorical attributes; cut type, color and clarity.

**Cut type**

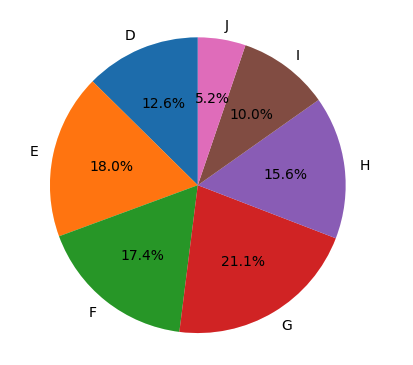
The diamonds were classified into five cut types as indicated in Figure 1 below. Most of the diamonds (40%) fall within the ‘Ideal’ cut type while 2.9% of the diamonds can be described as fair.



**Figure 1:** Distribution of diamond according to cut types.

**Color**

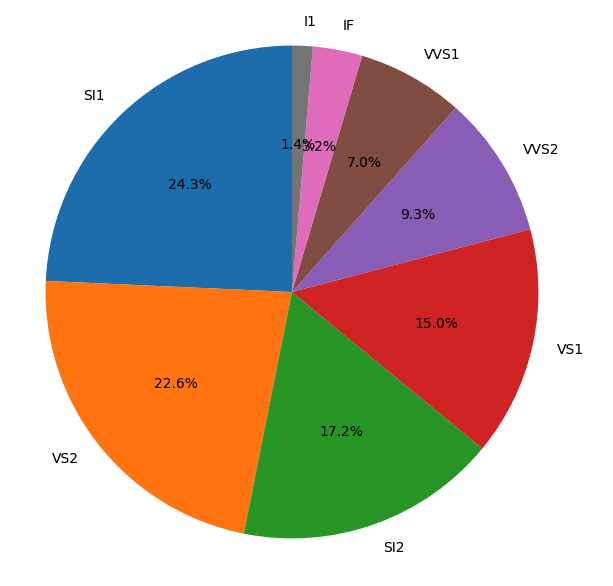
The diamonds were classified into seven colors, see Figure 2 below. Majority (21.1%) of the diamonds belonged to color ‘G’ while only 5.2% belonged to color ‘J’.



**Figure 2:** Distribution of diamonds according to color

**Clarity**

There were eight clarity categories with VS2 being the most common (22.6%) and I1 the least common (1.4%), Figure 3.

**Figure 3:** Distribution of diamonds according to clarity

### Correlation Analysis

A Pearson correlation matrix was calculated, and the results presented in a heatmap (Figure 4).

A screenshot of a graph

Description automatically generated

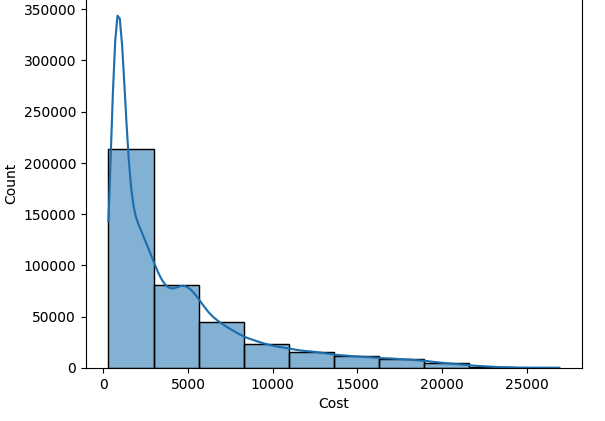
**Figure 4:** Pearson correlation matrix for numeric variables in the diamond dataset

One notable trend is the positive correlation between carat weight and diamond price “cost(dollars)” indicating that heavier diamonds tend to command higher prices. This observation aligns with the traditional understanding of the diamond market, where larger stones are often associated with increased rarity and value. A strong positive relationship was observed between carat and length, width and height of a diamond. Similarly, there was a strong positive correlation between cost (dollars) and length, width and height of a diamond sold. On the other hand, a negative correlation was observed between the table and depth attributes of a diamond.

### Key Trends

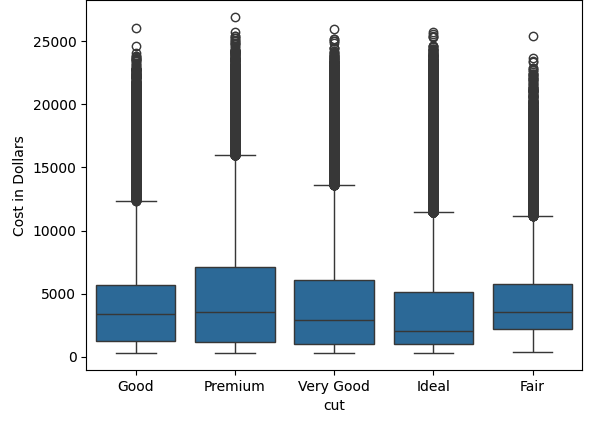
**Price distribution**

Majority of the diamonds sold for less than 5000 dollars. Only a handful sold for above 250,000 dollars, Figure 5, with a price variance of $20,257,077.69 over the ten years. Factors affecting the price of the diamonds included, cut type, color, clarity and year of sale.



**Figure 5:** Diamond price distribution over the ten year sale period

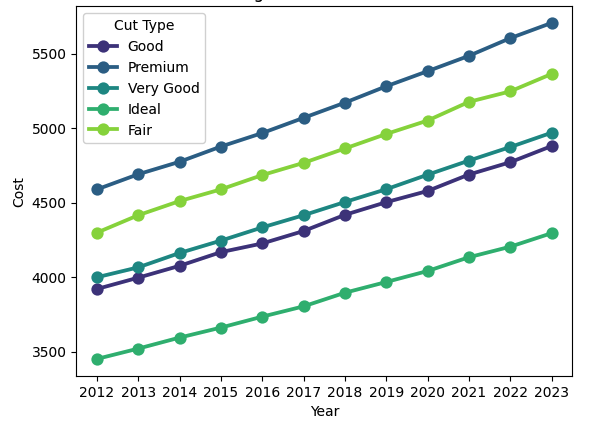
The prices varied across cut types, Generally, the ‘Premium’ cut type cost higher than any other cut type, Figure 6.



**Figure 6:** Average price distribution across cut types.

**Pricing over the years**

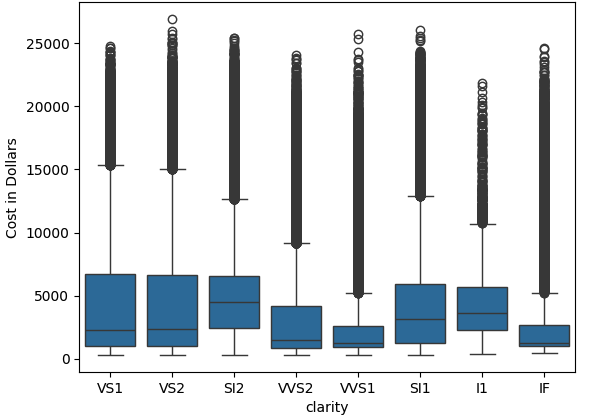
Over the years, diamond prices generally increased for each of the cut types, Figure 7.



**Figure 7:** Trend for diamond prices over the ten-year period.

**Diamond Clarity**

Diamond clarity also affected the price of the diamond with the SI2 category fetching a higher price compared to the rest of the categories, Figure 8.



**Figure 8:** Price variation of diamonds versus clarity types

**Carat distribution**

Most of the diamonds sold in the ten-year period weighed less than 1 carat, Figure 9.

A graph of a number of blue lines

Description automatically generated

**Figure 9:** Carat distribution for diamonds sold between 2012 and 2023.

**Pairwise Clustering of Diamonds**

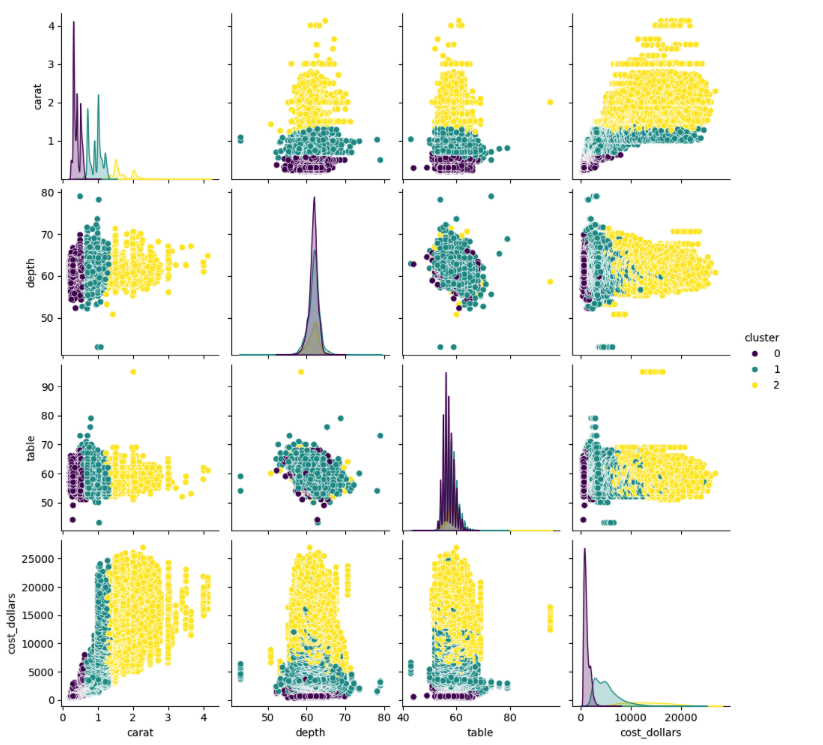
We conducted pairwise clustering to find natural clusters of diamonds considering attributes perceived to have some type of correlation with price as seen below, Figure 10.

Carat vs Cost: There is a positive correlation between the carat of the diamonds and their cost, confirming the expectation that larger diamonds tend to be more expensive.

Depth vs Cost: This relationship is more complex but there seems to be separation based on depth, but this suggests that this cannot be a strong predictor of cost.

Table vs Cost: This relationship is similar to Depth, indicating it is not a good attribute alone to predict cost.

Cluster 2 seems to have a tighter spread within all 3 attributes compared to cost, which could suggest that this group is more similar in physical characteristics.



**Figure 10:** Pairwise plots for Multiple Attributes with natural clusters