**Diamond Attribute Insights: Unveiling Value and Quality**

Exploratory Data Analysis

Name: Mariam Alquraan

1. **Target Variable:**

* **price (numerical):** Represents the diamond's value in USD. This will be the primary variable to predict.

2. **Important Feature Columns:**

* **carat (numerical):** Weight of the diamond; typically has a strong correlation with the price.
* The average carat size is **0.80**, and the average price is **$3,932**.
* **cut (categorical):** Quality of the cut, which impacts the diamond's brilliance. Categories include:
  + Fair
  + Good
  + Very Good
  + Premium
  + Ideal
* **color (categorical):** Diamond color grades, ranging from D (colorless, best) to J (noticeable tint, worst).
* **clarity (categorical):** Clarity grades describe the visibility of inclusions/blemishes. Ordered from best to worst:
  + 1. IF (Internally Flawless)
    2. VVS1 (Very Very Slightly Included 1)
    3. VVS2 (Very Very Slightly Included 2)
    4. VS1 (Very Slightly Included 1)
    5. VS2 (Very Slightly Included 2)
    6. SI1 (Slightly Included 1)
    7. SI2 (Slightly Included 2)
    8. I1 (Included 1)
* **depth (numerical):** Total depth percentage, which affects light performance.
* **table (numerical):** Width of the top of the diamond (table) relative to the widest point.
* **x (numerical):** Length in mm.
* **y (numerical):** Width in mm.
* **z (numerical):** Depth in mm.
* The z column had an extreme outlier (31.8 mm), which was removed for further analysis.
* The most common cut is **Ideal**, the most common color is **G**, and the most common clarity is **SI1**.

3. **Breakdown by Data Type:**

* **Numerical Columns (7):** price, carat, depth, table, x, y, z
* **Categorical Columns (3):** cut, color, clarity
* **No datetime columns** are present in this dataset.

4. **Key Column Relationships to Note:**

* **Size and volume relationships:** x, y, and z together represent the diamond’s physical dimensions, contributing to its perceived size.
* **Depth percentage:** Related to the ratio of z to the mean of x and y. It is an indicator of how "deep" or "shallow" the diamond is.
* **Quality metrics:** The combination of cut, color, and clarity captures the overall quality, which has a direct influence on price.
* **Weight-to-dimension relationship:** carat weight is closely tied to the x, y, and z dimensions and can be cross-validated during analysis.

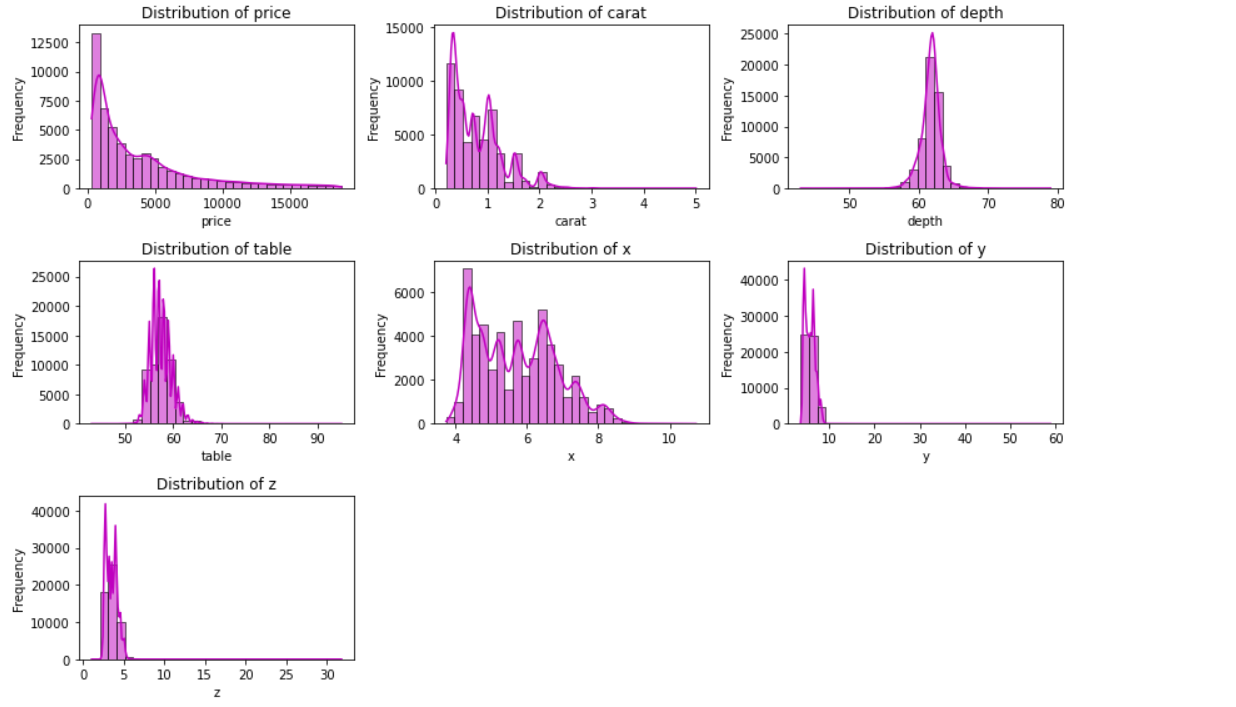
5. **Potential Relationships to Explore in EDA:**

* **Price vs. Carat:** Expect a positive correlation, with price increasing exponentially as carat weight rises.
* **Price vs. Cut/Color/Clarity:** Higher quality (better cuts, colorless, and fewer inclusions) diamonds typically command higher prices.

6. **Linking Columns Across Datasets:**

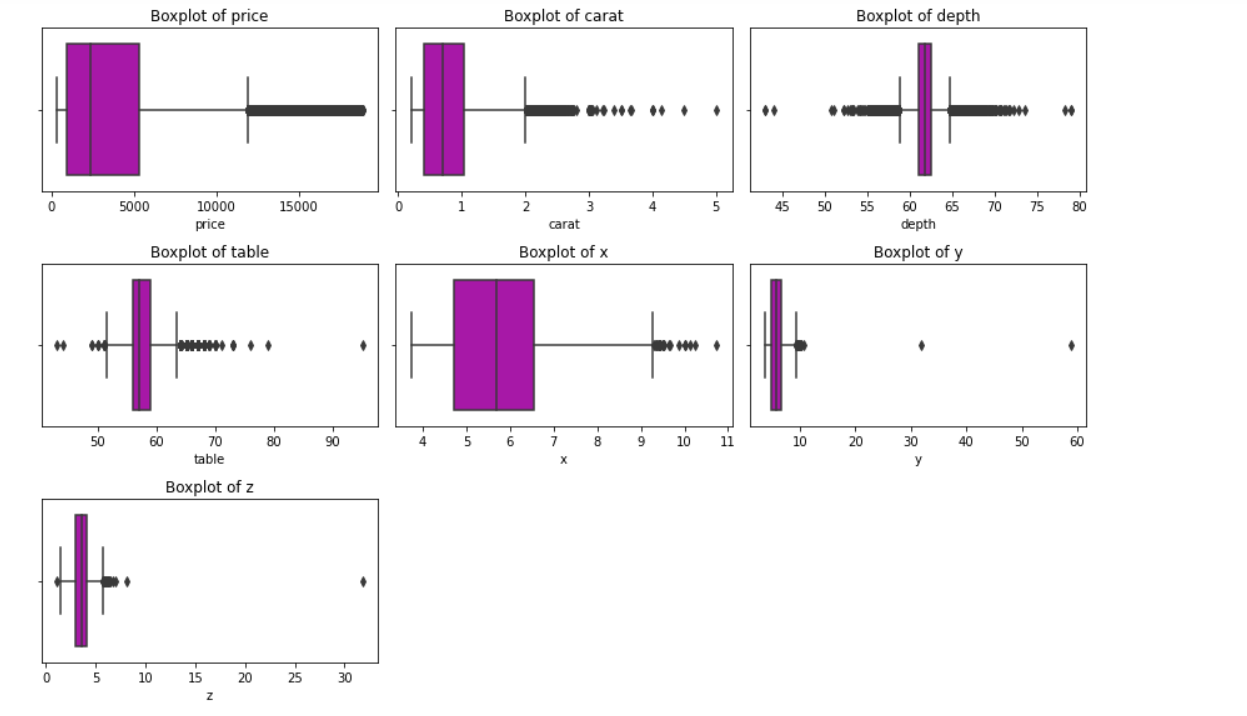
* Use lab (certification authority) to evaluate quality standards across datasets and how will effect on the price.
* Use shape and symmetry to understand how these characteristics relate to quality and pricing in different datasets.

## Charts:



**Distribution Types Observed:**

1. **Price:**
   * Right skewed (positively skewed).
   * Most diamonds have lower prices, with a long tail extending to higher values.
2. **Carat:**
   * Right skewed.
   * Most diamonds are small in carat weight, with a gradual decline for larger diamonds.
3. **Depth & Table:**
   * Near **normal distribution** but slightly skewed.
   * Most values are clustered around a central point.
4. **x, y, and z (dimensions of diamonds):**
   * Right skewed with some **discontinuities**.
   * Some values extend far to the right, indicating a few very large diamonds or **outliers**.



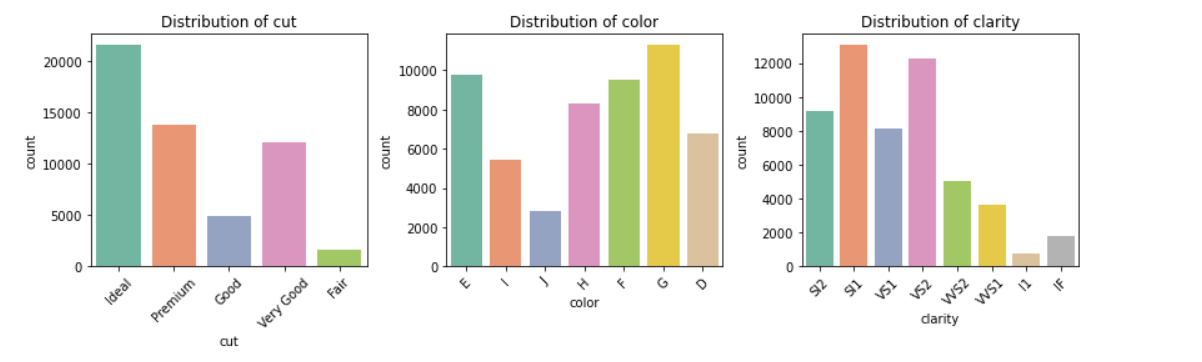
1️. **Outliers (Extreme Values)**

* **Most columns (price, carat, depth, table, x, y, z) have outliers.**
* The **whiskers** represent the range within 1.5 \* IQR (Interquartile Range). Anything beyond is considered an outlier.

2️. **Column-Wise Observations**

* **Price:** Right skewed, with many high-value outliers.
* **Carat:** Right skewed, with very large diamonds being outliers.
* **Depth & Table:** Many small and large values are outliers, indicating variations in diamond cut.
* **x, y, z (dimensions):** Extreme values in y and z columns suggest **possible data errors or rare large diamonds or outliers**.

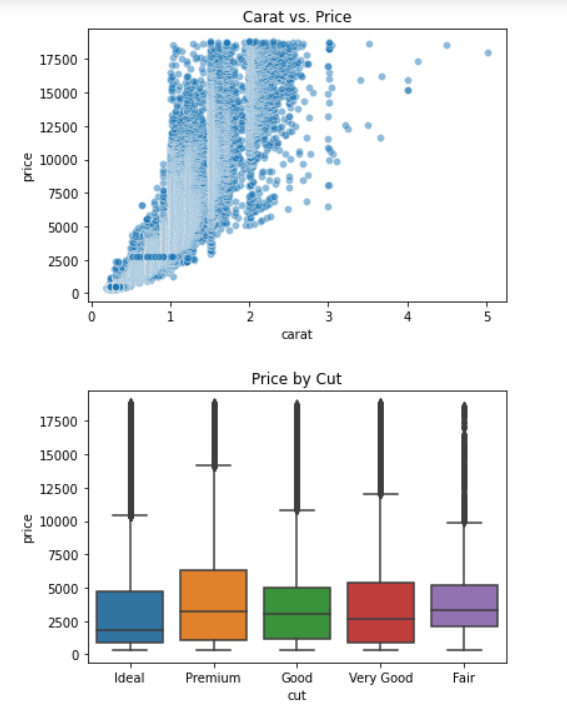
**Visualize Subgroup Size with a Bar Chart:**



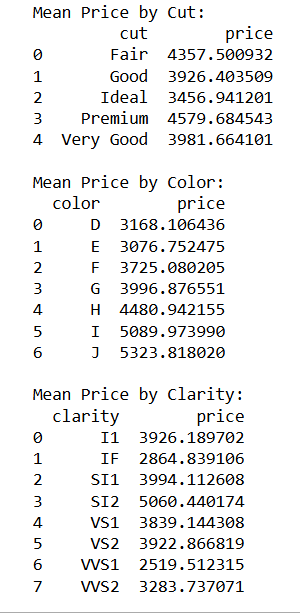
1. **Cut:** "Ideal" is the most common, followed by "Premium" and "Very Good." "Fair" is the least common.
2. **Color:** "G" is the most frequent, while "J" is the least.
3. **Clarity:** "VS2" and "SI1" dominate, while "I1" and "IF" are rare.

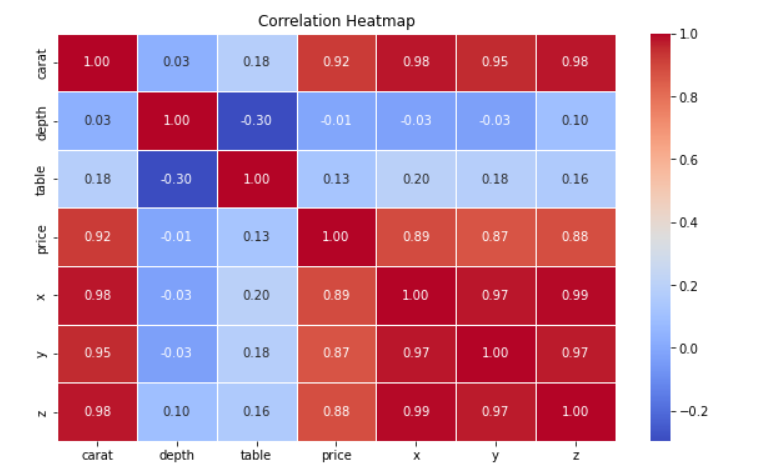
A comparison of colored boxes

Description automatically generated with medium confidence**Plot Two-Dimensional Distributions**

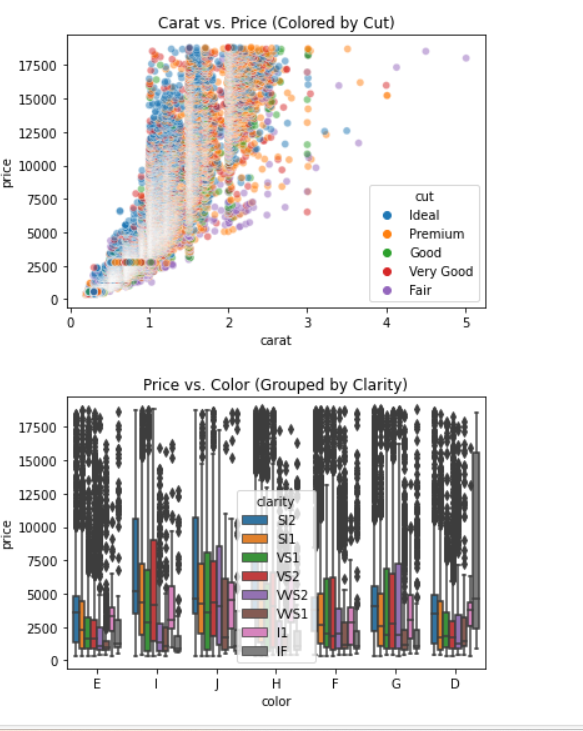


**Analyze How the Dependent Variable Changes Across Independent Variables**



****

**Identify Interesting Interactions**



**Carat vs. Price (Colored by Cut)**

* Cut Quality Effect:
  + Ideal and Premium cuts are densely clustered and often associated with higher prices at similar carat values.
  + Fair and Good cuts are spread more widely and seem to dominate the lower price ranges.

Stepped Pricing: Similar to before, the price does not increase smoothly, suggesting market-driving pricing categories.

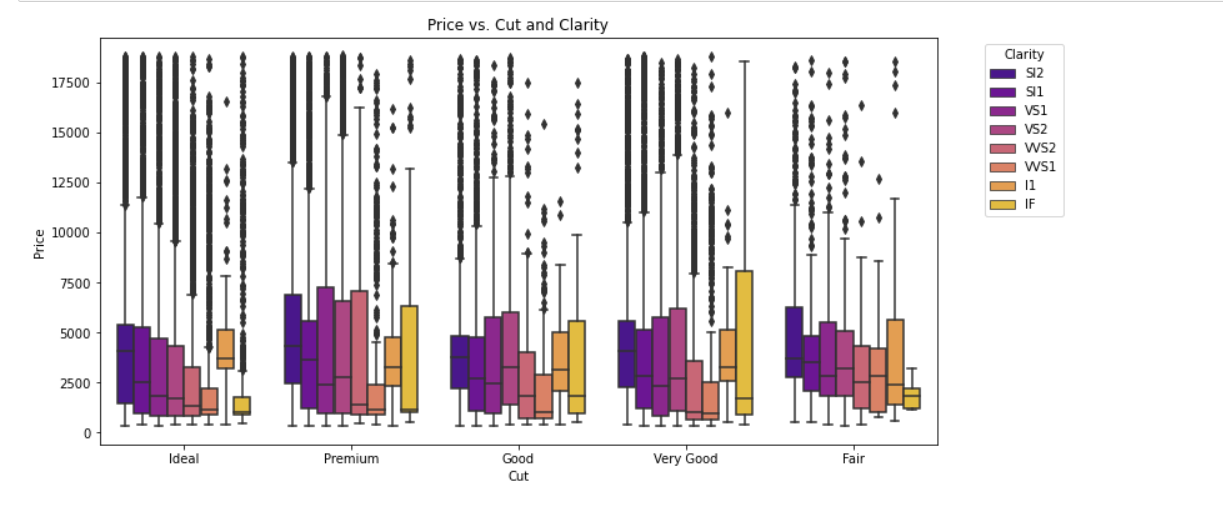
**Price vs. Color (Grouped by Clarity)**

* Price Variability:
  + Prices are highly scattered, suggesting that color alone does not strongly dictate price.
  + D-color diamonds (least yellow, most colorless) generally have higher median prices.
* Clarity Effect:
  + Diamonds with higher clarity grade (IF) generally appear in the higher price range.
  + Lower clarity diamonds (SI2, SI1, VS1) show a broader price distribution.

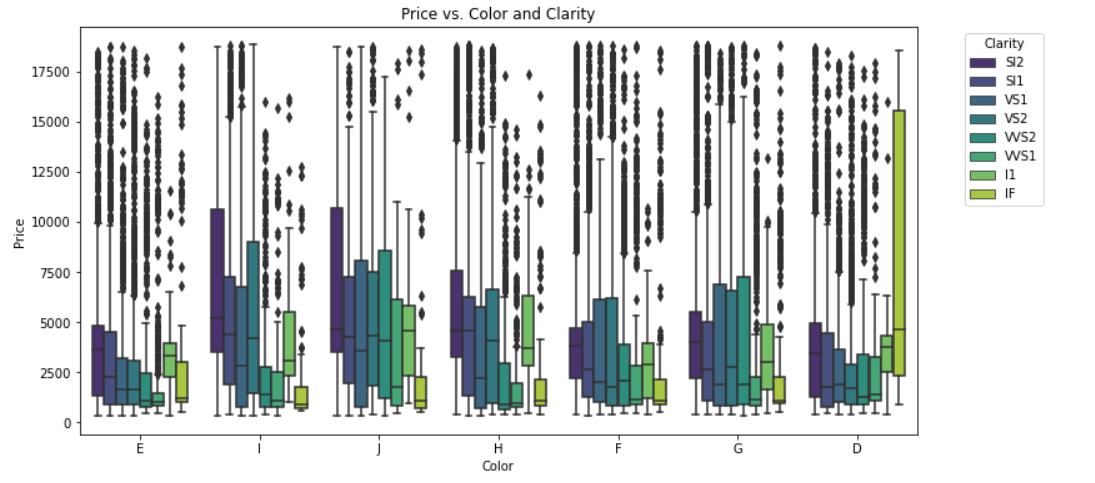
A chart showing a graph of colored dots

Description automatically generated with medium confidence

* Color plays a role but is secondary to carat weight in determining price.
* At lower carat weights, color variation leads to noticeable price differences.
* In higher carats, clarity and cut likely outweigh color as key price determinants.



* The combination of high-quality cut and clarity results in the highest prices.



* The combination of high-quality color and clarity results in the highest prices.

**Analyze Trends and Hypotheses**

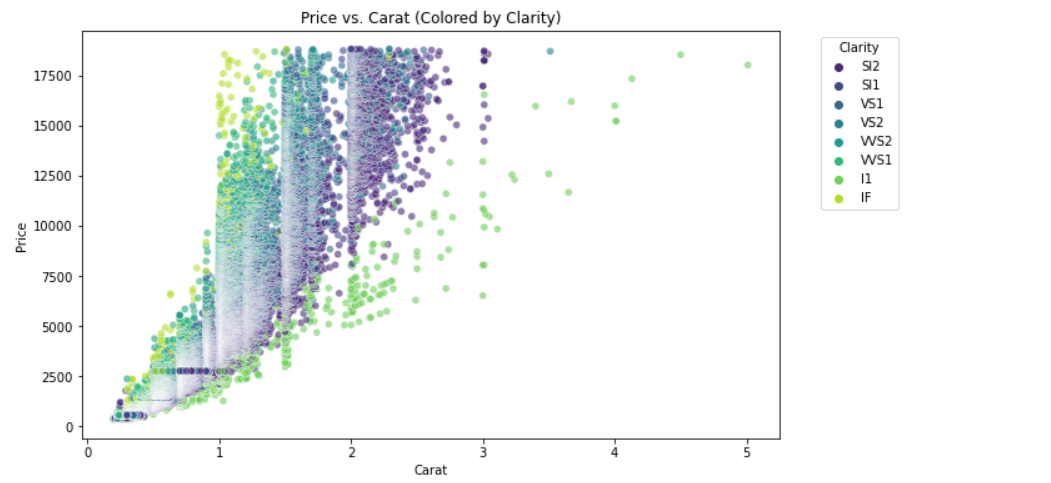
Based on the visualizations and group-by analysis, we can interpret trends and see if they support or contradict the problem statement.

**Observations:**

1. **Carat vs. Price**:
   * There is a strong positive correlation between carat and price. As carat increases, price generally increases.
   * However, the relationship is not perfectly linear, suggesting other factors (e.g., cut, clarity) also influence price.
2. **Price by Cut**:
   * Surprisingly, diamonds with better cuts (e.g., Ideal) tend to have lower average prices compared to lower-quality cuts (e.g., Premium, Very Good).
   * This could be because higher-cut diamonds are often smaller in carat size, which offsets the price.
3. **Price by Color**:
   * Diamonds with better color grades (e.g., D, E) tend to have higher prices.
   * This aligns with the expectation that higher-quality diamonds are more expensive.
4. **Price by Clarity**:
   * Diamonds with better clarity (e.g., IF, VVS1) tend to have higher prices.
   * This supports the hypothesis that clarity is a significant factor in determining price.
5. **Interactions**:
   * The interaction between carat and cut shows that while higher carat diamonds are generally more expensive, the cut quality can moderate this relationship.
   * The interaction between color and clarity reveals that diamonds with both high color and clarity grades command the highest prices.

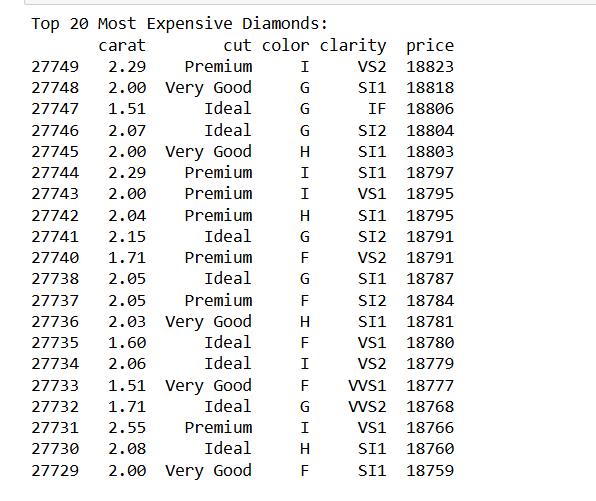
**Trends and Hypotheses:**

* **Hypothesis**: Higher-quality diamonds (better cut, color, clarity) are more expensive.
  + **Support**: The data supports this hypothesis for color and clarity, as higher grades in these categories are associated with higher prices.
  + **Contradiction**: The hypothesis is partially contradicted for cut, as higher-cut diamonds (e.g., Ideal) tend to have lower average prices. This could be due to the trade-off between cut quality and carat size.
* **Interesting Insight**: The interaction between carat and cut suggests that while carat size is a dominant factor in price, cut quality can moderate this relationship. Smaller, higher-cut diamonds may be priced lower than larger, lower-cut diamonds.



**Observations:**

* Diamonds with higher clarity grades (e.g., IF, VVS1, VVS2) tend to have higher prices.
* Clarity appears to have a significant influence on price.
* Some diamonds at **similar carat sizes** show **significant price variations**, possibly due to differences in cut, color, or brand influence.



**1. Carat Size:**

* The carat sizes of these diamonds range from **1.51 to 2.55**.
* Most of the top 20 diamonds are **above 2.00 carats**, indicating that larger diamonds tend to be more expensive.
* However, there are a few diamonds in the **1.5 to 1.7 carat range** that are also among the most expensive, suggesting that other factors (e.g., clarity, color) play a significant role in determining price.

**2. Cut Quality:**

* The cuts of these diamonds include **Ideal**, **Premium**, and **Very Good**.
* **Ideal** cuts are the most common among the top 20, appearing **8 times**.
* **Premium** cuts appear **7 times**, and **Very Good** cuts appear **5 times**.
* This suggests that **cut quality** is an important factor, but it is not the sole determinant of price.

**3. Color Grade:**

* The color grades of these diamonds range from **D (best)** to **I**.
* The most common color grades are **G** (appearing **6 times**) and **H** (appearing **5 times**).
* Surprisingly, **D-color** diamonds (the highest grade) do not appear in the top 20, indicating that **color alone** is not the primary driver of price for these diamonds.

**4. Clarity Grade:**

* The clarity grades of these diamonds range from **IF (Internally Flawless)** to **SI2 (Slightly Included 2)**.
* The most common clarity grades are **SI1** (appearing **8 times**) and **VS2** (appearing **4 times**).
* Only **1 diamond** has an **IF** clarity grade, and **2 diamonds** have **VVS1/VVS2** grades.
* This suggests that **clarity** is important but not the sole determinant of price, as diamonds with lower clarity grades (e.g., SI1, SI2) can still be among the most expensive.

**5. Price:**

* The prices of these diamonds are all very close, ranging from **18,759*to*18,823**.
* This indicates that there is a **price ceiling** in this dataset, and the top 20 diamonds are clustered near this ceiling.

The Notebook Link:

https://colab.research.google.com/drive/1FvLcuxvDC125aj4Gglq26TShEFyM-1J9?usp=sharing