Diamond Attribute Insights: Unveiling Value and Quality

Data Curation by Mariam Alquraan

**General Dataset Information:**

**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)

File Name: Diamonds.csv 🡪**The largest diamond dataset currently on Kaggle**

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

Dataset Details: 219,703 Rows & 26 Columns

Size: 35.43 MB

Source: Kaggle [The largest diamond dataset currently on Kaggle](https://www.kaggle.com/datasets/hrokrin/the-largest-diamond-dataset-currely-on-kaggle)

**File Name:**  diamonds (cleaned).csv **🡪 Diamond Online Marketplace**

**Description:** It also includes attributes like certification type, proportions, and symmetry, making it a great resource for analyzing how various factors influence diamond prices.

**Dataset Details**: 6,486 Rows & 18 Columns

**Size**: 737.75 kB

**Source**: Kaggle [Diamond Online Marketplace 💍](https://www.kaggle.com/datasets/beridzeg45/diamonds-prices-prediction)

**Data Profile:**

* Check the first and the last five rows in the Dataframe for a general insight.
* Get the total number of rows and columns (53,940 rows & 10 columns).
* Examine information about the null values and data type of each column using .info() function.
* Investigate the number of unique values in each column using the .nunique() function
* Check the total number of null values in each column using .isnull().sum()
* Get statistics of all columns (numerical and categorical) using .describe(include= 'all')
* Relationship discovery using .corr()

Observations:

* The columns don’t have null values, but x, y and z columns have zero values.

x column zero values: 8

y column zero values: 7

z column zero values: 20

* cut, color and clarity columns are of type object, but they should be transformed into a string data type.
* The cart and price columns have strong relations.
* For the second and third datasets they have commons columns with the first dataset.

# Data Wrangling:

For the first dataset:

1. Missing Data: No null values were present in the original dataset.
2. Zero Values in Dimensions:

* Rows with zero values in x, y, or z columns were removed (after removed 53,920 rows × 10 columns)
* This prevents potential issues in further analysis or modeling

1. Outlier Detection:

* Used Interquartile Range (IQR) method to identify outliers
* Provides count of outliers in each numeric column
* No need to remove all the outliers (just remove from z)

1. Data Type Conversion:

* Converted 'cut', 'color', and 'clarity' to string type

1. Add new column price\_per\_carat.
2. Add a binary column for high-quality diamonds based on specific thresholds for cut, color, and clarity is\_high\_quality.
3. Add length\_width\_ratio column to determine the shape characteristics.
4. Add volume column to calculate the volume of the diamond based on its dimensions (x, y, z).
5. Add is\_symmetric column to check if the diamond is symmetric by comparing x and y dimensions.
6. Add price\_category to create price ranges (bins) for better segmentation.

Note: I am working with the first dataset because the second and third datasets have the same columns as the first, along with additional columns. Since I plan to use only a subset of the features, I decided not to spend extra time cleaning them.

Open the notebook in Colab here: https://colab.research.google.com/drive/1PLM0YR20NxcZiJc90lwIutRzCDSvIZR7?usp=sharing

* **Data Table Schema:**

Table :Diamonds price Dataset

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Description |
| cart | float64 | Diamond weight in carat |
| cut | string | Diamond cutting quality |
| color | string | Diamond color from J (worst) to D (best) |
| clarity | string | A measure of diamond clarity *(from left to right is worst to best: I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF)* |
| depth | float64 | Percentage depth that is equal to z / mean(x,y) |
| table | float64 | The width of the widest point at the top of the diamond |
| price | Int64 | Diamond price |
| x | float64 | Diamond length in mm |
| y | float64 | Diamond width in mm |
| z | float64 | Diamond depth in mm |