Exercise: Diamonds Data — Detecting and Cleaning Duplicates

Objective

1. Download a real dataset from the web.
2. Inspect and explore its structure.
3. Detect and remove duplicate rows.
4. Perform simple descriptive statistics.
5. Visualize relationships between variables

Step 1 — Load the dataset

Dataset URL:

url = https://raw.githubusercontent.com/mwaskom/seaborndata/master/diamonds.csv

Step 2 — Inspect the data

diamonds.head()

diamonds.shape

diamonds.info()

diamonds.isnull().sum()

Step 3 — Check for duplicate

diamonds.duplicated().sum()

Step 4 — Basic statistics

1) Average price by cut
2) Average carat by color
3) Correlation between carat and price
4) Minimum and maximum price overall

Step 5 — Visualizations

1) Histogram of price
2) Boxplot of price by cut
3) Scatterplot of carat vs price

Step 6 — Interpretation

- Does removing duplicates change the averages or plots?
- What is the general relationship between carat and price?
- Which cut has the highest average price?