Task 5: Exploratory Data Analysis (EDA)

Objective: Extract insights using visual and statistical exploration.

Tools: Python (Pandas, Matplotlib, Seaborn)

Observations for Each Visual:

1. .describe(), .info(), .value_counts()

- **.describe() gives statistical summary:**
  - Score: The mean is high, indicating strong average performance.
  - Rank: Distributed from 1 to 23 (top participants).
  - SubmissionCount: Ranges from 1 to 46, with a few participants submitting more frequently.
- **.info() shows:**
  - No missing values.
  - Score is a float; other numerical fields are integers.
- **SubmissionCount.value_counts() shows:**
  - Most participants submitted only once.
  - A few participants submitted 13, 21, or 46 times, showing high effort or experimentation.

2. Pairplot: sns.pairplot(df[['Rank', 'Score', 'SubmissionCount']])

Observation:

- Rank vs Score: Clear negative correlation — higher score results in a better (lower) rank.
- SubmissionCount does not show strong correlation with either Score or Rank.
- Confirms that more submissions do not guarantee better performance.

3. Heatmap: sns.heatmap(...corr())

Observation:

- Rank and Score have a strong negative correlation (~ -0.99), meaning Rank improves with higher Score.

- SubmissionCount shows little to no correlation with Score or Rank.
- This reinforces the insight that performance outweighs persistence.

4. Scatter Plot: sns.scatterplot(x='Rank', y='Score')

Observation:

- Clear downward slope.
- Top-ranked teams achieved very high scores.
- Indicates that Score is the main driver for leaderboard Rank.

5. Boxplot (Score by Submission Count): sns.boxplot(x='SubmissionCount', y='Score')

Observation:

- Scores vary across different submission counts.
- Participants with few submissions can score just as high as those with many.
- No consistent upward trend in score with submission frequency.
- Some outliers present in low submission groups.

6. Histogram of Scores: df['Score'].hist()

Observation:

- Distribution is left-skewed, with most participants scoring near the top end.
- Indicates close competition at the top.

7. Single Boxplot of Score: sns.boxplot(y='Score')

Observation:

- Narrow interquartile range — most scores are close together.
- A few low-score outliers.
- Confirms a tight scoring distribution, with top scores being competitive.

Summary of Findings:

1. Score is the key determinant of Rank — high scores directly lead to better rankings.

2. SubmissionCount does not strongly influence Score — quality of submission matters more than quantity.
3. Most teams score quite high, as shown in the score histogram and boxplot.
4. There is a strong correlation between Rank and Score (negative), and weak/no correlation between SubmissionCount and performance.
5. Efficient teams (fewer submissions, higher scores) perform equally well or better than high-frequency submitters.
6. Visual analysis confirms the competition is tight at the top, with only a few underperforming teams.