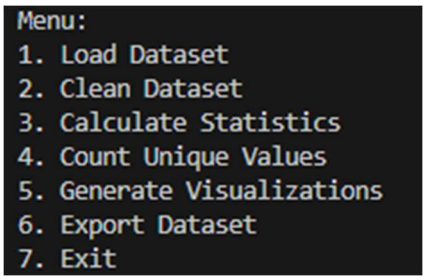
CODE REPORT

The provided an organized, user-friendly interface for carrying out crucial data analysis activities is provided by Python script, a mini data analytics platform. The script manipulates and visualizes data using the pandas, matplotlib, and seaborn libraries. A thorough explanation of the code's elements and operation may be found below.



Key Elements of the Code

Library Imports

- pandas: Utilized for manipulating and analyzing data.
- matplotlib.pyplot: Employed for generating static, animated, and interactive visualizations.
- seaborn: Used to improve data visualizations with more visually appealing and informative charts.
- os: Facilitates file system interactions, although this functionality is not utilized in the present script.

Function Definitions The script outlines multiple modular functions, each designed for a specific aspect of data analysis. These are described below:

1. **clean_dataset(dataset)**

Objective: Prepares the dataset by addressing missing entries and eliminating duplicates.

- Method:
- Missing Entries:
- Replaces missing values in numerical columns with the average.
- Replaces missing values in categorical columns with the most frequent value.
- Duplicates: Eliminates repeated rows.
- Result: A refined version of the dataset.
- Characteristics:

- Utilizes inplace operations for better memory efficiency.
- Displays information about the dataset after the cleaning process.
- Improvements:
- Could provide options for users on how to manage missing values (e.g., remove rows, fill with median).

```
RangeIndex: 2899 entries, 0 to 2898
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Unnamed: 0              2899 non-null   int64
 1   Batter                  2899 non-null   object
 2   Country                 2899 non-null   object
 3   No. of Innings          2899 non-null   int64
 4   No. of Not Outs         2899 non-null   int64
 5   Total Runs              2899 non-null   int64
 6   Total Balls Faced       2899 non-null   int64
 7   No. of Outs             2899 non-null   int64
 8   Batting Average         2899 non-null   float64
 9   Strike Rate             2899 non-null   float64
 10  Balls Faced per Innings 2899 non-null   float64
 11  Balls Faced per Wicket  2899 non-null   float64
dtypes: float64(4), int64(6), object(2)
memory usage: 271.9+ KB
None
```

2. **load_dataset(file_path)**

Objective: Imports a dataset from a designated CSV file.

- Input: file_path (str) - Location of the CSV file.
- Output: A pandas DataFrame that contains the dataset, or None if an error occurs.
- Features:
- Displays the initial 10 rows of the dataset.
- Manages errors effectively using a try-except block.
- Improvements:
- The function might incorporate checks to confirm the file's existence or prompt the user to enter a valid path.

```
Enter your choice: 1
Dataset loaded successfully!
   Unnamed: 0       Batter        Country  ...  Strike Rate  Balls Faced per Innings  Balls Faced per Wicket
0           0   A Ahmadhel       Bulgaria  ...    33.333333                 6.000000                6.000000
1           1      A Amado         Israel  ...    67.857143                14.000000               14.000000
2           2    A Andrews    Switzerland  ...    66.666667                 6.000000                6.000000
3           3    A Ashokan  Czech Republic  ...   115.565032                22.333333               31.266667
4           4      A Bagai         Canada  ...   107.142857                38.500000               51.333333
5           5  A Balbirnie        Ireland  ...   125.989783                18.000000               18.642857
6           6  A Bhadauria     Luxembourg  ...   129.032258                15.500000               15.500000
7           7    A Bhagwat      Hong Kong  ...    22.222222                 3.000000                3.000000
8           8   A Bhargava      Singapore  ...   116.666667                 3.000000                6.000000
9           9  A Bhattarai          Nepal  ...    66.666667                 3.000000                3.000000

[10 rows x 12 columns]
```

3. compute_statistics(data)

Objective: Computes fundamental statistics (average, middle value, standard deviation) for numerical columns.

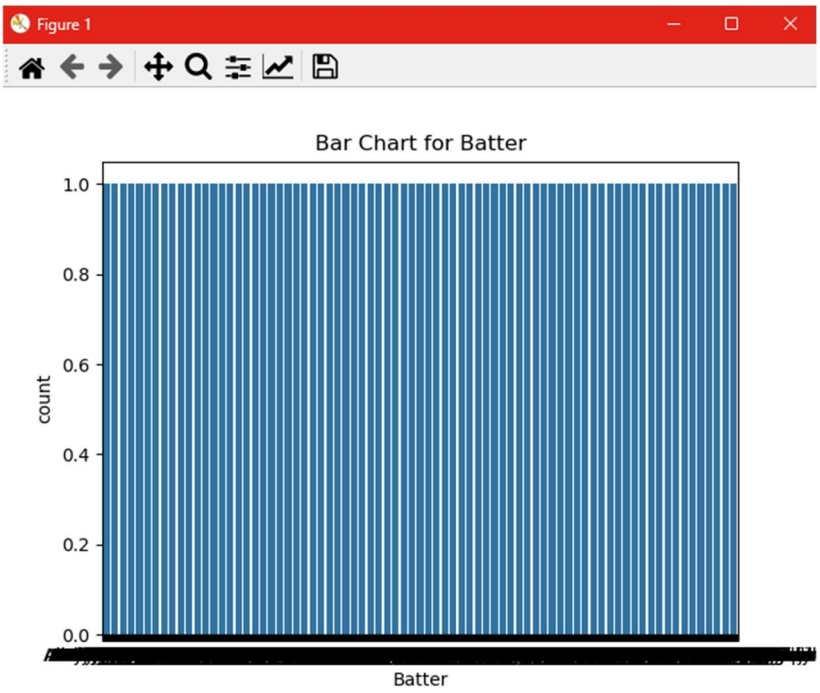
- Method: Loops through numeric columns and calculates the necessary statistics.
- Result: Displays statistics for every numeric column.
- Improvements:
- Could store the statistics in a file for later reference.
- Could incorporate visual representations of these statistics (e.g., boxplots).

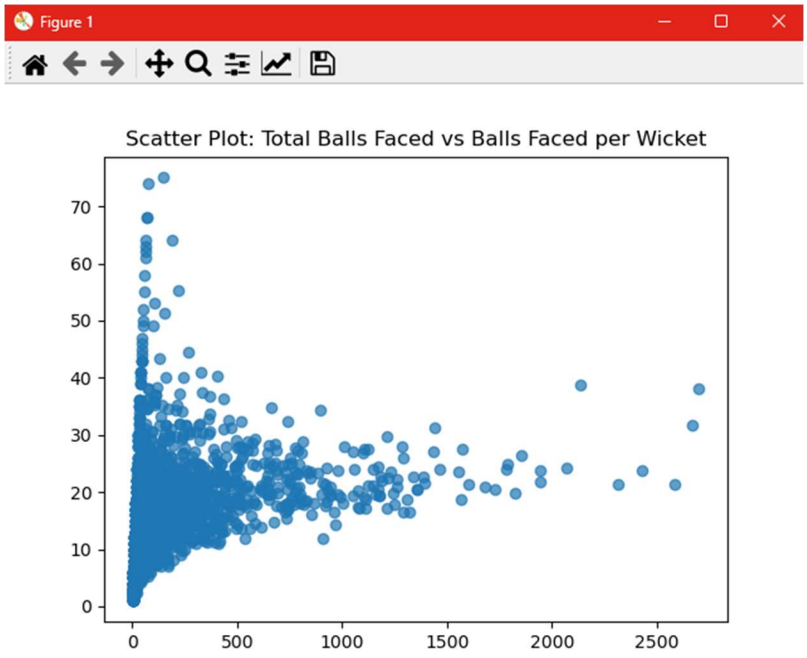
```
Menu:
1. Load Dataset
2. Clean Dataset
3. Calculate Statistics
4. Count Unique Values
5. Generate Visualizations
6. Export Dataset
7. Exit
Enter your choice: 3
Statistics calculated:
Unnamed: 0 -> Mean: 1449.0, Median: 1449.0, Std Dev: 837.0135403126204
No. of Innings -> Mean: 10.178337357709555, Median: 5.0, Std Dev: 14.367175604370242
No. of Not Outs -> Mean: 2.018282166264229, Median: 1.0, Std Dev: 3.1965750653000384
Total Runs -> Mean: 167.10037944118662, Median: 43.0, Std Dev: 345.02892117782244
Total Balls Faced -> Mean: 140.7499137633667, Median: 47.0, Std Dev: 262.6717451818266
No. of Outs -> Mean: 8.160055191445325, Median: 4.0, Std Dev: 12.20361154374188
Batting Average -> Mean: 14.153896153360813, Median: 12.2, Std Dev: 11.346766748633193
Strike Rate -> Mean: 95.00913442104587, Median: 100.0, Std Dev: 44.903864961039766
Balls Faced per Innings -> Mean: 10.772826412342187, Median: 9.65, Std Dev: 6.719508284154843
Balls Faced per Wicket -> Mean: inf, Median: 13.8, Std Dev: nan
```

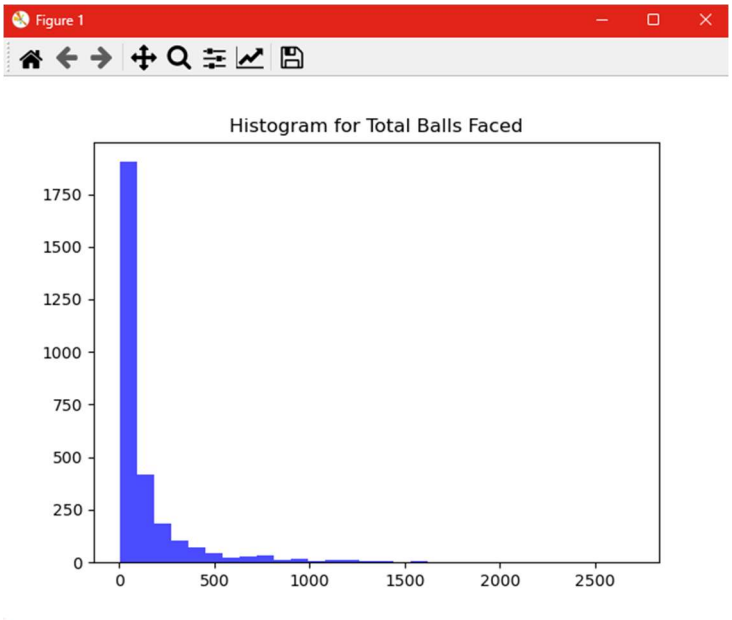
4. **generate_visualizations(dataset)**

Purpose: Provides interactive options to create visualizations.

- **Options**:
    1. Bar Chart for Categorical Columns.
    2. Histogram for Numeric Columns.
    3. Scatter Plot for Two Numeric Columns.
- **Process**:
    - Validates column names and types before creating plots.
    - Saves visualizations as .png files.
- **Output**: Displays plots and saves them for later use.
- **Features**:
    - User-friendly menu-driven interface.
    - Dynamic plot generation.
- **Enhancements**:
    - Could include more plot types (e.g., boxplots, heatmaps).
    - Might allow customization of plot parameters (e.g., colors, labels).







5. **export_dataset(dataset)**

Purpose: Exports the cleaned dataset to a user-specified CSV file.

- **Process**:
    - Prompts the user for a file name.
    - Saves the dataset using pandas.to_csv.
- **Output**: A CSV file with the cleaned dataset.

- **Enhancements**:
    - Could include a default file path or name if the user provides no input.

```
Menu:
1. Load Dataset
2. Clean Dataset
3. Calculate Statistics
4. Count Unique Values
5. Generate Visualizations
6. Export Dataset
7. Exit
Enter your choice: 6
Enter the file name to save the cleaned dataset (with .csv extension): batting_stats_T20I.csv
Dataset exported successfully as batting_stats_T20I.csv
```

```
Menu:
1. Load Dataset
2. Clean Dataset
3. Calculate Statistics
4. Count Unique Values
5. Generate Visualizations
6. Export Dataset
7. Exit
Enter your choice: 7
Exiting the program. Goodbye!
PS C:\Users\xlish\Downloads>
```