ENSF592 Project Group 11 Report

Dr. Emily Marasco
Balkarn Gill, Yajur Vashist
Tuesday June 13, 2023

The data set chosen for this project is a collection of Formula 1 racing statistics across the 2020, 2021, and 2022 seasons [1]. The data is stored across three separate CSV files, each corresponding to a different year. The data within each CSV file includes information on the race track, year, driver, team, and a variety of performance statistics. Furthermore, another spreadsheet representing driver ratings is included in the analysis.

The datasets for the different years were combined to create a comprehensive and continuous dataset. The combined dataset was then refined to include only races where drivers scored points. Additionally, a 'Positions Gained' column was added, representing the difference between the starting position and the final position of each driver for each race.

The solution offers an interactive, console-based user interface, where users can choose to view specific statistics or export the combined dataset to a CSV file. Here is a brief summary of the options and their respective outputs:

- Option 1 (Team Stats): Users input a specific team name. The program then prints the total points for the specified team for each year.
- Option 2 (Driver Stats): Users input a driver's name. The program then prints the total points for the specified driver for each year, and also displays a plot of points over the years using Matplotlib.
- Option 3 (Driver Ratings): Users input a driver's name. The program then prints the driver's ratings for all years, if available.
- Option 4 (General Stats): Users choose between printing a summary of all stats (sub-option 1) or a pivot table of all teams and their respective points (sub-option 2).
- Option 5: The program exports the combined dataset as a CSV file, 'F1_Summary.csv'.
- Option 0: The program terminates.

In the event of an invalid input, the system is designed to persistently solicit user input until a valid response is provided. This iterative approach ensures the mitigation of user errors and maintains the integrity of the input-output operations within the program.

The program meets all the specified requirements. This includes the usage of pandas functions such as describe, groupby, and masking operations for data analysis. It has user-defined functions with proper docstring documentation. It also features the use of Matplotlib for data visualization and generates an output CSV file.

The solution efficiently meets all of the specifications provided in the rubric. The merged dataset is well above the size requirement and the program does not modify the original data files and doesn't hard-code any data values, instead only referencing column names.

Three separate CSV files are imported into Pandas DataFrames and merged using the 'concat' operation. Duplicate rows arising from the merge are deleted. The dataset is multi-indexed on five levels: 'Track', 'Year', 'Driver', 'Team', 'No', and sorted by 'Year'. Null values are also handled as rows with NaN 'Points' are removed. All these operations take place within the main function, ensuring no global variables are used. Furthermore, an additional merge operation is implemented which integrates the primary dataset with a supplementary one that encompasses driver ratings. The merging process utilizes 'Year' and 'Driver' columns as common identifiers. As a result, a new comprehensive dataset is formed, amalgamating driver ratings with the corresponding racing statistics.

The program includes an interface that allows users to search based on the team or driver names. The user provides the required information through a numerical menu and additional inputs, following clear instructions given by the program. If an invalid entry is given, the error is handled using try/except statements, and the user is continuously prompted until valid input is given. All output information is clearly defined using printed headers or sentences.

The program uses the 'describe' method to print aggregate stats for the entire dataset. Two new columns, 'Team Points' and 'Positions Gained', are added to the dataset. Aggregation computation is done to calculate the sum of team points at each track. A masking operation is performed to remove rows where drivers did not score points. The 'groupby' operation is used to calculate team points at each track and total points for a driver across years. The program also includes a pivot table showing total points scored by each team each year. There are seven user-defined functions in the program.

The final DataFrame is exported to a CSV file in the working directory, complete with index and header values. The program uses matplotlib to create a plot of a driver's points over the years, satisfying the requirement for a graphical representation of data. The plot, however, is not saved as a .png file as per the provided code, which is an area that could be improved.

References:

[1] D. Y. Y. Y. Yenigun, “Touppercase78/formula1-datasets: Datasets &amp; Analyses for Formula 1 world championship,” GitHub, https://github.com/toUpperCase78/formula1-datasets (accessed Jun. 12, 2023).