**1. Project Diary Entry - [05/10/2023 - 12/10/2023]**

Team Members : Ashly Thomas Panangadan, Ans Mariya Joy

Task : Scraping and Wrangling Data from Formula1 Racing Website using Julia

Description :

Worked on the scraping and wrangling of data from the Formula1 racing website. Our goal is to collect data on Grand Prix races, race results, drivers, and starting positions for the years 2012 to 2022.

* We started by defining the URL of the Formula1 racing website, "https://www.formula1.com/en/results.html" and specifying the range of years we wanted to scrape, which is from 2012 to 2022.
* We created arrays to store data related to Grand Prix names, race result links, winners, winner IDs, race years, and winner starting positions. We also initialized an empty DataFrame named 'df.'
* For each year in the specified range, we implemented the 'scrape\_data' function. In this function :
* We constructed the URL for the specific year's race results page.
* Sent an HTTP GET request and parsed the HTML content using Gumbo.
* Extracted data from the table, such as the Grand Prix name and links to race results.
* For each Grand Prix, we followed the link to the race results page, retrieved the winner's driver ID, and added data to the arrays.
* For each winner, we took the starting position from the starting grid page using driver ID to identify the row.
* After scraping data for all years, we created a DataFrame 'df' containing columns for Grand Prix names, race years, driver IDs, driver names, and starting positions.
* We used the 'starting\_positions\_counts' function to count the occurrences of each starting position for each Grand Prix.
* The most common starting position for each Grand Prix was determined using the 'most\_common\_position' function.
* We grouped the DataFrame by 'Grand\_prix' and displayed a resulting DataFrame showing the common starting position and the number of occurrences for each Grand Prix.
* We defined a threshold value (3) and filtered the data to select Grand Prix races where the common starting position occurred more than the threshold times.

**2. Project Diary Entry - [13/10/2023 - 18/10/2023]**

Team Members : Ashly Thomas Panangadan, Ans Mariya Joy

Task : Data Analysis - Common Starting Positions in Formula1 Grand Prix Races

Description :

We focused on analyzing the data collected from the Formula1 racing website. Our aim was to identify the common starting positions for Grand Prix races and highlight any exceptions. This analysis provides insights into the starting positions that are most frequently associated with race winners.

* We started by analyzing the data to find the most common starting position for each Grand Prix race. Our analysis revealed that the majority of Grand Prix races had the starting position 1 as the most common. However, we identified exceptions where the most common starting position was not 1 but 2 as well.
* To perform this analysis, we used the data collected in the earlier stages of the project, where we scraped and cleaned data on Grand Prix races, winners, and their starting positions.

Notes:

The analysis revealed that, in most Grand Prix races, starting position '1' was the most common, with the exception of a few races where starting position '2' was more common. This insight can be valuable for understanding race dynamics and potentially developing race strategies in the sport of Formula1.

Report :

1. Introduction

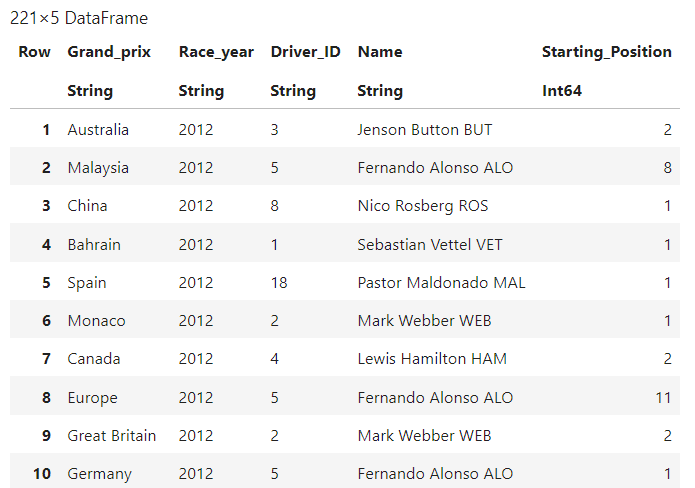
One of the fascinating aspects of Formula1 racing is the influence of the starting position on the race outcome. The Julia part of this project aims to explore and analyze the historical data to identify patterns related to the starting positions of race winners. The report has the following key steps :

* Data collection through web scraping.
* Analysis of common starting positions.

2. Methodology

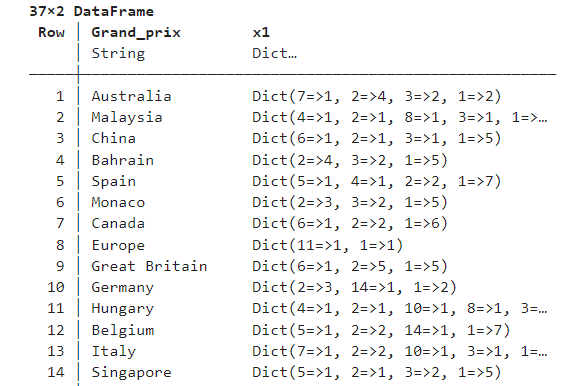
2.1 Data Collection

The data collection process involved scraping race results and starting grid information from the Formula1 website (https://www.formula1.com/en/results.html) for the years 2012 to 2022. The collected data included race year, grand prix, driver IDs, driver names, and starting positions. The Julia programming language, in combination with the HTTP, Gumbo, Cascadia and DataFrames packages, facilitated data retrieval. Upon data collection, the project leveraged Julia's capabilities for data manipulation. A DataFrame was created by collecting the winner's driver ID from the race\_result table and using this to identify their corresponding starting position from the starting\_grid table for analysis. Starting positions were identified for each race winner.



2.2 Data Analysis : Common Starting Positions

The data frame was grouped based on each country to find the frequency of each starting position that won the race. For each country, there is a dictionary with the starting position and its occurrences. This analysis focused on identifying the most common starting positions for race winners across various Grand Prix races.



3. Results

The analysis revealed that starting position '1' was the most common starting position for race winners across Grand Prix races. This dominance can be attributed to the strategic advantage of the pole position. However, there were exceptions where starting position '2' was more prevalent, indicating that factors beyond the starting position may influence race outcomes. The filtered results, focusing on races with starting position '1' and '2,' are presented below :

