**Data Wrangling Project Proposal**

Jack Wilson | 4/23/2025

**1. Introduction**

Formula 1 (F1) is a motorsport competition in which ten teams (constructors), each with two drivers, compete against each other in a series of races to earn points that determine individual and team standings.

The points system has evolved since the sport's introduction in 1950 and follows this timeline:

[List of Formula One Championship Scoring](https://en.wikipedia.org/wiki/List_of_Formula_One_World_Championship_points_scoring_systems#:~:text=In%202003%2C%20the%20FIA%20revised%20the%20structure,1991%20and%202009%2C%20and%2025%20since%202010.)

|  |  |  |
| --- | --- | --- |
| 1950 – 1959 | Top 5 finishers score | Winner earns 8 points |
| 1960 – 1990 | Top 6 finishers score | Winner earns 9 points |
| 1991 – 2003 | Top 6 finishers score | Winner earns 10 points |
| 2003 – 2009 | Top 8 finishers score | Winner earns 10 points |
| 2010 – Present | Top 10 finishers score | Winner earns 25 points |

The race winner receives the highest number of points while decreasing exponentially for lower finishing positions. These points are added to the driver’s individual total and their team’s total. After all the races are completed (~20), the driver with the most points wins the Drivers’ Championship, and the constructor with the most points wins the Constructors’ Championship.

Formula 1 is very dynamic. Drivers are constantly switching between teams, and old constructors are constantly being replaced by new ones. The competitive nature of F1 makes data analysis crucial for understanding performance trends, driver consistency, and team dominance.

This project aims to compile and analyze historical F1 data to uncover performance trends among drivers and teams. By integrating data from multiple sources, this analysis will provide a comprehensive look at how driver and constructor performance has evolved.

**2. Data**

This project utilizes scraped data from the official Formula 1 site along with four tables from a Kaggle dataset to build a comprehensive data frame.

Web Scraped Data:

* [Driver Performance Data for each Year (1950-2024)](https://www.formula1.com/en/results/2025/drivers):
  + Includes driver name, season position, season points, and team information.
* [Constructor Performance Data for each Year (1958-2024)](https://www.formula1.com/en/results/2025/team):
  + Includes team name, team position, and team points information.

[Kaggle Dataset](https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020?select=driver_standings.csv):

* drivers.csv
  + Includes driver name, code, date of birth, and nationality information.
* constructors.csv
  + Includes constructor name and nationality information.
* constructor\_standings.csv
  + Includes team points, team position, constructorId and raceId needed for merging with races.csv and later with scraped data
* races.csv
  + Includes year and round needed for merging with scraped data

Initially, the Kaggle data was integrated with the scraped data using team names, but mismatches arose due to variations in naming (e.g., McLaren Mercedes vs. McLaren). To resolve this, additional data frames were merged to create a unique set of columns—year, team position, and team points—to ensure proper integration with the scraped data.

These datasets will be horizontally integrated into a single dataset that includes driver rankings, team rankings, and additional attributes such as nationalities and personal information. The final dataset consists of 1,636 rows and 11 features, covering driver information from 1950 to 2024 and team information from 1958 to 2024, as the Constructors’ Championship began in 1958.

**2.1 Data Dictionary**

|  |  |  |
| --- | --- | --- |
| Field | Type | Description |
| year | Numeric | Year a season happened |
| driver\_name | Text | Full name of driver |
| driver\_nationality | Text | Nationality of driver |
| age | Numeric | Age of driver |
| individual\_points | Numeric | Total individual driver points for season |
| individual\_position | Numeric | Overall individual position for season |
| team\_name | Text | Team name |
| team\_nationality | Text | Nationality of team |
| team\_points | Numeric | Total team points for season |
| team\_position | Numeric | Overall team position for season |

**3. Proposed Analysis**

This project will focus on descriptive analytics, visualizations, and trends to answer research questions like:

* How have driver and constructor rankings correlated over time?
  + Is one team or driver consistently dominant?
  + If so, how could we level the playing field to give other teams success?
  + Line chart visualizations to compare driver/constructor ranks over multiple seasons
* Do certain nationalities dominate Formula 1?
  + F1 is based in Britain, so is there a surplus of British drivers?
  + If so, what ways could we branch out to acquire new and international talent?
  + Bar chart visualizations to see the number of drivers of each nationality
* What is the impact of age on driver performance?
  + Is there a common age where drivers are in their prime?
  + Is it better to hire an unstable young rookie or a steady tenured driver?
  + Scatter plot or line chart visualizations to compare drivers' success across their career