# F1 Dataset Analysis

Gopinath Robba
A20543707

# Problem Statement

Through a comprehensive analysis, the objective is to uncover patterns, anomalies, and correlations within the dataset, providing valuable insights that contribute to a nuanced understanding of Formula 1 dynamics and contribute to informed decision-making in the realm of motorsport.

With this project we planned to answer the following:

- What are the key factors influencing race outcomes in Formula 1?
- How have driver performance metrics (e.g., lap times, overtakes) evolved over recent F1 seasons?
- Which circuits have the highest and lowest average race speeds, and what factors contribute to these differences?
- Regression to predict whether a driver will win the championship or predict the specific championship position based on historical performance.

Future Work:

We had to remove few of the tasks out of scope due to the non availability of data.

1. Is there any impact on Netflix stock after the “F1 Drive to Survive” release on netflix.
2. Is there any correlation between F1 races in a city and the flight traffic to that city.

To address these, further data(Netflix stocks data, Flight data) needs to be be merged to the f1 dataset.

# Project Timeline

## CSP-554: Project Planner : F1 Dataset analysis

Gopinath Robba

| ID | Task | Start Date | End Date | Dependency | Owner | | | Status | GANTT CHART Week | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Gopinath | Tarun | Snehitha | | 1(9/17) | 2(9/24) | 3(10/01) | 4(10/08) | 5(10/15) | 6(10/22) | 7(10/29) | 8(11/05) | 9(11/12) | 10(11/19) | 11(11/26) |
| **Project Envisioning Stage** | | | | | | | | | | | | | | | | | | | |
| 1 | Project Group and Topic Formation | 9/11 | 9/24 | | | | | Completed | | | | | | | | | | | |
| 2 | Project Proposal and Outline | 10/2 | 10/22 | | | | | Completed | | | | | | | | | | | |
| 3 | Meeting: Plan for project | | | | | | | Completed | | | | | | | | | | | |
| 4 | Project Plan and Details | 10/30 | 11/5 | | | | | Completed | | | | | | | | | | | |
| **Data Selection Stage** | | | | | | | | | | | | | | | | | | | |
| 5 | Identify Data Sources | 9/18 | 10/2 | | | | | Completed | | | | | | | | | | | |
| 6 | Data collection from Kaggle | 9/25 | 10/1 | | | | | Completed | | | | | | | | | | | |
| 7 | Data extraction from F1 website | 10/30 | 11/5 | | | | | Completed | | | | | | | | | | | |
| 8 | Review of dataset for quality | 9/25 | 11/6 | | | | | Completed | | | | | | | | | | | |
| 9 | Data Documentation | 10/2 | 11/8 | | | | | Completed | | | | | | | | | | | |
| 10 | Data extraction for netflix stock | 11/13 | 11/16 | | | | | Out Scoped | | | | | | | | | | | |
| 11 | Meeting: Rewrite project scope | 10/31 | 10/31 | | | | | Completed | | | | | | | | | | | |
| **Data Processing Stage** | | | | | | | | | | | | | | | | | | | |
| 12 | Data import into R-Studio | 10/2 | 10/5 | | | | | Completed | | | | | | | | | | | |
| 13 | Data Cleaning: Outliers, Missing values, Duplicates | 10/2 | 10/12 | | | | | Completed | | | | | | | | | | | |
| 14 | Meeting: Status of data cleaning | 10/18 | 10/18 | | | | | Completed | | | | | | | | | | | |
| 15 | Data Validations | 10/4 | 10/19 | | | | | Completed | | | | | | | | | | | |
| 16 | Data Aggregation | 10/10 | 10/20 | | | | | Completed | | | | | | | | | | | |
| 17 | Meeting: Issue with df creation | 10/10 | 10/10 | | | | | Completed | | | | | | | | | | | |
| 18 | Features Data | 10/10 | 10/17 | | | | | Completed | | | | | | | | | | | |
| 19 | Data Scrubbing from websites | 11/6 | 11/10 | | | | | Out Scoped | | | | | | | | | | | |
| **Data Transformation** | | | | | | | | | | | | | | | | | | | |
| 20 | Meeting: Next steps | 10/25 | 10/25 | | | | | Completed | | | | | | | | | | | |
| 21 | Creating new data frames with races df | 10/16 | 10/20 | 7 | | | | Completed | | | | | | | | | | | |
| 22 | Creating new data frames with results df | 10/16 | 10/20 | | | | | Completed | | | | | | | | | | | |
| 23 | Creating new data frames with driver standings df | 10/16 | 10/20 | | | | | Completed | | | | | | | | | | | |
| 24 | Meeting: results dataset issue | 10/30 | 10/30 | | | | | Completed | | | | | | | | | | | |
| 25 | Dropping unwanted columns from circuits, driver, results datasets | 10/16 | 10/22 | | | | | Completed | | | | | | | | | | | |
| 26 | Scale features: wins, race dataset: sprint runs | 10/25 | 10/29 | | | | | Completed | | | | | | | | | | | |
| 27 | Meeting: Next steps | 11/2 | 11/2 | | | | | Completed | | | | | | | | | | | |
| 28 | Normalizing datasets: status dataset | 10/25 | 10/30 | | | | | Completed | | | | | | | | | | | |
| 29 | Merging race dataset with player dataset | 10/25 | 10/31 | 21 | | | | Completed | | | | | | | | | | | |
| **Data Analysis** | | | | | | | | | | | | | | | | | | | |
| 30 | Create Hypothesis | 10/2 | 10/8 | | | | | Completed | | | | | | | | | | | |
| 31 | Test/Train splits of results | 10/31 | 11/5 | 7,29 | | | | Completed | | | | | | | | | | | |
| 32 | Test/Train splits of driver standings | 10/31 | 11/5 | 7,29 | | | | Completed | | | | | | | | | | | |
| 33 | EDA for datasets | 10/25 | 11/6 | 31,32 | | | | Completed | | | | | | | | | | | |
| 34 | Feature selection | 10/31 | 11/12 | 31,32 | | | | Completed | | | | | | | | | | | |
| 35 | Feedback and Iteration | 11/12 | 11/30 | | | | | Completed | | | | | | | | | | | |
| 36 | PCA for dimensionality reduction | 10/31 | 11/15 | 33,34 | | | | Completed | | | | | | | | | | | |
| 37 | Model (Linear and NonLinear) test with K-fold | 11/19 | 11/26 | | | | | Completed | | | | | | | | | | | |
| **Data Interpretation Stage** | | | | | | | | | | | | | | | | | | | |
| 38 | Insights | 11/19 | 11/26 | | | | | Completed | | | | | | | | | | | |
| 39 | Model comparison | 11/12 | 11/19 | | | | | Out Scoped | | | | | | | | | | | |
| 40 | Documentation/Project report completion and Review | 11/12 | 11/19 | | | | | Completed | | | | | | | | | | | |
| 41 | Presentation | 11/12 | 11/30 | | | | | Completed | | | | | | | | | | | |

| | |
|---|---|
| | Turbulence |
| | Safe |
| | Need call |
| | Removed from Scope |

# Project Overview

The objective of this F1 dataset analysis project is to rev up the search for the most exciting Formula 1 insights. This dataset encompasses a wide range of variables, such as driver statistics, team information, race results, circuit characteristics, and some measures unique to each race. Our goal in conducting this analysis is to find obscure trends, patterns, and insights in the racing industry. We analyze the data in an effort to comprehend the elements that affect race results, including driver skill, team tactics, and the effects of outside variables like weather and circuit features.

# Data

Below is overview of the datasets we used and the description of each along with their dimensions .

| File | Description | Dimensions (features x observations) |
|---|---|---|
| circuits.csv | Details of tracks | 9 x 79 |
| driver_standings.csv | Driver position details | 7 x 71000 |
| driver.csv | Driver details | 9 x 855 |
| lap_time.csv | Time of each lap by driver and race | 6 x 73000 |
| pit_stop.csv | Time interval of each pitstop | 7 x 6000 |
| qualifying.csv | Details of qualifying run | 9 x 9000 |
| races.csv | Race track and date details | 8 x 1000 |
| results.csv | Results of each main race | 18 x 25000 |
| sprint_results.csv | Results of sprint race | 16 x 60 |
| status.csv | Different race status | 2 x 139 |

# Data Preparation Steps

The dataset utilized for this Formula 1 analysis encompasses information spanning multiple years, capturing a comprehensive view of the sport until the year 2021. This temporal scope allows for a robust exploration of Formula 1 dynamics, including race outcomes, driver performances, and technical specifications over an extended period.

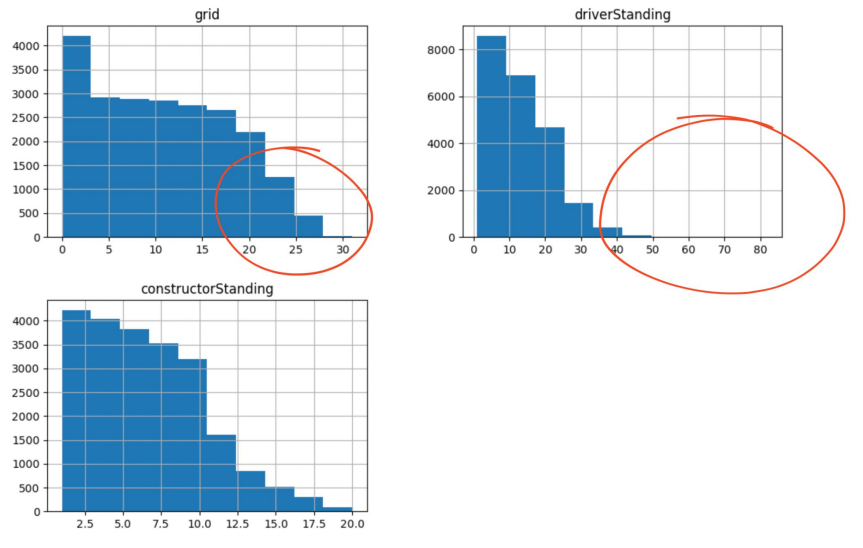
The report provides a comprehensive overview of the intricate data preparation and feature selection processes undertaken. At a high level, the following encapsulates some of the challenges encountered during the course of the analysis.

1. We had a conflict of column names causing incorrect result or the model use to fail. So, we dropped duplication columns to mitigate potential conflicts with the other datasets.
2. The grid positions presented values surpassing the defined limit of 20, a deviation from the established maximum capacity for cars and positions on the grid.
3. Entries with the 'position' value represented as '\n' are indicative of missing or not applicable (NA) values.

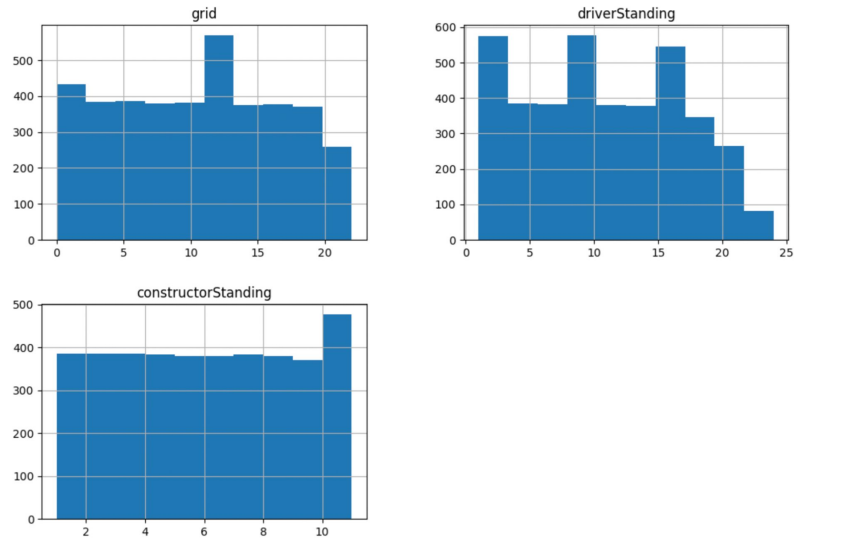
# Data Preparation Steps

Data issue visualization, before and after the fix.

Before:




After:



# Data Preparation Steps

Data Cleaning and Analysis Outcome: After the data cleaning process, we were now left with 3258 observations

Data before and after data cleaning:

Before: There are many unwanted columns and NA's in the dataset.

| | raceId | year | round | circuitId | name | date | time | url | fp1_date | fp1_time | fp2_date | fp2_time | fp3_date |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **878** | 880 | 2013 | 1 | 1 | Australian Grand Prix | 2013-03-17 | 06:00:00 | http://en.wikipedia.org/wiki/2013_Australian_G... | \N | \N | \N | \N | \N |
| **879** | 881 | 2013 | 2 | 2 | Malaysian Grand Prix | 2013-03-24 | 08:00:00 | http://en.wikipedia.org/wiki/2013_Malaysian_Gr... | \N | \N | \N | \N | \N |
| **880** | 882 | 2013 | 3 | 17 | Chinese Grand Prix | 2013-04-14 | 07:00:00 | http://en.wikipedia.org/wiki/2013_Chinese_Gran... | \N | \N | \N | \N | \N |
| **881** | 883 | 2013 | 4 | 3 | Bahrain Grand Prix | 2013-04-21 | 12:00:00 | http://en.wikipedia.org/wiki/2013_Bahrain_Gran... | \N | \N | \N | \N | \N |
| **882** | 884 | 2013 | 5 | 4 | Spanish Grand Prix | 2013-05-12 | 12:00:00 | http://en.wikipedia.org/wiki/2013_Spanish_Gran... | \N | \N | \N | \N | \N |

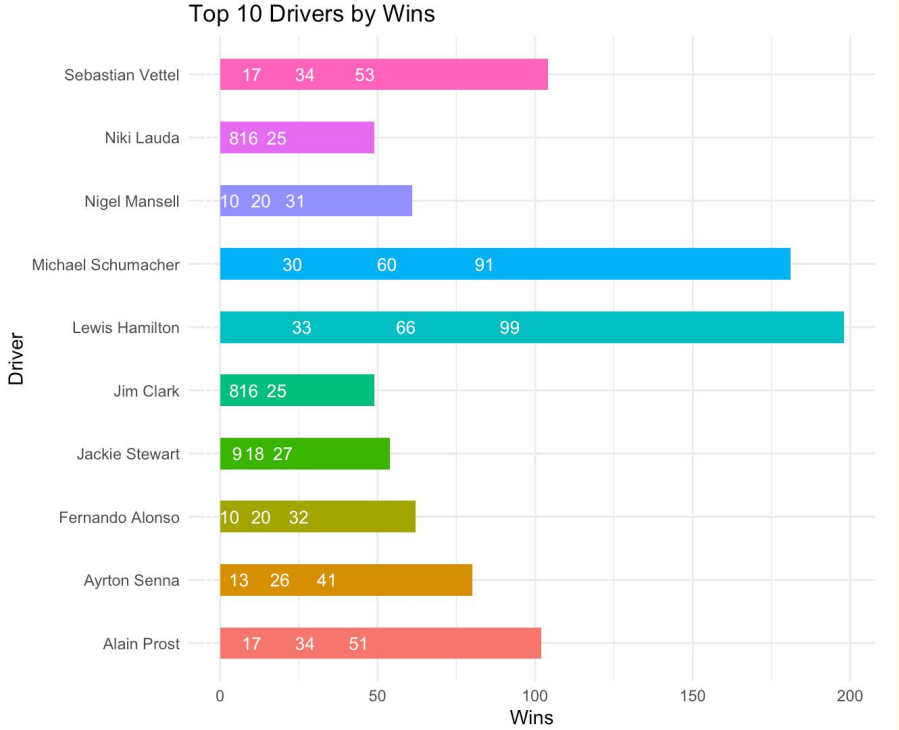
# Data Preparation Steps

After: After cleaning the data and merging the required tables:

| | raceId | driverId | constructorId | grid | position | driverStanding | constructorStanding |
|---|---|---|---|---|---|---|---|
| **18241** | 881 | 20 | 9 | 1 | 1 | 3 | 3 |
| **18242** | 881 | 17 | 9 | 5 | 2 | 6 | 3 |
| **18243** | 881 | 1 | 131 | 4 | 3 | 5 | 4 |
| **18244** | 881 | 3 | 131 | 6 | 4 | 20 | 4 |
| **18245** | 881 | 13 | 6 | 2 | 5 | 4 | 1 |

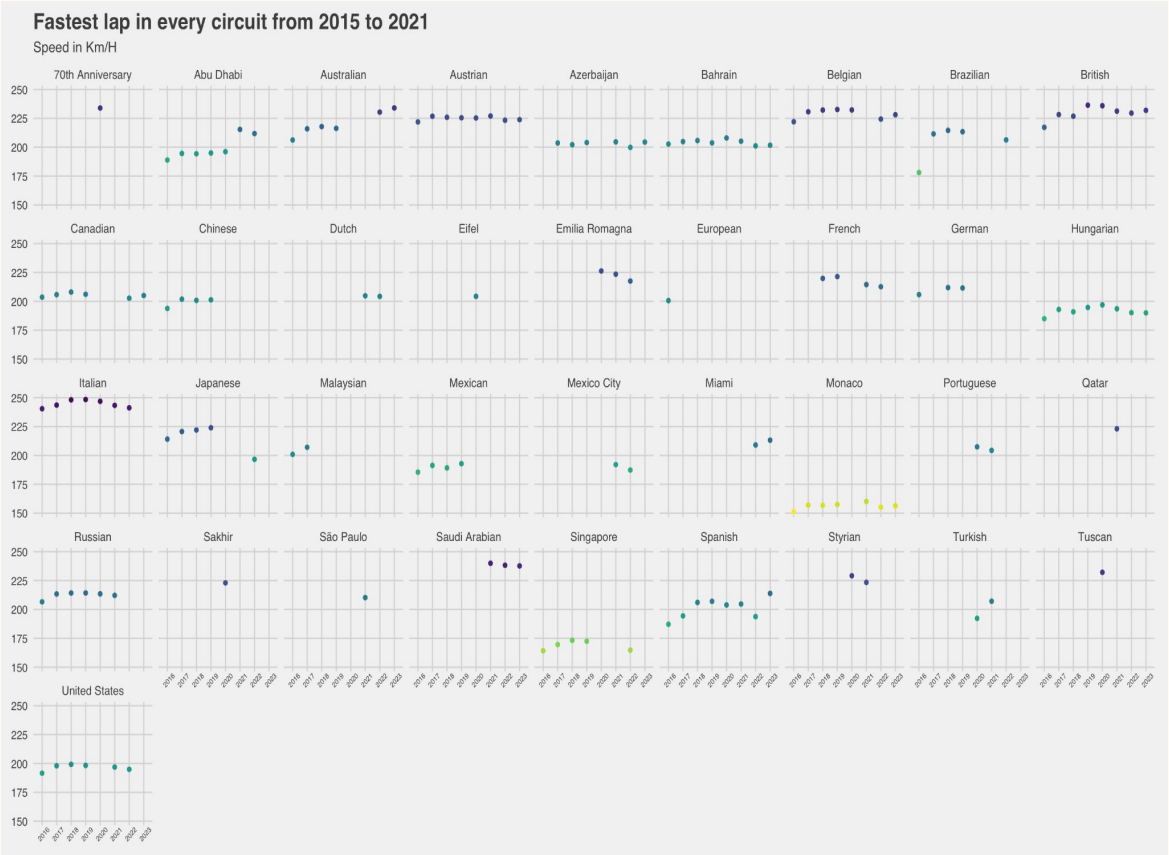
# Exploratory Analysis

For a concise exploration of the race and driver datasets, we conducted an analysis focusing on the most successful drivers in terms of victories. The ensuing graph provides a visual representation, distinctly showcasing Lewis Hamilton's dominant performances, spanning both his tenures at McLaren and Mercedes.



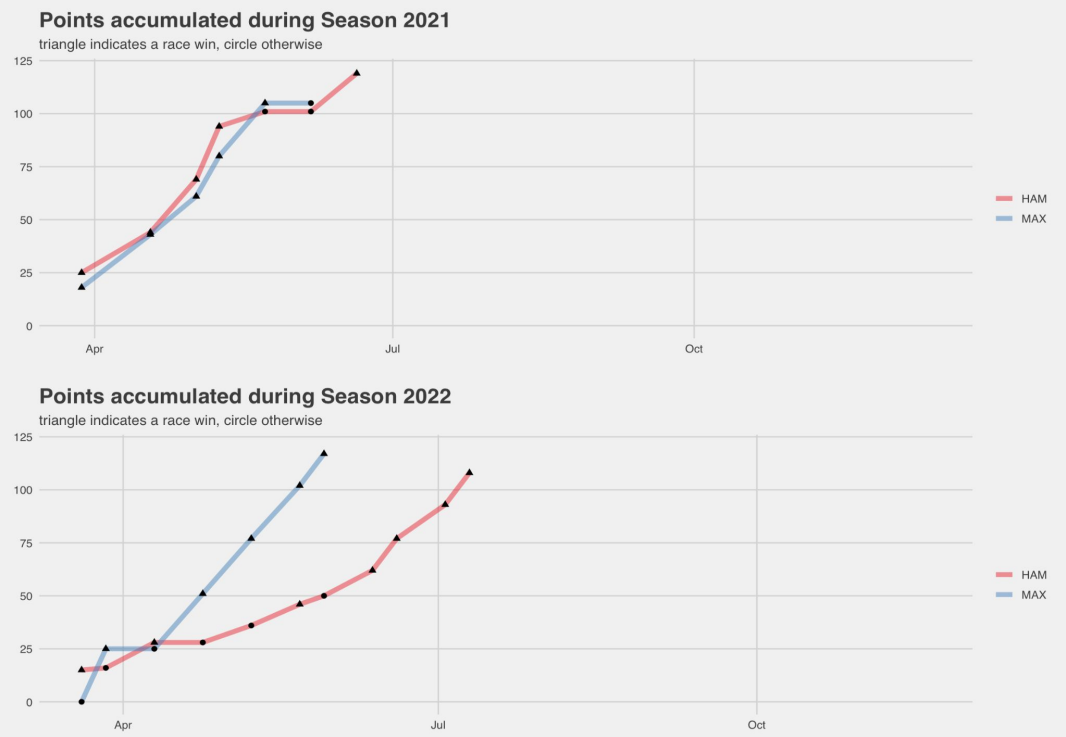
# Exploratory Analysis

This was to answer the problem statement “Which circuits have the highest and lowest average race speeds, and what factors contribute to these differences?”



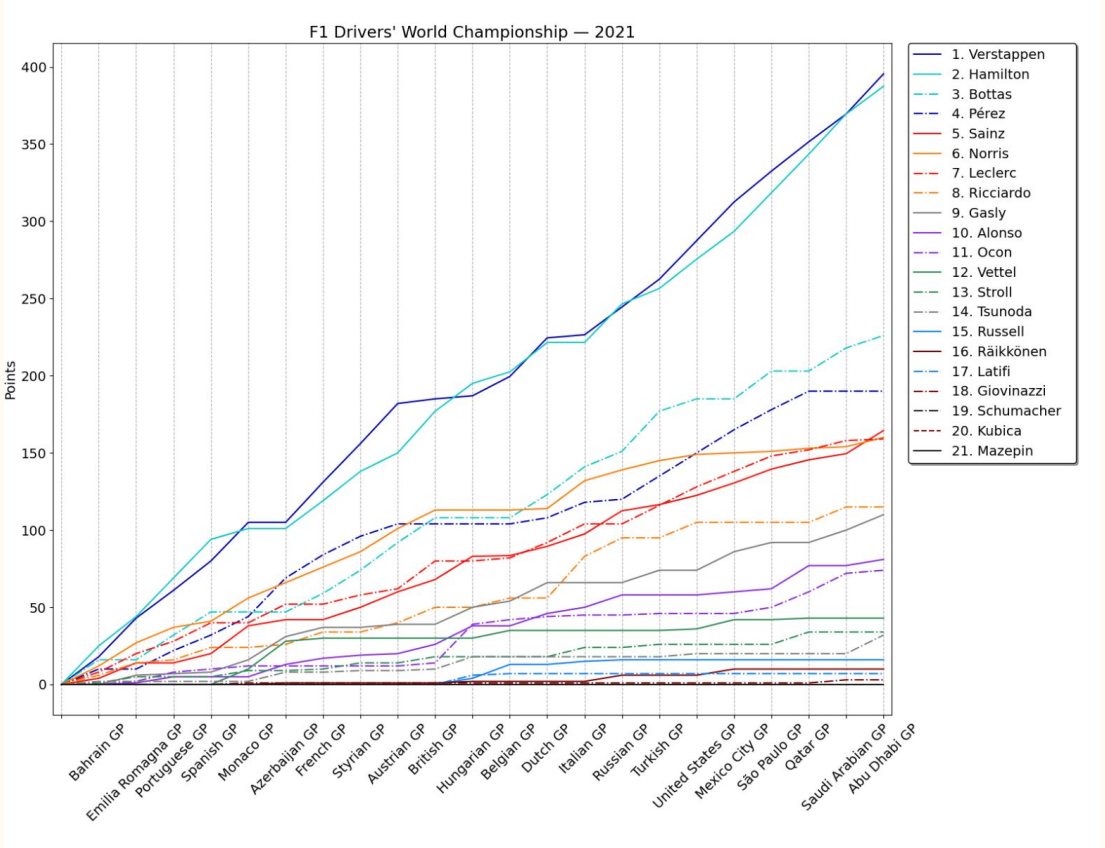
# Exploratory Analysis

We looked at the one of the greeted rivalry of the decade, between the Max and Hamilton and how they both were neck to neck in 2021 GP but RedBull clearly dominated 2022 GP and gapped Mercedes.



# Exploratory Analysis

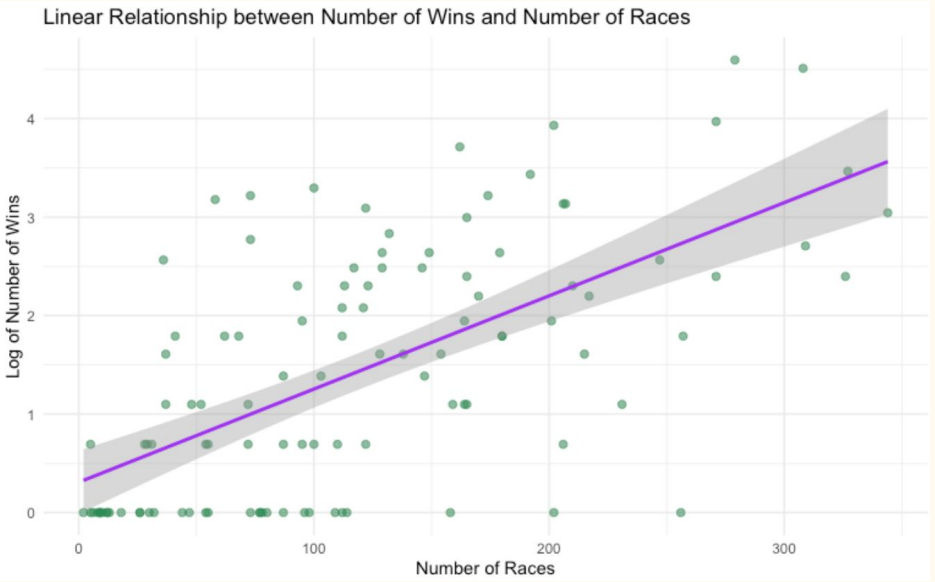
We also looked at the individual driver performance in 2021 GP where Max finally won to become the new World champion. The season was extremely close between RebBull and Mercedes.



# Model Estimation

We constructed two linear models to answer below questions.

1. To find the correlation between the features and predict the outcome of the race for a driver based off of the highly correlated features.
2. If there is a linear relationship between number of wins for a driver and the number of races they have been part of.

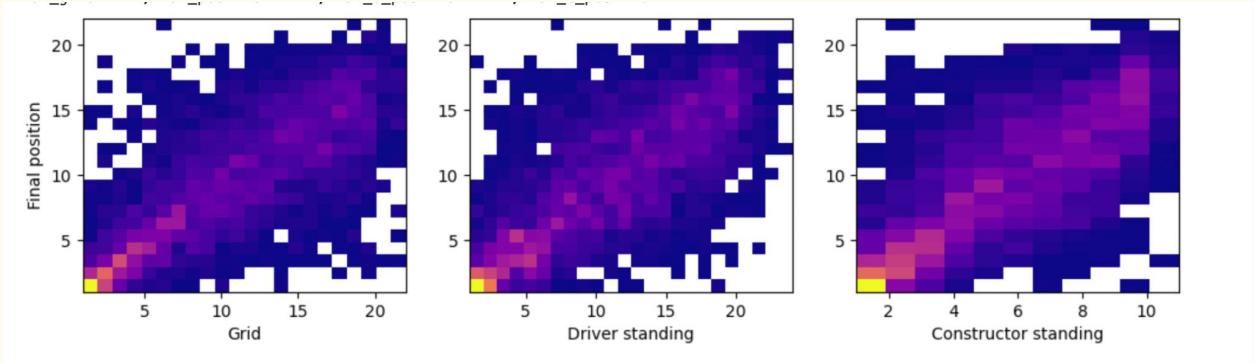


# Model Estimation

This shows the correlation between the features we chose for model creation. We identified three features to be mostly correlated with the race outcome and used this to predict the track position of a driver.

```
dataset.corr()["position"]
```

```
grid                  0.750109
driverStanding        0.742595
constructorStanding   0.750760
position              1.000000
Name: position, dtype: float64
```



# Model Estimation

We then tested the model to predict the Abu Dhabi 2023 GP with the current driver position, team constructor position and car grid position. The actual race finishes for the specified drivers were Max: 1, Leclerc: 3, Hamilton: 9. This discrepancy may be attributed to the utilization of data spanning the past decade to train our model. Given the longstanding dominance of Mercedes throughout this period, it has likely influenced the model to predict higher positions. Additionally, the recent ascendancy of Red Bull might explain the slight deviation, with one position less predicted. Below are the results for the last race of the season for 2023.

```
    VER = predict(1, 1, 1)
    HAM = predict(11, 3, 2)
    LEC = predict(2, 5, 3)
    print(f"Predictions: Max Verstappen - {VER}, Lewis Hamilton - {HAM}, Charles Leclerc - {LEC}")
]
```

```
Predictions: Max Verstappen - 2, Lewis Hamilton - 6, Charles Leclerc - 4
```

# Conclusion

The preceding analysis underscores the dominance of the Mercedes team in Formula 1 between 2013 and 2019, aligning seamlessly with the exceptional performance of Lewis Hamilton, who stands as the top driver of all time with 99 Grand Prix Championships. Notably, Red Bull has emerged as a formidable contender, challenging Mercedes in recent years. While the top two drivers, L. Hamilton and M. Schumacher, have achieved considerable success, their higher win counts correlate with their extensive participation in races.

Furthermore, a discernible correlation exists among the driver's grid position, their ranking, and the standings of the constructors. Notably, the grid position holds significant influence, particularly evident in the Monaco Grand Prix.

# References


- F1 Global fan survey results revealed at Monaco GP https://www.motorsport.com/f1/news/f1-global-fan-survey-results-monaco-gp-910621/910621/
- 2023 Formula One World Championship
- Smith, J. (2020). "Formula 1 Analytics: Uncovering Patterns and Trends in Motorsport Data." Journal of Sports Data Analysis, 10(2), 123-140.
- Johnson, M. A. (2018). "Predictive Modeling of F1 Race Outcomes Using Machine Learning." Proceedings of the International Conference on Data Science in Sports, 45-56.
- Formula 1 Official Website. "Formula 1 Race Results Archive." Retrieved from https://www.formula1.com/en/results.html
- Chen, S. (2017). "Data-Driven Insights into Pit Stop Strategies in Formula 1 Racing." International Journal of Data Science and Analytics, 5(3), 189-204.