## **1. Overview**

### **Objective:**

Aim is to analyze **Formula 1 racing data** to derive insights on driver performance, team strategies, and historical trends using **data science and machine learning** techniques.

### **Key Problem Statements:**

* Predicting **2025 season champions** using historical and current performance trends.
* Analyzing **pit stop efficiency, driver transitions, and circuit performance**.
* Evaluating **head-to-head driver rivalries** and **championship retention probability**.

## **2. Tools, Libraries, and Frameworks Used**

### **Development & Environment:**

* **Python 3.x** (Primary language)
* **Jupyter Notebook / VS Code** (Development environment)
* **Git & GitHub** (Version control and collaboration)
* **PowerBI**
* **Streamlit**

### **Libraries Used:**

* **Data Handling & Analysis:** pandas, numpy
* **Visualization:** matplotlib, seaborn, pyvis, networkx
* **Machine Learning:** scikit-learn, RandomForestRegressor
* **Statistical Analysis:** scipy.stats, statsmodels

## **3. Dataset Description**

The project leverages multiple datasets related to Formula 1 history, including:

| **Dataset** | **Description** |
| --- | --- |
| results.csv | Stores race results, driver positions, points, and fastest lap info. |
| races.csv | Contains race details such as year, round, and circuit information. |
| drivers.csv | Includes driver details (ID, name, DOB, nationality). |
| constructors.csv | Stores team details. |
| pit\_stops.csv | Data on pit stops (duration, timing). |
| driver\_standings.csv | Standings per season for each driver. |
| constructor\_standings.csv | Team rankings per season. |

## **4. Methodology & Approach**

### **Data Processing & Cleaning:**

* Handled missing values (\N replaced with -1 or driver/team averages).
* Merged relevant datasets using merge() for consolidated analysis.
* Encoded categorical variables using **One-Hot Encoding**.
* Mostly focused on data preprocessing and cleaning only when the need arises due to a shortage of time.

### **Statistical & Machine Learning Analysis:**

* **Markov Chains** for **team transition probabilities**.
* **Random Forest Regression** for **points prediction in 2025**.
* **ANOVA tests** to assess the impact of pit stops on race performance.
* **Stochastic Transition Matrix** for **driver team changes**.
* **Ensemble model - Random Forest regressor** used to predict points which were further used in predicting champions etc.,

### **Predicting 2025 Champions:**

* Built **ML model** to predict **driver’s points per match for a whole season** to form the driver and constructor rankings.
* Used **historical performance** trends to forecast results.

## **5. Key Insights & Findings**

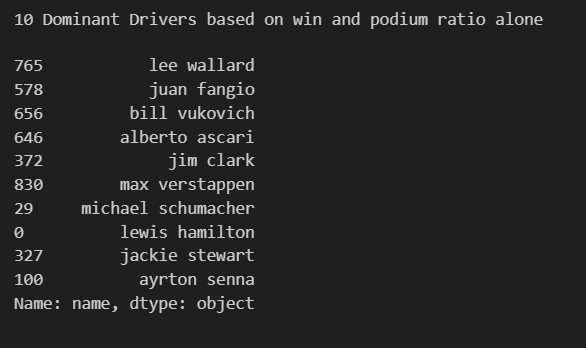
* **Champion Age Trends:** Most F1 champions win between **25 and 35**.
* **Head-to-Head Rivalries:** Drivers like **Jenson Button** and **Felipe Massa** show competitive patterns.
* **Pit Stop Efficiency:** Too many pit stops **negatively impact** the final position.
* **Team Movements:** **Red Bull & Mercedes** have the highest driver transition probabilities.

## **6. Visualizations**

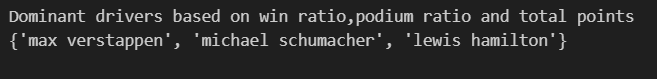
* **Bar plots** and **histograms** for **grid vs podium analysis** and **team performance comparison.**
* **Bar plots** for **average lap time constructor-wise.**
* **Network graphs** for **driver movements between teams**.
* **Heatmaps** for **pit stop efficiency and grid position impacts**.

**7. Question analysis :**

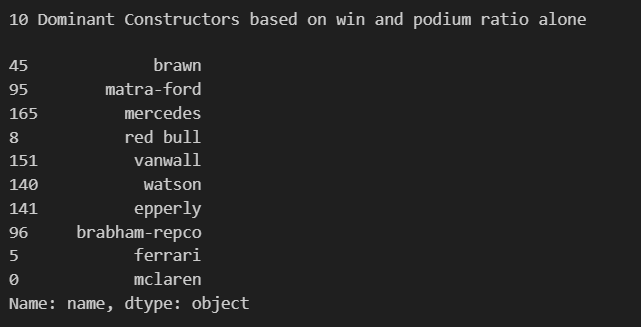
**1.****Driver & Constructor Performance**

****

As per the question we used the win ratio and podium finishes ratio to determine the dominant(top 10) drivers and the above are the results. But the above is sensitive to career length and thus we also include total points(career length high => total points high) to determine the dominant racers.

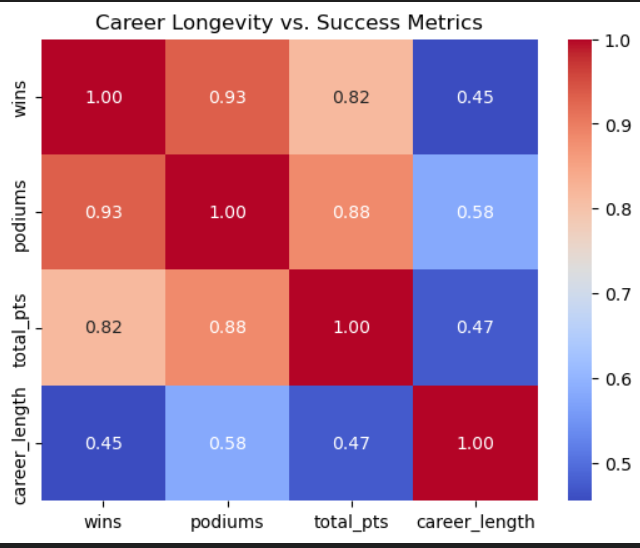
****

Similar strategy is applied to find dominant constructors

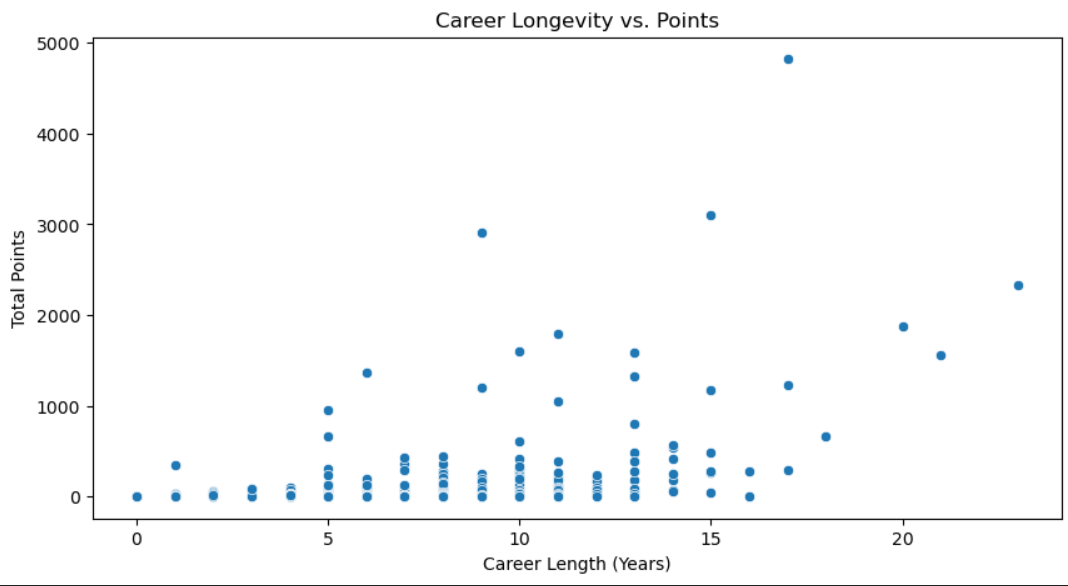
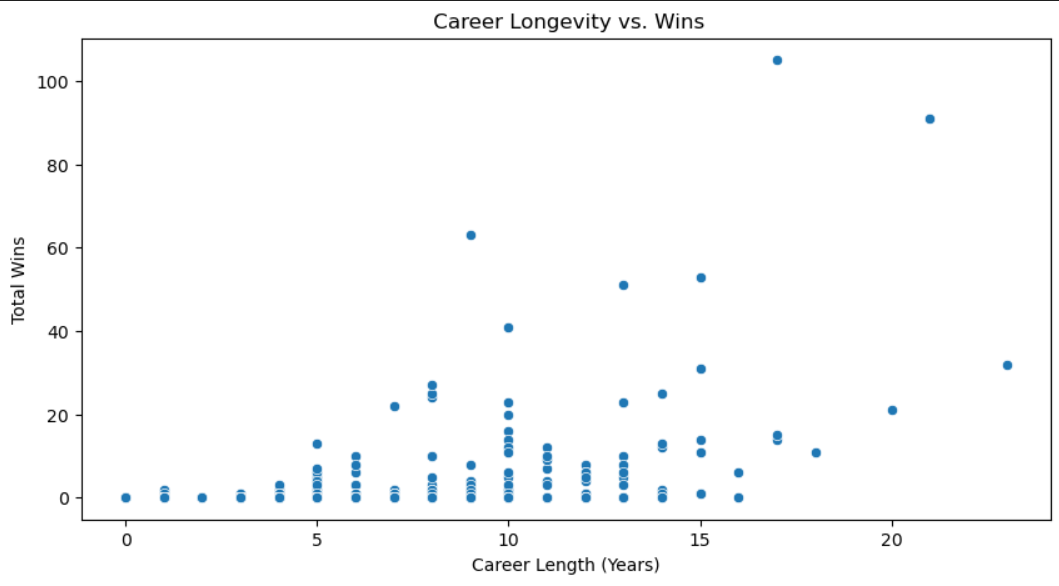




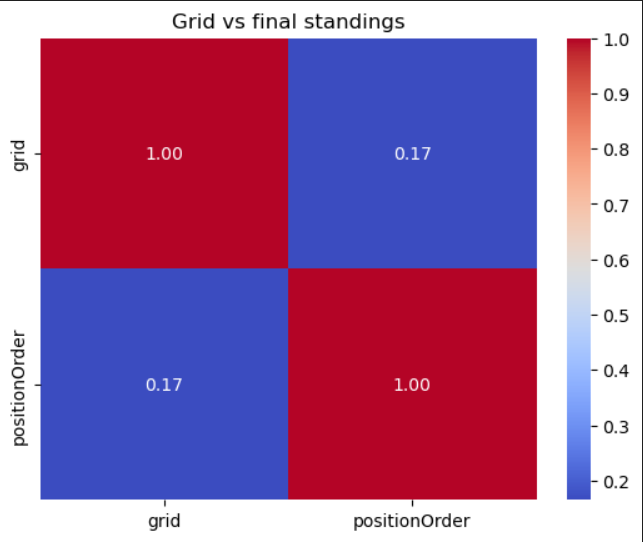
**Career longevity vs win measures [wins,podiums,totalpoints]**

****

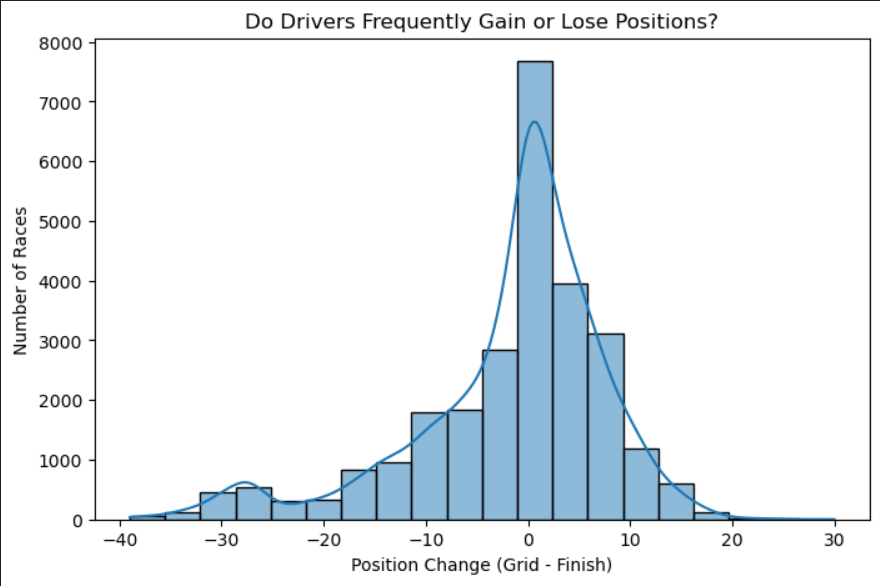
From the above heatmap (correlation matrix) we can see that there some positive correlation between career length and all win measures, which shows that a high career length corresponds to high win measures but need not be the case always as indicated by the below scatter plots.



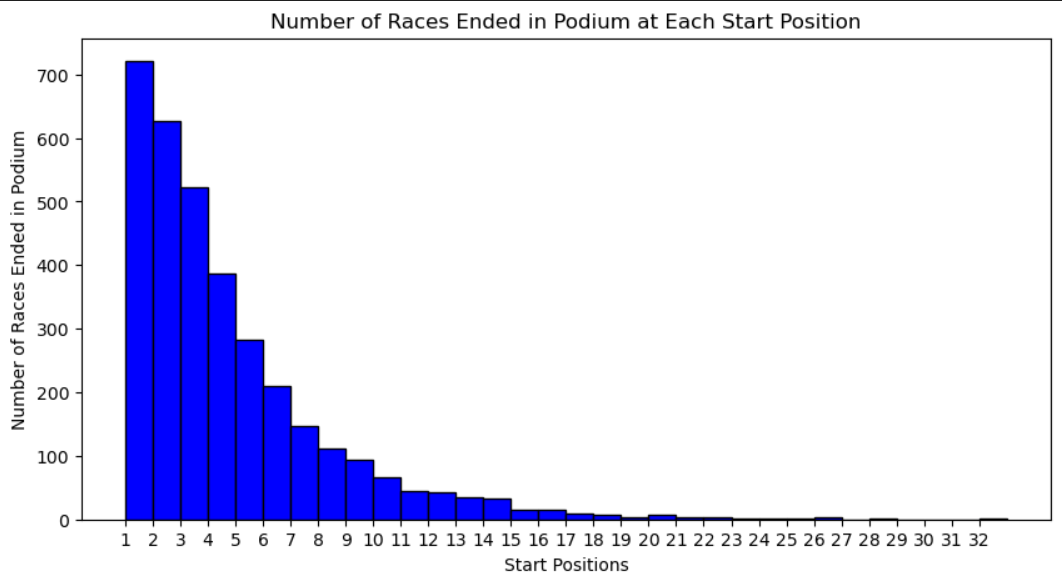
**2. Qualifying vs. Race Performance**

****

The above indicates that there does not exist much of a linear relationship b/w starting grid and finishing position.But we cant also ignore the fact that there is some(minimal) positive correlation which suggests that lower the starting position lower maybe the finishing position.



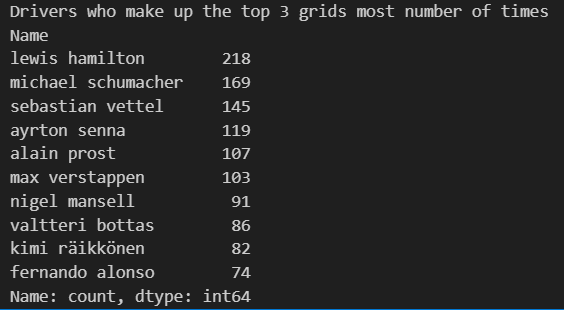
Indicates that mostly driver either gain or stay in same position.



Indiictaes that drivers having smaller starting positions or grids tend to finish at the top mostly.

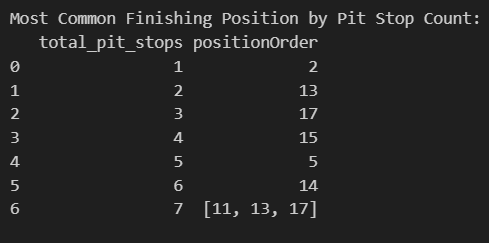


This indicates that some drivers tend to do better performance at races to have maximum position change (end-start) compared to others.But we cannot say if they are best as some drivers who start and finish near the same position gets a low position change.But the above are the people who does best when it comes to making up positions.

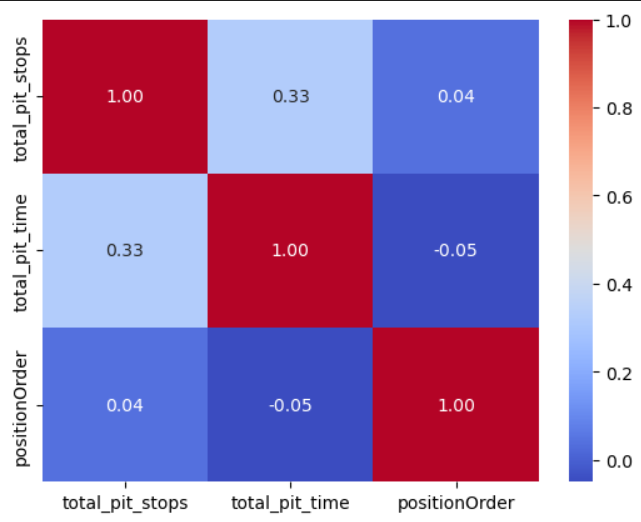


The above are the list of drivers who tend to perform well at qualifications to get top 3 grid locations.

**3.** **Pit Stop Strategies**

****

The above table shows the mode position attained by the drivers for each number of total pitstops.



No direct linear relationship between total pit stops, total pit stop time and position can be inferred from the correlation matrix.So we go for ANOVA test.

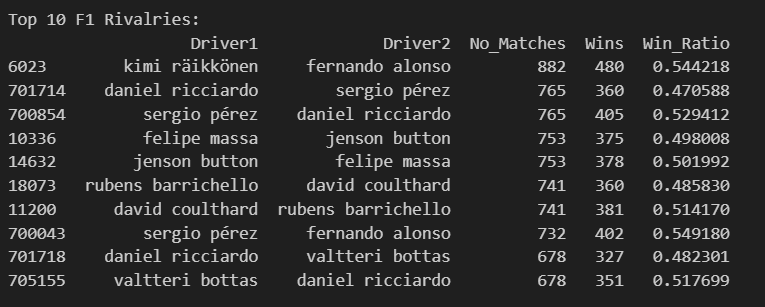
One Way ANOVA :

H₀ (Null Hypothesis): No difference in finishing positions across different pit stop time groups.

H₁ (Alternative Hypothesis): At least one group has significantly different finishing positions.



**4.** **Head-to-Head Driver Analysis:**

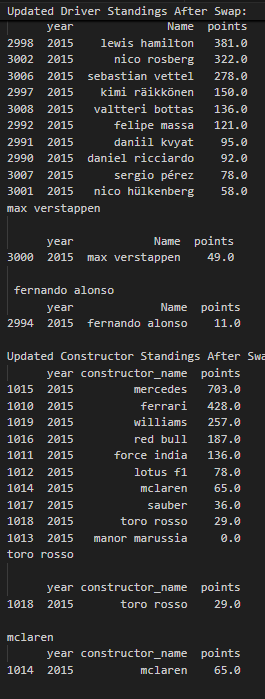
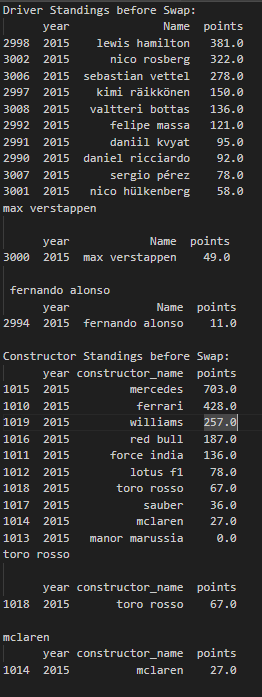
****

We have top rivalries when even after a considerable number of matches the win ratio or probability between the drivers must be almost equal indicating an equal potential or greater competition. Thus from the above table, we can infer the top rivalries filtered by win ratio in 0.50±0.05 interval.

**5.** **Hypothetical Driver Swaps:**

**Understanding the result of swapping two driver’s teams in a season**

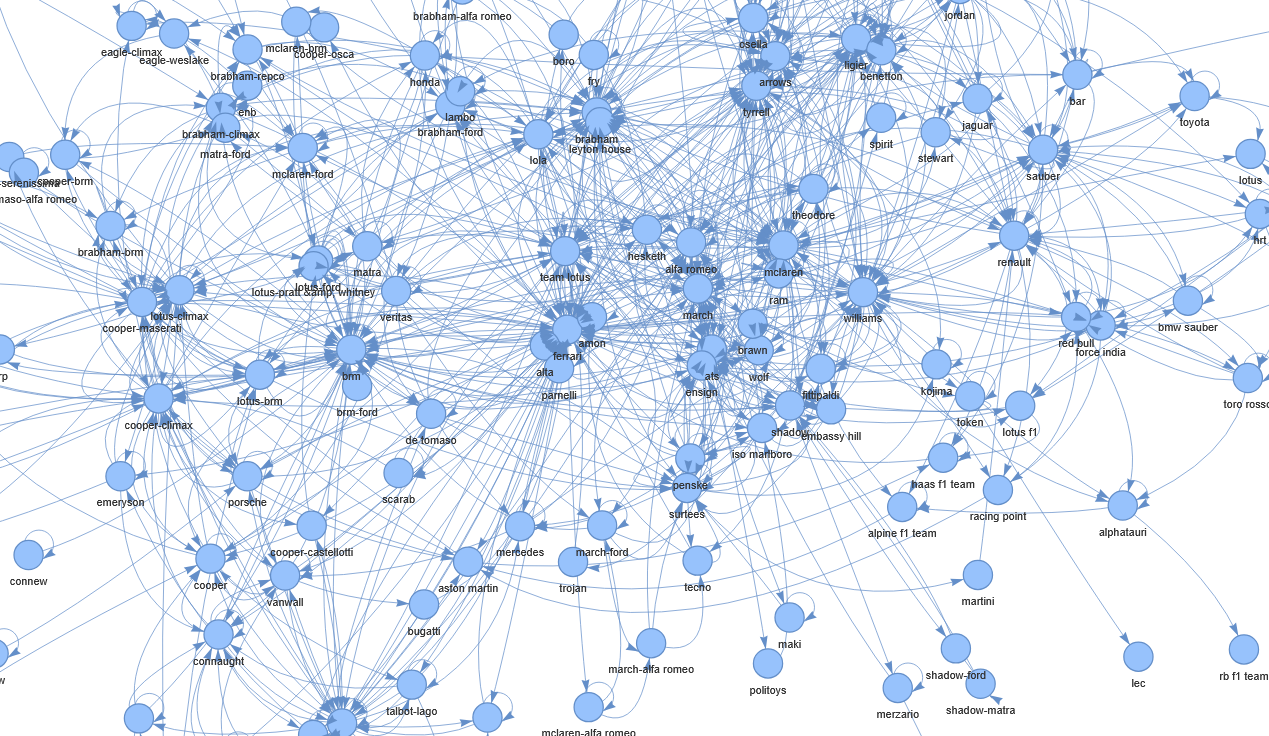
**Here we take driver1 = 'max Verstappen' driver2 = 'fernando alonso' in the year 2015**

****

It can be seen that swapping teams of players can lead to complete change in constructor rankings if a driver being swapped is of high dominance.From the updated stats we can clearly see that the driver standings does not change as we do not affect individuals performance by swapping teams and also assuming all other factors like pit stop strategies are almost same for the swapped teams. Also we can see that by swapping teams of a higher ranked driver like max verstappen with team of low ranked driver(comparitively) like fernandez leads to a drastic change in constructor standings.The rankings between the teams is inverted.

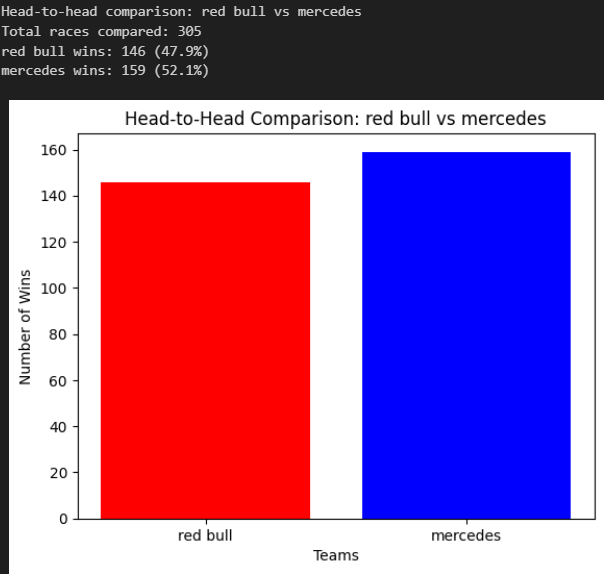
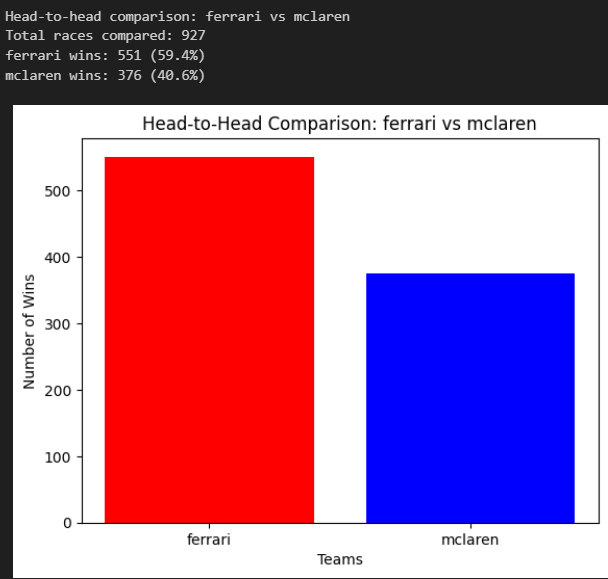
**6.****Driver Movements & Team Networks:**

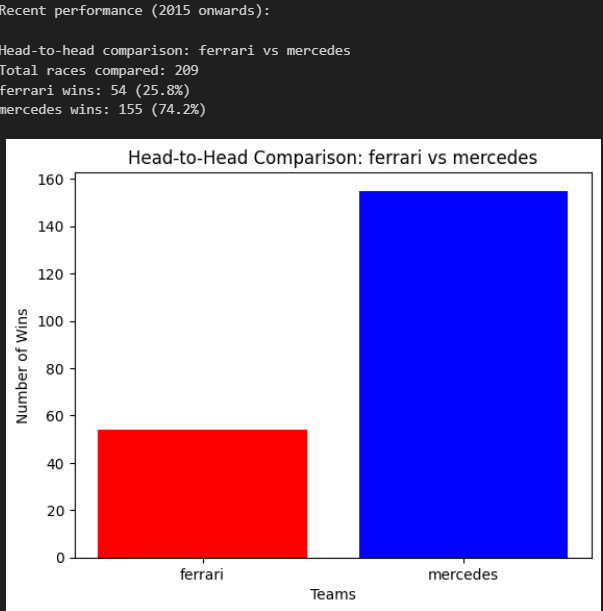
****

Framed network dataset  


Final network graph visualized using pyviz.This graph is directed with nodes indicating constructors and edges indicating a driver swap from one team to another.

**7.****Team Performance Comparison:**

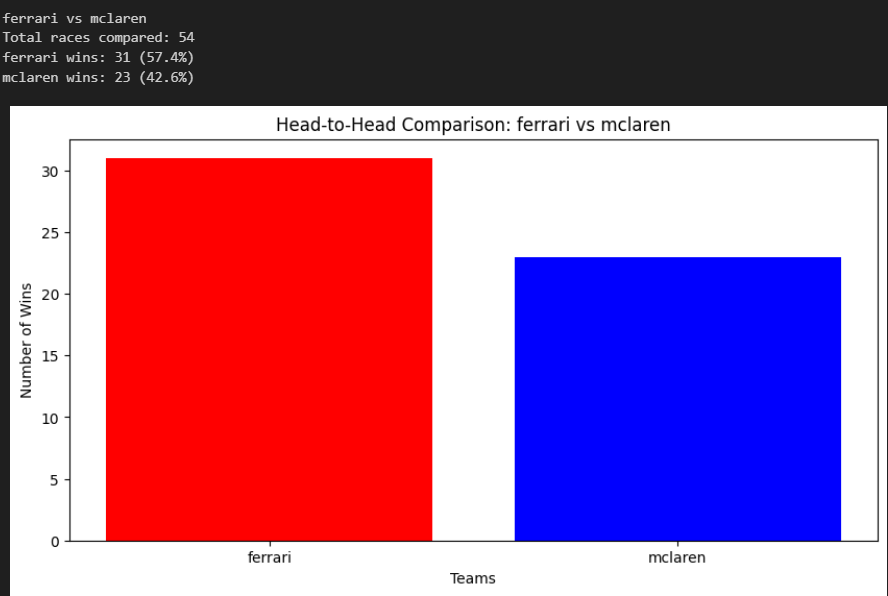
****

****

**With circuits :**

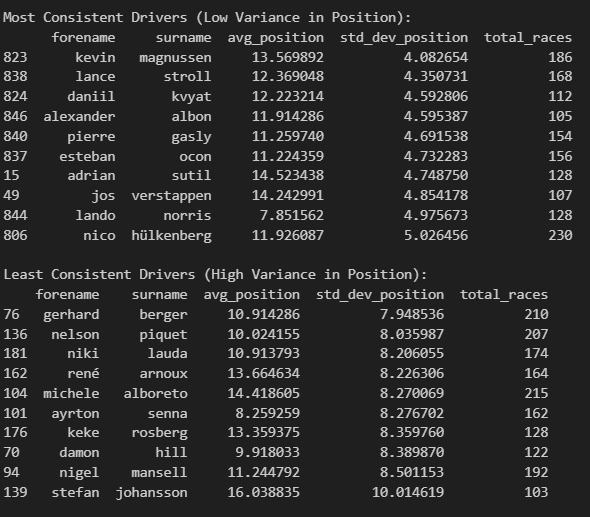
****

Here, the circuit factor **does not affect** the team performance , since the total circuit wins with circuit factor is same as without circuit factor. The circuit-wise factor is only a categorization and does not alter the calculations.



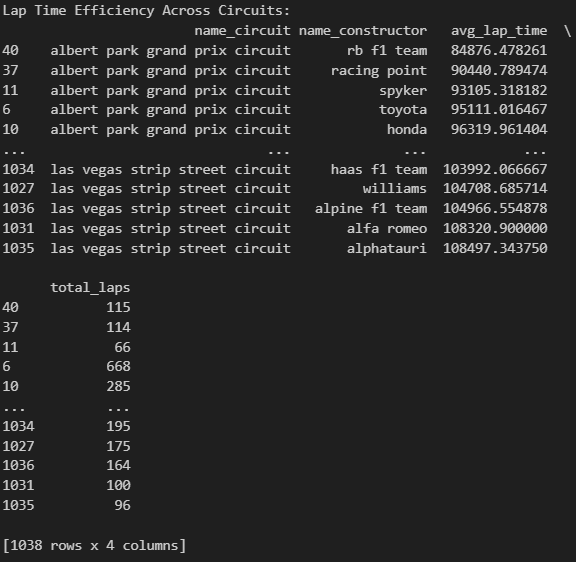
Comparioson done for "Ferrari" and "McLaren" in autodromo nazionale di monza circuit.

**8. Driver Consistency in Race Performance:**

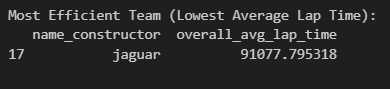
****

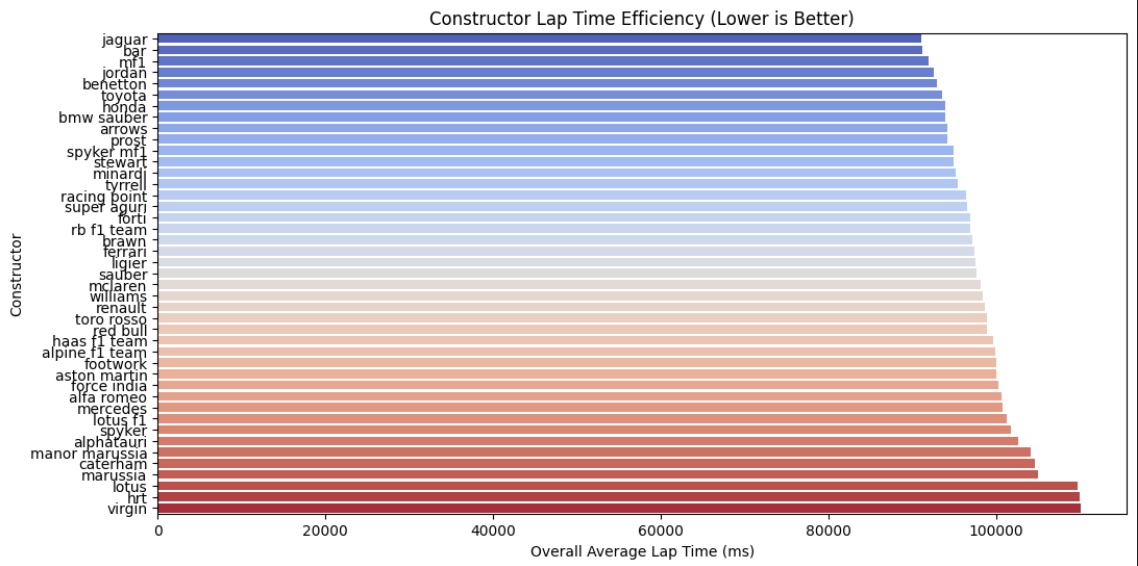
Here , **Kevin** is the **most consistent** since he has the lowest standard deviation . **Gerhard** is the **least consistent** since he has the highest standard deviation.

**9.****Lap Time Efficiency:**

****

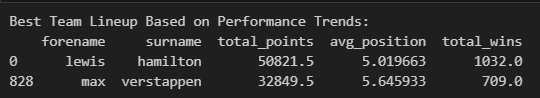
**Rb F1** Team has the **lowest average lap time** of 84876.47 ms in the **Albert Park Grand Prix Circuit.**

****

****

Overall average lap time for each constructor. This indicates that **Jaguar** has the lowest average lap time from which may indicate that they have **better pit stop strategies** and efficient **driver guiding** during the races.

**10.Best Team Lineup**

****

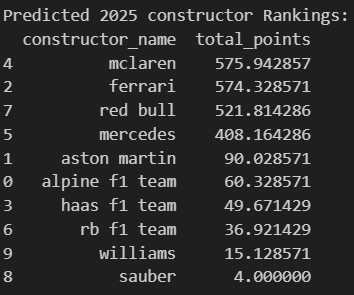
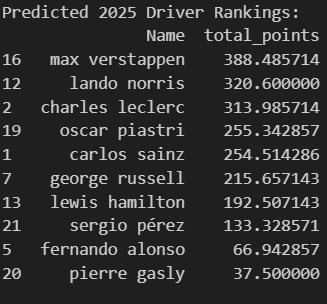
The best line up for a constructor based on various factors like **total points,average finishing position** and **total wins.**

From the above we can infer that the best line up possibly is a team made of

**Max Verstappen** and **Lewis Hamilton**

**11.****Predictions for 2025 Season:**

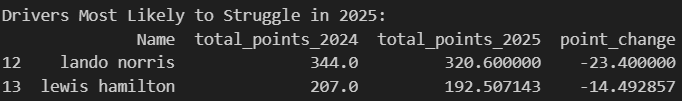
Assuming 2024 drivers alone and that they do not change their teams for the 2025 season. Also each season has many races and we will be aggregating the points earned in each match to get the final result and assuming their performance is similar to 2024.

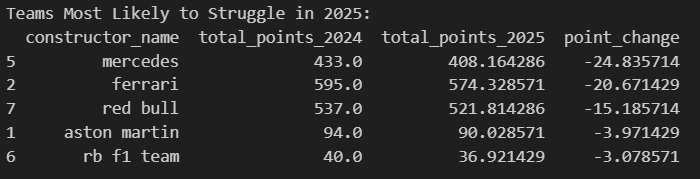
****

From the data, we can expect **Max Verstappen** to win the **2025 season**

As of constructors, **McLaren** is expected to win

**12.****Struggling Teams Analysis:**

****

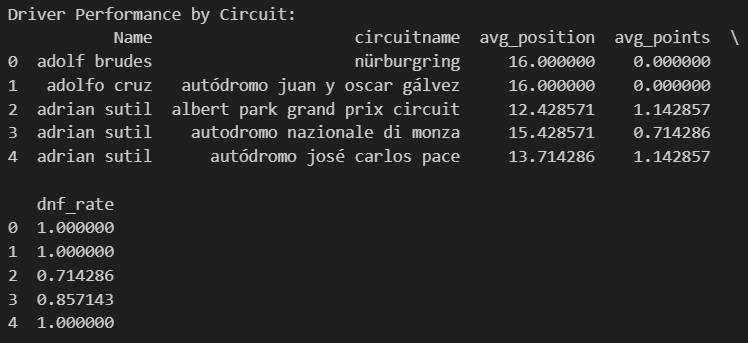
****

We have considered the change in performance from 2024 to 2025 to retrieve struggling teams and drivers in 2025 season. According to the data **Lando Norris** is expected to struggle more in 2025 season (performance compared with 2024 season).

Also when it comes to a team **Mercedes** is expected to struggle more in 2025 season (performance compared with 2024 season).

**13.****Driver-Specific Track Struggles:**

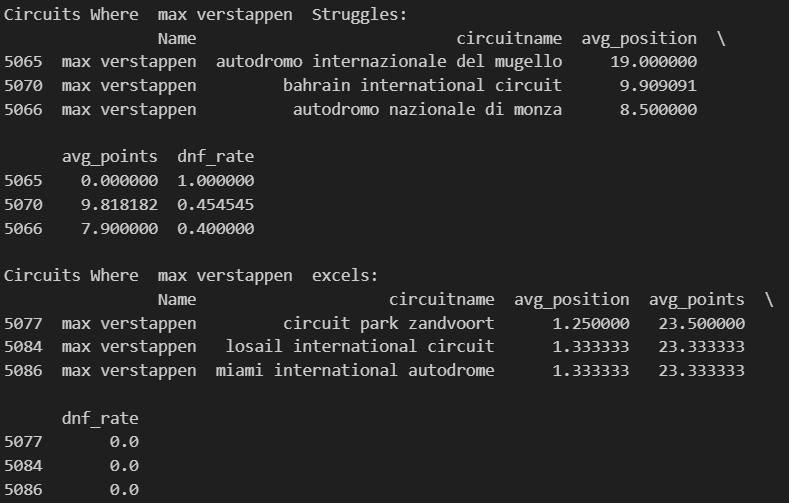
**General :**

****

From these we can say that **Adolf Brudes** struggles the most in the **nürburgring** circuit, concluded based on large avg position ,low average position and high DNF rate.

**Driver Specific :**

**driver = 'max verstappen'**

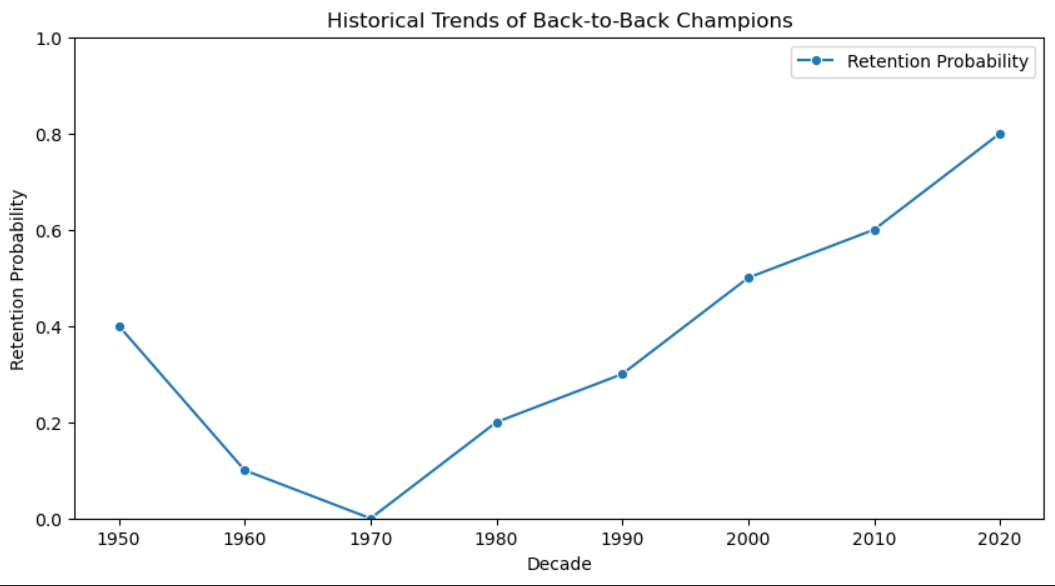
****

From these we can conclude that **max verstappen****struggles** most in autodromo **internazionale del mugello** circuit.

Also, he is known to **excel** most in **circuit park Zandvoort** .

**14.****Championship Retention Probability:**

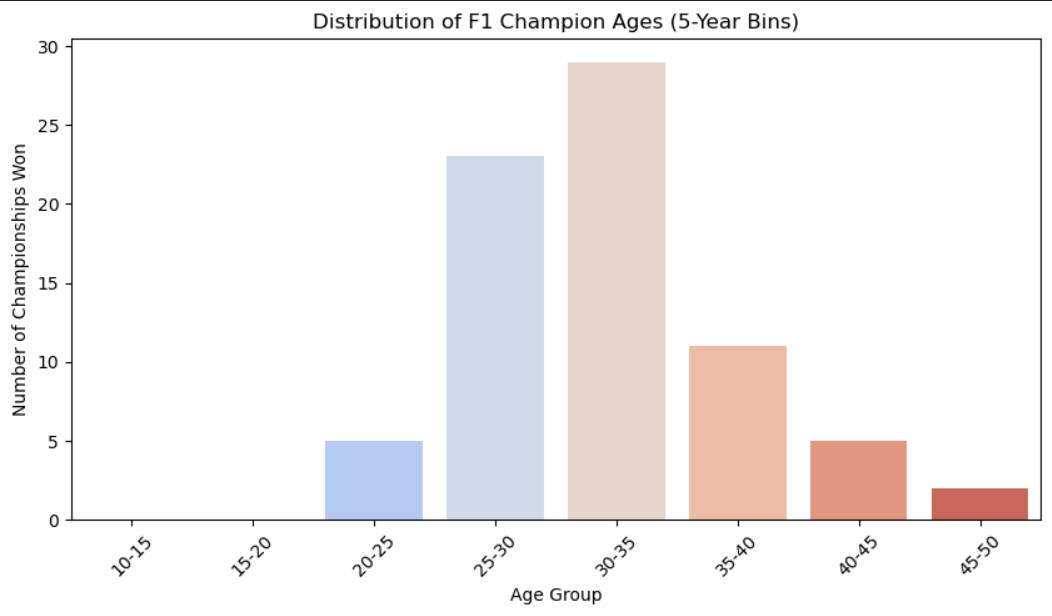
Overall Championship Retention Probability**: 33.78%**

****

From these, we can say that with probability **0.34** 2024 champion will retain their championship in 2025.

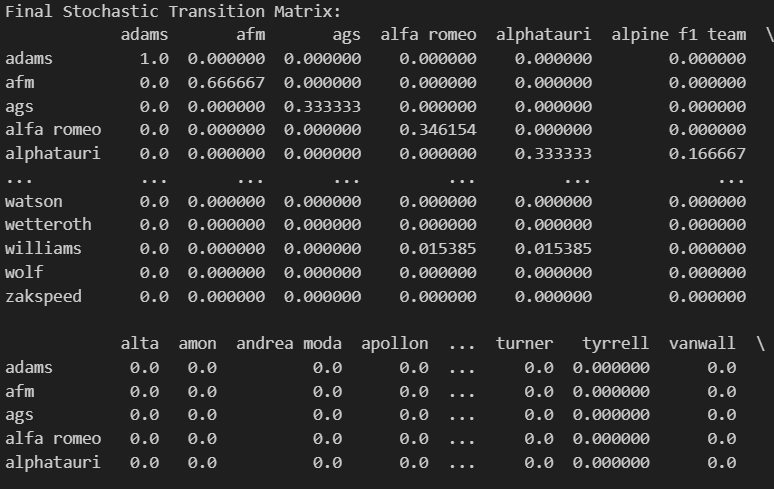
Also from the historical decade trends of championship retention, we can see that the probability of retaining championships is **increasing** since the 1970s.

**15.****Champion Age Trends:**

****

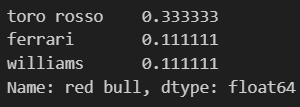
From this, we can see that drivers most frequently win championships at around **30-35** years of age followed by **25-30**. On the whole, we can say that most championships are won by drivers in the age group **25-35.**

**16.****Bonus Challenge (Optional):**

****

Above is the sample of the stochastic transition matrix used.

**driver = 'max verstappen'**



From the probability estimate it is seen that there is a probability of **0.33** for the driver to move from **red bull** to **toro rosso.**

## **8. How to Run the Project**

1️) **Clone the repository or download the zip file** :

git clone https://github.com/LalithKishore-S/DPL\_2025.git

2) **Jupyter Notebook** already has ready to view **outputs,** so we don't need to run it again.

3) For viewing the **results alone** you may **run** the **Streamlit app.py** file for compiled results and inferences.

**Command to run in terminal :** streamlit run app.py

**9. Git folder structure**

* Additional data
  + Network\_teams.csv
* DPL\_Dataset
  + Given dataset
* Final\_dataset
  + Partially processed and cleaned datasets
* Removing\_statuscsv\_bymapping.ipynb - used to map status id to status in results.csv and sprint\_results.csv
* Preprocess\_part1.ipynb - contains basic preprocessing for the first half of datasets.
* Data\_preprocessing\_part2.ipynb - contains basic preprocessing for the second half of datasets.
* Data Premier League-Problem Statement.docx - problem statement
* Analysis\_questions1\_5.ipynb - Qns 1 to 5 analysis
* Qn6\_network.ipynb - Qn 6 analysis
* Analysis\_questions\_7-10.ipynb - Qns 7 to 10 analysis
* Analysis\_questions11-12.ipynb - Qns 11 and 12 analysis
* Analysis\_questions13-15.ipynb - Qns 13 to 15 analysis
* Bonus\_challenge\_qn16.ipynb - bonus qn analysis
* Data\_streamlit - contains data to be presented in streamlit.
* DPL\_2025\_Documentation - Final compiled documentation.
* Team verification file - Team details
* Dashboard - Final PowerBI dashboard
* Basic.html and lib folders - for pyvis visualization(automatically created when run the cell)

## **9. Conclusion**

This project successfully provides **data-driven insights** into Formula 1 racing by leveraging **statistical analysis, machine learning, and network science**. The findings can help **teams, analysts, and fans** understand key factors influencing **race outcomes and driver performances**.

**GitHub Repo:** [DPL\_2025](https://github.com/LalithKishore-S/DPL_2025)