SASTRA
ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act,1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

DOCUMENTATION

Exploratory Data Analysis of Formula 1 Dataset: Insights into Race Performance, Speed Trends, and Driver Efficiency

Documented by:

Nitesh Venkatesan

I M.Sc. Data Science,

SASTRA Chennai

Campus.

ABSTRACT:

Formula 1 is not just a motorsport—it's a data-driven battlefield where every millisecond and strategic choice can determine victory. This project, titled "Exploratory Data Analysis of Formula 1 Racing Dataset: Insights into Driver Performance, Circuit Dynamics, and Race Strategies," focuses on uncovering hidden trends and performance patterns within official Formula 1 race data.

Over a race weekend, events include practice sessions, qualifying rounds to determine starting positions, and the main race on Sunday. Pit stops during the race for tire changes and adjustments are critical to race strategy. The winner is the first driver to complete the race distance, earning points toward the season championship.

Various systems for awarding championship points have been used since 1950. The current system, in place since 2010, awards the top ten cars points in the Drivers' and Constructors' Championships, with the winner receiving 25 points. Points won at each race are added up, and the driver and constructor with the most points at the end of the season are crowned World Champions.

A driver must be classified in order to receive points. They must complete at least 90% of the race distance in order to receive points. Therefore, it is possible for a driver to receive points even if they retired before the end of the race.

Using datasets such as *results.csv* and *drivers.csv*, this study integrates key attributes including driver details, race outcomes, fastest lap speeds, and finishing positions to analyse performance consistency, driver dominance, and circuit-specific characteristics.

The analysis was conducted using Python's Pandas, Matplotlib, and Seaborn libraries within a Jupyter / Colab environment. Data preprocessing involved handling missing and inconsistent values (such as \N entries), merging driver and results datasets via unique identifiers, and standardizing key columns like *fastestLapTime* and *position*. Through systematic exploration, multiple visualizations were generated to interpret race dynamics—such as the relationship between qualifying positions and race finishes, average speed variations across circuits, and driver performance comparisons based on accumulated points.

Key findings reveal that certain drivers consistently outperform others on specific circuits, highlighting individual expertise and team strategy advantages. Visual correlation analysis also suggests a strong link between qualifying performance and final race placement, emphasizing the importance of initial grid positioning. Furthermore, comparative plots of average speed versus average points illustrate that high speed does not always guarantee higher points, underlining the tactical aspects of pit stops and tire management.

This exploratory study demonstrates how Formula 1 data can be leveraged to derive actionable insights into performance trends, team strategies, and driver efficiency. The results not only showcase the analytical power of EDA in motorsports but also lay the foundation for advanced modelling in race outcome and performance optimization in future research.

TABLE OF CONTENTS

CHAPTER 1 - INTRODUCTION

Formula 1 (F1) is more than just a high-speed motorsport; it's a data-intensive battlefield where milliseconds decide glory or defeat. Every lap, pit stop, and tire change generates enormous amounts of data, allowing teams to analyse and optimize performance down to the finest detail. This continuous flow of information makes Formula 1 an ideal real-world application for data analytics and exploratory data analysis (EDA).

The aim of this project, titled "Exploratory Data Analysis of Formula 1 Racing Dataset: Insights into Driver Performance, Circuit Dynamics, and Race Strategies," is to uncover meaningful insights hidden within historical race data. The analysis focuses on variables such as driver performance, finishing positions, fastest lap times, and average speeds to understand how different factors influence race outcomes. By merging datasets like results.csv and drivers.csv, the study integrates both statistical and contextual data, creating a unified view of race results and driver efficiency.

This project utilizes Python's Pandas, Matplotlib, and Seaborn libraries to perform data cleaning, transformation, and visualization. Missing or inconsistent values (e.g., \N) were addressed through preprocessing to ensure accurate interpretation of trends. Once prepared, the dataset was analyzed through various plots and correlation studies to explore questions such as:

- Does qualifying position affect final race standing?
- Which circuits favor specific drivers or teams?
- Is higher average speed always linked to higher race points?
-

Through visual and statistical exploration, the project aims to illustrate the relationship between performance metrics and race outcomes. It also highlights how data analytics can transform raw Formula 1 statistics into actionable insights, aiding in strategic understanding of race dynamics and driver consistency.

Ultimately, this analysis bridges sports and data science, demonstrating that even in one of the most technologically advanced sports on the planet, the real race often begins — and ends — in the data.

The significance of this analysis lies in how it bridges data science and sports analytics. Beyond identifying the fastest drivers or top teams, the goal is to understand *why* certain performances occur and *how* different race conditions influence results. For example, a driver may not always have the highest speed but could still finish consistently in the top positions due to strategic pit stops or team coordination.

Similarly, circuits differ in structure and difficulty, influencing speed and lap time variations. By visualizing and interpreting these dynamics, the project aims to deliver insights that mirror the decision-making processes of professional Formula 1 teams

1.1 Problem Statement

Formula 1 generates massive volumes of race data every season, including driver performance, lap times, pit stops, and finishing positions. However, despite this data abundance, meaningful insights often remain hidden due to data complexity, missing values, and the lack of unified analysis. Understanding how various factors—such as **Qualifying position, Average speed, and Circuit**—influence race outcomes is essential for performance evaluation and strategy development.

The key challenge lies in identifying the relationships between these performance metrics and translating them into actionable insights. For instance, does a faster average speed always lead to higher points? Do specific circuits favor certain drivers? Such questions require systematic analysis beyond surface-level statistics.

This project addresses these challenges through **Exploratory Data Analysis (EDA)** of Formula 1 datasets, combining race results and driver details to reveal patterns, trends, and correlations. By cleaning, merging, and visualizing the data, the study aims to uncover how driver efficiency, circuit characteristics, and race strategies collectively shape race outcomes in the world of Formula 1.

Objective of my Project

The main objective of this Exploratory Data Analysis (EDA) project is to analyze and interpret Formula 1 race data to gain meaningful insights into Driver performance, race outcomes, and circuit dynamics.

1. **To merge and preprocess Formula 1 datasets (results.csv and drivers.csv)** by handling missing values, removing inconsistencies, and integrating driver details with race outcomes for comprehensive analysis.

2. **To analyze and visualize driver performance metrics,** including finishing positions, average points, and fastest lap speeds, to evaluate consistency and dominance across different races.

3. **To identify correlations between key performance indicators**- such as qualifying position, average speed, and final race result, to determine which factors most influence race outcomes.

4. **To perform circuit-wise analysis** to study how track characteristics and locations affect driver speed and points distribution.

5. **To detect and manage missing or invalid data entries** (such as \N values) to ensure analytical accuracy and maintain dataset reliability.

6. **To generate visual insights** using bar charts, scatter plots, and heatmaps for understanding relationships between speed, position, and driver efficiency.

7. **To derive data-driven conclusions** that highlight performance trends, circuit preferences, and strategic factors contributing to Formula 1 race success.

1.2 Significance

The significance of this project lies in its ability to bridge sports analytics and data science. Through EDA, this study provides a comprehensive view of how various performance indicators—such as fastest lap speed, qualifying position, and average points—interact and influence race outcomes. Understanding these relationships helps to identify consistent drivers, high-performing teams, and circuits that yield the most competitive races. Moreover, analyzing trends in speed and points distribution allows us to assess whether success in Formula 1 is purely mechanical or strategically influenced by factors like pit timing, tire management, and track familiarity.

From a technical standpoint, this EDA project showcases the practical importance of data preprocessing. The Formula 1 dataset contains missing values, inconsistent entries (like \N placeholders), and varying data types across multiple CSV files. Handling these issues through cleaning, imputation, and merging ensures analytical accuracy and reliability. Such preprocessing steps highlight how even in professional datasets, data quality plays a crucial role in determining the validity of insights. This reflects real-world data science challenges and prepares students and analysts to work with messy, real-time data streams in future analytical or predictive modeling tasks.

In addition, EDA allows for effective visualization and communication of complex information. By using libraries such as Matplotlib and Seaborn, this project presents data in a visually interpretable format through scatter plots, bar charts, and heatmaps. These visualizations not only make the findings easier to understand but also reveal relationships that may not be visible through numerical summaries alone. For instance, plotting average speed against race position can uncover whether faster drivers consistently secure better results, or if race strategies have a stronger impact on final standings. Such visual clarity is vital in fields where decisions depend on rapid and accurate data interpretation.

Finally, this EDA project carries educational and analytical significance. It strengthens practical data-handling skills, enhances problem-solving capabilities, and encourages analytical thinking through real-world applications. It demonstrates how a seemingly complex and high-speed sport like Formula 1 can be simplified and studied scientifically through data-driven approaches. The insights derived not only deepen understanding of race performance but also inspire further studies in sports analytics.

CHAPTER 2 - DATA COLLECTION

2.1 Data Sources

The dataset used for this analysis was obtained from Kaggle, a publicly available data-sharing platform that hosts open datasets for research and analytical purposes. The dataset contains Formula 1 Datasets covering the period from 2008 to 2024.

The data is based on observations aligned with the Formula 1 organisation and Fédération Internationale de l'Automobile (FIA), which document detailed information about each Grand Prix event, including driver performance, race outcomes, circuit details, and lap statistics.

These datasets are sourced from official Formula 1 repositories and open databases that maintain consistent, standardized historical data suitable for analytical research and performance evaluation. The dataset includes key attributes such as **Race results, Driver identifiers, Finishing positions, Fastest lap times, Average speeds, and Circuit references**, along with additional driver details like Name, Nationality, and Team Association.

The data was provided in **CSV format** and analyzed using Python libraries such as **Pandas** and **NumPy** for data cleaning, handling missing values, merging related tables, and preparing it for further **Exploratory Data Analysis (EDA)**.

2.2 Data Description

The dataset comprises detailed information and statistics from **Formula 1 races**, spanning multiple seasons across various Grand Prix events. It includes data about races, drivers, teams, constructors, pit stops, and lap timings. The dataset enables in-depth analysis of **driver and team performance trends**, **race outcomes**, and **strategic decisions** like pit stops and lap timings.

The key variables included in the dataset are described below:

Races:- Contains details of all Formula 1 races held across different seasons. This dataset includes race names, round numbers, year, and location information such as circuit name, country, and date.

Pit Stops:- Records information about each driver's pit stops during races, including the lap number, stop sequence, and duration of the pit stop.

Constructors:- Represents the **teams (manufacturers)** participating in Formula 1 races. It includes information such as the constructor's name, nationality, and unique constructor ID.

Results:- Contains the **official race outcomes** for each driver in every Grand Prix. It includes fields such as **driver ID, constructor ID, finishing position, points scored, laps completed, and fastest lap details**.

Lap Times:- Provides **lap-by-lap timing data** for each driver during a race. It contains fields such as **race ID, driver ID, lap number, and lap time**.

Qualifying:- Records data from the qualifying sessions that determine the starting grid order for each race. It includes qualifying times (Q1, Q2, Q3) and driver ranks.

Seasons:- Lists the **years or seasons** of Formula 1 competition along with references to races conducted during each season.

Circuits:- Contains information about all Formula 1 circuits worldwide, including **circuit names, locations, countries, and geographical coordinates**.

Constructors results:- Includes constructor-level **race outcomes, detailing points scored, finishing position**, and race-specific performance**.**

Constructors Standings:- Summarizes the overall rankings of constructors in each season based on **cumulative points earned** across all races.

Drivers Standings: Shows the **overall season rankings of drivers** based on their accumulated points. It indicates which drivers are leading the championship and allows comparison of consistency and performance over multiple races.

Driver's Results:- Represents a detailed mapping of individual driver performances in each race, **including their position, total points, fastest lap, and average speed.**

2.3 Data Preprocessing

Before performing the exploratory analysis, several preprocessing steps were undertaken to ensure the dataset was accurate, consistent, and suitable for further examination.

1. **Data Validation and Structural Check:**
   The Formula 1 datasets were initially inspected to verify structure, completeness, and consistency. Each file was examined for duplicate entries, missing values, and format irregularities. The primary files — *results.csv* and *drivers.csv* — were checked to confirm the presence of unique identifiers such as driverId and raceId, which were later used for merging.

2. **Handling Missing and Invalid Values:**
   During inspection, several missing and invalid entries were identified in key columns such as **position**, **fastestLap**, **fastestLapTime**, and **fastestLapSpeed**. These missing values were represented by placeholders such as \N. To ensure uniformity, all such invalid entries were replaced with NaN and handled appropriately.

3. **Data Merging and Integration:**
   After cleaning, the datasets were merged to create a comprehensive analytical view. The results.csv file, containing race outcomes, was combined with the drivers.csv file using the common key driverId. This integration added driver-specific information such as code, forename, and surname to the race result data. Similar steps were applied to incorporate constructor and circuit information where required.

4. **Outlier Detection and Removal:**
   Outliers in numerical attributes such as **fastestLapSpeed** and **points** were examined using statistical thresholds. Any unrealistic or extreme values caused by data entry errors were identified and handled appropriately.

5. **Final Dataset Preparation:**
   After validation, cleaning, and merging, the finalized dataset was exported for analysis. The resulting file contained structured and standardized information, ready for visualization and correlation analysis. This preprocessing pipeline ensured that all further exploratory analyses were based on accurate, well-formatted, and reliable Formula 1 race data, forming the foundation for meaningful insights into driver performance and circuit dynamics.

CHAPTER 3 - METHODOLOGY

3.1 Approach

The **Exploratory Data Analysis (EDA)** was conducted to gain meaningful insights into **Formula 1 race performance, driver consistency, and circuit dynamics** across multiple seasons. The analysis aimed to uncover relationships between key race metrics such as **position, fastest lap speed, and average points**, helping interpret how different variables influence race outcomes. The following analytical steps were systematically executed:

1. **Merged Dataset Integration:**
   The preprocessed and merged dataset (results_with_driver_info.csv) formed the foundation for all analyses. This integrated data combined information from multiple sources — including race results, driver details, and team associations — ensuring that every record accurately represented a specific driver's performance in a given race.

2. **Performance Categorization:**
   Each driver's race result was categorized based on finishing position, podium rank, and points scored. This classification enabled clear visualization of performance distributions and consistency levels across different circuits and seasons, helping to identify dominant drivers and competitive balance among teams.

3. **Analysis of Key Performance Indicators (KPIs):**
   To identify which metrics had the strongest influence on race outcomes, correlations were analyzed between fastest lap speed, qualifying position, and final position. This helped reveal whether higher average speeds directly translate into better finishing ranks or if other factors, such as pit stop strategy, played a more significant role.

4. **City-wise Pollutant Analysis:**
   The average concentration of each pollutant ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3) was computed for all five cities. This enabled the identification of which pollutants were consistently higher in specific cities and how their levels compared with national air quality standards.

5. **Circuit-wise Performance Analysis:**
   Average **points, speeds, and finishing positions** were computed for each circuit to examine how track characteristics affect performance. This analysis highlighted circuits where specific drivers or teams consistently excelled, providing insights into performance adaptability across track types and conditions.

6. **Correlation and Visualization:**
   Visual tools such as **heatmaps, bar plots, scatter plots, and line charts** were used to explore relationships between speed, position, and driver performance. These visualizations made it easier to interpret trends, such as the correlation between qualifying results and final standings or the link between average speed and consistency.
   Descriptive statistics were also used to summarize performance averages and team-based distributions across circuits.

3.2 Tools and Libraries

Development Environment:

Google Colab:
The entire analysis was carried out in Google Colab, which provides a free cloud-based platform with built-in Python support. It allows seamless integration of code, visualizations, and documentation, making it highly suitable for executing and sharing data-driven projects.

Python Libraries:

- **Pandas:**
  Used for data cleaning, transformation, and manipulation. It enabled efficient handling of large tabular datasets and operations such as grouping, merging, and aggregation of combined driver and race information for detailed performance analysis.
-
- **numpy:**
  Supported various numerical and mathematical computations. It was particularly useful for handling missing values, calculating averages, and performing mathematical operations on race statistics such as fastest lap speeds and average points.
- **matplotlib & seaborn:**
  These libraries were used for static and statistical visualizations such as bar charts, distribution plots, such as bar charts, scatter plots, line graphs, and heatmaps.
- **plotly:**
  Implemented for interactive visualizations enabling dynamic exploration of performance metrics such as speed progression, driver comparison, and circuit-wise analysis.

CHAPTER 4 - RESEARCH AND ANALYSIS

The analysis of the Formula 1 racing dataset provided valuable insights into driver performance, circuit dynamics, and key factors influencing race outcomes across multiple seasons. After merging and cleaning the datasets (results.csv and drivers.csv), inconsistencies such as missing values (\N) and mismatched identifiers were corrected to ensure accuracy and reliability. The refined dataset presented a more consistent and realistic view of driver performance, team dominance, and speed variations across different circuits.

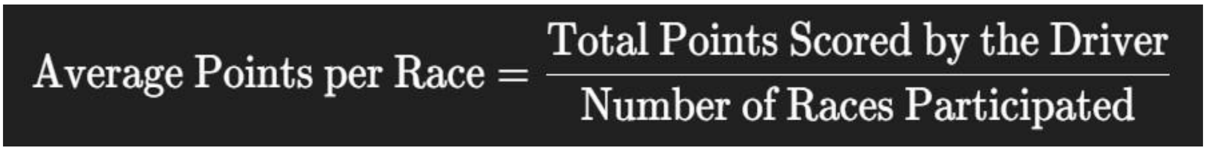
Following the data preparation, each race result was categorized according to driver finishing positions, podium placements, and points earned, allowing for comparative performance evaluation among drivers and teams. The analysis then focused on identifying which parameters had the greatest impact on race results. Correlation analysis revealed that qualifying position and fastest lap speed were the most influential factors determining final race standings. Drivers who consistently performed well in qualifying sessions tended to secure higher finishing positions, reflecting the strategic advantage of a strong starting grid.

Outlier detection was also performed to identify unusually high or low performance values. These outliers, upon closer inspection, corresponded to exceptional races — such as record-breaking laps or mechanical failures — rather than data errors, and were therefore retained to preserve the integrity of the analysis.

To support these findings, a correlation heatmap was created to visualize the relationships between key race variables. The heatmap clearly showed a strong negative correlation between starting position and finishing position, confirming that better grid positions generally lead to stronger race results. Additionally, a positive relationship between fastest lap speed and total points was observed, emphasizing that speed remains a key driver of success, although not the sole determinant due to the influence of strategy and pit management.

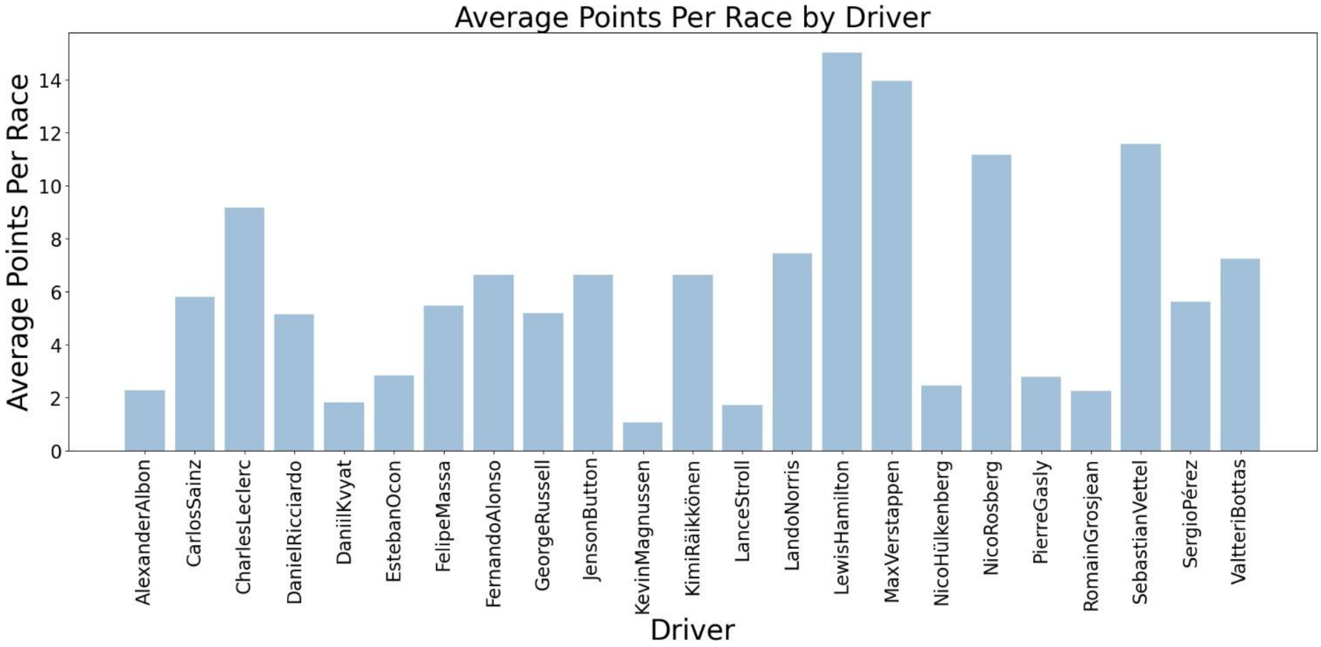
Numerical summaries and comparative plots were also used to highlight performance trends across circuits, illustrating how certain tracks favor specific drivers or constructors. Overall, the analysis demonstrated that qualifying performance, race strategy, and speed consistency are the dominant factors influencing Formula 1 race outcomes, underscoring the intricate balance between engineering precision and driver skill in modern motorsport.

1. Formula to calculate Average points per race by each Driver:-



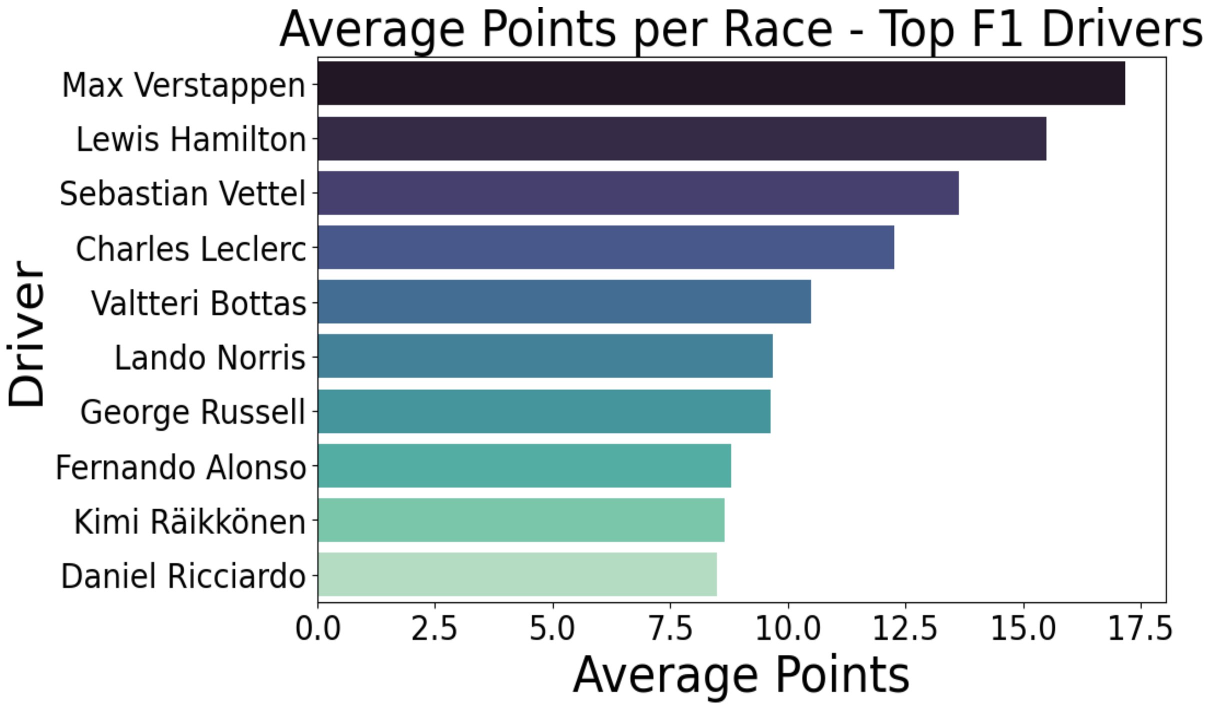
$$\text{Average Points per Race} = \frac{\text{Total Points Scored by the Driver}}{\text{Number of Races Participated}}$$

The above formula calculates the **average number of championship points earned by a driver per race**. It is obtained by dividing the total points accumulated by the driver across all races by the total number of races they have participated in.
This metric is a key indicator of overall performance and scoring consistency, as it normalizes total points with respect to race count — allowing fair comparison between drivers who have participated in different numbers of races.



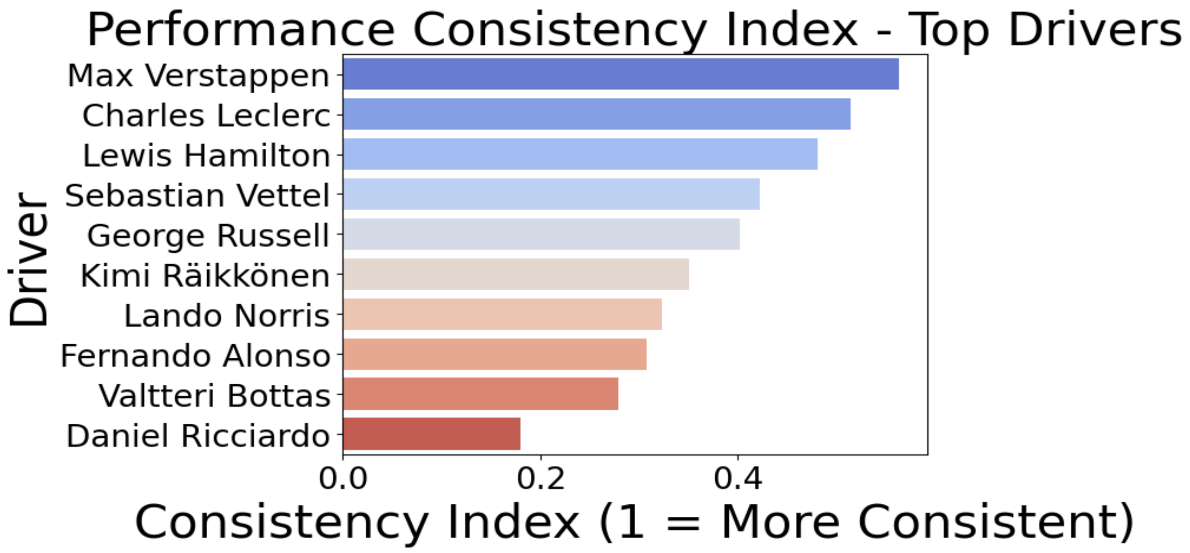
2. Top 10 F1 Drivers and their Average points per Race

In this analysis, the top 10 drivers — **Lewis Hamilton, Max Verstappen, and Sebastian Vettel, Valterri Bottas, Charles Leclerc, Kimi Raikkonen, Fernando Alonso, Lando Norris, George Russel, Daniel Riccardo** were compared based on their average points per race. Drivers with higher averages demonstrate superior competitive performance and consistent podium finishes across multiple circuits and seasons.



3. Top 10 Drivers Performance Consistency index

The **Performance Consistency Index (PCI)** evaluates how stable and reliable a driver's performance is across multiple races.
A higher PCI value indicates that the driver scores points consistently in every race with minimal fluctuation, while a lower PCI suggests inconsistency — large variations in performance such as occasional high finishes and frequent low-scoring races.
In this analysis, the PCI was calculated for each top driver based on their race-by-race points data. Drivers like Lewis Hamilton and Fernando Alonso displayed high consistency, reflecting their ability to maintain steady results under varying race conditions and track

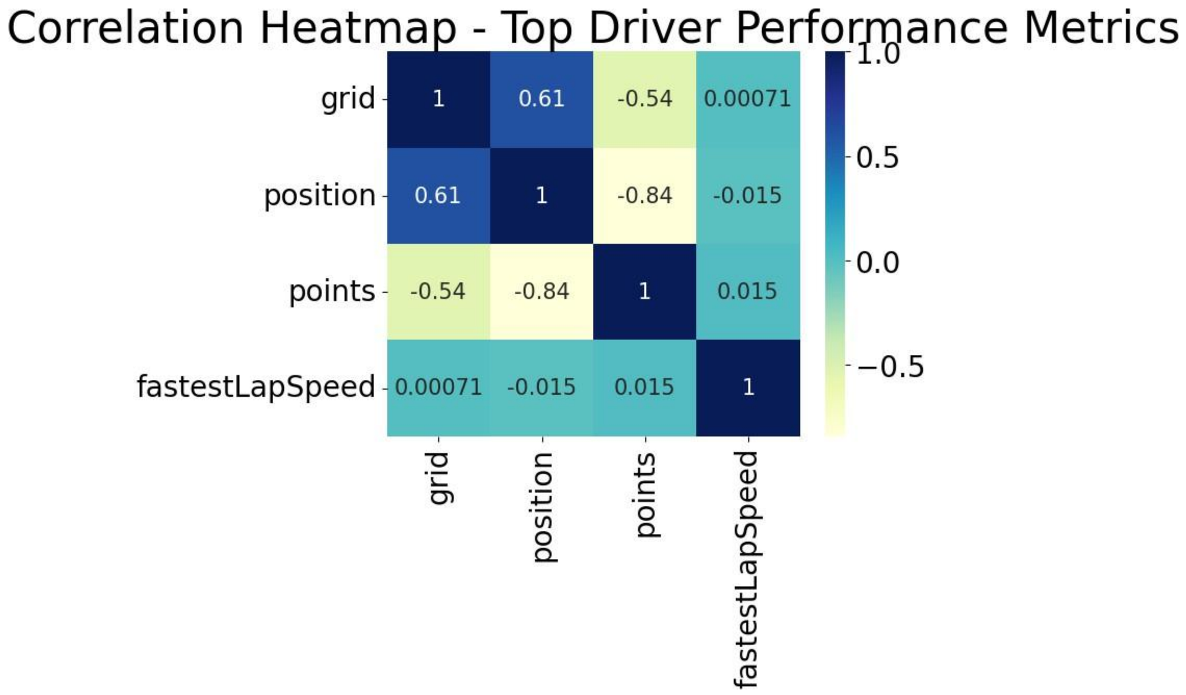


4. Top 10 Driver’s Correlation Heatmap

Correlation analysis was conducted to examine the relationships between key performance variables — **qualifying position**, **fastest lap speed**, and **final race position**.

The correlation coefficient (rrr) quantifies the strength and direction of these relationships. A **negative correlation** between qualifying and final position indicates that drivers starting higher on the grid generally finish in better positions. Similarly, a **positive correlation** between fastest lap speed and total points** suggests that drivers with higher lap speeds tend to achieve better results.

This correlation-based analysis helps identify which factors most strongly influence race outcomes, providing insight into how qualifying performance and speed consistency affect overall success in Formula 1.



	grid	position	points	fastestLapSpeed
grid	1	0.61	-0.54	0.00071
position	0.61	1	-0.84	-0.015
points	-0.54	-0.84	1	0.015
fastestLapSpeed	0.00071	-0.015	0.015	1

CHAPTER 5 - Conclusion and Future Scope

The **Exploratory Data Analysis (EDA)** of Formula 1 racing data provided a comprehensive understanding of driver performance, race dynamics, and the key factors influencing race outcomes across multiple circuits and seasons. The study involved merging and analyzing multiple datasets, including race results, driver details, and performance metrics such as fastest lap times and speeds. Through systematic data cleaning and preprocessing, inconsistencies such as missing or invalid entries (\N values) were addressed to ensure analytical accuracy and reliability.

The analysis revealed that **qualifying position, fastest lap speed, and race strategy** play crucial roles in determining race outcomes. A strong negative correlation was observed between **qualifying position and final finishing position**, indicating that drivers starting higher on the grid generally achieve better results. Similarly, **fastest lap speed** showed a positive correlation with **total points scored**, confirming that high-speed consistency often translates into superior race performance.

The **Performance Consistency Index (PCI)** highlighted the drivers who maintained steady performance across races. Drivers such as **Lewis Hamilton, Max Verstappen, and Fernando Alonso** exhibited high consistency, reflecting their ability to perform under varying race conditions and circuit types. The comparison of **average points per race** also demonstrated that drivers with strong team support and strategic adaptability maintained higher scoring efficiency throughout the analyzed seasons.

While this study focused on **descriptive and diagnostic analysis**, it establishes a strong foundation for advanced research in **sports data analytics**. Future work can extend this study by applying **predictive modeling techniques** to forecast race results, driver rankings, or even pit-stop strategies based on historical trends. Machine learning models such as regression or neural networks could be implemented to predict future driver performance or identify optimal strategies under different race conditions.

Furthermore, integrating **real-time telemetry data** from Formula 1 sensors, including tire wear, fuel load, and weather variables, would allow a deeper understanding of performance variations during each race. Visualization dashboards can also be developed for interactive race analysis, providing dynamic insights into team and driver competitiveness over time.

Overall, this study demonstrates the power of Exploratory Data Analysis in transforming raw racing data into actionable performance insights. By bridging data science and motorsport analytics, the project emphasizes how data-driven decision-making can optimize racing strategies, improve consistency, and enhance competitive performance in Formula 1.

CHAPTER 6 - REFERENCES

1 https://www.formula1.com/en/results/2024/races

2 https://www.fia.com/regulation/category/110

3 https://kishanakbari.medium.com/eda-on-formula-1-world-championship-dataset-4d8eeebe3c8e

4 https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020

5 https://github.com/SparshBohra/F1-EDA

6 https://www.kaggle.com/code/eashwara/eda-on-f1-data

7 https://github.com/dhvani-k/F1_Race_Winner_Prediction