**Formula 1 Performance Analysis**

Jack Wilson 5/4/2025

**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 |

Figure 1: Point Inflation

The number of points is an important measure in driver and constructor success. Therefore, these inconsistencies mean machine learning will primarily focus on the 2010 to 2024 data.

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 data scraped from the [Official Formula 1 Results](https://www.formula1.com/en/results/2025/races) site. Results from each individual race from 1950 to 2024 were collected, merged into a comprehensive dataframe, cleaned, and then aggregated by season.

The following tables were scraped:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table** | **Description** | **Columns** | **Source** |
| drivers\_info | Code and nationality for each driver | year, driver, driver\_code, nationality | [Driver Standings](https://www.formula1.com/en/results/2024/drivers) |
| races\_fastest\_laps | Fastest lap for each race | race\_url, driver, fastest\_lap | [Fastest Laps](https://www.formula1.com/en/results/2024/races/1229/bahrain/fastest-laps) |
| races\_pits | Total pit stop count for each driver in a race | race\_url, driver, pit\_count | [Pit Stop Summary](https://www.formula1.com/en/results/2024/races/1229/bahrain/pit-stop-summary) |
| races\_results\_raw | Finishing data for each driver in a race | race\_url, driver, end\_position, team\_name, laps\_completed, time\_gap, points | [Race Result](https://www.formula1.com/en/results/2024/races/1229/bahrain/race-result) |
| races\_start | Starting data for each driver in a race | race\_url, driver, start\_position, team\_name, qualifying\_time | [Starting Grid](https://www.formula1.com/en/results/2024/races/1229/bahrain/starting-grid) |
| races\_winner | Winning driver data for each race | race\_url, year, race, round, date, winner, winning\_team, laps, total\_time | [Results](https://www.formula1.com/en/results/2024/races) |

Figure 2: Scraped Data

These tables were merged using common keys like race\_url, year, and driver to create an extensive dataframe ready for cleaning.

**2.1 Cleaning**

Along with basic data cleaning (e.g., removing duplicates and imputing missing values), two key columns were engineered to improve interpretability and analysis.

* *position\_status*: Categorizes a driver race result based on the end\_position column
  + DQ = Disqualified
  + DNF = Did Not Finish
  + NC = Not Classified (did not complete enough laps or crashed)
  + CLAS = Classified (finished the race)
* Continent: Maps a driver to a continent based on their nationality (e.g., GBR >> Europe)

Additionally, constructors names were standardized to their common names (e.g., McLaren Mercedes >> McLaren) to simplify analysis and reduce noise caused by naming variations due to engine affiliations.

**2.2 Aggregation & Feature Engineering**

Since the raw data contains individual race results per driver across seasons, the data was aggregated at the season level. For each driver and season pair, statistics like avg\_finish\_pos and total\_points were computed.

To enhance machine learning and other analysis, additional feature were engineered. These include:

* Performance metrics
  + pct\_season\_points, lap\_completion\_rate, fastest\_lap\_rate, podium\_rate, points\_rate
* Experience and reliability
  + years\_experience, dnf\_rate, finish\_rate, consistency\_score, qualifying\_consistency, reliability\_score
* Qualifying vs Race Comparison
  + place\_differential, qualifying\_vs\_race, race\_improvement
* Impact Indicators
  + total\_podium\_finishes, total\_points\_finishes, championship\_impact

A binary label was also created to indicate whether or not a driver was retained by the same team in the following season.

The final analysis-ready dataset contains 3510 rows and 50 features.

**2.3 Data Dictionary**

|  |  |  |
| --- | --- | --- |
| **Column** | **Description** | **Type** |
| year | The season year | Numeric |
| driver | Driver name | Text |
| team\_name | Name of the constructor/team | Text |
| continent | Continent of the driver | Text |
| season\_rank | Driver’s final rank in the season | Numeric |
| avg\_start\_pos | Average start position across the season | Numeric |
| std\_start\_pos | Standard deviation of starting positions | Numeric |
| best\_start\_pos | Best start position | Numeric |
| worst\_start\_pos | Worst start position | Numeric |
| races\_participated | Number of races competed in | Numeric |
| avg\_finish\_pos | Average finishing position | Numeric |
| std\_finish\_pos | Standard deviation of finishing positions | Numeric |
| best\_finish\_pos | Best race finish | Numeric |
| worst\_finish\_pos | Worst race finish | Numeric |
| total\_pos\_gain | Positions gained from start to finish across all races | Numeric |
| avg\_pos\_gain | Average position gain per race | Numeric |
| total\_points | Total points scored | Numeric |
| avg\_points\_per\_race | Average points scored per race | Numeric |
| total\_pct\_season\_points | Percentage of total available season points earned | Numeric |
| avg\_pct\_season\_points | Average percentage of available season points per race | Numeric |
| worst\_pct\_season\_points | Lowest points percentage in a race | Numeric |
| best\_pct\_season\_points | Highest points percentage in a race | Numeric |
| total\_laps\_completed | Number of laps completed | Numeric |
| total\_laps | Total laps possible | Numeric |
| total\_fastest\_laps | Number of fastest laps earned | Numeric |
| CLAS\_total | Number of classified finishes | Numeric |
| DNF\_total | Number of Did Not Finish results | Numeric |
| DNS\_total | Number of Did Not Start results | Numeric |
| DQ\_total | Number of Disqualified results | Numeric |
| NC\_total | Number of Not Classified results | Numeric |
| backmarker | Backmarker indication | Bool |
| midfield | Midfielder indication | Bool |
| podium\_regular | Podium regular indication | Bool |
| points\_regular | Points regular indication | Bool |
| years\_experience | Number of years the driver has competed in | Numeric |
| total\_podium\_finishes | Total podium finishes in the season | Numeric |
| total\_points\_finishes | Total points finishes in the season | Numeric |
| dnf\_rate | Percentage of DNFs over races entered | Numeric |
| finish\_rate | Percentage of races finished | Numeric |
| lap\_completion\_rate | Percentage of laps completed | Numeric |
| fastest\_lap\_rate | Fastest laps as a percentage of races entered | Numeric |
| podium\_rate | Podium finishes as a percentage of races entered | Numeric |
| points\_rate | Points finishes as a percentage of races entered | Numeric |
| consistency\_score | Consistency in finish positions | Numeric |
| qualifying\_consistency | Consistency in start positions | Numeric |
| reliability\_score | Measures a drivers finish rate with lap completion rate | Numeric |
| qualifying\_vs\_race | Comparison of qualifying vs finishing positions | Numeric |
| championship\_impact | Strength of contribution to the championship outcome | Numeric |
| race\_improvement | Average improvement from start to finish positions | Numeric |
| driver\_retained | Driver retained indication | Bool |

Figure 3: Data Dictionary

**3. Analysis**

3.1 Driver Dominance

Formula 1 has been criticized for being dominated by a small group of drivers, especially with the recent domination of Lewis Hamilton. Understanding whether this pattern holds across decades can help assess the need for structural changes, such as budget caps or technical regulations. But is the perception of dominance statistically supported?

A graph with different colored lines

AI-generated content may be incorrect.To explore this, a line chart was plotted showing the rankings of the five most tenured drivers across their careers. This visualization highlights each driver’s “prime” period—Hamilton from 2014-2020, Vettel in 2017-2018, and Perez more recently in 2023.

Figure 4: Top 5 Tenured Drivers’ Rankings Over Time

To determine if dominance has increased over time, a Pearson Correlation test was conducted between driver rank and year. The result was a correlation coefficient of -0.103 and a p-value of 0.051. Since this p-value slightly exceeds an alpha level of 0.05, the correlation is not statistically significant. This suggests that while certain drivers have had dominant stretches, the overall trend in dominance has increased meaningfully over time.

3.2 Nationality Dominance

Despite Formula 1’s global popularity, the distribution of drivers by nationality reveals a heavy concentration from Europe. Analyzing the distribution of continents and their performance in the Drivers’ Championship provides information on potential cultural barriers to entry.

Pie charts clearly show that the majority of F1 drivers have come from Europe. While this reflects the sport’s historical roots and strong presence in the UK, it also signals the persistent geographic imbalance.

A pie chart with different colored circles

AI-generated content may be incorrect.Figure 5: Distribution of Drivers by Continent

This imbalance is reinforced through a bar chart, which reveals a large disparity in driver counts between Europe and all other continents.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 6: Count of Drivers by Continent

However, representation alone does not equate to performance. When performance metrics such as average season rank and average total percent points are analyzed by continent a surprising result emerges: drivers from Oceania outperform all other continents in terms of lowest average season rank and highest average total percent points.

A graph of different red bars

AI-generated content may be incorrect.

Figure 7: Average Driver Performance by Continent

3.3 Experience vs Performance

Team principals must decide whether to invest in new talent or retain experienced drivers. Understanding how performance evolves with experience can help them make recruitment and retention decisions.

A bar chart of average season rank by years of experience shows that drivers performance improves notably in the first 10 years, peaking around years 10-13. Drivers past year 13 tend to underperform compared to earlier years.

A graph of a number of years

AI-generated content may be incorrect.

Figure 8: Average Season Rank by Years of Experience

To test the strength of this relationship, Pearson correlation coefficients were calculated. The correlation between season rank and experience was -0.378 with a of p-value of 0.000, indicating a statistically significant negative relationship. As experience increases, rank improves.

Similarly, the correlation between percent season points and experience was 0.326 with a p-value of 0.000, indicating a statistically significant positive relationship. As experience increases, percent season points increase.

While neither correlation is particularly strong, both are meaningful. A scatterplot helps visualize the upward performance trend during a driver’s early and mid-career.

A comparison of red dots

AI-generated content may be incorrect. Figure 9: Correlation Between Experience and Performance

3.4 Driver Retention

Driver retention plays a critical role in team strategy. Constructors must weigh performance metrics against financial constraints and long-term goals when deciding whether to keep a driver for the following season. Identifying performance patterns associated with driver retention helps clarify the balance between consistency and competitiveness that teams value.

Multiple different classification models were tested using data from 2010 to 2024. The most accurate model was a Random Forest classifier, trained on an 80/20 training-test split. This model achieved an accuracy of 76.71% and an F1 score of 0.8046, indicating effective classification performance.

Feature importance scores identified average finishing position, qualifying consistency, and total season points as the most influential predictors of retention. These findings suggest that consistent performance over time is more valued than isolated strong results. The model’s confusion matrix shows a high false positive rate, indicating a tendency to over-predict retention and classify some drivers as retained when they were not.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 10: Random Forest Confusion Matrix

This tendency highlights the limitations of a purely performance-based model. Off-track factors like contract length and internal team dynamics are not captured by the current dataset. Merging or calculating this data would likely improve prediction accuracy and reduce misclassifications.

3.5 Total Points Analysis

Points serve as the primary performance metric in Formula 1. They directly determine driver and constructor standings and influence team funding. Accurately predicting total season points could help teams scout for new talent or adjust mid-season race strategies.

Several regression models were evaluated on data from 2010-2024. A Linear Regression model produced the best results, achieving an R2 score of 0.9639 using an 80/20 training-test split. This indicates that the model explained 96.39% of the variance in total season points

The most influential predictor was fastest lap rate, highlighting both car and driver competitiveness. While fastest laps do not always go to the top finishers, they generally indicate strong race pace. Additional significant predictors included DNF and finish rate, both of which reflect the critical role of reliability in point accumulation.

A visualization of predicted and actual results shows strong alignment, though the model cannot account for off-track factors such as weather or injuries that might cause a driver's performance to dip. Future modeling could explore these features to better capture mid-season shifts.

A graph with red dots

AI-generated content may be incorrect.

Figure 11: Actual vs Predicted Total Points

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

# Proposal / Check-in Challenges

- The initial data collection approach involved scraping season-level results (i.e., final standings for all drivers in a season) and merging that data with a downloaded dataset containing driver nationalities and birthdates. However, the downloaded dataset was incomplete, resulting in many missing values after the merge.

- Because only final standings were scraped, the dataset included limited numerical features (driver position, driver points, team position, and team points). This restricted analysis and led to weak insights.

- To address these issues, results from individual races were scraped instead. This provided more numerical variables that could later be aggregated by season. Since nationality could also be scraped and years of experience could be calculated, the downloaded dataset was ultimately unnecessary, and all data collection was done solely through scraping.

**4. Conclusion**