Predicting Formula 1 Podium Finishes: A Data-Driven Approach

Author: Matthew Lertsmitivanta

Course: OBA 465

Date: 05/14/2025

Team Position and Motivation:

This project was approached from the perspective of a Sports Analytics team working for a professional Formula 1 team. The goal was to provide evidence-based insights to optimize team strategy and driver recruitment. The motivation behind this analysis was to evaluate what determines performance in Formula 1—from raw speed to podium finishes. This included understanding how a driver's result is influenced by team quality, qualifying performance, and in-race execution. The ultimate aim was to support better, data-informed decisions for team management.

Executive Summary:

Formula 1 performance data from the past ten seasons (2014–2024) was analyzed to uncover what factors drive speed and success. The initial question was: "**Do top‑team drivers have significantly faster average lap times than other drivers across the last 10 seasons?**" To answer this, lap time data was used to compare top-team drivers with their midfield counterparts. The analysis showed that drivers from dominant teams (Mercedes, Red Bull, Ferrari, McLaren) do tend to be faster, supported by t-tests and regression analysis.

However, this finding confirmed a commonly accepted truth in Formula 1: better teams usually produce faster lap times. To generate more actionable insights, the focus shifted to a more practical and predictive question: "**Using qualifying times, sprint‑race results and pit‑stop efficiency, can each driver's probability of finishing on the podium be predicted?**" This reframed the project from descriptive comparisons to predictive modeling, providing greater value for decision-making. This insight can help the team make smarter choices about race strategy, driver prioritization, and expectations before the race even begins.

Description and Discussion of the Question:

The initial question, whether top-team drivers had significantly faster average lap times, helped establish a baseline understanding of performance differences across teams. The final and more impactful question asked whether podium finishes could be predicted using pre-race data such as qualifying performance, sprint results, and pit-stop efficiency. This tackled the core challenge in F1: combining race metrics to inform strategy before the race begins.

Variables Discussed:

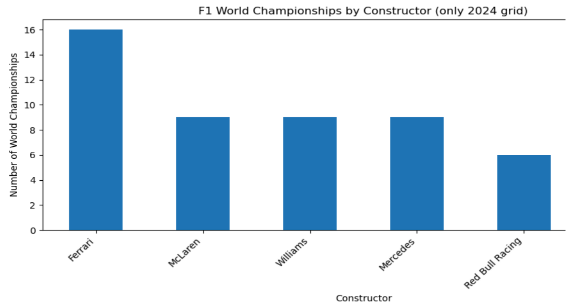
- **milliseconds:** Lap time recorded for a single lap.
- **team:** Constructor name.
- **is_top:** Whether the constructor is one of the top four.
- **circuitId**: The racetrack ID.
- **year:** The race season.
- **qualifying_position:** The driver's position in qualifying.
- **sprint_result:** Finishing place in the sprint race.
- **pit_stop_time:** Efficiency of pit stops.
- **podium_finish:** Binary yes/no indicator of podium outcome.

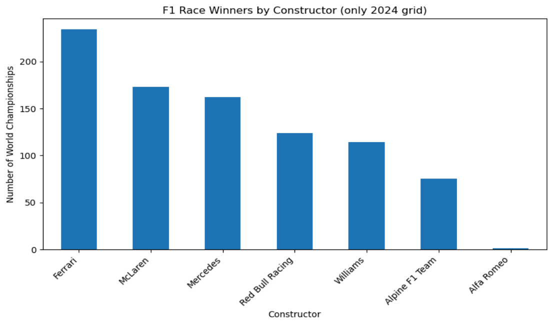
Constructed Variables:

- **canonical:** Standardized constructor names across time.
- **gap_ms:** Yearly lap time difference between top and midfield.
- **podium_finish:** Created from final race results.

Visuals

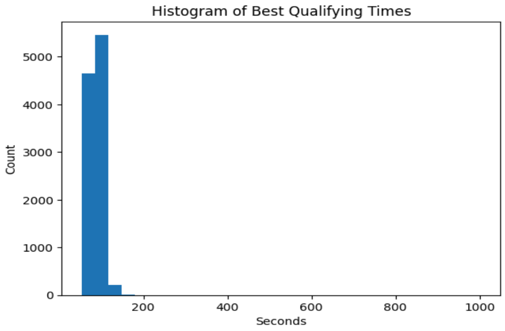
Determining Top Team Drivers (Initial Question)

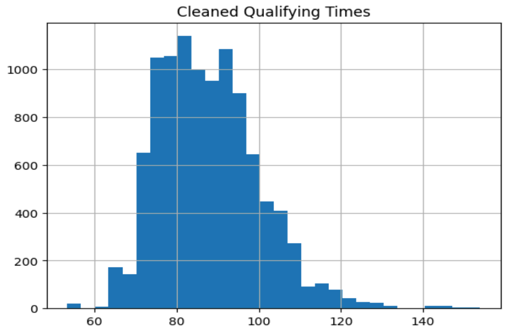




Top teams were identified by summarizing data on the number of world championships and races won by each constructor. The top teams for the analysis were: **Ferrari, McLaren, Mercedes, and Red Bull Racing.**

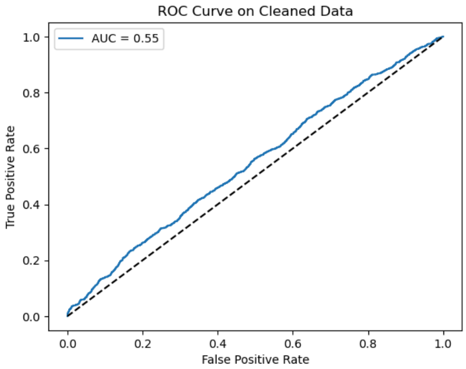
Cleaned Qualifying Times Distribution

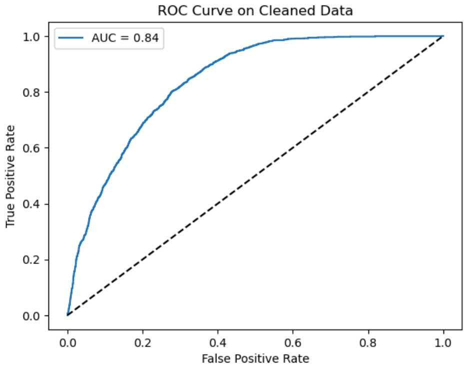




The initial histogram included uncleaned data with extreme outliers; some qualifying times were well beyond the realistic range (e.g., over 200 seconds), likely due to red flags, DNS (did not start), or placeholder values. After removing these outliers, the cleaned histogram showed a tight, normal distribution between 75–100 seconds, which aligns with typical F1 qualifying laps. This confirmed that the data was well-prepared for statistical analysis.

ROC AUC Curves





These two charts compare how well each model predicts podium finishes using ROC AUC, a metric that measures how effectively a model separates podium and non-podium drivers. A score of 0.5 indicates random guessing, while 1.0 reflects perfect accuracy. The first model, which used only race-day metrics such as qualifying time, sprint finish, and pit-stop duration, achieved an AUC of 0.55 and offered limited predictive value. When a driver's historical podium rate was added, model accuracy increased significantly. The second model reached an AUC of 0.84 and was able to correctly distinguish podium finishers in most cases.

Analytics Flowchart (Appendix Fig 1.1)

1. Evaluate Formula 1 historical race data
2. Develop program to access data from Kaggle
3. Get qualifying times, sprint-race results, and pit-stop efficiency features over all seasons
4. Split qualifying data into Q1, Q2, Q3, and merge all data together
5. Bootstrap data to obtain a normal distribution
6. Conduct statistical tests to find significance
7. Evaluate statistical findings
8. Conclude on the hypothesis
9. Communicate results to stakeholders

Rationale for Tools

Python was chosen for its flexibility and efficiency in handling large datasets. Pandas was used for data processing, Matplotlib for visualization, and SciPy/Statsmodels for statistical testing and modeling. The project moved from basic comparisons to logistic regression, a tool for estimating probabilities for binary outcomes like "Did this driver reach the podium?"

Understanding the Model's Role

The model uses inputs like qualifying time, sprint result, pit-stop duration, and a driver's podium rate to predict the likelihood—expressed as a percentage—that a driver will finish on the podium. This makes it a strategic tool for pre-race decisions. Race engineers and strategists can input known values before the race and use the model to estimate which drivers are most likely to score a podium, supporting decisions around team orders, pit strategy, and expectations.

Conversion from Logic to Tools:

The analysis began with descriptive comparisons to explore differences in average lap times. Logistic regression was then used to evaluate the impact of different race and historical features on the probability of finishing on the podium. The model structure allowed multiple predictors to be considered simultaneously—mirroring how real F1 strategists weigh several factors at once.

Primary Findings

- Top-team drivers are significantly faster on average.
- Qualifying time and sprint race results are strong predictors of podium finishes.
- Pit-stop efficiency adds further predictive value, especially when margins are tight.
- The final model provides probability estimates that can support real-time race strategy and driver comparison.

Key Results

Two-Sample t-Test: (t = –5.870, p = 5.08 × 10^{-9}).

Baseline Logistic Model

- Pseudo R^2 = 0.006
- ROC AUC = 0.547
- const: coef = 0.1954 (p = 0.498)
- best_qualifying_sec: coef = –0.0127 (p < 0.001)
- sprint_position_filled: coef = –0.0436 (p < 0.001)
- avg_pit_duration_imputed: coef = 0.0002 (p = 0.258)

Full Logistic Model (with driver_podium_rate)

- Pseudo R^2 = 0.212
- ROC AUC = 0.841
- const: coef = –2.0749 (p < 0.001)
- best_qualifying_sec: coef = –0.0078 (p=0.004)
- sprint_position_filled: coef = –0.0227 (p=0.059)
- avg_pit_duration_imputed: coef = 0.0002 (p=0.242)
- driver_podium_rate: coef = 6.9809 (p < 0.001)

Summary of Key Results

- **t-Test Result:** Drivers who make it to the podium qualify significantly faster than those who don't (t = –5.87, p < 0.00001).
- **Baseline Model (ROC AUC=55%):** Using only race-day data (qualifying, sprint, pit stops), the model performs slightly better than guessing: it explains less than 1% of the variation and predicts correctly just 55% of the time.
- **Full Model with History (ROC AUC=84%):** Once including how often each driver has finished on the podium in the past, the model gets much smarter. It correctly predicts podium finishes 84% of the time and explains over 20% of the variation. This doesn't mean the model perfectly predicts who will podium in each race, but rather that it correctly ranks podium vs. non-podium drivers 84% of the time.
- **Most Important Variable:** A driver's past podium rate is by far the best predictor, drivers who frequently finished in the top 3 before are far more likely to do it again.
- **Other Variables:** Qualifying speed still matters, but less so when a driver's history is known. Sprint placement and pit-stop efficiency have minimal added value when history is considered as well (known by the correlation coefficient).

What the Coefficients Mean in Plain Language

- A driver with a 10% higher podium rate (e.g., 0.30 vs. 0.20) is significantly more likely to finish on the podium, the biggest single influence on the model.
- Every additional second in qualifying slows a driver's predicted podium odds by about 0.7%.
- Sprint finish position and pit stop duration have smaller and less reliable effects in comparison.

Predictive Example: How Podium History Changes Probabilities

- Driver A qualified in 5th, had a sprint finish of 6th, a pit stop of 2.6 seconds, a podium rate of 0.15.
- Driver B had the same race-day stats but a podium rate of 0.40.
- The model gives Driver A a 33% chance of finishing on the podium. For Driver B, that rises to 72%.

This shows that historical performance dominates in prediction—even when same-day performance is equal.

How to Act on These Results

- When choosing between two similarly performing drivers, favor the one with a stronger podium track record.
- Use the model during pre-race briefings to align team expectations and prioritize support.
- Consider podium history when assigning team orders or riskier pit strategies—data shows it matters more than a sprint finish alone.

How to Use This Model in Practice

- **Inputs Needed:** Qualifying time (best lap), sprint result, pit stop time, and the driver's career podium rate.
- **When to Use:** Before the race or during qualifying to assess podium potential.
- **Who Should Use It:** Race strategists, performance engineers, and decision-makers setting race-day priorities.
- **What It Does:** Outputs a podium probability from 0% to 100% per driver. This supports decisions like pit strategy, team orders, and internal benchmarks.

Extensions and Recommendations

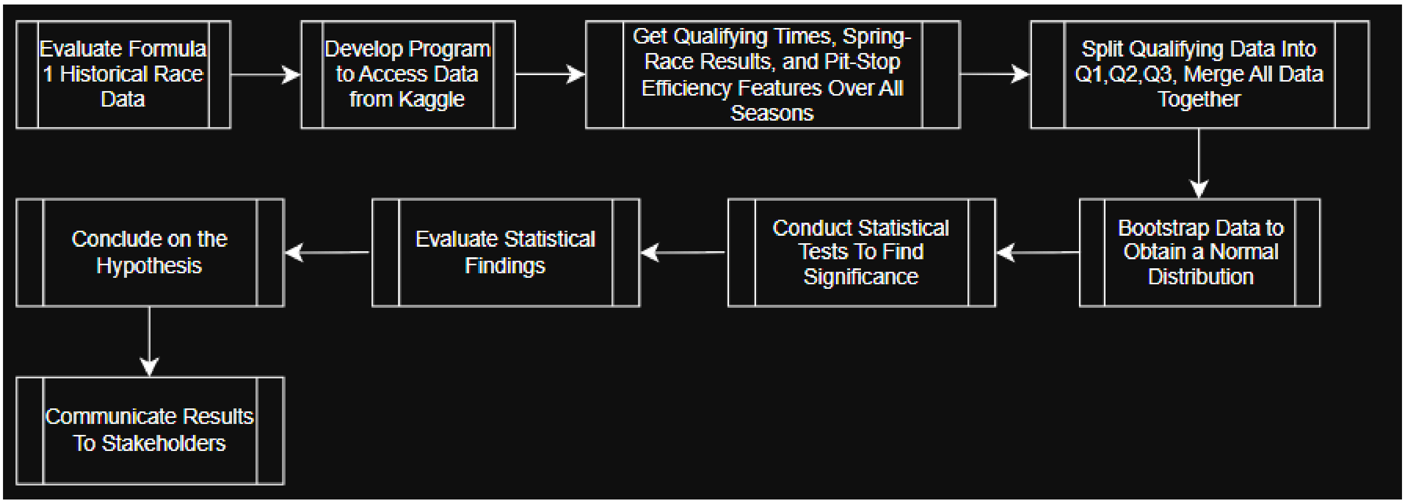
Future models could include weather conditions, tire choice, and real-time race variables like safety cars or position changes. Expanding this approach to predict full race position or season-long performance would also be valuable for strategic planning.

Sources

- Kaggle Formula 1 Dataset (constructor standings, races, results, lap times, qualifying)
- Canonical team mapping developed internally
- Qualifying and race dataset from qualifying_structured.ipynb & Formula_1_Project.ipynb

Appendix

Flowchart (Fig 1.1)


Evaluate Formula 1 Historical Race Data
Develop Program to Access Data from Kaggle
Get Qualifying Times, Spring-Race Results, and Pit-Stop Efficiency Features Over All Seasons
Split Qualifying Data Into Q1,Q2,Q3, Merge All Data Together
Bootstrap Data to Obtain a Normal Distribution
Conduct Statistical Tests To Find Significance
Evaluate Statistical Findings
Conclude on the Hypothesis
Communicate Results To Stakeholders