

Determining Best Predictors of F1 Race Results in the Hybrid Era

By Rody Bertolini

CSPB

University of Colorado Boulder

Denver, CO

pabe9903@colorado.edu

ABSTRACT

F1 is often categorized as a sport defined by engineering, where the winner of each race is primarily determined by the design and performance of the car, rather than the driver, the pit crew, race strategy or other factors. This project seeks to examine that commonly held wisdom by evaluating what aspect of a race has the greatest influence on the finishing position. This project specifically focuses on the Hybrid era (2014-Present) using more than 4,600 driver-race entries.

This Project uses the Formula 1 World Championship dataset derived from the Ergast API, and focus on nine relational tables, including results, qualifying, constructors, drivers, circuits, pit stops, and lap time, into a unified dataset with a granularity of one row per driver per race. Extensive preprocessing was used to standardize variable names, reconcile categorical identifiers, and remove missing values. Engineer features such as pitstop totals and lap time summaries, and then that data was then limited by year to remove older data with different rulesets which would cause fundamental differences in the data, such as modern pit stops not allowing refueling and thus greatly reducing pit times.

Exploratory data analysis and correlation assessment demonstrates that qualifying position is the strongest predictor of finishing position, which is followed by the constructor. This shows that modern F1 outcomes are dominated by grid position at the start of the race, which makes sense given the limited ability to pass in the modern era, with machine performance and strategy having lesser impacts.

INTRODUCTION

Formula 1 is widely considered to be the most technologically advanced sport in the world, with each race being determined by a combination of driver skill, car performance, team operations, and even race specific factors like slower corners vs. faster ones. It's a commonly held

belief that F1 is primarily car dependent, this is reflected in the current season by the assumption that Lando Norris will win the championship because he has the fastest car. However, despite having a demonstrably inferior car, Max Verstappen is only a few points behind Norris in the championship, demonstrating that there must be other factors in play. Therefore, this project seeks to analyze which factor has the strongest impact on race performance, to examine the commonly held belief against actual data.

This project focuses on the Hybrid era, which began in 2014, because the technical and sporting regulations prior to that dramatically alter the data. For instance, prior to 2009 cars would refuel during pit stops, dramatically increasing the length and relevance of the pit stop. Restricting the study to this period ensures consistency and avoids confounding effects from earlier seasons.

To address the central question, I constructed a unified dataset from publicly available Formula 1 information. This dataset integrates 9 tables and has over 4,600 entries. A comprehensive preprocessing pipeline was developed to standardize column names, and reduce inconsistencies between tables.

The primary objective of this study is not to predict race outcomes, but to determine which race specific attributes are most important. By applying exploratory data analysis and data-mining techniques, this project provides insight into the underlying structure of competitive performance in modern Formula 1, informing discussions surrounding the relative importance of engineering, strategy, and driver skill. This question provides a natural application for data-mining, it requires integrating datasets, engineering predictive features, and comparing the influence of multiple variables.

RELATED WORK

While there is academic research on Formula 1 in areas such as vehicle dynamics, aerodynamics, and even sociological studies of fandom, I was unable to find published research that directly analyzes which race-specific factors most strongly influence finishing position. Public data-oriented sites such as Tracing Insights (tracinginsights.com) provide race results, lap times, and basic exploratory visualizations, but they do not perform multivariate analysis or examine the comparative impact of qualifying performance, constructor strength, pit-stop execution, and lap-time characteristics in the way this project does. As a result, this work addresses a gap by focusing explicitly on factor importance within race outcomes.

DATASET

This project uses the publicly available Formula 1 World Championship dataset, which provides historical race information in a relational format. The dataset consists of nine core tables: races, results, drivers, constructors, circuits, status, qualifying, lap times, and pit stops, each supplied as an individual CSV file. Together, these tables describe every driver, car, circuit, and race in Formula 1 history.

Although the complete dataset spans from the beginning of the sport to the present, this project restricts analysis to the Hybrid era (2014–present). This filtering ensures comparability across seasons, as major regulatory changes prior to 2014, most notably the switch from refueling pit stops to non-refueling and the introduction of hybrid power units, fundamentally alter car behavior and race strategy. Early exploratory analysis confirmed that pre-2010 pit-stop patterns differed so dramatically from modern data that combining them would invalidate meaningful comparisons.

After filtering the tables to the Hybrid era and merging them into a unified structure, the final

dataset contains approximately 4,600 driver-race entries, with each row representing a single driver's performance in a single race. The integrated dataset includes roughly forty features capturing qualifying performance, race outcomes, constructor identity, circuit characteristics, pit-stop behavior, and lap-time statistics.

TECHNIQUES APPLIED

This project used data cleaning, preprocessing, table integration, feature engineering, EDA and simple modeling (multiple linear regression) to quantify the importance of various factors that influence a race. The Formula 1 dataset is divided across multiple tables, and so the first major step was constructing a unified dataset at the granularity of one row per driver per race. To do this I had to standardize column names across the tables, resolve inconsistent identifiers, and convert placeholders into proper missing values.

Initially after completing the process above I began EDA, and quickly realized that the failure to filter out data by rulesets led to unuseful results. I determined that the best way to proceed would be to remove the years where the rules would have a major impact on the factors I was measuring. That specifically meant removing all years that allowed for refueling during pit stops, dramatically increasing stop time. After removing all data before 2010 (the year refueling ended) I also realized that the power units changed in 2014 from full combustion to hybrid. I decided to focus only on the hybrid era (post 2014) so that I was only focusing on a single power unit type. Once these changes were made the data was more useable.

Feature engineering was also essential in the transformation of the raw tables into meaningful variables. The pit-stop table was aggregated to compute the total pit-stop time and the number of pit stops for each driver-race entry. Lap-time data was summarized into best lap time and average lap time, giving the essential two performance indicators. Qualifying data original

contained multiple sessions, but was simplified to a single qualifying position. These engineered features were more in line with the race factors being considered than the raw data that they were created from.

EDA was used to understand the structure of the dataset and to help identify which variables might influence the race results. This included summary statistics, null-value inspection, and distribution visualizations. Correlation analysis was used to assess linear relationships between variables, for example showing that qualifying had a strong positive correlation with the finishing result. Group level comparisons, such as finishing position for constructor or driver, revealed stable performance patterns.

Finally, a simple supervised learning model (multiple linear regression) was used to quantify the relative impact of qualifying, constructor, and pit stop metric on finishing position. The modeling step confirmed the EDA findings showing that qualifying is the strongest indicator in determining finishing position, followed by constructor and finally pit stops.

KEY RESULTS

This project was useful in revealing demystifying race results. Its findings help determine how qualifying performance, constructor strength, and other factors affect the finishing position of the driver. In the Hybrid era it is clear via exploratory, statistical, and modeling techniques, that qualifying position consistently is the strongest determinant of finishing position. Correlation analysis showed a strong linear relationship between qualifying position and finishing position, far exceeding that of pit stop metrics and lap-time aggregates. Scatterplots reinforced this pattern, showing a driver who starts at the front is likely to stay there for the entirety of the race with rare deviation. This behavior reflects the aerodynamic limitations of modern Formula 1, where dirty air affects cars' ability to follow, preventing passing. This makes it so track position confers a massive competitive advantage, allowing a driver to

maintain clear air and their own pace, avoiding mid-pack turbulence and racing incidents.

The car constructor also plays a meaningful role. Grouped averages reveal a consistent hierarchy of constructors, in which Mercedes, Ferrari, and Red Bull have a clear advantage. These trends reflect long-term differences in engineering resources, testing, and car development cycles. However, the constructor influence manifests as a baseline, such as Mercedes giving its drivers a 4 place advantage. This is distinct from a race-level determining factor, in which it was demonstrated that qualifying position is more significant. This is because a person who qualifies first is likely to finish first, while a Mercedes starting in 20th is likely to finish in 16th. It conveys an advantage, but it's lesser in the actual race. When included in a regression model alongside qualifying, the constructors' coefficients were smaller, which indicates while stronger teams do increase performance, they are still secondary to qualifying.

Pit stop metrics and lap time summaries had a minimal influence on finishing position. Both exhibited a correlation near 0. This makes sense, Formula 1 has optimized pit stop operations, usually taking 2 seconds in the pit itself and 25 seconds in the pit lane total. While lap time metrics had a higher correlation, the fact that so much of modern racing depends on tire strategy, where drivers are aiming for a specific pace to preserve the tires, means lap times are less important than strategy and track position.

A multiple linear regression model brought quantitative clarity to these relationships. The model achieved an R^2 of approximately 0.47 on both training and test sets, indicating that nearly half of all outcome variability can be explained by qualifying position, constructor identity, and pit-stop features. Within this model, the coefficient for qualifying position was by far the largest, confirming its dominant predictive importance. Constructor coefficients introduced predictable offsets for historically strong teams, but their magnitudes remained significantly

smaller than that of qualifying performance. Pit-stop and lap-time metrics contributed almost no predictive value. Taken together, these results demonstrate that in the Hybrid era of Formula 1, finishing position is driven primarily by track position at the start of the race, with constructor strength providing a secondary but less influential contribution and pit-stop dynamics playing only a minor role. This conclusion supports the broader understanding that modern F1 outcomes are shaped mainly by qualifying performance and the aerodynamic difficulty of overtaking, rather than by race-event factors such as pit strategy or marginal differences in race pace.

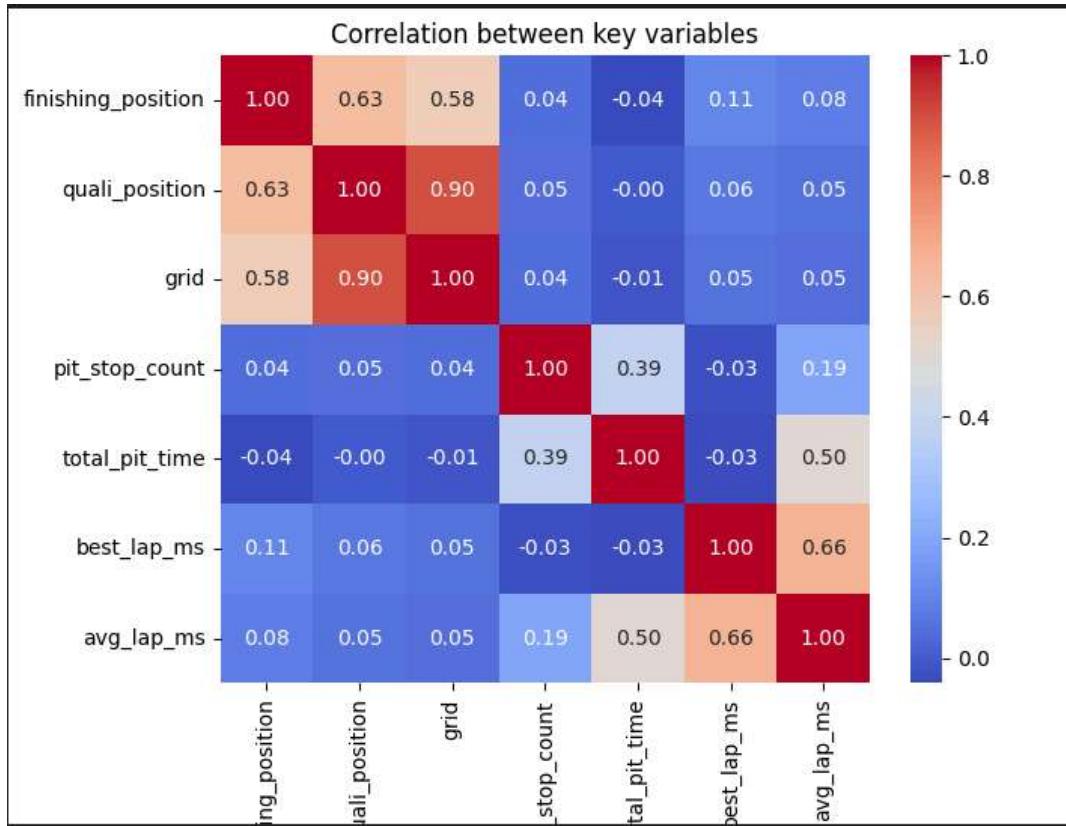
Applications

The finds of this project help understand the competitive dynamics in modern Formula 1. The most significant insight is that qualifying performance overwhelmingly determines race outcomes in the Hybrid era. This has direct applications for both analysts in media when explaining the race weekend and the strategies on display, but also the team engineers planning out race strategy. Clearly, a driver's starting position needs to be prioritized every weekend, even over things like tire strategy.

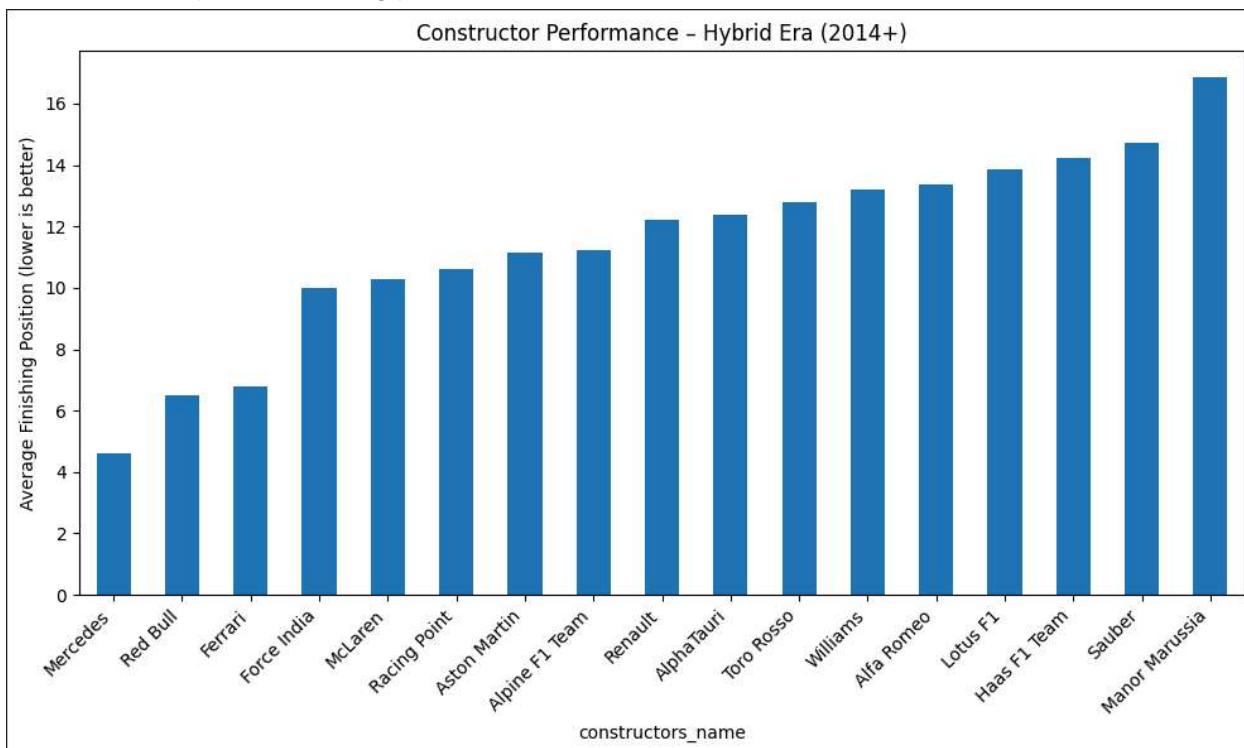
The secondary role the constructor plays was also made clear, especially in regards to teams long-term performance. While constructors like Mercedes and Ferrari can reliably elevate their drivers' expected finishing position, the influence isn't transformative. The 2025 season has been defined by poorer performing cars qualifying well, and thus remaining competitive in the points throughout the season, this analysis explains why.

VISUALIZATIONS

Heatmap showing various the relationship of various attributes with finishing position:



Constructors' Impact on finishing position



Qualifying and Finish Position Scatterplot:

