Final Project Proposal

Duncan Craine

STA334- Statistical Consulting

Professor Millie White

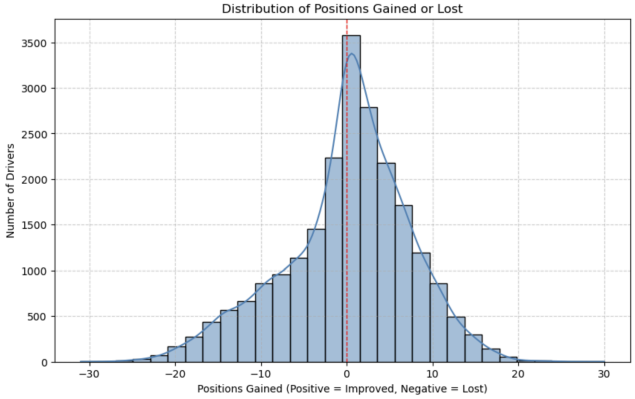
October 6, 2025

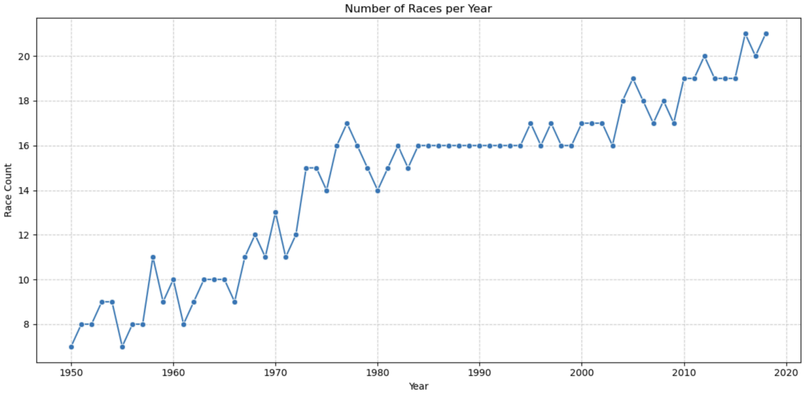
Introduction

Formula 1 (F1) has long been seen as the peak of motorsport and one of the most exciting sporting events worldwide. Each year, the F1 calendar features races held on city streets and famous circuits across the globe. F1 draws hundreds of thousands of fans in person and millions of viewers on TV from all around the world for each race weekend. Teams push technological limits, and drivers constantly risk their lives to get every bit of speed from the car and track. Why? To win millions of dollars, hear the crowd roar, and gain fame through total dominance. Even though there are rules that promote safety and close racing, every week teams introduce new car upgrades, and new tracks pose fresh challenges, making each race unpredictable. While technological innovation and team strategies attract the most devoted fans, most people are drawn to the drama that comes from uncertainty: unexpected driver performances, sudden mechanical failures, multi-million-dollar crashes, and rapidly changing weather. The unpredictability of F1 is a core part of its global appeal, impacting fan engagement, sponsorship value, and the competitive balance of the sport.

Methodology

In a data-driven sport, extensive data has been collected into 13 spreadsheets containing relational datasets with a total of 94 columns, covering the years from 1950 to 2017. Due to the mess, a lot of cleaning, preparation, and exploration are needed before analyzing the large amount of data. The data is very messy, but using a database with querying capabilities is the best way to retrieve the desired information. For each question, the answer may be hidden across multiple datasets, so querying allows subsets of data to be pulled together and made easier to work with. Many columns, such as lap times and dates, need to be converted into readable time formats rather than scraped strings to make them usable. Finally, with cleaned and prepared data, we can approach the questions. By using statistical tests, boxplots, and machine learning models, we can see how surprising the results are. Running basic models will likely show that wherever a driver starts, they are more than likely to finish there. This is reflected in the graph to the right, which shows it as the most common outcome. However, with more than 20 drivers on the grid and the ability to overtake over several hours, drivers shuffle positions through complex strategies, teamwork, and skill. Overtaking involves the risk of crashing, as these cars are moving at incredibly high speeds and the margin for error is very small. As the popularity has increased over the years, the demand for more races in more locations has increased. F1 now has 24 races in a year, and the growth is visible in the graph to the right. However, these new tracks introduce new challenges. Some tracks are narrower, making overtaking more difficult and reducing the margin for error, which leads to more crashes, red flags, safety cars, and many other race-changing scenarios. Some races are inevitably more or less exciting, so examining the change in position from the previous race, season standings, and qualifying position to the current race will provide insights into

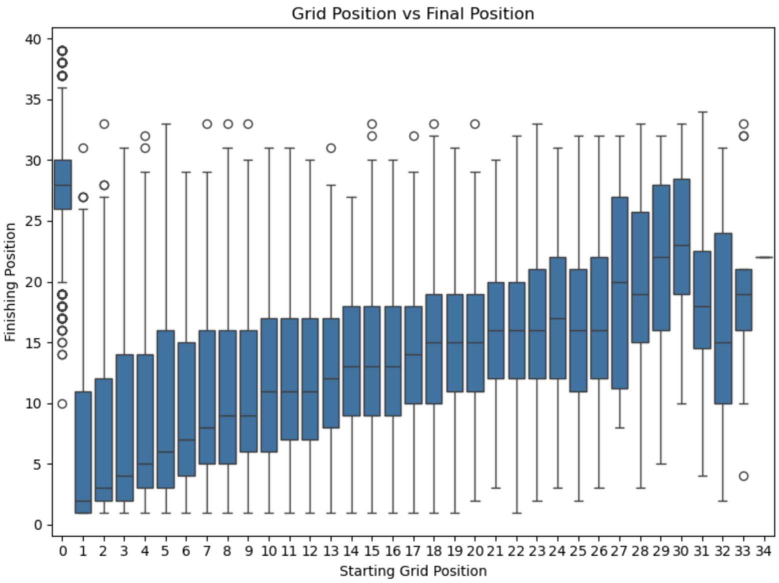




unpredictability. This analysis can be performed by sorting the data based on many potentially influencing factors.

Project Plan

This project aims to explore the question: What makes Formula 1 unpredictable and exciting? Using historical data from 1950 to 2017, the goal is to measure various sources of unpredictability, including inter-team competition, the impact of DNFs (Did Not Finish), how often positions are gained or lost from each driver's starting position, and the effect of circuits on race results. The figure to the right displays an incredible spread of finishes from each position, leaving fans on the edge of their seats, but still showing the incentive to qualify well each Saturday before the race. The analysis will identify when and where unpredictability has been highest, as well as whether the sport has become more or less predictable over time.



References

Formula 1 Race Data. (n.d.). Kaggle. Retrieved October 6, 2025, from https://www.kaggle.com/datasets/cjgdev/formula-1-race-data-19502017