

Rody Bertolini  
CSPB 4502 - Data Mining  
Project Proposal  
9/19/25

### **Project Proposal**

Project Title: "Determining How Various Factors Impact the Results of an F1 Race"

Description: Using historical Formula 1 race data I will quantify how much various factors (the driver, the car, the team, track characteristics, pit stops, race strategy, and weather) affect the outcome of a race. I will compare these aspects directly to determine which have the greatest influence in a race by measuring their marginal impact on finishing position and podium probability.

Prior Work:

My project is based off of prior data collection uploaded to Kaggle.

Data Sets: F1 Championship 1950 -2024

URL:

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

Created by user Vopani

Drivers (driver metadata)

Driver Standings (Ranking for each season)

Races (Date)

Circuits (Track)

Results (Race Results)

Qualifying (Qualifying Results)

Lap Times (Time for each lap in a race)

Proposed Work

1. Data cleaning
  - a. Handle missing pit stop times, convert lap times to second
2. Transform data with normalization
  - a. Normalize lap times, sort tracks into discrete attributes of short, medium, long
3. Data Integration
  - a. Merge multiple CSV files into a single table for modelling.
4. Describe patterns such as podium finishes vs. non podium finishes, driver ranking, etc.
5. Measure similarity.
6. Analysis and evaluate results.

Tools

1. Python
  - a. Pandas
  - b. Numpy

Evaluation:

- 1.Understand patterns via classification, regression and ranking.
2. Summarize and compare data distributions.