DSCI 510 Final Project Progress Report

Project: Formula One - Race Performance Analysis

Project Scope Update

The project scope remains unchanged from the original proposal. I am analyzing Formula 1 race performance data from the 2023-2024 seasons to identify key factors influencing race outcomes, focusing on qualifying positions, race finishes, pit stops, and lap times. Minor adjustment: prioritized OpenF1 and Ergast APIs for initial collection due to better documentation; F1 API Dev will be integrated later as supplementary source.

Data Sources

OpenF1 API (https://api.openf1.org/v1/)
- Endpoints: /sessions, /laps, /drivers, /pit.

Successfully collected 44 races,
~35,000 lap records,
~450 pit stops,
driver metadata.

Key fields: driver_number, lap_duration, pit_duration, session_key, meeting_key, circuit names, dates. Format: JSON converted to pandas DataFrames.

Ergast API (http://ergast.com/api/f1/)
- Endpoints: /[year]/results.json,
/[year]/qualifying.json,
/[year]/driverStandings.json.

Collected 880 race results, qualifying data (Q1/Q2/Q3 times), championship standings for 2023-2024.

Key fields: position, grid, points, status, fastest_lap, driver/constructor info. Format: JSON with nested structure, parsed and flattened.

Issues / Difficulties

Encountered: (1) API rate limiting on large requests - solved with retry logic and request delays. (2) Driver identifier inconsistencies between APIs - created mapping dictionaries for standardization. (3) Missing lap data from DNF/DSQ - added status filtering. (4) Different date formats - using pandas to_datetime() for normalization.

Anticipated: (1) Complex data merging between lap-level and race-level data may create large datasets. (2) Storage concerns for full lap data - may need aggregation. (3) Analysis scope prioritization with limited time remaining.

Next Steps: Complete merging pipeline, begin exploratory analysis, create visualizations, implement validation tests.