Visual Analysis of Formula 1 Racing Strategy and Performance

Project Overview

This F1 data science project conducts a comprehensive analysis of the 2023 Formula 1 season using simulated data that mirrors the real structure of the Ergast API. The goal is to extract performance trends, compare driver and team strategies, and build a pipeline that not only visualizes race insights but also predicts top performers using machine learning. By analyzing structured race data—such as lap times, qualifying results, pit stop details, and driver standings—we aim to simulate how F1 teams and ...

Using a well-defined pipeline, the project begins with data loading and preprocessing, followed by exploratory visual analysis and feature engineering. Key performance indicators such as qualifying consistency, lap time variability, and pit stop efficiency are extracted for each driver. These metrics are then used to build an interpretable model that attempts to predict which drivers are most likely to finish in the top three of the season standings.

Data Preprocessing & Exploration

The datasets used in this project include `race_schedule.csv`, `driver_standings_2023.csv`, `qualifying_results_2023.csv`, `lap_times_2023.csv`, and `pit_stops_2023.csv`. Each file simulates data from the respective domain of Formula 1 racing. The preprocessing stage focuses on cleaning and formatting each dataset for integration. Lap times and qualifying sessions are converted from string time format into floating-point seconds for easier comparison. Driver and team names are standardized across files...

Exploration begins with analyzing the driver standings data, which provides a summary of each driver's championship points, constructor, and final ranking. Visualizations such as horizontal bar plots were used to represent this data, allowing a quick view of dominance (e.g., Red Bull or Ferrari leading the table) or close midfield battles. Pit stop data was aggregated by driver to understand the frequency and average duration of pit stops—critical factors that can make or break race outcomes.

Next, lap-by-lap data was used to analyze consistency and position changes throughout races. This offers insights into whether a driver starts strong but fades, or gains ground lap by lap. Fastest lap times were extracted to identify peak performance windows, track affinity, or exceptional tire management. Additionally, qualifying results were analyzed across 23 rounds to observe how drivers evolve through the season, revealing those with improving form or decline in pace.

Predictive Modeling

A classification model was built using a Random Forest Classifier to predict whether a driver finishes in the top 3 of the season. To make this prediction meaningful, several features were engineered:

- Average Q3 Time: Collected from qualifying results, this reflects single-lap performance potential.
- Lap Variability: Standard deviation of lap times indicates how consistent a driver is across a race.
- Average Pit Duration: Represents how efficient the pit crew is and how long a driver is off track.
- Total Points: Championship points as a feature gives strong indication of overall performance.

The features were joined into a unified dataset per driver and split into training and test sets. The model showed good separation power between top-3 and non-top-3 finishers, confirming the hypothesis that qualifying speed and consistency strongly correlate with championship outcomes.

Key Visual Insights

• Driver Standings Bar Chart: Shows overall points per driver, grouped by constructor. This visually highlights team hierarchies and individual contributions.

• Pit Stop Frequency and Duration: Charts show which drivers have more frequent or longer pit stops. Combined with performance data, this reveals strategic trends such as undercutting or tire saving.

• Lap-by-Lap Position Evolution: Line plots help visualize whether drivers climb or fall back during races, signaling overtaking ability, tire management, and racecraft.

• Fastest Lap Time Trends: Comparing fastest laps per driver across multiple races reveals performance surges, qualifying pace carry-over, and circuit-specific performance.

• Qualifying Time Trends (Q3): Drivers' Q3 times were tracked across 23 rounds. This highlighted form slumps or improvement across the season, such as mid-season upgrades or confidence drops.

• Qualifying vs Championship Finish Correlation: By merging qualifying position with final standings, a scatter plot showed a moderately strong relationship between starting position and season finish.

• 3D Circuit Clustering: Using simulated circuit-level data (average speed, laps, pit stops), a KMeans model grouped circuits into three categories. The interactive 3D scatter plot helps visualize whether a circuit is more about speed (e.g., Monza) or complexity (e.g., Monaco).

Limitations & Future Scope

Since the project is based on simulated data, it lacks real-world nuances like tire compounds, weather changes, mechanical failures, or in-race penalties. These variables significantly affect race results and could provide a richer modeling dataset if integrated. In future iterations, real Ergast API data could be pulled to generate live analysis dashboards or telemetry-based performance prediction.

Other valuable extensions include:

- Predicting per-race results or podiums using time-series models like LSTMs
- Estimating pit stop strategies dynamically based on lap delta trends
- Creating race replays using animated visualizations from lap time data
- Applying PCA to reduce dimensionality and visualize team-level performance clusters
- Using Shapley values or feature importance heatmaps to interpret ML models

Conclusion

This project successfully simulates a professional-grade Formula 1 data analysis pipeline. It not only recreates key aspects of motorsport analytics—such as driver form, pit strategy efficiency, and circuit characterization—but also demonstrates how those metrics can be modeled for future prediction. The classification model, while simple, offers a baseline for how lap and qualifying performance can be translated into strategic projections.

With enhanced data sources, more advanced modeling, and real-time integration, this project could serve as a foundation for broadcast analytics, team strategy platforms, or fan engagement tools. Whether you are a data scientist or an F1 enthusiast, the structured insights provided by this notebook can help decode the complexities of race weekends.

All code, visualizations, and documentation are hosted in the GitHub repository:
https://github.com/Santhakumarramesh/f1-advanced-analysis

All code, visualizations, and documentation are hosted in the GitHub repository:

https://github.com/Santhakumarramesh/f1-insights-predictive-analysis