Visualization of Complex Data

DATS 6401

Final Term Project Proposal

Aakash Singh Sivaram

Project Title: Analyzing the Evolution of Formula 1 World Championship (1950-2024)

Introduction:

The Formula 1 World Championship dataset, spanning from 1950 to 2024, offers a comprehensive overview of the sport's rich history. This dataset is particularly intriguing due to its extensive coverage of races, drivers, constructors, circuits, and more over seven decades. Visualizing this data can uncover trends, performance patterns, and the evolution of the sport, providing valuable insights into how Formula 1 has transformed over time. The combination of numerical and categorical variables facilitates a multifaceted analysis, enabling the exploration of complex relationships within the data.

Dataset Overview:

The selected dataset is publicly available on Kaggle and comprises over 73,000 observations distributed over multiple tables.

List of Features:

- **Numerical Features (4+ features):**
    - **Race Duration:** Time taken to complete each race.
    - **Lap Times:** Recorded time for individual laps.
    - **Pit Stop Counts:** Number of pit stops made during a race.
    - **Driver Standings Points:** Points accumulated by drivers over a season.
- **Categorical Features (4+ features):**
    - **Driver Names:** Identifiers for each driver.
    - **Constructor Teams:** Names of the teams participating.
    - **Circuit Locations:** Venues where races are held.
    - **Race Seasons:** Specific years or seasons of the championships.

Feature Engineering:

To enhance the analysis, feature engineering will be employed to create additional variables such as:

- **Average Speed per Race:** Calculated by dividing the total distance by race duration.
- **Overtake Counts:** Estimating the number of position changes during a race.
- **Circuit Classification:** Categorizing circuits based on attributes like length, type (street vs. track), or historical significance.
- **Driver Experience Level:** Determined by the number of races participated in prior seasons.

Static Plots for Numerical Features:

- **Histograms:** To visualize the distribution of lap times across different races.
- **Box Plots:** To identify outliers and understand the spread in race durations.
- **Line Charts:** To track the progression of driver standings points throughout a season.
- **Scatter Plots:** To examine relationships between variables, such as pit stop counts versus race duration.

Static Plots for Categorical Features:

- **Bar Charts:** To display the number of wins per constructor or driver.
- **Pie Charts:** To show the proportion of races held at various circuit locations.
- **Heatmaps:** To illustrate the frequency of driver-constructor pairings over the years.
- **Stacked Bar Charts:** To compare race outcomes across different seasons or circuits.

Interactive Dashboard Draft:

The interactive dashboard will be designed using MS PowerPoint, featuring multiple tabs to provide a comprehensive analysis experience:

- **Data Cleaning:** Overview of methods such as handling missing values and ensuring data consistency.
- **Outlier Detection & Removal:** Visual tools and statistical summaries to identify and address anomalies.
- **Dimensionality Reduction (PCA):** Demonstrating how Principal Component Analysis can reduce data complexity while preserving significant variance.
- **Normality Tests:** Including tests like Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling to assess data distribution.
- **Data Transformation:** Applying techniques such as normalization and standardization to prepare data for analysis.
- **Data Loading & Visualization:** Interactive plots for numerical and categorical features, along with statistical summaries.

This approach aims to create a comprehensive and engaging platform for exploring the evolution of Formula 1, offering valuable insights into the sport's history and development.