Formula 1 Race Dashboard

Visual Analytics Project
A.Y. 2023/2024

Federico Perotto, 1701738

Introduction

Formula 1 is widely considered the pinnacle of motor sports and is currently at the peak of its popularity. The incredibly advanced technology employed by the teams results in huge amounts of data.

Unfortunately, most of this data is kept as a closely guarded secret by the teams, which means that the general public as well as journalists, pundits and content creators have to rely on limited or biased information coming from the teams or from the broadcast

Project's goal

This project has the goal of providing a useful and comprehensive tool to view and analyze the evolution and progression of F1 races and compare drivers' performances, allowing its users to create new opinions, check their existing assumptions against real data or also use the dashboard itself to present and accompany the content pieces.

NEXT

Data Sources

- FastF1 Python Package

  The FastF1 python package is utilized to access the results, timing data, telemetry and weather data for all the races of the 2023 season. The data is in the form of pandas DataFrames, which then get saved as CSV files.

- Pit Stop Data

  The FastF1 python package includes some information about pit stops, but not the precise duration. The information is available on the web site of Formula 1 official partner DHL. It provides precise measurement of pit stops duration for every race in the form of tables, which have been manually converted to CSV files.

Data Structure

- **Laps Files**: Every race has its own laps.csv file. Each row corresponds to one lap completed during the race, and each file has between 800 and 1500 lines. Of the 31 columns contained in each file, we use the following: `Time, Driver, DriverNumber, LapTime, LapNumber, Stint, Sector1Time, Sector2Time, Sector3Time, SpeedI1, SpeedI2, SpeedFL, SpeedST, Compound, TyreLife, Team, LapStartDate, TrackStatus, Position.`

- **Telemetry Files**: Each driver has his own file for every race, which contains information about the car telemetry. The data is sampled at a rate of around 4Hz, so every file contains between 30000 and 40000 lines. The columns of data that we use are: `Speed, nGear, Throttle, Brake, Time`

- **Results Files**: Every race has a results file. Each row corresponds to a single driver and contains information about a driver starting and finishing position, as well as some meta-data. The data that we use is: `DriverNumber, Abbreviation, TeamName, LastName, Position, ClassifiedPosition, GridPosition`

- **Weather Data Files:** For each race the API provides data about the weather conditions. This data is used as extra information in the dimensionality reduction process, which is described in the preprocessing section. The information used from this files are: `Time, AirTemp, TrackTemp, WindDirection, WindSpeed`

NEXT →

Pre-processing (1 / 2)

To visualize the result of the dimensionality reduction process, we generate dedicated CSV files for every race. To generate these CSV files some data aggregation, preprocessing, and clean up is performed.

- For each lap contained in the laps.csv file, we fetch the telemetry and weather data corresponding to that lap and that driver.
- From the data retrieved for every lap we compute the following metrics:
  - Average speed over the lap
  - Max speed
  - Min speed
  - Average throttle level
  - Percentage of the lap spent at 100% throttle
  - Percentage of the lap spent braking
  - Air temperature
  - Track temperature
  - Wind direction
  - Wind speed
- These metrics are then appended to the original laps data.

Pre-processing (2 / 2)

- Before performing the dimensionality reduction, the data is cleaned up. First we remove rows that would be obvious outliers (i.e., laps in which a driver did a pit stop), as well as rows with missing data. Columns with no relevance to the evaluation of a driver's performance (such as session times or DriverNumber) as well as categorical attributes are also removed (such as Deleted and IsPersonalBest).
- The remaining data is scaled using the StandardScaler provided by the sk-learn python library and then t-SNE is applied.
- Since one of the goals for the project is to be able to evaluate and compare drivers' performances, it made sense to use t-SNE as the dimensionality reduction technique and look for clusters or outliers in the new space.
- The results of the t-SNE process are then appended back to the data that generated them and saved to a CSV file.

Visualizations

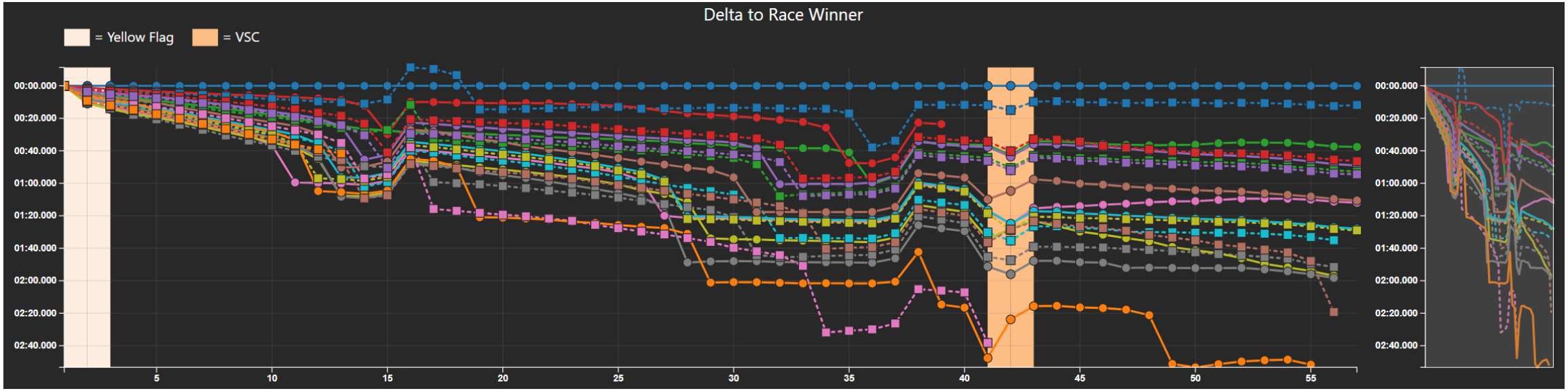
NEXT →

Drivers Legend

- This view serves as a global interactive legend. It displays all the drivers that participated in the selected race.
- This view also shows how each driver is encoded in the other visualizations. The color represents the team the driver drives for, and each driver is represented either with a continuous line and a circle, or with a dashed line and a square.

NEXT

Linechart + Context (1 / 2)



- This view shows the time delta between each driver and the race leader during the course of a race. Each line corresponds to a driver, and the time difference is computed at the start of every lap. Every lap of every driver is highlighted by either a circle or a square, according to their encoding shown in the drivers legend.

NEXT →

Linechart + Context (2 / 2)

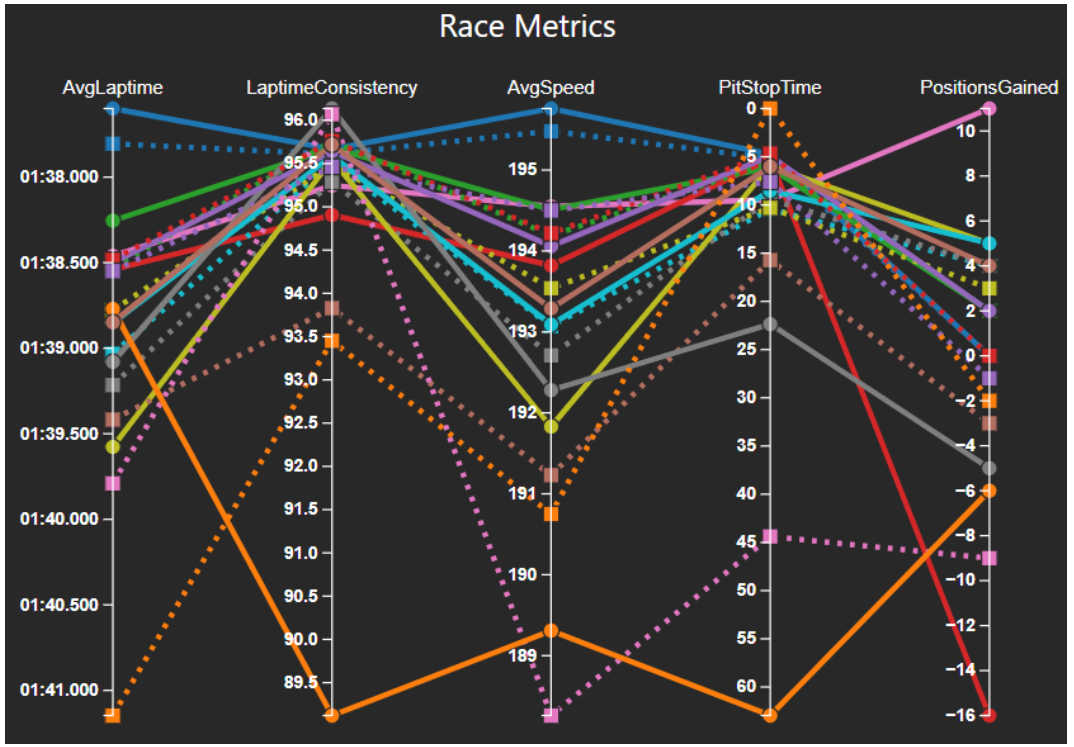
Insights

- This chart provides insight into the race evolution, the consistency of performances and the impact of strategies. The slope of the lines indicates whether a driver is gaining or losing time relative to the leader, but also relative to the other drivers, which allows the user to compare their performances, and the lines also represent the current position of the drivers, and their relative distance is proportional to their distance on track which adds context to the strategy chosen by teams and drivers.
- The colored rectangles shown in the background of the chart represent the status of the track and have a corresponding legend. These rectangles provide even further insight into the strategies chosen by drivers, as well explaining some of the sudden jumps in line trajectories.
- Next to the line chart, a context overview is provided. Through an interactable vertical brush, this section provides a simplified view of the line chart and what part of it is currently visible.

NEXT →

Parallel Coordinates Chart

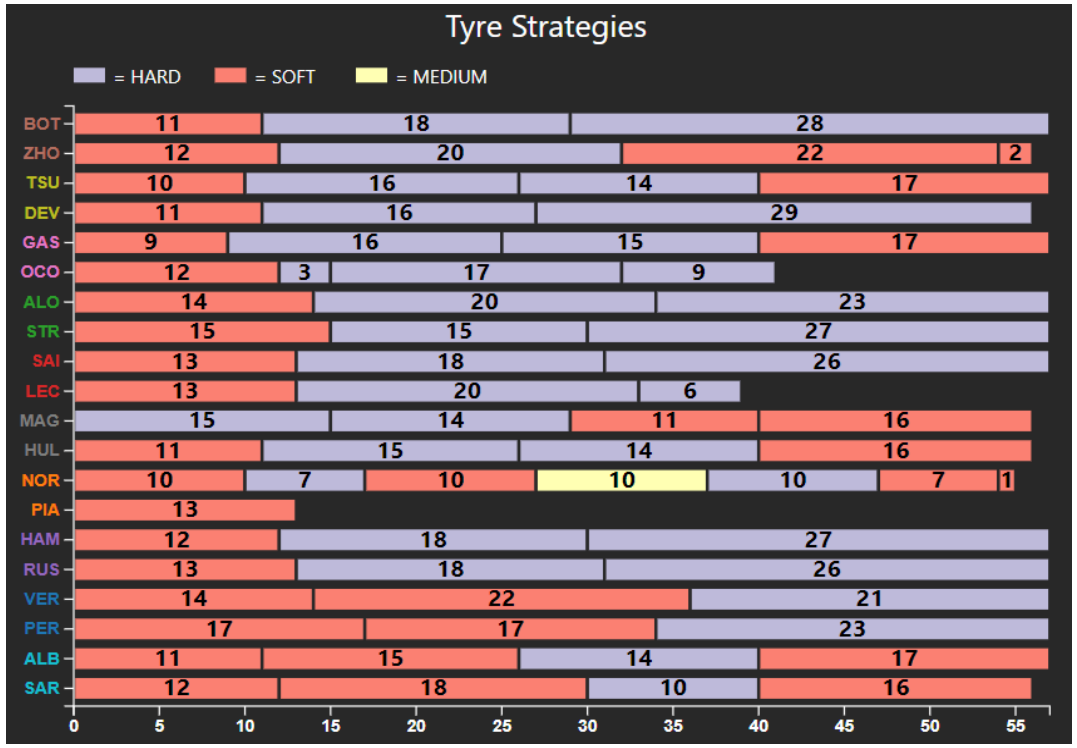
- The parallel coordinates plot allows to analyze and compare metrics relevant to the race. Each driver is represented by a line that intersects each axis at a point corresponding to their data value for that metric.
- The drivers are differentiated by using different colors, line styles and shapes, as shown in the drivers legend. The points of intersection are highlighted by either a circle or a square.
- The axis have been oriented so that ’good’ values for each metric are on the top side.
- This chart can be useful to identify or highlight patterns, clusters or outliers, which reflect in the evaluation of the drivers’ performances.



NEXT

Stacked Bar Chart (1 / 2)

- The stacked bar chart is used to visualize the tyre strategies chosen by the drivers and teams during the course of the race
- Each bar corresponds to the strategy of one driver, and its length corresponds to the number of laps the driver completed. Each bar is divided in multiple segments (stints) that represent the tyre compounds used by a driver over the course of the race.
- Each bar is divided in multiple segments (stints) that represent the tyre compounds used by a driver over the course of the race. The color of the segments indicates the compound of the tyre, and the length of the segments corresponds to the number of laps that the driver spent on that compound (stint length). The stint length is also written inside of the bar segment, for easier comparisons between stints.



NEXT

Stacked Bar Chart (2 / 2)

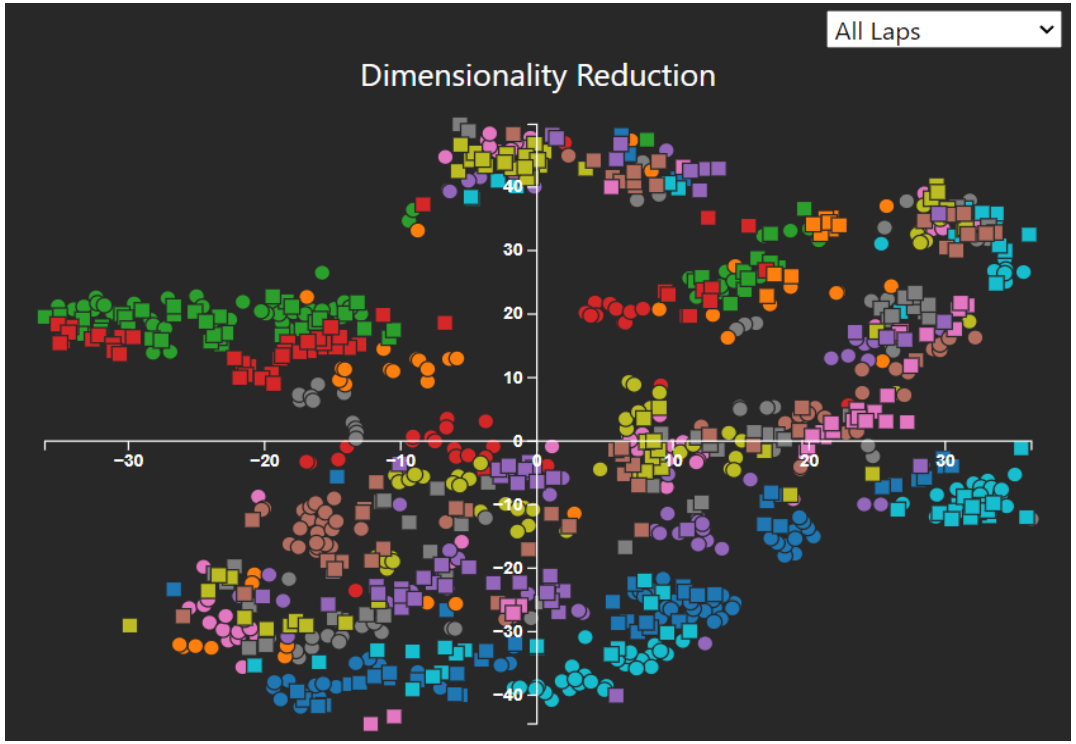
Insights

- This visualization allows the user to compare the different tyre strategies and to identify common choices or unique approaches, while also granting insight on the ability of drivers to manage or utilize the different compounds.
- By inspecting the various stints, the users understand and contextualize how the strategy choices impacted the drivers and their performances.
- By interacting with this chart, the analysis can be refined and enhanced.

NEXT →

Scatterplot

- The scatter plot is used to visualize the result of the t-SNE dimensionality reduction technique. The method used to obtain the data for this visualization is described in the preprocessing section.
- The goal for this chart is to allow the user to identify patterns or clusters in a two-dimensional representation of the data. Each element in the scatter plot represents a single lap from one of the drivers
- Although t-SNE can introduce both false positives and false negatives, the relative position of each element should correlate to a measure of similarity. Distant and isolated elements are outliers, and close elements are similar laps.



NEXT