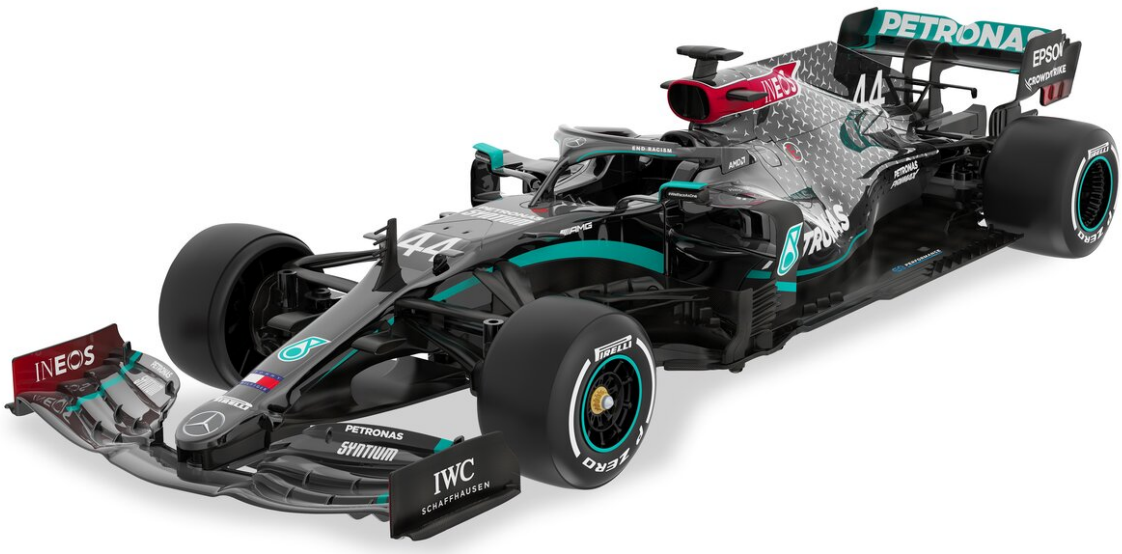
Data Science Project

Formula One



Team Members:

1- Tarek Yasser
2- Abeer
3- Mohamed Khaled
4- Kareem Taha

Problem description:

Formula One is a data-intensive sport where multiple teams compete together and all of them have their eyes on the win. The team that can infer more information from the data they have and utilize it in decision-making for the next races and seasons will have an edge over other teams. Even the slightest edge can have a great impact on the team's performance.

F1 teams use data from past events to improve future outcomes and harness the power of data from their drivers, ensuring that they are well-equipped to respond to various scenarios.

By sourcing historical data and using it in complex data science algorithms, F1 can predict race strategy outcomes with increasing accuracy for teams, cars, and drivers.

In our project, we plan to mimic an F1 data science team and try to utilize the data we have to better understand how different aspects change the outcome of the race.

Dataset:

- We found a dataset that contain lots of information about each race from 1950 to 2023

Questions

- Optimal pit stop count per circuit per race per year
    - Group the data by year, race and circuit and find the optimal pit stop count (number of pit stops for the winning driver)
- Performance analysis: driver performance (average position in all races per season)
  Group the data by season and by driver and average his positions (does the data even have all positions along the race or the final position only?)
- Does age influence driver performance (average position in all races per season)?
    - Group the data by season and by driver and average his positions along the years (get average position per circuit? If 20 drivers it will always be 10)
- How do different rulesets affect performance? (per lap per 4-5 years)
    - Need to collect data about ruleset year range (This ruleset was in effect in the years xxxx-xxxx)
    - Find how updating the rule set affects the speed of cars in same tracks at least (top speed per track per year, if found same driver better)
- Spending vs Performance for teams/companies.
    - Need to collect spending data.
    - Perhaps for the top 5 teams
    - Perhaps for 5 years (2010-2015).
    - Budget per team per year
- Does a team being the manufacturer of their engine affect performance?
    - Need to collect data for a subset of years and teams.
    - Average performance per 5 years per team
- Home vs away's effect on performance. (Does having a race in your country affect performance?)
    - Find a relation between the nationality of the driver and his performance in races that take place in his homeland(ex:: Lewis Hamilton performance in Silverstone track)
    - Find a relation between the performance of the constructor and the race they compete in (ex: Ferrari’s performance in Monza)
- How qualification results affect race outcome?

    - Analyze if the position taken in the qualification affects the position at the end of the race.
- Which tracks favor overtaking?
    - Analyze which tracks are easier to overtake cars in, which will really help the drivers choose their pitstop strategies, thus they will have shorter stints but faster pace.
    - This will be decided by seeing the start position of each driver and the end position at the end of the race, the track will favor overtaking if the positions of the drivers changed a lot from their position at the first of the race
- Average retirement age?
    - Get the last race for each driver and then get the average age of each driver on his last race
    - Would consider taking only drivers who raced their last race when they were 29 years old or older, because if they had their last race before that, most probably they left Formula one for another sport and did not retire
    - This info would be really helpful for constructors who want to buy drivers, in order to know the average prime age of drivers
- How does a track's altitude affect top speed? / average lap time.
    - It's believed by F1 commentators that the higher the altitude of the circuit the less oxygen there is, so the engine would not burn enough fuel.
    - We will prove that by calculating the average speed of each circuit and see the correlation with altitude.
- Which tracks have the most DNFs?
    - DNF: Did Not Finish
    - See which tracks are brutal for the cars

Need two more inferential and predictive questions.
Inferential: "Seek to infer (generalize or extrapolate) information about a large population of data using a smaller sample of data (e.g., where analysis was performed). Thus, can be used to test a hypothesis generated from exploratory data analysis"
Predictive: Seek to find patterns and make predictions about the future. Can answer by using historical or current data

- [Inferential] Do European tracks have a similar DNF rate to Germany (or any other European country)?
- [Inferential] Given that Mercedes drivers have above average top speed, does this apply to all German teams?
- [Predictive] Will having a different/safer ruleset lower top/average speeds?
- [Predictive] Does a team spending more affect/increase their performance in terms of average driver position / top speed?

Proposed work plan for each question:

- Understand the dataset and know how to utilize it well and utilize rows that describe a specific race among multiple CSVs, the CSVs act like a database where each entity has a unique id so it's important to have time to explore the data

- Some data have to be scraped from the internet like the country of each circuit and the nationality of each driver.
- Some of the questions need the dataset to be cleansed in some sort like the average retirement age has to consider only drivers who had their last race when they are older than 28 years
- Visualize the data in order to understand it better and determine whether there are some aspects that we neglected in our questions
- Answer our questions, and see how the answers we got will help us in decision-making for next seasons and races
- Keep iterating the questions till we make sure that we have all the info we need to support all the decisions we took