Task 1
https://github.com/karloskarl/F1Analysis

Task 2 - Business Understanding

Formula One is a form of racing that, ever since its beginnings in the 1950s, has entertained millions and pushed the boundaries of automotive engineering. Its purpose-built circuits and strict set of rules (the formula) for manufacturing participating cars have made it one of the most captivating and competitive racing series in the world. It is this nature of the competition that has spawned a devoted community and fanbase in which debate over pretty much every aspect of the racing is endless. While quantifiable data exists and is easily accessible (as discussed below), most of this debate, particularly regarding the drivers and constructors, is largely based on emotion - skewed as a result of steadfast personal loyalties and opinions. Hence, looking at the data and allowing data science to give us an undistorted answer to the most burning of questions is an interesting prospect.

With this project, our main drive is to provide the Formula One community with sound and compelling data science solutions to some of the hottest debates. It is important to us that these solutions are easily digestible and well-visualized so anyone interested in Formula One racing can understand the results of this project.

Our resources consist of a sizeable dataset of over 80 years of Formula One results, lap and pit times, constructors, drivers etc. This dataset should cover all our needs. We will be working on this dataset using Python tools such as pandas and scikit-learn - the software will be running on our personal computers.

The main requirement of this project is the deadline of 11.12.2023 at noon. Our requirements include the aforementioned legible visualization of the results and, of course, the correct use of the tools to produce accurate results. The main risk for a project such as this would be hardware failure, which is incredibly rare and can be sidestepped by both team members thanks to secondary PCs.

Some terminology:
Grands Prix - the races, which take place on a circuit or closed public roads.
Circuit - purpose-built physical racetrack location
Season/Championship (with year) - A single year's worth of Grands Prix, which result in points that ultimately crown the season's winners (for both drivers and constructors)
Constructor - Corporate entity that designs and builds Formula One cars

Our data-mining goals, to start, are machine learning-produced equations for predicting race/season results and various visualizations of said equations in practice. To understand whether our equations were genuinely accurate, we will compare the machine learning-produced results with genuine results from the test portion of the dataset to make a qualitative assessment of accuracy. Criteria is met when both team members unanimously agree upon the results' accuracy and only then the results are presented.

Task 3 - Data Understanding

The dataset needs to contain detailed race-by-race results and statistics for as many races as possible, ideally with minimal gaps. The dataset for the races should include at least starting position, finishing position, points gained, and time behind the winner. There should be complete results for every year's (1950-2023) drivers and constructors championships, even if there are gaps in race-by-race data. Any further data such as lap times for every lap will only be bonuses for the project. There is no specific format we are looking for the data to be in.

This data is available on kaggle as a dataset (https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020) which has all of the qualities we are looking for. There are complete race results for every race since 1995 and there is even more data available here than we were originally looking for, so we can use those to gain extra insights into the dataset. We can use the results.csv for race results by driver for every single race, as well as driver_standings.csv for more aggregated drivers championship finishing position data. For any qualifying comparisons we wish to make we can use qualifying.csv. For any constructor results we can use either constructor_results.csv for race-by-race constructor results, but constructor_standings.csv contains standings information aggregated. The status.csv file contains a reference table for status codes within the results.csv statusId field, such as "Finished", "Disqualified", or "Retired". In the positionText field in both results.csv and consturctors_results.csv the text can contain either an integer (finishing position), "R" (retired), "D" (disqualified), "E" (excluded), "W" (withdrawn), "F" (failed to qualify) or "N" (not classified).

We discovered several quality issues with the dataset, but none of them are big enough to make us use another dataset. We found that all of the non-values were represented by the character "\N" which means that those all need to be replaced by NaN. We also found that there are significant gaps in qualifying.csv, with all data from before 1994 being unavailable. There are 84 races missing in the timeframe of 1995 to 2022 as well. There are some aggregate data tables, but they will not be sufficient for our uses, so we will be forced to make aggregate data frames with combined statistics from multiple races to make comparisons easier. There are also smaller errors in results.csv, especially from anything pre-1995. It looks like the data from before 1995 was corrupted in some way, with more null values in those races than compared to the period after 1995.

Task 4 - Project Plan

Outline for two main questions - predicting results for both drivers and constructors
Task 1: Create aggregated drivers and constructors season result dataframe to train on

Otherwise it will be a big pain to try and train on a random mix of properties from different tables. This will contain metrics such as average result, average qualifying, average distance behind teammate, as well as anything else we come up with that would be necessary.

Task 2: Simple analysis

Use non-machine learning methods to analyze the data and try and find answers to the questions asked.

Task 3: Feature Engineering

possible features: driver total experience, constructor performance in recent races, average lap times.

Task 4: Machine Learning Model Development

Experiment with different models, model parameters and other tweaks to find the best way to predict future results.

Task 5: Model Evaluation and Interpretation

Evaluate the model based on mean absolute error and create visualizations to compare different models to each other

If we decide to work on additional questions for this project, then we expect to apply a similar outline of steps. As for our tools and techniques - we believe that for such a simple project, we are equipped with all the necessary data and tools.