Name: Muley, Tushar

Assignment: Milestone 1 for Week 3 and 4

Date: January 10, 2021

1. 3 data sources, along with a description of each one (links to each are fine, no need to submit the actual data)
   1. For the CSV files types. I found Formula 1 (F1) data in Kaggle.com <https://www.kaggle.com/rohanrao/formula-1-world-championship-1950-2020?select=races.csv>
   2. For website I have a table of F1 drivers that I found on Wikipedia.com <https://en.wikipedia.org/wiki/List_of_Formula_One_drivers>
   3. For API I am planning on using Ergast API. They provide a wealth of information for non-commercial use related to Formula 1

<https://ergast.com/mrd/>

1. The relationships between them, or the relationship you will make between them.

The data from Kaggle.com and Ergast has a relation but the data from Wikipedia does not. I am planning to make a relationship between drivers, circuits (race courses), and race results along with a few other tables. I will have to build an id that will help me bridge the table between driver, circuits, contractors, race results and some other information like pit stops and pole positions. Some of the data does have unique keys I can use like the Kaggle csv files and API from Ergast.com. The circuit information I am getting from Wikipedia will need to be link to the data from Kaggle csv files and Ergast API.

1. What you believe you will have to do to the data to accomplish all 5 milestones and what your interpretation is of what the data means (you could provide a data dictionary or a summary of what the data is) – should be at least 250 words.

What I believe I will need to do to accomplish all 5 milestones for the term project. Starting with the csv files I believe I am going to have remove special characters that are showing up in the data set. Replace some of the header columns with better readable names. I want to include images of race circuits for reference in the visualization but not sure how I will provide that yet.

Identifying bad data will be a little difficult since track and driver information changes season over season. I think I might use driver and track information to determine if lap times or other information is out of range or out of the expectation.

I believe I will need to bring in other tables or call a few more APIs to allow me to completely identify the needed data. For example, status table which provides the status of what occurred during the race that might have caused a driver and car to not complete the race. This will additional information that is need for the analysis.

The main data dictionary:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| driverId | Driver identification |
| driverRef | Driver reference based on last name |
| number | Current race car number |
| code | First three letters of the drivers name |
| forename | Drivers first name |
| surname | Drivers last name |
| dob | Drivers data of birth |
| nationality | Drivers official nationality |
| url | Driver Wikipedia page |
| Circuit | Race course name |
| Map | Image of the race course |
| Type | Type of race course, Street, Road, Race Street circuit - Public street converted to race course. Example Long Beach or Monoco. Road circuit - Normal road course are part race track part street. Race circuit - Purpose built for racing. Circuit of the Americas |
| Direction | Direction the cars race in. |
| Location | City and Country the race circuit is located. |
| Last length used | Length of circuit |
| Grands Prix | Name of the circuit |
| Season(s) | Years the circuit was used |
| Grands Prix held | Number of races held |
| resultId | Id of the results table index. |
| raceId | Race identification. |
| driverId | Driver identification. |
| constructorId | Contructor indentification. |
| number |  |
| grid | Position on the race grid during start. |
| position | Position finished at the end of race. |
| positionText | Text version of the position. |
| positionOrder | Final position. |
| points | Points earned at the end of race |
| laps | Number of laps completed |
| time | N/A |
| milliseconds | N/A |
| fastestLap | Which lap was the fastest lap. |
| rank | Overall ranking of the fast lap in that race to other competitors. |
| fastestLapTime | Time of the fastest lap. |
| fastestLapSpeed | Speed of the car on the fastest lap. |
| statusId | Status id of how the car completed the race. |