**Formula 1 Results Analysis & Prediction**

**Team GottaGoFast:**   
Zachary Lawell, Lezeh Foy, Deepak Verma

MSCS 5610 – 701 Data Mining (Spring 2019)

Marquette University

# Abstract

In this project, a Formula One (F1) dataset containing information about drivers, constructors, and race statistics is analyzed using various data mining techniques. The scope of the project includes, data pre-processing, exploratory data analysis, feature selection and model generation. We start off by combining various tables containing F1 data to create two complete datasets on constructors and drivers and then move on to generating charts and plots to find interesting patterns within the dataset. The main goal of this project is to find correlations between the features in the dataset and determine the key features that can be used to predict the outcome of a race. The dataset being used for this analysis contains information on drivers, races and constructors over the span of 68 years ( 1950-2018). To limit model generation and run time, about 60% of the data is selected and used for exploration, analysis, and model training. Data selected to use for model training takes into account the need to have representation of all decades within the data so as to accurately explore changes in race patterns, car speeds and constructor standing over time. Data pre-processing, visualization, analysis and modeling is performed using python and its various data analysis libraries using Jupyter notebook as the interface for visualization. Overview of results will be added to the abstract as we progress through the project.

# Introduction and Background

Formula One is the highest class of single-seater auto racing sanctioned by the Fédération Internationale de l'Automobile (FIA). The FIA F1 World Championship has been a premier event around the world since its inaugural season in 1950. The word "formula" in the name refers to the set of very strict rules and regulations to which all participants' cars must conform.

Tucked inside each car is around 1.25 kilometers of wiring and more than 200 sensors, some of which give readings up to 1,000 times per second. Sensors around the track and highly precise race statistics from each event have also been gathered since inception. This makes F1 a treasure trove of big data. Data is gathered and analyzed in real-time using high performance computer clusters. The results of which are utilized by the teams to provide real-time advice to drivers during a race to gain an edge. The teams competing in F1 are at the forefront of technological advancement and therefore huge investments can drive developments which make a difference on the track. Average mid-field teams spend around $125m to $150m per annum, whereas top teams spend $350m to $410m per annum[3]. The time difference between these teams are less than 0.06 seconds per lap. At the end of each season, the top 10 teams share a winning of approximately $810m (as of 2018)[4].

Gathering data and quickly analyzing it to predict outcomes via the application of machine learning techniques can help the teams gain an edge where a single position advancement results in millions of dollars of winnings. Analysis and predictions of data from prior race data could allow race organizers, media, and fans to quickly narrow down the likely outcomes of races. However too many factors affect the outcomes just as with any sport.

In this project we’ll utilize the techniques studied in the Data Mining class against the dataset chosen. This will include pre-processing the data, performing initial visualization and observatory analysis. We then intend to apply some unsupervised and supervised techniques to model the data and predict outcomes. Some of the initial goals include: Determining the fewest number of the most important attributes for modeling the data, utilizing unsupervised machine learning to discover any interesting patterns, utilizing supervised machine learning to predict the likelihood of a race happening or the winner of a race.

# Dataset

For this project we are exploring the dataset from Formula 1 racing available from http://ergast.com/mrd/. The Ergast Developer API is an experimental web service which provides a historical record of motor racing data for non-commercial purposes. The API provides data for the Formula One series, from the beginning of the world championships in 1950. This dataset contains data from 1950 all the way through the 2018 season, and consists of tables describing race results, constructors results, constructors, race drivers, lap times, pit stops, qualification results, and many more as per the scheme provided at <http://ergast.com/schemas/f1db_schema.txt>.

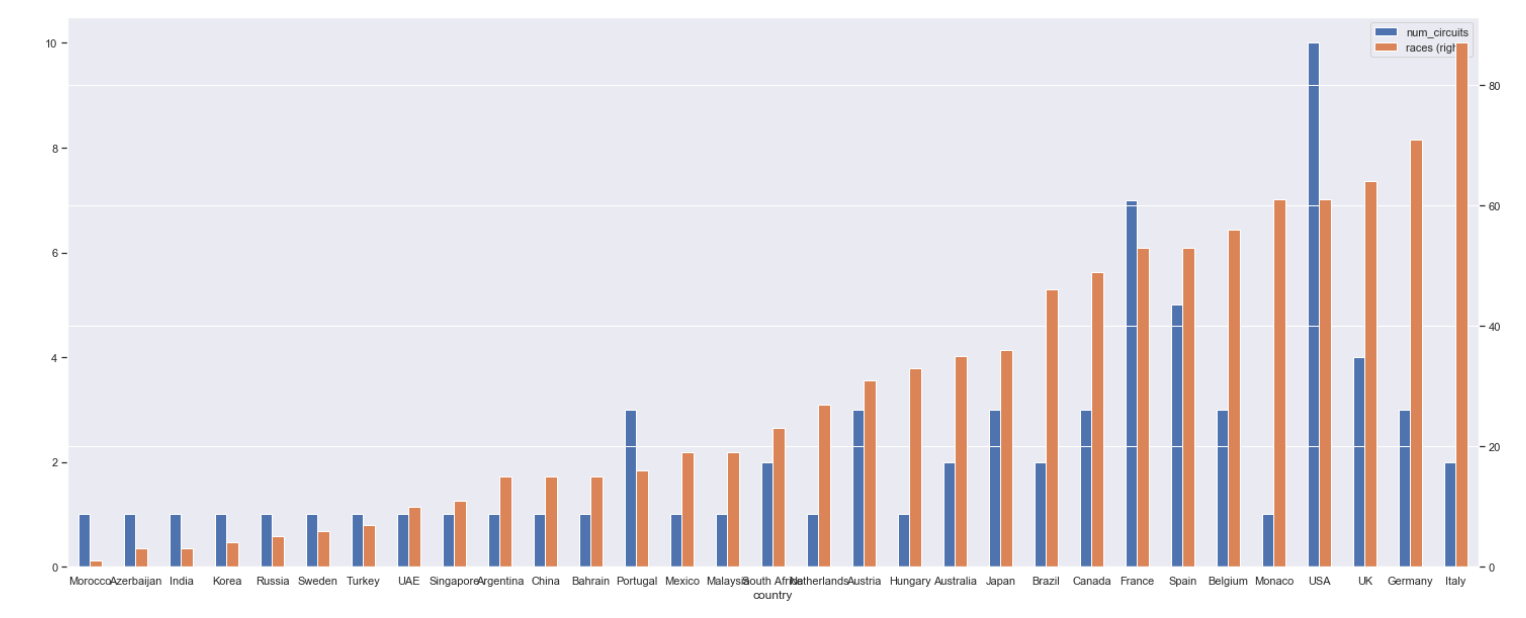
# Data Preprocessing

As with any data mining project, one of the first steps in analyzing this dataset was creating data frames and structures that will make the exploratory and analysis process easier. As indicated in the dataset schema, the information in this dataset was separated into multiple tables. Using the Constructor IDs and the Driver IDs, two main data frames are created by joining the tables above. The Constructor data frame contains information specific to each Constructor (vehicle manufacturer) while the Results date frame joins the driver performance information distributed across the tables. Parameters that aren’t of interest such as the url of each constructors wikipedia page or variables that were being duplicated but named something else are removed from the final tables as well. After preprocessing we end up with 2 data frames with the following parameters;

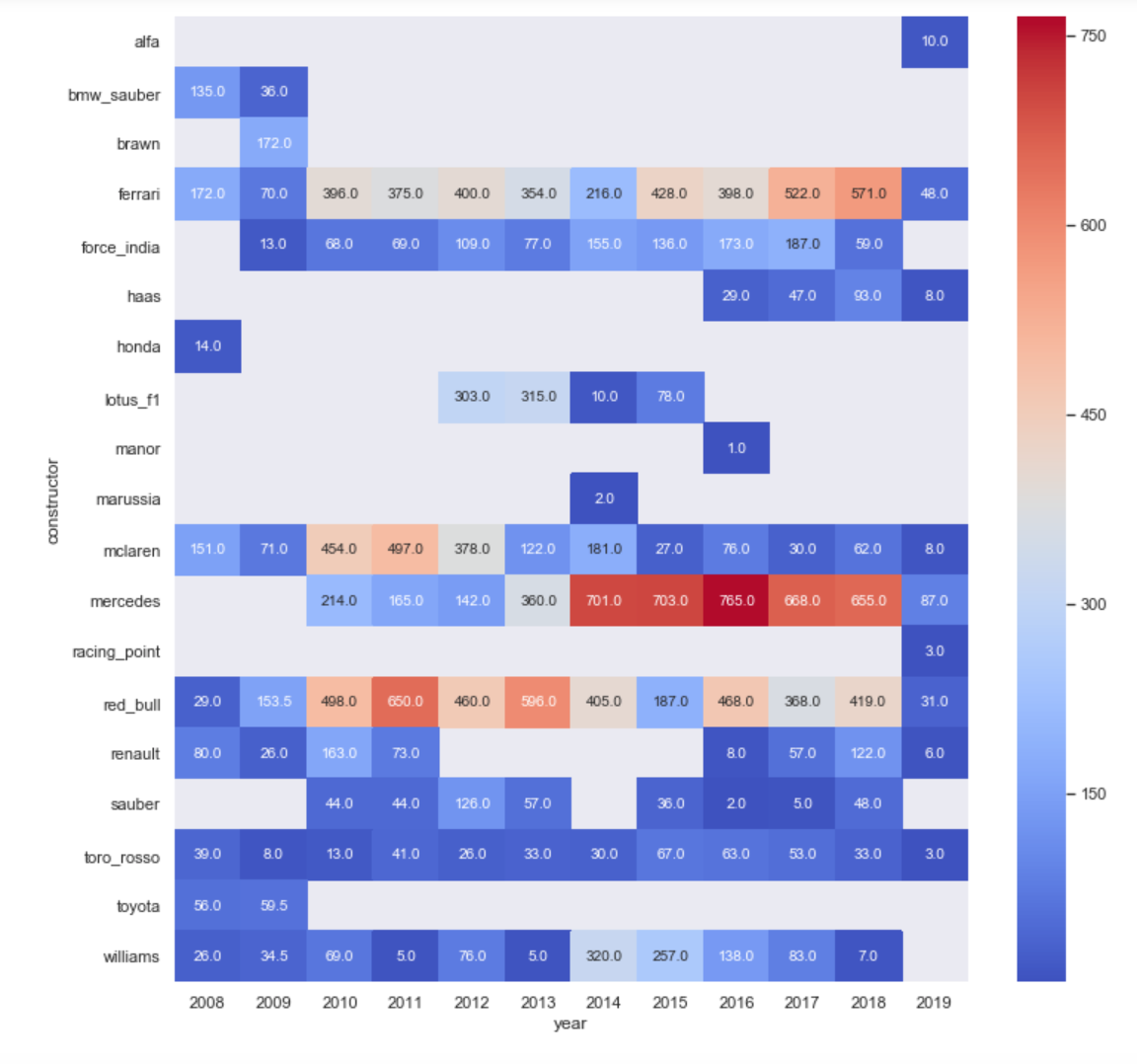
| **Dataset** | **Number of Variables** | **Sample size** |
| --- | --- | --- |
| Constructor\_full.csv | 18 | 12126 |
| Results\_full.csv | 25 | 24237 |

# Exploratory Data Analysis

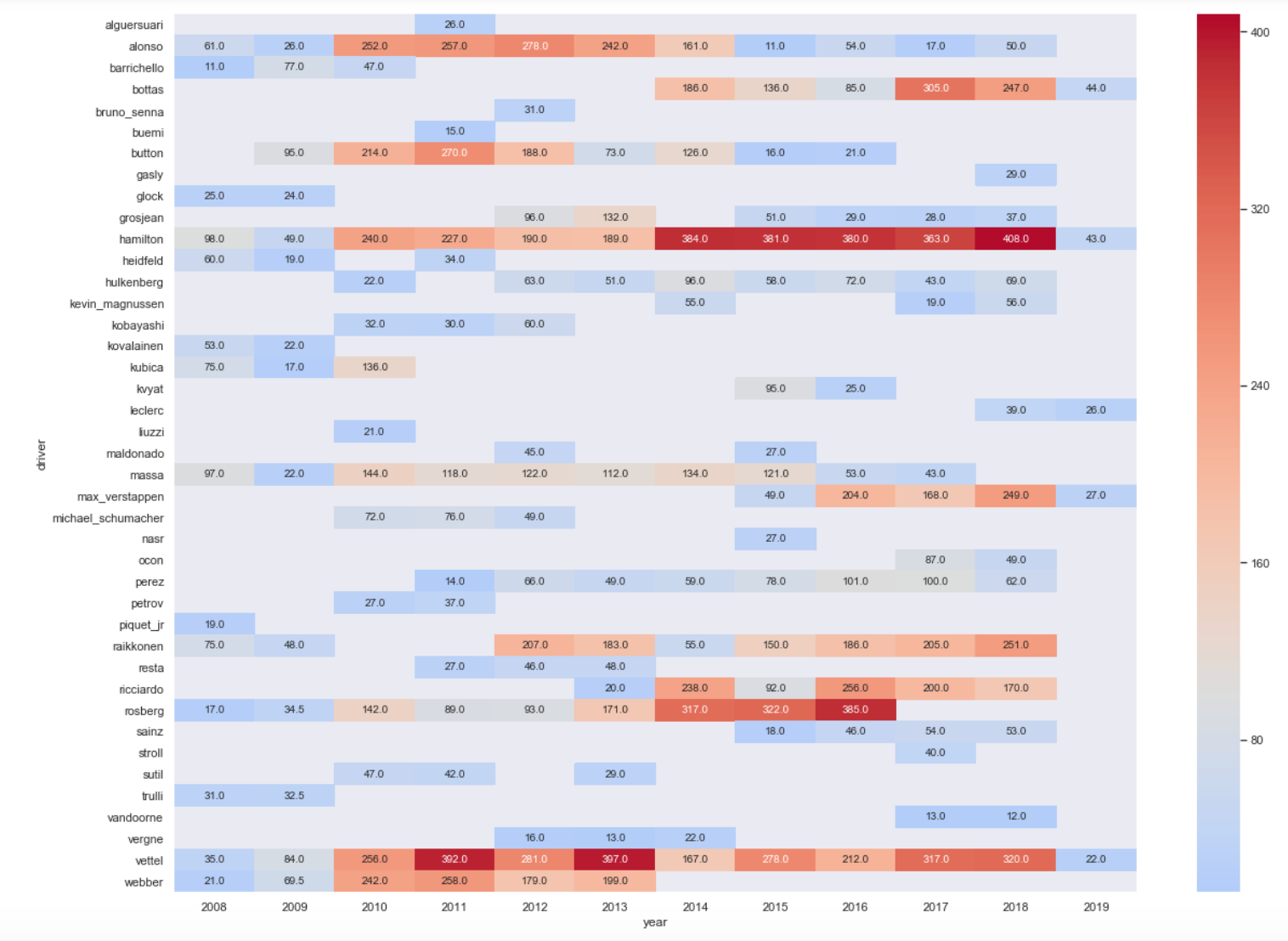
To begin understanding the contents of the dataset and relationship between constructors, drivers and circuits, the constructor and results data frames are imported into a Jupyter notebook and various plots and graphs are generated. The main goal at this stage of the project is to visualize the data and begin performing feature selection for modeling. Multiple plots and charts are developed a few of which are explained below.

From the constructor data frame a total x constructors from 23 countries have participated in Formula One racing between 1950 -2018 with Britain and Italy having the most number of constructors. A plot is created to show the number of races held in each country as well as the number of circuits in the country. As shown in Figure 1 below, the United States has 10 circuits which is the highest number of circuits amongst all the countries even though it ranks lower with the number of races held

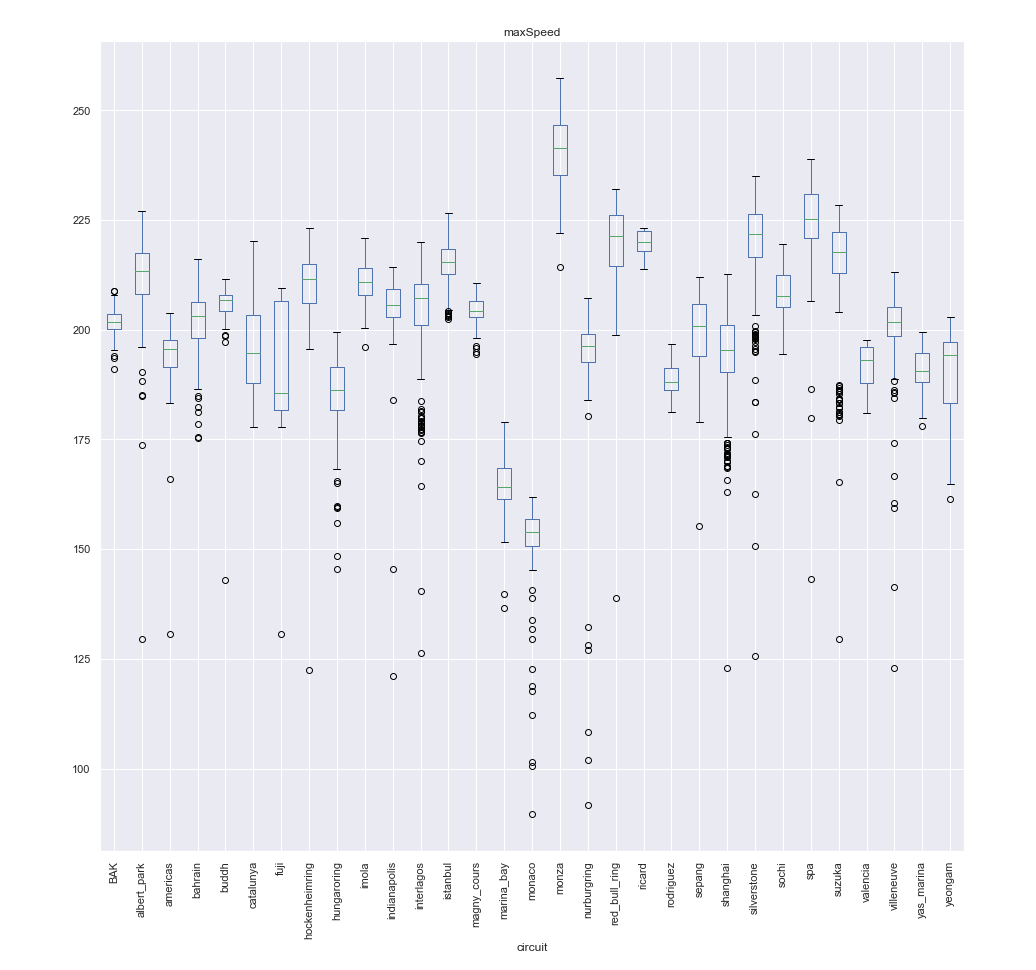
Limiting the scope to just the past 10 years, the distribution of points for the top 20 constructors is analyzed. A heat map generated to visualize this shows that although the total number of constructors is high, not every constructor participates in races every year. Ferrari, Mercedes and Redbull seem to be the most competitive constructors in the past decade with Mercedes having the overall highest point record of 765 in 2016.



A similar heat map is also generated to show the distribution of points accumulated for the top drivers in the last decade. Similar to the constructors, data isn’t available for every driver in every year. Although the data seems to be uniformly distributed for most of the drivers, there are a few hot spots showing drivers, Hamilton, Vettel, Rosberg and Webber having significantly consistent high point rankings for multiple years in a row.



Considering the fact that every Formula One circuit is designed differently we explore how the vehicle speeds vary by circuit. The box plot below shows the key statistics for the speeds recorded at the circuits that have held the most racing over the years. Monza is the circuit with the latest speeds with average speed being around 245mph.



In the next step of the project, features from the exploratory data analysis like the Driver ID, Constructor ID, Circuit, Year of the race and country being represented will be used to generate supervised and unsupervised machine learning techniques to build classification and predictive models.

# Model Creation and Evaluation

# Results

# Discusion

Most of the kernels on kaggle that have explored this dataset have limited their scope to simply exploratory analysis without any prediction. Our project goals differ from what other analyst have used this dataset for in that, in addition to performing exploratory data analysis to provide insights on the sports statistics, we will be building models to classify drivers, constructors and races as well as predict race outcomes. At this stage of the project, we have successfully preformed the following ;

* Normalized our data, by combining tables from the relational database schema.
* Charted some statistical analysis such as the team and drivers with the most historic wins, countries with the most wins and circuit speeds.

In the upcoming stages of the project we will continue to expand our analysis to include

* Finding correlations between data elements such as drivers and track, qualification position and race win by track
* Applying clustering techniques against various data elements such as drivers constructors, countries and circuits.
* Selecting features that can predict the winning race outcome and its probability
* Generating classification and predictive models.

# Conclusion

# Team Member Contributions

# References

[1] Bunker, Rory, and Fadi Thabtah. “A Machine Learning Framework for Sport Result Prediction.” Applied Computing and Informatics, Elsevier, 19 Sept. 2017, <https://doi.org/10.1016/j.aci.2017.09.005>

[2] Tulabandhula, Theja, and Cynthia Rudin. “Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing.” *Big Data,* vol. 2, no.2, June 2014, pp. 97–112., https://[doi.org/10.1089/big.2014.0018](http://doi.org/10.1089/big.2014.0018)