**Personal intro**

-Hello my name is Alejo Fernandez. I am a fourth year student at the University of California Los Angeles. I am pursuing a degree in Philosophy and minoring in Digital Humanities.

**Project intro:**

In this project I want to analyze Formula 1 data, at first I was focusing on the correlations between a pitstop and the final position for a driver in the same grand prix. The type of data that I wanted to analyze in Formula 1 is limited and not much of it is publicly available. Therefore, I had to transition the subject of my project from pit stop correlations to lap times and lap speed analysis in Formula 1 from 2014-2017.

Context:

What is formula 1

-Formula One (also Formula 1 or F1 and officially the FIA Formula One World Championship) is the highest class of single-seat auto racing that is sanctioned by the Fédération Internationale de l'Automobile (FIA). The FIA Formula One World Championship has been one of the premier forms of racing around the world since its inaugural season in 1950.

Why is this interesting?

- At first I decided to focus on this question because I think pit stops are the epitome of the team work factor that goes into a grand prix. It’s an instance in which all three aspects are working simultaneously. The driver, the pit crew, and the race strategist/engineers need to work harmoniously in order to have a successful race.

- The driver has to hit his marks, not exceed speed on pitlane entry and have a safe pit exit.

- The crew must be precise, a group of specialized pit crew that are in charge of the vehicles modifications. This occurs in under 5 seconds and even under 2 seconds.

- The strategist/engineers in the team have to be aware of everything occurring in the race from the start. Taking into consideration countless variables that are unpredictable. They have to time the pit strategy based upon what everyone else is doing, at some points it becomes of a game of who is going to make the first move to begin the race strategy. The importance of the pitstop for a driver is critical because a driver can have the perfect race and go into their MANDATORY Pit stop yet have a crew member leave a wheel nut loose which causes your wheel to fly off as you are entering the track. This ruins your race. If the race strategist make a premature call to enter the pit to change tire compound, it can ruin the race by having no grip at the end of the grand prix or by having too much grip and ruining the tires before the end of the race. Additionally, the engineers can order to make changes to the vehicles if the driver is complain. The engineer then communicates that with the crew and the crew execute the changes. There are countless changes but a change of camber, caster, toe, or wing angle are some of the more common adjustments effectiveness of data analytics in finding the insights from the data that are of high significance and efficient.

**pitstops, rules/procedures**

Once drivers enter the pitlane, they must slow down to 80kph (50 mph) in the race. This is to ensure the safety of the pit crews and everyone working in and around the pit lane. Drivers have a special button on their steering wheels that they press, which limits their speed automatically in the pit lane. However, if drivers fail to respect the speed limit, they will receive a time penalty.

Pit stops in Formula 1 are a complicated process and involve over 20 people. Teams practice them for hours during the offseason and in between track sessions on a race weekend. They must be as quick as possible as in Formula 1; mere split seconds can be the difference between coming out ahead or behind a chasing car.

Each of the 23 people involved in a pit stop has a specific role, and all are as crucial as the rest. A front jackman and a rear jackman are responsible for lifting the car off the ground to change the tires. The jacks they use cost over USD 300.000 as they are highly specialized pieces of equipment that need to lift the car with ease.

Some four-wheel gunners are responsible for unfastening the old wheel and fastening the new wheel. Someone removes the old tire to help the wheel gunners, and another person helps put the new wheel on. Therefore, for each wheel attachment on the car, there are three people responsible for the wheels. Two mechanics are also responsible for holding the car steady while the pit stop is taking place, and another two adjust the front nose cone if necessary. Once the car has been stopped, the mechanics will quickly change the tires. Ever since mid-race refueling was banned in 2010, the average pit stop time if no front wing change is required is around two to three seconds. However, if the front wing has been damaged and needs to be changed, pit stops will take well over 10 seconds as it takes much longer to remove and replace the front wing than it does with tires. Red Bull performed the fastest pit stop in Formula 1 history at the 2019 Brazilian Grand Prix, and they changed Max Verstappen’s tires in only 1.82 seconds. This does not include the time it takes for drivers to enter and exit the pit box. The entire process takes between 20-30 seconds, depending on the length of the pitlane at the specific track.

**Additional context on F1 2014-2017 season**

Formula 1 fans were stuck with races that were producing less than exciting racing because of how dominant Mercedes was and is. Additionally, Red Bull and Ferrari were not only struggling with their own cars but were also in the midst of changing their driver lineup. Pair a dominant duo of Lewis Hamilton and Nico Rosberg with Mercedes F1 W05 and the begging of a dominant era began while other teams were trying to catch up. From 2014-2021 Mercedes won the World Constructors’ Championship, in that time frame Red Bull went through multiple driver line ups. Along with other teams on the grid, the movement and dropping of drivers to replace another hinders the drivers ability to get familiar with the car. As a result I think this played to Mercedes advantage in having a pair of dominant driver contractually locked in for years when they began their Formula 1 domination.

**Motivation to study this topic**

Part of the motivation to study Formula 1 data sets is because I think Formula 1 is the epitome of motorsport, the biggest budgets and the newest technology. Not only are Formula 1 cars ridiculously fast around a track but drivers and crew members must perform at the highest level or risk getting fired. Additionally, I wanted to focusing my research project on two aspects:

1.) A topic that I find interesting

2.) A topic that is interesting to motorsport enthusiasts.

Since F1 is the highest form of motorsports I figured I could reach a larger audience if I focused on a series like F1 instead of smaller series that only have a niche audience.

**Any prior related work**

-Fun fact I am also a professional race car driver. I began my racing career at age 16, when I began with kart and at age 18 after high school I turned pro and began to compete in Rallycross.

-My first race was under the FIA Rallycross Championship in the RX lites division in Montalegre Portugal. My second race was at X-Games Austin 2014.

-I competed in Red Bull Global Rallycross championship were I obtained multiple podiums, fast laps, and pole positions.

**Methods//Data introduction**

The Data:

-The data I used was found on kaggle.com

-The data sets I gathered contains data from 2014 through the 2017 season, and consists of tables describing constructors, race drivers, lap times, pit stops and more.

-The files I used were CSV type. I created a .txt file that had drivers name with their id since the original data set used numbers as its primary form of reference for a driver.

Acknowledgements:

-The data was downloaded from http://ergast.com/mrd/ at the conclusion of the 2017 season.

-The data was originally gathered and published to the public domain by Chris Newell.

**Limitations of the data set**

## Limitations of the Data set

The limitations that I had with the data set data are the following:

-Formula 1 data in regards to performance is difficult to find

-The data that was public and available didn’t have pit stop information that I was looking for. I wanted to focus my project for the 2018-2021 Formula 1 seasons but I wasn’t able to find much data about those particular years or the data that I did find, didn’t have the data I was looking for such in regard to pit stop times.

-Therefore I decided to focus on what was available and see where my analysis would take me.

**Analytical process**

I analyzed the data in this project both quantitatively and qualitatively. I analyzed and viewed the data sets individually at first and then I merged different dataset with matching relevant features. At first I wanted the project to focus on pitstop correlations but that was difficult because wasn’t able to split the dataset by year, so the pitstop plot I created is cumulative. Because of that I wasn’t able to use the pitstop data how I wanted so I decided to see what type of plot I could create from the pitstop data and then figure out what other type of question I can infer from the data. I created a plot that demonstrated the average time duration (AKA: pitstop) it took drivers using the data from 2014-2017.

**Data cleaning**

-I went through trial and error trying to understand the data I had. So, I created a statistical analysis that included the relationship between all individual features in the dataset along with a correlation coefficient matrix, number of nan values, skewness of features, and heatmap to further analize the data I had.

I used I used numpy,pandas,matplotlib, and seaborn to analyze this project

#This was the first data frame I created, it was obtained from the csv file titled 14-17 races and has the following columns: raceid,year,round,circuitid,name,date,time, and a Wikipedia url

Number of Nan Values

-The number of Nan Values is critical in understanding and manipulating the data. This helped me analyze the missing values in the rest of the datasets that I had. In turn this allowed me to see the usability of the data since there were some errors in the data set. This boils down to the way that the data system operate.

Skewness:

The skewness explains the density of the data. This can be significant to look at if we use a specific year to observe high and low significance. This helps find the features that can see a decline or estimation over the years. If the data skews negatively then those are see as in an increase in their distribution over the coming years.

Heatmap:

The color in the heatmap represents the level of similarity between the features. The colors range from deep to light blue. If the color is a deep blue (1) it mean that the two features are highly correlated to one another.

# Now I am going to take a look at the pitstop CSV file so that I can create data frames like i did with the race file.

# The pitstop file contains the following columns:

#raceid, driverid,stop number, lap performed, time of day the pitstop was performed, the duration, and its equivalence in milliseconds.

**Methodology**

1. **Dataset**

The Data:

-The data I used was found on kaggle.com

-The data sets I gathered contains data from 2014 through the 2017 season, and consists of tables describing constructors, race drivers, lap times, pit stops and more.

-The files I used were CSV type. I created a .txt file that had drivers name with their id since the original data set used numbers as its primary form of reference for a driver.

Acknowledgements:

-The data was downloaded from http://ergast.com/mrd/ at the conclusion of the 2017 season.

-The data was originally gathered and published to the public domain by Chris Newell.

## Limitations of the Data set

The limitations that I had with the data set data are the following:

-Formula 1 data in regards to performance is difficult to find

-The data that was public and available didn’t have pit stop information that I was looking for. I wanted to focus my project for the 2018-2021 Formula 1 seasons but I wasn’t able to find much data about those particular years or the data that I did find, didn’t have the data I was looking for such in regard to pit stop times.

### Analytical process

-Therefore I decided to focus on what was available and see where my analysis would take me.

I analyzed the data in this project both quantitatively and qualitatively. I analyzed and viewed the data sets individually at first and then I merged different dataset with matching relevant features. At first I wanted the project to focus on pitstop correlations but that was difficult because wasn’t able to split the dataset by year, so the pitstop plot I created is cumulative. Because of that I wasn’t able to use the pitstop data how I wanted so I decided to see what type of plot I could create from the pitstop data and then figure out what other type of question I can infer from the data. I created a plot that demonstrated the average time duration (AKA: pitstop) it took drivers using the data from 2014-2017. After, was when my research question changed and i began to explore the data.

In depth explanation of analysis

Data cleaning

-I went through trial and error trying to understand the data I had. So, I created a statistical analysis that included the relationship between all individual features in the dataset along with a correlation coefficient matrix, number of nan values, skewness of features, and heatmap to further analize the data I had.

-I used numpy,pandas,matplotlib, and seaborn in this project.

Number of Nan Values:

-The number of Nan Values is critical in understanding and manipulating the data. This helped me analyze the missing values in the rest of the datasets that I had. In turn this allowed me to see the usability of the data since there were some errors in the data set. This boils down to the way that the data system operate.

Skewness:

The skewness explains the density of the data. This can be significant to look at if we use a specific year to observe high and low significance. This helps find the features that can see a decline or estimation over the years. If the data skews negatively then those are see as in an increase in their distribution over the coming years.

Heatmap:

The color in the heatmap represents the level of similarity between the features. The colors range from deep to light blue. If the color is a deep blue (1) it mean that the two features are highly correlated to one another.

1. **Missing value analysis**

-The two features' time and milliseconds have more than 50% nan values this means that it will be very difficult to understand the two columns with that many nan values.

-Two features that had more than 50%, one at about 19%, and the remaining are of less than 6%.

-I removed the features that had a nan value greater than 15% from the dataset.

-Now the data frame will not account for the other features that have any percent missing greater than 50%

-Therefore time and milliseconds will not be accounted for in the new data frames I create.

1. **Statistical Analysis**

-Even after I have cleaned the data, based on this statistical analysis, a normal person wouldn’t be able to deduct anything from this.

-That was part of the issue I was encountering with the pit stop data at first.

-This is another instance in which the systems and methods the people who gathered this data can be confusing.

-An clear example would be to look at the driverid, a driverid min is 1 which is correlated to Lewis Hamilton arbitrarily. The max driver 1d value is 843, so I’m assuming that’s driver ids account for the number of drivers that have ever participated in Formula 1. However, this is exactly my point, Hamilton was not the first driver in Formula 1 the majority of his competitors have a driver id higher than 800.

1. **Skewness Analysis**

-This skewness analysis helps paint a picture of the data distribution.

-The data set distribution has both negative and positive distribution of features.

-The feature status ID has a high positive distribution and the driver id has a negative distributions.

-The remaining features have very low positive index distributions between 0 to 1 and this will indicate the remaining features demonstrate a normalized distribution over dataset samples.

Chart, histogram

Description automatically generated

Skewness Analysis over the features

1. **Correlation Analysis**

-The Correlation analysis was very important as it provided insight into the internal relationship between the features in the dataset.

-This will help hone in on the features that are more relevant to the output feature. Meaning, this can also help eliminate the redundant features by producing the same coefficient score in each feature.

-The correlation analysis was carried out on all features and their relationship was projected in terms of coefficients.

- Heatmap:

The color in the heatmap represents the level of similarity between the features. The colors range from deep to light blue. If the color is a deep blue (1) it mean that the two features are highly correlated to one another.

A screenshot of a computer

Description automatically generated with medium confidence

Correlation Analysis and their coefficients presentation

1. **Analysis between Features**

### The analysis between features

-The analysis was performed to understand the effect of variation of one feature with another features and distribution between them along with density and sample measurement.

-I chose a few features and their density, joint density distributions were observed along with time and speed features. The feature's position and points were analyzed along with time and daytime distributions.

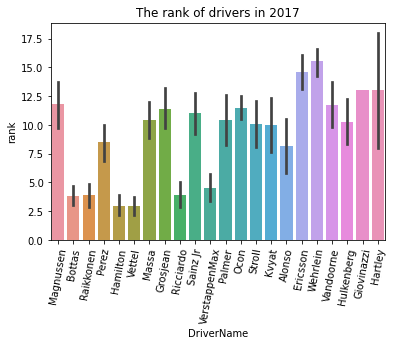
-For example look at resultid race id on the left upper corner correlate with the result id and race id on the x-axis. It demonstrates 1, showing the normalization of the distribution over the data samples.

-This type of analysis will help me to understand the internal relations between features.

Heatmap:

The color in the heatmap represents the level of similarity between the features. The colors range from deep to light blue. If the color is a deep blue (1) it mean that the two features are highly correlated to one another.

**Results:**



Drivers Rank in 2017 racing competitions

By analyzing the above plot, we can observe that Formula 1 driver, Wehrlein has the highest

rank among all of the drivers in the 2017 Formula 1 season.

NOTE. The lower the ranking number is for a driver means that they performed better than a higher ranking driver..

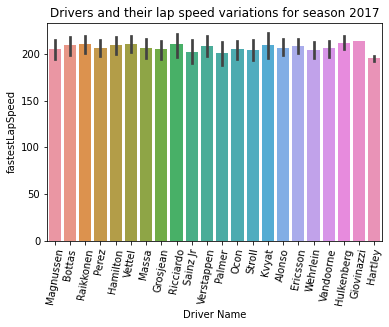
Just behind Wehrlein is Ericsson and this makes sense in comparison to all of the other drivers because based on the 2017 season both Wehrlein and Ericsson attempted the most GPs and had the

worst results out of their competitors, cumulatively.

On the other hand we can observe how Hamilton and Vettel have

the lowest rank in the plot. This further demonstrates

The drivers Hamilton and Vettel are in the lowest rank in 2017 racing challenges.



Plot between driver and their speed in season 2017

-The plot above demonstrates the each drivers fastest lap speed visually.

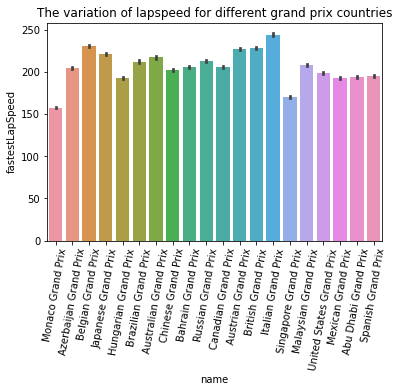
-Hartley had one the worst fastest lap speed. This is a combination of car and circuit.

-The next plot will demonstrate what I am alluding to here. Hartley wasn’t a part of the Italian Grand prix which is known for its speed.

-Hartley’s performance is not purely a result of being in a slow car it is also a result of being absent at one of the fastest Grand prix in the Formula 1 calendar.

-Furthermore, Giovinazzi was also a drivers performance that needs some explanation. He replaced Wehrlein in the 2017 season at the Australian Gp and at the Chinese Gp. He might've got lucky with a monster lap at the Australian Gp receiving a tow from his competitors.

-Notes: The Australian Gp is one of the faster tracks on the calendar and a tow is when a car breaks the slip stream in front of you. This causes you to draft behind them and gain an advantage in speed.



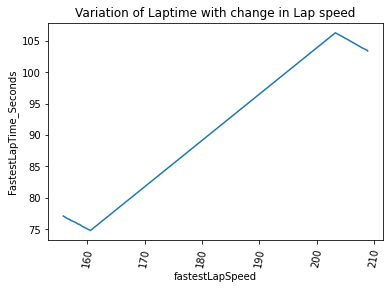
-Plot above demonstrates the lap speed variation of each Grand prix that took place in the 2017 Formula 1 season

-The above analysis describes the Grand prix that has the highest and lowest lap speed.

-The Italian Gp was the fastest speed while Singapore occupied last place in the lap speed.

-Deeper analysis for Singapore having such a drastic drop in speed found that Singapore Gp a race that began under wet conditions. This significantly reduces the speed of the vehicles. As a result of the treacherous conditions the race was cut short, drivers did not complete the scheduled race distance of 61 laps due to the 2-hour time limit being reached, so the chequered flag was shown on lap 58.

-Interesting feature to compare is the lap speed between the Singapore Gp and the Monaco Gp, it just goes to show how ridiculous Formula 1 cars can be. In dry conditions around the Monaco Gp the fastest lap was a little over 150mph. At the Singapore Gp where it was a rain race from the get-go they exceed speed of that dry conditions in Monaco. It was also a crash fest so a lot of the fast cars didn’t register a fast lap.



Plot between lap speed ad lap time

-This analysis was extremely interesting because of the way the values are plotted; there is a dip at the beginning and end

-The plot demonstrates that there is a dip in speed when the lap times are longer than 105 seconds. This could be due to the fact that race organizer need to have a track layout in mind that will accommodate for high speeds therefore the track distance becomes longer. Longer straights = longer time vehicle is in a straight line optimal to reach top speeds.

-The higher the speeds you’ll need more track area to decelerate or account for run off area in case of emergency (brake failure)



Plot between lap time variation for grand Prix countries

-The plot indicates the fastest time lap taken by each Grand prix in the 2017, F1 season.

-The results are really cool to analyze here because we can see the Belgian Grand prix takes place at Spa leng of 4.2 miles long

-Australian Gp has a legth of 3.2 miles

-Austrian GP is 2.6 miles long and unlike Monaco which is also a short cirucit, the Austrian Gp is much more of a race track that allows drivers to maintain a hgher lap speed unlike the Monaco Gp

Chart, bar chart

Description automatically generated

Pit stop Data analysis

-It's difficult to take gather information from this but I was able to deduce information from it.

-The first thing I was able to infer from the plot was Mercedes Consistency with pitstops. Looking at Hamilton bar, his seems to have the shortest duration in pitstops from 2014-2017. This has to be a significant factor in obtaining the championship for those seasons.

-Bottas was interesting because from 2014-2016 he raced under Williams, they were known for their quick pitstops. During his time at Williams his team mate was Felipe Massa, his bar also demonstrates quick pitstops.

-Harleys pitstop bar demonstrates that it was low in time duration however this is a result of his lack of participation in formula 1. Since the data frame holds all of the pitstops from 2014-2017, cumulatively this increases the average pit stop duration because some track have a longer pitstop, meaning the pitstop is linger in duration.

**Discussion**

### What my analysis shows:

-A variety of information in regards to Formula 1 lap speed and lap times during the 2017 season. I also analyze the pit stop duration for all drivers that competed in Formula 1 during the 2014-2017 race seasons.

-Personally one of the analysis that I enjoyed the most was the variation of lap time with change in lap speed. I thought this was an interesting plot and i think most people would be surprised to know that lap speed above 200mph will make your lap faster anything less will mean that the lap time will increase the duration of the lap time. Essentially there waw. so anything less than 160 mph and anything over 205mph will a faster lap time comparatively speaking in comparison to tracks that are stuck in between those ranges in speed.

### What is the big picture:

-I am eager to explore motorsports data in the future. I wanted to focus first on F1 because I thought i would be able to find interesting data. It had it's pros and cons. The big picture overall was to familiarize myself with the data that I was ablet to obtain and see what type of analysis I could create. I wanted this project to be the foundation of a YouTube channel I build in the future. One that is focused on motorsports, that gives a unique perspective on them by combining my python abilities to find data and my knowledge with race cars and racing. This will allow me to explore different topics than other Motorsport content creators on YouTube.

### How are these findings useful

-These findings can be useful in a number of ways.

-The findings in this project can be useful to create informational/educational content. For people that are wanting to find information in F1 from 2014-2017. I think this project will be helpful for other motorsport enthusiast since there aren't many Formula 1 analysis.

-This project will be useful for me as it will provide a foundation of what type of data i should look for and how to approach another research topic in the future.