The data comes from Formula 1 racing. There is an API called the Ergast Developer API (<http://ergast.com/mrd/>) which “is an experimental web service which provides a historical record of motor racing data…from the beginning of the world championships in 1950.” From this API comes the FastF1 package (<https://github.com/theOehrly/Fast-F1>) which is “a python package for accessing and analyzing Formula 1 results, schedules, timing data, and telemetry.” Due to constant technological advancements, not all seasons have the same data – the timing data, car telemetry, and position data are only available for seasons dating back to 2018. From the FastF1 package, I created hundreds of dataframes and consolidated them into four major dataframes: weather, laps, messages, and results.

A brief overview of Formula 1 should be in order now that the data sources are known. Formula 1, also known as Formula One and F1, cars are the fastest road-course racing cars in the world – often exceeding 200 mph seconds after breaking to below 50 mph. There are ten teams, each with two drivers, competing for two concurrent awards – best driver and best constructor.

The driver with the most points at the end of a season is awarded the Drivers' Championship Award.  Teams who have accumulated the most points from their drivers receive the Constructors’ Championship Award. The point breakdown is below:

| **Position** | **Points Scored** |
| --- | --- |
| 1 | 25 |
| 2 | 18 |
| 3 | 15 |
| 4 | 12 |
| 5 | 10 |
| 6 | 8 |
| 7 | 6 |
| 8 | 4 |
| 9 | 2 |
| 10 | 1 |
| 11-20 | 0 |

In 2021, Red Bull drivers finished 1st and 4th overall, with the two drivers scoring a combined 585 points. That same season, Mercedes drivers finished 2nd and 3rd overall with the two drivers scoring a combing 613 points. Red Bull had the top driver while Mercedes was deemed the top constructor.

There are a different number of races in each season, and a mix-mash of race track continuity. For example, Italy, which has two tracks, has been used 9 times in the past 5 years; meanwhile, the United Arab Emirates has only one track and has only been used 1 time in the past 5 years.

Let’s dig back into the dataframes and variable selection.

1. Weather – measures weather-related events every minute of each race
   1. RACE – the race track aka course
   2. AirTemp – the temperature of the air in Celsius
   3. Humidity – the measure of humidity
   4. Pressure – the measure of air pressure
   5. Rainfall – Boolean
   6. TrackTemp – the temperature of the track in Celsius
   7. WindSpeed – the measure of the wind
   8. Year – the year of the RACE
2. Laps – provides a lap-by-lap breakdown for each driver
   1. RACE - the race track aka course
   2. LapNumber – the foundation for the dataframe. Every observation in this dataframe is broken into a row by the lap number.
   3. PitInTime – the time, from the start of the race, a pit stop is made
   4. Year – the year of the RACE
3. Messages – communication from “Race Control”, the entity responsible for monitoring and supervising the race.
   1. RACE – the race track aka course
   2. Flag – shows if a flag has been issued. There are 10 different flags with their own meaning. For our purposes, we are only concerned with YELLOW, DOUBLE YELLOW, and RED flags which indicate hazards on the track which prevent drivers from passing another driver and slows down the speeds the drivers at which the drivers can travel. Further reading can be found here (<https://onestopracing.com/flags-in-f1-explained/#:~:text=Each%20flag%20used%20in%20F1,means%20a%20driver%20is%20disqualified>.)
4. Results – shows the final results of each race for all 20 drivers
   1. RACE – the race track aka course
   2. Year – the year of the RACE
   3. Driver – the three letter abbreviation of the driver
   4. FullName – the full name of the driver
   5. Points – the number of points won by the driver/team
   6. TeamName – the name of the team sponsoring the driver

The idea for this shiny app was to visualize the data for my thesis. Dr. Beregovska and I need to see what the data shows so we know what steps to take next. The idea of the thesis is to try and predict when a team will make a decision to have their driver come in for a pit stop. Pit stops are among the most strategic decisions a team can make, so being able to predict when a competitor will pit can completely alter one’s strategy and affect the outcome of the race.

The idea for predicting the decision to make a pit stop stemmed from a book by Dr. Bruce Bueno de Mesquita Game (<https://en.wikipedia.org/wiki/Bruce_Bueno_de_Mesquita>) called The Predictioneer’s. He is known within the field of selectorate theory (<https://en.wikipedia.org/wiki/Selectorate_theory>) and his model is classified, but is known to use the Schofield mean voter theorem (<https://polisci.wustl.edu/files/polisci/imce/z.1052006.3.respub.2007.pdf>) and solves for Perfect Bayesian Equilibrium (<https://en.wikipedia.org/wiki/Perfect_Bayesian_equilibrium>).

The goal of this app is to make passable graphs that a user can manipulate to show various changes and provide a story to the data. I didn’t want to lose track of the purpose of this class – which is the visualization of data – so I also generated maps of each race track in Python. The colorization of the track relates to the gear number of the driver who ran the fastest lap (<https://theoehrly.github.io/Fast-F1/examples_gallery/plot_gear_shifts_on_track.html>).

Does a story emerge?

The question is to simply see what the data shows us. To do this, we considered pit stops by lap #, driver, team, and course.

We looked at how various teams and drivers performed looking at their point accumulation.

We also looked at the weather data, taking the mean of five measurements. Those 5 measurements were considered both by race and as a comparison of the entire dataframe.

And the final consideration was the number of hazardous flags for each course since yellow/double yellow/red flags minimize time-loss when making a pit stop, thereby increasing the likelihood of a pit.

The course visualization is good to see what the course looks like - the number of turns or length of straightaways can affect tyre life.

Also, being in a higher gear indicates higher RPMs which would reduce tire tread further and cause more pit stops.

The next tab shows the lap number when a pit stop was made. The top graph shows it by the course, adjusting for the range of years.

The bottom graph does not reduce it by the selected course, but still does adjust for the range of years.

We can see, only looking at the 2018 year, that the most frequent time to get a pitstop was lap # 31 in China. There were 6 pit stops on lap 31.

However, looking at the entire 2018 season, only 7 pit stops were made on lap #31. Did something happen in the 2018 China race to cause many drivers to come in for a pit stop on lap 31?

Similarly, looking at Monaco 2018, the majority of pit stops (7) were made on lap 17. Across the entire 2018 season, lap 17 was the third most populous lap # to have a pit stop. What makes Monaco fall in-line as opposed to China’s result?

We can look at tyre life as an obvious strong indicator of pit stop frequency. Looking at the total tyre life histogram, it actually follows a pretty normal distribution with a center around 19 laps.

So, when we look at the tyre life for Italy across all 5 years (9 races), it is interesting to see a strong left-skew distribution with the four most frequent ages of tyres occurring at laps 23, 27, 30, and 40.

Looking back at the visualization, it seems the Italian races have relatively few turns so perhaps that played a big role in the tyre life.

On the flip side, in 2022, Japan’s two most frequent age of tyres before coming in for a pit were at laps 2 and 5. Weather played a major role in this race as there was a significant downpour. In 2022, laps 2 and 5 were both outside the top-10 most frequent age of tyres when coming in for a pit stop.  Japan accounted for 50% of all lap 2 pit stops and 33% of all lap 5 pit stops.

Speaking of weather, the next tab analyzes the weather. It is cler to see that Japan’s weather falls outside the norm across all years and courses.

In 2022, their average air temperature was 16 degrees Celsius. The average F1 race over the past 5 years is 24 degrees.

The humidiy, air pressure, and track temperature can also play a large role at vehicle handling – look at the discrepancies in Mexico compared to the norm.

The next tab covers hazardous flags by the user inputted course and years. Flags have already been discussed. This needs further development – perhaps overlapping the lap number plot with a yellow, red, or black bar when a flag is initiated.

Unfortunately that would require further dataset manipulation using timing data down to ms and was not in the scope of this assignment.

The next two tabs are similar. We are looking at the number of points accumulated by teams and drivers on the Points tab. And on the Pit tab, we are looking at the number of pit stops by teams and drivers.

Both the Points and Pit tab depend on the course selection and years.

An interesting observation is the change in driver’s teams as they can switch teams year-over-year. Also, there are many more drivers due to the F1 construction layout, retirements, promotions, and covering for ill-drivers.

There should be a better way of visualizing the pit data to indicate which season is covered by the pit stop data. For example ZHO has only been around since 2022 so it makes sense he only has 1 pit stop as compared to HAM who has been around for 10+ years who has 4 pit stops. A more apt comparison would be looking at Spain and comparing ALO to HAM.

And on the final tab is just a preview of the data that can be downloaded on the side panel, showing what we have to work with.

To answer the question, I believe that the story which emerges regarding the ability to predict pit stops is dependent upon the course, the weather, tyre life, and flag status.

Further considerations could be global variables such as the team and driver’s goal (only a few teams compete for first, a few compete for the next few spots, and the rest are just trying to field a team that occasionally gets points).

Analyzing the telemetry and actual lap data would be imperative. If a normal pit stop takes 20 seconds and the driver in 1st has a 30 second lead on older tyres, then a pit stop would be expected. However if the driver in 1st was only ahead 10 seconds, he would come out 10 seconds behind the driver in 2nd (removing all other drivers from the field) and a determination will need to be made is if the driver on new tyres can reclaim his position.