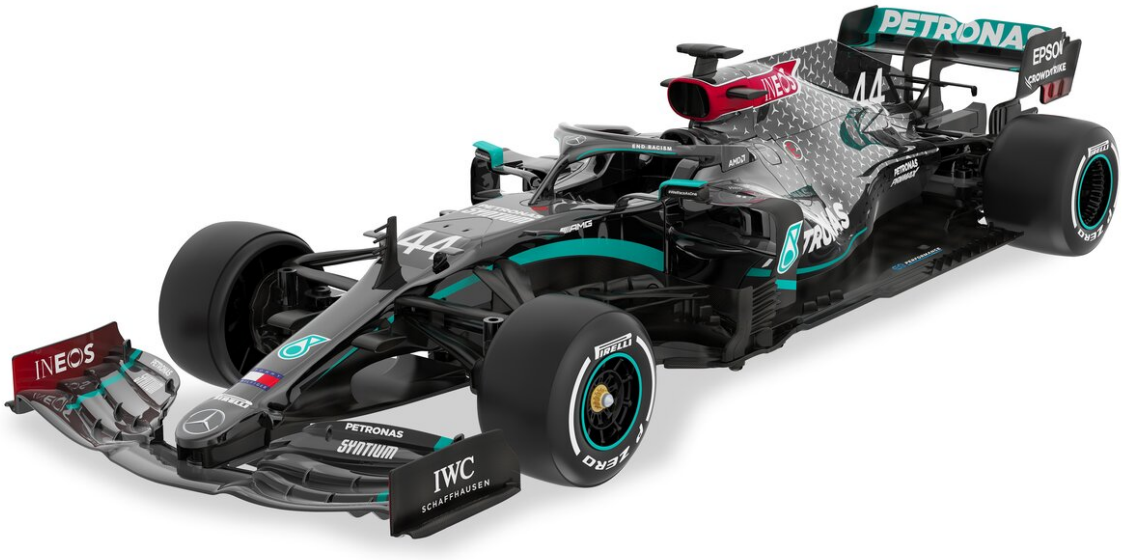
Data Science Project

Formula One Data Analysis



Under the supervision of: Dr. Dina Elreedy

1. Team Members and Contributions:

1- Tarek Yasser

- Question 1: “Does a team being the manufacturer of their engine affect performance?”
- Question 2: “Does having a race in your country affect performance?”
- Question 3: “Given that Mercedes drivers have above average top speed, does this apply to all German teams?”

2- Abeer Hussein

- Question 1: “How does a track's altitude affect top speed / average lap time?”
- Question 2: “Which tracks have the most DNFs?”
- Question 3: “Which tracks favor overtaking?”

3- Mohamed Khaled

- Question 1: “ Predict race result ”
- Question 2: “ driver average position ”
- Question 3: “optimal number of pitstops”

4- Karim Taha

- Question 1: “Average retirement age”
- Question 2: “Age Influence on Performance”
- Question 3: “Time-series analysis on the drivers’ standings”

2. Data analysis cycle and epicycle

Does a team being the manufacturer of their engine affect performance?

Stating and refining the question

a. Setting expectations
   One wouldn't be overreaching to expect that a team manufacturing their own engine would be beneficial to performance. Whether that's points, wins, or some other metric. Having a tightly integrated system in any context usually results in better performance.
b. Collecting data (questions, or results)
   Our dataset doesn't contain any information pointing to whether a team was manufacturing their engine at some point in time. **So we had to manually collect data for the time frame 2010-2020. The data sources can be found here (2012-2020), here (2011), and here (2010).**

   Some notes on the engine manufacturer data: some teams changed names during the analysis time frame (2010-2020), so we had to also manually look those up and preprocess the data such that the names are updated. We also had to do this to the rest of the dataset **before** exploratory data analysis so our tables can be joined properly (more on this later), we also updated constructor IDs since they are needed later in the question.

   A question that we encountered along the way is "What is performance?". We narrowed this down to a couple of things: points for drivers, and wins for constructors.
c. Matching expectations and data: N/A

Exploratory data analysis

a. Setting expectations
   All the data we have will be available. Each row will contain (driverId, constructorId, points, wins). (hint: this did not happen :))
b. Collecting data, questions, or results
   Our dataset seems to be normalized similarly to how a relational database works. The data we need spans multiple files: engine_manufacturers_processed.csv, races.csv, constructors.csv, constructor_standings.csv, and engine_manufacturers_processed.csv. Each of those files contains a part of the data we need and a foreign key to common data.

   We begin by filtering races such that we only have races in the year range (2010-2020) since this is the common table between drivers and their results / constructors and their results.

   We then do the aforementioned preprocessing on constructor names and IDs to match their new names. This is to prevent confusion when interpreting the results.

We then do a group of joins on races, constructor_standings, engine_manufacturers, and constructors so that we have all the data we need to figure out for a certain race whether a team was making their own engine or not.

Something to note is the wins and points per season are accumulative in our dataset. This means that we need to sort the data by race id and round, then group the data by constructor id and year, then subtract each row from the one preceding it such that we have the points for each race individually.

We then group the data by year, team name, and engine manufacturer, then sum to get the total points and wins per year. We also get a list of teams that manufacture their engines which we'll use later.

c. Matching expectations and data
Our expectation was thoroughly destroyed by the data being normalized. This required a fair bit of data wrangling to get all the needed pieces of data together and actually start working on the question.

Building Models

a. Setting expectations: N/A
b. Collecting data, questions, or results.
We used a linear regression model to predict the effect of a team using an engine from a specific manufacturer on their points and wins. The following are the coefficients for each engine manufacturer for wins and points

Engine manufacturer	Points coefficient	Wins coefficient
Cosworth	-8.618574	-0.092395
Ferrari	2.247538	-0.025009
Honda	-0.960243	-0.040221
Mercedes	6.367299	0.141857
Renault	0.963980	0.015768

This shows that for points, Renault, Ferrari, and Mercedes are more likely to end up with more points. While for wins, only Renault, and Mercedes are more likely to end up with more wins.

The data seemed to back our expectations up, but to be fully sure, we needed to test our hypothesis. Since the population mean is not known, we resorted to one sample T-tests.

We had 2 hypotheses to test:

1. Does a team manufacturing their engine cause them to score more points than average?

2. Does a team manufacturing their engine cause them to score more wins than average?
Both hypotheses had $\alpha = 0.05$. The null hypothesis is that the mean points/wins scored by teams that make their own engines is not greater than the mean of all data, while the alternative hypothesis was that the mean points/wins are **greater** than the mean of all the data.

The first hypothesis test returned a p-value of $0.003 < \alpha$, which indicates that we can reject the null hypothesis and thus the mean points for teams that manufacture their engine is indeed greater than the mean of all the data.

The second hypothesis test returned a p-value of 0.048, which also indicates that we can reject the null hypothesis.

c. Matching expectations and data: N/A

Result interpretation

a. Setting expectations: N/A
b. Collecting data, questions, or results
As seen in the previous section, our expectation matches the results we got out of our data. Manufacturing your own data indeed influences your chances of scoring more points or wins positively. Although the influence is more positively towards points than wins.
c. Matching expectations and data: N/A

Communicating Results

See:
- K&I: Does a team being the manufacturer of their engine affect performance?
- FF&R: Does a team being the manufacturer of their engine affect performance?

Does having a race in your country affect performance?

Stating and refining the question

a. Setting expectations.
At most, having a race in your country should increase your chances of scoring more points/wins, but not by a significant margin.
b. Collecting data, questions, or results.
Our dataset contains nationalities for drivers and constructors, but contains countries for races. So we need to figure out a way to map nationalities to countries or vice versa. This was eventually done through demonyms (nationality-to-country map).

Similar to the previous question, the definition of "performance" here is a bit ambiguous. For this question, we'll use points for drivers and wins for constructors.
c. Matching expectations and data: N/A

Exploratory data analysis

a. Setting expectations.
   We expect to see a difference in performance between drivers and teams racing inside or outside their country, albeit small.
b. Collecting data, questions, or results.

Driver Points

We first need to perform some preprocessing to prepare the data and gather it all into one table. We first begin by changing the nationalities of some drivers that have ‘American-Italian’ or ‘East German’ as their nationalities. For the American Italians, we manually inspected their racing careers and saw that they competed with Italian teams only. While East German drivers were simply changed to German.

Afterwards, we use our demonyms table to map the nationality of each driver to a country. We then do a bunch of joins such that we get each driver along with their race results. Similarly to the previous question, we need to subtract driver points since they are accumulative. This nets us about 33K rows.

We then filter the data to get a list of drivers that have race results both outside and inside their home country. We then group the results by driver ID and whether the race results were in their home country or not, then we compute the mean for each of the two groups per driver, we then remove any drivers that don’t have results remaining at home **and** outside home, or drivers with zero mean points at home or away, this is to reduce noise and aid in interpreting the data. This reduces the number of rows down to about 340.

After inspecting the results we have for each driver grouped by the home status, the data seems to contain a fair bit of outliers on the low and high ends, so do an IQR filtering on the home and away results separately, then again remove drivers who don’t have home **and** away results remaining. This reduces our row count to about 300. The plot below shows only the first 30.

It’s still not apparent that home scores are noticeably higher than away scores. To better inspect this, we compute the mean for each group:

- Home mean: 0.425
- Away mean: 0.504

Constructor Wins

As for the constructors, we do similar preprocessing and joining to get the results of each constructor for each race. We additionally convert nationalities to countries similar to how we did for drivers. Afterwards, we filter the data to get only the teams that race results at home and away. Since wins are also accumulative, we need to group by years and constructor IDs, before subtracting rows to get the wins per race. Then we compute the mean wins for each constructor for home and away races. Finally, we remove all constructors that achieved zero wins at home and away.

Quickly inspecting the data, we can see that not all teams have points at home, but for the ones that do, they appear to achieve more wins at home than away. With the mean wins at home being 0.169, while the mean away is 0.108.

c. Matching expectations and data: N/A

Building Models

Driver Points

Finally, to test our hypothesis, we perform a two sample T-test with $\alpha = 0.05$ on the set of home points and away points. Our hypothesis are a s follows:

- $H_0 = Having\ a\ race\ at\ home\ does\ not\ affect\ the\ points\ a\ driver\ scores\ on\ average$,
- $H_1 = Having\ a\ race\ at\ home\ results\ in\ more\ points\ for\ a\ driver\ on\ average$,

The hypothesis test returned a p-value of 0.875, which means that **we cannot reject the null hypothesis, and thus it is likely that having a race in your country does not affect your points.**

Constructor Wins

We already hypothesized that according to the means computed in the previous section, teams competing at home achieve more wins on average VS teams competing away from home. Similar to driver points. Formally, our hypothesis are as follows ($\alpha = 0.05$):

- $H_0 = Having\ a\ race\ at\ home\ does\ not\ affect\ the\ wins\ a\ \ constructor\ achieves\ on\ average$
- $H_1 = Having\ a\ race\ at\ home\ results\ in\ more\ wins\ for\ a\ constructor\ on\ average$

After performing a two sample T-test on the set of wins at home and wins away from home, the p-value was computed at 0.0296, which means that **we can reject the null hypothesis and conclude that when competing at home, a constructor is more likely to achieve more wins on average.**

Result interpretation

Contrary to our initial expectation, the number of points achieved on average by a driver competing in their own country is not necessarily greater than a driver competing away from home. While for constructors, the number of wins on average increases when competing in their home country.

Communicating Results

See

- K&I: Does having a race in your country affect performance?
- FF&R: Does having a race in your country affect performance?

Given that Mercedes drivers have above average top speed, does this apply to all German teams?

Stating and refining the question

a. Setting expectations.
   Mercedes is one of the teams with the most points and wins, so it's not completely out of the question that they would have above average top speed. However, there is no reason that all German teams would follow this.
b. Collecting data, questions, or results.
   The data we need for this question is a table that has constructorId, nationality, and top speed. Our dataset contains this data but it's scattered over multiple tables.
c. Matching expectations and data: N/A

Exploratory data analysis

a. Setting expectations.
   We expect that we'll have data for a few teams including Mercedes and a few other German teams, and that Mercedes is quite likely to have a mean fastest lap speed greater than the mean.
b. Collecting data, questions, or results.
   After inspecting the data, it seems like a lot of rows don't actually contain '\N' as fastestLapTime values, after consulting the source of the data (ergast), it seems like fastestLapTime values are only present for races starting 2004. The rows containing these values were filtered out.

   We store the list of German constructors, as we'll need that later on. We then compute the mean fastest lap speed value for each constructor. To get the mean of all constructors/results, we compute the mean over all constructors again.

   Looking at the remaining data, we found out that constructor ID 131 (Mercedes) indeed surpassed the mean fastest lap speed. (205.51 vs 202.58)

   Next, we filter the mean fastest lap speed values for german constructors using the list of german constructors that we stored earlier. This results in only two German constructors with IDs 131 and 2 which correspond to Mercedes and BMW Sauber. They have mean fastest lap speeds of 205.51 and 205.013 respectively.

c. Matching expectations and data.
   As seen in the previous section, our expectation is matched in that Mercedes is indeed above the mean fastest lap speed, the only other German team (ID = 2) also exceeding the mean.

Building Models

To test if our hypothesis is statistically significant, we'll conduct a one sample T-test. Our hypotheses are as follows, $\alpha = 0.05$:

- $H_o = The\ mean\ fastest\ lap\ speed\ of\ German\ teams\ is\ no\ different\ than\ the\ mean.$
- $H_1 = The\ mean\ fastest\ lap\ speed\ of\ German\ teams\ is\ greater\ than\ the\ mean.$

The p-value returned from the T-test is $0.0296 < \alpha$, which means that we can reject H_o and thus the mean fastest lap speed for German teams is greater than the mean.

Result interpretation

Given the results from the previous two sections, we can see that Mercedes has a mean fastest lap speed that's over the mean of all teams, and that other German teams do follow.

Communicating Results

- K&I: Given that Mercedes drivers have above average top speed, does this apply to all German teams?
- FF&R: Given that Mercedes drivers have above average top speed, does this apply to all German teams?

Best number of pitstops for a circuit?

Stating question

- Number of pitstops made in a race is a crucial aspect in the race that greatly affects the position of the driver.
- At first we wanted to analyze how the length of the track and its weather condition on race time (each track has a specific margin of time in the year to race on for example bahrain race must be during the winter)
- After researching we found out that we needed the following things: weather conditions, how different tyre compound wear out for each race, and how the downforce of vehicles affect the tyres also
- Sadly downforce of vehicles and its effect on tyres was confidential, and how tyres are affected by race conditions was really hard to quantitate it and measure it with a factor as tyre types always changes each season and even may change mid-season which made our analysis really hard

Refining question

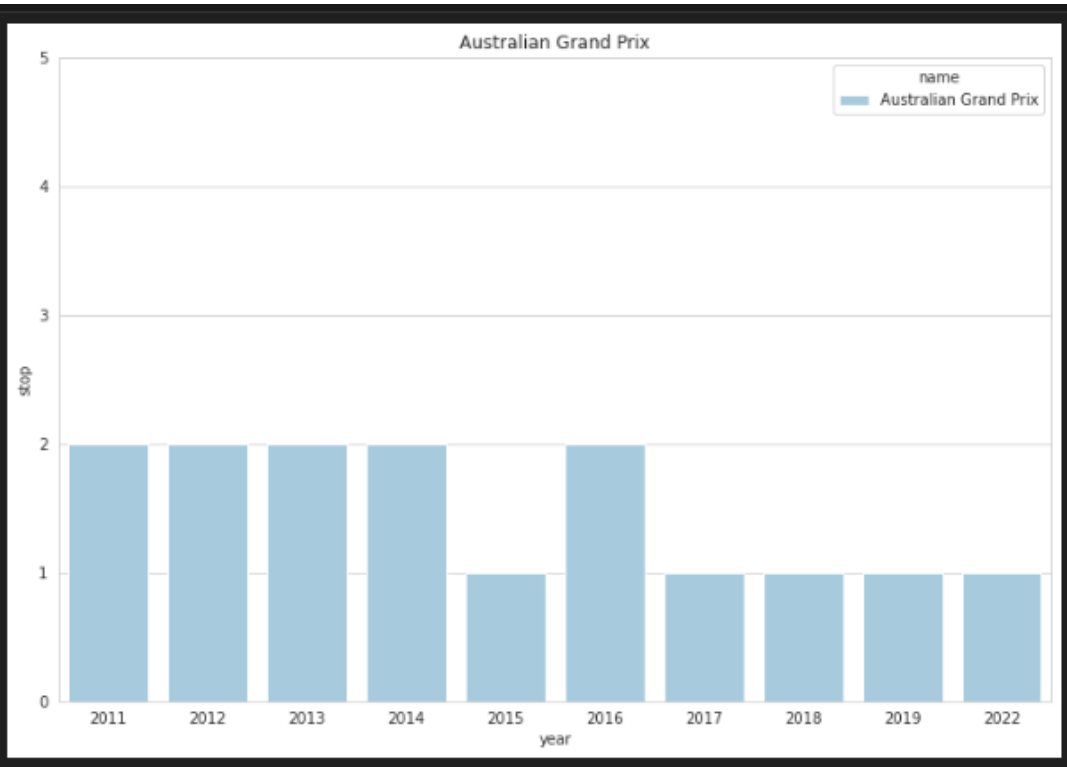
1. We had to figure out from the data we have how to solve our question especially after making sure that no data from the internet can help us
2. Our new assumption became that the number of pitstops of the winning driver of the race is the optimum number of pitstops

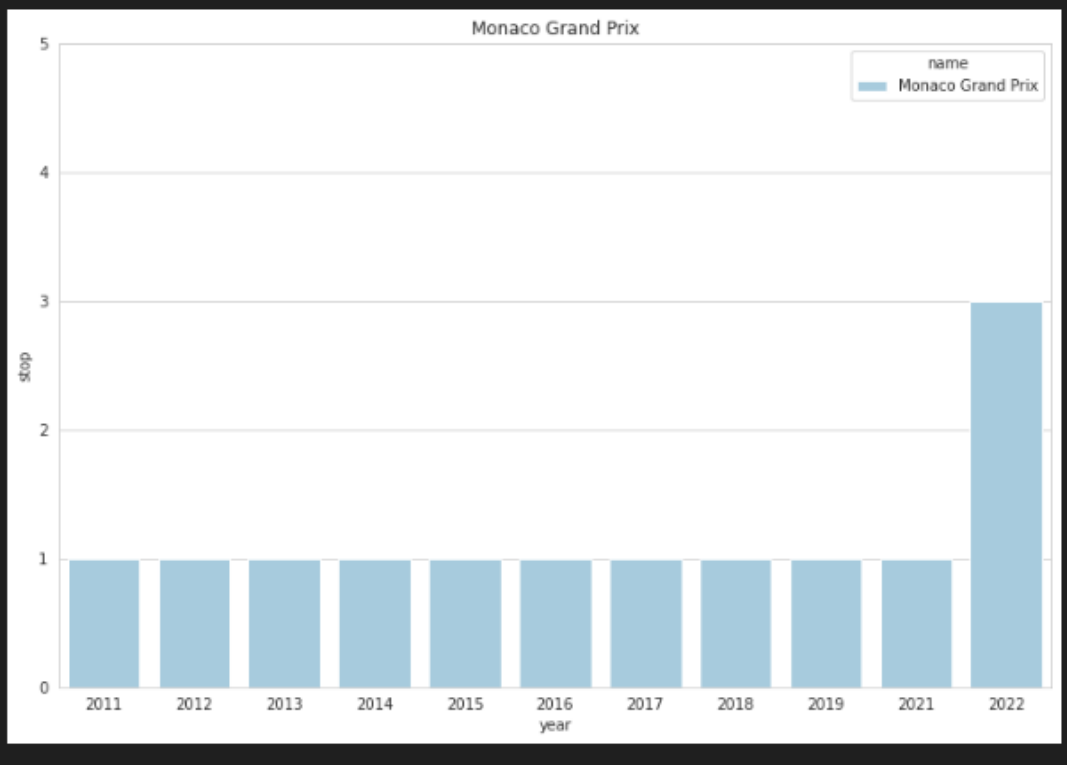
Exploratory Analysis

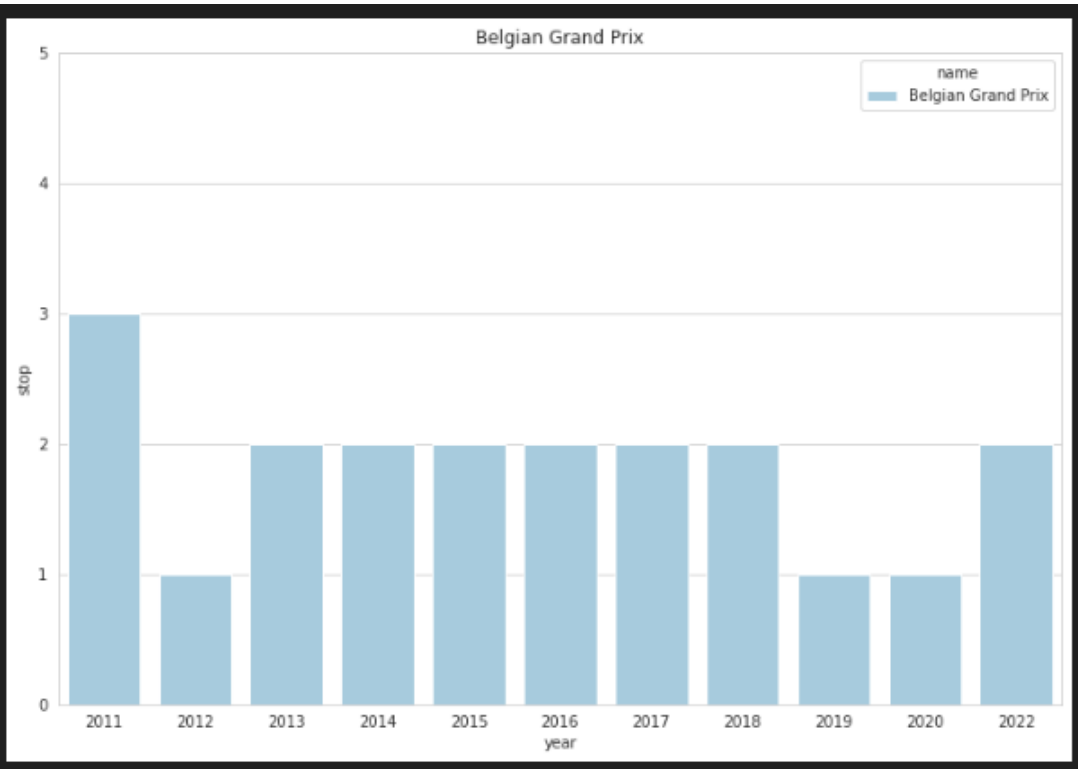
- We found out the pitstops was banned in 2005 and after that until 2010 it consisted of a pitstop and fuel change so we only took races starting from 2011
- We joined results_csv on pit_stops_csv on drivers_csv on _races_csv
- So now for each race we get the result of the driver and number of pitstops he did
- We then filtered out the drivers who did not win and only took the race winner
- We had the number of pitstops of the race winner for each year from 2011 to 2022
- There were races that had more than 4 stops and after research we found out that this happens when the race had to be stopped due to accidents or rain.
- For rain drivers had to pit and put on harder tyres and when the weather returns back normal drivers come back and put on softer tyres so rainy races tend to have more pitstops and this is really rare (like Canadian GP 2011)

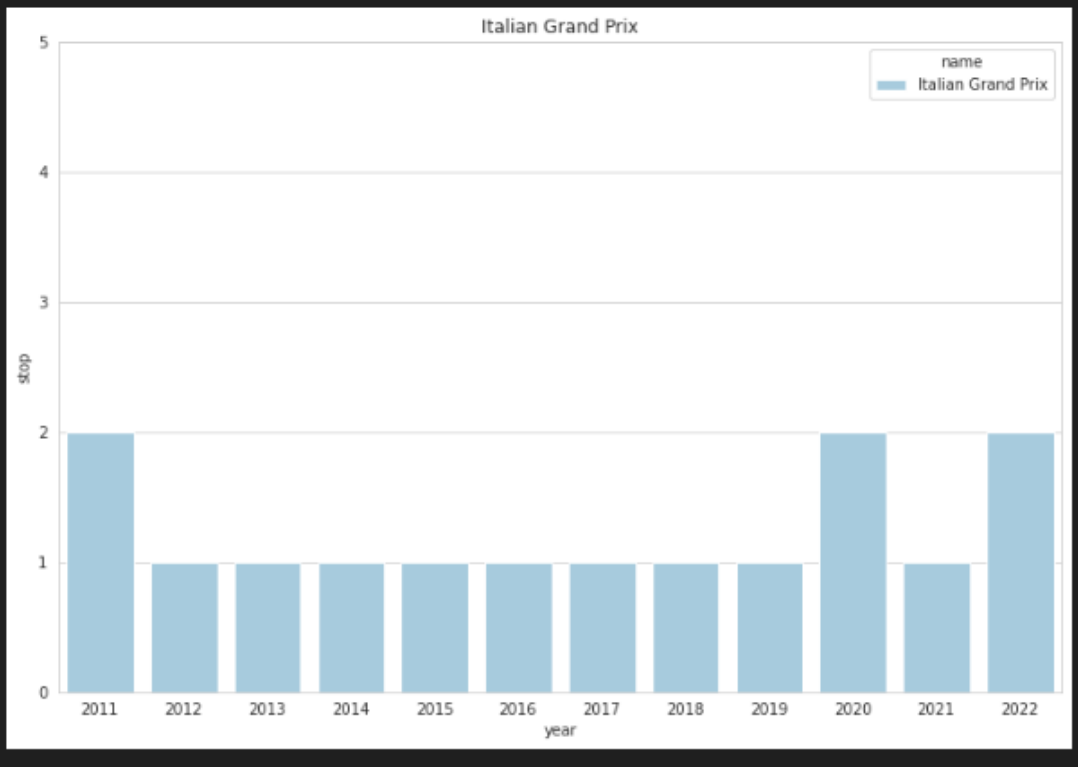
Insights Gained

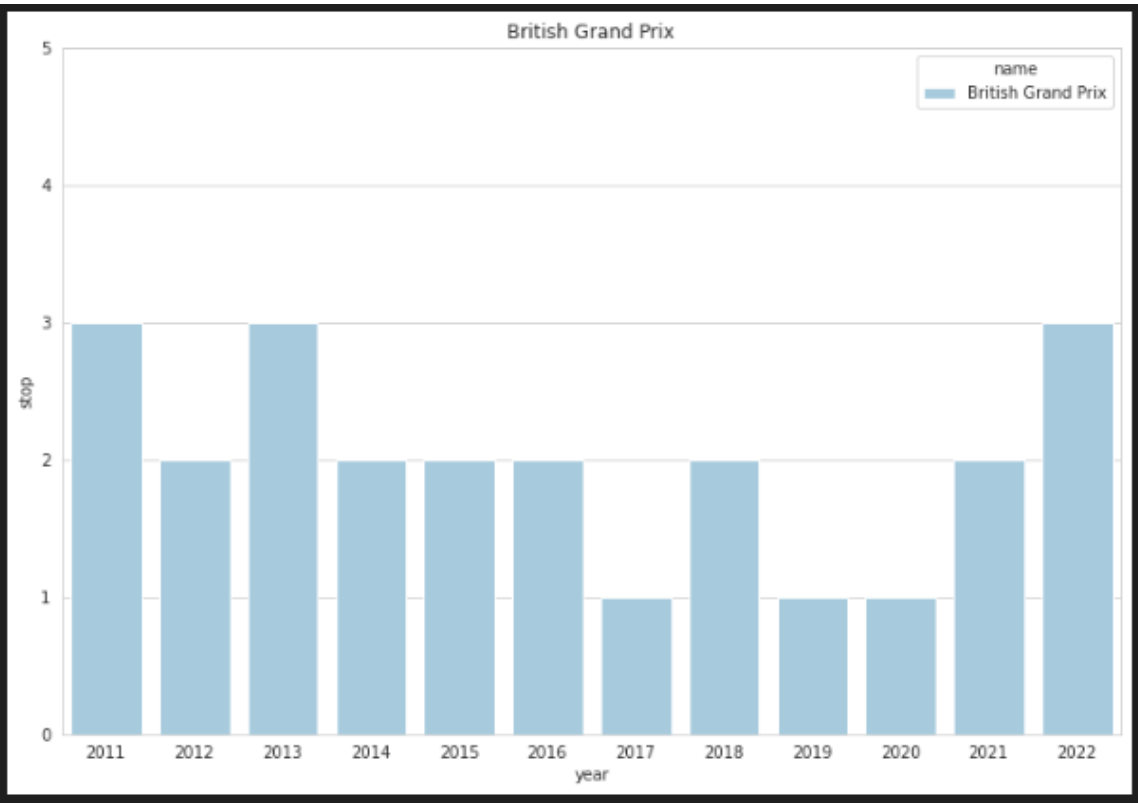
- We plotted the number of pitstops of the most famous races that must be held each season and that was the output, the plots show that almost the number of pistols don’t change :

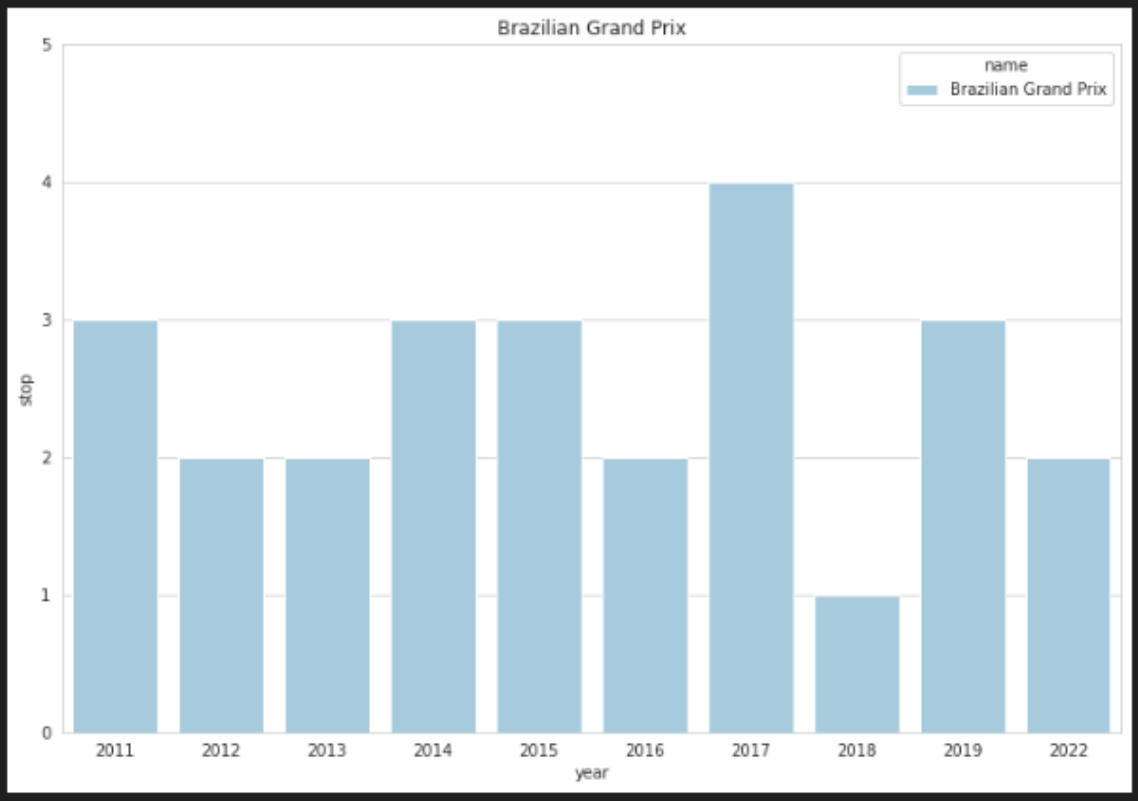

Australian Grand Prix
name
Australian Grand Prix
stop
year
0
1
2
3
4
5
2011
2012
2013
2014
2015
2016
2017
2018
2019
2022


Monaco Grand Prix
name
Monaco Grand Prix
stop
year
0
1
2
3
4
5
2011
2012
2013
2014
2015
2016
2017
2018
2019
2021
2022

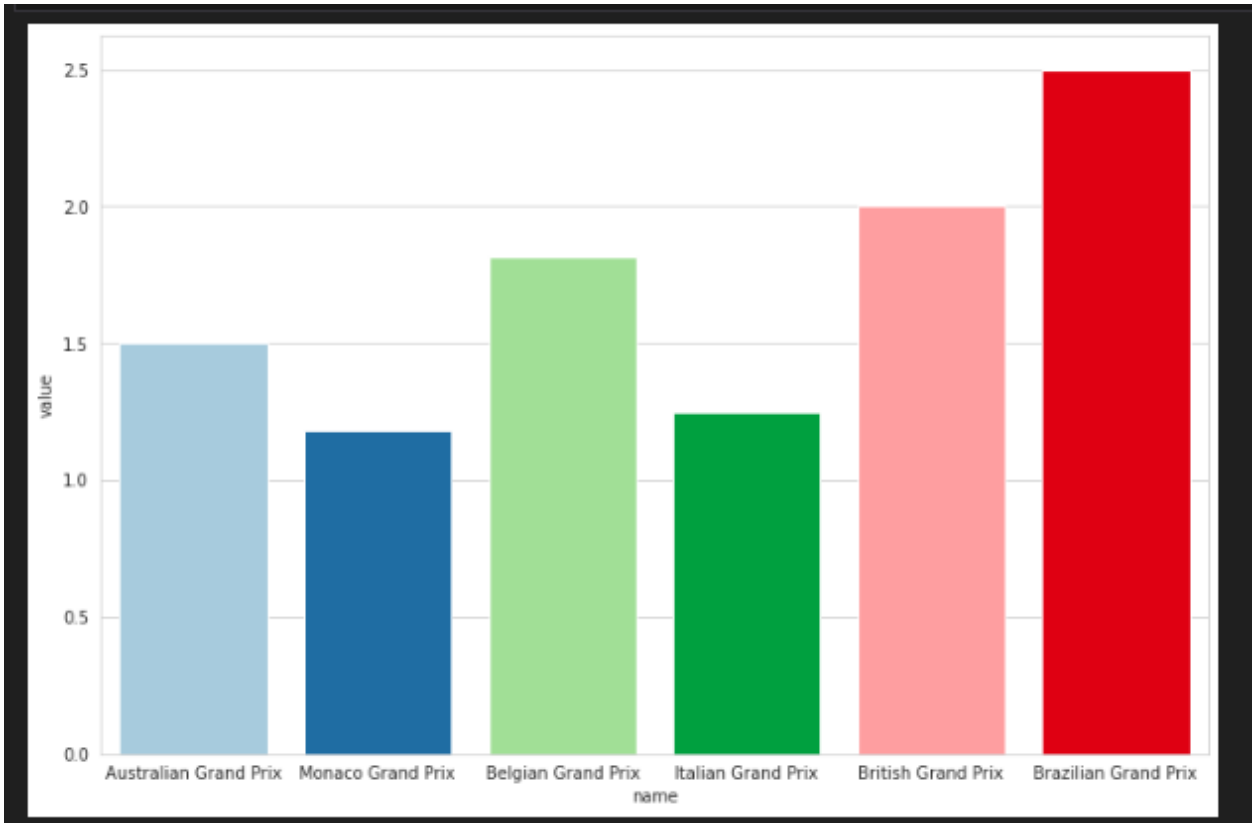

Belgian Grand Prix
name
Belgian Grand Prix
stop
year
0
1
2
3
4
5
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2022


Italian Grand Prix
name
Italian Grand Prix
stop
year
0
1
2
3
4
5
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022


British Grand Prix
name
British Grand Prix
stop
year
0
1
2
3
4
5
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022


Brazilian Grand Prix
name
Brazilian Grand Prix
stop
year
0
1
2
3
4
5
2011
2012
2013
2014
2015
2016
2017
2018
2019
2022

And the avg position of these races



The last graph shows that the Brazilian grand prix is the most crucial track in the is the BRazilian grand prix so if a driver started losing time to other drivers most probably changing the tyres would be beneficial, which makes sense really as brazil is fairly warm in November(race time) and warmer weather makes tyres wear out faster

Future Work :

- Having more data from races about the tyres wear out coefficients and parameters about the downforce effect on tyres would be better
- FIA gives more data about tyres across seasons and how they change year to year
- Having the data stated available would allow hobbyists to analyze more the effect of each track on tyres and get better insights to determine the best pitstop for each race

Driver Average Position Across Years

Goal of the question

- Having the average position of the driver shows us his real performance and can show us also the performance of the driver's vehicle if he was good at a season and then started to degrade

Data Handling and Pre-processing:

- At first we found that we have all the data needed so we started loading it and seeing what it does have.
- In results.csv not all racers finished the race so we had to drop results that have status other than 1,11,12,13 as these statuses are the only one who finished
- Also we dropped results that finished after position 29 as we rarely have racers who finished after that.
- Also we dropped all '\\N' from the position column

Operations On Data:

- We joined results on drivers on races using (driver_id and race_id)
- Thus each column had the result of a driver in a specific race
- Then we had to make a map of all driver where the key is the driver_id and value list of all his results of all races
- We then averaged the position of each driver

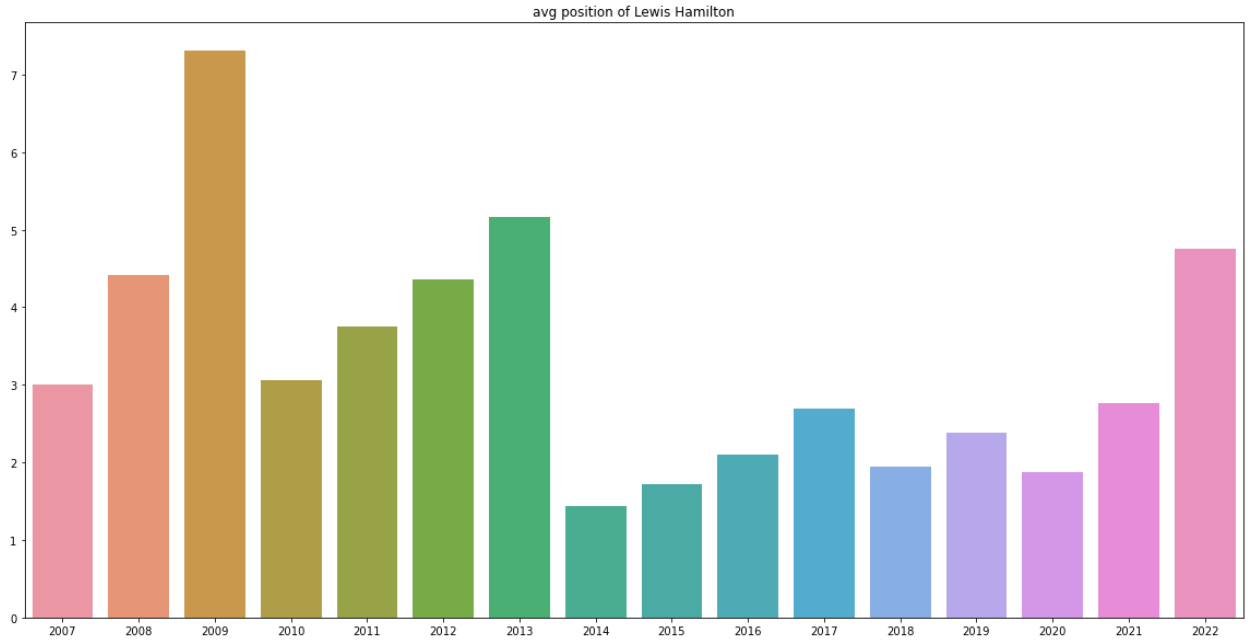
Insights:

- We could sort the avg position of all drivers to know the GOATS of F1

```
[18]  ✓ 0.1s

id=657, avg_pos=1.0, name=Bill Vukovich
id=579, avg_pos=1.7317073170731707, name=Juan Fangio
id=373, avg_pos=1.9210526315789473, name=Jim Clark
id=786, avg_pos=2.0, name=Luigi Fagioli
id=591, avg_pos=2.0, name=George Amick
id=701, avg_pos=2.0, name=Bill Holland
id=802, avg_pos=2.0, name=Dorino Serafini
id=647, avg_pos=2.1818181818181817, name=Alberto Ascari
id=102, avg_pos=2.3229166666666665, name=Ayrton Senna
id=475, avg_pos=2.46875, name=Stirling Moss
id=526, avg_pos=2.5, name=Troy Ruttman
id=642, avg_pos=2.56, name=Nino Farina
id=117, avg_pos=2.5703703703703704, name=Alain Prost
id=328, avg_pos=2.610169491525424, name=Jackie Stewart
```

- We then made a cell that take the driver id and shows his position over years



-

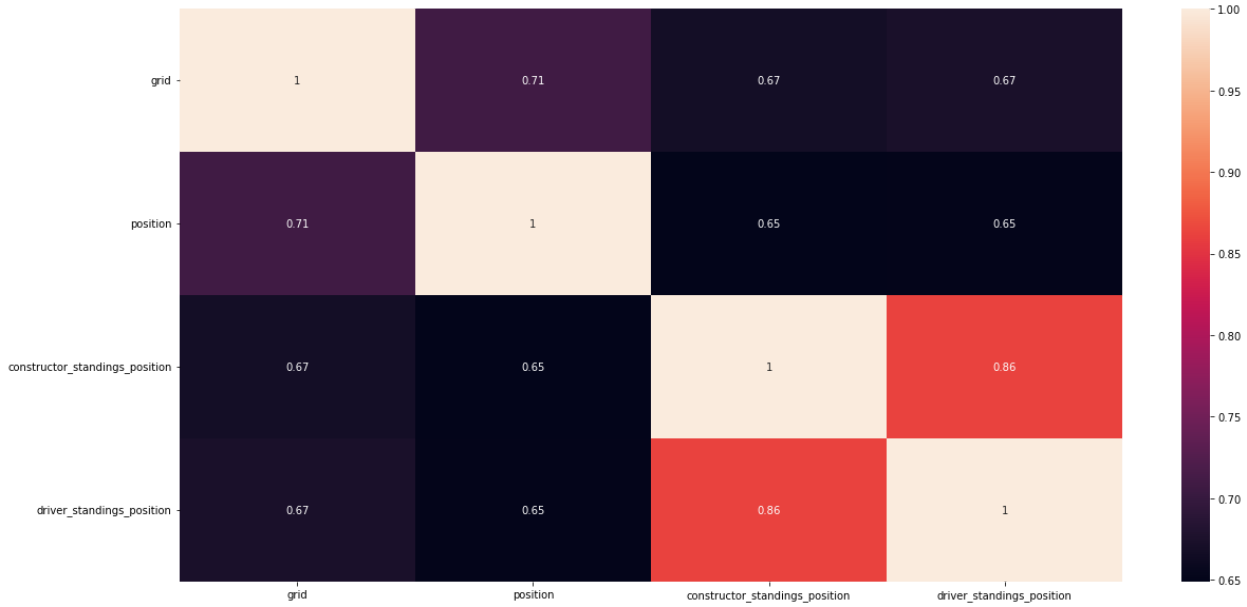
- This graph shows lewis Hamilton performance, one of the greatest drivers ever, he won 2008 and from 2014 to 2020, the graph shows that clearly and this also means that the problem was in the car he drove especially on 2011,2012,2013

Predict the position of the next race;

Exploratory data analysis

- Kept exploring the data until we found columns that highly correlate with each other which are grid_position, constructor_standing and driver standing and found that graph

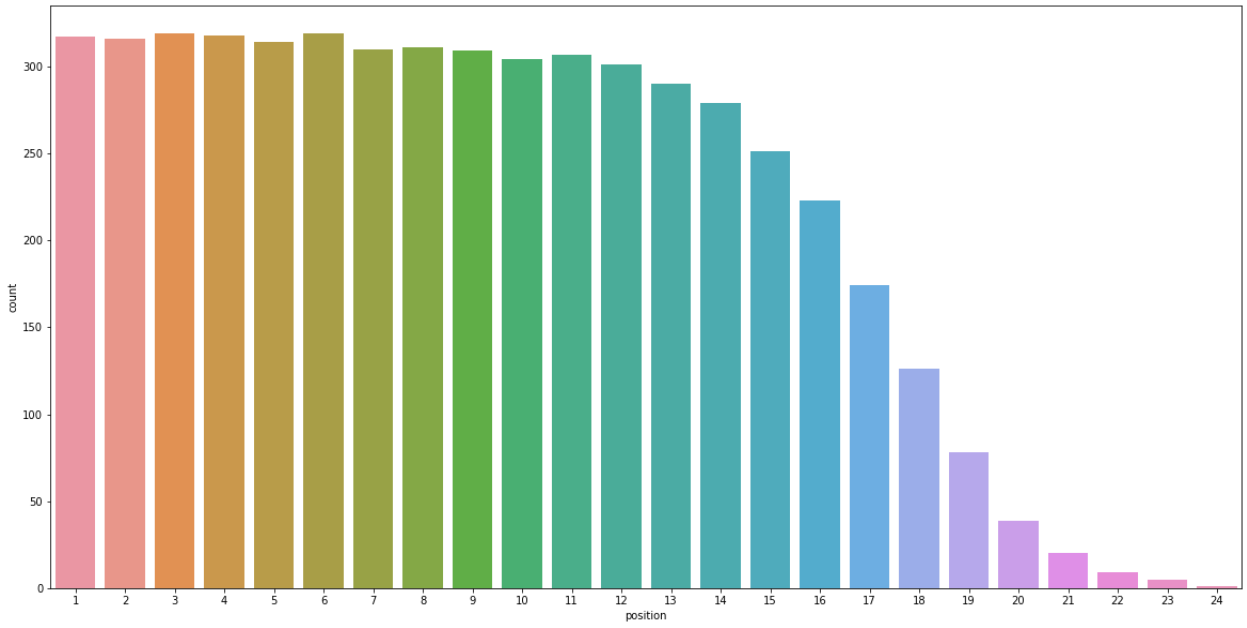
-



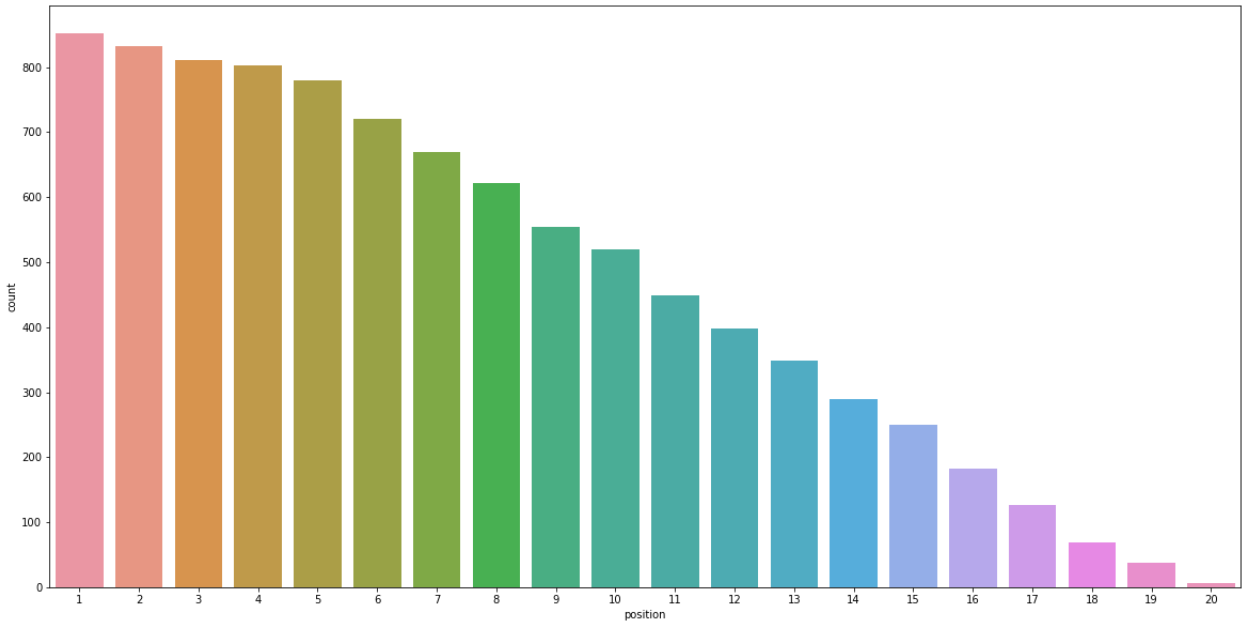
Data pre-processing

- The Constructor standing should generally lie from 1 to 10 as most seasons had only 10 constructors and grid pos, driver standing lies between 1 to 20 as most seasons had only 20 drivers
- The data did not have that so we did lots of cleansing to get that
- Position before :

-



- Position after pre-processing



Work done:

- Made the feature vector X : {grid_pos, racer_rank, constructor_rank} and y {race_result}
- Began to try different model and evaluate them to see what we can get from this problem, knowing that this problem is unique and complex as F1 race is really full of kaos and even best race analytics can't predict the race result.
- We tried the following
    - SVM model with 'rbf' kernel and we got acc around 19%
    - MLP using tensorflow and this got the best acc around '20'
    - Tried one hot encoding all the features as they are classes not continuous vars so to do that we had to implement a mlp from scratch using pytorch but sadly this did not improve also our acc
- In general we iterated a lot on training this model as we kept caleaning the data we train on and try again the model

Insights and Future Work:

- Clearly the parameters we chose affect the race reulst, but they are not enough, so this model proved that:
    - F1 is a really chaotic sport that lots of parameters affect the race result
- We may need more parameters like:
    - Aerodynamics of each car and how they get affected by wind speed for race
    - The race lenght of straights and corners and whether each vehicle favor corenrs or straights

    - How the racing style of the driver
- In general our model isn't that bad, random choice is 5% while our model is 20% acc, which can help the teams alot in putting exectaitons and race decisions.

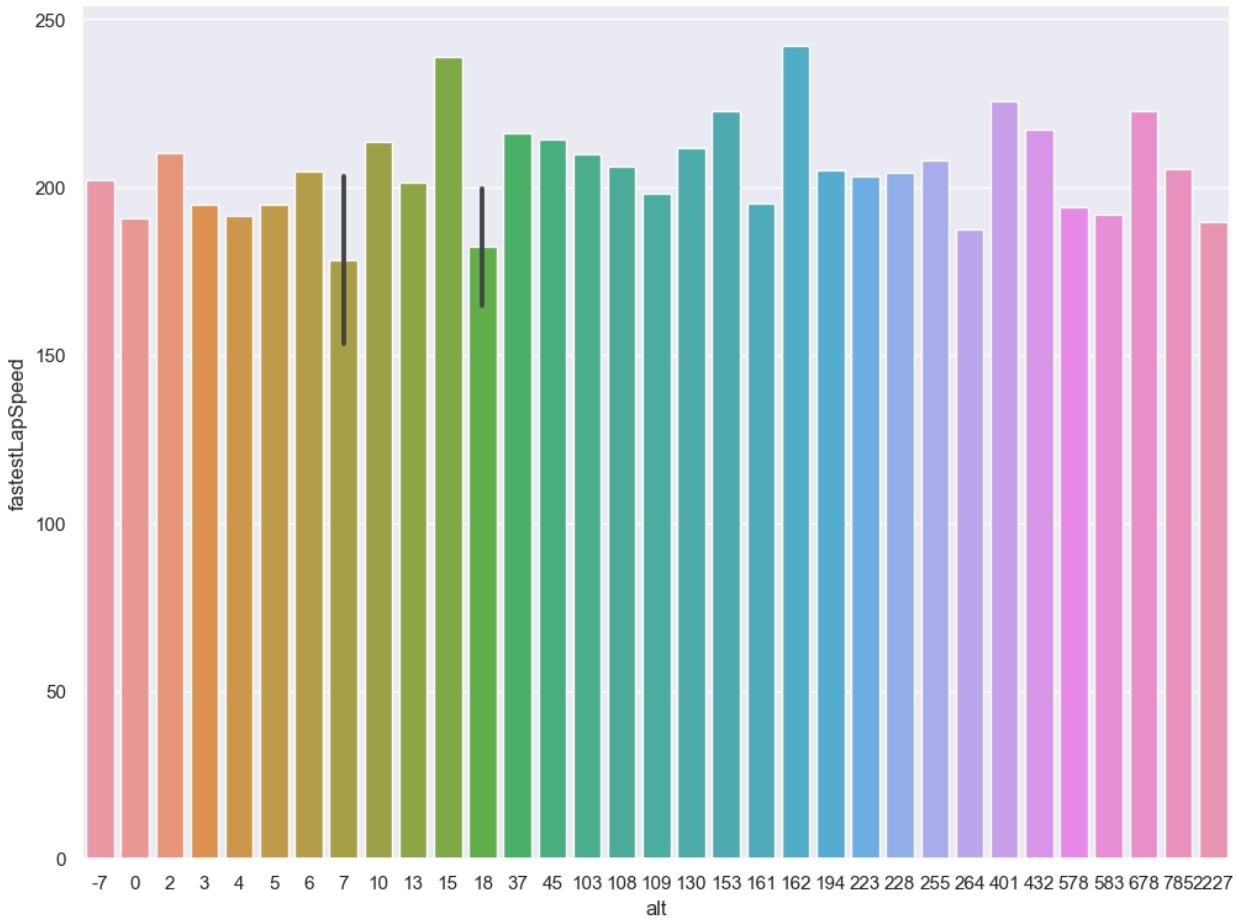
How does a track's altitude affect top speed / average lap time?

Exploratory data analysis

a. Setting Expectations
   It's believed by F1 commentators that the higher the altitude of the circuit the less oxygen there is, so the engine would not burn enough fuel.

b. Collecting data, questions or results
   I used the tables: circuits, races, results
   Then checked the expectation by calculating the average speed of each circuit and seeing the correlation with altitude.
   I merged the tables, cleaned them, then grouped by the circuit id and the altitude, then calculated the mean of the speeds

c. Matching expectations and data
   You can see there is no correlation, as the altitude increases, the speed doesn't necessarily decrease



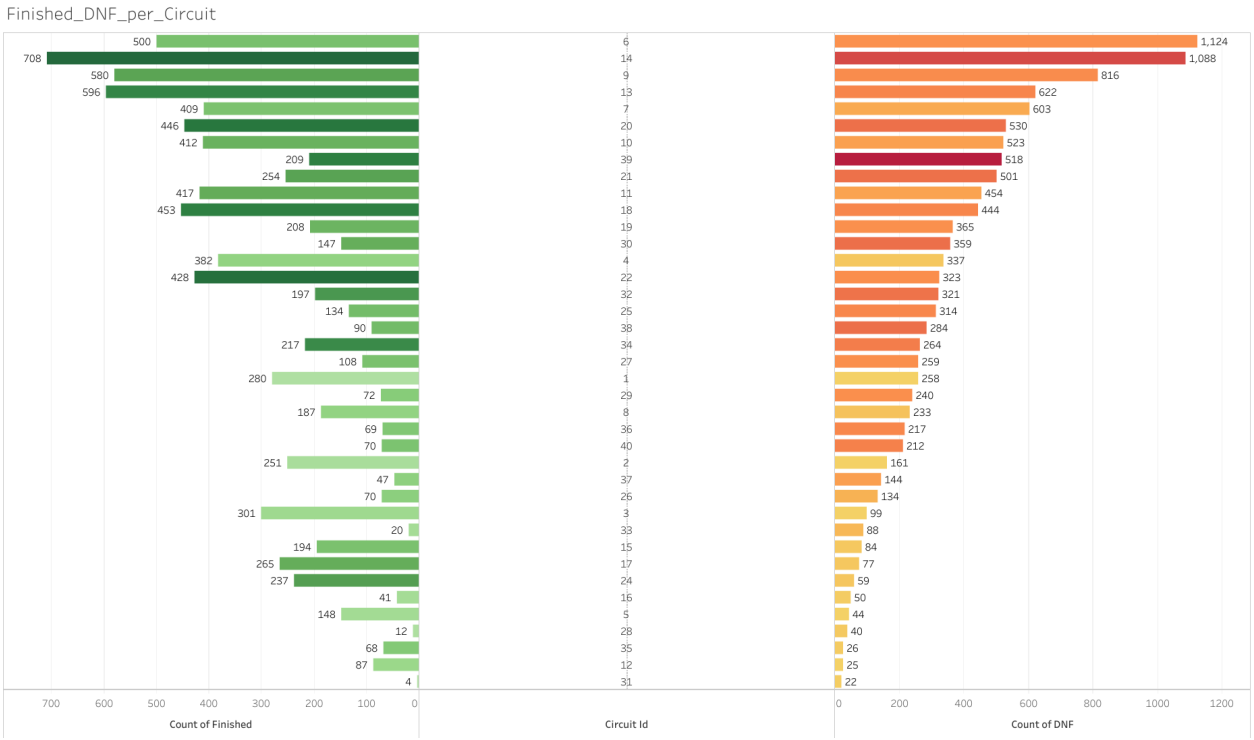
Which tracks have the most DNFs?

Stating question

Which tracks have the most number of drivers who did not finish the race
See which tracks are brutal for the cars

Result interpretation

Circuit 6 is the most with Did Not Finish drivers, then 14 and 9

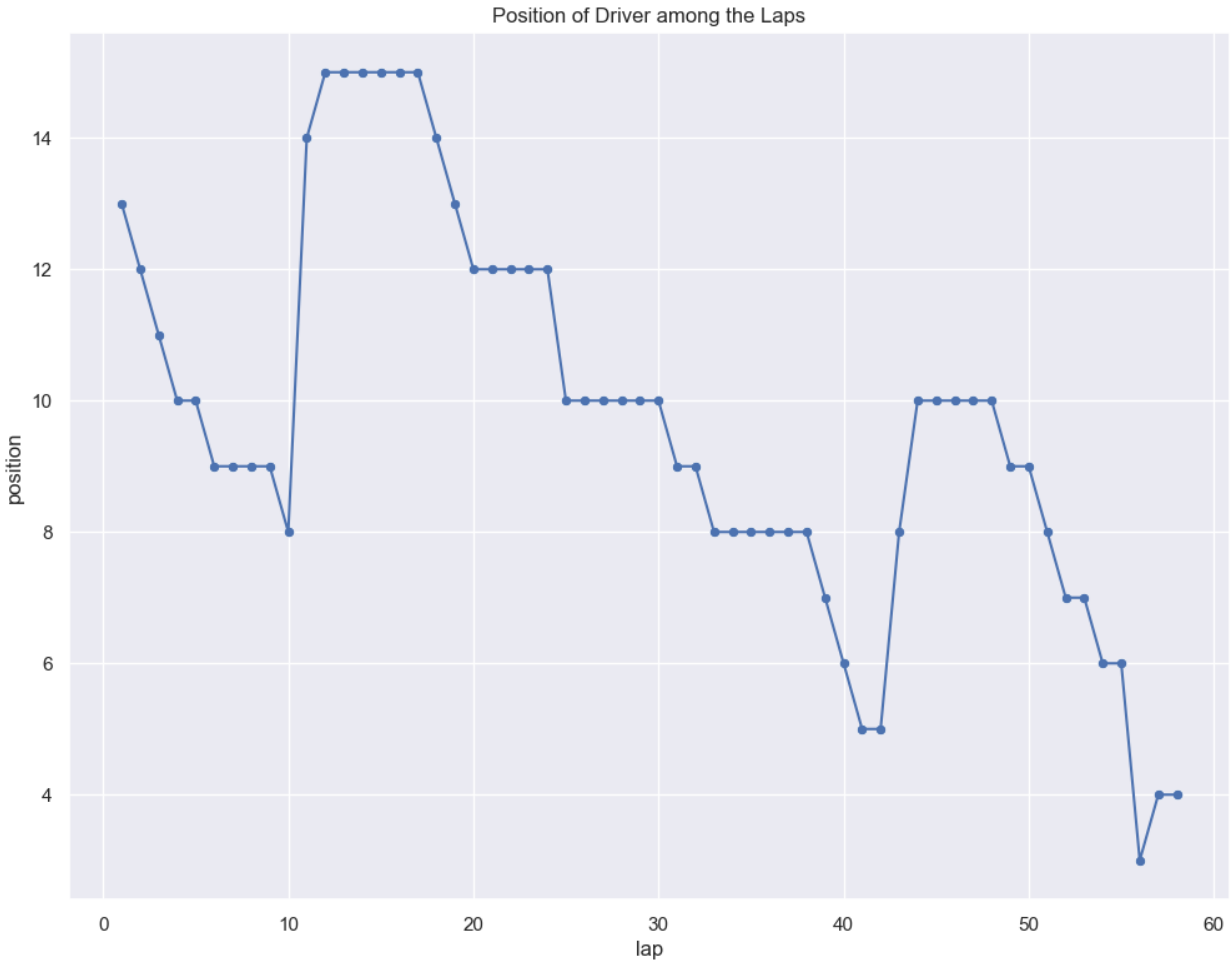


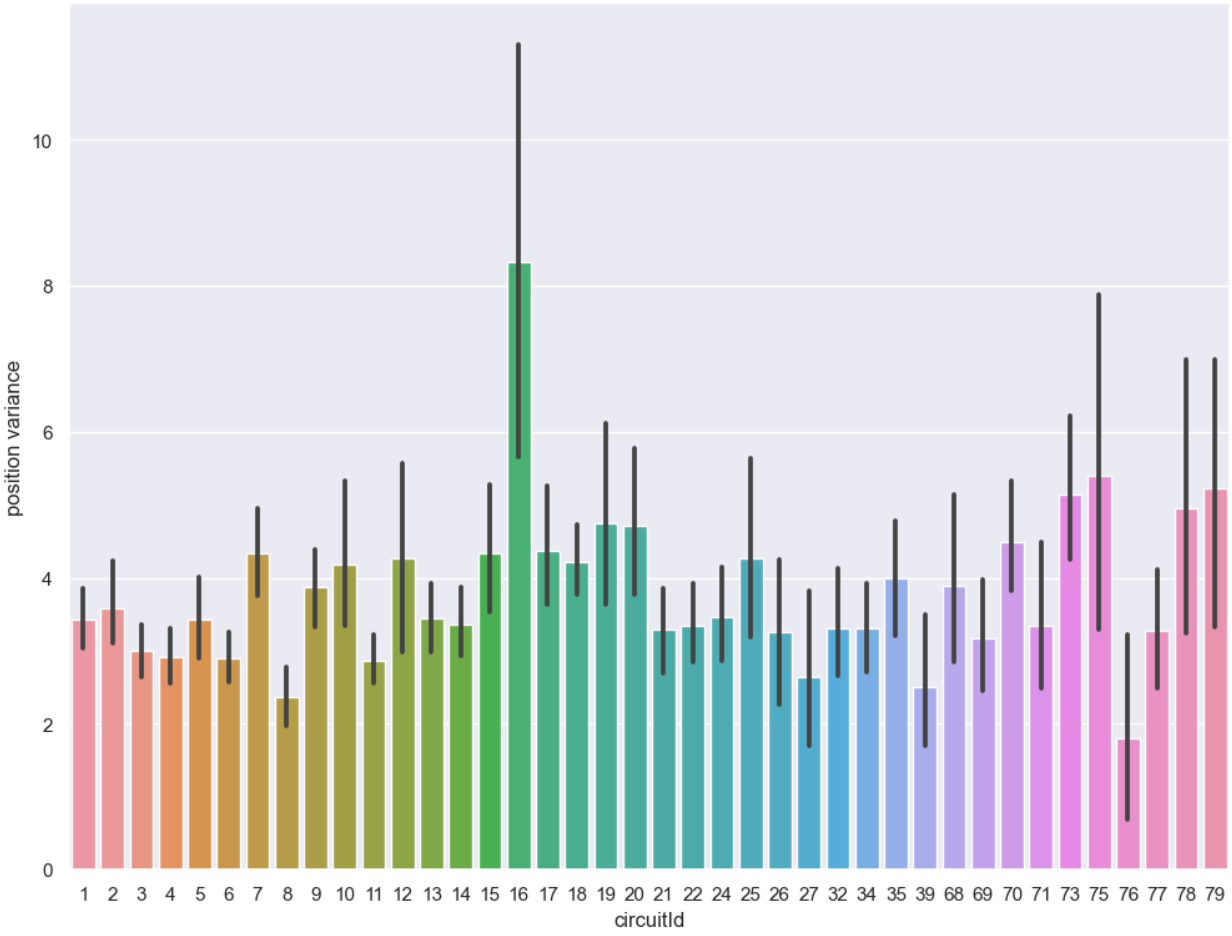
Which tracks favor overtaking?

Stating and refining the question

- Analyze which tracks are easier to overtake cars in, which will really help the drivers choose their pitstop strategies, thus they will have shorter stints but faster pace.
- This is decided by seeing the start position of each driver and the end position at the end of the race, the track will favor overtaking if the positions of the drivers changed a lot from their position at the first of the race

Result interpretation




position variance
0
2
4
6
8
10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 25 26 27 32 34 35 39 68 69 70 71 73 75 76 77 78 79
circuitId

What is the average retirement Age?

Stating and refining the question

a. Setting expectations.
The average retirement age should be in mid thirties, since most drivers who leave earlier are usually switching sports, like going to NASCAR, and in the last 30 years not so many drivers were able to compete in their early forties.

b. Collecting data, questions, or results.
The data needed should contain the driverId, driver dob and his retirement year, this data can be extracted from our dataset.

c. Matching expectations and data:
The results turned out to match the data with the average retirement age being 36.

Exploratory data analysis

a. Setting expectations.
We expect that drivers generally retire after 29 years.

b. Collecting data, questions, or results.
The drivers driverId and dob are already present in the data, what remains is the retirement age.

Retirement age:
We join the following data: the driverId,raceId and race year, on the driverId, then we group them by driverId, and reduce the resulting groups by taking the max race year.

We now have the year of the last race the driver participated in, subtracting the driver's dob would yield his retirement age. However, we must make sure that the driver really retired, so we filter out drivers whose last race was before they were 29, they are considered to have left the sport, and those who have competed in the last 3 years, since it's not clear whether they are retiring or taking a break.

c. Matching expectations and data: NA

Building Models

The average retirement age turned out to be 36 years. We tested if our hypothesis is statistically significant, using the one sample T-test.
Our hypotheses are as follows, $\alpha = 0.05$:

- $H_o = The\ average\ retirement\ age\ is\ 36.$
- $H_1 = The\ average\ retirement\ age\ is\ greater\ than\ 36.$

The p-value returned from the T-test is $0.877 > \alpha$, which means that we fail to reject H_o, i.e. The average retirement age is 36.

Age Influence on Performance

Stating and refining the question

a. Setting expectations.
   Age is expected to be an influential factor in a driver's performance in motorsports. As drivers age, they may experience declines in physical fitness, reaction time, and cognitive abilities, which can impact their ability to compete at a high level.
   Additionally, older drivers may have accumulated more racing experience, which can be an advantage in terms of strategy and decision-making.

b. Collecting data, questions, or results.
   The performance metric used is the driver's final race position.
   The data needed is the race standings for each race for each driver and the corresponding race year and the driver's final position.

c. Matching expectations and data: *NA*

Exploratory data analysis

a. Setting expectations.
   We expect the driver's performance to peak in between 24 and 35, since this is generally the time where drivers win their F1 titles. Many successful F1 drivers have won their championships during this age range, including Lewis Hamilton, Sebastian Vettel.

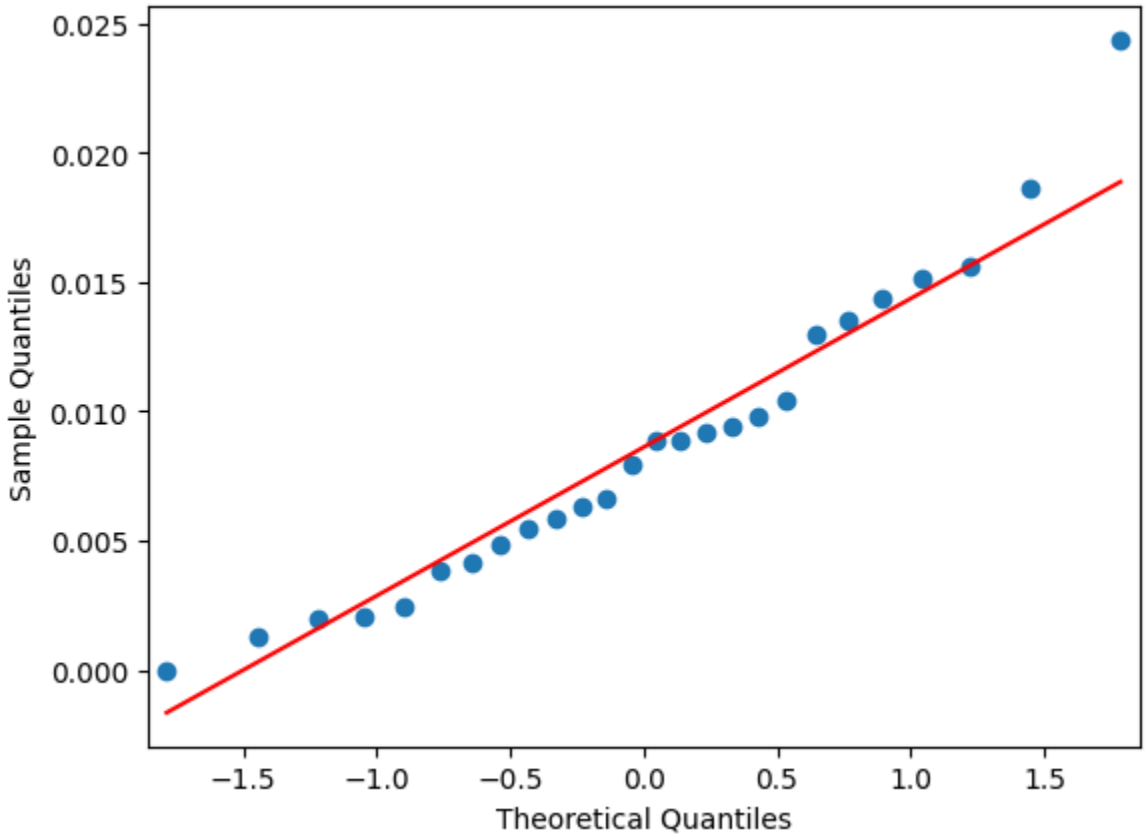
   This suggests a normal distribution pattern, where drivers before and after the age range are less performing than those inside the range, with the peak at the mean.

b. Collecting data, questions, or results.
   The data was collected by joining the driverId, driver's dob, raceId, race year, and driver's final position for the race with raceId, then adding an age column as the difference between the race year and the driver's dob, then grouping on the age, and taking the average position for each age group.
   To account for technological differences due to advancements in the automobile industry, we dropped out results before 2000.

c. Matching expectations and data. *NA*

Building Models

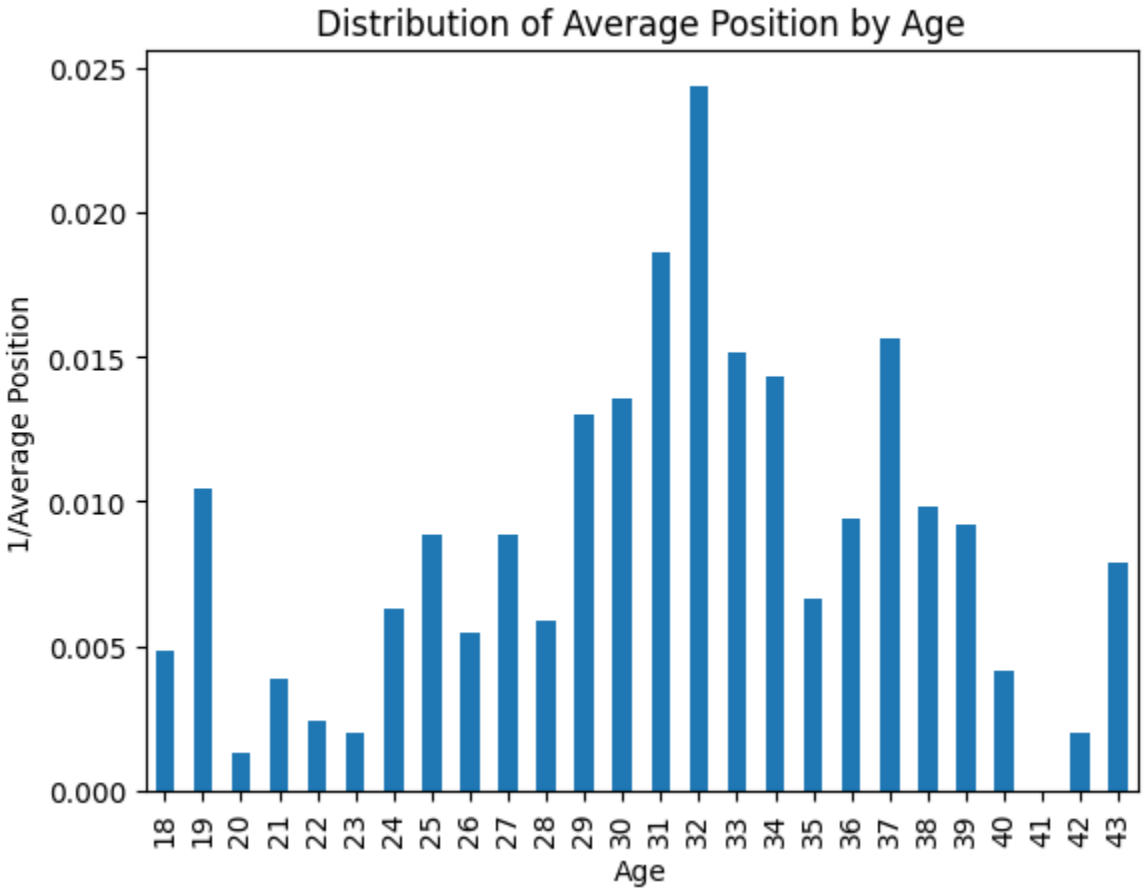
To test the normality hypothesis, we used the Q-Q plot. It confirmed our assumption.



Result interpretation

The results show that drivers achieve their peak performance in their early thirties, and better results at late thirties than early twenties.

Communicating Results



The performance is best in the age range 30 to 34, with the right hand side showing better performance, experience beats youth, even though youth can be an advantage in terms of physical fitness and reflexes, experience can give older drivers an edge in terms of racecraft, strategy, and decision-making.

Time series prediction of drivers' performance

Stating and refining the question

a. Setting expectations.
   We expect a weak correlation between the driver's performance across 30 races, taken in sequential order across time.
   We try to predict the driver's position in the next 5 races based on the previous standings of 25 races.
b. Collecting data, questions, or results.
   The data should have, for each driver in the model, a list of races positions in about 30 races. This is hard to achieve, since it's easy to find a 30-race streak of consistent performance, usually drivers are physically exhausted.

c. Matching expectations and data: *NA*

Exploratory data analysis

a. Setting expectations. *NA*
b. Collecting data, questions, or results.
   The data was collected by considering drivers who have participated in at least 20 races and at most 30, then each driver's vector was chopped at length 30, those with less than 30 races were interpolated, vector represents the race positions for some driver in 30 consecutive races.

   The problem can be formulated as a classification problem with 20 classes (possible positions in a race).

c. Matching expectations and data. *NA*

Building Models

We built a RacePredictor neural network, to predict the finishing position of a driver. It consists of 3 fully connected layers, each with 25 input features.

The first one applies a ReLU activation and has 64 output, the second one has 32 and the last has 5, the output vector, the predicted results for the next 5 races (training data).

The output is passed through a sigmoid layer, since this is a classification task, then multiply by 19 and add 1 to rectify the original classes.

Result interpretation

1. The model is not good at predicting the trend of the time series
2. As the racers' joint performance does not follow a trend, it is hard to predict the performance of the racer.

3. Each driver has a different performance, so it is hard to predict the performance of all drivers using only one model.
4. We can fine tune the model using the data of a single driver to predict the performance of that driver
5. So for each driver, we should fine tune the model for him to get a good prediction

Communicating Results

1. Before tuning for each driver:
   [3. 2. 2. 3. 4.]
   [11. 7. 6. 2. 2.]
   ==========
   [17. 8. 12. 16. 17.]
   [17. 17. 13. 19. 15.]
   ==========
   [8. 6. 11. 9. 15.]
   [17. 6. 16. 6. 8.]
   ==========
   [13. 15. 14. 17. 17.]
   [13. 9. 9. 14. 12.]

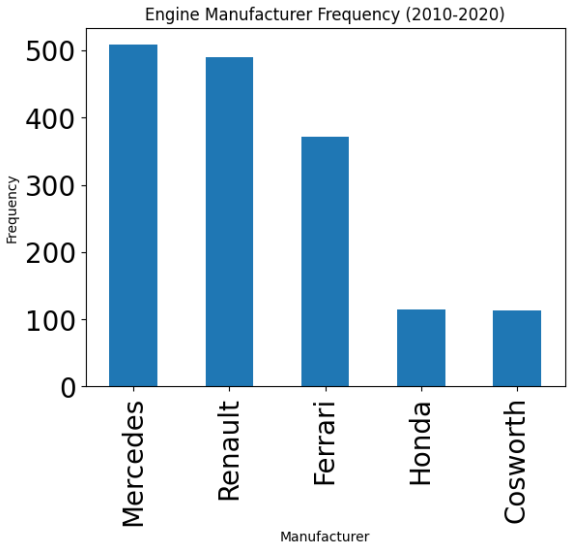
2. After tuning for Lewis Hamilton:
   predictes_positions: [2. 1. 2. 2. 2.]
   Ground truth: [2. 1. 1. 2. 1.]

   This is just a proof of concept of the fine tuning idea.

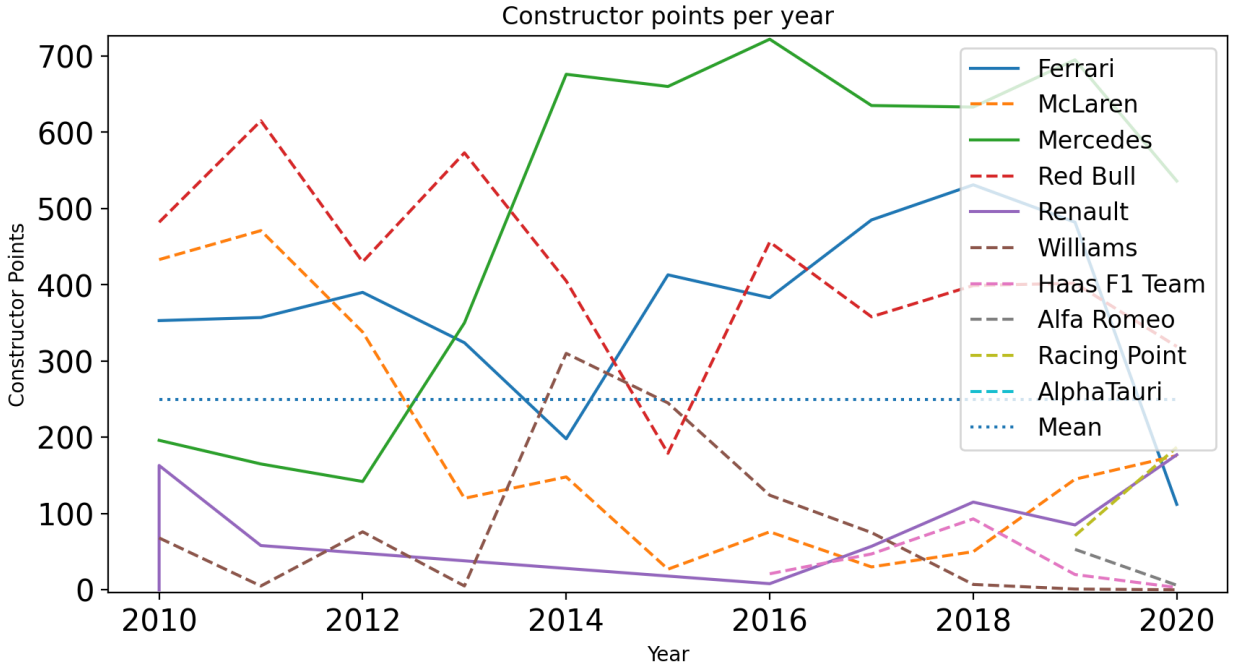
3. Knowledge and insights

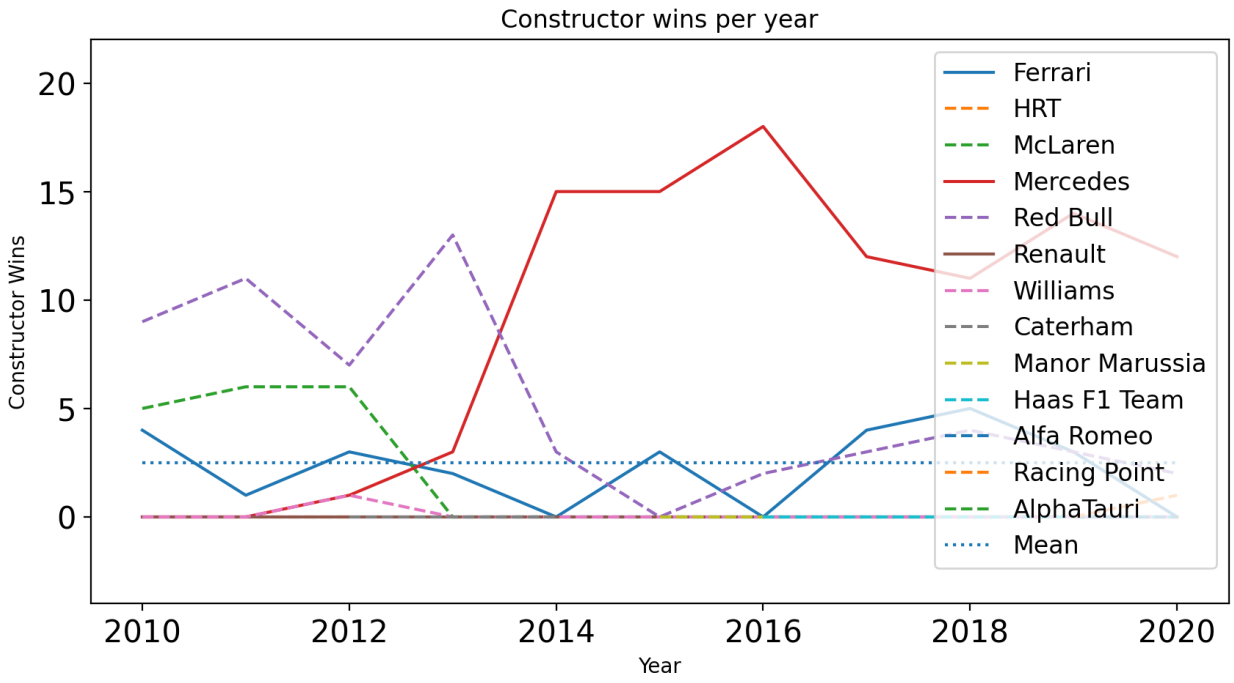
K&I: Does a team being the manufacturer of their engine affect performance?

We plotted the number of races each engine was used, yielding the following bar plot. We can see that Mercedes, Renault, Ferrari have the most number of races where their engines were used, by a wide margin.



We also plotted the points and wins for each team, with teams that make their own engines as solid lines, and other teams as dashed lines.



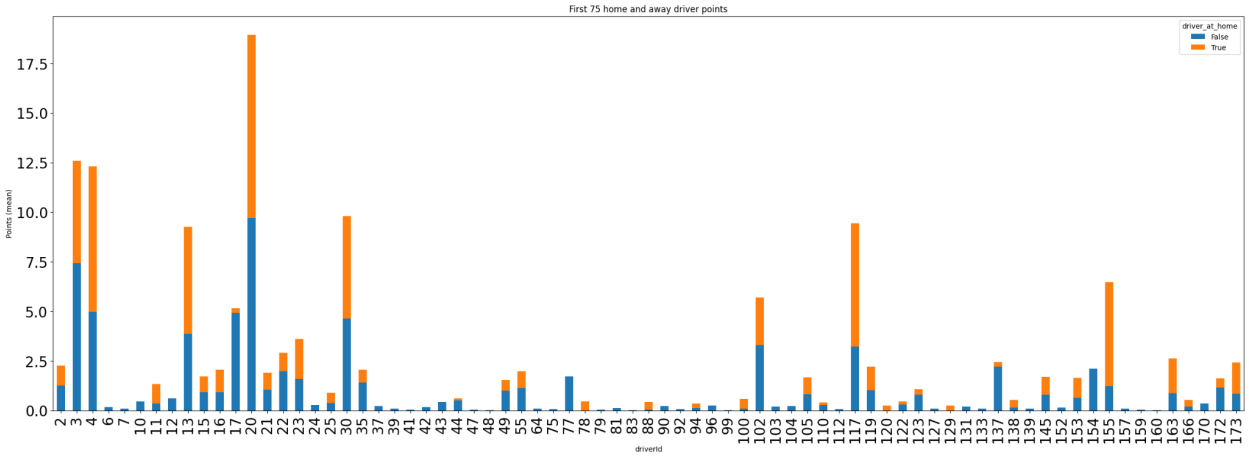


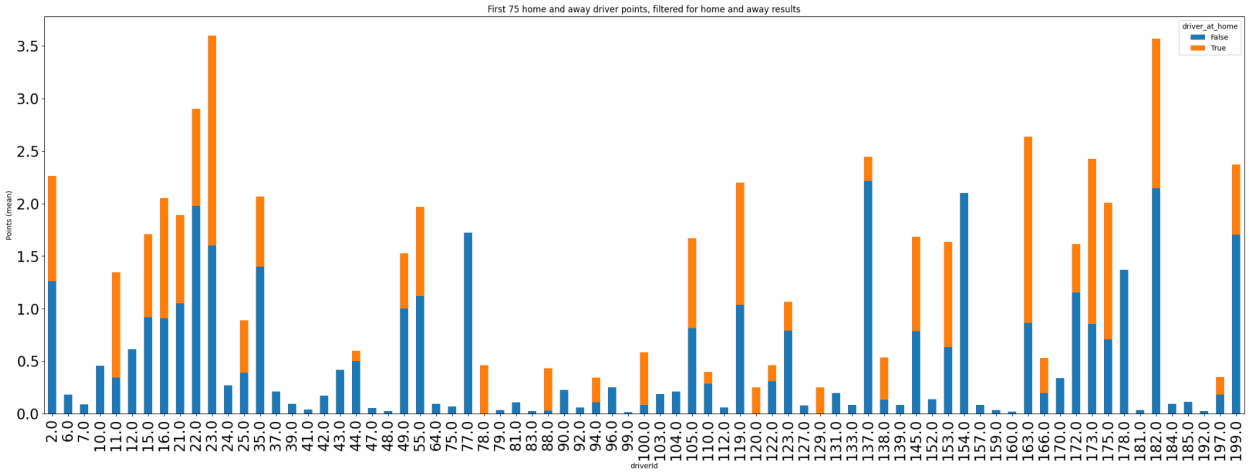
From the two plots above, we can see that teams that make their own engines seem to do well in terms of points, but a bit worse off for wins. We can also deduce that there's some correlation between the performance of engines and the frequency of their use as can be seen by the more performant engines of Mercedes, Ferrari, and Renault being used more than the rest.

- Teams that make their own engines score more points, and score more wins (to a lesser degree).
- Mercedes, Ferrari, and Renault are the engine manufacturers most linked to positive scores. Mercedes and Renault are the most linked to positive wins.

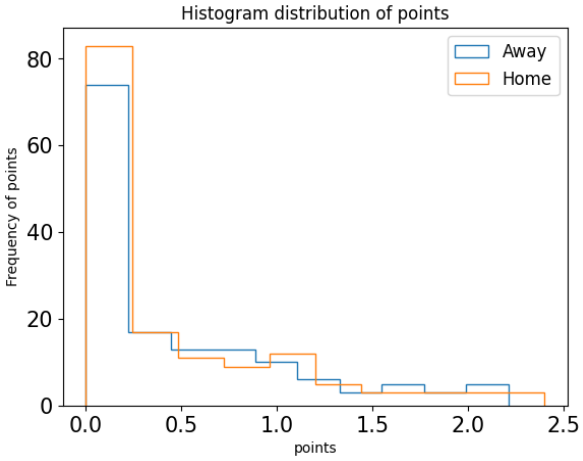
K&I: Does having a race in your country affect performance?

The following two plots show the first 75 values of the mean points at home and away for each driver. The first plot is filtered only such that it has only drivers that have both points for home and away races (even if they have zero points). While the second plot is IQR filtered to eliminate outliers for better interpretability. The X-axis is the driver ID, while the Y-axis is the mean points. The orange bars represent the mean points at home, while the blue bars represent the mean points away from home.





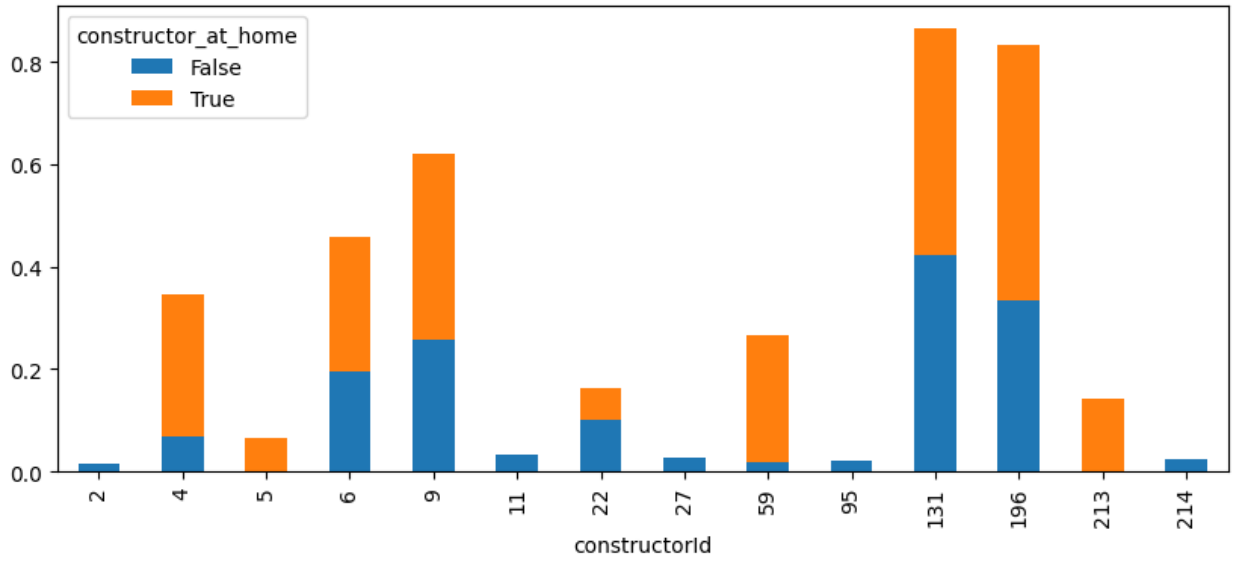
After filtering the data, it's quite obvious that for most drivers, the mean points the score at home is usually less than the mean points away from home.



Next up, we plot a histogram of the values so we can get a glance at the global distribution (not the first 75 values only). We can see that most home points fall within the lower range, although there are some that reside within the higher ranges.

Afterwards, we plot a similar bar plot but for the mean wins at home and away for each constructor (that has both home and away results). The only filtering we have is that we removed all constructors that have 0 means overall (at home **and** away). The resulting data is quite small so there's no need to filter it further. This plot shows that for most constructors, the mean number of wins at home exceeds the mean number of wins away

from home.



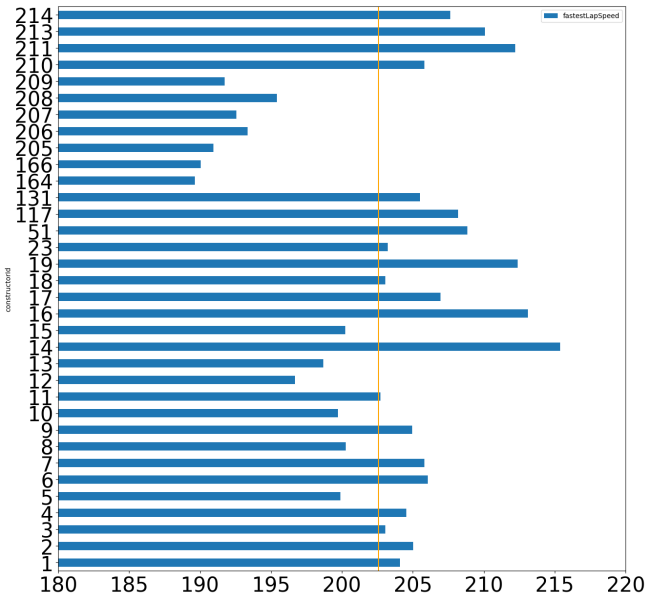
- Drivers competing at home do not achieve more points on average, while constructors competing at home do achieve more wins on average.

K&I: Given that Mercedes drivers have above average top speed, does this apply to all German teams?

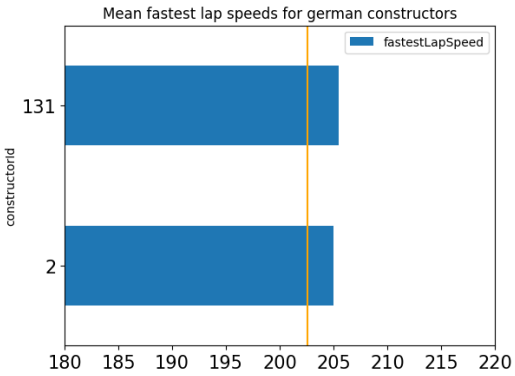
To gain some insights about the data we processed up until now, we'll visualize it.

First, we visualize the mean fastest lap speed for each constructor. The X-axis is the mean fastest lap speed, while the Y-axis is the constructor ID. The orange vertical line is the mean fastest lap speed over all constructors. We can clearly see that Mercedes (ID 131) passes this mean.

Then, we get the IDs of all German constructors so we can inspect them further. We only get two IDs (131 for Mercedes, and 2 for BMW Sauber). We plot them in the following figure.



The figure is quite empty, so it's easy to see that both constructors exceed the mean fastest lap speed value.



- Mercedes and other German teams achieve average fastest lap speed greater than the mean fastest lap speed.

4. Final findings and results.

FF&R: Does a team being the manufacturer of their engine affect performance?

- For advertisers, sponsors, or merchandisers, you should mostly consider teams that make their own engines. This info is not hard to source, and teams that make their own engines show a significant advantage in terms of scores/points, and a lesser advantage in terms of wins.
- Focus on teams like Mercedes with consistently high points and wins.

FF&R: Does having a race in your country affect performance?

- For advertisers, sponsors, or merchandisers, it's advised to support teams that compete at home frequently since they score more wins. This will draw more publicity and thus the team can be used as a strong advertising tool.

FF&R: Given that Mercedes drivers have above average top speed, does this apply to all German teams?

- German teams seem to go faster than the average non-German team. This could be a great piece of info for advertisers (ad ideas and stars) and merchandisers such as toy car manufacturers.

5. Future work and enhancements

- Like any other data science project, more data for more accurate results.
- More testing / models.
- More interpretations of some questions. (i.e. “performance” could also mean driver position)
- More plots to visualize and explore the data.