DATA VISUALUSATION INDIVIDUAL PROJECT

8th January 2018

Pedro A. López F.

Msc Data Science   Goldmiths, University of London

What has changed in almost 1,000 races of Formula 1 and what have been the major trends?

1. Introduction

Formula 1 is one of the world's most exciting events in the world of motorsport racing. The name comes from the set of rules from which all the drivers must comply to be able to race. Since the beginning, it has been a sport of high adrenaline, having drivers race in different weather conditions and compel to a consistent winning strategy to achieve the championship. Because the technical side of the sport is so important, many metrics can be scrutinized to gain interesting insights about Formula 1. Next year, the championship is going to celebrate a total of one thousand races held since 1950, and to commemorate this milestone an analysis of its evolution is presented ahead.

2. Research Questions

Through this exploratory data analysis, I would like to address interesting interrogations that have evolved in the championship since 1950.

Has the number of races per year changed?

Has the number of races increased over time? Are there any peak changes?

Has the number of drivers per team changed?

To understand the variation of teams per season and to understand the ratio of drivers per teams across time.

Which are the fastest tracks by average lap speed performed by winners?

Here I make a comparison of the average speed of the fastest lap of all the tracks in the championship.

Which are the most frequent breakdowns per decade, have the cars improved?

An analysis of the most important breakdowns or reasons to quit a race in the championship.

Which are the teams with most 1st place podiums?

Here is a comparison of the number of wins of all the teams that have participated in the championship.

Which country produces the most drivers?

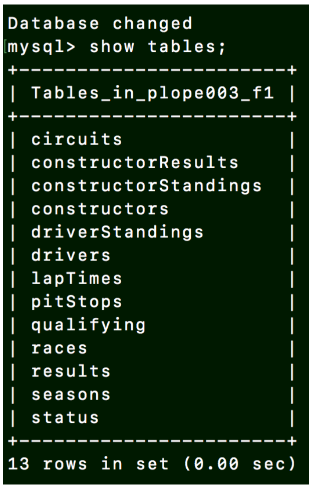
Data from the nationality of all drivers to find the driver factories of the world.

3. Data

Formula 1, like any other sport, has a vast amount of information such as standings, points, times, average speed, etc. All this information has different levels of granularity. The data can be viewed per driver, per team, per championship, and so on.

The data for this analysis was obtained from the website http://ergast.com/mrd/ which is an experimental web service that provides historical records of data from Formula 1 for non-commercial purposes. All the information for this report is retrieved from a MySQL 5.1 database dump which was restored in IGOR and connected directly to Python using sqlalchemy package which permits the user to query the data and create pandas data frame with it.

Below a brief description of the 14 tables contained in the database:



```
Database changed
mysql> show tables;
+-----------------------+
| Tables_in_plope003_f1 |
+-----------------------+
| circuits              |
| constructorResults    |
| constructorStandings  |
| constructors          |
| driverStandings       |
| drivers               |
| lapTimes              |
| pitStops              |
| qualifying            |
| races                 |
| results               |
| seasons               |
| status                |
+-----------------------+
13 rows in set (0.00 sec)
```

circuits: General information about all the circuits.

contructorResults: Information per race of each team.

constructorStandings: Information per championship of each team.

constructors: General information of each team.

driverStandings: Results per driver.

drivers: General Driver information.

lapTimes: Information relating the lap times of the drivers per race.

pitStops: Detailed information of the Pit Stops during races.

qualifying: Detailed information of the qualifying sessions.

races: List and information of all the races during the years.

results: Results of each race.

status: Status per driver after each race.

The data in this database is reliable because all information is available in the web. The benefit of using a database like this is that it collects and stores all information together. The only issue is that not all the data is consistent because the evolution of the championship, and also the internet has enabled more ways to collect uniform data across teams, drivers, tracks, etc.

Some data was also webscraped from the official www.Formula1.com archives.

For cleaning and pre-processing of the data, the pandas and numpy python libraries were the most used tools. This enabled to group data and retrieve descriptive statistics that are used in the report.

4. Relevant Research Articles

Williamson, Martin. (21 September 2017). A brief history of Formula One. ESPN UK. http://en.espn.co.uk/f1/motorsport/story/3831.html

5. Exploratory data analysis and visualisation

Has the number of races per year changed?

Has the number of races increased over time? Are there any peak changes?

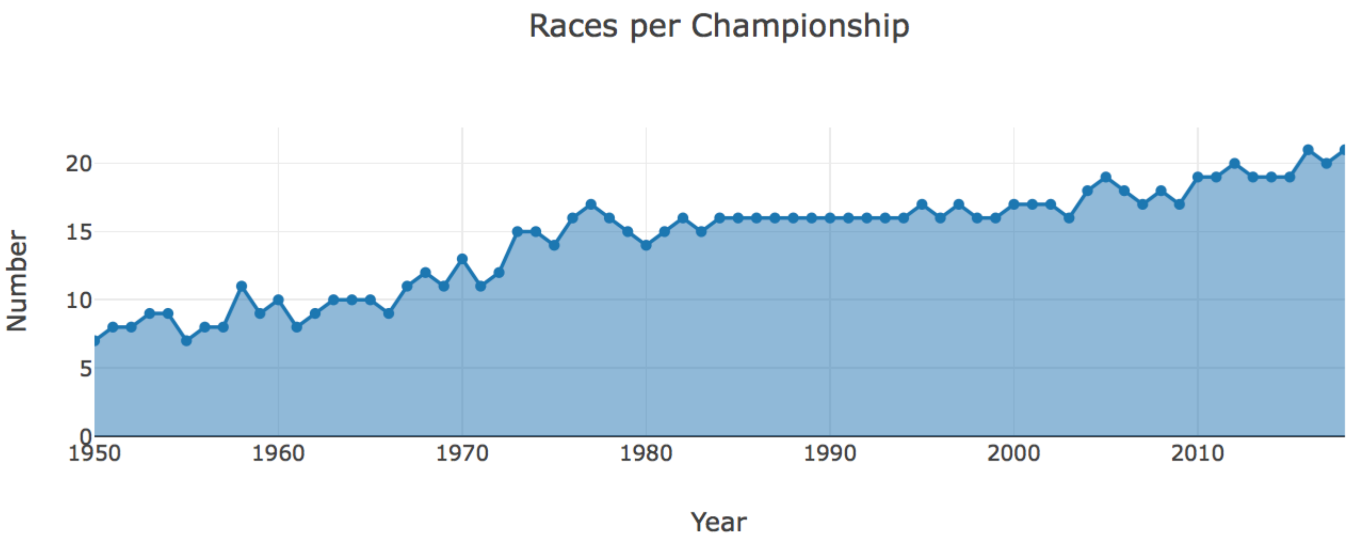


Figure 1: Races per Championship

Data: Number of official races held each year

Type of Visualization: Line Chart

Through the years, the number of official races has increased from a **minimum** of 7 races in the year 1950 to a **maximum** of 21 races in the year 2017. In terms of agenda and due to the complexity of the sport this means that each increase makes it harder for teams to keep up with the racing calendar, and this number has tripled. The turning point for number was after the year 1976 when the calendar reached the **mean** of 16 races per year, then in went down and stabilized from 1984 and 1993 to increase again. The time between races is the period where teams can make improvements to the cars. One hypothesis we can make is that in the past decades the sport has become more competitive because of the little time that teams have between races.

Has the number of drivers per team changed?

To understand the variation of teams per season and to understand the ratio of drivers per teams across time.

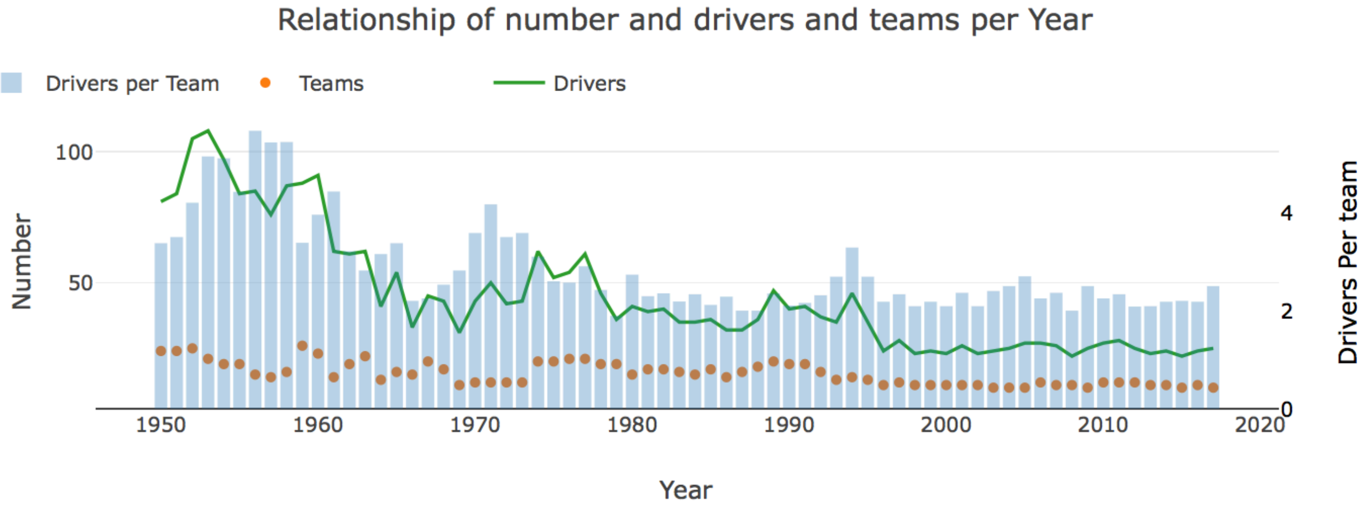


Figure 2: Relationship of number of drivers and teams per Year

Data: Number of Drivers per year, number of teams per year and Drivers/Teams

Type of Visualization: Combined Line, Scatter and Bar charts

When comparing the number of drivers per race we can observe a decreasing trend through the years changing from a **maximum** of 108 drivers in 1953 to a **minimum** of 22 in 2008. However, the number of teams has decreased as well from a **maximum** of 26 during the 1950s to a **minimum** of 10 in 2017. The most interesting fact is that the **average** drivers per team is 2.86 translating to 2 drivers per team through history. An explanation of the decrease in teams should be explained by the previous chart because with the higher number of races it is more expensive for a team to maintain a competitive performance through the racing calendar.

Which are the fastest tracks by average lap speed performed by winners?

Here I make a comparison of the average speed of the fastest lap of all the tracks in the championship.

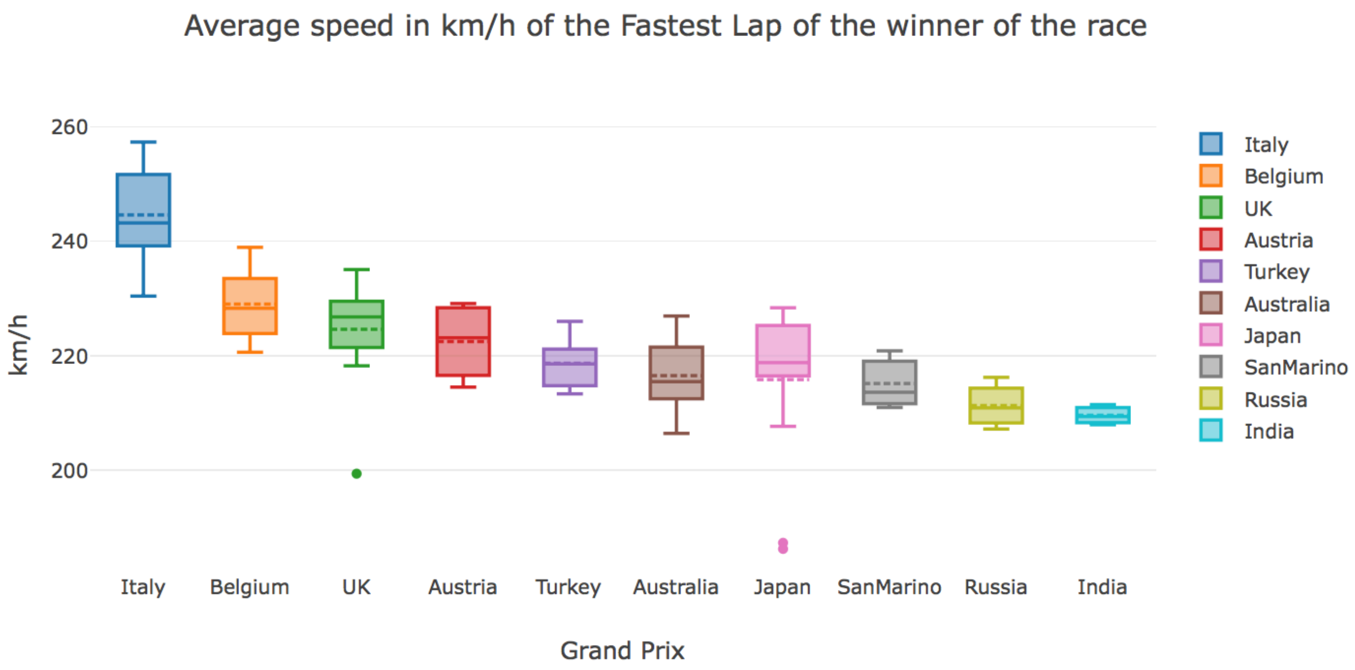


Figure 3: Average speed in km/h of the Fastest Lap of the winner of the race

Data: Average speed in km/h of the Fastest Lap of the winner of the race

Type of Visualization: Box Plot

Speed is one of the most impressive things that distinguish Formula 1 racing from other motorsports. The drivers get to impressive speeds along the tracks and additionally have the challenge to take over other competitors, take different types of turns and keep the performance until the end of the race to win. Here I took the average speed during the fastest lap of the winners for all the races that have accumulated data. I ended with a list of the top 10 tracks. The dash line in the middle of each boxplot shows the average of each track speed. The fastest track is “Autodromo Nazionale Monza” (Italy) with an imposing average speed of 244.5 km/h. This is outstanding. Just imagine driving at that average speed for a race of more than 300 kilometers [1][2][3]. The second fastest track was “Circuit de Spa-Francorchamps” (Belgium) with an remarkable average speed of 229 km/h. The numbers alone are impressive, but when compared with Italy the difference is 15km/h that translates to 6% less. On third place comes Silverstone (United Kingdom).

The range of the top 10 circuits is between a **maximum** of 244.5 km/h, a **minimum** of 209.5 km/h and an **average** of 220.7 km/h.

[1] http://www.monzanet.it

[2] https://www.spagrandprix.com/en/

[3] http://www.silverstone.co.uk

Which are the most frequent breakdowns per decade, have the cars improved?

An analysis of the most important breakdowns or reasons to quit a race in the championship.

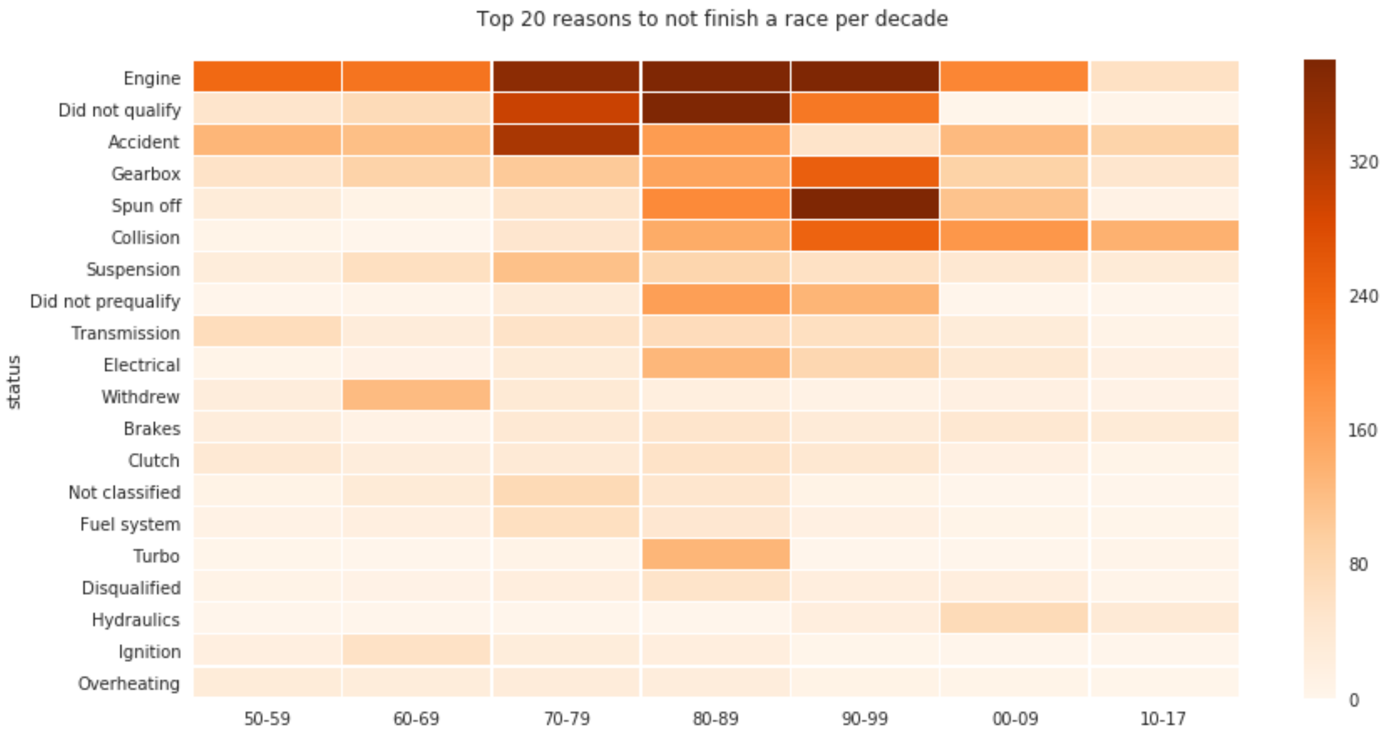


Figure 4: Top 20 reasons to not finish a race per decade

Data: Amount of times each reason occurred per decade.

Type of Visualization: Heatmap

The main objective of motorsports is to finish the race and get to the podium, but this accomplishment is not only in the hands of the Drivers. Teams must ensure reliability of the car in order to finish a race and give the pilot a chance to finish the race. From 100 reasons of not finishing I used a Paretto analysis to extract the top 20 reasons for a driver to not finish a race. Most of the reasons are related mechanical problems e.g.: Engine failure, Suspension problems, Gearbox problems. But there are some reasons related to events such as “Did not prequalify” and “Collision”. We can see an evolution through every decade, from the 70s to the late 90s the most frequent reasons where related to engine problems. Then in the 90s we can notice that “Spun off” becomes significant, and we can infer that this was due to problems with the traction of the cars. Then in the last two decades, engine failures diminished significantly and “collision” remained as a substantial reason. From this we can conclude that the car engines have improved but the human factor remains the same.

Which are the teams with most 1st place podiums?

Here is a comparison of the number of wins of all the teams that have participated in the championship.

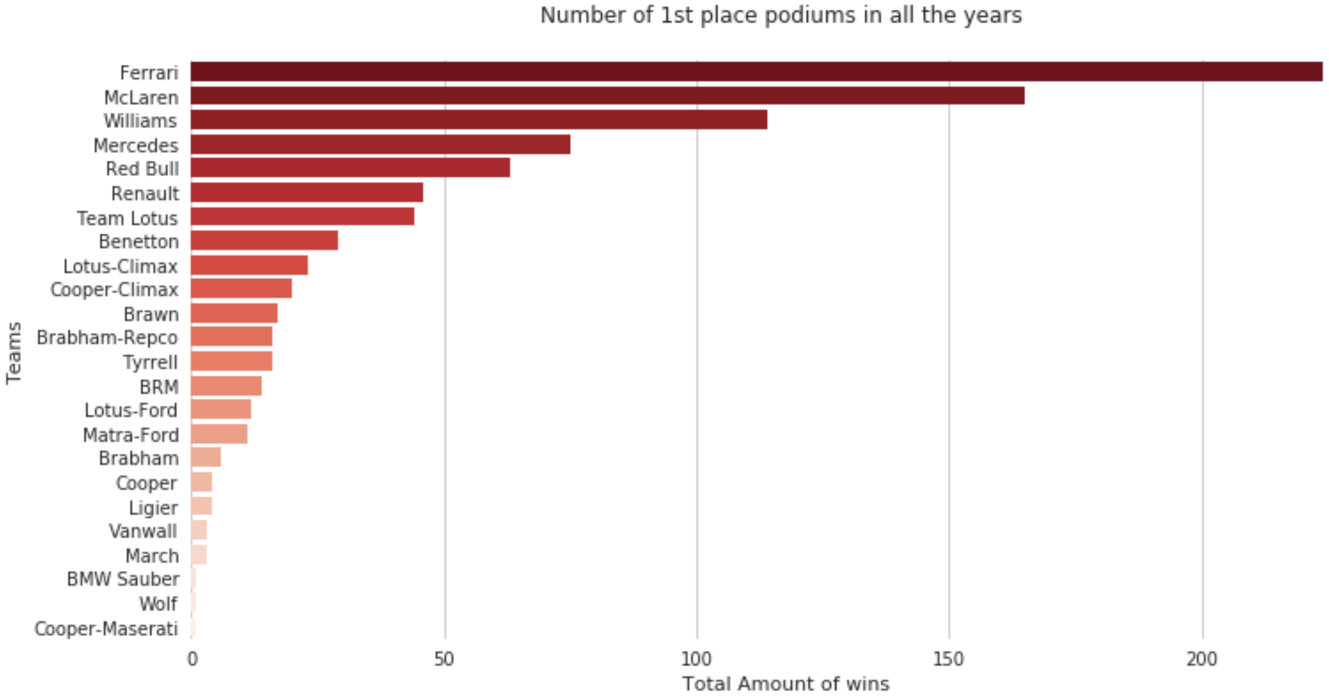


Figure 5: Number of 1st place podiums in all the years

Data: Number of 1st place wins per team.

Type of Visualization: Horizontal Bar chart.

Here are all the teams whose drivers have won at least one 1st place podium in the history of Formula 1. The top 3 contenders are Ferrari (Scuderia Ferrari) with 224 wins, McLaren (McLaren Racing Limited) with 165 wins and Williams (Williams Grand Prix Engineering Limited) with 114 wins. "How old are this teams?" is a question that comes to the mind. Ferrari is 68 years old (1950), McLaren is 52 years old (1966) and Williams is 41 years old (1977). This could explain the difference but if we look it the proportion of wins from the races each team has run the proportion is not very far apart:

Team	% Wins/Races
Ferrari	23.55%
McLaren	20.00%
Williams	16.64%

Which country produces the most drivers?

Data from the nationality of all drivers to find the driver factories of the world.

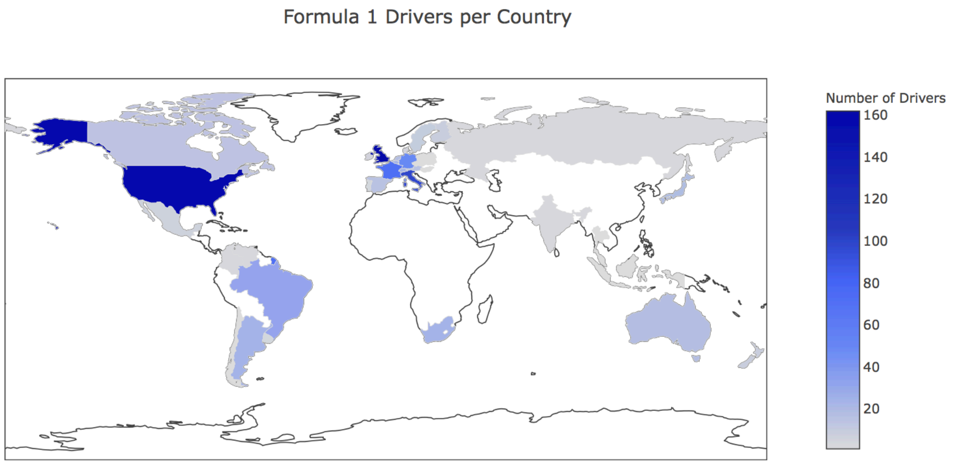


Figure 6: Number of Drivers per country

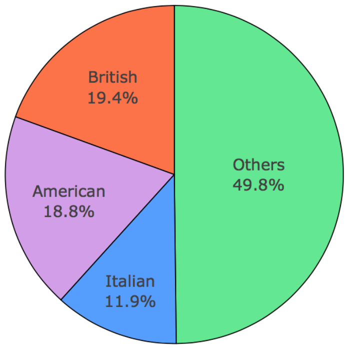


Figure 7: Proportion of Drivers per nationality

Data: Drivers per nationality

Type of Visualization: Choropleth map and pie chart.

Since the beginning Formula 1 has been an international championship, hosting races in almost all continents. Through the years, the teams must transport all their staff, cars, equipment from one track to the other to compete. But let's analyse where do the drivers come? It is very interesting that 3 countries represent more than half of the nationality origins from the drivers. United Kingdom, United States and Italy are the top three "factory of drivers", but are they the factory of champions as well?

When analysing the total Formula 1 champions, more than half come mainly from the United Kingdom, Germany and Brazil. The only country dominant in both number of champions and drivers is the United Kingdom.

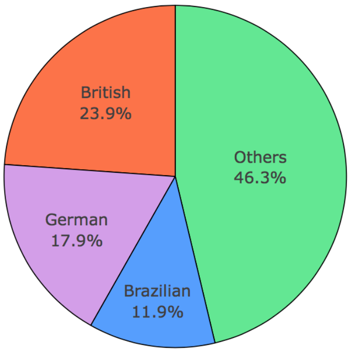


Figure 8: Proportion of Champions per nationality

6. References:

Code Reference:

http://docs.sqlalchemy.org/en/latest/core/connections.html

https://gist.github.com/stefanthoss/364b2a99521d5bb76d51

https://plot.ly/python/offline/

https://www.stackoverflow.com

https://plot.ly/python/box-plots/

https://plot.ly/python/choropleth-maps/

https://chrisalbon.com/#articles

https://seaborn.pydata.org/generated/seaborn.barplot.html

7. Conclusions

The purpose of this assignment was to analyse some relevant metrics of Formula 1, in order to identify the key changes in the evolution of the sport since 1950, considering almost one thousand races held so far. Therefore, six crucial examinations were considered to interpret what has happened.

The first finding is linked to the number of races over time. Races have tripled since the sport started, making it harder for the drivers and the engineering teams to survive during a championship. More races translate into less time between competitions, leading to a higher struggle for a team who travels from one place to another while having to fix any mechanical issue and also keep focused to win each race. It appears to have become a more popular sport, but it also requires more budget to complete a championship.

The next finding is about the number of drivers per team. Even though the amount of drivers has decreased almost 80%, the number of teams has also reduced around 61%. However, an interesting outcome is that the average drivers per team is 2 throughout history. A hypothesis for the decline in teams could be the increase in the number of races, which make the championship more expensive to complete.

Other considerations were the fastest tracks by average lap speed performed by winners. The range of the top 10 circuits is between a maximum of 244.5 km/h, a minimum of 209.5 km/h and an average of 220.7 km/h which is an impressive speed that has to be maintained during the approximately 300 km per race.

Additionally, an analysis of the most frequent breakdowns per decade was made to determine if cars have improved. Using Paretto to get the top 20 causes of failure to finish the race, it was spotted that from the 70s to the late 90s the most frequent reasons where related to engine problems. Then in the last two decades, engine failures were reduced significantly and “collision” became the highest problem. Therefore, we can infer that the car engines have improved over the years, but the human factor hasn’t changed.

Another finding was related to the teams with most 1st place podiums. The top 3 competitors are Ferrari, with 224 wins; McLaren, with 165 wins; and Williams, with 114 wins. However, the teams didn’t start at the same time, which could have explained the difference. Consequently, considering the ratio of wins to the races each team has run, the results are not very far apart, but the order remains unchanged.

Last but not least, a curious discovery was the nationality the drivers, in an attempt to locate the “driver factories” of the world. More than half of the total drivers come from 3 countries: United Kingdom, United States and Italy. Though most importantly, were are the winners from? Analysing the total Formula 1 champions, almost 54% are from the United Kingdom, Germany and Brazil. Hence, the United Kingdom is the country with the highest number of drivers and also of champions.

8. Learnings / Things to improve?

Since the beginning of the masters I’ve been exposed to python and the vast of flexibility and tools that the language provides. I agree that this module has covered the necessary tools to perform a profound analysis and exploration of a dataset.

The pandas and numpy libraries provide a very efficient way to analyse vast amount of data, but the most important use of them is the reusability of the analysis. Once a good code is made to analyse a data source, the data source size can increase and all the analysis will be automatically updated. The definition of the data types is a key element of each analysis not only to work with the data in python but to understand your data as well. I think the purpose of this work is scalable to other datasets, and it can be applied in the industry.

Another thing to point out is the flexibility to combine data sources is incredible. This is my first experience or doing a project with Relational Databases in this case MySQL and the process retrieve the information and summarize it to get the insights, was a very simple process.

One problem that I encounter with finite datasets is that in some instances the use of a spreadsheet software can result in faster data pre-processing. But this is overshadowed by the vast options of visualization available within the python package. The reason why I used plot.ly for most of the charts because this library has the flexibility to visualize the data interactively. The good thing about libraries like this is that they force the user to prepare the data in certain way makes the data standard and easy to share. The other library used was Seaborn, which has good options of visualizations and are very customizable as well.

I think with continuous practice and learning I should focus in more general and readable code that can be reusable for other datasets.

One thing to note is that not all the documentation is available for these packages and is very complicated to execute the ideas regarding the visualisations. It is still a process of trial and error to customize the different options of plots from these libraries.