National College of Ireland

Data Analytics Project

Using Historical Data to Identify the Best Driver in Formula 1 History

BSc(Hons) in Computing

Data Analytics

2021/2022

Adolfo Carvalho

17102871

x17102871@student.ncirl.ie

Acknowledgements

I would like to express my sincere gratitude to the people who cooperated and motivated me during the journey of this project:

- Tom, my son who was born while I was starting to develop this analysis and served as extra motivation for my graduation.
- Gracielle, my beloved and dedicated wife for coping with my short availability during the most important time of our lives.
- Stefano Petrini, PhD, my good friend for helping me with the development of the formulas used in the scores computations.
- Guilherme Mohor, PhD, one of my dearest friends, also passionate about Formula 1, for helping me in defining what makes a great driver.
- Pedro Ferraz, the friend with whom I speak the most about Formula 1, and many other sports, for providing me with his insights and assisting in data sourcing.
- My supervisor, Ms Divyaa Elango, for the patience and the good counselling and advice.

1. Executive Summary

Sports are a core aspect of human society. As a species we're not only content in playing, we enjoy the competition. In fact, we are the only species on the planet that play games with specific sets of rules while other animals just play randomly to test their physical skills and/or to assess their hierarchical position in their society (Fagen, 1974). As civilization evolved and modernised, so did sports. New tools and new artefacts opened the possibilities for new types of competitions.

Formula 1, in the media and this project referred to as F1, is one of the most technical sports in existence and it is an attraction for sports fans, car lovers, technology enthusiasts and people eager for adrenaline. Being so attractive to fans and bringing an enormous sum of money every year makes F1 the most successful motorsport competition in the world.

Apart from the competition, an aspect of following a sport that any fan enjoys is having arguments and discussions about the sport they follow. Since Muhammad Ali claimed to be "The GOAT", The Greatest of All Time, among every sports fan around the globe there's a discussion of who was the greatest in that sport and such debates are responsible for a considerable amount of media coverage.

As we phase through the Artificial General Intelligence stage of AI (Shevlin, Vold, Crosby and Halina, 2019) it is very important to be able to interpret subjective concepts in objective terms. The term GOAT is very subjective and every pundit will bring some emotional baggage to their argumentation. This project hopes to shine a light on how to properly identify what and how objective parameters can be used by an AI to construct a subjective concept.

Although the aim of this project is mainly academic, it is possible to extrapolate the knowledge gained from it into multiple commercial areas for example:

- Media and publicity: being able to rank competitors would allow the leagues to focus the coverage on the best performer athletes;
- Gaming: sports games are a huge industry and to achieve a realistic competitiveness gaming companies spend much of their efforts on ranking the characters;
- Betting houses: data analysis, DA, and AI are already vastly used by betting houses, this project might bring more ideas on how to apply statistical analysis to odds prediction;

2. Introduction

2.1. Background

In the early 90's the most popular athlete in Brazil was Ayrton Senna. He was a three-time world champion of F1 and the weekly family entertainment was watching F1 races on Sundays. Unfortunately on the 1st of May 1994, a tragic accident during a race took Senna's life. Years after his death his name is still associated with greatness and he attracts a lot of passion, as in any other sport, both loving and hating. Various reporters and analysts agree on the opinion that he'd have won many other races and championships if he had the chance to finish his career. On the other hand, it isn't unusual to hear that his death created an aura around his name and that he is unfairly regarded as a top driver.

Years later many other talents have emerged in the sport, Michael Schumacher, Fernando Alonso, Sebastian Vettel, Lewis Hamilton and Max Verstappen are the most known examples. Pointing out who performed better is a task that requires not only too much knowledge about the sport but also processing so much information that makes it infeasible to humans.

This debate occurs in any other sport and engaging in it is part of being a sports fan. A curious aspect of this debate is that it carries a lot of emotion but at the same time the arguments are armoured by numbers and statistical facts. For example, nobody expects a Portuguese to admit that Lionel Messi is a better footballer than Cristiano Ronaldo the same way an Argentinian is not expected to consider Cristiano Ronaldo better. But at the same time, very few Portuguese would say that Pepe is better than Messi, simply because there are no statistics that could be used to sustain this argument. So it is clear that such subjective opinion is in a certain manner based on objective data.

Two works revised in this project, and detailed further below, attempted to quantify the data available and answer the question of who's the best F1 driver. They are very popular among F1 fans, especially the one made by Amazon, who sponsors F1 and used it as a publicity stunt, and both works had many points in line with the general public opinion and some other points that raised eyebrows. In the end, these papers served as an inspiration for this project.

Does considering pure statistics create something too different from what people feel? Is it wise to rely solely on numbers to generate a ranking of F1 drivers? Are machines advanced enough to join the conversation about such human-made concepts? Such questions are the driving forces behind this project and hopefully, the answers to those questions will be unveiled in this report.

2.2. Aims

At the present state of AI, researchers are putting a massive effort into human-like intelligence. To achieve this the machines must be able to simulate and replicate subjective ideas. One interesting feature of AIs would be the ability to develop taste and predilection for sports, arts and culture.

The main goal of this project was to provide an objective ranking of the best drivers in F1 history and compare that ranking with the mainstream media opinion, so we can determine that the AI applied to this analysis is in line with human thoughts.

This ranking was generated by gathering the maximum amount of data available about all the races and drivers in F1 history. After the data was prepared and normalised, thirteen statistical attributes were calculated and with a help of a random forest algorithm it was possible to identify how much of a part each of these attributes played in the fact of a driver winning the championship in each year, that's how we got the weight for each attribute in a ponderate average that was used to calculate the *Base Score* for each driver. The *Base Score* was then used to compare each driver, year by year, against each teammate that they had raced in their career. This comparison was used using a second mathematical formula and generated the *Refined Score* which resulted in the final ranking.

This report also presents various charts to help better understand the ranking and the drivers' comparisons as well as mainstream media comparison with the ranking generated.

2.3. Technology

In deciding what technology to be used in this project some Reddit and Youtube videos were reviewed and along with "Python and R for the Modern Scientist" (Scavetta and Angelov, 2021) it was decided that both R and Python were gonna be used in different stages of the project.

Also reviewing various data storage alternatives, SQLite stood out for being portable and robust. The idea of keeping the data in CSV or TSV format was discarded as there were too many files and some complex queries would have to be performed, which was easily done with a SQL database but could be an issue with a CSV file.

In the first stage of the project, the data scraping was completed with R. Using RStudio, six R scripts were written to gather various datasets from three different websites. The data scraping had the aid of various R libraries and browser extensions.

The second and main stage was performed using Python. It involved pre-processing and normalising the datasets, building the SQLite database, mining and analysing the data. For the data mining, decision tree and random forest algorithms were considered and after carrying out a performance comparison between the two it was decided that the random forest algorithm was the best fit for this project.

Below is a list of the libraries and browser extensions used throughout the analysis:

- Rvest (R) — Web Pages Scraping
- Stringr (R) — Characters Manipulation
- Pandas (Python) — Data Manipulation
- Numpy (Python) — Data Manipulation
- OS (Python) — Managing files/directories
- Unidecode (Python) — Characters Manipulation
- Math (Python) — Calculations
- GDown (Python) — File Downloads
- Matplotlib (Python) — Data Visualisation

- ➢ Seaborn (Python) Data Visualisation
- ➢ SQLAlchemy (Python) ORM Database
- ➢ SKLearn (Python) Machine Learning
- ➢ SciPy (Python) Data Visualisation
- ➢ Selector Gadget (Google Chrome) Allows selecting page elements CSS attributes

The Python scripts were published in a Github repository and also put together in a Google Colab Notebook for easy access and sharing, allowing the non-technical audience to view the raw material. Other tools were used on different scales during the project, namely: Sublime Text, Google Chrome, Google Docs and Google Sheets.

2.4. Literature Review

The title "GOAT" lacks objectivity and is mainly used in casual environments, at most it is mentioned during event transmissions and media debates. For this reason, very little research has been found in that direction and that was an encouraging factor for this project as it can be a precursor to understanding how to translate human subjective terms into machine parameters.

Two worth mentioning works attempted to build a similar ranking in F1. Despite having some critique of those works they were somehow an inspiration for this project and their imperfections were taken into consideration in an attempt to have a more accurate analysis.

2.4.1. Amazon AWS: The fastest driver in Formula 1

This work was very popular among the F1 community at the time when it was published by the Amazon AWS team (2020).

They made it very clear that they attempted to only identify the driver's speed during the qualifying and, in their own words, this is 'one element of a driver's vast armoury', nonetheless it is a very important element of that armoury. AWS built a mathematical model to determine a list of the 50 drivers that, supposedly, have the better pace during a qualifying session.

The first critique of AWS's model reflects more a personal disappointment, also felt by numerous F1 fans, rather than a constructive commentary. It's difficult to understand why they limited their work only to the qualifying pace. With all the knowledge and resources that they had at their disposal, many people found it frustrating that they chose to self restrict to a single characteristic of a driver.

Another point to be made is that they only collected data from 1983 to 2020. Obviously, before that, the qualifying had a different format but it's understood that it would be possible to correlate the qualifying pace to the race pace for the pre-1983 era. By picking only a range of seasons, they purposely didn't consider data of highly estimated drivers like Juan Manuel Fangio, Jim Clark and Emerson Fittipaldi. Starting the analysis in 1983 would also affect the ratings of drivers who started their careers before that year and retired after, such as Nelson Piquet and Niki Lauda.

Finally, they output a ranking where the score is the time difference to the best time. For example, Ayrton Senna is the fastest and Michael Schumacher is the second with a time

difference of 0.114s. But what does that time difference mean? In long tracks like Spa Francorchamps, that means they are quite close whereas in short tracks like Interlagos it is a reasonable difference.

Overall it is a very interesting work and the opportunity to expand such work and bring more topics to the analysis was very inspirational.

2.4.2. Reiner Eichenberger and David Stadelmann: Who is the best Formula 1 driver? An Economic Approach to Evaluating Talent

Before the AWS work, this was a famous paper within the specialised media, not so much among the fans. In their article, they attempted to establish who was the fastest Formula 1 driver (Eichenberger and Stadelmann, 2009). They tried to apply economics models to the sports world which was an interesting idea but it led to some misleading findings.

Studying their work it is notorious that the most apparent flaw is that they over attempted to isolate what they called *Car Factor*. The *Car Factor* would be how a driver's achievement was possible thanks to, and only to, their equipment. But in the F1 community, there is a saying that 'great drivers drive great cars' (Mercedes-AMG PETRONAS F1 Team, 2020). It is quite intuitive that a top team wouldn't have interest in paying more to a driver if any other driver could deliver the same results for less money. Thus in their effort to eliminate the *Car Factor* Eichberger and Stadelmann ended up polluting the result.

Other than that it is a very concise work and it would be interesting to see it applied to current data as it dates back to 2009, meaning that it misses out on Sebastian Vettel and Lewis Hamilton who won 4 and 6 championships, respectively, since then.

2.4.3. The success equation: Untangling skill and luck in business, sports, and investing

Even though his book isn't specific about F1, Michael Mauboussin (2012) sets an example of how to quantify sports statistics.

This book lays out various standards and ideas for quantifying skills and success. For this project, the philosophy that was taken from the book and is decisive in concluding that the final ranking is relevant is that outliers athletes are a combination of great skills and great luck.

To understand how important that is first it is necessary to remember that more people have been to space than drivers who have won championships. It means that being an F1 world champion is being an outlier driver. That leads to the conclusion that when measuring the success of the world champions skill as well as luck is being measured. Therefore the ranking shown in this report is, in the final sense, a measurement of drivers' skills.

2.5. Structure

The following list gives an overview of this report's sections and briefly describes what is covered under each section:

- Data: details about the raw data that was used in the analysis. Characteristics of the file structure and source of data are expanded in this section;
- Methodology: an overview of the analysis methodology. It is broken down into subsections that better detail each stage of the project;

- Data Selection: an explanation of how the required data for the project was selected and sourced, it is an expansion of the *Data* section of this report. It also has a subsection with data exploration analysis giving better insight into the data that was acquired;
- Pre-processing: details of the actions taken to clean and prepare the data for the analysis;
- Transformation: an explanation of how the pre-processed data was transformed into its final form to be used by the machine learning algorithms;
- Data Mining: the actual core of the analysis, this covers all the techniques used to achieve the results and the justifications for using such techniques;
- Conclusion and Discussion: using the results obtained from the *Data Mining* a conclusion is reported including an evaluation of the tools and techniques used;
- Further Development: a suggestion of how this analysis could be improved, how it can serve future analysis and what kind of adjustments would be necessary to keep it up to date;

3. Data

3.1. Selection

Due to the nature of the data selected for this project, over three thousand datasets were used during the analysis. Despite the large number of files used they have a small number of records each. The datasets contain, for example, the starting grid for each race in history which amounts to over one thousand files. The datasets were split into eight categories:

3.1.1. Champions

- **Description:** This is a single file category and contains a single dataset that holds the history of all the world champions for each year in history.
- **Source:** The list was obtained from the F1 Fansite world title list page (2022a) and due to its simplicity it was just copied into Google Sheets and saved as a CSV file.
- **Structure:**

File Format	CSV
Data Format	Structured
Number of Columns	2
Number of Rows	72

Table 3.1: Champions dataset

3.1.2. Drivers Races

- **Description:** For each driver that participated in an F1 event a dataset was created and it contains one row per event that the driver was present. Each dataset was named with the driver's name and surname.
- **Source:** This dataset category was scraped from the F1 Fansite list of drivers (2022b) by an R script which iterates through a list with all drivers' names and opens the respective link to the driver's career details.
- **Structure:**

Files Format	CSV
Data Format	Structured
Number of Datasets	866
Number of Columns	6
Number of Rows	Varies

Table 3.2: Drivers Dataset

3.1.3. Races Results

- **Description:** These were very crucial for the project. They contain the official race results for each race that happened in F1. The datasets were organized in 72 folders, one per F1 season, and in each folder, there is one dataset per race of that season.
- **Source:** Another R script was created to iterate through the official F1 Archive website (2022) and source the race's information.
- **Structure:**

Files Format	CSV
Data Format	Structured
Number of Datasets	1057 (in 72 folders)
Number of Columns	5
Number of Rows	Varies

Table 3.3: Races Results dataset

3.1.4. Drivers Races Complement

- **Description:** The *Drivers Races* datasets had some inconsistencies that were found during the pre-processing and transformation stages. Those inconsistencies were found when searching for the driver within the *Race Results* datasets. For example, Alan Rees had three races in his dataset but when looking into those race results his name was not found because, in this situation, he was listed to race but never actually participated in the event.
- **Source:** Each case was manually investigated as there were multiple causes for the inconsistencies therefore it was not possible to automate the process. Once the problem was acknowledged a final dataset was manually downloaded from Stats F1 (2022b).
- **Structure:**

Files Format	CSV
Data Format	Structured
Number of Datasets	866
Number of Columns	6
Number of Rows	Varies

Table 3.4: Drivers Races dataset

3.1.5. Lineups

- **Description:** One of the key calculations that this project does is to compare the drivers against their teammates over the years thus it is important to have the teams' lineups per season: which team each driver raced for. Each dataset belongs to a

different team that competed in F1 and one record per driver per year, for instance, if a driver raced for team A in the years 1967 and 1968 the A dataset will contain two rows for this driver, plus the rows for other drivers-years.

- **Source:** A R script scraped the data from Stats F1 (2022a).
- **Structure:**

Files Format	CSV
Data Format	Structured
Number of Datasets	206
Number of Columns	2
Number of Rows	Varies

Table 3.5: Lineups dataset

3.1.6. Races Grids

- **Description:** This contains the starting grid for each race in F1. Similarly to the *Race Results* datasets they are placed in 72 folders, one per season, and one file per race in that season.
- **Source:** This data was also sourced from the official F1 website (2022) by an R script. This script loads the races for each season and searches for the starting grid information for all races in that season.
- **Structure:**

Files Format	CSV
Data Format	Structured
Number of Datasets	986 (in 72 folders)
Number of Columns	5
Number of Rows	Varies

Table 3.6: Grids dataset

3.1.7. Races Grids Complement

- **Description:** The official F1 website, for an unknown reason, doesn't provide the starting grid for all the races. Those races with missing starting grids were identified during the pre-processing and transformation stages by comparing the *Race Grids* and *Race Results* datasets.
- **Source:** After detecting the races which didn't have a starting grid dataset, the complementary dataset was scraped with R from the F1 Fansite results (2022c).

➢ **Structure:**

Files Format	CSV
Data Format	Structured
Number of Datasets	71 (in 9 folders)
Number of Columns	2
Number of Rows	Varies

Table 3.7: Grids Complementary dataset

3.1.8. Wet Races

➢ **Description:** This category contains a single dataset which records all races that were affected by rain in the history of F1.

➢ **Source:** A scraper was written in R to check the weather information of each race from Stats F1 (2022c) then manually, using Google Sheets, the drain weather races, such as 'Sunny' and 'Cloudy', were removed from the dataset.

➢ **Structure:**

Files Format	CSV
Data Format	Structured
Number of Datasets	1
Number of Columns	2
Number of Rows	Varies

Table 3.8: Wet Races dataset

3.2. Exploratory Data Analysis

The Exploratory Data Analysis, EDA, was performed on the transformed data and it will help to better understand the data that is being used in the analysis as well as point to early findings and statistics contained in our data.

Table 3.9 shows some drivers' statistics:

➢ Career Seasons: how many seasons each driver competed over their careers, only counting drivers who competed more than one season to exclude guests at specific events;

➢ Career Races: how many races each driver raced during their careers;

➢ Career Wins: number of times a driver won a race;

➢ Career Podiums: number of times drivers reached top 3 positions in a race;

➢ Career Pole Positions: number of races a driver started in front;

Stat	Seasons	Races	Wins	Podiums	Poles
Count	533	779	111	215	102
Mean	5.11	30.29	9.55	14.85	10.29
Average	4	7	4	3	4
Std	3.63	53.66	15.62	25.41	15.62
25%	2	2	1	1	1
75%	7	33	11	18.5	13
Min	2	1	1	1	1
Max	19	351	103	182	103

Table 3.9: Stats Summary

Figure 3.1 is a chart that demonstrates how important it is to evaluate the data year by year: it shows the average wins, podiums and pole positions per season and we can see that all three increase as the drivers get experience but fall at the end of their careers:

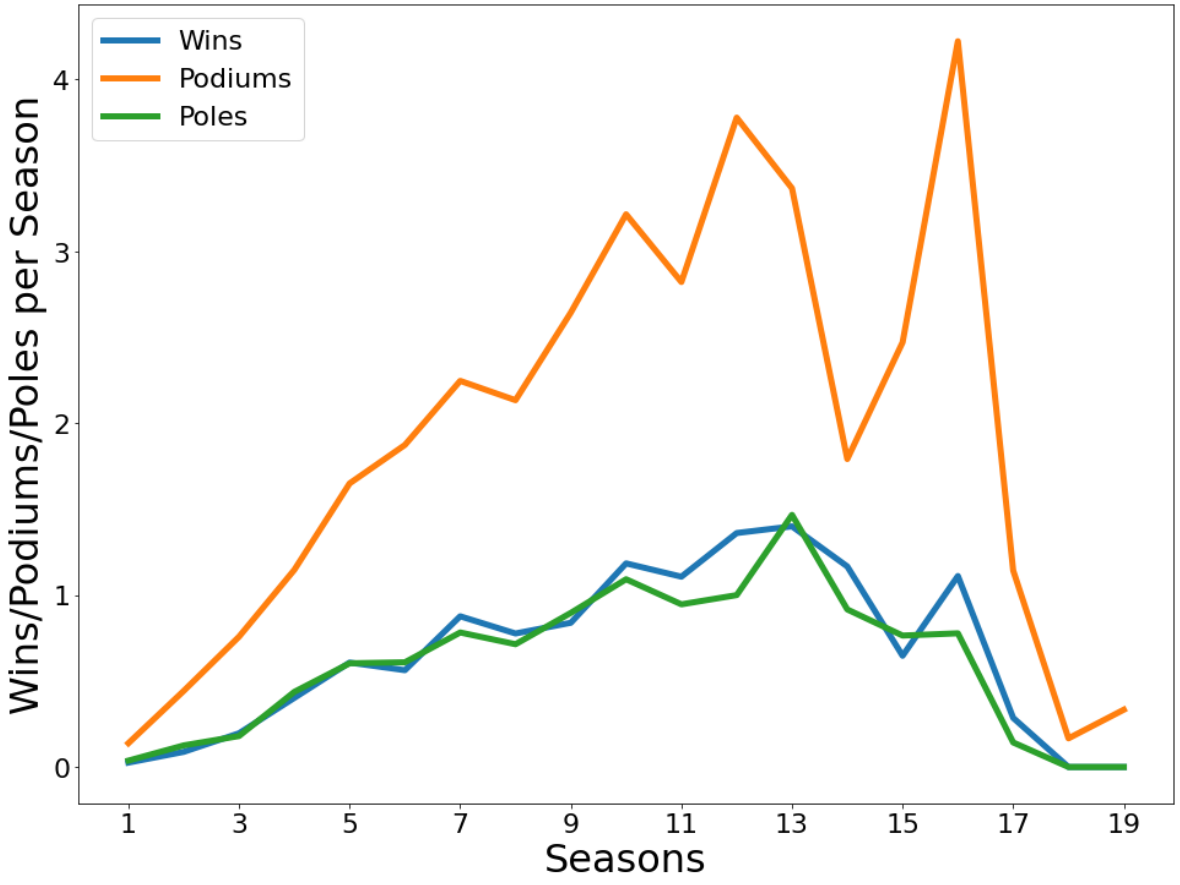

Figure 3.1: Wins, Podiums and Poles per Season

Furthermore, Figure *3.2* illustrates how consistent F1 drivers have been over the seasons. It plots the average points per race for all drivers over all seasons. From this plot, it is visible

that the distribution is pretty constant over the seasons. It debunks the argument that F1 has changed over the years and the different regulations and season lengths make it impossible to make a proper cross-season comparison:

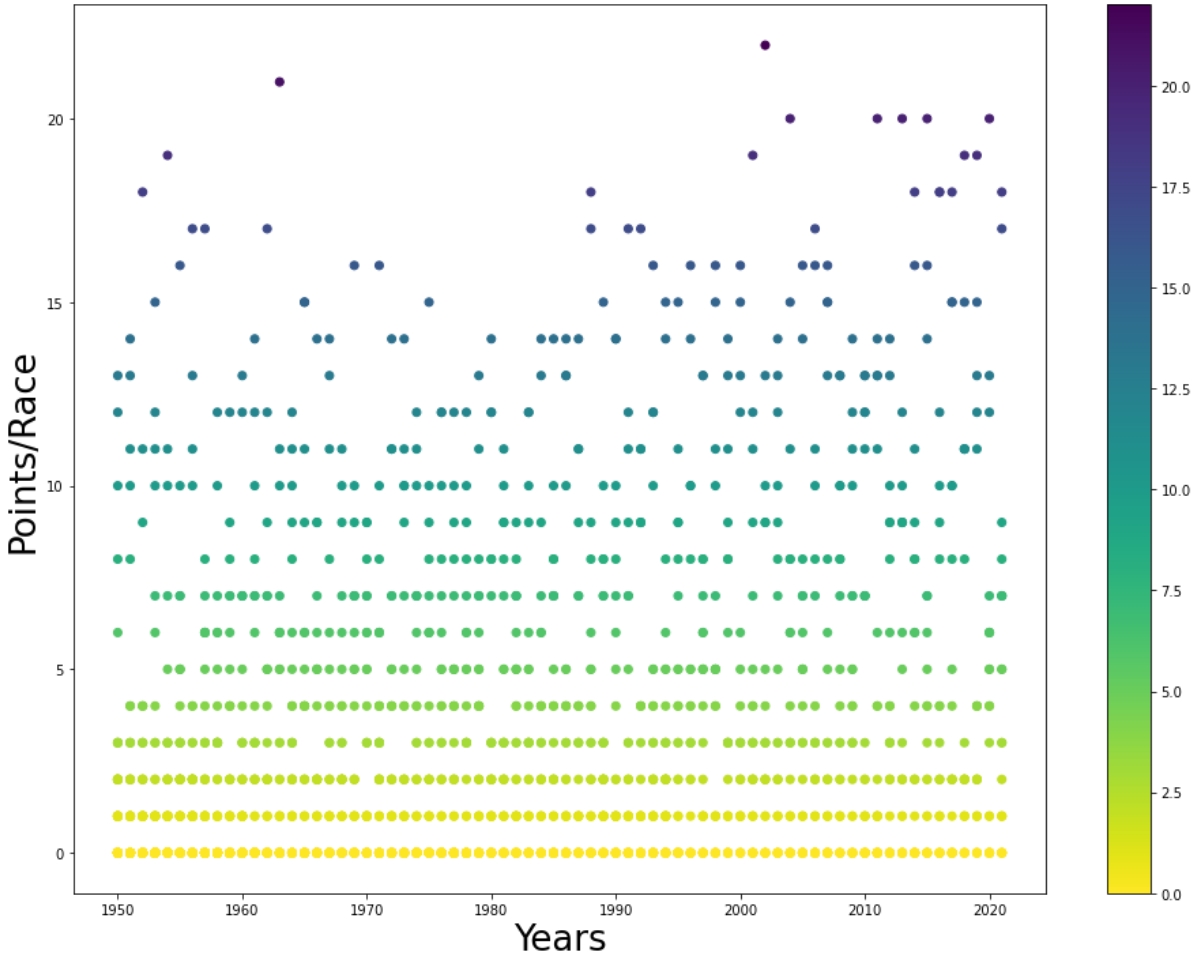

Figure 3.2: Scatter Points per Race over Years

Finally, *Figure 3.3,* is a violin plot which shows the distribution of the number of victories by the starting position. Starting the races from the first row is the better position to win the race, followed by the pole position and then 3rd or lower is, as many would imagine, the starting position with the least chances of winning the race:

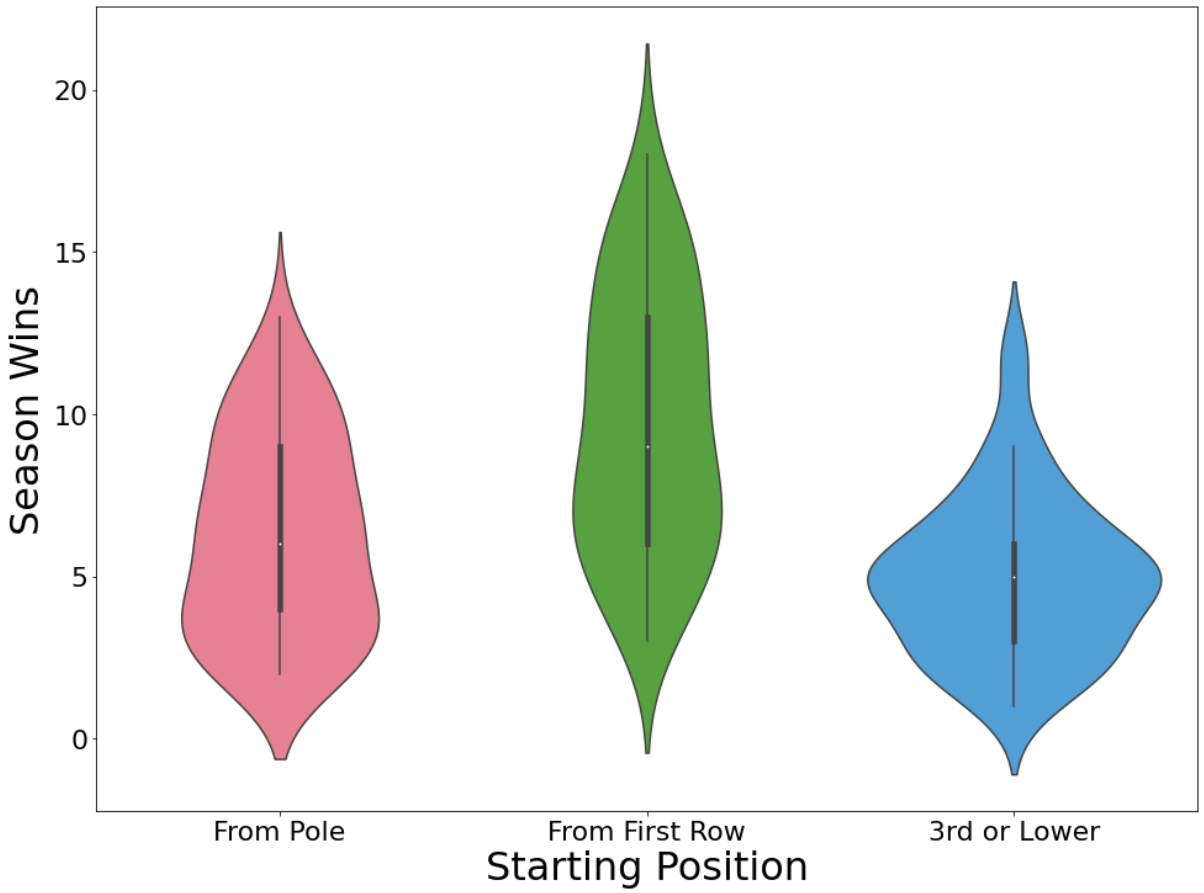

Figure 3.3: Winning probability by starting position

Some other charts and tables could be retrieved from the data, but these were the most relevant for this analysis and are enough to illustrate the data we are dealing with and to support the arguments that will be present in this report.

4. Methodology

This project was developed using a methodology known as Knowledge Discovery in Databases, KDD. As it is a widespread methodology in the data analysis sector it was found to be the best suited for this project as plenty of resources were available for studying and its structure also provides strong foundations for this analysis. As various datasets were gathered from different sources and had to be normalised and modified before being assembled into a single database, KDD seemed to be the methodology that was more aligned with the structure of the project.

In *Figure 4.1* we see the 5 stages of KDD (Fayyad, Piatetsky-Shapiro and Smyth, 1996):

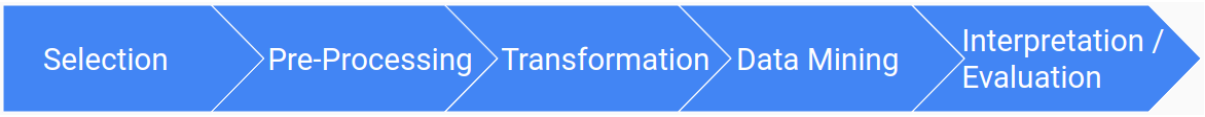

Figure 4.1: KDD Stages

4.1. Selection

The task of identifying the data needed for this project revolved around answering the question:

"If I were an F1 team boss looking for a driver, what qualities would I be looking for to select my driver?"

The first step to answer that question was to look into the most mentioned stats in F1 debates and records that seemed more relevant in the debate. Once the required stats for this analysis were properly identified it was a matter of going after the sources. The first source was the official F1 website mainly because it is the most respected source of data alongside with other two websites, F1 Fansite and Stats F1, all data necessary for the computations in this project was gathered.

The full details of the datasets used in this project are found in Section 2 of this report.

4.2. Pre-Processing

Following Wickham's (2014) Tidy Data paper suggestion, the structure for the datasets is as follows:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Such structure was achieved during the data selection stage, as most of the data were scraped by customised R scripts they were already gathered in the format intended to be used during the analysis.

Normalising the data was the task that required the most effort during the pre-processing stage. As the datasets were gathered from different sources it was necessary to ensure that all variables were in the same format. An example is the name of the races, which from

some sources had the track name and others had the country or even the city names where the races occurred. Normalising the data was done with the help of Python scripts.

Some normalisation tasks were automated and the script made the necessary changes to the datasets while others had to be manually checked. In such situations, the Python script generated an error output pointing to the information that needed to be checked and each issue was addressed accordingly.

Below is a summary of the pre-processing measures performed:

4.2.1. Remove Drivers' Names Accentuations

Some drivers' names contained special characters, for example, Jose Froilan González had to be renamed to Jose Froilan Gonzalez. This action had to be performed both in the *Drivers Races* datasets names and also in the records inside *Race Grids* and *Race Results.* Renaming the drivers was easily accomplished with a Python script which used the package Unidecode to leave only standard characters.

4.2.2. Normalise Drivers' Names

The second issue with drivers' names was that some sources would use three names for the driver whilst others would use only first and last name or even in some cases abbreviations were used instead of the actual names. To identify the issues a Python script scanned the datasets and datasets names for inconsistencies and generated an output file with the errors. The error had various causes and had to be manually investigated, the most common causes were:

- ➢ Driver missing in races or grid dataset: caused by a certain driver being listed for a race but at the last moment didn't participate in the event;
- ➢ Abbreviation: A.J. Foyt was used instead of Anthony Joseph Foyt, some other occurrences like this were also identified;
- ➢ Missing Middle Name(s): Certain sources used the full name of the drivers when others didn't include some of the middle names;
- ➢ Names Prefixes and Hyphen: surname prefixes, for instance 'van der', were in certain data sources ignored and removed from the drivers' names. Similarly, compost names were linked by a hyphen on some occasions and separated as two names on others;

In total 712 issues were reported by the Python script. Although this number seems too large for manual inspection there were a lot of multiple occurrences for the same driver that once fixed would fix many other issues.

4.2.3. Normalise Race Names

In the same way that the drivers' names didn't follow the same pattern for all the data sources the names of the races also had to be normalised. The race names didn't only depend on the source, as on some occasions the same source used different names for the same race event. The solution was to create a dictionary with the possible names and assign them to a key. This dictionary was imported in multiple Python files used during the analysis

and used when a normalised race name was necessary. An entry example of this dictionary is:

"United States" : ['United States', 'usa', 'america', 'american', 'COTA', 'Circuit of the Americas']

Meaning that races named 'United States', 'usa', 'america', 'american', 'COTA' or 'Circuit of the Americas' would all be treated as 'United States'. This dictionary was also used when seeding the datasets into the SQLite database so it only used the key values.

4.2.4. Look For Missing Starting Position

There were occasions when it was not possible to establish the starting position of a driver at a certain race. The causes for this issue varied from case to case, ranging from the driver starting from the pit lane to the driver not starting the race at all, and each cause had to be dealt with individually. Using the output of a Python script 135 errors were manually scrutinized and solved.

4.2.5. Inspect Wet Races Dataset

Even using the race names dictionary, there were inconsistencies with the *Wet Races* dataset which caused the race not to be identified in the other datasets. There were five errors and they were all caused by the race name having the year attached to the race name, for example, Indianapolis1950 instead of Indianapolis 1950. Despite all issues having the same principle, there were too few occurrences to justify the development of an automated script to deal with them and they were all fixed by hand.

4.3. Transformation

During the transformation stage of the analysis, the pre-processed data was transformed into its final state to be fed to the data mining algorithms. For this project transforming the data meant to compile all the datasets into a relational database, in SQLite.

All work done during the transformation was completed using Python and several scripts were written to seed the information into the database. The database was created with the SQLAlchemy toolkit for python, it is an ORM library which makes it much easier to deal with database queries and helps to improve the code readability.

The SQLite database contains nine tables and information that would be duplicated in the datasets now is related to the appropriate table using foreign keys. Out of the nine tables in total, six were created during the transformation process, they are:

4.3.1. Drivers

- **Name:** *drivers*
- **Description:** The table *drivers* is referenced by every other table except *races*. It contains basic information about the drivers. It was built with information from the datasets categories *Drivers Races* and *Drivers Races Complement*.
- **Number of records:** 779

➢ **Columns:**

Field	Type	Description
id	Integer, primary key	Unique ID for each driver
name	String	Driver name
surname	String	Driver surname
abbreviation	String	Abbreviation used by the FIA

Table 4.1: Drivers DB table

4.3.2. Races

➢ **Name:** *races*

➢ **Description:** As the header of the races, name and year, doesn't need to be replicated for each *race_results* (see next section), this table was created to hold such information and it is referenced by *race_results.* The information in this table came from the datasets categories *Races Results* and *Wet Races.*

➢ **Number of records:** 1057

➢ **Columns:**

Field	Type	Description
id	Integer, primary key	Unique ID for each race
name	String	The race's event name
year	Integer	Year in which the race occurred
rain_affected	Boolean	TRUE if it was affected by rain, FALSE on the contrary

Table 4.2: Races DB table

4.3.3. Races Results

➢ **Name:** *race_results*

➢ **Description:** This table contains the details of the races. Each registry is a driver that competed in a certain race, therefore it references both the *drivers* and the *races* tables. To compile this table the data was gathered from the datasets categories *Races Results, Races Grids and Races Grids Complement.*

➢ **Number of records:** 23594

➢ **Columns:**

Field	Type	Description
id	Integer, primary key	Unique ID for each result
driver_id	Integer, foreign key	ID of the driver that participated in this race
race_id	Integer, foreign key	ID of the race the result is being recorded
starting_position	Integer	Position in the grid in which the driver started the race
finishing_position	Integer/String	Position that the driver finished the race. Can contain 'DQ', 'EX' or 'NC' if not completed or disqualified
team	String	Team that the driver was racing for
retirement_reason	String	If a driver didn't complete the race, a detail of the reasons why

Table 4.3: Races Results DB table

4.3.4. Lineups

➢ **Name:** *lineups*

➢ **Description:** To identify the drivers' teammates the lineups is very important. This table is a one to one copy of the *Lineups* datasets, except that instead of containing the actual driver name it references it to the *drivers* table. Also the name of the team, that was used as the dataset name, became a column in this table.

➢ **Number of records:** 3259

➢ **Columns:**

Field	Type	Description
id	Integer, primary key	Unique ID
team	String	Name of the team
year	Integer	Year which the driver raced for the team
driver_id	Integer, foreign key	ID of the driver
total_races	Integer	Number of races the driver raced for the team this year

Table 4.4: Lineups DB table

4.3.5. Standings

➢ **Name:** *standings*

➢ **Description:** It records how many points each driver scored in a certain season. By selecting the *finishing_position* values from the *race_results* table, the points were calculated using the current F1 regulation, as the points system varied over the years. From first to the tenth place were assigned points in the order: 25, 18, 15, 12, 10, 8, 6, 4, 2 and 1.

- **Number of records:** 2001
- **Columns:**

Field	Type	Description
id	Integer, primary key	Unique ID
year	Integer	The year which the season happened
driver_id	Integer, foreign key	ID of the driver
points	Integer	How many points the driver scored in that season

Table 4.5: Standings DB table

4.3.6. Champions

- **Name:** *champions*
- **Description:** This table is just one record per champion in history. It is a copy of the *Champions* dataset.
- **Number of records:** 72
- **Columns:**

Field	Type	Description
id	Integer, primary key	Unique ID
driver_id	Integer, foreign key	ID of the driver
year	Integer	Year in which the driver won the championship

Table 4.6: Champions DB table

With the data normalised and in its final form, the analysis per se could begin. As the F1 statistics, like most sports, are mainly numerical data there was no need to perform any sort of conversions or to scale the data except for the weather conditions that were categorised in two categories: rain-affected and not rain-affected, this was converted as a boolean column in the *races* table.

5. Data Mining

To understand the steps taken in the Data Mining, DM, it is important to remember that the goal of this project is to build a ranking of the best drivers in history and this ranking would use statistical data found in the datasets and then comparing of each driver's stats with every teammate's they had over their careers.

To accomplish this goal the first step was to identify what sort of statistical data was present in the datasets, and once identified they were stored in a database table called *base_score_stats.* This table contains 2971 records and each record represents the stats of a driver during a certain season. It was generated by a Python script which iterated through all other tables counting the relevant stats per driver and then seeding them to the final table. The *base_scoe_stats* columns are listed in *Table 5.1*:

Field	Type	Description
id	Integer, primary key	Unique ID
driver_id	Integer	ID of the driver
year	Integer	Year in which the stats were obtained
races	Integer	Number of races the driver raced in the season
wins	Integer	How many races the driver won during the season
podiums	Integer	How many times has the driver reached the podium (top 3) in the season
poles	Integer	Number of races that the driver started from the first position on the grid
front_rows	Integer	Number of times that the driver started either from first or second place in the grid
wet_races	Integer	How many rain-affected races happened in the season
wet_wins	Integer	How many rain-affected races the driver won during the season
wet_podiums	Integer	How many times has the driver reached the podium of rain-affected races
gained_positions	Integer	Sum of how many positions the driver gained over the season. Eg: if started 3rd and finished 1st, then gained 2 positions
lost_positions	Integer	Similar to gained_positions, but it's the sum of lost positions over the season
retirements_for_collisions	Integer	How many times has the driver retired due to a crash

points	Integer	Sum of points the driver scored over the season
wins_not_from_pole	Integer	The number of times the driver won during the season when not starting the race from the pole position
starting_position	Integer	Average of starting position the driver had in the season
season_races	Integer	Total number of races in the season (not only the ones the driver participated in)
races_until_now	Integer	Number of races the driver participated in their career up to the current season
races_record_until_now	Integer	Maximum number of races by any driver until the current season
championships_until_now	Integer	Number of times the driver won a championship until the present season
championships_record_until_now	Integer	Maximum number of championships won by any driver until the current season
champion_this_season	Boolean	TRUE if the driver won the current championship and FALSE if not

Table 5.1: Base Score Stats DB table

An issue with using the data as they were saved in the table is that the ranges are too different and lack a standard, consequently, they all had to be computed in a certain way to result in a value that ranges from 0 to 1, where 0 is the worst and 1 the best result. Such computation produced thirteen statistical attributes that would be the core of the scores calculations. The rationale and the explanation of each of these attributes are as follows:

- Wins ratio: winning races is the achievement that most attracts media and attention to a driver not to mention that is the attribute that most increases the chances of winning a championship. Instead of using the raw number of victories a driver had, a ratio was calculated by dividing the number of victories by the number of races the driver competed in the season.
- Podiums Ratio: reaching the podium is in itself a great achievement by a driver, not as much as prestigious as winning but a great feat. Consistently being present on the podium is an accomplishment that most champions had over the years, it was also calculated by dividing the number of podiums a driver had by the number of races the driver raced in the year.
- Poles Ratio: qualifying doesn't award points but allows the drivers to start the race in the best position possible. During qualifications is also the moment when the drivers have the opportunity to demonstrate their raw speed, some drivers were really good during races but not so fast in a single lap. Dividing the number of pole positions by the number of races the driver raced generated the ratio of the pole positions used in the calculations.

- Front Rows Ratio: starting from the second place isn’t too bad and in some circuits is even preferable as it allows the driver to get a *toe* when a car behind gains speed by aerodynamic turbulence generated by the car in front and increases the chances of reaching the first corner at the front. Similarly to the previous attributes, it is also the division of the number of front rows starts by the number of races.
- Wet Wins Ratio: the rain reduces drastically the influence of the car in the competition and levels the field. Also by the increased difficulty of racing in the rain, it is a great benchmark of raw talent. Dividing the number of times a driver won races affected by rain by the number of rain-affected races they were present is how this ratio was obtained.
- Wet Podiums Ratio: In the same way podiums are a good indicator of talent, podiums in rain-affected races are also a very good indicator that the driver is highly skilled. During wet races is when drivers with not so good cars have greater chances of reaching the podium.
- Positioning: being able to retain the position and, most importantly, gain positions during the race is how the racecraft skills of a driver are judged. Defending, overtaking and taking advantage of strategic opportunities is crucial for a driver to succeed in F1. This is one of the most complex attributes to calculate and to ensure that it ranges from 0 to 1 it is calculated by the gained positions average (gained positions divided by the number of races) minus the average of the lost position (lost positions divided by the number of races) all divided by the starting position average (starting position divided by the number of races). This way if a driver raced twice in the championship, one starting 11th and finishing 5th and the other starting 5th and finishing 6th they gained 6 positions, lost one and on average started in 8th in two races which gives a final positioning score of: [(6/2) - (1/2)] / (8/2) = 0.625.
- Retirement for Collisions: it doesn't matter how skilled a driver is if they are involved in too many accidents. Crashes are part of the sport but many drivers didn't get to grow in their careers because they were too frequently involved in collisions. As this attribute is inversely proportional to the driver's skill it was calculated by subtracting the number of collisions the driver had from the number of races and then dividing the result by the number of races. Therefore if a driver didn't crash during the season they had a score of 1 and a driver who crashed all their races had a score of 0.
- Average Points: the final standing is what matters during the season and to feature at the top consistency is key. Great drivers manage to squeeze little points even when their race seemed to be a tragedy. It was calculated by dividing the number of points at the end of the season by the number of races the driver competed in and then again dividing by 25, the most points a driver can attain in a race, to ensure the 0 to 1 range.
- Wins Not From Pole Ratio: In parallel with the positioning, winning when not starting at the front shows good racecraft skills. Dividing the number of wins when the driver didn't start at the front by the number of total wins is how this ratio was obtained.

- ➢ Season Appearances: some drivers only appeared in a few events during the season, normally replacing an unavailable driver or sometimes as special guests. But this causes a problem: in some circumstances, it may increase the stats too much, for instance, if a driver only appeared in one race and won the race it may seem that the driver had a brilliant season when in fact they never really competed for the championship. Dividing the number of races the driver participated by the actual number of races in the season is a way of ensuring the accuracy of the score.
- ➢ Championships Ratio: in history, there are two ways a driver can be part of the fan's memory and media coverage: winning championships or having a long career. To give credit to drivers with a longer career, this and the next attributes use the driver record at the present season against the historical record at that particular season. The historical record was only calculated up to the current year, and year by year, in an attempt to achieve a better balance between newer and old drivers.
- ➢ Races Ratio: as the previous attribute it is used to award driver's experience. It is a division of how many races the driver had in their career by the record of races at the present season.

Once the attributes were established and balanced in common values it was necessary to verify if using those attributes makes sense for the analysis. It doesn't matter how many attributes or in which way historical data is used if it is pointless for the final argument.

To do such verification a Person Correlation test was performed with a simple average of the above attributes and the fact that a driver won the championship or not. Winning the championship was used as the dependent variable as being a champion is the ultimate goal of a driver in F1 and when talking about greatness it is the most important achievement.

The hypothesis for the test is:

H_o: *There is no* significant correlation between the attributes average and winning the championship

H_1: *There is* a significant correlation between the attributes average and winning the championship

Figure 5.1 shows that when winning the championship the drivers had a much greater *Season Base Score* than those who didn't win:

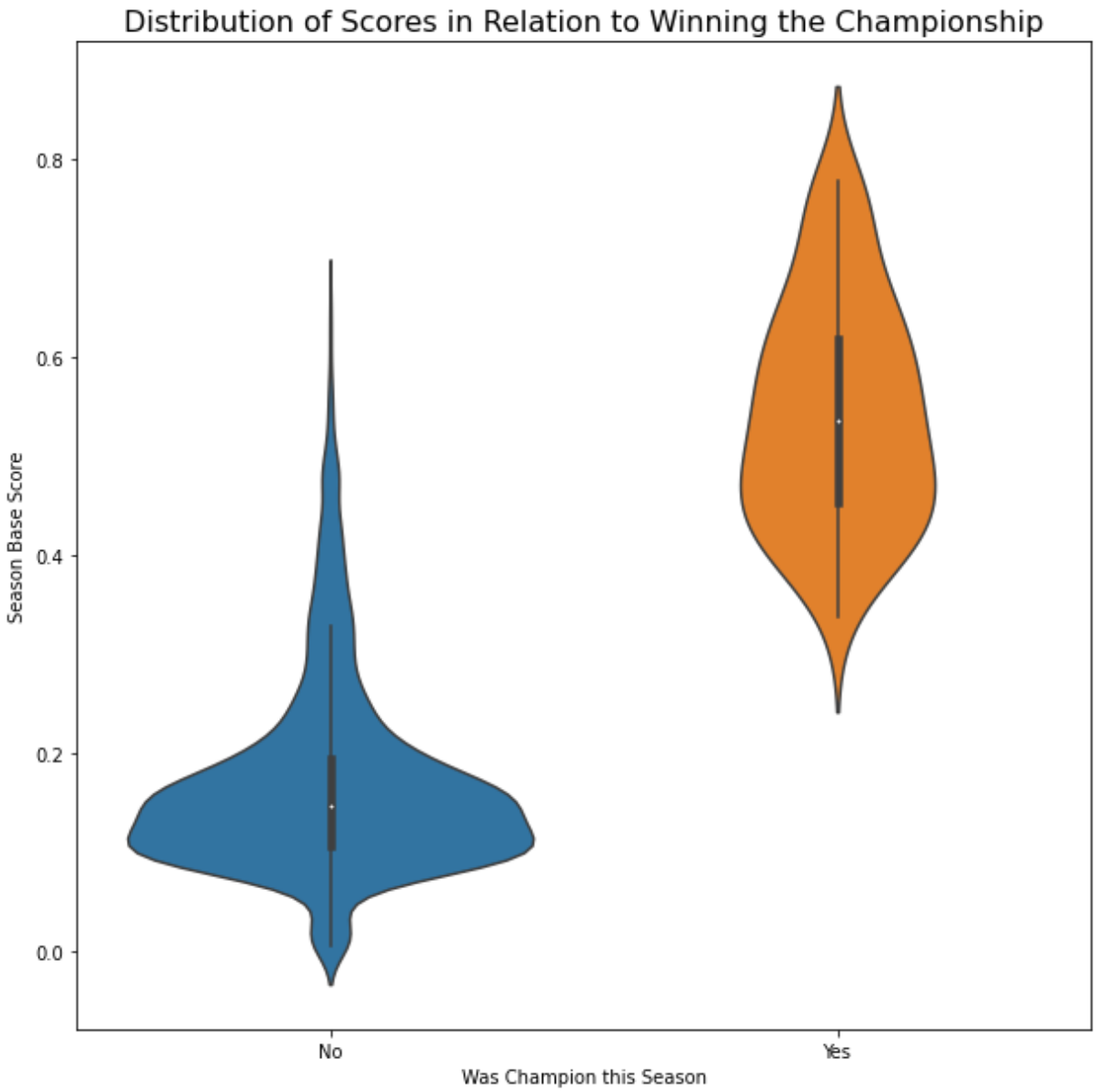

Figure 5.1: Distribution of Scores in Relation to Winning the Championship

The *Correlation Coefficient* obtained was 0.511 with a p-value of $4.45*10^{-198}$, thus we reject the null hypothesis and confirm that there is a correlation between the average of attributes and winning the championship.

To improve the correlation coefficient the solution was to apply different AI algorithms to establish the weight of each attribute. These weights are vital for the final calculation as clearly some attributes are more impactful than others.

The first algorithm that was considered for the analysis was Decision Tree but during the research stages, Random Forest seemed the best fit for the project. To select the most appropriate both were applied to the data and a table of metrics comparison was generated and showed that despite having very similar scores overall, Random Forest is slightly more accurate. The fitness of the model was also plotted to generate a visual representation of both algorithms.

This task was done in Python with the usage of the Scikit-Learn package, the metrics comparison table and the fitness plot generated are shown in *Table 5.2* and *Figure 5.2*:

Metric	Random Forest	Decision Tree
Accuracy Score	0.9896967755712149	0.9868673926969892
Balanced Accuracy Score	0.8652796494602438	0.8602344915118871
Precision Score	0.8157864810377166	0.7263516960762105
Recall	0.7346573565143596	0.7272179850661227

Table 5.2: AI Algorithms metrics comparison

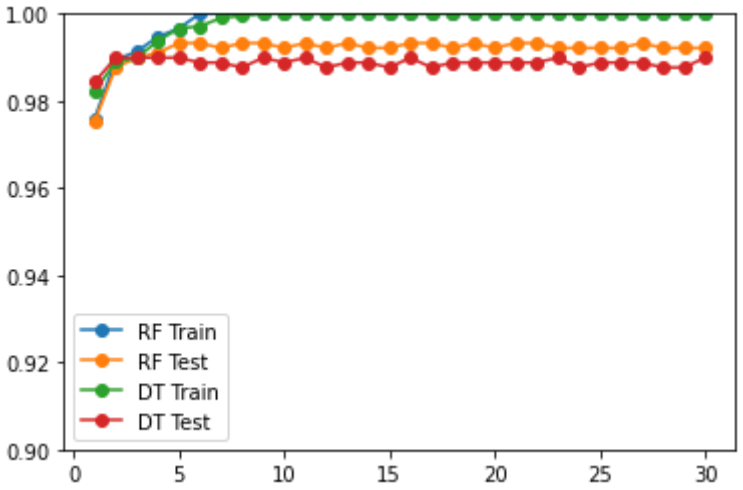

Figure 5.2: AI Algorithms fitness

Now that the attributes were demonstrated to be relevant to the analysis and that the Random Forest algorithm showed better accuracy it is possible to start calculating the drivers' scores. Firstly, also with help of the Scikit-Learn package, we split the dataset in two: 70% training and 30% testing. The ratio 70/30 was chosen because it is industry standard for a dataset of the size being used. The attributes weights were:

- Wins Ratio: 0.230710
- Average Points: 0.165837
- Championships Ratio: 0.143366
- Poles Ratio: 0.112597
- Podiums Ratio: 0.083959
- Front Rows Ratio: 0.052263
- Wins Not From Pole Ratio: 0.050193
- Races Ratio: 0.041287
- Season Appearances: 0.032841
- Positioning: 0.026928
- Wet Podiums Ratio: 0.023868
- Retirement For Collisions: 0.022006
- Wet Wins Ratio: 0.014143

As seen in *Figure 5.3:*

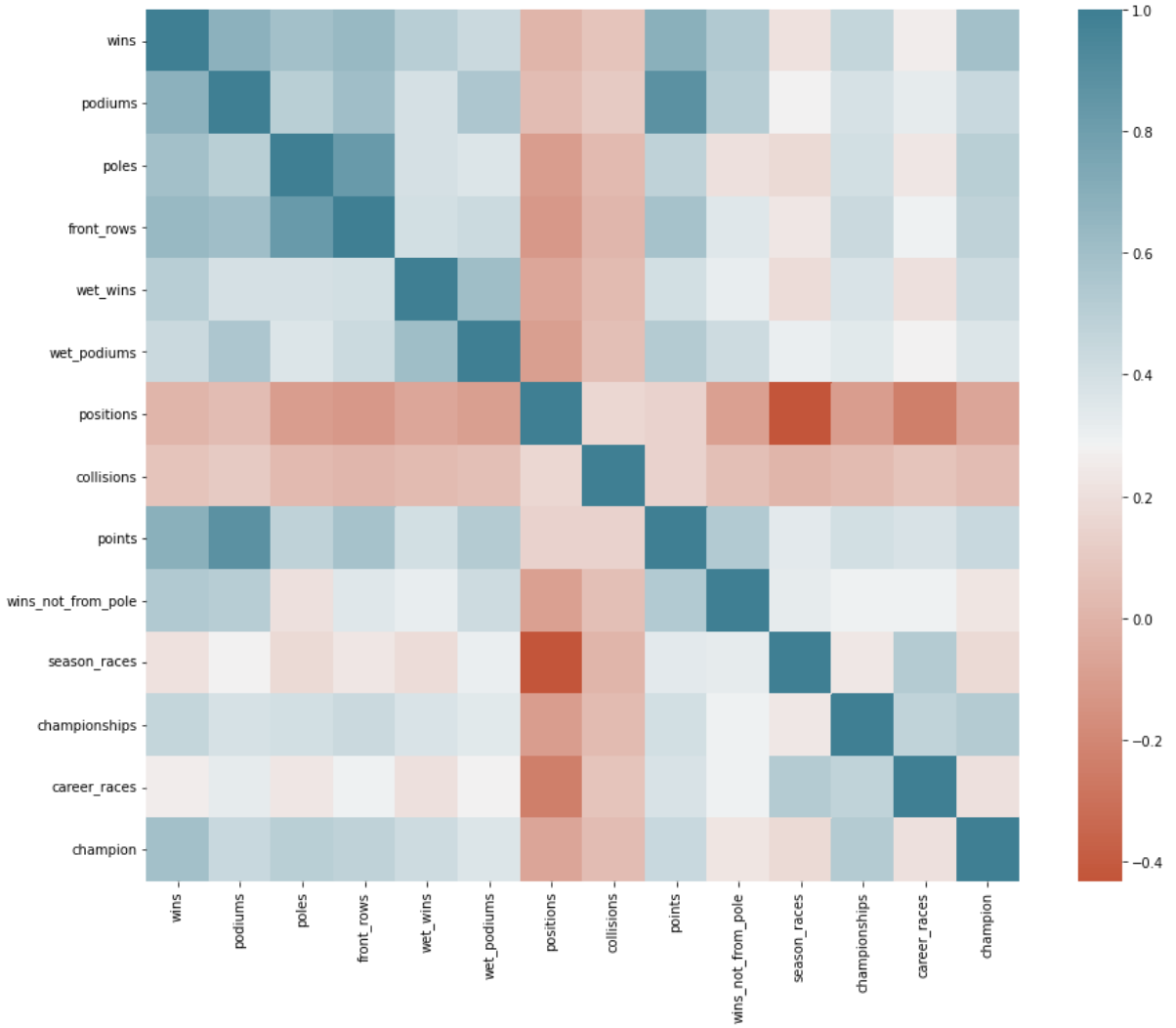

Figure 5.3: Attributes correlation heatmap

Using the attributes and their respective weights, the following formula was used to calculate the *Season Base Score* of each driver in each season:

$$S_i = \text{ɑ} * C_{\text{ɑ}} + \beta * C_{\beta} + \text{ɣ} * C_{\text{ɣ}} + \Delta * C_{\Delta} + \varepsilon * C_{\varepsilon} + \theta * C_{\theta} + \lambda * C_{\lambda} + \sigma * C_{\sigma} + \tau * C_{\tau} + \varphi * C_{\varphi} + \chi * C_{\chi} + \psi * C_{\psi} + \omega * C_{\omega} \quad (1)$$

Where:
i: the season for which the score is being calculated
S_i: ith season base score
C: the correlation coefficient for each attribute
ɑ: wins/races
β: podiums/races
ɣ: pole positions/races
Δ: front rows starts/races
ε: rain-affected wins/rain-affected races
θ: rain-affected podiums/rain-affected races
λ: (average gained positions-average lost positions)/average starting position
σ: (races-retirements for collision)/races
τ: career points/races/25
φ: (wins-pole positions)/wins
χ: races driver raced in the ith season/total races in the ith season

ψ: championships driver won until the ith season/most championships won by any driver until the ith season
ω: driver's career races until the ith season/races record at until the ith season

Using this formula the *Base Score* for each season for every driver was calculated and saved in the database table *base_score* that contains the columns:

Field	Type	Description
id	Integer, primary key	Unique ID
driver_id	Integer	ID of the driver
year	Integer	Year in which the score was calculated
career_score	Float	Overall score of driver's career up to the current season
season_score	Float	Score for the current season alone

Table 5.3: Base Score DB table

Only the *Base Score* could be used as a driver ranking as it is based on plenty of historical data and also has a strong correlation with the championship-winning variable. But as a driver is only as good as the car they drive, it is important to compare a driver's score against the score of their teammates over the years. For example, if a driver was racing against a rookie with a low *Base Score* and beating that rookie during the season it shouldn't be like the driver racing against another driver who has a high *Base Score* and lost the season by a small margin.

This was the final step of the DM and another two formulas were used, the first is just a summation of the drivers' *Season Base Scores* over their careers divided by the number of seasons they competed, given by the formula:

$$X_d = \frac{\sum_{i=1}^{N_d} S_{di}}{N_d} \quad (2)$$

Where:
d: the driver the score is being calculated
i: the *ith* season in the driver *d* career
X_d: the driver *d* overall career score
N_d: number of seasons the driver *d* had in their career
S_{di}: season score for the driver *d* in their *ith* season, obtained from (1)

It is important to distinguish between *Season Base Score* and *Base Score.* The first is a score calculating the attributes for a single season only whereas the latter is calculated using the entire driver's career stats.

The second formula is the most complex formula of the analysis and it includes another 3 attributes to the calculation:

- Score Factor: the square root of the driver's *Base Score* multiplied by their teammate's *Base Score*. This multiplication is squarely rooted because both the

multiplicand and the multiplier would have the same unit so the square root regulates the imbalance. It has a weight of 13 in the ponderate average but depending on the next attribute it may have a weight of 14.

- ➢ Inner Points Ratio: the ratio of how many points the driver scored by how many points the team scored. If the driver scored all the team's points in the season, then this ratio would be 1 whereas if the driver scored a third of the team's points this ratio is 0.333. It's important to note that on some occasions neither the team nor the driver scored points, in this case, the *Inner Points Ratio* is discarded to avoid division by zero and the *Score Factor* has a weight of 14 instead of 13.
- ➢ Races Ahead Ratio: number of races that the driver finished ahead of their teammate divided by the number of races they raced with the same car in that season. This is an important comparison as it shows the driver with greater consistency.
- ➢ Grid Ahead Ratio: number of races that the driver started ahead of their teammate divided by the number of races they raced with the same car in that season. It shows the driver with better raw speed.

The final formula is a ponderated average of the attributes above where the *Score Factor* has a weight of 13 when the team which the driver raced in that season scored one or more points or 14 when the team hasn't scored any points. This formula results in the final number on which the ranking will be based: the *Refined Score.*

As proposed, the *Refined Score* is, in the final sense, a way of quantifying driver talent and points out the greatest driver in history. Therefore it is only necessary to calculate the *Refined Score* for the drivers who have won at least one championship in their careers since it would be senseless to include someone who never won a championship in this list, and it is possible that drivers who always raced as a 'Driver B' for a certain team, meaning the other driver had the preference of development and strategy, would have a good *Refined Score* for driving with a great car for several seasons. A good example is Rubens Barrichello who raced for Ferrari against Michael Schumacher but Schumacher had all the preferences within the team and Barrichello never really aspired to win the championship yet won races, scored good points, etc. In addition to the drivers who have won at least one championship, the *Refined Score* was also calculated for drivers who are racing in 2022 as they can, in theory, win a championship in the future. The formula for the *Refined Score* is:

$$F_d = \frac{\sum_{i=1}^{N_d}\left(\sum_{b=1|P_{ti}\neq 0}^{M_{di}}\sqrt{X_{di}*S_{bi}}*13+\left(\frac{P_{di}}{P_{ti}}\right)+\left(\frac{A_{di}}{R_i}\right)+\left(\frac{G_{di}}{R_i}\right)+\sum_{b=1|P_{ti}=0}^{M_{di}}\sqrt{X_{di}*S_{bi}}*14+\left(\frac{A_{di}}{R_i}\right)+\left(\frac{G_{di}}{R_i}\right)\right)}{16*N_d*\sum_{i=1}^{N_d}M_{di}} \quad (3)$$

Where:
d: the driver the refined score is being calculated
i: the *ith* season in the driver *d* career
b: driver's teammate which is being compared with
t: team for which the driver *d* and their teammate *b* raced in the *ith* season
F_d: refined score
N_d: number of seasons in the driver's career
M_{di}: number of teammates the driver had in the ith season
X_{di}: career base score of the driver *d,* obtained from (2)

S_{bi}: base score of the driver's teammate *b,* obtained from (2)
P_{di}: points the driver *d* scored in the races they raced against the teammate *b* in the *ith* season
P_{ti}: points the team *t* scored in the *ith* season
A_{di}: number of races the driver *d* finished ahead of their teammate *b* in the *ith* season
G_{di}: number of races the driver *d* started in the grid ahead of his teammate *b* in the *ith* season
R_i: number of races the driver *d* competed for the team *t* in the *ith* season

As the calculations for the *Refined* Score took some time to run it was a good idea to store those values in another table called *refined_scores* in the SQLite database. This table contains 50 rows and the columns:

Field	Type	Description
id	Integer, primary key	Unique ID
driver_id	Integer	ID of the driver
refined_score	Float	Driver's refined score
won_championships	Boolean	TRUE if the driver has won championships over their career and FALSE if not

Table 5.4: Refined Scores DB table

6. Results

After obtaining the *Refined Scores* it was just a matter of sorting them from the highest to the lowest and filtering only the drivers who won championships as the *Refined Score* was also calculated for the driver enlisted for the 2022 season but were never champions.

According to the techniques and technologies applied to the data used in this analysis, the ranking of the greatest drivers in F1 history is as seen in *Table 6.1*:

Ranking Position	Driver	Refined Score	Ranking Position	Driver	Refined Score
1	Lewis Hamilton	0.4002253803	18	Nelson Piquet	0.263466511
2	Alain Prost	0.3813199806	19	Jim Clark	0.2582346297
3	Juan Manuel Fangio	0.3703172219	20	Mike Hawthorn	0.2557521468
4	Ayrton Senna	0.3503616093	21	Jenson Button	0.2486126619
5	Michael Schumacher	0.3377811846	22	Graham Hill	0.2479825686
6	Nico Rosberg	0.3238521903	23	Jochen Rindt	0.2420361198
7	Alberto Ascari	0.3205399785	24	Nigel Mansell	0.2413399301
8	Nino Farina	0.3157836326	25	Phil Hill	0.2349554564
9	Sebastian Vettel	0.3144763176	26	Jack Brabham	0.2285978506
10	Fernando Alonso	0.28946863	27	Jody Scheckter	0.213432992
11	Max Verstappen	0.2867344657	28	Keke Rosberg	0.2113506296
12	Niki Lauda	0.2855730313	29	Jacques Villeneuve	0.2112590096
13	Jackie Stewart	0.2803445337	30	Emerson Fittipaldi	0.1996404549
14	Mika Hakkinen	0.2792723704	31	John Surtees	0.185949284
15	Kimi Räikkönen	0.2779689993	32	James Hunt	0.1837095709
16	Denny Hulme	0.2749815759	33	Mario Andretti	0.1801142112
17	Damon Hill	0.2652064925	34	Alan Jones	0.1770245309

Table 6.1: Historical Ranking

Additionally, a second ranking was generated with the current drivers who never won a championship and are not rookies, as rookies would have a score of 0, and this ranking might serve as a predictor of who has greater chances of becoming a great driver in the future. This predictory list is as in *Table 6.2*:

Ranking Position	Driver	Refined Score
1	Valtteri Bottas	0.3097547288
2	Charles Leclerc	0.2631761081
3	Daniel Ricciardo	0.2226104268
4	Sergio Perez	0.1963956668
5	Lando Norris	0.186181167
6	Pierre Gasly	0.1844233293
7	Esteban Ocon	0.1774379937
8	Carlos Sainz	0.1734198612
9	George Russell	0.1725648143
10	Lance Stroll	0.1668386768
11	Robert Kubica	0.1611138853
12	Mick Schumacher	0.1449449965
13	Antonio Giovinazzi	0.1388568375
14	Yuki Tsunoda	0.109752144
15	Nicholas Latifi	0.0874302843
16	Nikita Mazepin	0.06824045104

Table 6.2: Predictor Ranking

There are further considerations about both rankings that will be debated and detailed with charts in the next section of this report.

7. Conclusion and Discussion

7.1. Predictor

The first topic that will be discussed is the prediction table. As we are trying to predict future talents we have to take into consideration that some drivers are in the final years of their careers. As seen in the EDA, the average season competed by a driver is 5.1 and the median is 4. So from our top 5: Valteri Bottas, Charles Leclerc, Daniel Ricardo, Sergio Perez and Lando Norris we predict that **Charles Leclerc** or **Lando Norris** will figure in the champions hall of fame soon. The reasons to discard the others are:

- Valteri Bottas: despite topping the *Refined Score* list of non-champions, in 2022 he'll be racing his 10th championship which means he is reaching the end of his career and won't be so competitive anymore. Additionally, he will be racing for Alfa Romeo, which is a mid-field team and probably won't be competing for victories;
- Daniel Ricciardo: Similarly to Valteri Bottas, Ricciardo is reaching the end of his career and will be competing in his 12th season in 2022. He is driving for McLaren, a traditional team, but they don't have title aspirations at the moment. Even if they were, he only scored 70% of the points of his teammate in 2021 and there are no indications that in 2022 it'd be different.
- Sergio Perez: He is also aiming toward the end of his career, 2022 will be the 12th season for Perez. He will be racing for a title-fighting team, but against the current champion, Max Verstappen. It will be the third season Perez and Max are racing for the same team, in the first year Perez scored only 58% of Max's points and in 2021 only 48%. There is no indication that Perez will be fighting for a title again.

Even though the aim of this project was not to make future predictions, the knowledge used for the analysis is very similar to what recruiters would be interested in when looking for a talented driver thus there is a good level of confidence in the prediction that Leclerc or Norris will win a championship in the following years.

7.2. Ranking

When talking about the historical ranking, the main accomplishment of this analysis, it is interesting to compare it with other rankings so we can discuss its accuracy. Considering only drivers who won championships from the above-mentioned works from Amazon AWS and Eichenberger, *Table 7.1* and *Table 7.2* show both top 10 lists, respectively:

Amazon	
Ranking	**Name**
1	Ayrton Senna
2	Michael Schumacher
3	Lewis Hamilton
4	Max Verstappen
5	Fernando Alonso
6	Nico Rosberg
7	Sebastian Vettel
8	Jenson Button
9	Alain Prost
10	Jacques Villeneuve

Table 7.1: Amazon's Top 10

Eichenberger, R. and Stadelmann	
Ranking	**Name**
1	Juan Manuel Fangio
2	Jim Clark
3	Michael Schumacher
4	Jackie Stewart
5	Mike Hawthorn
6	Fernando Alonso
7	Alain Prost
8	Graham Hill
9	Emerson Fittipaldi
10	Kimi Raikkonen

Table 7.2: Eichenberger's Top 10

It was also considered the ranking made by three different media sources from three different countries, to avoid biases. The rankings were taken from BBC TopGear (2022), Auto Motor und Sport (2022) and the Globo Esporte's blog Voando Baixo (Lopes, 2022), their lists are described in *Table 7.3, Table 7.4* and *Table 7.5*:

BBC TopGear	
Ranking	**Name**
1	Lewis Hamilton
2	Juan Manuel Fangio
3	Jim Clark
4	Ayrton Senna
5	Michael Schumacher
6	Jackie Stewart
7	Stirling Moss
8	Alain Prost
9	Niki Lauda
10	Fernando Alonso

Table 7.3: BBC's Top 10

Auto Motor und Sport	
Ranking	**Name**
1	Juan Manuel Fangio
2	Michael Schumacher
3	Lewis Hamilton
4	Sebastian Vettel
5	Alberto Ascari
6	Alain Prost
7	Jim Clark
8	Ayrton Senna
9	Jackie Stewart
10	Fernando Alonso

Table 7.4: Auto Motor's Top 10

Globo Esporte	
Ranking	**Name**
1	Lewis Hamilton
2	Ayrton Senna
3	Juan Manuel Fangio
4	Jim Clark
5	Alain Prost
6	Michael Schumacher
7	Niki Lauda
8	Jackie Stewart
9	Fernando Alonso
10	Nelson Piquet

Table 7.5: Globo's Top 10

Note that from the top 10 ranking generated in this report 6 drivers also appear in all specialised media rankings, namely: Lewis Hamilton, Alain Prost, Juan Manuel Fangio, Ayrton

Senna, Michael Schumacher and Fernando Alonso. In this project's top 5, all drivers figure in the top 10 lists of all specialized media and it is a good indication that our AI reached its goal to a certain extent and provided us with a reasonable result. In *Figure 7.1* a Venn Diagram illustrates the intersection between the 4 lists:

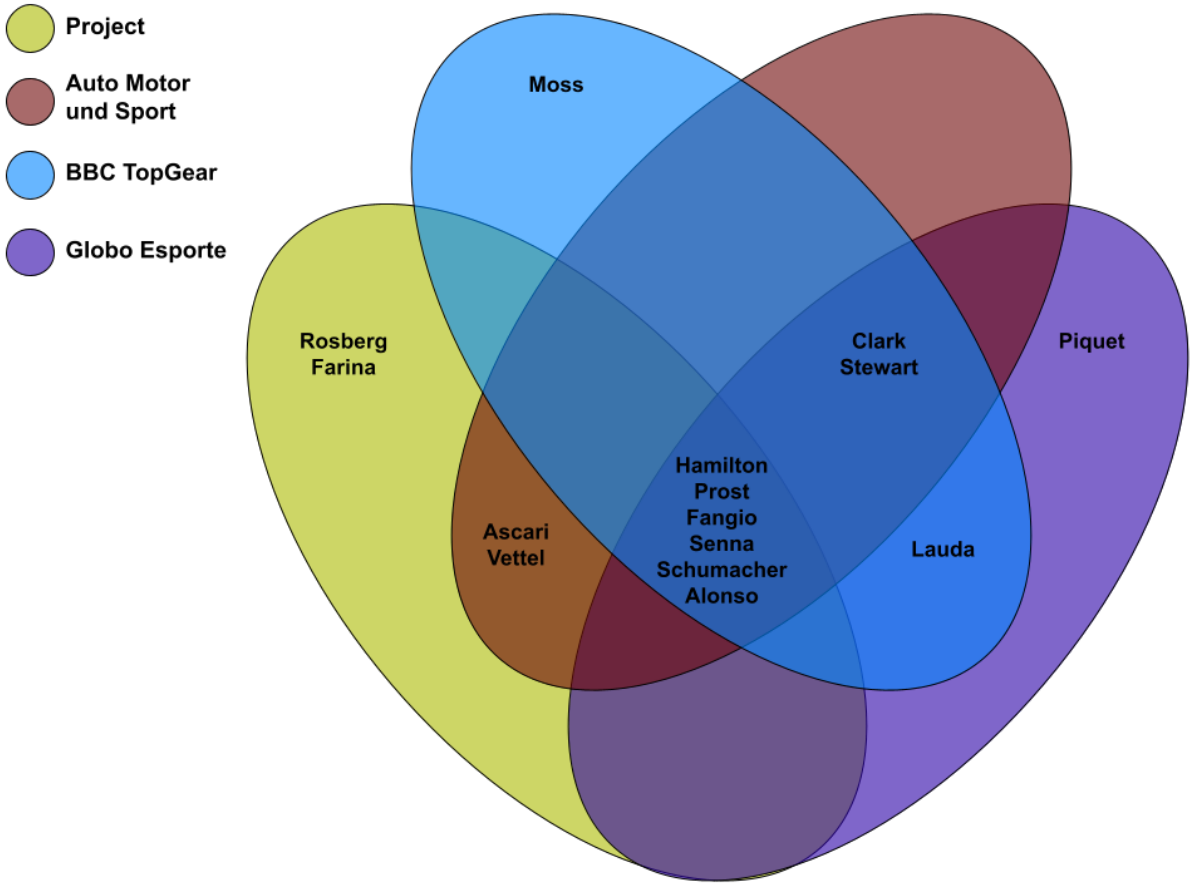

Figure 7.1: Top 10 Venn Diagram

When comparing this project's top 10 with the revised works, only 3 drivers figure in all lists but there is a simple reason: the Amazon project used post-1983 data, which automatically excluded some drivers who featured in this project's top 10, they are Juan Manuel Fangio, Alberto Ascari and Nino Farina. Whilst Eichenberger's top 10 was created in 2009, before Lewis Hamilton, Sebastian Vettel and Nico Rosberg won their championships and all three of them are in this analysis's top 10 list.

Another point to be made is that, as seen in *Figure 7.2*, the *Base Scores* tend to decrease towards the end of the drivers' careers.

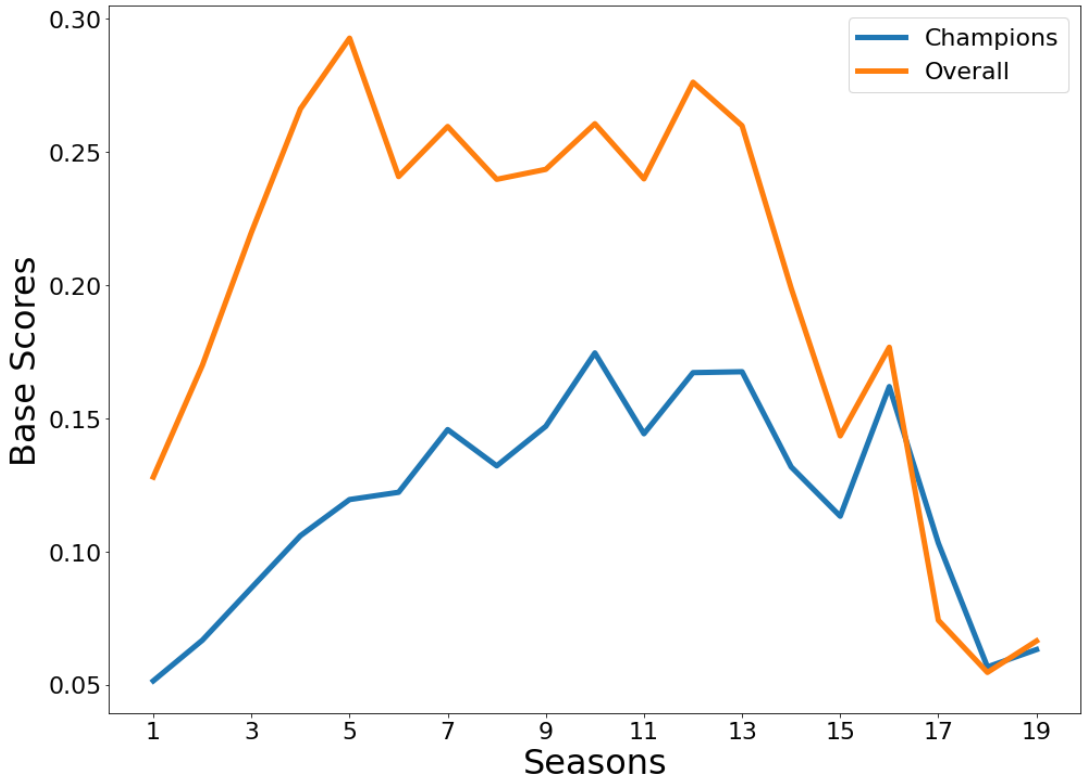

Figure 7.2: Season Base Scores over drivers' careers

This implicates in:

- Ayrton Senna probably would have a better score if he was still alive. He was still racing for a championship-fighting car and his score was still increasing;
- Michael Schumacher would also have had a better score if he hadn't come back from retirement for his last three seasons, which also influenced the next topic;
- Nico Rosberg is number 6 in the ranking because he raced three years against Michael Schumacher and was ahead of Schumacher all three years. He also raced very closely to Lewis Hamilton for Mercedes for four seasons, even winning one championship in one of them;
- Lewis Hamilton's score will probably drop in the following years as he reaches the end of his career and will, probably, be racing for less competitive teams;
- Max Verstappen was 12th on our list but he will be only 25 years old in 2022 and has, at least, another 10 years of top performance. We can expect him to climb the ranking as years go by;

The last observation to be made is how important it was to compare the drivers against their teammates. As seen in the chart above the *Season Base Score* is never higher, on average, than 0.3 even for championship-winning drivers but the final ranking with the *Refined Score* contains nine drivers with scores higher than 0.3.

To help picture how important the teammate's comparison was, we have two charts below: in *Figure 7.3* a radar chart shows how close the top 5 drivers are. In various attributes, the five were very similar, and Fangio actually beats Hamilton in most of them but probably didn't take the highest spot in the ranking because he had teammates with lower

scores than Lewis Hamilton's teammates. The second chart, *Figure 7.4,* is a bar chart, also for the top 5 drivers, with their *Career Base Scores*, in which Fangio is ahead of Lewis Hamilton and Michael Schumacher would be in third.

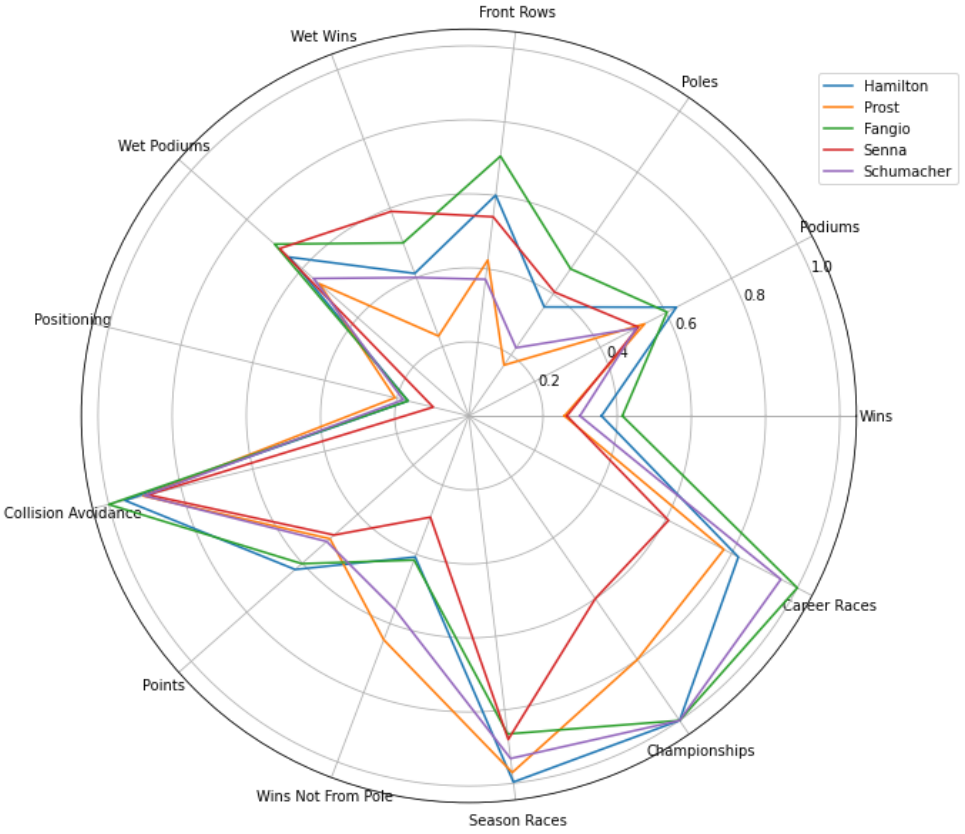

Figure 7.3: Top 5 Base Score Radar Chart

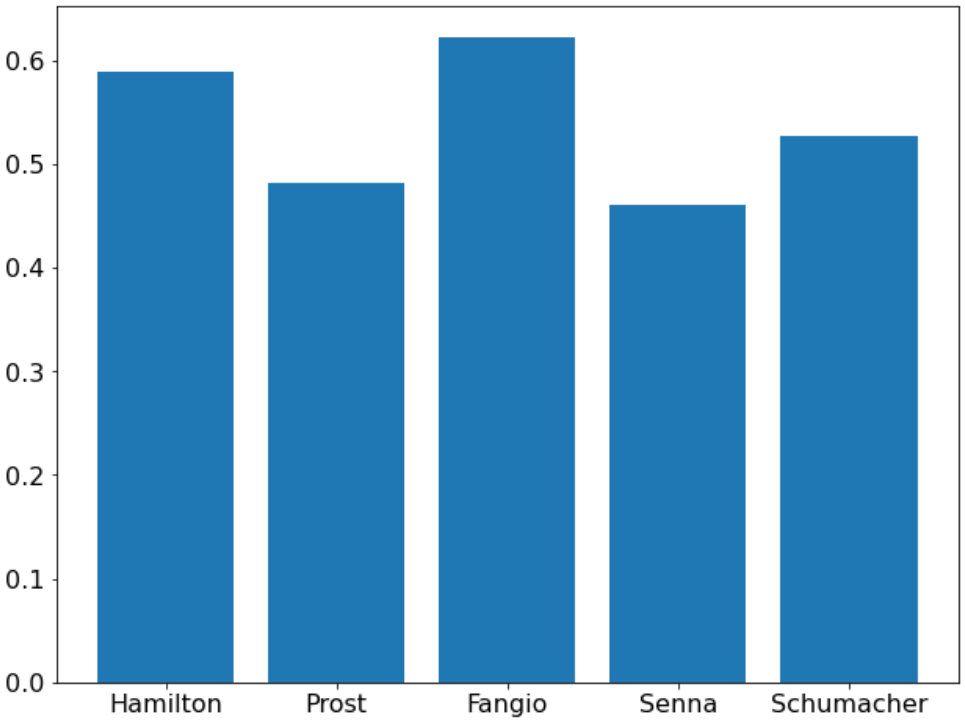

Figure 7.4: Top 5 Careers Base Scores

8. Further Development

This analysis should be performed yearly, at the end of each season, to re-calibrate the scores of the drivers that are still running and to update all the datasets necessary for the analysis. Performing this analysis periodically would also help to verify the precision of the predictory ranking.

In terms of improvements some additional developments could be implemented in the analysis which could have been done with more time:

- Apply Wisdom of Crowds: “Diversity and independence are important because the best collective decisions are the product of disagreement and contest, not consensus or compromise.” (Surowiecki, 2004, p. Intro XIX). As Giles (2007) explains in his article, Wikipedia is as accurate as the notorious printed encyclopedia Brittania. That's due to the Wisdom of Crowds: the more people contribute to an opinion, the more that opinion gets closer to the reality, this was also demonstrated by Galton (1907) when he observed visitors at a carnival attempting to guess the weight of a bull and the more guesses were registered the closer it was to the actual weight. Using scraping scripts it would be possible to scrape ranking and opinion polls from specialised F1 websites and forums. It wouldn't replace this analysis but could be an extra attribute or a calibrator for the AI algorithms.
- Refine Attributes: there are an infinite number of statistical attributes that could be added to the analysis, and the more the better. Finding more data, such as drivers' ages, could help in making a more accurate ranking.
- Quantify Wet Weather: it is well known among F1 fans that during rain-affected races it is easier to identify raw talent in drivers. But in this analysis, both attributes that were related to rain-affected races were very poorly correlated to the dependent variable. Ideally, more time should be spent in trying to better quantify these attributes and ensure that they had a higher weight in the formulas.

9. References

Auto Motor und Sport (2022) *Formel 1: Wer ist der beste Fahrer aller Zeiten?*. Available at: https://www.auto-motor-und-sport.de/formel-1/bester-f1-fahrer-aller-zeiten/ [Accessed 26 April 2022].

BBC TopGear (2022) *Here are the 10 best ever F1 drivers*. Available at: https://www.topgear.com/car-news/formula-one/here-are-10-best-ever-formula-1-drivers [Accessed 26 April 2022].

Eichenberger, R. and Stadelmann, D. (2009) 'Who is the best Formula 1 driver? An Economic Approach to Evaluating Talent', SSRN Electronic Journal. doi: 10.2139/ssrn.1017292.

F1.com (2022) *Archive 1950-2016*. Available at https://www.formula1.com/en/results/archive-1950-2016.html [Accessed 02 January 2022].

F1 Fansite (2022) *F1 Champions: All Time F1 World Title List*. Available at: https://www.f1-fansite.com/f1-results/f1-champions/ [Accessed 02 January 2022].

F1 Fansite (2022) *F1 Drivers archive of racers*. Available at: https://www.f1-fansite.com/f1-drivers/ [Accessed 02 January 2022].

F1 Fansite (2022) *Results: The All Time F1 Championship Seasons*. Available at https://www.f1-fansite.com/f1-results/ [Accessed 02 January 2022].

Fagen, R., 1974. Selective and evolutionary aspects of animal play. The American Naturalist, 108(964), pp.850-858.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). 'The KDD process for extracting useful knowledge from volumes of data', *Communications of the ACM*, 39(11), pp.27-34.

Galton, F. (1907). 'Vox populi (the wisdom of crowds)', *Nature,* 75: 450-451.

Giles, J. (2005). 'Special Report Internet encyclopaedias go head to head', *Nature*, 438(15), pp.900-901.

Lopes, R. (2022) 'F1: top 10 dos melhores pilotos da história na visão do Voando Baixo', *Voando Baixo*, 9 January 2022. Available at: https://ge.globo.com/motor/formula-1/blogs/voando-baixo/post/2022/01/09/f1-o-top-10-dos-maiores-pilotos-da-historia-na-visao-do-voando-baixo.ghtml [Accessed 26 April 2022].

Mauboussin, M. (2012) *The Success Equation: Untangling Skill and Luck in Business, Sports, and Investing.* Boston: Harvard Business Review Press.

Mercedes-AMG PETRONAS F1 Team (2020) [Twitter] 17 August. Available at: https://mobile.twitter.com/MercedesAMGF1/status/1295423365105225729 [Accessed 8 April 2022].

Scavetta, R. and Angelov, B. (2021) *Python and R for the Modern Scientist*. Sebastopol: O'Reilly Media.

Shevlin, H., Vold, K., Crosby, M. and Halina, M. (2019) 'The limits of machine intelligence: Despite progress in machine intelligence, artificial general intelligence is still a major challenge', *EMBO Reports*, 20(10), p.e49177.

Smedley, R., Wise, C., Enkhbayar, D., Price, G., Cheng, R. and Yang, G. (2020) 'The fastest driver in Formula 1', AWS Machine Learning Blog, 20 August. Available at: https://aws.amazon.com/blogs/machine-learning/the-fastest-driver-in-formula-1 [Accessed 16 December 2021].

Surowiecki, J. (2004) *The Wisdom of Crowds*. NewYork: Anchor Books.

Stats F1 (2022) *Constructors*. Available at: https://www.statsf1.com/en/constructeurs.aspx [Accessed 02 January 2022].

Stats F1 (2022) *Drivers*. Available at: https://www.statsf1.com/en/pilotes.aspx [Accessed 02 January 2022].

Stats F1 (2022) *Seasons*. Available at: https://www.statsf1.com/en/saisons.aspx [Accessed 02 January 2022].

10. Appendices

10.1. Project Plan

10.1.1. Proposed Plan (October 2021)

Step	Timeline	Description
Define data needed	Mid Nov 2021	Define what information will be necessary for the project
Gather data source	Late Nov/Early Dec 2021	Find which websites, forums, etc have the suitable data
Load the datasets	Early Jan 2022	Develop the programs that will read the internet from online sources and save them locally to be analysed
Prepare and transform data	Early Feb 2022	Clean the datasets, removing unnecessary data, adding missing information and merging duplicates/redundant information
Data Mining	Late Mar 2022	Use a learning machine algorithm to analyse the data and respond with the results
Data Interpretation	Early Apr 2022	Interpret the data from the data mining step and build reports and charts accordingly
Final Report	Early May 2022	Prepare and submit the final report, containing all the documentation about the course of the project

10.1.2. Revised Plan (December 2021)

Step	Timeline	Description	Status
Define data needed	Mid Nov. 2021	Define what information will be necessary for the project	Completed, continuously reviewed.
Gather data source	Late Nov./Early Dec. 2021	Find which websites, forums, etc have the suitable data	Completed, continuously reviewed.
Load the datasets	Late Dec. 2021	Develop the programs that will read the internet from online sources and save them locally to be analysed	Completed, continuously reviewed.

Prepare and transform data	Early Jan. 2022	Clean the datasets, removing unnecessary data, adding missing information and merging duplicates/redundant information	In progress.
Data Mining	Late Jan. 2022	Use a learning machine algorithm to analyse the data and respond with the results	
Data Interpretation	Late Mar. 2022	Interpret the data from the data mining step and build reports and charts accordingly	
Final Report	Late Apr. 2022	Prepare and submit the final report, containing all the documentation about the course of the project	

10.2. Reflective Journals

10.2.1. November 2021

What?

For personal reasons I couldn't spend much time on the project this month.

On top of that, deliverables for other modules took up most of the time I manage to allocate for college work.

The two main achievements this month were :

- finished the list of the datasets I'll need for the project and finding the source for them:

Tems Lineups (F1 archive)

Qualy Results (F1 archive)

Races Results (F1 archive)

Championship Results (F1 archive)

Drivers Results Per Race (F1 Fan Site)

Rain affected races (F1 Fan Site)*

Rain affected qualys (F1 Fan Site)*

**Will need manual adjustments*

- made a draft of how the data will be quantified in order to determine a ranking:

First: find the 'gross stats' per driver (wins/races, poles/races, wins/front-row-starts, podiums/front-row-starts, collisions/races, wins/rain, poles/rain, career points, career championships), and award them scores based on those stats;

Second: eliminate drivers who never won championships from the list, this is to prevent a driver who, by chance, had good stats (in a short career, for example);

Third: compare each driver against their teammates (as they drive the same car) per season (using the score from the first step to find a 'teammate difficulty'). The 'teammate difficulty' must be increased if the driver is a rookie and/or if it's their first car driving for that team;

So What?

It was vital for the project to have the structure designed, now the focus on the technologies needed can be optimized and hopefully catch up with the schedule.

Now What?

Midpoint presentation is the main goal for this month. Have to focus on the literature review and the report/proposal.

Alongside that will start scraping some data and study about the technologies needed for the project.

Only had a half decent talk with the supervisor on 30/11, in which we agreed that time is the main concern for this project. Definitely have to find a way of working closer with her.

10.2.2. December 2021

What?

There was good progress in the project in the first half of the month. Compared to other months it was probably the most productive.

The second half of the month the main focus was the Mid Point report and presentation, the Christmas and Exam period took over the schedule, but I'm on track.

The two main achievements this month were :

- Almost finished with the two formulas that are the heart of the analysis: drivers base scores and drivers refined scores. A friend who's a PhD in Engineering is giving some help with the mathematical accuracy of the formula. They may suffer some changes, add or remove parameters, but they have their final form.
- Made the scraping programs (in R) and all datasets are already downloaded.
- Made the preparers (in Python) to prepare the datasets and save the data in a SQLite database

So What?

The development seems on a good track. I'm weekly reporting to the supervisor who's then giving me feedback. This method has improved the development as I have a weekly target, rather than monthly.

Once the data is prepared, the analysis is just a matter of running the formulas on the data.

Now What?

As soon as exams are over, I hope to have the data prepared by the end of the month so I can start doing a ML on the data.

There's a lot of learning going on as both Python, SQLite and R are new to me, but I'm enjoying the process. Will make use of the break until semestre 2 to watch as most lessons I can.

10.2.3. January 2022

What?

With the exams out of the way I managed to dedicate a good time to the project. Most of the time was spent on Youtube videos about SQLAlchemy and Python tutorials.

The database was recreated using SQLAlchemy and a lot of Python scripts were written to validate and cross check the data among the datasets.

The final analysis is almost done, only the refined score script isn't completed.

So What?

After some thoughts I decided that the formula is gonna change, using AI techniques I'll identify which attributes seem to contribute more towards a world title and then I can onderate the weights of each attribute. This won't change the coding of the formula as I just have to add the weights and run again.

Now What?

Everything seems on good track. This month I'll dedicate myself to learning AI techniques that can help figuring out the weights of each attribute.

On the top of that I'm also generating some diagrams to be included on the final report to help with data visualization. This is also helpful for me to understand better the attributes I'm using on both formulas and will also save some time when I'm doing the final report.

10.2.4. February 2022

What? I decided to use a Random Forest algorithm to identify the attributes that played most part in a driver being champion. Speaking with the supervisor it came to my attention that other algorithms had to be considered and I was requested to study deeper into the subject.
So What? I have to consider other algorithms to be used and I have to learn about metrics that will help to decide which algorithm will be the final. Some data visualization is also being generated to be included in the final report.
Now What? The coding part of the project is almost completed, I just have to build a proper justification of which algorithm is gonna be used and tune the results into the formula, and then build the final ranking. This month we also had to create the showcase profile and find the pictures that are gonna be used on the college website. This was more challenging than it seems as a lot of F1 pictures are very expensive. I eventually found a free one and had to pay for the second, but I guess it was worthed.

10.2.5. March 2022

What? After discussion with the supervisor and evaluating 3 different algorithms, random forest was the best fit for the project. The formula was finalized and the code finished to build the ranking.
So What? The coding part of the project is concluded. Some more scripts might be necessary as the report is written, and I hope to have everything put into a Google Notebook.
Now What? The report has started, I'm calculating a report with over 15k words, which means there's a lot of work to do. It feels a little bit off track as there's only one and a half months to the delivery date. I plan to share the report with the supervisor and request weekly feedback as our weekly conversations have helped a lot during the build phase of the project.

10.2.6. April 2022

What?

April was a month mostly busy with TABA's for other modules so the report didn't evolve as much as I wished.

Before starting the report I put the code on a Google Colab Notebook for easy sharing.

In the report front the introduction and referencing are completed, despite some more referencing might arise while writing the other bits. I also completed a draft with key points that must be present in the report, so the writing would be facilitated.

Apart from the report we also must submit a banner for which I already put together a few layout ideas and I'm evaluating them.

So What?

The schedule is still feeling tight but the development definitely seems to be on track to be delivered on date. Not having the other modules to worry about should allow a speed improvement.

Now What?

The next 10 days will be dedicated to finishing the report, as I have the ideas clear in my mind and a draft with the key points I believe is doable. Apart from writing the report some Exploratory Data Analysis data/charts must be generated to be included in the report, it'll probably take a day or two.

Concurrently I'll slowly work on the banner, I'm not a great designer but it's just a simple image and 30-45min a day working on it should be enough to have it ready in two weeks.

The final 5 days after finishing the report will be dedicated to reviewing it, there are tools like Grammarly that I'd like to use to improve readability and ensure the grammar is accurate.

10.3. Resources Links

10.3.1. Github

For those who want to have bigger control and check/modify the scrapers the Github contains the entire project. It requires the environment to be setup: RStudio, Python, SQLite and the packages but it isn't too difficult:

https://github.com/adolfotcar/formula-1

10.3.2. Google Colab Notebook

For the general audience the Google Colab Notebook is ideal because it is possible to see all the analysis without having the environment installed on your computer. It only contains the Python part of it but it's enough as you don't need to scrape all the data again:

https://colab.research.google.com/drive/1VRg_wmoleik729wKcWyRxv8TI4bv8C7b?usp=sharing