Project Report

# GitHub URL

https://github.com/jphannaford/UCDPA\_justinhannaford

# Abstract

This Introduction to Data Analytics project will focus on Formula One racing between 1950 to 2022 inclusive. Given that this timeframe captures the sport in it’s entirety, the analysis will look to uncover trends in subjects like the global reach of the sport, safety, reliability and individual driver performance over the life of the sport itself.

# Introduction

Formula One racing is widely seen as the pinnacle of motor sport and has always interested me. My home city of Adelaide hosted the Australian Grand Prix between 1985 and 1996 and growing up, I attended many races.

Formula One is now a global sport, with annual racing at many historic tracks like Silverstone, Monza and most famous of all, Monaco. Interwoven within the fabric of the sport is the often long standing rivalry between teams and drivers. Sometimes that rivalry can boil over intra-team as we saw with 2 giants of Formula One, Ayrton Senna & Alain Prost, when they both raced for McLaren in 1989.

There is a rich catalog of 70 plus years of innovations that have eventually fed into the consumer market. Inventions like active suspension and hybrid technology have been widely adopted by the automotive industry and today are seen as standard.

# Dataset

The following 3 datasets were chosen for my analysis:

1. Formula One statistics encompassing historical driver, race, seasons, circuit and constructor data. This database covers Formula One’s inaugural season (1950) through to the completion of last season (2022) and is freely available for non-commercial usage. (<http://ergast.com/mrd/>). This is a comprehensive database maintained by an Formula One enthusiast who provides both API access and csv file download functionality.
2. Formula One Deaths (Wikipedia). This table contains all drivers who competed in Formula One and have died in driving accidents. These accidents are both Grand Prix and other races.
3. Formula One World Champions (Wikipedia). This table contains a list of all divers who have won a Formula One World Championship.

# Implementation Process

Python 3 is a powerful and flexible open-source programming language. Leveraging the structured workflow of Juypter Notebook my first step was to import all relevant Python packages that I needed within the project. Next, I imported 7 distinct datasets relating to Formula One: 2 via web scrapes and 5 csv files. Where I imported from the web, I created new DataFrames from the table. The following phase was dedicated to inspection and refinement of data.

(Web scrape) Fatalities - A list of every Formula One driver who has died in a motor racing accident

I checked the number and type of columns and if there were any null or duplicate values. Reviewing the first few columns I noticed that ‘Date of accident’ was in MM DD, YYYY format so I split this value using the comma as a delimiter to leave just the year. I then changed the type from object to int. I also observed that ‘Driver’ had the country in brackets after it so I split it using open bracket delimiter to leave just the name and reapplied.

The ‘Event’ column shows all Formula One driver deaths irrespective of the types of race, so non-Grand Prix races were filtered out and saved. Finally I dropped all columns except for ‘Driver’ and ‘Date of accident’ and reviewed ‘f1\_fatalities’

(Web scrape) World Champions - A list of every driver who has won the Formula One world title

I inspected the number and type of columns and checked if there were any null or duplicate values. Looking at the head and tail of the DataFrame, I noticed that there were 2 header and 2 footer rows. This was due to merged cells within Wikipedia, so I dropped one header row and both footer rows and reviewed to ensure that the issue was fixed.

As with ‘f1\_fatalities’ there was extra data within the ‘Season’ and ‘Driver’ columns so I split by open square bracket delimiter and reapplied to the DataFrame. Then I dropped all columns except for ‘Season’ and ‘Driver’, grouped by ‘Driver’ to create a new column, ‘TimesWorldChampion’, sorted descending and reviewed output of ‘f1\_world\_champions’

(csv) Circuits – Contains every race track, host city and country within Formula One

I checked the number and type of columns and if there were any null or duplicate values. I also dropped all columns except for ‘circuitId’, ‘name’ and ‘country’ and reviewed ‘f1\_circuits’

(csv) Status – A mapping table used for race events mechanical failure, accidents and disqualifications etc.

I checked the number and type of columns and if there were any null or duplicate values and reviewed ‘f1\_status’

(csv) Drivers – A full list of every driver who has ever driven in a Formula One race

I checked the number and type of columns and if there were any null or duplicate values. I also dropped all columns except for ‘driverId', 'forename', 'surname' & 'nationality'. I then added a new column called ‘fullname’ by concatenating 'forename' and 'surname’ and moved it’s position to between ‘surname’ and ‘nationality’.

To ensure that all values in ‘f1\_world\_champions’ ‘Driver’ column were in ‘f1\_drivers’ , I merged this DataFrame into ‘Drivers’ using a right join. This check showed that 2 World Champions did not exist in ‘f1\_drivers’ ‘fullname’. I identified these differences and renamed in ‘fullname’ as per ‘Driver in ‘f1\_world\_champions’. After this, I merged both DataFrames again, this time on an outer join and dropped the ’Driver’ column.

I reviewed ‘f1\_drivers’ and checked for NaN values. As these values existed, I replaced with a zero and changed the type to int and reviewed the final output.

Races (csv) – A comprehensive list of every race between 1950 to 2022 inclusive.

To begin with I merged ‘f1\_circuits’ DataFrame into ‘Races’ on an inner join. I ensured that I used suffixes as both DataFrames have ‘name’ columns. I checked the number and type of columns and if there were any null or duplicate values, dropping all columns except 'raceId', 'year' & 'country' and reviewed the final output.

Results (csv) – Every result, including accidents and retirements, from every race.

I merged ‘f1\_status, ‘f1\_drivers’ and ‘f1\_races’ into ‘f1\_results’ using ids on an inner join. I checked the number and type of columns and if there were any null or duplicate values, dropping all columns except 'positionText' , 'positionOrder', 'status', 'fullname', 'nationality' & 'year' columns and reviewed the output.

Although not undertaken, there are opportunities for machine learning within this data project. For example, supervised regression machine learning could be utilised to determine likelihood of a driver winning a race or World Championship. Or it could be leveraged to predict how often a driver is in involved in an accident or race retirement.

# Results

# Chart 1: Number of Formula One races Per Year

# Table Description automatically generated with medium confidence

# Chart Description automatically generated

# Creating a new DataFrame, ‘f1\_numberofraces’, I grouped ‘f1\_races’ by ‘year’ to create a new column called ‘NumberOfRaces’. I then plotted a line on this grouping and year, adding a line for the average number of races per year using numpy. To further demonstrate Insight #1, I displayed the countries that held races in both 1950 and 2022.

# Chart 2: Number of Formula One retirements per decade

# Chart, bar chart Description automatically generated

# I created a new DataFrame called ‘f1\_retirements’ by filtering ‘f1\_results’ on ‘positionText’ and ‘status’. I then grouped by year, creating a new column ’NumberOfRetirements’ and dropped all seasons post 2019 as there are only 3 years available in the 2020’s. Finally, I plotted a bar graph based on number of retirements and year collated per decade.

# Chart 3: Top 5 World Champions versus all other World Champions

# Chart, pie chart Description automatically generated

# Graphical user interface, text, application, email Description automatically generated

# Graphical user interface, text, application, email Description automatically generated

# To determine if there were multiple World Champions, I used a for loop to show all winners equal to or more than 4 times. Creating a new list, ‘f1\_world\_champions\_top\_5’, I summed up the number of times these 5 drivers had won titles versus the rest. To enable plotting, I converted the list to a DataFrame and plotted a pie chart based on percentage. To further demonstrate Insight #2, I counted all drivers and single time World Champions.

# Charts 4: Accidents and Fatalities per decade

# Chart, line chart Description automatically generated Chart, line chart Description automatically generated

# DataFrame ‘f1\_accidents’ was created from ‘f1\_results’ filtering on ‘positiontext’, ‘year’ and ‘status’. I grouped by year, creating a new column ’NumberOfAccidents’ and dropped all seasons post 2019. Finally, I plotted a line graph based on number of accidents and year collated per decade.

# To facilitate the Fatalities graph, I grouped ‘f1\_fatalities’ by ‘Date of accident’ to create a new column 'NumberOfFatalities’. As per the accidents graph, I plotted a line graph based on number of fatalities and year collated per decade.

# Charts 5: Most Formula One wins and best conversion rate of wins in Formula One

# Chart, funnel chart Description automatically generated

# Chart, bar chart Description automatically generated

# A new DataFrame ‘f1\_individual\_wins’ was created by filtering ‘f1\_results’ where ‘positionOrder’ = 1. This new DataFrame was grouped by ‘fullname’ to generate the sum of race wins per driver. Renaming this new column to ‘NumberOfWins’, I sorted descending to show the driver with the most wins. Plotting a horizontal bar chart on driver and number of wins showed the top 10 performing drivers.

# The second chart includes number of wins but also number of races as per ‘f1\_experience’ which is generated by grouping ‘f1\_results’ by ‘fullname’. To complete the data required, I captured the race wins per number of races within ‘f1\_conversion\_rate’ ‘Percent’ column. To ensure that the data was meaningful I had a qualification of a minimum 20 races, then dropped all drivers except the top 10 conversion rate. I plotted a horizontal bar chart showing the conversion rate at the top in descending, with number of wins and number of races.

# Insights

As part of this data analysis and mapping I was able to identify the following trends and insights within Formula One:

1. The inaugural season of 1950 had 7 races, 6 of which were hosted in Europe. Season 2022, had a record 22 races, of which only 10 were European based. This shows that Formula One has grown from small European centric roots into a truly global sport.
2. Formula One has always been about ingenuity, progress and development of cutting edge technology. The advent of Turbo engines in the late 1970's, with usage peaking in the mid 1980's is a prime example of this development. Wikipedia notes that, although exceedingly quick, these engines had appalling reliability and were eventually banned from 1989 onwards. This graph evidences Formula One's increased reliability from the 1980's onwards.
3. There have been 73 seasons of Formula One; each season crowning a new World Champion. Analysis shows that there have been 855 drivers to have taken part in Grand Prix racing. Of these drivers, 17 have won single World Championship titles. The best performing top 5 drivers of all time have won 28 World Championship titles or a total of nearly 38%. This shows that only very few drivers reach ultimate success and even further, attain legend status within the sport.
4. The first chart shows a surge in accidents during the 1970’s. The second chart shows a peak in fatalities in the 1960’s. Comparing the two charts, we can see that there has been a dramatic fall in the number of Formula One deaths since the 1960’s, however, there has been a marked increase in accidents since this decade that has remained high to this day. This demonstrates that although Formula One cars crash more regularly these days, the likelihood of dying in an accident is extremely rare. Therefore, Formula One is inherently a safer sport now than it was in the 1960’s.
5. Lewis Hamilton and Michael Schumacher have recorded the most Grand prix wins, but have also both competed in more than 250 races each. The Conversion chart shows that 2 drivers (Fangio & Clark) from a bygone age have the best conversion rate in the sport. Given their limited season schedule compared to present day drivers, they did not have the opportunity to win 103 races like Lewis Hamilton. Aside from wins, there are other important career markers, like this conversion, to recognise as criteria for all-time greatness.

# References

<http://ergast.com/mrd/>

<https://en.wikipedia.org/wiki/List_of_Formula_One_fatalities>

<https://en.wikipedia.org/wiki/List_of_Formula_One_World_Drivers%27_Champions>

<https://en.wikipedia.org/wiki/Formula_One_engines>

<https://en.wikipedia.org/wiki/Prost%E2%80%93Senna_rivalry>

<https://www.makeuseof.com/f1-tech-in-road-car/>