Project Report – F1 Driver Analysis from 2020 – 2022

Contents

[GitHub URL 2](#_Toc130331527)

[Abstract 2](#_Toc130331528)

[Introduction 2](#_Toc130331529)

[Dataset 2](#_Toc130331530)

[Implementation Process 3](#_Toc130331531)

[Packages Imported: 3](#_Toc130331532)

[Custom Function 3](#_Toc130331533)

[Data Sets Imported from Kaggle 3](#_Toc130331534)

[Datasets Imported from F1 Website 4](#_Toc130331535)

[Results 6](#_Toc130331536)

[Points Analysis 6](#_Toc130331537)

[Wins Analysis 9](#_Toc130331538)

[Points VS Position Analysis 12](#_Toc130331539)

[Insights 13](#_Toc130331540)

[References 13](#_Toc130331541)

# GitHub URL

<https://github.com/brianregan111/UCDPA-BrianRegan>

# Abstract

Analysis of the driver performance for seasons 2020 – 2022.

Looking to highlight the dominance of the 2 main drivers over these 3 seasons and show the lap between them and the rest of the drivers.

# Introduction

The purpose of the project is to analysis F1 Driver performance over the past 3 seasons. In past 3 seasons has been heavily dominated by 2 drivers, Max Verstappen (VER), and Louis Hamilton (HAM).

The purpose of the analysis and will be firstly to highlight and prove the dominance of these 2 drivers through the data.

The purpose will also be to visualize the performance of the top 10 drivers across the last 3 seasons.

This information could then be used for doing predictive analysis for the oncoming season and could use machine learning to predict the 2023 standings.

# Dataset

Link: [Formula 1 World Championship (1950 - 2023)](https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020)

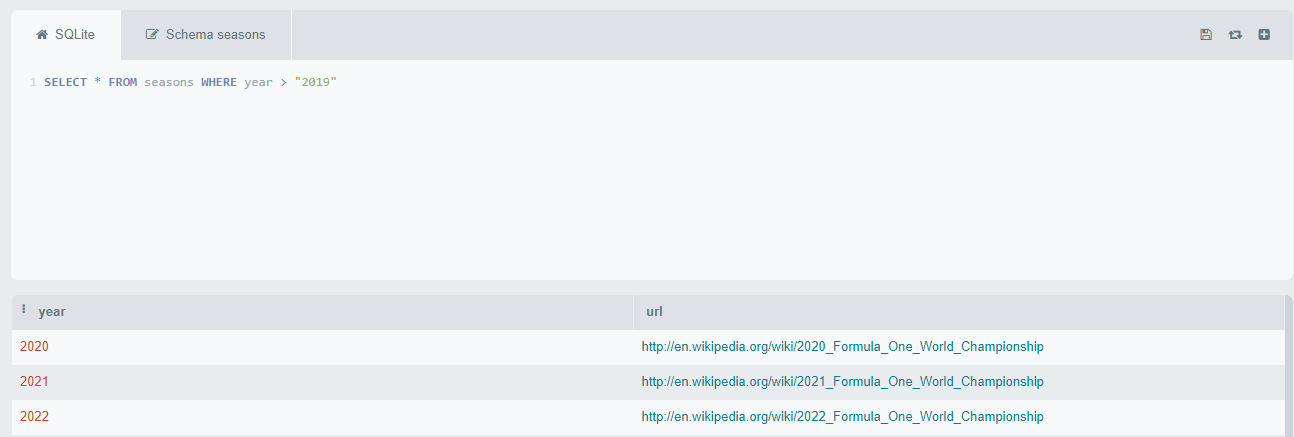
Expected update frequency: Weekly

Source: <http://ergast.com/mrd/> - API service

The dataset used was taken from Kaggle as it was a rich dataset with the opportunity to combine multiple tables. Of the 14 tables available in Kaggle the following were used:

1. Season.csv
2. Circuits.csv
3. Drivers.csv
4. Races.csv
5. Results.csv

Using SQLite the Season.CSV file was uploaded and using the following code the number of seasons was reduced to the relevant seasons required for the report:



This produced the Seasons.SQLLite.csv

Additional Data was Web scrapped from the Formula 1 official website for every relevant year:

1. [Standings2022](https://www.formula1.com/en/results.html/2022/drivers.html)
2. [Standings2021](https://www.formula1.com/en/results.html/2021/drivers.html)
3. [Standings2020](https://www.formula1.com/en/results.html/2020/drivers.html)

The reason this data was web scraped was due to the due to the data being rich, accurate and could easily be extended on with later seasons data if required.

# Implementation Process

## Packages Imported:

The following packages were imported for the project:

* pandas
* Seaborn
* requests
* matplotlib.pyplot
* numpy
* plotly.express

## Custom Function

A custom function is setup which is reusable for rounding numbers with more than 2 decimals down to 2.

## Data Sets Imported from Kaggle

Races.csv were imported, the years were filtered down to > 2019.

* The number of races per year are printed for later analysis.

Circuits.csv were imported.

Lists of countries and continents were created and turned into a dictionary.

* This is left merged with circuits to get the continent on the circuits table for graphing later.

A list of active drivers is created.

* Duplicates are removed.

Driver.csv is imported.

* Forename + Surname + code was added to one column to be used as a FK for joining.

Results.csv is imported.

* Results is inner joined with races(R2020) to bring the races data and limit the results to the relevant races.
* Results is left joined with active drivers, this could be used to limit results down to active drivers.
* Results is inner joined with active drivers, to bring in driver details.
* Results is inner joined with circuits to bring in the location and other circuit details to the result table.
* The Columns that are not needed are dropped.
* Points column was changed to an integer.

The expended results table (Res\_merged\_all) is then pivoted by driver and returns the mean points & starting position.

* The pivot created the bar/line Matplotlib chart:
* Line = Mean positionOrder (Starting Position)
* Bar = Mean Points
* The Res\_merged\_all is then used as a loop to count the number of drivers that won each year.

A winnersonly table is created from Res\_merged\_all returning the winning drivers only from the results table.

* winnersonly grouped by driver and sum of positions to generate a pie chart displaying the sum of wins by Driver (Matplotlib).
* Winnersonly grouped by Driver and year by sum of position order to display in a stacked bar chart to show winner per year by the driver (Matplotlib).
* Winnersonly grouped by driver and continent by sum of position order to show winners by continent (Matplotlib).

## Datasets Imported from F1 Website

The Standings for 2020,2021 and 2022 were imported from the F1 Website and concatenated into table StandingsCOMB

The average points were stored in a variable and a conditional field is added to show the drivers who were over and under the mean in the standings.

* The % of over and under is stored and printed for analysis.
* Created a seaborn pie chart to visualize the over under.

StandingCOMB is pivoted by driver, year and sum of points as the variable, it is then summed, sorted, and cleaned to required columns and assigned to clpvstandings. Using this table, I do the following:

* Added a total point column to pivot.
* Sorted by this column.
* Dropped the total row and column from the table.
* This is used to print the winning driver with their points using a loop.
* Difference between the winning driver and average per year using a loop.
  + Using Numpy to find the driver with the most points.
  + Return the average points per year
  + Created a dataframe for the winner and average points
  + Created a bar chart to show the difference between the average and the winner per year.
* This is used to generate a heatmap of the top 10 drivers by year (Seaborn).

StandingCOMB is grouped by driver and agg by points stored as drpoints

* % of total for each driver was added.
* Top 10 Points by drivers’ bar chart was generated from this drpoints (Seaborn).
* Percentage of Total Points by Top 10 Drivers Pie chart generated from drpoints (Seaborn).

# Results

## Points Analysis

**The drivers with the most points for the past 3 seasons are as follows:**

2020: Lewis Hamilton HAM, 347.0 points

2021: Max Verstappen VER, 395.0 points

2022: Max Verstappen VER, 454.0 points

**The average points per season are:**

Average points for 2020: 75.39

Average points for 2021: 105.05

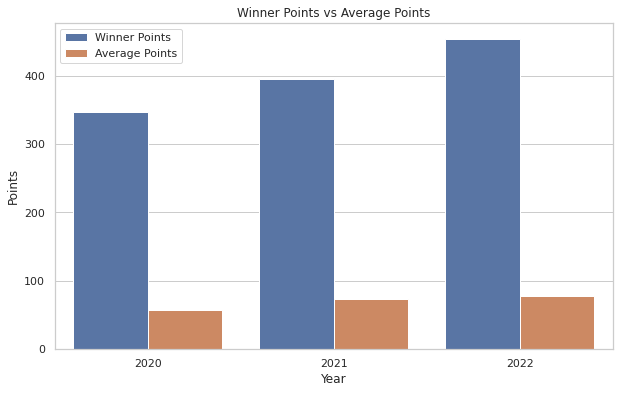
Average points for 2022: 106.82

**The differnce between the winners and the average:**

Difference in points between Lewis Hamilton and the average for 2020: 271.61

Difference in points between Max Verstappen and the average for 2021: 289.95

Difference in points between Max Verstappen and the average for 2022: 347.18



**Percentage of Driver Over/Under Average Points:**

Drivers Over the average points: 37.88%

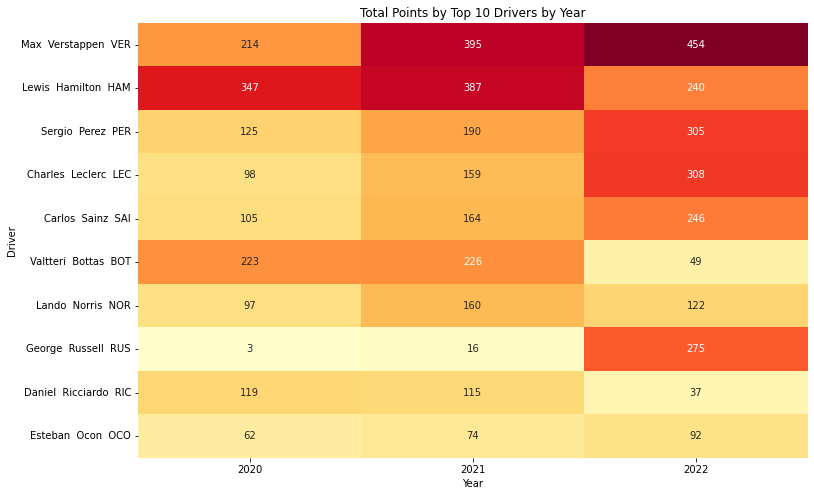
Drivers Under the average points 62.12%

Chart, pie chart

Description automatically generated

**Points By Top 10 drivers by Year(HeatMap):**

Looking at the top 10 drivers sorted by the total points over the past 3 years we can see the total points by Driver. Using the heatmap its clear to see who has won the most point in this time with Max Vertstappen and Louis Hamilton with the majortity of points in the darker red for at least 2 seasons consutively.



It should also be noted that the number of races person season increased after 2020:

2020: 17

2021: 22

2022: 22

This has resulted with an increase in total number of points availble to score in 2021 with a further increase avaible in 2022 due to addidaitonal points been allocated to sprint sessions:

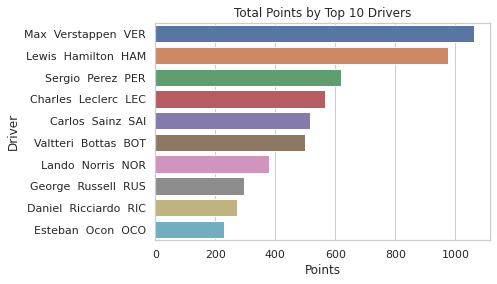
Total points availible in 2020: 1734

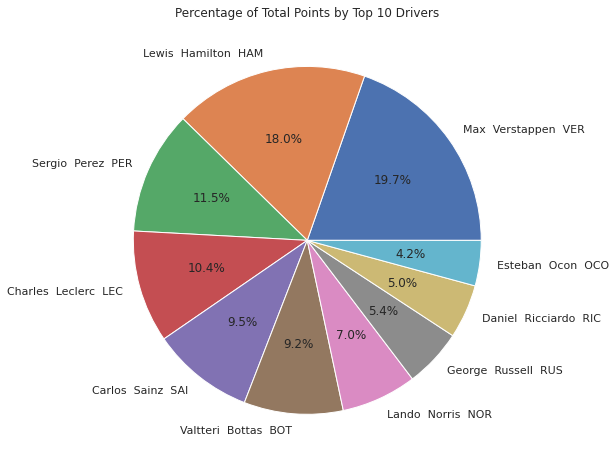
Total points availible in 2021: 2206

Total points availible in 2022: 2350

**Top 10 Drivers By Total Points & Percentage of Total Points:**

The top bar chart displays the top 10 drivers with their total points for the past 3 seasons. The pie chart displaying what percentage of the total points each driver earned.

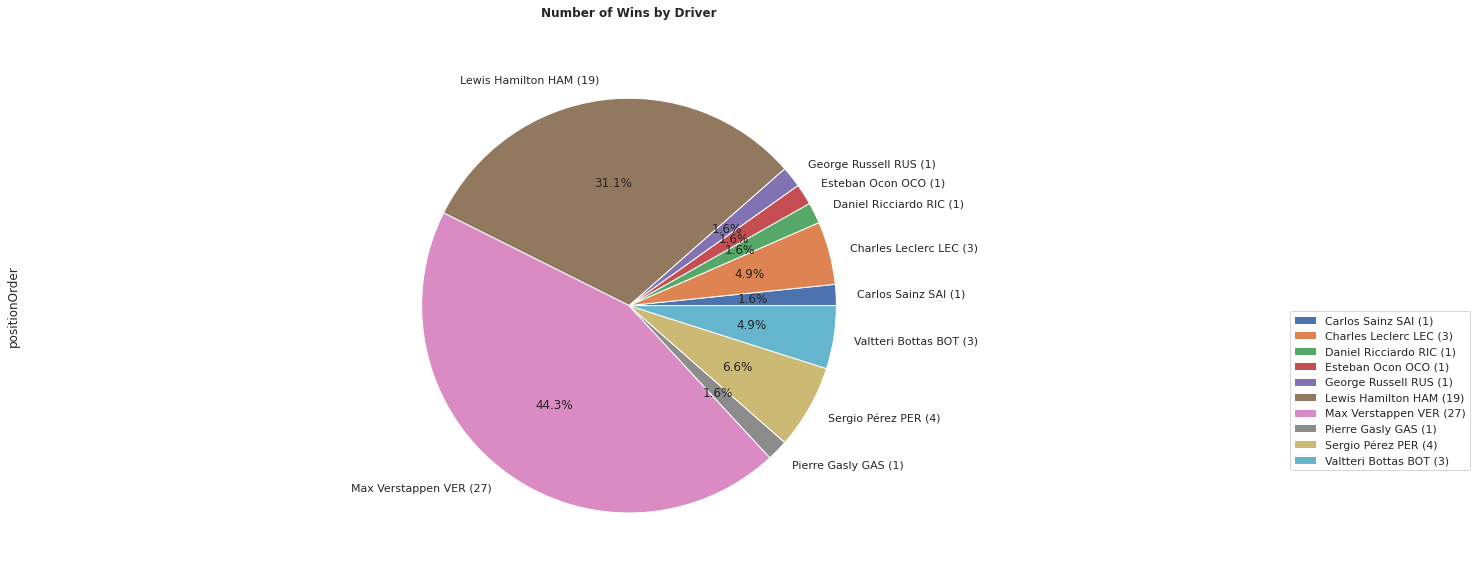




## Wins Analysis

**Number of Wins by Driver:**

In the below diagram again, the trend can be observed when comparing the number of wins by individual over the past 3 seasons. Both HAM & VER share a large portion of the wins.



**Number of Wins by Driver by Season:**

The number of drivers who won per season are as follow:

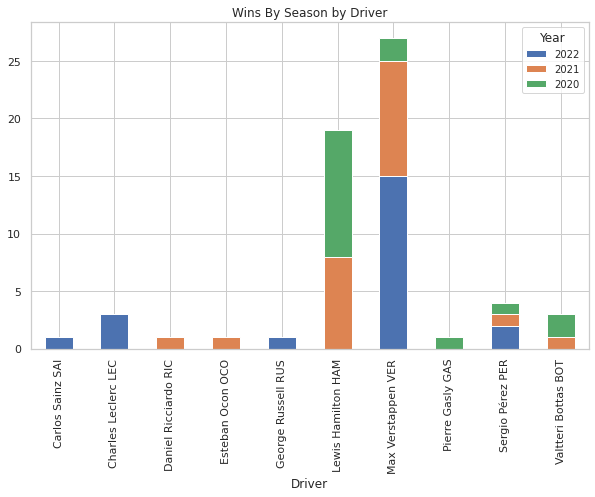
In 2020, 5 different drivers won.

In 2021, 6 different drivers won.

In 2022, 5 different drivers won.

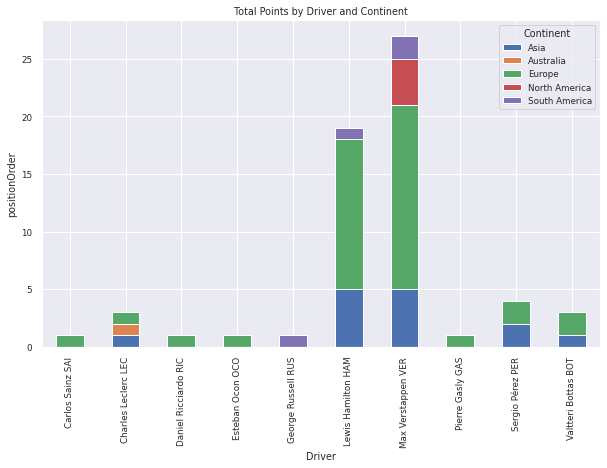
Which are shared between 10 drivers.

The stacked bar chart represents the winners and numbers highlighting the years that they won.



**Number of Wins by Driver by Continent:**

The following graph shows the split of wins by driver and continent.

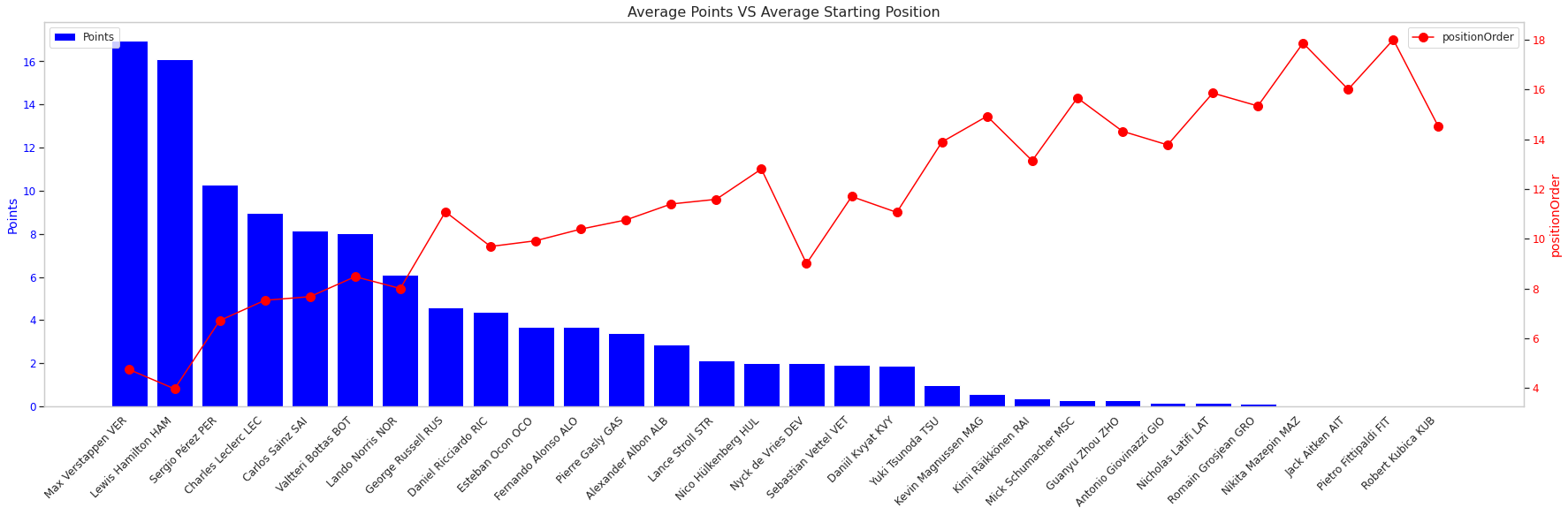


## Points VS Position Analysis

**Average Points Vs Average Starting Position:**

The bar chart below represents the average points scored by Drivers over the past 3 years.

The line represents the average starting position.



# Insights

From the data analyzed several key insights:

* Despite the growth in available points the gap between the top has grown, further demonstration the dominance of the top 2. From 2020 – 2022 there has been a growth of 616 points but as seen in the “Points By Top 10 drivers by Year(Heatmap)” VER has further increased his total in the past 2 seasons and pulled away from the average. This also highlighted by over the past 3 season only 37.88% of drivers having scored over the mean points total.
* A large gap between the winners is further highlighted by the Average points vs Winners.
* “Average Points Vs Average Starting Position” bar/line chart combined highlights a correlation between the average points scored and average starting position. Typically, the better the average position the better the average point scored. This is highlighted by VER and HAM high point and low starting position. The higher the average starting position typically represents a lower number of points scored.
* “Number of Wins by Driver by Season” the stacked bar chart below demonstrates this split of drivers who won, while supporting the domination of both VER and HAM over the past 3 seasons.
* “Top 10 Drivers By Total Points & Percentage of Total Points” Bar and pie chart show by total points scored and percentage of total points that both VER and HAM have displayed dominance, with winning 19.7% & 18% of the points when compared to the Top 10 in that order.
* The number of winners over the past 3 seasons has remained low in as shown in the Win Analysis section, “Number of Wins by Driver” pie chart, highlighting this and further highlighting the dominance of the top 2, collectively sharing 75% of the wins.

**Machine Learning**

Using the F1 Dataset a potential prediction use would be to work out the 2023 performance of drivers based on points scored over the past 3 season. To calculate this, regression methods could be used. An example of a of a method that could be used is linear regression, as the relationship between the total points and past performance variables can be used to make a prediction on what the points would be using the independent variables.

Classification could also be used to work out who would win each race of 2023 which should provide the overall winner, classification can take past results and driver performance inputs which should allow for a binary answer of win or not.

# References