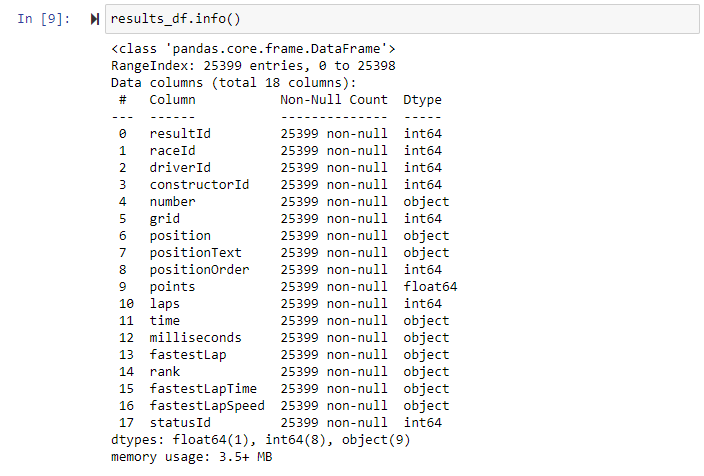
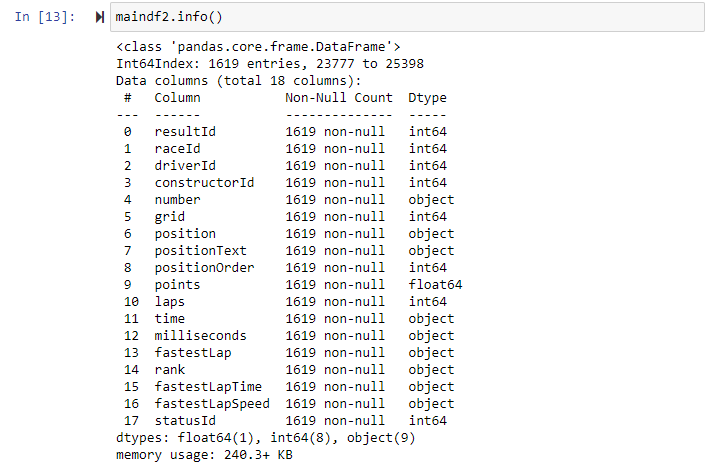
**F1 Correlation description**

To answer the research question, it was necessary to explore all the available tables to identify which variables could be used for the models.

The tables that were initially selected have information from 1950 to 2021. Carrying out a preliminary exploration of the tables, the table that has the greatest amount of information is results.csv, which has 18 columns and a total of 25,399 observations, being this the main table for this project.



However, due to the modifications in the formula 1 regulations, changes in the scoring and classification formats, the project period would cover from 2018 to 2021. Once the filter was made, the table was left with 1619 observations .



Regarding the information from this table, it can be identified that the variable that give an answer to the study question is the column called positionOrder (Final Position) that presents the final positions of each of the drivers for each of the races. Determining which are the most relevant factors to predict the result of the race requires knowing the final result.

To obtain the name of the constructors and drivers, we perform a merge of our main dataframe (maindf) with the constructors.csv and drivers.csv tables. In this way some patterns can be identified.

Image

From the graph it can be interpreted that the most successful teams in the study period are Mercedes, Red Bull and Ferrari, while those that obtain the most negative results are Williams, Haas and AlfaRomeo.

Image

Gráfico

Descripción generada automáticamenteGráfico, Gráfico de barras

Descripción generada automáticamente

Plotting the final results by driver, we can see that the drivers with the best results are Hamilton, Max Verstappen and Bottas. While the drivers with the worst ratings are Pietro Fittipaldi and Mazepin.

But what does the result of an F1 race depend on, the constructor or the driver?

From the original database, the starting position variable (Grid) is identified, which could be related to the driver's ability to start the race in a good position. Carrying out the first correlations, it is observed that the starting position is highly related to the final position.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

It can be seen that it is more common for a driver who starts in the first places to end up in the first places, it is even observed the difficulty that a driver who starts in the last positions get to the first positions, however, it is more common for drivers who start in the first places to finish in last places, so it is understood that in addition to the starting position there are other variables that could explain who is the winner of a race.

Referring to the weight of the constructor in determining the winner of a race, we can identify the rankfast variable (Rank) that ranks the cars according to the times of the fastest laps of the entire race. This variable refers to the maximum speed that a vehicle can reach and therefore it can be determined that it depends more on the constructor.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

The results of the correlation graph indicate a high relationship between the maximum speed of a car and the final position, where the slowest cars in general always finish in the last places, however, there are cases where fast cars end in last places, showing that there are other factors that affect the final result.

With this in mind, more variables are analyzed to explain the principal factor that determined the final result of a race. The variables were classified according to their relationship with the driver or constructor.

**Constructor**

* **Pit-stops strategy**

In the pit-stops variable, the relationship between the winner of the race is not determinant. This can be easily seen, since a constructor that finishes first and a constructor that finishes last can have the same number of pit stops. The variation between the number of stops per team is not representative in the graph. Likewise, in this graph you can see cases where there are more pit-stops than what is estimated on average among constructors.

Image

* **Constructor points per season**

During each race, each constructor gets a sum of point depending on the rank race, at the end of the season the constructor end with a total sum of points.

Regarding the correlation between the season points and the final position, it is shown that the constructor that has more points accumulated during the season, will be the constructor that end in the first place.

Image

* **Constructor points per circuit**

As mention before, the constructor gets points in every race, in this correlational chart is shown almost the same behavior that it is shown in the last variable (points per season). If a constructor gets more points in a circuit it will be the winner of that circuit, also it is important to mention, that this is a constructor analysis, meaning that the total of points is the sum of points that every pilot gets for the team.

Image

* **Constructor failures average per circuit**

In every race it is common that at least one constructor present a failure during a race, as shown in the next image, in average the majority of the constructor has between 0 and 1 failure during a circuit. In rare cases there are constructors that has more than 5 failures during a circuit, but this is not a normal behavior.

In the next graphics it is shown the number of failures each constructor had has, and also per driver.

**Gráfico, Gráfico de dispersión

Descripción generada automáticamente**

**Gráfico, Gráfico de barras

Descripción generada automáticamente**

**Constructor Failures Average per circuit to each driver (Built it with the base tables)**

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

**Constructor Failures Average per each driver (Number of failures attributed to the constructor)**

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

**Driver**

* **Experience**

To calculate the experience variable, the total number of years that each driver has driven was counted and the Rank() function was applied to order in descending order, where the value of 1 represents the runner with the most experience.

**Gráfico

Descripción generada automáticamente**

This correlation presents a relationship between the runners with the greatest number of years of competition and the final position of the race.

* **Total number of races**

To calculate this variable, we add the total number of races that each of the drivers has run, then the Rank() function is applied to order them in descending order, where the value of 1 represents the runner with the most experience.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

It is observed that the greater experience, the more possibilities a pilot has to remain in the first positions, but not so the pilots with less experience, who usually remain in the last places.

* **Most winning drivers by position**

This variable was calculated by filtering the runners who have won a race, then the sum per runner is obtained and the column is added to the maindf, then the Rank() function was applied to order in descending order, where the value of 1 represents the runner with the highest number of wins.

**Gráfico, Gráfico de dispersión

Descripción generada automáticamente**

In this case, although the relationship is not strong, it is observed that the runners who have won once have a good chance of winning again, however, the result is not totally determinant.

* **Track experience**

**Number of times driver has run each circuit (Built it with the base tables)**

To calculate this variable, we add the total number of races that each of the drivers has run on each of the tracks, then the Rank() function was applied to order in descending order, where the value of 1 represents the runner with the most experience. on each of the tracks.

**Gráfico, Gráfico de dispersión

Descripción generada automáticamente**

For this correlation, such a linear relationship is not observed, which can be explained because there are some tracks that have been run a few times and even so there is a winner or there are drivers who, despite having a lot of experience, can end up in the last places.

* **Times**

**Drivers Qualifying pole wins (Built it with the base tables)**

To calculate this variable, the sum of the times a driver has a qualifying position for all the races is obtained.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

It is observed that the pilots who have had the pole position at some time are more likely to be in the first places than those who have never won it.

* **Average of crashes per pilot**

In a F1 race, crashes are one thing that happen and affect the result of the race. In average a pilot has less than 0.14 crash, but as it is shown in the graphic as higher is the number of crashes, the result will be losing the race.

Image

* **Driver total points**

Points are awarded to drivers in every race they finish, the winner receives 25 points, the second finisher 18 points, and so on. The next image shows how the points are distributed by driver, and it is shown that the distribution of point per driver is divided between all drivers.

Image

**External factors**

**Weather (Built it with an external base)**

Taking in count the weather variable we can see that is not a determinant variable, it can affect the race, but as seen on the graph, the result remain the same.Gráfico, Forma, Rectángulo

Descripción generada automáticamente