INDICE

Sommario

[INTRODUZIONE 3](#_Toc167386824)

[1. I Dati 7](#_Toc167386825)

[1.1 Le Librerie 7](#_Toc167386826)

[1.2 Definizione delle Funzioni 8](#_Toc167386827)

[1.3 Definizione del Dataframe 9](#_Toc167386828)

[2. Analisi Descrittive 11](#_Toc167386829)

[2.1 Statistiche Semplici 11](#_Toc167386830)

[2.2 Distribuzione dei Tempi per Team 13](#_Toc167386831)

[2.3 Correlazioni 15](#_Toc167386832)

[3. Regressione 16](#_Toc167386833)

[3.1 Scatterplot 16](#_Toc167386834)

[3.2 Definizione Distanza D-Cook 17](#_Toc167386835)

[3.3 Indice di Bontà R2 17](#_Toc167386836)

[3.4 Retta di Regressione 19](#_Toc167386837)

[4. Telemetrie a confronto 21](#_Toc167386838)

[4.1 Estrazione delle Telemetrie 21](#_Toc167386839)

[4.2 Confronto delle Distanze Percorse 22](#_Toc167386840)

[4.3 Confronto delle Velocità 24](#_Toc167386841)

[4.3.1 I due piloti più veloci 24](#_Toc167386842)

[4.3.2 Gli altri piloti 28](#_Toc167386843)

[5. Le Performance sulle diverse mescole 30](#_Toc167386844)

[5.1 Prestazione dei Team divisa per mescola 30](#_Toc167386845)

[5.1.1 Hard 31](#_Toc167386846)

[5.1.2 Medium 33](#_Toc167386847)

[5.1.3 Soft 34](#_Toc167386848)

[5.2 Correlazioni con le temperature 35](#_Toc167386849)

[6. Conclusioni 37](#_Toc167386850)

[6.1 Riepilogo 37](#_Toc167386851)

[6.2 Le Vere Grandezze 38](#_Toc167386852)

[6.3 Considerazioni Finali 38](#_Toc167386853)

[RINGRAZIAMENTI 41](#_Toc167386854)

# INTRODUCTION

In a world full of data, there is an increasing search for solutions to leverage it to one's advantage, in any field that surrounds us. Specifically, sports are adapting, and year after year, more information is being collected to gain benefits.

One of the main areas where data collection and analysis are fundamental is motorsport. In particular, the most developed is Formula 1: the place where technology, strategy, and adrenaline merge together.

In 1947, the CSI, the International Sports Commission, which is now known as the FIA, established the first Formula 1 World Championship for drivers. The subsequent years were used to establish regulations, registrations, and the selection of circuits. Three years later, in May 1950, the first official Formula 1 Grand Prix was held, then still intended only for individual drivers. The event took place in Great Britain and was won by Nino Farina driving the Alfetta 158, with which he won the first edition of the World Championship.

Today, 74 years have passed since that first Grand Prix, years that have seen various winners, marking the evolution of increasingly high-performing, alternative, and sometimes even extravagant cars. Iconic are the scenes in today's films that recount the internal dynamics between drivers and mechanics in motorsport, and how analysts on the sidelines were armed simply with a stopwatch and a notebook. Over the years, the world of motorsport has evolved at an uncontrollable pace, testing the most brilliant engineers. With the evolution of cars, the importance of data and how it could lead to improvements was understood.

Data analysis is a fundamental pillar in the Formula 1 landscape, where even the smallest variation can determine victory or defeat. Each Grand Prix sees the production and collection of several gigabytes of data from both cars of the same team, a significant amount of information.

Indeed, data is used for various aspects, such as developing new technologies aimed at improving the performance of the cars. This involves an in-depth exploration of aerodynamics, engines, and suspensions through the use of available data. Analysis is crucial in refining existing cars, allowing teams to identify possible issues or areas for improvement. Another example is in race preparation, where teams use collected data to study the track and simulate race conditions. With the collected data, engineers spend their time analyzing the car's telemetry, thus helping to ensure a constant competitive advantage.

Each lap of the track produces 35 megabytes of data, translating to 30 gigabytes of information to analyze during a race weekend. This data, collected through a series of sensors placed on the car, is processed using specific software like ATLAS. All this data is transmitted in real-time to the team. The transmission takes place through the antenna placed on the nose of the car or via receivers along the circuit, allowing engineers to remotely monitor the car's performance and intervene promptly during the race.

The main objective of the thesis is to show how, through Python, it is possible to analyze past data in the world of motorsport and extract useful information from it. An example of research is shown, taking into account a specific Grand Prix. The analysis is structured as follows: initially, data is collected and transformed, an Excel file is created containing various variables divided by year and race, simple statistics are calculated, a correlation analysis is performed, and a linear regression model is estimated. In the next step, the analysis goes into more detail: first, the telemetry of the two fastest laps is compared and mapped within the circuit, then the comparison is repeated among other drivers, all through the use of various libraries. In the penultimate chapter, a very important aspect is analyzed, namely the performance of the 10 teams over the years divided by different tire compounds: first, the data is normalized for comparison, and then the correlation of performance with external variables such as temperatures is studied. In the last chapter, there are conclusions with a summary and final considerations.

# 1. The Data

Before conducting statistical analyses, a fundamental step must be carried out, which is the collection and transformation of data using specific tools. For the data selection, all laps of all drivers in the championships from 2019 to 2023 are extracted. This timeframe was chosen because new tire regulations were introduced in 2019: the number of different compounds was reduced from nine between “slick” and “wet” tires to five, with three for dry conditions and two for wet conditions. For dry conditions, the compounds range from Soft, which is the softest and most performance-oriented but has the shortest lifespan, to Medium, which offers a good compromise in terms of performance and durability. Finally, there are the Hard compounds, which are the hardest, allowing drivers to complete more laps before needing to change tires, at the cost of higher lap times.

## The libraries

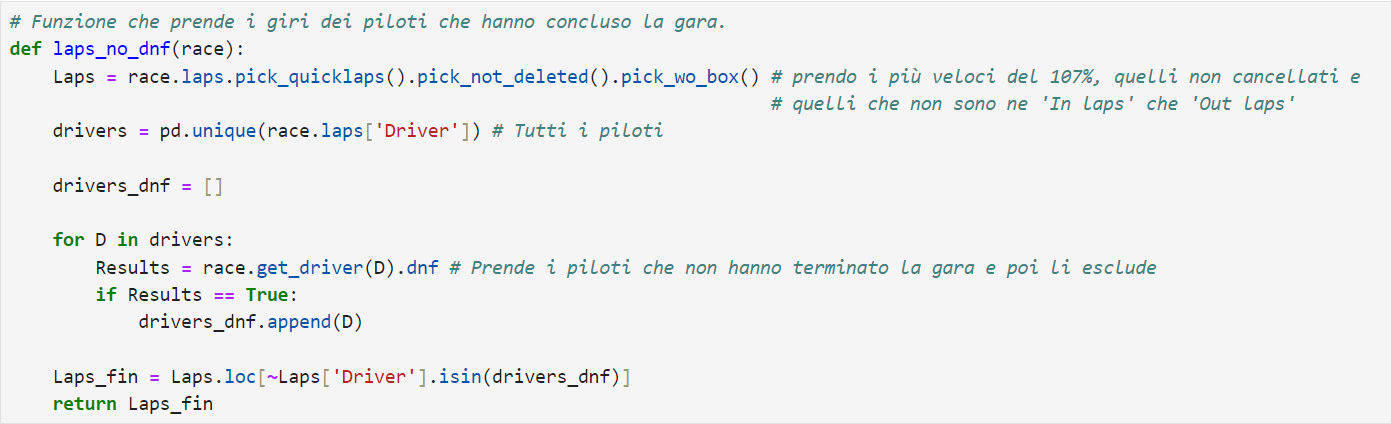
Information is extracted using a library called "FastF1". A library is a collection of predefined code that provides a set of functions or modules to perform specific operations. In this case, FastF1 is a library created by a team of developers that provides access to Formula 1 data, including real-time data, for races and qualifying sessions. Specifically, it allows access to all events, locations, and dates, session results, recorded times, track conditions, messages from the stewards (such as penalties and investigations), telemetry, and track characteristics. The main methods that will be used are:

* get\_session and load to extract and load a specific session of a particular race in a given year;
* laps to retrieve all laps of all drivers in that session;
* Finally, methods to filter the laps that meet certain characteristics.

Other libraries used include Pandas and Numpy for data analysis and dataframe creation, Matplotlib and Seaborn for creating charts, Scikit-learn for estimating regression models, and Statsmodels for assessing the goodness-of-fit of the model.

## Function Definitions

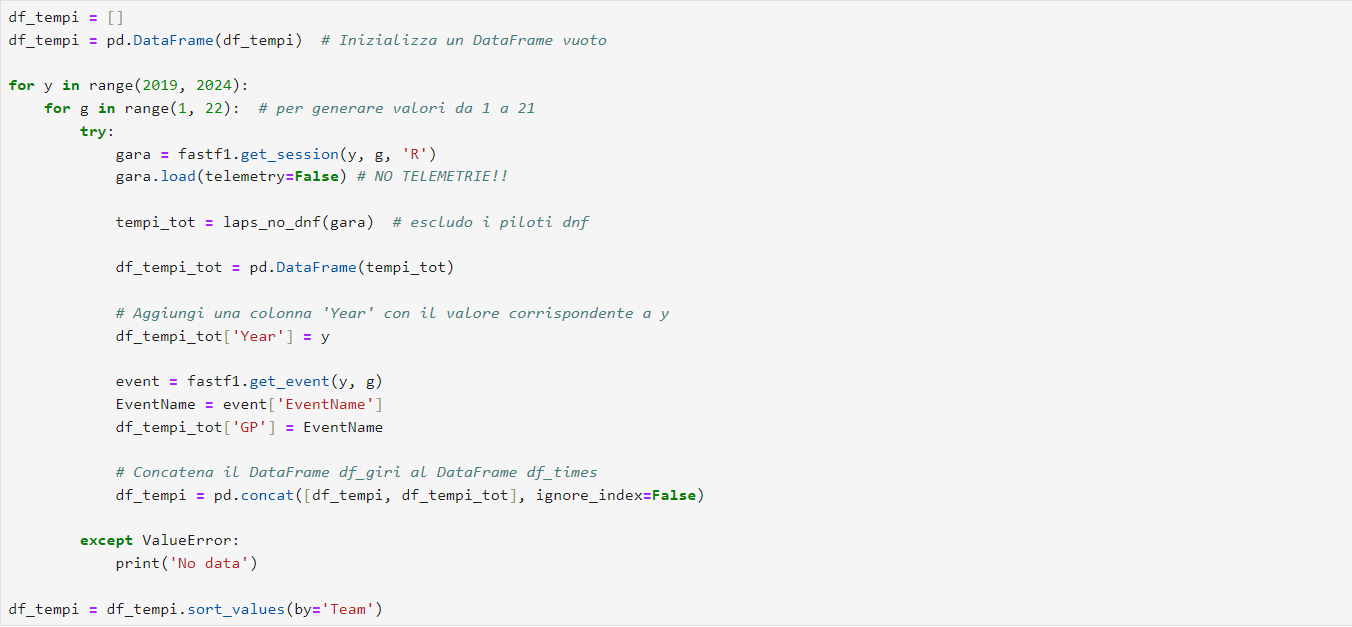
The step preceding the definition of the dataframe is to define a specific function that will be called during data extraction. The function, named "laps\_no\_dnf", takes the argument “race,” which is the loaded session. Subsequently, through the methods of FastF1, the laps are filtered as follows: the fastest laps within a certain limit, 107% (for example, if the fastest lap lasted 100 seconds, laps longer than 107 seconds are not considered), laps not invalidated due to an infraction, and laps in which the driver is entering or exiting the pits, since speed limits must be respected in the pitlane. This way, laps under the safety car, virtual safety car, or laps where a driver had car issues are also eliminated.



*Codice per la definizione della funione laps\_no\_dnf*

## DataFrame Definition

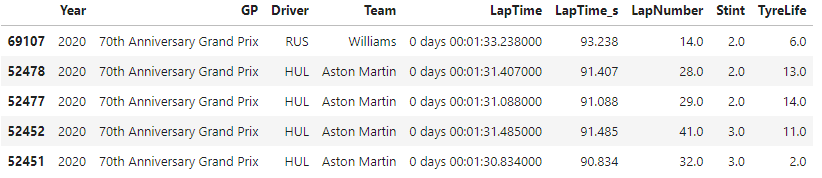
Through a "for loop", the sessions are loaded; for each iteration, the laps are filtered with the previously defined function, and a dataframe is created using Pandas, which is then concatenated with the others. Two columns are added to the dataframe: one containing the Grand Prix and one with the year in which it was held, so that the data can be grouped.



*Codice per la definizione del dataframe*

With the to\_csv method, the data is extracted into a CSV file, containing approximately 72,700 rows and 33 variables, the main ones being:

* Year: the year;
* GP: the name of the Grand Prix;
* Driver: abbreviation of the driver's name;
* Team: the driver's team;
* Stint: the period between the start and the first pit stop or between one pit stop and another;
* LapNumber: the lap number in which the time is recorded;
* TyreLife: the number of laps the tires have been used;
* LapTime: the time in minutes, seconds, and milliseconds taken to complete the lap in date format;
* LapTime\_s: created with methods, first converting LapTime into a timedelta object and then extracting the total seconds;
* Compound: the type of tire among SOFT, MEDIUM, and HARD;
* Sector1Time, Sector2Time, Sector3Time: the times recorded for the 3 sectors in a given lap;
* SpeedST: the maximum speed recorded in km/h of the car on the longest straight.

****

*Esempio di alcune variabili del dataframe*

# 2. Descriptive Analysis

To analyze the data, a specific Grand Prix is selected: the 2023 "Qatar Grand Prix". One of the features of this Grand Prix is that, being held at night, lower temperatures can lead to different and interesting dynamics during the race. However, the main reason it was chosen is the track's diversity, featuring a combination of fast corners, long straights, and technical sections.

## 2.1 Simple Statistics

For an initial descriptive analysis, some variables are selected, such as lap time, the time spent in the first, second, and third sectors, and finally, the top speed recorded on the longest straight.

Immagine che contiene testo, schermata, Carattere, numero

Descrizione generata automaticamente

*Statistiche semplici del gran premio*

In the dataframe, the times were of type “Time”, so they were first converted to “Time Delta” and then to seconds.

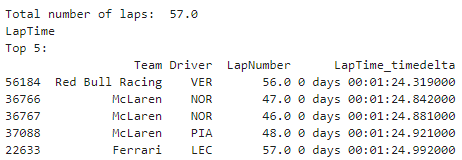
It shows that in 2023 the average lap time was 87.392 seconds, meaning the drivers averaged a lap time of 1:27.392, while the median was 2 tenths lower at 1:27.206.

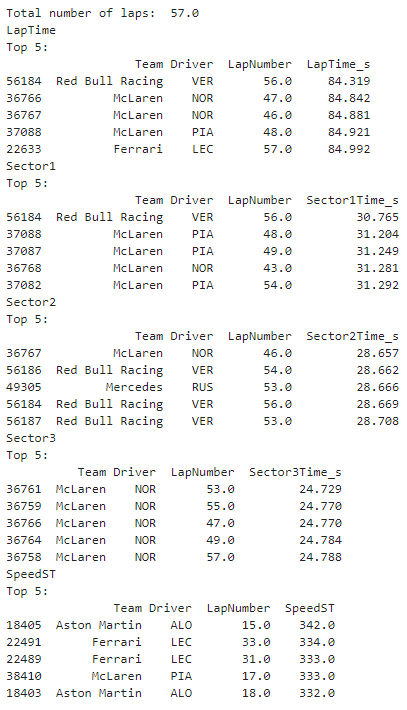
There is only a 3-second difference between the fastest and the slowest lap.

From the start/finish line to the first sector timing beam, the cars took an average of 32.224 seconds, 29.455 for the second sector, and 25.713 for the third.

As for the top speed recorded on the longest straight, the average is 308 km/h.

In this case, there is a significant difference in the top speeds recorded, as 75% of the observations are below 311 km/h, while the highest speed was 342 km/h.



Immagine che contiene testo, schermata, numero, menu

Descrizione generata automaticamente

*Le statistiche univariate del gran premio*

From the univariate analysis, the most performing drivers for the different variables are extracted. In the case of lap time, the fastest lap was achieved by Max Verstappen in his Red Bull with a time of 1:24.319, recorded on lap 56 of the Grand Prix. The fastest lap was half a second quicker than the second fastest, set by Lando Norris on lap 47 with 1:24.842. Although the Red Bull driver had the lowest lap time, he was only the fastest in the first sector, while Norris was the quickest in the other two sectors with 28.657 and 24.729. Regarding top speed, Fernando Alonso's Aston Martin, thanks to the use of DRS, passed the speed trap at 342 km/h on lap 15, 8 km/h faster than the second-fastest, a Ferrari driver at 334 km/h. In Formula 1, an 8 km/h difference is significant. The third and fourth highest speeds were 333 km/h, and the fifth was 332 km/h, highlighting a significant gap for the Spanish driver.

## Distribution of Times by Team

To visualize the distribution of times, two types of overlaid charts are used:

* The first is a “swarmplot” that allows the analysis of the distribution of a numerical variable against a categorical one, avoiding overlaps and showing data density.
* The second is the “violin plot” which, also concerning distribution, combines a box plot and a kernel density plot. It visualizes the presence of any peaks or outliers.

This allows for comparing different groups.

Immagine che contiene testo, diagramma, Diagramma, schermata

Descrizione generata automaticamente

*Distribuzione dei tempi divisi per scuderie ordinati per la mediana*

In this case, the distribution of lap times divided by team is visualized. Each observation is distinguished by a color indicating the compound used for the lap: red for soft, yellow for medium, and white for hard.

A purple line representing the median lap time of all teams was added to the chart. The median was chosen to compare it with each team's median, shown as a dashed line within the violin plot.

It shows that the fastest teams were McLaren and Mercedes, below the overall median.

Ferrari recorded times in line with the central value, while Red Bull and Aston Martin were above it.

In the case of Red Bull, despite Verstappen having an excellent race and winning, the team's times were penalized by his teammate, Sergio Pérez, who lapped at a slower pace and finished in tenth position.

From the chart, it is particularly evident which team distributions are more or less concentrated, as in the case of Aston Martin, where times have lower dispersion.

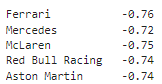
## 2.3 Correlations

As noted from the simple statistics, the fastest laps are recorded towards the end of the race. This is due to the fact that during the course of the Grand Prix, the cars consume fuel and reduce their weight.

Therefore, before proceeding with a correlation analysis for the Qatar Grand Prix, an analysis between the lap number and the corresponding lap time is performed. This analysis is divided by year, Grand Prix, and the five teams, extracting a dataframe in CSV format.



*Codice definizione del dataframe delle correlazioni*



*A sinistra le correlazioni dei diversi team nei diversi GP. A destra la mediana delle correlazioni per team.*

For each created group, a function is applied that selects only the “LapNumber” and “LapTime\_s” columns and calculates the correlation between them using the “corr()” method.

After applying the function to each group, the result is transformed into a matrix using the “unstack()” method. This reorganizes the data so that the unique values of the 'Team' column become the rows of the matrix index, the unique values of the 'Year' column become the columns, and the correlation values are placed in the corresponding cells.

As can be seen, there are strong negative correlations in most of the races. Therefore, as the number of laps increases, the average lap time recorded by the cars decreases.

# 3. Regression

In this chapter, we will delve into the relationship between the number of laps and the lap times of the cars, showing how it is possible to estimate regression lines with a high goodness of fit.

## 3.1 Scatterplot

Initially, a scatterplot (a graph of lap times versus laps) of two teams, Red Bull and McLaren, is shown. These two teams scored the most points in the Grand Prix.

Immagine che contiene testo, Diagramma, schermata, linea

Descrizione generata automaticamenteImmagine che contiene testo, schermata, Diagramma, linea

Descrizione generata automaticamente

*Scatterplot della Red Bull a sinistra e della McLare a destra.*

In particular, the graphs show how lap times decrease as the number of laps increases.

In the case of Red Bull, the values are more dispersed; whereas the two McLaren drivers had more consistent lap times.

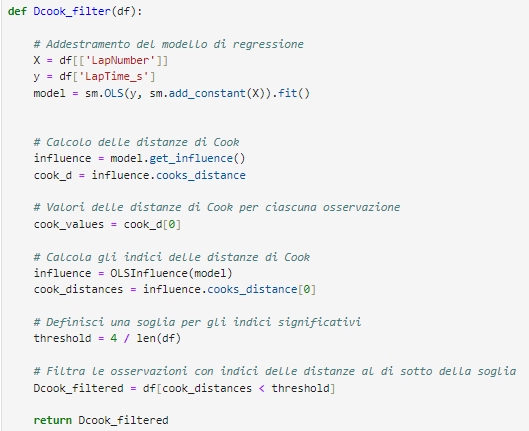
As seen previously, the fastest laps are recorded in the final phase of the race, where the cars are much lighter.

Therefore, observing the decreasing trend in times and their pattern, we can consider estimating a regression line passing through the points.

## 3.2 Definition of Distance D-Cook

Before estimating the line, it is necessary to exclude influential values from the model using the Cook's distance index.

This index is a measure used to evaluate the influence of individual points on the regression model parameters and is used to eliminate influential observations, which are values that might have a significant impact on the model construction.

****

*Codice per la definizione dell’indice D-cook*

In Python, the function Dcook\_filter is defined: this takes the dataframe as an argument, then, using methods from the scikit-learn and statsmodels libraries, the model is trained, the distance index is calculated, and a threshold for significant values is defined. All distances smaller than this threshold, defined by a calculation of 4 divided by the number of observations in the dataframe, are then taken.

Finally, the function returns the new filtered data.

## 3.3 Goodness of fit index R2

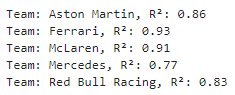
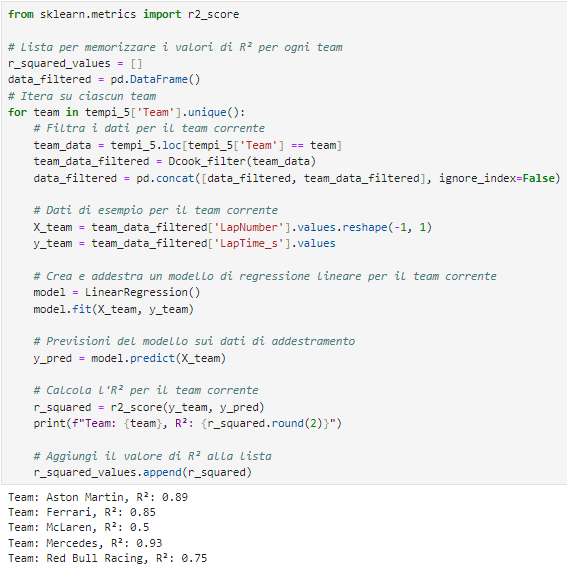
Once the function for calculating distances is defined, it will be used to obtain the R2 goodness-of-fit index for the model.

R2 is a measure that indicates how well the regression model fits the observations and provides a measure of the variance explained by the model, which is how much information is retained and not lost compared to the total data.

A value closer to 1 indicates a model that explains the data variation well.

At this point, we proceed with writing the code to calculate the index: a for loop is created that, at each iteration, takes the lap times divided by team, then applies the Dcook\_filter function so that the data is filtered based on the distance index.

On the new values filtered through the LinearRegression function, the regression model is created and trained, the variable containing the predicted values of the model is defined, and finally, through the r2\_score method, the goodness-of-fit is calculated for each team.



*A sinistra il codice che calcola gli indici R2 che si trovano a destra*

As can be seen from the results on the right, the 5 estimated models show a high index value, with Ferrari and McLaren being very close to 1, indicating that the estimated lines pass very close to the observations.

In the case of the Maranello team, however, only one driver was considered because the teammate could not participate in the race, so the observations were much more homogeneous.

For McLaren, it was an outstanding performance; the team provided two very similar cars to their drivers, showing great race pace for both. This resulted in a second and third place finish.

## 3.4 Regression Line

At this point, it is possible to analyze, through the regression lines, the lap time trends of the 5 teams during the GP.

Using the Seaborn library, we can plot the previously estimated model, color-coded for the respective teams.

Immagine che contiene schermata, testo, linea, Diagramma

Descrizione generata automaticamente

*Plot delle rette di regressione dei 5 team*

By studying the lines, the performance of the teams can be summarized and compared.

In the first phase of the race, Red Bull had the lowest pace but did not improve much as the laps increased.

Ferrari was the worst in the first part, but in the second part, the "Rossa" evolved more than the others, reaching nearly the same times as Mercedes in the final stages, which was competitive throughout the GP.

Aston Martin did not have great pace throughout the race, finishing with the highest times compared to the others.

Finally, McLaren was above both Mercedes and Red Bull for the first 15 laps, but it was the team that progressed the most, ending with the best times by a significant margin.

In the final part, Red Bull, Ferrari, and Mercedes were running at similar times.

Therefore, through the regression lines, a lot of information can be gathered, such as race pace, by observing the position of the line, or how much the teams improve during the GP, particularly by observing the slope of the line, and comparisons can be made by looking at where they are located relative to each other.

# 4. Telemetry Comparison

So far, a general analysis of lap times for the Grand Prix has been conducted, but it is possible to complete it by going into more detail.

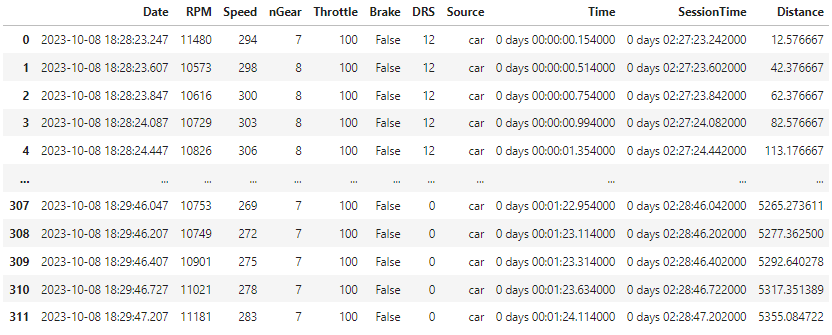
As previously explained, thanks to the FastF1 library, very valuable information from the motorsport world can be analyzed.

With this tool, you can extract, with just a few commands, the air and track temperatures and their evolution, as well as the telemetry data of the teams with very high precision.

## 4.1 Extraction of Telemetries

Once the session is loaded, as was done when the lap times dataframe was defined, the first step is to choose the laps from which to extract telemetry data. In this case, the two fastest laps of the entire race are taken: the first by Verstappen and the second by Norris, already analyzed previously. At this point, using the “get\_car\_data()” method, a new data table containing the telemetry data is created. Observations are extracted approximately every 200 milliseconds, allowing for precise information. The data is grouped into the following variables:

* **Date**, which contains the exact date and time the measurement was taken;
* **RPM**, "Revolutions Per Minute," shows the engine RPM of the car at that moment;
* **Speed**, the speed in km/h of the car;
* **nGear**, the number of the gear engaged;
* **Throttle**, the percentage of throttle applied;
* **Brake**, whether the brake is being used (True) or not (False);
* **Time**, the time elapsed since the start of the lap;
* **Distance**, the distance covered in meters at that moment.



*Immagine della tabella delle telemetrie di Verstappen*

A high number of details can be observed; indeed, 311 data points representing all the moments of a single lap have been extracted.

In addition to Verstappen's telemetry, Norris's telemetry is also collected for comparison.

## 4.2 Comparison of Distances Covered

The first comparison made is a simple comparison of the distances covered during the lap.

The X-axis represents the distance in meters covered by the drivers, while the Y-axis represents the time recorded at that moment.

Immagine che contiene testo, Diagramma, linea, diagramma

Descrizione generata automaticamente

Immagine che contiene testo, Diagramma, linea, diagramma

Descrizione generata automaticamente

*Plot dei giri a confronto tra Verstappen e Norris*

In the graph, Verstappen is shown in blue and Norris in orange, with dashed vertical lines marking the points where the turns are, to better understand the position on the track at that moment.

Analyzing the graph, one can observe the time gap that develops over the lap.

In the first turns of the lap, the drivers were close, but in the following turns, the power of the Red Bull increased the gap between the two. Subsequently, the difference remained constant until, around the last two turns, the British driver managed to reduce the gap, aligning with the Dutch driver's times.

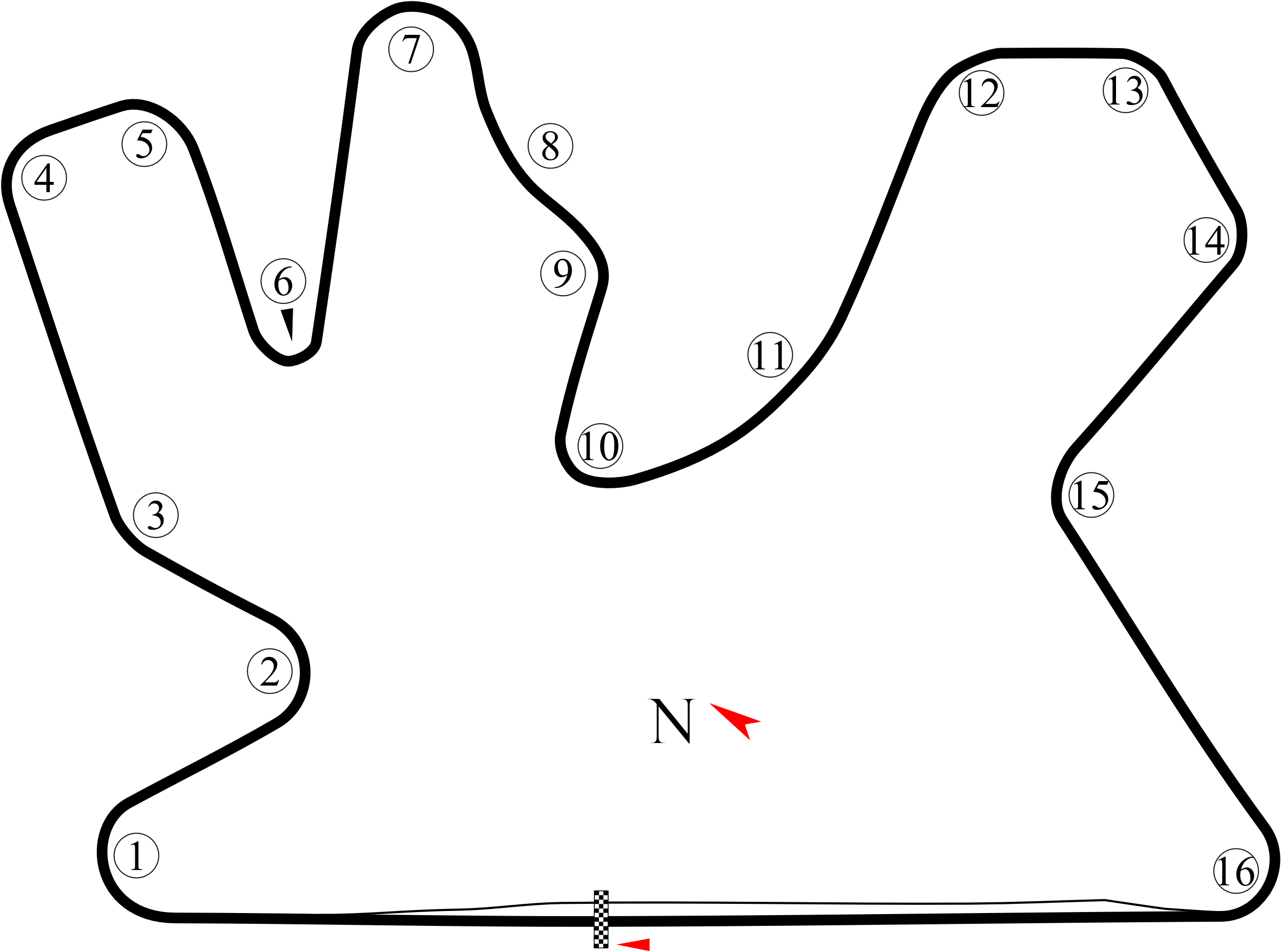
However, in the last kilometer, Verstappen exited the final turn better and, taking advantage of his engine, was able to widen the gap again.

The two lines being so close highlight how much difference there is between one car and another in the world of motorsport.

## 4.3 Comparison of Speeds

### 4.3.1 The two fastest drivers

Going into more detail, we study the speeds brought to the track by the two drivers.



*Circuito di Lusail con le curve numerate*

This figure shows the Lusail circuit where the Qatar Grand Prix was held. The corresponding number of each turn is marked, and you can see the fastest, the slowest turns, and the sections of the track where there are straights. The main sections are as follows: turn 1, which has a long, rounded shape to the right, but hard braking is needed as it comes from the long straight; turn 6, which is the slowest turn and is a very complicated uphill hairpin; turns 8-9, which are very fast; then the three right-hand turns 12-13-14, where a lot of speed is also carried; and finally turn 16, where drivers feel strong lateral G-forces while trying to exit as best as possible to reach the longest straight.[[1]](#footnote-2)

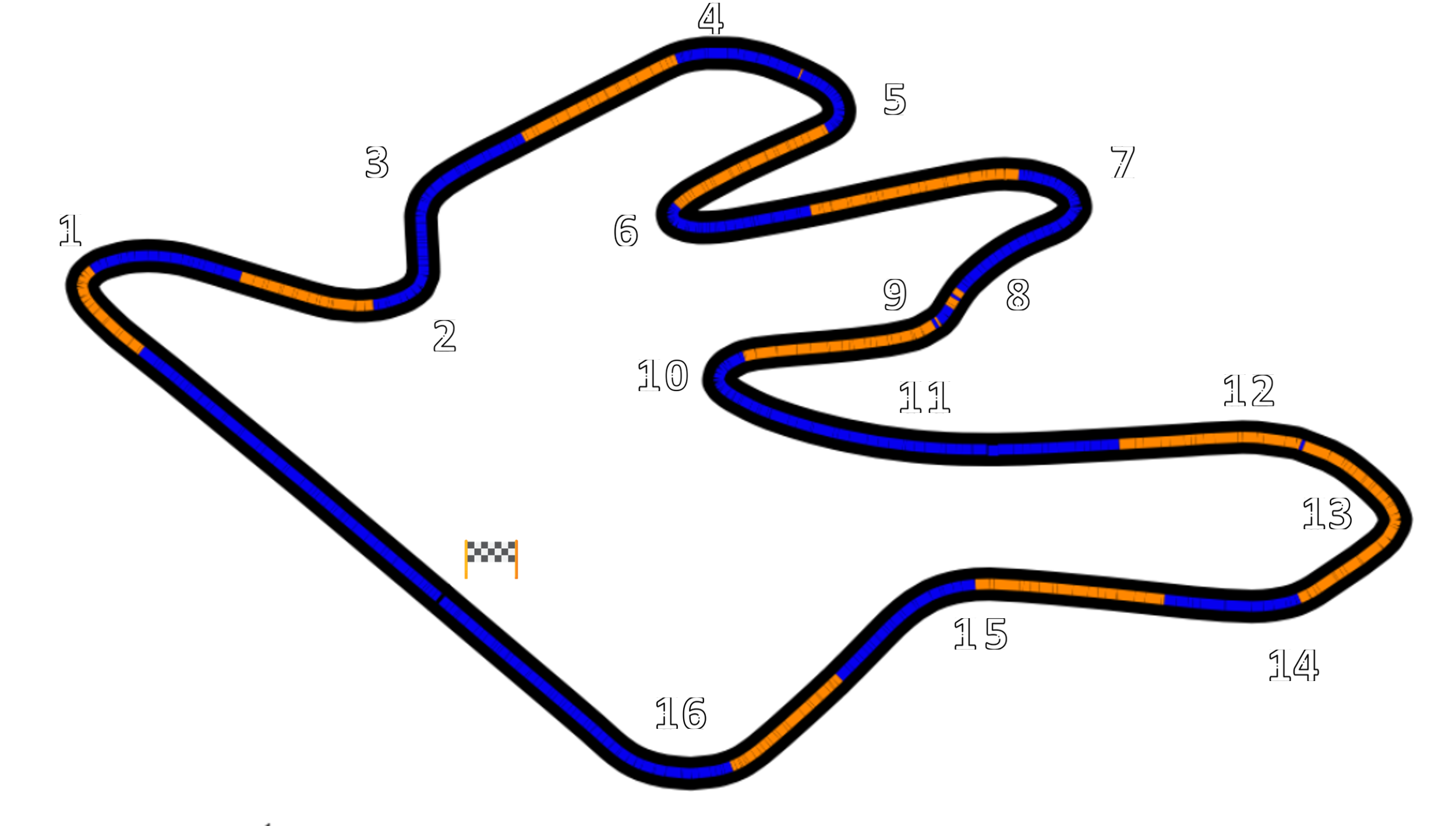
Immagine che contiene testo, Diagramma, linea, schermata

Descrizione generata automaticamente

Immagine che contiene testo, Diagramma, linea, schermata

Descrizione generata automaticamente

*Plot delle velocità dei piloti a confronto*

**

*Mappa con i punti con i tratti dei piloti più veloci*

These two images are the most used graphs by both Formula 1 analysts and motorsport journalists because they manage to communicate fundamental information to understand the difference between two or more drivers.

In the top image, the speed plot of the drivers is built so that they can be compared, and it is studied in which points one car prevailed over the other.

While the second graph highlights the sections, or portions of the track, with different colors, where one of the two cars was faster.

From both the first plot and the second mapping, it is possible to understand where improvements can be made, the strengths of one's car, and those of the competitor's car.

It can be observed that throughout the track, the speeds alternated at different points.

Starting from the first graph, the first important data point noted is the difference in speed recorded on the first straight, demonstrating the level Red Bull can reach on the straight before the braking for turn 1.

The two maintained similar speeds until the braking for turn 4, where the driver in blue braked a moment earlier, resulting in higher speed through turns 4 and 5 and gaining a slight advantage.

At the hairpin of turn 6, however, Norris managed to be faster, confirming greater agility in slow turns, but in acceleration, the two cars were equal.

Arriving at the rapid direction changes of turns 8 and 9, it is noted that the Red Bull had to lift slightly, while the McLaren did not lose speed before the braking for turn 10.

For the following turns, the three fast right-handers, there is an alternation of speeds between Verstappen and Norris, where the former has to slow down in the middle of the turn, while the latter maintains a more linear pace but with lower speeds.

Then comes turns 15 and 16, where the orange car showed great dominance.

The bottom mapping reflects well what was shown in the previous chapters: the first sector, mostly blue, is where Red Bull lapped with lower times than rivals; in the second, the two leading drivers were similar, while in the third, with faster turns, McLaren dominated.

Overall, the two cars used different approaches to setup.

McLaren, with higher aerodynamic downforce, proved to be very agile and stable, able to make direction changes much faster, thus closing the gap with the world champion mainly thanks to the third sector.

However, the Woking team had to sacrifice two things to have more downforce: the first is acceleration; in fact, despite being faster in the corners, its exit speed was at the same level as Red Bull or slightly lower; the second is top speed.

The Austrian team, on the other hand, opted for a more aerodynamically unloaded setup, allowing their cars to gain greater speed on the straights and in acceleration.

As seen earlier, at the end of the lap, Red Bull's setup allowed Verstappen to record a gap of about half a second.

### 4.3.2 The other drivers

Immagine che contiene testo, Diagramma, calligrafia, linea

Descrizione generata automaticamente

*Plot delle velocità di più piloti a confronto*

The telemetry data of other fastest drivers, such as Charles Leclerc with Ferrari, Russell with Mercedes, and Alonso with Aston Martin, are also compared to provide a more complete picture of the different characteristics of the cars.

It can be noted that in some parts of his best lap, Ferrari was faster than the drivers previously described, particularly showing peaks near braking points and higher values in fast corners.

This setup's result is clearly seen from the large gap that forms between turns 12 and 15.

Thus, the Maranello team opted for a car with even more aerodynamic downforce compared to McLaren, losing out primarily on the straights.

In contrast, the Silver Arrows chose a balance similar to Red Bull's; in fact, on the first straight, the two teams reached almost the same speeds, but in cornering speed, Russell was the slowest driver compared to all the others.

Finally, Fernando Alonso, with a configuration more in line with McLaren's but not at the same level, showed a different strength compared to other cars: the braking point. As seen especially in turn 1, he manages to brake a few meters later and more decisively. Another detail is in the last turn, as the Spanish driver, unlike his rivals, slows down more significantly to gain more speed on the long straight.

# 5. Performance on different tire compounds

As previously mentioned, teams must decide on their strategy, pit stop timing, and tire compound based on collected vehicle data and track temperatures before a race.

The choice of tire type is crucial, especially since not all cars perform equally on the same type of tire under the same track conditions.

For instance, some cars struggle to bring the Hard tires up to temperature, others wear out the Soft tires too quickly, while others record slower lap times on the Medium tires or experience excessive vibrations. All these factors also affect the driver's feel and grip on the track.

In this chapter, the normalized lap times of each team over the past five years, after the tire regulation change, will be analyzed.

## 5.1 Performance of Teams Divided by Tire Compound

Team performances depend not only on the tire compound but also on other factors such as different races and circuits, updates to the car over the years, and the stint in which a specific tire is used. Therefore, to compare team performances, some transformations are necessary.

The first step is the normalization of lap times: the average time is calculated for each compound, stint, team, GP, and year. Then, all lap times are divided by the corresponding average time, saved in the "Ratio\_to\_mean" variable.

In this way, all lap times are normalized and can be compared across different circuits in terms of length and structure.

Two plots are made for each compound, the first showing the performance of the 10 teams over these five years, and the second showing the performance trend over time.

### 5.1.1 Hard

Immagine che contiene diagramma, Rettangolo, schermata, quadrato

Descrizione generata automaticamente

*Boxplot dei tempi normalizzati per le Hard*

A boxplot is calculated for each team to observe how the lap times are distributed and where they stand relative to the average performance.

For example, with the Hard tires, Red Bull and Mercedes were the most performant teams, recording over 75% of their normalized times below the average. In fact, these two teams have shared the world championships over the past five years.

They are followed by Ferrari, which also performed better than the average.

McLaren, Aston Martin, and Alpine have their median distribution exactly at the average performance level. Williams was the team with the worst performance among the 10.

Immagine che contiene linea, diagramma, testo, Diagramma

Descrizione generata automaticamente

*Trend di crescita nel corso degli anni con le Hard*

Regarding performance growth, it is noted that Mercedes dominated in the first two years, while in 2021, Red Bull's continuous growth slightly surpassed them. That year, Mercedes won its eighth consecutive constructors' championship, but Max Verstappen became the world champion.

Ferrari experienced more inconsistent growth, although always below average performance, being beaten on the Hard tires by teams like Aston Martin and Alpine in 2020. However, in 2022, it managed to get closer to Red Bull, challenging them in the early races.

In 2023, the three strongest teams experienced a performance drop, probably due to the introduction of some safety regulations by the FIA.

### 5.1.2 Medium

Immagine che contiene testo, diagramma, schermata, Rettangolo

Descrizione generata automaticamente

*Boxplot dei tempi normalizzati per le Medium*

Immagine che contiene linea, diagramma, Diagramma, schermata

Descrizione generata automaticamente

*Trend di crescita nel corso degli anni con le Medium*

Regarding the Medium tires, the performances were similar to the Hard tires for almost all teams. Notably, McLaren was the third strongest team in 2020 and 2021, although with a significant gap from Mercedes and Red Bull.

### 5.1.3 Soft

Immagine che contiene testo, schermata, diagramma, Rettangolo

Descrizione generata automaticamente

*Boxplot dei tempi normalizzati per le Soft*

Immagine che contiene linea, diagramma, testo, Diagramma

Descrizione generata automaticamente

*Trend di crescita nel corso degli anni con le Soft*

Finally, the analysis of the softest and most performant compound shows that Mercedes was the top-performing team over these five years, surpassing Red Bull, followed closely by Ferrari, but with a smaller gap compared to the other compounds.

Regarding performance growth, the graph reflects the trends described earlier, with Ferrari's performance in 2021 standing out as better compared to other compounds.

## 5.2 Correlations with Temperatures

After analyzing the different compounds, the analysis continues by seeking a correlation between the tires and the temperatures recorded for each Grand Prix and year.

The hypothesis is that higher temperatures cause certain tires to wear out faster or record higher lap times.

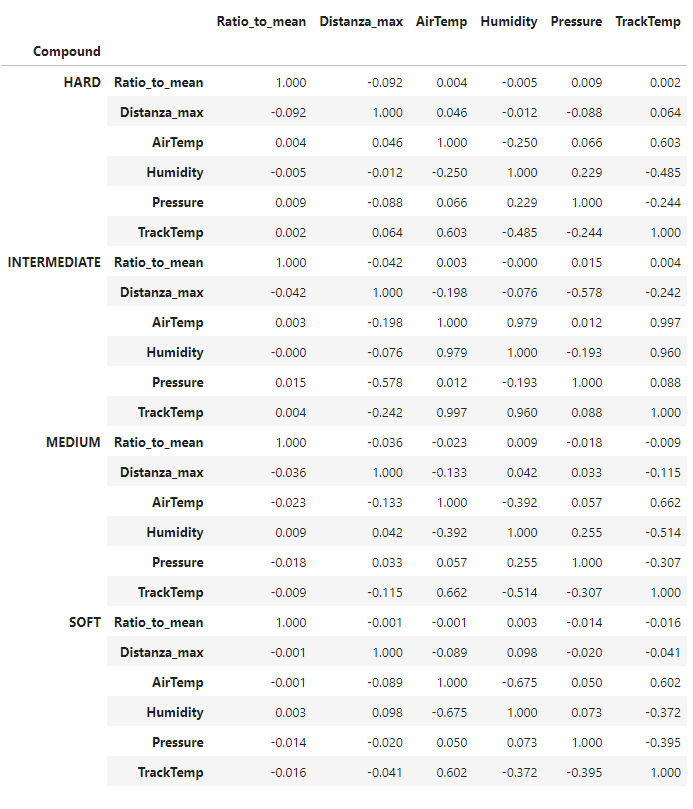
From the FastF1 library, a dataframe was created containing the average air and track temperatures, atmospheric pressure, and humidity for each year and GP.

After extracting the dataframe, a join was made on the year and GP to create a new table containing the average temperature of each race and other characteristics for each lap.

Since another aspect dependent on temperatures is tire duration, but each circuit has a different length, a new variable was calculated.

This variable takes the maximum number of laps the tire lasts in the stint before being changed and multiplies it by the track length to allow comparison in terms of distance traveled in kilometers rather than the number of laps.

At this point, the correlations are calculated.



From this output, it is noted that the two variables of interest, "Ratio\_to\_min" (normalized laps) and "Distanza\_max" (maximum distance traveled before changing the tire), are not correlated and thus do not depend on temperatures and pressure.

Only in the case of intermediate compounds, i.e., wet tires, is there a slight negative correlation between maximum distance and pressure; as pressure increases, the average distance traveled by the intermediate tires before needing to be changed decreases.

Generally, track temperatures affect car performance, lap times, and compounds, but it is a factor that requires considering other elements, making it difficult to quantify and predict even for Formula 1 teams.

# 6. Conclusions

The conclusions of this research represent the culmination of an in-depth investigation into the role of statistics in Formula 1. A brief summary of the analyses conducted will be provided, explaining the real magnitudes behind motorsport, followed by final considerations.

## 6.1 Summary

The information generated from a single lap is vast and can be utilized for various purposes. Initially, we explored data retrieval from FastF1, identified Python libraries to import and their primary purposes. We discussed how to select only data that met specific criteria, such as excluding laps slower than a certain threshold. This led to creating a dataframe containing all laps from 2019 to 2023 and various variables for subsequent analysis.

We proceeded with initial descriptive statistics like mean, median, and extreme values, and univariate analyses of lap times, sectors, and maximum speeds. Distributions of collected information were plotted by team for comparison, and correlation analysis demonstrated the relationship between increasing lap numbers and decreasing lap times. Correlation analysis was completed with a scatter plot and regression model estimation: initially defining the function for the D-Cook distances index to eliminate influential model values, calculating goodness of fit through R2 to observe retained information, and constructing line graphs for comparison.

The final step involved a detailed analysis starting with the two fastest laps of the race, those of Max Verstappen and Lando Norris, followed by inclusion of three additional drivers. A comparison of distances covered by the top two over time was conducted to determine where each gained an advantage on the track. Subsequently, speeds at different points of the lap were compared, and a map of dominant track segments was created to examine vehicle strengths, styles, and preferred types of corners. Finally, an additional comparison was repeated with three more drivers.

Finally, the performances of the 10 teams over these 5 years were analyzed based on tire compounds, followed by studying the relationship between performance and external environmental factors such as air and track temperatures.

## 6.2 The Real Measures

When comparing different driver times, we discussed gaps such as "half a second" or a few km/h more, but those deeply knowledgeable about motorsport understand the true value of 5 tenths of a second in motorsport. At a track like Qatar's, a lap spans 5.38 kilometers, typically completed in one minute and 27 seconds, navigating a variety of diverse corners. Therefore, drivers at the front must seek to gain an advantage over their competitors. In this specific instance, Max Verstappen, in his best lap, crossed the finish line half a second ahead of the next fastest time. On average, the Dutch driver gains about 0.1 to 0.2 seconds per lap over others.

Consider that humans take approximately 3 to 4 tenths of a second to blink, which is the same amount of time separating the top 5 drivers during qualifying to determine their starting order. Thus, while Formula 1 cars may appear extremely close in performance, the reality differs. Gaining approximately 0.1 to 0.2 seconds per lap holds significant value. This gap, without considering other factors, resulted in Red Bull dominating and winning every race of the 2023 World Championship except one, finishing with more than double the points of the second-place team.

## 6.3 Final Considerations

The FastF1 library offers additional information on aspects not covered. For example, it provides data on track and air temperatures throughout each Grand Prix, atmospheric pressure, and track structure. Each analysis allows for comparisons between different drivers, explores differences over time, and can analyze various real-time scenarios during a race.

Therefore, with this library, it is possible to conduct more advanced, precise, and high-level statistics. The creators of FastF1 offer the option to obtain some data in CSV or SQL format and have provided an API for conducting studies using other tools.

These chapters have demonstrated only a few examples of the analyses that can be performed with the extracted data. To conduct even more precise studies, it is necessary to consider additional factors such as tire compounds, tracks, grid positions, and the potential for safety cars.

The primary objective was to demonstrate how statistics applied to Formula 1 can reveal aspects not visible to the naked eye and illustrate that behind super-performing drivers and cars lies an even broader world that is often unknown to us.

1. <https://www.motosprint.it/news/motomondiale/moto-gp/2023/11/17-6855003/gp_qatar_curve_lunghezza_e_criticita_per_la_motogp> [↑](#footnote-ref-2)