A.a 2022-2023

DATA MANAGEMENT PROJECT REPORT

CREATION AND ANALYSIS OF A FORMULA ONE DATASET

Season 2013 vs. Season 2014 and Season 2021 vs. Season 2022

Eleonora Brambatti (858098), Matteo Pasotti (901810), Marta Privitera (898017)

18^{th} September 2023

Contents

1 INTRODUCTION

1.1 Formula 1 Championship

Formula One is one of the most globally renowned racing championship, where the most technologically advanced racing cars compete in different circuits around the world.
The first Formula One race took place on the 13^{th} of May 1950, from that moment onwards every team of racing cars have progressively evolved, thanks to increasingly sophisticated technologies and new regulations and rules have been implemented by the Fédération Internationale de l'Automobile (FIA), that is the association that organizes this championship; these rules and regulations cover various aspects, including car design, safety standards, and racing conduct, to ensure fair competition and driver safety. These changes made year after year are what make this sport increasingly thrilling and captivating for its fans. Our analysis will focus on the changes between year 2013 and year 2014 and then between year 2021 and 2022, but first we introduce briefly how is structured a Formula One Championship.

The Teams that participate to the competition are very famous, some of them are Ferrari, Mercedes, Red Bull and each Team is represented by two drivers for a total number of approximately 20 drivers per season. As we anticipated before, every year (season) a series of races called Grand Prix (GP) is held, generally from March to December. Every race is taken in a different circuit, the number of GP can change from a season to another one (in recent years, the number of circuits has been around 20). Before the race, practice and qualifying sessions are scheduled, all of them take place in different days.

Practice sessions: during the practice sessions, drivers and Teams can familiarize themselves with the track and test their car setups. In particular, they can collect data about the speed and time behaviour in order to analyze car's performance.

Qualifying sessions: qualifying sessions take place to determine the starting grid for the race. The actual format of the qualifying sessions is the following one: there are three segments called Q1, Q2 and Q3. In Q1, all drivers participate, and after a specific time, the slowest five or six drivers are eliminated and these drivers will start from the last positions on the starting grid. The remaining drivers move on to Q2. Similarly, in Q2, more drivers are eliminated, leaving the top ten to proceed to Q3. In Q3 the top 10 drivers compete for the pole position, that is the first starting position on the grid. The fastest driver in Q3 secures pole position on the grid and the other starting positions are determined based on the results of the qualifying sessions times.

Race: on race day, there is the true competition, in which drivers must complete a specific number of laps around the circuit to gain points. The winner is the first driver who crosses the finish line in the last lap. At the end of each GP, points are awarded to drivers based on their finishing position, for example, the winner receives 25 points, while the second driver receives 18 points and the third 15.

At the end of the season, after all the GPs have been contested, the final result is determined by summing up all the points accumulated in each GP by each driver. The driver with the highest number of points becomes the Formula 1 champion.
Simultaneously, the Team that wins the Constructors' Championship is the one that has obtained the highest total points by adding up the points earned by all its drivers throughout the season.

Now, as we anticipated, we will focus on two couples of years: 2013 and 2014 and then 2021 and 2022. We chose these particular years because in 2014 and in 2022 very important changes in regulations have been implemented. Let's see which are these changes in detail.

1.2 2013 vs. 2014

The new regulations and rules in 2014 are very important, here we show the most relevant changes between 2013 and 2014:

1. **Turbo engines**: the most relevant change concerns the engine, as turbo engines are introduced, and they are 1.6-liter V6 engines [1] (against the naturally aspirated 2.4-liter V8 engines used in 2013 [2].) This is a very important change because it marks the beginning of the "V6 turbo hybrid", until this year no significant changes on the engine were made, so this is a very significant news in the Formula One world.

2. **ERS**: the hybrid engine combined an internal combustion engine with an Energy Recovery System (ERS) [3]. The ERS is a novelty, because in 2013 was used a simpler and less powerful system called KERS (Kinetic Energy Recovery System) [4], a system that is designed to recover kinetic energy from the car during braking, store that energy and make it available to propel the car [5]. On the contrary, the new component is a system that is designed to recover energy from the car, store that energy and make it available to propel the car [6], this system is more advanced and more powerful. The more extensive energy recovery is allowed via a single MGUK (Motor Generator Unit - Kinetic) and/or a single MGUH (Motor Generator Unit - Heat) [3]. The MGUK has more or less the same function of the KERS: recovering kinetic energy generated during braking, but it is more efficient and provides higher energy recovery capabilities. The true novelty was the MGUH as it allows the recovery of waste heat from the turbocharger. It transformed this heat energy into electrical energy, further enhancing the overall energy recovery process.

3. **Number of available engines**: in 2013, during the entire season, each driver was allowed to change the engine 8 times, in 2014, the number of available engines during the season is just 5.

4. **Fuel limit**: In 2014, the limit of fuel is 100 Kg per race [7], while in 2013 there was no limit imposed on the teams. This is a very challenging change, in fact, with this restriction, teams are encouraged to to develop power units that could complete the race distance with the allocated fuel and, consequently, to elaborate new race strategies.

As we can see these are very important changes in F1 world, that can have a very strong impact on car performances, and that's why we are interested in them.

1.3 2021 vs. 2022

Now we focus on the other couple of years. As in 2014, even 2022 sees important technical changes in the competition[8], they are:

1. **Aerodynamics**: technical regulations of 2022 have changed the aerodynamics, as they eliminate the part of components that created violent turbulence destabilizing the chasing cars. The main difference between 2022 and the previous years is that aerodynamic downforce is no more generated uniformly starting from the front wing, but from the car's floor, favoring the drivers behind ground-effect cars, will have more opportunities to get closer to their rivals at any point on the circuit, and consequently, more overtaking maneuvers are expected.

2. **Fuel**: in 2022, there is a little transition to sustainable fuel. In fact, in 2022 the fuel used is called E10 fuel and it is a blend containing 10% ethanol. This is a choice made to reduce the environmental impact of Formula One.

3. **18-Inch Wheels**: there is a change in the inch wheels between 2021 and 2022, they move from 13 to 18. The new 18-inch wheels are provided by BBS to all the teams. The larger 18-inch wheels provided several advantages, such as better grip, improved tire performance, and enhanced handling characteristics, all of which contributed to a more modern and visually striking appearance of the Formula 1 cars.

4. **Brakes**: the braking system features larger carbon discs (diameter increased from 278 to 330 mm) with fewer cooling holes (from 1,500 to 1,100) but wider. The geometries and volumes of the calipers, pads, and

pumps have also been redesigned. Additionally, from the 2022 F1 Championship, the brake cooling ducts will be the same for all the teams. This is to prevent the brake ducts from being used – as in previous years – to direct flows and generate aerodynamic downforce.

5. **Weight**: the single-seaters of the 2022 F1 Championship will be the heaviest ever: with a minimum regulated weight of 795 kg (against the 752 kg in 2021 and previous years), to which the 80 kilograms of the driver-seat assembly and the approximately 110 kilograms of fuel must be added.

6. **Tyres**: the Pirelli tires mounted on the new 18-inch wheels will have lower and stiffer sidewalls compared to last year's tires, making them less prone to overheating. This will allow drivers to worry less about tire temperatures.
   All the tires must be heated to 70 degrees Celsius (in 2021, the maximum temperature for tire warmers was 100 degrees Celsius at the front and 80 degrees Celsius at the rear).

1.4 Research question

Now, after all these premises and this introduction, what we want to know is: how have the performances of cars changed from 2013 to 2014? And from 2021 to 2022? Which is the real impact in GP results with all these important technical changes? How do these new regulations effect the times of car during a race? Is there an improvement or a deterioration in results from one year to another?

These are our starting questions, which form the foundation for creating the dataset that will contain the necessary information enabling us to provide answers.

2 DATA ACQUISITION

To accomplish our objectives, we used data gathered from three sources: FIA.com that is the official website of Formula 1 competition, Wikipedia, which served as a valuable resource for obtaining data related to Formula 1 circuits and weather conditions and Pitwall.app website providing comprehensive information about motorsport events such as Formula 1. The datasets were obtained by accessing these sources and extracting the required information using appropriate acquisition methods.

Data we collected can be considered **historical data**. This data includes information about race and qualifying results, lap times of drivers, and other details about past races. As time goes by, this information becomes part of the history of Formula 1 and can be used to trace the evolution of the championship over the years. Similarly, meteorological data related to Formula 1 Grand Prix has been collected. This data includes information about weather conditions during the races, such as whether it was a rainy or sunny day for each race and qualifying sessions. This data is crucial for Formula 1 Teams and drivers, as weather conditions can significantly influence car performance and race strategies.

Both historical race data and meteorological data can be used for retrospective analysis, comparative studies between different seasons, and to assess the impact of environmental conditions on driver and car performances over the years of our interest. In particular, as we just said before, we focused on the Formula 1 data of 2013 and 2014, 2021 and 2022.

2.1 Benchmark Data Acquisition

The data acquisition process has the following structure (reported within the main.py file in the Data Acquisition folder):

- Acquisition of drivers, and for each driver races and qualifications data in the years 2013-2014 and 2021-2022.

- Acquisition of the car Team for each driver.
- Acquisition of circuit names for each year.
- Acquisition of the track length and the number of turns for each circuit.
- Acquisition of the maximum speed during the qualifications.
- Acquisition of meteorological data for qualifications and races.

The codes for the first four Data Acquisition processes are stored within the WS_Season*year*.py files, whereas the retrieval of maximum speeds during qualifications is performed using the PDFtoJSON_*year*.py files. Furthermore, the scripts responsible for capturing meteorological conditions during both qualifications and races can be found in the WS_Weather*year*_Qualifying.py and WS_Weather*year*_Race.py files, respectively (where *year* stands for 2013, 2014, 2021 or 2022).

2.2 Drivers, races and qualifications Data Acquisition

We collected data about qualification and race times of each driver. We used the Web Scraper tool to collect data from the Pitwall.app platform which provides all kind of Formula 1 statistics, in-depth analysis, lap time comparison, standings and race results. The web scraping gather data related to the 2013-2014 and 2021-2022 Formula 1 seasons in four different files. We used the BeautifulSoup library for parsing HTML and the Selenium library for web automation. In particular the following steps have been followed:

1. Importing necessary libraries and modules such as re, time, json, itertools, BeautifulSoup, selenium, and fuzzywuzzy.
2. Setting up the ChromeDriver for web scraping.
3. Defining lists for seasons, races, and drivers: these lists contain identifiers for the specific years, Gran Prix races, and drivers of interest for web scraping.
4. Specifying the year of interest.
5. The script iterates through all possible combinations of drivers, races, and seasons and visits the corresponding URLs using Selenium. It waits for elements to be present on the page and uses BeautifulSoup to extract data from specific sections of the website. In particular, we create a pattern to extract Year, Grand Prix Name, and Driver and we remove the sections that are not useful. Since information about various aspects of the race such as overtakes, pit stops, incidents, and yellow flags are not particularly relevant for the purposes of our analysis, which aims to primarily focus on studying the timings and speeds of the cars independent of specific and contingent race characteristics, we decide to remove the fields about "Race progression", "pit stops", "Summary", and "Practice".
6. We insert the Qualifying section and Laps section of each Driver and Grand Prix.
7. The extracted data are stored in a nested dictionary named *complete_data*.

At this point each driver's data is represented as a nested dictionary:

```
{
    "Driver_Name_1": {
        "Name_of_GP_1": {
            "Qualifying": {
                "Q3 Time": [
                    "Min:Sec.Millisec"
                ],
                "Q2 Time": [
                    "Min:Sec.Millisec"
                ],
```

```
                "Q1 Time": [
                    "Min:Sec.Millisec"
                ]
            },
            "Laps": {
                "1": "Min:Sec.Millisec",
                "2": "Min:Sec.Millisec",
                ...
            }
        },
        "Name_of_GP_2": {
            "Qualifying": {
                ...
            },
            "Laps": {
                ...
            }
        },
        ...
    }
        ...
}
```

The outermost key represents the driver's name, and the value is another dictionary containing information about each Grand Prix race the driver participated in. Within each Grand Prix race dictionary, there are two main sections:

- "Qualifying": this section contains the qualifying times for the driver in the Q3, Q2, and Q1 sessions of the race.
- "Laps": this section contains lap times for each lap completed by the driver during the race. The lap times are represented as "Min:Sec.Millisec" format.

The structure repeats for each driver and their respective Grand Prix races, forming a comprehensive dataset of performance metrics for the drivers in the given Formula 1 season. An example of the page website on which has been applied Web Scraping technique is provided in the Figure 1.

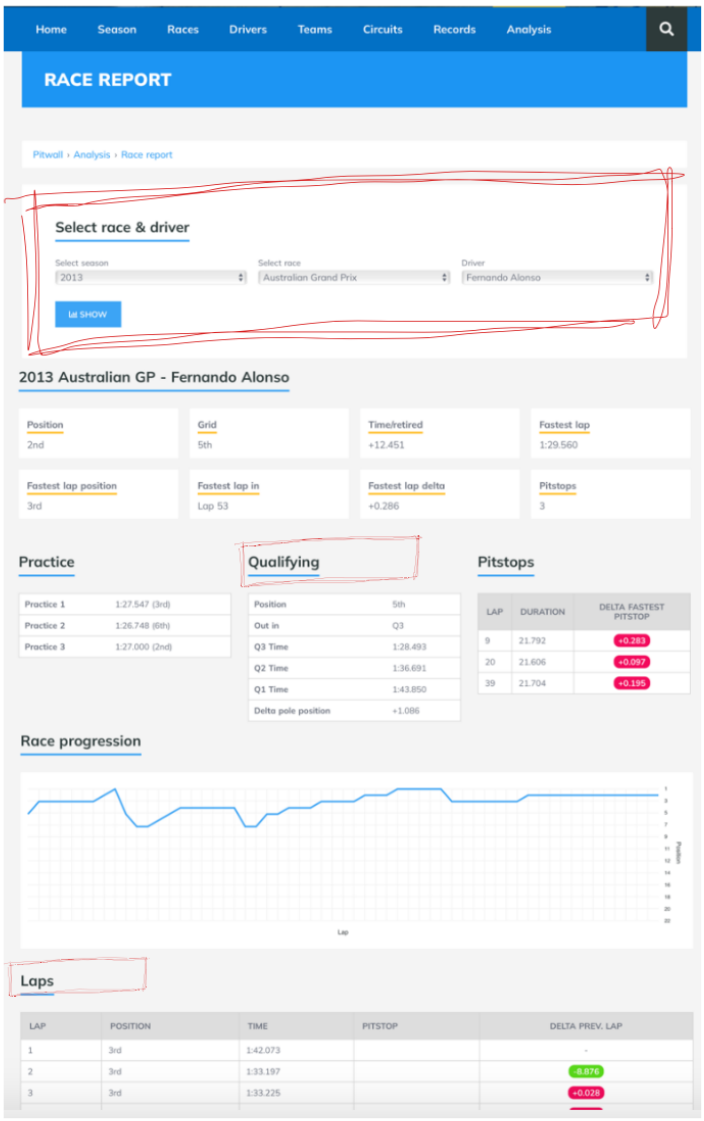


Figure 1: Example of website page taken from Pitwall.com: `https://pitwall.app/analysis/race-report?utf8=&season=7&race=122&driver=22&button=`

2.3 Driver Car Data Acquisition

We perform web scraping operations to obtain data about drivers and cars during Gran Prix races and Qualifying Sessions.

1. A specific webpage for each year (https://pitwall.app/seasons/year-formula-1-world-championship) is visited using the Chrome driver.
2. The HTML page is parsed using BeautifulSoup.
3. The table containing the data about drivers and cars is selected.
4. For each row of the table (excluding the first row, which contains the header), the data from the columns corresponding to the driver and car are extracted.
5. The data is saved in the *driver_car_data* dictionary in the format {"Driver": driver, "Car": car}. Subsequently, the car data is inserted into the main dictionary *complete_data* based on the corresponding driver's name. If a match of at least 75% similarity is found between the driver's name and an existing key in *complete_data*, the car is assigned to the driver.

Now each driver's data is represented as follows:

```
{
    "Driver_Name_1": {
        "Name_of_GP_1": {
            "Qualifying": {
                ...
            },
            "Laps": {
                ...
            }
        },
        "Name_of_GP_2": {
            ...
        },
            ...
        "Car": "Name_of_Car"
    }
    "Driver_Name_2": {
            ...
    }
        ...
}
```

- "Car": This field indicates the name of the car that the driver is racing with in the Formula 1 season.

The JSON data structure is organized as above, with driver names at the top level, followed by specific Grand Prix names attended by each driver, and further details such as qualifying and Laps data associated with each Grand Prix. We added the "Car" field which is common for each Grand Prix and represents the car they are driving throughout the season.

2.4 Circuit Data Acquisition

We perform the web scraping to obtain data about the circuit names for the Formula 1 season specified by the $year\_of\_search$ variable.

1. We construct the URL by filling the base Wikipedia URL with the $year\_of\_search$ value: "$https://en.wikipedia.org/wiki/year\_of\_search\_Formula\_One\_World\_Championship$".

2. The code visits the constructed URL using the Chrome driver. It waits until the page element with the CSS selector ".wikitable" is present before proceeding. The HTML source of the page is parsed using BeautifulSoup.

3. We select all tables with the class "wikitable" on the page. The third table is considered the table of interest (index 2 in the tables list).

4. The code extracts all rows from the table. An empty list called $Circuit\_Data$ is initialized to store the circuit data.

5. For each row in the table, the code extracts the columns' data and saves it in a dictionary. The "Grand Prix" text in the race name column is replaced with "GP" for compatibility with the previously collected data (and for future merging).

6. The circuit data (circuit name and race name) is appended to the $Circuit\_Data$ list as dictionary in the format "Circuit": $circuit\_name$, "Prix": $race\_name$. The code then inserts the circuit data into the main $complete\_data\_dictionary$ for each driver and Grand Prix based on the corresponding name of Grand Prix.

7. Finally, the circuit data is inserted into the *complete_data* dictionary under the respective driver and GP, with the circuit name stored as "Name" and the initial values of "Length" and "Turns" set to None. The structure of the *complete_data* dictionary is now:

```
{
    "Driver_Name_1": {
        "Name_of_GP_1": {
            "Qualifying": {
                ...
            },
            "Laps": {
                ...
            },
            "Circuit": {
                "Name": "Name_of_Circuit",
                "Length": None,
                "Turns": None
            }
        },
        "Name_of_GP_2": {
            ...
        },
            ...
        "Car": "Name_of_Car"
    }
    "Driver_Name_2": {
            ...
    }
        ...
}
```

The JSON dictionary now includes an additional field called "Circuit" for each Grand Prix attended by a driver.

- "Circuit": this field contains information about the specific circuit where the Grand Prix race took place.
  - "Name": this sub-field holds the name of the circuit where the Grand Prix was held.
  - "Length": this sub-field represents the length of the racing circuit in kilometers or miles.
  - "Turns": this sub-field provides the number of turns on the racing circuit.

  The "Circuit" field is nested within the Grand Prix data for each driver, meaning that the circuit information is associated with each specific Grand Prix they participated in.

2.5 Data Acquisition of the track length and of the number of turns

In order to fill "Length" and "Turns" fields we used Selenium. We scraped Wikipedia pages for circuit information, about circuit length and number of turns, and update the dictionary (*complete_data*) with this information for each driver and Grand Prix.

1. The URL associated with the specific circuit is constructed using a Wikipedia URL template:

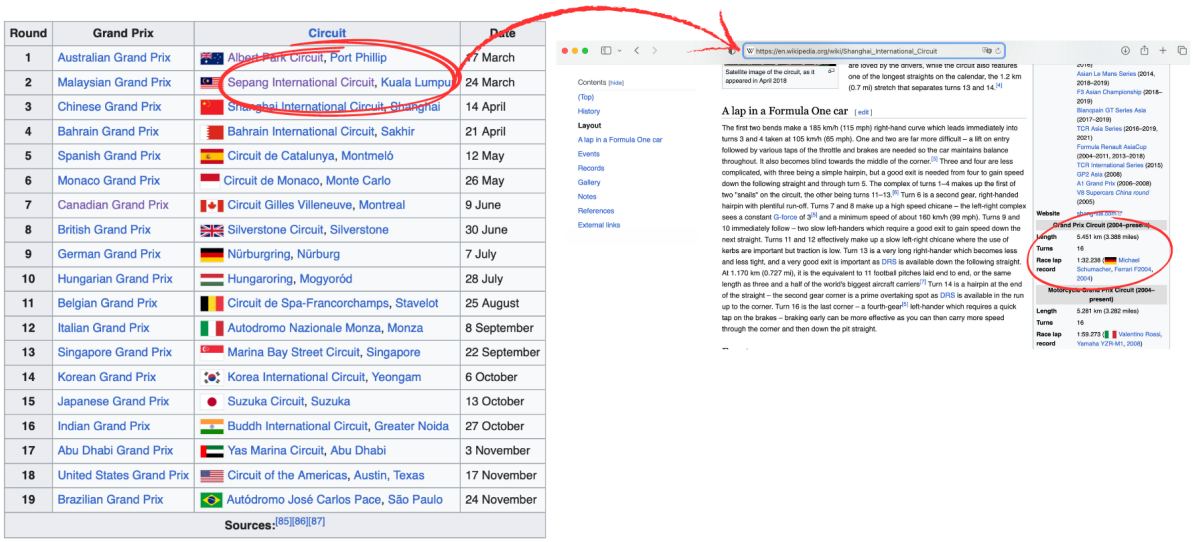
   "https://en.wikipedia.org/wiki/_url".

   Selenium's *chrome_driver* navigates to the constructed URL and the HTML source code of the webpage is parsed using BeautifulSoup to create a soup object.

2. The index of the first occurrence of the string "Grand Prix Circuit" is found within the table element of the webpage and manipulated to remove noise and unwanted variations.

3. An empty list called *date_matches* is initialized.
4. Year ranges and other date variations are extracted and added to the *date_matches* list.
5. The code checks each date range or single date in *date_matches* to see if it matches a reference year.
6. If there is a match, the circuit information (length and turns) is extracted and added to the *stats_circuit* list which has been previously initialized to store the circuit statistics.
7. The code iterates over the *complete_data* dictionary for each driver and GP. If the circuit name matches a circuit in *stats_circuit*, the circuit length and number of turns are added to the respective entry in *complete_data*.

An example of Web Scraping functioning is shown in Figure 2.



Round	Grand Prix	Circuit	Date
1	Australian Grand Prix	Albert Park Circuit, Port Phillip	17 March
2	Malaysian Grand Prix	Sepang International Circuit, Kuala Lumpur	24 March
3	Chinese Grand Prix	Shanghai International Circuit, Shanghai	14 April
4	Bahrain Grand Prix	Bahrain International Circuit, Sakhir	21 April
5	Spanish Grand Prix	Circuit de Catalunya, Montmeló	12 May
6	Monaco Grand Prix	Circuit de Monaco, Monte Carlo	26 May
7	Canadian Grand Prix	Circuit Gilles Villeneuve, Montreal	9 June
8	British Grand Prix	Silverstone Circuit, Silverstone	30 June
9	German Grand Prix	Nürburgring, Nürburg	7 July
10	Hungarian Grand Prix	Hungaroring, Mogyoród	28 July
11	Belgian Grand Prix	Circuit de Spa-Francorchamps, Stavelot	25 August
12	Italian Grand Prix	Autodromo Nazionale Monza, Monza	8 September
13	Singapore Grand Prix	Marina Bay Street Circuit, Singapore	22 September
14	Korean Grand Prix	Korea International Circuit, Yeongam	6 October
15	Japanese Grand Prix	Suzuka Circuit, Suzuka	13 October
16	Indian Grand Prix	Buddh International Circuit, Greater Noida	27 October
17	Abu Dhabi Grand Prix	Yas Marina Circuit, Abu Dhabi	3 November
18	United States Grand Prix	Circuit of the Americas, Austin, Texas	17 November
19	Brazilian Grand Prix	Autódromo José Carlos Pace, São Paulo	24 November
Sources:[85][86][87]			

Figure 2: Web Scraping of Circuit characteristics from Wikipedia page.

The dictionary now appears as follows:

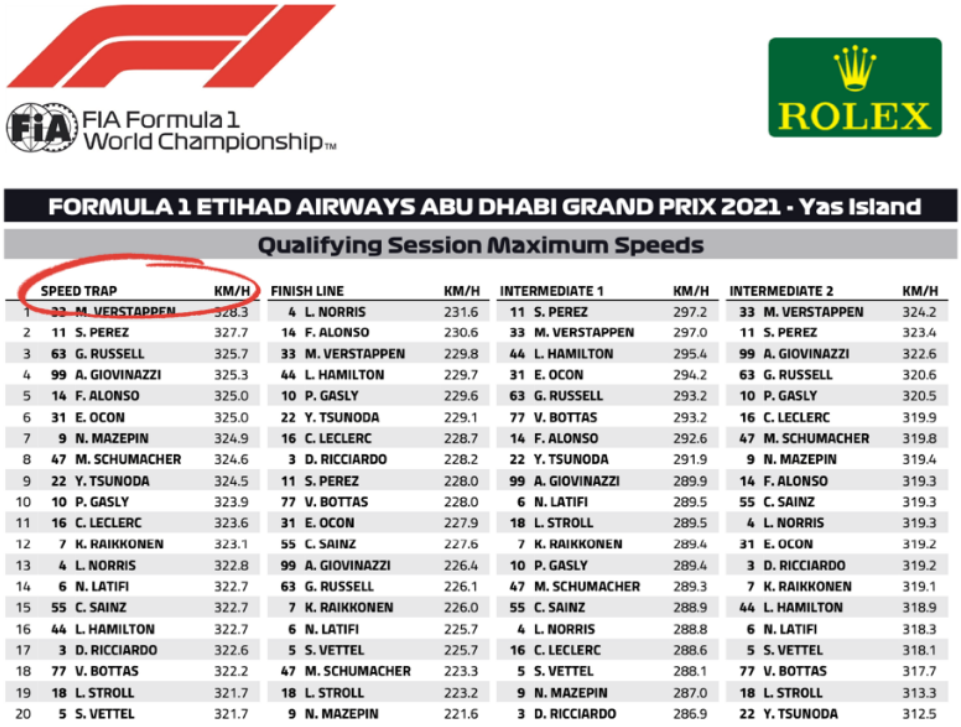
```
{
    "Driver_Name_1": {
        "Name_of_GP_1": {
            "Qualifying": {
                "Q3 Time": [
                    "Min:Sec.Millisec"
                ],
                "Q2 Time": [
                    "Min:Sec.Millisec"
                ],
                "Q1 Time": [
                    "Min:Sec.Millisec"
                ]
            },
            "Laps": {
                "1": "Min:Sec.Millisec",
                "2": "Min:Sec.Millisec",
```

```
                ...
            },
            "Circuit": {
                "Name": "Name_of_Circuit",
                "Length": Length_in_km,
                "Turns": Number_of_turns
            }
        },
        "Name_of_GP_2": {
            "Qualifying": {
                ...
            },
            "Laps": {
                ...
            },
            "Circuit": {
                ...
            }
        },
            ...
        "Car": "Name_of_Car"
    }
    "Driver_Name_2": {
            ...
    }
        ...
}
```

So now "Name", "Length" and "Turns" fields are no more empty.

2.6 Data Acquisition of the Maximum Speeds during Qualifying Sessions

We acquired the maximum speeds reached by the drivers during a specific Grand Prix Qualifiying session and year. The scripts use the PyPDF4 library to extract data from PDF files. This library takes a directory of PDF files (for example the one shown in the Figure 3) related to the year (Qualifying_Session_Max_Speed_*year*) as input and processes each file individually.



FIA Formula 1 World Championship™

ROLEX

FORMULA 1 ETIHAD AIRWAYS ABU DHABI GRAND PRIX 2021 - Yas Island

Qualifying Session Maximum Speeds

	SPEED TRAP	KM/H	FINISH LINE	KM/H	INTERMEDIATE 1	KM/H	INTERMEDIATE 2	KM/H
1	33 M. VERSTAPPEN	328.3	4 L. NORRIS	231.6	11 S. PEREZ	297.2	33 M. VERSTAPPEN	324.2
2	11 S. PEREZ	327.7	14 F. ALONSO	230.6	33 M. VERSTAPPEN	297.0	11 S. PEREZ	323.4
3	63 G. RUSSELL	325.7	33 M. VERSTAPPEN	229.8	44 L. HAMILTON	295.4	99 A. GIOVINAZZI	322.6
4	99 A. GIOVINAZZI	325.3	44 L. HAMILTON	229.7	31 E. OCON	294.2	63 G. RUSSELL	320.6
5	14 F. ALONSO	325.0	10 P. GASLY	229.6	63 G. RUSSELL	293.2	10 P. GASLY	320.5
6	31 E. OCON	325.0	22 Y. TSUNODA	229.1	77 V. BOTTAS	293.2	16 C. LECLERC	319.9
7	9 N. MAZEPIN	324.9	16 C. LECLERC	228.7	14 F. ALONSO	292.6	47 M. SCHUMACHER	319.8
8	47 M. SCHUMACHER	324.6	3 D. RICCIARDO	228.2	22 Y. TSUNODA	291.9	9 N. MAZEPIN	319.4
9	22 Y. TSUNODA	324.5	11 S. PEREZ	228.0	99 A. GIOVINAZZI	289.9	14 F. ALONSO	319.3
10	10 P. GASLY	323.9	77 V. BOTTAS	228.0	6 N. LATIFI	289.5	55 C. SAINZ	319.3
11	16 C. LECLERC	323.6	31 E. OCON	227.9	18 L. STROLL	289.5	4 L. NORRIS	319.3
12	7 K. RAIKKONEN	323.1	55 C. SAINZ	227.6	7 K. RAIKKONEN	289.4	31 E. OCON	319.2
13	4 L. NORRIS	322.8	99 A. GIOVINAZZI	226.4	10 P. GASLY	289.4	3 D. RICCIARDO	319.2
14	6 N. LATIFI	322.7	63 G. RUSSELL	226.1	47 M. SCHUMACHER	289.3	7 K. RAIKKONEN	319.1
15	55 C. SAINZ	322.7	7 K. RAIKKONEN	226.0	55 C. SAINZ	288.9	44 L. HAMILTON	318.9
16	44 L. HAMILTON	322.7	6 N. LATIFI	225.7	4 L. NORRIS	288.8	6 N. LATIFI	318.3
17	3 D. RICCIARDO	322.6	5 S. VETTEL	225.7	16 C. LECLERC	288.6	5 S. VETTEL	318.1
18	77 V. BOTTAS	322.2	47 M. SCHUMACHER	223.3	5 S. VETTEL	288.1	77 V. BOTTAS	317.7
19	18 L. STROLL	321.7	18 L. STROLL	223.2	9 N. MAZEPIN	287.0	18 L. STROLL	313.3
20	5 S. VETTEL	321.7	9 N. MAZEPIN	221.6	3 D. RICCIARDO	286.9	22 Y. TSUNODA	312.5

Figure 3: This pdf provide the 2021 Abu Dhabi Grand Prix data, which is then converted into JSON format.

Each file concerns a specific Grand Prix (example: Abu Dhabi GP - 2021.pdf).

1. The script iterates through each file in the input directory, checking if it has a ".pdf" extension.
2. For each PDF file, the code opens the file and extracts the text from each page.
3. Regular expressions are used to extract specific data from the extracted text: driver names, maximum speeds, Grand Prix names, and years. The extracted data are stored in a list of dictionaries.
4. The script creates a corresponding output file path by replacing the ".pdf" extension with ".json".
5. The extracted data is converted to a JSON string with proper indentation.
6. The JSON data is written to the output file.

An illustrative instance .pdf file transformed to a .json file is the following one concerning the 2021 Abu Dhabi Grand Prix:

```
[
  {
    "driver": "M. VERSTAPPEN",
    "km/h": "328.3",
    "gp": "Abu Dhabi GP",
    "year": 2021
  },
  {
    "driver": "S. PEREZ",
    "km/h": "327.7",
    "gp": "Abu Dhabi GP",
    "year": 2021
```

```
  },
        ...
]
```

2.7 Data Acquisition of Qualifying Sessions Weather

Qualification Data Acquisition is performed to obtain weather data for the qualifying sessions of Formula 1 in the year 2013-2014 and 2021-2022.

1. The code starts by defining a list of Grand Prix names and an empty dictionary to store the results. It then defines a function, *ottenere_testo_tra_sezioni*, which uses the BeautifulSoup library to extract text between specified sections on a given webpage.

2. Next, the code iterates through each Grand Prix name and constructs the URL for the corresponding Wikipedia page "https://en.wikipedia.org/wiki/*year*_nome_gran_premio_Grand_Prix".

3. In 2013, the *ottenere_testo_tra_sezioni* function directly extracts the text between the "Qualifying" and "Race" sections. For the year 2014, the code uses the same function to extract the text between the "Practice_and_Qualifying" and "Race" sections from the Wikipedia page. If the desired content is not found, it then searches for the text between the "Qualifying" and "Race" sections. A similar process is followed for the years 2021 and 2022, where the *ottenere_testo_tra_sezioni* function extracts the text between the "Qualifying" and "Sprint_qualifying" sections on the Wikipedia page. If the desired content is not found, it again looks for the text between the "Qualifying" and "Race" sections.

4. The code then examines the extracted text to identify weather-related keywords (e.g., "wet", "rain", "rainfalls", "Rain", " Rain", " Rainfalls", " wet") in order to determine the weather condition during the qualifying session. If any relevant keyword is detected, the weather condition is categorized as "rain"; otherwise, it is classified as "sun". However, in the specific case of the year 2014, if the sentence "tyres, took ninth." is also found in the text examined, the code automatically sets the weather condition to "sun".

5. For Grand Prix classified as "sun", the code proceeds to scrape the corresponding Italian Wikipedia page to check for additional weather-related keywords:

   parole_chiave_meteo_italiano = ["bagnata", "bagnato", "pioggia", "acqua", "gomme da bagnato", "pista scivolosa", "spray d'acqua", "visibilità ridotta", "aquaplaning", "pneumatici da pioggia", "safety car"]

   If any of these keywords are found in the extracted text from the Italian Wikipedia page, the code proceeds with another check. It looks for specific sentences to determine if the weather condition should be updated to "sun":

   "set up da bagnato" (just in 2014 code), "si svolge senza pioggia", "possibile arrivo della pioggia" , "non c'è minaccia della pioggia".

   If any of these sentences are found, the weather condition is updated to "sun". On the other hand, if none of these sentences are found, the code then uses the keywords from the *parole_chiave_meteo_italiano* list to determine if the weather condition should be updated to "rain".

   This multi-step process provides a comprehensive analysis of weather conditions during the qualifying sessions of Grand Prix races classified as "sun" based on information gathered from the Italian Wikipedia page. By checking for both specific sentences and additional weather-related keywords, the code enhances the accuracy of weather classification for these races.

6. After processing all the Grand Prix, the code formats the results and saves them in a JSON file named "Weather_Qualifying.json" in the specified directory related to the year 2013, 2014, 2021 or 2022.

   As an example, let's consider the JSON file containing data for the 2013 Grand Prix events:

```
{
    "Australian GP": "rain",
    "Malaysian GP": "rain",
    "Chinese GP": "sun",
    "Bahrain GP": "sun",
    "Spanish GP": "sun",
    "Monaco GP": "rain",
    "Canadian GP": "rain",
    "British GP": "sun",
    "German GP": "sun",
    "Hungarian GP": "sun",
    "Belgian GP": "rain",
    "Italian GP": "sun",
    "Singapore GP": "sun",
    "Korean GP": "sun",
    "Japanese GP": "sun",
    "Indian GP": "sun",
    "Abu Dhabi GP": "sun",
    "United States GP": "sun",
    "Brazilian GP": "rain"
}
```


2.8 Data Acquisition of Race Session Weather

We scraped weather data for the Grand Prix races of Formula 1 in the year 2013-2014 and 2021-2022. We used the BeautifulSoup library to extract weather information from Wikipedia pages.

1. The code begins by defining a list of Grand Prix race names for the specific *year* season and initializes an empty dictionary to store the results of weather data for each race.

2. It then defines a function called $ottenere\_testo\_tra\_sezioni(url)$, which is used to retrieve the weather information from a specific Wikipedia page:

   "https: //en.wikipedia.org/wiki/*year* _nome_gran_premio_Grand_Prix".

   The function sends an HTTP request to the provided URL, parses the content using BeautifulSoup, and searches for the table row containing the weather information. If found, it extracts the weather data from the corresponding table cell.

3. Next, the code iterates through each Grand Prix race name in the list and constructs the URL for the corresponding Wikipedia page. It calls the $ottenere\_testo\_tra\_sezioni(url)$ function to obtain the weather information for that race.

4. After obtaining the weather information, the code checks for specific keywords (e.g., "wet", "rain", "rainfalls") in the extracted weather data to determine the weather condition during the race. If any relevant keyword is found, the weather condition is classified as "rain"; otherwise, it is classified as "sun".

5. Finally, the code stores the race weather data in a dictionary with the Grand Prix race name as the key and the weather condition as the value. The resulting dictionary is then saved in a JSON file named "Weather_Race.json" in the specified directory of the year of reference.

The internal structure of the JSON is identical to the one used for the qualifications, and for brevity, it is not repeated here.

2.9 Reproducibility of the data

The data we have collected is static in time, unless the provider modifies the structure or content of the HTML pages. As a result, these aspects of the project are fully reproducible and will yield the same results as the ones we have stored. The data obtained using the Web Scraper tool or by converting PDF to JSON can be easily re-collected through the same procedures we employed. This can be done by accessing the pages through the links provided in the text above or by executing Python code to convert the files again.

2.10 Data Acquisition and Data Integration

To arrive at the creation of the all final JSON files after the data acquisition process, there was also a data integration process involved. However, we will delve into this data integration process in detail later within the chapter dedicated to Data Integration.

3 DATA QUALITY ON RAW DATA

3.1 Completeness

At this point, we had to check on the JSON file we created in order to see if the data we collected were correct, in particular we had to check that all the fields we needed were in the right place with the corresponding values.

3.1.1 Season JSON

We started from the four Season JSON files that contained a lot of fields and values. To check if everything with the code had worked correctly and if we got what we wanted in terms of creation of fields and completeness of values, we followed the following schema that reflects the JSONs structure:

→ checking the number of drivers
→ checking the number of GPs per driver
→ checking the existence of the "Car" field for each driver with respective value
→ checking the existence of the "Qualifying", "Laps", "Circuit" fields for each GP
→ checking the existence of "Q1 Time", "Q2 Time" and "Q3 Time" in the "Qualifying" field with respective values
→ checking the existence of all numbers of laps in the "Laps" field with respective values
→ checking the existence of "Name", "Length" and "Turns" in the "Circuit" field with respective values

We wrote a code that iterates over each year and that gives us various output useful to state if our collected data are good or not. We opened our JSONs file in order to read the elements contained in them.

1. Number of drivers. We obtained this data measuring the length of the file, which means measuring the number of dictionaries contained in each file, due to the fact that each dictionary refers to a driver, counting the number of the dictionaries is the same as counting the number of the drivers. We obtained the following results:

Year	Number of drivers
2013	23
2014	24
2021	21
2022	22

Table 1: Number of drivers per year

All of these results showed in Table 1 are the correct numbers of drivers for these four years. So we are sure we have a dictionary per driver in which the name of the driver is the primary key.

2. Number of GPs. For counting the number of GPs for each driver we iterated over each driver, in this way we had access to each driver dictionary were the keys are the name of the GPs. We measured the length of each driver dictionary (minus 1, because the last of the dictionary should be "Car" and not a GP name and so we have not to count it) in order to obtain the number of the elements (GPs) for each driver. We found out that for each year all of the drivers had the same number of elements in their dictionaries, so we concluded that none of the GPs was missing for none of the drivers. In particular, we obtained the following results:

Year	Number of GPs
2013	19
2014	19
2021	22
2022	22

Table 2: Number of GPs for each driver per year

The results showed in Table 2 are reported correctly for each year. At this point we could say that in each driver dictionary there are the same and correct number of GPs. So, also this data was collected correctly and no data was missing.

3. "Car" field. To check if the acquisition of the car for each driver worked correctly in the loop that iterates over each driver if the "Car" field is found a new key (the driver name) with its value (the "Car" field value) is added in a smaller dictionary *driver_car* that only contains these information.
When the code ends the iterations over each driver we counted the number of the keys and the number of the values of the *driver_car* dictionary, in this way we could know if any "Car" field and its values was missing. Of course we expected that the number of the "Car" fields and its values were the same of the number of drivers that we previously found. These are the results we found out:

Year	Number of *Car* fields and corresponding values
2013	23
2014	24
2021	21
2022	22

Table 3: Number of *Car* fields and corresponding values per year

As we can notice comparing Table 1 and Table 3 the results are exactly the same. This suggests us that there is no missing "Car" field and that they are all complete with their values.

4. "Qualifying", "Laps" and "Circuit" fields. To check the existence of these fields, in every iteration over each driver we added an iteration over each GP. For each driver and GP we counted how many times the "Qualifying", "Laps" and "Circuit" fields were found and how many were not. For each driver we showed the results, of course we expected that, if all had worked in the right way, the number of these three fields for each driver corresponded to the number of GPs we have found before.
The results we obtained showed us that for the "Circuit" field all worked correctly and for each driver and GP this field exists, while for few drivers we had some exceptions for the "Qualifying" and "Laps" field. We show them in Table 4, 5, 6 and 7.

Driver	Missing "Qualifying" fields	Missing "Laps" fields
Jules Bianchi		1
Paul di Resta		1
Nico Hülkenberg		1
Kimi Räikkönen	2	3
Adrian Sutil		1
Giedo van der Garde		1
Heikki Kovalainen	17	17

Table 4: Number of missing "Qualifying" and/or "Laps" field in 2013 Season

Considering Table 4, in order to understand why there are missing fields we checked on internet the possible explanation for each driver.
We found out the following things:

- Jules Bianchi and Giedo var der Garde: during the Japanese GP race there was a collision between the two cars[9].
- Paul di Resta: he had a collision during the Italian GP race[10].
- Nico Hülkenberg: during the Australian GP race he had an alimentation problem[11].
- Kimi Räikkönen and Heikki Kovalainen: during the Abu Dhabi GP race Kimi Räikkönen had a collision[12]. For the two last GPs, United States and Brazilian, the Lotus F1 Team substituted Kimi Raikkonen with Heikki Kovalainen, that's why Raikkonen misses 17 qualifying GP sessions (United States and Brazilian) while Kovalainen misses 17 qualifying GP sessions (all but United States and Brazilian), the same situation happens for the race session of these two GPs[13].
- Adrian Sutil: during the United States GP he had an incident.

So, we can explain the absence of the "Qualifying" and "Laps" fields for these drivers and we can consider our data as correct without errors.

Driver	Missing "Qualifying" fields	Missing "Laps" fields
Kamui Kobayashi	3	5
Pastor Maldonado		1
Felipe Massa		3
Sergio Pérez		2
Kimi Räikkönen		1
Daniel Ricciardo	1	
Will Stevens	18	18
Adrian Sutil		1
Sebastian Vettel	1	
Jules Bianchi	4	5
Marcus Ericsson	3	3
Max Chilton	3	4
André Lotterer	18	18

Table 5: Number of missing "Qualifying" and/or "Laps" field in 2014 Season

Considering now Table 5 we did the same researches as before for 2013 on the internet in order to check the reason of the missing fields. That's what we found:

- Kamui Kobayashi: during the Australian GP race he had a collision[14]. He did not take part to the Singapore GP race because of a technical problem. He did not take part to Belgian GP and skipped the United States and Brazilian GPs qualifying and race sessions[15][16][17].
- Pastor Maldonado: he did not take part to the Monaco GP race because of a technical problem[18].
- Felipe Massa: during the Australian GP and German GP races he had a collision[14][19]. During the British GP race he had an accident at the beginning of the race[20].

- Sergio Pérez: he did not take part to the Malaysian GP race because of a technical problem**??**. He also had a collision during the Monaco GP race[18].
- Kimi Räikkönen: during the British GP race he had an accident at the beginning of the race[20].
- Daniel Ricciardo: during the Abu Dhabi GP qualyfing session his car was found irregular[21].
- Will Stevens: he only partecipated to the qualifying and race sessions of the last GP of Abu Dhabi substituting Marcus Ericsson[21].
- Adrian Sutil: he had a collision during the United States GP race[15].
- Sebastian Vettel: during the Abu Dhabi GP qualyfing session his car was found irregular[21].
- Jules Bianchi: he had a collision in the Canadian GP race[22]. He had a serious accident during the Japanese GP race, so he did not take part to the last four GPs: Russian, United States, Brazilian and Abu Dhabi[23].
- Marcus Ericsson: he skipped the last three GPs (United States, Brazilian and Abu Dhabi) because of financial problems of the Caterham Team[24].
- Max Chilton: he had a collision in the Canadian GP race[22]. He ended his season with the Russian GP race, not taking part to the last three GPs: United States, Brazilian and Abu Dhabi[23].
- André Lotterer: he only participated to the Belgian GP qualyfing and race sessions substituting Kamui Kobayashi[17].

So even in this case, the missing "Qualifying" and "Laps" fields are correct.

Driver	Missing "Qualifying" fields	Missing "Laps" fields
Valtteri Bottas		1
Nicholas Latifi		1
Charles Leclerc		2
Nikita Mazepin		2
Esteban Ocon		1
Sergio Pérez		1
Kimi Räikkönen	2	2
Mick Schumacher	1	1
Lance Stroll		1
Yuki Tsunoda		2
Max Verstappen		1
Robert Kubica	20	20

Table 6: Number of missing "Qualifying" and/or "Laps" field in 2021 Season

We did the same analysis based on the Table 6 through the internet for the 2021 Season. These are the results:

- Valtteri Bottas: during the Hungarian GP race he had a collision[25].
- Nicholas Latifi: he had a collision during the Emilia Romagna GP race[26]
- Charles Leclerc: during the Hungarian GP race he had a collision[25]. He had a technical problem during the Monaco GP race[27]
- Nikita Mazepin: he had an accident during the Bahrain GP race[28]. He did not take part to the Abu Dhabi GP race because of resulting positive to the SARS-CoV-2[29].
- Esteban Ocon: he had a collision during the Austrian GP race[30].
- Sergio Pérez: during the Hungarian GP race he had a collision[25].
- Kimi Räikkönen: he did not take part to the qualifying and race sessions of the Dutch and Italian GP because of resulting positive to the SARS-CoV-2[31][32].

- Mick Schumacher: he did not take part to the qualifying session of Monaco GP because of a problem during free practice session, but he participated to the Monaco GP race starting from the last position in the grid[27]. He had a collision during the Mexico City GP race[33].
- Lance Stroll: during the Hungarian GP race he had a collision[25].
- Yuki Tsunoda: he had a brake issue during the Italian GP race[32]. He had a collision during the Mexico City GP race[33].
- Max Verstappen: he had a collision during the British GP race[34].
- Robert Kubica: he only took part to the Dutch and Italian GP qualifying and race session substituting Kimi Räikkönen after he resulted positive to the SARS-CoV-2[35].

This analysis is a confirmation of the correct collection of our qualifying and race data for 2021 season.

Driver	Missing "Qualifying" fields	Missing "Laps" fields
Alexander Albon	1	3
Lewis Hamilton		1
Nico Hülkenberg	20	20
Kevin Magnussen		1
Daniel Ricciardo		1
George Russell		1
Carlos Sainz		2
Mick Schumacher		1
Yuki Tsunoda		1
Guanyu Zhou		1
Sebastian Vettel	2	2
Nyck de Vries	21	21

Table 7: Number of missing "Qualifying" and/or "Laps" field in 2022 Season

Table 7 shows us the last results for season 2022, even in this case we went to check on the internet the reasons for the missing fields. We found out:

- Alexander Albon: he had a collision during the British GP race[36]. He did not take part to the Italian GP because he had developed appendicitis[37]. He had a technical problem during the Japanese GP race[38]
- Lewis Hamilton: he had a collision during the Belgian GP race[39].
- Nico Hülkenberg: he only participated to the Bahrain and Saudi Arabian GPs substituting Sebastian Vettel resulting positive to the SARS-CoV-2[40].
- Kevin Magnussen: he had a collision during the Brazilian GP race[41].
- Daniel Ricciardo: he had a collision during the Brazilian GP race[41].
- George Russell: he had a collision during the British GP race[36].
- Carlos Sainz: he had an accident during the Japanese GP race[38]. He had a collision during the Emilia Romagna GP race[42].
- Mick Schumacher: he had a problem during the qualifying session of Saudi Arabian GP and for this reason he did not take part to the Saudi Arabian GP race[43].
- Yuki Tsunoda: he had an engine problem during the Saudi Arabian GP race[43].
- Guanyu Zhou: he had a collision during the British GP race[36].
- Sebastian Vettel: he did not take part to the Bahrain and Arabian GPs because of resulting positive to the SARS-CoV-2[40].
- Nyck de Vries: he only participated to the Italian GP substituting Alexander Albon[37].

Even in this case, data are completed.

5. "Q1 Time", "Q2 Time and "Q3 Time" fields. Then, we went in each of the previous fields in order to see if what they contain is correct. We started with the "Qualifying" field and we had to check if it contains the three sub-keys "Q1 Time", "Q2 Time" and "Q3 Time" and their respective values. In this case we had to consider the following thing: in each GP all the drivers should have the data for the "Q1 Time", then five or six drivers are deleted and do not participate to the second qualifying, so we should have a lower number of data for "Q2 Time" and just 10 records for the "Q1 Time" per GP. This is exactly what we had checked on our JSON file. For each driver and each GP if the "Qualifying" field exists we checked if "Q1 Time", "Q2 Time" and "Q3 Time" fields exist (all of them), otherwise the code returns an error. If all the three fields are found, a new key (the name of the GP) is added to a new dictionary called *qtime*, for each key there are four sub-keys "Q1", "Q2", "Q3" and "NQT". The values corresponding to these four fields are the number of the times that the "Q1 Time", "Q2 Time" and "Q3 Time" contain a value for that GP (in other words "Q1" contains the number of drivers that have a time for the first qualifying of that GP, "Q2" contains the number of drivers that have a time for the second qualifying of that GP and "Q3" contains the number of drivers that have a time for the third qualifying of that GP). The "NQT" field contains the number of times that for that GP none of the three fields "Q1 Time", "Q2 Time" and "Q3 Time" contain a value (in other words it contains the number of drivers that have no qualifying time for that GP).
   Through the *qtime* dictionary we found out that some values that should be present were missing, we show them through Table 8, 9, 10 and 11.

GP	Q1	Q2	Q3	NQT
Malaysian GP	22	15	10	0
Chinese GP	22	16	8	0
Bahrain GP	22	16	9	0
Monaco GP	20	16	10	2
German GP	22	16	8	0
Hungarian GP	22	16	9	0
Singapore GP	22	16	9	0

Table 8: GPs in which at least one time is missing in the Qualifying session in season 2013.

For 2013 season we expected 22 values for Q1 Time, 16 values for Q2 Time and 10 values for Q3 Time. In fact, even if the number of drivers is 23, we know from the previuos analysis that Heikki Kovalainen only took part to the season substituting Kimi Räikkönen during two GPs, for this reason we expected to have only 22 times for Q1 session. Then the last six positions are assigned and only 16 drivers participate to the Q2 session and finally in Q3 are assigned the first ten positions, that's why we expected to find 10 values. Of course we expected that the NQT columns did not contain any value but 0, but we can see that there is an exception.
As we can see from Table 8 in these GPs we have some different values from the ones we expected, so we checked on the internet if our data were correct,that's what we found:

- Malaysian GP: during Q2 Pastor Maldonado did not set valid time[44].
- Chinese GP: during Q3 Sebastian Vettel and Nicolas Hülkenberg did not set valid times[45].
- Bahrain GP: during Q3 Jenson Button did not set any times[46].
- Monaco GP: Jules Bianchi and Felipe Massa had some problems with their car and did not set any time during Q1[47].
- German GP: during Q3 Jenson Button and Nico Hülkenberg elected not to set a time, saving their tyres for the race[48].
- Hungarian GP: during Q3 Mark Webber did not complete any timed laps due to the KERS issue encountered in Q2[49].
- Singapore GP: during Q3 Esteban Gutiérrez opted to save his tyres for the race[50].

So the data we collected were correct and there are no error in the number of qualifying times for season 2013.

GP	Q1	Q2	Q3	NQT
Australian GP	21	16	10	1
Chinese GP	21	16	10	1
Spanish GP	21	14	9	1
Monaco GP	22	15	10	0
Canadian GP	21	16	10	1
Austrian GP	22	16	8	0
British GP	22	15	10	0
German GP	21	15	10	1
Hungarian GP	20	16	9	2
Russian GP	21	16	10	0
United States GP	18	14	10	0
Brazilian GP	18	13	10	0
Abu Dhabi GP	18	13	8	0

Table 9: GPs in which at least one time is missing in the Qualifying session in season 2014.

For season 2014 we expected the same values as in 2013 season: 22 values for Q1, 16 values for Q2 and 10 values for Q3. In fact, even if the number of drivers for this season is 24, through previous analysis we had known that Will Stevens only participated to one GP substituting Marcus Ericsson and similarly André Lotterer only took part to one GP substituting Kamui Kobayashi, so the total number of drivers who took part to each GP is 22. Then in Q2 six drivers did not participated because they yet have the starting position assigned, and in Q3 just the first ten positions are assigned. As before in "NQT" we did not expect to find any values different from 0. From Table 9 we can see where the expected values are not found. As we did for season 2014 we searched on the internet the reasons why this happened:

- Australian GP: during Q1 Pastor Maldonado did not set valid times[51].
- Chinese GP: Pastor Maldonado did not not participate to qualifying sessions[52].
- Spanish GP: during Q1 Pastor Maldonado did not set valid times, during Q2 Jean-Éric Vergne opted to save his tyres for the race and Kevin Magnussen couldn't complete any valid laps due to the issue with his propulsion system, during Q3 Sebastian Vettel, due to a gearbox issue, did not achieve any timed lap[53].
- Monaco GP: during Q1 Felipe Massa was rear-ended by Marcus Ericsson, preventing him from participating in Q2[54].
- Canadian GP: Esteban Gutiérrez couldn't take part to qualifying sessions due to an accident during free practice[55].
- Austrian GP: Nico Hülkenberg and Lewis Hamilton during Q3 did not set valid times[56].
- British GP: during Q2 Adrian Sutil lost control of his car's rear and got beached in a gravel trap[57]
- German GP: Marcus Ericsson did not take part to qualifying sessions, during Q1 Lewis Hamilton damaged his car and for this reason he did not take part to Q2 session[58].
- Hungarian GP: during Q1 Pastor Maldonado and Lewis Hamilton had problems with their cars, during Q3 Kevin Magnussen crashes into the barriers[59].
- Russian GP: Jules Bianchi did not take part to the GP because he had a serious accident during the previous GP[60]
- United States GP: Marussia and Caterham Team drivers (Jules Bianchi, Max Chilton and Kamui Kobayashi, Marcus Ericsson) did not take part to the GP, this time with Q1 were decided the last four position, then with Q2 the position from 14^{th} to 11^{th} and with Q3 as always the first ten starting positions[61].

- Brazilian GP: Marussia and Caterham Team drivers (Jules Bianchi, Max Chilton and Kamui Kobayashi, Marcus Ericsson) did not take part to the GP, this time with Q1 were decided the last four position, then with Q2 the position from 14^{th} to 11^{th} and with Q3 as always the first ten starting positions. During Q2 Daniil Kvjat opted to save his tyres for the race[62].
- Abu Dhabi GP: the Marussia Team drivers (Jules Bianchi and Max Chilton) did not take part to the GP qualifying session, Daniel Ricciardo and Sebastian Vettel were found with irregularities with their cars, they were disqualified and their times were considered not valid[63].

So even in this case the unexpected values were correct.

GP	Q1	Q2	Q3	NQT
Emilia Romagna GP	19	15	9	0
Monaco GP	19	15	10	0
Azerbaijan GP	18	15	10	2
French GP	18	14	10	2
Hungarian GP	19	14	10	1
Belgian GP	20	15	9	0
Russian GP	19	13	10	1
Turkish GP	20	14	10	0
United States GP	20	14	10	0

Table 10: GPs in which at least one time is missing in the Qualifying session in season 2021.

Considering 2021 season we expected 20 values for Q1, 15 values for Q2 and 10 values for Q3. In this season as we have seen before, even if the number of drivers who participated is 21, Robert Kubica only took part to two GPs substituting Kimi Räikkönen, that's why the correct number of drivers who partecipated to each GP should be 20. After Q1 five drivers did not pass to the following session Q2 which, consequently, only counts 15 drivers and as always Q3 session is reserved for the 10 first drivers that try to win the pole position. As in the previous case, we didn't expect any value different from 0 in the "NQT" section. What we can see from Table 10 is that there are some GPs in which our expectation are not satisfied. On the internet we found the reasons that can explain these unexpected values:

- Emilia Romagna GP: Lance Stroll did not set a lap during Q3[64].
- Monaco GP: Mick Schumacher did not take part to the qualifying session after an accident during the free practice session[65].
- Azerbaijan GP: Lance Stroll and Antonio Giovinazzi had and accident during Q1[66].
- French GP: during Q1 Yuki Tsunoda had an accident and Lance Stroll did not set valid times. Mick Schumacher did not take part to Q2 because of his damaged car[67].
- Hungarian GP: Mick Schumacher did not take part to the qualifying session. Carlos Sainz during Q2 had an accident[68].
- Belgian GP: during Q3 Lando Norris had an accident[69].
- Russian GP: during Q1 Max Verstappen did not set valid time, Nicholas Latifi and Charles Leclerc did not take part to Q2[70].
- Turkish GP: Carlos Sainz did not take part to Q2 session[71].
- United States GP: George Russell did not set any valid times during Q2 session[72].

Our researches let explain us the reason why there were missing values and let confirm us that data we collected were correct. For season 2022 we expected the same values for Q1, Q2 and Q3 (respectively: 20, 15 and 10) as in 2021 season. In fact, as we saw before, Nico Hülkenberg participated to only two GPs substituting Sebastian Vettel and Nyck de Vries took part to only one GP substituting Alexander Albon. So, after Q1 in which should participate 20 drivers, only 15 drivers took part to the second session and the first 10 positions are decided with Q3. As always, we expected no values in "NQT" column. In Table

GP	Q1	Q2	Q3	NQT
Saudi Arabian GP	19	15	10	1
Australian GP	19	15	9	1
Emilia Romagna GP	19	15	9	1
Miami GP	19	15	10	1
Canadian GP	20	13	10	0
Austrian GP	20	15	9	0
French GP	20	15	8	0
Dutch GP	20	15	9	0
Italian GP	20	14	9	0
Brazilian GP	20	15	9	0

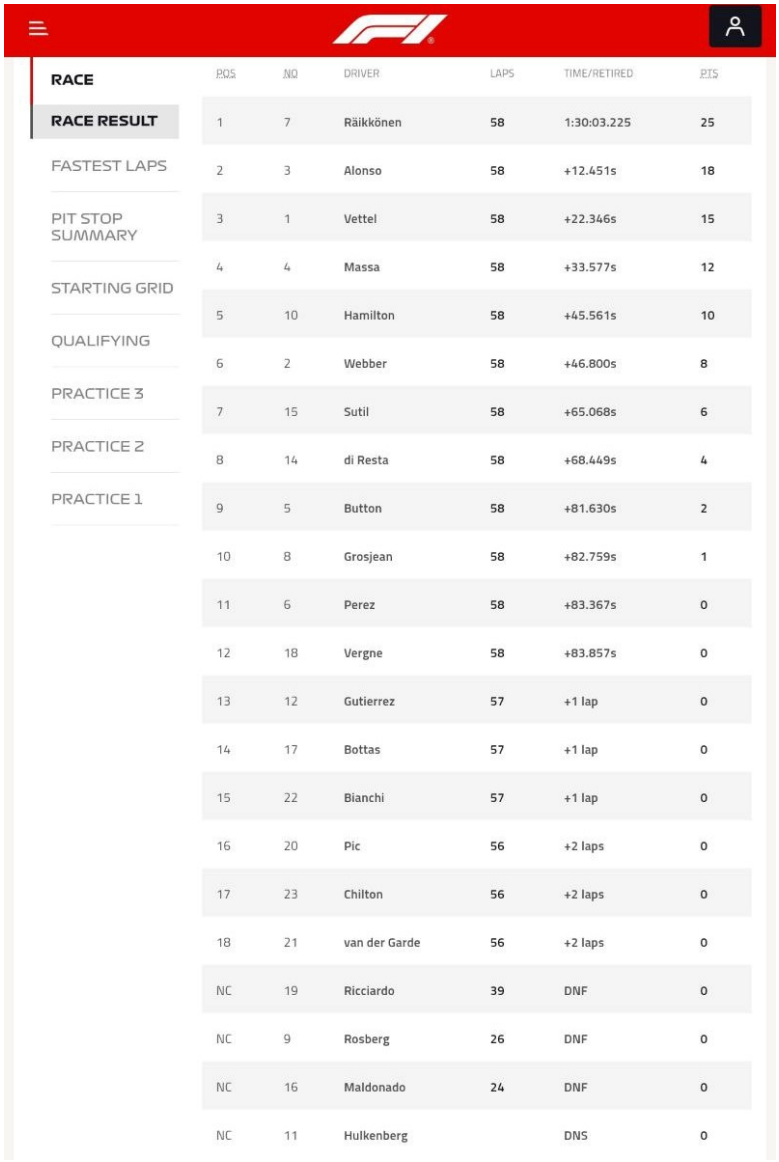
Table 11: GPs in which at least one time is missing in the Qualifying session in season 2022.

11 we can see the GPs in which there are different values from the expected ones. On the internet we found the reasons of these unexpected values:

- Saudi Arabian GP: Yuki Tsunoda had a problem with his car during Q1[73].
- Australian GP: Lance Stroll had an accident during Q1, Fernando Alonso during Q3 had an accident[74].
- Emilia Romagna GP: Alexander Albon had an accident during Q1. Carlos Sainz had an accident during Q3[75].
- Miami GP: Esteban Ocon had an accident during the free practice session and for this reason did not take part to qualifying sessions[76].
- Canadian GP: during Q2 Charles Leclerc and Lando Norris did not set any valid time[77].
- Austrian GP: during Q3 Sergio Pérez committed irregularities and for this reason had no valid time[78].
- French GP: Carlos Sainz and Kevin Magnussen decided not to set times in this qualifying phase[79].
- Dutch GP: Lance Stroll had a technical problem during Q3[80].
- Italian GP: Yuki Tsunoda did not take part to Q2 session. During Q3 Fernando Alonso committed irregularities and his time was cancelled[81].
- Brazilian GP: during Q3 Charles Leclerc had a problem with his tyres[82].

We considered these data complete.

6. Laps number field. For each season, to check if everything worked correctly with this field we counted the number of existing sections within the "Laps" field for each driver and GP (the number of sections should correspond to the number of laps done by the driver) and then we count the number of values within the keys in order to check if there were any empty section. In order to check is this count obtained for each driver and GP was correct we take into account the official Formula 1 website (formula1.com) and in the section *Results* and then *Archive 1950-2022*, selecting the year of interest and the races results we found the *race results* tables for each GP in which there were, among the others the *Driver* and the *Laps* columns, in which for each driver were reported the number of laps he did in that GP. We can see an example of one of these tables in Figure 4.
   For each year, we have done web scraping on the internet pages of Formula 1 website for all the GPs and we extracted the data contained in *Driver* and *Laps* columns and we saved them into a JSON file. Then we compared these web scraped data to our original data in order to see if the number of laps for driver for each GP was correct.
   The results obtained showed us that the majority of the data correspond to the Formula 1 website data, a part from some exceptions. We found out that sometimes the number of laps resulted by counting our data was different from the web scraped data by 1 or 2 units. This difference should not cause alarm because



POS	NO	DRIVER	LAPS	TIME/RETIRED	PTS
1	7	Räikkönen	58	1:30:03.225	25
2	3	Alonso	58	+12.451s	18
3	1	Vettel	58	+22.346s	15
4	4	Massa	58	+33.577s	12
5	10	Hamilton	58	+45.561s	10
6	2	Webber	58	+46.800s	8
7	15	Sutil	58	+65.068s	6
8	14	di Resta	58	+68.449s	4
9	5	Button	58	+81.630s	2
10	8	Grosjean	58	+82.759s	1
11	6	Perez	58	+83.367s	0
12	18	Vergne	58	+83.857s	0
13	12	Gutierrez	57	+1 lap	0
14	17	Bottas	57	+1 lap	0
15	22	Bianchi	57	+1 lap	0
16	20	Pic	56	+2 laps	0
17	23	Chilton	56	+2 laps	0
18	21	van der Garde	56	+2 laps	0
NC	19	Ricciardo	39	DNF	0
NC	9	Rosberg	26	DNF	0
NC	16	Maldonado	24	DNF	0
NC	11	Hulkenberg		DNS	0

Figure 4: Example of website page taken from formula1.com: `https://www.formula1.com/en/results.html/2013/races/879/australia/race-result.html`

the explication is very simple: sometimes the time laps after a lap in which the driver has completed a pit stop is deleted because there could be some irregularities[83], such as:

- Pit Lane speeding: During the pit stop, drivers must adhere to a predetermined speed limit in the pit lane to ensure the safety of mechanics and other participants. If a driver exceeds the speed limit, the time after the pit stop can be deleted.
- Unsafe release: An unsafe release from the pit lane can lead to deletion of the time following the pit stop. If the driver is released in a dangerous manner that jeopardizes the safety of other drivers or team members, the time can be canceled, and additional penalties may be imposed.
- Pit Stop regulation violations: There are specific rules governing pit stop procedures, such as how many team members can work on the car at a given time or what equipment can be used. Violations of these rules can result in the deletion of the time after the pit stop.
- Rejoining the track unsafely: After a pit stop, if a driver rejoins the track in a dangerous manner and creates a hazardous situation for other participants, the time following the pit stop can be canceled.
- Illegal technical modifications: If illegal modifications are made to the car during the pit stop, such as the use of unauthorized equipment or excessive fueling, the time following the pit stop can be eliminated.

- Ignoring flags or race official instructions: If a driver ignores flags or instructions from race officials during the pit stop, they may be penalized with the deletion of the time following the pit stop.
- Penalties imposed after the Pit Stop: If a penalty is imposed on the driver during the pit stop, such as for improper on-track behavior, and this penalty must be served after the pit stop, the time following the pit stop can be eliminated.

Our source of data took into account also that eliminated lap times, this is the reason why sometimes we had more data than the ones found on the Formula 1 website pages. For our analysis, that is purely related to the car performances changes over years, having these lap times is not a problem or an error, we will consider also them for the last part of the analysis. So with the even with the creation of "Laps" field no errors have been made.

7. "Name", "Length" and "Turns" fields. To check if everything worked correctly also with the sub-keys of "Circuit" field, for each driver and GP we count the keys and the values contained in the "Circuit" field. We expected that the sub-keys were "Name", "Length" and "Turns", so we first check if all of these three voice where found, if it happened, the number of sub-keys and values were counted.
We found out that for each driver and GP the "Circuit" field always contained 3 sub-keys ("Name", "Length" and "Turns") and each of them contained a value. Exactly what we expected if all worked correctly with the code.

So, the first completeness file check was ended, the we had to proced with the other JSON file we created.

3.1.2 Maximum speed JSON

Then we checked the JSON file of maximum speed. As we had seen before, we had a JSON file for each GP for each year and all these JSON files were formed by some dictionaries, one for each driver that participated to that GP. In each small dictionary we expect to find 4 keys: "driver", "km/h", "gp", "year" and we expect to find a value for each of these keys. So, for each year we open each JSON file corresponding to each GP of that specific year and we counted the number of elements (number of small dictionaries) present in the JSON files, then in each dictionary if all the four keys "driver", "km/h", "gp", "year" were found, we counted the number of keys and the number of values.
For each file we expected to find the number of elements equal to the number of drivers that participated to that GP qualifying sessions. If some drivers did not take part to that GP qualifying sessions of course couldn't have the maximum speed data recorded during qualifying session, while, even if a driver did not have a recored time during qualifying, it is possible that the has a value of maximum speed, because it could have had a problem during the session. So, we have to take into account these aspects during the analysis.

- 2013: all the JSON files of 2013 season but one counts 22 elements (dictionaries) with 4 keys and 4 values. The number 22 corresponds to the number of drivers who partecipated to the qualifying sessions, as we found out before. The only one exception is the Monaco GP JSON file, that only counts 20 drivers, but this data is in line with the number of drivers who took part to the qualifying session (Table 8).

- 2014: in the 2014 season we expected 22 elements for each GP as 22 drivers participated to each GP. In some JSON file we did not find this results, we list them and we considered Table 9 to make analysis:

  - Abu Dhabi GP: it only counts 20 dictionaries but we have seen that the two Marussia Team drivers did not take part to the GP. qualifying session.
  - Brazilian GP: two Marussia and two Caterham Team drivers did not take part to qualifying session of Brazilian GP, that's why the Json file only contains 18 elements.
  - Canadian GP: the file only counts 21 drivers and that's because Esteban Gutiérrez did not participated to the qualyfing sessions.
  - Chinese GP: only 21 dictionaries are counted in the JSON file because Pastor Maldonado did not take part to the qualyfing session.

  - German GP: Marcus Ericsson did not take part to qualifying sessions and the JSON file did not contain his dictionary, but only the 21 remaining ones.
  - Hungarian GP: the JSON file only contains 21 elements because Pastor Maldonado had to stop his first qualifying session too soon to record a max speed value.
  - Russian GP: Jules Bianchi did not take part to the qualifying session and in fact the JSON file contains only 21 elements.
  - United States GP: as in Brazilian GP, two Marussia and two Caterham Team drivers did not take part to qualifying session of United States GP, that's why the Json file only contains 18 elements.

- 2021: for this season we expected to have 20 dictionaries for each JSON file (GP) and our expectation was not satisfied only with two GPs: Hungarian GP and Monaco GP that count 19 elements. This result should not surprise us because, as we have seen before through Table 10, in both GP, Mick Schumacher did not participate to qualifying session. The number of keys and values for each dictionary is 4.

- 2022: we counted the number of elements for each JSON file (so, for each GP) and we expected to find 20 elements for each GP as 20 is the number of drivers that participated to each GP, as we have seen before. We obtained that result for each GP but one: Miami GP that counts 19 elements. Referring to Table 11 and its analysis we have seen that one driver (Esteban Ocon) did not take part to the qualifying sessions of Miami GP, so the data 19 is correct. All of the dictionaries contain 4 keys and each key contains a value, so there are no missing data.

We could conclude that all the JSON files where completed.

3.1.3 Qualifying and Race Session Weather JSON

The analysis of completeness on these JSON files (both qualifying and race sessions) was quite rapid. We should have had two files per year, one for qualifying and the other for race sessions and we knew that these files should have been structured with the names of the GPs of the year as the keys and the word "sun" or "rain" as the value. For each one of the 8 files we checked the number of keys (GPs) and if each key contained a value corresponding to "rain" or "sun". We found out that each file contained the correct number of GP and the value for each key was the word "rain" or "sun". All these files were complete.

3.2 Accuracy

3.2.1 Syntactic accuracy

For certain database fields, we obtained a reference vocabulary from a reliable source to verify syntactic accuracy. Below is a table displaying the fields under analysis, the corresponding sources for comparison, and the percentage of inaccuracies identified for each field. Calculating the percentage of inaccuracy between two strings involves comparing their characters and determining how many characters are different. This method provides a way to quantify how dissimilar two strings are in terms of their content.

Analyzed Fields	Comparative Sources	Reference file name	Percentage of inaccurate records
Driver name	`https://pitwall.app/analysis/race-report?utf8=âœŞ&season=7&race=122&driver=22&button=;` *	Season2013.json-Season2014.json-Season2021.json-Season2022.json	0%
Grand Prix name	`https://pitwall.app/analysis/race-report?utf8=âœŞ&season=7&race=122&driver=22&button=;` *	Season2013.json-Season2014.json-Season2021.json-Season2022.json	10%
Qualifying fields	`https://pitwall.app/analysis/race-report?utf8=âœŞ&season=7&race=122&driver=22&button=;` *	Season2013.json-Season2014.json-Season2021.json-Season2022.json	0%
Laps fields	`https://pitwall.app/analysis/race-report?utf8=âœŞ&season=7&race=122&driver=22&button=;` *	Season2013.json-Season2014.json-Season2021.json-Season2022.json	0%
Circuit fields	`https://en.wikipedia.org/wiki/2013_Formula_One_World_Championship`; *	Season2013.json-Season2014.json-Season2021.json-Season2022.json	0%
Maximum speed	*Grand Prix name*; 2013.pdf-*Grand Prix name*-2014.pdf; *Grand Prix name*-2021.pdf; *Grand Prix name*-2022.pdf	*Grand Prix name* - 2013.json; *Grand Prix name* - 2014.json; *Grand Prix name* - 2021.json; *Grand Prix name* - 2022.json	0%
Car name	`https://pitwall.app/seasons/2013-formula-1-\world-championship`; *	Season2013.json-Season2014.json-Season2021.json-Season2022.json	0% *

Table 12: The asterisks next to the links and below the "Comparative sources" field indicate that the URL address is valid for other years as well, replacing the year 2013 with the desired year for the search. The asterisk near the inaccuracy percentage of car names is placed to indicate that the "Car" field is present on the website in uppercase letters.

Considering table 12 the following considerations can be done:

- The accuracy percentage is evaluated solely for fields present within the JSON files; it does not take into account any missing fields.
- The **Grand Prix name** field exhibits a certain percentage of inaccuracy. Below, an example illustrates the procedure through which this percentage was derived. Please note that the example is generalizable to all Grand Prix names present in the JSON files, and for this reason, the percentage is consistent across all.

To begin, we take the string that we consider as the reference or baseline. For example in the case of "Australian Grand Prix", which has 20 characters we compare it with the other string, "Australian GP", which has 12 characters. By examining each character in both strings, we can identify the differences between them. In this example, the characters 'G' and 'P' differ in the second string compared to the reference string. The formula for calculating the percentage of inaccuracy divides the number of different characters by the total number of characters in the reference string. This result is then multiplied by 100 to express it as a percentage. In the case of "Australian GP" versus "Australian Grand Prix," there are 2 different characters out of 20, which leads to a percentage of inaccuracy of 10%.

Regarding the fields related to the weather conditions in the qualifying and race sections, it's not feasible to conduct an analysis of syntactic accuracy. This is due to the fact that the weather conditions "rain" or "sun" result from textual processing and aren't directly attributable to a specific "Weather Race" or "Weather Qualifying" entry on the website.

3.2.2 Semantic Accuracy

Another important feature in a database is that data reflects appropriately the phenomena it represents. In our case, we asked ourselves the following questions:

- Does the performance statistics (race times and speed) accurately mirror the on-track reality? This relies on the sources from which we compiled the data and the methodology used by the data collectors (which could include estimations to some extent).
- Can the data regarding Formula 1 circuits, drivers and their race results be considered reliable? Data extracted from the Formula 1 website Pitwall.app holds credibility because the data come from the official website of Formula 1 (i.e https://www.fia.com); Wikipedia also generally serves as a dependable reference.
- What about Weather Conditions Data? The reliability of the Weather field, both for qualifying and races, depends on the type of textual analysis that has been conducted. As previously mentioned in the Data Acquisition section, the string "rain" is assigned to the "Weather" field only if a set of specified keywords are found within the designated paragraph or table. Table 13 specifies the section under analysis, the Wikipedia site from which information is extracted, the specific paragraph or table on the Wikipedia page, and the keywords used for assigning the "rain" field. If none of these words are found, the "sun" field is assigned to the "Weather" section. In the event that the english Wikipedia page does not contain any of the keywords present in the "keywords" field, we have chosen to conduct an additional check on the Italian version of Wikipedia. This approach aims to mitigate potential errors arising from informational gaps that might exist in the English version of the platform. Table 14 illustrates the verification conducted in cases where the "sun" field was assigned in the previous step. Regarding the Race section, no such additional check was conducted. This is due to the fact that the meteorological information within a table is placed under a specific "Weather" field, leaving no room for ambiguity.

Weather	Website Source	Source format	keywords checked
Qualifying	https://en.wikipedia.org/wiki/2013_Grand_Prix_name_Grand_Prix;	paragraph between "Qualifying" and "Race"	"wet", "rain", "rainfalls"
	https://en.wikipedia.org/wiki/2014_Grand_Prix_name_Grand_Prix	paragraph between "Practice_and_qualifying" and "Race" or "Qualifying" and "Race"	" wet", " rain", " rainfalls", "Rain", " Rain", " Rainfalls", " wet"
	https://en.wikipedia.org/wiki/2021_Grand_Prix_name_Grand_Prix	paragraph between "Qualifying" and "Sprint_qualifying or "Qualifying" and "Race"	" wet", " rain", " rainfalls", "Rain", " Rain", "Rainfalls", " wet"
	https://en.wikipedia.org/wiki/2022_Grand_Prix_name_Grand_Prix	paragraph between "Qualifying" and "Sprint_qualifying or "Qualifying" and "Race"	" wet", " rain", " rainfalls", "Rain", " Rain", " Rainfalls", " wet"
Race	https://en.wikipedia.org/wiki/2013_Grand_Prix_name_Grand_Prix; *	table about Grand Prix of the specific year *	"wet", "rain", "rainfalls" *

Table 13: The asterisks serve as indicators: the first signifies that the website can be adjusted for researching a different year than 2013 by merely modifying the year within the link. The second highlights that the Source format and keywords remain constant across other years.

.

Weather	Website Source	Source format	keywords checked
Qualifying	https://it.wikipedia.org/wiki/nome_gran_premio_italiano_2013;	paragraph between "Qualifiche" and "Gara"	"bagnata", "bagnato", "pioggia", "acqua", "gomme da bagnato", "pista scivolosa", "spray dácqua", "visibilità ridotta", "aquaplaning", "pneumatici da pioggia", "safety car" *
	https://it.wikipedia.org/wiki/nome_gran_premio_italiano_2014	paragraph between "Qualifiche" and "Gara"	"bagnata", "bagnato", "pioggia", "acqua", "gomme da bagnato", "pista scivolosa", "spray dácqua", "visibilità ridotta", "aquaplaning", "pneumatici da pioggia", "safety car" *
	https://it.wikipedia.org/wiki/nome_gran_premio_italiano_2021	paragraph between "Qualifiche" and "Qualifica_Sprint" or "Qualifiche" and "Gara"	"bagnata", "bagnato", "pioggia", "acqua", "gomme da bagnato", "pista scivolosa", "spray dácqua", "visibilità ridotta", "aquaplaning", "pneumatici da pioggia", "safety car" *
	https://it.wikipedia.org/wiki/nome_gran_premio_italiano_2022	paragraph between "Qualifiche" and "Qualifica_Sprint" or "Qualifiche" and "Gara"	"bagnata", "bagnato", "pioggia", "acqua", "gomme da bagnato", "pista scivolosa", "spray dácqua", "visibilità ridotta", "aquaplaning", "pneumatici da pioggia", "safety car" *

Table 14: The asterisks mean that Further verification was then carried out on the Italian keywords, which need to be contextualized. If any of the phrases "si svolge senza pioggia" , "possibile arrivo della pioggia" , or "non c'è minaccia della pioggia" appear in the text, it indicates that the qualifiers did not take place under adverse weather conditions. Consequently, the correct condition to attribute to the "Weather" field remains "sun".

3.3 Consistency

This parameter allows for verifying if the data comprehensively depict all facets of the reality of interest. The table 15 illustrates the adjustments that have been implemented to ensure dataset consistency.

File Dataset	Consistency Adjustments
Season2013.json - Season2014.json - Season2021.json - Season2022.json	Lap times and qualifying times are expressed in minutes and seconds. To ensure consistent and unique unit of measurement and to manage mathematical computations, we will convert these times into seconds.
"Name_GrandPrix" GP - 2013.json; "Name_GrandPrix" GP - 2014.json; "Name_GrandPrix" GP - 2021.json; "Name_GrandPrix" GP - 2022.json	In these files, driver names appeared in the format: "First Letter of Name. Full Surname (uppercase)". For this reason, the names will be transformed to conform to the "Full Name + Full Surname (lowercase)" format of the dataset where the integration will take place;
Weather_Qualifying.json (2013,2014,2021,2022) and Weather_Race.json (2013,2014,2021,2022)	In the JSON files associated with each year, a weather condition, "rain" or "sun", is assigned to each Grand Prix. In the section dedicated to Semantic Accuracy, it is clarified that the term "rain" is associated with the "Weather" category even when conditions implicitly refer to such a situation. For instance, expressions like "wet" or "rainfalls" are understood as indicative of rainy conditions. This ensures that the "rain" data remains consistent with the actual meteorological situation during the qualifying sessions (Weather_Qualifying.json) or race session (Weather_Race.json). Regarding rain, this does not necessarily imply a continuous rain condition (or similar) during the qualifying phase or race. Instead, it indicates that there was at least one occurrence of rain (or similar) during the three qualifying sessions or during laps;

Table 15

3.4 Currency and timeliness

The temporal quality is measured through two parameters: timeliness and currency. Timeliness measures how data is up to date with respect to a particular moment, while currency measures how quickly the data is updated compared to the corresponding phenomenon in the real world. Websites like Pitwall.app strive to provide real-time or near-real-time updates, ensuring that the data they present are reflective of what was happening in the actual races. We know that since Pitwall.app was created in 2020, the data concerning drivers and Grand Prix events that occurred before 2020 were added around that time. Conversely, data for the subsequent years, 2021 and 2022, were added as the actual races took place. Regarding Wikipedia pages, many of them are subject to constant updates by various users. As for data about the maximum speeds achieved by drivers during qualifying sessions, it appears that these were added in the years when the races took place and loaded into the FIA official site. It has been found that such data were included in PDF documents that haven't been updated further.

4 DATA INTEGRATION

The acquired datasets during the Data Acquisition phase have been merged to form four JSON files corresponding to each analyzed year: FinalDataset2013.json, FinalDataset2014.json, FinalDataset2021.json, and FinalDataset2022.json. These JSON files are the result of integrations between the dataset of Grand Prix seasons with data related to the drivers' maximum speeds during qualifying sessions and the weather conditions during races and qualifying sessions.
Due to the fact that the data integration code is strongly tied with the data acquisition code, in this paragraph we need to refer again to the data acquisition process that we have explained before.

4.1 Dataset about Seasons

First step:

- Dictionaries *complete_data*, *driver_car_data*, and *Circuit_Data* are initialized. *complete_data* is the main dictionary that will contain all the merged data about driver and Grand Prix. *driver_car_data* holds information about drivers and their respective cars, while *Circuit_Data* contains data about race circuits.
- For each driver and Grand Prix, a default dictionary is created inside *complete_data* to store the data for that specific GP. The code then iterates through all sections of the webpage (on Pitwall.app) except the first one (Summary) and the data for each section (Practice, Qualifying, Pitstops, Race progression, and Laps) is stored in the respective GP dictionary (*complete_data*) under the driver's name. Then, Pitstops, Race progression and Practice sections are deleted.

Second step:

- Merging Driver and Car Data:

  The code performs web scraping on a table that provides information about drivers and their cars. The extracted data (driver and car information) is saved as a dictionary with keys "Driver" and "Car", and appended to the *driver_car_data* list. After collecting all the driver car data from the table, another loop (for item in *driver_car_data*) goes through each entry. For each entry, it tries to match the driver's name with the keys in the *complete_data* dictionary (which contains the GP data collected earlier). The similarity score is calculated using the fuzz.ratio() function, which measures the similarity of two strings as a percentage. If the similarity score between the current driver name and a driver name in the *complete_data* is greater than or equal to 75 (a threshold chosen in the code), it is considered a close enough match. In that case, the code proceeds to insert the car information for that driver into the *complete_data* dictionary. The code adds the car information to the existing data for that driver under the "Car" key. If, instead, the driver's name does not have a suitable match in *complete_data*, the code creates a new entry in the dictionary using the driver's name as the key and stores the car information under the "Car" key. What we have noticed is that no new driver is added to the dictionary; therefore, the similarity condition is met by all the names of the drivers to whom the "Car" is added. An example of this pipeline is reported in Figure 5 where the merging process has been done for 2013 data.

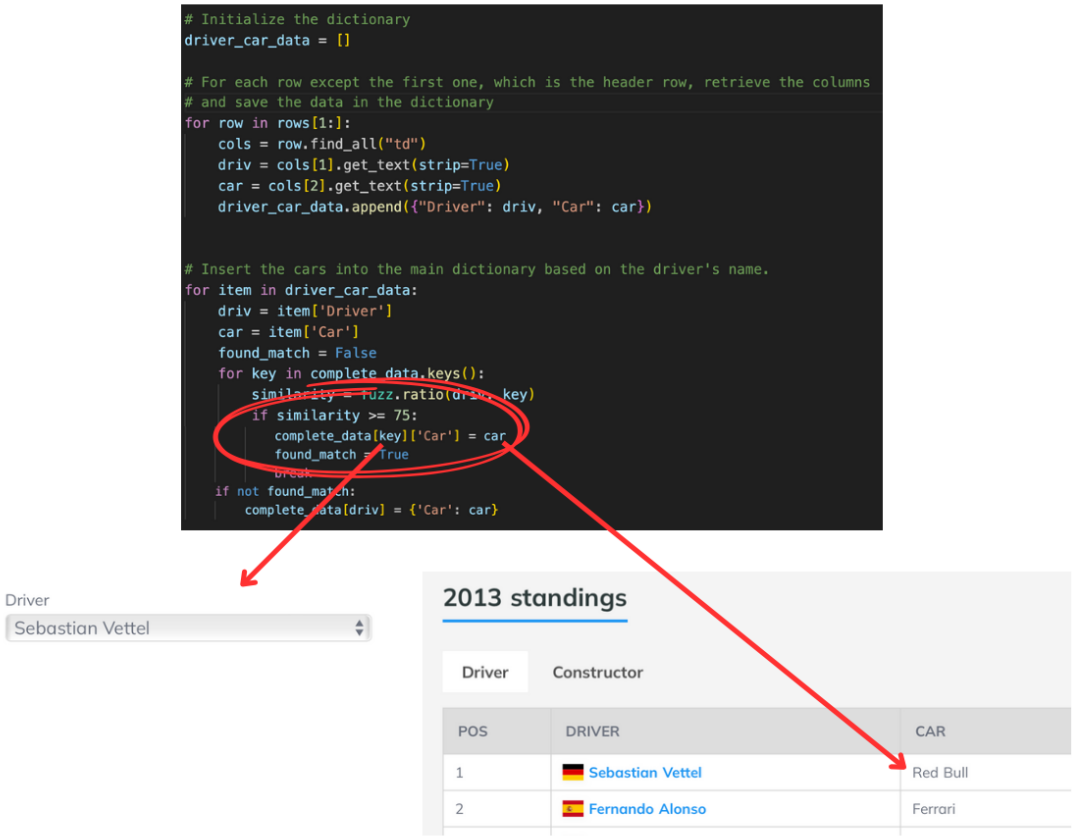


Figure 5: This image shows how the integration of data related to the drivers' cars was implemented using Python.

- Enrichment of Grand Prix with the corresponding circuit:

  For each GP name in the *Circuit_Data* dictionary (which contains data about GP circuits), the code searches for an approximate match within the keys of the *complete_data* dictionary, where Grand Prix names are used as keys to store circuit information for each driver. An intermediate step involves changing the name from "Grand Prix" to "GP" to ensure compatibility with the *complete_data* dictionary. The similarity score is calculated using fuzz.ratio() on GP names in *complete_data* dictionary and GP names in *Circuit_Data* dictionary. If this measure exceeds the specified threshold (75 in this case), it is considered a match. The code then proceeds to insert the circuit information for that GP into the corresponding driver's data within the *complete_data* dictionary. The code adds a new dictionary under $complete\_data[pilot][prix]['Circuit']$, representing the circuit data. This dictionary contains the circuit name under the 'Name' key and sets the 'Length' and 'Turns' keys to None for now. This step associates the circuit information with the specific race for the driver in the *complete_data* dictionary. After this process, the *complete_data* dictionary will have been updated with the circuit information for each driver and their respective races, making the data more comprehensive and complete. The length and number of turns for each circuit will be filled in later during the "Fourth Webscraping" of the code, which fetches additional data about the circuits from Wikipedia. At the end, a json file is created for each year with the data of the corresponding Season: Season_*year*.json. An example of this pipeline is reported in Figure 6 where the merging is done for 2013 data.

Figure 6: This image shows how the integration of data related to the circuits was implemented using Python after data acquisition and check of similarity.

4.2 Dataset about Qualifying Maximum Speeds

- Enrichment of the Drivers data with their Maximum Qualifying Speeds:

  In this code we combine data from different JSON files. The first type of JSON file contains a list of dictionaries corresponding to the maximum speeds achieved by drivers during the qualifying sessions of a specific Grand Prix and in a specific year (e.g Qualifying_Max_Speed/2013/Abu Dhabi GP - 2013). The

second type of JSON file contains nested dictionaries resulting from previous (the content of *complete_data* stored in Season_2013.json for example)

The goal of this enrichment is to enhance the data in the first JSON file with the qualifying speeds data, using the driver's name as the key.

In the code we store the first type of JSON data in the variable $first\_json$ and the second type of JSON data in the variable $second\_json$.

- The code iterates through each dictionary in $first\_json$ (loaded from a folder indicating the year 2013, 2014, 2021 and 2022) to extract the driver's name (driver) and the Grand Prix (gp) for which the qualifying speed data is available.
- To find approximate matches of the driver's name in the second JSON file, the code uses fuzzy string matching methods (fuzz.token_sort_ratio and process.extract). The limit=3 argument specifies that it will only consider the top 3 matches. If a driver's name is found to have a match with a similarity score of 70 or higher (which means that the function found three matches along with their similarity scores), it retrieves the corresponding driver's name from second_json. It then proceeds to find approximate matches for the Grand Prix name within the matched driver's data in second_json. If a Grand Prix name is found to have a match with a similarity score of 99 or higher, it inserts the "km/h" data from first_json into the corresponding location in second_json.
- After merging all the data, the code writes the final merged JSON file with the updated data to a new file named "$Qualifying\_Season\_Merged\_YEAR.json$" in a folder named "$Merged\_Data$".

An example of this data integration process is illustrated in Figure 7. In this example, the enrichment was performed on Max Verstappen's maximum speed during the qualifying sessions of the 2021 Bahrain Grand Prix. The pipeline successfully matched the driver's name "Verstappen" and the Grand Prix "Bahrain" from the first JSON file to the corresponding entries in the second JSON file. Subsequently, the "km/h" data for Max Verstappen's qualifying speed was added to the appropriate location in the merged JSON file.

Note: The fuzzywuzzy library's fuzz.ratio() function calculates the similarity between two strings by using the Levenshtein distance algorithm, which measures the number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other. The resulting similarity score is expressed as a percentage, with higher values indicating greater similarity.

By using this approach, the code can handle slight variations or typos in driver and race names and still successfully approximate the correct matches, ensuring that the collected data is appropriately assigned to the corresponding drivers and races.

Figure 7: This image depicts how the data enrichment process was carried out for the driver 'Max Verstappen' during the Bahrain GP in 2021.

4.3 Dataset about Weather Condition

- Enrichment of Grand Prix data with the weather condition during qualifying Sessions
  - The script takes two JSON files, the first one with the data for the current season's qualifying races (e.g Qualifying_Season_Merged_2013.json) and the second one with weather data for the same Season (e.g 2013/Weather_Qualifying.json), and merges the weather information into the corresponding Grand Prix details.
  - Once the data from both files is loaded into variables *data*1 and *data*2, the script starts iterating through the data from the first JSON file (*data*1). This data is organized in a nested structure, with information about drivers, races, and corresponding details.
  - For each race in *data*1, the script checks whether the same race exists in the weather data (*data*2). This check is necessary to ensure that there is weather information available for that particular Grand Prix.
  - If the script finds a match, it proceeds to check whether the GP's details in *data*1 include a "Qualifying" section. This section is essential because the weather condition during the qualifying session is what needs to be enriched.
  - When both conditions are met, the script adds the "Weather Condition" from *data*2 to the "Qualifying" section of the corresponding race's details in *data*1. By doing this, it combines the original data from *data*1 with the additional weather information from *data*2.
  - After processing all races and enriching the qualifying details with weather information, the script saves the updated data (*data*1) into a new JSON file (named for example FinalDataset2013.json).

An example of this data enrichment is shown in Figure 8.

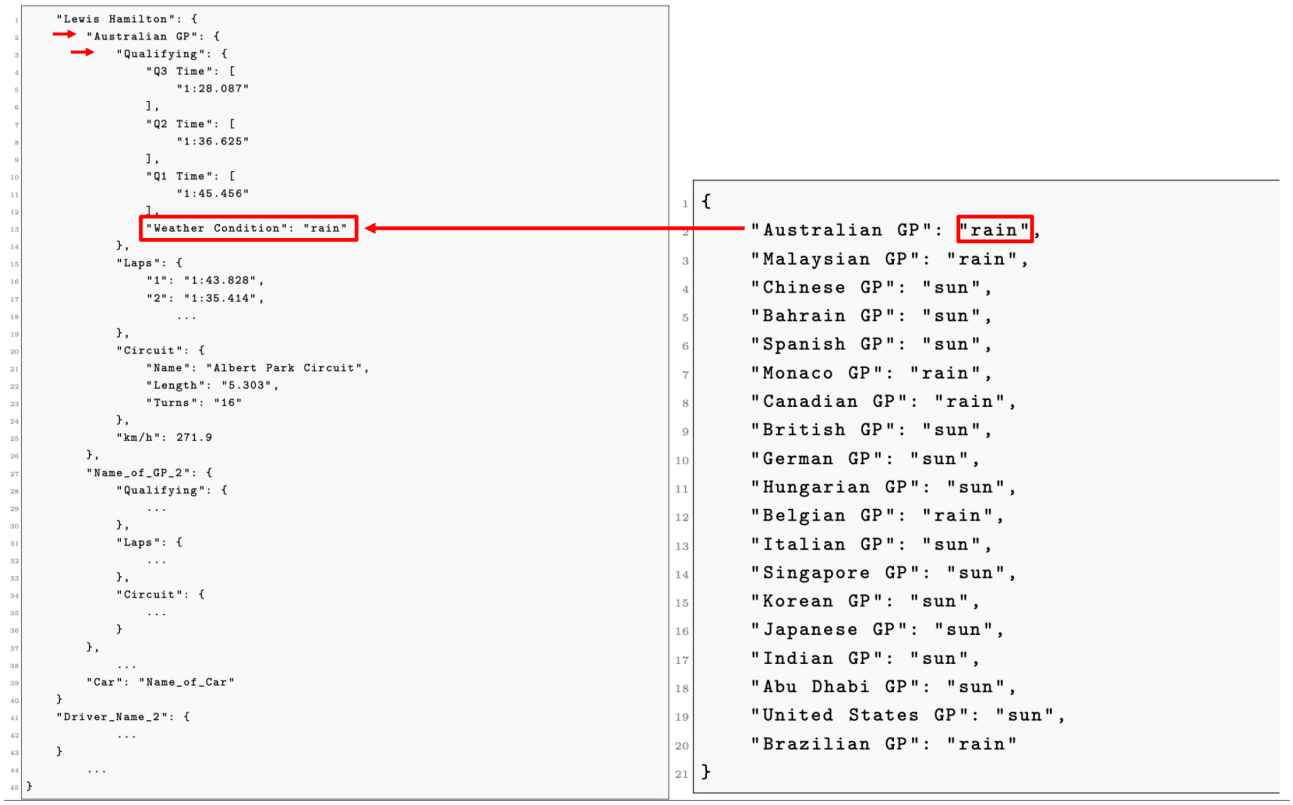


Figure 8: This image depicts how the data enrichment process of weather condition during qualifying session was carried out for the driver Lewis Hamilton during the Australian GP in 2013.

- Enrichment of Grand Prix data with the weather condition during race sessions.
  - The script starts by loading the contents of the first FinalDataset JSON file corresponding to the year we are analyzing (for example FinalDataset2013.json). This file contains a comprehensive set of data about the races that took place during the current season.
  - Next, the script loads the contents of the second JSON file (in 2013/Weather_Race.json for example). This file specifically contains weather-related data pertaining to each race lap during the season.
  - The script iterates through the data from the first JSON file, accessing each driver's performance in their respective Grand Prix and the associated race details.
  - For every Grand Prix in the first JSON file (*data*1), the script looks for a corresponding match in the weather data from the second JSON file (*data*2). This step is essential to ensure that there is weather information available for the Grand Prix "Laps" section.
  - When the script identifies a race with matching weather data, it proceeds to check whether the race details in *data*1 include information about individual laps ("Laps" section). This check is critical because it allows the script to enrich the data specifically with weather conditions during each race lap.
  - If both conditions are met, i.e a GP with matching weather data exists, and the grand prix details in *data*1 contain lap information, the script merges the "Weather Condition" from the second JSON file (in *data*2) into the corresponding "Laps" section of the first JSON file (in *data*1).
  - Finally, after successfully processing all the Grand Prix and enriching the "Laps" details with weather information, the script saves the updated data (*data*1) back into the original final JSON file (e.g FinalDataset_2013.json).

An example of this data enrichment is shown in Figure 9.

Figure 9: This image depicts how the data enrichment process of weather condition during race session was carried out for the driver Lewis Hamilton during the Australian GP in 2013.

5 DATA QUALITY ON MERGED DATA

5.1 Completeness

At this point we had to check if with the merge everything worked correctly, in order to not have errors in the data that we are going to store in the next step.

5.1.1 Maximum Speed Merge JSON

We started by analizing the first merge between the Seasons JSON files and the Maximum Speed JSON files. Due to the fact that the modification applied were just the addition of a new key "km/h" in the "Circuit" field of each driver and GP if the dictionary of that driver exists in the Maximum Speed JSON and so the "km/h" key is found. We expected to find the same results as in subsubsection 3.1.2 per each year and fortunately we obtain were what we expected: for each year, each GP counts the correct number of "km/h" that corresponds to the number we found out during the analysis of the MAximum Speed JSON file. We could say that the merge process have worked correctly.

5.1.2 Qualifying and Race Session Weather Merge JSON

Then we had to check the last merge step between the season files updated with the maximum speed data and the weather condition JSON file. We expected to find for each driver and each GP a new field in both sections (if exist) "Qualifying" and "Laps" called "Weather Condition" with the "sun" or "rain" value.
We checked our results with Tables from 4 to 7. We found the results we expected:

- 2013: for each GP there are 22 "Weather Condition" sections in "Qualifying" field that corresponds to the 22 drivers who took part each qualifying session. Considering the race sessions of 2013, we found out that there were some GPs with a lower number than 22 of "Weather Condition", in fact, for Australian, Italian, Abu Dhabi and United States GPs we counted the existence of only 21 "Weather Condition" in the "Laps" section, while Japanese GP only counts 20 "Weather Condition" sections. This data is in line with the previous results obtained (see Table 4 and its analysis).
- 2014: considering the qualifying part, United States, Brazilian and Abu Dhabi GPs count 18 "Weather Condition" sections and the Russian GP counts 21 "Weather Condition" fields, all the other GPs count 22 sections, in line with what we have seen through Table 5. Also the results of race sessions are in line with what we found in previous analysis (Table 5): instead of 22 fields, Malaysian, German, Singapore and Russian GP only count 21 "Weather Condition" sections, Australian, Monaco, Canadian, British and Abu Dhabi GPs count 20 "Weather Condition" fields, Brazilian GP counts 18 fields and United States only 17.
- 2021: in this season, the "Qualifying" field counts 20 "Weather Condition" sections for all GPs but one (Monaco GP) that counts 19 fields as we could imagine from Table 6 and its analysis. The "Laps" field contains 20 times the "Weather Condition" field for all GPs, apart from these exeptions: Bahrain, Emilia Romagna, Monaco, Austrian, British, Italian and Abu Dhabi GP count 19 "Weather Condition" sections in the "Laps" field, Mexico City GP counts 18 sections and Hungarian GP only 16.
- 2022: the "Weather Condition" sections counted in the "Qualifying" field are 20 for each GP, while for the race part the Belgian and Emilia Romagna GPs count 19 "Weather Condition" fields, the Saudi Arabian, Japanese and Brazilian GPs 18 and the British GP only 17, instead of the 20 of all the other GPs. These results are in line with the one analyzed before from Table 7.

We could conclude that the merging process worked correctly and that we could proceed with the storage process.

6 DATA STORAGE

After completing the data collection and data integration phase, and conducting initial quality improvement checks, we proceeded to import each data file as a separate MongoDB collection. In the end, we obtained four collections, each corresponding to a specific Formula 1 season. Each file is organized following a hierarchical structure, where the primary key of the dictionaries is represented by the driver's name, while the corresponding value serves as the secondary key of another dictionary, and so on.

The decision to use the driver's name as a key rather than as a value in a generic dictionary such as "driver_name": "Fernando Alonso" stems from the fact that our analysis focuses on pilot data from the Formula 1 seasons of 2013, 2014, 2021, and 2022, which remain constant over time. As driver names are unique identifiers, using them as keys enhances data integrity and ensures a clear and unambiguous representation of driver-specific data across multiple seasons.

We made the decision to use MongoDB as our NoSQL database for storing Formula 1 data due to two primary reasons:

1. Flexible Document Structure:

   MongoDB allows us to create collections containing documents with different schemas, which is beneficial for storing diverse Formula 1 data. For instance, we can efficiently store information about drivers and races without the need to allocate memory for data that doesn't apply universally. For example, if a certain driver did not participate in a specific GP, the data related to the race and qualifying sessions would not be included in that GP document. Instead, the GP document would only contain the structural characteristics of the circuit, such as the name, length, and curves of the track. Additionally, if a driver

was disqualified during the qualifying session, the "Qualifying" field would not even be added to the nested dictionary.

This data structure also allows us to include the laps completed by a certain driver during a specific Grand Prix without being constrained by the maximum number of laps for that particular race session. This approach avoids the issue of missing data; if a driver didn't complete the total number of laps, those laps would simply not be inserted. In a relational database, this wouldn't be possible, as 'nan' values would need to be added within the table where the driver didn't complete a lap.

Furthermore, if a driver did not compete in a specific Grand Prix or was disqualified, the weather conditions related to that race or qualifying session are not added, thus avoiding the problem of missing data that would naturally arise in relational models.

The maximum speeds achieved by drivers are included in the GP dictionaries only if the driver participated in the qualifying session. As previously mentioned, this approach ensures that the speed data is included even if the qualifying times are missing (i.e., the "Qualifying" dictionary is either empty or not inserted). This is because the absence of qualifying times does not imply that the driver's maximum speed was not recorded, for example, before an incident or any irregularities leading to disqualification. By adopting this strategy, we maintain data completeness and avoid potential inaccuracies in our analysis.

2. Single File Storage:

   With MongoDB, we can store all the Formula 1 data in a single .json file for each year, which simplifies data management and prevents the need to join multiple tables for every query. This design choice streamlines the querying process, avoiding potential errors during the joining phase that could arise in traditional relational databases. By keeping all the data in one place, we ensure better data integrity and easier data handling for users.

7 DATA QUALITY ON STORED DATA

The last data quality we need to do is on the stored data. The only check needed was the one related to the completeness of the final dataset.

7.1 Completeness

We had to performe some checks on the Final_dataset collection on MongoDB, to make sure that all the data were joined correctly. First of all, we opened the tree of a few documents, to check that the tree's structure was the intended one. Fortunately, it was. Then, we applied the same query codes we used before for the data quality of raw data and merged data but this time we have combined them to a unique code following the final structure of our database, so this time the order we followed to do the check was this one:

→ checking the number of drivers
→ checking the number of GPs per driver
→ checking the existence of the "Car" field for each driver with respective value
→ checking the existence of the "Qualifying", "Laps", "Circuit" and "km/h" fields for each GP
→ checking the existence of "Q1 Time", "Q2 Time" and "Q3 Time" in the "Qualifying" field with respective values
→ checking the existence of "Weather Condition" in the "Qualifying" field with "sun" or "rain" value
→ checking the existence of all numbers of laps in the "Laps" field with respective values
→ checking the existence of "Weather Condition" in the "Laps" field with "sun" or "rain" value
→ checking the existence of "Name", "Length" and "Turns" in the "Circuit" field with respective values

The results obtained were exactly the same as the previous ones on the JSON files, so the storage process worked correctly without the loss of any data. Now we could star with our analysis on our final dataset.

8 DATA EXPLORATION

At this point we had the necessary data to try to answer to our initial research questions. We divided the data exploration section in three part: the first one dedicated to the qualifying times data, the second one focused on the race times data and the third and last one takes in consideration the maximum speed data. Our main goal is to compare this data over years, considering first seasons 2013 and 2014 and then 2021 and 2022.

Before beginning these three analysis we used the circuit data contained in the "Length" and "Turns" fields to show through a graph the characteristic of each GP circuit. First, we considered season 2013 and 2014. Due to the fact that our analysis is based on the comparison between the two years, we only wanted to consider the common GPs, that in this case are: Spanish, Abu Dhabi, United States, Japanese, German, Bahrain, Singapore, Italian, Monaco, Brazilian, Canadian, Hungarian, Australian, Chinese, Malaysian, British, and Belgian GP (so, 17 GPs, in spite of the 19 totals for each of the two years).
In Figure 10 it is possible to see the results of the scatterplot.

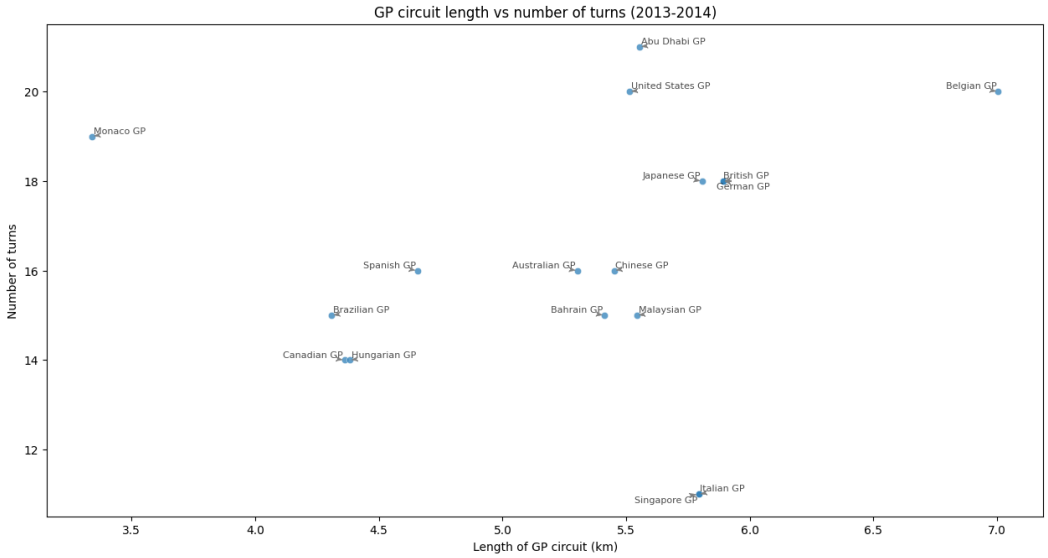


Figure 10: Scatterplot of length values and number of turns for common GPs in 2013 and 2014.

We can see that there are only few circuits with very similar characteristics like Singapore and Italian GP, Japanese and German GP, Canadian and Hungarian GP, but the other cases show different structures.

Then we created the same graph for 2021 and 2022 seasons taking in consideration only that GPs in common for both years. We can see them through Figure 11.

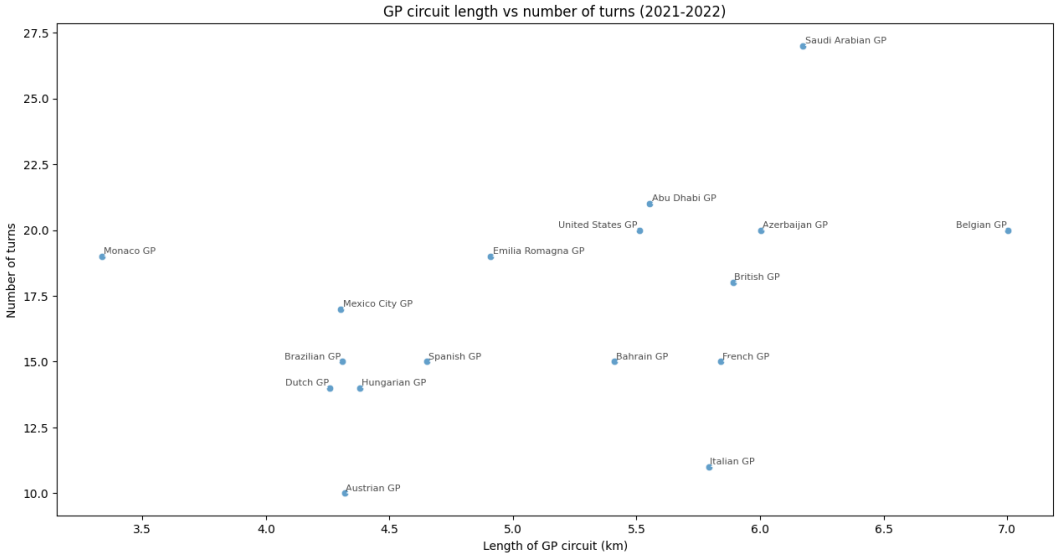


Figure 11: Scatterplot of length values and number of turns for common GPs in 2021 and 2022

This time the remaining ones are: Monaco, Mexico City, Brazilian, Dutch, Hungarian, Spanish, Austrian, Emilia Romagna, Abu Dhabi, United States, Azerbaijan, British, Bahrain, French, Italian, Saudi Arabian and Belgian GP (17 GPs in common, in spite of the 20 GPs for each year). This time there are no GPs that show very similar structure.

Finally we could say that every GP has his own characteristic and maybe this aspect could help us to find some patterns in the following analysis.
Now we can start with the first section of data exploration.

8.1 Qualifying times

As anticipated, we started our analysis with a query about the Qualifying times. We wanted to make a comparison first between the two years 2013-2014 and then between 2021-2022 in order to answer, in part, to our research questions, so we wanted to understand which was the impact of the changes applied between the two different season (we have to specify that we considered the two couples of years separately, everything we did for years 2013 and 2014 we also did later for years 2021 and 2022). In order to do that, we divided in two parts this first query, focusing on two different aspects:

- **First Focus - Qualifying times comparison in each GP for each driver**: We decided to first consider each GP separately from each other and for each one show the results of qualifying times for each driver, to see if there was clearly behaviour or patterns for the two different seasons.
- **Second Focus - Overall average qualifying times comparison for each Team**: then we wanted to plot an overall results and comparison among all the GP qualifying times and we decided to do that considering the Teams and not each single driver anymore.

We wrote a code that could show us the desired results. Of course the code worked on our final and stored data on MongoDB. In order to simplify the code explanation and understanding we divided it into four parts:

Part I: Data Processing and Filtering

1. The code connects to the databases stored in MongoDB system, in this way it has access to the four collections containing the databases relative to each year.

2. It is defined a function that can convert the time from the format "minutes, seconds, milliseconds" to "seconds" and we can apply it later to qualifying times when the code accesses to them.

3. After having established a connection with the databases, for each of them the code iterates over each driver and GP. In each GP it searches the "Qualifying" field and if it found it controls if the value correspondent to the "Weather Condition" field contained in the "Qualifying" section is equal to "sun". This passage eliminates all the GP in which the weather condition was no good. This selection was made because bad weather condition couldn't allow us to make good comparison, in fact, the rain or the wet track influences the times and the performances of cars. That's why we would not consider the GP in which bad weather condition would not give us usable results.

Part II: Qualifying time selection for driver

1. If the good weather condition is found, the code searches for the "Q1 Time", "Q2 Time" and "Q3 Time" and its values and if it found any of them it saves them into the variables *q1_ time*, *q2_ time* and *q3_ time*. Considering our previous analysis done during the data quality process, the possible results for each driver and GP of this passage are:

   - no values found;
   - value relative to *q1_ time* found;
   - values relative to *q1_ time* and *q2_ time* found;
   - values relative to *q1_ time*, *q2_ time* and *q3_ time* found.

   So, for each driver and GP we could have no values, one value, two values or three values.

2. Considering the fact that we wanted just one value per year for each driver, the code selects the minimum value among the found ones. In this way we knew that we were working with the best result among the recorded ones for each driver in each specific GP. Considering that the three qualifying sessions are taken separately we could consider that each time is not influenced by the others, so it is possible to consider one of them separately from the others.

3. Then a dictionary for each year called *data_ year* (where the word year must be substituted with 2013, 2014, 2021 and 2022) is created and the primary keys are the name of the GPs (the ones with only "sun" weather condition in the "Qualifying" section), and then for each GP there are secondary keys corresponding to each driver name with the qualifying time selected before as the value.

Part III: Average qualifying time calculation for Teams

1. For the second focus of our analysis the *data_ year* dictionary assumes a different structure, in fact, each driver of each GP contains two sub-keys instead of one value. The keys are "Time" which contains the value of the minimum qualifying time and then "Car", which contains the value relative to the Team of the driver.

2. Then a new dictionary is created started by the *data_ year* dictionary. This new dictionary called *aggregated_ data_ year* contains the name of the Cars (or Teams) as primary keys and it contains all the qualifying times relative to the Teams driver.

3. Finally, a new dictionary called *average_ data_ year* is created where for each Car the mean of the qualifying times is calculated (in this way we have for each Team just one value that is the mean of the qualifying time values of the drivers of that Team).

Part IV: Data Representation with Scatterplot

1. Starting from the scatterplots of each GP where the driver data are shown, we used the *data_ year* dictionaries created. We first consider the *data_ 2013* and *data_ 2014* and we create a list called *lines* in which we inserted three values: the drivers there were in common in the GP of both years and the qualifying time for 2013 and for 2014 (if they were not None).

2. Then we used that list to create one scatterplotfor each GP in which, on the x-axes there are the names of the drivers, on the y-axes there are the seconds and in the graph are represented by points, the qualifying times for each driver (the two years are distiguible by different colors of points).

First Focus - Qualifying times comparison in each GP for each driver

The GPs that "passed" the filtering phase for years 2013 and 2014 were just eight: Abu Dhabi, Bahrain, German, Italian, Japanese, Singapore, Spanish and United States.
The graphs of the GP qualifying times obtained for year 2013 and 2014 showed three types of behaviour:

1. 2013 qualifying times lower than 2014 qualifying times for all the drivers (or the vast majority);
2. 2014 qualifying times lower than 2013 qualifying times for all the drivers;
3. a mixture of 2013 qualifying times lower and higher than 2014 qualifying times.

Considering the first case, in the majority of the GP it is possible to see this behaviour. As an example, we show the Spanish GP graph in Figure 12.

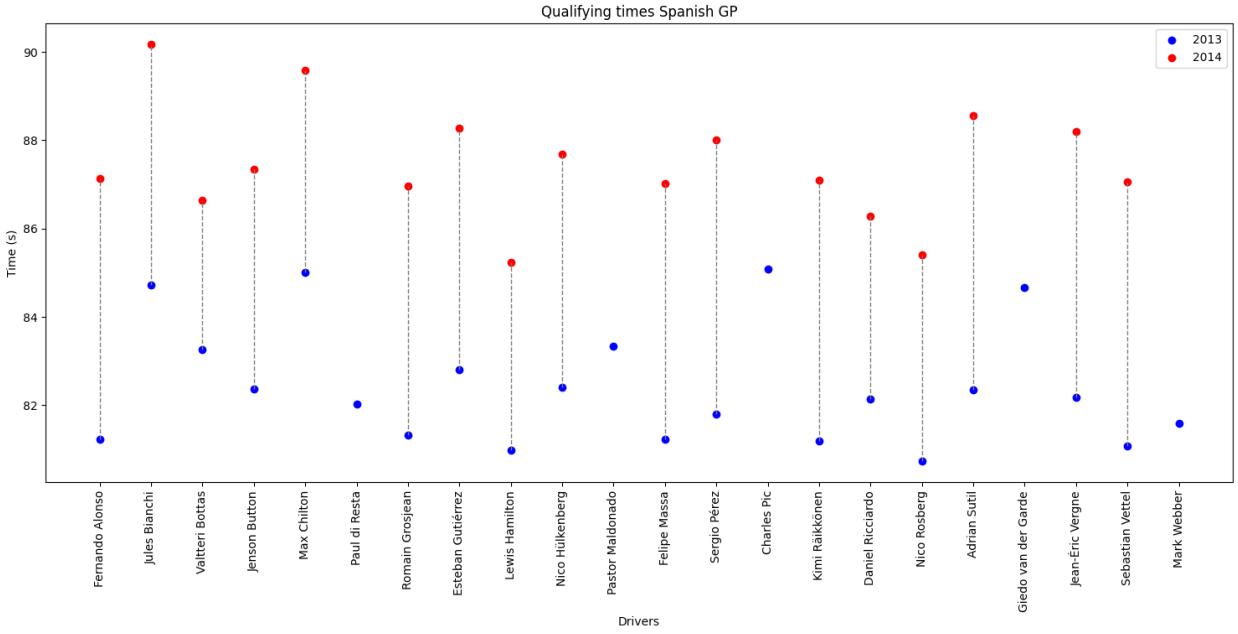


Figure 12: Graph of Spanish GP qualifying times per driver (2013 vs 2014).

The Spanish GP graph above (Figure 12) shows a clear separation between the values of the two years and despite the fact that the other GP graphs don't show this clear division, we could find the same behaviour for which every driver recorded higher qualifying times in 2014 (red points) with respect to 2013 (blue points) also for Japanese, Singapore and Bahrain GP. We could also add the Italian and Abu Dhabi GP in which just for one or two drivers the opposite behaviour of qualifying times is shown (the qualifying time relative to 2013 is higher than the one relative to 2014). Considering the graph in Figure 10 we could notice that Spanish GP is the one among the eight GPs analyzed in this section that has the lowest length.
Considering the second situation we only have one case of this behaviour, that is with the German GP, as we can see in Figure 13.

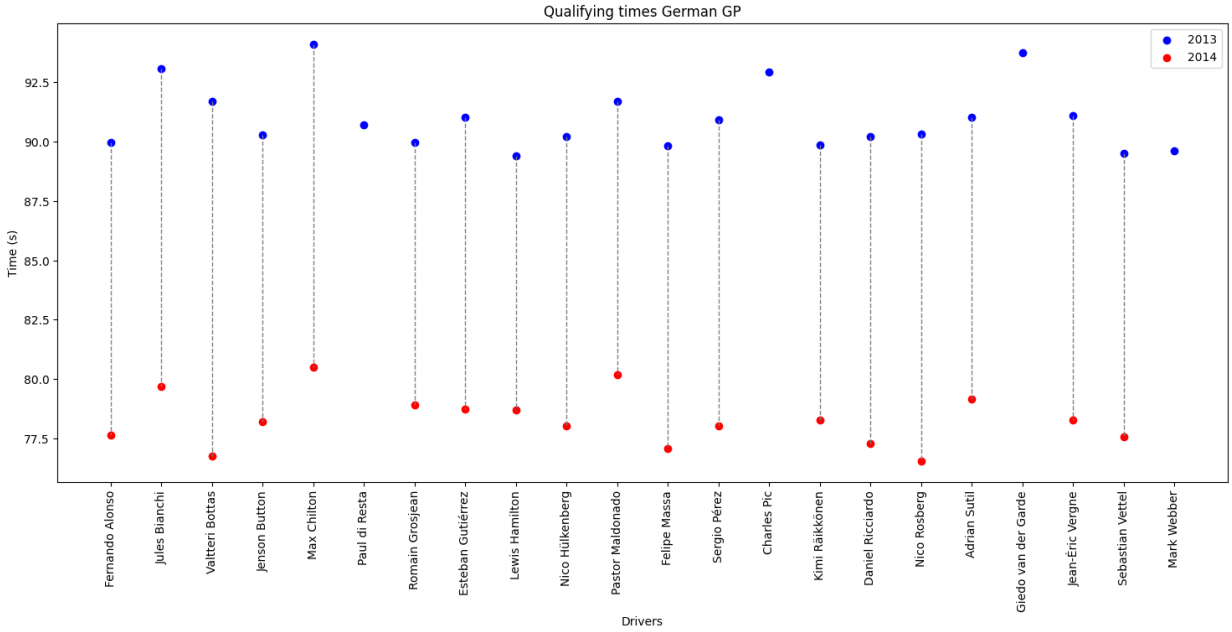


Figure 13: Graph of German GP qualifying times per driver (2013 vs 2014).

Here the situation is exactly the opposite of the the previous cases. Making a reference to the graph in Figure 10 we can see that German GP is the one, among the eight considered in this analysis, that is the longest in terms of kilometers. This could led us thinking that with a longer circuit, the changes between 2013 and 2014 in terms of engine could give better results, but of course we can't be sure about it and considering the fact that Japanese GP that shows the opposite behaviour with respect to the German GP but has very similar length to that GP suggested that this is a weak assumption. The last GP graph, relative to United States, is the only one that assumes the third behaviour, as we can see through Figure 14.

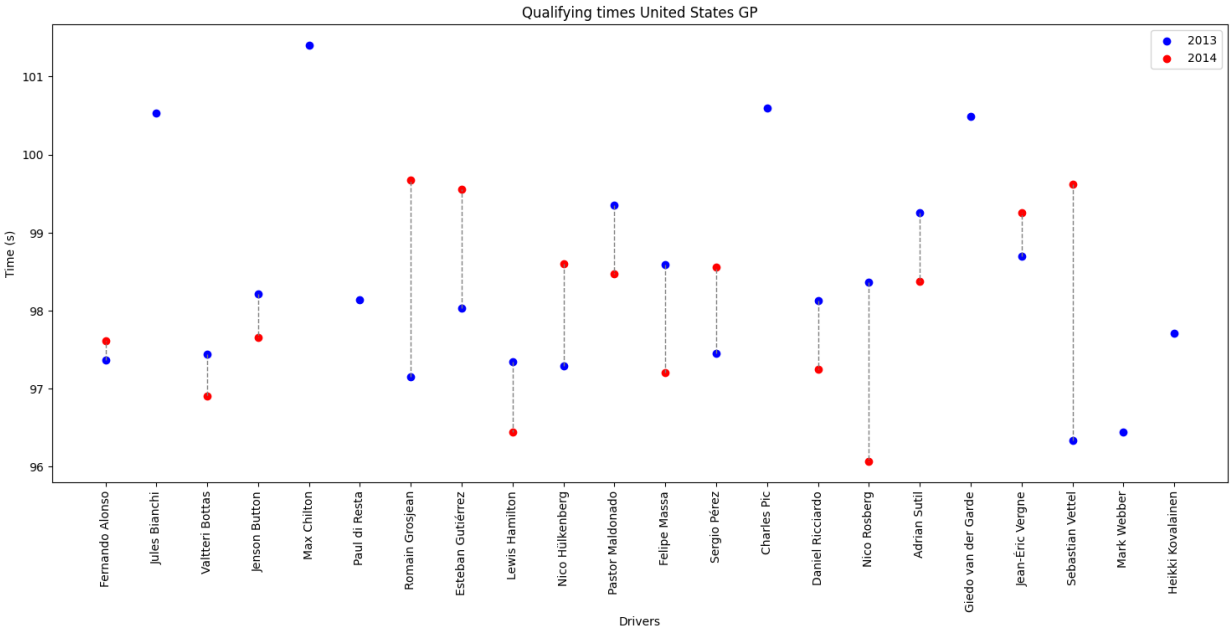


Figure 14: Graph of United States GP qualifying times per driver (2013 vs 2014).

From Figure 14 we can see that there is a clear separation between the values of the two years and that there are many drivers for which the qualifying time relative to 2013 is lower than the one relative to 2014 but many

drivers show the opposite behaviour. Considering the graph in Figure 10 is not easy to assume positions about this data. The United States GP is the one with the highest number of turns, but the fact that the new engine can be related to the number of turns in a circuit to explain performances of cars is not sustained by the other cases of GP results.

Now we can consider the other couple of years 2021 and 2022.
This time the number of GPs that showed good weather condition in both years is twelve: Abu Dhabi, Austrian, Azerbaijan, Bahrain, Dutch, French, Hungarian, Italian, Mexico City, Monaco, Saudi Arabian and Spanish. For these two seasons only one behaviour of qualifying times was shown by the graphs and it is the one that we can see in Figure 15 relative to the Monaco GP.

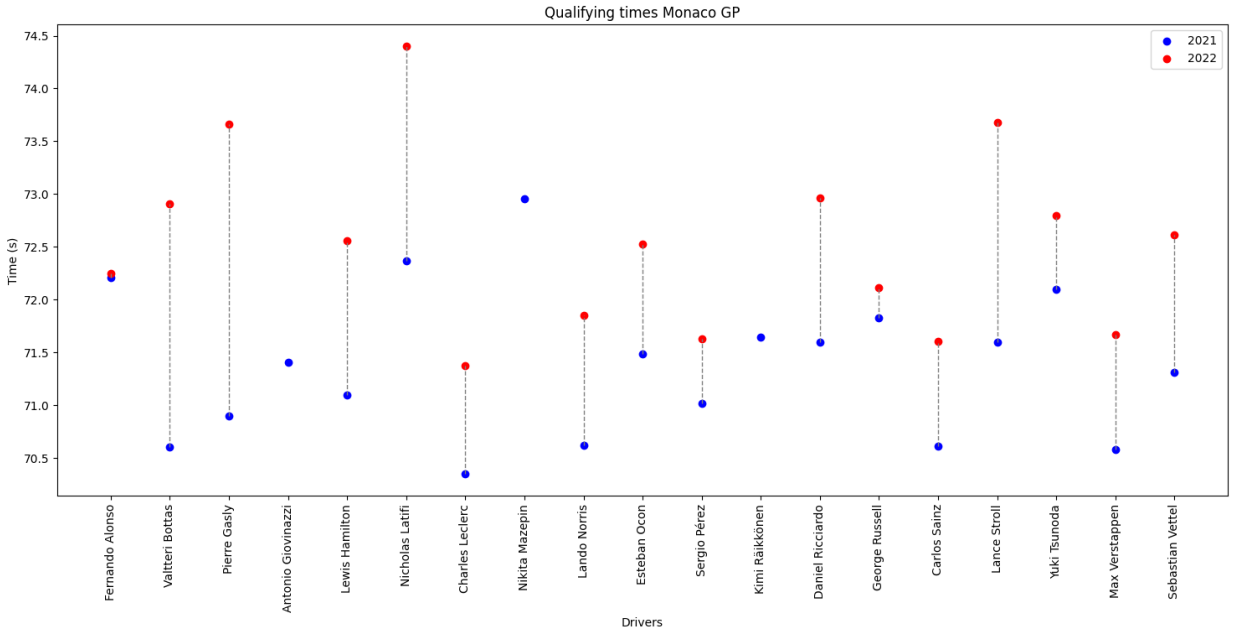


Figure 15: Graph of Monaco GP qualifying times per driver (2021 vs 2022).

All the twelve GP graphs showed higher qualifying times during season 2022 with respect to season 2021, even if the values of the two years cannot be clearly separated as it happens in Figure 12.This time making references to the graph in Figure 11 is not useful, because the behaviour is the same for all the GP independently from the length or the number of turns in a circuit.

Second Focus - Overall average qualifying times comparison for each Team

After the previous analysis we could expected that when we calculate the average of the qualifying times of all the GPs per Teams the trend showed by the graphs was lower qualifying times for 2013 and 2021 and higher qualifying times for 2014 and 2022.

From Figure 16 we can see that our though was correct apart from one exception: the Williams Team, but as we had said before, there are some graphs for these two years in which the behaviour of the qualifying times is the opposite of the predominant one, that explains this exception.

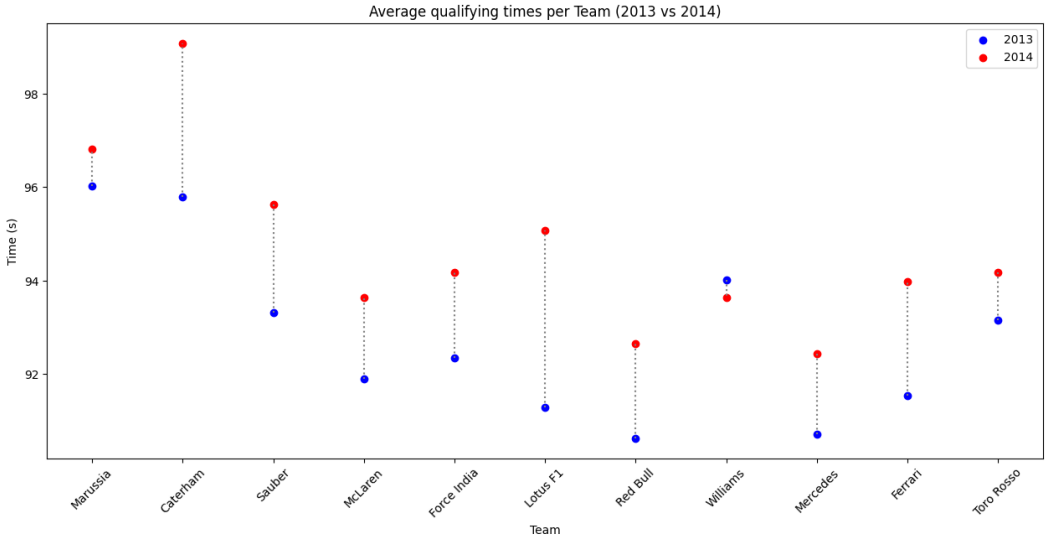


Figure 16: Graph of average qualifying times per Teams (2013 vs 2014).

For seasons 2021 and 2022 there were no doubt of what the overall average qualifying graphs could show: all the GP graphs shows a unique trend that sees the qualifying times relative to 2021 lower than the ones relative to 2022 and so does the graph in Figure 17.

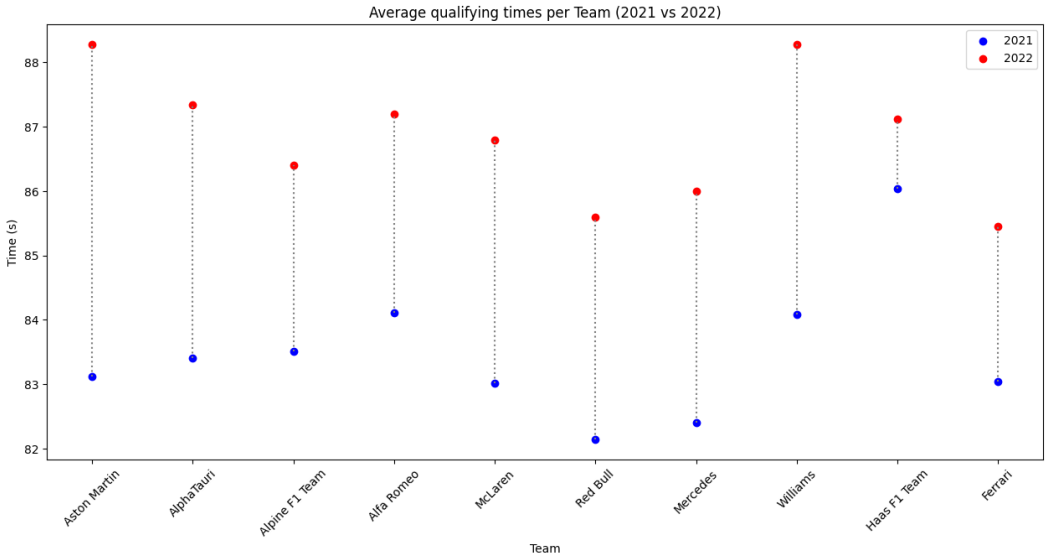


Figure 17: Graph of average qualifying times per Teams (2021 vs 2022).

At this point what can we answer about our research questions? It seems that the new changes applied between 2013 and 2014 and the ones applied between 2021 and 2022, in general, do not take improvement on the performances of cars, the vast majority of drivers and Teams had better results during 2013 with respect to 2014.

Does it seem strange? Maybe we did not expect that, because over the years we are let to think about an improvement of things, but we have to consider that years 2014 and 2021 were the first time that the new important changes were applied and the first time that the driver could try the cars on each track with the news was at the beginning of the GP, this should not surprise us if the first try with the new changes did not bring better results.

Comparing the graph in Figure 16 and 17, we can note that on the y-axes the reference values change: in 2013-2014 seasons the lowest average was around 90 seconds, while the highest average in 2021-2022 seasons was around 89 seconds, a very big improvement, but it happened in 7 years.

8.2 Race times

In this second query, our goal is to determine whether there has been a shift in the trend of the average lap times. These times are calculated across three distinct sections, into which the Formula 1 race times have been divided. We will examine whether there have been changes in the lap time trends from 2013 to 2014 and then from 2021 to 2022, mirroring the observations regarding the qualifying times discussed in the previous paragraph. Our analysis has two main focuses:

- **First Focus - Average times across sections for single drivers**: our primary focus lies on the average times obtained by the individual drivers during each Grand Prix race session, in order to determine whether there has been an increase or decrease in these times.
- **Second Focus - Average race times across sections for teams**: this time our analysis considers teams instead of distinguishing between individual drivers. Moreover, the averages are no longer computed per each individual Grand Prix; instead, they are calculated across all races within each year. This broader approach allows for a comprehensive perspective on the performance trends of the teams over the years. Before proceeding with these two analyses, we followed a series of steps within the query code to ensure that the data concerning average times were consistent with each other and could be compared with those of other years.

To do so, a Python code has been designed to analyze lap time data from Formula 1 races across different years (2013 and 2014, 2021 and 2022) and stored in a MongoDB database. The code can be divided into 3 main sections as follows:

Part I: Data Processing and Filtering

1. The code establishes a connection to the MongoDB cluster and accesses different collections within the specified database for each year to retrieve lap time data.
2. Within each document (correponding to a specific year), the code extracts lap time information for individual drivers and their corresponding laps in a specific Grand Prix. This involves navigating through the nested structure of the document to access lap time data. If the "Laps" section is not found, the analysis proceeds to the next driver.
3. For each driver and Grand Prix, the lap times are converted from their original format ("minutes:seconds") to seconds for consistency and ease of analysis. The lap time data is then stored in appropriate data structures, such as dictionaries.
4. The code also tracks the maximum number of laps completed in each Grand Prix.
5. The code extracts and stores information about the car used by each driver. This information is associated with the driver's lap time data.

6. For each driver participating in each Grand Prix, the code generates box plots using the filtered lap time data. Box plots consist of a box representing the interquartile range (IQR), which contains the middle 50% of the data. Outliers, which fall outside the "whiskers" of the box, are also shown on the plot. Outliers on these box plots are detected. Any lap time data point falling below the lower whisker or above the upper whisker is considered an outlier. These outliers are identified and subsequently removed by means of the IQR-based function (*remove_ outliers*) from the filtered lap time data, ensuring that the remaining data points are more representative of the overall distribution.

7. The filtered lap time data is then stored in dictionaries (*lap_ info_ data*) to preserve the structure of lap times per driver, Grand Prix, and lap number. Once the filtered data based on the box plots is obtained, further filtering is performed.

8. For each driver within each document, the code checks if there are Grand Prix data and if the weather conditions are "sun".
   Note: This check is fundamental for our analysis, as adverse weather conditions would compromise race times and then our analysis. Moreover times recorded under rainy weather condition would not be comparable to those achieved under "good" weather conditions.

9. The code then checks if the driver and Grand Prix have lap data available in *lap_ info_ data* dictionary.
   Note: This is also an important check, as the driver might not have participated in the race.

Part II: Execution of Averages Calculations per Section

1. The code calculates the maximum number of laps completed by the driver in that Grand Prix.

2. It calculates average lap times divided into three sections for each driver and for each Grand Prix. The sections are defined based on the driver's maximum completed laps. Three lists (section_1, section_2, section_3) are defined to store lap time data for each section.

3. The code calculates the average times for each section by summing up lap times and dividing by the number of times for that section. If there are no lap times for a section, a value of 0.0 is assigned.

4. The average times for each section are stored in dictionaries: *averages_ 2013*, *averages_ 2014*, *averages_ 2021*, *averages_ 2022*. Each dictionary is structured such that the Grand Prix is a key, and the driver is a sub-key, with the corresponding values being the average times for the sections.

Part III: Data Representation with Scatterplot

1. Common drivers between different seasons are found by comparing the lists of drivers created from *averages_ 2013*, *averages_ 2014*, etc. dictionaries.

2. For each Grand Prix in the 2013 data (averages_2013), the code checks if the same Grand Prix exists in the 2014 data (averages_2014). The same is done for 2021 and 2022.

3. It creates lists to store average times for each section for common drivers between 2013 and 2014 (and for common drivers between 2021 and 2022).

4. Scatterplots are generated using the Matplotlib library, and each point represents a driver's average time for a specific section of the race (**First Focus**). The colors and markers differentiate the sections and seasons.

5. A function *rearrange_ sections* is defined to reorganize data into a more convenient format. The function rearranges the sections of average lap times for each driver within each Grand Prix into a dictionary with Grand Prix keys and driver-car section values.

6. The 2013 and 2014 data are transformed using this function and assigned to *averages_car_2013* and *averages_car_2014*. The same is done for 2021 and 2022 data which are transformed and assigned to *averages_car_2021* and *averages_car_2022*.

7. For each year and for each car, cumulative values for lap times and counts for each section (Section 1, 2, 3) are calculated. In this way average lap times for each section are calculated for each car.

8. Scatterplot is generated to compare average lap times for different sections between car teams for the 2013 and 2014 seasons and for the 2021 and 2022 seasons (**Second Focus**). Each point on the plot represents a team's average time for a specific section, and points are grouped by sections and colored by season.

Analyzing the graphs obtained through the code described above, we can now answer the question regarding changes in average race times.

First Focus - Average times across sections for single drivers:

With specific reference to the years 2013 and 2014, the graphs depict the trends of average times depending on the Grand Prix that the drivers participated in, in both years. Therefore, each Grand Prix can be assigned to one of the following categories based on whether there was an overall improvement or worsening in driver performance from one year to the next:

- **Lower average times in 2013 compared to 2014**: Australian GP (range of times: approximately 90 - 104 seconds), Spanish GP (range of times: approximately 88 - 96 seconds), Italian GP (range of times: approximately 87 - 93.5 seconds), Chinese GP (range of times: approximately 100 - 110 seconds). Among the Grand Prix belonging to this category, there is one that cluster into ranges corresponding to the reference section: Chinese GP.
  Note: Cases in which there has been an increase in each of the three 2014 sections (at least for the 2/3 of them in the graph) are also included in this category.

- **Lower average times in 2014 compared to 2013**: German GP (range of times: 92.5 - 101 seconds), Brazilian GP (range of times: 74 - 82 seconds).
  Note: Cases in which there has been a decrease in each of the three 2014 sections (at least for the 2/3 of them in the graph) are also included in this category.

- **Average times neither lower nor higher from 2013 to 2014**: United States GP (range of times: 100 - 103 seconds), Singapore GP (range of times: 112 - 120 seconds), Bahrain GP (range of times: 98 - 106 seconds), Canadian GP (range of times: 77 - 82.5 seconds), Belgian GP (range of times: 112 - 120 seconds), Abu Dhabi GP (range of times: 105 - 111 seconds). Among the Grand Prix belonging to this category, there are some that tend to cluster into ranges corresponding to the reference section. They are: Chinese GP, Bahrain GP, Canadian GP, Belgian GP, Abu Dhabi GP, United States GP.

Three illustrative graphs have been provided, representing each of the mentioned categories (fig.18 for the first, fig. 19 for the second and fig. 20 for the third). These graphs show different colors, indicating the reference sections, and symbols denoting the corresponding years of the data.

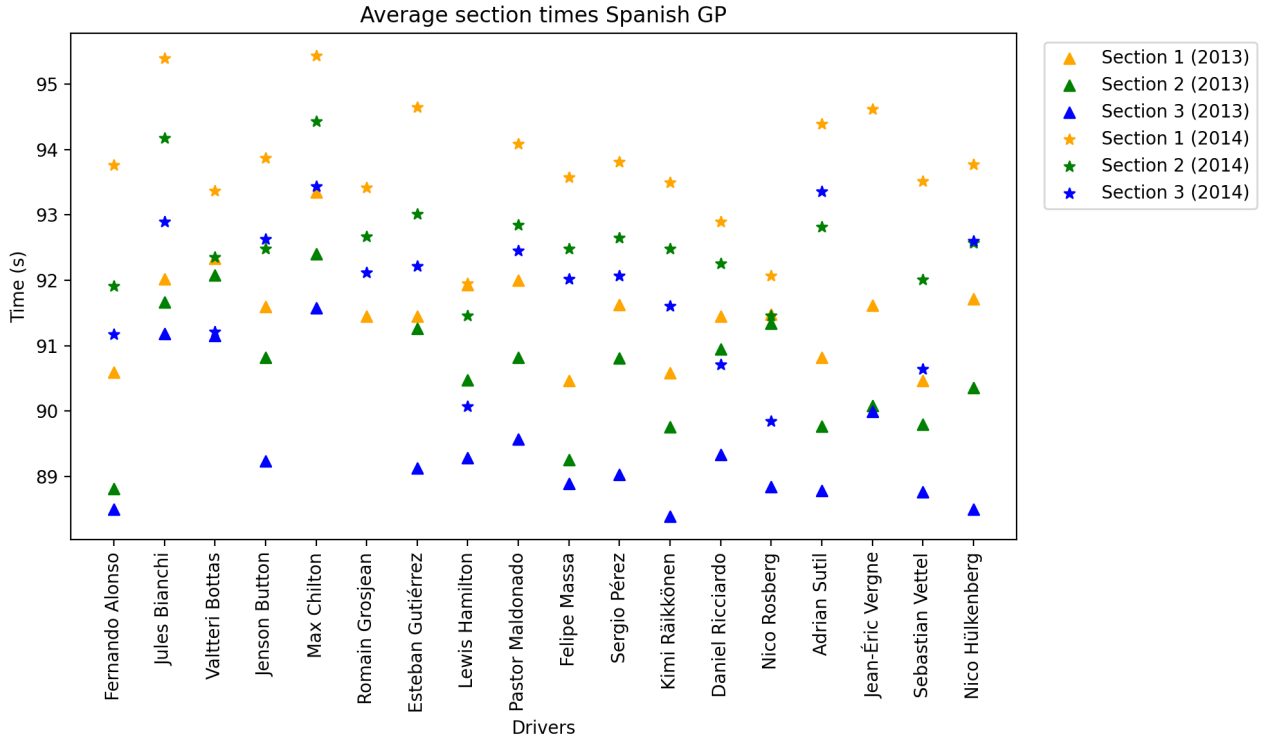


Figure 18: Graph of Spanish GP race times per driver (2013 vs 2014).

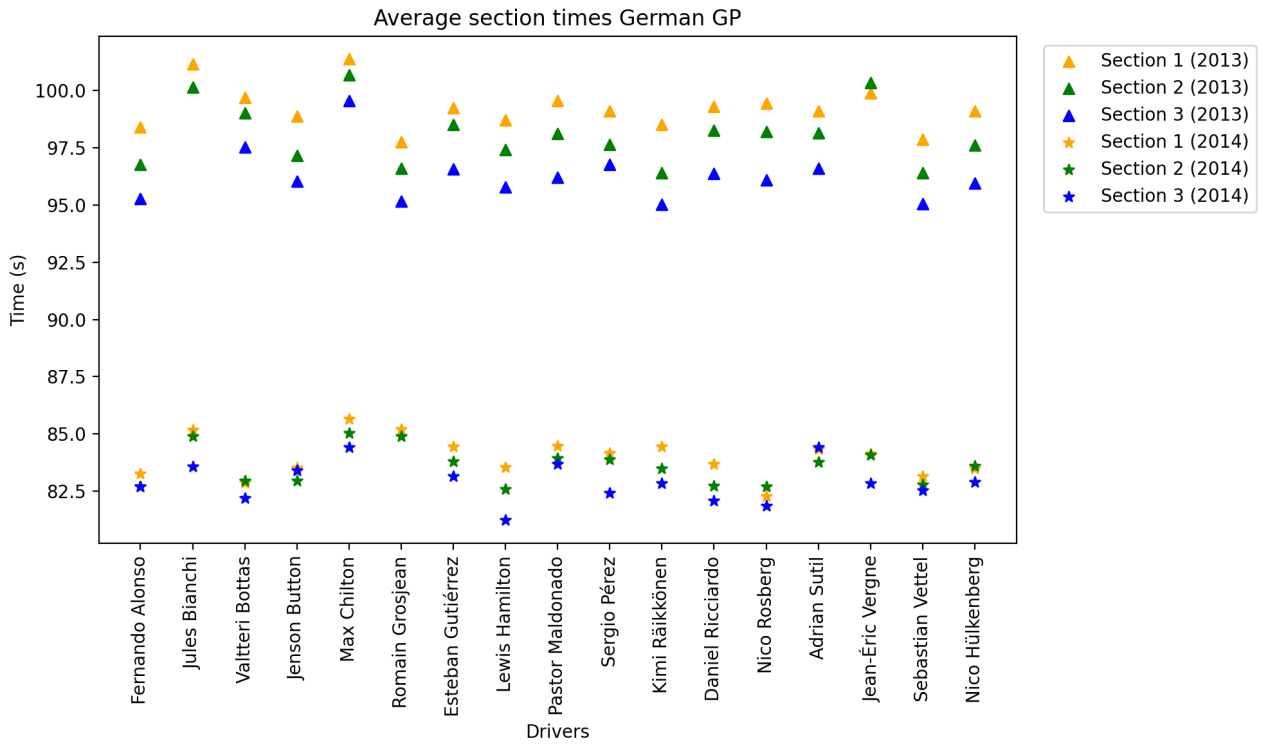


Figure 19: Graph of German GP race times per driver (2013 vs 2014).

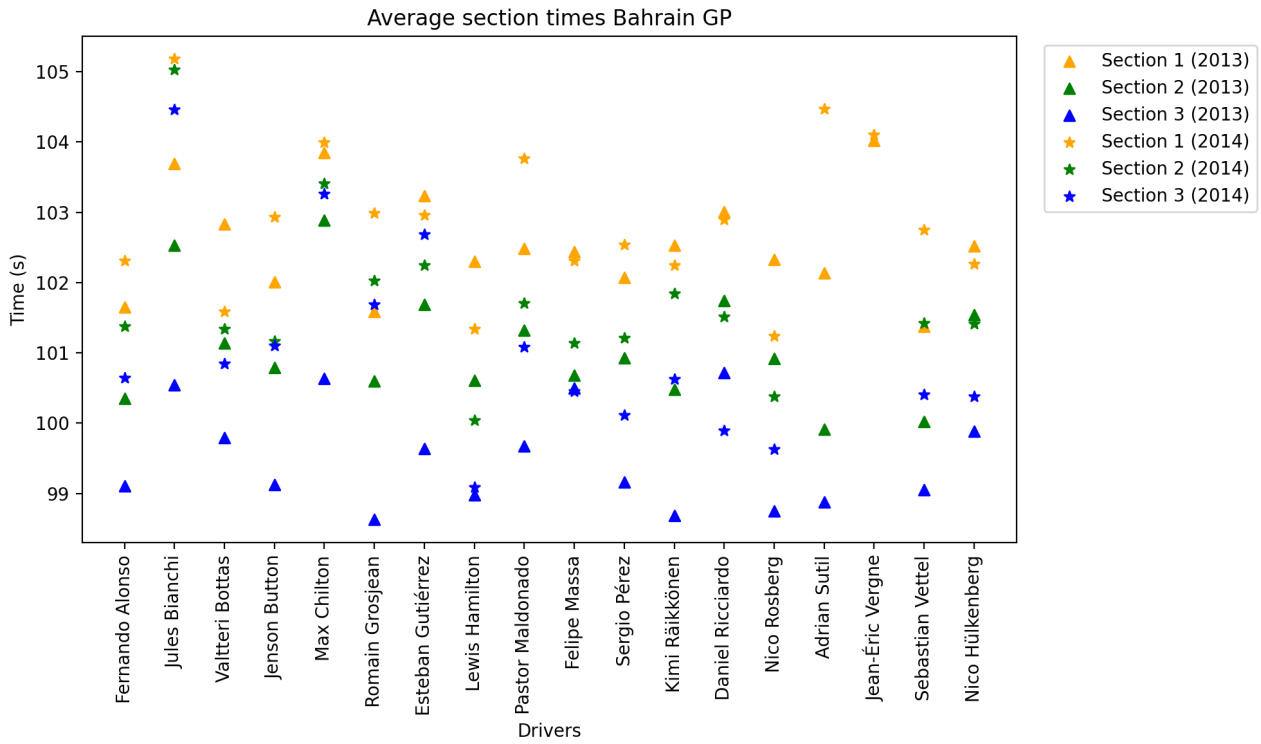


Figure 20: Graph of Bahrain GP race times per driver (2013 vs 2014).

From these plots, we can infer that changes in regulations had a significant effect in only two Grand Prix, benefiting the drivers times. In four cases, the average times were better in 2013 compared to 2014. In most instances, the overlapping of times from different sections makes it difficult to definitively define an overall trend. However, it is observed that generally, the trend of average times in a specific section remains relatively consistent in the subsequent year as well. This could be due to strategies employed by individual drivers or the characteristics of the circuit. We can notice that the lowest times of each year generally correspond to the third sections, the times just above the lowest times correspond to the second sections, and the highest times correspond to the first sections. By focusing solely on the third sections, it was noticed that in all Grand Prix, including those with mixed conditions, the majority of drivers recorded lower times in 2013 compared to 2014. This could be due to the higher speeds used by the pilots during the final phase of the race.

Examining the graph depicted in Figure 10, a notable observation arises: the Australian GP, Spanish GP, and Chinese GP all feature an identical number of turns. While this might prompt the speculation that a circuit with 16 turns, like those, could yield unfavorable outcomes due to the alterations in engine regulations between 2013 and 2014, such a conjecture remains uncertain.

Further analysis reveals that the German GP stands out as one of the lengthiest circuits in terms of kilometers. Consequently, one could tentatively posit that a longer track might have benefited from the engine changes in 2014 compared to 2013. Nevertheless, it's essential to exercise caution in embracing this notion, given that the Brazilian GP shares a similar profile with the German GP but possesses substantially lower track length.

Moreover, the behavior of the Belgian GP, the lengthiest circuit among the tracks, is intricate and doesn't align neatly with the earlier assumption. Hence, the preliminary inference made about longer circuits and the impact of the 2013-2014 engine changes appears tenuous.

We now focus on 2021 and 2022 years. The graphs depict the trends of average times depending on the Grand

Prix that the drivers participated in, in both years. Therefore, each Grand Prix can be assigned to one of the following categories based on whether there was an overall improvement or worsening in driver performance from one year to the next:

- **Lower average times in 2021 compared to 2022**: Bahrain (range of times: 94-103 seconds), Spanish GP (range of times: 81-91 seconds), Mexico City GP (range of times: 80-86 seconds), Brazilian GP (range of times: 73-79 seconds), Abu Dhabi GP (range of times: 87-93 seconds),Azerbaijan GP (range of times: 105-113 seconds), Austrian GP (range of times: 68 - 73 seconds), British GP (range of times: 90 - 97 seconds).
  Note: Cases in which there has been an increase in each of the three 2022 sections (at least for the 2/3 of them in the graph) are also included in this category.

- **Lower average times in 2022 compared to 2021**: Saudi Arabian GP, except for the the third section of 2021 times which is always lower than at least one of the sections of the 2022 times (range of times: 93-120 seconds).

- **Average times neither lower nor higher from 2021 to 2022**: French GP (range of times: 97-102 seconds), Dutch GP (range of times: approximately 74 - 80 seconds), Italian GP (range of times: approximately 85 - 90 seconds), United States GP (range of times: 100 - 103 seconds). Among the Grand Prix belonging to this category, there are some that tend to cluster into ranges corresponding to the reference section. They are: French GP, Dutch GP and United States GP.

Three illustrative graphs have been provided, representing each of the mentioned categories (fig.21 for the first, fig. 22 for the second and fig. 23 for the third). As for 2013 and 2014 graphs, these graphs show different colors, indicating the reference sections, and symbols denoting the corresponding years of the data.

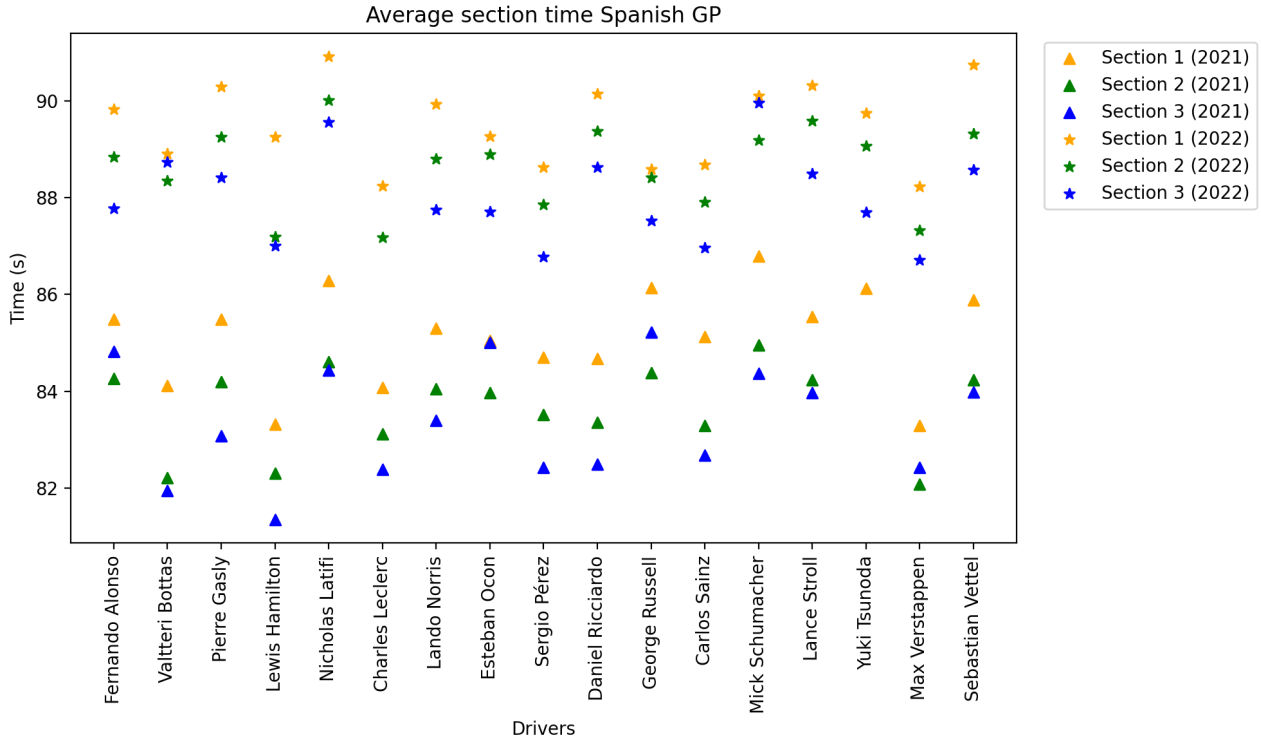


Figure 21: Graph of Spanish GP race times per driver (2021 vs 2022).

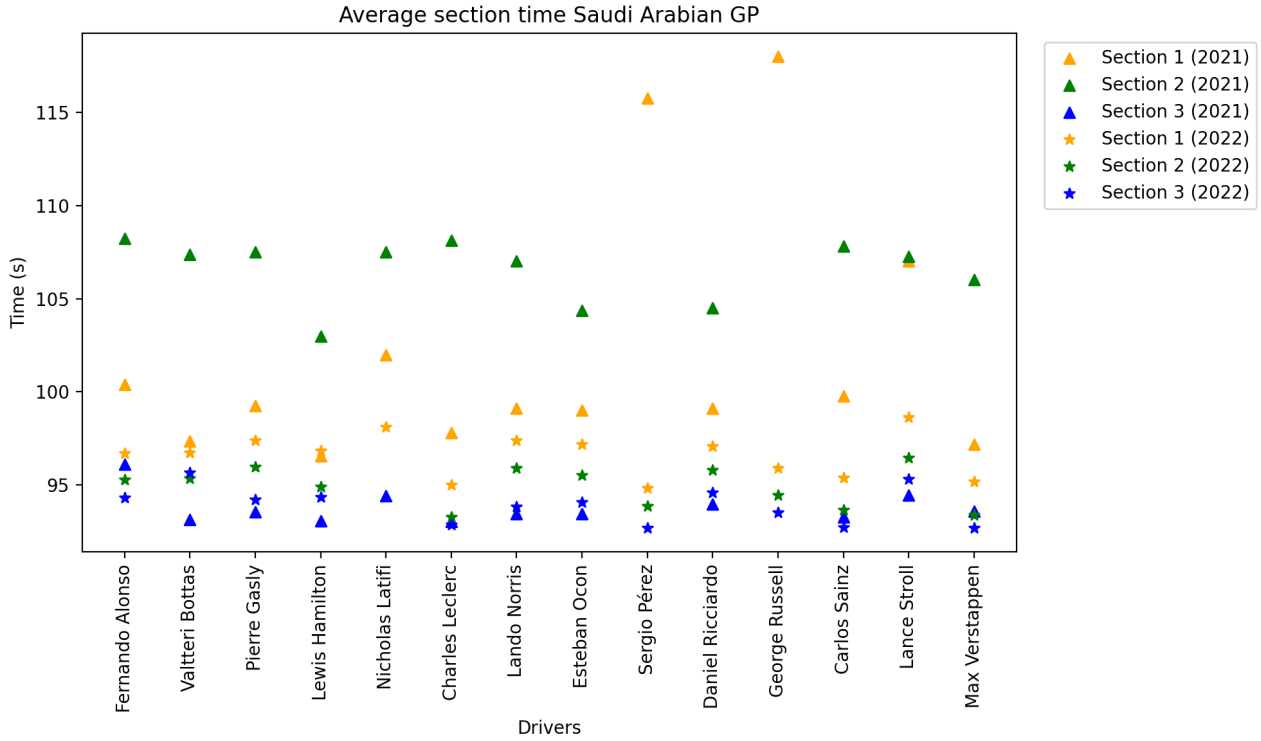


Figure 22: Graph of Saudi Arabian GP race times per driver (2021 vs 2022).

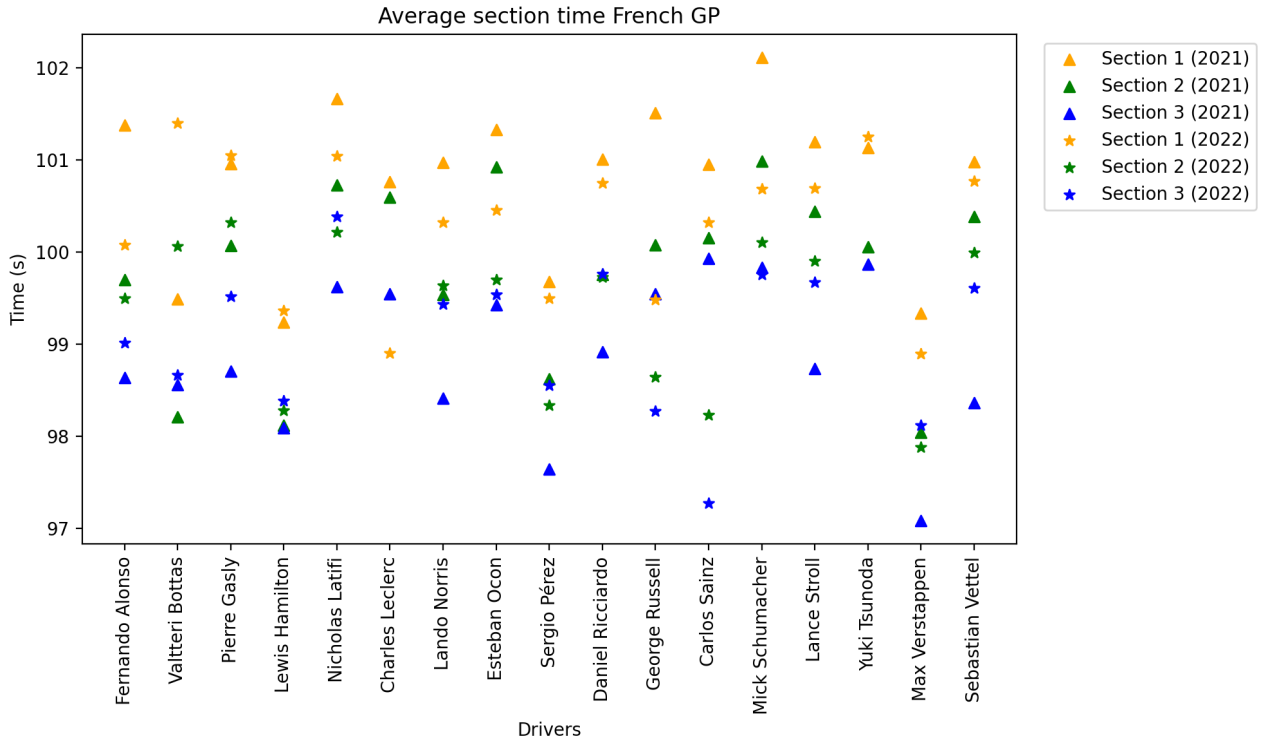


Figure 23: Graph of French GP race times per driver (2021 vs 2022).

From these plots, we can infer that changes in regulations had a notable impact in just a single Grand Prix, resulting in improved race times for the drivers. In five cases, the average times were better in 2021 compared to 2022. On the other hand, in most instances (with mixed trend), the overlapping of times from different sections makes it difficult to definitively define an overall trend. In those cases it is observed that generally, the trend

of average times in a specific section remains relatively consistent in the subsequent year as well. This could be due for example to strategies employed by individual drivers or the characteristics of the circuit. From the analysis of the graph in fig. 11, we can observe that the only Grand Prix where there were improvements in lap times from 2021 to 2022 was the one held in Saudi Arabia. This Grand Prix takes place on a circuit with the highest number of turns among all the circuits. However, we cannot easily hypothesize a direct correlation between the changes made to the Formula 1 regulations between 2021 and 2022 and the improved lap times achieved on circuits with a high number of turns.
Indeed, the Azerbaijan and Abu Dhabi Grands Prix seem to contradict this hypothesis. Despite these circuits having a considerable number of turns, there was a worsening of race times during these events. This suggests that factors other than the number of turns may influence the performances of teams and drivers. The Bahrain, Brazilian, Spanish, and Mexico City Grands Prix, which have a relatively low number of turns (15) and a circuit length not exceeding 5.5 km, belong to the category of circuits that experienced a decline from 2021 to 2022. However, there are also circuits with a considerable number of turns, such as Abu Dhabi (20 corners), Azerbaijan (19), and British (18), among the longest in this analysis, for which there was a worsening in lap times during races. Due to all these reasons, it is not possible to establish a clear relationship between the circuit length and turns and the worsening of race times between 2021 and 2022. Another proof of that is the Austrian Grand Prix, which has the lowest number of turns and is also among those with shorter lengths, falls into this category.

Second Focus - Average race times across sections for teams:

In this case, the graphs provide a more general overview of the average race times for each racing team. For each section, the average times are calculated by taking a comprehensive yearly average (without distinguishing between individual Grand Prix races) across all the times associated with a team. Firstly, with specific reference to the years 2013 and 2014, we notice from fig 24 a general trend of the data clustering into sections (excluding Sauber, Toro Rosso and the Marussia team where 2014 Section 2 is closer to Section 3 of the same year than it is to 2013 Section 2).

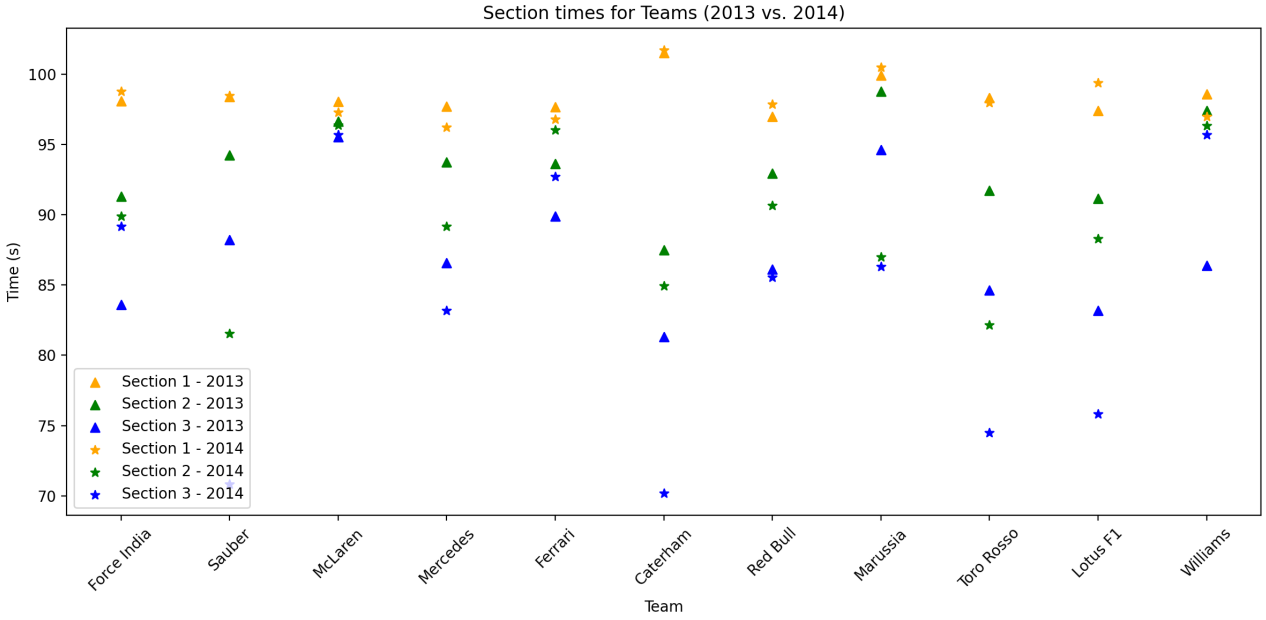


Figure 24: Graph of times per Team (2013 vs 2014)

Overall, we do not observe a clear improvement or worsening in the times associated with the teams from 2013 to 2014 (except for Mercedes which experiences an improvement overall section times). This behavior is not surprising, considering that, based on the collected data related to average driver times, we have seen a "mixed"

behavior in the majority of Grand Prix races. Nonetheless what we do observe is that, in general, the lower times are associated with the third sections of both years, the higher times with the first sections of both years, and the intermediate times with the second sections. This could be linked to the teams' race strategies as well as the specific characteristics of the circuits.

Concerning 2021 and 2022 we can point out from fig 25 that the majority of teams (namely 7 out of 10) experienced a significant decline from 2021 to 2022 in terms of average time per section.

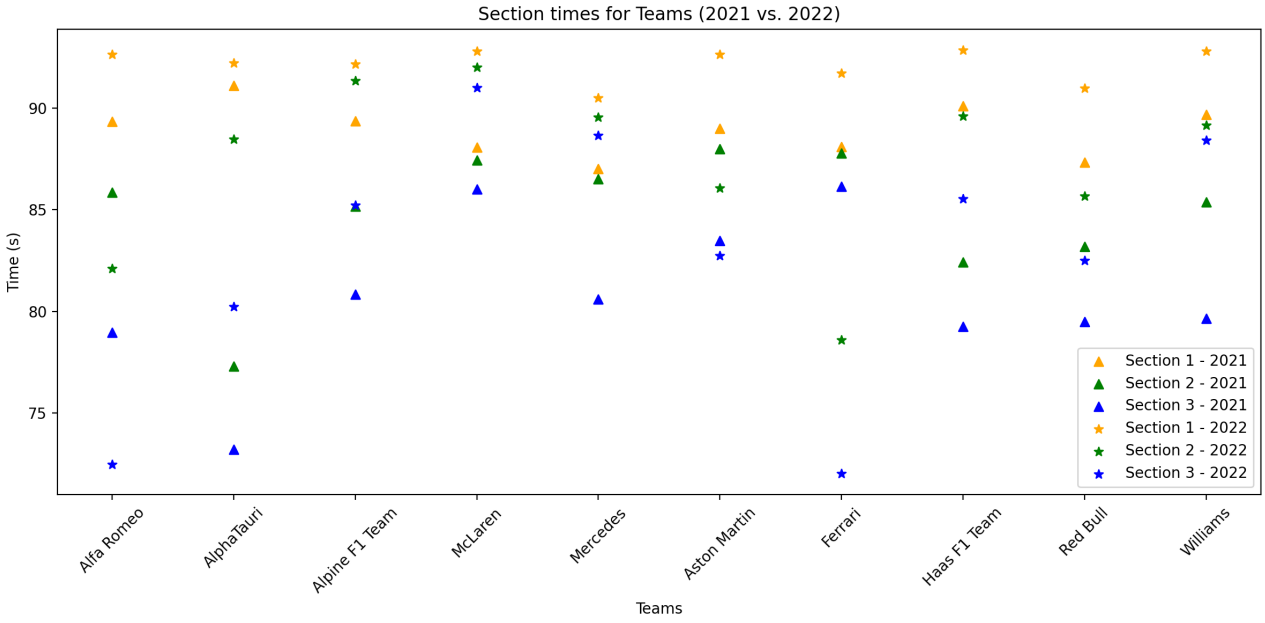


Figure 25: Graph of times per Team (2021 vs 2022).

More specifically, these teams are: McLaren (with a gap of 3/4 seconds from a section to the corresponding one in the following year), Mercedes (with a few seconds' gap between the first and second sections, and an almost 7-second gap between the third section and its corresponding one in the following year), Red Bull and Alpine F1 team (with a gap of 3/4 seconds from a section to the corresponding one in the following year), Alpha Tauri (with a gap of nearly 1 second from section 1 (2013) to section 1 (2014) and then wider gaps of about 7/8 seconds on the other sections),Williams (with a gap of nearly 3 second from section 1 (2013) to section 1 (2014) and then wider gaps of 8 seconds on the other sections), Haas F1 Team (with a gap of 3 seconds from section 1 (2013) to section 1 (2014), 5 seconds from section 2/3 (2013) to section 2/3 (2014)).On the other hand, Alfa Romeo, Ferrari and Aston Martin experienced an improvement in times for the second and third sections, from 2021 to 2022.

In summary, has there been a shift in the trend of the average lap times?

It appears that the modifications implemented between 2013 and 2014, as well as those between 2021 and 2022, generally did not lead to enhancements in car performance (looking both at the Teams and drivers race times). Specifically, while the transition from 2013 to 2014 does not exhibit a clear improvement or decline in drivers' race times (due to a significant number of Grand Prix lacking a distinct and defined trend), the same cannot be said for the 2021-2022 period. In the latter case, it can be stated that there was no enhancement in drivers' performances; rather, a noticeable decline occurred in a substantial number of Grand Prix races. In general, we arrive at the same conclusions as the qualifying times, whereby we must consider that the years 2014 and 2021 marked the first instances when significant new changes were implemented and also the maiden occasion for drivers to test the cars with these updates on each track, right from the beginning of the Grand Prix. Given this context, it should not be surprising if the initial attempt with these fresh changes didn't yield better results. On the contrary, comparing the graphs in Figures 24 and 25, we observe a decrease in the average values for each section from the 2013-2014 to 2021-2022 period. Specifically, for the first section, the overall average value drops

from about 98 to 91 seconds. For the second section, there is a decrease from 92 to approximately 86 seconds of average values. As for the third section, a decline is noticeable from the previous average of 86 seconds to 81 seconds. Overall, there has indeed been an improvement in the race results (as well as in the average qualifying times, as previously mentioned) but after several years since the introduction of the new regulations.

8.3 Maximum speed

As part of our ongoing research, we are now exploring the analysis of our third query, which focuses on the Maximum Speed data. This query builds upon the methodology we established in our previous two analyses. Our main goal remains the same: to conduct a comparative investigation, initially between the periods 2013-2014 and subsequently between 2021-2022 in order to gain insights into the implications of the alterations made through these years.
Just as we did in our previous analyses, we continue to treat each set of years—2013 and 2014, as well as 2021 and 2022—as distinct entities. The analytical procedures that were meticulously applied to the earlier years are replicated in the later years, maintaining consistency in our approach.
In line with our research objectives, we have divided this third query into two focused sections, mirroring the structure of our previous analyses:

- **First Focus - Maximum Speed Comparison in each Grand Prix for each Driver:** Our primary step was to individually examine each Grand Prix in isolation. For every Grand Prix, we presented a breakdown of maximum speeds achieved by each driver. The purpose was to identify discernible trends or patterns that might have emerged across the different seasons (2013 vs 2014 and then 2021 vs 2022).
- **Second Focus - Overall Average Maximum Speed Comparison for each Team:** This time, the scope shifts from individual drivers to the collective performance of teams. Through visual representations, we offer a comprehensive view of the average maximum speeds achieved by each team throughout all the races.

We developed a code that enabled us to display the desired results and analysis. Naturally, this code was executed on our conclusive dataset and subsequently stored within MongoDB. To enhance the clarity of both explanation and comprehension, we divided the code into four distinct sections.

Part I: Data Processing and Filtering

1. The code connects to the databases stored in MongoDB system, in this way it has access to the four collections containing the databases relative to each year.
2. After having established a connection with the databases, for each of them the code iterates over each driver and GP. In each GP it searches the "Qualifying" field and if it found it controls if the value correspondent to the "Weather Condition" field contained in the "Qualifying" section is equal to "sun". This passage eliminates all the GP in which the weather condition was no good. Once again, this choice was made because unfavorable weather conditions hindered the possibility of conducting a meaningful comparison. Rain or a wet track significantly impact speed and car performances. Consequently, we opted to exclude Grand Prix races with adverse weather conditions, as they would not provide us with reliable and useful results.
3. If the good weather condition is found, the code populates the speed dictionary with maximum speeds for each driver and Grand Prix while Car details for each driver are stored in the car dictionary (in which the key is the driver name).

Part II: Comparing Maximum Speeds of drivers:

1. The code first identifies common Grand Prix (GP) races that occurred in both years (in the dictionary $gp\_drivers\_speed$ and in $gp\_drivers\_speed1$ if the years are 2013 and 2014).For each common GP, the code extracts the list of drivers who participated in both years.

2. Then, for each common GP, scatter plots are generated to compare the speeds of common drivers in the two years. The X-axis represents the common drivers, the Y-axis represents their speeds, and two scatter plots (one for each year) are overlaid on the same graph. Dotted lines connect the speeds of the same driver in the two years, providing a visual comparison.

The code involves multiple sections, each dedicated to a specific year's data. For example, after accessing the collection for the year 2013, the code performs data extraction, analysis, and plotting specific to that year. This is then repeated for years 2014, 2021, and 2022.

Part III: Comparing Average Speeds of Teams

The code calculates average speeds for each team based on driver speeds and then creates scatter plots to compare these averages across different years.

1. First, it constructs a dictionary (*velocita_per_macchina*) where each car is associated with a list of speeds of the drivers using that car.

2. It iterates through the dictionary of driver speeds (speed) for a specific year, extracting the car associated with each driver. If available, it adds the driver's speeds to the list associated with that car.

3. Then, it calculates the average speed for each car based on the collected driver speeds and stores these averages in a separate dictionary (*media_velocita_per_macchina*).

4. Finally, it compares average speeds across different years for cars that are common between the two years, creating scatter plots that overlay average speeds for each car in the given years.

First Focus - Maximum Speed for single drivers:

We obtained that during eleven GPs in both years 2013 and 2014 there was good weather: Canadian, Spanish, United States, Singapore, Monaco, Bahrain, Brazilian, German, Abu Dhabi, Japanese and Italian. All the eleven graphs of the GP maximum speeds obtained for year 2013 and 2014 show just one behavior: 2013 maximum speeds lower than 2014 maximum speeds for all the drivers. As an example, we show the German GP in Figure 26 or the Spanish GP graph in Figure 27.

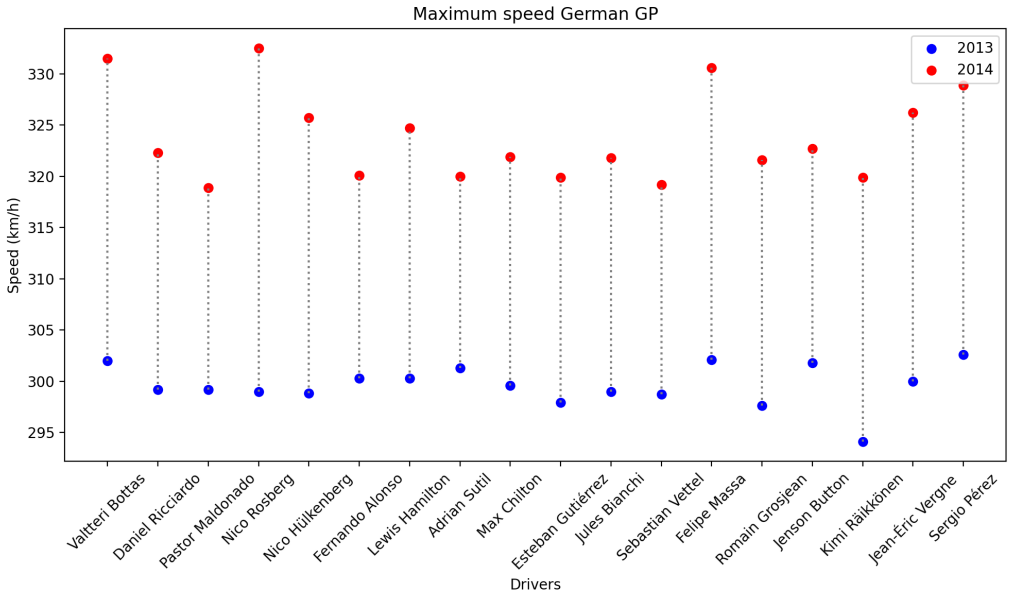


Figure 26: Graph of Maximum Speed in German GP, per driver (2013 vs 2014)

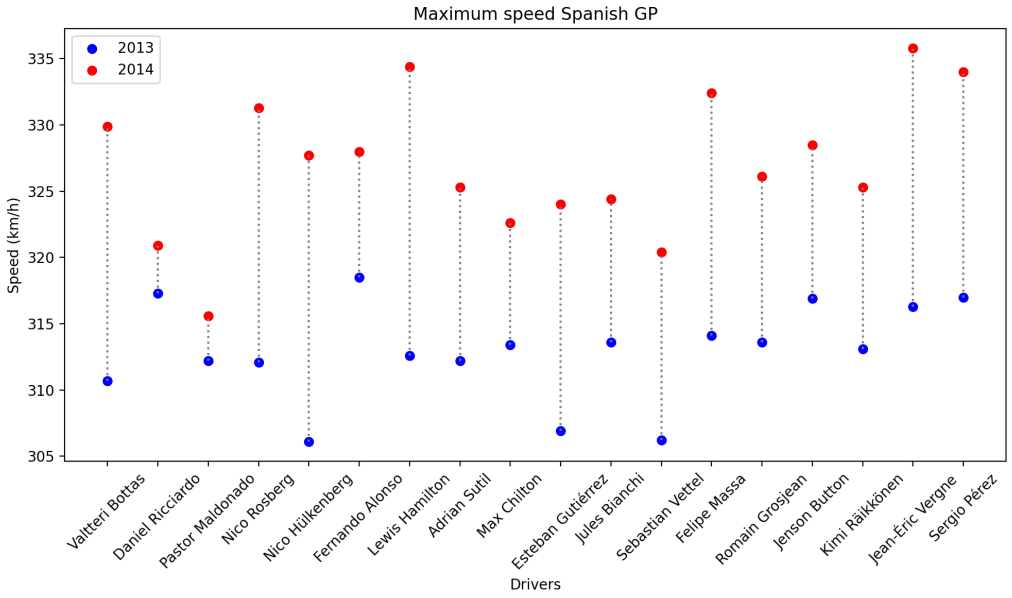


Figure 27: Graph of Maximum Speed in Spanish GP, per driver (2013 vs 2014)

Considering now years 2021 and 2022, there are two types of behavior of data in different GPs. The first one can be seen in Figure 28, in which the data relative to 2021 and 2022 are clearly separated and, in particular, the speeds recorded in 2022 are higher than the ones recorded in 2022, so, the cars reached a faster speed during 2022. Just two GPs of the eleven with good weather condition in both years have this characteristic: Dutch and French GP.

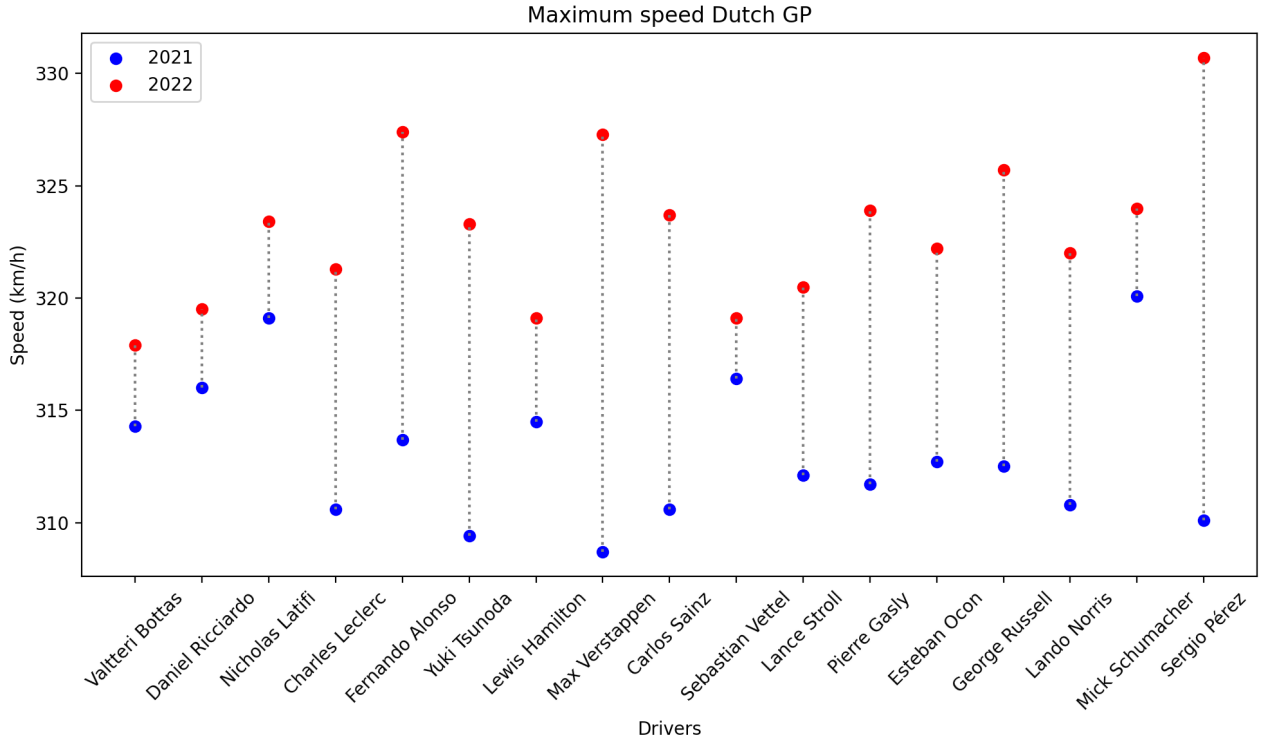


Figure 28: Graph of Maximum Speed in Dutch GP, per driver (2021 vs 2022)

The other nine GPs (Saudi Arabian, Hungarian, Monaco, Bahrain, Azerbaijan, Austrian, Abu Dhabi, Spanish and Italian) don't show this clear separation among 2021 and 2022 speeds data, in fact, in the same GP, for some drivers, 2021 speeds are higher than 2022 speeds and for the other drivers the opposite happens. In Figure 29 we can see an example of one of the nine GP in which the speed data have this behaviour.

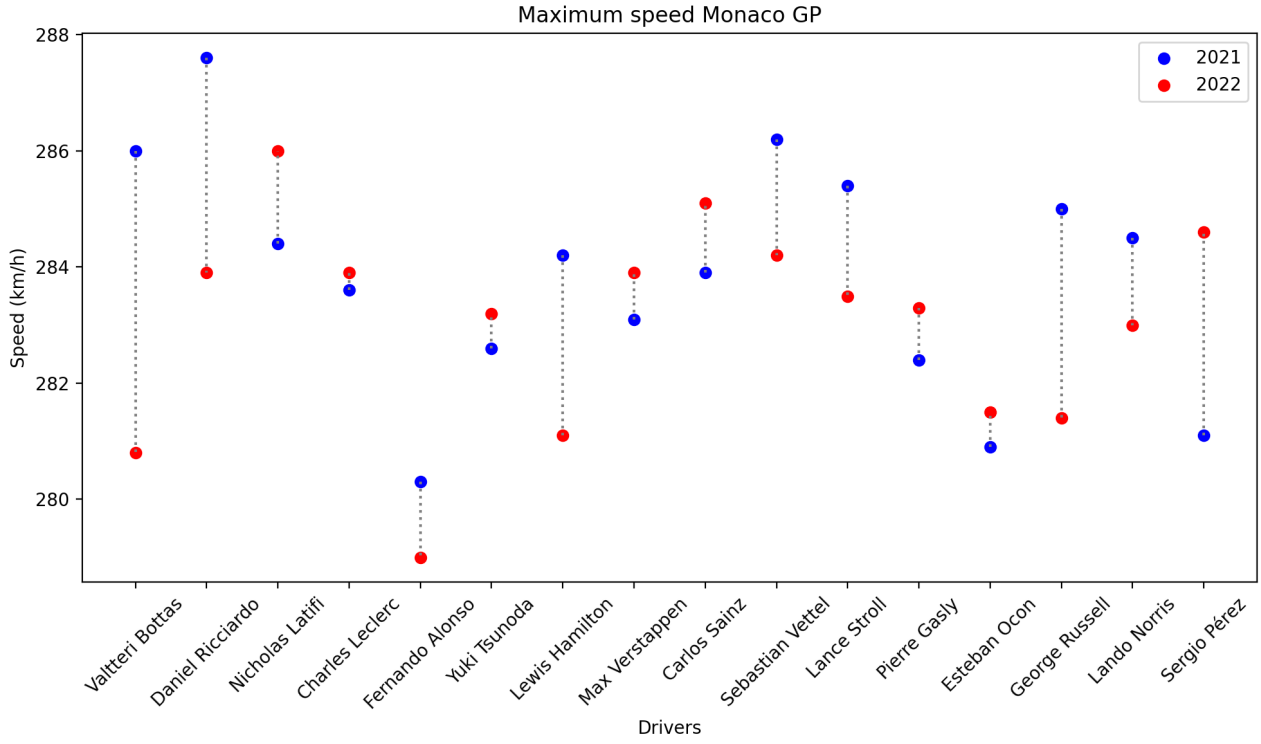


Figure 29: Graph of Maximum Speed in Monaco GP , per driver (2021 vs 2022).

Second Focus - Overall Maximum Speeds for each Team

Considering now a global view of 2013 and 2014 speeds data (for each Team) we should not be surprised by the fact that, as in each single GP for each single driver, for each single Car, on mean, 2013 speeds are lower values than 2014 speeds. We can see this fact in Figure 30 in which the separation between the two years data is clear.

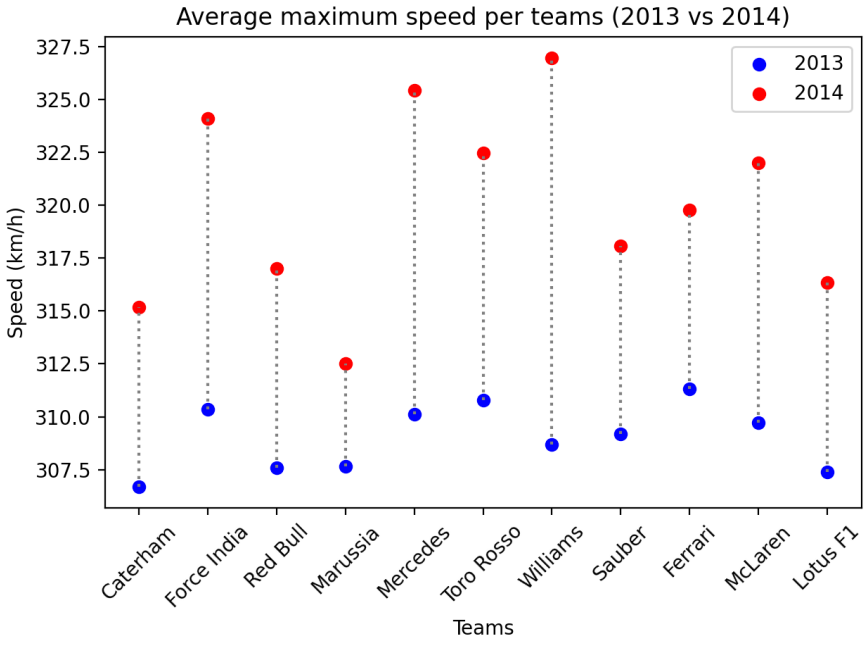


Figure 30: Graph of Maximum Speed per Team (2013 vs 2014).

Considering just for a second the results obtained with the Qualifying data in the previous paragraph this behaviour could seem strange because the performance of the qualifying times is better in 2013 than in 2014, but the performance of speed is better in 2014 than in 2013. We have to underline that this data are different and not necessarily linked to each other.
The maximum qualifying speed data are recorded only in a particular section of the circuit (and they could not be the maximum speeds reached during the session)that is a straight line and even if they have improved with 2014 it doesn't mean that the entire qualifying time, that is recorded over the entire circuit with the turns and the other parts in addition to the straight, line was lower and so that the driver overall went faster during the qualifying session. This fact can suggest us that maybe the change in the engines used could improve the speed that can be reached during a race or a qualifying or free practice session, but does not help in reducing the times recorded during a lap.

For years 2021 and 2022 we have a mix of higher 2021 speeds than 2022 speeds and higher 2022 speeds than 2021 speeds. As we can see from Figure 31 Red Bull, Alpine, Williams, Ferrari and Alpha Tauri reached, on mean, a higher speed during 2022, while Haas, Mercedes, Aston Martin, Alfa Romeo and McLaren have a higher speed data for year 2021. This should not surprised us due to the fact that, as we have seen before, the majority of the GPs for these two years showed a mixed behaviour of speeds data.

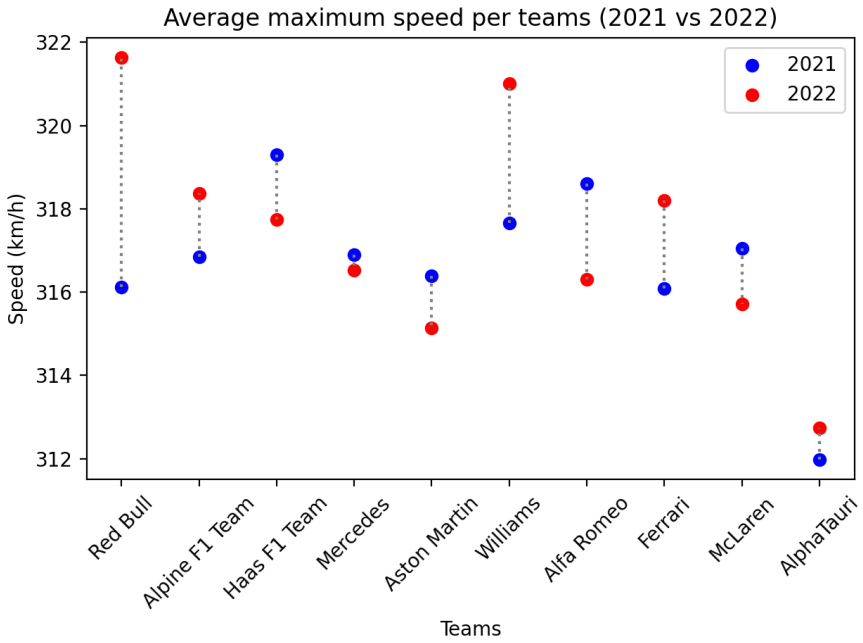


Figure 31: Graph of Maximum Speed per Team (2021 vs 2022).

This time we can think that the news introduced in year 2022 are not directly related with the possible speed data improvement and that only some of the Teams were able to exploit the changes in order to reach a higher speed during the straight line in each Qualifying GP session.

Considering at the same time Figure 30 and Figure 31 we can see that the highest maximum speeds reached on mean are the ones relative to year 2014 while the lowest are relative to year 2013, the values relative to 2021 and 2022 are included between these two limits.

In conclusion, in the analysis results, several key findings emerged:

For the years 2013-2014, there is a consistent pattern of higher maximum speeds in 2014 compared to 2013, both at the individual driver level and the Team level. This indicates an improvement in speed over these

two years. In contrast, for the years 2021-2022, the speed data shows a mixed behavior. Some Grand Prix races exhibit higher speeds in 2022, while others have higher speeds in 2021. This suggests that the changes introduced in 2022 did not universally translate to increased speeds, and some Teams adapted better than others.

From Figure 10 and the data on maximum speeds recorded during qualifying sessions, the following significant conclusions can be drawn :

In the case of the Italian Grand Prix, which takes place on a circuit with the absolute minimum number of turns, drivers show exceptional performances compared to all other circuits. If we look a the graphs of the maximum speed per driver, during the 2013-2014 period, the speed range (from 325 to 355 km/h) during qualifying sessions is notably shifted upwards compared to other circuits (see table 32).

2013-2014	~ Max speed (km/h)	~ Min Speed (km/h)	Max-Min Speed (km/h)
Italian GP	355	325	30
Bahrain GP	332	305	27
United States GP	338	305	33
Japanese GP	316	293	23
German GP	333	294	39
Singapore GP	309	285	24
Spanish GP	336	305	31
Abu Dhabi GP	343	309	34
2021-2022	~ Max Speed	~ Min Speed	
Bahrain GP	324	310	14
Italian GP	348	332	16
Austrian GP	321	312	9
Saudi Arabian GP	321	311	10
Hungarian GP	312	296	16
Spanish GP	338	310	28
Abu Dhabi GP	333	315	18
Azerbaijan GP	325	305	20
Monaco GP	288	279	9
Dutch GP	332	309	23
French GP	340	320	20

Figure 32

This suggests that drivers achieve higher maximum speeds during qualifying in Italy than at any other location. A similar finding emerges for the 2021-2022 period (see always fig 32). During these seasons, drivers continue to record high maximum speeds in qualifying sessions at the Italian Grand Prix. Specifically, the highest minimum speed among the maximum speeds recorded at all circuits turn out to be the highest at the Italian Grand Prix. At the same time, the absolute maximum speed achieved in these qualifying sessions (348 km/h) also belongs to the Italian Grand Prix. These data clearly indicate that the Italian circuit has a significant impact on driver performances, resulting in extraordinary maximum speeds during qualifying.

Now, all the statements made subsequently are relative to the data presented in table 32. It is interesting to notice that one of the longest circuits, where the German Grand Prix took place, show the most significant improvement in terms of speed between 2013 and 2014. The difference between the lowest maximum speed recorded in 2013 and the highest speed achieved by the drivers during the qualifying sessions in 2014 is approximately 40 km/h for the German Grand Prix, while for all other circuits, it is about 10 km/h or less.
When we compare Grand Prix graphs about the period between 2021 and 2022, we notice that the difference between the maximum speed and the minimum speed is on average much more limited compared to what is observed between 2013 and 2014. While in the 2013-2014 period we see speed variations in Grand Prix ranging from 23 km/h to 40 km/h, in the 2021-2022 period, these variations are much more contained, ranging from 9 to 28 km/h.

This observation could, albeit purely hypothetically, suggest that regulatory and technological changes between 2013 and 2014 had a significant impact on performance, resulting in an increase in speeds for all drivers.
Another noteworthy observation is that the maximum speeds reached during the Monaco Grand Prix in 2021 and 2022 exhibit both one of the smallest variations in terms of maximum speed among all the Grand Prix races and the lowest absolute values in terms of maximum speeds. This could be attributed to the fact that the Monaco circuit features a high number of turns and is significantly shorter than the other circuits (as we can see through figure 11). Similarly, the Austrian Grand Prix also exhibits a limited variation in terms of speed (9 km/h, just like the Monaco Grand Prix). The circuit is characterized by a lower number of turns compared to the circuits of the 2021 and 2022 seasons but it shows one of the shortest track lengths among all the circuits of the two seasons.

9 CONCLUSION

In our analysis of Formula 1 data from the years 2013, 2014, 2021, and 2022, we have gathered a comprehensive dataset for comparative insights.
For each of the specified years, we meticulously collected the following data:

- **Drivers** & **Teams**: we collected data on the drivers and the Teams they were associated with, in each Formula 1 season.
- **Grand Prix**: we documented details about each Grand Prix race. This information is essential for contextualizing race results.
- **Laps** & **Qualifications** : we acquired the number of laps completed in each race, providing insights into race duration and driver performance. We gathered detailed information on the Formula 1 qualifying sessions, covering three phases: Q1, Q2, and Q3. These sessions play a crucial role in determining a driver's performance enhancement or deterioration throughout the specified Formula 1 seasons.
- **Maximum Speeds During Qualifying Sessions**: we recorded the highest speeds achieved by drivers during the qualifying sessions. This data offers insights into the aerodynamic efficiency and power of the cars.
- **Circuit with Turns** & **Length**: we gained data about the racing circuits themselves, including their names and lengths and we took the number of turns at each racing circuit. Understanding the circuit layout is critical for comprehending the technical challenges faced by drivers and Teams.
- **Weather Conditions**: we meticulously documented the weather conditions during qualifying sessions and races because weather can significantly impact race outcomes and strategy.

During the data collection phase, we primarily obtained data through web scraping, with the exception of the PDF files containing maximum speeds, which were downloaded and converted to JSON format.
To verify the completeness of the acquired dataset we conducted an in-depth analysis of missing data.
We assessed data consistency by examining the update status of websites, their official sources (as in the case of the fia.com for speed data), and cross-referencing multiple sources (as in the case of weather conditions gathered from both Italian and English Wikipedia pages). Data accuracy was evaluated using analytical tool, the percentage of inaccuracies, and through semantic considerations. Additionally, we conducted semantic considerations, which included evaluating the reliability of data sources. In the case of weather conditions, our assessment was based on words or sentences found within the texts from which we extracted the data, and this evaluation was conducted across multiple languages to ensure consistency and accuracy.
Subsequently, we integrated all the collected data from the specified years (2013, 2014, 2021, and 2022) into a consolidated dataset. This dataset is stored into JSON files named "Final_Dataset*year*.json". Once the dataset was completed, we performed three queries on the data, to show three analyses that it is possible to perform with them:

1. **Query 1**: In our analysis, we embarked on a comparison of Qualifying times to address research questions regarding the impact of changes made in Formula 1 racing between different seasons. While the initial transition to new regulations in 2014 and 2021 did not immediately lead to better results, the data suggests that the sport experienced significant improvements in lap times over a more extended period (from 2013 upto 2021).

2. **Query 2**: In our second query, we shifted our focus to average lap times, considering the data from three distinct sections of Formula 1 races. We aimed to identify trends and changes in lap times between 2013 and 2014 and then between 2021 and 2022. This analysis paralleled our observations regarding Qualifying times.

3. **Query 3**: In this phase of our research, we delved into the analysis of Maximum Speed data in qualifications in Formula 1. Our analysis revealed a clear improvement in maximum speeds from 2013 to 2014, while the effects of changes in 2022 on speed were more variable and team-dependent. The highest average maximum speeds were observed in 2014, while the lowest were in 2013, with values for 2021 and 2022 falling in between these extremes. Moreover, it is important to note that maximum speed data during qualifying sessions represent only a specific section of the circuit (usually a straight line) and may not necessarily correlate with overall lap time improvements or deterioration.

In this particular case, our objective was to analyze the progression of driver and team performances over a span of four years: 2013, 2014, 2021, and 2022. However, it's important to acknowledge that there is room for enhancing this analysis to yield more comprehensive and detailed insights. In fact, for further analysis we can consider different data and strategies to obtain good results, for example:

1. To gain a comprehensive perspective on the evolution of time and speed data in Formula 1, we might extend our analysis to encompass a more extensive timeframe. This could involve examining data from a span of 10, 15, or even 20 years, or possibly from the inaugural year of the championship in 1950 to the most recent year for which we have data. By doing so, we would be able to attain a holistic view of how these metrics have evolved over a more extensive historical period, providing valuable insights into the trends and changes that have occurred across multiple decades of Formula 1 racing.

2. In addition to the data we have already collected, we might explore additional factors that could influence race outcomes. For instance, we could examine data related to tyre changes during a race or any modifications made to the car. These additional insights could help us better understand how such variables impact race results.

3. We could explore alternative analyses to gain insights into different aspects of racing. For instance, we might investigate the trends in lap times when drivers make pit stops. This could reveal whether there have been improvements or changes in this specific phase of the race over the years.

4. We could conduct a more in-depth analysis of driver and team performances to explore potential relationships. For instance, we might investigate whether belonging to a specific team consistently leads to better results, regardless of the driver. Alternatively, we could examine whether certain drivers consistently perform well regardless of the team they are associated with.

These are just some of the examples of different analyses we could make in order to improve our results. Formula 1 is a very big world with many data useful to make analysis. Starting with the dataset we created it is possible to make further enrichment in order to improve the quality of the data itself and of potential analyses.

10 ROLE OF GROUP PARTICIPANTS

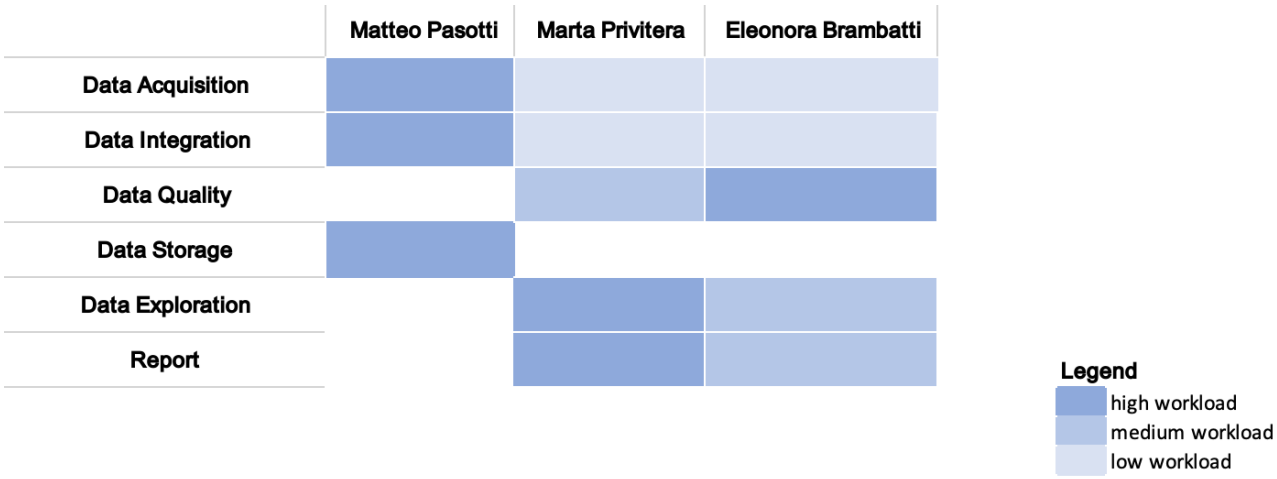


Figure 33: All decisions concerning data and their management were thoroughly discussed among all members of the group, and every effort was made to reach a consensus among all parties involved.

References

[1] FIA. "Article 5.1". In: *2014 FORMULA ONE TECHNICAL REGULATIONS* (2014).

[2] FIA. "Article 5.1". In: *2013 FORMULA ONE TECHNICAL REGULATIONS* (2013).

[3] FIA. "Article 5.2". In: *2014 FORMULA ONE TECHNICAL REGULATIONS* (2014).

[4] FIA. "Article 5.2". In: *2013 FORMULA ONE TECHNICAL REGULATIONS* (2013).

[5] FIA. "Article 1.20". In: *2013 FORMULA ONE TECHNICAL REGULATIONS* (2013).

[6] FIA. "Article 1.24". In: *2014 FORMULA ONE TECHNICAL REGULATIONS* (2014).

[7] FIA. "Article 5.1.4". In: *2014 FORMULA ONE TECHNICAL REGULATIONS* (2014).

[8] 2023 Mondadori Media S.p.A. "F1 2022 – Cosa cambia nel regolamento". In: (2023).

[9] "Risultati". In: *Gran Premio del Giappone 2013* (Wikipedia).

[10] "Risultati". In: *Gran Premio d'Italia 2013* (Wikipedia).

[11] "Risultati". In: *Gran Premio d'Australia 2013* (Wikipedia).

[12] "Risultati". In: *Gran Premio d'Australia 2013* (Wikipedia).

[13] "2013: Le ultime gare da titolare alla Lotus F1 Team". In: *Heikki Kovalainen* (Wikipedia).

[14] "Risultati". In: *Gran Premio d'Australia 2014* (Wikipedia).

[15] "Risultati". In: *Gran Premio degli Stati Uniti d'America 2014* (Wikipedia).

[16] "Risultati". In: *Gran Premio del Brasile 2014* (Wikipedia).

[17] "Aspetti sportivi". In: *Gran Premio del Belgio 2014* (Wikipedia).

[18] "Risultati". In: *Gran Premio di Monaco 2014* (Wikipedia).

[19] "Risultati". In: *Gran Premio di Germania 2014* (Wikipedia).

[20] "Risultati". In: *Gran Premio di Gran Bretagna 2014* (Wikipedia).

[21] "Risultati". In: *Gran Premio di Abu Dhabi 2014* (Wikipedia).

[22] "Risultati". In: *Gran Premio del Canada 2014* (Wikipedia).

[23] "Risultati". In: *Gran Premio di Russia 2014* (Wikipedia).

[24] "Caterham 2014". In: *Marcus Ericsson* (Wikipedia).

[25] "Risultati". In: *Gran Premio d'Ungheria 2021* (Wikipedia).

[26] "Risultati". In: *Gran Premio dell'Emilia Romagna 2021* (Wikipedia).

[27] "Risultati". In: *Gran Premio di Monaco 2021* (Wikipedia).

[28] "Risultati". In: *Gran Premio del Bahrain 2021* (Wikipedia).

[29] "Risultati". In: *Gran Premio di Abu Dhabi 2021* (Wikipedia).

[30] "Risultati". In: *Gran Premio d'Austria 2021* (Wikipedia).

[31] "Risultati". In: *Gran Premio d'Olanda 2021* (Wikipedia).

[32] "Risultati". In: *Gran Premio d'Italia 2021* (Wikipedia).

[33] "Risultati". In: *Gran Premio di Città del Messico 2021* (Wikipedia).

[34] "Risultati". In: *Gran Premio di Gran Bretagna 2021* (Wikipedia).

[35] "2021". In: *Robert Kubica* (Wikipedia).

[36] "Risultati". In: *Gran Premio di Gran Bretagna 2022* (Wikipedia).

[37] "2022". In: *Alexander Albon* (Wikipedia).

[38] "Risultati". In: *Gran Premio del Giappone 2022* (Wikipedia).

[39] "Risultati". In: *Gran Premio del Belgio 2022* (Wikipedia).

[40] "2022". In: *Nico Hülkenberg* (Wikipedia).

[41] "Risultati". In: *Gran Premio di San Paolo 2022* (Wikipedia).

[42] "Risultati". In: *Gran Premio dell'Emilia Romagna 2022* (Wikipedia).

[43] "Risultati". In: *Gran Premio d'Arabia Saudita 2022* (Wikipedia).

[44] "Risultati". In: *Gran Premio della Malesia 2013* (Wikipedia).

[45] "Risultati". In: *Gran Premio di Cina 2013* (Wikipedia).

[46] "Risultati". In: *Gran Premio del Bahrein 2013* (Wikipedia).

[47] "Risultati". In: *Gran Premio di Monaco 2013* (Wikipedia).

[48] "Q3". In: *2013 German Grand Prix* (Wikipedia).

[49] "Resoconto". In: *Gran Premio d'Ungheria 2013* (Wikipedia).

[50] "Q3". In: *2013 Singapore Grand Prix* (Wikipedia).

[51] "Risultati". In: *Gran Premio d'Australia 2014* (Wikipedia).

[52] "Risultati". In: *Gran Premio di Cina 2014* (Wikipedia).

[53] "Risultati". In: *Gran Premio di Spagna 2014* (Wikipedia).

[54] "Risultati". In: *Gran Premio di Monaco 2014* (Wikipedia).

[55] "Risultati". In: *Gran Premio del Canada 2014* (Wikipedia).

[56] "Risultati". In: *Gran Premio d'Austria 2014* (Wikipedia).

[57] "Qualifying". In: *2014 British Grand Prix* (Wikipedia).

[58] "Risultati". In: *Gran Premio di Germania 2014* (Wikipedia).

[59] "Risultati". In: *Gran Premio d'Ungheria 2014* (Wikipedia).

[60] "Aspetti sportivi". In: *Gran Premio di Russia 2014* (Wikipedia).

[61] "Aspetti sportivi". In: *Gran Premio degli Stati Uniti 2014* (Wikipedia).

[62] "Aspetti sportivi". In: *Gran Premio del Brasile 2014* (Wikipedia).

[63] "Resoconto". In: *Gran Premio di Abu Dhabi 2014* (Wikipedia).

[64] "Qualifying". In: *2021 Emilia Romagna Grand Prix* (Wikipedia).

[65] "Resoconto". In: *Gran Premio di Monaco 2021* (Wikipedia).

[66] "Resoconto". In: *Gran Premio d'Azerbaigian 2021* (Wikipedia).

[67] "Resoconto". In: *Gran Premio di Francia 2021* (Wikipedia).

[68] "Resoconto". In: *Gran Premio d'Ungheria 2021* (Wikipedia).

[69] "Resoconto". In: *Gran Premio del Belgio 2021* (Wikipedia).

[70] "Resoconto". In: *Gran Premio di Russia 2021* (Wikipedia).

[71] "Resoconto". In: *Gran Premio di Turchia 2021* (Wikipedia).

[72] "Resoconto". In: *Gran Premio degli Stati Uniti 2021* (Wikipedia).

[73] "Resoconto". In: *Gran Premio dell'Arabia Saudita 2022* (Wikipedia).

[74] "Resoconto". In: *Gran Premio d'Australia 2022* (Wikipedia).

[75] "Resoconto". In: *Gran Premio dell'Emilia-Romagna 2022* (Wikipedia).

[76] "Resoconto". In: *Gran Premio di Miami 2022* (Wikipedia).

[77] "Resoconto". In: *Gran Premio del Canada 2022* (Wikipedia).

[78] "Resoconto". In: *Gran Premio d'Austria' 2022* (Wikipedia).

[79] "Resoconto". In: *Gran Premio dei Francia 2022* (Wikipedia).

[80] "Resoconto". In: *Gran Premio d'Olanda 2022* (Wikipedia).

[81] "Resoconto". In: *Gran Premio d'Italia 2022* (Wikipedia).

[82] "Resoconto". In: *Gran Premio del Brasile 2022* (Wikipedia).

[83] "Art. 28". In: *2021 FORMULA ONE SPORTING REGULATIONS* (2021).