Wacky Races

João Manuel de Almeida Rodrigues (m20180550@novaims.unl.pt, m20180550); Nuno Jorge Pinto Lopes 20211056@novaims.unl.pt, 20211056); Ricardo Daniel Pacheco de Almeida (m20201335@novaims.unl.pt, m20201335); Soraia Andreia Anselmo Alves (m20201840@novaims.unl.pt, m20201840); Teresa Paula de Sousa Brazão Albuquerque Barroso (T2003041@novaims.unl.pt, T2003041).

**Group.** I **TP.** 3

**Abstract.** Wacky Races is a group project that aims to provide great insights on Formula 1 (F1), to both newcomers and veteran fans of the sport. To do that, data related to the sport was extracted, scraped, and stored both locally and in a cloud-based service. The project required the usage of Python, HTML and T-SQL, and the utilization of multiple modules not learned in class. As a result, multiple visualizations were created, as well as a map with geospatial information for introducing new fans to the last 7 years of this sport.

**Keywords.** F1, data visualization, geospatial visualization, web scrapping

# Introduction

The decision to focus this group work on Formula One (F1) data stemmed not only from a desire to present this motor sport to those unfamiliar with it, but for the opportunity to identify correlations and obtain new insights on what goes on in the news and social media around the F1 global community.

Formula One is the highest class of international racing for open-wheel single-seater formula racing cars, sanctioned by the Fédération Internationale de l'Automobile (FIA), since 1950. These are the fastest regulated road-course racing cars in the world, thanks to their top-of-the-line technology. The average annual cost of running an F1 team – designing, building, and maintaining cars, pay, transport – is approximately US$247 million. It is the group’s belief that this topic connects with data science since it allows for the use of analytical tools and methodologies along with challenging data extraction. 2015 marked the biggest update in F1 cars’ motorization unit, which combines a combustion engine with an electrical component, first proposed in the year before. During this time, Drag Reduction System (DRS) technology was also introduced, having first been deployed in 2011. DRS reduces the drag that the rear wing creates while increasing the speed that a car can achieve on a straight road. Thus, the group decided to focus on the period between 2015 and 2022.

# Data and Methods

Web scrapping was performed on several sites, as well as API accesses. Data is stored in local folders, an Azure database, and on several .csv and .xlsx files described in the ‘README.txt’ document.

Several libraries were used, that allowed for exploring different ways of storytelling with data. Some of those were taught in the classroom: pandas (the data science module to perform manipulations, grouping and transform data), matplotlib (to create plots and customize visualizations), and NLTK with Wordcloud (to do some natural language processing and visualization). Others were discovered elsewhere. As stated by Jake VanderPlas “No matter what type of scientific, numerical, or statistical problem you are facing, it’s likely there is a Python package out there that can help you solve it.” (VanderPlast, Jake, 2016). Packages not mentioned or learned in-depth in classroom include: Folium (to create dynamic maps integrating multiple languages, e.g. HTML, CSS and JavaScript), Pyodbc (to connect to azure SQL DB), Wikipedia (to extract information from Wikipedia), Base64 (to convert files to base64), os (to access files within folders), Urlimage (to infer an image given an url), decimal (to handle float datatypes with more precision for using geospatial coordinates), PIL (from which Image was imported to add image processing capabilities to the Python interpreter) and io (from which BytesIO was imported to provide Python with main facilities for dealing with various types of I/O).

To increase the level of control over the extracted data, and since the website structures change frequently, four different functions were prepared to help creating these operations: ‘Links\_Extraction’, ‘fia\_f1\_data’, ‘fia\_f1\_session’ and ‘f1\_gp\_circuits’. The ‘Links\_Extraction’ function was created to extract all the relevant links from the fia.com website. It has 2 parameters: ‘url’, representing the website page that is to be scraped, (i.e. <https://www.fia.com/f1-archives?season=1108>), and ‘url\_string’, that takes a string contained in the ‘url’, to filter extraction results.

The ‘f1\_gp\_circuits’ function was created to extract the Grand Prix’s (GP) dates and names from the espn.com website (i.e., https://www.espn.com/f1/schedule/\_/year/2022).

The ‘fia\_f1\_data’ and ‘fia\_f1\_session’ functions both use the same principle. They receive a season, and the related lists ‘gp\_city’ and ‘gp\_links’ that were previously prepared to collect the needed data. ‘fia\_f1\_data’ was extracted from the race classification page. It includes data on teams, drivers, fastest laps, best sector times, speed traps, maximum speeds and pit stops for each GP. ‘fia\_f1\_season’ extracts data from the season classification page, including initial qualification and grid for each GP. Both functions use BeautifulSoup and request packages to extract information from tables available in each web page. That information is selected and renamed appropriately for each GP link. The result is a data set that holds all the data mentioned above per season, per GP. A connection was set from our python notebooks to an Azure SQL database to retrieve extra information from <http://ergast.com/mrd/db/#csv> (database schema available on folder “1. Data Preparation”, please refer to the ‘README.txt’ file). MySQL (5.7) F1-related data dumps were accessed and converted it into T-SQL to store on a relational Azure DB (for access, refer to the ‘README.txt’ file). API functionality was also used to ensure access to the most recent data and for obtaining circuits’ locations.

# Results and Discussion

The group’s first goal was to provide some global statistics about F1, for the period between 2015 and 2021, which represents the hybrid era. The top 3 teams with more points (Figure 1), the top 3 drivers with more wins (Figure 2), and the top 3 fastest tracks, per season (Figure 3), were:

Figure - Top Constructors per Season

A screenshot of a computer

Description automatically generated with medium confidence

Figure - Top Driver per Season

Graphical user interface

Description automatically generated

Figure - Top Fastest Tracks per Season

Graphical user interface

Description automatically generated with medium confidence

Next, using Python’s natural processing language abilities, the group wanted to obtain unbiased data-driven perspective on Nikita Mazepin’s popularity as a disastrous racer and meme persona, due to his concurrent accidents, spins and collisions. To do that, the group drew a plot using the ratio between the number of races attended per each driver and the number of Did Not Finish (DNF). Results prove that F1 fans are right: Nikita Mazepin has a 90% ratio of DNF’s in his 1 year long F1 career (Figure 4).

Figure - DNF Ratio per Driver

Chart, sunburst chart

Description automatically generated

To understand how unpredictable F1 raced can be, a correlation was plotted between the pole position – driver starting position – and the final position, in each race. The idea stemmed from F1 driver Valtteri Bottas, who recently won a record for being the driver with most pole positions without a championship title (Figure 5).

Figure - Pole Position vs Final Race Position

Shape, arrow

Description automatically generated

The F1 community is very large and active in social media. In 2019, F1 races were broadcasted to an audience of 1.9 billion people worldwide, a number still increasing. Text mining techniques were applied on a Kaggle dataset with a collection of more than 500,000 F1 tweets from 2021, from which 2000 were used due to performance concerns. With the NLTK package and a Wordcloud visualization, the most frequent words in the world of F1 were identified. Sebastian Vettel, an ex-world champion, got the most mentions. Finally, taking advantage of the data available on F1 circuits, drivers, and constructors, and explore a new avenue for visualizing data, a Folium map was developed. It was used to display information on a global map using geospatial coordinates, personalized map markers, and a simple HTML popup with more information and images, to introduce those unfamiliar with the sport.

# Conclusions

The main goals were achieved, based on the data collected. The main challenge was web scrapping, because, even though there is a plenitude of F1 data, it is hard to select the most important features right away. Furthermore, the loads of data are huge, which required the definition of a reduced scope to guarantee the work was completed until its due date.

The next steps would include collecting more data and from more seasons, to be able to create more correlations and create new problematics: e.g. understand the impact of good and bad weather in the performance of each driver/team/track; understand the impact of fast and slow pit stops in each team’s classification; compare performances before and after the introduction of DRS. Finally, this work’s follow-up would be to apply a supervised machine learning algorithm, to try to predict this season’s winner.

# Statement of contribution & Acknowledgments

Overall, the entire work was very well distributed among the different elements, however each of us was mostly in charge of specific tasks, to increase the efficiency of the project.

Nuno and João oversaw the data collection, using web scrapping, Azure and SQL databases, as well as the data cleaning. Ricardo, Teresa and Soraia contributed with the analysis and visualizations. The presentation materials and report deliverables were created with the support of all the group members.

# References

Ergast developer API. (2022). Retrieved May 22, 2022, from <http://ergast.com/mrd/>;

F1 - the official home of Formula 1® racing. Formula 1® - The Official F1® Website. (2022). Retrieved May 22, 2022, from <https://www.formula1.com/>;

Federation Internationale de l'Automobile. (2022). Retrieved May 22, 2022, from <https://www.fia.com/>;

ESPN Internet Ventures. (2022). Formula 1 news, live grand prix updates, videos, drivers and results. ESPN. Retrieved May 22, 2022, from <https://www.espn.com/f1/smarturl.it/gab-juls-show>;

Wayback Machine. (2022). Retrieved May 22, 2022, from <https://web.archive.org/>;

VanderPlas, Jake. “A Whirlwind Tour of Python.” *O’Reilly Media*, 7 Oct. 2016, https://www.oreilly.com/content/a-whirlwind-twour-of-python7.