Formula 1 Analytics and Prediction

Data Mining CSCI 5002

Rohan Sai Nalla
Data Science, University of Colorado Boulder, rohan.nalla@colorado.edu
Kaushik Narasimha Bukkapatnam
Data Science, University of Colorado Boulder, kabu9868@colorado.edu
Sriram Arabelli
Data Science, University of Colorado Boulder, srar4232@colorado.edu
Sashank Gangadharabhotla
Data Science, University of Colorado Boulder, saga8160@colorado.edu

Abstract

“In the current epoch, the influx of data has influenced the processes in every domain that exists. With the introduction of data and analytics, every domain has transformed the way it works by integrating the data into its process pipeline. A major shift in paradigm is observed in sports as well. Formula 1 which has been the top racing sport event in terms of viewership has dramatically transformed the way the teams perform. The current era in formula 1 is called the Turbo Hybrid Era, owing to its new sustainable engines but there's also a significant addition to the cars that belong to this era, which is the numerous sensors added to the car that transmit millions of data points to be analyzed to the paddocks at every moment of time. This enormous data transformed the way races undergo, in a quantified sense, racing is only 40% effort whereas the other 60 is attributed to the calculated strategies based on the collected data. In this project, we leverage the data that is available to us through Eargast API and intend to develop a real-time analytical dashboard that is capable of summaries any event in the current timeline. Additionally, we intend to focus a significant portion of time on developing a Machine Learning model that can predict lap times based on the circuit that is selected. The analytics is segregated into two aspects one is for a Fan's perspective and one's for a potential investor who's looking out for historic information about a particular team

to place strategic investments. The same two perspective approach will be applicable to the dashboard. Finally, the dashboard will be cloud-hosted and openly available to anyone who's willing to understand and study a race."

INTRODUCTION

Formula 1 has been the pinnacle of Motor Sport from its inception to the current Turbo-Hybrid era. It is a sport where ten teams compete with 20 drivers, with the most sophisticated machinery. Every minute detail has the most value when it comes to F1. Races vary with microsecond differences, and it all depends on the synergy between all the parts that make that machinery work and the driver's ability that integrates with that machinery. Congregation of all these technical and psychological elements works concurrently to helm immense delight for the spectators. F1, in recent years, has gained much traction, pulling more viewers yearly. To quantify, the current year averages over 1.4 Million viewers per race. For 2022, the viewership has increased by over 49% compared to the previous year and a 131% increase from the year before.

A lot has evolved over the years from a data perspective. Data collection has become streamlined and temporal, with a multitude of sensors constantly collecting and transmitting data on each car. According to Amazon Web Services, each F1 car has over 300 sensors transmitting approximately 1.1 Million telemetry data points per second from the cars to the pit walls. In the contemporary era, Formula 1 is a balancing act of both Data-Driven Strategy making and the Skill of Driving. If one fails, the other suffers. On the other side of the coin, apart from the technical aspect of F1, team economy is an essential parameter for overall performance. Rationally, adequate capital investment can lead to better results for a particular team. Given that investors stake their money on outcomes and team performance, collecting performance data may help both teams and investors regain and create confidence.

The primary objective of this project is to analyze the intricate details of the sport both in a temporal and quantified sense. The chief result of this project is to provide an analytical summary of a team and a driver based on their performance in previous and current races and mine data about their efficiency, skill set, errors, and temporal visualization of their performance. This information could deem valuable for a potential investor to finance a team or a driver and foresee returns with high probability. Apart from this aspect of the project, considering the vast and versatility of the data, we also plan to use Statistical Machine Learning, Deep Learning models and Time Series to predict certain aspects of the race. Upon studying the correlation and feature importance, it is feasible to predict Race Positions, Tier Degradation, and Race Strategies to be used. Though in Formula 1, every parameter is disconnected from a previous race and has no pattern, we plan to mine information logically while sticking closely to the science of F1 itself.

2 RELATED WORK

Considering the prevalence of Formula 1 and the vast amount of data generated, a lot of analytic work has been conducted in this area. Due to the individuality of each session, it is very difficult to predict the future aspects of a particular race by leveraging historical data. However, there have been some attempts made to model a system to predict certain outcomes. The accuracy in these cases has been low due to the disconnect between each event. Below are some of the noted work done in the area that we are focusing on.

2.1 Title: Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing; Authors: Theja Tulabandhula and Cynthia Rudin

The objective was to refer to the previous analysis in the professional sports domain and expand on in-game predictions and decisions for professional car racing. Their objective was to forecast how the ranking would change as a result of tire-changing decisions. They were primarily focused on NASCAR racing and data related to it. Their work entailed EDA, feature generation, modeling, and data mining. They used ridge regression, support vector regression (SVR), using a linear kernel, LASSO (least absolute shrinkage and selection operator), random forests for regression, and two baselines in their methods for prediction. They were able to establish that the baseline approaches are significantly outperformed by ridge regression, SVR, LASSO, and random forests. Furthermore, utilizing the number of tire-change decisions made during a NASCAR session, they were able to anticipate the drop and climb in ranking using their model.

2.2 Title: Formula 1 Race Predictor; Author: Veronica Nigro

The author's work revolves around analyzing the F1 race data and coming up with a model to predict the race winner. The information was gathered from the Ergast F1 data repository and the official Formula 1 website, which includes details on every championship and race from 1950 through 2019 along with their locations. The objective was to examine different parameters over time and develop a machine-learning model that would work in tandem with neural networks to forecast who would win each race. The model classification using neural networks and SVM gave the best results, accurately predicting the winners of 62% of the races in the next year or roughly 13 out of the 21 grand prix races.

In addition to the work stated above, some advancement in this field have also been done by a variety of tech enthusiasts, including the use of machine learning models to predict lap times, pit stop techniques using 20 years' worth of data, and historical pit stop data starting from 2012. Another attempt was made to forecast the crashes; however, the predictions were not reliable or accurate because the event was random.

3 METHODOLOGY

For the project, we used a three fold approach segmenting the work into the multiple elements which where then aggregated and integrated into the dashboard. The segments of the approach are disccused in this section.

3.1 Acquiring the data

The data for this project is sourced from two sources, an API called Eargast and a python library called FastF1. The Eargast API is an open-sourced online API that collects and relationally stores the data from the official F1 website.It is an experimental online service called Ergast Developer API offers historical data on auto racing for non-commercial uses. Please take the time to read the terms and conditions. Since the 1950 introduction of the world championships, the API offers data for the Formula One series.It also has a real-time feature where users can collect Lap-by-Lap timing and telemetry data for a live race. On the other hand, Fastf1 is a python library with most of the functions in it that cater to efficient data fetching and selective data collection. It uses specific calls through an event-based function which just takes parameters like year, location, and the session and fetches the data related to the entire session. In the back end, fastf1 also collects data from Eargast API so, if the API is down for some reason it hinders the data collection. FastF1 processes large data chunks (about 50–100 MB every session), hence the majority of the data is cached locally (be aware).

The Data for most of the analysis is fetched using FastF1, we used the original API to collect historic data for teams and their standings over the years. We constrained the historic data collection to only the Turbo Hybrid Era because the telemetry data is only available for the years 2014 and later. Through fast f1, we collected Race, and Qualifying data for a particular session using the get_session() function in the library. The majority of the analysis and visualizations are written dynamically anticipating the change in parameters for each circuit.

The data fetched from the library is comprehensive with multiple datasets representing certain aspect of either race of the car. For instance, the telemetry data called laps has the data for lap times, compounds used, PitInTime, PitOutTime etc. There's also results, weather, Car position data, Car Speed, RPM, nGear within the same session object. We made use of the majority of the data elements to perform sophisticated analysis where understandability ranges from a beginner's to a high level.

3.2 Predictive Modelling

The model used in this project for prediction of lap times was Recurrent Neural Network. A recurrent neural network (RNN) is a family of artificial neural networks whose connections between nodes can generate a cycle, allowing output from some nodes to effect subsequent input to the same nodes. This model is most useful in the cases, where the data is sequential or only its current input or need to memorize from previous inputs. One example can be that they are frequently used in prediction of upcoming words in a sentence ("I was born in Mumbai, I am fluent in -------" ; Prediction: "Hindi"). Here we use the model to predict lap times using previous lap times as all the other parameters changes continuously. The model we use here utilized LSTM layers to give better prediction. The main benefit of LSTM is that it can analyze whole data sequences in addition to single data points. Intentionally, LSTMs are created toprevent the long-term, short-term reliance issue. They don't struggle to learn; rather, remembering information for extended durations is essentially their default behavior. The main difference of this over normal RNN is the function's complexity they use, normal ones use single tanh, but LSTM utilizes 3 more functions apart from **tanh**.

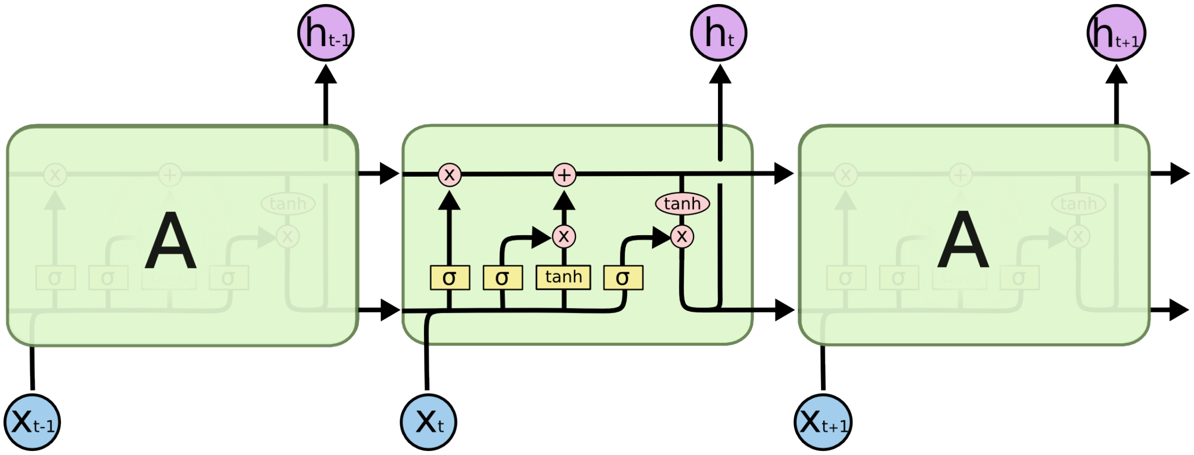


Figure 1: A single LSTM Module

Here the lap times from a specific race and a specific constructor is used to train for the forecasting model, here we have used Mercedes data on Grand Prix. The reason for this is as the car changes from season to season, the model becomes redundant if we use all the previous year data. Then the lap times are normalized using a scaler function & made into an array such that two successive lap times are in a list. It was then split into train & test sets for further analysis (70-30 split)

The structure of the model created here is, the input layer consists of one LSTM layer, Hidden Layer consists of (Dropout, Dense, Dropout, LSTM, Dropout, LSTM, Dropout Layers) and Output Layer consist of (Dense Layer). The training data was used to train the model and to find its performance it was tested against test data. To get a prediction out of this model we have to give 2 lap times as input and descale the output to get the forecasted lap time. The accuracy we achieved for this model is 87.4 %

3.3 Dashboard Development

To host all of our findings we devised a few factors based on which we tested multiple hosting services along with their codebase. The factors were, easy of integration, low latency, high server loads, support for python and visually appealing. Hosting a python project the go-to application code base is Flask and html, but this take a fair bit of new codebase creation and is not at all easy to integrate all the plots we had. Also, the plots we made in the analysis are interactive and dynamic. Choosing flask will loose the dynamic nature because the input stream should be typecasted into the ones understandable to Javascript. The consideration of streamlit came into existence, when all of our factors of integration are well aligned with its properties. Streamlit is a python library which has internal functions to generate UI. The codebase is python itself, resulting in ease of integrating of any piece of analytics code. We could also make it interactive with all the available widgets functionalities. Streamlit also provides a great cache system called the 'experimental singleton' where the data fetched can be cached in the server and doesn't have to fetch the data everytime the page is reloaded. This significantly reduces latency, and server load problem.

The dashboard was developed in such a manner that we could see the full data from two perspectives: one from the viewpoint of a fan and one from the perspective of an investor.

In comparison to the investors, we focused on a little more technical aspects of the race, including an in-depth analysis of grand prix, covering not only the actual races but also the practice and qualifying sessions. Some of the dashboard's key offerings include comparative analysis between two drivers, their gear shift patterns, lap times, and the tire strategy employed by the team in each part of the race. Furthermore, we have included a visual representation of the fluctuations in driver standings over the course of a season. A fan can also get the schedule of the sessions for the entire season. There is also a separate section for displaying points table including the driver standings as well as the constructor standings. A fan can also get the details of the final grid positions following a qualifying session, as well as the details on the knocked out drivers. This fan section of the dashboard not only includes individual and comparative driver performance analysis, but also team performance analysis, where certain key aspects of the session and team strategy can be viewed. This includes information on the drivers' fastest laps and sector timings, the tire compound used, and the weather conditions during a session.

Concerning the sponsor area of the dashboard, we have included data from the Turbo Hybrid Era of Formula One, which began in 2014 and continues to the present day. By default, a sponsor would arrive on the current season, where they would have the choice of seeing the whole chronology or a team's annual performance. When selecting a team in the turbo hybrid era, the whole driver history over the years, the points earned throughout the seasons, and the number of podium finishes achieved by a team would be presented. According to the annual summaries, a sponsor would be able to choose a team of interest. This will display the current year's car model, followed by the current standing and points secured up until the current timeline of the season, as well as the number of wins secured by each driver of the team and the individual points scored by the drivers. The end user can also view the progression of the driver standings through the sessions within the year, as well as the cumulative standings through the season for all drivers from all teams.

Ultimately, after developing the dashboard we looked into multiple hosting services which are free and have the flexibility to host any script. Streamlit, in its native form, can be hosted on Heroku, AWS EC2 and Streamlit Cloud. Out of the three, we tried hosting it on Heroku and Streamlit's Native Cloud hosting service. The load times in standard Heroku service is really long and data-fetch operation fails due to the server load issues. On the other hand Streamlit's native hosting service is efficient when it comes to handling all the aforementioned issues. The hosting service is given on invitation basis, a standard allocation gets a 800Mb of CPU space. Hosting a project is simple and takes only the public github repository link with a dedicated 'requirements.txt' file with all the python library requirements listed. It takes only a few minutes to load all the requirements and boot up the application. As far as the latest version of Streamlit is concerned, the web application's URL is now indexible with a normal Google search.

4 EVALUATION

4.1 Model Evaluation

To be able to produce a model with the intention of predicting the lap times has always been one of the objectives of this project. To understand the specificities of different parameters impacting the lap time we have looked into a lot of parameters ranging from engine performance to tire degradation. But most of these parameters gave out less correlation between themselves & the lap times, one of the reasons for this is having very minute difference in lap times.

Hence for prediction of lap times we have tested 3 models and compared their accuracies to finalize on the model. The 3 models were, Multiple Linear Regression, Time Series, Recurrent Neural Network. The Multiple Linear Regression is a statistical method for forecasting the outcome of a response variable that makes use of several explanatory variables. For the first model, the (float) variables with correlation greater than 0.5 were chosen & scaled accordingly to a normal distribution and were fed into the Multiple Linear Regression model. The maximum accuracy received on 30% of the data used as a testing data as 76.8%, which was bad and came as expected. This was more of the minimum requirement for the upcoming models

The next model we have used is Time Series, here only the lap time parameter was used, and the models used here were, ARIMA & PROHPET. Making scientific projections based on data with historical time stamps is known as time series forecasting. It entails creating models through historical study, using them to draw conclusions and guide strategic decision-making in the future. Both models were giving nearly similar accuracies, 82-84%, but on a average basis ARIMA was producing better result on hyper-parameters tuning.

The last model we used was Recurrent Neural Network, they operate on the tenet that each layer's output is saved and fed back into the system's input in order to forecast that layer's output. To create a single layer of recurrent neural networks, the nodes from several layers of the neural network are compressed. The reason we felt that this might be more useful in prediction is because it uses last lap times as an input to forecast the current lap time. The accuracy we achieved for this forecasting model is 87.4 % and we felt that this might be more competent model compared to the rest of the models.

Table 1: Comparative Model Performance

Model	Accuracy
Multiple Linear Regression	76.80%
Time Series (ARIMA & Prophet)	82-84%
Recurrent Neural Network (LSTM)	87.42%

4.2 Dashboard

Considering the dynamic nature of the dashboard, we cross verified the results obtained from the dashboard to the real-world data. The inconsistencies found are then corrected through code evaluation.

5 RESULTS

5.1 Predictive Analytics

The data collected here is the lap times taken from the races conducted in Grand Prix in 2022, and to be more specific the lap times of both the cars in the Mercedes constructor team. The dataset's shape after preprocessing is (235,2).

This model was trained on 25 epochs and received an accuracy of 87.4%. The reason for not using further epochs or add more layers, is because they caused the model to overfit. When trained on other ML models the predictions came out be constantly around 85 seconds, as the majority of the difference between laps was in milliseconds, but LSTM was able to spot the minute difference, and most impressively it was able to predict areas which would spike (Increased lap duration) on new data.

5.2 Dashboard

The web-hosted dashboard link can be found here. link-to-the-dashboard

6 CONCLUSION

In this entire project, we comprehensively worked on the Formula 1 Data. Analyzing the various segments of the Sport, we got a deeper insight into the how teams perform in a data-centric perspective. The data used in this project is segmented into multiple sessions, Practice, Qualifying and Race. One thing is clear that with the complexity of the data,

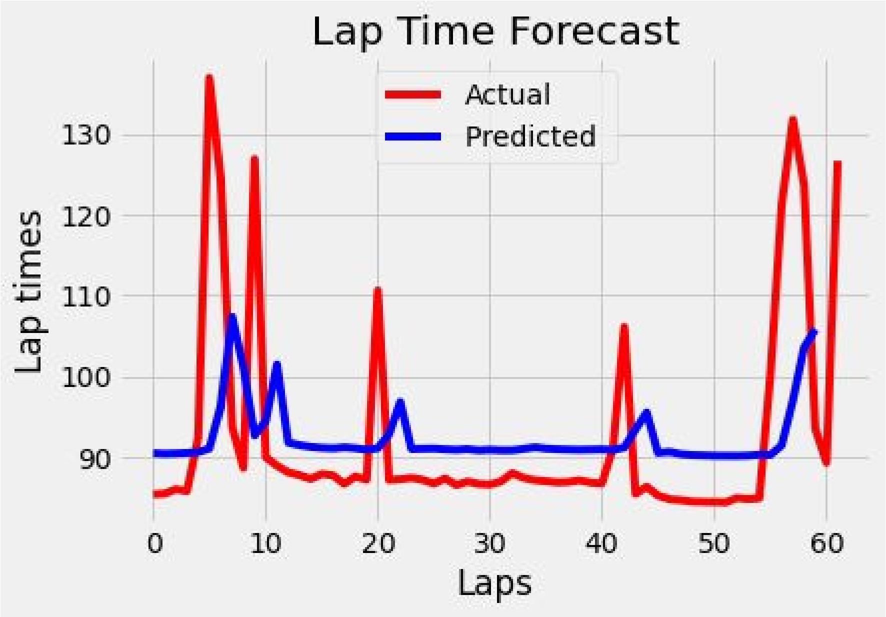


Figure 2: True vs Predicted

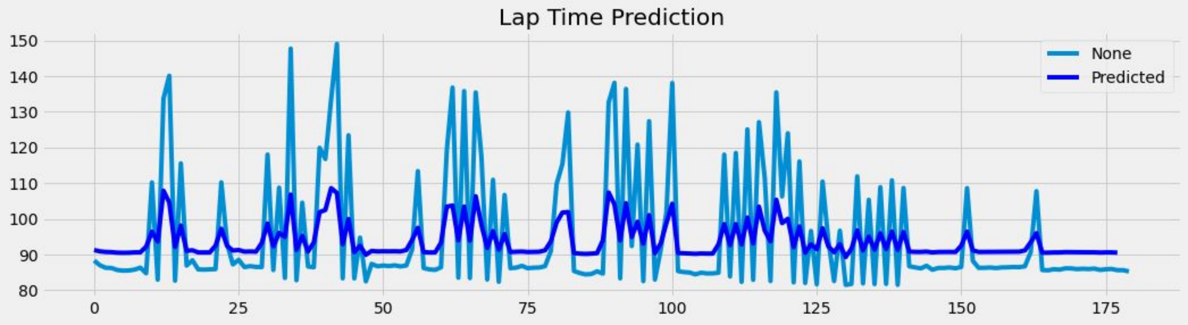


Figure 3: Model Performance

we need exhaustive efforts and substantial amount of time to to include the multitude of other parameters (presently ignored) to improvise the existing analysis. For the predictive analysis, since the parameters are ever changing and are not consistent in every race, It was challenging to account every attribute to improve the model performance. The results obtained are not ideal in context to the Formula 1 standards where a 2second difference has a significant influence. In the entire duration of the project there are a lot of learnings at every moment. We intend to work more on this project futher to improve on the current limitations.

7 FUTURE WORK

The future of our project looks bright as our first goal is to publicize our dashboard for a wider audience. We plan to use sources such as Reddit forums to spread the word so that it comes to light for other F1 enthusiasts. We plan to have our product as a centralized repository for Formula 1. This way our product will be one of its kind dashboard to hold the nooks and crooks of F1 racing sport.This will also fill the current void wherein there's no existing dashboard with the functionalities included in this application. The same of sort of approach can be applied to other racing sport as well. NASCAR is another form of racing which is quite popular in the United States of America.

ACKNOWLEDGEMENTS

We would like to convey our heartfelt gratitude to Professor Di Wu (Data Science Department, University of Colorado Boulder) for his constant encouragement and for his tremendous support in completing this project. His motivation has been a strong push for us right from the beginning of this project up until the final submission. We would also like to thank the Teaching Assistant's for their constant advice and useful suggestions which were very useful for us in completing this project.

We would like to also acknowledge that this project was entirely completed by us and not by someone else.

References

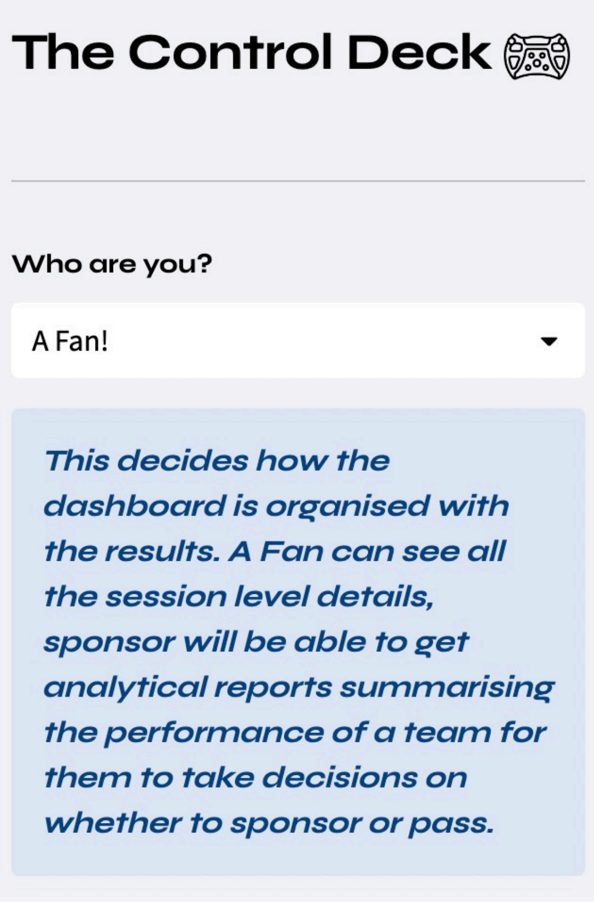
1. Heilmeier, A., Thomaser, A., Graf, M., & Betz, J. (2020). Virtual Strategy Engineer: Using Artificial Neural Networks for Making Race Strategy Decisions in Circuit Motorsport. Applied Sciences, 10(21), 7805.

2. Nigro, V. (2020). Formula 1 Race Predictor. [online] Medium.Availableat: https://towardsdatascience.com/formula-1-race-predictor-5d4bfae887da.

3. Tulabandhula, T. and Rudin, C. (2014). Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing. Big Data, 2(2), pp.97–112. doi:10.1089/big.2014.0018.
4. Jared (2022). Formula One Race Lap-by-Lap Prediction with Machine Learning. [online] GitHub. Available at: https://github.com/Jared-Chan/f1ml Accessed 19 Sep. 2022].
5. SICOIE, H. MACHINE LEARNING FRAMEWORK FOR FORMULA 1 RACE WINNER AND CHAMPIONSHIP STANDINGS PREDICTOR (Doctoral dissertation, tilburg university).
6. ESPN Press Room U.S. (2022). Formula 1 Miami Grand Prix on ABC Attracts Record 2.6 Million Viewers; Largest Live F1 U.S. TV Audience Ever. [online] Available at: https://espnpressroom.com/us/press-releases/2022/05/formula-1-miami-grand-prix-on-abc-attracts-record-2-6-million-viewers-largest-live-f1-u-s-tv-audience-ever/.
7. Amazon Web Services, Inc. (2018). F1 Insights powered by AWS. [online] Available at: https://aws.amazon.com/f1/. Ergast Developer API. (n.d.).
8. Ergast Developer API. [online] Available at: http://ergast.com/mrd/. theoehrly (n.d.).
9. FastF1 documentation — Fast F1 2.3.0 documentation. [online] Available at: https://theoehrly.github.io/Fast-F1/ [Accessed 19 Sep. 2022].Conference Short Name:WOODSTOCK'18
10. Olah, C. (2015, August 27). Understanding LSTM Networks – colah's blog. Github.io. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

APPENDIX

Figure 4: Title



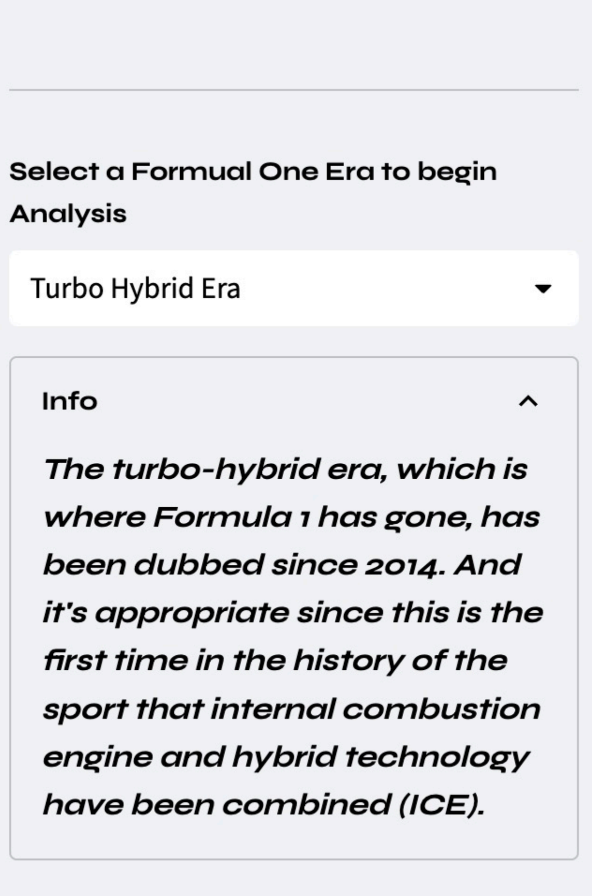
The Control Deck

Who are you?

A Fan!

This decides how the dashboard is organised with the results. A Fan can see all the session level details, sponsor will be able to get analytical reports summarising the performance of a team for them to take decisions on whether to sponsor or pass.

Figure 5: Control Deck



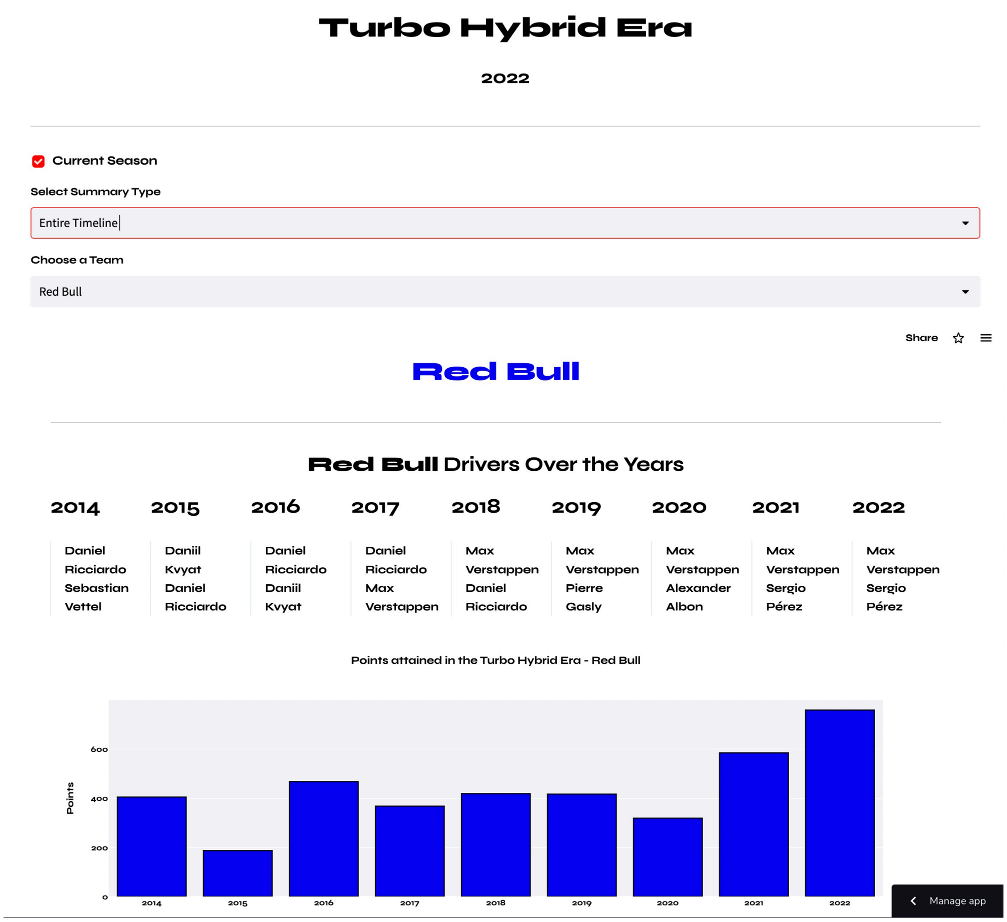
Select a Formual One Era to begin Analysis

Turbo Hybrid Era

Info

The turbo-hybrid era, which is where Formula 1 has gone, has been dubbed since 2014. And it's appropriate since this is the first time in the history of the sport that internal combustion engine and hybrid technology have been combined (ICE).

Figure 6: Information



Turbo Hybrid Era

2022

☑ Current Season

Select Summary Type

Entire Timeline

Choose a Team

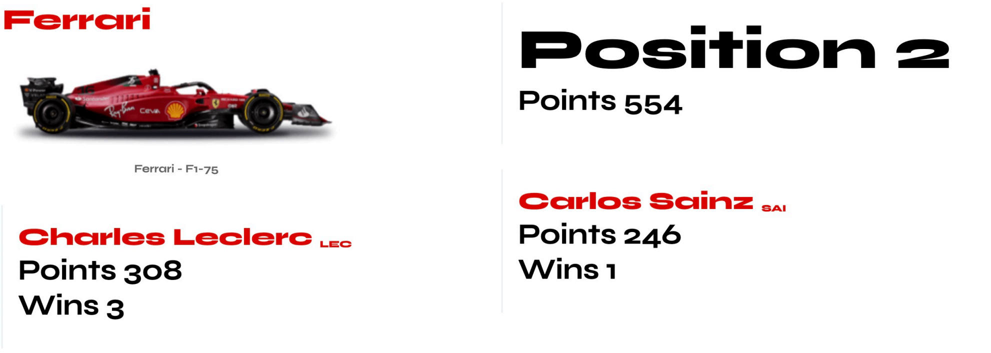
Red Bull

Red Bull

Red Bull Drivers Over the Years

2014	2015	2016	2017	2018	2019	2020	2021	2022
Daniel Ricciardo Sebastian Vettel	Daniil Kvyat Daniel Ricciardo	Daniel Ricciardo Daniil Kvyat	Daniel Ricciardo Max Verstappen	Max Verstappen Daniel Ricciardo	Max Verstappen Pierre Gasly	Max Verstappen Alexander Albon	Max Verstappen Sergio Pérez	Max Verstappen Sergio Pérez

Points attained in the Turbo Hybrid Era - Red Bull



Ferrari

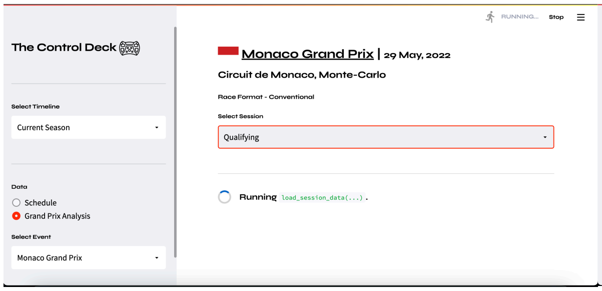
Ferrari - F1-75

Position 2

Points 554

Carlos Sainz SAI
Points 246
Wins 1

Charles Leclerc LEC
Points 308
Wins 3



The Control Deck

Select Timeline

Current Season

Data

○ Schedule

● Grand Prix Analysis

Select Event

Monaco Grand Prix

RUNNING... Stop

Monaco Grand Prix | 29 May, 2022

Circuit de Monaco, Monte-Carlo

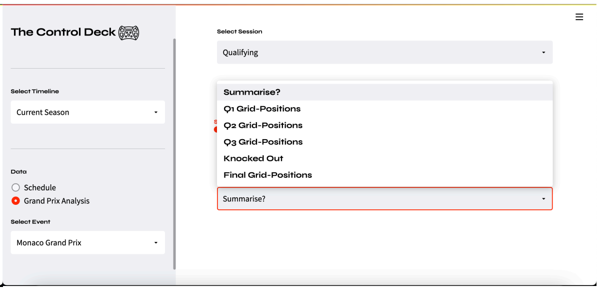
Race Format - Conventional

Select Session

Qualifying

Running `load_session_data(...)`.

Figure 7: Session Select



The Control Deck

Select Timeline

Current Season

Data

○ Schedule

● Grand Prix Analysis

Select Event

Monaco Grand Prix

Select Session

Qualifying

Summarise?

Q1 Grid-Positions

Q2 Grid-Positions

Q3 Grid-Positions

Knocked Out

Final Grid-Positions

Summarise?

Figure 8: Session Details

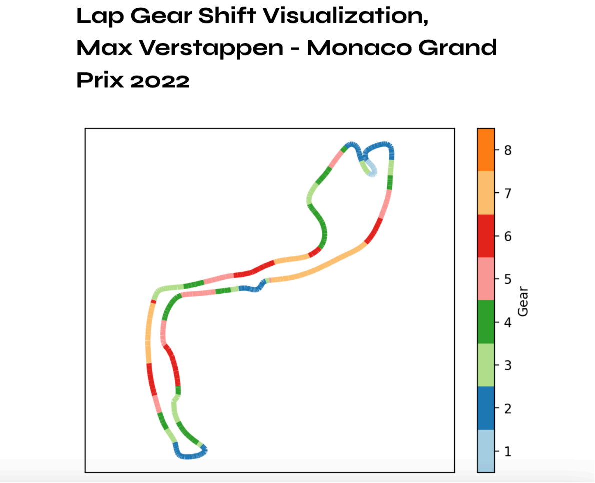
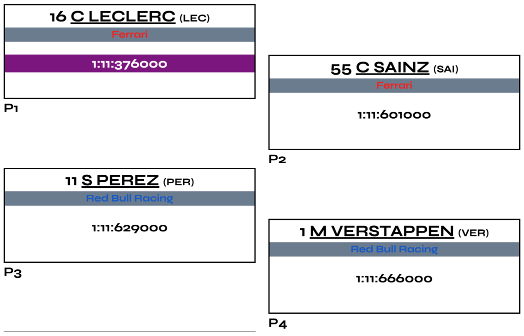


Figure 9: Gear Map



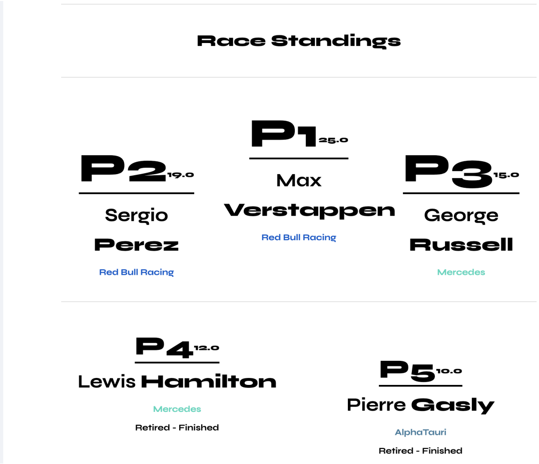
16 C LECLERC (LEC)
Ferrari
1:11:376000
P1

55 C SAINZ (SAI)
Ferrari
1:11:601000
P2

11 S PEREZ (PER)
Red Bull Racing
1:11:629000
P3

1 M VERSTAPPEN (VER)
Red Bull Racing
1:11:666000
P4

Figure 10: Grid Positions



Race Standings

P2 19.0
Sergio Perez
Red Bull Racing

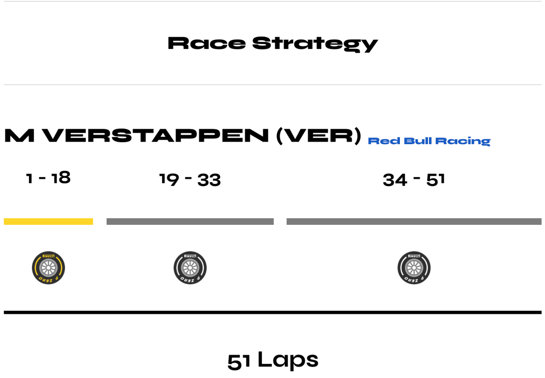
P1 25.0
Max Verstappen
Red Bull Racing

P3 15.0
George Russell
Mercedes

P4 12.0
Lewis Hamilton
Mercedes
Retired - Finished

P5 10.0
Pierre Gasly
AlphaTauri
Retired - Finished

Figure 11: Race Standings



Race Strategy

M VERSTAPPEN (VER) Red Bull Racing

1 - 18

19 - 33

34 - 51

51 Laps

Figure 12: Race Strategy