Formula 1 Analytics

Ayush Praveen
Department of Computer Science and Engineering
PES University
Bangalore, India
ayushpr.ap@gmail.com

Yash Vardhan Soni
Department of Computer Science and Engineering
PES University
Bangalore, India
dhan9576@gmail.com

Anirudh Chandrasekar
Department of Computer Science and Engineering
PES University
Bangalore, India
ani16may2002@gmail.com

Pracheth Thamankar
Department of Computer Science and Engineering
PES University
Bangalore, India
pracheth.thamankar@gmail.com

***Abstract*—Formula 1 is the greatest racing spectacle on the planet, with more than half a billion fans following its unrivaled mix of guts, grit, and glory worldwide. Since 2018, F1 has seen a steady increase in average viewership per race in the U.S. — from half a million in 2018 to almost 1.5 million in the 2022 circuit. The 47 percent increase from 2021 alone has helped spur F1's financial success. In this report, we focus on using statistical analysis methods to gain insights into the various factors that affect pit stops, lap times, and driver performance. Exploratory Data Analysis is performed on a broad spectrum of features such as pit stop duration, wins at pole position, drivers' standings, mechanical failures, effect of different circuits on these features, etc. We intend on running a correlation analysis between drivers and tracks to obtain each driver's best performance on different tracks. Predictions on driver's race results are performed using different classification models and the results are discussed.**

***Index Terms*—Formula One, Data Analytics, Pole Position, Best F1 Driver**

I. Introduction

The sports market has exploded in the 21st Century with the introduction of new media formats such as streaming broadcasting and social media. With the growth of sports and the monetary value associated with it sports analytics and sports team management have become popular topics.

Game outcome prediction plays a major role in sports performance analysis. Game prediction significantly influences many parts of the sports market such as viwerbase, team management, strategy, and also sports betting.

Formula One (also known as Formula 1 or F1) is the highest class of international racing for open-wheel single-seater formula racing cars sanctioned by the Fédération Internationale de l'Automobile (FIA). The World Drivers' Championship, which became the FIA Formula One World Championship in 1981, has been one of the premier forms of racing around the world since its inaugural season in 1950. The word formula in the name refers to the set of rules to which all participants' cars must conform. A Formula One season consists of a series of races, known as Grands Prix, which take place worldwide on both purpose-built circuits and closed public roads. [1]

The Formula 1 Championship is held annually with each season comprising of 20 races held in circuits around the world. Each race has ten teams with two drivers per team, competing against each other. Each team is responsible for building and tuning their cars, planning out strategies, training drivers, etc.

In March 2007, F1 Racing published its annual estimates of spending by Formula One teams. The total spending of all eleven teams in 2006 was estimated at $2.9 billion US. This was broken down as follows: Toyota $418.5 million, Ferrari $406.5 m, McLaren $402 m, Honda $380.5 m, BMW Sauber $355 m, Renault $324 m, Red Bull $252 m, Williams $195.5 m, Midland F1/Spyker-MF1 $120 m, Toro Rosso $75 m, and Super Aguri $57 million. [2] Given the monetary value involved in F1's logistics, team planning, technology, and training, analyzing performance and finding ways to make racing more efficient has become critical. This has paved way for Data Analytics in F1.

Data analytics has become an integral part of the F1 environment. Formula 1 is one of the most data-driven sports in history. The data analytics in F1 is based on telemetry. Telemetry is the collection of measurements and other data at remote or inaccessible points and their automatic transmission to receiving equipment for monitoring. For example the sector times for each track, tire temperatures, braking points, etc. Telemetry started to be used in its modern form in the late 80s. Each car has from 150 to 300 sensors. The number isn't exact because from track to track they add and remove sensors. In Formula 1 racing, even the most subtle improvements can lead to victory. These improvements not only include fine-tuning the F1 cars, but also training the drivers and the pit crew.

Performance is a multi-faceted word in Formula 1. It can be associated with lap time, driveability, top speed, tyre degradation, downforce, power unit output and efficiency, overall reliability, component stiffness, aerodynamic drag, resource efficiency in cost, time, energy, and much more. The various areas of performance can influence each other, so measuring them depends on the data collected and the analysis undertaken. Each Formula 1 car carries around 300 sensors onboard, producing 1.5 terabytes of data throughout a race weekend. For a race season, a two-car team produces 11.8 billion data points. These must all be filtered and analyzed to look for performance gains, reliability issues, or strategies for the team to make better decisions or to work out their competitors' actions.

”Formula 1 is incredibly data-rich,” says Dan Keyworth, director of business technology at McLaren Racing, one of the leading F1 teams of all time.”Although we are a race team, we're really a technology business at heart.”

II. Review of Literature

A. Will the pole sitter be able to use their advantage and dominate the race from the start to the finish line ?

This paper [3] seeks to analyse the effect of Pole Position [4] on the outcome of the race through Formula One's history A front position on the grid, or the so called Pole Position is considered to be the best. It gives the driver a head start to the first along with some significant benefits such as flexible boxing strategies, 'clean air' which affects aerodynamic efficiency of the car and the safety of avoiding mid-field collisions. The authors considered that implementing OLS is inappropriate since the outcome variable 'Final Position' is a limited dependent variable and varies between 1 and 24. Therefore it cannot be treated as a continuous variable. The paper uses two models the Logit model and the Probit model to determine a dummy variable (Will the Pole sitter win(or not win the race)) and the Poisson model to determine the value of the limited Dependent variable. The authors analysed these response variables against several potential explanatory variables such as the driver's skill, performance of the constructor, rain on race day, as well as track characteristics The paper takes into account that over the course of Formula 1's history the procedure of awarding pole sitter position to drivers has changed many times. In general, before the main event, drivers participate in qualifying sessions which involve consequent knockouts and a final timed trial to determine the driver that wins the pole sitter position. The fastest driver/constructor pair would usually get the pole position. But other factors such as fuel usage, time into which car boxed, etc also have an effect.

The paper finds that the effect of pole position is statistically significant at the per cent level (p¡0.01) and positive. Also a position further from the start reduces probability of winning the race (p¡0.01). Rain does not seem to effect the probability of winning. The paper highlights an interesting trend, the effect of Pole Position is increasing over time. In the 1950s the effect is insignificant, from the 1970s the factor becomes statistically significant (p¡0.05). The effect is largest during the 2000s. However due to change in regulations there are slight variations in the effect of holding Pole position at start of the race. The paper also finds that 5 drivers (Michael Schumacher, Aryton Senna, Alain Prost being 3 of them) have a consistently high win probability. This probability also has a life-cycle affected by the drivers general skills. The model used explains 29 percent of the variation in the winning variable . Also the pole position has a significant effect on finishing position (p¡0.01).

The paper concludes that Yes, the pole position does provide a significant advantage over the other drivers in the grid. This advantage is about two positions at the finish line or about a 10 percent point higher probability of winning the race. The paper captures the effect of features such as the driver's ability, constructor and track characteristics and the effect or the lack thereof of rainy conditions.

B. Who is the best Formula 1 driver?

This paper [5] is the first to try to evaluate the true talent of a Formula 1 driver by separating it from the performance of his car. The authors found that most rankings today represent a simple sum of metrics. However, 20 of achieved points and do not reflect a driver's true talent. The racing position of every Formula 1 driver is a function of several factors such as their individual talent, the quality of their cars as well as other race-specific variables, such as weather conditions, characteristics of the track and home advantage, among others.

The authors define the dependent variable

$$Y_{ij}$$

which represents the classification of a driver i in race t. It is found that points is not a good choice to choose as the dependent variable since points gives us no information on the difference in skills of drivers. Also, the number of points awarded has changed over the course of F1's history hence it is unreliable.Racing times was found to depend on racing strategies and hence is not a good dependent variable.Also in later stages of a race advanced drivers tend to slow down on purpose to hold their positions.Similarly the authors eliminated some other prospective dependent variables based on the fact they were influenced by the team and not just the driver. The authors developed the following functional form for the linear regression model

$$y_{it} = \alpha_i + \gamma_{s,i} + X_{\beta} + U_{it} \tag{1}$$

where α_i is a dummy variable capturing quality of driver i and γ_{si}, represents car-year- specific effects. X is the design matrix of the other control variables and β its corresponding coefficient vector. The design matrix contains 1291 dummy variables of all drivers and all cars plus additional control variables.

stands for the error term. By using linear regressions and controlling for driver and car dummies the authors separate the talent of Formula 1 stars from what their car contributes to success. Michael Schumacher has had the most absolute wins and is among the TOP-10 drivers. However, he is not the top-ranked driver. The best Formula 1 driver ever is Juan Manuel Fangio.

Their analysis shows that Formula 1 data are not only of interest when trying to evaluate a driver's talent or in order to establish a world ranking. Additional economic and non-economic applications can be envisioned. By analyzing changes in the rules of Formula 1 driving we could quantify incentive effects. The analysis of dropouts also provides interesting insights on risk-taking.

III. Dataset and Pre-processing

The Formula 1 World Championship dataset [6] contains 14 csv files. This data has been collected through 71 seasons of formula 1 from 1950 till the latest 2021 version. This data has been consolidated from ergast Developer API. It contains data about the circuits, constructor results, constructor standings, constructors, driver standings, drivers, lap times, pit stops, qualifying results, races, race outcomes, seasons, sprint results, and the status of the driver at the end of the race. Constructor results have data regarding the points scored by each team in different races throughout the years. Constructor standings shows information regarding the position of each constructor after a race. Pit stops contain's data about each stop made by a driver in a particular race. Race outcomes, as the name suggests, contains information about the outcome of each race and the position of a driver before and after the race. Sprint results, similar to race outcomes, has the data after each sprint and including the positions of drivers and their lap times.

These data values have been pre-processed across 14 different csv files by various cleaning and imputation methods. Time and duration columns in multiple files were converted to timedelta format for easier analysis and visualization. Lap times and fastest lap times columns were also converted to a time delta format and the null values were replaced with time delta of zero since they had null values for the fastest lap speed. These null values could have been present because the drivers might not have finished these races or they might not have completed enough laps for them to be considered. In some columns, the missing values were filled with the mean of those columns.

IV. Data Analysis

We merged certain dataframes and formed a single dataframe for the analysis. The race outcomes, drivers, driver standings, constructors, and status were merged. There were three columns with missing values namely time taken to finish the race in milliseconds, fastest lap, maximum speed. We filled the missing values in the fastest lap with 0 as we can not assume on which lap they could have gone the fastest. For the other two columns we filled the missing values with the mean of their respective columns.

While checking for the skewness of the columns and found that wins, points and race time in milliseconds show high positive skewness while fastest lap and maximum speed have high negative skewness.

Next, to deal with outliers we specified the Inter-Quartile Range and considered only those data points which lie within the range. To find the correlation between the features, we plot a correlation heatmap[1].

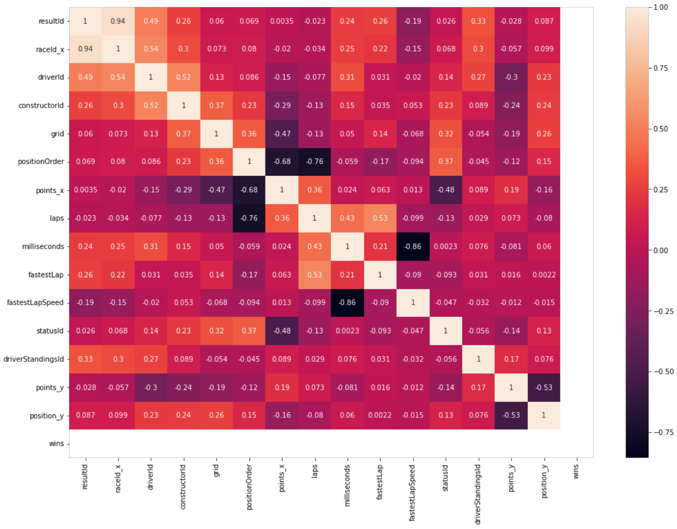


Fig. 1. Correlation Heatmap

As the correlation plot suggests, drivers who have greater maximum speed finish races faster which is indicated by milliseconds having a negative correlation with fastestLapSpeed. It is observed that drivers with a higher position order (numerically low) have completed more laps as our data contains lap-wise information, indicated by a negative correlation. It is also observed that drivers with a better grid position(numerically lower) at the start of the race tend to score higher points which is indicated by a negative correlation between the columns. It does not have a strong correlation though as there might be other factors affecting race outcomes.

A. *Pit Stop Analysis*

The objectives of this section were to analyse the following: 1. How has pit stop duration changed over the years? 2. Relationship between pit stop duration and race circuit. 3. Relationship between pit stop count and race circuit. 4. What are the best pit stop windows in different circuits?

1) Objective 1: How pit stop durations have varied over time? 1. Pit stop durations saw a significant increase from 2012 to 2014, more specifically from 2013 to 2014 with the beginning of the Turbo-Hybrid era. 2. Pit stop durations have remainded fairly stable from 2014 onwards. 3. Majority of the durations are centered around 20-30s. 4. Variance drastically increases from 2019.

2) Objective 2: Relationship between Pit Stop duration and race circuit. 1. Race circuits do have an impact on pit stop durations. 2. Number of laps (and hence length of the circuit) doesn't seem to have any correlation with pit stop durations. 3. Some circuits have larger variances then others, but overall the variances appear to be fairly consistent.

3) Objective 3: Relationship between Pit Stop count and race circuit. 1. Race circuits have a significant impact on pit stop count. We see that Mugello has seen by far the highest number of pit stops on average. 2. Number of laps (and hence, length of circuit) do not seem to have any correlation with pit stop count. 3. Variance also looks to be dependent on race circuit. This means that circuits with higher variance do not have a definite pit stop strategy that every team counts on.

4) Objective 4: What are the best pit stop windows at different circuits? For the sake of relevance, we'll only consider the circuits where races are currently held and we'll only consider pit stop data pertaining to the turbo hybrid era of F1, i.e, 2014 onwards.

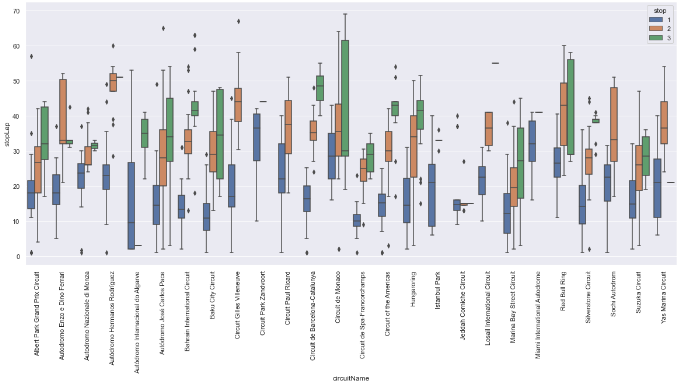


Fig. 2. Best Pit Windows

B. DNF Analysis

The objectives of this section were to analyse the following: 1. What percentage of drivers finish races every year? 2. How has the reliability of cars changed over the years? Is this result consistent with the previous statistic? 3. Who are the drivers with the most number of MDNFs? (Unluckiest drivers) 4. Most reliable and least reliable drivers. 5. Most reliable constructors. 6. Circuits with most number of crashes (Most dangerous circuits).

For this analysis we are working with the following datasets that we cleaned and pre-processed : drivers,constructors,circuits,races, status,results. The results file contains all the codes from above files and needs to be worked on. The dataframes are merged to obtain a single dataframe to perform analysis on.

1) Objective 1: How many drivers finished a race that they started? To figure this out, we can look at the number of drivers across all races with the race status as 'Finished' or '+1/2/3 laps'. This is done because drivers can finish on the lead lap (i.e. they haven't been lapped by the race winner) or they can finish after being lapped. In some cases, drivers may retire in the last couple of laps of the race.

In modern F1, they are still classified in the results as having finished provided they have run 95% of the race distance - usually anything over +3 laps would count as a Did Not Finish (DNF). Historically, all drivers who completed the race (regardless of the number of laps they were behind the winning driver) have been classified. Due to this, all finishers have been considered.

In all of F1 history, only 55.374% of race starts have been finished. But, if modern rules are considered, only 51.068% of race starts have been finished.

Now, we determine the percentage of drivers finishing in each race over the years. Car reliability could be one of the factors which impact these results. Better reliability could mean more drivers finish the race.

We can see from the below plot, the general downward trend in average finishes from 1972 until 1989 which has been the worst year where on average, only 29.5% of drivers would finish the each race. We see an upward trend from 1989 onwards until now where we see around 80% of drivers finishing each race. We investigate in further sections to see if car reliability is a major factor in determining these results.

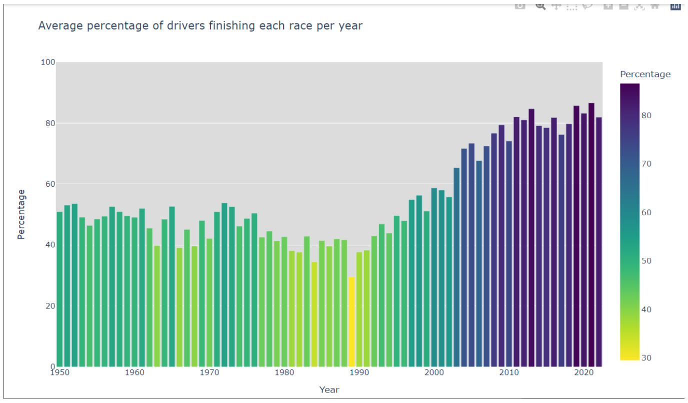


Fig. 3. Average percentage of drivers finishing each race per year

2) Objective 2: How has the reliability of cars varied over the years?

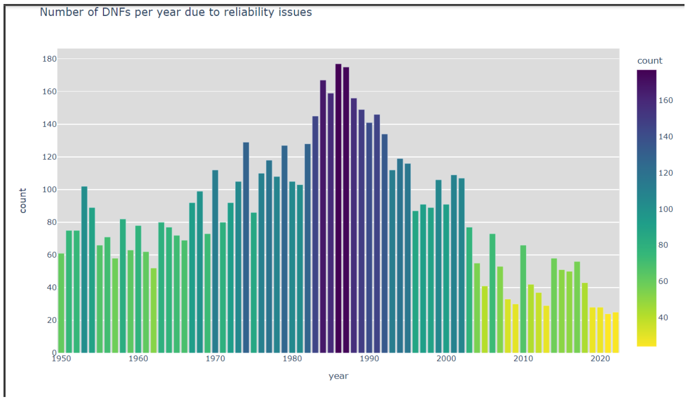


Fig. 4. Number of DNFs per year due to reliability issues

The above plot clearly shows that the 1980s and early 1990s saw the most number of mechanical retirements from races. It is important to note that in the early days of F1, a season had 7 races. In the mid to late-1980s, there were 16 races per season. Additionally, more experimentation and increasing regulations meant the teams had to try out new approaches - often resulting in a DNF result.

So let's normalize the statistic to the number of races that happened in the respective seasons to get a better picture.

Now we see that average DNFs per race increased between 1969 and 1989 with 1989 being the worst year. This seems

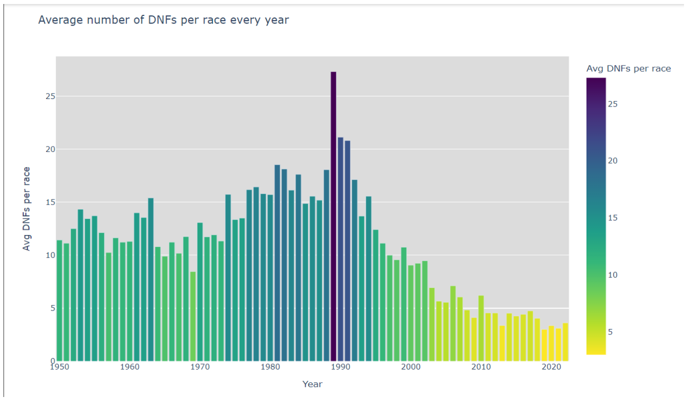


Fig. 5. Average Number of DNFs per race every year

fairly consistent with the previous statistic where we looked at the average percentage of drivers who finished races where we saw 1989 being the year in which the least proportion of drivers finished races.

Next, we need to consider how many drivers took part in each F1 season.

The current F1 driver grid consists of 10 teams with 2 main drivers each. Teams sometimes have to use reserve drivers due to factors such as driver injury, illness, etc. A good example of this is Nico Hulkenberg - who has filled in for Sergio Perez, Lance Stroll and Sebastian Vettel with Racing Point/Aston Martin Racing due to Covid 19 over 2020-2022.

However, F1 has had many more drivers in a single season in the past. 108 different drivers representing 41 teams took part in the 1953 F1 season - the most drivers in a single season ever. Due to thse factors, the following plot looks at the average number of mechanical DNFs in a season per driver.

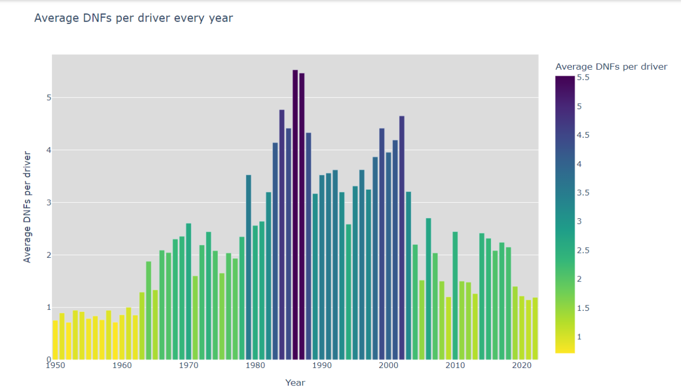


Fig. 6. Average DNFS per driver every year

Since we have only considered DNFs caused by reliability issues. This could be another way to look at how car reliability has varied over the years. Also for the next objective, we look at the unluckiest drivers in F1 who have had most of their unfinished races due to reliability issues not under their control. Looking at this graph, we can assume that many of the unluckiest drivers would have raced between 1983 and 2002.

3) Objective 3: Who are the drivers with the most number of MDNFs? (Unluckiest drivers)

Mechanical DNFs and accidents by driver Now we can look at which drivers have had the most number of mechanical DNFs (i.e. the worst luck) and accidents/collisions. These figures will also be considered in relation to the total number of Grand Prix races they have started so as to get a good understanding.

This analysis makes use of the mdnf_df dataframe created earlier containing all mechanical DNF results in F1 history as well as additional data from the all_results dataframe to account for accidents

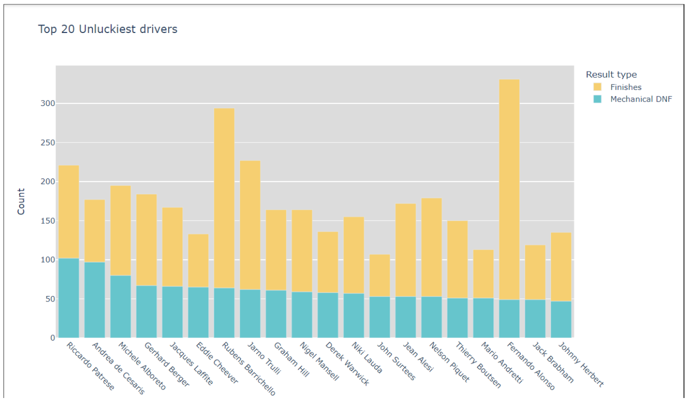


Fig. 7. Top 20 unluckiest drivers

The top 5 drivers in this chart were all active in F1 in the mid to late-1980s - the period of F1 with the highest number of mechanical DNFs (lowest car reliability) as per the 'Number of DNFs per year due to reliability issues' chart above.

4) Objective 4: Most Reliable drivers have been plotted in Fig. 7.

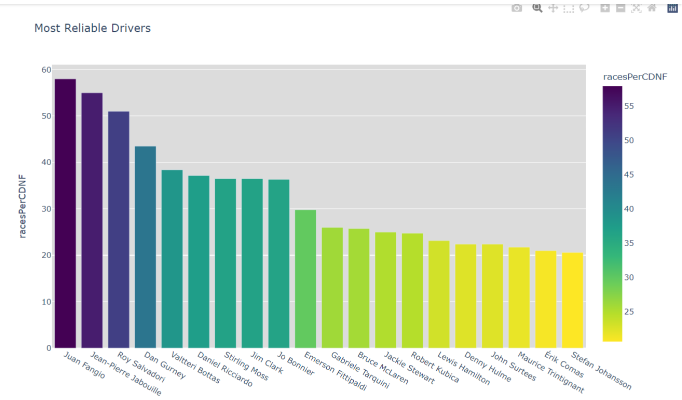


Fig. 8. Most reliable drivers

5) Objective 5: Most Reliable Constructors

This section looks at the number of mechanical DNFs for each constructor. These numbers has been compared with the number of starts made by each constructor. In modern F1, each constructor has 2 starters per race. However, in the early days of F1, a team could enter more than 2 drivers in a race event.

Crashes per constructor have not been considered as accidents and collisions are mostly caused by driver error, weather conditions and other external factors. These numbers have been considered in the drivers section above.

Only constructors with over 100 starts in F1 have been considered so as to have a meaningful indication of reliability over the period of a few seasons.

The number of starts have been divided by number of mechanical DNFs. A higher number indicates higher reliability.

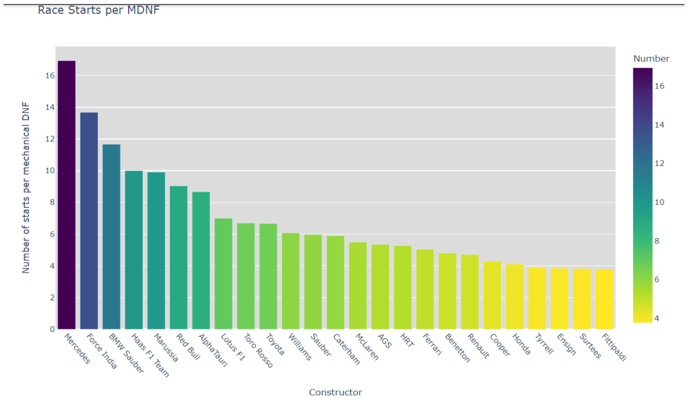


Fig. 9. Race Starts per DNF

6) Objective 6: Circuits with the most crashes

Finally, we can consider the circuits which have had the most accidents and collisions.

Accidents and collisions are more common when weather conditions change during a race. Accidents are also common at narrow street circuits such as Monaco, Baku, Jeddah and Singapore with many drivers touching the wall and compromsing their laps/races.

The number of accidents at each circuit will also be compared against the number of Grands Prix held at the circuit to get a better picture of the dangers a certain circuit poses. An example of why this is necessary is Jeddah - a street circuit which (as of 2022) is very new to F1 (only 2 races held so far) but has had multiple crashes.

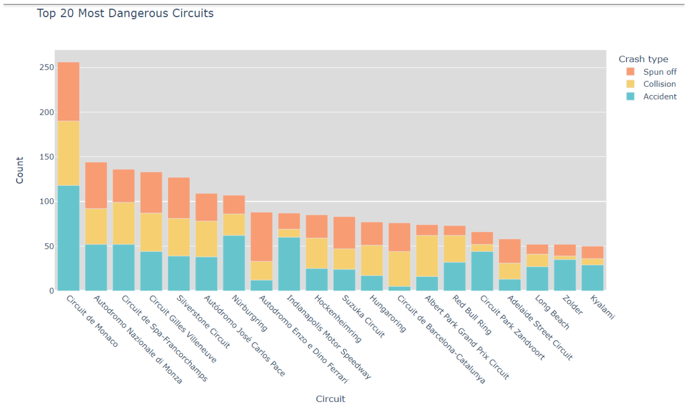


Fig. 10. Top 20 Most dangerous circuits

Finally, we can look at the number of crashes per race at different circuits. A higher number indicates that a circuit is more dangerous. Fig 12 shows this as a graph.

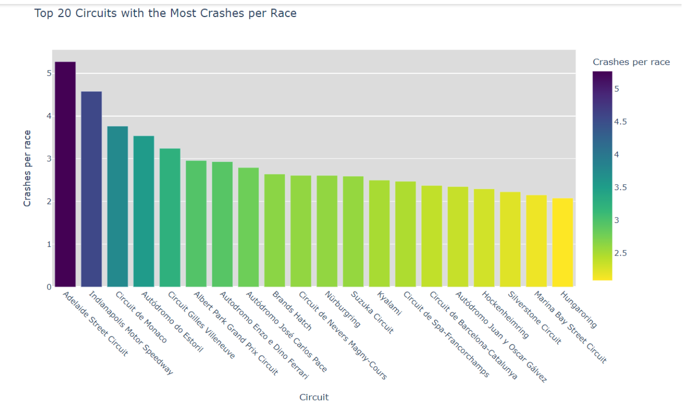
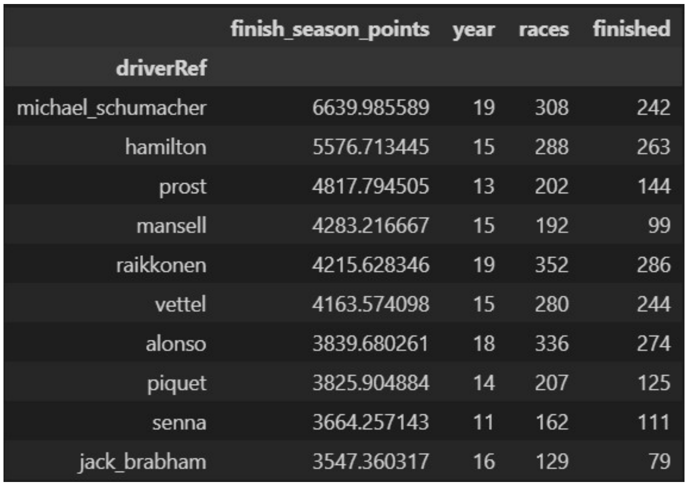


Fig. 11. Top 20 circuits with most crashes per race

C. Driver Analysis

In this section, we try to find the best and most dominant drivers throughout f1 history. Initially, this is calculated by looking at the cumulated points of each driver through history. But there are many assumptions in this: 1. The points system has changed in recent years and hence the list is dominated by recent drivers. 2. The number of races has increased per year which again favors recent drivers. 3. The f1 cars have become more reliable in recent years which favors recent drivers.

Firstly we adjust the points per race and use the current scoring system to calculate the total points cumulated by each driver. Then we normalize points of all drivers keeping the number of races same for each year. Lastly, we normalize points to their race finishes.



	finish_season_points	year	races	finished
driverRef				
michael_schumacher	6639.985589	19	308	242
hamilton	5576.713445	15	288	263
prost	4817.794505	13	202	144
mansell	4283.216667	15	192	99
raikkonen	4215.628346	19	352	286
vettel	4163.574098	15	280	244
alonso	3839.680261	18	336	274
piquet	3825.904884	14	207	125
senna	3664.257143	11	162	111
jack_brabham	3547.360317	16	129	79

Fig. 12. Normalized Points Table

D. Pole to Win Analysis

In this section, We calculate the pole-to-win conversion rates on 2 factors: 1. At different circuits - This gives us an idea as to on which circuits pole positions matter more 2. For different drivers - This is helpful in designing championship strategies. We use the results and races csv files to find drivers who started at the pole position(1st position) and also ended the race at the first position. We set a minimum threshold for both these factors so that we aren't biased towards circuits or drivers where there were only a couple of races where the driver started and ended at the first position. The Pole-to-Win ratio is the ratio of the number of races where the driver started at the pole position and ended first to the total number of races the driver started at the pole position.

The top 20 pole-to-win ratios for circuits are as follows:

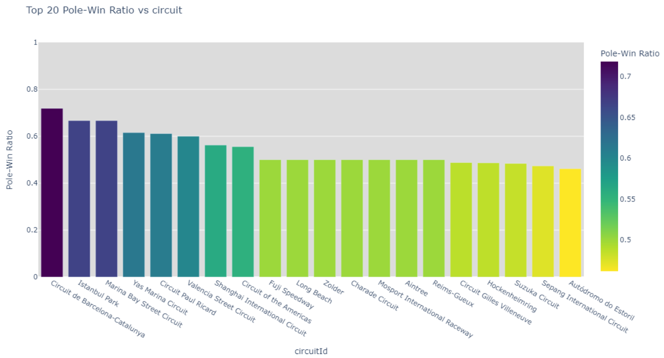


Fig. 13. Top 20 circuits with the highest pole-to-win ratios

V. Experimental Results

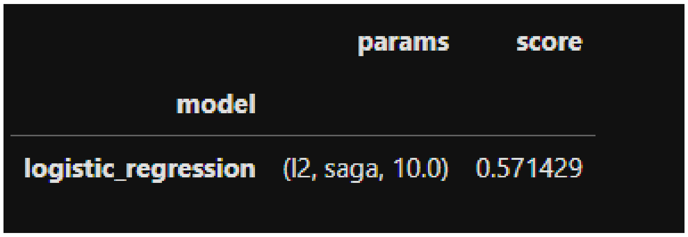
The data we use to run prediction models is the final_df csv file. In this approach we tried to predict the race winners of the 21 races in the 2019 F1 year. The model used for this was the logistic regression model and we have treated the problem as a classification problem. We trained our model on the years before 2019 making our X_train contain data pertaining to the years pre 2019. We drop the driver name and the podium position from X_train. The y_train contains podium position.

Our test data was the 21 races in the year 2019 with X_test being 2019 races with the driver name and podium columns dropped and y_test being the podium position.

The precision score acheieved by the logistic regression model with parameters penalty='l2',solver='saga',C=10.0,max_iter=10000 turns out to be 0.571429 which is the maximum that can be achieved by the model. This accounts for 12 race winners being predicted correctly.

Acknowledgment

We would like to acknowledge our Data Analytics Course Professor Dr. Gowri Srinivasa for encouraging and providing us with this opportunity to get a hands-on experience in the field, and guiding us along the way. We would also like to acknowledge our assistant professors who have prepared the course content and also the teaching assistants who have been constantly providing resources to practice the learnt concepts.



	params	score
model		
logistic_regression	(l2, saga, 10.0)	0.571429

Fig. 14. precision score

VI. Contribution of Team Members

1.Pit-Stop Analysis - Ayush
2.Pole-to-Win Analysis - Yash and Pracheth
3.DNF Analysis - Anirudh and Ayush
4.Driver Analysis - Yash and Ayush
5.ML Models: Logistic Regression - Anirudh and Ayush

VII. Peer-Review Questions

What are your recommendations to the team for their final review? List any questions you would want them to answer in the coming weeks

1.Since the project is more analytical show more graphical representations
Ans: The project currently includes various graphs that have been added since the peer-review.
2.Can show inference on drivers pole positions and how far ahead they are from their competition, fastest pit stops, etc
Ans: Pole positions have been covered.
3.Reason for performing pole to win conversion rates?
Ans: They might not have had the fastest car.
4.Why does it take longer to do pit stops in some circuits over others?
Ans: This has also been covered in our analysis.
5.Which are the most accident prone circuits?
Ans: This has been represented graphically in the report.

References

[1] https://en.wikipedia.org/wiki/Formula_One
[2] Budgets and Expenses in Formula1. F1scarlet. Retrieved 30 August 2015.
[3] Wesselbaum, D. and Owen, P.D. (2021), The Value of Pole Position in Formula 1 History. Australian Economic Review, 54: 164-173.
[4] https://f1experiences.com/blog/f1-glossary-a-z-most-commonly-used-terminology
[5] Stadelmann, David. (2007). Who is the Best Formula 1 Driver - An Econometric Analysis (Wer Ist Der Beste Fahrer in Der Formel 1? Eine Ökonometrische Analyse). SSRN Electronic Journal. 10.2139/ssrn.1017292.
[6] https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/metadata?select=lap_times.csv