Predicting outcomes in Formula 1

Adrian Astăluș

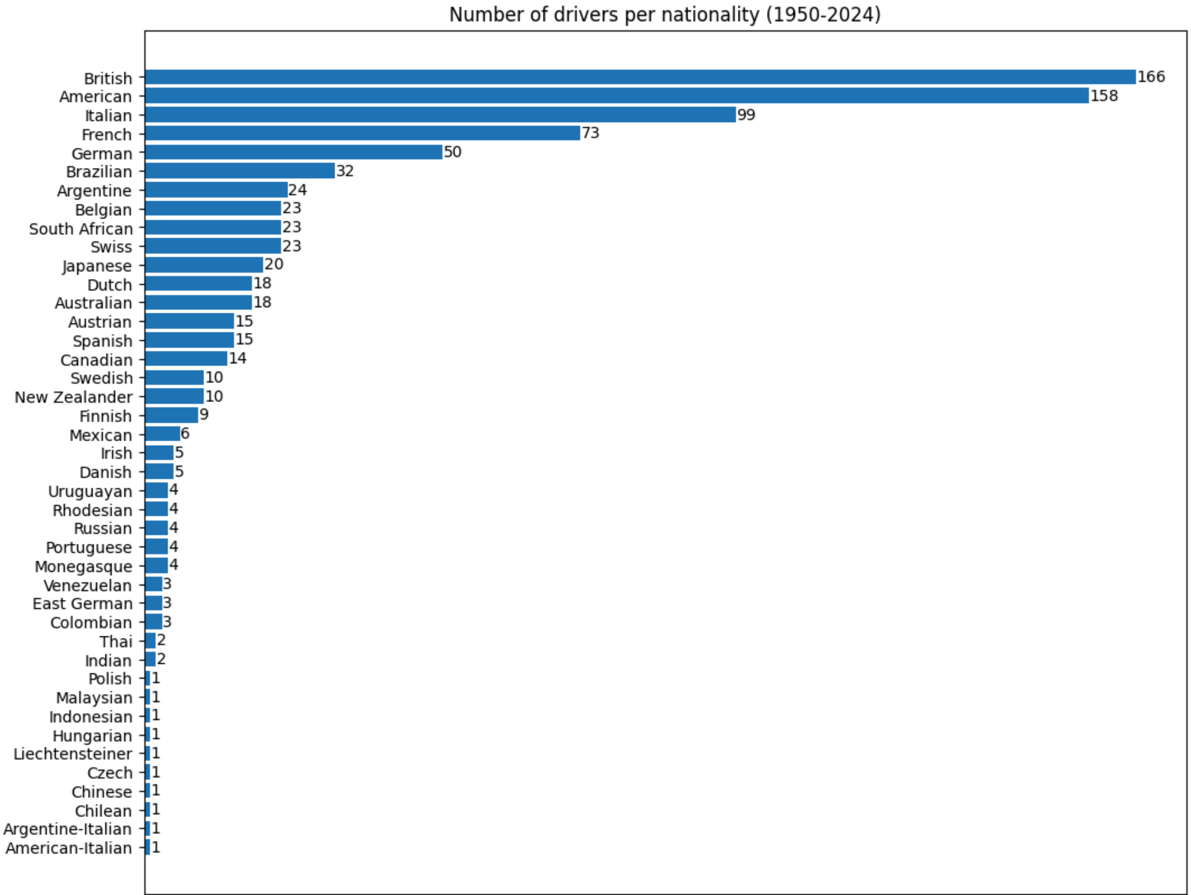
1. Introduction

In the last couple of decades, Formula One has experienced almost a rebirth as the sport has managed to become bigger than ever, attracting millions of people around the world, whether on track, or behind the screens.

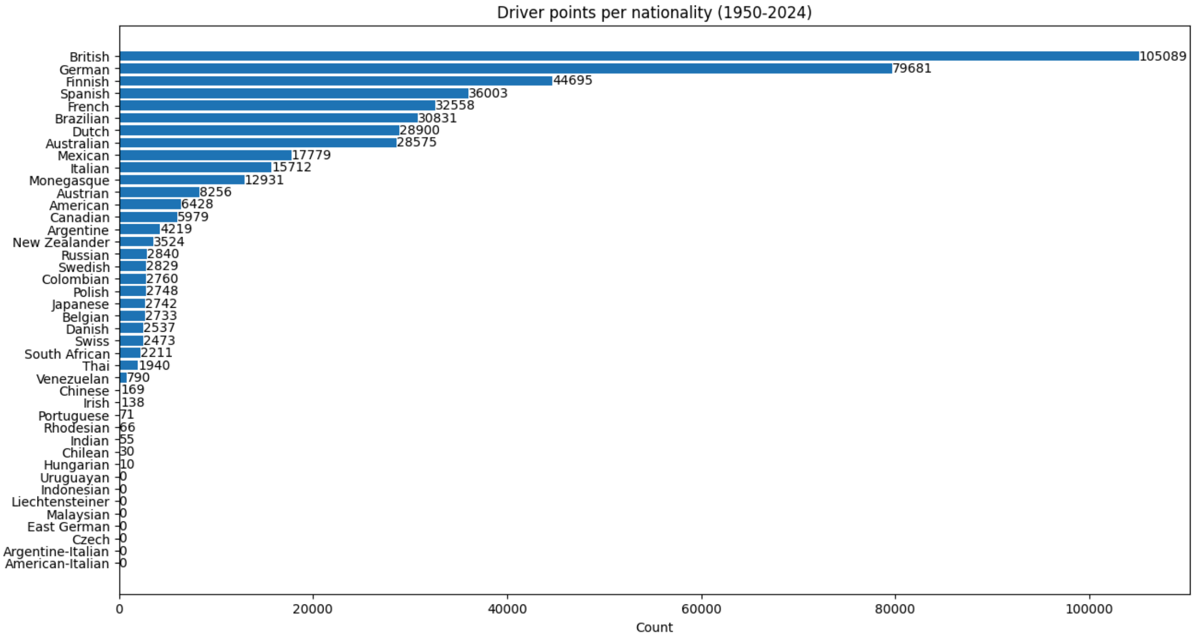
One of the major factors that contributed to this rise is credited to the rapid growth of technology. Besides having at disposal many methods of propagating the sport all around the world, from social media, to broadcasting networks, F1 teams have also heavily invested into fields such as data analytics and machine learning, which have significantly improved car performance, driver performance and race strategies. These investments have managed to make the races the most exciting they have ever been, the sport now being considered the pinnacle of motorsport.

2. Dataset analysis

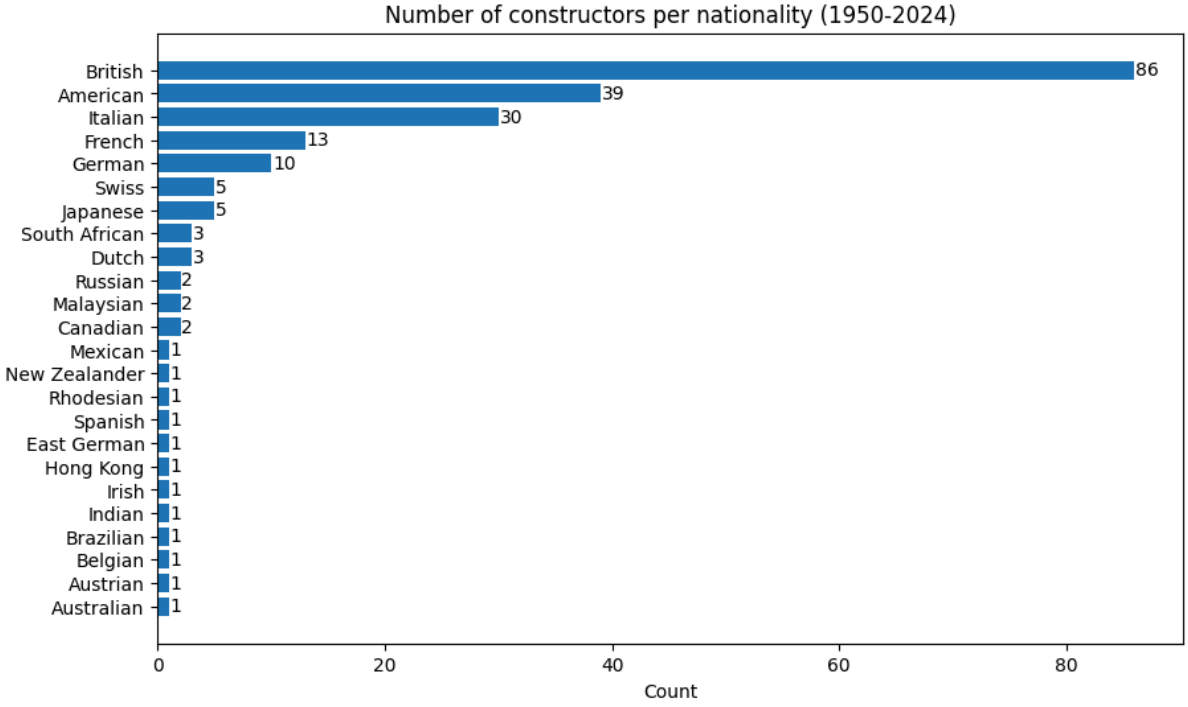
The adoption of the latest technologies allowed the teams and the sport itself to produce amounts of data that were never thought of before, from lap times and driver standings to telemetry data and driver vitals. Some of these data are protected by teams for obvious reasons, but general data is often available online for people to use. One such database is the Formula 1 Championship data, which offers a wide variety of information regarding teams and drivers standings, race results, pit stop times, and many others, making it an excellent data set for analysis.



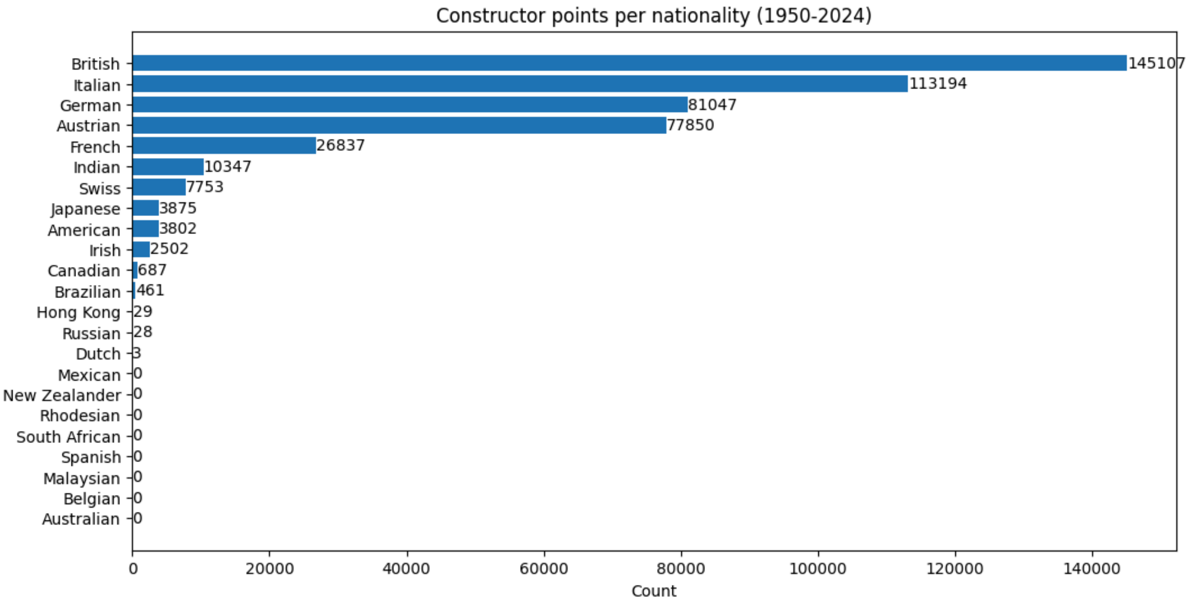
From a drivers' perspective, we can see that this dataset contains information about almost, if not all, drivers that have ever raced in F1. Originating in Great Britain, we can clearly see that a large number of drivers are British, followed by American and Italian drivers. When it comes to the points scored by these drivers, the numbers tell a different story. American drivers, although in great numbers, have failed to score as much as drivers from Germany or Finland, even if there were fewer drivers from these countries.



The same storyline can be seen when we take a look at the teams that have existed throughout the decades in F1. American teams, although great in number, have constantly been outperformed by teams from countries with fewer representatives.



One highlight of the analysis of the teams is, to no surprise, Ferrari, the Italian team which has been present in all the F1 seasons ever. Therefore, the longevity and the incredible success of this team is probably the reason why Italian teams place so high in the ‘leaderboard’. Other notable nationalities that appear in this top are the German teams and the Austrian teams, which have been dominant in the last two decades, through Mercedes and Red Bull respectively.



To solidify the claim that Formula One is an internationally renowned sport, we can take a look at the locations where races have taken place throughout the years to see that the sport has raced at venues from all parts of the globe, on every continent possible.

3. Forecasts using ARIMA

Looking at the dataset, it can be observed that the 2024 season is incomplete. Considering the fact that the season is over, this represented an excellent opportunity to create several predictions and forecasts, having the actual results as comparison.

The dataset stops after round twelve. The first most obvious prediction that can be made is regarding the driver standings. Everyone is interested in who the championship winner will be. This being said, after gathering the necessary data to form a time series, an ARIMA model has been applied on this data in order to predict how the championship standings will look after the following five races.

The ARIMA model has three parameters: ‘p’ (autoregressive order) represents the number of lag observations included in the model. It determines how many previous data points influence the current predicted value. This implies that a bigger parameter value accounts for longer-term dependencies. The mathematical formula for this is: $y_t=\phi_1 y_{t-1}+\phi_2 y_{t-2}+\cdots+\phi_p y_{t-p}+\epsilon_t$, where y_t is the current value, ϕ_i is the autoregressive coefficient and ϵ_t is the error term.

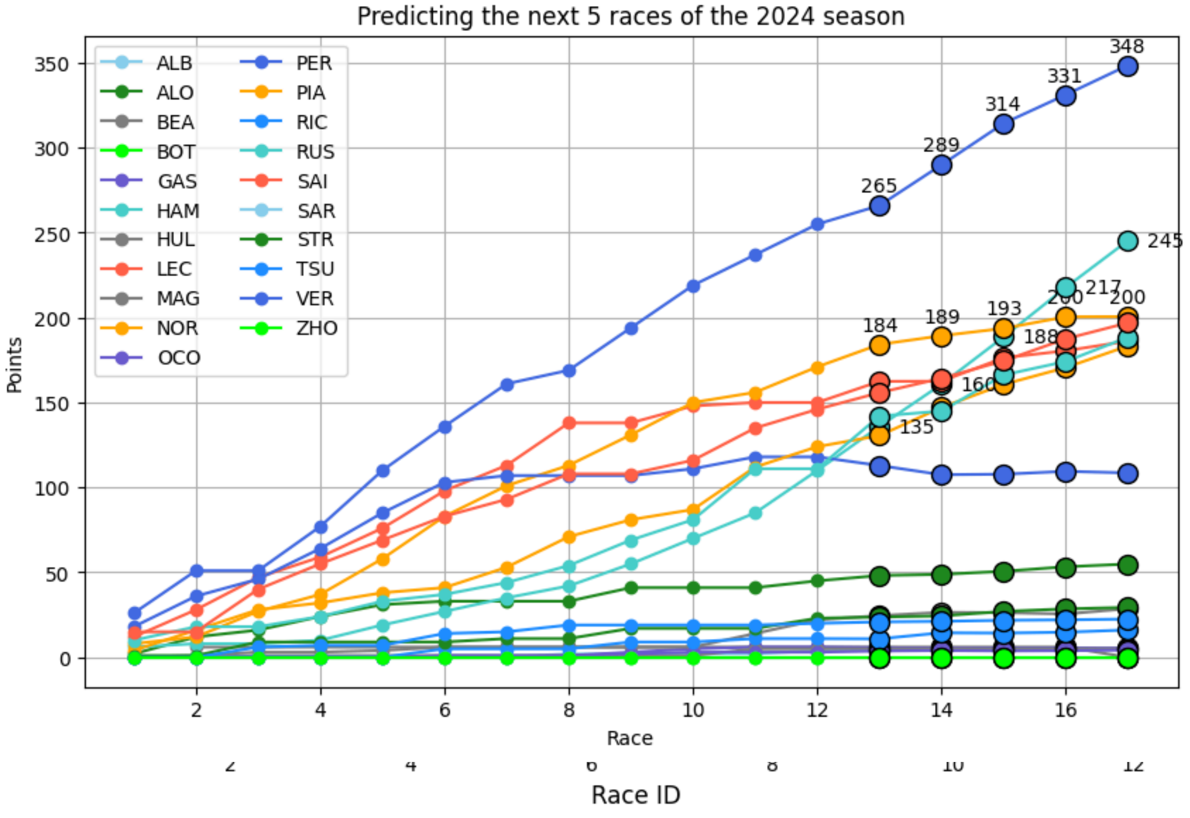
The 'd' parameter (differencing order) refers to the number of times the data is differenced to make it stationary. It is used when the data shows a trend (i.e. is increasing or decreasing over time). The dataset used for this prediction is such a dataset, so differencing will be applied in this case. The first-order difference formula is: $y_t' = y_t - y_{t-1}$. If the first difference applied does not provide a stationary time series, a second difference will be applied, the second-order differencing having the formula: $y_t^* = y_t' - y_{t-1}'$.

Lastly, the third parameter, 'q' (Moving Average Order), refers to the number of past forecast errors included in the model. Mathematically, this parameter can be written as: $y_t=\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+\cdots+\theta_q\epsilon_{t-q}+\epsilon_t$, where ϵ_{t-i} is the previous forecast error and θ_i represents the moving average coefficient.

For this forecast, the (p,d,q) parameter values of 5, 1 and 0 respectively have provided a relatively accurate prediction of championship standings over the next five races.

We can see even from the initial data that the leading driver, Verstappen, looks like the championship winner from the middle of the season already, the next places on the podium announce some surprises. Although starting the season quite slow, Hamilton seems to overtake Norris for the second place on the podium. Further down the field, other changes in the standings are predicted as well.

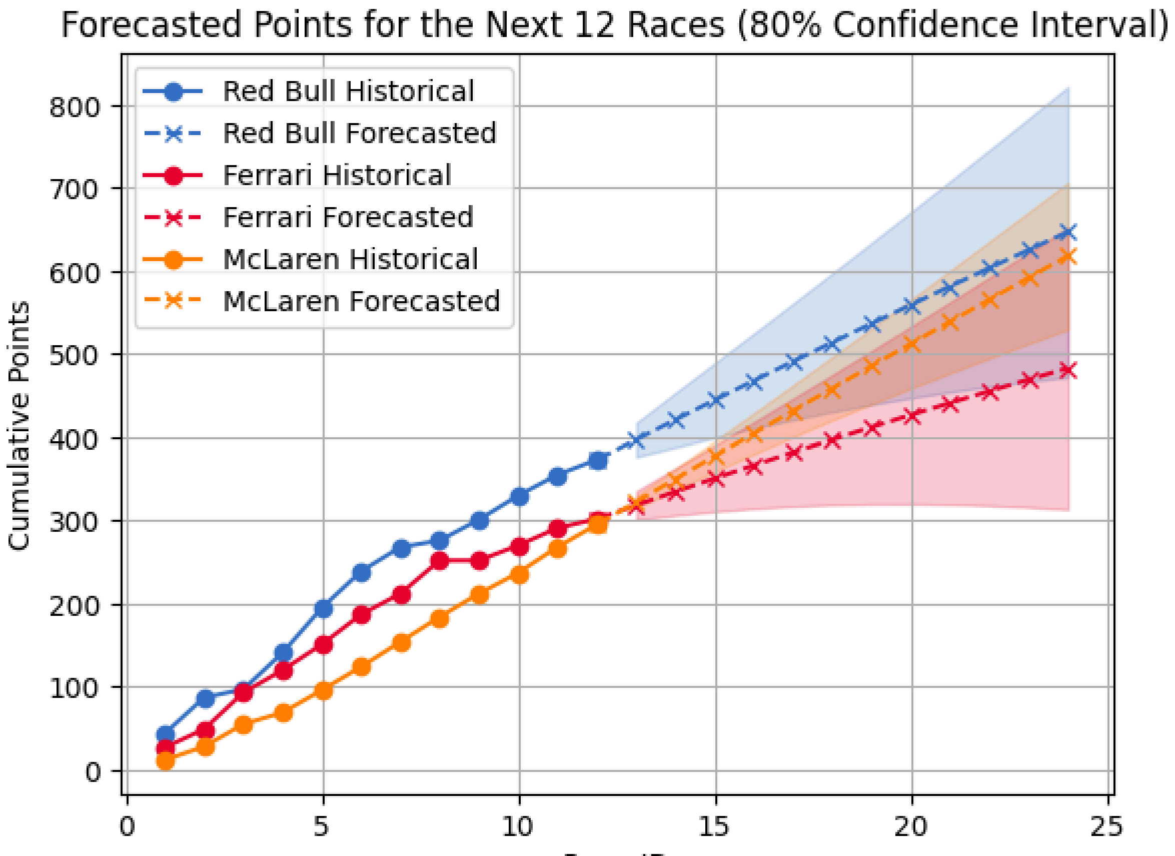
Going down the order, the standings do not seem to change at all, signaling a big discrepancy in points gained by top teams and the teams at the lower end of the grid.



This discrepancy can also be seen when looking at how the constructors championship is unfolding.The top four teams, Red Bull, McLaren, Ferrari and Mercedes seem to be picking the majority of the points available each race, with the other teams battling for the lower places. Even if Verstappen has a considerable advantage in the Drivers Championship, his team is not that far away from the other teams. This being a team game, both drivers need to score consistent points, but this season, Verstappen's teammate, Perez, is very inconsistent. This allows for the other teams to stay relatively close to the leaders. Despite winning two races so far, Mercedes are way too behind to be considered title challengers unfortunately. In contrast, McLaren, having only one win, are so consistent that they are close to overtaking Ferrari for the second place.

This being said, forecasting how the season ends could be of high importance, as the prizes earned by each team depend on the place they finish, and every extra penny counts. Using ARIMA once again, we are able to determine which team ends on top at the end of the season. For this forecast, the chosen ARIMA parameters were (p,d,q) = (1,1,1). Because there are still twelve races remaining in the dataset, and the championship battle looks to be very tight, we cannot be too sure of what the final outcome will be, therefore we also introduce confidence intervals, an 80 percent confidence interval more precisely.

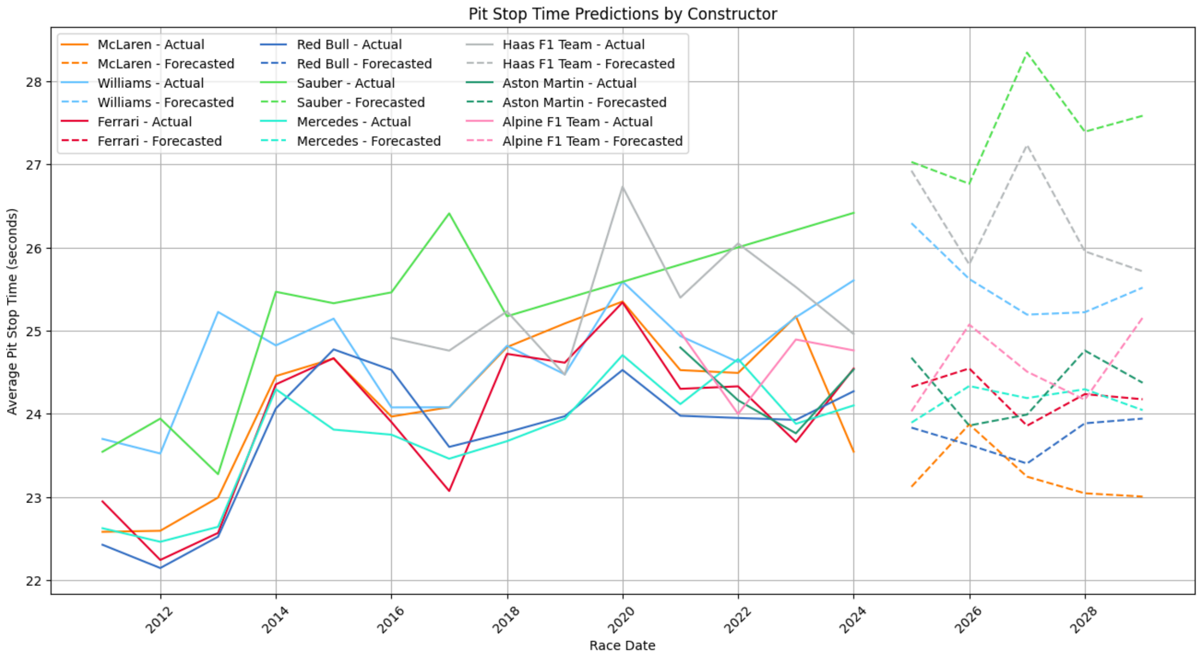
The resulting plot shows that, although Red Bull currently hold an almost 100-point lead over the other two teams, there is still a possibility that McLaren, with their very strong recent form, can overtake them in the standings. Ferrari, on the other hand, seem less likely to overtake them, but are still looking to fight for second place until the last races.



In order to have a championship-winning team, every member of the team needs to perform at their best, whether we are talking about the driver itself, or the mechanics and workers back at the teams’ factories, who come up with upgrades for the car. One part of the team that is oftentimes overlooked, but plays an essential role in the team, is the pit stop crew. In a sport where every second counts, being able to change the car’s tires as quickly as possible can make the difference between winning and losing a race. Obviously, there are human limitations to how quick a pit stop can be performed. But, the pit stop crew consists of more than ten people, so coordination and cohesion are of high importance.

As mentioned before, a quick pit stop could make the difference between a win and a loss, the average pit stop time for each team could be used as an indicator of how well that team performs during that season, although this has a smaller impact that other factors, such as the upgrades brought by the teams over the season.

Nevertheless, we can still use the pit stop data to create a general image on how the teams will perform in the following seasons. For this prediction, we take the mean pit stop time for each time, dating back from the 2011 season and. This information is then used to create another ARIMA model (with p,d,q parameters of 5, 1 and 0 respectively) to predict the mean pit stop time for each of the nine teams over the next five years. Plotting the forecasted values next to the historical data, we can see that the top teams, who have invested in pit stop equipment and personnel consistently over the decade, will still have an advantage over the other teams, some of which are predicted to have even worse mean pit stop times in the following years. From this information, we could say that there is a small probability for the top teams

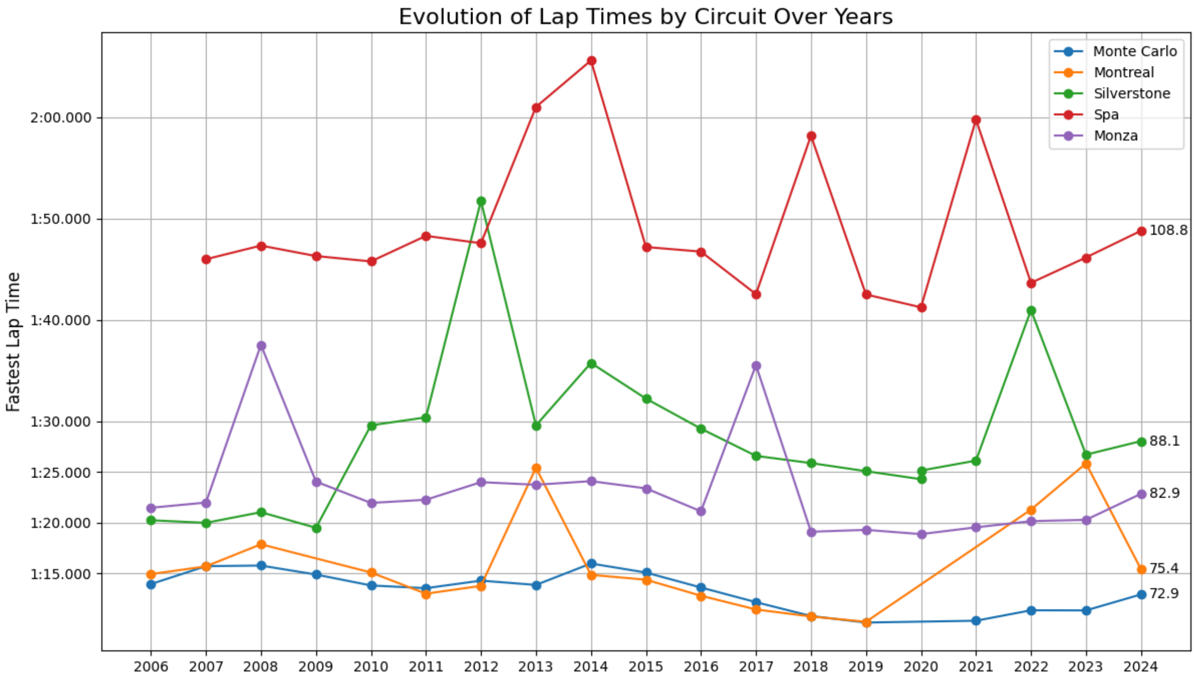


to fall back, or that the smaller teams will have an upsurge in performance, but, given the unpredictable nature of the sport and the regulation changes happening every few years, every prediction should be taken with a pinch of salt.

4. Linear regression in predicting pole times

The data provided by this dataset can be analysed from a historical point of view as well, where trends can be observed over the years. Trying to analyse the performance of a team could prove to be a difficult task, because of the high unpredictability that exists. For example, a team that is clearly the best team in the sport at the time could interpret new rules and regulations wrongly, allowing for other teams to overtake them in the standings. Therefore, information such as pole position times for a specific track could be used, because this metric provides the best qualifying time for each race, without taking into consideration the team that got the time.

In order to analyse the pole position evolution of specific tracks, a linear regression model can be used. A linear regression model fits a straight line relative to the data points and can be used to forecast short term trends. This model helps us provide an estimation of how the pole position times will evolve linearly based on the previous years' results.



The prediction of pole position times could be more precise if more variables were introduced in the linear regression but, considering the fact that the races on these tracks usually take place in similar conditions, historical data can be enough for a prediction with relatively high confidence.

5. Conclusion

In conclusion, this report provides an investigation on how forecasting methods can be applied on Formula 1 data, focusing on predicting key metrics as driver and constructor championship standings, pit stop times evolution and track performance improvements over the years.

The results highlight the fact that the use of forecasting methods such as ARIMA and linear regression provide valuable insights into some of the trends in the sport. However, the available data and its quality proved to be the main challenges, particularly in the area of detailed car performance data.

This being said, coming up with significant discoveries is a tough task. Nevertheless, the use of more advanced models could improve the accuracy of the forecasts and could allow for more in depth analysis of the data.