PROJECT REPORT

AND

LITERATURE SURVEYS

Formula 1 World Championships

Submitted by:

Sl no	Name	SRN	Section
1	Vishnu Anand	PES2201800067	F
2	Vaishnavi Kini	PES2201800253	F
3	Khushdeep Kaur	PES2201800063	F
4	Sudha Guru Priyanka	PES2201800393	F

Abstract:

Formula 1 (a.k.a. F1 or Formula One) is the highest class of single-seater auto racing sanctioned by the Fédération Internationale de l'Automobile (FIA) and owned by the Formula One Group. The FIA Formula One World Championship 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 Grand Prix, which take place worldwide on purpose-built circuits and on public roads. A multitude of drivers and constructors competed every season while fighting for wins, points and titles.

The dataset consists of all information on the Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, championships from 1950 till the latest ended 2020 season. With the amount of data being captured, analyzed and used to design, build and drive the Formula 1 cars is astounding. It is a global sport being followed by millions of people worldwide and it is very fascinating to see drivers pushing their limit in these vehicles to become the fastest racers in the world!

Problem Statement:

We aim to use the chosen dataset to perform a couple of predictions. These include predicting the top 10 drivers of the 2020 season and further, and also the top constructors of the same years. 2020 season was interrupted and many races were cancelled due to the Coronavirus pandemic. Our goal is to predict who would have been the drivers and constructors champion if the season had progressed normally.

About the dataset:

The dataset we have chosen has 13 files, which will go hand in hand as the columns are all interconnected. There are cross-references to various attributes from other tables, and hence removing attributes as a whole was not a valid option.

The dataset obtained was also clean, and did not require major preprocessing. The columns that needed cleaning were checked and dealt with on RStudio, by using the ‘is.na’ function. The columns are compiled as and when required by using the R library ‘sqldf’ that lets us write SQL-like queries on the R platform.

Literature Survey 1:

Regression analysis for prediction: Understanding the process

Palmer,B.P, PT, PhD,, Associate Professor, CSCS, FASCM, Professor and Shelton-Lacewell Endowed Chair

O'Connell, D. G. PT, PhD,, Associate Professor, CSCS, FASCM, Professor and Shelton-Lacewell Endowed Chair

Abstract: **This is a paper about regression analysis for prediction and how exactly that is done. In this paper, the authors have used a dataset about cardiorespiratory fitness.**

Keywords—regression, model, variance, analysis

1. INTRODUCTION

We shall be tweaking the conclusions and results obtained from this paper for our study. The authors have explained about how regression analysis and prediction aids understanding this type of study and how it contributes to better research.

2. DATA COLLECTION

Regression analysis is essentially trying to understand the relationship between a dependent variable and one or more independent variables. There is more to it than just understanding the relationship, we can use regression analysis to predict values for desired values of the dependent variable and also verify our model.

It is always important that we choose a relevant dependent variable. The dependent variable should have acceptable measurement qualities like reliability and validity and so on.

The aim of model selection is to minimize the number of independent variables which account for the maximum variance in the criterion. A more efficient model will only increase the value of Coefficient of Determination(R^2). Higher the value of R^2, lesser the unexplained variance and hence better prediction.

Since determining the best model for prediction is of paramount importance for good prediction, a partial F-test can be performed to test the significance of adding one or more variables into the model. It is always a good thing to test if the chosen variables are significant enough or not and to understand what the outcome will be if we add more variables.

After building the model, it is important to assess the accuracy of the predictions we make from the model. To do that, analysis of standard error of estimate (SEE) should be done. SEE represents the degree to which the predicted scores vary from the observed scores on the dependent variable. Lower values of SEE contributes to the model being more accurate in predictions.

Assessing the stability of the model is also a very important step once we are sure that the model produces accurate predictions. A model can be claimed as "stable" if it produces similar results for different samples from the same population without losing its accuracy. Cross-validation is done on the model to understand if the model is stable. This kind of validation uses data sets that are split into training data groups and validatory data groups. The R^2 values are compared from both the sets and assessment of shrinkage is done to determine if the model is stable or not.

3. CONCLUSION

After studying this paper, we have come to a good understanding of how and why regression analysis is important especially when Formula 1 is an extremely data driven sport.

Drivers find pace from the car and the track in specific corners and in a matter of a few hundredths of a second. If the team knows what racing line to follow and what lap times they have to maintain, it can prove very beneficial for them.

In real life, teams collect and analyse data from previous years as well as data from their simulators to compare and understand what pace they have in their cars and what they should be doing to get a good result in any given race. Each year, the car is different in some ways as teams try to optimize their results from their previous year. So, they collect and analyze data at insane speeds to optimize the performance of their cars.

From this paper, we have understood the procedure of regression analysis. We have also understood the methods of testing model accuracy and model stability and check if it meets the standards of reliability and validity.

Literature Survey 2:

Who Is The Best Formula 1 Driver? An Economic Approach to Evaluating Talent

Eichenberger. R,
Department of Economics,
University Freiburg,
CH-1700 Fribourg, Switzerland

Stadelmann.D.,
Department of Economics,
University Freiburg,
CH-1700 Fribourg, Switzerland

Abstract: **The observable performance of a driver depends both on his talent and the quality of his cars. In this paper, the author separates driver talent from car quality by econometrically analysing data covering 57 years of Formula 1 racing.**

Keywords: Formula 1, quality, racing

1. INTRODUCTION

The estimates also control for the number of drivers finishing, technical breakdowns and many other variables that influence race results. Individual success is determined to a large extent by factors such as the competitors' talents and the quality of their cars, the number of competitors in a race, weather conditions during the race, and pure racing luck. A talent estimate can be obtained by multiple regressions.

Data: The internet database "FORIX" by the magazine "Autosportatlas" represents the main source of information used by the author. Additional information and variables were coded using the official Formula 1 website formula1.com. This paper performs the analysis for the 302 drivers who achieved at least one point during their career. In the dataset, most of the descriptive statistics are evident such as length of the race, circumference of the track, rounds in grand prix, weather conditions, age of the drivers at their career start and end, number of races per driver, successful participations of drivers in wins, podium positions of drivers, and car changes.

2. DATA

The racing position of every Formula 1 driver is a function of a number of important impact 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.

From the linear regression model, the author obtains a unique driver coefficient αi of the dummy variable for every racer. This coefficient serves as an indicator for a diver's talent. The lower the value of the coefficient the better the Formula 1 driver.

The more drivers finishing a race, the more difficult it is to achieve a good classification. The interaction terms of the dropout periods with this variable are insignificant. A Wald-Test for their joint significance rejects the null hypothesis (p-value 0.001; F-value of Wald-Test 3.997). The control for technical dropouts TECHOUT is positive and highly significant as expected. Weather conditions have a negative and significant influence. As there are more human dropouts when the weather is bad, the average classification increases during bad weather. The length of a Grand Prix is also negative but does not have a significant influence (11-%-level). Thus, we have the following results for the control variables of specification.

3. CONCLUSION

Formula 1 drivers are faster the more talented they are and the higher the quality of their car. This paper is the first to try to evaluate the true talent of a Formula 1 driver by separating it from the performance of his car. Most rankings today represent a simple sum of achieved points and do not reflect a driver's true talent. The author treats talent as independent of the cars used and a number of other characteristics.

By using linear regressions and controlling for driver and car dummies we can separate the talent of Formula 1 stars from what their car contributes to success. The analysis of dropouts also provides interesting insights on risk-taking.

Literature Survey 3:

Formula 1 Race Predictor

Nigro, V. (2017)
PhD, Data scientist with a
Background in Finance and photography

Abstract: **This paper deals with the specific problem we have in hand, namely the prediction of the top drivers in Formula 1. It uses a varied approach, by including classification and regression both.**

Keywords: Regression, formula 1, statistics

1. INTRODUCTION

• Data Collection:

The author, starts out by describing how the collection of data was done. The data in this paper, has been gathered from two resources, namely the Ergast F1 data repository and the official Formula 1 website. Essentially the two have the same data, but using multiple datasets increases accuracy and completeness.

Dataframes have been created with the available datasets, factoring only the important columns in. Queries such as if a correlation could be present between the age of the drivers and their performance, if racing in their home country could have any psychological impact, or if some drivers are more prone to crash than others are answered.

2. DATA PREPROCESSING

As this paper uses a Machine Learning intensive model, we shall be refraining from using those tactics, and focus more on the regression parts of the paper. It also uses logistic and linear regressions, random forests, support vector machines and neural networks for both regression and classification problems.

The author ended up with 6 dataframes, after data preparation, and these were namely races, results, weather, driver and team standings and qualifying times from 1983 to 2019. The age of drivers and the cumulative difference in qualifying times is also calculated so that an indicator of how much faster is the first car on the grid compared to the other ones for each race is made.

The author has also taken pains to compile the data from various sources, whereas in our case we have used a pre-made dataset.

Our research would not focus on prediction using neural networks, but will stick to the base idea of linear regression, and multiple linear regression.

3. CONCLUSION

The predictions made by the author, and the predictions for the 2019 races are in fact accurate. As Formula 1 is a data intensive sport, predictions like these are quite useful as it can prove very advantageous to the constructors and the drivers alike.

Literature Survey 4:

Formula 1 race car performance improvement by optimization of the relationship between the front and rear wings

Bhatnagar.U.R
2016
The Pennsylvania State University
The Graduate SchoolCollege of Engineering

Abstract: **The sport of Formula 1 has been one of exhilaration, with every driver wanting to go faster and beat the previous time. Aerodynamic performance of a F1 car is currently one of the vital aspects of performance gain, as marginal gains are obtained due to engine and mechanical changes to the car.**
Keywords: Grand Prix, front, rear, aerodynamics

1.INTRODUCTION

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 Grand Prixs, which take place worldwide on purpose-built circuits and on public roads.

2. DATA COLLECTION

This paper explains that the aspect for potential development in the speeds lies in the modifications of the front and rear wing sections of the car. To perform the optimization, an algorithm called Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is used. A deep study of the drag and lift coefficients is made.

In order to obtain information, a lap simulation tool called AeroLapis used by the author. For simulation, the Sepang F1 race track, which annually hosts the Malaysian Grand Prix(GP), is selected. This track provides a perfect conundrum of whether to design the car for high downforce or low drag configuration, as it contains fast-turning corners and long straights.

The paper delves into mechanical details regarding the aerodynamic aspects, much of which we shall not be applying in our study. Nevertheless, the design of the most optimum car includes a factor of increasing its speed, and hence this paper might be crucial to that.

3. CONCLUSIONS

It was established that the total value of aerodynamic forces can be calculated as superposition of the aerodynamic forces generated by individual components, namely, front wing assembly, rear wing assembly, and body without wings. The study was conducted further, and it was found that the fluctuations in lap times were more sensitive to change in viscous drag of the front wing, rather than the one in the rear.

Finally, the drag reduction strategy was opted out to be the one that got the best lap time out of the car. The main purpose of the paper was to present information signifying the effects of change of the aerodynamic coefficients on the lap times.

The drawbacks of this model that the author has drawn out, is that interaction between different components of the car were assumed to be incorporated into the contributions suggested. But based on the location and design of the wings, the contributions might undergo change.

For our study, these conclusions can be factored in while we are calculating lap times, and while optimizing speed in general.

A critique of others' approach:

1. ASSUMPTIONS MADE

The observable performance of a driver depends both on his talent and the quality of his cars. Formula 1 drivers are faster the more talented they are and the higher the quality of their car. This paper is the first to try to evaluate the true talent of a Formula 1 driver by separating it from the performance of his car. Current rankings of Formula 1 racers provided by racing magazines and on the internet do not separate the qualities of the drivers and their cars, nor do they recognize the influence of other determinants of race outcomes. Usually, such rankings represent the simple sum of points, races won, podium positions achieved or similar measures. The resulting rankings are often not even corrected for the number of races a driver participated in, even though it is evident that competing in more races leads, ceteris paribus, to more points, podium positions and wins.

We treat talent as independent of the cars used and a number of other characteristics.

2. APPROACH USED:

The author separates driver talent from car quality by econometrically analyzing data covering 57 years of Formula 1 racing. The estimates also control for the number of drivers finishing, technical breakdowns and many other variables that influence race results.

The main aim of a driver and a team is to achieve a good classification and to obtain many points during a season. However, points are not a good choice as a dependent variable. Firstly, points are only attributed to the first six or eight classifications and thus differences in the performance of drivers without points could not be distinguished although they make an important difference for drivers and their teams. Secondly, the sum of points achieved also depends more on luck than the classification achieved. Thirdly, the number of points per classification was adjusted over time due to changes in racing rules. This makes comparisons using this measure complicated and unreliable. Racing time is neither an appropriate measure for success because it depends heavily on racing strategies. Finally, training times are also not a reliable measure of performance.

While they contain information on the overall speed of a driver they also depend largely on a team's strategy. In recent years it was forbidden to refuel the vehicle after training and before the race. Thus, cars with significant differences in fuel and therefore in weight participated in the qualification training, biasing training times.

As technical dropouts are not directly linked to a driver's talent, we control for such dropouts with a dummy variable. For human dropouts, the author calculates a hypothetical classification. There is no information available on the ranking of a driver during the time of dropout. Thus he sets counter factual rankings for human dropouts which equals the classification of the last driver arriving plus the number of total dropouts divided by two. We then test whether our results react robustly to variations in the treatment of human dropouts. From the linear regression model, we obtain a unique driver coefficient αi of the dummy variable for every racer. This coefficient serves as an indicator for a diver's talent. The lower the value of the coefficient the better the Formula 1 driver.

At the beginning of Formula 1 racing, many rather inexperienced drivers participated in Formula 1 racing without clear career perspectives. They often remained in Formula 1 for a short time. Thus, their results depended heavily on fortune. Consequently, they may bias estimates. Moreover, using 719 drivers would lead to a data matrix which could only be handled with computational difficulty. Thus, the author only analyzes the 302 drivers who achieved at least one point during their career. When presenting the results we consequently focus on drivers who participated in at least 40 races.

3. RESULTS OBTAINED:

In this paper the author analyzes a dataset from the start of Formula 1 racing in 1950 up to 2006 and calculate talent estimates for every driver. Thereby, establishing a historical world championship ranking which is based on the true talent of Formula 1 drivers. According to our results Michael Schumacher has been the fastest driver of the last three decades but he is not better than Formula 1 superstars of times gone by, such as Juan Manuel Fangio and Jim Clark. Apart from Schumacher, more recent

drivers such as Fernando Alonso and Kimi Räikkönen enter the all time TOP-10 champion's list.

The best racer in history, Juan Manuel Fangio, is not only the best concerning our talent ranking, he is also the best when considering relative measures of races in points and wins. Juan Manuel Fangio has achieved point ranks in 84.3 % of the races he participated in and in 47.1 % of his races he actually won. Thus, out of the 51 races he participated in, he won 24, was 35 times on the podium and all of his dropouts were due to technical reasons. Thus, Juan Manuel Fangio is also the best driver when considering only relative measures. Michael Schumacher has had the most absolute wins and is among the TOP-10 drivers. However, he is not the topranked driver. The best Formula 1 driver ever is Juan Manuel Fangio.

4. LIMITATIONS:

Well-organized teams were not common at the beginning of Formula 1 racing. While two or more drivers used the same car, the racing heroes concentrated on their own success and less on the team's success. Thus, the influence of team orders was probably negligible. Today, the team's success is important too and weaker drivers in a team sometimes make room for their team partners. Unfortunately, we cannot control our estimates for such team orders.

Sometimes people state that a driver's contribution to success depends not only on his own driving talent but also on his talent to improve the team's car. In the author's regressions, he only controls separately for a driver's capability and a car's capability and it is not possible to estimate the influence of a driver on his car. Drivers and cars are treated as dummy variables in the estimates.

5. PROPOSED PROBLEM STATEMENT:

An economic approach to evaluate the best Formula 1 driver:In our case robustness comprises changes in coefficient values and thus changes in the ranking which is a far stricter robustness measure than generally applied

In robustness tests the author analyzes the sensitivity of our statistical assumptions and include the racing experience of a Formula 1 driver in several specifications.

All evaluations, rankings and ranking lists suffer from random influences. Talent and capacities of the athletes are often not directly observable. Thus, other rankings hide the high uncertainty and the relative instability by not showing the error probability and the standard

deviations of their estimates. In comparison to other rankings which are based on points or wins, the author's analysis permits us to estimate an error probability and provide estimates of the variance of the talent coefficients. Thus, we do not only explicitly calculate the talent of a driver but also seriously take into account the error probability.

When including additional control variables, the coefficients of the drivers change only slightly and they change symmetrically for all drivers.

Our 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.

Visualization:

An open-source software called 'Tableau' is used to obtain most of our visualizations. It is an easy to use interface that is exclusively made for data visualization and uses the drag-drop feature. We have also used R to visualize some aspects, and the results are as shown below.

In Formula 1, each constructor races 2 drivers in 2 different cars. The points both the drivers score are added and awarded to the constructor.

To better understand the data set, we have plotted the points scored by the constructors and drivers to see who is or was the most decorated driver and constructor.

Figure 1 shows the driver who has scored the most number of points in his career. According to this chart, the top 3 drivers are Lewis Hamilton, Sebastian Vettel and Fernando Alonso.Figure 2 below shows the constructor with the most points scored. The conclusions we can draw from the chart is that the top 3 Constructors are Ferrari, McLaren and Mercedes.

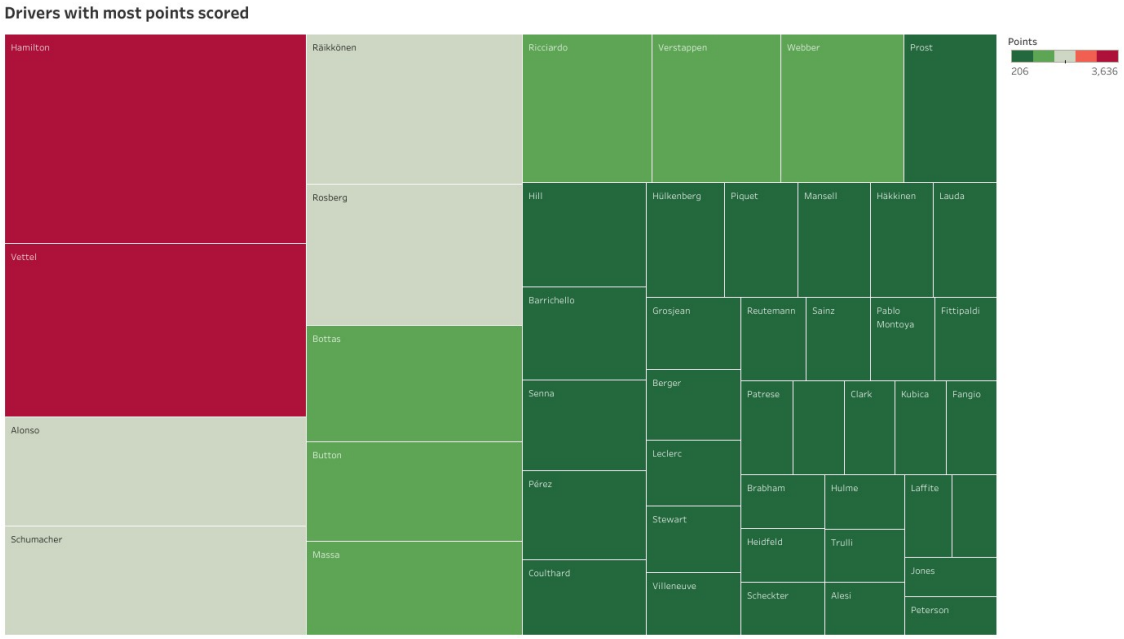

Figure 1: Drivers with most points

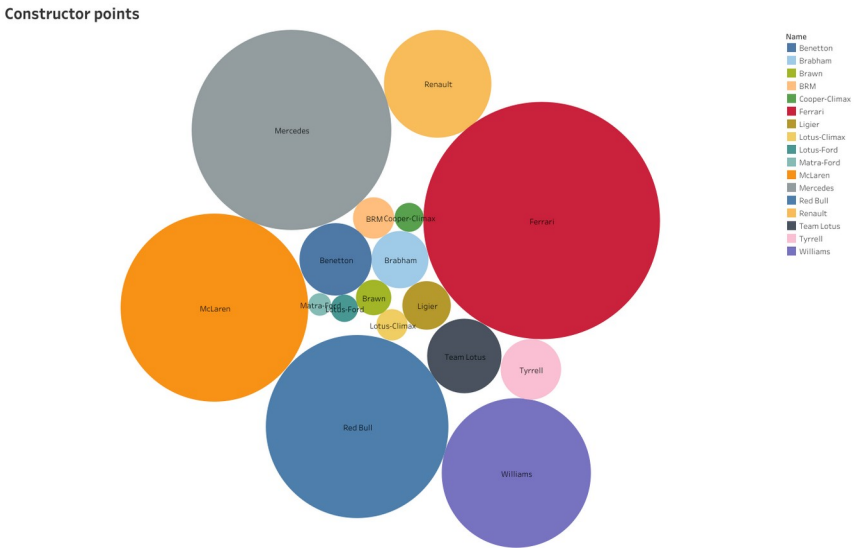

Figure 2: Constructor with most points

Winners and Nationality

As a part of EDA on the data set, we wanted to understand which nationality had produced the most drivers. From Figure 3 it is evident that a big chunk of winning drivers hail from Britain and Germany.

Similarly Figure 4 shows which nationality produced the most number of constructors as well. Again Britain has the upper hand, followed by Italy and Germany for the most number of constructors entered in the sport.

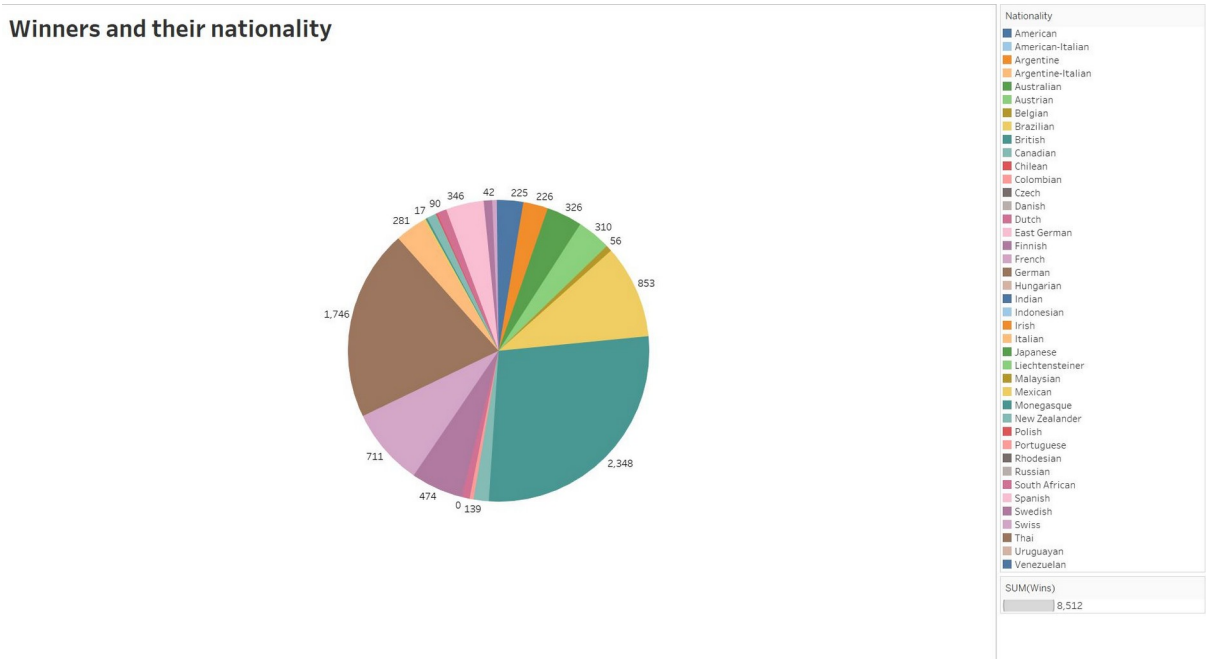

Figure 2: Winners and their nationality

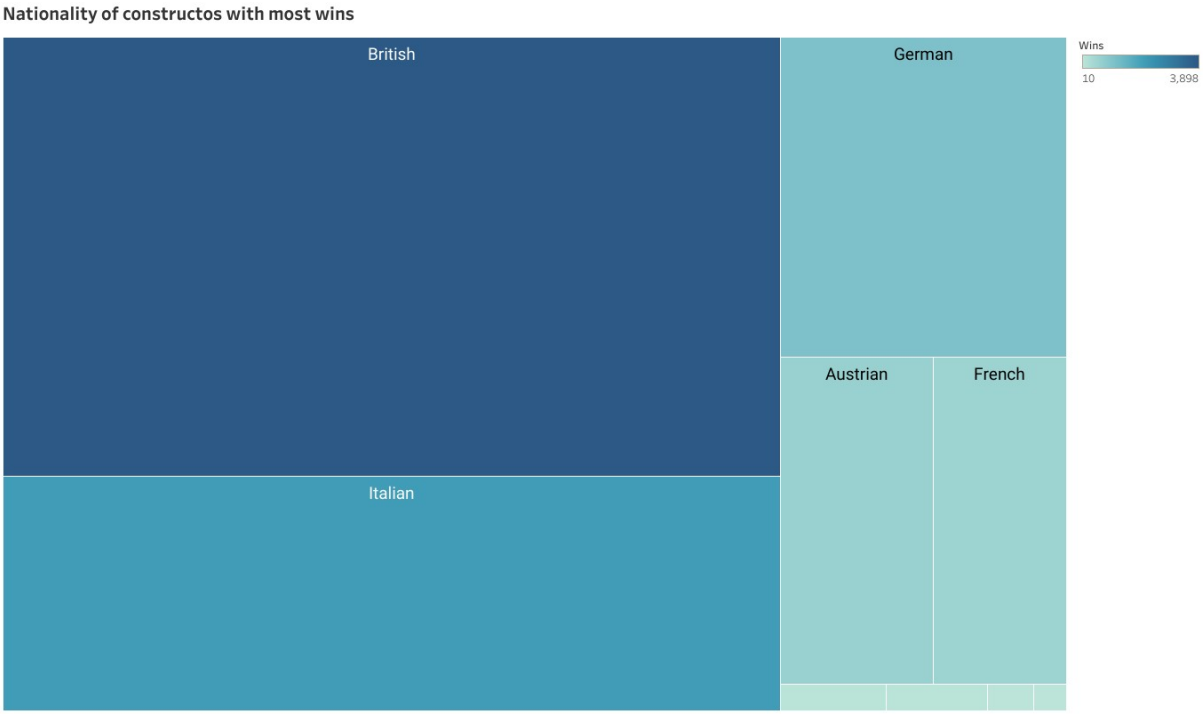

Figure 3: Winning constructors and their nationality

Top 10 Analysis

To understand the time period of the dataset, the following charts were plotted to get an understanding of the top 10 drivers and constructors of all time. This is done because in the long duration of championships that were conducted, many teams left, many teams came under new ownerships and other different reasons. But some of the top teams stayed for consecutive championships like Ferrari, Mercedes and McLaren.

It is important to know which constructors and drivers contributed the most to the statistics. So that we have a range of constructors and drivers to check for when the prediction is done for the 2020 season.

Figure 5 depicts the top 10 drivers of all time, based on their wins, with Lewis Hamilton (HAM), Michael Schumacher (MSC) and Sebastian Vettel (VET) being the top 3 drivers with the most number of wins.

Figure 6 shows the top 10 constructors of all time based on their wins with Ferrari absolutely dominating the sport followed by McLaren and Mercedes.

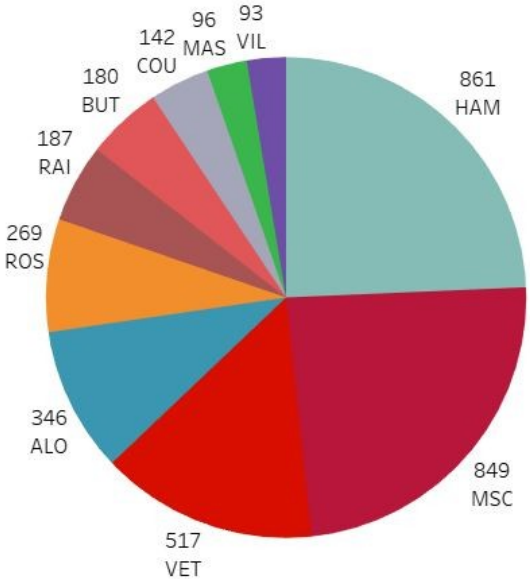

Figure 4: Drivers with most wins - All time

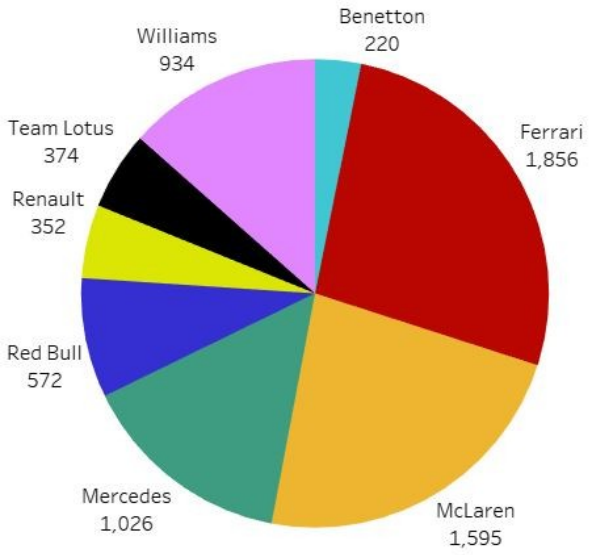

Figure 5: Constructors with most wins - All time

Hybrid V6 era analysis

Formula 1 moved from a noisy V10 engine to V6 engine configuration, a more reliable and powerful power unit (Engine) in the 2014 season. This came about as a major regulation change to curb carbon emissions and improve car reliability and safety. This also paved way for new technologies like the Energy Recovery System (ERS) and re-introduction of the turbocharger.

These changes produced a new era of cars that were faster and more reliable then ever. Figures 7 and 8 below represent which driver and team came out on top after this regulation change.

From Figure 7, it is evident that Lewis Hamilton (HAM) dominated this era and is currently holding the most number of points. Nico Rosberg (ROS) beat him once in 2016 and retired. Hamilton continues to score the maximum points in the other seasons.

Figure 8 shows the constructor championship stats from 2014 to 2019. It is very clear that Mercedes has completely dominated this regulation change and has got everything right every season and have consecutively won all the constructors championship since 2014. Ferrari and Redbull are very close but not close enough to the champions.

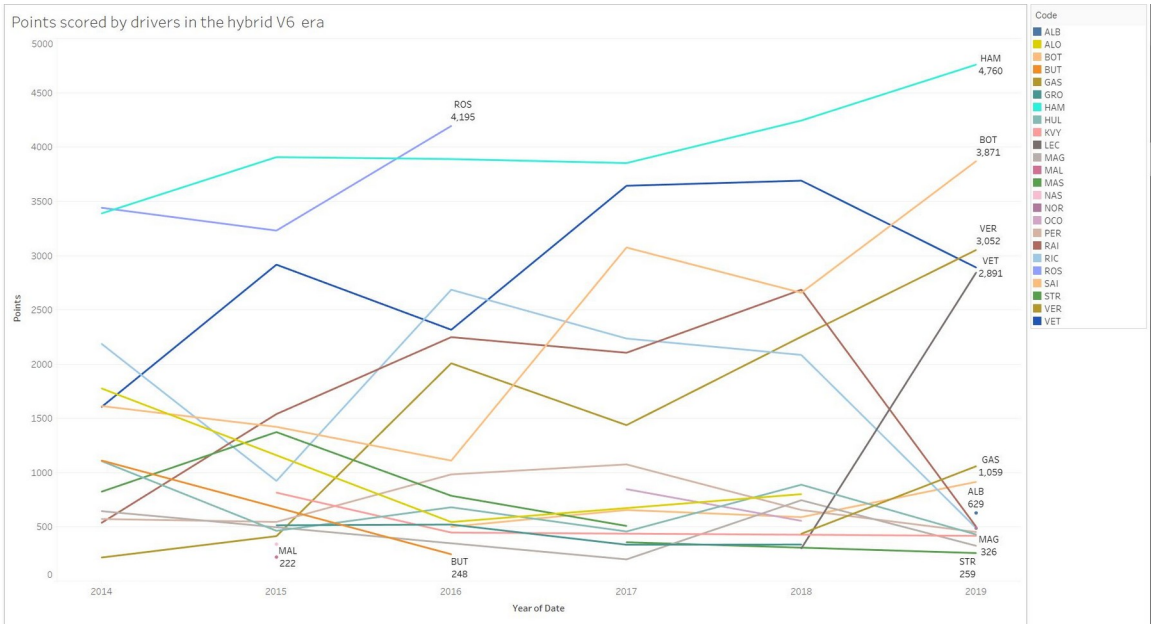

Figure 6: Drivers championship from 2014

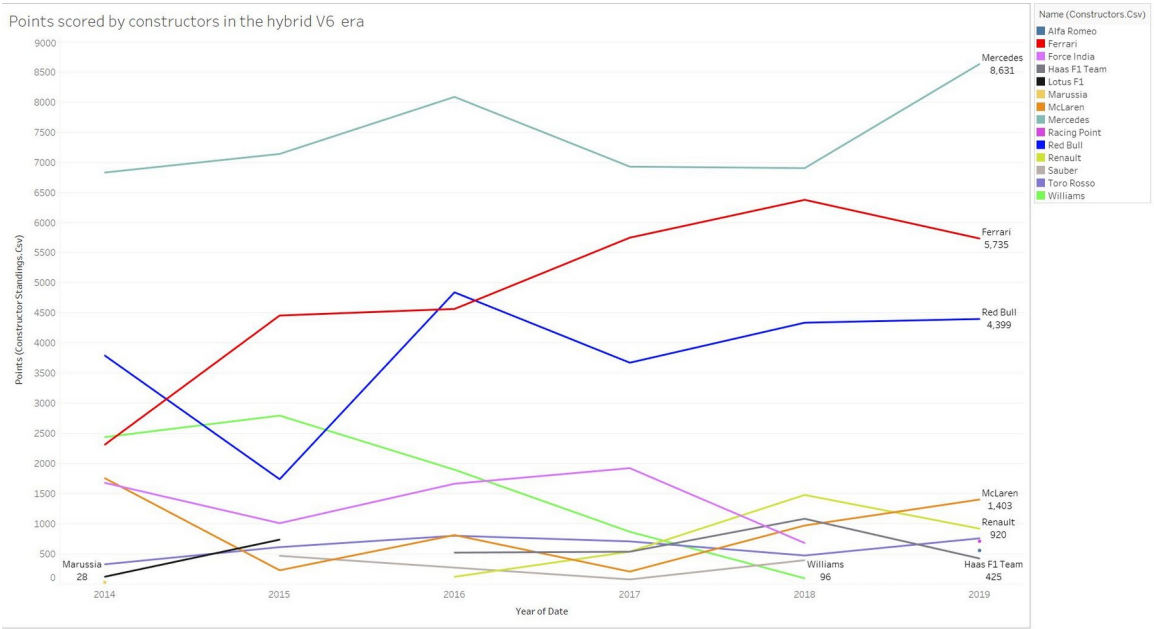

Figure 7: Constructors championship from 2014

2019 Season Analysis

The 2020 season was interrupted even before it could start due to the Coronavirus outbreak. Many races were cancelled until June and Formula 1 released a rather unusual calendar for the 2020 season with new and old tracks to conduct the championship.

=Our goal is to predict what would have been the case of the championship if the normal calendar had taken place. To do that we need to understand how the championship ended the previous season. The following charts aid in visually understanding what went down in the 2019 season.

Figure 9 shows the points that each driver scored in each race. From the figure it is evident that Lewis Hamilton (HAM) won the drivers title followed by Valtteri Bottas (BOT) and Max Verstappen (VER).

This information is important as there have not been any major regulation change since 2014 and the championship is equal. This information will be crucial for the prediction of the 2020 season results.

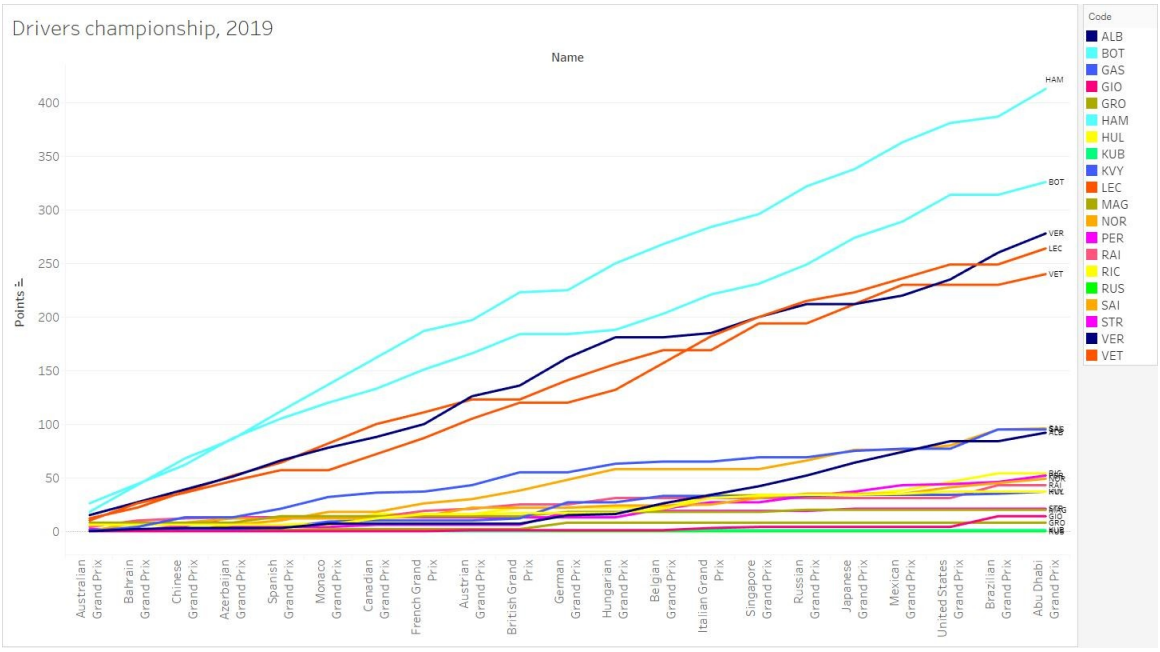

Figure 8: 2019 Drivers Championship

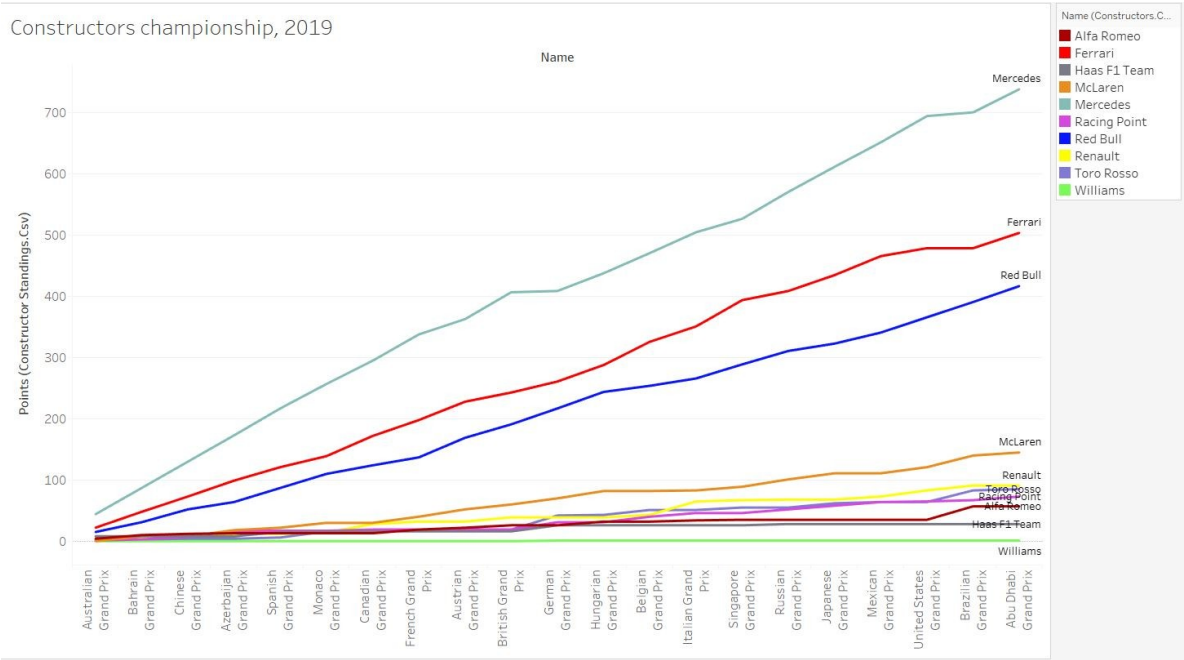

Figure 9: 2019 Constructors Championship