Evaluating the performance of different machine learning algorithms in predicting the outcome of Formula One races, and identifying which models are most effective.

Paul O’Mahony

Thesis submitted in partial fulfilment of the requirements for the degree of Masters of Science in Data Analytics at CCT College Dublin

## Declaration

I hereby certify that this material submitted to CCT College Dublin for the award of Master of Science in Data Analytics is entirely my own work, except where otherwise stated, and to the extent that such work has been cited and clearly acknowledged by references. Furthermore, this work has not been submitted for any other degree.

Signed: Paul O’Mahony

Student Number: sba22425

Date: 22/09/2023

## Acknowledgements

I have received a great deal of support and assistance in reaching this stage of academic research. Firstly, I would like to thank my primary school of Bunscoil Bothar Na Naomh in Lismore for providing the foundations for learning in my adolescence. This development continued to my secondary education in Blackwater Community School, before graduating with a Bachelor’s of Science degree in the National University of Galway, where my knowledge and skills for a professional environment were elevated to a new level.

I would like to extend this gratitude to the College of Computing Technology who have made this thesis possible through hard work and invaluable expertise. This research could not have been undertaken without the belief and inspiration of our lecturers, supervisors and librarians.

Lastly, I am deeply indebted to my parents and siblings for their love and unwavering support throughout all stages of my learning journey. Your belief in me has ensured that my personal growth and motivation remained high during this process and for that, I will always be grateful.

## 

## 

## 

## 

## 

## 

## 

## 

## 

## Abstract

Formula One racing is a fast clock speed industry which pitches the best motorsport teams against one another to produce their best possible product, the grand prix car. The teams must adapt and innovate each season based on factors such as new technology, human resources and regulation changes to remain competitive each season.

The goal is to design a prediction and decision support system for potential use in a professional environment. These challenges were initially addressed with the help of expert knowledge of the domain

By using a combination of the fundamental principles of data processing, statistical analysis and machine learning algorithms, the model focuses on data such as results, drivers, circuits and weather to predict the winner of the races presented.

**Table of contents**

## Introduction

### *Background and Context*

Formula One is considered the pinnacle of motorsport in Europe, where twenty high-speed cars compete against one another on a given circuit. These races have been governed by the FIA (Fédération Internationale de l'Automobile) since its inception in 1950, when the first Formula One race was held. Since then, Formula One has become a multibillion-dollar industry. The ten Formula One teams travel from one side of the globe to another on a weekly basis, transporting their entire set-up which consists from staff to equipment through various modes of transport. These demands on all involved have risen year on year due to the worldwide fanbase for the sport, with the Netflix series ‘Drive to Survive’ being an obvious contributor to the increased exposure.

The series helped to highlight the importance each team places on utilising speed and minimising their lap-time. The best lap-time is invariably achieved by following the optimal racing line. This consists of the shortest path around a racetrack and the points within the track configuration that enable a driver to achieve the car’s highest possible speeds. Within the domain of formula one racing, there are different scenarios and variables which can dictate the outcome of a race and fastest lap-time. Weather conditions will always influence a team’s strategy for a race and how they approach tyre changes during the race.

The Formula One rules require a driver to use two different types of tyres and this rule is waived if there is rain or inclement weather conditions at any stage of the race, where wet or intermediate tyres are needed. Tyre compounds consist of soft, medium and hard. A softer compound tyre usually results in faster lap times but degrades faster than a harder tyre compound. Pit stops are primarily used to replace the tyres on the car due to tire degradation (worn-out tyres) or weather changes, repair broken parts under the assumption that the car does not have to be retired, or to complete a time penalty which may have been issued by the race director. For safety reasons, there is a restricted speed limit when driving through the pit lane which results in a further time loss compared to driving past the lane on a race track.

Other key variables in the outcome of a driver's finishing position in a race include the limit of the car and engine performance, driver’s experience and skill, along with the driver and team’s strategy throughout a race and the decision making process that determines this. The sport has certainly become more data driven in this respect as the more reliable these data predictions are, the better the strategies become. Reliable data is an important factor in strategic decision making and can provide the underlying basis for deploying machine learning techniques to gain a competitive edge for a given team.

### *Problem Statement*

When looking to address the problem of determining which algorithms are most reliable and accurate in predicting the race outcome, the relevant factors must be explored through the historical data available. These include previous driver and team performances at a given track, weather conditions per race, number of pit stops taken and the length of time per stop, to name but a few.

The research problem is, therefore, defined as the lack of clarity around which machine learning algorithms provide the greatest accuracy in predicting race outcomes. These techniques are used by teams to drive their race strategies and gain a competitive edge where possible, but the exact algorithms used or structure of these models have not been divulged to the public.

### *Research Questions*

1. What variables have the most significant impact on the accuracy of the predictive model for Formula One races?
2. What are the potential benefits of using machine learning algorithms for predicting the outcome of Formula One races, beyond improving the accuracy of actual outcome predictions?
3. Could machine learning algorithms be used to help level the playing field between teams with different budgets or resources?
4. Could the use of machine learning algorithms introduce new ethical or privacy concerns in Formula One racing, and if so, how could they be addressed to mitigate any potential consequences?

### *Research Objectives*

Research objectives define the most important tasks completed throughout the research and are motivated by a desire to understand how machine learning algorithms can interpret and predict formula one outcomes based on the historical data fed to the models.

1. Identify the key features that can influence the result of a Formula One race and the weight or importance of each.
2. Transform the Formula One data gathered by applying Machine Learning techniques with respect to these key features to predict the outcome of a race based on historical data.
3. Critique the outcomes achieved through these techniques and analyze their effectiveness in being deployed in a professional environment for a competitive edge.

## Theoretical Research

### *Literature Review*

Formula One is a sport of fine margins where driver and car performance is differentiated to a tenth of a second. A competitive edge for any of these teams over their rivals can pave the way to success throughout the season and improve the finishing position of their cars per race. The introduction of data analytics and more specifically, advanced machine learning algorithms has enabled Formula One teams to process real time data and improve their strategic decision making on a race [1]. The prominence of data analytics in the sport poses the question to whether a race outcome can be accurately predicted and this exploration can be generally divided into three different spaces: psychological, mathematical and machine learning techniques.

When looking at the key concepts and theories around the prediction of sport results, Artificial Neural Networks (ANNs) seem to be the most common unsupervised approach to sport result predictions [2]. An ANN uses interconnected components to translate several inputs into the desired learning output and its power comes from the non-linearity of the hidden neurons, where the weight adjustment here contributes to the final model decision [3] [4]. A study by Bunker and Thabtah [5] [6] examined the application of unlabelled data for sports winner prediction. The ANNs attained a greater accuracy score of 71% once the most suitable metrics for the data had been chosen. [7]

Passfield [8] delves into the emergence of sport analytics and the mix of specialties involved such as mathematics, human physiology and big data. The study provides positive sentiment towards the ability sport science has in reforming knowledge based on player behaviour and sports in general.

The study of Neural Networks in the sport of greyhound racing conducted by Johansson and Sönströd [9] found that the methods performed better at the harder formats and strived for greater value rather than reverting to the simple choice of the favourite in a greyhound race, even when the odds were not supplied to the model for this study. The study demonstrated the complexity at which the Neural Networks perform and their desire to strive for the best possible outcomes when provided with the necessary data to do so.

Another study on greyhound racing predictions is that of Chen, Rinde, Sutahjo, Sommer [10] where two machine learning approaches were deployed, decision-tree and artificial neural networks. This was complemented with three human track experts to compare the algorithm's accuracy against the experts. With the assistance of the experts, the performance variables were reduced from 50 to 10. The results showed that both the decision tree and neural network methods outperformed the human experts for expected profit.

Davoodi, Khateymoori [11] also applied artificial neural networks to predict horse racing results. Data from 100 races at the Aqueduct racetrack in New York from 2010 were selected, where one neural network was used for each horse in a race. The eight features consisted of the horses weight, class of race, horse trainer, jockey, race distance, number of horses in the race, track conditions and weather [12]. These eight features made up the input layer, two hidden layers, and the output layer defined as the horse's finishing time. With over 400 epochs in place, an accuracy of 77% was achieved.

When looking at studies based on athletics and how this data is interpreted based on feature selection, Ofoghi, Zeleznikow, MacMahon, Rehula, and Dwyer [13] published a study regarding the prediction of triathlon results. It emphasised the importance of the data processing phase for this research and the relevant factors that can influence the winner’s performance. When importing the data, the researchers learnt that converting the time variable from HH:MM:SS format to raw seconds helped to emphasize the dominant performance of the leader and other participants against the field.

In the paper of Przednowek, Iskra and Przednowek [14], a similar study is conducted on the athletics discipline and specifically, a 400-metres hurdle race. The data of 21 athletes from the Polish National Team is examined through nonlinear methods [15] and the artificial neural network returns a prediction error of 0.72 seconds, with common predictors for the results also specified for these results such as age, speed endurance and other impacting factors.

Edelmann-Nusser, J.; Hohmann, A.; Henneberg, B. [16] explored the performance of female swimmers in the finals of the 2000 Olympic Games in Sydney by also using nonlinear neural networks (multi-layer perceptrons) through training data from 19 competitions prior to the event, which they found was insufficient in accurately training such networks. [17] The neural models were validated using the ‘leave-one-out’ method, resulting in effective risk-aversion and an error prediction of 0.05s. [18] [19]

Abut, Akay, Daneshvar, Hei [20] undertook the challenge of predicting the racing time accuracy of cross-country skiers. The study concluded that the 3 artificial neural networks used; Multilayer Feed-Forward Artificial Neural Network (MFANN) [21], General Regression Neural Network (GRNN) and Radial Basis Function Neural Network (RBFNN) [22], are fit for making predictions for skiing events due to acceptable error rates from these models.

Another study conducted by Sankaranarayanan, Sattar, Lakshmanan [23] analysed the prediction of a cricket game outcome through regression models while the game was in progress. The application of Ridge regression and attribute bagging algorithms helped to provide a prediction accuracy of 70%, higher than any reported studies in cricket mining. Factors that could influence the game outcome such as the toss, batsmen historical data and other parameters through the use of linear regression and nearest neighbour clustering algorithms [24].

Bailey, Clarke [25] also explored machine learning methods to predict the outcome of a cricket match between two teams and their respective features. The feature selection within this study is relatively small and does not offer a direct comparison between this study where ten individual teams will be examined and the feature selection is expected to be higher. [26]

Various studies are available within the sphere of European football and the machine learning methods applied to this data. Tax and Joustra [27] researched a result prediction system for the Dutch Eredivisie league by testing different combinations of dimensionality reduction techniques and classification algorithms. Joseph, Fenton, Neil [28] published a study based solely on the prediction of game results for the Tottenham Hotspur team during the years of 1995-1997 using a Bayesian network which was constructed by a domain expert and found that the domain knowledge [29] of this expert helped the network to outperform machine learning models that are constructed based on data analysis.

Goddard [30] approached the football prediction challenge from a different angle to most by comparing the modelling of goals scored versus the modelling of win-draw-lose results, rather than just using the previous match results. He concluded that a hybrid model achieved the best possible prediction based on the 25 years worth of data collated from English football matches. The features chosen here include the importance of certain games and the geographical distance between the two opponents.

Hucaljuk [31] used data from the Champions League tournament to set a target accuracy of 60% from their Naive Bayes, K-nearest neighbour and Random Forest techniques. The results outperformed this target accuracy by almost 9% but Hucaljuk outlines the possibility of improving feature selection to include individual player data in the quest to achieve greater results in future works. This would obviously increase the workload significantly and require more resources to achieve this but has greater implications for overall accuracy return from the algorithms.

Broich [32] used data from 153 games within the German Bundesliga to highlight the importance of efficiency when looking at the significance parameters for a match outcome [33]. The number of shots and passes taken by a team, along with the number of touches in the game were deemed important variables also. The study also provided insight that the expected impact of the distance covered by a team did not significantly influence the winning attribute.

Kampakis published two studies which looked at data in relation to the English Premier League. The first study [34] used Twitter posts to predict the outcome of these Premier League matches and proved that the social media platform has useful information for this objective. The author reported that the Twitter-based model performed better than historical data and statistics but increased its accuracy further when mixed with other prediction models.

The second study [35] involved collaborating with two Premier League clubs to predict player injuries in football. This included using machine learning algorithms to predict the recovery time needed for a specific injury, where correlation-based feature selection was used to improve the performance and reliability of this model. The prediction of a player getting an injury was studied through the relationship between the hours of training and matches played by the player with the number of injuries recorded. With the consent of the club, the GPS player data was analysed to see if overtraining or fatigue were contributors to an injury and also to predict if a player is threatened by injury based on the collated data.

Both Rue [36] and Pollard [37] deployed the iterative simulation technique Markov Chain Monte Carlo [38] to predict the outcome of English football matches. A dynamic structure was used in their models to facilitate the updating of parameters once more game results were fed into the existing dataset.

Ulmer, Fernandez, Peterson [39] discussed the use of machine learning algorithms within Premier League data to predict match outcomes based on the 2012-2014 seasons. The collated results confirmed that the outcome of a draw within these results were hard to predict and these predictions were detrimental to the accuracy of the prediction model. The algorithms used included stochastic gradient descent [40], linear support vector machines and random forest [41].

McCabe, Trevathan [42] took a broader approach to their sports prediction results through four sports; Football, Australian Rules Football, Rugby Union and Rugby League. The artificial neural network had 20 input nodes, 10 nodes in the hidden layer and 1 output node. The feature selection consisted of features that were common to all sports included, with any specific to just the one sport not included [43]. The algorithm returned an average performance of 67.5% in this study.

Haghighat, Rastegari, Nourafza [44] conducted a similar study into the data mining techniques for predictions of sport results. They discuss various classification methods such as Artificial Neural Networks, Support Vector Machines, Bayesian Method, Decision Trees, Fuzzy System and Logistic Regression [45] [46] [47]. The study outlined the advantages and disadvantages of each method used and could not provide an overall best method due to the different sports tested in the paper.

The use of American sport data has also been included in this review due to its wide availability and both Purucker [48] and Khan [49] used statistical data from the National Football League in America to deploy Artificial Neural Networks for the prediction of NFL games, to which the outcomes proved both equal and better than the experts respectively through a combination of different classification algorithms.

At college level, Leung and Joseph [50] challenge the use of traditional statistical comparison between the two competing teams of that given year, as players' skillsets are more interchangeable due to the turnover of players at this level. They adopt a prediction method based on the historical results of American college football teams overall. Another college football study by Delen, Cogdell, Kasap [51] found that regression-based predictions were better than classification-based ones by roughly 11% with the data provided.

Talukder [52] explored the possibility of injury within the National Basketball Association (NBA) players and used machine learning techniques to create a model for injury prevention during a game. The model achieved high accuracy figures for short-term injury prediction and ranked injuries based on the economical expense caused and the impact a player would have on the team’s performance if he was not strategically rested instead.

Kuehn [53] devised a framework for evaluating NBA players skillset and how these compliment the overall performance of the team. A probabilistic model for possessions, which considered factors such as the opponents lineups and in-play events revealed that players are mostly considered on their individual statistics in the sport and does not place enough emphasis on their complementary contributions to the overall team performance.

Wiseman [54] used various regression algorithms to predict the winning golf score on a PGA event based on the scores after round 1. Through linear regression, neural network, Bayesian linear regression and decision tree regression, the prediction of the winning score was accurate 67% of the time [55]. Feature selection included the round 1 leading score, round 1 average score, course par, major event and course yardage. The results were validated on 2016 scores based from the models training of data from 2004 to 2015.

Although the study of Formula one is limited, there is motorsports literature available through the examination of American NASCAR racing to support this. Pfitzner & Rishel [56] delved into the possibility of reliable predictors existing in the outcome of a race through correlation analysis and found that certain factors such as qualifying speed, pole position and driver experience. One of the main conclusions from this study was the effect of team membership i.e., number of team members being two or three, however, this will be irrelevant in this study due to the number of drivers per team being fixed to two. A similar study performed by Allender [57] found a correlation between the starting position of each driver and their experience to their finishing position in a race through the regression model used. This study questions the element of flair over experience which will be an interesting point of exploration for this research.

Depken, Mackay [58] offer an insight into the structure of a team motorsport like Formula One when they analysed the NASCAR Sprint Cup Series from 2005 to 2008, using multiple regression models. The conclusion from this was multi-car teams perform better in overall team finishing positions and cooperation.

The research papers so far have focused on data generated from the players within their respective sports. In Formula One, however, the car itself has more than 250 sensors attached when competing in both qualifying and championship races. In a recent video created by the Mercedes-AMG Petronas team on their YouTube channel [59], they discuss their typical data generation output over a race weekend. Up to 1TB (terabyte) of data is recorded per car and they estimate 30MB of live data is generated per lap. The sensors also vary in function to comply with regulatory requirements for a tyre pressure monitoring system, along with providing real-time data on temperature, speed, tyre degradation and pressure.

An article published by Alex Woodie [60] also features the head of IT with the Mercedes team who highlights the volume of data generated through these sensors over a given season. With over 10TB of data being collected during this time, the data is stored using cloud storage array solutions and poses an issue with how many eyes can sift through this data. Instead, the team has adopted an approach to look at the anomalies within this data to gain greater insight and drive their decision making throughout the season.

To continue the focus on the collective contribution, Bell, Smith, Sabel and Jones [61] delve into the efforts of both driver and team and reach the conclusion that the team effect greatly outweighs the drivers effect (86% contribution) and quantifies the increase in driver changes, where as team effects have a ‘legacy effect’ and are steady based on accumulated experience.

Eichenberger, Stadelmann [62] published a study on the best Formula One driver over a 57 year period, from 1950 to 2006. The authors took the view that driver performance is dependent on the capability of their car and their innate talent. Through the application of linear regression techniques, variables such as collisions, engine failures and disqualification were removed from the study. Two control variables were also introduced to provide balance and robustness, that being the team partner driver and home advantage for a driver when racing in their own country.

A comparative analysis between both Formula One and NASCAR has been conducted by Silva, Silva [63] to outline both the similarities and differences between both disciplines. Qualifying and race data from the 2009 season is explored to achieve this. One clear difference between the sports is the fact that NASCAR has double the number of races and drivers in a season. The similarities include the point scoring system for driver and constructor championships, as well as the qualifying laps scheduled before race day to determine each driver's starting position on the grid. Four key variables were defined to gain insight into race predictions; Qualifying, Practice, Points and Results. The research confirmed that the Qualifying variable, which held the results of each driver’s qualification time, had the greatest significance in predicting Formula One results. To compare this with NASCAR, a combination of the Practice and Result data formed the strongest prediction for NASCAR. One key difference to note here is the rules around rain occurring during a NASCAR race, where the race event is cancelled.

Graves, Reese, Fitzgerald [64] used a Bayesian hierarchical framework to analyze the NASCAR racing results. They assess the driver’s abilities and predict their future standings based on the rate at which their skills have improved throughout a given season. The study also provides insight into the existence of track specialists - some drivers perform much better on certain tracks than others.

When looking at factors that can influence the outcome of a Formula One race, the study performed by Biemann, Liu, Zeng and Huang [65] used reinforcement learning to identify recurrent patterns and LSTM (long short-term memory) architecture to gain achievements in the model outcome. This study offered an alternative approach to addressing the weather calculations for this study which was not approached in the other literature.

Choo, C.L.W. [66] published a research paper which compared the insights gained from the work of Tulabandhula, Rudin [67] which focused on pit stop strategy and tyre changes based on a given track within the season. The suggestion of adding more features surrounding pit stops and pit crew performance for further insight was a novel suggestion here.

Aversa, Cabantous, Haefliger [68] collated data from the infamous final race of the 2010 Formula One season, where Ferrari’s decision support system provided incorrect conclusions and in turn, influenced their race strategy to deny Fernando Alonso a world title as expected. The Ferrari team chose to align their pit stop strategy with their rivals at the time, Red Bull, and pitted the car shortly after their competitor. This non-data driven strategy forced Alonso to re-join the circuit in the midst of a condensed field and hinder his ability to progress past other cars and ultimately, lose the expected championship title.

Looking at the relationship between Formula One and technology and referencing the research question of the potential benefits machine learning algorithms can have on the decision making in the sport, Jenkins [69] investigates the relationship between a team’s competitive advantage and their ability to adapt to changes in technological knowledge or discontinuities. The study finds that competitive advantage is achieved through a team's capability to pivot through successive technological discontinuities. The two capabilities are defined as dynamic and sustaining capabilities.

Aversa, Furnari, & Haefliger [70] link to this study when analyzing the relationship between the business model and racing performance. The business models in Formula One are broken down into four; Internal Knowledge Transfer, External Knowledge Transfer, Formula One Supply and Talent. The four models evaluate the collaboration between Formula One teams and manufacturers, the sales of technology externally to other industries, the transferability between Formula One teams regarding supply and the scouting systems set for investing in both existing and future talent.

The authors focus specifically on the Red Bull and Williams racing teams and place the spotlight on their key factors such as financial, technological development and knowledge and human resources. The conclusion offers positive sentiment to the Red Bull strategy to invest in younger drivers and allocate costs to technological advancements instead.

Rosso, Rosso [71] examines the relationship between various race factors such as the weather and tyre types from the data distributed after the 2016 Monaco Grand Prix. They applied quantile regression [72] to study these factors in greater detail and the role the weather conditions had to play on the outcome of this specific race.

Sundar [73] provides a great insight into the 2005 Monaco Grand Prix.Kimi Raikkonen, a Finnish driver. was the fastest driver on Saturday’s qualifying and started the race from first place. He was competing with the title contender Fernando Alonso at the time in second place. It was on lap 23 when Minardi’s driver Christijan Albers crashed into a wall and the Safety Car was deployed. It is an unwritten rule in F1 that when the Safety Car is on track, the team strategy should always be to pit their cars. This is due to the fact that cars have to slow down once a safety car has been issued and follow this car around the track at a reduced pace, thus, reducing the impact of time in the pit. Alonso pitted immediately as it was expected, but surprisingly Raikkonen stayed on the track. This decision was made by McLaren chief strategist Neil Martin, after assessing all the data available. This decision to stay out on track was made within a minute of receiving the notification that a safety car had been deployed. Through the use of data analytics in this scenario, the team chose to take a risky decision and go against the grain of pitting the car. In the end, the decision bore fruit as Raikkonen gained a 35 second lead on Alonso and could afford to pit later in the race for the advantage of new tyres and still keep a 13 second lead when doing so. It proved an excellent decision post-race and is a prime example of data-driven decision making within sports.

Finally, a recent study conducted by Amazon Web Services [74] provided an insight into their partnership with Formula One and how the use of Machine Learning techniques and high performance computing are transforming the sport. It highlights the millions of data points being generated per second from the 300 sensors attached to a Formula One car and how the sport is becoming data-driven. The study also provides background analysis to factors which have been identified for this research such as tyre and driver performance. However, the machine learning algorithms used have not been disclosed due to their partnership with the sport and do not give guidance to the setup, training and evaluation of these models.

The literature mentioned above emphasizes the importance of feature generation and having domain knowledge in order to successfully apply this. The use of Neural Networks were applied in most of these studies and has increased in popularity for a prediction model similar to the one being proposed here. Regarding the literature available for Formula One, there are few studies published but the NASCAR motorsport studies provide some insight into the algorithms that could be used. The NASCAR studies do not provide the necessary parallels in terms of whether a driver should make a pit stop since a single tire compound is used and this would lead to a new approach being taken for the application of Formula One racing.

**Experimental Research**

### *Research Methodology*

Finding the appropriate methodology for this thesis involves considering all existing methodologies and the first of these is the KDD (Knowledge Discovery in Databases) methodology. This iterative process involves extracting novel and potentially valuable information from datasets, while establishing patterns and relationships within the data. The advantage of this approach is improved decision making once the process techniques have been applied;

1. Data Cleaning
2. Data Integration
3. Data Selection
4. Data Transformation
5. Data Mining
6. Pattern Evaluation
7. Knowledge Representation/Visualization

As mentioned above, the focus within these stages of the process is discovering useful information from the data and domain expertise is a key facilitator in choosing the appropriate data and results.

The SEMMA methodology is almost identical to the KDD methodology in that every stage of KDD corresponds to a stage of SEMMA. The SEMMA development tool for data mining enables smooth data mining execution through the following steps;

1. Sample
2. Explore
3. Modify
4. Model
5. Assess

In order to get an understanding of the datasets for this research, the sampling process is initiated. Exploration of trends within the data are investigated to gain further insight and any necessary data modification is performed, resulting in variable creation. Once the data has been cleaned and variables refined, a projected model is produced through a variety of data mining techniques. From here, the model is assessed for its reliability and tested for performance efficacy.

The third methodology to consider is CRISP-DM (CRoss-Industry Standard Process for Data Mining). This methodology proposes a standardization approach to assess the potential of data mining projects and identify potential irregularities or flaws in the model in a timely fashion. The model puts forward the following steps;

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

The model is mostly effective in a complex environment where a fast rate of change occurs.

The final methodology to propose is the Sports Results Prediction CRISP-DM framework of Bunker and Thabtah (2019). As expected, the methodology is based on the above CRISP-DM framework but specializes in its adaptations to a sports environment. The steps are as follows;

1. Domain Understanding
2. Data Understanding
3. Data Preparation & Feature Engineering
4. Modelling
5. Model Evaluation
6. Model Deployment

The ‘Domain Understanding’ step has been renamed to highlight the study of a sport domain instead of a specific business. ‘Data preparation & Feature Engineering’ has also been renamed in the methodology to focus on similar features in the subsets of sports data. The final three modelling steps advise splitting the data between training and test sets for model prediction accuracy purposes and retraining the models to predict new sports results in the process.

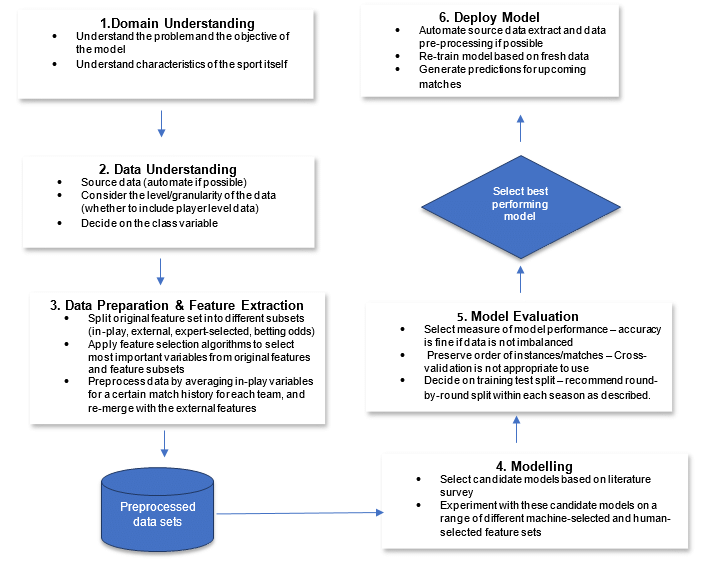


Figure 1. SRP-CRISP-DM methodology

Given the specific fit that SRP-CRISP-DM provides for sport analytics and its focus on creating race-related features within the data, this methodology was chosen for the dissertation.

### *Domain Understanding*

Bunker and Thabtah (2019) have defined the domain understanding step as “comprehending the problem, the goal of the modeling, and the specific characteristics of the sport itself”. To comprehend the problem of predicting the outcome of a race through machine learning techniques, knowledge of the sport and what factors potentially determine the race results are important.

As a keen Formula One fan for the last eight years, the release of the Netflix series ‘Formula One: Drive To Survive’ provided fans with exclusive access to the Formula One season both on and off the track. From a domain understanding point of view, the series highlights the work of all members of a constructor to gain a competitive advantage and the tactical prowess of the team principals when gaining a favourable result.

### *Data Understanding*

The first step to begin this research was to identify and validate historical data sources for Formula One racing. Ergast (<http://ergast.com/mrd>) is an open-source database where users can extract structured race data from as far back as 1950. The Ergast API allows for communication between applications to retrieve this data and a supporting database guide provides the attributes of each available table and their numerical or categorical variables within. The data consists of important data such as race circuits, race results, driver performance and team standings throughout a season.

The API did not include an important factor in race prediction which is the weather conditions on the day of a race. Not only does the weather conditions impact the overall speed of a Formula One car around a circuit, but can influence the strategy of each team for pitting their drivers at a certain time during a race and deciding which tyre compound to use also. The tyre compounds provided to each Formula One team consist of soft, medium and hard, along with wet tyres, for severe conditions.

The datasets append a URL link from Wikipedia as a unique identifier for each race. This enabled the use of the Python library BeautifulSoup to extract keywords from the race information such as wet, cloudy, warm, cold and dry. For example, if BeautifulSoup detected keywords rainy, pouring or slippery, a binary value of 1 was assigned to the weather\_wet column created in the dataset. Otherwise, a value of 0 was encoded into the column for this race.

#### *Data Dictionary*

| Dataset | Dataset Content |
| --- | --- |
| Circuits | The country and city where each race took place, along with an identifier for each circuit |
| Constructor Results | The results of each team in respect of their two cars per race. |
| Constructor Standings | The standings of each Formula One team after a race |
| Constructors | Provides the team name, country of origin and unique identifier to link the team drivers to the constructor |
| Driver Standings | The standings of each driver after a Formula One race |
| Drivers | Information on each driver such as their name, nationality, date of birth |
| Lap Times | Begins with data from 1996, the dataset provides each racers lap time per round. |
| Pit Stops | The length of time taken to perform each pit stop per driver in a race. |
| Qualifying |  |
| Races |  |
| Results | Shows the results of each race and the podium finishes. The podium finishes and race winners are important for this study. |
| Seasons | Includes all the racing seasons from 1950 and a unique URL to the Wikipedia link of each. |
| Sprint Results |  |
| Status | Provides confirmation if the driver completed the race or experienced a race-ending issue such as an accident, collision or engine failure. |
| Weather | Created using BeautifulSoup to identify keywords in the Wikipedia page to establish wet, cloudy, warm conditions. |

#### *Data Exploration*

Exploring and visualising the data gathered can unearth insightful areas or patterns from the start. Given the vast amount of information contained in these datasets, data exploration and visualisation can quickly determine the most important areas for analysis. Variable identification and analysis, along with detecting outliers and missing values, form a strong basis of this descriptive analytics journey.

The first figure presented is the number of circuits per country since the sport’s inception in 1950. This provides an insight into the global scale of the sport and the number of race tracks allocated per country to support the industry.

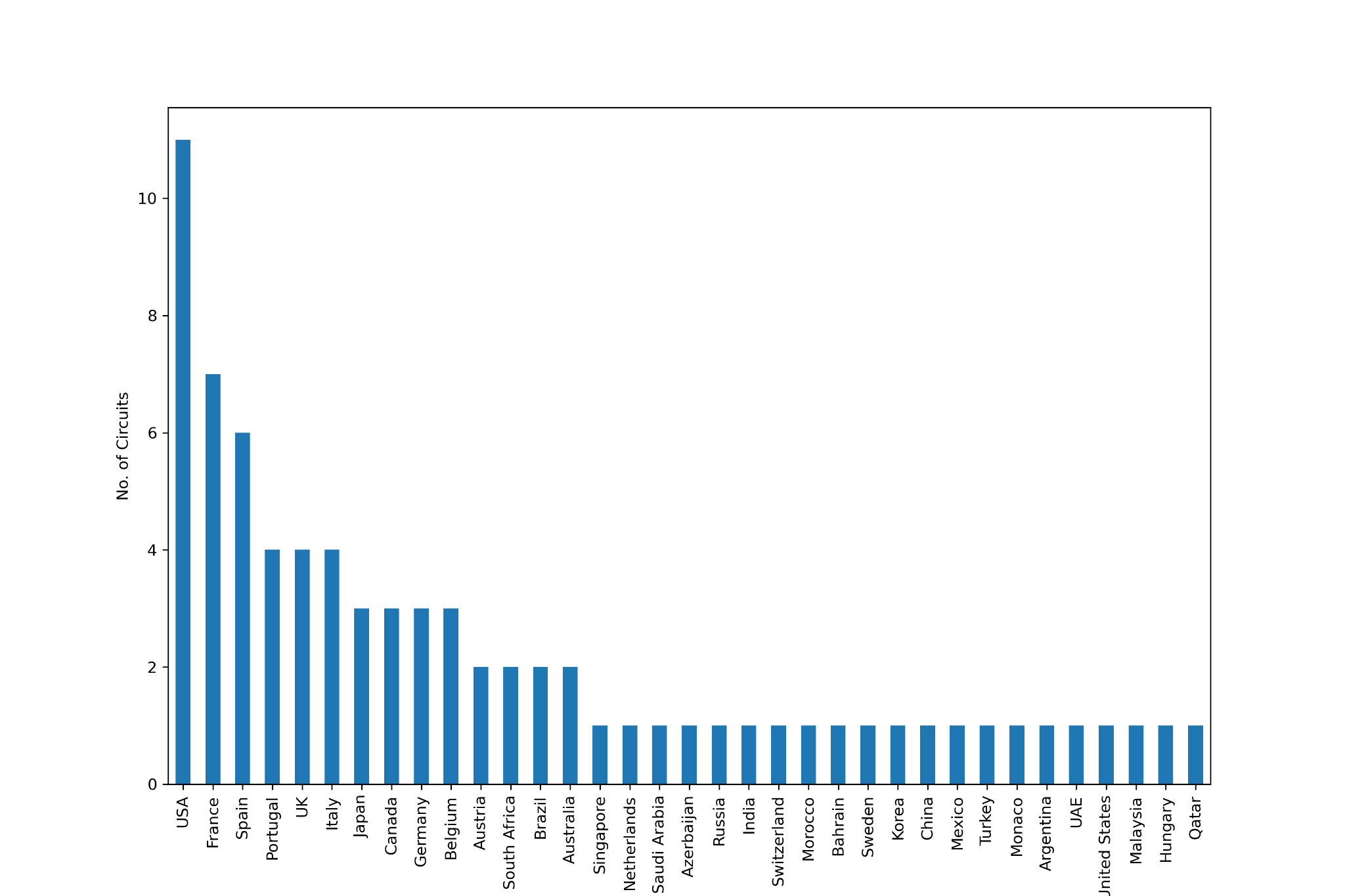


Figure 2. Number of circuits per country

The USA has hosted Formula One races in 11 different circuits and is clear of France and Spain respectively. For the 2023 season, the introduction of the Miami GP in 2022 has remained and the Las Vegas GP has been added to the calendar - a track which has not hosted a race in over 40 years.

Some of these circuits have only seen a handful of races being hosted throughout their existence and fans of the sport would be expected to name some iconic tracks such as Monaco and Monza if asked to name one circuit in the Formula One calendar. The figure below confirms that these two race tracks are the highest in the number of races hosted since the sports inception. Silverstone in the United Kingdom has ranked third highest in this figure, with the Spa race track in Belgium following in fourth.

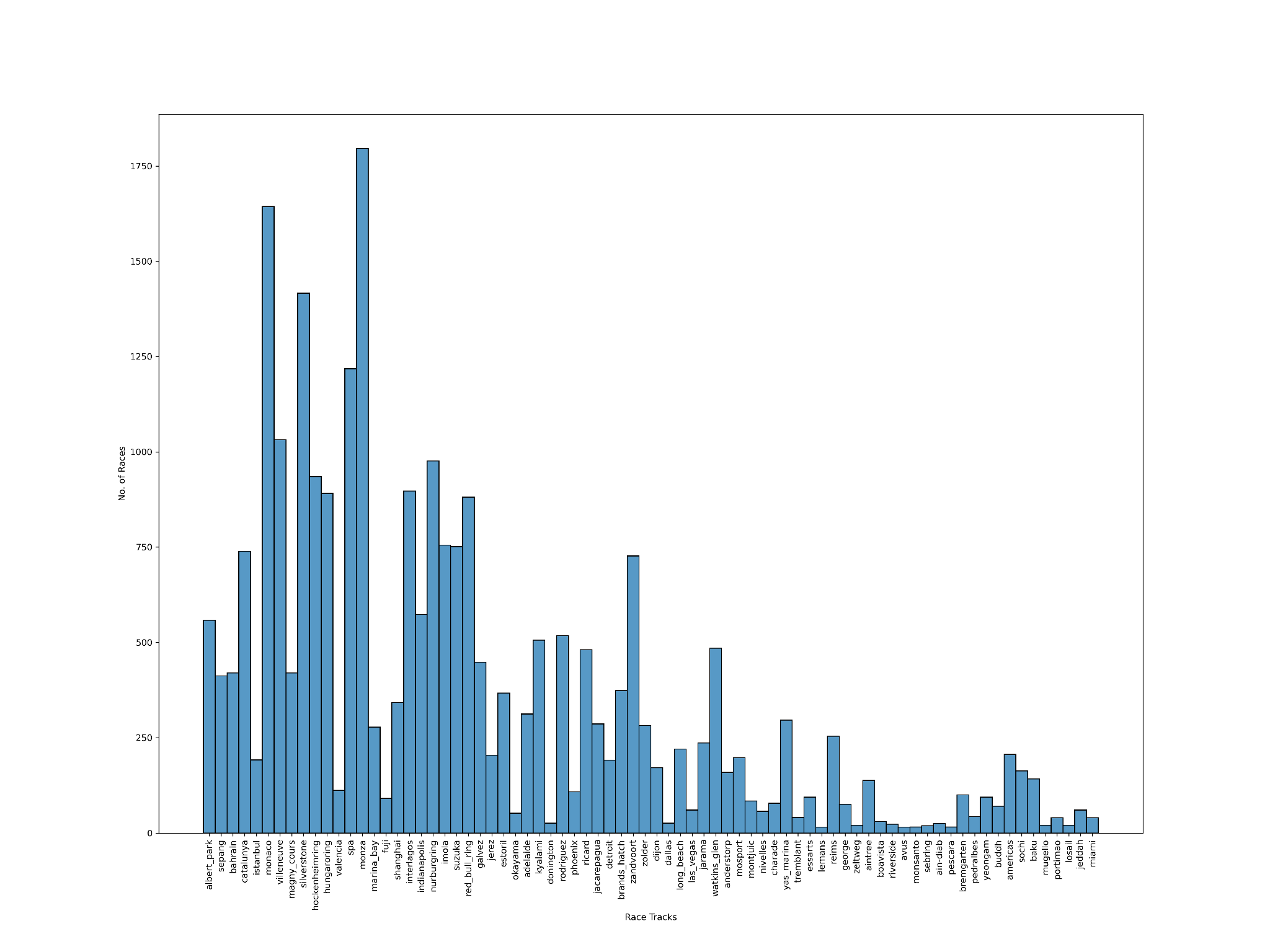


Figure 3. Number of races per track

The Silverstone track is the oldest track in the Formula One calendar and the United Kingdom is currently home to six out of the ten teams - Red Bull, Mercedes, McLaren, Aston Martin, Alpine and Williams. The reason for this is due to the abandoned airspaces post World War II and ex-RAF engineers being available to begin working in Formula One factories.

The number of races per season were raised as a point of concern for driver and team welfare last year. The demands on travelling the globe and transporting all necessary equipment require high-level planning and very little margin for error. The figure below emphasises the steady increase in the number of races throughout the years and the reduction of races in 2020 to seventeen as a result of the COVID pandemic. It also highlights that the current 2023 season is joint-highest with last season's calendar of 22 races.

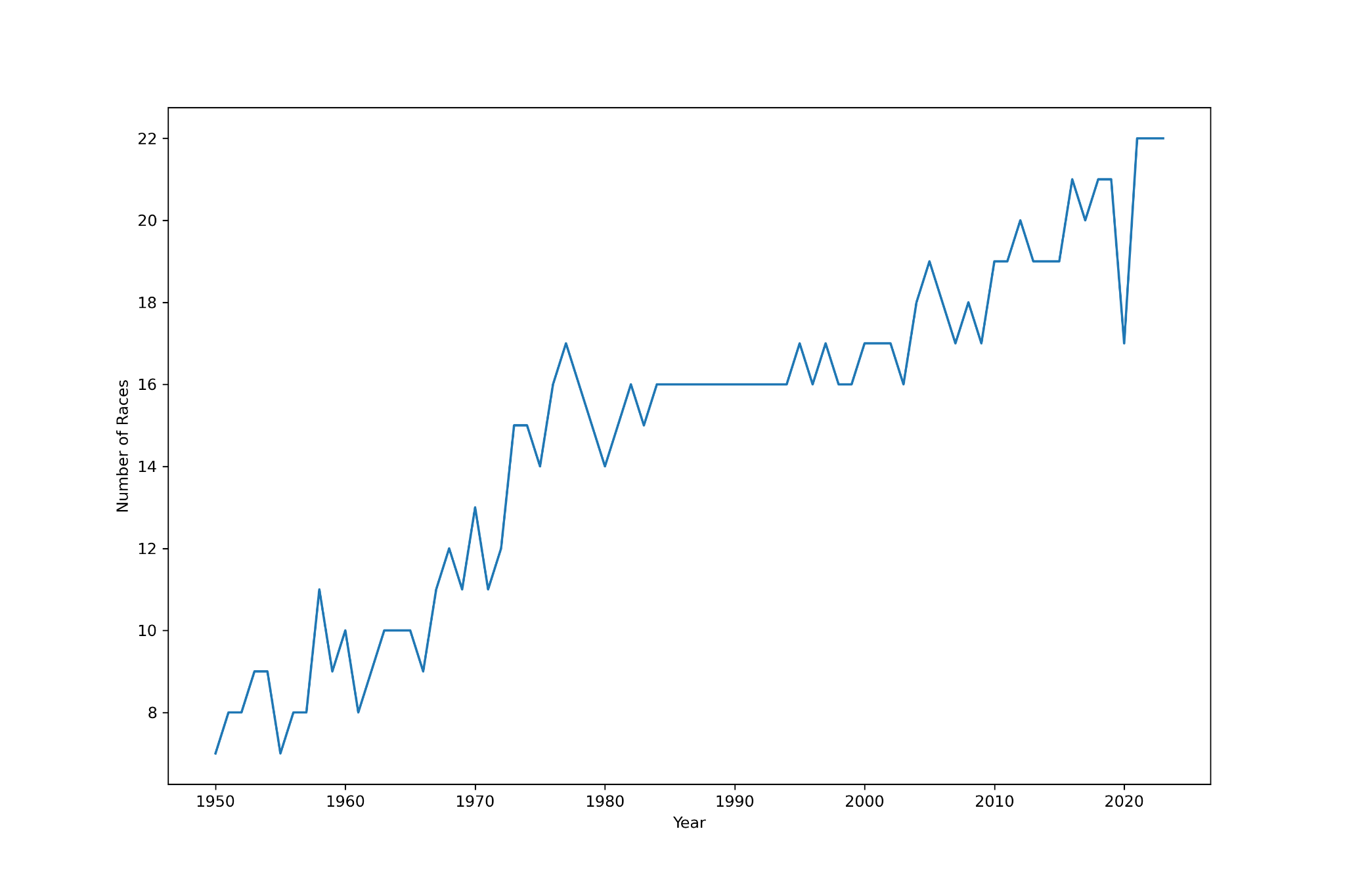


Figure 4. Number of races per season

In Figure 5, we observe the importance of starting from pole position in a race and the results earned by drivers when doing so. In most cases, starting from first place results in winning the race or otherwise, achieving second or third place.

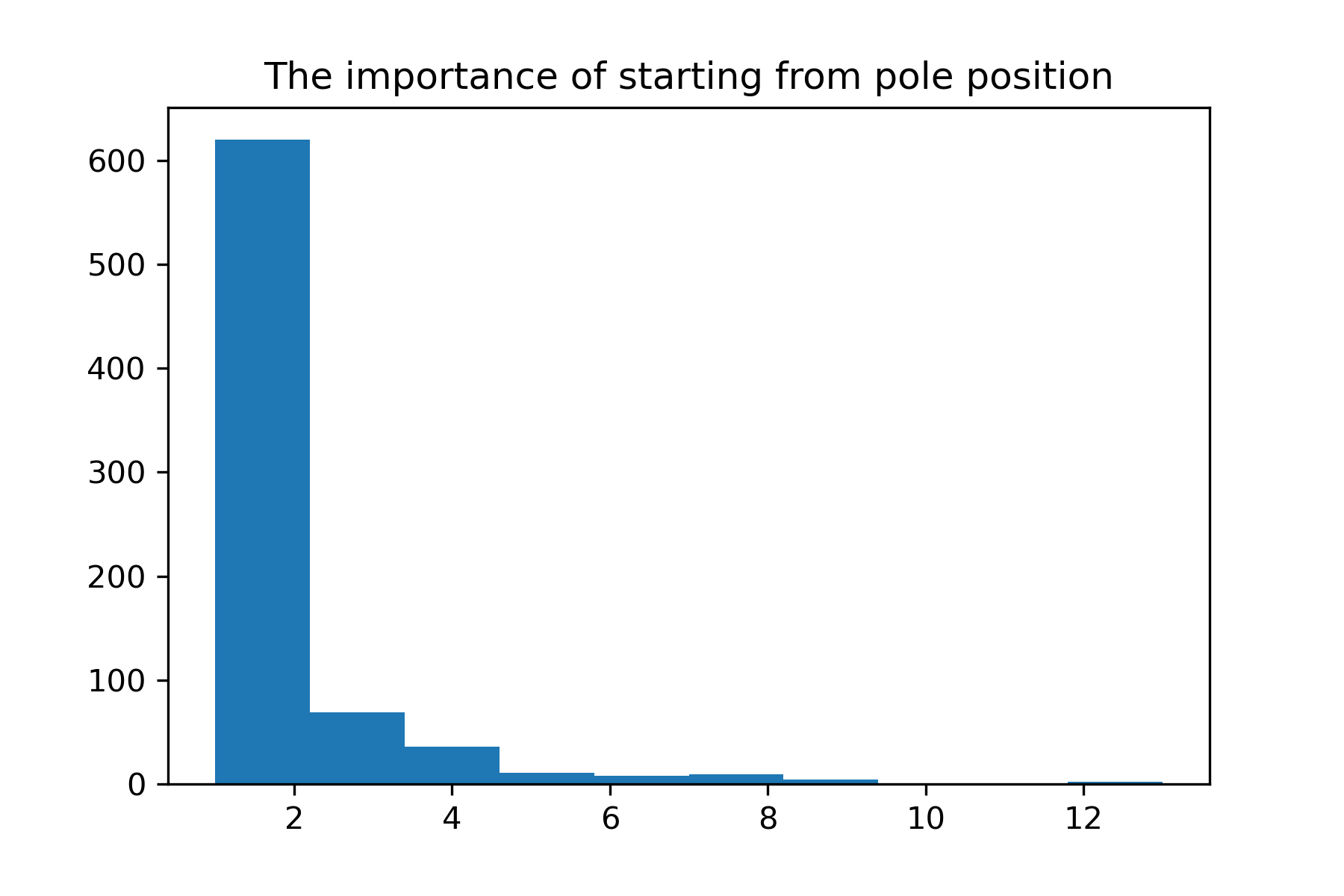


Figure 5. The importance of starting from pole position

Starting in pole position means that no other cars are in front of you to overtake and certain tracks are known to be difficult to execute overtakes during a race.

One such track is the Circuit de Monaco, which is a street circuit laid out around the city of Monte Carlo. The track is narrow and has tight corners as expected and is considered to be one of the greatest tests a Formula One driver will face.

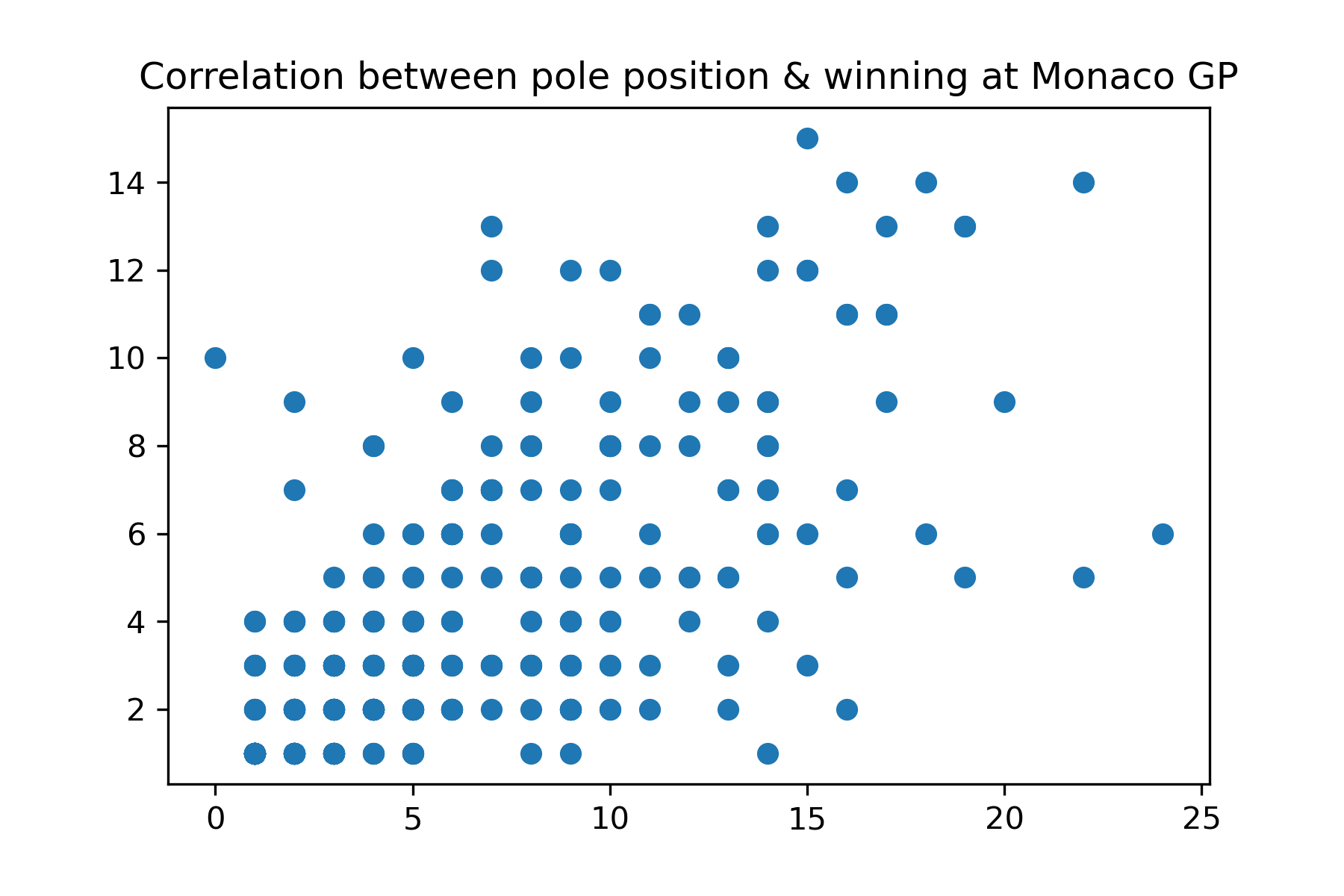


Figure 6. Correlation between pole position & winning at Monaco

The positive correlation between the starting position and results position shows the difficulties this track poses for drivers who need to overtake rivals to gain a position. However, the scatter plot shows variance in the results which can be attributed to the tight characteristics of the track and the likeliness of accidents or collisions occurring.

Figure 7 emphasises this as the Monaco race track has exactly 31 accidents or collisions more than the second highest track, Monza. These two tracks were in focus on Figure 3 as they topped the number of races run on a track so the accident rate is almost expected to be high in that respect. However, the horizontal bar plot shows Monaco to be clear of its race track partners and reinforce the testing nature of this track for drivers of all experience levels.

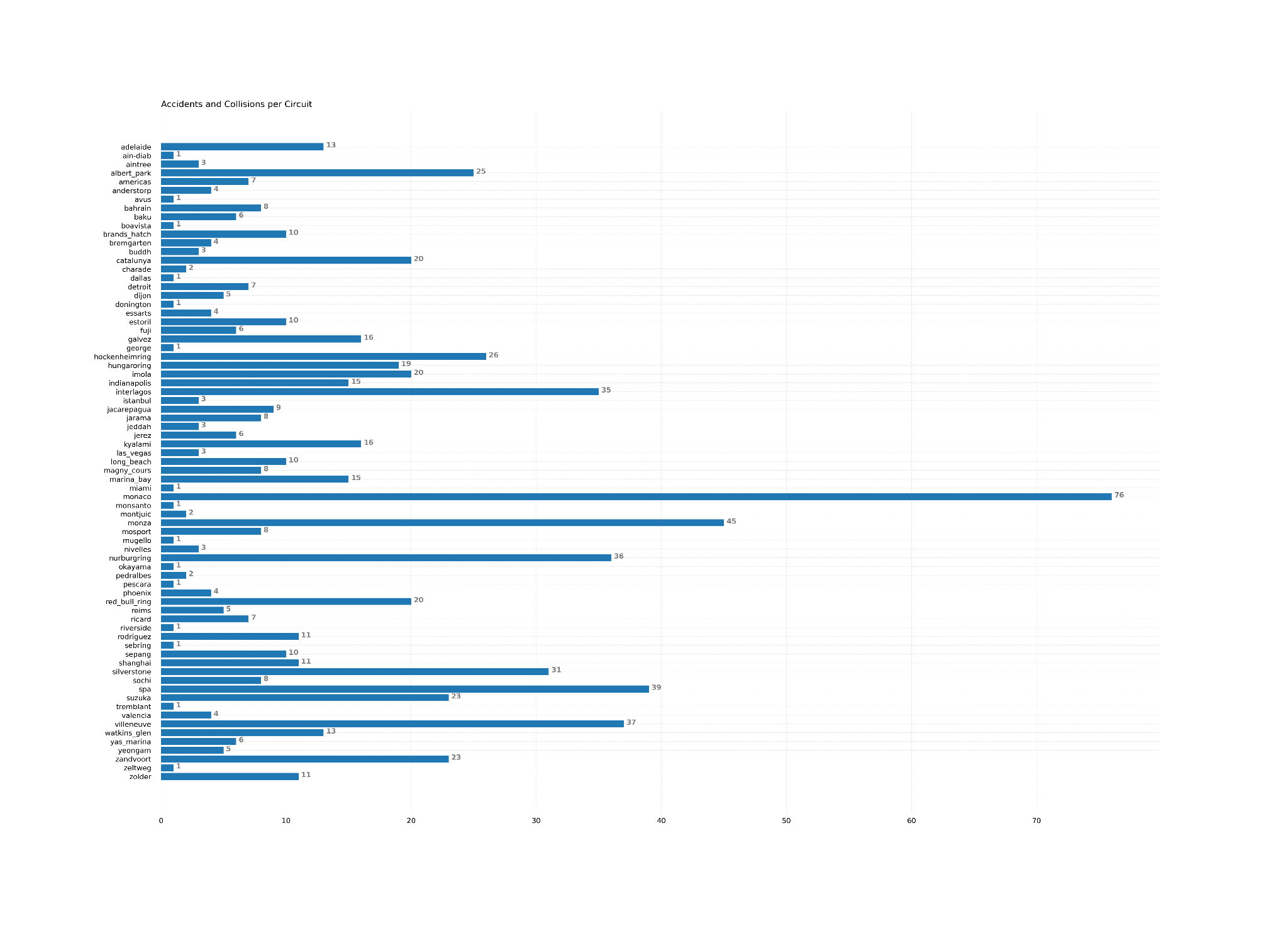


Figure 7. Accidents and Collisions per circuit

Formula One has seen notable regulation changes to reduce the number of accidents and collisions in a season. The introduction of the Halo safety device in 2018 marked a significant change in the safety of drivers from car debris or cars landing on top of one another during a race. The halo is a curved-bar that protects the driver’s head in the cockpit and weighs almost 10 kg to withstand impact and pressure where needed. The figure below shows a spike in accidents between the 2007 - 2010 seasons, with 2008 resulting in the highest number of accidents and collisions recorded in a Formula One season yet. Both the 2008 and 2010 seasons had 22 accidents and collisions recorded but a driver count of 368 in 2008 compared to 456 in 2010 led to the 2008 season having an accident count of 0.06.

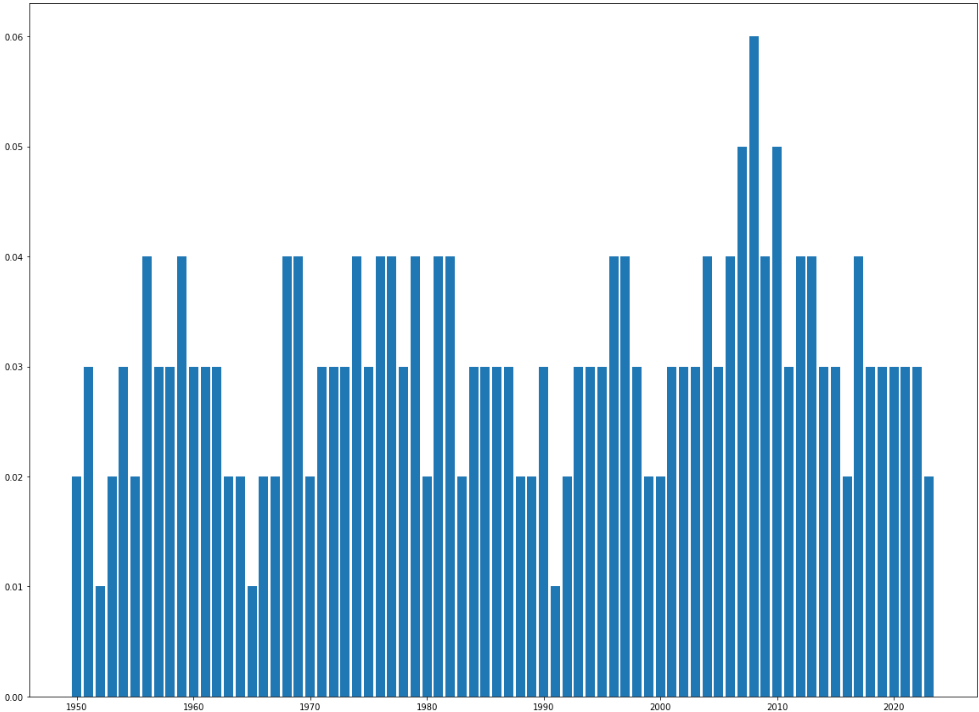


Figure 8. Accident rate per Year

Evaluating driver performance and consistency will be a key feature of the machine learning model’s ability to predict a race outcome. The figure below takes into consideration the number of wins, relative to the number of races competed, for each driver. The data shows that a five percent margin separates Lewis Hamilton, Michael Schumacher and Max Verstappen respectively. A similar gap is displayed between Max Verstappen’s win ratio compared to Fernando Alonso and after this, the percentage variance begins to decrease.

Lewis Hamilton is now the current all-time record holder with 103 wins but the data also highlights the explosive impact Max Verstappen has had on the sport given his debut was in 2016. The data also does not take into account that Max Verstappen is currently undefeated this season at the time of writing.

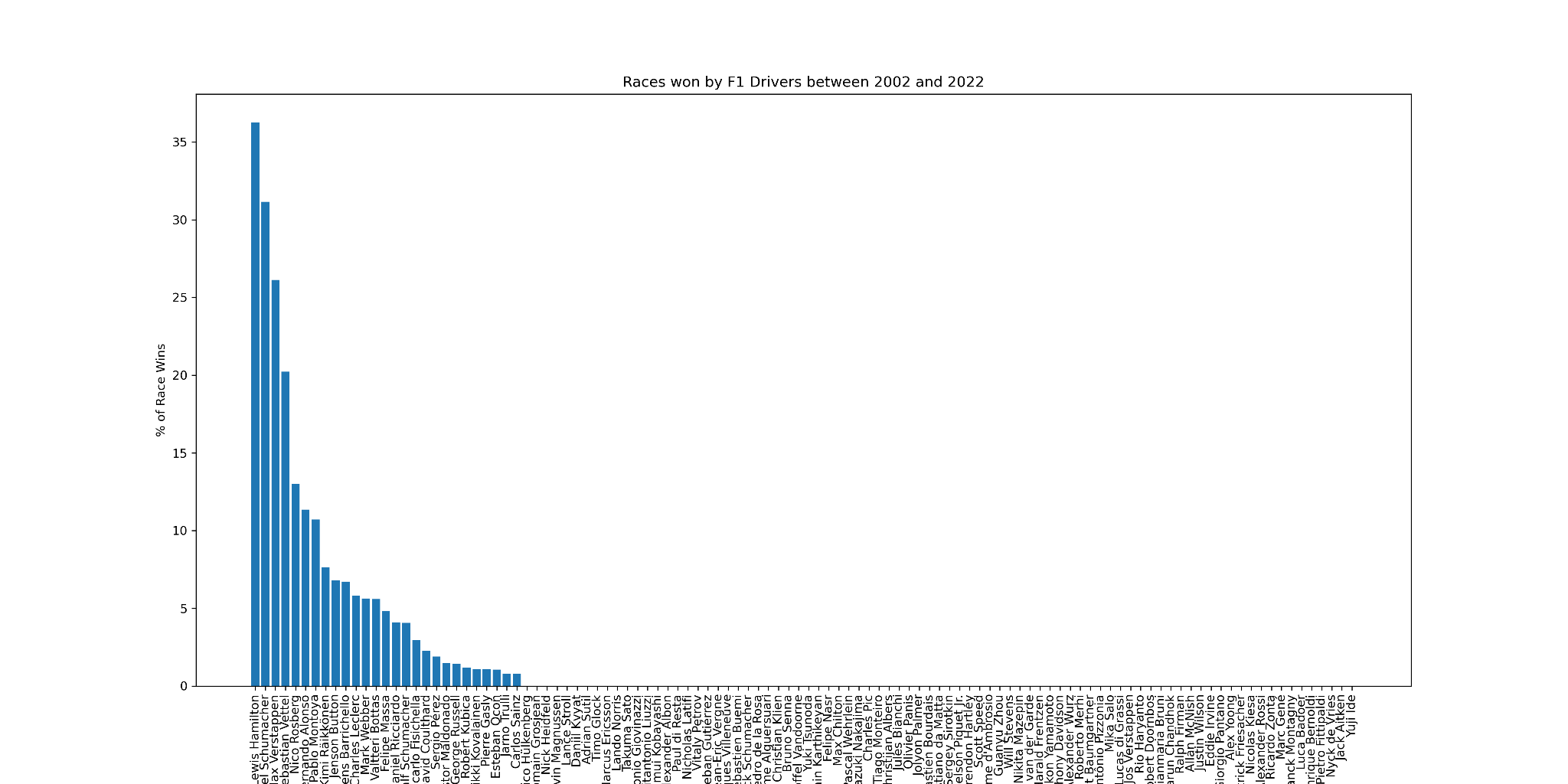


Figure 9. Percentage of wins per drivers (2002 - 2022)

#### *Missing Values*

Missing or corrupt data is all too common with real-world datasets and can induce biases in the data analytics if not addressed. There are ways to handle these potential problems such as treating them as null values or using imputation which predicts their value based on the existing values in the data.

The ‘Qualifying’ table shows missing values for the Q1, Q2 and Q3 variables. These represent the three stages allocated for grid position qualification for a driver before a race. Qualifying session one lasts for 18 minutes, session two lasts for 15 minutes and session three lasting 12 minutes. A break of seven and eight minutes are assigned to each car between these sessions and the fastest lap time recorded by each driver is taken as their qualifying time. The missing values here can be accounted for since a driver may crash the vehicle during qualification or suffer a performance failure in the car which results in forced retirement and incomplete qualifying rounds. An additional point to note here is the five slowest times in Q1 are eliminated from competing in Q2 and Q3, with the same rules applying for drivers moving from Q2 to Q3.

The figure below shows the ‘duration’ variable containing the highest percentage of missing data amongst all data tables. This represents the length of time for each car’s pit stop in a race and unlike the qualifying data, should have a record of each instance as the pit stop either occurs or not.

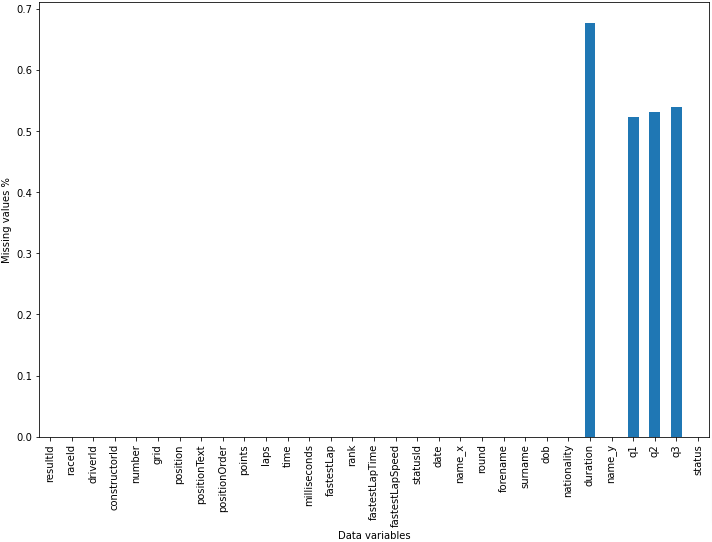


Figure 10. Percentage of missing data

### *Data Preparation & Feature Extraction*

The data preparation process

### *Modelling*

### *Model Evaluation*

### *Model Deployment*

### *Further Research*

Tyre change - this represents an important strategic decision by the team in any Formula One race. Although the speed of a Formula One car is more dependent on the car's capabilities and driver ability, the choice of tyre compounds at both the beginning of a race and during a race can yield significant advantage. The application of fresh tyres lead to greater performance and various factors such as weather, tyre degradation (wearing of tyre) and nearest rival’s pit strategy can influence the stage at which these changes are made in the pit stop during a race.

The Formula One regulations state that every driver must use at least two tyre compounds during a race unless wet weather plays a part in a race. This can result in a driver starting with soft tyre compounds to set faster lap times early in the race, before reverting to hard tyre compounds towards the end of a race and hoping to maintain their position against drivers which have reversed this strategy. The inclusion of tyre data such as the lap number where each driver entered the pit stop and which tyre compounds were used would offer greater predictive power to the models presented.

### *References*

[1] Saitta L, Neri F. Learning in the ‘‘real world’’. Machine Learning 1998; 30:133–163.

[2] J. Schmidhuber. Deep learning in neural networks: An overview (2015)

[3] Mohammad, Thabtah, McCluskey. An improved self-structuring neural network. (2016)

[4] Anil K.Jain, Jianchang Mao and K.M.Mohiuddin; Artificial Neural Networks: A Tutorial; Computer March 1996

[5] Rory P. Bunker, Fadi Thabtah, A machine learning framework for sport result prediction, Volume 15, Issue 1, (2019)

[6] P.J. Werbos. Applications of advances in nonlinear sensitivity analysis (2005)

[7] Qabajeh, Thabtah, Chiclana. A dynamic rule-induction method for classification in data mining. (2015)

[8] L. a. H. J. G. Passfield, "A Mine of Information: Can Sports Analytics Provide Wisdom From Your Data?," International Journal of Sports Physiology and Performance, vol. 12, no. 7, pp. 1-17, December 2016

[9] U. Johansson and C. Sonstrod; Neural Networks Mine for Gold at the Greyhound Track ; International Joint Conference on Neural Networks, Portland, OR. (2003)

[10] H. Chen et al.; Expert Prediction, Symbolic Learning, and Neural Networks: An Experiment on Greyhound Racing; 1994-12, 9(6):21-27 IEEE Expert

[11] Davoodi, E.; Khanteymoori, A.R. Horse racing prediction using artificial neural networks. In Proceedings of the 11th WSEAS International Conference on Neural Networks, Evolutionary Computing and Fuzzy Systems, Iasi, Romania, 13–15 June 2010; pp. 155–160

[12] F. Kamalov, F. Thabtah. A feature selection method based on ranked vector scores of features for classification (2017)

[13] Ofoghi, B., Zeleznikow, J., Macmahon, C., Rehula, J., & Dwyer, D. B. (2016). Performance analysis and prediction in triathlon. Journal of sports sciences, 34(7), 607–612

[14] K.Przednowek. J.Iskra and K.H.Przednowek; Predictive modeling in 400-metres hurdles races; Proceedings of the 2nd International Congress on Sports Sciences Research and Technology Support 2014, Pages 137-144

[15] S. Grossberg. Nonlinear neural networks: principles, mechanisms, and architectures (1988)

[16] Edelmann-Nusser, J.; Hohmann, A.; Henneberg, B. Modeling and prediction of competitive performance in swimming upon neural networks. Eur. J. Sport Sci. 2002, 2, 1–10

[17] M. A. Sartori and P. J. Antsaklis; A simple method to derive bounds on the size and to train multilayer neural networks; IEEE Transactions on Neural Networks, Vol. 2, pp. 467–471, 1991

[18] Pratt, J. Risk Aversion in the Small and in the Large. Econometrica 1964, 32(1-2), 122-136

[19] Langley P, Simon HA. Applications of machine learning and rule induction. Commun. ACM 1995; 38:54–64

[20] Abut, F., Akay, M. F., Daneshvar, S., & Heil, D. (2017, September). Artificial neural networks for predicting the racing time of cross-country skiers from survey-based data. In 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 117-121)

[21] Daniel Svozil, Vladimír Kvasnicka, Jir̂í Pospichal,Introduction to multi-layer feed-forward neural networks,Chemometrics and Intelligent Laboratory Systems, Volume 39, Issue 1, 1997 (Pages 43-62)

[22] Du, Ke-Lin & Swamy, M.N.s. (2014). Radial Basis Function Networks

[23] Sankaranarayanan VV, Sattar J, and Lakshmanan LVS. Auto-play: A data mining approach to ODI cricket simulation and prediction. In: Proceedings of the 2014 SIAM International Conference on Data Mining, 1064– 1072. SIAM 2014

[24] Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In Classification, clustering, and data mining applications (pp. 639-647). Springer, Berlin, Heidelberg

[25] Bailey M, and Clarke SR. Predicting the match outcome in one day international cricket matches, while the game is in progress. Journal of Sports Science & Medicine 2006; 5:480

[26] Marcano-Cedeno, Quintanilla-Dominguez, Cortina-Januchs, Andina. Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. (2010)

[27] N. Tax and Y. Joustra; Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach; Transactions on knowledge and data engineering, Vol. 10, No. 10, pp. 1-13, 2005

[28] Joseph, A.; Fenton, N.; Neil, M. Predicting football results using Bayesian nets and other machine learning techniques. Knowl.-Based Syst. 2006, 19, 544–553

[29] Fayyad U, Piatetsky-Shapiro G, and Smyth P. From data mining to knowledge discovery in databases. AI Magazine 1996; 17:37–54

[30] J. Goddard, "Regression models for forecasting goals and match results in association football," International Journal of Forecasting, vol. 21, no. 2, pp. 331-340, 2005

[31] J. a. R. A. Hucaljuk, "Predicting football scores using machine learning techniques," in MIPRO, 2011 Proceedings of the 34th International Convention, Opatija, Croatia, 2011

[32] Z. Y. Broich, "Statistical Analysis for the First Bundesliga in the Current Soccer Season," Progress in Applied Mathematics, vol. 7, no. 2, pp. 1-8, 2014

[33] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(Feb), 281-305

[34] S. Kampakis, "Using Twitter to predict football outcomes," 2014

[35] S. Kampakis, "Predictive modelling of football injuries," arXiv:1609.07480, London, UK, 2016

[36] S. O. Rue, "Predicting and Retrospective Analysis of Soccer Matches in a League," Department of Mathematical Science, NTNU, Norway, 2000

[37] R. M. Crowder, "Dynamic Modelling and Prediction of English Football League Matches for Betting," Journal of the Royal Statistical Society, vol. 51, no. 2, pp. 157-168, 2002

[38] S. Raychaudhuri, "Introduction to Monte Carlo simulation," 2008 Winter Simulation Conference, Miami, FL, USA, 2008, pp. 91-100

[39] Ulmer, B., Fernandez, M., & Peterson, M. (2013). Predicting soccer match results in the XI English Premier League (Doctoral dissertation, Ph. D. thesis, Doctoral dissertation, Ph. D. dissertation, Stanford)

[40] Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378

[41] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies IX for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), e1301

[42] McCabe, A.; Trevathan, J. Artificial Intelligence in Sports Prediction. In Proceedings of the Fifth International Conference on Information Technology: New Generations (itng 2008), Las Vegas, NV, USA, 7–9 April 2008.

[43] M.A. Hall. Correlation-based Feature Subset Selection for Machine Learning (1999)

[44] Haghighat, M.; Rastegari, H.; Nourafza, N. A Review of Data Mining Techniques for Result Prediction in Sports. ACSIJ Adv. Comput. Sci. Int. J. 2013, 2, 7–12

[45] Platt, J. C., (1999). Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in Kernel Methods: Support Vector Learning. B. Scholkopf, C. Burges and A. Smola. Cambridge, MA, MIT Press: 185-208

[46] J. Jane, E. N. Ganesh. A Review On Big Data With Machine Learning and Fuzzy Logic For Better Decision Making (2019)

[47] P. Melin, O. Casstillo. Hybrid Intelligent Systems for Pattern Recognition Using Soft Computing: An Evolutionary Approach for Neural Networks and Fuzzy Systems

[48] M.C. Purucker; Neural network quarterbacking; IEEE Potentials, 15(1996), pp. 9-15

[49] J. Kahn; Neural network prediction of NFL football games; World Wide Web Electronic Publication, 2003 (2003), pp. 9-15

[50] C.K. Leung and K.W. Joseph; Sports data mining: predicting results for the college football games; Procedia Computer Science 35, pp. 710-719, 2014

[51] Delen, D.; Cogdell, D.; Kasap, N. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. Int. J. Forecast. 2012, 28, 543–552

[52] S. Talukder, "Preventing in-game injuries for NBA players," in MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA, 2016.

[53] J. Kuehn, "Accounting for Complementary Skill Sets When Evaluating NBA Players’ Values to a Specific Team," in MIT Sloan Sports Analytics Conference 2016, Boston, MA, USA, 2016.

[54] Wiseman, Using Machine Learning to Predict the Winning Score of Professional Golf Events on the PGA Tour (2016).

[55] Witten, I. H. & Eibe, F. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco:2005, Morgan Kaufmann.

[56] Pfitzner CB, Rishel TD. Do reliable predictors exist for the outcomes of races? The Sport Journal, March 4, 2008.

[57] M. Allender; Predicting The Outcome Of NASCAR Races: The Role Of Driver Experience; Journal of Business & Economics Research, Vol. 6, No. 3, pp. 79-84, March 2008

[58] C. Depken and L. Mackey; Driver success in the Sprint Cup series: The impact of multi-car teams; Social Science Research Network. Available online at: <http://ssrn.com/abstract=1442015> 2017

[59] Mercedes-AMG Petronas Formula One - YouTube. How Much Data Does an F1 Car Generate? 2022

[60] A. Woodie. Go Fast and Win: The Big Data Analytics of F1 Racing (2018)

[61] Bell, A., Smith, J., Sabel, C. E., & Jones, K. (2016). Formula for success: multilevel modelling of Formula One driver and constructor performance, 1950–2014. Journal of Quantitative Analysis in Sports, 12(2), 99-112.

[62] Eichenberger, R., & Stadelmann, D. (2009). Who Is The Best Formula 1 Driver? An Economic Approach to Evaluating Talent. Economic Analysis & Policy, 39(3)

[63] Silva, K. M., & Silva, F. J. (2010). A tale of two motorsports: A graphical-statistical analysis of how practice, qualifying, and past success relate to finish position in NASCAR and Formula One racing.

[64] Graves, T., Reese, C. S., & Fitzgerald, M. (2003). Hierarchical models for permutations: Analysis of auto racing results. Journal of the American Statistical Association, 98(462), 282-291.

[65] Biemann, Liu, Zeng and Huang. Addressing partial observability in reinforcement learning for energy management (2021)

[66] Choo, C.L.W. Real-Time Decision Making in Motorsports: Analytics for Improving Professional Car Race Strategy. Master’s Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2015.

[67] Tulabandhula, T.; Rudin, C. Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing. Big Data 2014, 2, 97–112.

[68] Aversa, P.; Cabantous, L.; Haefliger, S. When decision support systems fail: Insights for strategic information systems from Formula 1. J. Strateg. Inf. Syst. 2018, 27, 221–236.

[69] Jenkins, M. (2010). Technological discontinuities and competitive advantage: A historical perspective on Formula 1 motor racing 1950–2006. Journal of Management Studies, 47(5), 884-910

[70] Aversa, P., Furnari, S., & Haefliger, S. (2015). Business model configurations and performance: A qualitative comparative analysis in Formula One racing, 2005–2013. Industrial and Corporate Change, 24(3), 655-676

[71] Rosso, G., & Rosso, A. F. (2016). Statistical Analysis of F1 Monaco Grand Prix 2016. Relations Between Weather, Tyre Type and Race Stints.

[72] Koenker, Roger, and Kevin F. Hallock. 2001. "Quantile Regression." Journal of Economic Perspectives, 15 (4): 143-156

[73] A. Sundar, "Here is everything you want to know about sports analytics," 24 August 2018. [Online]. Available: <https://dataconomy.com/2018/08/here-is-everything-you-wantto-know-about-sports-analytics>

[74] Amazon Web Services, Inc. F1 Insights Powered by AWS. Available online: <https://aws.amazon.com/de/f1> 2020