Sean Morris, Jacob Meyer, Jordan Rinaldi, Kehan Wu
Professor Jacobs
DSAN 5100-03
Due 10 December 2024

Predicting Success in Formula One

Introduction and Motivation:

The intersection of data science and sports has only grown in recent years. In almost every major professional sport, teams and individuals take advantage of data-driven insights in every facet of their sport, from training, to formulating a game plan strategy, to in-game decision-making, and even after the game is over. Whether an individual or team sport, there is still an abundance of opportunity to utilize data to gain an advantage over one's competition. With Formula One, this is no exception, as every team has a group of data scientists, engineers, and other data-driven strategy professionals to help teams get smarter in every aspect of racing.

Formula One (F1) is a unique case study, as it consists of a good combination of team and individual elements, namely that an individual driver is responsible for the outcome of each race, but they rely heavily on the support of many teams to have success on the grid. It was founded in 1946, and since then, it has been perhaps the most popular and successful global motor racing circuit. It is a symbol of technology, innovation, speed, and precision. It has undergone many changes to rules, tracks, and cars themselves, but the original premise has remained the same.

Since its founding, Formula One has evolved significantly, reflecting changes in technology, safety, and racing formats. In its early years, F1 cars were relatively basic compared to the technological marvels seen today, with limited safety features and rudimentary engineering. Over time, advancements such as the introduction of aerodynamic designs, turbocharged engines, and high-performance tires have revolutionized the sport. One significant milestone came in 1981 when the Concorde Agreement formalized the structure of F1, leading to greater consistency in regulations and financial stability.

The 21st century has seen a wave of safety and format innovations. For instance, the introduction of the "halo" device in 2018 drastically improved driver safety by protecting their heads from flying debris, a change that undoubtedly saved the life of Romain Grosjean during his dramatic crash at the 2020 Bahrain Grand Prix. Another major change was the ban on refueling during races in 2010, which shifted the focus to fuel efficiency and pit stop strategies. In 2021, the sprint qualifying format was introduced, adding an extra layer of excitement and strategy to race weekends. Throughout F1's history, legendary drivers like Ayrton Senna, Michael Schumacher, and Lewis Hamilton have pushed the boundaries of what is possible on the track, contributing to the sport's prestige and global appeal. With F1 currently featuring 10 teams and 20 drivers competing on the grid, it continues to balance tradition with innovation, ensuring its place as a premier global motorsport.

To have success in Formula One, drivers must be adept at making lightning-quick decisions, as well as having good knowledge about the car they are driving, the track they are driving on, and the other racers they are driving alongside. While most of the burden falls upon the driver, there are countless other members involved that contribute to the success of a racer. Pit crew teams take care of the car during the race to ensure that the car is able to maintain optimal efficiency. Other members of the crew communicate with the driver during the race to help drivers adjust their racing game plan, as well as alert them to any potential issues or things that must be addressed. Large teams off the track aid in development of the car itself, as every single detail is given extreme attention with the hopes of making the cars faster and more efficient. Strategists work with the drivers throughout race week to make a strategy unique to each race course.

In each of these areas, as well as in many more, members of race teams are looking at data to inform decision-making. Teams acquire data from every measurable aspect of a race, from lap times to pit stops, and they are constantly analyzing and forecasting the data to determine what works for an individual gameplan. In this analysis, we hope to achieve a similar outcome. We want to be able to think like a Formula One team, which has some of the most modern and advanced data science teams in the world of professional sports. We hope to determine what metrics matter the most for racers to achieve the ultimate goal in each race, which is winning. While there are many other measures of performance in the world of F1, we determine that winning is the most important goal that teams are striving for, and in the effort to approach this data science problem exactly like an F1 team would, this will be our measure of performance. Overall, this analysis will focus on how we can predict who will have success on the modern F1 grid by determining the factors and trends that contribute to winning a race in the 21st century.

Data Collection and Initial Analysis:

Data Collection

For this study, the team relies on an open-source dataset available online. The data contains in-depth information on Formula 1 races - including drivers, constructors (teams), qualifying races, track information, pit stop and lap times, and championships. While the data traces back to 1950, our study focuses on records from 2000 and beyond. The rationale behind this truncation lies in the dynamic nature of Formula One, where antiquated engines, strategies, engineering techniques, and rules are quickly phased out upon the advent of newer, more efficient ones. The inclusion of data that extends past 2000 provides a more consistent and relevant analysis by reducing the risks of confounding factors.

The overall dataset has a relational structure (**figure 1**), where each of the fourteen tables are tied together using primary and foreign keys. In other words, each table within the dataset contains information that is relevant to entries in other tables. For example, the *races* table offers a record of individual F1 events (e.g. circuit, date, name, year), where the column, *'raceId'*, links to the *results* table - containing the start and finish positions based on *driverId*. Similarly, the

drivers and *constructors* tables provide details on information like driver age and nationality, and constructor nationality. In one part of the analysis, we join the *results* and *constructors* tables to conduct analysis on the varying placements of F1 teams across different races. In most cases during our study, table joins are a necessary first step before conducting any sort of analysis.

Constructors

One of the overarching questions posed during this study is whether Formula One "constructors" differ in their rate of placement across races. Constructors (teams) are individuals or corporations that take part in building and maintaining the cars over the course of a season. Analysis on constructors begins with truncating the data to only records after the year 2000. The rationale behind this cutoff is to only include modern day races, and therefore ignore any potential confounding factors related to more antiquated engines, race rules, and strategies. Once a clean data frame is constructed, we created a plot displaying a bar graph of top constructors by wins since 2000. The plot reveals a clear trend in constructor winning patterns, where three teams hold a majority of F1 wins since 2000. The teams (**figure 2**) are Red Bull, Ferrari, and Mercedes - each of which hold over 112 race wins since 2000. The clear trend in constructor winning patterns naturally leads to a closer examination of what we label as "the big three" teams. For this, a boxplot showing position distributions for each of these teams, faceted from 2020 to 2024, is constructed to analyze how the final placements of Red Bull, Ferrari, and Mercedes have varied throughout the past 5 years. The plot (**figure 3**) shows a clear dominating trend, where each of 'the big three' teams has had around 50% of their placements within the top five over the last 4 years. Faceting the boxplots by year allows for additional temporal analysis of placements by these teams. Over the past 5 years, there has been a clear upwards shift in the distribution of placements by each of the big three teams - meaning their average placements in races has consistently improved year over year.

Following the zoomed-in analysis of placements by 'the big three,' the study then shifts its focus toward analyzing the aggregate placement of all teams since 2000 (**figure 4**). For this, we elect to construct a box plot displaying the distribution of placements across all teams since 2000. This plot offered some interesting insights into performances league-wide. Interestingly, the plot shows a similar distribution of final placements for both high-performing and low-performing teams. In each case, the final placements for both of these subsets are skewed. High-performance teams (big three) have a right-skewed distribution of final placements, meaning most of their outliers are due to abnormally high (poor) placements. On the other hand, low-performance teams (HRT, Virgin) have left-skewed distributions of final placements, meaning their outlier placements are better (lower) than usual. Most teams in the data fall in the middle ground of these two subsets. Middle of the pack teams for the most part have normally distributed final race placements, with much larger variances compared to the high- and low-performing teams.

As a final part of examining whether constructors differ in their placement across races, we incorporate a brief analysis of the financial power of 'the big three.' For this section, we look at

the budgets over time for some of the largest teams (**figure 5**) . Unsurprisingly, Red Bull, Ferrari, and Mercedes have the top three largest F1 budgets as of 2019. While not a catchall factor, we believe that money matters in F1 races, and the massive budgets of the big three teams is most certainly not a detriment to their continued success in the modern era of F1.

Sprints

Our next question of the study lies in whether or not the performance in Formula One sprint races has any influence on performance in actual races. Formula One sprints hit the scene in 2021, offering audiences an additional piece of high octane entertainment before the Gran Prix. Sprint races are shorter than traditional F1 circuits, typically stretching out less than 100 km - or about a third of traditional races. Sprints do not require any pit stops, therefore lowering the amount of strategy required, and placing more emphasis on pure speed. Unlike qualifiers, sprint races award championship points to high performers, and therefore offer teams greater incentive to place well[1].

For our analysis, we first look at a scatter plot of average sprint final position versus average final race position by constructor (**figure 6**). Unsurprisingly, the so-called 'big three' teams all have high placement across both, placing them in the lower left quadrant of the plot. In addition to this finding, the plot demonstrates a clear linear relationship between sprint performance and race performance. At first, this begs the question as to whether a simple linear regression could properly fit the two together. However, after running the model and looking at the residual plot, we observed a clear trend in the residuals that suggested clear issues with our modeling approach (**figure 7**). In order to better diagnose the issue, we construct a boxplot showing the distribution of final race performance based on final sprint position (**figure 8**). Similar to the constructor performance data, both sprint performance and race performance are discrete and ordinal in nature. Furthermore, the boxplot also hints at a nonlinear relationship between the two variables, as evidenced by the slight curvature of mean race placements as sprint placement begins to deteriorate. We revisit these findings in the statistical methods section, where specific models and tests are employed to account for the discoveries brought forth by our analysis here.

Grid Position

Our final question asks whether starting position in the race has any influence on performance in the race itself. F1 uses three qualifying rounds to determine starting positions. During qualifying, a driver completes a timed lap of the race course three times in three rounds to determine a position 1 to 20 for the field. This position is known as a driver's "grid position" and starting at the front of the grid is critical for success on race day. Overtaking (or passing) is not easy in Formula One due to the high speeds and typically narrow, windy courses, so there is a big advantage in starting in the top 5 compared to 10-15th position. The best drivers usually battle it

[1] https://shorturl.at/Qtcoz

out for the top spot known as the pole position with a large advantage in the actual race given to the driver who secures it.[2]

As we have throughout our analysis, we limited our dataset to 2000 onwards due to the evolution of Formula One throughout the years. The cars, engines, tires and all the minute components of today's races are a world away from the 1950s and 60s. The strategies and factors that aid a driver in winning in today's Formula One world were not necessarily the same 30 or 50+ years ago. Our goal is to predict the winner of a race in this era, and we do not want our results swayed by older races.

We looked at 7,688 individual driver race results that spanned across 124 drivers and 467 races from 2000 to 2024. We first created a heatmap **(figure 9)** of finish and grid positions which shows a linear relationship between the two. Unsurprisingly, a lower start position seems to correlate to a lower finish position in the race. This matches our expectation that a driver who qualifies in a top position is more likely to finish in a top position. The same characteristics of the driver and their car that lead them to perform well in qualifying translates to finish position in the race. We do notice that the spread of finish positions appears to be wider in the middle grid positions compared to the first and last few grid positions. This brings up an interesting question of why that might be. Those drivers that qualify in the top and bottom groups appear to be more consistent with how they finish compared to where they qualify whereas the larger spread in the middle shows less predictability on their finish position based upon grid position. This aligns with our constructor findings that middle of pack teams have more variability in their final position compared to high or low performing teams.

Statistical Methods:

This analysis will rely on multiple different statistical models and techniques, from machine learning to hypothesis testing.

Machine Learning:

Analysis on how final positions in sprint races impact final race positions is supplemented through the use of a proportional odds model, or ordinal logistic regression. Ordinal logistic regressions are designed to predict outcomes of a dependent variable that has a natural order or ranking. In the case of sprint and race positions, each of these data points are discrete and ordinal in nature, making them ideal candidates for the model. Unlike linear regressions, OLRs do not assume any difference between labels (i.e. first place versus second place are equivalent across the scale). To further reinforce this point, when looking back at the plot constructed during our analysis of race position distributions based on sprint position (**figure 8**), there are clear, equal spacings between outlier points - helping to reinforce the notion that this data is indeed ordinal. As a final point, OLRs are equipped to handle non-linear relationships (whereas OLS holds this as a core assumption). Again, when referencing the aforementioned plot, there is a clear non-linear trend in final race position distributions by sprint position.

[2] https://www.redbull.com/us-en/how-does-f1-qualifying-work

Hypothesis Testing:

We utilize multiple hypothesis tests within this analysis. The first test will be an analysis of variance, or ANOVA test. ANOVA is a statistical method that attempts to determine if the means of multiple categorical groups differ from one another. It requires a categorical labeling variable to divide the observations into groups, as well as a quantitative variable. This test will analyze the sample means of the numeric variable for each of the categories to determine whether the population means of each category differs, and this will be done by comparing the variability of individual observations to the sample means of each category.

In our analysis, in lieu of a one-way ANOVA test, we will use a non-parametric version of ANOVA called a Kruskal-Wallis test. This will be necessary for two different variables we will be testing, which are pit stop times and individual constructors. The advantage of using the Kruskal-Wallis test is that it does not require the data to be normally distributed. Because it is a non-parametric test, it does not make any assumptions about the mean, variance, or distribution of the sample data. Therefore, it can be used on datasets with smaller sample sizes and with different levels of variance across categorical groups. The null hypothesis of this test is that the means of the distributions are equal across categories, and the alternative hypothesis is that one of the means is different from the others, and thus there is a correlation between the numeric and categorical variables.

Looking first at pit stop times, **figure 10** depicts the histogram of average pit stop times within a race for a given driver. It is immediately clear that the data are not normally distributed. Moreover, this is with outliers removed, as there are some outliers that are more than 4 times the value of the average. This is due to the fact that some pit stops are far longer than others if the car has sustained some damage during the race that must be addressed before the car can return to the track. In the effort of understanding goal pit stop performance, we remove those observations from our hypothesis testing samples. We will also only look at top 10 finishes, because some of the later finishes have a lot of noise, due to the fact that the drivers often have little to no chance to win late in the race, and are thus less incentivized to have fast average pit stop times and instead attempt to get the fastest lap award for a race, which involves a longer pit stop to ensure that the car can achieve top speeds. Now that the data are cleaned, they are optimal for the Kruskal-Wallis test, due to the fact that each category will have the same number of samples. This is because at least 10 drivers finished in every race that we sampled.

Continuing forward into our analysis of constructor performance across Formula One races, we again elect to use the non-parametric Kruskal-Wallis test. In the initial preprocessing and EDA steps, it became clear that data on constructor performance is both discrete and ordinal. Kruskal-Wallis excels in testing differences between ranking-based data across groups, which exactly characterizes constructor performance. The final and perhaps most important reason why Kruskal-Wallis was chosen over ANOVA lies in distributions of constructor placements. Two of the largest assumptions when using ANOVA tests are that the data must be normally distributed and have equal variances. When looking at the distributions of race placements across

constructors (**figure 4**), it becomes clear that both of these assumptions are violated. As stated previously, the distribution of final positions across high- and low-performing constructors are skewed in opposite directions, while the middle-ground teams all possess normally distributed final placements. Further, the plot shows that the variances in constructor performance are not at all uniform. Rather, we observe much larger variances for middle-of-the-pack teams, and tighter spreads for high and low-performers. With all of this in mind, the decision on whether Kruskal-Wallis test over ANOVA becomes clear.

In addition to using a proportional odds model to examine the relationship between sprint performance and race performance, we also elect to run a spearman rank correlation test on the two variables. Again, the rationale behind using this test over others lies in two observations made during our initial analysis. The first being that both final race position and final sprint position are both discrete and ordinal, making the spearman rank test an ideal candidate for analysing the relationship between these two variables. The second reason for employing the spearman rank correlation test lies in the non-linear relationship between the two variables. Whereas Pearson correlation tests for linear correlations, spearman tests for monotonic relationships - meaning it is able to capture whether better sprint performance (lower positions) is consistently associated with better race performance (lower positions) without assuming a linear relationship. Finally, using this test in tandem with the OLR model allows us to capture the relationship between sprints and races in both a probabilistic and clearcut way.

Similarly to our sprint analysis, we used a Spearman Rank Correlation and hypothesis test to analyze the connection between grid position and finish position. We did so because these variables are ordinal and discrete making them a perfect candidate for a Spearman test. Additionally, a Spearman Correlation test does not assume normal distribution or a linear relationship between the variables unlike a Pearson correlation test. To investigate our conjecture that there is more variability between grid position and finish position for grid positions in the middle, we used the rollapply function in R to apply a Spearman Correlation test to a moving window of grid position and finish position.[3] This allowed us to determine whether there truly is a stronger relationship between grid and finish position for the top and bottom portion of the race.

Another hypothesis test that was utilized in this analysis was the chi-square test for independence. This test measures two categorical variables in a contingency table to determine if they are associated with one another, or if they are independent. To set up this test, one must make a frequency table with two categorical variables making up both axes, and the values within the table are the intersecting frequencies of occurrence within the sample. This is the best hypothesis test for measuring if two categorical variables have a correlation with winning. The null hypothesis of this test is that there is no significant relationship between the two categorical variables, and the alternative hypothesis is that the two variables are correlated.

We want to compare drivers to race courses, as well as constructors to race courses. To do this, we will subset the data to only include observations where the driver finished first in a race.

[3] Roll Apply in R: https://shorturl.at/5dyds

We will build the contingency table with our two variables of choice, and then use our chi-square test for independence to determine the relationship between the two variables. We also want to compare the results between constructors and drivers, to try and see if one of them has a better correlation with tracks to determine success. In other words, when a driver/team combination has a lot of success on a track, can it be more attributed to the driver or to the constructor? The chi-square test for independence will help us answer that question.

Results:

Random Forest:

Random forest models show significant advantages in analyzing complex non-linear relationships, especially in terms of the impact of starting position, pit strategy and track characteristics on race results. From the feature importance analysis **(as shown in Figure 11**), it can be seen that the total pit time (total_pit_time) is the most critical variable, with an importance of 54.7% for model performance. In addition, track altitude (alt) and driver score (points) contributed 33.9% and 29.8% importance, respectively, indicating that strategies and external environmental factors also have a profound impact on race results. The model also revealed that starting position (grid) and the number of pit stops (pit_stop_count) had a smaller impact, but still affected performance under certain conditions. Overall, the random forest model achieved an accuracy rate of 84.5%, capturing the complex interactions between multiple variables. However, the model performed relatively weakly in predicting the results of mid-to-back-row drivers, which may be due to the fact that these positions are more susceptible to random factors such as accidents or weather.

The random forest model further shows that a driver starting in the top five has a more than 70% chance of finishing in the top ten, which confirms the crucial role of the starting position in the race. Regarding pit strategies, the data shows that the average single pit stop time for the top ten drivers is 2.8 seconds, while the pit stop time for the bottom ten drivers generally exceeds 4 seconds. This difference may directly determine the final ranking of the drivers. Teams usually make their final pit stop between 55% and 60% of the race to balance the speed and durability of the tires. In addition, lap times at high-altitude circuits are relatively poor, with an average lap time about 6 seconds slower than at low-altitude circuits, reflecting the impact of environmental factors on car performance.

Linear Regression:

Linear regression models, on the other hand, focus on analyzing the linear relationship between variables. In the analysis of starting position and final score, the regression coefficient (**as shown in Figure 12**) indicates that starting position has a significant negative correlation with the driver's score, and the higher the starting position, the higher the score. The correlation coefficient of the regression analysis is -0.725, indicating a close relationship between starting

position and driver score. The average score of drivers starting in the top three positions was 15.6 points, while the average score of drivers starting in the 15th to 20th positions was less than 2 points. In addition, the adjusted R^2 value of the linear regression model was 0.0489 (**see Figure 13**), indicating that although the model can explain the linear relationship between some variables, the overall effect is limited, especially in capturing complex nonlinear relationships. Combining the results of the linear regression model and the random forest model further reveals the importance of pit stop efficiency and starting position for race performance. The significant coefficients in the linear regression model show that every additional second of pit time increases lap time by 2.23 seconds. This finding is consistent with the feature importance analysis of the random forest (**as shown in Figure 11**), in which total pit time has the greatest impact on lap time, with an increase in node purity of 1889.89. increased by 1889.89. In addition, the random forest model also confirmed the significant impact of track altitude (Altitude) and driver score (Points) on lap time, ranking second and third with an increase in node purity of 1688.91 and 1054.48 respectively. In contrast, the impact of the number of pit stops (Pit Stop Count) was minimal, with an increase in node purity of only 486.14.

From the linear regression model (**see Figure 13**), the regression coefficient of the starting position is -4.14 milliseconds, indicating that for every position behind the starting position, the lap time will increase by 4.14 milliseconds, while every 1 point increase in championship points can reduce the lap time by 7.27 milliseconds. These data further verify the importance of pit stop efficiency and starting position for the final result, and also show the high degree of consistency between the results of the random forest and linear regression in some respects. By optimizing these key variables, teams can significantly improve their race performance.

Pit Stops:

Due to the results of our random forest model, we performed an extensive analysis of the relationship between pit stops and final performance. **Figure 14** shows the boxplot of average pit stop time (in seconds from start to finish when they were not on the track) grouped by final position. As mentioned in the prior section, this portion of analysis has large time outliers removed, and it only includes observations where the driver finished in the top 10. This eliminates a bunch of variance that would be detrimental for our analysis, as many of the pit stop times are skewed from crashes and late finishes where the focus is not on a fast pit stop time, but rather on preserving the health of the car or attaining race points in some other method besides final position. We also only show data from 2010 and forward, as the rest of our project only studies data since 2000, but the dataset that we had only contained pit stop data after 2010. This was also beneficial for our analysis, as the Formula One rules changed substantially prior to the 2010 season. Previously, teams had to fuel their cars during the race, which introduced different strategies, such as leaving less fuel in the cars to make them lighter and thus faster. However, in 2010, F1 removed fueling from the race as a safety precaution. This benefits our analysis because pit stop times prior to 2010 were likely far higher on average, due to the fact that pit

crews had to spend more time refueling the cars. This would have added a temporal bias to our analysis, so we are happy to only analyze pit stop times after this rule change.

As mentioned, we used a Kruskal-Wallis test for this analysis. The null hypothesis of this test is that the average pit stop times for each final position value are equal for the population. We get a p-value of 1.676e-6, with a chi-squared value of 43.59 and 9 degrees of freedom. Our p-value is significantly low at a 95% significance level, which means that we can reject the null hypothesis and conclude that there is a correlation between average pit stop time for a race and final race position. Looking back at **figure 14**, we see a general increase in average pit stop time as race finish gets worse. The top finishing positions average nearly a full second slower per pit stop than teams that finish towards the middle of the pack. Our hypothesis test supports the results from our machine learning test, as it is clear that pit stop times contribute to winning and can be the difference between winning and losing.

We took this analysis a step further and looked at the number of times a driver went for a pit stop during a race. We wanted to find out if drivers who perform better in races pit more or less, on average. We performed the same hypothesis test as above, and the test yielded a p-value of 0.9414, with a chi-squared value of 3.4952 and 9 degrees of freedom. Based on this p-value, it is clear that we cannot reject the null hypothesis, which was that there is no significant difference between the number of pit stops and the final race position. **Figure 15** shows exactly why, as the histogram shows that there is not a major difference between the number of pit stops, as almost every driver pits either once or twice for the duration of a race.

Finally, we looked at when drivers make their final pit stop during a race to determine if that had a correlation to winning. In Formula One, tires are crucial, as they can wear out fast, and they have a major impact on how fast a car can go. Teams deploy different types of tires based on track conditions and desired speed/durability tradeoff (some tires can increase top speeds at the cost of durability). Because teams have these options, we wanted to see if winning teams had different strategies for when to make their final pit stop to get the last fresh set of tires for their final push to the top of the podium. We will use the percentage of the total number of laps, as each race may have a different number of total laps, and each lap may not be exactly the same distance for each course. Once again using the Kruskal-Wallis test, we get a p-value of 0.327, with a chi-squared value of 10.297 and 9 degrees of freedom. Therefore, we again fail to reject the null hypothesis and conclude that there is not enough evidence to support a correlation between last pit stop time and finishing position. **Figure 16** shows the boxplot for this categorical distribution, and we see that almost every racer makes their final pit stop around 55-60% of the race duration. This is because pit stops are often correlated, as each racer obtains a similar amount of wear to the tires and thus needs to pit roughly around the same time during a race.

There is not a major difference in race position when it comes to pit strategy, namely number of pit stops and when teams pit for the last time in a race. However, average pit stop time is something that clearly correlates with winning, as shown in our Kruskal-Wallis hypothesis test.

Therefore, we can confidently reinforce our earlier findings that pit stops are a key measure of performance for teams to improve if they want an edge in final race position.

Constructors:

As stated previously, our analysis into whether constructors differ in their final race position led us to utilize the Kruskal-Wallis test over ANOVA. Our decision to elect for this test lies in three observations made during the EDA stage. The first observation lies in discrete ordinal structure of placement data. That is, race placements are not continuous (i.e. a driver cannot finish anywhere in between first and second, second and third, and so on). While ANOVA is equipped to handle discrete data on its own, the next to assumptions make its use in this context obsolete. The final two observations that prevented us from using ANOVA to model differences in constructor performance across races lies in the distributions of said performances. First, ANOVA assumes normality across all group distributions, which is not the case when looking at constructor-wide final race position distributions. Finally, ANOVA also relies on equal variances between groups, which again, is violated by our data.

Moving now into the results of our Kruskal-Wallis test (**see figure 17**), our null hypothesis for the test states that there are no significant differences in the final placement of constructors across races, while the alternative hypothesis states that there are in fact differences. Our results show a chi-squared value of 267.1 on 37 degrees of freedom, yielding a p-value of 2.2e-16. This p-value, being far smaller than our rejection threshold of 0.05, allows us to reject the null hypothesis and conclude that there are significant differences in constructor performance across Formula One races.

Chi-Square Test for Independence:

Our analysis has shown that certain constructors have dominated since the turn of the century, showing how important car-building and team maintenance are in F1. We also know intuitively that individual drivers are important, as they are the ones actually performing during the race. We hope to determine if either of these fields correlate with individual tracks when it comes to winning. Since 2000, Formula One has seen 74 unique drivers, 23 unique constructors, and 35 unique circuits. We will use a chi-square test to find out if certain drivers/constructors do better on certain tracks, and if it is a useful predictor for race performance. We followed the data subsetting method discussed in the “Statistical Methods” section, and performed the test with the null hypothesis being that there is no significant relationship between the two variables. We achieve a similar outcome for both. First, looking at the relationship between drivers and circuits, we get a p-value of 1 when we include all 74 drivers, but when we filter the contingency table to only include the 18 drivers who have won a race in the sample, we get a new p-value of <2.2e-16. These numbers are completely backwards, and they show the sample imbalance for F1 drivers. Since 2000, there are 3 drivers (Verstappen, Hamilton, and Vettel) who have won a large portion of the races. Moreover, these drivers often win the same races year after year. For example, Lewis Hamilton has won the British Grand Prix at Silverstone 9 times, and Max

Verstappen has won the Dutch Grand Prix at Zandvoort three times in a row since he has been racing. These results are shown in the heatmap (**figure 18**). Therefore, it is clear that there is a strong relationship between driver and track when it comes to winning, but this is because the winners of races are a small sample of the total drivers that enter a race.

We get a similar result when doing a chi-square test for independence for constructors and circuits. When including all constructors, we get a p-value of 1, but when filtering for just the 9 constructors who have won a race since 2000, we get a new p-value of 5.336e-08. Therefore, we fail to reject the null hypothesis when looking at all constructors, but reject the null hypothesis when just looking at past winners. **Figure 19** reflects the same trend that we saw when looking at drivers, which is that there is a correlation between constructor and track, specifically for the "Big Three" constructors that we analyzed earlier. Both Red Bull and Mercedes have their tracks where they have found a clear advantage over the rest of the field. This is information that can be used by other teams to determine what strategies are working for certain tracks, and if it is more the racing style of the driver, or the game plan of the team. Because the p-values for the two tests were so similar, it is not possible to conclude whether the driver or the constructor is more influential on winning at a certain course.

Sprints:

When analyzing the relationship between final sprint position and final race position, we elected to take a two-pronged approach. The first approach included the use of an ordinal logistic regression (OLR) to predict the transition probabilities between final sprint positions and final race positions. Our reasons for using OLR over a linear regression are outlined in the machine learning section of this report, but as a brief summary, we elected to use this method instead due to the discrete ordinal structure of the placements data, and their respective non-linear relationship.

The OLR results (**see figure 20**) reveal that our primary predictor, final sprint position, has a significant effect on final race positions, particularly in the better (lower) final positions. For instance, the coefficient for *positionOrder_sprint_L* is 5.95 ($p < 0.001$), which indicates a strong positive relationship between lower (better) sprint positions and lower (better) race positions. However, as the sprint positions begin to lag to higher values, the coefficients begin to decrease in magnitude, which reflects a weaker association between the two variables as we go down the podium. The intercept values in the results table help to further highlight the transition points between our two variables. For example, the threshold between a first place finish and second place finish (1|2) has a log-odds value of around -4.09, while that of a tenth and eleventh place finish (10|11) increases to around 0.39. This dramatic increase indicates that as sprint placements worsen, the odds of finishing higher in races worsen as well. As a final note, our model achieves a residual deviance of around 1664. When comparing this value to our models degrees of freedom of 261, there is clear evidence of a poor fit. The most likely reason for this lies in the fact that our model was unable to capture additional variation in the data. It is difficult to say with accuracy what this additional variation is, but we conjecture that it must have something to do with external factors like strategy, track conditions, and weather.

The second part of our analysis into the relationship between final sprint position and final race position led us to using Spearman's Rank correlation to quantify the relationship between these two variables. Spearman's correlation was chosen to model this relationship due to the ordinal nature of both final race and final sprint positions. Additionally, we noticed a slightly non-linear relationship between the two variables, which a Spearman Correlation value can account for. The null hypothesis for our test ($\boldsymbol{\rho = 0}$) posits that there is no association between sprint race performance and final race performance, while the alternative ($\boldsymbol{\rho \neq 0}$) states that there is an association between the two. The test results (**see figure 21**) yielded a correlation ($\boldsymbol{\rho}$) of 0.485, indicating a moderately positive relationship between final sprint position and final race position. In other words, better performance in sprint races (lower placements) is typically associated with better performance in final races (lower placements). The test yielded a p-value of 2.2e-16, which is far lower than our rejection threshold. Therefore, we are able to reject the null hypothesis and conclude that there is indeed a significant, positive relationship between final sprint position and final race position. While the test yielded significant results, it should not be accepted as a perfect predictor of race performance. Rather, we stress that the effects of other factors like race strategy, weather conditions, and driver skill are all likely contributors to the relationship between sprint and race performance.

Starting Grid Position:

For analyzing the relationship between grid position and finish position, we decided to use the Spearman Rank Correlation test based upon our ordinal rank data. We set up our null hypothesis ($\varrho = 0$) to be that there is no monotonic relationship between grid position and finish position with our alternative hypothesis ($\varrho \neq 0$) being that there is a monotonic relationship. With our dataset as discussed previously, our Spearman test yielded a correlation (ϱ) of 0.725 that indicates a high correlation between grid position and finish position. This strongly suggests that a better starting position from qualifying leads to a higher race finish. Our test output a p-value < 2.2e-16 (**See figure 22**). Since this is well below our significance level, our correlation is statistically significant, and we can reject the null hypothesis that there is no monotonic relationship between finish position and grid position. This model does explain a substantial portion of the variation in finish position with a higher ϱ value than we found in our Sprints analysis but it is still not a perfect predictor for race performance. There is still substantial variation in finish position not explained purely by grid position where other factors such as car, race strategy and driver ability come into play.

To investigate our theory that there is a strong relationship between finish position and grid position for the top and bottom for start position, we used rollyapply in conjunction with our Spearman correlation test again to test the correlations more locally. With a window setting of 5, we calculated the correlation coefficients for our entire dataset and then took the average for each grid position and plotted them (**see figure 23**). The graph shows a higher correlation between finish position and grid position for grid positions 1-4 and 18+ than the middle of the pack which supports our hypothesis that there is a strong relationship between finish position and grid

position for the top and bottom for start position. The reasons for this could be that drivers that qualify for the top few positions are the best based upon their abilities and car and perform more consistently than middle of the pack drivers. Similarly, those at the back for qualifying are more likely to be in the back at the finish due to a variety of factors like driver ability and car quality. Based upon this, we would need to use additional information to improve the accuracy of our model to predict the middle of the pack driver's finish position. Our research question focuses more on winning or placing highly but it is still important to note that the correlation between finish position and grid position does vary greatly depending on the driver's grid position.

Conclusion:

Winning in Formula One is a challenging task. There are a myriad of factors that could affect a driver's performance on race day including their car, race strategy, pit stops, track type, driving abilities, grid start position, weather and much more. In our paper, we analyzed many of these factors with the goal of predicting success on the track and found quite a few that correlate strongly with success.

The constructor (team) is an important factor in determining race performance. We found that the top three teams consistently outperformed the rest over the last four years with roughly 50% of their driver places in the top 5 positions. Additionally, these top three constructors and the 18 drivers who have won a race in our data sample perform better on certain tracks with particular drivers routinely winning the same race year after year like Lewis Hamilton winning the British Grand Prix 9 times.

Driver sprint finish, pit stop times and grid position are additional predictors of race performance. We found a moderately strong correlation between sprint finish and overall finish position for drivers with a statistically significant correlation value of .485. For pit stops, we determined that a higher finish position is correlated with a lower average pit stop time. This suggests that higher performing teams have smoother, faster pit stop processes as well as cars that are highly dependable. Finally, grid (start) position provided one of our strongest correlations with finish position based upon our statistically significant correlation value of .725. We determined that this relationship is weaker for the middle of the pack drivers who have more variability in their finish position than drivers in the top or bottom of the grid.

Predicting success in Formula One cannot be done with simply one factor. The features we analyzed showed promise in predicting success but none were perfect. Combining several features into one prediction tool would likely yield more accurate predictions. We can hypothesize that a driver from the big three constructors with the same grid position as a driver from another constructor should have a higher predicted finish based upon the influence the constructor has on final position. Using more complex models would allow us to confirm this suspicion and more precisely predict race outcomes in Formula One.

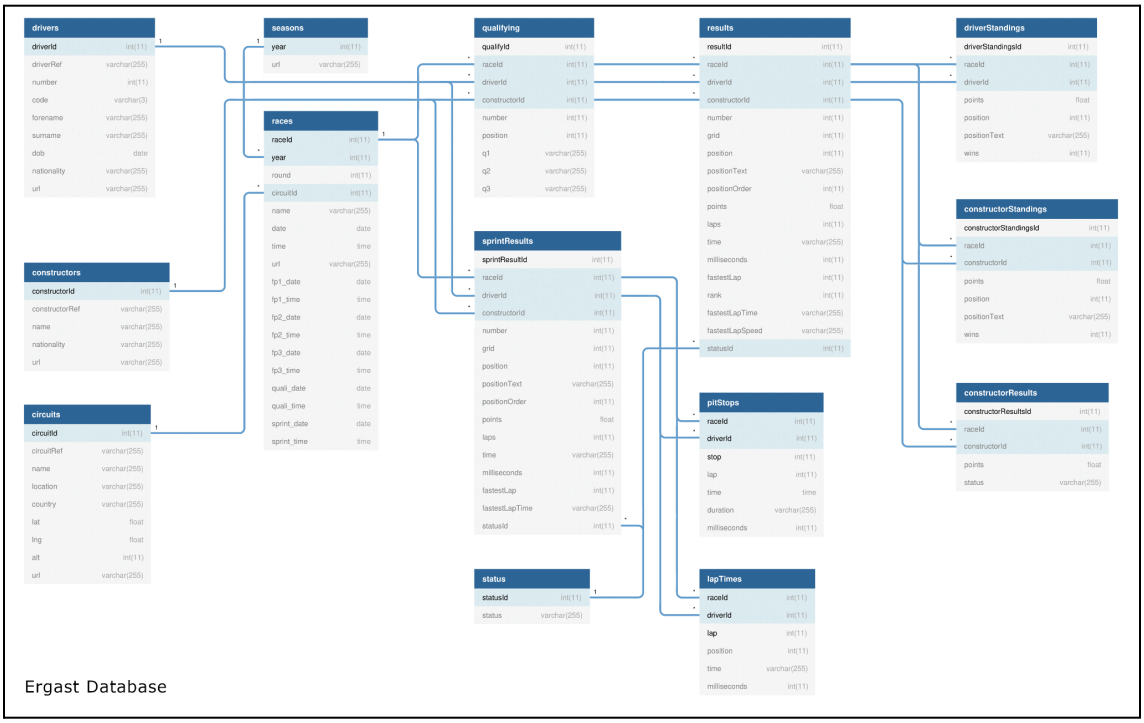
Future work could incorporate a time series approach as well since the sport of Formula One is ever changing. The cars have evolved over the years driven by technological innovations and rule changes around safety and refueling during races. These car innovations have increased

safety and handling while not necessarily increasing raw speed due to imposed engine limitations. Drivers are using more sophisticated training techniques to heighten their reflexes and skills. There are further factors that we did not investigate such as weather, altitude and road quality which could affect constructors and drivers differently. Incorporating as many relevant components that are relevant into a future model would likely lead to better predictions.

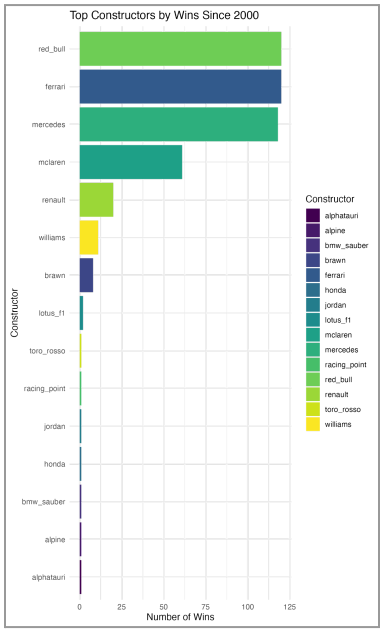
Appendix

(Figure label located above the figure)

- Figure 1: ER-Diagram of F1 Data

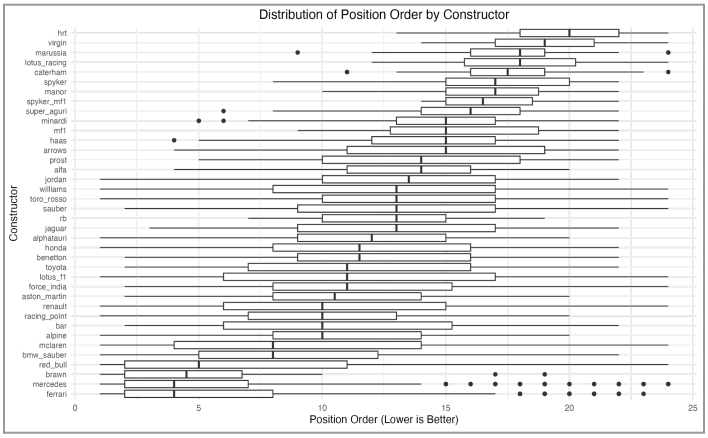


- Figure 2: Top Constructors by Wins



- Figure 3: Position Order Distribution of ‘Big Three’ Since 2000

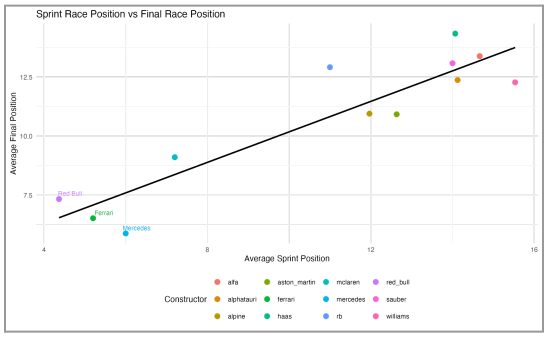
- Figure 4: League-wide position order distribution since 2000



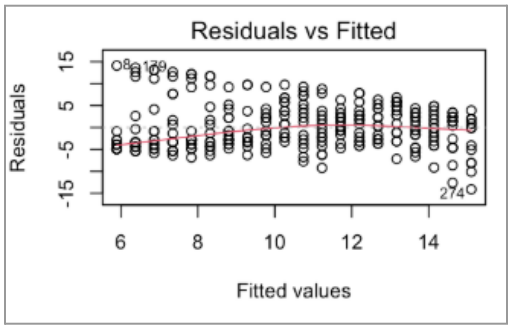
- Figure 5: Top Constructor Budgets (2015-2019)

F1 Teams	**Budget Between 2015 & 2019**				
	2015	2016	2017	2018	2019
Mercedes	$527.6M	$352M	$352.1M	$400M	$484M
Ferrari	$474.7M	$483.3M	$295.3M	$410M	$463M
Red Bull	$532.5M	$286.2M	$284M	$310M	$445M
McLaren	$528.3M	$246.4M	$240.8M	$220M	$269M

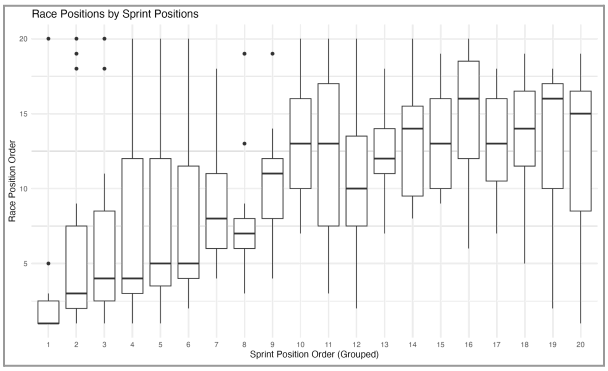
- Figure 6: Scatterplot of Sprint vs. Race Position (By Constructor)



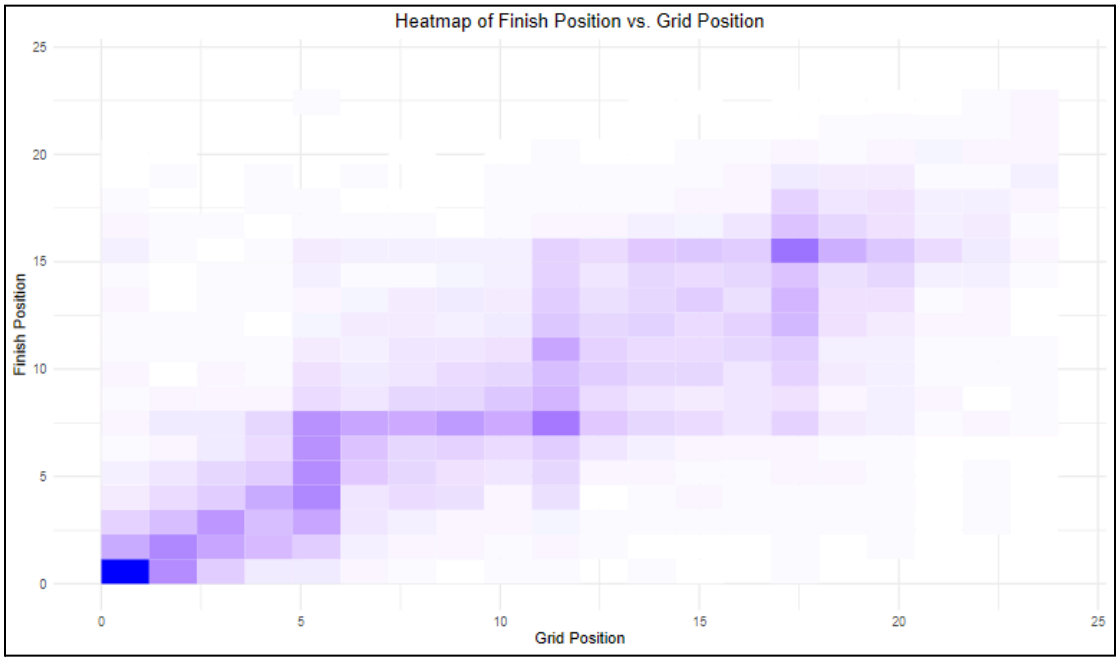
- Figure 7: Residuals Plot of Regression (Sprint ~ Final)



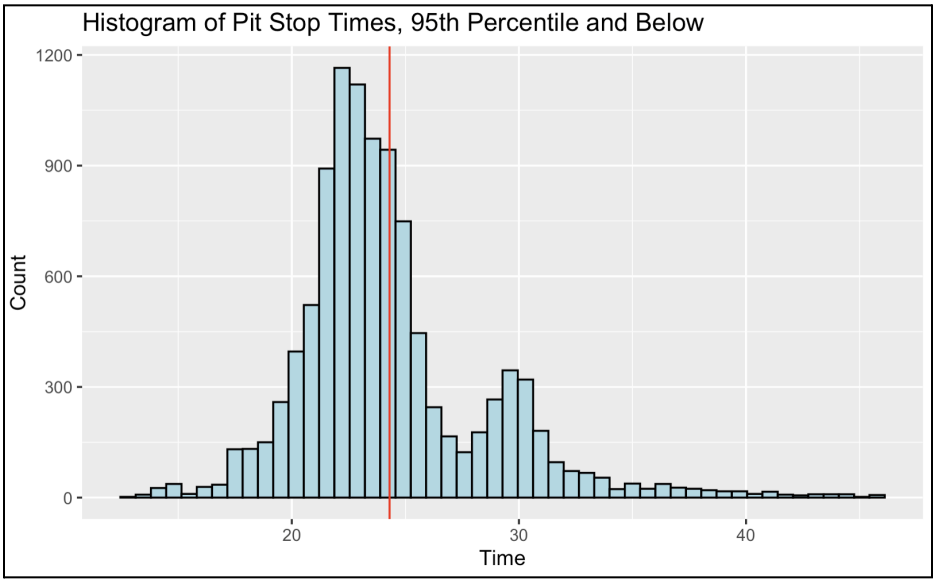
- Figure 8: Boxplot of Race Positions by Sprint Position



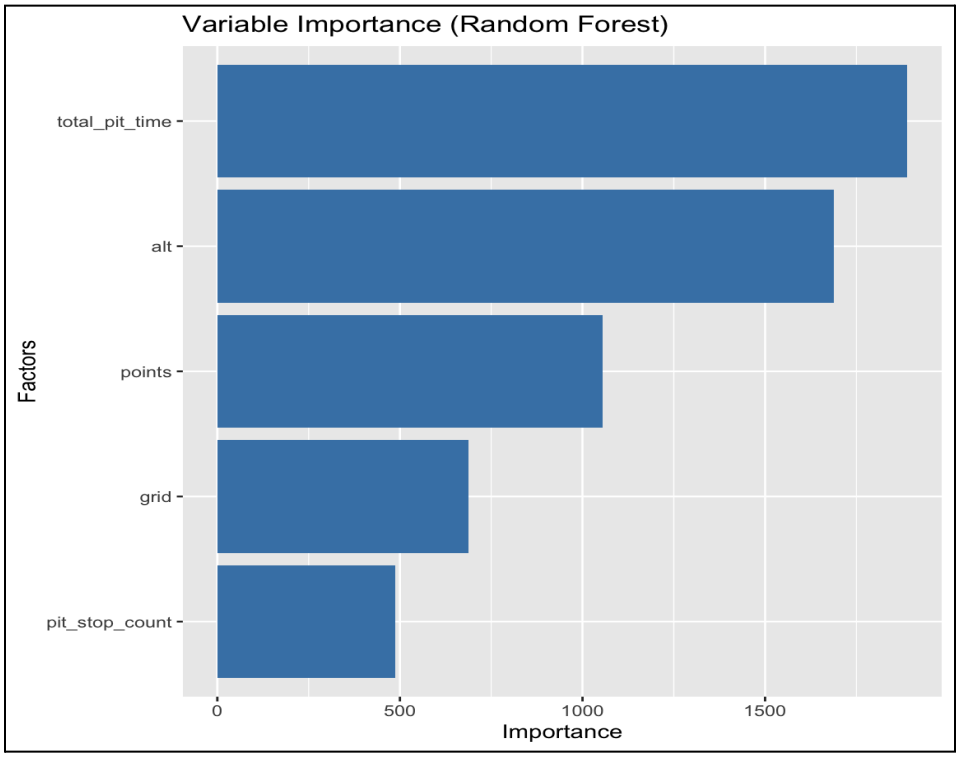
- Figure 9: Heatmap of finish position versus grid position



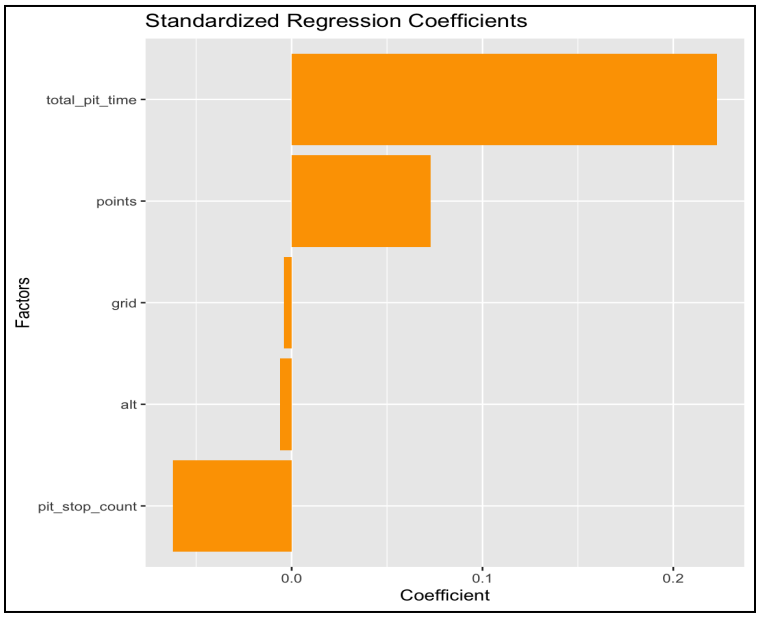
- Figure 10: Histogram of average pit times for a given driver, for a given race, with outliers removed from right tail. Red vertical line is the mean of the distribution.



- Figure 11: Random Forest Feature Importance



- Figure 12: Linear Regression and Random Forest Feature Importance



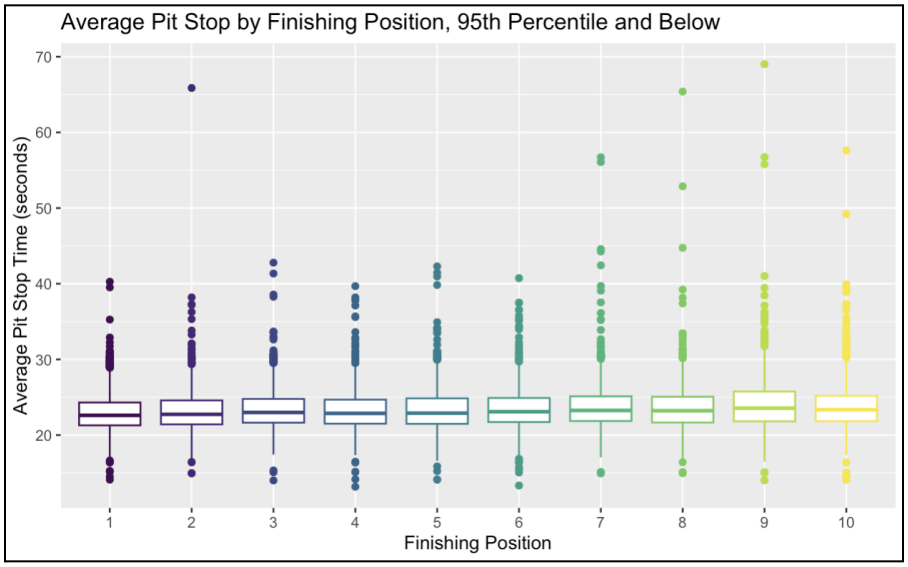
- Figure 13: R-Squared Value for Multiple Regression

```
##
## Call:
## lm(formula = milliseconds ~ alt + grid + pit_stop_count + total_pit_time +
##     points, data = full_data_scaled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5125 -0.0995 -0.0905 -0.0794 10.2295
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.451e-14  5.989e-03   0.000    1.000
## alt            -5.990e-03  6.007e-03  -0.997    0.319
## grid           -4.141e-03  6.501e-03  -0.637    0.524
## pit_stop_count -6.231e-02  6.929e-03  -8.992   <2e-16 ***
## total_pit_time  2.230e-01  6.576e-03  33.915   <2e-16 ***
## points          7.273e-02  6.850e-03  10.616   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9752 on 26513 degrees of freedom
## Multiple R-squared:  0.04907,    Adjusted R-squared:  0.04889
## F-statistic: 273.6 on 5 and 26513 DF,  p-value: < 2.2e-16
```

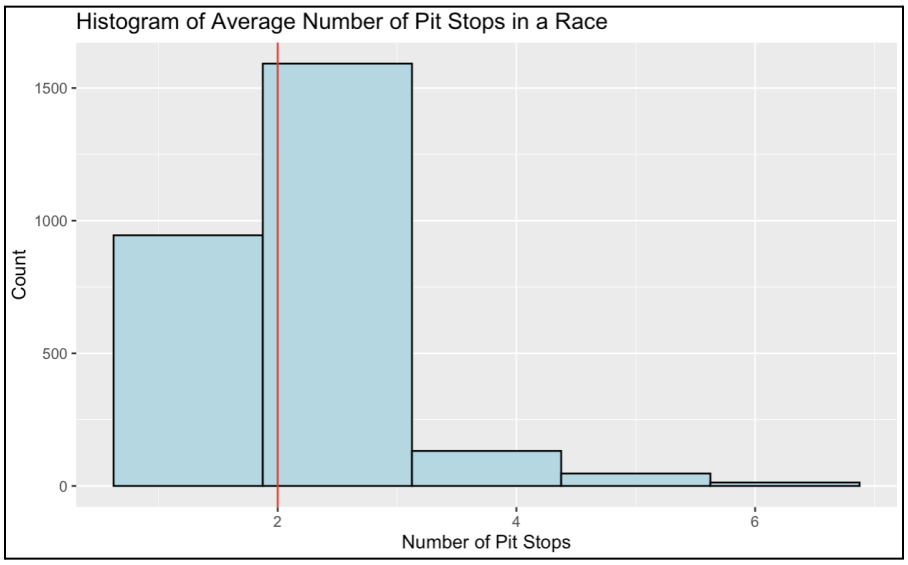
```
rf_importance
```

```
##                %IncMSE IncNodePurity         Factor
## total_pit_time 54.74657     1889.8984 total_pit_time
## alt            33.90366     1688.9101            alt
## points         29.83644     1054.8486         points
## grid           22.21784      688.7080           grid
## pit_stop_count 30.14769      486.1489 pit_stop_count
```

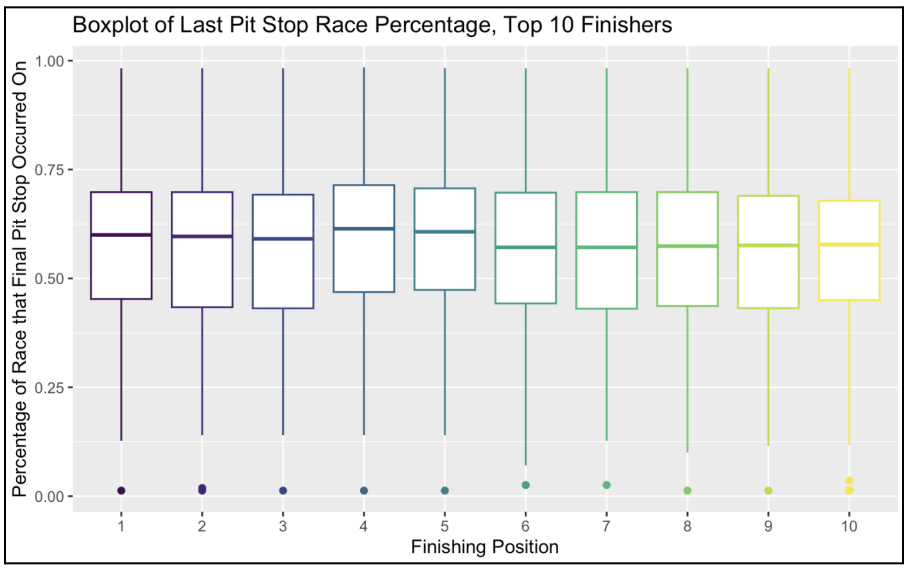
- Figure 14: Boxplot of average pit stop time by finishing position. Only includes top 10 finishes, and has outlier removal on pit stop times.



- Figure 15: Histogram of number of pit stops in a race, per driver. Red line is the cumulative average.



- Figure 16: Boxplot of percentage of race duration where drivers made final pit stop, grouped by finishing position.

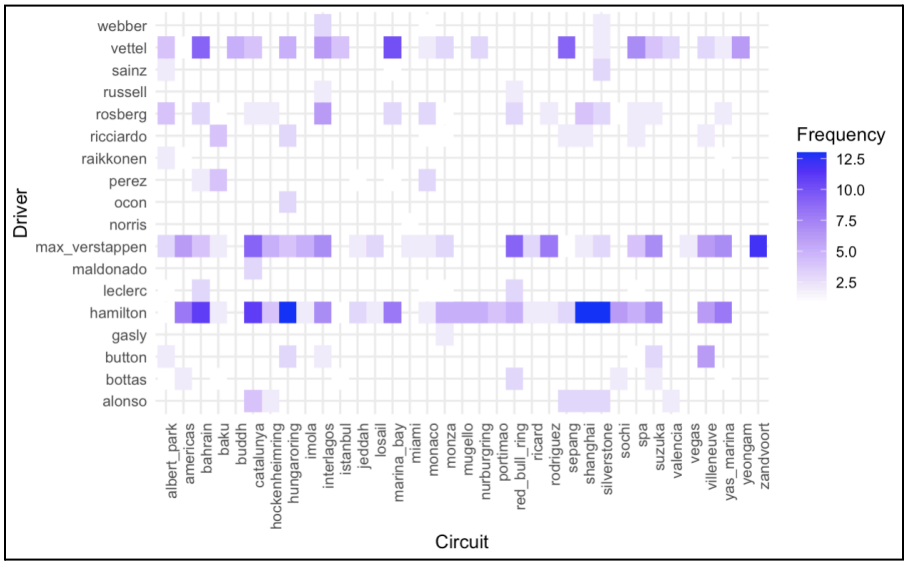


- Figure 17: Kruskal-Wallis Test of Final Race Placement By Constructor

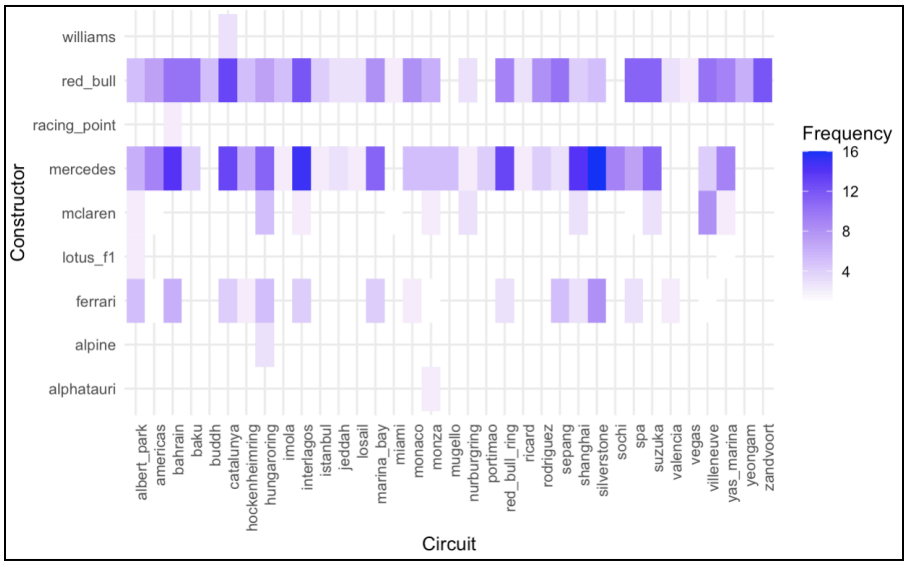
```
        Kruskal-Wallis rank sum test

data:  positionOrder by constructorId
Kruskal-Wallis chi-squared = 2761.1, df = 37, p-value < 2.2e-16
```

- Figure 18: heatmap of winning driver and winning circuit. Only shows drivers with at least 1 win in the sample.



- Figure 19: heatmap of winning constructor and winning circuit. Only shows constructors with at least 1 win in the sample.



- Figure 20: OLR Regression of Final Sprint Position on Final Race Position

```
Coefficients:
                          Value Std. Error  t value
positionOrder_sprint.L   5.950862     0.5943 10.01320
positionOrder_sprint.Q  -2.621266     0.5010 -5.23189
positionOrder_sprint.C   0.439519     0.4719  0.93131
positionOrder_sprint^4  -0.700482     0.4653 -1.50555
positionOrder_sprint^5   0.453042     0.4511  1.00439
positionOrder_sprint^6  -0.457235     0.4567 -1.00119
positionOrder_sprint^7   0.400806     0.4529  0.88501
positionOrder_sprint^8   0.127216     0.4529  0.28092
positionOrder_sprint^9  -0.304955     0.4514 -0.67563
positionOrder_sprint^10 -0.458352     0.4495 -1.01978
positionOrder_sprint^11  0.085483     0.4477  0.19093
positionOrder_sprint^12  0.368286     0.4467  0.82437
positionOrder_sprint^13  0.037985     0.4468  0.08501
positionOrder_sprint^14 -0.217313     0.4465 -0.48667
positionOrder_sprint^15  0.638604     0.4453  1.43399
positionOrder_sprint^16  0.507698     0.4420  1.14868
positionOrder_sprint^17 -0.032675     0.4321 -0.07562
positionOrder_sprint^18  0.007111     0.4207  0.01690
positionOrder_sprint^19 -0.107429     0.4337 -0.24771

Intercepts:
     Value    Std. Error t value
1|2   -4.0927   0.3534   -11.5825
2|3   -3.0143   0.2570   -11.7273
3|4   -2.2841   0.2079   -10.9848
4|5   -1.7160   0.1786    -9.6104
5|6   -1.2575   0.1610    -7.8106
6|7   -0.8582   0.1495    -5.7392
7|8   -0.5004   0.1422    -3.5198
8|9   -0.1801   0.1383    -1.3026
9|10   0.1165   0.1368     0.8519
10|11  0.3906   0.1370     2.8518
11|12  0.6494   0.1383     4.6947
12|13  0.9046   0.1408     6.4251
13|14  1.1629   0.1445     8.0459
14|15  1.4326   0.1499     9.5556
15|16  1.7224   0.1576    10.9298
16|17  2.0420   0.1685    12.1196
17|18  2.4100   0.1848    13.0421
18|19  2.8831   0.2133    13.5156
19|20  3.6295   0.2812    12.9068

Residual Deviance: 1663.734
AIC: 1739.734
```

- Figure 21: Spearman Rank Correlation Between Final Sprint Position and Final Race Position

```
        Spearman's rank correlation rho

data:  analysis_data$positionOrder_sprint and analysis_data$positionOrder_race
S = 2317418, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4850125
```

- Figure 22: Spearman Rank Correlation between Finish Position and Grid Position

$H_0 : \rho = 0$ (No Monotonic relationship between grid position and finish position)
$H_1 : \rho \neq 0$ (There is a monotonic relationship between grid position and finish position)

$$\rho = 0.725$$

$$\text{p-value} < 2.2 * 10^{-16}$$

- Figure 23: Average Correlation between finish position and grid position by grid position