Strategic Optimization From Qualifying to Pole Position in Formula 1*

An Analysis on Driver's Qualifying Lap Time for the Last 2 Decades (2006 - 2023)

Zhijun Zhong

April 17, 2024

In this study, I created predictive model from predicting the optimal lap time for Q2 eliminations that minimizes tire wear to using data from the first 2 round of the qualifying sessions to estimate the pole position lap time in the third qualifying round available in Kaggle. My finding revealed that it is possible to accurately estimate the pole position lap time and the optimal elimination time in the second qualifying. By using the predictive models, it can bring a better knowledge to teams and drivers on others while still own the ability to alter strategies in the following sessions. My aim is to be both competitive and preserving resources in qualifying in order to have the best outcome.

Table of contents

*Code and data are available at https://github.com/JerrZzzz/Strategic-Optimization-in-F1.

1 Introduction

Formula One has been the most advanced and well-known auto sport around the world in the last few decades. We have witnessed the development of technology from the immense power of V8 to V6 engine to complex aerodynamics but have unlimited potential. Every single team on the grid has only one goal which is to win races and every single driver on the grid has only one aim which is to win the Formula One driver world champion. Winning races in Formula One consists of skill, strategies, performances, and luckiness. There are 15 to 20 Grand Prix every year taking place all around the globe. Every single one of the Grand Prix contains 3 practice sessions, 3 qualifying sessions, and one big race session.

Some people generally believe that only the race session matters in a Grand Prix since it is the only one that decides the points board. However, 3 qualifying sessions are equally important because they can directly decide the grid order. Dennis Wesselbaum et al found in their research paper that the first grid position, AKA pole position, will give the driver a significant amount of advantage from the start of the race(Wesselbaum and Owen 2021). They also found that there is about a 10% increase in the probability for a pole position driver to win the race(Wesselbaum and Owen 2021). So, when we are watching Formula One races, we are not given the information about how quick the car can be around this circuit which is important information that they can use to make the qualifying sessions more interesting for

viewers. Thus, I decided to use a Kaggle dataset published by Rohan Rao to build a model and predict the lap time for the second and third qualifying sessions (Rao 2020).

According to Formula 1 sporting regulations article 39.2(Fédération Internationale de l'Automobile (FIA) 2024), the slowest 5 cars in the first qualifying session, aka Q1, will be "prohibited" from taking further part in the qualifying session. The slowest 5 cars in the second qualifying session, aka Q2, will be "prohibited" from taking further part of the qualifying session leaving 10 cars to the final qualifying session, aka Q3.

Moreover, in article 30.2 (Fédération Internationale de l'Automobile (FIA) 2024), there are only 13 sets of dry weather tires for every Grand Prix including 8 sets of soft compound tires, 3 sets of median compound tires, and 2 sets of hard compound tires. The harder the tire is the longer the tire is going to last and the slower the lap time is. As A. Tremlett and D. Limebeer stated tire saving whilst optimal in the lap time is considered the most important strategy in Formula 1(Tremlett and Limebeer 2016). The teams must make choices on what tire they should use in practice and qualifying sessions in order to make sure that there are enough tires for them to change in race sessions. Thus, I think that it is best to save tire in the second session of the qualifying for the race. I created my first model to find the slowest estimated lap time to enter Q3. It means that any driver who can be quicker than this lap time will be able to enter Q3 using Q1 as the predictor. Therefore, drivers can both save tires and enter Q3 safely.

Furthermore, the competition for the pole position is very important in Q3 for all teams. Thus, I created another model to find the estimated pole position time using Q1 and Q2 as predictors to estimate pole position lap time. After using the models, we found that there is a significant linear relationship between Q1 and Q2 which means that we can use a linear model on Q1 to estimate Q2. Moreover, pole position lap time mainly relies on Q2 rather than Q1, but we still need Q1 variable on the model. Our analyze uses historical race data and statistical modeling techniques to estimate those, thereby providing valuable insights for race strategy optimization.

The remainder of this paper is structured as follows. Section 2 is a section about the dataset I used in this paper. I carefully examined every variable of in the dataset and created graphs to visualize every variable. In Section 3, I presented two significant models for estimating Q3 entering lap time and pole position lap time. I also explained the prediction of these models. Section 4 visualize these models with graphs which can bring our understanding to a second level, while in the Section 5, I thoroughly discussed the application of my models and further action to strengthen using other currently unavailable variables like weather conditions and tire usage. Moreover, in the Section A, it consists of a graphic analysis of the quality of my models by using residual plots, qq-plots, and so on. Finally, Section B.2 includes all the resources and references I used in this paper.

2 Data

2.1 Data sources

The data used in this paper is extracted from Kaggle created by Rohan Rao(Rao 2020). This dataset is created with real Formula One data from real races. It contained from every circuit which had a Formula One race in history to every race in history. In this paper, I only used the races dataset and qualifying dataset to perform the model.

This paper is entirely made with the R program (R Core Team 2023) along with other packages which help me with graphing, table building, and so on. I used tidyverse(Wickham et al. 2019), dplyr(Wickham et al. 2023), and knitr(Xie 2023) to clean data from the original dataset. By using ggplot2(Wickham 2016), modelsummary(Arel-Bundock 2022), stringr(Wickham 2023b), and rstanarm(Goodrich et al. 2022), I am able to create the model and graph them. When it comes to downloading data for Kaggle, I used httr(Wickham 2023a), jsonlite(Ooms 2014), usethis(Wickham et al. 2024), devtools(Wickham et al. 2022), kaggler(Kearney 2024). readr(Wickham, Hester, and Bryan 2024) helped me to read CSV files to my workspace while styler styled my scripts(Müller and Walthert 2024). Some other used package are listed, janitor(Firke 2023), lubridate(Grolemund and Wickham 2011), gridExtra(Auguie 2017), here(Müller 2020).

2.2 Datasets

Table 1: Probability of each qualifying position to winning grand prix

Position	Wins	Percentage
1	176	51.612903
2	86	25.219941
3	37	10.850440
4	14	4.105572
5	10	2.932551
6	6	1.759531

The raw data contained a lot of datasets which are cross-referenced. For example, in drivers dataset, includes the detailed information of all drivers in Formula One history where each one of them is assigned to a driverId where we can use these driverId in later datasets without too much detail. In this way, it can effectively avoid oversizing one dataset, but it hardens our cleaning process. We combined the results dataset which includes all Grand Prix results and the qualifying dataset which contained all qualifying results in Formula One history. By combining them, we can count the number of different qualifying results which end up winning the Grand Prix. We are able to calculate the probability of each qualifying result that eventually

wins the Grand Prix in Table 1. Moreover, if we combine races, qualifying and drivers dataset and calculate the percentage for a driver to secure pole position and then win the Grand Prix, we will be able to watch the pole-to-win rate for each driver in Table 2.

Table 2: Dual Victory Rate for Each Driver

names	count	dual_victories	percentage
Hamilton	107	62	57.94393
Massa	16	8	50.00000
Rosberg	30	15	50.00000
Vettel	57	31	54.38596
Alonso	14	9	64.28571
Button	6	5	83.33333

Table 3: Q1 Position 11 vs Q2 Position 11 Lap Time

raceId	q1sec	q2sec
Id number assigned to every unique grand prix in history Range: 1-1100	Q1 11th fast driver lap time in Q1 Range: 54.388-130.529	Q2 11th fast driver lap time in Q2 Range: 53.995-129.377
1	86.026	85.504
2	95.260	94.769
3	96.443	95.975
4	93.479	93.242

In the race dataset, it contained the race information for every raceId. I can identify the year of the race which is very important to limited to recent races. In qualifying dataset, consists of all the lap times for Q1 to Q3 for all races from 2006 to 2023. However, the dataset only includes the position in the third qualifying meaning that the position for the first qualifying is not included. Thus, ordering the lap time for every race from the fastest to the lowest is necessary to identify the position for the first qualifying session. By using the qualifying dataset, I am able to create another dataset that involves the 11th fastest driver for Q1 and the 11th fastest driver in Q3. In this way, we can use the 11th fastest time in Q1 to predict the 11th fastest time in Q2(Table 3). There are 341 observations in this dataset where raceId go from 1 to 1110 and lap time varies from 54 seconds a lap to 130 seconds a lap. Moreover, I deviated from the original dataset to create a new one that records the total pole position each driver has in their career and collects the number of races where they turn their pole position into a race win. So, I can use this information to calculate the ratio of pole position to win for each driver.

Table 4: Q1 fastest driver vs Q2 fastest driver vs Q3 fastest driver Lap Time

raceId	q1sec	q2sec	q3sec
Id number assigned to every unique grand prix in history Range: 1-1100	Q1 fastest driver lap time in Q1 Range: 53.904-127.130	Q2 fastest driver lap time in Q2 Range: 53.647-126.609	Q3 fastest driver lap time in Q3 Range: 53.377-125.591
1	78.143	77.328	76.609
2	88.917	87.702	86.720
3	65.116	64.951	64.391
4	72.386	71.908	71.365

By using the same technique, I also create a dataset to record the fastest time in Q1, Q2, and pole position time in Table 4. There are in total of 340 observations in this dataset. Notice that I have changed the lap time of both datasets to a unit of seconds which will make it easier to visualize the change in the Section 3.

2.3 Measurements

This dataset is intended to present users with detailed information about every race in F1 history including race details, driver details, and even lap time for every driver in the race. We mainly used the recorded lap time for drivers in qualifying for races from 2006 to 2023. Initially, the lap time detail is obtained by the high-quality sensors on the Formula One race car and presented to the audience on the live broadcast. Eagest API where Rohan Rao extracts the dataset from collects the data directly from it without manual error. The lap time is stored in a format of "minutes:seconds.milliseconds". We transfer the format into a seconds-only format where it is a numeric value that we can compare and apply model on it. The outliers happen when the weather changes in any of the 3 qualifying rounds making any rounds particularly slow or fast. Since significant variables like tire usage and weather are not included in the raw datasets, we are unable to clearly find the outliers in our dataset. We decide to use a 95% quantile of the difference between any 2 qualifying sessions to reduce the impact of weather information on our model.

2.4 Visualizing Lap times

To take a close look at the distribution of the lap times, I created a histogram based on the median of q1, q2, and q3 lap times using the 10 fastest drivers from each qualifying session data(Table 4). We can see that the difference in the distribution is obvious. On average, the driver in the second qualifying session is about 0.7 seconds faster than those in the first

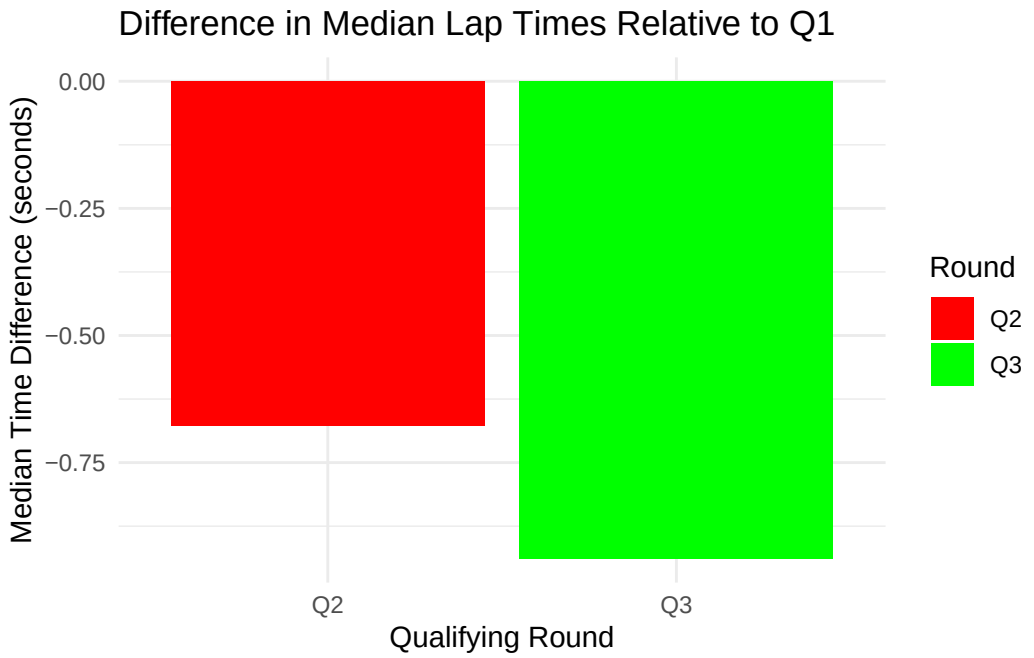

Figure 1

qualifying session. The gap is even bigger when it comes to the third qualifying session where it is close to 1 sec faster than those in Q1.

2.5 Missing data

The first missing data type in our dataset observed as "na" represents any mechanical failure or personal issue where the driver is not able to do a valid lap within a specific time. So, we decided that we could drop this kind of missing data so that it would not lead to any form of bias towards our model. We notice that this kind of missing data is not common in our dataset because those drivers who are not able to do a valid lap will count as infinitely large lap time meaning that they will always be at the bottom of the grid. However, in the paper, we focused on the 11th fastest driver in the first qualifying and the 11th fastest driver in the second qualifying which will not involve those missing values. The second missing value type is observed as "\N" which represents the eliminated driver's lap time in further qualifying sessions. If a driver had been eliminated in the first or second qualifying, we expect them not to take part in any further qualifying meaning they cannot have any lap time in these sessions. Thus, the author of the dataset decides to use "\N" to represent those missing lap times. So, by this nature of the missing value, we believe that dropping them will not bias our model.

2.6 Similar datasets

There are a lot of self-made datasets on different statistical websites like Kaggle where fans pull out the dataset all by themselves. I have done a lot of research and the dataset I used in this paper is by far the most comprehensive and accurate. However, Formula One data are strictly limited to each team. Further information is confidential to the team's database. So, it is very hard for us to collect detailed information like tire usage, car setup, and engine output mode where these factors are crucial when we analyze specific lap times for pole position and so on. Since the lack of that information, our predictions cannot be considered as accurate at a Formula One team level. But it is sufficient for the use of television broadcasts or building cliffhangers.

3 Model

The two models shown in this section below aim to predict lap times using different predictors and different types of models to fit. This is motivated by the fact that more and more teams didn't take qualifying sessions seriously since overtaking in race on track is easier and easier. Ending on top nowadays seems to have a smaller and smaller advantage over others. Thus, By creating these models, I want to raise the importance of qualifying sessions. After building models, I created different graphs to verify that the model that I built is a good fit for most observations in Section B including residual plots and distribution, and qq-plots.

3.1 Model set-up

I set up a null hypothesis and alternative hypothesis to test if there is a relationship between my predictor and response variable defined as follows. I plan to use the p-value to justify if we can reject the null hypothesis. I set a threshold of 0.5 where if the p-value is below 0.5, then we can reject the null hypothesis, and if the p-value is above 0.5, then we do not have sufficient evidence to reject it meaning that the relationship between predictor and response variable are weak and vice versa.

- Null Hypothesis: There is no relationship between X and Y, the model between X and Y does not have meaning.
- Alternative Hypothesis: There is a strong relationship between X and Y, the model between X and Y can predict general cases.

We can define our basic linear regression as Equation 1 while we use coefficients β_1 to create a slope for our linear regression line and β_0 to be the value of y when x is zero. We can also add a x^2 to create Equation 2 in order to make it a non-linear model so that we can compare it to the linear model and choose a better fit for our model. Moreover, by adding a second variable x_2(Equation 3), we can create a 3-dimensional linear model in which we use

2 predictors to estimate the response variable. With the application of these equations, we can properly measure our estimated lap time for both second qualifying elimination and pole position.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{2}$$

$$y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{3}$$

3.2 Model Uncertainties

We used at most 2 variables in our models to estimate a complex situation. We believe that both models are oversimplified with only 2 variables to generalize all circumstances. It will cause high uncertainties and failure to estimate lap time precisely. As we further discussed in Section A, because of the lack of important variables, the model will prefer to over-estimate when predicting pole position lap time and under-estimate when predicting second qualifying elimination time.

4 Results

4.1 Predict Q2 Elimination Time Model

When I used the 11th fastest driver in qualifying 1 to predict the elimination time for Q2, I introduced a linear model and a nonlinear model. In Table 5, the column on the left is the linear model while on the left is the nonlinear model. We define the linear model Equation 4 where $\beta_0 = 0.0723$ and $\beta_1 = 0.9937$. We define y as our response variable which is our estimated goal while x as our predictor which is the 11th fastest time in qualifying 1, and ϵ is a random error that occurs when there is variability in the response variable that isn't explained by our predictor or the randomness in the real data.

$$y = 0.0723 + 0.9937x + \epsilon \tag{4}$$

The nonlinear model is defined by Equation 5 where where $\beta_0 = 87.4653$, $\beta_1 = 207.4629$, and $\beta_2 = -0.1304$. So the full equation becomes Equation 5 while x, y, and ϵ are defined the same as above.

Table 5: Model of Elimination Time for Q2

	Linear Model Summary	Non-linear Model Summary
(Intercept)	0.0723	87.4653***
	(0.1232)	(0.0166)
q1sec	0.9937***	
	(0.0014)	
poly(q1sec, 2)1		207.4629***
		(0.2902)
poly(q1sec, 2)2		−0.1304
		(0.2902)
Num.Obs.	306	306
R2	0.999	0.999
R2 Adj.	0.999	0.999
AIC	114.4	116.2
BIC	125.6	131.1
Log.Lik.	−54.206	−54.104
RMSE	0.29	0.29

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

$$y = 87.4653 + 207.4629x + -0.1304x^2 + \epsilon \tag{5}$$

As suggested in the table Table 5, any coefficient with 3 stars behind it means that the p-value of this coefficient is smaller than 0.001. Any coefficient with a p-value smaller than 0.001 suggests that there is very little chance of this coefficient being zero. So in our model, the coefficient for the predictor as suggested with 3 stars tells us that the predictor will rarely be zero meaning that we can reject the null hypothesis where there is no relationship between the 11th fastest lap time in qualifying 1 and the elimination time. It is the same as the nonlinear model, where our predictor has a p-value smaller than 0.001 suggesting a strong relationship between the 11th fastest lap time and elimination time.

In Table 5, the value for R^2 and adjR^2 are both 0.999 with a sample size of 306. R^2 and adjR^2 suggested the fit of the model on our data. It means that 99.9% of my observations can be explained by our model. Thus, it is a good fit for both the linear model and the nonlinear model. However, an extremely high R^2 value often suggests a lack of sample size or over-fitting. With a 306 sample size, we cannot conclude with a lack of sample size. In consideration of our qualifying 1 lap time in unit second, we expect the R^2 value to be higher than normal since the lap time difference is not significantly high often resulting in a 0.3 or a 0.5-second difference. So it is normal to have such a high R^2 value because our prediction has to be accurate to the third decimal place.

We plan to compare two models using Bayesian Criterion and Akaike Criterion which is the AIC and BIC columns in Table 5 to find a better fit. Generally, if the value of AIC and BIC are low, it indicates a better model. However, there are no scales to measure how good the model is. It is often applied to both models and compares which one is lower suggesting a good fit. In Table 5, by comparing AIC and BIC, the difference is obvious suggesting the linear model being a good fit. It matched the principle of model parsimony which stated that all else being equal, the simpler the model is the more complete the test of hypothesis is. So a linear model is what we prefer when it comes to predicting elimination time.

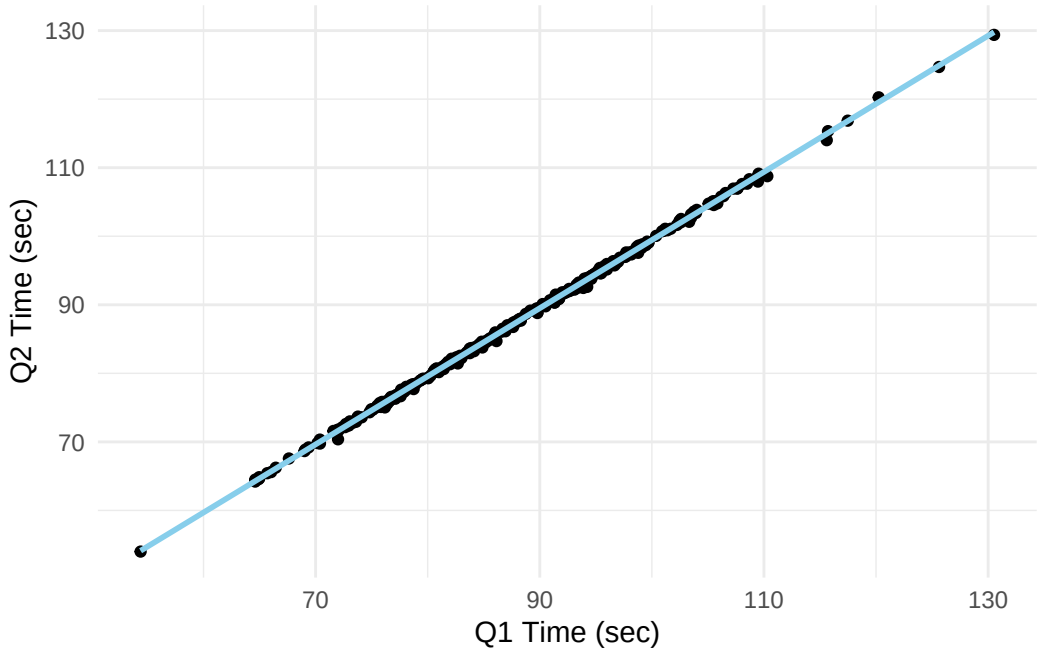

Figure 2: Fittness of the Model for Elimination Time in Q2

After creating a figure on the model for predicting the second qualifying elimination time, we found that the relationship between the 11th fastest time in qualifying 1 and the elimination time in Q2 is very strong. For every one-second gain in Q1 time, we expect to get a 0.99-second gain in Q2 time. However, our prediction is not always accurate due to the track conditions and so on. We can create a table to measure the uncertainty of our model in Table 6.

Table 6: Uncertainty Calculation on Model of Elimination Time for Q2

fit	lower_bound	upper_bound	uncertainty	input
69.63007	69.05674	70.20339	0.5733268	70
74.59848	74.02616	75.17080	0.5723207	75
79.56689	78.99525	80.13853	0.5716392	80
84.53530	83.96402	85.10659	0.5712835	85

Table 6: Uncertainty Calculation on Model of Elimination Time for Q2

fit	lower_bound	upper_bound	uncertainty	input
89.50372	88.93246	90.07497	0.5712542	90
94.47213	93.90058	95.04368	0.5715513	95
99.44054	98.86837	100.01272	0.5721744	100
104.40895	103.83583	104.98208	0.5731224	105
109.37736	108.80297	109.95176	0.5743937	110

In this Table 6, We have input a range of first qualifying time from 75 to 110 seconds. We discovered that the average uncertainty of predicted time is 0.572 seconds. It suggested that all 95% of the points will lay in the range of ± 0.572 seconds. So our predicted Q2 elimination time can be various in a range of 1 second or more. However, the fit of our model is reasonable since most of the differences of the position 11th between Q1 and Q2 are usually 0.6.

4.2 Predict Pole Position Time Model

A more complex model in Table 7 is used when we are trying to predict pole position time for every race where we used 2 variables both the fastest time in the first qualifying and the fastest time in the second qualifying. This model is defined as a multivariable Equation 6 where $\beta_0 = -0.0783$, $\beta_1 = -0.0541$, and $\beta_2 = 1.0529$. x_2 is the main variable we use to form the regression line which is the fastest lap time in the second qualifying while x_1 is defined as the support variable to x_2 to adjust its position which is the fastest lap time in the first qualifying session, and ϵ is the random error term representing the variability and real-life errors.

$$y(x_1, x_2) = -0.0783 + -0.0541x_1 + 1.0529x_2 + \epsilon \tag{6}$$

From model summary Table 7, it suggests again like our first model that the R^2 and R^2Adj are very high at 99.8%. It means that 99.8% of our current data can be explained using this built model. With a sample size of 274, we cannot be confident that our sample size is big enough to prevent it from over-fitting. However, as mentioned above, we expect R^2 and R^2Adj to be big since the y scale is small. Moreover, 3 stars behind the coefficient of x_2 (fastest lap time in second qualifying) mean its p-value is smaller than 0.001 which suggests an extremely weak possibility of it being zero. It suggests a strong relationship between the fastest lap time in second qualifying and our pole position time. However, there are no stars behind the coefficient of x_1 (fastest lap time in the first qualifying session) showed us that the p-value is probably higher than 0.1 meaning there is a great chance it can be zero. Furthermore, the intercept row displays no star meaning the intercept value can probably be zero. Thus, to sum up, this model does suggest that there is a linear relationship between Q1 and Q2 time

Table 7: Model of Pole Position Time

	Predicting pole position lap time model
(Intercept)	−0.0783
	(0.2386)
q1sec	−0.0541
	(0.0764)
q2sec	1.0529***
	(0.0770)
Num.Obs.	274
R2	0.998
R2 Adj.	0.998
AIC	411.8
BIC	426.3
Log.Lik.	−201.916
RMSE	0.51

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

with pole position time. However, it also tells us that the relationship between the fastest lap time in the first qualifying is weak to the pole position lap time. From the p-value of the interception, we understand that when we achieve a very quick lap in the second qualifying, the pole position lap is still predicted as a fast lap.

Table 8: Uncertainty Calculation on Model of Pole Position

fit	lower_bound	upper_bound	uncertainty	inputq1	inputq2
69.44370	68.43699	70.45041	1.006711	70	69.63007
74.40412	73.39935	75.40888	1.004769	75	74.59848
79.36453	78.36098	80.36809	1.003555	80	79.56689
84.32495	83.32188	85.32802	1.003072	85	84.53530
89.28537	88.28205	90.28869	1.003320	90	89.50372
94.24578	93.24148	95.25008	1.004300	95	94.47213
99.20620	98.20019	100.21221	1.006009	100	99.44054
104.16662	103.15817	105.17506	1.008442	105	104.40895
109.12703	108.11544	110.13863	1.011597	110	109.37736

Since the linear model for estimating pole position used 2 variables, it is very hard to visualize it in the paper, but we can still look at the uncertainties applied to this model. I have simulated the same time for the first qualifying as our first model Section 4.1. I also assumed that the first position followed the same pattern discussed in the first model Section 4.1. Comparing

it to the model above, we can see that the range for my uncertainty goes from approximately ± 0.572 to about ± 1.00. Even though that our first model is used to model the 11th position lap time, it can also be possible to predict position 1. Thus, the uncertainty brought by the multivariable model is significantly higher than a simple linear model in model 1 Table 6.

5 Discussion

As we discussed above, we expect to use the early sessions of qualifying to estimate both the elimination time for the second qualifying session and the pole position time. We found out that there is a positive correlation between the 11th-fastest driver in the first qualifying session and the 11th-fastest driver in the second qualifying session. Using our model in Section 4, we are able to predict a time for a driver to beat in the second qualifying. The importance of knowing the elimination time is to minimize tire wear in order to save tires for the race session in the future. Moreover, we also created a model using the top driver in the first two sessions in qualifying to estimate the potential pole position time in the last session. By knowing this, drivers will have a much bigger motive for contention for the pole position. In the following discussion sessions, we will discuss deeper into the importance of both models and give weaknesses and future steps to strengthen our models.

5.1 Analyzing Resource Management in Formula One

In the world of Formula 1, the factor that determines success or failure is not only the speed of a designed race car but also the strategy that a team uses. There are a lot of examples in history that fast cars fail because of the use of strategy. It includes the tire usage, the pit stop timing, and so on. In a real race, a set of tires that had not been used before can be much faster than a set of tires that had been used for a few laps. The degradation happens when tires have to handle extreme conditions such as high-speed corners, heavy braking, and downforces which increases tire wear. Over some time, as the rubber heats and expands, it will lead to blistering, visibly deteriorating the tire's surface(Reddy et al. 2022). The degradation of the tire happens not only in races but also in qualifying where cars have to perform on the absolute limit of itself and increase the degradation level.

The tire usage in a modern Formula One race weekend is limited by rules for only 13 sets of dry tires can be used by each driver in each team including 8 sets of soft compound tires, 3 sets of medium compound tires, and 2 sets of hard compound tires (Fédération Internationale de l'Automobile (FIA) 2024). The softer compound of the tire is, the more grip a car can have meaning faster lap times. However, a race weekend contained 3 practice sessions for the team and driver to get familiar with the track, 3 qualifying sessions to determine the grid position for the race, and one race session to help the team score points. As rule 30.5(h) stated in F1 at the end of each practice session, the team has to return 2 sets of their tires to the manufacturer (Fédération Internationale de l'Automobile (FIA) 2024). It means that at the end of the third

practice session, each team will have at most 7 sets of unused tires. Moreover, rule 30.5(h)(ii) suggested that each driver should always have 2 different compounds of tires by the start of the race. It restricts the team to return all the medium compound and hard compound tires before the race. Thus, a team can at most return 4 sets of medium or hard tires leaving at least one set of either medium or hard. As shown in Table 9, it demonstrates the strategy a team might use to manage the resources they have. The most commonly used strategy is the third one which gives a great variability on race strategy to any unforeseen situation that might arise in the race. Thus, we can see that for the third strategy, there is only one fresh soft tire for the driver to use in the third qualifying session. It means that drivers will only have one chance to get the best out of the car and compete with the rest of the grid.

Table 9: Qualifying tyre Usage Strategy

Sessions	Tyre_strategy_1	Tyre_strategy_2	Tyre_strategy_3
Tyres left from Practice	1M+5S	1H+5S	1H+1M+4S
Q1	1S	1S	1S
Q2	1S	1S	1S
Q3	2S	2S	1S
race	1M+1S	1H+1S	1H+1M+1S

Note:
S = Soft Tyre, M = Medium Tyre, H = Hard Tyre

We discovered that minimizing our lap time in the second qualifying session is the optimal option for us. If we can save our tires in the second qualifying session, we will be able to have another set of soft tires that we can use in the third qualifying session in order to fight for the pole position, when our driver makes a mistake on his first lap which they can have a second attempt on fighting the pole position. If we use the driver's practice lap time to be the predictor to estimate the first qualifying session in order to achieve the same goal for minimizing tyre wear will have too many factors and uncertainties to address since a lot of the teams hide their true speed till the qualifying session. Thus, using the first qualifying lap times to predict the elimination time will bring a team various benefits not only in the third qualifying session but also in the race later.

5.2 The Advantages of Securing a Pole Position

Some people always consider the race as the key to success in Formula One. When we look at the structure of every race, we notice that only the ones who achieved in the top 10 score points for themselves and their team. However, it does not mean that the other sessions are not as important as the race itself. There are a lot of drivers in history disappointed when they are not able to have a good lap in a qualifying session especially when they are fighting

for the pole position. The significance for us to build such a model to predict the time for pole position is not only to give the driver a better chance of winning points but also to have a great commercial value.

5.2.1 The Correlation Between Pole Position and Race Wins

When we look at a map of a circuit for Formula One, we can often see the grid position drawn as a white line at the start of the circuit. After the qualifying session, the fastest driver among the rest of the grid will start the race behind the first white line marking the pole position. Notice that the distance between the pole position and the second white line is small often within half to one car length designed to pack the cars together at the standing start of the race. However, even though the distance is small, in real races, it can have a big difference. Top motor racing drivers have the fastest reaction speed in the world. They will release the clutch within 0.2 seconds after the light went out meaning that the reaction speed is often smaller than 0.2 seconds. Let's use statistics to back up our argument. A Formula One car often travels from 0 to 100 in 2.6 seconds. If I am on pole position and react 0.1 seconds slower than the second car, when the time I get to 160km/h, I cover the distance of 53.42m, while the second car covers a distance of 57.78m resulting in a difference of 4.36m. However, 4.36m is about the same as a car's length which is the same as the difference between pole position and second position. Thus, it suggested that if I lost the pole position, I had to react 0.1 seconds faster than the pole position driver to get into the same start line as them till I accelerated to 160km/h.

Moreover, if we take a look at the rate for a driver to achieve pole position and eventually win the grand prix on Figure 3, we will see that a lot of them have a percentage over 50% of it. It suggested that if a driver Secured the pole position, there was half a chance that they would win the race. For drivers like Max Verstappen, they have a staggeringly over 75% of pole-to-win rate meaning that there is almost no one able to win the Grand Prix if they won the pole position. If we disregard the drivers themselves, from 2006 to 2023, there are over 50% of all victories of the grand prix came from pole position in qualifying in Figure 4. 93% of all Grand Prix wins are born in the top 6 of the qualifying. Thus, we can conclude the higher position a driver achieves in qualifying will give the driver a better chance of winning the race, especially achieving the pole position.

In addition, different drivers will show different patterns of their performance under the high-pressure environment of qualifying, where milliseconds decide the grid positions. For instance, some drivers can be strong in the early sessions when the track is less crowded, allowing them to deliver strong lap times in less competitive environments while other drivers tend to perform under pressure in the third qualifying round, where the track is sticker and have a better grip. Thus, analyzing physiological data can illustrate how drivers' body figures, like adrenaline and heart rate, will impact their behaviors. Combining these personal elements into predictive models can improve the accuracy of estimating lap times and provide teams with personalized

strategies for each driver's unique strengths, where they can adapt to the driver's strengths and avoid their weak spots.

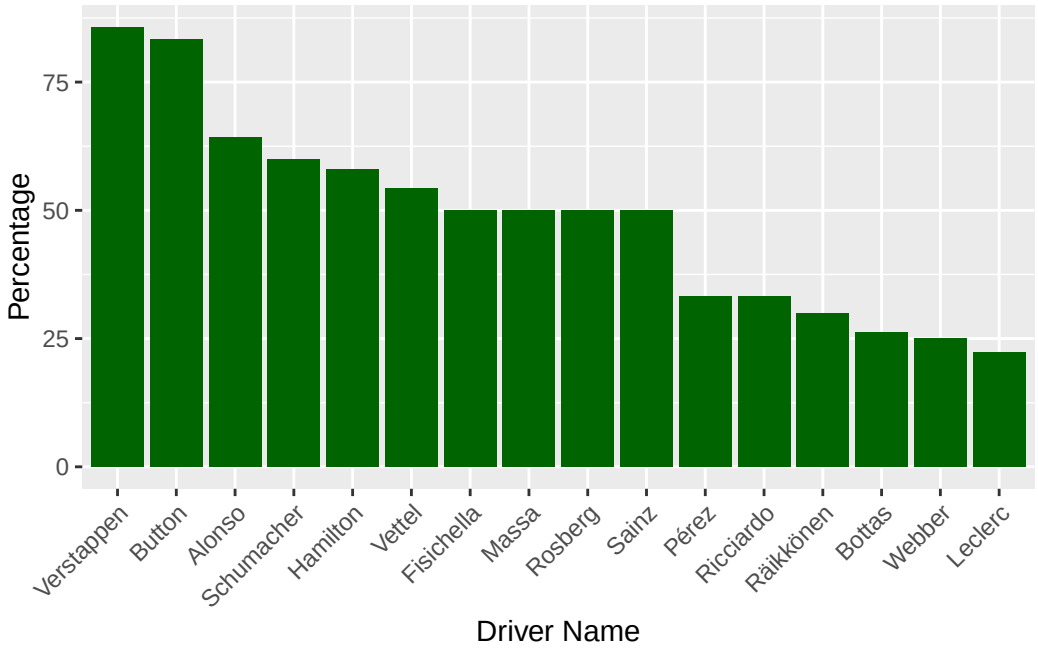

Figure 3

5.2.2 The Financial Implications of Achieving Pole Position

Achieving pole position is not just a matter of prestige but also a significant financial advantage. Teams that secure the top spot on the grid benefit from increased exposure and visibility, which often translates into enhanced sponsorship opportunities. For modern Formula One races, not only the top 3 of the race winners have a podium celebration, but also a podium for those drivers who achieve the top 3 of the qualifying. All 3 drivers will be interviewed lively around the globe along with cars which filled with sponsorship. Thus, many sponsors gave their team a target to reach in order to keep the exposure and advertising them. This visibility is much more important for smaller teams than big teams where financial sponsorship and prize money are linked to race performance. A study by Loughborough University found out that the sponsors have more exposure to cars and the clothes drivers wear(Tan and Pyun 2018). It suggested that if a team wants more exposure on their car and driver, the most obvious way is to finish on a podium that gives the longest clip of the car, driver, and the team. Thus, pole positions and strategies are significant to achieve them. Moreover, for teams that achieve the constructor championship, FIA will present prizes to teams at the end of each season. The prizes are not fixed every year which depends on the total income of FIA. Moreover, for a team to improve their car performance for 0.1 seconds will cost them 10M dollars on average.

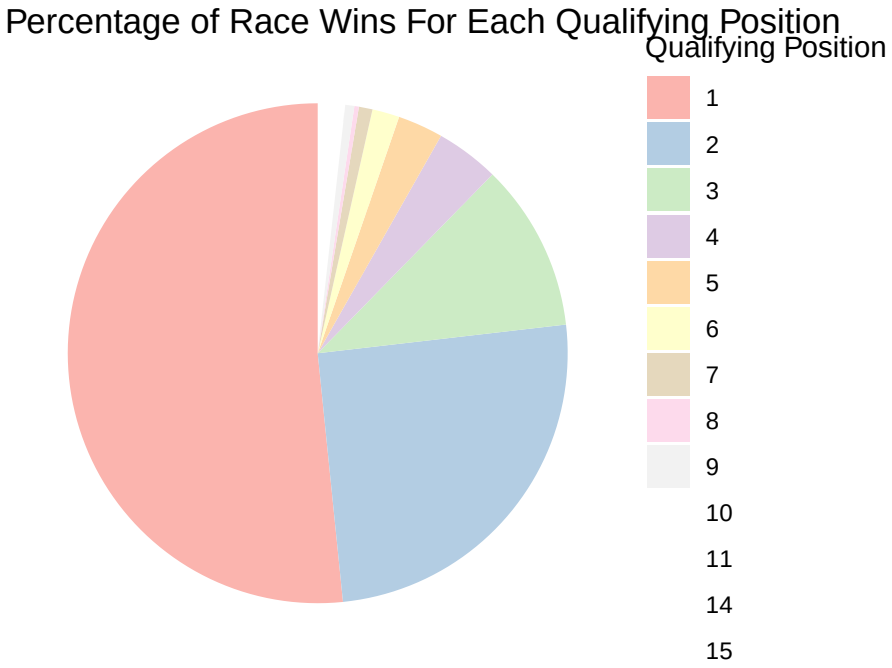

Figure 4

It means that if we are able to accurately predict pole position time while keeping a good tire strategy using second qualifying elimination time, we are possible to save a lot of money and keep our car in the front of the grid while having a slower car.

Therefore, both models can bring the team a bigger chance of achieving a pole position which can result in a podium finish while giving the team the best strategy to compete in the following races and financial advantages.

5.3 Weaknesses and next steps

Even though there are a lot of advantages when we can predict lap time for second qualifying elimination and pole position, there are still a lot of weaknesses in our model and even in our paper. Although our original dataset contained a lot of information in F1, in order to predict the time precisely need more variables like weather information, and track temperatures to help us understand these underlying patterns behind the model. Moreover, we used qualifying information from 2006 to 2023 which is a long period. There are a lot of regulation changes during this period of time meaning that the car's performance is not accurate which we think it can cause bias when we use them. The results show that the uncertainty for our model is high. We also considered using a more recent period, but since there are only about 20 races each year, our model will not be properly trained with such a small sample size which will result in overfitting and bias when estimated with poor generalization. Moreover, the uncertainty we

have discussed in the Section 4 is large. The uncertainty of 95% of our data points will lay in a range of ± 1 second. It means that our prediction will have an uncertainty of 2 seconds. In a Formula One race, the difference is from the top of the grid to the bottom of the grid. So, our prediction here can only be used as entertainment rather than a precise lap time. In addition, the Eagest API website where Rohan Rao created the dataset from announce that they will stop updating after 2024 suggesting that the dataset will not be up to date since then.

Thus, our next step is to improve the dataset as much as we can. We can manually input the dataset for each race with specific track conditions, tire usage, weather, and so on in order to improve our model. We can also select a specific driver instead of selecting a period to help us avoid bias. In addition, using track conditions, weather, and tire can help us filter possible outliers in our dataset to improve it. We can also use the updated information for Eagest API in the future to test and further train our model to make it more precise. Therefore, by taking these steps, we are able to address the above weaknesses and improve our model.

A Appendix

This section provided further graphs to support the main aim of this research paper. It includes further detail on addressing the qualify of the model and any further experiments or data that provide additional contexts to the paper.

B Model details

In the following section, I will discuss the quality of my model by using residual plots, residual distribution plots, QQplots along with scale location plots. By analyzing these graphs, we are able to discuss the further action that need to be taken and how to address the issue with our model.

B.1 Residual distribution check for estimating elimination time for second qualifying session

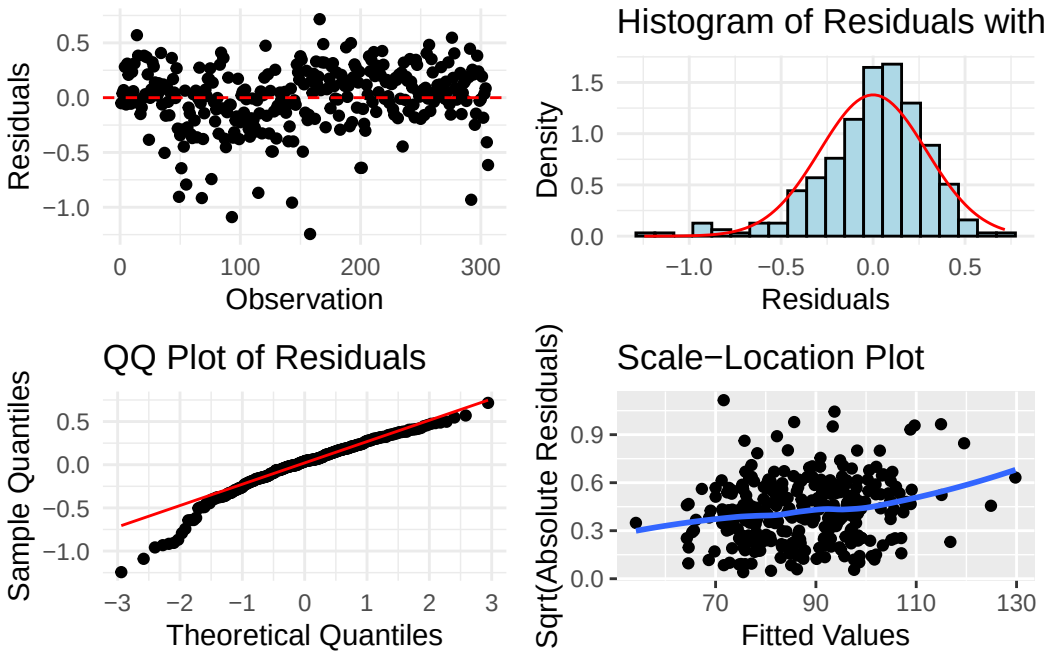

Figure 5: Examining how first model fits

By fitting the residual plot on the top left corner of Figure 5, we can see that all the points are randomly distributed along the y=0 line. It suggested that the residual for my data points is a good fit. However, we can see that the points below the red line have higher residuals than

the points above the red line. Some of these points are potential outliers meaning that they should be removed, but because of the nature of our dataset where it lacks important variables like weather information. So removing potential outliers is very hard in these circumstances. Thus, we can say that our model here will tend to underestimate rather than overestimate. Moreover, from the qq plot and the residual distribution plot, we can see that they did not fit as a normal distribution meaning that there are higher residuals on one side of the fit creating a potential underestimate. In addition, for the scale location plot, we found out that the blue line should be a horizontal line representing a constant variance. However, a curved line represents a different variance across the fitted values. It suggested that a different model can be considered to fit the values. A more complex model should be used to address the issue here, but the lack of variables in our dataset shows an insufficient variable in building models. So further steps need to be taken to improve models.

B.2 Residual distribution check for estimating pole position time for third qualifying session

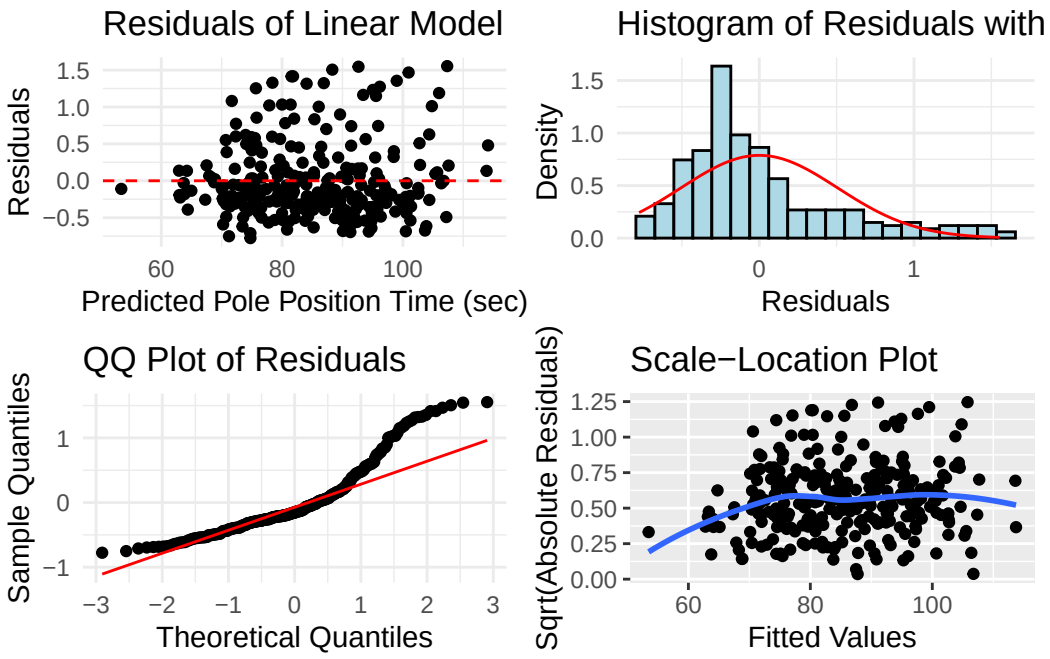

Figure 6: Examining how second model fits

We can see that through this Figure 6, the situation is the same as we mentioned above Section B.2. The residuals deviate to the overestimate side of the graph. Our model here tends to overestimate rather than underestimate in Section B.2. It suggested that there are more outliers where drivers do well in the second qualifying session and behave not so well in the third qualifying session. The reason behind this is that without critical information

like weather and track conditions. It is very hard for us to clean the outliers. For those high residual points that can be considered to be influenced by weather factors, they should be removed from our model, but we took them into account which will affect the training of the model. For the scale location plot, the blue line suggested that the variation is not consistent over the fitted values. So same conclusion should be drawn as above where a more complex model can be fitted.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*.

Fédération Internationale de l'Automobile (FIA). 2024. "2024 Formula One Sporting Regulations." https://www.fia.com/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Kearney, Michael Wayne. 2024. *Kaggler: Kaggle API Client*.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Müller, Kirill, and Lorenz Walthert. 2024. *Styler: Non-Invasive Pretty Printing of r Code*. https://github.com/r-lib/styler.

Ooms, Jeroen. 2014. "The Jsonlite Package: A Practical and Consistent Mapping Between JSON Data and r Objects." *arXiv:1403.2805 [Stat.CO]*. https://arxiv.org/abs/1403.2805.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rao, Rohan. 2020. "Formula 1 World Championship (1950-2020) Dataset." https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020.

Reddy, G Dinesh, Saksham Rastogi, Harshit Mahajan, Durvesh Chavan, Manit Gandhi, and Clifton Fernandes. 2022. "Operation Research in Tyre Management and Strategy Building of a Formula One Car." *International Research Journal of Modernization in Engineering Technology and Science* 04 (11): 901–7. https://doi.org/10.56726/IRJMETS31293.

Tan, Shi Ying, and Do Young Pyun. 2018. "The Effectiveness of Sponsorship of the F1 Singapore Grand Prix: Recall and Recognition." *International Journal of Asian Business and Information Management* 9 (1): In Press. https://doi.org/10.4018/IJABIM.2018010101.

Tremlett, A. J., and D. J. N. Limebeer. 2016. "Optimal Tyre Usage for a Formula One Car." *Vehicle System Dynamics* 54 (10): 1448–73. https://doi.org/10.1080/00423114.2016.1213861.

Wesselbaum, Dennis, and P Dorian Owen. 2021. "The Value of Pole Position in Formula 1 History." *Australian Economic Review* 54 (1): 164–73.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023a. *Httr: Tools for Working with URLs and HTTP*. https://httr.r-lib.org/.

———. 2023b. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://stringr.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan,

Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jennifer Bryan, Malcolm Barrett, and Andy Teucher. 2024. *Usethis: Automate Package and Project Setup*. https://usethis.r-lib.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Jim Hester, Winston Chang, and Jennifer Bryan. 2022. *Devtools: Tools to Make Developing r Packages Easier*. https://devtools.r-lib.org/.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.