An Analysis of the 2023 Formula 1 Singapore Grand Prix

Danny Satterthwaite, Bern Adu, and Masood Ali Khan

April 23, 2024

Contents

1 Abstract **1**
1.1 Abbreviations 2

2 Introduction **2**

3 Methods **3**
3.1 Data collection and preprocessing 3
3.2 Analytical Approaches 4

4 Exploring the Data **5**
4.1 Practice EDA 5
4.2 Qualifying EDA 7
4.3 Race EDA 8

5 Analysis **9**
5.1 Qualifying Analysis 9
5.2 Race Analysis 10

6 Results and Discussion **14**

7 Conclusion **15**

1 Abstract

In this paper, we explored the various components of a Formula 1 weekend, the practice sessions, qualifying, and the race. Through our exploratory data analysis, we observed the importance of many factors including tire compound, tire life, track status, and number of laps completed when used to predict lap times. Using these factors, as well as others, we used various analytical approaches such as multiple linear regression, logistic regression, and Poisson regression to model race lap time, when to pit during the race, and the frequency

of fastest sector times during the race. We found that tire life and tire compound appeared to have a significant effect in all models. Ultimately, the strength of our models varied, yet some were quite strong in their predictive ability.

1.1 Abbreviations

1. EDA: Exploratory Data Analysis
2. FP1, FP2, FP3: First Practice 1, First Practice 2, First Practice 3

2 Introduction

Formula 1, featuring 24 races, 10 teams, and 20 drivers, alongside stakes involving hundreds of millions of dollars, ranks among the world's most followed sports series, attracting an average of 70 million viewers per race. Many may recognize it from Netflix's "Drive to Survive," which has offered a behind-the-scenes look for the last five years, or perhaps through its most iconic and statistically greatest driver, Sir Lewis Hamilton. With immense wealth and prestige at stake, victory is the ultimate aim for the teams and drivers, who utilize vast numbers of engineers and data scientists to craft the swiftest cars and fine-tune performance. Our objective was to simulate the role of these data scientists during a race weekend, optimizing car speed across three practice sessions prior to qualifying (which determines the race's starting lineup) and the race itself. What can we learn from these practice sessions to inform our strategies for qualifying and the race? And with real-time lap data during the race, how can we adjust strategies and influence the race's outcome (noting that teams must make at least one pit stop to switch to a different tire compound, with the trade-off that harder tires last longer but are slower, and vice versa)?

We begin exploring the data by examining relationships between teams, drivers, lap times, sector times (each track is split into 3 sectors to give a more detailed sense of times), tire choice etc. within the practice sessions, qualifying, and race. From there we further analyzed the practice laps to see what insight they can provide to predict qualifying lap times. Lastly, we analyzed several aspects of the race. We conclude with a discussion and wrap up of our results and findings.

Regarding related studies, the most relevant research we found is a blog that introduces the package we used and details some features, like comparing two drivers [1]. This was invaluable for initiating our code and conducting exploratory data analysis, but our ambitions go further as we aim to narrate the story of a race weekend through our analyses and forecasts. Additionally, George Washington University performed an analysis of Formula 1 using XGBoost, binary classification, and other advanced methods. However, their

approach differed from ours as they had access to data spanning 30 years for all races, allowing them to employ a broader range of techniques [2]. While there is considerable research in this area, much of it is conducted in-house by teams and remains confidential due to financial incentives, making it inaccessible. Nevertheless, F1 itself generates analytics used in strategy articles and during race broadcasts [3], and AWS provides strategic insights and analyses [4]. Our project contributes to this field by exploring the data in depth and predicting tactical decisions like next lap pit stops and fastest sector times, aiming to enhance understanding and strategic planning in Formula 1 racing.

3 Methods

3.1 Data collection and preprocessing

The data is accessed through the fastf1 package, which allows users to access all available Formula 1 data for every race. Once accessed, the data takes the form of Pandas DataFrames. The package grants access to many key variables such as the individual sessions, laps, drivers and allows for the creation of custom visualizations, although we plan to create our own graphs and attempt to replicate those we see on TV.

While there are hundreds of variables in our DataFrames, we will focus on generally across all areas, Driver, Team, Lap Time, Lap Number (for during the race), Stint (how many times pitted), and the variables relating to sector times (each lap is a broken down into sectors so analyzing these may be key, i.e. some cars are good in straight lines but bad in sectors with lots of turns), and qualifying variables like Q1, Q2, Q3 which are the qualifying times (the field is shrunk down after Q1 and Q2). Despite the hundreds of variables we have access to, the package does not disclose fuel data. For context, unlike in practice sessions and qualifying, in which teams have the freedom to use and replenish fuel as desired, allowing for flexible strategy and testing (a lighter car is a faster car, so teams tend to run the bare minimum amount of fuel) during the race the cars cannot refuel. In our analysis we attempt to account for this by including the Lap Number variable as there is clearly a strong correlation between lap time and fuel load. Further, in addition to this sample of the data/ rows, we also have DataFrames of the results for each session which we use to make comparisons based on our predictions and analysis. See below Figures 1,2,3 for sample tables.

In regard to cleaning and preprocessing, the fastF1 package stores all timing variables (LapTime etc) as timeDelta objects which are/were incompatible with many of graphing libraries we used. Thus we converted timing variables to numeric values to store them in seconds. We also removed missing values before

our machine learning approaches.

	DriverNumber	BroadcastName	Abbreviation	DriverId	TeamName	TeamColor	TeamId	FirstName	LastName	FullName	...	CountryCode	Position	ClassifiedPosition	GridPosition	Q1	Q2	Q3	Time	Status	Points
10	10	P GASLY	GAS	gasly	Toro Rosso	469bff	toro_rosso	Pierre	Gasly	Pierre Gasly	...		NaN		NaN	NaT	NaT	NaT	NaT		NaN
11	11	S PEREZ	PER	perez	Racing Point	f596c8	racing_point	Sergio	Perez	Sergio Perez	...		NaN		NaN	NaT	NaT	NaT	NaT		NaN
16	16	C LECLERC	LEC	leclerc	Ferrari	dc0000	ferrari	Charles	Leclerc	Charles Leclerc	...		NaN		NaN	NaT	NaT	NaT	NaT		NaN
18	18	L STROLL	STR	stroll	Racing Point	f596c8	racing_point	Lance	Stroll	Lance Stroll	...		NaN		NaN	NaT	NaT	NaT	NaT		NaN

Figure 1: Practice Sample Data

	Time	Driver	DriverNumber	LapTime	LapNumber	Stint	PitOutTime	PitInTime	Sector1Time	Sector2Time	...	FreshTyre	Team	LapStartTime	LapStartDate	TrackStatus	Position	D
0	0 days 00:20:47.262000	BOT	77	NaT	1.0	1.0	0 days 00:19:02.300000	NaT	NaT	0 days 00:00:39.892000	...	True	Mercedes	0 days 00:19:02.300000	2021-09-10 16:04:02.311	1	NaN	
1	0 days 00:22:08.066000	BOT	77	0 days 00:01:20.804000	2.0	1.0	NaT	NaT	0 days 00:00:26.987000	0 days 00:00:27.093000	...	True	Mercedes	0 days 00:20:47.262000	2021-09-10 16:05:47.273	1	NaN	
2	0 days 00:23:56.231000	BOT	77	0 days 00:01:48.165000	3.0	1.0	NaT	0 days 00:23:50.008000	0 days 00:00:32.720000	0 days 00:00:33.869000	...	True	Mercedes	0 days 00:22:08.066000	2021-09-10 16:07:08.077	1	NaN	
3	0 days 00:30:03.506000	BOT	77	NaT	4.0	2.0	0 days 00:28:20.973000	NaT	NaT	0 days 00:00:37.951000	...	False	Mercedes	0 days 00:23:56.231000	2021-09-10 16:08:56.242	1	NaN	
4	0 days 00:31:24.191000	BOT	77	0 days 00:01:20.685000	5.0	2.0	NaT	NaT	0 days 00:00:27.142000	0 days 00:00:26.915000	...	False	Mercedes	0 days 00:30:03.506000	2021-09-10 16:15:03.517	1	NaN	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
285	0 days 00:27:41.267000	MAZ	9	0 days 00:01:22.897000	5.0	2.0	NaT	NaT	0 days 00:00:27.550000	0 days 00:00:27.742000	...	True	Haas F1 Team	0 days 00:26:18.370000	2021-09-10 16:11:18.381	1	NaN	
286	0 days 00:29:16.811000	MAZ	9	0 days 00:01:35.544000	6.0	2.0	NaT	0 days 00:29:13.034000	0 days 00:00:29.624000	0 days 00:00:31.156000	...	True	Haas F1 Team	0 days 00:27:41.267000	2021-09-10 16:12:41.278	1	NaN	

Figure 2: Qualifying Sample Data

	Time	Driver	DriverNumber	LapTime	LapNumber	Stint	PitOutTime	PitInTime	Sector1Time	Sector2Time	...	FreshTyre	Team	LapStartTime	LapStartDate	TrackStatus	Position	Delete
0	0 days 01:04:16.958000	GAS	10	0 days 00:02:02.171000	1.0	1.0	0 days 01:02:22.702000	0 days 01:00:55.481000	NaT	0 days 00:00:31.018000	...	True	AlphaTauri	0 days 01:02:14.632000	2021-09-12 13:03:16.241	267	19.0	Fals
1	0 days 01:05:45.963000	GAS	10	0 days 00:01:29.005000	2.0	1.0	NaT	NaT	0 days 00:00:28.594000	0 days 00:00:30.877000	...	True	AlphaTauri	0 days 01:04:16.958000	2021-09-12 13:05:18.567	1	18.0	Fals
2	0 days 01:07:35.136000	GAS	10	0 days 00:01:49.173000	3.0	1.0	NaT	0 days 01:07:31.414000	0 days 00:00:28.479000	0 days 00:00:37.644000	...	True	AlphaTauri	0 days 01:05:45.963000	2021-09-12 13:06:47.572	1	18.0	Fals
3	0 days 01:04:05.139000	PER	11	0 days 00:01:50.352000	1.0	1.0	NaT	NaT	NaT	0 days 00:00:31.553000	...	True	Red Bull Racing	0 days 01:02:14.632000	2021-09-12 13:03:16.241	267	8.0	Fals
4	0 days 01:05:32.745000	PER	11	0 days 00:01:27.606000	2.0	1.0	NaT	NaT	0 days 00:00:29.036000	0 days 00:00:29.775000	...	True	Red Bull Racing	0 days 01:04:05.139000	2021-09-12 13:05:06.748	7	7.0	Fals
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	

Figure 3: Race Sample Data

3.2 Analytical Approaches

We employed various statistical models such as multiple regression, binary logistic regression, and Poisson regression to do further analysis on the race data. We first used multiple regression to examine the relationship between lap time, tire life, and laps completed (as a substitute for fuel load). We then used logistic regression to model the probability of whether a driver would pit the following lap or not. As our response is binary, that is, whether a driver pit the following lap or not, we use logistic regression to model our research question. The model incorporates covariates and interactions among variables. The model is:

$$ln(\frac{P(Y=\frac{1}{x})}{1-P(Y=\frac{1}{x})}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

where $ln(\frac{P(Y=\frac{1}{x})}{1-P(Y=\frac{1}{x})})$ is the log odds of the dependent variable, Y is the dichotomous outcome, β_0 is the intercept, and the B_k's are the model coefficients and X_k's are the predictor variables. Also, a Poisson regression was used to model the count of the fastest sector times achieved by a driver as a function of track status, tire life, and other variables. The Poisson distribution has an average expected value of λ. The variance of a Poisson distribution is also λ. Since our

interest here is model the count of the fastest sector times. The model is:

$$log(\hat{Y}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

where β_p reflects the amount of change in the logarithm of the predicted number of events for a unit change in X_p.

4 Exploring the Data

4.1 Practice EDA

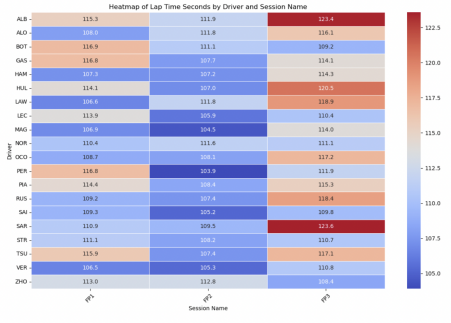

Figure 4: Lap Time By Practice Session

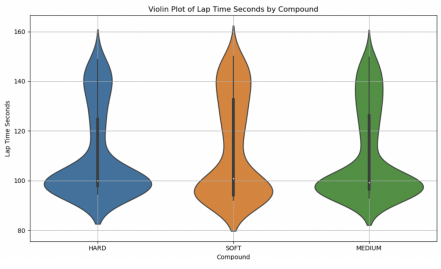

Figure 5: Practice Lap Time by Tire Compound

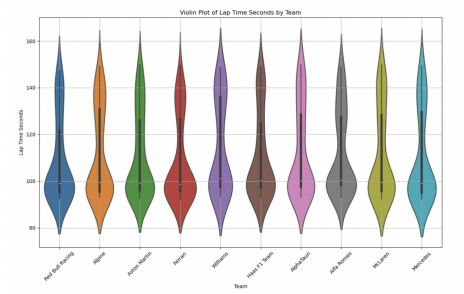

Figure 6: Lap Time by Team in Practice

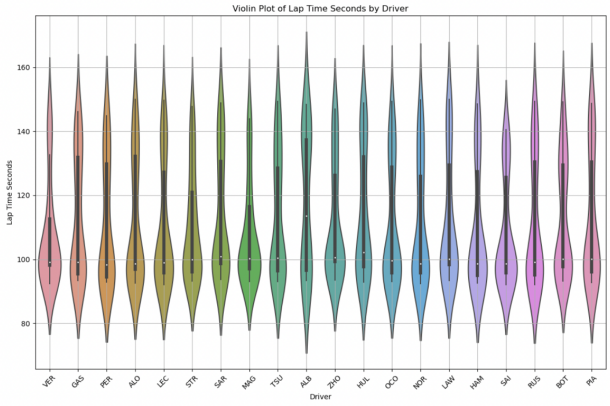

Figure 7: Lap Time by Driver in Practice

First, looking at the distribution of lap time between the three practice session (see Figure 4 above), we see that FP2 had the fastest lap times. This is not surprising as typically FP2 occurs at the same time as qualifying so teams are likely trying to simulate qualifying by setting up and driving the car as quick as possible.

Next, exploring lap time by tire compound for all three sessions, we see (see Figure 5 above) that the soft compound tires are clearly the fastest of the compounds. Further the distribution of lap times for the soft tires is slightly less narrow around the median with the soft tires having the fullest tails suggesting that while they are the quickest, under certain circumstances (like using them for too many laps that the rubber erodes too much), they can also be as slow as medium and hard tires. The distribution of the medium and hard tires appears similar with the median lap time for the medium tires being slightly quicker. This is all useful information that we will consider for predicting race lap times

and race pit stops.

In examining the distribution of lap times by team, we can see (see Figure 6 above) that Ferrari has the quickest median lap time while Williams has the slowest median lap times. That being said, the teams all appear closely matched but we suspect more nuanced differences will emerge in our analysis.

Lastly, looking at the distribution of lap time across all sessions for each driver we can see (See figure 7 above) that these distribution all appear quite similar. This is not particularly surprising as each driver and team are experimenting with different car setups and run plans. However, we see that Hamilton, Norris, and Sainz have the quickest median lap times. This is interesting because they are the three highest finishers in the race.

4.2 Qualifying EDA

Driver	Time in Seconds (Q3)	Difference to Next Driver
Carlos Sainz	90.984	0
George Russell	91.056	0.072
Charles Leclerc	91.063	0.007
Lando Norris	91.27	0.207
Lewis Hamilton	91.485	0.215
Kevin Magnussen	91.575	0.09
Fernando Alonso	91.615	0.04
Esteban Ocon	91.673	0.058
Nico Hulkenberg	91.808	0.135
Liam Lawson	92.268	0.46

Figure 8: Qualifying Result, Top 10

In looking at the top ten qualifiers and the gaps between them, we see (see Figure 8 above) how incredibly close, thousandths of seconds, many of the drivers are to each other. This suggests the drivers are optimizing their car's performance given the track. Further, as we will discuss in our analysis, these incredibly tight margins make attempting to forecast qualifying results quite challenging.

4.3 Race EDA

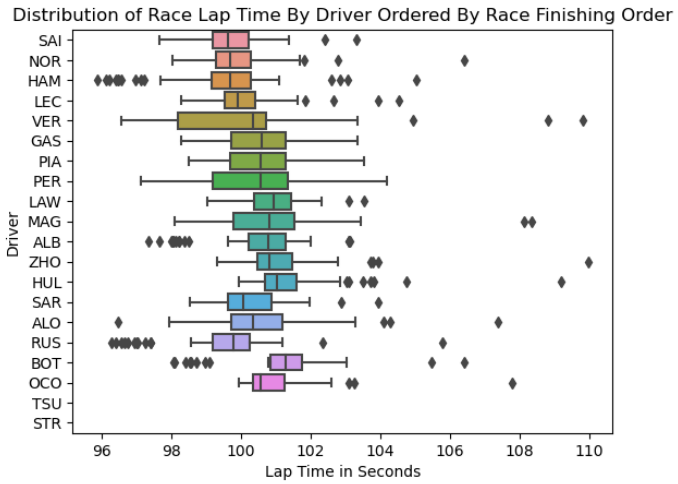

Figure 9: Race Lap Time by Driver

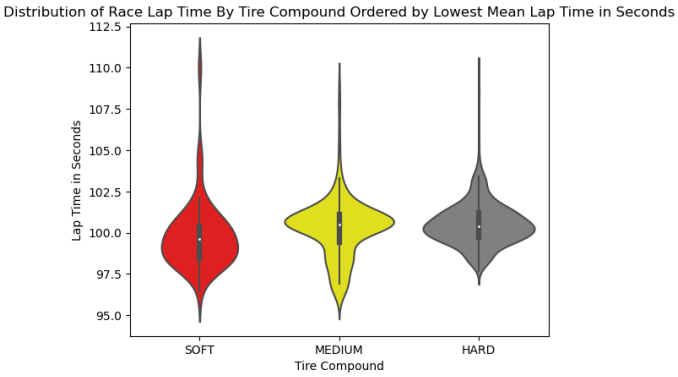

Figure 10: Race Lap Time By Tire Compound

In looking at the distribution of race lap times for each driver, we can see (see Figure 9 above) that the winner, Sainz, also had the quickest median race lap time (after removing the obvious outlier laps). This holds true for the top 5 finishers except for Russell, who crashed in the final laps of the race. We see that for Sainz and Leclerc, the two Ferrari drivers, that their lap time distributions, as we would expect as they are driving the same car, are incredibly similar, with Sainz having the slightly quicker median lap time which makes sense as he finished ahead of his teammate. Stepping back, we can see that while there is a strong correlation between median lap time and finishing position, given the

crashing, mistakes, and poor driving of other drivers, median and average lap time do not completely match/support the race results.

Continuing with the retrospective analysis of the race, we wanted to confirm our understanding and suspicions on tire performance by compound. After removing the laps that were clearly pit stops (and thus much much slower), we see that (see Figure 10 above), as we suspected, the softer the tire compound, the quicker on average the lap is. While the distributions are quite wide, likely due to mistakes or being slowed up by other drivers, the majority of laps on soft tires are less than 99 seconds. For medium tires, the distribution around the mean is much tighter, suggesting less variance in performance but we can also see they appear slower by approximately 1.5 seconds. The hard tires are once again slower on average (graph is arranged left to right by mean) but the distributions between the hard and medium tires have a strong amount of overlap suggesting similar median performance is achieved. But the hard tires appear to have an even tighter distribution, with very few laps being quicker than 99 seconds. This is likely why the mean lap time for hard tires is higher than that of the medium tires (and thus soft tires).

5 Analysis

5.1 Qualifying Analysis

Team	Team Fastest Qualifying Time (Seconds)	Team Best Practice Time (Seconds)	Team Theoretical Fastest Time (Seconds)	Difference Between Qualification Time and Fastest Practice Time (Seconds)	Difference Between Qualification Time and Theoretical Fastest Time (Seconds)
Ferrari	90.984	92.12	91.965	1.136	0.981
Mercedes	91.056	92.355	92.349	1.299	1.293
McLaren	91.27	92.711	92.594	1.441	1.324
Haas	91.575	93.017	92.965	1.442	1.39
Aston Martin	91.615	92.478	92.965	0.863	1.35
Alpine	91.808	93.361	93.361	1.553	1.553
AlphaTauri	92.166	93.285	93.133	1.119	0.967
Red Bull	92.17	92.812	92.659	0.642	0.489
Williams	92.668	94.327	93.768	1.659	1.1
Alpha Romeo	92.809	93.105	93.066	0.296	0.257

Figure 11: Qualifying Analysis. Teams Top Time vs Best Practice Time and Theoretically Fastest Time from Sectors in Practice

In analyzing qualification results (see Figure 11 above), it is important to remember the goal and format of qualification. The goal is to set the fastest lap. However, there are 3 sessions, and the 5 drivers with the slowest fastest laps after the first 10 minutes (session 1) are eliminated, then the 5 drivers with the slowest fastest laps after the next 10 minutes (session 2) are eliminated, and then the fastest lap of the final session (last 10 minutes) takes pole position and starts first. Qualifying is critically important as it sets the order for the start of the race. This can be even more critical for tracks/races that are narrow (like this race) which makes passing hard. Thus, using practice data to predict qualifying lap times/ performance are critical. Thus, as each team has the same car, we look at the single fastest practice lap time for each team and the theoretically

fastest possible lap time from the practice lap sectors. We then compare these to the drivers fastest qualifying laps. First looking at Ferrari, we see that their fastest single lap in any of the practice sessions was 92.12 seconds while their theoretical fastest possible lap based on the sectors comprising all their practice laps was 91.965 seconds. Carlos Sainz, the fastest qualifying Ferrari driver, had a fastest qualifying lap of 90.984 seconds. While these appear surprisingly close, the difference between his qualifying lap and theoretical fastest lap was +0.984 second which would have made him qualify 9'th. Thus continuing, we see, that in general while yes, the theoretical fastest lap per team is within +0.2 and +1.6 seconds of the teams fastest qualifying driver which seems quite good, given the incredibly tight margins of the final qualifying laps/results, this theoretical fastest time is not a particularly useful predictor, aside from giving us a rough ballpark and ordering of fastest teams.

5.2 Race Analysis

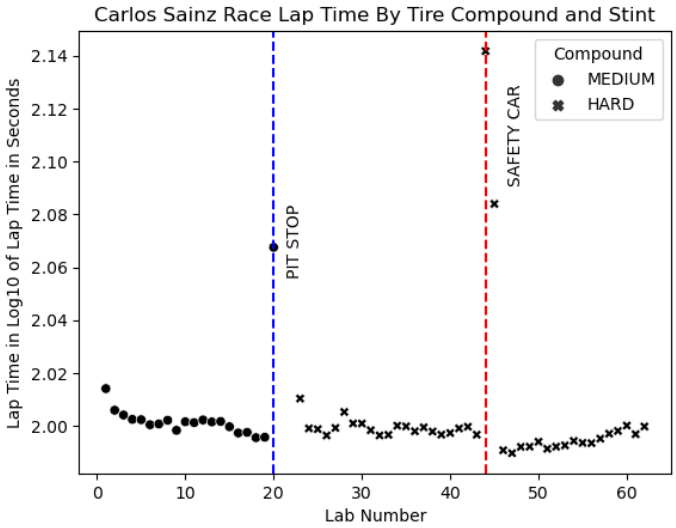

Figure 12: Carlos Sainz Race Analysis Graph

Continuing our analysis, we revisit our exploratory data analysis to explore the potential relationship between tire life and fuel load. In retrospectively looking at Carlos Sainz, the race winner's, lap times across his pit stop, (blue line in Figure 12 above), and accounting for the safety car which slows all cars down, (red line in Figure 12 above) and two tire compounds, we see (see Figure 12 above), that after doing a log transformation of lap time to make the pattern more clear, that for his first stint, the more laps he did on his medium compound tires, the quicker he went. But after changing to a new set of hard tires, the more laps he did on the tires, the slower he went. But, conceptually, the newer

(fresher) the tires a car has, the quicker it will go (due to the rubber compound and tire construction), and as the car does more laps on the set of tires, the tires degrade, loosing grip, and the car gets slower. On the other hand, as the cars cannot refuel during the race (even during the pit stops), the more laps a car has completed, the lighter the car is and the quicker the car will go. However, as our dataset does not provide a variable to track fuel load, we estimate fuel load with the Lap Number variable. Thus, motivated by this and Figure 12 where we can see there is a clear interaction between these effects, and we turn to modeling their relationship. In looking at potential variables of interest for this relationship, we can see that the following variables could have an effect: Stint, Compound, Tire Life, Fresh Tire, and Lap Number. We use these to attempt to predict lap time Lap Time in Seconds where each row will represent a lap. Results: After processing and encoding our variables, we began by using a 5-fold multiple linear regression, iterating over all potential combinations of predictors and examining the models using MSE and R^2. In doing we saw that our best model used the predictors Stint, Lap Number, Hard Compound Tires, which had coefficients 0.668, -0.0682, and 1.474 respectively. Further, we see that only Lap Number (0.003) and Hard Compound Tires (0.006) were statistically significant. This model had an average MSE of 69.63 and R^2 of 0.01009 across all five folds. Given these incredibly poor values, we conclude that the relationship between Stint, Compound, Tire Life, Fresh Tire, and Lap Number, is likely to not be linear.

Next we use logistic regression to look at the likelihood a driver pits in a given lap.

classification report :

	precision	recall	f1-score	support
0	0.98	1.00	0.99	214
1	0.00	0.00	0.00	4
accuracy			0.98	218
macro avg	0.49	0.50	0.50	218
weighted avg	0.96	0.98	0.97	218

Figure 13: Logistic Regression Classification Report

Independent variable	Coefficient (s.e.)	Odds ratio	95% CI	P-value
Intercept	-15.0284 (18.44)			0.42
Lap Number	-0.1244 (0.03)	0.88	0.83 – 0.93	<0.001
Tire Life	0.2025 (0.04)	1.22	1.12 – 1.33	<0.001
Track Status	0.0838 (0.04)	1.09	1.01 – 1.17	0.03
SpeedL1	0.0299 (0.03)	1.03	0.97 – 1.09	0.32
SpeedL2	0.0183 (0.03)	1.02	0.97 – 1.07	0.48
SpeedFL	-0.0382 (0.03)	0.96	0.91 – 1.01	0.16
SpeedST	0.0310 (0.03)	1.03	0.97 – 1.10	0.35
Lap Time (seconds)	-0.0228 (0.10)	0.98	0.24 – 0.51	0.83

Abbreviations: CI-confidence interval; s.e.- standard error. A 2-sided p-value <0.05 indicate statistically significant*Overall p-value<0.001 was based on a likelihood ratio test.R-squared of 18.77% was obtained from the model.

Figure 14: Multiple Logistic Regression Results Predicting Next Lap Pit Stop.

After exploring the data and discovering various trends and patterns. One of the few questions that piqued our interest was predicting the next lap pit stop. We conducted a multiple logistic regression to assess the association between pit stop and the other predictors. See Figures 13 and 14 above. Figure 14 shows the results of the logistic regression analysis performed including the coefficients, odds ratio, and confidence intervals. The results show that lap number, ire life and track status were significantly associated with lap pit stop.

The odds ratio associated with lap number is 0.88, this indicates that with each additional lap, the odds of a pit stop in the next lap decrease by about 11%. This suggests a decreasing likelihood of pit stops as the race progresses. The odds ratio of 1.22 suggests that for each additional lap that a set of tires is used, the odds of a pit stop in the next lap increase by about 22.4%. This aligns with expectations that older tires are more likely to be changed. For track status, the odds ratio of 1.09 suggests that with each increase in the track status number (potentially reflecting deteriorating conditions or other variables), the likelihood of a pit stop in the upcoming lap rises by approximately 9%. The other predictors, SpeedL1, SpeedL2, SpeedFL, SpeedST, and Lap Time is not statistically significant in this model. The likelihood ratio test for model yielded a significant result($p < 0.05$) which indicates that there is a significant overall effect on predicting the next lap pit stop.

Next we use Poisson regression to model the count of the fastest sector times achieved by a driver.

Independent variable	Coefficient (s.e.)	Exp(coef)	95% CI	P-value
Intercept	3.5491 (1.38)			0.01
Tire Life	0.0341 (0.01)	1.03	1.02 – 1.05	<0.001
Track Status	0.3381 (0.07)	1.40	1.22 – 1.61	<0.001
SpeedL1	0.0002 (0.01)	1.00	0.99 – 1.00	0.92
SpeedL2	0.0002 (0.01)	1.00	0.99 – 1.00	0.92
Position	0.0022 (0.01)	1.00	0.99 – 1.01	0.65

Abbreviations: CI-confidence interval; s.e.- standard error. A 2-sided p-value <0.05 indicate statistically significant. R-squared of 99.95% was obtained from the model.

Figure 15: Poisson Regression Results Predicting the Fastest Sector Time

We conducted another analysis to model the count of fastest sector times achieved by a driver as a function of track status, tire life, and other variables using Poisson regression. To identify the fastest sector times for each session, the data was aggregated to count how many times each driver achieved the fastest sector time.

We performed a Poisson regression using the aggregated count of fastest times across all sectors. The predictors in the model include track status, tire life, speedl1, speedl2 and position. Figure 15 shows the results of the Poisson regression analysis performed including the coefficient, log count, and confidence intervals. The analysis indicates that both track status($p < 0.05$) and tire life($p < 0.05$) significantly affect the ability of a driver to achieve the fastest sector times, with improvements in both predictors leading to higher counts of fastest times. For a 1 unit increase in track status, the fastest sector times are expected to increase by a factor of 1.40. The coefficient (0.0339) indicates that longer tire life is associated with a higher log count of fastest times. For a unit increase in tire life, fastest sector times are expected to increase by a factor of 1.034.

This analysis indicates that both track status and tire life significantly affect the ability of a driver to achieve the fastest sector times, with improvements in both predictors leading to higher counts of fastest times. This model can help in understanding and predicting performance based on these variables in future races.

6 Results and Discussion

In discussing our findings and results, beginning with our EDA, for practice, we observed very similar distributions amongst teams and drivers. However, interestingly, the three drivers who had the quickest median times across all practice sessions, Hamilton, Norris and Sainz, finished the race the best - third, second, and first respectively. We also saw very close qualifying times amongst the fastest drivers which suggests that the drivers were both maximizing their cars and the speed the track can be driven; but their differences are likely down to car performance. From the race EDA, we observed that Sainz was not only fastest in qualifying, he also won the race. Further, a majority of the top ten finishers in the race were top ten qualifiers, but the exceptions either crashed (like Russel) or are known to drive slow cars on average like Hulkenberg.

Furthermore, as we saw in our machine learning approaches, tire life and tire compound were important predictors. Additionally, from the EDA of tire compound, the softer the tire, the quicker the average race lap time. This is to be expected given that a softer tire should be quicker than a firmer tire, albeit with a shorter lifetime. In exploring the relationship between practice lap times and qualifying lap times, we compared theoretical fastest lap times amongst teams for all practice sessions and saw that during qualifying all teams were quicker. This is likely to be due to unknown factors not in our dataset, such as engine settings and power modes. In our logistic regression model where we looked at whether a driver pits or not, we concluded tire life and track status were significant in predicting whether a driver pitted or not. This makes sense given that as you drive, the structure of the tire erodes, loosing grip, decreasing lap time, triggering teams to pit. Furthermore, as track status likely denotes whether there is a safety car (and thus the cars slow down and loose their advantages) and the net loss from pitting falls from roughly 30 seconds to 15 seconds. Thus, it is no surprise track status is significant as we would expect teams to prefer, if the timing works out, to pit during a safety car so they loose less time. Furthermore in doing a Poisson regression to model fastest sector time frequency, not only did our model account for nearly all of the variability in fastest sector frequency, but we also observed that, similarly, tire life, lap number, and track status, were influential. The addition of lap number is interesting as we know the cars are lighter, and thus faster towards the end of the race, we hypothesize that faster sector times occur during later laps, suggesting the importance of the lap number variable. From our EDA we suspected this to be the case so we used multiple linear regression to model the impacts of tire life, tire compound, and lap number on lap time. However, after exploring and adjusting the model, our best model only accounted for a very small amount of variability in lap time hence it would be better to look at this using another form of predictors or perhaps a different type of model. Overall, while it was interesting to focus on just one race in particular, we suspect that using multiples races, would improve our models predictive abilities by

accounting for average pace and consistency of the teams and drivers.

7 Conclusion

The study aimed to explore the various components of Formula one weekend using the qualifying, practice and race data. From the result of the analysis, our findings suggest that tire life, track status and tire compound significantly influence the predictive ability of a race whether looking at pit stop, fastest sector times or lap time. This generally makes sense as these variables are major factors in the racing sector.

References

[1]: Jasper, J. (2022, March 31). How to analyze Formula 1 telemetry in 2022: A Python tutorial. Method. Retrieved from https://medium.com/towards-formula-1-analysis/how-to-analyze-formula-1-telemetry-in-2022-a-python-tutorial-309ced4b8992

[2]: Staub, S. (n.d.). Formula One race prediction model. FORMULA ONE RACE PREDICTION MODEL. Retrieved from https://datasci.columbian.gwu.edu/sites/g/files/zaxdzs4746/files/2023-02/f1-final-presentation.pdf

[3]: Medland, C. (2022, October 23). Strategy guide: What are the possible race strategies for the 2022 United States Grand Prix? Formula 1. Retrieved from https://www.formula1.com/en/latest/article.strategy-guide-what-are-the-possible-race-strategies-for-the-2022-us-grand.5tiDhCCh8nay336LaMic9X.html

[4]: F1 Insights powered by AWS — Alternative Strategy. (n.d.). Amazon Web Services. Retrieved from https://aws.amazon.com/sports/f1/