Driver Performance Prediction

Jenny Dang
30153821
jenny.dang@ucalgary.ca

Azariah Francisco
30085863
azariah.francisco@ucalgary.ca

Mary Mo
10131867
ga.mo@ucalgary.ca

Tiffany Okura
30163305
nana.okura@ucalgary.ca

Diane Doan
30052326
diane.doan1@ucalgary.ca

Abstract –Formula 1 (F1 or Formula One), a competitive motor racing series, is a data-driven sport that generates a large amount of data during races, including car telemetry, performance metrics, driver details, and race results. This study aims to build machine learning models to evaluate the impact of certain variables on driver performance prediction such as qualifying position, starting grid position, lap times, pit stops, historical performance (points) and constructors. Data was collected from an API and analyzed using data preprocessing, Exploratory Data Analysis (EDA) and a machine learning model. The machine learning model used in this study was linear regression. Our model was able to achieve an R^2 value was 1.0, and an RMSE value of $2.2x10^{-14}$ was obtained.

Keywords–exploratory data analysis, machine learning, linear regression, R^2, RMSE

I. PREAMBLE

This project was done by five students: Jenny Dang, Mary Mo, Diane Doan, Azariah Franciso and Tiffany Okura. Everyone contributed to the project, each completing about 20% of the work. Each member equally worked on the exploratory data analysis and final report. While Jenny worked on the experiment setup, Mary worked on the machine learning model.

The public repository containing our notebook can be found here: https://github.com/jennydang61/Driver-Performance-Prediction

I, Jenny Dang, hereby sign this declaration on December 16th, 2024

I, Mary Mo, hereby sign this declaration on December 16th, 2024

I, Azariah Francisco, hereby sign this declaration on December 16th, 2024

I, Tiffany Okura, hereby sign this declaration on December 16th, 2024

I, Diane Doan, hereby sign this declaration on December 16th, 2024

II. INTRODUCTION

Formula 1 is an international auto-racing sport that features high-speed cars competing on various tracks around the world. It is recognized by the Fédération Internationale de l'Automobile (FIA), the sport's governing body, as the most prestigious motor racing competition in the world. The series consists of a set of races, known as Grands Prix, which takes place on purpose-built circuits and in some cases closed city streets [2]. Predicting driver performance can be challenging due to many factors like driver's skill, car performance, and race conditions. In this study, we aim to develop a predictive model for driver performance using multiple regression models. Predicting driver performance in Formula 1 is crucial, as it helps teams plan their race strategies, such as when to pit or change tires, which can make a difference in whether they win or lose. Furthermore, analyzing driver data helps improve safety by determining risky situations before

an accident happens. Understanding each driver's strengths and weaknesses also allows for more effective training and development.

This study uses data on Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops and championships. The dataset covers the years 1950 to the latest 2024 season and was obtained from Kaggle, a data science competition platform and online community for data scientists and machine learning [1]. This dataset is a useful resource for studying the history and trends of Formula 1.

Formula 1 is a popular topic that many people do for data analyses. Certain projects utilize data on races that date back to 1950 up to 2024, similar to this dataset that Vopani provided in Kaggle. Kaggle's notebook, "F1 Exploratory Analysis" by Johnny Mcclorey [5] examines various aspects of Formula 1 racing including driver performance, team statistics, and race outcomes. With this data, Mcclorey was able to extrapolate which driver and constructor can be considered to be the greatest of all time. Another project that can be seen on Kaggle is "Formula 1 Pit Stops Analysis" by Kevin Kwan [6], which focuses on analyzing the impact of pit stops on race outcomes in Formula 1 racing. Pit stops influence the overall performance and standings of a driver, making pit stops crucial to race strategies. F1_Race_Win_Predictor, a GitHub project by Bakshi Akshat [7] aims to develop a machine learning model that predicts the winner of Formula 1 races. It is done so by extrapolating historical race data, driver statistics, etc., to predict race outcomes.

Our Formula 1 project aims to enhance existing race predictions by incorporating a broader range of metrics that other models often overlook. By utilizing machine learning, we will predict a driver's final race position based on their qualifying position, starting grid position, and pit stop details, including the duration and number of stops, among other metrics.

While analyzing the data, the main questions that we wanted to answer were:

1. How does the qualifying position influence the final race position?
2. How does the difference between qualifying and final positions affect the driver's final position?
3. How does the average number of pit stops influence the final race position?
4. How does the average lap times influence the driver's final position?
5. How does the average pit stop duration influence the driver's final position?
6. How does accumulating points over a year influence the driver's final position?
7. How do constructor points influence the final position of a driver?
8. Are drivers with a higher number of wins more likely to achieve better final positions?
9. Does starting grid position influence a driver's final race position?

This project proposes a regression model that predicts the driver's performance based on several variables such as qualifying position, number of pit stops, lap times, stop duration, constructor points, driver points, starting grid position and final position as our target variable. The main finding found in our study is that we could predict the driver's performance quite well as we calculated an R^2 of 1 and an RMSE value of 2.2×10^{-14} respectively. The R^2 value demonstrates that our regression model explains all the variance in the data, suggesting a perfect fit, while the extremely low RMSE value tells us there are minimal discrepancies between the predicted and actual values.

III. METHODOLOGY

The methodology outlined in this paper is organized into five components. Setup, Exploratory Data Analysis, Experimentation Factors, Experiment Process and Performance Metrics. These components are detailed as follows:

A. Experiment Setup

We set up our environment in Jupyter Notebook with Pandas and PySpark because of its ease of use for collaboration. Furthermore, we decided to use Python as our programming language to take advantage of its wide range of libraries, including those for machine learning and visualization. As our Big Data Framework, we are using Apache Spark as it's efficient in dealing with Machine Learning applications. Spark's efficiency in processing large datasets makes it efficient to train and deploy models effectively.

The first step in our setup is extracting our dataset from an API. The dataset covers performance metrics, driver

information and race results from 1950 to the latest 2024 season. After examining the dataset, we decided to combine the results, qualifying, lap times, pit stops, driver standings, drivers, races, constructors, constructor_standings and status data. Furthermore, we will be using data from 2011 to 2024 for more accurate results. This approach ensures we have all the necessary information to answer our analysis questions and identify features for machine learning. Before merging our data, we inspected each data file to ensure a smoother process. In each table, we adjusted data types, replaced strings, added and deleted columns, and renamed others. Once we've merged all the data into one data frame, we removed all null values and applied a condition that only displays drivers who had completed a race and those whose starting grid was one or higher.

Following the merging of the data, we conducted an Exploratory Data Analysis to gain a deeper understanding of the dataset. This analysis helped us understand the relationships within our data, enabling us to identify and extract the feature variables that correlate best with our target variable: final position.

B. Exploratory Data Analysis

Our experimental process started with an exploratory data analysis, using tools such as Pandas and SQLite to examine our data. Firstly, we conducted descriptive statistics on our dataset by using the describe() function to summarize characteristics of the data including mean, median, count, minimum value, maximum value, and standard deviation, as well as the 25th, 50th, and 75th percentiles. Performing bivariate analysis was the next step in our analysis [4].

Bivariate Analysis:

The analysis questions we wanted to answer all focused on how one specific factor affected a driver's final race position. Here are our findings:

1. How does the qualifying position influence the final race position?

This analysis was done to determine if the qualifying position influences the final race position. We utilized data on qualifying and final positions, and calculated the average of the qualifying positions to conduct this analysis. The results show a strong positive correlation between the average qualifying position and the final race position (Figure 1). This showcases that drivers who qualify in better positions (lower number) tend to also finish the race in better positions. For example, drivers with an average qualifying position of around 2 generally finish near the top.

Correlation between Average Qualifying Position and Final Position: $r = 0.9890957766899194$

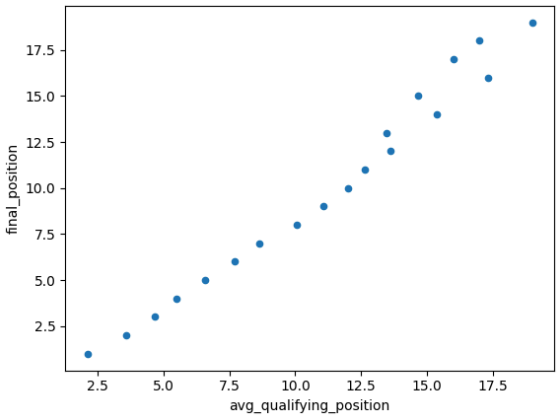

Figure 1: Average Qualifying Position Compared to Final Position

2. How does the difference between qualifying and final positions affect the driver's final position?

This analysis was done to determine if the variance in the qualifying and final positions of a driver would influence the driver's final position in the race. We looked at the difference between the qualifying and final position of a driver at each race and compared it to the driver's final position from the same race. The resulting graph (Figure 2) shows a slight downward trend, indicating that a correlation may exist between the two. The downward trend suggests a negative correlation, meaning a more negative difference indicates a higher final position.

Correlation between Average Difference in Qualifying and Final position and Average Final Position: $r = -0.6579683353593053$

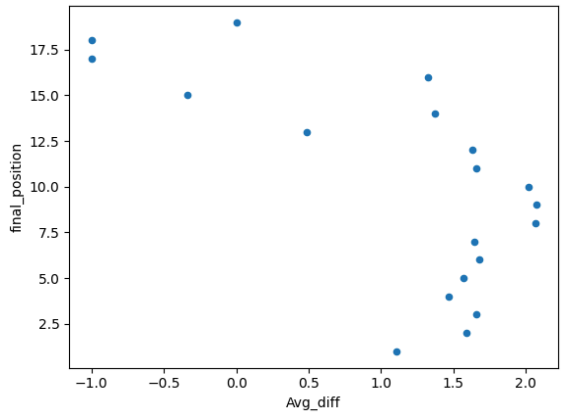

Figure 2: Average Difference in Position Compared to Final Position

3. How does the average number of pit stops influence the final race position?

This analysis was done to determine if the average number of pit stops influences the final race position. We utilized data on the pit stops and final positions, and calculated the average number of pit stops to conduct this analysis. The results demonstrate a moderate positive correlation between the average number of pit stops and the final race position (Figure 3). This demonstrates that drivers who make fewer pit stops tend to finish the race in better positions (lower number). For example, drivers who make an average number of pit stops of approximately 1.02 generally finish near the top.

Correlation between Average Number of Pit Stops and Final Position: $r = 0.3161343620376219$

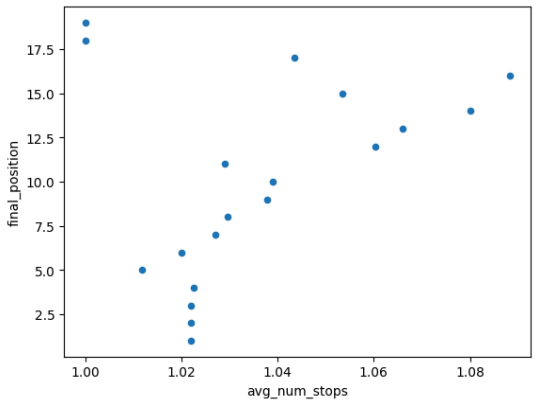

Figure 3: Average Number of Pit Stops Compared to Final Position

4. How does the average lap times influence the driver's final position?

This analysis was done to determine if the average lap time influences the final race position. We utilized data on lap time, and final position, and then calculated the average of the lap time to conduct this analysis. The results show a strong positive correlation between average lap times, and the final position (Figure 4). The figure below showcases that drivers who have a lower average lap time in the race, also have a better final position in the race. For example, drivers that had an average lap time of less than 85000 milliseconds, end the race in the top 3.

Correlation between Average Lap Time (in milliseconds) and Final Position: $r = 0.0012423901311608516$

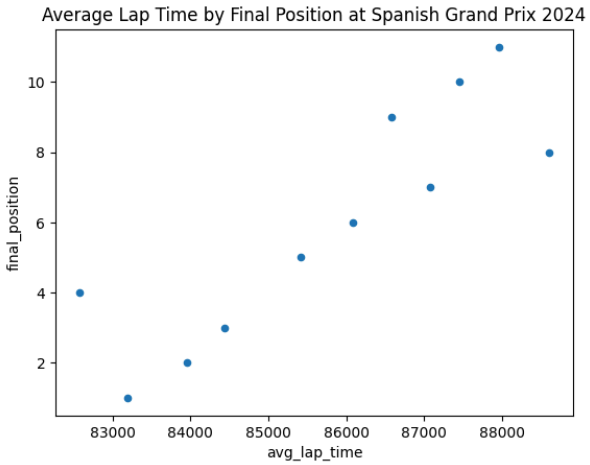

Figure 4: Average Lap Time Compared to Final Position

5. How does the average pit stop duration influence the driver's final position?

This analysis was done to determine if the average pit stop duration influences the final race position. We utilized data on the pit stop duration and final positions, and calculated the average pit stop duration to conduct this analysis. The results demonstrate a strong negative correlation between the average pit stop duration and the final race position (Figure 5). This demonstrates that drivers who spend a longer time at pit stops tend to finish the race in better positions (lower number). For example, drivers who spend an average pit stop

duration of approximately 90000 milliseconds generally finish near the top.

Correlation between Average Pit Stop Duration (in milliseconds) and Final Position: r = -0.813915348786526

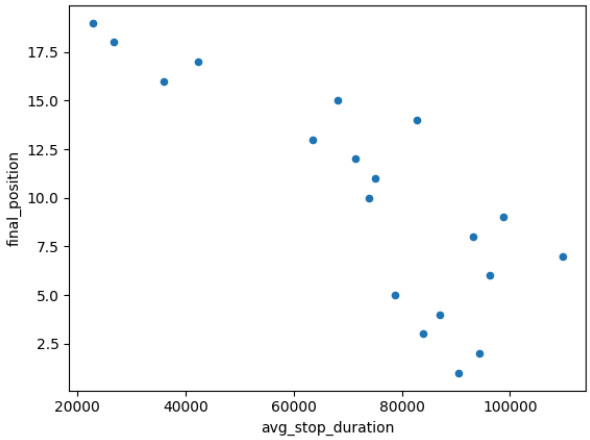

Figure 5: Average Stop Duration Compared to Final Position

6. How does accumulating points over a year influence the driver's final position?

This analysis was done to determine if the points a driver accumulates over a year influence the final race position. We utilized data on driver points according to the year and final positions. The results demonstrate a moderate positive correlation between the driver points and the final race position (Figure 6). This demonstrates that drivers who accumulate more points in the year tend to finish the race in better positions (lower number). For example, drivers who have accumulated points of approximately 20 generally finish near the top.

Correlation between Accumulated Driver Points in a Year and Final Position: r = -0.4311887317881362

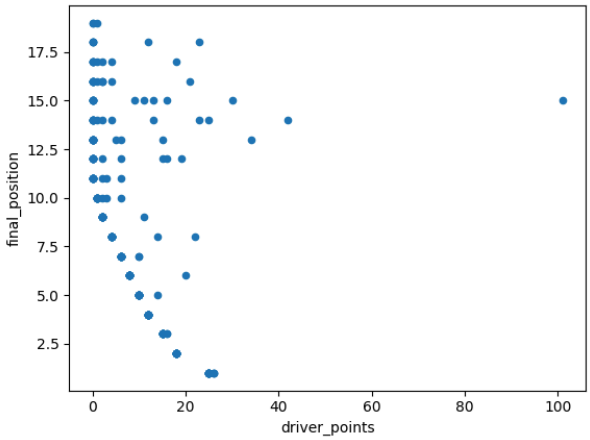

Figure 6: Driver Points Compared to Final Position

7. How do constructor points influence the final position of a driver?

This analysis was done to determine if constructor points influence the final position of the driver. We utilized data on constructor points, and final position, to conduct this analysis. The findings suggest a moderate negative correlation between the constructor points and the final position, implying that higher constructor points generally correlate with better final positions in the standings. As shown in Figure 7, the scatter plot shows a noticeable trend where the final position decreases (better position) as constructor points increase.

Correlation between Constructor Points and Final Position: r = -0.5434195174463154

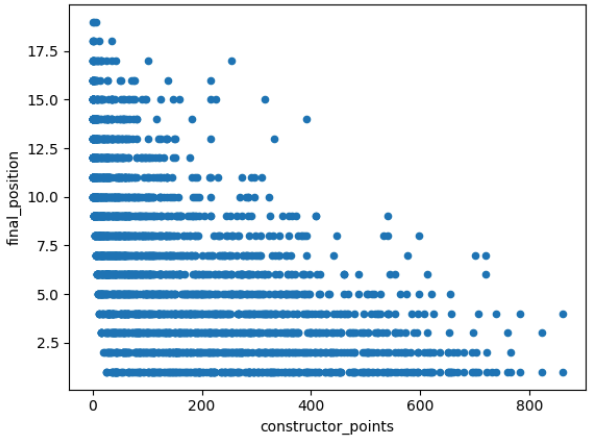

Figure 7: Constructor Points Compared to Final Position

8. Are drivers with a higher number of wins more likely to achieve better final positions?

This analysis was done to determine if a driver with a high number of wins is more likely to achieve a better final position. We analyzed the final position and the count of total wins per driver. We grouped the information based on driver ID/name and final position. From our findings, we concluded that drivers with a higher number of wins are more likely to achieve better final positions. In Figure 8, we see a trend downward suggesting that as the final position value decreases (better position), the driver wins increase. For example, drivers with more than 20 wins, predominantly finish in top positions. Furthermore, this trend showcases a negative correlation.

Correlation between Driver Wins and Final Position: r = -0.41882579729250646

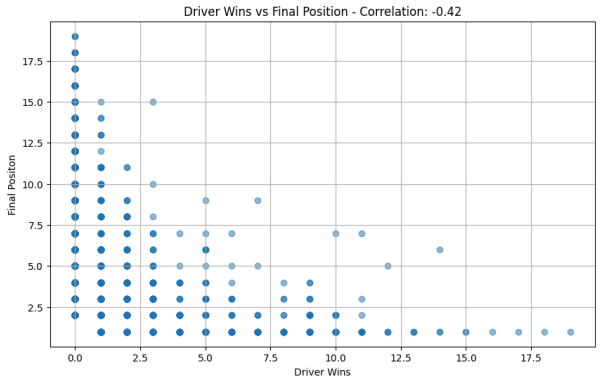

Figure 8: Driver Wins to Final Position

9. Does starting grid position influence a driver's final race position?

This analysis was done to determine the impact of starting grid positions on drivers' final race positions, an important factor in racing. We conducted the analysis using starting grid position and final position data and finding the correlation between the two variables. The calculated correlation value revealed a strong positive correlation between the two. Indicating that drivers starting near the front of the grid tend to finish the race in higher standings (position 1 being the highest). Figure 9 illustrates this relationship with an upward trend, showing that as the final position value increases, the starting grid value also increases, indicating a move toward lower positions.

Correlation between Starting Grid Position and Final Position: r = 0.9864921673137025

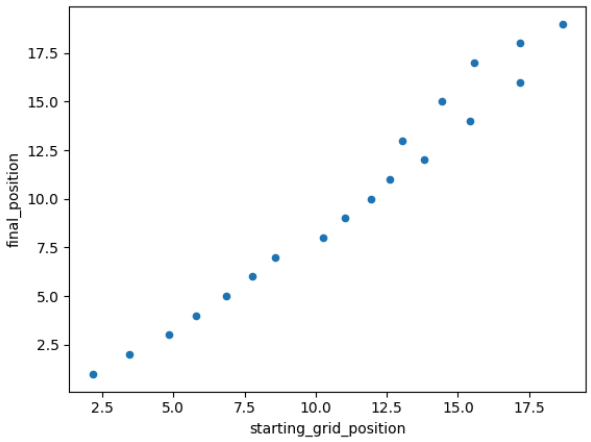

Figure 9: Starting Grid Position Compared to Final Position

Definition of Correlation:

The main decision factor we used to determine if a value was going to be used as a feature variable was correlation. The Pearson correlation coefficient *r*, which measures the linear dependence between two variables, was calculated to determine how strongly certain features relate to the final race position. A positive coefficient represents a positive correlation and a negative coefficient represents a negative correlation between two variables. We considered $|r| \geq 0.5$ to be a strong correlation, $0.3 \leq |r| < 0.5$ to be a moderate correlation, $0.1 \leq |r| < 0.3$ to be a weak correlation, and $r < 0$ to be unrelated [3].

C. Experimentation Factors

Linear Regression was chosen for the machine learning model as it is inherently suited for predicting continuous numerical outcomes, such as the final position of a driver. It is also a good starting point for understanding relationships between features (e.g. qualifying position, lap times) and the target (final position) as it is a simple yet effective model. The coefficients of the regression model provide clear insights into how each feature influences the predicted outcome.

From the above analysis, we decided to exclude some features in the final model such as driver wins data, as it proved rather challenging to extract meaningful data from the raw dataset.
The dataset was split into 70% training and 30% testing data.
Another analysis that was excluded but was obtained is, which drivers are the most consistent in their races. This was excluded because a driver being consistent does not help determine the outcome of the race. Although a driver can be consistently bad or consistently good, which can help predict the outcome, it is not reliable data to use to predict the final positions of each driver.

D. Experiment Process

In preparation for machine learning, the dataset underwent minimal preprocessing, as all selected features were numerical and most processing had already been done before the exploratory data analysis part. This simplicity allowed for direct use in the model without requiring categorical encoding, such as one-hot encoding. Each feature was scaled and vectorized to ensure compatibility with the PySpark MLlib Linear Regression model. This step involved, converting the raw data into a feature vector using Spark's VectorAssembler. The target variable, the driver's final position, was maintained as-is for prediction tasks. This streamlined preprocessing pipeline ensured efficiency while preserving the integrity of the data for analysis.

E. Performance Metrics

This model has a root mean squared error (RMSE) of 2.2×10^{-14} and an R^2 value of 1. The very low RMSE value indicates that the predicted value and the actual values are small, indicating that the model's predictions are accurate. The R^2 value indicates that there is a perfect positive correlation.

TABLE I. P-VALUE

Features	**p-value**
Qualifying Position	0.0000
Difference Between Qualifying & Final Position	0.0000
Number of Pit Stops	0.0000
Driver Points	0.0164
Lap Time	0.0414
Stop duration	0.6773
Constructor Points	0.0024
Starting Grid Position	0.0000

The features with p-values of 0, the qualifying position, the difference between the qualifying & final position, the number of pit stops and the starting grid position all play an important role in our data. We can conclude all significantly affect the outcome because the p-value is 0. From the table, we can also infer that features with a higher p-value such as the stop duration do not play a significant role in our data.

IV. RESULTS

A. Key Findings

Based on the extrapolated data, it can be inferred that: 1. Drivers that start at better positions tend to achieve better final positions. This can be seen from Sebastian Vettel, who started at grid position 1 and ended the race in first. This is indicative of the qualifying position being a critical factor in influencing race outcomes because it can give drivers advantages. 2. Constructors that have more competitive vehicles like Red Bull and McLaren, have a better chance of securing top positions in the race. It can be seen that Sebastian Vettel with Red Bull finished first, while Lewis Hamilton who has McLaren as their constructor finished second. 3. Drivers who can maintain minimal pit stops or can have efficient pit stops tend to have better final positioning. This can be seen with the drivers that finished in the top 5 all doing one pit stop. The amount of times a driver has a pit stop, and the duration of a driver's pit stop all play a significant role in determining whether the driver will secure a higher position or lower position. 4. Drivers with faster lap time averages also tend to finish in higher positions. This can be seen with Sebastian Vettel, having an average lap time of 98109 ms finishing in first place. While external factors can affect the outcome of the race, Formula 1 racing is still highly dependent on drivers performing consistently well, and

the pit stop crew to perform well as well, for the team to secure a better position.

TABLE II. SUMMARY OF RESULTS

Model	R^2	RMSE
Linear Regression	1.0	2.2×10^{-14}

V. CONCLUSION

A. Conclusion

In this study, we were tasked to predict the driver's performance in Formula 1. To predict driver performance, we evaluated various variables and compared them to the driver's final position by calculating the correlation between the two variables. Moderate and strong correlations were chosen as a feature for machine learning in the linear regression model. We took eight features from the dataset to use in the linear regression model. This includes the qualifying position, the difference between the qualifying and final position, the number of pit stops, driver points, lap time (ms), stop duration (ms), constructor points, and starting grid position. We decided not to use driver wins because it was difficult to infer data that would accurately represent the total driver wins.

Our linear regression model has a RMSE value of 2.2×10^{-14} indicating high accuracy in the model's predictions and an R^2 value of 1 indicating a perfect positive correlation. This suggests that linear regression is an effective model for predicting driver performance based on the features we used. In summary, the data suggests variables such as qualifying position, constructors, pit stop efficiency, and lap time are key factors that influence driver performance.

B. Future Work

This project has provided insightful information on various factors that can influence driver performance in Formula 1. However, there are areas in this project that could be improved for future studies. The topic of driver performance in Formula 1 can be developed and explored further by considering other variables and factors that may potentially influence driver performance.

Firstly, the circuit characteristics are a major factor that should be considered in future projects. Circuits with higher altitudes indicate thinner air, which means air resistance is much weaker. This, in turn, affects the acceleration and handling of a Formula 1 driver. A car can accelerate faster with less air resistance, making stopping and handling much harder. Higher altitudes also lead to less oxygen, affecting the driver's performance. The design of the circuit is also important. For instance, tracks that have more high-speed straights tend to favour cars with top speeds, whereas tracks with many corners and turns require more skill from the driver to precisely accelerate and brake.

Secondly, weather is another crucial factor that plays a pivotal role in Formula 1 races and can shift performance outcomes drastically. With higher temperatures, tire degradation occurs more quickly due to hotter roads, causing the rubber to melt quicker. At the same time, colder temperatures can cause the air tire to compress making the tires smaller. Rain causes the circuits to have less traction making it much harder to handle these vehicles, thus creating higher chances of errors or crashes. Rain also impacts the visibility of the road, reduced visibility affects the driver's ability to react and see incoming accidents. Incorporating weather data into prediction models enables us to simulate how different conditions will affect performance for individual drivers and teams.

Addressing these areas in future studies will further deepen and broaden our understanding of driver performance in Formula 1 and contribute towards advancements in race strategies, driver development, and predictive modelling.

REFERENCES

[1] Kaggle. "Formula 1 World Championship (1950 - 2024)." [Online], https://www.kaggle.com/datasets/rohanrao/formula-1-world-championship-1950-2020/data?select=drivers.csv. [Accessed 19 November 2024].
[2] Howson, George. "What Is Formula 1 In Simple Terms?" F1 Chronicle, https://f1chronicle.com/what-is-formula-1/. [Accessed 16 December 2024].
[3] Turney, Shaun. "Pearson Correlation Coefficient (r) | Guide & Examples." Scribbr, 13 May 2022,

https://www.scribbr.com/statistics/pearson-correlation-coefficient/. [Accessed 18 December 2024].
[4] Bivariate Analysis, GeeksforGeeks, 20 May 2024 [Online], https://www.geeksforgeeks.org/bivariate-analysis/. [Accessed 19 December 2024].
[5] Mcclorey, Johnny, “F1 Exploratory Analysis”, Kaggle, 22 April 2022, [Online] https://www.kaggle.com/code/johnnymcclorey/f1-exploratory-analysis.[Accessed 16 December 2024].
[6] Kwan, Kevin. “Formula 1 Pit Stop Analysis” 02 April 2022 [Online], https://www.kaggle.com/code/kevinkwan/formula-1-pit-stops-analysis.[Accessed 19 December 2024].
[7] Bakshi Akshat. “F1_Race_Win_Predictor” GitHub, 6 December 2023, [Online] https://github.com/akshat448/F1_Race_Win_Predictor/blob/main/f1-racepredictor.ipynb [Accessed 16 December 2024].