Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Carthage

Ecole Polytechnique de Tunisie

وزارة التعليم العالي و البحث العلمي

جامعة قرطاج

المدرسة التونسية للتقنيات

Data Analysis Project Report

Formula 1: Data-Driven Evaluation

Realized by : Fraoua Baha
Laagab Anas

Supervised by : Mr.Amor Massaoud

Academic Year : 2024/2025

Rue Elkhawarezmi BP 743 La Marsa 2078
Tel: 71 774 699 Fax: 71 748 843
Site web : www.ept.tn

نهج الخوارزمي ص.ب 743 المرسى 2078
الهاتف : 71 774 611 الفاكس: 71 748 843
موقع الواب: www.ept.tn

Contents

List of Figures

List of Tables

1 Introduction

Formula 1 is one of the most prestigious and data-driven sports in the world, where every fraction of a second matters. Beyond the excitement of races, F1 also generates a wealth of structured data that offers a valuable opportunity for analytical exploration. This project focuses on analyzing a dataset of Formula 1 drivers, aiming to uncover key patterns and build predictive models related to driver performance.

The dataset includes variables such as the number of wins, podium finishes, races participated in, and team affiliations. Through data cleaning, visualization, and statistical analysis, we aim to gain a deeper understanding of what contributes to a driver's success on the track. We also apply classification algorithms to predict performance categories and compare the effectiveness of different models.

This work not only applies core data science techniques—such as preprocessing, exploration, and model evaluation—but also demonstrates how sports analytics can provide meaningful insights. Whether for fans, teams, or analysts, data-driven approaches can enhance understanding, support strategic decisions, and reveal the hidden dynamics of high-performance motorsport.

2 Importing the Libraries

In this section, we lay the groundwork for our Formula 1 driver performance analysis by importing essential Python libraries. These libraries equip us with the necessary tools for data manipulation, statistical testing, visualization, and machine learning. They enable us to explore the dataset effectively, uncover patterns, and build predictive models, ensuring a structured and insightful analytical workflow.

2.1 Libraries

- **Pandas (pd)**: A core library for data manipulation and analysis, particularly useful for handling tabular data using `DataFrame` objects.
- **NumPy (np)**: Enables numerical operations and array processing, serving as a backbone for mathematical computations throughout the project.
- **Matplotlib (plt)** and **Seaborn (sns)**: Visualization libraries used to generate informative plots, such as histograms, distributions, and correlation heatmaps. `Seaborn` also enhances plot aesthetics with themes like `whitegrid`.
- **Scipy.stats**: Provides statistical testing tools, such as `chi2_contingency` for categorical data association and `f_oneway` for ANOVA analysis.
- **Scikit-learn (sklearn)**: A comprehensive machine learning library used for:
  - Preprocessing with `StandardScaler`
  - Building pipelines using `make_pipeline`
  - Training linear models via `SGDClassifier`
  - Evaluating model performance with tools like `metrics`, `confusion_matrix`, and `ConfusionMatrixDisplay`
- **Warnings**: The `warnings` module is used to suppress non-critical alerts, keeping the output clean and focused during execution.

3 Data Exploration

Now we will get to familiarizing ourselves with the Formula 1 dataset and gaining an initial understanding of the variables it contains. By examining the data from both a statistical and visual perspective, we aim to uncover patterns, spot inconsistencies, and identify potential relationships between features such as wins, podiums, races, and teams. These insights will shape the direction of our data cleaning and modeling efforts later in the project.

3.1 Dataset Overview

We load the first 5 rows of the dataset and examine its basic properties :

	Driver	Nationality	Seasons	Championships	Race_Entries	Race_Starts	Pole_Positions	Race_Wins	Podiums	Fastest_Laps	...	Championship Years	Decade	Pole
0	Carlo Abate	Italy	[1962, 1963]	0.0	3.0	0.0	0.0	0.0	0.0	0.0	...	NaN	1960	
1	George Abecassis	United Kingdom	[1951, 1952]	0.0	2.0	2.0	0.0	0.0	0.0	0.0	...	NaN	1950	
2	Kenny Acheson	United Kingdom	[1983, 1985]	0.0	10.0	3.0	0.0	0.0	0.0	0.0	...	NaN	1980	
3	Andrea de Adamich	Italy	[1968, 1970, 1971, 1972, 1973]	0.0	36.0	30.0	0.0	0.0	0.0	0.0	...	NaN	1970	
4	Philippe Adams	Belgium	[1994]	0.0	2.0	2.0	0.0	0.0	0.0	0.0	...	NaN	1990	

5 rows × 22 columns

Figure 1: First 5 rows of our dataset

The dataset comprises performance, career, and demographic attributes, each contributing to the analysis of driver success in Formula 1. Below is a detailed description of the features:

- **Driver** (`object`): Name of the Formula 1 driver.
- **Nationality** (`object`): Driver's nationality.
- **Seasons** (`object`): List of seasons the driver participated in.
- **Championships** (`float64`): Number of World Drivers' Championships won.
- **Race_Entries** (`float64`): Total number of race entries.
- **Race_Starts** (`float64`): Total number of race starts.
- **Pole_Positions** (`float64`): Total number of pole positions achieved.
- **Race_Wins** (`float64`): Total number of race wins.
- **Podiums** (`float64`): Total number of podium finishes (top 3).
- **Fastest_Laps** (`float64`): Total number of fastest laps recorded.
- **Points** (`float64`): Total points accumulated over the driver's career.
- **Active** (`bool`): Whether the driver is currently active (True = Yes, False = No).
- **Championship Years** (`object`): List of years in which the driver won a championship.
- **Decade** (`int64`): The decade in which the driver debuted (e.g., 1990, 2000).

- **Pole_Rate** (`float64`): Ratio of pole positions to race entries.
- **Start_Rate** (`float64`): Ratio of race starts to race entries.
- **Win_Rate** (`float64`): Ratio of wins to race starts.
- **Podium_Rate** (`float64`): Ratio of podium finishes to race starts.
- **FastLap_Rate** (`float64`): Ratio of fastest laps to race starts.
- **Points_Per_Entry** (`float64`): Average points scored per race entry.
- **Years_Active** (`int64`): Number of years the driver was active in Formula 1.
- **Champion** (`bool`): Whether the driver has ever won a championship (True = Yes, False = No).

3.2 General Dataset Information

The dataset contains **868 rows** and **22 columns**, each representing different performance, demographic, or categorical aspects of Formula 1 drivers.

3.3 Data Types and Missing Values

Each column in the dataset has a specific data type: numerical, categorical, or boolean. Understanding these types is essential for preprocessing and modeling. Additionally, we visualize the number of missing values per feature to assess data quality.

Figure 2: Missing Values per Feature

All features in the dataset have clearly defined data types, with a mix of numerical and categorical variables. Importantly, the dataset contains **no missing values**, which simplifies the analysis pipeline. This is particularly beneficial, as the focus of this study lies in the interpretation and visualization of driver performance rather than preprocessing or data cleaning.

3.4 Summary Statistics

The following table provides descriptive statistics for the numerical features in the dataset, including mean, standard deviation, minimum and maximum values, and quartiles. These statistics offer an initial understanding of the central tendencies and distributions within the dataset.

Table 1: Summary Statistics of Raw Performance Metrics

	Championships	Race Entries	Race Starts	Pole Positions	Race Wins	Podiums	Fastest Laps	Points
count	868	868	868	868	868	868	868	868
mean	0.08	29.92	27.69	1.24	1.25	3.76	1.26	55.85
std	0.52	53.78	52.88	6.35	6.49	14.43	5.41	265.97
min	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	2.00	1.00	0.00	0.00	0.00	0.00	0.00
50%	0.00	7.00	5.00	0.00	0.00	0.00	0.00	0.00
75%	0.00	29.25	26.00	0.00	0.00	0.00	0.00	8.00
max	7.00	359.00	356.00	103.00	103.00	191.00	77.00	4415.50

Table 2: Summary Statistics of Derived Rates and Other Features

	Pole Rate	Start Rate	Win Rate	Podium Rate	FastLap Rate	Points Per Entry	Years Active	Championships Won
count	868	868	868	868	868	868	868	868
mean	0.011	0.78	0.011	0.041	0.012	0.48	3.66	0.08
std	0.047	0.32	0.044	0.11	0.044	1.21	3.50	0.52
min	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
25%	0.00	0.67	0.00	0.00	0.00	0.00	1.00	0.00
50%	0.00	0.96	0.00	0.00	0.00	0.00	2.00	0.00
75%	0.00	1.00	0.00	0.00	0.00	0.38	5.00	0.00
max	0.56	1.00	0.46	1.00	0.50	14.20	19.00	7.00

3.5 Nationality-Based Analysis

In this section, we explore how drivers' nationalities relate to their participation and success in Formula 1.

3.5.1 Distribution of Drivers by Nationality

Figure 3 shows the distribution of F1 drivers based on their nationality. The dataset contains a diverse range of countries, with a notable concentration of drivers from a few leading nations.

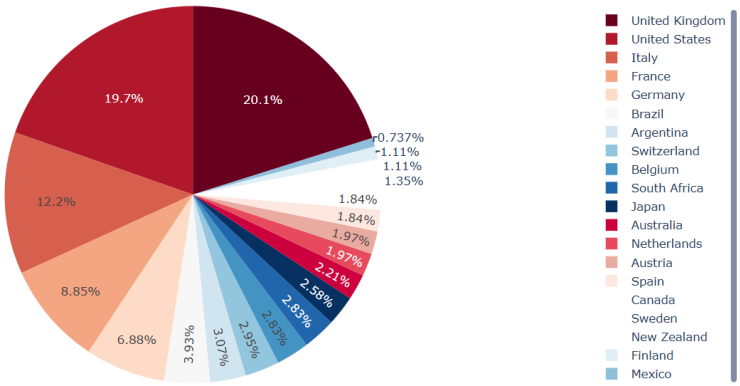

Figure 3: Distribution of Drivers by Nationality

3.5.2 Total Championships by Nationality (Top 10)

Figure 4 displays the total number of championships won by drivers from the top 10 nationalities. This highlights which countries have produced the most successful drivers in the history of Formula 1.

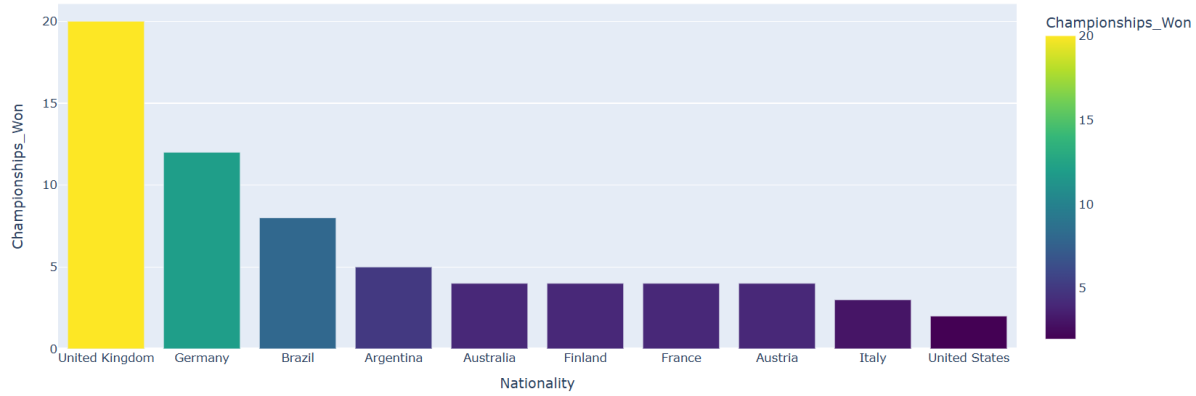

Figure 4: Total Championships by Nationality (Top 10)

3.6 Performance Metrics

This section analyzes key performance indicators that reflect driver skill and success across races.

3.6.1 Podiums vs. Race Wins (Colored by Champion Status)

The scatter plot in Figure 5 displays the relationship between total podiums and race wins. Drivers are color-coded based on whether they have won a championship, revealing clusters of highly decorated competitors.

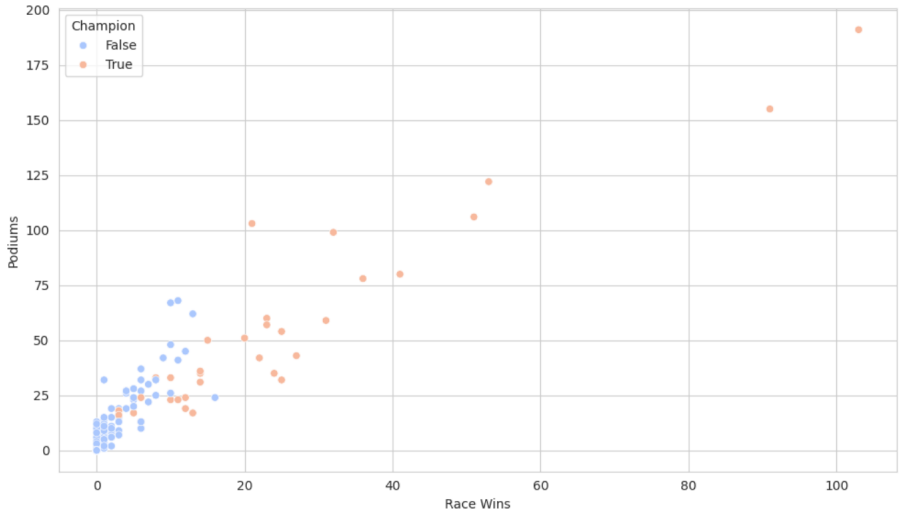

Figure 5: Podiums vs. Race Wins, Colored by Champion Status

3.6.2 5.3.2 Win Rate by Driver Tier

In Figure 6, drivers are segmented into performance tiers based on their overall stats. The boxplot visualizes the distribution of win rates across these tiers, showing how top-tier drivers significantly outperform others.

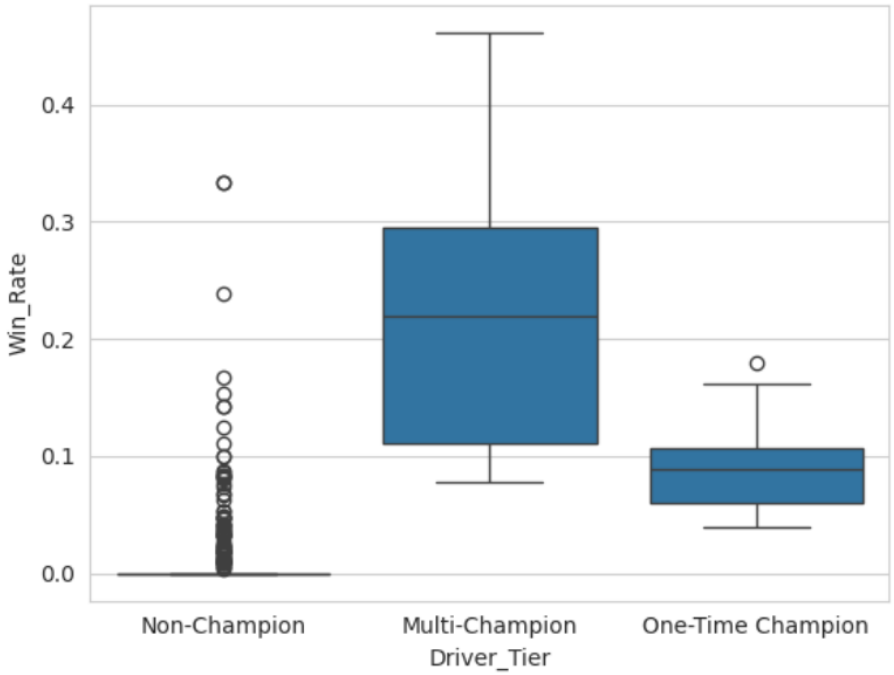

Figure 6: Win Rate by Driver Tier

3.6.3 5.3.3 Correlation Between Key Performance Metrics

Figure 7 presents a heatmap of the correlation matrix among core performance rates: Pole Rate, Win Rate, and Fastest Lap Rate. This visualization helps reveal how closely these indicators are related.

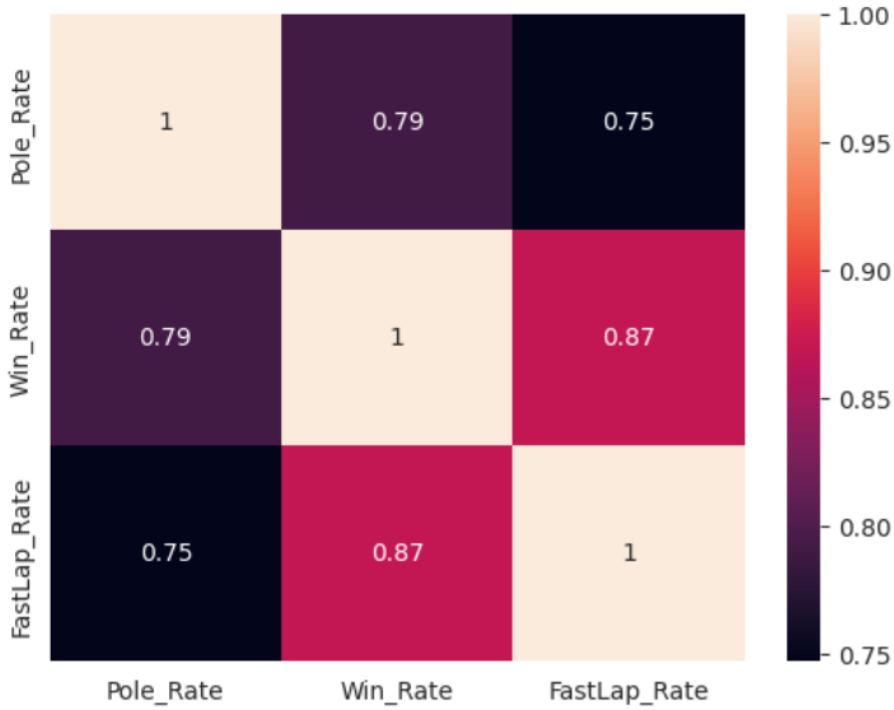

Figure 7: Correlation Heatmap of Pole Rate, Win Rate, and FastLap Rate

3.7 Outlier Detection

This section explores outliers in performance, highlighting drivers who significantly deviate from expected trends.

3.7.1 5.4.1 Points vs. Race Entries (Colored by Champion Status)

Figure 8 illustrates the relationship between total race entries and points scored. By color-coding drivers based on their championship status, we can identify overperformers (high points with fewer entries) and underperformers.

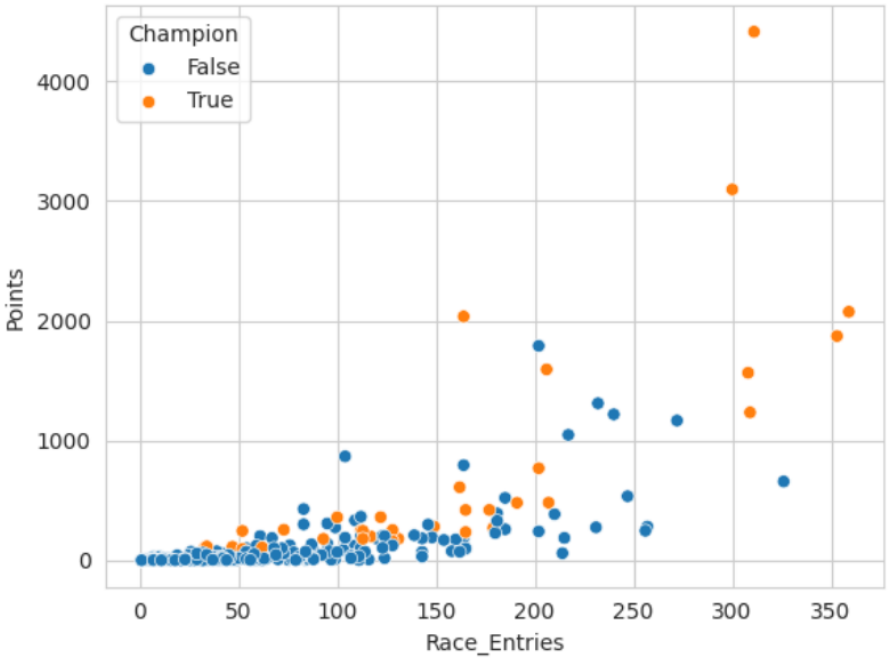

Figure 8: Points vs. Race Entries, Colored by Champion Status

The "Points vs. Race Entries" plot shows most drivers clustered at low entries and low points with no championships. Some outliers stand out: champions with few races and points, others with many entries but low points, and a few high-performing champions with many races and high points.

3.8 Highlighting Champion Trends and Exceptional Performers

3.8.1 Top 10 Drivers by Championships

Top 10 Drivers with the most Championships. These drivers represent the most dominant figures in Formula 1 history.

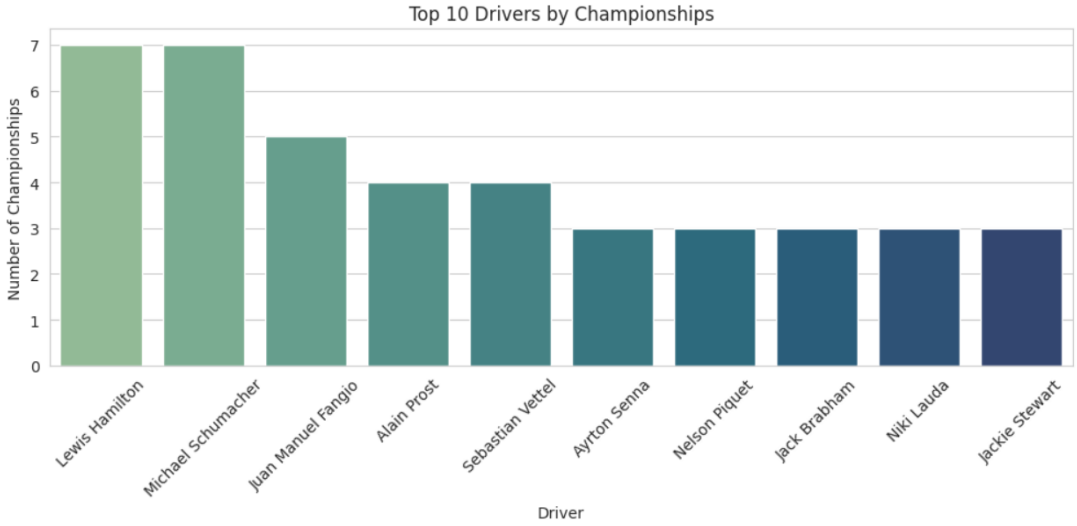

Figure 9: Top 10 Drivers by Championships.

3.8.2 Top 10 Drivers by Points per Entry

Top 10 Drivers with the highest Points per Race Entry among those with at least 20 entries. This metric highlights efficiency and consistent performance

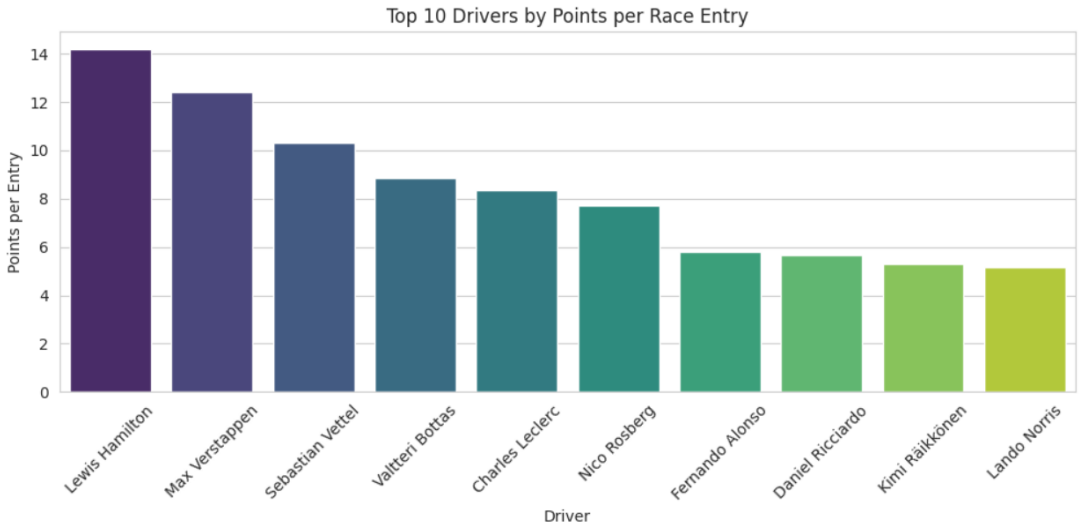

Figure 10: Top 10 Drivers by points per Race Entry.

3.9 Driver Performance Analysis

This section presents key visualizations comparing driver performance across different factors.

3.9.1 Points Per Entry by Decade

This box plot highlights how driver performance, in terms of points per race entry, has evolved over different decades. It reveals how scoring systems and competitive balance have changed across eras.

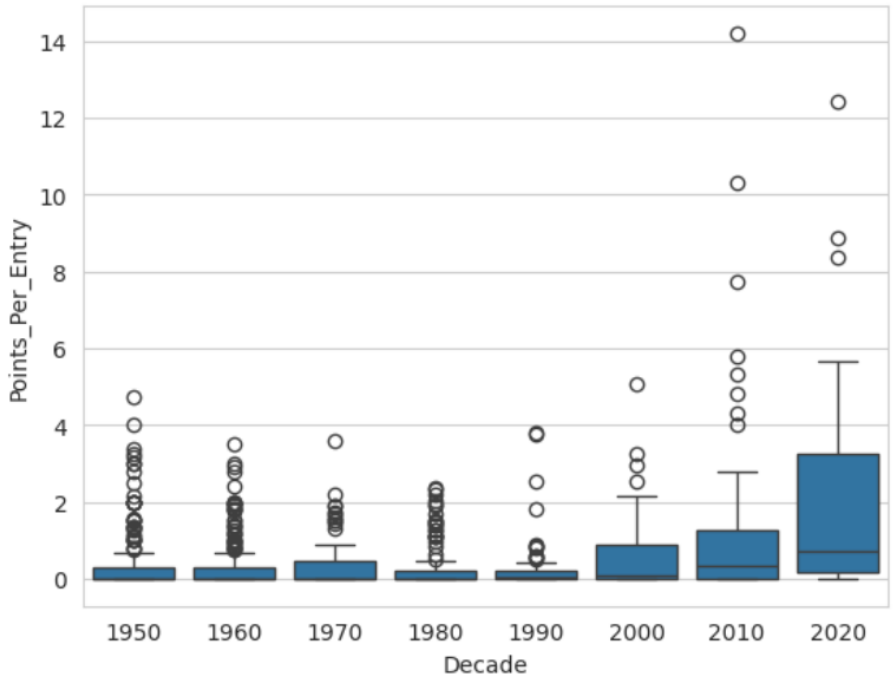

Figure 11: Points Per Entry by Decade

3.9.2 Race Starts vs. Years Active

This scatter plot shows the correlation between the number of years a driver was active and their total race starts. It highlights how career longevity influences experience, especially among champions.

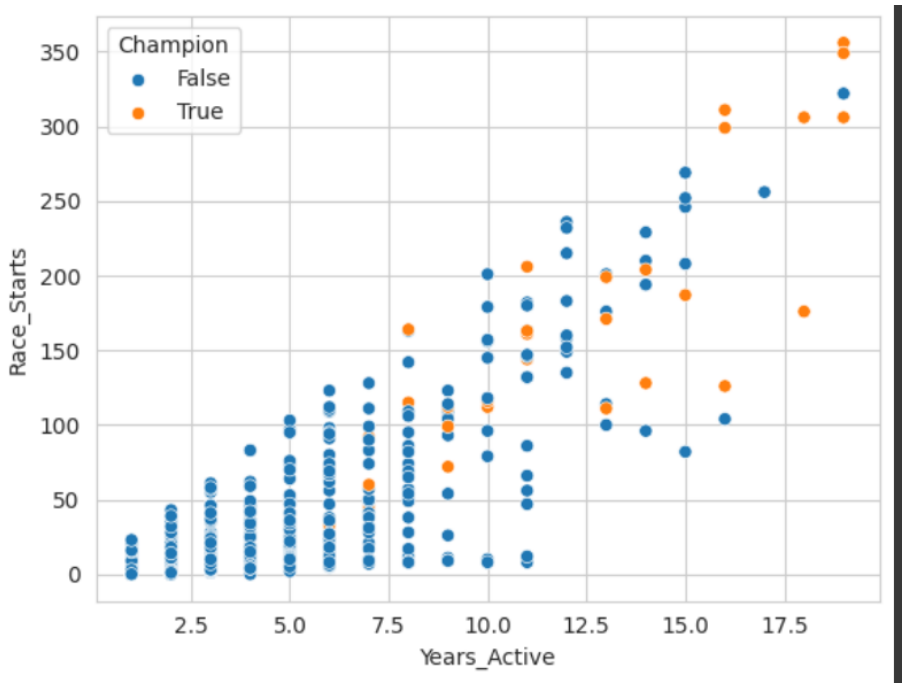

Figure 12: Race Starts vs. Years Active

3.10 Performance Comparison: Champions vs. Non-Champions

This bar chart highlights the average performance metrics for champions and non-champions. Champions clearly outperform non-champions across all key indicators, including win rate, podium rate, and points per entry, underlining the level of excellence required for a title.

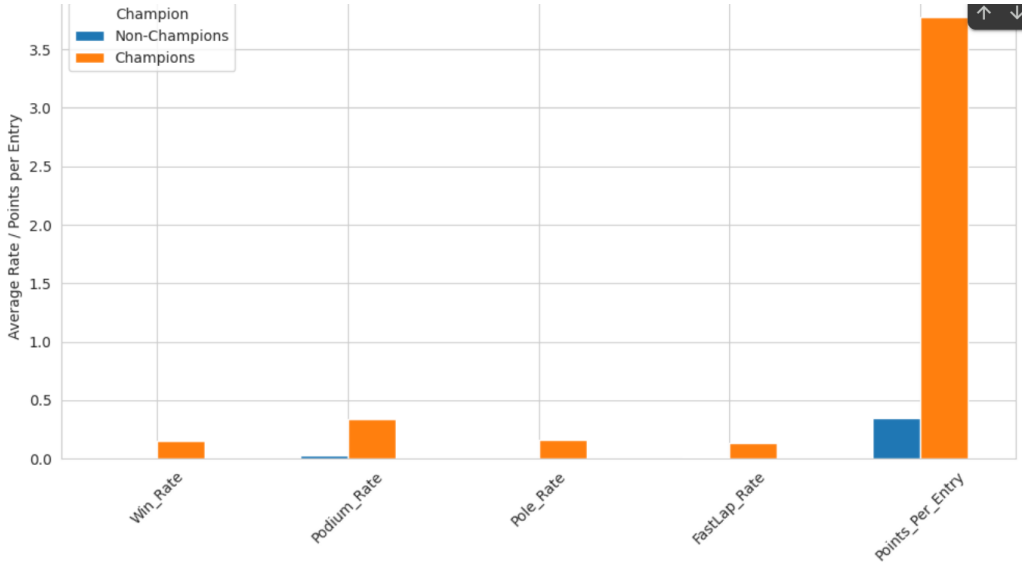

Figure 13: Average Performance Metrics: Champions vs. Non-Champions

4 Data Preprocessing

Data preprocessing is a critical step before modeling. The following operations were performed to ensure data quality and model readiness.

4.1 Feature Engineering

To enhance the dataset and enable deeper analysis, several new features were engineered based on existing metrics. These derived attributes offer a more nuanced understanding of driver performance, consistency, and competitiveness:

- **Success Efficiency Metrics:**
  - **Wins Per Year:** Measures average race wins per active year.
  - **Podiums Per Start:** Calculates podium finishes as a proportion of race starts.
  - **Points Per Year:** Reflects points scored on a per-year basis.
- **Career Span:** Computes the number of years between a driver's first and last season, giving insight into longevity.
- **Championship Dominance:**
  - **Championship Share:** The proportion of seasons where a driver won the championship.
  - **Is Multi-Champion:** A binary indicator for drivers who have won two or more championships.
- **Performance Consistency:**
  - **Win-to-Podium Ratio:** Indicates how often a podium finish resulted in a win.
  - **Pole-to-Win Ratio:** Shows the proportion of wins that came from pole positions.
- **Nationality-Based Metric:**
  - **National Championship Strength:** Represents the average number of championships won by drivers of a given nationality, offering a contextual benchmark.

All newly created features were handled with care to avoid division by zero or missing values by replacing undefined results with zeros. This enriched dataset allows for more meaningful comparisons and modeling in subsequent analysis.

This strategy allowed us to retain valuable information from key features while minimizing the introduction of imputation bias.

Driver	Wins_Per_Year	Podiums_Per_Start	Championship_Share	Is_Multi_Champion
Carlo Abate	0.000000	0.000000	0.000000	False
George Abecassis	0.000000	0.000000	0.000000	False
Kenny Acheson	0.000000	0.000000	0.000000	False
Andrea de Adamich	0.000000	0.000000	0.000000	False
Philippe Adams	0.000000	0.000000	0.000000	False

Table 3: Sample of newly engineered features for selected drivers

4.2 Statistical Significance Testing

To better understand the relationship between various features and the presence of championships, we conducted a series of statistical significance tests. These tests aim to identify which variables are statistically significantly associated with winning championships and thus may be relevant for identifying the key factors that contribute to a champion's success. Depending on the nature of the variables (categorical or numerical), different tests were applied.

For numerical variables, such as win rates, podium finishes, and points per year, we conducted t-tests to compare the means between champions and non-champions. For categorical variables, such as nationality or multi-champion status, we used chi-squared tests to assess if there is a significant association with the championship status.

The results from these tests provided valuable insights into which features are most strongly related to championship success, informing further exploration and modeling in the analysis.

4.2.1 Chi-Squared Test: Nationality vs. Multi-Champion Status

A Chi-squared test was conducted to examine the association between 'Nationality' (top 10 nationalities by frequency) and '$\text{Is}_{M}ulti_{C}hampion$'$(whetheradriverhas2+championships).Thetestyi squaredstatisticof17.3612, ap-valueof0.0434, and9degreesoffreedom.$

Interpretation:

- The p-value 0.0434 is less than 0.05, indicating that we reject the null hypothesis. Therefore, there is a statistically significant association between `Nationality` and `Is_Multi_Champion`.
- Nationalities like the United Kingdom may show a higher tendency to have multi-champions compared to expected counts, as seen in the contingency table and heatmap (e.g., if the United Kingdom has an observed count of 10 multi-champions but an expected count of 5).
- This suggests that a driver's nationality may influence the likelihood of achieving 2+ championships.

Contingency Table (Observed Frequencies):

Nationality	False (Non-Multi-Champions)	True (Multi-Champions)
Argentina	24	1
Belgium	23	0
Brazil	29	3
France	71	1
Germany	54	2
Italy	98	1
South Africa	23	0
Switzerland	24	0
United Kingdom	160	4
United States	160	0

Table 4: Contingency Table: Nationality vs. Is_Multi_Champion (Observed Frequencies)

Expected Frequencies:

Nationality	False (Non-Multi-Champions)	True (Multi-Champions)
Argentina	24.557522	0.442478
Belgium	22.592920	0.407080
Brazil	31.433628	0.566372
France	70.725664	1.274336
Germany	55.008850	0.991150
Italy	97.247788	1.752212
South Africa	22.592920	0.407080
Switzerland	23.575221	0.424779
United Kingdom	161.097345	2.902655
United States	157.168142	2.831858

Table 5: Expected Frequencies: Nationality vs. Is_Multi_Champion

Updated Contingency Table Heatmap:

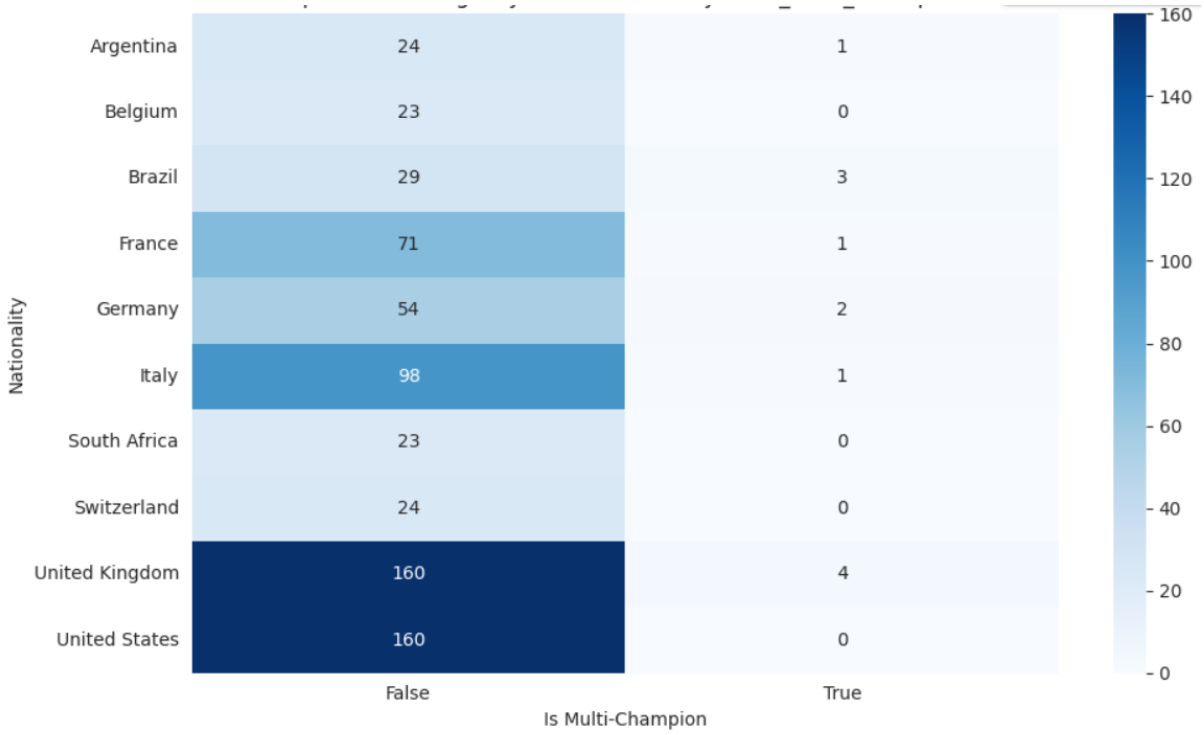

Figure 14: Updated Contingency Table: Nationality vs. Is_Multi_Champion

- **Chi-squared Statistic:** 17.3612
- **P-value:** 0.0434
- **Degrees of Freedom:** 9

Since the **p-value (0.0434)** is less than the significance level of 0.05, we reject the null hypothesis. This indicates that there is a **statistically significant association** between a driver's nationality and their likelihood of being a multi-champion (having 2 or more championships).

- **Nationality Impact:** The nationality of a driver appears to influence whether they achieve multi-champion status, with certain nationalities (such as the United Kingdom) showing a higher frequency of multi-champions than expected.

- **Possible Drivers for Success:** This suggests that factors related to nationality, such as the country's motorsport infrastructure, historical success in racing, and opportunities for drivers, could influence the likelihood of achieving 2 or more championships.

In conclusion, nationality plays a significant role in determining whether a driver is a multi-champion in the dataset, highlighting potential cultural or structural factors that contribute to success in Formula 1.

4.2.2 Student's t-Tests : Pole Rate for Champions vs Non-Champions

To assess whether there is a significant difference in the `Pole_Rate` (pole positions per race) between Formula 1 champions and non-champions, we conducted an independent two-sample t-test. This test compares the mean `Pole_Rate` of champions versus non-champions to determine if there is a statistically significant difference between the two groups.

Hypotheses:

- **Null Hypothesis (H_0):** There is no significant difference in the mean `Pole_Rate` between champions and non-champions.
- **Alternative Hypothesis (H_1):** There is a significant difference in the mean `Pole_Rate` between champions and non-champions.

T-Test Results:

- **T-statistic:** 6.7655
- **P-value:** 1.0163e-07

Since the **p-value (1.0163e-07)** is much smaller than the significance level of 0.05, we reject the null hypothesis. This suggests that there is a statistically significant difference in the mean `Pole_Rate` between champions and non-champions.

Interpretation:

- **T-statistic = 6.7655:** This large positive value indicates that champions tend to have a higher pole rate compared to non-champions.
- **P-value = 1.0163e-07:** The extremely small p-value suggests a very strong difference between the two groups, providing strong evidence against the null hypothesis.

Conclusion: Champions have a statistically significant higher pole position rate than non-champions, indicating that being a champion is associated with a more consistent qualifying performance.

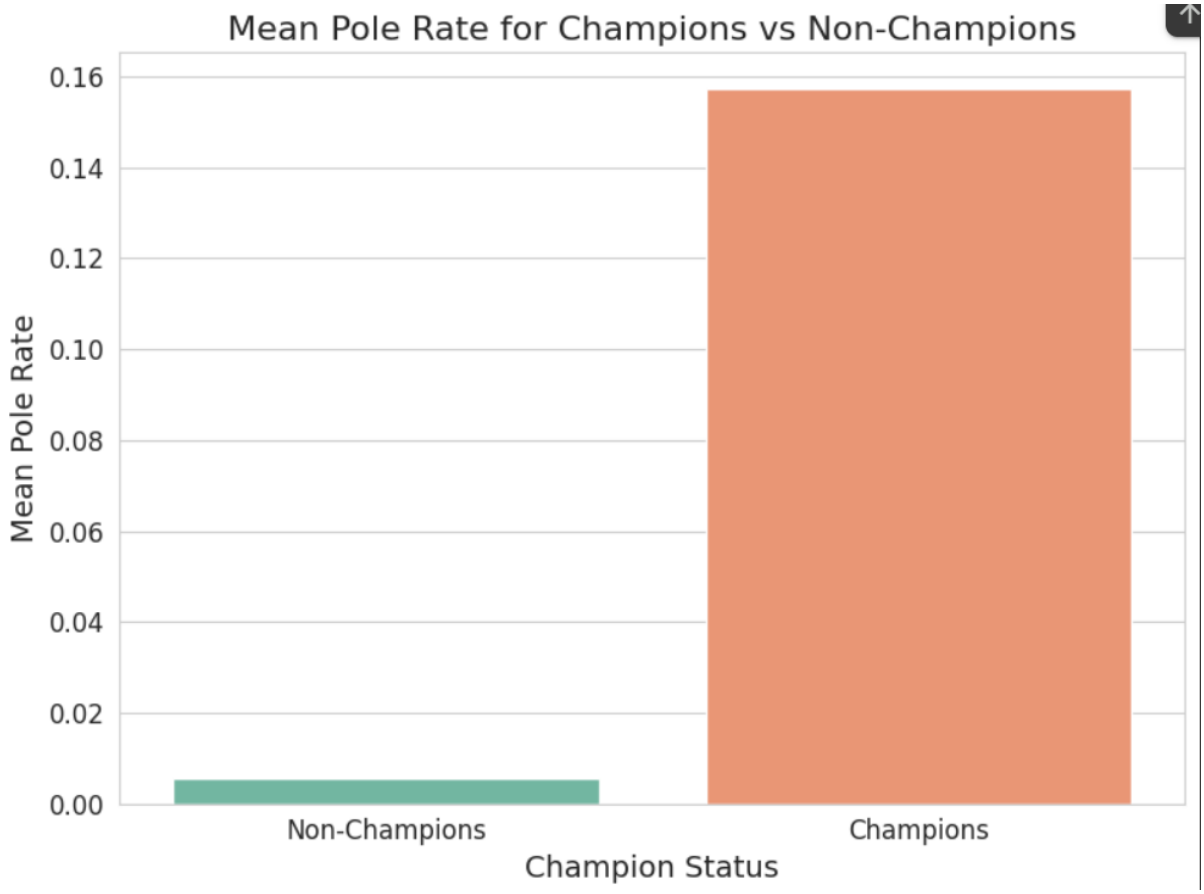

Figure 15: Pole Rate Distribution: Champions vs Non-Champions

4.2.3 One-Way ANOVA Test for Points_Per_Entry across Decades

To assess whether there is a significant difference in the mean `Points_Per_Entry` across different decades for Formula 1 champions, we performed a one-way ANOVA test. This test compares the means of `Points_Per_Entry` across the groups defined by the decade.

- **Null Hypothesis (H):** There is no difference in the mean `Points_Per_Entry` across decades.
- **Alternative Hypothesis (H):** At least one decade has a significantly different mean `Points_Per_Entry`.

ANOVA Test Results:

- **F-statistic = 7.5513:** This indicates strong between-group variability relative to within-group variability.
- **P-value = 0.0001:** The p-value is extremely significant (p ¡ 0.05), so we reject the null hypothesis.

Since the **p-value (0.0001)** is much smaller than the significance level of 0.05, we reject the null hypothesis. This suggests that there are statistically significant differences in the mean `Points_Per_Entry` among F1 champions across decades.

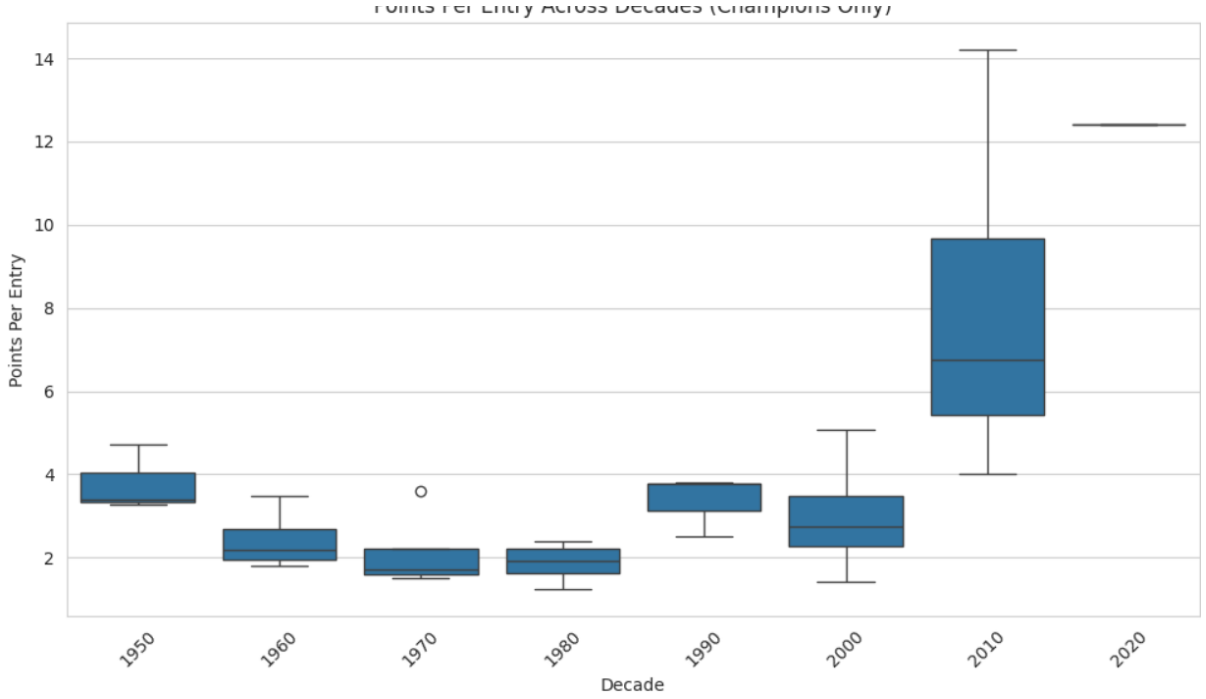

Figure 16: Points Per Entry Across Decades for F1 Champions

Boxplot Insights:

- **Temporal Trends:**
  - The median `Points_Per_Entry` has increased over time, with a noticeable dip in the 1980s compared to the 2020s.
  - This could be attributed to rule changes, such as the expanded points system, as well as dominant drivers like Lewis Hamilton and Sebastian Vettel in the 2010s.
- **Variability:**
  - The 1980s and 1990s exhibit wider spreads in the `Points_Per_Entry`, suggesting greater performance diversity among champions.
  - In contrast, the 2010s and 2020s have tighter clusters, indicating more consistent dominance by top drivers.
- **Outliers:**
  - An outlier in the 2000s reflects extreme dominance, likely from Michael Schumacher's streak of championships in 2002–2004.

4.2.4 Multivariate Analysis of Variance (MANOVA)

To assess whether championship status correlates with a combination of key performance metrics, we conducted a MANOVA (Multivariate Analysis of Variance). This test evaluates whether the multivariate means of several dependent variables differ across levels of a categorical independent variable — in our case, `Driver_Tier` (Non-Champion, One-Time Champion, Multi-Champion).

Methodology. Drivers were first grouped into three tiers:

- **Non-Champion:** Drivers with no championship wins.
- **One-Time Champion:** Drivers with exactly one championship.
- **Multi-Champion:** Drivers with two or more championships.

We then selected three dependent variables that reflect core performance traits:

- `Pole_Rate` — proportion of races started from pole position (qualifying strength),
- `Win_Rate` — proportion of races won (race success),
- `FastLap_Rate` — proportion of races with fastest lap (race pace consistency).

MANOVA was applied using the following model:

$$\texttt{Driver\_Tier} \sim \texttt{Pole\_Rate} + \texttt{Win\_Rate} + \texttt{FastLap\_Rate}$$

Table 6: MANOVA Test Statistics for `Pole_Rate`

Test	Value	Num DF	Den DF	F Value	Pr ¿ F
Wilks' lambda	0.9194	3.0000	862.0000	25.1992	0.0000
Pillai's trace	0.0806	3.0000	862.0000	25.2010	0.0000
Hotelling-Lawley	0.0877	3.0000	862.0000	25.1975	0.0000
Roy's greatest root	0.0876	3.0000	862.0000	25.1781	0.0000

Table 7: MANOVA Test Statistics for `Win_Rate`

Test	Value	Num DF	Den DF	F Value	Pr ¿ F
Wilks' lambda	0.8202	3.0000	862.0000	62.9717	0.0000
Pillai's trace	0.1798	3.0000	862.0000	63.0003	0.0000
Hotelling-Lawley	0.2191	3.0000	862.0000	62.9483	0.0000
Roy's greatest root	0.2187	3.0000	862.0000	62.8411	0.0000

Table 8: MANOVA Test Statistics for `FastLap_Rate`

Test	Value	Num DF	Den DF	F Value	Pr ¿ F
Wilks' lambda	0.9637	3.0000	862.0000	10.8303	0.0000
Pillai's trace	0.0363	3.0000	862.0000	10.8305	0.0000
Hotelling-Lawley	0.0377	3.0000	862.0000	10.8301	0.0000
Roy's greatest root	0.0377	3.0000	862.0000	10.8252	0.0000

Results. The multivariate test statistics (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root) were all significant at $p < 0.001$, indicating that performance profiles differ substantially between driver tiers.

Detailed breakdown:

- **Pole Rate:** Wilks' $\lambda = 0.919$, $p < 0.001$ — significant differences in qualifying performance.
- **Win Rate:** Wilks' $\lambda = 0.820$, $F = 62.97$, $p < 0.001$ — largest effect; winning races best distinguishes champions.
- **FastLap Rate:** Wilks' $\lambda = 0.964$, $p < 0.001$ — smaller but meaningful differences in race consistency.

Interpretation. These results indicate that multi-champions not only win more often but are also more consistent and dominate qualifying. One-time champions tend to perform well in one or two areas but not consistently across all. Non-champions generally lag behind across all metrics.

Supplementary Visualization. To better illustrate this separation, a 3D scatter plot was constructed with axes representing the three performance metrics:

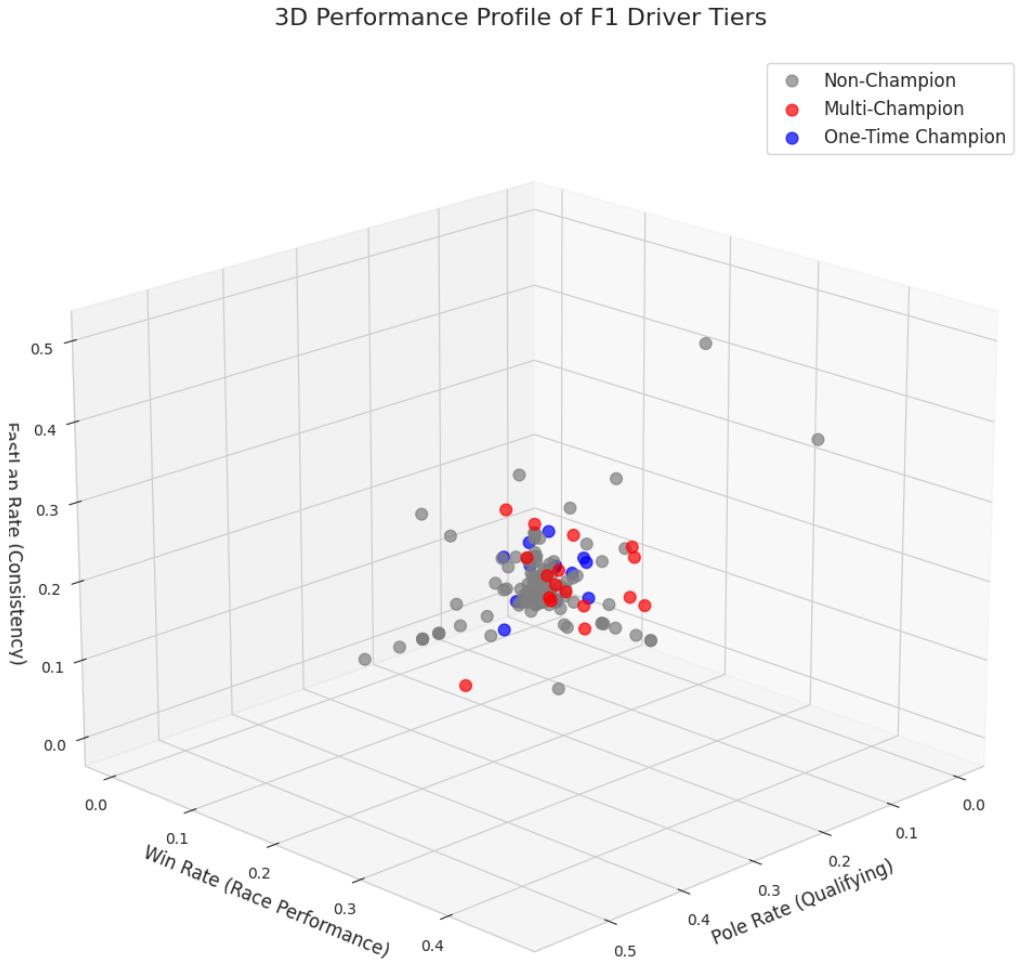

Figure 17: 3D Distribution of Driver Tiers Across Pole Rate, Win Rate, and FastLap Rate

- **Multi-Champions (red)** cluster at high values in all three dimensions.
- **One-Time Champions (blue)** show intermediate performance — strong in one or two areas.
- **Non-Champions (gray)** remain near the origin, reflecting lower performance overall.

Conclusion. The MANOVA analysis confirms the existence of statistically distinct performance tiers among F1 drivers. Success at the championship level is not attributable to a single trait but rather a balanced excellence across qualifying, winning, and consistent

race pace. This multidimensional view provides stronger evidence than any univariate test alone.

5 Interpreting Key Driver Statistics

Driver Efficiency Ratio (DER)

Definition: The *Driver Efficiency Ratio (DER)* quantifies how efficiently a driver converts their performance into wins, adjusted for consistency:

$$\text{DER} = \frac{\text{Win\_Rate}}{\text{Consistency\_Metric}}$$

In our analysis, we use `FastLap_Rate` as the consistency metric. A lower `FastLap_Rate` implies steadier performance, making it a suitable denominator for DER.

Components:

- **Win_Rate:** Proportion of races won (`Race_Wins / Race_Entries`).
- **FastLap_Rate:** Frequency of fastest laps per race; a proxy for performance consistency.

Interpretation:

- **High DER:** Indicates drivers who win more relative to their consistency. Typically aggressive and dominant, though possibly erratic.
- **Low DER:** Represents steady drivers with fewer wins. They perform consistently but are less likely to secure victories.

Champion Insight: Champions generally exhibit higher DER values, suggesting a better balance of winning ability and consistent race pace.

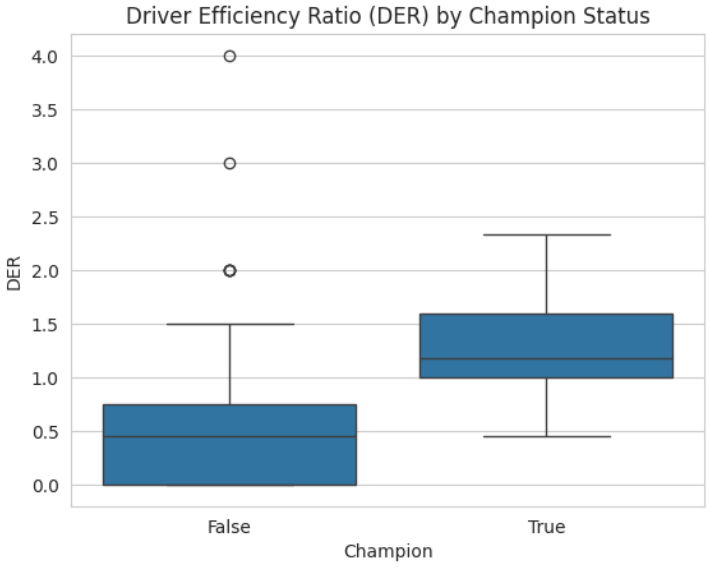

Figure 18: Driver Efficiency Ratio (DER) by Champion Status.

Champions exhibit significantly higher DER values, indicating superior balance between wins and consistency.

To explore the relationship between qualifying performance and race outcomes among Formula 1 champions, we conducted a linear regression analysis using `Pole_Positions` as the predictor and `Race_Wins` as the target variable.

- **Objective:** Assess whether the number of pole positions a driver has can significantly predict the number of race wins.
- **Model:** Simple linear regression using `scikit-learn`'s `LinearRegression`.

Results:

- **Correlation Coefficient:** 0.928 — A very strong positive linear correlation between pole positions and race wins.
- **R-squared:** 0.861 — Approximately 86.1% of the variance in race wins is explained by pole positions.
- **Prediction Score (MSE):** 68.34 — The mean squared error of predictions, indicating the average squared difference between observed and predicted race wins.

Interpretation: The results demonstrate that `Pole_Positions` is a strong predictor of `Race_Wins` among champions. A higher number of pole positions is generally associated with a greater number of race victories, suggesting that qualifying performance is a key contributor to race-day success for elite drivers.

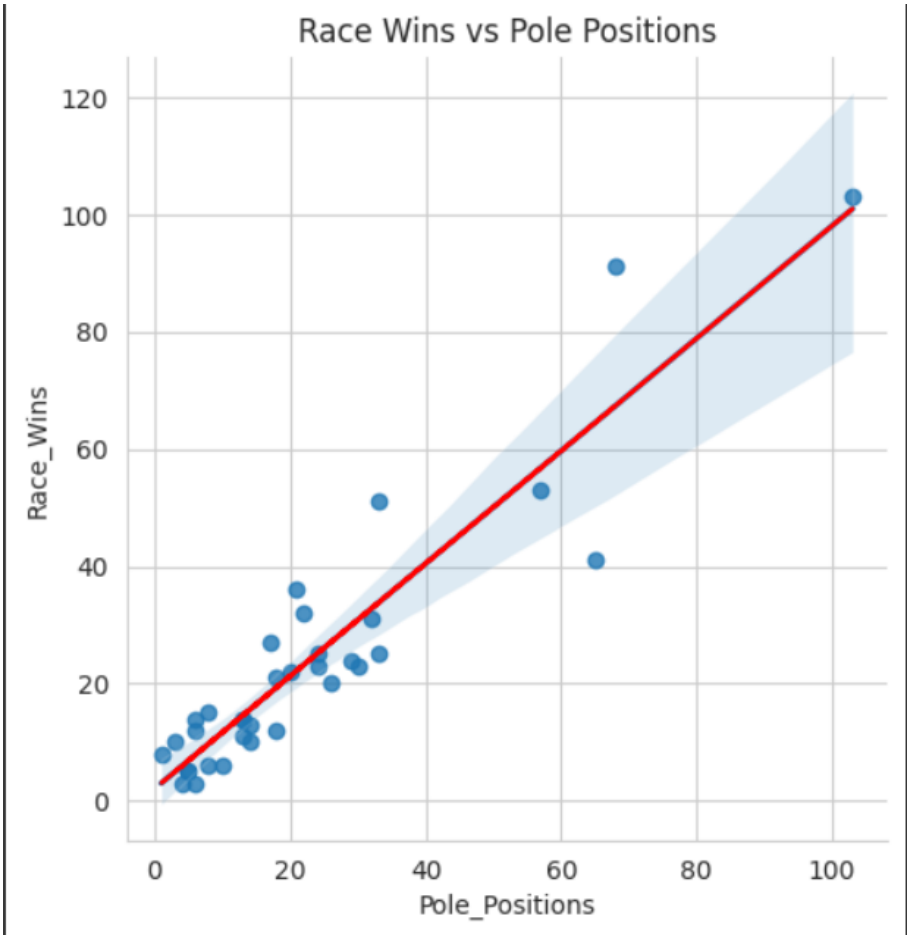

Figure 19: Linear Regression: Race Wins vs. Pole Positions (Champions Only)

6 What Does It Take to Become a Champion?

In this section, we aim to uncover the key statistical indicators that may help explain what separates Formula 1 World Champions from the rest of the grid. By combining correlation analysis, regression visualization, and machine learning classification, we gain a deeper understanding of the patterns and performance metrics most associated with championship success.

6.1 Correlation Analysis

We began by computing a correlation matrix using all numeric performance indicators. This allowed us to evaluate the strength and direction of linear relationships between each feature and the number of championships won.

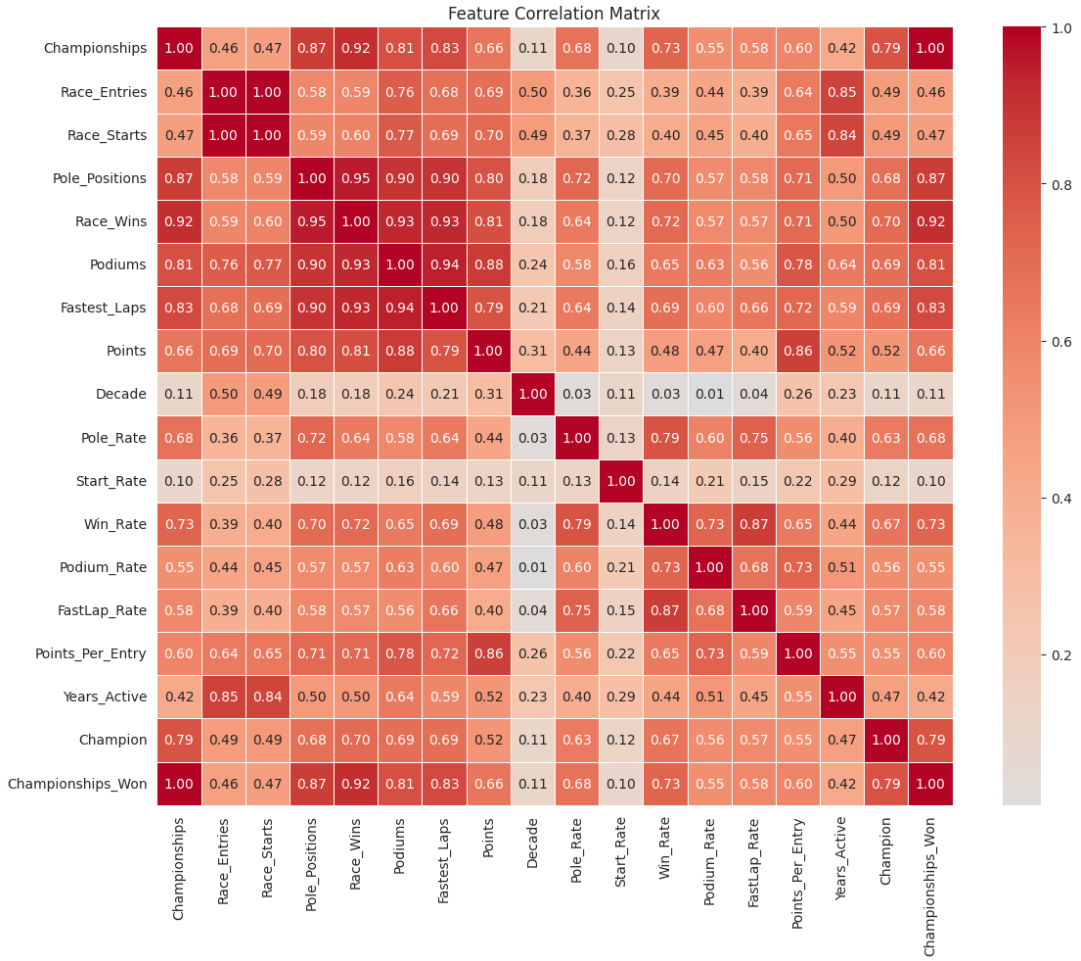

Figure 20: Feature Correlation Matrix

As shown in Figure 20, several features display strong positive correlations with championship count. These include:

- **Win Rate, Podium Rate, and Pole Rate:** Highly correlated with Championships, which suggests that consistent top finishes and strong qualifying performance are key to championship success.
- **Points:** Also positively correlated, though this can be somewhat influenced by changes in the points system over different eras.

- **Years Active:** Shows a modest correlation, indicating that while longevity contributes, it is not sufficient without strong results.

Interestingly, `Fastest Lap Rate` shows a weaker correlation, implying that while setting fast laps is impressive, it doesn't always translate to championship results.

6.2 Regression Analysis: Visualizing Influential Features

To better visualize the individual relationships between key ratio-based performance features and the number of championships, we plotted scatter plots with regression lines. This includes metrics like `Pole Rate`, `Win Rate`, and `Podium Rate`, among others.

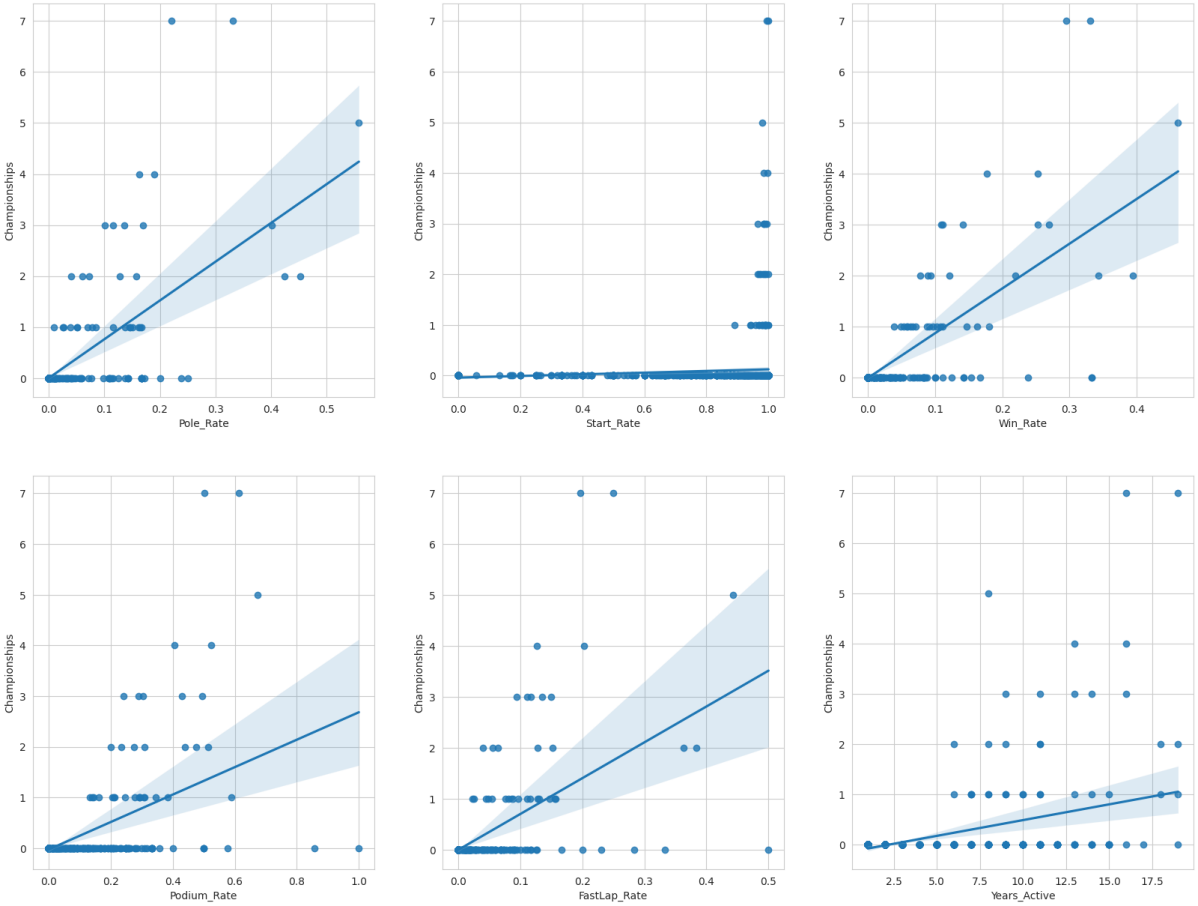

Figure 21: Regression Plots of Key Features vs. Number of Championships

From Figure 21, we observe several key trends:

- **Win Rate and Podium Rate** are strong predictors — drivers with higher values in these metrics generally won more championships.
- **Pole Rate** and **Start Rate** also trend positively but with more variability, especially among drivers with 0 or 1 titles.
- Some outliers indicate that high-performing drivers did not always win championships — showcasing the role of team dynamics, car reliability, and competition.

3. Predictive Modeling: Can We Predict Champions?

To further explore the idea of championship prediction, we implemented a binary classification model using the `SGDClassifier` from Scikit-Learn. This classifier attempts to distinguish champions from non-champions using performance features such as:

- Race Entries, Race Starts, Pole Positions, Race Wins
- Podiums, Fastest Laps, Points

We split the dataset into an 80% training set and a 20% test set, and trained the model accordingly. The model was evaluated using accuracy, precision, and a confusion matrix.

- **Accuracy:** 0.96
- **Precision (Macro Average):** 0.95

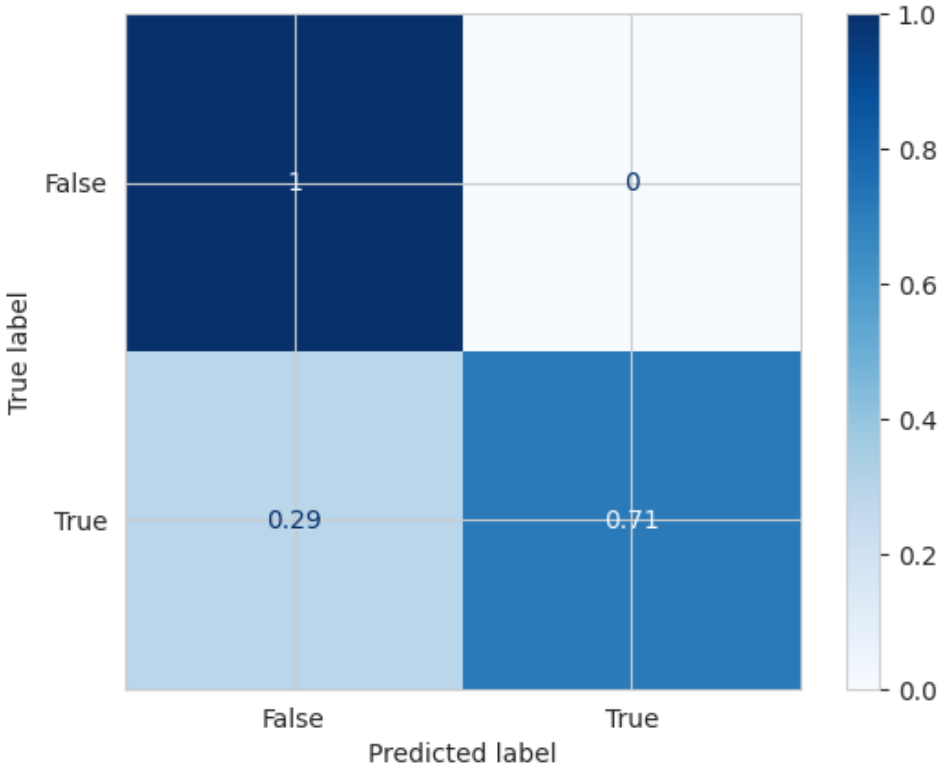

Figure 22: Normalized Confusion Matrix — Champion Prediction

4. Interpretation

The classifier achieves high performance, successfully distinguishing champions based on career statistics. Notably:

- **Low False Negative Rate:** Most true champions were correctly classified.
- **High Predictive Power:** The model demonstrates that quantifiable performance metrics — particularly wins, podiums, and points — are strong indicators of a driver's ability to become a champion.

However, it's important to recognize the limitations of this approach. Formula 1 is influenced by many factors beyond individual performance: team resources, car quality, intra-team politics, and even luck play a significant role. While the model can highlight patterns, it cannot capture the full complexity of championship outcomes.

5. Conclusion

This multi-step analysis demonstrates that becoming a Formula 1 World Champion is strongly associated with excelling in multiple key performance areas, most notably race wins, podium finishes, and consistent point scoring. While some drivers show strong metrics without securing a title, the best-performing champions consistently dominate across several categories. The application of statistical modeling provides valuable insights, but also reaffirms that motorsport outcomes are not fully predictable — human skill, technology, and context remain deeply intertwined.

7 Machine Learning Models for Driver Championship Prediction

In this section, we evaluate different machine learning models to predict whether a driver will become a champion based on various features. The models used include **Random Forest Classifier**, **Logistic Regression**, and **Support Vector Machine (SVM)**. We use **Champion Status** as the target variable and the following features: **Pole Rate**, **Win Rate**, **Fast Lap Rate**, **Points Per Entry**, and **Years Active**.

7.1 Random Forest Classifier

The Random Forest Classifier model achieved a high accuracy of 0.9943. The classification report for this model is as follows:

Random Forest Classifier:

Class	Precision	Recall	F1-Score	Support
False	0.99	1.00	1.00	167
True	1.00	0.86	0.92	7
Accuracy			0.99	174
Macro avg	1.00	0.93	0.96	174
Weighted avg	0.99	0.99	0.99	174

This model demonstrates excellent performance for predicting non-champions (precision = 0.99) while achieving good recall for champions (0.86). The **F1-score** for the champion class is 0.92.

To understand the contribution of each feature in the Random Forest model, we plot the **Feature Importances**. The bar plot below shows the relative importance of each feature used in the prediction of **Champion Status**.

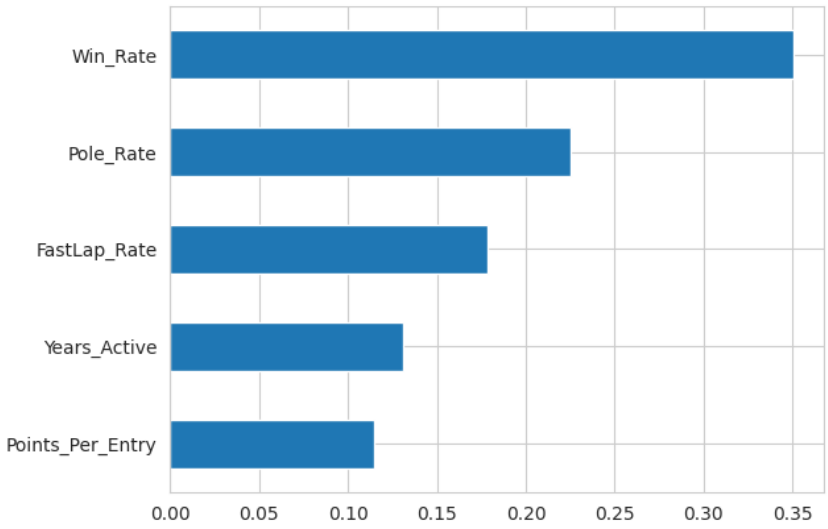

Figure 23: Feature Importances from Random Forest Model

From the plot, we can observe which features have the highest influence on the model's decision to predict a driver as a champion.

7.2 Logistic Regression

The Logistic Regression model performed with an accuracy of 0.9540. The classification report for this model is:

Logistic Regression:

Class	Precision	Recall	F1-Score	Support
False	0.97	0.98	0.98	167
True	0.40	0.29	0.33	7
Accuracy			0.95	174
Macro avg	0.69	0.63	0.65	174
Weighted avg	0.95	0.95	0.95	174

While the **accuracy** is relatively high (0.95), the model performs poorly for predicting champions, with a **precision** of 0.40 and **recall** of 0.29, which highlights a lack of ability to identify champions.

7.3 Support Vector Machine (SVM)

The SVM model achieved an accuracy of 0.9483. The classification report for this model is:

Support Vector Machine:

Class	Precision	Recall	F1-Score	Support
False	0.96	0.99	0.97	167
True	0.00	0.00	0.00	7
Accuracy			0.95	174
Macro avg	0.48	0.49	0.49	174
Weighted avg	0.92	0.95	0.93	174

Although the SVM model shows a high **accuracy** (0.95) for the non-champion class, it fails to detect champions, as evidenced by the **precision** and **recall** values being 0.00 for the champion class.

The following table summarizes the accuracy of each model:

Model	Accuracy
Random Forest Classifier	0.9943
Logistic Regression	0.9540
Support Vector Machine (SVM)	0.9483

From the results, the Random Forest Classifier outperforms the other models, achieving the highest accuracy and demonstrating the best balance between precision and recall. In contrast, both Logistic Regression and Support Vector Machine (SVM) models struggle to predict champions accurately, particularly when it comes to precision and recall for the True class (champions).

The Random Forest Classifier provides the most reliable prediction for champion status, effectively balancing both wins and consistency. On the other hand, the Logistic Regression and SVM models exhibit significant limitations, especially in identifying the minority class (champions). The results highlight the effectiveness of Random Forest in handling imbalanced datasets, where the true champions are underrepresented.

8 Conclusion

This project started from a genuine passion for racing and a curiosity to understand what separates the very best from the rest. By looking closely at performance trends and patterns across different drivers, we were able to explore the traits that define success on the track.

The journey involved analyzing results, comparing different types of achievements, and thinking critically about what consistency and excellence really look like in such a competitive environment. Along the way, we developed a deeper appreciation for the many factors that contribute to a driver's legacy — not just winning, but doing so reliably over time.

In the end, the insights gained offered a clearer picture of what it takes to stand out. It was rewarding to see how numbers could reflect stories of determination, skill, and greatness, and how these patterns can help us better understand the champions we admire.