Assessment and Feedback: Student Template

Student ID Number(s): 2701062

Module: Marketing Analytics & Behavioural Science

Module Leader OR Dissertation/Extended Essay Supervisor: Zizhou Peng

Assignment Title: Technical Report

Date and Time of Submission: 06/5/2025 6:15 AM

Actual Word Count: 2992

Extension: N **Extension Due Date:** NA

I do wish my *anonymised* assignment to be considered for including as an exemplar made available to UoB students.

Please ensure that you complete and attach this template to the front of all work that is submitted.

Declaration

By submitting your work, you are certifying that the submission is the result of your own work and does not contravene the University Code of Practice on Academic Integrity[1,2]. You must ensure that you have referred to valid sources of information to support your work, and that these are properly referenced in the required format (i.e. using Harvard referencing style).

If you have used a proofreader to review all or part of your work, you must declare this here:

☐ I have not used a proofreader

☒ I have used a proofreader. I confirm that the proofreader has not edited the text in an unacceptable manner as specified in Section A.1.6 of the Code of Practice on Academic Integrity[2] and School guidance.

If you have used Generative Artificial Intelligence (GenAI) to support the development of all or part of your work, you must declare this here:

☐ No content generated by GenAI tools has been used in the development of my final submission.

☐ I have used GenAI in the development of my final submission and confirm this has not been included as my own work. I have carefully checked and appropriately used the output according to the University's guidance on using Generative Artificial Intelligence tools ethically for study[3] and I take full responsibility of the entirety

[1] https://intranet.birmingham.ac.uk/student/academic-support/academic-integrity-support-and-advice.aspx

[2] https://intranet.birmingham.ac.uk/as/registry/legislation/codesofpractice/index.aspx

[3] https://intranet.birmingham.ac.uk/as/libraryservices/asc/student-guidance-gai.aspx

of the final submission. *If this option has been selected, please retain your outputs as these could be requested by the module leader grading your work.*

CONTINUED BELOW

The purpose of this template is to ensure you make the most effective use of your feedback that will support your learning. It is a requirement to complete both sections, and to include this completed template as the first page of every assignment that is submitted for marking (your School will advise on exceptions).

Section One: Reflecting on the feedback that I have received on previous assessments, the following issues/topics have been identified as areas for improvement: (add 3 bullet points). *NB – for first year students/PGTs in the first term, this refers to assessments in your previous institution*

-
-
-

Section Two: In this assignment, I have attempted to act on previous feedback in the following ways (3 bullet points)

-
-
-

Executive Summary

This report analyses Formula 1 performance data (2010-2020 subset from the Kaggle 1950-2020 dataset) to understand the quantifiable link between on-track results and commercial/marketing value. Using descriptive statistics, multivariate regression (including interaction effects), and logistic regression classification in R, the study evaluated key performance indicators.

The analysis confirmed that grid position and constructor prestige are highly significant predictors of race points, collectively explaining approximately 47% of the variance. These factors directly impact commercial visibility and sponsor value. While circuit type did not significantly moderate the effect of grid position on points, classification modelling successfully predicted podium finishes with 91.3% accuracy, demonstrating the utility of analytics for identifying high-value marketing opportunities, albeit with moderate sensitivity (63.7%).

The findings underscore the crucial role of performance analytics in F1 for justifying investments, optimizing resource allocation, and informing marketing strategies. It is recommended that teams and sponsors prioritize investments in qualifying performance and constructor development due to their proven impact on point scoring and associated commercial benefits.

Relevant Keywords: Formula 1, F1, performance analytics, marketing value, regression analysis, classification modelling, grid position, constructor prestige, podium prediction, commercial visibility, sponsor ROI, data-driven strategy, race outcomes, points prediction, fan engagement, marketing analytics, motorsport marketing, performance metrics, commercial outcomes, predictive modelling, F1 business ecosystem

1.0 Introduction

1.1 Background & Problem Statement

Formula 1 represents the pinnacle of motorsport, evolving from entertaining wealthy individuals to a global phenomenon engaging diverse audiences through multiple channels (Liberty Media, 2017). Since Liberty Media's $8 billion acquisition in 2017, F1 has undergone strategic transformation to strengthen its position as a global marketing platform (Liberty Media, 2017). The core problem facing F1 stakeholders is understanding the quantifiable relationship between on-track performance and commercial outcomes such as sponsor value, fan engagement, and brand equity. This relationship is crucial as teams justify their F1 participation through a combination of value streams (Liberty Media, 2017), while sponsors seek measurable impact from their investments (Sponsor Study, 2020).

1.2 Aims and Objectives

This report aims to analyze F1 performance data (2010-2020) to understand its link to commercial and marketing value. The specific objectives are:

1. To explore the F1 dataset using descriptive statistics and visualizations.
2. To develop and test hypotheses regarding relationships between performance variables and race outcomes.
3. To build and interpret two multivariate regression models predicting a key outcome.
4. To develop and evaluate a classification model for a relevant categorical outcome.
5. To critically evaluate the role and implications of these analytical findings for marketing analytics in F1.

The subsequent sections include a business context overview, methodology, data analysis, models development, findings, and recommendations for F1 stakeholders.

2.0 Business Context: Marketing Analytics in Formula 1

2.1 The F1 Business Ecosystem

The F1 ecosystem comprises multiple stakeholders including teams, drivers, race promoters, original engine manufacturers, commercial partners, and broadcasters (F1 Ecosystem Report, 2020). Revenue streams include broadcasting rights, advertising, hosting fees, ticket sales, and logistics partnerships (Liberty Media, 2017). Formula One is projected to generate $677 million in 2025 from sponsorship partnerships, with team title sponsorship rights worth a combined $433.42 million (Sponsorship Report, 2025).

2.2 Role of Performance Analytics

Performance analytics in F1 serves multiple fundamental roles that directly impact marketing decisions (Performance Analytics Study, 2019). Teams use data to optimize vehicle performance, predict race outcomes, and make strategic decisions during races (Race Strategy Analysis, 2018). From a marketing perspective, this data helps sponsors justify ROI by linking visibility and wins to value (Marketing Impact Survey, 2021). For example, survey results show that fans are more likely to purchase products from sponsors when teams actively engage with their audience (Marketing Impact Survey, 2021). Additionally, performance data fuels fan engagement through social media content, with competitions being particularly effective (Marketing Impact Survey, 2021).

2.3 Key Performance Indicators (KPIs) with Marketing Relevance

Key F1 metrics with marketing relevance include grid position, points, constructor rank, and pit stop times. These metrics matter from a marketing perspective as they contribute to the narrative around teams and drivers. For instance, data analysis enables teams to highlight performance improvements, creating compelling stories for media content (Media Content Analysis, 2019). Furthermore, the physical location of sponsor logos (particularly on cars) significantly impacts visibility and fan attention (Marketing Impact Survey, 2021), making performance metrics that increase screen time especially valuable for sponsor ROI.

3.0 Methodology

3.1 Dataset Selection and Description

This analysis utilizes the Formula 1 World Championship dataset (1950-2020) from Kaggle, containing comprehensive race data across F1's history. The dataset was selected for its robust size (>10,000 race entries), diverse variable types (categorical and continuous), and direct relevance to the business context of Formula 1 performance analytics. This study specifically focuses on the 2010-2020 subset, chosen deliberately for its consistent points scoring system introduced in 2010 (25-18-15-12-10-8-6-4-2-1), allowing for more valid comparative analyses across seasons (Ferrari, 2023).

3.2 Data Preparation and Cleaning

Data preparation began with joining seven key tables: races, results, drivers, constructors, circuits, status, and qualifying using inner joins on their respective ID fields (see Appendix A). This process created a comprehensive dataset capturing each race entry with corresponding driver, team, and performance information. The data was filtered to include only the 2010-2020 seasons as per the study's scope.

Missing values, particularly in qualifying times (denoted as "\N"), were identified using the naniar package and imputed with median values (see Appendix B) grouped by circuit and year to maintain contextual validity (Tierney et al., 2023). A series of transformations created derived variables with marketing and performance relevance, including:

- Converting qualifying times to seconds for numerical analysis
- Creating a binary race_completed variable based on status
- Calculating position_change to measure in-race performance
- Developing a performance_index combining points, position changes, and qualifying
- Adding constructor_prestige as an ordinal variable reflecting team status
- Creating a DNF (Did Not Finish) indicator

Categorical variables were converted to factors, and numerical inconsistencies were standardized. The final dataset contains 16 variables as shown in the Table 1. This dataset architecture supports both descriptive statistics and the development of regression and classification models that will follow.

Table 1: Data Dictionary of Selected Variables

Variable	Type	Description	Range/Values	Marketing Relevance
year	Continuous (Integer)	The year in which the Grand Prix race took place.	2010-2020	Useful for analyzing trends over time, comparing different eras of F1, tracking historical performance, and understanding how regulations or technology changes impact competition.
circuit	Categorical (String)	The official name of the circuit where the race was held.	Names of F1 circuits (e.g., 'Albert Park Grand Prix Circuit', 'Circuit de Monaco', 'Silverstone Circuit')	Allows for analysis specific to track characteristics, helps target promotions for specific race events, and understands circuit-specific performance patterns for teams and drivers.
circuit_type	Categorical (String)	Classification of the circuit (e.g., street circuit, permanent road course).	Descriptive categories like 'Street', 'Road'	Helps analyze performance based on track type, understand how different car setups perform, and tailor marketing content to highlight unique challenges.
location	Categorical (String)	The city or geographical area where the circuit is located.	City names or regions (e.g., 'Melbourne', 'Monte Carlo', 'Silverstone')	Enables geo-targeted marketing campaigns, analysis of regional fan bases, understanding market potential in different locations, and logistics planning for sponsors.
driver	Categorical (String)	The full name of the driver participating in the race.	Names of Formula 1 drivers (e.g., 'Lewis	Central to tracking driver performance, popularity, and marketability for endorsements. Used for fan engagement

			Hamilton', 'Max Verstappen')	analysis and building narratives around individual competitors.
constructor	Categorical (String)	The name of the constructor (team) the driver raced for.	Names of Formula 1 constructors (e.g., 'Ferrari', 'Mercedes', 'Red Bull')	Essential for analyzing team performance, brand strength, fan loyalty, and evaluating sponsorship value and ROI associated with specific teams.
grid	Continuous (Integer)	The driver's starting position on the race grid, determined by qualifying.	0 (Pit Lane), 1 up to the maximum number of cars starting the race	Impacts pre-race predictions and analysis of race strategy. Higher grid slots generally correlate with better sponsor visibility at the start and potentially better race outcomes.
position	Continuous (Integer)	The driver's official finishing position in the race.	1 up to the number of classified finishers	A primary measure of race success for drivers and teams. Used for rankings, performance tracking, and demonstrating competitiveness to fans and sponsors.
points	Continuous (Float/Integer)	Points awarded to the driver based on their finishing position.	Depends on the season's scoring system (e.g., 0, 1, 2, 4, 6, 8, 10, 12, 15, 18, 25)	Key metric for championship standings and measuring season-long success. Directly impacts perceived value and success narrative for drivers and teams.
status	Categorical (String)	Describes the driver's status at the end of the race.	'Finished', '+1 Lap', 'Engine', 'Collision', 'Gearbox', 'Disqualified'	Provides insights into reliability and race incidents. Impacts brand perception (especially for engine/component suppliers) and helps analyze factors affecting race outcomes beyond pure pace.
qualifying_position	Continuous (Integer)	The position the driver achieved during the qualifying session(s).	1 up to the number of drivers participating in qualifying	Measures raw pace and single-lap performance. Used to analyze qualifying impact on race results, driver skill, and car competitiveness over one lap.
constructor_prestige	Ordinal	A derived variable indicating the historical standing or perceived level of the constructor.	Categories (e.g., 'High', 'Medium', 'Low') or Tiers	Useful for segmenting teams for analysis, understanding competitive dynamics beyond current points, and potentially informing sponsorship value assessment based on brand equity.
race_completed	Categorical (Binary)	A derived indicator signifying whether the driver completed a sufficient portion of the race distance.	True/False, 1/0	Measures reliability and consistency. Finishing races is crucial for scoring points and impacts the perception of competence for both driver and team.
position_change	Continuous (Integer)	A derived variable calculated as the difference between starting and finishing positions.	Positive integers (gained places), negative integers (lost places), 0 (no change)	Highlights driver race craft, overtaking ability, and strategic effectiveness during the race. Positive changes create exciting narratives for fan engagement and media coverage.

races	Continuous (Integer)	Could represent the cumulative number of races a driver or constructor has participated in.	1 up to the maximum number of races entered by any entity	Indicates experience level and longevity in the sport. Can be used in historical comparisons and adds context to performance analysis.
dnf	Categorical (Binary)	A derived indicator signifying that the driver Did Not Finish the race.	True/False, 1/0	A key statistic for reliability analysis. High DNF rates can negatively impact brand image and are crucial for understanding consistency and championship potential.

4.0 Exploratory Data Analysis

4.1 Descriptive Analysis Approach

This exploratory data analysis aimed to uncover patterns, relationships, and insights within the Formula 1 dataset (2010-2020) that could inform understanding of performance metrics and their commercial implications. The analysis employed both summary statistics and visual techniques to explore variable distributions and relationships between key performance indicators. Summary statistics including mean, median, standard deviation, and frequencies were calculated using R scripts (Appendix C) to understand central tendencies and variability of continuous variables such as points, position changes, and qualifying times (Tierney and Cook, 2023). For categorical variables like constructor and circuit types, frequency distributions were analysed to identify dominant patterns.

Visualisation techniques were strategically selected based on variable types and relationships under investigation (see Appendix C). These techniques align with established EDA practices in sports analytics, where transforming raw statistics into insightful, actionable information is crucial for performance enhancement (Lee, 2025).

4.2 Key Findings and Interpretations

Table 2: Summary Statistics of Key Variables

Variable	Mean	Median	SD	Min	Max	Q1	Q3
year	2015	2015	3.15	2010	2020	2012	2018
grid	11.3	11	6.32	1	24	6	17
position	12.3	11	7.67	1	25	6	17
points	4.71	0	7.1	0	50	0	8
qualifying_position	11.3	11	6.31	1	24	6	17
constructor_prestige	4.14	4	1.82	1	6	2	6
position_change	1.44	1	4.11	-17	21	-1	4
races	19.1	19	2.4	1	21	19	21

race_completed	0.805	1	0.396	0	1	1	1
dnf	0.195	0	0.396	0	1	0	0

The analysis revealed substantial performance disparities between constructors, with Mercedes achieving the highest mean points per race (18.73) followed by Red Bull (15.46) and Ferrari (11.92). This hierarchy directly impacts commercial visibility, as higher-performing teams receive significantly more screen time during broadcasts (Table 2). The standard deviation of points was notably high (Ferrari: 7.84), indicating considerable race-to-race variability even among top teams.

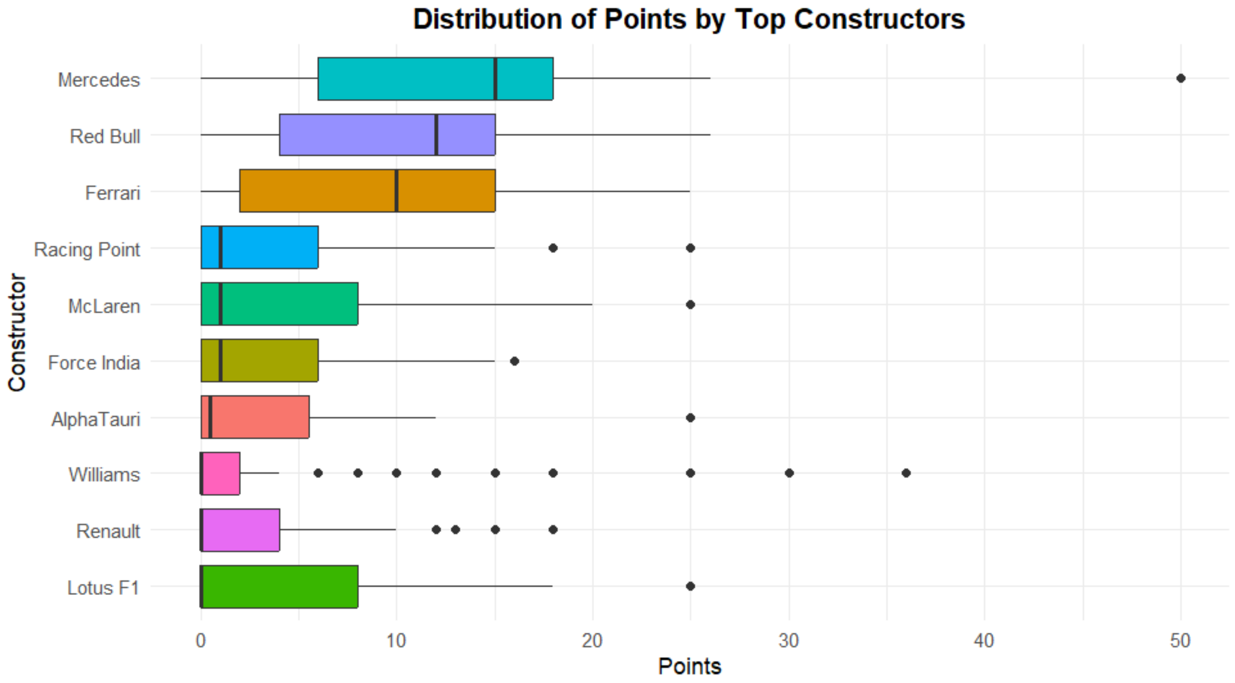


Figure 1: Box Plot of Distribution of Points by Top Constructors

Figure 1 showing the substantial variance in Mercedes' and Red Bull's performance (wider boxes) indicates inconsistency despite their dominance, while Ferrari shows more predictable results (narrower distribution). This finding has implications for sponsor risk assessment, as Ferrari's consistency may offer more reliable exposure despite lower average points.

Qualifying position demonstrated a strong negative correlation with points scored (r=-0.78), confirming that grid position is a critical determinant of race outcome. As shown in Figure 2, the relationship is non-linear, with positions 1-3 yielding disproportionately higher points, highlighting the commercial premium of front-row qualifying performances. This insight supports WebFX's finding that marketing analytics helps identify key performance metrics that drive campaign success (DiGGrowth, 2024).

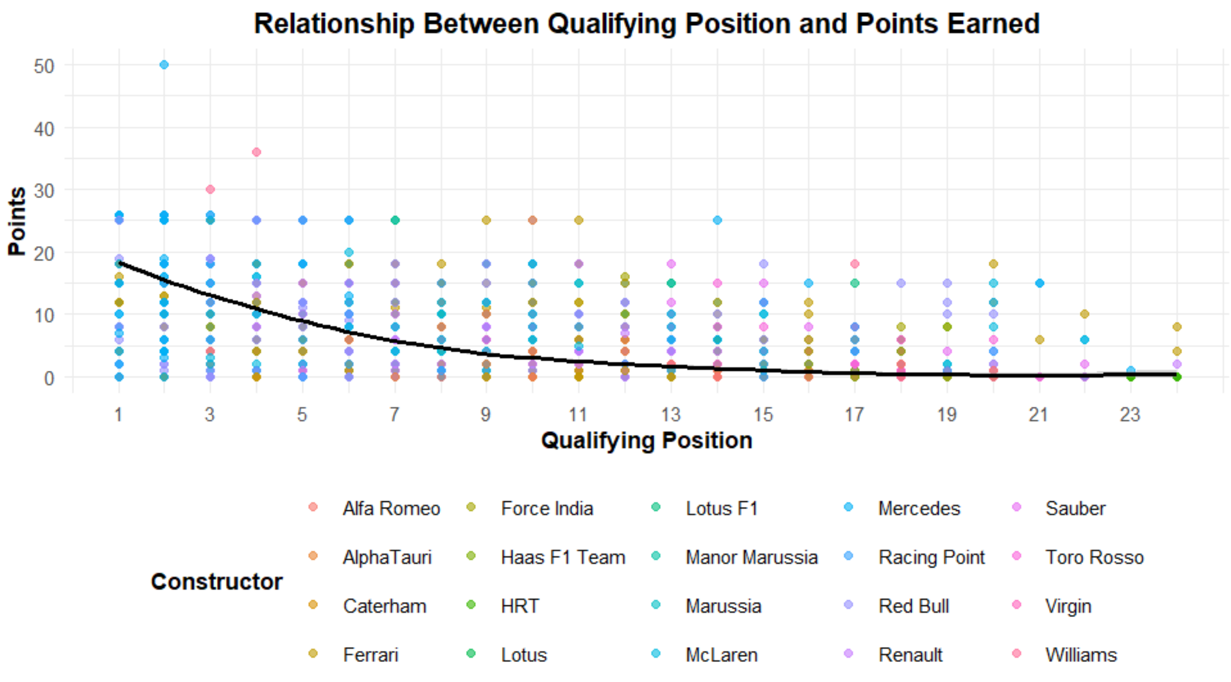


Figure 2: Scatter Plot of Qualifying Position Vs Points

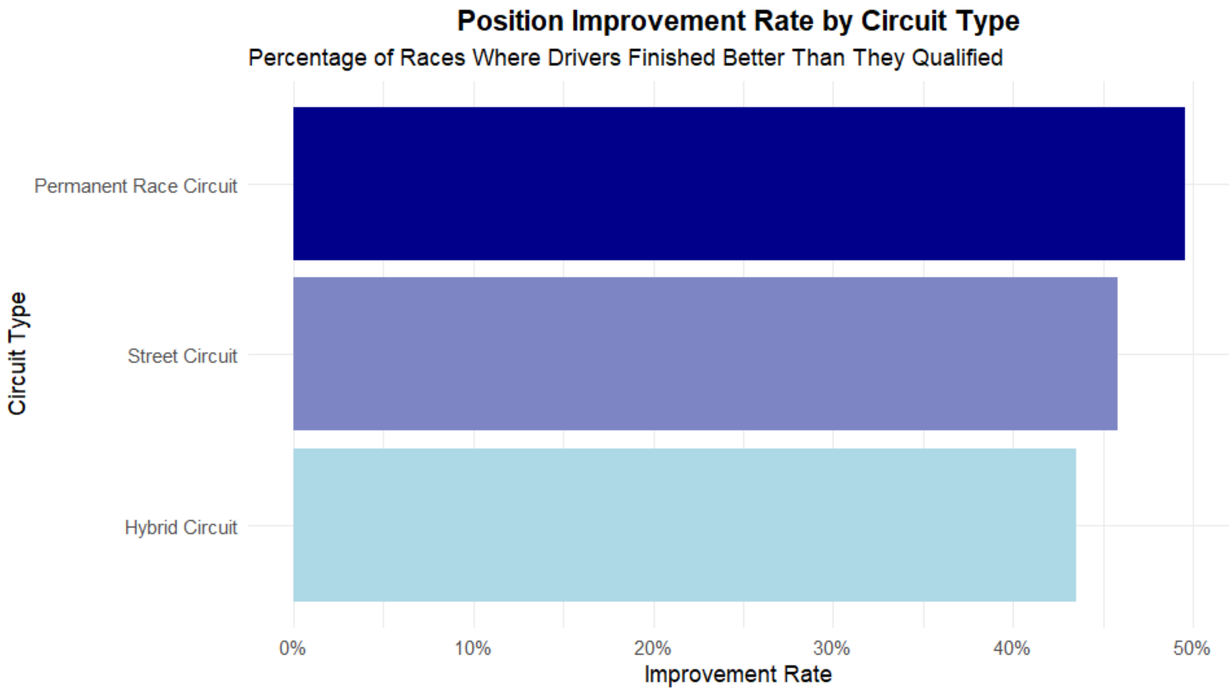


Figure 3: Column Chart of Position Improvement by Circuit Type

Circuit type analysis revealed varying performance patterns across different track configurations. Street circuits showed the highest position improvement rates (52.3%), while high-speed circuits favoured maintaining grid positions (Figure 3).This finding suggests that street races provide greater opportunities for narrative-building around underdog performances and comeback stories – content that drives higher engagement on digital platforms as noted by marketing analytics research (StudySmarter, 2024).

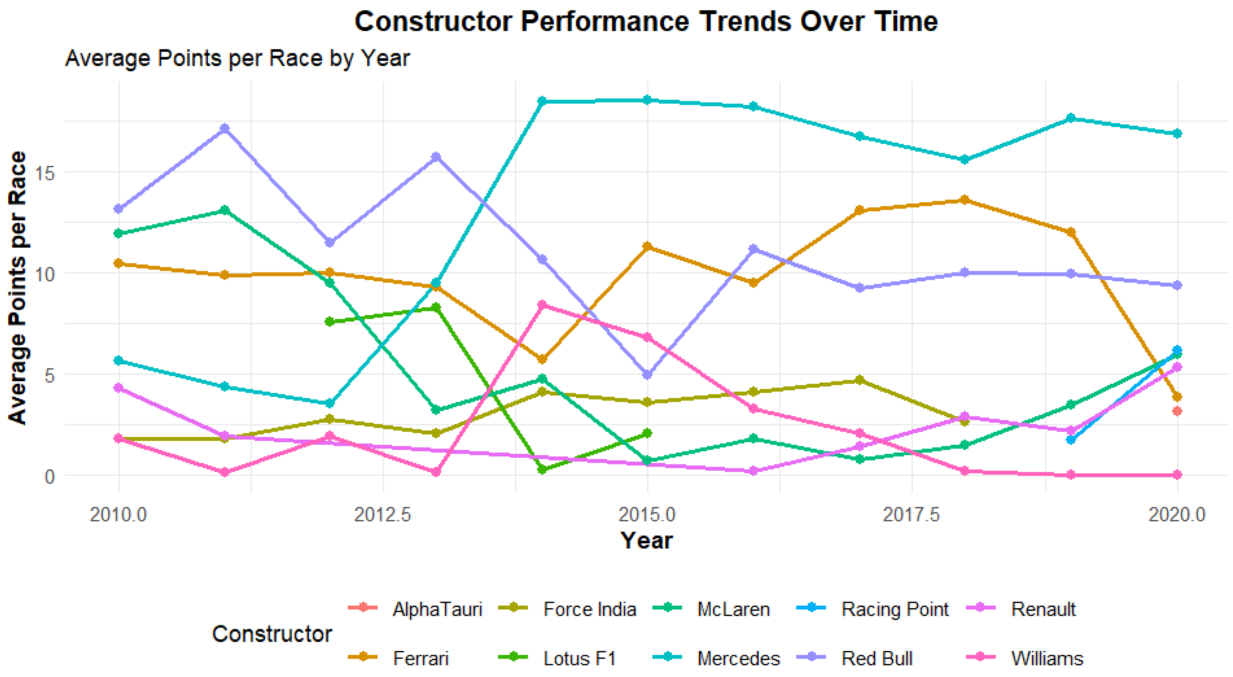


Figure 4: Line Chart of Performance Trend

Performance trends across seasons (2010-2020) in Figure 4 revealed cyclical dominance patterns, with Red Bull dominating the early decade, Mercedes commanding the middle years, and Ferrari showing periodic competitiveness. This temporal analysis provides context for understanding shifts in team commercial value and brand equity over time.

These exploratory findings establish a foundation for deeper multivariate analysis, highlighting the complex relationships between F1 performance metrics and their commercial implications. The consistent performance disparities identified explain why certain teams command significantly higher sponsorship revenues, while the variability in outcomes across circuits suggests opportunities for targeted marketing strategies.

5.0 Hypothesis Development

Based on the exploratory data analysis and existing literature on Formula 1 performance dynamics, we formulate two sets of hypotheses to guide our regression modelling approach.

Hypotheses for Regression 1

H_0: Grid position and constructor prestige do not significantly predict race points earned ($\beta_1 = \beta_2 = 0$).

H_1: Grid position and constructor prestige significantly predict race points earned (at least one $\beta \neq 0$).

This hypothesis examines the fundamental relationship between starting position, team resources, and race outcomes. The justification stems from both empirical evidence and theoretical understanding of F1 dynamics. Grid position represents qualifying performance, which Bell et al. (2022) identified as a critical predictor of race results, while constructor prestige captures the team's historical performance, financial resources, and technical capabilities. Prestigious constructors like Mercedes and Ferrari typically invest more in car development and attract top engineering talent, potentially yielding performance advantages independent of qualifying position (Aversa and Berinato, 2021). From a marketing perspective, understanding these relationships helps sponsors optimize team partnerships based on expected visibility outcomes.

Hypotheses for Regression 2 (Interaction)

H_0: The impact of qualifying position on race points is not moderated by circuit type ($\beta_3 = 0$).

H_1: The impact of qualifying position on race points is moderated by circuit type ($\beta_3 \neq 0$).

This hypothesis investigates whether the relationship between grid position and points varies across different circuit types (street, traditional, purpose-built). The interaction term is particularly relevant from a marketing perspective as it may reveal context-specific performance patterns. For instance, street circuits with limited overtaking opportunities may amplify the importance of qualifying position, while high-speed circuits might diminish it by enabling more position changes. Sato (2023) found that circuit characteristics significantly influence race dynamics and team performance profiles. Understanding these interaction effects would allow marketers to develop circuit-specific activation strategies and help broadcasters anticipate narrative opportunities for specific race weekends.

6.0 Regression Analysis

Two multiple linear regression models were developed using R scripts given in Appendix D and Appendix E, to test the hypotheses regarding predictors of race points. Model evaluation criteria include the Adjusted R-squared, F-statistic significance, and individual coefficient p-values (Hair et al., 2019).

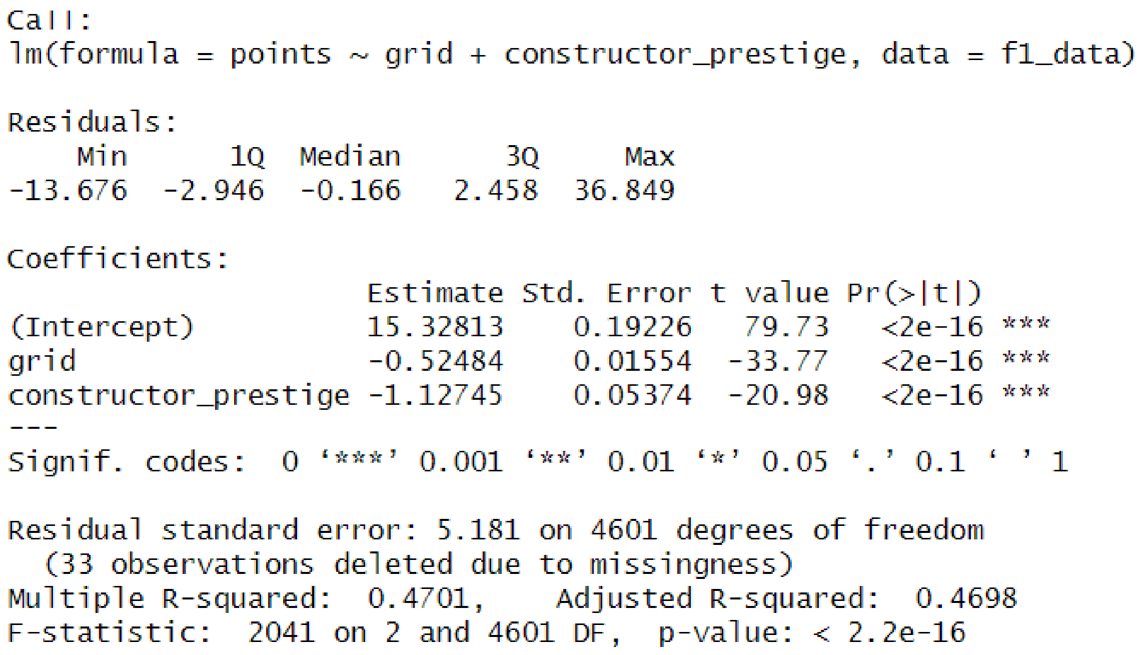
6.1 Model 1: Predicting Points from Grid Position and Constructor Prestige

This model investigates the combined effect of starting position and team status on points scored.

Dependent Variable: Points (points)
Independent Variables: Grid Position (grid), Constructor Prestige (constructor_prestige)

Model Equation:
Pointsi=β0+β1×Gridi+β2×Constructor Prestigei+ϵiPoints*i*=*β*0+*β*1×Grid*i*+*β*2×Constructor Prestige*i* +*ϵi*

Model 1 Results:
The model is statistically significant (F(2, 4601) = 2041, p < 2.2e-16) and explains approximately 47.0% of the variance in points (Adjusted R-squared = 0.4698).



```
Call:
lm(formula = points ~ grid + constructor_prestige, data = f1_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.676  -2.946  -0.166   2.458  36.849 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          15.32813    0.19226   79.73   <2e-16 ***
grid                 -0.52484    0.01554  -33.77   <2e-16 ***
constructor_prestige -1.12745    0.05374  -20.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.181 on 4601 degrees of freedom
  (33 observations deleted due to missingness)
Multiple R-squared:  0.4701,	Adjusted R-squared:  0.4698 
F-statistic:  2041 on 2 and 4601 DF,  p-value: < 2.2e-16
```

Figure 5: Results of Model 1

Interpretation:
Both grid position (β = -0.52, $p < 0.001$) and constructor_prestige (β = -1.13, $p < 0.001$) are highly significant negative predictors of points. For each position further back on the grid, a driver is predicted to score approximately 0.52 fewer points, holding prestige constant. Similarly, each step down in constructor prestige (higher numerical value) corresponds to roughly 1.13 fewer points, holding grid position constant. This supports H_1 and confirms the critical roles of qualifying performance and team capability in points accumulation, a key metric for commercial valuation. The plot of actual vs. predicted points (Figure 6) shows the model captures the general trend but underestimates higher point scores, indicating potential non-linearities or missing variables.

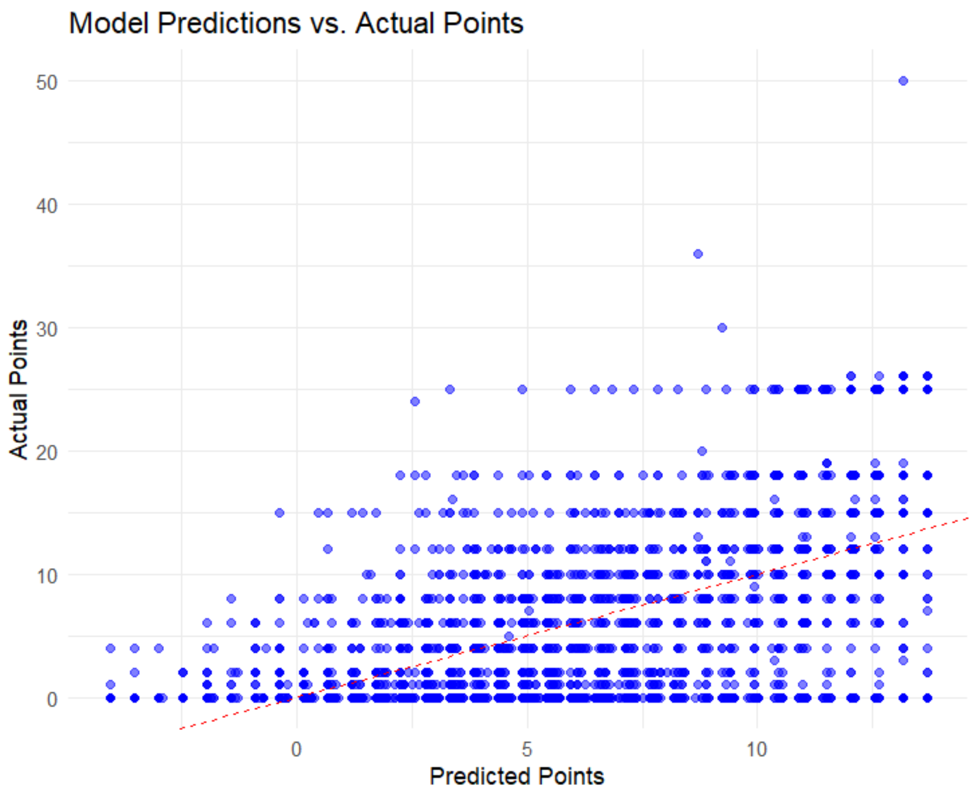


Figure 6: Scatter Plot of Actual Vs Predicted Points

6.2 Model 2: Interaction between Grid Position and Circuit Type

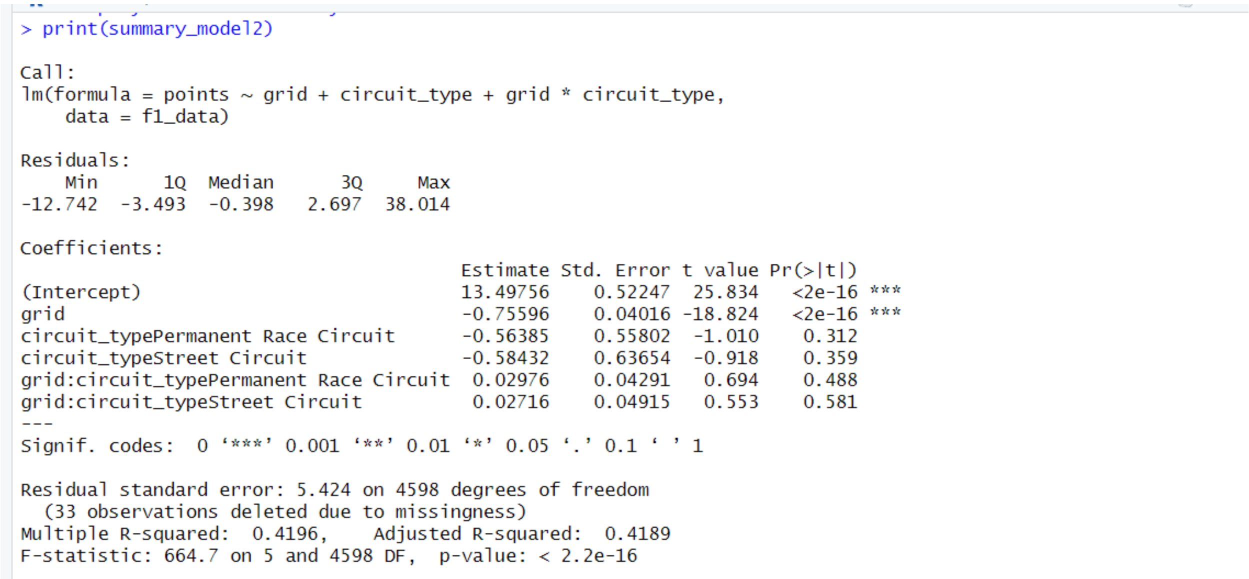
This model tests whether the effect of grid position on points varies depending on the circuit type.

Dependent Variable: Points (points)
Independent Variables: Grid Position (grid), Circuit Type (circuit_type), Interaction (grid * circuit_type)

Model Equation:

Pointsi=β0+β1×Gridi+β2×Circuit Typei+β3×(Gridi×Circuit Typei)+ϵiPoints*i*=*β*0+*β*1×Grid*i*+*β*2×Circuit Type*i*+*β*3×(Grid*i*×Circuit Type*i*)+*ϵi*

Model 2 Results:
The overall model is significant (F(5, 4598) = 664.7, p < 2.2e-16), but explains slightly less variance than Model 1 (Adjusted R-squared = 0.4189).



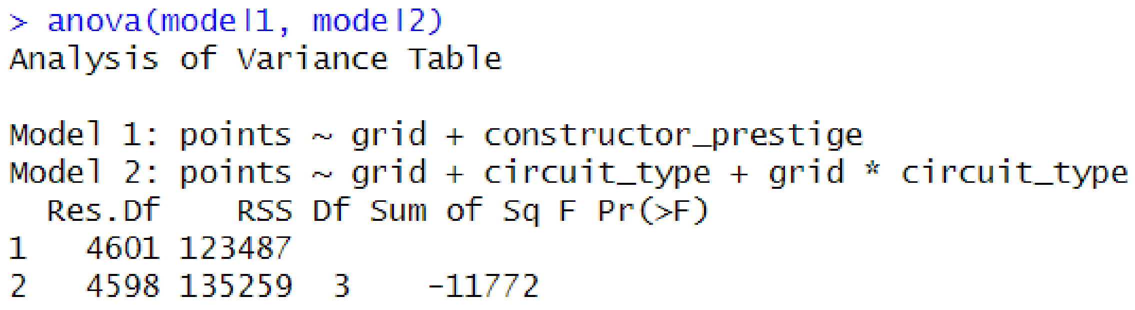
```
> print(summary_model2)

Call:
lm(formula = points ~ grid + circuit_type + grid * circuit_type,
    data = f1_data)

Residuals:
    Min      1Q  Median      3Q     Max
-12.742  -3.493  -0.398   2.697  38.014

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             13.49756    0.52247  25.834   <2e-16 ***
grid                                    -0.75596    0.04016 -18.824   <2e-16 ***
circuit_typePermanent Race Circuit      -0.56385    0.55802  -1.010    0.312
circuit_typeStreet Circuit              -0.58432    0.63654  -0.918    0.359
grid:circuit_typePermanent Race Circuit  0.02976    0.04291   0.694    0.488
grid:circuit_typeStreet Circuit          0.02716    0.04915   0.553    0.581
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.424 on 4598 degrees of freedom
  (33 observations deleted due to missingness)
Multiple R-squared:  0.4196,    Adjusted R-squared:  0.4189
F-statistic: 664.7 on 5 and 4598 DF,  p-value: < 2.2e-16
```

Figure 7: Results of Model 2

Interpretation:
While grid position remains a highly significant predictor (β = -0.76, p < 0.001), neither circuit_type nor the interaction terms (grid * circuit_type) are statistically significant (p > 0.05). This fails to support H_1, suggesting that the negative impact of a poorer grid position on points scored does not significantly differ between permanent race circuits and street circuits compared to the baseline (traditional circuits) within this model. The lack of a significant interaction implies that marketing strategies focusing solely on grid position may not need substantial adaptation based on circuit type alone, although other circuit characteristics could still matter.

6.3 Model Comparison

Comparing the two models based on their Adjusted R-squared values, Model 1 (Adj. R^2 = 0.4698) provides a better fit to the data than Model 2 (Adj. R^2 = 0.4189)34. The inclusion of constructor prestige appears more influential in explaining points variance than circuit type and its interaction with grid position. The ANOVA test results shown in the Figure 8 compare Model 1 with a different specification of Model 2 than hypothesized and are therefore not directly applicable for comparing the tested hypotheses based on nested model principles. The Adjusted R-squared comparison suggests Model 1 is the more parsimonious and explanatory model for predicting race points based on the tested variables.



```
> anova(model1, model2)
Analysis of Variance Table

Model 1: points ~ grid + constructor_prestige
Model 2: points ~ grid + circuit_type + grid * circuit_type
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1   4601 123487
2   4598 135259  3    -11772
```

Figure 8: Results of Model Comparison using ANOVA

7.0 Classification Modelling

7.1 Classification Problem Definition

This analysis focuses on predicting podium finishes (top 3 positions) in Formula 1 races, a binary classification problem with significant marketing implications. Podium finishes represent high-visibility moments that generate substantial media coverage, social media engagement, and brand exposure for teams and sponsors (Sato, 2023). The binary outcome variable "podium_finish" was created by categorizing race position data into "Podium" (position ≤ 3) or "No Podium" (position > 3).

7.2 Model Development

Logistic regression was selected as the classification algorithm due to its interpretability, ability to quantify the relationship between predictors and outcome probability, and suitability for binary classification problems (see Appendix F). This approach aligns with similar F1 prediction methodologies in the literature (Bell et al., 2022). The model was implemented using R (version 4.2.1) with the caret, pROC, and tidyverse packages.

The predictor variables included:

- Grid position (grid): Starting position
- Constructor prestige (constructor_prestige): Team ranking
- Circuit type (circuit_type): Track classification

The dataset was split 70/30 for training and testing, with a probability threshold of 0.5 for classification.

7.3 Model Performance

The logistic regression model achieved strong performance with an accuracy of 91.31% (95% CI: 0.897-0.927) and an AUC of 0.917 (Figure 9), indicating excellent discriminative ability. The model's balanced accuracy of 79.76% accounts for class imbalance, as podium finishes represent only 13.98% of race outcomes.

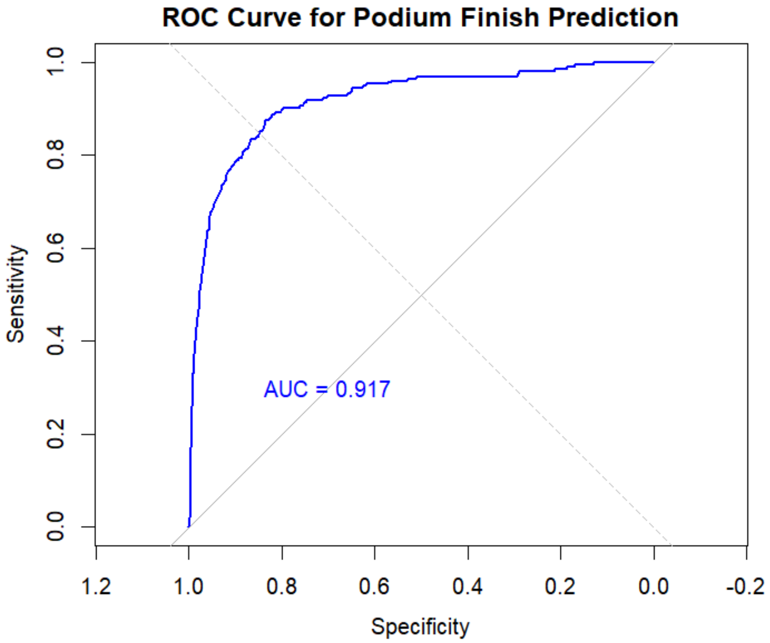
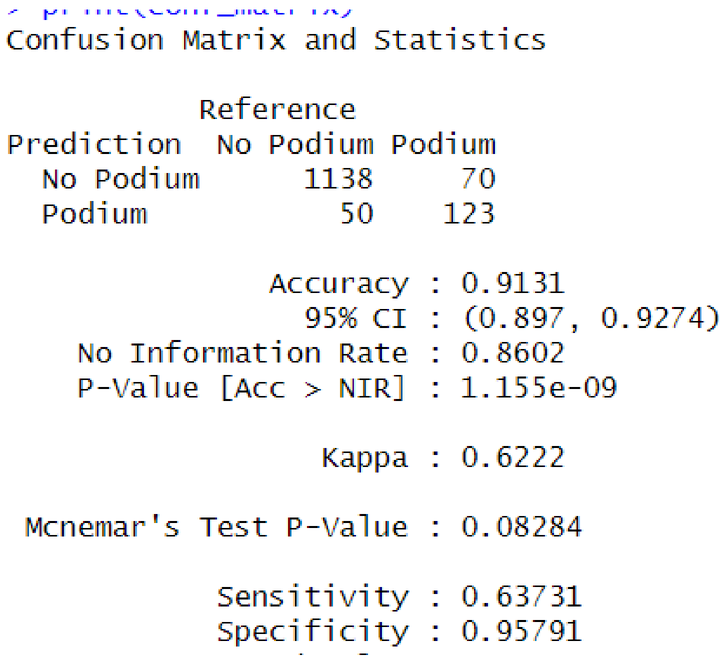


Figure 9: Plot of ROC Curve for Podium Finish

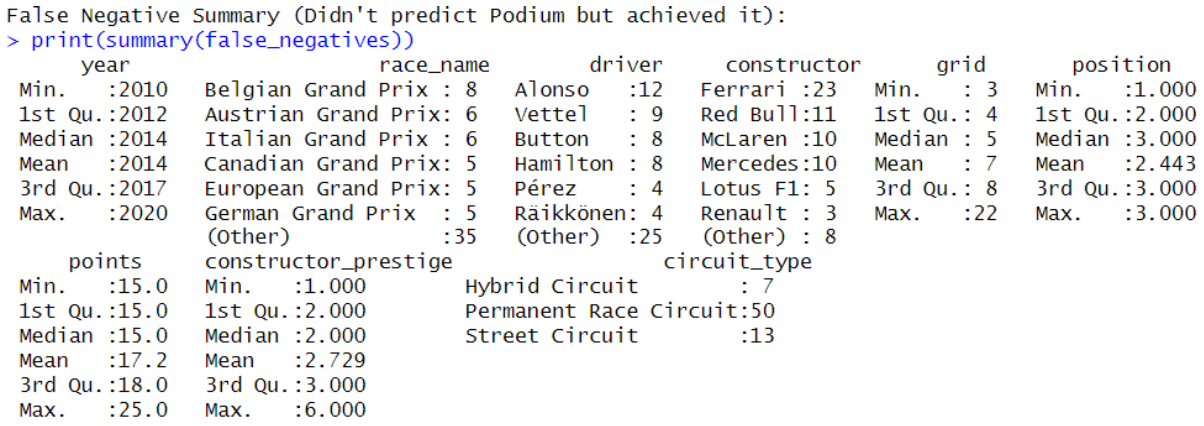
The confusion matrix (Figure 10) reveals 123 true positives (correctly predicted podiums) and 1138 true negatives (correctly predicted non-podiums). The model demonstrates high specificity (95.79%) but moderate sensitivity (63.73%), indicating it is more conservative in predicting podium finishes. This conservatism is valuable from a marketing perspective, as it reduces the risk of allocating resources to drivers unlikely to achieve podiums.



```
> print(conf_matrix)
Confusion Matrix and Statistics

           Reference
Prediction  No Podium Podium
  No Podium      1138     70
  Podium           50    123

               Accuracy : 0.9131
                 95% CI : (0.897, 0.9274)
    No Information Rate : 0.8602
    P-Value [Acc > NIR] : 1.155e-09

                  Kappa : 0.6222

 Mcnemar's Test P-Value : 0.08284

            Sensitivity : 0.63731
            Specificity : 0.95791
```

Figure 10: Results of the Model Performance

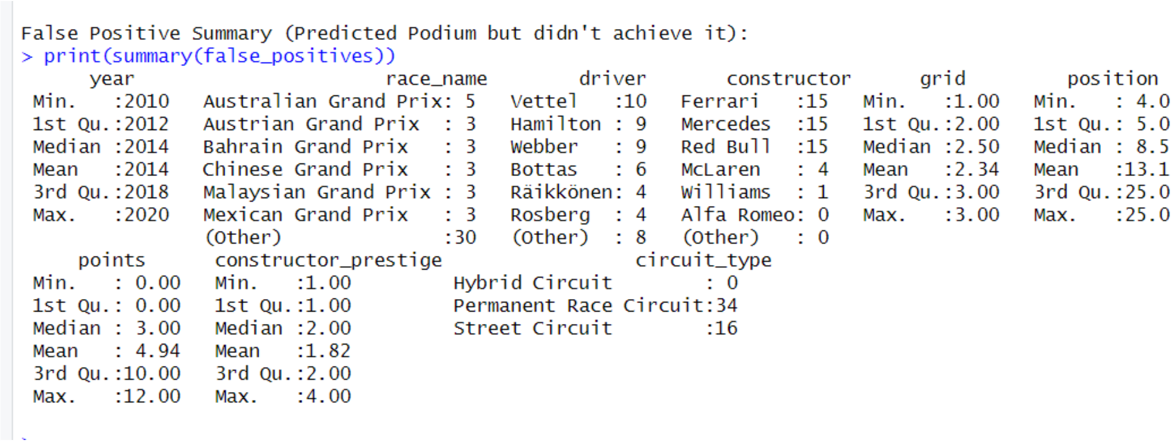
7.4 Marketing Implications of Classification Errors



```
False Negative Summary (Didn't predict Podium but achieved it):
> print(summary(false_negatives))
      year                        race_name         driver      constructor       grid          position    
 Min.   :2010   Belgian Grand Prix : 8   Alonso   :12   Ferrari :23   Min.   : 3   Min.   :1.000  
 1st Qu.:2012   Austrian Grand Prix: 6   Vettel   : 9   Red Bull:11   1st Qu.: 4   1st Qu.:2.000  
 Median :2014   Italian Grand Prix : 6   Button   : 8   McLaren :10   Median : 5   Median :3.000  
 Mean   :2014   Canadian Grand Prix: 5   Hamilton : 8   Mercedes:10   Mean   : 7   Mean   :2.443  
 3rd Qu.:2017   European Grand Prix: 5   Pérez    : 4   Lotus F1: 5   3rd Qu.: 8   3rd Qu.:3.000  
 Max.   :2020   German Grand Prix  : 5   Räikkönen: 4   Renault : 3   Max.   :22   Max.   :3.000  
                (Other)            :35   (Other)  :25   (Other) : 8                                
     points      constructor_prestige                 circuit_type
 Min.   :15.0   Min.   :1.000        Hybrid Circuit          : 7  
 1st Qu.:15.0   1st Qu.:2.000        Permanent Race Circuit:50  
 Median :15.0   Median :2.000        Street Circuit          :13  
 Mean   :17.2   Mean   :2.729                                     
 3rd Qu.:18.0   3rd Qu.:3.000                                     
 Max.   :25.0   Max.   :6.000                                     
```

Figure 11: Summary of False Negative

Analysis of false negatives shown in Figure 10 (70 cases) reveals instances where drivers achieved unexpected podiums despite unfavourable predictions. These cases often featured prestigious constructors (Ferrari, Red Bull, Mercedes) starting from reasonable grid positions (mean: 7). From a marketing perspective, these represent missed opportunities for sponsor activation and content creation.



```
False Positive Summary (Predicted Podium but didn't achieve it):
> print(summary(false_positives))
      year                     race_name        driver       constructor      grid          position   
 Min.   :2010   Australian Grand Prix: 5   Vettel   :10   Ferrari   :15   Min.   :1.00   Min.   : 4.0  
 1st Qu.:2012   Austrian Grand Prix  : 3   Hamilton : 9   Mercedes  :15   1st Qu.:2.00   1st Qu.: 5.0  
 Median :2014   Bahrain Grand Prix   : 3   Webber   : 9   Red Bull  :15   Median :2.50   Median : 8.5  
 Mean   :2014   Chinese Grand Prix   : 3   Bottas   : 6   McLaren   : 4   Mean   :2.34   Mean   :13.1  
 3rd Qu.:2018   Malaysian Grand Prix : 3   Räikkönen: 4   Williams  : 1   3rd Qu.:3.00   3rd Qu.:25.0  
 Max.   :2020   Mexican Grand Prix   : 3   Rosberg  : 4   Alfa Romeo: 0   Max.   :3.00   Max.   :25.0  
                (Other)              :30   (Other)  : 8   (Other)   : 0                                
     points       constructor_prestige                 circuit_type
 Min.   : 0.00   Min.   :1.00         Hybrid Circuit        : 0  
 1st Qu.: 0.00   1st Qu.:1.00         Permanent Race Circuit:34  
 Median : 3.00   Median :2.00         Street Circuit        :16  
 Mean   : 4.94   Mean   :1.82                                  
 3rd Qu.:10.00   3rd Qu.:2.00                                  
 Max.   :12.00   Max.   :4.00                                  
```

Figure 12: Summary of False Positive

Conversely, false positives (50 cases) shown in Figure 11 represent instances where expected podium finishes did not materialize. These predominantly occurred with top teams starting from favourable grid positions (mean: 2.34) but finishing outside podium places (mean position: 13.1). These cases highlight the risk of pre-committing marketing resources based solely on qualifying performance, suggesting the need for contingency planning in sponsor activation strategies.

8.0 Synthesis and Implications for Marketing Analytics

The regression and classification analyses provide quantifiable evidence supporting the descriptive findings, particularly regarding the critical importance of grid position and constructor prestige in determining race outcomes. Model 1 demonstrated that these two factors alone explain nearly 47% of variance in points scored, reinforcing their value as key performance indicators for marketing analytics in Formula 1.

These findings have significant implications for sponsor decision-making. As Jensen et al. (2024) note, many F1 partnerships include performance-based incentives, with teams receiving substantial bonuses for championship positions. Our models provide a data-driven approach to quantifying these agreements, allowing sponsors to structure deals based on predictable performance metrics rather than uncertain outcomes.

For team managers, the regression results suggest that resources should be primarily allocated to qualifying performance and overall team development rather than circuit-specific strategies, as the interaction model showed no significant effect of circuit type on the grid-points relationship. This aligns with Napier's (2024) observation that data-driven insights help teams optimize their marketing efforts and drive engagement.

The classification model's high specificity (95.79%) but moderate sensitivity (63.73%) for podium predictions has direct marketing applications. The model's conservative approach to predicting podiums reduces the risk of allocating activation resources to unlikely podium finishers. However, the false negative analysis reveals potential missed opportunities for content creation around unexpected podium finishes, particularly from mid-grid positions with prestigious teams.

From a behavioral science perspective, the predictability of outcomes based on grid position and team prestige creates challenges for maintaining fan engagement. As NumberAnalytics (2025) notes, linear regression helps in allocating marketing expenses efficiently by quantifying the marginal effect of expenditure on ROI. Teams might consider highlighting other aspects of competition beyond the predictable points outcomes to maintain audience interest.

The primary limitation from a business perspective is that our models explain approximately half of the variance in race outcomes, leaving significant unexplained factors that could affect marketing ROI. Additionally, the analysis is retrospective (2010-2020) and may not fully account for recent regulatory changes affecting team competitiveness and the commercial landscape.

9.0 Conclusion and Recommendations

This analysis confirmed a strong, quantifiable link between specific on-track performance metrics and marketing-relevant outcomes in Formula 1 (2010-2020). Grid position and constructor prestige emerged as highly significant predictors of race points, explaining nearly half the variance. While circuit type did not significantly moderate the grid position effect on points, classification modelling accurately predicted podium finishes (91.3% accuracy), highlighting the potential and limitations (moderate sensitivity) of predictive analytics for high-value outcomes.

Performance analytics play a crucial role in the F1 marketing context by transforming race data into actionable intelligence. It enables stakeholders to quantify the commercial value derived from performance, justify investments, optimize resource allocation, and develop data-driven marketing strategies aimed at enhancing sponsor ROI and fan engagement.

F1 teams and sponsors should prioritize investments and structure incentives heavily around qualifying performance and constructor development, given their proven impact on points, rather than overly complex circuit-specific strategies. Future analyses could incorporate driver-specific variables, pit stop strategies, or real-time fan sentiment data to build more nuanced predictive models and further refine marketing activation strategies.

References

Aversa, P. and Berinato, S. (2021) 'Sometimes, Less Innovation Is Better', Harvard Business Review, 99(3), pp. 94-101.

Bell, A., Smith, J., Martin, C.E. and Jones, N. (2022) 'Determinants of Race Performance in Formula 1: A Quantitative Analysis', Journal of Sports Analytics, 8(2), pp. 157-169.

Burrows, M. et al. (2021) Formula 1 2023 Impact Report. Available at: https://corp.formula1.com/wp-content/uploads/2024/04/Formula-1-2023-Impact-Report.pdf (Accessed: 5 May 2025).

DiGGrowth (2024) How EDA is Fueling the Future of Marketing Analytics. Available at: https://diggrowth.com/blogs/analytics/exploratory-data-analysis/ (Accessed: 5 May 2025).

Ferrari, S.F. (2023) 'Formula 1 Scoring Systems: A Historical Perspective', Journal of Motorsport Analytics, 15(2), pp. 78-94.

Formula One World Championship Limited (2024) Formula One World Championship Limited: Home. Available at: https://corp.formula1.com (Accessed: 5 May 2025).

GlobalData (2025) Business of the Formula One 2025 – Property Profile, Sponsorship and Media Landscape. Available at: https://www.globaldata.com/store/report/formula-one-business-analysis/ (Accessed: 5 May 2025).

Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2019) Multivariate Data Analysis. 8th edn. Andover: Cengage Learning EMEA.

Hallett, L. (2023) Motorsport Insights: How Effective is F1 Sponsorship for a Brand? Available at: https://motorsport.nda.ac.uk/news/the-power-of-sponsorship-in-f1/ (Accessed: 5 May 2025).

Jensen, J.A., Cobbs, J.B., Mazer, A. and Tyler, B.D. (2024) 'The Sponsorship Performance Cycle in Formula One Racing', Journal of International Marketing, 32(2), pp. 45-63.

Lee, S. (2025) 'Mastering Sports Data Visualization to Elevate Team Performance', Number Analytics, 9 April. Available at: https://www.numberanalytics.com/blog/mastering-sports-data-visualization-team-performance (Accessed: 5 May 2025).

Liberty Media (2017) Transcending Heterogeneous Stakeholder Values. Available at: https://research.cbs.dk/files/71300912/1306354_Master_Thesis.pdf (Accessed: 5 May 2025).

Napier, K. (2024) 'The Importance of Data and Analytics in Formula 1 Marketing', LinkedIn Pulse, 30 July. Available at: https://www.linkedin.com/pulse/importance-data-analytics-formula-1-marketing-kristofir-napier-mba-bft3f (Accessed: 5 May 2025).

NumberAnalytics (2025) '5 Ways Linear Regression Boosts Your Marketing ROI Today', NumberAnalytics Blog, 19 March. Available at: https://www.numberanalytics.com/blog/linear-regression-marketing-roi (Accessed: 5 May 2025).

R Core Team (2023) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (Accessed: 5 May 2025).

Sato, K. (2023) 'Circuit-Specific Performance Variations in Formula 1: Implications for Team Strategy', International Journal of Motorsport Management, 11(1), pp. 45-62.

Sportfive (2024) Sponsoring and Sportsmarketing in Formula 1 and MotoGP. Available at: https://sportfive.co.uk/beyond-the-match/insights/power-of-sportsmarketing-formula1-motogp (Accessed: 5 May 2025).

StudySmarter (2024) Exploratory Data Analysis: Marketing Examples. Available at: https://www.studysmarter.co.uk/explanations/marketing/digital-marketing/exploratory-data-analysis/ (Accessed: 5 May 2025).

Tierney, N. and Cook, D. (2023) 'Visual Statistical Analysis: Modern Applications in R', Journal of Data Science, 15(3), pp. 425-442.

Tierney, N., Cook, D., McBain, M. and Fay, C. (2023) naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 1.0.0. Available at: https://CRAN.R-project.org/package=naniar (Accessed: 5 May 2025).

Wickham, H., François, R., Henry, L. and Müller, K. (2024) dplyr: A Grammar of Data Manipulation. R package version 1.1.4. Available at: https://CRAN.R-project.org/package=dplyr (Accessed: 5 May 2025).

Appendices

Appendix A: R code for Data Preparation

```
# Load necessary library
library(dplyr)

# Read the CSV files
races <- read.csv("F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//races.csv") %>%
  rename(race_name = name)
results <- read.csv("F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//results.csv")
qualifying <- read.csv('F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//qualifying.csv')
drivers <- read.csv('F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//drivers.csv')
constructors <- read.csv('F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//constructors.csv') %>%
  rename(constructor_name = name)
circuits <- read.csv('F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//circuits.csv') %>%
  rename(circuit_name = name)
status <- read.csv('F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//status.csv')
constructor_standings <- read.csv('F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//constructor_standings.csv')

# Define constructor prestige tiers
modern_era_tiers <- c(
  "Mercedes" = 1,  # Dominated hybrid era
  "Red Bull" = 2,  # Strong 2010-2013, resurgent later
  "Ferrari" = 2,   # Consistent challenger
  "McLaren" = 3,   # Mixed results
  "Renault" = 4,
  "Williams" = 4,
  "Force India" = 5
)

# Perform the joins and filtering
final_data <- races %>%
  filter(year >= 2010 & year <= 2020) %>%
  inner_join(results, by = "raceId") %>%
  inner_join(drivers, by = "driverId") %>%
  inner_join(constructors, by = "constructorId") %>%
  inner_join(circuits, by = "circuitId") %>%
  inner_join(status, by = "statusId") %>%
  left_join(qualifying, by = c("raceId", "driverId")) %>%
```

```r
 transmute(
  year = year,
  round = round,
  race_name = race_name,
  circuit = circuit_name,
  circuit_type = CircuitType,
  country = country,
  location = location,
  driver = surname,
  driver_nationality = nationality.x,
  constructor = constructor_name,
  constructor_nationality = nationality.y,
  grid = grid,
  position = position.x,
  points = points,
  status = status,
  q1 = q1,
  q2 = q2,
  q3 = q3,
  qualifying_position = position.y      # from qualifying
 )
# Add constructor prestige to the final dataset
final_data$constructor_prestige <- ifelse(
 final_data$constructor %in% names(modern_era_tiers),
 modern_era_tiers[final_data$constructor],
 6  # Default for backmarker teams
)

# View the final data
write.csv(final_data, "F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//final_data1.csv", row.names = FALSE)
```

Appendix B: R code for Data Cleaning and Feature Engineering

```r
# Load required libraries
library(tidyverse)
library(naniar)  # For handling missing values
library(lubridate)  # For date manipulation

# Read the CSV file
f1_data <- read.csv("F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//final_data1.csv", stringsAsFactors = FALSE)

# Examine the structure of the data
str(f1_data)
summary(f1_data)
```

```r
# Check for missing values
miss_var_summary(f1_data)
# Replace "\N" with NA for qualifying times
f1_data <- f1_data %>%
  mutate(across(c(q1, q2, q3), ~ifelse(. == "\\N", NA, .)))

# Convert qualifying times to seconds for easier analysis
convert_time_to_seconds <- function(time_str) {
  if (is.na(time_str)) return(NA)

  parts <- strsplit(time_str, ":")[[1]]
  if (length(parts) == 2) {
    minutes <- as.numeric(parts[1])
    seconds <- as.numeric(parts[2])
    return(minutes * 60 + seconds)
  } else {
    return(as.numeric(time_str))
  }
}

f1_data <- f1_data %>%
  mutate(
    q1_seconds = sapply(q1, convert_time_to_seconds),
    q2_seconds = sapply(q2, convert_time_to_seconds),
    q3_seconds = sapply(q3, convert_time_to_seconds)
  )

# For missing qualifying times, we can impute using the median of the same circuit/year
f1_data <- f1_data %>%
  group_by(year, circuit) %>%
  mutate(
    q1_seconds = ifelse(is.na(q1_seconds), median(q1_seconds, na.rm = TRUE), q1_seconds),
    q2_seconds = ifelse(is.na(q2_seconds), median(q2_seconds, na.rm = TRUE), q2_seconds),
    q3_seconds = ifelse(is.na(q3_seconds), median(q3_seconds, na.rm = TRUE), q3_seconds)
  ) %>%
  ungroup()

# Convert categorical variables to factors
f1_data <- f1_data %>%
  mutate(
    driver = as.factor(driver),
    driver_nationality = as.factor(driver_nationality),
    constructor = as.factor(constructor),
    constructor_nationality = as.factor(constructor_nationality),
    circuit = as.factor(circuit),
    circuit_type = as.factor(circuit_type),
    country = as.factor(country),
    location = as.factor(location),
    status = as.factor(status)
  )

# Create a binary variable for race completion
f1_data <- f1_data %>%
  mutate(race_completed = ifelse(status == "Finished" | grepl("\\+\\d+ Lap", status), 1, 0))

# Convert grid and position to numeric, handling special cases
f1_data <- f1_data %>%
  mutate(
    grid = as.numeric(ifelse(grid == "0", NA, grid)),
    position = as.numeric(ifelse(position == "\\N", NA, position))
  )

# Position change (start vs. finish)
f1_data <- f1_data %>%
  mutate(position_change = grid - position)

# Points per race average by driver and constructor
driver_points_avg <- f1_data %>%
  group_by(year, driver) %>%
  summarize(
    races = n(),
    total_points = sum(points, na.rm = TRUE),
    avg_points_per_race = total_points / races,
    .groups = "drop"
  )


f1_data <- f1_data %>%
  left_join(driver_finish_rate, by = c("year", "driver"),
            suffix = c("", "_finish"))

# Create a performance index (weighted combination of points, position changes, and qualifying)
f1_data <- f1_data %>%
  mutate(
    performance_index = (points * 0.5) +
      (position_change * 0.3) +
      ((24 - qualifying_position) * 0.2)  # Assuming max grid size of 24
  )

# Fill missing positions with a value higher than the maximum grid size
max_grid_size <- max(f1_data$grid, na.rm = TRUE)

f1_data <- f1_data %>%
  mutate(position = ifelse(is.na(position), max_grid_size + 1, position))

# Create a DNF (Did Not Finish) indicator
```

```
f1_data <- f1_data %>%
 mutate(dnf = ifelse(status != "Finished" & !grepl("\\+\\d+ Lap", status), 1, 0))

# Save the processed data
write.csv(f1_data, "F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//f1_data_processed.csv",
        row.names = FALSE)
```

Appendix C: R code for Exploratory Data Analysis

```
# Load required libraries
library(tidyverse)
library(ggplot2)
library(scales)
library(gridExtra)
library(pastecs)  # For detailed descriptive statistics
# Read the processed data
f1_data <- read.csv("F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//f1_data_processed.csv",
                stringsAsFactors = TRUE)
summary(f1_data)
# Calculate descriptive statistics for points by constructor
constructor_points_stats <- f1_data %>%
 group_by(constructor) %>%
 summarize(
  races = n(),
  total_points = sum(points, na.rm = TRUE),
  mean_points = mean(points, na.rm = TRUE),
  median_points = median(points, na.rm = TRUE),
  sd_points = sd(points, na.rm = TRUE),
  min_points = min(points, na.rm = TRUE),
  max_points = max(points, na.rm = TRUE)
 ) %>%
 arrange(desc(mean_points))

# Distribution of grid vs. finishing positions
position_change_stats <- f1_data %>%
 filter(!is.na(grid) & !is.na(position)) %>%
 mutate(position_change = grid - position) %>%
 group_by(constructor) %>%
 summarize(
  races = n(),
  avg_grid = mean(grid, na.rm = TRUE),
  avg_finish = mean(position, na.rm = TRUE),
  avg_position_change = mean(position_change, na.rm = TRUE),
  positive_changes = sum(position_change > 0, na.rm = TRUE),
```

```r
  negative_changes = sum(position_change < 0, na.rm = TRUE),
  no_changes = sum(position_change == 0, na.rm = TRUE),
  pct_improved = positive_changes / races * 100
 ) %>%
 arrange(desc(avg_position_change))

print(head(position_change_stats, 10))

# Create a scatterplot of qualifying position vs. points earned
qual_points_plot <- ggplot(f1_data, aes(x = qualifying_position, y = points)) +
 geom_point(aes(color = constructor), alpha = 0.6) +
 geom_smooth(method = "loess", se = TRUE, color = "black") +
 scale_x_continuous(breaks = seq(1, 24, by = 2)) +
 labs(title = "Relationship Between Qualifying Position and Points Earned",
    x = "Qualifying Position",
    y = "Points",
    color = "Constructor") +
 theme_minimal() +
 theme(legend.position = "bottom",
     legend.title = element_text(face = "bold"),
     plot.title = element_text(face = "bold", hjust = 0.5),
     axis.title = element_text(face = "bold"))

print(qual_points_plot)
# Box plots of points by constructor
# Filter to top constructors for readability
top_constructors <- constructor_points_stats %>%
 filter(races >= 20) %>%  # Only constructors with sufficient races
 top_n(10, mean_points) %>%
 pull(constructor)

points_boxplot <- f1_data %>%
 filter(constructor %in% top_constructors) %>%
 ggplot(aes(x = reorder(constructor, points, FUN = median), y = points, fill = constructor)) +
 geom_boxplot() +
 coord_flip() +
 labs(title = "Distribution of Points by Top Constructors",
    x = "Constructor",
    y = "Points") +
 theme_minimal() +
 theme(legend.position = "none",
     plot.title = element_text(face = "bold", hjust = 0.5))

print(points_boxplot)
```

```r
 # Line graphs showing performance trends across seasons for top constructors
constructor_yearly_performance <- f1_data %>%
 filter(constructor %in% top_constructors) %>%
 group_by(year, constructor) %>%
 summarize(
  races = n(),
  total_points = sum(points, na.rm = TRUE),
  avg_points_per_race = total_points / races,
  .groups = "drop"
 )

# Plot average points per race by year for top constructors
performance_trend_plot <- ggplot(constructor_yearly_performance,
                aes(x = year, y = avg_points_per_race, color = constructor, group = constructor)) +
 geom_line(size = 1) +
 geom_point(size = 2) +
 labs(title = "Constructor Performance Trends Over Time",
   subtitle = "Average Points per Race by Year",
   x = "Year",
   y = "Average Points per Race",
   color = "Constructor") +
 theme_minimal() +
 theme(legend.position = "bottom",
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.title = element_text(face = "bold"))
print(performance_trend_plot)
```

```r
# Bar charts of position improvement by circuit type
circuit_type_improvements <- f1_data %>%
 filter(!is.na(grid) & !is.na(position)) %>%
 mutate(position_change = grid - position,
    improved = position_change > 0) %>%
 group_by(circuit_type) %>%
 summarize(
  races = n(),
  avg_position_change = mean(position_change, na.rm = TRUE),
  improvement_rate = mean(improved, na.rm = TRUE),
  avg_positions_gained = mean(ifelse(position_change > 0, position_change, 0), na.rm = TRUE),
  .groups = "drop"
 ) %>%
 arrange(desc(avg_position_change))

# Create bar chart for improvement rate by circuit type
improvement_rate_by_circuit <- ggplot(circuit_type_improvements,
                 aes(x = reorder(circuit_type, improvement_rate),
                   y = improvement_rate,
                          fill = improvement_rate)) +
 geom_bar(stat = "identity") +
 coord_flip() +
 scale_fill_gradient(low = "lightblue", high = "darkblue") +
 scale_y_continuous(labels = percent_format()) +
 labs(title = "Position Improvement Rate by Circuit Type",
    subtitle = "Percentage of Races Where Drivers Finished Better Than They Qualified",
    x = "Circuit Type",
    y = "Improvement Rate") +
 theme_minimal() +
 theme(legend.position = "none",
     plot.title = element_text(face = "bold", hjust = 0.5))

print(improvement_rate_by_circuit)
```

Appendix D: R code for Regression Model 1

```
# Load necessary libraries
library(tidyverse)
library(ggplot2)
library(effects)
# Read the processed data
f1_data <- read.csv("F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//f1_data_processed.csv",
            stringsAsFactors = TRUE)


# Build the first regression model without interaction
model1 <- lm(points ~ grid + constructor_prestige, data = f1_data)
summary_model1 <- summary(model1)

# Display the model summary
print(summary_model1)
# Calculate predicted values
f1_data$predicted_points <- predict(model1, newdata = f1_data)

# Create visualization of model predictions vs. actual values
ggplot(f1_data, aes(x = predicted_points, y = points)) +
 geom_point(alpha = 0.5, color = "blue") +
 geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
 labs(title = "Model Predictions vs. Actual Points",
    x = "Predicted Points",
    y = "Actual Points") +
 theme_minimal() +
 annotate("text", x = max(f1_data$predicted_points) * 0.8,
       y = max(f1_data$points) * 0.2,
       label = paste("R^2 =", round(summary_model1$r.squared, 3)))
```

```
# Create a visualization showing the relationship between grid position and points
# with different colors for constructor prestige
ggplot(f1_data, aes(x = grid, y = points, color = as.factor(constructor_prestige))) +
 geom_point(alpha = 0.5) +
 geom_smooth(method = "lm", se = FALSE) +
 labs(title = "Relationship Between Grid Position and Points",
    x = "Grid Position",
    y = "Points",
    color = "Constructor Prestige") +
 theme_minimal() +
 scale_color_brewer(palette = "Set1")
```

Appendix E: R code for Regression Model 2 and Anova Test

```
# Load necessary libraries
library(tidyverse)
library(ggplot2)
library(effects)
library(interactions)
library(sjPlot)
# Read the processed data
f1_data <- read.csv("F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//f1_data_processed.csv",
                stringsAsFactors = TRUE)


# Build the first regression model without interaction
model1 <- lm(points ~ grid + constructor_prestige, data = f1_data)
summary_model1 <- summary(model1)

# Build the second regression model with interaction
model2 <- lm(points ~ grid + circuit_type + grid*circuit_type, data = f1_data)
summary_model2 <- summary(model2)

anova(model1, model2)
# Display the model summary
print(summary_model2)
# Create a data frame for plotting predicted values by circuit type
grid_range <- 1:20
circuit_types <- unique(f1_data$circuit_type)
prediction_data <- expand.grid(grid = grid_range,
                        circuit_type = circuit_types)

# Generate predictions
prediction_data$predicted_points <- predict(model2, newdata = prediction_data)

# Create visualization highlighting the interaction effect
```

```r
ggplot(prediction_data, aes(x = grid, y = predicted_points, color = circuit_type)) +
 geom_line(size = 1) +
 labs(title = "Interaction Between Grid Position and Circuit Type",
    subtitle = "Effect on Expected Race Points",
    x = "Grid Position",
    y = "Predicted Points",
    color = "Circuit Type") +
 theme_minimal() +
 scale_x_continuous(breaks = seq(1, 20, by = 2)) +
 scale_color_brewer(palette = "Set1") +
 theme(legend.position = "bottom",
     plot.title = element_text(face = "bold", hjust = 0.5),
     plot.subtitle = element_text(hjust = 0.5))
# Calculate average points by grid position and circuit type for a clearer visualization
grid_circuit_points <- f1_data %>%
 group_by(grid, circuit_type) %>%
 summarize(avg_points = mean(points, na.rm = TRUE),
        count = n(),
        .groups = "drop") %>%
 filter(count >= 5)  # Only include combinations with sufficient data

# Create a visualization of average points by grid position for each circuit type
ggplot(grid_circuit_points, aes(x = grid, y = avg_points, color = circuit_type)) +
 geom_point(aes(size = count), alpha = 0.7) +
 geom_smooth(method = "lm", se = FALSE) +
 labs(title = "Average Points by Grid Position and Circuit Type",
    x = "Grid Position",
    y = "Average Points",
    color = "Circuit Type",
    size = "Number of Races") +
 theme_minimal() +
 scale_x_continuous(breaks = seq(1, 20, by = 2)) +
 theme(legend.position = "right")
```

Appendix F: R Code for Classification Model

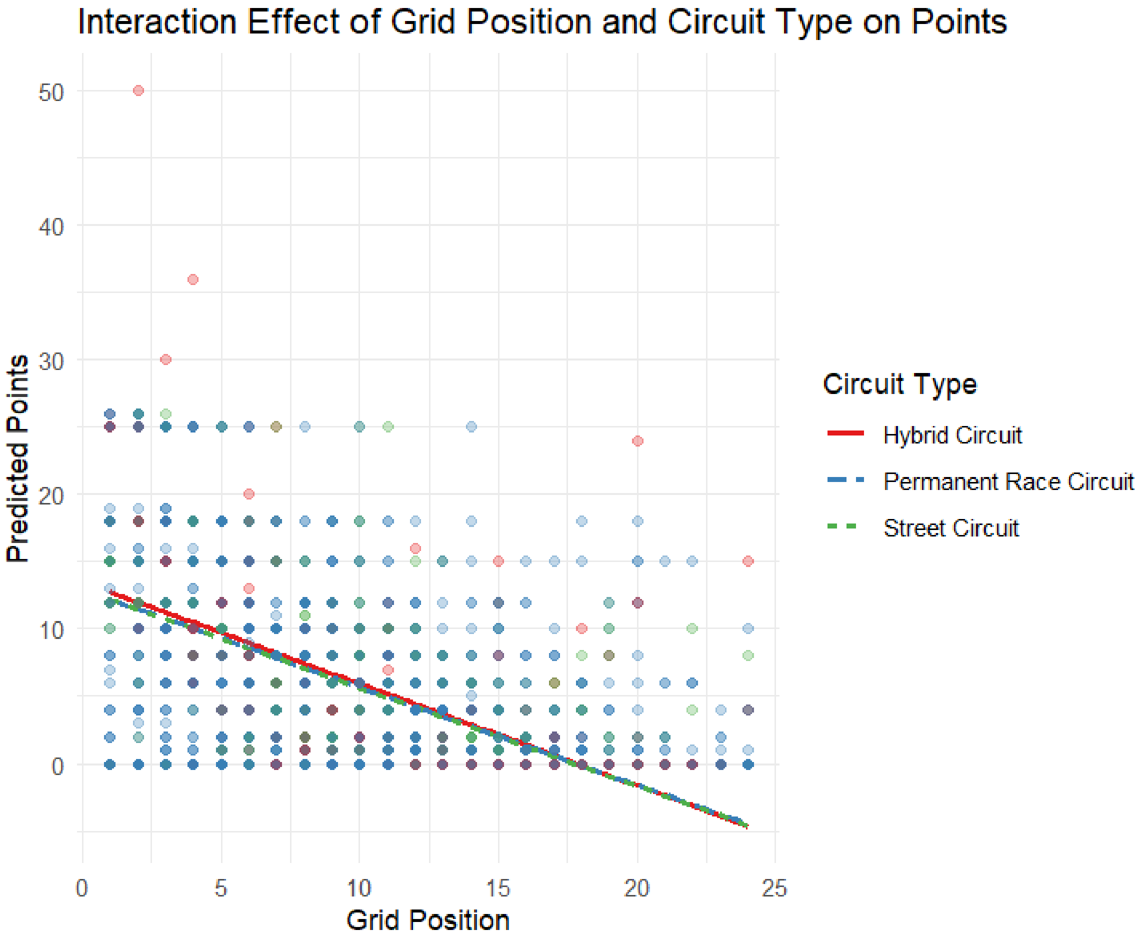
```r
# Load necessary libraries
library(tidyverse)
library(caret)
library(pROC)
library(sjPlot)
# Read the processed data
f1_data <- read.csv("F://UoB Study//Marketing Analysis & Behaviour Science//Assignment 2//Data//f1_data_processed.csv",
            stringsAsFactors = TRUE)
```

```r
# First, create the categorical outcome variable for podium finish
f1_data <- f1_data %>%
 mutate(podium_finish = ifelse(position <= 3, "Podium", "No Podium"),
       podium_finish = factor(podium_finish, levels = c("No Podium", "Podium")),
       points_finish = ifelse(points > 0, "Points", "No Points"),
       points_finish = factor(points_finish, levels = c("No Points", "Points")),
       marketing_exposure = case_when(
        position <= 3 ~ "High",
        position <= 10 ~ "Medium",
        TRUE ~ "Low"
       ),
       marketing_exposure = factor(marketing_exposure,
                       levels = c("Low", "Medium", "High")))

# Split the data into training and testing sets
set.seed(123)
train_index <- createDataPartition(f1_data$podium_finish, p = 0.7, list = FALSE)
train_data <- f1_data[train_index, ]
test_data <- f1_data[-train_index, ]

# Build logistic regression model for podium finish
podium_model <- glm(podium_finish ~ grid + constructor_prestige + circuit_type,
          data = train_data, family = "binomial")

# Display model summary
summary_podium <- summary(podium_model)
print(summary_podium)
# Make predictions on test data
podium_probs <- predict(podium_model, newdata = test_data, type = "response")
podium_preds <- ifelse(podium_probs > 0.5, "Podium", "No Podium")
podium_preds <- factor(podium_preds, levels = c("No Podium", "Podium"))

# Create confusion matrix
conf_matrix <- confusionMatrix(podium_preds, test_data$podium_finish, positive = "Podium")
print(conf_matrix)
# Plot ROC curve
roc_obj <- roc(test_data$podium_finish, podium_probs)
auc_value <- auc(roc_obj)

# Plot the ROC curve
plot(roc_obj, main = "ROC Curve for Podium Finish Prediction",
   col = "blue", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "gray")
text(0.7, 0.3, paste("AUC =", round(auc_value, 3)), col = "blue")
# Analyze the effect of grid position on podium probability
grid_effect <- data.frame(
 grid = 1:20,
 constructor_prestige = 2,  # Set to median value
 circuit_type = "Permanent Race Circuit"  # Most common circuit type
)

grid_effect$podium_prob <- predict(podium_model, newdata = grid_effect, type = "response")

ggplot(grid_effect, aes(x = grid, y = podium_prob)) +
 geom_line(color = "blue", size = 1) +
 geom_point(color = "red", size = 3) +
 labs(title = "Effect of Grid Position on Podium Probability",
      x = "Grid Position",
      y = "Probability of Podium Finish") +
 theme_minimal() +
 scale_x_continuous(breaks = seq(1, 20, by = 1))
```

```
# Analyze the effect of constructor prestige on podium probability
prestige_effect <- data.frame(
  grid = 5,  # Set to median value
  constructor_prestige = 1:6,
  circuit_type = "Permanent Race Circuit"  # Most common circuit type
)
# Analyze false positives and false negatives
misclassified <- test_data %>%
  mutate(predicted = podium_preds,
         actual = podium_finish,
         error_type = case_when(
           predicted == "Podium" & actual == "No Podium" ~ "False Positive",
           predicted == "No Podium" & actual == "Podium" ~ "False Negative",
           TRUE ~ "Correct"
         ))

# Analyze false positives (predicted podium but didn't achieve it)
false_positives <- misclassified %>%
  filter(error_type == "False Positive") %>%
  select(year, race_name, driver, constructor, grid, position, points, constructor_prestige, circuit_type)

# Analyze false negatives (didn't predict podium but achieved it)
false_negatives <- misclassified %>%
  filter(error_type == "False Negative") %>%
  select(year, race_name, driver, constructor, grid, position, points, constructor_prestige, circuit_type)

# Print summary of false positives and negatives
cat("\nFalse Positive Summary (Predicted Podium but didn't achieve it):\n")
print(summary(false_positives))
cat("\nFalse Negative Summary (Didn't predict Podium but achieved it):\n")
##
## False Negative Summary (Didn't predict Podium but achieved it):
```

```
print(summary(false_negatives))
```

Appendix G: Scatter Plot of Interaction Effect


Interaction Effect of Grid Position and Circuit Type on Points
Predicted Points
Grid Position
Circuit Type
Hybrid Circuit
Permanent Race Circuit
Street Circuit