Word count: 1698

Predicting Race Outcomes in Formula 1 between 2014 and 2021

Introduction

Formula 1 (F1) is a globally renowned motorsport in which marginal gains and data-driven strategies heavily determine competitive advantage. Teams operate under highly regulated conditions while navigating variable externalities such as track design and weather. Understanding the factors contributing to race outcomes has significant implications for team strategy and performance analytics. (Msakamali, 2024)

This study seeks to quantify the extent to which starting grid position, constructor tier, and track type (street vs. permanent circuit) influence final driver results in a Grand Prix. The selected period of 2014–2021 ensures consistency in team structures and constructor performance tiers, avoiding confounding changes introduced by the major 2022 technical regulation overhaul. (Méndez, 2023)

The primary objective of this project is to build a multiple linear regression model that examines how grid position, team tier, and track type collectively affect finishing position. The analysis combines exploratory data visualisation with statistical modelling to validate assumptions and quantify effects. The broader significance lies in interpreting whether strategic elements like qualifying performance or structural elements like team constructor tier outweigh environmental randomness introduced by track types

Research Aims and Hypotheses

The current paper seeks to investigate the question: *Does a driver's qualifying position significantly predict their final race result in Formula 1, and do factors such as track type, starting grid position, and team tier modify this relationship?*

Table 1: Aims and Hypotheses

Aim	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
Aim 1: Investigate the relationship between grid position and final race result, regardless of context.	There is no significant relationship between grid position and final race result.	There is a significant relationship between grid position and final race result.
Aim 2: Investigate whether the relationship between grid and race result differs between street circuits and traditional circuits.	There is no significant difference in the relationship between grid and race result across circuit types.	There is a significant difference in the relationship between grid and race result across circuit types.
Aim 3: Investigate whether the impact of grid position varies across team tiers.	No hypotheses stated.	No hypotheses stated.

For the statistical testing in the paper, alpha is set at the conventional level of 0.05

Word count: 1698

Literature Review

A Formula 1 (F1) Grand Prix weekend typically consists of three practice sessions, a qualifying session, and the main race. Qualifying performance determines the starting grid, and drivers compete for points awarded based on final race positions. Qualifying at the front offers an opportunity to start in clean air and avoid early collisions. The race is influenced by tire strategy, pit stop efficiency, and track-specific factors like overtaking opportunities. Circuits are categorised as either street (e.g., Monaco, Singapore) or traditional (e.g., Silverstone, Monza), each presenting unique challenges that affect driver performance and race dynamics.

Previous studies have examined the relationship between starting position and race outcomes in motorsport, with F1 as a popular context due to its data-rich environment. Previous research (Wesselbaum and Owen, 2021) has consistently shown a strong positive correlation between qualifying position and final placement, with front-row starters more likely to finish in top positions. This is often attributed to the importance of track position, especially on circuits where overtaking is limited.

The role of constructor performance has also been explored. Studies by (Budzinski and Feddersen, 2019) categorise teams into tiers based on long-term performance metrics, showing that drivers from top-tier teams consistently outperform lower-tier competitors regardless of starting grid position. This reflects underlying resource differences, including car development budgets, technical staff, and in-race strategy.

Track type is an area that has received growing interest. (Marion *et al.,* 2015) Research suggests that street circuits introduce more variability due to narrow lanes, limited overtaking zones, and higher safety car probability. These factors can undermine the advantage of a strong qualifying position, making final outcomes less predictable than traditional circuits.

I will determine if these factors, such as qualifying performance or team constructor tier, are more significant than each other and how final race position can most accurately be predicted.

Methodology

A multiple linear regression tested the relationship between grid position and final race result, examining the effects of constructor tier and track type. This analysis evaluates how multiple independent variables (IVs) influence a single dependent variable (DV), with each IV assessed while controlling for the others (Tranmer et al., 2020). This study uses quantitative methodology to determine if a driver's starting position, constructor tier, and track type significantly predict their finishing position. The model evaluates both continuous and categorical predictors.

The dataset for this project, compiled from the Ergast Developer API ('Ergast Developer API', 2024), includes driver-level data from all F1 races from 2014 to 2021. This period was selected for regulatory consistency, avoiding disruptions from the 2022 regulation changes (Méndez, 2023). It features each driver's qualifying position, final race classification, team, circuit name, and race year. A binary variable distinguishes street from traditional circuits, based on FIA

classifications and an additional variable segments drivers by grid band (Top/ Midfield/Backmarker)

Before analysis, the dataset was cleaned by removing drivers who did not finish (DNFs) or were disqualified and excluding sprint races and non-championship events.

The final race position is the DV, while the IVs include grid position , constructor tier, and track type. Multiple linear regression allows non-continuous IVs (Tranmer et al., 2020). Based on performance from 2014 to 2021, the constructor tier ranks teams as 1 for top, 2 for midfield, and 3 for backmarkers. Track type is coded as 0 for traditional circuits and 1 for street circuits.

Presenting Data

Firstly, I examined summary statistics to understand the distribution and scale of our key variables.

Table 2: Summary Statistics of Final Race Position, Grid Position, Constructor Team Tier and Circuit Type (Street/Circuit)

Variable	Mean	Standard Deviation	Min	25%	50%	75%	Max
Grid Position	10.35	5.95	1.0	5.0	10.0	15.0	22.0
Final Race Position	9.06	5.09	1.0	5.0	9.0	13.0	22.0
Constructor Tier	1.81	0.64	1.0	1.0	2.0	2.0	3.0
Circuit Type (Street/Circuit)	0.23	0.42	0.0	0.0	0.0	0.0	1.0

All table values rounded to 3 significant figures

These statistics confirm that the dataset is balanced across grid and finishing positions and that most races were held on traditional circuits (as Circuit Type has a mean < 0.5). The spread in Constructor Tier also indicates a fair mix of Top, Midfield and Backmarker teams.

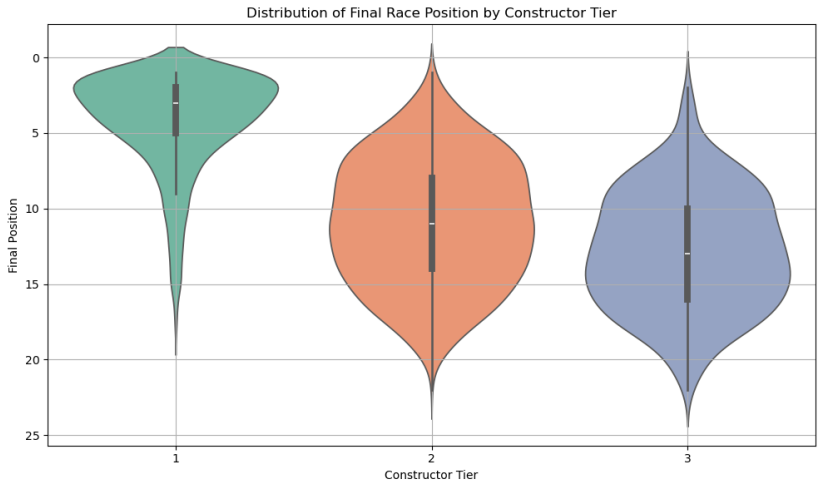
Our analysis begins with visualising how the outcome variable behaves across key categories.

Constructor Tier

Figure 1 shows a violin plot of finishing positions by constructor tier. This helps us understand how performance varies between the top and bottom of the grid. Figure 1 demonstrates that drivers from Tier 1 constructors consistently achieve better final positions, with a tighter distribution. The spread increases across Tiers 2 and 3, highlighting how the tiers influence race consistency. These patterns, illustrated visually in Figure 1, are later confirmed through regression analysis (Table 5), where the constructor tier significantly affects the final position.

Word count: 1698

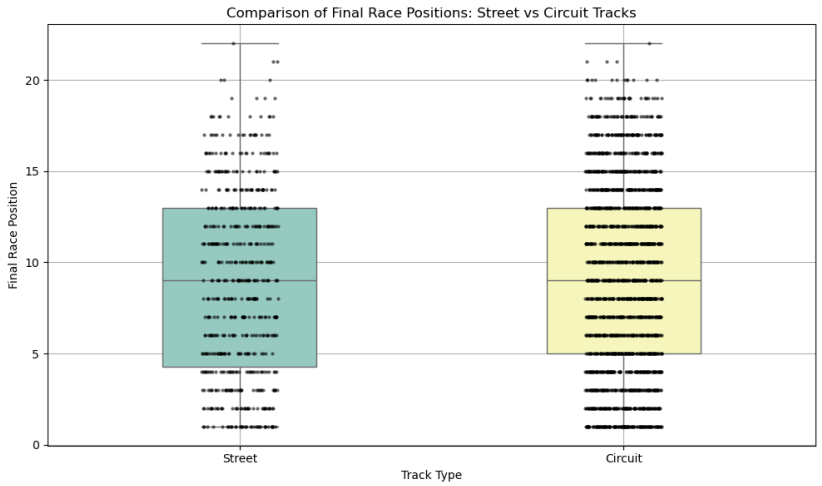
Figure 1: Violin plot to show the Distribution of Final Race Position by Constructor Tier



Circuit Type

Figure 2 displays the distribution of finishing positions on street vs. circuit tracks. Street circuits show more variability and a wider interquartile range, suggesting increased unpredictability. This supports findings in the literature and the significance of the 'Circuit Type' variable in the regression model.

Figure 2: Boxplot to show the Comparison of Final Race Positions: Street vs Circuit Tracks



Results

Table 3: Results of the Fit of the Multiple Linear Regression Model to the Data

R^2	0.611
Adjusted R^2	0.611
F-statistic	1404.0
Probability of F-statistic	< 0.001

All table values rounded to 3 significant figures

Word count: 1698

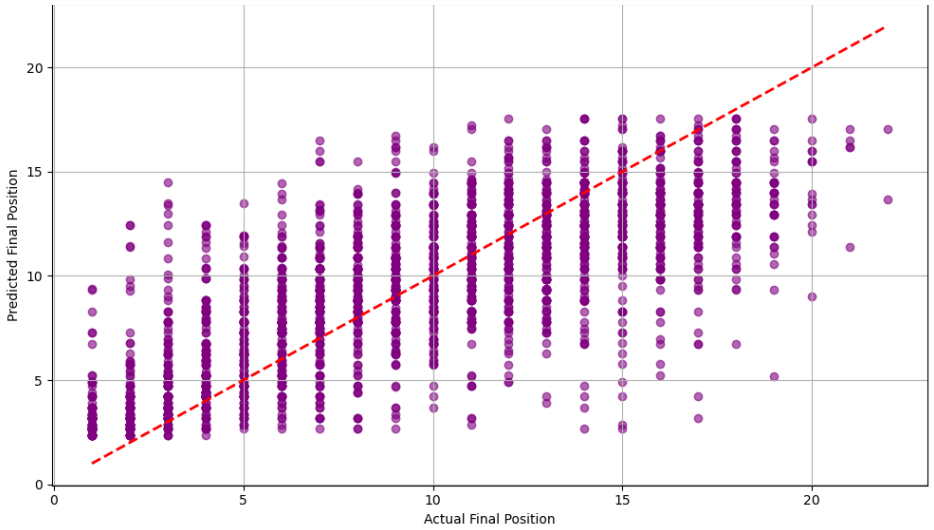
Table 4: Multiple Linear Regression Coefficients, and their Associated t-values and p-values

Variable	Coefficient	t-value	p-value
Intercept (Constant Term)	0.119	0.633	0.527
Grid (x_1)	0.516	40.2	<0.001
Tier (x_2)	2.027	16.9	<0.001
Street (x_3)	-0.318	-2.18	0.029
$y = 0.516x_1 + 2.027x_2 - 0.318x_3 + 0.119$ Where y is the final race position, (x_1) is the grid position, (x_2) is the constructor tier, and Grid (x_3) is the binary categorical variable circuit type (street/circuit)			

All table values rounded to 3 significant figures

The regression model returned an R^2 value of 0.611, indicating that approximately 61.1% of the variance in final race positions is explained by the combination of grid position, constructor tier, and track type. Each variable was statistically significant at the 5% level. Grid position emerged as the most influential predictor, with each place lost in qualifying corresponding to an expected drop of approximately 0.52 places in the final result. Constructor tier also played a critical role, as drivers from Tier 3 teams were predicted to finish over two positions higher than those from Tier 1 teams, all else being equal. While less impactful, track type showed a modest yet statistically significant effect, with drivers generally performing slightly better on street circuits—likely reflecting the unpredictability of such environments, which can diminish the advantage of top teams.

Figure 3: Model Fit of Multi-Linear Regression: Actual vs Predicted Final Race Position



This plot assesses the predictive accuracy of the multiple linear regression model by comparing actual race outcomes to predicted finishing positions. Each point represents a race result. The red dashed line denotes perfect prediction, where actual and predicted values align. The tight clustering of points around the line—particularly for front-runners—indicates strong predictive validity, while wider dispersion in the midfield and backmarker regions suggests greater variability in outcomes for lower grid starters.

Word count: 1698

Discussion

The regression findings support long-standing intuitions in F1 strategy. Grid position emerged as the strongest predictor of the final outcome, reinforcing the importance of qualifying performance. For every place lost on the grid, a driver is expected to finish roughly half a place lower, all else being equal. The constructor tier had the second strongest influence, with drivers from lower-tier teams predicted to finish approximately two places behind their higher-tier counterparts, while controlling for grid and track. This highlights structural disparities in car performance and technical resources. The effect of track type was statistically significant but more modest. Street circuits were associated with slightly better finishing positions on average, potentially due to chaotic race dynamics that occasionally benefit lower-tier drivers through incidents, safety cars, and reduced overtaking. While the model explains 61.1% of the variance, the remaining 38.9% may be attributed to in-race incidents, driver skill, team strategy, weather conditions, and other unmeasured factors. This points to the value of future models incorporating additional data such as weather, pit stops, and DNF (Did Not Finish) status.

Limitations and Considerations

However, several limitations must be acknowledged. The model does not incorporate dynamic variables such as weather conditions, tyre strategies, pit stop execution, or safety car deployments, which could materially affect race outcomes. Moreover, while acknowledged, driver-specific skills and team orders were beyond this analysis's scope. The treatment of final position as a continuous variable, rather than ordinal, presents an additional modelling constraint. Ethical considerations were minimal as the analysis relied entirely on publicly available and anonymised data.

Conclusion

This investigation demonstrates that an F1 driver’s grid position is a significant predictor of race outcomes and constructor tier, although less so, and both factors show strong practical relevance. Including track type revealed its limited influence but added variability to the model, reflecting the dynamics of street circuits. The study’s strengths include structured, multi-season race data, suitable statistical techniques, and clear visualisations supporting key findings. Future research could integrate time-series data and fixed-effects models to distinguish driver impact from team influence. Overall, this project establishes a credible, replicable foundation for predictive modelling in F1 and paves the way for advanced performance analytics in high-variance motorsport contexts.

Word count: 1698

References

Budzinski, O. and Feddersen, A. (2019). 'Measuring Competitive Balance in Formula One Racing'. doi: 10.2139/ssrn.3357687. (Accessed 1 May 2025)

Marion, A., Aversa, P., Luiz, Mesquita, and Jaideep, Anand. (2015). 'Driving performance via exploration in changing environments: Evidence from formula one racing'. *Organization Science*, 26 (4), pp. 1079–1100. doi: https://doi.org/10.1287/orsc.2015.0984. (Accessed 2 May 2025)

Méndez, L. A. (2023). 'Quantifying the Impact of the 2022 Formula One Technical Regulations on Wake Turbulence: A Numerical Analysis'. *University of Dayton*. (Accessed 2 May 2025)

Msakamali, B. (2024). 'F1 Data Analysis and Tactical Insights: Exploring Formula 1 Race Performance Strategies.' *Tampere University of Applied Sciences*. (Accessed 1 May 2025)

Wesselbaum, D. and Owen, P. (2021). 'The Value of Pole Position in Formula 1 History'. *Australian Economic Review*, 54, pp. 164–173. doi: 10.1111/1467-8462.12401. (Accessed 2 May 2025)

Word count: 1698

Appendix

A. Classification of Race vs Street Circuits and binary value assigned to them

Grand Prix	Type	
Australian Grand Prix, Monaco Grand Prix, Canadian Grand Prix, Singapore Grand Prix, Russian Grand Prix, Azerbaijan Grand Prix, Saudi Arabian Grand Prix	Street	1
Malaysian Grand Prix, Bahrain Grand Prix, Chinese Grand Prix, Spanish Grand Prix, Austrian Grand Prix, British Grand Prix, German Grand Prix, Hungarian Grand Prix, Belgian Grand Prix, Italian Grand Prix, Japanese Grand Prix, United States Grand Prix, Brazilian Grand Prix, Abu Dhabi Grand Prix, French Grand Prix, Mexican Grand Prix, Styrian Grand Prix, 70th Anniversary Grand Prix, Tuscan Grand Prix, Eifel Grand Prix, Portuguese Grand Prix, Emilia Romagna Grand Prix, Turkish Grand Prix, Sakhir Grand Prix, Qatar Grand Prix, Dutch Grand Prix, Sao Paulo Grand Prix, San Marino Grand Prix	Circuit	0

B. Classification of Constructor Team Tier and values assigned to them

Teams	Rank
Ferrari, Red Bull, Mercedes	Top (1)
McLaren, Williams, Renault, Force India, Sauber, Alfa Romeo, Aston Martin, Haas F1 Team, Racing Point, AlphaTauri, Alpine F1	Midfield (2)
Toro Rosso, Marussia, Caterham, Lotus F1	Backmarkers (3)

C. Python code snippet of violin plot

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load your data
df = pd.read_csv("raceScatter.csv")
output_filename = 'violin_figure.png'

# Clean the data
df['position'] = pd.to_numeric(df['position'], errors='coerce')
df['tier'] = pd.to_numeric(df['tier'], errors='coerce')
df = df.dropna(subset=['position', 'tier'])

# Create violin plot
plt.figure(figsize=(10, 6))
sns.violinplot(x='tier', y='position', data=df, inner='box', palette='Set2')
plt.title("Distribution of Final Race Position by Constructor Tier")
plt.xlabel("Constructor Tier")
plt.ylabel("Final Position")
plt.gca().invert_yaxis()
plt.grid(True)
plt.tight_layout()
plt.savefig(output_filename)
plt.show()
```

Word count: 1698

D. Python code snippet of box plot

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

output_filename = "boxplot"
# Load data
df = pd.read_csv('raceScatter.csv')
df['position'] = pd.to_numeric(df['position'], errors='coerce')
df['Street'] = df['Street'].map({0: 'Circuit', 1: 'Street'})
df = df.dropna(subset=['position', 'Street'])

# Plot
plt.figure(figsize=(10, 6))
sns.boxplot(x='Street', y='position', data=df, width=0.4, palette='Set3', showfl:
sns.stripplot(x='Street', y='position', data=df, color='black', size=3, alpha=0.(

plt.title("Comparison of Final Race Positions: Street vs Circuit Tracks")
plt.xlabel("Track Type")
plt.ylabel("Final Race Position")

plt.grid(True)
plt.tight_layout()
plt.savefig(output_filename)
plt.show()
```

E. Python code snippet of multi variable linear regression

```python
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

outputFile = "multilinear"
# Load data
df = pd.read_csv("raceScatter.csv")

# Clean relevant columns
df['position'] = pd.to_numeric(df['position'], errors='coerce')
df['grid'] = pd.to_numeric(df['grid'], errors='coerce')
df['tier'] = pd.to_numeric(df['tier'], errors='coerce')
df['Street'] = pd.to_numeric(df['Street'], errors='coerce')

# Drop missing values
df = df.dropna(subset=['position', 'grid', 'tier', 'Street'])

# Define X (predictors) and y (target)
X = df[['grid', 'tier', 'Street']]
y = df['position']

# Add constant (intercept)
X = sm.add_constant(X)

# Fit model
model = sm.OLS(y, X).fit()
y_pred = model.predict(X)

#  Plot actual vs predicted
plt.figure(figsize=(10, 6))
plt.scatter(y, y_pred, alpha=0.6, c='purple')
plt.plot([min(y), max(y)], [min(y), max(y)], 'r--', linewidth=2)  # 45° line
plt.title("Actual vs Predicted Final Race Position")
plt.xlabel("Actual Final Position")
plt.ylabel("Predicted Final Position")
plt.gca().invert_xaxis()  # Optional: Lower actual position is better
plt.gca().invert_yaxis()  # Optional: Lower predicted position is better
plt.grid(True)
plt.tight_layout()
plt.show()
plt.savefig(outputFile)

# Print regression summary (optional)
print(model.summary())
```

Word count: 1698

F. Dataset sample

resultId	year	Street		raceId	driverId	tier	constructorId	number	grid	position
22130	2014	1	Australian Gr	900	3	1	131	6	3	1
22131	2014	1	Australian Gr	900	825	2	1	20	4	2
22132	2014	1	Australian Gr	900	18	2	1	22	10	3
22133	2014	1	Australian Gr	900	4	1	6	14	5	4
22134	2014	1	Australian Gr	900	822	2	3	77	15	5
22135	2014	1	Australian Gr	900	807	2	10	27	7	6
22136	2014	1	Australian Gr	900	8	1	6	7	11	7
22137	2014	1	Australian Gr	900	818	3	5	25	6	8
22138	2014	1	Australian Gr	900	826	3	5	26	8	9
22139	2014	1	Australian Gr	900	815	2	10	11	16	10
22140	2014	1	Australian Gr	900	16	2	15	99	13	11
22141	2014	1	Australian Gr	900	821	2	15	21	20	12
22142	2014	1	Australian Gr	900	820	3	206	4	17	13
22152	2014	0	Malaysian Gr	901	1	1	131	44	1	1
22153	2014	0	Malaysian Gr	901	3	1	131	6	3	2
22154	2014	0	Malaysian Gr	901	20	1	9	1	2	3
22155	2014	0	Malaysian Gr	901	4	1	6	14	4	4
22156	2014	0	Malaysian Gr	901	807	2	10	27	7	5
22157	2014	0	Malaysian Gr	901	18	2	1	22	10	6
22158	2014	0	Malaysian Gr	901	13	2	3	19	13	7
22159	2014	0	Malaysian Gr	901	822	2	3	77	18	8
22160	2014	0	Malaysian Gr	901	825	2	1	20	8	9
22161	2014	0	Malaysian Gr	901	826	3	5	26	11	10
22162	2014	0	Malaysian Gr	901	154	3	208	8	15	11
22163	2014	0	Malaysian Gr	901	8	1	6	7	6	12
22164	2014	0	Malaysian Gr	901	155	3	207	10	20	13
22165	2014	0	Malaysian Gr	901	828	3	207	9	22	14
22166	2014	0	Malaysian Gr	901	820	3	206	4	21	15