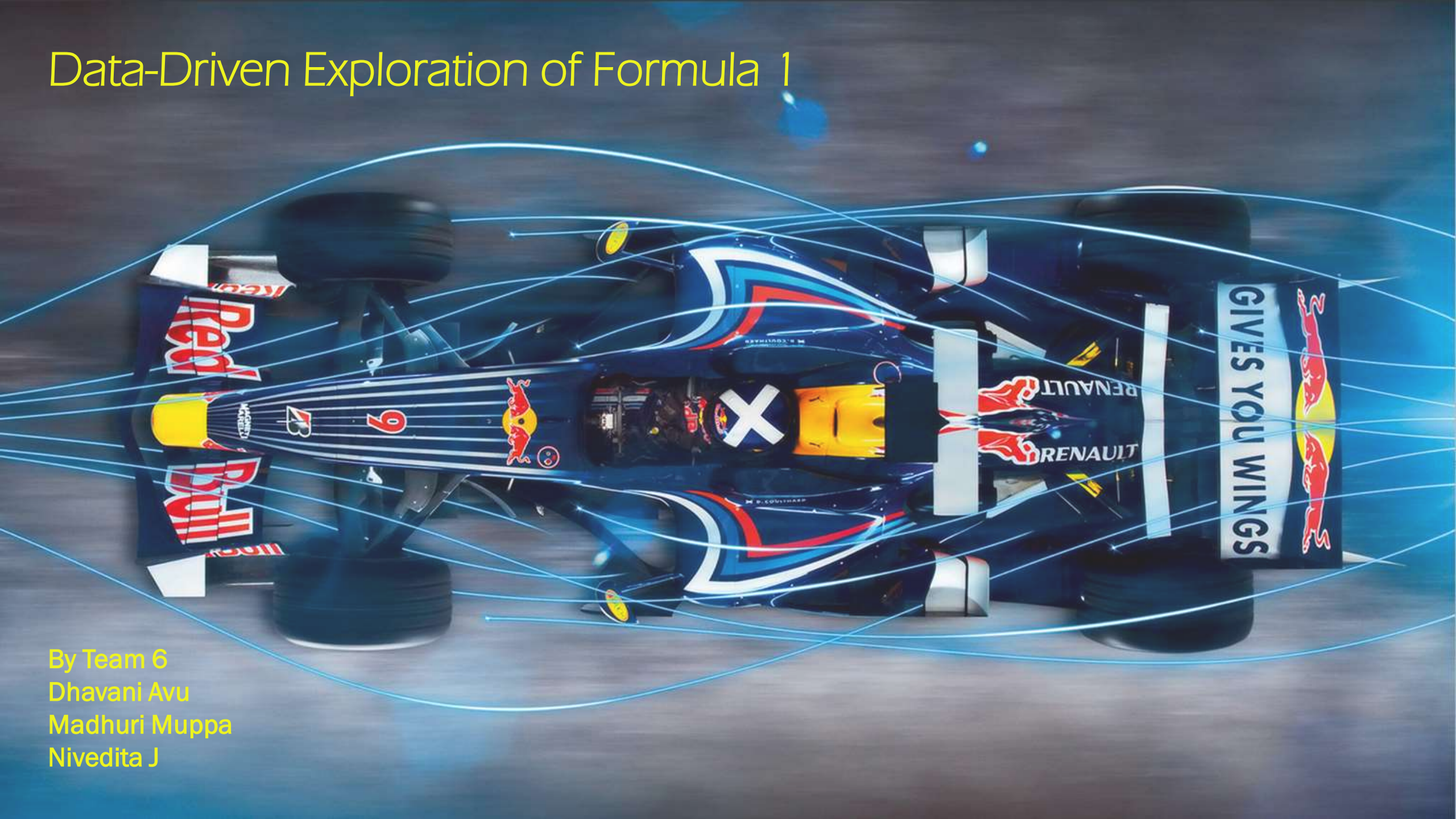
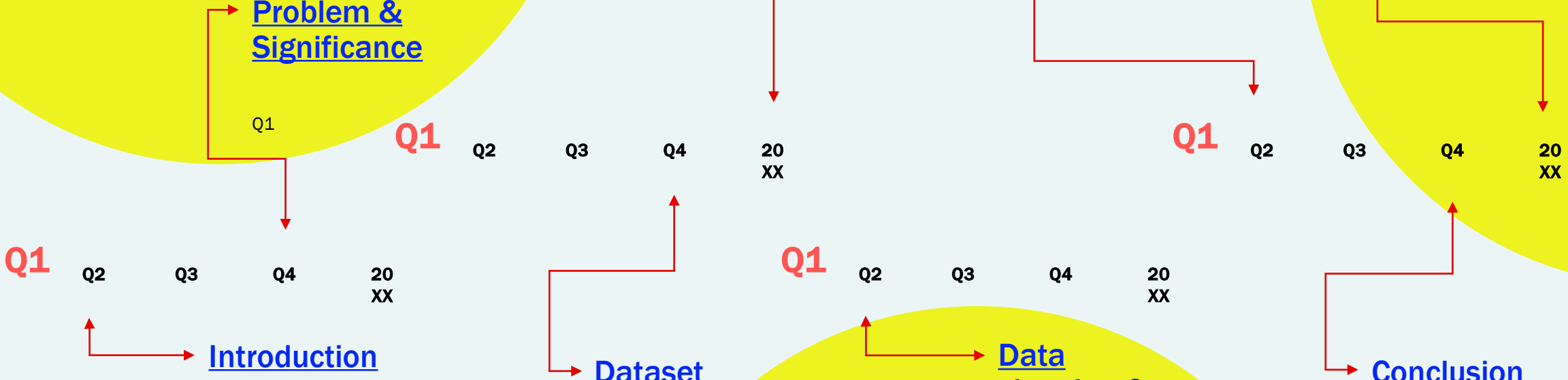
Data-Driven Exploration of Formula 1



By Team 6
Dhavani Avu
Madhuri Muppa
Nivedita J

CONTENTS



Introduction

- Formula 1 is the pinnacle of motorsport, featuring the fastest racing cars globally, capable of reaching speeds over 200 mph.
- Teams employ state-of-the-art technology in aerodynamics, materials, and engines, pushing the boundaries of innovation.
- A truly global sport, Formula 1 hosts races on iconic tracks around the world, from Monaco's city streets to Monza's high-speed straights.
- The sport attracts the best drivers globally, showcasing exceptional skills, reflexes, and physical fitness.
- Each team consists of engineers, strategists, and support staff working collaboratively to optimize car performance and race strategies.
- Races are not just about speed; strategic decisions, pit stops, and tire choices play a crucial role in achieving victory.
- Beyond the racing, Formula 1 embodies luxury and glamour, attracting a global audience and high-profile sponsors.
- With a history dating back to the 1950s, Formula 1 has evolved, witnessed legendary rivalries, and set benchmarks for motorsport excellence.
- Modern Formula 1 relies heavily on data analytics, from car performance optimization to race strategy planning.
- Boasting a passionate and diverse fan base, Formula 1 captivates audiences with thrilling races and iconic moments.

Problem & Significance

Problem Statement:

- Formula 1 teams face the challenge of optimizing race strategies for enhanced performance in dynamic conditions.
- Research Questions:
  - How can machine learning models be used to predict the outcomes of Formula 1 races
  - How do different Formula 1 circuits impact race outcomes.
  - Are there particular years with significant shifts in constructors' performance?
  - Do demographic factors of the driver have any impact on the race outcome?

Significance of the Problem:

- Efficient race strategies contribute to success in Formula 1, impacting team standings and driver achievements.
- Helps to unravel the complexities, providing valuable insights into optimizing race outcomes through data-driven analysis.
- The significance extends beyond racing, showcasing the applicability of data science and machine learning in real-world scenarios.

Dataset description

- **Ergast Motor Racing Data API**: Our dataset is sourced from the Ergast Motor Racing Data API, a comprehensive open-source repository focused on Formula 1 races, drivers, and constructors.
- **Rich Historical Data**: Spanning from 1950 to the present, the dataset provides a rich historical perspective, enabling in-depth analyses of Formula 1 dynamics over time.
- **Comprehensive Information**: The Ergast dataset includes a diverse range of columns, offering essential information on race circuits, constructors, drivers, race-specific details, and comprehensive results and standings.
- **Granularity**: Achieving a high level of granularity, the dataset includes lap-by-lap details, qualifying results, pit stop information, and fine-grained race-specific statistics.
- **Reliability and Openness**: As an open-source API, Ergast is a reliable and transparent resource, supporting our research with accurate and detailed data.
- **Geographical Data**: The dataset also incorporates geographical data, including latitude and longitude, facilitating the exploration of relationships between circuit locations and race outcomes.

Proposed approach

DATA COLLECTION

DATA CLEANING

EXPLORATORY DATA ANALYSIS

FEATURE ENGINEERING

MODEL BUILDING AND EVALUATION

Data Cleaning

The dataset utilized for our project was obtained from ergast.com/f1, encompassing comprehensive data from Formula 1 seasons from 1960 to present.

The dataset comprised 14 distinct datasets interconnected by foreign keys, and we successfully merged them to align with our project specifications.

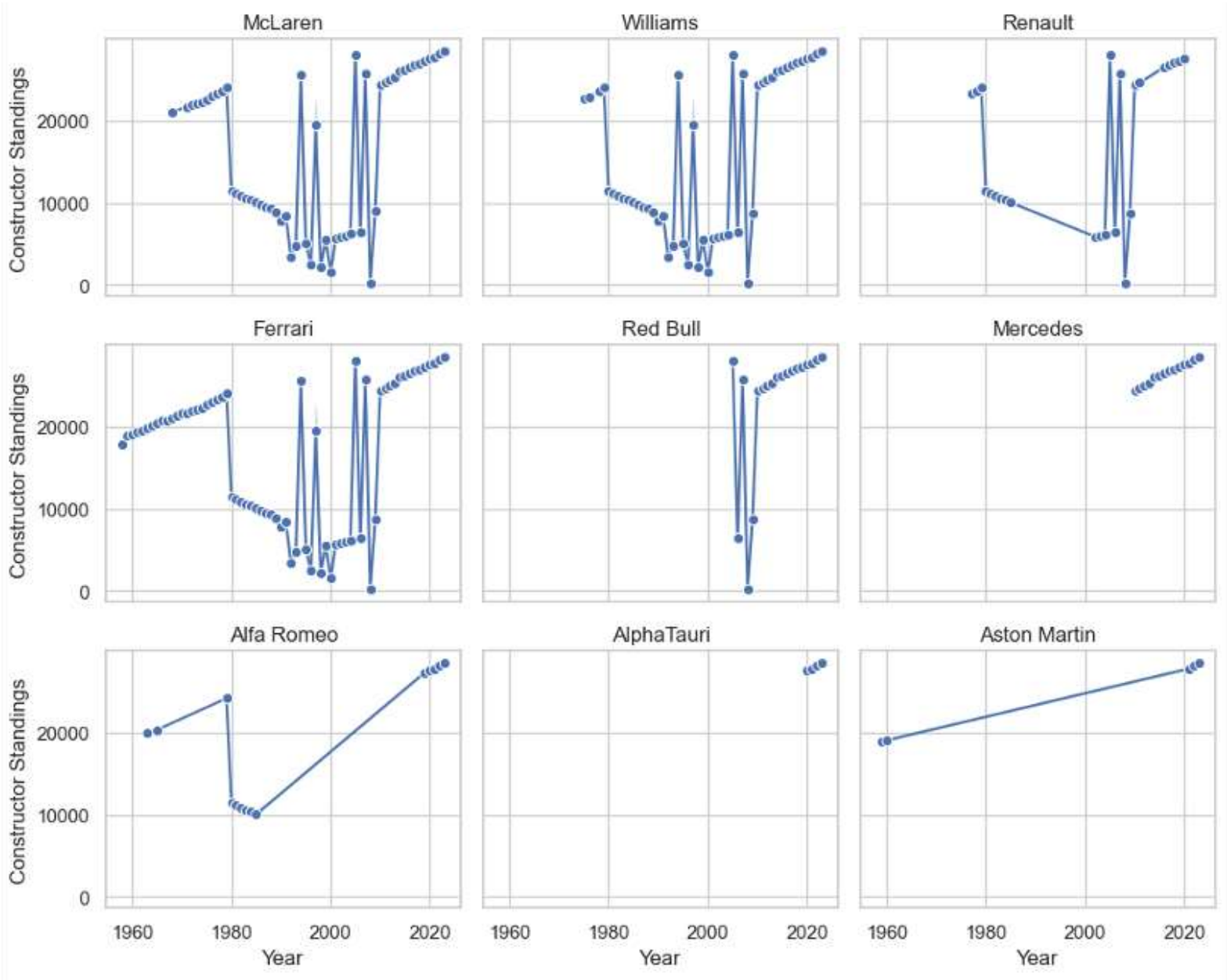
Within these datasets, null values were denoted by "\N."

To uphold data integrity, we addressed these null values by either removal or replacement with average values as deemed appropriate.

Constructors performance from 1960- Present

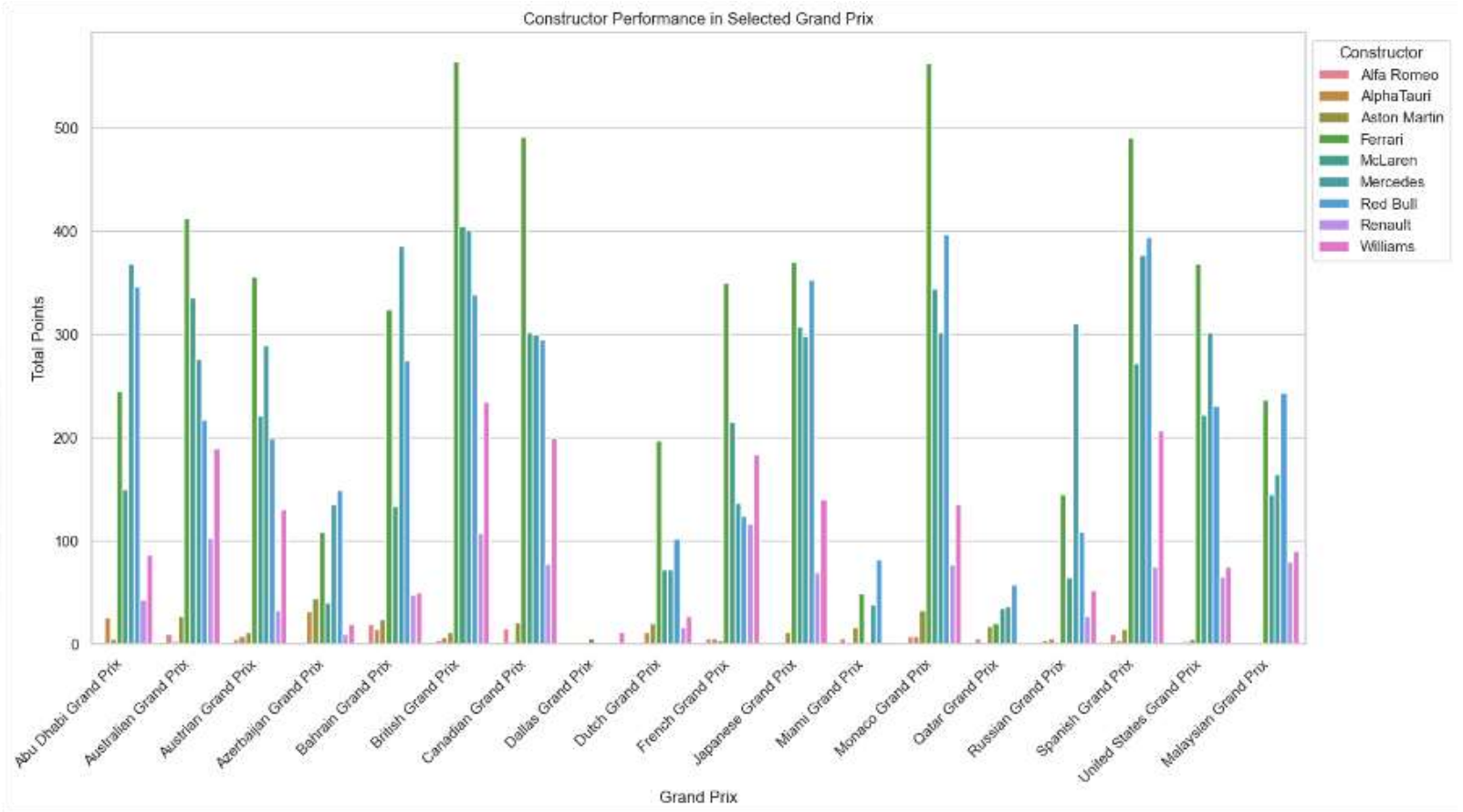
- According to this graph you can clearly see that no particular team has had a consistent improvement in the performance over the years.

- Teams like McLaren and Williams have been in the sport since the beginning of the sport and their performances have had significant drops and progress.

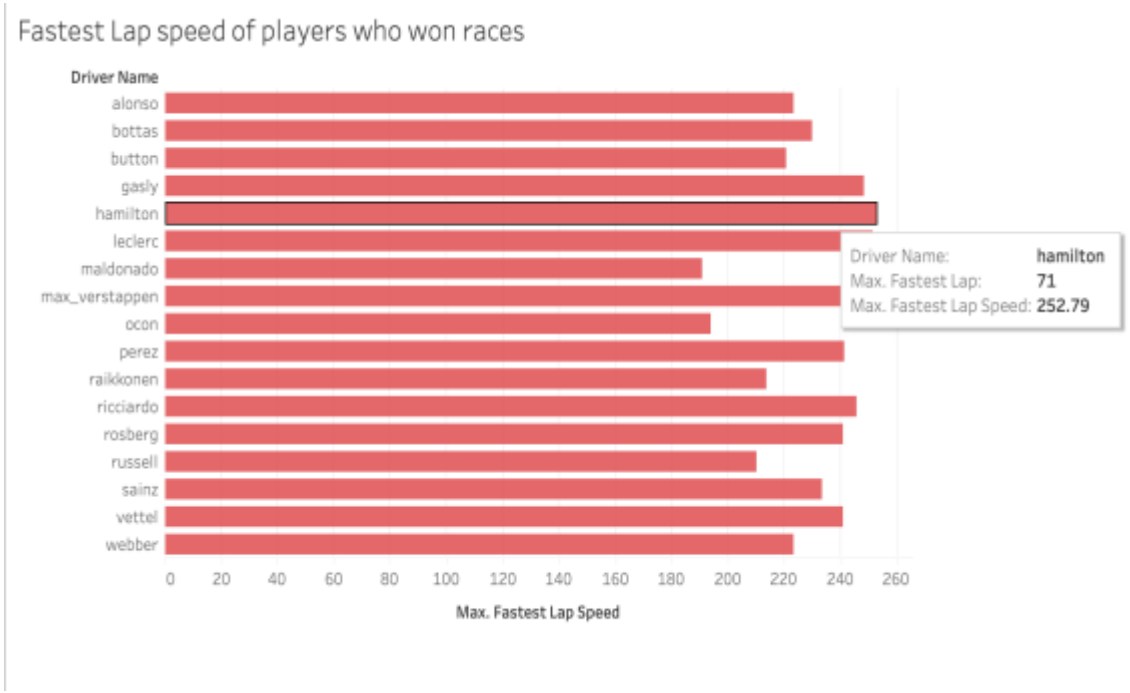


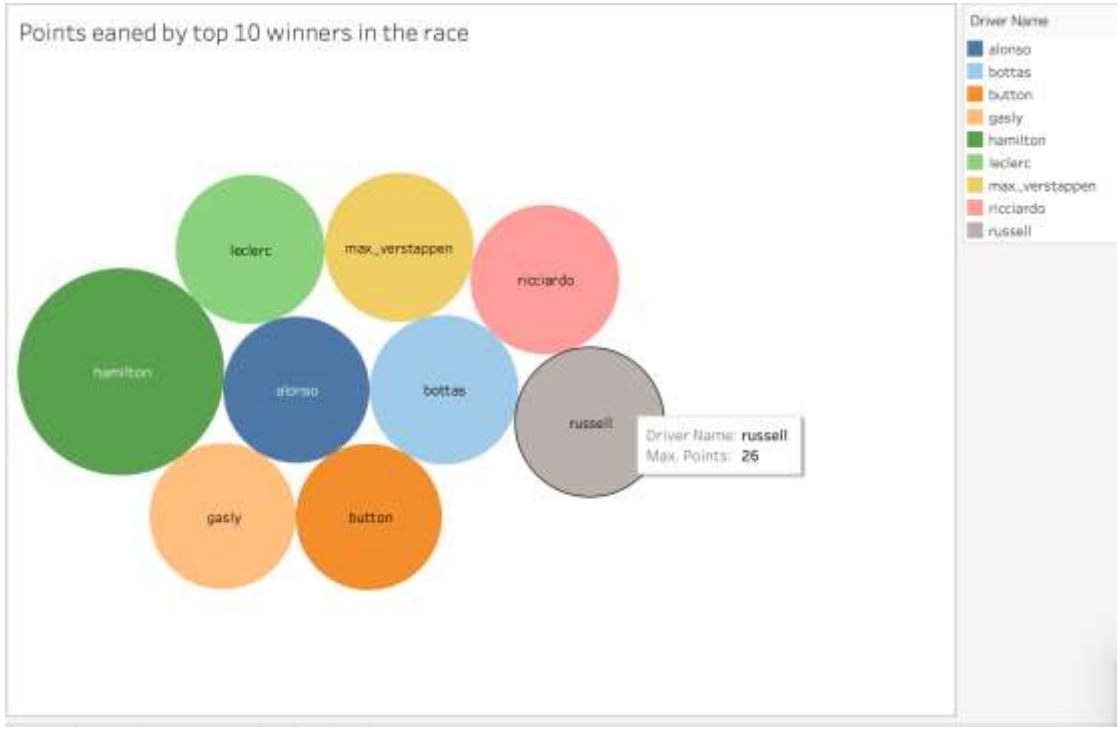
Circuits impact on race outcomes

- Every car is not suitable for every circuit. Some cars perform better in high-speed corners while some perform better in low-speed corners, and some are best on straight lines.
- Looking at the graph you can say t hat few constructors perform best at few race circuits.

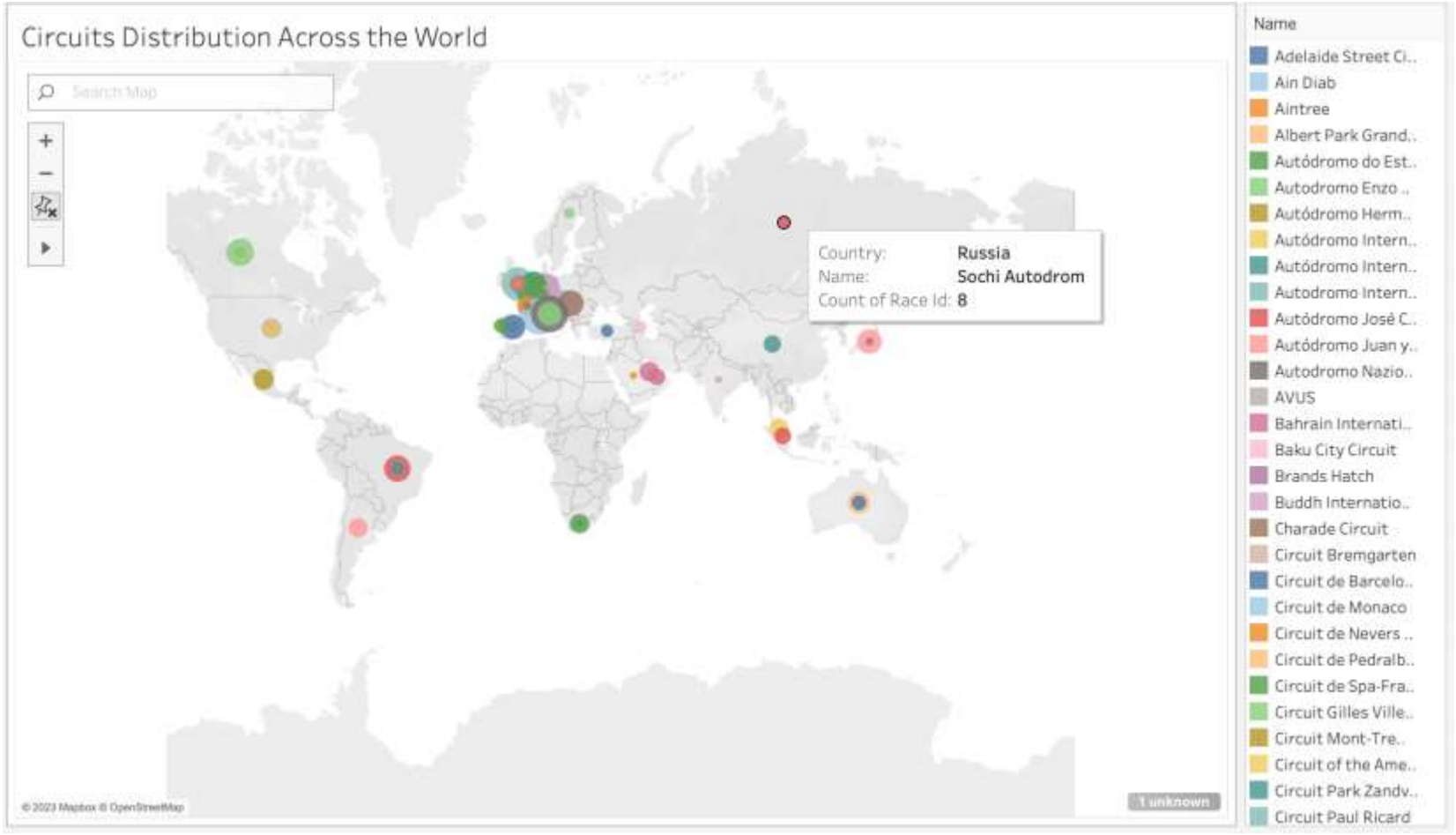


Driver Standings





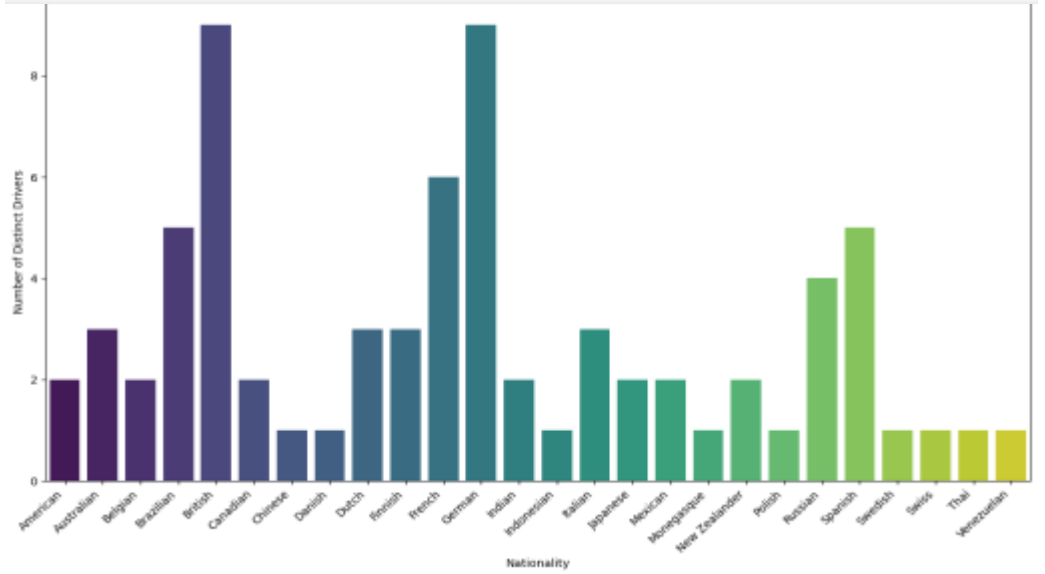
- These formula 1 races happen all over the world.
- We typically have 20 races or so and we have around 77 racetracks in the world.
- This is an illustration of the circuits around the world.



Driver Demographics

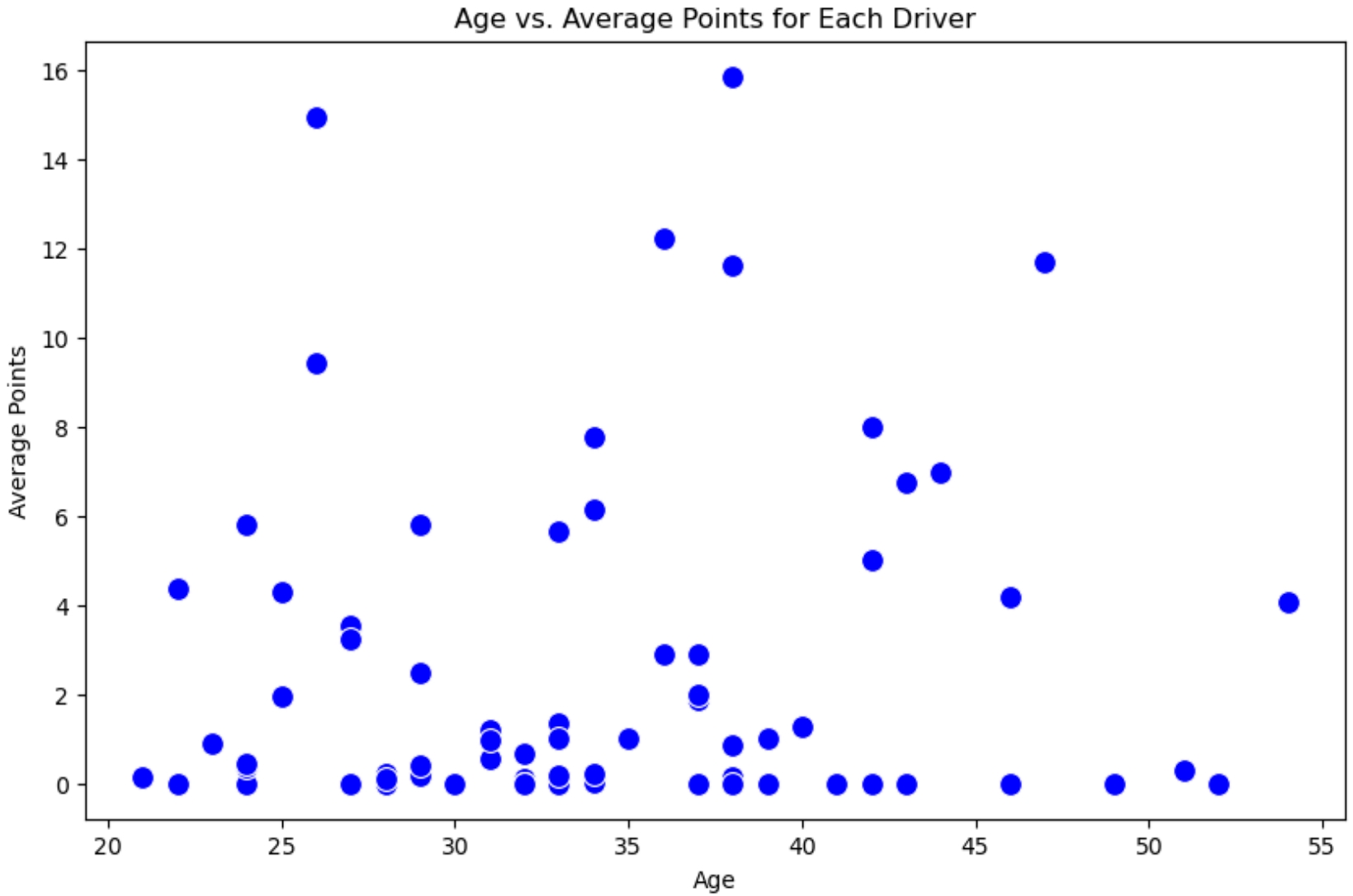
- As per our graphs, the highest number of wins acquired are by the drivers from United Kingdom and the second highest number of wins acquired were by the drivers from Germany.

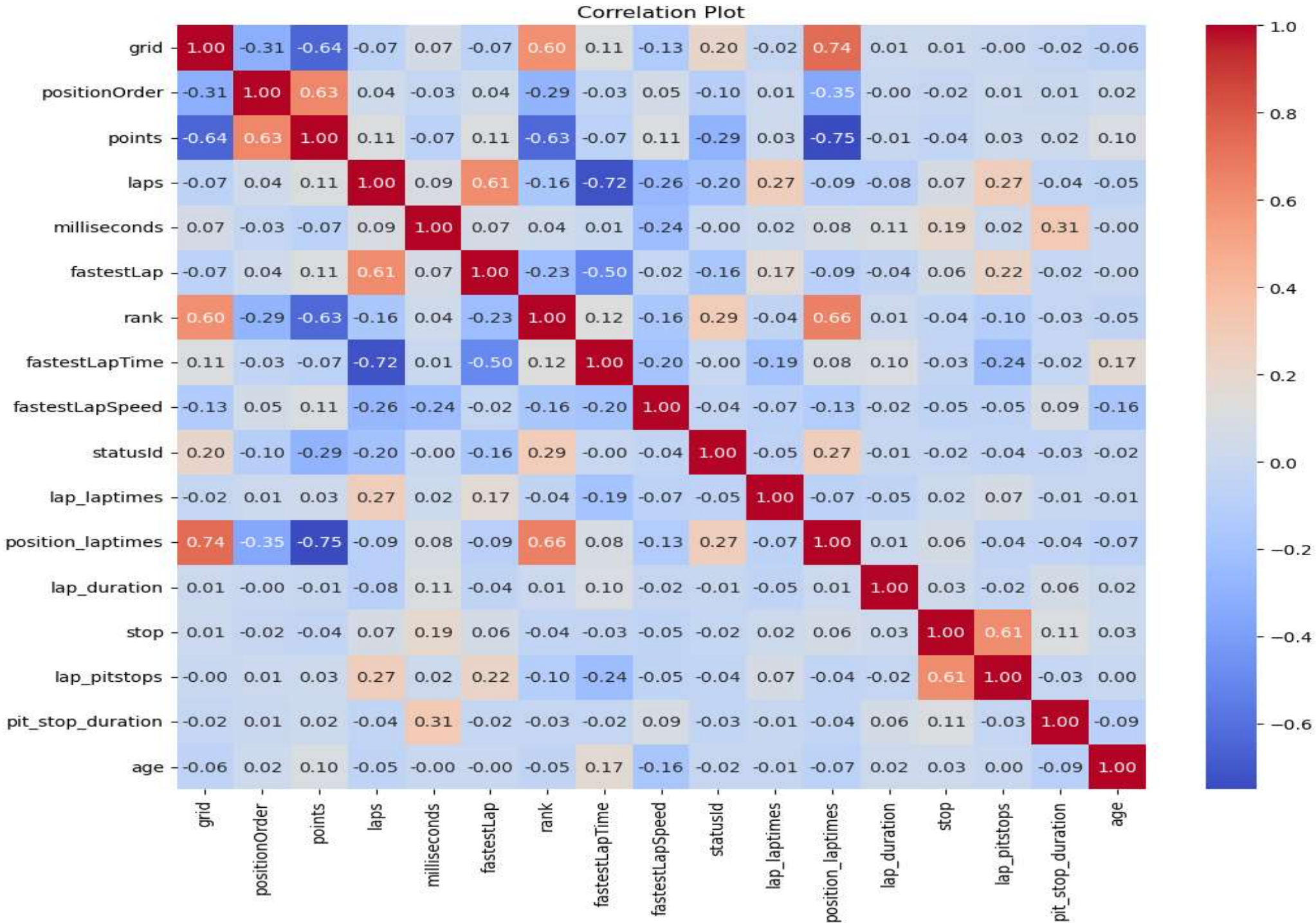
- This could also be because, these two countries contribute highest to the total number of drivers in Formula 1.



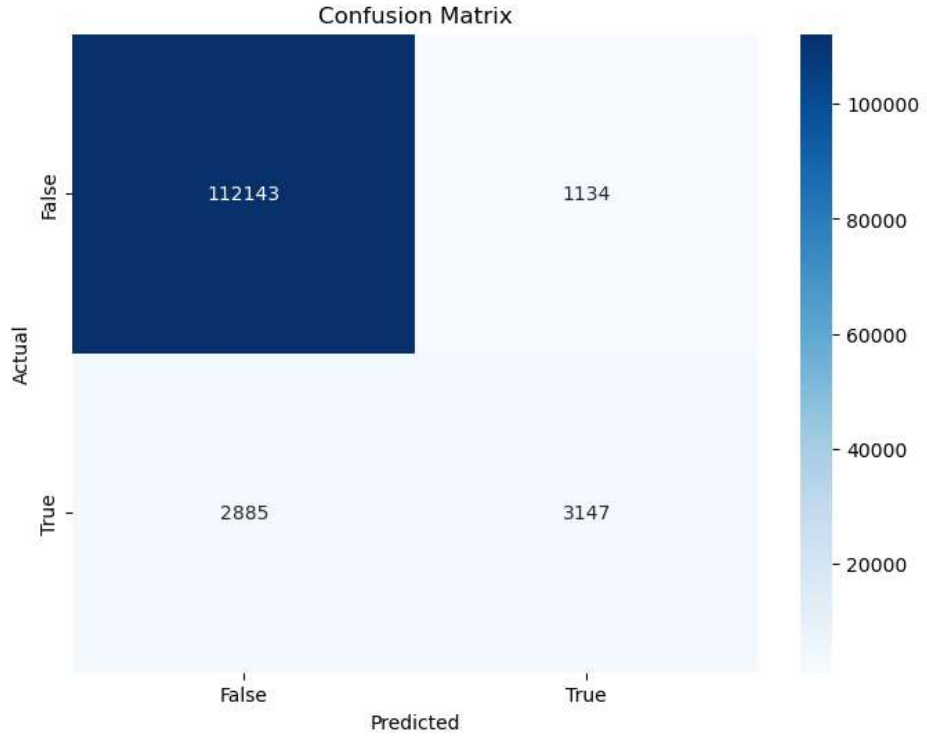
Is age a factor for driver performance?

- Based on the current analysis, we do not find a strong technical basis to assert a direct relationship between the age of drivers and the average points they earn.




Correlation Plot
grid 1.00 -0.31 -0.64 -0.07 0.07 -0.07 0.60 0.11 -0.13 0.20 -0.02 0.74 0.01 0.01 -0.00 -0.02 -0.06
positionOrder -0.31 1.00 0.63 0.04 -0.03 0.04 -0.29 -0.03 0.05 -0.10 0.01 -0.35 -0.00 -0.02 0.01 0.01 0.02
points -0.64 0.63 1.00 0.11 -0.07 0.11 -0.63 -0.07 0.11 -0.29 0.03 -0.75 -0.01 -0.04 0.03 0.02 0.10
laps -0.07 0.04 0.11 1.00 0.09 0.61 -0.16 -0.72 -0.26 -0.20 0.27 -0.09 -0.08 0.07 0.27 -0.04 -0.05
milliseconds 0.07 -0.03 -0.07 0.09 1.00 0.07 0.04 0.01 -0.24 -0.00 0.02 0.08 0.11 0.19 0.02 0.31 -0.00
fastestLap -0.07 0.04 0.11 0.61 0.07 1.00 -0.23 -0.50 -0.02 -0.16 0.17 -0.09 -0.04 0.06 0.22 -0.02 -0.00
rank 0.60 -0.29 -0.63 -0.16 0.04 -0.23 1.00 0.12 -0.16 0.29 -0.04 0.66 0.01 -0.04 -0.10 -0.03 -0.05
fastestLapTime 0.11 -0.03 -0.07 -0.72 0.01 -0.50 0.12 1.00 -0.20 -0.00 -0.19 0.08 0.10 -0.03 -0.24 -0.02 0.17
fastestLapSpeed -0.13 0.05 0.11 -0.26 -0.24 -0.02 -0.16 -0.20 1.00 -0.04 -0.07 -0.13 -0.02 -0.05 -0.05 0.09 -0.16
statusId 0.20 -0.10 -0.29 -0.20 -0.00 -0.16 0.29 -0.00 -0.04 1.00 -0.05 0.27 -0.01 -0.02 -0.04 -0.03 -0.02
lap_laptimes -0.02 0.01 0.03 0.27 0.02 0.17 -0.04 -0.19 -0.07 -0.05 1.00 -0.07 -0.05 0.02 0.07 -0.01 -0.01
position_laptimes 0.74 -0.35 -0.75 -0.09 0.08 -0.09 0.66 0.08 -0.13 0.27 -0.07 1.00 0.01 0.06 -0.04 -0.04 -0.07
lap_duration 0.01 -0.00 -0.01 -0.08 0.11 -0.04 0.01 0.10 -0.02 -0.01 -0.05 0.01 1.00 0.03 -0.02 0.06 0.02
stop 0.01 -0.02 -0.04 0.07 0.19 0.06 -0.04 -0.03 -0.05 -0.02 0.02 0.06 0.03 1.00 0.61 0.11 0.03
lap_pitstops -0.00 0.01 0.03 0.27 0.02 0.22 -0.10 -0.24 -0.05 -0.04 0.07 -0.04 -0.02 0.61 1.00 -0.03 0.00
pit_stop_duration -0.02 0.01 0.02 -0.04 0.31 -0.02 -0.03 -0.02 0.09 -0.03 -0.01 -0.04 0.06 0.11 -0.03 1.00 -0.09
age -0.06 0.02 0.10 -0.05 -0.00 -0.00 -0.05 0.17 -0.16 -0.02 -0.01 -0.07 0.02 0.03 0.00 -0.09 1.00
grid
positionOrder
points
laps
milliseconds
fastestLap
rank
fastestLapTime
fastestLapSpeed
statusId
lap_laptimes
position_laptimes
lap_duration
stop
lap_pitstops
pit_stop_duration
age
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Model	Accuracy
Naïve Bayes	93.6%
Decision Tree with cross validation	92.6%
Logistic Regression	94.9%
Logistic Regression with feature scaling	95.2%
Gradient Boosting	96.6%



Modeling & Evaluation

Conclusion

Our exploration showcased the power of machine learning in predicting Formula 1 race outcomes. The Gradient Boosting algorithm, with an impressive 96.6% accuracy, demonstrated its potential as a robust tool for anticipating race winners.

Geographical visualizations uncovered the global influence of Formula 1 circuits on race dynamics. Our maps provided a unique perspective on how different circuits contribute to the sport's diverse and dynamic landscape.

Analyzing constructors' points over the years unveiled pivotal moments, highlighting years with significant shifts in the competitive balance. This exploration opens avenues for understanding the factors driving these transformative periods in Formula 1 history.

Exploring driver demographics through visualizations offered nuanced insights. While no clear linear trends emerged, scatter plots and bar charts provided a unique lens into the diverse backgrounds of Formula 1 drivers and their potential impact on race outcomes.

References

- Bekker, J., Lotz, W. Planning Formula One race strategies using discreteevent simulation. J Oper Res Soc 60, 952–961 (2009). https://doi.org/10.1057/palgrave.jors.2602626
- Garcia Tejada, L. (2023). Applying Machine Learning to Forecast Formula 1 Race Outcomes. Aaltodoc.aalto.fi. https://aaltodoc.aalto.fi/handle/123456789/122937
- Jasper. “Visualizing Formula 1 Race Strategies in Python Using Fastf1, Pandas and Matplotlib.” Towards Formula 1 Analysis, 17 July 2022, medium.com/towards-formula-1-analysis/visualizing-formula1-racestrategies-in-python-using-fastf1-pandas-and-matplotlib-95fe6b3298fa. Accessed 19 Sept. 2023.
- N. Saijo, R. Nishizono and M. Kashino, "Relationship between predriving heart rate and driving performance in formula car racing: a case study," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 4957-4960, doi: 10.1109/EMBC46164.2021.9630288.
- P. Azzoni, D. Moro and G. Rizzoni, "Time-frequency signal analysis of the acoustic emission of Formula 1 engines," Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis (Cat. No.98TH8380), Pittsburgh, PA, USA, 1998, pp. 441-444, doi:10.1109/TFSA.1998.721456.
- Patil, Ankur, et al. A Data-Driven Analysis of Formula 1 Car Races Outcome. Jan. 2023, pp. 134–46, https://doi.org/10.1007/978-3-031-26438-2_11.
- T. Lampprecht, D. Salb, M. Mauser, H. Van De Wetering, M. Burch and U. Kloos, "Visual Analysis of Formula One Races," 2019 23rd International Conference Information Visualisation (IV), Paris, France, 2019, pp. 94-99, doi: 10.1109/IV.2019.00025.
- Terms & Conditions – Ergast Developer API. (n.d.). Retrieved October 16, 2023, from http://ergast.com/mrd/terms/
- What is Formula 1? (n.d.). Www.rookieroad.com. https://www.rookieroad.com/formula-1/what-is/
- Y. Liu et al., "Dynamic Path Planning for Formula Automous Racing Cars," 2021 40th Chinese Control Conference (CCC), Shanghai, China, 2021, pp. 6087-6093, doi: 10.23919/CCC52363.2021.9550038.