Title: Sports Analytics for Enhanced Performance Prediction: Premier League and Formula 1 Player and Driver Analysis

Authors: Akshay Krishnan, Jaya Rishita Pasam

Github: https://github.com/akrishnan96/Sports-Analytics-Performance-Prediction

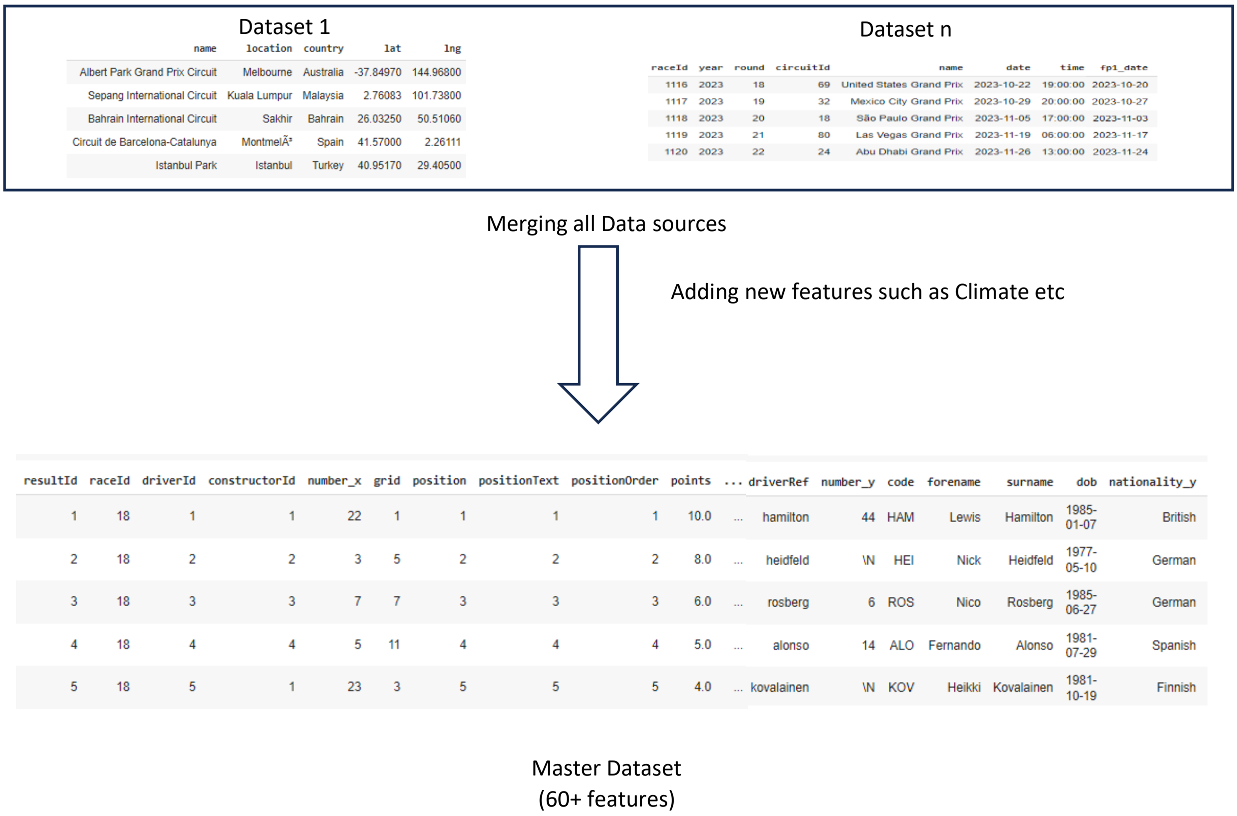
Summary: This project aims to improve performance prediction in the field of sports, with a focus on the Premier League and Formula 1. It does so by leveraging the power of data science and machine learning. The main goal is to analyse driver performance in Formula 1 and predict player performance in the Premier League to shed light on the complex aspects influencing these athletes' achievement. This study is based on historical datasets covering Formula 1 driver performance indicators, such as race positions and lap times, as well as football player statistics, such as goals scored and assists.

Through ethical scraping techniques, data is collected from different data sources, creating extensive datasets that support our analytical procedures. For the Premier League, this data includes past player statistics, match outcomes, and player performance indicators that were obtained from reliable sources, such as the official Premier League website, football analytics websites, and maybe open data repositories. Data related to player participation in other tournaments during the same year will also be collected thus providing a holistic view of their performance. The official data from the Formula 1 website, which is renowned for its dependability in presenting motorsport information, will serve as the dataset's anchor for our analysis of the Formula 1 race. This dataset will include a variety of data about racers, drivers, lap times, race results, and other important contextual information.

F1 Driver Performance

Data Collection

For the initial analysis of F1 data, data was obtained from the Ergast data source(http://ergast.com/mrd/). The data consisted of 14 datasets which contain information on the Formula 1 races, drivers, constructors, qualifying, circuits, lap times, pit stops, and championships from 1950 till the 2023 season. The methodology for the data collection for the F1 data is as follows.



Dataset 1

name	location	country	lat	lng
Albert Park Grand Prix Circuit	Melbourne	Australia	-37.84970	144.96800
Sepang International Circuit	Kuala Lumpur	Malaysia	2.76083	101.73800
Bahrain International Circuit	Sakhir	Bahrain	26.03250	50.51060
Circuit de Barcelona-Catalunya	MontmelÃ³	Spain	41.57000	2.26111
Istanbul Park	Istanbul	Turkey	40.95170	29.40500

Dataset n

raceId	year	round	circuitId	name	date	time	fp1_date
1116	2023	18	69	United States Grand Prix	2023-10-22	19:00:00	2023-10-20
1117	2023	19	32	Mexico City Grand Prix	2023-10-29	20:00:00	2023-10-27
1118	2023	20	18	São Paulo Grand Prix	2023-11-05	17:00:00	2023-11-03
1119	2023	21	80	Las Vegas Grand Prix	2023-11-19	06:00:00	2023-11-17
1120	2023	22	24	Abu Dhabi Grand Prix	2023-11-26	13:00:00	2023-11-24

Merging all Data sources

Adding new features such as Climate etc

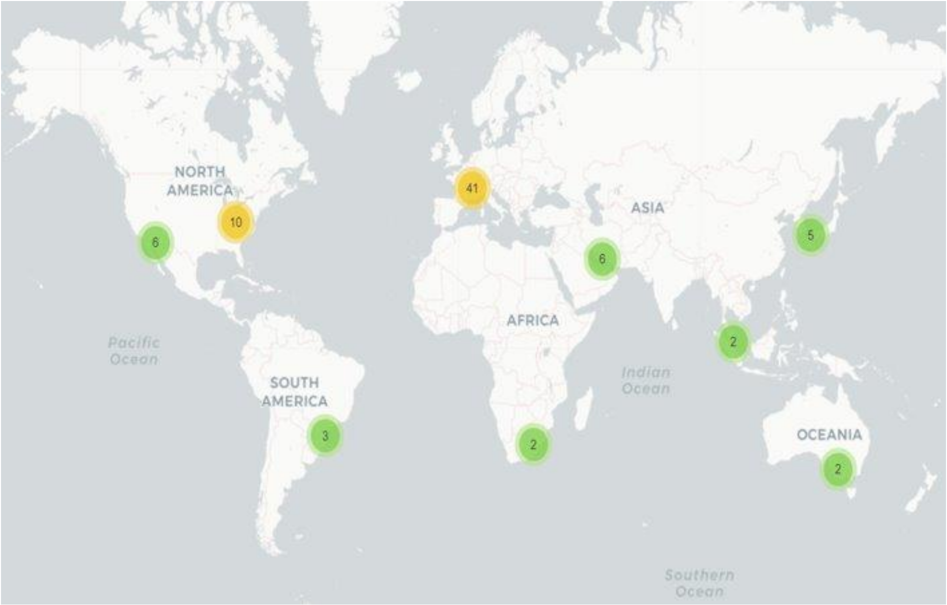
resultId	raceId	driverId	constructorId	number_x	grid	position	positionText	positionOrder	points	...	driverRef	number_y	code	forename	surname	dob	nationality_y
1	18	1	1	22	1	1	1	1	10.0	...	hamilton	44	HAM	Lewis	Hamilton	1985-01-07	British
2	18	2	2	3	5	2	2	2	8.0	...	heidfeld	\N	HEI	Nick	Heidfeld	1977-05-10	German
3	18	3	3	7	7	3	3	3	6.0	...	rosberg	6	ROS	Nico	Rosberg	1985-06-27	German
4	18	4	4	5	11	4	4	4	5.0	...	alonso	14	ALO	Fernando	Alonso	1981-07-29	Spanish
5	18	5	1	23	3	5	5	5	4.0	...	kovalainen	\N	KOV	Heikki	Kovalainen	1981-10-19	Finnish

Master Dataset
(60+ features)

Exploratory Data Analysis:

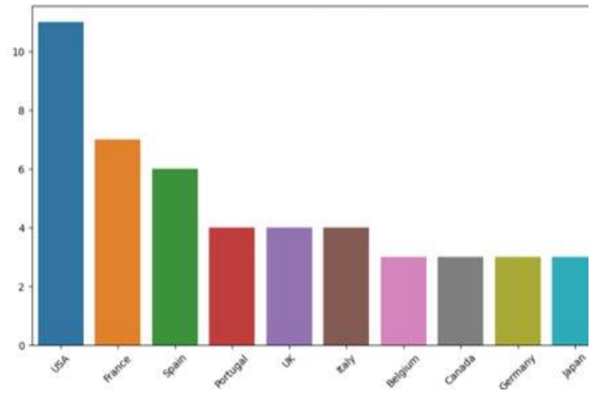
Several visualizations provided insight into our analysis of The Formula 1 data. The goal of these visualizations is to highlight significant trends and conclusions based on the information we have gathered from the Ergast data source. We will examine the variables that have affected the drivers' performances as we navigate these graphs and charts, compare the drivers' lap times, and find interesting patterns that help us understand the variables that affect the drivers' performances and the successful finishes of the teams. Our visualizations will aid in illuminating the narrative underneath the data and offer insightful information regarding championships and their driver's wins.

Plot-1: Circuits around the world

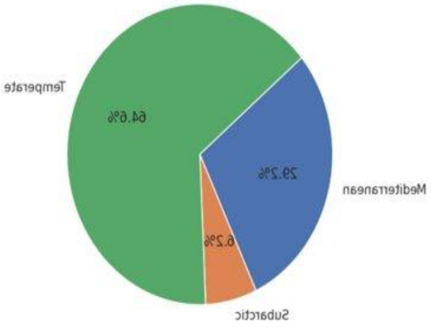


This visualization is essential for understanding the geographic distribution of circuits. It gives a brief description of the locations of circuits, which is useful for evaluating regional coverage and locating regions with a lot of circuits. It is a helpful tool for geographic analysis and circuit placement decision-making.

Plot2: F1 Circuits by Country.



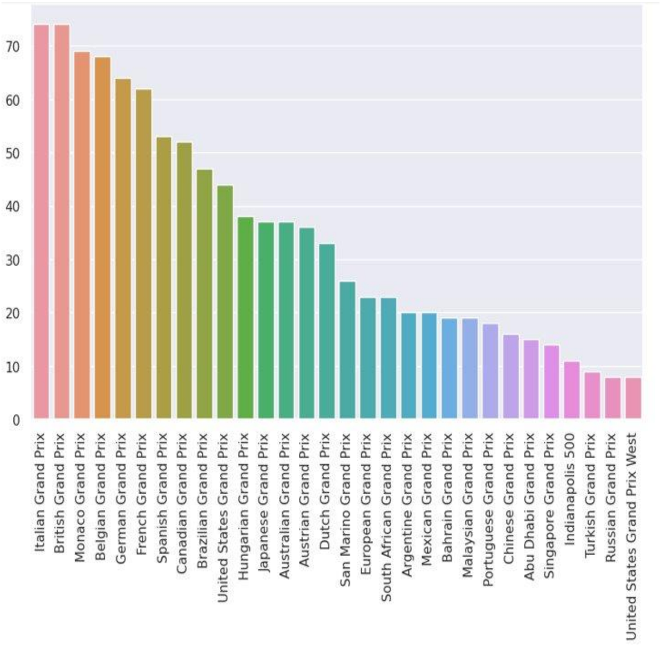
Plot3 : F1 Circuits by Climate:



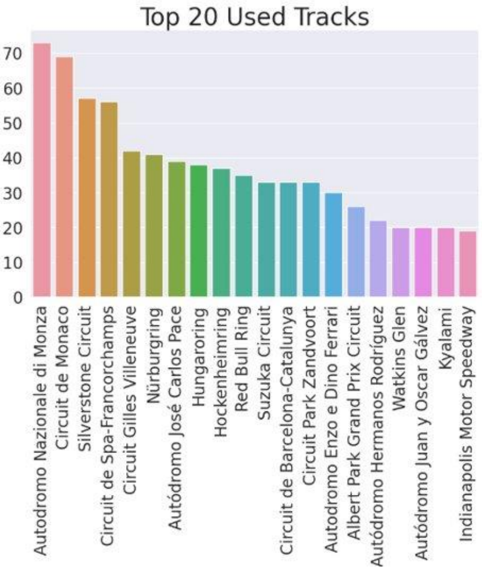
The **PLOT2** a bar graph gives a clear and comprehensive summary of those countries with the most F1 circuits. The USA has the most circuits, with 11, followed by France with 7, Spain with 6, and a number of other countries with varied numbers. The geographic distribution of F1 circuits can be better-understood thanks to this visualization, which can also help with scheduling and marketing decisions.

The **PLOT3** pie chart gives a general summary of the differences in climate amongst F1 circuits. It is clear that the majority of circuits (64.6%), have temperate climates, followed by Mediterranean climates (29.5%) and subarctic climates (6.2%). This knowledge can be useful for planning races, choosing tires, and other possibly climate-dependent decisions.

Plot4: Most Hosted Grand Prix:



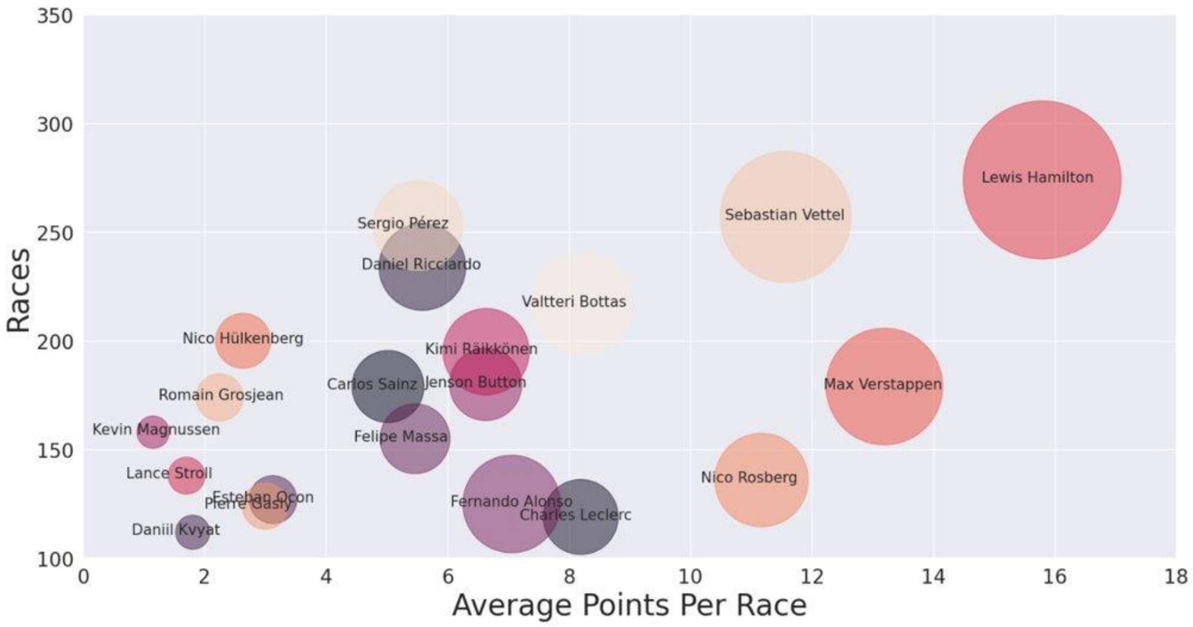
Plot5: Top 20 Used Tracks



The top 30 Formula One Grand Prix venues that have organized the most races are shown in PLOT 4. The height of each bar, which symbolizes a Grand Prix event, shows how many times the Grand Prix has been hosted. Along the x-axis, the Grand Prix races are labelled. Understanding which Grand Prix events have a long history and have been hosted the most frequently is the aim behind this visualization. The Italian Grand Prix and the British Grand Prix both have a long history, having each been hosted 74 times.

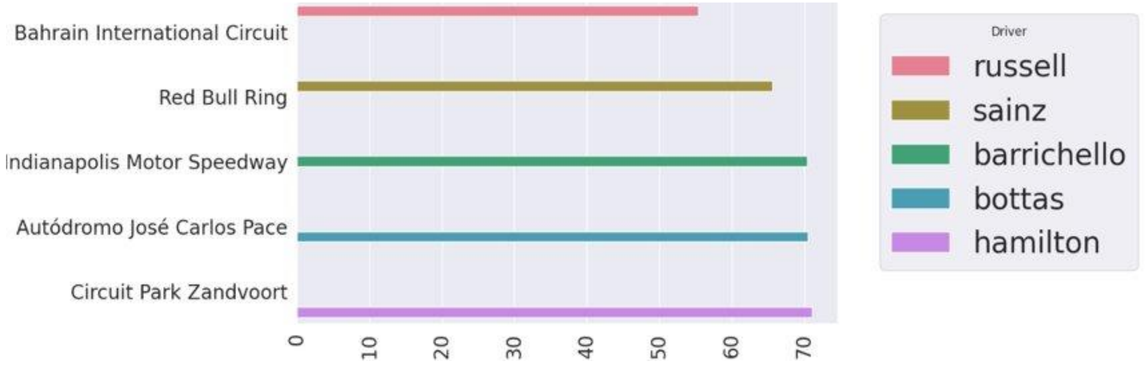
The PLOT5 shows the top 20 Formula 1 tracks in terms of usage over the course of the sport's history. Each bar represents a distinct track, and the height of the bar represents the number of seasons in which that track has hosted F1 competitions. The x-axis contains labels for the tracks.Understanding which tracks have been often used in Formula 1 over the years is made easier with the help of this visualization. With 73 years of use, it is certain that the Autodromo Nazionale di Monza is the most used track.

Plot6: Avg Points by Drivers per Race in the last 15 years:



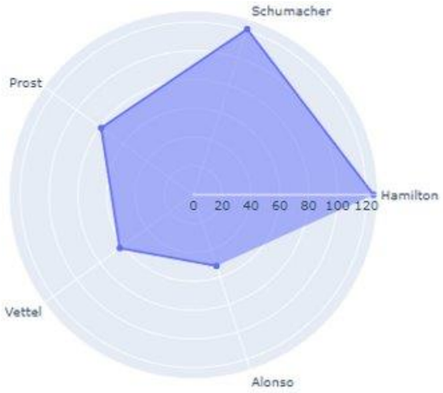
The top 20 Formula 1 drivers over the previous 15 years are shown in this bubble chart by their average points earned per race. Each bubble is a driver, and the number of points they have scored over the course of the session are represented by the size of the bubble. The y-axis displays the number of races the driver has competed in, and the x-axis displays the average points per race.

Plot7: Fastest Lap Times of All Time

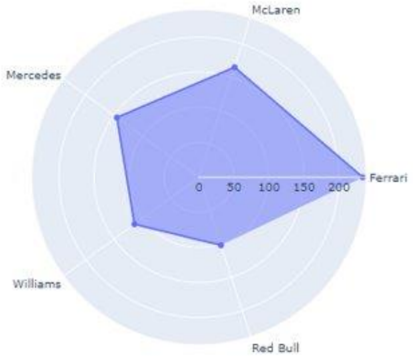


The top 5 quickest laps are shown in this bar graph with their respective lap times by driver and track. The length of each bar, which symbolizes a track, corresponds to the lap duration in seconds. The outside of the plot contains a legend that lists the drivers that each color of the bars represents. When examining driver performance and track characteristics, this information is useful.

Plot8: Top 5 Drivers:



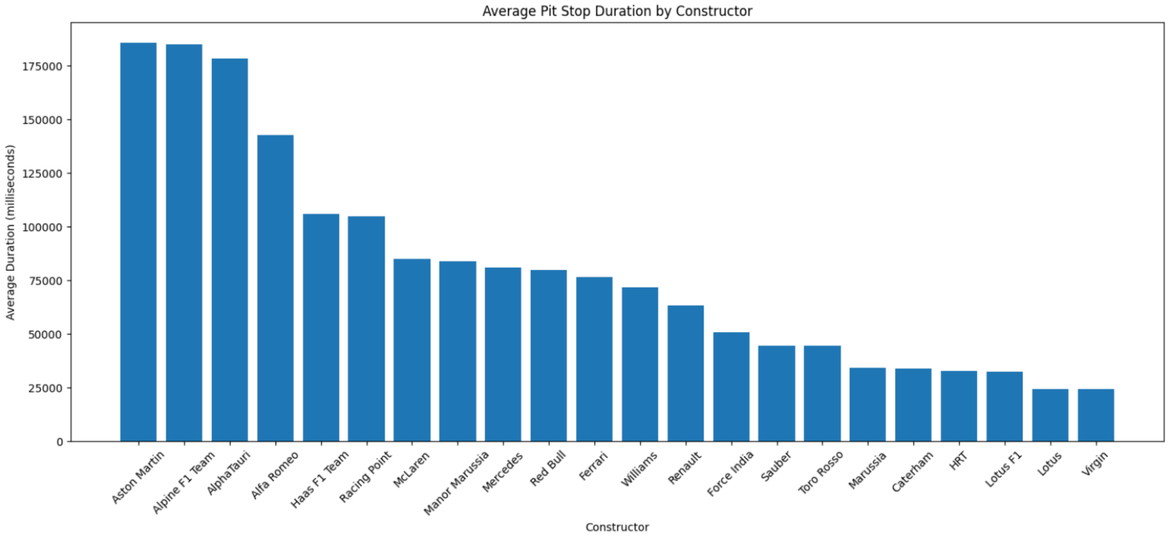
Plot9: Top 5 Constructors:



In these spider plot, The top 5 drivers and teams with the most race is displayed using Plotly.

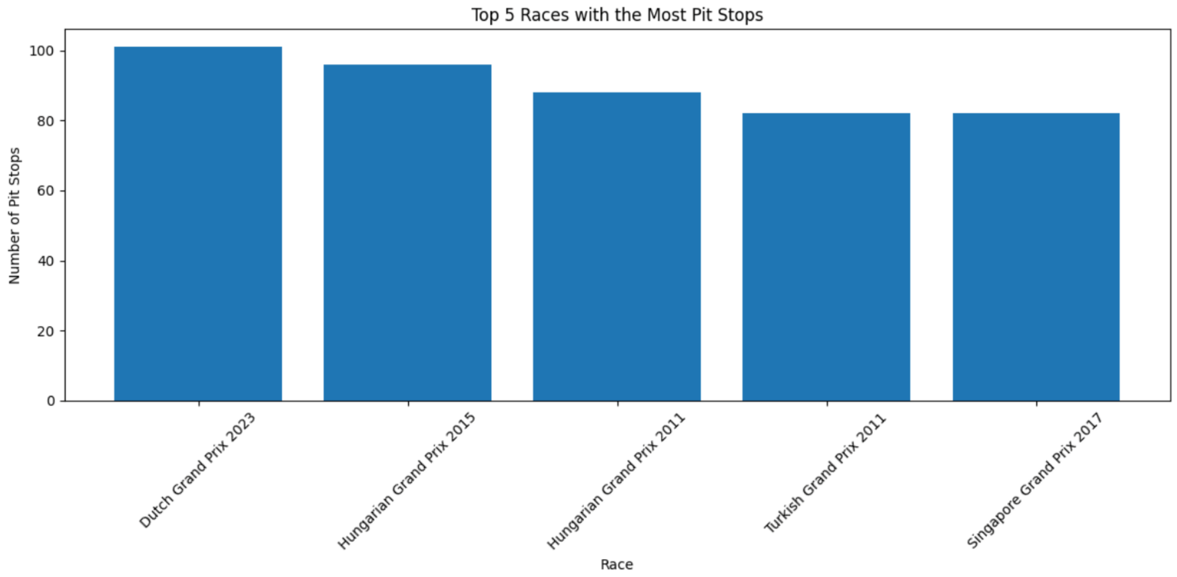
The top 5 drivers with the most appearances in the data are analysed in this visualization. It visually emphasizes the variations in the frequency of appearances among different drivers. For instance, based on the length of their respective arms on the spider plot, it is evident that Lewis Hamilton and Micheal Schumacher have made a substantial number of appearances. The top 5 teams based on the number of constructors championships are shown in this visualization. It shows the differences between the team's performance. For instance, you can see that Ferrari is the constructor with the most championships.

Plot10: Average Pit Stop Duration by Constructor



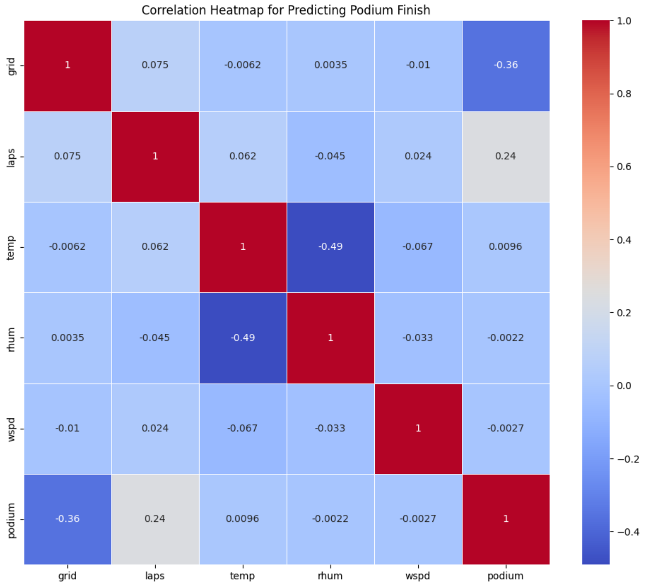
As we know the time taken in a pitstop is determined mostly by the mechanics, this analysis can be helpful in identifying teams that have the most efficient and effective pitstop operations and provides insight into their competitive edge in the racing.

Plot11: Top 5 Races with the Most Pit Stops



This analysis focuses on determining the Formula 1 races with the most pit stops. This bar chart can be used to explore the dynamics of race tactics, tire management, and how race circumstances affect the frequency of pit stops. It provides useful insights into the obstacles and strategies that drivers and teams confront in these races.

Plot12: Correlation HeatMap for Predicting Podium Finish



Grid vs. Podium (-0.361800): A negative correlation indicates that a driver's starting position in front of the grid (lower grid numbers) is significantly linked to a better chance of finishing on the podium. The result is consistent with the competitive advantage obtained by starting at the front of the grid.

Laps vs. Podium (0.240635): A positive correlation means that the more laps a driver completes, the more probable it is that they will finish on the podium. This is an obvious result, as more laps indicate a driver's ability to avoid wrecks or technical difficulties.

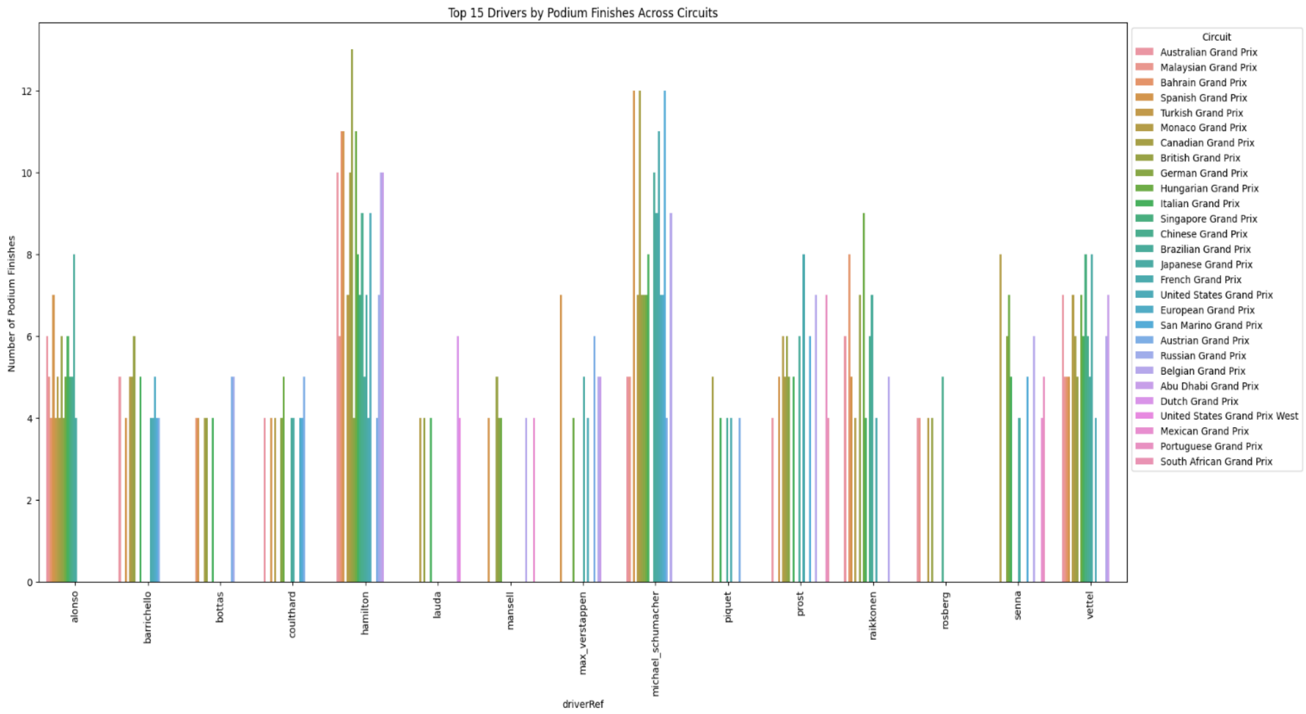
Temperature vs. Podium (0.009579): The almost zero correlation suggests that temperature has a minimal direct effect on podium results, implying that temperature alone is not a good predictor of success.

Rhum (Relative Humidity) vs. Podium (-0.002224): Similar to temperature, relative humidity does not appear to have a substantial bearing on the likelihood of a podium result.

Wspd (Wind Speed) vs. Podium (-0.002695): Wind speed has little to do with podium finishes.

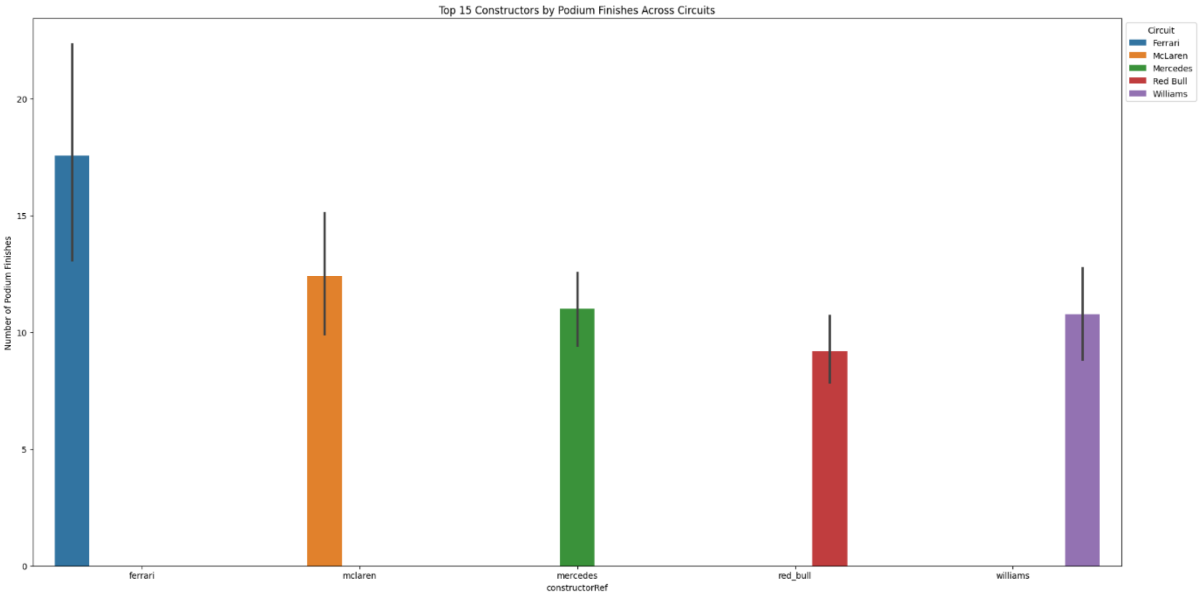
According to the findings, grid position is a strong predictor of podium finishes. The better a driver's starting grid position, the more likely he will finish on the podium. Completed laps also correspond with podium finishes, implying that drivers who complete more laps are more likely to finish the race.

Plot13: Top 15 drivers by Podium Finishes Across Circuits



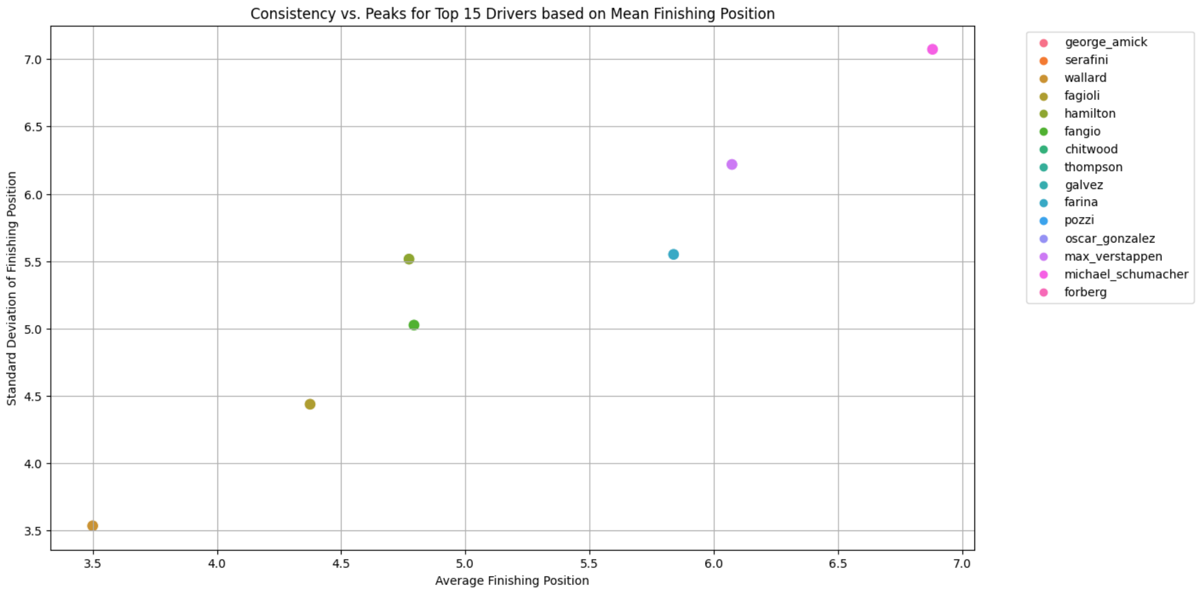
This analysis focuses on the performance of Formula 1 drivers on various circuits, with the goal of determining whether individual drivers or teams excel on specific racetracks. The bar plot shows the top 15 drivers and podium finishes across multiple circuits. It emphasizes drivers' specialization in various circuits, indicating their success in those areas. It sheds light on the strategic strengths of drivers and teams on certain tracks, as well as insights into circuit specialization and variables contributing to competitive advantages.

Plot14: Top 15 Constructors by Podium Finishes Across Circuits



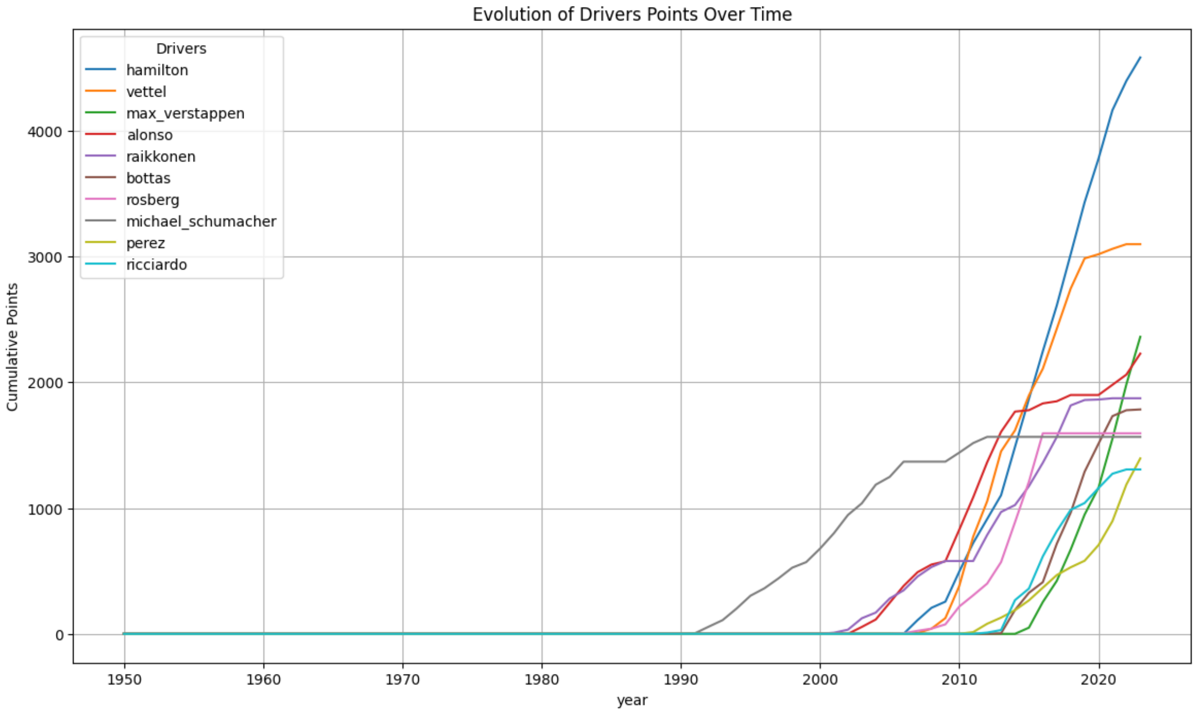
This analysis looks at the performance of Formula One constructors on different circuits to see if any constructors excel on specific tracks. The bar plot, shows the top 15 constructors and podium finishes on various circuits. This graph reveals constructor specialization in specific tracks, showcasing their exceptional performance at these circuits. .

Plot 15: Consistency Vs Peaks for Top 15 Drivers based on Mean Finishing Position



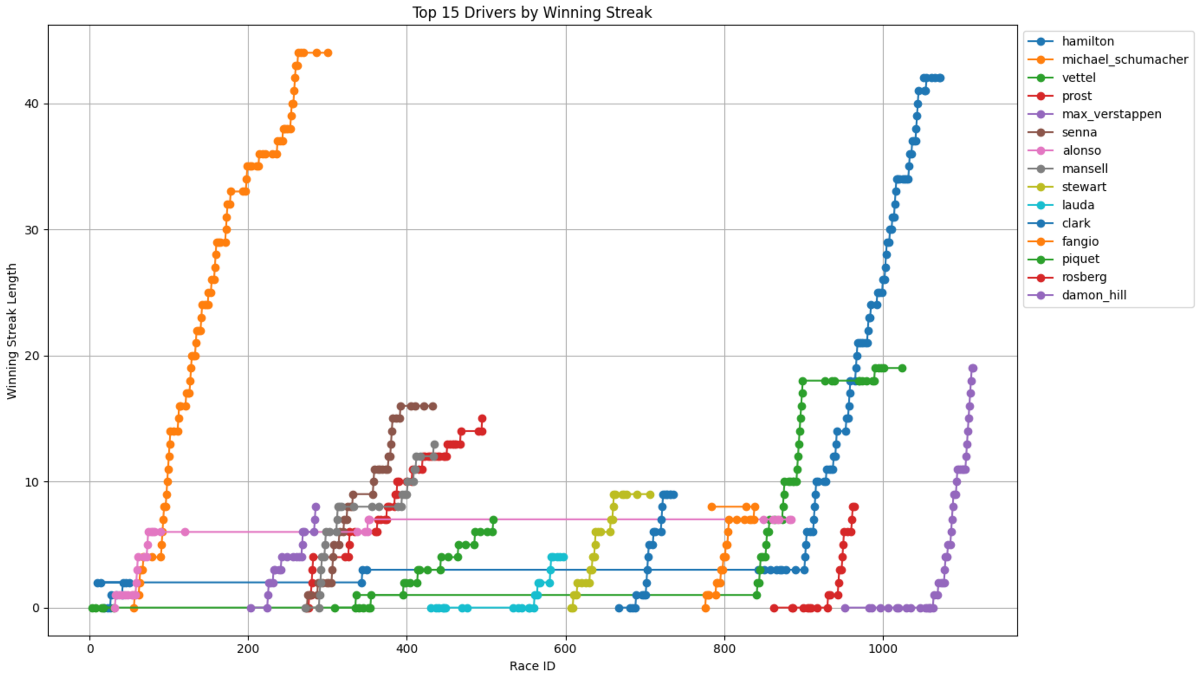
According to the scatter plot, drivers in the bottom-left quadrant are both high-performing and consistent, putting them in the top tier of rivals. Drivers in the top-left quadrant thrive in terms of performance, although their results may have peaks and troughs. Drivers in the bottom-right quadrant are extremely consistent, but their average finishing position is lower. Finally, drivers in the upper-right quadrant had lower average finishes and more unpredictable performance outcomes.

Plot16: Evolution of Drivers Points Over Time



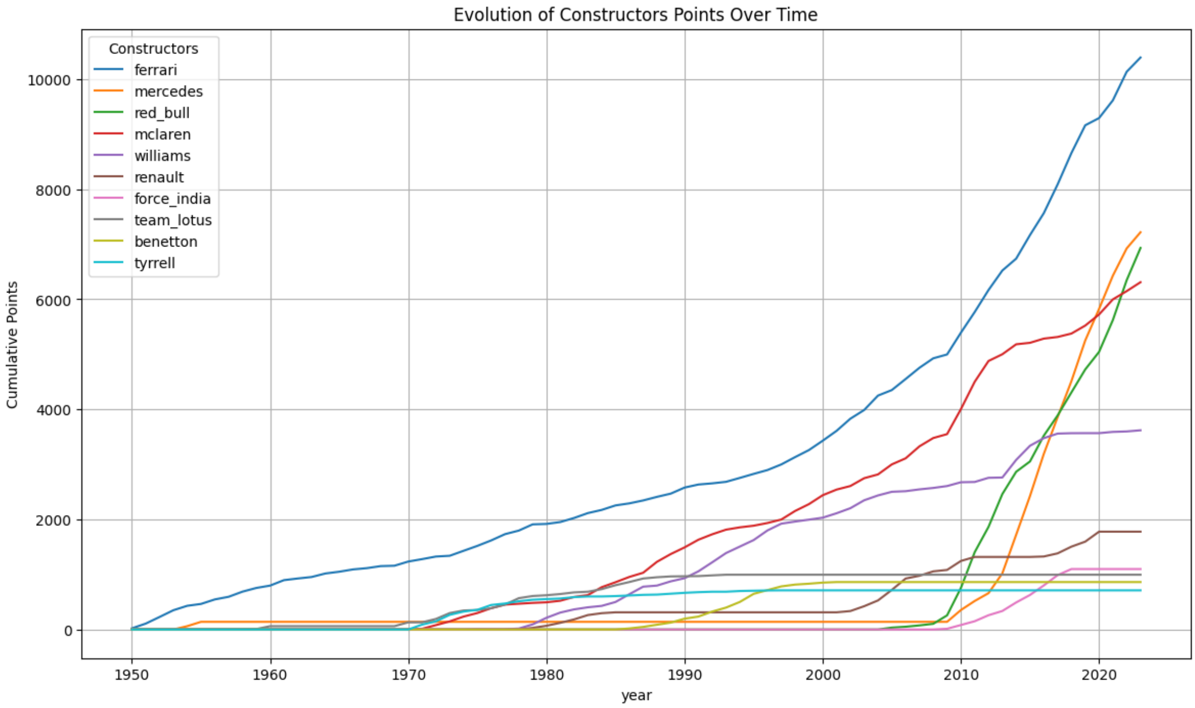
The resulting line chart clearly highlights the best drivers based on their total points, with a focus on the most recent year accessible in the dataset. This shows how drivers' performances have changed over time, emphasizing those who have consistently amassed considerable points throughout their careers

Plot17: Top 15 Drivers by Winning Streak



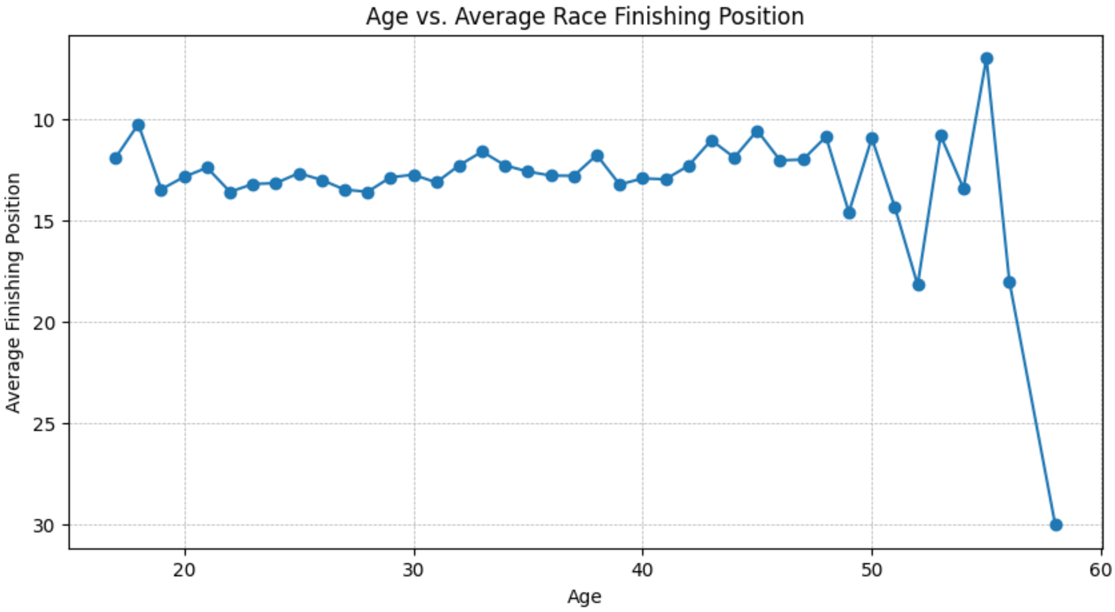
This analysis explores into the winning streaks of Formula 1 drivers, examining the consecutive victories attained by the sport's finest performers. The analysis focuses on finding the top drivers with the most impressive streaks and visualizing their victories. Each driver's name is paired with a different line in the line chart, which depicts their winning streaks over numerous races.

Plot18: Evolution of Constructor Points Over Time



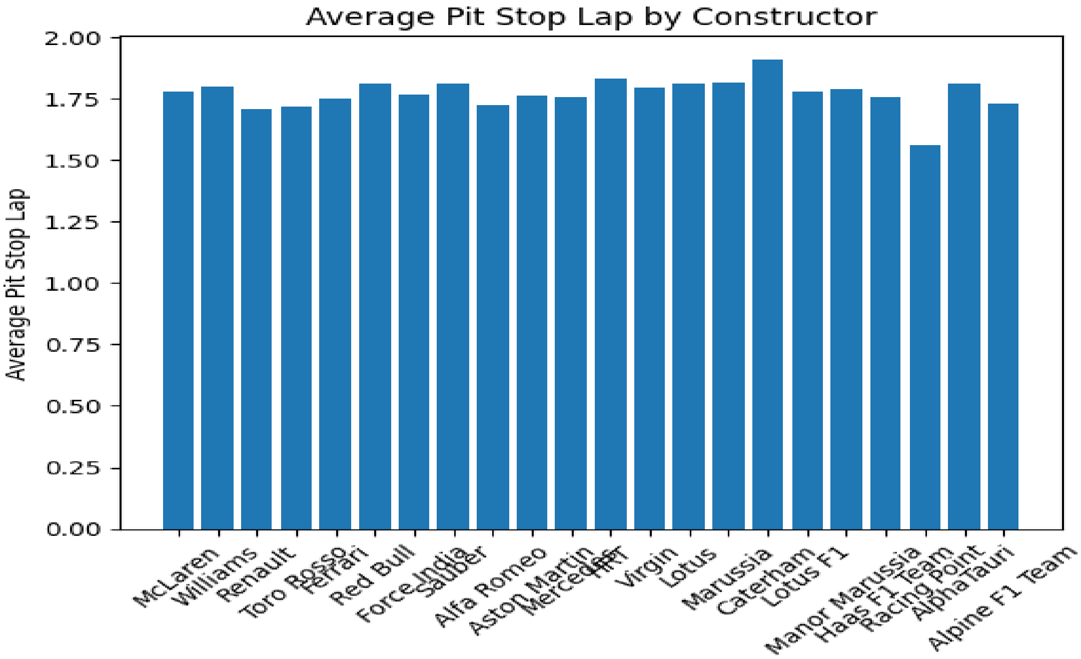
The analysis of the evolution of Formula 1 constructor performance throughout time provides useful insights. The cumulative points graphs clearly demonstrate growth trajectories, reflecting a consistent upward tendency for specific constructors. Another interesting factor shown by this analysis is dominance. Periods in which a constructor routinely finishes first or accumulates significantly more points than competitors emphasize their dominant position in Formula 1.

Plot19: Age vs Average Race Finishing Position



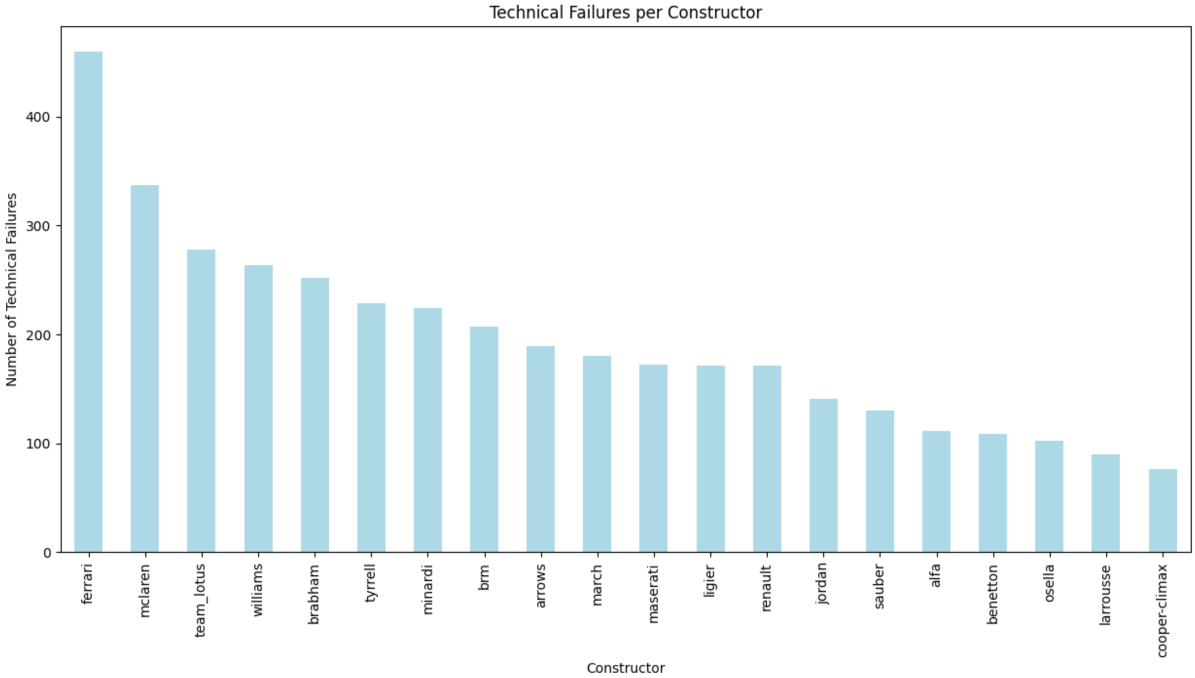
The line chart depicts the relationship between a driver's age and their average race-finishing position. reveals significant insights into the age-performance dynamics in Formula One. It is clear that age plays a role if there is a huge difference in the driver's age.

Plot20: Average Pit Stop Lap by Constructor



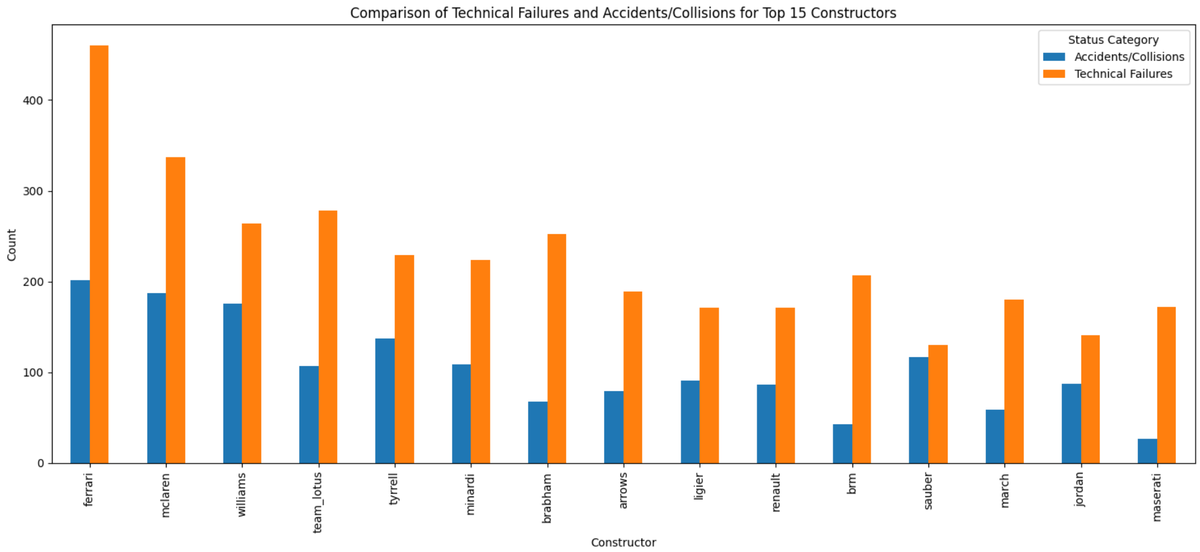
The analysis enables the breakdown of Formula 1 pit stop strategy. Early pit stops may suggest aggressive tactics to obtain a competitive edge via the "undercut." Late pit stops, on the other hand, may indicate endurance-focused strategies or an attempt to profit on tire degradation among competitors.

Plot 21: Technical Failures per Constructor



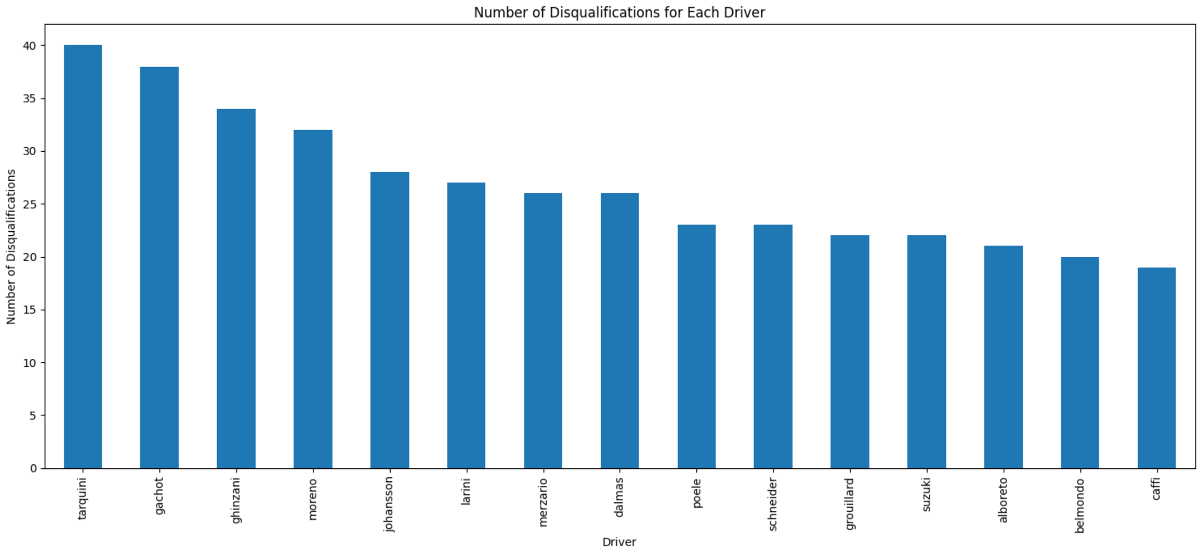
The plot shows the top 20 constructors with the most technical failures, with the number of technical failures on the y-axis and the constructor on the x-axis. This visualization provides insights into the reliability and technical issues that Formula One constructors face during races. It can be a useful tool for teams and fans to evaluate the dependability and performance of various constructors in the sport.

Plot22: Comparison of Technical Failures and Accidents/Collisions for Top 15 Constructors



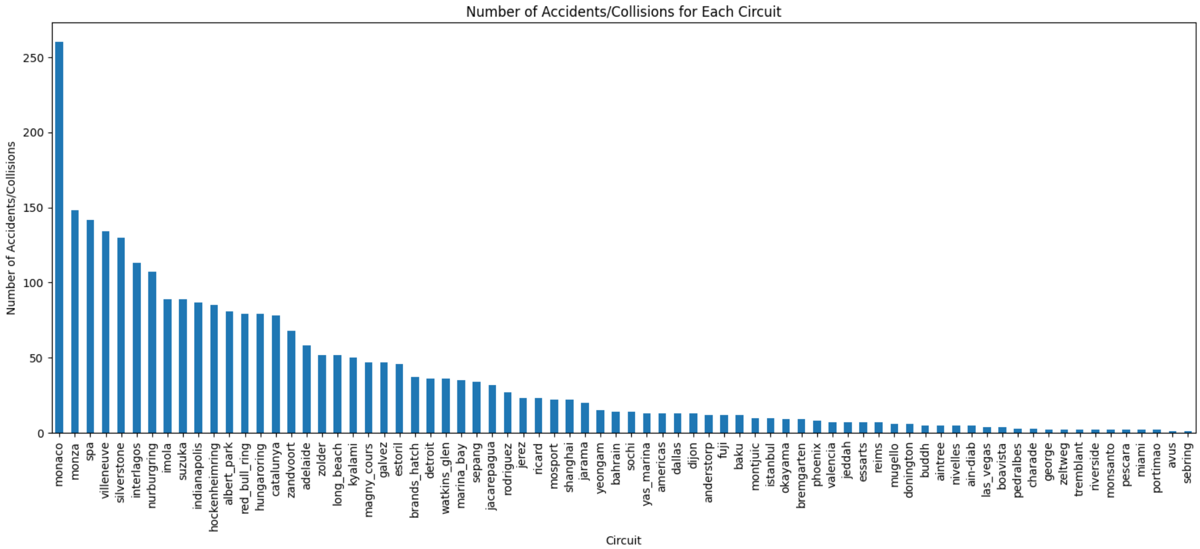
The analysis is a stacked bar plot in which each bar represents a top constructor and the categories "Technical Failures" and "Accidents/Collisions" are layered to show the distribution of these problems. The y-axis represents the count, while the x-axis represents the top builders. This visualization aids in comparing the frequency of technical faults and accidents/collisions for the top 15 Formula 1 constructors.

Plot23: Number of Disqualifications for Each Driver



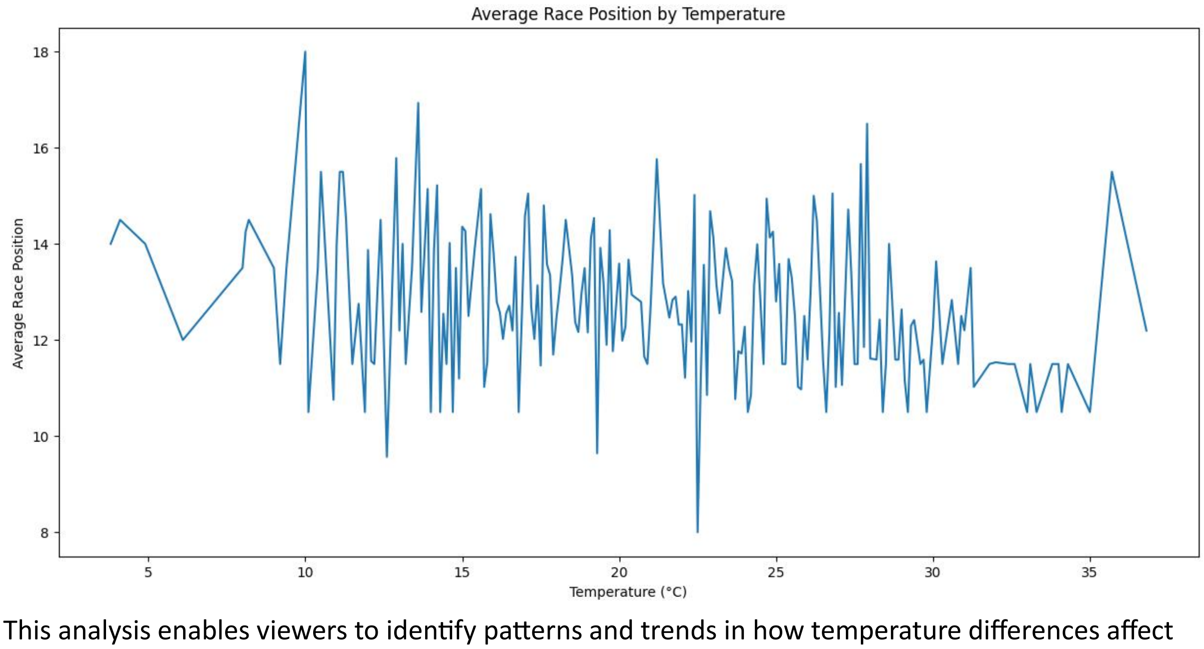
Each driver is depicted on the x-axis of the bar plot, while the number of disqualifications is shown on the y-axis. The bar plot shows the extent of each driver's disqualifications, which can be useful in evaluating their performance and adherence to sports regulations. Understanding the historical records of driver disqualifications is the main goal of this visualization.

Plot 24: Number of Accidents/Collisions for each circuit



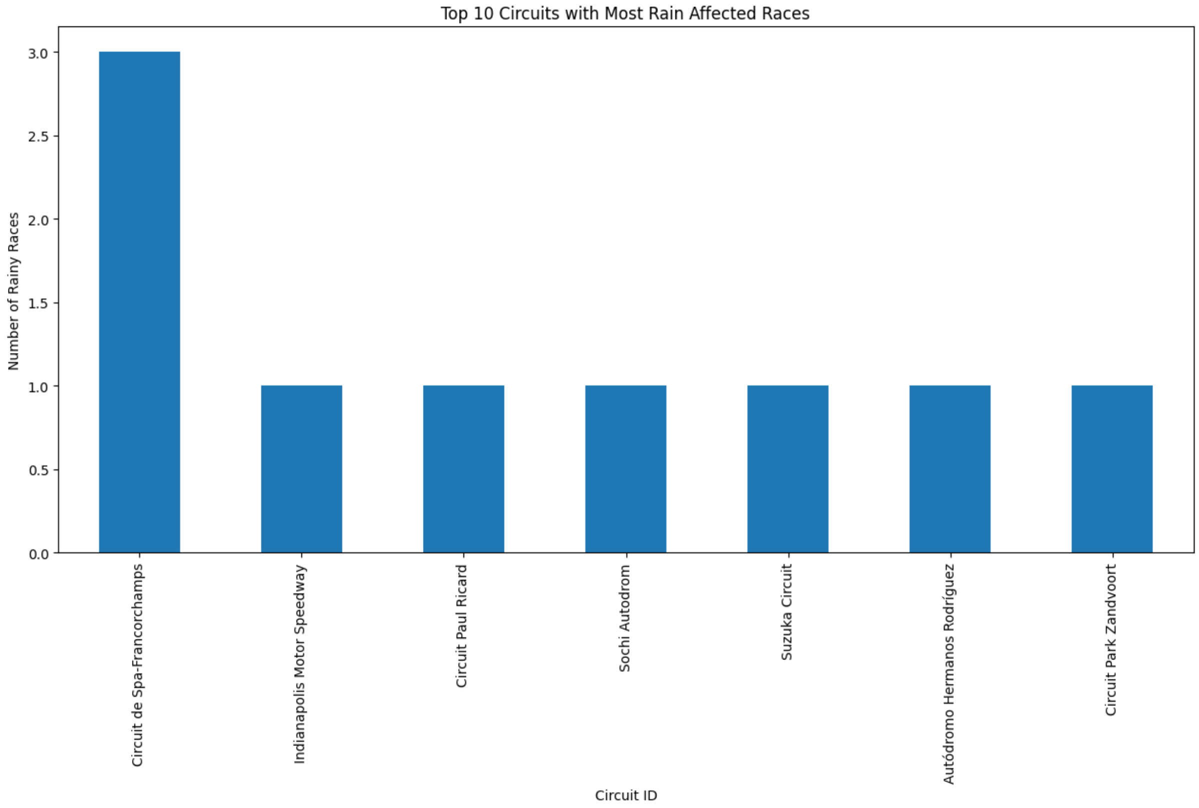
The x-axis of the bar plot represents the circuit, while the y-axis represents the number of accidents and collisions. This bar plot provides a thorough overview of circuit-specific accident statistics, allowing for insights into the safety and performance characteristics of each Formula 1 circuit. It's a great tool for understanding and improving safety in racing.

Plot 25: Average Race Position by Temperature



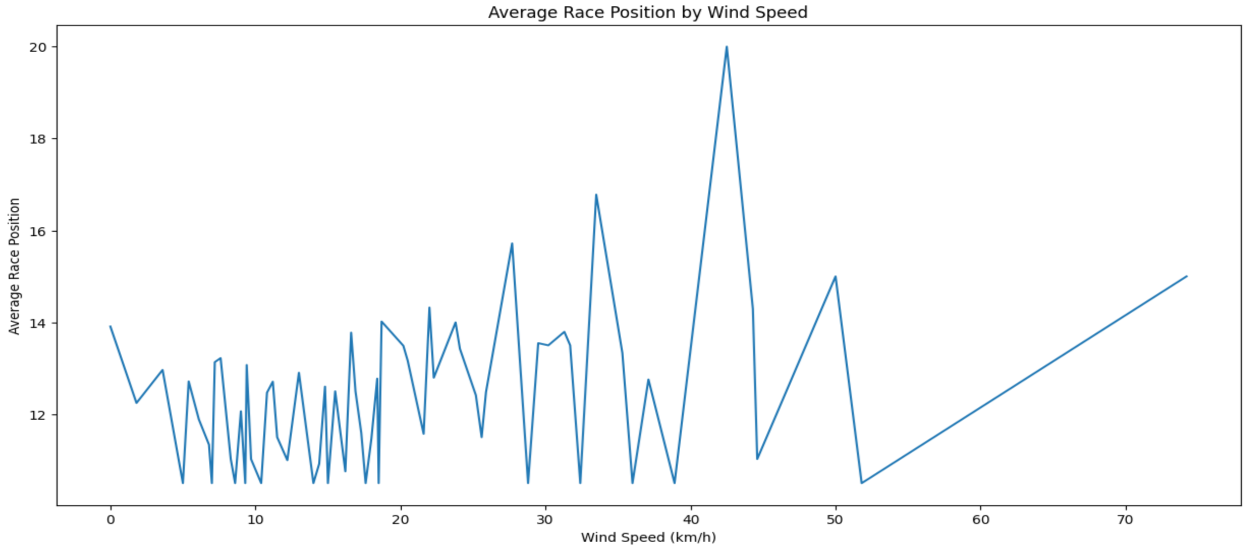
This analysis enables viewers to identify patterns and trends in how temperature differences affect race outcomes. It's a useful tool for understanding the impact of weather on Formula One races, assisting teams and viewers alike in anticipating potential performance differences under various temperature circumstances.

Plot 26: Top 10 Circuits with Most Rain Affected Races



This barchart is intended for analyzing the frequency of rainy races on various Formula One circuits.. The top ten circuits with the most rain-affected races are displayed in a bar chart, providing a visual picture of the circuits that saw the most rain during Formula One events.

Plot 27: Average Race Position by Wind speed



This visualization allows viewers to see if high or low wind speeds have a perceptible impact on Formula 1 driver performance, providing valuable insights into the function of wind conditions in motorsport events.

Data Pre-processing:

1. **Incorporation of Weather Data**

   Influence on Race: Weather conditions can have a considerable impact on race performance and strategy. Variables such as temperature and rainfall play a critical role in determining crucial factors such as tire selection, wear rate, driver strategy, and overall visibility during the race.

   Data Integration: To include the weather conditions into the primary dataset, the 'merge' function was utilized. This integration was performed using the 'raceId' column, which served as a common identifier between the datasets.

2. **Handling Missing Data**

   Circuit-Specific Medians: To address missing weather data, a circuit-centric approach was adopted. Median values of each weather feature based on individual racing circuits was computed. This method ensured that the different weather characteristics inherent in each track were preserved.

   Column Removal: Columns such as snow, wpgt, and tsun, had a large number of missing data. These columns were dropped to prevent any unexpected biases.

   Residual Missing Values: Following these measures, any remaining data gaps in the dataset were filled by imputing the column's median value.

Feature Generation & Engineering:

1. **Driver's Past Performance:**

   To gain insight into the recent form of drivers, several new features were introduced which captured their recent race positions. Example: `driver_avg_position_last_2` and `driver_avg_position_last_5`. These metrics allow for an assessment of both the short-term form and consistency of drivers over a longer period.

2. **Constructor's Past Performance:**

   Similarly, the recent performance trends of constructors were also calculated. Features such as `constructor_avg_position_last_2` and `constructor_avg_position_last_5` were formulated. These metrics illuminate the current form of constructors, serving as valuable predictors for forthcoming races.

3. **Past Podium Finishes:**

   The `past_podiums` metric denotes the number of times a driver or constructor has secured a podium finish. This serves as a testament to their expertise, skill level, and historical achievements in the sport.

Significance of Engineered Features

These newly-engineered features provide a vital historical context, reinforcing the notion that past race outcomes can be indicative of future performances. Through Recursive Feature Elimination (RFE) analysis, the significance of certain features, such as `dwpt`, was further highlighted.

Role of Historical variables

Historical performances are crucial in race predictions. Often, the recent form of a driver or constructor serves as a telling predictor for their prospects in imminent races.

Feature Engineering Insights

The development of these intricate features required an in-depth knowledge of F1 racing and a understanding of the significance of various performance metrics.

Model Development & Evaluation:

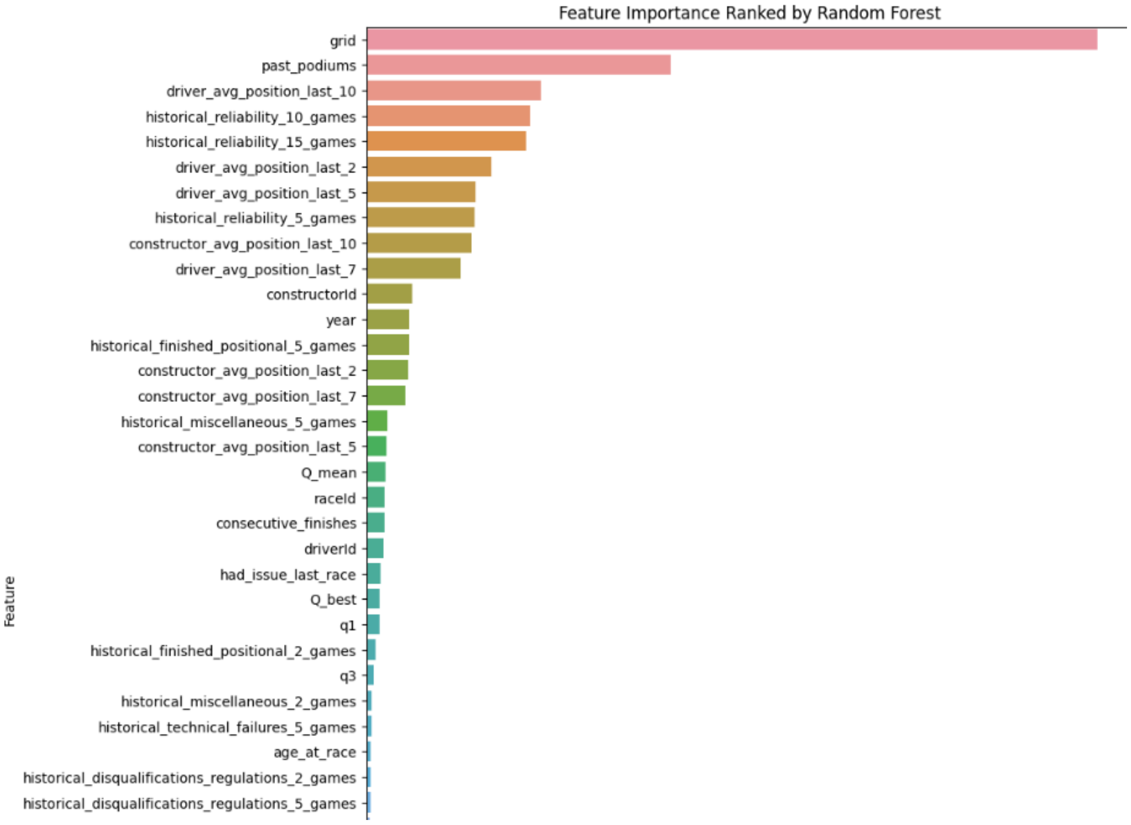
1. **RandomForest Classifier**

   Initial Model Configuration: In the preliminary stages of model creation, the issue of class imbalance by setting the class_weight parameter to 'balanced' was addressed. This approach yielded an accuracy of 81%, with a a notable recall for the minority class.

   SMOTE Implementation: By oversampling the minority class, model performance was optimized, emphasizing the necessity of addressing data imbalance

   Feature Importance Visualization: The significance of the features was visualized using a bar graph, ranking them based on their importance as determined by the RandomForest model. A brief analysis of these results is provided below:

Plot 28: Important variables from RF



Top Features:

- Grid emerges as the key feature, illustrating that a driver's starting position in a race holds significant predictive power for achieving a podium finish
- Past Podiums follows closely, underscoring the notion that historical successes can be indicative of future achievements
- The Driver's Average Position in the Last 10 Races stands out as a vital predictor, mirroring a driver's recent form
- Historical Reliability over intervals of 10 and 15 races resonates with the emphasis on the consistent technical prowess and performance in racing

Other Notable Features:

- Average standings of both drivers and constructors over varied durations (last 2, 5, 7, and 10 races) exhibit relevance, pointing to the predictive worth of both recent and extended performance trends.
- While Qualifying Positions (q1, q2, q3) don't surpass the grid position in terms of importance, they remain influential in the prediction paradigm.
- Metrics such as Consecutive Finishes and Had Issue Last Race serve as markers of reliability, with potential implications for anticipating future race outcomes.

Least Important Features:

- Aspects like Historical Disqualifications/Regulations and Historical Technical Failures over a span of 5 races appear less crucial. The infrequency of such events could account for their diminished predictive capacity
- Age at Race features lower on the scale, implying that a driver's age might not be a potent predictor for top-tier race finishes.

Features to Potentially Exclude:

- Features positioned at the lower end of the importance graph, especially those nearing negligible importance, could be considered for removal when refining the model. It's essential, however, to weigh their potential contextual value as occasionally, seemingly less crucial features might bolster model robustness.

Considerations for Model Building:

- Highly ranked features warrant meticulous handling during the data preprocessing phase to ascertain their integrity.
- To curb overfitting and boost computational efficiency, it might be beneficial to contemplate excluding those features with diminished importance, contingent on model performance.

Weather Data:

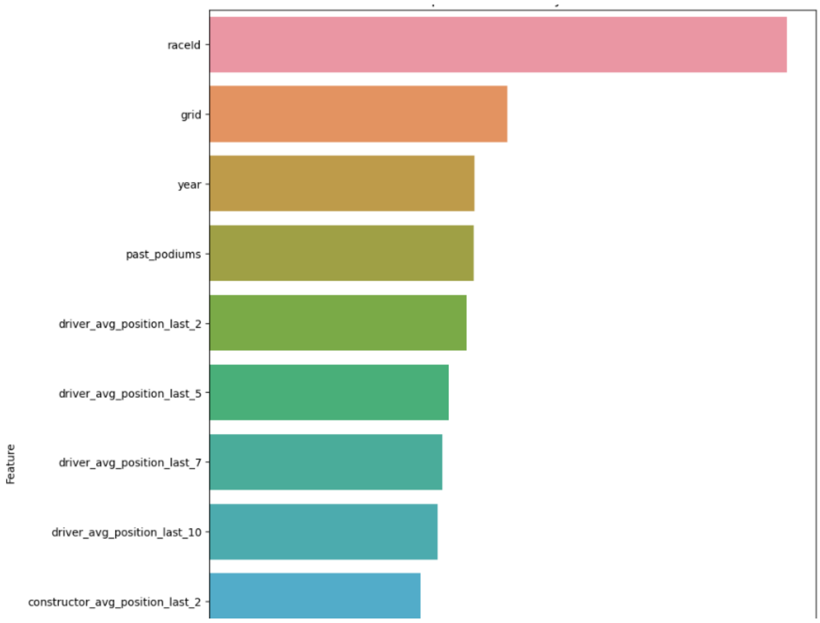
- An intriguing observation is the non-prominent ranking of weather-related features. This could signify a nominal influence of weather on race outcomes, or it might reflect the model's inability to adeptly harness the weather-race relationship.

Keep in mind that feature importance is model-specific. If you change the model or the hyperparameters, the feature importance could change. Additionally, high feature importance does not imply causality—it just means there is a strong association with the target variable within the context of the model.

2. XGBoost Classifier

- SMOTE Balancing: XGBoost was fine-tuned using SMOTE-balanced data. The `scale_pos_weight` parameter was adjusted based on class distribution.
- Weighted Model Approach: A method that prioritizes the minority class by assigning more weight during training, an alternative to oversampling.
- Feature Selection using Recursive Feature Elimination (RFE)
    - A RandomForest classifier was chosen for its ensemble nature, providing averaged insights from multiple decision trees.
    - 13 crucial features were pinpointed, revealing the most influential factors for predicting podium finishes.
    - Model Evaluation with RFE Features: By solely using the 13 significant features identified by RFE, the model attained a 90% accuracy, advocating for feature relevance over quantity.

Plot 29: Important variables from XGBoostClassifier



- In this updated bar graph showing feature importance after RFECV (Recursive Feature Elimination with Cross-Validation), it seems that the number of features has been reduced, focusing on the most relevant ones determined by the Random Forest model in combination with RFECV.

Top Features:

- raceId stands out as the most important feature now, which suggests that the specific race has a high impact on the outcome. This could be due to the unique characteristics of each race, like the track, weather conditions, or other race-specific variables.
- Grid retains its position as a key feature, affirming the importance of the starting position
- Year appears as a significantly important feature, indicating that the time factor or seasonality is crucial—perhaps reflecting changes in teams, cars, regulations, or driver experience.

Performance Features:

- Past Podiums continues to be an important feature, along with the Driver's Average Positions in the Last 2, 5, 7, and 10 Races. These demonstrate that both the historical success of drivers and their recent performance trends are strong predictors.
- Constructor's Average Position in the Last 2 Races also emerges as important, which suggests the recent performance of the team is a significant predictor of race outcomes.

Analysis After Feature Selection:

- The model has become more streamlined with a focus on the most influential features. RFECV has successfully eliminated redundant or less informative features, potentially leading to a more robust and generalizable model.
- Simplifying the model in this way can also make it more interpretable and often improves performance by reducing the risk of overfitting.

Considerations for Model Refinement:

- It is important to consider the possibility of data leakage, especially with features like `raceId`, which should generally not have predictive power unless it is encoding information such as time (for time series) or other race-specific information that would be known before the race starts.
- Further model validation is necessary to ensure these features truly have predictive power and are not artifacts of the feature selection process or data-specific circumstances.
- This condensed view of feature importance after RFECV should guide you in further refining the model, potentially improving both its performance and interpretability.

Discussion

1. Influence of Weather Data: While weather plays a vital role, not all weather features proved valuable. Some were excluded due to excessive missing values.
2. Challenges: Class imbalance posed significant challenges, necessitating techniques like `class_weight` balancing and SMOTE for improved model results.
3. Significance of Feature Selection: The RFE method was instrumental in filtering out noise and centering on essential features, resulting in enhanced model performance and interpretability.

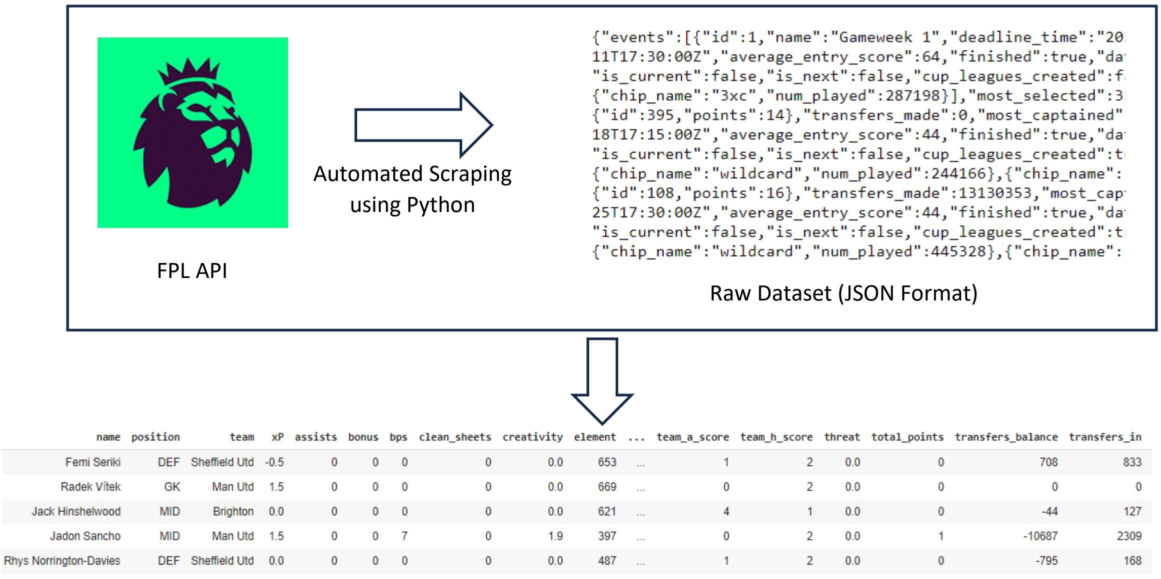
Conclusion

Using sophisticated machine learning methods, a predictive model for F1 racing podium finishes was successfully crafted. The endeavor underscored the value of meticulous data preprocessing, managing class imbalance, and judicious feature selection. Tree-based models, namely RandomForest and XGBoost, demonstrated their efficacy in this predictive task.

Premier League Player Performance

Data Collection and Pre-Processing

Data was obtained from the Official Fantasy Premier League API where we obtained player stats such as Games Played, Goals Scored, Minutes Played, Goals, Assists, Shots Taken and Shots on Goal. More player performance and team performance related variables are currently being obtained from multiple other sources such as commentary, other league websites and so on. The current methodology for extracting the data is as shown below.



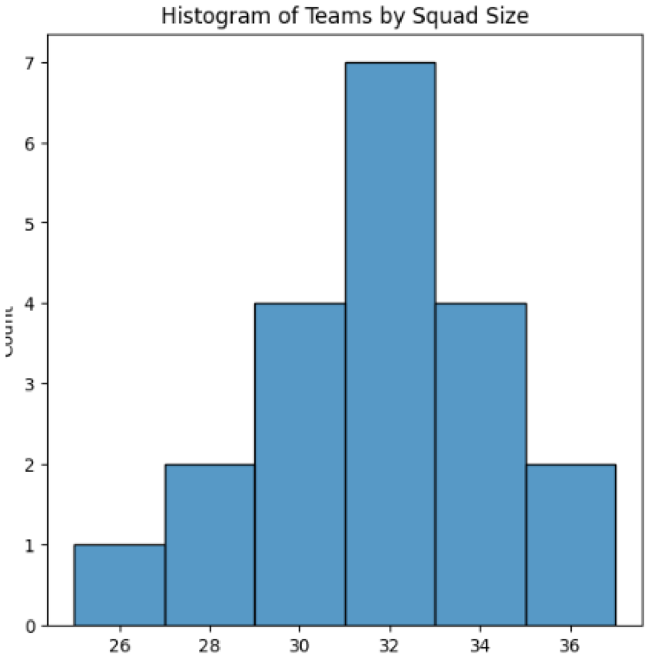
```
{"events":[{"id":1,"name":"Gameweek 1","deadline_time":"20
11T17:30:00Z","average_entry_score":64,"finished":true,"da
"is_current":false,"is_next":false,"cup_leagues_created":f
{"chip_name":"3xc","num_played":287198}],"most_selected":3
{"id":395,"points":14},"transfers_made":0,"most_captained"
18T17:15:00Z","average_entry_score":44,"finished":true,"da
"is_current":false,"is_next":false,"cup_leagues_created":t
{"chip_name":"wildcard","num_played":244166},{"chip_name":
{"id":108,"points":16},"transfers_made":13130353,"most_cap
25T17:30:00Z","average_entry_score":44,"finished":true,"da
"is_current":false,"is_next":false,"cup_leagues_created":t
{"chip_name":"wildcard","num_played":445328},{"chip_name":
```

name	position	team	xP	assists	bonus	bps	clean_sheets	creativity	element	...	team_a_score	team_h_score	threat	total_points	transfers_balance	transfers_in
Femi Seriki	DEF	Sheffield Utd	-0.5	0	0	0	0	0.0	653	...	1	2	0.0	0	708	833
Radek Vítek	GK	Man Utd	1.5	0	0	0	0	0.0	669	...	0	2	0.0	0	0	0
Jack Hinshelwood	MID	Brighton	0.0	0	0	0	0	0.0	621	...	4	1	0.0	0	-44	127
Jadon Sancho	MID	Man Utd	1.5	0	0	7	0	1.9	397	...	0	2	0.0	1	-10687	2309
Rhys Norrington-Davies	DEF	Sheffield Utd	0.0	0	0	0	0	0.0	487	...	1	2	0.0	0	-795	168

Master Dataset
(40+ features)

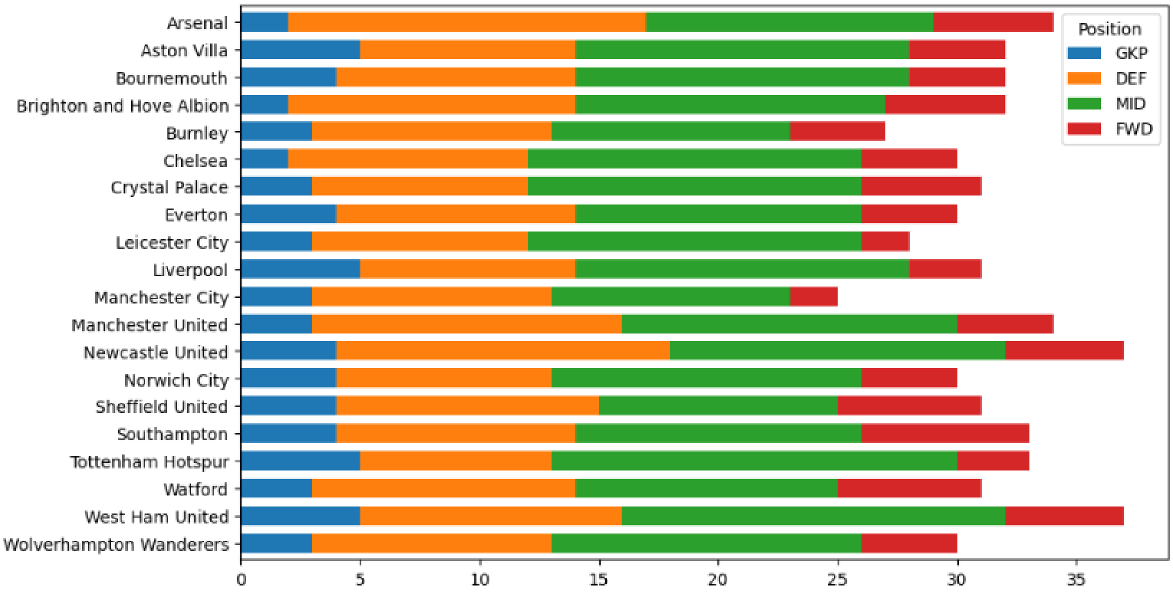
Exploratory Data Analysis

Plot30: Distribution of Squad Size across the 20 teams



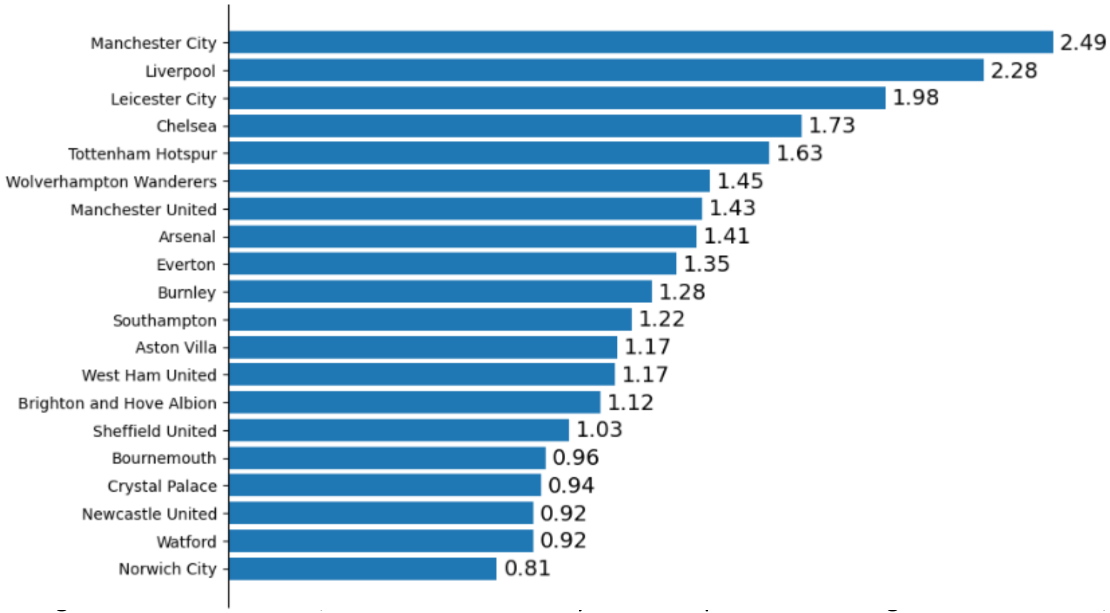
On average, teams typically maintain squads comprising approximately 30 players, which is an essential context to consider. It is worth noting that for each game, teams field a 'match-day squad' consisting of 18 players, including 11 starters and 7 substitutes. Nonetheless, there is noticeable variability among teams, as some operate with as few as 25 players in their squads. This information provides valuable insights into the diverse squad sizes across different teams.

Plot31: Distribution of Player Position by team



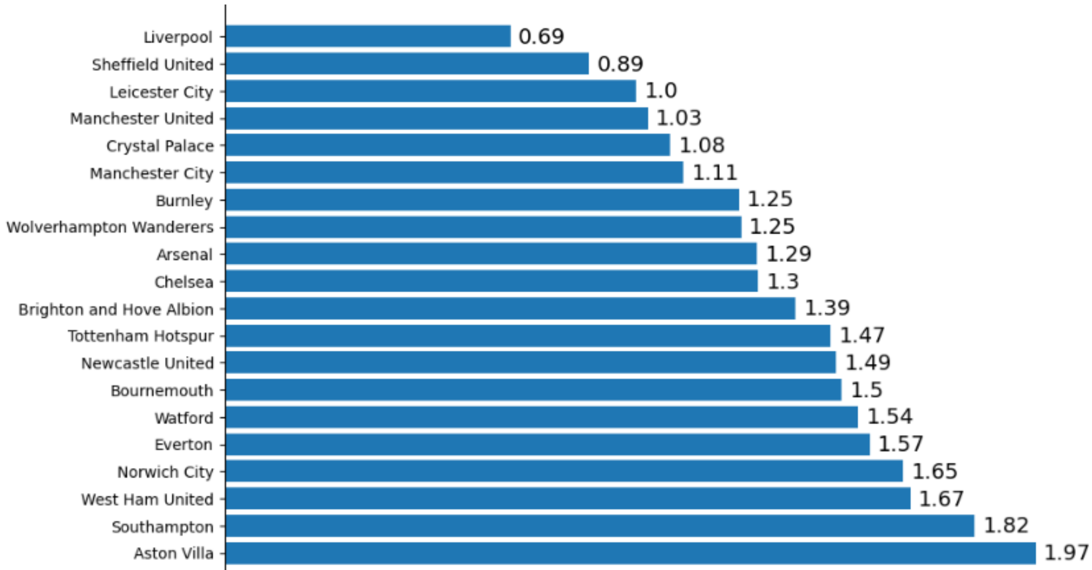
When analysing the composition of player squads, it becomes evident that midfielders and defenders are the prevailing roles within these squads. This observation aligns with expectations, as goalkeepers are understandably the least frequent player type since each team can field only one goalkeeper per game. This shows the distribution of positions within teams and the significance of maintaining a balance between various player roles for a successful team performance.

Plot32: Mean Goals Scored per Game



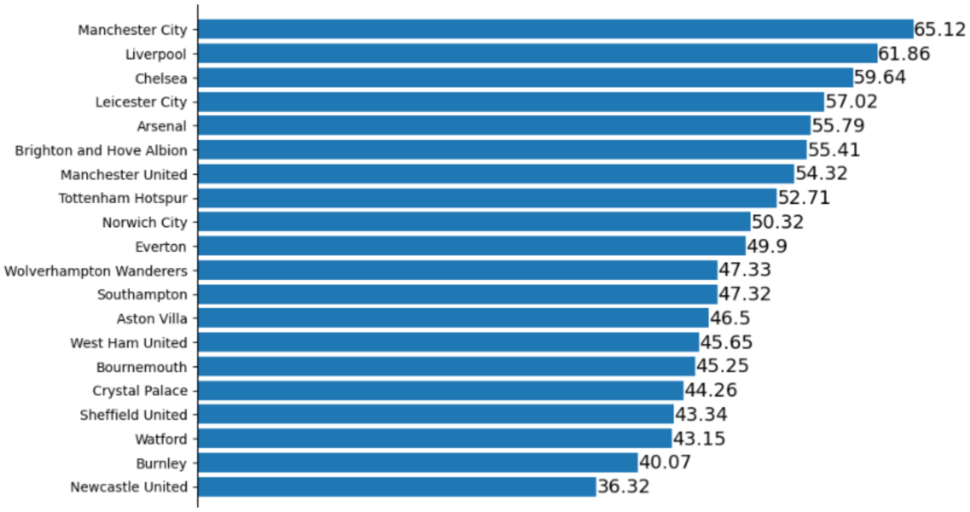
ıester
; per
h just
ss
vhile
Watford and Norwich face more challenging circumstances in this aspect.

Plot33: Mean Goals Conceded per Game



The grapn provides insignts into tne average goals conceded per game, revealing tnat Liverpool stands out with an impressive defensive record of just 0.69 goals per game. Conversely, Sheffield United concedes slightly more at 0.89, while Aston Villa struggles defensively, allowing an average of 1.97 goals per game. Southampton also faces defensive challenges, conceding 1.82 goals per game, and Norwich City concedes an average of 1.65 goals per game. Clearly, maintaining a solid defence is crucial in football, and Liverpool and Sheffield United have excelled in this aspect. In contrast, newly promoted Norwich City and Aston Villa have faced challenges in keeping goals out. These statistics shed light on the contrasting defensive performances of these teams, highlighting the significance of defensive prowess in the league.

Plot34: Possession per Game



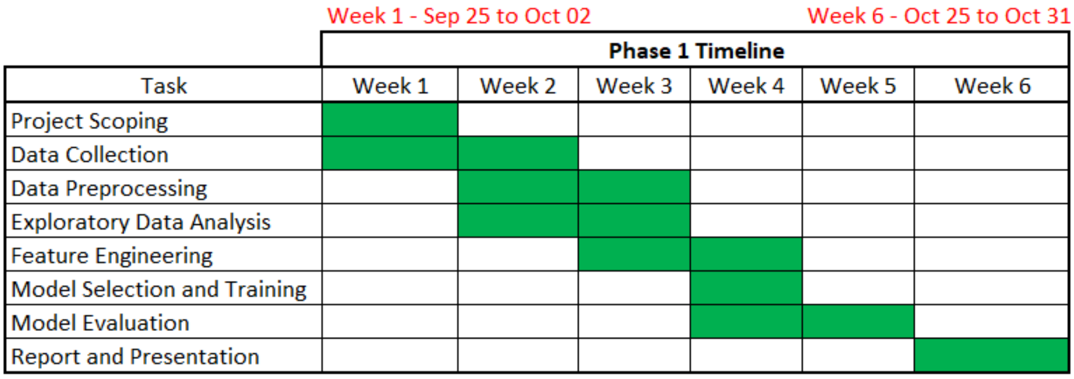
The data presents the average possession per game for several teams, with Manchester City leading at 65.12%, followed closely by Liverpool at 61.86% and Chelsea at 59.64%. In contrast, Newcastle holds 36.32%, Burnley at 40.07%, and Watford at 43.15%. Possession plays a pivotal role in a team's ability to create goal-scoring opportunities, as, without the ball, chances are limited. Various tactical approaches have emerged in the game that diminishes the emphasis on ball possession, often favoured by lower-tier teams. These teams may focus on counterattacking strategies, capitalizing on opportunities after absorbing pressure from their opponents. It's evident that possession and the number of shots taken in a match are closely correlated, but the link between possession and actual goals scored is less straightforward. Wolves, and to some extent, Liverpool, exemplify this by exhibiting clinical efficiency in converting possession into goals, outperforming what their possession statistics might predict. This highlights the complexity of football tactics and the diverse strategies employed by different teams to achieve their goals on the field.

Next Steps - Phase Two

1) Extracting more player related information from other leagues as well since player form depends across different stages of the game
2) More EDA for the PL Data and Feature Engineering
3) Build base model for predicting the form of the PL players
4) Scraping Twitter Data as a part of Sentiment Analysis for both Premier League player data and F1 driver
5) EDA based on the extracted sentiment data

6) Incorporate the sentiments within the existing features and then start the model building and evaluation phase

Project Progress



Week 1 - Sep 25 to Oct 02

Week 6 - Oct 25 to Oct 31

	Phase 1 Timeline					
Task	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6
Project Scoping	🟩					
Data Collection	🟩	🟩				
Data Preprocessing		🟩	🟩			
Exploratory Data Analysis		🟩	🟩			
Feature Engineering			🟩	🟩		
Model Selection and Training				🟩		
Model Evaluation				🟩	🟩	
Report and Presentation						🟩

References:

1) Formula 1 World Championship dataset (1950 - 2020) Retrieved from: http://ergast.com/mrd/

2) Premier League. (2023). Premier League Data. Retrieved from https://www.premierleague.com/stats

3) Fantasy Premier League. (2023). Premier League Data. Retrieved from https://fantasy.premierleague.com/api/bootstrap-static/

4) R. Pariath, S. Shah, A. Surve and J. Mittal, "Player Performance Prediction in Football Game," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 1148-1153, doi: 10.1109/ICECA.2018.8474750

5) S. Manish., V. Bhagat and R. Pramila, "Prediction of Football Players Performance using Machine Learning and Deep Learning Algorithms," 2021 2nd International Conference for Emerging Technology (INCET), Belagavi, India, 2021, pp. 1-5, doi: 10.1109/INCET51464.2021.9456424

6) Stoppels, Eloy. Predicting race results using artificial neural networks. MS thesis. University of Twente, 2017.

7) Accessing Fantasy Premier League data using Python. Retrieved from https://medium.com/analytics-vidhya/getting-started-with-fantasy-premier-league-data-56d3b9be8c32