**Football Matches Analysis Report.**

**Introduction.**

Football is the world’s most popular sport and is often described as a game of passion, skill, and strategy. Beneath the excitement on the pitch lies a wealth of data that can reveal patterns that are not always visible to the casual fan. Every match produces numbers such as goals scored, fouls committed, possession percentages, and passes completed. Together, these statistics tell the story of how teams win, lose, or draw.

This project analyzes a dataset of football matches to uncover these stories through the lens of data science. By examining performance indicators such as goals, possession, fouls, cards, and shooting accuracy, the analysis seeks to answer important questions. Do home teams truly have an advantage? Which teams are most efficient in converting chances into goals? How do discipline and possession influence match outcomes?

Through exploratory data analysis, visualizations, and statistical insights, the project highlights key trends in team performance and discipline. The goal is to understand what drives success in football while also showing how data science techniques can be applied to sports analytics, transforming raw numbers into meaningful insights about the beautiful game.

**Problem Statement.**

The objective of this project is to analyze football match data to understand the factors that influence team performance and match outcomes. The aim is to explore patterns within the dataset, identify the most significant performance indicators, and evaluate how elements such as goals, possession, fouls, cards, and shooting accuracy shape the results of football matches.

Through this process, the project demonstrates the application of the data science workflow, including data exploration, cleaning, feature creation, and visualization. The final goal is not only to extract insights into what drives success on the football pitch but also to showcase how data science methods can transform raw match statistics into meaningful conclusions that explain the dynamics of the game.

**Scope of the Analysis.**

This project focuses on analyzing football match statistics to evaluate patterns in performance and match outcomes. It explores variables such as goals, possession, fouls, yellow and red cards, passes, and shooting accuracy to identify how these factors contribute to success or failure on the pitch. The analysis is limited to the variables provided in the dataset and does not include external data sources such as player-level information or additional league data. The scope is restricted to descriptive and exploratory analysis rather than predictive modeling, with the aim of uncovering insights and trends that explain how teams perform under different conditions.

**Dataset Description.**

**Overview of the Data**.  
The dataset used in this analysis contains detailed records of football matches, capturing both home and away team statistics. It includes information such as goals scored, shots attempted, shots on target, possession percentages, fouls committed, yellow and red cards, passes completed, offsides, corners, and match dates. Each record represents a single match with corresponding statistics for both teams.

The dataset consists of 20 matches and 25 variables. These variables provide a comprehensive view of team performance, covering attacking efficiency, defensive discipline, and control of play through possession and passing. The structured format of the data makes it suitable for exploratory data analysis and visualization, enabling the identification of key patterns and performance indicators across matches.

**Data Characteristics.**

The dataset includes both numerical and categorical variables. Numerical variables consist of match statistics such as goals scored, shots attempted, shots on target, possession percentages, fouls committed, passes completed, and corners. Categorical variables include team names, match results, and league identifiers. The dataset is complete, with no missing values or duplicates, which allowed for direct analysis without extensive imputation.

To enhance the analysis, additional engineered features were created. These included total goals scored in a match, categorical classification of results as home win, away win, or draw, and shooting accuracy metrics derived from shots and shots on target. These engineered variables provided new perspectives for evaluating performance and identifying the most influential factors shaping match outcomes.

**Data Preprocessing.**

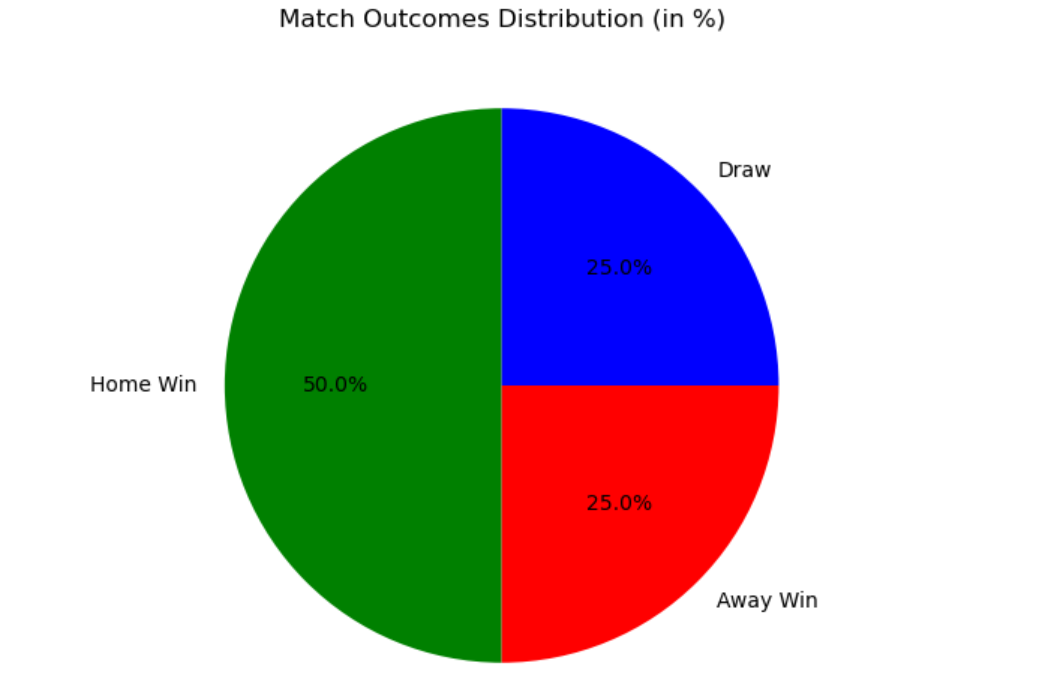
Before analysis, the dataset was prepared through several preprocessing steps to ensure data quality and consistency. The match date column was converted into a proper datetime format to allow for time-based analysis. Duplicate entries were checked, and no inconsistencies were found. New variables were engineered, including total goals per match, categorical match results classified as home win, away win, or draw, and shooting accuracy calculated as the ratio of shots on target to total shots. These steps ensured that the dataset was clean, well-structured, and ready for exploratory data analysis and visualization.

**Visual Analysis with Bar Plots.**

The first bar plot shows the distribution of match outcomes. It highlights that home teams won half of the matches, while away teams won a quarter, and the remaining quarter ended in draws. This demonstrates the clear advantage that home teams often have in football.

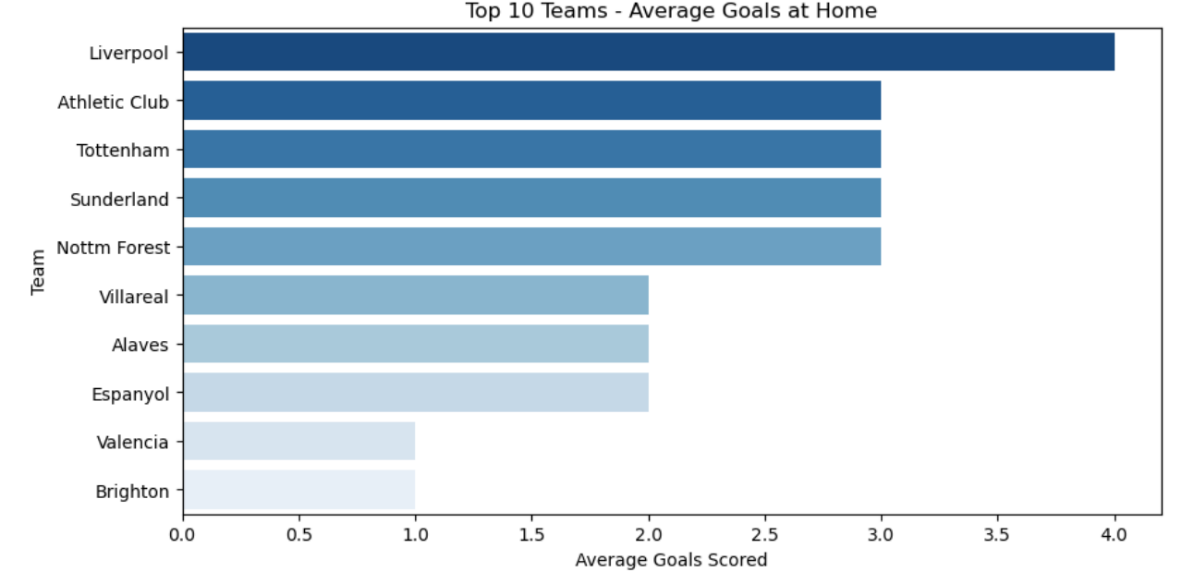
*Distribution of Match Outcomes.*

The pie chart that shows Home Wins, Away Wins, and Draws percentages.



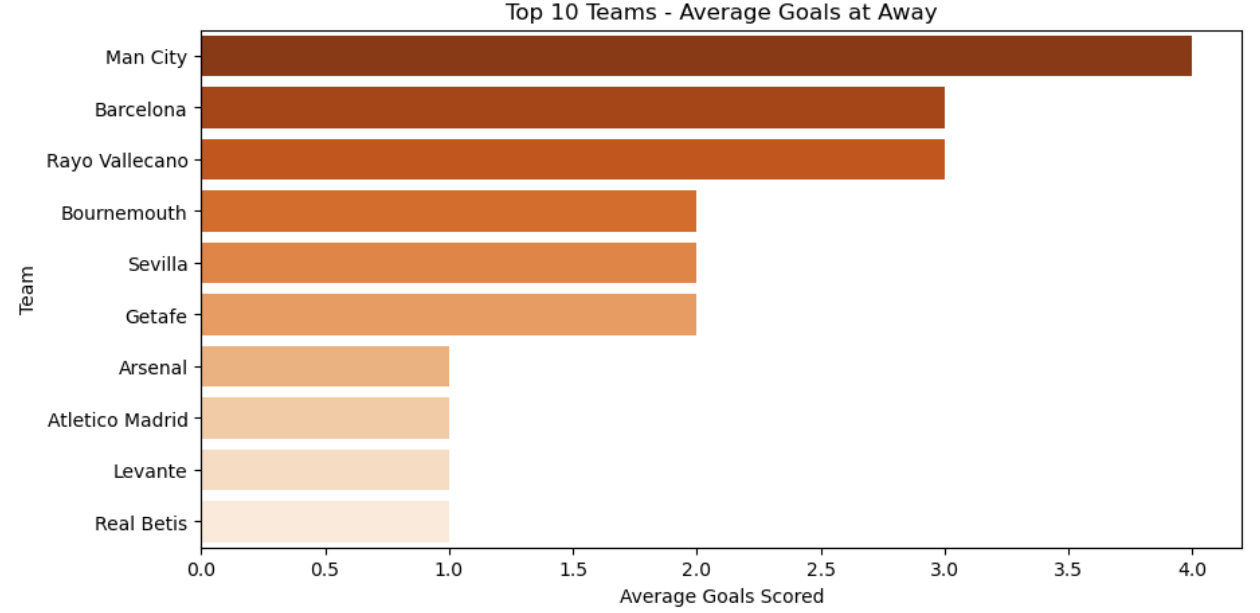
*Average Goals Scored by Home Teams.*

The bar chart with average home team goals (Liverpool highest).



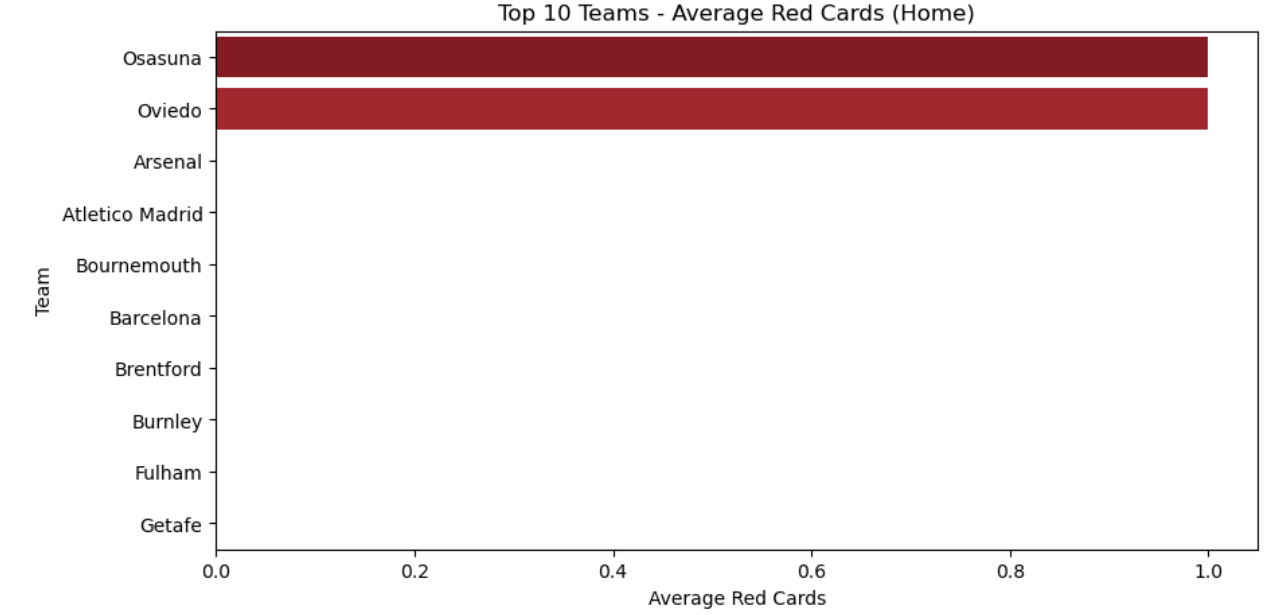
*Average Goals Scored by Away Teams.*

The bar chart with average away team goals (Man City highest).



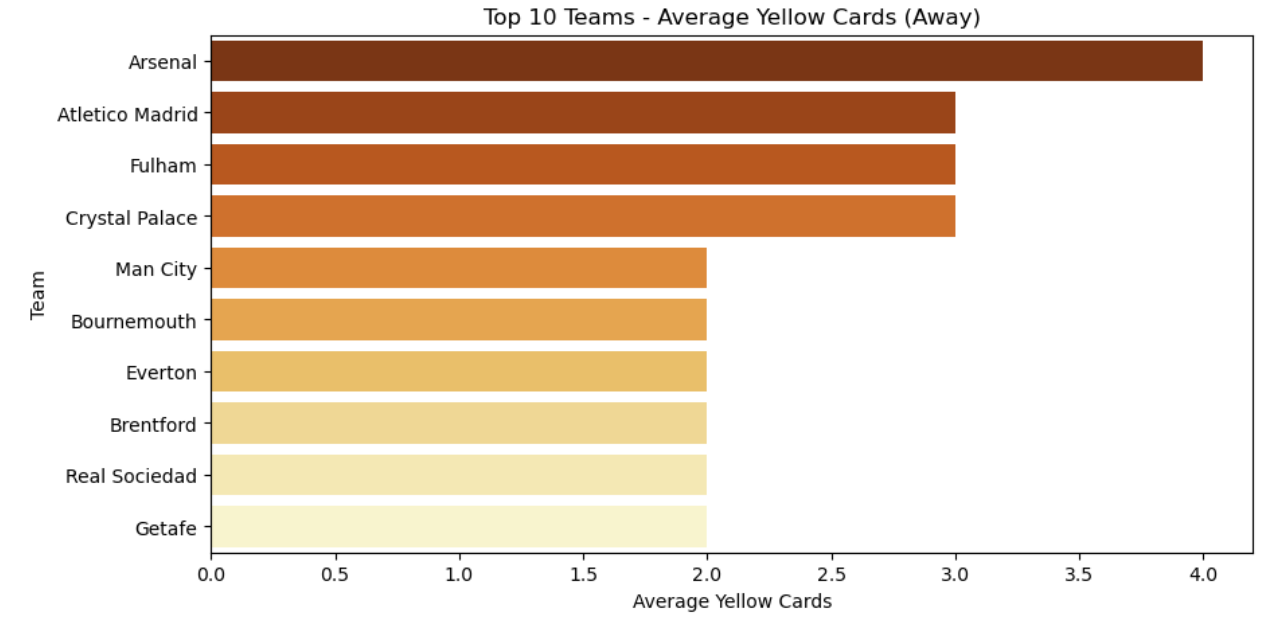
*Average Red Cards by Home Teams.*

The bar chart showing red card averages for home teams (Osasuna top).



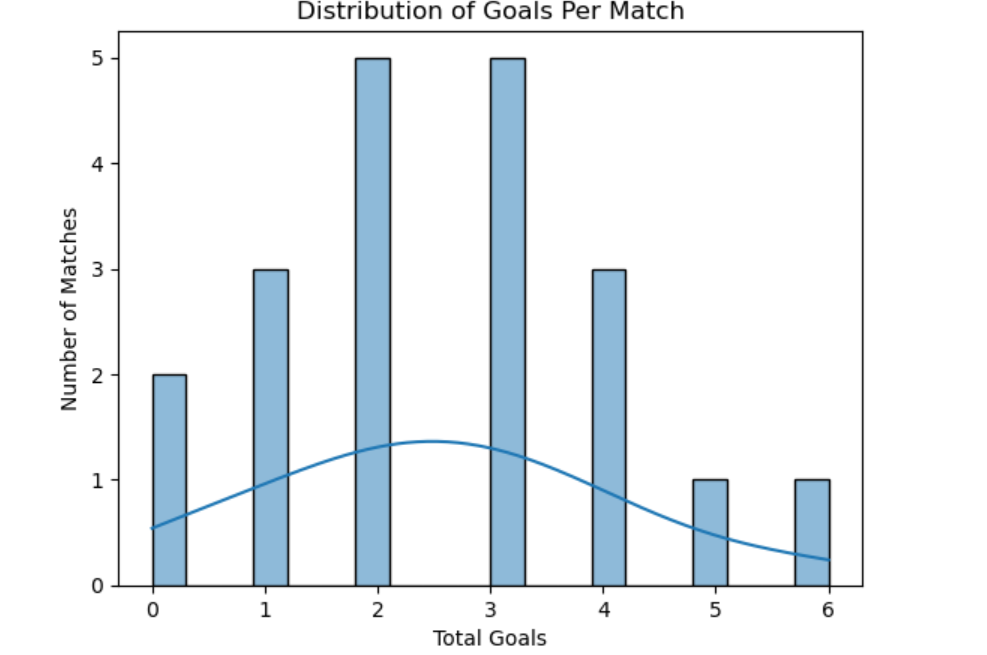
*Fouls and Yellow Cards by Away Teams.*

The bar chart comparing fouls and yellow cards for away teams (Arsenal and Atletico Madrid).



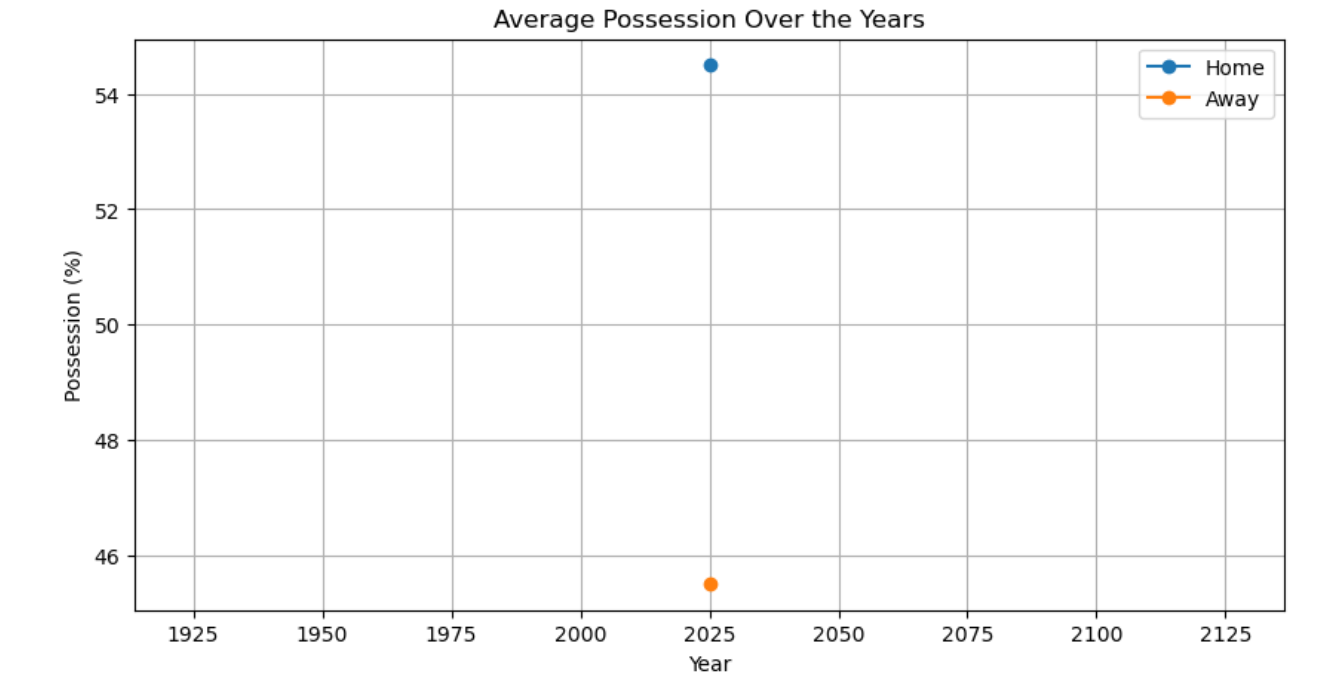
*Goals Distribution.*

This histogram shows the distribution of total goals scored across matches. Most matches recorded between two and three goals, while very few ended with extremely high or low goal counts. The visualization highlights that although some teams displayed strong attacking performances, the majority of matches followed a balanced scoring pattern.

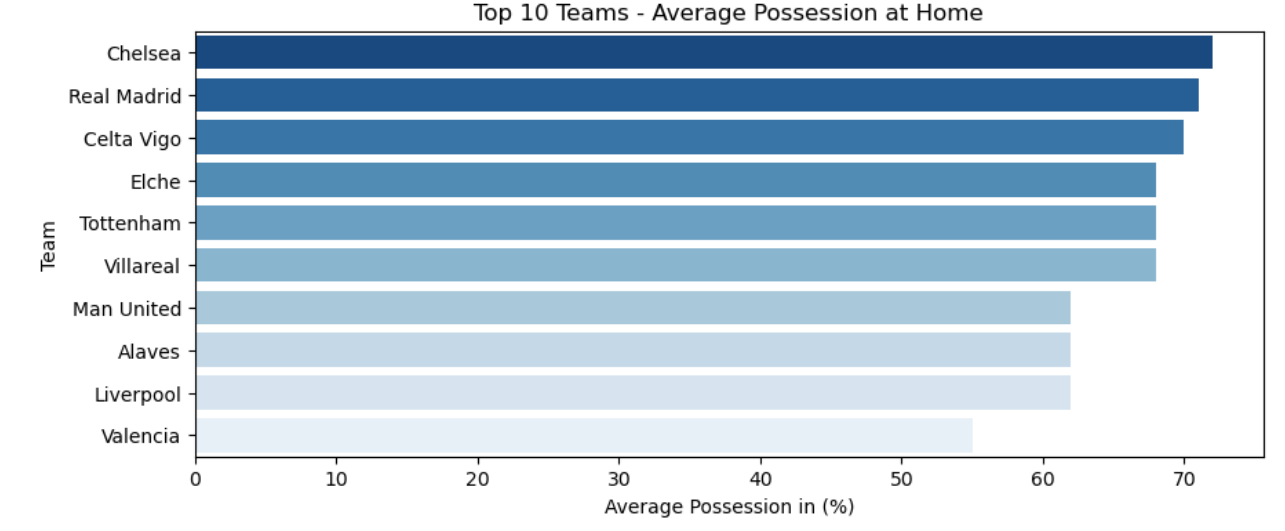


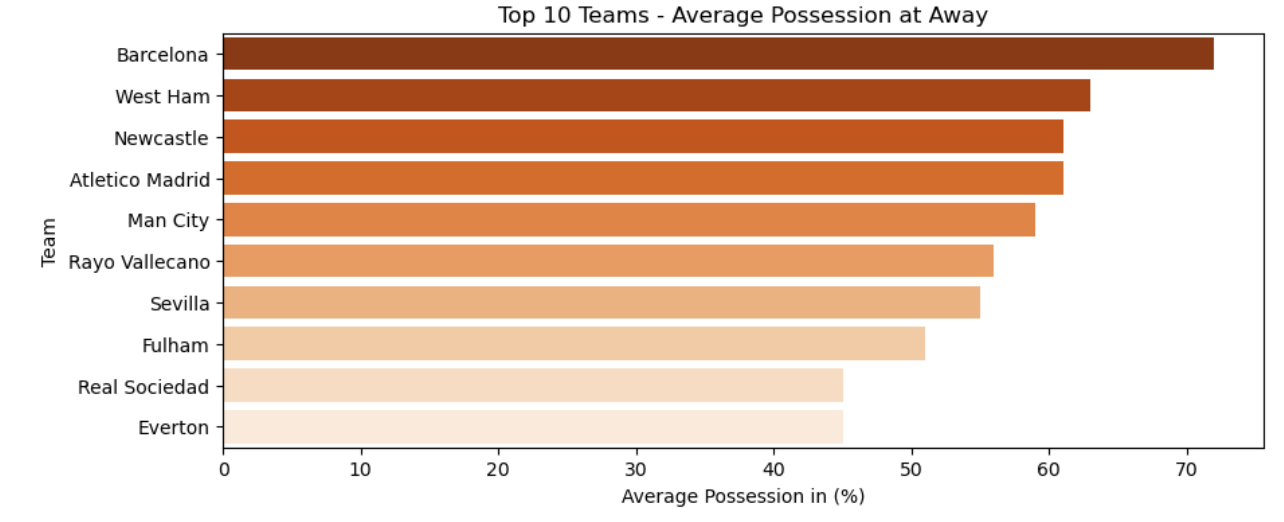
*Possession Distribution.*

This histogram illustrates the distribution of possession percentages among teams. The analysis shows that most teams consistently maintained possession in the range of 50 to 60 percent, reflecting balanced control of the ball across matches. However, outliers such as Chelsea at home and Barcelona away reached levels as high as 72 percent, demonstrating their ability to dominate possession beyond the typical average. Over the years, this pattern reinforces the idea that while many teams share possession relatively evenly, possession-heavy teams like Barcelona and Chelsea have historically built their strategies around controlling the flow of the game.



*Possession Comparison Home vs Away.*

The boxplot comparing home team possession vs away team possession ranges.



**Conclusion**

**Summary:**  
This project applied data science techniques to explore patterns and performance indicators in football matches. Through data preprocessing, visualization, and descriptive analysis, the study uncovered clear trends related to home advantage, attacking efficiency, discipline, possession, and shooting accuracy.

The findings showed that home teams generally performed better, with Liverpool and Manchester City emerging as leaders in offensive performance at home and away respectively. Possession-oriented teams such as Chelsea and Barcelona demonstrated their ability to dominate matches by controlling the ball, while disciplinary issues, particularly for teams like Mallorca, Arsenal, and Atletico Madrid, highlighted the negative impact of fouls and cards. Shooting accuracy further reinforced the importance of efficiency in front of goal, with teams like Aston Villa and Oviedo making the most of their chances.

Unlike predictive modeling projects, this analysis focused solely on descriptive and exploratory methods. The goal was not to build a forecasting system but to transform raw football statistics into meaningful insights that explain match outcomes and team behavior.

In conclusion, the project demonstrates how data science can be used to better understand the dynamics of football. The insights gained provide a foundation for future work, which could include the development of predictive models to forecast match results or advanced performance metrics at the player level.

**Future Work and Limitations**

While this project provided meaningful insights into football match outcomes and team performance, it was limited in scope to the variables available in the dataset. Player-level statistics such as individual goals, assists, passing accuracy, and defensive actions were not included, which restricted the depth of analysis. Incorporating player data in future studies would enable a more detailed understanding of how individual contributions shape team results.

Another limitation is that the dataset covered a relatively small number of matches, which may not capture long-term trends across different seasons or competitions. Expanding the dataset to include multiple leagues, seasons, and international competitions would improve the generalizability of the findings.

This project also focused on descriptive and exploratory analysis rather than predictive modeling. Future work could extend the analysis by building machine learning models to forecast match outcomes, predict goal probabilities, or evaluate player performance. Additionally, advanced techniques such as clustering could be applied to group teams with similar playing styles, while regression models could be used to quantify the impact of possession, discipline, and shooting accuracy on winning probabilities.

Overall, the project lays a strong foundation for football analytics using exploratory data science. With larger datasets and advanced modeling approaches, future research could provide deeper insights into tactical patterns, player efficiency, and predictive performance in the sport.

**References**

Kaggle Football Team Stats 2022/2023 : <https://www.kaggle.com/datasets/vivovinco/20222023-football-team-stats>

Kamau Johnson. Documentation and analysis blog post on Medium. Available at: <https://medium.com/@Kamau_Johnson/football-matches-analysis-report-25ef1e9f7e28>