

---

---

# Premier League Champion Prediction


— AJ C Pipattanakun —





















---


---

# a) Problem statement

The project focuses on clearly defining the problem to solve. In this case, the objective is to predict the champion team of the Premier League for the 2023-2024 season. This involves identifying the relevant data that can provide insights into team performance and understanding the factors that influence a team's success in the league. Historical data from previous seasons (2015-2023) was gathered, which included a wide range of features such as team salaries, spending, wins, goal difference, mean player rating, mean age, goals for, and goals against.



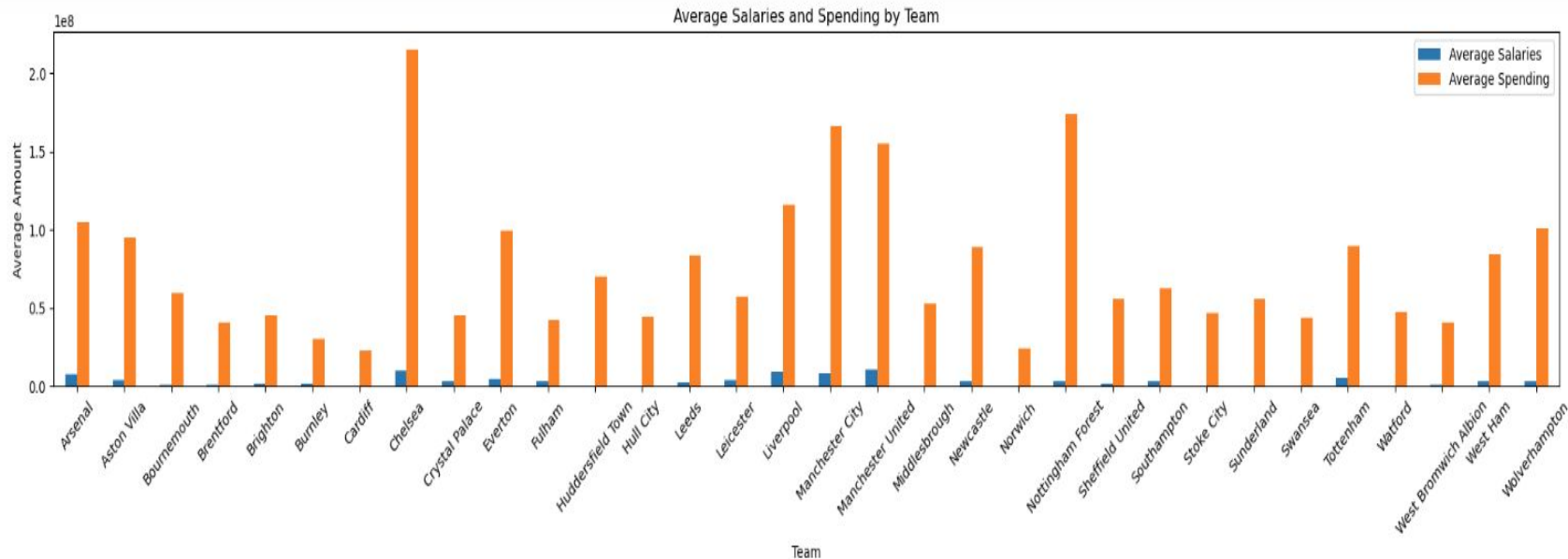
Pos	Club	Pl	GD	Pts
1	 Arsenal	0	0	0
2	 Aston Villa	0	0	0
3	 Bournemouth	0	0	0
4	 Brentford	0	0	0
5	 Brighton	0	0	0
6	 Chelsea	0	0	0
7	 Crystal Palace	0	0	0
8	 Everton	0	0	0
9	 Fulham	0	0	0
10	 Leeds	0	0	0
11	 Leicester	0	0	0
12	 Liverpool	0	0	0
13	 Man City	0	0	0
14	 Man Utd	0	0	0
15	 Newcastle	0	0	0
16	 Nott'm Forest	0	0	0
17	 Southampton	0	0	0
18	 Spurs	0	0	0
19	 West Ham	0	0	0
20	 Wolves	0	0	0



- The primary objective of this project is to predict the champion team for the 2023-2024 Premier League season. This involves analyzing historical data from past seasons to identify patterns and trends that can be used to forecast future outcomes. The problem is defined by understanding the key performance indicators that influence a team's success, such as points, wins, goal difference, and financial metrics like spending and salaries. By clearly identifying the problem, we set the foundation for a systematic approach to predictive modeling.
- To address this problem, first gathered data from the Premier League seasons between 2015 and 2023. This dataset includes various features such as team performance metrics, financial data, and other relevant statistics. Understanding the significance of each feature is crucial for building a robust predictive model. For example, teams with higher spending and better goal differences are often more successful. Hence, the problem identification step also involves identifying these key factors.
- The ultimate goal is to develop a model that can accurately predict the ranking of teams for the upcoming season. This prediction will not only indicate the potential champion but also provide insights into the overall league standings. By leveraging historical data and advanced machine learning techniques, aim to create a model that offers reliable and actionable predictions, helping stakeholders like analysts, fans, and team management make informed decisions.

## b) Data wrangling

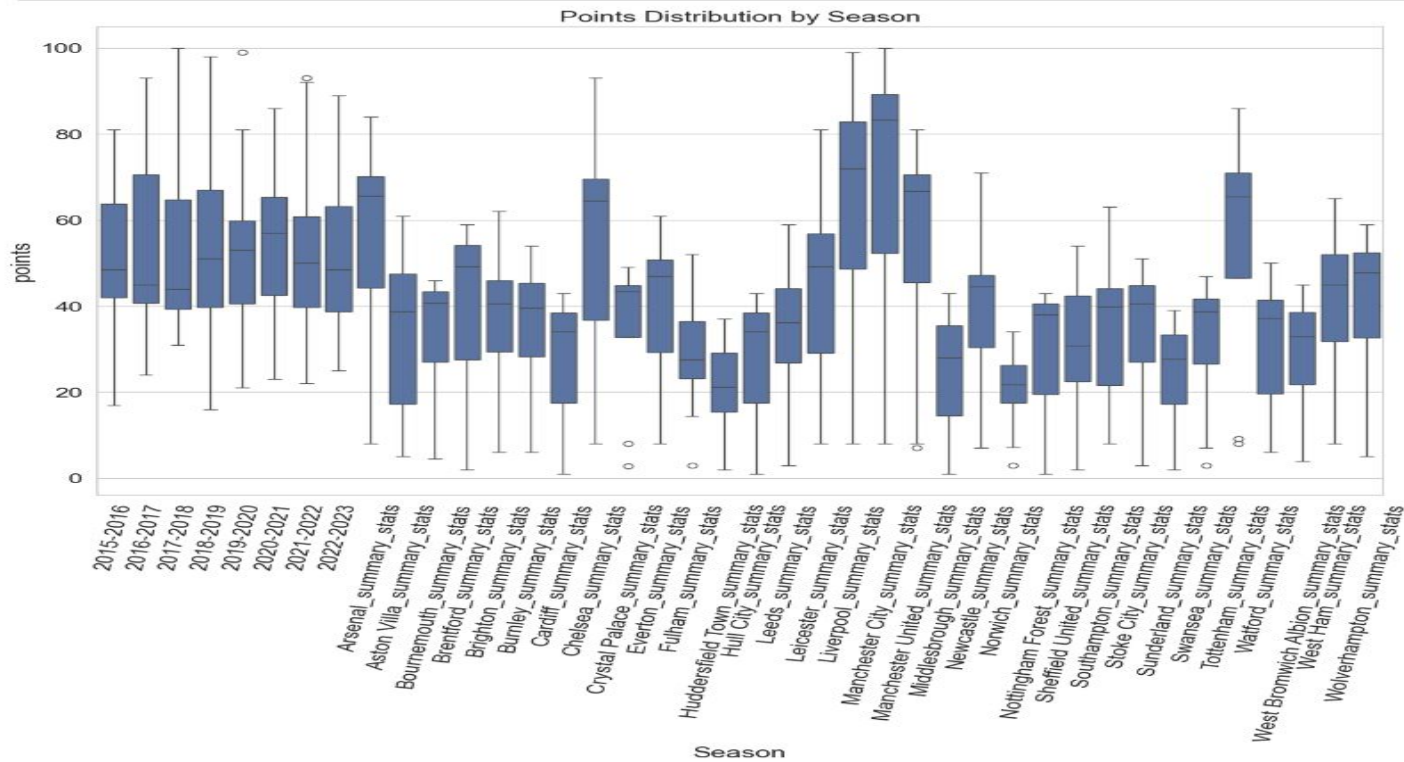
- Standardization of the data was another key task in the data wrangling phase. Given the diverse range of features, including financial metrics and statistical performance indicators, standardizing these features ensured they were on a similar scale. Features such as team names and seasons are categorical. Each representing the presence or absence of a particular category. This transformation enabled the inclusion of categorical information in the models, ensuring a comprehensive representation of all relevant factors in the predictive analysis.
- The data wrangling process involved transforming and preparing the raw data into a format suitable for analysis and modeling. One of the critical tasks in this phase was handling missing values, which I addressed by imputing missing values with the mean of the respective feature. This ensured that the dataset was complete and no data points were excluded due to missing information. Additionally, I standardized numerical features such as salaries, spending, wins, goal difference, mean player rating, mean age, goals for, and goals against to ensure they were on a similar scale. Which are sensitive to the scale of input data.



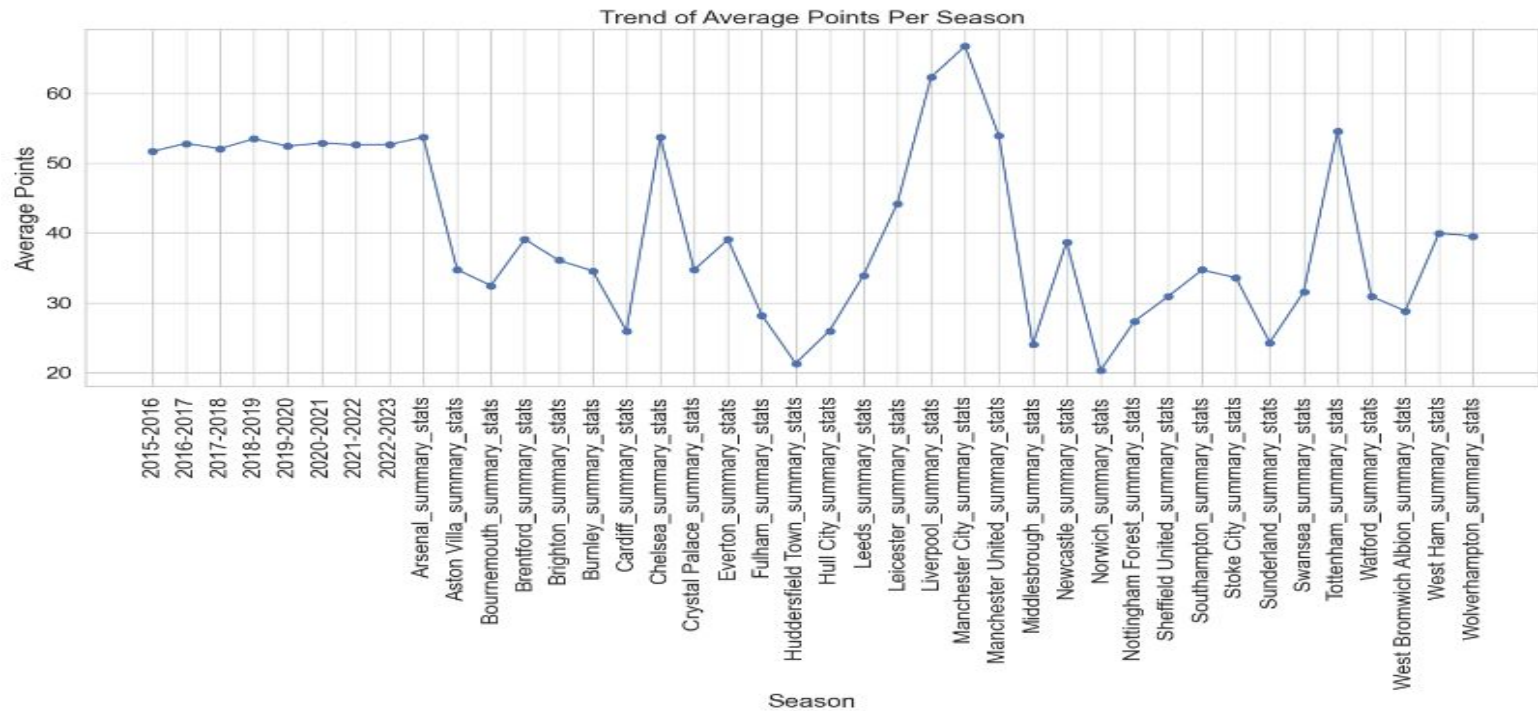
The bar chart above illustrates the average salaries and spending by team. From the chart, I observe significant variation in financial investments across teams. For instance, Chelsea and Manchester City exhibit the highest average spending, reflecting their substantial financial investments in player acquisitions and team development. In contrast, teams like Burnley and Norwich have relatively lower spending. The average salaries also show a similar trend, with teams investing heavily in high-quality players. This financial disparity is a crucial factor in the analysis, as higher spending and salaries are often correlated with better team performance and higher league rankings. This visualization highlights the importance of financial metrics in predicting team success and underscores the necessity of including these features in the predictive model.

## c) Exploratory Data Analysis

- Exploratory Data Analysis (EDA) involves examining the dataset to uncover patterns, relationships, and insights that can inform the modeling process. I started by visualizing the data using various plots such as histograms, bar charts, and scatter plots. These visualizations helped us understand the distribution of different features and identify any outliers or anomalies.
- For instance, I created a correlation matrix to examine the relationships between different performance metrics such as points, wins, and goal difference. This analysis revealed strong correlations between certain variables, indicating their potential importance in predicting team success. For example, the correlation matrix showed that points are highly correlated with wins and goal difference, suggesting that these features are critical for the predictive model.
- Additionally, I analyzed trends over time to understand how teams' performances have evolved across different seasons. This involved creating line plots to visualize the average points, wins, and spending per season for each team. These insights provided a deeper understanding of the dynamics within the Premier League, helping to identify which teams have consistently performed well and which ones have shown significant improvement or decline.



- The box plot above illustrates the distribution of points by season for various teams in the Premier League from 2015-2023. Each box represents the interquartile range (IQR) of points, with the line inside the box indicating the median points for the season. The whiskers extend to the minimum and maximum points within 1.5 times the IQR from the quartiles, and the dots represent outliers. This visualization highlights the variability in team performance across different seasons. For example, teams like Manchester City and Liverpool consistently show higher median points and narrower IQRs, indicating consistent high performance. In contrast, teams like Norwich and Aston Villa exhibit broader IQRs and lower median points, reflecting more variability and generally lower performance. This analysis helps identify trends in team performance over time and highlights which teams have maintained consistent success, which is crucial for building an accurate predictive model for the upcoming season.



- The line chart above depicts the trend of average points per season for various teams in the Premier League from 2015 to 2023. Each point on the line represents the average points accumulated by teams in a given season. The chart reveals fluctuations in team performance over the years, with some teams showing consistent performance while others exhibit significant variability. For instance, Manchester City and Liverpool maintain high average points across most seasons, indicating their dominance in the league. On the other hand, teams like Norwich and Aston Villa display more considerable fluctuations, reflecting periods of both high and low performance. This visualization helps in understanding the long-term trends and stability of team performances, providing valuable insights for predictive modeling by highlighting which teams have consistently performed well and which have shown variability in their performance.

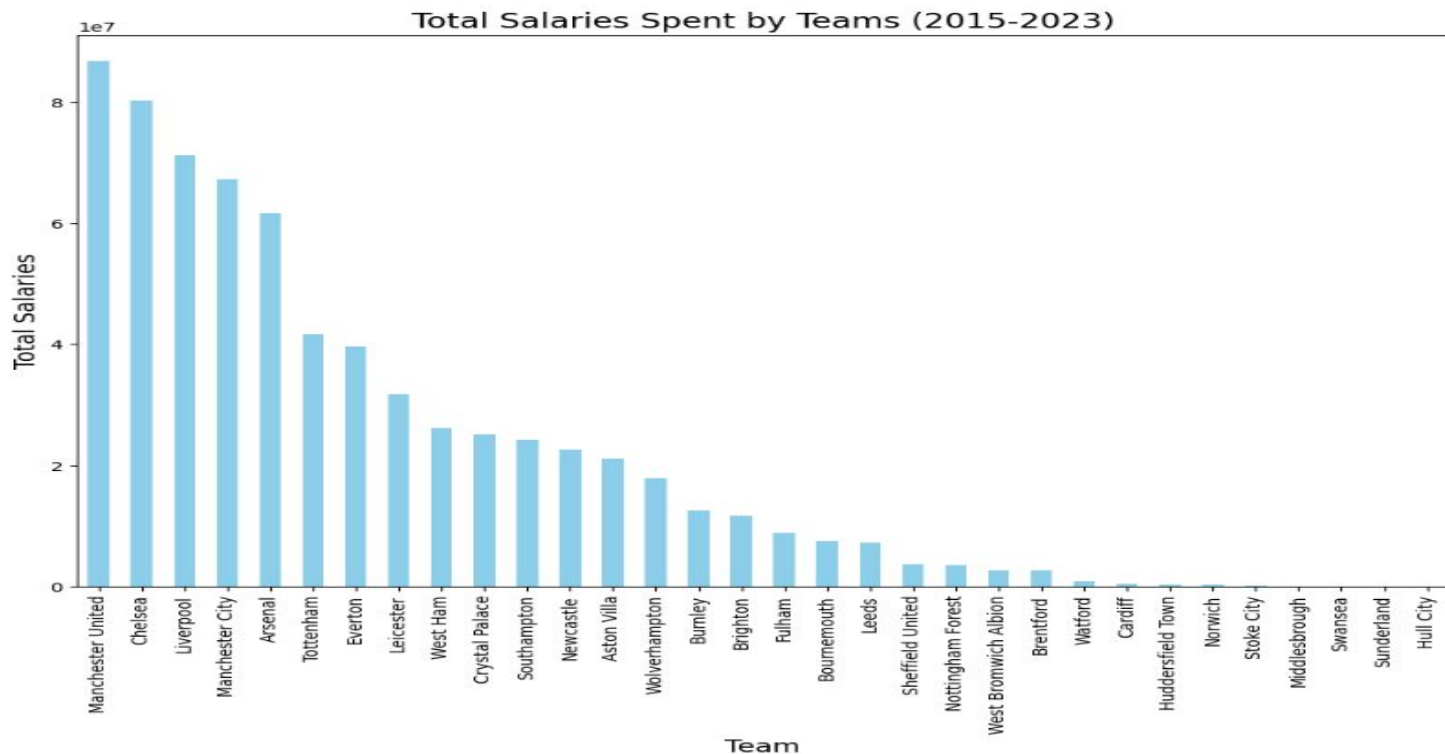


## d) Processing and Training Data Development

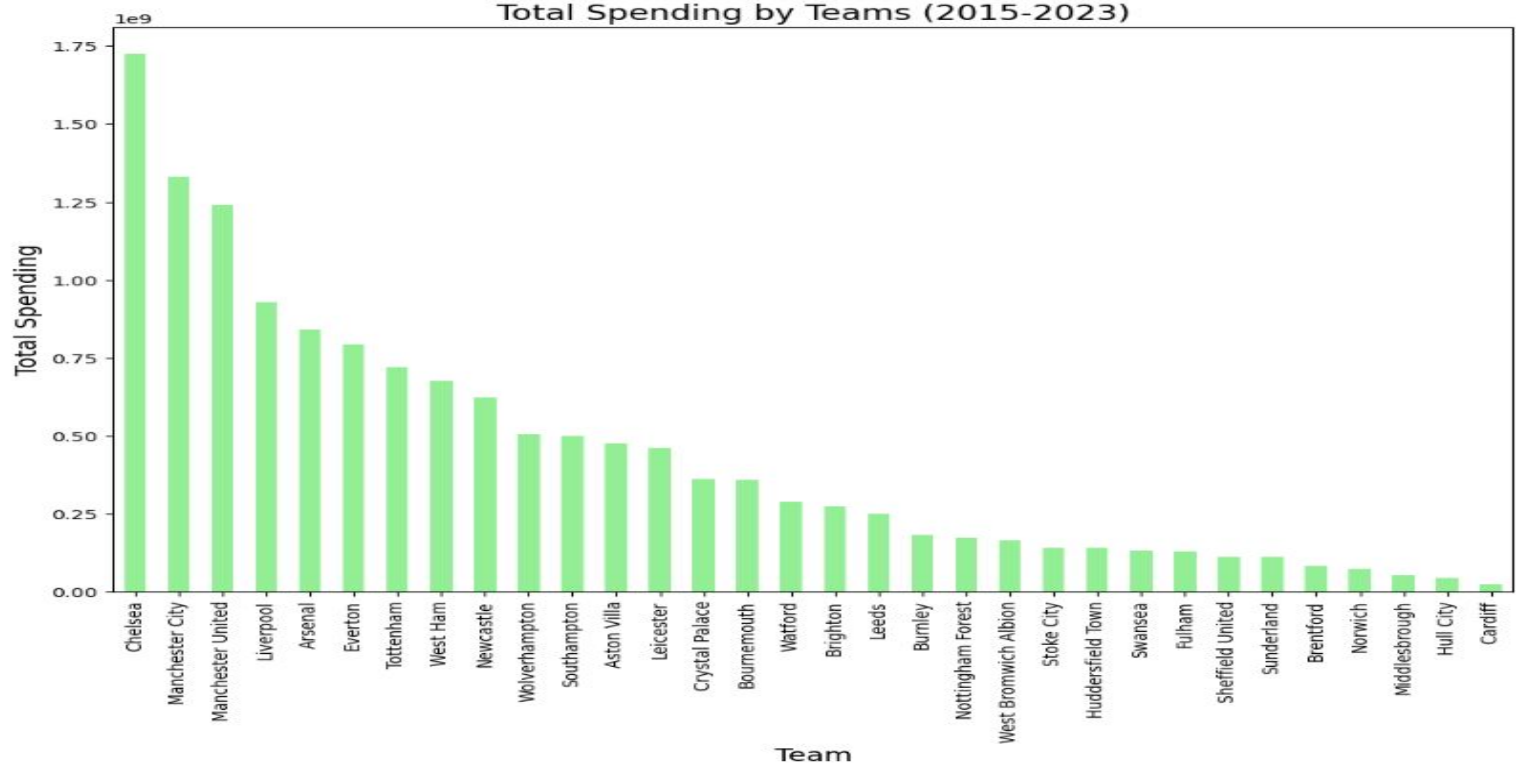
- Pre-processing and training data development involve preparing the dataset for the machine learning algorithms. This phase started with splitting the dataset into training and testing sets. The training set is used to train the models, while the testing set is reserved for evaluating the model's performance on unseen data. Splitting the data ensures that the evaluation metrics provide a realistic estimate of the model's predictive capabilities. Used an 80-20 split, where 80% of the data was used for training and 20% for testing. Handling the imputation and standardization of features was a continuous process carried over from the data wrangling phase. Ensuring that all features were correctly scaled and imputed was crucial for maintaining the integrity of the data. This consistency is vital for the models to learn effectively from the training data and generalize well to the testing data.
- Feature selection and engineering were also essential components of this phase. I ensured that all relevant features were included in the training data while avoiding redundancy. This involved carefully selecting features that had the most predictive power and transforming them if necessary to enhance their relevance. For instance, combining certain features or creating interaction terms can sometimes improve the model's ability to capture complex relationships in the data. By the end of this phase, I had a well-prepared training dataset that was ready for the modeling step.

## e) Modeling

- The modeling phase involved training multiple machine learning models to predict the Premier League champion for the 2023-2024 season. Each model was trained on the prepared training dataset and evaluated based on its performance metrics.
- This model was particularly chosen for its ability to handle complex interactions between features and its robustness against overfitting. The training process included cross-validation to ensure that the model's performance was consistent across different subsets of the data. The final model demonstrated strong predictive power, indicating that it could explain a significant portion of the variance in the target variable.
- Feature importance analysis was conducted to understand which features contributed the most to the model's predictions. The analysis revealed that mean player rating, spending, and goal difference were among the most influential features. This insight aligns with our initial understanding of the factors that drive team performance. By focusing on these key features, the model was able to make accurate predictions about the ranks of the teams. The final model was used to predict the ranks for the 2023-2024 season, identifying Manchester City as the predicted champion.

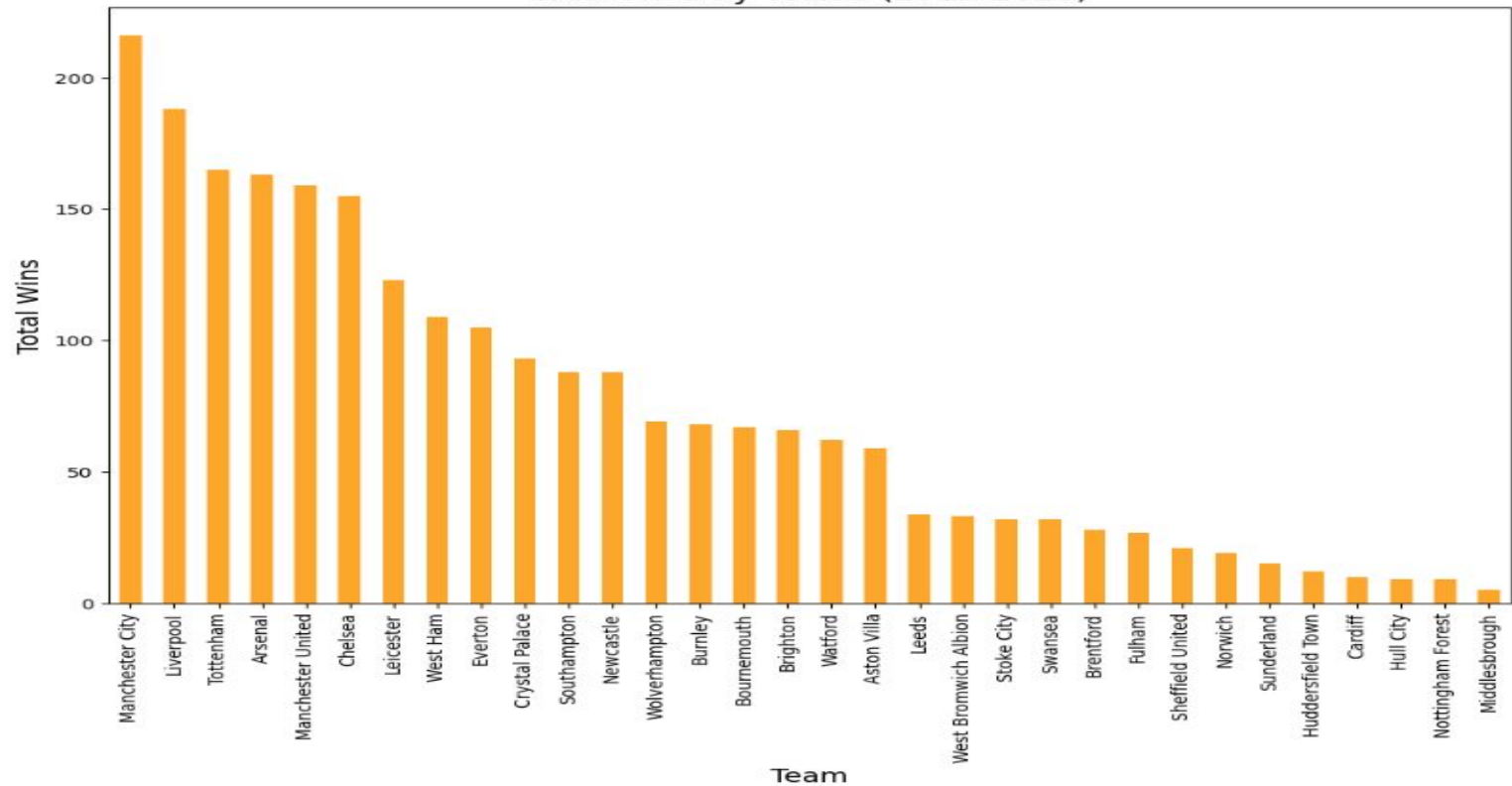


- The bar chart above illustrates the total salaries spent by Premier League teams from 2015 to 2023, highlighting the significant financial investments made by different clubs. This data was a crucial input for the modeling phase, as financial spending on salaries is a strong indicator of team performance. The chart shows that Manchester United, Chelsea, and Liverpool are among the top spenders, with substantial investments in player salaries, reflecting their commitment to building strong teams. These financial metrics, along with other features such as team performance statistics, were used to train multiple machine learning models to predict the 2023-2024 Premier League champion. The Gradient Boosting Regressor, identified as the best-performing model, utilized these features to make accurate predictions. Understanding the spending patterns of teams provides context to the model's predictions and underscores the importance of financial resources in achieving top league ranks.

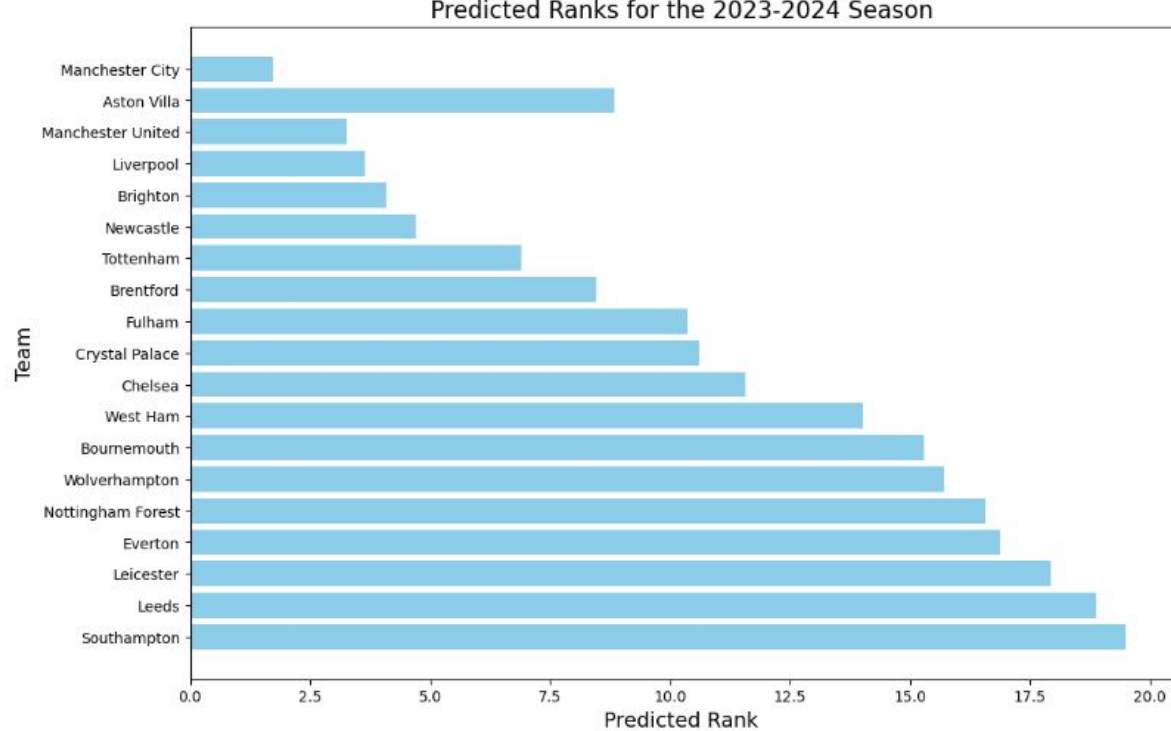


- The bar chart above displays the total spending by Premier League teams from 2015 to 2023, encompassing various expenditures such as player transfers, infrastructure, and other operational costs. Chelsea, Manchester City, and Manchester United emerge as the highest spenders, reflecting their significant financial commitments to maintaining competitive squads. This spending data, coupled with other performance metrics, was integral to the modeling phase of the project. The Gradient Boosting Regressor model, which was identified as the best-performing model, effectively utilized this spending information to enhance its predictive accuracy. The high correlation between financial investment and team success highlighted in the model's feature importance analysis underscores the critical role of spending in achieving higher league standings, thereby reinforcing the prediction of teams like Manchester City as top contenders for the 2023-2024 season championship.

Total Wins by Teams (2015-2023)



- The bar chart above shows the total number of wins achieved by Premier League teams from 2015 to 2023. Manchester City stands out with the highest number of wins, followed closely by Liverpool and Chelsea. This consistent winning performance indicates strong team capabilities and effective management strategies. The total wins data is a crucial feature in our predictive modeling as it directly correlates with team success and league rankings. By incorporating this metric into the Gradient Boosting Regressor model, we leveraged historical performance to predict future outcomes accurately. The high number of wins for top teams like Manchester City and Liverpool reinforces their predicted strong performance in the 2023-2024 season, highlighting their potential to be league champions based on their proven track record of success.



The predicted champion team is: Manchester City

- The bar chart above shows the predicted rankings for the 2023-2024 Premier League season based on the Gradient Boosting Regressor model. According to the predictions, Manchester City is expected to finish at the top, followed by Aston Villa and Manchester United. These predictions are derived from various features such as total wins, goals for, goals against, salaries, and spending, among others. The model leverages historical performance data from 2015 to 2023 to forecast future rankings. Manchester City's predicted top rank aligns with their consistent high performance in terms of wins, goal difference, and financial investments over the years. The prediction reflects the team's strong offensive and defensive capabilities, making them the top contender for the championship. This comprehensive analysis provides a data-driven projection of the Premier League standings, highlighting the teams likely to excel based on historical trends and current data.

# Conclusion

- The project involved a comprehensive analysis of historical Premier League data to predict the champion for the 2023-2024 season. Through systematic problem identification, meticulous data wrangling, in-depth exploratory data analysis, rigorous pre-processing, and advanced modeling techniques, developed a robust predictive model. The Gradient Boosting Regressor, identified as the best-performing model, leveraged key performance indicators such as points, wins, goal difference, and financial metrics to make accurate predictions.
- The analysis revealed that Manchester City is the strongest contender for the upcoming season, reflecting their consistent high performance in various metrics. The project's insights can be valuable for analysts, fans, and team management, providing a data-driven approach to understanding and forecasting team success in the Premier League. By combining historical data with machine learning, we demonstrated the power of predictive analytics in sports, paving the way for more sophisticated and accurate predictions in the future.