Capstone Project
Football Data Analysis

Mentor : Munna Pandey

Submitted By: Deepak Venkat Sairam Dasari

INTRODUCTION:

Title:

Football Analytics: Unlocking Insights for Strategic Success

Subtitle:

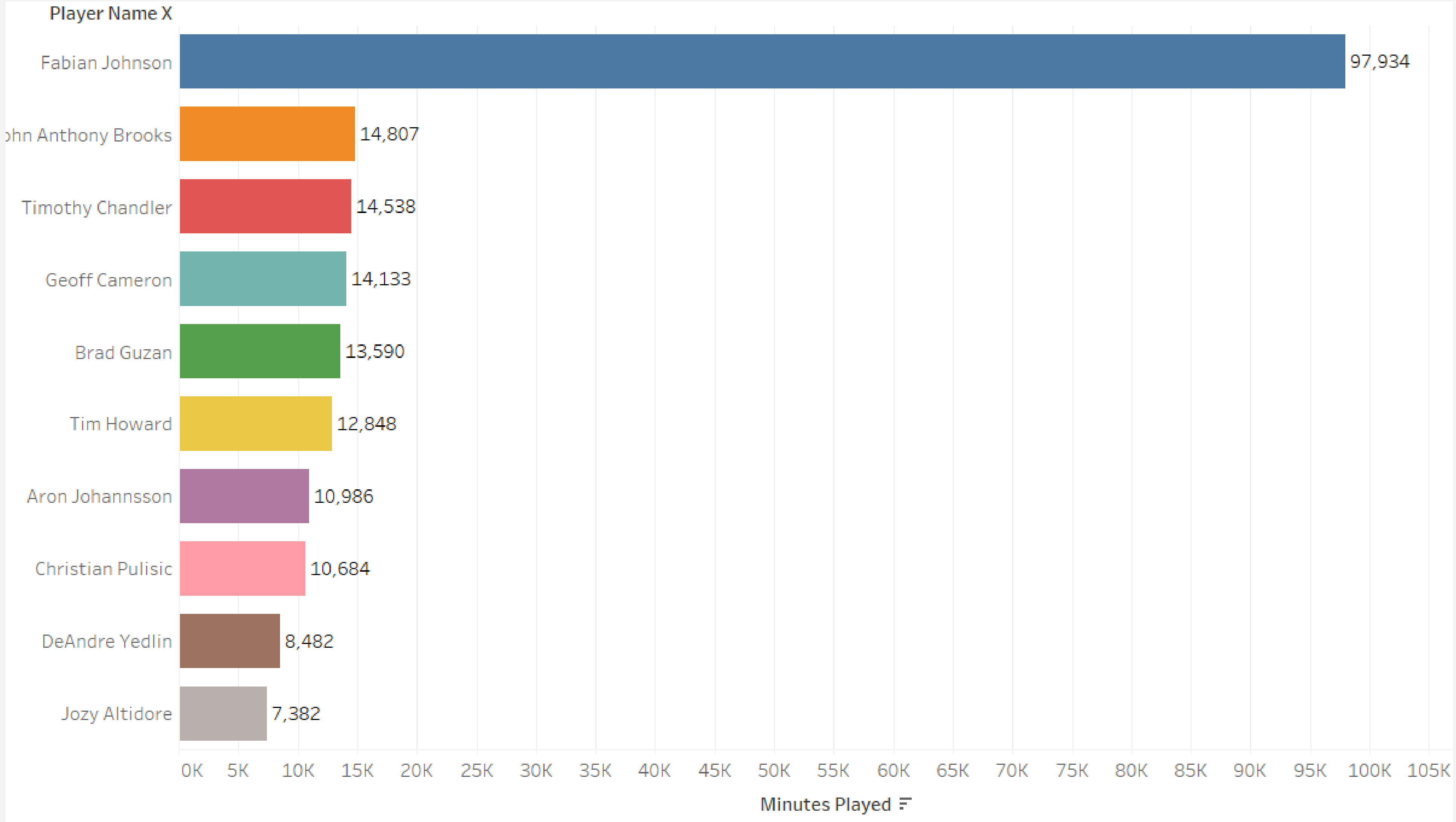
A Comprehensive Analysis of Football Data Using Advanced Analytical Techniques

Content:

- Football is more than just a game; it’s a global phenomenon that unites fans, drives business decisions, and showcases talent.
- This project aims to analyze a comprehensive football dataset to uncover actionable insights across critical areas such as player performance, team comparisons, market trends, and match events.
- Using tools like Python, MySQL, Tableau, and Excel, we delve into data preprocessing, exploratory and descriptive analysis, and predictive modeling to support decision-making in the football industry.
- The objective is to equip stakeholders with data-driven insights that enhance performance, optimize strategies, and drive competitive advantages.
- This project demonstrates the power of data analytics in transforming raw information into meaningful solutions, shaping the future of football.

Performance Analysis

Top 10 players with maximum minutes played.



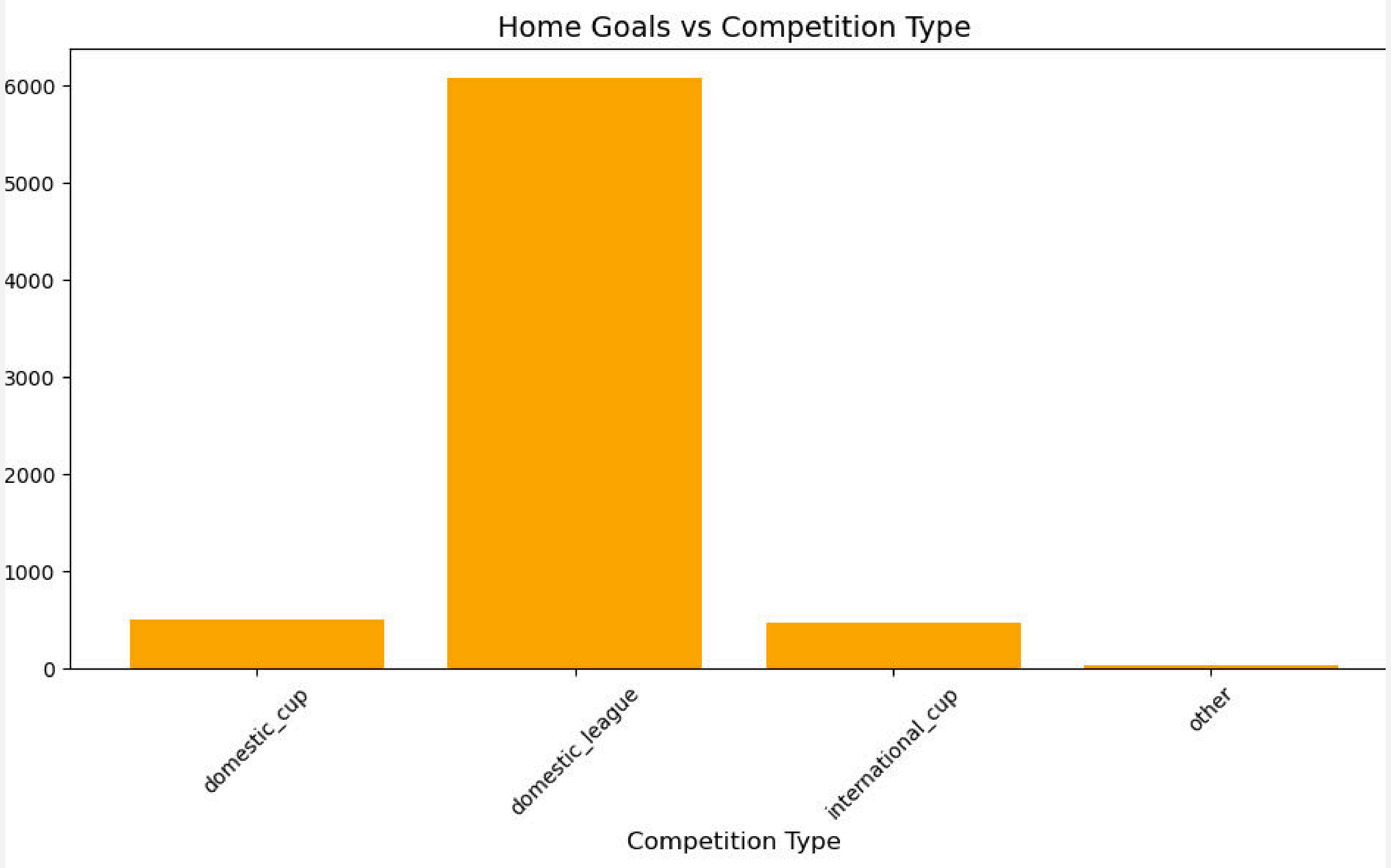
m of Minutes Played for each Player Name X. Color shows details about Player Name X. The marks are labeled by sum of Minutes Played. The view is filtered on Player Nan which has multiple members selected.

Interpretation:

- **Fabian Johnson** has played the most minutes, with a total of 97,934 minutes.
- **John Anthony Brooks** and **Timothy Chandler** come in second and third, respectively, with 14,807 and 14,538 minutes played.
- **Jozy Altidore** has played the least minutes among the listed players, with only 7,382 minutes.

Performance Analysis

What are the distribution of home goals across different competition types?



Interpretation:

- The number of goals scored is highest in the "Domestic League" competition.
- The "Domestic Cup" and "International Cup" competitions have a similar number of goals scored, which is significantly lower than the "Domestic League".
- "The "Other" category has the lowest number of goals scored.

Performance Analysis

Logistic Regression for Classification

Can we predict whether a football club will win or lose a game based on factors such as the number of goals scored, home/away status, and other match-related variables?

Out come:

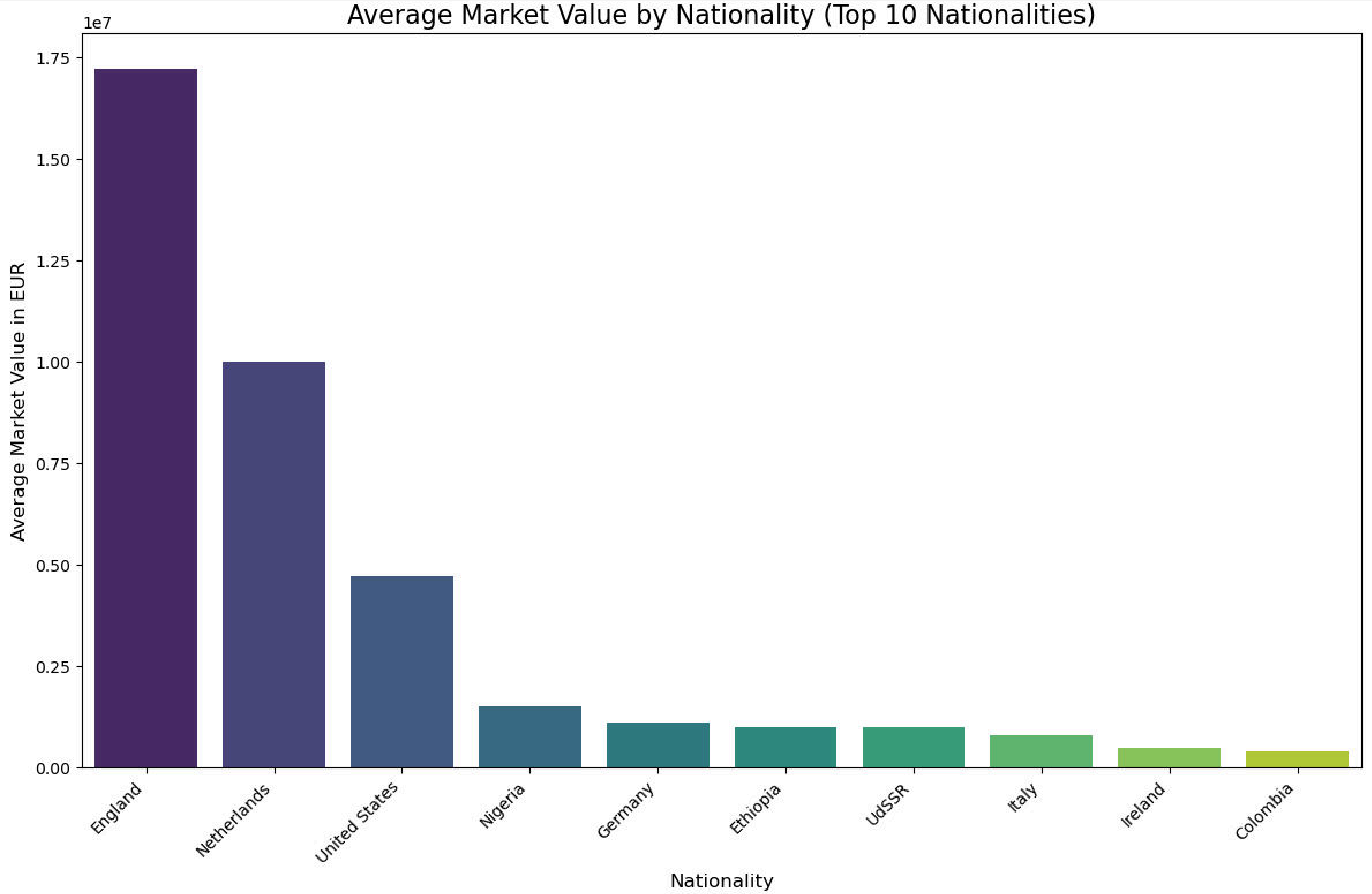
```
X = df[['home_club_goals', 'away_club_goals', 'attendance']] # Independent variables
y = df['match_outcome'] # Target variable (win or loss)
Accuracy: 1.0
Cross-validation scores: [1. 1. 1. 1. 1.]
Mean cross-validation score: 1.0
```

Interpretation:

- **Accuracy**:The model achieved an accuracy of 1.0 on the test set, indicating that it correctly predicted every match outcome for the entire test dataset.
- **Classification Report:**The precision, recall, and F1-score for both classes were 1.00, indicating that the model is perfectly identifying both home wins and away wins without any errors.
- **Cross-Validation**:The cross-validation scores across all folds were also 1.0, further confirming that the model performs perfectly when evaluated on different subsets of the data.

Player Profile and Market Value

Top 10 Average Market Value by Nationality

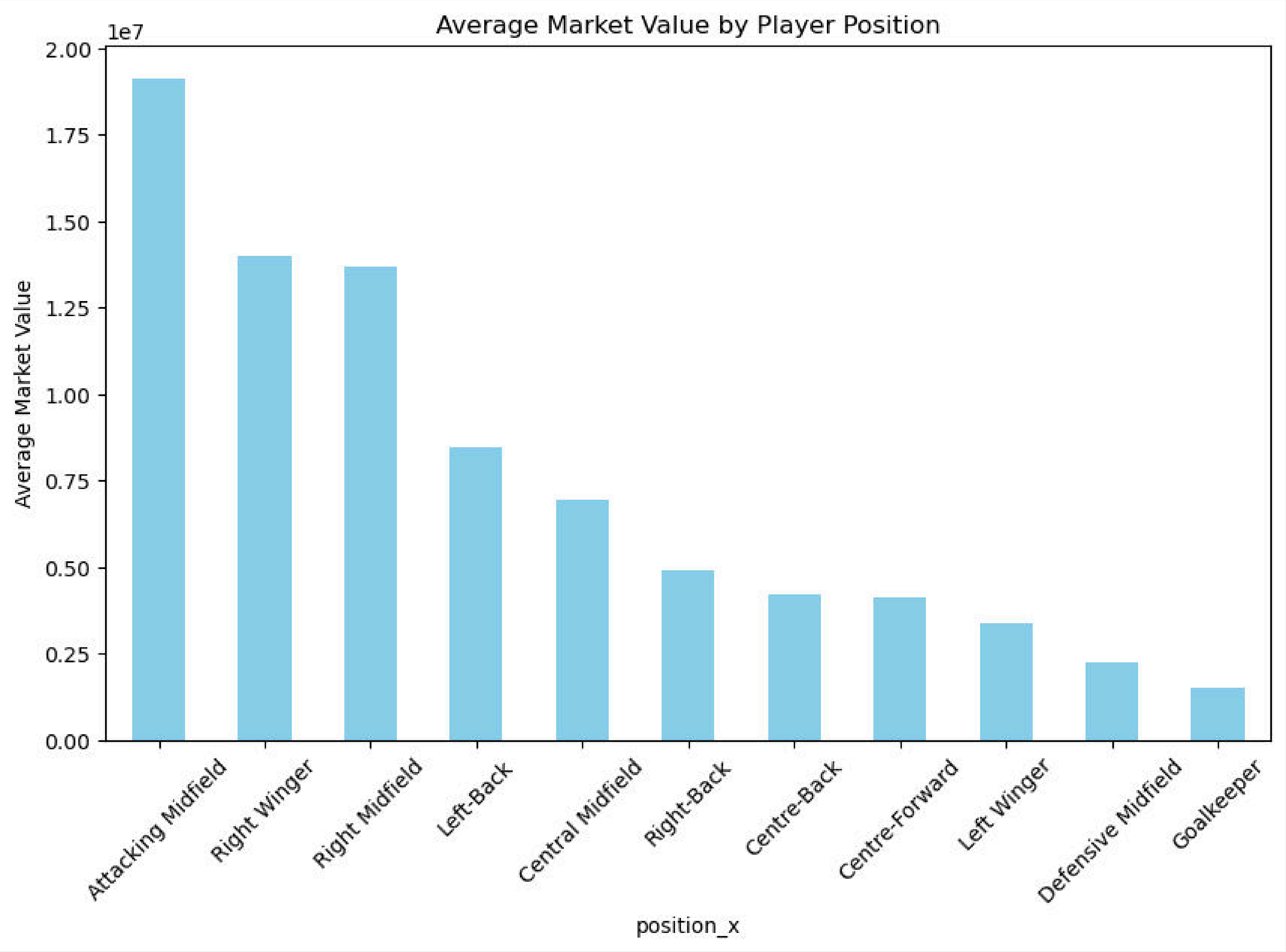


Interpretation:

- **England** has the highest average market value among the top 10 nationalities.
- The average market value decreases significantly after England, with **Netherlands** and **United States** having the next highest values.
- The remaining nationalities (Nigeria, Germany, Ethiopia, USSR, Italy, Ireland, and Colombia) have considerably lower average market values.

Player Profile and Market Value

Dose position of players influence the average market value



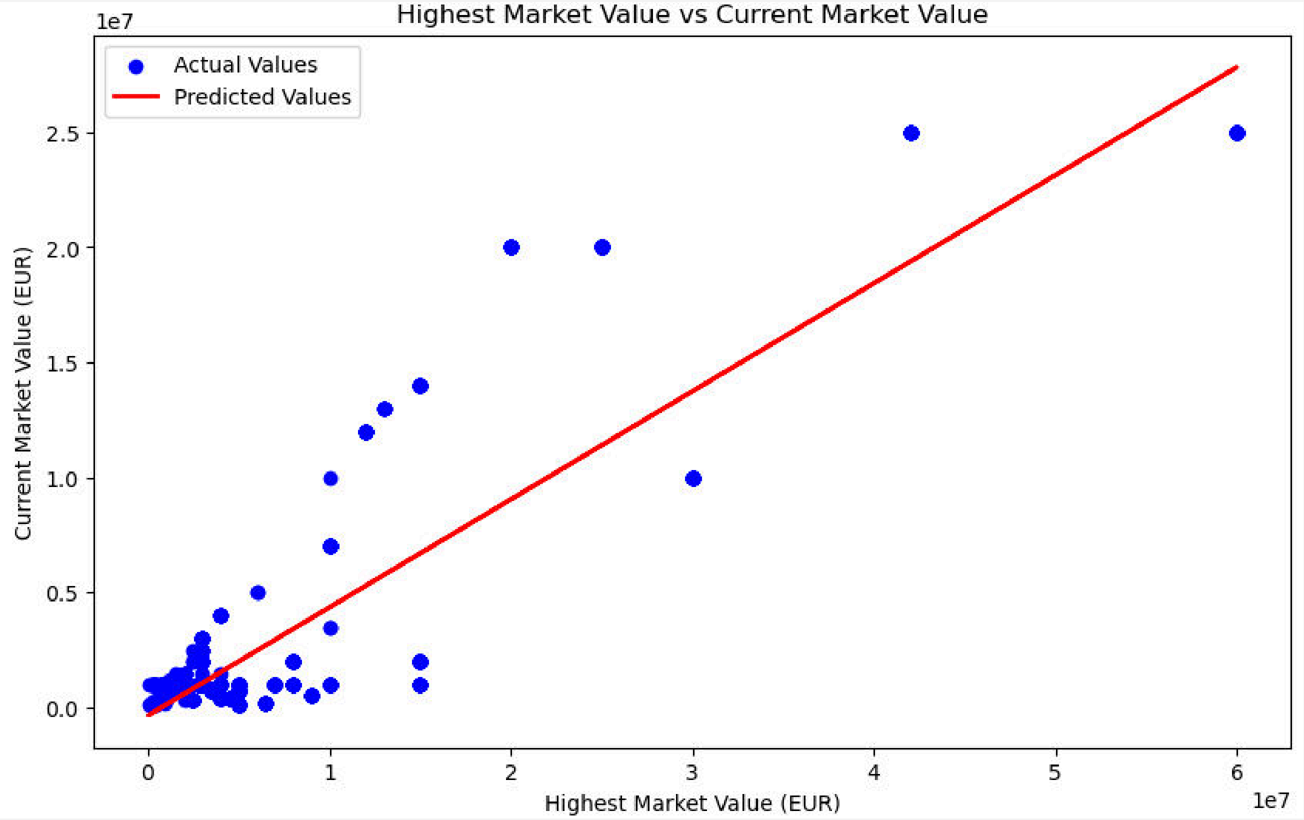
Interpretation:

- **Attacking Midfielders** have the highest average market value among the player positions.
- **Right Wingers** and **Right Midfielders** have the second and third highest average market values, respectively.
- **Left-Backs** and **Central Midfielders** have the next highest average market values.
- **Right-Backs**, **Centre-Backs,** Centre-Forwards, and Left Wingers have similar average market values, which are lower than the top-ranked positions.
- **Defensive Midfielders** and **Goalkeepers** have the lowest average market values.

Player Profile and Market Value

Simple linear regression

How does the highest market value of a player impact their current market value?



Out come :

X = df[['highest_market_value_in_eur']] # **Independent variable**
y = df['market_value_in_eur'] # **Target Variable**
R-squared: 0.7974076999430394

Interpretation:

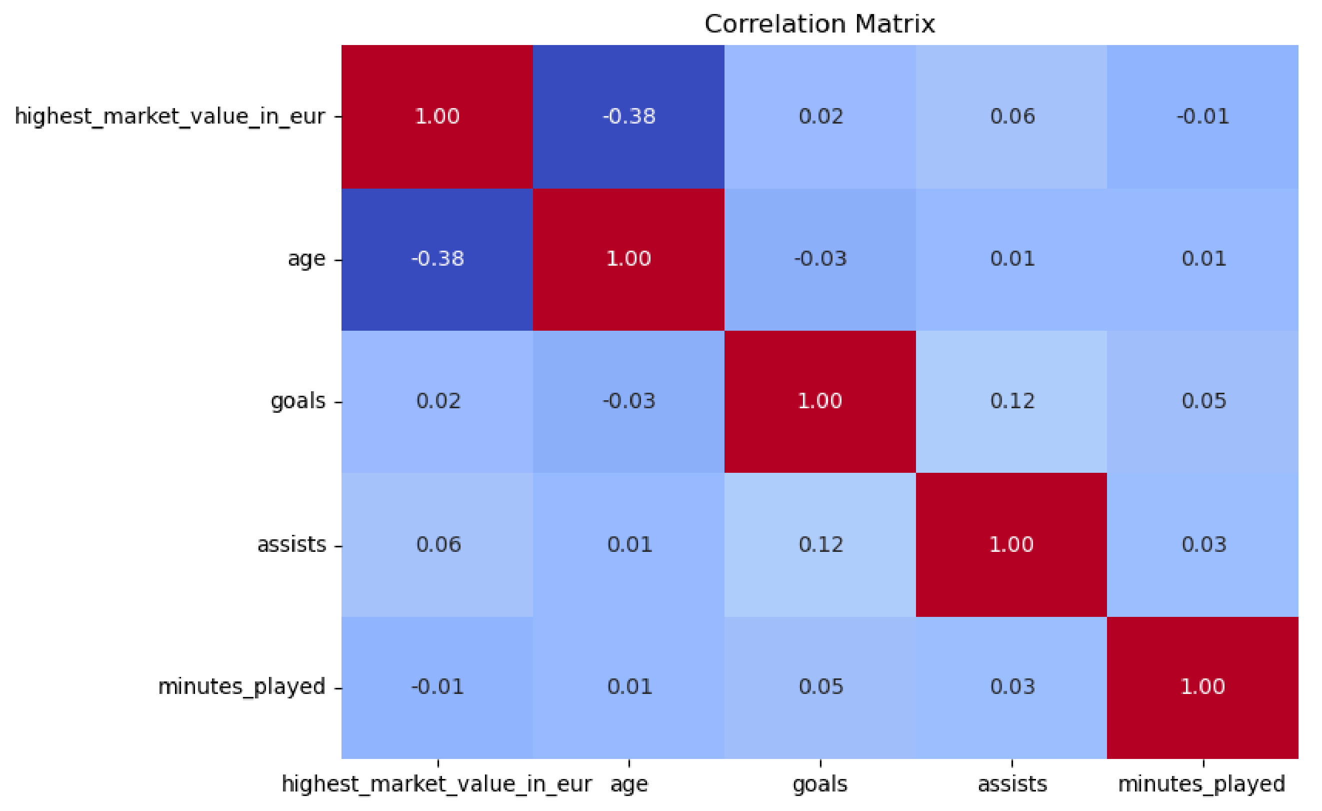
R-squared (0.80): This value indicates that approximately 80% of the variance in the target variable (market value) can be explained by the model using the selected features (such as goals, assists, minutes played, and age). This is a relatively strong fit, suggesting that the model is doing a good job of capturing the relationship between the predictors and the market value.

Based on the visual representation, we can infer that there is a positive relationship between the "Highest Market Value" and the "Current Market Value". As the "Highest Market Value" increases, the "Current Market Value" tends to increase as well.

Player Profile and Market Value

Multiple linear regression:

How can we predict a player's market value by analyzing key performance metrics such as goals, assists, and minutes played, and which features should we prioritize to ensure an accurate prediction



Out come :

X = df[['highest_market_value_in_eur', 'age', 'goals', 'minutes_played']] **#Independent variables**
y = df['market_value_in_eur'] # **Target Variable**
R-squared: 0.8443352084411382

Interpretation:

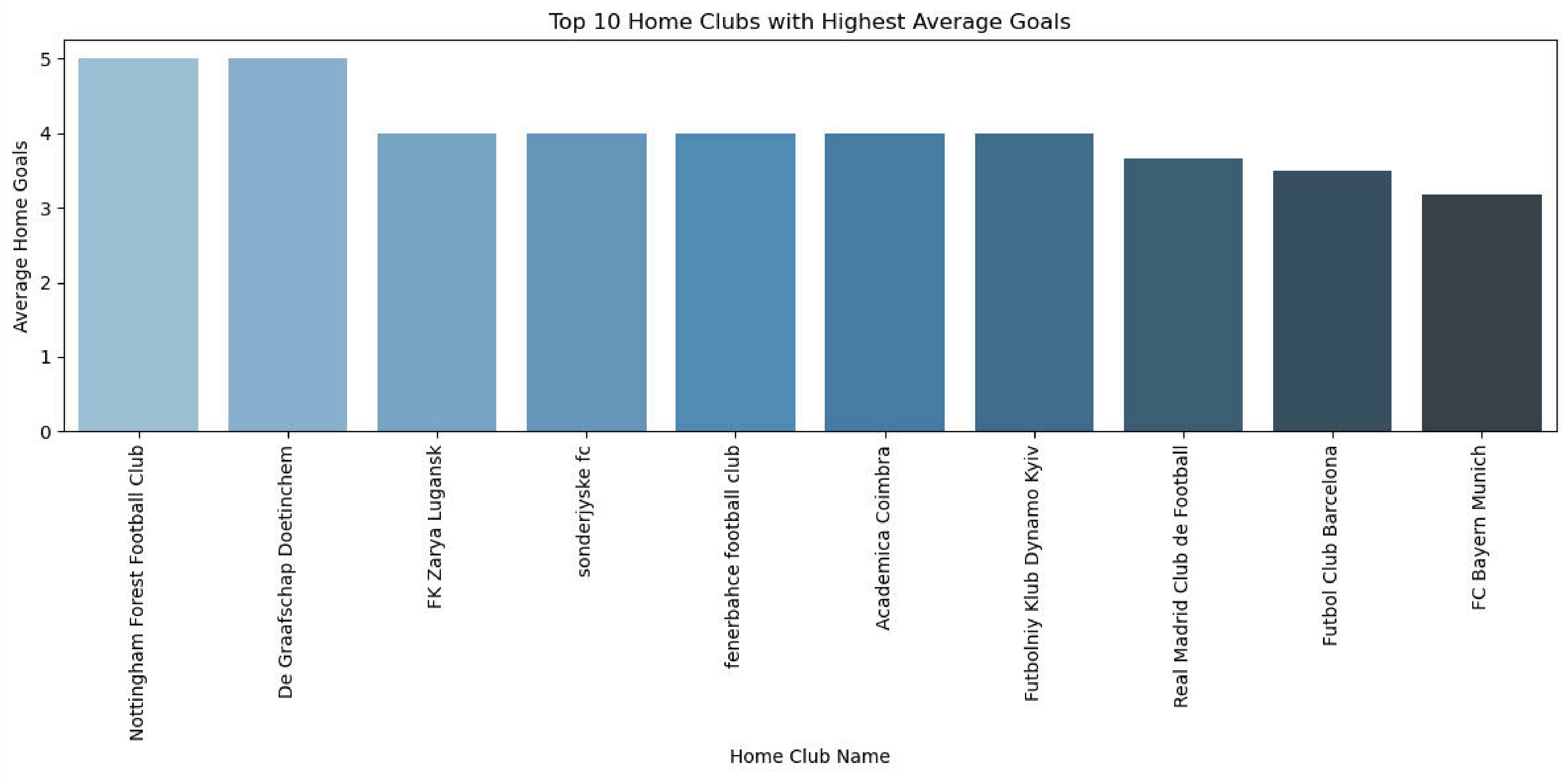
R-squared (0.8443) indicates that approximately 84.43% of the variance in the dependent variable (likely "Current Market Value") can be explained by the independent variables in the model. This suggests a good fit of the model to the data.

Goals vs. Assists: Strong correlation (0.12), suggesting players scoring more goals often provide more assists

- The feature **assists** showed strong multicollinearity with other feature **goals**
- The column was removed to avoid redundancy and improve model interpretability.
- Reducing multicollinearity ensures the model is less prone to overfitting and improves performance.

Team Comparison

Top 10 Home Clubs with Highest Average Goals

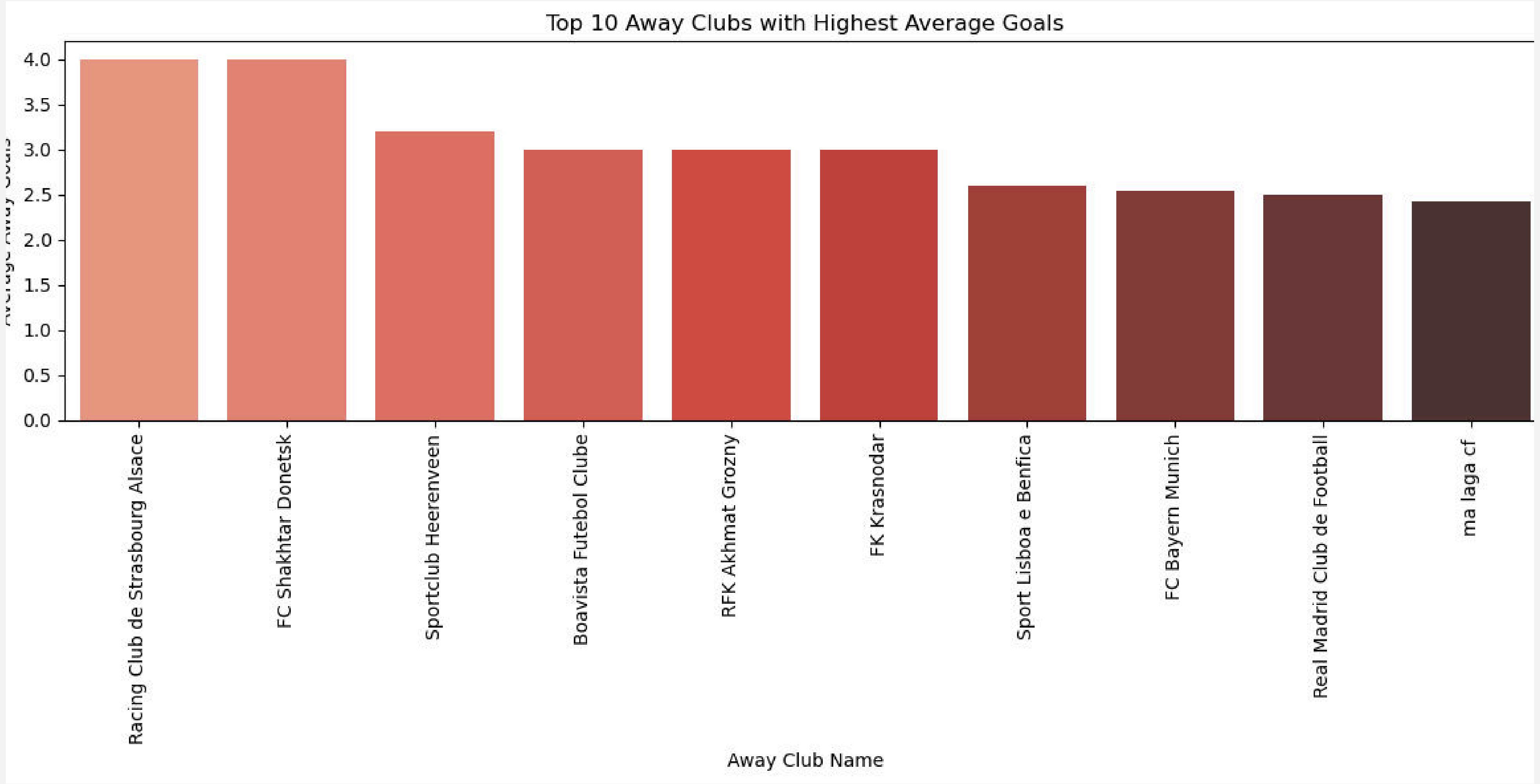


Interpretation:

- **Nottingham Forest Football Club** leads with the highest average number of home goals.
- The next few clubs, including De Graafschap Doetinchem, FK Zarya Lugansk, and Sonderjyske FC, have similar average home goals.
- The remaining clubs, such as Fenerbahce Football Club, Academica Coimbra, and Futbol'nyy Klub Dynamo Kyiv, have slightly lower average home goals.
- Real Madrid Club de Futbol, Futbol Club Barcelona, and FC Bayern Munich have the lowest average home goals among the top 10.

Team Comparison

Top 10 Away Clubs with Highest Average Goals



Interpretation:

- **Racing Club de Strasbourg Alsace** leads with the highest average number of away goals.
- The next few clubs, including FC Shakhtar Donetsk and Sportclub Heerenveen, have similar average away goals.
- The remaining clubs, such as Boavista Futebol Clube, RFK Akhmat Grozny, and FK Krasnodar, have slightly lower average away goals.
- Sport Lisboa e Benfica, FC Bayern Munich, Real Madrid Club de Futbol, and Malaga CF have the lowest average away goals among the top 10.

Team Comparison

T-test for Two-Sample Comparison between the goals scored by home teams and away teams
Assumptions:
Null Hypothesis (H0): There is no significant difference in the average goals scored by home and away teams.
Alternative Hypothesis (H1): There is a significant difference in the average goals scored by home and away teams.

Outcome:

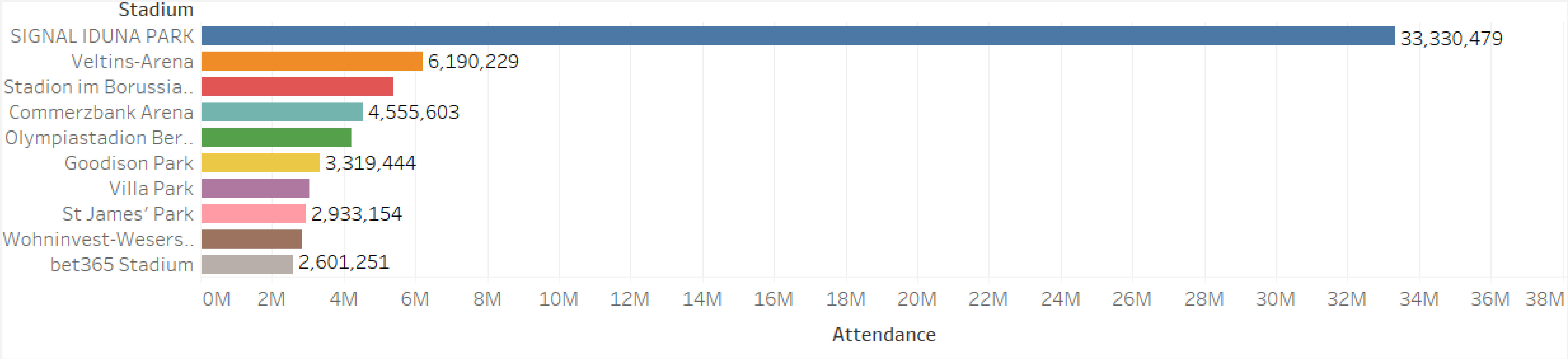
T-test for Two-Sample Comparison (Home vs Away Goals):
T-statistic: 10.168129153700171
P-value: 3.6791444496545895e-24
Reject the null hypothesis: There is a significant difference in the goals scored by home and away teams.
Result:
If P-value < 0.05, we reject the null hypothesis, meaning there is a significant difference.
If P-value ≥ 0.05, we fail to reject the null hypothesis, meaning no significant difference.

Interpretation:

Since the p-value is less than 0.05, we reject the null hypothesis, which means there is strong evidence to conclude that there is a significant difference in the goals scored by home and away teams.

Attendance and Stadium Analysis

Top 10 Stadiums with Highest Attendance



Interpretation:

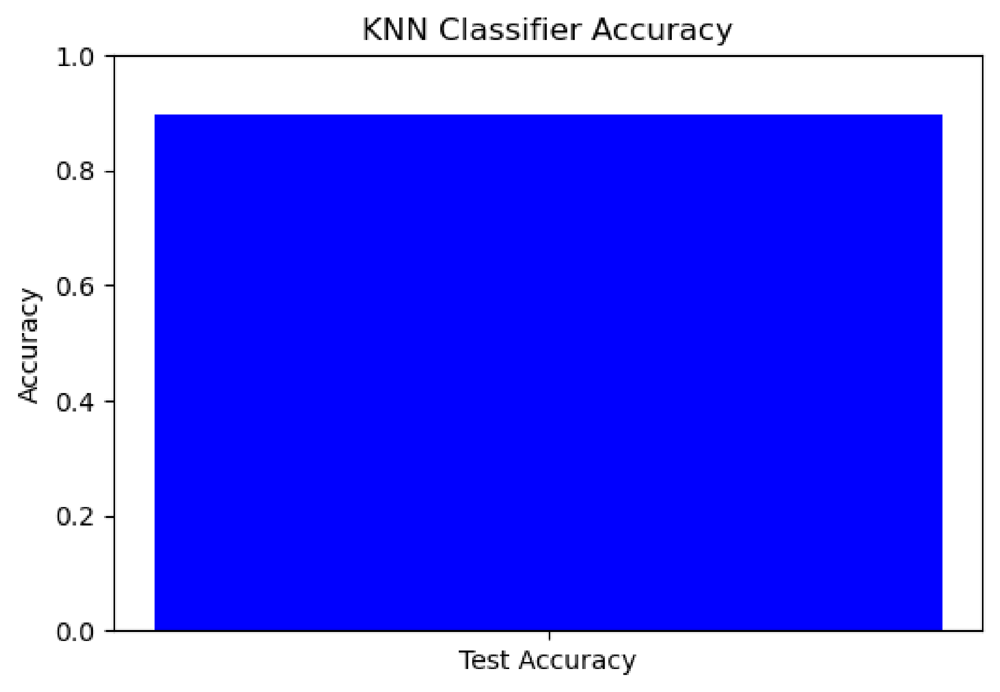
- **Signal Iduna Park** has the highest attendance with a total of 33,330,479.
- Veltins-Arena comes in second with 6,190,229 in attendance.
- Stadion im Borussia-Park has 4,555,603 in attendance.
- The remaining stadiums on the chart have attendance ranging from 2,933,154 to 2,601,251.

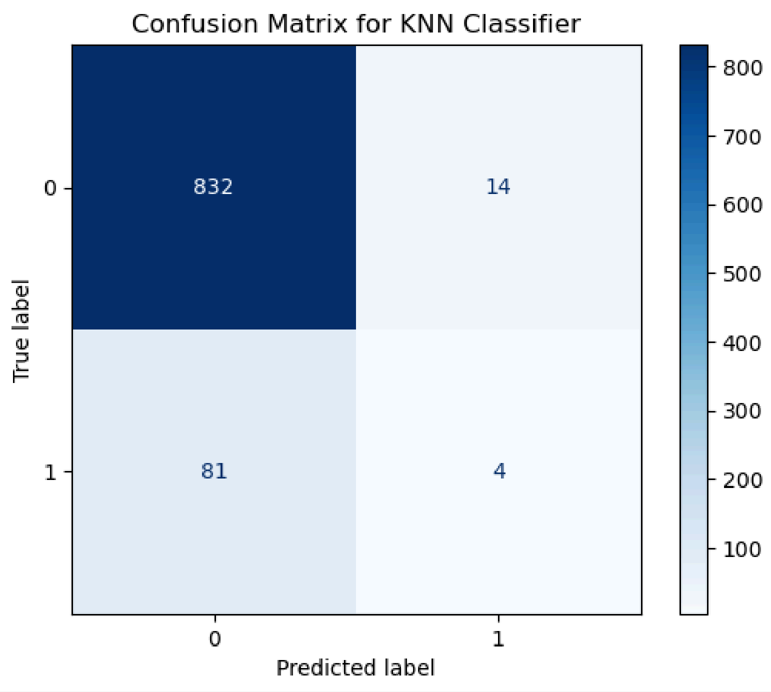
Attendance and Stadium Analysis

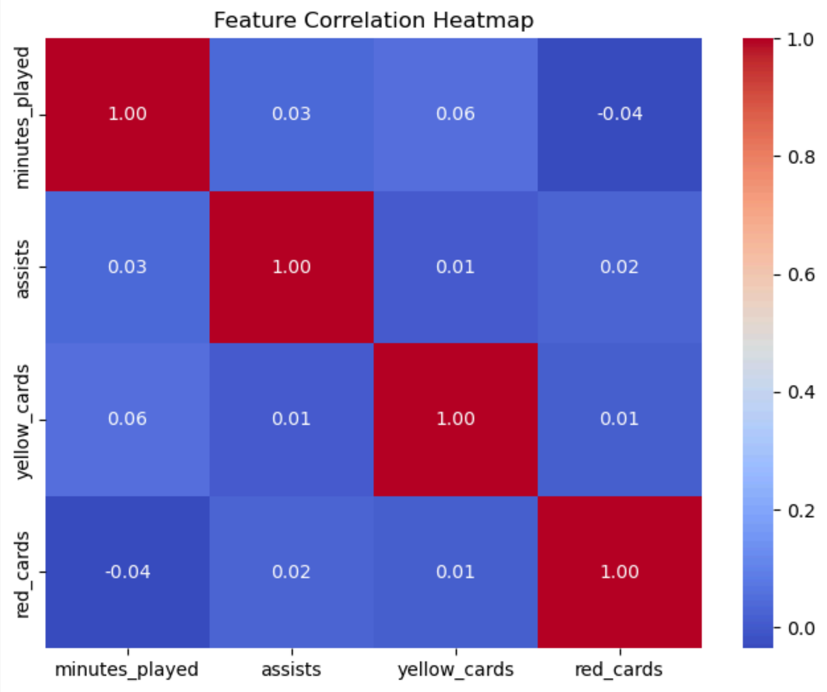
Using K-Nearest Neighbor (KNN) Classification

Can we predict if a player will score a goal based on previous performance data such as minutes played, assists, and goals?

features = df[['minutes_played', 'assists', 'yellow_cards', 'red_cards']]
target = df['goal_scored']







Interpretation:

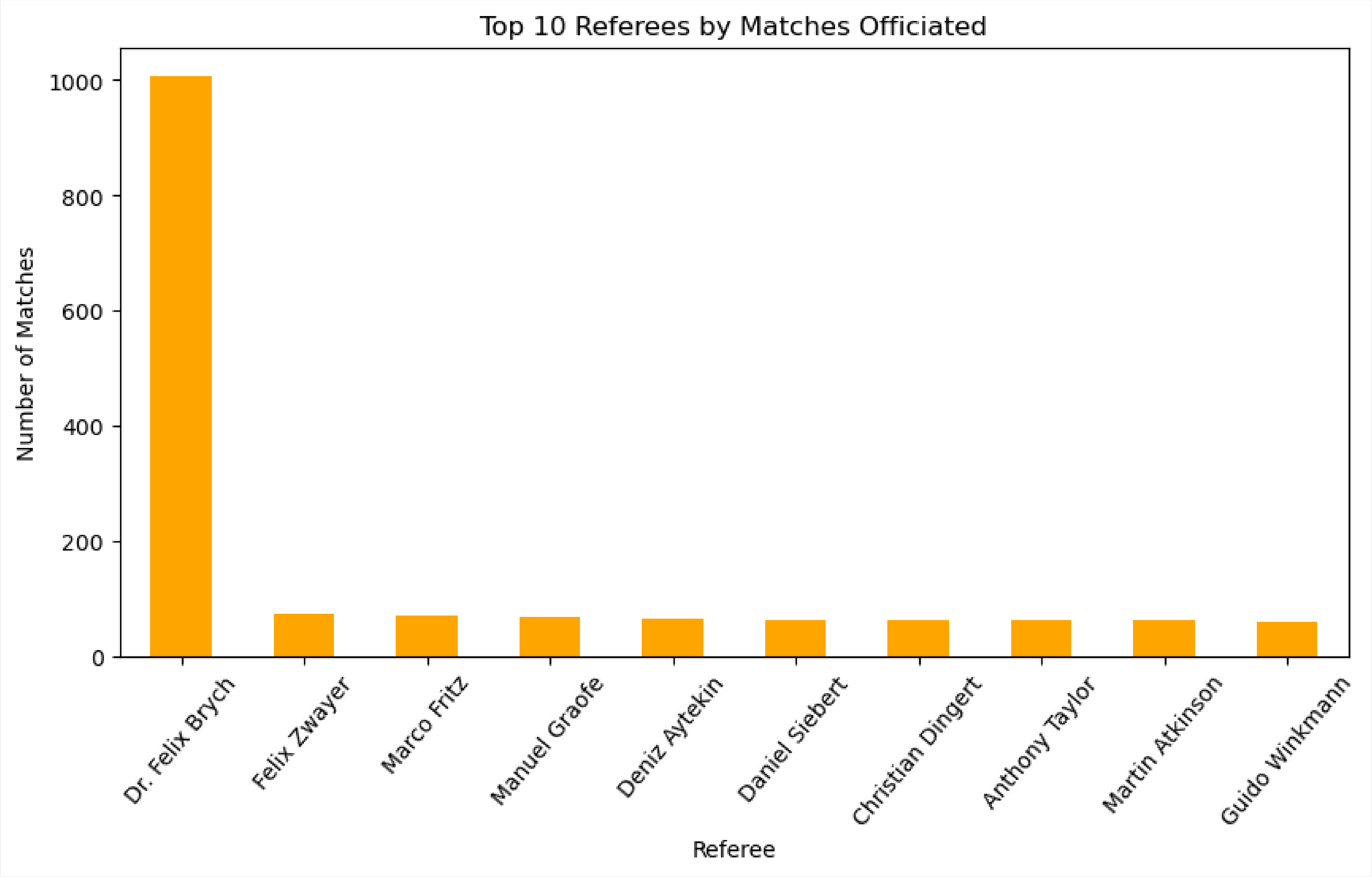
Accuracy: The accuracy score of 0.8979 represents a high level of performance. This means that the KNN classifier correctly predicted the class labels for approximately 89.79% of the instances in the test dataset.

Confusion Matrix:

- True Positives (TP): 832 - The model correctly predicted that 832 players would score a goal.
- True Negatives (TN): 4 - The model correctly predicted that 4 players would not score a goal.
- False Positives (FP): 14 - The model incorrectly predicted that 14 players would score a goal when they did not.
- False Negatives (FN): 81 - The model incorrectly predicted that 81 players would not score a goal when they did.

Referee Analysis

Top 10 referees by number of matches

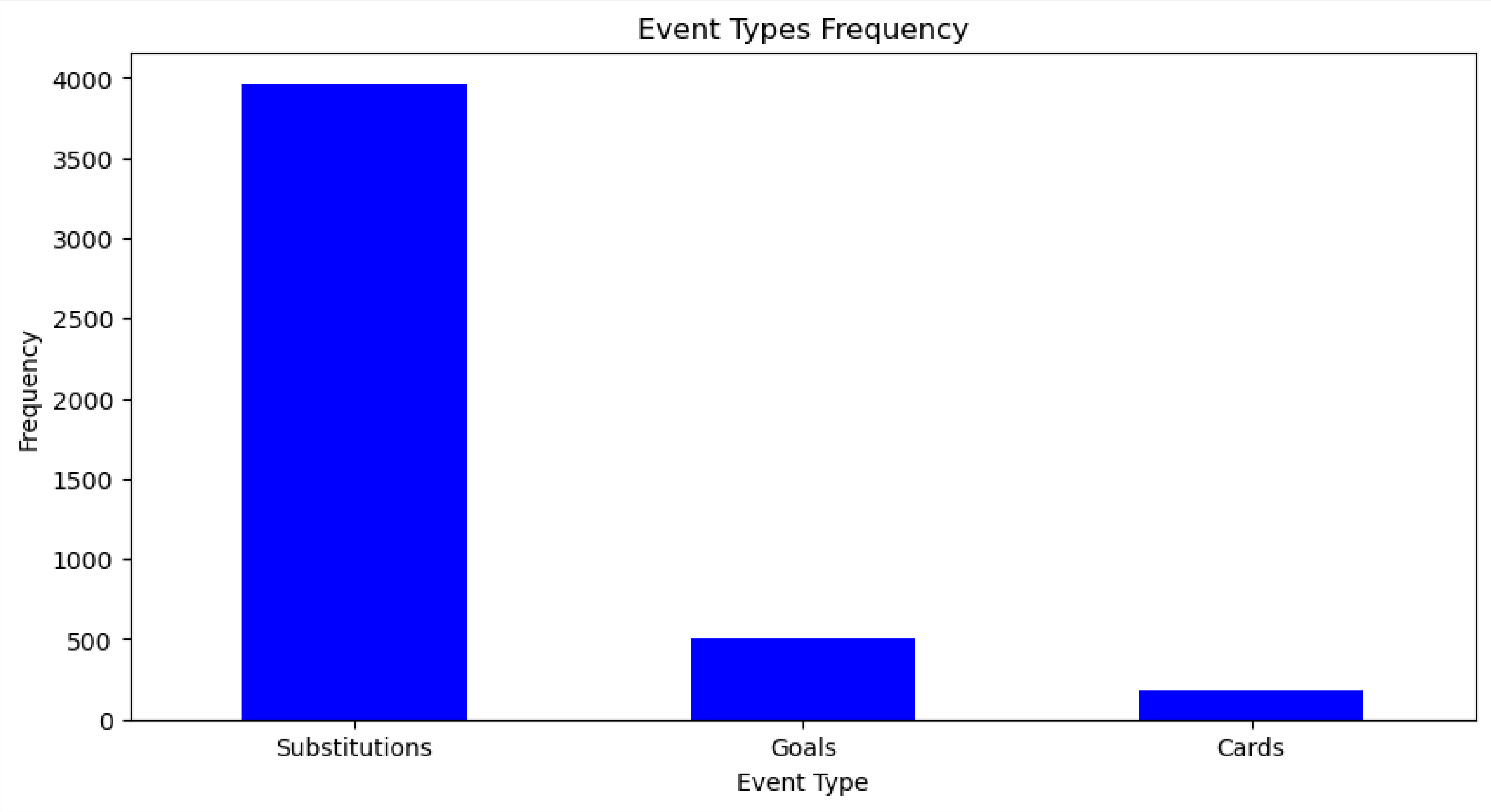


Interpretation:

- **Dr. Felix Brych** has officiated the most matches among the top 10 referees.
- Felix Zwayer and Marco Fritz have officiated the second and third highest number of matches, respectively.
- The remaining referees, including Manuel Graefe, Deniz Aytekin, Daniel Siebert, Christian Dingert, Anthony Taylor, Martin Atkinson, and Guido Winkmann, have officiated fewer matches compared to the top three.

Event frequency

Event type frequency

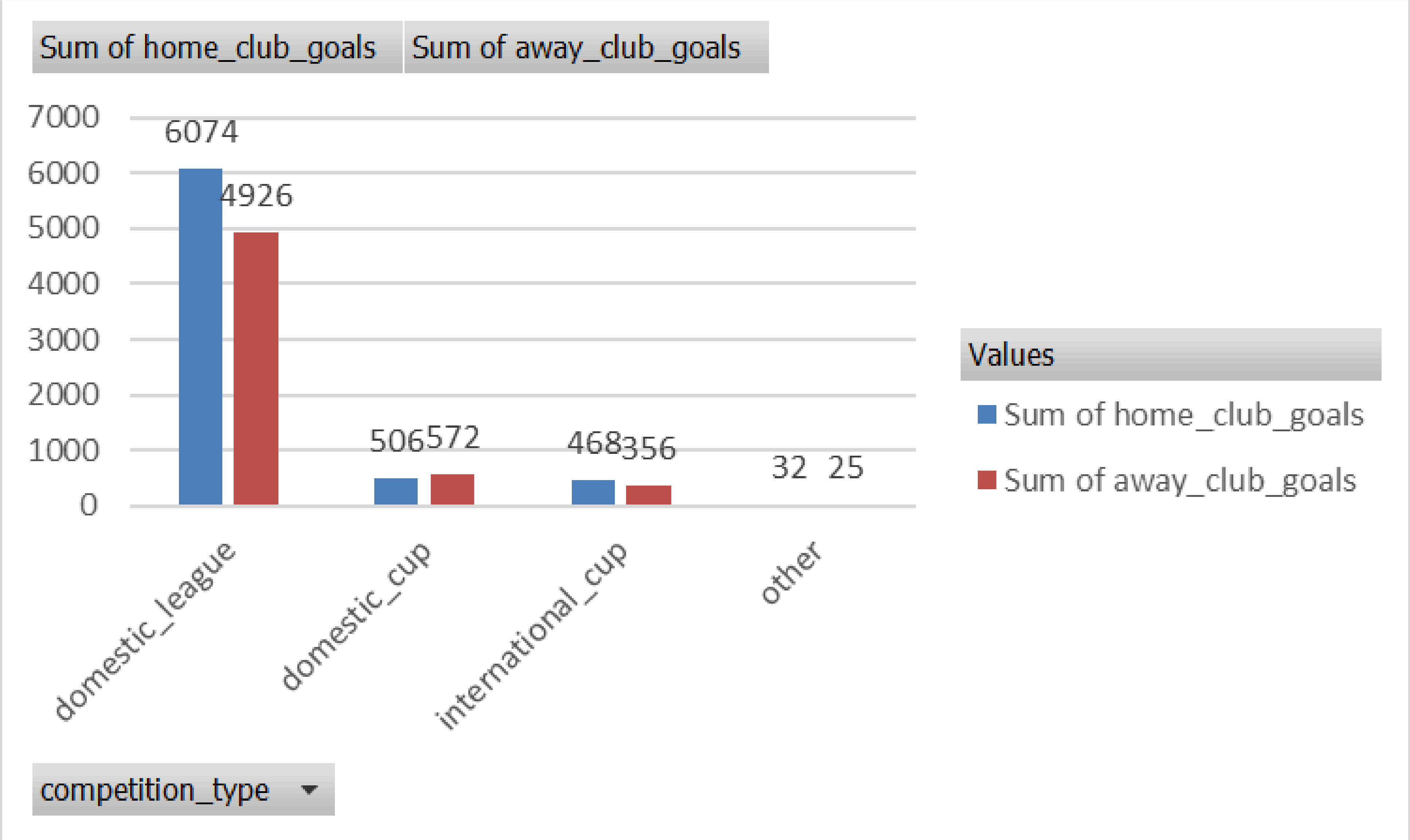


Interpretation:

- Substitutions have the highest frequency, with the bar reaching close to the 4000 mark on the y-axis.
- Goals have a significantly lower frequency compared to substitutions, with the bar height around 500.
- Cards have the lowest frequency among the three event types, with the bar height around 250.

Competition Analysis

Which competition type has highest goals scored in home club and away club?

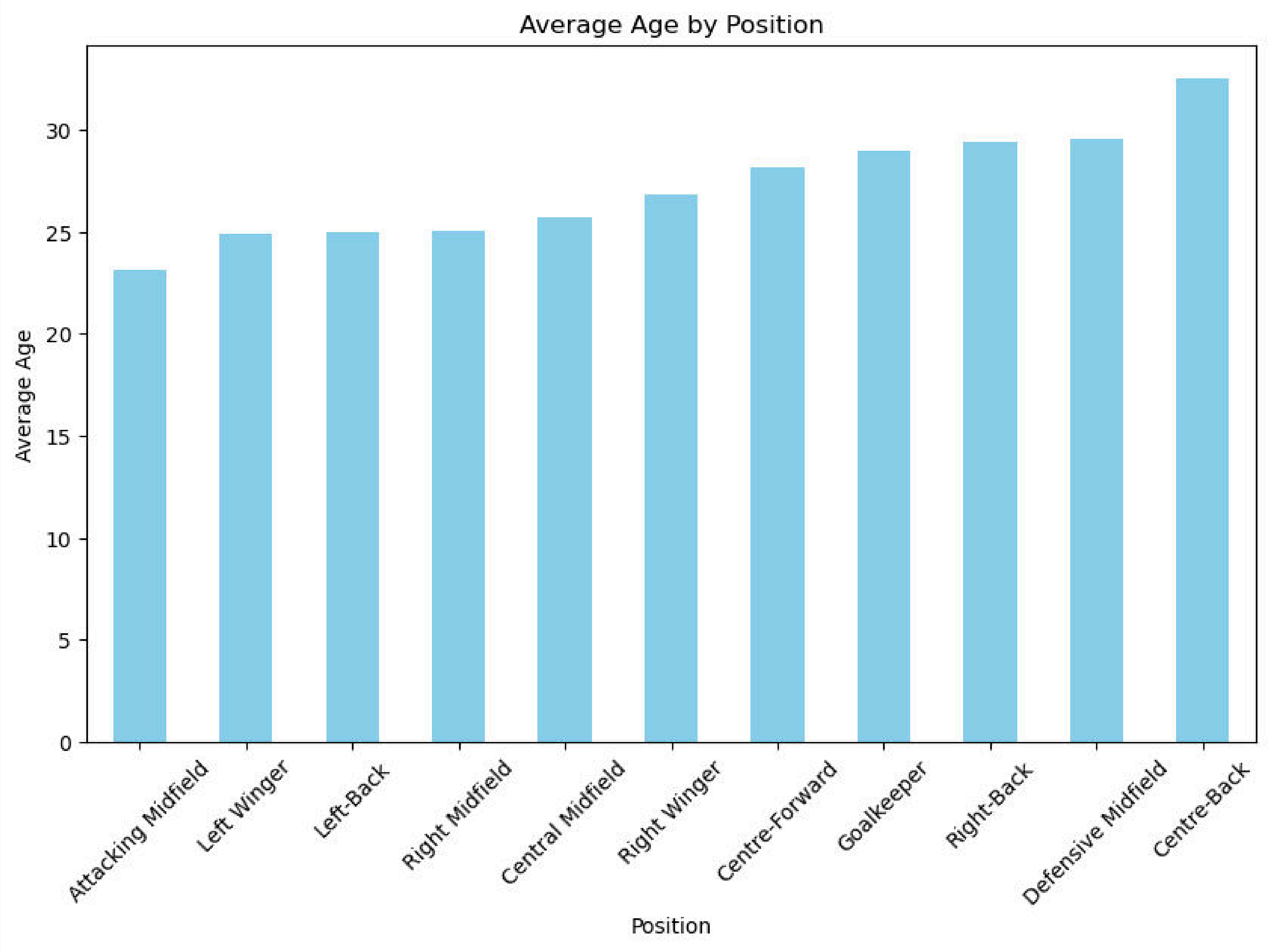


Interpretation:

- **Domestic League:** This competition type has the highest number of both home and away goals.
- **Domestic Cup:** Shows a significant number of goals, though lower than the domestic league.
- **International Cup:** Has a considerably lower number of both home and away goals compared to the domestic competitions.
- **Other:** Shows the least number of goals, both home and away.

Player Attributes and Demographics

Average Age of Players by Position

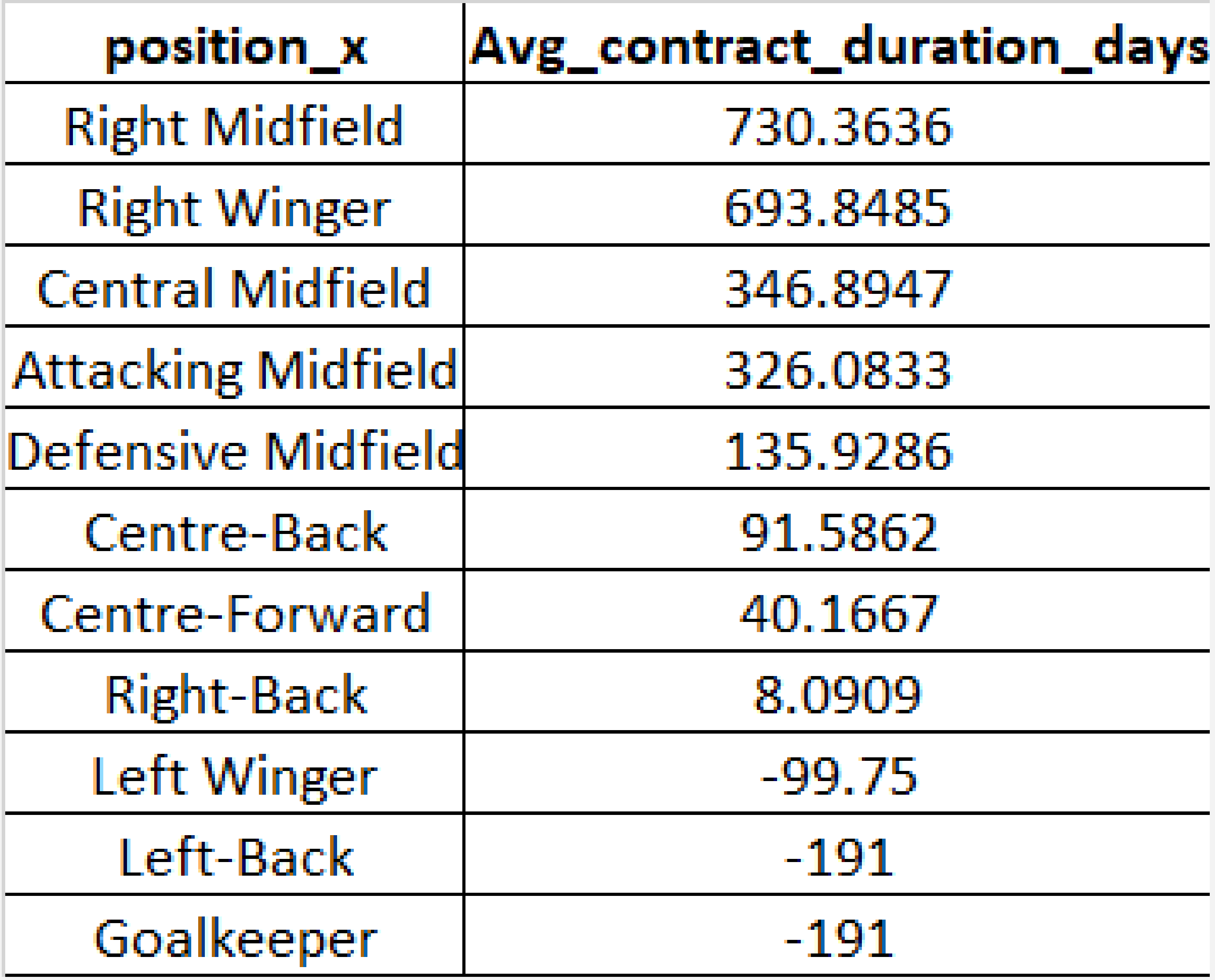


Interpretation:

- Centre-Back has the highest average age among the player positions.
- Defensive Midfield is the second highest.
- Attacking Midfield has the lowest average age.

Contract Management:

Which positions have the highest average contract duration?

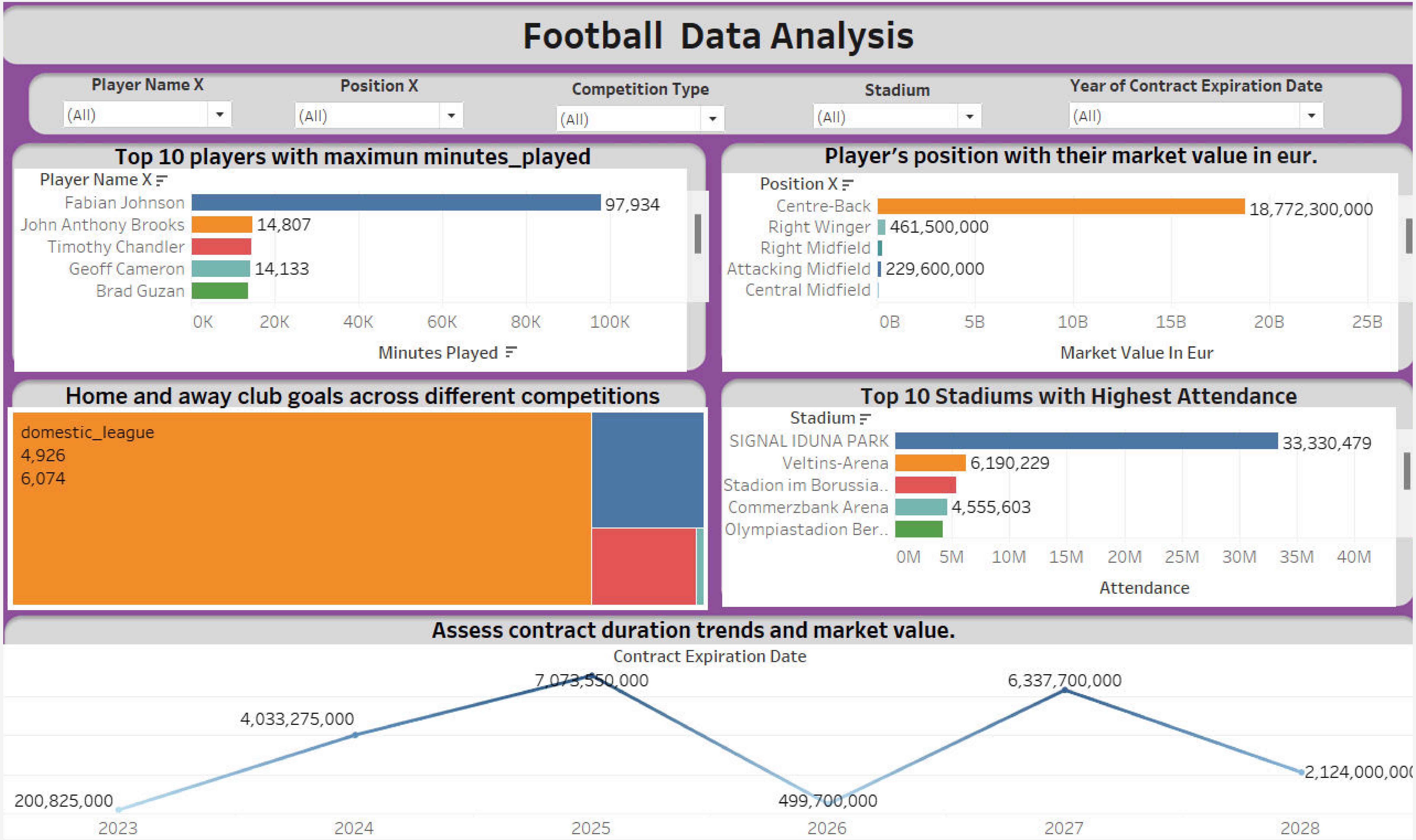


position_x	Avg_contract_duration_days
Right Midfield	730.3636
Right Winger	693.8485
Central Midfield	346.8947
Attacking Midfield	326.0833
Defensive Midfield	135.9286
Centre-Back	91.5862
Centre-Forward	40.1667
Right-Back	8.0909
Left Winger	-99.75
Left-Back	-191
Goalkeeper	-191

Interpretation:

- Right Midfielders have the longest average contract duration at 730.36 days.
- Right Wingers come in second with an average contract duration of 693.85 days.
- Central Midfielders have an average contract duration of 346.89 days.
- Attacking Midfielders have an average contract duration of 326.08 days.
- Defensive Midfielders have an average contract duration of 135.93 days.
- Centre-Backs have an average contract duration of 91.59 days.
- Centre-Forwards have an average contract duration of 40.17 days.
- Right-Backs have an average contract duration of 8.09 days.
- Left Wingers and Left-Backs have negative average contract durations (-99.75 and -191 days, respectively).
- Goalkeepers also have a negative average contract duration of -191 days.

Dashboard



Football Data Analysis

Player Name X: (All)

Position X: (All)

Competition Type: (All)

Stadium: (All)

Year of Contract Expiration Date: (All)

Conclusion:

- **Player Impact**: Goals, assists, and disciplined play significantly influence team success.
- **Market Value Trends**: Player attributes like height, position, and leadership (e.g., captaincy) correlate with market value.
- **Team Comparisons**: Goal-scoring patterns and performance vary across teams and competitions.
- **Fans and Stadium Influence**: High attendance positively impacts home team performance.
- **Referee Decisions:** Discipline patterns affect match outcomes, emphasizing referee impact.
- **Strategic Insights**: Substitution timing reveals tactical approaches in critical moments.

- This analysis provides a comprehensive view of individual player performance, team dynamics, and market trends.
- By leveraging these insights, clubs can make informed decisions on player acquisitions, match strategies, and long-term planning.
- The findings also highlight the importance of attributes like position, discipline, and fan engagement in shaping both on-field and market success.

Thank You!