English Premier League Insights: Decoding the Game Through Data

Kamal Aggarwal | English Premier League (EPL) Dataset

The English Premier League – A Legacy of Football Excellence

Founded in 1992, the Premier League stands as the pinnacle of English football. Each year, 20 clubs battle it out for the coveted title, showcasing intense rivalries and captivating a global fanbase.

Our analysis delves into common match statistics such as Goals (Full Time Home Goals/Full Time Away Goals), Match Results (Full Time Result), Fouls, and Cards to uncover hidden patterns.

What We Aimed to Discover

Turning Raw Match Data into Actionable Insights

Explore Trends & Patterns

Identify significant trends and anomalies within historical match statistics.

Understand Outcome Factors

Determine which factors contribute most to wins, draws, or losses.

Predict Match Outcomes

Develop and test Machine Learning models to forecast game results.

Design Custom Metrics

Create novel metrics, like 'match intensity', for deeper insights into game dynamics.

About Dataset

EPL Match Statistics (1993–2022)

This dataset captures match-level data from the **English Premier League** over nearly **three decades (1993 to 2022)**. It includes match outcomes, scoring statistics, team behaviors, and referee decisions.Sourced from **Kaggle**, the dataset provides a robust foundation to explore patterns, predict results, and understand football through data.

Dataset Structure

- 🧾 **Total Matches:** 11,000+
- 📅 **Seasons Covered:** 1993 to 2022
- 🔢 **Columns:** 23
- 📁 **File Format:** CSV

Inside the Dataset – Structure & Preparation

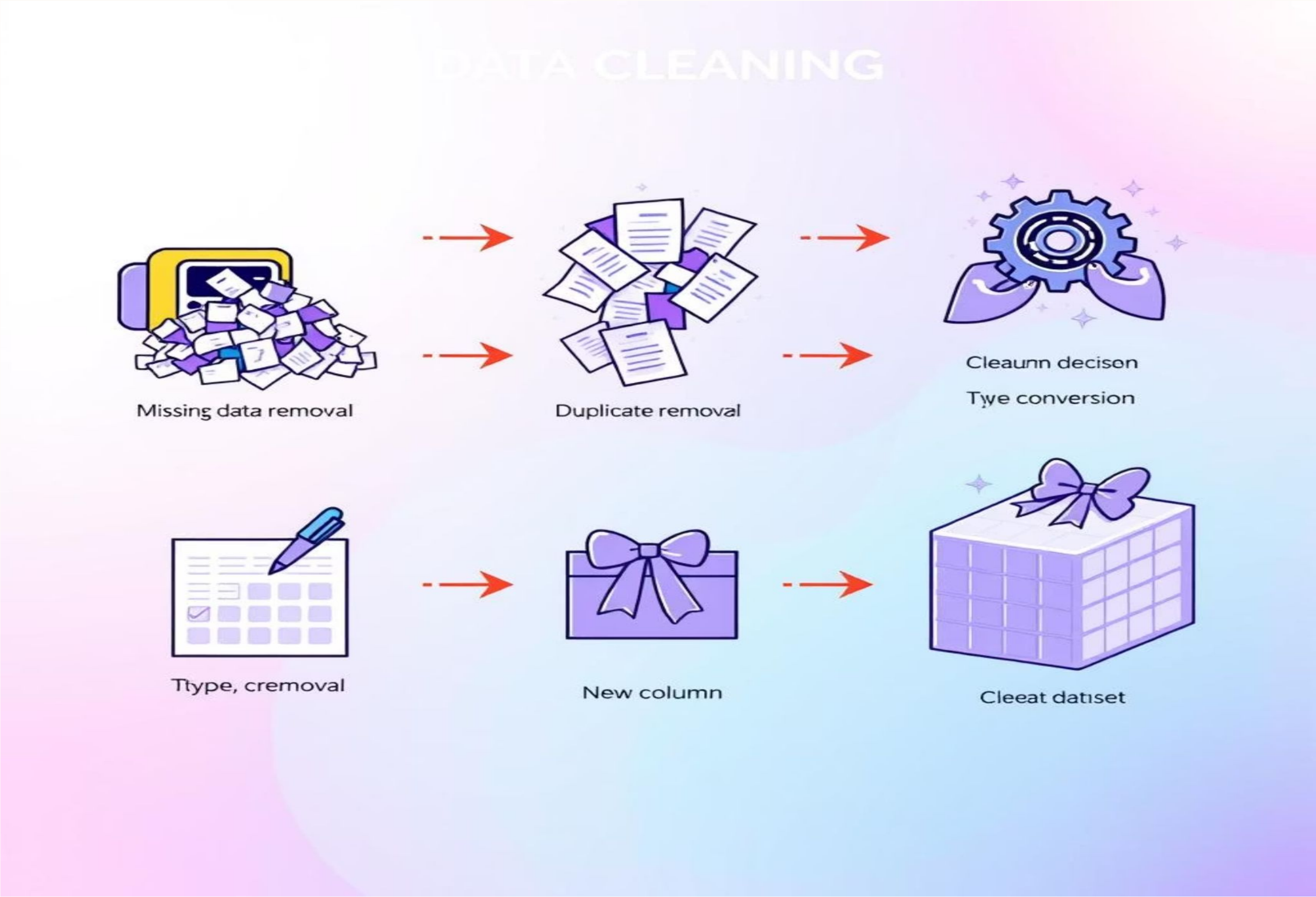
11,000+ Matches, 29 Seasons, 23 Features

Our dataset comprises over 11,000 matches spanning 29 seasons, featuring 23 distinct statistical features per game. Rigorous preprocessing was essential to ensure data quality and reliability.

- Cleaned approximately 25% missing data entries.
- Converted data types, notably 'DateTime' to datetime objects.
- Removed duplicate entries to maintain data integrity.
- Generated new columns like 'Year' and 'Date' for time-series analysis.

Code Snippet: Data Cleaning

```
df['DateTime'] = pd.to_datetime(df['DateTime'])df['Year']
= df['DateTime'].dt.yeardf.dropna(inplace=True)
```



Descriptive Statistics – What's the Game Made Of?

Mean, Median, Mode Tell a Story Too

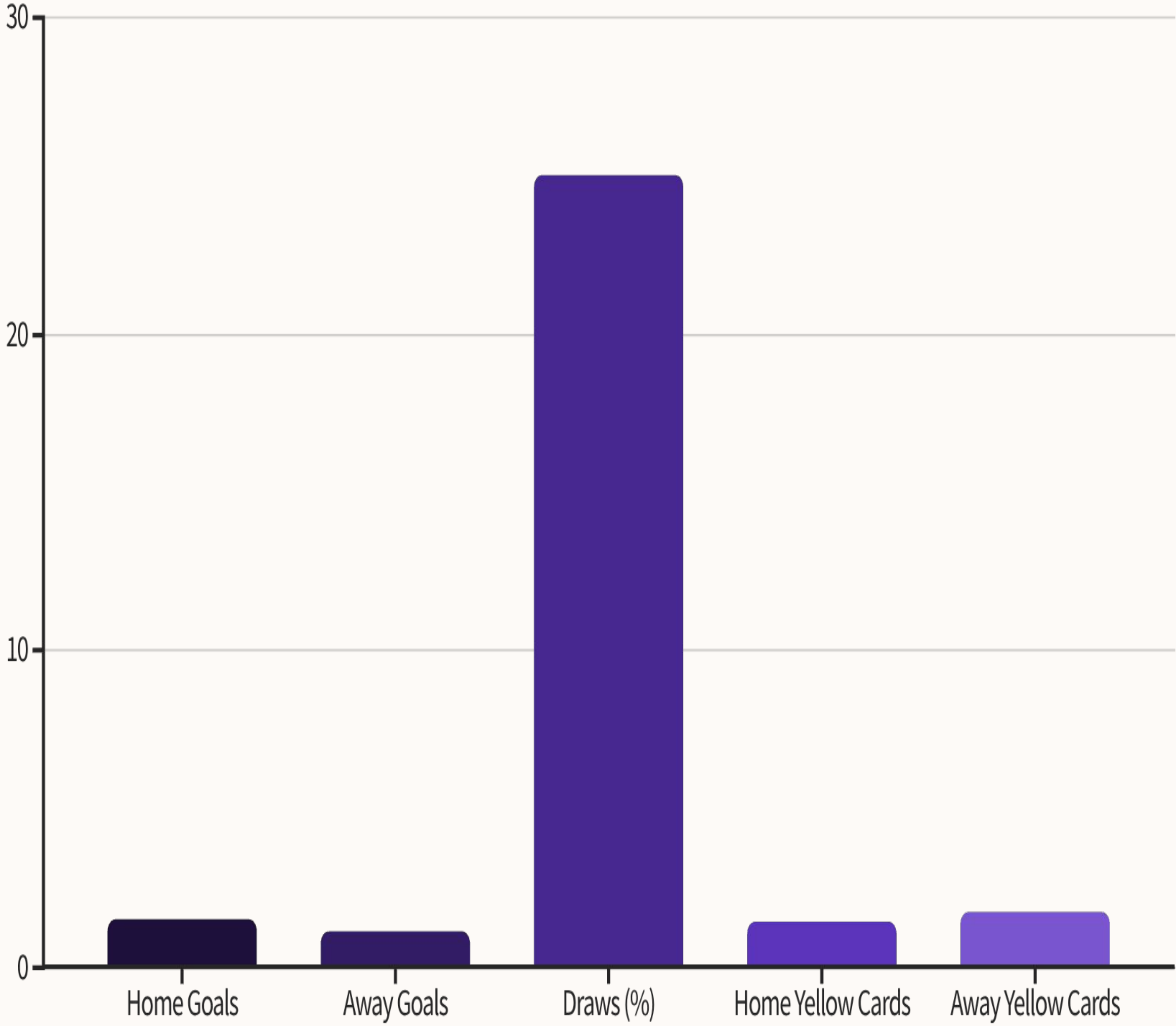
Average Home Goals: Approximately 1.51 goals per match.

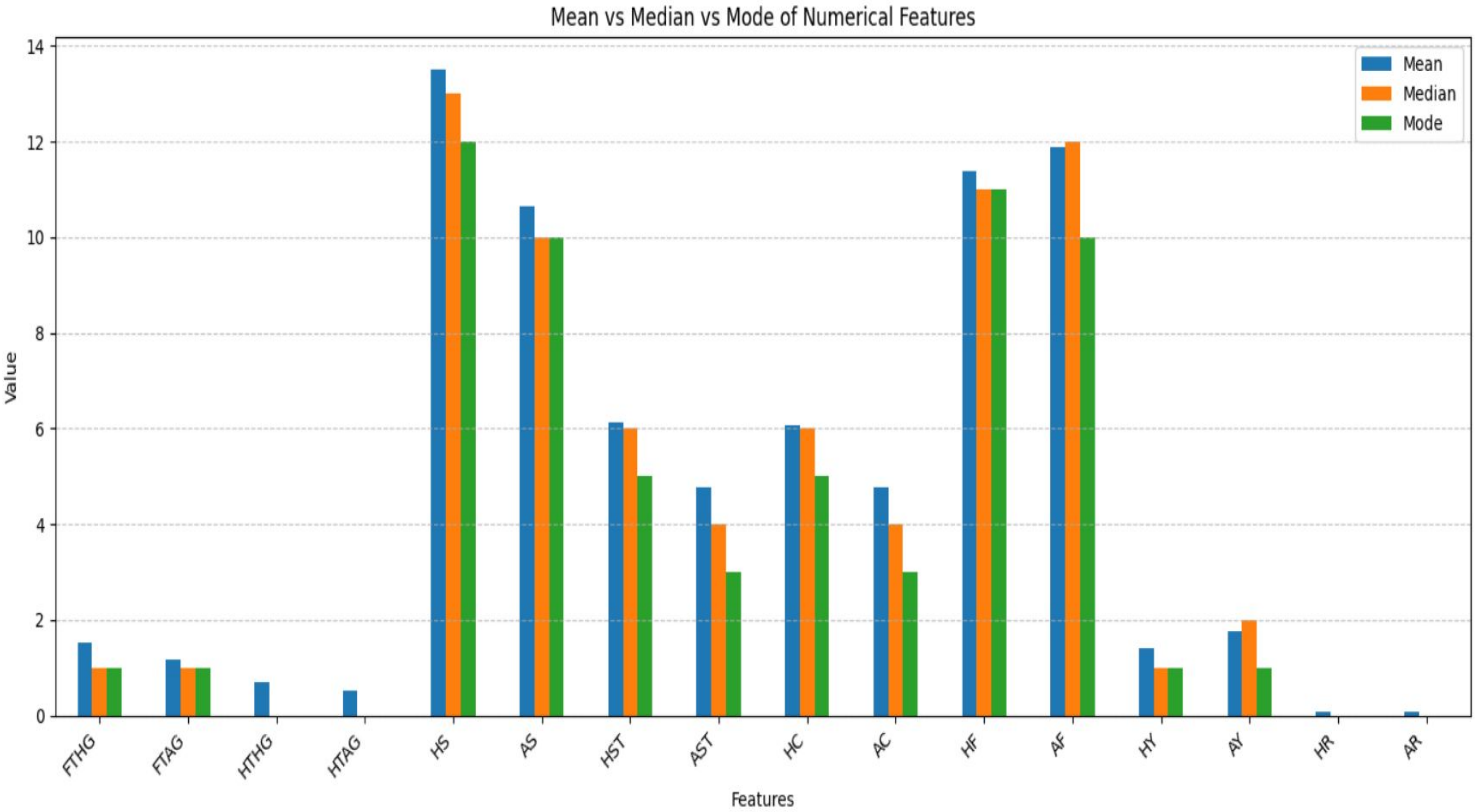
Average Away Goals: Around 1.14 goals per match.

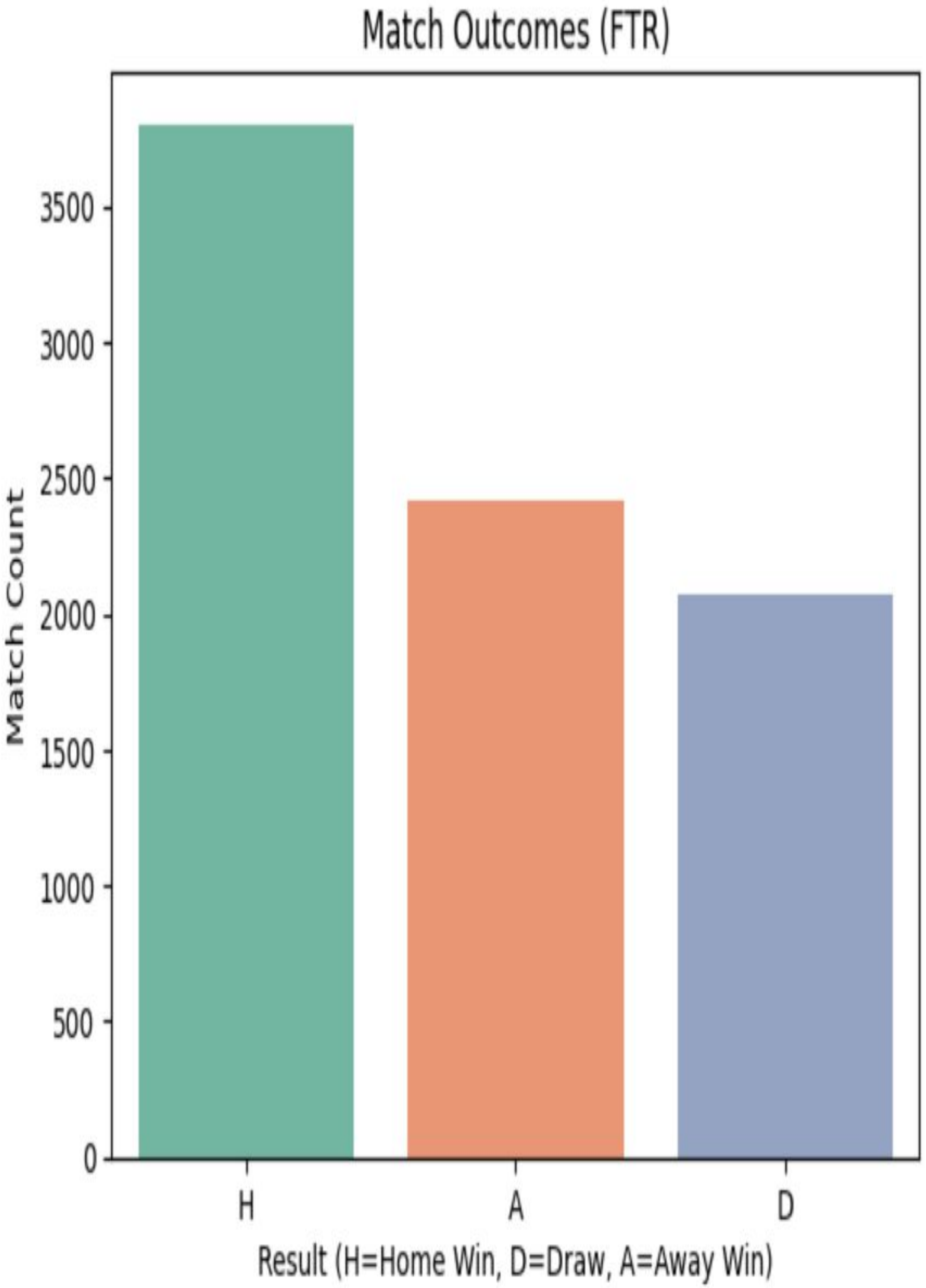
Draws: Constitute about 25% of all games played.

Average Yellow Cards: Home teams average 1.4, while away teams average 1.75.

These initial statistics highlight fundamental patterns in game dynamics, suggesting a slight home advantage in scoring and more disciplinary actions for away teams.




Mean vs Median vs Mode of Numerical Features
Mean
Median
Mode
Value
0
2
4
6
8
10
12
14
FTHG
FTAG
HTHG
HTAG
HS
AS
HST
AST
HC
AC
HF
AF
HY
AY
HR
AR
Features



What the Charts Say About the Beautiful Game

Home Advantage, Scoring Trends, Discipline Disparity

Home Field Dominance

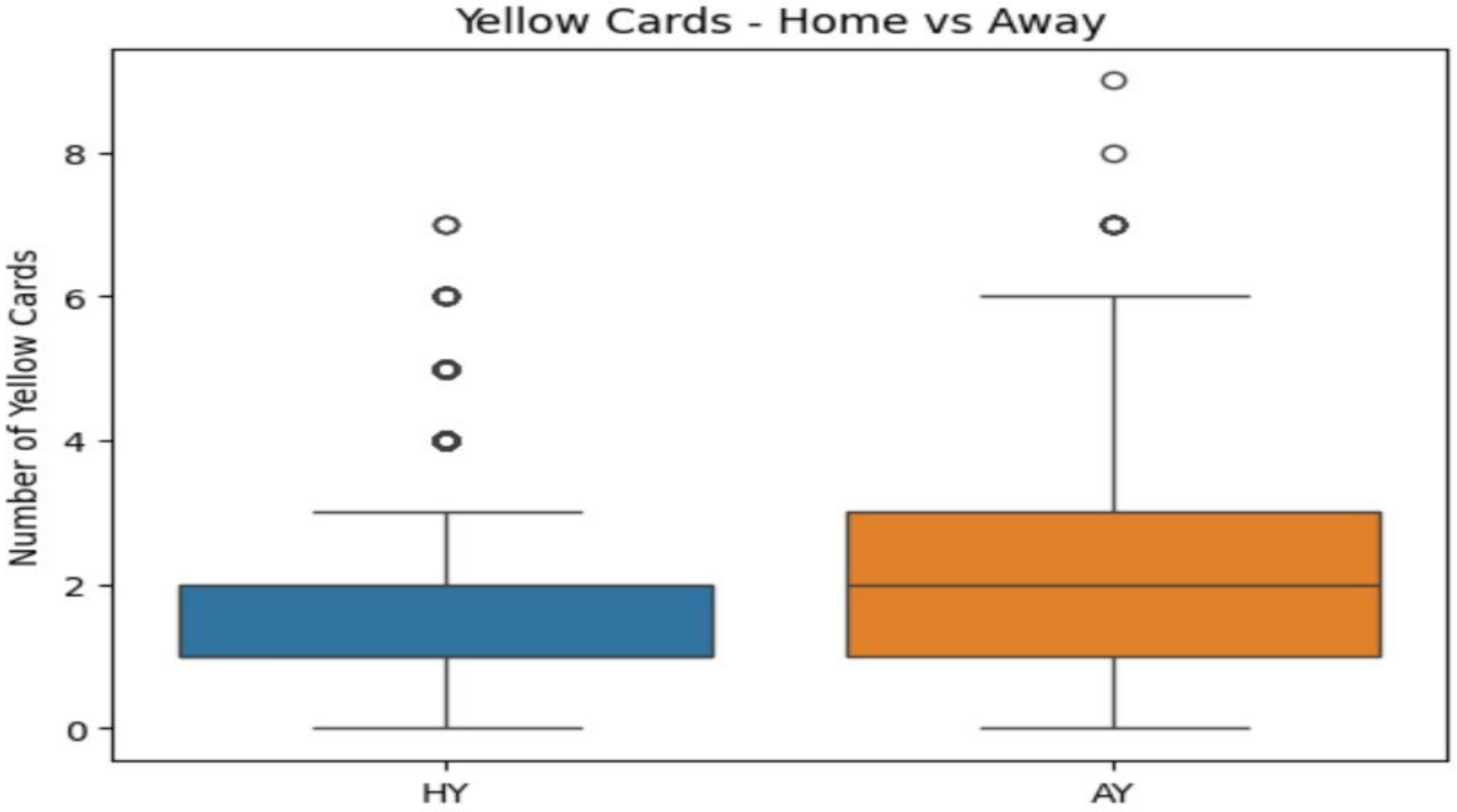
Home teams secure victory in approximately 46% of matches, highlighting a significant advantage when playing in front of their fans.

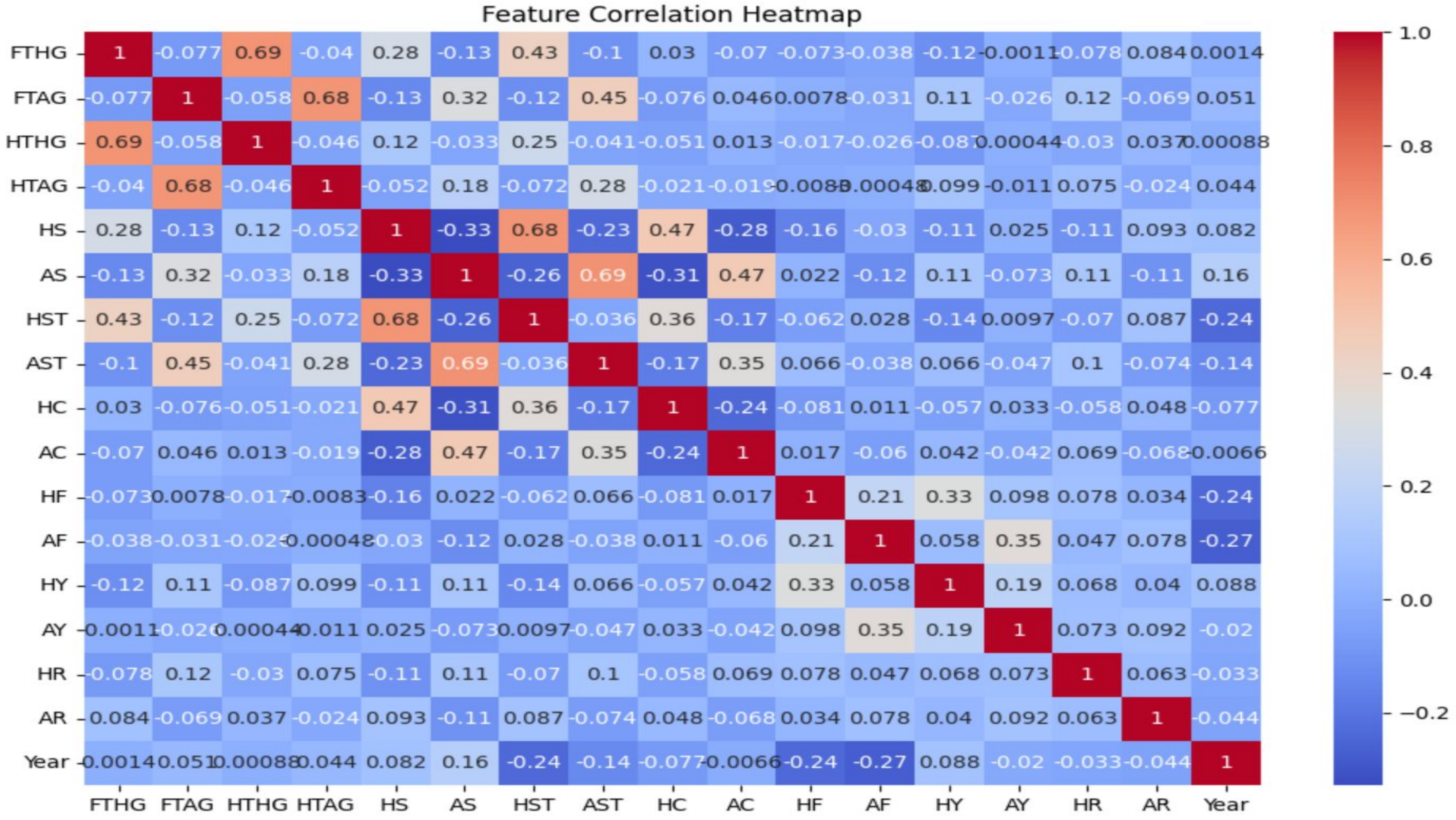
Away Team Fouls

Away teams tend to commit more fouls, suggesting a more aggressive or defensive approach on the road.

Shots Correlate with Wins

A clear positive correlation exists between the number of shots taken and the likelihood of winning a match.

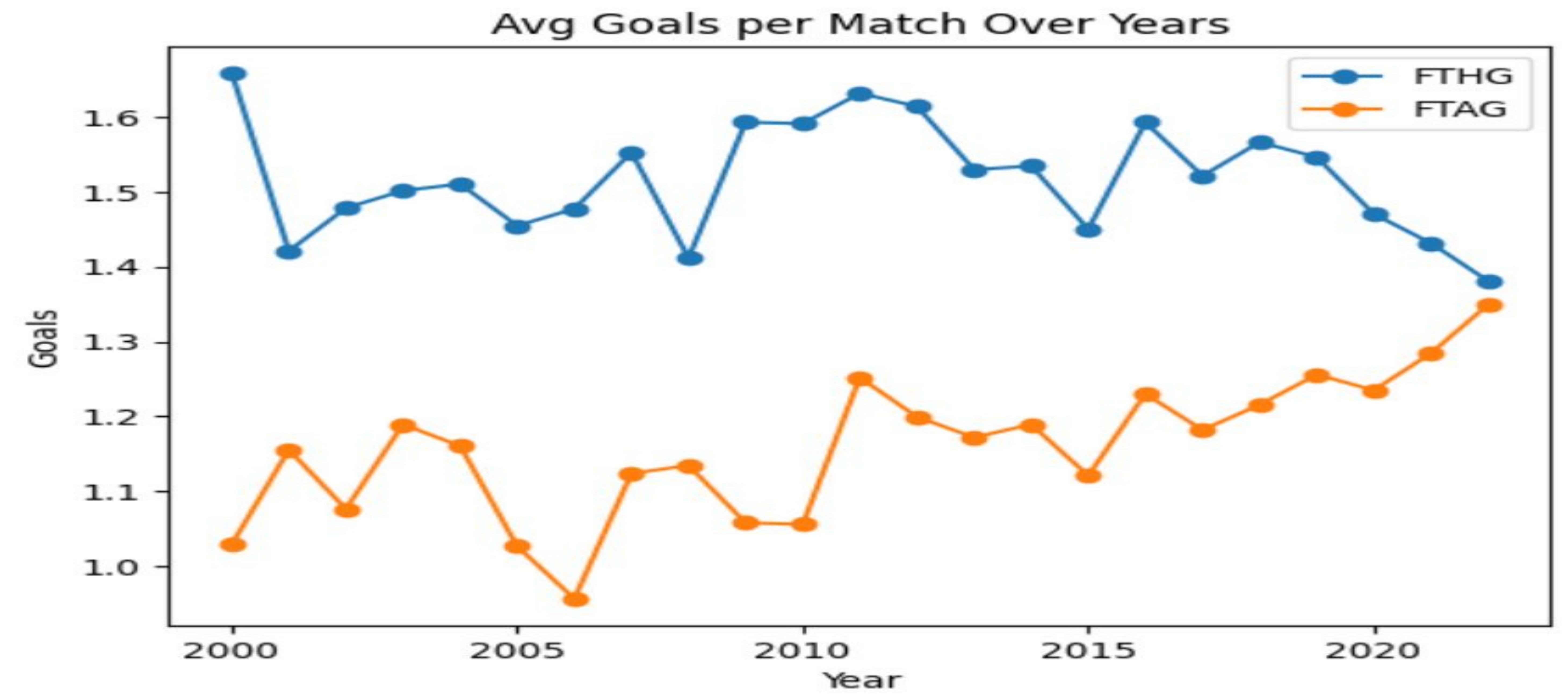

Yellow Cards - Home vs Away
Number of Yellow Cards
0
2
4
6
8
HY
AY



Feature Correlation Heatmap

Evolution of the League – Goals & Fouls by Year

A Shift Towards Clean & Attacking Football



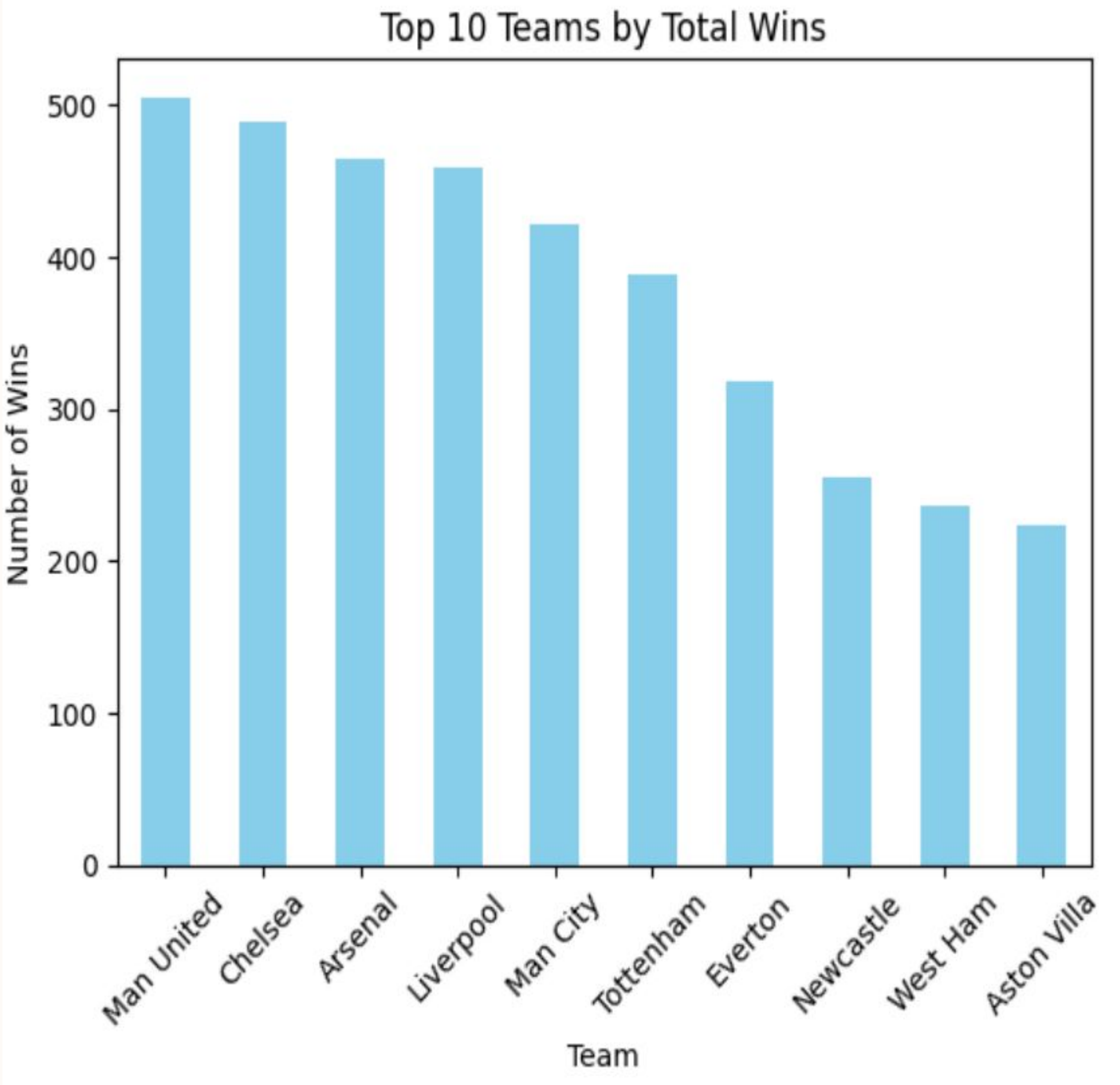
Measuring Match Intensity with a Custom Formula

Combining Goals, Fouls, Shots, and Cards

To quantify the excitement and competitiveness of each game, we developed a custom 'Match Intensity' metric. This formula aggregates various in-game actions, offering a holistic view beyond just the score.

$$MatchIntensity = FTHG + FTAG + HS + AS + HY + AY + HF + AF + HR + AR$$

Our analysis of this metric revealed that the top 10 most intense matches often involved classic rivalries, known for their fierce competition. Interestingly, high-intensity games don't always equate to high-scoring affairs, emphasizing the role of defensive battles and midfield struggles in determining game flow.



Can Stats Predict the Outcome?

Training ML Models to Classify Match Results (H/D/A)

We leveraged various Machine Learning models to predict match outcomes (Home win, Draw, Away win) based on pre-match and in-game statistics. Our goal was to identify the most effective predictive algorithms for football analytics.

58%

Logistic Regression

A baseline model providing a foundational understanding of feature impact on outcomes.

55%

Random Forest

Demonstrated improved accuracy by combining multiple decision trees, capturing complex relationships.

55%

XGBoost

Exhibited the highest performance, showcasing the power of gradient boosting for complex datasets.

Code Snippet: Model Training

```
model = RandomForestClassifier()model.fit(X_train, y_train)
```

What Drives Match Outcomes?

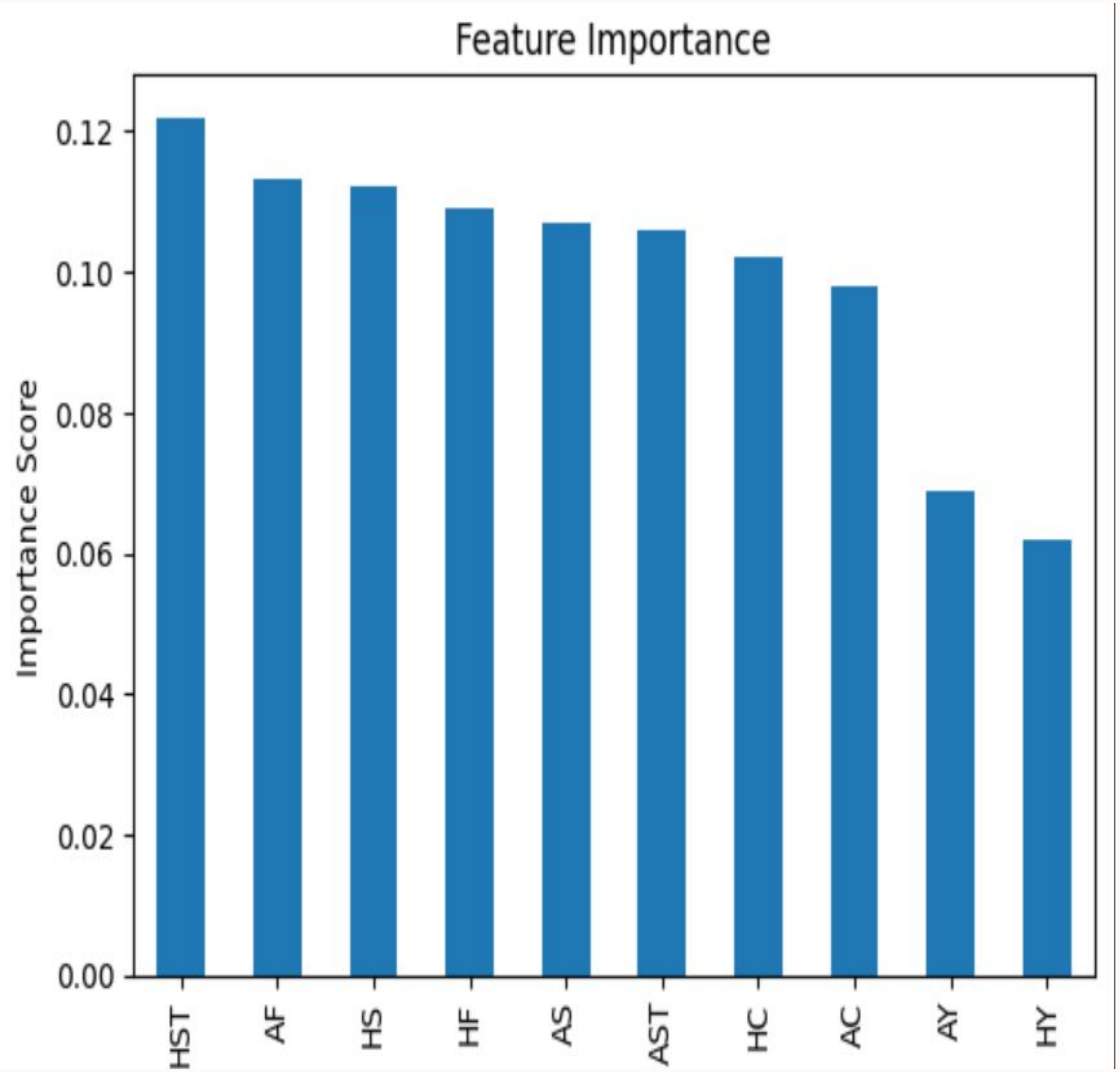
Shots on Target, Fouls, and Cards Play a Big Role

Delving into feature importance, our models highlighted several key variables that significantly influence match results, offering actionable insights for teams and analysts.

Home/Away Shots on Target: Unsurprisingly, a direct indicator of attacking prowess and goal-scoring opportunities. The more shots on target, the higher the chance of winning.

Fouls (especially away): Fouls, particularly those committed by away teams, play a crucial role. This suggests that disciplinary actions and aggressive play can disrupt flow and impact results.

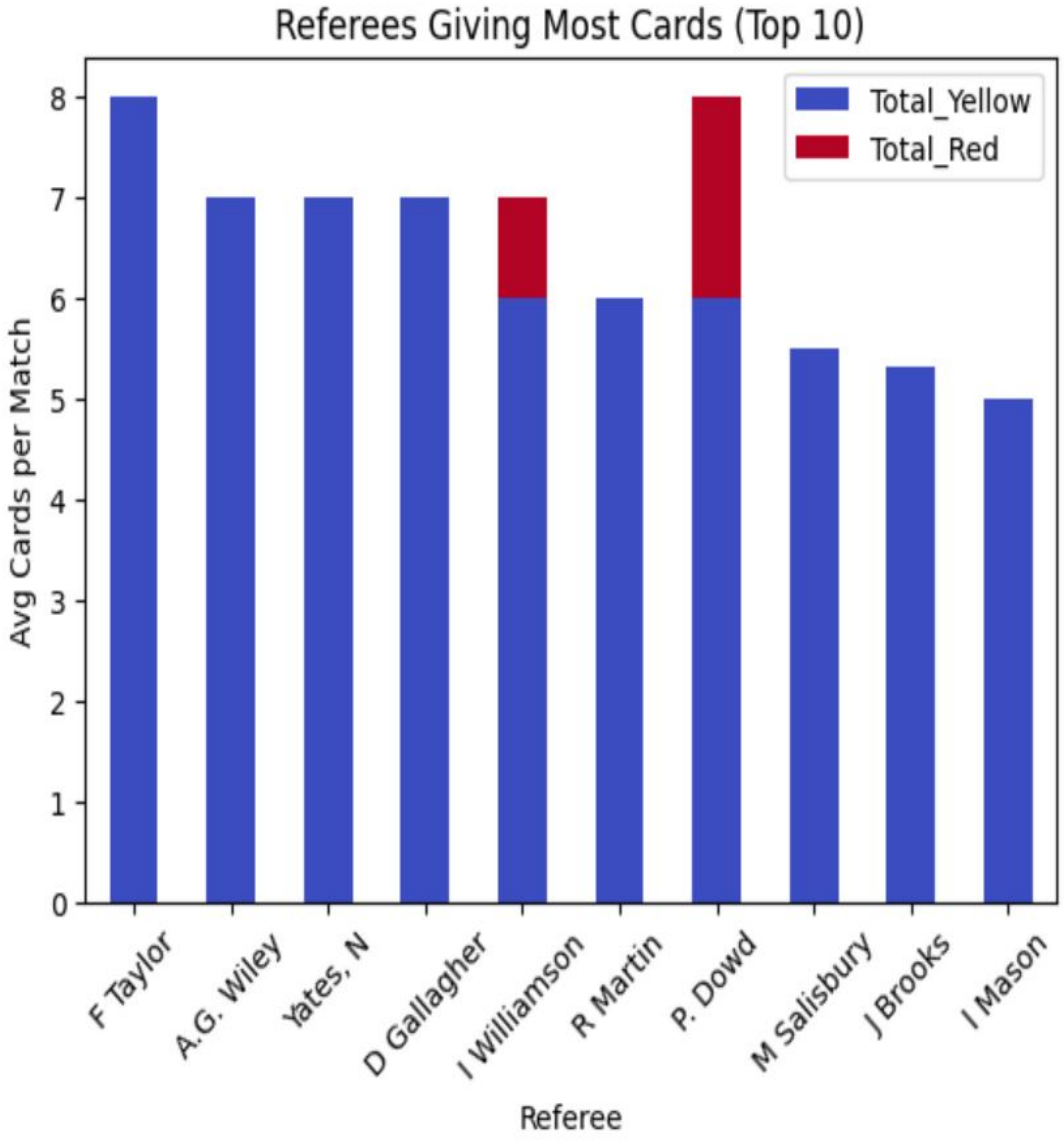
Corners: The number of corners awarded serves as a strong proxy for sustained attacking pressure, indicating a team's ability to keep the ball in dangerous areas.



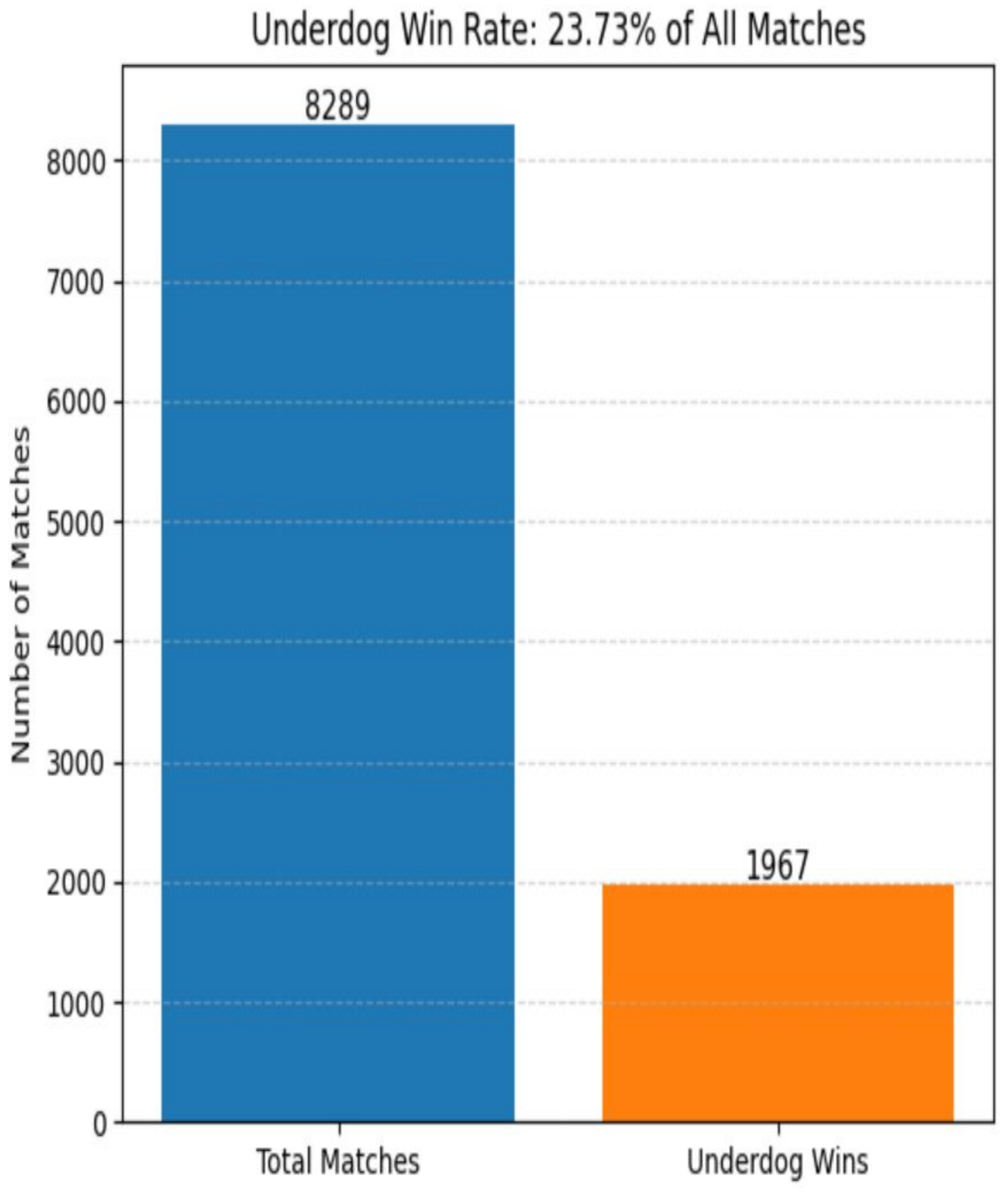
Uncovering Unusual Patterns

Delve into the subtle yet significant trends emerging from the English Premier League (EPL) data, focusing on surprising referee behaviors and the unpredictable nature of underdog victories.

- Referee X issues ~5 cards/match
- 13% of matches saw fewer shots but still a win.

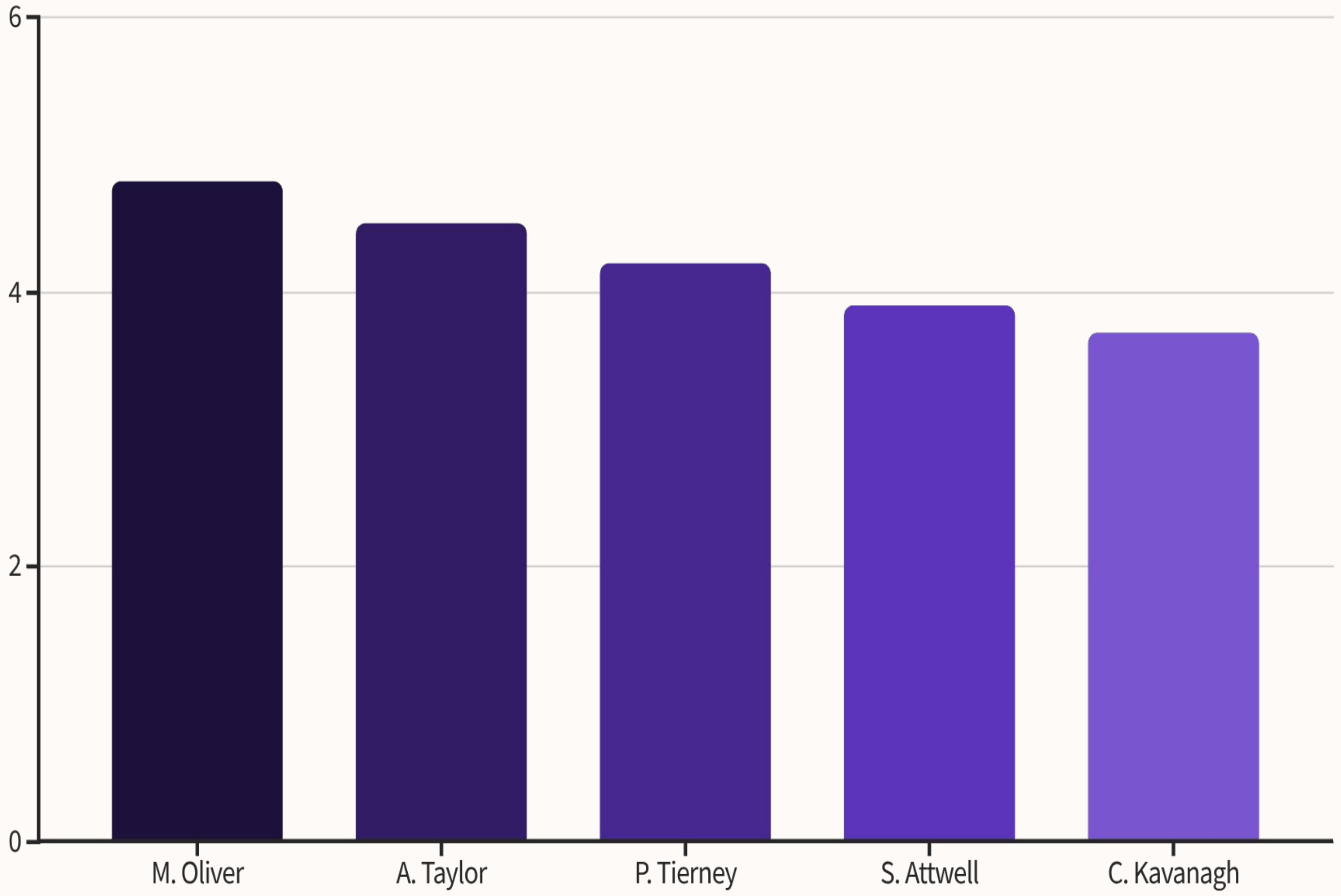


Referees Giving Most Cards (Top 10)



Underdog Win Rate: 23.73% of All Matches

Referee Tendencies: Card Distribution



Certain referees, like Michael Oliver, consistently issue more cards per match, impacting game flow and team strategy. Understanding these biases can offer a competitive edge.

Underdog Upset: Fewer Shots, More Wins

Unpredictable Outcomes

Approximately 13% of matches in the EPL witnessed the winning team having fewer shots on goal than their opponent.

Implications

challenges conventional wisdom, suggesting factors beyond dominant possession and shot count contribute significantly to match results.

Key Takeaways & Recommendations

Actionable Insights

Leverage these insights for sharper fantasy soccer picks and more informed betting strategies.

Strategic Coaching

Coaches can adapt tactics by analyzing referee tendencies and the dynamics of upset victories.

CONCLUSION

From Numbers to Strategy: Unlocking Football's Deeper Truths

Home Advantage Exists, but is narrowing over time.

Aggression Doesn't Guarantee Victory — teams with fewer fouls and cards tend to perform better.

Shots on Target Correlate Strongly with match wins, more than total possession or corners.

Referees Influence Game Flow — a few referees consistently issue more cards than others.

Underdog Wins Are Real — nearly **~13–15%** of matches are won by teams with fewer shots.

Custom Metrics (e.g., Match Intensity) add richer storytelling than just goals and results.

"This analysis proves that football isn't just a game of goals — it's a game of data, decisions, and discipline."

THANK YOU