Introduction

This project aimed to explore and analyze the performance of soccer players from the English Premier League 2023 season. Using a comprehensive dataset of 421 rows and 28 columns, we investigated factors influencing goal-scoring ability and identified key predictors for winning Player of the Match (POTM). Goalkeepers were excluded, focusing solely on outfield players. Data cleaning, exploratory data analysis (EDA), and machine learning techniques were employed to achieve the objectives.

Dataset Overview

The dataset was sourced from the official Premier League website (www.premierleague.com/stats). Key variables included performance metrics such as Goals, Assists, and Minutes Played, along with efficiency metrics like POTM and Suspensions. Columns such as Unnamed, Clean Sheets, and Goals Conceded were removed for relevance. The dataset focuses on 27 columns including:

- Performance Metrics (e.g., Goals, Assists, Playing Time)
- Efficiency Metrics (e.g., POTM, Suspensions, Yellow Cards)

Dataset Variables and Descriptions

Variable Name	Description
Nation	Nationality of the player
Local/Foreigner	Indicates if the player is English or not
Team	Team the player plays for
Position	Position (e.g., Attack, Midfield, Defence)
Age	Age of the player
MP	Number of matches played
Starts	Number of matches started
Min	Number of minutes played
90s	Number of 90-minute periods played
Gls	Number of goals scored
Ast	Number of assists
G+A	Goals plus assists
G-PK	Non-penalty goals
PKMade	Penalty goals scored
PKAttempt	Penalty attempts
CrdY	Number of yellow cards
CrdR	Number of red cards
Suspension	Number of suspensions
Gls/90	Goals scored per 90 minutes
Ast/90	Assists per 90 minutes
G+A/90	Goals plus assists per 90 minutes
G-PK/90	Non-penalty goals per 90 minutes
Gls/GM	Goals per game
Ast/GM	Assists per game
G+A/GM	Goals plus assists per game
G-PK/GM	Non-penalty goals per game
POTM	Binary flag indicating if the player won 'Player of the Match'

Research Objectives

1. Predict how features like Minutes Played and Starts affect the number of goals scored.
2. Classify whether a player wins Player of the Match (POTM) based on their performance metrics.
3. Use machine learning models to provide actionable insights for coaching and recruitment decisions.

Importance

Understanding how performance metrics such as goals, assists, playing time, and positional roles influence a player's likelihood of winning POTM is crucial for:

- Evaluating Player Performance: Identifying key metrics for success.
- Strategic Team Management: Informing decisions on substitutions and game strategies.
- Reducing Bias: Ensuring POTM awards are based on measurable performance.
- Enhancing Fan Engagement: Promoting transparency in awards criteria.
- Improving Recruitment Decisions: Identifying high-impact players.

Methodology

The project was executed in the following steps:

1. Data Cleaning:

Data cleaning is a crucial step in any data analysis project to ensure the quality and reliability of results. Here's why it was necessary:

1. **Relevance**: By dropping irrelevant columns and duplicate rows, we eliminated unnecessary noise in the dataset, focusing only on meaningful variables.

2. **Consistency**: Renaming columns and ensuring uniform data formats (e.g., converting categorical variables to proper types) helped streamline analysis, reducing errors during modeling and interpretation.

3. **Completeness**: Handling missing data ensured that the analysis could proceed without biases or inaccuracies stemming from incomplete records.

2. Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is essential to uncover patterns, detect anomalies, and understand the relationships within the data. Specifically, we used EDA to:

1. **Gain Initial Insights**: Visualizing the Age distribution by Position using Box Plots allowed us to identify trends like the older average age for Defenders and Midfielders.
2. **Highlight Key Differences**: The Bar Graph showing Goals scored by each Team revealed the dominance of teams like Manchester City and Liverpool, as well as the struggles of Wolves and Southampton.
3. **Understand Relationships**: The Correlation Heatmap helped us identify moderately positive relationships, such as between Goals and Assists (0.49) or Minutes Played (0.38), which informed our feature selection for modeling.
4. **Spot Independence**: Scatter plots for Age and Yellow Cards indicated no strong correlation, helping refine assumptions about discipline being age-independent.
5. **Prepare for Modelling**: EDA allowed us to confirm that our selected features had potential predictive power, guiding us in choosing and fine-tuning machine learning models.

3. Machine Learning Models:

- Linear Regression for predicting Goals based on performance metrics.
- Decision Tree for classifying POTM.
- Logistic Regression for binary classification of POTM.

Machine Learning Models and Justification

Machine learning models were selected to address specific research questions and provide actionable insights. Here's why each model was used:

1. **Linear Regression**:
    - **Purpose**: To predict the number of goals scored based on quantitative performance metrics such as Minutes Played, Assists, and Position.
    - **Why**: Linear Regression is effective for modelling relationships between continuous variables, making it suitable for predicting goals.
2. **Decision Tree**:
    - **Purpose**: To classify whether a player wins Player of the Match (POTM) based on their performance metrics.
    - **Why**: Decision Trees provide a clear and interpretable way to understand decision boundaries and thresholds, such as the number of Goals required for POTM.
3. **Logistic Regression**:
    - **Purpose**: To predict whether a player wins POTM as a binary classification problem.
    - **Why**: Logistic Regression is robust for binary outcomes, and its coefficients provide interpretable insights into the relative importance of features like Goals and Assists.

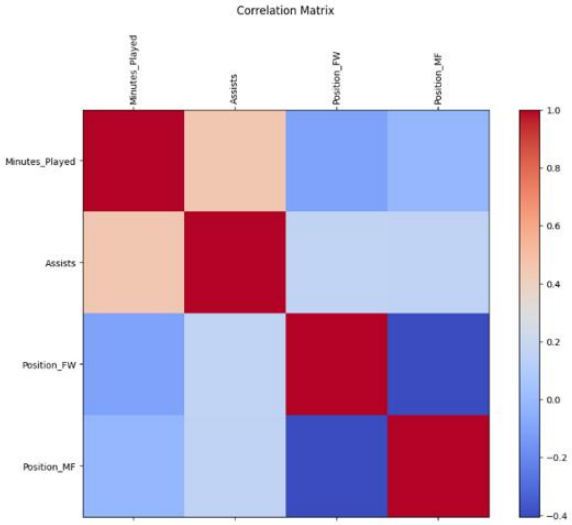
Results and Insights

1. Linear Regression:

- R-squared: 0.48, Adjusted R-squared: 0.46.
- Goals were positively influenced by Assists (coefficient: 0.35) and Position (Forward: 5.36).
- Moderate fit, suggesting potential for feature refinement.

```
Model Evaluation Metrics:
Root Mean Squared Error: 2.3766291829134096
Mean Absolute Error: 1.8160954144205055
R-squared: 0.4824244195314227
Adjusted R-squared: 0.4562180610266846
```

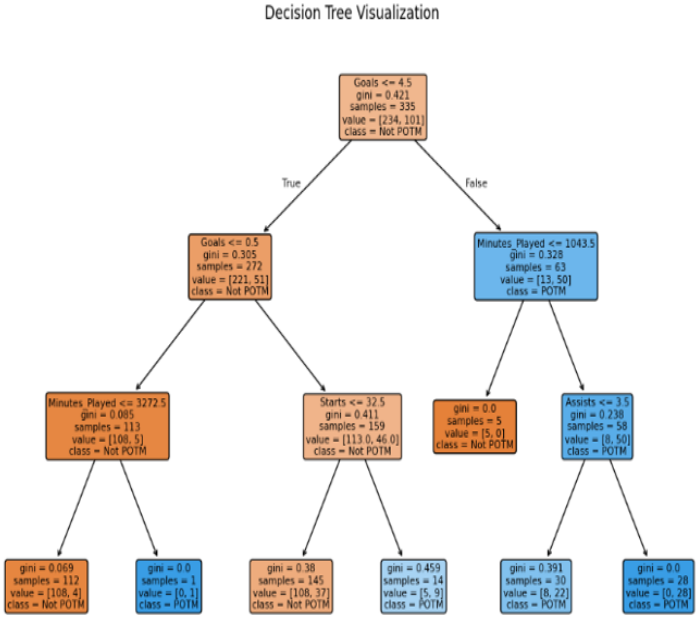
```
Model Coefficients:
                Coefficient
Minutes_Played     0.001623
Assists            0.350063
Position_FW        5.359270
Position_MF        1.786317
```



2. Decision Tree:

- Accuracy: 81%, with Goals as the most significant predictor.
- Key splits were based on Goals and Minutes Played.
- Visualization revealed clear thresholds for POTM classification.

```
Accuracy: 0.8095238095238095
Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.89      0.87        62
         1.0       0.65      0.59      0.62        22

    accuracy                           0.81        84
   macro avg       0.75      0.74      0.75        84
weighted avg       0.80      0.81      0.81        84

Confusion Matrix:
 [[55  7]
 [ 9 13]]

Feature Importances:
              Importance
Goals           0.757381
Minutes_Played  0.144498
Starts          0.063852
Assists         0.034269
```



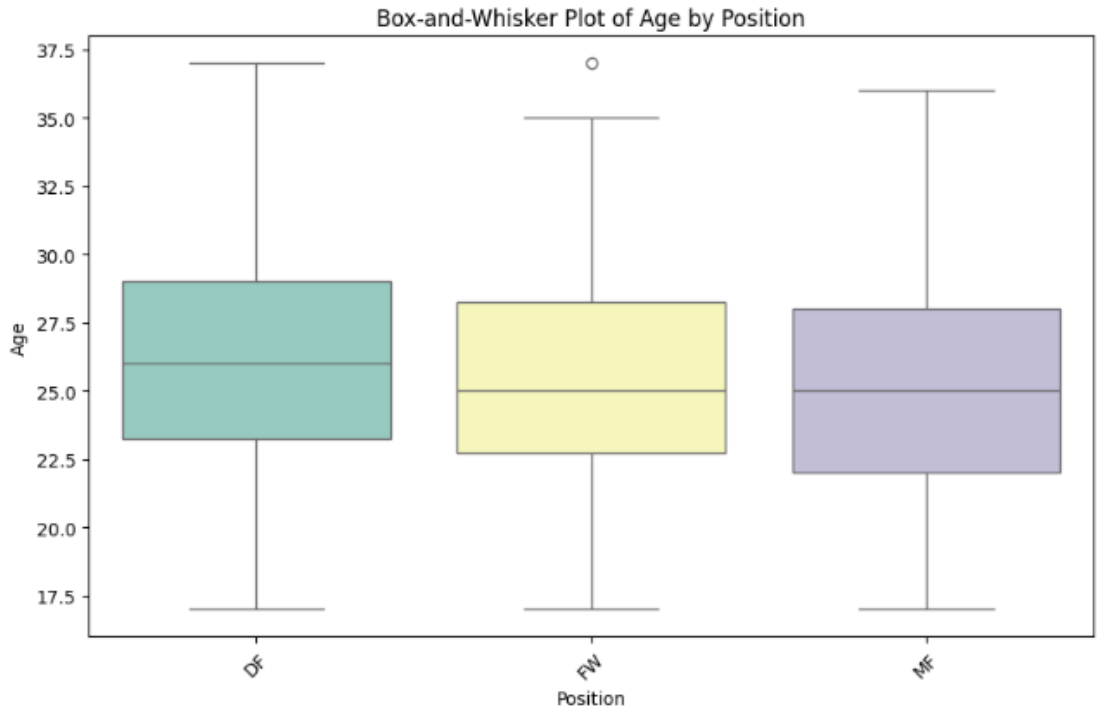
3. Logistic Regression:

- Accuracy: 84.5%.
- Goals had the highest odds ratio (1.50), indicating its dominant impact on POTM.
- Assists and Minutes Played contributed moderately.

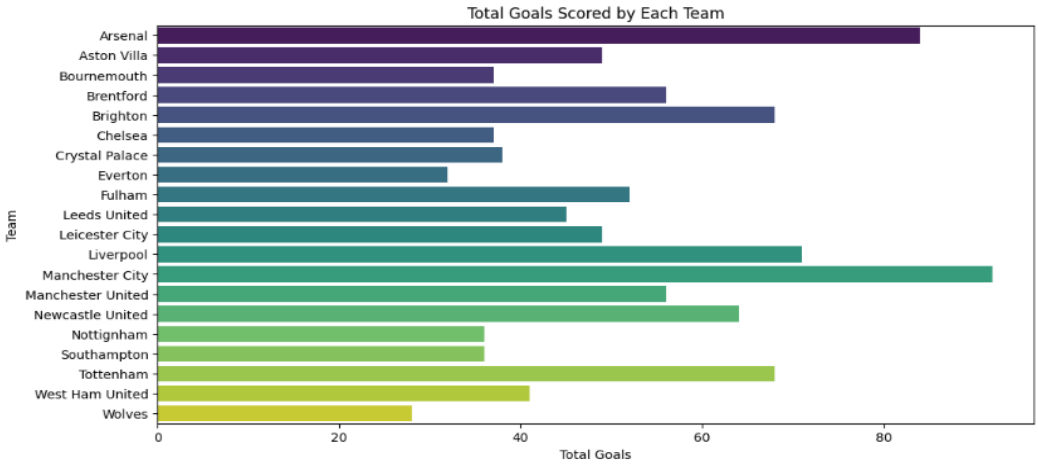
```
Accuracy: 0.8452380952380952

Classification Report:
               precision    recall  f1-score   support

         0.0       0.86      0.95      0.90        62
         1.0       0.80      0.55      0.65        22

    accuracy                           0.85        84
   macro avg       0.83      0.75      0.77        84
weighted avg       0.84      0.85      0.83        84


Confusion Matrix:
 [[59  3]
 [10 12]]

Feature Importance (Odds Ratios):
                Odds Ratio
Goals             1.500652
Assists           1.162794
Starts            0.846275
Minutes_Played    1.002373
```

Key Visualizations

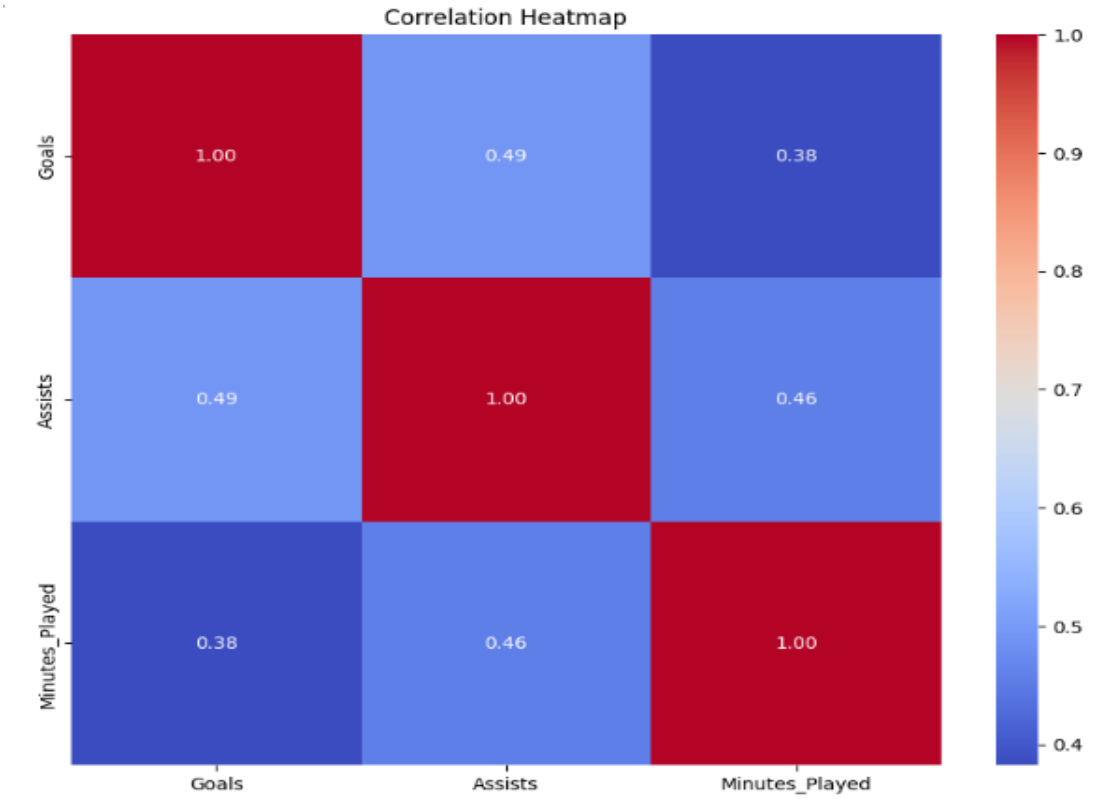
1. Box Plot: Age distribution by Position revealed that Defenders and Midfielders tend to be older than Forwards.



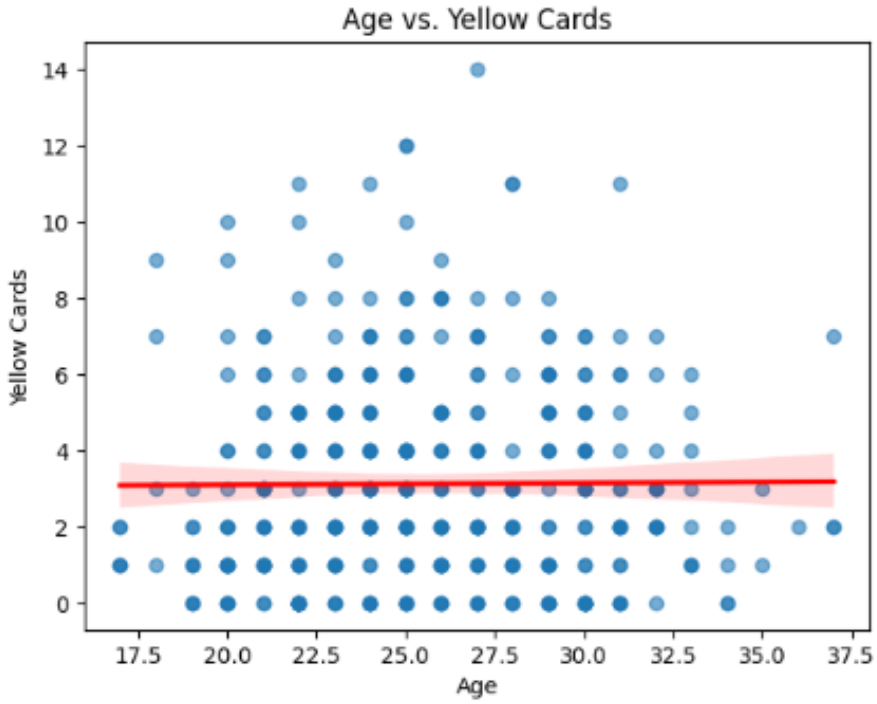
2. Bar Graph: Manchester City and Liverpool dominated in Goals scored, while Wolves and Southampton struggled.



3. Correlation Heatmap: Moderate correlations observed among Goals, Assists, and Minutes Played.



4. Scatter Plot: No strong correlation between Age and Yellow Cards, indicating discipline is independent of age.



Conclusion and Future Work

This project provided valuable insights into player performance in the English Premier League. Goals and Assists emerged as the strongest predictors for POTM. Logistic Regression was the preferred model for its accuracy and interpretability. Future improvements could include addressing class imbalances, exploring ensemble methods (e.g., Random Forests), and incorporating additional features like Shots on Target or Pass Accuracy.