CSP 571- DATA PREPARATION AND ANALYSIS

Fall 2023

Department of Computer Science

Project Report

Topic: Player Performance Analysis

PROJECT GROUP

- ADITYA SHIVAKUMAR - A20513527 (ashivakumar@hawk.iit.edu)
- PUNALRAJ PARTHASARATHY - A20519421 (pparthasarathy@hawk.iit.edu)
- CHENGLONG WU YANG - A20518979 (cwuyang@hawk.iit.edu)

PROJECT GROUP LEADER- ADITYA SHIVAKUMAR

Abstract

Our project analyzes the five major soccer leagues between 2021-2022. The five clubs participating in our analysis are Premier League, Ligue 1, Bundesliga, Serie A, and La Liga. We take a deep dive into the dataset that contains a vast amount of statistics. We aim to preprocess the data for data handling like missing values, and outliers, and perform exploratory data analysis to extract patterns and insights. Another goal of ours is to create a model that forecasts future player's positions based on statistical attributes, assisting football clubs, coaches, and talent scouts in making educated decisions about player recruitment, squad composition, and strategic planning.

Table of Contents

Overview

Problem Statement

The realm of professional football is characterized by the need for continuous performance optimization at both individual and team levels. A critical challenge in this context is effectively leveraging the wealth of statistical data available on players to enhance decision-making processes related to player performance. The primary problem addressed by this project is the development and implementation of an analytical framework capable of analyzing and interpreting extensive football player data. This framework aims to identify key performance indicators, categorize players based on positions and continents, and derive meaningful insights through exploratory data analysis and visualization techniques. Additionally, the project seeks to utilize predictive models to predict player performance metrics and player's position.

Relevant Literature

Player Performance Prediction in Football Game by Richard Pariath

Encompasses approximately 21,280 football players. Preprocessing techniques were applied to handle missing values. Dimensionality reduction was performed through a heatmap analysis to identify key performance-related attributes. Principal Component Analysis (PCA) was also utilized to optimize dimensionality. A supervised learning model was built for player performance prediction with separate models for different player positions. Linear regression was employed for market value prediction. The model for overall player performance exhibited good accuracy with an accuracy rate of 84.34%.

Sports Analytics for Football League Table and Player Performance Prediction.

The research focused on analyzing football data related to teams and players, addressing data quality issues like missing values and duplicates through preprocessing. It consisted of two parts: firstly, using machine learning models, notably Random Forest with over 70% accuracy, to predict the relative performance of football teams in upcoming seasons. Secondly, it involved simulating matches for a specific season to forecast final league standings, integrating various datasets. The models demonstrated effectiveness in predicting match outcomes and league standings, particularly for the English Premier Division, with 57% accuracy in match predictions and a low RMSE of 9 for team points.

Dataset Summary

The dataset, sourced from Kaggle, encompasses detailed 2022-2023 football player statistics per 90 minutes, spanning five major leagues: Premier League, League 1, Bundesliga, Serie A, and La Liga. It comprises 2,500 rows and 124 columns, encapsulating a broad spectrum of player data, including personal information (like name, age, and nationality), in-game metrics (such as goals, shots, passes, and tackles), and advanced analytics (like shot-creating actions, goal-creating actions, and various types of passes and touches). This comprehensive dataset offers an extensive insight into individual player performances and is suitable for in-depth football analysis.

Data Processing - Pipeline details

The data processing pipeline for the football player statistics project includes loading the dataset from a CSV file, checking its dimensions, and identifying and handling missing values. Duplicate rows are removed for data integrity. Position names are standardized for consistency, and players' nationalities are categorized by continent. The pipeline features exploratory data analysis with `ggplot2` visualizations, statistical tests like ANOVA, and correlation analysis. Custom functions calculate performance metrics for different player positions, and linear regression analysis is applied. Finally, advanced analysis is conducted using machine learning models like Random Forest and PCA.

Data Issues

Our data processing involves checking for missing values across the entire dataset. Given the nature of the dataset, which is focused on football player statistics, it was determined that replacing any missing values with zero was the most appropriate strategy. This decision was based on the rationale that in the context of player statistics, the absence of a record translates to a non-occurrence. Using other statistical imputation methods like mean or median substitution could potentially distort the data, introducing inaccuracies. Upon conducting this check, it was observed that the dataset has no missing values.

In our data preprocessing, duplicate rows were identified and removed, ensuring data accuracy. This step was crucial for further analysis.

Assumptions and Adjustments

Position Generalization

We have made adjustments to our dataset due to the amount of different possible positions, where every position is generalized to main four positions:

- DFFW, DFMF → DF
- FWDF, FWMF → FW
- MFDF, MFFW → MF

For GK adjustments is not necessary because there is only one goalkeeper in a football match.

Grouping countries by continent

In the dataset, players' nationalities are represented by a wide range of countries, posing challenges due to the extensive diversity and data scarcity for certain countries. To streamline the analysis, a new feature that groups players' nationalities by their respective continents such as Africa, Asia, Europe, etc, was introduced. This approach not only simplifies the analysis but also addresses the problem of the underrepresentation of nations with fewer players. Additionally, considering the significant concentration of leagues in Europe, this continent-based grouping facilitates a more balanced examination of trends, enabling a better understanding of regional influences in the distribution of players.

Introduction of a new column- age group

Under 25	25-32	32-35	35+
1295	1174	169	51

The majority of players in the dataset are under 25, with 1,295 under 25 and 1,174 between 25 and 32 years old. This shows that professional football is primarily a sport for young people, with fewer players as they get older. Only a few players (169) are between the ages of 32 and 35, and only a few (51) are 35 or older. This distribution most likely reflects the sport's physical demands, which favor younger athletes, and it predicts a steep fall in professional play as players reach their mid-30s.

Data Analysis

Summary statistics

Position Count:

DF → 964
FW → 683
GK → 164
MF → 878

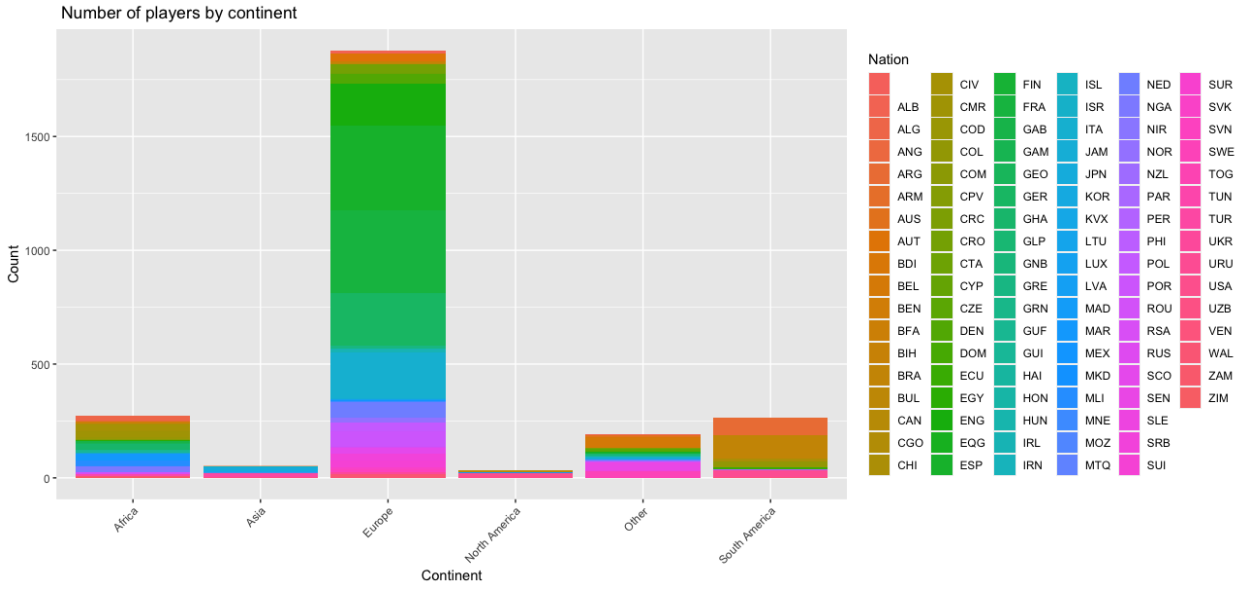
For each position statistics were calculated on some of the most important variables such as goals, assists, and touches.

POS	mean_goals	median_goals	sd_goals	mean_assists	median_assists	sd_assists	mean_touches	median_touches	sd_touches
DF	0.40	0	0.78	0.05	0	0.15	63.09	62	16.01
FW	2.31	1	3.09	0.12	0	0.26	41.22	38.80	16.01
GK	0	0	0	0.002	0	0.01	37.31	36.75	6.8
MF	0.90	0	1.54	0.1	0	0.38	56	54.55	19.18

Visualization

Number of Player distribution by Continent and Nationality

The geographical distribution analysis of football players by continent in the dataset reveals the sport's global footprint, highlighting disparities in representation. This analysis is key in identifying which continents are under or overrepresented, thereby aiding teams in making strategic decisions. Colored bars for each continent display the diversity of player origins, helping to uncover patterns.

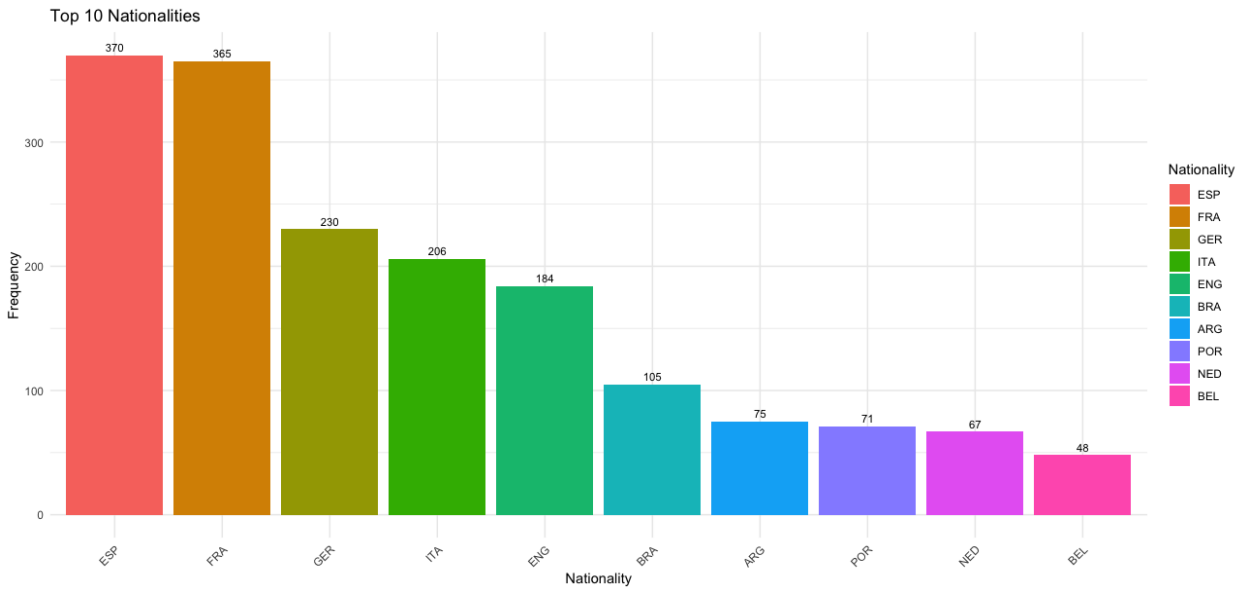


Inference

The graph suggests a dominance of European players, likely due to the dataset's focus on Europe-based leagues and a preference for local European talent over international players. This trend may be influenced by factors such as familiarity, logistical ease, and a well-established football infrastructure in Europe. Consequently, a European player with strong statistics is more likely to be recruited by a team, reflecting a potential bias towards local talent.

Ranking of Top 10 Nationality among all the players

It is intended to highlight the top ten nationalities among football players in the dataset, demonstrating the most prevalent countries of origin. By exhibiting the frequency of each country, it's useful for determining which countries have the most representation, which can indicate the strength and popularity of football or the success of their development programs.

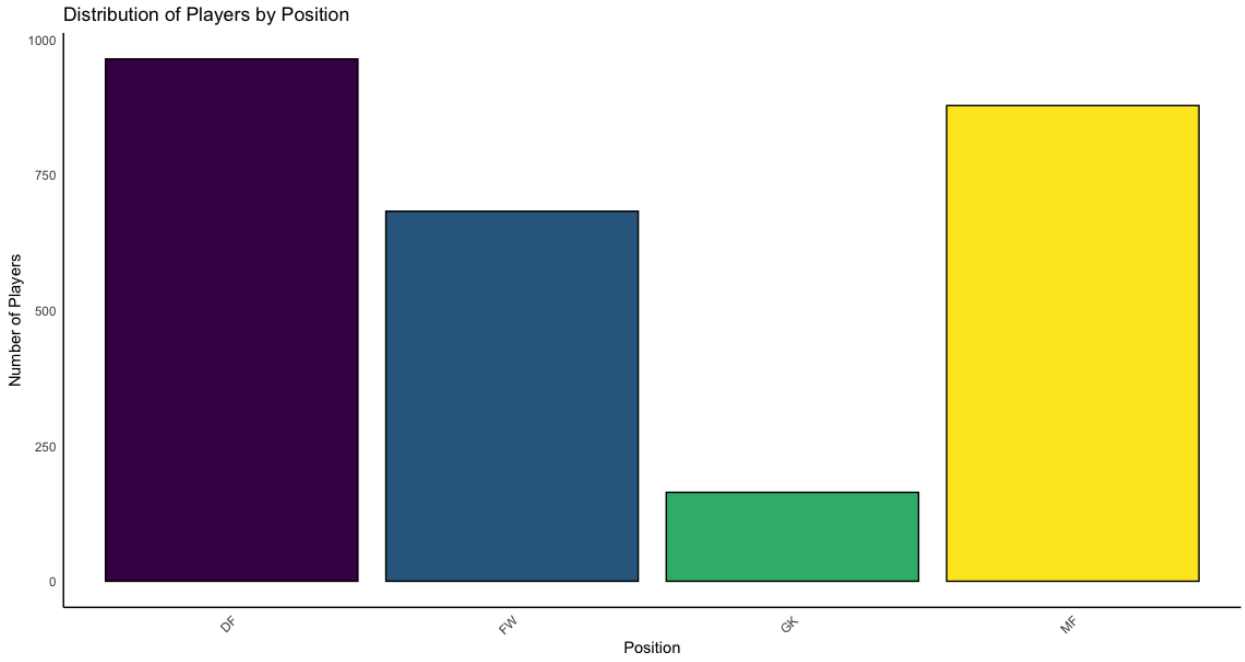


Inference

The data distribution represents a strong presence of European nationalities, particularly from Spain and France. The dominance of these regions, especially countries with their own prominent leagues, suggests a tendency to prefer players not only from across Europe but also specifically from these leading football nations. This trend might be influenced by factors such as national league popularity, player familiarity, and regional football traditions.

Distribution of Players by Position

Looking at the distribution of football players across various positions will be useful to identify which positions are most prevalent and which areas are potentially saturated or in demand. Such visualization can help clubs and sports analysts discover patterns in player roles and guide judgments about training focus, recruitment needs, and strategic team composition planning.

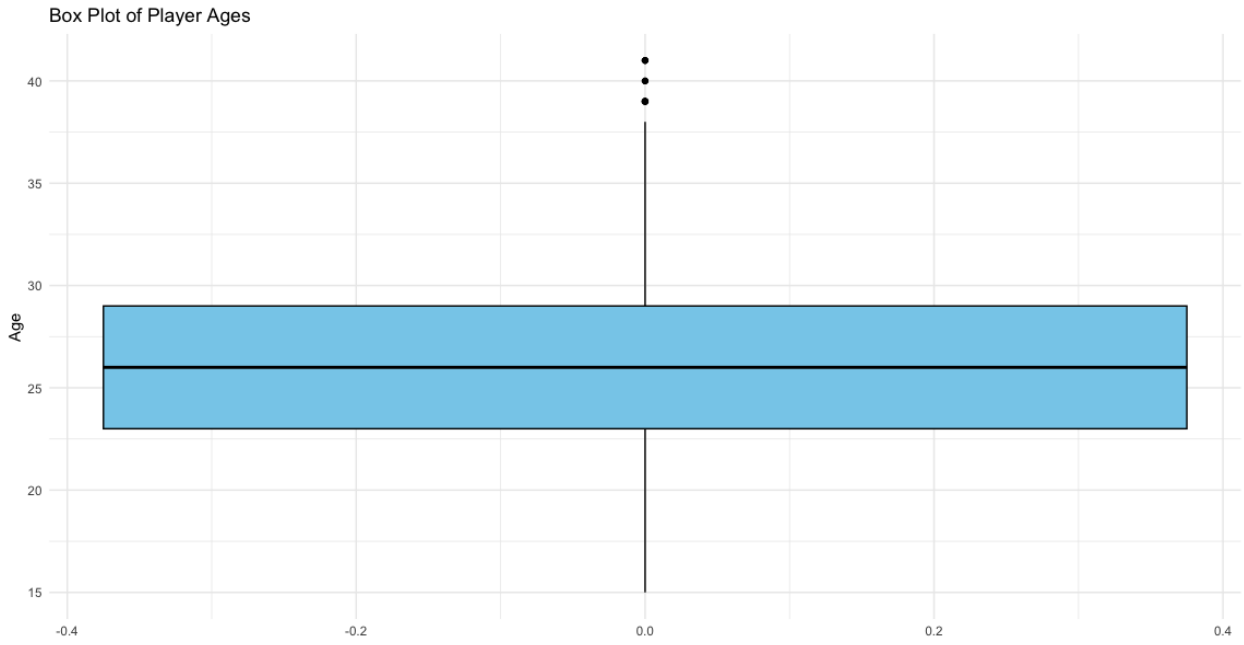


Inferences

The dataset reveals that Defenders (DF) are the most prevalent group, suggesting a preference or necessity for versatile players in football teams. Following closely are Midfielders (MF) and Forwards (FW), with Goalkeepers (GK) being the least represented. This reflects typical player composition in a football team, where a larger number of defenders and midfielders is often required. Midfielders and defenders are pivotal in controlling the game's flow and covering extensive ground on the field, making them more susceptible to fatigue. Consequently, teams may maintain a higher number of these players to allow for effective rotation and substitution strategies during matches. This approach ensures that the team can sustain high-performance levels throughout the game, highlighting the importance of these roles in football.

Distribution of Players by Age

Illustrating the median, quartiles, and outliers in the age distribution of football players. It paints a clear image of the sport's age dynamics, which can be critical for understanding the career span and the recruitment potential of younger stars or the experience level of older players.

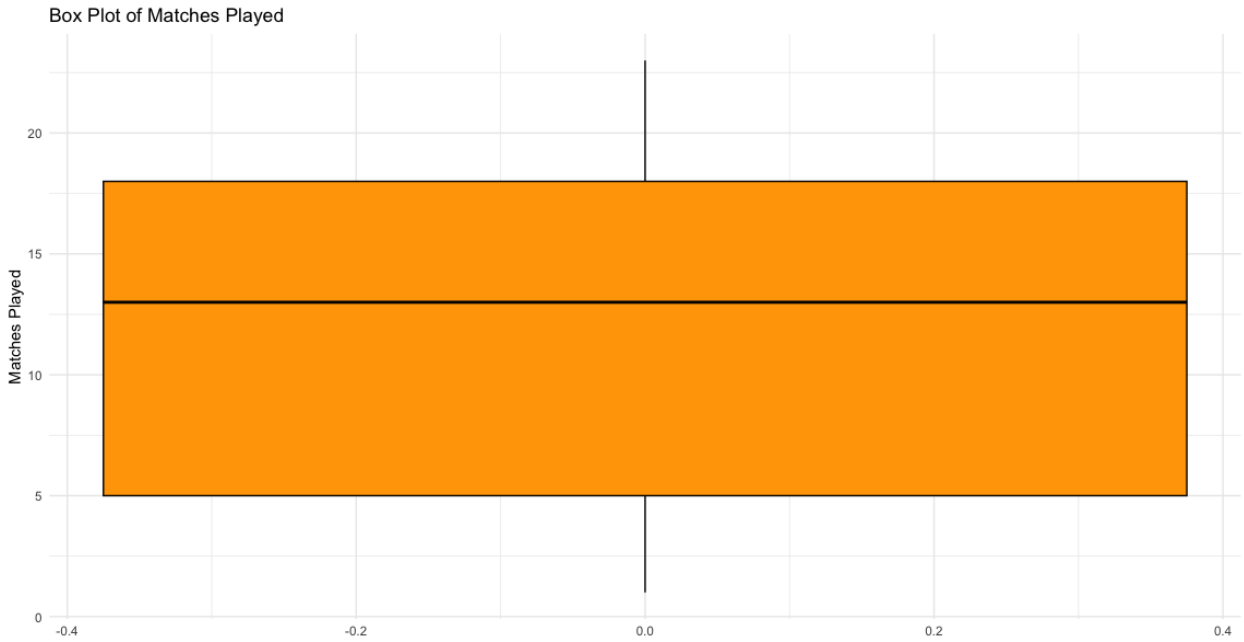


Inferences

The box plot depicts a concentration of participants in their mid-to-late twenties, with outliers reaching their forties. This shows that there is a prime age range for players, with few continuing to play professionally after a particular age. The outliers in their forties cannot be disregarded as the seasoned players continue to compete professionally. The box plot's age distribution shows a relatively young workforce in professional football.

Distribution of Matches Played

A box plot to check for the distribution and outliers in the matches played column.

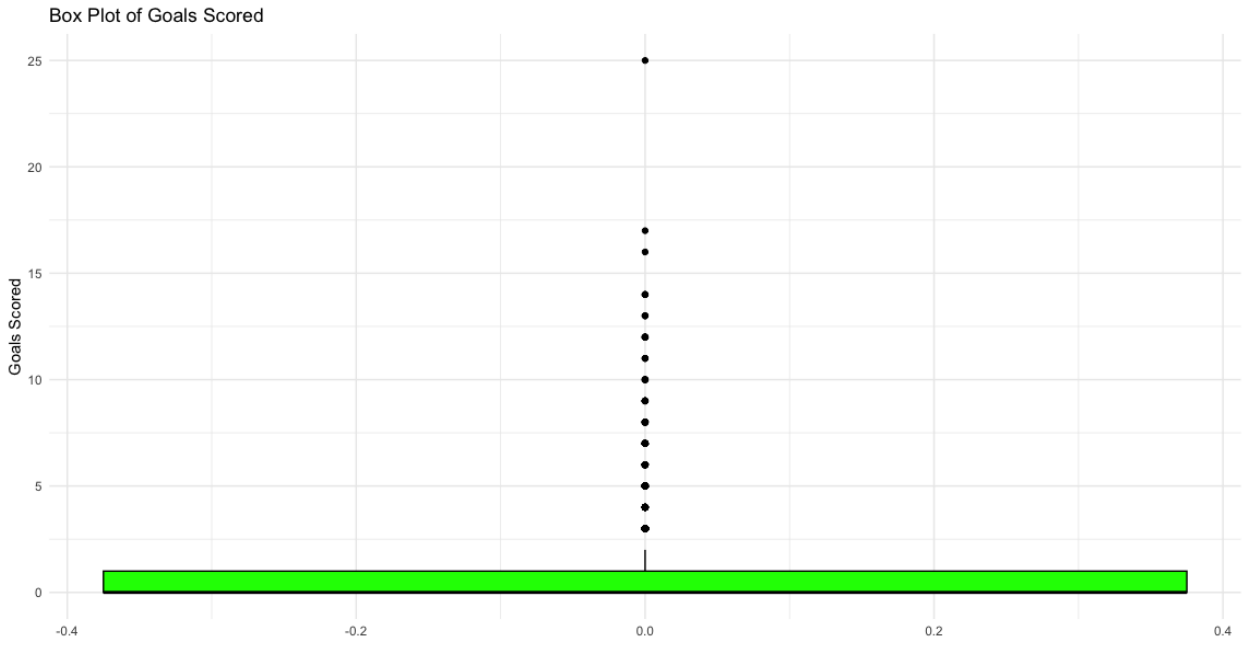


Inferences

The box plot indicates that most players participate in a moderate number of matches, with only a few engaging in exceptionally high or low counts, reflecting differences in career spans, injury rates, or lineup changes. This pattern also offers a perspective on the usual range of matches played per season, highlighting the variability in player participation.

Distribution of Goals Scored

Examining goal-scoring data across players by displaying the median, range, and outliers. It's critical for determining scoring frequency and trends. The plot assists clubs and analysts in assessing scoring potential across players and aiding them in strategies.

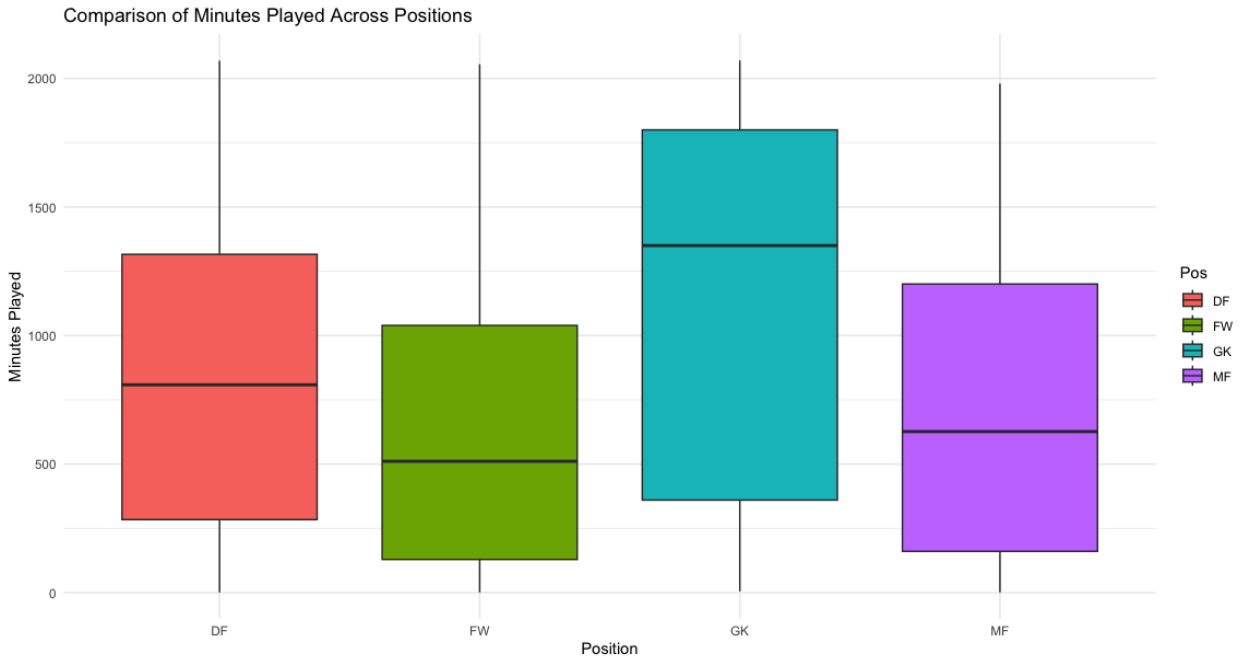


Inferences

The concentration toward the lower end of the plot shows that most players score only a few goals, while a significant number of outliers highlight exceptional goal scorers. This disparity likely mirrors the distinct roles players hold within teams and the rarity of prolific goal-scoring talents. These outliers, representing high-scoring individuals, cannot be removed as they are crucial for understanding the exceptional capabilities and contributions of top performers in the sport, providing valuable insights into the characteristics that differentiate average players from elite goal scorers. It can also indicate that in some teams are only focused on one player and playing their tactics around them.

Visualization of Minutes Played per Position

Comparing average playing time across positions, which might show positional demands. It's critical to determine how much game time certain roles get. This analysis aids in the management of player workloads and the dynamics of playing time distribution within a team.

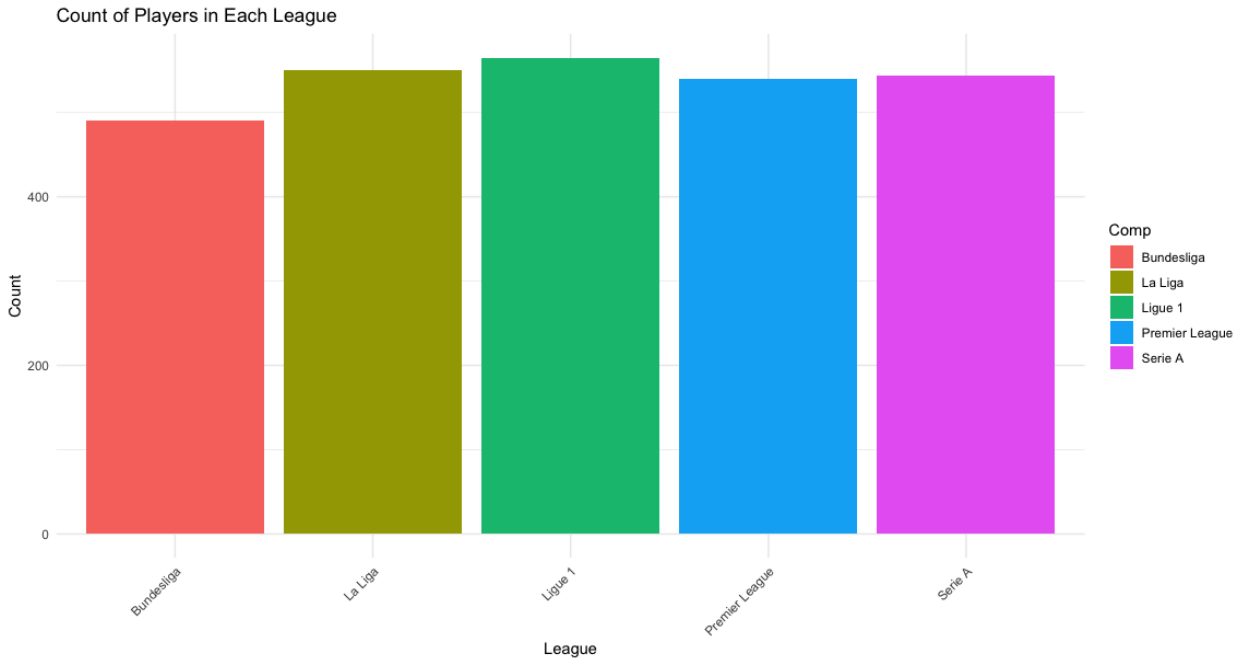


Inferences

Goalkeepers had higher median minutes played, as expected given their specialized function and less frequent rotation, according to the box plot. Other positions have varying percentages, owing to tactical replacements and the physicality of outfield roles, which may necessitate more frequent player changes.

Distribution of Players per League

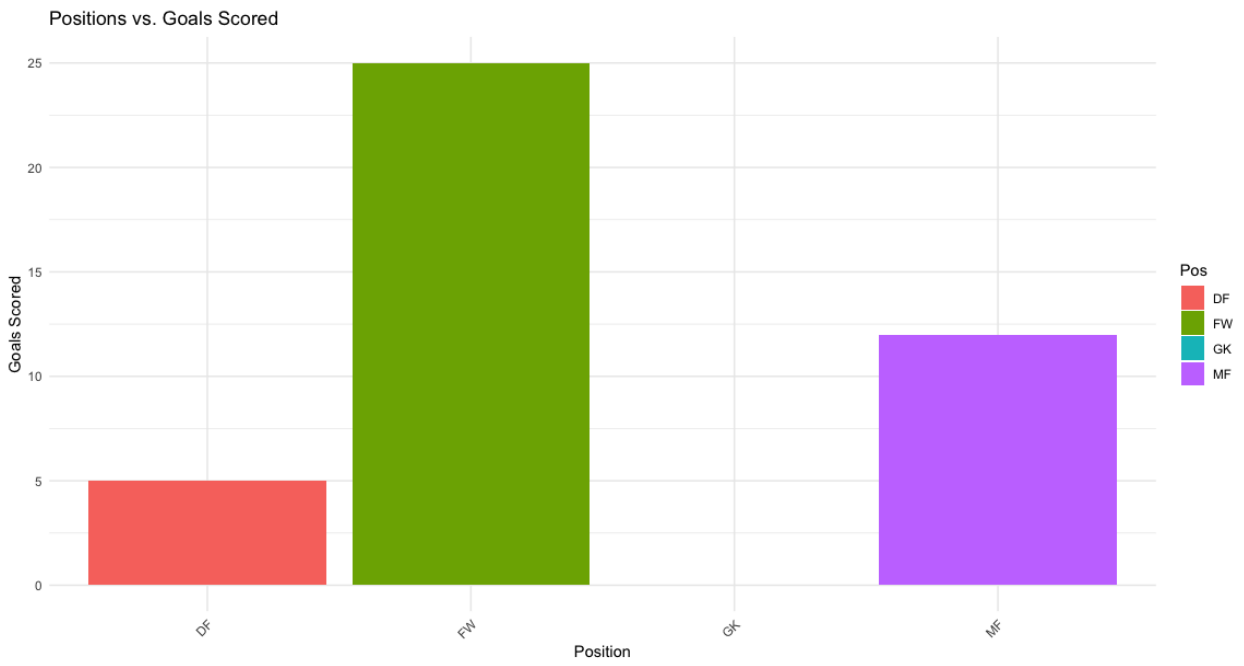
Comparing the number of players in the top European football leagues. It depicts player distribution visually, illustrating the size and magnitude of each league. Such data can be used to analyze league popularity.



Inferences

The distribution reveals a generally equal count of players across leagues, implying that major European leagues operate on a comparable scale. This balance may indicate the leagues' competitiveness and market equity.

Visualizations of Goals Scored per Position

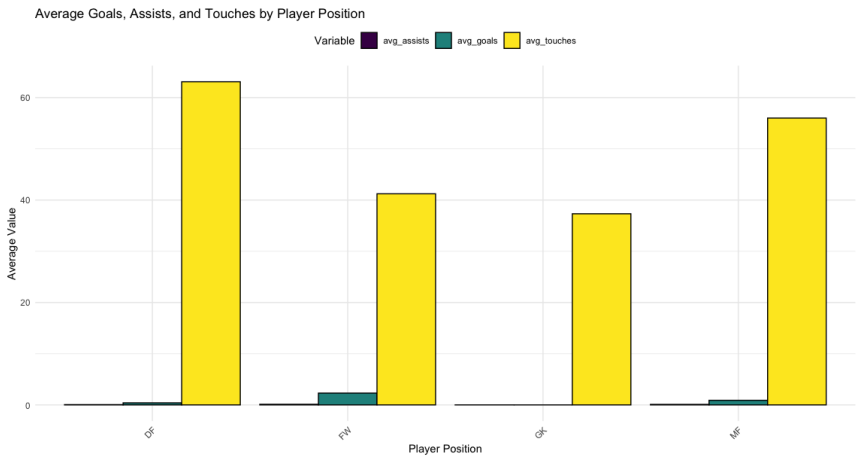


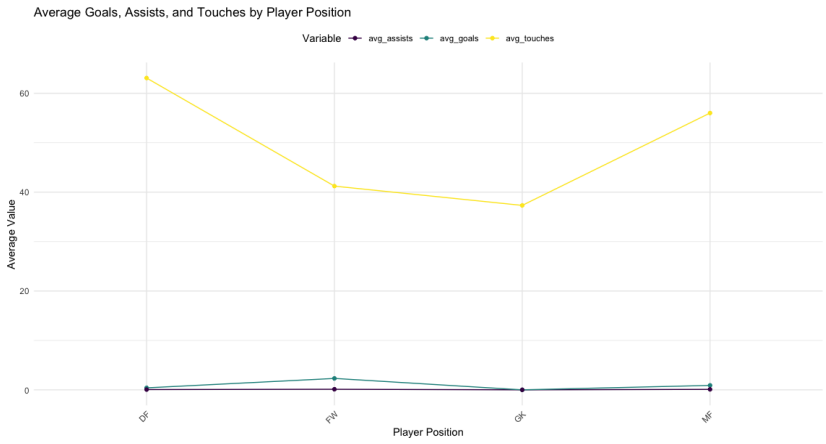
Inferences

Forwards (FW) have the highest average number of goals scored, which aligns with their primary role in scoring. Midfielders (MF) also contribute a fair amount of goals, likely due to their involvement in offensive plays and passing. Defenders (DF) have the fewest goals, as their main role is to prevent goals rather than score them. Goalkeepers (GK) are typically not involved in scoring and thus have the least amount of goals, if any.

This chart aligns with typical football strategies where forwards are more engaged in scoring and midfielders in helping forwards and vice versa, while defenders and goalkeepers focus on preventing them.

Different types of visualization of Average of Touches, Goals, and Assists





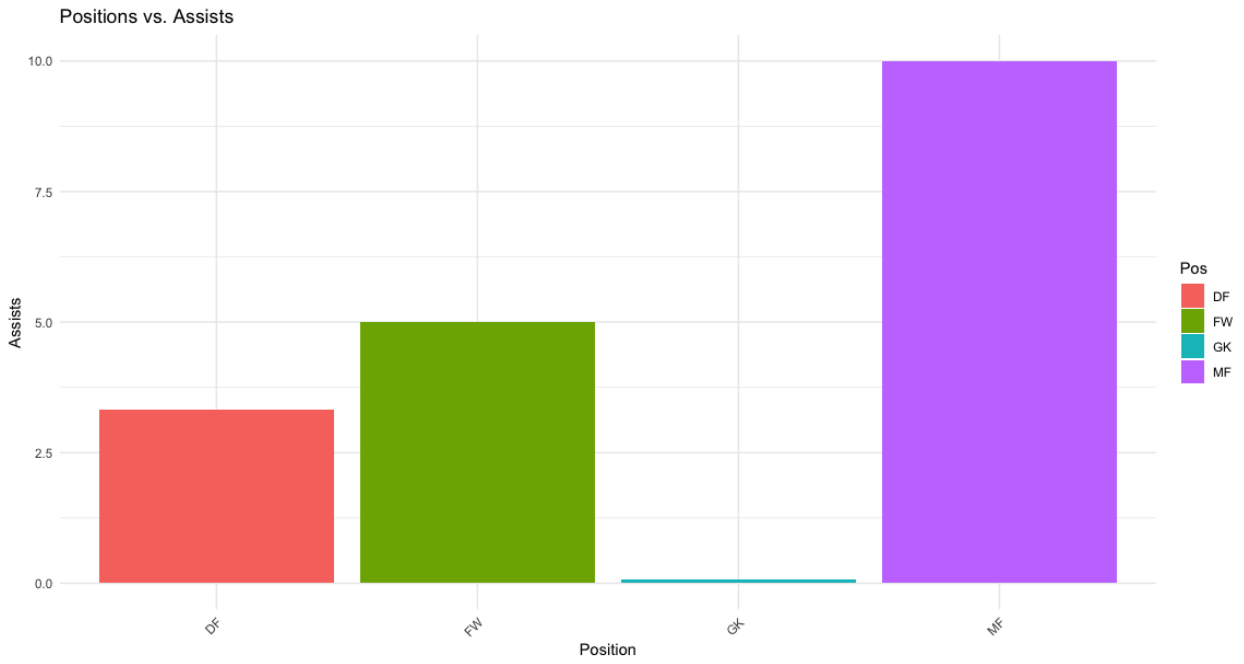
Inference

Midfielders (MF) have the highest average values for assists and touches, which is consistent with their role in game control and opportunity creation. Forwards (FW) have a high average amount of goals, which corresponds to their primary goal of scoring.

Defenders (DF) and goalkeepers (GK) have low goal and assist statistics, which is to be anticipated considering their defensive positions. When compared to midfielders, their average number of touches is similarly smaller, indicating less frequent involvement in ball possession.

This chart provides insights into the typical involvement of each position in offensive actions and ball possession during games. Some teams are heavily focusing on possession and we can deduct their play style.

Visualization of Assists per Position



Inferences

Midfielders (MF) have the highest average assists, which is expected as they often play a key role in setting up goals. Forwards (FW) also have a significant number of assists, likely due to their involvement in attacking plays. Defenders (DF) have fewer assists, which aligns with their primary role in defense rather than attack. Goalkeepers (GK) have the least assists, which is consistent with their specialized position and rare involvement in goal-scoring opportunities.

Custom Performance Metrics

The performance metrics are calculated by considering the player's position and assigning weighted importance to features specific to that role, be it forward, defender, or midfielder. The code provided illustrates the differential weighting assigned to various features, tailored to each position. Additionally, the features incorporated into the player performance metrics vary across positions, as they are selected through manual analysis to include key attributes that are most relevant and impactful for each specific role on the field. This tailored approach ensures that the performance metrics accurately reflect the unique contributions and skill sets required by different positions in a football match.

```
calculate_midfielder_metric <- function(x) {
  # Assigning weights
  w_tackles = 0.15
  w_interceptions = 0.15
  w_passes_total_cmp = 0.2
  w_passes_assisted = 0.2
  w_ppa = 0.15
  w_sca = 0.15
```

```
calculate_defender_metric <- function(x) {
  # Assigning weights
  w_clearances = 0.2
  w_aerial_duels_won = 0.2
  w_blocks = 0.2
  w_tackles_def_3rd = 0.2
  w_tackles = 0.1
  w_interceptions = 0.1
```

Performance for each player is unique based on the player's respective position.

Evaluation metrics:

We rated players based on their features and respective positions and assigned scores in the performance metrics column.

We have used the calculated performance metrics and analyzed several features to find relationships between player performance and other significant features that may affect or improve a player's performance.

```
Player: Brenden Aaronson | Position: MF | Performance Metric: 7.191968
Player: Yunis Abdelhamid | Position: DF | Performance Metric: 11.584
Player: Himad Abdelli | Position: MF | Performance Metric: 11.13176
Player: Salis Abdul Samed | Position: MF | Performance Metric: 9.55133
Player: Laurent Abergel | Position: MF | Performance Metric: 9.348032
Player: Oliver Abildgaard | Position: MF | Performance Metric: 9.202128
Player: Matthis Abline | Position: FW | Performance Metric: 7.08
Player: Matthis Abline | Position: FW | Performance Metric: 11.48667
Player: Abner | Position: DF | Performance Metric: 10.465
Player: Zakaria Aboukhlal | Position: FW | Performance Metric: 11.37333
Player: Tammy Abraham | Position: FW | Performance Metric: 13.35333
Player: Francesco Acerbi | Position: DF | Performance Metric: 8.878
```

Correlation Analysis

ANOVA (Analysis of Variance) + Tukey HSD(Honestly Significant Difference)

The **ANOVA** is testing whether there are statistically significant differences in the number of assists among different player positions.

```
              Df Sum Sq Mean Sq F value   Pr(>F)
Pos            3   3.13  1.0420   13.82 6.11e-09 ***
Residuals   2685 202.50  0.0754
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The triple asterisks (***) next to the p-value signify that the results are highly statistically significant, indicating strong evidence against the null hypothesis. Thus, there are significant differences in the number of assists among different positions.

Tukey HSD is used to determine exactly which positions differ from each other after finding a significant F-statistic from the ANOVA.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Goals ~ Pos, data = player_data)

$Pos
            diff        lwr          upr     p adj
FW-DF  1.9112607  1.6731903  2.149331208 0.0000000
GK-DF -0.4035270 -0.8055970 -0.001456923 0.0487788
```

Inference

The Tukey HSD test results show that there are substantial disparities in goals scored between player positions. Forwards score 1.91 goals more than defenders on average, reflecting their offensive role. Midfielders also outscore defenders, owing to their simultaneous attacking and defensive responsibilities. Goalkeepers have fewer goals than all other field players, emphasizing their defensive expertise. These studies highlight the varied functions and contributions of each football position.

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Assists ~ Pos, data = player_data)

$Pos
             diff         lwr          upr     p adj
FW-DF  0.06712607  0.03181767  0.102434473 0.0000065
GK-DF -0.05260905 -0.11224034  0.007022248 0.1058525
MF-DF  0.04520319  0.01226959  0.078136790 0.0024000
GK-FW -0.11973512 -0.18112398 -0.058346258 0.0000034
MF-FW -0.02192288 -0.05794124  0.014095486 0.3991527
MF-GK  0.09781224  0.03775784  0.157866637 0.0001711
```

We can infer that FW-DF, MF-DF, and MF-GK indicate forwards and midfielders have significantly more assists than defenders and goalkeepers, respectively. GK-DF, GK-FW, and MF-FW indicate goalkeepers have fewer assists compared to defenders and forwards, and midfielders have fewer assists compared to forwards. All p-values are very low (essentially zero), indicating these differences are statistically significant after adjusting for multiple comparisons.

Correlation between differences between features

	Goals	Assists	MP	Touches
Goals	1.0000000	0.10500561	0.40654530	-0.14311366
Assists	0.1050056	1.00000000	0.04122202	0.01054092
MP	0.4065453	0.04122202	1.00000000	0.02650752
Touches	-0.1431137	0.01054092	0.02650752	1.00000000

Goals and Assists: A positive but weak correlation (r 0.11) exists, showing that players who score more goals have somewhat more assists, although the association is not strong.

Goals and Matches Played (MP): A somewhat favorable correlation (r 0.41) implies that players who play more matches tend to score more goals, which may reflect the enhanced scoring possibilities that come with more playing time.

Goals and touches: have a weak negative association (r -0.14), showing that more touches on the ball do not always result in more goals. This could imply that simply having the ball more frequently does not boost the chances of scoring.

Assists and Matches Played (MP): There is a moderate positive association (r 0.41), indicating that players who have more assists often play more matches, most likely because those who are good at creating scoring opportunities are kept on the field longer.

Assists and Touches: The correlation is nearly zero (r 0.01), indicating that there is no relevant association between the number of touches and assists made by players.

Touches and Matches Played (MP): There is a very modest positive correlation (r 0.03), indicating that playing more matches does not considerably boost a player's number of touches.

```
        Pearson's Chi-squared test

data:  table(player_data$Nation, player_data$Pos)
X-squared = 373.47, df = 312, p-value = 0.009592
```

The test results indicate that player positions in the dataset are not distributed irrespective of nationality. various positions may be more or less prevalent among various ethnicities, indicating probable cultural or training differences in the sport among countries.

Pearson Correlation Coefficient

Correlation Analysis	Pearson Correlation Coefficient
FW Age ←→ Performance Metric	0.121365038276124
MF Player MP ←→ Performance Metric	0.104349751521524
DF Player MP ←→ Performance Metric	0.527263431467159

Inference

There is a positive but weak correlation between Forward (FW) players' age and their performance metric (r 0.12), implying that as forwards age, their performance may improve slightly, however the association is not substantial.

A comparable minor positive connection (r 0.10) is shown between the number of matches played by Midfielder (MF) players and their performance metric, indicating that greater playing time for midfielders may be connected with a slight gain in performance.

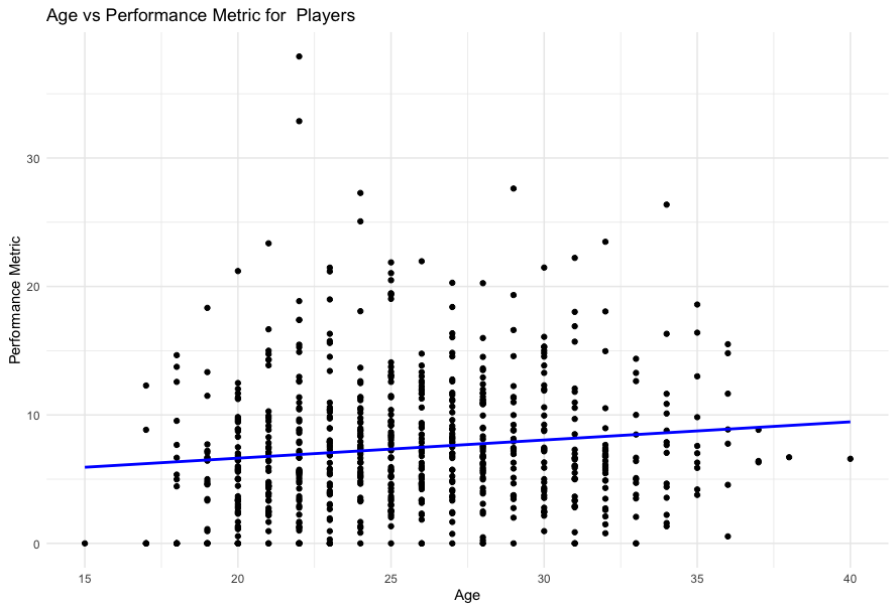
There is a moderate positive correlation (r 0.53) between Defender (DF) players' matches played and their performance metric, indicating a more substantial link in which defenders who play more matches tend to have better performance metrics.

Feature engineering

We chose the most significant features for each position and categorized players based on their stats in those features.

Regression analysis between age and performance metrics:

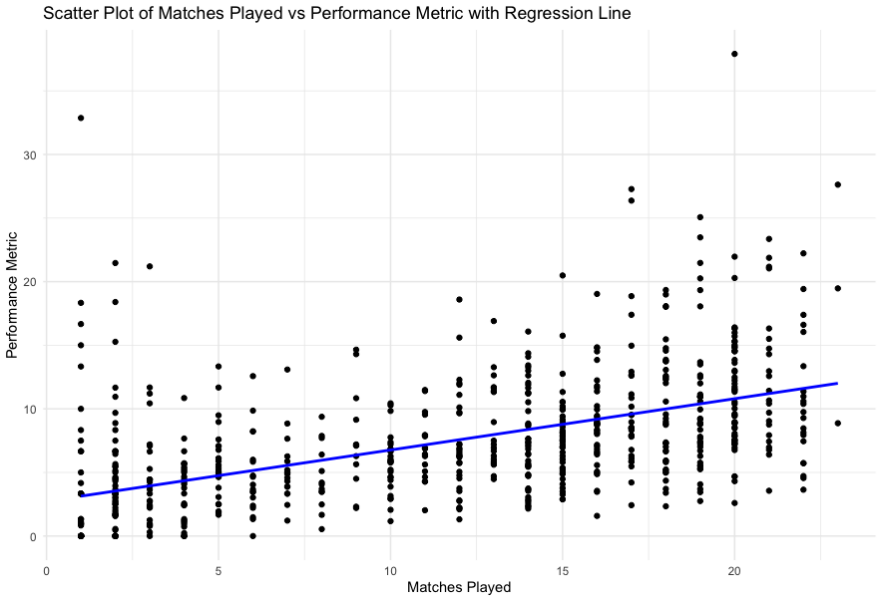
```
Residual standard error: 5.229 on 681 degrees of freedom
Multiple R-squared:  0.01473,   Adjusted R-squared:  0.01328
F-statistic: 10.18 on 1 and 681 DF,  p-value: 0.001484
```



The scatter plot suggests that there is a statistically significant, albeit weak, relationship between age and the performance metric for players. Given the low R-squared values, age alone does not strongly predict the performance metric. The presence of many data points spread out from the regression line also suggests high variability that is not captured by age alone. The significant p-value indicates that age is a predictor of performance, but the effect size is small, meaning other factors likely also play a significant role in determining the performance metric.

Regression analysis between performance metrics and Matches Played

```
Residual standard error: 4.476 on 681 degrees of freedom
Multiple R-squared:  0.278,     Adjusted R-squared:  0.2769
F-statistic: 262.2 on 1 and 681 DF,  p-value: < 2.2e-16
```



The positive slope of the regression line in the scatter plot shows that as players participate in more matches, their performance metric tends to increase. However, since the R-squared value is not very high (less than 0.3), there are other factors not included in the model that also affect the performance metric. Despite this, the strong F-statistic and low p-value confirm that matches played are a significant predictor of the performance metric in the dataset.

Model selection

In this project, two predictive models were selected to forecast player positions:

1. Random Forest
2. SVM

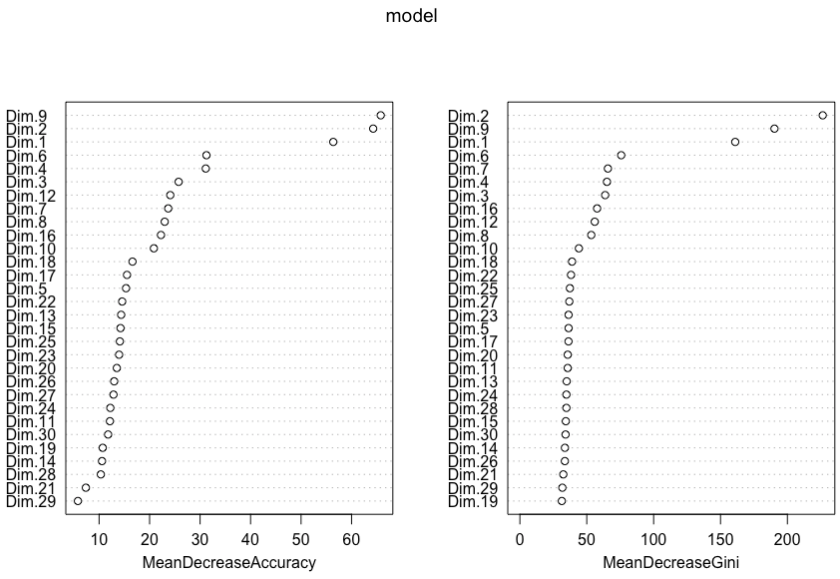
Random forest model Building and evaluation:

```
             Df Sum Sq Mean Sq F value Pr(>F)
Pos           3   1694   564.6   164.7 <2e-16 ***
Residuals  2685   9206     3.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Random forest model test results:

```
Confusion Matrix and Statistics

          Reference
Prediction  DF DFFW DFMF  FW FWDF FWMF  GK  MF MFDF MFFW
      DF   162    7   17   1    0    3   0   7    4    3
      DFFW   0    0    0   0    0    0   0   0    0    0
      DFMF   0    0    0   0    0    0   0   0    0    0
      FW     0    0    2  59    2   26   0   3    1    7
      FWDF   0    0    0   0    0    0   0   0    0    0
      FWMF   0    0    0   6    1    4   0   1    1    3
      GK     0    0    0   0    0    0  32   0    0    0
      MF     3    0    1  12    1    9   0 110    6   22
      MFDF   0    0    0   0    0    0   0   0    0    0
      MFFW   0    0    0   3    2    6   0   0    0    6

Overall Statistics

               Accuracy : 0.6998
                 95% CI : (0.6589, 0.7385)
    No Information Rate : 0.3096
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6124

 Mcnemar's Test P-Value : NA
```

The confusion matrix indicates the performance of the model has an overall accuracy of 69.98%, which falls within a 95% confidence interval of approximately 65.89% to 73.85%. This accuracy is significantly better than random chance, as suggested by the p-value of less than 2.2e-16. The Kappa statistic of 0.6124 further indicates a substantial agreement between the predicted and actual classifications, correcting for chance. However, the model seems to struggle with certain classes which represent misclassifications.



The variable importance plots from the Random Forest model reveal that certain features, notably `Dim.2`, `Dim.9`, and `Dim.1`, are critical to the model's performance, as indicated by their high Mean Decrease in Accuracy and Gini values. These plots suggest that while the model considers a range of features, a select few have a more significant impact on the model's ability to make accurate predictions.

SVM model building and evaluation:

An SVM (Support Vector Machine) model has been implemented to predict player positions.

```
          Reference
Prediction  DF DFFW DFMF  FW FWDF FWMF  GK  MF MFDF MFFW
      DF   159    5   15   1    0    0   0   3    3    0
      DFFW   0    0    0   0    0    0   0   0    0    0
      DFMF   0    0    0   0    0    0   0   0    0    0
      FW     1    0    0  57    2   23   0   2    1    6
      FWDF   0    0    0   0    0    0   0   0    0    0
      FWMF   0    0    1   6    3    7   0   0    1    3
      GK     0    0    0   0    0    0  32   0    0    0
      MF     5    1    2  14    0    9   0 112    7   23
      MFDF   0    0    0   0    0    0   0   0    0    0
      MFFW   0    1    2   3    1    9   0   4    0    9

Overall Statistics

               Accuracy : 0.7054
                 95% CI : (0.6647, 0.7438)
    No Information Rate : 0.3096
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6227

 Mcnemar's Test P-Value : NA
```

The classification model exhibits an overall accuracy of 70.54%, with a substantial Kappa statistic of 0.6227, indicating significant agreement beyond chance. The model demonstrates strong predictive performance for classes like "DF" and "GK," achieving high sensitivity and specificity. However, challenges are observed in correctly identifying instances of certain classes, such as "DFFW" and "FWMF," where sensitivity is notably low. While the balanced accuracy is generally high, further investigation and potential adjustments may be required to enhance performance, particularly for classes with lower sensitivity.

CONCLUSION

This project's extensive data analysis on football player statistics meticulously prepared the dataset for analysis, uncovering key positional features and the impact of age and matches played on player performance. The exploration revealed distinct patterns in player distribution and led to the development of a unique 'performance metric', quantifying each player's performance based on their statistical data. Predictive models, including Random Forest and SVM, were employed, achieving commendable accuracy. However, the imbalance within the dataset indicates the need for alternative methods to achieve a more balanced analysis.

The study uncovered a positive correlation between age and matches played, suggesting that increased experience enhances performance. It also highlighted a preference for European players and those in their prime age, shedding light on recruitment trends in professional football. While the findings offer significant insights, they also reveal limitations due to the dataset's extensive scope and potential feature exclusions and inclusion, calling for further research to refine the models and fully leverage the available data for strategic decision-making in football management.

Data Sources

1. Kaggle Dataset: https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats
2. GitHub repository: https://github.com/Aditya-Shivakumar0301/Player-Performance-Analysis--CSP5711

References

[1] Kaggle Dataset: https://www.kaggle.com/datasets/vivovinco/20212022-football-player-stats
[2] R. Pariath, S. Shah, A. Surve and J. Mittal, "Player Performance Prediction in Football Game," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 1148-1153, doi: 10.1109/ICECA.2018.8474750.
[3] Sports Analytics for Football League Table and Player Performance Prediction: https://www.researchgate.net/publication/344438913_Sports_Analytics_for_Football_League_Table_and_Player_Performance_Prediction
[4] Analyzing player performance with animated charts: https://www.r-bloggers.com/2021/12/analyzing-player-performance-with-animated-charts/
[5] NBA Player Performance Analysis: PCA, Hierarchical Clustering, and K-Means Clustering: https://rpubs.com/HassanOUKHOUYA/NBA_Player_Performance_Analysis_PCA_Hierarchical_Clustering_and_K-Means_Clustering