# Data Visualization Project 1
# Final Report

Prepared by

Tran Le Hai

Thai Ba Hung

Dang Duc Dat


COLLEGE OF ENGINEERING AND COMPUTER SCIENCE

**VINUNIVERSITY**

# Contents


**1 INTRODUCTION** **3**

**2 QUESTION 1** **3**

2.1 Introduction . . . 3

2.2 Approach . . . 3

2.3 Analysis . . . 3

2.4 Discussion . . . 6

**3 QUESTION 2** **6**

3.1 Introduction . . . 6

3.2 Approach . . . 7

3.3 Analysis . . . 7

3.3.1 Forwards . . . 7

3.3.2 Defenders . . . 7

3.3.3 Midfielders . . . 8

3.4 Discussion . . . 9

3.4.1 Forwards . . . 9

3.4.2 Defenders . . . 10

3.4.3 Midfielders . . . 11

**4 CONCLUSION** **11**

# 1 INTRODUCTION

The English Premier League (EPL) is widely regarded as one of the most competitive and globally followed football leagues. While fans often engage in subjective debates about which teams are most entertaining or which players fail to meet expectations, data science provides a structured approach to quantify such assessments. This project aims to explore two key questions using real-world data from the 2021–2022 EPL season: **(1)** Which teams exhibited the least attractive playstyles? and **(2)** Which players underperformed relative to their positional responsibilities?

The primary data sources for this analysis are the Premier League Match Data 2021–2022 and the EPL 21–22 Matches Players dataset, both publicly available on Kaggle and curated by Evan Gower. The match-level dataset includes 380 fixtures with 22 attributes per game, capturing variables such as goals, fouls, shots, and possession. The supplementary player-level dataset provides more granular insights, including individual statistics on substitutions, goals, and disciplinary actions. This dual-dataset approach enables a comprehensive analysis of both team and player performances throughout the season. However, the datasets are not without limitations—certain advanced metrics such as expected goals (xG), passing maps, and pressing actions are absent, which constrains the depth of tactical analysis. Nonetheless, the breadth of available attributes supports meaningful statistical exploration and visualization.

# 2 QUESTION 1

## 2.1 Introduction

As a football lover, an audience like me would prefer football matches which consist of numerous goals and actions, thrilling back-and-forth matches or fast-paced matches with unpredictable results. Moreover, as we become busier with our work, we do not want to spend our precious time watching an "unattractive" match. So what does it mean for an "unattractive" playstyle team? To us these teams will tend to be more practical, play more defensive with lots of ways to interrupt the match like delaying tactics or creating many set-piece situations. It is an optimal playstyle with few shots, less possession, more fouls but still achieves acceptable results, more likely to be drawn. To address this, we focus on key statistics that reflect a team's playstyle: the number of shots (HS, AS) and shots on target (HST, AST) as indicators of attacking intent and excitement, and the number of fouls (HF, AF), yellow cards (HY, AY). We also consider the draw result matches stats and compare the playstyle as home or as against of every team.

## 2.2 Approach

To answer the first question, we try to follow this process. Firstly, we will try to normalize the stats. As there are many matches in which a very dominant team like Man City or Liverpool play against a much weaker team like Norwich or Watford that can generate significant and extreme numbers on goals or shots. By this way it will be more fair to evaluate other teams. Secondly, we want to investigate the "unattractive" matches which are matches with less than 2 goals or 1-1 draw. We expect to see the "unattractive" teams play more in these matches. Then, we will view the "unattractive" week which has the most number of "unattractive" matches. We want to see if the result of the season can affect the performance of each team since we believe that when a team has no motivation they will play less intriguingly. Moreover, we need to consider the playstyle of each team when they are at home compared with when they are as against. This indicates how practical teams are when they have no benefit from the home. Finally we try to create a metric call "dirty score" to find the least "unattractive" teams.

## 2.3 Analysis

We will normalize the stats before visualizing the data.

```
data['HomeShotsNormalized'] = data['HST'] / data['HST'].max()
data['HomeFoulsNormalized'] = data['HF'] / data['HF'].max()
data['HomeYellowCardsNormalized'] = data['HY'] / data['HY'].max()
data['HomeGoalsNormalized'] = data['FTHG'] / data['FTHG'].max()

data['AwayShotsNormalized'] = data['AST'] / data['AST'].max()
data['AwayFoulsNormalized'] = data['AF'] / data['AF'].max()
data['AwayYellowCardsNormalized'] = data['AY'] / data['AY'].max()
data['AwayGoalsNormalized'] = data['FTAG'] / data['FTAG'].max()
```

In this analysis, we define "unattractive" matches based on a combination of normalized metrics, including low goal counts, high fouls, and low shot frequency. We first examine which teams most frequently appear in these matches. After normalizing the relevant variables, we calculate the percentage of each team's appearances in such fixtures and visualize the results using a bar chart.

```
# Step 1: Find the teams that participated in 1-1 draw matches
draw_1_1 = data[(data['FTHG'] == 1) & (data['FTAG'] == 1)]
teams_in_draw_1_1 = pd.concat([draw_1_1['HomeTeam'], draw_1_1['AwayTeam']])

# Step 2: Find the teams that participated in matches with total goals <= 1
total_score_less_than_equal_1 = data[data['FTHG'] + data['FTAG'] <= 1]
teams_in_total_goals_1_or_less = pd.concat([total_score_less_than_equal_1['HomeTeam'], total_score_less_than_equal_1['AwayTeam']])

# Step 3: Calculate the total number of matches each team participated in (home and away)
total_matches = pd.concat([data['HomeTeam'], data['AwayTeam']]).value_counts()

# Step 4: Count how many times each team appears in the filtered matches
draw_1_1_count = teams_in_draw_1_1.value_counts()
total_goals_1_or_less_count = teams_in_total_goals_1_or_less.value_counts()
```

We come up with the pie chart:

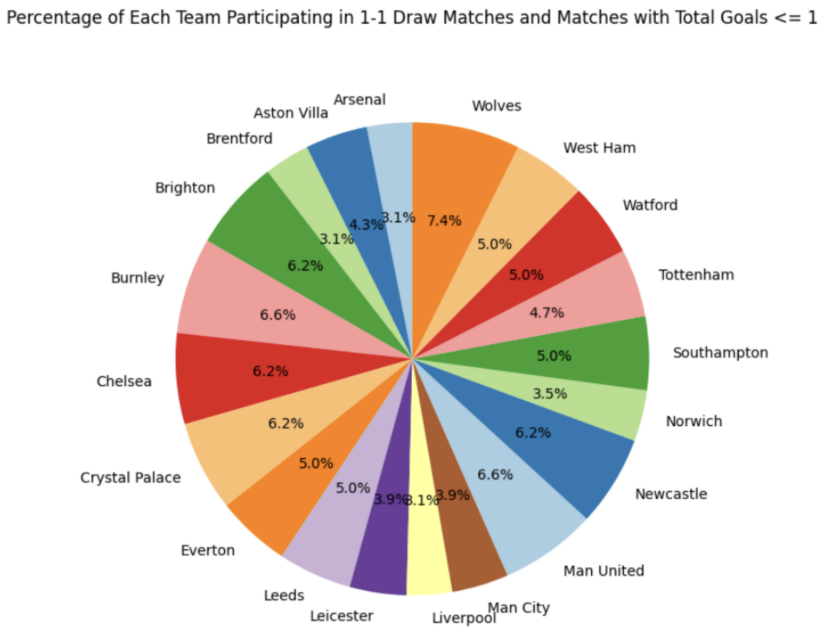



Figure 1 shows that teams such as Manchester United and Wolverhampton Wanderers have the highest percentage of appearances in unattractive matches. Notably, most teams register over 6% participation, indicating that such matches are relatively widespread across the league.

Next, we examine how unattractive matches are distributed throughout the season. By aggregating the number of such matches per week, we aim to uncover any temporal trends, particularly in the final stages of the league.

```
# Step 1: Convert the dates into a datetime format
data['Date'] = pd.to_datetime(data['Date'])

# Step 2: Extract the week number from the Date column
data['Week'] = data['Date'].dt.isocalendar().week

# Step 3: Filter for 1-1 draw matches and matches with total goals <= 1
draw_1_1 = data[(data['FTHG'] == 1) & (data['FTAG'] == 1)]
total_score_less_than_equal_1 = data[data['FTHG'] + data['FTAG'] <= 1]

# Step 4: Combine both filtered datasets into one
combined_matches = pd.concat([draw_1_1, total_score_less_than_equal_1])

# Step 5: Group by week to count the number of combined matches per week
combined_matches_weekly = combined_matches.groupby('Week').size()
```

We come up with the bar chart:

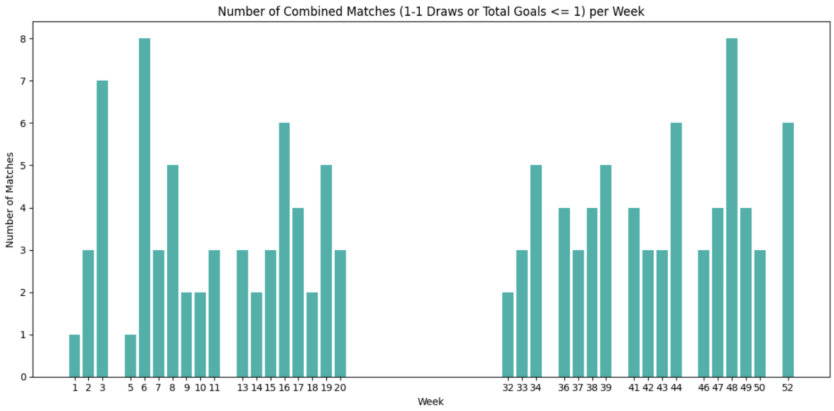



Figure 2 reveals a clear increase in the number of unattractive matches during the final weeks of the season. This trend may be attributed to a lack of competitive motivation among mid-table teams who are neither in contention for European qualification nor at risk of relegation.

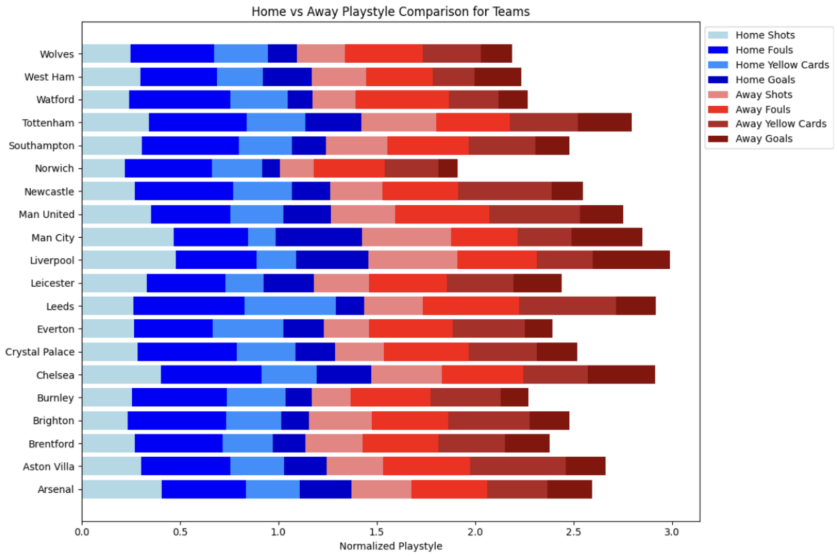
To understand team behavior in different environments, we compare their performance metrics when playing at home versus away. Normalized statistics for fouls, yellow cards, shots, and goals are computed separately for home and away contexts.

```
# Step 2: Aggregate playstyle stats for each team (both home and away)
team_stats = data.groupby('HomeTeam').agg({
    'HomeShotsNormalized': 'mean',
    'HomeFoulsNormalized': 'mean',
    'HomeYellowCardsNormalized': 'mean',
    'HomeGoalsNormalized': 'mean',
}).reset_index()

# Aggregate away stats for the same teams
away_stats = data.groupby('AwayTeam').agg({
    'AwayShotsNormalized': 'mean',
    'AwayFoulsNormalized': 'mean',
    'AwayYellowCardsNormalized': 'mean',
    'AwayGoalsNormalized': 'mean',
}).reset_index()

team_stats = pd.merge(team_stats, away_stats, left_on='HomeTeam', right_on='AwayTeam', how='left')
```

This visualization illustrates a noticeable shift in playstyle between home and away fixtures. Most teams adopt a more cautious or defensive approach when playing away, as reflected by decreased goal attempts and increased fouls or bookings.




## 2.4 Discussion

The three visualizations collectively provide important insights into identifying the most "unattractive" teams in the league. Top-performing teams such as Liverpool and Chelsea exhibit balanced and consistent statistics across all charts, supporting their reputation for dynamic and entertaining football. In contrast, mid-table teams—particularly those ranked between 12th and 16th—appear more frequently in low-tempo or defensive matches.

To quantify unattractiveness further, we introduce a composite metric called the "Dirty Score", calculated as follows:

$$\text{Dirty Score} = \text{Fouls}_{\text{norm}} + \text{YellowCards}_{\text{norm}} + (1 - \text{Shots}_{\text{norm}}) - \text{Goals}_{\text{norm}}$$
$$+\text{UnattractiveMatches}_{\text{norm}} + \text{OpponentStrength}_{\text{norm}} + \text{NoAchievementsLeft}_{\text{factor}}$$

Teams with the highest Dirty Scores—those exhibiting high fouls, low shot volume, and fewer goals—tend to cluster around mid-table rankings. This supports our hypothesis that these teams contribute disproportionately to less engaging football.

While the findings are consistent with expectations, the analysis is constrained by the simplicity of available metrics. Future research should incorporate more sophisticated data such as expected goals (xG), pressing intensity, and passing network structures to better quantify playstyle. Additionally, contextual variables like injuries, squad rotation, or managerial tactics could refine the evaluation of team and player performances.

# 3 QUESTION 2

## 3.1 Introduction

In professional football, not all players meet performance expectations. While clubs aim to field top talent, underperforming players can hinder progress and weaken competitiveness. Data-driven analysis enables clubs to make informed decisions, especially when considering player transfers. This study analyzes individual player performance in the English Premier League 2021–2022 season across three key tactical roles: forwards, midfielders, and defenders. Using two datasets—`all_players_stat.csv` and `soccer21-22.csv`—we assess contributions, goal metrics, and disciplinary records to identify the bottom

four performers in each role. A position-specific approach ensures fairness in evaluating players, guiding management decisions regarding potential offloading.

## 3.2 Approach

To assess underperformance effectively, we developed custom performance scores tailored to each position. Visualizations were central to this analysis. Bar charts rank players based on custom metrics, while scatter plots explore relationships between key performance variables, with color and size encoding additional dimensions. Forwards were assessed using goal metrics (e.g., goal contribution, minutes per goal, and penalty reliance), defenders by appearances, goals conceded, and fouls, and midfielders by assists, passes, and discipline. This approach offers both qualitative and quantitative insights to highlight inefficiencies that may not be obvious in raw statistics.

## 3.3 Analysis

### 3.3.1 Forwards

A bar chart ranks forwards by a custom score derived from goals, contribution percentage, and penalty dependency. Fabio Silva scored the lowest (-36.0), with Armstrong, Weghorst, and Ayew also underperforming. A scatter plot of MinutesPerGoal vs. GoalContribution shows Silva with zero impact despite considerable playtime. A grouped bar chart contrasts penalty dependence and goal efficiency, showing weak conversion rates across the group.

```
forwards['ForwardScore'] = (
    forwards['Goals'] * 10 + forwards['GoalContribution'] * 2 +
    (forwards['GoalEfficiency'] - 1) * 20 - forwards['MinutesPerGoal'] * 0.05 -
    forwards['SubstitutionRate'] * 0.1 - forwards['Penalties'] / forwards['Goals'].clip(lower=1) * 10
)
```

```
# Plot 3: Scatter plot of Minutes Per Goal vs. Goal Contribution
ax3 = plt.subplot(gs[1, 0])
scatter = ax3.scatter(
    x='MinutesPerGoal',
    y='GoalContribution',
    s=100,
    c=worst_4_forwards['Goals'],
    cmap='YlOrRd',
    data=worst_4_forwards
)
```

### 3.3.2 Defenders

A bar chart based on a composite Defender Score (e.g., tackles, interceptions, and blocks) identifies Ngakia, Lyanco, Gomez, and Holding as the least effective. A bubble plot of appearances vs. team goals conceded reveals players linked with poor defensive outcomes. Another bar chart compares fouls and successful tackles, exposing those whose aggressive play doesn't translate to defensive reliability.

```
# Plot 1: Bar chart of the 4 worst defenders with their scores
ax1 = plt.subplot(gs[0, 0])
sns.barplot(x='Player', y='DefenderScore', data=worst_4_defenders, palette='Reds_r', ax=ax1)
ax1.set_title('4 Worst Performing Defenders (Lower Score is Worse)', fontsize=14)
ax1.set_xlabel('Player', fontsize=12)
ax1.set_ylabel('Defender Performance Score', fontsize=12)
ax1.tick_params(axis='x', rotation=45)
```

```
# Plot 2: Bubble plot for Appearances vs. Goals Conceded
ax2 = plt.subplot(gs[1, 0])
scatter = ax2.scatter(
    x='Apearances',
    y='GoalsConceded',
    s=worst_4_defenders['SubstitutionRate'] * 10,
    c=[colors[i] for i in range(len(worst_4_defenders))],
    alpha=0.7,
    data=worst_4_defenders
)
```

```
# Plot 3: Enlarged Bar Chart of Disciplinary Issues vs. Shot Conversion Against
ax3 = plt.subplot(gs[:, 1])
bars1 = ax3.bar(x - width/2, worst_4_defenders['DisciplinaryIssues'], width, label='Disciplinary Issues', color='#FF9999')
bars2 = ax3.bar(x + width/2, worst_4_defenders['ShotConversionAgainst'], width, label='Shot Conversion Against (%)', color='#66B2FF')

ax3.set_title('Disciplinary Issues vs. Shot Conversion Against', fontsize=16)
ax3.set_xticks(x)
ax3.set_xticklabels(worst_4_defenders['Player'], rotation=45, fontsize=12)
ax3.legend(fontsize=12)
```

### 3.3.3 Midfielders

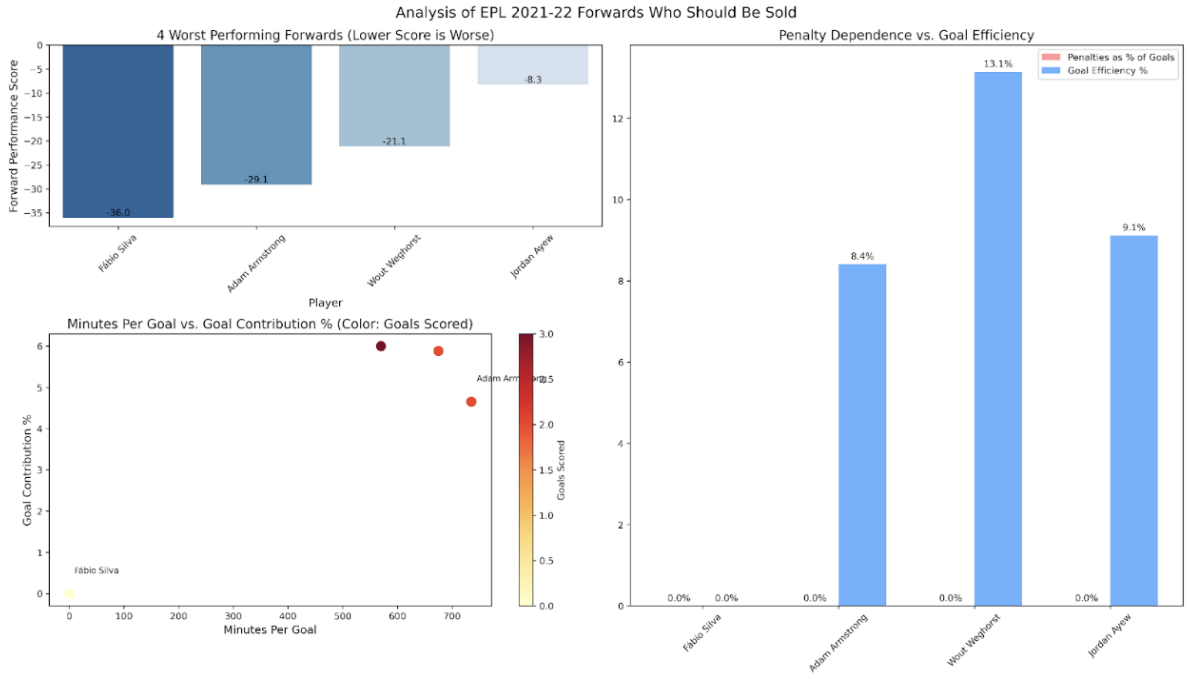
Midfielders were evaluated based on offensive and defensive contributions. A horizontal bar chart shows the lowest assist and key pass counts for El Neny, Nakamba, Young, and Milivojevic. A scatter plot maps total passes vs. yellow cards, with pass accuracy shown via color. Though some had positive performance scores, all had zero goals and poor defensive impact when on the field.

```
# Plot 1: Main Performance Score Chart
ax1 = plt.subplot(gs[0, 0])
bars = ax1.bar(
    worst_midfielders['Player'],
    worst_midfielders['MidfielderScore'],
    color=sns.color_palette("viridis", 4)
)
```

```
# Plot 2: Game Contribution Chart
ax2 = plt.subplot(gs[1, 0])
ax2.bar(r1, worst_midfielders['Apearances'], width=barWidth, edgecolor='grey', label='Appearances', color='lightblue')
ax2.bar(r2, worst_midfielders['Goals'], width=barWidth, edgecolor='grey', label='Goals Scored', color='green')
ax2.bar(r3, worst_midfielders['Apearances'] * worst_midfielders['PlayerDefensiveContribution'], width=barWidth,
       edgecolor='grey', label='Goals Conceded Contribution', color='red')
```

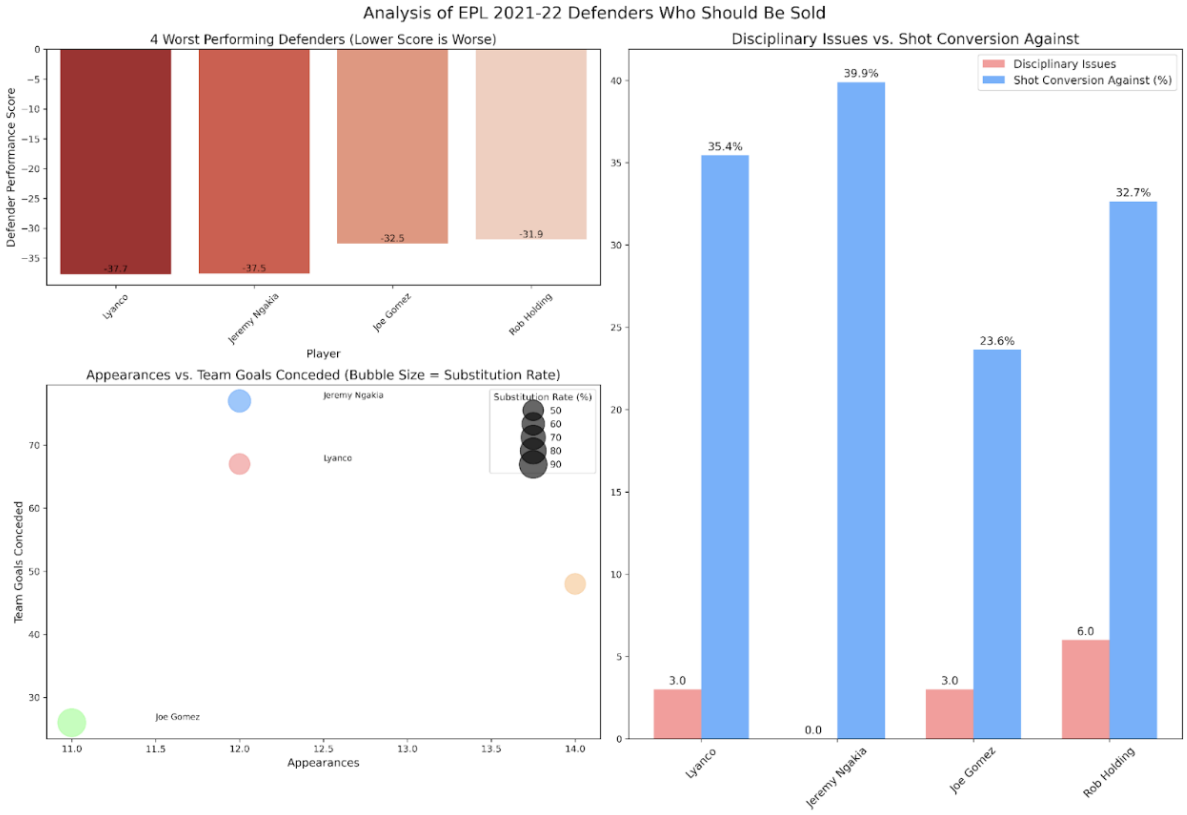
## 3.4 Discussion

### 3.4.1 Forwards



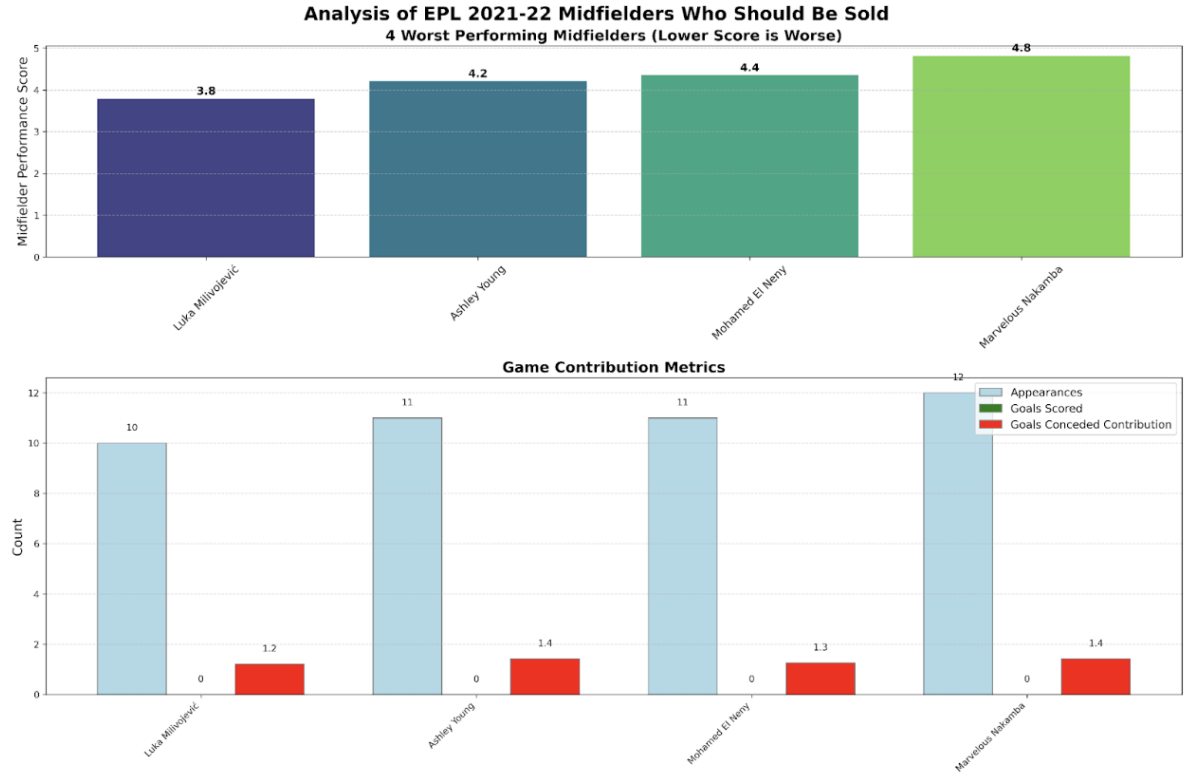

The analysis of forward players indicates significant underperformance among the bottom four. Fabio Silva stands out as the most inefficient attacker, recording a performance score of -36.0 with no goal contribution across numerous minutes played. This suggests a concerning mismatch between playtime and productivity. Adam Armstrong and Wout Weghorst, although managing some goals, performed inefficiently relative to their time on the pitch, with little influence on overall team success. Jordan Ayew showed slightly better goal efficiency but remained below acceptable standards for a forward in a top-flight league. Additionally, none of these players compensated for poor scoring output by creating opportunities for teammates. This highlights a fundamental lack of offensive value, and from a management standpoint, these players could be replaced by more efficient, cost-effective options in the transfer market.

### 3.4.2 Defenders




Among defenders, Jeremy Ngakia and Lyanco emerge as the weakest links based on their highly negative performance scores and poor defensive metrics. The bubble plot clearly shows their presence on the pitch correlated with high team goals conceded, suggesting direct contributions to defensive instability. Joe Gomez and Rob Holding, while slightly better, also exhibited below-average defensive performance, with Holding committing the most fouls among the group. Ngakia and Lyanco also suffered from high shot-conversion rates against, indicating lapses in positioning or marking. These issues, compounded by limited contributions in other defensive areas, justify recommending these players for transfer. Defensive reliability is essential for team consistency, and these players appear to pose a risk rather than offer security.

### 3.4.3 Midfielders




Midfielders are typically expected to bridge the gap between defense and attack, contributing to both phases of play. However, the four lowest-performing midfielders—Nakamba, El Neny, Young, and Milivojevic—failed to meet those expectations. Despite making 10–12 appearances, none recorded a goal, and their assist and key pass numbers were notably low. The scatter plot shows that even their pass volume and accuracy did not compensate for disciplinary issues, such as frequent yellow cards. Furthermore, their “Goals Conceded Contribution” metrics suggest they were on the pitch during periods of defensive weakness. Although their scores remained slightly positive, these players offered minimal impact and may be occupying squad space that could be used more productively. From a strategic standpoint, replacing them with more dynamic midfielders could enhance team balance and creativity.

# 4 CONCLUSION

This project provided an analytical overview of team strategies and player effectiveness within the 2021–2022 English Premier League season. By constructing a "Dirty Score" based on metrics such as fouls, shots, and goal involvement, we identified teams that consistently contributed to less dynamic, more defensively-oriented matches—particularly those positioned mid-table (12th to 16th place). These teams exhibited conservative tactics, especially during away matches or low-scoring fixtures, leading to gameplay that was, by our criteria, less attractive to spectators.

On an individual level, role-specific evaluations allowed for the identification of underperforming players across different positions. Forwards like Fabio Silva and Adam Armstrong registered minimal attacking output relative to their playtime. Defenders such as Lyanco and Jeremy Ngakia displayed weak defensive performance paired with high goals conceded. Similarly, midfielders like Nakamba and El Neny contributed little offensively while offering limited defensive support. While factors such as limited minutes or injuries may explain some of these patterns, the overall trends suggest suboptimal contributions from these individuals.

In summary, this analysis demonstrates the value of structured football data in extracting actionable insights for fans, analysts, and club decision-makers. Future work could incorporate more advanced performance metrics (e.g., xG, key passes, defensive pressures) or use machine learning models to forecast player development and team outcomes. Such enhancements would further improve the robustness and interpretability of performance assessments in professional football.