Finding Competitive Advantages in Fantasy Soccer

By Argy Sakti (das9669) and Abhi Vachani (apv9448)

1. Abstract

The purpose of this study is to identify competitive advantages for participants in the Fantasy Premier League, also known as FPL, an online fantasy soccer competition, by employing a data-driven approach to improve FPL scores. The research problem in this project concerned the determination of soccer metrics that are significant in predicting Premier League players' future performances. To investigate the research problem, the study utilized a quantitative data analysis based on historical data on the English Premier League. Quantitative data analysis involved the collection, cleaning, and examination of player performance metrics and match reports. The major findings of this study were the identification of statistically significant metrics that contribute to a player's FPL scoring actions. According to these findings, the study determined that a data-based approach can provide FPL managers with a competitive edge in decision-making. Consequently, this can ultimately lead to improved overall performance and higher rankings in the Fantasy Premier League.

2. Introduction

This study focused on the Fantasy Premier League, a Fantasy Soccer game with over 11 million users. FPL participants compete every match week to select a squad of 15 players with 11 starters and earn points based on the starting players' minutes, goals, assists, goals conceded, and other significant actions. This study is important to find competitive advantages over millions of other competitors by utilizing available data on the English Premier League.

Considering the global popularity of FPL, there is a wide array of information online that participants can seek to attempt to gain an advantage. The information can be both qualitative and quantitative, such as researching team lineups, injury news, recent statistical performances, and the difficulty of upcoming matches. Further, the Premier League collaborates with a sports analytics company, Opta, to gather performance data. On top of having traditional metrics, their data collection involves sophisticated data analysis in the form of underlying metrics.

Underlying metrics are advanced statistical measures that provide a nuanced understanding of players and team performance by looking at particular actions and situations that transpire during a match. The level of detail in the data on the Premier League opens up the possibility of implementing a refined data-based approach to enhance decision-making in FPL. As such, this study attempts to synthesize the available data to determine the metrics that FPL managers should examine when deciding on their teams. By identifying the most significant metrics for success in FPL, this study will advance knowledge on the topic by helping FPL managers make more informed decisions.

Figure 1 illustrates the pipeline of our pre-analysis of our Premier League datasets. The pre-analysis pipeline involves data acquisition, data ingestion, data cleaning, and data profiling. We utilized NYU Dataproc and the Apache Hadoop Distributed File System to handle our data ingestion step. Then, we used Apache MapReduce and Apache Spark to clean and profile the different datasets. These steps ensure that the datasets used in the study are relevant to answering our research problem. Then, Figure 2 shows the pipeline of our analysis following the pre-analysis stage. In this step, we did different types of analyses using MapReduce according to the appropriate datasets. We analyzed defensive metrics and performed regression analysis between underlying metrics and goal-scoring actions to determine the significance of those metrics as predictors for future performances. Finally, we visualized our findings using Tableau to communicate the results and key insights from our analyses. By understanding the relationship between these metrics and FPL points-scoring actions, this study derived meaningful insights into optimizing FPL performance with a data-based approach.

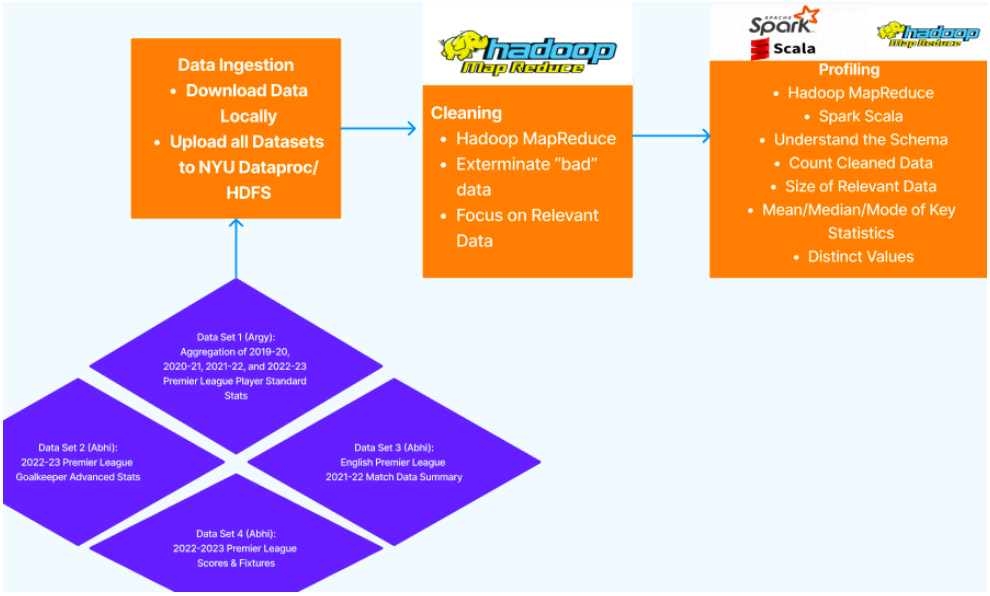

Figure 1: Pipeline of Pre-Analysis

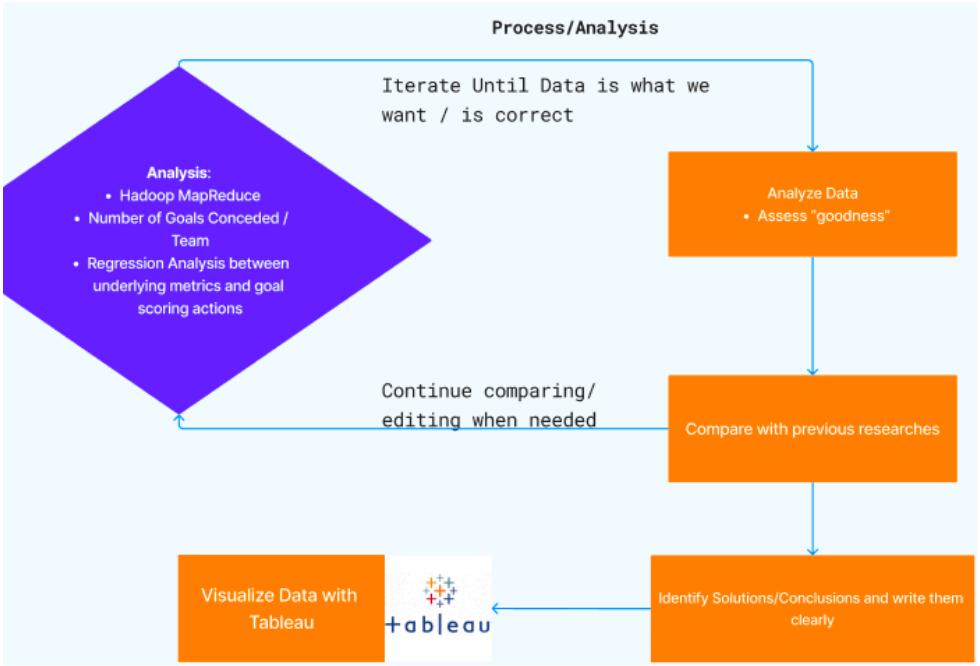

Figure 2: Pipeline of Data Analysis

3. Motivation

Sports analytics is a field with ever-increasing popularity. According to Markets and Markets, the Global Sports Analytics Market is expected to reach $8.4 Billion by the end of 2026. While money was NOT our motivation, this just shows how large the market for sports analytics is.

The reason we chose this topic is because of our love and passion for soccer. We both really enjoy the sport and love to learn about the statistics behind the game. Fantasy Premier League is a game that we both play that also uses current stats to base their scoring system. Naturally, it was a perfect fit for our project. Finally, as we both have a strong base level of knowledge in this subject, a data analysis project made a lot of sense for us because we would not have to take time to understand the scope of the field.

4. Related Works

4.1. Player Recommendation System for Fantasy Premier League using Machine Learning:

The goal of our project was to compare current and past players' real-life performances in games to the predefined points system in Fantasy Premier League. Based on this, we were able to see which players performed better in FPL based on real-life statistics. This article was extremely relevant to our project because they are also analyzing FPL performance based on in-game statistics. However, a key difference is that the related article uses a Machine Learning model to predict the points a certain player would earn in FPL. While they used in-game statistics similar to ours, the approach was more focused on using Machine Learning to predict which players would earn the most points.

4.2. Identification of skill in an online game: The case of Fantasy Premier League:

We used this study to better understand the nuances of FPL. This article is solely focused on the online game as opposed to the athletes themselves. We used this article as a related work because the intricacies of FPL are difficult to understand, and the points-scoring system is something that is fundamental in our research before we could attempt to correlate real-life performance to FPL points.

5. Description of Datasets

5.1. Aggregation of 2019-20, 2020-21, 2021-22, 2022-23 Premier League Player Standard Stats

This dataset is a CSV file provided by FBref and contains the statistics of Premier League players from the 2019-20, 2020-21, 2021-22, and 2022-23 seasons. This dataset contains 2157 rows and 38 columns, and has a size of 344,109 bytes. FBref is a third-party source that collects statistics gathered by Opta. Since Opta works directly with the English Premier League, we were assured that this dataset is reliable. Figure 3 illustrates the schema of this dataset.

Column	**Data Type**	**Description**
Rk	Integer	Rank based on the default sort of the dataset (alphabetical order of player's name)
Player	String	Player's name
Nation	String	Player's nation
Pos	String	Player's position(s)
Squad	String	Player's squad
Age	String	Player's age
Born	Integer	Player's year of birth
MP	Integer	Matches played
Starts	Integer	Number of starts
Min	Integer	Number of minutes played
90s	Float	Minutes played divided by 90
Gls	Integer	Goals scored
Ast	Integer	Goals assisted
G+A	Integer	Total number of goals scored and assisted
G-PK	Integer	Non-penalty goals
PK	Integer	Penalty kicks scored
Pkatt	Integer	Penalty kicks attempted

CrdY	Integer	Yellow cards
CrdR	Integer	Red cards
xG	Float	Expected goals
npxG	Float	Non-penalty expected goals
xAG	Float	Expected assisted goals
npxG + xAG	Float	Non-penalty expected goals plus expected assisted goals
PrgC	Integer	Progressive carries
PrgP	Integer	Progressive passes
PrgR	Integer	Progressive passes received
Gls per 90	Float	Goals scored per 90 minutes
Ast per 90	Float	Goals assisted per 90 minutes
G+A per 90	Float	Total number of goals scored and assisted per 90 minutes
G-PK per 90	Float	Non-penalty goals per 90 minutes
G + A - PK per 90	Float	Total number of non-penalty goals scored and assisted per 90 minutes
xG per 90	Float	Expected goals per 90 minutes
xAG per 90	Float	Expected assisted goals per 90 minutes
xG + xAG per 90	Float	Expected goals plus expected assisted goals per 90 minutes
npxG per 90	Float	Non-penalty expected goals per 90 minutes
npxG + xAG per 90	Float	Non-penalty expected goals plus expected assisted goals per 90 minutes

Figure 3: Schema of Dataset 5.1

As shown by Figure 3, the dataset contains traditional metrics such as goals and assists, as well as underlying metrics such as expected goals and expected assists. With the aim of understanding

metrics that can be predictive of future Premier League performances, this dataset is vital in helping us quantify the relationship between performance metrics and FPL-scoring actions. Additionally, the dataset is a substantial sample size because it spans multiple seasons. As a result, this dataset is robust against major outliers.

5.2. 2022-23 Premier League Goalkeeper Advanced Stats

This dataset is also provided by fbref.com. It contains the statistics of 2022-23 Premier League goalkeepers. Since these are considered 'advanced' statistics, they need careful human analysis as well because some of the metrics may not be as clear as others. They contain facts and metrics, just like the dataset for the players. The schema is shown below:

Column	Data Type	Description
Rk	Integer	Rank based on the default sort of the dataset (alphabetical order of player's name)
Player	String	Player's name
Nation	String	Player's nation
Pos	String	Player's position
Squad	String	Player's squad
Age	String	Player's age
Born	Integer	Player's year of birth
90s	Float	Number of Full Matches played
GA	Integer	Number of Goals Against (Conceded)
PKA	Integer	Number of Penalty Kicks Allowed
FK	Integer	Number of Free Kick Goals Against
CK	Integer	Number of Corner Kick Goals Against
OG	Integer	Number of Own Goals Scored Against Goalkeeper
PSxG	Float	Post-Shot Expected Goals

PSxG/SoT	Float	Post-Shot Expected Goals on Target
PSxG+/-	String	Post-Shot Expected Goals Minus Goals Allowed
/90	String	Post-Shot Expected Goals Minus Goals Allowed per 90 mins
Cmp	Integer	Passes Completed (Longer than 40 yards)
Att	Integer	Passes Attempted (Longer than 40 yards)
Cmp%	Float	Pass Completion Percentage (Longer than 40 yards)
Att	Integer	Passes Attempted (Not Including Goal Kicks)
Thr	Integer	Throws Attempted
Launch%	Float	Percentage of Passes That Were Launched (Not Including Goal Kicks & Longer than 40 yards)
AvgLen	Float	Average Length of Passes in yards (Not Including Goal Kicks)
Att	Integer	Goal Kicks Attempted
Launch%	Float	Percentage of Goal Kicks that were Launched (Passes Longer than 40 yards)
AvgLen	Float	Average Length of Goal Kicks
Opp	Integer	Crosses Faced
Stp	Integer	Crosses Stopped
Stp%	Float	Crosses Stopped Percentage
#OPA	Integer	Defensive Actions Outside Penalty Area
#OPA/90	Float	Defensive Actions Outside Penalty Area per 90 minutes
AvgDist	Float	Average Distance of Defensive Actions
Matches	String	Link To All Matches Played (Match Log Dataset)

Figure 4: Schema of Dataset 5.2

As seen in the figure above, this dataset contains key actions and underlying metrics performed by all Goalkeepers in the English Premier League. As a part our project focuses solely on current Goalkeepers, this dataset is helpful for understanding the relationship between in game actions and FPL points scoring.

5.3. English Premier League 2021-22 Match Data Summary

This Dataset contains information from every match during the 2021-22 English Premier League season. Data includes teams, referee, and stats by home and away side such as fouls, shots, cards, and more. The schema is shown below:

Column	Data Type	Description
Date	Date	Date the Match was Played
Hometeam	String	Home Team
Awayteam	String	Away Team
Fthg	Integer	Full Time Home Goals
Ftag	Integer	Full Time Away Goals
Ftr	String	Full Time Result
Hthg	Integer	Half Time Home Goals
Htag	Integer	Half Time Away Goals
Htr	String	Half Time Result
Referee	String	Referee during Match
Hs	Integer	Number of Home Shots
As	Integer	Number of Away Shots
Hst	Integer	Home Shots on Target
Ast	Integer	Away Shots on Target
Hf	Integer	Home Fouls Committed

Af	Integer	Away Fouls Committed
Hc	Integer	Home Clearances
Ac	Integer	Away Clearances
Hy	Integer	Home Yellow Cards
Ay	Integer	Away Yellow Cards
Hr	Integer	Home Red Cards
Ar	Integer	Away Red Cards

Figure 5: Schema of Dataset 5.3

As shown in the figure above, this dataset contains all game information and statistics from every match during the 2021-22 English Premier League season. As we wanted to compare and analyze our work as it relates to last season, this dataset was important to us. This dataset was especially useful in calculating the number of goals conceded per team for an entire season.

5.4. 2022-23 Premier League Scores and Fixtures

Our final Dataset contains information from every match during the 2022-23 English Premier League season. Similar to the previous dataset, key metrics such as the teams playing and the scoreline of the games are included. However, the schema is different, as shown below:

Column	**Data Type**	**Description**
Wk	Integer	Matchweek Number
Day	String	Day of Week
Date	String	Date of Match
Time	String	Time of Match
Home	String	Home Team
xG	Float	Expected Goals of Home Team

Score	String	Score of Match
xG	Float	Expected Goals of Away Team
Away	String	Away Team
Attendance	Integer	Number of Fans in Crowd
Venue	String	Stadium Match was Played at
Referee	String	Referee of Match
Match Report	String	Link to Match Report (fbref.com)
Notes	String	Additional Notes

Figure 6: Schema of Dataset 5.4

As shown in Figure 6, Dataset 5.4 contained information for every match during the 2022-23 English Premier League Season, the current season. Since we analyzed data from athletes who are currently playing, this dataset was helpful for us to explore every match of the current season.

6. Analytics Stages

For all the datasets specified above in section 5, we used NYU Dataproc and, subsequently, the Apache Hadoop Distributed File System, abbreviated as HDFS, to perform data ingestion. We stored the datasets in custom directories on HDFS. Following the data-ingestion phase, we performed data cleaning on each of the datasets using Apache MapReduce. This step involves removing null columns and redundant data in all the datasets. Then, however, there are dataset-specific issues that were required to resolve.

In dataset 5.1, there were two significant cleaning issues, both of which were related to the format of the dataset. Since the dataset is an aggregation of statistics from multiple seasons of the English Premier League, there were additional headers within the row data that should not be there. Then, we also needed a column that informs us of the season that applies to any particular column. To resolve these issues, we used the additional headers within the dataset as a marker to

distinguish between the different seasons. For every header that we encountered, we decrement the season year by one, starting from 2022-23 until 2019-20. As such, we added a "Year" column to the dataset. Further, we also removed all the headers but the top header from the dataset.

In datasets 5.2 – 5.4, there were a few cleaning issues. To begin with, the issue of bad rows had to be solved during this phase. This included rows that contained extra commas and were poorly formatted. As a result, these had to be fixed in order for us to continue working with this dataset. Additionally, there existed header rows that needed to be ignored while performing the analysis. Finally, we wanted to focus our data on key columns, and thus, our cleaning phase also emphasized narrowing down the datasets.

To answer our research problem of determining metrics that are most significant to predicting Premier League players' future performances, we identified the relationship between the metrics in our dataset with the total amounts of goals scored. As FPL focuses on individual players' performances, this analysis focused on each player individually before being aggregated together as a whole. Intuitively, when trying to predict future goal statistics, historical goal statistics seemed to be a logical metric to investigate. We used MapReduce and found that the regression for each individual player between the 2020-21 and 2021-22 seasons was 0.65. We did not have much understanding of the context of this particular value, and as such, we looked at the relationship of other metrics with the number of goals scored.

From all the available metrics, we found the expected goals metric to be the best predictor for future goals, even having a higher regression value than historical goals scored. With our analysis on MapReduce, we determined that the regression value between expected goals and the number of goals scored from the 2019-20 to 2022-23 Premier League seasons is 0.85. With this finding, we wanted to investigate the nuances of the expected goals metric and how it might apply to each individual player. Thus, we looked at the difference between actual goals scored and expected goals for each individual player. We found that there is a wide range for that statistic. So, while players might have similar expected goals, the number of goals they actually score could have a large disparity, which ultimately has an impact on their value as FPL assets. Figure 7 visualizes the goals minus the expected goals metric. It shows that most players tend to

gravitate around the zero value, meaning they score as many goals as expected. However, there are players that significantly outperform and underperform their expected goals. At the top, we have Son Heung-Min, who has a goals - expected goals value of 17.30, which determines that he is extremely efficient when given a scoring opportunity. On the other hand, we have Neal Maupay at the bottom, who underperformed his expected goals by 12.60. Consequently, this shows that while the expected goals metric is a good predictor of future goal performance, it is also important to note the ability of the player selected, as their skill to finish chances in front of goal can have a large impact on their FPL points.

On the other hand, another part of our analysis focused on defensive actions by goalkeepers. Arguably the most important piece of data for a goalkeeper is the number of goals conceded. In turn, the best goalkeepers and defensive teams have the least number of goals conceded. Therefore, we chose to use our datasets to calculate this. By running our MapReduce program on dataset 5.3, we used the Mapper class to map scores from every game with the respective home and away teams and, subsequently, their team names. Following this, we used the Reducer class to aggregate all home and away scores as well as the team names that these numbers belonged to. Once they were all together, the final part of the Reducer class sums the total number of goals conceded per team. Finally, since our project focused on individual actions, we had to map goals conceded per team to goalkeepers themselves. Our spark program was run on dataset 5.2 to understand what teams the goalkeepers played for. This led to our final conclusion that Manchester City, Liverpool, and Chelsea conceded the fewest goals during the 2021-22 season. By combining our profiled data with the results above, we learned that Ederson Moraes, Alisson Becker, and Édouard Mendy, respectively, should be the highest performers in Fantasy Premier League based on goals conceded.

7. Visual Representation of Analytics

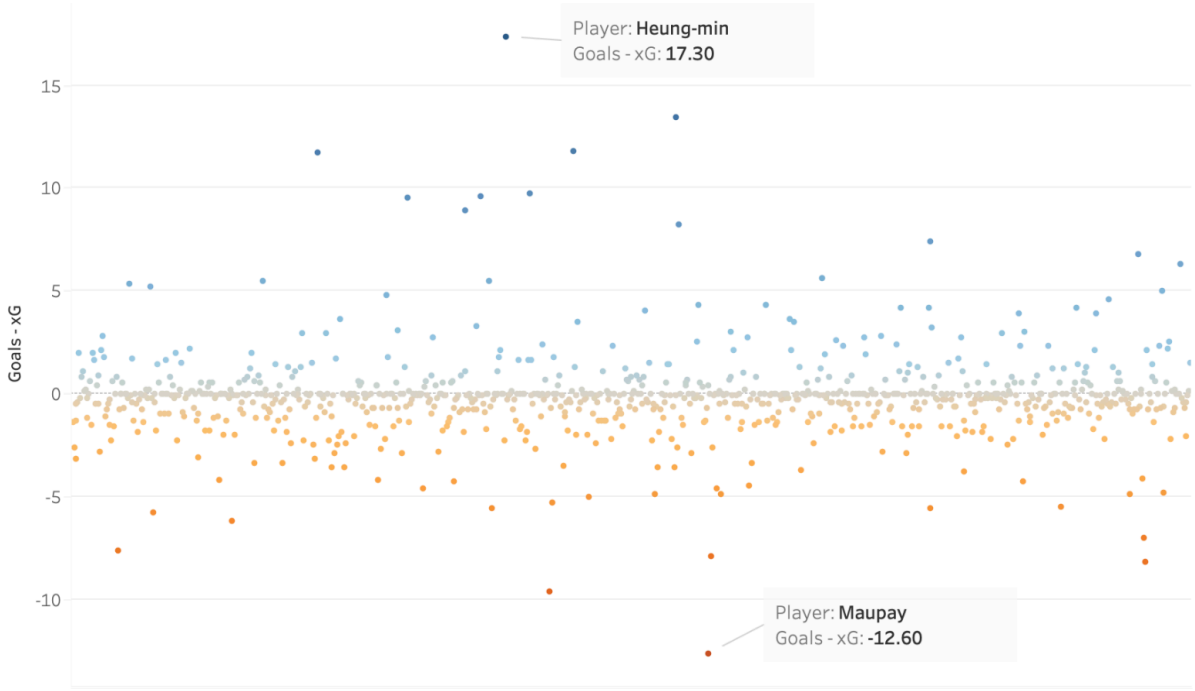

Figure 7: Goals Scored - Expected Goals of English Premier League Players From 2019-20 to 2022-23

8. Conclusion

This study has explored the significance of various metrics in predicting the future performance of English Premier League players in the context of Fantasy Premier League. Through the application of NYU Dataproc, Apache Hadoop Distributed File System (HDFS), and Apache MapReduce, we successfully ingested, cleaned, and analyzed multiple datasets to uncover key insights.

Our findings revealed that the expected goals metric is the strongest predictor for future goals, with a regression value of 0.85. However, we also found that the disparity between actual goals scored and expected goals for individual players can vary widely, highlighting the importance of a player's ability to finish goal-scoring opportunities when selecting them for Fantasy Premier League teams.

Additionally, our analysis of defensive actions by goalkeepers demonstrated that the number of goals conceded is a critical data point for determining a goalkeeper's performance. Manchester City, Liverpool, and Chelsea emerged as the top teams in terms of fewest goals conceded during the 2021-22 season, with Ederson Moraes, Alisson Becker, and Édouard Mendy identified as the highest performers according to this metric.

This study provides valuable insights for Fantasy Premier League participants to optimize their team selection. We can build on this study by continually refining our understanding of the relationships between the different metrics as more data becomes publicly available to us. Additionally, we can use the findings in this study as the foundation for more in-depth predictive models, taking into account other factors such as fixture difficulty and home advantage.

Acknowledgements

We would like to thank NYU HPC for letting us process our data on their servers, Tableau for the free trial access, and FBref and data.world for making their data available for public access. Also, we would like to thank Professor Malavet for her teaching and help throughout the Spring 2023 semester.

References

"English Premier League 2021-22 Match Data." *Data.world*, Data.world, https://data.world/evangower/english-premier-league-2021-22-match-data.

Opta. "Opta-Football." *Stats Perform*, Opta, 24 Feb. 2021, https://www.statsperform.com/opta-football/.

O'Brien JD, Gleeson JP, O'Sullivan DJP (2021) Identification of skill in an online game: The case of Fantasy Premier League. PLoS ONE 16(3): e0246698. https://doi.org/10.1371/journal.pone.0246698

"Premier League Player Stats." *FBref.com*, FBref, 7 May 2023, https://fbref.com/en/comps/9/stats/Premier-League-Stats#all_stats_standard.

“Sports Analytics Market Overview.” *MarketsandMarkets*, MarketsandMarkets, Mar. 2022, https://www.marketsandmarkets.com/Market-Reports/sports-analytics-market-35276513.html#:~:text=The%20global%20Sports%20Analytics%20Market,27.3%25%20from%202021%20to%202026.

V. Rajesh, P. Arjun, K. R. Jagtap, S. C. M and J. Prakash, "Player Recommendation System for Fantasy Premier League using Machine Learning," 2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE), Bangkok, Thailand, 2022, pp. 1-6, doi: 10.1109/JCSSE54890.2022.9836260.