Abstract

This document seeks to analysis the metrics of both players and clubs with respect to their performance in the 2020/21 EPL

SOCCER ANALYSIS

Analysis of the English Premier League



STUDENT NUMBER: B1680944

AGBOKA, JUDE SEDEM KOJO (Student)

B1680944@live.tees.ac.uk

Contents

[EXECUTIVE SUMMARY 5](#_Toc124124362)

[Key Findings 5](#_Toc124124363)

[Recommendations 10](#_Toc124124364)

[INTRODUCTION 10](#_Toc124124365)

[DATA SOURCE AND DESCRIPTION 11](#_Toc124124366)

[Source 11](#_Toc124124367)

[Description of Dataset 11](#_Toc124124368)

[OBJECTIVE, AIM, AND OVERVIEW OF BUSINESS INTELLIGENCE QUESTIONS 13](#_Toc124124369)

[Analysis and Visualisation of the dataset 14](#_Toc124124370)

[Insight into the influencer of a player by position 14](#_Toc124124371)

[Influencer and segment section by Defender 15](#_Toc124124372)

[Influencer and segment section by Forward 20](#_Toc124124373)

[Influencer and segment section by Goalkeepers 26](#_Toc124124374)

[Influencer and segment section by Midfielder 30](#_Toc124124375)

[Overview of Goals and penalty Goals by position 36](#_Toc124124376)

[Provide an overview of the average age of players and the number of matches played. 37](#_Toc124124377)

[Along with the average minutes and average age by position 38](#_Toc124124378)

[Understanding of how these metrics affect club performance by comparing Average Assists, average goals, and average Xa and xG by club 39](#_Toc124124379)

[SUMMARY – Finding based on analysis and evaluation 40](#_Toc124124380)

[Recommendations 41](#_Toc124124381)

[Conclusion 42](#_Toc124124382)

[APPENDIX 42](#_Toc124124383)

[DATA PRE-PROCESSING 42](#_Toc124124384)

[Loading Data 42](#_Toc124124385)

[Data Cleaning 43](#_Toc124124386)

[Removing blanks 44](#_Toc124124387)

[Replacing Misspelt Names/Values 45](#_Toc124124388)

[RENAMING COLUMNS 46](#_Toc124124389)

[DATA MODELLING 47](#_Toc124124390)

[Dimension Tables 48](#_Toc124124391)

[Creating Relationship 49](#_Toc124124392)

[INTRODUCTION 54](#_Toc124124393)

[Dataset 55](#_Toc124124394)

[Data Model 55](#_Toc124124395)

[Dax and m language 56](#_Toc124124396)

[Home page and navigation 61](#_Toc124124397)

[Player Profile 65](#_Toc124124398)

[Club Profile 73](#_Toc124124399)

[AI & Q&A 75](#_Toc124124400)

[Reference 76](#_Toc124124401)

[Figure 1. Average age by club 6](#_Toc124124404)

[Figure 2.average minutes and age by position 7](#_Toc124124405)

[Figure 3. goals by club 7](#_Toc124124406)

[Figure 4. Key influencer by position DF 8](#_Toc124124407)

[Figure 5. KEY INFLUENCER 9](#_Toc124124408)

[Figure 6. Screenshot of the EPL Dataset before pre-Processing 13](#_Toc124124409)

[Figure 7. KEY INFLUENCER CHART 14](#_Toc124124410)

[Figure 8. KEY INFLUENCER DF 1 16](#_Toc124124411)

[Figure 9. KEY INFLUENCER DF 2 16](#_Toc124124412)

[Figure 10. TOP SEGMENT DF 1 17](#_Toc124124413)

[Figure 11. TOP SEGMENT DF 2 18](#_Toc124124414)

[Figure 12. TOP SEGMENT DF 3 19](#_Toc124124415)

[Figure 13. KEY INFLUENCER FW 1 20](#_Toc124124416)

[Figure 14. KEY INFLUENCER FW 1 21](#_Toc124124417)

[Figure 15. TOP SEGMENT FW 1 22](#_Toc124124418)

[Figure 16. TOP SEGMENT FW 2 23](#_Toc124124419)

[Figure 17. TOP SEGMENTS FW 3 24](#_Toc124124420)

[Figure 18. TOP SEGMENT FW 4 25](#_Toc124124421)

[Figure 19. KEY INFLUENCER GK 1 26](#_Toc124124422)

[Figure 20. KEY INFLUENCERS BY GK 2 27](#_Toc124124423)

[Figure 21. TOP SEGMENT GK 1 28](#_Toc124124424)

[Figure 22. TOP SEGMENTS GK 2 29](#_Toc124124425)

[Figure 23. KEY INFLUENCERS MF 1 30](#_Toc124124426)

[Figure 24. KEY INFLUENCERS MF 2 31](#_Toc124124427)

[Figure 25. KEY INFLUENCER MF 3 32](#_Toc124124428)

[Figure 26. TOP SEGMENTS MF 1 33](#_Toc124124429)

[Figure 27. TOP SEGMENT MF 2 34](#_Toc124124430)

[Figure 28. TOP SEGMENTS MF 3 35](#_Toc124124431)

[Figure 29. GOALS AND PENALTY GOALS BY POSITION 36](#_Toc124124432)

[Figure 30.AVERAGE AGE BY MATCHES 37](#_Toc124124433)

[Figure 31. AVERAGE MINS AND AGE BY POSITION 38](#_Toc124124434)

[Figure 32. AVR xG,xA, Mean of goals and assist by club 39](#_Toc124124435)

[Figure 33. EPL FILE BEING LOADED 43](#_Toc124124436)

[Figure 34. COLUMN REMOVAL 44](#_Toc124124437)

[Figure 35. REMOVING BLANKS 45](#_Toc124124438)

[Figure 36. REPLACING MISSPELLED NAMES 46](#_Toc124124439)

[Figure 37.Shows a screenshot of the unit of the column which needs to be changed 46](#_Toc124124440)

[Figure 38.Screenshot of changed unit 47](#_Toc124124441)

[Figure 39.Screenshot before column renaming 47](#_Toc124124442)

[Figure 40.Screenshot after renaming a column 47](#_Toc124124443)

[Figure 41.Screenshot of the table before creating the Dimension and Fact table 48](#_Toc124124444)

[Figure 42.To create a relationship an id has to be created using the index column 49](#_Toc124124445)

[Figure 43.IDs linking the fact table to the dimension table 49](#_Toc124124446)

[Figure 44.Shows the duplicated datasets used for the fact and dimension table 50](#_Toc124124447)

[Figure 45. FACTS DATA 50](#_Toc124124448)

[Figure 46. CLUB DIMENSION DATA 1 50](#_Toc124124449)

[Figure 47. PLAYER DIMENSION DATA 2 51](#_Toc124124450)

[Figure 48. CREATING RELATIONSHIP 1 51](#_Toc124124451)

[Figure 49. CREATING RELATIONSHIP 2 52](#_Toc124124452)

[Figure 50.Autodetect CAN ALSO BE USED to build a relationship between the fact and dimension TABLE 53](#_Toc124124453)

[Figure 51. The relationship created between the fact and dimension table (Star Schema model) 53](#_Toc124124454)

[Figure 52. DATA MODEL 55](#_Toc124124455)

[Figure 53. CREATING CONDITIONAL COLUMN FOR XG PERFORMANCE 57](#_Toc124124456)

[Figure 54. CONDITIONAL COLUMN FOR xG PERFORMANCE 57](#_Toc124124457)

[Figure 55. CREATING CONDITIONAL COLUMN FOR xA PERFORMANCE 58](#_Toc124124458)

[Figure 56. CONDITIONAL COLUMN FOR xA PERFORMANCE 58](#_Toc124124459)

[Figure 57. M LANGUAGE 1 59](#_Toc124124460)

[Figure 58.M LANGUAGE 2 59](#_Toc124124461)

[Figure 59. SUSPENSION QUERY 60](#_Toc124124462)

[Figure 60. FINAL RELATIONSHIP OF TABLES 60](#_Toc124124463)

[Figure 61. Shows a screenshot of the homepage 61](#_Toc124124464)

[Figure 62.Shows the names of the buttons to access the pages 61](#_Toc124124465)

[Figure 63.Shows a navigation button on all the pages to help in the navigation of the pages 62](#_Toc124124466)

[Figure 64. PLAYER PROFILE DASHBOARD 1 62](#_Toc124124467)

[Figure 65.. PLAYER PROFILE DASHBOARD 2 63](#_Toc124124468)

[Figure 66. CLUB PROFILE DASHBOARD 63](#_Toc124124469)

[Figure 67. AI DASHBOARD 64](#_Toc124124470)

[Figure 68. ASK QUESTIONS DASHBOARD 64](#_Toc124124471)

[Figure 69. MINUTES 65](#_Toc124124472)

[Figure 70. GOALS 66](#_Toc124124473)

[Figure 71. ASSISTS 66](#_Toc124124474)

[Figure 72. GOALS AND PENALTY GOALS BY POSITION 66](#_Toc124124475)

[Figure 73. AVERAGE AGE BY MATCHES 67](#_Toc124124476)

[Figure 74. AVERAGE xA,xG, by position 68](#_Toc124124477)

[Figure 75. RED AND YELLOW CARDS BY POSITION 68](#_Toc124124478)

[Figure 76. PERC PASSES COMPLETED & GT COUNT PASS ATTEMTD BY POSITION 69](#_Toc124124479)

[Figure 77. AVERAGE RED AND YELLOW CARDS BY CLUB AND MATCHES 70](#_Toc124124480)

[Figure 78. AVERAGE GOALS 71](https://liveteesac-my.sharepoint.com/personal/b1680944_live_tees_ac_uk/Documents/Power%20bi%20ICA.edited%20(1).docx#_Toc124124481)

[Figure 79. AVERAGE ASSISTS 71](https://liveteesac-my.sharepoint.com/personal/b1680944_live_tees_ac_uk/Documents/Power%20bi%20ICA.edited%20(1).docx#_Toc124124482)

[Figure 80. AVERAGE MINS & AVERAGE AGE BY POSITION 72](#_Toc124124483)

[Figure 81. AVERAGE xG, xA, Mean of gOALS BY CLUB 72](#_Toc124124484)

[Figure 82. AVERAGE AGE BY CLUB 73](#_Toc124124485)

[Figure 83. GOALS BY CLUB 74](#_Toc124124486)

[Figure 84. DECISION TREE 74](#_Toc124124487)

[Figure 85. FUTURE FORCAST OF GOALS AND ASSIST BY POSITION 75](#_Toc124124488)

[Figure 86. Q & A 76](#_Toc124124489)

[Figure 87. SLICERS 76](#_Toc124124490)

# EXECUTIVE SUMMARY

The English Premier League is well watched over the world because of the excitement it brings to our homes. The league has transformed how the game is played from a tactical point of view. From the Alex Ferguson era to the Jose Mourinho years and now to the tactical masterclass of Pep Guardiola. All this is due to their genius minds and smart ways of analysing their team and opposition players.

Let us put ourselves in the shoes of these geniuses and seek to analyse this dataset by asking the following:

.

* How to determine the average goals and assist per minute per player.
* Insight into the average xA and xG by position.
* Overview of Goals and penalty Goals by position.
* Provide an overview of the age of the player and the number of matches played.
* Comparison of Average Goals and Assist by position.
* Understanding of how these metrics affect club performance by comparing Goals, average goals, and average xA and xG by club.

## Key Findings

* The average age of the highest performing clubs was 25. This shows that it is good to have a good balance between youth and older players to bring about experience and youthful athleticism.
* Experience is key as a goalkeeper. Goalkeepers tend to improve as the years go by. Among all the positions Goalkeepers had the most minutes and had an average age high above the normal.
* The top 4 clubs that season had the highest number of goals. This proves the importance of goals in winning games and points.
* The average xA and xG do not determine the position of a club come the end of the season.
* Player position is a key influencer in determining the average xG and xA of a player.
* The average age of 25 had the highest number of matches and minutes.
* Shockingly Midfielders occupied the top 3 positions with the highest number of red and yellow cards.
* Defenders had the highest percentage of passes completed against the passes attempted.

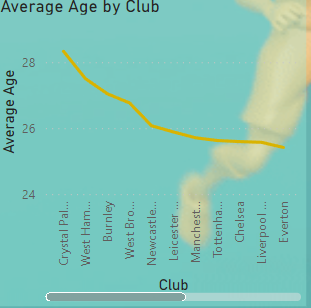


Figure 1. Average age by club

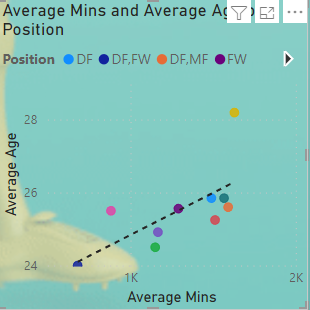


Figure 2.average minutes and age by position



Figure 3. goals by club

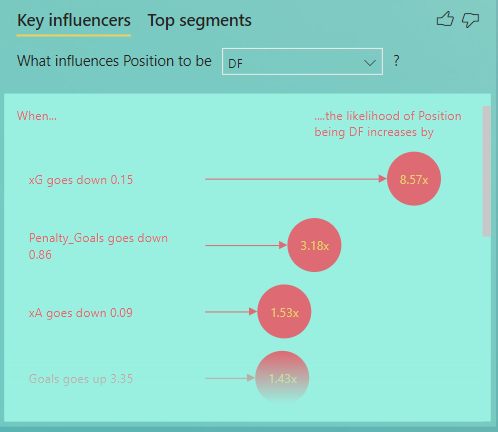


Figure 4. Key influencer by position DF

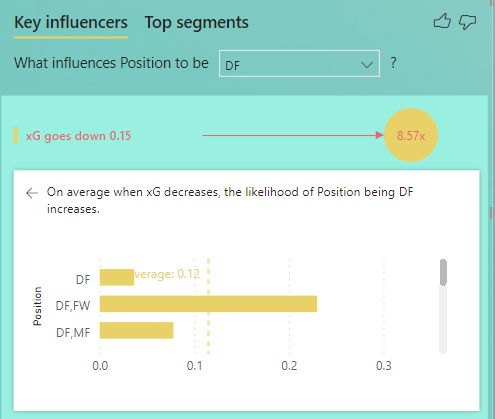


Figure 5. KEY INFLUENCER

## Recommendations

* A different metric other than xA and xG should be used to rate the goalkeeper's performance because xG and xA are reflective of a player's position.
* More effort should be put in place to increase the average goals of midfielders.
* Clubs should balance between being defensive and offensive to get a high chance of winning the league.
* Offensive players should be efficient and ruthless in front of goal and making good use and being aware of their positioning in a tactical setup to counteract the effect of the opposition thereby increasing their xG and xA.
* Even though the league is progressive, there is a need for improvement in different positions concerning clubs to make it more competitive

# INTRODUCTION

The game of football is one that we all love to watch. Over the years the game has become more tactical than physical. The game has evolved, and the English premier league is spearheading this evolution. Player metrics have been used to analyse opposition players along with club performance in recent years.

This project seeks to analyse and visualize English Premier League 2020/21 dataset and gain insights into the factors and metrics affecting players and club performance. This is done through analysis of the EPL dataset through pre-processing, Data visualization, and Artificial Intelligence.

## DATA SOURCE AND DESCRIPTION

### Source

The dataset was sourced from Kaggle an online open-source data library via https://www.kaggle.com/datasets/rajatrc1705/english-premier-league202021. Previous research has been made on this dataset using a few programming languages like Python and R to analyse and visualize the data. This paper aims to seek to analyse and visualize the dataset using Power bi. The dataset has 533 rows and 18 columns before pre-processing.

Reason for choosing this Dataset

The dataset was selected due to several reasons.

* First, it was sourced from a credible website (Kaggle). The credibility of the data source is very important in data analysis and visualization. Data sourced from a non-credible website may be biased and contain several disparities. Therefore, getting a very credible dataset is very essential in data analysis and visualization because it prevents data skewness.
* Secondly, it has understandable attributes and sufficient information which is very essential in creating a relationship and finding a correlation between the data through Data modelling (Star Schema) in addition to finding several questions and solutions through AI (Intelligent questions), linear regression (DAX), and visualization.
* Moreover, the dataset has maximum rows and columns, along with a few outliers and missing data (NA’s) to enable pre-processing steps.

### Description of Dataset

Table 1. Shows the column's names and description of the dataset

|  |  |  |
| --- | --- | --- |
| No. | Attributes | Description |
|  | Name | Name of the Football player |
|  | Club | The football club of the player |
|  | Nationality | The country the player is from |
|  | Position | The player's position (Forward, Midfielder, Defender, Goalkeeper) |
|  | Age | The age of the player |
|  | Matches | Number of matches played by the player |
|  | Starts | Number of matches the player started |
|  | Minutes | Minutes of matches the player played |
|  | Goals | Goals per player |
|  | Assist | Assists per player |
|  | Passes Attempted | Number of passes attempted by a player |
|  | Percentage of Passes Completed | The percentage of Passes completed by the player (Cross, Non-Cross, Header, through balls, corner, long balls, etc) |
|  | Penalty Goals | Number of Penalty Goals by player |
|  | Penalty Attempted | Number of penalties attempted by a player |
|  | Xg (Expected Goal) | The average likelihood that the player scored from a particular position on the field. It ranges from 0 to 1. Zero represents impossible to score and one represents expected to score by a player |
|  | xA (Expected Assist) | The average Quality of passes and creative productive output via chances created from open play leading to scoring opportunities (ranges from 0 to 1. Zero represents a pass leading to the goal which will not result in a goalscoring assist and one represents an assist successful and results in a goal from every event by a player. |
|  | Yellow card | Number of yellow cards per player |
|  | Red Card | Number of red cards(sanctions) per player |

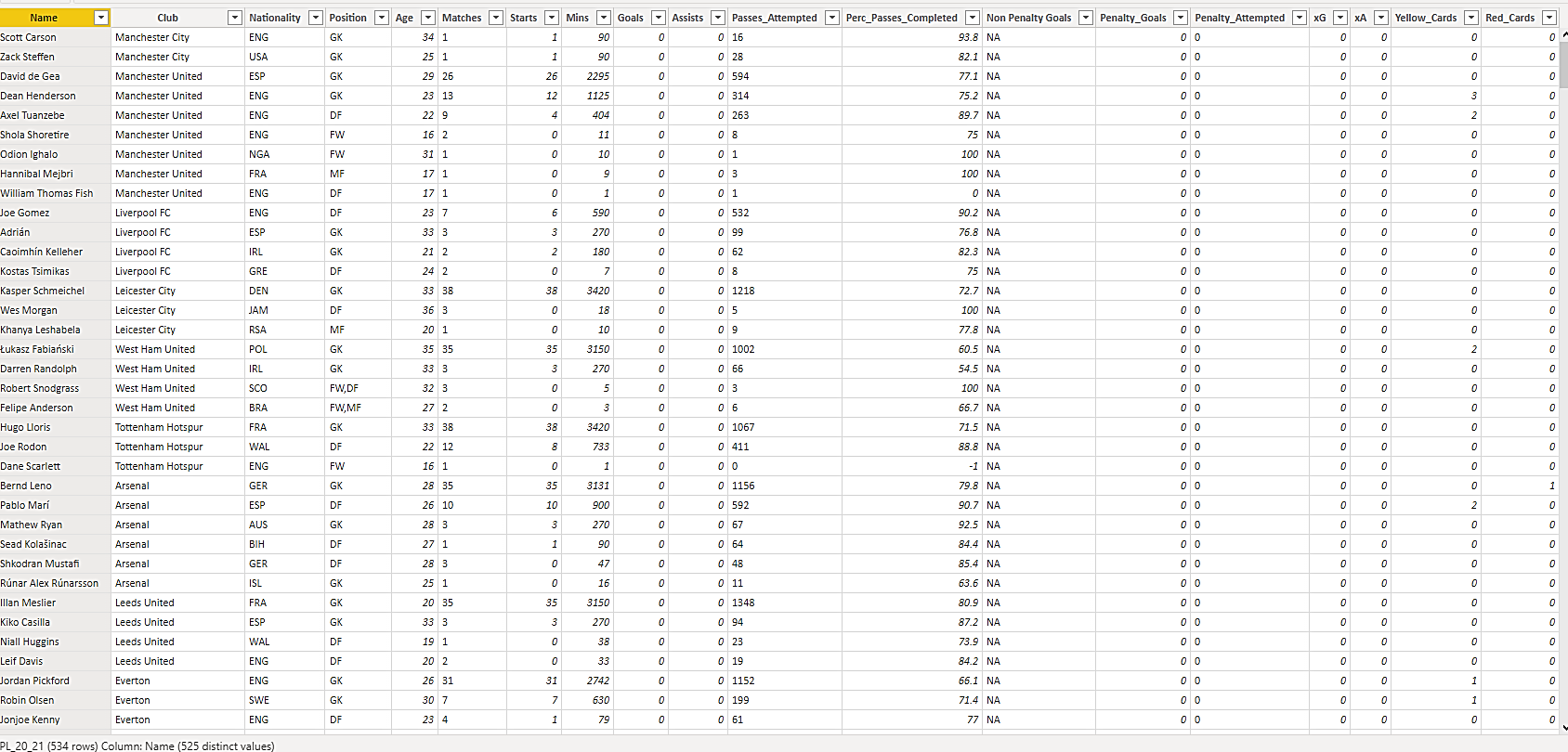


Figure 6. Screenshot of the EPL Dataset before pre-Processing

## OBJECTIVE, AIM, AND OVERVIEW OF BUSINESS INTELLIGENCE QUESTIONS

This project seeks to address several objectives using the dataset

The analysis is focused to address specific problems and ask the following questions by player, club, and position:

PLAYER, CLUB, POSITION

* To determine the average goals and assist per minute per player.
* Insight into the average xA and xG by position.
* Overview of Goals and penalty Goals by position.
* Provide an overview of the age of the player and the number of matches played.
* Comparison of Average Goals and Assist by position.
* Understanding of how these metrics affect club performance by comparing Average Assist, average goals, and average xA and xG by club.
* To offer a clear view of the metrics which affect player performance through AI-generated questions.
* And lastly, the number of red and yellow cards per player to determine the number of matches a player was suspended.

The power bi projects aim to answer these questions both analytically and visually to give a broader view of the performance of the EPL players and club and ways to improve performance.

# Analysis and Visualisation of the dataset

The visualization seeks to answer the questions in the findings above

## Insight into the influencer of a player by position

The dashboard below seeks to answer the question above

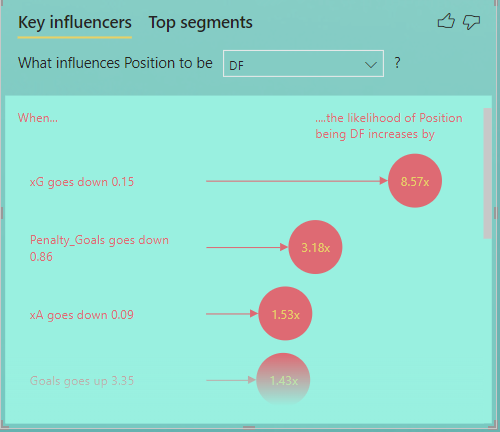
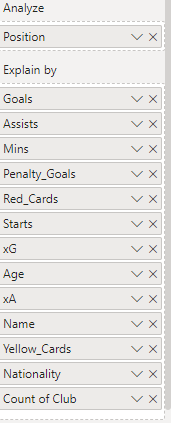


Figure 7. KEY INFLUENCER CHART

This chart is used to visualize the key influencers based on a player’s position based on the ‘Explain by section’ in the figure below.



This chart was created using the key influencer visual of Power BI. This chart was chosen because it showcases every single analysis visually in one chart based on the players' position

### Influencer and segment section by Defender

The graph shows that when xG goes down by 0.15, penalty Goals goes down by 0.86, xA goes down 0.09 and goals goes up by 3.35, Red card goes up 0.30, Starts goes up by 11.87, Yellow cards goes up 2.27, Age goes down 5.47 and assists go up by 2.10 the likelihood of the position being a defender increases by 8.57, 3.18, 1.53, 1.43, 1.29, 1.17, 1.10, 1.07, 1.04 respectively.

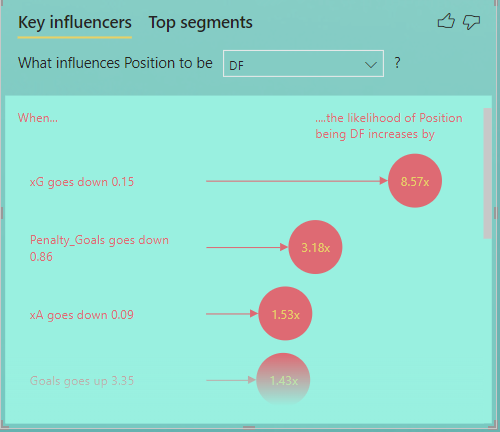


Figure 8. KEY INFLUENCER DF 1



Figure 9. KEY INFLUENCER DF 2

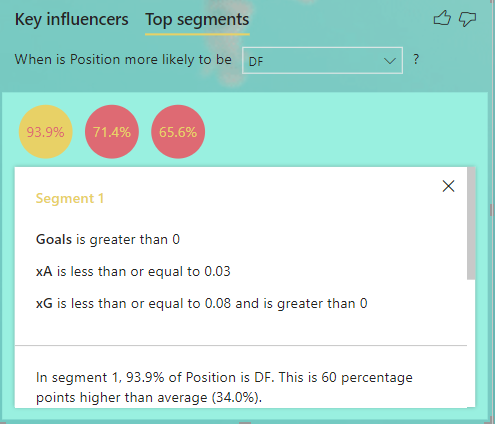


Figure 10. TOP SEGMENT DF 1

The size of the bubble represents the percentage of the position (93.9%) of Segment One as the Defender position. This is 60 percentage points higher than the average (34.0%). The population count is 49. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins.

With Segment 1. The goals of Defenders are greater than 0. The xA is less than or equal to 0.03 and xG is less than or equal to 0.08 and greater than 0

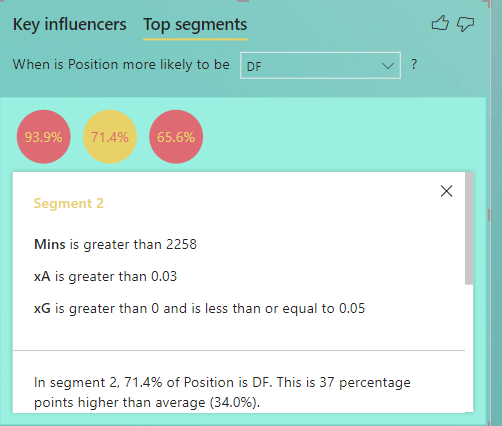


Figure 11. TOP SEGMENT DF 2

The size of the bubble represents the percentage of the position (71.4%) of Segment Two as Defender position. This is 37 percentage points higher than the average (34.0%). The population count is 28. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins

With Segment 2. Minutes of Defenders is greater than 2258. The xA is greater than 0.03 and xG is greater than 0 and is less than equal to 0.05

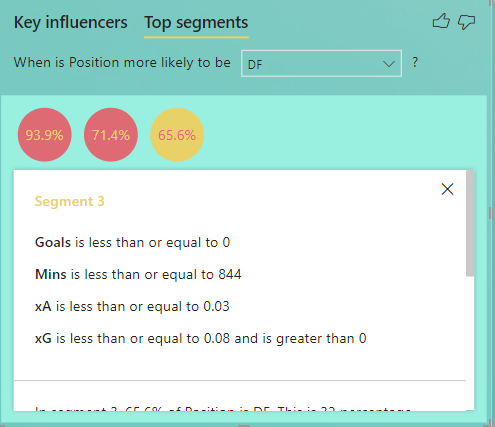


Figure 12. TOP SEGMENT DF 3

The size of the bubble represents the percentage of the position (65.6%) of Segment Three as Defender position. This is 32 percentage points higher than the average (34.0%). The population count is 32. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins

With Segment 3. Goals are less than or equal to 0. Minutes are less than or equal to 844. xA is less than or equal to 0.03 and Xg is less than or equal to 0.08 and greater than 0.

### Influencer and segment section by Forward

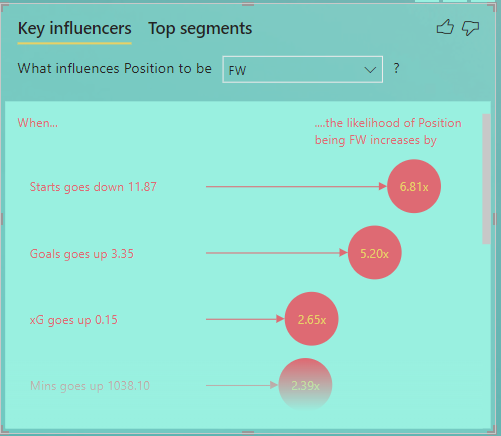


Figure 13. KEY INFLUENCER FW 1

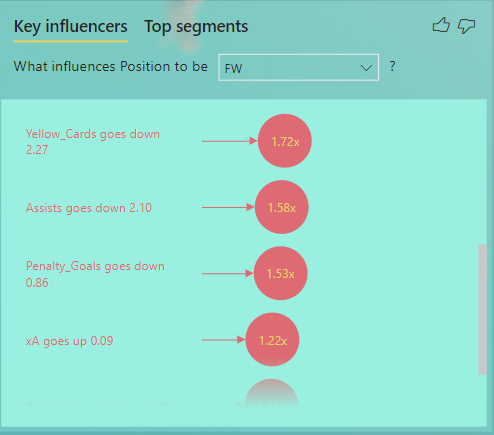


Figure 14. KEY INFLUENCER FW 1

The graph shows that when starts goes down by 11.87, Goals goes up by 3.35, xG goes up by 0.15 and Minutes goes up by 1038.10, Yellow card goes down 2.27, Assists goes down by 2.10, Penalty Goals goes down by 0.86, xA goes up 0.09 and red cards go down 0.30 the likelihood of the position being a Forward increase by 6.81, 5.20, 2.65, 2.39, 1.72, 1.58, 1.53, 1.22, 1.18 respectively.

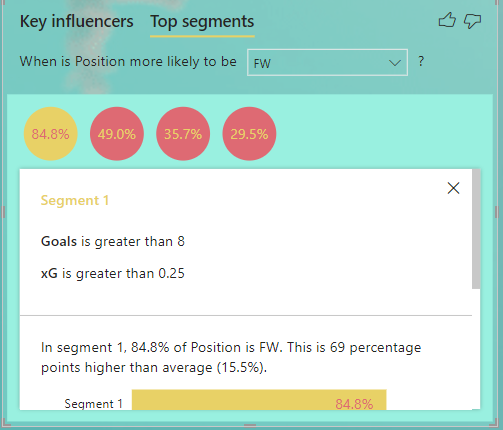


Figure 15. TOP SEGMENT FW 1

The size of the bubble represents the percentage of the position(84.8%) of Segment One in the Forward position. This is 69 percentage points higher than the average (15.5%). The population count is 33. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins

With Segment 1. Goals are greater than 8. xG is greater than 0.25.

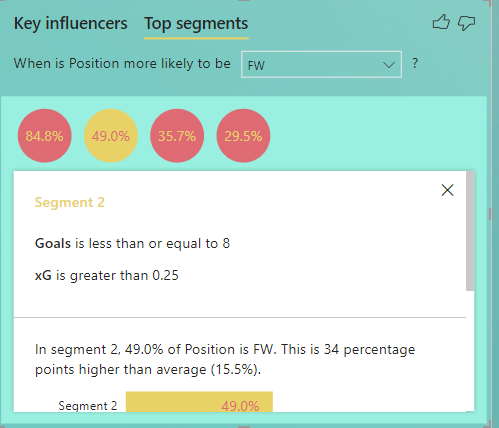


Figure 16. TOP SEGMENT FW 2

The size of the bubble represents the percentage of the position(49.0%) of Segment two is the Forward position. This is 34 percentage points higher than the average (15.5%). The population count is 49. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins.

With Segment 2. Goals are less than or equal to 8. xG is greater than 0.25.

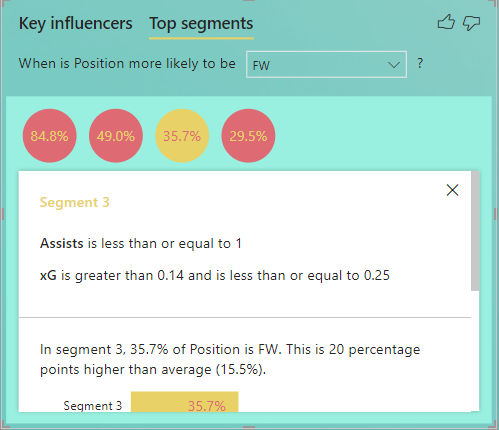


Figure 17. TOP SEGMENTS FW 3

The size of the bubble represents the percentage of the position (35.7%) of Segment three is the Forward position. This is 20 percentage points higher than the average (15.5%). The population count is 28. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins.

With Segment 3. Assists are less than or equal to 1. xG is greater than 0.14 and less than or equal to 0.25.

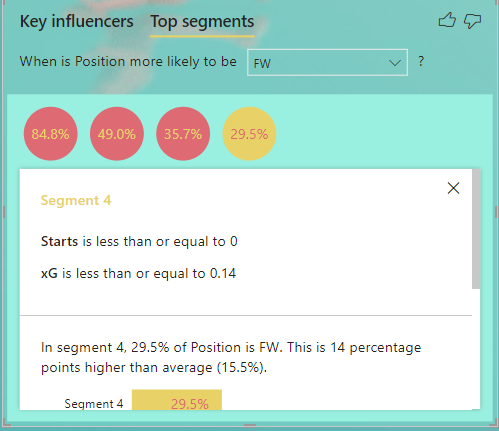


Figure 18. TOP SEGMENT FW 4

The size of the bubble represents the percentage of the position (29.5%) of Segment Four is the Foward position. This is 14 percentage points higher than the average (15.5%). The population count is 44. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins.

With Segment 3. Starts is less than or equal to 0. xG is less than or equal to 0.14

### Influencer and segment section by Goalkeepers

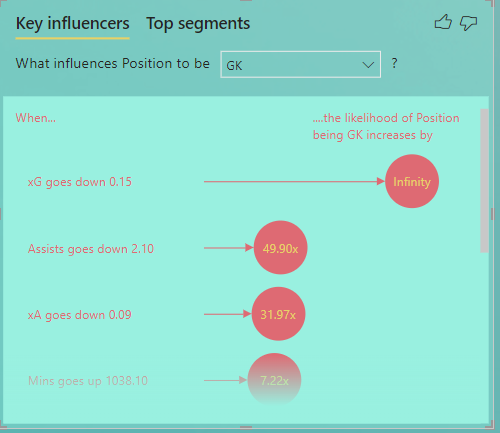


Figure 19. KEY INFLUENCER GK 1

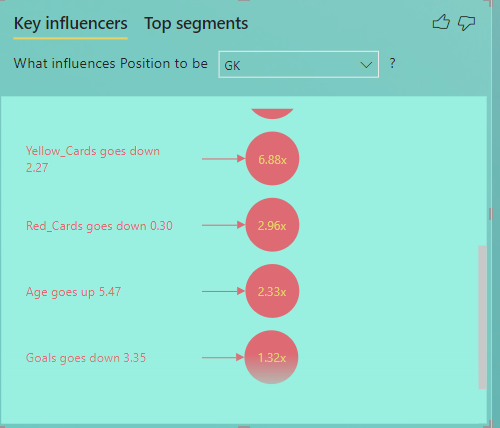


Figure 20. KEY INFLUENCERS BY GK 2

The graph shows that when xG goes down by 0.15, Assists goes down by 2.10, xA goes down 0.09 and Minutes goes up by 1038.10, Yellow card goes down 2.27, Red Cards goes down 0.30, Age goes up 5.47, Goals goes down 3.35 the likelihood of the position being a Goalkeeper increase by Infinity, 49.90, 31.97, 7.22, 6.88, 2.96, 2.33, 1.32 respectively.

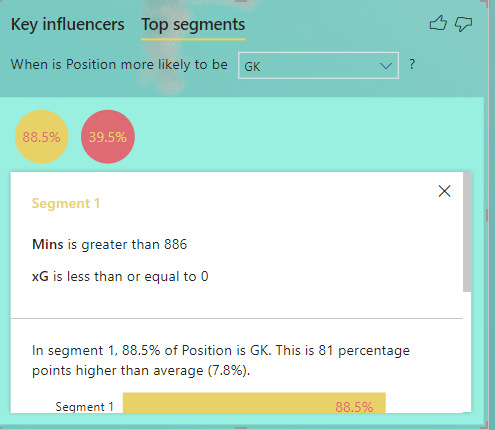


Figure 21. TOP SEGMENT GK 1

The size of the bubble represents the percentage of the position (88.5%) of Segment One is the Goalkeeping position. This is 81 percentage points higher than the average (7.8%). The population count is 26. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins

With Segment 1, Minutes is greater than 886. xG is less than or equal to 0.

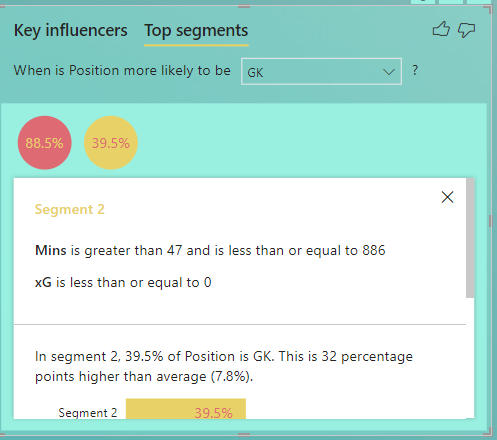


Figure 22. TOP SEGMENTS GK 2

The size of the bubble represents the percentage of the position (39.5%) of Segment Two is the Goalkeeping position. This is 32 percentage points higher than the average (7.8%). The population count is 43. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins.

With Segment 2. Minutes are greater than 47 and less than or equal to 886. xG is less than or equal to 0.

### Influencer and segment section by Midfielder

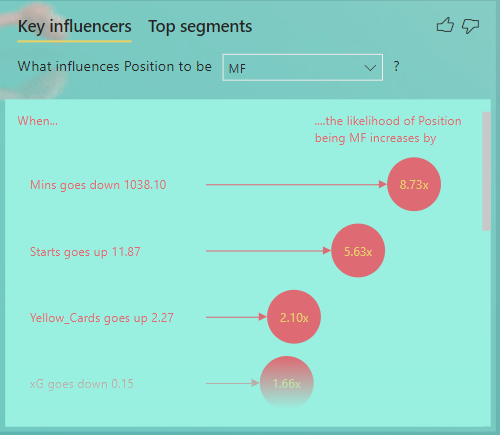


Figure 23. KEY INFLUENCERS MF 1

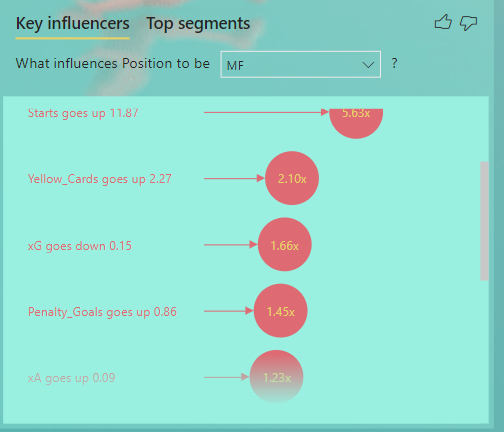


Figure 24. KEY INFLUENCERS MF 2



Figure 25. KEY INFLUENCER MF 3

The graph shows that when minutes is down by 1038.10, starts goes up by 11.87, Yellow cards go up 2.27 and xG goes down by 0.15, Penalty Goals go up 0.86, xA goes up 0.09, Red cards go down 0.30, Assists goes up 2.10, Goals goes down 3.35, Age goes down 5.47 the likelihood of the position being a Midfielder increase by 8.73, 5.63, 2.10, 1.66, 1.45, 1.23, 1.15, 1.12, 1.09, 1.07 respectively.

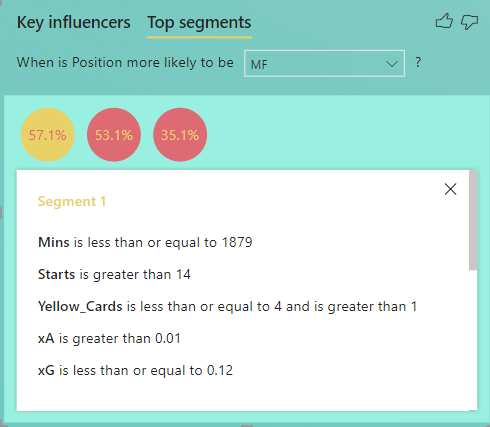


Figure 26. TOP SEGMENTS MF 1

The size of the bubble represents the percentage of the position(57.1%) of Segment One is the Midfield position. This is 37 percentage points higher than the average (20%). The population count is 28. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins.

With Segment 1. Minutes is less than or equal to 1879, Starts is greater than 14, Yellow cards is less than or equal to 4 and is greater than 1, xA is greater than 0.01, xG is less than or equal to 0.12.

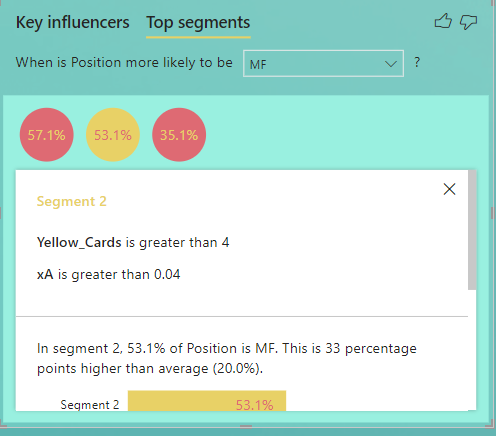


Figure 27. TOP SEGMENT MF 2

The size of the bubble represents the percentage of the position (53.1%) of Segment Two is the Midfield position. This is 33 percentage points higher than the average (20%). The population count is 49. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins

With Segment 2. Yellow cards are greater than 4 and xA is greater than 0.04.

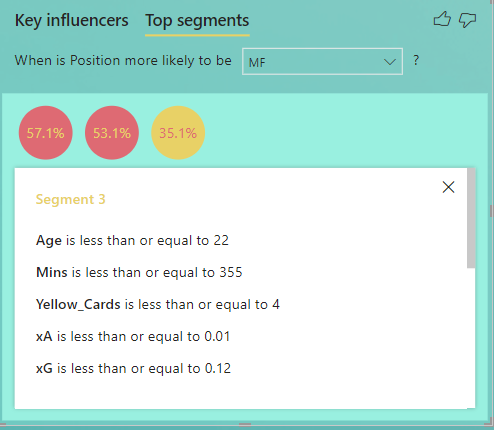


Figure 28. TOP SEGMENTS MF 3

The size of the bubble represents the percentage of the position (35.1%) of Segment Three is the Midfield position. This is 15 percentage points higher than the average (20%). The population count is 37. The bigger the bubble the bigger the percentage of the population of the dataset. The height represents the average position of that segment. The AI suggests the different segments and bins.

With Segment 3. Age is less than or equal to 22, Minutes is less than or equal to 355, Yellow cards is less than or equal to 4, xA is less than or equal to 0.01, xG is less than or equal to 0.12.

## Overview of Goals and penalty Goals by position

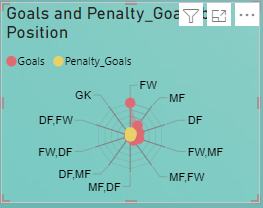


Figure 29. GOALS AND PENALTY GOALS BY POSITION

This chart was created using the radar chart visual of Power BI derived from ‘Get more visuals’ This chart was chosen because it showcases every player's position at a go.

It was created by placing Position in the category column and Goals and Penalty goals in the Y-axis column.

Penalties are very essential in the game of football. People may say penalties are a game of luck. This chart is meant to prove them wrong by showcasing the positions who are good penalty takers for the likelihood of scoring a goal hence winning a match.

The graph showed that Forwards have a high chance of converting or scoring a goal with 48 penalties and 446 non-penalty goals, followed by Midfielders with 36 penalty goals and 171 Goals, followed by defenders with 129 goals and 1 penalty goal. This shows that defenders are not prolific penalty takers hence the best option for any club is to make their centre forward take penalties with a midfielder next in line with the least likely being a defender.

## Provide an overview of the average age of players and the number of matches played.

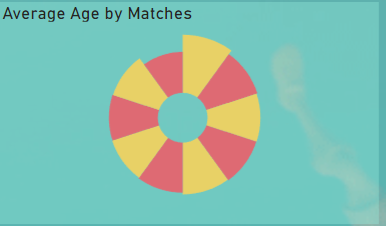


Figure 30.AVERAGE AGE BY MATCHES

This chart was created using the aster plot visual of Power BI derived from ‘Get more visuals.’ This chart was chosen because of its simplicity and its ability to simplify the data visually.

It was created by placing Matches in the category column and average age on the Y-axis.

The average age was calculated using a measure in DAX. This measure calculates the average age of the dataset. Since it is the average of the whole data a calculated column cannot be used. This is done using the formula below ;



and comparing it by ‘Matches’ in the EPL\_20\_21 table

The graph showcased that the average age had the highest number of matches played. This showcases that during a player's prime the player will be given more playing time since the player will likely be having a balance of youth and experience.

## Along with the average minutes and average age by position

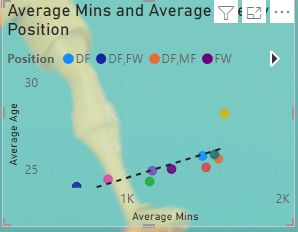


Figure 31. AVERAGE MINS AND AGE BY POSITION

This chart was created using the scatter plot along with a trend line to know the type of correlation of the graph.

It was created by placing Average minutes which was created using a measure in DAX. This measure calculates the average minutes of the dataset. Since it is the average of the whole data a calculated column cannot be used. This is done using the formula below ;



and comparing it by 

by position.

## Understanding of how these metrics affect club performance by comparing Average Assists, average goals, and average xA and xG by club

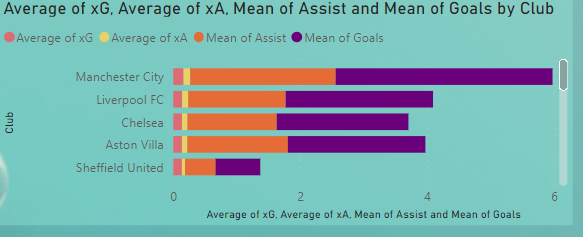


Figure 32. AVR xG,xA, Mean of goals and assist by club

This chart was created using the stacked bar chat. This chart was chosen because of its simplicity and its ability to simplify the data visually.

It was created by placing Club in the x-axis column and Average xG, Average xA, Mean of Assist, and Mean of Goals in the Y-axis

DAX was used to create the Mean of assists and Mean of Goals. This measure calculates the average goals and assists of the dataset. Since it is the average of the whole data a calculated column cannot be used. This is done using the formula below ;





The top 3 teams league position is reflective of the importance of average goals, assists, xA, and Xg in team performance.

# SUMMARY – Finding based on analysis and evaluation

A Power BI Dashboard was made to respond to the main questions in the report. The important conclusions are outlined below.

* It was deduced that the average age of 26.8 had the highest number of matches. It was then deduced that most players get more playing time when they get to mid age.
* Forwards had the highest average of xA and xG with the least being goalkeepers.
* The outlier was the Goalkeeper position which had the highest average Minutes of 1,628.50 and Average age of 28.19. This shows how reliable goalkeepers are.
* The top 3 clubs with the highest of these metrics finished in the top 3 positions in the 2020/21 EPL season. That shows how these metrics are significant for club position at the end of a season.
* The highest average age by the club was Crystal palace with Manchester United having the least.
* The highest number of Goals was scored by Manchester City who won the league that season and the least number of Goals were scored by Sheffield United who got relegated by the end of that season. This shows the importance of Goals in earning points which then determine the club standings at the end of the season.
* Forwards scored the highest number of Goals and Penalty Goals with the least being Goalkeepers with no Goals.
* According to Infographic, there is a total of 986 goals scored in the 2020/21 EPL season.
* The same procedure was used for the infographic visual for the number of assists. Assist was put in the measured field. There was a total of 685 assists in the EPL.
* According to Infographic, there is a total of 750K plus minutes played in the EPL

Recommendation

# Recommendations

* A different metric other than xA and xG should be used to rate the goalkeeper's performance because xG and xA are reflective of outfield player's position.
* More effort should be put in place to increase the average goals of midfielders.
* Clubs should balance between being defensive and offensive to get a high chance of winning the league.
* Offensive players should be efficient and ruthless in front of goal and making good use and being aware of their positioning in a tactical setup to counteract the effect of the opposition thereby increasing their xG and xA.
* Even though the league is progressive, there is a need for improvement in different positions in the various clubs to make it more competitive.

# Conclusion

I knew very little about Power BI before this project, but it has enabled me to advance both my data analytics and Power BI skills. The problems I faced while working on this project ranged from selecting the appropriate data set to data cleaning/pre-processing to the actual analysis, but I was able to get through them with the support of my lab tutor Magda and by conducting my research and learning.

I am confident that Power Bi will help resolve issues throughout my career, I am looking forward to using the skills I have learned in this module to real-world obstacles.

As a result, I was particularly interested in several of the features of Power BI, such as the reusable queries for data cleaning the increased amount of dataset rows supported by Power BI, and an add-in for artificial intelligence and machine learning tools that uses decision trees and Q & A.

# APPENDIX

# DATA PRE-PROCESSING

## Loading Data

Firstly, the dataset was loaded via ‘Get Data’ in the data section on the Power bi-home tab. Different options are on the type of file to be loaded into Power bi. The file for this project is in a CSV format that is a comma-separated file hence Text/CSV option was selected, The EPL\_2020/21.csv file was then loaded.

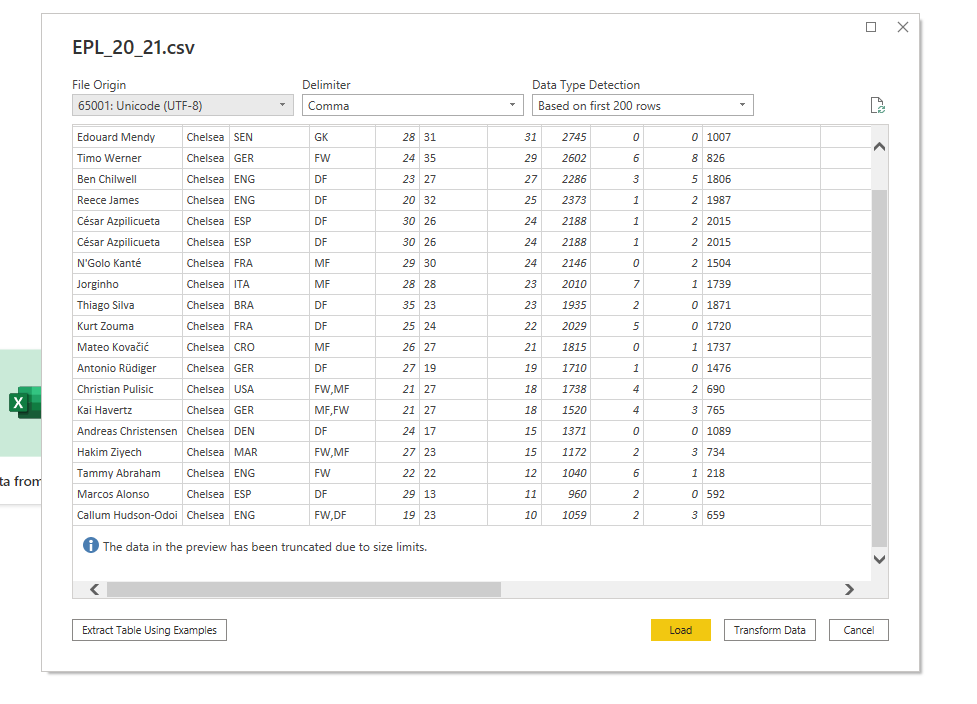


Figure 33. EPL FILE BEING LOADED

Fig.2 Loaded CSV file

## Data Cleaning

To perform the data cleaning steps the transform data at the queries section of the home tab was selected and opened the Power Query editor.

Removing Columns

The dataset contained a column named Referees which showcased the name of referees who officiated during the season. The column was not needed for our analysis hence was removed.



Figure 34. COLUMN REMOVAL

### Removing blanks

The dataset contains a few blanks. The percentage of blanks and NAs were checked on the view button on the home tab of the query editor under the checking quality box. Checks the percentage of empty rows in each column of the dataset. Most columns had less than 9 percent of blank rows hence are removed since it would not affect the result and visualization of the data.

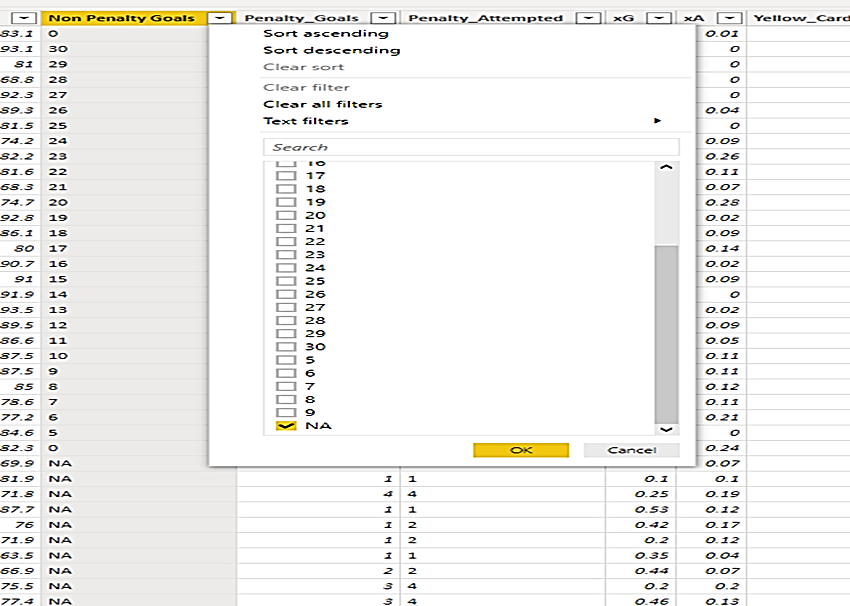


Figure 35. REMOVING BLANKS

### Replacing Misspelt Names/Values

The players' column had wrongly spelled players' names with signs and symbols. The Replace Value option in Power Query is used to replace the name.

In this column, Cesar Azpilicueta was spelled as C@esar\_Azpilicueta. The column is highlighted, then on the home tab, Transform is chosen. Under this, the Replace Values dropdown is selected, and 'Replace Values' is selected. This pops up a dialog box as shown.

In the ‘Values to Find’ field, C@esar\_Azpilicueta was typed in and in the ‘Replace With’ field, Cesar Azpilicueta was typed in. This replaces all the wrongly spelled fields. This was done for other misspelled names in the players' column.

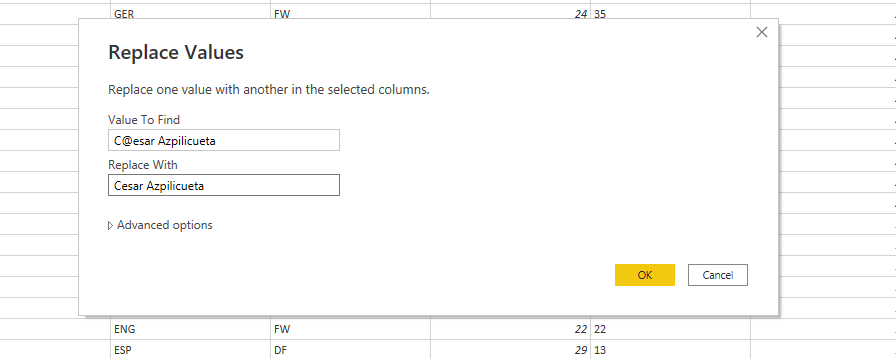


Figure 36. REPLACING MISSPELLED NAMES

### RENAMING COLUMNS

A few columns were renamed. This is done by double-clicking on the header and typing the new name.

The table shows the before and after pre-processing steps.

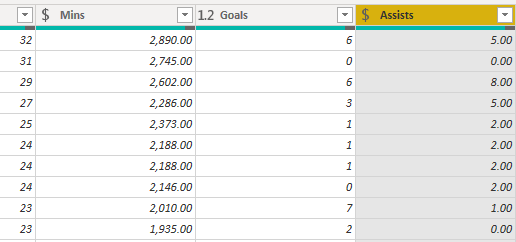


Figure 37.Shows a screenshot of the unit of the column which needs to be changed

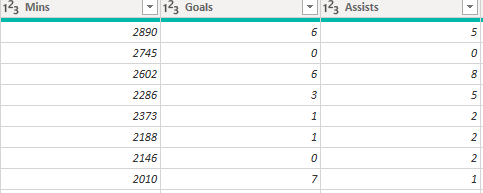


Figure 38.Screenshot of changed unit

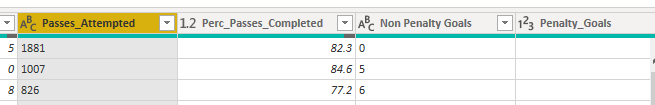


Figure 39.Screenshot before column renaming

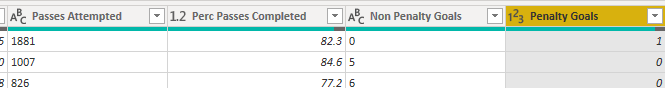


Figure 40.Screenshot after renaming a column

# DATA MODELLING

The table below shows what the data model looks like after pre-processing (One table containing all the information)

The table is then divided into a fact and dimension table. A fact table is quantitative, and a dimension table is qualitative.

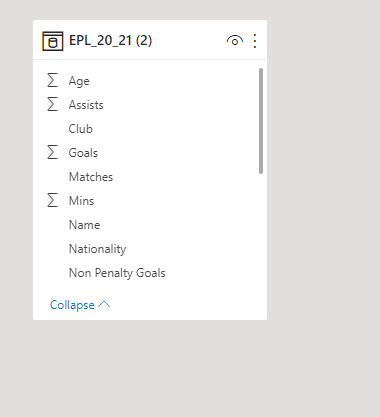


Figure 41.Screenshot of the table before creating the Dimension and Fact table

## Dimension Tables

A two-dimension table was created for the model. This was done by duplicating the original dataset (EPL 2020/21.csv). This duplicated dataset is renamed ‘CLUB’.

On the CLUB duplicate table, all other columns are removed except club. Since the club has duplicate Club names. Remove duplicate was used to remove multiple club names. Next, an index or ID column was created by clicking ‘Add column’ on the home tab and clicking on ‘index' and select from 1. The new column is renamed Club ID.

On the second duplicate table renamed as Player all other columns are removed except Name, Nationality, player id, and position. Next, an index or ID column was created by clicking 'Add column' on the home tab and clicking on 'index' and select from 1. The new column is renamed player ID.

Columns with the same names on the original table that is EPL 2020/21 as that of the columns on the Dimension table are then removed creating a fact table before merging.

## Creating Relationship

The index or ID columns are then added to the fact table to create autodetect relationship or by clicking on the merge queries on the home tab and connecting the tables on their shared columns to create a relationship.

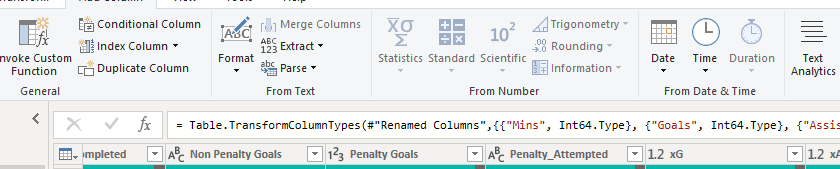


Figure 42.To create a relationship an id has to be created using the index column



Figure 43.IDs linking the fact table to the dimension table

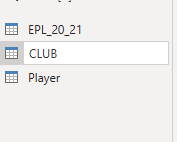


Figure 44.Shows the duplicated datasets used for the fact and dimension table

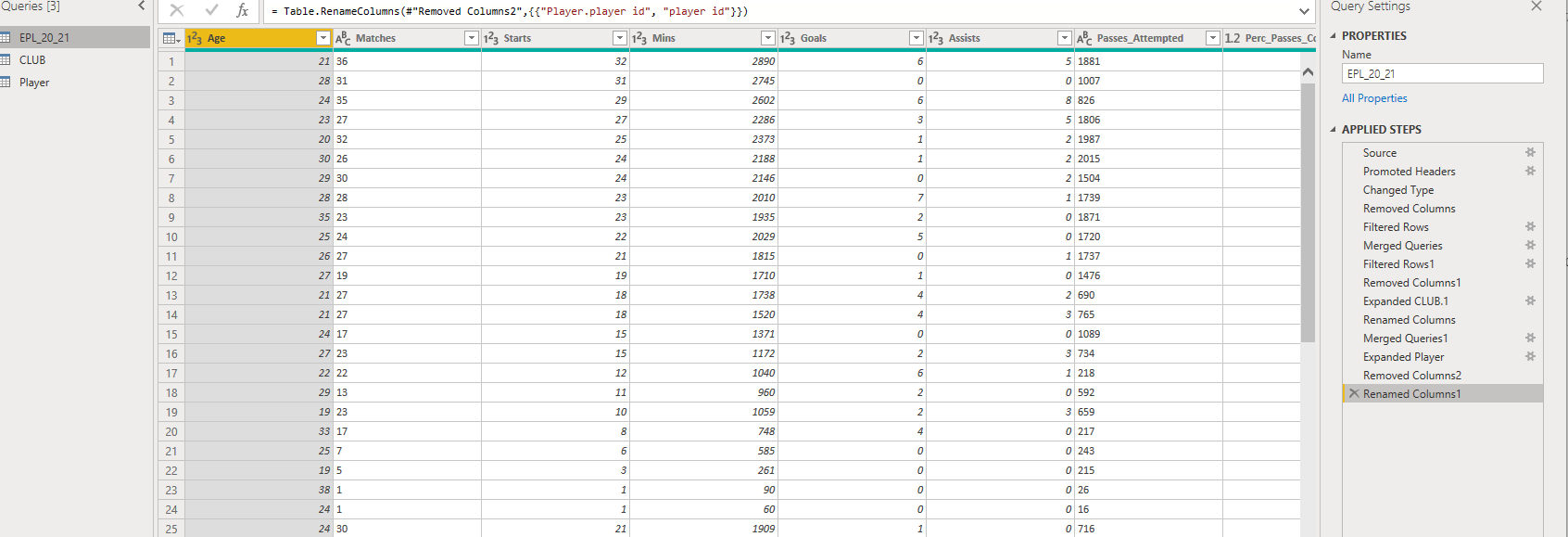


Figure 45. FACTS DATA



Figure 46. CLUB DIMENSION DATA 1

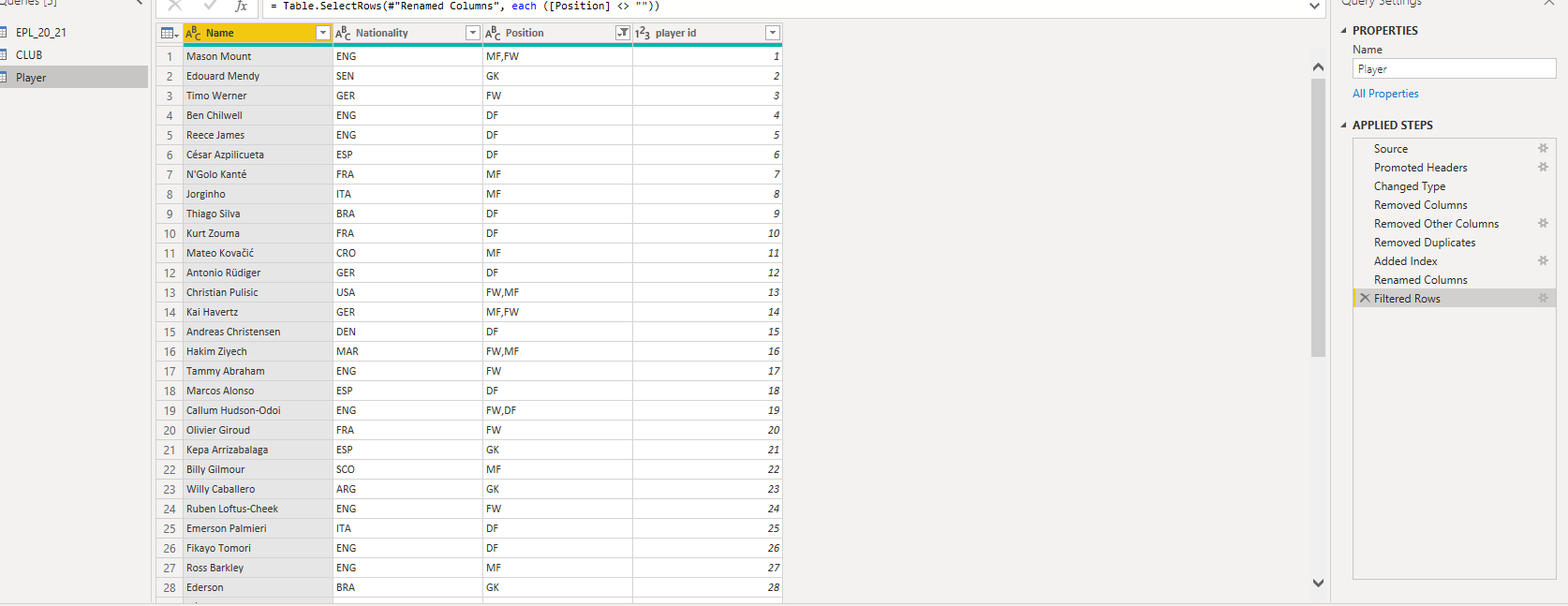


Figure 47. PLAYER DIMENSION DATA 2



Figure 48. CREATING RELATIONSHIP 1

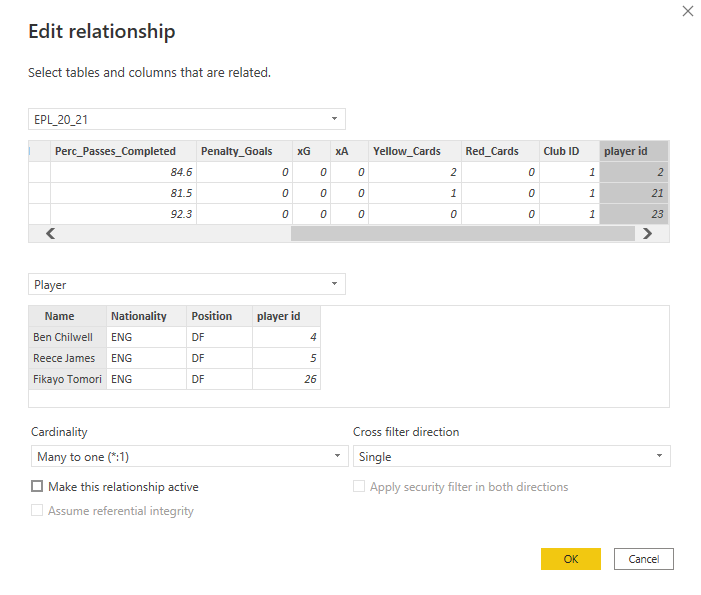


Figure 49. CREATING RELATIONSHIP 2

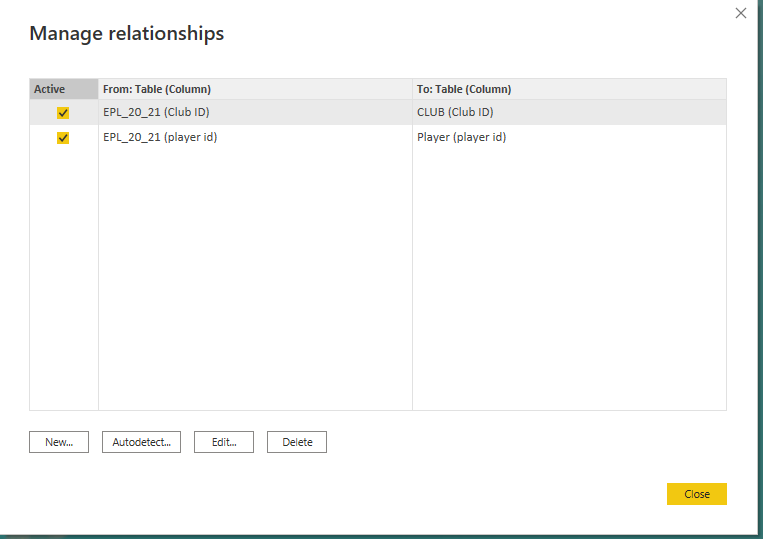


Figure 50.Autodetect CAN ALSO BE USED to build a relationship between the fact and dimension TABLE

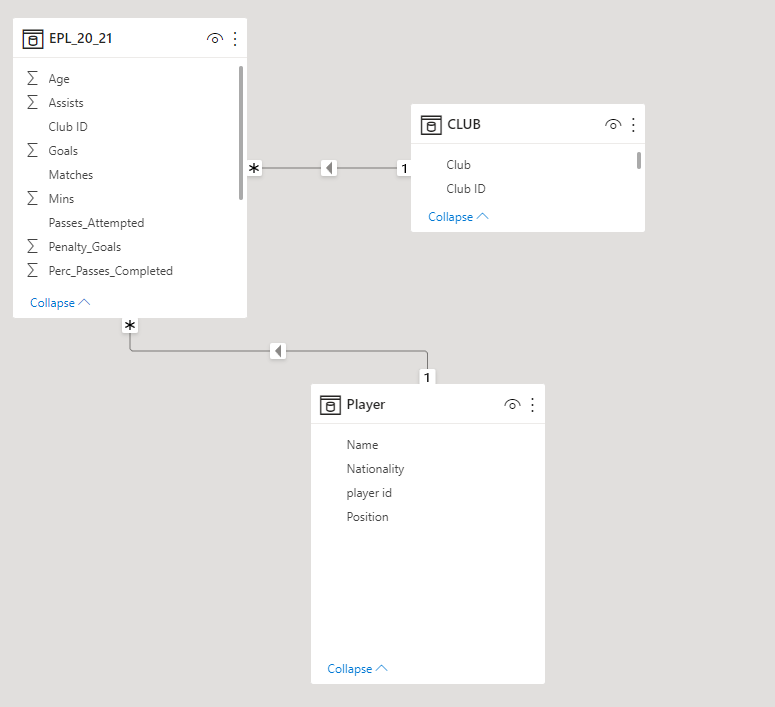


Figure 51. The relationship created between the fact and dimension table (Star Schema model)

# INTRODUCTION

Football, as we all know, is the world-leading sport. This report seeks to know how the following metrics like Goals and assists affect the xG and xA of a player and club and its relation to the player's position.

This metric is a very recent metric used to give a broader perspective on a player and club's performance and ways to improve upon it. Even though it is not a true reflection of the general performance of a team or player it gives a sense of player analysis.

Expected Goals as we know as xG is the average likelihood that the player scored from a particular position on the field. It ranges from 0 to 1. Zero represents impossible to score and one represents the expected to score by the player.

Whiles Expected Assist (xA) is the average quality of passes and creative productive output via chances created from open play leading to scoring opportunities (it ranges from 0 to 1). Zero represents a pass leading to a goal that will not result in a goalscoring assist and one represents an assist successful and results in a goal from every event by a player.

Even though this metric is discriminatory to Goalkeepers I believe soon there will be a metric used especially for goalkeepers to rate their performance.

It is expected that this analysis will give insights into the minutes, matches, goals, assist, xG, and xA with relationship to age and position and ways to improve upon this performance.

The analysis seeks to answer the following:

* Insight into the influencer of a player by position
* Overview of Goals and penalty Goals by position.
* Provide an overview of the age of the player and the number of matches played.
* Comparison of Average Goals and Assist by position
* Understanding of how these metrics affect club performance by comparing Goals by the club, average goals, and average Xa and xG by club

# Dataset

The dataset was sourced from Kaggle an online open-source data library via https://www.kaggle.com/datasets/rajatrc1705/english-premier-league202021 A full description of the dataset is in the data source and description section above.

# Data Model

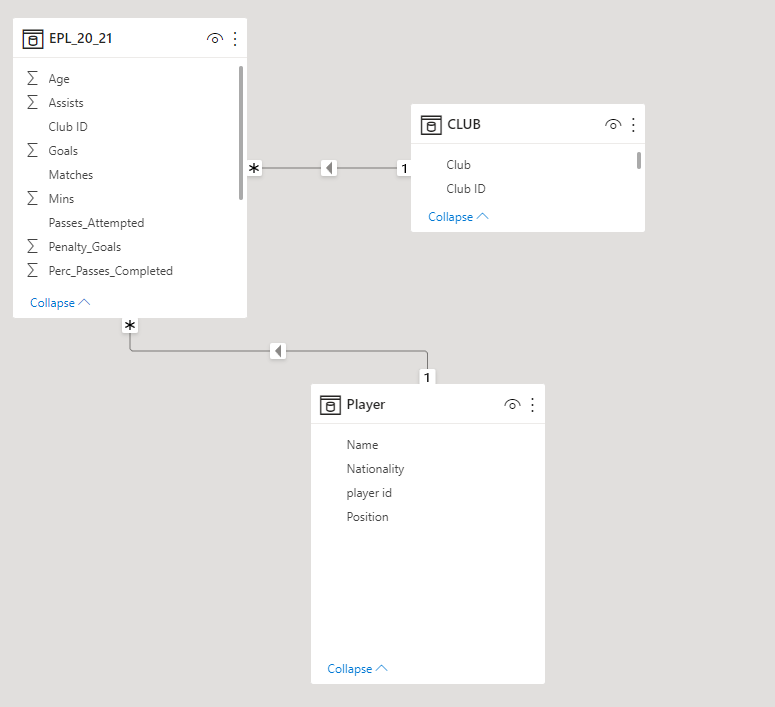


Figure 52. DATA MODEL

The figure above shows the dataset normalized into fact and dimension table

One fact table called EPL 20\_21- and two-dimension tables called ‘CLUB’ and ‘player’.

# Dax and m language

DAX and M language was used in my report. The DAX was used to create measures which were then used to create visuals.

Here is a list of DAX measures used since our data does not make room for calculated columns.

Mean of xA = AVERAGE(EPL\_20\_21[Assists])

Mean of xG = AVERAGE(EPL\_20\_21[Goals])

Average Age = CALCULATE(AVERAGE('EPL\_20\_21'[Age]), FILTER('EPL\_20\_21', 'EPL\_20\_21'[Age] ))

Average Assists = AVERAGE(EPL\_20\_21[Assists])

Average Goals = AVERAGE(EPL\_20\_21[Goals])

Average Matches = CALCULATE(AVERAGE('EPL\_20\_21'[Matches]), FILTER('EPL\_20\_21', 'EPL\_20\_21'[Matches] <> 18 ))

Average Mins = AVERAGE(EPL\_20\_21[Mins])

Along with DAX I also used M language to create 2 columns to rate the players xA and xG performance using conditional column in the power query editor in addition to creating query to know the amount of suspension a player got using M language. All these can be found in the figures below.

let

suspension = Table.FromRecords({

[category= "Missed 2 matches", Yellow cards=10],

[category="Missed 1 match", Yellow cards=5]

})

in

suspension

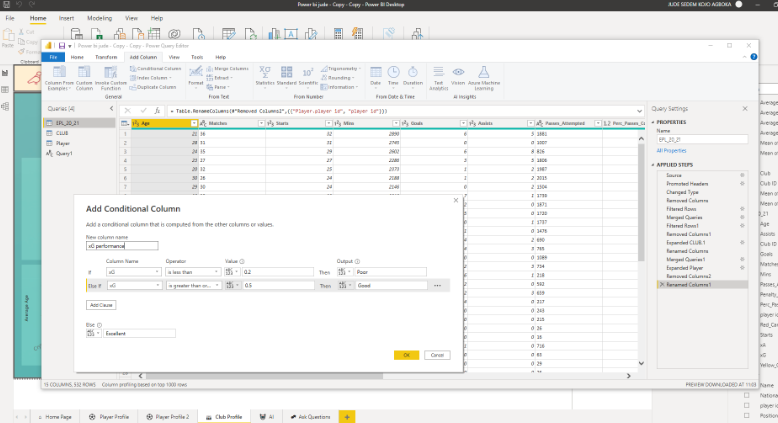


Figure . CREATING CONDITIONAL COLUMN FOR XG PERFORMANCE

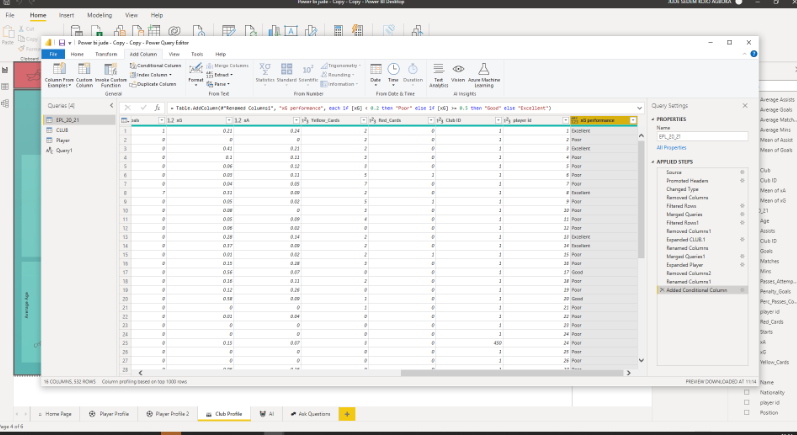


Figure . CONDITIONAL COLUMN FOR xG PERFORMANCE



Figure . CREATING CONDITIONAL COLUMN FOR xA PERFORMANCE

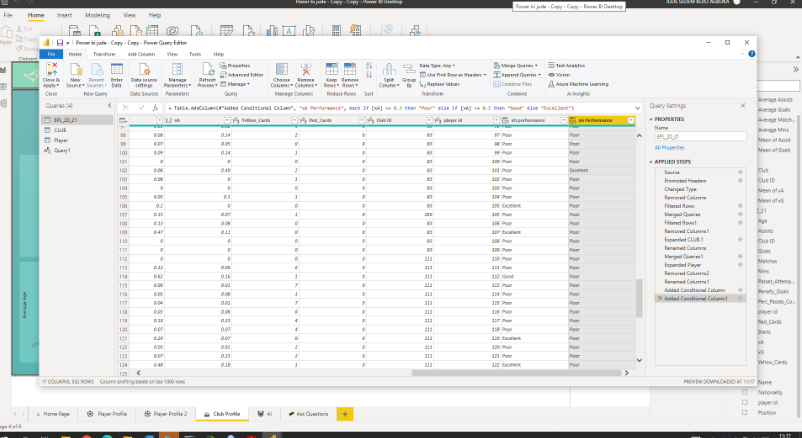


Figure . CONDITIONAL COLUMN FOR xA PERFORMANCE

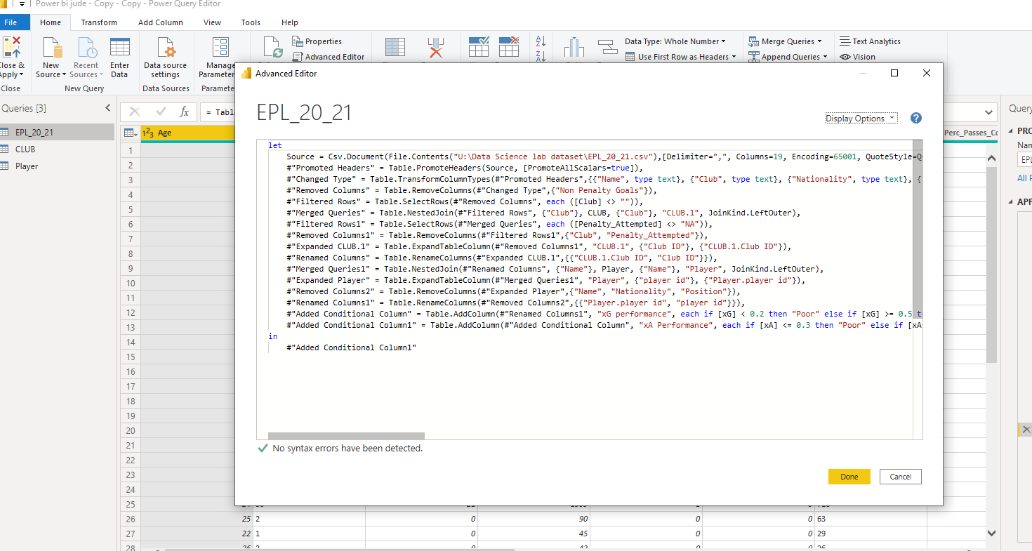


Figure . M LANGUAGE 1



Figure .M LANGUAGE 2

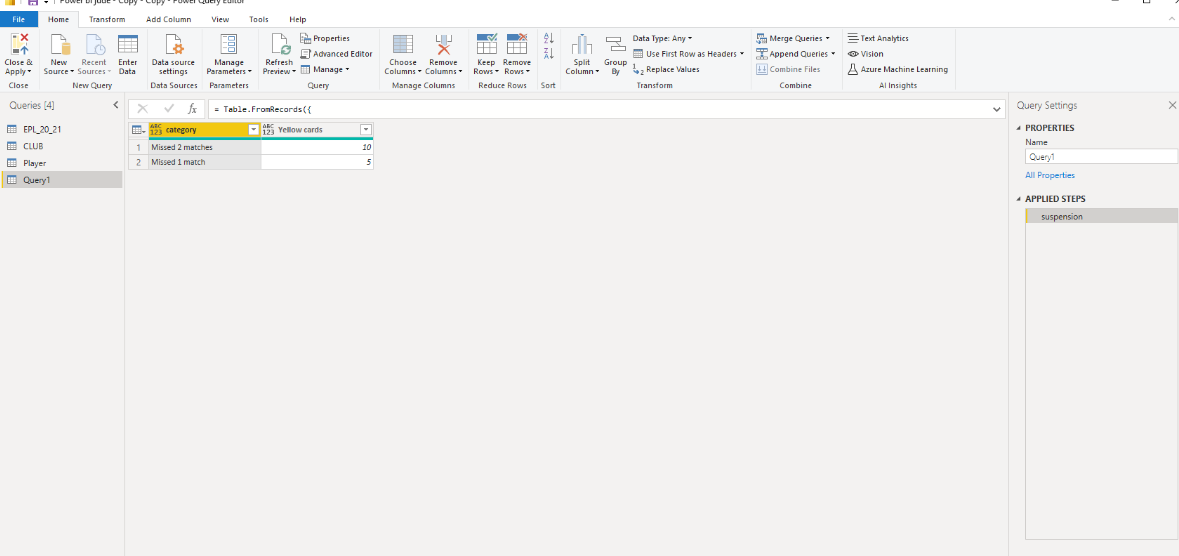


Figure . SUSPENSION QUERY

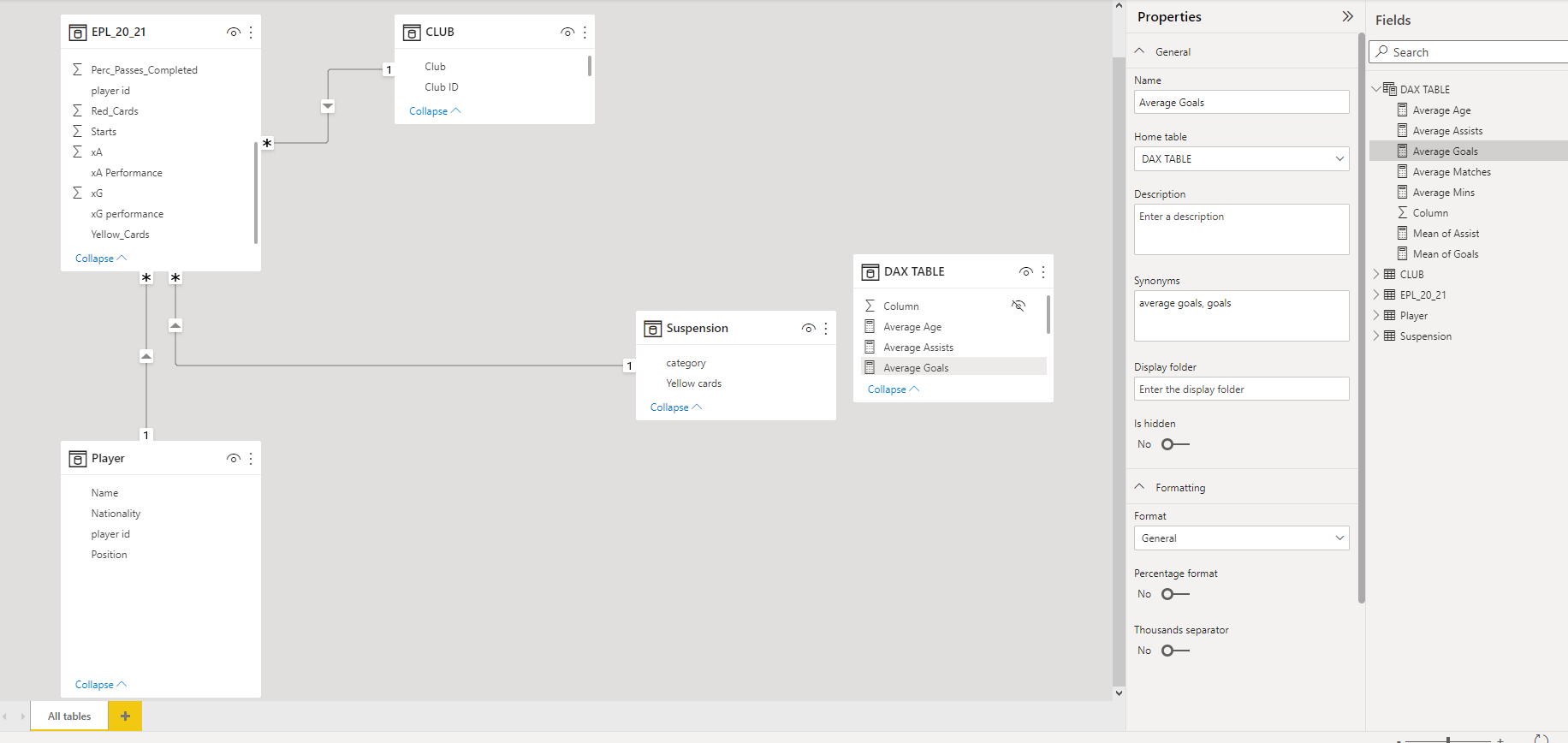


Figure . FINAL RELATIONSHIP OF TABLES

# Home page and navigation



Figure 61. Shows a screenshot of the homepage

You can quickly travel through the report pages by using the Homepage as a navigation page. These were created by using a navigation button on the element section of the insert tab, and further formatting was used to ensure that the navigation button is vertical. The navigation button automatically refreshes whenever a page is removed or added to the report.

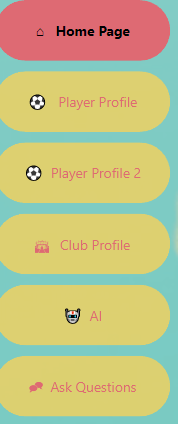


Figure 62.Shows the names of the buttons to access the pages

By using Ctrl + Right Clicking the button one can access the pages



Figure 63.Shows a navigation button on all the pages to help in the navigation of the pages

|  |  |
| --- | --- |
| Button | Description |
| 1st Button | Navigate to Home Page |
| 2nd Button | Navigate to Player Profile Page |
| 3rd Button | Navigate to Player Profile 2 Page |
| 4th Button | Navigate to club profile Page |
| 5th Button | Navigate to AI Page |
| 6th Button | Navigate to Ask Questions Page |

DASHBOARD



Figure . PLAYER PROFILE DASHBOARD 1



Figure .. PLAYER PROFILE DASHBOARD 2



Figure . CLUB PROFILE DASHBOARD

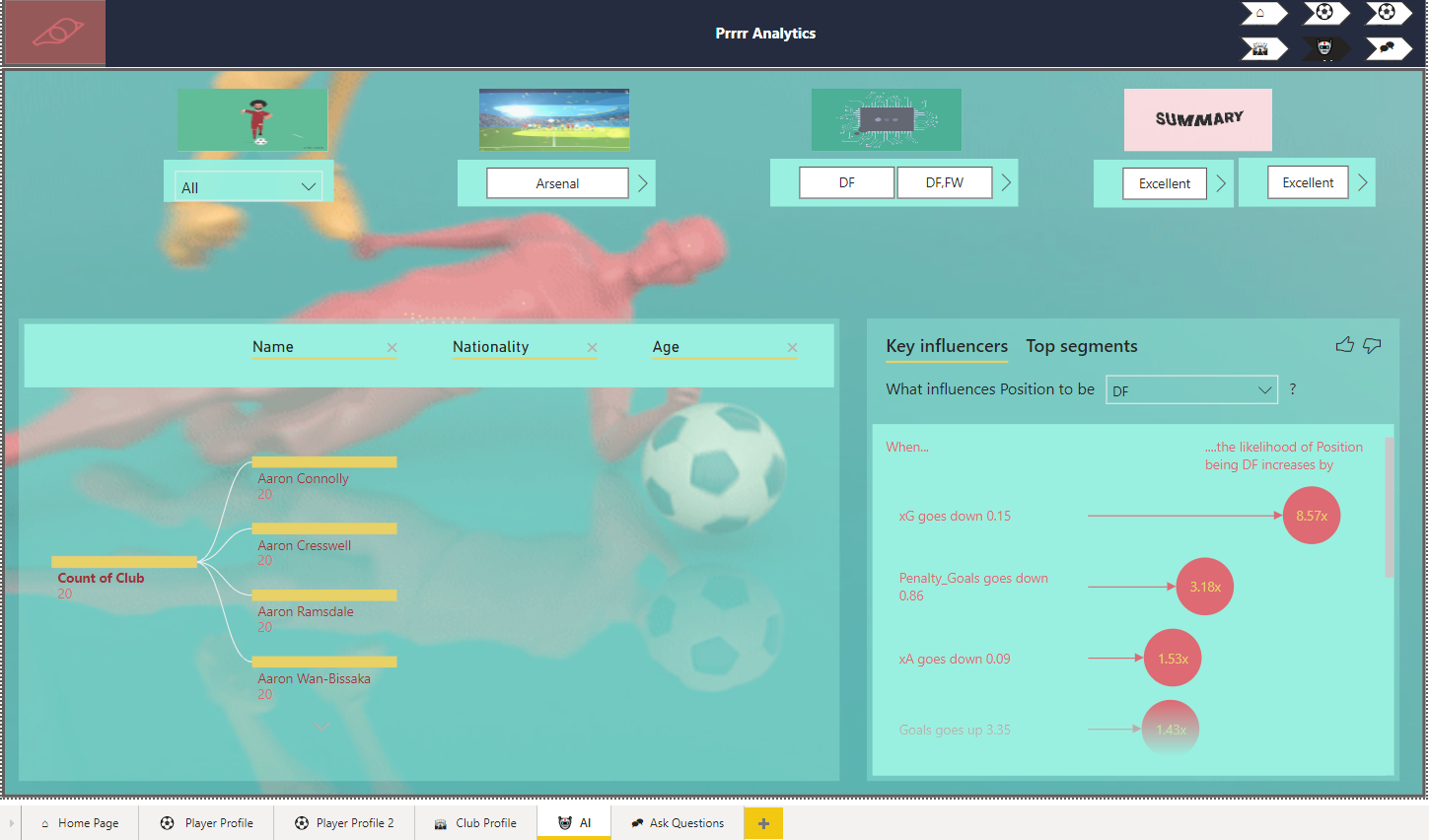


Figure . AI DASHBOARD

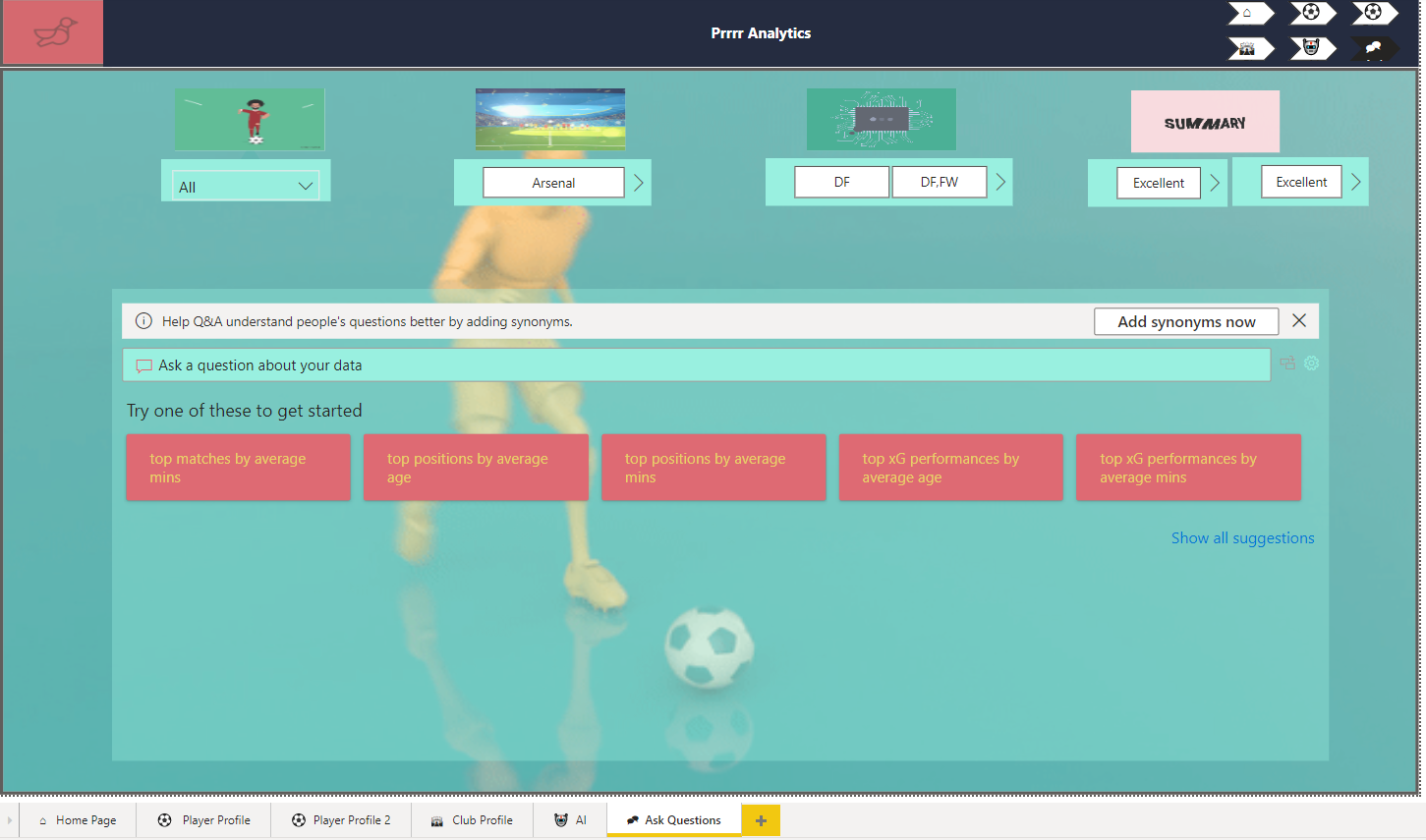


Figure . ASK QUESTIONS DASHBOARD

Charts, cards, and page navigation are all mixed in the dashboard. Additionally, two slicers have been introduced, allowing the user to slice the dashboard by player name, club, and position.

# Player Profile



Figure 69. MINUTES

The Infographic Designer visual was imported into this chart using the 'Get More Visuals' option. It was made using the Minutes measured in the measured field. According to Infographic, there is a total of 750K plus minutes played in the EPL.

The same visual was used to create goals and assist in Infographic design. It was made using the Goals measured in the measured field. According to Infographic, there is a total of 986 goals scored in the 2020/21 EPL season.

The same procedure was used for the infographic visual for the number of assists. Assist was put in the measured field. There was a total of 685 assists in the EPL.

Hover over the infographic to display the figure.

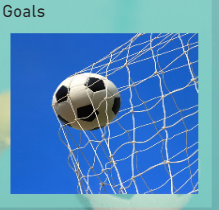


Figure 70. GOALS



Figure 71. ASSISTS

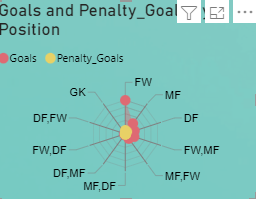


Figure 72. GOALS AND PENALTY GOALS BY POSITION

The chart was created using a Radar chart imported from ‘get more visuals’. This visual was used because it summarises Goals and Penalty Goals by position in a simple and precise visual format. This was done by placing Position in the category section and Goals and Assist on the y-axis. The Visual states that Forward score the highest number of Goals and Penalty Goals with the least being Goalkeepers with no Goals

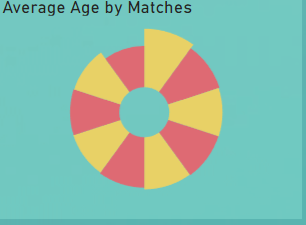


Figure 73. AVERAGE AGE BY MATCHES

The chart was created using an aster plot imported from ‘get more visuals’. DAX was used to calculate the Average using the formula below and placed on the Y-axis whiles Matches was placed in the category section to be plotted.



It was deduced that the average age of 26.8 had the highest number of matches. It was then deduced that most players get more playing time when they get to a mid-age.

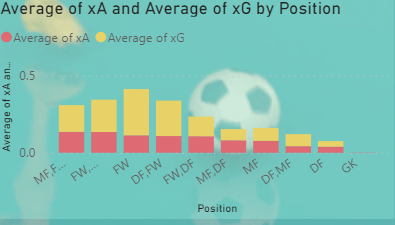


Figure 74. AVERAGE xA,xG, by position

The chart was created using a stacked column chart. This was done by placing Position on the x-axis and Average xA and xG on the y-axis. The average was created by clicking on the drop-down arrow on the Y-axis section and choosing Average from the popped-up box. Forwards had the highest average of Xa a and xG with the least being goalkeepers.

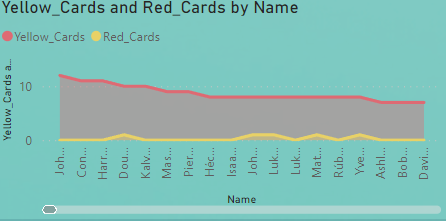


Figure 75. RED AND YELLOW CARDS BY POSITION

The chart was created using an area chart. This was done by placing Name on the x-axis and yellow cards and red cards on the y-axis. The highest number of cards was recorded by John McGinn.

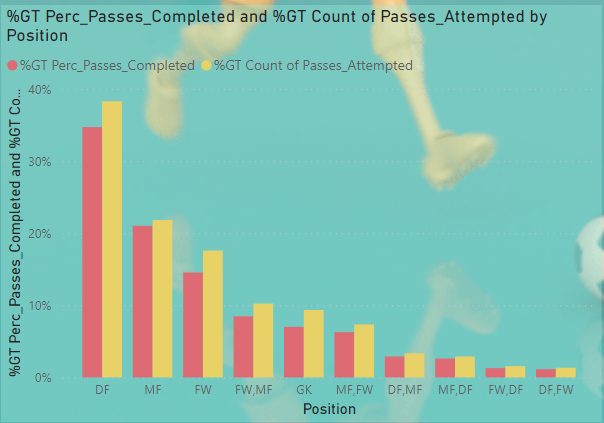


Figure 76. PERC PASSES COMPLETED & GT COUNT PASS ATTEMTD BY POSITION

The chart was created using a clustered column chart. This was done by placing Position on the x-axis and % GT Percentage of Passes completed and %GT Count of passes attempted on the y-axis. Defenders had the highest % GT Percentage of Passes completed and %GT Count of passes attempted followed by Midfielders etc

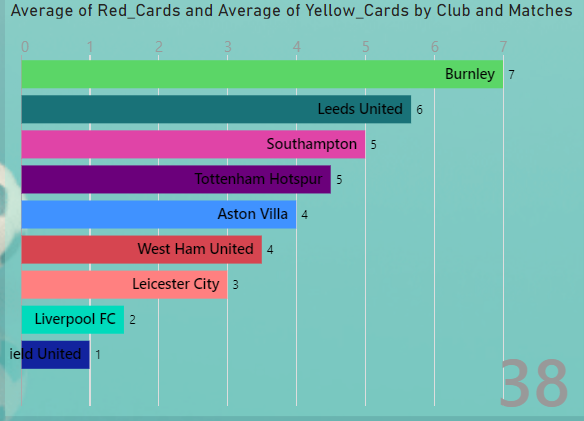


Figure 77. AVERAGE RED AND YELLOW CARDS BY CLUB AND MATCHES

The animated chart was created using an Animated bar chart imported from 'get more visuals’. This visual was used because it summarises Average Red and Yellow cards by club and number of matches (38 matches). This was done by placing the club in the name section with Average red and Average Yellow cards in the value section. The average was created by clicking on the drop-down arrow on the value section and choosing Average from the popped-up box. The animated chart states that Burnley had the highest average of cards followed by Leeds, Southampton, Tottenham, Aston Villa, West Ham, etc

 This card visual was created by placing average Goals in the Fields section

Figure 78. AVERAGE GOALS

The average Goals were created using the DAX formula below



This card visual was created by placing the average Assist in the Fields section

Figure 79. AVERAGE ASSISTS

The average Goals were created using the DAX formula below



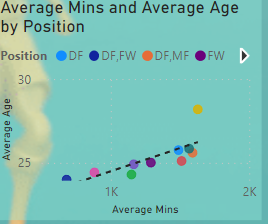


Figure 80. AVERAGE MINS & AVERAGE AGE BY POSITION

The chart was created using a scatter chart. This was done by placing Average minutes on the x-axis, Average Age on the y-axis and Position in the Legend section. A trendline was added to show the correlation. It showed a positive correlation between Average Age and Average Minutes by position. The outlier was the Goalkeeper position which had the highest average Minutes of 1,628.50 and Average age of 28.19. This shows how reliable goalkeepers are.

The DAX formula for average age is below;



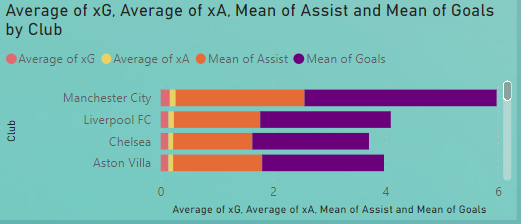


Figure 81. AVERAGE xG, xA, Mean of gOALS BY CLUB

The chart was created using a stacked bar chart. This was done by placing the club on the x-axis and Average xG, Average xA, Mean of Assists, and Mean of Goals on the y-axis. The top 3 clubs with the highest of these metrics finished in the top 3 positions in the 2020/21 EPL season. That shows how these metrics are significant for club position at the end of a season.

# Club Profile



Figure 82. AVERAGE AGE BY CLUB

The chart was created using a line chart. This was done by placing the club on the x-axis and the Average Age on the y-axis. The highest average age by the club was Crystal palace with Manchester United having the least.

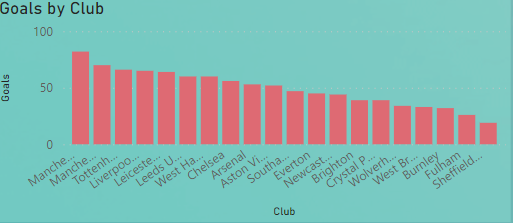


Figure 83. GOALS BY CLUB

The chart was created using a stacked column chart. This was done by placing Club on the x-axis and Goals on the y-axis. The highest number of Goals was scored by Manchester City who won the league that season and the least amount of Goals were scored by Sheffield United who got relegated by the end of the season. This shows the importance of Goals in ending points which then determine the club standings at the end of the season.

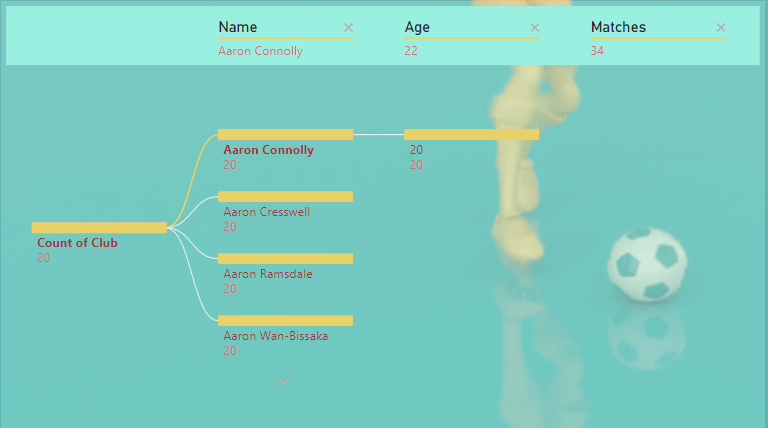


Figure 84. DECISION TREE

The chart was created using a Decomposition Tree. This was done by placing the club in the Analyse section and Name, Age, Matches, Red cards, yellow cards, Goals, Penalty Goals, xA, xG, and Assist in the Explain section. This explains everything in a simple format based on these metrics given.

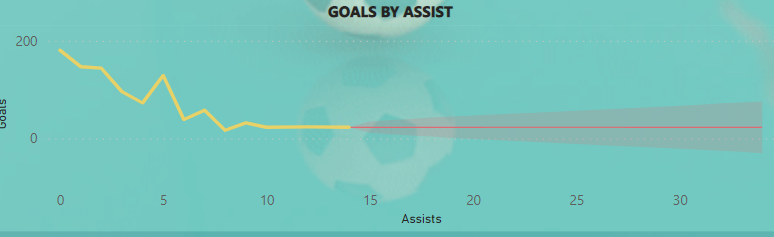


Figure . FUTURE FORCAST OF GOALS AND ASSIST BY POSITION

# 

# AI & Q&A

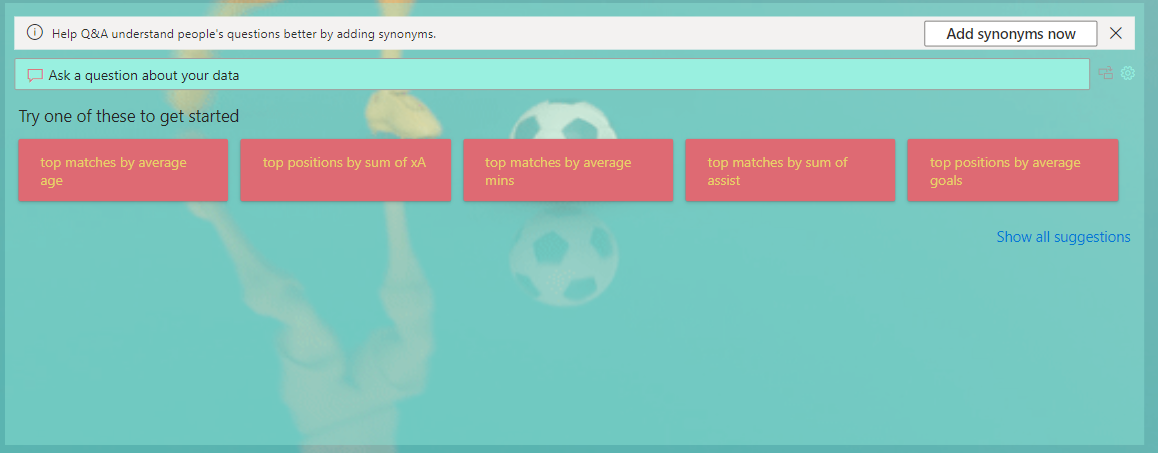


Figure 86. Q & A

The chart was created using Q&A. The questions are AI-generated. These help the user ask a certain question based on the data for easy analysis and understanding. The user can also ask and generate questions.



Figure 87. SLICERS

The chart was created using a slicer. Five slicers were created. The first slicer was filled with club data in the field section. Similar procedures were done with the four other slicers with club ,position, xG performance and xA performance placed in the Field section in their respective slicers.

# Reference

https://www.kaggle.com/datasets/rajatrc1705/english-premier-league202021