BIG DATA AND BUSINESS INTELLIGENCE MODULE

2020/2021 ENGLISH PREMIER LEAGUE ANALYSIS BUSINESS INTELLIGENCE SOLUTION

Adejubu Praise Peace

B1220791@live.tees.ac.uk

School of Computing,

Engineering & Digital

Technologies

Teesside University

TABLE OF CONTENT

EXECUTIVE SUMMARY

The English Premier League is a highly competitive and lucrative football league that attracts millions of fans and billions of pounds in revenue each year. Using data from the 2020-2021 season, this report aims to provide a business intelligence solution to help understand the trends and patterns within the league to make informed decisions.

This analysis was done using Power BI and it includes visuals such as bar graphs and scatter plots, as well as measures and calculated columns created using DAX and M language. I analyzed data from the kaggle.com [1] website on team and player statistics such as goals scored, assists, and passes completed. Our analysis found that Manchester City and Manchester United were the most successful teams in the 20/21 season, ranking top in most team stats. In terms of individual player performance, Manchester City's Kevin De Bruyne and United's Bruno Fernandes were the standout performers, leading the league in goals and assists in that Season. Other top performers included Liverpool's Mohamed Salah, who scored 22 goals, and Tottenham's Harry Kane, as we say "Football Is all about goals".

Key findings include:

- Fulham and Liverpool both have the most count of different nationals playing for them in their clubs, with 17 each. This does not affect thier team game play.
- We can see that higher attempts to pass the ball create a higher goal ratio. High scoring clubs like Manchester city, Manchester united, and Liverpool all have the highest pass attempts.
- 23-30 is the prime age for footballers in the EPL, Players in this range are first team players, and they tend to have the highest playing time and score more goals.
- The most sought after posiition in the EPL is DF, and MF next, based on the fact that teams with higher pass rate, score more goals. So most managers now play possession style football from the defence line.
- The most aggressive age in the premier league is 21-29 years, being the most booked.

Recommendations: Overall, our analysis showed that both teams were consistent and dominant in their performance, and this was key factor in their success. Based on these findings, it is recommended that clubs focus on improving their attack to increase their chances of success in the

Premier League by investing in top attacking players. We also recommend that players aim to improve their goal-scoring and assist statistics to stand out in the league.

INTRODUCTION

The English Premier League is a professional football league in England and Wales [2]. It is the top tier of the English football league system and is contested by 20 teams. The 2020/2021 season saw a total of 380 matches played. The business intelligence requirements for this analysis were to identify the key performance indicators for each club and to understand the factors that contribute to their success in the league. The English Premier League is highly competitive and lucrative football league that attracts millions of fans and billions of pounds in revenue each year. To better understand the performance of teams and players in the 20/21, we conducted a business intelligence analysis using Power BI visuals.

Data Source:

The data for this analysis was sourced from the Kaggle website [1], which includes statistics for all clubs in the league. The dataset includes information on points scored, goals scored, goals conceded, and shots on target for each club. I also sourced for the club logos from dream league soccer kits website [3].

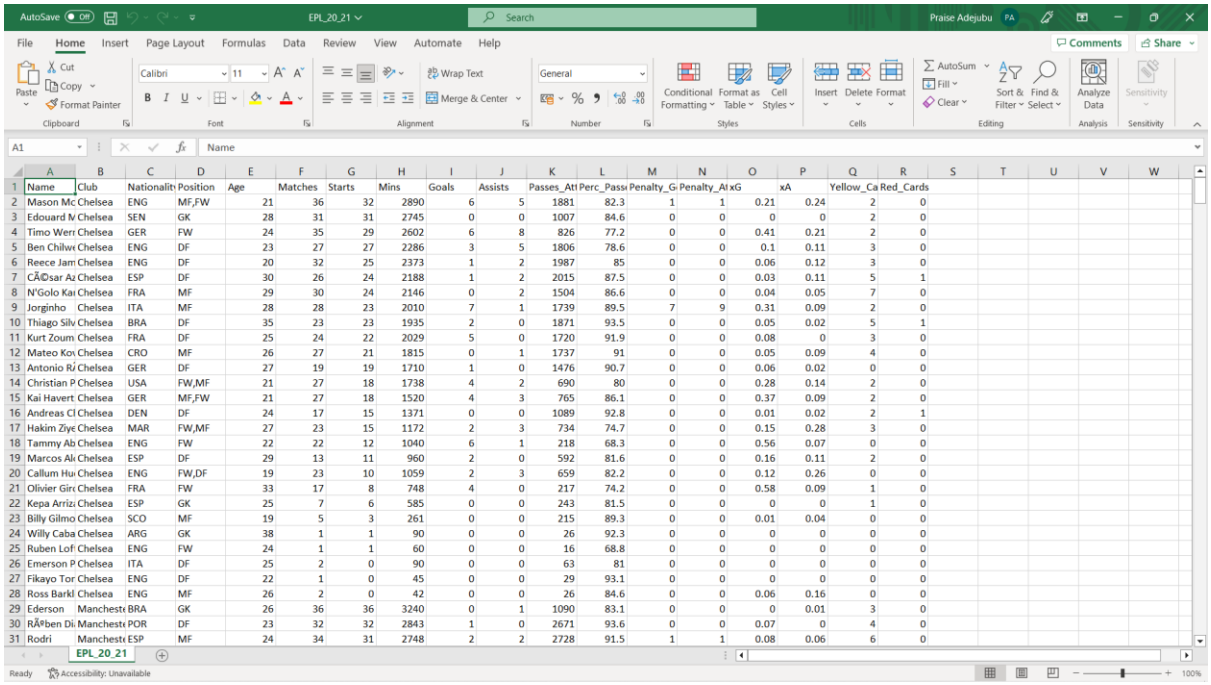
Index	Columns	Description
1	Position	Each player has a certain position, in which he plays regularly. The position in this dataset are, FW - Forward, MF - Midfield, DF - Defensive, GK - Goalkeeper
2	Starts	The number of times the player was named in the starting 11.
3	Mins	The number of minutes played by the player.
4	Goals	The number of Goals scored by the player.
5	Assists	The number of times the player has assisted other player in scoring the goal.
6	Passes_Attempted	The number of passes attempted by the player.
7	Perc*Passes*Completed	The number of passes that the player accurately passed to his teammate.
8	xG	Expected number of goals from the player in a match.
9	xA	Expected number of assists from the player in a match.
10	Yellow_Cards	The players get a yellow card from the referee for indiscipline, technical fouls, or other minor fouls.

Index	Columns	Description
11	Red Cards	The players get a red card for accumulating 2 yellow cards in a single game, or for a major foul.
12	Name	The name of the player.
13	Club	The club a player is signed to.
14	Nationality	The country a player nationalizes with.
15	Matches	The number of matches played by the player.
16	Penalty_Goals	The number of goals a player has scored from a penalties kick.
17	Penalty_Attempt	The number of attempted penalties a player has taken.
18	Age	The players age

Table 1: Column description

Clubs	URL link
▪ Arsenal Logo PNG	▪ https://i.ibb.co/MGzp8F5/Arsenal.png
▪ Aston Villa Logo PNG	▪ https://i.ibb.co/Z82PRMB/Aston-Villa.png
▪ Brighton Hove Albion Logo	▪ https://i.ibb.co/hC6p0QZ/Brighton-Hove-Albion.png
▪ Burnley FC Logo	▪ https://i.ibb.co/SdYsbpt/Burnley.png
▪ Chelsea FC Logo	▪ https://i.ibb.co/8Krn4jf/Chelsea.png
▪ Crystal Palace Logo	▪ https://i.ibb.co/0K0gD0b/Crystal-Palace.png
▪ Everton Logo	▪ https://i.ibb.co/pyrjGNr/Everton.png
▪ Fulham United	▪ https://i.ibb.co/1Myd3WP/Fulham.png
▪ Leeds United FC	▪ https://i.ibb.co/mbWVMtZ/LeedsUtd.png
▪ Leicester FC	▪ https://i.ibb.co/vLDCCK7/Leicester-City.png
▪ Liverpool FC	▪ https://i.ibb.co/Sfv6h5T/Liverpool-FC.png
▪ Manchester City	▪ https://i.ibb.co/F49gQQg/Manchester-City.png
▪ Manchester United	▪ https://i.ibb.co/zJyLGx0/Manchester-United.png
▪ New Castle United	▪ https://i.ibb.co/7YTwXDx/Newcastle-United.png
▪ Sheffield United	▪ https://i.ibb.co/r5VMyZx/Sheffield-United.png
▪ Southampton	▪ https://i.ibb.co/G0QR2HP/Southampton.png
▪ Tottenham Hotspurs	▪ https://i.ibb.co/wSKbNmf/Tottenham-Hostpur.png
▪ West Bromwich Albion	▪ https://i.ibb.co/R2YkkGd/West-Bromwich-Albion.png
▪ Westham United	▪ https://i.ibb.co/RpX17Gp/West-Ham-United.png
▪ Wolverhampton United	▪ https://i.ibb.co/MMM13gS/Wolverhampton-Wanderers.png

Table 2: Club logo data



Name	Club	Nationality	Position	Age	Matches	Starts	Mins	Goals	Assists	Passes_Att	Perc_Passe	Penalty_G	Penalty_At	xG	xA	Yellow_Ca	Red_Cards
Mason Mc	Chelsea	ENG	MF,FW	21	36	32	2890	6	5	1881	82.3	1	1	0.21	0.24	2	0
Edouard M	Chelsea	SEN	GK	28	31	31	2745	0	0	1007	84.6	0	0	0	0	2	0
Timo Werr	Chelsea	GER	FW	24	35	29	2602	6	8	826	77.2	0	0	0.41	0.21	2	0
Ben Chilwe	Chelsea	ENG	DF	23	27	27	2286	3	5	1806	78.6	0	0	0.1	0.11	3	0
Reece Jam	Chelsea	ENG	DF	20	32	25	2373	1	2	1987	85	0	0	0.06	0.12	3	0
CÃ©sar Az	Chelsea	ESP	DF	30	26	24	2188	1	2	2015	87.5	0	0	0.03	0.11	5	1
N'Golo Kan	Chelsea	FRA	MF	29	30	24	2146	0	2	1504	86.6	0	0	0.04	0.05	7	0
Jorginho	Chelsea	ITA	MF	28	28	23	2010	7	1	1739	89.5	7	9	0.31	0.09	2	0
Thiago Silv	Chelsea	BRA	DF	35	23	23	1935	2	0	1871	93.5	0	0	0.05	0.02	5	1
Kurt Zoum	Chelsea	FRA	DF	25	24	22	2029	5	0	1720	91.9	0	0	0.08	0	3	0
Mateo Kov	Chelsea	CRO	MF	26	27	21	1815	0	1	1737	91	0	0	0.05	0.09	4	0
Antonio RÃ	Chelsea	GER	DF	27	19	19	1710	1	0	1476	90.7	0	0	0.06	0.02	0	0
Christian P	Chelsea	USA	FW,MF	21	27	18	1738	4	2	690	80	0	0	0.28	0.14	2	0
Kai Havert	Chelsea	GER	MF,FW	21	27	18	1520	4	3	765	86.1	0	0	0.37	0.09	2	0
Andreas Cl	Chelsea	DEN	DF	24	17	15	1371	0	0	1089	92.8	0	0	0.01	0.02	2	1
Hakim Ziye	Chelsea	MAR	FW,MF	27	23	15	1172	2	3	734	74.7	0	0	0.15	0.28	3	0
Tammy Ab	Chelsea	ENG	FW	22	22	12	1040	6	1	218	68.3	0	0	0.56	0.07	0	0
Marcos Al	Chelsea	ESP	DF	29	13	11	960	2	0	592	81.6	0	0	0.16	0.11	2	0
Callum Hu	Chelsea	ENG	FW,DF	19	23	10	1059	2	3	659	82.2	0	0	0.12	0.26	0	0
Olivier Gir	Chelsea	FRA	FW	33	17	8	748	4	0	217	74.2	0	0	0.58	0.09	1	0
Kepa Arriz	Chelsea	ESP	GK	25	7	6	585	0	0	243	81.5	0	0	0	0	1	0
Billy Gilmo	Chelsea	SCO	MF	19	5	3	261	0	0	215	89.3	0	0	0.01	0.04	0	0
Willy Caba	Chelsea	ARG	GK	38	1	1	90	0	0	26	92.3	0	0	0	0	0	0
Ruben Lof	Chelsea	ENG	FW	24	1	1	60	0	0	16	68.8	0	0	0	0	0	0
Emerson P	Chelsea	ITA	DF	25	2	0	90	0	0	63	81	0	0	0	0	0	0
Fikayo Tor	Chelsea	ENG	DF	22	1	0	45	0	0	29	93.1	0	0	0	0	0	0
Ross Barkl	Chelsea	ENG	MF	26	2	0	42	0	0	26	84.6	0	0	0.06	0.16	0	0
Ederson	Manchest	BRA	GK	26	36	36	3240	0	1	1090	83.1	0	0	0	0.01	3	0
RÃºben Di	Manchest	POR	DF	23	32	32	2843	1	0	2671	93.6	0	0	0.07	0	4	0
Rodri	Manchest	ESP	MF	24	34	31	2748	2	2	2728	91.5	1	1	0.08	0.06	6	0

`Fig. 1: EPL 20/21 Original csv.data

The dataset was selected because it contains information about all different players that played in the EPL 20/21 season with their respective clubs.

The size of this dataset also makes it useful in aiding the development and demonstration of business skills like data cleaning and pre-processing, data modelling and relationship creation, it contains 533 row and 18 columns.

The standard stats such as Goals, Assists, xG, xA, Passes Attempted, Pass Accuracy and more [4], useful for this analyse and get useful trends and patterns that can help clubs improve their investment strategies.

BI Requirements/Questions:

The BI requirements for this analysis were to identify the key performance indicators for each club and to understand the factors that contribute to their success in the league. Specifically, the following questions were addressed:

- Which club team is the most diverse in nationality? And does this affect their game play in anyway?
- Do more passes correlate to clubs scoring more goals?
- At what age range does a player peak and gain the managers trust?
- What position is most sought after and why?
- What is the most booked position? How does booking influence game stat?, Also evaluate each team booking stat and penalty conversion rate?.

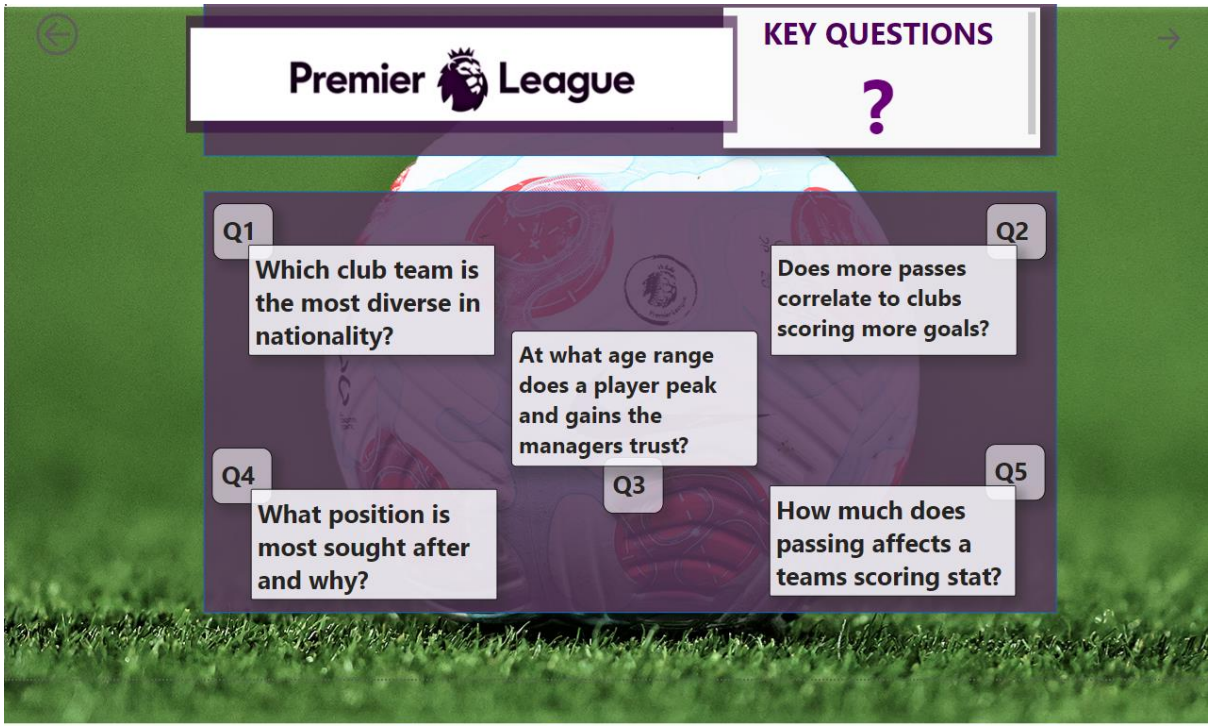


Fig. 2: Visual showing Key Question dashboard

FINDINGS BASED ON ANALYSIS AND EVALUATION

Using Power BI, several visualizations were created to analyse the data, answering the questions above.

1. The first dashboard is to answer the question: Which club team is the most diverse in nationality? And does this affect their game play in any way?

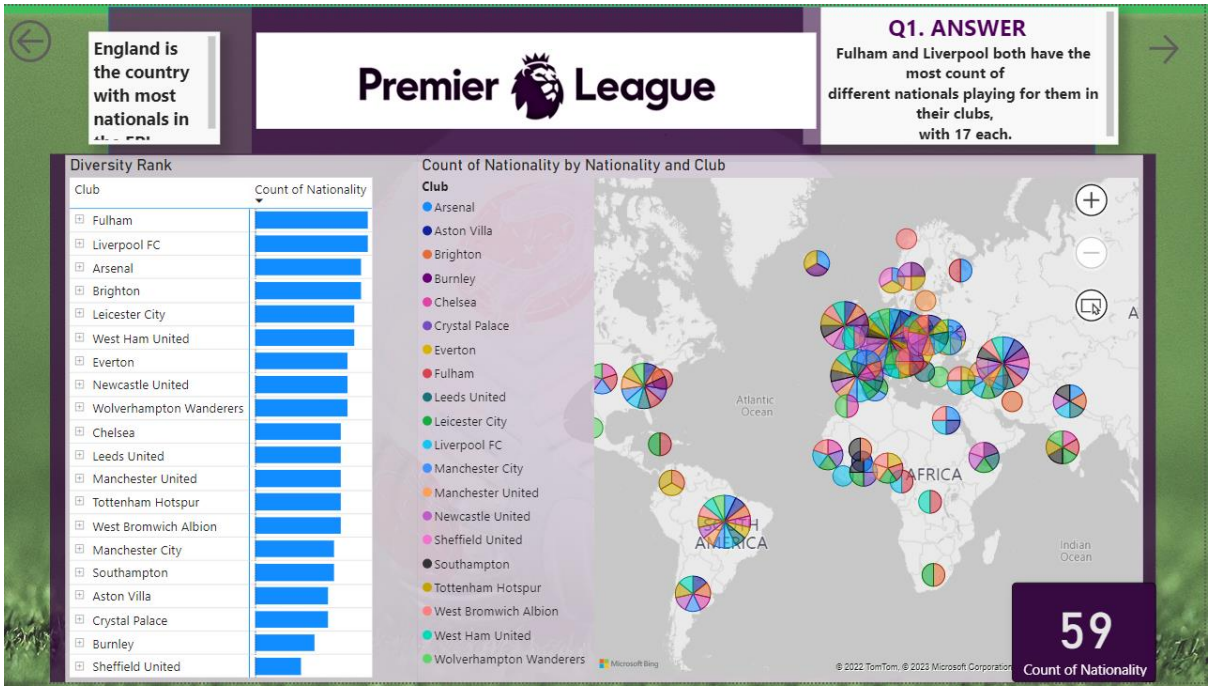


Fig. 3: Team Diversity Rank Dashboard

To visualize the diversity of nationality in an English Premier League (EPL) club team, I used a matrix table and a Map bubble chart to show the number of players on the team from different countries. The team with the highest number of players from the most countries would be considered the most diverse.

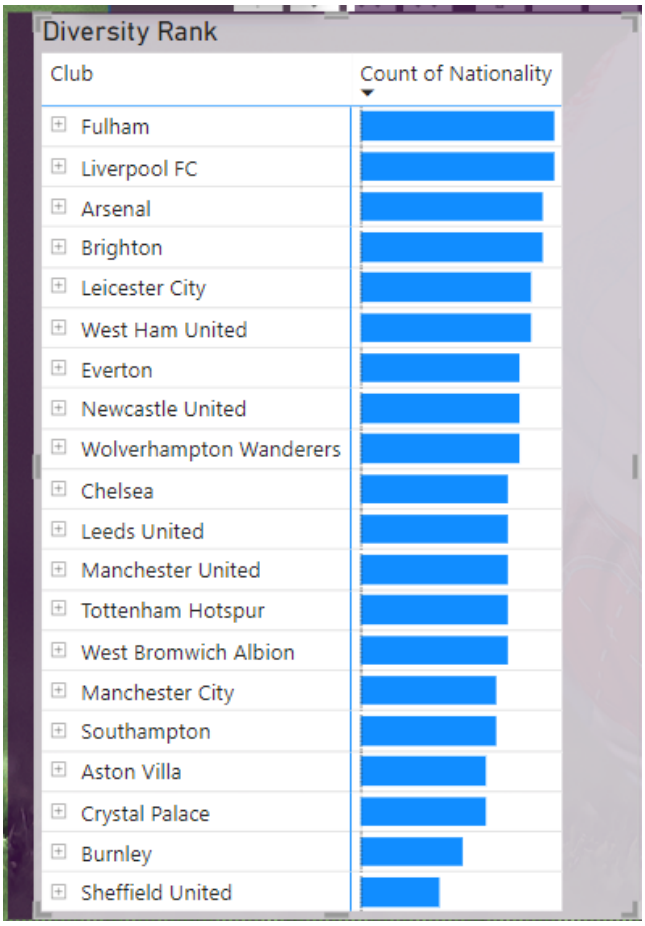


Fig. 4: Team Diversity Rank matrix table

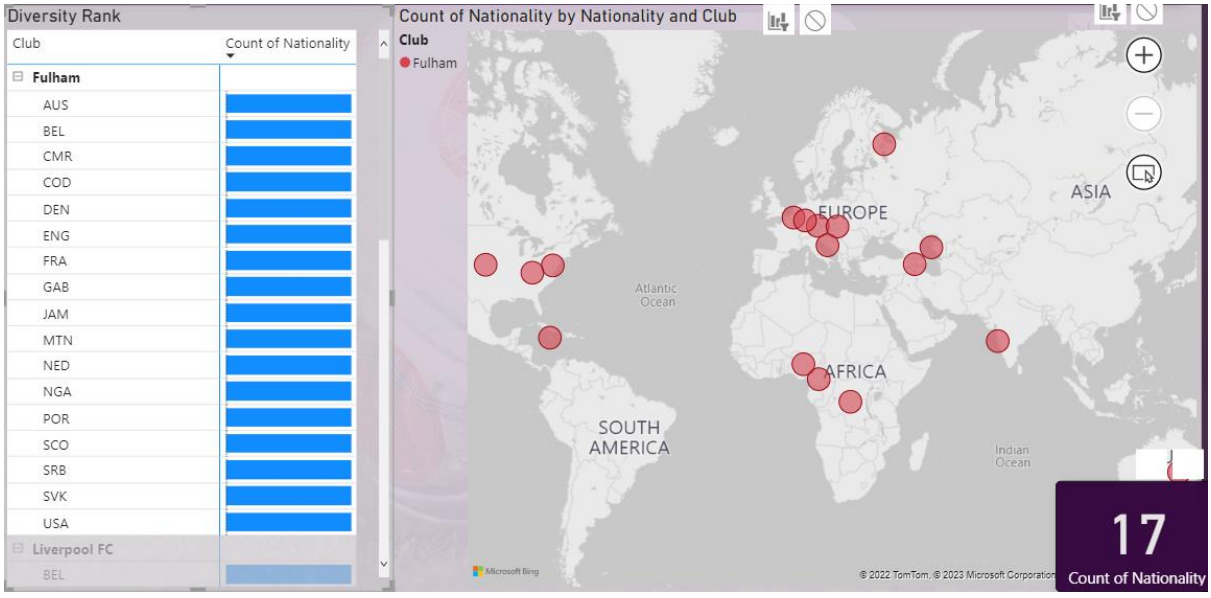


Fig. 5: Fulham Diversity Rank matrix table/ map

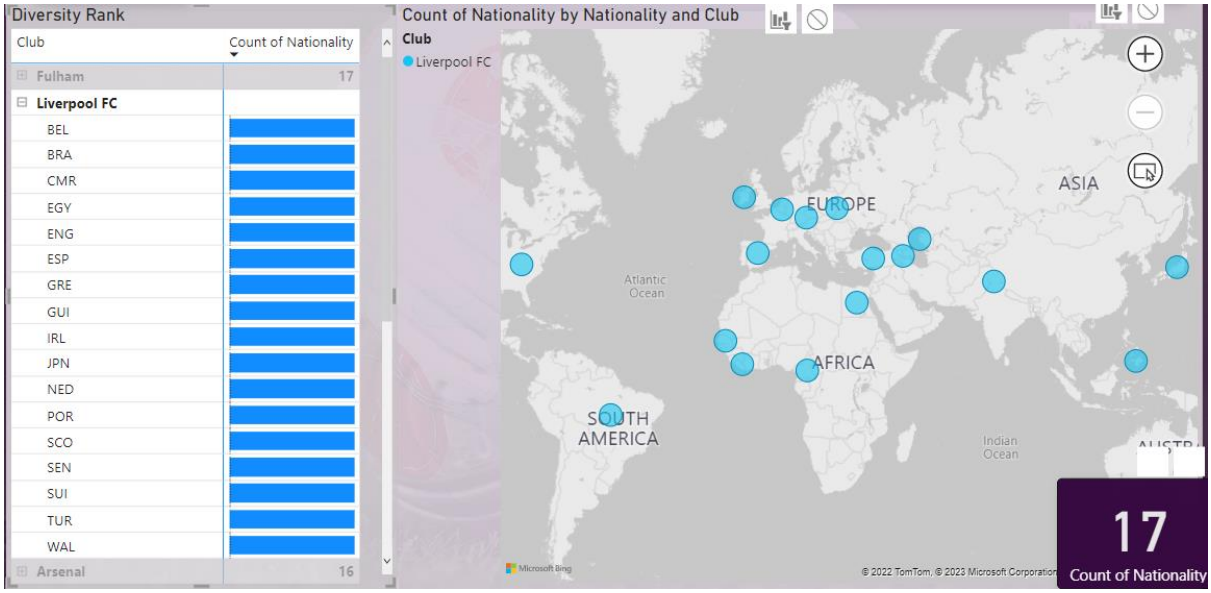


Fig. 6: Liverpool Diversity Rank matrix table/ map

Here on the Matrix table, Fulham and Liverpool ranks joint first with 17 different nationals being represented in their team. The table has horizontal bars stringed to the front that ranks each team with how much diverse they are, with 17 being the max and 0 being the least on the bar.

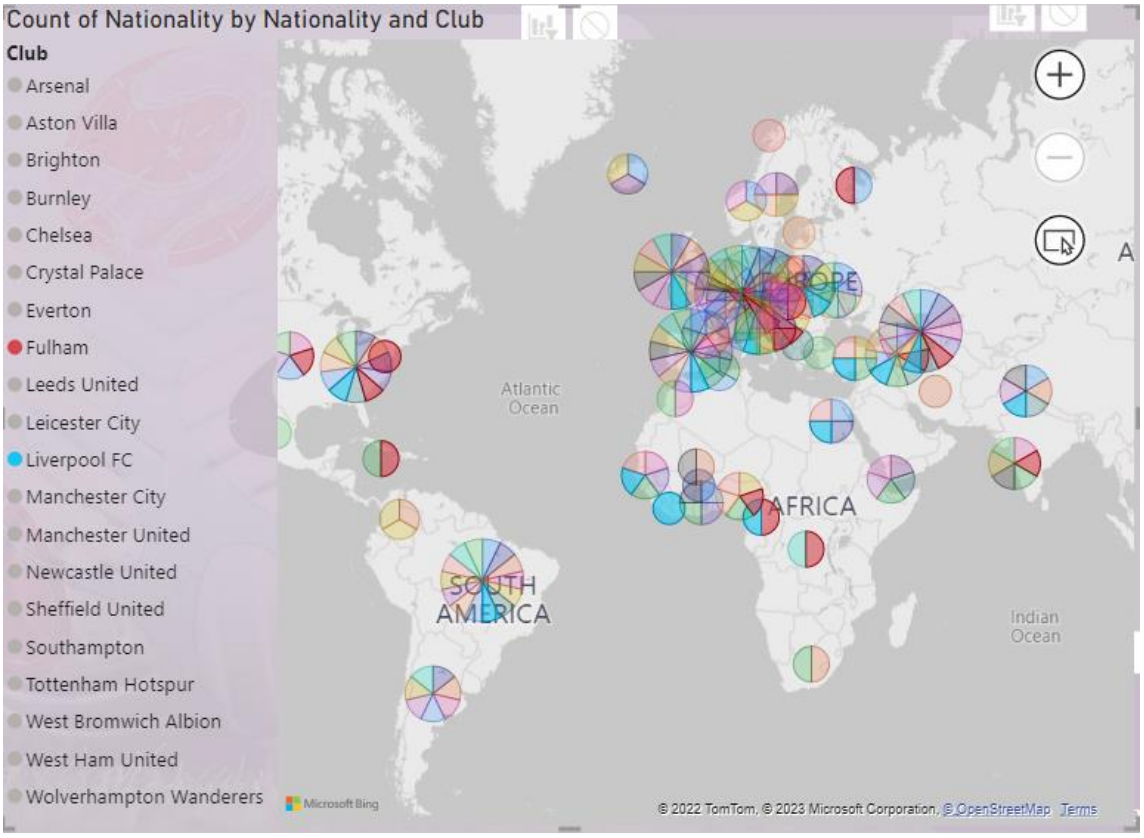


Fig. 7: bubble geographical map for club national diversity

I also considered the geographical map with bubbles to visualize, such that the countries with more players involved in the league has bigger sized bubble, and each club has its own colour representing them.

Fig. 8: clubs Nationality count

A counter was also attached using the card visual to show how many nationalities are represented by the clubs and the league in total. This would provide additional context and allow for more meaningful comparisons between teams.

Key Findings: So, we can see that from the illustrations above, there is no correlation of a club's team being nationally diverse and playing well. This is because Fulham and Liverpool are currently at opposite levels on stats, Fulham are currently amongst the least successful team in the league, and Liverpool is the third-ranking team in across all performance stats. The only fact I can draw is diversity will increase the club's social image which could also increase the teams merchandise sale across the world.

2. The second dashboard reveals the relationship between the number of passes a club makes and the number of goals they score in each match. This visualization allows us to see whether there is a correlation between the two variables, meaning that as one increases, the other tends to increase as well.

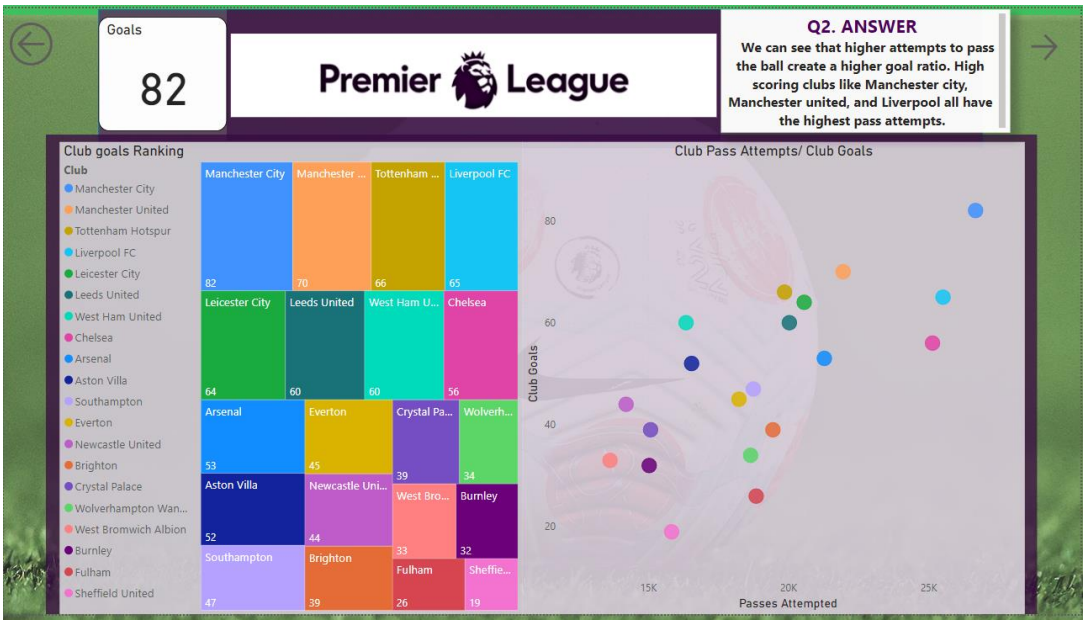


Fig. 9: Dashboard Evaluating Passes to Goal relationship

A calculated column was created using DAX to achieve this visual:

```
1 Club Goals = SUMX(DISTINCT('Clubs'[Club]),CALCULATE(SUM('Performance stat'[Goals])))
```

Fig. 9: DAX calculated column “Club Goals”

Fig. 10: Treemap and Scatter chart showing club goal rank and relationship of pass to goals

Scatter chart and tree map were adopted in this dashboard to show the prove that the higher attempts to pass the ball the more goals are scored. High scoring clubs like Manchester city, Manchester united, and Liverpool all have the highest pass attempts.

From the visualization, it appears that there is a positive correlation between passes and goals scored. This means that clubs that make more passes tend to score more goals. This is likely because more passes typically lead to more possession and control of the ball, which in turn can lead to more opportunities to score goals.

Key Findings: Factors Contributing to Team Success: Our analysis found that Manchester City and Manchester United’s consistent passing and dominant performance was key factor in their success, as they had the most passes and goals in the league. Manchester city scored 82 and had over 25k passing attempts during the season, trumping the goal rank with only United coming in second.

3. The third visualization is a useful tool for understanding the age at which a player is most likely to be a key member of the first team and play most games. It can help managers identify when a player is likely to be at their peak performance level, both in terms of overall contribution to the team and goal-scoring ability.

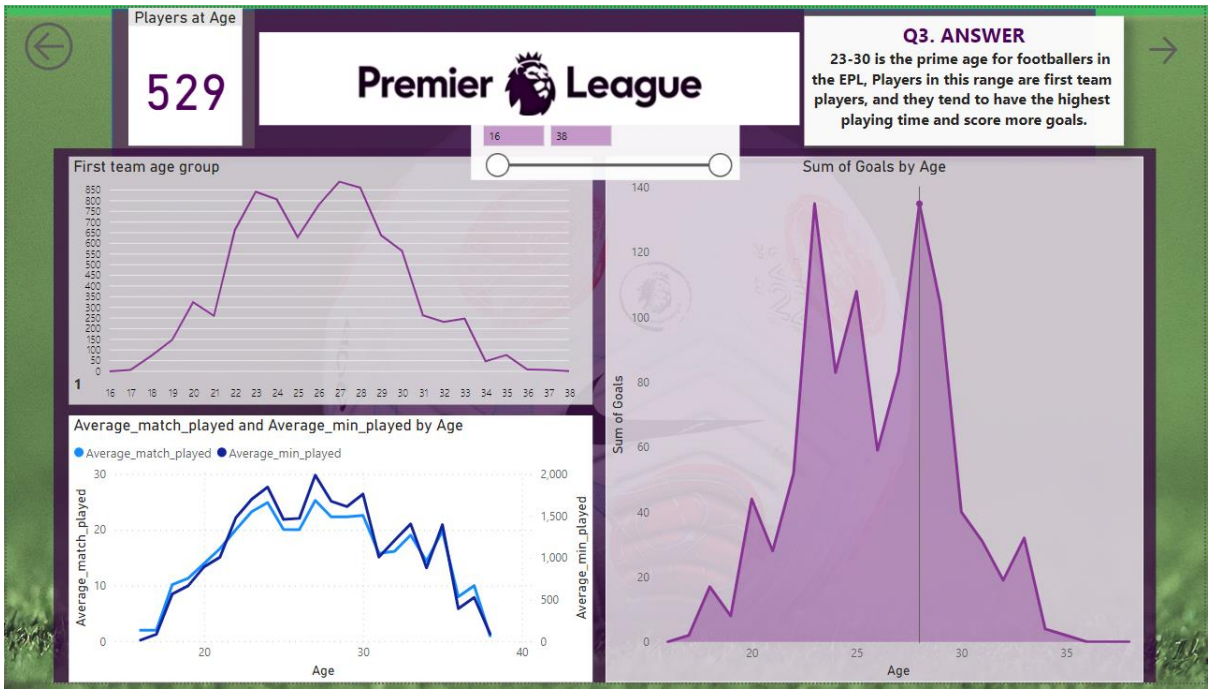


Fig. 11: Dashboard Evaluating age player's peak

A DAX was created to compute some calculated measures "Average_match_played", and "Average_min_played" ;

The "Average_match_played" measure is used to calculate the average number of matches played by a player over a given period of time. This can be useful for understanding how often a player is likely to be available for selection, and for making decisions about playing time and squad rotation.

```
1  Average_match_played = AVERAGE('Match statistics'[Matches])
```

Fig. 12: DAX calculated measure of "Average_match_played"

The "Average_min_played" measure, on the other hand, is used to calculate the average number of minutes played by a player over a given period of time. This can be useful for understanding how much playing time a player is likely to receive, and for making decisions about tactics and formations.

```
1  Average_min_played = AVERAGE('Match statistics'[Mins])
```

Fig. 13: DAX calculated measure of "Average_min_played"

Both of these calculated measures can be useful for managers and coaches in terms of understanding the performance and availability of their players. They can also be useful for fans and analysts in terms of understanding the contributions of different players to their teams.

A table visual was created to check the DAX solution in relation to players' age. As shown in the line chart visual.

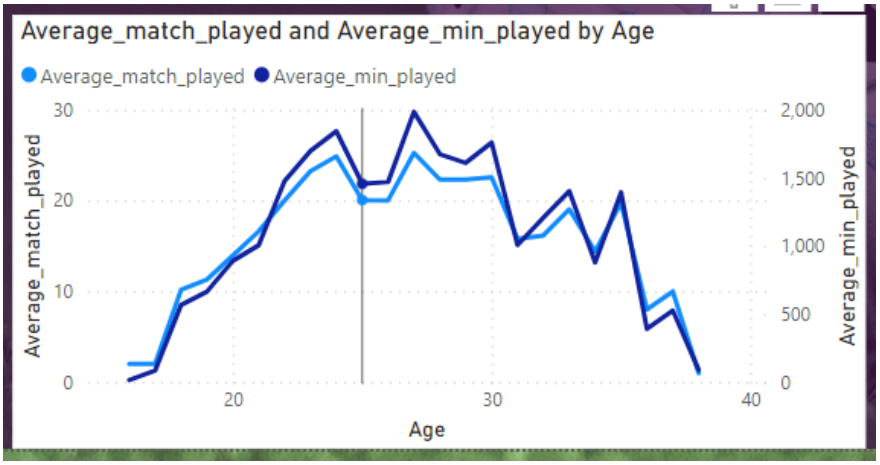


Fig. 12: Line chart comparing average match and minutes played to players age.

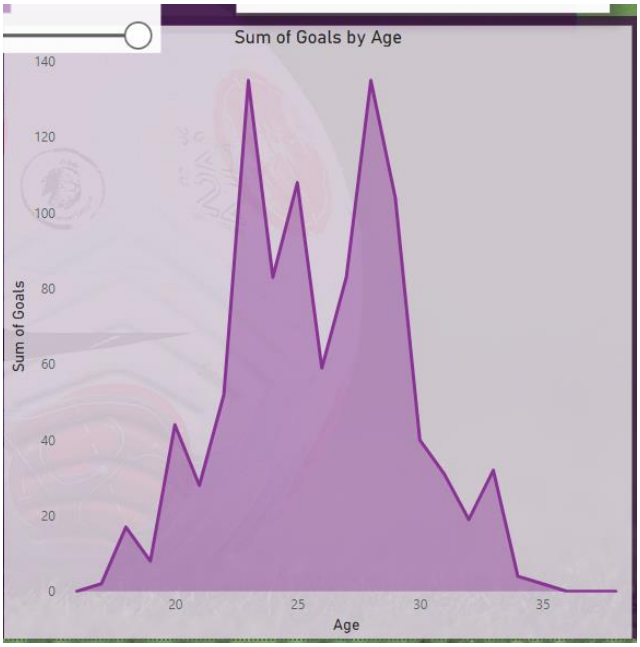


Fig. 13: Stacked area chart evaluating goal scorers by age group.

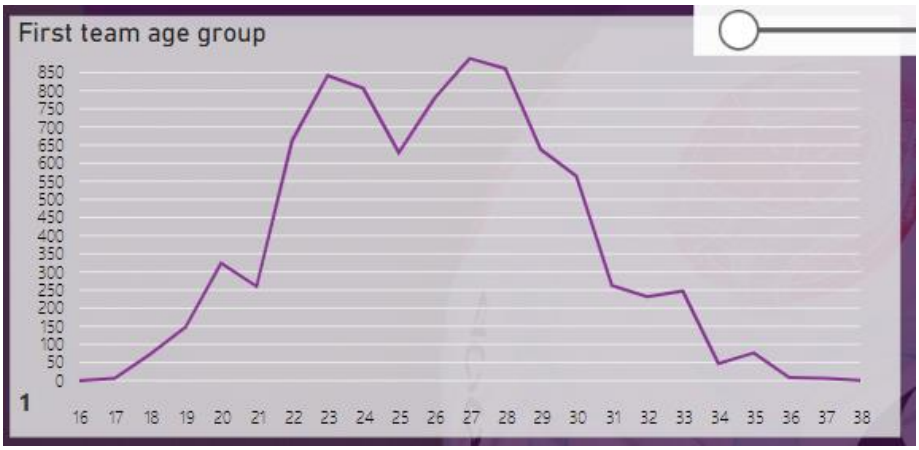


Fig. 14: Power KPI visual showing age range of first team players.

Key Findings: One key aspect of this visualization is that it allows managers to see the age range in which players are most trusted to start and play significant game time. This is important for a number of reasons. Firstly, it helps managers to understand the age at which players are most likely to be trusted and relied upon by their teammates and coaches. This can be useful in terms of selecting players for the first team and allocating playing time.

Secondly, the visualization can provide insights into the prime age of players in the league. This is the age at which players are most likely to be at their best, in terms of both physical and mental performance. Knowing this age range can be useful for managers in terms of identifying the right players to sign and build their teams around.

23-30 is the prime age for footballers in the EPL, Players in this range are first team players, and they tend to have the highest playing time and score more goals.

4. The fourth answer to the question: What position is most sought after in the EPL and why? This will help further understand the answer to the second question, how passing aids goal scoring.

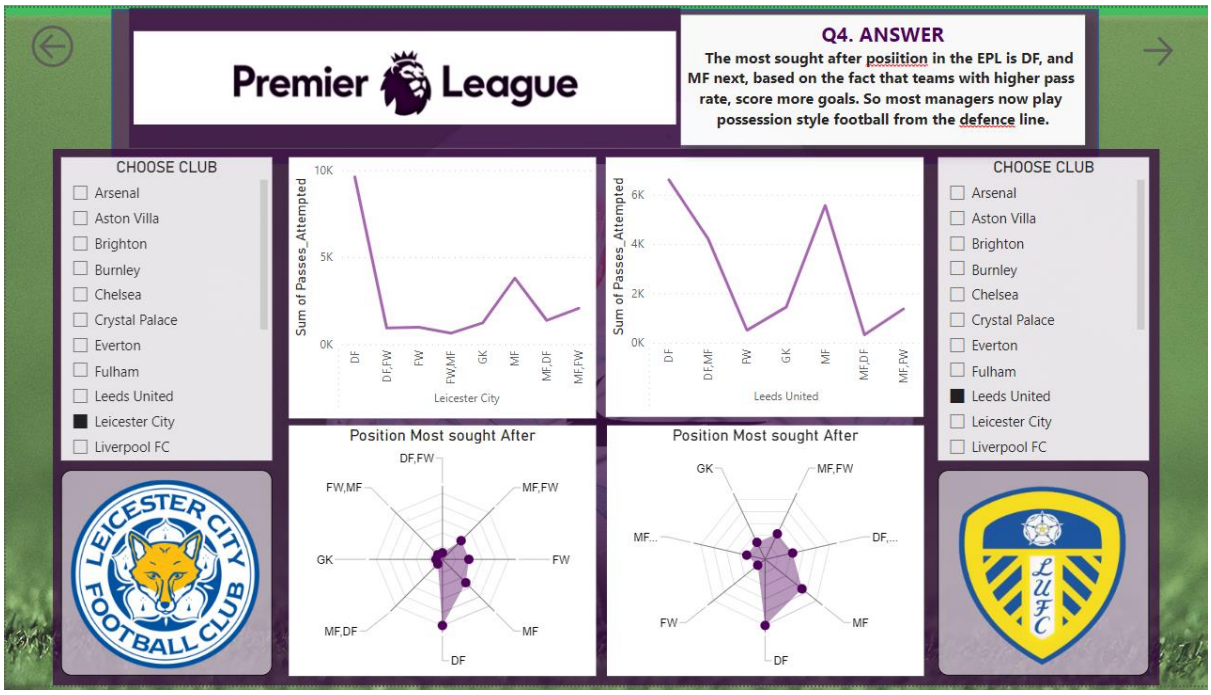


Fig. 15: Dashboard comparing and evaluating team positions priority

Key Findings: The position of defender (DF) and midfielder (MF) are the most sought after in the English Premier League (EPL) for a few reasons. Firstly, the role of these players is crucial in the success of a team's possession-based style of play. This means that they need to be skilled at passing the ball to their teammates, both to maintain possession of the ball and to create scoring opportunities. As teams with higher pass rates tend to score more goals, it makes sense that managers would prioritize these positions in order to improve their team's chances of winning.

Additionally, defenders and midfielders also play a key role in preventing the opposing team from scoring goals. They are responsible for marking and tracking the movements of opposing players, as

well as intercepting passes and tackling the ball away. Therefore, having strong defensive players is essential for any team that wants to have success in the EPL.

In this dashboard I setup a comparison table for any two teams, so we can compare two teams and see their success rate as correlating to their passing attempts. This visual shows where each team places priority in signing players.

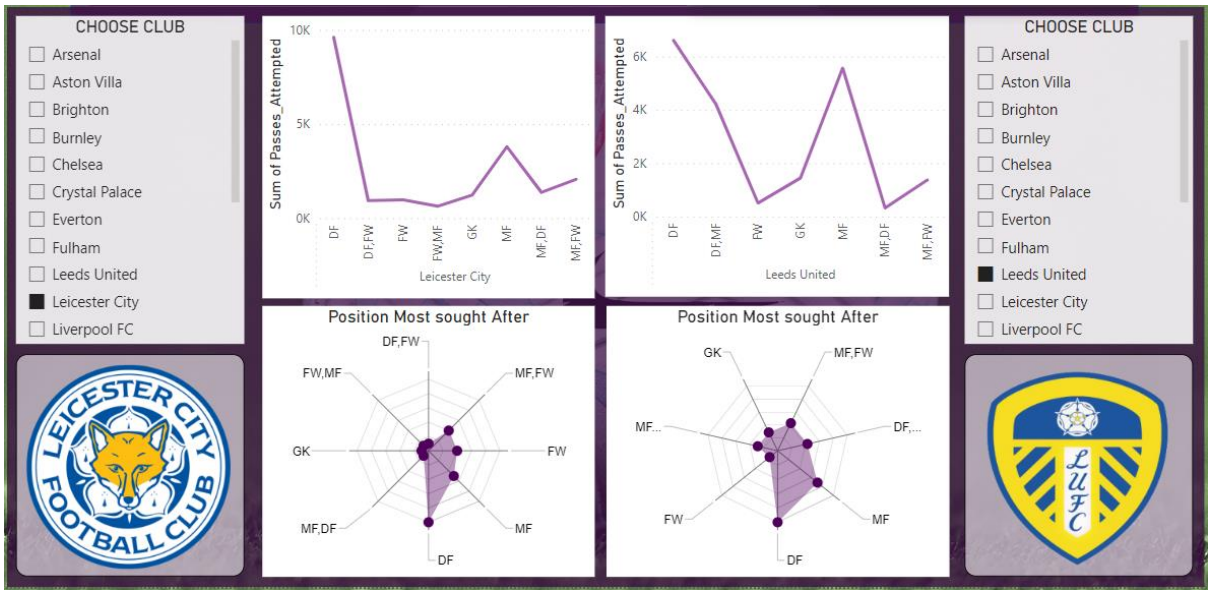


Fig. 16: showing visual setup for comparing club position/passing data

The dashboard is made of two equal halves and two slicers to choose whichever teams to be compared using the radar chart and the line chart as shown in Fig. 15 and Fig.16 respectively.

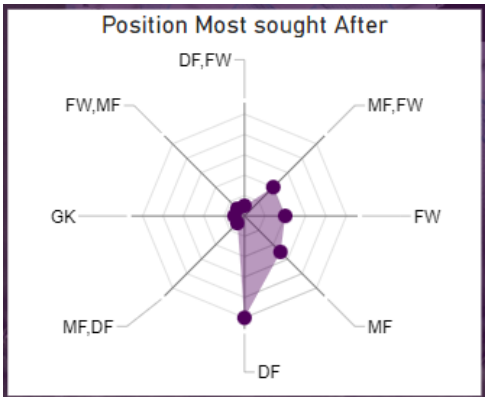


Fig.17 Radar chart

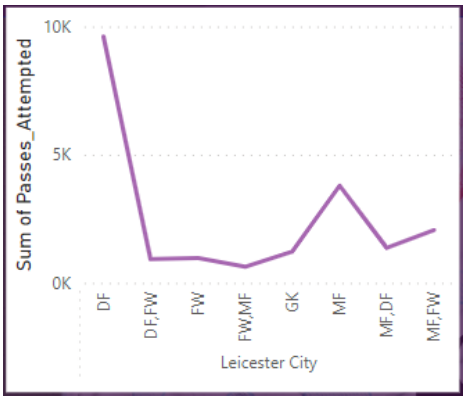


Fig.18 Line chart

The most succesful teams in the season, Manchester city and United mostly prioritised DF and MF positions and had higher pass rate which led to high goal tallies.

5. This final visualization addresses multiple questions in one dashboard and can be useful for quickly understanding trends and patterns in data. The visualization might show the most booked position in the league, along with how booking (e.g., a yellow card given to a player) influences game statistics such as scoring or aggression. The visualization also shows the booking and penalty conversion rate for each team, allowing us to see which teams are more aggressive or disciplined in their play and clinical in their finish. Overall, this type of visualization can provide a comprehensive view of the data, making it easier to draw insights and make informed decisions.

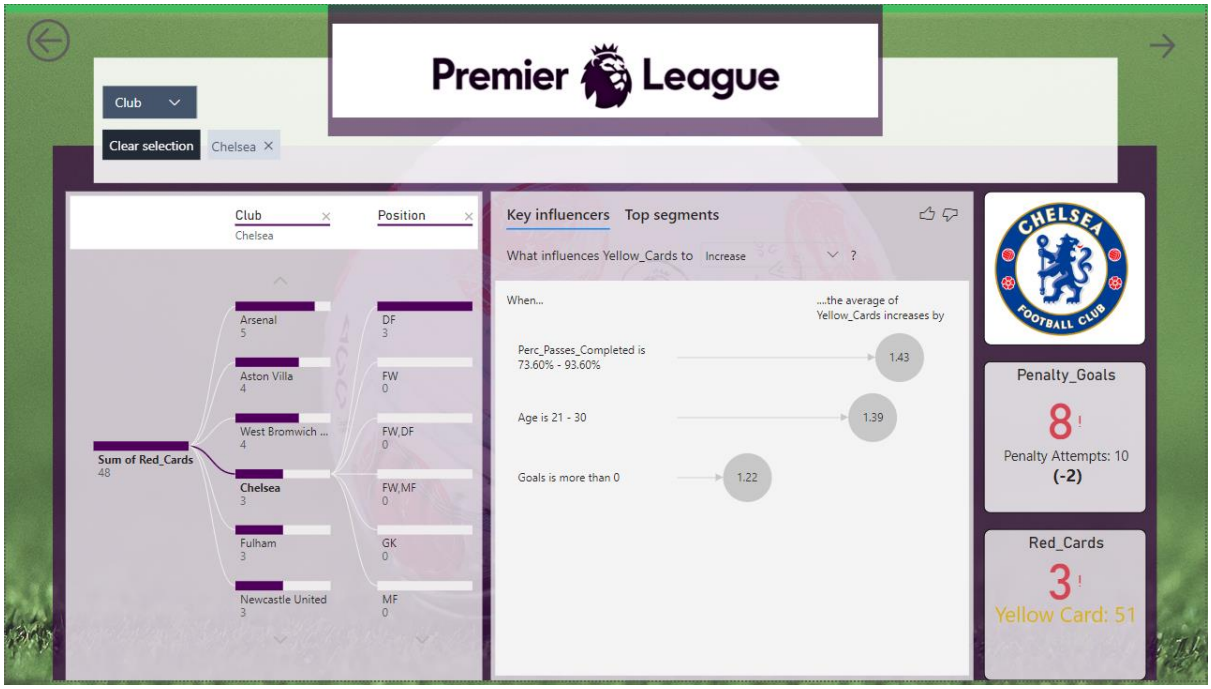


Fig. 19: Dashboard evaluating clubs booking and penalty stats

Fig 20. Clubs Penalty stats.

Fig 21. clubs Booking stats

The KPI card is used to show statistics for each club's penalty attempt and conversion rate, and also the number of penalties missed. If any is missed the value turns red. The KPI card was also used as seen in fig. 21, to show the number of red and yellow card given to each club.

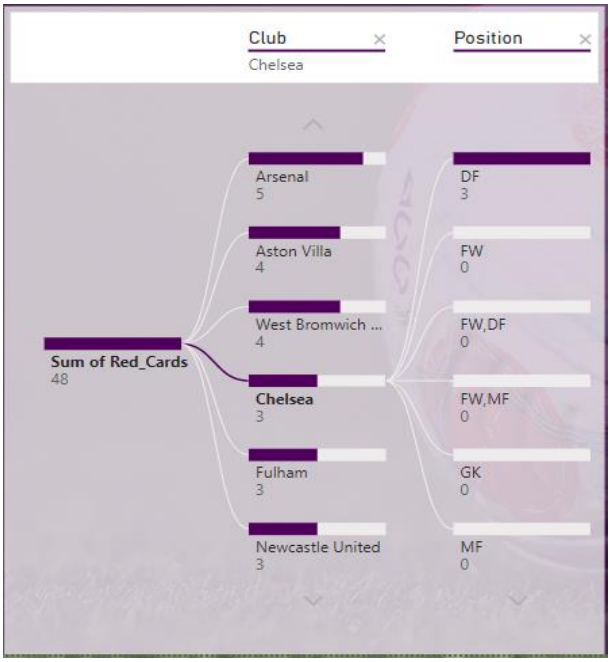


Fig 22. Decomposition tree evaluating red card by club/position

Key influencers were also adopted to show what influences yellow card to increase.

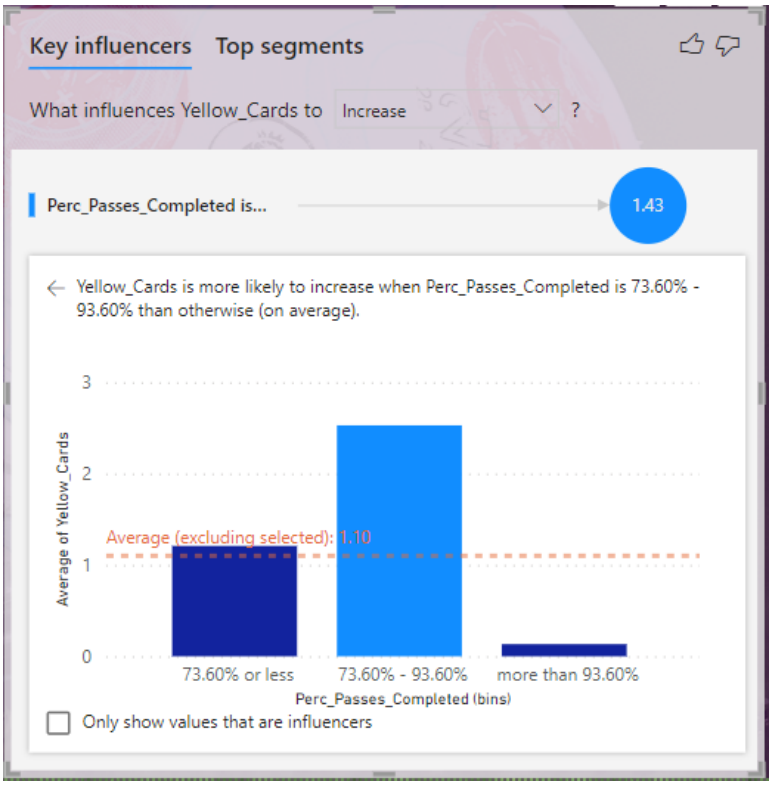


Fig 23. Visual showing passes completed Influence on overall yellow card stat

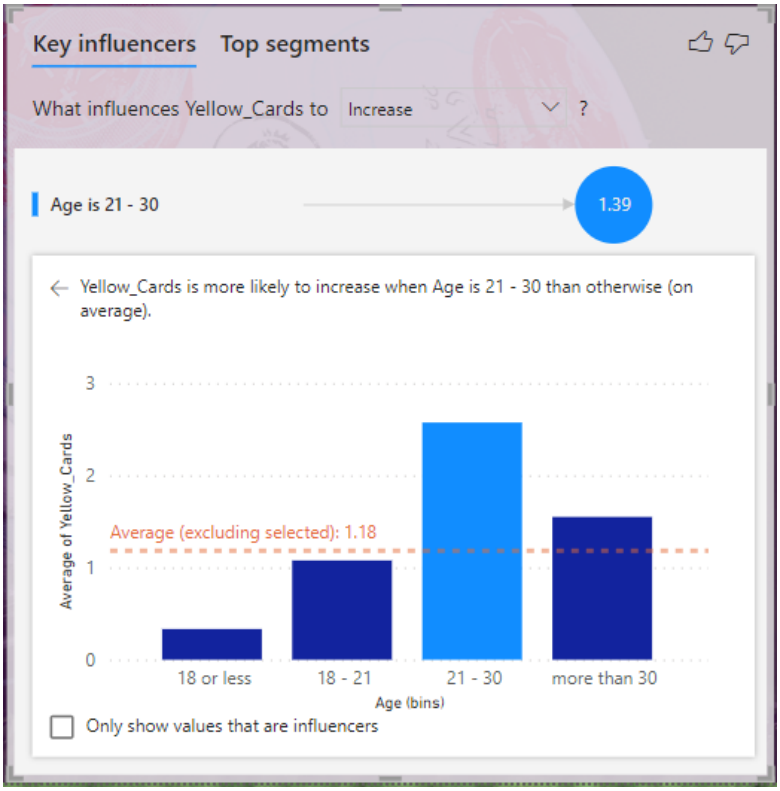


Fig 24. Visual showing age Influence on overall yellow card stat

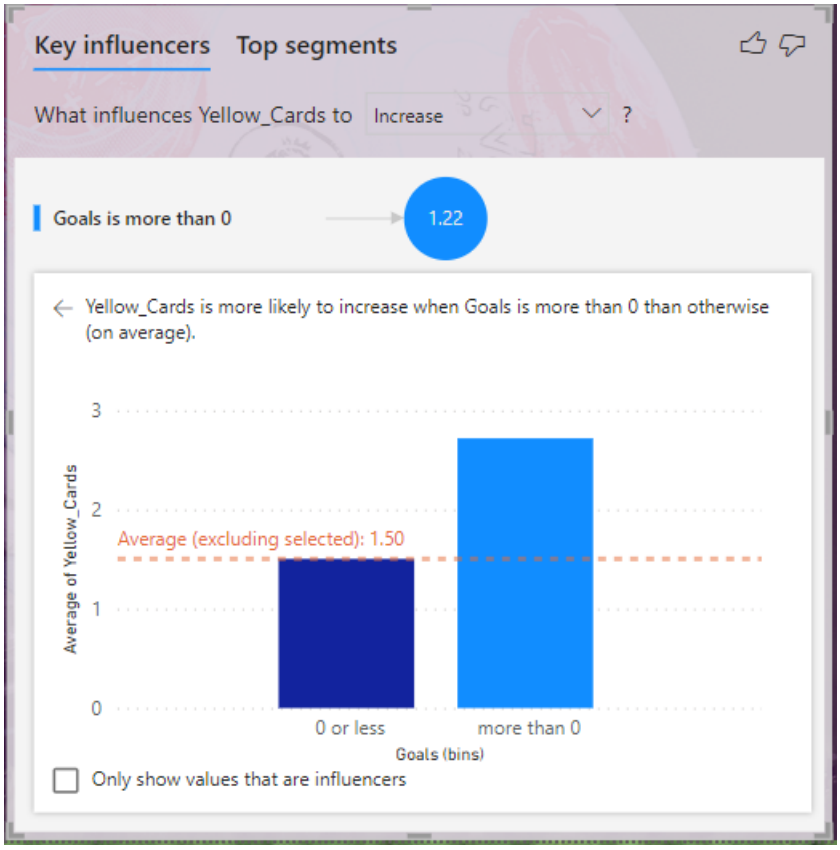


Fig 25. Visual showing goals Influence on overall yellow card given

A measure “Average_Age_Players” was created using DAX for the next visualisations on the dashboard :

```
1  Average_Age_Players = AVERAGE('Player information'[Age])
```

Fig. 26: DAX calculated measure of “Average_Age_Players”.

The measure "Average_Age_Players" was calculated using the DAX (Data Analysis Expression) language in order to display the next visual on the dashboard. This measure was used in visualizations on the dashboard to display the average age of players in a specific data set. This measure can be helpful for analyzing trends in the age of players that collected yellow card more.

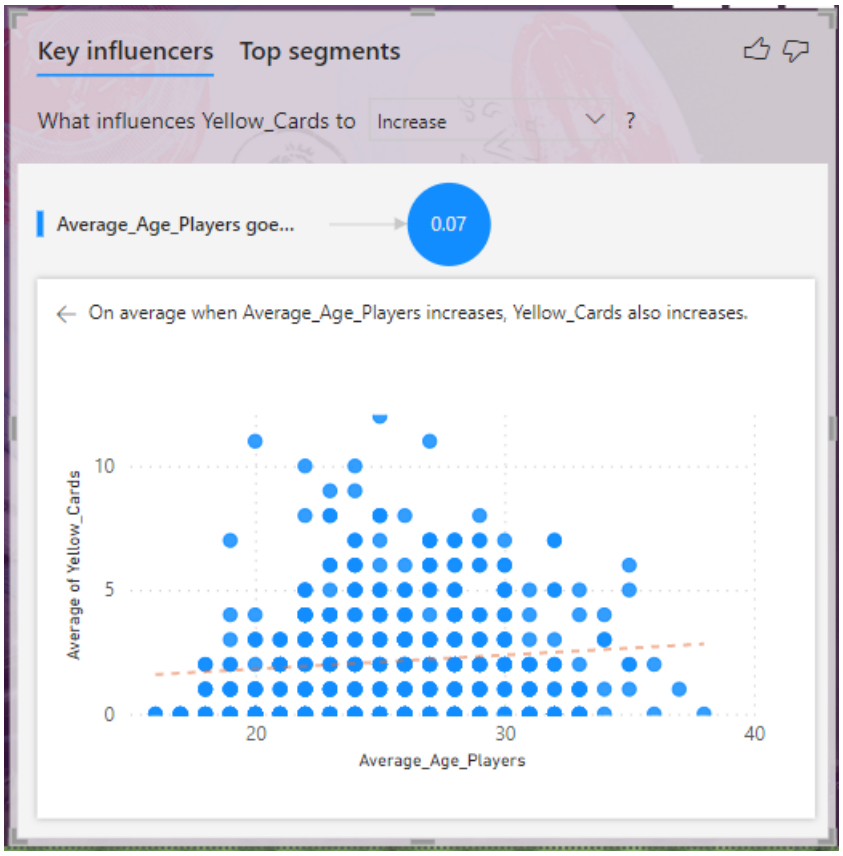


Fig 27. Visual showing players average age Influence on overall yellow card given

Key Findings: Finally, the dashboard includes a visual that shows the booking and penalty conversion rates for each team. This allows coaches and players to compare their performance to that of their opponents and identify areas where they may need to improve. Additionally, it helps fans understand how each team is faring in terms of discipline and penalty management. Overall, the final visualization provides a comprehensive view of booking and penalty data that can be used to inform strategy and improve performance.

RECOMMENDATIONS

Based on these findings, it is recommended that clubs focus on improving their defence and midfield positions in order to increase their chances of success in the Premier League. This could involve investing in players who excel in these areas or implementing strategies to improve their performance in these areas. Additionally, clubs should consider investing in players with a high number of shots on target, as this can lead to a higher number of goals scored.

The importance of passing in possession-based football and the need for strong defensive players are the main reasons why the positions of DF and MF are so sought after in the EPL. By prioritizing these positions, teams can improve their chances of success on the pitch and better control the flow of the game.

Finally, the visualization can help managers to identify the age at which players are most likely to be at their goal scoring best. This is important for a number of reasons, including the fact that goals are a key factor in determining the success of a team. Knowing the age at which players are most likely to score goals can help managers to identify key players to build their teams around, and to allocate playing time and resources accordingly.

CONCLUSIONS

In conclusion, during this analysis, I found out that I enjoy analysing and visualizing datasets, the 2020/2021 English Premier League data in particular is something that resonates with me as I am a football lover, and I follow the league here in England. Furthermore, I am a faithful fan of Manchester united and their performance during this season was great.

Personally, Prior to starting this project, I had limited knowledge of Power BI. However, during the course of this project, I was able to significantly improve my skills in Power BI and data analytics. There were several challenges that I encountered while working on this project, such as selecting an appropriate dataset, cleaning and processing the data. I was able to overcome these challenges through guidance from my tutors and by conducting my own research and learning. I am excited to apply the skills I have gained from this module to real-world situations and believe they will be useful in addressing problems throughout my career.

REFERENCES

[1]https://www.kaggle.com/datasets/rajatrc1705/english-premier-league202021?select=EPL_20_21.csv

[2] En.wikipedia.org. 2023. *English Premier League - Wikipedia*. [online] Available at: < https://en.wikipedia.org/wiki/English_Football_League > [Accessed 8 January 2023].

[3] https://dreamleaguesoccerkits.com/2021/05/premier-league-club-logo/

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

APPENDIX:

DATA PRE-PROCESSING OR DATA CLEANSING

Loading the Data:

The pre-processing of the EPL 20/21 dataset was first initiated by importing it into the Microsoft Power BI tool. There were two options for accomplishing this. Manually by selecting the file type from the "Get Data" menu and then importing it, or by creating a blank query and importing it using M-Language. In this report, the former was utilized; nonetheless, a screenshot of the M language input is included.

I also used external source to get the clubs logo for the visualizations. I will also be illustrating how I achieved this;

- Starting up Microsoft Power BI application:

Fig 28. Power BI starting up

- Click on the "Get Data" button in the ribbon at the top of the Power BI interface.

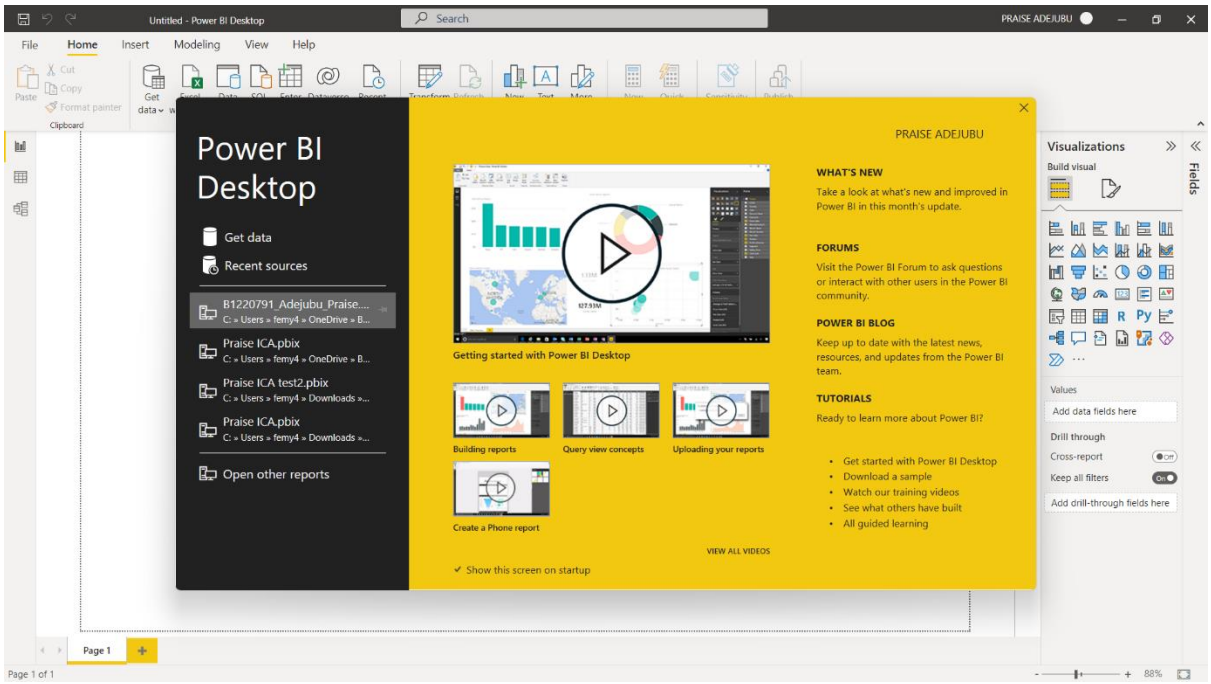


Fig 29. Power BI Start page

- In the "Get Data" window, select "File" from the list of data sources on the left side.
- From the list of file types, select "CSV" and click "Connect."

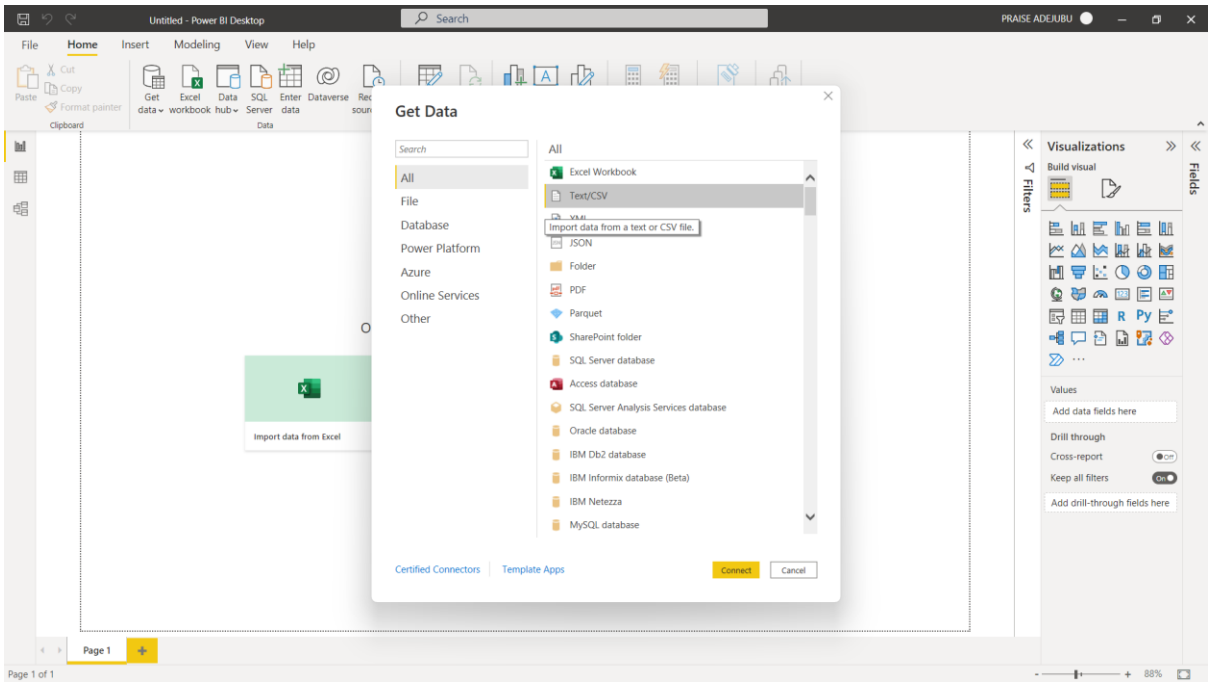


Fig 30. Power BI modelling process

- In the "Import Data" window, navigate to the location of the CSV file on your computer and select it.

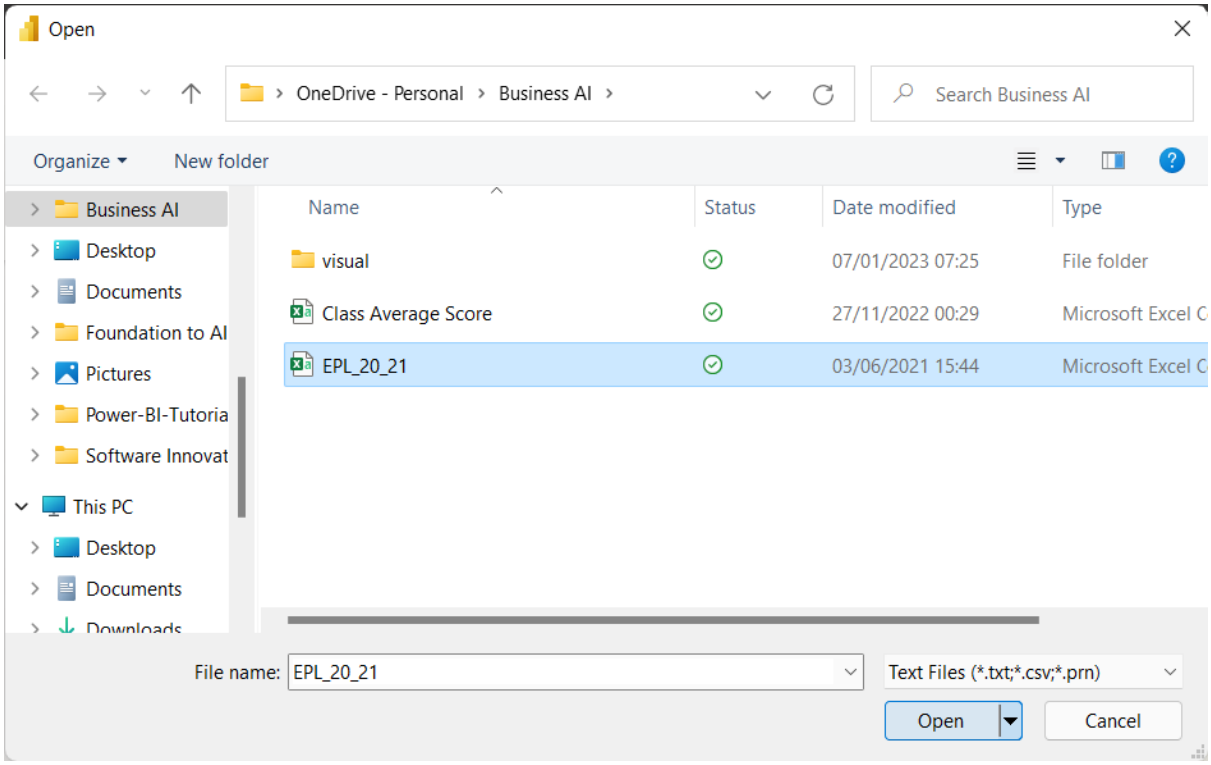


Fig 31. Power BI modelling process

- Click "Load" to import the data into Power BI.

When loading the data in, we need to attention to some key info on the dialog box:

i. the data type is calibrated based on the first 200 rows by default; however, this may be adjusted to use the complete dataset or no data type at all.

ii. The delimiter for a csv file is commas. This can vary also depending on the file. When importing a text file separated by “Tab”, the delimiter would be changed accordingly.

iii. File origin/encoding. The encoding defaulted to 1252: Western European (Windows).

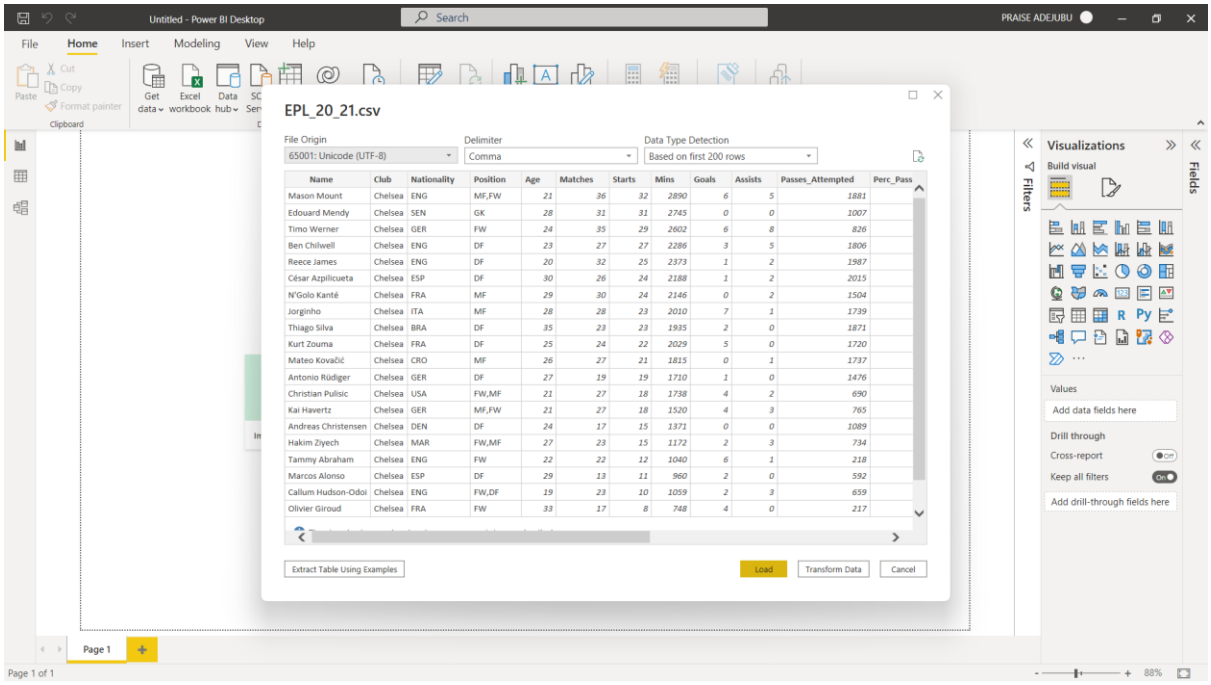


Fig 32. Power BI modelling process

- The data will be imported as a table.

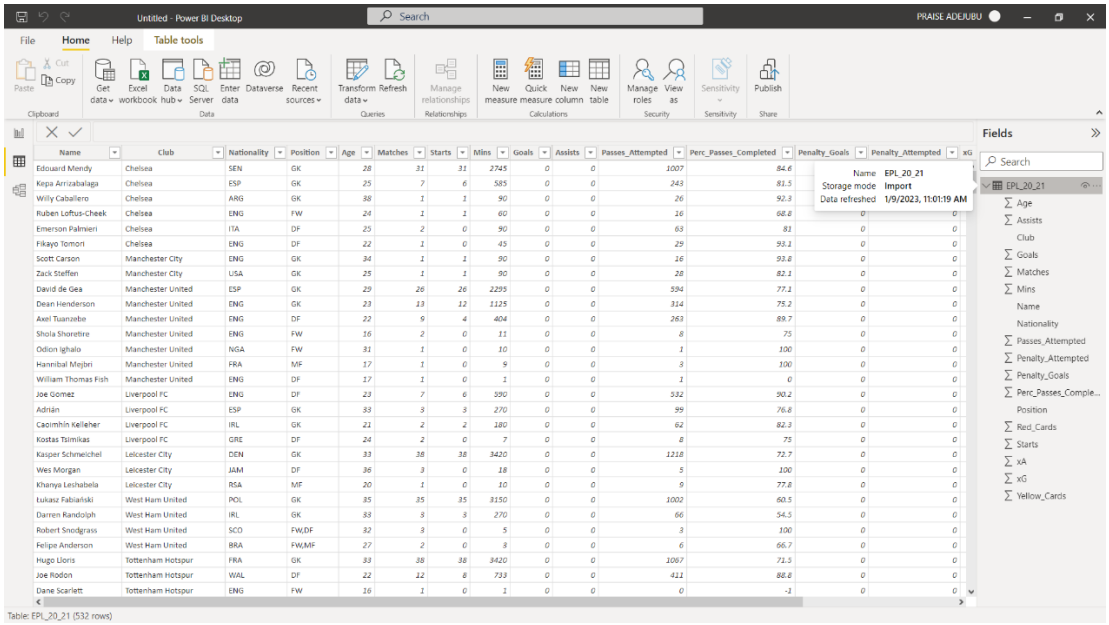


Fig 32. Power BI modelling process

Now for using M Language to import the dataset into power BI;

- Click on the "Get Data" button in the ribbon at the top of the Power BI interface and from the list of file types, select "Blank Query".

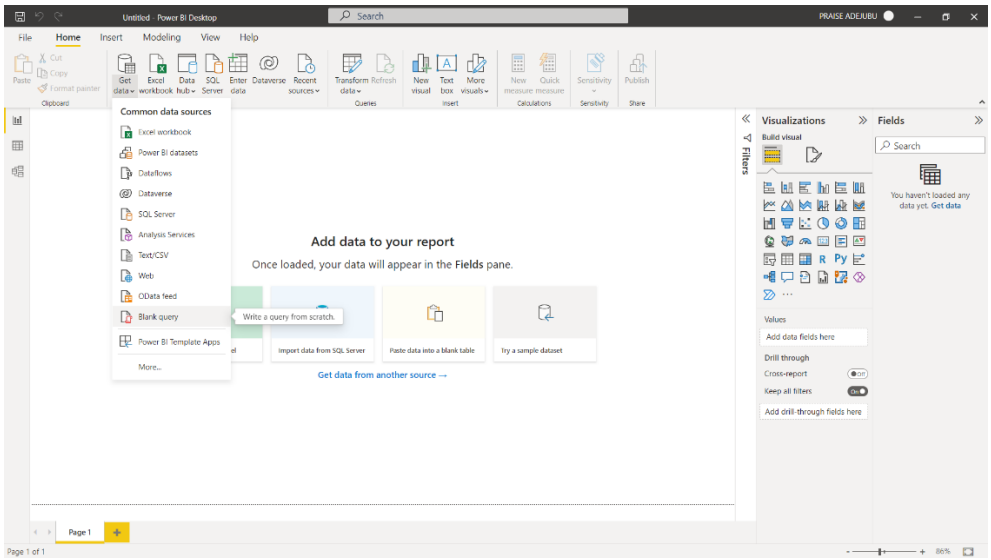


Fig 33. Power BI modelling process

- Next click on Advanced Query:

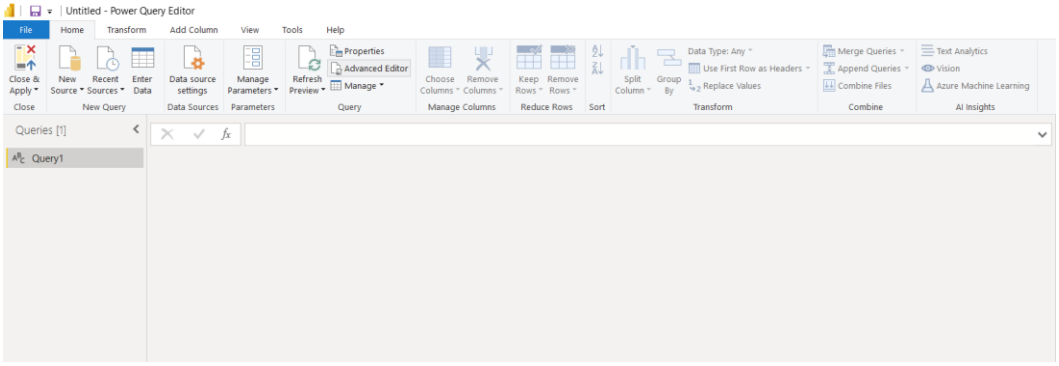


Fig 34. Power BI modelling process

- Type in the source destination as follows:

Fig 35. Power BI modelling process

Now for creating the clubs logo table, I sourced the data from dream league soccer kits website, https://dreamleaguesoccerkits.com/2021/05/premier-league-club-logo/ :

- Click on the "Enter Data" on the Home tab:

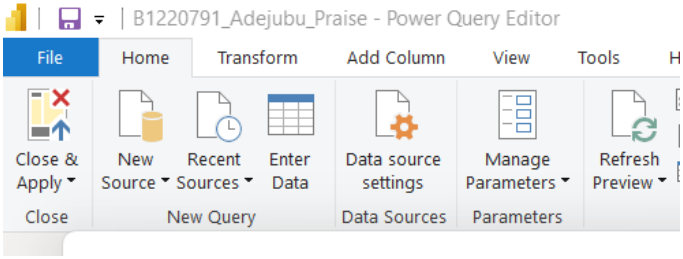


Fig 36. Power BI modelling process

- Go to the URL links and copy them:

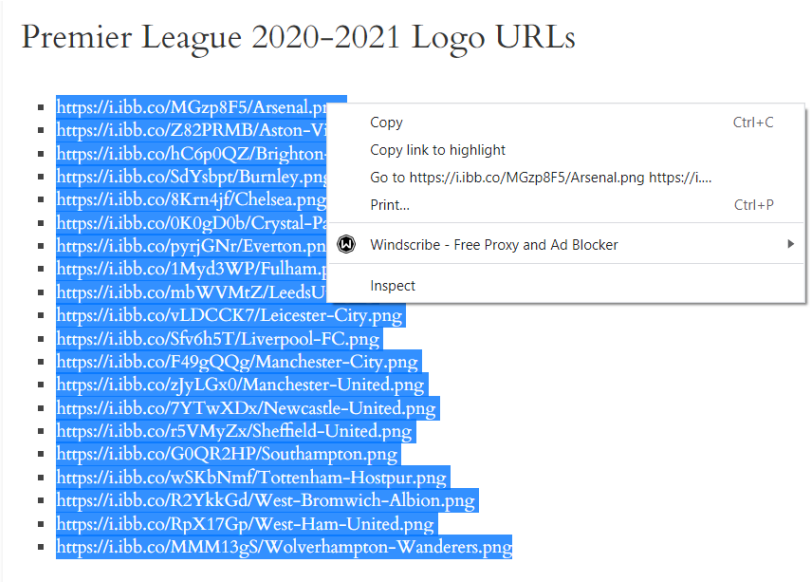


Fig 37. Power BI modelling process

- Paste them in the empty column Table and rename it to Logos, then click "Okay".

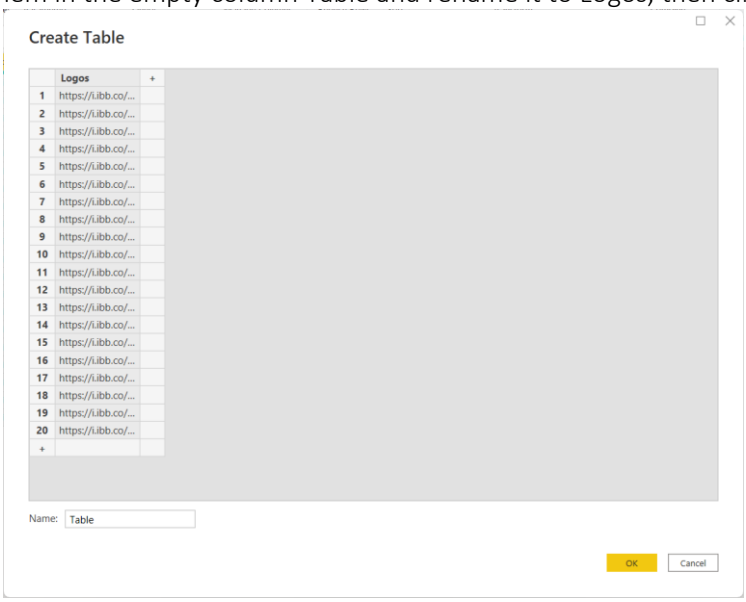


Fig 38. Power BI modelling process

The new table comes in the queries.

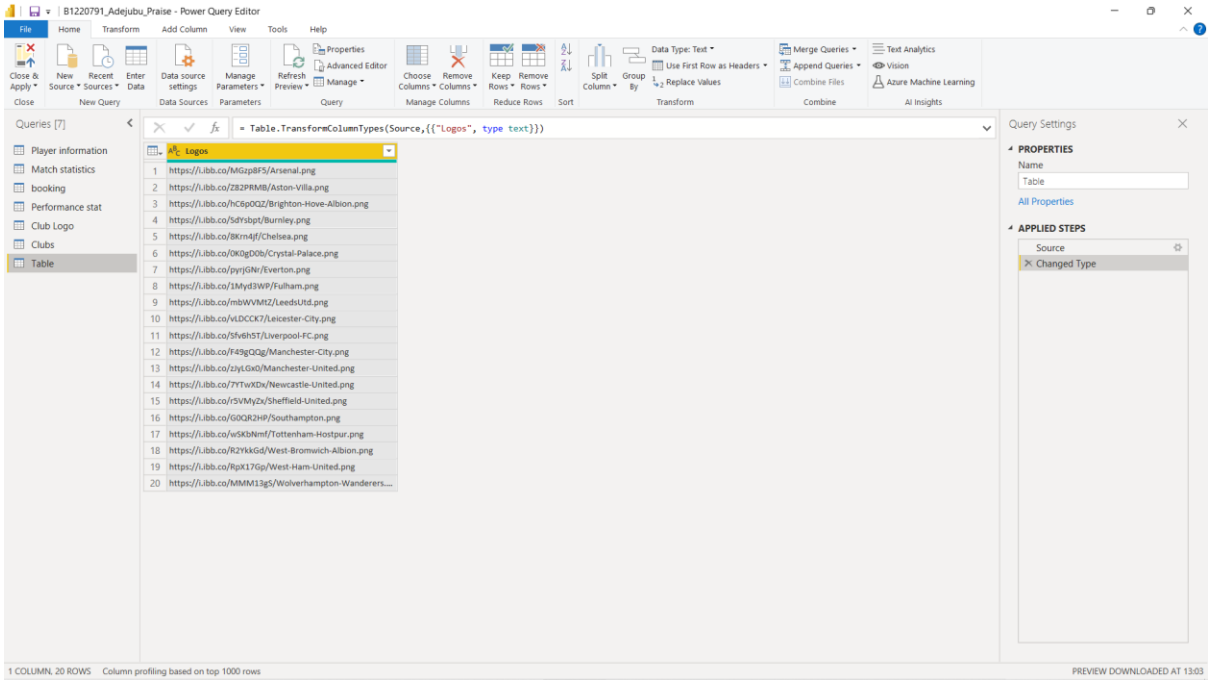


Fig 39. Power BI modelling process

Data Cleaning:

The next step was to clean up the pre-process the data. This will was done in Power query, to do this,

- I clicked on the “Transform data”, on the home tab:

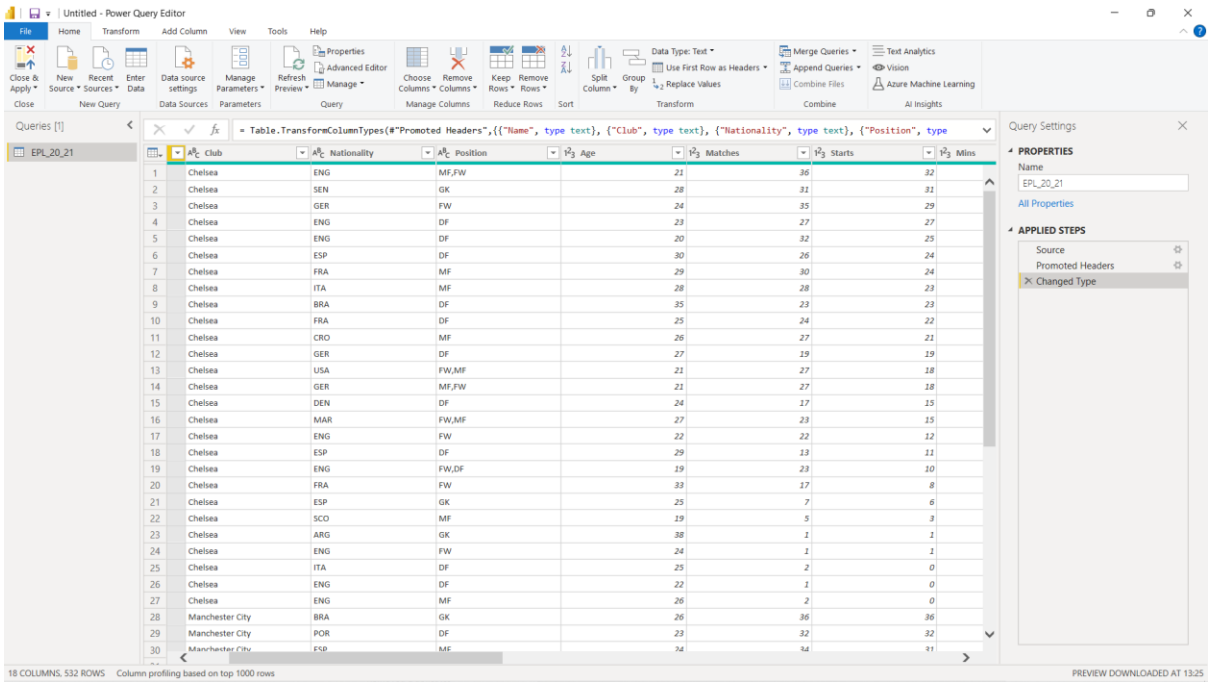


Fig 40. Power BI modelling process

The dataset was cleansed once the previous step of importing it into Power BI was completed, allowing us to use it for our analysis. The cleaning procedure followed the thinking process below.

Promoted Headers: The program recognized the next row on the data as the main headings, and thus Transfers the First row to be the header of the table. In this case it happened automatically given that the original csv file was imported with column 1, column 2… as the headings. This step can be done manually:

- After the table is imported, select the “Use First Row as Header” option on the home tab:

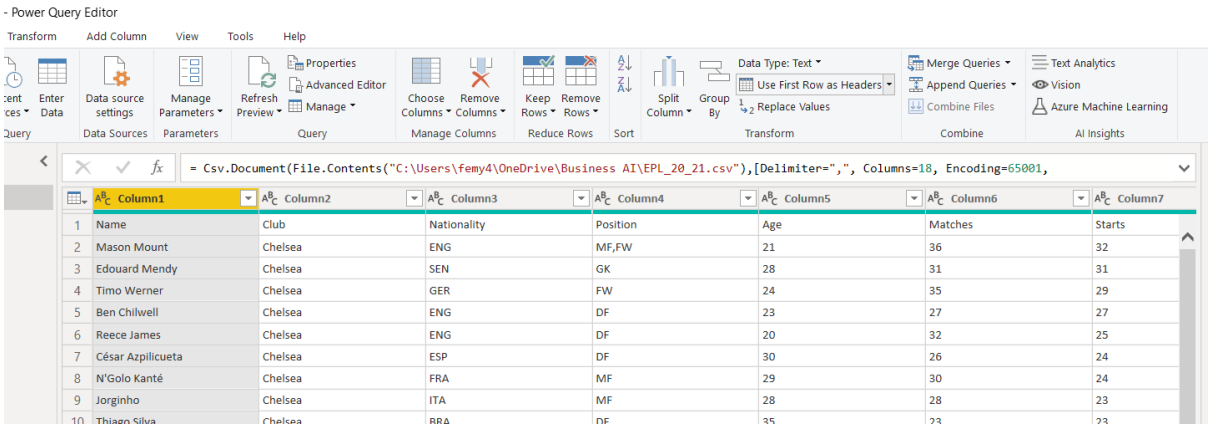


Fig 41. Power BI modelling process

Changed Type: The next cleansing was to change the data types, this is to correctly recognize what each column type it is. This step was also done automatically, but It can be done alternateively using this M formula:

= Table.TransformColumnTypes(#"Promoted Headers",{{"Name", type text}, {"Club", type text}, {"Nationality", type text}, {"Position", type text}, {"Age", Int64.Type}, {"Matches", Int64.Type}, {"Starts", Int64.Type}, {"Mins", Int64.Type}, {"Goals", Int64.Type}, {"Assists", Int64.Type}, {"Passes_Attempted", Int64.Type}, {"Perc_Passes_Completed", type number}, {"Penalty_Goals", Int64.Type}, {"Penalty_Attempted", Int64.Type}, {"xG", type number}, {"xA", type number}, {"Yellow_Cards", Int64.Type}, {"Red_Cards", Int64.Type}})

- This text above should be inputted to the “formula bar” to transform the columns type.

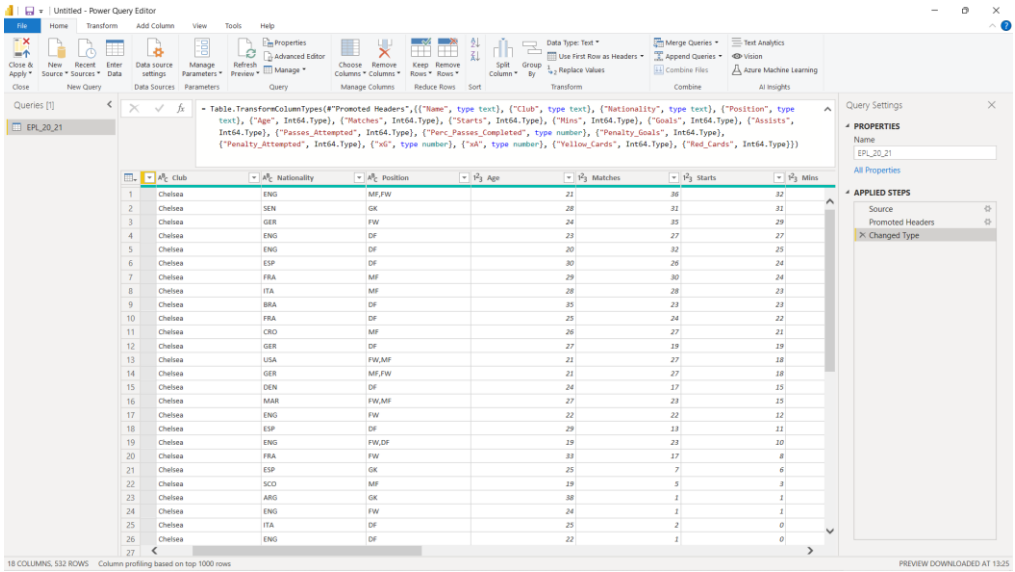


Fig 42. Power BI modelling process

Checking for Nulls, Blanks from the dataset: I checked for empty values in all columns but there were none, all columns were 100 percent.

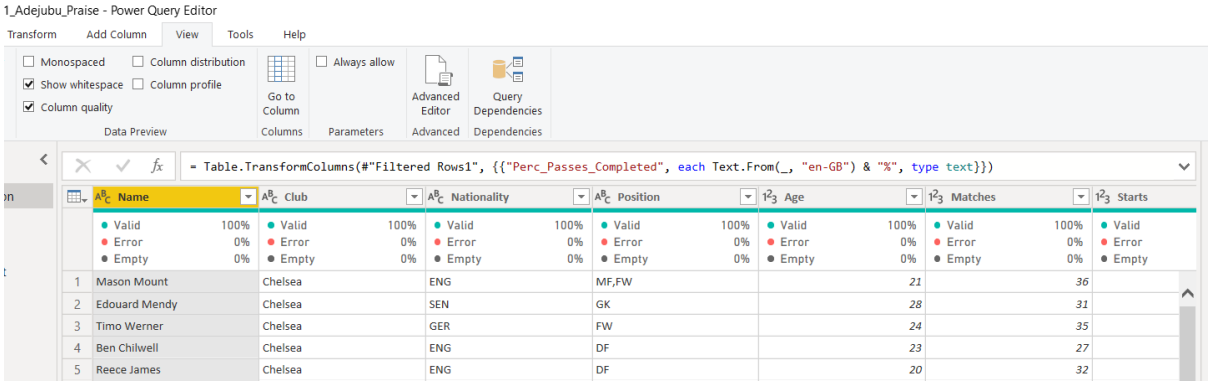


Fig 43. Power BI modelling process

- If there are null values (empty cells) in the dataset, you can replace them with a default value using the "Replace Values" function. To do this, click on the column header with the null values, then select "Replace Values" from the ribbon at the top of the Power BI interface.

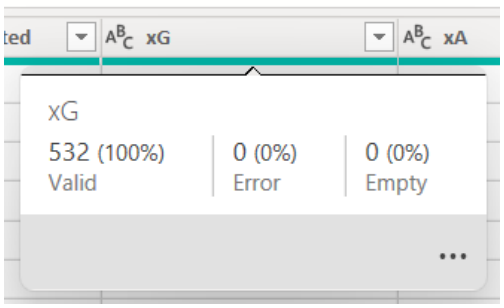
Remove unnecessary columns: If there are columns in the dataset that you do not need for your analysis, you can remove them to reduce the size of the dataset and make it easier to work with. To do this, click on the column header, then select "Remove Columns" from the ribbon at the top of the Power BI interface.

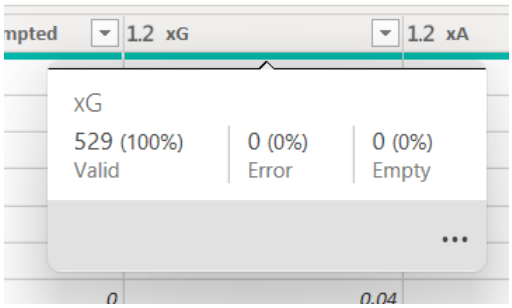
Check and Remove errors from the dataset: I noticed the “xG” and the “Perc_Passes_Completed” columns had wrong values, in that; Research online explains that “xG”, Expected Goals is a metric designed to measure the probability of a shot resulting in a goal. An xG model uses historical information from thousands of shots with similar characteristics to estimate the likelihood of a goal on a scale between 0 and 1, [5] as explained in https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/. So any number above 1 is an error in the xG column.

Also in the “Perc_Passes_Completed” column, there was a -1 value, which is impossible to rate a pass a -1.

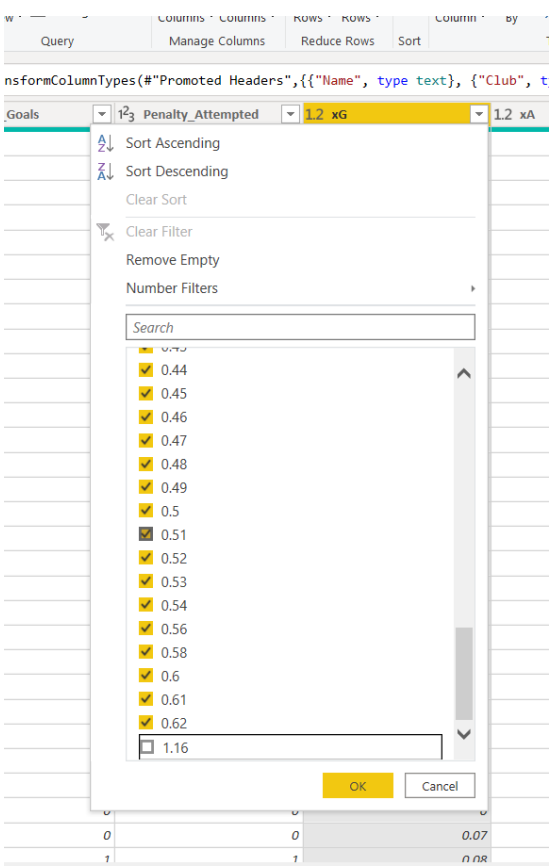
- So the next action was too filter the data using M Language:

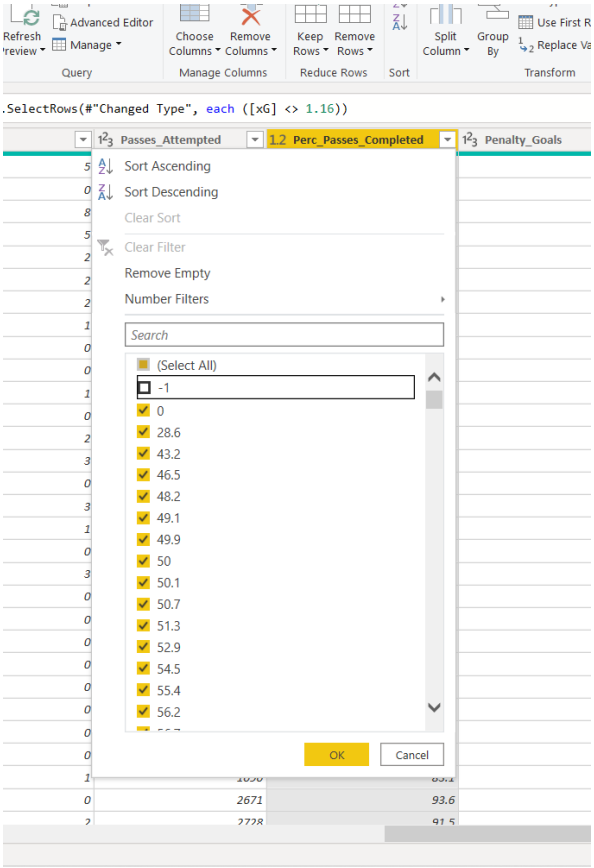
= Table.SelectRows(#"Changed Type", each ([xG] <> 1.16) and ([Perc_Passes_Completed] <> -1))





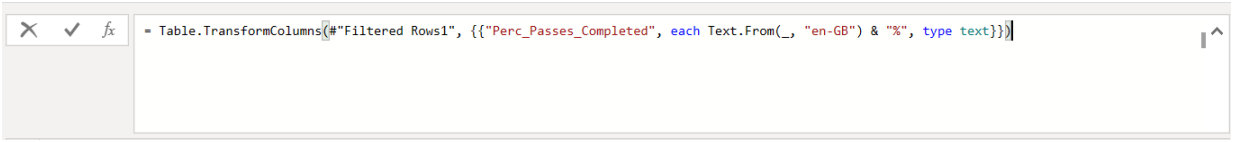
- This can also be done manually, as shown in the illustrations below:





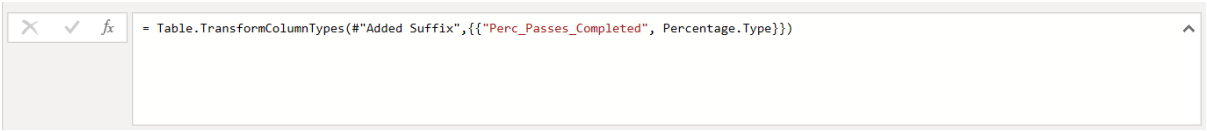
- Next step was to manually change the percentage column by adding “%” to the value and then changing the column type to percentage. The M formula for this is:

= Table.TransformColumns(#"Filtered Rows1", {{"Perc_Passes_Completed", each Text.From(_, "en-GB") & "%", type text}})



- And then to change the column type, type this M formula into the Formular bar and press enter:

= Table.TransformColumnTypes(#"Added Suffix",{{"Perc_Passes_Completed", Percentage.Type}})

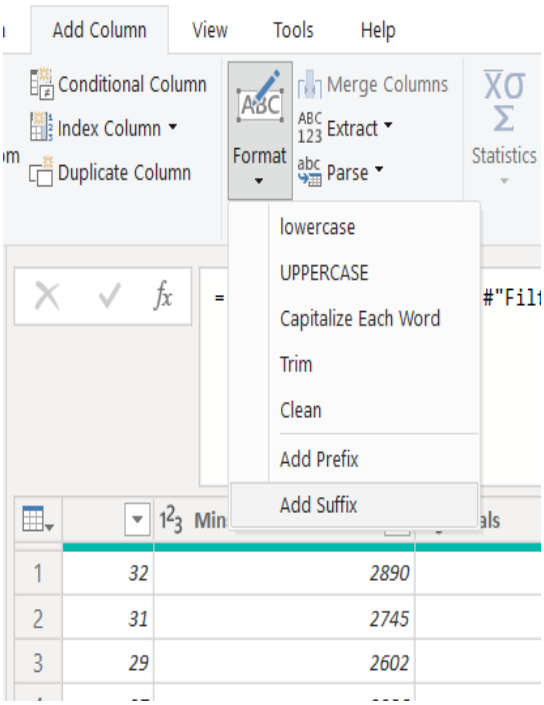


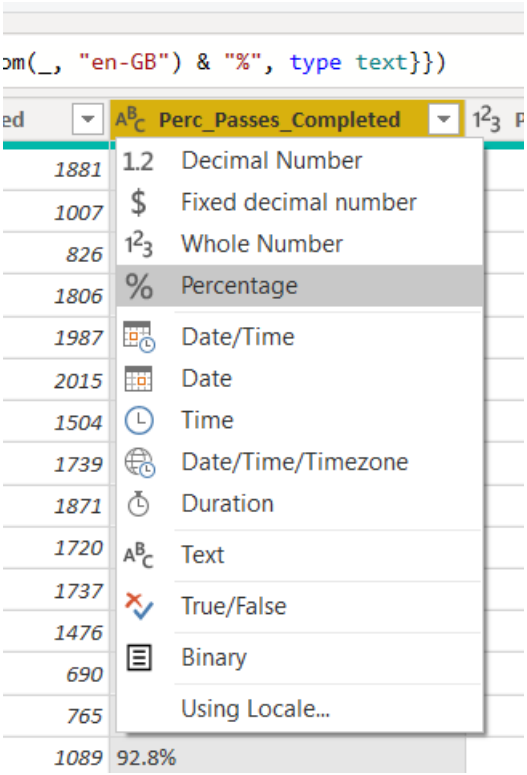
1.2 Perc_Passes_Completed
82.3
84.6
77.2
78.6
85
87.5
86.6
89.5
93.5
91.9
91
90.7

% Perc_Passes_Completed
82.30%
84.60%
77.20%
78.60%
85.00%
87.50%
86.60%
89.50%
93.50%
91.90%
91.00%
90.70%

This can be done manually by firstly selecting the drop menu on “Format” from the Add column bar, them scrolling down to “Add Suffix”.

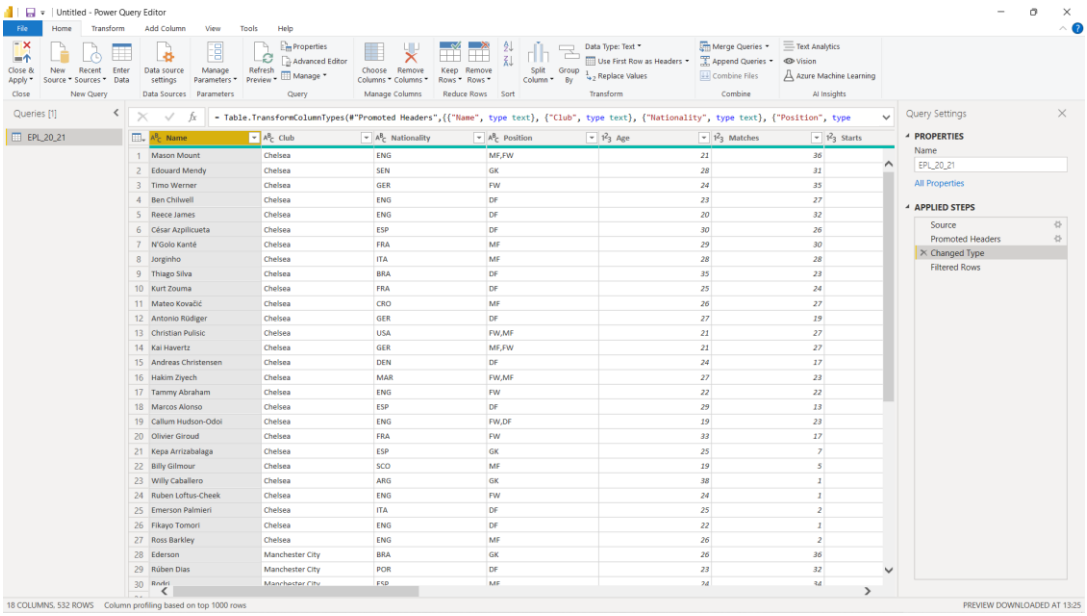
- Next go to the “Perc_Passes_Completed” column Header and click on the left side symbol icon, then choose “Percentage from the list.





Checking for Duplicate rows: I checked for any rows that were repeated in the table datand found none. This means all data on the table are unique and almost ready for analyzing. The steps I used are illustrated below;

- Firstly select the first colum “Name”:



- Then hold ‘Shift’ key on the keyboard and select the last column “Red_Card”:

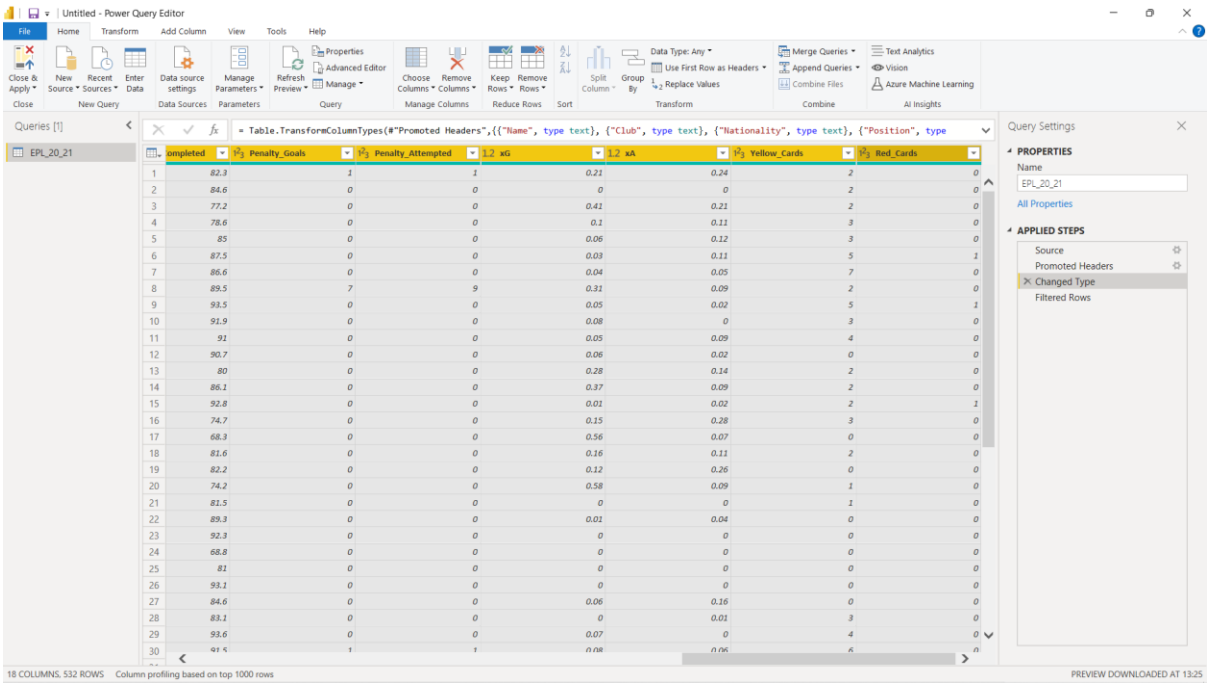
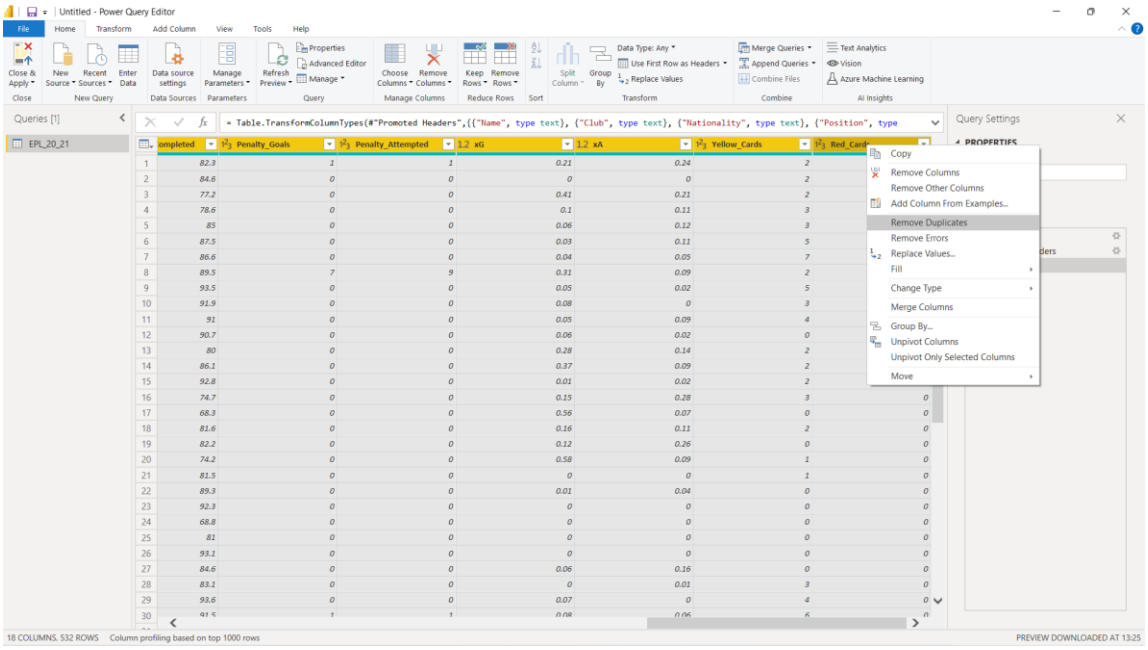


Fig 45. Power BI modelling process

- Next ‘Right click’ the mouse and select “Remove Duplicates” from options:



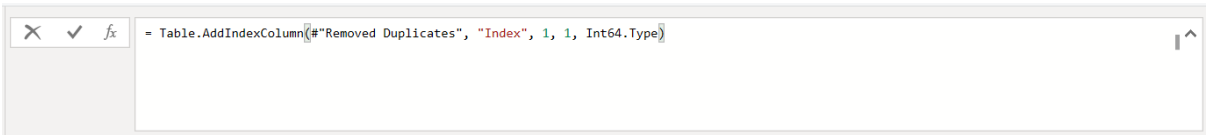
If there are duplicate rows in the dataset, these steps would remove them and the number of Rows would reduce in number.

BI DATA MODELLING VIA STAR SCHEMA - FACTS AND DIMENSIONS

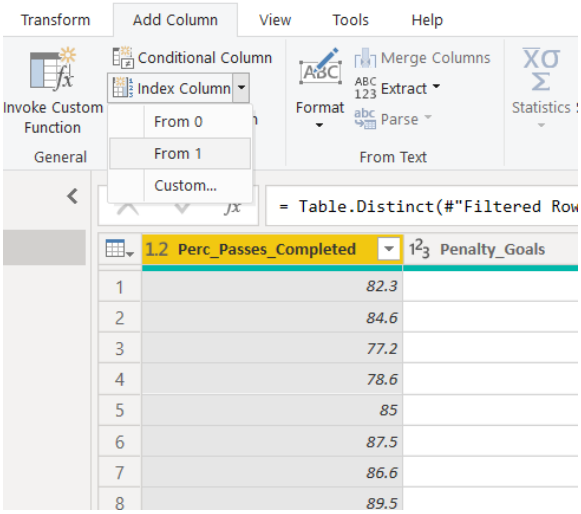
Adding Index column:

The next step was to introduce a column that would help with my relationship throughout my model. A common shared data that would help in the relationship chart.

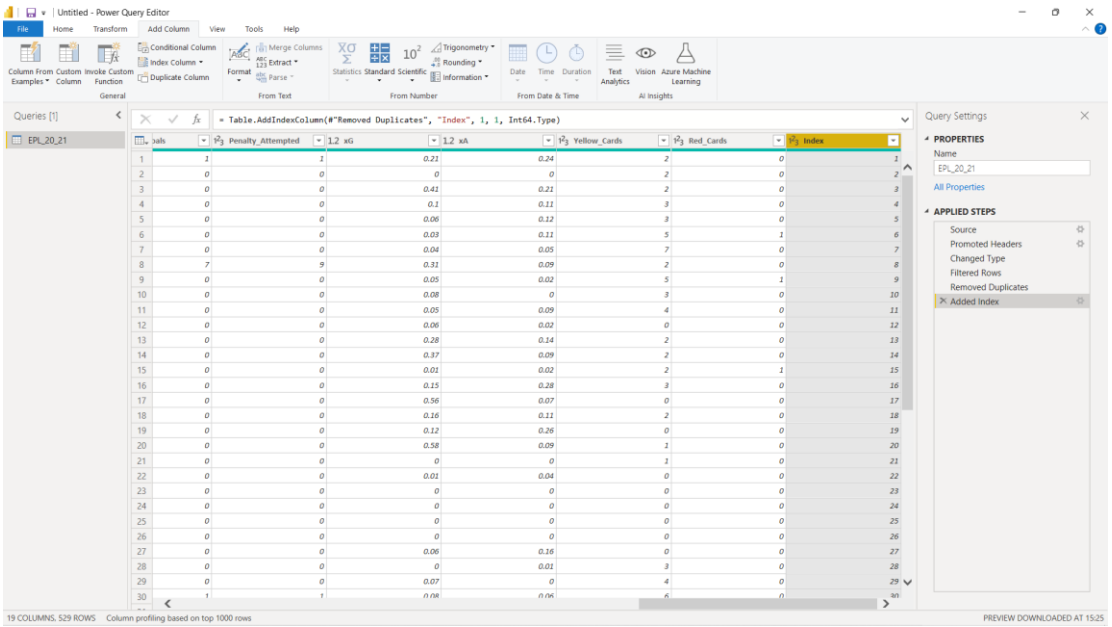
- The M Language formula for this is: = Table.AddIndexColumn(#"Removed Duplicates", "Index", 1, 1, Int64.Type)



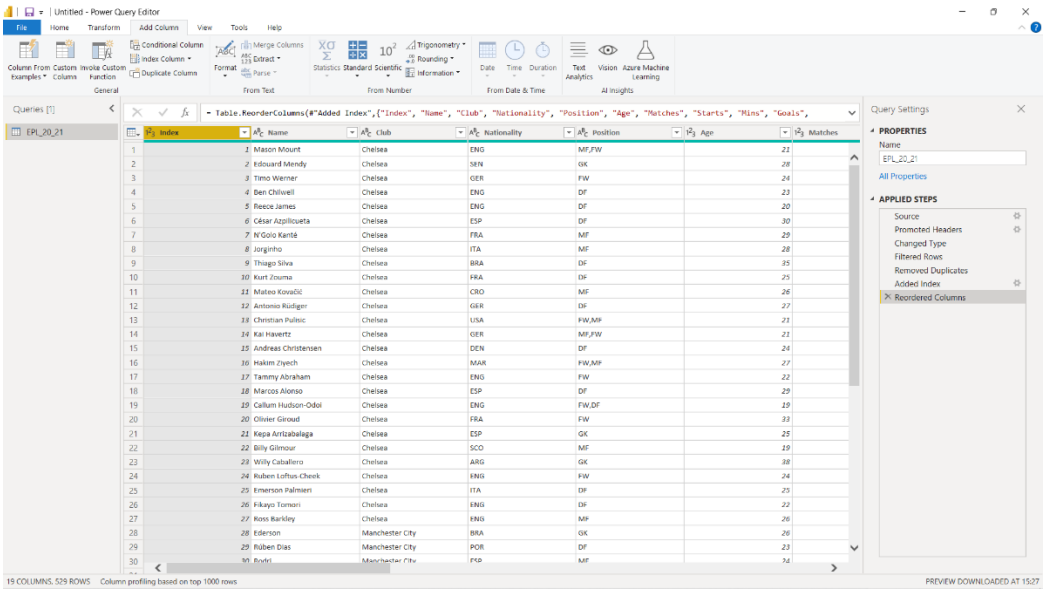
- Alternatively, this can also be done: goto “IndexColumn”, on the “Add Column” bar, select “From 1”.



- A new column “Index”, will appear at the end of the table, as shown below:

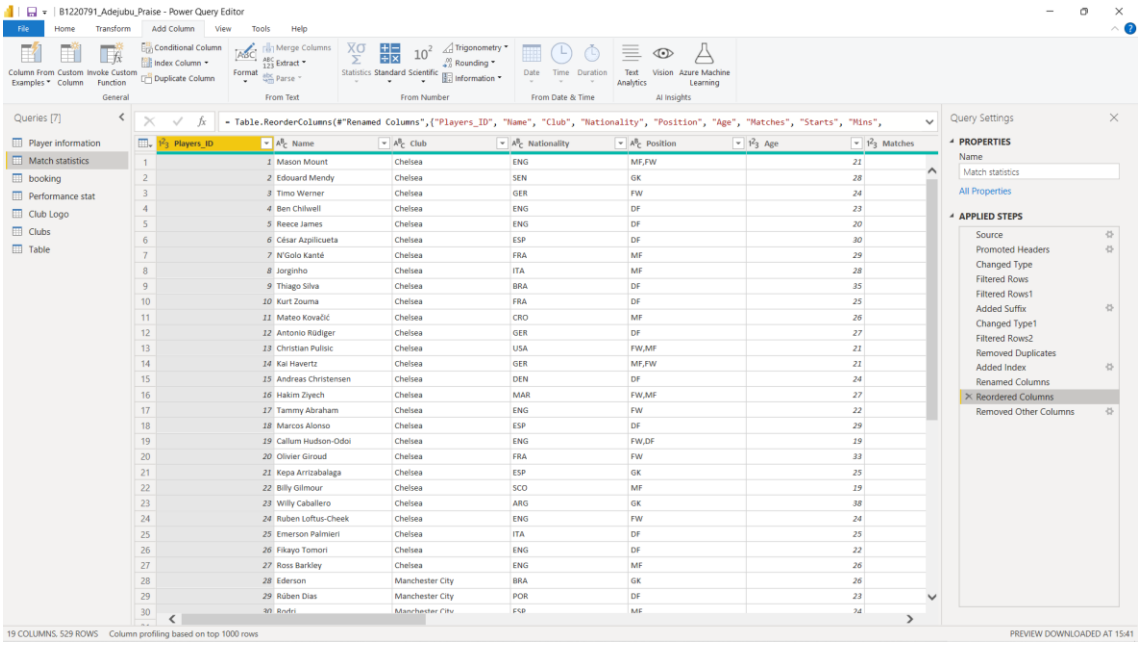


Reordering column: Manually drag the Index column to the start of the table, just before the “Name” column.



Rename columns: Now the “Index” column is not descriptive or easy to understand, I renamed it to “Player_ID”, to make the data easier to work with. To do this, ‘Right click’ on the “Index” column header, then select "Rename" from the list.

- This can easily be done using M language also with the formula: = Table.RenameColumns(#"Added Index",{{"Index", "Players_ID"}}).



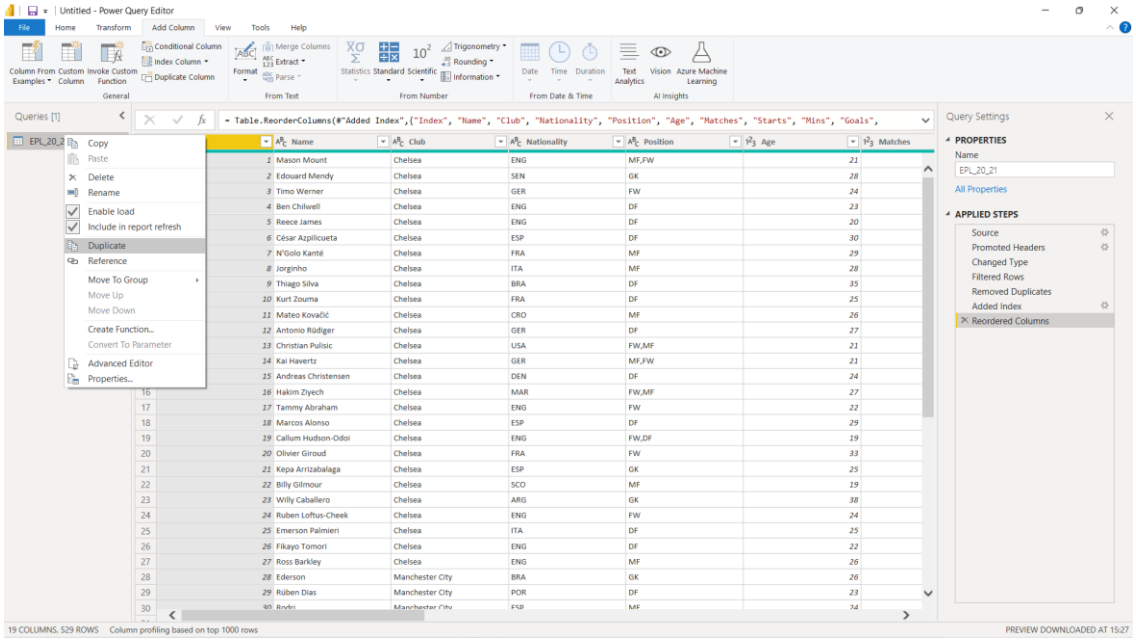
These were the steps I used in cleaning the data to be ready for analysing and by following these steps, I effectively cleaned this dataset for modelling and analysis in Power BI.

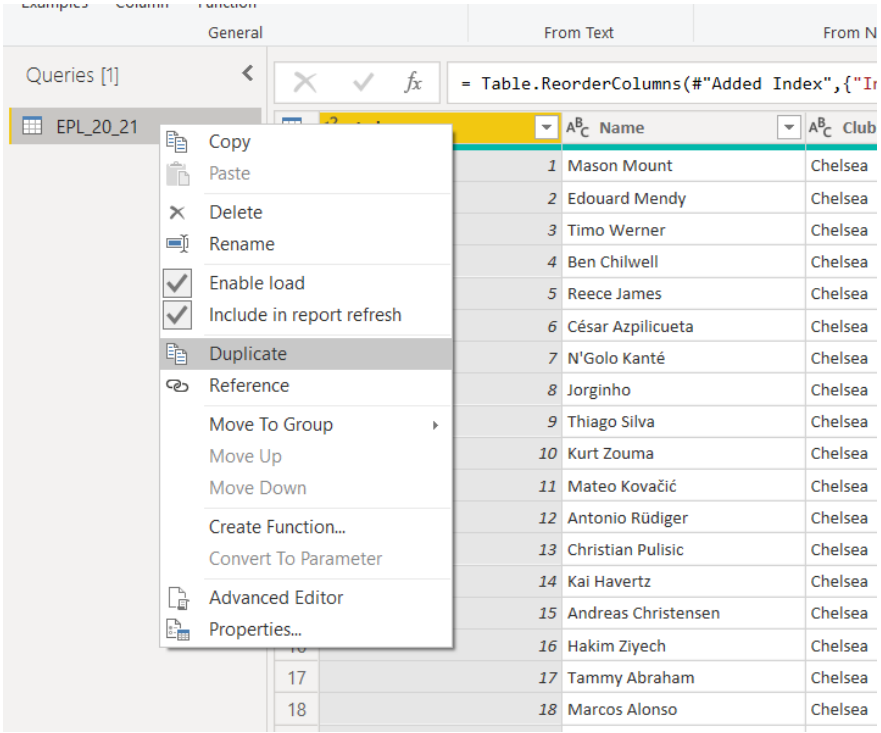
Creating Dimension tables

I duplicated the table severally and removed unwanted columns to create my Fact and Dimension tables.

Duplicating Tables: To create the dimension tables, after completing the pre-processing, I duplicated the main table now and the steps are shown below:

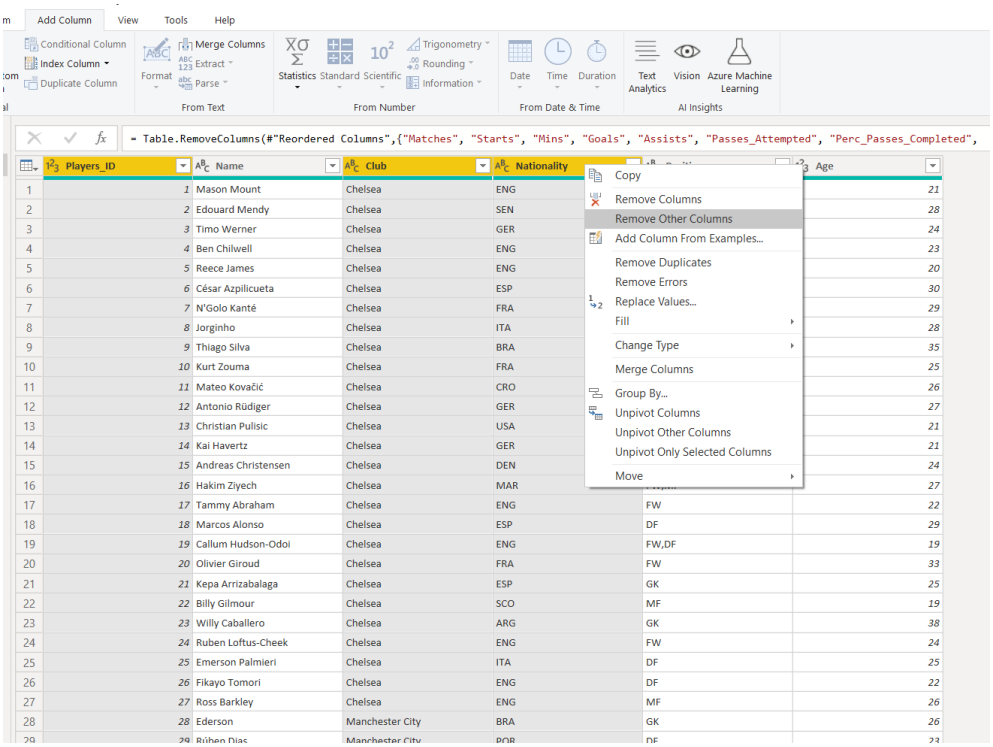
- 'Right' click on the Table "EPL_20_21!" in the "Queries" Station, by the far left, and select "Duplicate" from the list drop down.





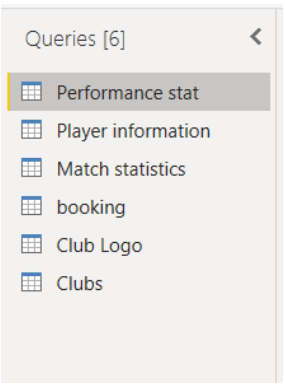
Removing unwanted Columns for Dimension tables: This is how I removed columns from tables is illustrated using M Language: I used this formula: = Table.RemoveColumns(#"Reordered Columns",{"Matches", "Starts", "Mins", "Goals", "Assists", "Passes_Attempted", "Perc_Passes_Completed", "Penalty_Goals", "Penalty_Attempted", "xG", "xA", "Yellow_Cards", "Red_Cards"})

- Adding any column into the list above to remove them. To do this, just select the columns wanted or the once not wanted (by holding down 'Ctrl' key and 'Right click' on one of them, select "Remove Other Columns or Remove Columns in whichever preference.

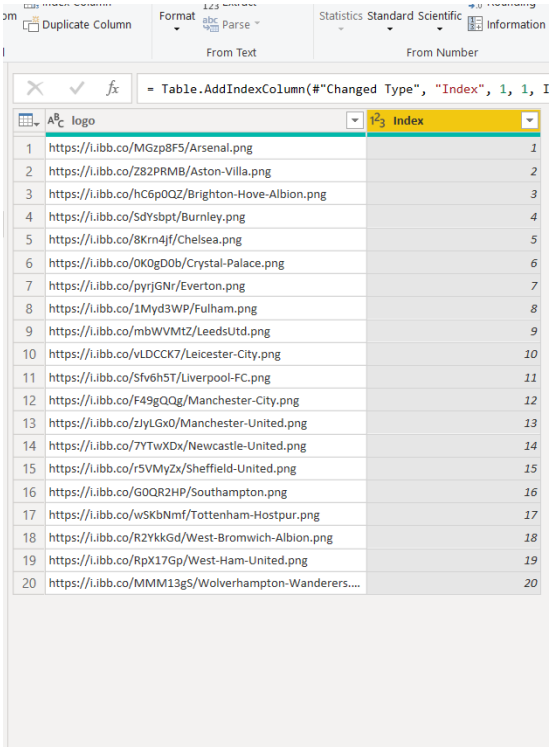


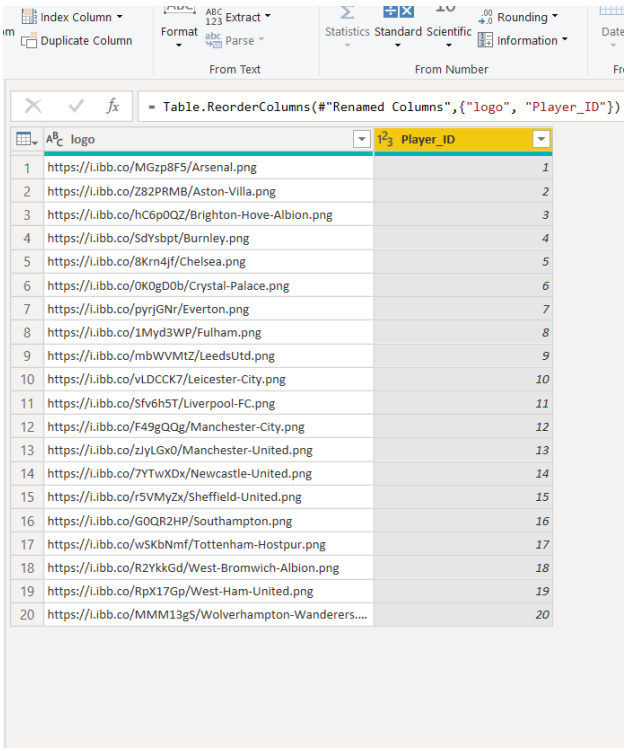
	Players_ID	Name	Club	Nationality		Age
1	1	Mason Mount	Chelsea	ENG		21
2	2	Edouard Mendy	Chelsea	SEN		28
3	3	Timo Werner	Chelsea	GER		24
4	4	Ben Chilwell	Chelsea	ENG		23
5	5	Reece James	Chelsea	ENG		20
6	6	César Azpilicueta	Chelsea	ESP		30
7	7	N'Golo Kanté	Chelsea	FRA		29
8	8	Jorginho	Chelsea	ITA		28
9	9	Thiago Silva	Chelsea	BRA		35
10	10	Kurt Zouma	Chelsea	FRA		25
11	11	Mateo Kovačić	Chelsea	CRO		26
12	12	Antonio Rüdiger	Chelsea	GER		27
13	13	Christian Pulisic	Chelsea	USA		21
14	14	Kai Havertz	Chelsea	GER		21
15	15	Andreas Christensen	Chelsea	DEN		24
16	16	Hakim Ziyech	Chelsea	MAR		27
17	17	Tammy Abraham	Chelsea	ENG	FW	22
18	18	Marcos Alonso	Chelsea	ESP	DF	29
19	19	Callum Hudson-Odoi	Chelsea	ENG	FW,DF	19
20	20	Olivier Giroud	Chelsea	FRA	FW	33
21	21	Kepa Arrizabalaga	Chelsea	ESP	GK	25
22	22	Billy Gilmour	Chelsea	SCO	MF	19
23	23	Willy Caballero	Chelsea	ARG	GK	38
24	24	Ruben Loftus-Cheek	Chelsea	ENG	FW	24
25	25	Emerson Palmieri	Chelsea	ITA	DF	25
26	26	Fikayo Tomori	Chelsea	ENG	DF	22
27	27	Ross Barkley	Chelsea	ENG	MF	26
28	28	Ederson	Manchester City	BRA	GK	26
29	29	Rúben Dias	Manchester City	POR	DF	23

- I repeated this step for all four Dimension tables and the one bridge table.

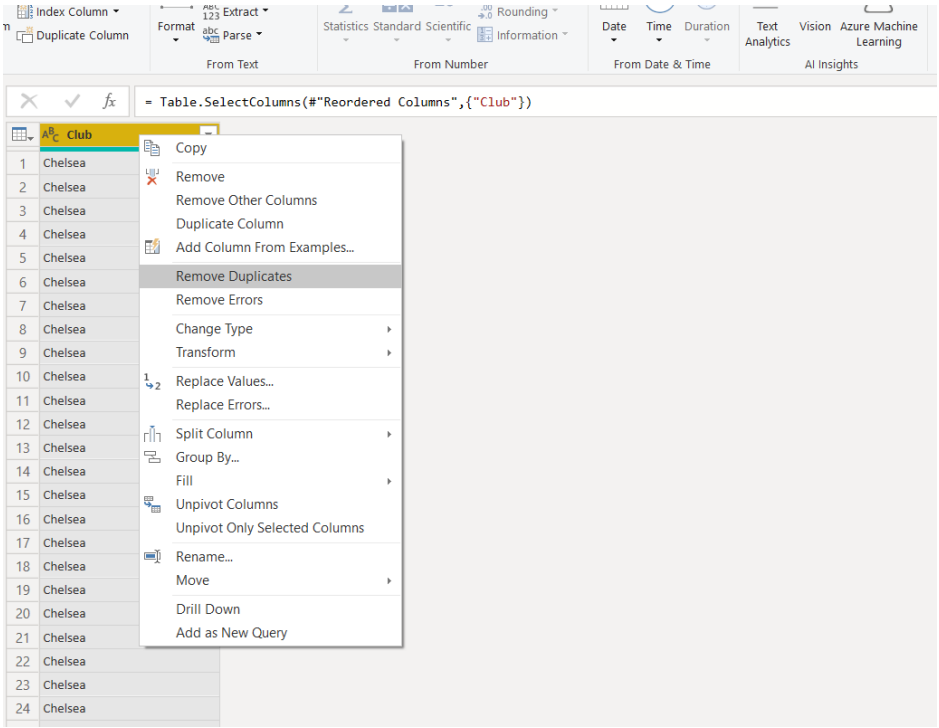


- There is one table that I created using external data to get the Logos of each clubs. After loading the url's into the table, I just added an "index" table to it and renamed it "Players_ID" in order to the share data with the "Clubs" data.



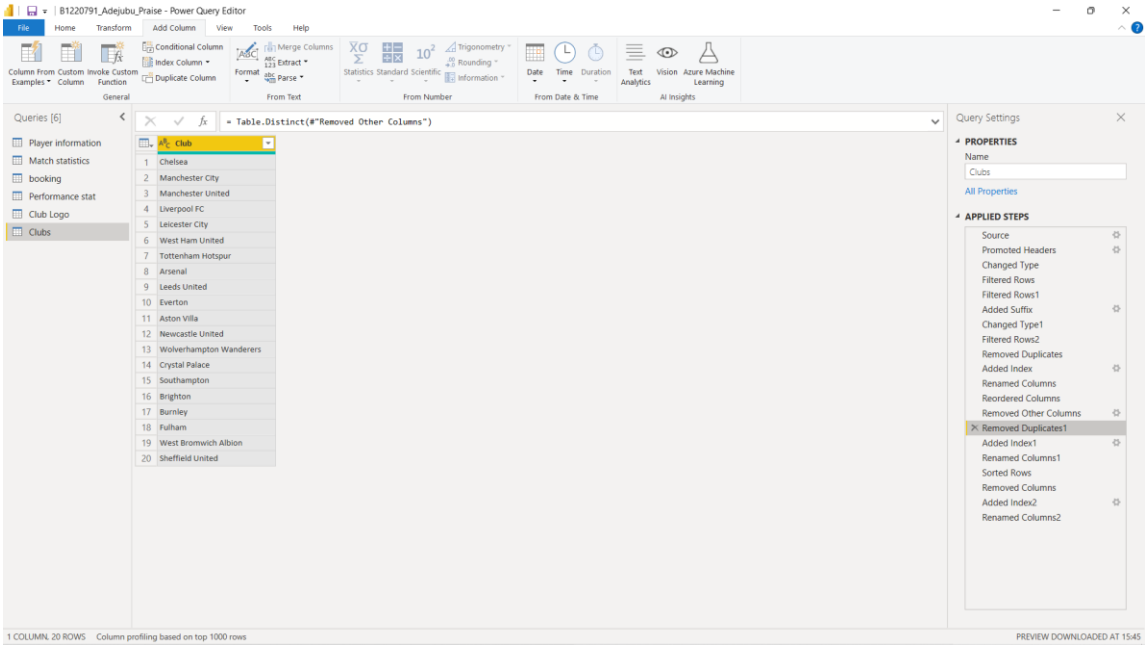


The next and final process was to create a separate "Clubs" table to Bridge/link the "Club logos" table data to the rest of the model. I achieved this using these steps:



I duplicated the original table and removed all columns except the “Club” column as shown in the column above.

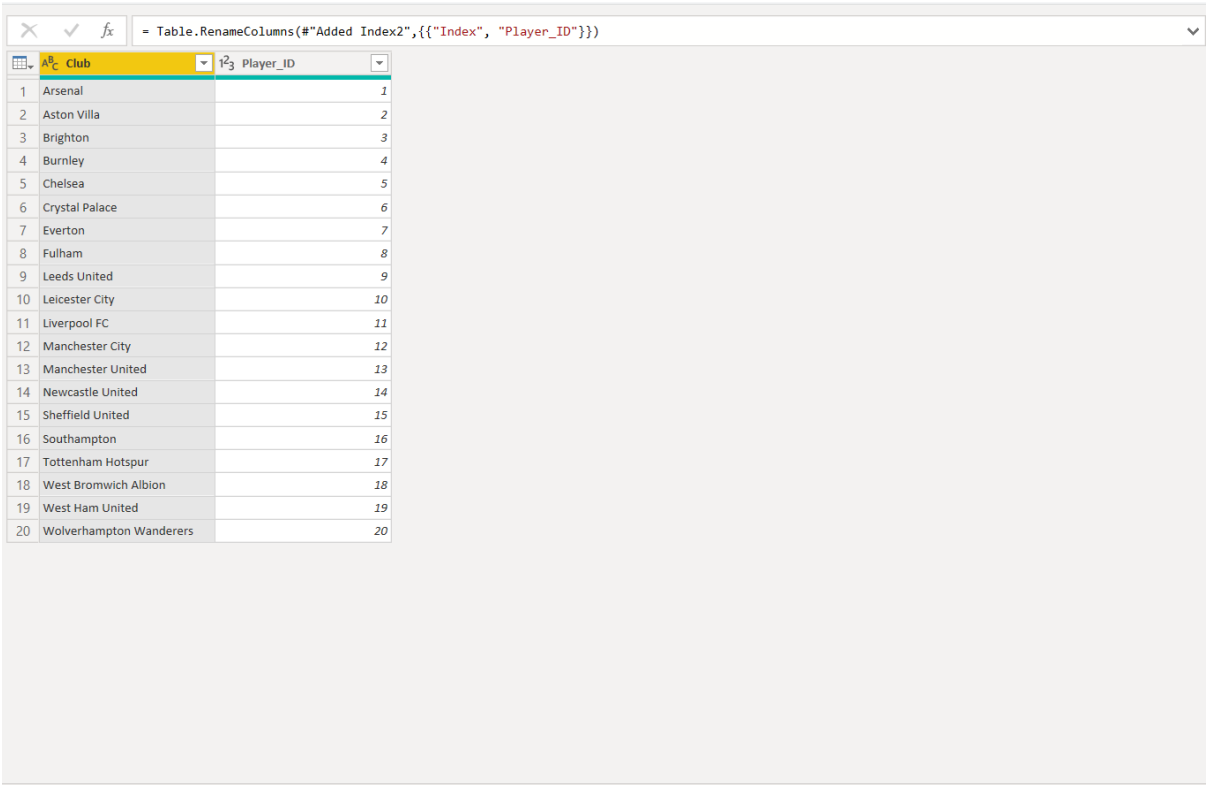
- Next, I sorted the “club” column in Ascending order, using M Language: = Table.Sort(#"Renamed Columns1",{{"Club", Order.Ascending}})



- Finally, adding an “Index” column and renaming it “Player_ID” with these M Formulas:

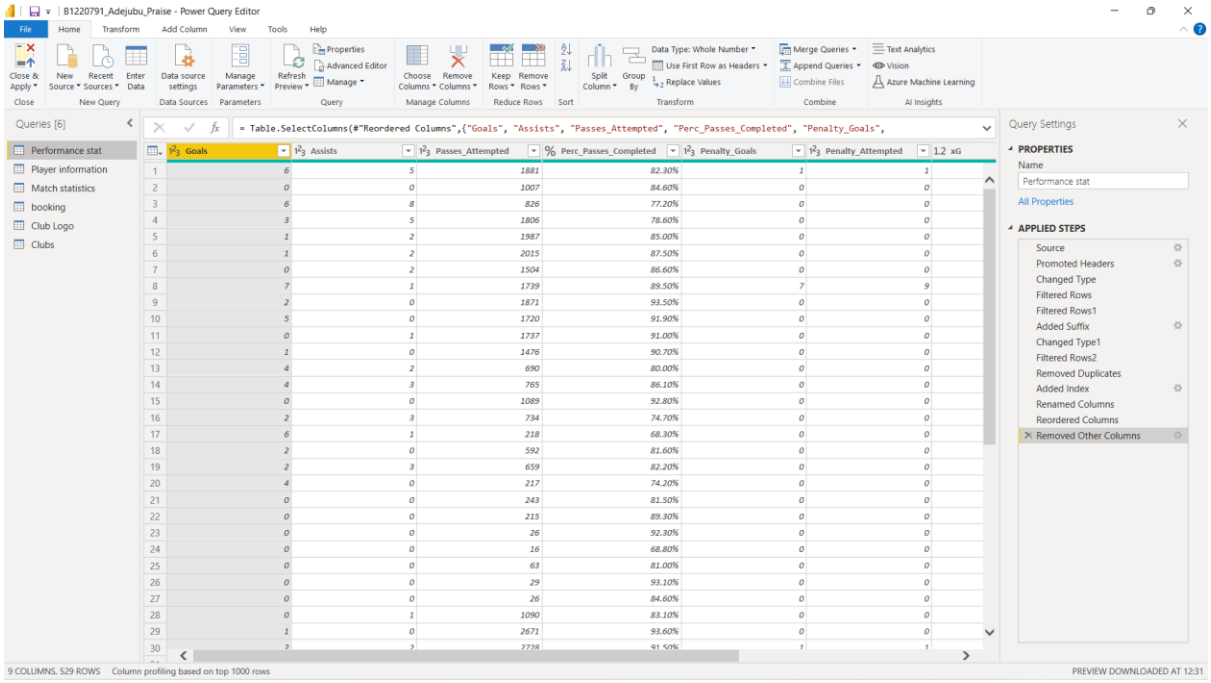
#"Added Index2" = Table.AddIndexColumn(#"Removed Columns", "Index", 1, 1, Int64.Type),

#"Renamed Columns2" = Table.RenameColumns(#"Added Index2",{{"Index", "Player_ID"}})

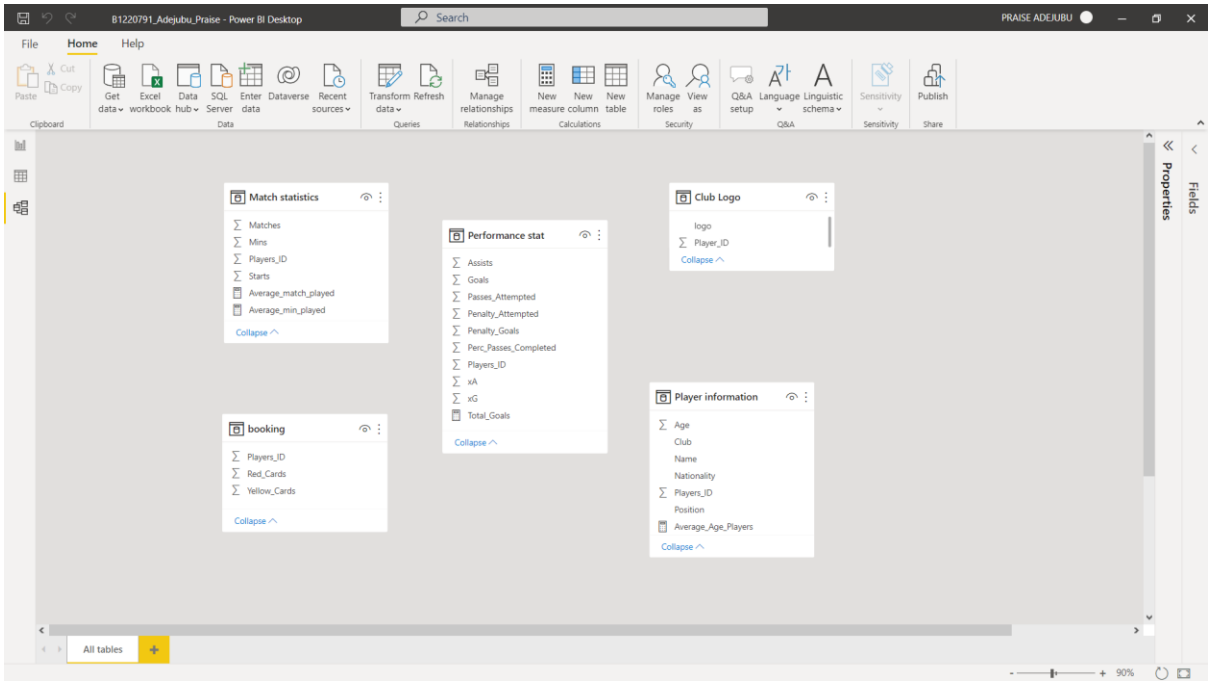


	Club	Player_ID
1	Arsenal	1
2	Aston Villa	2
3	Brighton	3
4	Burnley	4
5	Chelsea	5
6	Crystal Palace	6
7	Everton	7
8	Fulham	8
9	Leeds United	9
10	Leicester City	10
11	Liverpool FC	11
12	Manchester City	12
13	Manchester United	13
14	Newcastle United	14
15	Sheffield United	15
16	Southampton	16
17	Tottenham Hotspur	17
18	West Bromwich Albion	18
19	West Ham United	19
20	Wolverhampton Wanderers	20

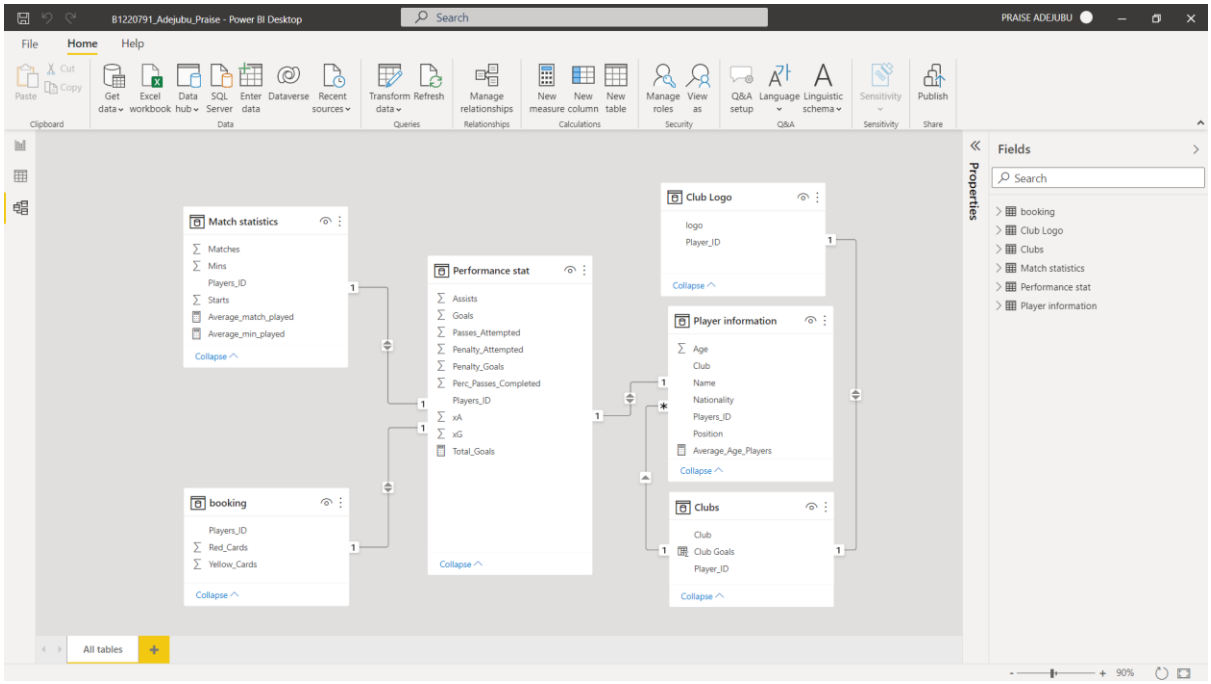
This created a way for my model to share a relationship, using the “Player_ID” table.



- To analyze the data for this project, relationships must be established between the tables in the model. To create these relationships, we will select "Manage relationships" from the home tab of the data model section. This will open a dialog box, and we will select "New" to create a new relationship.



Once all the relationships have been created, we will have a Star schema model where one fact table (Performance stat) is connected to several dimension tables (Player Information, Match statistics, Bookings, Club Logo, and Clubs). The Clubs table will serve as a bridge table between the Club Logo table and other tables.



C. DAX and M Language:

Several measures were created using DAX to calculate the rankings for each metric in the visualizations. The following DAX measures and column were created:

	DAX Description	DAX Formula
1	DAX calculated column "Club Goals" This is to calculate the total number of goals for each club and create a column.	`Club Goals = SUMX(DISTINCT('Clubs'[Club]),CALCULATE(SUM('Performance stat'[Goals])))`
2	DAX calculated measure of "Average_match_played". This is to create a measure to get every players average number of matches played that can be called for visualization without taking computing space.	`Average_match_played = AVERAGE('Match statistics'[Matches])`

3	DAX calculated measure of "Average_min_played" This is to create a measure to get the average number of minutes played by each player.	`Average_min_played = AVERAGE('Match statistics'[Mins])`
4	DAX calculated measure of "Average_Age_Players". This measure calculates the average age of players across the clubs in the league	`Average_Age_Players = AVERAGE('Player information'[Age])`

Table. 3 DAX Formulas and descriptions:

M Language was used to clean and transform the data before creating the star schema. These are some of the M Language formulas used in the model:

	Description	M Language Formula sample in model
1	To filter errors from the data using M Language from the ([xG] and "Perc_Passes_Completed" table at the same time.	= Table.SelectRows(#"Changed Type", each ([xG] <> 1.16) and ([Perc_Passes_Completed] <> -1))
2	Adding suffix "%" to a "Perc_Passes_Completed" column	= Table.TransformColumns(#"Filtered Rows1", {{"Perc_Passes_Completed", each Text.From(_, "en-GB") & "%", type text}})
3	Changing the "Perc_Passes_Completed" column type to percentage.	= Table.TransformColumnTypes(#"Added Suffix",{{"Perc_Passes_Completed", Percentage.Type}})
4	Creating an index column to the table. In this case the "Player_ID" column was introduced to the table as a relationship tool.	= Table.AddIndexColumn(#"Removed Duplicates", "Index", 1, 1, Int64.Type)
5	Renaming columns: Here the "Index" column is not descriptive or easy to understand, I renamed it to "Player_ID", to make the data easier to work with.	= Table.RenameColumns(#"Added Index",{{"Index", "Players_ID"}}).

Table. 3 DAX Formulas and descriptions:

DASHBOARDS

The pages were organized by metric, with each page containing visualizations showing general and vital information's for the 2020/2021 English premier league season for each club.

The Power BI dashboard for this analysis included the following pages:

Fig. Power Dashboard page 1

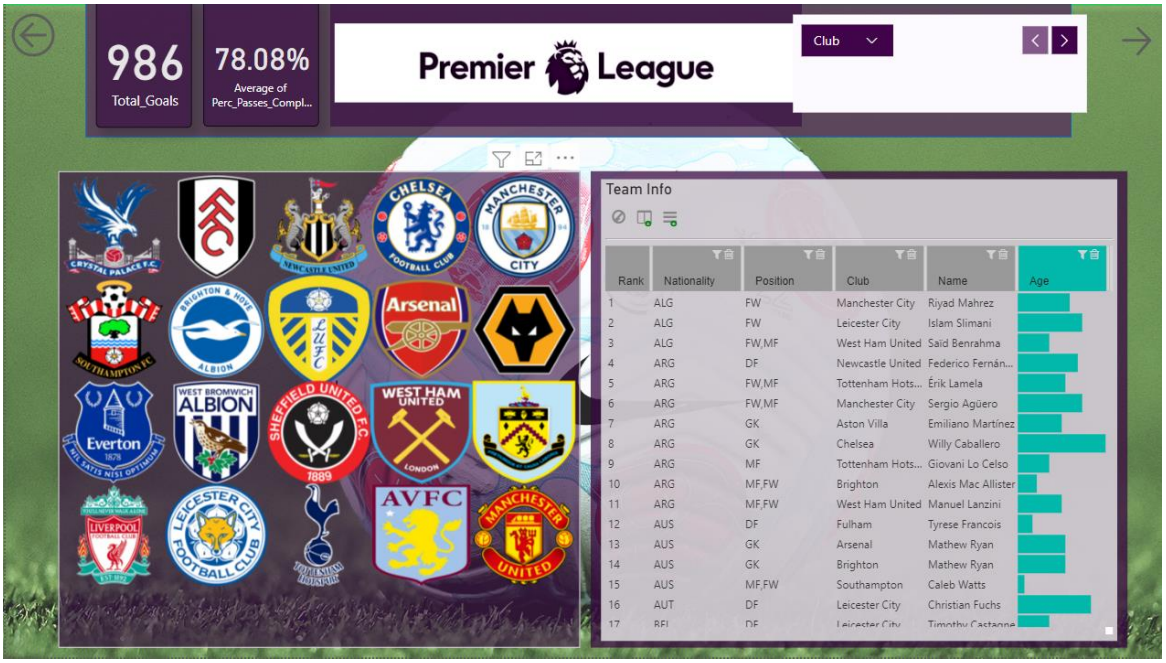


Fig. Power Dashboard page 2

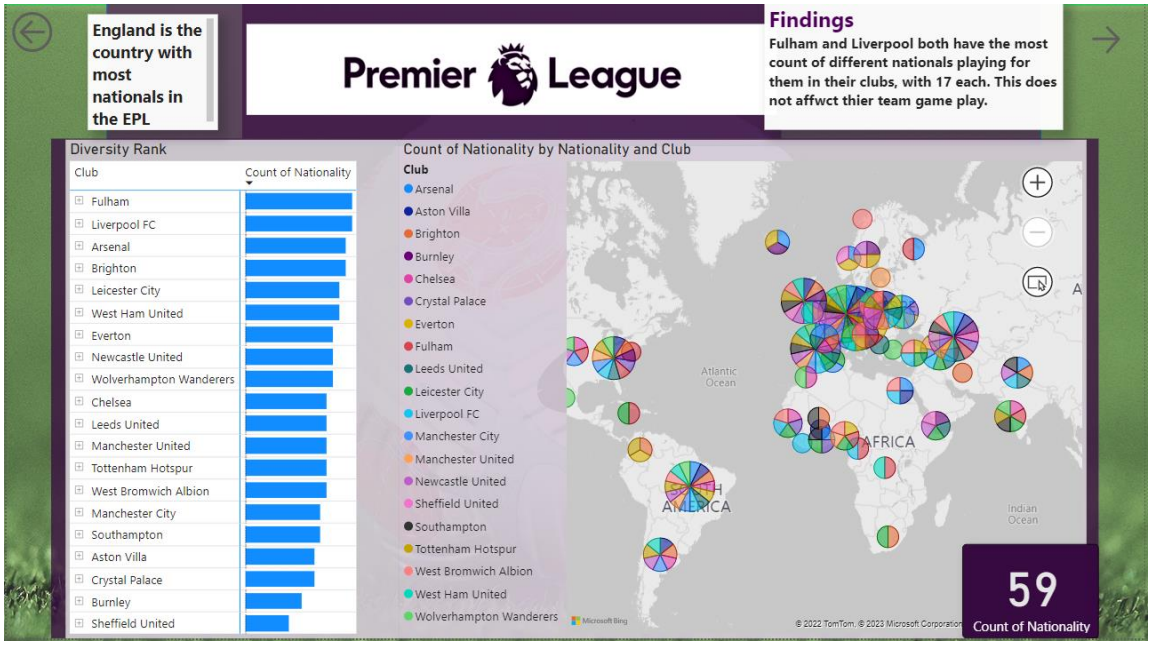


Fig. Power Dashboard page 3

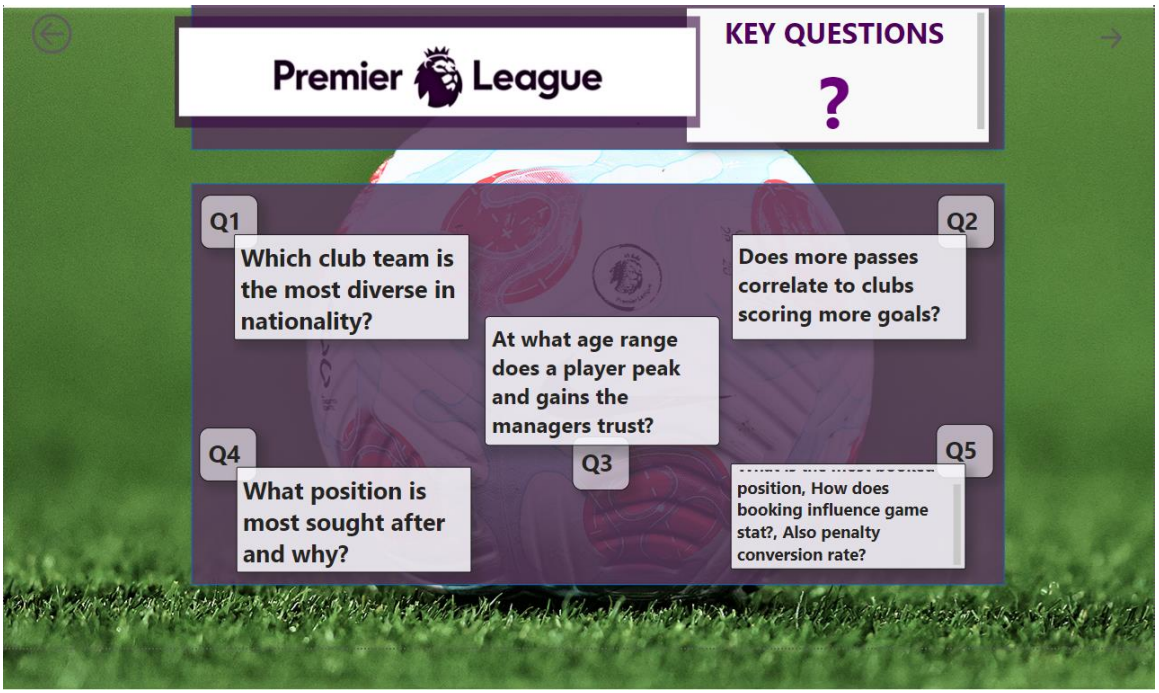


Fig. Power Dashboard page 4

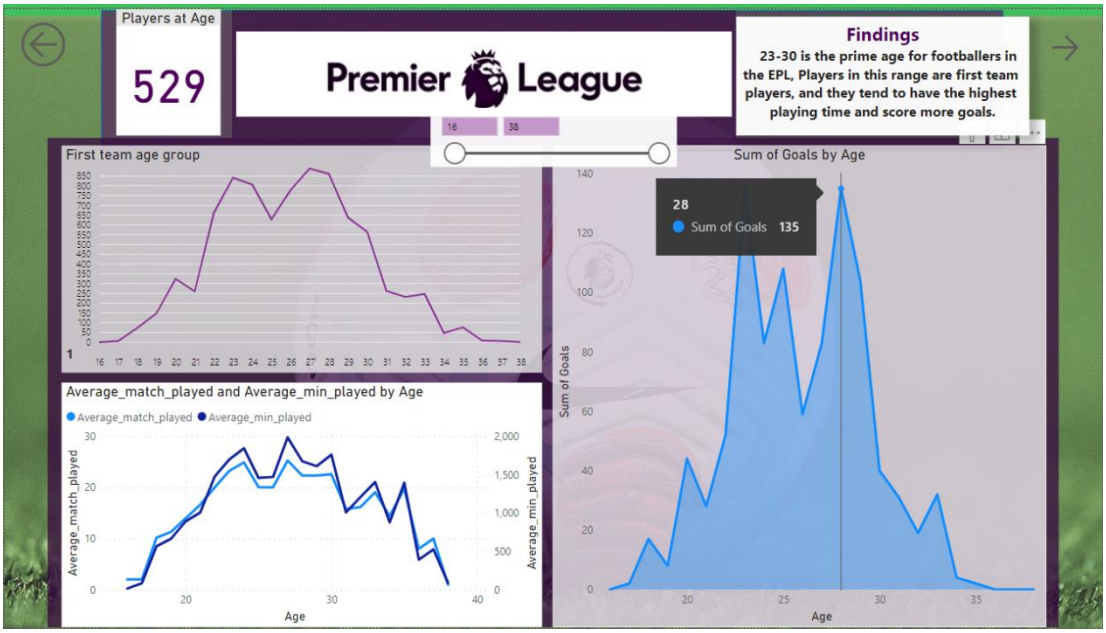


Fig. Power Dashboard page 5

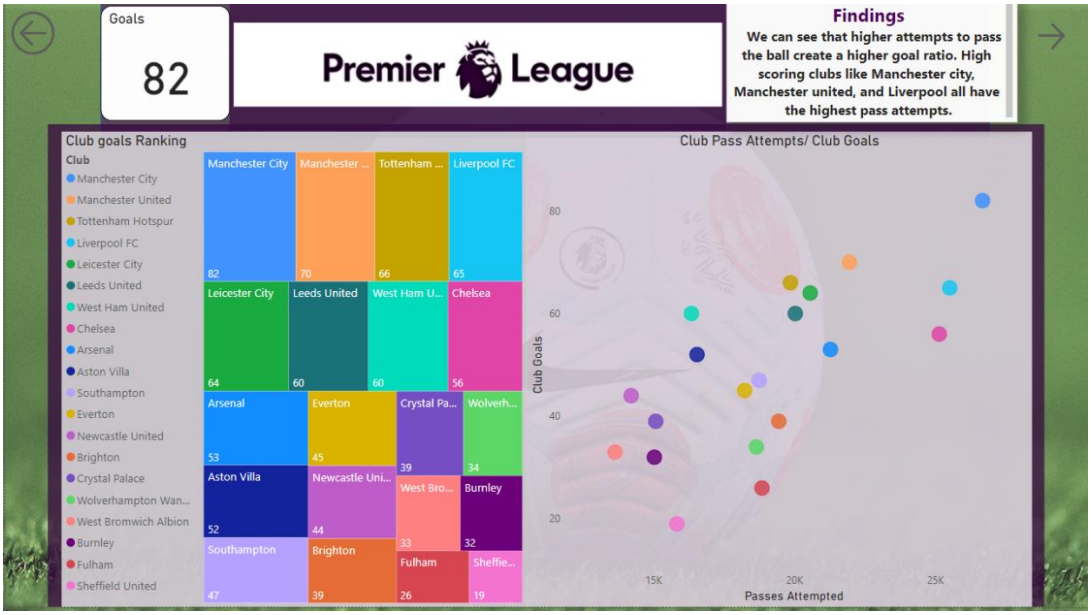


Fig. Power Dashboard page 6

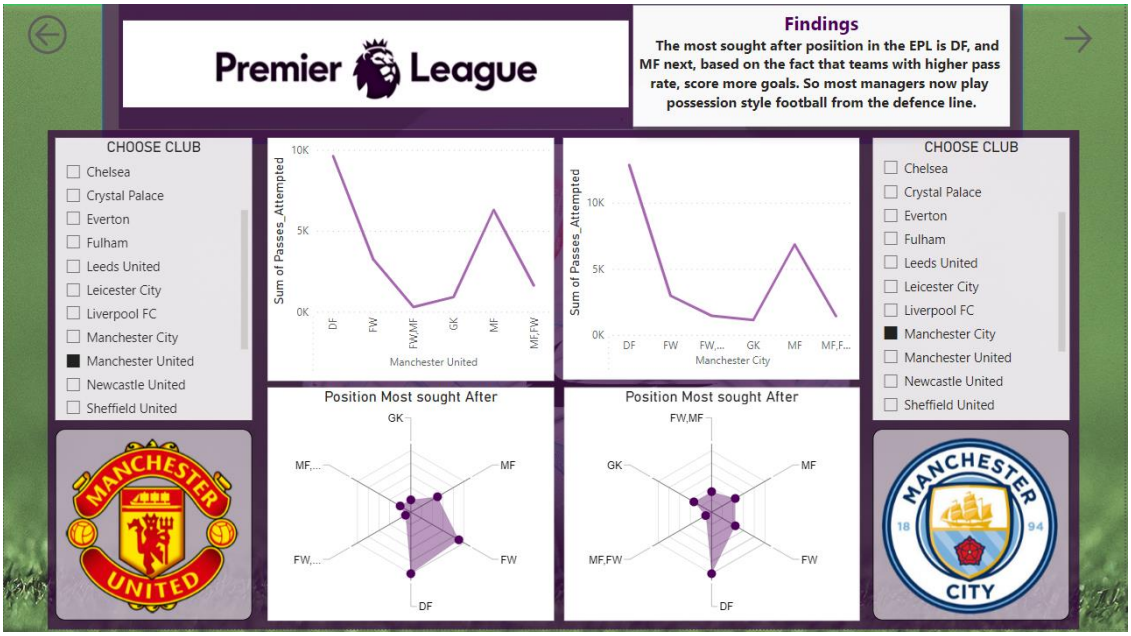


Fig. Power Dashboard page 7

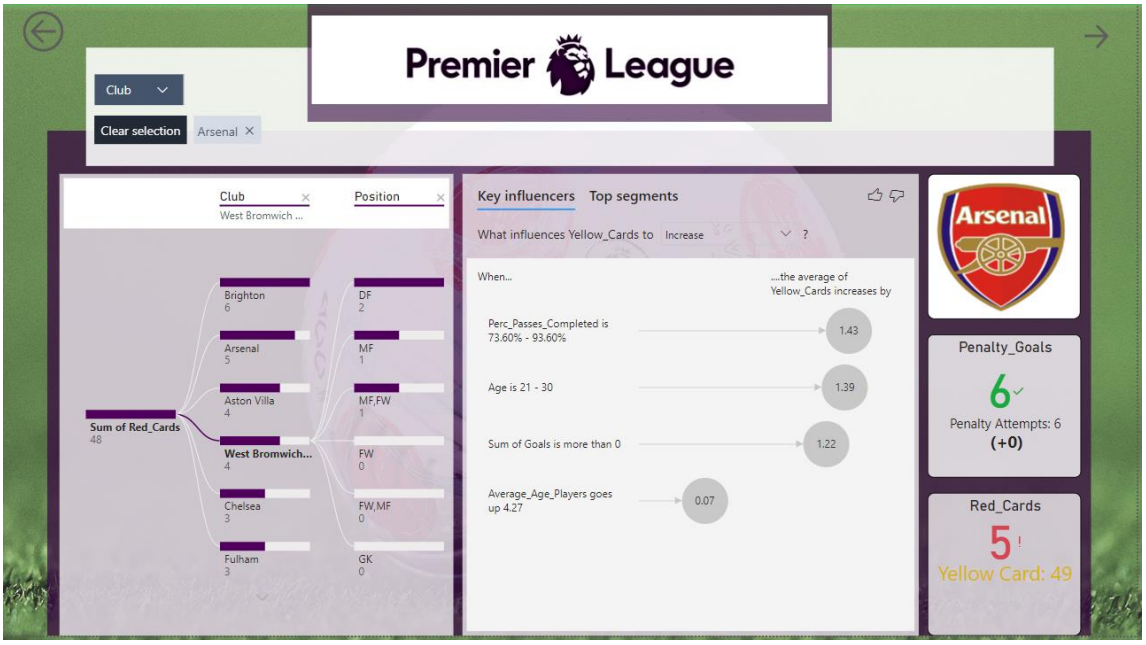


Fig. Power Dashboard page 8

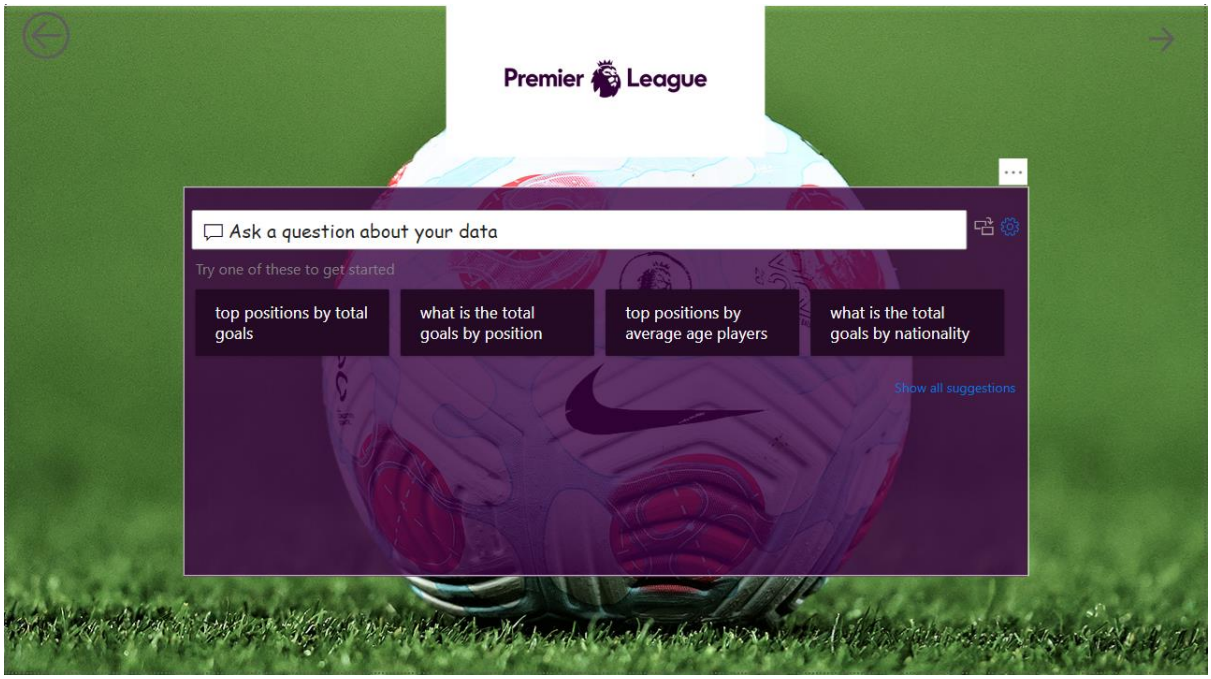


Fig. Power Dashboard page 9

Fig. Power Dashboard page 10

Note: All descriptions of Dashboards are made in the “Findings based on evaluation” Section.

Thank you.