COSC3000 - Visualization, Computer Graphics & Data Analysis

Data Visualisation Project Report

Name: Sebastian Narloch

Student Number: 44345714

What is the best football league in the world?

A deep dive analysis of the differences between topflight football with visualisations and analysis to determine the best league in the world best 2008 and 2016

Table of Contents

1.0 Introduction

Association Football, more colloquially known as Soccer, Football or the 'The world game' is one of the most followed and beloved sports in the world. The sports inception in the late 19th century has seen domestic leagues formed all across the globe as well as international competitions which attract billions of viewers. The nature of football and many other sports is to prize the most successful team over the course of a season or competition period, due to this basic model, patterns emerge of dominant teams. With Europe having a high proportion of successful leagues and successful teams, followers of the sport have been arguing the question for decades, which league is the 'best'?

1.1 Aims

The project aim is to use data visualisation techniques on a football database which spanned from 2008 to 2016 and determine which league was the best in the world based on the inferences of the data visualisations. These visualisations will also allow us to view previously unseen patterns in football results and hypothesise as to why a particular league is more successful, dominant or competitive.

1.2 Background

The Champions League is an annual inter-league, European competition created in 1955 and sanctioned by EUFA. The competition is a round-robin style tournament where teams are seeded based on their ranking in the prior domestic season and play through a traditional World Cup style format. The Champions League is as close to a litmus test of determining the best league in Europe; however, the competition is swayed by many variables such as scheduling issues, luck and pressure. The Champions League does not particularly highlight which league is the best in the world, but rather which team is the best. Since there is no league style competition (the marathon equivalent of football) to put the most successful football leagues in competition with one another, data analysis can be used to determine which league is statistically better than the rest.

The five leagues which will be compared in this report will comprise of the Premier League (England/Wales), La Liga BBVA (Spain), Bundesliga (Germany), Serie A (Italy) and Ligue 1 (France). These five leagues were chosen based on their UEFA coefficient, a seeding system based on various variables, which determines how many teams from each league are allocated a position in the Champions League. Since these teams were placed in the top five of the co-efficient laddering, these leagues are considered to be the best, however which one is really the best?

Ranking			Member association	Coefficient						Teams	Places in 2021–22 season			
2020	2019	Mvmt	(L: League, C: Cup, LC: League cup[1])	2015–16	2016–17	2017–18	2018–19	2019–20	Total		CL	EL	ECL	Total
1	1	—	Spain (L, C)	**23.928**	**20.142**	19.714	19.571	16.017	**99.426**	5/7	4	2	1	7
2	2	—	England (L, C, LC[1])	14.250	14.928	**20.071**	**22.642**	**16.285**	88.176	4/7				
3	4	+1	Germany (L, C)	16.428	14.571	9.857	15.214	15.571	71.641	5/7				
4	3	–1	Italy (L, C)	11.500	14.250	17.333	12.642	12.642	68.367	5/7				
5	5	—	France (L, C, LC[1])	11.083	14.416	11.500	10.583	9.333	56.915	2/6	3			6
6	7	+1	Portugal (L, C)	10.500	8.083	9.666	10.900	10.300	49.449	0/5				

Fig 1: UEFA Coefficients [1]

2.0 Methods

2.1 Data Collection and processing

The data source used for this project was sourced solely from a database repository on the data science website Kaggle.com. The name of the repository used is the 'European Soccer Database' [2] which is a gold rated data source on the site and has over 1400 kernels analysing the data set. The database which stores all the information was quite large at 299 MB and held over 1.08 million data entries over 9 table. The contents of the database stored real life football match results over 11 different leagues between 2008 and 2016 as well as statistics for the video game FIFA over the same time period.

The processing of the data to produce the visualisations was done solely in Python with various different libraries such as Matplotlib, Seaborn and Pandas just to name a few. Due to the sheer size of the database, the data did not come in CSV format but rather in an SQLite relational database. The database' structure was not as intuitive as was hoped for since some of the tables were designed poorly and missed vital information which would make data analysis easier. This meant that in order to work with data effectively, SQL queries had to be performed to create views which would allow manipulating the data more straight forward for most cases but not all.

3.0 Analysis and Discussion

3.1 Metrics for analysis

Due to football being a sport where 'success' is determined via a number of various factors, a baseline metric had to be established by what would set apart the leagues being analysed and compared. The metric which was established to compare the leagues was goals scored; goals at the end of the day decided who won, lost or drew the game. A general rule of thumb in almost all sports is that the more you or your team scores, while also stopping the opponent scoring, the more likely you are to win. A hypothesis was made that higher scoring leagues with higher average scores per game meant that a league was more competitive than other leagues. This will now be determined and analysed. The following data visualisations look at all the leagues within the 8 year period (2008-2016) and analysis.

3.2 Average Total Goals Per Game

The first way to gauge the competitiveness of a particular league that will be looked at, is to look at the total amount of goals scored within a game (cumulative). This will show the normal distribution of score totals across different leagues.

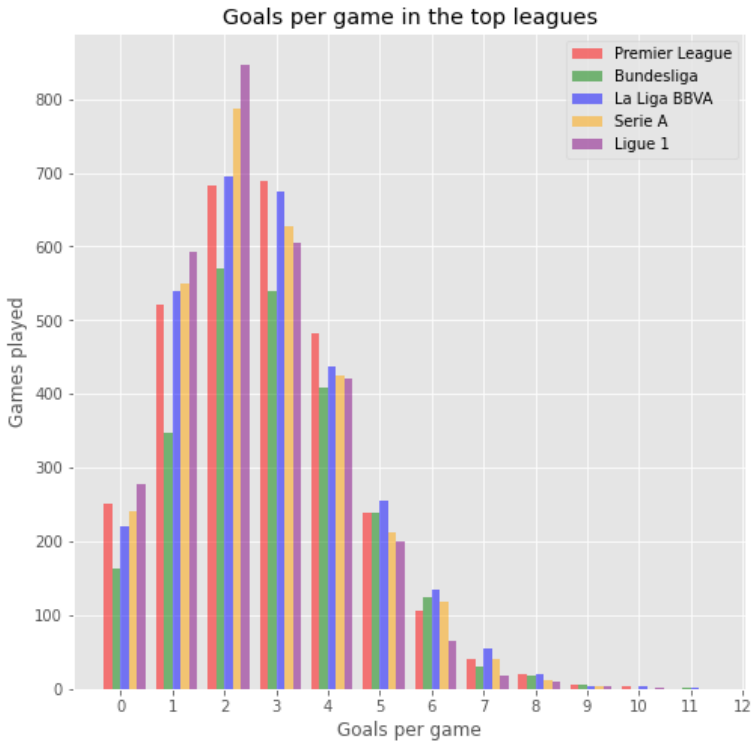

Fig 2: Goals per game in the top leagues (2008 – 2016 cumulative)

Premier_League_Total_Games	LaLiga_Total_Games	Bundesliga_Total_Games	SeriaA_Total_Games	Ligue1_Total_Games
3040	3040	2448	3017	3040

What can be inferred from this graph is that the spread of goals among the top five leagues follows a positively skewed normal distribution curve, where the bulk of the cumulative scores range from 0 to 6. At the peak point of the distribution (2 goals) it can be seen that the French Ligue 1 and the Italian Serie A have the highest amount of total games with a cumulative score of 2, a score which can only be comprised of either 2-0 victories or 1-1 draws. However, past this point it can be seen that the English Premier League leads the race for cumulative goals up to 5 goals per game and is then overtaken by La Liga which leads for the remaining score totals. An interesting observation can be seen in the Premier Leagues distribution; the Premier Leagues is the only leagues which has its peak past the 2 goal cumulative, with 3 goals being the most common cumulative score. A notable mention in this graph is the cumulative scores of the German Bundesliga, which has far less in each category of goals scored. This is mainly due to the fact that the Bundesliga follows a shorter schedule and two fewer teams than the other leagues and thus payed nearly 600 less games in the 8 year period. However, the Bundesliga can be seen having relatively a large amount of high scoring games, despite the league playing less games.

3.3 Score Breakdown

The previous graph showed the cumulative scores across each league and compared them in a histogram style. In order to get a deeper understanding of all the leagues, a breakdown of the cumulative scores needs to be analysed to gauge what type of league it is; this will be done using contour plots.

Ligue 1

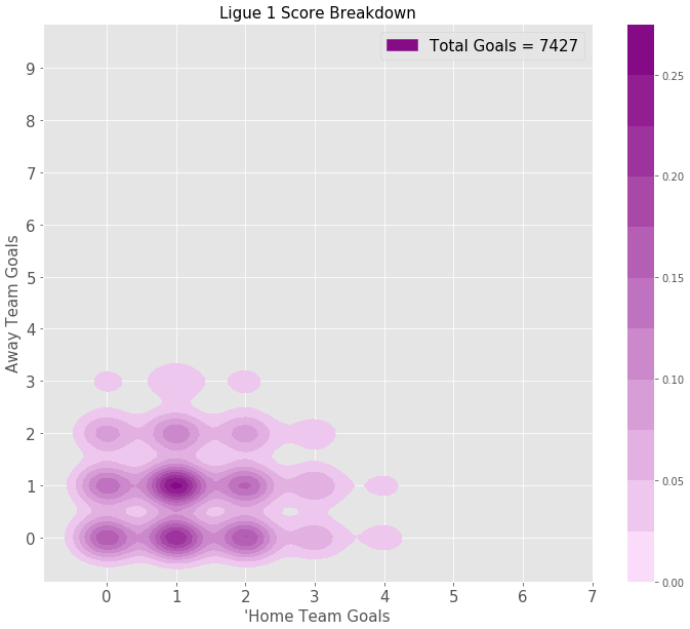

Fig 3: Ligue 1 Score Breakdown

In figure 2 it was seen that Ligue 1 recorded the highest number of games with a cumulative score of 2, at around 850 games. This breakdown of the Ligue 1 scores shows that in fact the most common score is 1-1 draw with roughly 27% games finishing with this result. Another interpretation of this contour plot shows that the bulk of the games are between 3 home team goals and 2 away team goals, with total goals in the 8 year period being 7427.

Serie A

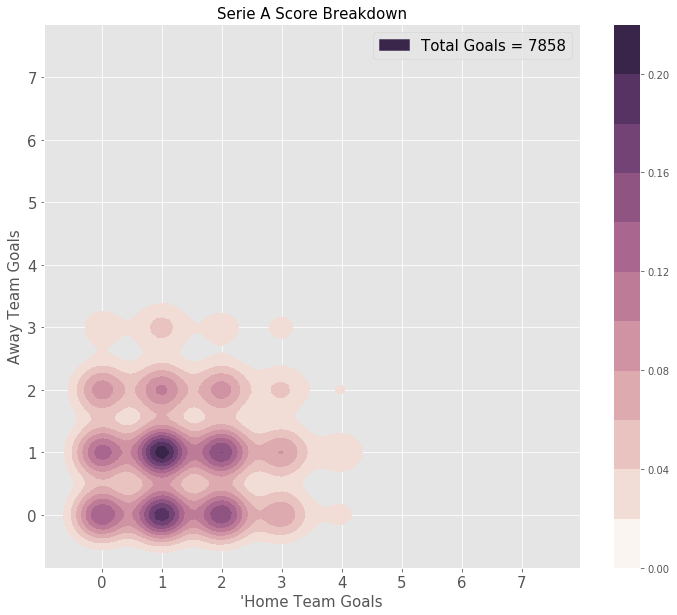

Fig 4: Serie A Score Breakdown

The Serie A much like Ligue 1 had a large amount of total games ending with a cumulative score of 2 goals with just under 800 total games and 7858 total goals. The 1-1 draw was also a very common result, occurring roughly 22% of the time. The Serie A however compared to Ligue 1 has a slightly more diverse goal breakdown with away teams scoring more often and games resulting in a wider range of results for both home and away goals. The Serie A's score breakdown definitely shows a league with a bit more fight from both teams this is due to the fact that 1-1 draws occurs less frequently and thus means that the league has greater attacking threats.

Bundesliga

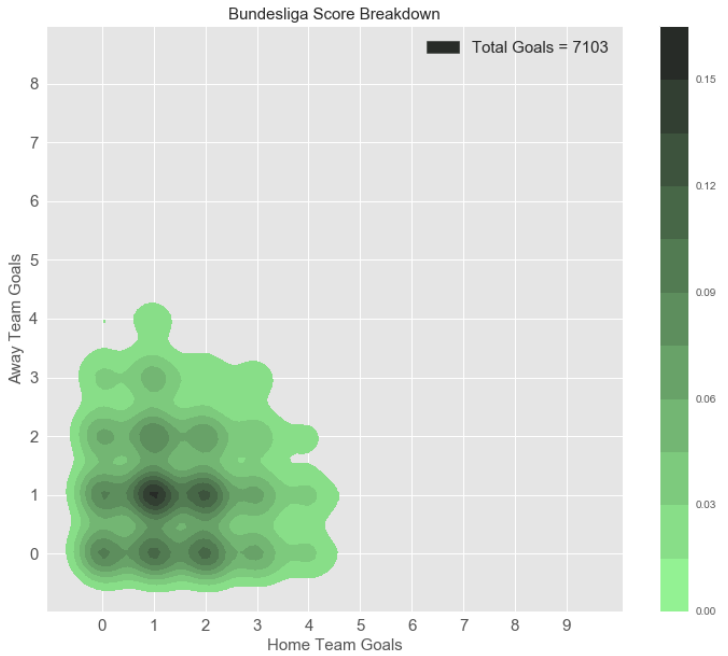

Fig 4: Bundesliga Score Breakdown

At a first glance it is quite apparent that the Bundesliga is a very diverse league in terms of match outcomes compared to both Ligue 1 and Serie A. Considering the Bundesliga had played almost 600 fewer games in the same time period as the Serie A and Ligue 1 while scoring 7103 goals in the same period is quite a feat. What stands out in the Bundesliga score breakdown is that although the 1-1 draw still the most common result, it only occurred roughly 16% of the time. Another interesting thing about the Bundesliga is that the league has away teams scoring more often as well as in larger amounts more often. This plot could signify that the Bundesliga appears to be a quite competitive league with away teams pulling their weight more often and giving the home team a run.

La Liga BBVA

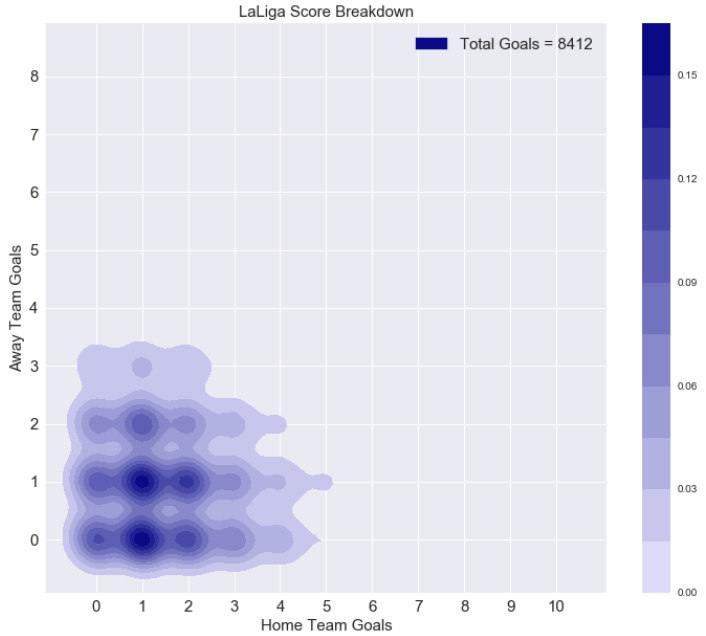

Fig 5 La Liga Score Breakdown

The Spanish La Liga follows the same pattern as all other leagues with the 1-1 draw being the most common result (16% of the time) as was examined in the cumulative score graph. A standout of La Liga comes from its very impressive goal total which is in fact the largest of all the leagues at 8412. The 1-0 home victory also proved to be quite common and shares similarities with Serie A and Ligue 1. The leagues score distribution is very close to that of the Bundesliga, however La Liga is shown to have more games where the home team scores upwards of 4 goals. La Liga's score breakdown could signal that it is quite a competitive league since both the home teams and away teams tend to score more goals, however compared to the Bundesliga, La Liga's away teams score fewer goals.

Premier League

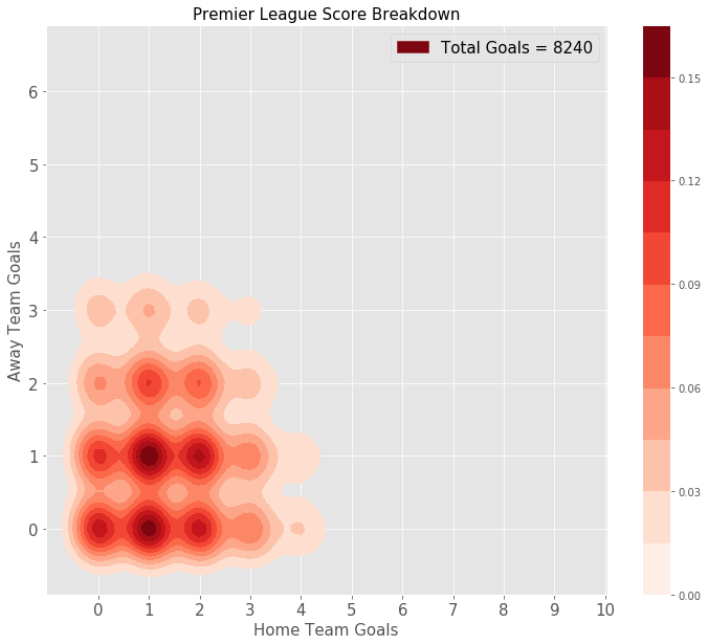

Fig 6: Premier League score breakdown

The Premier League, the world most known league in the world shares many commonalities with the other leagues in terms of its goal breakdown. In figure 2 it was seen that the Premier League was the only league with a peak average cumulative score of above two. Although the 1-1 draw (16%) and 1-0 (15%) home win are quite common, the games with cumulative results of 3 cumulative (3-3 diagonally) are shown to be more common in the Prem compared to other leagues, with the 2-1 and 1-2 results most common. A league with higher scoring games, particularly when the home team scores, can signify that both the away team and home team have more equal playing field and thus can be deemed more competitive.

3.4 Home Advantage by League

A common trend which was discovered while analysing the score breakdowns between the leagues was that the home team tended to score more goals compared to their visiting opponent. Home advantage is a real phenomenon which occurs in sports and can be attributed to a psychological edge which the home team gains by having fan base echoing support. A way to measure competitiveness within a league could be to see how often the away team can overcome this disadvantage and cause an 'upset' at the home team's grounds.

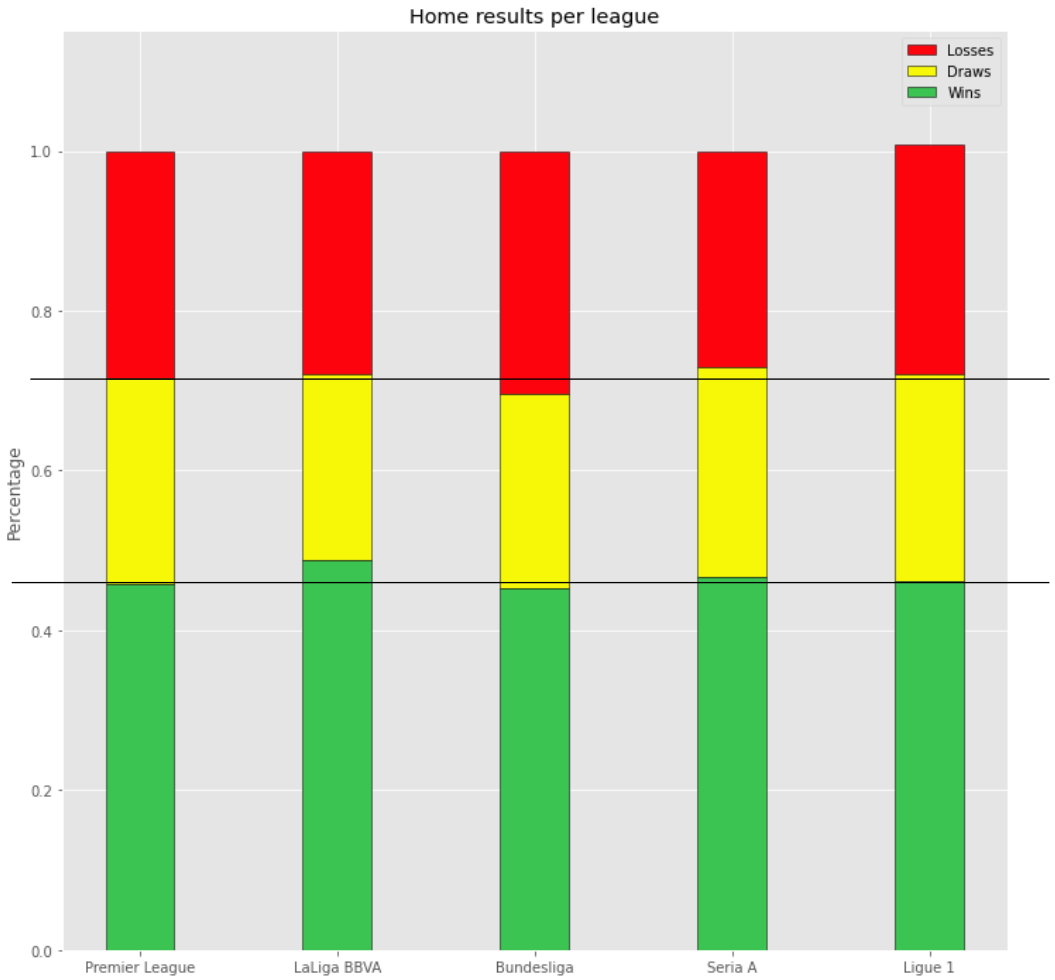

Fig 7: Home advantage of the different leagues

What is observed in figure 7 is that the Premier League, Bundesliga, Serie A and Ligue 1 all have home win rates of between 45% and 46%, La Liga however sees a home win rate closer to 50%. Due to this high win rate in La Liga, draws are shrunken quite drastically compared to the other leagues. The Serie A sees the lowest rate of 'upsets' but instead sees a higher rate of draws compared to other leagues, this bodes well for competition. The largest shock however comes from the Bundesliga where it can be seen that away teams have a higher chance of beating the home team compared to the other leagues. The fact that an away team within the Bundesliga is more likely to pull off an 'upset' is an interesting observation and could be an indication of a more competitive league.

3.5 Goals Per Team in Each League

In order to gauge a league on its goal scoring merit, a breakdown of the goals by each team within the specific league needs to be analysed. This analysis will give insight into the dynamics within the different leagues and show how many teams within a particular league are competing for the top spot. The following scatter plots will be displaying all the teams to have played within the topflight league within their country between 2008 and 2016 and display their goal total within that period.

Ligue 1

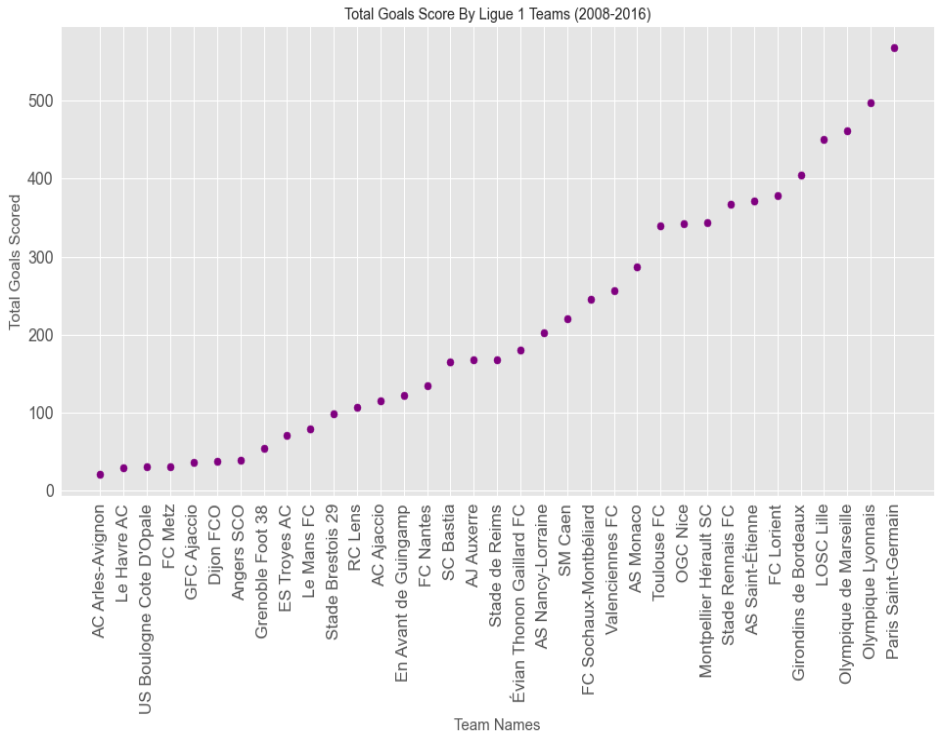

Fig 8: Ligue 1 goals by team

Ligue 1's goals per team distribution follows a relatively linear distribution with Paris Saint Germain, the leagues most notable team leading the scoring at 600 goals between 2008 and 2016. What is quite surprising about Ligue 1's distribution is the lack of 'chunking', a phenomenon observed where there is a drastic gap in scoring quantity by a group of teams. This linear pattern can signify that a leagues quality is distributed quite evenly and thus can signify a less competitive league since there are fewer teams of equal ability.

Serie A

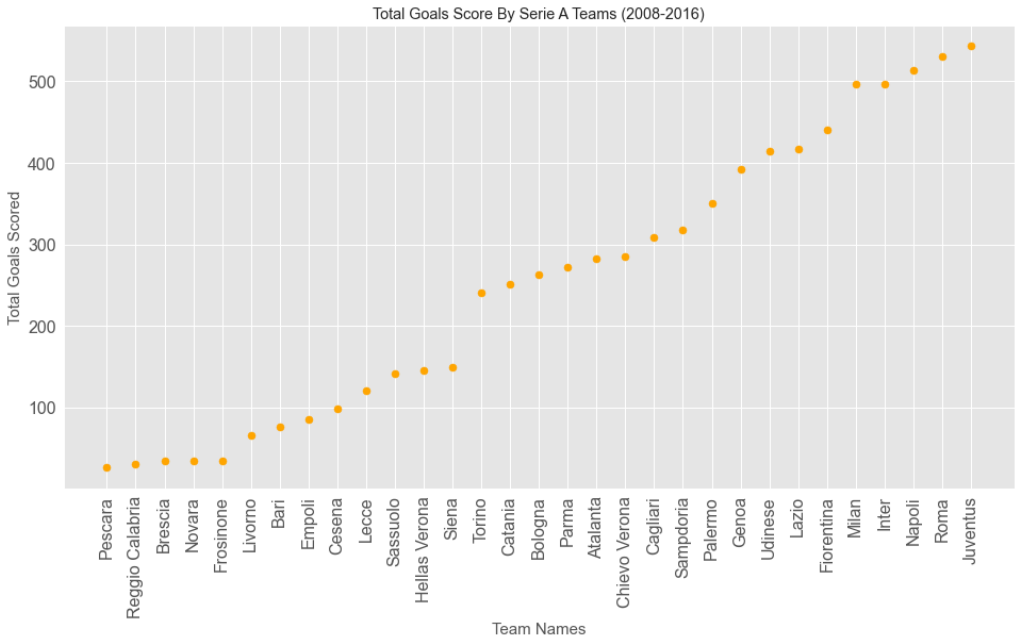

Fig 9: Serie A goals by team

The Serie A goals per team breakdown shows the aforementioned 'chunking' which occurs in various leagues; this is quite apparent by the noticeable quality gap between Siena and Torino. A very interesting thing about the Serie A is the fact there is a chunk of five teams from Milan to Juventus who all have 490+ goals. This large amount of top end quality combined with a sizeable mid table pack would have made the Serie A very competitive between the 8 year period.

Bundesliga

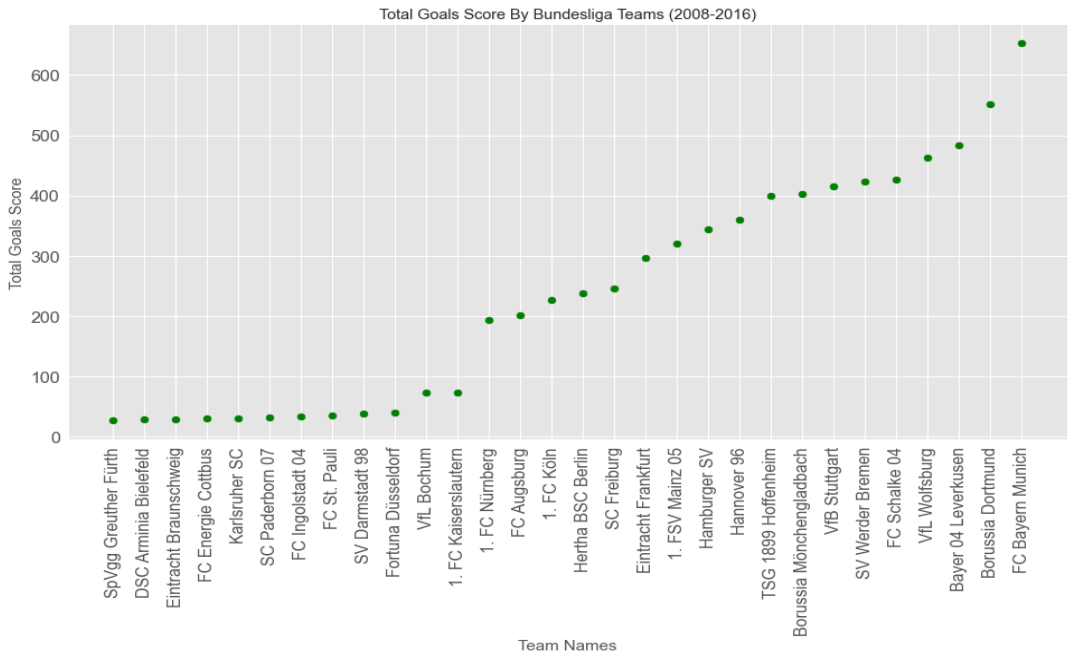

Fig 10: Bundesliga goals by team

The Bundesliga is a quite peculiar league in that there is large divide in quality; the league is split into two chunks with the bottom twelve teams barely hatching over 100 goals in an 8 year period. This large number of teams with so few goals to account for could signify that the Bundesliga is a heavily rotated league with many different teams coming up from lower divisions. Above the twelve lower placed teams, the growth of goals per team grows at a linear rate until VfL Wolfsburg where the growth resembles an exponential growth to the peak where Bayern Munich heads the league with approximately 680 goals. The noticeable divide in quality could be the reason why the score breakdown for the Bundesliga had a large spread and had a greater proportion of high scoring away teams. With more lower quality teams, the goal breakdown of the Bundesliga was likely skewed, since teams who barely scored over 50 goals within an 8 year period likely got pummelled in both home and away games.

La Liga

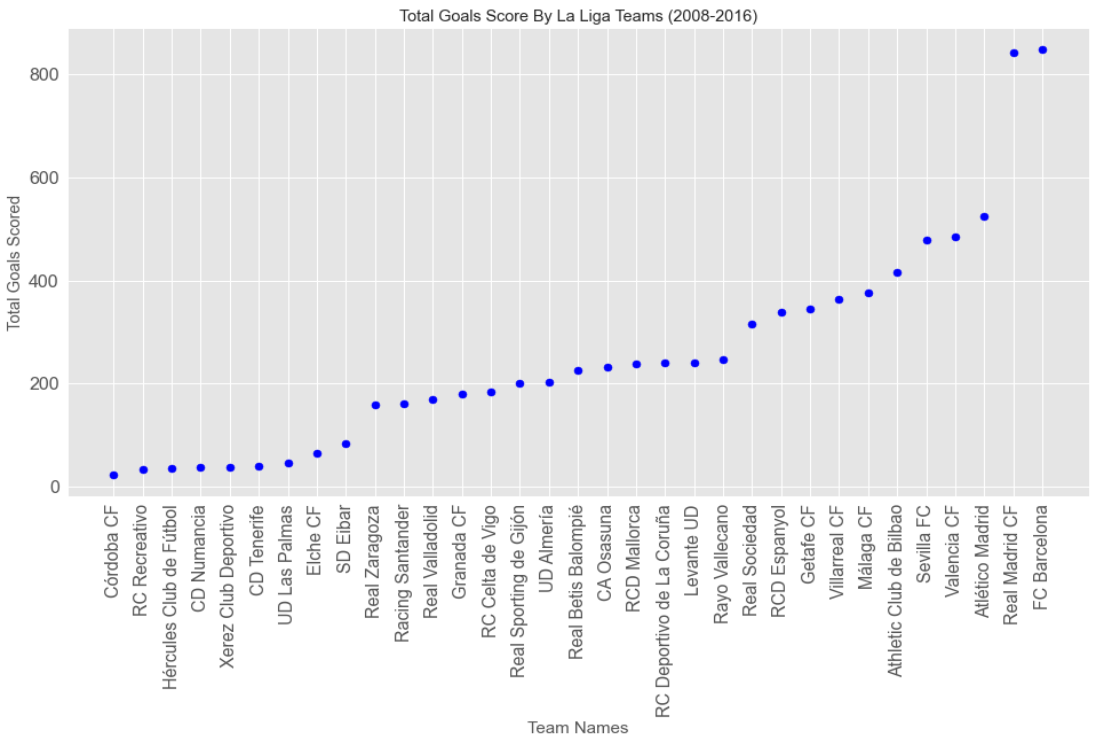

Fig 11: La Liga goals per team

La Liga's goals per team breakdown reveals a lot about the league, the general trend of the goals per team follows a linear curve up until Atletico Madrid where there is a sudden shift towards Real Madrid and Barcelona. The shear dominance of Real Madrid and Barcelona in the Spanish league makes the assumption that the league is a two-horse race very easy to believe. With both Real Madrid and Barcelona having scored 800+ goals over an eight year period, the leagues title winning competition is relatively low and reserved to only Barcelona and Real Madrid. However, the mid relatively equal mid table is more competitive and likely fighting for a Champions League position.

Premier League

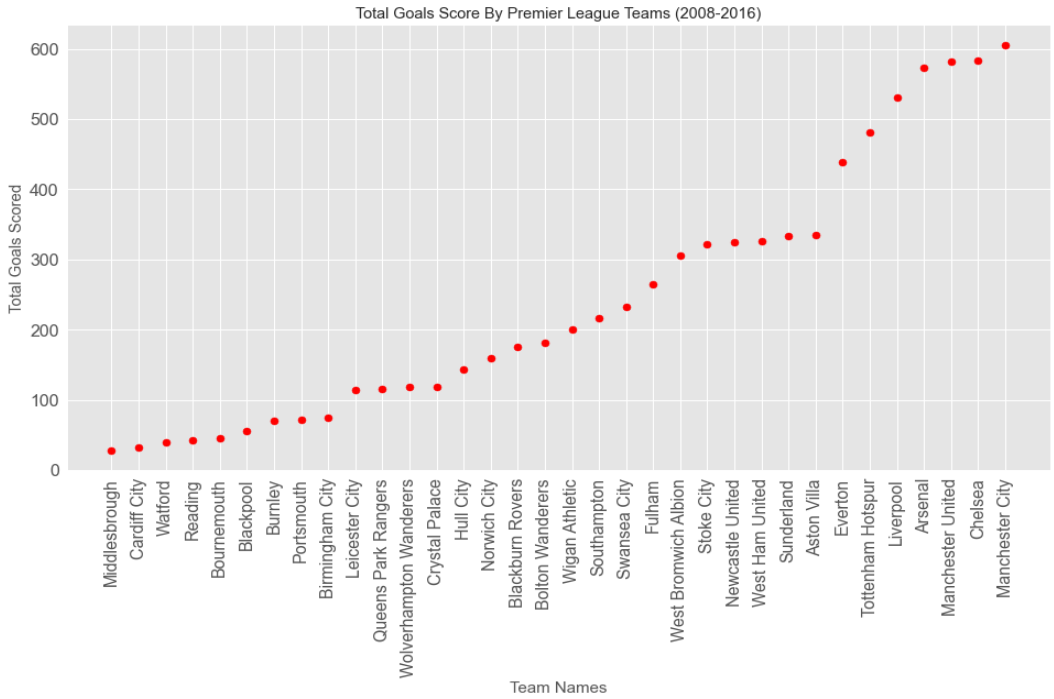

Fig 12: Premier League goals per team

The Premier Leagues goals per team breakdown shows a linear trend of goals per team from Middlesbrough up to Fulham. After Fulham there is a plateau ranging from West Bromwich Albion until Aston Vila, this is quite interesting as no other leagues have displayed a plateau so high and could be an insight into the competitiveness of the Premier League. After the mid table plateau there is a noticeable jump in goals from Everton to Arsenal, the latter of which begins another plateau of the highest scoring teams. The Premier League's goal scoring patterns which display several plateaus can be an indicator that the league has several teams with equal ability and thus means that there is a high level of competition within the Premier League.

3.6 The Top 30 Teams

Having just looked at the goals per team within each team's respective league, some clear patterns and insights have been observed. Further insights can be established by looking at how teams from other leagues fair against one another. The following table will display the top 30 goal scoring teams across the five leagues observed and display them in ascending order as was done for the league comparisons.

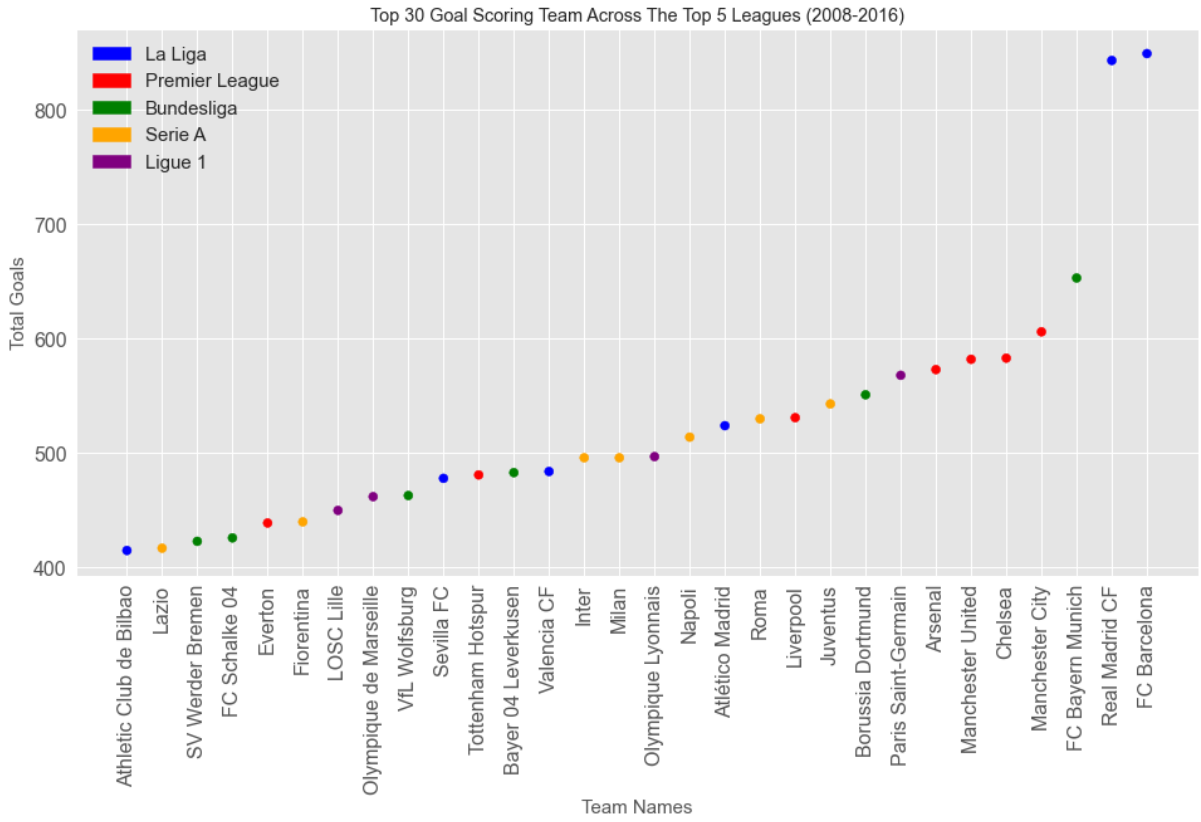

Fig 13: Top 30 teams ranked by goals scored

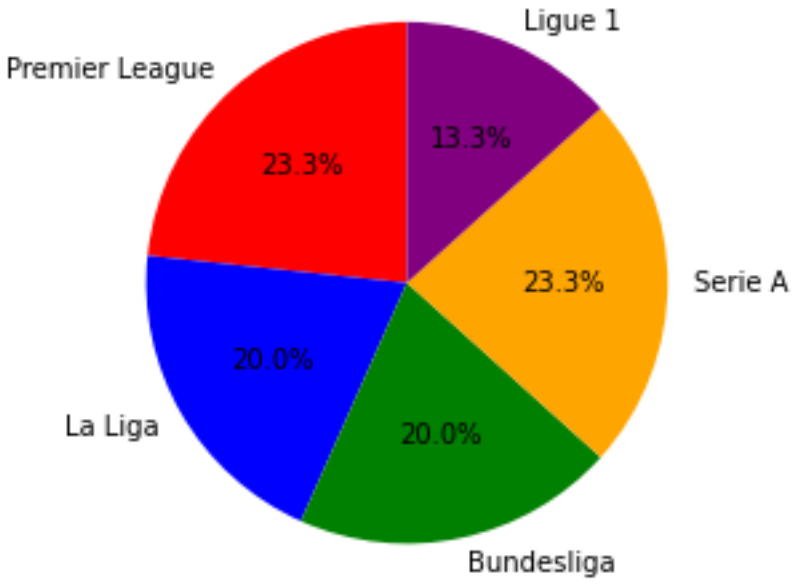

Fig 14: A pie chart breakdown of the top 30 teams by league representation

Several things are quite apparent on first glance when looking at figure 13 above, with the most jarring observation being Barcelona's and Real Madrid's dominance over other high scoring teams.

Looking at figure 14 it can be seen that there is an almost even split by league representation in the top 30, with the Premier League and Serie A leading representation at 23.3% each. However, the Premier Leagues representative teams can be weighted more heavily since 4 of the top 7 teams are Premier League teams whereas Serie A's highest team come in at 10th (Juventus).

With the Premier League having 4 teams ranked so highly, this shows that the Premier League has a high level competition within Europe and domestically.

While observing figure 13, a correlation was observed between goals scored and the level of wealth generated by the teams and their respective leagues. Deloitte, a very well-known professional services network, complies an annual report of wealth in football called the *Deloitte Football Money League* where it ranks the 30 teams which bring in the most revenue.

Rank in 2016	Club	Revenue (€ million)	Country	Rank in 2015	Change
1	Real Madrid	577.0	Spain	1	–
2	Barcelona	560.8	Spain	4	+2
3	Manchester United	519.5	England	2	–1
4	Paris Saint-Germain	480.8	France	5	+1
5	Bayern Munich	474.0	Germany	3	–2
6	Manchester City	463.5	England	6	–
7	Arsenal	435.5	England	8	+1
8	Chelsea	420.0	England	7	–1
9	Liverpool	391.8	England	9	–
10	Juventus	323.9	Italy	10	–
11	Borussia Dortmund	280.6	Germany	11	–
12	Tottenham Hotspur	257.5	England	13	+1
13	Schalke 04	219.7	Germany	14	+1
14	Milan	199.1	Italy	12	–2
15	Atlético Madrid	187.1	Spain	15	–
16	Roma	180.4	Italy	24	+8
17	Newcastle United	169.3	England	19	+2
18	Everton	165.1	England	20	+2
19	Internazionale	164.8	Italy	17	-2
20	West Ham United	160.9	England	21	+1
21	Galatasaray	159.1	Turkey	18	-3
22	Southampton	149.5	England	25	+3
23	Aston Villa	148.8	England	22	-1
24	Leicester City	137.2	England	31+	
25	Sunderland	132.9	England	27	+2
26	Swansea City	132.8	Wales	29	+3
27	Stoke City	130.9	England	30	+3
28	Crystal Palace	130.8	England	31+	
29	West Bromwich Albion	126.6	England	31+	
30	Napoli	125.5	Italy	16	-14

Fig 15: Deloitte Money League 2016 and contributions by league [3]

Appearances by Country

Ranking	Country	Number of Teams	Total Revenue (€ million)
1		16	3939.8
2		5	993.7
3		3	1324.9
		3	974.3
5		1	480.8
		1	159.1
		1	132.8

Fig 16: Appearances by league

The Deloitte Money League report of 2016 shares a very close resemblance to the top 30 scoring teams in figure 13; in fact, 19 of the top 30 scoring teams are in the money league. The fact that the breakdown of appearances by league in the top 30 in figure 14 also resembles the appearances by countries in the money league in figure 16 is uncanny. Another observation that was made is that a large portion of the English teams on the lower part of the money league report also fall in the mid table plateau zone in the Premier Leagues breakdown, a zone which likely signified a more competitive league. This correlation can be used to infer that wealth is a major benefactor in the competitiveness within a league and that either wealth drives competition or competition drives wealth.

3.7 Conclusion

Having observed and analysed the previous data visualisations, a conclusion of which league is the best needs to be made. Based on the patterns and observations made, the most competitive league in the world between 2008 and 2016 was the English Premier League. The reason for this is because the league has shown through its goal statistics that the league sees more competitive and exciting games, has more teams of equal ability and delivers more wealth to increase the competition within the league. The fact that the Premier League shows a high level of competition is a variable which likely contributes to its global popularity.

4.0 Self Reflection

4.1 What I learned

I learned a lot of about data science and data analysis through this project; having spent hours and hours trying to manipulate the data in a certain way, I learnt many different skills and brushed up on many others. My SQL skills definitely improved since SQL was vital to manipulating the data in a way to convey a specific message. By analysing the data I was also able to gather insights about football that I never knew and this was very interesting to me. Having never used any Python data analysis libraries before, I am quite impressed by the amount of clear visualisations I was able to produce in such a short time of learning. This project has given me many insights about data science and has made me more interested in the field than I was before.

4.2 What I wish I had done

My original idea was to compare real life football statistics and compare them to the player and team statistics in the FIFA video games. The purpose of that analysis was to see how FIFA ratings and rankings compared to real life football. The idea was out of my reach since the database was not designed in a way to make that analysis easy, thus this project idea was devised. The things which I wish I had done to improve this project was to use more mathematical methods to actually prove my hypothesis, however I couldn’t do that since I don’t have the statistics background to do that effectively.

5.0 References

[1] Mathien, H., 2016. European Soccer Database. [online] Kaggle.com. Available at: <https://www.kaggle.com/hugomathien/soccer> [Accessed 14 March 2020].

[2] En.wikipedia.org. 2020. UEFA Coefficient. [online] Available at: <https://en.wikipedia.org/wiki/UEFA_coefficient> [Accessed 24 June 2020].

[3] En.wikipedia.org. 2020. *Deloitte Football Money League*. [online] Available at: <https://en.wikipedia.org/wiki/Deloitte_Football_Money_League> [Accessed 26 June 2020].