The Premier League Project Summary

The purpose of this project is to showcase my ability to collect data from various sources and gain some statistical insights from the data.

I have made 5 datasets which contain information about premier league teams. The tables along with the sources in which I found the information is found below:

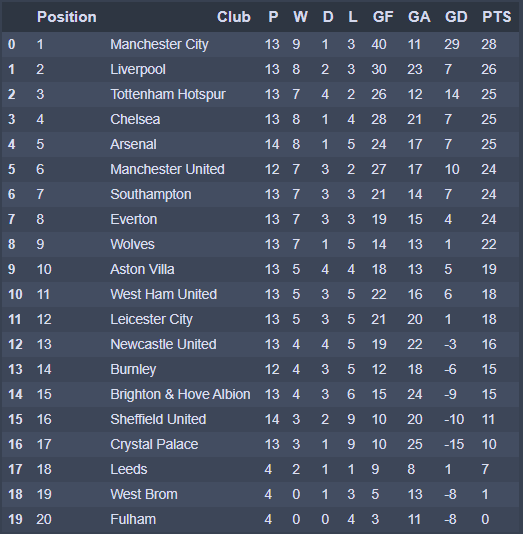
Form – This is the premier league table stats from when the premier league season 19/20 resumed, right up until match week 4 before the first 20/21 international break. <https://www.givemesport.com/1586210-liverpool-6th-man-utd-2nd-premier-league-table-based-on-games-since-lockdown-only> (Lockdown to the end of the 19/20 season) plus the premier league table up until match week 4, which can be found on Google.

H2h – This table shows us the number of wins each team in big 6 (Liverpool, Manchester United, Manchester City, Arsenal, Chelsea and Tottenham) has had vs other members of the big 6. I have taken the data from <https://www.11v11.com/> (to clarify, this is only for the premier league and no other competition).

Placements – This is where each current premier league team has finished in the last 5 seasons, data here has been collected from Google. (e.g. Premier league table 18/19).

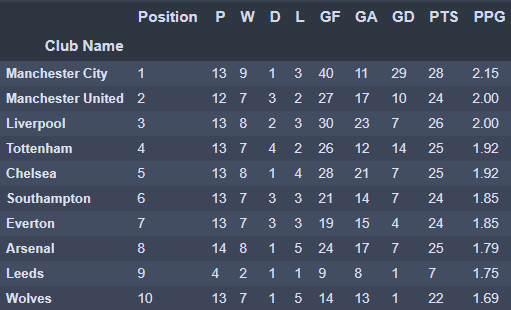
Squad Values – This is a table of the financial values of each current premier league squad since the 18/19 season until 19/20 season and the rankings of their squad values compared to each other from the 15/16 season to the 20/21 season. Data here has been taken from <https://www.transfermarkt.com/>.

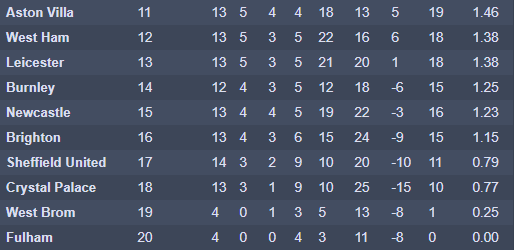
Transfers – This is the net spend of each current premier league team, showing is how much more or less a club has spent for their player transfers, incomings and outgoings. The data here was also from <https://www.transfermarkt.com/>.

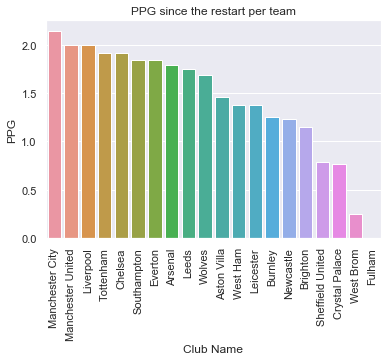


Current Form

This is the table which tells us the form of each premier league team, we can see that this dataset has some problems. The newly-promoted teams, Leeds, West Brom and Fulham don’t have enough games played to show an accurate amount of form as the rest of the teams. We fixed this by ordering the table in terms of points per game, this will give us a more accurate representation of the form between the newly promoted sides. We’ll also arrange the index and change the positions relative to the new points per game column.

This is now what new form table looks like, here are the top 10 performers since the restart during the 19/20 season. If you already know what the 19/20 league table looked like, you’d notice that Liverpool, the champions had begun to decline slightly and Tottenham, Southampton, Manchester United have improved relative to their entire league campaign.

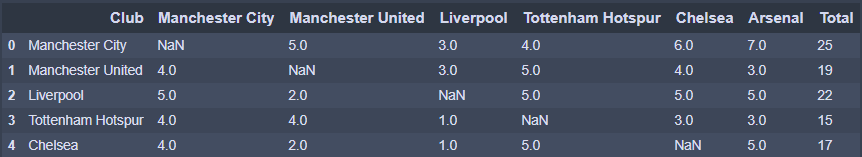
The bottom ten are represented here, Aston Villa and West Ham narrowly escaped relegation that year, which means since the restart they had improved that much better to avoid it. Surprisingly Sheffield United finished very slow considering that they finished 9th last season, this goes to show how well their season began in this year.



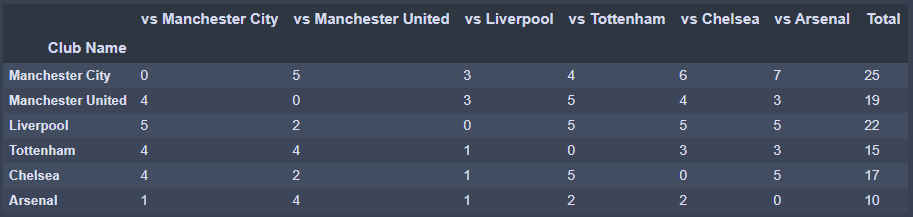
Here’s a visual representation of the table of the points per game per club, we can see there is a sharp dip between Wolves and Villa and an even sharper dip between Brighton and Sheffield. Villa has won all their games this season, so that will boost them up largely in this table from where you would expect.

Some errors here would be that the data for Fulham, West Brom and Leeds are quite inaccurate as they were in the championship before the current season. This means they haven’t had a fair sum of games to play before being evaluated.

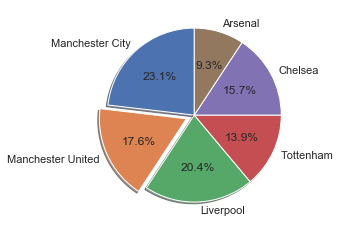
Head to Head Matchups



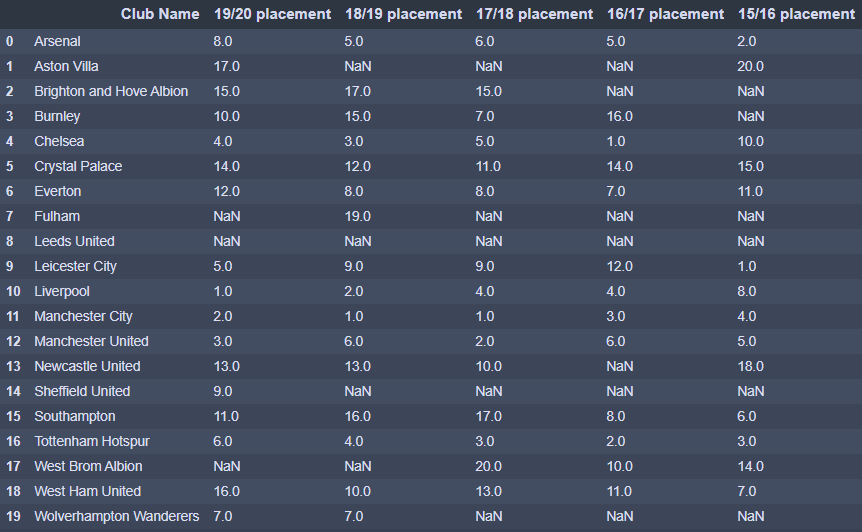
Above is the initial table of head to head wins the big 6 have won against each other. To clean this table we need to rename the columns and rows to match the same names as those in the previous dataset, change our values from floats to integers as none of these units are decimal values and fill in our null values with 0s as no team can win against themselves.



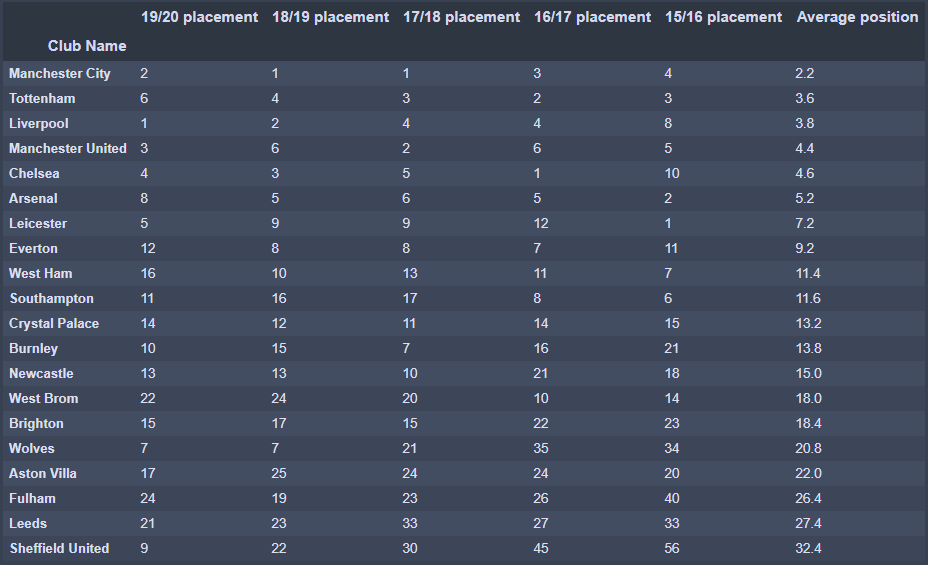
Some noticeable points we can see are that within a premier league season, you can only play each other twice and this dataset is covered over 4 seasons, which means each team has played 8 games against one another. Manchester City, therefore, has 7 wins in 8 games vs Arsenal in the premier league, more than any other team against another in the big 6. We can also see that the northern English teams tend to do better in the big fixtures than what the London teams do. Every team has had a win against each other over the last 4 seasons. The London teams have only had 1 win each against Liverpool in the last 4 seasons.



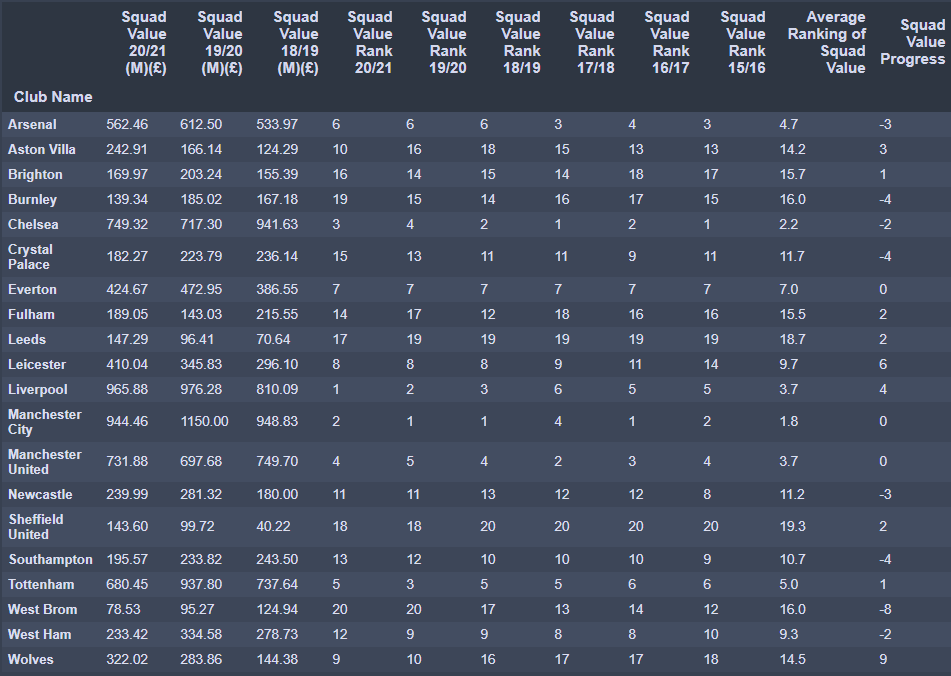
Here is a pie chart of the percentage share of wins each team has had. Liverpool and City take up more than 43% of the circle whilst the London clubs combined are a little shy from 40%. This tells us that over the last 4 seasons, Arsenal will average a win against the big six less than 1 out of 10 times.

League placements since the 2015/2016 season

On the right here, we have five columns for each season from 15/16 to 19/20 which details the placement that each current premier league team has finished in the given season. A problem here is that many of the current premier league teams were not in the premier league for some of these years.

To fix this issue I calculated their league finishes as if the leagues below the premier league were placements continuing from the bottom of the premier league. For example, Sheffield United finished 10th in the championship in the 17/18 season, therefore their placement is 30th. I did this for each of the null values within this table and fixed other issues, such as numerical format and selecting the correct index.

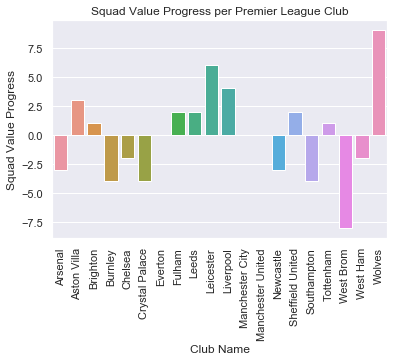
This is the final product of the placements table; I added another column to show each club average placements over the last 5 seasons. The table is ordered in the highest average position to lowest, as we can see the most frequent winner is in first place. It is surprising to see Tottenham in second place given that they have only finished there once, however, some of the bigger clubs have had some low placements over the years, Manchester United have had three 5th/6th place finishes. Liverpool and Chelsea had an 8th and 10th place finish respectively in the same season which has taken a toll on their average.

Squad Values (\* only between current premier league teams)

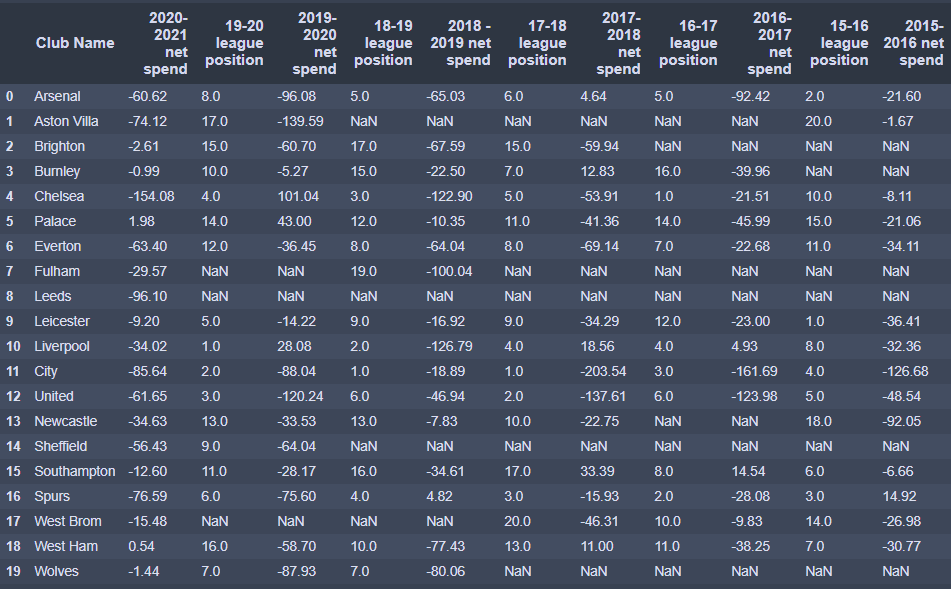
This already cleaned table represents squad values, the combined market price of a given team’s players for seasons 18/19 to 20/21, I have placed three columns which the exact price of the squad. On the right side of the table, we have the rank each team places among each other in terms of the most expensive squad (1st) and least expensive squad (20th). For example, Arsenal in the 20/21 season, are the 6th most expensive squad among current premier league teams.



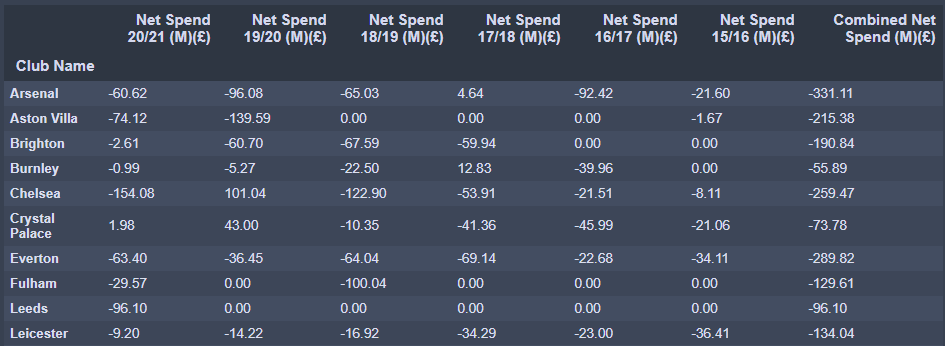
Image 1 shows us how much a team has changed from the 15/16 season to the 20/21 season. Wolves in first place with 9 represents that they have moved up 9 places among their rivals in terms of most valuable squads. This reflects how much a team has improved, Wolves finished 14th in the championship during the 15/16 season, and in the 19/20 season, they were in the Europa League. Arsenal is one of the current big six who have lost 3 places since 2015 when they were last in the Champions League, since then they haven’t competed, so we can expect their squad to not be as good as then. Below, there is a visual representation of this.



Transfers

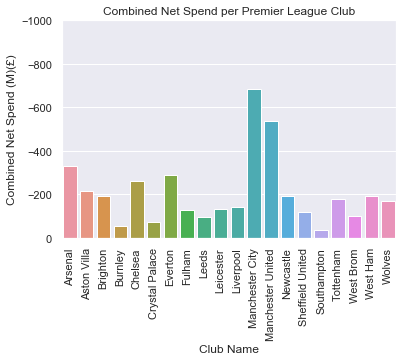


The final messy dataset above called transfers tells us the net spend per current premier league club over the past 5 seasons. Here is the final table so that we can see what was fixed:

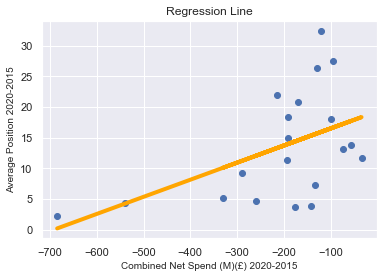


The top 10 are shown above, fixes included:

* Null values were changed to 0 as their spending hadn’t been spent for the premier league at the time as the team were in a lower league, therefore this represents net spend specifically for the premier league.
* I had added a column for combined net spend which is the sum of values across the seasons for each current premier league team.
* I dropped the columns for league finishes as I already have a dataset for that.
* I renamed columns and row names and set the index to club names.

Here is a visualisation of the clubs with the highest net spend:

Manchester City, the club that has won the most league titles over the last 5 seasons has spent the most amount on transfers compared to their sales of players. Manchester City rarely sells their most valuable players, unlike Liverpool who sold Coutinho for over £100M which largely decreases their net spend. The big clubs also do have a much better chance of convincing better players to join them because of the chances of success. We can see this as Tottenham get outspent by West Ham, Everton and Brighton, yet those clubs do not compete at the same level. This concludes that there are more factors to making successful teams than transfer costs.



To prove this, here is a scatter plot with a regression line of combined net spend against a team’s average league position.

The relationship between the two is weak with an R-Squared value of 0.25. All we can tell from this graph is that teams with very low average rankings, don’t spend big money on transfers every year.