Implementing a visualisation with R:

Exploratory analysis of football statistics for the EPL

Daniel Audcent

20316239

Fundamentals of Information Visualisation

April 23, 2021

Word count: 3021

Contents

1. Glossary

EPL	English Premier League
Rk	Rank of Squad in the EPL table (1-20)
# Pl	Total number of players which have been used in games
90s	Total minutes played divided by 90
SCA	Shot Creating Actions. The two offensive actions leading to a shot. E.g. Passes, dribbles, drawing fouls
SCA90	Shot Creating Actions per 90 minutes
PassLive	Completed live-ball passes that lead to a shot attempt
PassDead	Completed dead ball passes that lead to a shot attempt
Drib	Successful dribbles that lead to a shot attempt
Sh	Shots that lead to another shot attempt
Fld	Fouls drawn that lead to a shot attempt
Def	Defensive actions that lead to a shot attempt
GCA	Goal Creating Actions. The two offensive actions directly leading to a goal, such as passes, dribbles and fouls.
GCA90	Goal Creating Actions per 90 minutes
PassLive	Completed live-ball passes that lead to a goal
PassDead	Completed dead ball passes that lead to a goal
Drib	Successful dribbles that lead to a goal
Sh	Shots that lead to another goal-scoring shot
Fld	Fouls drawn that lead to a goal
Def	Defensive actions that lead to a goal
OG	Actions that directly lead to an opponent scoring an own goal
Poss	Possession calculated as the percentage of passes attempted
MP	Matches played by the player or squad
Gls / GF	Goals scored or allowed
GA	Goals conceded
GD	Goal difference. E.g. goals scored – goals conceded
Ast	Assists
G-PK	Non-penalty goals
PK	Penalty kicks converted
PKatt	Penalty kicks attempted
CrdY	Number of yellow cards
CrdR	Number of red cards
G+A	Goals and Assists
G+A-PK	Goals and Assists minus Penalty kicks converted
xG	Expected goals (not including penalty shootouts)
xGA	Expected goals against
xGD	Expected goal difference
xA	Expected goals assisted
xP	Expected points

2. Explanation of the initial questions posed

In recent years, there has been a renewed interest towards the use of football statistics which are believed to hold insights into reviewing team expectations, performance analysis and even predicting match outcomes. In the past, these tools have been briskly dismissed by many football fans, for a number of reasons stemming from their emotional ties to the traditions of the sport, as well as the intrusive nature of transitioning statistics into the footballing world. The aim of this project has been to investigate a number of the controversial football statistics such as expected goals and points (xG and xP), for the English Premier League teams over the past few seasons, through the creation of visualisations using R.

One of the benefits with the introduction of these more advanced statistics, is the potential of measuring team performance against their expectations in a systematic way, even though many of the humanistic and circumstantial factors will be discounted. It also allows sports teams the opportunity to identify areas of strengths and weaknesses, as well as a loose method of tracking improvement in those departments too. With the coronavirus epidemic causing huge ripples in the 2020 – 2021 footballing season, team performances have wildly fluctuated, which has caused many managers to point the finger at the lack of fans present in stadiums. However, it can also be argued that in this situation the scales are now levelled, with many clubs feeling the financial squeeze through the lack of commercial income, which is mostly down to the clubs fanbase. This caused a huge amount of hesitation in the transfer market, whereas in the past clubs have willingly spent large sums of money on big signings to help them reach their targets set. With many clubs opting to place their trust in existing players, how have team performances been affected? How have the clubs with the luxury of flexing their financial power in the transfer market, fared this season so far? Are the teams with the most possession actually seeing better results? This project aims to investigate similar types of questions, using the power of visualisation.

The questions initially focussed on during this project are:

- How does the EPL football teams xP compare to their actual points currently and in the previous season?
- Which teams tend to have the most ball possession in the EPL this season?
    - Are there any trends between an EPL team’s possession and their attacking threat in terms of GCA?
    - Are there any trends between an EPL team’s possession and the number of goals scored?
    - Are there any trends between an EPL team’s possession and the number of goals conceded?
- Has having a higher club value translated into better performances looking at the current season's data only?
    - Has a club’s market value had any effect on the expected performance of teams in terms of xG and xGA?

3. Comments on the data sources

Given the popularity of the sport there was an abundance of data sources available for this project, however there were a number of factors which needed to be taken into consideration while investigating the potential sources. One of the most important was aspects, was the validity and reliability of the data to be collected, which can be closely tied to a data sources overall reputation. As the data had to be extracted via web scraping, each potential data source was thoroughly examined before a decision was made on whether the data was appropriate for this project. Keeping the number of data sources to a minimum was also taken onboard, as this would greatly reduce the amount of pre-processing required, whilst also helping prevent confusion between sources holding similar features, such as the various formats found in club names. With these aspects in mind, three sources of data were suitably identified for this project:

- TransferMarkt: One of the most popular German – based football information websites containing a wide variety of information such as football fixtures and results, player and match statistics, also club and player market values. For this project, only club market values were required from this source as more accurate information for other statistics could be found through other sources. Market value estimations are often believed to have slight inaccuracies, however obtaining accurate information for this metric can be challenging. Existing papers and journals are available where this source is placed under the microscope:

1. Beyond crowd judgments: Data-driven estimation of market value in association football
2. Towards Data-Driven Football Player Assessment
3. Information precision in online communities: Player valuations on www.transfermarkt.de

- Understat: Contains a huge number of football statistics including their very own xG and xP predictor algorithms. These values have been calculated after training a neural network on a very large dataset containing relevant information on over 100,000 shots taken in Europe's top five leagues. This project solely focuses on at most, data from the past five seasons in the English Premier League, however statistics are available for the previous seven seasons through their website. Data was extracted for every match for each English Premier League team over the past five seasons, with a strong interest being placed upon their xG and xP predictions alongside the number of points gained per game.

- FBRef: Another source which contains huge amounts of football statistics provided by a key player in the football analytics industry called StatsBomb, who have had their work endorsed by several high-profile figures at top football clubs. Similar to Understat, StatsBomb have also managed to create their own algorithms for football statistics, making strong claims that they have "the most accurate xG modelling in the industry". The datasets used from this source in this project, the key features of interest, along with any other features present can be observed in Appendix A.

4. Data pre-processing

As three separate data sources were involved in this project, there were a few hurdles to overcome initially. First of all, the statistics had to be extracted from each of the online tables, hence bespoke python scripts were produced for each source. The beautifulsoup, requests and json libraries came in useful here, offering a flexible solution which could be tailored to each web page scraped. The programs created submit and receive a single response request from each website, which is then parsed according to the information required, often located through a unique html class name. These programs were also capable of reformatting that information into a neat data frame which had a similar appearance to the existing table found on each website, before being output to a csv file with a custom name. Several parameters were created to smoothen out the process, allowing for specific seasons data to be extracted without any major adjustments to the code.

After the csv files had been created, the data was read into R. Given the nature of the information acquired, only minimal cleansing was necessary such as renaming club names to be consistent across all sources and reformatting columns to the required data type. Other preliminary actions included initialising a vector of team colours based on each club's primary colour as well as creating a vector containing the club logo images for later use in the visualisations.

The majority of the initial questions used a similar methodology, which was to create a new data frame for each question which contained all the required information pooled from each of the data files, which is then reformatted or transformed in such a way that would produce optimal results for the visualisation required. This process made it easy to track which columns had been used from each dataset whilst still allowing plenty of customisation with regards to column names and datatypes, as these could be altered on demand.

For the first question posed, two data frames were produced from the existing data by looping through all the teams for both the 2019 and 2020 seasons. These data frames were then merged together by the club's name to obtain the points and xP for all EPL teams over the past two seasons. These values were then used to calculate the points difference for each season which was then appended at the end of the data frame.

The second question posed started with a similar process to the first, however additional steps were required for the further sub questions. This involved creating vectors of defined classes for each of the features to be studied, so that the data frame could be split into subset based of a condition. A count field could then be appended to these subsets according to which class the row was designated as, essentially becoming a separate class frequency table holding the answers to each of the questions proposed.

The final initial question proposed (and an additional sub question) was fairly straight forward in terms of preparation, following a similar procedure to the first question. Small additions were included to enhance the quality of the visualisation, such as reordering variables, performing an average value calculation, as well as merging with the club logos data frame.

5. Visualisation strategies considered for the initial questions

A scatter plot was the original idea for the first question, as this type of plot is very well known for identifying patterns in data as well as comparing a data points position in relation to the rest of the data. This was then taken a step further through transforming into a bubble plot, enabling a third variable to be present responsible for scaling the data point size, in this case being the change in points difference over the two seasons. Visual aids were then added in the form of datapoint labels and a red or green colour assignment for negative and positive values, to form the final visualisation which can be found in Appendix B. This makes it easy for the interpreter to group teams with similar characteristics, and also recognise teams with more unusual circumstances such as Liverpool, Everton, and Aston Villa. The teams which are not present in both seasons had been omitted from the data during the pre-processing stage.

For the first stage of the second question investigated, a bar chart was chosen as the most appropriate form of visualisation. The final visualisation for this question can be viewed in Appendix C. Primary team colours have been assigned to each bar to allow faster comparisons for the more attentive football fans. The bars have also been ordered from lowest to highest and horizontal dotted cut-off lines have been added, assisting with interpreting which teams fall into the different possession categories in the next sub questions.

The visualisations for the sub parts to the second question can be observed in appendices D, E and F. Each of these graphics were produced using the same strategy. This involved picking a target feature, and producing donut and pie plots separately for each of three groups classified from the visualisation in Appendix C. The range of values contained within each group, have been colour coded consistently across all three visuals, allowing an interpreter to make simple cross comparisons building a common identity for each possession group.

Question 3 returned to a scatter plot (refer to appendix G); however, the data points are substituted for images of each club’s logo, and a trend line was drawn to emphasise the difference in market values between the biggest, medium valued and smallest teams. Visualising the sub question for this part required a different approach, with radial charts proving useful for drawing comparisons between the two features, visible in Appendix H. A dotted radial line was also included showing the average for all EPL teams, aiding the interpreter in making deductions about each club.

6. The process of generating further questions

A common theme found within the initial questions proposed was the data being limited to only the current or previous two seasons. By introducing additional historic data to the process, it would be possible to gain a better insight into how the EPL teams are current performing relative to their previous standards set, along with how each clubs' expectations have changed over time.

The following questions were then proposed to provide further insight:

- How have each of the EPL clubs values changed over the 2017, 2018, 2019 and 2020 seasons?
- Have the EPL teams been meeting their calculated total xP at the end of every season?
    - Has the xP been a good predictor of actual points for all the EPL teams?
- How consistent has each EPL team been in terms of their finishing position over the past few seasons?

7. Visualisation techniques considered for the further questions

At this stage, a lot more data had just been introduced, so alternative ways of visualising the data now need to be considered, so the visualisations would be not become overly cluttered. Appendix I shows the visualisation produced for the fourth question, where the team logos have been ordered for each season showing the ranking of club values at the end of each transfer window over the past four seasons. Each team's series is joined by a line plot in the club's main colour, making it easier to follow the trends for individual teams.

For question 5 a box plot was found to be well suited to the data, shown in appendix J. Every team involved in the EPL in the most recent 5 seasons, has its own box plot showing the minimum, maximum and median values for the point difference between the expected points and the actual points which were achieved. This effectively allows one to make a judgement on the accuracy of the xP predictor over the past 5 seasons, on an individual club level, but also generally for the whole league as well so the sub question for this query has been embedded in this visualisation too.

The final visualisation in this project, visible in appendix K, is a heat map based on the number of times a team has finished in each position. Darker shades of grey indicate that a club has finished several times in that position over the last four seasons, lighter shades are for finishing in this position once, and white reserved for never finishing in a specific position. This means that some teams who have recently been promoted to the EPL, have been discounted from the graph for this visualisation.

8. Critical discussion of the visualisation designs

Overall, nearly all of the visualisations produced fit their purpose well, however there are still some areas of improvement. One case is the visualisation produced for question 3.0 (see appendix G) where a line graph had been chosen. Even though this visualisation manages to express the patterns in the data, a better choice may be an ordered bar chart alongside the current EPL standings in the form of a table. This would allow an interpreter to group teams of similar club values together, enabling them to quickly make decisions whether each club are under or overperforming.

Another potential improvement could be either reducing the number of teams focused upon in appendix I, which was the visualisation produced for question 4. This plot is very busy, which makes it rather difficult to focus on individual team performance which was the purpose of this question. Another alternative solution could be implementing animations to this chart, where the interpreters mouse cursor hovering over a specific team would reduce the visibility of the other teams included, whilst enhancing the visibility of current team being studied. Due to the time constraints with this project this option had to be dismissed, however with more time this would certainly improve the quality of the visualisation.

One of the simpler improvements noted was for question 1.0 (Appendix B), where the text size of the team names often overlaps making it difficult to associate a few bubbles with their clubs. This could be resolved by a combination of making the text size smaller and also having custom positions for each of the labels, so they are not too closer together.

Finally, the colour gradient background present in the visualisation for question 5.0 (Appendix J) can be quite confusing to interpret, in terms of whether the positive and negative events are focusing on the clubs or whether the xP predictor’s performance is being assessed. The intent behind this implementation was to assess the quality of the xP predictor against the actual points achieved by each team, so it may be best to remove the gradient background, and as an alternative include horizontal interval lines showing how far the xP predictor is away from being perfect.

9. Reflection of the development process

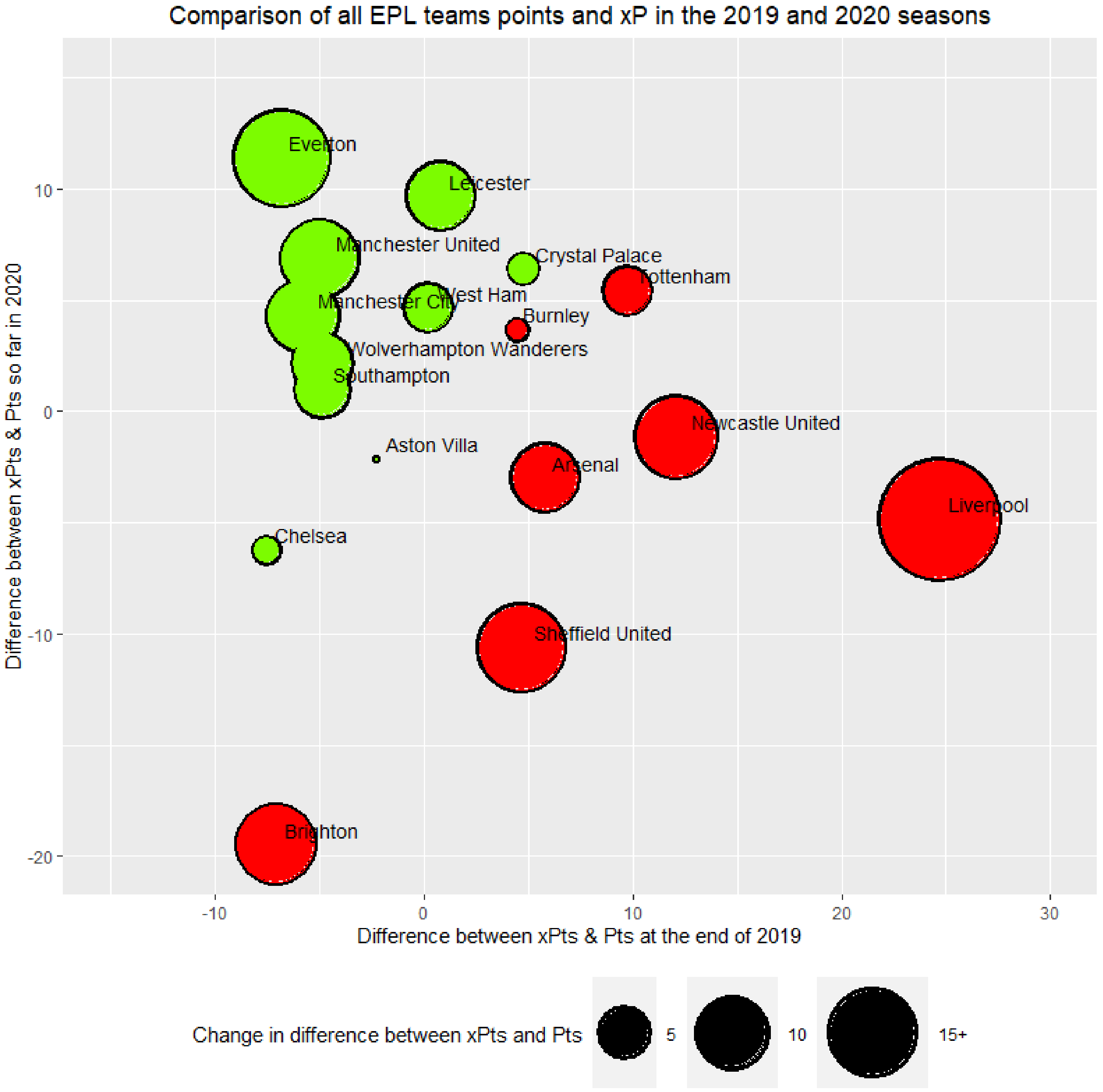
During this project, it was very important to keep an open mind when producing the visualisations for each question. There were a number of challenges faced throughout, such as deciding on which type of chart was most suitable for the data, generating ideas on how to present the graphics so that the information could easily be extracted, along with figuring out which R libraries work best for what transformations or visualisations are required. It was very common during this project, to use the method of trial and error to solve any problems encountered and assess the effectiveness of visualisations. Even after, there are still further improvements which can be made to the final visualisations produced which is all part of the developmental process. This worked well however the process was still very time consuming. It really showed, how much work is put into even the simplest looking visualisations in order to make them as effective as possible at communicating the data.

The ggplot library was so expansive, it was very easy to get carried away with all the customisations options. To prevent this from happening, I asked myself the following question frequently at the start of the process: “By adding this feature, am I improving the effectiveness of the visualisation, or am I just making the visualisation more complex?”. However, as I learned more about the field of data visualisation, I was able to make much quicker judgements on whether adding extra features was appropriate leading to it nearly becoming instinctive.

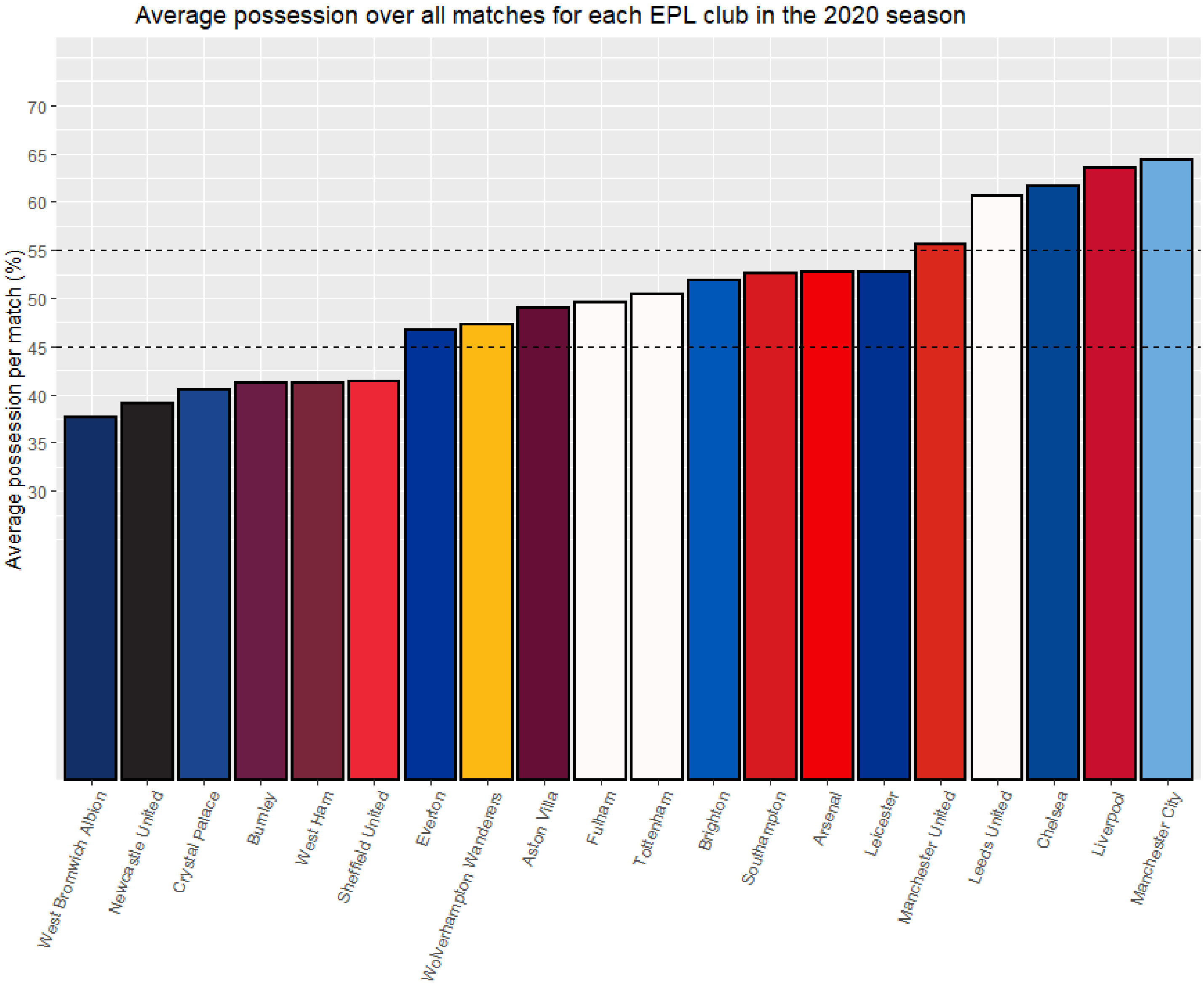
10. Appendix

Dataset name	Key features of interest	Other features present
Squad_Goal_and_Shot_Creation_2020.csv	Squad, GCA, GCA90	# Pl, 90s, SCA, SCA90, SCAPassLive, SCAPassDead, SCADrib, SCASh, SCAFld, SCADef, GCAPassLive, GCAPassDead, GCADrib, GCASh, GCAFld, GCADef, GCAOG
Standard_Squad_Stats_2020.csv	Squad, Poss, Gls, xG	# Pl, Age, Ast, G-PK, PK, Pkatt, CrdY, CrdR, Gls90, Ast90, G+A90, G-PK90, G+A-PK90, npxG, xA, npxG+xA, xG90, xA90, xG+xA90, npxG90, npxG+xA90
EPL_League_table_2017.csv	Rk, Squad, GF, GA, Pts, xG, xGA	MP, W, D, L, GD, xGD, xGD/90
EPL_League_table_2018.csv	Rk, Squad, GF, GA, Pts, xG, xGA	MP, W, D, L, GD, xGD, xGD/90
EPL_League_table_2019.csv	Rk, Squad, GF, GA, Pts, xG, xGA	MP, W, D, L, GD, xGD, xGD/90
EPL_League_table_2020.csv	Rk, Squad, GF, GA, Pts, xG, xGA	MP, W, D, L, GD, xGD, xGD/90

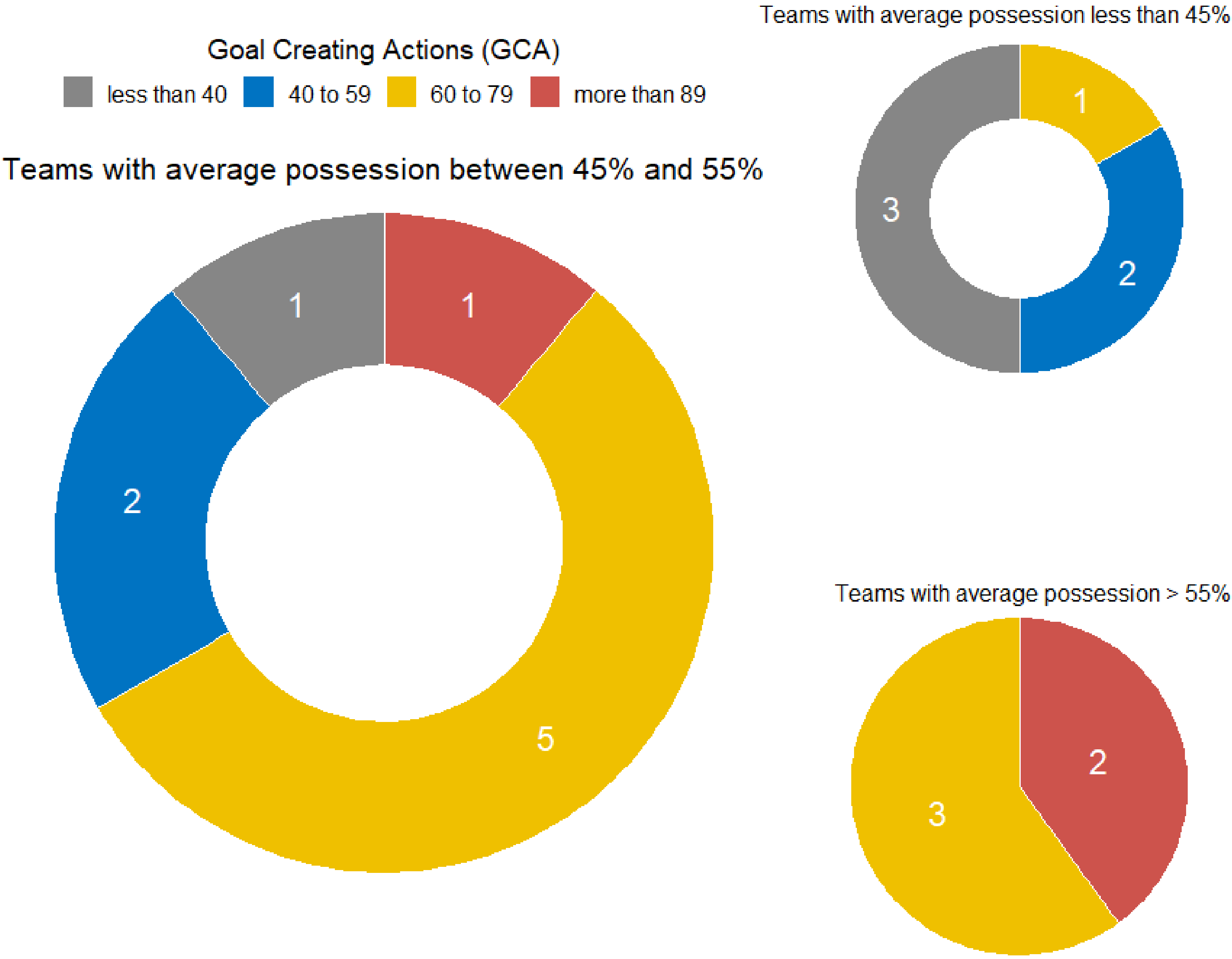
Appendix A: A list of the csv datasets used and their features, retrieved from online tables found on https://fbref.com



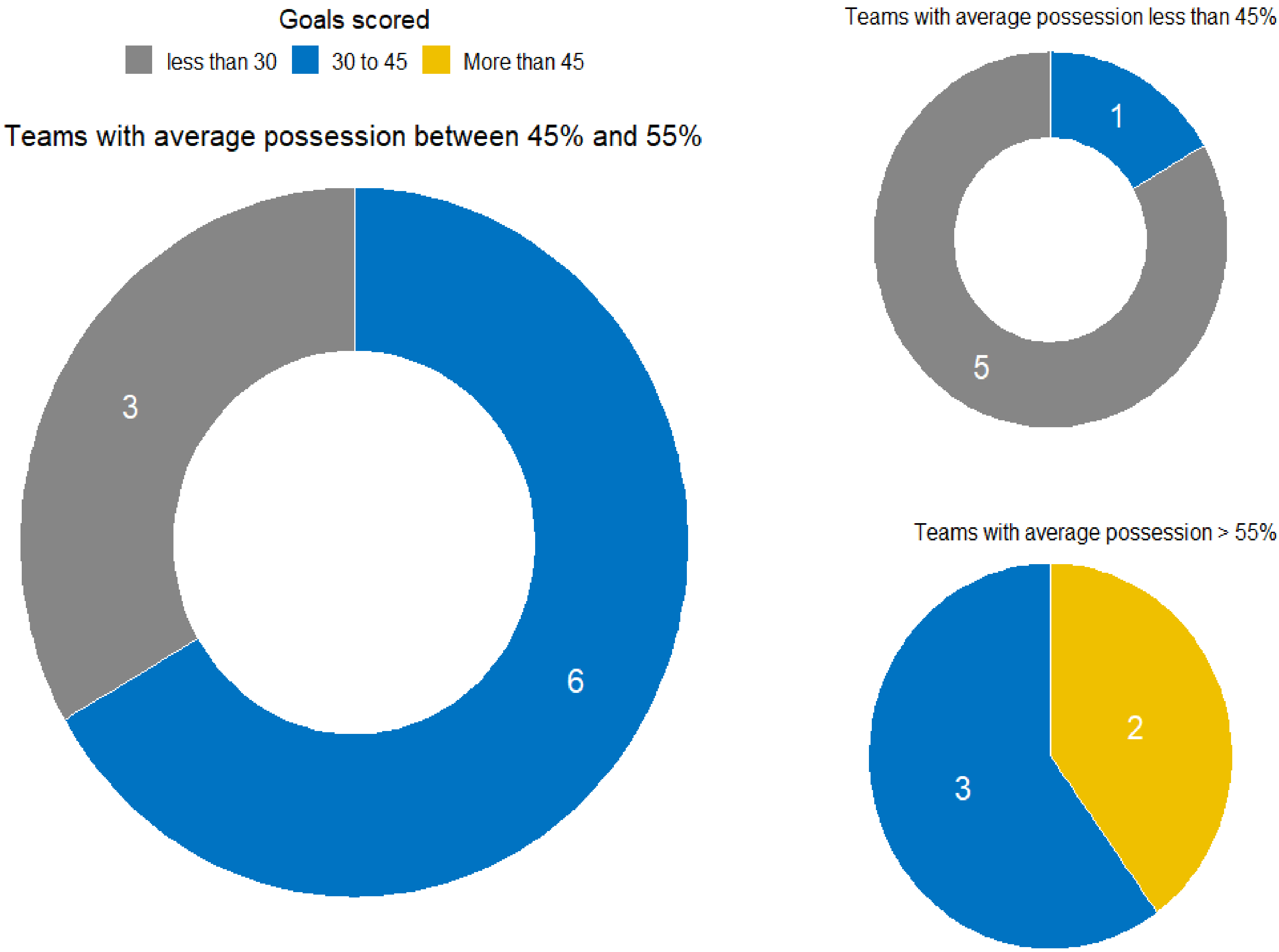
Appendix B: Visualisation for question 1.0. Green datapoints denote positive differences, whereas red datapoints show a negative difference in values.



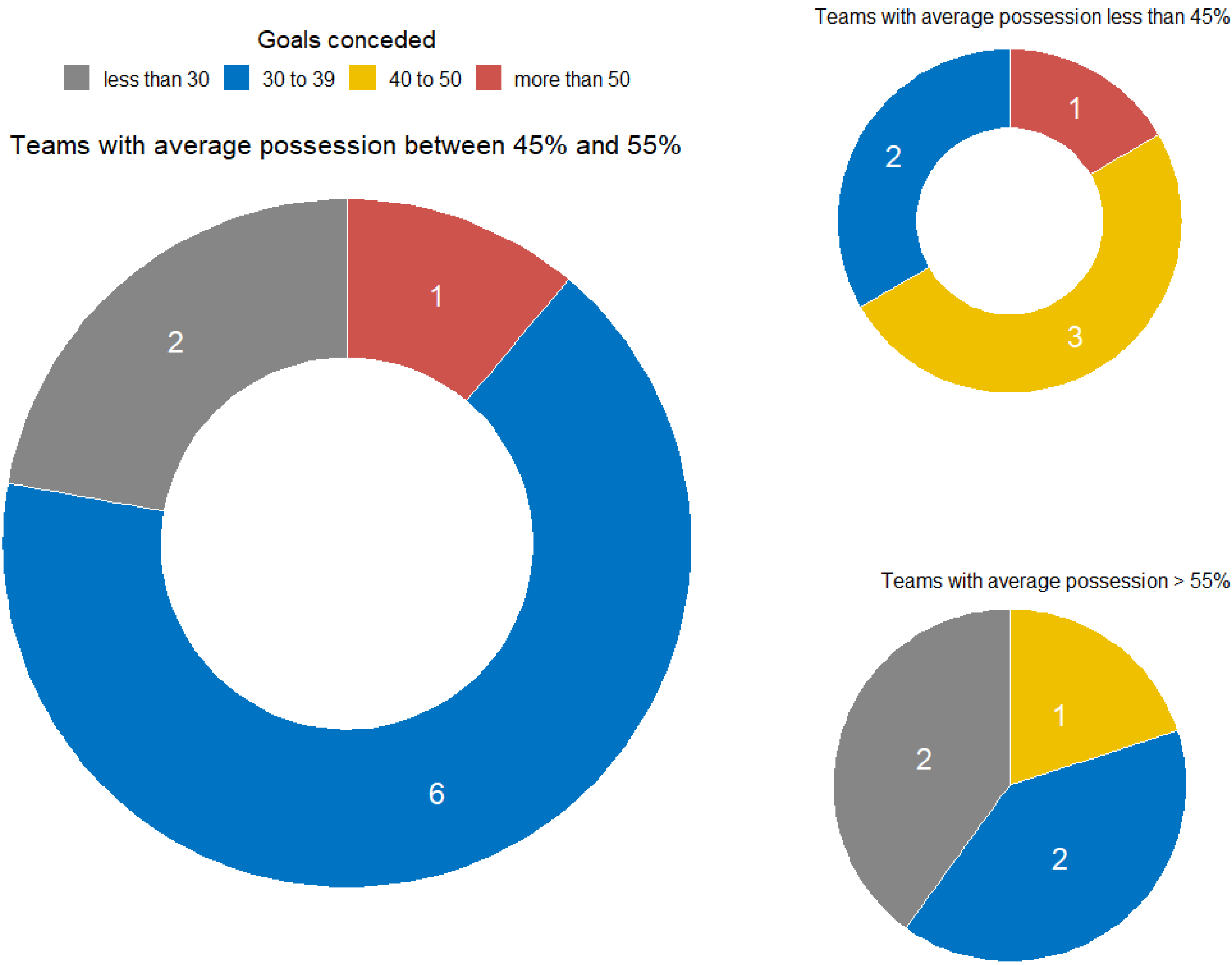
Appendix C: Visualisation for question 2.0. The clubs have been ordered from lowest to highest possession, with horizontal dotted lines found at 45% and 55%, indicating the cut-off regions for each of the possession groups categorised for sub questions 2.1, 2.2 and 2.3.



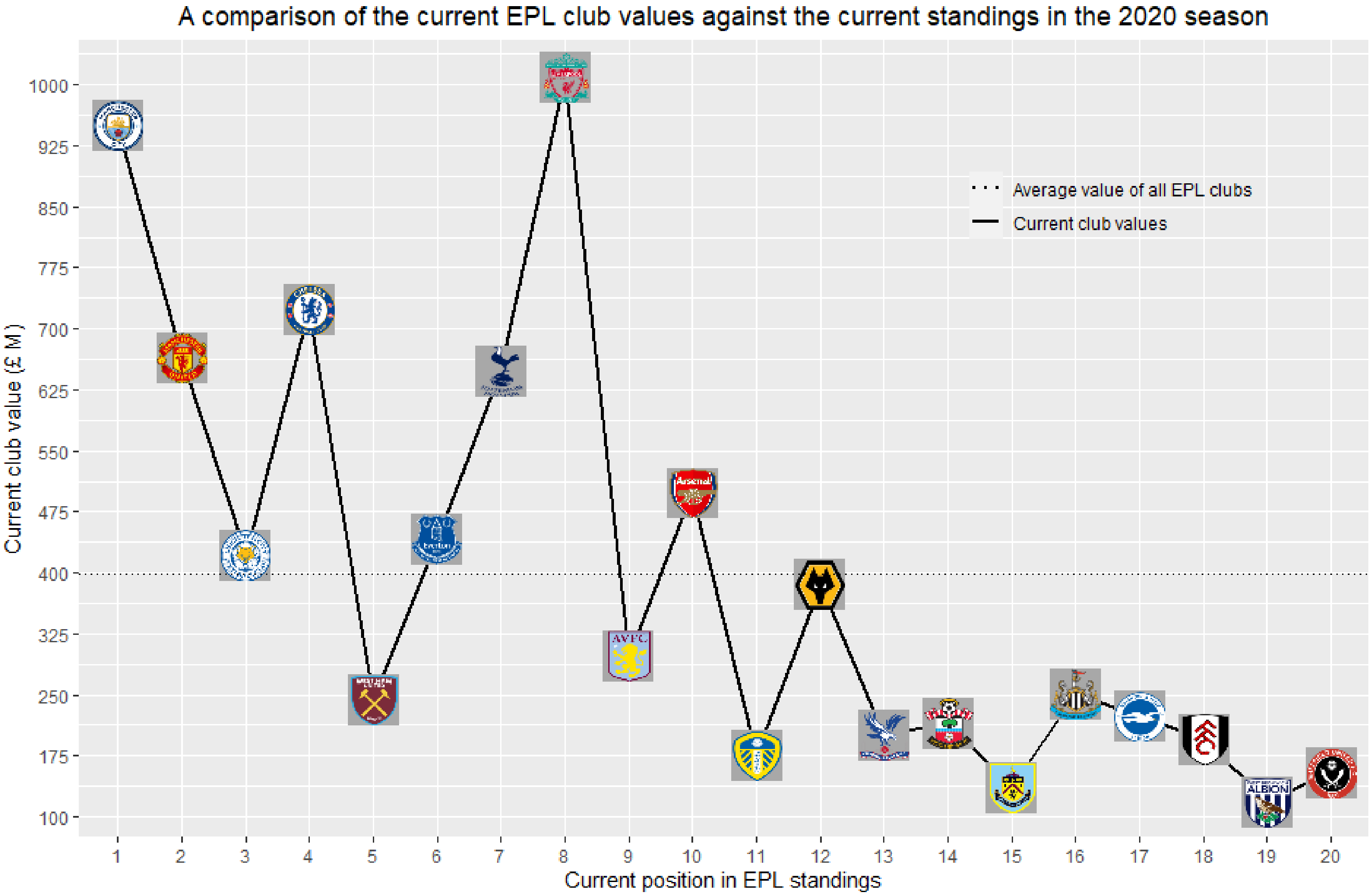
Appendix D: Visualisation for sub question 2.1. This visualisation uses two donut plots and a pie plot to display a frequency count for the number of EPL teams in each range of values in terms of GCA, for each of the three possession groups categorised. It appears that the teams with a higher average possession tend to have a higher number of GCA, although clearly there are still a few anomaly cases which could be investigated further.



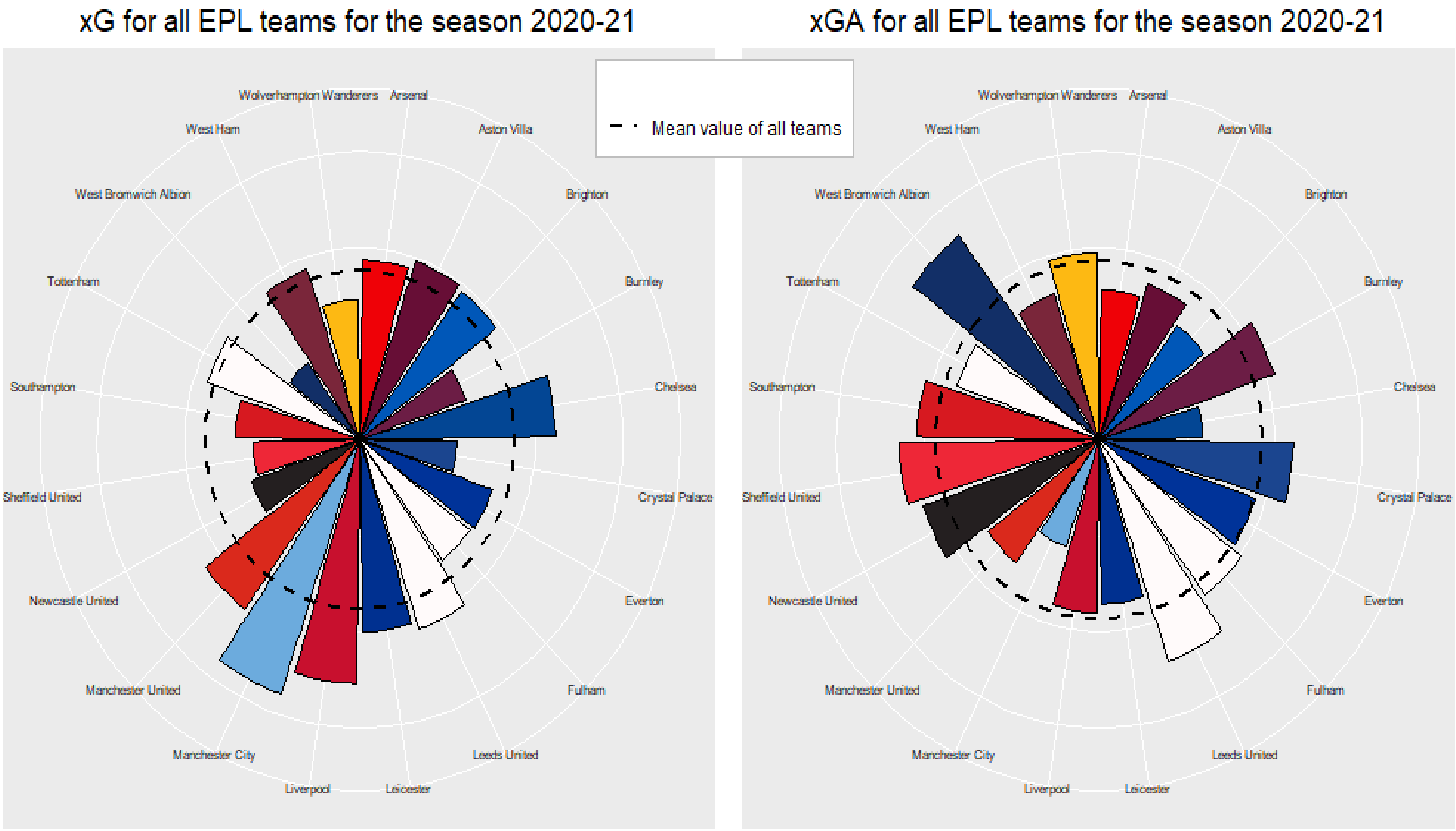
Appendix E: Visualisation for sub question 2.2. The same format has been used previously in question 2.1, however this visualisation is considering the number of goals scored by EPL teams in the 2020 season. A clear trend is shown as the teams with a higher average possession tend to score more goals, allowing for some overlap between the groups.



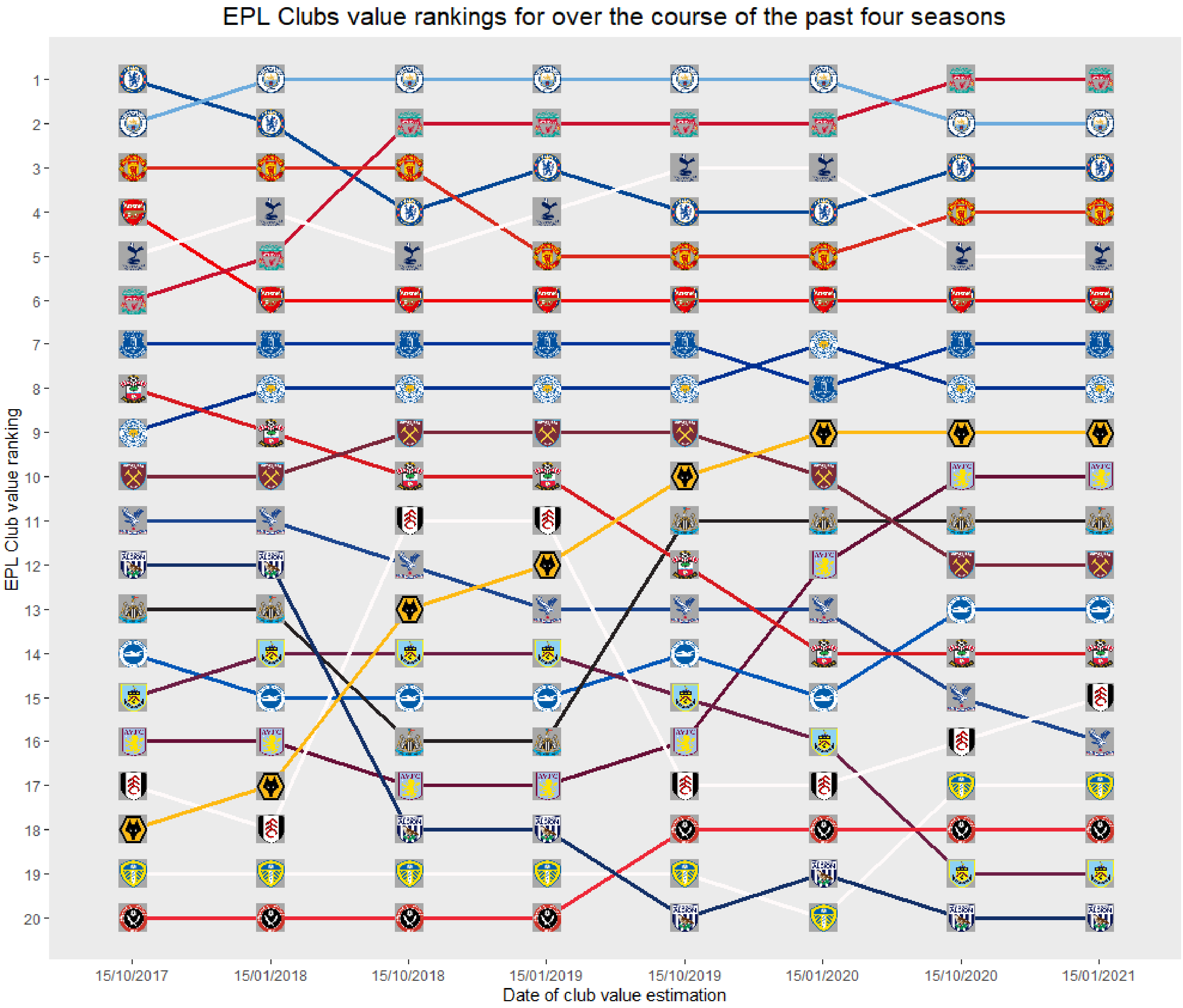
Appendix F: Visualisation for sub question 2.3. Alike to questions 2.1 and 2.2, this graphic concerns the number of goals each EPL team concedes, grouped by the possession classes subset during question 2.0. It is much more difficult to spot any trends within this visualisation, due to teams from different possession groups having a fairly even spread of the number of goals conceded with lots of overlapping.



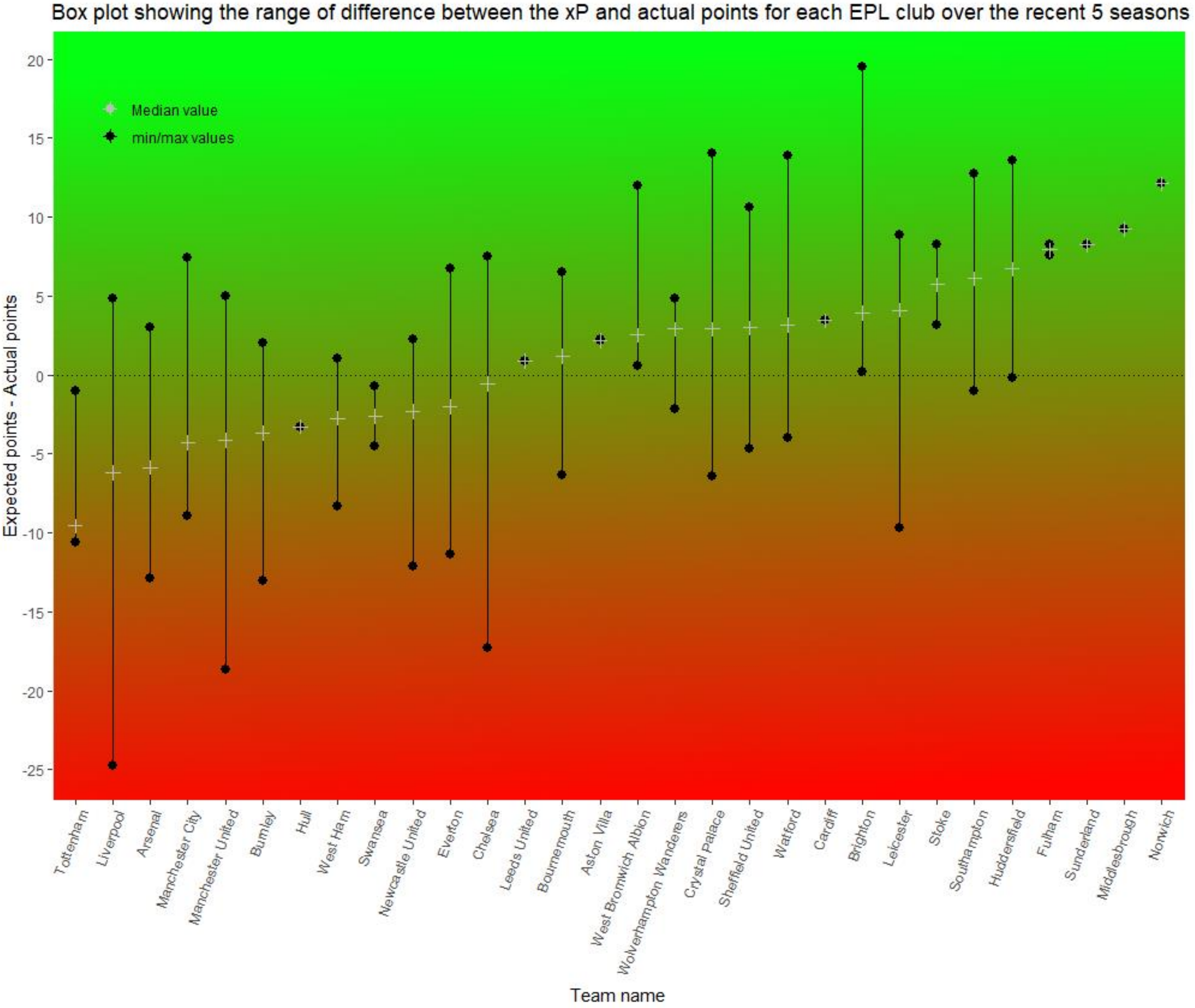
Appendix G: Visualisation for question 3.0. This graphic makes it relatively easy to see how club market value does not always translate into a better league standing, even though a general trend is shown, where the bigger clubs finish higher up the league table. Considering the first 10 positions in the EPL table, it appears quite atypical with several mid-valued teams beating the highest valued teams to the top spots. On the contrary, the last 10 positions in the standings are very much expected with many of the low-valued teams occupying these spots.



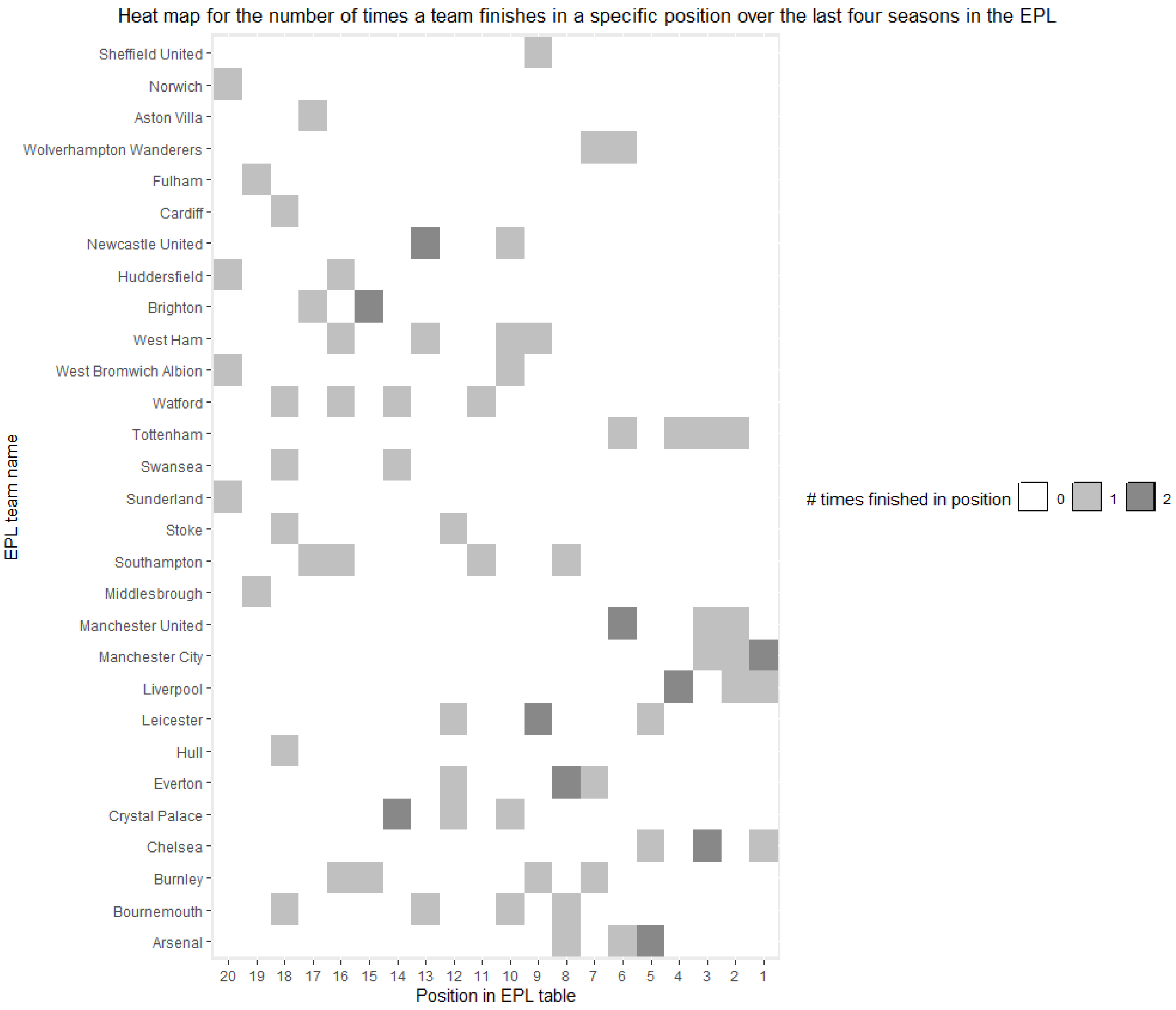
Appendix H: Visualisation for sub question 3.1. Two radial charts are shown side by side, allowing quick comparisons between individual teams. The average line also assists in classifying teams which are expected to perform well and teams which are not performing as expected, against the rest of the competition. It is interesting to note that Leeds United appear quite anomalous, being expected to outscore, and concede more than the average of all clubs, a property no other EPL club holds.



Appendix I: Visualisation for question 4.0. A series has been plotted for every current EPL team over the last 4 seasons at the end of each transfer window. There are a number of interesting observations which can be extracted from this visualisation, such as wolves rapid climb up the rankings, the consistency shown within the top 6 teams, along with the constant movement within the bottom half of the rankings.



Appendix J: Visualisation for question 5.0. A box plot has been produced for each of the EPL teams which have competed over the past 5 seasons, with their median, maximum, and minimum point differences between the predicted points and actual points achieved being shown. The background also displays a colour gradient insinuating positive values being a good event, and negative values being bad events.



Appendix K: Visualisation for question 6.0. A gradient heat map has been implemented for every EPL team over the past four seasons, with darker shades symbolising more times finished in that position. It is evident that not many teams have been consistent in terms of finishing position, with only half the league having finished in a position more than once and no teams finishing in the same spot three times.