Football Insights Report

In this report, I will outline my analysis of data from Premier League Games during the seasons: 2016/17, 2017/18, 2018/19. These datasets were sourced from datahub.io. Each season was a separate csv file, and I downloaded the three most recent that they had. I chose this as data has become an integral part of all sports games, and I felt it would be worthwhile to engage in some analysis myself.

The datasets consisted of 65 different data points for each game in each of the three seasons, of which there were 380 games a season. This gave me 65 columns and 1140 rows. The data consisted of:

- the teams playing,
- the goals scored,
- the shots taken,
- the corners awarded,
- the fouls committed,
- cards shown
- various different odds giving from different betting companies.

There were also quite a few columns that were not included, such as "Betbrain average over 2.5 goals". These can be found in the excel files, and their column name keys can be found on the Kaggle link below.

There were four insights that I attempted to gain from this data, as set out in the project brief. These four insights were selected in order to showcase a wide variety of programming techniques, and, on occasion, this approach led to somewhat of a decline in the pure statistics that I was calculating. But again, the goal was to showcase good statistical programming, opposed to being statistically sound. I also printed the outcomes so that the report would have a section for code, a section for plots and a section for outcomes.

The four insights were as follows:

1. I looked at the impact that being home and away has on the probability of outcomes when a team has had a player sent off.
2. I looked at the overlap between goals being scored in the first half and goals scored at the end of the game within the big six clubs.
3. I looked at how similar the odds that each betting company were giving for the different games.
4. I looked at different statistics arising from shots taken on goal:
    a. I looked at shots taken on target and off target while both home and away to investigate any differences.
    b. I looked at the correlation between corner kicks and shots that were on target and not scored.
    c. Finally, I looked at the proportion of corner kicks coming as a result of shots on target not scored.

I wanted to investigate these four insights, using programmatic techniques to analyse the data, print outcomes and give good visualisations. I should note that when naming my various different labels, I used a key to make it easier to comb through, as each variable was related to the different insights and there was no variable that was related to two different insights. This key was: card – 1st insight, goal – 2nd insight, bet – 3rd insight, shot – 4th insight.

I started by pre-processing the data. I wanted to clean the data up as well as possible. For the most part, the data was very well put together. Anything that was missing was missing in season sized chunks (for example, Ladbrokes Betting Odds were missing for season 18/19). As a result, this mainly consisted of deleting the columns that I wouldn't be using and renaming the ones I would be using, to make it easier to access.

Insight 1

For my first insight, I looked at the impact of being home and away in games where a team has had a player sent off. I must caveat this by saying that, from a statistical perspective, this insight doesn't really hold up to scrutiny. For example, I didn't take into account: the goals scored before the player was sent off, what time the player was sent off at, etc.

To begin, I created a new dataframe consisting of the sum of home and away red cards. I used this to then check the maximum red cards in a given game, which was 2 here. This meant that a caveat had to be added whereby I had to extract out games where there were 2 red cards given, but both teams had a player sent off. I wanted to only investigate games where there was a disadvantage. Had there been games with three red cards, I would have had to add an extra caveat whereby I would have to factor out games where there was 2 red cards given to one team, and 1 red card to the other, as there would be games where a team was given a red card and still had an advantage. This also gave me my first issue, as I initially didn't include the [0] part, and so when I printed, python was returning the type aswell, which I didn't want.

I then created two new subsets, containing the games where the home team received a red card and when the away team received one. I had an issue here in that I tried to just create these datasets without making them copies. This gave me a "SettingWithCopyWarning". After some googling, I found a solution[3]. I then changed the value of each result from: 'Draw' to 0, 'Home' to 1 for home red cards and -1 for away red cards, and 'Away' to -1 for home red cards and 1 for away red cards. This gave me all the results in numerical form. Following this, I conducted a T test to see if there was a significant difference in the means of the outcomes of the games. I then created an auto outcome response based on the p-value of the t test. The p value did not meet the threshold, and so it appears that being home or away has no impact.

Insight 2

In this insight, I looked at the overlap between goals being scored in the first half and goals scored at the end of the game within the big six clubs. The goal here was to create a linear regression line to study this. Ultimately, this was a study into how important momentum was for each of the big six teams.

For this insight, given that I would be focusing on all the games of specific teams, I knew I would have trouble in combining the team's home and away games. Hence, I created a function[4], using some guidance from Google, that swapped the rows of different columns. I then isolated the games that Liverpool played (Liverpool being the first team I chose), both home and away, and created dataframes for each. I then swapped the 'Away Team' and the 'Home Team' columns, the 'Full Time Home Goals' and 'Full Time Away Goals' columns and the 'Half Time Home Goals' and 'Half Time Away Goals' columns, and switched the names around for each. So, for example, in the Liverpool Away Teams dataframe, I would have Liverpool in the Home Team column and Liverpool's half time and full time goals being in the home half time and home full time goals columns.

This was so I could amalgamate both dataframes into one and I would have all of Liverpool's Games, Half Time Goals and Full Time Goals all in the same columns. Who was Home or Away in this insight didn't really matter.

I then created a linear regression line and plotted it on a graph. The graph showed me various different results of games, but did not show the frequency at which each result was given. The linear regression line gave me a good indicator of how many goals would be scored at full time, given the goals scored at half time. I also added some labels and titles to the graph.

I did the same thing for Man United, Man City, Chelsea, Arsenal and Tottenham. I then added each linear regression line into a graph and plotted it, giving each time a different colour to tell them apart. This showed some interesting results. For example, Arsenal are likely to score more than Chelsea if they haven't scored at half time. However, if there either side scored two goals, Chelsea would be more likely to end up with more goals at full time.

Insight 3

In this insight, I looked at how similar the odds that each betting company were giving for the different games. These were given by odds to 1 of each outcome happening. For this insight, I had the issue that was mentioned previously in that there was no data for Ladbrokes for the 18/19 season. This left me with a choice on whether to impune the data, delete Ladbrokes from my dataset or complete the insight on only the first two seasons. Given the quantity of data that was missing, and the fact that I still had a lot of games to work with, I felt that it would be better if I only looked at the seasons 16/17 and 17/18. So, I deleted all the data from the 18/19 season.

I then computed the correlation between each of the six betting agencies for home games, away games and draws, and plotted them using the seaborn library. As could have been expected, there was an extremely high correlation rate between each – with all being above 0.95.

However, I wanted to see if there was any significant difference between the odds given. So I undertook an ANOVA test, to see if there was any difference between the six betting agencies, and created an automatic conclusion based on the p value given. Surprisingly, there were significant differences between the draws and the away win odds, but no difference in the home game odds. This would appear to imply that there is more consensus on the likelihood of home teams winning, while less so on away teams.

Insight 4

Finally, I looked at different statistics arising from shots taken on goal. I created two empty lists, of shots not on target home and away. I then used iteration to manually compute these values – taking the shots on target away from the total shots. I then turned these into dataframes and changed the names of the columns. I had some issue in changing the name from the auto name given when turning a list into a dataframe[5]. I then combined both dataframes into one dataframe so as to make it easier to compute.

I plotted the shots taken on target vs off target for both home and away using a violin plot, which gave me some good information about the whole of their data. Interestingly, there is seemingly an advantage towards home teams as they scored higher on nearly every metric. I had issues in being able to accurately label the violin plot, and I had to settle for a text plot. I used this to help me[6].

I then did a similar thing as before to get shots on target not scored for both home and away, however I used a while loop instead of a for loop. I did this to compare shots on target not scored to corner

kicks. This is because, as one could imagine, a lot of shots on target that are not scored would result in a corner kick. I wanted to look at how many corner kicks could possibly come from shots on target.

I firstly computed the correlation, which was quite low. There was 41% for home games and 32% for away games. I then summed the total shots on target not scored, the total corners and the proportion between them, which was around 56%. Finally, I printed out some extrapolations of this data, in that a large plurality of corner kicks must come from non-shot related incidents.

Finally, I deleted some of the datasets that were already contained in other datasets, for ease of exploring through them.

Ultimately, some of the insights I received were quite surprising. I was especially surprised at the ANOVA test for the betting odds. I do realise that some of the data I completed weren't as statistically sound as they could have been, but I'm happy that I adequately showed a good range of programming techniques and that I showed that I understood the statistical programming to a good extent.

References:

1. https://datahub.io/sports-data/english-premier-league
2. https://www.kaggle.com/datasets/caesarlupum/betsstrategy
3. https://www.dataquest.io/blog/settingwithcopywarning/
4. https://www.statology.org/swap-columns-pandas/
5. https://www.kdnuggets.com/2022/11/4-ways-rename-pandas-columns.html
6. https://stackoverflow.com/questions/33864578/matplotlib-making-labels-for-violin-plots