CO3093 COURSEWORK 1 Report

Big Data & Predictive Analytics - Simulation-based & Regression Models

Ihtasham Chaudhry

Department of Informatics
University of Leicester
23rd February 2018

Question 1

1.1

It is important to consider missing values in our data set and to filter out the columns based on this information so that we have all the information we need to make predictions and have *clean* data. In the case of our problem (predicting a winner), the most important factors that we must consider are win/loss statistics and goals scored/conceded statistics.
We can then draw some conclusions from the data such as:

1. The average number of goals scored per match throughout the tournament by each team playing at home is 1.49 and away is 1.18. We also notice that the median of these values falls close to the minimum values which indicates that the data on this column is positively skewed and has a longer tail towards higher values.

2. From this data we can also observe that on average teams win more games playing home than they do playing away, The mean (μ) of wins at home for each team is larger than the median of the dataset, this indicates that the distribution is skewed towards large values. The standard deviation (σ) is very small relative to the min and max values, this indicates that the distribution has "long tails".

3. Following from the previous point, the same observation can be made regarding the number of "FTHG" or "Full Time Home Goals" by each team per match. The data again is skewed towards larger values.

1.2

As we are only considering two teams; Manchester United and Manchester City, we can further filter the data and extract only the games played by both of those teams where they are playing either home or away. When we accumulate the data by teams and their home and away games to see how they perform for each category. After doing this we can draw some analysis from the data.

Table 1: Mean goals scored per game over the season (higher is better)

	Home	Away
Man Utd	2.25	1.83
Man City	3.50	2.33

Table 2: Mean goals conceded per game over the season (lower is better)

	Home	Away
Man Utd	0.41	0.91
Man City	0.75	0.75

We can see that over 24 games played by both teams over the course of the season, Man City has a higher average of goals both in the home and away side compared to Manchester United, and have conceded a higher average of goals home but a lesser average of goals home compared to Manchester United.

1.3

We can visualise this information and compare the offensive and defensive performance of the two teams. To do this we will consider goal scores. To measure the defensive performance of a team we can see how many goals were conceded by the team playing home, lower is better in this case, i.e we want to see that the distribution is skewed towards less goals. And in contrast to this to measure offensive performance we can see how many goals they scored away, higher is better in this case and alternatively we want to see that the data is skewed more towards the right to see which is the better team.

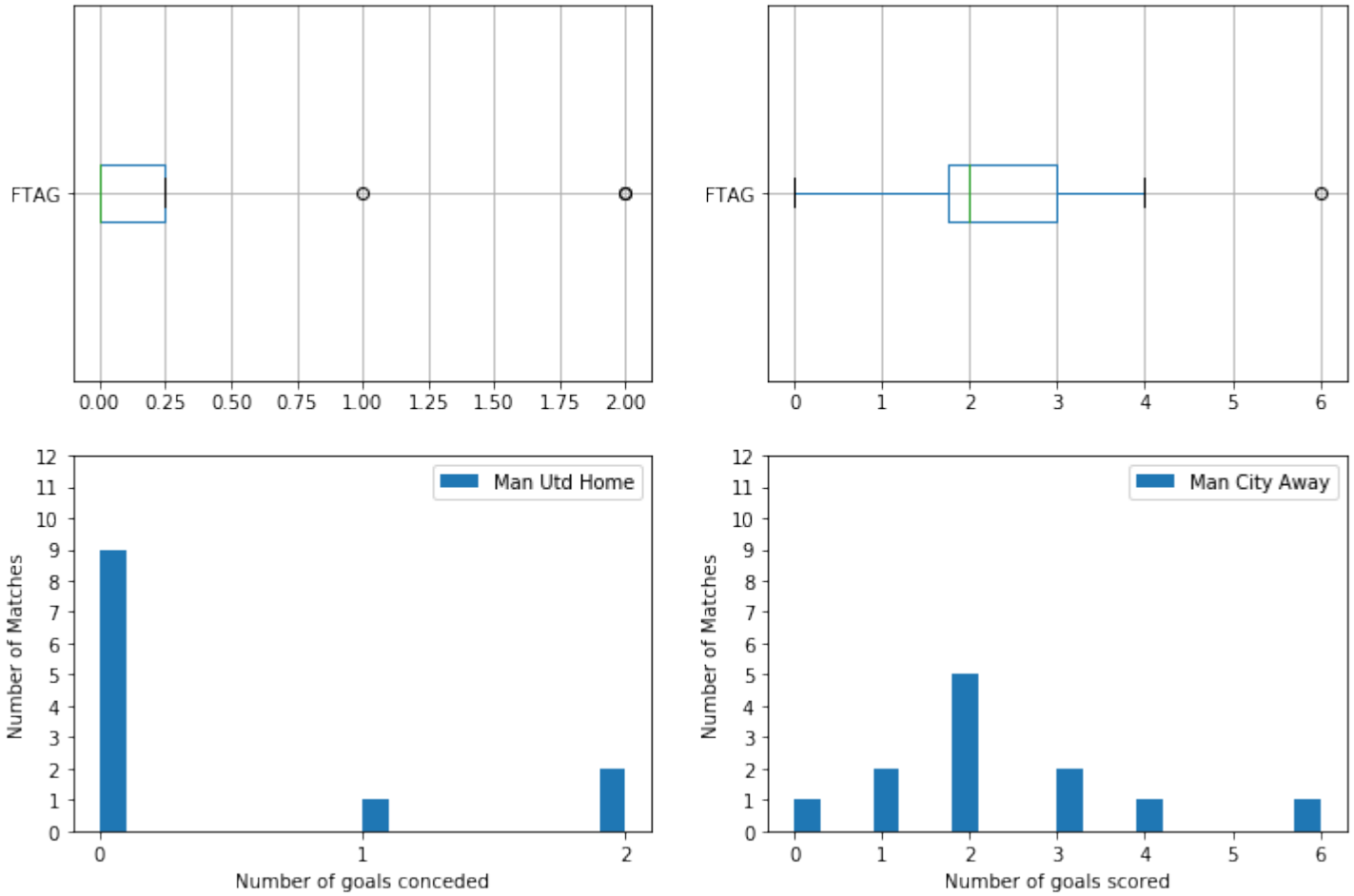

Figure 1: A comparison of M.U's Home defence against M.C's away offence.

From this graph we can see that Manchester United's defence is quite strong, this is because out of 12 games playing at home they only conceded goals in 3 out of those 12 games. Furthermore, we can also see that Manchester City's offence is fairly strong and we can tell that the number of goals they score on the away side is almost distributed in a Poisson distribution manner. From the boxplots we can see that Manchester City's goals away were between 2 and 3.

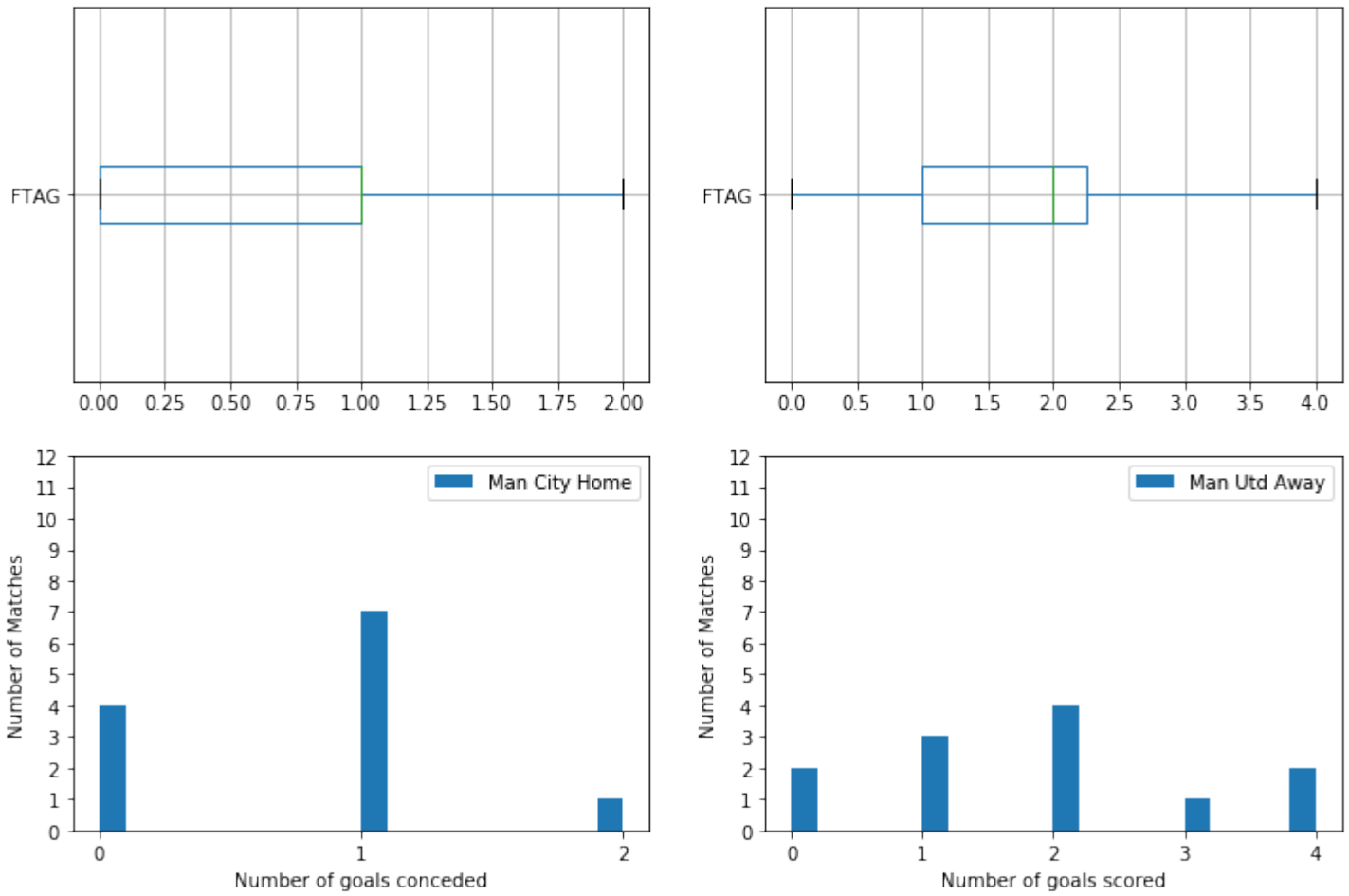

Figure 2: A comparison of M.C's Home defence against M.C's away offence.

In contrast to the above graph, we can see that Manchester City's home defence is worse than Manchester United's home defence. We can assert this by seeing that Manchester City conceded at least 1 goal in 8 matches compared to the 3 that Manchester United conceded at home. Furthermore, we can also see that Manchester United's goals scored away are more spread out and have a higher mean than Man City's. However we can also see that at most Man Utd only scored 4 goals in any one game over the whole season compared to the 6 scored by Man City which could indicate a higher level of offence.