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.
We can then draw some conclusions from the data such as:

1. Manchester City has the highest amount of wins (11) playing home compared to any other city. While West Brom has the lowest (2).

2. 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.

3. The maximum number of goals scored in the tournament in one match was 7 goals and the maximum number of shots taken in one match was 35 shots.

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 better win rate of 87.5%, and a higher average of goals both in the home and away side compared to Man Utd.