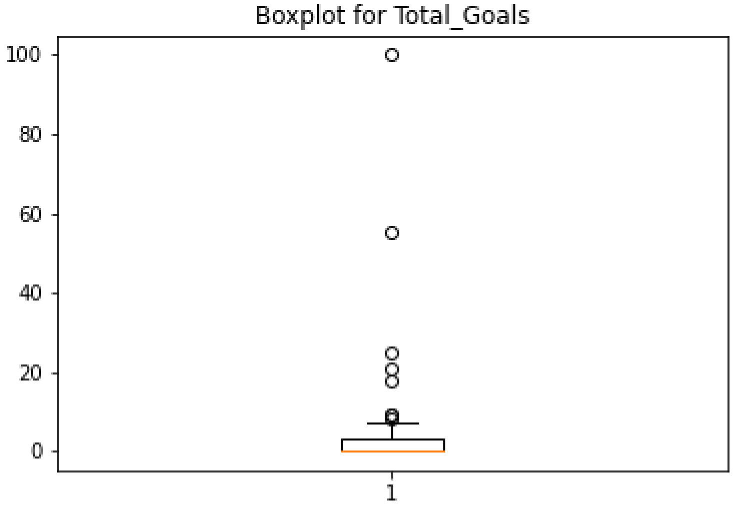
1. A discussion of the appropriateness of the regular expression used in Task 3, including some examples of where you might expect it to perform poorly

   My ans: r' (\d+) [-] (\d+) '

   I used the capturing groups so that I was able to extract the number from each side and sum them together to get the total_scores. I avoided the consideration of ‘-’ while doing the calculation.

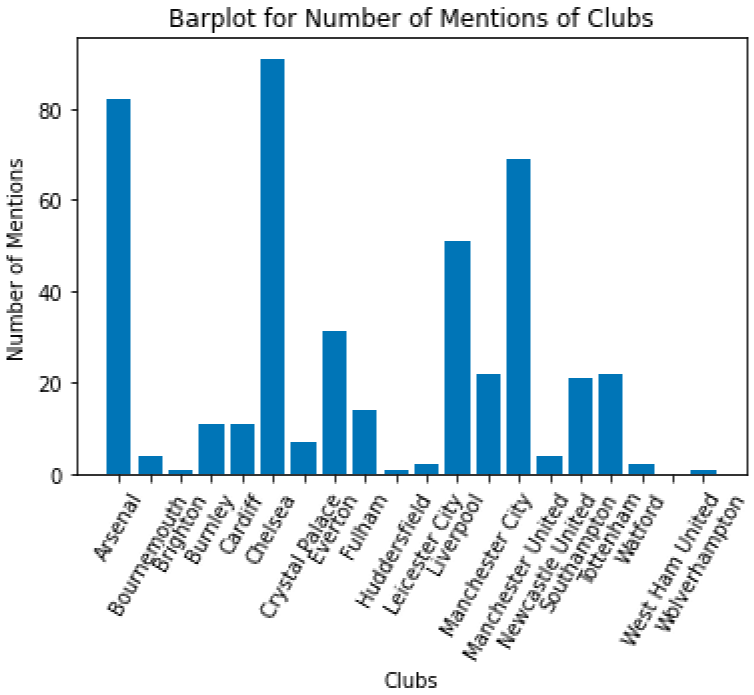
   Where expect to perform poorly
   However, there might be spaces before or after the hyphen, ie. 5-   2, 4   - 2;
   Even with matching the pattern correctly, the match score might be for bets instead of results, ie. bets on 4-2 (instead of real match scores).

2.
   Task4



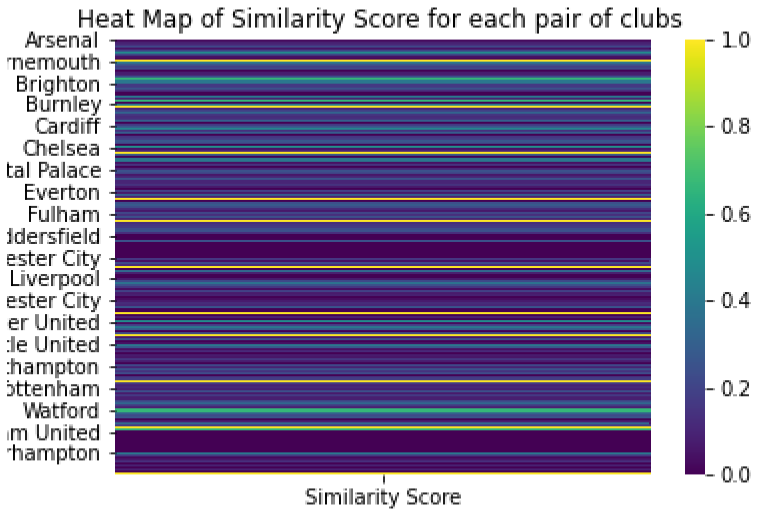
   The boxplot indicates that the distribution of values of total-goals is significantly skewed towards the lower values, except for several (around 5) outliers.

Task5



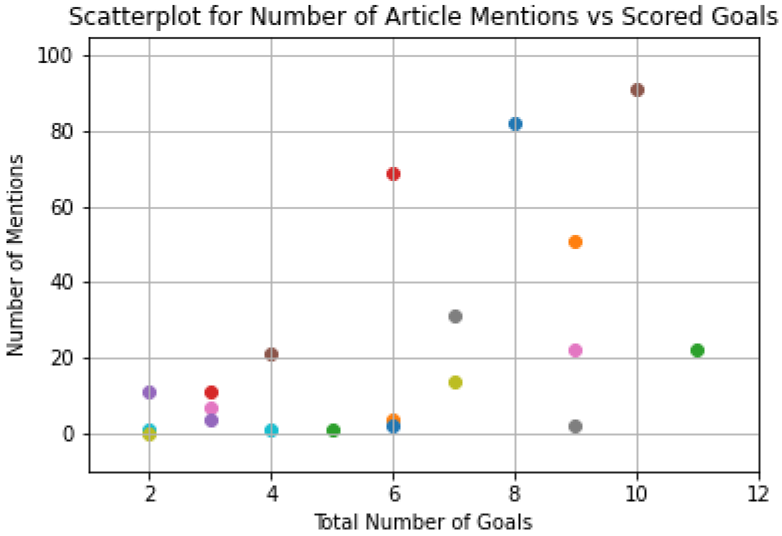
The above barplot, which is multimodal, shows us the distribution of the number of mentions of each club,
We can tell that around 4 clubs have significantly high mentions, while half of the rest (around 8) are not deviating too much from each other, and the left half (around 8) are extremely low.

Task6



There are a lot of distinct yellow lines, where the similarity score is 1 (to its own). These lines divide each club into areas, within which there are many different colors of lines indicating the degree of similarity between the club and other clubs. The lighter the color is, the stronger the similarity between clubs is.

Task7



In the scatterplot, there's the potential relationship between Total Number of Scored Goals and the Articles Mentioning. Around 14 teams with a higher number of goals tend to have more articles mentioning. However, we can’t ensure the significance of the relationship.