**Assignment 1 – Task 10**

**Task 3**

While able to extract with perfect recall any match scores formatted as two positive numbers lower than 100 adjoined via a hyphen, the constructed regular expression fails to locate orthographic scores or any that does not strictly adhere to the above stipulation, both of which being potentially likely occurrences, and as such is limited in score retrieval power. Additionally, and although complemented with code that will identify any two numbers separated by hyphens as potential score candidates and subsequently discard those whose length is greater than two digits – thereby ensuring that the last and first two digits of two adjacent years, for instance, shall not be misinterpreted as a match score – the present regular expression will incorrectly select number ranges, ratios, probabilities, dates, year abbreviations and such in the described format as scores, and evidently did so for several articles upon which the program function was tested.

**Task 4**

Table

Description automatically generated

The data is heavily skewed such that the median recorded soccer match score is 0, with over 75% of scores deemed as legitimate (those less than 1.5 times the interquartile range above the third quadrant) below 5 while all are below 10. Due to the limitations in score extraction pertaining to the regular expression in use, 5 of the 7 outliers identified are of such exceptional magnitude that they are undoubtedly erroneous, the majority of which are clustered around 20 with the largest recorded as 100.

**Task 5**

Chart, bar chart

Description automatically generated

Roughly half of the recorded teams were mentioned at least once in less than 20 articles, another approximate half of which had less than 10 mentions, with 8 teams in total having less than 5 mentions and 1, West Ham United, with 0. Of the 5 teams with more than 50 mentions, 3 had approximately 70 or more, the largest of which, Chelsea, has slightly over 90 and thus was greater in magnitude than the median number of mentions, around 10, by a factor of over 9.

**Task 6**

Graphical user interface

Description automatically generated

The overwhelming majority of similarity scores between soccer team pairs – quantified as the quotient of the number of articles mentioning both teams together, over the summation of the number of articles mentioning the first and second teams separately, with the entire calculation subsequently multiplied by 2 – falls within the range of roughly 0.1-0.4, primarily 0.2-0.3, with only 3 distinct team pairs possessing larger scores, all of which approximate 0.7. 2 had scores of 0 across nearly all of their corresponding pairs. Due to the alphabetisation of team names, the data shape is symmetrical across the diagonal from the top-left to bottom-right corners, signified by a bright red line indicating a maximum similarity score of 1 between each team and itself.

**Task 7**

Chart, scatter chart

Description automatically generated

There seemingly exists a moderate positive linear correlation between the number of total goals scored by each team and article mentions, as for each data point an increase in magnitude across the x-axis appears generally to correspond to an increase in magnitude across the y-axis. Contrary to such hypothesis, however, is a cluster of four points occupying the top-right corner for which a moderate linear increase evidently does not apply, perhaps due to the teams in question’s popularity or noteworthy performance warranting more article mentions than the average team.