The regular expression used in Task 3: [^\d](\d{1,2}-\d{1,2})[^\d]

This regular expression searches for a hyphened pair of numbers, limiting the number on either side of the hyphen to be less or equal to 2 digits (0-99). This expression also checks that such hyphened pair of numbers is enclosed by non-numerical elements on the left and right boundaries. The search conducted by re.findall produces the following outcome:

\*xx-xx\* or \*x-x\*, where x is between 0-9, and \* is any non-numerical character.

The disadvantage of this regular expression may include the “greediness” of the capture group, where all combinations of the number appear in the search result (e.g. 23-10 may appear as 2-1, 2-10, 3-1, 23-1 .etc), requiring further steps to filter the result obtained.

Another significant disadvantage is the fact that dates are sometimes expressed in similar forms, and this expression cannot distinguish between these dates and actual scores. e.g. 12-1 may refer to first of December, or 12th of January, or the score of 12-to-1.

Task 4:

Chart

Description automatically generated

It’s quite clear that the median is 0 in this boxplot, with a small interquartile range. This suggests that most articles either did not mention the match score, or the reported scores are to be quite small (<10 based on visual estimation). Such information implies that it may be uncommon for soccer matches to result with a great number of total goals, leading to the conclusion that it’s difficult to goal in soccer matches so that fewer goals are expected.

Task 5:

No certain pattern can be stated from the bar chart generated in task 5, but it’s obvious that

Chart, bar chart

Description automatically generatedcertain clubs get way more media exposure than the other. Arsenal and Chelsea were mentioned in more than 80 articles, whereas West Ham United appeared in 0. The reason for such inconsistency is unclear here.

Task 6:

A picture containing chart

Description automatically generatedThis visualisation provides a measure of how common a club appears in the same article as another club. The similarity scores seem to lie within the purple range (0.2-0.4) for most comparisons. However, club CAR and HUD seem like they never appear in articles alongside other any clubs, implying the clubs being quite unique.

WAT always gets mentioned with club BHA and BOU, indicated by their high similarity scores, suggesting they might share things in common.

Task 7:

Chart, scatter chart

Description automatically generatedThis plot compares media exposure to the performance of the teams. Although no clear trend can be identified, it seems like better performance can lead to more media exposure. As the goals scored by team increase, the scatter plot becomes more loosely distributed as some ends up with more exposure, and some remains unchanged. This suggests that better performance unnecessarily means more mentions in articles, but it brings chance to increase media exposure or potential chance to fame, although unguaranteed.