TASK 10

1. A discussion of the appropriateness of the regular expression

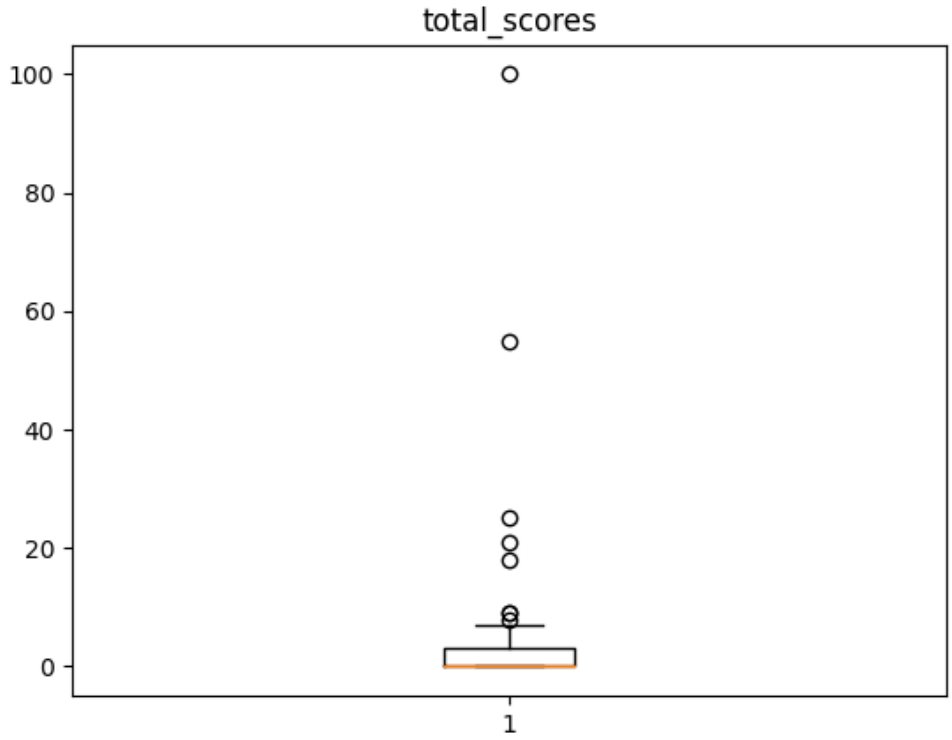
The regular expression I finally used is `[0-9]+-[0-9]+`. There I totally have tried three expression. The first was `[0-9]*\-[0-9]*`. However, it performs poorly when it met 'non number - non number'. For example, in 001.txt, there has 'quarter-finals' which would be matched by the first formulation. Because the '*' represents zero or more. Then I tried `[0-9]+\-[0-9][0-9]+`. But it still didn't work well, especially when it met '2009-07', it would match it with '09-07'. Finally I used `[0-9]+-[0-9]+`. It was not best, but easy to subsequent process. There I found all scores were composed of 'number-number'. But the years was similar. So I just used `[0-9]+-[0-9]+` and eliminated the year. The code is as follows.

```
for filename in os.listdir(articlespath):
    data_path = os.path.join(articlespath, filename)
    # print(data_path)
    data_txt = open(data_path).read()
    # expression error '[0-9]*\-[0-9]*'
    data_tmp = re.findall('[0-9]+-[0-9]+', data_txt)

    if len(data_tmp) == 0:
        total_data.append((filename, 0))
    else:
        max_score = 0
        for tmp in data_tmp:
            score = (tmp.split('-'))
            assert len(score) == 2
            if int(score[0]) <= 99 and int(score[1]) <= 99 and int(score[0]) +
int(score[1]) > max_score:
                max_score = int(score[0]) + int(score[1])
                # if filename == '014.txt':
                #     print(max_score)
        total_data.append((filename, max_score))
total_data = sorted(total_data, key=lambda t: t[0])
total_scores = [data[1] for data in total_data]
```

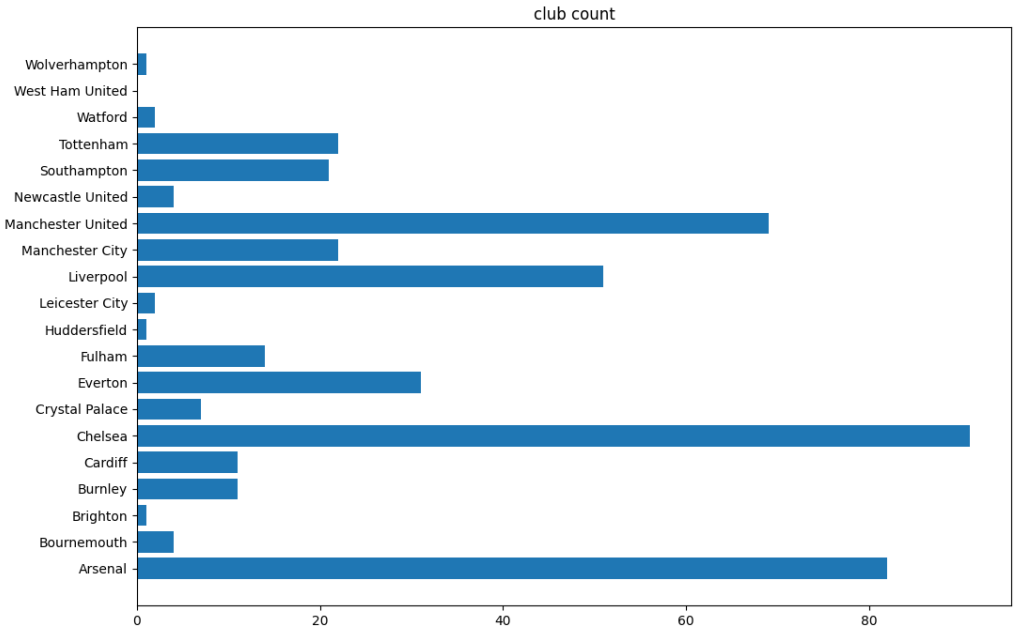
2. An analysis of the visualisations produced in Tasks 4, 5, 6 and 7

The image of Task4 is as follows.

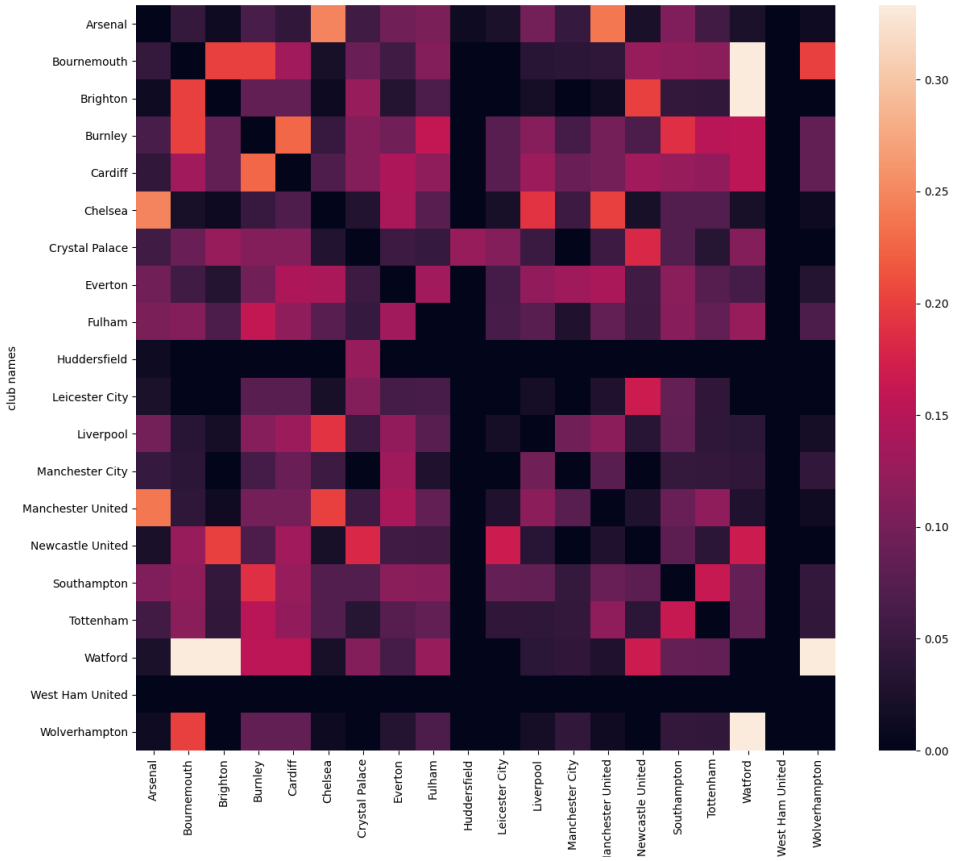


From the image, we can find the most scores are in the range of IQR. Only little is outlier.

The image of Task5 is as follows.

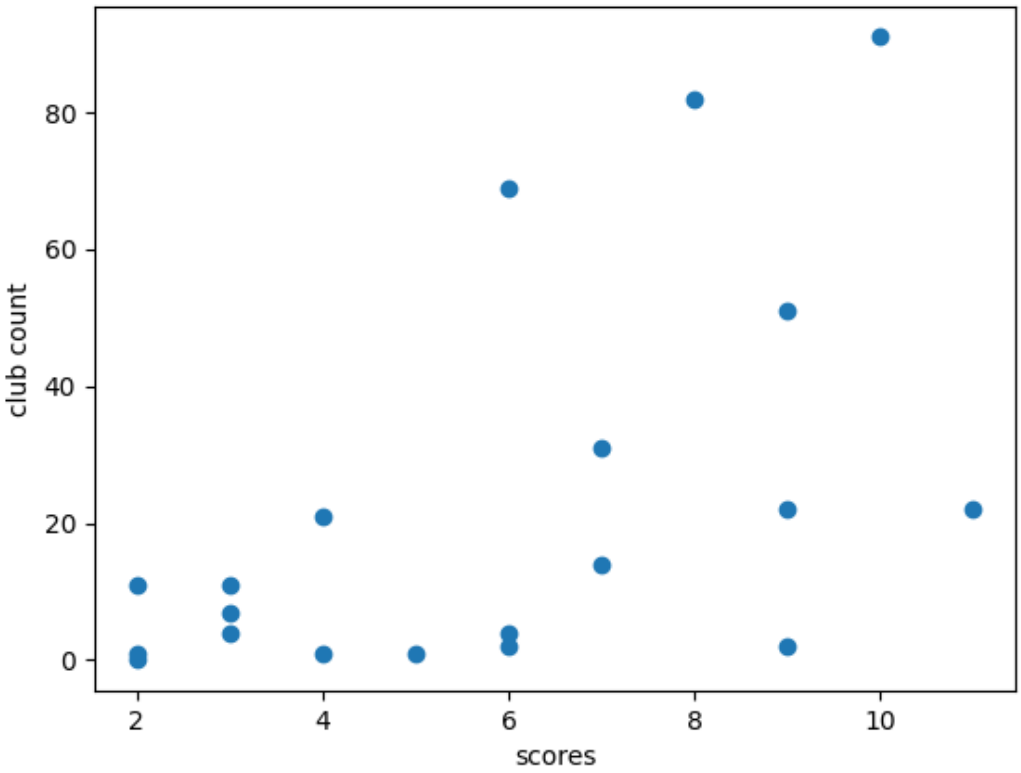


The image of Task6 is as follows.



We can find that there are some teams that the Huddersfield and West Ham United is different with others.

The image of Task7 is as follows.



We can find that the number of times a team is mentioned in the media is related to its performance. For example, when scores are more than 6, the count of clubs is more than 20 mostly. But when scores are less than 6, the count of clubs is less than 20 mostly.