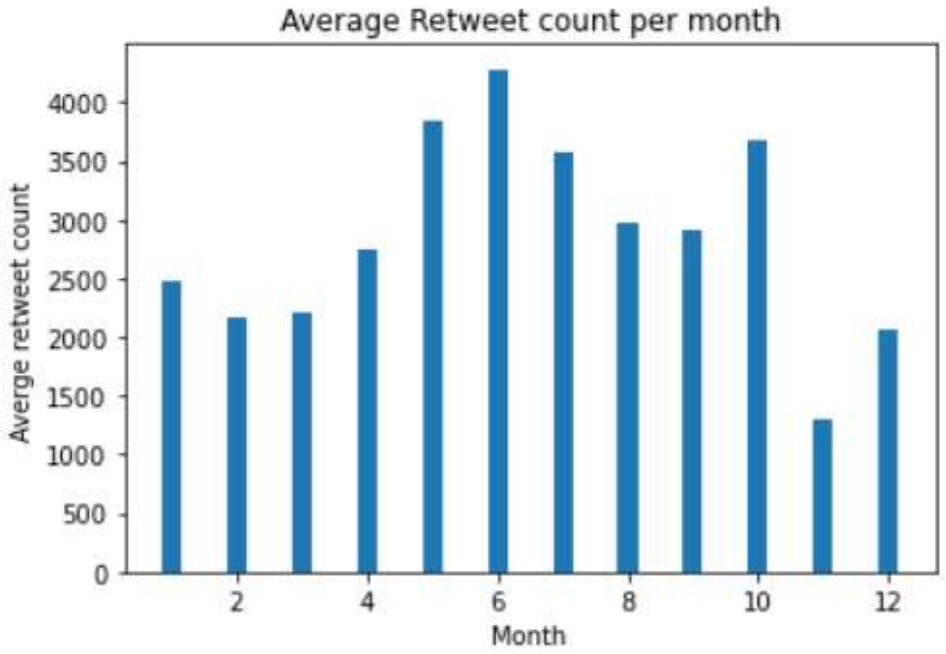
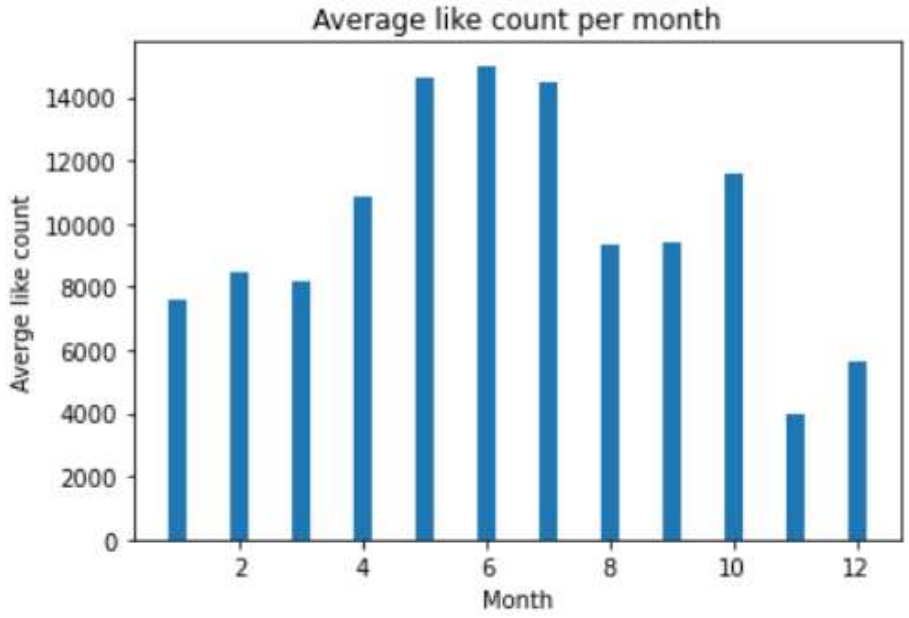
Analysis Report

After wrangling and cleaning the datasets to produced a clean dataset, analysis was performed on it to find some insights. The analysis was answer the following questions:

1. Which month has the most retweets on average?
2. Which month has the most likes(favorites) on average?
3. Is there a correlation between the average retweet, average like and year?
4. Is there a correlation between the rating and the average number of retweets and likes?

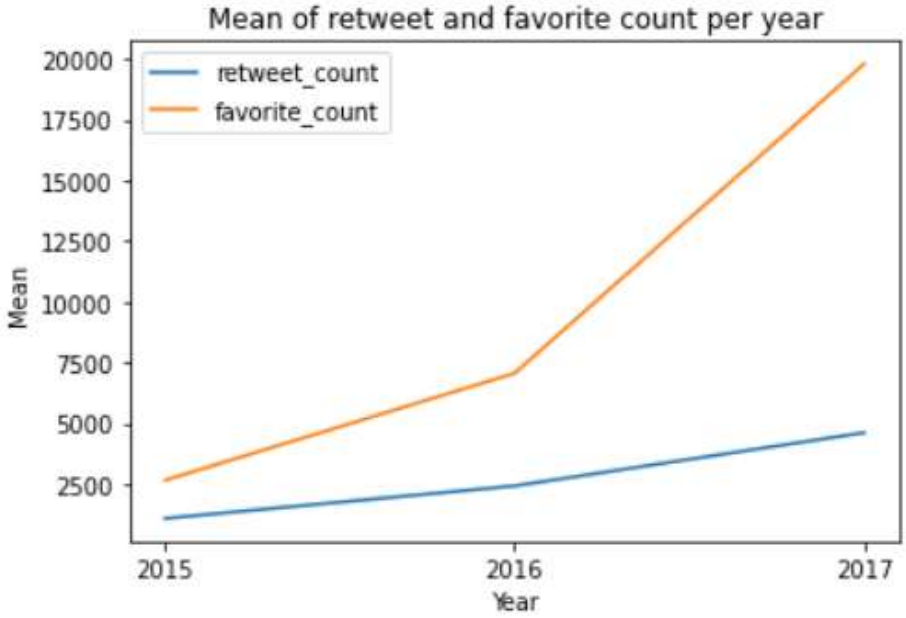
By using the describe method on the dataframe of the dataset, we are able to quickly see that on average the rating given to dogs in our dataset is approximately 11 with the minimum been 10 and the maximum been 17. Then proceeded to perform some analysis on the retweet_count and favorite_count on a monthly basis, below show its visualisation using a bar chart





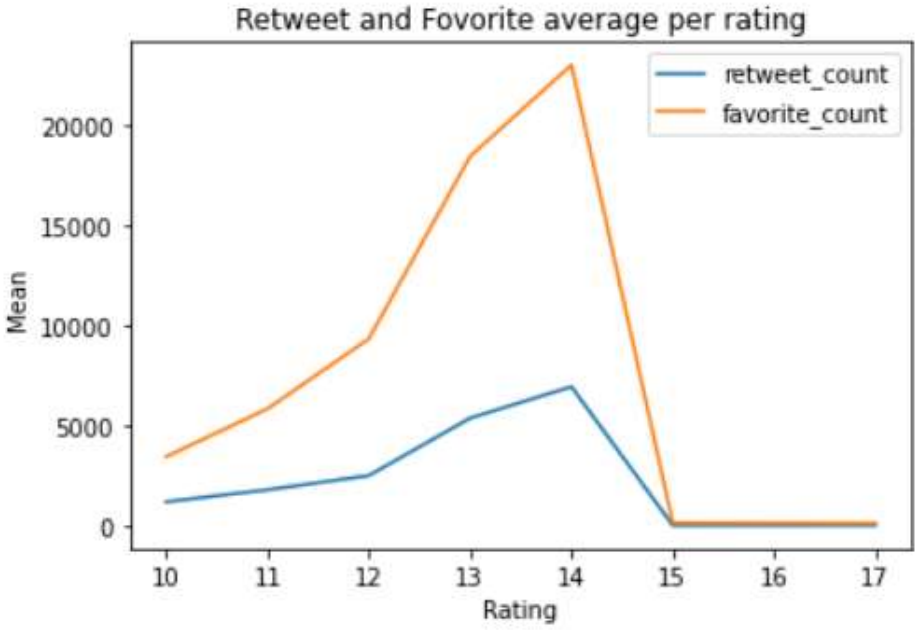
From this we can see that the month with the most retweets and likes is the 6th month which is June.

Then proceeded to do the same check but on a yearly basis to see if there is a correlation



From this visualisation, we can see that from our dataset between 2015 and 2017, there is an increase in the number of retweets and number of likes. This could signify an increase in public participation with the tweets from WeRateDogs and also a growing popularity as well.

To examine the if there is a correlation between the ratings and average retweet and like, the dataset was grouped by the ratings and then a line graph for plotted to examine this visually



From this initial check we can see that there is an increase in retweet count and favorite count with an increase in the rating. However, when the rating gets to 15 there is a massive drop. To examine this further, we take a look at the count for each of the rating and we get this

rating_numerator	retweet_count	favorite_count
10	440	440
11	425	425
12	498	498
13	303	303
14	43	43
15	1	1
17	1	1

From this table we can see that the reason for this massive drop is that the number of rows/samples with a rating of 15 and 17 is 1 respectively which would have a huge impact on the result plotted above. Hence, we can take the ratings 15 and 17 as outliers since we do not have enough samples with those ratings.