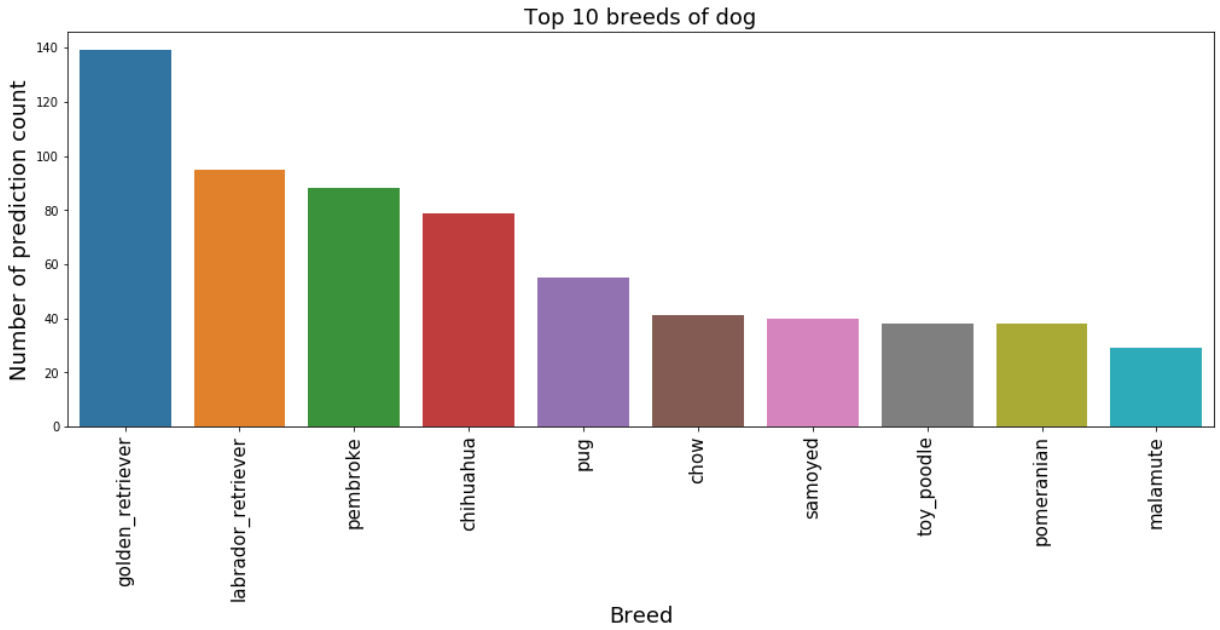
Report: act_report

After cleaning the data, we perform analysis with it to bring out at least 3 insights and 1 visualization. The following gives a detailed report on the insight derived from the data.

> Steps:
>
> 1. Read the master data
> 2. Perform analysis
> 3. visualize

1. Most Popular Dog Breed (Top 10)

```
golden_retriever      139
labrador_retriever     95
pembroke               88
chihuahua              79
pug                    54
chow                   41
samoyed                40
pomeranian             38
toy_poodle             38
malamute               29
Name: 1st_prediction, dtype: int64
```



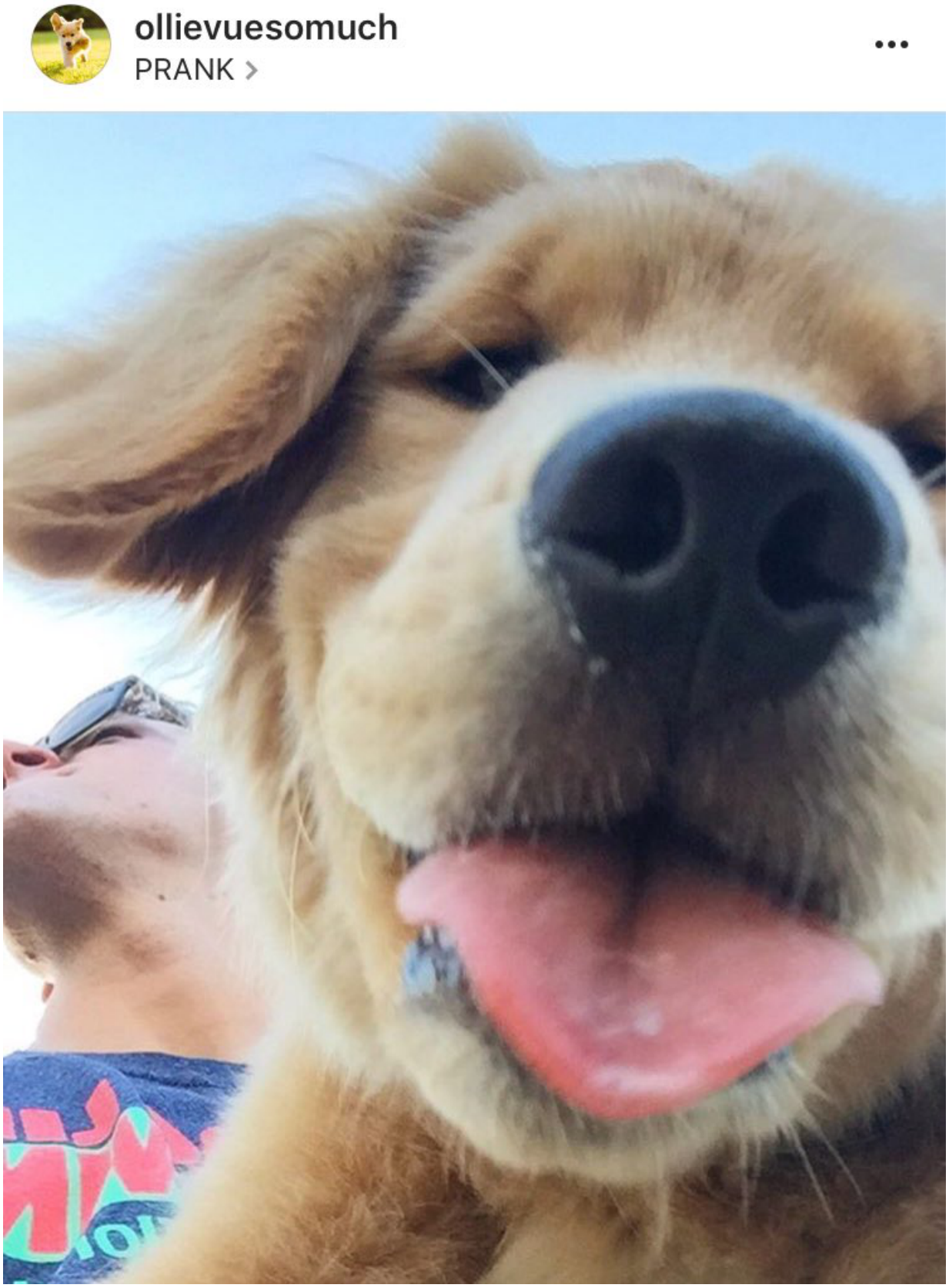
From the above figure, it is evident that **Golden Retriever** is the most popular dog breed amongst WeRateDogs's tweets in terms of the number of image_predictions having a total of 139 dogs. The second most popular dog breed is Labrador Retriever (95 dogs) followed by Pembroke (88 dogs)

and the 10th most popular is malamute with (29 dogs). Further analysis, shows that there are many dog breeds with 1 dog prediction.

Furthermore, we found out that the majority of the golden retriever has a rating of 12/10 (54 out of 139) while 29 of the golden retriever dogs have a rating of 13/10. From this, we can say that being popular doesn't necessarily mean that it would have the highest rating.

2. golden_retriever dog breed with the highest rating

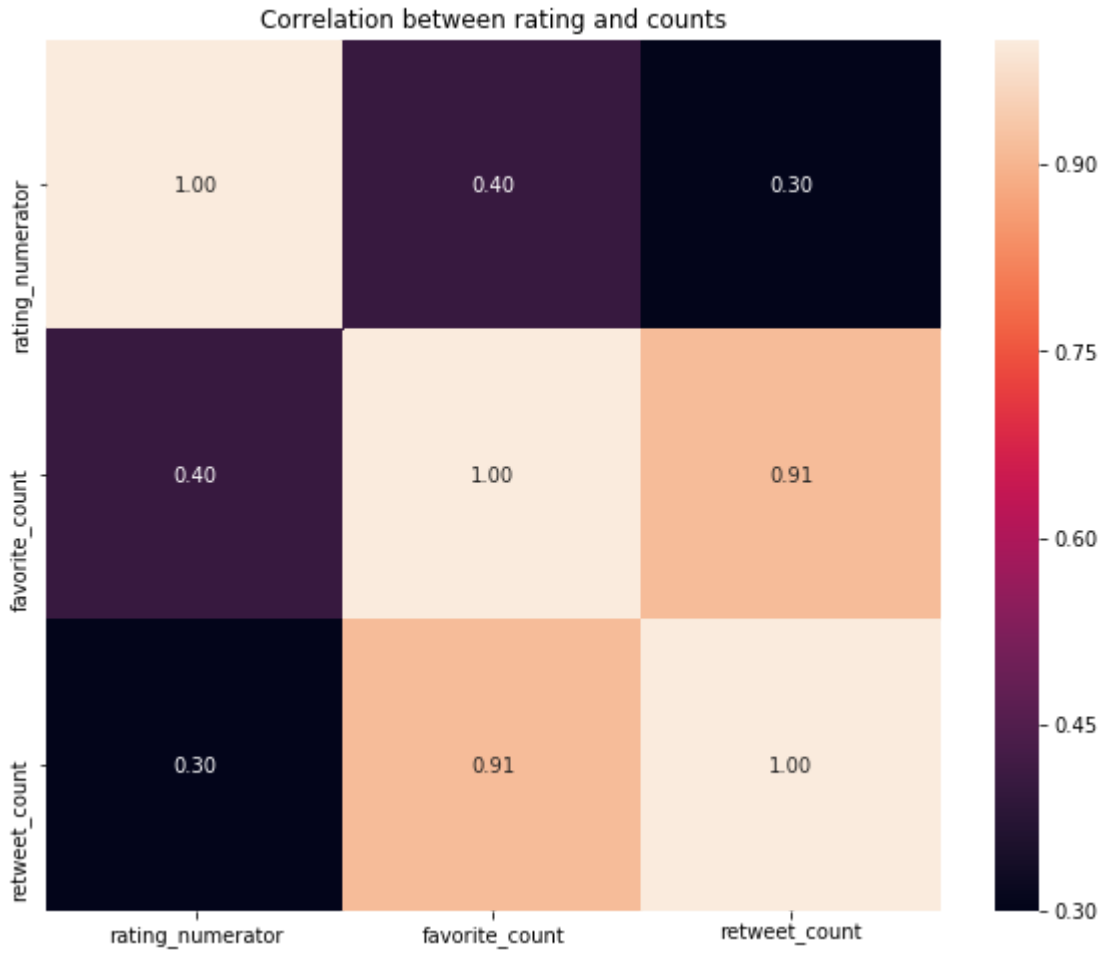
After cleaning, it was discovered that the highest golden_retriever dog rating is 14/10. From the analysis using a simple calling at the highest rating with `golden_retriever[golden_retriever['rating_numerator'] == 14]` , this golden_retriever dogs are not with the highest like and retweet count. This shows that having a high ratings doesn't guarantee high number of count and retweet

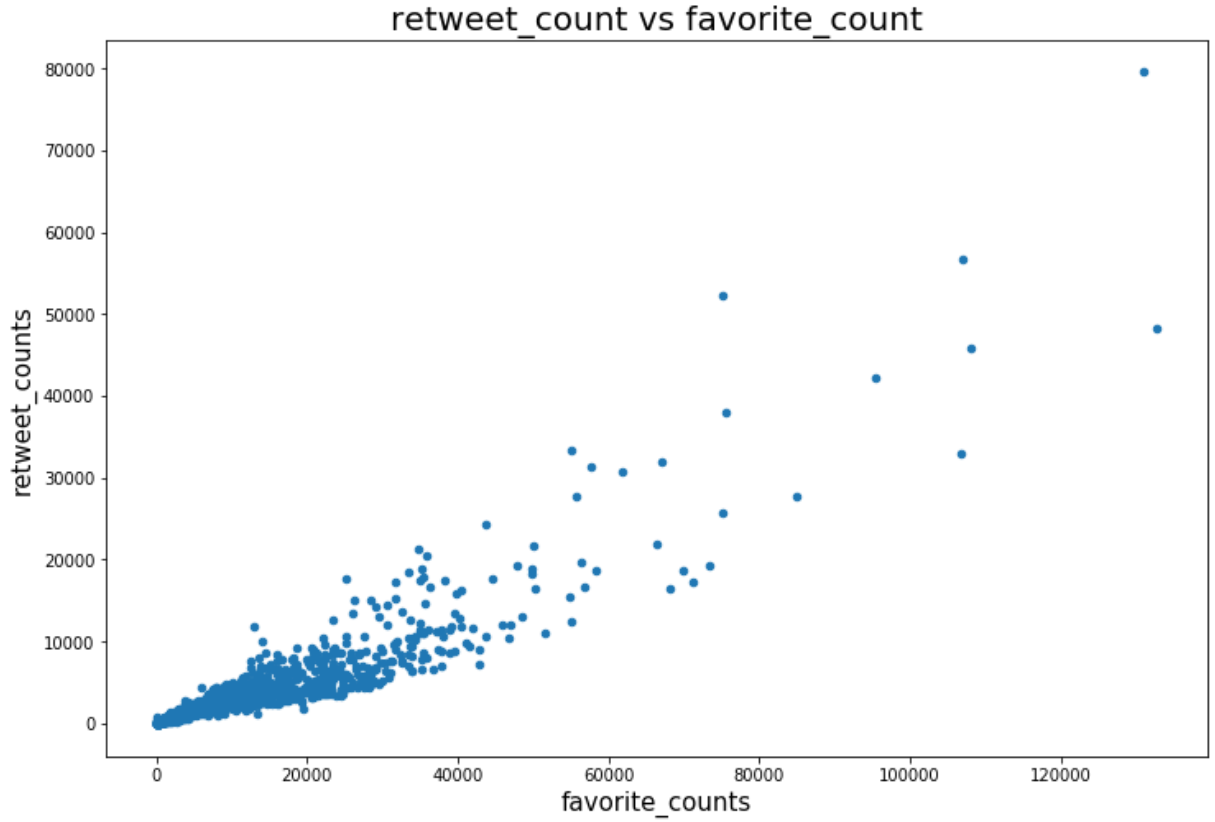


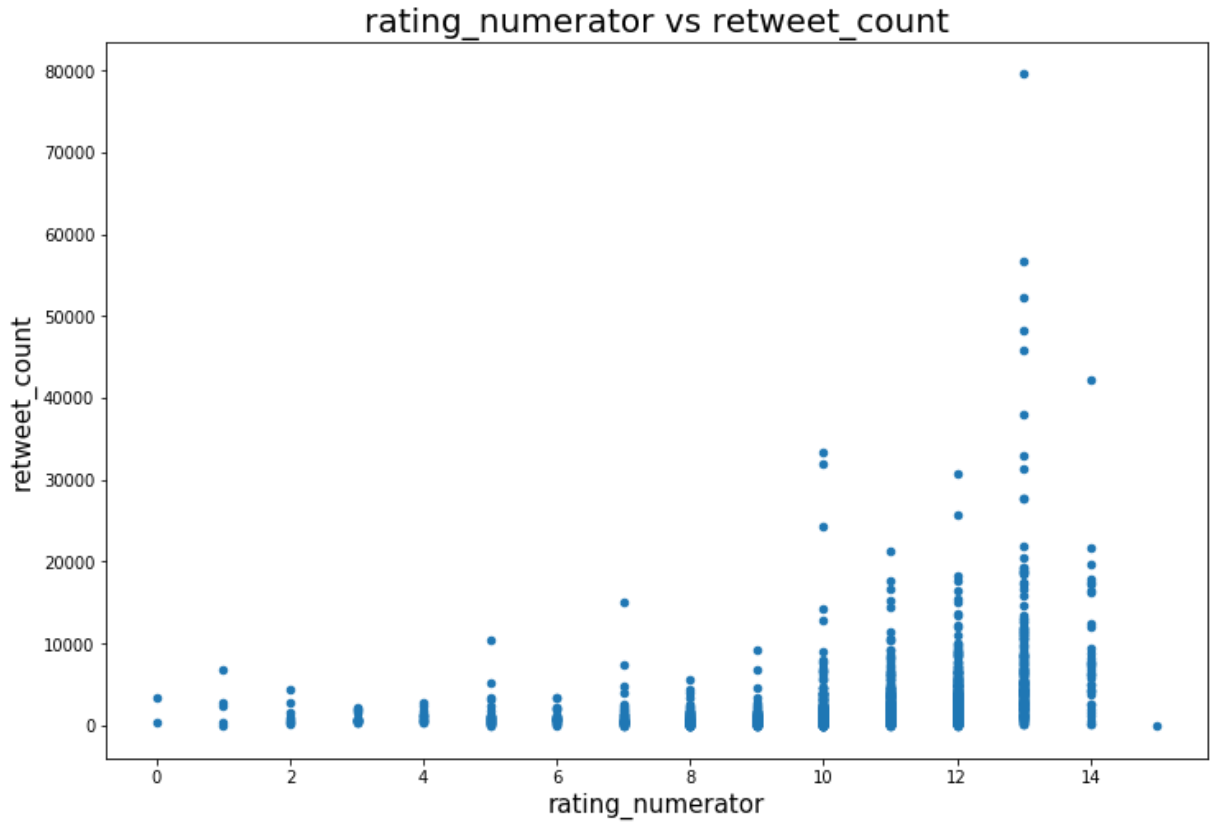
Correlation between ratings, retweets, and favorite (like)

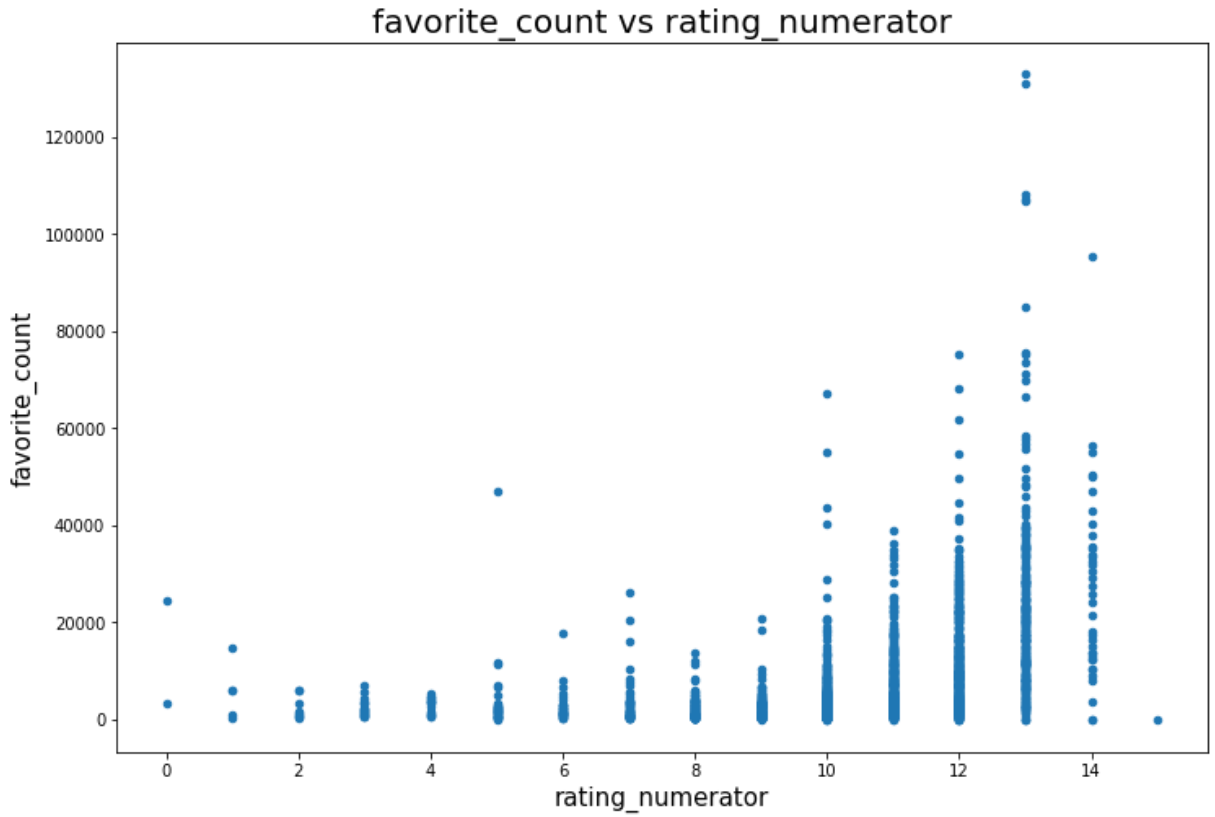
Using the pandas.corr() method to get the correlation between the dog ratings, retweet_count, and favorite_count.

The heatmap below show the correlation between the 3 variables. The scatter plots shows the relationship between two of the variables picked individually.




retweet_count vs favorite_count
retweet_counts
favorite_counts
0
10000
20000
30000
40000
50000
60000
70000
80000
0
20000
40000
60000
80000
100000
120000


rating_numerator vs retweet_count
retweet_count
rating_numerator
0
10000
20000
30000
40000
50000
60000
70000
80000
0
2
4
6
8
10
12
14



From the above analysis, we found out that there is a strong positive correction (0.91) between the favorite count and the retweet count, which means that the value of the favorite count would likely determine how high or low the retweet count is. However, there is a weak correlation between these counts and the dog ratings (0.4 and 0.3 respectively) i.e the outcome of counts is not necessarily determined by how high or low the rating is.