Analysis report

Once we combine all the cleaned dataframes into one, we have the following variables

```
tweet_id                1685 non-null int64
in_reply_to_status_id   20 non-null float64
in_reply_to_user_id     20 non-null float64
timestamp               1685 non-null datetime64[ns, UTC]
source                  1685 non-null object
text                    1685 non-null object
expanded_urls           1685 non-null object
rating_numerator        1685 non-null float64
rating_denominator      1685 non-null int64
name                    1205 non-null object
stage                   261 non-null object
jpg_url                 1685 non-null object
img_num                 1685 non-null int64
dog_species             1685 non-null object
retweet_count           1685 non-null int64
favorite_count          1685 non-null int64
```

Of these, the following are quantitative variables

1. retweet_count
2. favorite_count
3. rating_numerator (discreet)

Categorical variables

1. dog_species
2. name
3. stage

Timestamp is also a continuous numeric variable.

First, I looked at the ratings over time. A simple scatter plot as shown in figure 1 can help us look at this relationship. From the plot it is clear that *after 2016/11 most of the ratings were >10, and before that, there were non-negligible ratings less than 10.* However, most of the ratings are generally over 10.

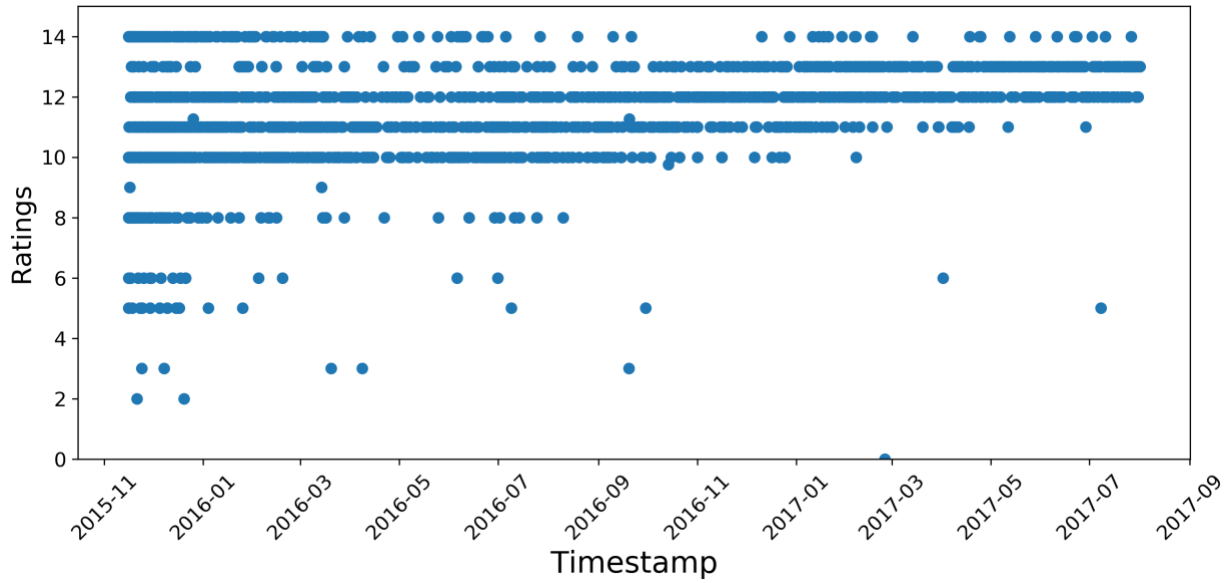

Figure 1. Ratings over Time

Next, I look at the relationship between retweet and favorite counts. There is a clear positive linear relationship. However, we see that generally the higher ratings have higher retweets and higher favorite counts.

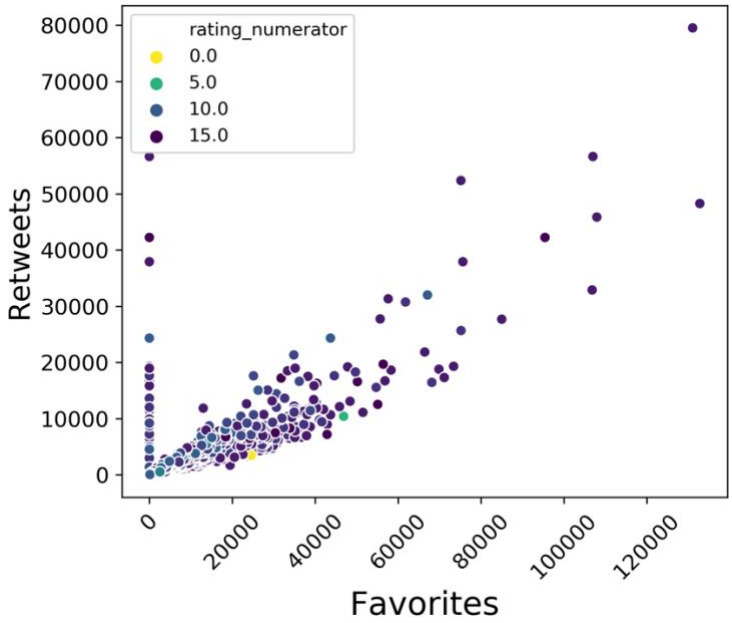

Figure 2. Retweets vs Favorite Counts

After that I analyzed the counts of the ratings and dog stages. As we can see in the ratings plot, most ratings are over 10, and the higher number around 10-12. There are very few ratings less than 10. Analyzing the stage counts, we see the most common stage is pupper followed by doggo, while floofer and blep are the least common stages.

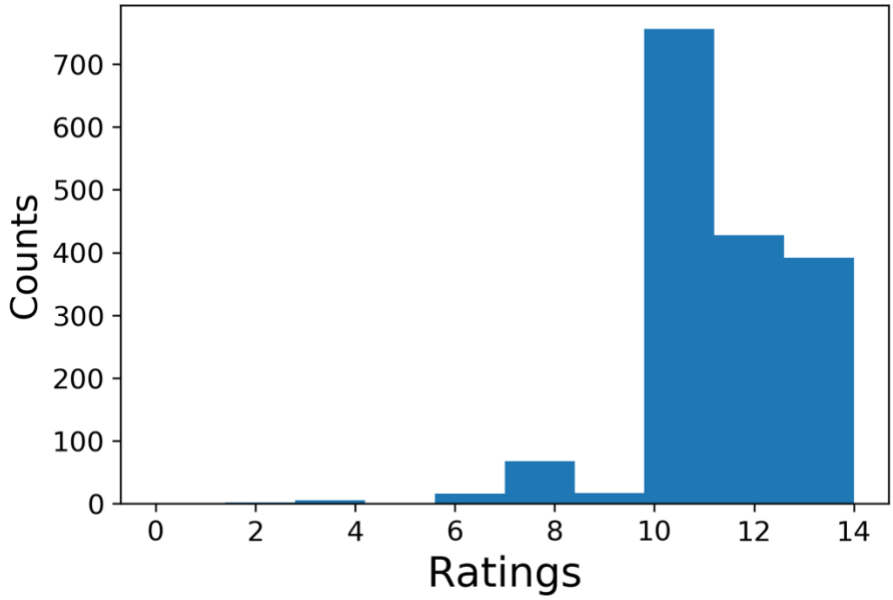

Figure 3. Ratings Count

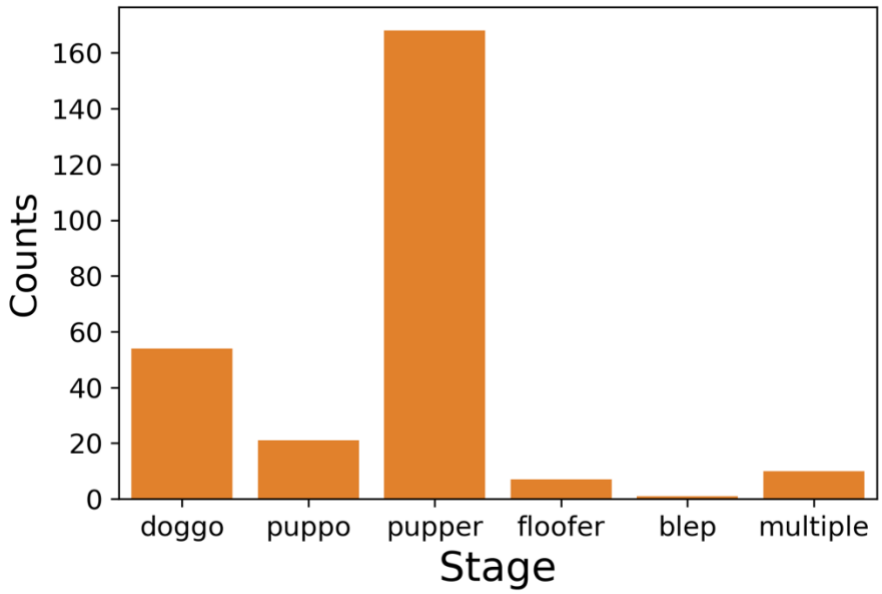

Figure 4. Stage Count

Lastly, to analyze which species and names are most popular, I used word clouds, because there are several names and species, and using count plots would get too cluttered.

Figure 5. Species Word Cloud

Figure 6. Name Word Count