ANALYSIS AND VISUALIZATION OF WERATEDOGS TWITTER DATA

INTRODUCTION

We will be analyzing and Visualizing WeRateDogs Twitter data. Data Analysis and Visualization are the final phases where the cleaned data is mined for information and insights.

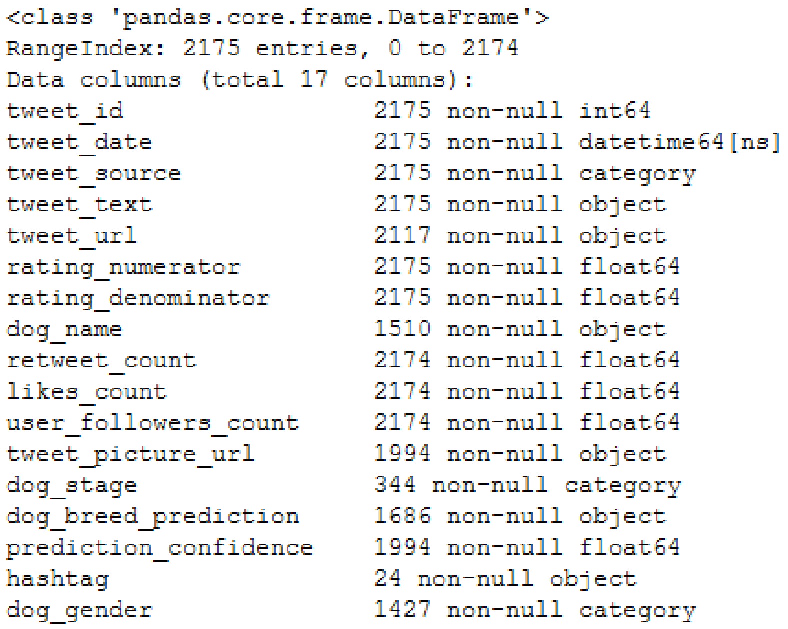
ABOUT DATASET

We will analyze the curated data of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 7 million followers and has received international media coverage. The data was wrangled by gathering from different sources, they were assessed and cleaned and stored, ready for analysis.

DATA LOAD

IMPORT DATA

We have stored the wrangled data in twitter_archive_master.csv file. Once loaded, the structure of the dataset is shown below. The dataset has *2175 rows* and *17 columns*.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2175 entries, 0 to 2174
Data columns (total 17 columns):
tweet_id                 2175 non-null int64
tweet_date               2175 non-null datetime64[ns]
tweet_source             2175 non-null category
tweet_text               2175 non-null object
tweet_url                2117 non-null object
rating_numerator         2175 non-null float64
rating_denominator       2175 non-null float64
dog_name                 1510 non-null object
retweet_count            2174 non-null float64
likes_count              2174 non-null float64
user_followers_count     2174 non-null float64
tweet_picture_url        1994 non-null object
dog_stage                344 non-null category
dog_breed_prediction     1686 non-null object
prediction_confidence    1994 non-null float64
hashtag                  24 non-null object
dog_gender               1427 non-null category
```

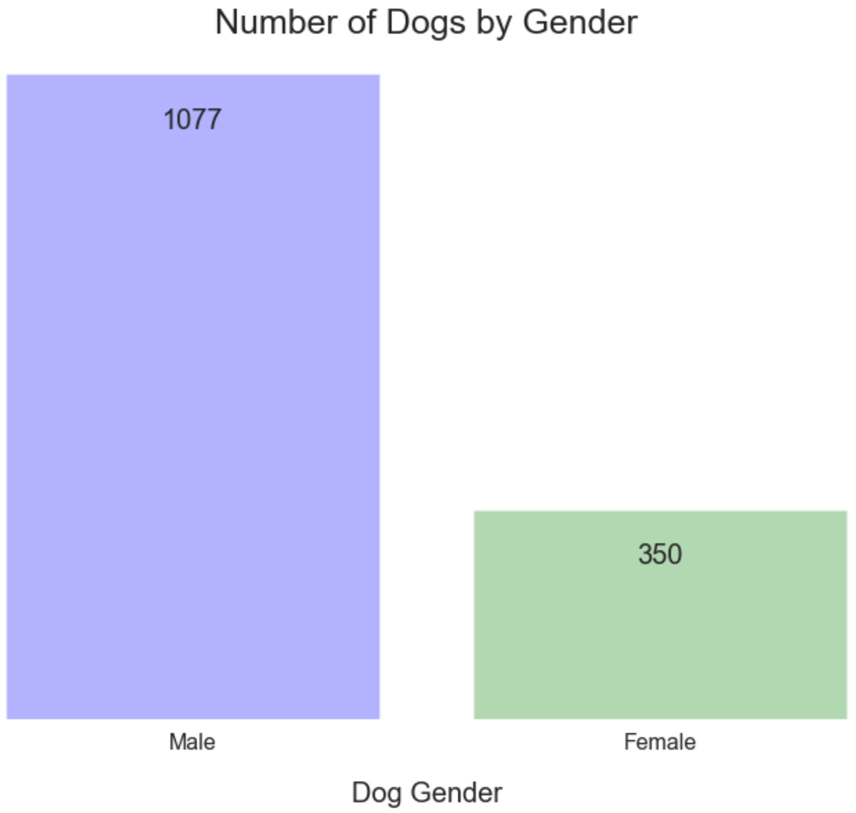
UNI-VARIATE ANALYSIS

We will start with single variable analysis

GENDER

The first question that comes to our minds, how many dogs are Male? and How many dogs are Female?

We can see that there are 1077 Male dogs and 350 Female dogs in our dataset.



Observation:
1. There are more 'Male' dogs than 'Female' dogs in our dataset.

DOG STAGE

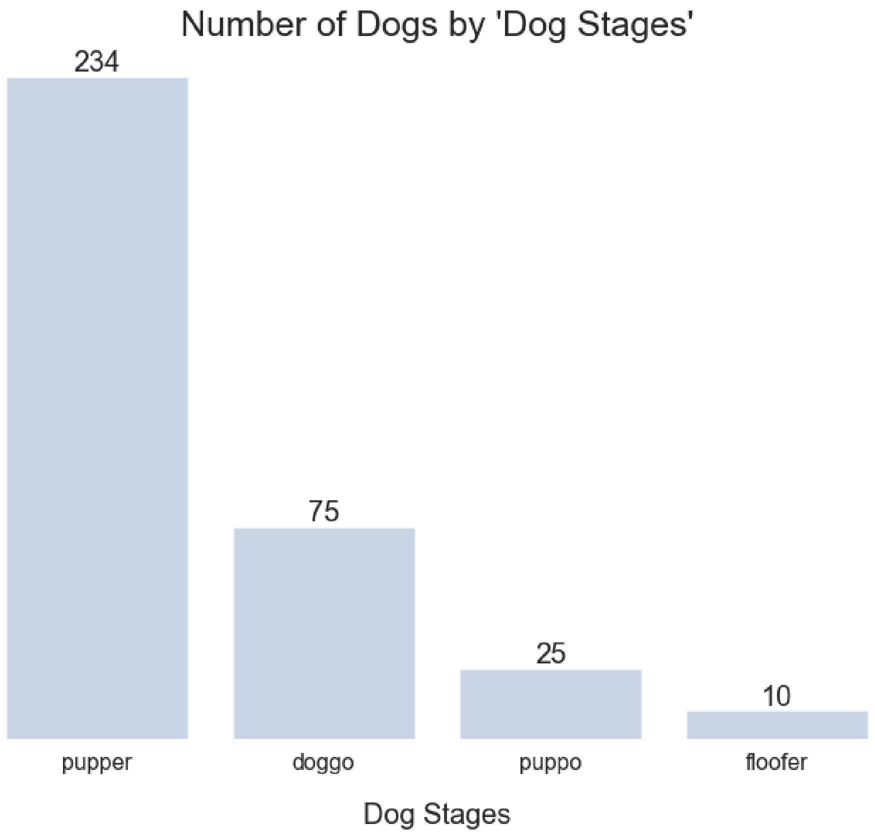
'Dog Stages' in our dataset represent various stages in dog's growth/life. In our dataset, the stages are specified in Doggo Lingo.

DoggoLingo tends to operate in terms of free indirect discourse; the speaker is a human admiring (or mirin') a dog, but the vocabulary is implied to be the dog's own idiom. In keeping with old cultural beliefs about what goes on in the canine brain, the vocabulary of DoggoLingo is upbeat, joyful, and clueless in a relentlessly friendly way. - Oxford Dictionary

In various stages of growth, Dogs are called by different names. Some that are found in our dataset are

- **pupper** - younger dogs
- **puppo** - adolescent dogs
- **doggo** - mature dogs
- **floofer** - dogs with excess fur or old dogs

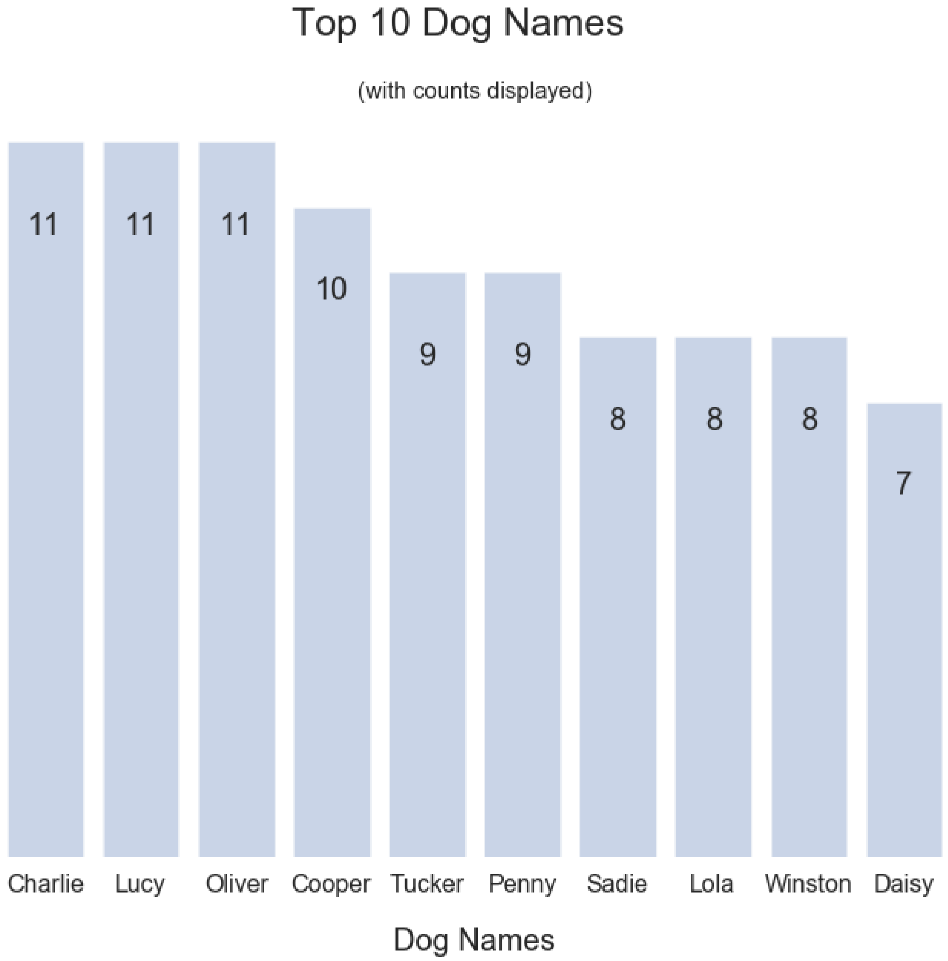
If we plot the number of dogs by dog stages, this is what we get



There are 'puppers' with a count of 234, more than the combined count of the other three categories

DOG NAME

Dog names are the names given to dogs. Its how they're called. We love naming them, playing with them, scolding them. What's the most common dog name? Does your dog have a name that is in this list?



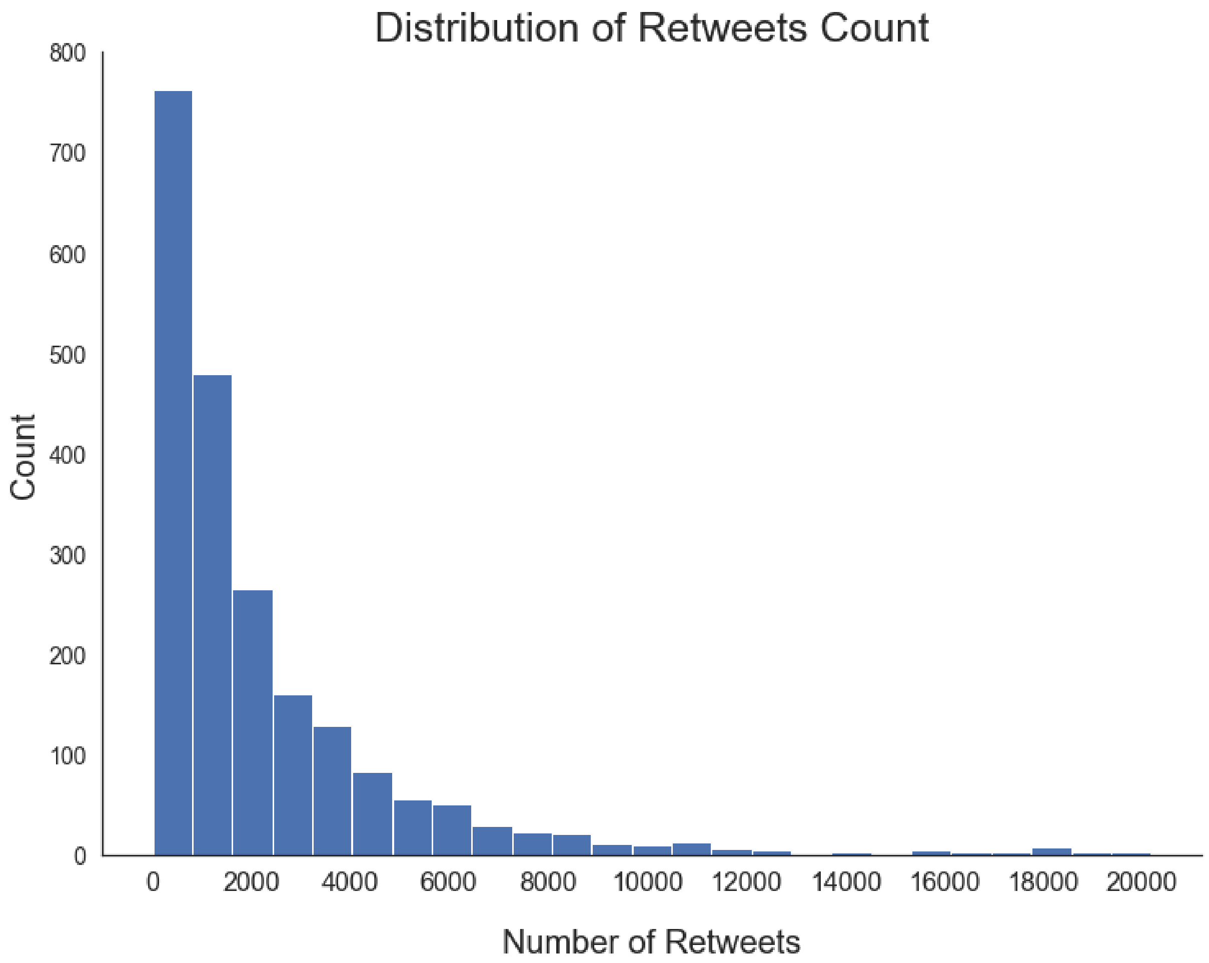
The most common dog names are **Charlie, Lucy & Oliver**. There are 11 dogs with that name in each case.

RETWEETS

What is it?

A Retweet is a re-posting of a Tweet. Twitter's Retweet feature helps you and others quickly share that Tweet with all of your followers. You can Retweet your own Tweets or Tweets from someone else. -Twitter FAQ

The Retweet count shows how many times the particular Tweet was retweeted by others.

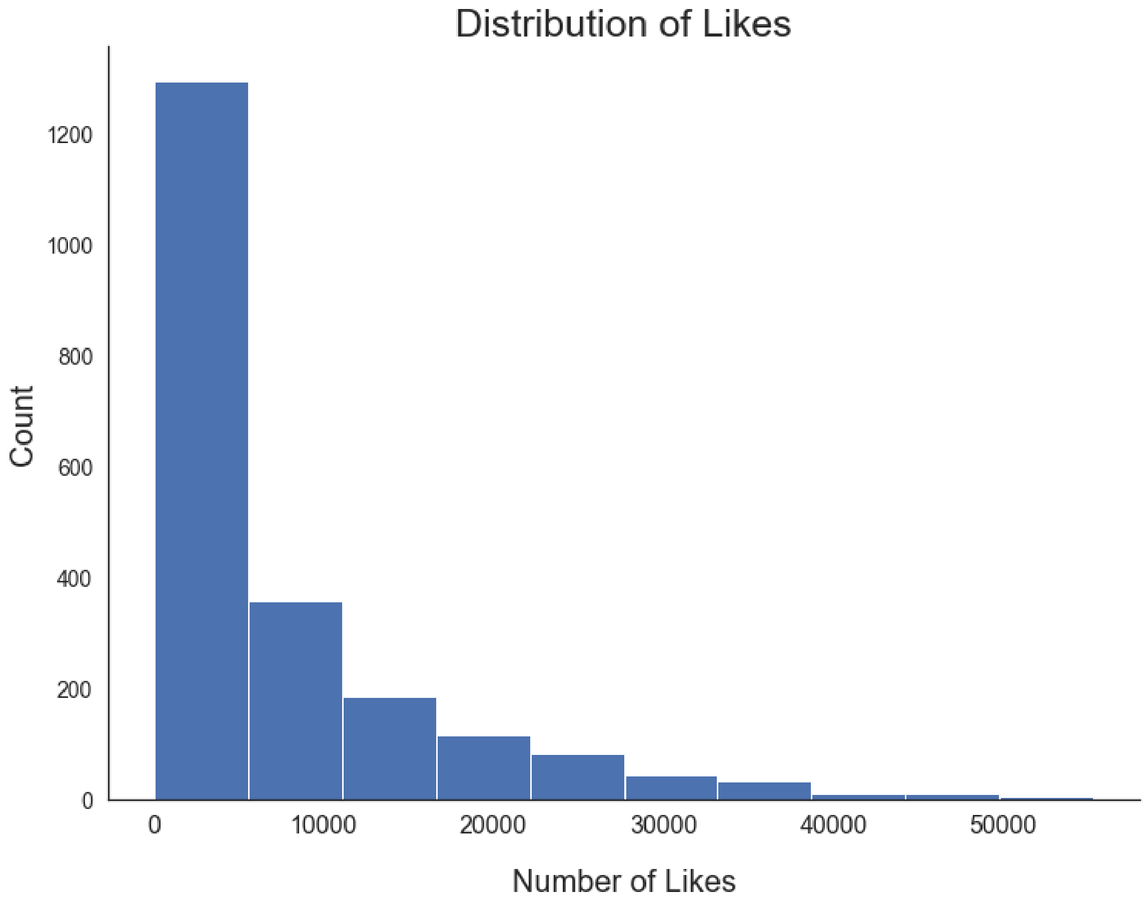


If we check the distribution of Retweets, the distribution is skewed. Most Tweets are retweeted less than 2000 times(but that is still a huge number!). But there are cases where there are more than 20000 Retweets!

LIKES

Likes are represented by a small heart and are used to show appreciation for a Tweet or a Moment. -Twitter FAQ

Likes count in our dataset shows how many times the particular Tweet was liked by others

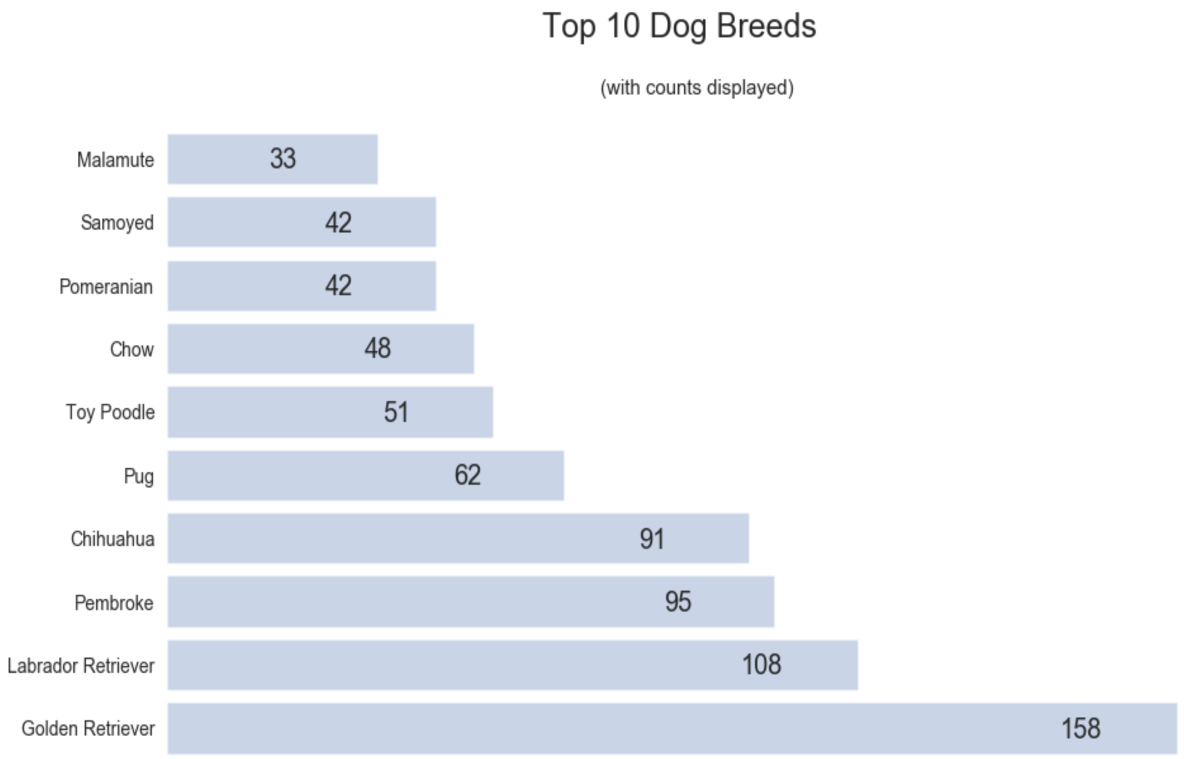


The distribution is similar to Retweets. It is skewed right with more values in the lower end of the spectrum.

DOG BREED

Every Tweet image in the Twitter archive for WERATEDOGS was passed through a neural network that can classify breeds of dogs and its outcome is what we have in the consolidated breed prediction and prediction confidence columns. So the question is how many dogs are there in each breed? The number of breeds in our dataset is huge. So we will consider only the top 10 breeds.

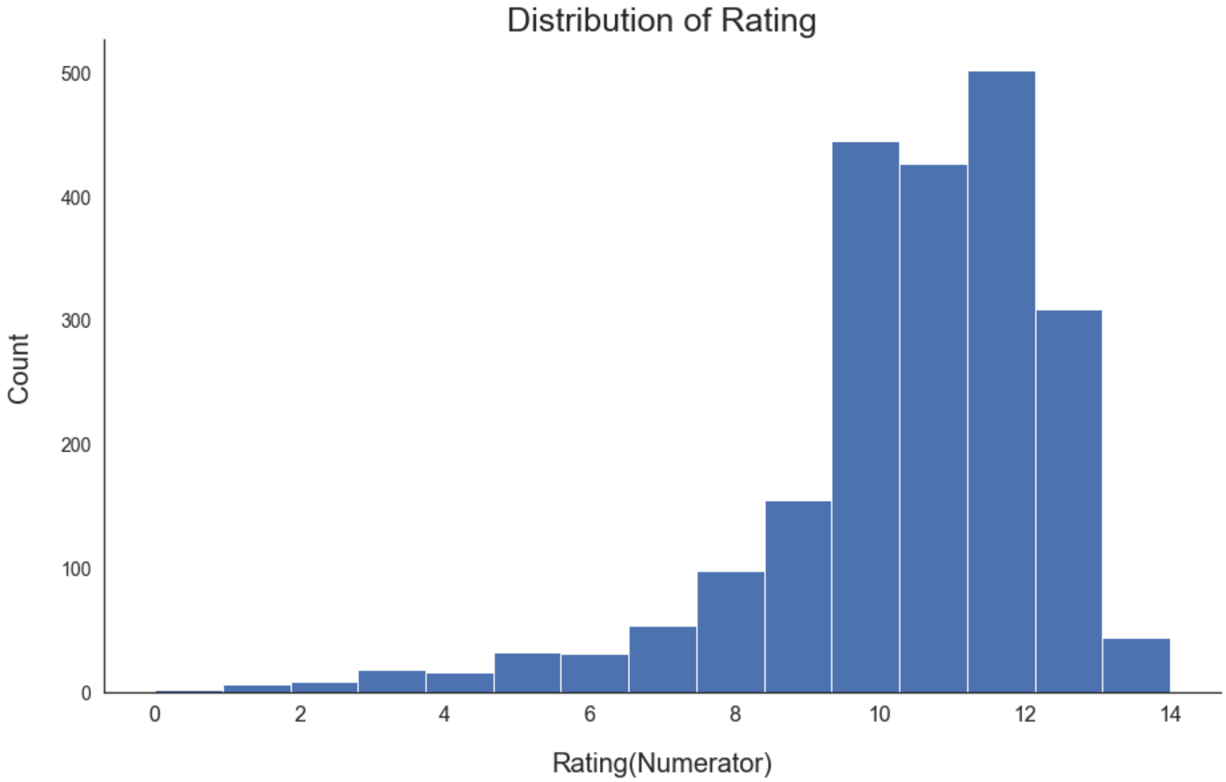
First place goes to 'Golden Retriever' dogs with a count of 158. Then we have Labrador, Pembroke in 2nd and 3rd places. The top 10 breeds are plotted below.



RATING

It is the rating given to doggo in the Tweet picture!

Ratings are given in the format *numerator/denominator* with denominator mostly 10 and numerator almost always greater than 10. Why? **They're all good Dogs Brent!**



As we can see from the above plot, most dogs will get a rating above 10. The most common value seems to be - *12*

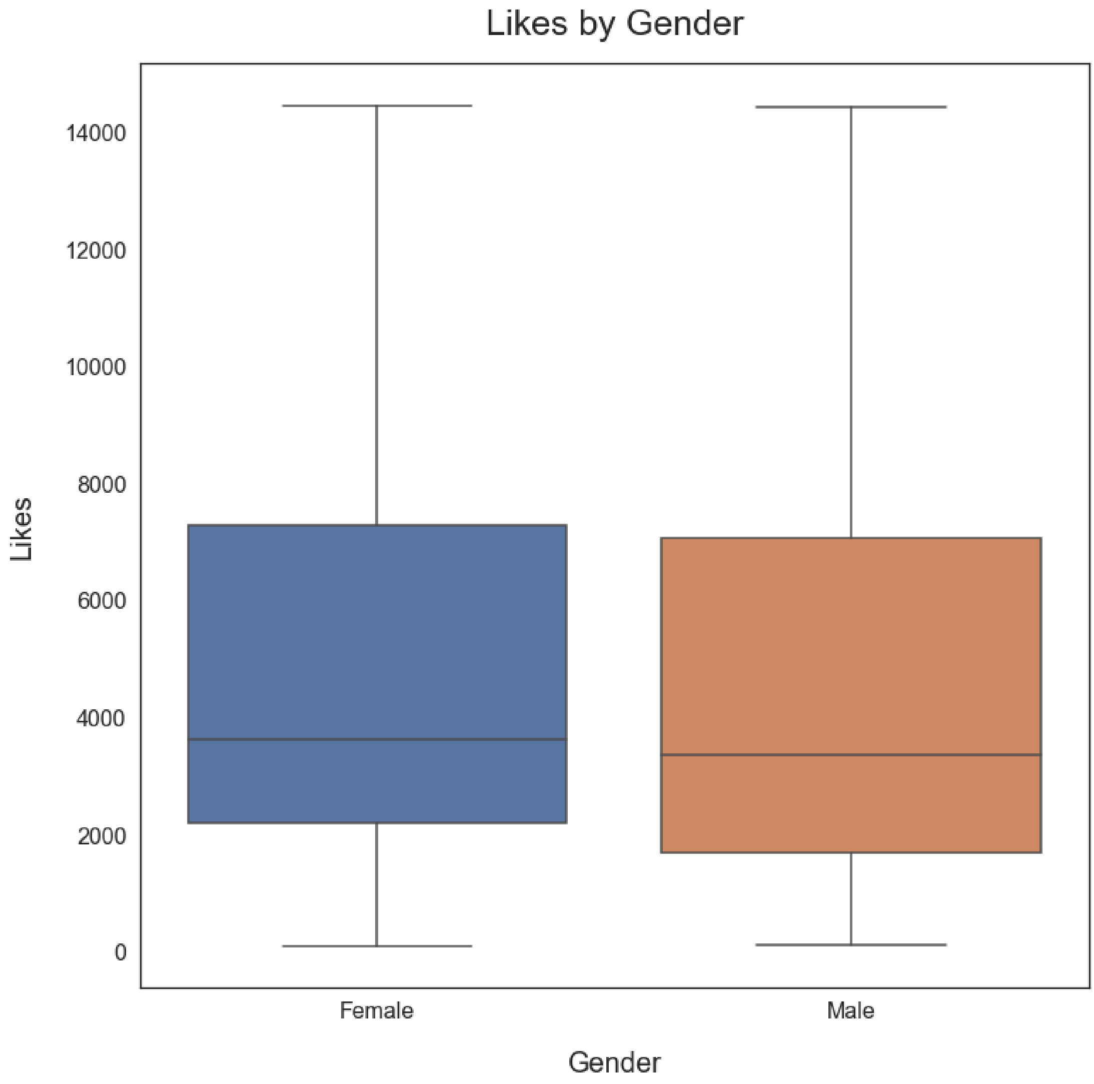
BI-VARIATE ANALYSIS

We have seen single variables and their properties. Let’s check how they relate to each other

LIKES BY GENDER

Does gender influence the number of likes a dog gets? Does a female dog get more likes than male dogs?

When we restrict the data removing outliers, we get a plot like the one below

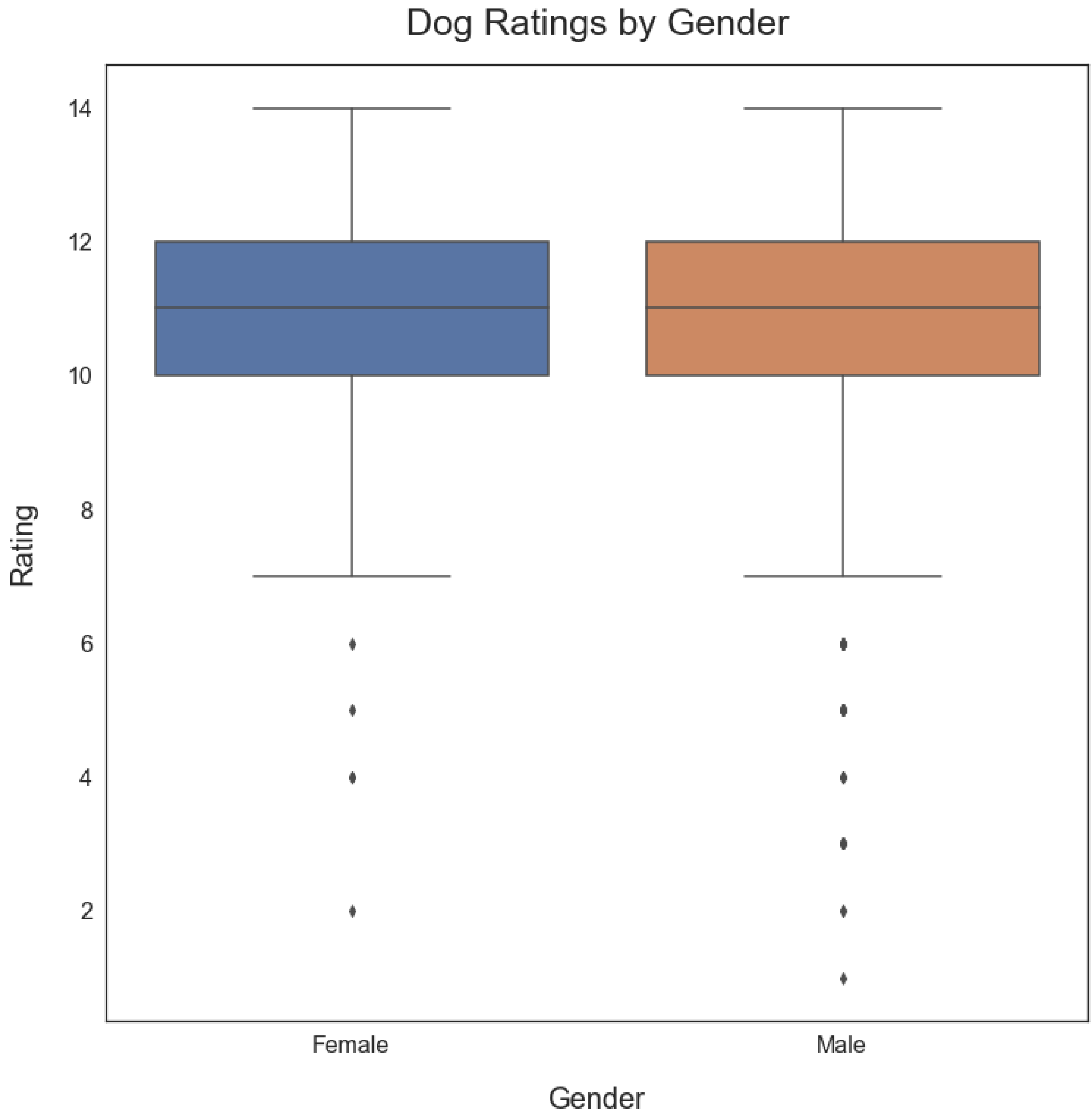


Although the difference is not much, 'Female' dogs get slightly more likes than male dogs on average!

RATING BY GENDER

We saw that female dogs have a bit more advantage than male dogs in likes. Does this mean, they also get more rating than male dogs ? Does rating vary by gender?

We'll remove outliers and concentrate on IQR which is what we are interested in.

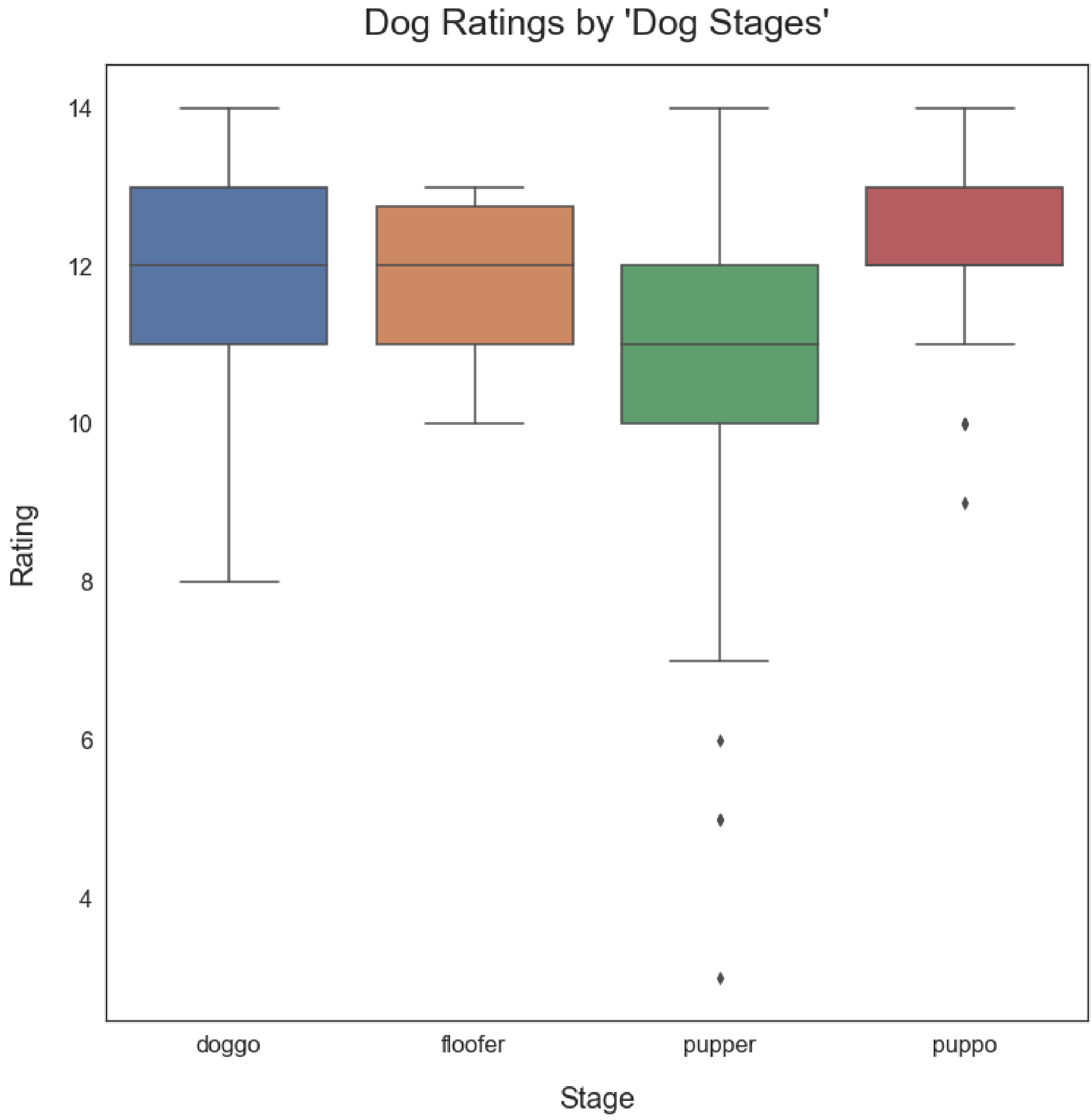


Hmm...the boxplot reveals that rating doesn't change by gender. Both genders get an equal rating on average. Thats good to hear!!

RATING BY 'DOG STAGE'

Does 'Stage' influence the ratings given? We all love a cute *puppo*! who doesn't! but does that mean good ol'floofers dont get equal rating?

The plot reveals the variation in rating for each dog stage. On average, *puppers* have less rating than doggos and floofers. *puppos* get a better rating than other 3 categories as seen below

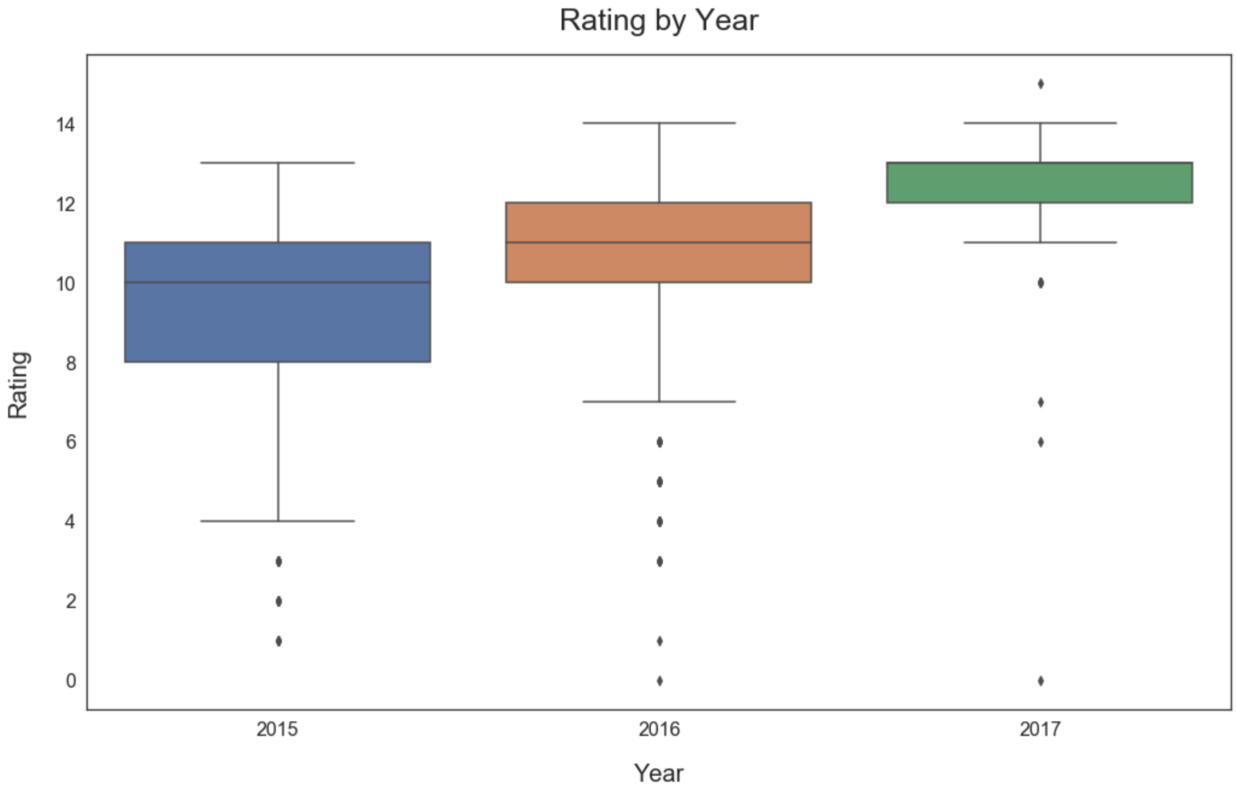


RATING BY TWEET DATE

We identified that ratings vary by dog stage. Are there any other factors that influence ratings? Do dogs get a better rating in certain months of the year? how does rating vary by year?

We have Tweets from 2015 - 2017 in our dataset. We will first check rating values by 'Year' and then 'Month'

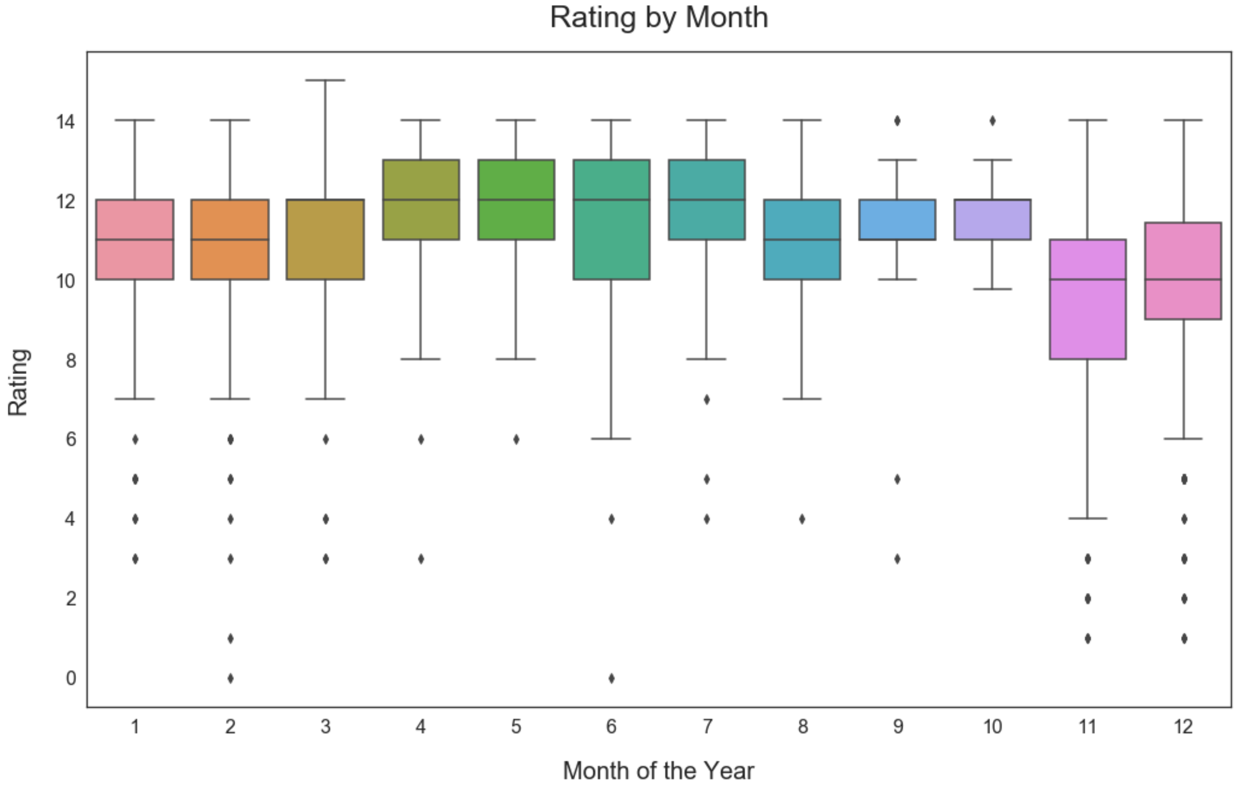
BY YEAR



Surprisingly, rating values keep increasing as year progresses. Dogs are getting better every year! The median rating value in 2016 is nearly equal to the 3rd quartile value in 2015.

However, the first quartile value in 2017 is nearly equal to the 3rd quartile. We can see the pattern repeating.

BY MONTH

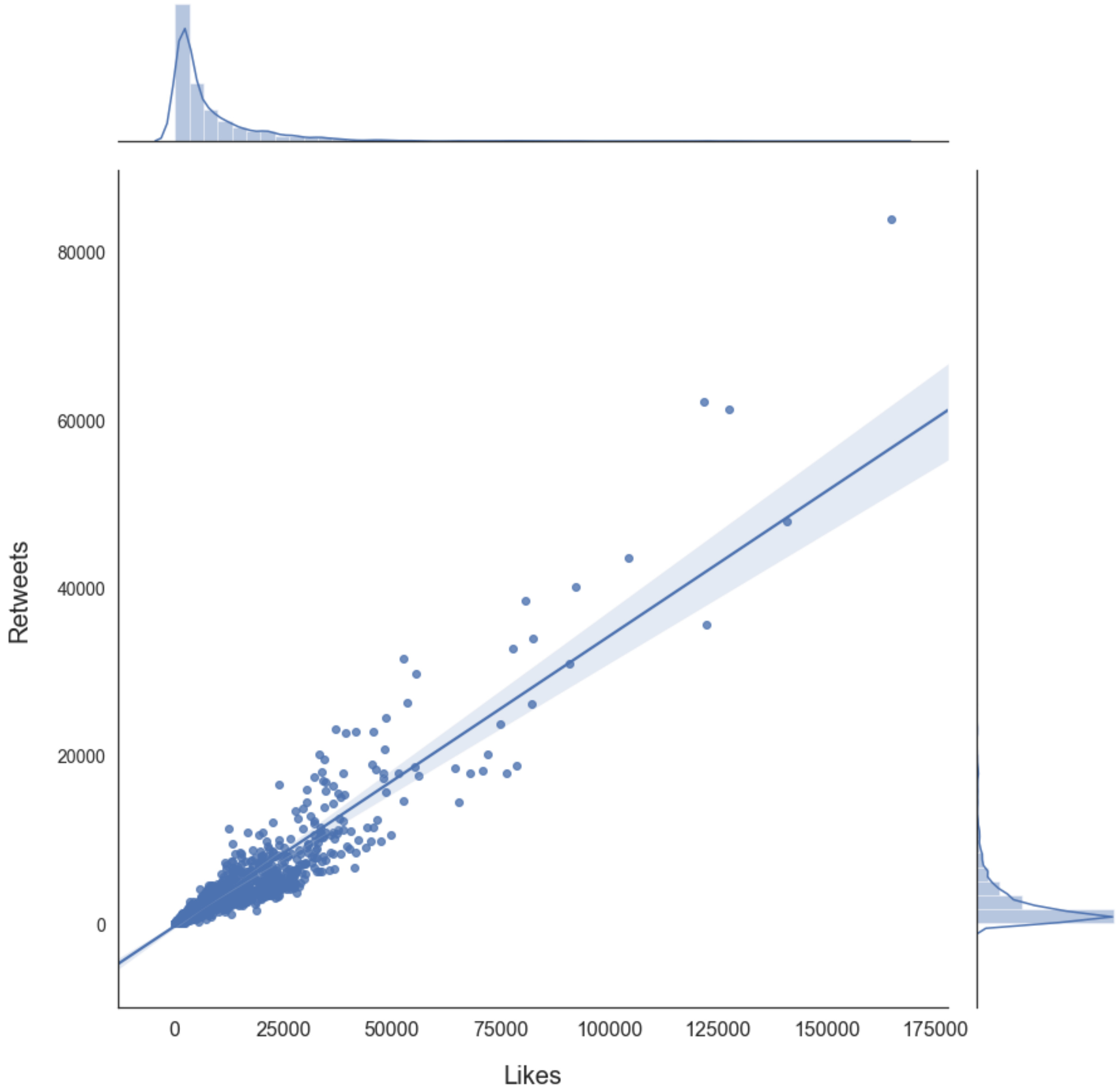


The months April, May, June and July seem to have higher ratings on average than other months as seen from the box plots above.

LIKES AND RETWEETS

'Likes' and 'Retweets' have many similarities. There should be **high correlation** between the 2 columns, because if a person likes a tweet, its more probable that they will share the tweet. But can we find evidence in our dataset for this?

The answer is 'Yes!' Look at the plot for Likes and Retweets.



There seems to be a high positive correlation between Likes and Retweets as expected. When the correlation coefficient is calculated, we have the below output

Pearson's r = 0.93 and p=0.0

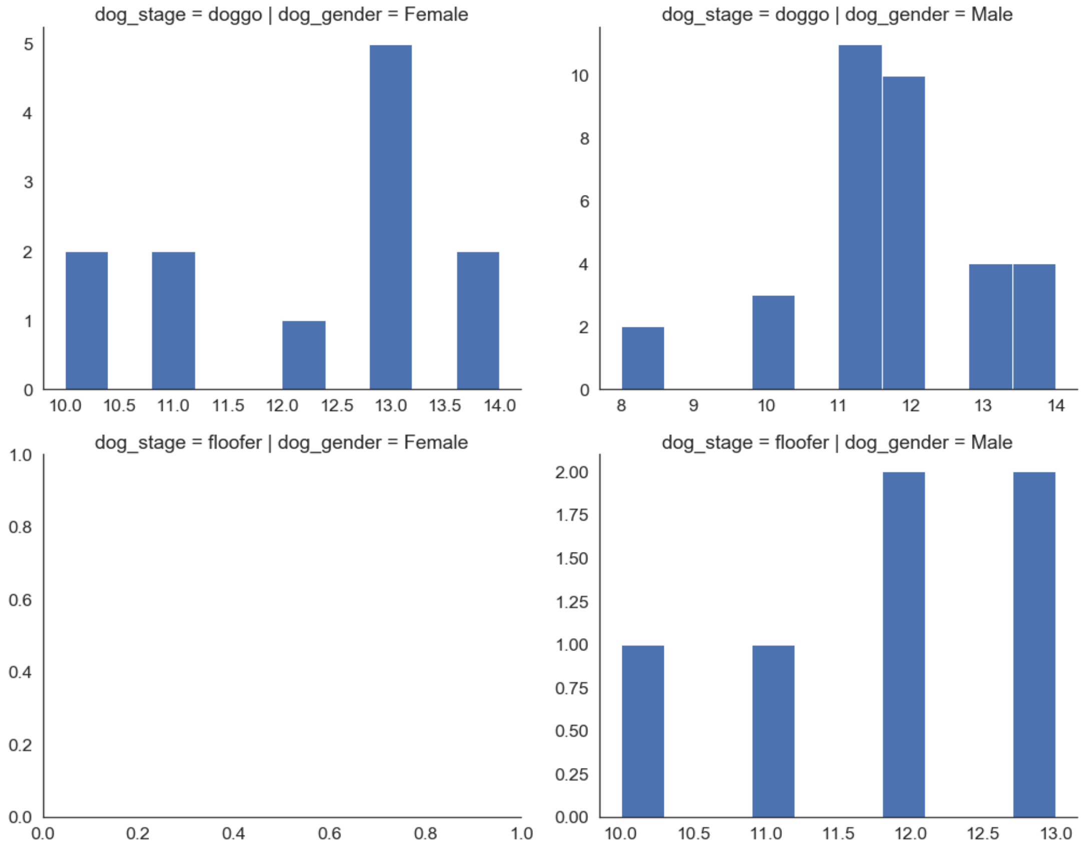
This is statistically significant

MULTI-VARIATE ANALYSIS

We have seen single variable characteristics, we've also seen how 2 variables relate to each other. Now we'll pull in more variables and check out their relationships

RATING BY GENDER AND DOG STAGE

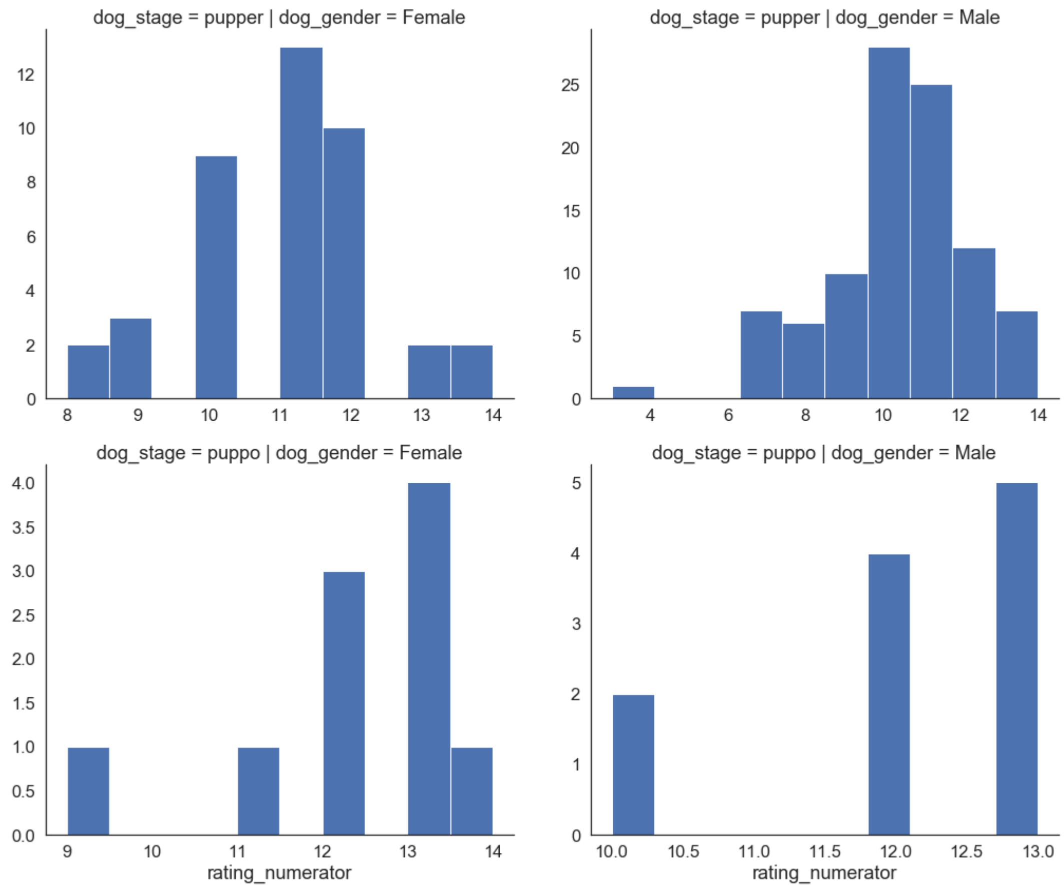
We have seen that ratings vary by dog stage. In that case, we will also check how dogs from each gender get rated for each dog stage. The below plot show the distribution of rating for each gender in each dog stage.



The above plot is distribution of rating for doggo and floofer. For puppo and pupper, the plot is available below.

The plot on a whole, convey a lot of information. A few we could identify by glancing the plots

- We can see *puppers* got low ratings. they've been bad bois!
- There are no plots for 'Female' floofers
- 'Female' doggos and puppos have better ratings than their counterparts



CONCLUSION

Planning to get a dog? get 'Female' puppos. They are not very young but they are not matured too. You can watch them grow, spend more time with them, play with them and then take pictures and send it to @dog_rates. You'll have a better chance of getting a better rating in that case.

But hey!! Who cares about rating! Get any dog you want, cause they're all good dogs friend!