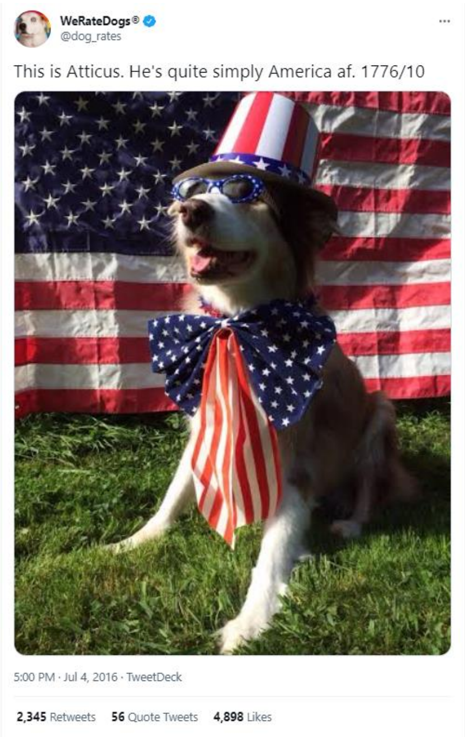
WeRateDogs: Twitter Data Analysis

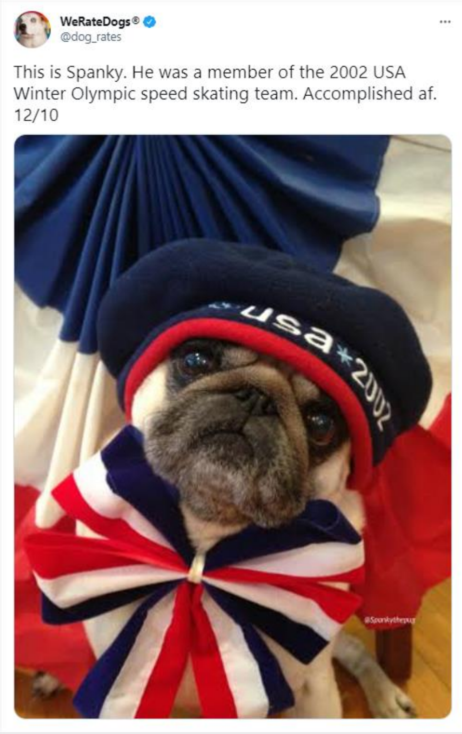


WeRateDogs® @dog_rates

This is Atticus. He's quite simply America af. 1776/10

5:00 PM · Jul 4, 2016 · TweetDeck

2,345 Retweets 56 Quote Tweets 4,898 Likes

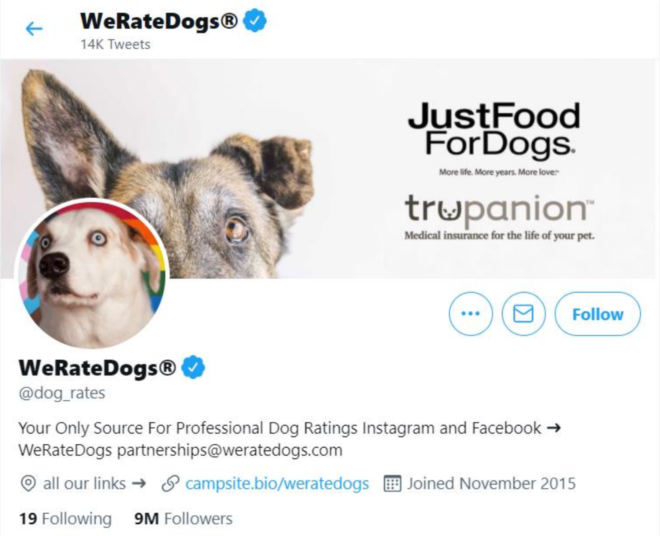


WeRateDogs® @dog_rates

This is Spanky. He was a member of the 2002 USA Winter Olympic speed skating team. Accomplished af. 12/10

Let’s join me, Atticus and Spanky to celebrate the happy 4th July, shall you? Don’t you think they are both really excited for the day and so well-prepared by looking at Atticus’s big smile and his super stylish American flag outfit, and the USA 2002 hat that Spanky is wearing? I personally think that they both deserve extremely high ratings for their cuteness as I adore them both so much. Are you now curious about who they are and what the ratings they got in each picture are?

If yes, you should take a look at a Twitter account WeRateDogs @dog_rates where you can find thousands of tweets rating pictures of cute dogs with a humorous comment about the dog. WeRateDogs has over 9 million followers who read this page, rate dogs, retweet tweets and add to favorites. An interesting thing about the dog rating system is that the score can be granted almost always greater than 10 out of 10. 11/10, 12/10, 13/10, or even 1776/10 just like what Atticus got. Why? Because "they're good dogs Brent."



WeRateDogs®
14K Tweets

JustFood ForDogs.
More life. More years. More love.
trupanion
Medical insurance for the life of your pet.
Follow

WeRateDogs®
@dog_rates

Your Only Source For Professional Dog Ratings Instagram and Facebook → WeRateDogs partnerships@weratedogs.com

all our links → campsite.bio/weratedogs Joined November 2015

19 Following 9M Followers

Because of such popularity of this page, we decided to dive into some number and extract statistics from the tweets. I gathered data from different sources including given dataset by Udacity, using python commands to download data from a specific link and scarping data from Twitter's API. After that, I assessed my data against quality and tidiness issues and then fixed all the identified issues to produce one clean and tidy master dataset. This twitter archive master data contains basic tweet data (tweet ID, timestamp, text, etc.), rating, dog name, and dog "stage", retweet count as well as favorite counts. It also consists of breeds of dogs based on image predictions. I had a lot of fun with doing analysis on this clean data and I would like to share with you my interesting findings through the following visualizations and insights.

My approach to analysis was to ask a question and then try to answer the question with the data that I had. Before going to six research questions in details, I will briefly walk you through some basic descriptive statistics of the data first.

1. **Basic descriptive statistics:**

I would like to point out one important fact that we assume that the maximum value of the rating numerator is 15 and everything above are outliers. The outliers will be ignored while producing Figure 1. Similar corrections were made for rating denominator, favorite count and retweet count. Without further delay, let's look at plots and numbers

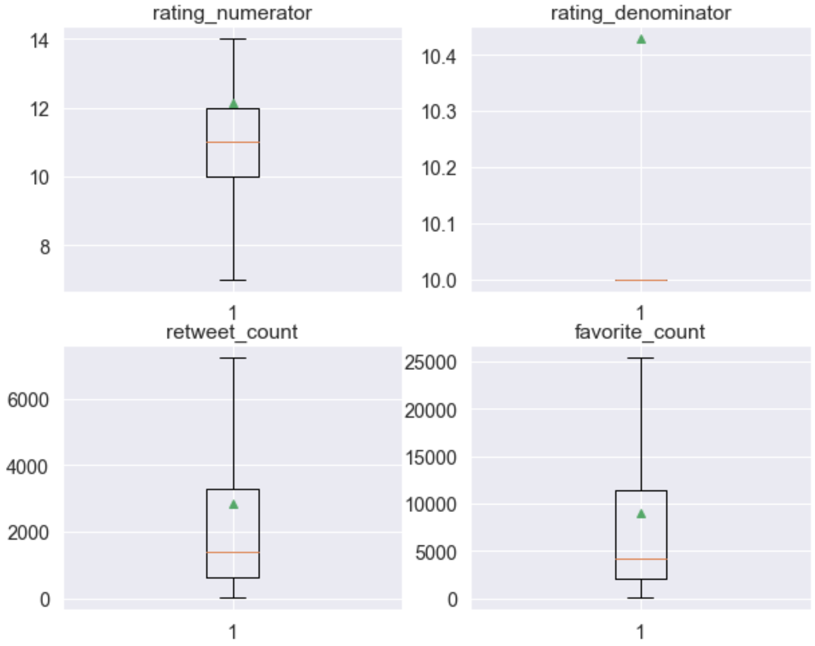


Figure 1: Mean, box and whisker plots of rating numerator, rating denominator, retweet and favorite count

Figure 1 shows the mean illustrated by a green triangle dot, and the box and whisker plot (including the median, upper and lower quartiles, and highest and lowest observations) of four variables: rating numerator, rating denominator, retweet and favorite count.

As can be seen, the mean and median rating numerators are around 12.14 and 11 respectively. At 75 percentiles, most dogs get at the scale of 12 on rating out of 10. The mean and median retweet count are around 2842 and 1405 respectively. 75 percent of the tweets got up to around 3300 retweets and 11416 favorites. The median and mean favorite count are 4196 and 8981 respectively. The mean values are much higher than the median values due to extremely high outliers.

2. Visualization analysis:

In this part, visualizations were created from the data using Python commands in Jupyter Notebook to answer the six following research questions.

Research question 1: How are correlated the variables?

To answer this question, I drew a correlation heatmap (Figure 2) for four different numerical variables, including rating numerator, rating denominator, retweet and favorite count.

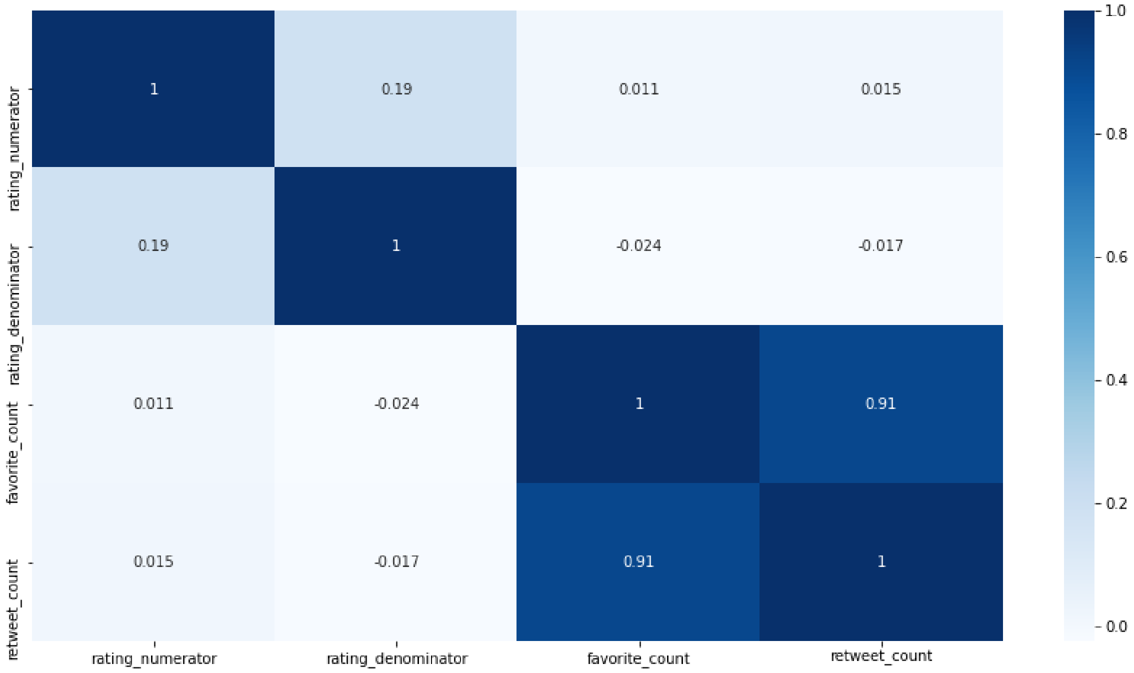


Figure 2: Correlation heatmap

According to Figure 2, there is a strong correlation between the number of favorites and retweets, which is expected. Moreover, there are no correlations between dog ratings and the number of retweets/favorites. Figure 3 below shows that there is a strongly positive correlation (+0.91; shown by correlation heat map) between the number of retweets and favorites.

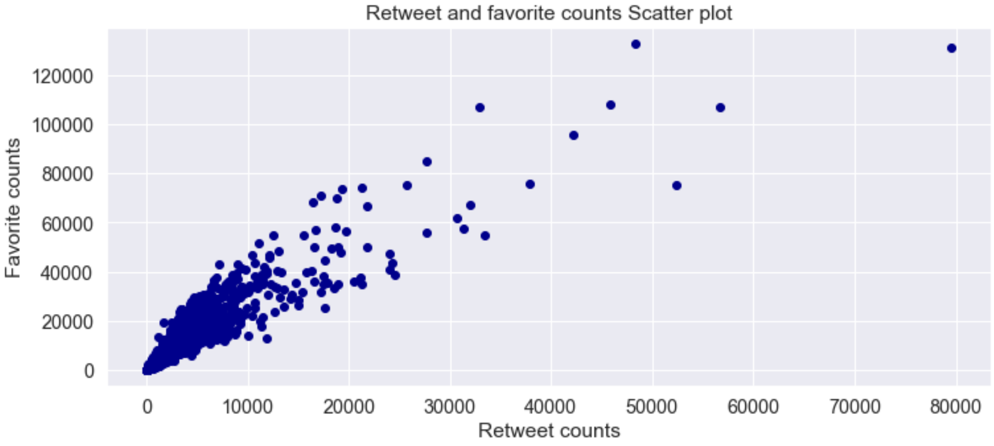


Figure 3: Retweet and favorite counts scatter plot

Research question 2: How did the retweet count and favorite count improve over time?

In Figure 4, in the beginning, the favourite counts and the retweet counts are at a similar level, yet the number of tweets per time is more. As the 2016 and 2017 progress, the number of tweets per time decrease (seen via the density of blue and orange dots), but then the number of the favourite counts and retweet counts becomes higher and higher.

Another trend noticed is that favourite counts seem to increase drastically going up to 120000 for a few tweets, yet the retweet counts remain less than 50000 for most of the time.

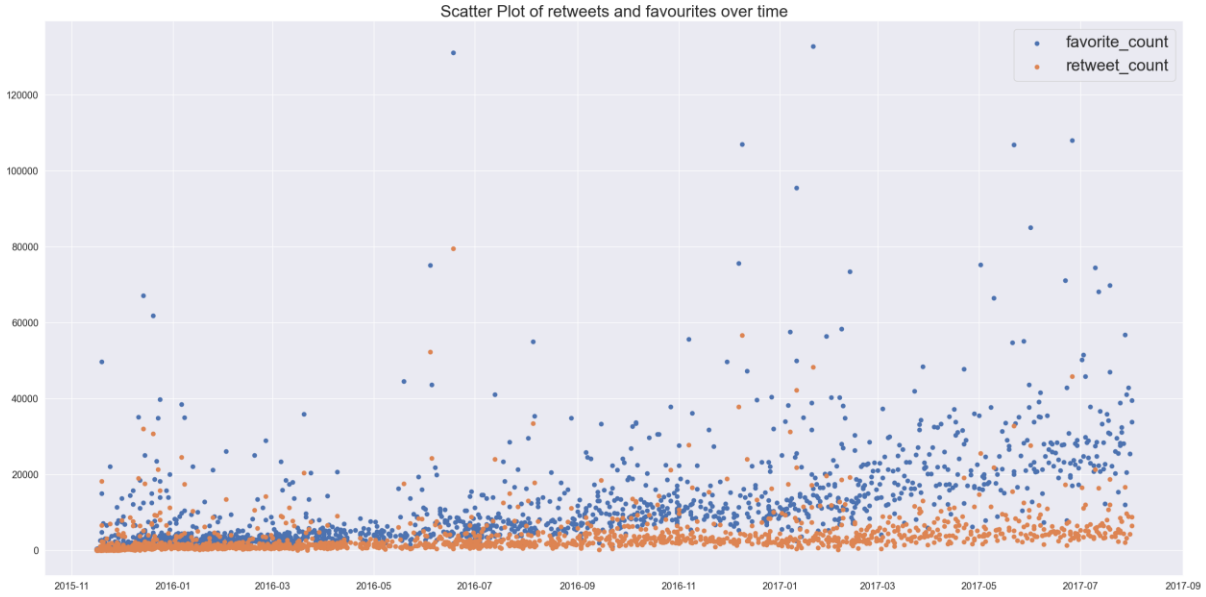
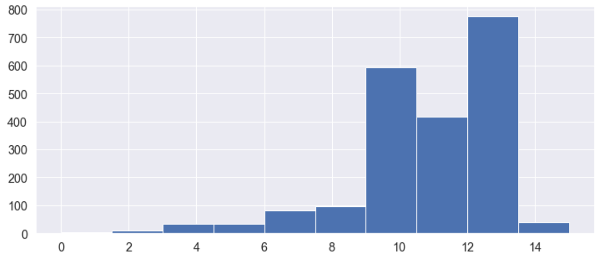


Figure 4: Scatter Plot of retweet and favorite count over time

Research question 3: How is the rating distribution?



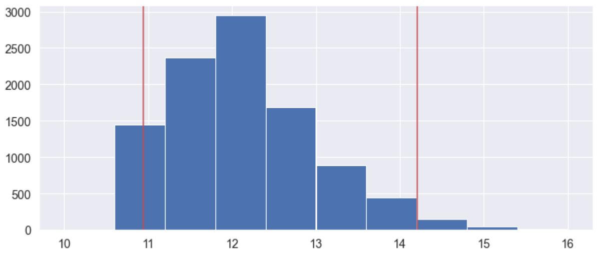


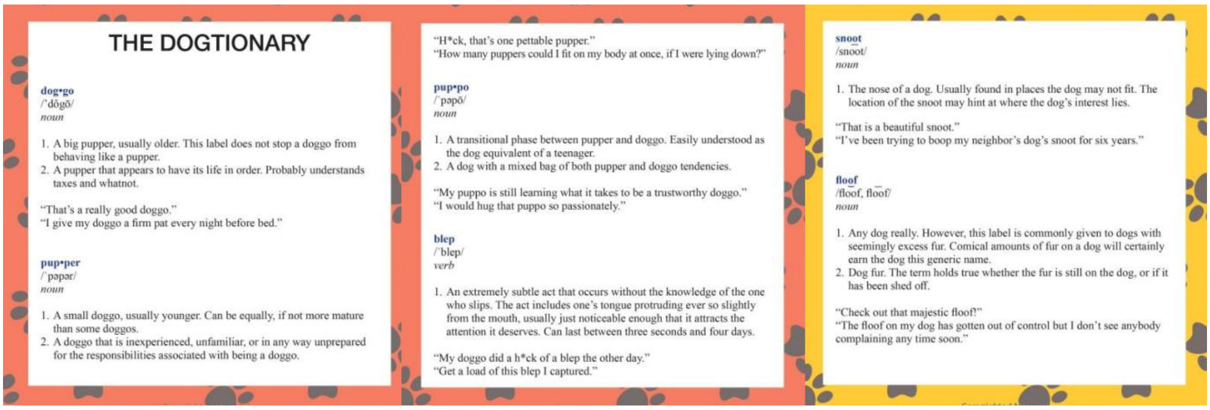
Figure 5: Rating numerator histogram

Figure 6: Confidence interval of mean for rating numerator values

We can see that the rating numerator histogram is skewed to the left. Moreover, we defined the confidence interval and with the certainty 95% we can say that the mean value of the numerator will be in between 11 and 14.

Research question 4: What stages of dog are in the tweets?

Only 336 out of 2097 tweeted dogs were identified with their stage of life (i.e. doggo, pupper, puppo, floofer and multiple). The definition of each stage is in Figure 7 below. For the stage of multiple, it happens when there are two or more dog stages mentioned in a tweet.



THE DOGTIONARY

dog•go
/ˈdōgō/
noun

1. A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
2. A pupper that appears to have its life in order. Probably understands taxes and whatnot.

"That's a really good doggo."
"I give my doggo a firm pat every night before bed."

pup•per
/ˈpəpər/
noun

1. A small doggo, usually younger. Can be equally, if not more mature than some doggos.
2. A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

"H*ck, that's one pettable pupper."
"How many puppers could I fit on my body at once, if I were lying down?"

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

"My puppo is still learning what it takes to be a trustworthy doggo."
"I would hug that puppo so passionately."

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one's tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

"My doggo did a h*ck of a blep the other day."
"Get a load of this blep I captured."

snoot
/snoot/
noun

1. The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog's interest lies.

"That is a beautiful snoot."
"I've been trying to boop my neighbor's dog's snoot for six years."

floof
/floof, floof/
noun

1. Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
2. Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

"Check out that majestic floof!"
"The floof on my dog has gotten out of control but I don't see anybody complaining any time soon."

Figure 7: The Dogtionary explains the various stages of dog: doggo, pupper, puppo, and floof(er) (via the #WeRateDogs book on Amazon)

According to Figure 8, among 336 tweets that mentioned dog stages, dogs in ***pupper*** stage of dog life cycle get most tweets, which is expected as they are adorable.

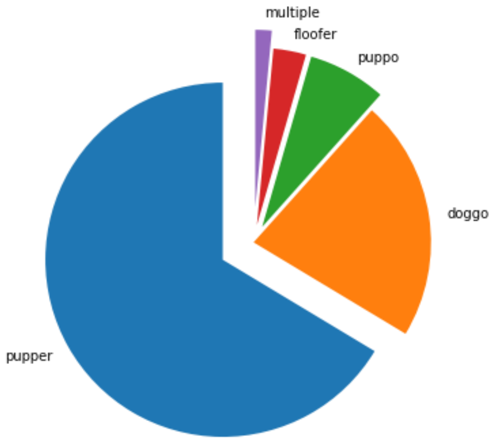


Figure 8: Pie chart of dog stages

Research question 5: What type of dogs are there in the tweets?

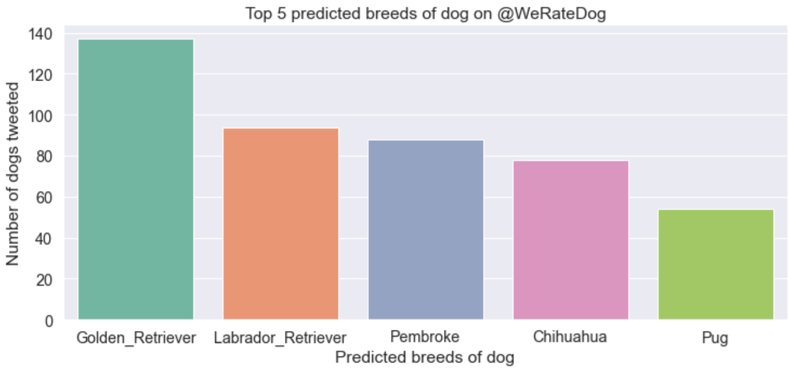


Figure 9: Bar chart of top 5 breeds of dog

The breeds of dogs were not extracted directly from the texts of the tweets in the dataset. They were identified and classified based on the image predictions. Out of 11 breeds of dog identified, The most popular dog, based on image predictions, is a Golden Retriever (137), followed by Labrodor Retriever (94) and Pembroke (88). It is a mix of small and big dogs!

Research question 5: What are the most common names of dog mentioned in the tweets?

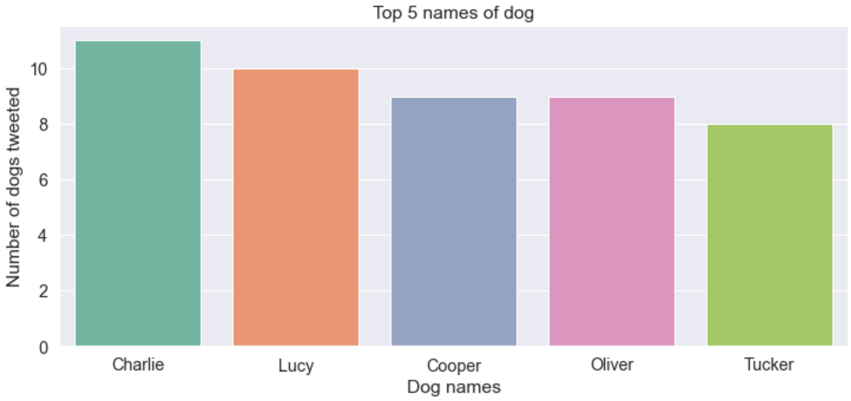


Figure 10: Bar chart of top 5 names of dog

The bar chart shows the top five names people have kept for their dogs. The top favorite names are Charlie, Lucy, Oliver, Cooper, and Tucker.

Research question 6: What are the tones of language and most common words used on this @WeRateDogs Twitter account?

One fun and easy way to visualize the texts from the tweets in our dataset is the word cloud as it shows the most frequently used words from the tweet texts and display in a customizable image. To illustrate the spirit of @WeRateDog Twitter account, I use dog paw print for the outline of the word cloud in this project.

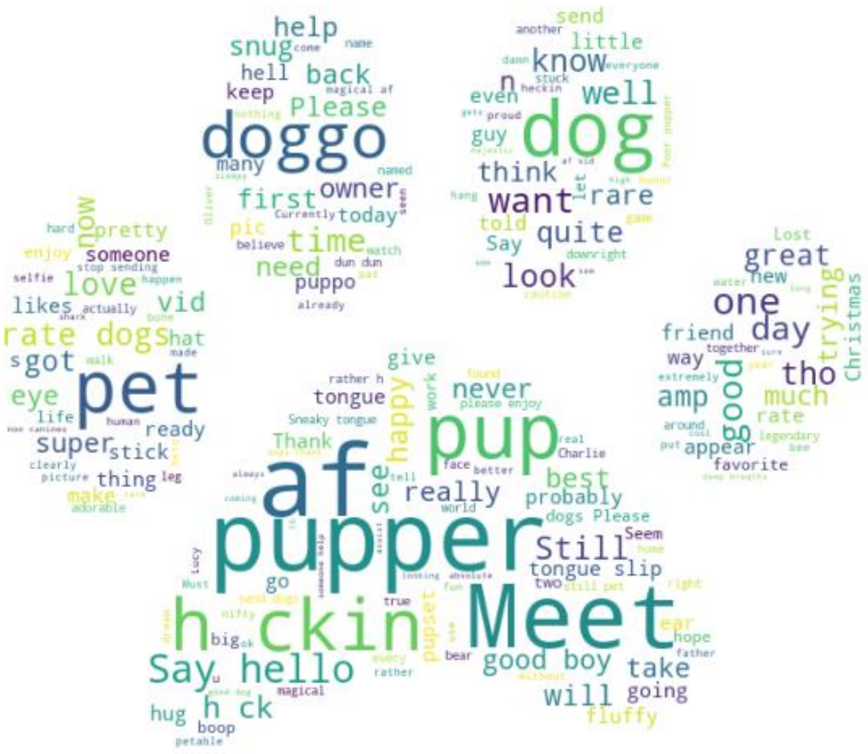


Figure 11: Word Cloud with a dog paw print

The larger the word, the more frequently it is used in the tweets in this dataset. From the word cloud, several most frequently words used in the tweet texts are *'pupper', 'dog', 'doggo', 'pup','pet', 'say hello', 'meet', and 'af'. We can tell that the language used in these tweets are very 'doggy' (of course), fun, happy and casual.

Conclusion:

After cleaning all the datasets, merged all three datasets into one single dataset. I get a dataset with basic tweet data of 2097 tweets and some other dog-related data such as dog names, dog stages and predicted breeds of dog.

Here is the summary of all the insights I have analyzed:

- At 75 percentiles, most dogs get at the scale of 12 on rating out of 10.
- On average, there are more favorite counts than retweet counts with 8981 and 2842 respectively.
- As time goes on, more people are favoriting a tweet than retweeting a tweet.
- There is a strongly positivie correlation between the number of retweets and favorites.
- Top 5 dog names are Charlie, Lucy, Oliver, Cooper, and Tucker.
- Top 5 breeds of dogs based on image predictions are Golden Retriever, followed by Labrodor Retriever and Pembroke.
- The pupper stage has the highest number of dogs among all other stages.
- Several most frequently words used in the tweet texts are 'pupper', 'dog', 'doggo', 'pup','pet', 'say hello', 'meet', and 'af'. The language used in these tweets are very 'doggy' (of course), fun, happy and casual.

The data wrangling project was one of the most fun projects that I have done to date. It is always fun to sit down with a data set and try to gauge what it is telling you. Along with the fun bit, there was also quite a bit of learning in this project. It will be available on my Github, and I would love to hear what you think of it!