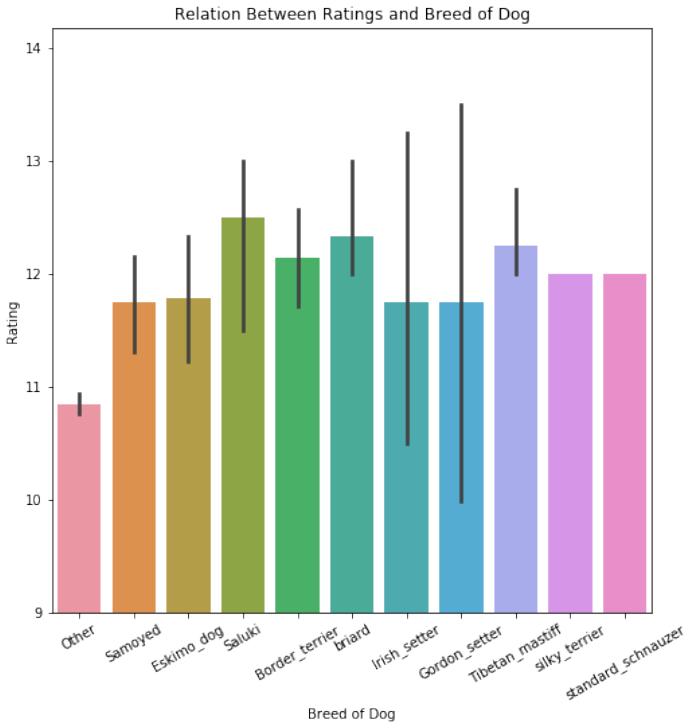
Data Analysis & Visualization for WeRateDogs Dataset

This report is the data analysis and visualization for WeRateDogs Dataset. WeRateDogs is a twitter user who rate the dog in picture and leave with comment. This dataset includes the tweets sent by WeRateDogs from 2015-11-15 to 2017-8-01. Dataset is created after some data cleaning, some unuseful information was removed.

This dataset includes some basic information of the tweets like tweet id, sent time, text content. Also, information extracted from each tweet's text are included, like ratings, dog name, dog stage. In addition to these, the image prediction for each image in tweets is provided. After some data wrangling analysis to this interesting dataset, several insights are come up.

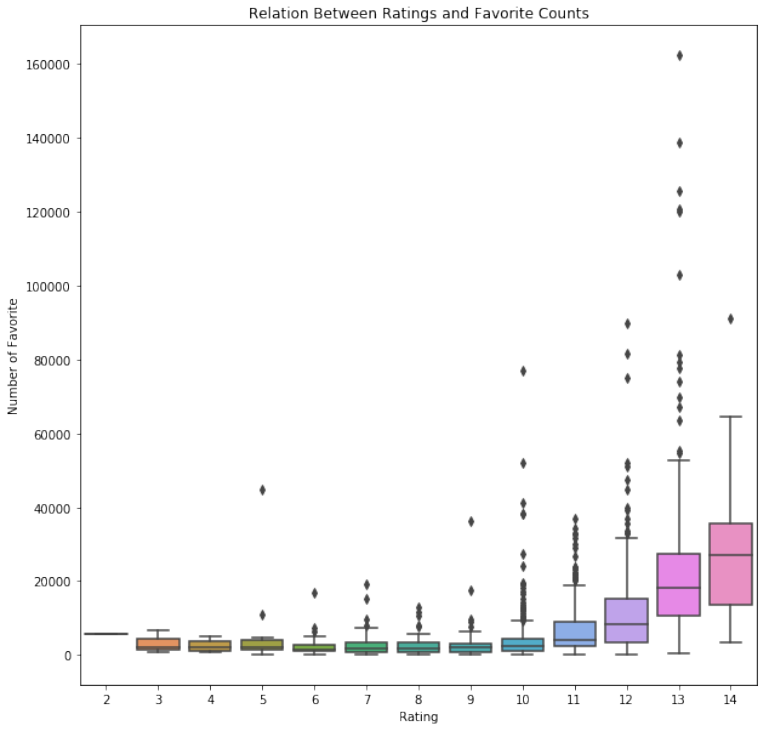
1. Which Breed of Dog get higher ratings?

There are 111 breeds of dogs according to the image prediction. After calculating the mean value of the ratings for each breed, we get the top 10 highest rated breed: Saluki, briard, Tibetan mastiff, Border terrier, silky terrier, standard schnauzer, Eskimo dog, Irish setter, Gordon setter, Samoyed. A visualization for this result was created as shown on the right. We can see that, saluki get the highest rating, approximately 12.5. Except the top 10 highest rated breeds, all other breeds' ratings are calculated under "Other" with average rating about 10.8.



2. Relation between ratings and favorite counts.

In this dataset, favorite counts show how many people clicked the heart and show their like to the specific tweet. We are interested in whether the number of likes, that is how popular the dog is, is consistent with the rating. So the relation of the ratings and favorite counts is draw in the second bar plot. As we can see from the figure, with the rating increases, more likes are added to the tweet, especially when ratings are over 11. That's because WeRateDogs rarely give ratings less than 10.



3. Dog Stage Distribution

There are four stages for dog: doggo, floofer, pupper, puppo. A pie plot is created to show how the 4 stages dog tweets distributed. It seems like the number of tweets for pupper is the largest. Floofer and puppo take only a small part of the distribution.