Aigerim Shopenova
Data Analyst Nanodegree Program

Act Report

Project: Wrangle and Analyze Data

Data

1) The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced."

2) Image Predictions file a table full of image predictions (the top three only) using neural networks alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

3) JSON file with tweets includes retweet count and favorite count. Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

Data Wrangling

Before data analysis process, I gathered and cleaned the data in order to make it easier to understand and find some interesting insights. I assessed the data both visually and programmatically for quality and tidiness. Data quality has several dimensions such as completeness, accuracy, validity and consistency, which were kept in mind during data wrangling stage.

I got 1657 tweets with good quality data, which were further analyzed.

Motivation

I was curios about dog stages and their different characteristics such as the number of tweets in the dataset, average rating, the number of retweets and favorites.
Also, I was curious about the general image of the data using descriptive statistics to see the results on rating, number of retweets and favorites and confidence of prediction data.

Insights

- Pupper is the most popular among defined dog stages. However, the majority of dog stages is not defined, what creates some uncertainties (image 1).

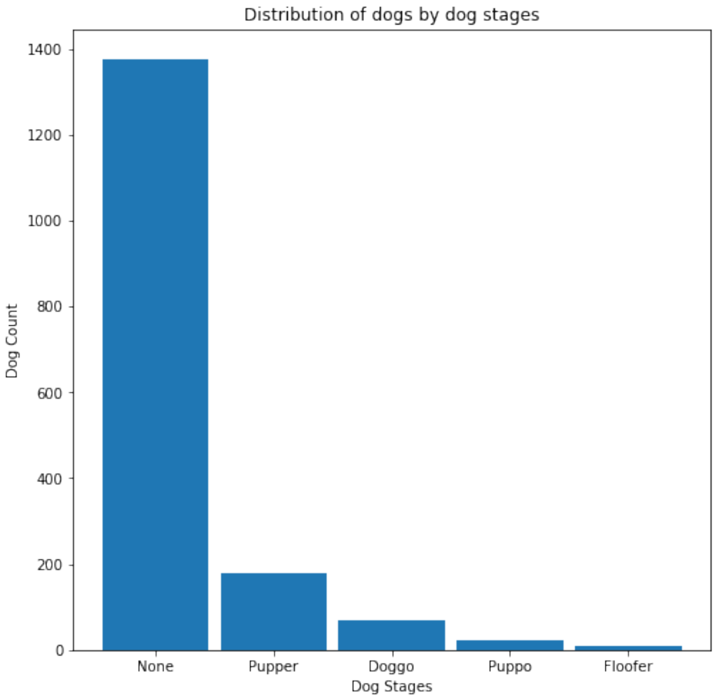

Image 1. Distribution of dogs by dog stages

- The average rating for each dog stage is around 12. The most popular dog stage by the rating is Puppo. However, None values are presented and it creates some uncertainties (image 2).

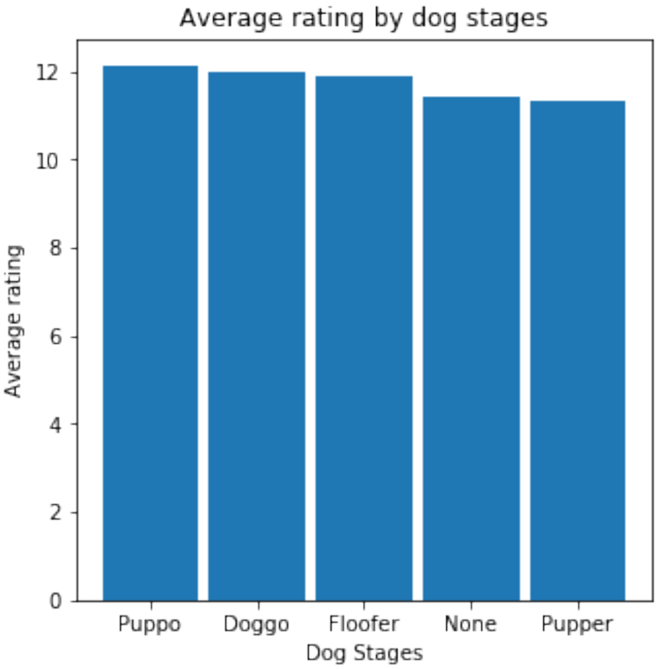

Image 2. Average rating by dog stages

- According to the number of retweets, Doggo and Puppo are the most popular among other dog stages (image 3).

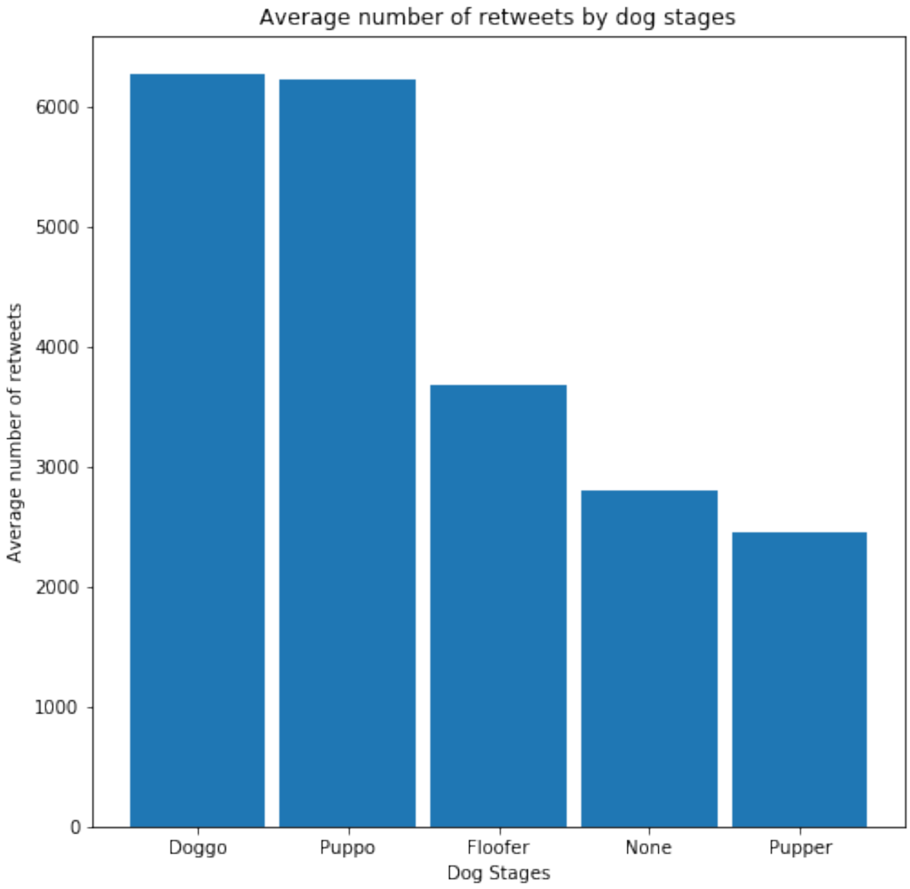

Image 3. Average number of retweets by dog stages

- Puppo and Doggo are the most favorite dog stages, which have a significant number of favorite counts in comparison with other dog stages (image 4).

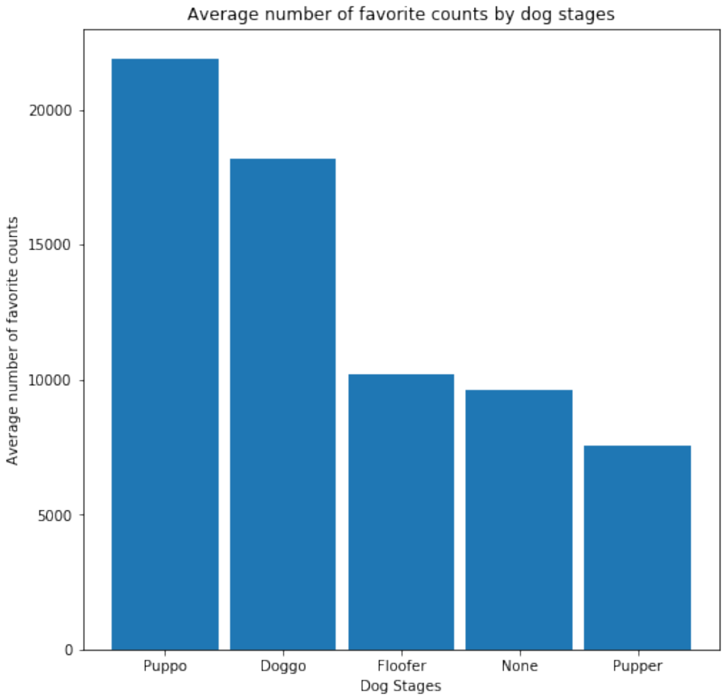

Image 4. Average number of favorite counts by dog stages

- As it was explained in the project, dog rating should >= 10, which is correct after data cleaning. Also, it should be noted that the majority of ratings belongs to the interval between 10 to 12 (image 5).

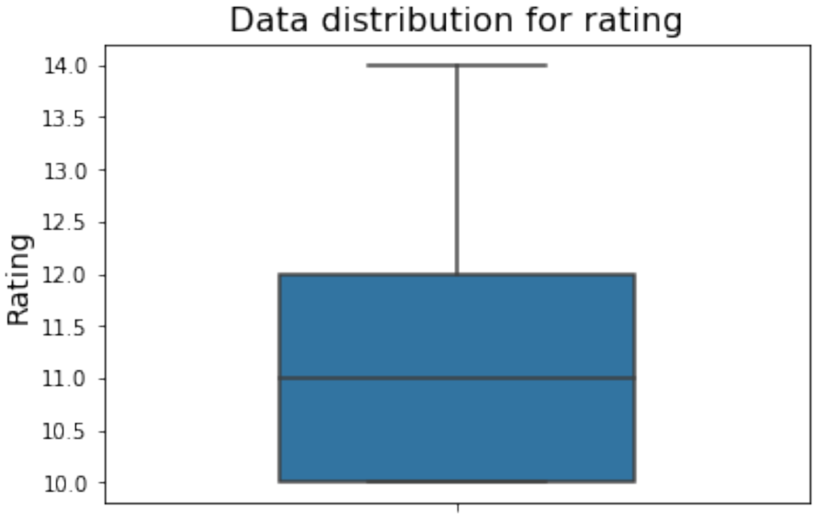

Image 5. Data distribution of tweets for rating

- It's clear that retweets lie in the interval between 0 and less than 10,000. However, there are some number of outliers for tweets, which can be between 10,000 and 80,000 (image 6).

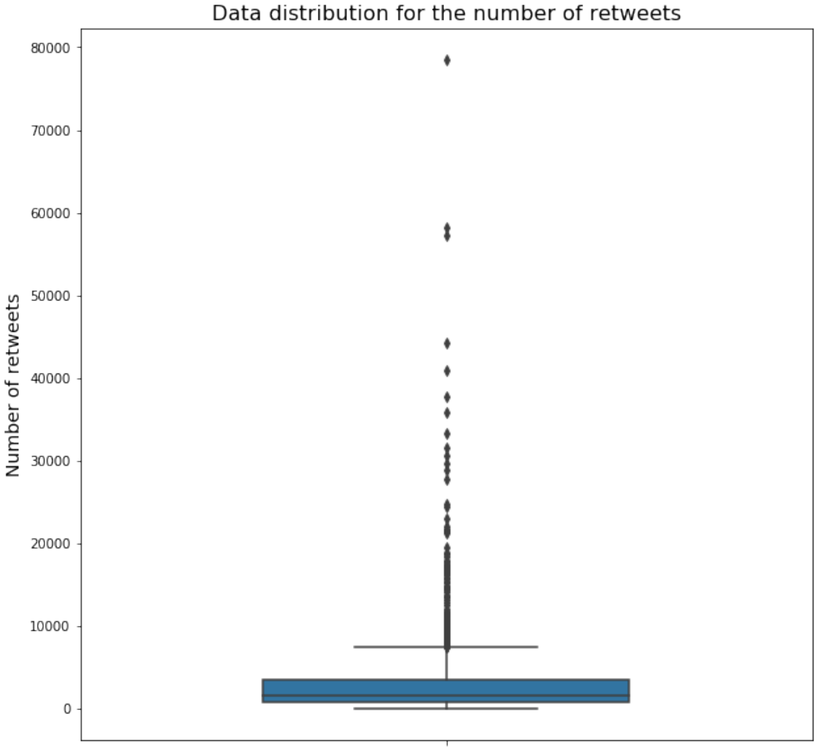

Image 6. Data distribution for the number of retweets

- The number of favorite tweets vary between 0 to 3,000. However, there are some outliers, where the number of favorites lie between >3,000 and 16,000 (image 7).

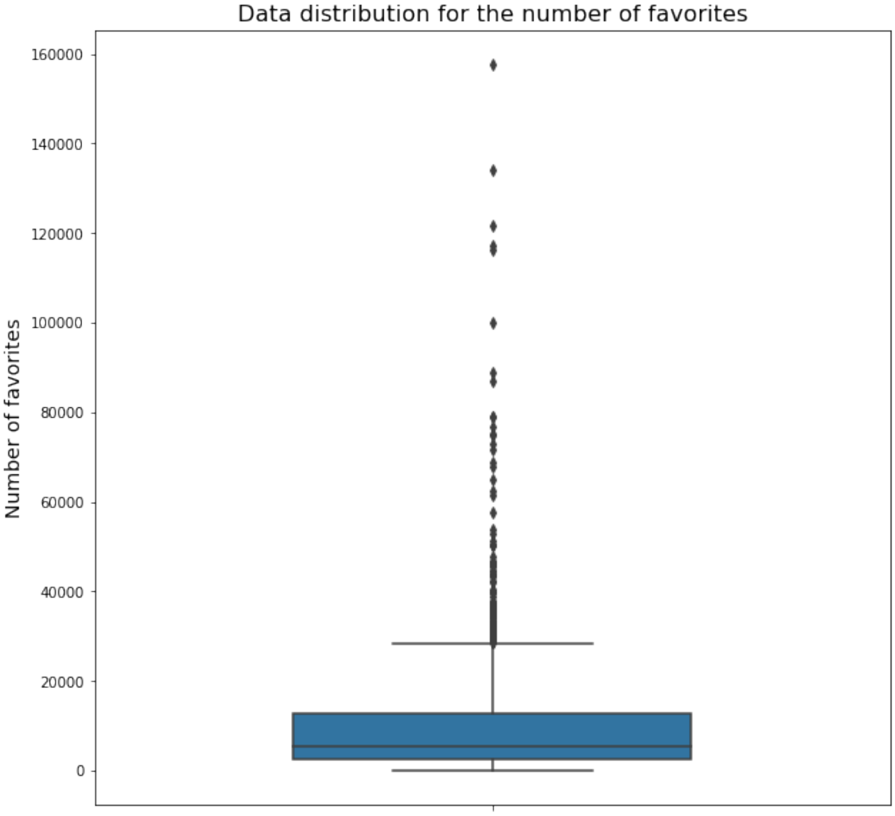

Image 7. Data distribution for the number of favorites

- 50% of data for the confidence of prediction of a dog breed belongs to the interval between 0.4 to 0.85, where the mean probability is around 0.6 (image 8).

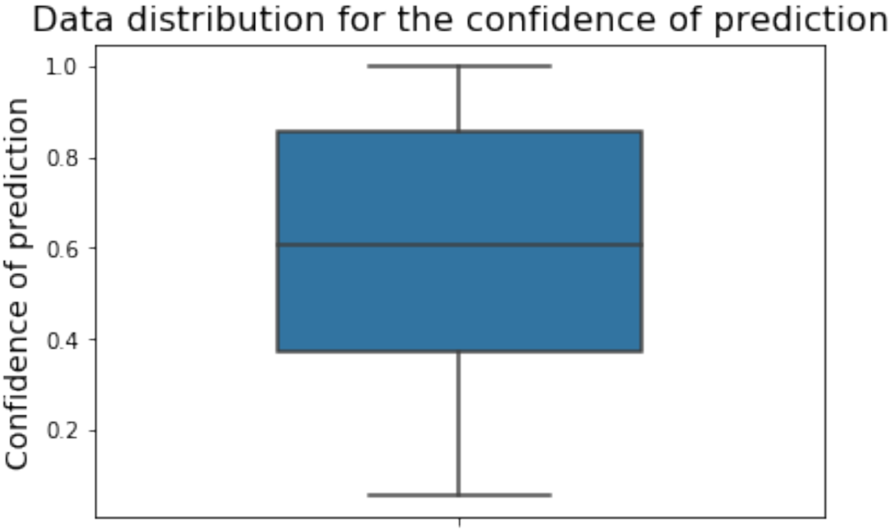

Image 8. Confidece of prediction and its distribution in the tweets

Overall, Doggo and Puppo dog stages are the most popular among defined dog stages. Hovewer, the majority of dog stages does not have data. This creates some incertanties in the real image on data. 50% of images were prdicted in 0.4 to 0.85 confidence interval, where the mean probability is around 0.6.