Act Report

Introduction:

This Wrangle and Analyze Data Project is Uadacity's Data Analyst Nanodegree project. This project involves wrangling of data from various sources associated with Twitter account for @WeRateDogs. It was launched in 2015 by college student Matt Nelson, and has received international media coverage for its popularity. The data describe all rating and used unusual rating method where Numerators are mostly above 10. After gathering the data, quality and tidiness issues were assessed and then cleaned. Finally, three visualizations were created and insights can be found below.

Background:

Before I started visualization, I wrangled these three datasets, created new variables, drop low quality data and merged all variables, which were interesting to me into one dataset. Initially upon gathering we had 2356 observations but after cleaning only 1928 entries and 13 columns present in the final data. These columns are: tweet_id, timestamp, source, text, expanded_urls, name (of the dog), stage (pupper, puppo, dogger and floofer), rating, jpg_url, prediction_algorithm, confidence_level, favorite_count and retweet_count.

Real-world data rarely comes clean. Some columns have missing values such as: name, stage and prediction_algorithm. This should be keeping in mind while looking at the visualizations below.

Analysis and Visualization:

Part I: Visualizing the three most frequent dog stages

In order to present a high-quality data, I created a new column named "stage" and put the status of all puppies under column. After cleaning the data, I would like to know which three stages are the most common for dogs. According to the graph, we know that most of the dogs have pupper, doggo and puppo as their stages.

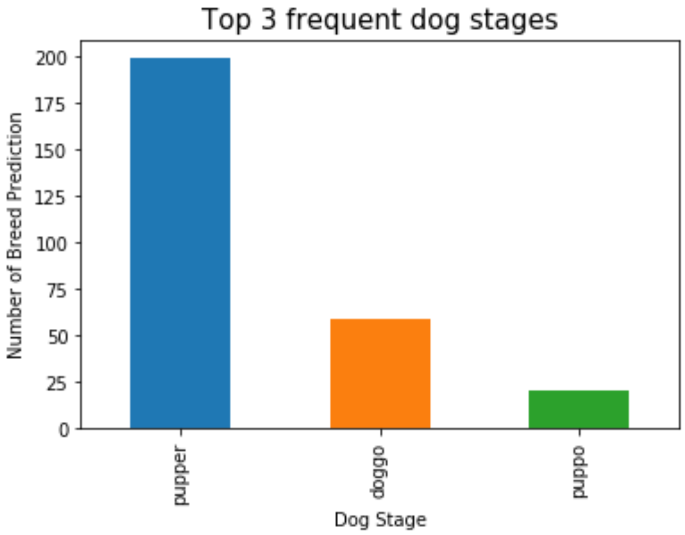

Figure 1: Top 3 frequent dog stages

Part II: Visualizing the five most frequent dog names

Although I didn't sort out the name column, I am curious what kind of name the breeder would like to give to the their puppies. The result shows that most of the dogs are named Charlie, Lucy, Oliver, Cooper and Penny.

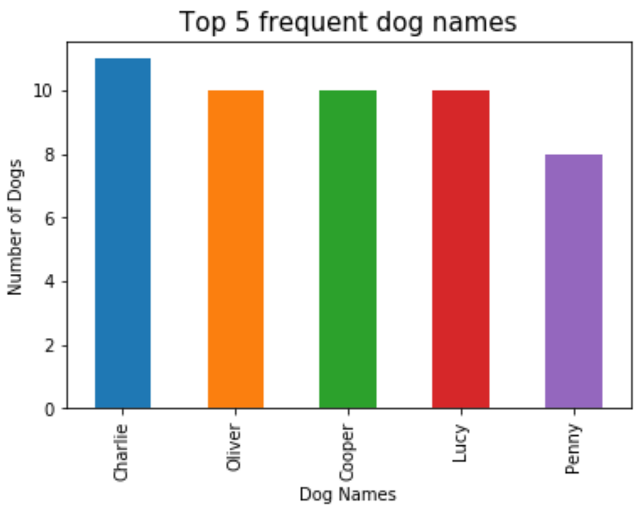

Figure 2: Top 5 frequent dog names

Part III: The Distribution of Favourite Count compare to Retweet Count

WeRageDogs had over 4 million followers; their tweets are likely to get many favourites and retweets. In figure 3, it can be seen that favourite and retweet counts are highly positively correlated. The higher the favourite count the higher the retweet count. More than that, we can found that the distribution of favourite count is located to the right of the distribution of retweet count. Therefore, people favour the tweets more often than retweet them.

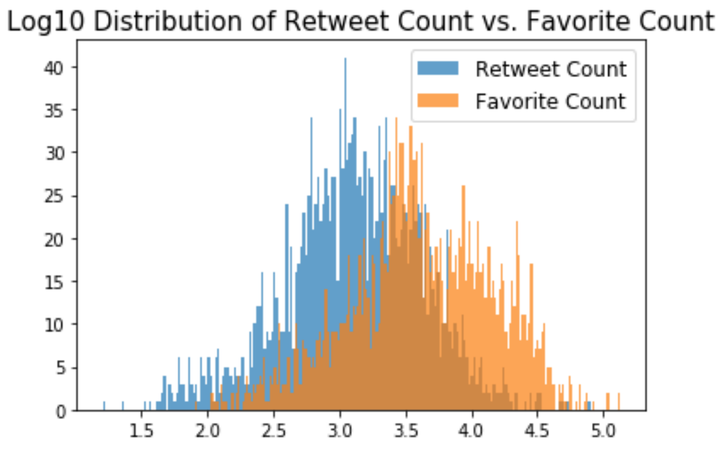

Figure 3: The Distribution of Favourite VS. Retweet