Act Report

The goal of this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

Enhanced Twitter Archive

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." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

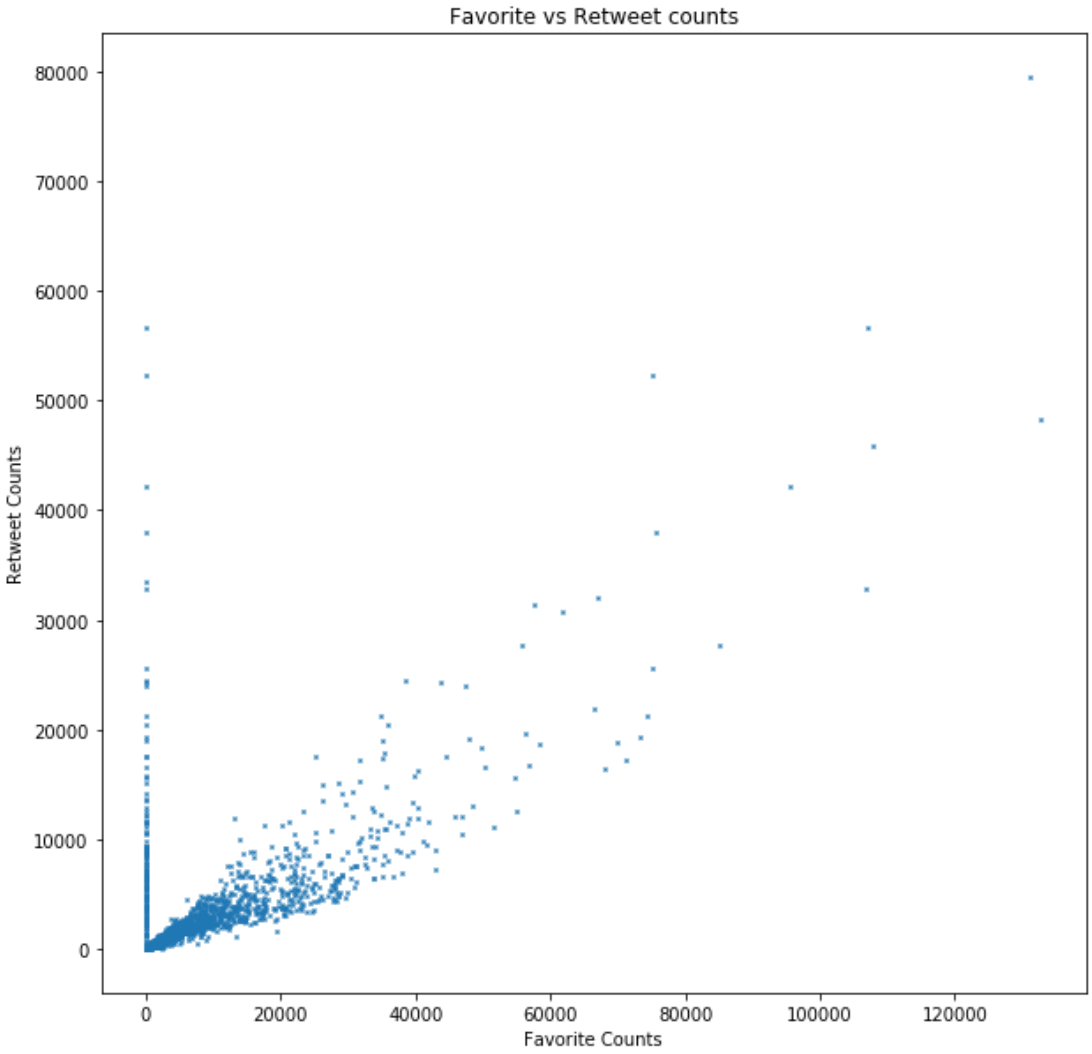
Image Predictions File

One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) 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).

Analyzing

From df.describe() function, we can see that there is a huge difference between 75% and max for column rating_numerator and rating_denominator (rating_numerator: 12 for 75% and 1776 for max, rating_denominator: 10 for 75% and 170 for max). There might be some typos in the dataset. In this project, we will not continue to analyze this.

We could get the most common dog breeds by using df.p1.value_counts(), df.p2.value_counts() and df.p3.value_counts() functions.



The above plot is the relationship between favourite and retweet account, we could know that there seems to be a strong positive relationship between favourite and retweet.