Act Report

Data exploration stems from questions about our dataset. After our wrangling effort, we had clear tables with clean observation.

One of our wrangled dataset contains information about tweets on WeRateDogs twitter page, such as: name of the dog, the dog stage, rating, number of favorites, and number of retweets.

The other dataset contains information about and a dog classifier, that recognize what kind of dog is in a photo, it gives three predictions, each contain predicted object's name, is it a dog or not, and prediction confidence.

Among the questions we asked about the data are, how many each stage of dog is in our first dataset, what's the relationship between number of favorites and number of retweets, what's the relationship between number of retweets and rating, and finally, in our second dataset, how reliable is our classifier across all predictions.

- **how many each stage of dog is in our first dataset?**

To answer this question, we had to plot a histogram of stage of dog count in our dataset. Turns out, Puppers seems mentioned the most on it's own, with doggos second behind, but doggos seem to most be paired with other stages in our dataset, and most of the data had missing values in our dataset. So in terms of popularity from tweets only in our dataset, puppers stand out the most.

- **what's the relationship between number of favorites and number of retweets?**

By plotting a scatter plot, we noticed that they have a strong positive correlation, which is not surprising, because a tweet retweeted many times must be likable, people would favorite for the same reason.

- ## what's the relationship between number of retweets and rating?

This one was interesting, since that twitter page have a special rating system (i.e. 11/10, 14/10, 12/10) we had to find a way to translate their rating to our understanding, so we just calculated the rating shown, which seems to have worked, but some of the ratings was exaggerated with different high numbers which wasn't their know trademark rating, so in most cases it didn't get a high number of retweets. And that explains the few outliers that had high translated rating and low number of retweets.

Same as before, we used a scatter plot for analysis, and as expected, the higher the rating the higher the number of retweets.

- ## how reliable is our classifier across all predictions?

We plotted three histograms for each prediction the classifier makes, between all 3 predictions of the algorithm, most of its probability is below 50% confidence, but as we can see, the algorithm's top prediction rarely go below 30% which is pretty good for image recognition, and it has many instances with high confidence. So our classifier seems moderately reliable.