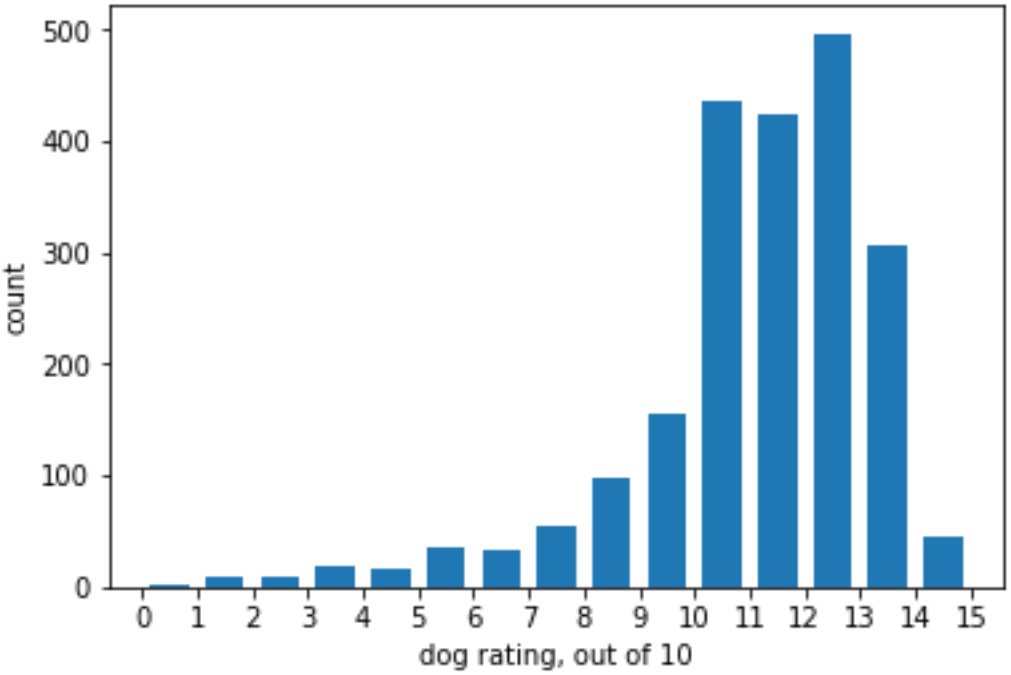
Udacity Data Analyst Nanodegree

Project "Wrangle and Analyze Data"/ Brief Analysis of Data

Richard Steele

Insight #1

The ratings, while popular, are of course highly subjective. Setting aside the numerator-greater-than-10 observation (after all, we are informed that "they're good dogs"), quite a few outliers exist in the unedited dataset, ranging from the low 40s all the to 1,776. In order to introduce some reasonableness into the analysis, I created a smaller frame with all ratings under 30. Given the nature of the system, I expected a distribution of ratings skewed to the left, and was not disappointed:



As seen above, the ratings drop off dramatically around 14 or 15. Programmatically, I found the number of "reasonable" ratings to be 2,144, leaving 20 outliers to be excluded.

Insight #2

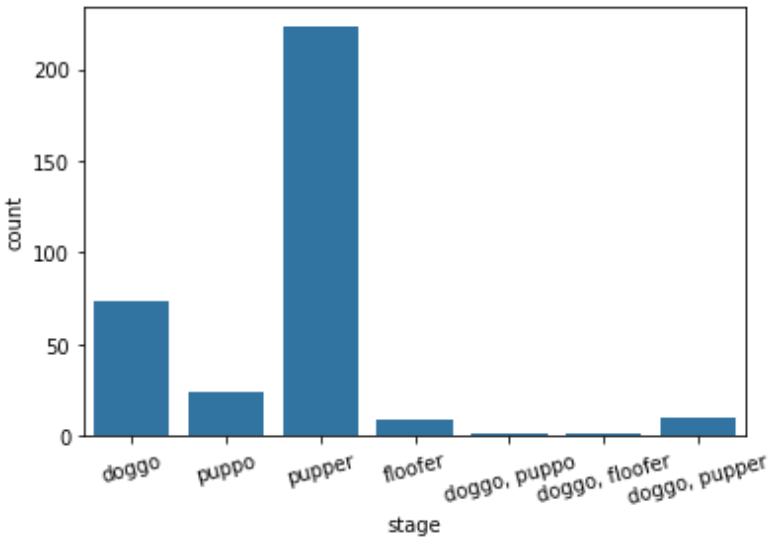
According to the Project Motivation, the image-prediction data are generated by an algorithm which "predicts" the breed of dog referenced in the images. For each of the three predictions the algorithm provides a "confidence" figure between 0 and 1, and then assigns "True" or "false as whether the algorithm correctly predicted that the image is that of a dog. I compared the number of "True" and "False" outcomes with the mean confidence figures for each prediction:

PREDICTION	MEAN OF CONFIDENCE LEVEL	COUNT OF TRUE PREDICTIONS	COUNT OF FALSE PREDICTIONS
Prediction 1	0.594548	1,532	543
Prediction 2	0.134589	1,553	522
Prediction 3	0.0603241	1,499	576

I am not well-versed in machine learning or deep learning, and even if I was rudimentarily familiar therewith, I am not familiar with the algorithm in question. Having said this, I still wonder how the confidence level could drop so precipitously (on average) and still show prediction success rates of 73.8%, 74.8%, and 72.2% respectively. Given the whimsical nature of the prediction exercise (the algorithm made predictions such as "otter" and "bath towel"), I can only report what I see.

Insight #3

Although 1,822 out of the 2,163 rows (84%) of the rating archive dataset do not contain data in the 'stage' column, I wanted to know which of the "maturity stages" (doggo, floofer, pupper, and puppo) is the most prevalent. In a manner like that in Insight #1, I created a subset of the archive wherein the 'stage" column contained one of the maturity descriptors. An ordered plot produced the following:



Clearly, the "pupper" is the most frequent. A "pupper" is defined as "as small doggo, usually younger. Can be equally, if not more mature, than some doggos. (Also a) doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo." Doggo, the highest level, came in a distant second.