Udacity Data Analyst Project 4: Analysis

January 10, 2021

J. W. Thacker

For the first small analysis I decided to see what the most common type of predicted dog is since each tweet is comes with a picture of a dog. I decided to use the 'p1' feature which is the neural networks number 1 prediction of the breed of the animal in the picture. However, while sifting through the data in image_preds.csv I noticed that not all the predicted categories are dog breeds. For example, the particular dog below is, I believe, a shar pei pup that was mislabeled as a loaf of bread. However, these misclassifications were more infrequent than frequent but still definitely present in the predictions by the neural network, so I decided to look at the top ten most commonly predicted dog breeds instead. The most common dog breeds can be seen in the barplot but I think it's notable that retrievers are the top two breeds.

For the second analysis I decided to look at the distribution of 'dog stage'. I am not sure how Twitter determines the dog stage, but my guess is that they find some kind of evidence in the text of the tweet itself. I think this is supported by the fact that in figure 3 it was way more common that a dog stage was not assigned. According to the plot, 'pupper' was the most common dog stage that was assigned when one was assigned. According to Dogtionary the definition of a 'pupper' is : A small dogg, usually younger. Can be equally if not more mature than some doggos. Since the definition of doggo appears in the definition of pupper, the definition of doggo is: A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.

For the third analysis I decided to see if I could determine the distribution of the 'retweet count' variable, which I'll call X. This variable is actually a discrete variable, but since the values have such a wide range I decided

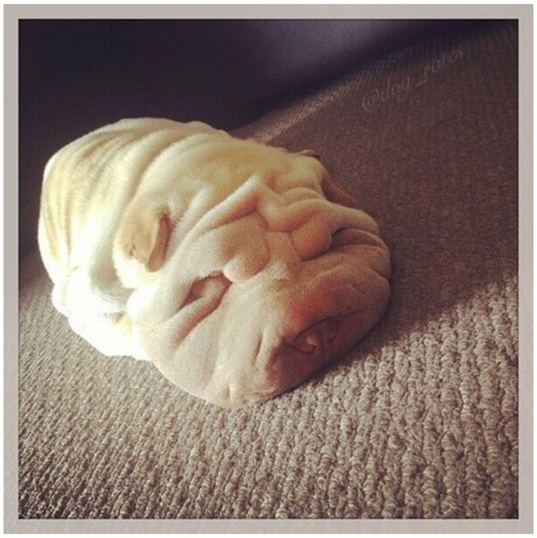

Figure 1: This shar pei puppy was mislabeled as a 'loaf of bread" :)

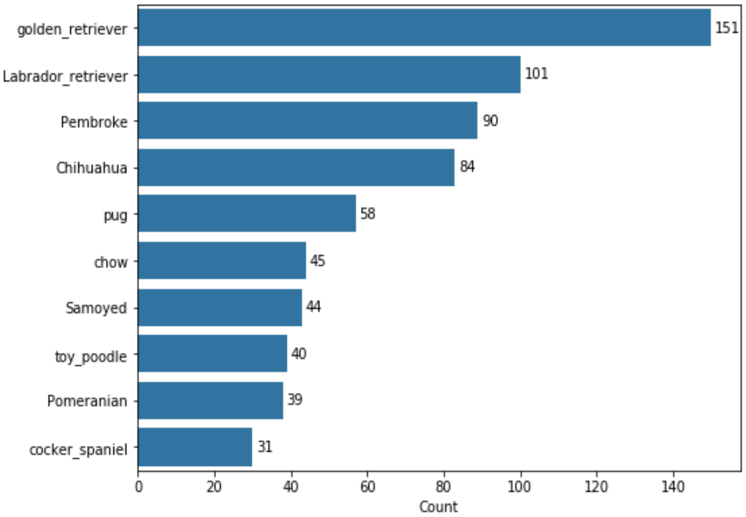

Figure 2: The most common predicted dog breeds from the neural network run by Udacity.

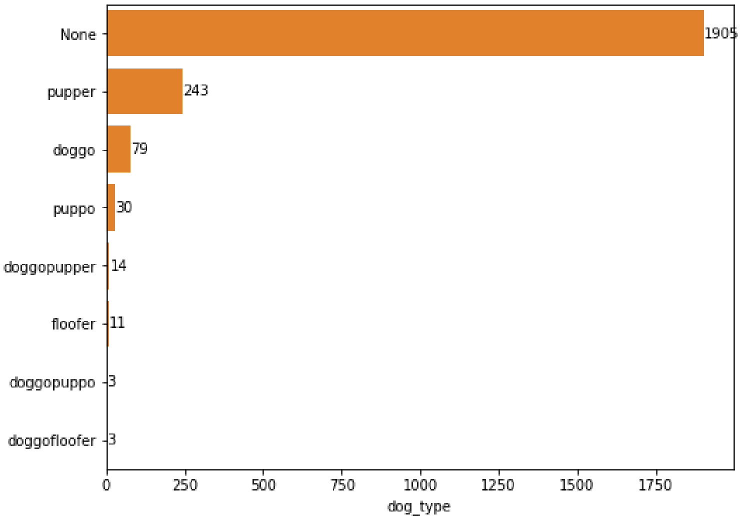

Figure 3: The most common dog stages as found the in tweets by Twitter.

to approximate it as a continuous variable. I know that 'count' variables are typically heavily skewed right. Therefore I had a feeling that the retweet count variable would have a log-normal distribution. To verify this I replotted the data on a log plot, namely $Y = log(X)$, and if the log of the data is normally distributed, then retweet count is normally distributed. The histogram of $Y = log(X)$ is seen in figure 4. In addition, the median (orange line) and the mean (red line) are plotted on the histogram as well. In order for me to say that, without a doubt, that the distribution of 'retweet count' is log-normal I would like to see that those values are approximately equal. BUT, the histogram is still very symmetric and the mean of the data is not too far away from the median, so I would say that it's still plausible that 'retweet count' is log-normally' distributed.

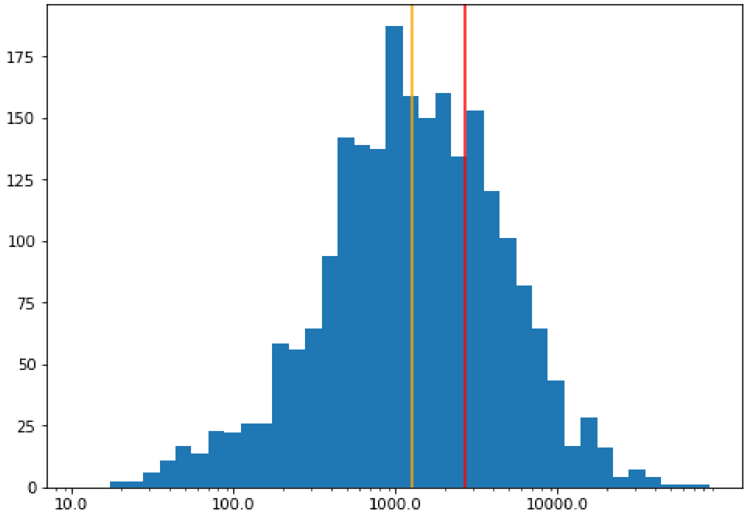

Figure 4: The histogram of $Y = log(X)$. The red line is the mean of Y and the orange line is the median of Y.