Analysis Report

Abel Flores Prieto

April 1, 2018

After gathering, assessing and cleaning the data, I was able to analyze the data more easily and more accurately having taken out unnecessary information. In this document, I go over three insights on the cleaned data.

The confidence percent for the number one prediction from a picture has a flat distribution, with a mean of 60% and a median of 60%. Figure 1 shows the histogram for the confidence percent for all three predictions. As we can see, the histogram for the number one prediction is pretty flat until we reach confidence levels of about 95% and more. From this histogram it is clear that the number one prediction is considerably more accurate than the other two predictions. Noting that from the number one prediction only 74% is actually a dog breed, I can think of two possibilities. Either the dog in the picture was wearing a costume or the dog simply looked like another thing.

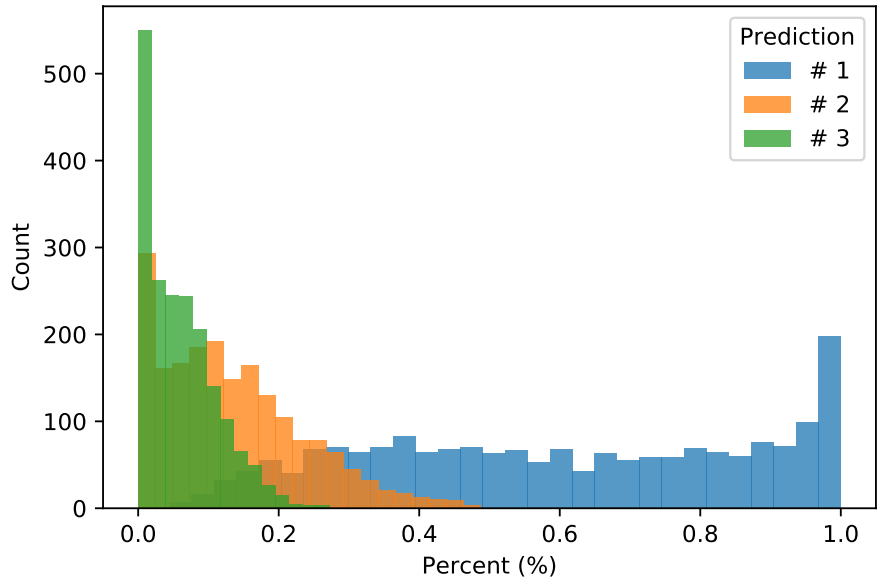

Figure 1: Histogram for the confidence percent for the prediction of tweet image.

Another insight comes from the rating for each tweet. Since a couple of tweets do not have a rating denominator of 10, either because there is more than one dog in the picture and 'WeR-

ateDogs' added all the ratings together, I created a new variable to obtain the ratio between the rating numerator and the rating denominator. The rating ratio has a mean value of 1.05 and a median on 1.1, which means that on average more dogs obtain a higher rating numerator than denominator. This is common for the Twitter account 'WeRateDogs', since they usually give scores such as 11/10 or 13/10.

Lastly, while analyzing the data I realized that the favorite count was very much correlated to the retweet count. Looking at Figure 2, we can see that as the favorite count increases, the retweet count increases as well. It is also good to note that tweets, on average get more favorites than are being retweeted. These two variables have a positive correlation coefficient of 0.94.

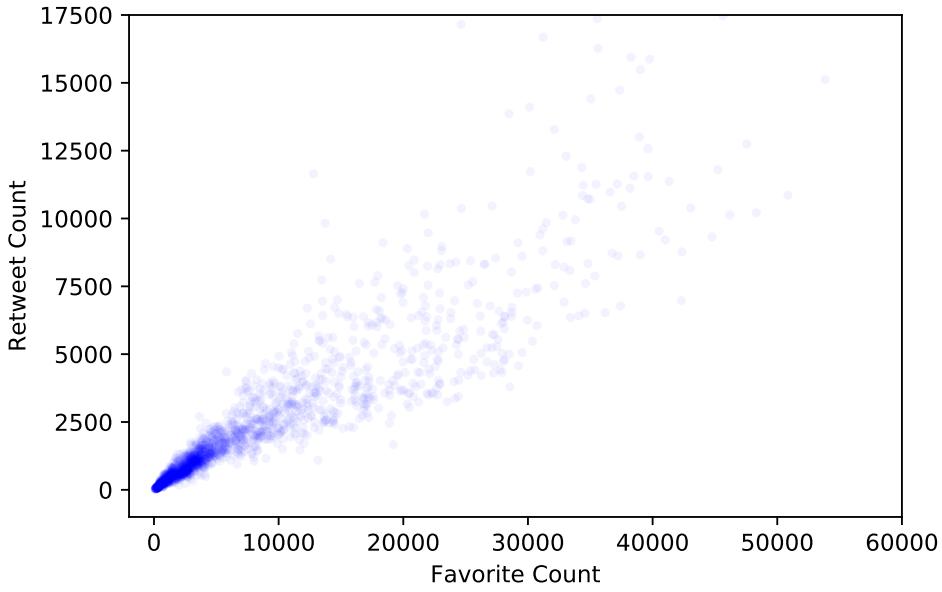

Figure 2: Plot for retweet count versus favorite count. On average, tweets have a higher favorite count than retweet count.