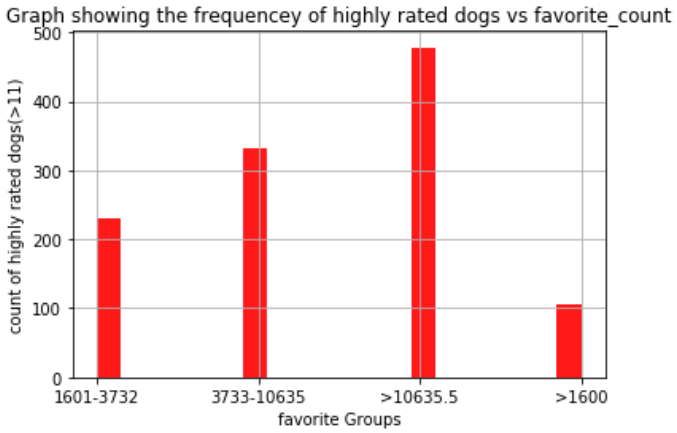
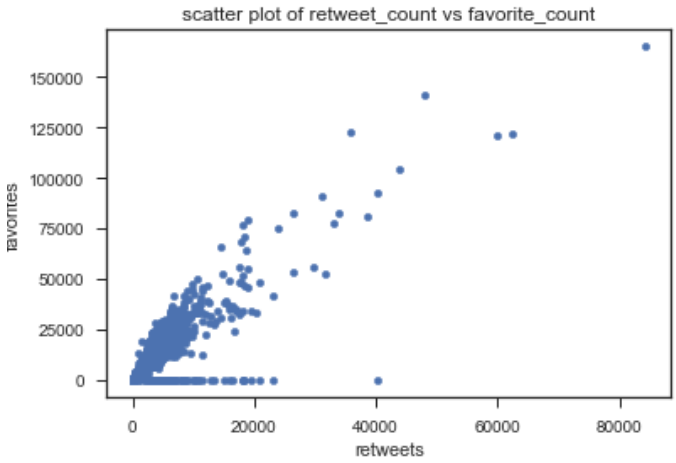
Miriam Wanjohi - Data Wrangling Report

Analysis

For analysis part of the project, I created dummy variables for 'favorite_count' and created a new categorical folder. I then plotted the categorical variable against all the dogs that rated over 11 (above the median). From the graph (right) 478 dogs that were rated above 11 had over 106355.5 favorite likes. Only about 106 dogs highly rated dogs (>11) had less than 1600 likes.



There was also a positive correlation between favorite and retweet counts as shown in the scatter plot with a correlation coeffient of 0.85. This makes sense since pictures with a lot of likes would most likely be more retweeted.



Using the clean data, I queried the dataset for the dog with the most retweets 84243 vs the one th at had the least retweets (only 12).

The most retweeted dog was a labrador retriever, with a caption "Here's a doggo realizing you can stand in a po…". It had a rating of 13/10 with 165045 likes. The url for the most Labrador did not hence I was not able to retrieve the dog image.

The dog that was retweeted the least was an English setter with a rating of 11/10 and only 12 retweets. The commentaty on it was “oh my. Here you are seeing an adobe setter giv…”.

Here is the picture (right) for the least retweeted dog: (I would think maybe the fake puppies made the picture less appealing hence the low retweet count).

I was curious about the outlier dog with a rating score of 1776/10. I am not sure if it was typo or not, but it had only 2654 retweets and 5428 likes. The classifier classified it wrongly and the caption was “This is Atticus. He’s quite simply America af…”. Looking at the dog picture though, I can see why it has such an exaggerated rating; maybe it was not a typo – perhaps the person rating the dog tweets that day was feeling a bit patriotic.