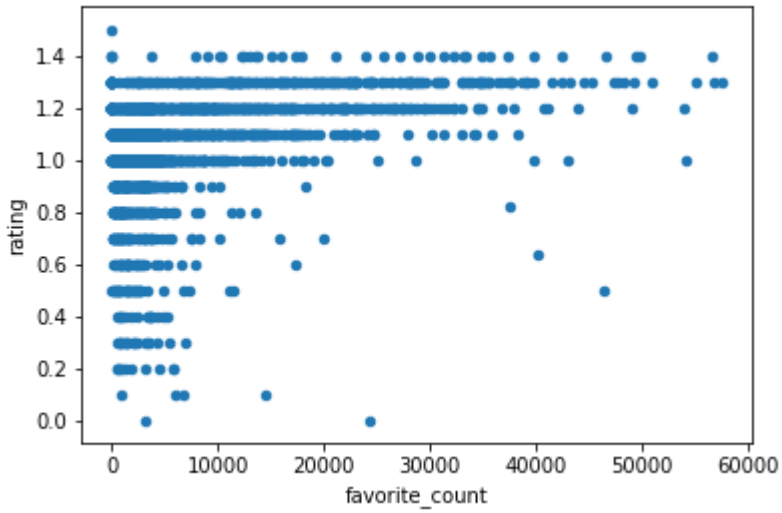
Project report – Summary of result

By Tu Duong

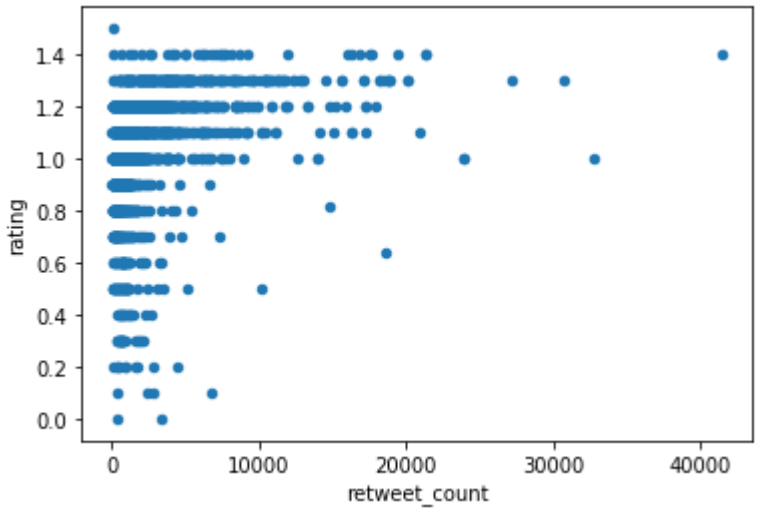
I wrangled, and explored, and prepared analysis for a dataset contains dog rating information from user WeRateDogs on Twitter. The main purpose of the analysis is to evaluate if there is any relationship among the variables that we collected. The dataset was prepared based on three sources of data: from a provided twitter archived, a text file from Internet contain image identification results, and from Twitter's API. Data was cleaned, merged, and saved as "twitter_archive_master.csv".

Comments:

- The dataset contains ~2000 tweets with ratings and pictures.
- The median of the rating is 1.057. That means on the scale of 10, there are many dogs were rated higher than 10.
- There are a variety in the breeds of the dogs. Total number of dog breeds is 375. Golden retrievers and labrador retrievers account for more than half of dogs rated on WeRateDogs.
- In general, the number of likes has positive correlation with dogs' rating. The higher the number of likes (favorites), the higher the dogs' rating. However, the relationship does not appear to be strong.



- Similarly, the number of retweets has positive correlation of the dog's rating. The higher the number of retweets, the higher the dogs' rating.



- Image reading prediction is not very accurate, which results in the incorrect names of the breeds such as Prison, Millitary Uniform, Electric Fan etc. Some dog breeds have higher median rating than others.

- Future work could go deeper in to exploratory data analysis. For example, looking closely at the distribution of the rating, reviewing the accuracy of the image prediction algorithm, and eventually building a model to predict the rating based on various factors such as likes, retweet, breed type.