Act Report

(Derrick Xu, 07 June 2020)

This write-up documents the insights and visualizations generated by analyzing the wrangled data in twitter_archive_master.csv. The dataframe is created by wrangling two WeRateDogs datasets and one neural network predictions dataset. Data wrangling steps are documented in wrangle_report.pdf.

Image 1: Example Post on WeRateDogs

The first insight is that the most common dog breeds on WeRateDogs are: Golden Retriever, Labrador Retriever, Pembroke Golden Retriever. From Table 1, Golden Retriever is the most common dog breed, and is featured on 155 separate tweets.

golden_retriever	155
Labrador_retriever	105
Pembroke	96

Table 1: Top 3 Most Common Dog Breeds on WeRateDogs

The second insight is that there are surprisingly large number of breeds. In fact, there are 1679 breeds from the wrangled data. This goes to show that there are far more breeds than is recognized by the Fédération Cynologique Internationale, and the American Kennel Club, which recognize 340 and 167 breeds respectively. Figure 1 shows a histogram created by plotting breeds vs number of dogs for that breed. There are typically <20 dogs per breed for a total number of 1679 dogs. This confirms the large variety of dog breeds in the wrangled data.

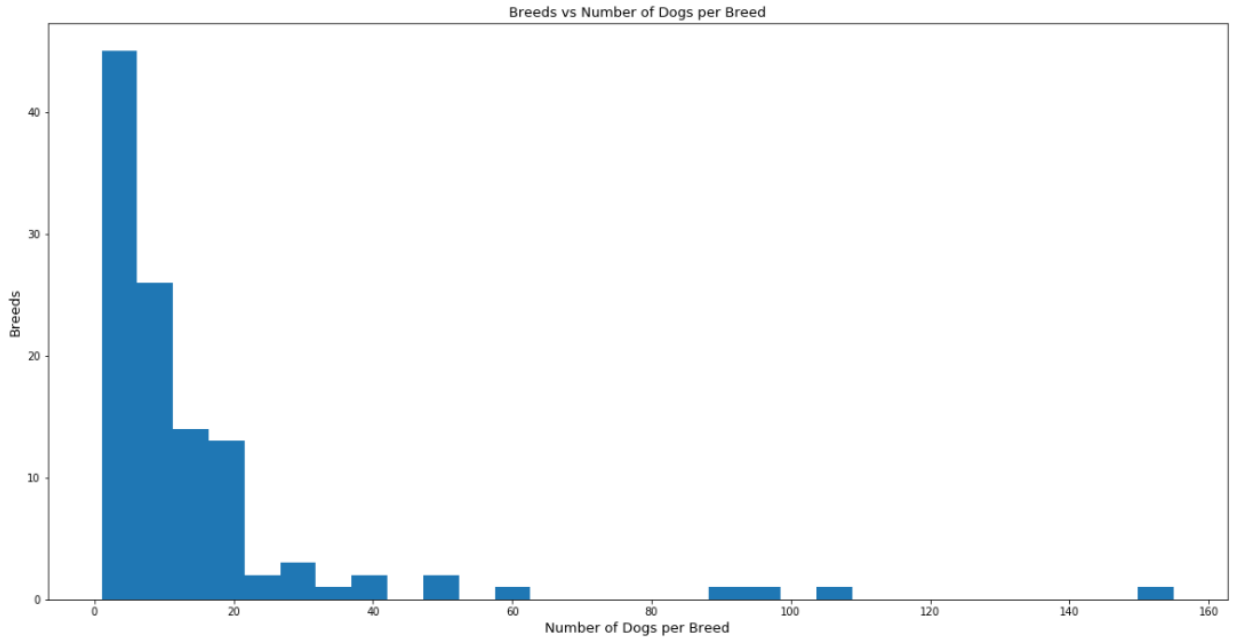

Figure 1: Histogram of Breed vs Number of Dogs per Breed

The third insight is that Bouvier_des_Flandres is the best rated dog, and Japanese_spaniel the worst rated dog. Bouvier_des_Flandres has an average rating of 13/10, while Japanese_spaniel has an average rating of 5/10.

The fourth insight is that favorite_count and retweet_count are very positively correlated (correlation = 0.93), but the two variables are only moderately correlated with rating_numerator (correlation = 0.54). In fact, tweet_id is slightly more correlated to favorite_count than rating_numerator (correlation = 0.60). This is tested for robustness using different methods for correlation (i.e. pearson, kendall and spearman). A hypothesis for the positive correlation between tweet_id and favorite_count is the increase in number of WeRateDogs users over time, since it is likely that the tweet_id is created in an increasing order over time.

	tweet_id	rating_numerator	rating_denominator	favorite_count	retweet_count
tweet_id	1.000000	0.544693	NaN	0.599247	0.372188
rating_numerator	0.544693	1.000000	NaN	0.408259	0.308682
rating_denominator	NaN	NaN	NaN	NaN	NaN
favorite_count	0.599247	0.408259	NaN	1.000000	0.930683
retweet_count	0.372188	0.308682	NaN	0.930683	1.000000

Table 2: Correlation Matrix (Method = Pearson)