act_report

July 14, 2020

1 WeRateDogs Twitter Archive Act Report

1.0.1 Summary

The current project investigates the data wrangling capabilities of Python and its libraries by investigating the WeRateDogs Twitter account archive. This contains basic tweet data for the 2,356 tweets containing ratings (out of their 5,000+ twitter count) as of August 1, 2017.

However, data in this archive is incomplete, and so is suplemented with additional data. In particular, retweet count and favourite count are missing from the archive, and these are gathered by querying Twitter's API via the particular tweet IDs from the archive. Simimlarly, neural network results classifying breeds of dogs were also included by programmatically downloading said image prediction data hosted in the Udacity servers.

Objectives

With this in mind, the project objectives revolved around wrangling the WeRateDogs Twitter data to generate clean and easily accessible data from which analyses and visualizations can be created to extract meaningful insights from the tweet metrics.

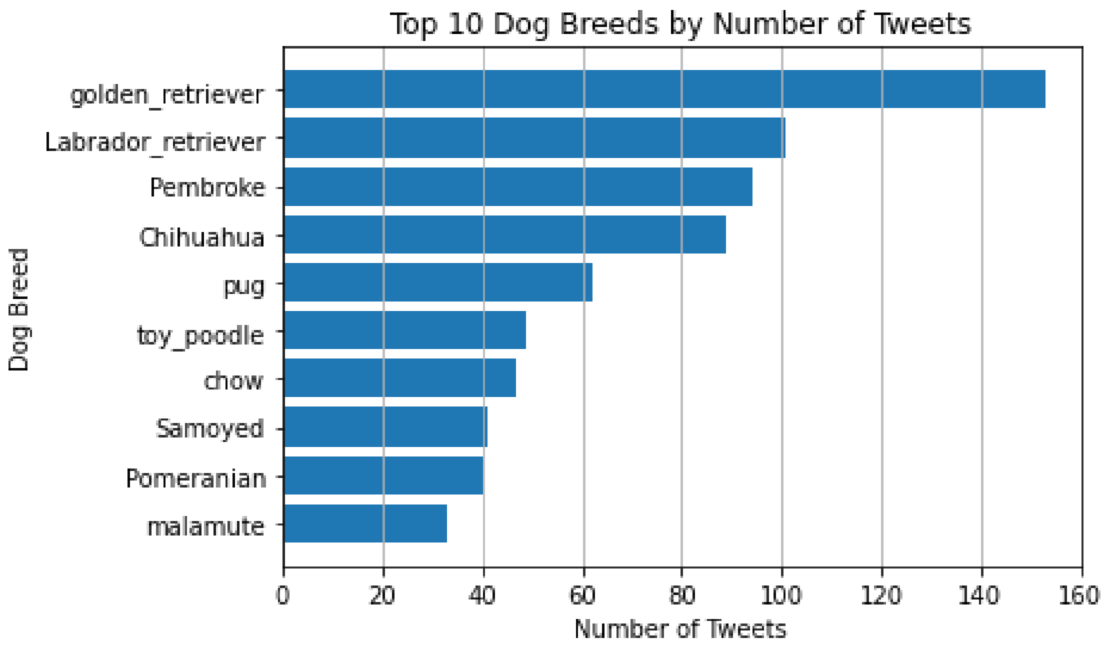
1.0.2 Data Wrangling

Fulfillment of the objectives requires wrangling data to a state in which it is easily accessible and meaningful. This involves gathering all necessary data, assessing it, and finally cleaning it. Visual and programmatic assessment was performed, focusing on four dimensions of clean data, namely completeness, validness, accuracy and consistency. Quality and tidiness issues were then addressed by making use of Python and its libraries, after which the clean dataset with all meaningful metrics merged was stored for further analysis.

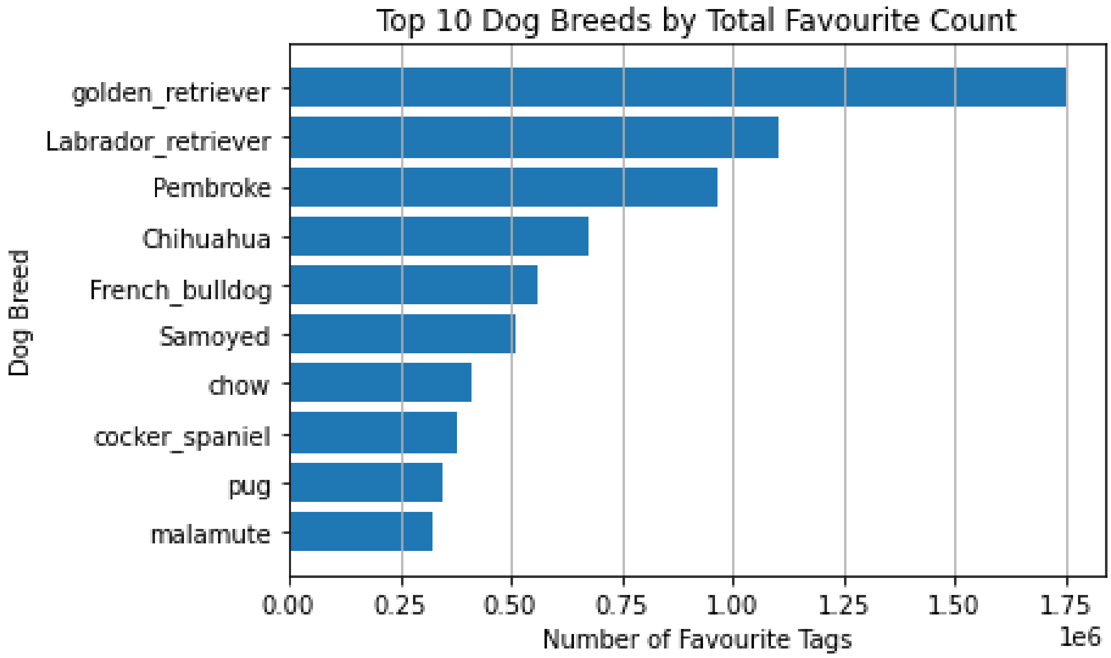
1.0.3 Data Analysis and Insights

Retriever Dog Breed is the Most Popular When grouping tweets by dog breeds based on the neural network image predictions, the results shown in the horizontal bar plots below were obtained. Interestingly, the top 4 dog breeds by total number of breeds coincide in exactly the same order of popularity with those by total number of favourite tags. Retrievers are evidently the most popular dog breed (Golden Retriever first and Labrador Retriever second), followed by Pembroke in third place. A total of 8 out of the top 10 dog breeds in both scenarios coincide.

from IPython.display import Image Image("images/top_breeds_by_number.png")

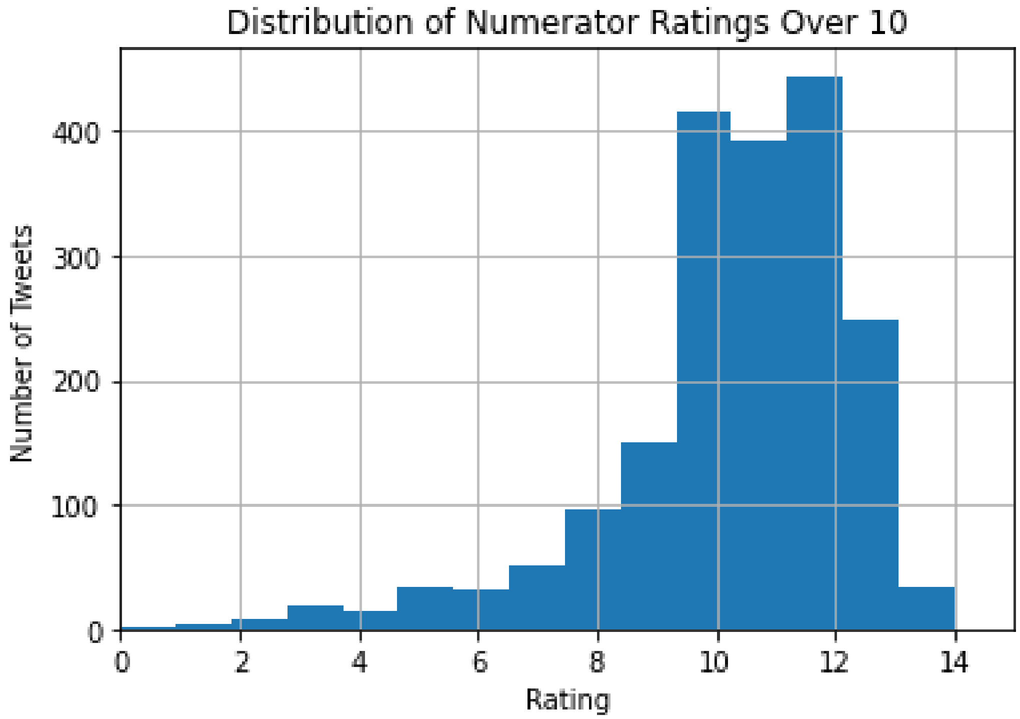


from IPython.display import Image Image("images/top_breeds_by_favourites.png")



Dog Ratings are Consistently High Data for numerator ratings is heavily skewed towards the left, with 75% of ratings exceeding a value of 10, and mean rating located at 10.5 (above the denominator standard value of 10).

from IPython.display import Image Image("images/rating_distribution.png")

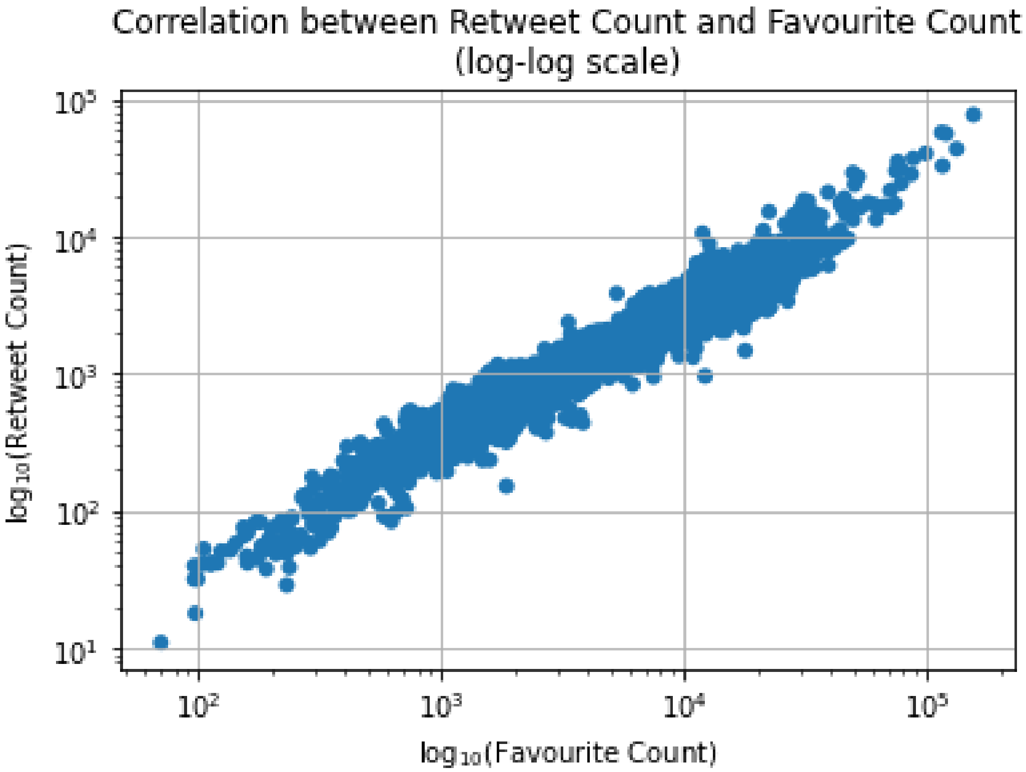


WeRateDogs Tweets Obtain More Favourite Tags than Retweets on Average Overall, both the mean and median values for favourite count are higher than that for retweet count. This indicates that, on average, people are more likely to favourite a WeRateDogs tweet than to retweet it.

Parameter	Mean	Median
Retweet Count	2446	1183
Favourite Count	8271	3729

Retweet Count and Favourite Count are Strongly Related Further from the insight that favourite tags are more likely than retweets, a strong positive correlation was identified between the retweet counts and favourite counts. As shown in the figure, the higher the favourite count, the higher the retweet count, with a Pearson's correlation factor of r=0.929.

from IPython.display import Image Image("images/retweet_vs_favourite.png")


Correlation between Retweet Count and Favourite Count
(log-log scale)
10^5
10^4
10^3
10^2
10^1
$\log_{10}$(Retweet Count)
10^2
10^3
10^4
10^5
$\log_{10}$(Favourite Count)

[ ]: