WeRateDogs – Insights into the @dog_rates Twitter page

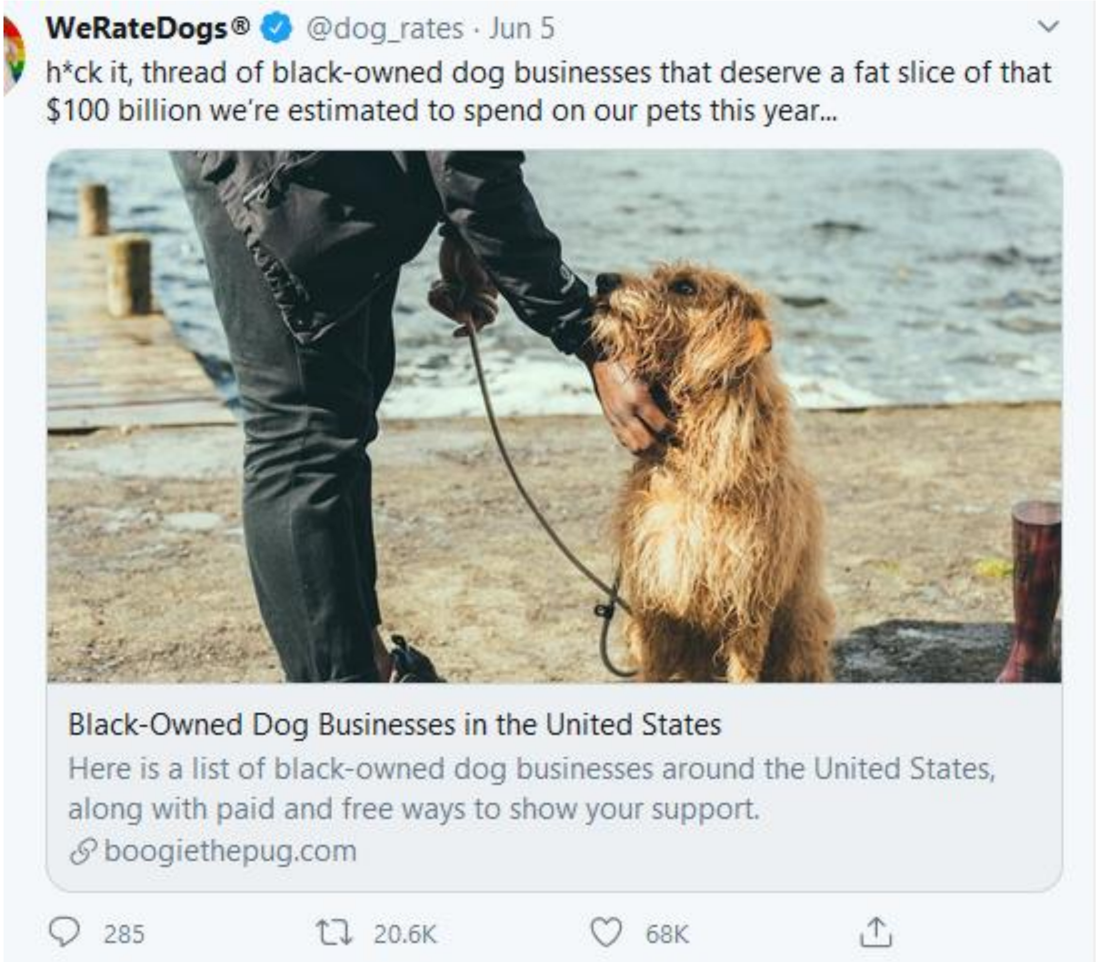
By Ayoade Olayiwola

Introduction and Background

I guess every analyst would love a clean data; real-world data however rarely comes clean. This means that some form of wrangling will almost always be needed

The dataset wrangled for this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter page with about 8.7million followers that regularly shares pictures of dogs along with a catchy description and often a rating typically out of 10 for the dog in the picture, sometimes exceeds 10.

Typical posts/tweets on @dog_rates are put below



WeRateDogs® @dog_rates · Jun 5

h*ck it, thread of black-owned dog businesses that deserve a fat slice of that $100 billion we're estimated to spend on our pets this year...

Black-Owned Dog Businesses in the United States

Here is a list of black-owned dog businesses around the United States, along with paid and free ways to show your support.

boogiethepug.com

285 20.6K 68K

WeRateDogs® @dog_rates · Jun 12

This is Monk. His human founded the Minnesota Freedom Fund, which has received so much support recently they've pawsed donations. Monk says thank you and would like to redirect your attention to other orgs below keeping targeted communities safe. 13/10

docs.google.com/document/d/1yL...

Minnesota Freedom Fund and Reclaim the Block

This project works through the data wrangling process, focusing on the gathering, assessing, and cleaning of data. There are visualizations and observations from the analysis provided as well.

Gather

This project gathered data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers.
- The Udacity-provided tweet-json.txt

Assess and Cleaning the Data

Assessing data requires data analysts to evaluate a data set on quality and tidiness issues. This resulted in the following quality and tidiness issues being fixed in the project

- Removal of 181 Retweets

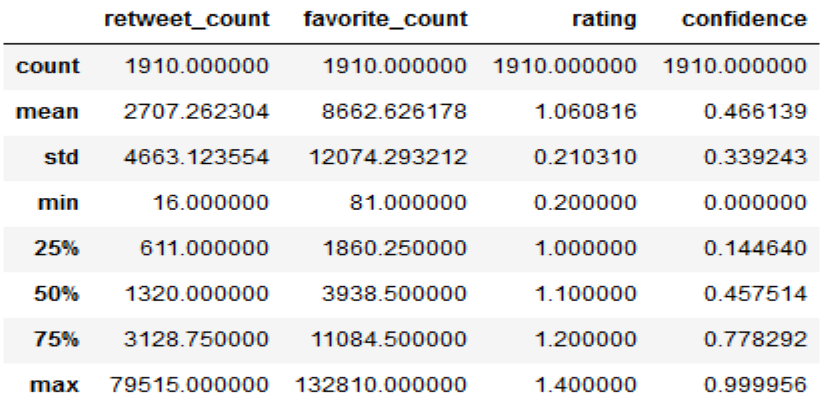
- Fixing of missing data in expanded_urls (Tweets without images)
- Addressing the fact that not all images are dog images
- Fixing incorrect ratings and as well incorrect dog names
- Fixing missing values in dog names (represented as None)
- Fixing erroneous datatype (tweet_id, timestamp)
- Merging three data frames
- Dropping unneeded columns.

The issues where each defined and fixed programmatically

Analysis and Visualization

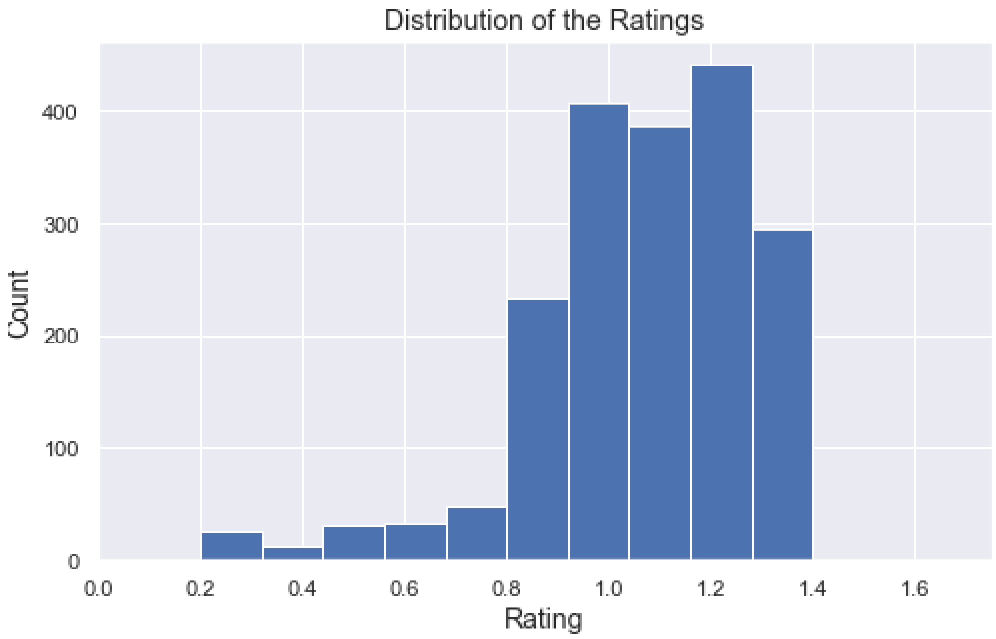
The following analysis with attendant visualization was performed on the date

1}. Descriptive Analysis and the distribution of the ratings

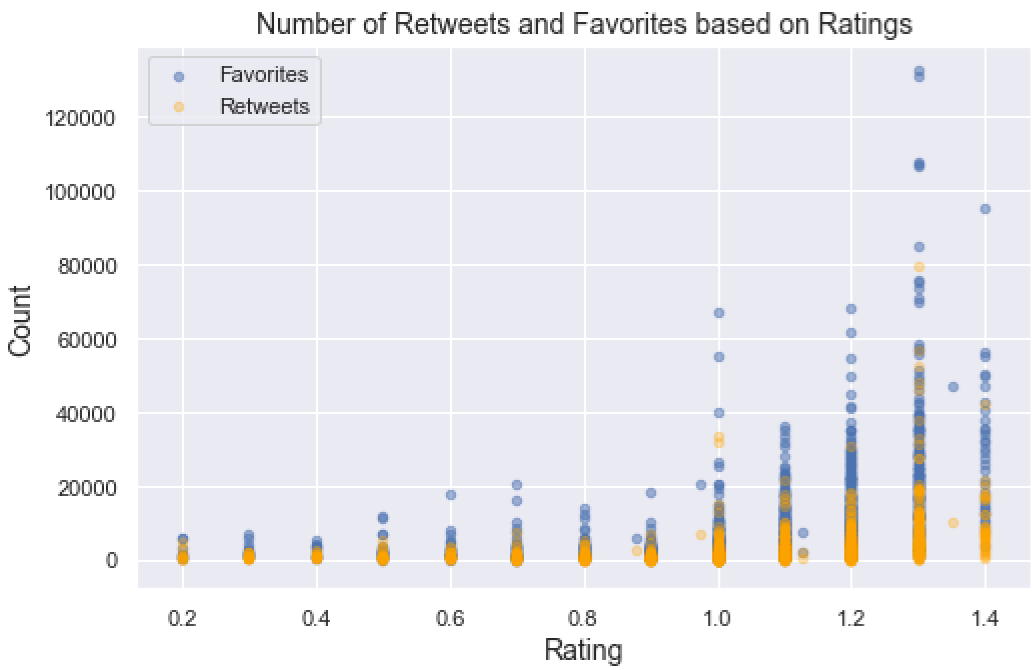


	retweet_count	favorite_count	rating	confidence
count	1910.000000	1910.000000	1910.000000	1910.000000
mean	2707.262304	8662.626178	1.060816	0.466139
std	4663.123554	12074.293212	0.210310	0.339243
min	16.000000	81.000000	0.200000	0.000000
25%	611.000000	1860.250000	1.000000	0.144640
50%	1320.000000	3938.500000	1.100000	0.457514
75%	3128.750000	11084.500000	1.200000	0.778292
max	79515.000000	132810.000000	1.400000	0.999956

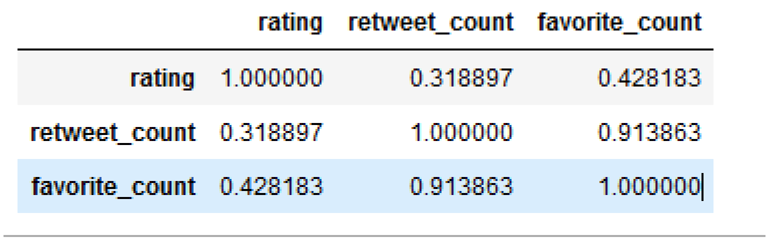
The mean dog rating is 1.06 and the ratings are more frequent between 1 and 1.3. The most frequent rating is 1.2



2}. Relationship between favourite_count, retweet_count and rating



Because retweets and favorites may be assumed to be positively correlated and also that retweets and favorites increase with ratings, I analyzed the relationship between these variables. The correlation matrix is below



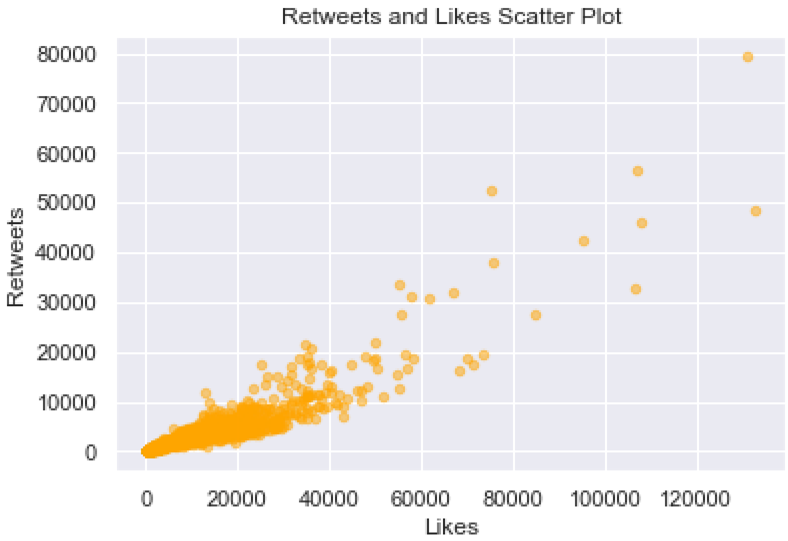
	rating	retweet_count	favorite_count
rating	1.000000	0.318897	0.428183
retweet_count	0.318897	1.000000	0.913863
favorite_count	0.428183	0.913863	1.000000

There is a strong positive correlation between retweet counts and favorite counts (0.91). The correlation coefficent between rating and retweet counts is rather too weak. It seems to suggests that people's retweets are largely independent of the rating; probably the picture or some other factors triggers influence retweet

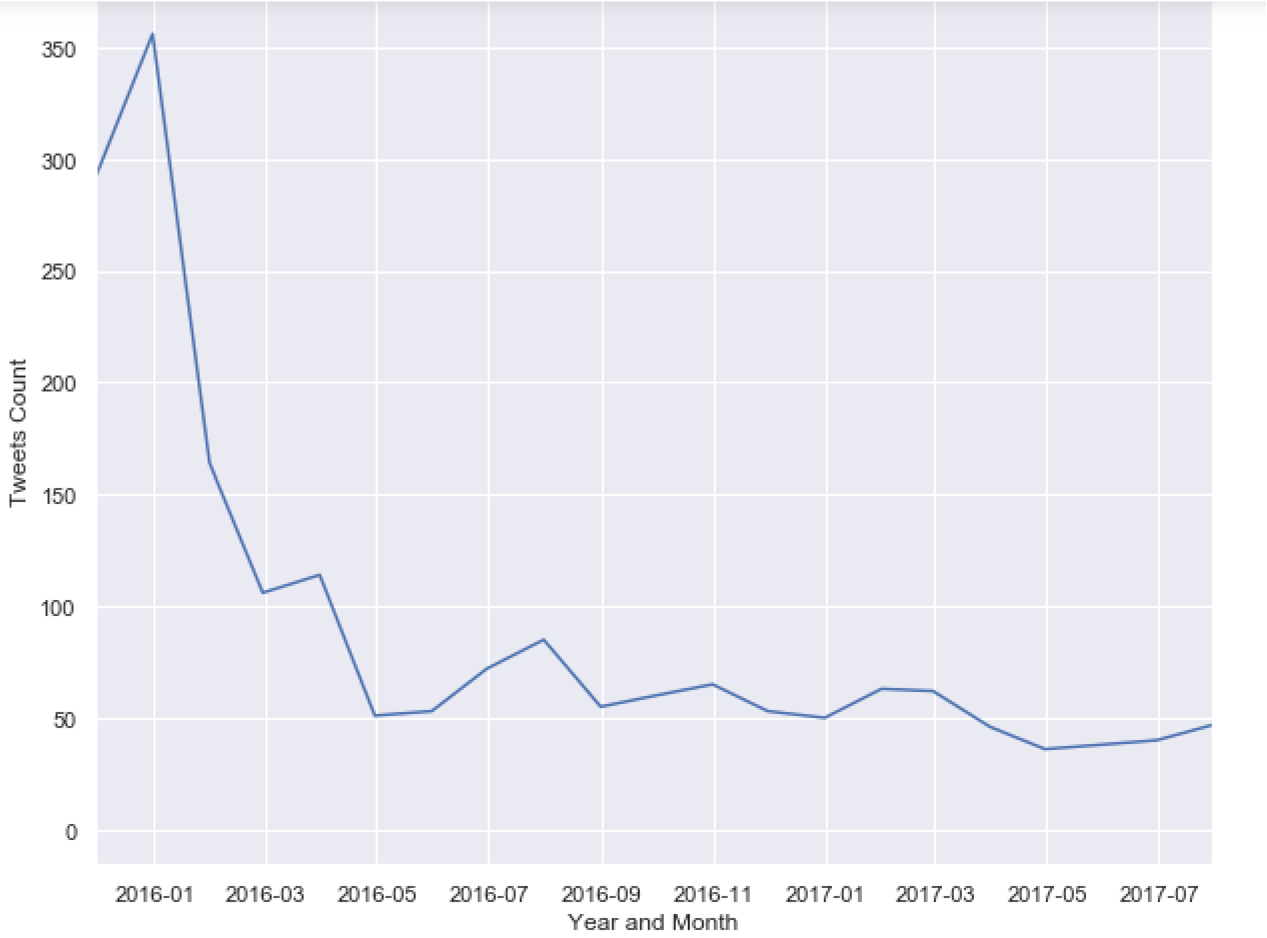
3}. The relationship between likes and dislikes

It is expected that the relationship between these two variables will be positive. So a scatter plot was plotted to examine the nature of this relationship in the tweeter archive data.

Retweets are positively correlated with Likes.

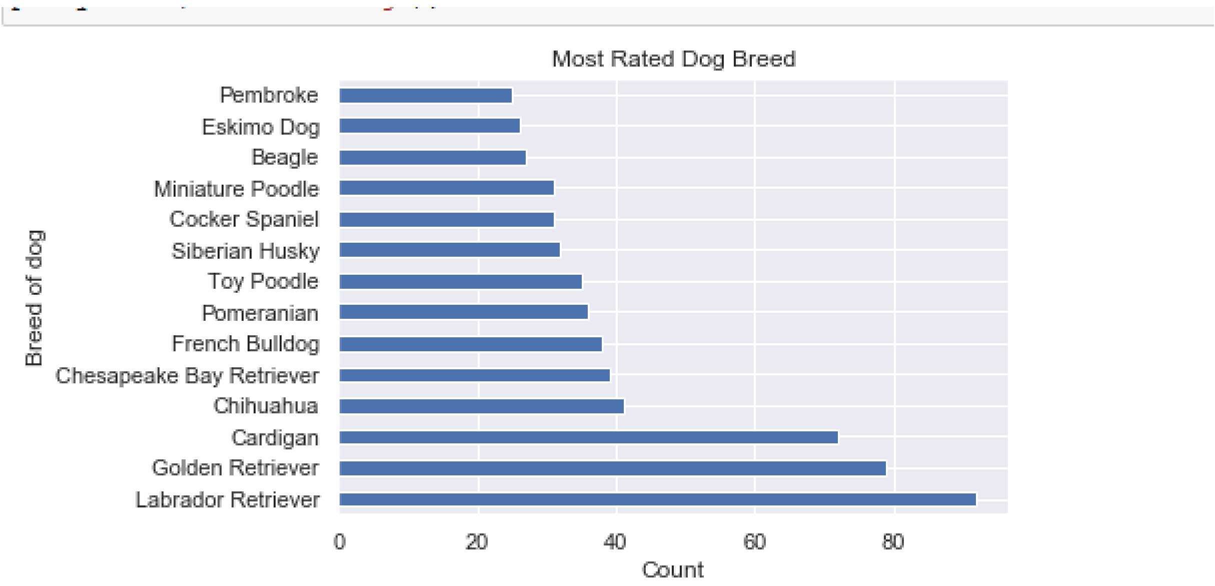


4}. The trend of tweets over time



This trend shows that over time tweets decreased sharply, with the peak in early 2016, spikes in activity during the early spring of 2016, mid-summer of 2016, and generally decreasing from there. The reason for this sharp decline cannot be deduced from the dataset.

5}. Most Rated Dog Breed



The most popular dog breed is a labrador retriever, followed by golden retriever. This analysis does not factor in the None label column

Conclusion

This write-up offers a straightforward look at the data wrangling process. Much more can still be done on and with the dataset. I will definitely still work on this outside of the projects; I encourage others to do the same too.