

WeRateDogs – Insights into the @dog_rates Twitter page

May 25, 2020

Udacity – Data Analyst Nanodegree

Saikiran Bikumalla

Project- 4 (Wrangle and Analyse Data)

Hyderabad, India

Introduction :

Real-world data rarely comes clean. The dataset wrangled for this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Here's an example:



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 involved gathering data from three different sources:

1. The WeRateDogs Twitter archive. The `twitter_archive_enhanced.csv` file was provided to Udacity students which can be downloaded directly.
This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the [Requests](#) library and the following URL:
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Using Twitter API (which is authorised) and extracting the contents from twitter API such as tweet_id , retweet count and favorites count using python tweepy library and storing it in tweet_json.txt file.

Assess:

Assess data involves the evaluation and assessment of the dataset to get to know the quality and tidiness issues of the dataset.

Quality Issues : The four main data quality dimensions are:

- Completeness: missing data
- Validity: does the data make sense?
- Accuracy: inaccurate data? (wrong data can still show up as valid)
- Consistency: standardization?

Tidiness Issues : Three requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

Clean:

Cleaning data is tedious, and often iterative. Just when an analyst believes they have found all quality and tidiness issues, there are often additional issues that arise. The cleaning process involves three steps:

1. Define: determine exactly what needs to be cleaned, and how
2. Code: programmatically clean the code
3. Test: evaluate the code to ensure the data set was cleaned properly

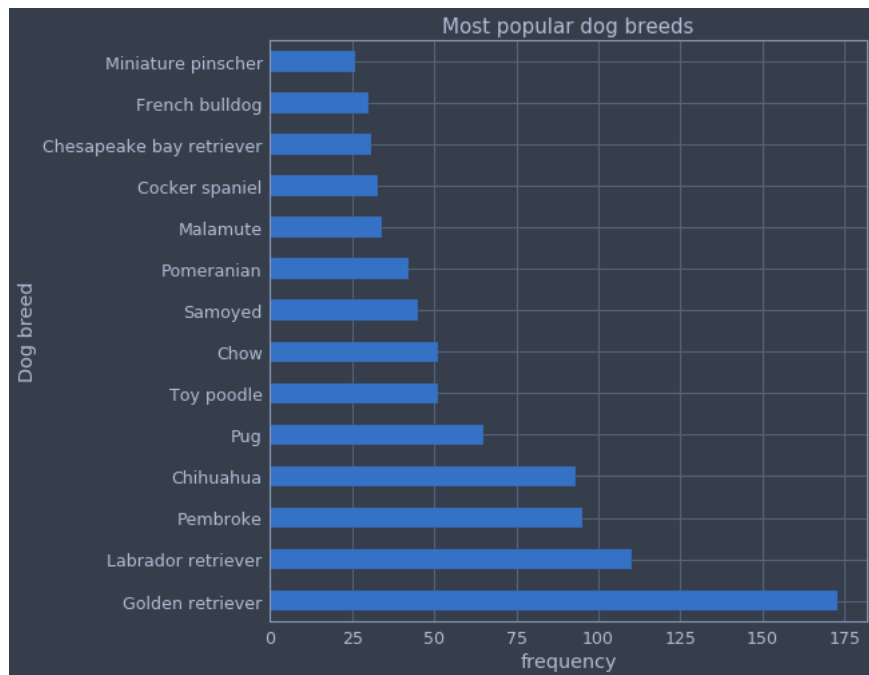
Storing , Analysis and Visualization:

After the dataset is cleaned properly, it is stored in a csv file named twitter_archive_master.csv

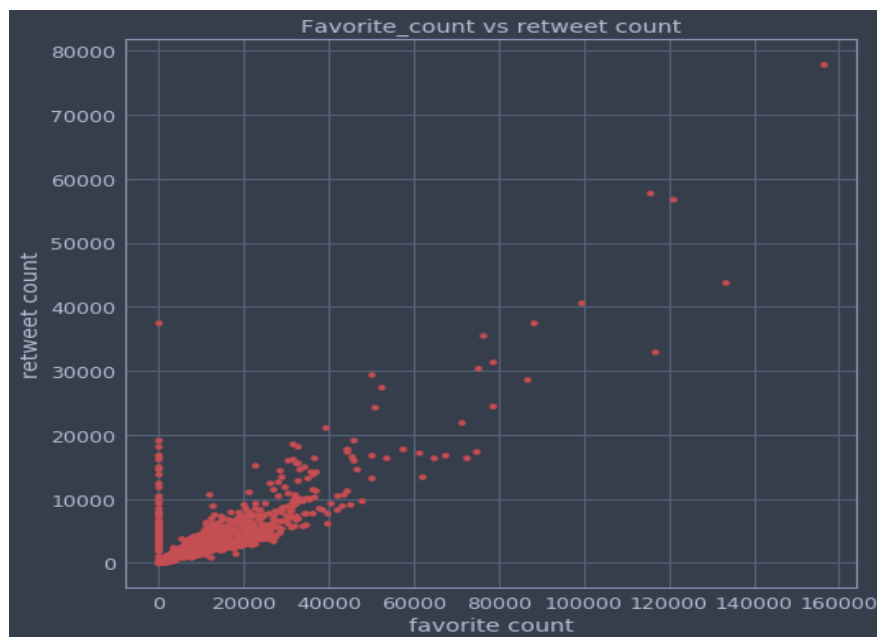
I have analysed the following five insights on this data:

Popularity of Dog Breeds:

The most popular dog breed is Golden retriever, with Labrador retriever coming in the second most popular dog breed. Pembroke and chihuahua aren't far behind. The page owner could use these insights to provide more posts on these, because they are proven to be popular to drive more traffic.



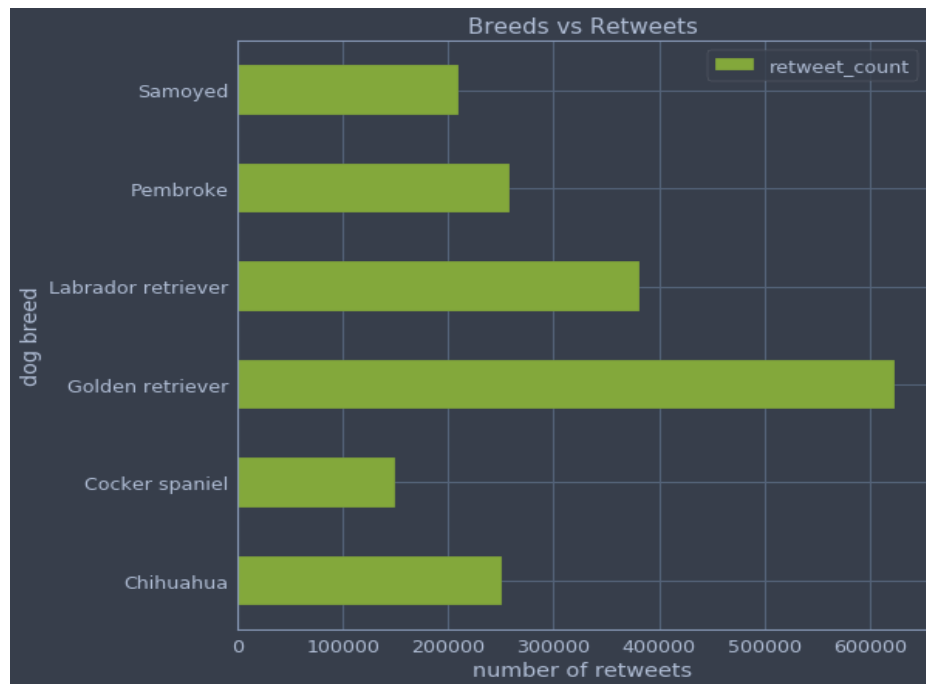
Relationship between favorites and retweets:



There is a positive relationship between favorite count and retweets count with a correlation coefficient of 0.86 which means a strong positive correlation exists. This insight could be used by the owners and data analysts to popular posts.

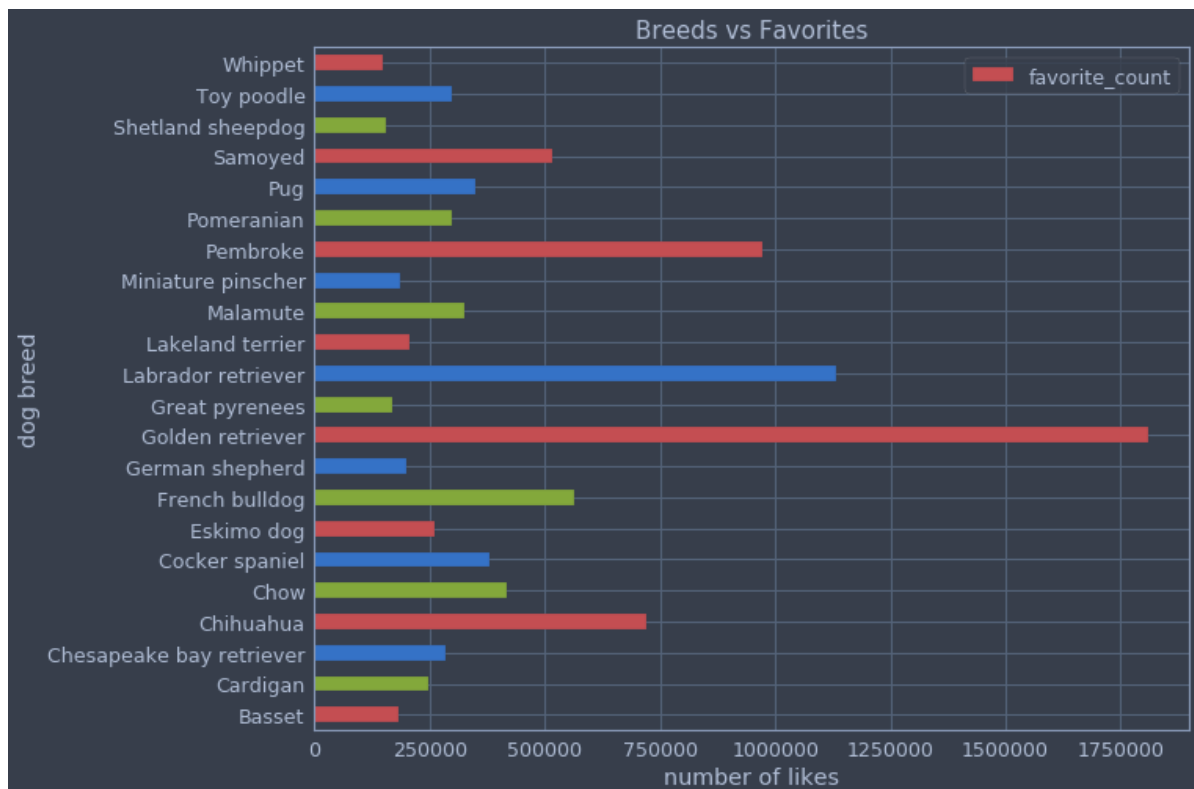
Dog breeds with most number of retweets:

The most retweeted dog breed is Golden retriever, with Labrador retriever the next most and pembroke , Chihuahua and samoyed aren't far behind. The page owner could use this insight to provide with the above breeds to get more retweets and drive more traffic to the page.



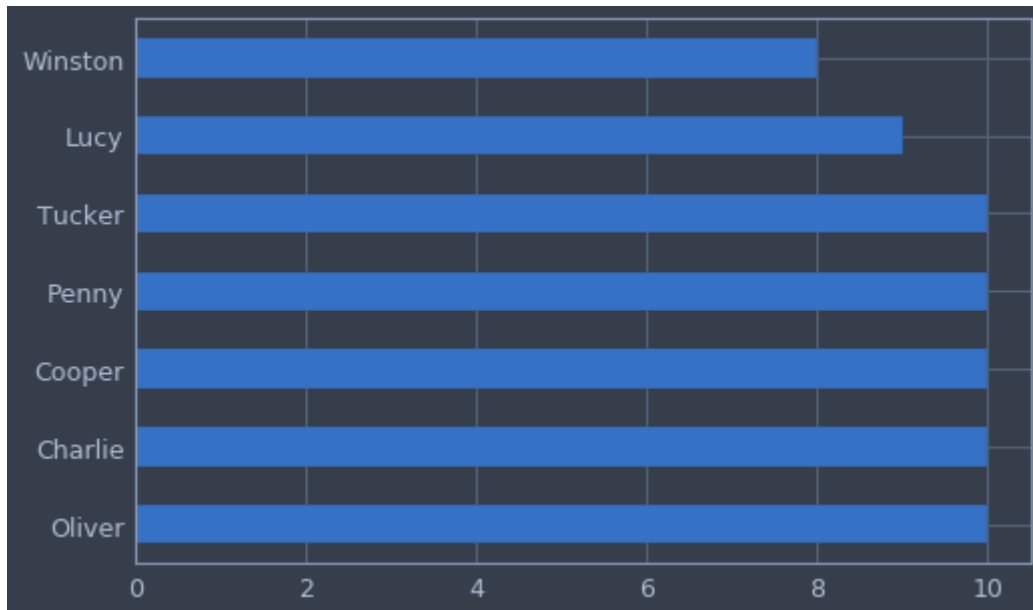
Dog breeds with most number of favorites:

The most favorited (liked) dog breed is Golden retriever, with Labrador retriever the next most and pembroke, Chihuahua and samoyed aren't far behind. The page owner could use this insight to provide with the above breeds to get more retweets and drive more traffic to the page. It can be observed that the same breeds had most retweets and the same have most likes. This strengthens our insight regarding strong positive correlation between favorite and retweet counts.



Dog name commonality:

Names are important, especially for dogs. They sound sweet and good for them. The most popular dog names in this dataset are Oliver, Cooper, Charlie , Tucker and Penny for a tie of 10 each.



Conclusions:

This write up provides a straight forward look for the data wrangling process essentials and data analysis and visualisation insights. There are so many more insights that can be assessed and analysed from this dataset, I highly encourage to deep dive into this data set to get what else we can find from it!