DATA WRANGLING

MULUNEH ABRHAM

Introduction:

The dataset wangling in this project is the tweet archive of twitter user @dog_retes also known as WeRateDogs is a twitter account rates people's dog with humorous comment an out the dog.

Gather:

This project gathered data from the following sources:

1. The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file. This archive contains basic tweet data (tweet ID, timestamp, text, ...) 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 another object, animal, etc.) is present in each tweet according to a neural network.
3. The tweeter json text file downloaded from the Udaity class room

Assessing Data:

After gathering the information, I started to evaluate it for both quality and neatness

This are the main issue in quality dimensions:

1. Completeness: Missing data
2. Validity: Does the data make sense
3. Accuracy: Inaccurate data
4. Consistency: Standardization

There are three main requirements for tidiness:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observation unit forms a table

Clean:

Data cleaning is laborious and frequently iterative. When data analysts think they have solved all quality and organization problems, they frequently discover new ones. The cleaning process involves three steps:

1. **Define:** Determine exactly what needs to be clean and how.
2. **Code:** Programmatically clean the code

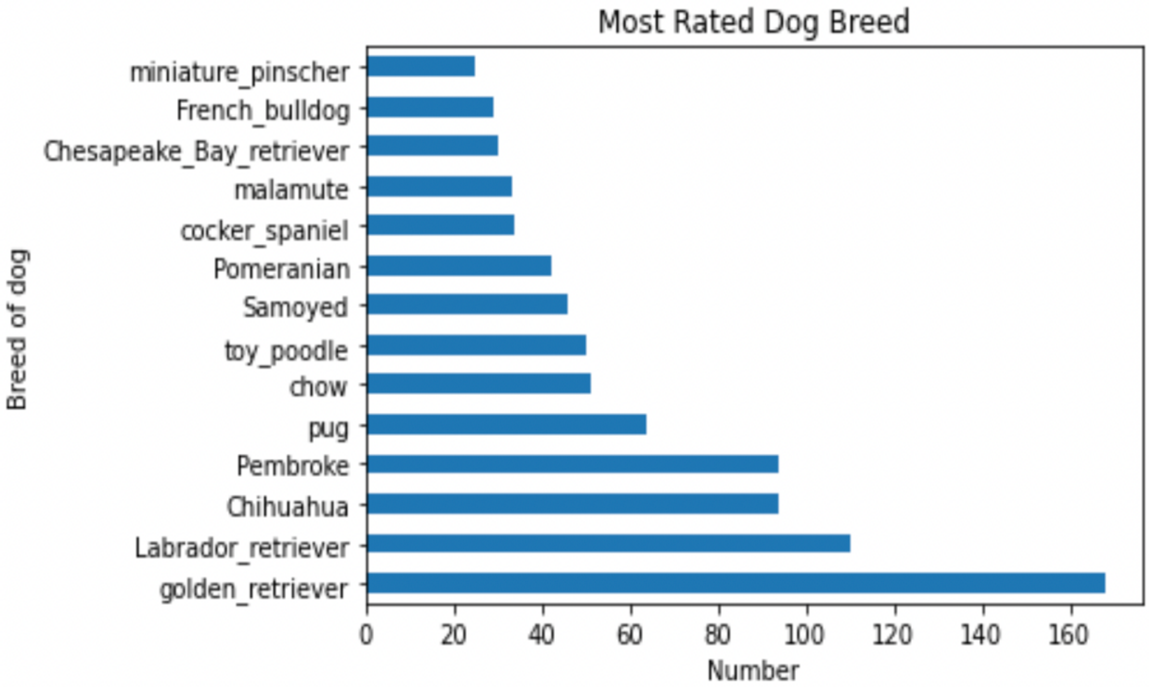
3. **Test:** Evaluate the code to ensure the data set was cleaned properly

Analysis and Visualization:

There is serval analysis, which I have done and those are in following:

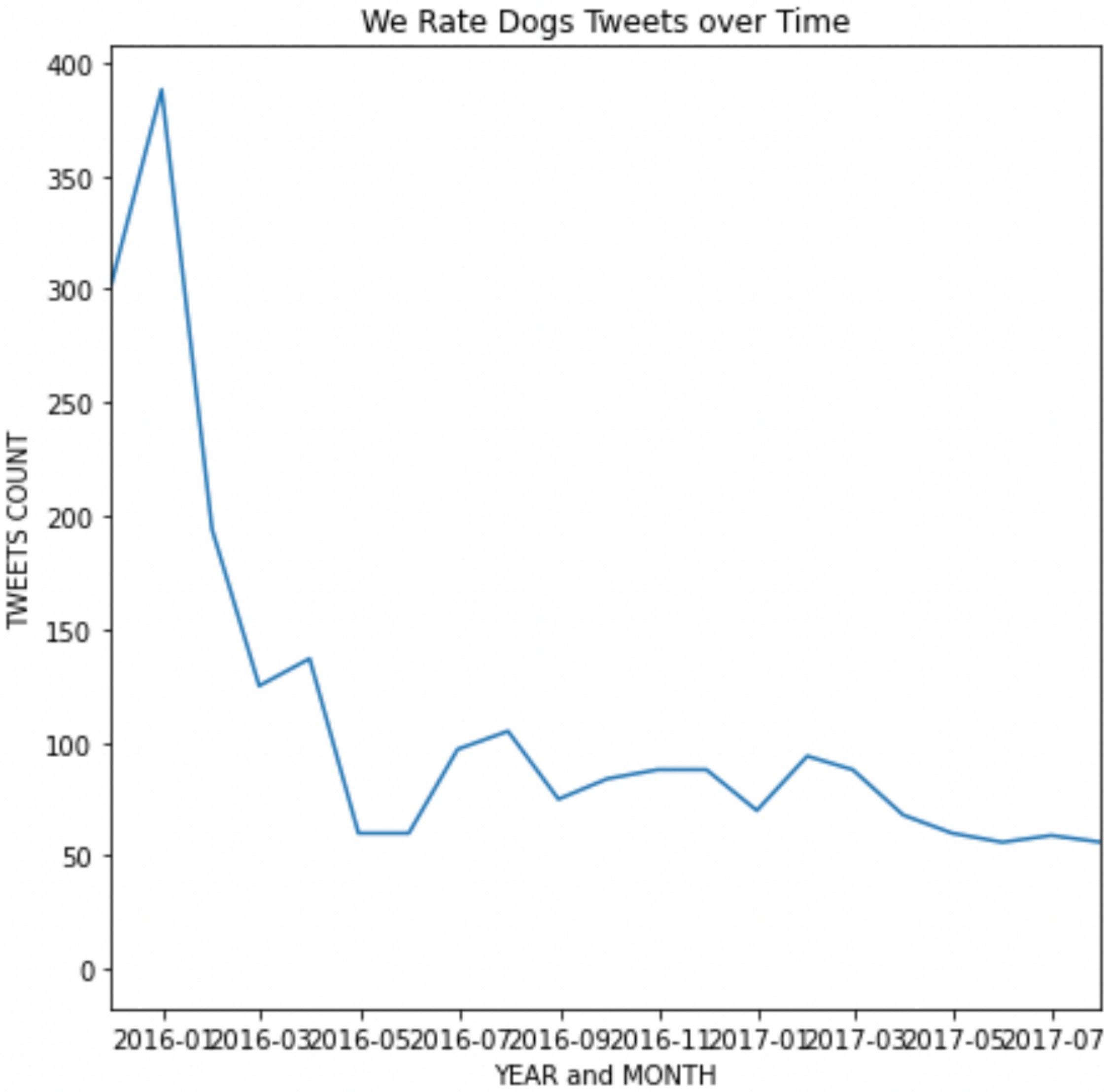
- Dog Breed Popularity:

Golden retrievers are the most popular dog breed (ignoring the none classification), followed by Labrador retrievers as the second most popular breed. Chihuahua is not so far away. The owner of the page might use this data to develop targeted marketing campaigns for particular breeds that aren't popular in order to boost their popularity while also utilizing breeds that have been shown to be popular in order to enhance user traffic to the page.



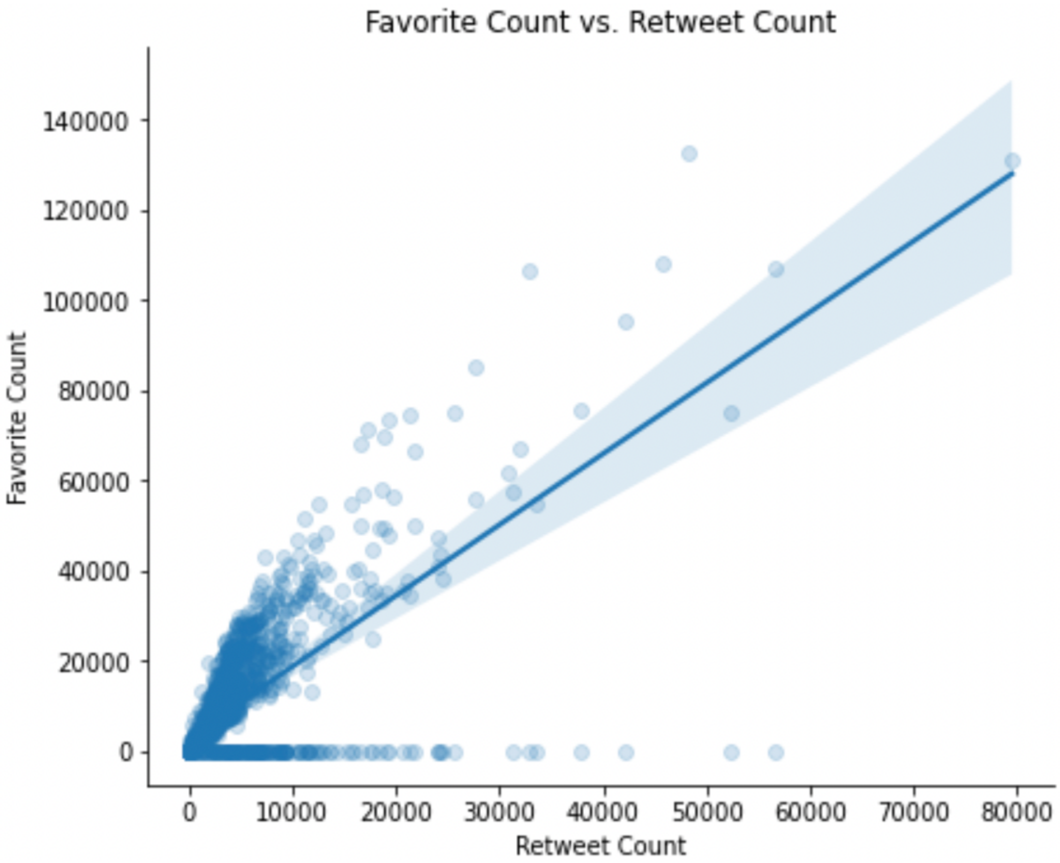
- Tweets Over Time:

Tweets dropped significantly over the course of the tweets gathered for this dataset, starting in early 2016. (i.e. is 2016-01). While the number of tweets has usually declined over time, there have been peaks in activity in early 2016 (2016-01) and mid-summer 2016 (i.e., between 2016-03 and 2016-05). The owner of the WeRateDogs Twitter account ought to be aware of this development and think of ways to drive more users to the Page.



- Favorite vs Retweet Counts:

Favorite ("like") counts and the volume of retweets for a post are positively correlated. When deciding how to improve users' traffic to the page, the owner of the WeRateDogs twitter account must take into account this correlation. A data analysis team might suggest old posts with lots of favorites or retweets so the page owner can model new posts after previously well-liked ones.



Conclusion:

The write up offers a straight look at the data wrangling process. There is so much more that can be done with this data set.