16-06-2020

Report

On

Analysis and Insights into Final Data

By:

Abhishek Tiwari

Introduction:

Real-world data rarely comes clean. The dataset that you will be wrangling (and analyzing and visualizing) 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.

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

Gather:

This project gathered data from the following sources:

- The WeRateDogs Twitter archive. The 'twitter-archive-enhanced-2.csv' file was provided to Udacity Students (Like me).
  - ➢ 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 prediction, i.e., what breed dogs (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students (Like me).
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favourite ("like") count at minimum and any additional data I find interesting.

Assessing Data:

Once the data was gathered, I began to assess the data on both quality and tidiness issue:

There are four main issue in quality dimensions:

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

And 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:

Cleaning data is tedious and often iterative. Just when data analyst believe they found all quality and tidiness issue, they often found additional issue arises. The cleaning process involves three steps:

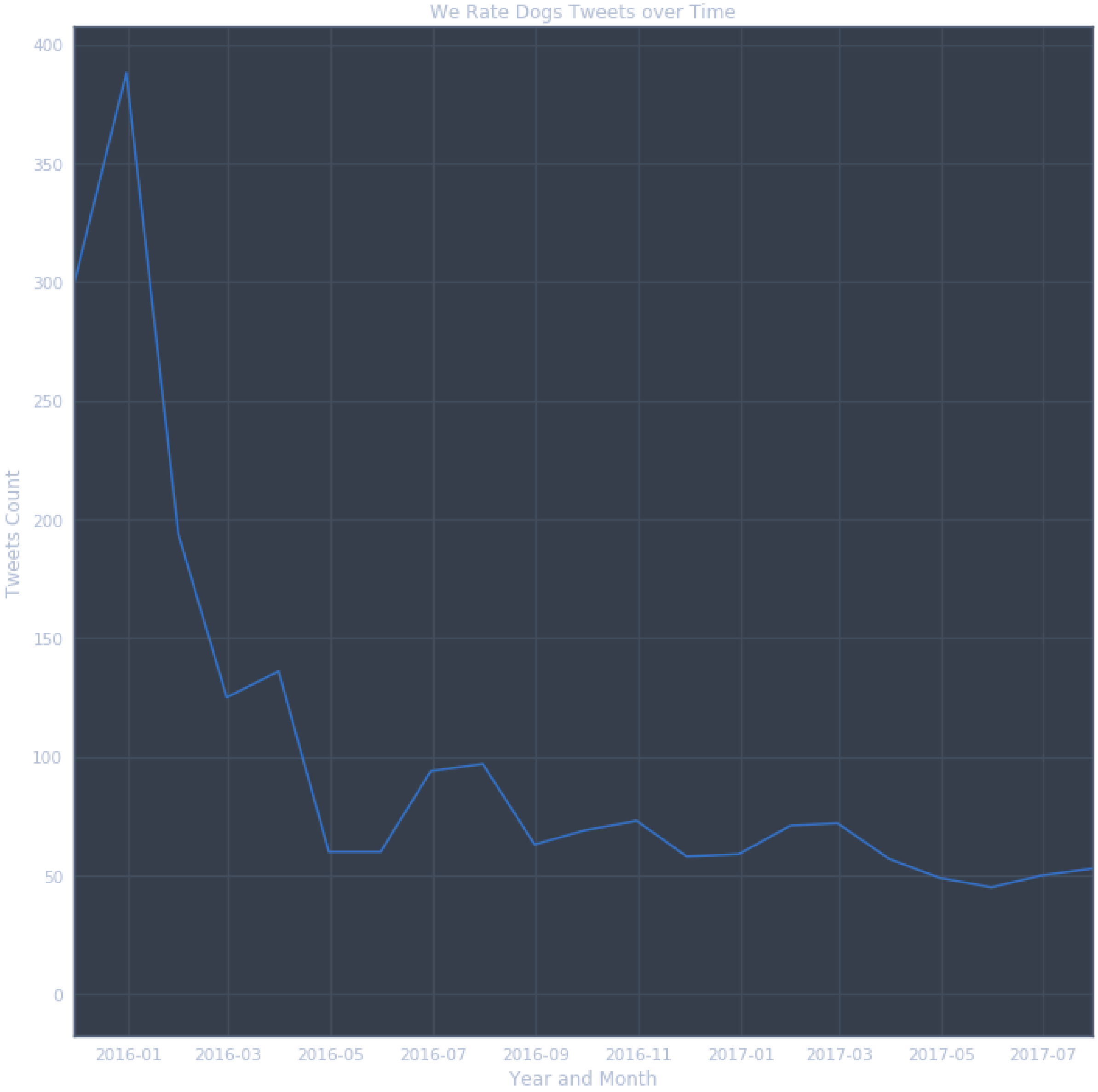
1. **Define:** Determine exactly what needs to be clean and how.
2. **Code:** Programmatically clean the code
3. **Test:** Evaluate the code to ensure the data set was cleaned properly.

Analysis and Visualization:

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

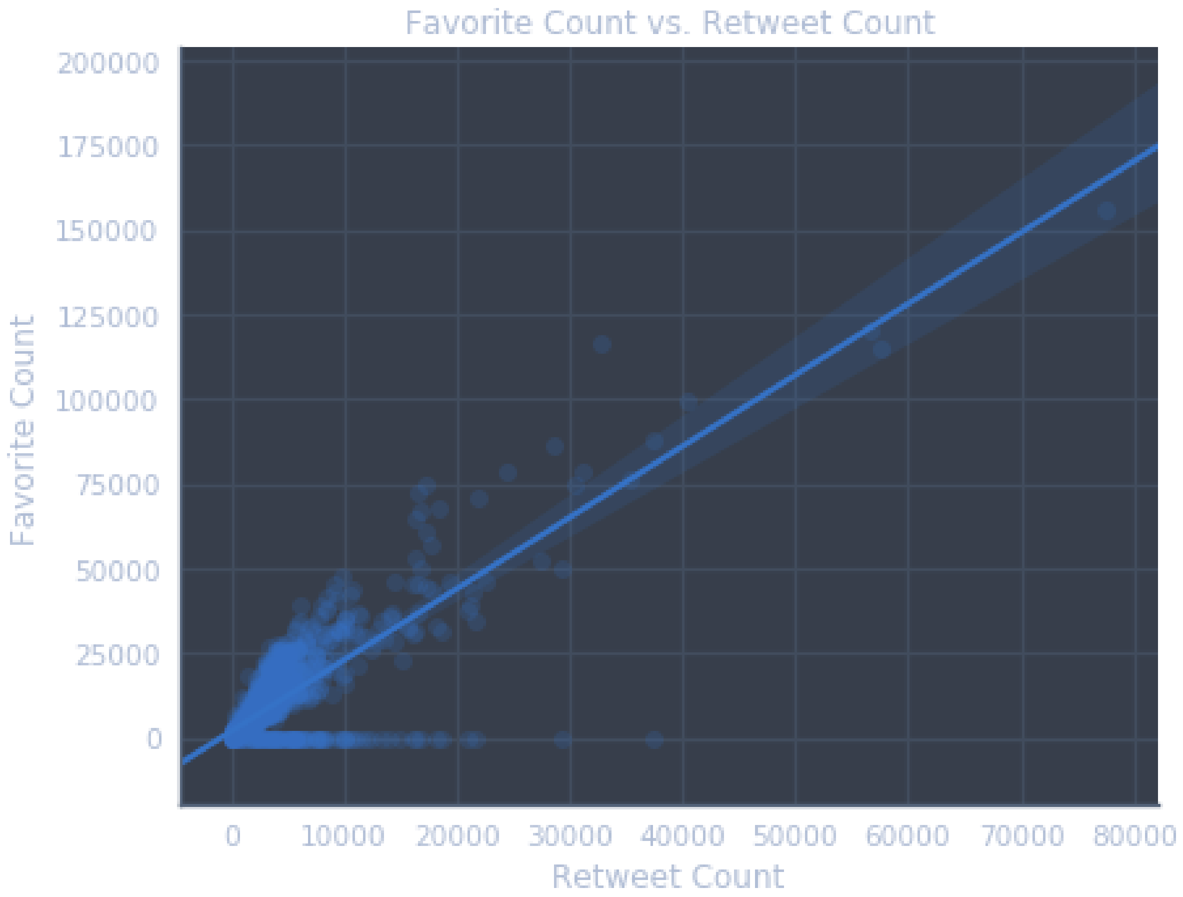
- **Tweets Over Time:**

  Over the time period of the tweets collected for this dataset, tweets decreased sharply starting in early 2016 (i.e. is 2016-01). While the tweets continue to decline over the time, there are spikes in activity during early 2016 (i.e. 2016-01) and in mid-summer of 2016 (i.e. in between 2016-03 to 2016-05), but continues to generally decreased from there. The owner of the WeRateDogs Twitter account should be aware of this trend and consider way to increase users' traffic on the Page.



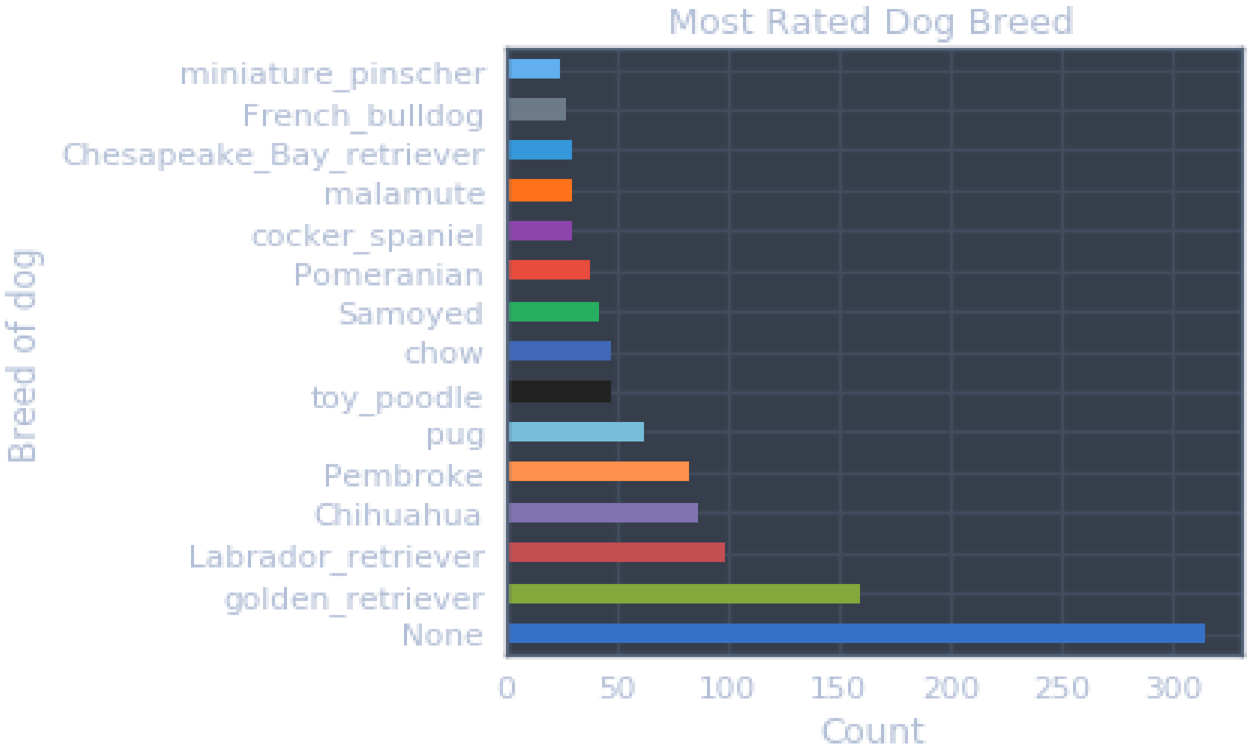
- **Favorite vs Retweet Counts:**

  There is a positive correlation between favorite ("like") counts, and how much a post was retweeted. This correlation is important for the owner of the WeRateDogs twitter account to understand when determining method to increase users' traffic on the page. A data analysis team could recommend previous posts with either a high retweet counts or high favorite count so that page owner could model future posts off historically popular posts.



- **Dog Breed Popularity:**

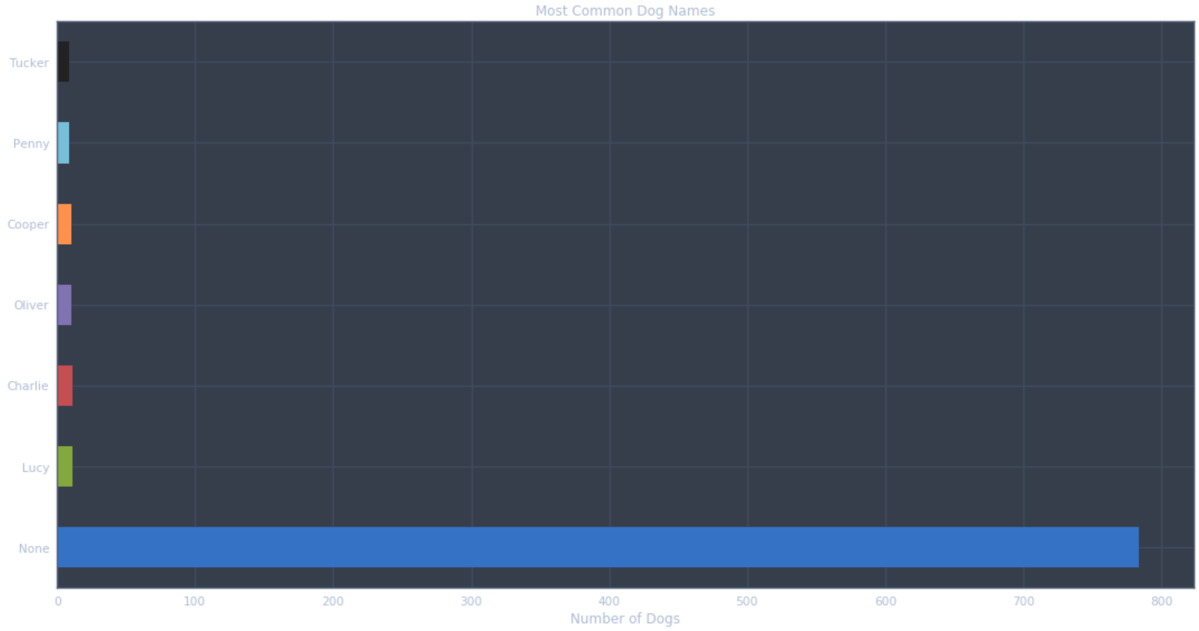
The most popular dog breed is golden retriever (ignoring the none label), with a Labrador retriever coming in as the second most popular breed. Chihuahua isn't far bind. The page owner could use this information to create targeted marketing efforts for certain breed that aren't popular to increase their popularity, but also utilize the breed that are proven to be popular to drive user traffic to the page.



- **Dog Name Commonality:**

Names are important, especially for Dogs. The First four dog names (ignoring the None Label) are:

1. Lucy
2. Charlie
3. Oliver
4. Cooper



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.