14-09-2022

Report

On

Analysis and Insights into the Final Dog Data

By: Calvin Baraka

Introduction:

The good thing about real world raw data is it rarely comes while clean almost ever. I mean isn't that where all the fun lies haha. The dataset that we are looking at is the twitter archive of Twitter user @dog_rates, also known as WeRateDogs. This is a twitter account rates people's dogs
This project works through the data wrangling process, focusing on the gathering, assessing and cleaningof 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:

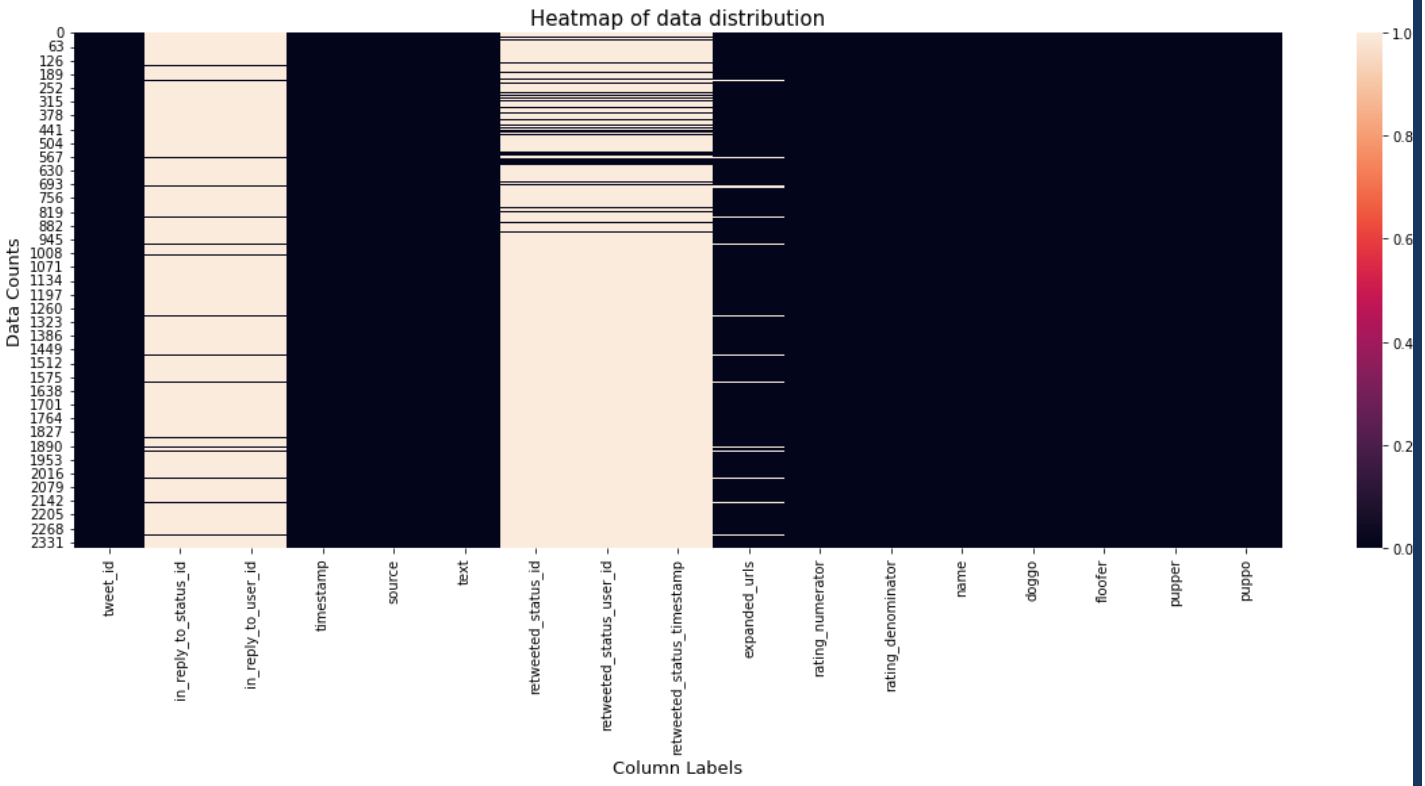
There are four main issue in quality dimensions:

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

In here we looked at a heatmap of data distribution to see how severe the missing data issue was:



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.

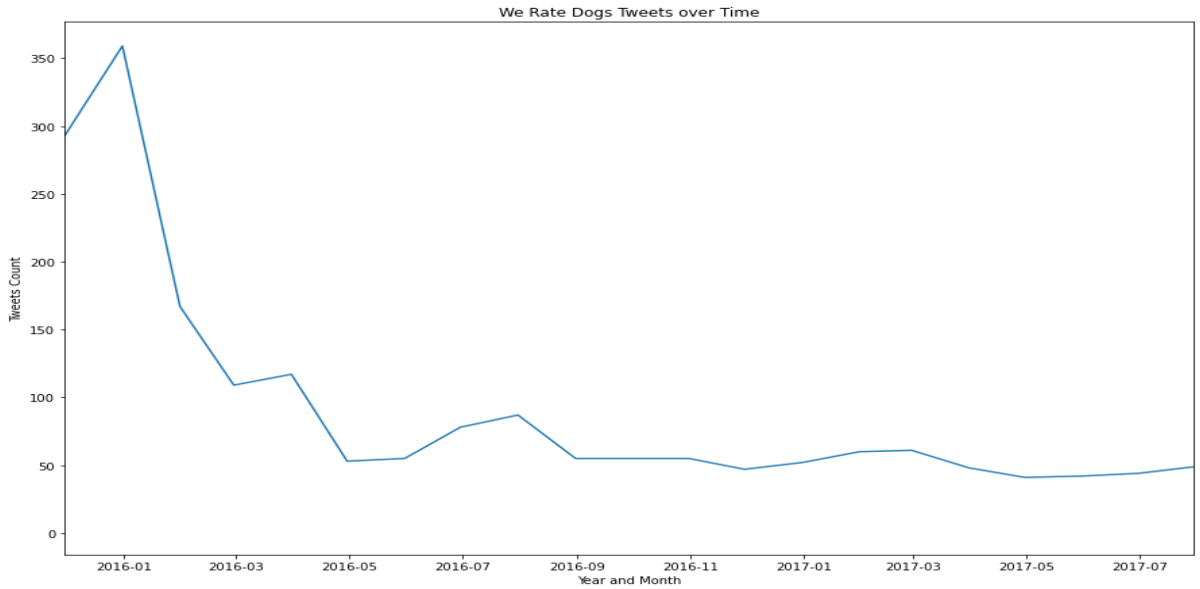
Analysis and Visualization:

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

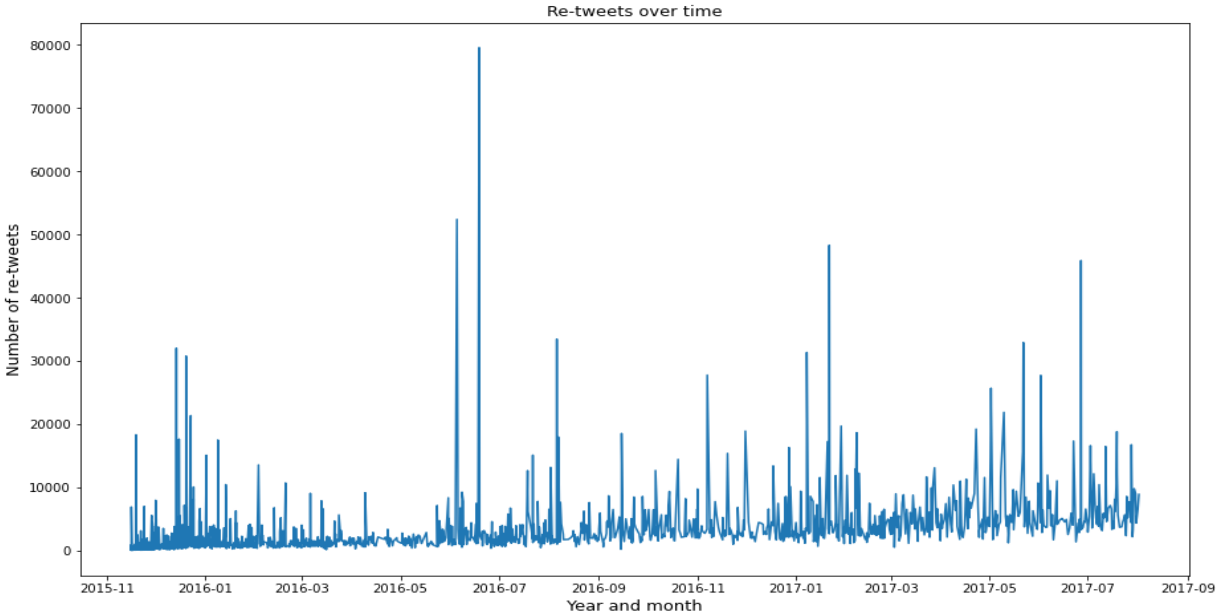
- **Tweets Over Time:**

  Over the time period of the tweets collected for this dataset, tweets decreased sharply starting in early 2016.

  Although the tweets continue to decrease over timet, here arespikes 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..



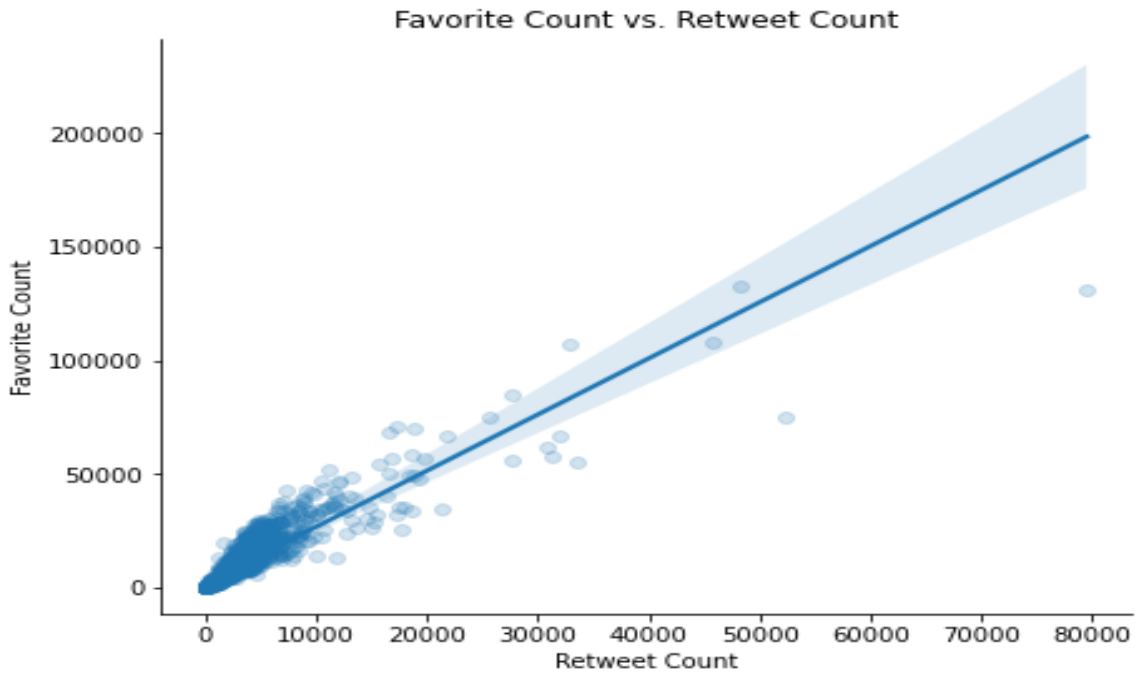
  An analysis of the retweets also showed spikes in activity at random intervals as shown below:



- **Favorite vs Retweet Counts:**

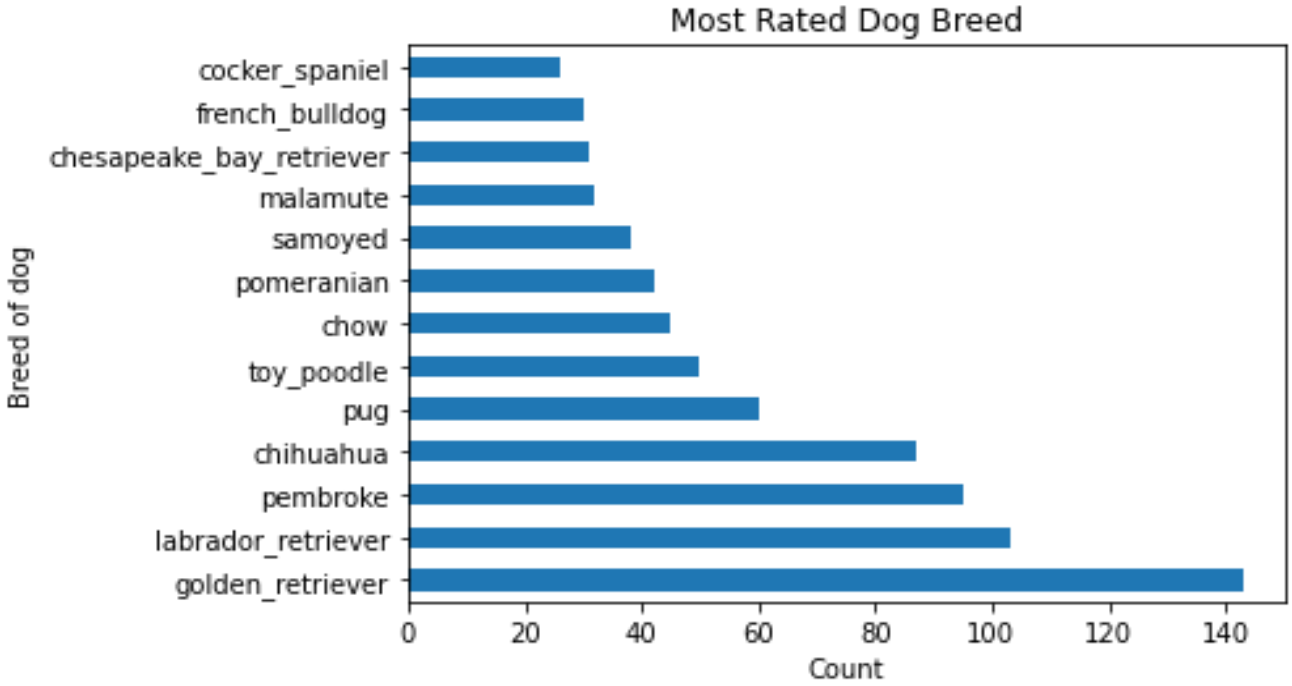
  There was generally a positive correlation in the relationship highlighted above and the same can be seen from the graph below. That is between the retweets and likes

  This correlation is important for the owner of the WeRateDogs twitter account to understand when determining method to increase users' traffic on the page. This could also assist the owner determine which dog breeds typically attract more traffic.



- **Dog Breed Popularity:**

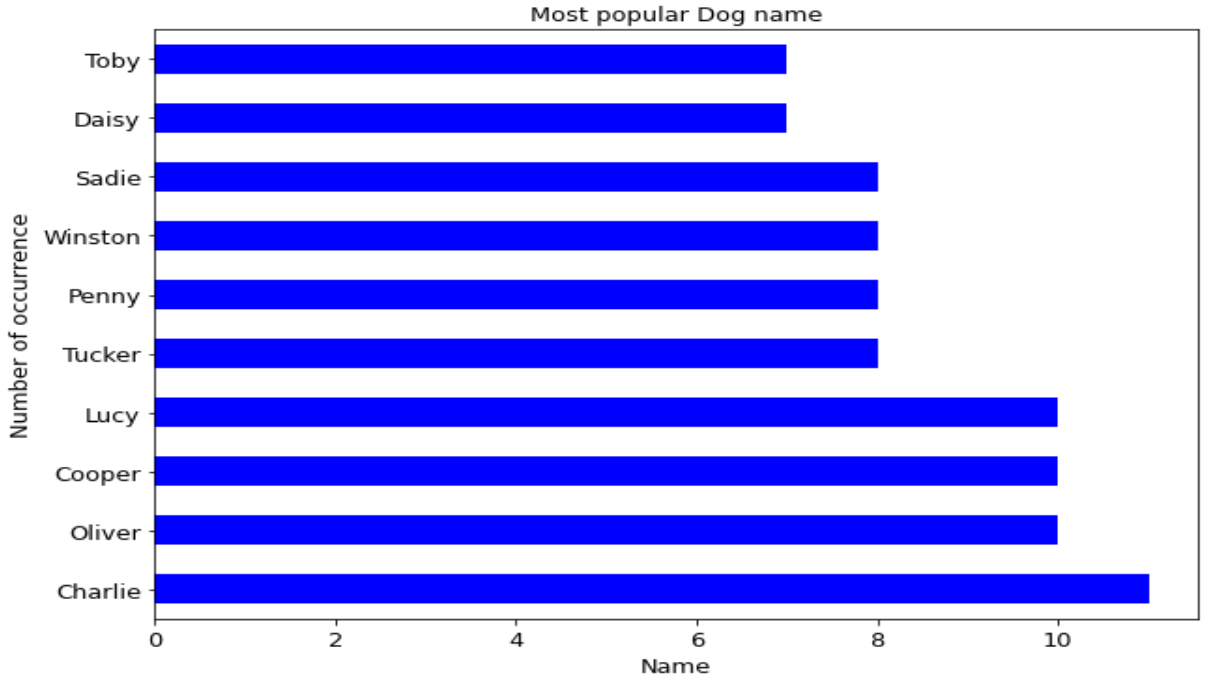
The most popular dog breed is golden retriever with the Labrador retriever coming in as the second most popular breed. The page owner knowing this should try to improve and angle his tweets towards the target audience.



- **Dog Name Popularity:**

  Names are important, especially for Dogs.
  The top 5 most popular dog-names are:

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



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.