Analysis Report for the WeRateDogs Project:

By Faith Omohodion

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 issues 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:
5. Each variable forms a column
6. Each observation forms a row
7. 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:

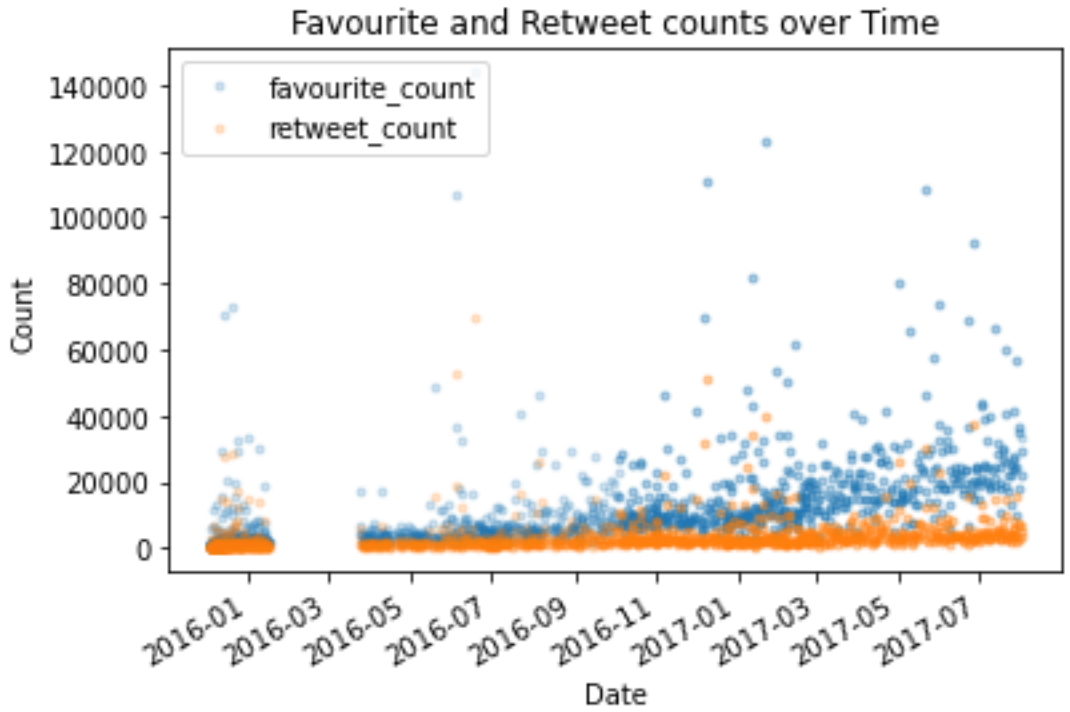
1. Define: Determine exactly what needs to be clean and how.
2. Code: Programmatically clean the code
3. Test: Evaluate the code to ensure the data set was cleaned properly.

Analysis and Visualization: There is several 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,

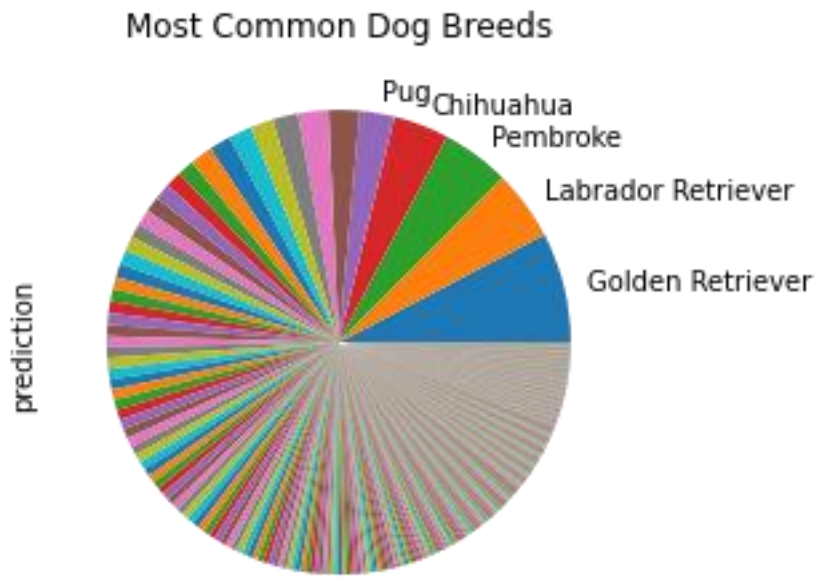
there are spikes in activity during early 2016 (i.e. 2016-01) and in mid-summer of 2016 (i.e. in between 2016-03 to 2016-05), but continues to generally decreased from there. The owner of the WeRateDogs Twitter account should be aware of this trend and consider way to increase users' traffic on the Page.

- Features in the dataset: The features of interest to me were retweet count, favourite count, rating, dog type, prediction and confidence percentage. To begin with I found out that Golden Retriever, Labrador Retriever, Pembroke, Chihuahua and Pug were the most common breeds predicted in this dataset. But when it came to breeds with the highest mean retweet and favourite counts these breeds did not even make the top 10. Breeds like Afghan Hound, Saluki, French Bulldog, Standard Poodle, English Springer and Cardigan were in the top 10 for both the highest mean count lists. Most of the breeds with the highest mean dog ratings weren't that high on mean retweet and favourite counts. Only exception in this list was the Saluki which was high on ratings as well as counts. Also, the two types of Retrievers namely the Labrador and the Golden which were two of the most common breeds did have high ratings on average. In terms of dog types, floofers and puppos had higher average ratings as compared to doggos and puppers. Though this can also be attributed to the fact that the number of dogs classified as floofers and puppos was a lot less than the number of dogs classified as doggos and puppers.



A lot of the tweets have very high retweet or favourite counts as represented by the outliers in the scatter-plot above.

I also created a pie chart showing the proportions of all the different predictions in the dataset.



This shows the prediction given and out of all the most common dog predictions 5 of them were dog breeds which are Golden Retriever, Labrador Retriever, Pembroke, Chihuahua, and Pug, accounted for around 25% of the total number of tweets in the dataset. Given the fact that my analyses were conducted from a clean data I can say that these numbers and visualizations are quite accurate.