WeRateDogs Twitter Archive - Act Report

WeRateDogs:

The dataset used for the analysis and visualisation 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

This Twitter archive contains basic tweet data (tweet ID, timestamp, text, etc.) for 2356 of the 5000+ of their tweets as they stood on August 1, 2017. The Twitter archive data does not contain the retweet count and favourite count for each tweet and this has taken from another data file provided by Udacity. Also image predictions file from Udacity containing the predictions for dog breeds from three different resources based on the images from the tweets has also been used for the analysis.

Data Wrangling

The initial data received is not clean and has some quality and tidiness issues. During the data wrangling process 8 quality issues and 2 tidiness issues have been identified using visual and programmatic assessment and cleaned the data. After cleaning the quality and tidiness issues found during the assessment, there were about 1951 tweets with good quality data.

Insights

The following three questions were analysed in the analysis

1) What are the top 30 most tweeted dog breeds by WeRateDogs?
2) Which stage of dogs got the highest retweet counts and favorite counts?
3) What is the relation between WeRateDogs Ratings Vs Retweet and Favourite count?

1) What are the top 30 most tweeted dog breeds by WeRateDogs?

Analysed the data based on the predictions with highest confidence and found the top 30 most tweeted dog breeds by WeRateDogs and the plot is shown in Figure 1.

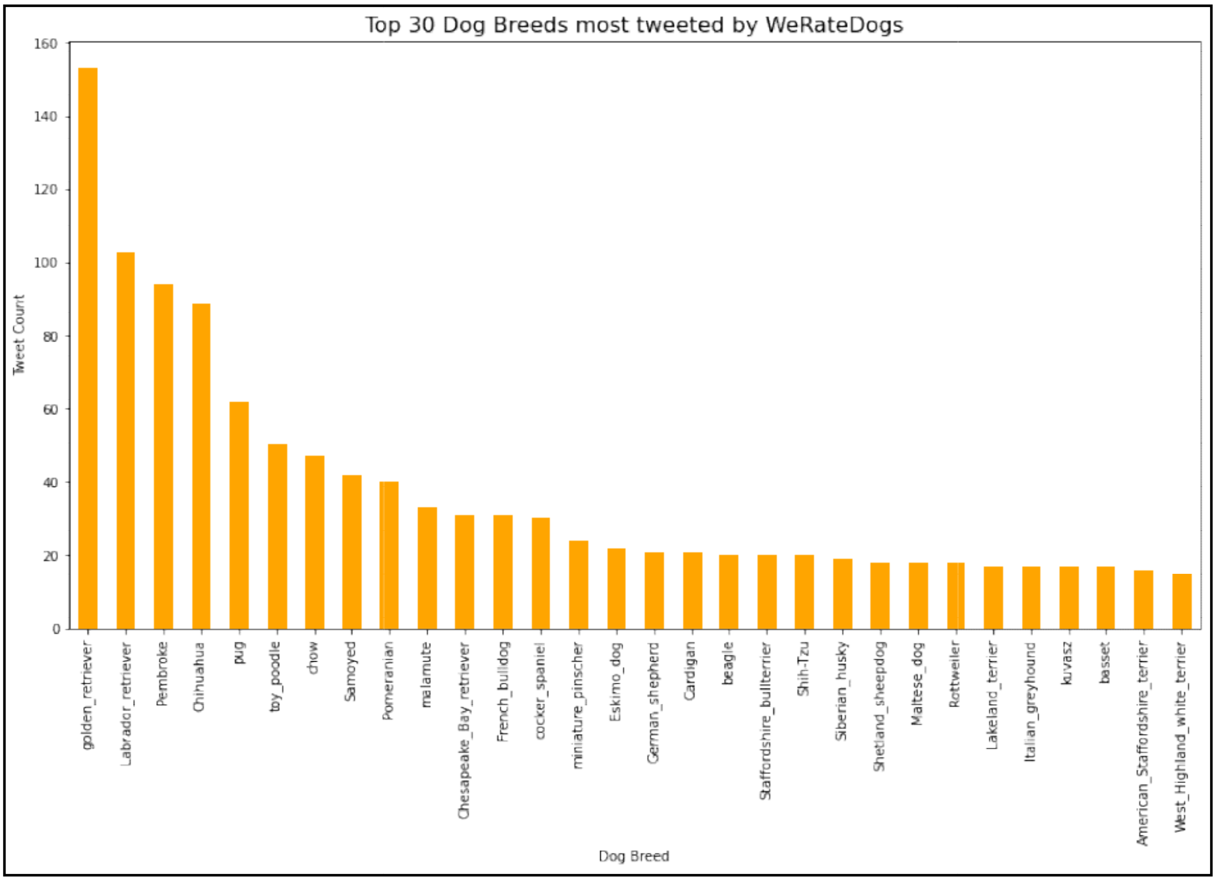

Figure 1: Top 30 dog breeds most tweeted by WeRateDogs

From Figure 1 the list of top 30 dog breeds most tweeted by WeRateDogs can be seen and also Golden Retriever is the most popular dog. Using python libraries the top 5 Golden Retriever dogs images extracted and presented in Figure 2 to Figure 6.

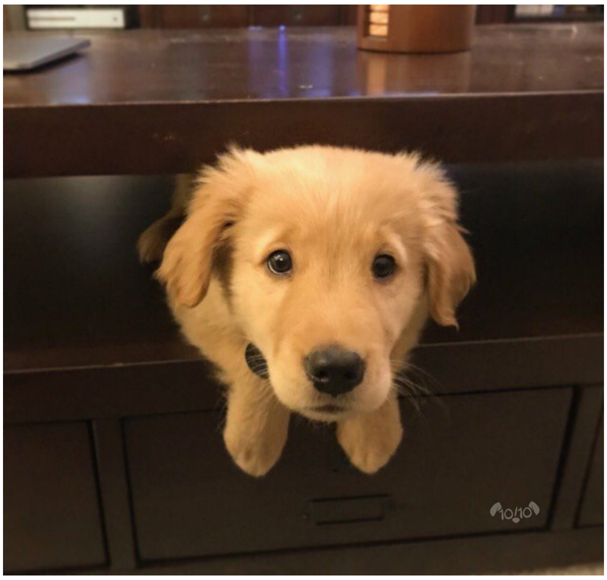

Figure 2: First most tweeted dog by WeRateDogs in Golden Retriever breed

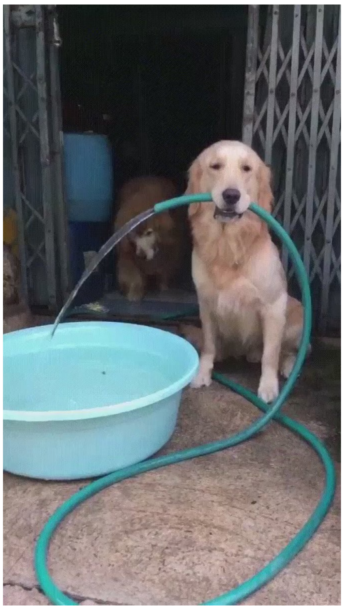

Figure 3: Second most tweeted dog by WeRateDogs in Golden Retriever breed

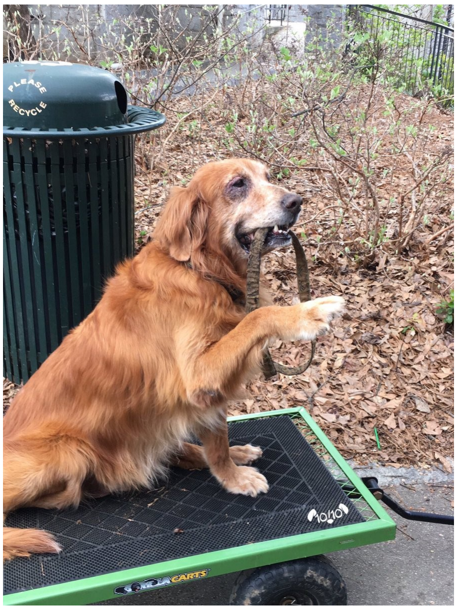

Figure 4: Third most tweeted dog by WeRateDogs in Golden Retriever breed

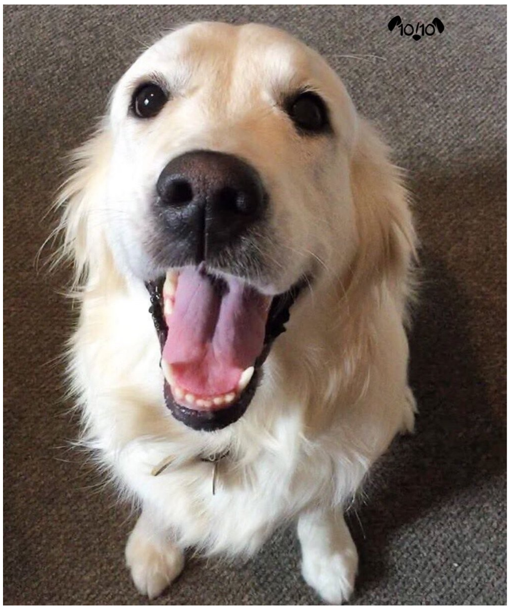

Figure 5: Fourth most tweeted dog by WeRateDogs in Golden Retriever breed

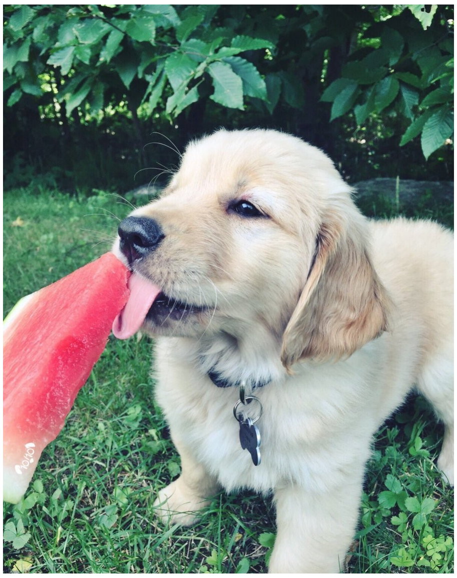

Figure 6: Fifth most tweeted dog by WeRateDogs in Golden Retriever breed

2) Which stage of dogs got the highest retweet counts, favorite counts and the highest average rating?

Plots of dog stages Vs total favourite counts, retweet counts and average rating are shown from Figure 7 to Figure 9.

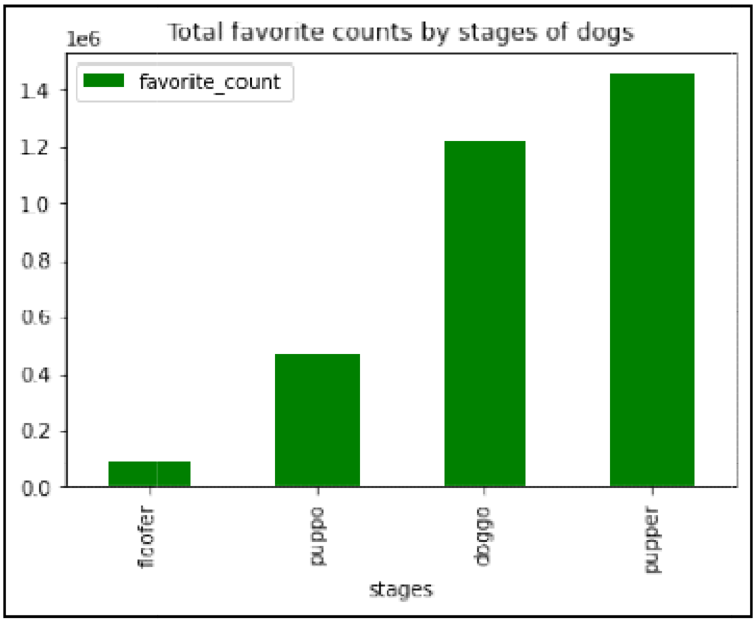

Figure 7: Total favorite counts by stages of dogs

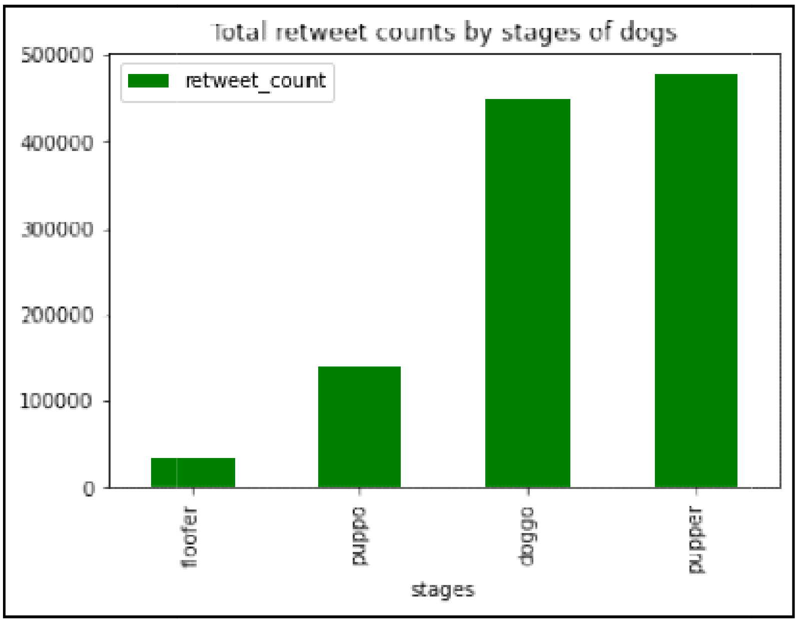

Figure 8: Total retweet counts by stages of dogs

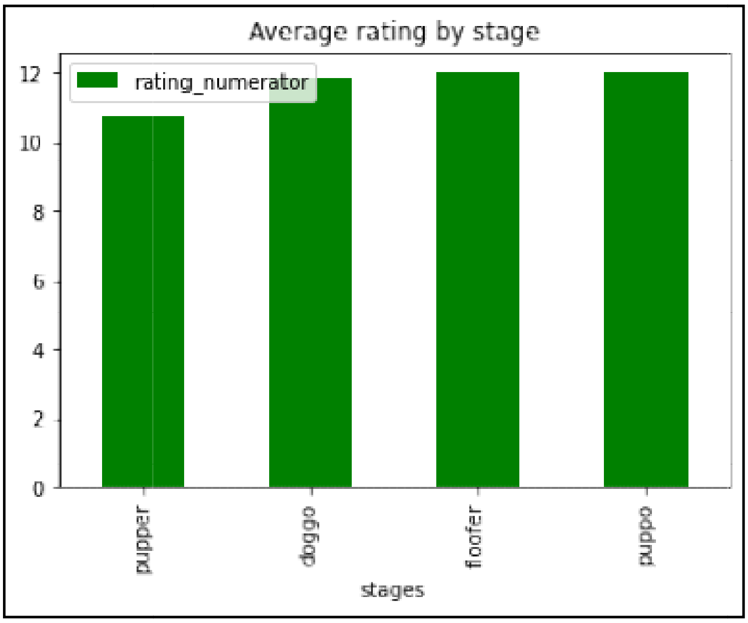

Figure 9: Average rating by stage

The dog stage 'Pupper' has got the highest favorite and retweet counts the average rating of the dog stage 'Pupper' is slightly lower than the remaining 3 dog stages.

3) What is the relation between WeRateDogs Ratings Vs Retweet and Favourite count?

Plots of dogs ratings by WeRateDogs Vs favourite counts and retweet counts are shown from Figure 10 to Figure 11.

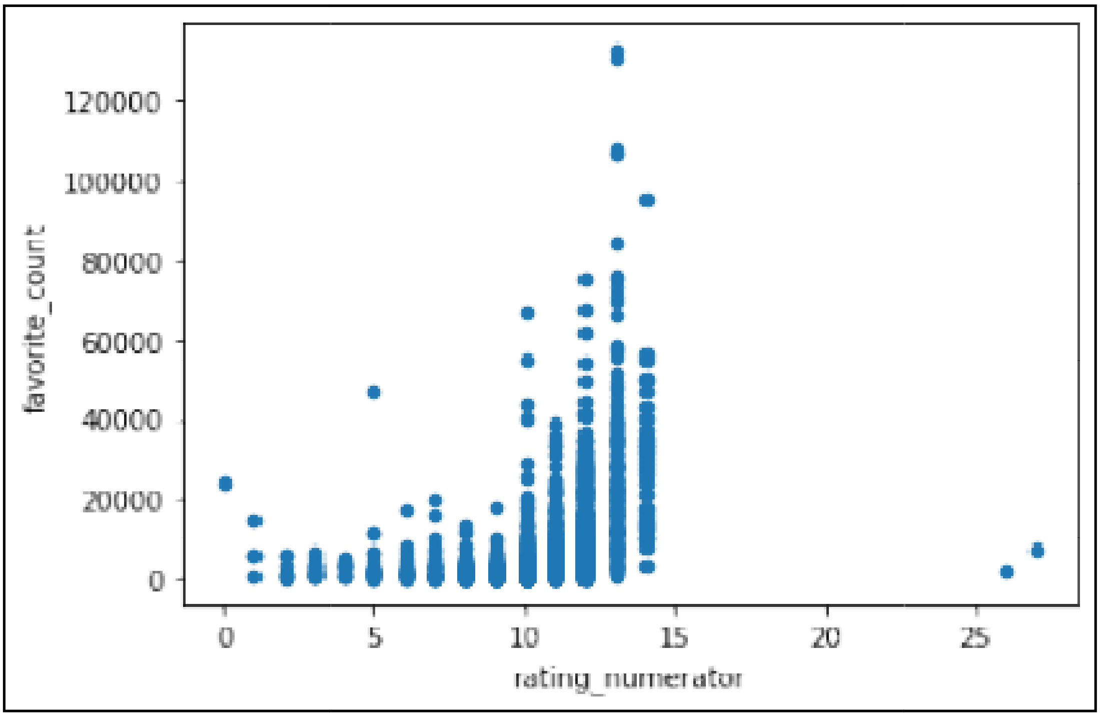

Figure 10: Dogs Rating Vs Favorite count

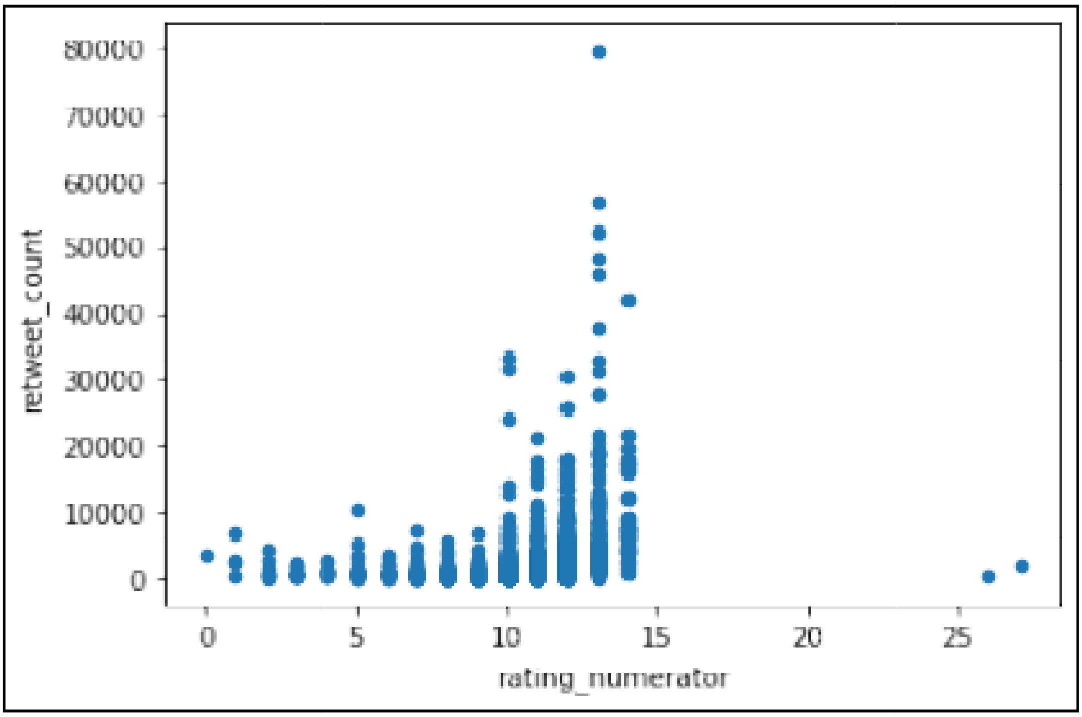

Figure 11: Dogs Rating Vs Retweet count

There is no clear and direct relation between dogs ratings by WeRateDogs Vs Favorite and Retweet count (i.e. the dogs with highest rating does not necessarily have highest favourite or retweet count). In general the dogs with rating 14 (which is close to the highest rating of 15, excluding the outliers) have the highest favourite and retweet count.

The aim of the project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. This report contains the steps performed and with a brief summary of the data wrangling efforts in this project.
The key steps performed in this project were:

1. Data wrangling
    - Gathering data
    - Assessing data
    - Cleaning data
2. Storing the data
3. Analyzing and visualizing the data (not covered in this report)
4. Reporting
    - Wrangling efforts (current report)
    - Analyses and visualisations (not covered in this report)

Data Wrangling:

Real-world data rarely comes clean. Using Python and its libraries, I have gathered data from a variety of sources and in a variety of formats, assessed its quality and tidiness, then cleaned it, so it can be used for analyses and visualisation.

Gathering data:

The data for this project has been gathered from three different sources.

1) Twitter archive data (file: twitter_archive_enhanced.csv)

This file has been downloaded manually from Udacity website and read the data into pandas dataframe. The file contains the information such as twee id, the dog's rating and its stage along with some additional information like timestamp, urls, source etc.

2) Tweet image predictions (file: image_predictions.tsv)

The file has been downloaded programmatically using Requests library, the download link is provided by Udacity. The file contains the tweet id and dog breed predictions data from 3 different prediction sources.

3) Twitter API additional data (File: tweet_json.txt)

This file has been downloaded manually from Udacity website and read the data into pandas dataframe. The file contains the information such as twee id, retweet count and favourite count etc.

Assessing data:

After gathering all the required data from three different resources, the data has been assessed visually and programmatically for quality and tidiness issues. Identified at least eight quality issues and two tidiness issues in the data gathered.

The visual and programmatic assessment showed several quality and tidiness issues such as redundant rows, columns, incorrect data type, very suspicious dogs rating, multiple columns with similar information where it is possible to present the information in one single column and data in multiple files than in a one single file etc.

Cleaning data:

The data has been cleaned for the quality and tidiness identified during the assessment phase. The data cleaning has been done in three steps Define, Code and Test.

Before cleaning the data, copies of the original dataframes have been created, so the original data will always remain available and if there are any issues in cleaning process the original data can be still accessed.

Twitter archive data (file: twitter_archive_enhanced.csv):

- The dog stages 4 columns 'doggo', 'floofer', 'pupper', 'puppo' has been converted to one single column 'stage' and dropped these 4 columns.
- Removed the redundant data: Dropped all rows containing retweets where these columns will be non-null and then dropped the retweet columns. Similarly dropped all rows containing ‘in_reply_to’ entries where these columns will be non-null and then dropped the ‘in_reply_to’ columns.
- Dropped all rows with missing data in 'expanded_urls' column
- Dropped all rows with 'rating_denominator' coulmn is not equal to 10 and also the rows with ‘rating_numerator' value greater than 30.
- Changed the 'timestamp' to datetime data type

Twitter API additional data (File: tweet_json.txt):

- Merged the 'retweet_count' and 'favorite_count' columns from the Twitter API additional data to Twitter archive data and removed any rows with missing information and dropped the redundant columns.

Tweet image predictions (file: image_predictions.tsv)

- Merged the Tweet image predictions dataframe to Twitter archive data and removed any rows with missing information and dropped the redundant columns.

- Created a single column 'Breed' covering the breed of the dog (from p1, p2, p3) and another column 'Confidence' covering the highest confidence from p1_conf, p2_conf, p3_conf and dropped the redundant columns

Storing the data:

A single dataframe was created by cleaning the quality and tidiness issues and was saved to a new ‘twitter_archive_master.csv’ file.