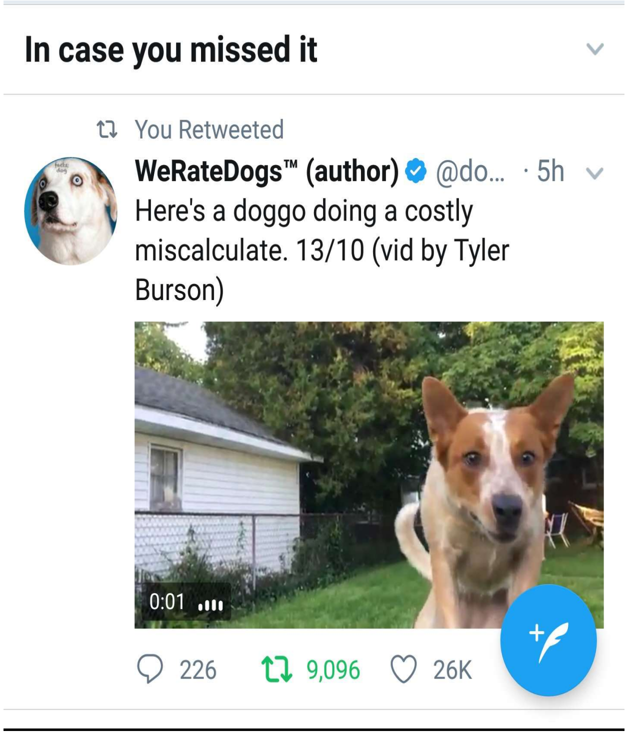
# WeRateDogs – Analytical Insights for the @dog_rates Twitter page

## Introduction and Background

Real-world data rarely comes clean. We have used tweet archive of Twitter user @dog_rates as the project data set. @dog_rates or “WeRateDogs” is a Twitter account which rates people's dogs with a humorous comment about the dog.

We can cite few examples as the following:



In case you missed it

You Retweeted

WeRateDogs™ (author) @do... · 5h

Here's a doggo doing a costly miscalculate. 13/10 (vid by Tyler Burson)

*Source: https://twitter.com/Kathryn_CC/status/895840796104163329/photo/1*

This project encompasses the processes such as gathering, assessing, and cleaning of data. Later various types of visualizations and analytical reporting were performed based upon the cleaned data.

## Gathering Data

This project gathered data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students.
- WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets.

- The tweet image predictions, i.e., what breed of dog is present in each tweet, as per the neural network predictions. This file was provided to Udacity students.
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data that I find interesting.

*Source: https://twitter.com/dog_rates/status/1235633333704384513*

## Assessing Data

The purpose of data assessment is to evaluate a data set on quality and tidiness issues.

The four (4) main data quality dimensions are:

- Completeness: missing data?
- Validity: does the data make sense?
- Accuracy: inaccurate data? (wrong data can still show up as valid)
- Consistency: standardization?

And there are three (3) requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

As we scan through the gathered data, we should identify the type of data to be visualized and obviously, the filtered (cleaned) proportion of data that would be used programmatically, to convey important insights to the stakeholders.

## Cleaning Data

The cleaning of data usually takes of the following steps:

1. Define: identify exactly what needs to be cleaned, and the process to clean it

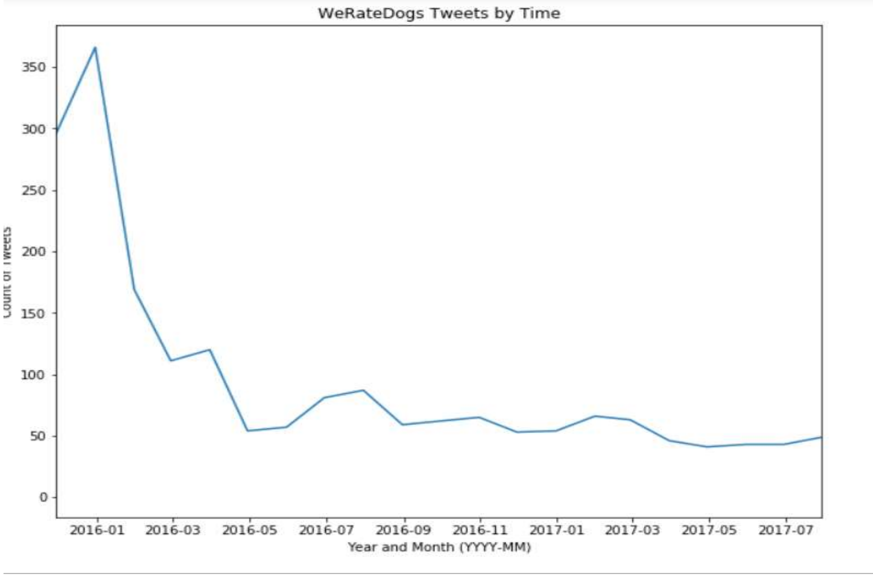
2. Code: programmatically cleaning the code

3. Test: evaluating the code to ensure that the data frame was cleaned properly

## Analysis and Visualization

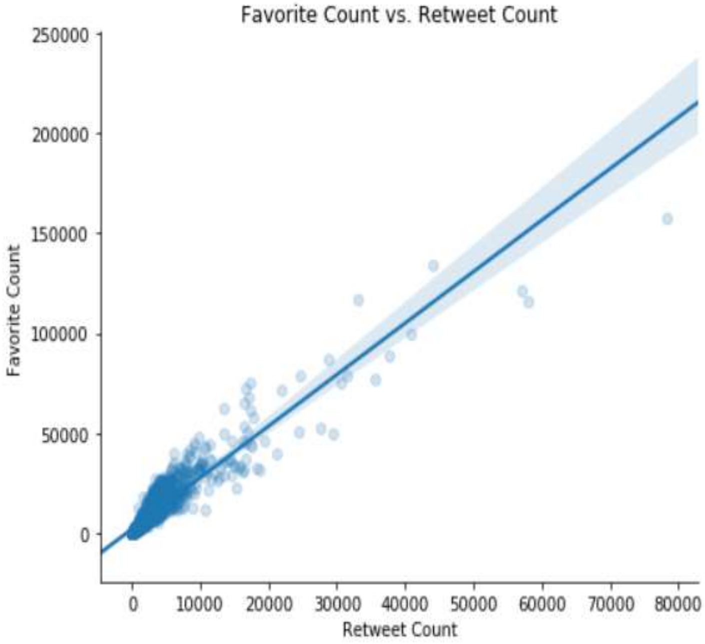
Four different data characteristics are presented as insights, from the cleaned data.

### Tweets vs Time



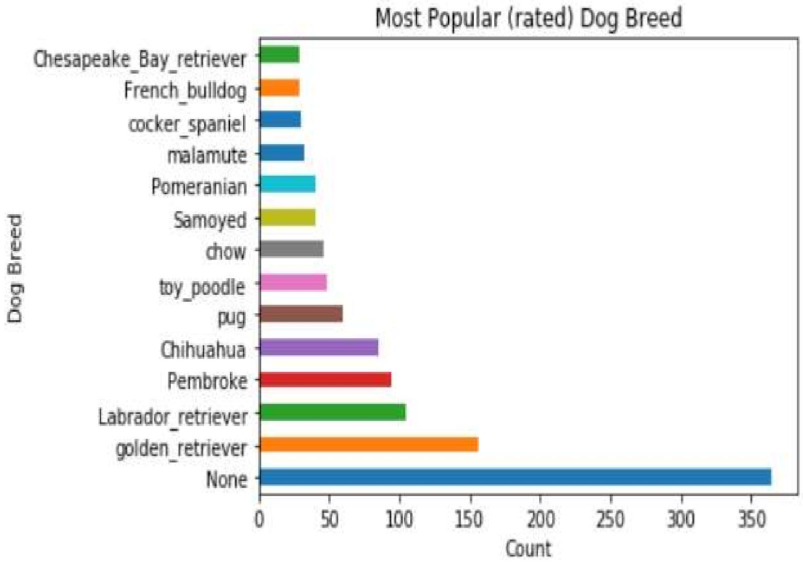
If we follow the "Tweets by Time" graph at the right, we can see that, the tweets decreased sharply starting in early 2016. While the tweets continue to decline over time, spikes were identified during the early spring of 2016, mid-summer of 2016, but later continues to generally decrease from there. The data set doesn't contain data that would clarify such trend, but the owner of the data set should see the visuals and try to re-vitalize the decreasing Tweet trend, since 2016.

### Favorite vs. Retweet Counts



Positive co-relation is identified between favorite ('like") counts, and how much a post was retweeted. This co-relation is going to be important for the owner of the WeRateDogs Twitter account in order to formulating methods to increase user traffic on the page. A data could provide data of historically popular posts, so that the page owner could model future posts off historically popular posts.

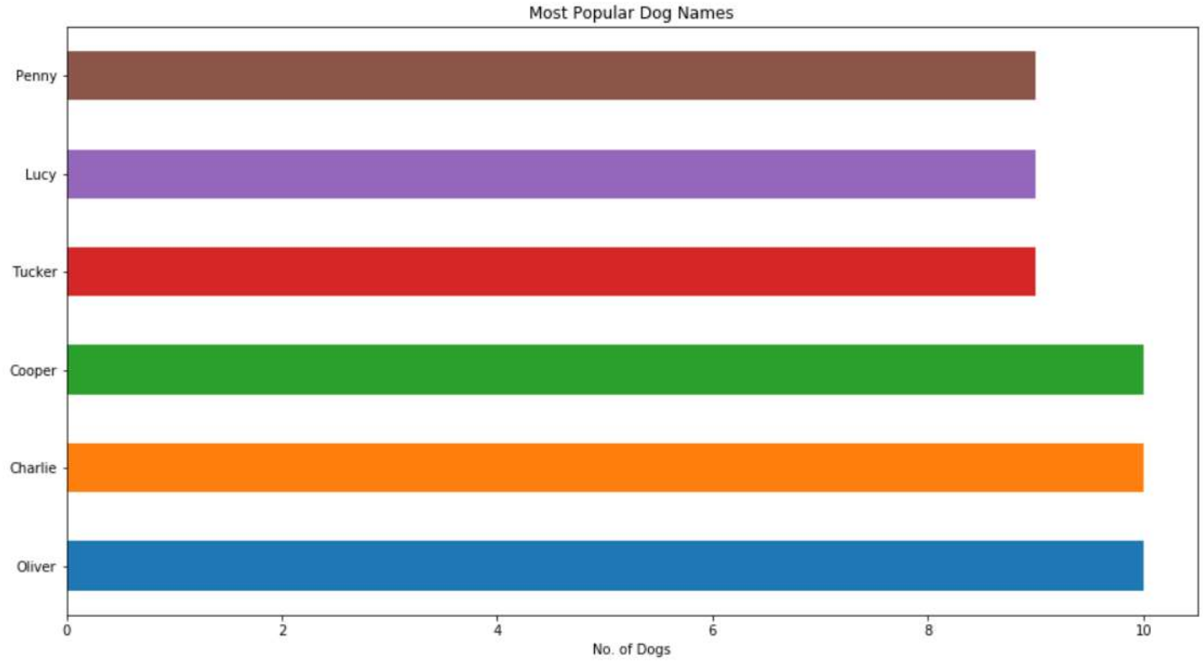
## Dog Breed Popularity



Golden Retriever happens to be the most popular dog breed, followed by Labrador Retriever. Pembroke and Chihuahua are also having comparable popularity like Labrador Retriever. The page owner could use this information to create targeted marketing efforts for certain breeds that aren't as popular to increase their popularity (for example, French bulldog, Cocker Spaniel, Chesapeake Bay Retriever etc.) but also utilize the popular breeds to divert user traffic to those.

## Dog Name Commonality

The three most popular dog names are Oliver, Charlie and Cooper. While the next popular set contains Tucker, Lucy and Penny.



## Conclusion

The above data wrangling exercise could find a number of distinctive features of the data set which could be utilized to increase future traffic, or to increase popularity over time. By carefully scanning through the data over and again, we would possibly be able to identify more diverse set of attributes that would determine the inherent insights present within the data.