WeRateDog Twitter Data Analysis
Wrangling Report
By **Brian Luk**

Data Source
The 3 data sources are gathered in 3 ways – .csv file manually added to the locally machine, data file retrieved from the Python Tweepy's library, and .csv file download from the web using Python's Requests library.

Assess/Clean Data
I first did a manual visual assessment of the data, then a programmatic assessment.
The first task is to check for data type, null values, duplicates and incorrect values. I have identified and fix the data type of the *Timestamp, Tweet ID, rating_denominator and rating_numerator* column. There is also an issue with the null valued represented by a "None" String value, particularly in the dog stages and name columns. Idenified an incorrect dog stage name, I changed the column name of *floofer* to *floof*, as well as the value of the column.

Then, I joined the 3 data frames a single data frame. Conveniently, during the process of joining the *info_df* and *archive_df*, I excluded all the retweets or reply-tweets, leaving only the original ones.

I un-pivots some columns into a single column as well so that the data frame better consolidated and data is stored in way more readable for the computer. I have merged the *rating_denominator* and *rating_numerator* columns to a *rating* column by changing the data type of the 2 columns for int to string and adding a "/" between the 2 Strings. I also merged for the dog stage columns into a single column called `stage`.

Lastly, I save the cleaned data frame into one single .csv file.

When I am performing data analysis and visualization, I notice that I have done the data type converted of the 2 rating columns and merged them by mistake. Particularly as I was doing visualization, having the denominator and numerator as *int* turned out to be very useful for data filtering in order for the visualization to make sure. And I realized that merging the 2 columns will only be an aesthetic improvement.