# Wrangle WeRateDogs Twitter Data

In this data wrangle project, I followed the three steps of data wrangling.

In the gathering data stage, I loaded csv file into pandas dataframe. Also, I was able to download a tsv file from the url provided and read the file into pandas dataframe. The most difficult task was to gather additional data from Twitter API. Firstly, I created a Twitter developer account, this allowed me to create consumer key and access token and get access to Twitter API. Then I run the script to get tweet JSON data using tweet ID. The data was saved in a text file, with each row corresponding to one tweet ID. Lastly, I read the text file line by line to get the retweet count and favorite count numbers and store them in a pandas dataframe.

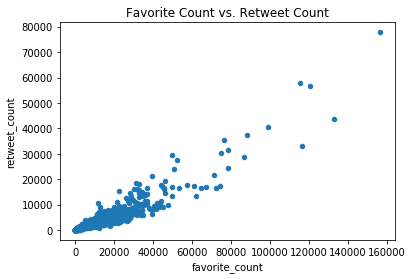
In the assessing date stage, I used both visual and programmatical method to assess the quality and tidiness of the three dataframes. I identified 6 qualities issues from “archive” data and 1 each from “img\_pred” and “additional” data. They were missing values, wrong data types, case sensitivity, and unreliable data, etc. I also found two tidiness issues from the three data sets, such as too many columns forming one variable, and tables should be joined into one.

In the cleaning data stage, I made a copy of each dataframe and worked on the copy in the following cleaning steps. I first dropped the columns in “archive” that were not useful in the analyzing stage. This also helped me to remove missing values in the table. Then I changed timestamp column to datetime type, dog stages columns to Boolean type; removed rows that had timestamp before August 1st, 2017, and rows with 0 rating numerator and denominator; changed predicted breed columns to lowercase in “img\_pred”. Lastly, I joined the three cleaned table, removed rows with missing values, created a rating column for further analysis.

# Act Report - WeRateDogs Twitter Data

As we enjoy surfing on WeRateDogs pages, it is even interesting to get insights on data behind the scenes. After analyzing the cleaned WeRateDogs Twitter data, I have three interesting insights to share.

First, I found favorite counts have positive relationship with retweet counts, that is, tweets with higher favorite count tend to have higher retweet count, tweets with lower favorite count tend to have lower retweet count. This is intuitively true, as the more people like the tweet, the higher chance it gets retweeted. The relationship between these two variables are shown in the scatter plot chart below.



Furthermore, I am interested in finding who is the one with the highest rating score. Atticus, the one wiring sunglasses and an American flag bow tie, is the superstar. He got 1776/10 rating, which is much higher than any other dogs. However, he is not the one with the highest favorite count. Atticus only has 5,123 favorite count, whereas a Labrador Retriever has the highest favorite count of 156,164. So high rating does not necessarily come with high favorite count.

Finally, I explored which breed of dog has the highest average favorite count. It turned out to be Saluki, with an average favorite count of 22,193, which is much higher than the second place French Bulldog’s 17,475. This result is based on the predicted breed with the highest confidence.

The data can tell us much more than the three insights above. It was fun wrangling and analyzing the WeRateDogs Twitter data**!**