# Wrangling and analyzing Using Tweepy for weratedogs account tweets

**By: Mahmoud El-Sayed**

- **This report describes the effort done in wrangling for this project**

Wrangling consists of 3 main steps

1- Gathering

2- Assessing

3- Cleaning

## Gathering

We collected Data from 3 main sources that created 3 datasets to work with.

1- we have a file at hand (**twitter-archive-enhanced**)

2- The second Dataset was a file produced from processing images from a neural network model that created predictions and scores for those predictions, we pulled this file using Requests library

3- Finally, the last dataset was pulled using twitter API (tweepy) to obtain valuable info like retweets count and favorites count

## Assessing

Then we had to assess those 3 datasets for problems to clean for further analyzation.

Assessing can be broken down to 2 main problems

### - Quality issues

- name column contains 'a' instead of an actual name

- timestamp,retweet_status_timestamp should be of type datetime

- None in Doggo,floofer,pupper and puppo columns should be changed to NANs

- column names are not illustrative

## - tidiness issues

- Doggo,floofer,pupper,puppo should be melted into one column

- text column contains both link and text which breaks first rule of tidiness

- df_1, df_2 and df_3 should be merged into 1 dataset

## Cleaning

In this step we solve the problems we documented in the assessing step programmatically ste[ by step by first defining the problem then coding then testing to see if our solution fixed the problem correctly or not.

**And at last, we have clean dataset(s) that we can work on and analyze and look for useful info and insights.**

Examples for some useful info that we can see now

- Most retweeted tweet
- Most favorited tweet

And a plot of the retweets and favorite count with respect to time.

**Output of this project:**

Table with Shape of (2356,26)