**Wrangling WeRateDogs Twitter Data**

Data wrangling includes three parts:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

# 1.Gathering Data

I gathered data from three different sources in a Jupyter Notebook titled ​**wrangle\_act.ipynb**​.

* The WeRateDogs Twitter archive were already given as **twitter\_archive\_enhanced.csv**​*.* ​I downloaded it manually and created **df\_*twitter\_archive***​ ​​data frame.
* I downloaded **image\_predictions.tsv**​ file programmatically using the Requests library and the given url and created ​**df\_image\_predictions**​. This consists tweet image predictions.
* To obtain the ​**df\_tweet**​​data frame with tweet ID, retweet count, and favorite count I did the following: I requested a developer account from twitter but there was no respond, so I provided the necessary code and used the ​**tweet\_json.txt**​ file that was provided and convert it into a data frame,

# Assessing Data

I assessed the data in ​**wrangle\_act.ipynb** both visually and programmatically. I used *head(),*

*.value\_counts(), info(), .isnull(), .notnull()*​ functions of Pandas to assess the data programmatically. I also exported the data into Excel to have deeper look. Then by taking into account the ‘Key Points’ which are stated in the project motivation, I detected and documented quality & tidiness issues as the following:

***Quality***

* df\_twitter\_archive table:

*- Missing data in expanded\_urls (Tweets without images).*

*- 181 Retweets.*

*- Incorrect dog names.*

*- Missing values in dog names.*

*- Wrong datatype for tweet\_id.*

*- Wrong datatype for timestamp.*

*- Unclear data in the source column.*

*- Dog stage's type is not categorical.*

* *df\_image\_predictions table.*

*- Missing records (2075 instead of 2356).*

*- Lowercase breed names in p1, p2, p3 and '\_' is used instead of space.*

***Tidiness***

- Merge all three data frames.

- Combine dog "stage" columns (i.e. doggo, floofer, pupper, and puppo).

- Combine rating\_numerator and rating\_denominator columns.

- Drop unneeded columns.

# 3.Cleaning Data

First, I created copies of the data frames before cleaning data. I cleaned data by documenting the define, code, and test steps of the cleaning process. I started with dealing the missing data and then merged three data frames to work on a single data frame named as **df\_twitter\_archive\_clean.**​ After that, I tried to solve the other quality and tidiness issues in a logical order.

I mostly used functions of Pandas, loops. I also cleaned some data manually for incorrect dog ratings. I spent quite time on re-extracting, cleaning and correcting names, ratings, dog stages, and cleaning the tweets with the non-dog images. Overall, I believe that this project challenged me to improve my data wrangling skills.

Finally the cleaned master data set which will be used in data analysis is stored in a csv file named ​**twitter\_archive\_master.csv**​.