# WeRateDogs Twitter Data Report

## Interdiction:

The WeRateDogs dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

By using Python and its libraries, we will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called Data Wrangling. we will document our wrangling efforts in this Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries).

Our goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations , in this report briefly describes my wrangling efforts.

## Project details:

### The tasks of this project are as follows:

1- Gathering data.

2- Assessing data.

3- Cleaning data.

## Steps:

### 1- Gathering data

Data was gathered from 3 different sources:

**Twitter archive file:**

the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

**The tweet image predictions:**

i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information.

**Twitter API & JSON:**

Retweet/Favorite"like" count by query the Twitter API for each tweet's JSON data (programmatically using Python's Tweepy library) and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Then reading this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

### 2-Assessing data

In this step, we assessed all datasets that we have by various ways in order to get Quality/Tidiness issues, and we noticed these following issues:

**Quality :**

1- timestamp and **retweeted_status_timestamp** should be datetime not Strings.

2- Create a function where I keep the first true prediction along the confidence level as new columns.

3- rename id in tweets to **tweet_id** so can merge later.

4- Delete columns that won't be used for analysis.

5- missing some **expanded_urls**.
6- column timestamp sepreate them into two columns Date and Time.

7- drop duplicate **jpg_url**.

8- change the values in name from None ,an ,a or the to null.

9- Create 1 column for image prediction and 1 column for confidence level.

10- Keep original ratings (no retweets) that have images.

**Tidiness :**

-last four columns are stages for the dog, better we make it one column called stage using melt function.

- Merge 'tweets' and 'image_prediction' into '**twitter_archive**'.

**Cleaning data :**

It is the process of fixing and resolving issues identified in the Cleaning process. The (code, and test) steps were used in the cleaning process. First, copies of the Data Frames were created before cleaning. Then, the steps of cleaning were applied iteratively on all issues.

**Conclusion :**

In the end Data wrangling is a core skill that whoever handles data should be familiar with. I have used Python programming language and some of its packages. There are several advantages of this tool (as compared to e.g. Excel) that is used by many data scientists.

For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases ,It is strong in dealing with big data (much better than Excel), It can deal with a large variety of data (unstructured data like JSON (Tweets) or also structured data from ERP/SQL databases, It is easy to document each single step and if needed re-run each single and in.

Handling, assessing, cleaning and visualizing of data is possible programmatically using code.