# Wrangle Report

## Introduction:

This project aims to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

project are as follows:

Data wrangling, which consists of:

- Gathering
- Assessing data
- Cleaning data
- Storing, analyzing, and visualizing your wrangled data

## 1- Gathering Data

Gather each of the three pieces of data as described below in a Jupyter Notebook titled wrangle_act.ipynb:

1- The WeRateDogs Twitter archive which stored in the file named “twitter_archive_enhanced.csv”

2- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3- The tweet json file I download from Udacity as the twitter account still in the review process

## 2- Assessing data :

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues.

**issues :**

- **Quality**

1- twitter archive data frame :

- source column is not clear contain html tag
- name column contain wrong names and lower charcter
- rating_denominator column is greater than 10 in some entries and there one contain 0 and it wrong

- some entries are retweets
- there expanded_urls are missing
- wrong data type for tweet_id column should be object not int
- wrong data types (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, timestamp and retweeted_status_timestamp) should be datetime
- many missing values in more than column

**2- image predictions data frame :**

- wrong data type for tweet_id column should be object not int
- missing data
- p1, p2, p3 should be  lower
- there a  duplicated in  jpg_url

3-  tweet json data frame :

- wrong data type for tweet_id column should be object not int
- missing data

**<u>-  Tidiness</u>**

- Merge all datasets to one.
- The last columns of twitter archieve can be merged to one as dog's type.

- Merge rating_numerator and rating_denominator columns to one.
-

- **<u>Cleaning data:</u>**
  In this phase I solved all the issues that discovered in the assessing data phase