# Data Wrangling Report

By Islam Soliman

As an assignment for Udacity Data Analysis Professional Nano-Degree; This report illustrates the main steps I've followed in the data wrangling of Twitter account "WeRateDogs".

## Data Gathering

This is the initial step in which we collect the data for this project from three main sources:

1. Twitter_archive_enhanced.csv file, this file was downloaded manually and uploaded to the project workspace and finally imported to our working environment using Pandas function "pd.read_csv()" creating "archive_df".
2. Image_predictions.tsv file, this is the second file that was hosted on a webpage and imported from its URL using Requests library function "get()" and Pandas function "read_csv()" creating "image_predictions_df". (Note: This file contains image predictions for the rated dogs).
3. The final dataset was gathered from Twitter API via tweepy library, alternatively via the provided "tweet_json.txt" creating "api_df".

## Data Assessment

This is where I investigated the three imported datasets both visually and programmatically fo quality and tidiness issues:

**Tidiness findings:**

1. Dog stage data are not combined in one categorical column.
2. The three DataFrames should be combined together for better analysis.

**Quality Findings:**

1. There are 181 retweets indicated in retweeted_status_id (Keep original tweets only).
2. There are 745 tweets without dogs names.
3. Invalid tweet_id data type (int instead of str)
4. Invalid timestamp data type (str instead of datetime)
5. There are 281 tweets with missing photo URL.
6. Underscores were used in multiple locations at p1, p2, and p3 instead of using space.
7. Inconsistencies in names between lower and upper case letters.
8. There are some duplicated values in the dataframe.
9. Wrong image predictions by the API (entry 444).
10. Some tweets contain more than one rating.
11. Columns that won't be used for analysis should be deleted.

## Cleaning Data

This part of the data wrangling was divided in three parts: Define, code and testing the code. These three steps were on each of the issues described in the data assessment section.

First I've create a copy of the three original DataFrames. I wrote the codes to manipulate the copies. This allowed me to create a new copy from the original whenever needed in case of errors or mistakes, I could create another copy of the original DataFrames and continue working on the cleaning part.

This stage has challenged me in many ways and made me feel more comfortable cleaning data programmatically on my own.

## Conclusion

Data wrangling is a core skill to any Data Analyst.

I have used Python programming language and many of its packages. There are many advantages of this tool (as compared to e.g. Excel) that is used by many Data Analysts.

• For gathering data there are several packages that assists in web scraping, and using APIs to collect data (Tweepy for Twitter) or communicating with SQL databases.

• It is a stronger and faster tool in dealing with large amount of data.

• It can assess data with a large variety of tools and libraries.