# A BRIEF REPORT OF THE DATA WRANGLING PROCESS ON THE WeRateDogs RATINGS DATASET

By Joseph Osuntoki

## 1.0 INTRODUCTION

This project's objective was to gather WeRateDogs Twitter data in order to produce insightful and reliable analysis and visualizations.

This project is primarily focused on manipulating data from the WeRateDogs Twitter account using Python. A final notebook (wrangle act.ipynb) was produced after the data wrangling process.

The project is the second project of the Udacity Data Analyst Nanodegree program.

## 2.0 PROJECT REQUIREMENTS

The tweet history of Twitter user @dog rates, better known as WeRateDogs, is the dataset that used in this project.

WeRateDogs is a Twitter account that rates users' dogs and adds a lighthearted comment.

It is required in this project that the following objectives are met:

- Gather the data needed
- Assess the data
- Clean the data
- Store the data
- Analyzing and visualizing the data
- Reporting

### *Tools and Libraries Needed*

A jupyter notebook is the main tool needed to complete this project. After which, some specific libraries were imported for the data wrangling process. These libraries make the process easier and more efficient.

1. pandas - For effective data manipulation
2. numpy - For performing arithmetic operations on arrays
3. requests - To download a file from the internet programmatically
4. json - To read the json file that was queried from Twitter
5. matplotlib - For data visualization
6. seaborn - An advanced data visualization library
7. os - Provides functions for modifying folders and fetching data from them
8. tweepy - To query the twitter API

## 3.0 PROCESS OVERVIEW

### *Gathering the Data*

Three different forms of data were used for this project, and they were acquired as described below:

- **WeRateDogs Twitter Archive File:** Udacity programmatically extracted this and made twitter archive enhanced.csv available for usage.
- **Image Predictions File:** According to a neural network, each tweet's image predicts the breed of dog that is present. The URL https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2adimage-predictions/imagepredictions.tsv was used to programmatically download the file (image predictions.tsv), which was hosted on Udacity's server.
- **Twitter API & Tweet JSON File:** Using the tweet IDs from the WeRateDogs Twitter archive, I used Python's tweepy module to query the Twitter API for each tweet's JSON data, and I saved the whole set of JSON data for each tweet in a file called tweet json.txt.

### *Assessing the Data*

The three files were assessed both visually and programmatically. Both assessments are useful for detecting data quality issues and data tidiness issues. Data quality issues is mainly concerned completeness, validity, accuracy, and consistency. Data tidiness, on the other hand is concerned with structureal issues that make analysis difficult.

### *Cleaning the Data*

**Data Quality Issues**

I discovered the following issues during the wrangling process:

1. Repetitive columns: having both retweet_status_id and retweet_status_user_id in the twitter_enhanced dataset creates redundancy. They are not necessary for our analysis
2. Incorrect data types in date (twitter_enhanced), all id columns (twitter_enhanced, image, df)
3. Incorrect data types in retweet_count and favorite_count in df table (should be integers, not float)
4. Standard denominator value (twitter_enhanced) is 10, others should be investigated and corrected
5. Missing values in the dog_class (twitter_enhanced)
6. Duplicated values in jpg_url column (Image)

7. Rating_numrator (twitter_enhanced) - values such as 1776, 666, 960, 420 are high unlikely
8. Non-descriptive column names in image table
9. Inconsistent format in twitter_enhaned name column (first letter should be in capital letter). Same for first, second, and third predictions in image table
10. Iphone seems to be the major source of the data (in twitter_enhanced and df table), the rest cannot be properly interpreted

**Data Tidiness Issues**

1. "doggo","floofer","pupper","puppo", should be melted into a single column (Twitter_enhanced)
2. Source and tweet columns duplicated in twitter_enhanced and df table.
3. Created_at (df) - day of the week should be separated from the time of occurrence. Created_at should be dropped as well to prevent date duplication (in tw_enhanced and df)
4. Separate date from the hours,minutes in twitter_enhanced
5. Let tweet_id be the first column in the df table
6. Tweet_id in twitter_enhanced duplicated in the image and df tables
7. Tweet_JSON (df) shoud be part of twitter_enhanced. Infact, if possible, all three should be combined into one.

### *Storing the Data*

The three datasets were merged together using the merge() function on the "tweet_id" column (the only column common to all three) to create a master csv file named twitter_archive_master.csv

## 4.0 CONCLUSION

A skilled data wrangler is able to gather data from many sources, handle data quality and tidiness issues, and able to transform, manipulate to data to generate insightful findings. I can say, I accomplished the same in this project using all the wonderful Python libraries. I thoroughly enjoyed this project as it drove to me some uncomfortable zones but I went through it all and I'm now proud to say I'm better skilled and equipped to take up a role as a data analyst.  Thanks to **ALX** and the **Udacity** team for this opportunity.