**WRANGLE ACT REPORT: On WeRateDogs Twitter Archive**

This is the wrangle act report of the tweeter act of WeRateDogs. WeRateDogs is a tweeter account that rates people’s dogs based on their pictures, videos, breeds and names. These ratings where gathered, cleaned, assessed and visualized.

Three data source where gathered to achieve this report. This was how the report where gotten

1. **Gathering of Data**
2. The first file I download was the (twitter\_archive\_enhanced.csv). After I downloaded it, I uploaded it and read the data into a pandas DataFrame.
3. The second one was the (image\_predictions.tsv). this file was downloaded programmatically using my request library and URL code that was provide by Udacity
4. The third files were, tweet\_json.txt and twitter\_api.py, which I download from the link provided in my Udacity classroom.

After I have gathered these files, I uploaded them in my Jupiter notebook and imported the libraries I will be using for this project.

1. **Assessing Of Data**

After gathering all three pieces of data, I assessed them both visually and programmatically to check for quality and tidiness issues. I then went ahead to detect and document  **eight (8) quality issues** and **two (2) tidiness issues** in the "Accessing Data" section in the wrangle\_act.ipynb of my Jupiter Notebook

Here are some of the quality and tidiness issues I assessed

**Quality Issues**

1. They were some incorrect datatypes on tweet\_id and time stamps

2. They were 181 entries in (retweeted\_status\_id, reweeted\_status\_user and reweeted\_status\_timestamp) instead of 2356 entries

3. They were 78 entries in (reply\_to\_status\_id and reply\_to\_user\_id) making it to have about 2278 missing entries.

4. The expanded\_urls had some missing values too. It had just 2297 entries instead of 2356

5. Some of the breeds of the dogs where not properly capitalized in p1, p2 and p3.

6. Some of the tweets where retweeted, making must of them not original tweets

7. Some of the tweets had missing images

8. Both the rating\_numerator and rating\_denorminator had incorrect ratings

9. Some of the tweets in the name column valid name

# **Tidiness Issues**

1. The three different data set had one (1) common column, which is "tweet\_id" they all need to be merged and put in just one dataset.

2. There are four different columns for dog names (doggo, floofer, pupper and puppo) and they ought to be put in a single column

1. **Cleaning of Data**

In this part, I first made copies of each of the following data files (twitter\_archive\_enhanced.csv, image\_predictions.tsv and tweet\_json.txt) and merge them in single data file called (df\_merge)

I did this to enable me clean, analyze and visualize them properly.

I handle all the issues raised from the quality issue to the tidiness issues. I used this following step by

Defining the issue, writing codes to handle the issue, and then testing my codes by running them on my Jupiter notebook to check if it runs properly without errors.

**The last step was storing my data, analyzing and visualization them.**

**Storing**: I stored my initial file as (df\_merge as twitter\_archive\_master.csv)

**Analyzing**: I ran this new stored file (twitter\_archive\_master) by checking the file using some of this codes (.head(), .info(), .describe()) .

**Visualizing**: I created some tables and used statistical analyzation like bar chat, pie chat to answer some of the questions.