## Bashair Altalhi

# Wrangle Report

In this project, I used python and its libraries to gather data from different sources and formats. Then, assess its quality and tidiness, after that clean it, analyses, and visualizations.

### Data Gathering:

We used three types of data.

- **Twitter archive**

It is a csv file that is provided in the project. I upload it to Jupter Notebook by using pandas.read_csv.

- **Image predictions**

It is a tsv file that is provided in the project. I read the file with open as file then upload it by using pandas.

- **Twitter json**

It is a txt file. I download the data from Project page and open it with pandas.

### Assessing:

- **Visual assessment:**

I explored the data by looking at each column and row by excel and python. This step can be done by display each data.

- **Programmatic Assessments:**

I investigated the data deeply by using different commands such as describe() that display a statistical summary. Also, the command .sample() which illustrate a sample of data. The command .duplicated() to check for duplicate in information. Also, the command .info() shows the number of non-null count and the type of each row.

Then, I documented some of the quality issues and tidiness issues of the three data.

### Quality:

- **Twitter archive data**
  - Missing value in the following:
    - 'in_reply_to_status' the total of values are 78 and it is supposed to be 2356
    - 'in_reply_to_user_id' the total of values are 78 and it is supposed to be 2356.

        - 'retweted_status_id' the total of values are 181 and it is supposed to be 2356.
        - 'retweted_status_user_id' the total of values are 181 and it is supposed to be 2356.
        - 'retweted_status_timestamp' the total of values are 181 and it is supposed to be 2356.
        - 'expanded_urls' the total of values are 2297 and it is supposed to be 2356.
    - 'timestamp' is object not datetime.
    - 'retwetted_status_timestamp' is object not datetime (removed later no need to change).
    - ' name' some names are invalida 'Nane', 'a', 'the', 'an', 'very', just', 'quit', 'one','not', 'mad', 'getting', 'actually', 'unacceptable','this', 'old', 'space', 'such', 'light', 'infuriating', 'incresibly', 'all', 'by', 'my', 'his', 'officially', and 'life'.
    - 'rating_denominator' type is int not float.
    - 'dog_stages' should be catagory.
    - 'rating_numerator' type is int not float.
    - 'rating_numerator' has incorrect ratings.
        - Tweet ID 810984652412424192. 24/7 isn't a valid rating it should not be any rating.
        - Tweet ID 835246439529840640. 960/00 isn't a valid rating should be 13 and 10.
        - Tweet ID 883482846933004288 5/10 but it should be 13.5/10.
        - Tweet ID 832215909146226688 and 786709082849828864 75/10 it should be 9.75/10.
        - Tweet ID 778027034220126208 27/10 it should be 11.27/10.
        - Tweet ID 681340665377193984 5/10 it should be 9.5/10.
        - Tweet ID 680494726643068929 26/10 it should be 11.26/10.
- **Twitter json**
    - 'id' is int not object.
    - 'retweet_count' and 'favorite_count' is floats not integers.
- **Image predictions**
    - 'id' is int not object.

## Tidiness:

- 'puppo', 'doggo', 'floofer', 'pupper' are dog "stage" should be in one cloumn.
- Merge df_archive, df_image, and df_tweets into one by tweeter Id.

## Cleaning:

In this section of the project I cleaned each issue by first define the issue then write the code to clean it finally test if the code worked.

## Analyzing, and visualizing the wrangled data:

The last step of the project after storing the wrangled data, I asked three question and answered it by visualizing.