**Data Wrangling Project Report (Case study Twitter WeRateDogs Post)**

The Objectives of this project is to Gather, Access, Clean and Visualize to analysis the twitter page of WeRateDogs. WeRateDogs is a twitter account that rates dog’s people’s dog with everyone commenting on the dogs.

Below are the processes carried out in achieving the objectives of this projects:

I downloaded the csv file of twitter\_archive on Udacity website which was read on Jupiter note using pandas.

I also download image prediction data from this url <https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv>' using request.get method.

The third method of gathering used was the use of Twitter API which I downloaded the JSON API files what was uploaded by Udacity and which was uploaded on my Jupiter notebook and read the file.

During my assessment I discover a lot of data quality problems and tidiness which makes the data to be very dirty and messy which I had to seriously clean the data for it to be good for the analysis. Some of the cleaning I did was Missing data, duplicated data, incorrect data type, inaccurate data, and some columns that are not needed for the analysis had to be dropped. After cleaning the dataset had to be merge together so as to have a proper tidy dataset.

Some of the functions I used in accessing the dataset programmatically are

1. unique()
2. isnull()
3. duplicated()
4. dtypes()
5. describe()
6. info()
7. shape()
8. list()

Also, some of the function used for the cleaning are as follows:

1. drop() which was used to drop some column that are not needed for the analysis
2. dropna() which was used to drop the missing data’s
3. pd.to\_datetime() which was used to convert the date data type from object to datetime format.
4. rename() which was used to rename some column for proper column description
5. insert() which was used to move some column for proper constituency
6. merge() which was used to merge the 3 dataset together for proper tidy format
7. astype(float) which was used to change the data type to float
8. sort\_value() which was used to sort the dataset
9. drop\_duplicated() it was used to drop the duplicated values in the dataset
10. melt() it was used to merge and create another column for dog stage

Then visualization was made so as to analysis the dataset using some insight that was derived from the dataset using python. It was discovered from my relationship plot that there is a strong correlation between the 3-prediction dog and their prediction confidence as well as favorite count and retweet count. Also, it was discovered that the higher the rating number the higher the favorite count and the same thing applies to retweet count. Retweet count also increase slowly due to the more the people get use to the channel the slower the retweet count.