Data Wrangling Report:

About Dataset:

The Dataset I wrangled is a twitter archive of a user known as **WeRateDogs** who rates the dogs with a amusing comment. They dont rate dogs with a specific fixed denominator , most of the times the rating numerator is more than denominator. this data consists of 2356 rows and it consists of all the tweet data from November 2015 to August 2017.

Based on the above images data we have another Dataset which contains the image predictions along with tweet ID and image URL and image number. It has top three predictions of dog breeds along with respective confidence in those predictions for every tweet.

Data Gathering:

Gathering from a csv file:

I just imported the **csv** file which is a twitter archive of WeRateDogs.I manually downloaded it from udacity projects lesson.The name of the file is **twitter_archive_enhanced.csv**.

Gathering from a server:

I downloaded the image predictions tsv file from the server using python **requests** library .the link to file hosted in Udacity server is https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv I loaded the data from the tsv file into a dataframe using **read_csv()** and **sep** attribute of that function.The name of this file is **image_predictions.tsv**

Gathering using Twitter API:

I gathered the required twitter data from **Twitter API** using **tweepy** access library in python and the tweet ID's from other the **twitter_archive_enhanced.csv** . By default it is JSON data and stored it in **tweet_json.txt** using **json** library in python. and extracted the attributes(retweet_count, favorite_count and tweet ID) data of all the tweets from their respective tweet objects.

Data Assessment:

Visual Assessment:

I went through the whole DataFrame visually and found some issues in some datasets.

Here are the issues I found in each dataset visually:

twitter_enhanced:

- None values in **name** column.
- hyperlinks in source column.
- None values in all the **doggo**, **floofer**, **pupper**, **puppo** columns.
- Multiple columns for dogstage__(tidyness issue)__

- Retweets are Present in the Dataset.

image_pred:

- Multiple DogBreeds are possible__(tidyness issue)__
- Unnecessary columns for analysis
- Dog breed names starting with small case alphabets

Programatic Assessment:

I assessed all the dataframes using some pandas and numpy functions and found some more issues in the datasets.

Here are the issues I found in each dataset programatically:

twitter_enhanced:

- Erroneous names in **name** column.
- incorrect Numerators and Denominators for some of the tweets
- Datatypes of **tweet_id**,**in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id**.

image_pred:

- Datatype of **tweet_id** is Int

Cleaning Dataset:

As a first step I made copies of all the three original datasets so as to enable us to access the original datsets wherever required.I followed a organised **Define,Code and Test** programmatical clean process.

Storing Data:

I joined all the dataframes performing **"Inner Join"** on tweet_id programmatically using **merge()** function.Then, I stored the cleaned dataframe in a csv file named **twitter_archive_master.csv**