Data Gathering

We collected our data from three different resources:

1- CSV file :Collected from the twitter data base and downloaded from Udacity classroom
2- TSV file: Created a file from the URL provided and imported the data in TSV data frame.
3- json file: Collected Twitter API data for the Retweet and Favorite count provided by Udacity

We used multiple techniques to import our data in the following manner:

1- CSV: opened the file by using the pandas read_csv function
2- TSV: downloaded the data from the provided url using the requests library and created and TSV file and then used pandas read_csv function and changed separator
3- json: Created a list and started a loop to extract the required data from the json file using the os library commands and the transformed the list into a pandas dataframe

Output:

The output file are:

1- CSV: tweets_df
2- TSV: img_df
3- Json: api_df

Data Assessment:

Started with some visual assessment and then run programmatic assessment

Visual assessment:

Using the MS Excel to take an overview of the data

Programmatic assessment:

Using the jupyter notebook importing the pandas,re and numpy libraries using their main function to display the data and recheck on the visual assessment output and introduced new output.

Main functions used:

1- Df.info() for showing the general info for the dataframe
2- Df.describe() for having a general descriptive view over the dataframe
3- Df.value_counts() for having the number of values repeated and assessing the validity of these values
4- Re.findall(pat) using the regulae expression method to extract the required data

Output:

The quality issues found in the datasets are :

- rating numenator (over range values like 1776, has 2 zero value, has 438 values X<10 and 25 values X>20)
- rating denominator(over range values like 170 and 23 values != 10)
- most dogs stages has a value None
- some dogs has multiple dog stages need to be combined
- Dogs names show some typo errors and none values
- clean all the images with no dog in
- Drop all entires with no images
- Removing the false predictions

The untidy data isssues found in datasets are :

- all dog stages must be in one column under the name of dog stage
- remove all the retweets and replies in the dataframe
- tiding the the img_df by resizing the dataframe size in columns

Data Cleaning:

- **What is the structure of the this process (work flow):**
    1- Started by defining the issue to work on
    2- Start coding the dataframe to enhance its output
    3- Testing the results outcome.

- **What are the techniques used to fix the issues?**

Table Name	Quality Issues	Solution
Archive_df	○ **Wrong values shown (extraction got the wrong numbers)** ○ **Float values** ○ **Some tweets had combined dog ratings** ○ **Dog stages was including a None value** ○ **Some dog names were extracted by the wrong means and some had a none value**	○ **Extracted the right values** ○ **Changed the column type to float** ○ **Divided the rating over the number of dogs** ○ **Replaced all none values with NaN** ○ **Extracted the dog names and replaced post with no name with NaN**
Img_df	○ **Some entries had no images** ○ **false predictions**	○ **Removed all entries with no images** ○ **Removed all false predections**

Table Name	Tidiness Issues	Solution
All tables	**Retweets and replies**	**Removed all retweets and replies**
Archive_df	**dog stages in multiple columns**	**Made one column**
Img_df	**Multiple columns for prediction**	**Made one column**

Output:

The output files was three tables: archive_clean, img_clean, api_clean are collected into one data frame (archive_master) and made as CSV