Introduction Of project:

The dataset used in the wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Data Wrangling of this project consists of:

1. **Gathering Data**- Extract data from different sources
   a. Loaded data from csv file.
   b. Using request library to download tsv file hosted on Udacity server.
   c. Using Tweepy to get data from Twitter's API.

2. **Accessing data**- After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues as shown below:

Quality Issues:

1. As there are 78 records for " in_reply_to_status_id"   and "in_reply_to_user_id" columns which is very less in count , so both columns are not helpful in analysis, that's why drop these columns.

2. All columns related to 'retweets' contains only 181 records ,so they are not helpful in analysis, that's why drop these columns.

3. Change datatype of timestamp from object to datetime

4. Extract the string between from Source field between a href tag.

5. Missing value in expanded_urls columns

6. Exclude zero values from the rating_denuminator.

7. Replace the value 'None' with the NaN (missing value), so that it become easy to find missing values directly by using built-in functions

8. Missing value in name field..

9. Change the column names of image prediction table from p1 ,p2,p3 to prediction_1,prediction_1,prediction_1 respectively.

Tidiness Issues:

1. change last 4 columns 'doggo', 'floofer','pupper', 'puppo' into one column as dog_stage

2. Merge all the tables to make only one master table.

3. Cleaning Data:

Fixed both quality and tidiness issues as described while assessing data using python functions. Each issue is resolved by following the below steps;

a. **Define –** Define the issue in words properly.

b. **Code**- Write a code to fix the issue.

c. **Test –** Check if the fixed is done or not.

4. **Storing, Analyzing and Visualizing:**

Stored the clean DataFrame in CSV file with the main one named tweets_master_data.csv. After that prepared the insights and visualization in act_report.html