Wrangling Report¶

Note: This is the report of my wrangling effort My definition of data wrangling is like doing a chore. Imagine I have a messy and dirty room. What do I need first? I get a box to gather all my items. I am looking around the room and detects the clutters and dirt to plan. Then, I organize them by dividing my clothes, papers, books, and electronics.
In this project, I was doing chores on the data with my computer. Three things that I am doing is to gather, assess, and clean data for the purpose to complete my assignment at Data Analysis Udacity Nanodegree.

Introduction:¶

I wrangled the data from a popular Twitter account called, WeRateDogs, where people check or share cute and funny pictures of their dogs. This report consisted of the details of my wrangling effort for this project.
https://twitter.com/dog_rates?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eautho r

Gather :¶

There are three files that I have gathered.

Twitter_Archive : I received the twitter_archive.csv from Udacity through this assignment and can get access through pandas.

Images-prediction : is downloaded programmatically by myself. It is also provided or hosted from Udacity server. All I need is to use 'request' library to scrap it.

Tweet_json.txt : First, I have queried Twitter API with the keys and code pass provided by Twitter developers website. The second is that I filter between no_tweet and yes_tweet to gather only *the tweets with tweet_id*. The third is append the list of tweets with tweet_id in a dictionary --meaning I converted the list into a dictionary. The fourth step is to open a new Json file called tweet_json.txt. Then, dump the dictionary in tweet_json.txt. In the last step, open and load the tweet_json.txt. Then create the dataframe in the tweet_json.txt with the help of the loop and collect only 'tweet_id,' 'favorite_count,' and 'retweet_count' columns to make them appear in the table.

Assess:¶

When I begin to assess the data, I usually look through Jupiter Notebook and a little bit of Excel. When it comes to coding, I have checked with the name of data frames along with info, counts, drop, and more. All I did is to check or find anything that looks wrong, unappealing, dirty, and messy in my twitter_archive file and images-prediction tables. --Not to forget, I took few notes on what I need to clean next.

Clean:

Before I begin, I created a copy of my two data frames: ta_clean and images_clean for not messing up from the original. Also, I added ‘define’ to indicate what I need to do. Additionally, I included ‘code’ to show my process to clean up my data. Then, ‘test’ to demonstrate the final appearance of my two tables. I followed through my ‘Assess’ notes on what to clean. I did delete a few inappropriate data or unnecessary columns and converted few columns from float to object