Data Wrangling Report

The dataset:

The dataset 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.

The wrangling fallowed these tree steps:

I- Gathering_Data:

I gathered three pieces of data:

- Downloading the file " twitter_archive_master.csv " manually from the Udacity's classroom.

- Downloading the file " image_predictions.tsv " programmatically from the Udacity's servers using the Requests library.

- Gather each tweet's retweet count and favorite ("like") count from the "tweet_json.txt" file.

II, III – Assessing and Cleaning Data:

While working on this dataset, several observations were made :

- Quality issues:

Dataset	Issue	Solution
tweets_api	- Original tweets mixed with retweets.	- Remove rows that are retweets with drop() function.
t_archives	- The "name" column has None values.	- Drop rows in the "name" column that has None values with drop() function.
	- The "rating_denominator" column is not fixed at 10.	- Fix the "rating_denominator" column at 10.
	- The "timestamp" column is a string instead of a datetime type.	- Turn the "timestamp" column into a datetime with 'to_datetime()' function.
	- The URL in the "source" column is unclear.	- Only get the URL from the "source" column with str() and split() function.
	- The values in the "rating_numerator" column are not reasonible.	- Make the values in the "rating_numerator" column more reasonable.
image_predictions	- Some predictions are all false.	- Drop the predictions that are all false.
	- Some of the dog bread's names are lowercase, while others are uppercase.	- Make all of the dog bread's names lowercase with the str() function and lower() method function.

- Tidiness issues:

Dataset	Issue	Solution
t_archives	- The colomuns : 'retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp','in_reply_to_status_id', 'in_reply_to_user_id' has a lot of NaN values.	- Drop The colomuns : 'retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp','in_reply_to_status_id', 'in_reply_to_user_id'.
	- The dog stage names "doggo","floofer","pupper" and "puppo" each has a column.	- Melt the columns "doggo","floofer","pupper" and "puppo" into one column called "dog_name" using the melt function.