# Data Wrangling Report

**1. Gathering Data**

About the Dataset(s)

The dataset that you will be 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Data had been gathered as below sourced and loaded to pandas data frames:

1- Twitter archive data as the form of csv format (twitter-archive-enhanced.csv)

2- loaded Image Predictions File (image_predictions.tsv) also tried to automate loading but didn't complete

3- Additional Data via the Twitter API since there was issue in credential and account approval so i used tweet-json.txt as source to avoid delays however i added relevant portion of code for later execution

4-Created a dataFrame with tweet ID, retweet count, favorite count mainly to get followers

**2- Assess Data**

- Started Displaying data captured from data frames through samples
- Started checking metadata
- build some insights to capture as much as possible quality and tidiness issues

Data Wrangling Report

1. Gathering Data

About the Dataset(s)

The dataset that you will be 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. These ratings almost always have a denominator of 10. The

numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Data had been gathered as below sourced and loaded to pandas data frames:

1- Twitter archive data as the form of csv format (twitter-archive-enhanced.csv)

2- loaded Image Predictions File (image_predictions.tsv) also tried to automate loading but didn't complete

3- Additional Data via the Twitter API since there was issue in credential and account approval so i used tweet-json.txt as source to avoid delays however i added relevant portion of code for later execution

4-Created a dataFrame with tweet ID, retweet count, favorite count mainly to get followers

**3-Assess Data**

- Started Displaying data captured from data frames through samples
- Started checking metadata
- build some insights to capture as much as possible quality and tidiness issues

1-Work on 3 Data Quality issues related to datatypes

2-For tidness make sure that all tweets ids in archive clean consistent with image_df

3-Discrading 3 fields that will not lead to solid analysis as a results of lot's of missing values

4-Dropping retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp from metadata

5-Fill missing data to allow chaning metadata for reply_to_status_id & in_reply_to_user_id

6-Change in_reply_to_status_id & in_reply_to_user_id to integer type

7-Change timestamp to datetime data type

8-Exclude zero values from the numertor and denuminator ratings

9-Replace the value 'None' with the NaN to show that it is missing values for 4 coulmns doggo,floofer,pupper and puppo

10-Discrading additional 4th  field expanded_urls that will not lead to solid analysis currently

11-As part of tidness will include all dogs classifications doggo,floofer,pupper and puppo to be merged into one column  dog_classification

12-Convert the dog_classification datatype to categorical

13-Drop the all dogs classifications colmns : doggo, floofer, pupper and puppo

14-One more Data quality issue to Change Name data type string to be able to analyze archive_clean

15-for better tidiness will rename 6 columns to have better meaningful visibility in image_clean

16-for better tidness will rename id column to have better meaningful visibility in tweets_clean

17-One more Data quality issue to unity tweet if data type as string

**4-  Clean:**

- Create a copy of archive_df data to cleanup data (archive_clean)
- Create copy from image_df to cleanup data (image_clean)
- Create copy from tweets_df to cleanup data (tweets_clean)
- fixed majority of identified quality and tidness

**5-Store**

- Stored cleaned data for archive_df in archive_clean.csv
- Stored cleaned data for image_df in image_clean.csv
- Stored cleaned data for tweets_df in tweets_clean1-Work on 3 Data Quality issues related to datatypes