Wrangle Report

- This report describes the data wrangling processing which represent in 3 parts:
  1- Gathering data
  2- Assessing data
  3- Cleaning data

- About data: It is the tweet archive of Twitter user **@dog rates**, also known as **Waterdogs**. Waterdogs 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**." Waterdogs has over 4 million followers and has received international media coverage.

- First part is gathering data, I have gathered this data from 3 sources
  1- Csv file: The WeRateDogs Twitter archive
  2- From internet: The tweet image predictions
  3- From tweeter through API: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's **Tweepy** library and store each tweet's entire set of JSON data in a file called `tweet_json.txt` file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas Data Frame with (at minimum) tweet ID, retweet count, and favorite count.

- Second part is assessing data, this part divided into two ways
  1- Visual assessment: each piece of gathered data is displayed in the Jupiter Notebook for visual assessment purposes. Once displayed, data i assessed in an external application (Excel).
  2- Programmatic assessment: pandas' functions and/or methods are used to assess the data.

- Third part is cleaning data, I have cleaned the issues which I find when I was assess the data , this issues divide into 2 kinds

  1- Tidiness which was 2 issues:

  - doggo, floofer, pupper, and puppo should be in one column not 4
  - Data frames were divided into 3 tables

  2- quality wich was 14 issues:

  - tweet_id should be "string" not "int"
  - some values in rating_denominator column isn't "10"
  - some values in rating_numerator column less than "10"
  - some values in rating_numerator column = zero
  - timestamp should be "data time" not "str"
  - retweeted_status_id should be removed because we interest in tweet
  - retweeted_status_user_id should be removed because we interest in tweet
  - retweeted_status_timestamp should be removed because we interest in tweet
  - Nulls represented as (none) in name column
  - names p columns have some upper letter and some lower letter