Introduction:

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.

My tasks in this project are as follows:

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing, and visualizing data

Step 6: Reporting

Data wrangling is about gathering the right pieces of data, assessing your data's quality and structure, then modifying your data to make it clean. But the assessments you make and convert to cleaning operations won't make your analysis, visualization, or model better, though. The goal is to just make them possible, i.e., functional.

Gathering

In this phase of our project, we will collect data from three different files sources:

- First I gather **Twitter_Archive_enhanced.csv**,This archive contains basic tweet data (tweet ID, timestamp, text, etc.). It extracted programmatically.
- Then collected data from **image_predictions.tsv** is present in each tweet according to a neural network. It is hosted on Udacity's servers and should be downloaded programmatically using the Requests library.

Variable Name	Definition
tweet_id	the last part of the tweet URL after "status/"
p1	the algorithm's #1 prediction for the image in the tweet
p1_conf	how confident the algorithm is in its #1 prediction
p1_dog	whether or not the #1 prediction is a breed of dog
p2	the algorithm's #2 prediction for the image in the tweet
p2_conf	how confident the algorithm is in its #2 prediction
p2_dog	whether or not the #2 prediction is a breed of dog
p3	the algorithm's #3 prediction for the image in the tweet
p3_conf	how confident the algorithm is in its #3 prediction
p3_dog	whether or not the #3 prediction is a breed of dog

- Last part was **tweet_json.txt,** it gathers each tweet's retweet count and favorite ("like") count at the 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 file. using Python’s tweepy library.

Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. Note: do not include your Twitter API keys, secrets, and tokens in your project submission.

After we completed Gathering step and make sure that all collected data are loaded successfully, we go to the second step which is assessment.

Assessing

Assessing is the second step in the data wrangling process. Assessment phase divided into two main parts

- **Data quality issues (dirty data):** Data that has **quality issues** have issues with content like missing, duplicate, or incorrect data. This is called *dirty data*.

- **Lack of tidiness (messy data):** Data that has specific **structural issues** that slow you down when cleaning and analyzing, visualizing, or modeling your data later.

I search for the issues by two ways:

- Visually by scrolling: I just opened the three files looked through the data in its entirety using pandas and spreadsheet.
- Programmatically using code:I used some methods such as .info, .shape , .

cleaning

Quality issues

Twitter _archive file

- redundant retweets rows (181 retweeted_status_id,181retweeted_status_user _id and 181 retweeted_status_timestamp)
- redundant "in reply to users' tweet" rows (78 in_reply_to_status_id and 78 in _reply_to_user_id)
- tweet_id must be string instead of integer
- timestamp must be in datetime format instead of string and any timestamp later than August 1st, 2017 should be removed.
- Some names are wrong and some have values of 'None' instead of NaN.
- Some rating denominators do not equal to 10.
- Some rating numerators are unacceptable
- Removing html code from source.

image _predictions

- Replace underscore with space in p1, p2, and p3
- Remove tweets in predictions where all probabilities of p_dog are False, where the image might not be a dog.
- The columns p1, p2, p3 in predictions are inconsistent in their capitalization.

Tidiness issues

Twitter _archive file

- There are 4 different columns (doggo, floofer, pupper, and puppo) for dog stages. The different dataframes should be merged into a single one.
- All three dataframes should be merged into one clean dataframe since they all hold information about the same entity; tweet.

Tweets_data

- Rename the column id to be tweet_id to facilitate merging.

After the whole wrangling data steps has been finished, storing the data is start.

We store the cleaned table in new csv file called **twitter_archive_master.csv.**