Introduction:

This report is for step-wise description of Udacity Data Analyze Nano-Degree data wrangling project. Using Jupyter Notebook with Python 3, I gathered, assess, and clean data from CSV, URL, and JSON API sources to perform analysis on Twitter account: WeRateDogs. I will explain how I perform my data wrangling act for each step.

Preparation:

I imported all the modules/ package I need for this project: pandas, numpy, requests, tweepy, json, functools, matplotlib.pyplot, seaborn.

Gather:

CSV:

Gather from CSV file that I downloaded from Udacity project page and uploaded in the Jupyter Notebook.

URL:

Use requests module to open URL and save the content as tsv file.

JSON API:

Registered an developer account on Twitter, use Tweepy to get all Json information and saved them in tweet_json.txt file, then extracted 'id', 'retweet_count', 'favorite_count',' retweeted_status' columns from the file and save them as a data frame.

Assess:

Performed info, sample, unique, and head methods to assess these 3 data frames.

For Quality issues, I found that:

twitter_archive_enhanced.csv:

- in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id, retweeted_status_user_id should be object instead of float
- timestamp, retweeted_status_timestamp should be timestamp instead of object
- inconsistence with null object: none and NaN
- invalid names: 'a','not','one','an','quiet','very','my','his','unacceptable','this','all','old','the','by'
- Some records have retweet_status_id, need to exclude

image_predictions.tsv:

- for the sake of consistency, p1, p2, p3 format need to be all lowered cases

tweet_json:

- id should be tweet_id
- some of the retweet_status is a link, need to be exclude

For Tidiness Issues, I found out that:

twitter_archive_enhanced.csv:

- doggo, floofer, pupper, and puppo should reshape as one column under "type" instead of 3 columns

Overall:

- should compile these 3 tables as 1 table

Clean:

- Made copies for each data frame as tae_clean, ip_clean, df_tweets_clean.
- Performed clean act by using astype, to_datetime, replace('None',np.nan), notnull, rename, str.extract, drop, reduce methods to clean the previously mentioned problems. Later used info, sample, unique, and head methods to test the results.
- Again, I iterated in the process of data wrangling.

Store:

I saved the cleaned master data frame as twitter_archive_master.csv by using to_csv method.

Analyze & Visualize:

I used matplotlib.pyplot and seaborn modules to plot the data in three sections.

- For the first section, I create a data frame called df_rftime, containing 'timestamp', 'retweet_count', 'favorite_count'. Plotted ‘Retweet Counts by Timestamp and Favorite Counts by Timestamp’ and 'Favorite Counts by Timestamp' graphs.
- For the second section, I created a data frame, df_rfratio, containing 'retweet_count', 'favorite_count', 'rating_numerator', 'rating_denominator'. Then I calculated the ratio of rating and dropped the numerator and denominator columns. Plotted ‘Retweets and Favorite Counts Colored by Rating Ratio’ with xlimit 0 to 10,000 and ylimit 0 to 30,000, colored by rating ratio.
- For the last section, I created a data frame, df_rftxt, containing 'text', 'retweet_count', 'favorite_count'. Then calculated the text length beyond 10 words, stored as a new column named ‘ text_length’. Plotted 'Retweets and Favorite Counts Colored by Text Length with 10 or More Words’, colored by ‘text_length’ .

Reflection:

This is a very challenging project for me. I had no experience with Pandas before, since I started from Term 2. By doing researches on githubs, stackflow, and documentations along with asking my Udacity mentors, I successively solved the problem and finished the project. From this project, I gained a better sense of data wrangling using Python with different modules and the process of analysis from rudimentary state.