This report describes the wrangling efforts involved in completing the "WeRateDogs" project as part of Udacity's Data Analysis professional Nanodegree. With FWD-Egypt
The Data Wrangling process consists of:
1. Gathering the data
2. Assessing the data
3. Cleaning the data

1. Gathering
Gathering Data for this Project involved obtaining three different datasets from three different sources. Each one testing a different way to obtaining a dataset.
The first was to download a file manually and be able to open a csv file("twitter-archive-enhanced.csv). the file consisting of the largest amount of data.
The second was to be able to download a file programmatically using Python Requests library. The file contained image predictions on the breed of the dog coming from a neural network on some of the tweets already downloaded in the archive file. The file was in tsv format and tested your ability to open this type of file successfully.

The third data is the resulting data from twitter_api.py. I proceed with the following part of "Gathering Data for this Project" on the tweet-json.txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count."

I've the code ability to query Twitter's API and use a Python library called Tweepy to obtain further data on the tweets in the archive file using the tweet id. The Tweepy library returned the data in json format, from which it was possible to iterate through and append data to a file

2. Assessing the data
The three saved data frames were then assessed visually inside a jupyter notebook with pandas.That allow me to quick scanning it and filter it and i do it with programming assesment also inside jupyter with pandas functions ,df.info(),df.head(),df.describe(),df.value_counts
and this process was iterator.
i found qouality and tidiness issue:
quality Issue:
remove notnull in retweeted_status_id / in_reply_to_status_id
(remove columns from archive_df:
in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp)
fix the "None" string issue in the those entries by replacing it with empty string ""
convert timestamp to datetime
classfication of dog to be readable by seprate it with "-"
"" values with Nan in dog_breed
filter archive_df_clean
filter api_df_clean
filter image_df_clean
in archive_df_clean , image_df_clean , api_df_clean reset index
type float and also it should be correctly extracted (rating_numerator) in archive_df_clean

extract source in archive_df_clean

Tidiness Issue:
classfication of dog: doggo,floofer,pupper,puppo in one column(Column headers are values, not variable names.)
tidiness issue Column headers are values, not variable names prediction_1 ,confidence_1,breed_1 prediction_2, confidence_2, breed_2 prediction_3, confidence_3, breed_3
archive_df_clean, api_df_clean in one table to solve tidiness issue A single observational unit is stored in multiple tables.

3. Cleaning the data
firstly i made a copy from my data set .The cleaning followed the standard process of define, code and test for each of the issues.All of the cleaning was performed using programmatic tools, such as def functions or pandas built-in functions (wide_to_long, regex,replace ,drop,extract etc), and i test all of this issue.

Conclusion
Data wrangling provides a clean data frame for future analysis and visualization, in our case we concluded with the "twitter_archive_master.csv" "image_df_clean.csv". Those files can also be shared with others without having to wrangle the data.