Wrange Report

1. Data Gathering

The wrangling active began by gathering different but related data from three different sources to successfully carry out the projects as given. The project was related to dog rating on Twitter. The first data was twitter_archive_enhanced.csv that first downloaded and loaded into the iPython notebook workspace. The second data was requested using *request library* on the Udacity website as provided in the project description and the final end-product of the data is saved and loaded into Jupyter workspace as "image_predictions.tsv". And finally, Tweepy library was used to query third data from Twitter with the help of Twitter API and were further processed and read in the workspace as tweet-json.csv.
This brings me to the end of data gathering.

2. Data Assessment

After the gathering was concluded and part of second data was embedded in third data which is tweet-json.csv, this data set was both used for visual data assessment and programmatical data assessment to identify the messy and dirty part of the data, that is quality issues and tidiness issues.
Assessing the data for quality, I was able to identify that what made the data to be of low quality and/or dirty were:

a. Via visual assessment, entities and extended entities are more or less the same. That is, one of them is redundant.
b. Via programmatical assessment, coordinates have all its values as missing values
c. Via programmatical assessment, contributors have all its values as missing values
d. Via programmatical assessment, created_at columns should be datetime neither object not text
e. Via visual assessment, retweeted, retweet count and status are redundant and not needed
f. Via itereirative programmatical assessment, there are overlapping columns between the tweet data table and user JSON columns.
g. Some tweets were identified to be retweets and they are not part of the original data.

In my attempt to ensure that the data is quite fit for further analysis, the tidiness issues that were pinpointed are:

a. Via visual assessment, entities columns consist of many columns that need to be break into individual columns or be separated.
b. Via visual and programmatical assessment, the three data set should form a single observational unit and thus need to be combined.
c. Via visual assessment of twitter archive data, field names like doggo, floofer, pupper and puppo ought to be under one field name instead of three

3. Data Cleaning

In this section, I cleaned all the issues in terms of quality and tidiness documented while assessing the data using visual and programmatical assessment techniques.

I first made a copy of the original data before cleaning and begun by then flatten out the entities columns of JSON object to have individual columns. But before doing that, I created a function that flatten the JSON column objects into their individual columns.

I then used the *flatten function* created to resolve issue 1 and move to issue 2 which was removing all columns that had no single or barely contained any meaningful in them. After that, I moved to another issue by convert the created_at date to proper Python date while modifying lang column to its full language name in the columns. I then added prefix to potential overlapping columns that might bring about any errors in collapsing any JSON column into meaningful individual columns which was last completed as final issue to be resolved. And the end, some columns and rows data were removed before all three data set were merged into a single observational unit.