Wrangling Report

The dataset that will be wrangled 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.

The Data Wrangling consists of three main steps:

1) Data Gathering
2) Data Assessment
3) Data Cleaning

A) Gathering the data:

There are 3 files that is needed to be downloaded and inserted into pandas DataFrame for further analysis

I. **twitter-archive-enhanced.csv:** is download manually from udacity resource files presented in the project

II. **image_predictions.tsv :** is downloaded programmatically using the link provided in the course .. this should be done using the requests library

III. **tweet_json.txt :** 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

B-C) Data Assessment and Cleaning :

Data Assessment Issue	Cleaning Strategy
retweeted_status_id and **in_reply_to_status_id** contains values which represents a retweet or reply.	delete the rows that represents a retweet or reply from the DataFrame so that only original tweets is subjected to analysis
Unnecessary columns for assessment: **in_reply_to_status_id,in_reply_to_user_id ,retweeted_status_user_id, retweeted_status_timestamp.**	Remove these columns using pandas.drop() method to drop these columns with additional argument inplace=True
Incorrect datatype of some columns	Change dtype in tweet_id column from int to string and timestamp column from object to datetime
multiple columns [doggo, floofer, pupper, puppo] in twitter archive tables that has its values duplicated and present in both rows and columns	Use pd.melt() method with these 4 columns as value_vars and drop the duplicate values based on id_tweet using drop_duplicates() method
rating_denominator column has values of more than 10 .. conversion of all values to 10 for better assessment is required instead of deleting the data	Multiplication of each value with a factor to convert all values in this column to 10 for better assessment of values
image_prediction DataFrame : the prediction values is presented in more than one column that is hard to assess.	Combine all these values on one columns using np.where method for conditional extraction of all dogs image prediction that is True
All Three DataFrames have tweet_id columns.	Use pd.merge method to merge all 3 dataframes into one dataframe.

After Data Wrangling, the file is ready to perform EDA (Exploratory Data Analysis) on it.