Wrangle report

Data collection:

The first phase is to gather data from 3 different sources, twitter_archive_enhanced.csv is WeRateDogs Twitter archive data downloaded directly and its type is csv, image_predictions.tsv a tsv file containing the tweet image prediction and downloaded by using requests library, lastly tweet_json.txt which is a json text file has additional data of tweets and obtained via the twitter API by using Tweepy library.

Assessing Data:

In this phase we use both programmatic and visual assessment, visually we can see missing values in the first table twitter_archive, the expanded_urls format is "https://twitter.com/dog_rates/status/ "+ tweet_id, since we have the id already, we can dispose of this column. Programmatically we find that the source column has four unique values included in html tag hence it is irrelevant, 181 statuses are retweeted, erroneous data type in several columns mainly the timestamp column who also has "+0000" at the end of each timestamp, and NaN values are considered as "None" string.

All issues mentioned above are quality issues, the table additional_tweet also have erroneous datatype, unneeded retweets, and a number of empty columns. As for the tidiness issues they are generally located in additional_tweet: a break of the rule one variable one column for instance entities column elements are json type and contain either repeated or irrelevant elements, the second tidiness issue is that there are several columns have the same contents but in different datatype, finally the tweet_id label must be the same for all tables to be able to use them lately.

Cleaning Data:

Now after assessment we make a copy of each table, in order to clean our dataset, we begin by removing retweeted statuses of the table twitter_archive_clean, then removing empty and incomplete columns, the same process is done in additional_tweet_clean in which we fix the tidiness issues by deleting repeated or irrelevant columns, then we rename the columns id to tweet_id and display_text_range to text_length to change later its format to integers instead of a list of range.

The next step in twitter_archive_clean is to format and clean the timestamp at the same time replacing "None" strings by NaN values, after that we merge the two dataframes twitter_archive_clean and additional_tweet_clean by their column tweet_id into one dataframe new_twitter, then as a final modification we remove incomplete columns and leaving only rows in common between new_twitter and image_predic_clean dataframes, plus resting index and changing the type of tweet_id to string.