Project: Wrangling and Analyze Data

- I started by importing the necessary datasets from 3 different sources. The first one is twitter_archive which is a CSV file. The second one is image_predictions which is a TSV file. The third one is additional data from twitter API as JSON data in a TXT file. After that, I started to take a closer and deep look into the data in those 3 datasets. I checked their data types and the number of null values they have. Also, I selected only some columns that I will need in my analysis from JSON file. To make datatypes checking easier and faster, I created a function that shows a sample row of any dataframe and in the second row, it shows the datatype of each column. This function is helpful because it is useful for getting an overview about any dataframe and its datatypes without writing repetitive code for each dataframe. Then, I checked whether all denominator values are 10 as they should be. Moreover, I checked whether all numerators are greater than 10 as this is a rule in the We Rate Dogs twitter account. I found out that there are some wrong values in numerator and denominator columns so, I dropped the rows in which they are present. I also checked the values in image number column to make sure that there is no illogical value (such as any number less than 1 or float numbers). Then, I checked whether there are duplicated rows in the JSON dataframe. I found that there are no duplicated rows in the JSON dataframe.

- By using both of the programmatic assessment and visual assessment, I detected 14 issues in the datasets which are stated below:

- Quality issues

In twitter_archive dataset

1. timestamp column has 'object' data type
2. tweet_id column has 'integer' data type
3. in_reply_to_status_id column has 'float' data type
4. in_reply_to_user_id column has 'float' data type
5. retweeted_status_id column has 'float' data type
6. retweeted_status_user_id column has 'float' data type
7. rating_denominator column has values other than 10
8. rating_numerator column has values less than 10
9. doggo, floofer, pupper and puppo columns have null values that are considered actual values because they are written as 'None' which is shown as a value instead of empty value.

In image_predictions dataset

10. tweet_id column has 'integer' data type

In json_data dataset

11. id column has 'integer' data type

- Tidiness issues

In twitter_archive dataset

12. doggo, floofer, pupper and puppo columns are one variable (observational unit) as dog stage but it is spread across 4 columns

13. timestamp has year, month and day in one column

In general

- 14. Information about one type of observational unit (tweets) is spread across three different files/dataframes

● Here are the solutions I made to each issue:

First, I made a copy of each dataset I have (twitter archive, image_predictions and JSON data)

1. Converted timestamp column datatype to datetime
2. Converted tweet_id column datatype to string
3. Converted in_reply_to_status_id column datatype to string
4. Converted in_reply_to_user_id column datatype to string
5. Converted retweeted_status_id column datatype to string
6. Converted retweeted_status_user_id column datatype to string
7. Dropped rows that have denominators other than 10
8. Dropped rows that have numerators less than 10
9. Replaced 'None' in stage columns with empty string
10. Convered tweet_id column datatype to string
11. Converted id column datatype to string
12. combine stage columns in one column
13. Splitted the year, month and day in 3 separate columns then added them to the main dataframe. And converted their datatype to integer.
14. Created a master dataframe that contains the 3 datasets in one CSV file.

In addition, I created a CSV file that contains a dataframe that has the rows of tweets only (without retweets).