Data wrangling

At the very begging , I sought the requirements of the project in order to be able to follow the data wrangling steps : Data Gathering, Data Assessment, Data Cleaning.

I. Data Gathering

- I started by getting and saving image_predictions.tsv from the Udicity servers, then write its contents on the workplace to be read as a Pandas DataFrame.
- Secondly, I read the 'twitter-archive-enhanced.csv' as a pandas dataframe.
- Then, I sought using Tweepy, but got many errors regarding its use, so had to use the already attached 'tweet-json-copy.txt' file. I faced many issues when trying to load the whole file as a Json file, thus, I split the tweets and converted each one to a json object in order to read its data. After doing so , I saved their contents in a dictionary to be saved as a pandas dataframe.

II. Data assessment

I divided the data assessment into two main groups : quality issues and tidiness issues as taught.

Qulaity Issues

'twitter_archive'

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, and retweeted_status_user_id are float , and not complete.
- retweeted_status_timestamp and timestamp are string.
- retweeted_status_timestamp is not complete.
- rating_denominator is changing from 10.
- some pictures are for other animals , not dogs
- many data are NaN
- unecessary data of retweets

'image_predictions'

- there are 2075 entries instead of 2356
- 66 duplicate jpg_urlentries

Tidiness Issues

- drop unnecessary columns. i.e. retweet columns
- reduce four type columns doggo , floofer , pupper, and puppo into one column : type
- Merge the three dataframes together

III. Data Cleaning

After detecting the issues regarding the datasets, I initiated the data cleaning. Started by making copies of the datasets, and reassessing them quickly again. Since, I won't be 100% accurate when performing the assessment from the first time, I had to iterate over the data assessment again due to finding new issues while trying to clean the data. After attempting and cleaning the datasets whether on the quality or the tidiness sides, I saved the clean dataset on a .csv.

After which , I started data visualization in order to draw some conclusions from the clean data we have.