Data Wrangling Report

Data gathering

In this process, 3 different data were gather using different method of data gathering.

WeRateDogs Twitter archive

This data was manually downloaded through a link provided by the tutor. The data was then uploaded and read into a pandas DataFrame using the traditional pd.read_csv() method in pandas.

The tweet image predictions

This file (image_predictions.tsv) is present in each tweet according to a neural network. It is hosted on Udacity's servers. The data was downloaded programmatically using the Requests library and the supplied URL. This data was also read into a pandas DataFrame using the traditional pd.read_csv(). However, it is important to note that since the data is a 'tsv' file, the delimiter was specified in the process.

Additional data from the Twitter API

This data contains each tweet's retweet count and favorite ("like") count and can be gathered by using the tweet IDs in the WeRateDogs Twitter archive to query the Twitter API for each tweet's JSON data using Python's Tweepy library. Each tweet's entire set of JSON data can then be stored in a file called tweet_json.txt file. However, since this .txt file was provided, the tweet_json.txt file was read file line by line into a pandas dataframe.

Data Assessing

After the necessary data has been fully gathered, the data were assessed both visually and programmatically. Some of the programmatic assessment performed on the data includes:

- df.info()
- df.describe()
- df.sample()
- df.duplicates()
- df. Isnull()

and so on. Where df represent the data frame.

Data Cleaning

The data appeared to be very messy and untidy. Therefore, I proceeded into cleaning some of the quality and tidiness issues identified via both visual and programmatic assessment of the data. The Define, Code and Test framework were followed throughout this process. Some of the issues addressed are:

Quality issue

- The df_twitter_archive data contains lot of missing values.
- Missing values on the df_twitter_archive data are represented as 'none'.
- The url was type more than ones in the expanded_url column of the twitter_archive data.
- In the 'names' column of the twitter_archive data, some names were represented with just 'a'.
- The timestamp should be a datetime data type and not object.
- Inconsistent naming convection in the p1,p2 and p3 columns of the image prediction data(some started with upper case while some started with lower case).
- More than 70% of data in the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp hence these columns should be dropped since we don't need them.
- The twitter_id should be object(string) not integer.
- The rating_denominator has values other than 10.
- Convert retweet_count and favorite_count from float to integer

Tidiness issues

- The different dog "stage" (i.e. doggo, floofer, pupper, and puppo) shoould appear under the column 'dog stage'.
- The retweet count and favorite count should be part of the twitter_archive dataframe hence the twitterApi and twitter_archive should be merged
- The Image prediction dataframe and twitter_archive dataframe should also be merged

Saving the data

After the cleaning operation has been performed on the data, the combined data frame was saved as a csv file with the name **twitter_archive_master**.