Wrangling We Rate Dogs Twitter Data

Data Wrangling for this project can be divided into 3 steps :

- Gathering data
- Assessing data
- Cleaning data

Gathering Data:

In this phase, we gather data from 3 sources:

- Data from a .csv file - loaded into workspace as df_twitter using pandas read functionality
- Data from a .tsv file - downloaded programmatically from Udacity's servers using requests module and store as df_img
- Data from twitter API using tweepy module

Twitter API is accessed using consumer and access credentials of twitter developer account.

Tweets are accessed using the tweet_id from the csv file gathered and are stored in json format in a text file. The stored json data from the text file is then read line by line and a dataframe df_twitter_new is built. The json module is used to store and retrieve data in json format.

Assessing Data:

Data is assessed visually using google sheets and programatically using pandas and the following issues were found :

Quality Issues

df_twitter

- tweet_id is of type int instead of str
- source column has text content within anchor tag
- timestamp is of type datatype object instead of datetime

- retweeted_status_id and retweeted_status_user_id are of type float instead of string
- retweeted_status_timestamp is of type datatype object instead of datetime
- in_reply_to_status_id and in_reply_to_user_id are of type float instead of string
- Values other than 10 in rating_denominator
- Extreme values of 10 and 1776 in numerator
- Animals other than dogs are also present in the tweets. Many of them are also given ratings

df_twitter_new

- tweet_id is of type int instead of str
- Rows are lesser than that of df_twitter.Total rows = 2342. No of rows in df_twitter = 2356

df_img

- tweet_id is of type int instead of str
- img_num is of type instead of category
- Rows are lesser than that of df_twitter. Total rows = 2075. No of rows in df_twitter = 2356

Tidiness Issues

- According to the rules of tidy data, df_twitter_new can be merged with twitter archive dataframe to form a single observational unit
- Dog stages doggo,floofer, puppo, pupper are in separate columns

Cleaning Data:

A copy of all dataframes are made with _clean as suffix in dataframe names and cleaning operations are performed on the copies.

- Datatype issues are fixed using astype and to_datetime methods
- Beautiful soup module is used to clean the ‘source’ column of df_twitter_clean
- Since we only need original tweets, rows that contain replies to tweets and retweets and related columns are identified and dropped from the dataframe using pandas drop method
- Rows with tweets that rate non dog entries are identified using the string ‘we only rate dogs’ and are dropped

- Dog stages - doggo, floofer, pupper and puppo that are in separate columns are organised into a single column using string manipulations
- Twitter archive data in df_twitter_clean and additionally collected data in df_twitter_new_clean are grouped as a single observational unit using pandas merge functionality. Only rows with common tweet ids are merged.
- Values other than 10 in rating_denominator are identified and corrected
- Incorrect values in rating_numerator are identified and corrected

Storing the Cleaned Data:

The cleaned data in df_twitter_clean and df_img_clean are merged into df_twitter_master dataframe using pandas merge functionality.Only rows with common tweet ids are merged.

The merged master dataframe is then stored in a csv file using pandas to_csv method.