Project stages

This project will go through 3 stages (Gathering, Assessing and Cleaning) in order to gain a clean and combined dataset and identify at least eight quality issues and two tidiness issues.

Data wrangling, which consists of:

1- Gathering data.

2- Assessing data.

3- Cleaning data.

Storing, analyzing, and visualizing your wrangled data.

Reporting on:

1- your data wrangling efforts.

2- your data analyses and visualizations.

Gathering Data

1) Enhanced Twitter Archive

The WeRateDogs Twitter archive. (twitter_archive_enhanced.csv)

I read the file directly using pandas read function.

2) Image Predictions File

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. (image_predictions.tsv)

I downloaded the data using request libraries and read it by pandas.

3) Additional Data via the Twitter API

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line.

Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Because the process of obtaining a developer account and extracting tweets using twitter API takes time I will not go with the first option and will just download the json file that is already created and provided by Udacity then read the file and load the required data then create a data frame.

Quality issues

twitter_archive tabel

1- The table includes retweets and replies while we need only original tweets. Rmove the rows those are not original tweets.

2- Most of the values in the columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp) are missing and there is no interest to have these columns in our analysis. Remove them.

3- Wrong data type (timestamp). Change the type.

4- tweet_id in the type of int instead of object in all tables. Change them all.

5- rating_denominator column includes values with 0 while it shold not.

6- rating_numerator column includes values less than 10 while it should not.

7- name column includes values with "None". Replace them with NaN.

image_predictions table

8- Some of the data in the colmns p1, p2, p3 in uppercase and some in lowercase. Change them all to uppercase.

9-The columns names p1, p2, p3 are not descriptive. Change them to more descriptive names.

Tidiness issues

10- doggo, floofer, pupper, puppo in seperated columns in the twitter_archive table. Add another column for Dog class and remove the seperated columns.

11- All tables are separated and must be merged together. Merge them all on "tweet_id".

Save the clean data frame to (twitter_archive_master.csv).