Data Wrangling

Project Details

Fully assessing and cleaning the entire dataset would require exceptional effort so only a subset of its issues (eight quality issues and two tidiness issues at minimum) needed to be assessed and cleaned.The tasks for this project were:

- Data wrangling, which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Storing, analyzing, and visualizing our wrangled data

Gathering Data for this Project

Enhanced Twitter Archive

The WeRateDogs Twitter archive provided by Udacity. This contains basic tweet data for all 5000+ of their tweets, but not everything.I manually downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv

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. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: image_predictions.tsv

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.

Assessing data

Tideness issue

- (doggo , floofer ,pupper ,puppo) colums it is must be a value for a new column dog_type not as separated column
- Combine the three datasets together.

Quality issue

'twitter_archive' table

- 1) columns ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id' , 'retweeted_status_user_id' ,'retweeted_status_timestamp') have too many missing values and might not be needed in this analysis so i drop it
- 2) in columns (doggo , floofer ,pupper ,puppo) replace 'None' with NAN to calculate the missing values in this columns
- 3) There are 2297 tweets with expanded_urls (links to images) indicating 59 tweets with missing data
- 4) drop 'expanded_urls' columns i will not use in analysis as it repeat in image_predictions table as "jpg_url"
- 5) Change tweet_id to an object datatype in the 3 tables
- 6)the rating_denominator = 10 will be used & other value will neglected
- 7) the rating_numerator < 20 will be used & other value will neglected
- 8) Some of the rows from the tail() output above have invalid strings in -- the name column, e.g. "a", "an", "in". These words are all the 3rd word in the tweet these will change to 'none '
- 9) 'timestamp' column convert its data type from string to data time
- 10) Drop all rows containing retweets, where these columns will be non-null: retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp.

Cleaning Data

- Remove columns ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id' , 'retweeted_status_user_id' ,'retweeted_status_timestamp') as it will no longer provide any useful information.
- define drop 'expanded_urls' columns i will not use in analysis as it repeat in image_predictions table as "jpg_url"
- correct datatype of 'tweet_id' to be string instead of integers in 3 tables
- correct datatype of 'timestamp' to be datatime instead of string
- want to kknow the missing data exactly in columns doggo, floofer ,pupper ,puppo so i will replace 'None' with nan
- Removing multiple cases of where the denominator of rating != 10.
- Multiple cases of where the numerator of rating < 20. These entries will be removed.
- It looks like the dog names are all capitalized, so words that begin in lowercase are probably not names, #like "a", "the" and "an". Here's the list of these "names".
- list all names that are NOT capitalized
- All these entries were changed to "none"
- use loc to add a new column dog_class = doggo, floofer, pupper or poppo. NaN will be used if not any of the previous
- dropping unneded doggo, floofer, pupper or poppo columns
- # Merging documents to form a working dataframe
- # drop 'rating_denominator' column
- Only want original ratings (Delete retweets and replies).Select the rows from twitter_archive_df that retweeted_status_id and in_reply_to_user_id columns that is null
- There are 181 retwe`ets, and we’re only interested in “original tweets”.