Dog Ratings- Data Wrangling

This data wrangling focuses on gathering , assessing, cleaning and storing of various data sets related to “dogratings" twitter feed. Data is ultimately cleaned and put in a format appropriate for analysis and visualization.

Data Gathering:

- “twitter_archive_enhanced.csv” file is downloaded from Project folder given by Udacity. This data is loaded into a Pandas data frame. This data set consists information such as tweet ids,ratings, breed type, dog name, timestamp of the tweet, url of the tweet, source etc
- “image_predictions.tsv” file is downloaded programmatically from the URL

URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv using Python Requests library.

This dataset is loaded into a pandas data frame. This data set is the result of neural network algorithm that classifies the names of the dogs based on the images, it also hold prediction numbers and their corresponding confidence matrix.

- “tweet_json.txt” file is downloaded programmatically by calling twitter API using tweepy. This data is stored into a pandas data frame. This data set consists of retweets and favorite counts for each of the tweets in the original “twitter_archive_enhanced.csv”

These three datasets are later loaded into three different data frames for assessment of the data.

Data Assessment:

After assessing the data manually and programmatically using various Pandas functions that following conclusions are made.

Quality:

- “twitter_archive_enhanced.csv” has columns “in_reply_to_status_id”, “in_reply_to_user_id”, “retweeted_status_id” , “retweeted_status_user_id”, “retweeted_status_timestamp” columns have majority null values.
- “twitter_archive_enhanced.csv” has “retweeted_status_id” populated for retweets. To prepare a simple dataset to join with tweet counts lets remove the retweets.
- “in_reply_to_status_id” in “twitter_archive_enhanced.csv” is populated for tweet replies.Lets remove these tweet replies.
- “name” column in “twitter_archive_enhanced.csv” has majority of the rows (745) have None populated.There are also 109 rows that have invalid values for name with values like ('the','a','an','such' etc).
- Invalid values in ratings numerator and rating denominator in “twitter_archive_enhanced.csv”. There are 16 records that have ratings numerator greater than 14 and ratings denominator not equal to 10.
- There are some records with decimal rating numerator. These are incorrectly extracted from the text column.
- timestamp column in “twitter_archive_enhanced.csv” is not in date type.
- When one of the columns in columns doggo, floofer, pupper, puppo is populated, other columns are populated with a value "None" . This should be represented as a Null value.
- name column has majority of the rows (745) have None populated.There are also 109 rows that have invalid values for name with values like ('the','a','an','such' etc).
- Source column data in “twitter_archive_enhanced.csv” is not very useful for the analysis. Delete this column.
- column names in “image-predictions.tsv “must be renamed to something more descriptive about the data it holds
- Some breed name values in “image-predictions.tsv” are populated in p1_dog are starting with upper case and some are starting with lower case letter.

Tidiness

- Dog breed values in “twitter_archive_enhanced.csv” can be restructured to be stored in one column instead of four different columns.

- Drop the columns that are unnecessary in “image-predictions.tsv”. such as img_num,p2, p2_conf,p2_dog,p3,p3_conf,p3_dog

3. “twitter_archive_enhanced.csv” has 2356 rows and “image-predictions.tsv” has 2075 rows. After removing the retweets merge the data in these datasets with “tweet_counts” to have the data in single data frame

Data Cleansing:

First take a copy of the original data frames before making any changes to the original data set.

The following steps are performed programmatically on the copied data frames using various python functions to clean the data set

On the twitter_archive_enhanced data set

- Delete the rows that are tweet replies
- Delete columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id, retweeted_status_timestamp after removing the retweets
- Identify the rows that are invalid in the name column and delete them
- Change the timestamp column to date data type
- Filter out the rows where ratings denominator is not equal to 10
- Extract decimal rating numerator values from text column and update the rating numerator with correctly extracted values.
- Filter out extreme outliers in Ratings numerator i.e delete rows where ratings numerator is (1776,420) any ratings greater than equal 420
- Delete source and expanded_urls column
- Create a new column called breed_type and populate it using the data in the columns doggo, floofer, pupper and puppo.

On image_predictions data

- Drop the columns img_num,p1_dog,p2, p2_conf, p2_dog, p3, p3_conf,p3_dog
- Rename the columns jpg_url, p1, p1_conf and p1_dog to descriptive column names such as image_url, prediction, confidence_level

- Join the three data sets into one single data frame

Data Store:

Final copies of each data frame is exported into css file and stored in the folder.

Final master copy of the combined data set is exported into twitter_archive_master.csv