Project#4 - Wrangle and Analyze Data

Developed By: Rakan210 on June 17, 2019 - as psrt of Udacity Data Analyst Nanodegree program (DAND)

Project Steps

1. **Gathering data**
2. **Assessing data**
3. **Cleaning data**

Part#1: Gathering Data:-

This project required 3 source of data to be gathered:

- The WeRateDogs Twitter archive, which is provided in the following link: [twitter_archive_enhanced.csv].
- 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 .
- 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. Note: do not include your Twitter API keys, secrets, and tokens in your project submission.

Part#2: Assessing Data:-

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues in your wrangle_act.ipynb Jupyter Notebook.

Quality issues: (completeness, validity, accuracy, and consistency issues)

- All 'None' values should be changed to NP.NaN.
- Delete all Retweets/Replys records since they shouldn't be considered. ([retweeted_status*] & [in_reply_to*] columns)
- Delete all records that missing values in [expanded_url] column as they indicate ratings without an image.
- Delete all records that have inaccurate value in [rating_denominator] as it should fixed to "10".
- Delete all records that have extreme rates in [rating_numerator] column (i.e. 0 & >20), as the over all rating value should not exceed "10".
- Replace invalid dogs' names with NP.NaN.
- Rename column labels to a descriptive ones.
- Not all tweets have predictions photo (Archive Tweets "df_archive": 2356 rows , Predection Images "df_images": 2075 rows), so only tweet with images will be considered.
- Insert NumPy.NaN to indicate Null values.

Tidiness issues: (structural issues)

df_archive:

- Remove unrequired columns for the analysis.
- Combine dog stages (doggo, floofer, pupper and puppo) into one column called [dog_stage].
- Combine rating columns into one column called [dog_rating]

df_master:

- Combine all DataFrames into one master DataFrame called [df_master]. Its content should be stored in twitter_archive_master.csv.

Note: Bold columns will be marged in the master dataframe

#df_archive:

- **tweet_id** - Required for Analysis
- in_reply_to_status_id - Required for Clean up
- in_reply_to_user_id - Required for Clean up
- timestamp - Not required
- source - Not required
- text - Not required
- retweeted_status_id - Required for Clean up
- retweeted_status_user_id - Required for Clean up
- retweeted_status_timestamp - Required for Clean up
- expanded_urls - Required for Clean up
- **rating_numerator** - Required for Anaysis & Clean up
- **rating_denominator** - Required for Anaysis & Clean up
- **name** - Required for Anaysis & Clean up
- doggo - Required for Anaysis & Clean up
- floofer - Required for Anaysis & Clean up
- pupper - Required for Anaysis & Clean up
- puppo - Required for Anaysis & Clean up

#df_images:

- **tweet_id** - Required for Anaysis
- **jpg_url** - Required for Anaysis
- **img_num** - Required for Anaysis
- **p1** - Required for Anaysis
- **p1_conf** - Required for Anaysis
- p1_dog - Not required
- p2 - Not required
- p2_conf - Not required
- p2_dog - Not required
- p3 - Not required
- p3_conf - Not required
- p3_dog - Not required

#df_collected_tweet:

- **tweet_id** - Required for Anaysis
- **favorites** - Required for Anaysis
- **retweets** - Required for Anaysis
- **date_time** - Required for Anaysis

Part#3: Cleaning Data:-

===| Summary of Cleaning Data Stage |===

#1# All 'None' values has been changed to NP.NaN
#2# Deleted Records: 259 -> Reason: Retweets/Replys
#3# Deleted Records: 3 -> Reason: Ratings without an image
#4# Deleted Records: 17 -> Reason: Inaccurate Rating (denominator)
#5# Deleted Records: 6 -> Reason: Inaccurate Rating (numerator)
#6# [dog_rate] column has been created contains calculated dog's rating
#7# Invalid dogs' names {'a','the','an','not','one','Mo','O','Al','my','his','this','all'} were replaced with NP.NaN
#8# [dog_stage] column has been created contains value of 'doggo','floofer','pupper','puppo' columns
#9# Unrequired columns in [df_archive_clean] have been deleted (listed above)
#10# Unrequired columns in [df_images_clean] have been deleted (listed above)
#11# Combined all DataFrames into Master one called [df_master]
#12# Renamed [df_master]'s column labels to a descriptive ones
#13# Saved [df_master] to a CSV file called 'twitter_archive_master.csv'

df_master.info()

Data columns (total 11 columns):
time_stamp 1943 non-null datetime64[ns]
tweet_id 1943 non-null int64
favorites 1943 non-null int64
retweets 1943 non-null int64
dog_name 1356 non-null object
dog_rate 1943 non-null float64
dog_stage 301 non-null object
image_url 1943 non-null object
images_count 1943 non-null int64
img_prediction 1943 non-null object
prediction_conf 1943 non-null float64