Project#4 - Wrangle and Analyze Data

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

In this project, all the analysts were requested unleash their analytic skills by walking through all data analysis process [Gather, Asses, Clean, Analyze, and Report]. The focus is at the data wrangling part, which includes gathering, assessing and cleaning the data. Below are the detailed requirements:

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)

Note: Visual assessment (VA) & Programmatic assessment (PA)

df_archive:

1. (PA) Change datatype of [timestamp] to datetime64 and [rating_numerator] to float.
2. (VA) Rename column labels to a descriptive ones.
3. (VA) Change all 'None' values should be changed to NP.NaN.
4. (PA) Delete all Retweets/Replys records since they shouldn't be considered. ([retweeted_status*] & [in_reply_to*] columns)

5. (VA) Modify some records that have inaccurate value in [rating_denominator] as it reasonibly to be fixed to or be multiple of "10" (rating_denominator %10 = 0).
6. (PA) Modify some records that have inaccurate value in [rating_numerator] due to the existing of the dicimal.
7. (VA) Replace invalid dogs' names with NP.NaN.
8. (PA) 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.
9. (PA) Delete all records that missing values in [expanded_url] column as they indicate ratings without an image.

Tidiness issues: (structural issues)

df_archive:

1. Combine dog stages (doggo, floofer, pupper and puppo) into one column called [dog_stage].
2. Combine rating columns into one column called [dog_rating]

df_master:

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

Part#3: Cleaning Data:-

Detailed step-by-step cleaning process was done to overcome all the mentioned issues above in ` wrangle_act.ipynb`. In general, each data frame was handled and cleaned by its own before merging them into a master data frame. After that, another cleaning process was done on the master data frame to finally polish the data and start the analysis and reporting stages.