Wrangle Report

By: Hamdy Abdel-Shafy

Project overview:

The objective of this project is to practice what have been learned in the section of data wrangling and anaysis from "Udacity Data Analysis Nanodegree Program". Through this project, the dataset wrangled, analyzed, and visualized is the tweet archive of Twitter user @dog_rates which is also known as WeRateDogs. Briefly, WeRateDogs is a Twitter account that rates people's dogs with humorous comments about the dog. Further information about this issue can be seen at Wikipedia or at Twitter.

Project steps:

The project is conducted by the following steps:

- **Gathering data**

Three different dataset were obtained and used as described below in a Jupyter Notebook:

a. The WeRateDogs Twitter archive: this data (twitter_archive_enhanced.csv) was provided by the Udacity.

b. The tweet image predictions: this data (image_predictions.tsv) is hosted at Udacity's server and was programmatically obtained by using a suitable library and URL information (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv).

c. Twitter API & JSON: using the tweet IDs in the WeRateDogs Twitter archive file, the Twitter API for each tweet's JSON data should be queried using Python's tweepy library and each tweet's entire set of JSON data should be stored in the file tweet_json.txt. This file should be line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url. An alternative way, which is the procedure I used in the current project, is to access the data without a Twitter account using information provided Udacity from the file: twitter_api.py. It is a Twitter API code to gather some of the required data for the project. The code is red and

executed in my notebook. Also, I have tried using the data from tweet-json.zip file provided by Udacity.

- **Assessing data**

  Once the three tables were obtained I assessed the data, I had a look on each part of the data using some statistical parameters through Jupyter notebook.

- **Cleaning data:**

  It is very important issue to check the quality of data and clear unexpected errors before analysis step to make the outputs valid, consistent and more accurate. A first and very helpful step in this regard, was to create a copy of the three original dataframes to save the original data from any unexpected problems that could happen. There were a couple of cleaning steps that were very challenging. I put detailed explanation of each step my wrangle_act.ipynb.

- **Storing and analyzing data**

  The filtered datasets are combined together according to tweet ID and stored on local directory in two different formats to be used later for further analyses and visualizing the outputs.

- **Presenting outputs in suitable format**

  I created different plots using different variables, e.g. box plot, bar chart, and histogram. Finally, I created a funny word cloud for the tweets text.

Conclusion:

The subject of data wrangling and analysis is considered a core programming skills that whoever handles data should be acquired. Beside the skills we acquired through this experiment, additional independence advantage was obtained through looking for solutions for programming errors that one might be made by using a libraries documentations and/or helpful websites, e.g. Stak Overflow.