Name: Amr Elbana
Date: 15/ 03/ 2022

Wrangle Report

This is the 4th project in Udacity Data Analyst Nanodegree. This project helps the student to put the newly acquired skills into practice. The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The project is consisting of three parts which are under the wrangling process, and they are as follow:

1) Data Gathering
2) Data Assessing
3) Data Cleaning
4) Data Storing
5) Data Analyzing and visualization

a) Data Gathering

In this project, I will work on three files, each one of them will be gathered in a different way.

1) The first file is “twitter_archive_enhanced.csv”, this file is provided, and I will just import it into the jupyter notebook using the pandas package from python.

2) The second file is a downloadable file from the Udacity server and is also it just being imported to the Jupyter notebook using pandas.

3) The third file is the true challenge between all files as this file will be pulled from Twitter API, I will 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 panda Data Frame with (at minimum) **tweet ID, retweet count, and favorite count**.

b) Data Assessing

In this part I used the initial visual inspection to inspect the three datasets after reading into the notebook. I found many aspects that need further inspection which is done programmatically.

After the inspection, I decided which points I am going to work on and categorize them into quality issues that are related to the completeness and accuracy of the data and the tidiness of the data which is related to the structure of the data.

c) Cleaning Data

After assessing the data, the cleaning step come

- Quality issues
    1) Tweet_id should be an object, not an int
    2) Timestamp should be Date Time, not object
    3) There are 66 images that are duplicated
    4) There are some columns that contain a lot of null values
    5) the most common value in the rating_denominator column is 10, so other values should be either removed or changed
    6) some cells in expanded_urls contains more than 1 URLs
    7) There are around 745 None values in the name of the dogs which are Null values
    8) The number of tweet id is not consistent across the three data frames
    9) Extract the breed of the dog from the prediction tables

- Tidiness issues
    1) Maybe these columns ['doggo', 'floofer', 'pupper', 'puppo'] should be transformed to dummy variable

    2) The data should be in one table