Data Wrangle Report

By Dina Essam, Jan. 2021

Data Analysis Nanodegree Project :Wrangle and Analyze Data

The dataset that will be wrangled , analyzed and visualized is the tweet archive of WeRateDogs. WeRateDogs is a twitter account that people rate their dogs.
wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.
The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356)

Purpose of this analysis :

The purpose is to create analyses and visualizations as it is very basic information from twitter ,that need gathering, then assessing and cleaning .

Data wrangling, which consists of:

- Gathering data :- In this step we collect data from three main sources first source is twitter_archive_enhanced.csv that is downloaded manually , the second source is image_prediction.tsv that is downloaded from its relevant URL by using request library , final dataset was gathered from tweeter API via tweepy library .
- Assessing data :- In this step we investigate our datasets to detect the quality and tidiness issues , as missing and empty data , delete useless data columns , tidy data as combine the three datasets in one master dataframe .
- Cleaning data :- Clean each data issue that documented while accessing , the project should be cleaned after this step , for example completeness as some records missing statistical fields , another quality issue is validity there are some records are not actually original tweets .then tidiness issue three columns are present for variable dog_stage .
- Storing, analyzing, and visualizing our wrangled data :- Store the clean DataFrame(s) in a CSV file with the main one named twitter_archive_master.csv. Analyze and visualize wrangled data in wrangle_act.ipynb Jupyter Notebook. It explained briefly in act_report.pdf .