Data Wrangling Project

WeRateDogs Twitter data wrangling and analysis

Prepared by: Ahmed Senam Mostafa

March 2021

This report briefly describes the wrangling efforts done on the data of WeRateDogs, by go through the 3 stages of data wrangling:

1- Gathering data
2- Assessing data
3- Cleaning data

1- Gathering data:

Gathering data is the first step in data wrangling. Before gathering, we have no data, and after it, we do.

In this project we have 3 data frame

- twitter-archive-enhanced.csv ➔ which is provided by Udacity, and contains large amount of tweets information as (tweet_id, time, rating, …)
- image-predictions.tsv ➔ File is downloaded from given URL, and contains (images, dog breeds,…)
- tweet-json ➔ I used the one provided by Udacity as my tweeter developer account still not active. This file has information related to number of retweets and favorites for each tweet_id.

2- Assessing data:

Jupyter notebook and pandas have been used to assess the three data frames. Data assessment done using visual and programmatic checks.

Below are sample of pandas commands used for assessment:

- df.head
- df.info()
- df.describe()
- df.value_count()
- df[df[‘column'].isnull()]

The output of the assessment stage is divided into:

- Quality issues ➔ which are related to validity, consistency, accuracy… of data
- Tidiness issues ➔ which are related to structure of data

3- Cleaning data

Based on the output and classification of the previous stage (assessment). I will work in this stage to clean the data.

- Quality issues:
    - tweet-json file
        - Keeponly needed columns "id_str" , "retweet_count" , "favorite_count"
        - Convert "id_str" from int to string
        - Rename "id_str" to "tweet_id"

    - twitter-archive-enhanced.csv
        - convert "timestamp" to datetime
        - convert "tweet_id" to string
        - removing "retweeted status_id" rows
        - remove "in_reply_status" rows

    - image-predictions.tsv
        - make all dog names start with small letter
        - Removing rows with no dog picture (all are false)
        - convert type of "tweet_id" from int to string

- Tidiness issues:
    - twitter-archive-enhanced.csv
        - add one column for dog stage
        - Drop 4 columns (puppo, pupper, floofer, doggo)
    - Merge 3 data frames in 1 data frame as all are related to same things

Finally, all the above steps are stored in:

- twitter_archive_master.csv ➔ contains the clean and tidy data for anaylsis
- wrangle_act.ipynb ➔ contains all work done on data frames