Wrangling Report of the WeRateDogs Project

By: Cynthia Obojememe

23/May/2022

This Waring report briefly explains the steps observed in gathering the needed dataset for the WeRateDogs tweet project, analyzing the data and to provide interesting and trustworthy analyses and visualizations.

The Wrangling process of the data is divided into three steps:

1. Gathering data.
2. Assessing data.
3. Cleaning data

Gathering data

The gathering data step involves obtaining the needed dataset from 3 different sources and using 3 different methods to load the dataset to the pandas dataframe.

The first was to manually download the csv file title Twitter_archive_enhance csv provided by Udacity by simply clicking on the provided link to download the data. The data consist 2356 entries of WeRateDogs archived tweet which has already been extracted and filtered (to a limited degree) from the 5000+ archived for tweets with ratings only

The second step was to be able to download a tsv file from the server link provided by Udacity programmatically using Python Requests library. The file contained image predictions which is a flat file(tsv)on the breed of the dogs coming from a neural network on some of the tweets already downloaded in the archive file.

The third step was to query the twitter API using the Tweepy library for WeRateDogs archive file to obtain tweet_Id, retweet count,favorite count stored. The tweepy query then returns the content of the file into a JSON file format. This was loaded using the Json library.

All 3 sorted datasets were then loaded into a pandas dataframe in my jupyter notebook.

Assessing the data

To evaluate the already gathered dataset and save the files a dataframe, I had to assess the file visually and programmatically to detect and filter problems with the dataset which are either Tidiness issues (These are structural issues) or Quality issues (contain related issues). Due to the large contain of the files, i had to analyze the dataset programmatic assessment was made inside jupyter with pandas using the following functions, df.info(),
df.head(), df.sample() (several different samples were taken), df.value_counts(), df.sort_values(), loc functions,describe(),unique() etc.

Several tidiness and quality issues were noted in the wrangling act ipynb to be fixed in the cleaning stage. Some issues stated are, the duplicate Tweet_id on the Twitter_archive_enhance csv, image predictions and json.txt data frames, The structure issues of the dog stage in the Twitter_archive_enhance csv, etc.

The basis for assessment is that each variable forms a column, each observation forms a row and each type of observational unit forms a table. After assessing the three datasets, it was decided to work on the tidiness issues posed from the assessment , reducing superfluous columns that wouldn't be needed in any future analysis.

Cleaning the data:

This is the final step in the wrangling process is cleaning all the outlined tidiness and quality issues.The cleaning step has to adhere to the standard process of define, code and test for fixing each of the issues. We are advised to first make a copy of the dataset before the cleaning operations will be conducted. I utilized the pandas library which has a lot of functions to fix all of the issues listed in the data. Most of the cleaning was performed using programmatic tools, such as def functions for appending some multiple columns, and pandas built-in Functions such as (merge, extract, regex etc).

In conclusion, The Data wrangling provides a clean data frame for future analysis and visualization, in our case, we stored the cleaned data into a new csv file titled 'twitter_archive_master.csv'.