Wrangle Report

By Murtala Umar Adamu

November 2022.

This is a report that covers the wrangling effort carried out in the course of the project. The project was in three phases which include the following:

1. Gathering
2. Assessing
3. Cleaning

Gathering

Data was gathered from three sources in the course of the project. In obtaining the data, different methodologies were used.

The first dataset which is called *twitter-archive-enhanced* is a csv file manually downloaded. This dataset contained the most data of the three with 2356 rows and 17 columns.

The second dataset was downloaded programmatically from Udacity servers using python's request library. The dataset which is called *image_predictions* is a tsv file containing image predictions. The dataset has 2075 rows and 12 columns.

Lastly, the *tweet_json* dataset was downloaded as txt file. The dataset was to be downloaded programmatically via the twitter API or manually in the event setting up a tweeter developers account proved abortive. Following the challenge faced in setting up the developers account, the dataset was downloaded manually. In order to work with the dataset, a for loop is used to iterate over the *tweet_json* file which is saved in a list called *tweets_details*. The dataset contains 2350 rows and 31 columns.

Assessing

The datasets were independently assessed. In doing this, both visual and programmatic methodologies were employed.

Visually, the three datasets were exported and opened in excel notebook. The datasets were inspected using filters. This made it easily to scan through the rows and columns. Programmatically, various commands which include, *.head(), .tail(), .sample(), .describe(), .isnull()*, and *.shape* amongst others were used to inspect the datasets.

Quality and tidiness issues were detected in the course of the assessment. The quality issues detected include dog names mixed in upper case and lower cases, contents of text column cut short, some columns which were irrelevant and were best dropped amongst others.

The tidiness issues detected were basically two. The three (3) datasets will be merged to one and the four (4) categories of dogs will be merged to one column.

Cleaning

The cleaning of the dataset entails dealing with all the issues identified in the assessment stage of the data wrangling.

Copies of the datasets were made following best practice. The datasets were then merged to one. Following that, all issues related to quality were then dealt with. It is important to note that all the issues were taken care of programmatically.

Conclusion

In conclusion, the data wrangling process was carried out in a step by step process starting with gathering the data, assessing it for quality and tidiness issues and then cleaning following all the issues identified.