Report on The Process of Wrangling WeRateDogs Datasets

Introduction

As it is generally known, the Process of wrangling a data or datasets involves a compilation of skill and know-how, in order to attempt to achieve to get a clean data to use for analysis. In this report, I explain the process of gathering and assessing WeRateDogs dataset from three different source using different methods.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage. The dog ratings always have a denominator of 10 and the numerators is almost always greater than 10. Why? Because "they're good dogs Brent!"

Importing The Libraries

Before any wrangling process can be made, one needs to import python libraries that make working in Jupyter Notebook easier and smooth. Different libraries are sourced and combined to allow for accurate and efficient analysis. Before a library or libraries can be imported and used in Jupyter Notebook, it must first the installed.

Data Gathering Process

The three datasets gathered for this project:

1. A CSV file named twitter-archive-enhanced.csv, read with the regular pandas function. The file contains 2356 rows and 17 columns.
2. A TSV file named image_predictions.tsv, was read directly from it URL source using the pandas 'with open()' method. It has 2075 rows and 12 columns.
3. A JSON file named tweet_json.txt, which was read also with the 'with open()' method as shown below. Its is worthy of note that a Twitter API could also be used to gather this particular data. This dataset contains 2354 rows and 3 columns, as shown below.

Data Assessment

My data assessment process involved the use of visual and programmatic methods in discovering the issues of the datasets. The visual assessment involved exploring the dataset in a spreadsheet or using pandas methods to discover underlying messiness or untidiness in the dataset while Programmatic assessment, on the other hand, was done using pandas methods.

However, due to the rigour of assessing datasets, I restricted the issues detected in this project to nine (9) quality issues and three (3) tidiness issues, using both visual and programmatic assessment.

Also, I ensured I worked on copies of the datasets so as to preserve the original datasets in reading into memory.

Data Cleaning

For this project, I employed a three-stage method The method used for data cleaning are:

- Define
- Code
- Test

The Define stage focuses on how the overview of the issue and how I plan to solve it. The Code stage is where I actually write the python codes to resolve the issue. The Test stage is where I visualize my solution, to confirm that the issue has been resolved.

In this section, I cleaned all of the issues documented while assessing.

Merging and Storing Data

The clean datasets were then merged into one final dataset that has rows and columns. The final dataset was then stored in a master data frame called twitter_archive_master.csv with utf-8 encoding.