WeRateDogs

Data-Wrangling Project Report

Data wrangling is the key part in data analysis. Data wrangling is the process of gathering your data, assessing its quality and structure, and cleaning it before you do things like analysis, visualization, or build predictive models using machine learning. In this project, I analyse the WeRateDogs Data, which is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 9 million followers and has received international media coverage.

The first step in this project is the gathering of data and those data are obtained using 3 different method that is manual downloading of twitter_archive_enhanced.csv file, programmatically downloading of image_predictions.tsv from the provided url using Request library and query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. The needed information is extracted from the text file and stored in a dataframe.

The next step which involve assessing of the gathered data visually and programmatically. Visually involve open the whole on jupyter notebook or external application, scroll through to detect any inconsistent or abnormal data such as missing data or wrong entering of data. Programmatically involves use of different panda functions and/or method to assess detect issues. These issues include quality and tidines issues which are documented for cleaning stage.

The next step is cleaning of the data that is the issues documented during assessing stage. Before cleaning, a copy of each data was made to avoid data being loss. To clean, the define-code-test framework was to use and it help clarifies each step of the cleaning and proper documentation. Define part of the cleaning step is the converting the documented issuses into defined cleaning tasks which serve as an instruction list to other or for me in the future so that they can look at my work and be able to reproduce it. The code part convert these instructions to code and then test the code visually to see if the issue was resolved.

After the cleaning operation the final data was high-quality and tidy master pandas DataFrame store in csv file with the main one named twitter_archive_master.csv