wrangle_report

October 24, 2022

0.1 Reporting: wragle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

0.2 Data Wrangling Report

This report provides a summary of the wrangling steps and process used in the second project under the Udacity Nanodegree program. The purpose of the project was to wrangle and analyze data gathered from the "WeRateDogs" twitter account. The process involved three main steps which are: gathering, assessing and cleaning

0.2.1 Step 1: Gathering Data

This is the very first step in data wrangling and invloves collecting or acquiring data from different sources. This data must be relevant to whatever analysis that is being done.Data can be gotten from numerous sources which are usually in different formats, the most important thing is that the data must be credible and you must be able to work with the data in its original format or convert it to a more suitable format required for the analysis.

For the purpose of this project, data was gathered from the WeRateDog twitter account. The dataset was gathered in three different formats, A csv file, A tsv file and a json file.

0.2.2 Step 2: Assessing Data

Data assessment is the second step of the data wrangling process. When you have gathered your data from various sources, the next step is to examine the data for possible issues that can affect your analysis. This assessment can be carried out either **visually** and or **programmatically**.

Visual assessment involves scanning the data set for as many issues that can be spotted by merly looking through the data set. While you may be able to spot a few data issues, it is quite an ineffecient way of assessing data as there usually many more issues you may be unable to spot especially when working with larger data sets.

Programmatic assessement is however a very efficient way in assessing data. It invloves the use of programmatic languages and code is assessing data issues.

Data issues can be categorized into **Quality** and **Tidiness**. Quality issues has to do with the content of the data while Tidiness issues has to do with the structure of the data.Below are the quality and Tidiness issues identified during the assessment step of the proeject

Quality Issues

1. tweet_id is of integer data type
2. The dataset contains some retweet data
3. The name column for animals includes articles such as "a", "an" and "the"
4. The timestamp has +0000 after the seconds
5. rating_numerator and rating_denominator is int dtype instead of object
6. Some of the tweets are replies and are not original tweets
7. Timestamp is object data type instead of datetime
8. in_reply_to_status_id is of float datatype

Tidiness Issues

1. The three data sets should be merged into one
2. Columns doggo, floofer, pupper, and puppo should me merged into a single column

0.2.3 Step 3: Data Cleaning

This is the final step in the data wrangling process. It involves addressing issues identified in the data assessment step. It is advisable to create copies of the orginial data set before commencing the cleaning process, so that the original data set is not altered and can always be referred to.

The data cleaning step for this project followed a 3 step methodology: - Define: Stating how the issues will be cleaned - Code: Writing codes to solve the issue - Test: Checking if the code worked

All quality and tidiness issues identified above were cleaned using pandas in python and other libraries and stored in a master csv file.