Data wrangling, or the process of cleaning raw data, is the most important part of end-to-end data science projects because it allows for valuable insights and accurate decision-making. In this project, we planned to use data-wrangling concepts to clean data from weRatedogs and derive some useful insights. This project followed three data wrangling steps: gathering, cleaning, and assessing.

Three datasets were collected in three different ways. To begin with, the Twitter archive enhanced dataset was manually downloaded, and it includes information such as tweet id, retweet status, dog rating, dog name, and dog categories. Additionally, the image prediction dataset was gathered programmatically and saved using a URL. It contains information such as image URLs and other data. JSON data, which contains every detail about the tweets, was collected using the Twitter API and read line by line to extract necessary columns such as tweet id, retweet count, and favorite count.

Following the collection of all datasets, Pandas, the most commonly used library in this project, was used to load those distinct datasets into the Jupyter notebook. quality issues and tidy issues were identified by assessing them visually with a simple view of those datasets and programmatically with different python libraries such as info, describe, head, tail, and others.

Quality issues

- Columns with incorrect data types
- Inconsistent in data such as mixed upper case and lowercase and underscores in P1, P2, and P3 columns.
- Missing observations
- Nulls represented as none in some columns
- Rows where a dog has more than one category

Tidy issues

- Four categories of dogs should be combined to form a single categorical column
- Three datasets could be merged into one.

In the cleaning part, three datasets were copied and following the define, code, and test process, the aforementioned issues were cleaned whereby we started by what to be done in code or simply how to handle the issue and solve the issue in the code part and test if the issue is solved. Started by addressing all quality issues and continued to tidy issues. In this part, we ended up having a cleaned dataset combined with all cleaned datasets and stored it as twitter_archive_master.

Limitations

Due to the absence of metadata, I was limited in not knowing the meaning of some columns such as p1, p1_conf, p2,p2_conf, and others. This results in not providing some analysis regarding those columns.