wrangle_report

May 29, 2025

0.1 Reporting: wragle_report

1 Here are the folowing steps I took:

- Data Gathering
- Data Assessing
- Data Cleaning

2 First, Data Gathering

2.0.1 What is data gathering?

Data gathering, also known as data collection, is the systematic process of collecting and measuring information on variables of interest. It's a crucial step in research and decision-making, allowing businesses and researchers to gain insights, answer questions, and evaluate outcomes.

2.0.2 How I applied it on this project:

For the first dataset, I was already given a CSV with the data I wanted, so I went ahead and used the *.read()* function and copied the address of the fie after upoading it to the workspace

For the second dataset, I was given the URL to a website, and had to use request to get the content of the second dataset and saved it as a .tsv file,

For my third and last dataset, I was given a JSON file, for which I used *.read_json()* then converted the result into a CSV file, and that was it!

3 Second, Data Aseessing

Data assessing, or data assessment, is a structured evaluation of an organization's data to understand its current state, identify issues, and develop a data strategy. It involves exploring data sources, data quality, and data management practices to ensure accuracy, reliability, and overall effectiveness.

3.0.1 How I applied it in this project

One Visual assessment I made right off the bat, is that In the second database, there appears to be capital and small letters are in the same coulmn, that is clearly a quality issue.

I was asked in this project to define eight quality, and two tidiness issue

After assessing the issues by code, I was asked to write them down in a brief dewscreption, and so I did, here are some of the issues I encountered: * tweet_id should be object * some values in lang coulmn are out dated

that takes us to the last step of Data Wrangling

4 Data Cleaning

4.0.1 What is Data Clening?

Data cleaning, also known as data cleansing or scrubbing, is the process of identifying and correcting errors, inconsistencies, and inaccuracies within a dataset. It ensures that data is accurate, complete, consistent, and usable for analysis or decision-making.

4.0.2 How I applied it in this project

So, now that I have all my problems written down, I can start defining the issues I written down, as in giving them more detail in the part of what I'm going to do to fix these issues, here are exampes for defining your issues:

- Ids are not used for mathematical operations and therefore are best converted into object data types
- The indonesian language should be reffered to as Id instead of In for this had become outdated

Then we start coding a soloution to the problem I documented and defined

Lastly we merge the three datasets to they become one, since there were very few coulmns in each, This was soving a tidiness issue

I also did some cleaning at the end for some coumns to use them for analysis later like source coulmn

And so we were left with the final csv: twitter_archive_master.csv