wrangle_report

May 22, 2025

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.

For this project, I had three separate datasets, The first contains basic tweet data for all 5000+ of their tweets, but not everything.

The second contains a set of images, and three diffrent predections using an ai module featuring what kind of dog it is, but some of them werent real or even dogs

The third and final dataset is actually a JSON file, it contains more details than the last two, some things like the language used in the tweet are featered in this dataset

1 so, here are the folowing steps I took:

- Data Gathering
- Data Assessing
- Data Cleaning

2 The libraries I'm using for the data wrangling

- Numpy: NumPy is a foundational open-source library for scientific computing in Python, providing high-performance multi-dimensional array objects (ndarray) and a wide array of mathematical functions

- Pandas: Pandas is a Python library that provides tools for data manipulation and analysis. It is built on top of the NumPy library and provides data structures like DataFrames and Series that are designed to make working with structured data, such as tables and time series, easier and more intuitive.

- request: is primarily used for making HTTP requests in Python. It simplifies the process of sending and receiving data over the web, offering a user-friendly interface for web-related operations compared to manually dealing with network protocols. Essentially, it's a tool that makes it easier for Python developers to interact with web services and APIs.

3 First, Data Gathering

3.0.1 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.

3.0.2 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!

4 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.

There are also two types of assessing: Visual assessing, and assessing using code

4.0.1 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.

Most of my other assessments are using code, and my most used sodes for assessing are *head(), info(), sample(), .duplicated().sum(), .isnan().sum()* And *nunique()*

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

5 Data Cleaning

5.0.1 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.

5.0.2 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

5.0.3 Limitations on Data Wrangling

I didn't really have any limitaions except in the cleaning step, the predections were the hardest to clean in my opinion