Asking Questions with CSV Data

Conceptual Data Exploration with Native Python or Pandas

Goal

This assignment emphasizes conceptual data literacy over advanced Python syntax. Students will examine how structured, file-based datasets represent information and how relationships emerge from their organization, laying the groundwork for future coursework with relational databases. Through hands-on exploration of a CSV dataset, students will ask and answer meaningful questions and critically assess the limitations and possibilities of the data.

Dataset

Choose a CSV dataset from NYC Open Data or Kaggle:

https://opendata.cityofnewyork.us/data/

https://www.kaggle.com/datasets

Your dataset must have at least 100 rows, 5 columns, and include at least one categorical column.

Implementation Choice

Choose ONE approach (You only need to implement one approach.):

Option A — Native Python using csv.DictReader:

https://medium.com/@3valuedlogic/using-python-csv-3-dictreader-e4814ce2e44
https://docs.python.org/3/library/csv.html

Option B — pandas using DataFrames
https://codesignal.com/learn/courses/basics-of-numpy-and-pandas-with-titanic-dataset/lessons/mastering-pandas-a-deep-dive-into-dataframes-and-data-manipulation
https://medium.com/data-science/olympics-kaggle-dataset-exploratory-analysis-part-2-understanding-sports-4b8d73a8ec30

Required Tasks

Using your dataset, your code must:
1. Print the first 2 rows
2. Print the first row
3. Print rows 10–19
4. Print column names
5. Print the first 10 values of one column
6. Print the first 10 rows of three columns

*7. Your three python statements (code) to answer the three questions below.

***save your code as analysis.py**

Three Data Questions

Write three questions your dataset can answer. For each question include:
- The question in plain English as a comment
- The output
- A short explanation of why the data structure supports this question

***save this as a README.md file**

What the Data Cannot Answer

Write one paragraph explaining:
- A question you wish you could answer but cannot
- What data is missing
- Any assumptions that would be misleading

***save this in your README.md file**

Why You Chose This Dataset

Write a few sentences explaining why you chose the dataset you did.

***include this in the top of your README.md file**

Submission

Submit:
- analysis.py

- README.md with dataset explanations
- CSV file you used

In your github repo. Submit the link to your github repo and edstem repos in Brightspace.

Grading

25% — Data loading & inspection tasks
25% — Filtering and counting
25% — Two-condition breakdown
25% — Reflection on limits