RELEVANT RESEARCH

Data Processing and Analysis Exercise

The exercise below is meant to help us get a sense of what your strengths are as a data scientist. They are also a good example of what you would be required to do if you were to join our company.

Important: You are allowed to use Google, ChatGPT, and other internet resources when completing this exercise. However, whenever you use these resources, please indicate that you do so when you document your work. If you decide to use something you find online, include the URL where you found the method that helped you. *If you resolve an issue using ChatGPT or another LLM, please include a link to your prompt page so that we can see how you modified ChatGPT's response to fit the parameters of this exercise.*

It's not a problem to use these resources. We just want to know *how* you accomplished the task.

Part 1. Tidying and Visualizing an Immigration Detention Dataset

This is an evaluation for your comfort using R or Python programming languages to clean, wrangle, and visualize administrative data. A completed exercise using R would be our preference, but deliverables using Python will be accepted. Please read the instructions very carefully before proceeding through the exercise.

Input: "messy_ice_detention.csv"

This is an example of a dataset that we use very frequently. It contains the basic information of the detention facilities holding immigrants in the United States and is updated semimonthly. An updated, clean version of this messy dataset can be found in the "Facilities" sheet at the bottom of this webpage.

Goal: Clean, analyze, and then visualize the top 10 largest detention facilities in the dataset.

A. Clean

This is a messy dataset. It is likely that this step will take you the longest time.

The first thing you will need to do is remove the header in the first few rows. Then, you will see characters inserted where they shouldn't be, blanks, and misformed dates. You may need to google the address of a detention center. Wherever possible, use code to resolve issues. However, if you decide to manually change a value in the CSV, please indicate where you chose to do so.

Table 1. Overview of relevant columns in detention dataset

Column Name(s)	Description
Name	Name of Detention Facility
City	City of Detention Facility
State	State of Detention Facility
Level A: Level D	Average Populations in Each Category
Last Inspection End Date	Date of Last Inspection

B. Analyze

Sum the “Level A”, “Level B”, “Level C”, and “Level D” columns to make a new “Total Population” column. Then subset that column so that you only have the top ten largest detention facilities in the dataset.

C. Visualize.

Choose whatever type of visualization you prefer (bar graph is probably most straightforward) and show the top ten facilities. The ggplot2 package is recommended, but if you prefer to use base R, that’s fine, too. Matplotlib and Seaborn packages can be utilized here if you are using Python. Save this as a png image.

Part I Output:

1. Your R/Python script(s) with comments about your methods and resources that you used.
2. If you used an AI or LLM, please include the prompt page.
3. An image of your visualization.

Part 2: Github

A. Create Repository

1. Name: “data-screening-exercise”
2. Make it **public.**
3. Add a README.md file and write how to execute and interpret your work.

B. Commit Your Changes

- Upload your R script output and other important files.
- Use Git to commit your work regularly.

C. Submit

- Share the GitHub link in an email to Adam (adam@relevant-research.com) once everything is done.