TRG

Junior Data Engineer Assessment

For this assignment, the datasets you will be working with consist of UK reported street crimes.

1. You can download the relevant datasets here: https://data.police.uk/data/
2. On the page select the following for the respective fields:
   - Date Range: at least 6-12 months worth of data i.e March 2021 to March 2022
   - Forces: All forces
   - Data sets: Include crime data and Include outcomes data
3. Once you have downloaded the zip file and extracted the data, you will notice for each district has two files.
   - As an example for “Avon and Somerset” we have:
     - 2019-01-avon-and-somerset-street.csv and;
     - 2019-01-avon-and-somerset-outcomes.csv

Your task is to create and ETL pipeline that:

1. Extracts the following fields from each csv:
   a. crimeID
   b. districtName
      i. Can be extracted from the filename
   c. latitude
   d. longitude
   e. crimeType
   f. lastOutcome
      i. The last outcome should be taken from the <district>-outcomes.csv file where the crime IDs match. If there is no matching data use the data listed in the original <district>.csv file.
   g. As an example the final data structure should look like **Table_1** below
2. Stores the final structured data in a .csv file or a database of your choice i.e MongoDB or Postgres
3. [Bonus] Provide some insight into the data i.e Where do most crimes happen? Whats is the most common crime?
4. [Optional] Use docker-compose to orchestrate the set up of your final solution project.

5. Make sure to include a detailed README.md file outlining the setup instructions and description of what steps you took to reach your final solution.
6. Upload your project to GitHub/GitLab and include a link when responding.

crimeId	**districtName**	**latitude**	**longitude**	**crimeType**	**lastOutcome**
98096d1a69205691a56b89c1182eadd6aaf15400ea18da134e0023f20aba5cdb	avon and somerset	51.419357	-2.515072	Criminal damage and arson	Under investigation
7984cd127f0fa49c7fc6de29e042b51881910a716de1d12c49f7bbe9a809ecd4	avon and somerset	none	none	Vehicle crime	Suspect charged

Table_1: final data structure

If there are any details that are unclear please do not hesitate to ask any questions.

Good luck! We look forward to reviewing your solution.