DTC Data Engineering Zoomcamp

Capstone Project

Erik Huß, Cohort 2023

French Radioactive Waste

The dataset to work with is from the [Open platform for French public data](#), and it's publicly available in the open data European Commission [website](#), and the raw data can be seen [here](#).

Problem Statement

This dataset contains information on the location, type, and description of various radioactive waste materials. However, there is a significant issue of uncertainty regarding the type and description of some entries in the dataset. This lack of information can impede the accurate assessment of potential health and environmental risks associated with these materials. Therefore, there is a need to address this uncertainty and ensure that the dataset provides complete and reliable information on radioactive waste materials.

Analysis

The analysis of the radioactive waste dataset revealed several key findings. First, we identified the most common isotopes present in the dataset, which included Cobalt, Caesium and Tritium . This information is crucial for assessing the proper storage and potential health/environmental risks associated with these materials.

Additionally, the facilities with the highest volume of radioactive waste were identified.By understanding which facilities possess the most radioactive waste and its origin, we can develop targeted interventions to reduce the amount of waste produced and minimize potential risks or auditions to ensure the proper storage or disposal of such materials.

Furthermore, we found that the Essonne department has the most residues in the dataset. This information is important for local authorities to take appropriate measures to ensure the safe handling and disposal of radioactive waste in the area.

Finally, our analysis revealed a significant level of uncertainty in the dataset, particularly with regard to missing information on the type and description of some entries. To address this issue, we recommend implementing measures to improve the accuracy and completeness of the dataset. This may include conducting additional surveys, establishing standardized reporting requirements, and improving communication between facilities and regulatory agencies to ensure that all relevant information is captured or by simply updating the dataset when the next iteration of it is published. By addressing these uncertainties, we can ensure that the dataset provides reliable and comprehensive information on these waste materials.

Review Criteria & Explanation

- Problem Description
- Cloud: Google cloud services are used, for access to the project please read the "Google Cloud Project" file in the "README" folder.
- Data Ingestion: Batch/Workflow is used, several steps with Luigi for download, transformation and upload, the code used is in the "Pipeline" folder, Terraform is used to configure and upload to the Google Cloud Services.
- Data Warehouse: Tables need to be created manually. Data is manipulated with BigQuery, both clustered and partitioned for better understanding the "Counties" and "Isotopes" data; code used is both saved in Google BigQuery and the "BigQuery" folder.
- Transformation: Just simple SQL transformation, code used is both saved in Google BigQuery and the "BigQuery" folder, Jupiter was used for most of the data exploration, the notebook can be found in the "Data Exploration" folder, and the data used and transformed can be found in the "Data" folder.
- Dashboard: It can be found here.
- Reproducibility: The whole project is uploaded in GitHub, and instructions are in the "Readme" file from the "README" folder.