Electricity Sector Data Integration & Augmentation

- COMP5339 Project Assignment 1 -

In this data engineering assignment, your team will acquire, integrate and augment electricity sector emissions and generation data from the Australian National Greenhouse and Energy Reporting website, and integrate it with Census data from the ABS. The project aims to provide practical experience in core data engineering skills such as data retrieval, integration, cleaning, transformation, storage, and visualization.

This is a **group assignment**, comprising teams of 2 to 3 students. **The groups formed in Assignment 1 will remain the same for Assignment 2,** which will build on Assign1.

These two assignments represent two stages with separate submissions:

- **Stage 1 (Week 8)**: Submit your Stage 1 report & code by **23:59, Friday 26 September**.
- Stage 2 (Week 12): the instructions for Assignment 2 will be provided later.

All submissions will undergo plagiarism checks.

Assignment 1 Tasks

This Assignment 1 concentrates on acquiring and integrating static datasets about electric energy production and emissions in Australia for historic data and trend analysis. Your team needs to develop Python programs to complete the following tasks:

1. **Data Acquisition**

   Retrieve the following datasets:

   - the electricity sector emissions and generation data for the last ten years (2014 to 2024) from the National Greenhouse and Energy Reporting (NGER) website (https://data.cer.gov.au/datasets/NGER/ID0243);
   - data about the proposed and planned large-scale renewable power stations from the Clean Energy Regulator (https://cer.gov.au/markets/reports-and-data/large-scale-renewable-energy-data); and
   - recent data from the Australian Bureau of Statistics (ABS) Data by Regions portal (https://www.abs.gov.au/methodologies/data-region-methodology/2011-24#data-downloads). Here you have a choice: Out of the data cubes present in ABS, select either ‘population and people’ or ‘economy and industry’ datasets for your analysis and integrate it.

2. **Data Integration and Cleaning**

   Combine the retrieved data into a single, consolidated database. During this process, you may need to clean and pre-process the data to ensure consistency and reliability. Tasks may include handling missing values, converting data types, and filtering out irrelevant or inconsistent data.

3. **Data Augmentation**
   Augment your integrated dataset about large-scale power stations with their geo-location by programmatically querying the geographic coordinates using a public geocoding API (such as Google Maps or OpenStreetMap/Nominatim) for all the energy facilities present. Document methods and API usage.

4. **Data Transformation and Storage**
   Transform the processed and augmented data into a structured format suitable for analysis and visualization. Specifically, you should
   - design a suitable database schema for storage in database, and
   - implement this schema and store your data in either DuckDB or a PostgreSQL database. Whichever system you choose to install, make sure you include the spatial extensions so that we can run some spatial queries in Assignment 2. This should be straight-forward for DuckDB, but when choosing PostgreSQL, make sure PostGIS is included in the chosen install package.

   Important Note: Clearly justify your database design decisions (e.g., normalized or deformalized schema) in your project report. If your group encounters significant difficulties working with a database, you may alternatively store your data in separate CSV files; however, choosing CSV storage will result in a mark penalty.

5. **Documentation and Reporting**
   Maintain detailed documentation throughout your project workflow. Your final report should clearly summaries your methods for data retrieval, integration, cleaning, augmentation, and transformation. Specifically:
   - Highlight key insights and findings.
   - Describe challenges encountered and how you overcame them.
   - Provide recommendations for future improvements.

Deliverables

Submit the following three files via Canvas:

1. Python files (**.py** or **.ipynb**)
   - Consolidate your Python scripts or Jupyter notebooks into **as few files as possible**.
   - You may also submit the additional dataset that you obtained in **Data Augmentation** as a **.txt** file, if it is time-consuming to retrieve this dataset.
2. **requirements.txt** file
   - Include a requirements.txt file listing all Python packages required to run your program within a clean Python virtual environment.
3. Project Report (**.pdf**)
   - Provide a clearly structured report of **up to 4 pages** (with optional additional appendix).

- Recommended sections includes, but are not limited to:
  - ***Dataset Description***: Explain your data retrieve and preprocessing methods.
  - ***Data Exploration***: Describe key exploration activities and findings, including at least one data visualization such as a bar-chart or a map visualization of one chosen aspect of the dataset.
  - ***Data Augmentation***: Identify additional web APIs or data sources used, describe how you accessed these APIs and/or datasets, and justify their relevance.
  - ***Database Design***: Provide a clear diagram of your database schema and justify your design choices.

- You report must include a short paragraph not exceeding 100 words, detailing the contribution of individual team members, example:
  - AXX: Led data collection and preprocessing, …
  - BXX: Developed the data model …
  - …

Submission Deadline

All Assignment 1 deliverables are due by **23:59, Friday 26 September.** Late submissions will incur penalties according to university policy.

Group Member Participation and Marking

This is a group assignment. Please declare the level of participation of each team member as part of your report (see 'Deliverables' above), but also note that the mark awarded for your assignment is conditional on you being able to explain any of your answers to your tutor or the lecturer if asked.

If your group is experiencing difficulties with the content, you should ask on Ed (use a private post if you need to discuss code or report writing directly).

Level of contribution	Proportion of group mark received
No participation.	0%
Minor contributor with at least some understanding of the group submission.	50%
Major contributor to the group's submission.	100%