Task

Objectives of the task

- Ingest the datasets we provide into a database of your choice (although we would prefer a graph database);
- Second, demonstrate the functionality of the database by running queries that we provide

- **Datasets:** The following two datasets need to be ingested into the database (extract the latest version in whatever format you deem appropriate for the pipeline you build)
    - NPASS: https://bidd.group/NPASS/downloadnpass.html
    - NAEB: https://naeb.louispotok.com/

- **ETL pipeline:**
    - **Outline:** Provide an outline of the ETL pipeline you will implement to ingest the data into a database (preferably a graph database)
        - Provide a convincing argument for the data model you choose for the database
        - Provide a lucid database schema for the data model you choose

    - **Implement the ETL pipeline:**
        - Preferably in Python (or a language of your choice) on your machine
        - ETL pipeline should be connected with the database that is ready to be queried
        - Make sure that the code is well written in an object oriented manner with test cases.

Queries to run: Demonstrate the database usability for the following queries

1. Q1: List the compounds associated with 'Analgesic' usage by reporting their pubchem_cid, iupac_name and SMILES
2. Q2: List the compounds (pubchem_cid) associated with a single unique usage; additionally report the associated plants (species names) for each such compound.