ETL Project Report – Exploring Hazardous Substances in Chicago

By Karen Gutzman & Itunu Oyeyipo

Extract:

- Extracted our data from the U.S. National Library of Medicine Hazardous Substances Database (HSDB), and Environmental Protection Agency's Toxic Release Inventory (TRI)
    - https://www.nlm.nih.gov/databases/download/hsdb.html
    - https://www.epa.gov/enviro/data-downloads
- Explored the datasets to better understand the data structure and data type
- Downloaded different formats of data files such as csv and xml then imported them into Jupyter notebook and cleaned it up.
- Encountered challenges in choosing critical information related to the research questions because of enormous data files and limited expertise in the topic area.

Transform:

- Utilized pandas to rename columns that were too long, dropped columns that weren't necessary, and joined relevant columns from all of the files for succinctness.
- Parsed a large xml file using xml.etree.ElementTree as ET and identified the structure of the data to access the critical information.
- Used python *for loop* to append data into lists, and created a data frame using the lists.
- Created data frames and identified columns for connecting data across tables such as CAS numbers & Facility IDs.
- Investigated discrepancies in the CAS number format, and found that Chem_ID is substituted when CAS is unavailable.
- Ensured that the null in the data didn't affect critical columns.

Load:

- Loaded the files into three separate tables in MySQL database using sqlalchemy
- Observed the output to ensure its desired form
- Experienced difficulty connecting sqlalchemy in Jupyter notebook to MySQL database because the packages weren't installed in the virtual environment, rather it was installed globally.
- Worked to install large file storage in GitHub because the HSDB file exceeded normal file size requirement. Troubleshooted by installing LFS package for GitHub
- Due to the richness in data content, these datasets can serve as a comparison for other data sources (such as comparing distance between schools and TRI facilities with carcinogenic substances using latitudes and longitudes)