# Project 2 (ETL) – Proposal

**Group Members**

- Nandhini Nallathambi

- Patricia Colomer

- Razvan Radu

- Siyuan Liang

- Salma Abdirahman

**(E) Data Source:**

The dataset is comprised of 9 CSV files from Kaggle.com, compiled by the CDC (Centres for Disease Control and Prevention), which provides a record of several major historical disease outbreaks. The files include data on aspects such as the type of disease/outbreak, the location and timing of such outbreaks, its death toll and a number of other information.

Link to dataset:<https://www.kaggle.com/datasets/thedevastator/a-comprehensive-history-of-major-disease-outbrea>

**Files:**

* df\_1.csv
* df\_11.csv
* df\_16.csv
* df\_2.csv
* df\_22.csv
* df\_24.csv
* df\_25.csv
* df\_4.csv
* df\_5.csv

**(T) Transformation of Data**

* Decide on what data to use. Drop copies of data or data not required.
* Combine data from CSVs that share similar data.
* Clean up CSV tables, removing unnecessary columns/rows.

**(L) Load of Data**

* Data loaded onto PostgreSQL which is a relational database.

**Outcome:**

From the data, we collected, we hope to create tables that allow us to successfully analyse data of historical pandemics.