Group 3, Project 3 Data Engineering

# **Death Rate of countries and its causes (1990-2019)**

Workflow:

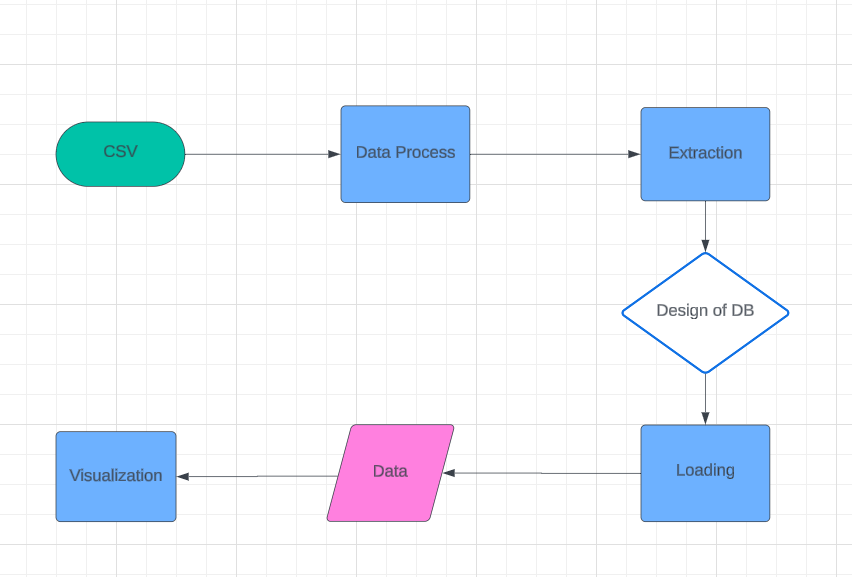
**-> Overview**

Source of the data set: https://www.kaggle.com/datasets/bilalwaseer/death-rate-of-countries-and-its-causes

**-> Understanding the data:**

The dataset is provided in CSV format and was loaded into Jupyter Notebook using Pandas for analysis.

**-> Flowchart:**

****

**-> Data Processing Steps:**

1. **Columns Analysis:**
   * Analyzed columns to identify those falling into the same category.
   * Discovered an additional space in the "High in sodium" column, causing initial errors during column merging.
2. **Deleting Redundant Columns:**
   * Removed redundant columns containing duplicate or similar data.
3. **Renaming Columns:**
   * Renamed columns to adhere to SQL naming standards (no spaces in column names).
4. **Checking for Duplicates and Empty Cells:**
   * Checked for and handled duplicate entries and empty cells in the dataset.
5. **Handling Empty Cells within Country Code:**
   * Identified several empty cells within the "Country code" column.
   * Added country codes for England, Scotland, Wales, and Northern Ireland.
   * Moved regions of the world to a separate DataFrame and deleted them from the main DataFrame.

**-> DB Selection:** Relational Database

**-> Identifying Entities:**

* Country
* Health Indicator

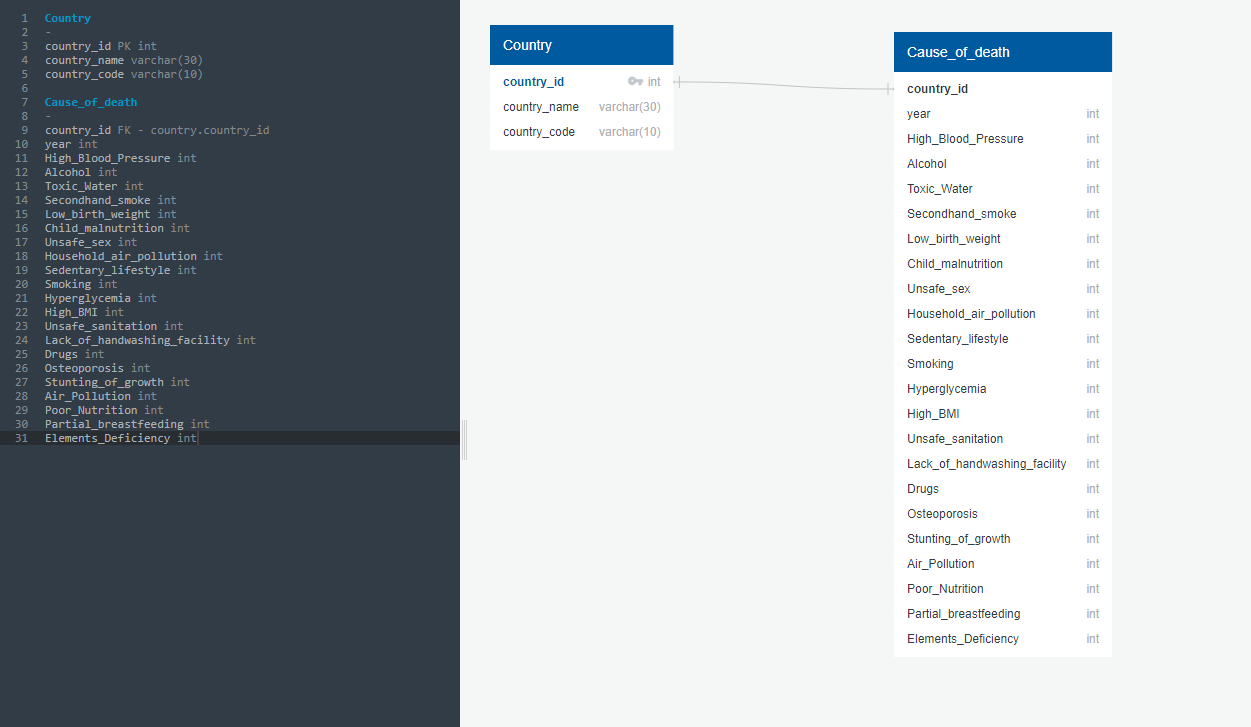
**-> Creating Tables:**

* **Country Table:** Contains information about countries.
* **Health\_Indicator Table:** Contains information about health indicators.

**-> Establishing Primary and Foreign Keys:**

* Primary Key: country\_id

**Entity-Relationship Diagram (ERD):**

****