Lab 6 | 5% of the final Grade | Individual Assignment

ETL Data Cleaning and Loading

Objective:

This assignment aims to test your skills in extracting, transforming, and loading (ETL) data from a CSV file into a SQL database. You will work with a given CSV file and perform a series of data cleaning tasks, storing the cleaned data into different tables in a SQL database called Lab6ETL.

CSV File:

The provided CSV file is named **Lab6ETL.csv** and contains the following columns:

- Region Code
- Region Name
- Country Code
- Country Name
- Year
- Sex
- Age group code
- Age Group
- Number
- Percentage of cause-specific deaths out of total deaths
- Age-standardized death rate per 100,000 standard population
- Death rate per 100,000 population

Tasks:

1. **Task 1: Remove Null Values**
    - Clean the data by removing any rows that contain null values in any of the columns.
    - Load the cleaned data into a table named **CleanedData_NoNulls** in the Lab6ETL database.
2. **Task 2: Filter by Year**
    - From the cleaned data, filter the records to only include data from the year 2020.
    - Load this filtered data into a table named **CleanedData_2020** in the Lab6ETL database.
3. **Task 3: Standardize Age Group Names**
    - Standardize the names of the Age Group column by converting all entries to lowercase and removing any leading or trailing whitespace.
    - Load the cleaned data into a table named **CleanedData_AgeGroup** in the Lab6ETL database.

4. **Task 4: Remove Outliers**
    - Identify and remove outliers in the **Death rate per 100,000 population** column (e.g., values that are significantly higher than the mean).
    - Load the cleaned data into a table named **CleanedData_WithoutOutliers** in the Lab6ETL database.
5. **Task 5: Aggregate Data**
    - Aggregate the data by **Country Name** and calculate the total number of deaths for each country.
    - Load this aggregated data into a table named **CleanedData_Aggregated** in the Lab6ETL database.

Submission Requirements:

1. **Python Scripts:**
    - Submit a Python script for each task **(5)** that includes the code used to clean the data and load it into the respective SQL table.
2. **SQL Database Backup:**
    - Include a backup of the Lab6ETL database after completing all tasks (**You Should have 5 tables**).
3. **Screenshots:**
    - Provide screenshots demonstrating the successful execution of each ETL task, including the creation of tables and loading of data.

Grading Criteria:

- Correctness of the data cleaning tasks.
- Successful loading of data into the SQL tables
- Clarity and organization of Python scripts
- Completeness of submissions