**Clean Water and Sanitation Database Project**

**Problem Definition:**   
There is still **unequal access to clean drinking water** in rural and urban areas. Despite efforts to improve water infrastructure, many rural areas still lack access to clean water, leading to health issues and lower quality of life.

**Objective:**   
Project focus on analyzing the **disparities in access to clean drinking water** between rural and urban areas. The goal is to identify regions with the greatest need for improved water infrastructure and to provide data-driven recommendations for resource allocation.

**Database Design**

**ERD (Entity-Relationship Diagram)**

**Schema Design**

The following are SQL statements to create the database schema:

**CREATE DATABASE water\_sanitation;**

**USE water\_sanitation;**

**CREATE TABLE WaterManagement (**

**RegionID INT,**

**RegionName VARCHAR(100),**

**Type VARCHAR(50), -- Rural or Urban**

**Population INT,**

**SourceID INT,**

**SourceType VARCHAR(50), -- E.g., Well, Piped Water, etc.**

**Capacity INT, -- E.g., Liters per day**

**AccessRate DECIMAL(5, 2), -- Percentage of population with access**

**TestID INT,**

**TestDate DATE,**

**ContaminantLevel DECIMAL(5, 2), -- E.g., mg/L of contaminants**

**TestResult VARCHAR(50), -- E.g., Pass or Fail**

**TotalPopulation INT,**

**ChildrenPopulation INT,**

**AdultPopulation INT,**

**ElderlyPopulation INT,**

**Year INT,**

**AccessPercentage DECIMAL(5, 2), -- Percentage of population with access to clean water**

**PRIMARY KEY (RegionID, SourceID, TestID, Year)**

**);**

**Data Retrieval Queries**

Retrieve all regions with their corresponding water sources and test results:

**SELECT**

**RegionName,**

**SourceType,**

**TestDate,**

**TestResult,**

**ContaminantLevel**

**FROM**

**watermanagement;**

Retrieve regions where the water test failed, along with contaminant levels:

**SELECT**

**RegionName,**

**SourceType,**

**TestDate,**

**ContaminantLevel**

**FROM**

**watermanagement**

**WHERE**

**TestResult = 'Fail';**

Retrieve the total population, including children, adults, and elderly, for each region:

**SELECT**

**RegionName,**

**Population,**

**ChildrenPopulation,**

**AdultPopulation,**

**ElderlyPopulation**

**FROM**

**watermanagement;**

Retrieve the access percentage and capacity for water sources in urban regions:

**SELECT**

**RegionName,**

**SourceType,**

**AccessPercentage,**

**Capacity**

**FROM**

**watermanagement**

**WHERE**

**Type = 'Urban';**

**Data Analysis SQL Queries**

Analyze the average contaminant levels by region type (Urban vs. Rural):

**SELECT**

**Type,**

**AVG(ContaminantLevel) AS AvgContaminantLevel**

**FROM**

**watermanagement**

**GROUP BY**

**Type;**

Analyze the correlation between water source capacity and access percentage:

**SELECT**

**Capacity,**

**AccessPercentage**

**FROM**

**watermanagement;**

Identify regions with the highest and lowest access rates to clean water:

**SELECT**

**RegionName,**

**AccessRate**

**FROM**

**watermanagement**

**ORDER BY**

**AccessRate DESC;**

Analyze the population distribution (children, adults, elderly) in regions with failed water tests:

**SELECT**

**RegionName,**

**ChildrenPopulation,**

**AdultPopulation,**

**ElderlyPopulation**

**FROM**

**watermanagement**

**WHERE**

**TestResult = 'Fail';**

**Importing Data into Excel**

* **Export Data from the Database**:
  + Use SQL queries to export data from the database into CSV or Excel format.
  + Ensure that all necessary tables (e.g., water quality tests, population data, and source types) are included in the export.
* **Import Data into Excel**:
  + Open Excel, go to the **Data** tab, and select **Get Data** > **From File** > **From Workbook**.
  + Navigate to the exported file, select it, and click **Import**.
  + In the import wizard, choose the appropriate data range and sheet where you want to place the data.
  + Verify that the data types (e.g., text, numbers, dates) are correctly imported.

**Ensuring Consistency**

* **Data Formatting**:
  + Review the imported data to ensure that all columns are formatted correctly.
  + For numerical data (e.g., population, contaminant levels), ensure that the columns are formatted as numbers.
  + For date fields, ensure that the data is formatted as dates in Excel.
* **Data Cleaning**:
  + Identify and remove any duplicate records.
  + Check for and resolve any missing or incomplete data.
  + Ensure that all data values are consistent with the expected ranges (e.g., population numbers should not be negative, percentage values should be between 0 and 100).
* **Data Validation**:
  + Set up data validation rules in Excel to prevent incorrect data entry. For example, restrict certain columns to accept only numeric values or dates.

**2. Testing for Accuracy and Consistency**

**Verify Data Accuracy**

* **Cross-check with Source Data**:
  + Compare a sample of the imported data against the original database or raw data to ensure accuracy.
  + Perform spot checks on key metrics (e.g., contaminant levels, population counts) to verify that they match the database.
* **Recalculate Derived Metrics**:
  + If your dataset includes calculated fields (e.g., percentage of the population with access to clean water), recalculate these in Excel to confirm that they are accurate.

**2.2. Test Data Consistency**

* **Check for Data Anomalies**:
  + Use Excel functions like COUNTIF or SUMIF to identify any anomalies (e.g., regions with zero population, negative contaminant levels).
  + Create pivot tables to analyze and visualize the data, making it easier to spot inconsistencies.
* **Test Pivot Tables and Charts**:
  + Verify that the pivot tables and charts created from the imported data behave as expected.
  + Apply different filters and slicers to the pivot tables to ensure that the data updates consistently across all visualizations.