Public Schools Dataset

1. The Problem of the Dataset

This dataset highlights the distribution and accessibility of public schools, which can reveal disparities in education access, particularly in underserved regions.

Additional Problem Identified:
The dataset does not include student demographic information, which could provide deeper insights into how well public schools are serving diverse populations. Without this data, it may be harder to analyze equity issues in access to quality education across different ethnic, socioeconomic, or gender groups.

2. Reason Behind Selection

It provides critical insights into how well public schools are distributed across different regions, making it ideal for spatial and accessibility analysis.

Additional Reason for Dataset Selection:
The availability of geolocation data (longitude and latitude) allows for advanced spatial analyses, such as calculating the distance between schools and student populations or identifying underserved areas that lack adequate access to public education facilities. This makes the dataset highly valuable for urban planning and education policy development.

3. Problem Being Solved

The analysis will focus on understanding public schools' geographic distribution and accessibility to students. It can also help identify areas needing more educational resources.

Additional Problems Solved by Analysis:

- **Resource Allocation: The dataset can help determine which schools are overcrowded or under-resourced, allowing policymakers to allocate funds and resources more efficiently.**
- **Access to Specialized Schools: Analyzing the availability of different school categories (elementary, middle, high) in relation to population density can reveal whether students in certain areas have adequate access to the level of education they need.**
- **Impact of School Location on Student Performance: By comparing school locations to student performance data (if available), the dataset could also be used to analyze the impact of travel distance or school accessibility on student outcomes.**

4. Key Columns:

- **School Name (Categorical):**
    - **Identifies each public school.**
- **Category (Categorical):**
    - **School type (high, middle, elementary).**
- **ZIP Code (Categorical):**
    - **Helps analyze distribution by region.**
- **Longitude/Latitude (Numerical):**
    - **Used for mapping and spatial analysis.**
- **Phone (Categorical):**
    - **Useful for contacting schools.**
- **Address (Categorical):**
    - **Useful for mapping school locations.**

Additional Column Ideas:

- **School Capacity (Numerical):**
    - **Describes the maximum number of students each school can accommodate, which would be useful for determining overcrowded schools.**
- **Student Enrollment (Numerical):**
    - **Tracks the current number of students, allowing for an analysis of capacity vs. actual enrollment to identify underused or overcrowded schools.**

5. Data Cleaning Techniques

- **Geolocation Cleaning:**
    - **Verify and correct longitude and latitude data to ensure schools are accurately mapped.**
- **Handling Duplicates:**
    - **Removing duplicate entries for schools to maintain data integrity.**
- **Standardization:**
    - **Ensuring consistency in categories (e.g., "High School" vs "HS").**

Additional Data Cleaning Techniques:

- **Address Standardization:**
    - **Normalize address formats for consistency, especially when conducting geospatial analyses.**
- **Handling Missing Data:**
    - **Use imputation or removal techniques for missing fields, such as ZIP codes or school categories.**
- **Validation of Contact Information:**
    - **Check and validate phone numbers and addresses to ensure they are up-to-date, especially for use in communication with the schools.**