Dataset : Public Schools Dataset

1. The Problem of the Dataset

This dataset highlights the distribution and accessibility of public schools across various regions. It can reveal significant disparities in education access, particularly in underserved areas where resources are limited. Understanding these disparities is crucial for policymakers and educators aiming to improve educational equity and ensure that all students have access to quality education.

2. Reason Behind Selection

The dataset provides critical insights into the geographic distribution of public schools, allowing for spatial and accessibility analysis. By examining factors such as school density, regional demographics, and socioeconomic status, stakeholders can identify patterns that inform decisions about resource allocation, infrastructure development, and educational policy. This makes the dataset ideal for researchers, educators, and policymakers focused on enhancing educational access and equity.

3. Problem Being Solved

The analysis aims to address several key issues:

- **Geographic Distribution:** Understanding how public schools are distributed geographically and identifying areas with limited access.
- **Accessibility Analysis:** Evaluating how easily students can access schools, factoring in transportation, socio-economic barriers, and population density.
- **Resource Allocation:** Identifying regions that require more educational resources, such as funding, facilities, and staff, to improve educational outcomes.
- **Policy Recommendations:** Providing data-driven recommendations to policymakers to target interventions in underserved communities, ensuring equitable access to education.

4. Key Columns:

- **School Name (Categorical):**
    - Identifies each public school, allowing for easy reference and communication.
- **Category (Categorical):**

    - Describes the type of school (e.g., high school, middle school, elementary school), aiding in categorization for analysis.
- **ZIP Code (Categorical):**
    - Facilitates the analysis of school distribution by region and helps correlate school accessibility with community demographics.
- **Longitude/Latitude (Numerical):**
    - Essential for mapping school locations and conducting spatial analyses to visualize geographic disparities.
- **Phone (Categorical):**
    - Provides contact information for each school, useful for inquiries and communication with school administrators.
- **Address (Categorical):**
    - Detailed school address for mapping purposes and to ensure accurate geographic representation.
- **Total Enrollment (Numerical):**
    - Indicates the number of students enrolled in each school, which can help assess school capacity and resource needs.
- **Free/Reduced Lunch Percentage (Numerical):**
    - Reflects the socio-economic status of the student body, providing insight into the demographics of the schools served.
- **Academic Performance Metrics (Numerical):**
    - Data on standardized test scores or graduation rates to evaluate school performance relative to resources.

5. Data Cleaning Techniques:

- **Geolocation Cleaning:**
    - Verify and correct longitude and latitude data to ensure accurate mapping of schools, eliminating errors that could skew spatial analysis.
- **Handling Duplicates:**
    - Remove duplicate entries for schools to maintain data integrity and avoid inflating counts during analysis.
- **Standardization:**
    - Ensure consistency in category naming (e.g., "High School" vs. "HS") to facilitate accurate comparisons across data points.
- **Missing Value Imputation:**

    - Identify and handle missing values, especially in critical columns like total enrollment and academic performance, using appropriate imputation methods.
- **Address Normalization:**
    - Standardize address formats (e.g., street abbreviations) to ensure consistency and accuracy in mapping.
- **Categorical Encoding:**
    - Convert categorical variables to numerical formats (if needed) for machine learning models while maintaining interpretability.
- **Validation Against External Data:**
    - Cross-reference the dataset with official educational databases or census data to validate accuracy and completeness.