Dataset 1: Kwena-Public Schools Survey

1. The Problem of the Dataset:

This dataset contains detailed information about various public schools, including high, middle, and elementary schools. The dataset's main issue revolves around how schools are distributed across regions, their accessibility to students, and how well school resources are managed. Challenges may arise in analyzing the relationship between school locations, student demographics, and the quality of education. Understanding geographical disparities in education is a critical issue for policymakers and local governments.

2. Reason Behind the Selection of This Dataset:

This dataset is ideal for educational analysis because it contains valuable details such as school names, addresses, cities, and geolocation data (longitude and latitude). These features make it suitable for studying accessibility and equity in school placement, urban planning, and community resource allocation. Additionally, the dataset can be used for enhancing transportation planning, determining areas for new school construction, and providing insights for public service improvements. The inclusion of contact information and website links enables researchers to evaluate digital accessibility as well.

3. Problems Being Solved by Analysis of the Dataset:

- **Geographical Distribution of Schools:** Investigate whether schools are evenly spread across different cities and regions, identifying areas that might lack proper educational facilities.
- **Accessibility for Students:** Analyze the proximity of schools to students' residences to understand if some regions are underserved.
- **School Facilities and Services:** Using URLs, schools' online resources or services can be assessed to identify gaps in digital accessibility.
- **School Infrastructure and Community Needs:** Address disparities in school infrastructure, especially in low-income neighborhoods.
- **Urban Planning and School Placement:** Provide insights for city planners on optimizing school placements based on population density, future growth, and transportation routes.

4. List of Columns and Their Descriptions:

- **CATEGORY (Object):** The type of school (e.g., Elementary, Middle, High) represented in the dataset.
- **SCHOOL NAME (Object):** The official name of the school.
- **ADDRESS (Object):** The street address where the school is located.
- **CITY (Object):** The city in which the school operates.
- **ZIP CODE (Integer):** The postal code for the school's location.

- **PHONE (Object):** The contact phone number for the school.
- **URL (Object):** The school's official website URL, providing access to additional information and resources.
- **LONGITUDE (Float):** The longitudinal geographic coordinate of the school for mapping purposes.
- **LATITUDE (Float):** The latitudinal geographic coordinate of the school for precise location tracking.
- **LOCATION (Object):** A combined field including the address and geolocation data (longitude, latitude) for complete geographical identification.

5. Data Cleaning Techniques:

- **Removing Duplicates:** Ensures there are no repeated entries of the same school, preventing skewed results in analysis.
- **Validating Geographical Coordinates:** Check longitude and latitude for correctness, ensuring they fall within the expected geographic range to avoid misplacement.
- **Standardizing Address Formats:** Ensure consistency in how addresses are represented (e.g., using abbreviations like "Rd." instead of "Road"), which improves the accuracy of location-based analysis.
- **Phone Number Standardization:** Format phone numbers consistently (e.g., using the international format +1-xxx-xxx-xxxx), making it easier for users to contact schools.
- **Handling Missing or Incorrect Data:** Identify missing values in critical columns like ZIP CODE, LONGITUDE, and LATITUDE, filling in gaps with external resources or using averages.
- **URL Validation:** Check all URLs to ensure they are working and link to the appropriate school websites, flagging or correcting any that are broken or outdated.