Name: B.kranthi kumar

Roll no: 2211CS010069

SECTION: S6

SEM ID: 6

REPORT OF THE PROJECT

1. Dataset Description

The dataset analyzed is named Bangalore Water Supply and Sewerage Board "BWSSB_Water_Consumption_August_2025_Extended_Filled.csv", loaded into a Spark session for scalable processing. It comprises both categorical and numerical columns, suitable for government-led analytics tasks such as urban water management, resource allocation, ward-level comparisons, efficiency performance, and consumption pattern analysis.

Domain Description:

- **The domain is focused on urban water utility management and municipal water consumption analytics, specifically for Bengaluru city under BWSSB (Bangalore Water Supply and Sewerage Board) jurisdiction.**
- **The data represents water consumption patterns across 2,653 wards, including both real municipal wards and synthetic wards for extended analysis.**
- **Data is cleaned to remove nulls and ensure completeness, with columns processed for analytical convenience and feature engineering applied for enhanced insights.**
- **Data types are strategically separated for analysis: categorical columns (ward names, income levels, ward types) vs. numerical columns (consumption metrics, connection counts, efficiency scores, household sizes).**

2. Observed Insights & Hidden Facts

The notebook conducts multifaceted exploratory data analysis using:

Summary statistics generation to discover means, ranges, distributions, and variations across all numerical columns, revealing consumption patterns from 0-499.88 ML and efficiency scores from 0.5-1.0.

Correlation analysis utilizing heatmaps and scatter plots to identify strongly associated fields, revealing underlying trends such as:

- **Strong positive correlation (0.65) between number of connections and consumption volume**
- **Moderate relationships between household size and consumption patterns**

- **Independent relationship between efficiency scores and total consumption volumes**

Visualization gallery with comprehensive plotting functions:

- **Box plots and violin plots to show consumption distribution across income levels and ward types**
- **Bar charts and comparative plots to analyze consumption patterns by demographic factors**
- **Scatter plots and correlation matrices to identify consumption-efficiency relationships and outliers**
- **Distribution histograms and cumulative plots to highlight consumption inequality and resource distribution**
- **Geographic-style comparisons between real and synthetic ward performances**

Hidden facts and non-obvious insights include:

- **The discovery of significant consumption inequality where top 20% of wards consume 52% of total water resources while bottom 50% consume only 18%**
- **The efficiency paradox revealing no direct correlation between high consumption and low efficiency, suggesting behavioral factors beyond technical efficiency**
- **Income-level consumption patterns showing high-income wards average 298.7 ML vs low-income wards at 217.9 ML, with middle-income wards demonstrating most balanced patterns**
- **Optimal household size identification at 4 persons for best water efficiency across consumption levels**
- **Ward-type analysis showing real wards exhibit more consistent patterns while synthetic wards display wider variance and extreme values**
- **Regional consumption clusters with eastern Bengaluru showing higher consumption density and northern regions demonstrating better efficiency**

3. Recommendations

Based on the findings from the dataset and analytical workflow, the following actionable recommendations are proposed:

Targeted Infrastructure Interventions:
Direct additional monitoring resources and efficiency programs to high-consumption wards identified through outlier analysis, particularly those with consumption above 400 ML and efficiency below 0.7. Implement tiered water pricing for consumption above 300 ML to encourage conservation in high-usage areas.

Resource Allocation Optimization:
Use observed correlations between connection counts and consumption volumes to optimize future infrastructure planning and resource distribution. Allocate maintenance resources based on consumption patterns and efficiency metrics to maximize operational returns and reduce non-revenue water.

Strategic Monitoring and Reporting:
Establish regular data analytics dashboards using the visualization patterns developed in this analysis for BWSSB stakeholders. Enable quick assessment of ward-level performance, early detection of consumption anomalies, and real-time policy adjustments based on emerging patterns.

Continuous Data Quality Improvement:
Maintain systematic data cleaning procedures as demonstrated (null handling, type validation, feature engineering) and standardize data collection formats across all wards for easier integration and longitudinal analysis. Implement automated data validation pipelines for ongoing data quality assurance.

Foster Data-Driven Water Governance:
Encourage BWSSB departments and urban planning authorities to utilize this automated analytics workflow, ensuring water management decisions are based on empirical consumption patterns, efficiency correlations, and performance data rather than assumptions or traditional approaches.