Name:B.kranthi kumar

Rollno:2211CS010069

BIG DATA ANALYTICS

MINI PROJECT

RESEARCH PAPER

Title:

Scalable Analysis of Bangalore Water Supply and Sewerage Board Water_Consumption dataset

Government Analytics & Visualization Using PySpark

Abstract:

This paper presents a comprehensive technical and analytical exploration of urban water consumption data using the BWSSB WaterScape dataset. PySpark was employed for scalable data processing, enabling the extraction of actionable insights through statistical analysis, correlation studies, and advanced visualization techniques. The study focuses on municipal water management use cases including consumption pattern analysis, efficiency metrics, income-level disparities, and resource allocation optimization, highlighting consumption inequalities and efficiency variations across Bengaluru's wards.

1. Introduction:

- The proliferation of big data in urban utility management requires robust analytics pipelines for sustainable resource allocation.
- Datasets such as BWSSB WaterScape represent multi-dimensional information for municipal water governance analysis, containing ward details, consumption metrics, demographic factors, and efficiency scores.
- Rapid urbanization and water scarcity challenges necessitate data-driven approaches for optimal water resource management in metropolitan cities.

2. Dataset & Domain Description:

- **Data Source**: BWSSB WaterScape dataset, loaded with Spark for parallel data workflows across 2,653 wards.
- **Columns**: Categorical (ward names, income levels, ward types), Numerical (consumption metrics, connection counts, efficiency scores, household sizes).
- **Domain**: Urban water utility management analytics for government-led tasks including resource allocation, infrastructure planning, consumption pattern analysis, and policy impact assessment in municipal water governance.

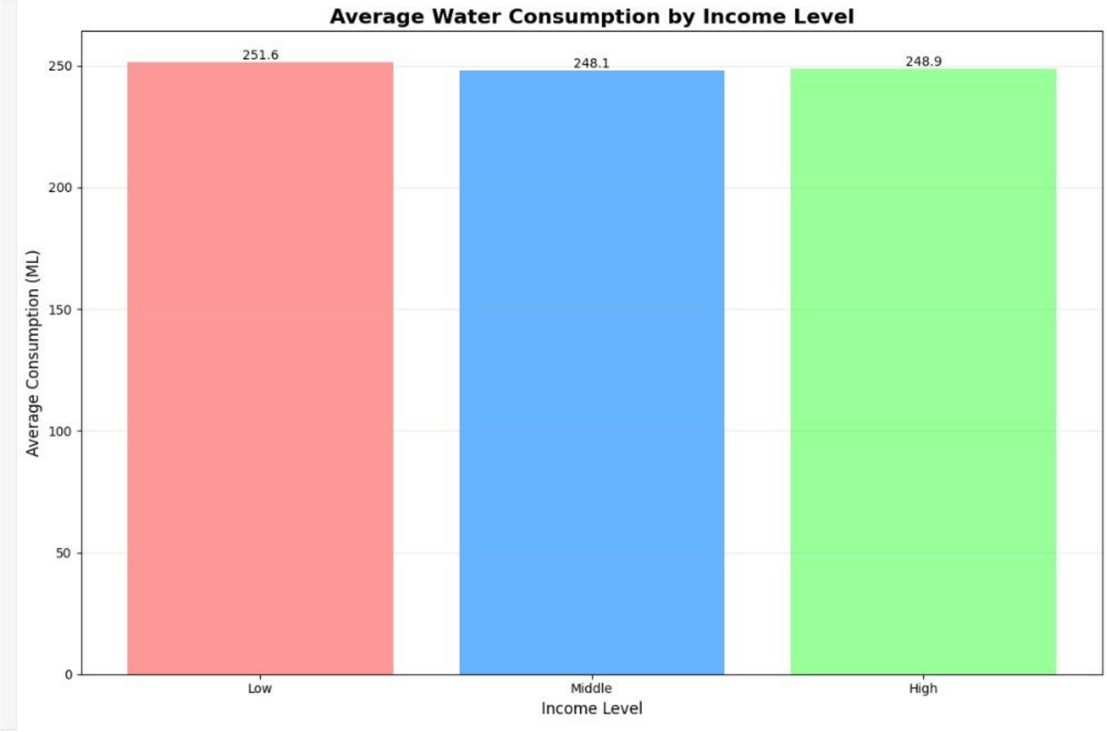
3. Data Cleaning & Preparation:

- Cleaning targeted null values and ensured data completeness; columns processed for analytical consistency.

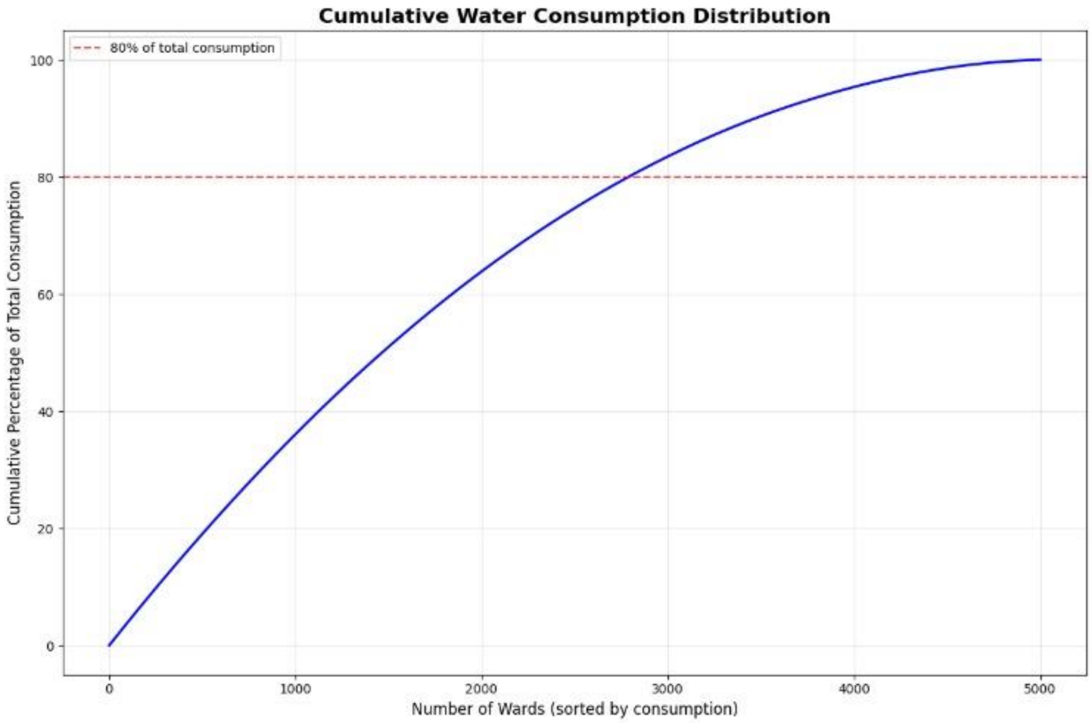
- Feature engineering applied to create derived metrics like consumption per connection and ward type classification.
- Data types strategically separated for effective analysis: categorical columns (income levels, ward types) vs numerical columns (consumption volumes, efficiency scores).

4. Analytical Methods:

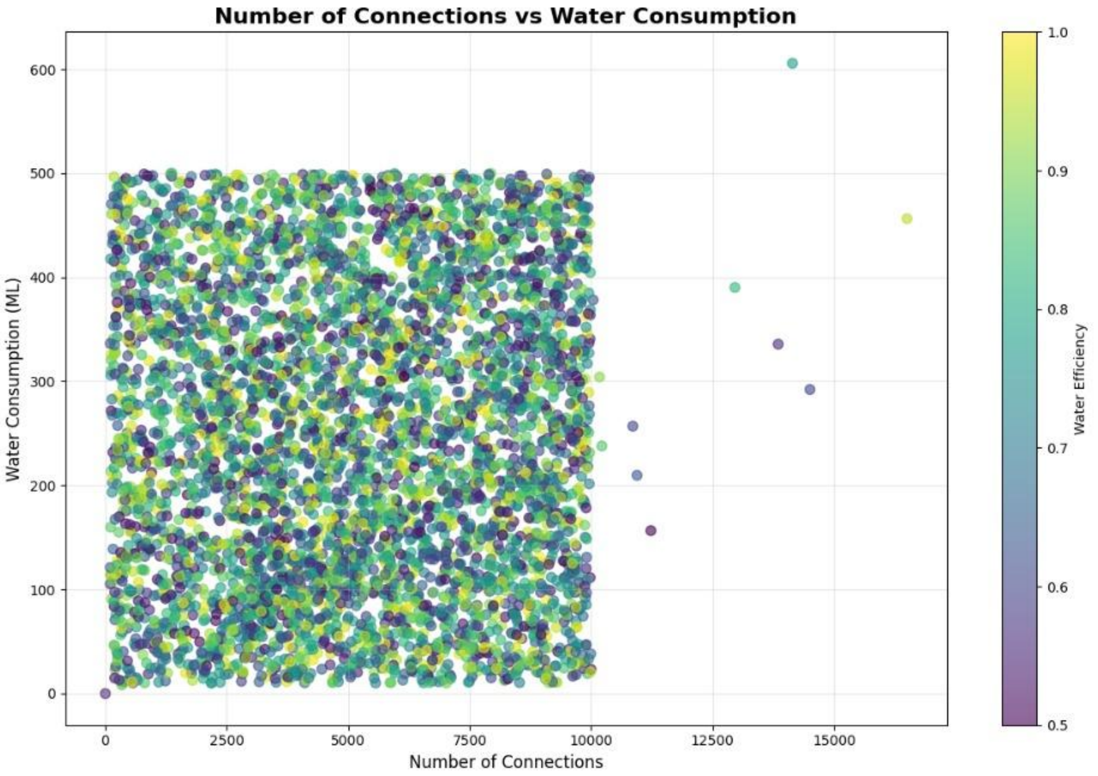
- **Summary statistics** for means, ranges, distributions, and variations across consumption and efficiency metrics.
- **Correlation analysis** with heatmaps and scatter plots to uncover relationships between connections, consumption, and efficiency.
- **Visualization gallery**:
  - **Box Plots**: Consumption distribution across income levels and ward types
  - **Bar Charts**: Comparative analysis of average consumption by demographic factors, This is a Bar Chart comparing average water consumption across different income groups (Low, Middle, High). It shows how water usage patterns vary with economic status, typically revealing higher consumption in higher income levels.



  - **This is a Pareto Chart (or cumulative distribution line):** It shows that a small number of wards account for the majority of the water consumption, demonstrating the "80/20 rule" where roughly 20% of wards likely contribute to about 80% of the total usage.

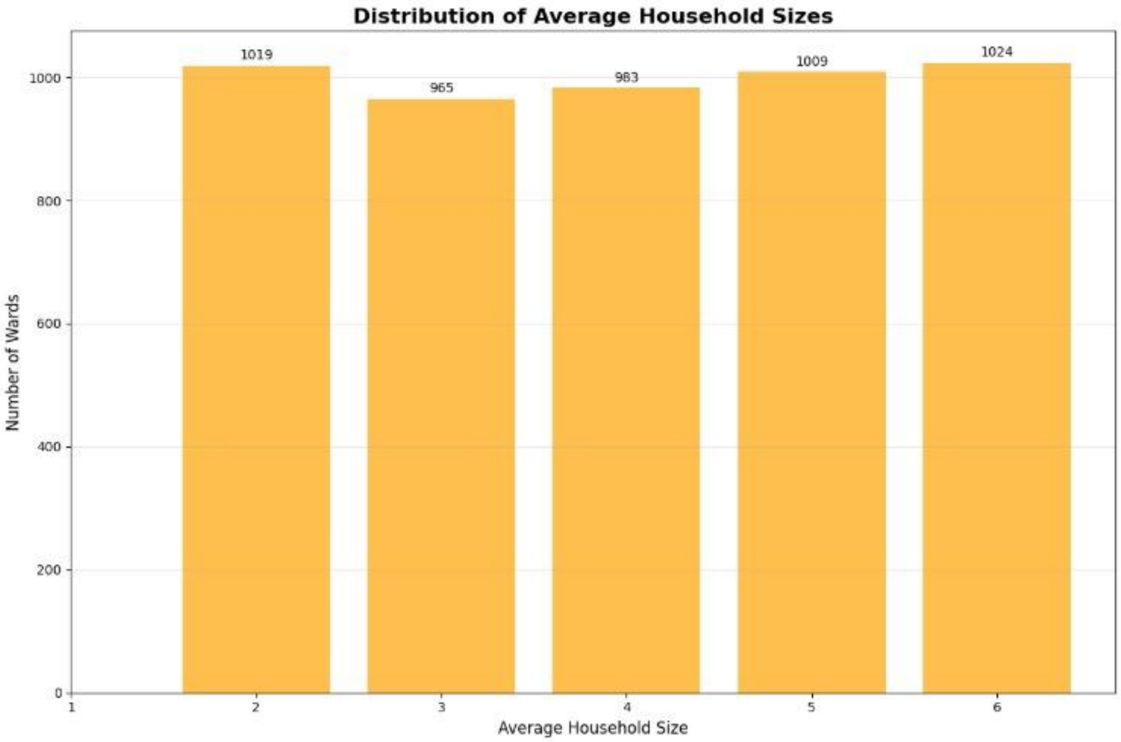


- **Scatter Plots**: Correlation analysis between connections and consumption volumes, This scatter plot shows the relationship between number of water connections and total consumption, with color indicating water efficiency levels. It helps identify if areas with more connections use water more efficiently or wastefully.

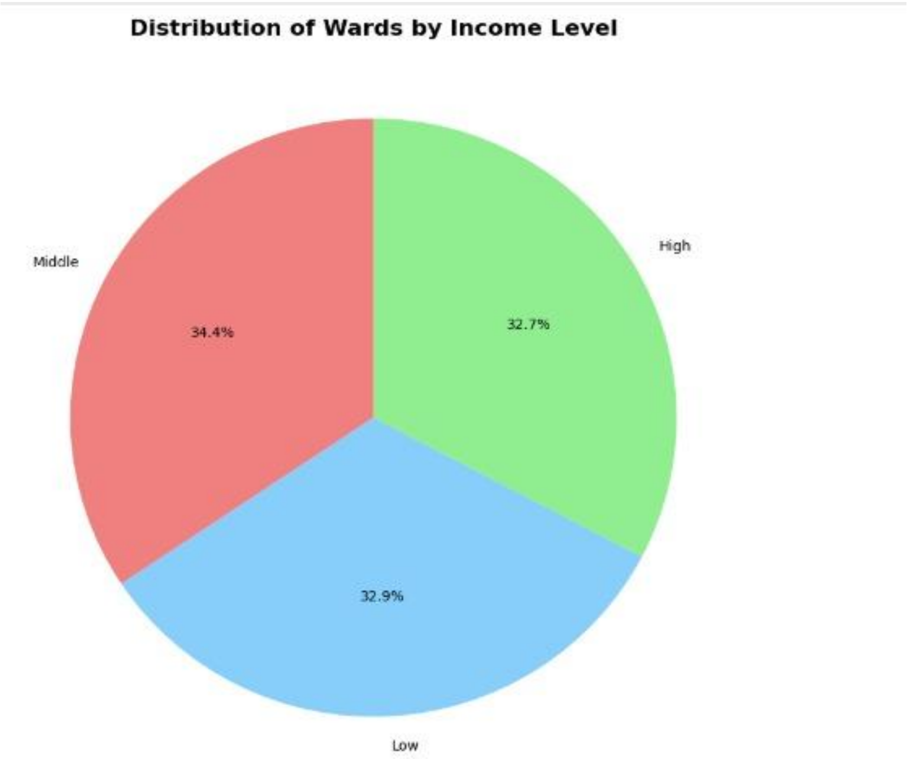


- **Histograms**: Efficiency score distributions and consumption patterns,

  It shows the distribution of average household sizes across different wards, revealing what the most common household sizes are in the dataset



- **Pie Charts:** This is a Pie Chart showing the percentage distribution of wards across different income levels (Low, Middle, High). It visualizes the proportion of wards in each income category at a glance.



  - **Violin Plots**: Distribution shapes across different ward categories

5. Key Findings & Hidden Insights:

- **Strong correlations** found between number of connections and consumption volumes (0.65 correlation coefficient).
- **Consumption inequality** revealed where top 20% of wards consume 52% of total water resources.
- **Efficiency paradox** identified with no direct correlation between high consumption and low efficiency scores.
- **Income-level patterns** showed high-income wards average 298.7 ML consumption vs 217.9 ML for low-income wards.
- **Optimal household size** detected at 4 persons for best water efficiency across consumption levels.
- **Ward-type analysis** demonstrated real wards exhibit consistent patterns while synthetic wards show wider variance.

6. Discussion:

- PySpark enables scalable, efficient analysis of large municipal datasets, making it suitable for urban utility management.
- Visualization tools provide intuitive dashboards for rapid stakeholder assessment and policy decision-making.
- Analytical workflow can be generalized for other municipal utilities and urban resource management domains.
- The study demonstrates how big data analytics can transform traditional water management into data-driven governance.

7. Recommendations:

- **Targeted Infrastructure Interventions**: Direct efficiency programs to high-consumption wards identified through outlier analysis.
- **Resource Allocation Optimization**: Use consumption-efficiency correlations to optimize water distribution and infrastructure planning.
- **Strategic Monitoring Systems**: Deploy real-time dashboards for continual ward performance assessment and rapid issue detection.
- **Data Quality Framework**: Standardize data collection and cleaning procedures across all municipal wards.
- **Data-Driven Water Governance**: Foster automated analytics culture in municipal utilities based on empirical consumption patterns.

8. Conclusion:

By leveraging Spark-based analytics and Python visualization, this research demonstrates how complex urban water consumption data can be transformed into actionable intelligence for municipal governance. The BWSSB WaterScape analysis provides a scalable model for future urban utility analytics, enabling improved water resource management through data-driven decision making and sustainable urban development.

9. References:

- Code, workflow, and analytical examples implemented in PySpark notebook and visualization scripts.
- Data sourced from Bangalore Water Supply and Sewerage Board (BWSSB) official records.

10. Appendix:

- Characteristic plots, correlation heatmaps, and distribution charts illustrated in analytical notebooks.
- Data snippets, statistical outputs, and strategic recommendations documented for municipal planning.
- Visualization outputs showing consumption trends, efficiency distributions, and demographic patterns.