**OPTMIMIZING WATER DISTRIBUTION IN HARARE**

**About the dataset**

**1. Introduction**

This document provides a detailed overview of the **Water Distribution Dataset for Harare**, covering the years 2020 through 2024. The dataset simulates water availability and demand across 50 unique areas in Harare, Zimbabwe, on a monthly basis. It is designed to aid researchers, urban planners, and policymakers in understanding water distribution priorities and challenges in the region.

**2. Dataset Description**

**2.1 Purpose and Scope**

The dataset aims to provide a comprehensive synthetic representation of water distribution factors in Harare, enabling:

* Prioritization of water allocation based on demand and availability.
* Analysis of population and industrial impacts on water resources.
* Support for decision-making in urban water management.

**2.2 Geographical and Temporal Coverage**

* **Geographical Coverage:** 50 distinct areas within Harare, identified by unique numeric AreaIDs (1 to 50). These areas represent diverse urban, suburban, and industrial zones.
* **Temporal Coverage:** Monthly data from January 2020 to December 2024, totaling 5 years of data.

**3. Dataset Structure and Content**

**3.1 Data Format and Volume**

* **File Format:** CSV (Comma-Separated Values)
* **Total Records:** 3,000 (50 areas × 5 years × 12 months)
* **Data Types:** All numeric columns are integers for consistency and ease of analysis.

**3.2 Detailed Column Descriptions**

| **Column Name** | **Description** | **Data Type** | **Range / Values** |
| --- | --- | --- | --- |
| **AreaID** | Unique identifier for each geographical area within Harare. | Integer | 1 to 50 |
| **YearID** | Numeric identifier for the year, where 1 = 2020, 2 = 2021....... 5 = 2024. | Integer | 1 to 5 |
| **month** | Month of the year for the data record. | Integer | 1 (January) to 12 (December) |
| **water\_availability** | Percentage of water availability in the area relative to demand. | Integer | 10 to 100 (%) |
| **population\_density** | Number of people per square kilometer in the area. | Integer | 50 to 2000 people/km² |
| **industrial\_activity** | Index representing the intensity of industrial activity influencing water demand. | Integer | 0 to 100 (index) |
| **high\_priority\_class** | Scale indicating availability of water-dependent jobs/businesses (1 = very low, 5 = very high). | Integer | 1 to 5 |
| **distribution\_priority** | Priority level for water distribution based on combined factors. | Integer | 0 = Low, 1 = Medium, 2 = High |

**Explanation of Columns and Values**

**AreaID**

* Represents a unique area within Harare.
* Numeric IDs from 1 to 50 correspond to distinct neighborhoods or districts.
* Used as a key to link with other geographic or demographic datasets.

**YearID**

* Encodes the calendar year for easier numerical processing.
* YearID = 1 corresponds to 2020, incrementing by 1 each year up to 2024.
* Facilitates time series analysis without string date parsing.

**Month**

* Indicates the month of the year (1 for January through 12 for December).
* Enables seasonal and monthly trend analysis.

**Water Availability**

* Integer percentage representing the proportion of water supply available relative to demand.
* Values range from 10% (very low availability) to 100% (full availability).
* Critical for identifying areas facing water scarcity.

**Population Density**

* Number of residents per square kilometer.
* Reflects urban density and potential water demand pressure.
* Higher values indicate densely populated areas.

**Industrial Activity**

* An index from 0 to 100 quantifying industrial water demand intensity.
* Higher values indicate more industrial operations consuming water.
* Useful for understanding non-domestic water usage.

**High Priority Class**

* A categorical scale from 1 (very low) to 5 (very high).
* Represents the availability of jobs or businesses that rely heavily on water.
* Helps identify economic zones where water supply is critical.

**Distribution Priority**

* Computed priority for water distribution based on combined factors.
* Values:
  + 0 = Low priority (normal conditions)
  + 1 = Medium priority (moderate concern)
  + 2 = High priority (urgent need)
* Assigned using the following rules:
  + High priority if water availability < 30% **and** population density > 1000, **or** high priority class ≥ 4.
  + Medium priority if water availability < 50% **and** industrial activity > 60.
  + Otherwise, low priority.

**4. Data Generation Methodology**

**4.1 Synthetic Data Generation Process**

* Data values were generated using pseudo-random number generators with fixed seed (np.random.seed(42)) to ensure reproducibility.
* Ranges for each variable were chosen based on typical urban water distribution scenarios.
* The dataset simulates realistic but synthetic conditions to support modeling and analysis without revealing sensitive real-world data.

**4.2 Priority Assignment Logic**

* The priority classification combines multiple variables to reflect water distribution urgency.
* The logic balances scarcity (low water availability), demand pressure (population and industrial activity), and economic importance (high priority class).

**4.3 Assumptions and Limitations**

* The dataset is **not** based on real-time measurements or official statistics.
* Area identifiers are anonymized and do not correspond to official administrative boundaries.
* Population density and industrial activity are static within each month and do not account for intra-month fluctuations.
* Users should treat the dataset as a synthetic benchmark for modeling rather than a definitive source of water distribution data.

5. Usage Guidelines

**5.1 Accessing and Loading the Dataset**

The dataset is stored as a CSV file named csv\_water\_distribution.csv in Google Drive.

**Example: Loading in Python (Google Colab)**

*f* ***rom*** *google.colab* ***import*** *drive*

***import*** *pandas* ***as*** *pd*

*drive.mount('/content/drive')*

*csv\_path = '/content/drive/My Drive/csv\_water\_distribution.csv'*

*data = pd.read\_csv(csv\_path)*

***print****(data.head())*

**5.2 Example Applications**

* **Water Distribution Modeling:** Prioritize water delivery based on area needs.
* **Urban Planning:** Analyze how population and industry affect water demand.
* **Policy Simulation:** Test impact of interventions on water scarcity.
* **Machine Learning:** Train predictive models for water shortages or demand forecasting.

**6. Glossary**

* **Water Availability:** Percentage of water supply relative to demand.
* **Population Density:** Number of people per square kilometer.
* **Industrial Activity:** Index of industrial water consumption.
* **High Priority Class:** Scale indicating critical water-dependent economic activity.
* **Distribution Priority:** Urgency level for water allocation.

**7. Contact Information**

For questions or support, please contact:

**Name:** Angeline Tsatsa  
**Email:** [ATsatsa@HarareCity.co.zw](mailto:ATsatsa@HarareCity.co.zw)  
**Organization:** City Of Harare  
**Date:** 05 May 2025