**Overview:**

The Omdena project focuses on predicting dengue fever outbreaks in Sri Lanka using machine learning models. This summary document provides an overview of the data collection process, including weather data, dengue case data, geographical location data, and population data. This document is intended for the data pre-processing and data analysis teams to ensure a smooth transition and understanding of the data collected.

1. **Dengue Cases**

* **Source**: <https://www.epid.gov.lk/weekly-epidemiological-report>
* **Time Period:** Data spans from Jan 2007 to Aug 2024.
* **Frequency:** Weekly.
* **Geographical Coverage:** Data covers all 25 districts in Sri Lanka.
* **Data Points:**
* Dengue Cases
* **Data Quality Consideration**
* **Missing Data -** Given the time period, all weekly data is available for each district.
* **Granularity** - The dataset provides weekly data at the district level.
* **Temporal coverage** - The dataset covers a long time span (2007-2024), making it highly valuable for analyzing long-term trends and patterns.

1. **Weather Data**

* **Sources:** [**https://open-meteo.com/en/docs**](https://open-meteo.com/en/docs)
* **Time Period:** Data spans from Jan 2007 to Aug 2024.
* **Frequency:** Weekly.
* **Geographical Coverage:** Data covers all 25 districts in Sri Lanka.
* **Data Points:**
* **Avg Max Temp (°C) - temperature\_2m\_max**: Average daily maximum temperature.
* **Avg Min Temp (°C) - temperature\_2m\_min**: Average daily minimum temperature.
* **Avg Apparent Max Temp (°C) - apparent\_temperature\_max**: Average daily maximum apparent temperature (feels-like temperature), considering factors like humidity and wind speed.
* **Avg Apparent Min Temp (°C) - apparent\_temperature\_min**: Average daily minimum apparent temperature (feels-like temperature), considering factors like humidity and wind speed.
* **Total Precipitation (mm) - precipitation\_sum**: Sum of daily precipitation, including rain, snow, and other forms of moisture.
* **Total Rain (mm) - rain\_sum**: Sum of daily rainfall amounts.
* **Avg Wind Speed (km/h) - wind\_speed\_10m\_max**: Average daily wind speed.
* **Max Wind Gusts (km/h) - wind\_speed\_10m\_max**: Maximum wind gusts recorded daily.
* **Weather Code**: Numerical code representing specific weather conditions (e.g., clear, cloudy, rainy).
* **Avg Daylight Duration (hours)**: Average number of daylight hours per day.
* **Avg Sunrise Time**: Average time of sunrise.
* **Avg Sunset Time**: Average time of sunset.
* **Data Quality Considerations**
* **Missing Data -** Given the time period, all weekly data is available for each district.
* **Granularity** - The dataset provides weekly data at the district level.
* **Temporal coverage** - The dataset covers a long time span (2007-2024), making it highly valuable for analyzing long-term trends and patterns.

1. **Population Data**

* **Source:** [**https://www.rgd.gov.lk/web/images/statistic/Mid-year-population-by-district--Sex.pdf**](https://www.rgd.gov.lk/web/images/statistic/Mid-year-population-by-district--Sex.pdf)
* **Dataset location -** [**https://drive.google.com/file/d/1rgKnGoelrNyzm\_EJ7LY1iA44ZVNU1aZW/view**](https://drive.google.com/file/d/1rgKnGoelrNyzm_EJ7LY1iA44ZVNU1aZW/view)
* **Time Period:** Data spans from 2014 to 2023.
* **Frequency:** Yearly
* **Geographical Coverage:** Data covers all 25 districts in Sri Lanka.
* **Data Points:**
* male
* female
* **Data Quality Consideration**
* **Missing Data -** (2007- 2013) and 2024
* **Granularity** - yearly data at the district level.
* **Temporal coverage** - covers a time span of 2014-2023)

1. **Geographical Data**

* Source :
* Dataset Location: <https://docs.google.com/spreadsheets/d/1J1b-UpBYxcxGZuZd1p01K9dwYMO-ZzsE/edit?gid=800997410#gid=800997410>
* **Time Period /Frequency:** One-time data collection.
* **Geographical Coverage:** Data covers all 25 districts in Sri Lanka.
* **Data Points:**
* **Latitude**: Exact latitude coordinates for each district.
* **Longitude**: Exact longitude coordinates for each district.
* **Area (km²)**: Total area of each district.
* **Population**: Total population count for each district.
* **Urbanization (%)**: Percentage of the population living in urban areas.
* **Climate Type**: General climate classification for each district.
* **Major Rivers**: Names of significant rivers within or near the district.
* **Major Cities/Towns**: Key urban centers or towns within the district.
* **Land Use (% Forest)**: Percentage of land covered by forests.
* **Land Use (% Agriculture)**: Percentage of land used for agriculture.
* **Land Use (% Urban)**: Percentage of land used for urban purposes.
* **Population Density (people/km²)**: Number of people per square kilometer.
* **Average Annual Rainfall (mm)**: Average yearly rainfall for the district.
* **Temperature Range (°C)**: Range of temperatures observed in the district.
* **Natural Disasters**: Historical data on natural disasters affecting the district (e.g., floods, landslides).
* **Dengue Incidence Rate (per 100,000)**: Rate of dengue cases per 100,000 people.
* **Mosquito Breeding Sites**: Known mosquito breeding sites within the district.
* **Rainfall Patterns**: Typical rainfall patterns, including seasonal variations.
* **Temperature and Humidity**: Average temperature and humidity levels.
* **Public Health Infrastructure**: Availability and accessibility of healthcare facilities.
* **Water Storage Practices**: Common practices for water storage in households and communities.
* **Population Mobility**: Patterns of population movement within and across districts.
* **Vector Control Measures**: Strategies and practices in place for controlling mosquito populations.
* **Waste Management**: Waste disposal and management practices in the district.
* **Number of Campaigns (last 5 years)**: Number of public health campaigns conducted in the last 5 years.
* **Focus Areas**: Key areas of focus for public health interventions (e.g., sanitation, vector control).
* **Distance to Nearest Hospital/Clinic (km)**: Proximity to the nearest healthcare facility.
* **Healthcare Facility Density (facilities per 100,000 people)**: Number of healthcare facilities available per 100,000 people.
* **Data Quality Consideration**
* **Missing Data -** all required geographical data is available for each district.
* **Granularity** - The dataset provides a one-time collection of data, capturing key geographical, demographic, and public health-related information at the district level.
* **Temporal coverage** - While the geographical data is static (one-time collection), it provides a comprehensive snapshot necessary for contextual analysis in combination with dynamic weather and disease incidence data.