EBAC/Semantix Project

Arboviruses Recife 2024

Data Collection

Data Source:

The chosen data sources are public, structured datasets regarding cases of Dengue, Zika Virus, and Chikungunya in the year 2024, available on the Recife city hall's website:

http://dados.recife.pe.gov.br/ro/dataset/casos-de-dengue-zika-e-chikungunya

Datasets

Dengue

- Link: http://dados.recife.pe.gov.br/.../download/dengon-2024.csv
- Data type: Structured (CSV)
- Access method: Direct download
- Alternate link: Google Sheets

Zika

- Link: http://dados.recife.pe.gov.br/.../download/zika-2024.csv
- Data type: Structured (CSV)
- Access method: Direct download
- Alternate link: Google Sheets

Chikungunya

- Link: http://dados.recife.pe.gov.br/.../download/chikon-2024.csv

- Data type: Structured (CSV)

- Access method: Direct download

- Alternate link: Google Sheets

Objectives

This project aims to conduct an Exploratory Data Analysis (EDA) on arbovirus cases reported in Recife in 2024 with the following goals:

- Analyze the distribution by gender, age group, neighborhood, and month

- Identify seasonal and demographic patterns

- Compare data across the three diseases

Technologies Used

- Language: Python

- Environment: Google Colab

- Libraries: pandas, numpy, rapidfuzz, unidecode

- Visualization Tool: LookerStudio

Data Modeling

Data Cleaning and Preprocessing:

- Removal of null records or those with special characters

- Conversion of data to numeric format or removal of inconsistent entries

- Standardization of dates and column names

- Extraction of text to create a "month" column

- Normalization of textual categories

- Creation of sorting columns

- Adjustments in encoding and delimiters

- Use of rapidfuzz for approximate string matching

Identification of Key Variables and Correlations:

- Descriptive analysis

- Analysis of relationships between demographics and disease type

- Case distribution by gender, age group, disease type, neighborhood or street, and notification month

- Count of confirmed cases

- Month-by-month seasonality

- Comparison between diseases based on age, gender, and location

- Calculation of statistical age correlations

- Identification of spatial and temporal patterns

Insights Report

1. Cases by Month:

- March had the highest number of notifications, followed by April and January

- Peak in late summer/early autumn due to favorable climate for Aedes aegypti

2. Cases by Age Group:

- Age groups 20-29, 30-39, and 40-49 had the highest case numbers

- Fewer notifications among seniors (60+) and infants under 1 year

3. Cases by Gender:

- Most cases occurred in females

4. Neighborhoods with Most Cases:

- Ibura, Cohab, and Várzea were the most affected neighborhoods

5. Age x Death Correlation:

- Positive correlation between age and death

- Age group 40-49 had the most deaths related to the diseases

Conclusions

- Greater impact on young adults and women

- Age is a risk factor for death

- Dengue leads in both cases and deaths

- Most affected neighborhoods: Ibura, Cohab, Várzea, Nova Descoberta, Boa Viagem, and Iputinga (>500 cases)

- Ibura requires urgent attention with nearly 1,000 cases in 2024

- Most affected street: Avenida Vereador Otacílio with 127 confirmed cases

- Social action and municipal investigation are necessary in this area