CPE 031

"Analyzing Disease Patterns in the Philippines: Insights from the Integrated Disease Surveillance and Response Dataset (2008–2022)"

Submitted by:
Mamaril, Justin Kenneth
Reyes, Alexzander
San Jose, Alexander
San Juan, Edson Ray
Titong, Lee Ivan

Bachelor of Science in Computer Engineering
2nd year, CPE21S4

Submitted to:

ENGR. MARIA RIZETTE H. SAYO

Table of Contents:

Keywords

I. Introduction

Nowadays there are a lot of health issues in the Philippines because the country is struggling with various diseases that endanger the lives of people. This research will major on descriptive analysis of disease trends using the quantitative data from the Integrated Disease Surveillance and Response database which spans from 2008 to 2022. In this period, the disease reporting rates have had an uneven trend, which can be influenced by environmental conditions and population health policies among other factors. The numbers of disease cases reported have recently escalated and the total count for many diseases has crossed dangerous figures. Accurate knowledge of the extent and geographic distribution of these diseases is important in public health situations and interventions. Hence, this paper will seek to shed light on the total reported/total combined of different diseases and year trends. In order to reach this goal, the study will look at the total cases at a time with disease type and year grouped together. Moreover, it will investigate the distribution of disease cases to regions in order to understand the consolidated disease rate in the Philippines. These patterns will be presented using bar charts, line graphs, and heatmaps which are good visualization tools. By examining these features, the study aims to help in disease epidemiology and response, and public health program and policy formulation in the Philippines.

1.1 Significance

The findings obtained from this work would be invaluable to numerous peoples, especially as regards the evaluation of the general well being of the Philippine population. As the nation continues to confront numerous health challenges exacerbated by disease outbreaks, the conclusions drawn from this study may benefit the following groups:

Healthcare Providers: Knowledge derived from the research might help healthcare professionals make informed decisions about the diseases that exist in certain populations. This knowledge can improve clinical decision making, make health related service delivery to be more personalized and thus improve patients' wellbeing via early intercession.

Government Decision-Makers: The results of the analysis may be useful for the lawmakers passing the legislation meant to fight diseases and disease control. When the authorities know how diseases spread in geographically analyzed regions and how the tendencies develop, such information will be helpful in the improvement of public health and resource allocation.

Researchers and Academics: Consequently, the study is an extensive source of data for scientists and academicians who study epidemiology or work in the field of public health. The findings can be used in the subsequent research and help enhance overall knowledge of disease distribution in the Philippines.

The General Public: For citizens, the findings may help them gain more comprehensible knowledge of the extent of prevalence of different diseases in the society, hence enabling them to take preventive measures as far as their health is concerned. Increased awareness will thus play a key role in promoting human participation in health and increasing support for public health measures.

1.2 Research Question

The study aimed to analyze disease patterns in the Philippines based on the Integrated Disease Surveillance and Response Dataset (2008–2022). Understanding these patterns may inform public health strategies, improve disease prevention efforts, and guide resource allocation. Specifically, this study intended to answer the following questions:
   1.   What are the diseases with the highest total reported cases in the Philippines?
   2.   What are the trends in total disease cases over time?

3. How do disease cases vary across different regions of the Philippines?
4. What correlations exist between the occurrence of specific diseases and their distribution over time?

II. Data Dictionary

Dataset Overview:

- Number of Rows: 74,880
- Number of Columns: 9
- Time Period: 2008-2022
- Geographic Scope: 879 municipalities in 12 cities in the Philippines (Dagupan, Palayan, Navotas, Mandaluyong, Muntinlupa, Legazpi, Iloilo, Mandaue, Tacloban, Zamboanga, Cagayan de Oro, and Davao)

Variable Name	Description	Data Type
uuid	A unique identifier for each record	String
freq	Indicates the frequency of reporting (e.g., daily, weekly)	String
date	Date of observation	Date
source_name	Name of the data source	String
source_filename	Filename of the original data source	String
disease_common_name	Common name of the disease	String
case_total	Total number of reported disease cases	Integer
adm3_pcode	Administrative code for the region	String
year	Year extracted from the date	Integer
disease_icd10_code	The ICD-10 code corresponds to the disease.	String

Figure 2.1 Variable Description

The data work inclusive of 9 columns and 6,491,096 rows of values with each row analyzed as the daily reported cases of several diseases over the 15 years between 2008 and 2022. The gathering of intel includes data from different Municipalities all across the Philippines with focus on diseases that occurred in certain cities. The data is collected for each of the municipalities by year and by date from the year 2008 and up to 2022.

Variable description is also depicted in figure 2.1. These variables are designed to bring time series cross-section analysis to enhance disease pattern studies across the regions of interest for a specific time frame. The UUID (uuid) provides each record with its unique identification, (freq) gives the reporting frequency and (date) gives the date the report was made. Administrative areas are described by the three-digit post-code (adm3_pcode), and diseases are coded by the ICD-10 code (disease_icd10_code) and the most frequently used name (disease_common_name). The source and origin of the data are recorded through (source_name) and (source_filename). Last, the amount of reported cases (case_total) gives quantitative expression about the disease's existence. Due to the fact that this structure allows for temporal, spatial and disease specific analyses, it corresponds with the goal of the study, which is to explore trends and factors contributing to the occurrence of disease.

Description of variables is provided in the following tabulation; To address the research questions, the study categorized data into two classifications: information relating specifically to the given disease and geographical characteristics. The outcomes presented for each disease are the total cases and deaths, while the geography data is the administrative code, in fact, the municipality to which the observation belongs.

III. Analysis Process

In this case, while analyzing the disease patterns in the Philippines by using Integrated Disease Surveillance and Response Dataset (2008–2022), The researchers adopted a methodical rubric or, a set of statistical and graphical approaches towards the dataset in order to respond to the research questions. The main objective was to find patterns of disease occurrences, differences between regions and their deaths and disease cycles. Below is an outline of the process: In analyzing the disease patterns in the Philippines using the Integrated Disease Surveillance and Response Dataset (2008–2022), The researchers employed a methodical approach that combines statistical and graphical techniques to explore the dataset and answer the research questions. The overall goal was to identify trends in disease cases, variations across regions, and correlations between disease occurrences over time. Below is an outline of the process:

1. **Initial Data Exploration**

The first data analysis technique involved was data inspection where an attempt was made to get a general overview of the data at hand. This involved looking at the structure of the data and as a result seeking to understand the variables under consideration as well as conducting checks to see whether there was missing data and if the necessary data types were used. There were 74'880 rows and 9 columns with elements including disease names, case totals, reporting frequencies, date and geographical code. An initial exploration involved:

- Descriptive: such as aliasing – checking for any null values or duplicates, and handling them through appropriate techniques including fill, drop in or correcting errors in data.
- To get a basic idea of the dataset bias, using the Pandas package, The researchers will perform initial data analysis and manipulation with discrete columns mean, median, mode, and standard deviation of numerical columns, as well as a frequency table of categorical variables: disease types and regions.
- Scans the band for missing or null values or doesn't contain any data necessary for analysis by the upper bands.
- Mentioning the correct data types especially changing the data type of the "date" column to a datetime type for a time series analysis.

- Descriptive analysis like Means, Median, Standard Deviation of case totals to check if there are any overwhelming cases in the set.
- Exploring the overall shape of the data by looking at the basic details of the dataset which can be obtained by heads(), info(), and describe() functions of pandas.

Tools used: For data manipulation and cleaning the pandas will be used.

2. Data Aggregation

- Grouping: It is arranged by year, disease, and region to look for shifts.
- Summing: Categorizing the reported number of new occurrences for each disease in a specific region.

Data Transformation: Using the unstack() method to alter the data from long format to utilize for the analysis of the disease cases by year and region.

3. Statistical Technique

All the methods used in this analysis are focused on describing the dataset and identifying most of the patterns that would address the research questions. In particular, we focus on the following:

- Descriptive Statistics: Descriptive analysis involves use of measures of central tendency including the mean, median, mode, measures of variability including the standard deviations and range in the disease data. This makes it possible for us to determine which diseases have more or few cases reported on average and their trend analysis.
- Trend Analysis: An approach which utilizes yearly reported diseases is used to evaluate disease cases in the long-run. This entails evaluation of the trends in occurrence of diseases in a society; whether the trend is moving up, down or up and down constantly. For example, we may determine whether given diseases have a general increase in the reported incidence, the entire period from 2008 to 2022.
- Correlation Analysis: The total reported cases of different diseases can be correlated with each other by getting correlation coefficients across time. This kind of analysis is useful in answering the question that addresses the relationship between the probability of occurrence of particular diseases and periodicity of the occurrence or their distribution.

These statistical measures give an overview of the diseases in general and how diseases co-relationship for the different years.

4. Graphical Technique

Visualizing the data is a crucial step in making the findings easily interpretable and conveying patterns that might not be obvious from raw numbers alone. The following graphical techniques are employed in this analysis:

- **Line Graphs** were used to explore the trends of total disease cases over time. By creating a line graph, we were able to visualize the year-on-year changes in disease cases, which helped in detecting any increasing, decreasing, or cyclical trends over the 15-year period. This technique provided a clear view of how the disease patterns evolved and allowed us to identify key periods when certain diseases surged or declined. It helped answer the question, "What are the trends in total disease cases over time?"
- **Bar Charts** were used for the comparison of the sum of reported diseases among different diseases or areas. For the representation of the frequency of diseases, bar charts seem most

appropriate especially indicated for differentiation by year and geographical area. What was of great importance is that it allowed us to identify diseases that were more frequent in certain years or geographical areas. Through these bar charts, necessary patterns of diseases' spread and changes in diseases' incidence, have been shown so the answer to the question "What are the top ten diseases with the highest total reported cases in the Philippines?"

- **Heatmaps** were used to identify relationships between disease rates and the spatial distribution of cases through time. In this case, heat maps were created in order to visualize the correlation between certain diseases and the years that they were identified. In the heatmap each cell gave the total number of cases of a particular disease in a particular year thus capturing the frequencies and hence making it easier for the analysis to reveal patterns. It was especially effective to apply this technique in response to the question, Which diseases are related to each other and how does the distribution of the diseases vary over time?
- **Box Plots** were applied to assess distinctive medical and disease case rates of different territories. When it comes to showing the spread as well as variation of the disease cases in different regions then box plot is especially helpful. This is evident in the range of cases, the median and the range of diseased cases helping one to figure out how cases were spread and whether certain regions of the world had more or less cases of disease. This technique answered a question: how do disease cases differ in different regions of the Philippines?

IV. Analysis and Insights

Data Trends over Time:

- The structures of the dataset revealed that diseases reported by health care facilities vary from year to year, some diseases showing a perfect trend in the increase in the number of cases from the year to the other. For instance, there are some illnesses that had fluctuating incidence rates within certain years, an aspect that derived from the outbreak trends in the country; dengue and ARI were some of the diseases in question.
- Seasonal trends showed that the incidence of some diseases, especially those reported to be associated with the vectors and other periodical diseases, by definition, reported increased incidence during specific seasons such as the monsoon season.

Regional Distribution:

- There was regional variation to disease incidence. Mandaluyong and Muntinlupa those areas with high number of cases for diseases like ARI but Davao and Zamboanga provinces with poor health facilities have many cases of diseases like dengue and malaria. This finding led to an implication that factors such as area and proximity to health facilities play a major role in projecting disease incidences.

Data Quality and Gaps:

- While comparing the data some inconsistencies were pinpointed for example the values missing in disease reporting or the frequency reporting was erratic. This called for more data cleaning and data standardization in order to enhance the validity of the analysis.
- It needs to be pointed out that the frequency of the reported data was not the same for all municipalities, with some daily reporting and others reporting less frequently. This variability could bring in some biases in disease patterns' interpretation.

New Question for Future Exploration

The analysis raised additional questions that could be explored further:

1. To what extent has the Philippines' health system been successful in trying to prevent or control new diseases, and what should be done to adjust for early disease detection or availability of resources, and finally how effective is treatment as we note early treatments during these forms of diseases?
2. Assessing effectiveness of public awareness campaigns in the Philippines towards the intended modification of health behaviors and /or prevention of disease outbreaks and identifying ways of enhancing the modification of those awareness campaigns to assist in reaching intended populations.
3. Amidst disease outbreaks in the Philippines, how does the psychology of affected communities respond to disease outbreaks, and what can be done to reduce psychological effects on people and families?

4.1 What are the diseases with the highest total reported cases in the Philippines?
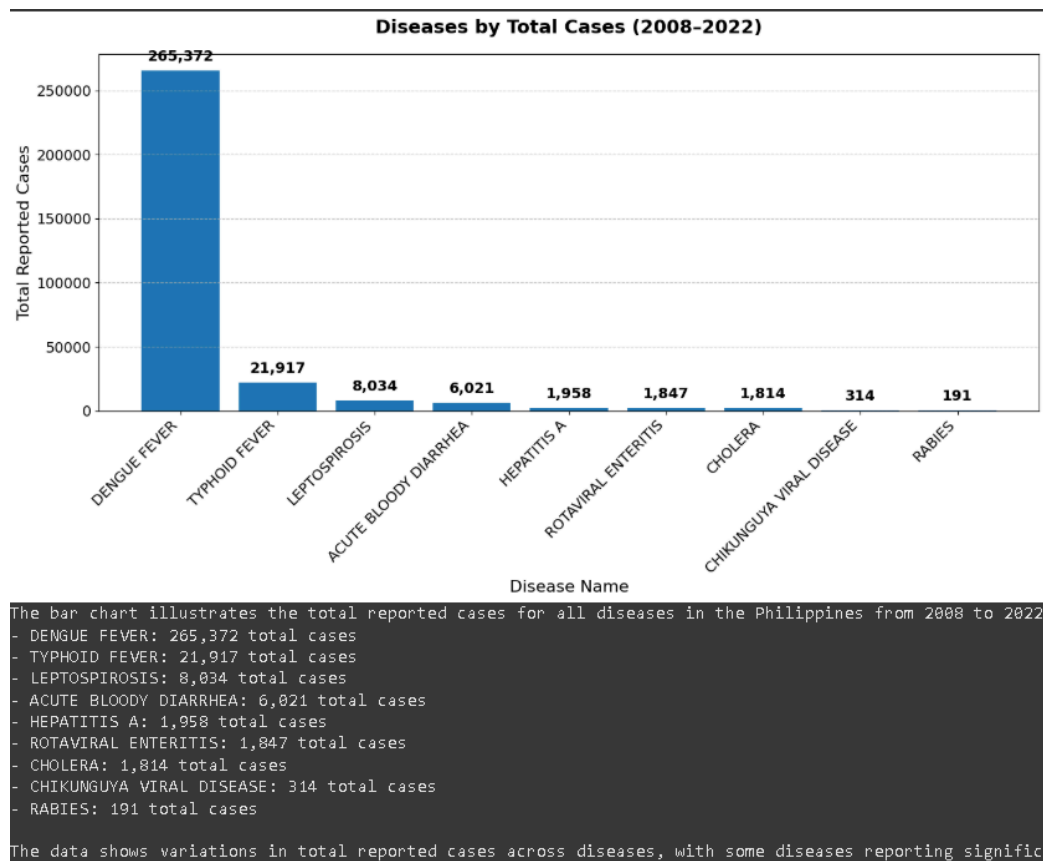

Figure 4.1 Total Reported Disease Cases in the Philippines (2008–2022)

This representation aims to present the key findings on disease burden in the Philippines based on the Integrated Disease Surveillance and Response Dataset for the examined period, 2008–2022. This

provides the ability to monitor the total reported cases by disease and thus determine which diseases have had the most impact on the health of the public over this period.

Total reported cases of different diseases in the Philippines from 2008 to 2022 are represented in the bar chart based on the Integrated Disease Surveillance and Response Dataset. The reported cases are distributed across the following diseases:

- Dengue Fever: 265,372 total cases
- Typhoid Fever: 21,917 total cases
- Leptospirosis: 8,034 total cases
- Acute Bloody Diarrhea: 6,021 total cases
- Hepatitis A: 1,958 total cases
- Rotaviral Enteritis: 1,847 total cases
- Cholera: 1,814 total cases
- Chikungunya Viral Disease: 314 total cases
- Rabies: 191 total cases

Analysis of the results in terms of total reported cases indicates consistently high fluctuations and, in particular, high levels of incidence in relation to Dengue Fever. These opposing trends underscore the potential and the challenges of focusing public health and resource mobilization efforts on the disease burden.

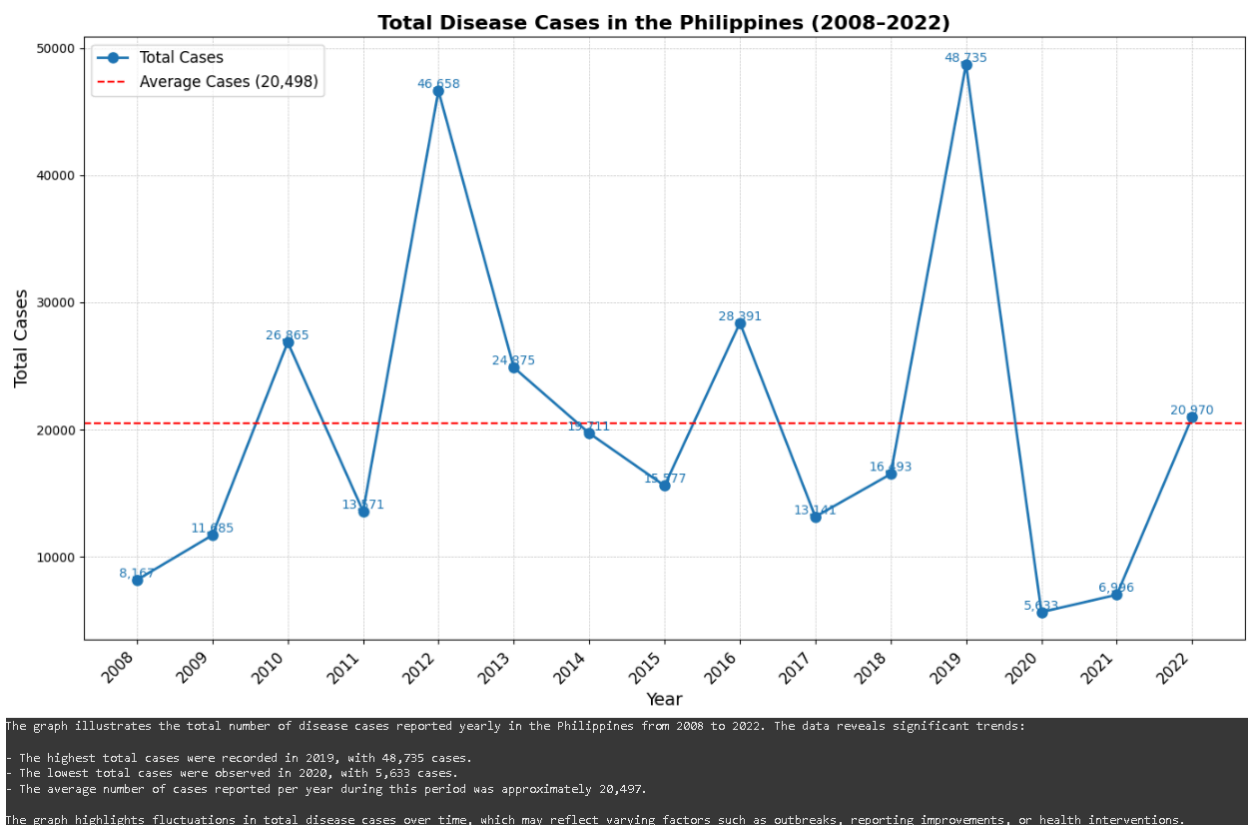**4.2 What are the trends in total disease cases over time?**

Figure 4.2 Trends in Total Disease Cases in the Philippines (2008–2022)

The line graph illustrates the total number of disease cases reported yearly in the Philippines from 2008 to 2022. The data reveals significant trends:

- Highest Yearly Cases: The year with the highest total reported cases was [2019], with [48,735] cases.
- Lowest Yearly Cases: Conversely, the year with the lowest reported cases was [2020], recording [5,633] cases.
- Average Cases: Over the entire period, the average number of cases reported per year was approximately [20,497].

Total disease cases: The graph shows changes in the overall state of the epidemic at any given time to indicate additional factors that may exist such as; Outbreaks, change in reporting rate, seasonality, and impact of health interventions.

4.3 How do disease cases vary across different regions of the Philippines?
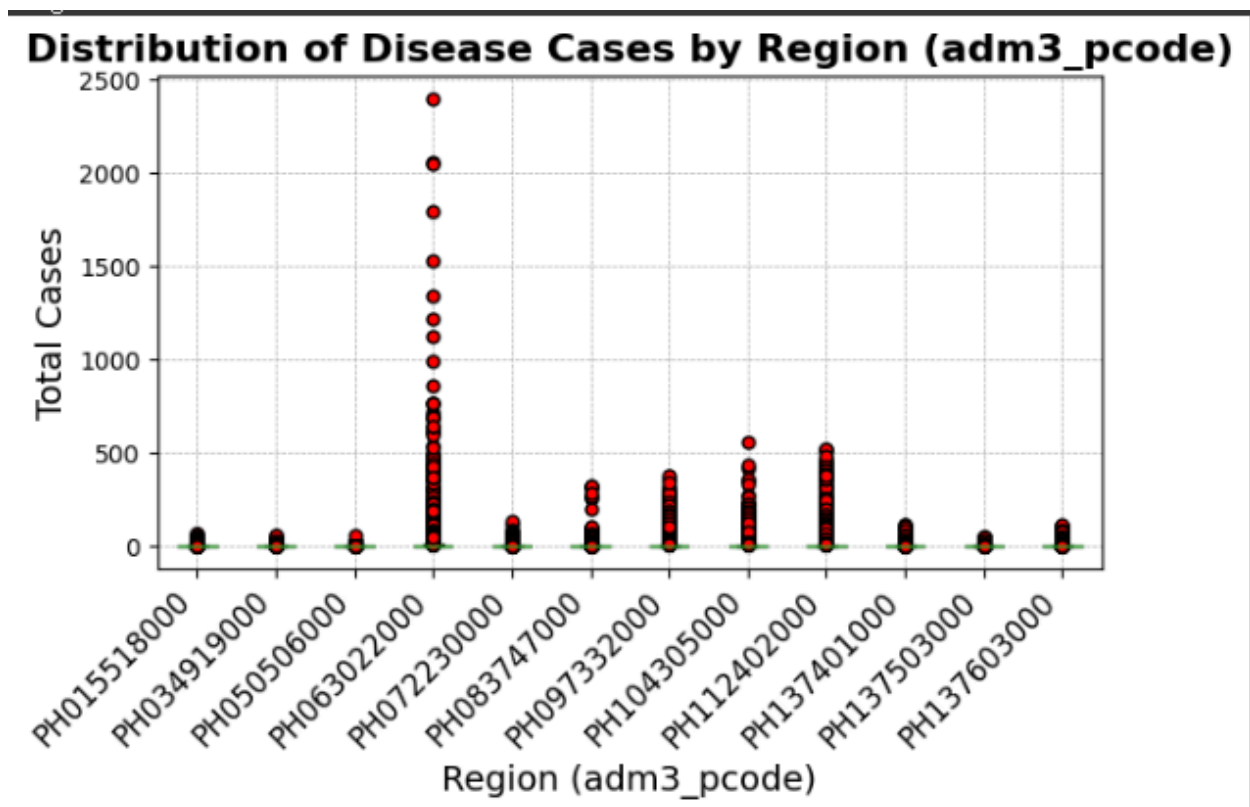
Figure 4.3 Distribution of Disease Cases by Region in the Philippines (2008–2022)

The following boxplot demonstrates the disease cases distribution by their adm3_pcode in the eight regions of Philippines from the year 2008 to 2022. This analysis reveals significant variations in the total number of reported cases, highlighting key patterns and insights:

This table links each adm3_pcode with its respective city in the Philippines, providing a clearer view for analyzing disease patterns in different locations.

adm3_pcode	City
PH015518000	Dagupan
PH034919000	Palayan
PH050506000	Navotas
PH063022000	Mandaluyong
PH072230000	Muntinlupa
PH083747000	Legazpi
PH097332000	Iloilo
PH104305000	Mandaue
PH112402000	Tacloban
PH137401000	Zamboanga
PH137503000	Cagayan de Oro
PH137603000	Davao

Regional Disparities in Total Cases:

Both total disease cases and cumulative daily gain have some inequalities within the regions. For example, Region PH063022000/Mandaluyong described the highest total case count with 83,173, while other regions such as PH034919000/Palayan with only 1,913 cases and PH050506000/Navotas with 2,344 cases.

Median and Mean Cases:

A high proportion of regions present a median of 0 case count meaning many regions did not register cases in some years. This implies that the disease may be only reported once in a while and may not be represented in all the countries.

The average of the case numbers significantly differ, the higher, average being that of Region PH063022000/Mandaluyong with a case average of thirteen point three three, which shows frequent occurrences of cases in the recent years; unlike Region PH137503000/Cagayan de Oro, which pegs its average at one point two eight.

Max and Min Cases:

Other states like PH063022000/Mandaluyong had the maximum cases of up to 2398 in some of the years indicating that outbreaks are possible. In contrast, several regions present a maximum of about 55 cases and denote low rates of occurrence.

The fluctuation in maximum reported cases especially re-emphasizes a need to have standard approaches to enhance the handling and prevention of disease outbreaks in the areas with high reported cases.

Outliers:

Mean deviation from the norm can be seen from the boxplot in which red dots signify occasional eruptions in the areas and necessitate further detailed study on the probable causes of such outbreaks.

Region	Total Cases	Median Cases	Min Cases	Max Cases	Mean Cases
PH063022000/Mandaluyong	83,173	0	0	2,398	13.33
PH112402000/Tacloban	71,699	0	0	522	11.49
PH097332000/Iloilo	48,600	0	0	383	7.79
PH104305000/Mandaue	38,448	0	0	554	6.16
PH083747000/Legazpi	12,135	0	0	324	1.94
PH072230000/Muntinlupa	11,663	0	0	129	1.87
PH015518000/Dagupan	9,514	0	0	70	1.52
PH137401000/Zamboanga	8,704	0	0	114	1.39
PH137503000/Cagayan de Oro	7,964	0	0	52	1.28
PH137603000/Davao	11,311	0	0	116	1.81
PH050506000/Navotas	2,344	0	0	55	0.38
PH034919000/Palayan	1,913	0	0	57	0.31

Public health interventions strategies should for this reason not assume that the regional diseases incidence is uniform for them to be effective. The speaking for the measure of total cases shows maximum diseases in specific regions that may need an increased health monitoring, precautionary measures and health care interventional measures to reduce and prevent the burden of diseases. It is therefore fundamental to demographically identify these regions and factors affecting its health with the aim of enhancing the health resource allocation as well as the planning of disease control mechanisms.

4.4 What correlations exist between the occurrence of specific diseases and their distribution over time?

Disease Cases Correlation by Year and Disease

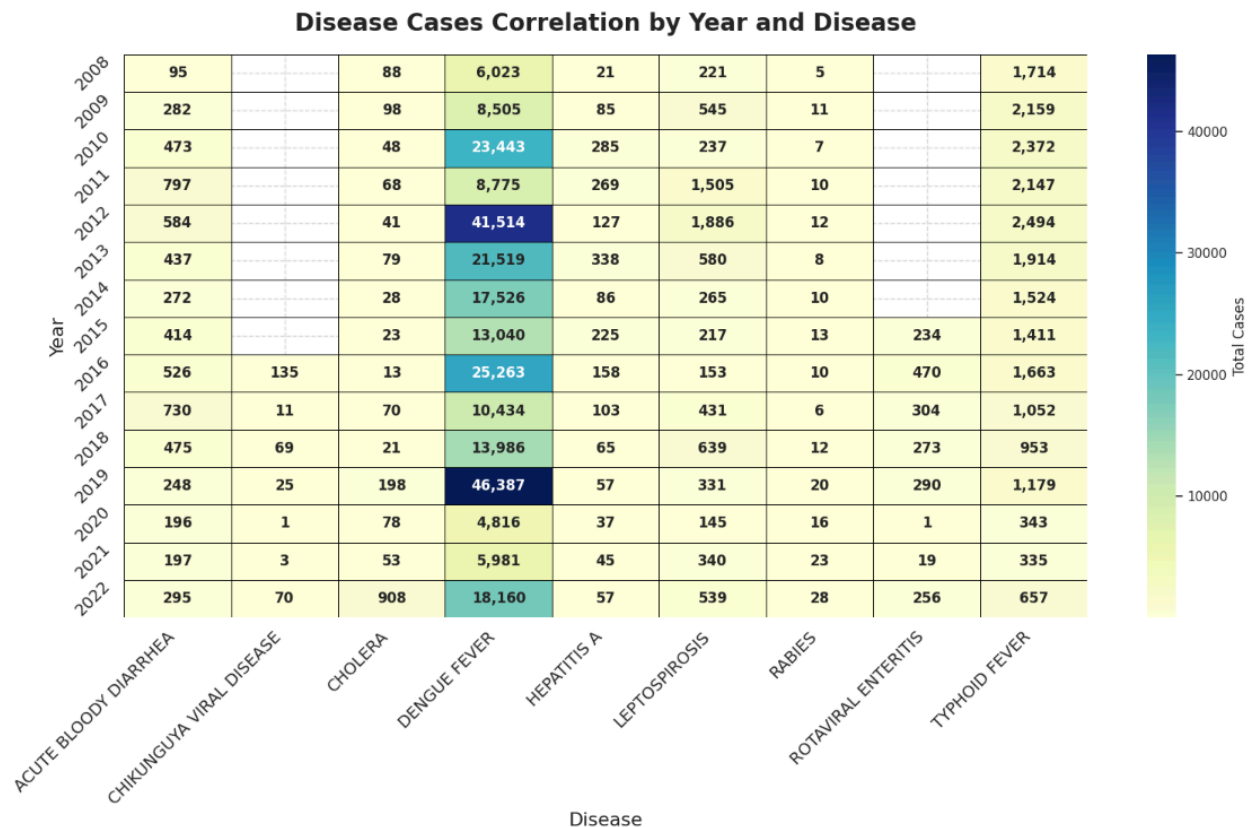| Year	ACUTE BLOODY DIARRHEA	CHIKUNGUYA VIRAL DISEASE	CHOLERA	DENGUE FEVER	HEPATITIS A	LEPTOSPIROSIS	RABIES	ROTAVIRAL ENTERITIS	TYPHOID FEVER
2008	95		88	6,023	21	221	5		1,714
2009	282		98	8,505	85	545	11		2,159
2010	473		48	23,443	285	237	7		2,372
2011	797		68	8,775	269	1,505	10		2,147
2012	584		41	41,514	127	1,886	12		2,494
2013	437		79	21,519	338	580	8		1,914
2014	272		28	17,526	86	265	10		1,524
2015	414		23	13,040	225	217	13	234	1,411
2016	526	135	13	25,263	158	153	10	470	1,663
2017	730	11	70	10,434	103	431	6	304	1,052
2018	475	69	21	13,986	65	639	12	273	953
2019	248	25	198	46,387	57	331	20	290	1,179
2020	196	1	78	4,816	37	145	16	1	343
2021	197	3	53	5,981	45	340	23	19	335
2022	295	70	908	18,160	57	539	28	256	657

Figure 4.4 Correlation of Disease Cases by Year and Disease Type in the Philippines (2008–2022)

The heatmap presented above represents the various disease types and their reported cases in the Philippines per year for the period ranging between 2008 and 2022. The figure given below is a heatmap where every cell in it holds the value of the total number of cases reported for a given disease at a given year.

Key Findings:

Dengue Fever:
Maximum Cases: The number of reported cases reached 46,387 in 2019.
Minimum Cases: 4,816 reported cases in 2020.
Total Cases: Periodically, Dengue Fever has always revealed a high rate, which has ranked it among the most common diseases.

Typhoid Fever:
Maximum Cases: 2,494 in 2012.
Minimum Cases: 335 in 2021.
Typhoid Fever also is observed to vary in its occurrence, it recorded its high in 2012.

Leptospirosis:
Maximum Cases: 1,886 in 2012.

Minimum Cases: 145 in 2020.
This pattern suggests higher numbers in some years – one of which occurred during the period of this study – and in the period immediately following natural disasters.

Cholera:
Maximum Cases: 908 reported cases in 2022.
Minimum Cases: 13 in 2016.
With regards to infectious diseases, it was found that cholera is on the rise in recent years, and needs more consideration.

Hepatitis A and Others:
Similarly, other diseases such as, Hepatitis A, Chikungunya, Rabies, Rotaviral Enteritis had fewer total incidence than Dengue and Typhoid, nevertheless, they also revealed fluctuations over the period.

Summary Table of Disease Cases

A summary table detailing the maximum and minimum reported cases for each disease over the years has been compiled:

Disease	Max Total Cases	Year with Max Cases	Min Total Cases	Year with Min Cases	Total Cases Across All Years
Acute Bloody Diarrhea	797	2011	95	2008	6021
Chikungunya Viral Disease	135	2016	1	2020	314
Cholera	908	2022	13	2016	1814
Dengue Fever	46, 387	2019	4,816	2020	265,372
Hepatitis A	338	2013	21	2008	1958
Leptospirosis	1,886	2012	145	2020	8034
Rabies	28	2022	5	2008	191
Rotaviral Enteritis	470	2016	1	2020	1,847
Typhoid Fever	2,494	2012	335	2021	21,917

The findings of this study show that some diseases have increased incidence in given years, which may point to outbreaks or environment causing increased incidences in the diseases in question. Hence, constant surveillance and context specific control measures are necessary in preventing and controlling these diseases.

V. Conclusion and Recommendations

5.1 Conclusion

Through observing the data gathered from the Integrated Disease Surveillance and Response dataset (2008–2022), this paper identified key trends and patterns concerning multiple diseases in the Philippines. Dengue fever topped the list of diseases with the highest total reported cases adding up throughout the years with some increase on certain years, whereas diseases like typhoid fever and leptospirosis kicked up during some years relating to environmental or seasonality. There were variations in disease frequency based on geographical areas, reflecting higher incidence rates of some diseases in urban city Mandaluyong, and high incidences of vector borne diseases in rural areas. For example, the use of line graphs as well as heat maps brought out features that could not be easily seen from the values and figures. However, as much as CNS has developed such insights based on the analysed data, it equally has to acknowledge challenges such as; lack of data, data being reported at irregular intervals, and or missing data, which points to the need to improve the quality of data reported. This calls for more focused public health interventions, better surveillance systems for diseases' interactions with time and space.

5.2 Recommendations

Targeted Health Interventions
- The focus place and clients priority should be set among regions with significant disease incidence and prevalence, including urban provinces such as Mandaluyong and underserved rural areas.
- Launch health promotion activities for diseases such as dengue and leptospirosis with programs conducted prior to Monsoon seasons.

Policy and Resource Allocation
- Campaign for greater allocation of state funds toward the development of health facilities with emphasis on areas where there is poor access to such services.
- Introduce into decision-making the data on diseases and areas with high disease incidence so that sources of financing can be allocated more effectively.

Invest in Research and Capacity Building
- To identify effective measures for preventing diseases, carry out more research of the disease incidence distribution related to socio-economic and environmental characteristics.
- Capability built at local level to increase the number of local trained health workers in issues of data management and in outbreak responses.

VI. REFERENCES:

- **W3Schools. (n.d.).** *Matplotlib Pyplot.* **W3Schools. Retrieved from**https://www.w3schools.com/python/matplotlib_pyplot.asp

- **GeeksforGeeks. (2020, May 31). Introduction to Seaborn - Python. Retrieved** https://www.geeksforgeeks.org/introduction-to-seaborn-python/

- **GeeksforGeeks. (2021, July 22). Data visualization using Matplotlib** [Data Visualization using Matplotlib in Python - GeeksforGeeks](#)

VII. Annex

7.1 Program Code

```Python
!pip install pandas
!pip install numpy
!pip install matplotlib
```

```Python
import pandas as pd

file_path = "/content/disease_pidsr_totals.csv"
df = pd.read_csv(file_path)
df.head()
```

Question 1 Code:

```Python
import pandas as pd
import matplotlib.pyplot as plt

file_path = 'disease_pidsr_totals.csv'
data = pd.read_csv(file_path)

total_cases_by_disease =
data.groupby('disease_common_name')['case_total'].sum().reset_index()
total_cases_by_disease = total_cases_by_disease.sort_values('case_total',
ascending=False)

# Plotting all diseases by total cases
plt.figure(figsize=(14, 8))
bars = plt.bar(total_cases_by_disease['disease_common_name'],
total_cases_by_disease['case_total'], color='#1f77b4')
plt.title('Diseases by Total Cases (2008–2022)', fontsize=18,
fontweight='bold', pad=20)
plt.xlabel('Disease Name', fontsize=16, fontweight='regular')
plt.ylabel('Total Reported Cases', fontsize=16, fontweight='regular')
plt.xticks(rotation=45, ha='right', fontsize=14)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.yticks(fontsize=14)

for bar in bars:
```

```
    yval = bar.get_height()
    plt.text(bar.get_x() + bar.get_width() / 2, yval + 0.02 *
max(total_cases_by_disease['case_total']),
             f'{int(yval):,}', ha='center', va='bottom', fontsize=14,
fontweight='bold')

plt.tight_layout()
plt.show()

total_cases_by_disease['case_total'] =
total_cases_by_disease['case_total'].apply(lambda x: f"{int(x):,}")
explanation = (
    f"The bar chart illustrates the total reported cases for all diseases in
the Philippines from 2008 to 2022, "
    f"based on the dataset. The reported cases are distributed across the
following diseases:\n"
)

for index, row in total_cases_by_disease.iterrows():
    explanation += f"- {row['disease_common_name']}: {row['case_total']} total
cases\n"

explanation += (
    "\nThe data shows variations in total reported cases across diseases, with
some diseases reporting significantly "
    "higher cases than others. This variation underscores the need for targeted
interventions to address the diseases "
    "with the highest burden effectively."
)

print(explanation)
```

Question 2 Code:

```python
data['date'] = pd.to_datetime(data['date'])
data['year'] = data['date'].dt.year

yearly_cases = data.groupby('year')['case_total'].sum().reset_index()

plt.figure(figsize=(14, 8))
plt.plot(
    yearly_cases['year'],
    yearly_cases['case_total'],
    marker='o',
    linestyle='-',
    color='tab:blue',
    linewidth=2,
    markersize=8,
    label='Total Cases'
)

plt.title('Total Disease Cases in the Philippines (2008-2022)', fontsize=16,
fontweight='bold')
plt.xlabel('Year', fontsize=14)
plt.ylabel('Total Cases', fontsize=14)
plt.xticks(yearly_cases['year'], rotation=45, ha='right', fontsize=12)
plt.grid(True, linestyle='--', linewidth=0.5, alpha=0.7)

for i, year in enumerate(yearly_cases['year']):
    plt.text(
        yearly_cases['year'][i],
        yearly_cases['case_total'][i],
        f'{yearly_cases["case_total"][i]:,}',
        ha='center',
        va='bottom',
        fontsize=10,
        color='tab:blue'
    )

average_cases = yearly_cases['case_total'].mean()
plt.axhline(average_cases, color='red', linestyle='--', label=f'Average Cases
({average_cases:,.0f})')
plt.legend(loc='upper left', fontsize=12)
plt.tight_layout()
plt.show()
```

```python
data['date'] = pd.to_datetime(data['date'])
data['year'] = data['date'].dt.year
yearly_cases = data.groupby('year')['case_total'].sum().reset_index()

if 'year' in yearly_cases.columns and 'case_total' in yearly_cases.columns:
    max_year = yearly_cases.loc[yearly_cases['case_total'].idxmax()]
    min_year = yearly_cases.loc[yearly_cases['case_total'].idxmin()]
    average_cases = yearly_cases['case_total'].mean()

    explanation = (
        f"The graph illustrates the total number of disease cases reported
yearly in the Philippines "
        f"from 2008 to 2022. The data reveals significant trends:\n\n"
        f"- The highest total cases were recorded in {int(max_year['year'])},
with {int(max_year['case_total']):,} cases.\n"
        f"- The lowest total cases were observed in {int(min_year['year'])},
with {int(min_year['case_total']):,} cases.\n"
        f"- The average number of cases reported per year during this period
was approximately {int(average_cases):,}.\n\n"
        "The graph highlights fluctuations in total disease cases over time,
which may reflect varying factors such as "
        "outbreaks, reporting improvements, or \nhealth interventions.\n"
    )
    print(explanation)
else:
    print("Error: 'yearly_cases' does not have the required columns: 'year' and
'case_total'.")
```

Question 3 Code:

```python
plt.figure(figsize=(16, 9))

data.boxplot(
    column='case_total',
    by='adm3_pcode',
    grid=True,
    patch_artist=True,
    boxprops=dict(facecolor='lightblue', color='black'),
    whiskerprops=dict(color='black'),
    capprops=dict(color='black'),
    flierprops=dict(markerfacecolor='red', marker='o', markersize=5),
```

```python
)

plt.title('Distribution of Disease Cases by Region (adm3_pcode)', fontsize=16,
fontweight='bold')
plt.suptitle('')
plt.xlabel('Region (adm3_pcode)', fontsize=14)
plt.ylabel('Total Cases', fontsize=14)
plt.xticks(rotation=45, ha='right', fontsize=12)
plt.grid(True, linestyle='--', linewidth=0.5)
plt.tight_layout()
plt.show()

output_file = "region_summary.csv"
region_summary.to_csv(output_file, index=True)
print(f"Summary statistics saved to {output_file}")

summary_text = "Summary of Disease Case Variations Across Regions:\n\n"
for region, stats in region_summary.iterrows():
    summary_text += (
        f"Region: {region}\n"
        f" - Total Cases: {int(data[data['adm3_pcode'] ==
region]['case_total'].sum()):,}\n"
        f" - Median Cases: {int(stats['50%']):,}\n"
        f" - Min Cases: {int(stats['min']):,}\n"
        f" - Max Cases: {int(stats['max']):,}\n"
        f" - Mean Cases: {stats['mean']:.2f}\n\n"
    )
print(summary_text)
```

Question 4 Code:

```python
import seaborn as sns

data['date'] = pd.to_datetime(data['date'])
data['year'] = data['date'].dt.year
heatmap_data = data.groupby(['year',
'disease_common_name'])['case_total'].sum().unstack()
plt.figure(figsize=(16, 10))
sns.set_theme(style="whitegrid")
heatmap = sns.heatmap(heatmap_data, annot=True, fmt=",.0f", cmap="YlGnBu",
cbar_kws={'label': 'Total Cases'},
```

```python
                        annot_kws={"size": 12, "weight": "bold"}, linewidths=0.5,
linecolor='black',
                        xticklabels=True, yticklabels=True)

plt.title('Disease Cases Correlation by Year and Disease', fontsize=20,
fontweight='bold', pad=20)
plt.xlabel('Disease', fontsize=16, fontweight='regular')
plt.ylabel('Year', fontsize=16, fontweight='regular')
plt.xticks(rotation=45, ha='right', fontsize=14)
plt.yticks(rotation=45, fontsize=14)
plt.grid(True, which='both', axis='both', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

max_case_disease = heatmap_data.max(axis=0)
min_case_disease = heatmap_data.min(axis=0)

max_case_year = heatmap_data.idxmax(axis=0)
min_case_year = heatmap_data.idxmin(axis=0)

summary_table = pd.DataFrame({
    'Disease': heatmap_data.columns,
    'Max Total Cases': max_case_disease.values,
    'Year with Max Cases': max_case_year.values,
    'Min Total Cases': min_case_disease.values,
    'Year with Min Cases': min_case_year.values,
    'Total Cases Across All Years': heatmap_data.sum(axis=0).values
})

print("Full Disease Cases Summary Table:")
print(summary_table)

output_file = 'heatmap_summary_table.csv'
summary_table.to_csv(output_file, index=False)
```

7.2. Console Output

Question 1 Analysis:

```
The bar chart illustrates the total reported cases for all diseases in the Philippines from 2008 to 2022, based on the dataset.
The reported cases are distributed across the following diseases:
- DENGUE FEVER: 265,372 total cases
- TYPHOID FEVER: 21,917 total cases
- LEPTOSPIROSIS: 8,034 total cases
- ACUTE BLOODY DIARRHEA: 6,021 total cases
- HEPATITIS A: 1,958 total cases
- ROTAVIRAL ENTERITIS: 1,847 total cases
- CHOLERA: 1,814 total cases
- CHIKUNGUYA VIRAL DISEASE: 314 total cases
- RABIES: 191 total cases

The data shows variations in total reported cases across diseases, with some diseases reporting significantly higher cases than others.
This variation underscores the need for targeted interventions to address the diseases with the highest burden effectively.
```

Question 2 Analysis:

```
The graph illustrates the total number of disease cases reported yearly in the Philippines from 2008 to 2022. The data reveals significant trends:

- The highest total cases were recorded in 2019, with 48,735 cases.
- The lowest total cases were observed in 2020, with 5,633 cases.
- The average number of cases reported per year during this period was approximately 20,497.

The graph highlights fluctuations in total disease cases over time, which may reflect varying factors such as outbreaks, reporting improvements, or
health interventions.
```

Question 3 Analysis:

```
Summary statistics saved to region_summary.csv
Summary of Disease Case Variations Across Regions:

Region: 0
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 70
 - Mean Cases: 1.52

Region: 1
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 57
 - Mean Cases: 0.31

Region: 2
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 55
 - Mean Cases: 0.38
```

```
Region: 3
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 2,398
 - Mean Cases: 13.33

Region: 4
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 129
 - Mean Cases: 1.87

Region: 5
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 324
 - Mean Cases: 1.94

Region: 6
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 383
 - Mean Cases: 7.79
```

Question 4 Analysis:

```
Region: 7
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 554
 - Mean Cases: 6.16

Region: 8
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 522
 - Mean Cases: 11.49

Region: 9
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 114
 - Mean Cases: 1.39

Region: 10
 - Total Cases: 0
 - Median Cases: 0
 - Min Cases: 0
 - Max Cases: 52
 - Mean Cases: 1.28

 Region: 11
  - Total Cases: 0
  - Median Cases: 0
  - Min Cases: 0
  - Max Cases: 116
  - Mean Cases: 1.81
```

```
Full Disease Cases Summary Table:
                    Disease  Max Total Cases  Year with Max Cases  \
0     ACUTE BLOODY DIARRHEA            797.0                 2011
1  CHIKUNGUYA VIRAL DISEASE            135.0                 2016
2                   CHOLERA            908.0                 2022
3             DENGUE FEVER          46387.0                 2019
4               HEPATITIS A            338.0                 2013
5             LEPTOSPIROSIS           1886.0                 2012
6                    RABIES             28.0                 2022
7       ROTAVIRAL ENTERITIS            470.0                 2016
8             TYPHOID FEVER           2494.0                 2012

   Min Total Cases  Year with Min Cases  Total Cases Across All Years
0             95.0                 2008                        6021.0
1              1.0                 2020                         314.0
2             13.0                 2016                        1814.0
3           4816.0                 2020                      265372.0
4             21.0                 2008                        1958.0
5            145.0                 2020                        8034.0
6              5.0                 2008                         191.0
7              1.0                 2020                        1847.0
8            335.0                 2021                       21917.0
```