**Analysis of Monthly Counts of Deaths by Select Causes from 2014 to 2019**

**Arun Mathew**

**October 24, 2023**

**Introduction**

**Repository** - <https://github.com/ArunMathew7/DAV-5400/tree/main/Project1>

The dataset contains information on mortality in the United States for the years 2014 to 2019, broken down by month. The data includes the number of deaths for various causes, such as natural causes, specific diseases (e.g., septicemia, malignant neoplasms, diabetes mellitus, Alzheimer's disease, etc.), and external factors like accidents, suicides, and drug overdoses. Researchers can use this dataset to analyze mortality trends, identify potential public health issues, and explore the impact of various factors on the causes of death in the United States.

**Data Description:**

1. **Year**: Represents the year in which the data was recorded.
2. **Month**: Indicates the specific month within the year when the data was collected.
3. **All Cause**: The total number of deaths due to all causes in a given year and month.
4. **Natural Cause**: The number of deaths attributed to natural causes, which typically includes climatic conditions or disasters.
5. **Septicemia**: The number of deaths caused by septicemia, a severe bloodstream infection.
6. **Malignant Neoplasms**: Represents deaths caused by malignant neoplasms, which are cancerous tumors.
7. **Diabetes Mellitus**: The number of deaths resulting from diabetes mellitus, a metabolic disorder characterized by high blood sugar levels.
8. **Alzheimer Disease**: Indicates deaths related to Alzheimer's disease, a neurodegenerative condition that affects memory and cognitive function.
9. **Influenza and Pneumonia**: Represents deaths due to influenza (flu) and pneumonia, typically respiratory infections.
10. **Chronic Lower Respiratory Diseases**: The number of deaths resulting from chronic lower respiratory diseases, which include conditions like chronic obstructive pulmonary disease (COPD).
11. **Other Diseases of Respiratory System**: Deaths attributed to other diseases of the respiratory system.
12. **Nephritis, Nephrotic Syndrome, and Nephrosis**: Indicates deaths related to kidney diseases, including nephritis and nephrotic syndrome.
13. **Symptoms, Signs, and Abnormal Clinical and Laboratory Findings, Not Elsewhere Classified**: Represents deaths where the primary cause was not a specific disease but rather abnormal clinical or laboratory findings.
14. **Diseases of Heart**: The number of deaths caused by heart diseases, including conditions like coronary artery disease and heart failure.
15. **Cerebrovascular Diseases**: Indicates deaths due to cerebrovascular diseases, such as strokes.
16. **Accidents (Unintentional Injuries)**: Represents deaths resulting from unintentional accidents.
17. **Motor Vehicle Accidents**: The number of deaths specifically related to motor vehicle accidents.
18. **Intentional Self-Harm (Suicide)**: Indicates deaths due to suicide, where individuals intentionally take their own lives.
19. **Assault (Homicide)**: Represents deaths due to assaults or homicides, where one person intentionally causes the death of another.
20. **Drug Overdose**: Indicates deaths resulting from drug overdoses, often associated with the abuse or misuse of drugs or substances.

**Research Question to answer with this data**

Are there seasonal variations in mortality for specific diseases or causes, and if so, what might explain these variations?

**Steps to solve the research questions**

* **Data Preparation**: Import the dataset and necessary libraries (e.g., pandas, matplotlib, seaborn). Make sure the data types are appropriate for analysis.
* **Data Exploration**: Examine the dataset for any missing values, outliers, or inconsistencies.
* **Seasonal Variation Analysis**: Group the data by disease causes and calculate the mean or sum of deaths for each cause for each month.
* **Visualization**: Create line plots or bar plots for specific diseases or causes over the months to visualize seasonal variations.
* Compare seasonal trends across different diseases.
* Examine external factors that might explain the variations, such as weather patterns (e.g., flu season in colder months), public health measures, or other relevant factors.

**Data Summary**

**Dataset Used** - <https://raw.githubusercontent.com/ArunMathew7/DAV-5400/main/Project1/Monthly_Counts_of_Deaths_by_Select_Causes__2014-2019.csv>

**Source**

This is one of the datasets from US Department of Health and Human services. I took the Monthly Counts of Deaths by Select Causes, 2014-2019. This dataset provides information on mortality by various causes over time, has several valuable use cases, particularly for epidemiologists, public health professionals, researchers, and policymakers.

**Use Cases**

**Epidemiological Studies**: Researchers can use this dataset to conduct epidemiological studies to understand the patterns and trends in mortality over the years. This can help identify areas where public health interventions are needed.

**Public Health Policy Development**: Public health officials and policymakers can use this data to inform the development and implementation of health policies and interventions. For example, identifying causes of death with increasing trends could lead to targeted public health campaigns.

**Disease Surveillance**: This dataset can be part of a broader disease surveillance system. It helps monitor the occurrence and spread of specific diseases or health conditions. This can be especially relevant in tracking infectious diseases.

**Seasonal Variations in Mortality**: The dataset can be used to study seasonal variations in mortality, which may reveal trends in diseases like influenza or heat-related illnesses, helping public health authorities prepare for seasonal variations.

**Comparative Analysis**: Researchers can compare trends across different jurisdictions, years, or causes of death. This can provide insights into variations in healthcare access and quality.

**Impact of Interventions**: It can be used to assess the impact of public health interventions or medical advancements on reducing mortality from specific diseases.

**Data Validation**: The dataset can be used to validate other datasets, especially those related to healthcare and mortality. It can serve as a reference for data quality checks.

**Research and Publications**: Researchers in the field of public health and epidemiology can use this data for research projects, academic publications, and presentations.

**Resource Allocation**: Public health agencies can use the data to allocate resources, such as medical staff and equipment, to areas or times of the year with higher mortality rates.

**Predictive Modelling**: The dataset can be used to develop predictive models for estimating future mortality trends, which can assist in healthcare resource planning.

**Risk Assessment**: Researchers can use the dataset to assess the risk factors associated with various causes of death. This can be useful in identifying areas for preventative measures.

**Analysing dataset**

Reading of dataset, cleaning and interpretation, data types of each attributes will be present in the data\_summary.py and the outputs will be in the all\_result.ipynb both the files are attached in the below URL.

Data\_summary.py - https://github.com/ArunMathew7/DAV-5400/blob/main/Project1/Data\_Summary.py

all\_result.ipynb - https://github.com/ArunMathew7/DAV-5400/blob/main/Project1/all\_\_results.ipynb

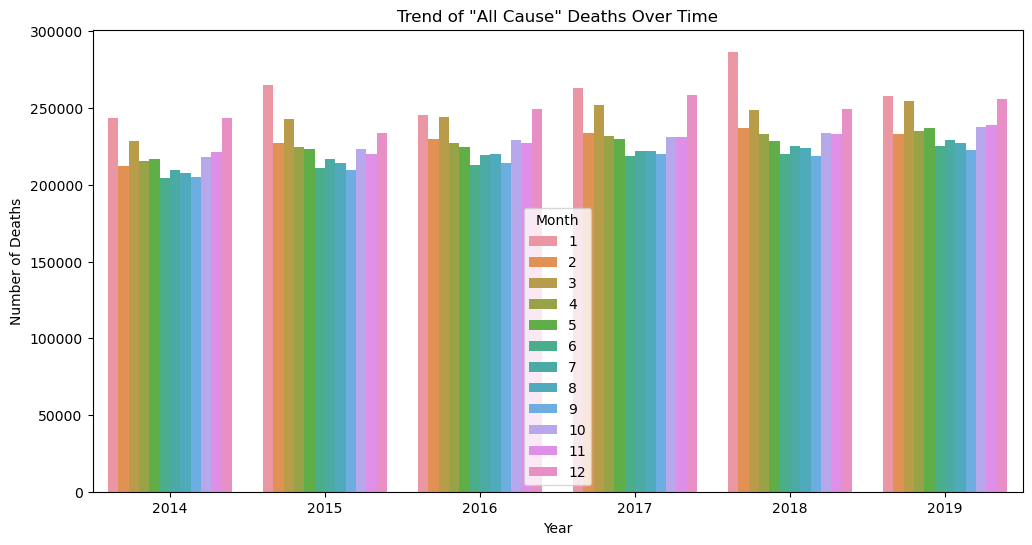
**Exploratory data analysis**

The initial steps in an Exploratory data analysis which are Reading of dataset, cleaning and interpretation, data types of each attributes, summary are done in the data summary part so I am not repeating those steps in this.

Python code for this EDA – <https://github.com/ArunMathew7/DAV-5400/blob/main/Project1/EDA.py>

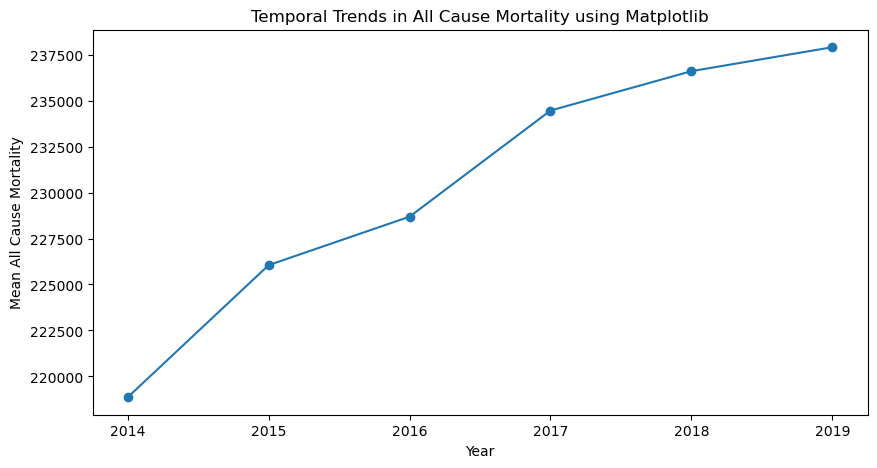
Outputs or graphical visualisation – <https://github.com/ArunMathew7/DAV-5400/blob/main/Project1/all__results.ipynb>

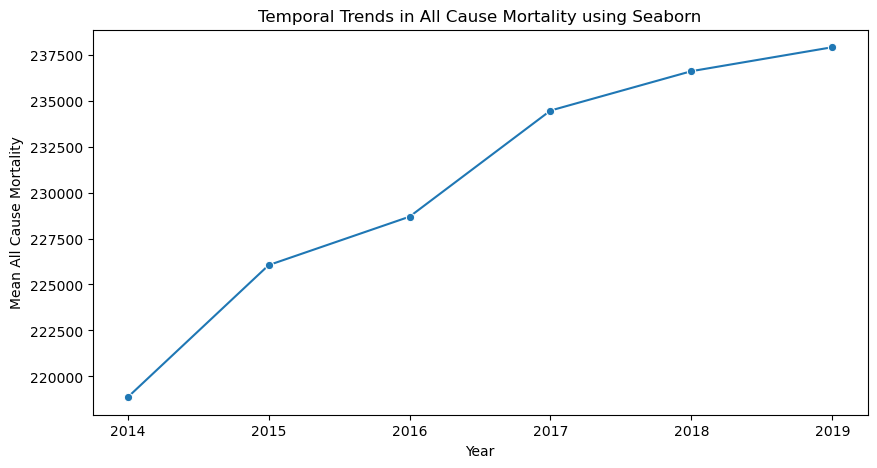
1. **Visualize the trend of "All Cause" deaths over time.**



The resulting plot will show how the number of "All Cause" deaths changes over the years, with each year's data separated by different colors for each month. This visualization helps in understanding the temporal trends in "All Cause" mortality. Analyzing this graph, we can see a similar dip in all the years in the month of June which means the number of death count is less is June and there is a highest death count is in the month of January from years 2014 to 2019.

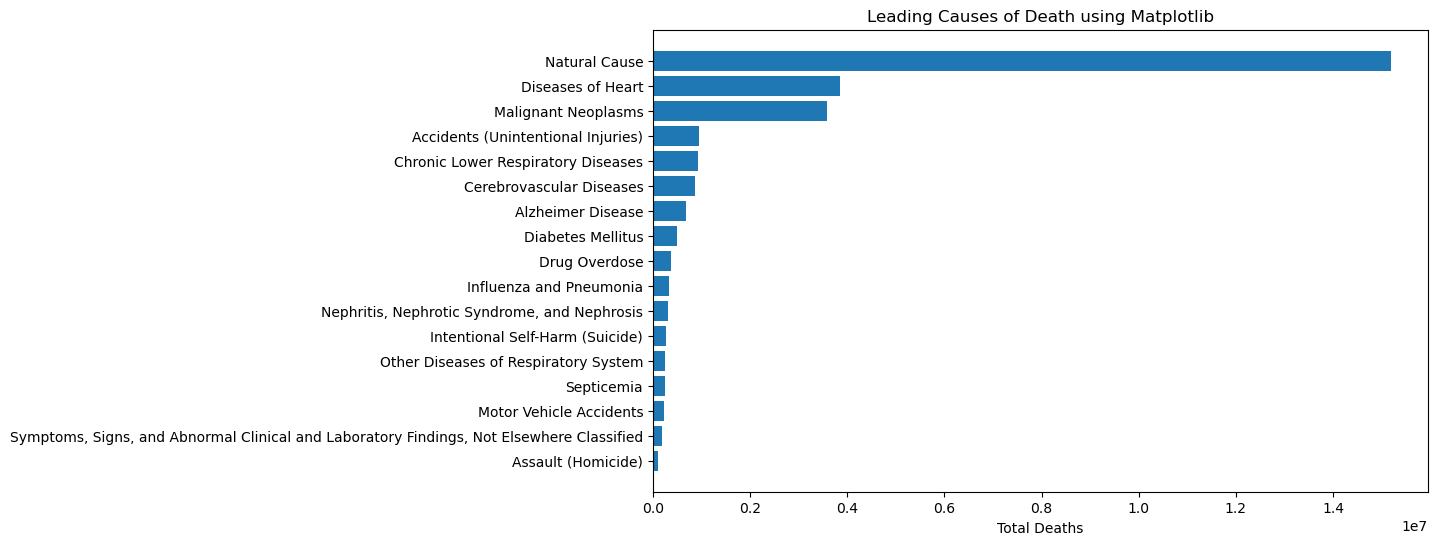
1. **Temporal trends in “All Cause” mortality**

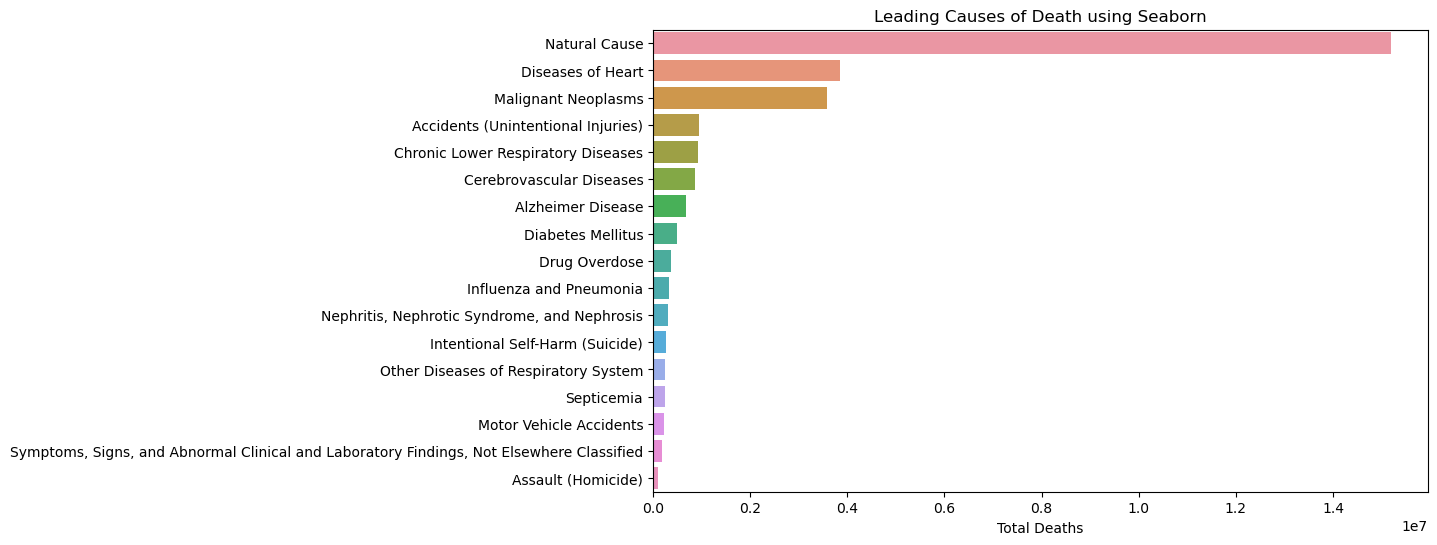




The resulting plot illustrates how the mean "All Cause" mortality varies over different years, helping to visualize temporal trends in this specific health metric. In this plot it clearly shows the death count is increasing as we move through the years.

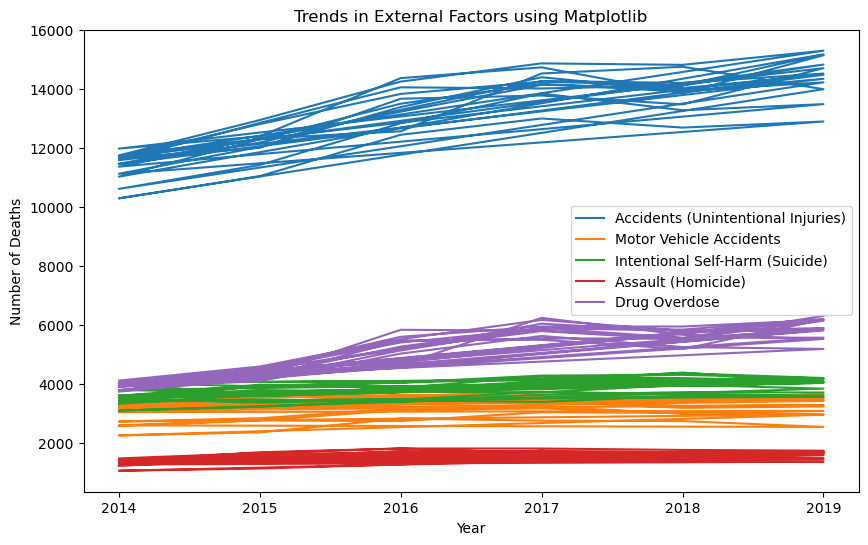
1. **Analyzing which what is the cause in leading deaths.**

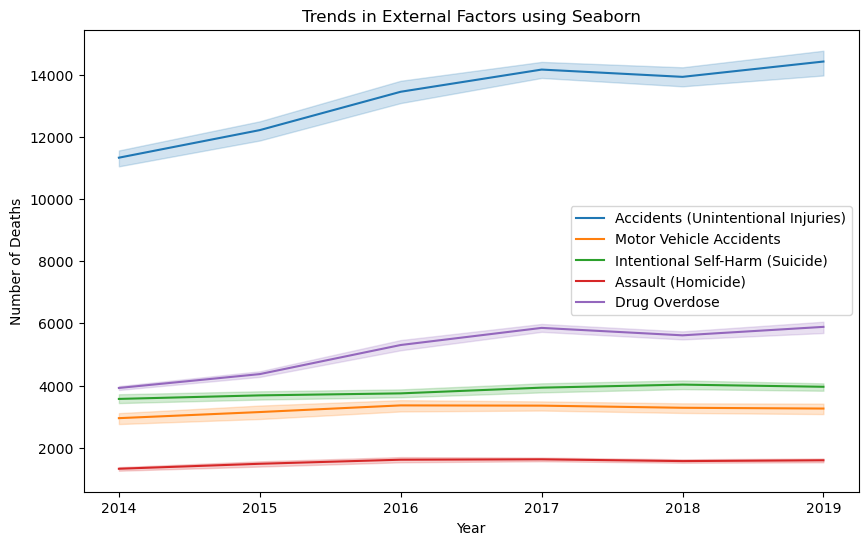




These plots illustrate the leading causes of death and their respective total death counts, making it easy to identify the primary causes of mortality in the dataset. The causes are displayed as horizontal bars, allowing for a clear comparison of their impact. This graph clearly shows that the major cause of death is natural cause and the least number deaths are by homicide.

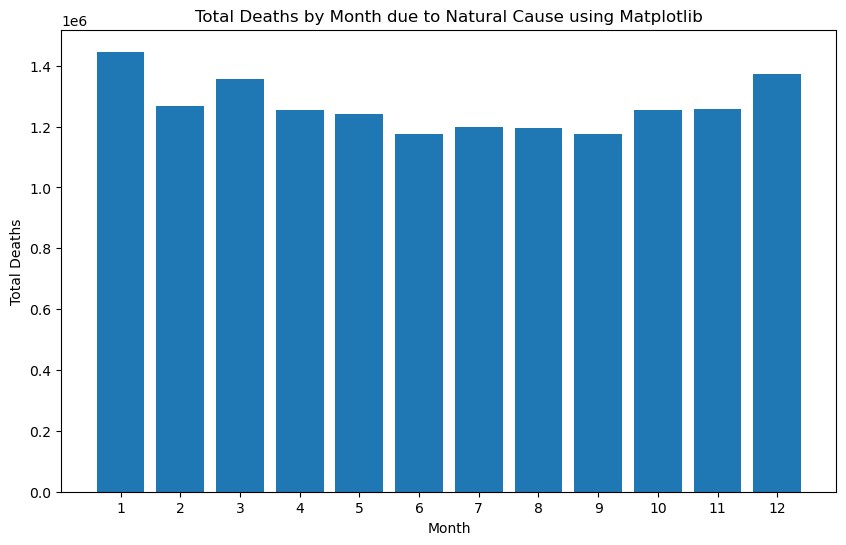
1. **Analyzing trends in deaths due to external factors.**

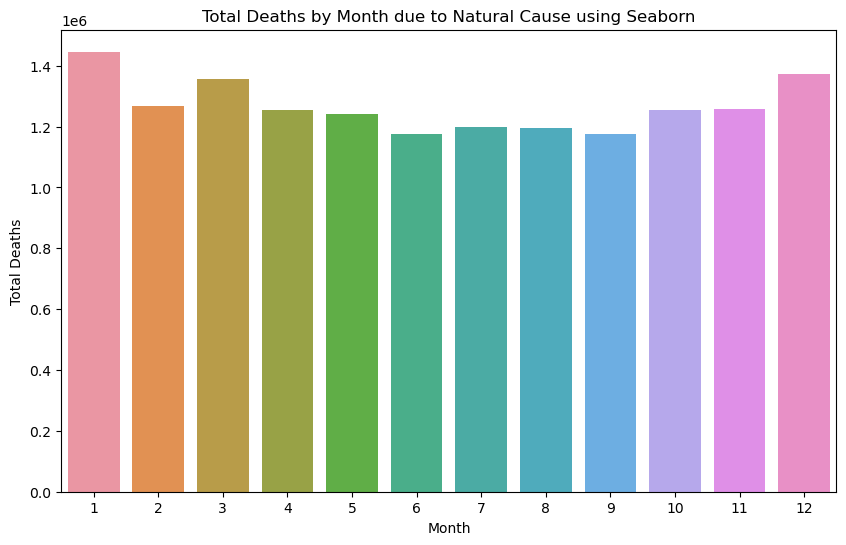




The resulting plot displays the trends over time for accidents (motor vehicle and unintentional), suicides or homicides and drug overdoses, allowing for a visual comparison of how these external factors have evolved. Considering the external factors most of the deaths are caused due to accident by unintentional injuries.

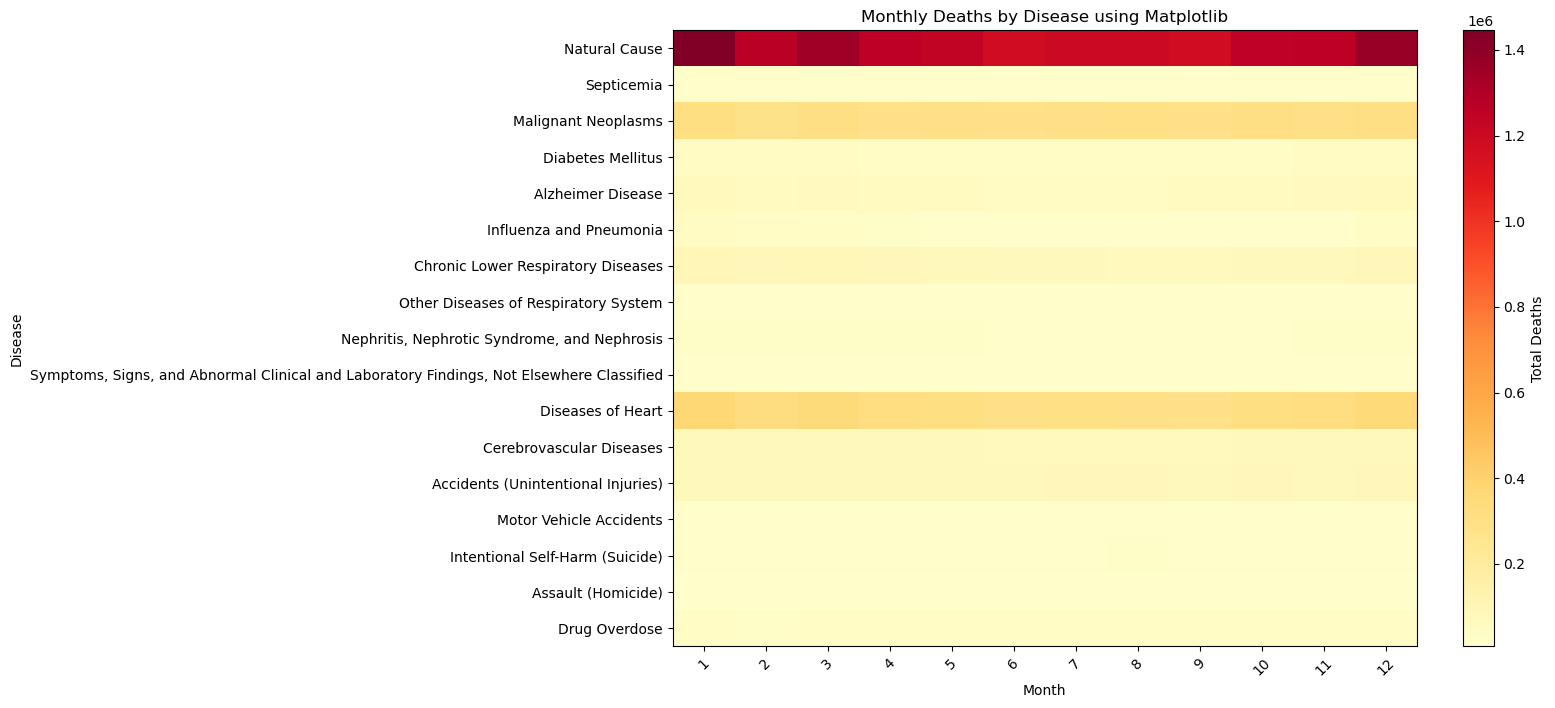
1. **As “Natural Cause” is the leading cause of death analyzing the count of death for each month for the years 2014 to 2019.**

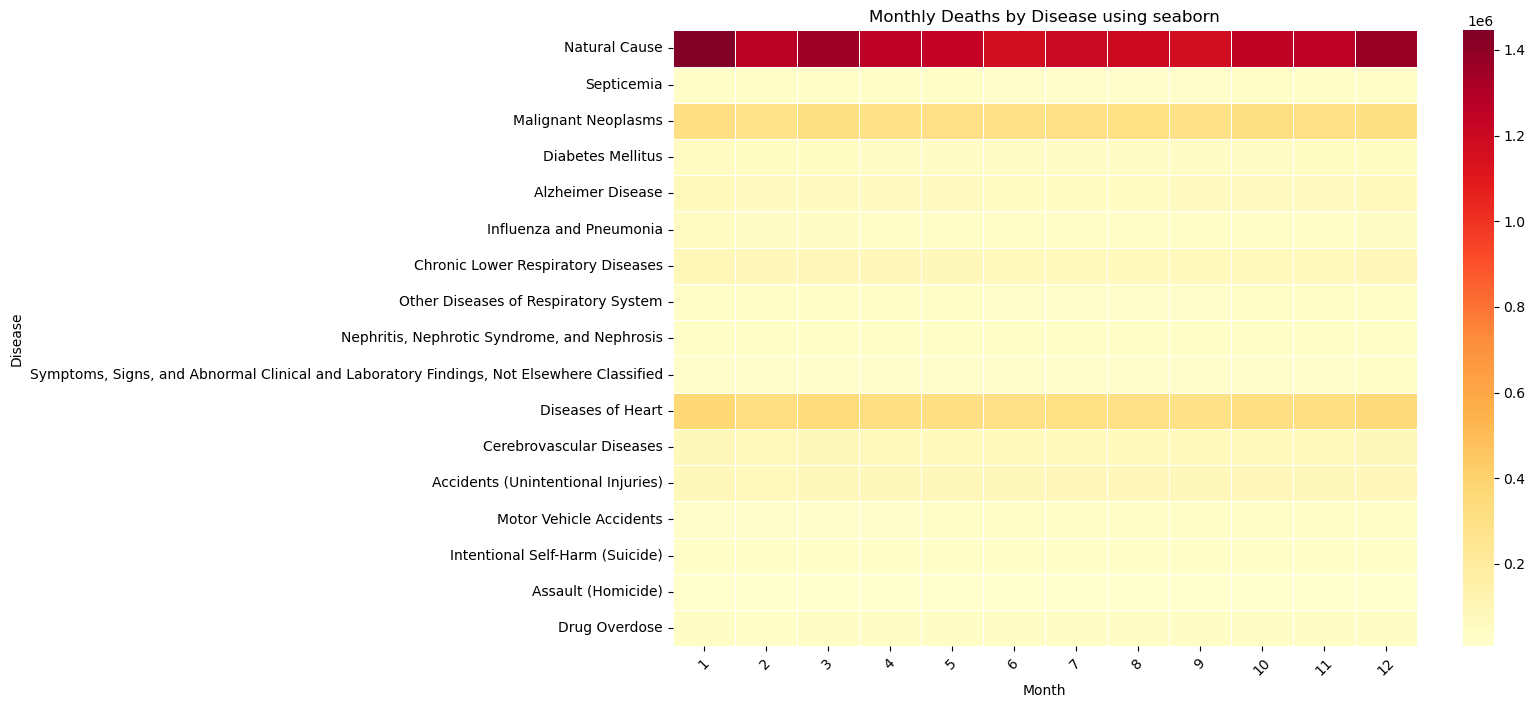




A visualization of the total deaths by month due to natural causes and identifies the month with the highest number of deaths. As we can see through the analysis for every year the number of deaths is high for the month of January for natural causes, through this we can conclude that January followed by December has the worst climatic conditions or disasters.

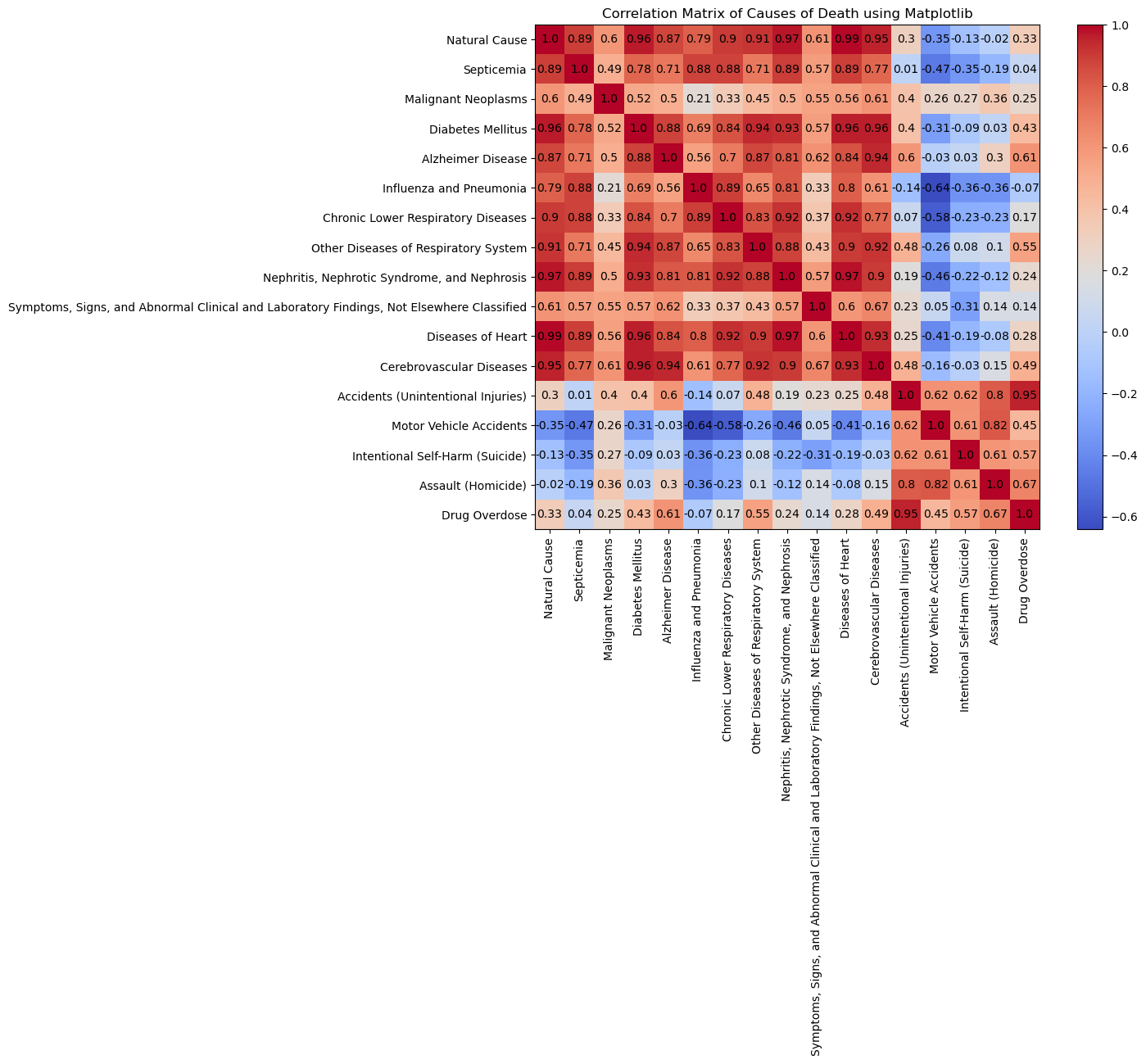
1. **A heatmap showing monthly death count by disease**

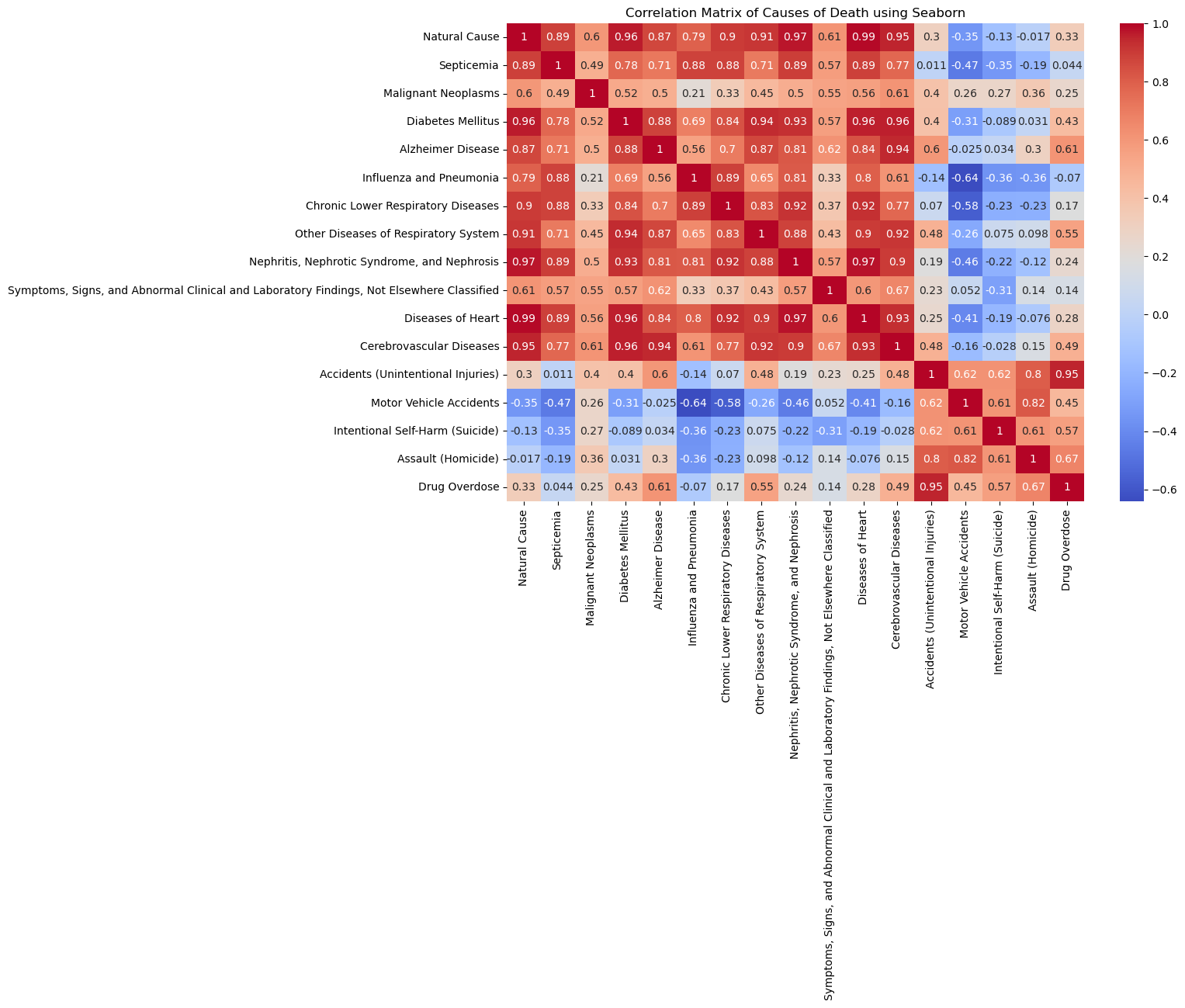




This is a visualization of monthly deaths by various diseases using a heatmap, making it easy to identify patterns and trends in disease-related deaths over time. In the previous plot, it is only about the death count due to the natural cause but here the visualization is about all the diseases vs month from the years 2014 to 2019.

1. **Creating a correlation matrix heatmap of different causes of death**





Finally, I am concluding my EDA using a correlation matrix which visualizes the correlation between different causes of death using a heatmap created with Matplotlib and Seaborn. It helps identify relationships and associations between various causes of death in the dataset.

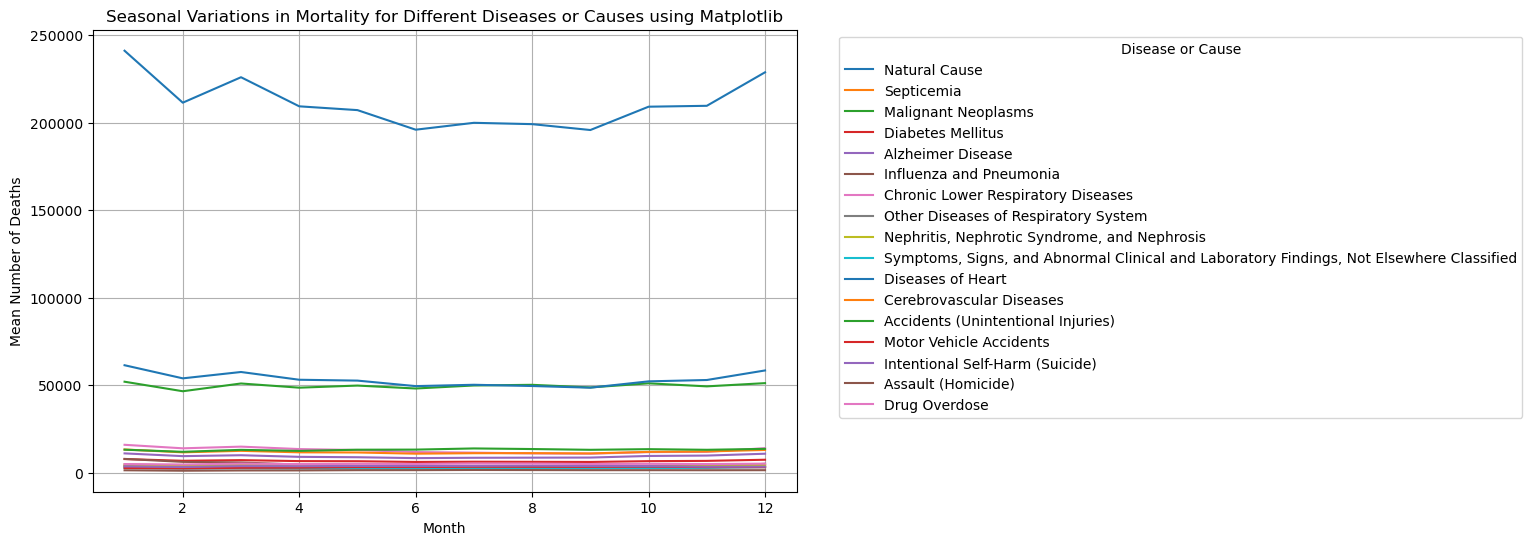
**Inference**

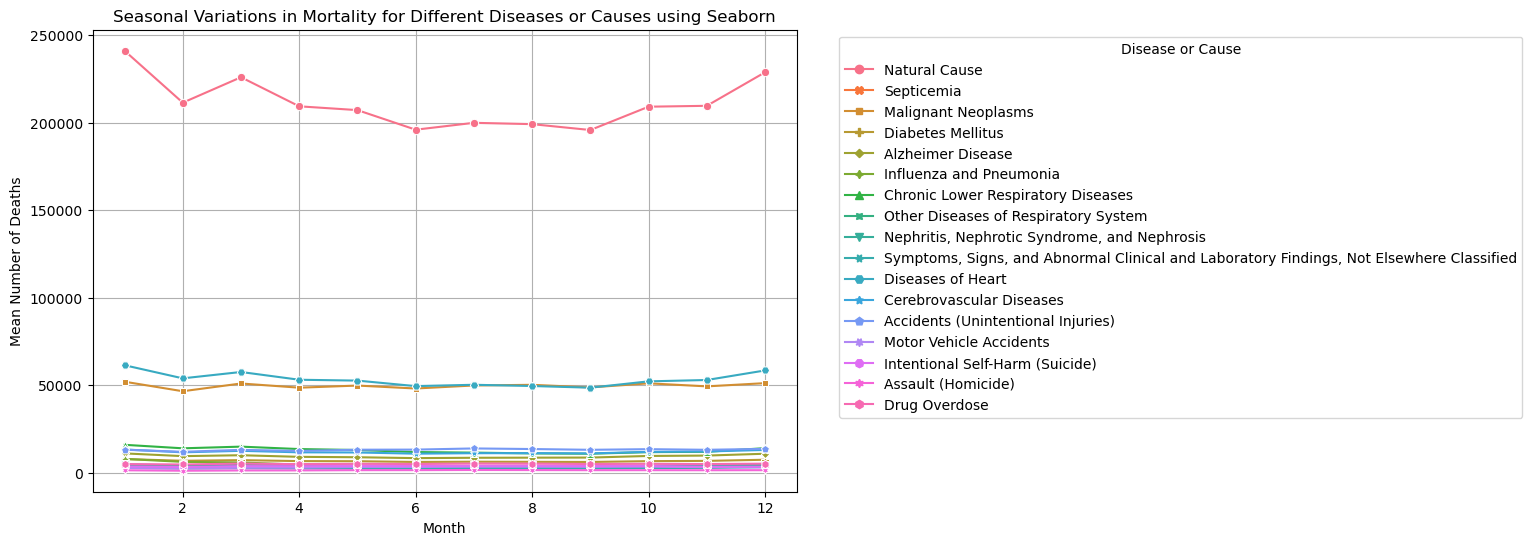
**Research Question**

Are there seasonal variations in mortality for specific diseases or causes, and if so, what might explain these variations?

**Steps to analyse this research question**

* Import the necessary libraries.
* Load the dataset into a pandas Data Frame.
* Select relevant columns which are disease columns
* The dataset is then grouped by the 'Month' column. For each month, it calculates the mean number of deaths for each of the diseases or causes. This grouping operation aggregates the data, summarizing the mean mortality statistics for each disease over the months.
* A line plot is generated using Seaborn and Matplotlib with plot title, legend and label.





We can see seasonal variations in deaths due to 3 causes such as natural causes, diseases of heart and malignant neoplasms. Another insight is the number of deaths of natural cause is much higher than all the causes of death and that is affected by the seasonal variations.

inference.py - <https://github.com/ArunMathew7/DAV-5400/blob/main/Project1/inference.py>

Output - <https://github.com/ArunMathew7/DAV-5400/blob/main/Project1/all__results.ipynb>

**Conclusion**

Analysing this plot visualizes seasonal variations in mortality for different diseases or causes over a span of months. By plotting each disease or cause separately, the viewer can observe how each one's mortality rate changes throughout the year. Patterns and trends specific to certain diseases or causes can be identified. For example, some diseases may exhibit seasonal spikes or declines.

The answer to my research question is yes, there are seasonal variations in mortality for specific diseases or causes. Taking a deep look into the graph there are 2 conclusions:

1. We can see seasonal variations in deaths due to 3 causes such as natural causes, diseases of heart and malignant neoplasms. Death counts during the months of January and December are higher for these diseases. During these months its winter and the temperatures are low this will weaken the immune system due to winter diseases and the blood vessels are more constrained during this time. So, people with diseases of heart and malignant neoplasm have more chances to die.
2. Another insight is the number of deaths of natural cause is much higher than all the causes of death and that is affected by the seasonal variations. So obviously this will increase the whole death count of all causes. It is due to severe climatic conditions in winter and natural disasters.

Combining these conclusions, it is confirmed that there is seasonal variations in mortality for specific diseases or causes.

**References**

Dataset Used - <https://catalog.data.gov/dataset/monthly-counts-of-deaths-by-select-causes-2014-2019-da9df>