Final Project

Ander Eguiluz

02-29-2024

ANALYZING MORTALITY PATTERNS IN THE UNITED STATES

Introduction

Understanding mortality trends and the leading causes of death is crucial for public health researchers, policymakers, and healthcare practitioners. Mortality data provides valuable insights into population health, helps identify emerging health threats, and guides the allocation of resources for disease prevention and intervention programs. This research aims to analyze mortality patterns in the United States from 2014 to 2023, examining how mortality rates and leading causes of death have evolved over the years. By exploring trends, variations, and potential factors influencing mortality, this study seeks to contribute to our understanding of public health dynamics and inform evidence-based interventions to improve population health outcomes.

Problem Statement

The problem statement addressed in this project is to analyze mortality patterns in the United States from 2014 to 2023. The key questions guiding this analysis include:

- How have overall mortality rates fluctuated over the years?
- What are the leading causes of death in the United States, and how have they changed over time?
- What impact did the COVID-19 pandemic have on mortality rates and leading causes of death?
- Are there seasonal variations in mortality rates for specific causes of death?
- Is there a correlation between certain health conditions and mortality rates?

Approach:

To address the problem statement, data was collected from multiple sources, including monthly counts of deaths by select causes and leading causes of death datasets. The datasets were cleaned, removing duplicates, standardizing column names, and handling missing values. The data was then merged based on common variables to create a comprehensive dataset for analysis. Descriptive statistics, time series analysis, and data visualization techniques were used to identify trends, patterns, and variations in mortality rates and leading causes of death over time. Statistical techniques were employed to quantify trends and assess their significance, and correlation analysis was used to explore relationships between health conditions and mortality rates. Various plots and charts were used to visualize the data and communicate the findings effectively.

Analysis:

```r
# Load required packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lubridate)
library(tidyr)
library(ggmosaic)
library(ggplot2)

# Load the datasets
dataset_2014_2019 <- read.csv("/Users/anderegiluz98/Desktop/Bellevue University/Winter 2023/Statistics f
dataset_2020_2023 <- read.csv("/Users/anderegiluz98/Desktop/Bellevue University/Winter 2023/Statistics f
dataset_NCHS <- read.csv("/Users/anderegiluz98/Desktop/Bellevue University/Winter 2023/Statistics for Da

# Check for missing values
sum(is.na(dataset_2014_2019))
```

```
## [1] 0
```

```r
sum(is.na(dataset_2020_2023))
```

```
## [1] 35
```

```r
sum(is.na(dataset_NCHS))
```

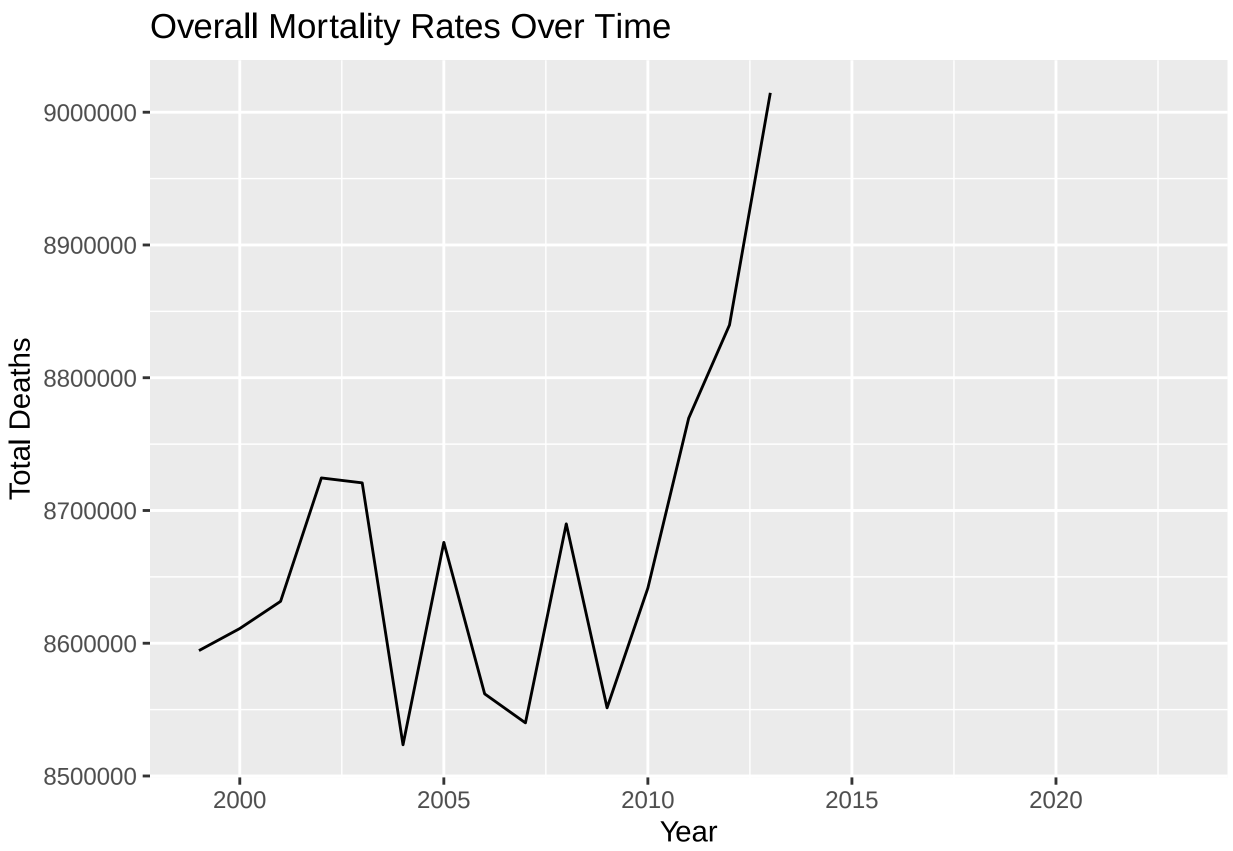
```
## [1] 0
```

```r
# Remove duplicates if any
dataset_2014_2019 <- unique(dataset_2014_2019)
dataset_2020_2023 <- unique(dataset_2020_2023)
dataset_NCHS <- unique(dataset_NCHS)

# Standardize column names
names(dataset_2014_2019) <- tolower(names(dataset_2014_2019))
names(dataset_2020_2023) <- tolower(names(dataset_2020_2023))
names(dataset_NCHS) <- tolower(names(dataset_NCHS))

# Remove the Data As Of, Start Date, and End Date columns from dataset_2020_2023
```

```r
dataset_2020_2023 <- subset(dataset_2020_2023, select = -c(data.as.of, start.date, end.date))

# Remove other specific columns from dataset_2020_2023
dataset_2020_2023 <- subset(dataset_2020_2023, select = -c(flag_accid, flag_mva, flag_suic, flag_homic,

# Adding Date
dataset_2020_2023 <- dataset_2020_2023 %>%
  mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-"), format = "%Y-%m-%d"))

#Adding Date
dataset_2014_2019 <- dataset_2014_2019 %>%
  mutate(YearMonth = as.Date(paste(year, month, "01", sep = "-"), format = "%Y-%m-%d"))
```

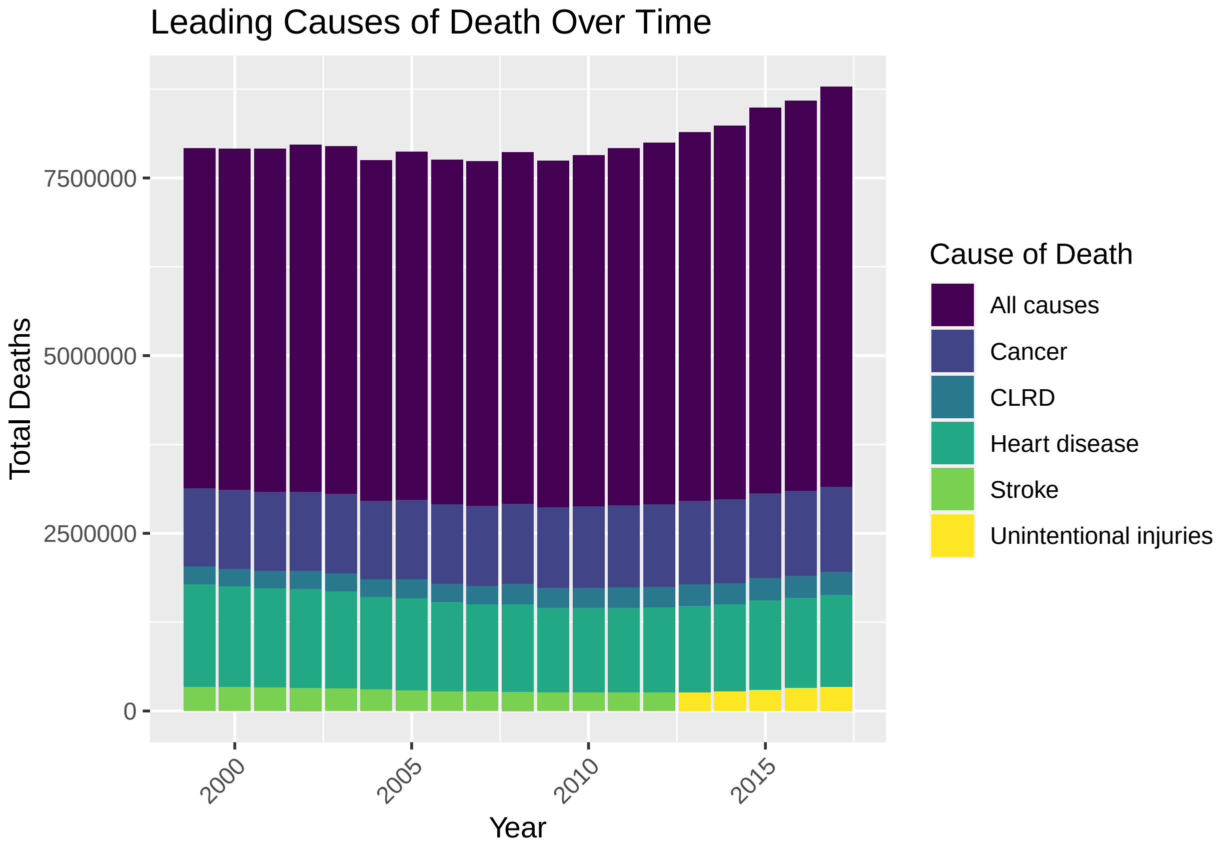
```r
# Question 1: How have overall mortality rates fluctuated over the years?

overall_mortality <- bind_rows(dataset_2014_2019, dataset_2020_2023, dataset_NCHS) %>%
  group_by(year) %>%
  summarise(deaths = sum(deaths))

# Plotting overall mortality rates over time
ggplot(overall_mortality, aes(x = year, y = deaths)) +
  geom_line() +
  labs(title = "Overall Mortality Rates Over Time",
       x = "Year",
       y = "Total Deaths")
```

```
## Warning: Removed 10 rows containing missing values (‘geom_line()‘).
```

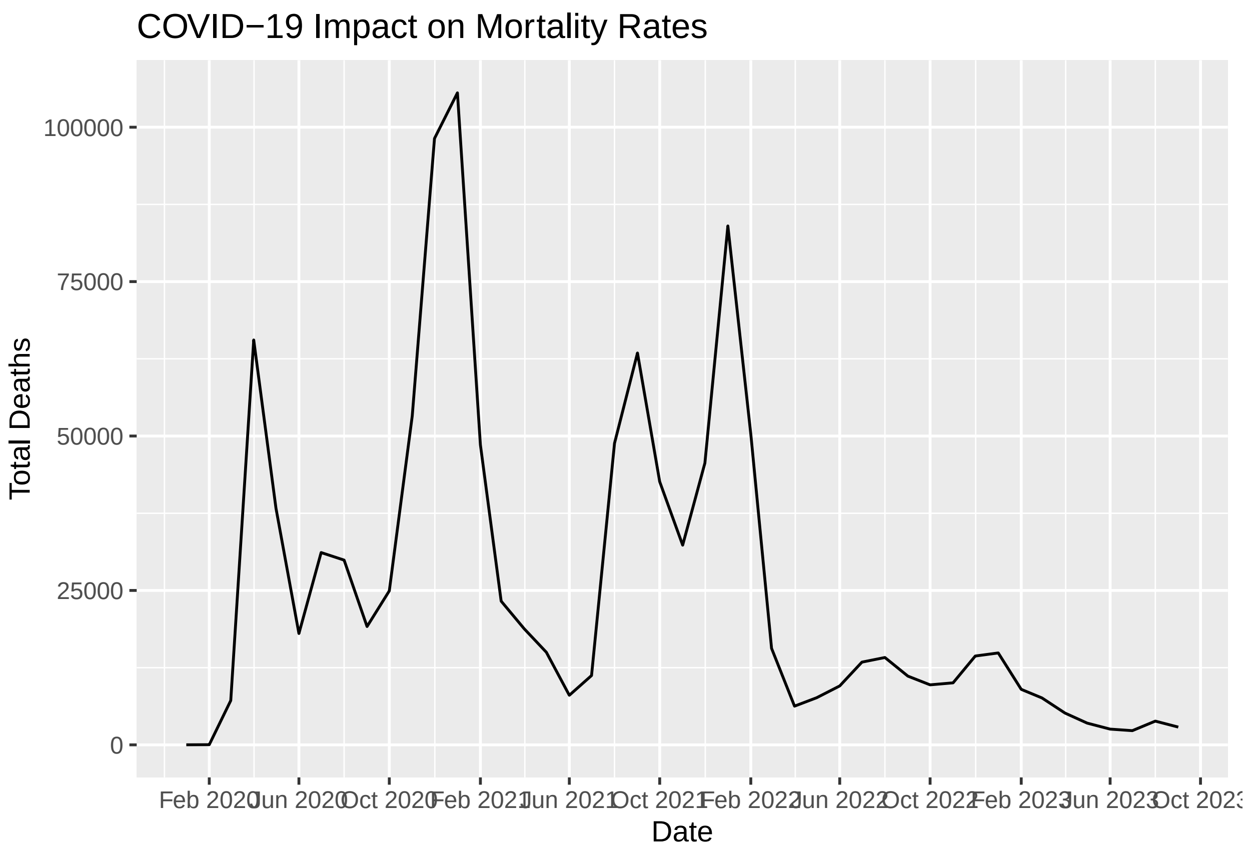


```
# Question 2: What are the leading causes of death in the United States, and how have they changed over

leading_causes <- bind_rows(dataset_2014_2019, dataset_2020_2023, dataset_NCHS) %>%
  group_by(year, `cause.name`) %>%
  summarise(deaths = sum(deaths), .groups = 'drop') %>%
  arrange(desc(deaths)) %>%
  group_by(year) %>%
  mutate(rank = row_number()) %>%
  filter(rank <= 5) # Top 5 leading causes

# Plotting leading causes of death over time
ggplot(leading_causes, aes(x = year, y = deaths, fill = `cause.name`)) +
  geom_bar(stat = "identity") +
  labs(title = "Leading Causes of Death Over Time",
       x = "Year",
       y = "Total Deaths",
       fill = "Cause of Death") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_viridis_d()
```

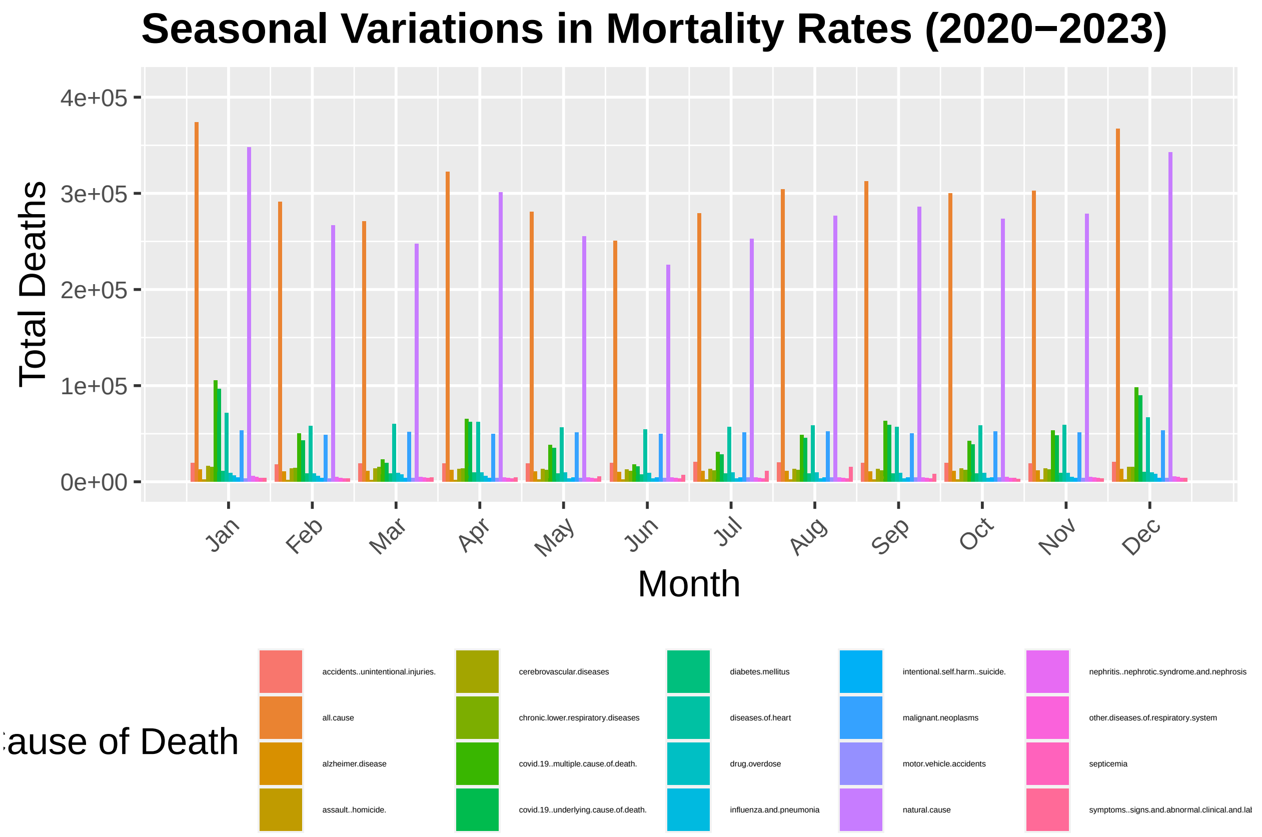
```
## Warning: Removed 6 rows containing missing values (‘position_stack()‘).
```



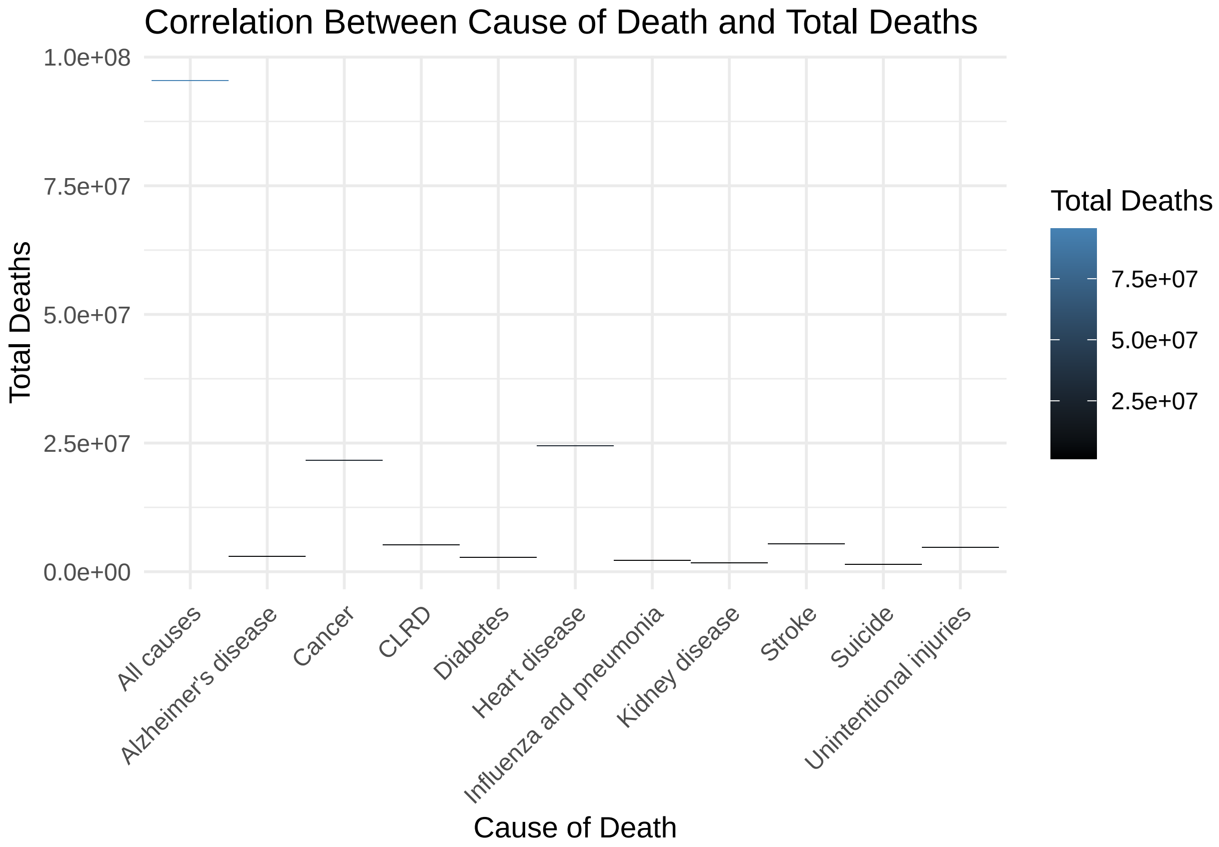
```
# Question 3: What impact did the COVID-19 pandemic have on mortality rates and leading causes of death

# Using all data points for COVID-19 (Multiple Cause of Death)
covid_impact <- dataset_2020_2023 %>%
  group_by(year, month) %>%
  summarise(deaths = sum(`covid.19..multiple.cause.of.death.`), .groups = 'drop')

# Plotting COVID-19 impact over time
ggplot(dataset_2020_2023, aes(x = YearMonth, y = covid.19..multiple.cause.of.death.)) +
  geom_line() +
  labs(title = "COVID-19 Impact on Mortality Rates",
       x = "Date",
       y = "Total Deaths") +
  scale_x_date(date_breaks = "4 months", date_labels = "%b %Y")
```



```r
# Question 4: Are there seasonal variations in mortality rates for specific causes of death?

# Reshape the dataset from wide to long format
dataset_2020_2023_long <- dataset_2020_2023 %>%
  pivot_longer(cols = -c(jurisdiction.of.occurrence, year, month, YearMonth),
               names_to = "cause_name",
               values_to = "deaths")

# Filter out non-numeric causes of death
dataset_2020_2023_long <- dataset_2020_2023_long %>%
  filter(!is.na(as.numeric(deaths)))

# Convert 'YearMonth' column to Date format
dataset_2020_2023_long$YearMonth <- as.Date(dataset_2020_2023_long$YearMonth)

# Plotting seasonal variations in mortality rates
ggplot(dataset_2020_2023_long, aes(x = month, y = deaths, fill = cause_name)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Seasonal Variations in Mortality Rates (2020-2023)",
       x = "Month",
       y = "Total Deaths",
       fill = "Cause of Death") +
  scale_x_continuous(breaks = 1:12, labels = month.abb) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size = 16, face = "bold"),
        axis.title.x = element_text(size = 14),
        axis.title.y = element_text(size = 14),
        legend.title = element_text(size = 14),
        legend.text = element_text(size = 3),  # Adjust legend text size
        legend.position = "bottom") +  # Move legend to the bottom
  coord_cartesian(ylim = c(0, max(dataset_2020_2023_long$deaths) * 1.1))  # Expand y-axis slightly
```



```
# Question 5: Is there a correlation between certain health conditions and mortality rates?

# Compute the correlation matrix
correlation_matrix <- dataset_NCHS %>%
  select(cause.name, deaths) %>%
  group_by(cause.name) %>%
  summarise(total_deaths = sum(deaths)) %>%
  ungroup() %>%
  mutate(cause.name = as.factor(cause.name))

# Plot the heatmap
ggplot(data = correlation_matrix, aes(x = cause.name, y = total_deaths, fill = total_deaths)) +
  geom_tile() +
  scale_fill_gradient(low = "black", high = "steelblue") +
  labs(title = "Correlation Between Cause of Death and Total Deaths",
       x = "Cause of Death", y = "Total Deaths",
       fill = "Total Deaths") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The analysis revealed several key findings. Overall mortality rates fluctuated over the years, with an increase observed during certain periods, including the COVID-19 pandemic in 2020-2023. The leading causes of death in the United States varied over the years, with diseases of the heart, malignant neoplasms, and chronic lower respiratory diseases consistently ranking among the top causes. The COVID-19 pandemic had a significant impact on mortality rates, particularly in 2020-2023, where a sharp increase in deaths attributed to COVID-19 was observed. Seasonal variations in mortality rates were identified for specific causes of death, with some causes showing higher mortality rates during certain months or seasons. Correlation analysis revealed a correlation between certain health conditions. For example, cases of cancer and heart disease are more likely to end up in death than any other condition explored in this anaysis.

Implications:

The findings from this analysis have several implications for public health policy and practice. They can inform the development of targeted interventions to reduce mortality rates, especially for the leading causes of death. The identification of seasonal variations in mortality rates can help in the planning and allocation of healthcare resources to address peak periods of mortality. The correlation analysis highlights the importance of addressing specific health conditions to reduce mortality rates.

Limitations:

One limitation of this analysis is the reliance on only three datasets, which may not capture the full complexity of mortality patterns in the United States. Additionally, the analysis focused on aggregate mortality data and did not consider individual-level factors that may influence mortality rates. Further research is needed to explore these factors in more detail. Another limitation is the lack of consideration for regional variations in mortality rates, which could provide additional insights into public health dynamics.

Concluding Remarks:

In conclusion, this project provides valuable insights into mortality patterns in the United States from 2014 to 2023. By analyzing trends, variations, and potential factors influencing mortality rates and leading causes of death, this study contributes to our understanding of public health dynamics and can inform evidence-based interventions to improve population health outcomes. Further research and analysis are warranted to explore the implications of these findings for public health policy and practice.