HW2

Cynthia Xu

2024-10-20

In this portfolio project, I explored suicide rates in the United States by visualizing data across several dimensions, including age groups, sex, and race. Using a combination of Shiny app and static visualizations like heatmaps, pair plots, and ridgeline plots, I was able to extract meaningful insights from the dataset, which originated from the National Center for Health Statistics (NCHS). These visualizations shed light on key trends and revealed surprising findings in how suicide rates vary across different population groups over time.

```
# Load required libraries
library(shiny)
library(ggplot2)
library(dplyr)
```

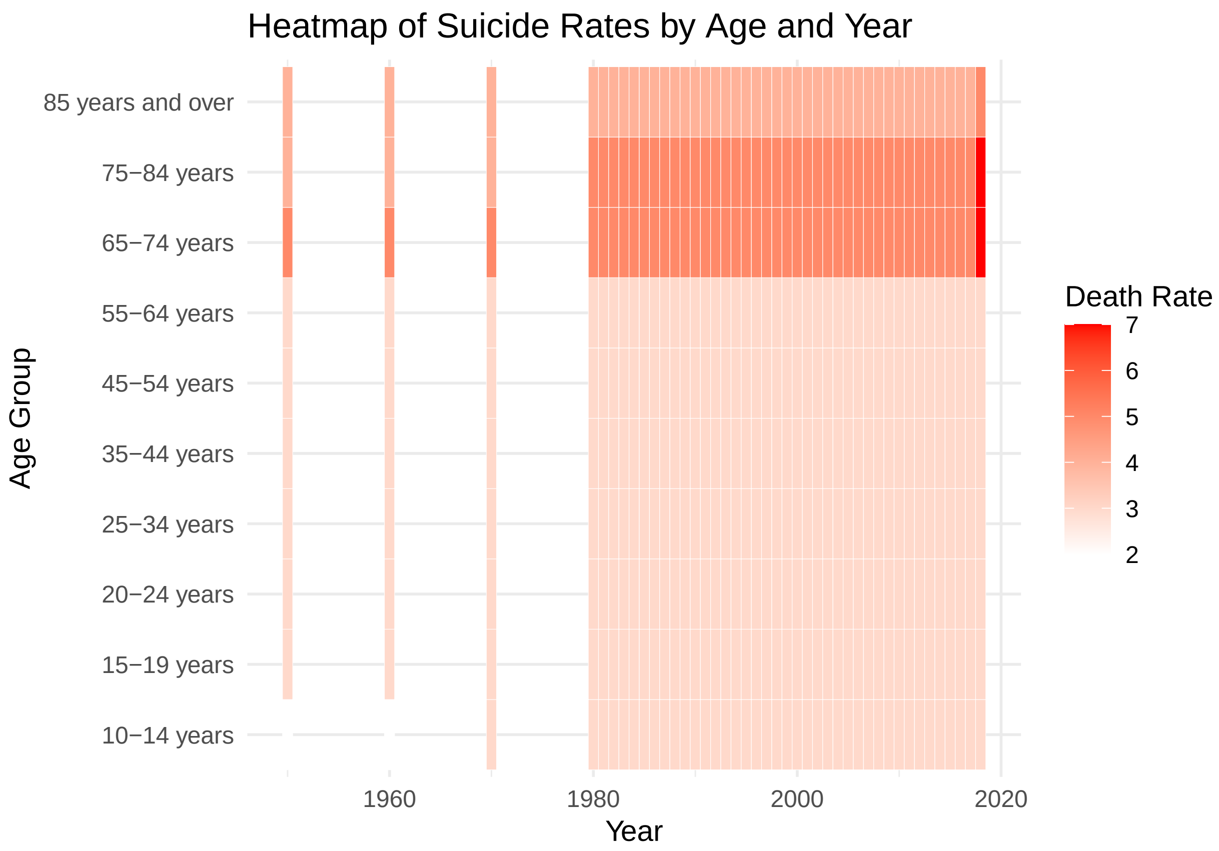
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Link to the publicly hosted CSV file
data_url <- "https://raw.githubusercontent.com/Cynthiaxu7/STAT436_HW2/main/HW2.csv"

# Load data from the public link
data <- read.csv(data_url)
# Data cleaning
data_cleaned <- data %>%
  filter(!is.na(ESTIMATE)) %>%   # Remove rows with missing estimates
  filter(!AGE %in% c("All ages", "15-24 years","25-44 years", "45-64 years", "65 years and over"))  # R
```

```
library(reshape2)

heatmap_data <- data_cleaned %>%
  dcast(YEAR ~ AGE, value.var = "ESTIMATE")
```

```
## Aggregation function missing: defaulting to length
```

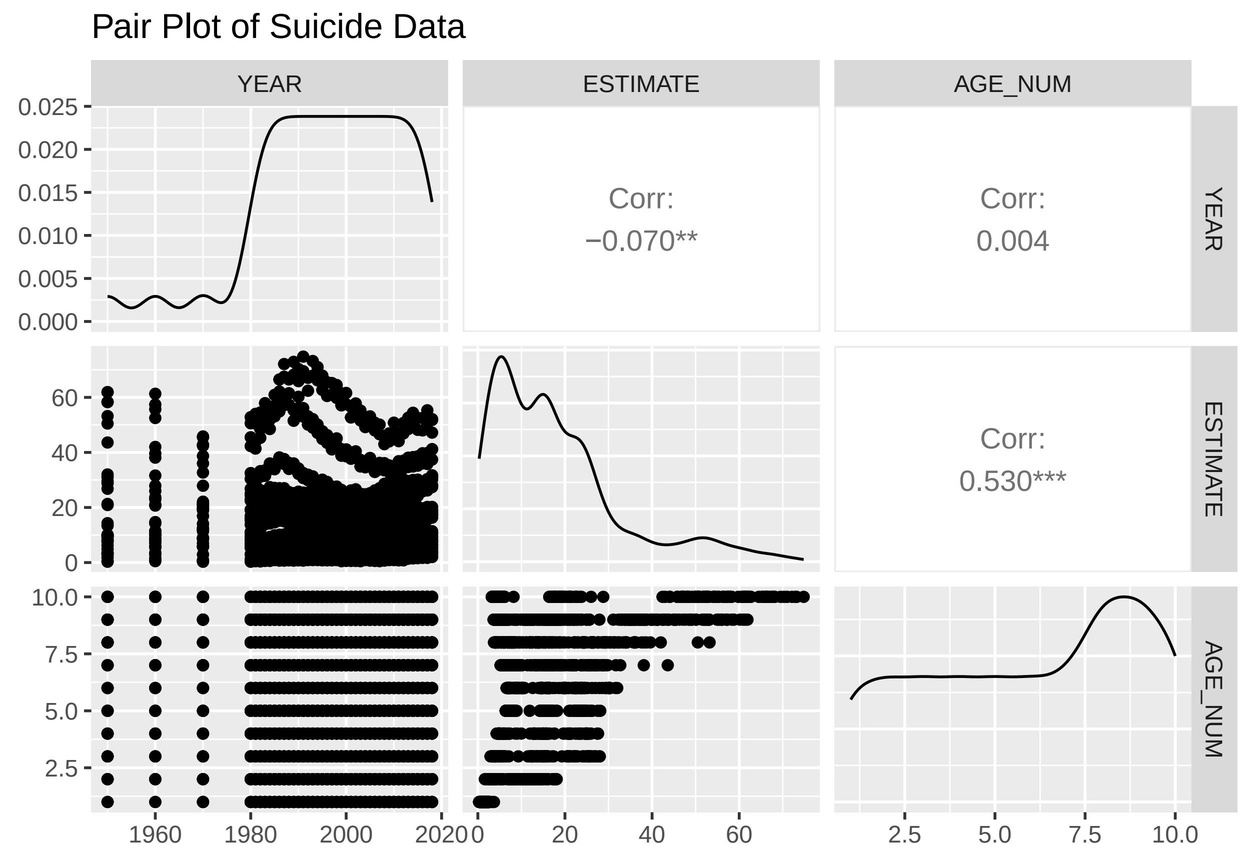
```
# Create the heatmap
ggplot(melt(heatmap_data, id.vars = "YEAR"), aes(x = YEAR, y = variable, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "white", high = "red") +
  labs(title = "Heatmap of Suicide Rates by Age and Year", x = "Year", y = "Age Group", fill = "Death Ra
  theme_minimal()
```



```
library(GGally)
```

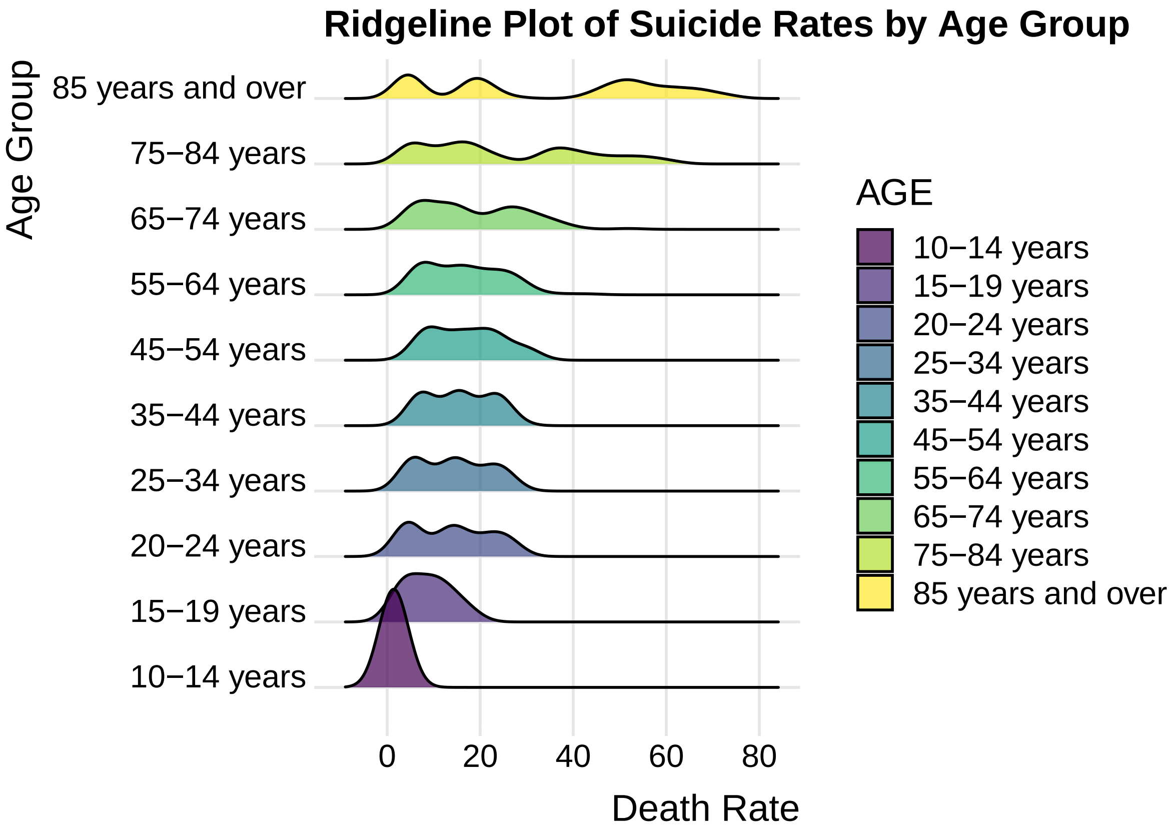
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
# Subset the data to include only the relevant numeric columns for pairwise plotting
pairplot_data <- data_cleaned %>%
  select(YEAR, ESTIMATE, AGE_NUM)

# Convert the `AGE` to a numeric factor for easier plotting
pairplot_data$AGE_NUM <- as.numeric(as.factor(pairplot_data$AGE))

# Generate the GGpairs plot
ggpairs(pairplot_data, title = "Pair Plot of Suicide Data", columns = 1:3)
```



```
library(ggridges)

# Create the ridgeline plot
ggplot(data_cleaned, aes(x = ESTIMATE, y = AGE, fill = AGE)) +
  geom_density_ridges(scale = 1.5, alpha = 0.7) +
  labs(title = "Ridgeline Plot of Suicide Rates by Age Group", x = "Death Rate", y = "Age Group") +
  theme_ridges() +
  scale_fill_viridis_d()
```

```
## Picking joint bandwidth of 3.1
```



Interesting Facts and Unexpected Findings

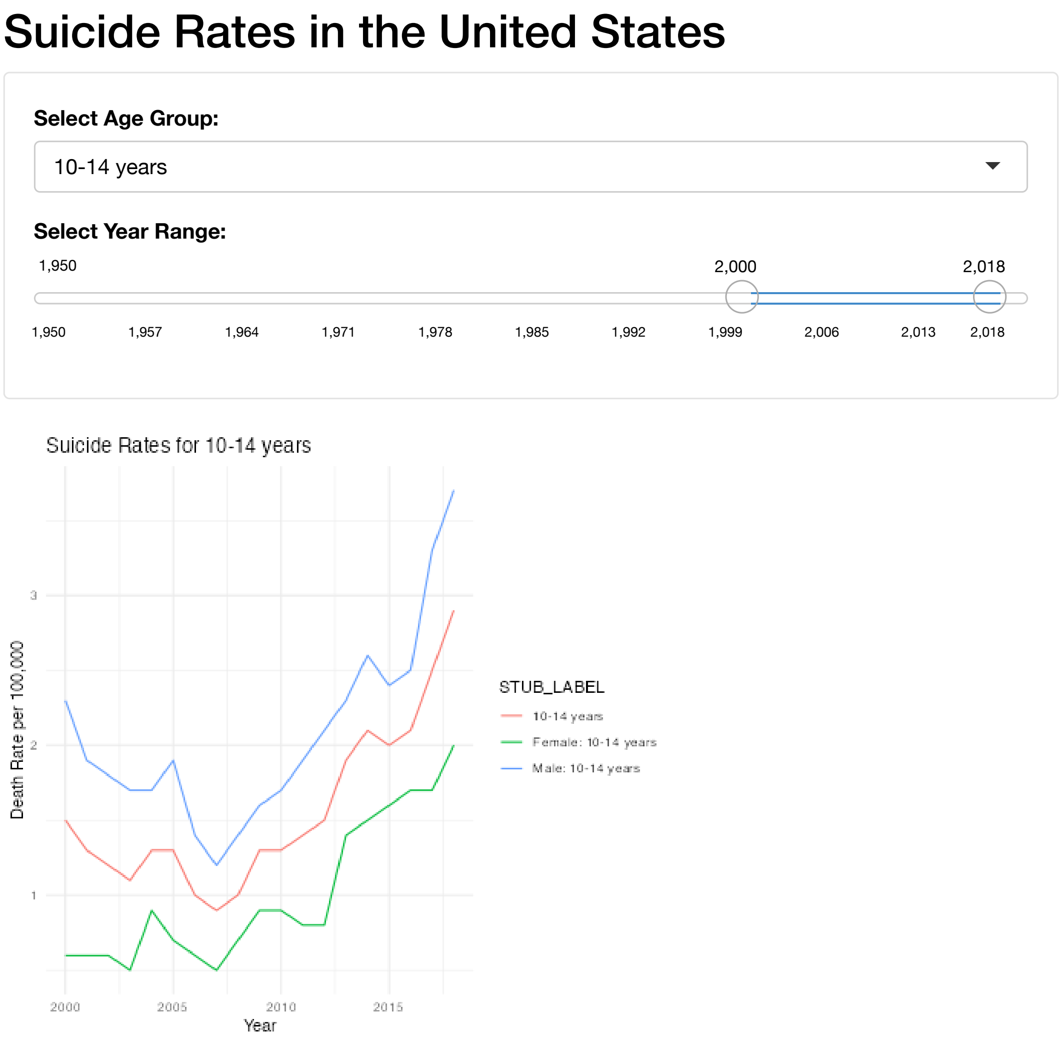
One of the more unexpected findings I encountered was the significant variability in suicide rates across different age groups based on the dataset I used. In particular:

- The heatmap clearly shows that older age groups (e.g., 65-74 years, 75-84 years) experience higher suicide rates compared to younger groups.
- However, despite lower rates, the ridgeline plot reveals a notable increase in suicide rates for younger age groups, especially from the late 1990s to the 2010s. This highlights a rising trend of suicides among younger populations, which was an alarming discovery.
- The pair plot revealed an interesting relationship between year and suicide estimates, showing some degree of correlation between age and suicide rates, though not as strong as expected (correlation values around 0.530). This suggests that while suicide rates change over time, they don't always follow a linear pattern for every age group.

Additionally, the density plot revealed that certain age groups (e.g., 65-74 years) have a wide distribution of suicide rates, implying significant variation in the rates across this group over different years. This highlights that suicide rates within a group may not be uniform and may require further investigation to understand the contributing factors.

```r
# UI for the application
ui <- fluidPage(
  titlePanel("Suicide Rates in the United States"),
  sidebarLayout(
    sidebarPanel(
      # Dropdown for selecting age group (choices from the cleaned data)
      selectInput("age_group", "Select Age Group:", choices = unique(data_cleaned$AGE)),

      # Slider for selecting a year range
      sliderInput("year_range", "Select Year Range:",
                  min = min(data_cleaned$YEAR), max = max(data_cleaned$YEAR), value = c(2000, 2018))
    ),

    mainPanel(
      # Output plot
      plotOutput("suicidePlot")
    )
  )
)

# Server logic
server <- function(input, output) {

  # Reactive expression to filter the data based on user inputs (age group and year range)
  filtered_data <- reactive({
    data_cleaned %>%
      filter(AGE == input$age_group,
             YEAR >= input$year_range[1], YEAR <= input$year_range[2])
  })

  # Render the plot based on the filtered data
  output$suicidePlot <- renderPlot({
    ggplot(filtered_data(), aes(x = YEAR, y = ESTIMATE, color = STUB_LABEL)) +
      geom_line() +                              # Line plot for suicide rates over time
      labs(title = paste("Suicide Rates for", input$age_group),
           x = "Year", y = "Death Rate per 100,000") +
      theme_minimal()                            # Use a clean minimal theme for the plot
  })
}

# Run the application
shinyApp(ui = ui, server = server)
```

```
##
## Listening on http://127.0.0.1:8791
```



Suicide Rates in the United States

Select Age Group:

10-14 years

Select Year Range:

2,000 – 2,018

Interface Design and Data Preparation

The Shiny app interface was designed to allow users to dynamically filter the data based on age group, year range, and race. This was achieved using dropdowns and sliders, which provided a smooth user experience and allowed users to interactively explore the data in different dimensions. I prioritized simplicity in the design while ensuring users could:

- 1. Select age groups and year ranges using the slider input.
- 2. Visualize line plots showing trends over time for different age and race groups .

Data Preparation was an essential part of this process. Initially, the dataset contained some redundant age groups like "All ages" and overlapping categories like "65 years and over". These were cleaned to prevent duplications and to ensure the plots were clear and meaningful. I also removed rows with missing values in the `ESTIMATE` column to avoid incomplete data affecting the visualizations.

In terms of style and layout, I opted for a minimal theme across all plots. This helped reduce distractions and focused the attention on the key trends in the data. For the heatmap , I used a red gradient color scale to highlight the intensity of suicide rates, making it easy to spot areas of concern (higher rates).

Reactive Graph Structure of the Shiny App

The Shiny app's reactive structure is centered around a filtered dataset based on user inputs:

- A reactive expression filters the data dynamically based on the selected age group and year range . This means that the visualizations update automatically whenever the user changes the input values.
- The core graph is a line plot that visualizes the suicide rate trends for the selected age group. The plot dynamically adjusts both the x-axis (year) and y-axis (suicide rate) based on the filtered data.

In addition to the basic line plot, the static visualizations (like heatmaps , ridgeline plots , and pair plots) were crucial for providing a deeper understanding of the dataset as a whole. These visualizations were pre-generated based on the cleaned dataset and allowed for a more in-depth comparison of trends across different dimensions (age, year).

Conclusion

The combination of the interactive Shiny app and various static visualizations provided a comprehensive way to explore suicide rates in the United States. Through this analysis, I was able to uncover surprising trends, such as the rising suicide rates among younger populations, and the significant variation within older age groups. These insights are valuable not only for understanding historical trends but also for informing potential interventions and public health strategies moving forward.