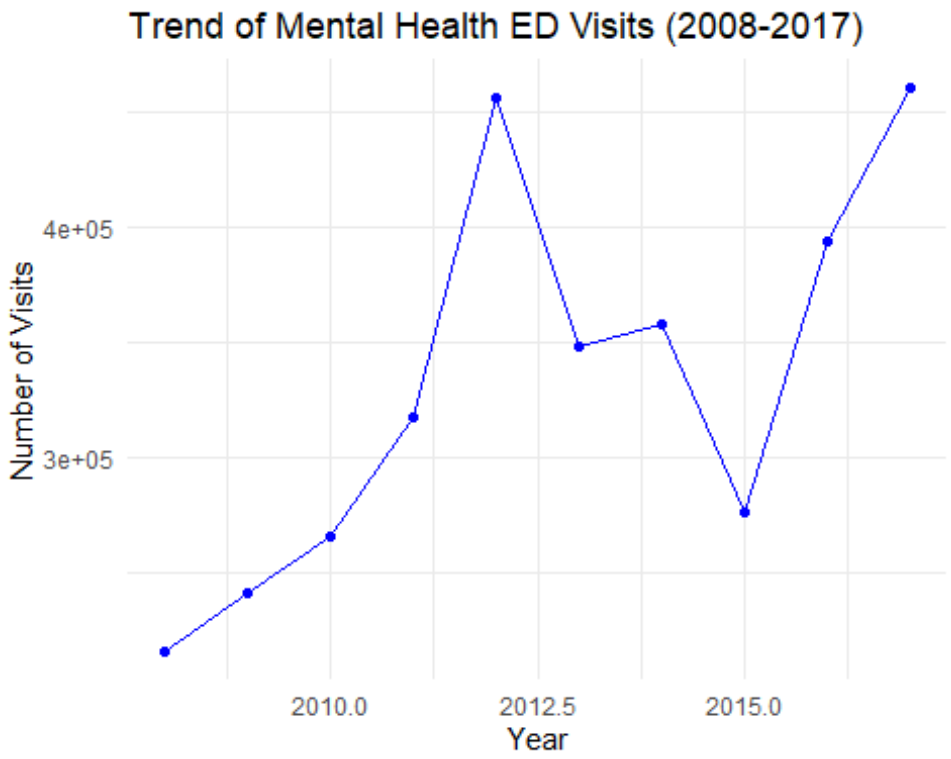
Week 4-8 Project

Vineet mehta

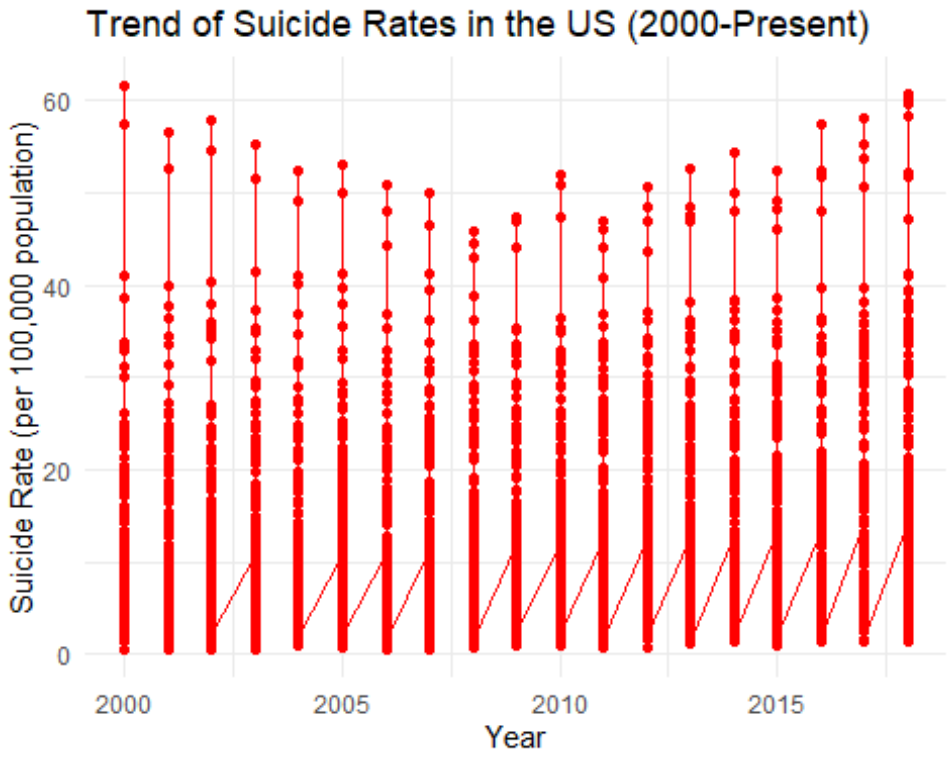
2024-10-20

```r
knitr::opts_chunk$set(echo = TRUE)
# Load libraries
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

library(ggplot2)

# Load datasets
suicide_rates <-
read.csv("Death_rates_for_suicide__by_sex__race__Hispanic_origin__and_age__Un
ited_States.csv")
mental_health_visits <-
read.csv("SHIP_Emergency_Department_Visits_Related_To_Mental_Health_Condition
s_2008-2017.csv")

# Descriptive statistics for suicide rates
suicide_summary <- suicide_rates %>%
  group_by(STUB_NAME, AGE) %>%  # Grouping by available demographic columns
  summarise(Mean_Rate = mean(ESTIMATE, na.rm = TRUE),
            SD_Rate = sd(ESTIMATE, na.rm = TRUE),
            Max_Rate = max(ESTIMATE, na.rm = TRUE))

## `summarise()` has grouped output by 'STUB_NAME'. You can override using
the
## `.groups` argument.

print(suicide_summary)

## # A tibble: 56 × 5
## # Groups:   STUB_NAME [12]
##    STUB_NAME AGE         Mean_Rate SD_Rate Max_Rate
##    <chr>     <chr>           <dbl>   <dbl>    <dbl>
##  1 Age       10-14 years      1.41   0.489      2.9
```

```
##  2 Age       15-19 years      8.74   1.94      11.8
##  3 Age       15-24 years     11.3    2.07      14.5
##  4 Age       20-24 years     13.8    2.20      17.4
##  5 Age       25-34 years     14.3    1.75      17.6
##  6 Age       25-44 years     15.0    1.19      17.9
##  7 Age       35-44 years     15.5    0.954     18.2
##  8 Age       45-54 years     17.0    2.29      20.9
##  9 Age       45-64 years     16.7    2.44      23.5
## 10 Age       55-64 years     16.3    3.02      26.8
## # ℹ 46 more rows

mental_health_data <-
read.csv("SHIP_Emergency_Department_Visits_Related_To_Mental_Health_Condition
s_2008-2017.csv")

# Convert the 'Value' column to numeric
mental_health_data$Value <- as.numeric(gsub(",", "",
mental_health_data$Value))

# Aggregate the data by year
mental_health_yearly <- aggregate(Value ~ Year, data = mental_health_data,
sum, na.rm = TRUE)

# Plotting the trend of mental health visits over time
ggplot(mental_health_yearly, aes(x = Year, y = Value)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(title = "Trend of Mental Health ED Visits (2008-2017)",
       x = "Year",
       y = "Number of Visits") +
  theme_minimal()
```




```
suicide_data <-
read.csv("Death_rates_for_suicide__by_sex__race__Hispanic_origin__and_age__Un
ited_States.csv")

# Convert 'ESTIMATE' to numeric
suicide_data$ESTIMATE <- as.numeric(suicide_data$ESTIMATE)

# Filter data from 2000 onwards
suicide_data_recent <- subset(suicide_data, YEAR >= 2000)

# Plotting the trend of suicide rates over time
ggplot(suicide_data_recent, aes(x = YEAR, y = ESTIMATE)) +
  geom_line(color = "red") +
  geom_point(color = "red") +
  labs(title = "Trend of Suicide Rates in the US (2000-Present)",
       x = "Year",
       y = "Suicide Rate (per 100,000 population)") +
  theme_minimal()

## Warning: Removed 143 rows containing missing values (`geom_point()`).
```

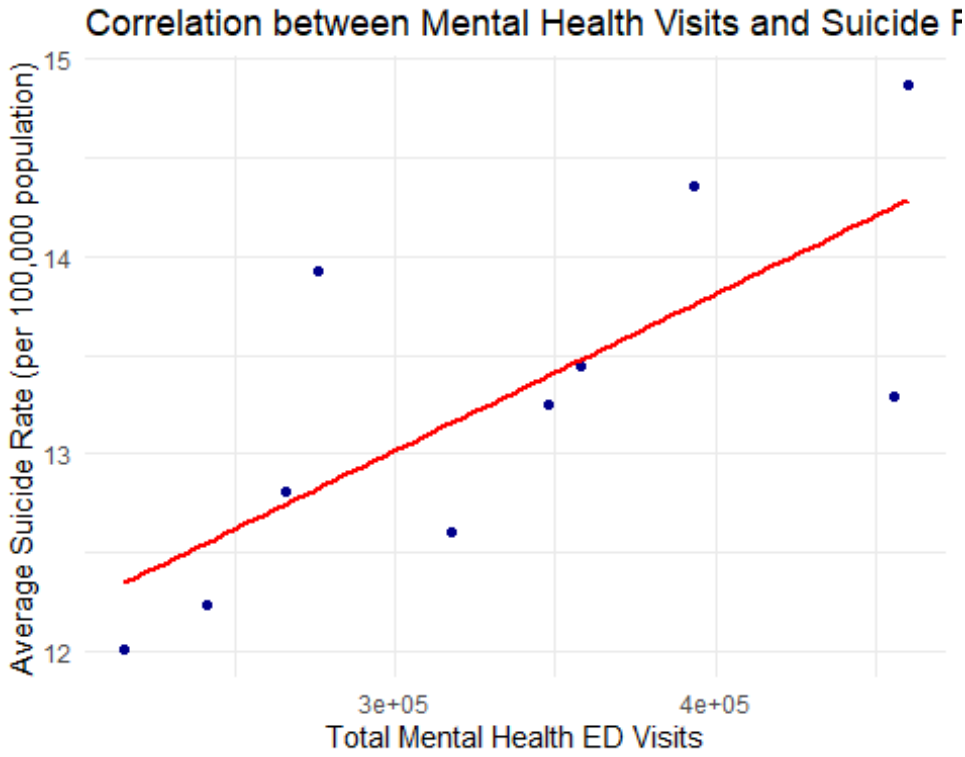



```
# Preprocess the mental health data
mental_health_data$Value <- as.numeric(gsub(",", "",
mental_health_data$Value))
mental_health_yearly <- mental_health_data %>%
  group_by(Year) %>%
  summarise(Total_Visits = sum(Value, na.rm = TRUE))

# Preprocess the suicide data (focus on total estimates from 2008 onwards)
suicide_data <- suicide_data %>%
  filter(YEAR >= 2008) %>%
  group_by(YEAR) %>%
  summarise(Average_Suicide_Rate = mean(ESTIMATE, na.rm = TRUE))

# Merge the datasets on common years
combined_data <- merge(mental_health_yearly, suicide_data, by.x = "Year",
by.y = "YEAR")

# Calculate Pearson correlation
correlation_result <- cor(combined_data$Total_Visits,
combined_data$Average_Suicide_Rate, method = "pearson")

# Print the correlation result
print(paste("Pearson Correlation Coefficient:", round(correlation_result,
3)))
```

```
## [1] "Pearson Correlation Coefficient: 0.742"
```

```r
# Plot the relationship between mental health visits and suicide rates
library(ggplot2)

ggplot(combined_data, aes(x = Total_Visits, y = Average_Suicide_Rate)) +
  geom_point(color = "darkblue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Correlation between Mental Health Visits and Suicide Rates",
       x = "Total Mental Health ED Visits",
       y = "Average Suicide Rate (per 100,000 population)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# Load the datasets
mental_health_data <-
read.csv("SHIP_Emergency_Department_Visits_Related_To_Mental_Health_Conditions_2008-2017.csv")
suicide_data <-
read.csv("Death_rates_for_suicide__by_sex__race__Hispanic_origin__and_age__United_States.csv")

# Preprocess the mental health data
mental_health_data$Value <- as.numeric(gsub(",", "",
mental_health_data$Value))
mental_health_yearly <- mental_health_data %>%
  group_by(Year) %>%
  summarise(Total_Visits = sum(Value, na.rm = TRUE))
```

```r
# Preprocess the suicide data (focus on total estimates from 2008 onwards)
suicide_data <- suicide_data %>%
  filter(YEAR >= 2008) %>%
  group_by(YEAR) %>%
  summarise(Average_Suicide_Rate = mean(as.numeric(ESTIMATE), na.rm = TRUE))

# Merge the datasets on common years
combined_data <- merge(mental_health_yearly, suicide_data, by.x = "Year",
by.y = "YEAR")

# Build a linear regression model
model <- lm(Average_Suicide_Rate ~ Total_Visits, data = combined_data)

# Calculate residuals
combined_data$residuals <- residuals(model)

# Plotting residuals
ggplot(combined_data, aes(x = Total_Visits, y = residuals)) +
  geom_point(color = "purple") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals of the Linear Regression Model",
       x = "Total Mental Health ED Visits",
       y = "Residuals") +
  theme_minimal()
```