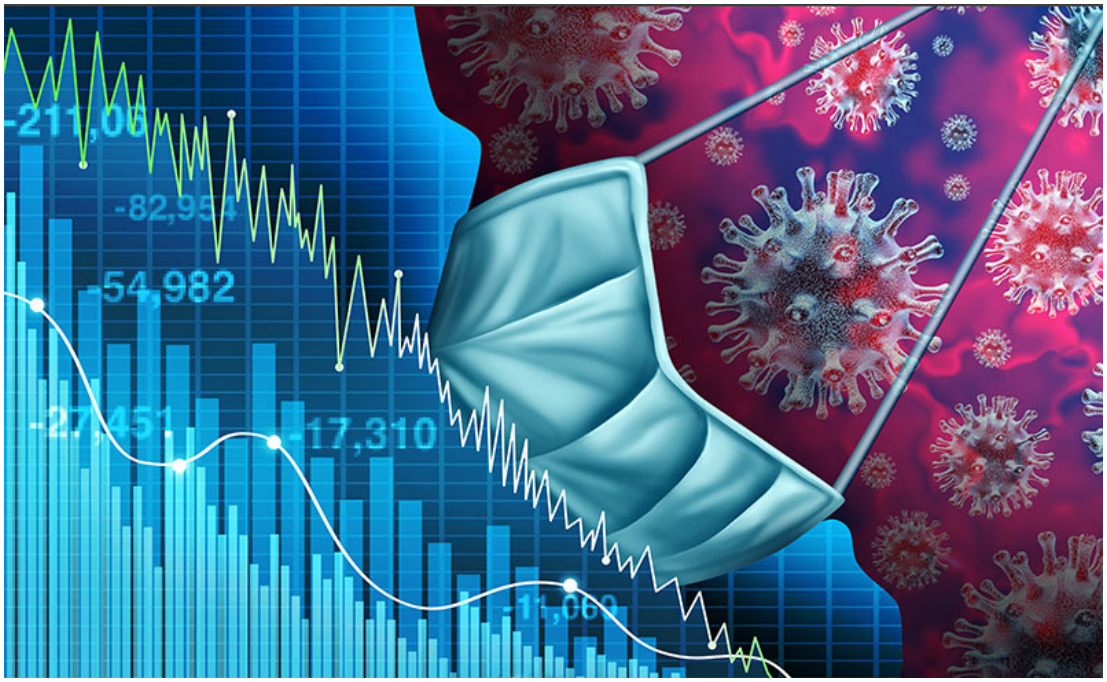
# Regional Contrasts in the Face of Pandemic: A Thorough Exploration of COVID-19 Impact Across U.S. Regions

Varun Putta

2024-01-20

- 


***Synopsis :***

The COVID-19 pandemic, also known as the coronavirus pandemic, is a global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The novel virus was first identified in an outbreak in the Chinese city of Wuhan in December 2019, and spread to other areas of Asia and then worldwide in early 2020. The World Health Organization (WHO) declared the outbreak a public health emergency of international concern (PHEIC) on 30 January 2020. The WHO ended its PHEIC declaration on 5 May 2023. As of 19 January 2024, the pandemic has caused 774,144,371 cases and 7,013,140 confirmed deaths, ranking it fifth in the list of the deadliest epidemics and pandemics in history.

***Objective :***

To conduct a thorough comparative analysis of COVID-19 impact across U.S. regions, considering death per case ratios as the primary metric. The project aims to rate regions based on both the prevalence of cases and the severity of outcomes, followed by a statistical analysis by correlation coefficient to provide robust insights.

```
## install and load the necessary libraries
#install.packages("tidyverse")
#install.packages("readr")
#install.packages("ggplot2")
#install.packages("janitor")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("lubridate")
#install.packages("sqldf")
```

```
suppressMessages(library(tidyverse)) ## To easily install and load the 'Tidyverse'
suppressMessages(library(readr)) ## To easily read rectangular data.
suppressMessages(library(tidyr)) ## To create tidy data
suppressMessages(library(ggplot2)) ## For mapping and ploting the data
suppressMessages(library(lubridate)) ## Dates and times made easy with lubridate
suppressMessages(library(janitor)) ## For cleaning and examining data
suppressMessages(library(dplyr)) ## Data Manipulation
suppressMessages(library(r02pro))
suppressMessages(library(usmap)) ## To plot US State maps
```

***Data Preparation*** :

One essential part of this project is to import and clean the data as needed. The Dataset provided contains the US COVID-19 data till 01/13/2021. The data is originally taken from The New York Times github repository: https://github.com/nytimes/covid-19-data

I. Importing data and creating the Dataframe

Setting up the working directory; 'US_counties.csv' Dataset has already been downloaded from Brightspace (University's academic platform)

```
setwd("/Users/varun/Desktop/Intermediate R project")
us_counties_r <- read_csv("us-counties.csv", show_col_types = FALSE)
```

II. Examining the DataFrame

```
head(us_counties_r)
```

```
## # A tibble: 6 x 6
##   date       county    state      fips  cases deaths
##   <date>     <chr>     <chr>      <chr> <dbl>  <dbl>
## 1 2020-01-21 Snohomish Washington 53061     1      0
## 2 2020-01-22 Snohomish Washington 53061     1      0
## 3 2020-01-23 Snohomish Washington 53061     1      0
## 4 2020-01-24 Cook      Illinois   17031     1      0
## 5 2020-01-24 Snohomish Washington 53061     1      0
## 6 2020-01-25 Orange    California 06059     1      0
```

Now we can see each column from this DataFrame, which are 1 Date, 3 characters( County, Fips & State) and 2 Doubles( Cases & Deaths)

```
str(us_counties_r)
```

```
## spc_tbl_ [927,008 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date  : Date[1:927008], format: "2020-01-21" "2020-01-22" ...
##  $ county: chr [1:927008] "Snohomish" "Snohomish" "Snohomish" "Cook" ...
##  $ state : chr [1:927008] "Washington" "Washington" "Washington" "Illinois" ...
##  $ fips  : chr [1:927008] "53061" "53061" "53061" "17031" ...
##  $ cases : num [1:927008] 1 1 1 1 1 1 1 1 1 1 ...
##  $ deaths: num [1:927008] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   date = col_date(format = ""),
##   ..   county = col_character(),
##   ..   state = col_character(),
##   ..   fips = col_character(),
##   ..   cases = col_double(),
##   ..   deaths = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
unique(us_counties_r['state'])
```

```
## # A tibble: 55 x 1
##    state
##    <chr>
##  1 Washington
##  2 Illinois
##  3 California
##  4 Arizona
##  5 Massachusetts
##  6 Wisconsin
##  7 Texas
##  8 Nebraska
##  9 Utah
## 10 Oregon
## # i 45 more rows
```

```
n_distinct(us_counties_r$county)
```

```
## [1] 1930
```

There are 927,008 rows and 6 columns. The variables in the data are date, county, state, fips, cases and deaths. The columns Deaths and Cases have cumulative values according to the Date column. There are 55 states and 1930 Counties we'll be studying on.

III. Create Tidy Data :

- The raw data has been loaded, now we need to pre-process it in order to get the data into a tidy format. To begin with, we need to find if there are any missing values and duplicate columns or rows.

```
#Checking duplicate data

us_counties_r[duplicated(us_counties_r),]
```

```
## # A tibble: 0 x 6
## # i 6 variables: date <date>, county <chr>, state <chr>, fips <chr>,
## #   cases <dbl>, deaths <dbl>
```

```
#There are no duplicate points in this DataFrame.
```

```
#Checking missing values

colSums(is.na(us_counties_r))
```

```
##   date county  state   fips  cases deaths
##      0      0      0   8664      0  19775
```

- There are 8664 and 19775 Missing values in Fips and deaths

```
distinct_state <-unique(us_counties_r$state[is.na(us_counties_r$deaths)])

## To know which state has missing value in deaths
distinct_state
```

```
## [1] "Puerto Rico"
```

- The variable "Fips" is not useful for the project, hence we will remove that variable. We also know that missing values in terms of death variable correspondes to the state of "Puerto Rico". We wil drop Puerto Rico from the dataset when needed

```
us_counties_r <- us_counties_r %>%
  select(-fips) %>%
  ## To make sure all the column names are inline with naming convention
  clean_names()
```

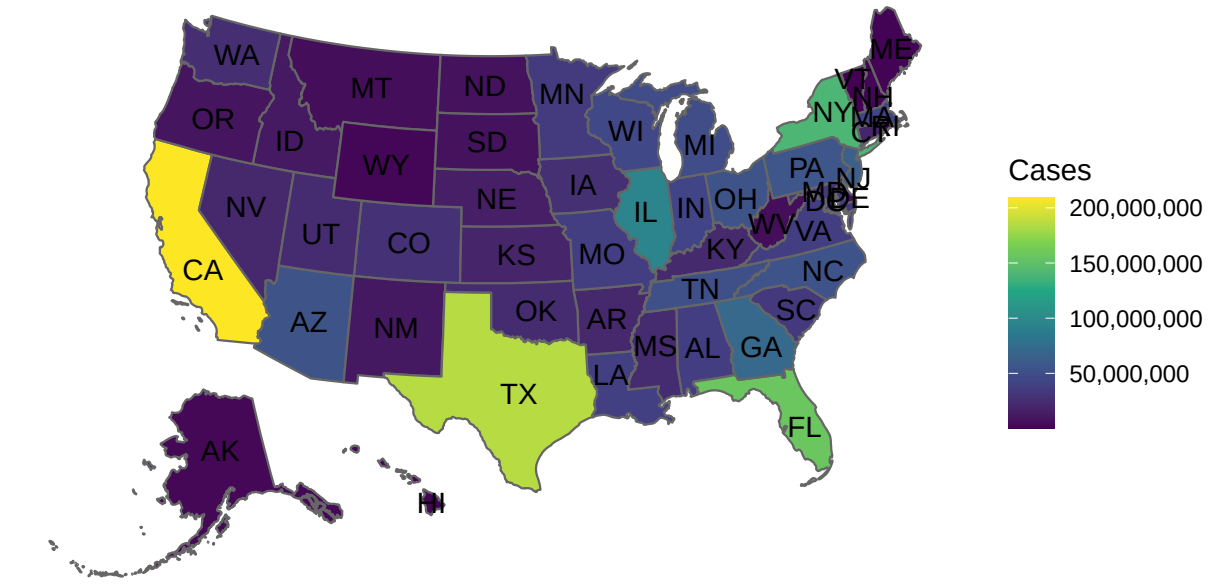
- For the project, we need to add a variable "region" to the "us_counties_r" dataframe. For this "State.region" data will be used which is pre-set in Base R.

```
state.info <- tibble(state = state.name, state.region)
state.info <- state.info %>%
  rename(region = state.region)
us_counties_r <- inner_join(us_counties_r, state.info, by = "state")
glimpse(us_counties_r) #To get a glimpse of data
```

```
## Rows: 904,814
## Columns: 6
## $ date   <date> 2020-01-21, 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-24, 20~
## $ county <chr> "Snohomish", "Snohomish", "Snohomish", "Cook", "Snohomish", "Or~
## $ state  <chr> "Washington", "Washington", "Washington", "Illinois", "Washingt~
## $ cases  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ deaths <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ region <fct> West, West, West, North Central, West, West, North Central, Wes~
```

***Exploratory Data Analysis:***

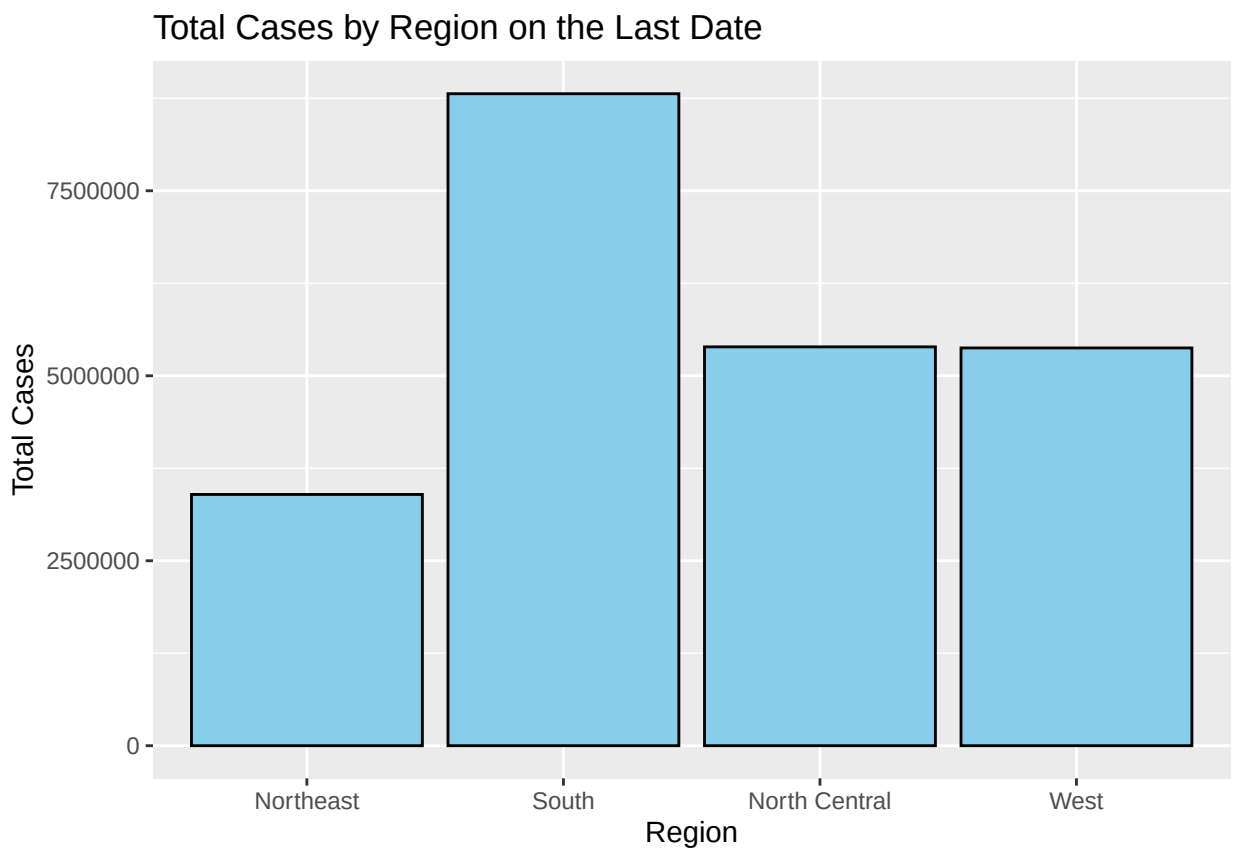
```
#Let us analyze the scattering of total cases over states in the US

us_counties_rt <- us_counties_r %>%
  group_by(state) %>%
  summarize(cases_total_main = sum(cases),
            deaths_total_main = sum(deaths),
            average_cases=mean(cases),
            average_deaths=mean(deaths)) %>%
  arrange(desc(cases_total_main))

us_counties_rt$state <- factor(us_counties_rt$state,
            levels = us_counties_rt$state[
              order(us_counties_rt$cases_total_main)])

#change the gradient type
#add title
plot_usmap(data = us_counties_rt,
  values = "cases_total_main",
  color = "grey40", labels = TRUE) +
  scale_fill_continuous(type = 'viridis', label = scales::comma) +
  labs(title = "COVID-19 - Total Number of Cases for All States",
    fill = "Cases") +
  theme_classic() +
  theme(
    panel.background = element_blank(),
    panel.border = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "right",
    axis.line = element_blank(),
    axis.ticks = element_blank(),
    axis.text.x = element_blank(),
    axis.text.y = element_blank(),
    axis.title.x = element_blank(),
    axis.title.y = element_blank()
  )
```




```
# Filter data for the last date
latest_data <- us_counties_r %>%
  filter(date == max(date))

# Summarize cases by region for the last date
all_region_cases <- latest_data %>%
  group_by(region) %>%
  summarize(`Total Cases` = sum(cases)) %>%
  arrange(desc(`Total Cases`))

# Now you have the top state.region with the highest number of cases on the last date
all_region_cases
```

```
## # A tibble: 4 x 2
##   region        `Total Cases`
##   <fct>                 <dbl>
## 1 South               8811189
## 2 North Central       5390115
## 3 West                5375633
## 4 Northeast           3395548
```

```
# plot this data into Bar chart

ggplot(all_region_cases, aes(x = region, y = `Total Cases`)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  labs(title = "Total Cases by Region on the Last Date", x = "Region", y = "Total Cases")
```

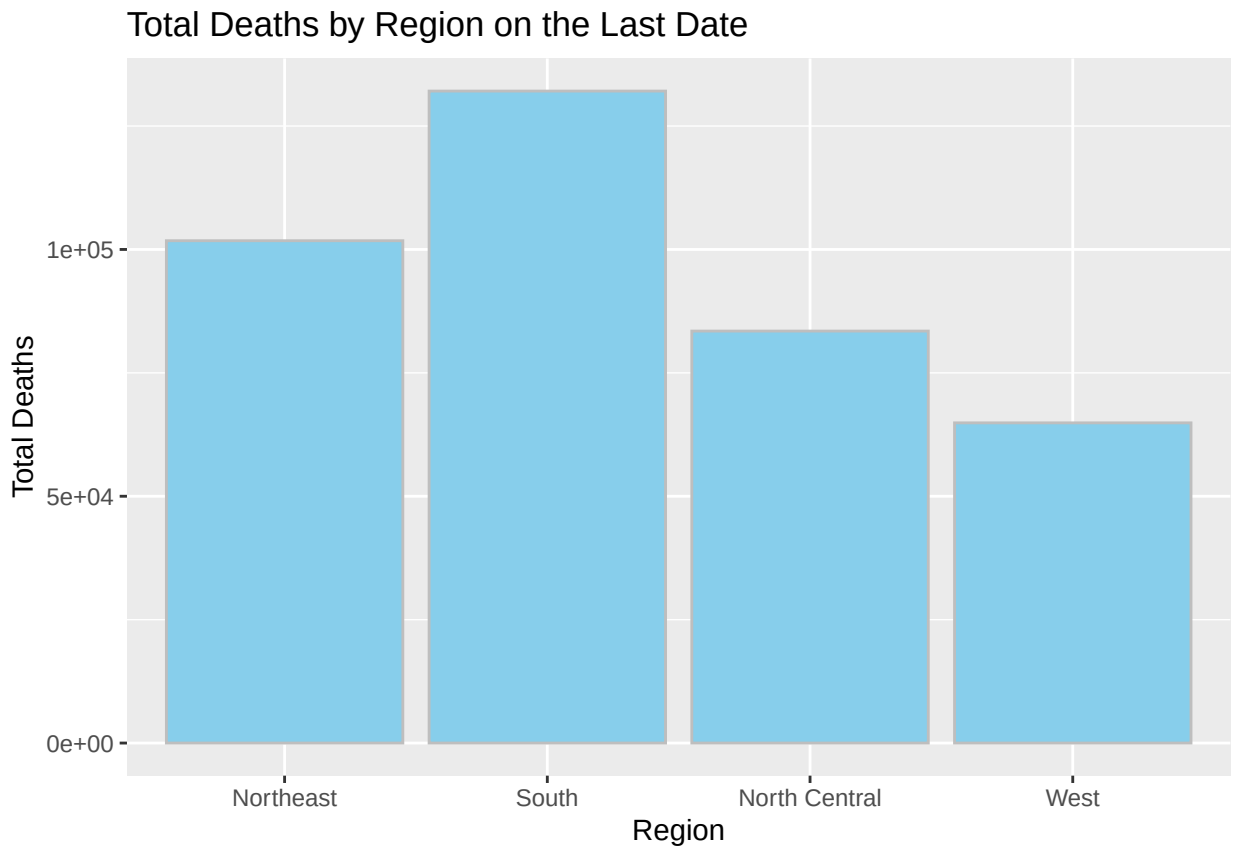



- As we can see, South region leads with 8811189 total cases but there isnt a significant difference between North Central and West.

```
# Summarize cases by region for the last date
all_region_deaths <- latest_data %>%
  group_by(region) %>%
  summarize(`Total Deaths` = sum(deaths)) %>%
  arrange(desc(`Total Deaths`))

# Now you have the Region with the highest number of cases on the last date

all_region_deaths
```

```
## # A tibble: 4 x 2
##   region        `Total Deaths`
##   <fct>                  <dbl>
## 1 South                 132081
## 2 Northeast             101762
## 3 North Central          83460
## 4 West                   64868
```

```r
#Lets plot this data into Bar Plot

ggplot(all_region_deaths, aes(x = region, y = `Total Deaths`)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "grey") +
  labs(title = "Total Deaths by Region on the Last Date", x = "Region", y = "Total Deaths")
```

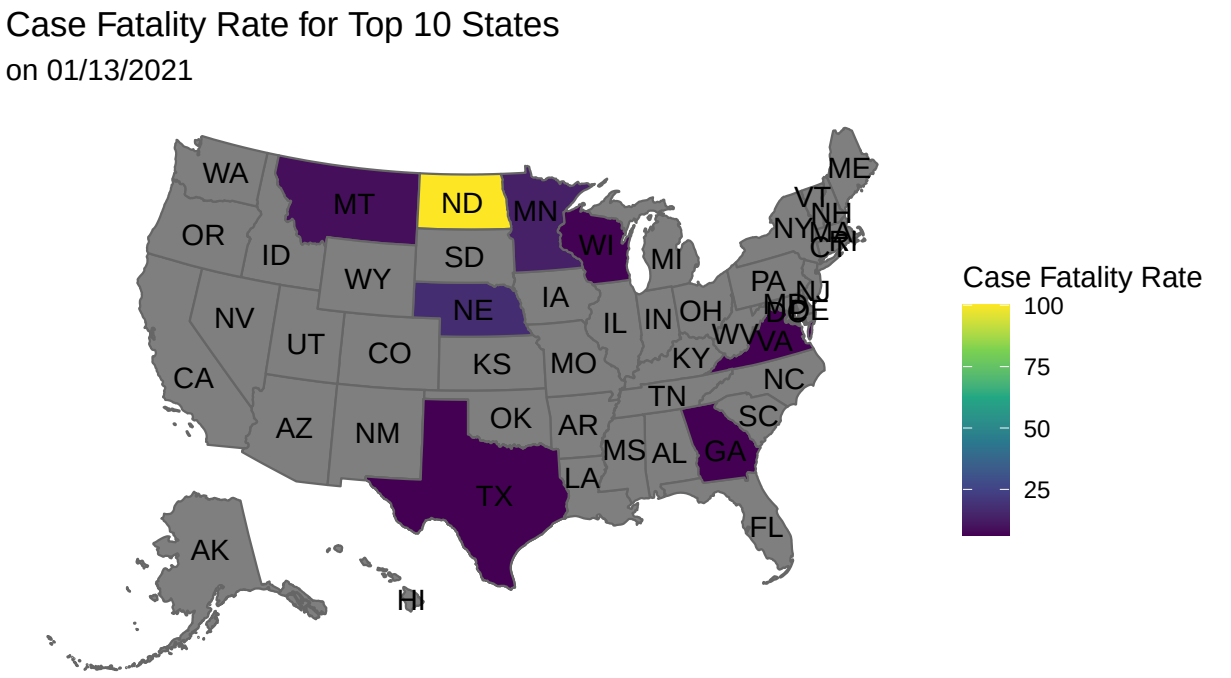



- As we can see, south region leads again with 132081 Deaths followed by Northeast.
- Next, we will calculate Case Fatality rate(Deaths per cases in %). We are also aware of the fact that "deaths" variable has missing values. To ease our calculation, we will turn these values to zero.
- We create a new variable "Case FR" in %

```r
## Assigning 0 to all missing value of variable "Death"
latest_data[is.na(latest_data)] = 0

## Creating a variable "`Case FR`".
latest_data$`Case FR` <- latest_data$deaths/latest_data$cases * 100

## Removing Infinite values
latest_data$`Case FR`[!is.finite(latest_data$`Case FR`)] <- 0

Highest_CFR_tp <- latest_data %>% arrange(desc(`Case FR`)) %>%
  head(10)
Highest_CFR_tp
```

```
## # A tibble: 10 x 7
##    date       county       state        cases deaths region        `Case FR`
##    <date>     <chr>        <chr>        <dbl>  <dbl> <fct>             <dbl>
##  1 2021-01-13 Unknown      North Dakota     7      7 North Central    100
##  2 2021-01-13 Grant        Nebraska        23      4 North Central     17.4
##  3 2021-01-13 Unknown      Minnesota      497     68 North Central     13.7
##  4 2021-01-13 Sherman      Texas          116     11 South              9.48
##  5 2021-01-13 Petroleum    Montana         11      1 West               9.09
##  6 2021-01-13 Kenedy       Texas           29      2 South              6.90
##  7 2021-01-13 Iron         Wisconsin      546     36 North Central      6.59
##  8 2021-01-13 Randolph     Georgia        474     31 South              6.54
##  9 2021-01-13 Knox         Texas          184     12 South              6.52
## 10 2021-01-13 Emporia city Virginia       463     30 South              6.48
```

```
# Lets plot this data in us map

usmap::plot_usmap(data = `Highest_CFR_tp`,
                  values = "Case FR",
color = "grey40", labels = TRUE) + #change the gradient type, add comma on legend
  scale_fill_continuous(type='viridis', label = scales:: comma) +
#add title, subtitle, caption, and legend title
  labs(title = "Case Fatality Rate for Top 10 States",
       subtitle = "on 01/13/2021",
       fill = "Case Fatality Rate") +
   ## removing all the axes from the plotting.
       theme_classic()+
  theme(panel.background = element_blank(),
        panel.border = element_blank(),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        legend.position = "right",
        axis.line = element_blank(),
        axis.ticks = element_blank(),
        axis.text.x =element_blank(),
        axis.text.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())
```




```
CFR_region <- latest_data %>%
  group_by(region) %>%  # Group by region before calculating the mean
  summarize(avg_Case_FR = mean(`Case FR`)) %>%  # Use summarize() to calculate mean
  arrange(desc(avg_Case_FR))
CFR_region
```

```
## # A tibble: 4 x 2
##   region        avg_Case_FR
##   <fct>               <dbl>
## 1 Northeast            2.16
## 2 South                1.84
## 3 North Central        1.76
## 4 West                 1.32
```

***Statistical Analysis:***

- Let us find the correlation between Total Cases, Total Deaths and mean CFR in regions. This analysis can provide insights into how these variables are associated and whether there are patterns or relationships between them.

```
all_region_cases
```

```
## # A tibble: 4 x 2
##   region        `Total Cases`
```

```
##   <fct>                <dbl>
## 1 South               8811189
## 2 North Central       5390115
## 3 West                5375633
## 4 Northeast           3395548
```

```
all_region_deaths
```

```
## # A tibble: 4 x 2
##   region        `Total Deaths`
##   <fct>                  <dbl>
## 1 South                 132081
## 2 Northeast             101762
## 3 North Central          83460
## 4 West                   64868
```

```
CFR_region
```

```
## # A tibble: 4 x 2
##   region        avg_Case_FR
##   <fct>               <dbl>
## 1 Northeast            2.16
## 2 South                1.84
## 3 North Central        1.76
## 4 West                 1.32
```

```
# Merge the data based on the 'region' column

merged_data <- merge(merge(CFR_region, all_region_cases,
by = "region"),
all_region_deaths,
by = "region")

colnames(merged_data)
```

```
## [1] "region"       "avg_Case_FR"  "Total Cases"  "Total Deaths"
```

```
correlation_matrix <- cor(merged_data[, c("avg_Case_FR", "Total Cases", "Total Deaths")])

correlation_matrix
```

```
##              avg_Case_FR Total Cases Total Deaths
## avg_Case_FR    1.0000000  -0.2312283    0.6324547
## Total Cases   -0.2312283   1.0000000    0.5848101
## Total Deaths   0.6324547   0.5848101    1.0000000
```

***Correlation between avg_Case_FR and Total Cases***:

- The correlation coefficient is approximately -0.23. There is a weak negative correlation between the average case fatality rate (avg_Case_FR) and total cases suggesting that regions with higher total cases tend to have a slightly lower average case fatality rate, but the relationship is not very strong.

***Correlation between avg_Case_FR and Total Deaths***:

- The correlation coefficient is approximately 0.64. There is a moderate to strong positive correlation between the average case fatality rate (avg_Case_FR) and total deaths implying that regions with higher average case fatality rates tend to have higher total deaths.

***Correlation between Total Cases and Total Deaths:***

- The correlation coefficient is approximately 0.58. Interpretation: There is a moderate to strong positive correlation between total cases and total deaths suggesting that regions with higher total cases tend to have higher total deaths.

***Summary :***

- "Regional Contrasts in the Face of Pandemic: A Thorough Exploration of COVID-19 Impact Across U.S. Regions" reveals intriguing patterns in the data. The South region emerges as a focal point, leading both in total COVID-19 cases and deaths, signaling potential challenges in managing the outbreak. However, the nuances become apparent when examining the case fatality rate, where North Dakota in the North Central region takes the lead, indicating regional variations in healthcare and response effectiveness.
- These findings prompt a deeper exploration into the distinct strategies and policies implemented by different regions' governments in navigating the pandemic. The weakly negative relationship between the average case fatality rate and total cases suggests a complex interplay of factors influencing outcomes. Meanwhile, the moderately to strongly positive relationships between the average case fatality rate and total deaths, as well as between total cases and total deaths, underscore the need for nuanced regional analyses.
- This multi-dimensional approach provides a comprehensive understanding of the COVID-19 impact, offering valuable insights for policymakers, public health officials, and researchers. It emphasizes the importance of considering diverse factors when evaluating the effectiveness of regional responses to the ongoing pandemic.

***Limitations :***

- This exploration provides valuable insights, but it has limitations. These include potential data inaccuracies, current data, lack of consideration for population dynamics, a fixed time frame, variability in policies and reporting, and regional heterogeneity. Acknowledging these limitations is important for a nuanced interpretation and highlights areas for further research.

***References :***

1. Ritchie H, Mathieu E, Rodés-Guirao L, Appel C, Giattino C, Ortiz-Ospina E, et al. (2020–2022). "Coronavirus Pandemic (COVID-19)". *Our World in Data*. Retrieved 19 January 2024.
2. https://www.who.int/healthtopics/coronavirus#tab=tab_1.