Analysis on Johns Hopkins Covid19 Data

Charles

2024-02-06

The Johns Hopkins COVID-19 dataset offers and extensive overview of global COVID-19 cases and fatalities. The is exploratory analysis aims to uncover regional disparities in COVID-19 cases within the United States. By employing logistic regressions, I will model the relationship between geographical regions and the prevalence of COVID-19 cases, seeking to understand how regional factors contributed to the spread of the virus. This approach not only highlights the variability across regions but also provides insights into the dynamics of COVID-19 transmissions.

Load Libraries

```
library(tidyverse)
library(ggplot2)
library(lubridate)
library(dplyr)
```

Import Data

```
url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
file_names <- c("time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_global.csv",
                "time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")
urls <- str_c(url,file_names)

global_cases <- read_csv(urls[1])
global_deaths <- read_csv(urls[2])
us_cases <- read_csv(urls[3])
us_deaths <- read_csv(urls[4])
us_cases
```

```
## # A tibble: 3,342 x 1,154
##         UID iso2  iso3  code3  FIPS Admin2   Provi~1 Count~2   Lat Long_ Combi~3
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>    <chr>   <chr>   <dbl> <dbl> <chr>
##  1 84001001 US    USA     840  1001 Autauga  Alabama US       32.5 -86.6 Autaug~
##  2 84001003 US    USA     840  1003 Baldwin  Alabama US       30.7 -87.7 Baldwi~
##  3 84001005 US    USA     840  1005 Barbour  Alabama US       31.9 -85.4 Barbou~
##  4 84001007 US    USA     840  1007 Bibb     Alabama US       33.0 -87.1 Bibb, ~
##  5 84001009 US    USA     840  1009 Blount   Alabama US       34.0 -86.6 Blount~
##  6 84001011 US    USA     840  1011 Bullock  Alabama US       32.1 -85.7 Bulloc~
##  7 84001013 US    USA     840  1013 Butler   Alabama US       31.8 -86.7 Butler~
##  8 84001015 US    USA     840  1015 Calhoun  Alabama US       33.8 -85.8 Calhou~
##  9 84001017 US    USA     840  1017 Chambers Alabama US       32.9 -85.4 Chambe~
```

```
## 10 84001019 US    USA     840  1019 Cherokee Alabama US       34.2 -85.6 Cherok~
## # ... with 3,332 more rows, 1,143 more variables: `1/22/20` <dbl>,
## #   `1/23/20` <dbl>, `1/24/20` <dbl>, `1/25/20` <dbl>, `1/26/20` <dbl>,
## #   `1/27/20` <dbl>, `1/28/20` <dbl>, `1/29/20` <dbl>, `1/30/20` <dbl>,
## #   `1/31/20` <dbl>, `2/1/20` <dbl>, `2/2/20` <dbl>, `2/3/20` <dbl>,
## #   `2/4/20` <dbl>, `2/5/20` <dbl>, `2/6/20` <dbl>, `2/7/20` <dbl>,
## #   `2/8/20` <dbl>, `2/9/20` <dbl>, `2/10/20` <dbl>, `2/11/20` <dbl>,
## #   `2/12/20` <dbl>, `2/13/20` <dbl>, `2/14/20` <dbl>, `2/15/20` <dbl>, ...
```

The data set is well-structured but requires additional tidying. First, we need to reshape the date columns using pivot_longer() to help us analyze and visualize the data. Next, we need to deal with our "NA" values. We can also remove or deselect Lat/Long since we won't be including it in our analysis.

Our data set consists of 1154 columns, 1144 of which are date columns. We need to use the pivot_longer() function to reshape our data and make it easier to analyze and visualize.

```
us_cases1 <- us_cases %>%
  pivot_longer(cols = -c(UID,iso2,iso3,code3,FIPS,Admin2,Province_State,Country_Region,Lat,Long_,Combin
               names_to = "date",
               values_to = "cases") %>%
  dplyr::select(-Lat,-Long_)
us_cases1
```

```
## # A tibble: 3,819,906 x 11
##         UID iso2  iso3  code3  FIPS Admin2  Provin~1 Count~2 Combi~3 date  cases
##       <dbl> <chr> <chr> <dbl> <dbl> <chr>   <chr>    <chr>   <chr>   <chr> <dbl>
##  1 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/22~     0
##  2 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/23~     0
##  3 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/24~     0
##  4 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/25~     0
##  5 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/26~     0
##  6 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/27~     0
##  7 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/28~     0
##  8 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/29~     0
##  9 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/30~     0
## 10 84001001 US    USA     840  1001 Autauga Alabama  US      Autaug~ 1/31~     0
## # ... with 3,819,896 more rows, and abbreviated variable names
## #   1: Province_State, 2: Country_Region, 3: Combined_Key
```

Ok, looks like we can reduce the number of columns by deselecting superfluous information. Let's do it!

```
us_cases2 <- us_cases1 %>%
  dplyr::select(-UID,-iso2,-iso3,-code3,-FIPS)
us_cases2
```

```
## # A tibble: 3,819,906 x 6
##    Admin2  Province_State Country_Region Combined_Key         date    cases
##    <chr>   <chr>          <chr>          <chr>                <chr>   <dbl>
##  1 Autauga Alabama        US             Autauga, Alabama, US 1/22/20     0
##  2 Autauga Alabama        US             Autauga, Alabama, US 1/23/20     0
##  3 Autauga Alabama        US             Autauga, Alabama, US 1/24/20     0
##  4 Autauga Alabama        US             Autauga, Alabama, US 1/25/20     0
##  5 Autauga Alabama        US             Autauga, Alabama, US 1/26/20     0
```

```
##  6 Autauga Alabama        US             Autauga, Alabama, US 1/27/20     0
##  7 Autauga Alabama        US             Autauga, Alabama, US 1/28/20     0
##  8 Autauga Alabama        US             Autauga, Alabama, US 1/29/20     0
##  9 Autauga Alabama        US             Autauga, Alabama, US 1/30/20     0
## 10 Autauga Alabama        US             Autauga, Alabama, US 1/31/20     0
## # ... with 3,819,896 more rows
```

Much better, now we need to change the date column from character to date type.

```
us_cases3 <- us_cases2 %>%
  mutate(date = mdy(date))
us_cases3
```

```
## # A tibble: 3,819,906 x 6
##    Admin2  Province_State Country_Region Combined_Key         date       cases
##    <chr>   <chr>          <chr>          <chr>                <date>     <dbl>
##  1 Autauga Alabama        US             Autauga, Alabama, US 2020-01-22     0
##  2 Autauga Alabama        US             Autauga, Alabama, US 2020-01-23     0
##  3 Autauga Alabama        US             Autauga, Alabama, US 2020-01-24     0
##  4 Autauga Alabama        US             Autauga, Alabama, US 2020-01-25     0
##  5 Autauga Alabama        US             Autauga, Alabama, US 2020-01-26     0
##  6 Autauga Alabama        US             Autauga, Alabama, US 2020-01-27     0
##  7 Autauga Alabama        US             Autauga, Alabama, US 2020-01-28     0
##  8 Autauga Alabama        US             Autauga, Alabama, US 2020-01-29     0
##  9 Autauga Alabama        US             Autauga, Alabama, US 2020-01-30     0
## 10 Autauga Alabama        US             Autauga, Alabama, US 2020-01-31     0
## # ... with 3,819,896 more rows
```

Next, we tidy our us_deaths data frame.

```
us_deaths1 <- us_deaths %>%
  pivot_longer(cols = -c(UID:Population),
               names_to = "date",
               values_to = "deaths") %>%
  dplyr::select(Admin2:deaths) %>%
  mutate(date=mdy(date)) %>%
  dplyr::select(-Lat,-Long_)
us_deaths1
```

```
## # A tibble: 3,819,906 x 7
##    Admin2  Province_State Country_Region Combined_Key  Popul~1 date       deaths
##    <chr>   <chr>          <chr>          <chr>           <dbl> <date>      <dbl>
##  1 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-22      0
##  2 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-23      0
##  3 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-24      0
##  4 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-25      0
##  5 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-26      0
##  6 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-27      0
##  7 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-28      0
##  8 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-29      0
##  9 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-30      0
## 10 Autauga Alabama        US             Autauga, Ala~   55869 2020-01-31      0
## # ... with 3,819,896 more rows, and abbreviated variable name 1: Population
```

To ensure comprehensive analysis, we must merge the us_deaths and us_cases data frames, which are similarly structured. The key distinction is the presence of a population column in us_deaths that is absent in us_cases. Merging these data frames will allow us to consolidate all relevant columns for analysis.

```
us <- us_cases3 %>%
  full_join(us_deaths1)
us
```

```
## # A tibble: 3,819,906 x 8
##    Admin2  Province_State Country_Region Combi~1 date       cases Popul~2 deaths
##    <chr>   <chr>          <chr>          <chr>   <date>     <dbl>   <dbl>  <dbl>
##  1 Autauga Alabama        US             Autaug~ 2020-01-22     0   55869      0
##  2 Autauga Alabama        US             Autaug~ 2020-01-23     0   55869      0
##  3 Autauga Alabama        US             Autaug~ 2020-01-24     0   55869      0
##  4 Autauga Alabama        US             Autaug~ 2020-01-25     0   55869      0
##  5 Autauga Alabama        US             Autaug~ 2020-01-26     0   55869      0
##  6 Autauga Alabama        US             Autaug~ 2020-01-27     0   55869      0
##  7 Autauga Alabama        US             Autaug~ 2020-01-28     0   55869      0
##  8 Autauga Alabama        US             Autaug~ 2020-01-29     0   55869      0
##  9 Autauga Alabama        US             Autaug~ 2020-01-30     0   55869      0
## 10 Autauga Alabama        US             Autaug~ 2020-01-31     0   55869      0
## # ... with 3,819,896 more rows, and abbreviated variable names 1: Combined_Key,
## #   2: Population
```

Now, we visualize the data

```
US_by_state <- us %>%
  group_by(Province_State,Country_Region,date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths *1000000 / Population) %>%
  dplyr::select(Province_State,Country_Region,date,
         cases,deaths,deaths_per_mill,Population) %>%
  ungroup()

US_by_state
```

```
## # A tibble: 66,294 x 7
##    Province_State Country_Region date       cases deaths deaths_per_mill Popul~1
##    <chr>          <chr>          <date>     <dbl>  <dbl>           <dbl>   <dbl>
##  1 Alabama        US             2020-01-22     0      0               0 4903185
##  2 Alabama        US             2020-01-23     0      0               0 4903185
##  3 Alabama        US             2020-01-24     0      0               0 4903185
##  4 Alabama        US             2020-01-25     0      0               0 4903185
##  5 Alabama        US             2020-01-26     0      0               0 4903185
##  6 Alabama        US             2020-01-27     0      0               0 4903185
##  7 Alabama        US             2020-01-28     0      0               0 4903185
##  8 Alabama        US             2020-01-29     0      0               0 4903185
##  9 Alabama        US             2020-01-30     0      0               0 4903185
## 10 Alabama        US             2020-01-31     0      0               0 4903185
## # ... with 66,284 more rows, and abbreviated variable name 1: Population
```

Prior to analyzing regional trends, it's essential to examine overarching US patterns.

```r
US_totals <-  US_by_state %>%
  group_by(Country_Region,date) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths *1000000 / Population) %>%
  dplyr::select(Country_Region,date, cases,deaths,
         deaths_per_mill,Population) %>%
  ungroup()
```

```r
US_totals %>%
  filter(cases > 0) %>%
  ggplot() +
  geom_line(aes(x = date, y = cases, color = "Cases")) +
  geom_line(aes(x = date, y = deaths, color = "Deaths")) +
  scale_y_log10() +
  scale_color_manual(values = c("Cases" = "blue", "Deaths" = "red")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "US COVID-19 Data",
       subtitle = "Cases and Deaths Over Time",
       y = "Count",
       color = "Metric")
```

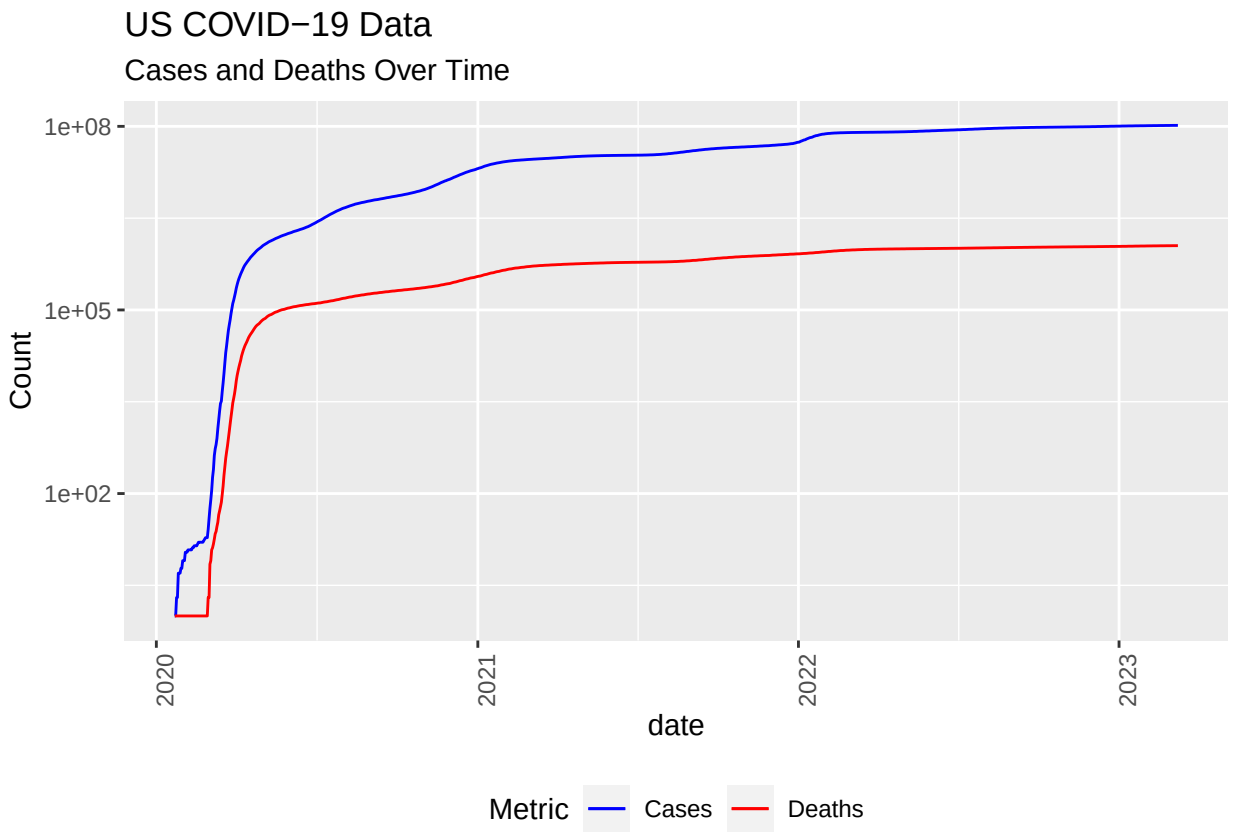


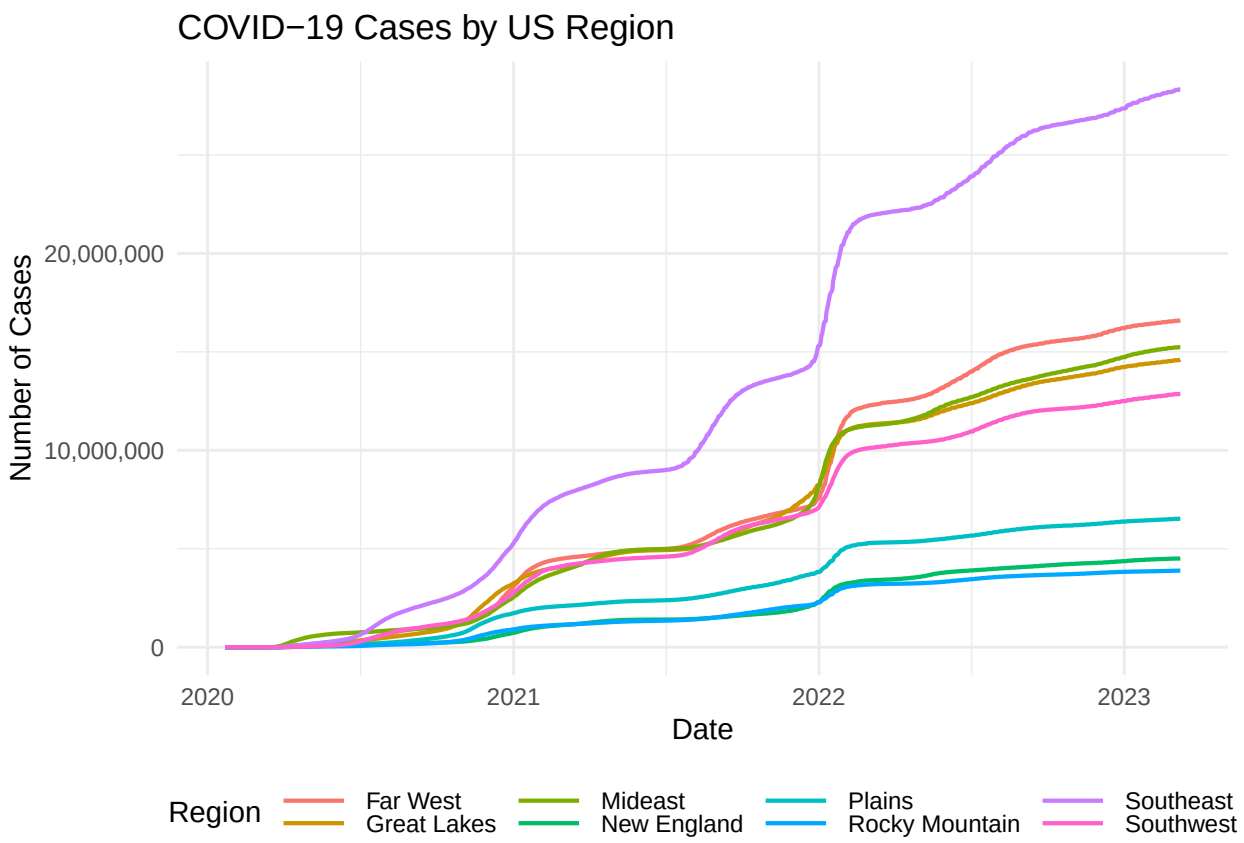
Figure 1 illustrates the dynamic trajectory of COVID-19 cases and deaths from 2020 to 2023. Initially, both metrics surged rapidly during the early stages of the pandemic in 2020. The following year, 2021, witnessed a more gradual increase in infections and fatalities, culminating in a stabilization of numbers by 2022.

In the following section, we introduce a new column titled "Region" to facilitate the analysis of COVID-19 cases and deaths by geographical region. This addition is crucial to understanding the spatial distribution of the pandemic's impact and enables a more nuanced exploration of regional trends and patterns.

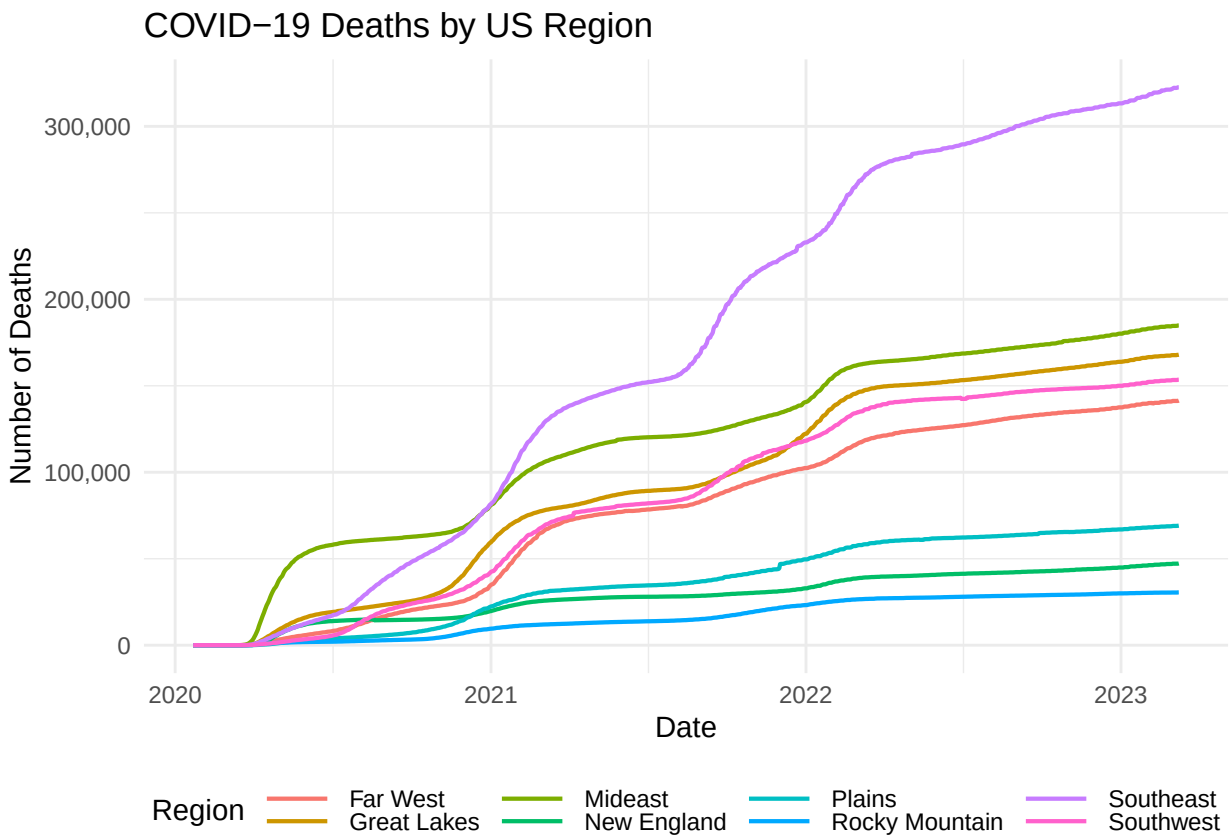
```
regions <- US_by_state %>%
  mutate(Region = case_when(
    Province_State %in% c("Alabama", "Arkansas", "Florida", "Georgia", "Kentucky", "Louisiana", "Mississ
    Province_State %in% c("Arizona", "New Mexico", "Oklahoma", "Texas") ~ "Southwest",
    Province_State %in% c("Alaska", "California", "Hawaii", "Nevada", "Oregon", "Washington") ~ "Far We
    Province_State %in% c("Illinois", "Indiana", "Michigan", "Ohio", "Wisconsin") ~ "Great Lakes",
    Province_State %in% c("Connecticut", "Maine", "Massachusetts", "New Hampshire", "Rhode Island", "Ver
    Province_State %in% c("Delaware", "District of Columbia", "Maryland", "New Jersey", "New York", "Pen
    Province_State %in% c("Iowa", "Kansas", "Minnesota", "Missouri", "Nebraska", "North Dakota", "South
    Province_State %in% c("Colorado", "Idaho", "Montana", "Utah", "Wyoming") ~ "Rocky Mountain"
  )
  )
unique(regions$Region)
```

```
## [1] "Southeast"      "Far West"       NA               "Southwest"
## [5] "Rocky Mountain" "New England"    "Mideast"        "Great Lakes"
## [9] "Plains"
```

```
library(scales)
regional_summary <- regions %>%
  group_by(Region,date) %>%
  filter(!is.na(Region)) %>%
  summarise(deaths = sum(deaths, na.rm = TRUE),
            cases = sum(cases, na.rm = TRUE),
            population = sum(Population,na.rm=TRUE))

regional_visualization <- regional_summary %>%
  ggplot(aes(x = date, y= cases, color = Region)) +
  geom_line(size = 0.75) +
  scale_y_continuous(labels = label_comma()) +
  labs(title = "COVID-19 Cases by US Region",
       x = "Date",
       y = "Number of Cases",
       color = "Region") +
  theme_minimal() +
  theme(legend.position = "bottom",
        legend.key.width = unit(1,"cm"),
        legend.key.height = unit(.05,"cm"))

regional_visualization
```



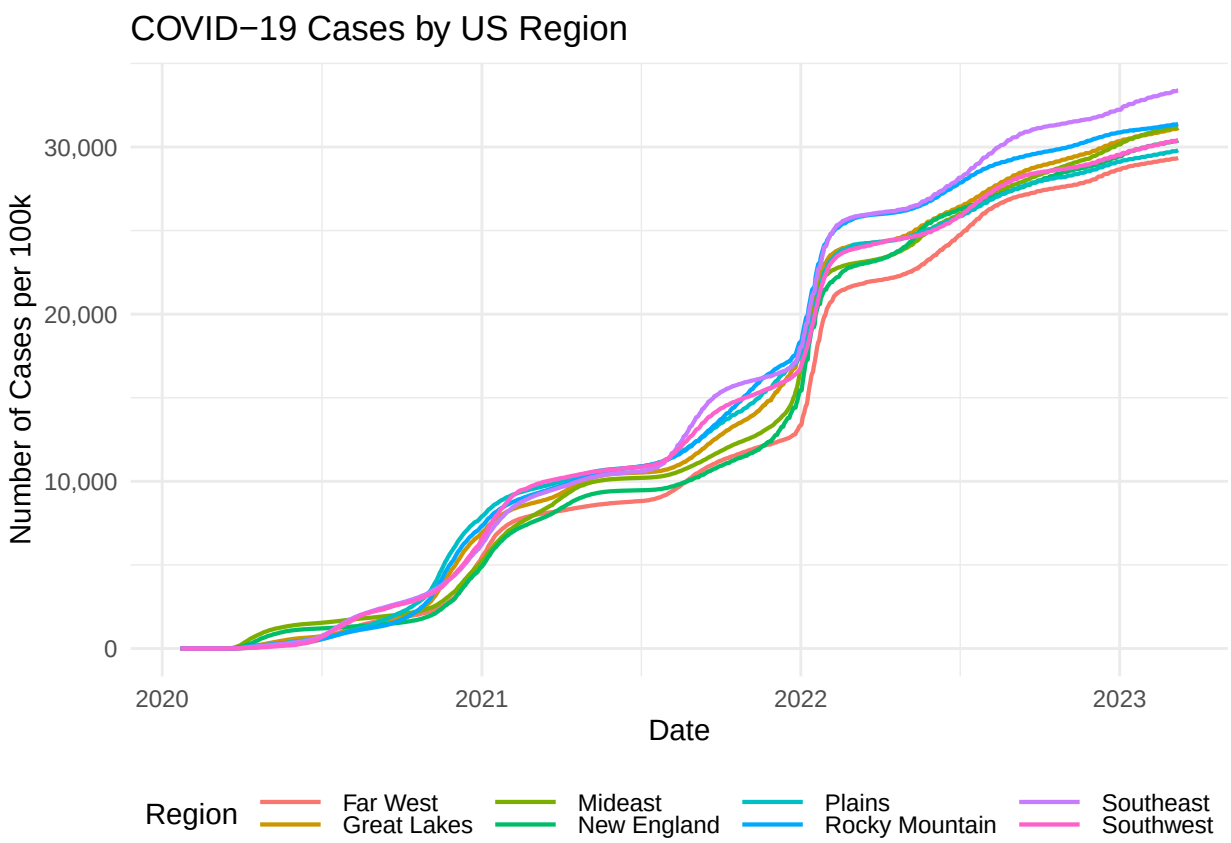

```
regional_visualization_2 <- regional_summary %>%
  ggplot(aes(x = date, y= deaths, color = Region)) +
  geom_line(size = 0.75) +
  scale_y_continuous(labels = label_comma()) +
  labs(title = "COVID-19 Deaths by US Region",
       x = "Date",
       y = "Number of Deaths",
       color = "Region") +
  theme_minimal() +
  theme(legend.position = "bottom",
        legend.key.width = unit(1,"cm"),
        legend.key.height = unit(.05,"cm"))

regional_visualization_2
```



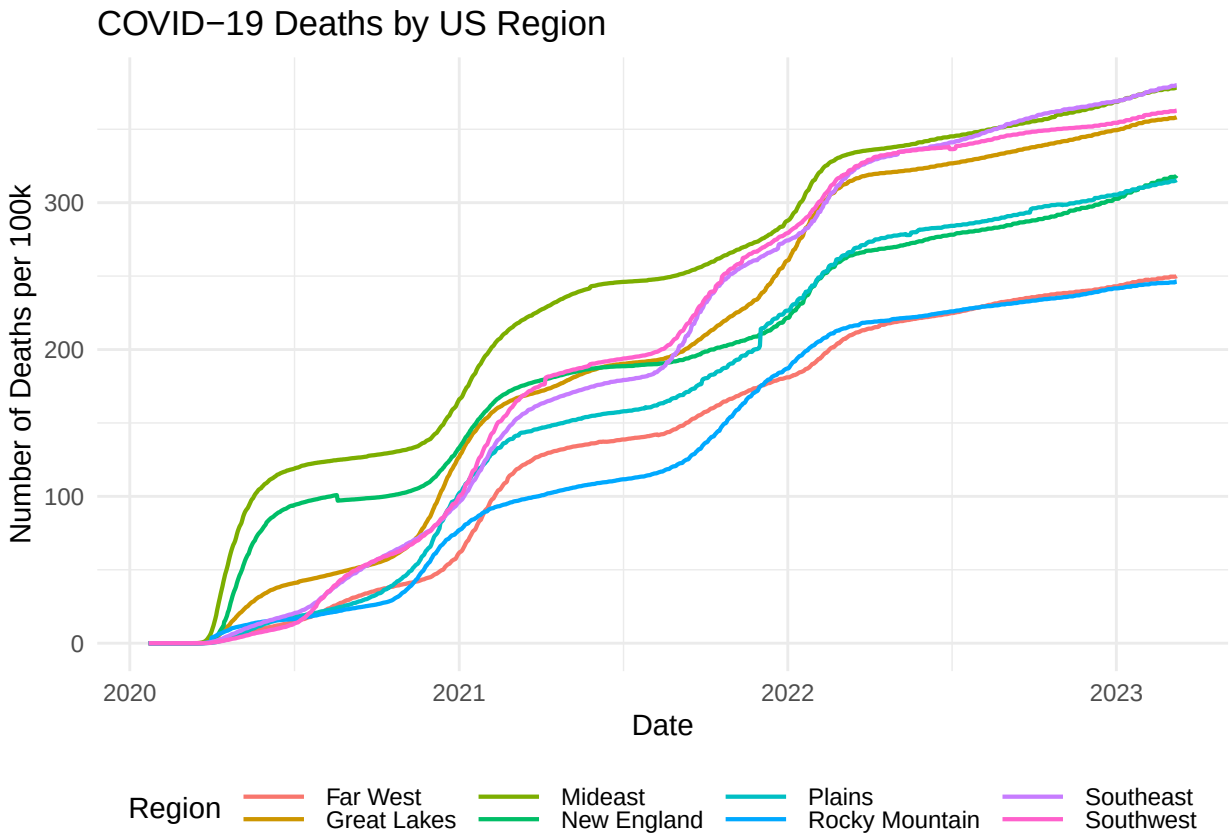
Figures 2 and 3 provide a macro overview of regional trends for COVID-19 cases and deaths. Given that our analysis is predisposed to bias towards larger population centers, it becomes imperative to normalize the data. This normalization will facilitate more equitable comparative analysis across regions, allowing for adjustments based on population size to ensure the accuracy and relevance of our findings.

```
regional_summary_normalized <- regional_summary %>%
  mutate(cases_per_100k = (cases / population) * 100000) %>%
  mutate(deaths_per_100k = (deaths / population) * 100000)

regional_normalized <- regional_summary_normalized %>%
  ggplot(aes(x = date, y = cases_per_100k, color = Region)) +
  geom_line(size = 0.75) +
  scale_y_continuous(labels = label_comma()) +
  labs(title = "COVID-19 Cases by US Region",
       x = "Date",
       y = "Number of Cases per 100k",
       color = "Region") +
  theme_minimal() +
  theme(legend.position = "bottom",
        legend.key.width = unit(1,"cm"),
        legend.key.height = unit(.05,"cm"))

regional_normalized_2 <- regional_summary_normalized %>%
  ggplot(aes(x = date, y = deaths_per_100k, color = Region)) +
  geom_line(size = 0.75) +
  scale_y_continuous(labels = label_comma()) +
  labs(title = "COVID-19 Deaths by US Region",
```

```
       x = "Date",
       y = "Number of Deaths per 100k",
       color = "Region") +
  theme_minimal() +
  theme(legend.position = "bottom",
        legend.key.width = unit(1,"cm"),
        legend.key.height = unit(.05,"cm"))

regional_normalized
```




```
regional_normalized_2
```



Our normalized data in figures 4 and 5 provide a more refined analysis of COVID cases and deaths by region, revealing patterns and convergence in cases per 100k, with notable trends in the Southeast region. To delve deeper into the statistical significance of our observations and understand the relationship between our predictor variables and the count of COVID cases, we employ a Generalized Linear Model (GLM). GLMs are versatile, allowing us to model different types of response variables. For our case, where the response variable is the count of COVID cases(count variable) we use Poisson regression. Poisson regression is apt for modeling count data, enabling us to explore how changes in predictor variables affect the rate of COVID cases. It calculates the expected log count of events (cases or deaths) given the predictors in the model, such as region or time. By analyzing the coefficients produced by this model, we can interpret the impact of each predictor, where the exponentiated coefficients give us rate ratios. This means we can quantify how the presence or change in a predictor variable influences the rate of COVID cases, adjusting for other factors in the model.

```
glm_cases <- glm(cases_per_100k ~ Region + date + offset(log(population)), data = regional_summary_norma
summary(glm_cases)
```

```
##
## Call:
## glm(formula = cases_per_100k ~ Region + date + offset(log(population)),
##     family = poisson(), data = regional_summary_normalized)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -75.847  -39.736    6.973   23.167   60.127
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)          -5.818e+01  6.219e-03 -9355.5   <2e-16 ***
## RegionGreat Lakes     2.876e-01  3.596e-04   799.9   <2e-16 ***
## RegionMideast         2.181e-01  3.620e-04   602.3   <2e-16 ***
## RegionNew England     1.378e+00  3.648e-04  3776.6   <2e-16 ***
## RegionPlains          1.048e+00  3.596e-04  2915.1   <2e-16 ***
## RegionRocky Mountain  1.660e+00  3.561e-04  4661.3   <2e-16 ***
## RegionSoutheast      -2.388e-01  3.540e-04  -674.5   <2e-16 ***
## RegionSouthwest       3.956e-01  3.591e-04  1101.6   <2e-16 ***
## date                  2.622e-03  3.252e-07  8063.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 139576804  on 9143  degrees of freedom
## Residual deviance:  13156751  on 9135  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(glm_cases))
```

```
##          (Intercept)    RegionGreat Lakes        RegionMideast
##         5.400954e-26         1.333259e+00         1.243662e+00
##    RegionNew England         RegionPlains RegionRocky Mountain
##         3.966120e+00         2.852600e+00         5.257783e+00
##      RegionSoutheast      RegionSouthwest                 date
##         7.875937e-01         1.485260e+00         1.002626e+00
```

Our analysis has identified statistically significant variations in COVID-19 case rates across different regions, as evidenced by p-values below the significance threshold of 0.001. This statistical significance is further elucidated through the analysis of the coefficients' magnitudes. When these coefficients are exponentiated, they reveal the relative differences in case rates per 100,000 individuals across regions in comparison to a designated baseline region. Specifically, the directionality of these coefficients (positive or negative) signifies whether the case rates are higher or lower relative to the baseline. Notably, the Rocky Mountain, New England, and Plains regions exhibit the most elevated exponentiated coefficients, highlighting a higher incidence rate. Conversely, the Southeast region, with an exponentiated coefficient of 0.78 and the sole negative coefficient, indicates a lower case rate per 100,000 individuals compared to the baseline.

To further substantiate our findings, we will extend our analysis through the application of the Negative Binomial model. This model, akin to the Poisson model, is particularly adept at accommodating over-dispersion, a scenario where the variance significantly surpasses the mean. The inclusion of an additional parameter in the Negative Binomial model addresses this excess variability, rendering it an adept and flexible tool for analyzing count data that may not conform to the Poisson model's assumptions.

```
library(MASS)

nb_model <- glm.nb(cases_per_100k ~ Region + date + offset(log(population)), data = regional_summary_no

coef(nb_model)
```

```
##          (Intercept)    RegionGreat Lakes        RegionMideast
##         -58.18574368           0.28764155           0.21806937
```

```
##    RegionNew England         RegionPlains RegionRocky Mountain
##           1.37778982           1.04825782           1.65972988
##      RegionSoutheast      RegionSouthwest                 date
##          -0.23875662           0.39561436           0.00262273
```

Excellent, the results from our Negative Binomial model are consistent with those obtained from the Poisson Model. Notably, the Rocky Mountain, New England, and Plains regions exhibit the highest coefficients, whereas the Southeast region is distinguished by its negative coefficient. This concordance between model outputs not only validates our analytic approach but also underscores the critical role of employing multiple statistical models for verification purposes. Such a methodology is particularly invaluable in complex scenarios, such as elucidating the dynamics of infectious disease spread, where accuracy and reliability of findings are paramount.

Sources of Potential Bias

1. The US COVID-19 data was biased towards larger population centers, potentially obscuring more detailed regional insights. To address this issue, I normalized the data, thereby facilitating a more nuanced comparative analysis across regions. This normalization process ensures that our findings account for population size variations, enabling a more equitable assessment of COVID-19's impacts.