final_report_danning_tian

Danning Tian (dt2354)

1/26/2022

Introduction

Since 2019, COVID-19 has taken many people's lives and largely affected people's regular life routine.

In this report, two questions are focused based on COVID-19 data from January 2020 to January 2022. The first question is *which state has highest prevalence rate and which state has lowest prevalence rate since the beginning of COVID-19 pandemic? Are there any differences when compared data in 2020 to 2021?* Such analysis is important because differences of severity in each state is an significant measurement of effectiveness of interventions; policies level and community level interventions launched in states with less severity may be applied to other states to reduce the severity of the spread of the disease.

The second question is *whether the relationship between cases and deaths stays constant from from Janurary 2020 to December 2021?* This is correlation may illustrate the effectiveness of intervention to prevent death, understand the nature of COVID-19 virus, and predict potential mutation of the COVID-19 virus from the differences.

This report would majorly focus on monthly incidence rate of COVID-19 cases in each US state to visualize the severity of such disease by month from January 2020 to January 2022 to address question 1. Also, potential correlation between number of cases and deaths would be illustrated by graphs to trace back the spread of the disease in the US.

Packages Required

In this report, following packages are needed:

```
library(tidyverse)
library(tidyr)
library(ggplot2)
library(ggrepel)
library(dplyr)
library(tibble)
library(readxl)
library(usmap)
```

"tidyverse", "tidyr" and "dplyr" are very powerful packages that make data analysis and date manipulation process easier.

"ggplot2" and "ggrepel" are powerful packages to make clear and easy-to-read graphs.

"tibble" package allows manipulate lists in a tibble form to get cleaner data

"readxl" package allows importing data from excel, in this report 2020 US population data is stored as excel.

"usmap" gives possibility to perform data visualization on US map

Data Preparation

Load data

The COVID-19 data used in this report is from: https://github.com/nytimes/covid-19-data

```
us_states <-  read.csv("us-states.csv")
us_data <- read.csv("us.csv")
# since we are not focus on county-level analysis, I did not import us_county.csv
```

These data sets are reported and collected by New York Times and updated daily, in an attempt to provide a complete record of the ongoing outbreak.

For the us_states data, there are four columns, namely date, state, fips, cases and deaths. For the us_data, there are three columns, namely date, case, and death.

Date indicated data collected data, state and fips indicated location, case shows cumulative cases of COVID-19 and death shows cumulative deaths caused by COVID-19. The range of the data is from 2020-01-21 to 2022-01-11, data are collected from a total of 56 states and regions.

The US population data in 2020 by state is from: https://www.census.gov/en.html

```
us_population_2020 <- read_excel("/Users/suerga/Desktop/final COVID-19 project/2020_us_population.xlsx")
```

The data set is reported by 2020 US census, the major purpose is showing regular population data of the country. Data are collected from a total of 51 states and regions. Since US census runs every 10 years, in this report, the population for COVID-19 data from 2020 to 2022 will based on the same population data collected in 2020.

check NA

```
# check na in all data sets
unique(is.na(us_states))
unique(is.na(us_data))
unique(is.na(us_population_2020))
```

Clean US_steta data

```
# for us_states data set
# make a copy of data
us_states_1 <- us_states

# since we focus on monthly reported data, we shall separate date to three columns, namely year, month,
us_states_1 <- separate(us_states_1, "date", c("year", "month", "day"), sep = "-")


# group data by year, month, and state, calculate cases per month and deaths per month
us_states_1 <- us_states_1 %>%
  group_by(year, month, state) %>%
  summarise(cases_month = sum(cases),
            deaths_month = sum(deaths))
```

```r
# combine population data with with COVID-19 states data
us_states_1 <- us_states_1 %>%
  right_join(us_population_2020, by = c("state" = "states"))

# rename 2020 to population
us_states_1 <- rename(us_states_1, "population" = "2020")

# using mutate function of calculate prevalence_rate_per_million, death_rate_per_million, and deaths_ov
us_states_1 <- us_states_1 %>%
  mutate(prevalence_rate_per_million = (cases_month / population) * 1000000,
         death_rate_per_million = (deaths_month / population) * 1000000,
         death_over_cases = deaths_month / cases_month)

# change data type from character to numeric for year and month
us_states_1$year <- as.numeric(us_states_1$year)
us_states_1$month <- as.numeric(us_states_1$month)

head(us_states_1)
```

```
## # A tibble: 6 x 9
## # Groups:   year, month [2]
##    year month state      cases_month deaths_month population prevalence_rate_pe~
##   <dbl> <dbl> <chr>            <int>        <int>      <dbl>                <dbl>
## 1  2020     1 Arizona              6            0    7151502                0.839
## 2  2020     1 California          14            0   39538223                0.354
## 3  2020     1 Illinois            10            0   12812508                0.780
## 4  2020     1 Washington          11            0    7705281                1.43
## 5  2020     2 Arizona             29            0    7151502                4.06
## 6  2020     2 California         285            0   39538223                7.21
## # ... with 2 more variables: death_rate_per_million <dbl>,
## #   death_over_cases <dbl>
```

Clean US data

```r
# for us_data data set
# make a copy of data
us_data_1 <- us_data

# calculate us_populationdata based on state population data
us_population <- sum(us_population_2020$`2020`)

# using mutate function of find prevalence_rate_per_million, death_rate_per_million, and deaths_over_ca
us_data_1 <- us_data_1 %>%
  mutate(prevalence_rate_per_million = (cases / us_population) * 1000000,
         death_rate_per_million = (deaths / us_population) * 1000000,
         deaths_over_cases = deaths/cases)

# remove cases and deaths columns by select function
us_data_2 <- us_data_1 %>%
  select(date, prevalence_rate_per_million, death_rate_per_million, deaths_over_cases)
```

```r
# for future plot, change data type of date from character to date
us_data_2$date <-  as.Date(us_data_2$date, format =  "%Y-%m-%d")

head(us_data_2)
```

```
##         date prevalence_rate_per_million death_rate_per_million
## 1 2020-01-21                0.002987436                       0
## 2 2020-01-22                0.002987436                       0
## 3 2020-01-23                0.002987436                       0
## 4 2020-01-24                0.005974873                       0
## 5 2020-01-25                0.008962309                       0
## 6 2020-01-26                0.014937182                       0
##   deaths_over_cases
## 1                 0
## 2                 0
## 3                 0
## 4                 0
## 5                 0
## 6                 0
```

```r
# check na in all data sets
unique(is.na(us_states_1))
unique(is.na(us_data_1))
```

Exploratory Data Analysis

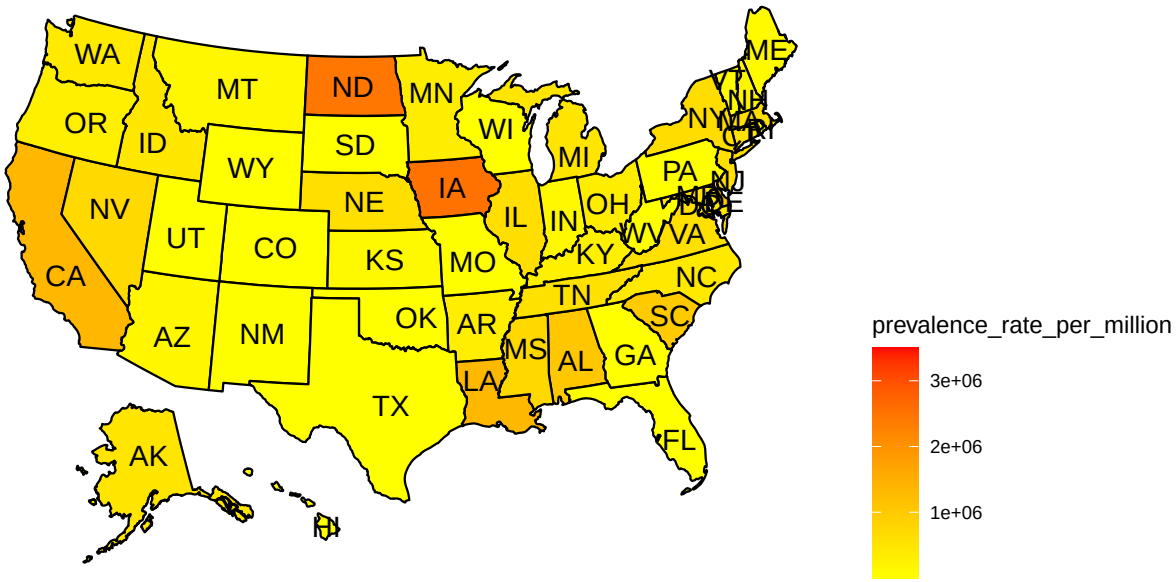
Question 1

In the first part of data exploratory data analysis, we focused on prevalence rate and death rate of the each states from January 2020 to December 2021 to address the following questions:

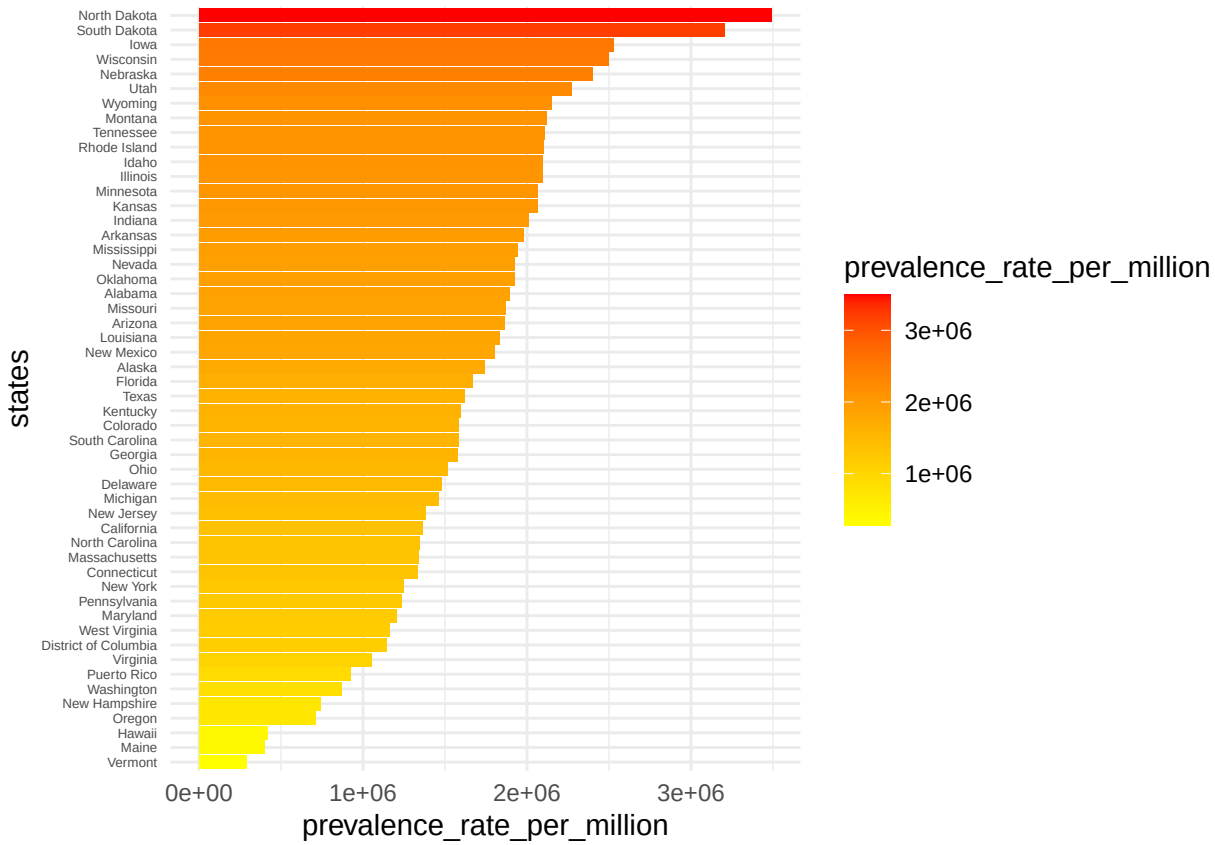
which state has highest prevalence rate and which state has lowest prevalence rate since the beginning of COVID-19 pandemic? Are there any differences when compared data in 2020 to 2021?

```r
# create a new data set with 2020 data
us_states_2020 <- us_states_1 %>%
  filter(year == "2020")

# create a new data set with 2021 data
us_states_2021<- us_states_1 %>%
  filter(year == "2021")


# 2020 prevalence data analysis
plot_usmap(data=us_states_2020, values = "prevalence_rate_per_million", labels = TRUE) +
  scale_fill_gradient(name = "prevalence_rate_per_million",
                      low = "yellow", high = "red") +
  theme(legend.position = "right") +
  ggtitle("prevalence rate per million in 2020")
```

prevalence rate per million in 2020

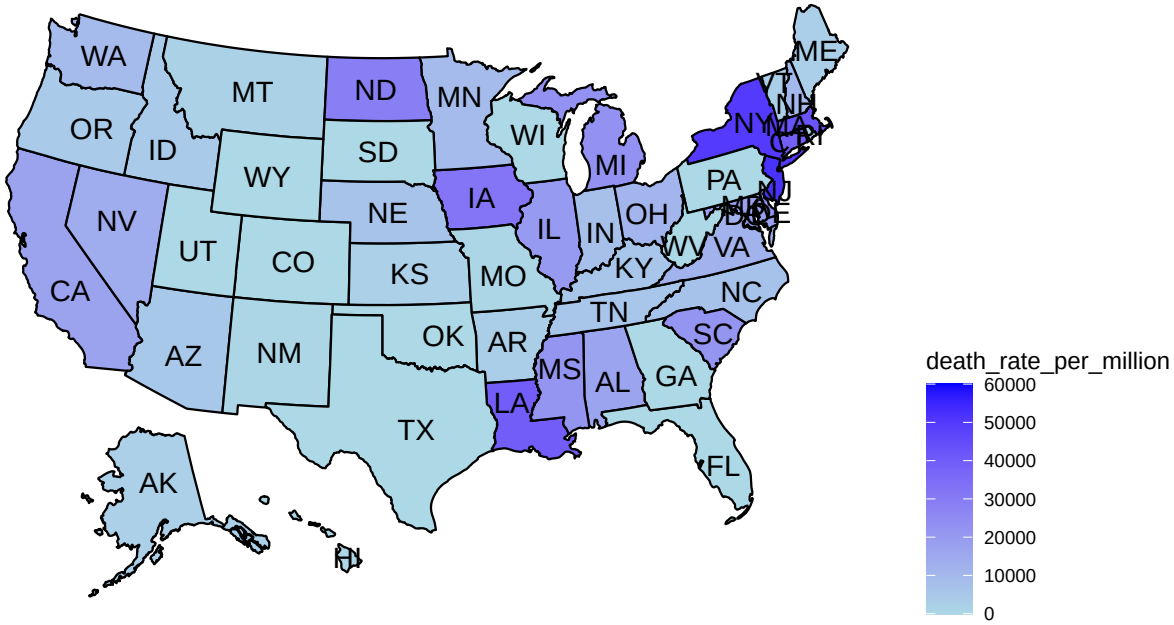



```
us_states_2020 %>%
  filter(month == "12") %>%
  ggplot(aes(x = reorder(state, prevalence_rate_per_million),
             y = prevalence_rate_per_million)) +
  geom_bar(aes(fill = prevalence_rate_per_million),
           stat = "identity")+
  scale_fill_gradient(name = "prevalence_rate_per_million",
                      low = "yellow", high = "red")+
  labs(x = "states") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 5)) +
  coord_flip()
```

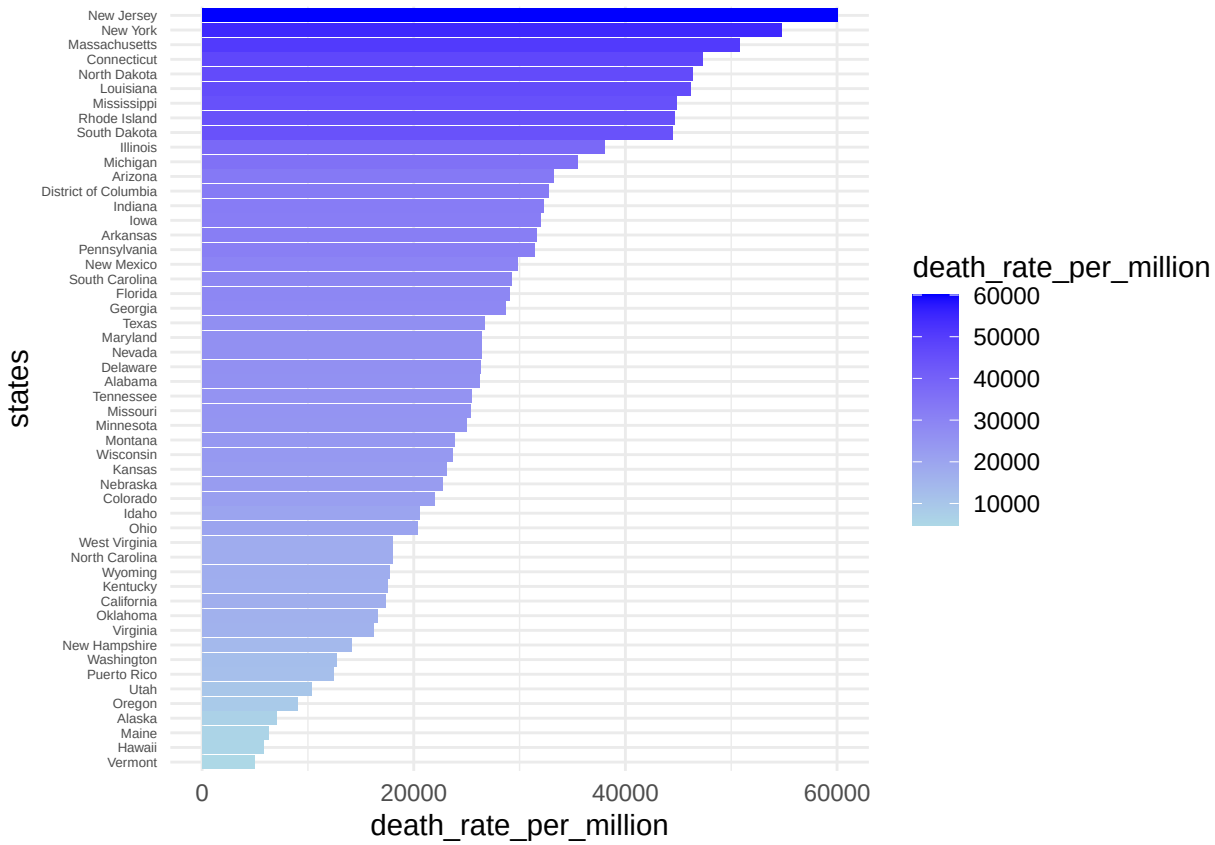



```
# 2020 death rate data analysis
plot_usmap(data=us_states_2020, values = "death_rate_per_million", labels = TRUE) +
  scale_fill_gradient(name = "death_rate_per_million",
                      low = "light blue", high = "blue") +
  theme(legend.position = "right") +
  ggtitle("death rate per million in 2020")
```

death rate per million in 2020

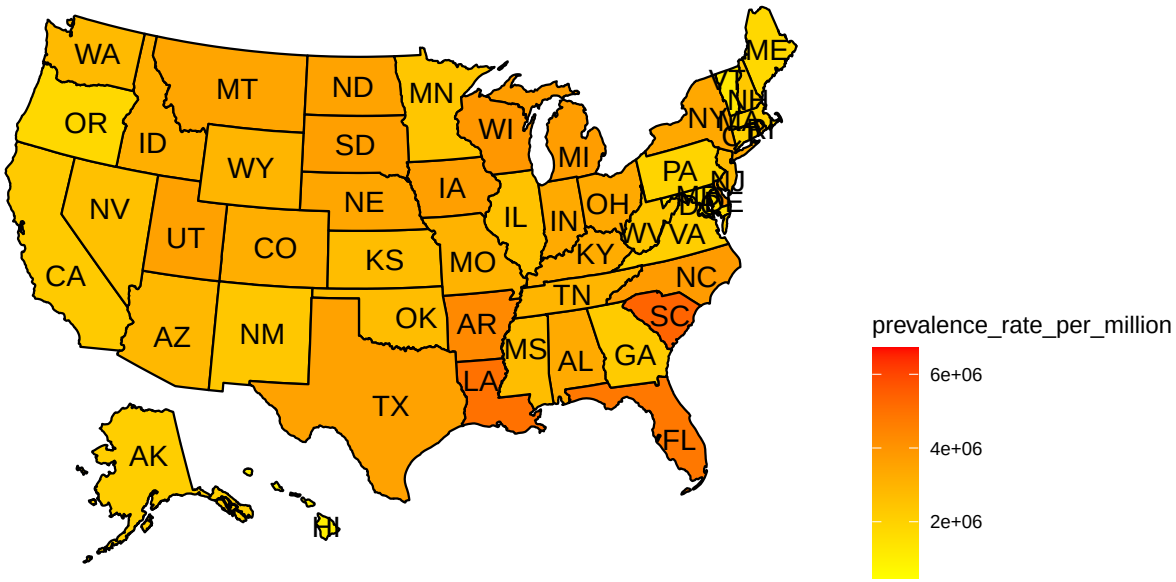



```
us_states_2020 %>%
  filter(month == "12") %>%
  ggplot(aes(x = reorder(state, death_rate_per_million),
             y = death_rate_per_million)) +
  geom_bar(aes(fill = death_rate_per_million),
           stat = "identity")+
  scale_fill_gradient(name = "death_rate_per_million",
                      low = "light blue", high = "blue") +
  labs(x = "states") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 5)) +
  coord_flip()
```

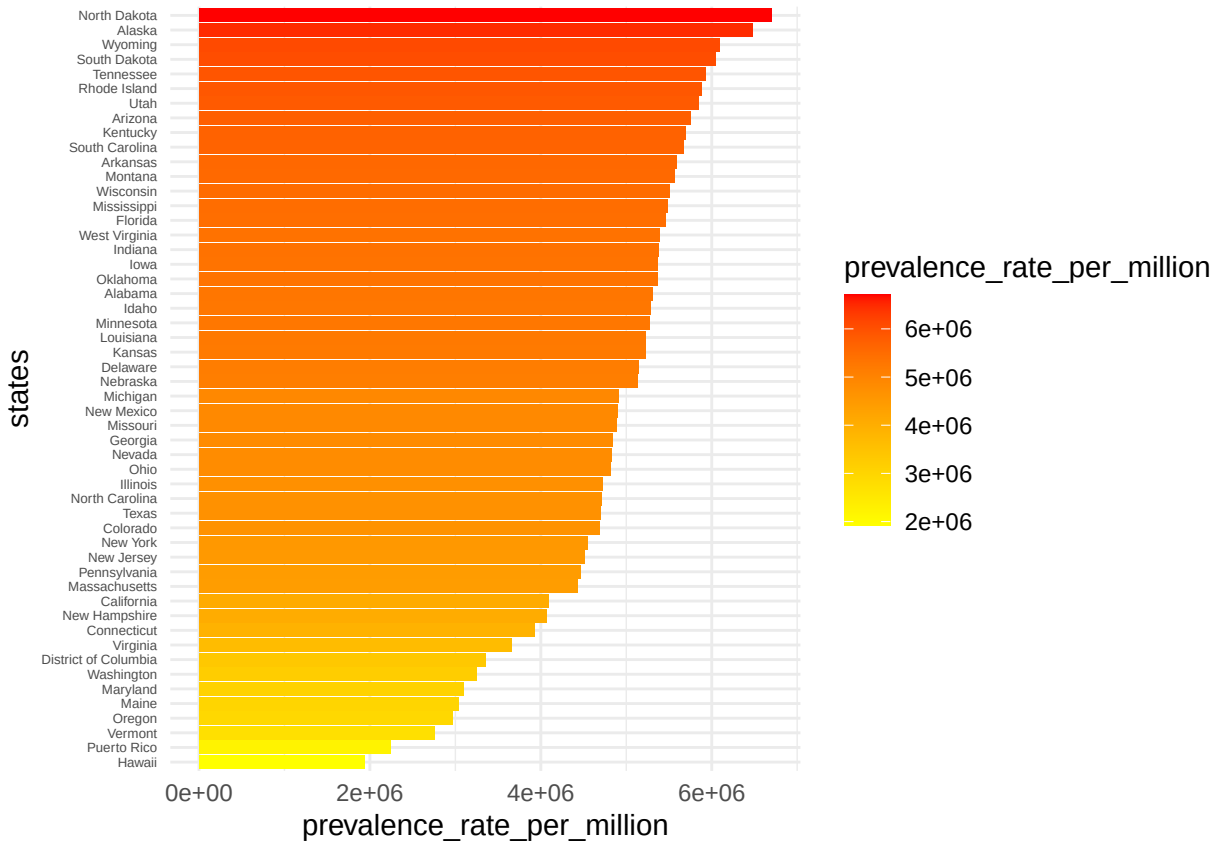



```
# 2021 prevalence data analysis
plot_usmap(data=us_states_2021, values = "prevalence_rate_per_million", labels = TRUE) +
  scale_fill_gradient(name = "prevalence_rate_per_million",
                      low = "yellow", high = "red") +
  theme(legend.position = "right") +
  ggtitle("prevalence rate per million in 2021")
```

prevalence rate per million in 2021

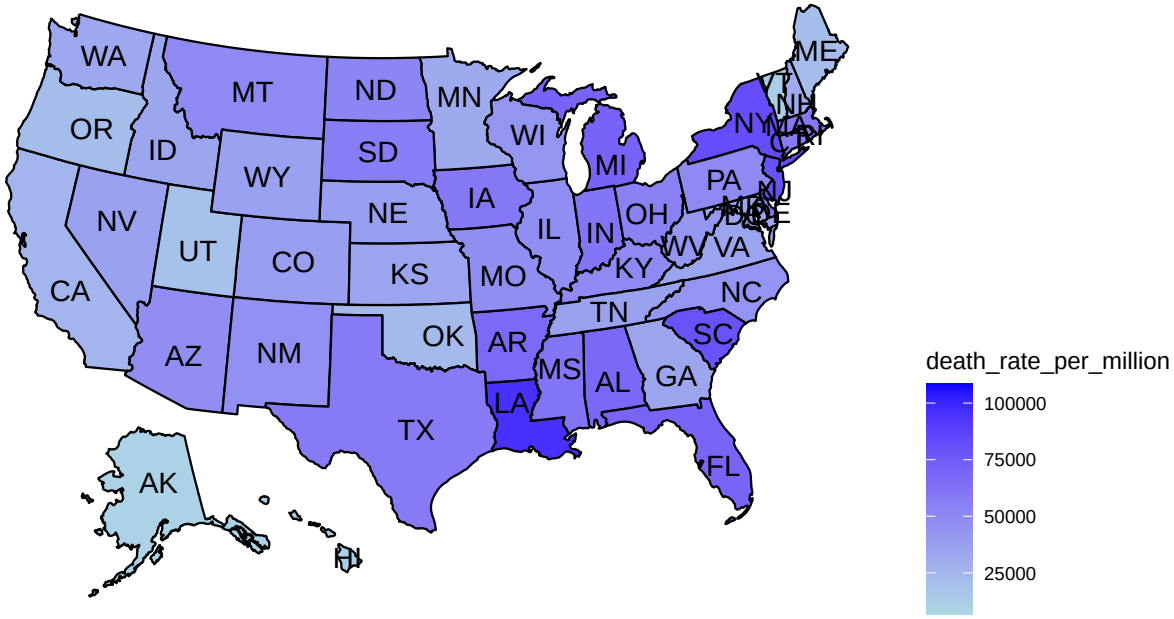



```
us_states_2021 %>%
  filter(month == "12") %>%
  ggplot(aes(x = reorder(state, prevalence_rate_per_million),
             y = prevalence_rate_per_million)) +
  geom_bar(aes(fill = prevalence_rate_per_million),
           stat = "identity")+
  scale_fill_gradient(name = "prevalence_rate_per_million",
                      low = "yellow", high = "red")+
  labs(x = "states") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 5)) +
  coord_flip()
```

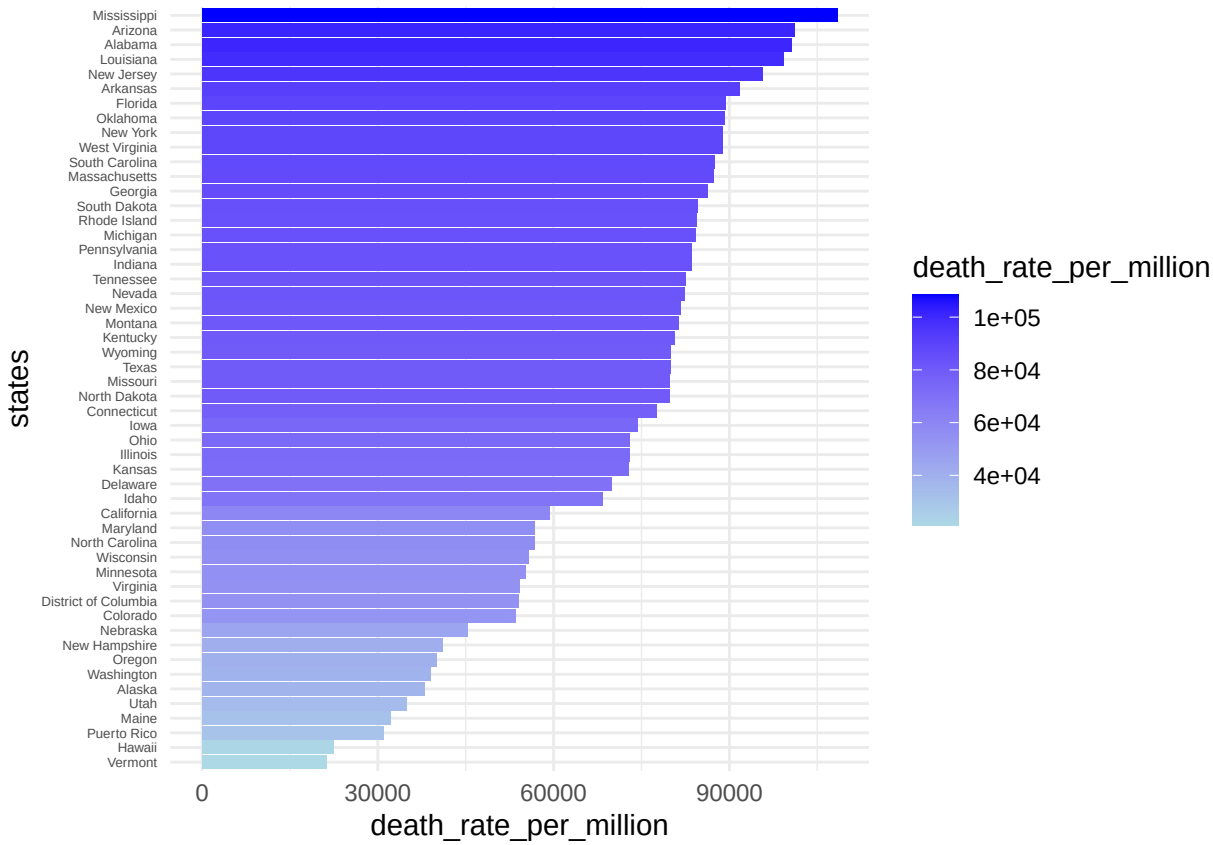



```r
# 2021 death rate data analysis
plot_usmap(data=us_states_2021, values = "death_rate_per_million", labels = TRUE) +
  scale_fill_gradient(name = "death_rate_per_million",
                      low = "light blue", high = "blue") +
  theme(legend.position = "right") +
  ggtitle("death rate per million in 2021")
```

death rate per million in 2021




```
us_states_2021 %>%
  filter(month == "12") %>%
  ggplot(aes(x = reorder(state, death_rate_per_million),
             y = death_rate_per_million)) +
  geom_bar(aes(fill = death_rate_per_million),
           stat = "identity")+
  scale_fill_gradient(name = "death_rate_per_million",
                      low = "light blue", high = "blue") +
  labs(x = "states") +
  theme_minimal() +
  theme(axis.text.y = element_text(size = 5)) +
  coord_flip()
```

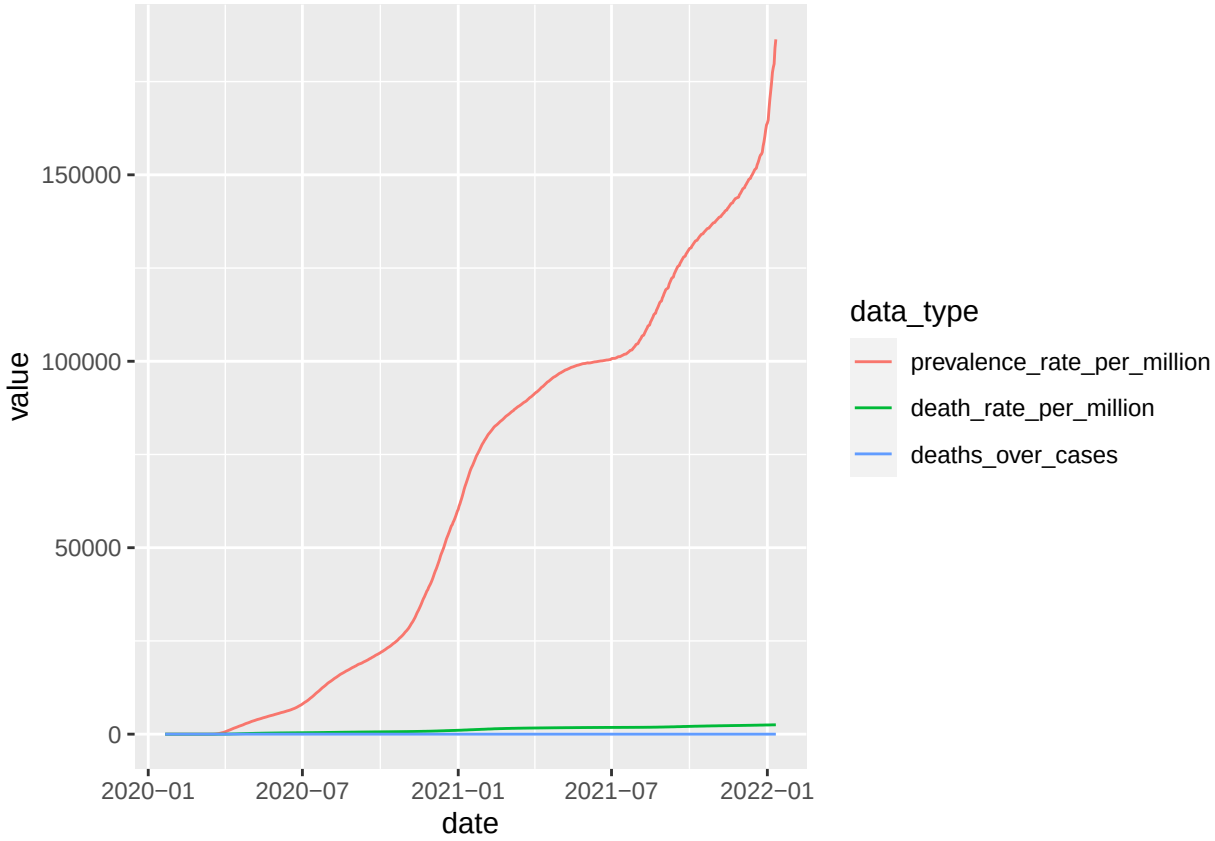


Based on the analysis, it is shown that certain state seems to have lower prevalence rate and death rate, it is important for states such as North Dakota and South Dakota to learn from State of Vermont and Maine to reduce prevalence rate policy-wise if possible.

Question 2

In the second part of data exploratory data analysis, we focused on prevalence rate and death rate ratio in the us_data to see if interventions may reduce deaths over cases.

```
us_data_3 <- us_data_2 %>%
  pivot_longer(!date, names_to = "data_type", values_to = "value")

us_data_3$data_type <- as.factor(us_data_3$data_type)

us_data_3 %>%
  mutate(data_type = fct_relevel(data_type, "prevalence_rate_per_million", "death_rate_per_million", "d
  ggplot(mapping = aes(x = date, y = value, color = data_type))+
  geom_line()
```

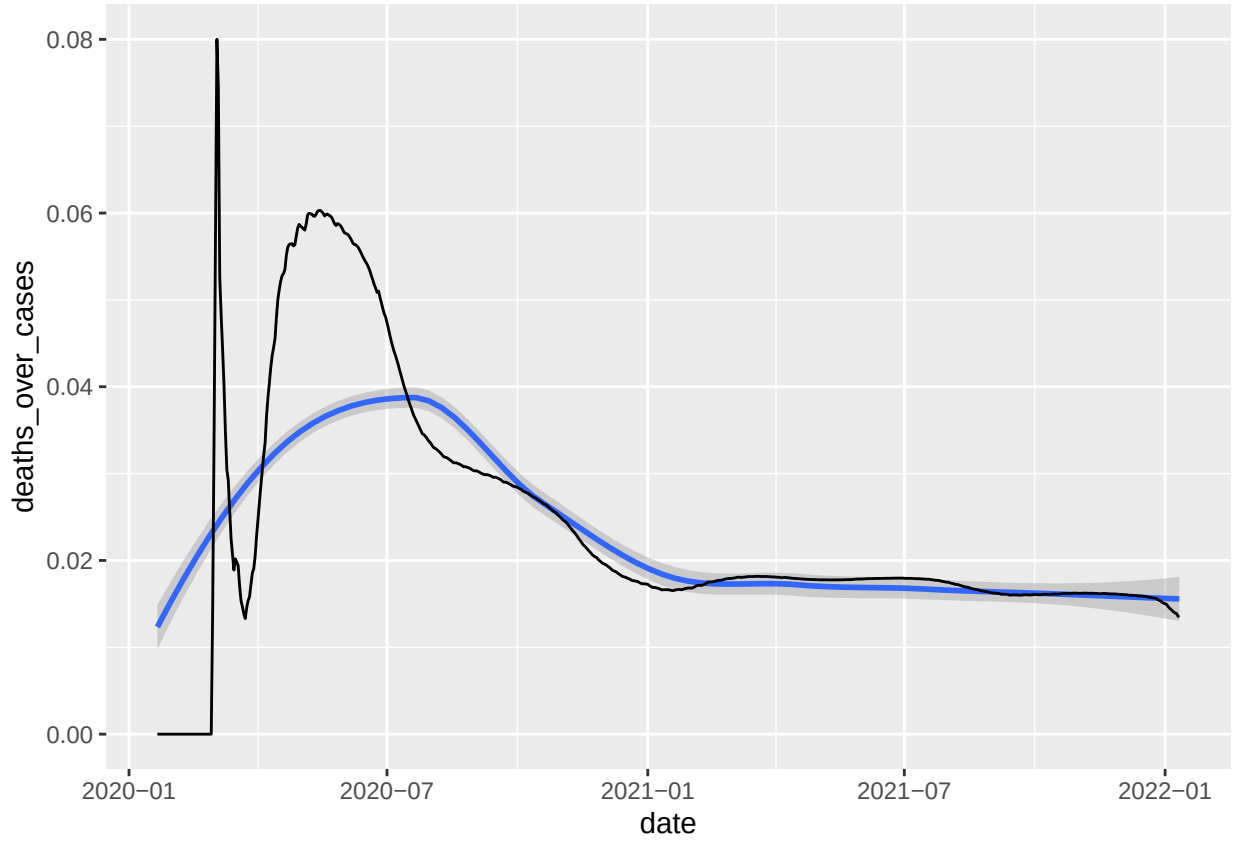


Form the plot, we can see that the prevalence rate is increasing sharply but death reat per million is controlled well when compared to the prevalence rate.However, there is also no specific changes in death_over_cases from the graphs.This might because the data was too small when compared to the prevalence rate data in the same scale.

In this case, when look at death_over_cases data alone:

```
us_data_2 %>%
  ggplot(mapping = aes(x = date,
                       y = deaths_over_cases)) +
  geom_smooth()+
  geom_line()
```

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



By looking at the data closely, it is easy to detect a drop around 2020-03 and remain constant ater 2021-01 based on the data despite of prevalence rate. Trace back to what happened in those points, on March 2020, quarantine policy was launched in multiple states which seems like an very effective intervention to reduce deaths over cases. From 2021, vaccine start to be available to US population, however, the correlation needs more data to support.

Summary

Based on my simple data analysis, differences between prevalence rate in different states is obvious. Even though there are possible confounders such as number of people tested, it is still important for those states with significantly high prevalence rate and death rate such as North Dakota to be aware of current intervention launched.

Also, when compared data in 2021 with 2020, although the number of cases and deaths are increasing, the prevalence death rate for certain states such as New York and New Jersey dropped significantly. This illustrated that potential interventions launched in these states might be more effective when compared to other states, which could be closely analyzed further.

To address these problem, prevalence rate and death rate are calculated to reduce bias of population and density map and death rate density map is illustrated to compare data of each state in 2020 and 2021 respectively to come up with the conclusion.

When considered deaths over cases rate, there are a few significant drops from Janurary 2020 to December 2021. This trend shows that potentially, current death over cases is stable. This might contributed to interventions launched to overcome the pandemic.

In this case, deaths over cases rate is calculated daily to plot a line chart that may illustrate the changes easily.

Overall, from current data analysis, humans seems still far away from overcome the pandemic completely. However, we shall not lose faith since we definitely know more about the COVID-19 virus from valuable data collected.

To further improve this analysis, data of intervention policies of each states and effectiveness measurement of current interventions would be significant to further consolidate the findings of this report.

Bibliography

Bureau, U. S. C. (2022, January 24). 2020 census. Census.gov. Retrieved January 29, 2022, from https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html

The New York Times. (2020, April 1). COVID-19 Data. The New York Times. Retrieved January 29, 2022, from https://www.nytimes.com/interactive/2021/us/new-york-covid-cases.html