FinalProject

Sarvar Khamidov

2024-12-07

Synopsis

For the Final Project for Statistical Programming in R course I have chosen to analyze the case fatality rate (CFR) in the US and in each state. I'm trying to identify if the CFR has risen and lowered over time and identify top 5 and bottom 5 states with highest and lowest CFRs how they differ in each state. Questions like this usually being asked in the government as well for resource allocation, policy making and public awareness.

For this problem I would calculate CFR for whole US and each individual states by each month. I would use use federal-level for overall trend CFR and state-level data for granual analysis of CFR at the state level. I would need to clean, focusing on key variables (date, state, cases, deaths), identify missing values. I would also calculate CFR monthly to smooth the daily fluctuations.

I will build plot and map plots to show my findings. The Plot would show the trend of the CFR over time with top 5 and bottom 5 states for each year. Trend plot for Building a map visulization with every state would help too see the whole picture at once allowing viewers to compare each state.

Packages Required

```
# read csv files
library(readr)
# data manipulation, filter()
library(dplyr)
# year / month extraction from date variable
library(lubridate)
# for visualization (plot and map)
library(ggplot2)
# for annotation of the plot
library(ggrepel)
# for visualization of the map
library(maps)
```

Data Preparation

This project uses two Coronavirus (COVID-19) Data in the United States datasets from NYTIMES github, the link for which can be found below. Two of the data are us-states.csv and us.csv. https://github.com/nytimes/covid-19-data?tab=readme-ov-file#geographic-exceptions

NYTIMES has been collecting the number of cases and deaths in every county in US for more than three years from March 2020 to March 2024. The data contains daily cases and deaths for the whole United States and each state individually overtime. The data was compiled from state and local governments and health departments. The dataset contains the date, state name, number of cases and number of deaths. In the deaths and cases contain probable and confirmed. Confirmed cases / deaths are based on confirmatory laboratory testing, and probable cases are based on on specific criteria for testing, symptoms and exposure. Each date contains cumulative number of cases and deaths as announced death and case.

Since all the data is given in csv format, I have imported them to R dataframes using read_csv function from readr package. After I investigated the data looking at the the columns, values, number of columns and rows, see if the data is messy. Then I checked for missing values for each data trying to understand what is the meaning of them.

```
us_states <- read_csv("us-states.csv")
```

```
## Rows: 61942 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): state, fips
## dbl  (2): cases, deaths
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
us <- read_csv("us.csv")
```

```
## Rows: 1158 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (2): cases, deaths
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Counting how many na values has the state data
sum(is.na(us_states))
```

```
## [1] 0
```

```
# Counting how many na values has the data
sum(is.na(us)) # Total # of na values
```

```
## [1] 0
```

```
head(us_states)
```

```
## # A tibble: 6 x 5
##   date       state      fips  cases deaths
##   <date>     <chr>      <chr> <dbl>  <dbl>
```

```
## 1 2020-01-21 Washington 53        1      0
## 2 2020-01-22 Washington 53        1      0
## 3 2020-01-23 Washington 53        1      0
## 4 2020-01-24 Illinois   17        1      0
## 5 2020-01-24 Washington 53        1      0
## 6 2020-01-25 California 06        1      0
```

```
head(us)
```

```
## # A tibble: 6 x 3
##   date       cases deaths
##   <date>     <dbl>  <dbl>
## 1 2020-01-21     1      0
## 2 2020-01-22     1      0
## 3 2020-01-23     1      0
## 4 2020-01-24     2      0
## 5 2020-01-25     3      0
## 6 2020-01-26     5      0
```

```
str(us)
```

```
## spc_tbl_ [1,158 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date  : Date[1:1158], format: "2020-01-21" "2020-01-22" ...
##  $ cases : num [1:1158] 1 1 1 2 3 5 5 5 5 6 ...
##  $ deaths: num [1:1158] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   date = col_date(format = ""),
##   ..   cases = col_double(),
##   ..   deaths = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

The US data contains 3 variables (date, cases and deaths) with 1158 rows. “date” is the date of the collection in “yyyy-mm-dd” format. Also “deaths” and “cases” contain number of deaths and cases respectevly for each day.

```
str(us_states)
```

```
## spc_tbl_ [61,942 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date  : Date[1:61942], format: "2020-01-21" "2020-01-22" ...
##  $ state : chr [1:61942] "Washington" "Washington" "Washington" "Illinois" ...
##  $ fips  : chr [1:61942] "53" "53" "53" "17" ...
##  $ cases : num [1:61942] 1 1 1 1 1 1 1 1 2 ...
##  $ deaths: num [1:61942] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   date = col_date(format = ""),
##   ..   state = col_character(),
##   ..   fips = col_character(),
##   ..   cases = col_double(),
##   ..   deaths = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

The US data contains 5 variables (date, state, fips, cases and deaths) with 61,942 rows. “date” is the date of the collection in “yyyy-mm-dd” format. State indicates the name of the state whereas the “fips” column is coded value for the state. Also “deaths” and “cases” contain number of deaths and cases respectevly for each day.

Exploratory Data Analysis

The primary purpose of the project is to compare the Case Fatality Rate for each state and nation wide. So we would need to calculate CFR, I started by calculating for each dataset which is total number of death divided to total number of cases. Since the data contains daily number of cases and deaths, I decided to calculate the aggregated death rate for each month by finding the total number of death and cases for each month.

```
state_per_month_cr <- us_states %>%
  mutate(month = months(date),
         year = year(date)) %>%
  group_by(state, month, year) %>%
  summarize(total_cases = sum(cases),
            total_deaths = sum(deaths)) %>%
  mutate(CFR_per = ifelse(total_deaths > 0 & total_cases > 0,
                      (total_deaths/total_cases)*100,
                      0),
         month_n = as.integer(factor(month, levels = month.name)))
```

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

```
head(state_per_month_cr)
```

```
## # A tibble: 6 x 7
## # Groups:   state, month [2]
##   state   month   year total_cases total_deaths CFR_per month_n
##   <chr>   <chr>  <dbl>       <dbl>        <dbl>   <dbl>   <int>
## 1 Alabama April   2020      125166         4044    3.23       4
## 2 Alabama April   2021    15647349       322507    2.06       4
## 3 Alabama April   2022    38938066       583766    1.50       4
## 4 Alabama August  2020     3355858        58883    1.75       8
## 5 Alabama August  2021    19824481       367172    1.85       8
## 6 Alabama August  2022    45122723       620378    1.37       8
```

```
us_per_month_cr <- us %>%
  mutate(month = months(date),
         year = year(date)) %>%
  group_by(month, year) %>%
  summarize(total_cases = sum(cases),
            total_deaths = sum(deaths)) %>%
  mutate(CFR_per = ifelse(total_deaths > 0 & total_cases > 0,
                      (total_deaths/total_cases)*100,
                      0),
         month_n = as.integer(factor(month, levels = month.name)))
```

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

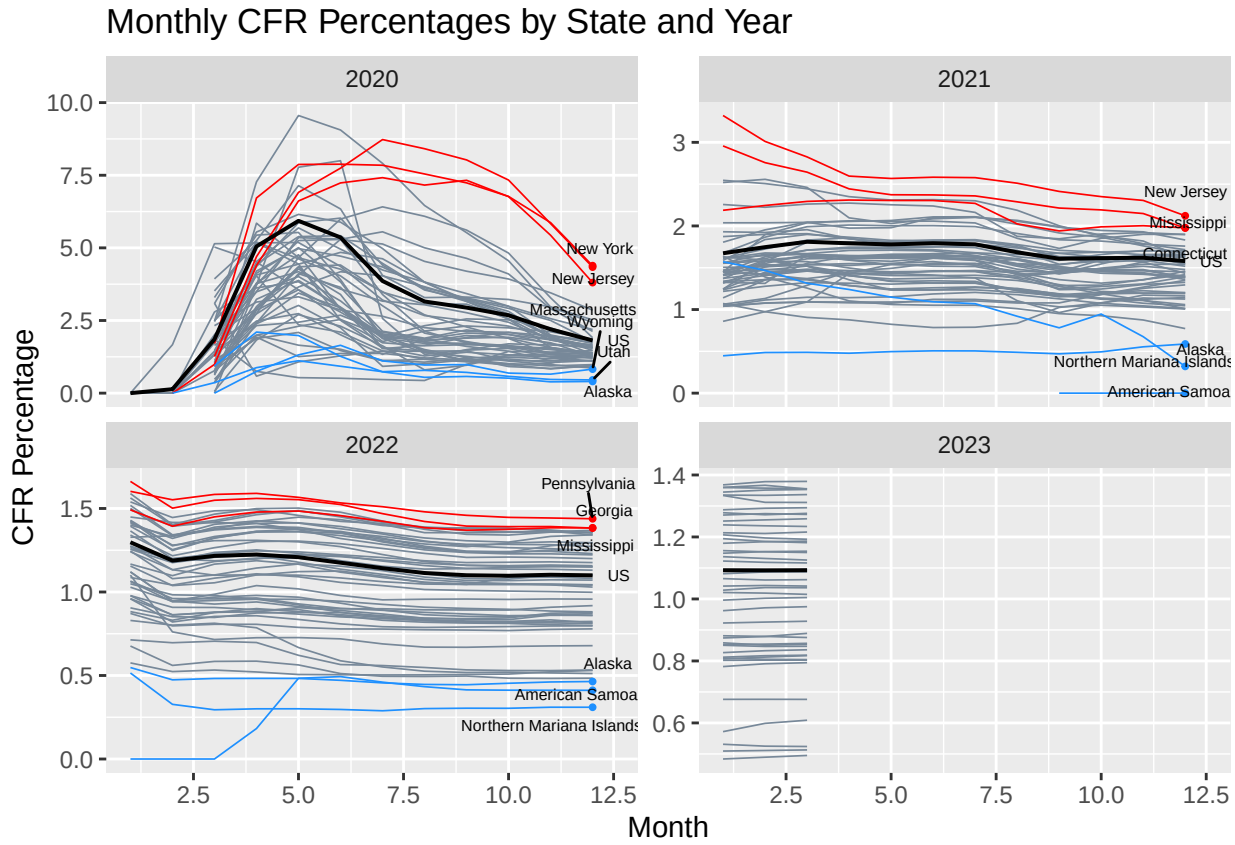
```
head(us_per_month_cr)
```

```
## # A tibble: 6 x 6
## # Groups:   month [2]
##   month   year total_cases total_deaths CFR_per month_n
##   <chr>  <dbl>       <dbl>        <dbl>   <dbl>   <int>
## 1 April   2020    19611708       990492    5.05       4
## 2 April   2021   944997168     16938795    1.79       4
## 3 April   2022  2416616245     29590230    1.22       4
## 4 August  2020   166758528      5259820    3.15       8
## 5 August  2021  1148978762     19355736    1.68       8
## 6 August  2022  2879851473     32068656    1.11       8
```

To visualize the trend of the CFR over time I have decided to create a line plot, where x axis is time in months, y axis is CRF in percent and each line represents a state with top 3 of that year colored in red and bottom 3 in blue. Also the black line represents the data for the whole US.

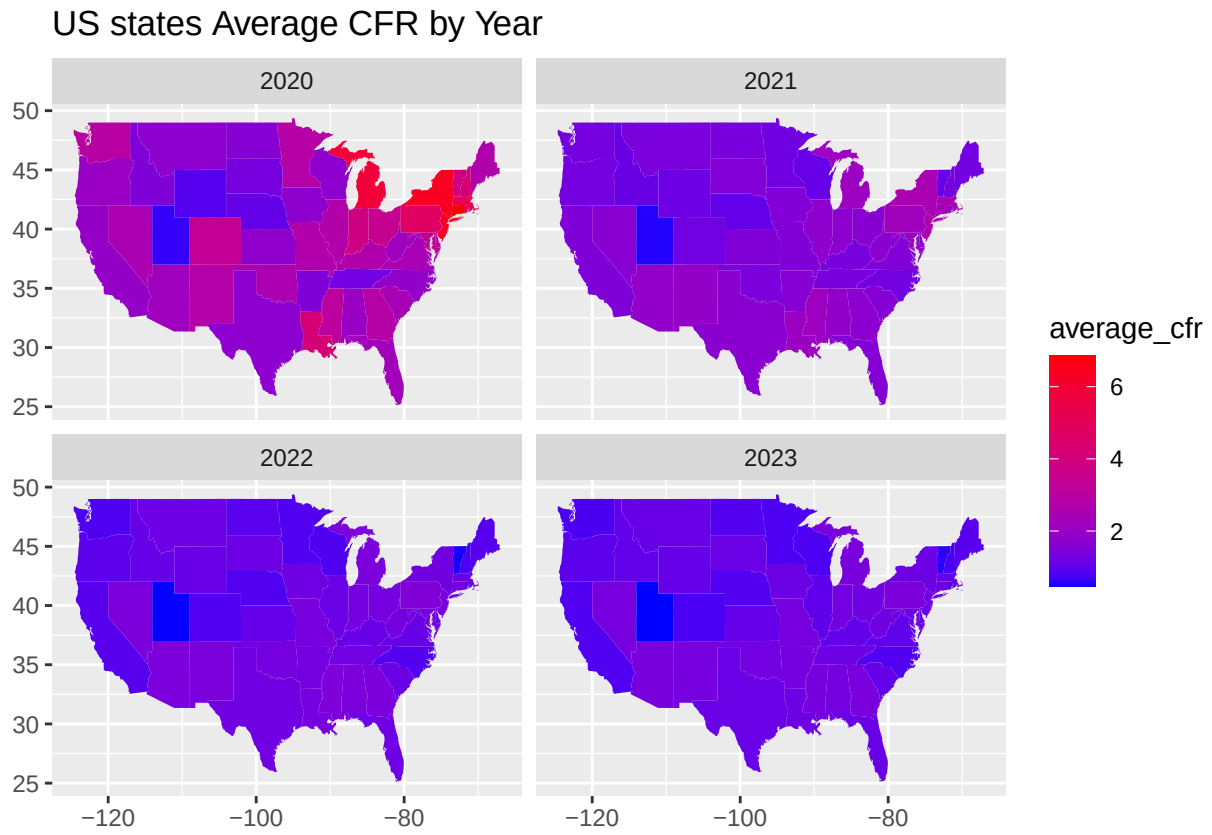
```
top3_per_year <- state_per_month_cr %>%
  filter(month_n == 12) %>%
  group_by(year) %>%
  arrange(desc(CFR_per), .by_group = TRUE) %>%
  slice(1:3) %>%
  ungroup()

bottom3_per_year <- state_per_month_cr %>%
  filter(month_n == 12) %>%
  group_by(year) %>%
  arrange(CFR_per, .by_group = TRUE) %>%
  slice(1:3) %>%
  ungroup()

# Combine all data
total_combined <- rbind(us_per_month_cr %>% mutate(state = "US"), state_per_month_cr)

# Plot
ggplot(total_combined, aes(x = month_n, y = CFR_per, group = state)) +
  # Background lines for all states
  geom_line(data = filter(total_combined, !(state %in% c(top3_per_year$state, bottom3_per_year$state))),
            aes(x = month_n, y = CFR_per, group = state),
            colour = "lightslategrey",
            size = 0.3) +
  # Highlight top3 for each year
  geom_line(data = total_combined %>% semi_join(top3_per_year, by = c("state", "year")),
            aes(x = month_n, y = CFR_per, group = state),
            colour = "red",
            size = 0.3) +
  geom_point(data = top3_per_year, aes(month_n, CFR_per), color = "red", size = 0.7) +
  geom_text_repel(data = top3_per_year,
                  aes(label = state),
                  nudge_x = 0.5,
```

```
                size = 2,
                direction = "y",
                vjust = 0.5,
                hjust = -1) +
# Highlight bottom3 for each year
geom_line(data = total_combined %>% semi_join(bottom3_per_year, by = c("state", "year")),
          aes(x = month_n, y = CFR_per, group = state),
          colour = "dodgerblue",
          size = 0.3) +
geom_point(data = bottom3_per_year, aes(month_n, CFR_per), color = "dodgerblue", size = 0.7) +
geom_text_repel(data = bottom3_per_year,
                aes(label = state),
                nudge_x = 0.5,
                size = 2,
                direction = "y",
                vjust = 0.5,
                hjust = -1) +
# Highlight US line
geom_line(data = total_combined %>% filter(state == "US"),
          aes(month_n, CFR_per),
          color = "black",
          size = 0.7) +
geom_text_repel(data = filter(total_combined, state == "US" & month_n == 12),
                aes(label = state),
                nudge_x = 0.5,
                size = 2,
                direction = "y",
                vjust = 0.5,
                hjust = -1) +
# Facet by year
facet_wrap(~ year, scales = "free_y") +
# General theme adjustments
theme(legend.position = "none") +
labs(title = "Monthly CFR Percentages by State and Year",
     x = "Month",
     y = "CFR Percentage")
```




```
state_per_month_cr %>%
  group_by(state, year) %>%
  summarise(average_cfr = mean(CFR_per)) %>%
  ungroup() %>%
  mutate(region = tolower(state)) %>%
  right_join(map_data("state"), by = "region") %>%
  ggplot(aes(x = long, y = lat, group = group)) +
  geom_polygon(aes(fill = average_cfr)) +
  scale_fill_gradient(low = "blue", high = "red") +
  facet_wrap(~ year) +
  labs(title = "US states Average CFR by Year",
       x = element_blank(),
       y = element_blank())
```

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



The first graph highlights the monthly trends in Case Fatality Rate (CFR) percentages across U.S. states from 2020 to 2023. In 2020, states like New York and New Jersey experienced significantly higher CFRs, peaking early in the pandemic. Over time, CFRs showed a general decline, with fewer states exhibiting extreme values in subsequent years. The bottom-performing states, such as Alaska and Utah, maintained consistently lower CFRs. By 2023, CFR variability diminished significantly, suggesting improved healthcare responses and outcomes across the board.

The second graph maps the average CFR by year across U.S. states. In 2020, northeastern states and some central regions experienced the highest CFRs, reflected by redder tones. Over the following years, CFRs generally decreased, with most states displaying a predominantly blue-to-purple gradient, indicating lower fatality rates. By 2023, nearly all states had CFRs in a similar, reduced range, suggesting a convergence toward improved survival rates and uniformity in pandemic response efforts nationwide.

Summary

In this project I analyzed and compared the case fatality rate (CFR) trend of COVID-19 in the US and in each state. To address the problem I had to calculate the CFR using number of deaths and cases in US and each state by each year using the given data for COVID-19 data for 2020-2023. To create map I used lat and long data of states that is available in map package.

Overall trend for CFR for all state and US has been declining since the beginning of the pandemic which was easily shown in two of the graphs. States from northeast had highest CFR percentages at the beginning of the pandemic.

These difference in death rate in each state might not been very visible to the common person. However, some of the reader might have felt the difference in state's regulations during the pandemic which might have been influenced because of the Case Fatality Rate of that state.

Some of the limitations of my analysis that there was not explanation to any sudden changes in trend. To improve it I would add any big COVID-19 related news dates on and see if matches with trend changes. Another limitation was that the analysis was done US based only, I would also try to other countries data.