Covid-19 Burden by US State

2022-08-02

Introduction

From the onset of covid-19, different states in the US have collected and generated a massive amount of data related to cases. An individual living in the US may have greater likelihood of contracting covid-19 depending on the state of residence. In addition, although all states were affected, some were affected at far greater levels than others. An analysis of the total number of cases in each state is necessary to begin to investigate and compare the burden covid-19 has caused on each state. Furthermore, to account for differences in population, total covid-19 cases per a certain number of residents will be evaluated.

Data source and collection

The first step of this data analysis process was to gather the data using the Johns Hopkins Covid-19 data as a primary source and saving the CSV file to the workbook. This analysis only utilizes information for US cases. In the US cases data set, I retained only data for cases, states, and dates. I then changed the dates from columns to rows.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi
file_names <- c("time_series_covid19_confirmed_US.csv")
urls <- str_c(url_in, file_names)
US_cases <- read_csv(urls[1])
```

```
## Rows: 3342 Columns: 947
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (941): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20,...
##
## 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.
```

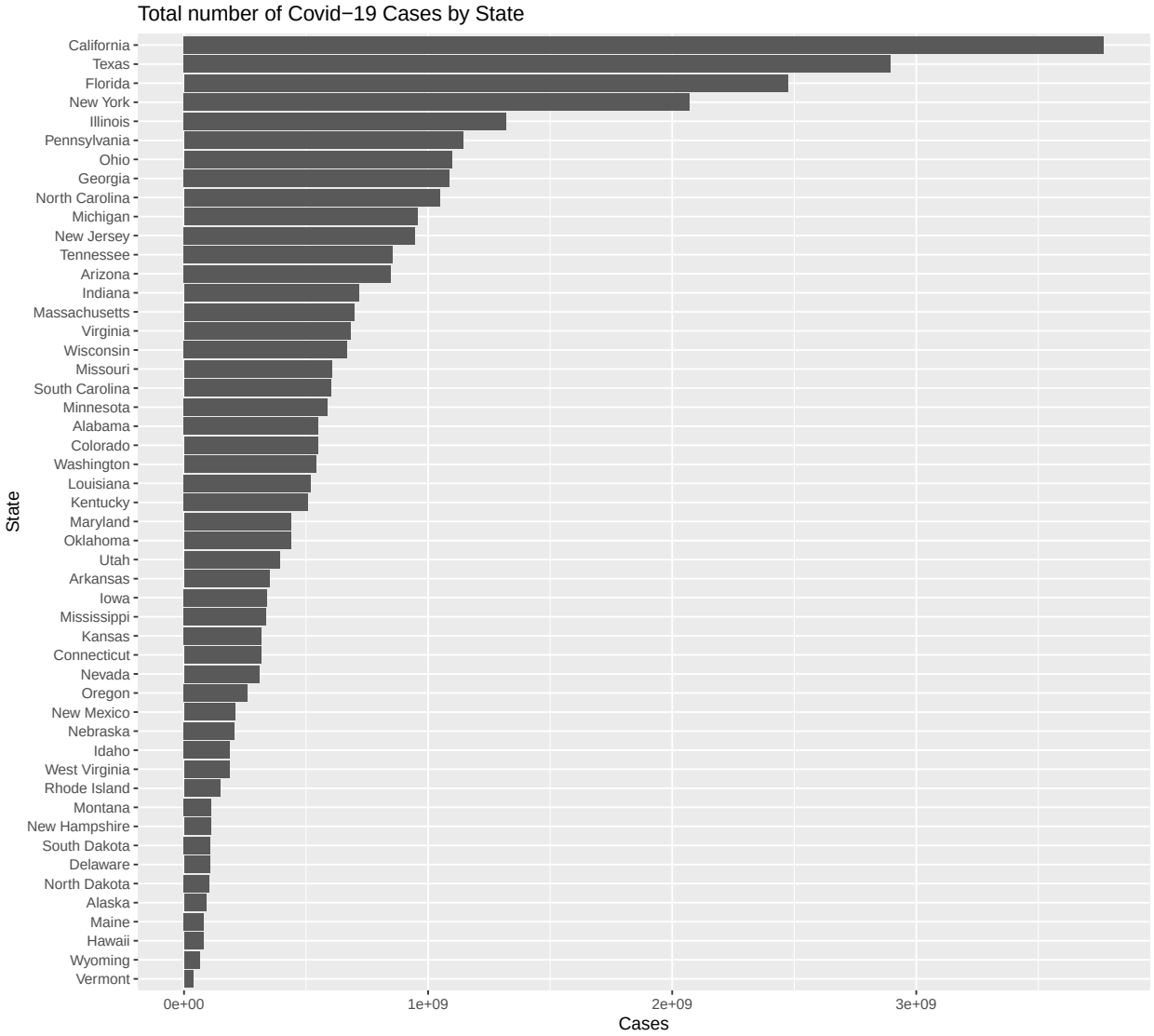
```
US_cases <- US_cases %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Country_Region, Lat, Long_, Admin2, Combined_Key))

US_cases <- US_cases %>%
  pivot_longer(cols = -c('Province_State'),
               names_to = 'Date',
               values_to = 'Cases'
  )
US_cases <- US_cases %>%
  rename(State = 'Province_State')
```

Total number of cases for each state are obtained below and cleaned up to contain only the 50 US states and exclude information for non-states such as American Samoa, Diamond Princess, and Washington DC. This information is then sorted based on states with the greatest number of cases.

```
US_cases <- US_cases %>%
  group_by(State) %>%
  summarise(Cases = sum(Cases))
US_cases <- US_cases %>%
  filter(!row_number() %in% c(3, 10, 11, 14, 15, 40, 45, 53))
```

States Ranked by Total Number of Cases

```
ggplot(US_cases, aes(x= Cases, y=reorder(State, Cases)))+
 geom_col() + labs(title = "Total number of Covid-19 Cases by State", y = 'State')
```



Analysis of States with Highest Number of Cases

The above visualization ranks states by the highest number of Covid-19 cases from the beginning of the pandemic until August 2022. It's clear that California has the highest number of cases followed by Texas, Florida, and New York. The state with the lowest number of cases was Vermont. However, this graph does not account for the populations of states. California has a far higher population than Vermont and would be expected to have a higher number of cases if they had similar infection rates.

Secondary Data Collection and Analysis: State Population

To account for differences in population, I utilized data from the US Census populations by state for 2021 and excluded Washington DC and Puerto Rico. I loaded in data for each state's population below. To obtain the total number of cases by 100,000, I divided the total number of cases for each state by population and multiplied by 100,000.

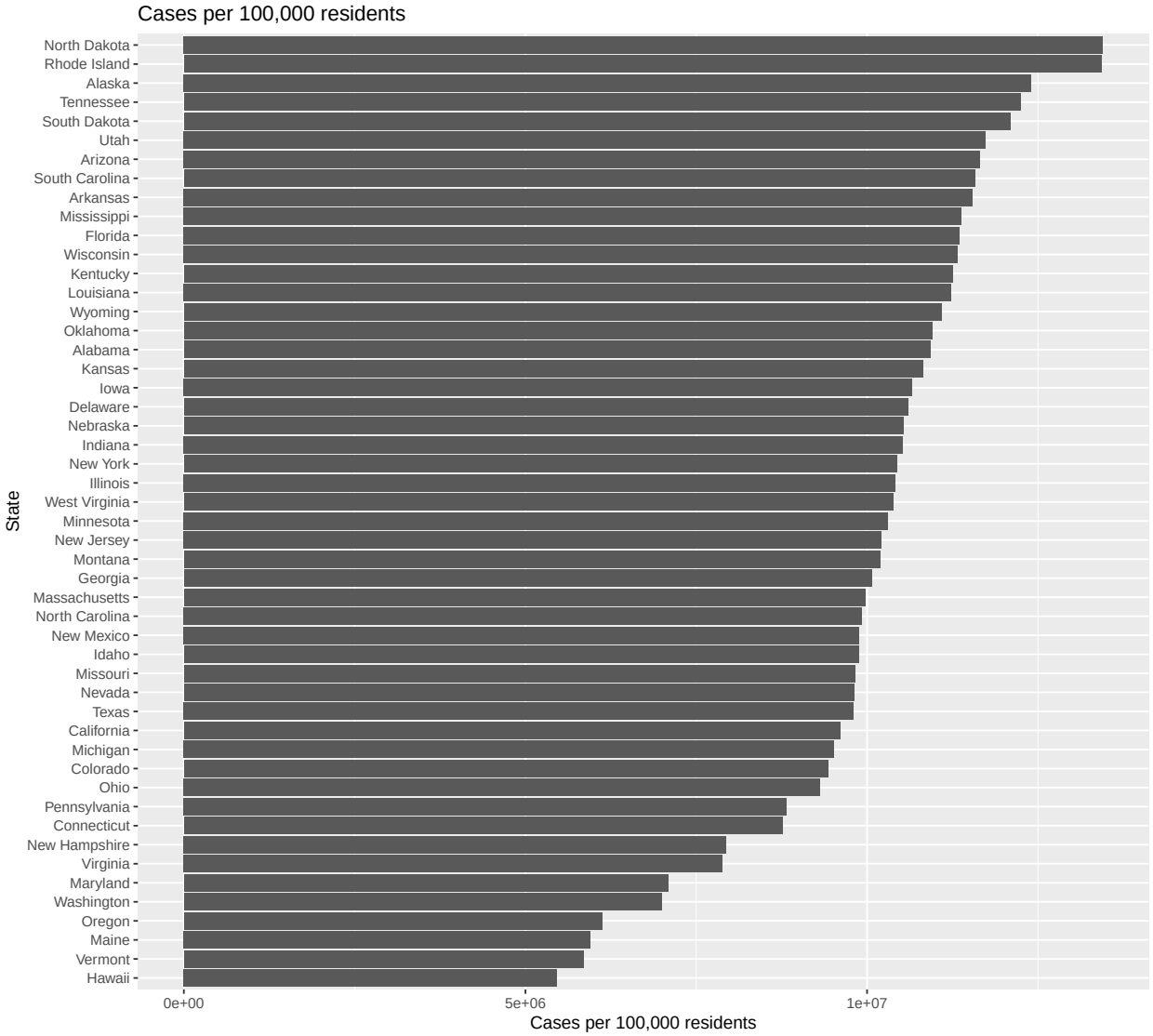
```
url_in2 <- "https://www2.census.gov/programs-surveys/popest/datasets/2020-2021/state/totals/NST-EST2021
pop_data <- read_csv(url_in2[1])
```

```
## Rows: 57 Columns: 30
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): SUMLEV, REGION, DIVISION, STATE, NAME
## dbl (25): ESTIMATESBASE2020, POPESTIMATE2020, POPESTIMATE2021, NPOPCHG_2020,...
##
## 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.
```

```
pop_data <- pop_data %>%
  select(NAME, POPESTIMATE2021)
pop_data <- pop_data %>%
  filter(!row_number() %in% c(1, 2, 3, 4, 5, 14, 57)) %>%
  rename(State = 'NAME', Population = 'POPESTIMATE2021')

US_cases_pop_combined <- cbind(US_cases, pop_data)
US_cases_pop_combined <- US_cases_pop_combined %>%
  rename(state2 = 'State') %>%
  select(-c(State)) %>%
  rename(State = 'state2')

US_cases_pop_combined$Population <- as.numeric(gsub(",", "", as.character(US_cases_pop_combined$Populati

US_cases_pop_combined <- US_cases_pop_combined %>%
  mutate(cases_per_100000 = Cases / Population * 100000)
```

States Ranked by Cases per 100,000 Residents

```
ggplot(US_cases_pop_combined, aes(x= cases_per_100000, y=reorder(State, cases_per_100000)))+
 geom_col() + labs(title = "Cases per 100,000 residents", x = 'Cases per 100,000 residents', y = 'State
```



Analysis

When looking at not the total cases of each state but instead the total cases per 100,000 residents, the visualization tells a different story. California had the greatest total amount of cases, but actually had the 14th lowest amount of cases by population compared to other cases.

The top three states with the highest amounts of cases based on population were North Dakota, Rhode Island, and Alaska. The states that had the lowest amounts of cases based on population were Hawaii, Vermont, and Maine.

Further analysis is necessary to determine the cause for these lower proportions of cases. Possible causes could include mask mandates, stay at home orders, immunizations, and closures of places of high density such as restaurants and concerts. In all likeliness, these protocols have varying effectiveness in different states. Multiple states could have the same mask mandates, for example, but 90% may adhere to the mandate in one state while only 40% adhere in another. This would indicate complex reasoning behind the number of cases per 100,000 residents, assuming that all states reported cases accurately.

Bias and Incompleteness

Bias in this analysis could come from my own geological location. I'm located in Colorado which from my analysis, is 39th/50 by cases per 100,000 residents. I attempted to only use official data that was reported to Johns Hopkins and the US census bureau and did not add anecdotal information related to mask usage or social distancing adherence in my area. I stated in my analysis that further analysis is necessary to determine the true cause for lower cases and stated some possible causes but did not list them as facts.

Incompleteness in this dataset can come from a multitude of reasons. First of all, not every state may have the same accuracy in reporting their numbers of covid cases. This can come from reasons ranging from inadequate resources around health departments and testing/reporting to the public simply not getting tested when they have symptoms and/or not reporting results of at-home tests. It's likely states have varying levels of accurateness in their reporting of covid numbers which could change the ranking of covid cases by state and population.

Conclusion

Looking at the total number of cases by state can be a misleading metric because it does not tell the whole story. Calculating and showing cases based on a standardized number of people normalizes the data and shows that some states had a far greater proportion of their population affected by covid than others.