COVID-19 in the Continguous United States

William Wilson

2023-07-11

The following packages are used in this report:

```
library(tidyverse)
library(stringr)
library(ggplot2)
library(ggthemes)
library(maps)
library(mapdata)
library(modelr)
library(utils)
library(gt)
library(scales)
```

The data comes from COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University at https://github.com/CSSEGISandData/COVID-19#covid-19-data-repository-by-the-center-for-systems-science-and-engineering-csse-at-johns-hopkins-university

```
#set base url
url_in <- str_c(
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19",
  "/master/csse_covid_19_data/csse_covid_19_time_series/")
#indicate individual files
file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_US.csv",
                "time_series_covid19_deaths_global.csv",
                "time_series_covid19_recovered_global.csv")

urls <- str_c(url_in,file_names)

#retrieve data
us_cases <- read_csv(urls[1])
global_cases <- read_csv(urls[2])
us_deaths <- read_csv(urls[3])
global_deaths <- read_csv(urls[4])
global_recovered <- read_csv(urls[5])
```

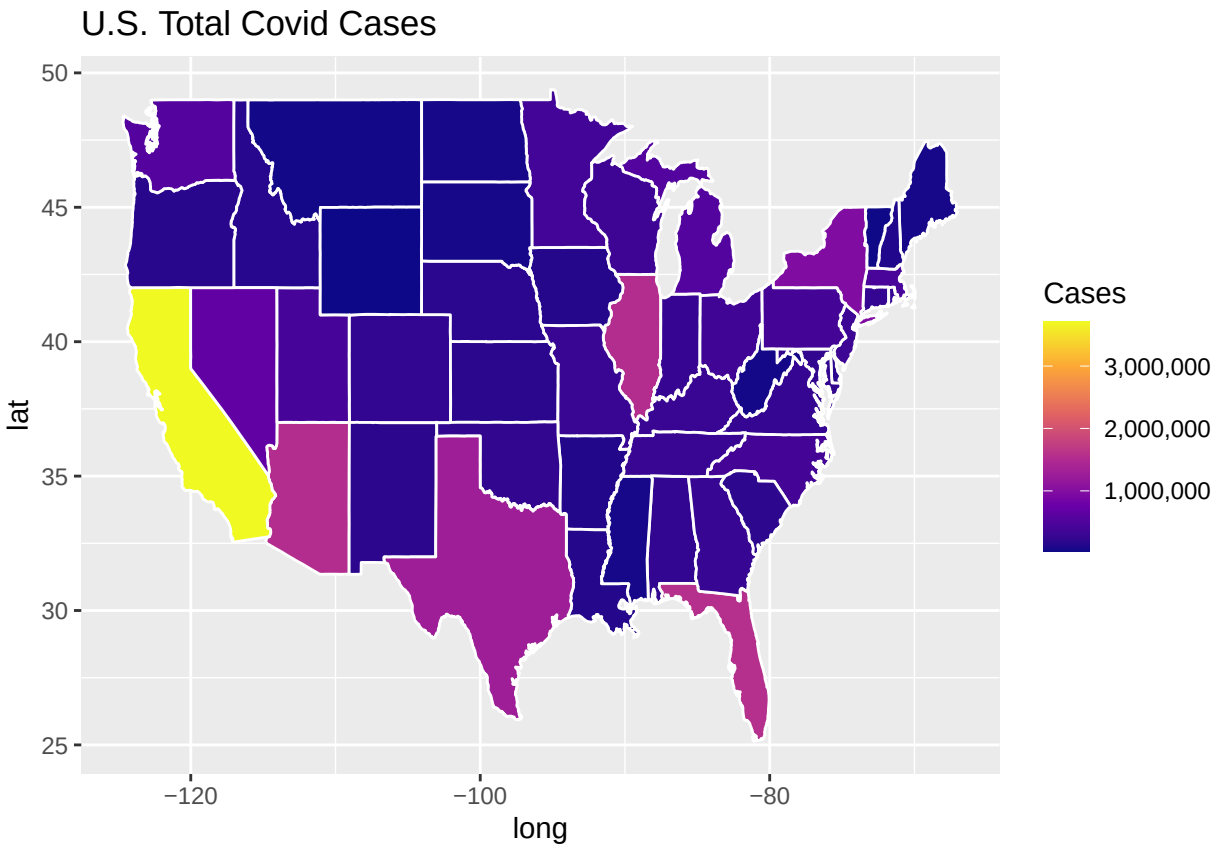
COVID-19 Case Data

We will analyze the amount of cases by state within the contiguous United States using the data provided.

```
#convert using pivot_longer to make more "R friendly"
us_cases_longer <- us_cases %>% as_tibble() %>% select(-(UID:FIPS))  %>% pivot_longer(cols = -(Admin2:C
#Summarize by State
us_cases_by_state <- us_cases_longer %>% group_by(Province_State) %>%  summarise(cases = max(cases), La

#make compatible with the map data we will join
us_cases_by_state <- us_cases_by_state %>% rename(region = Province_State)
```

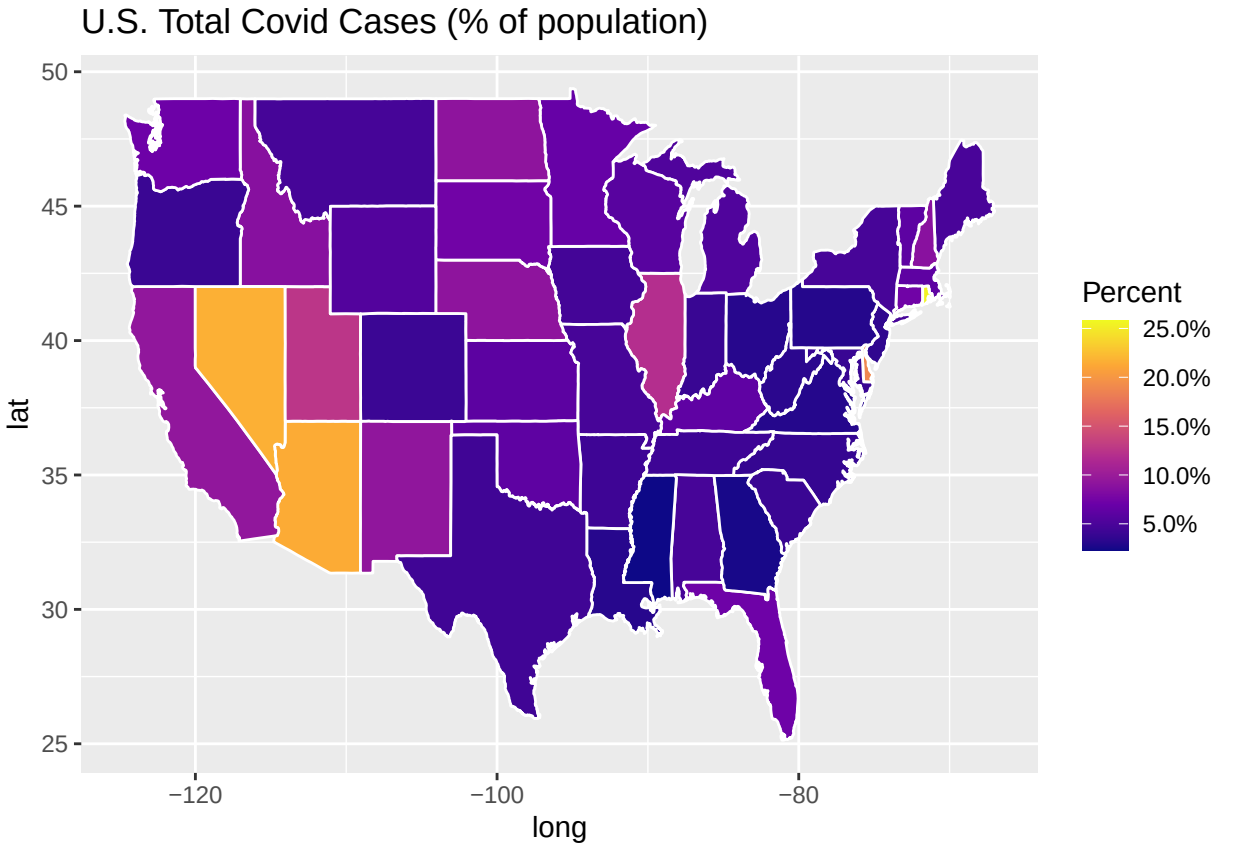
First we look at the total cases by state. This will largely reflect the population of the state as a whole. For instance, California shows as the state with the most cases because it also has the largest population.

```
us_cases_by_state$region <- tolower(us_cases_by_state$region)

states_map <- map_data("state")
covid_map <- left_join(states_map, us_cases_by_state, by = "region")

ggplot(covid_map, aes(long, lat, group = group))+
  geom_polygon(aes(fill = cases), color = "white")+
  scale_fill_viridis_c(option = "C",labels=label_comma()) + labs(title = 'U.S. Total Covid Cases', fill=
```



In order to better show the trend we will now map the cases as a percentage of the population. This will allow us to see how common Covid is within the population. The population data was taken from the census.gov website as noted in the comments below. Interestingly California is no longer the top state.

```r
#Keeping things easy, manually entering population data found at
#https://www2.census.gov/programs-surveys/popest/datasets/2020-2022/state/totals/
states <- c(
  "alabama","alaska","arizona","arkansas","california",
  "colorado","connecticut","delaware","district of columbia",
  "florida","georgia","hawaii","idaho","illinois","indiana",
  "iowa","kansas","kentucky","louisiana","maine","maryland",
  "massachusetts","michigan","minnesota","mississippi",
  "missouri","montana","nebraska","nevada","new hampshire",
  "new jersey","new mexico","new york","north carolina",
  "north dakota","ohio","oklahoma","oregon","pennsylvania",
  "rhode island","south carolina","south dakota","tennessee",
  "texas","utah","vermont","virginia","washington",
  "west virginia","wisconsin","wyoming","puerto rico")
pops <- c(5024356,733378,7151507,3011555,39538245,5773733,
          3605942,989957,689546,21538226,10711937,1455273,1839092,
          12812545,6785668,3190372,2937847,4505893,4657749,1362341,
          6177213,7029949,10077325,5706504,2961288,6154920,1084197,
          1961489,3104624,1377518,9289031,2117527,20201230,10439414,
          779091,11799374,3959346,4237291,13002689,1097371,5118429,
          886677,6910786,29145428,3271614,643085,8631384,7705247,
          1793755,5893725,576837,3285874
)
#create a tibble with state names and population
state_pops <- tibble(states,pops)
state_pops <- state_pops %>% rename(region = states)
#join the populatio data to the Covid Map data and add percentage calculation
covid_pop_map  <- left_join(covid_map, state_pops, by = "region")
covid_pop_map <- covid_pop_map %>% mutate(perc_inf = cases/pops)
#Change the state names to look nice when printed
covid_pop_map$region <- str_to_title(covid_pop_map$region)

#Plot a map showing percentage of population who have had Covid
ggplot(covid_pop_map, aes(long, lat, group = group))+
  geom_polygon(aes(fill = perc_inf), color = "white")+
  scale_fill_viridis_c(option = "C",label=label_percent(.1)) + labs(title = 'U.S. Total Covid Cases (%
```



The data is now shown in a table. The percentage indicates the number of cases divided by the population of the state. Notable states / regions include, the District of Columbia at 25.81% and Rhode Island with 25.40% in the East, and Arizona with 21.40% and Nevada with 21.62% in the west.

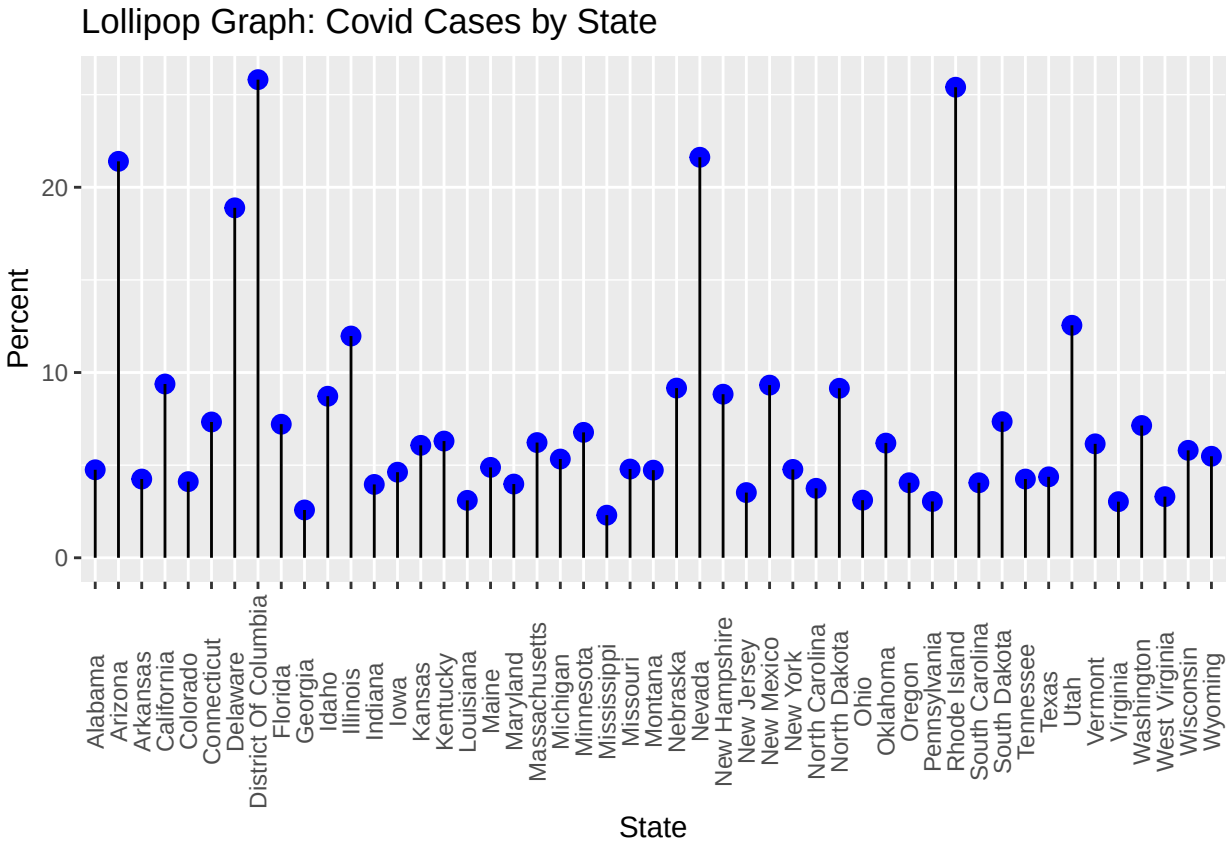
```
covid_state_summary <- covid_pop_map %>% group_by(region) %>% summarize(Percent = mean(perc_inf))
covid_state_summary$Percent <-  covid_state_summary$Percent * 100
covid_state_summary$Percent <- round(covid_state_summary$Percent, digits = 2)
covid_state_summary <- covid_state_summary %>% rename(State = region)
covid_state_summary %>% gt()
```

State	Percent
Alabama	4.75
Arizona	21.40
Arkansas	4.25
California	9.38
Colorado	4.11
Connecticut	7.33
Delaware	18.89
District Of Columbia	25.81
Florida	7.21
Georgia	2.58
Idaho	8.72
Illinois	11.97
Indiana	3.96
Iowa	4.62

Kansas	6.07
Kentucky	6.30
Louisiana	3.10
Maine	4.88
Maryland	3.98
Massachusetts	6.22
Michigan	5.33
Minnesota	6.77
Mississippi	2.30
Missouri	4.79
Montana	4.73
Nebraska	9.16
Nevada	21.62
New Hampshire	8.83
New Jersey	3.52
New Mexico	9.32
New York	4.77
North Carolina	3.75
North Dakota	9.15
Ohio	3.11
Oklahoma	6.19
Oregon	4.05
Pennsylvania	3.04
Rhode Island	25.40
South Carolina	4.05
South Dakota	7.35
Tennessee	4.25
Texas	4.37
Utah	12.55
Vermont	6.15
Virginia	3.03
Washington	7.14
West Virginia	3.30
Wisconsin	5.80
Wyoming	5.48

Now we visualize the same data using a "lollipop" graph. Here we can clearly see the states that have the highest Covid rates, Arizona, the District of Columbia, Nevada, and Rhode Island.

```
ggplot(covid_state_summary, aes(x = State, y = Percent)) +
  geom_point(size = 3, color = "blue") +
  geom_segment(aes(x = State, xend = State, y = 0, yend = Percent)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
  labs(title='Lollipop Graph: Covid Cases by State')
```



Model

Can the latitude of a state predict the number of Covid cases? The following model uses latitude to predict Covid Cases.

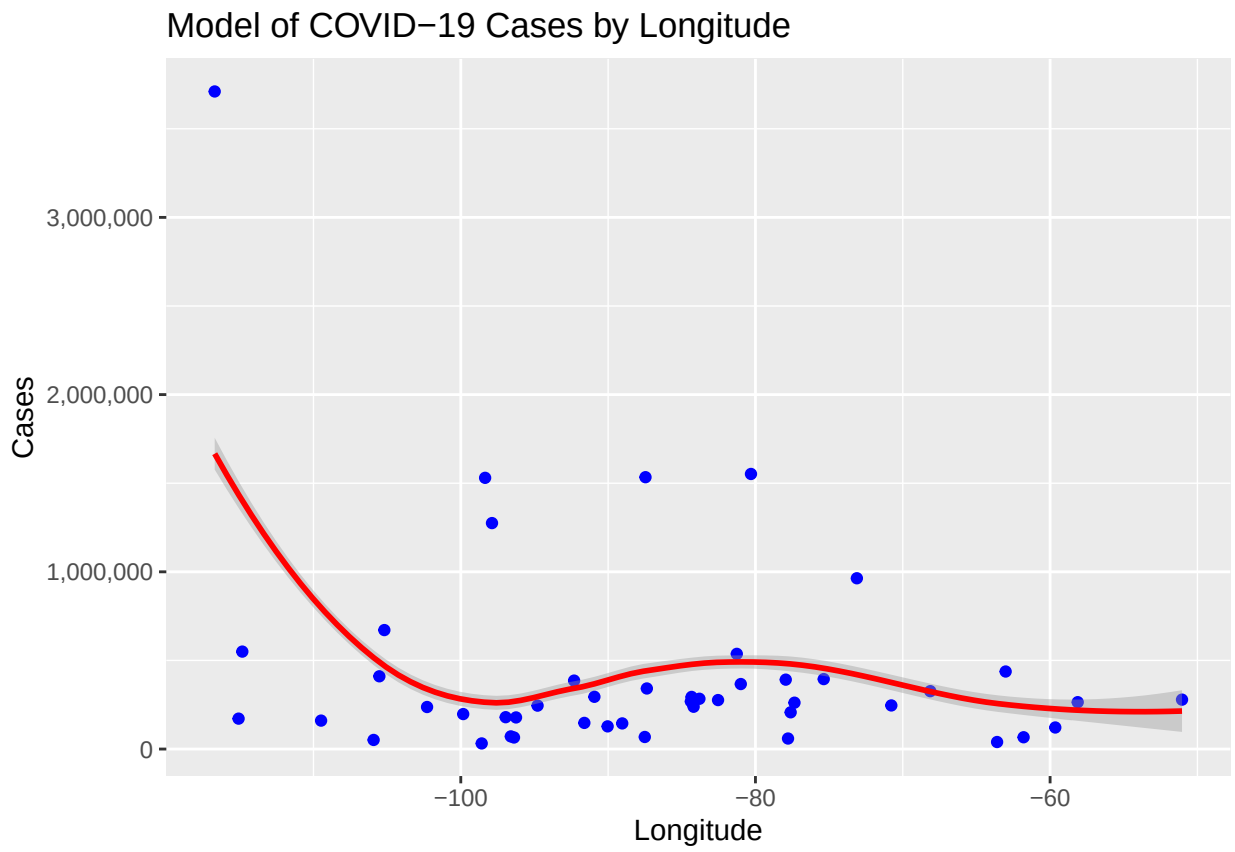
```
#filter outliers
us_cases_by_state <- us_cases_by_state %>% filter(Lat>25) %>% filter(Lat<50)
#we create a model
mod <- lm(cases ~ Long_ +  I(Long_^2) + I(Long_^3) + I(Long_^4), data = us_cases_by_state)
#show a summary of the model's performance
summary(mod)
```

```
##
## Call:
## lm(formula = cases ~ Long_ + I(Long_^2) + I(Long_^3) + I(Long_^4),
##     data = us_cases_by_state)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1286590  -217669  -103007    37357  1883774
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.468e+07  4.061e+07   1.593   0.1188
## Long_       3.462e+06  2.029e+06   1.706   0.0954 .
```

```
## I(Long_^2)  6.801e+04  3.724e+04   1.827   0.0749 .
## I(Long_^3)  5.793e+02  2.977e+02   1.946   0.0584 .
## I(Long_^4)  1.810e+00  8.764e-01   2.065   0.0451 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 554100 on 42 degrees of freedom
## Multiple R-squared:  0.2724, Adjusted R-squared:  0.2031
## F-statistic:  3.93 on 4 and 42 DF,  p-value: 0.008462
```

```
#let's add a new column to show prediction based on month
lon_pred <- us_cases_by_state %>% mutate(pred = predict(mod))



ggplot(data=lon_pred) + geom_point(aes(x = Long_, y=cases),color='blue') +
  geom_smooth(aes(x=Long_, y=pred),color='red') + labs(title='Model of COVID-19 Cases by Longitude',x='
```

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



In the graph we see a bump on the left side to accommodate the large number of cases in California, Arizona, Nevada. Overall the model is accurate with a p-value of 0.008.

Conclusions

As expected states with the highest population also have the highest number of Covid cases. However, when analyzing the percentage of the population with Covid the picture is more varied. Centers of high rates of the disease exist both in the Southwest US with Arizona, Nevada, and Utah at very high rates, as well as on the East Coast centering around the District of Columbia, Rhode Island, and Delaware. It would be interesting to compare the states with high and low levels of Covid at a deeper level. Are there similarities and differences between these two centers? What factors might contribute to these higher or lower rates?

Bias

One possible source of bias within my report would be my geographical location. As a resident of California (which has the highest raw number of cases) I need to guard against my personal experience of the disease effecting my judgement. I believe I was succesful in an objective analysis, not showing favor to any particular geographic region.