Analysis of COVID-19 cases and deaths in the U.S.

2023-02-24

Contents

Introduction

Data source

The data used in this analysis is provided by Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). It is connected with an article in The Lancet titled "An interactive web-based dashboard to track COVID-19 in real time" which can be found at: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30120-1/fulltext.
The data repository is maintained for the 2019 Novel Coronavirus Visual Dashboard that is maintained by JHU CSSE. The data files can be found at: https://github.com/CSSEGISandData/COVID-19. The data sets specifically used here come from the time series folder in the repository and include both the U.S. confirmed case and death counts.

Retrieving the data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
```

```
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.4     v forcats 1.0.0
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
url_root <- "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time
filenames <- c("time_series_covid19_confirmed_US.csv", "time_series_covid19_deaths_US.csv")
urls <- str_c(url_root, filenames)
init_us_cases <- read_csv(urls[1])
```

```
## Rows: 3342 Columns: 1142
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1136): 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.
```

```
init_us_deaths <- read_csv(urls[2])
```

```
## Rows: 3342 Columns: 1143
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1137): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## 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.
```

Data description

The confirmed cases and COVID deaths data sets include daily listings starting in January 2020 and detailed per U.S. state and county.

```
head(init_us_cases, 10)
```

```
## # A tibble: 10 x 1,142
##         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 1,131 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>, `2/16/20` <dbl>, ...
```

```
head(init_us_deaths, 10)
```

```
## # A tibble: 10 x 1,143
##         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 1,132 more variables: Population <dbl>, `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>, ...
```

Goals of the analysis

Goals of this analysis include:
1. Visualize the overall trend of confirmed cases and deaths
2. Zoom in on a specific state to see its trend
3. Determine the U.S. states with the highest and lowest rates
4. Model the relationship between cases and deaths
5. Visualize the total cases across the U.S. geographic regions and divisions
6. Model the relationship between case rate and political bias
7. Model the relationship between case rate and state GDP

Pivoting and joining the data

The data is provided in "wide" format and is converted to "long" format here. Additionally, unnecessary columns are removed. The confirmed cases and number of deaths combined into a single data frame so they can be analyzed together.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
us_cases <- init_us_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
```

```
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long_))

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

```
## Joining with `by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)`
```

Initial Analysis and Visualization

Viewing U.S. totals

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

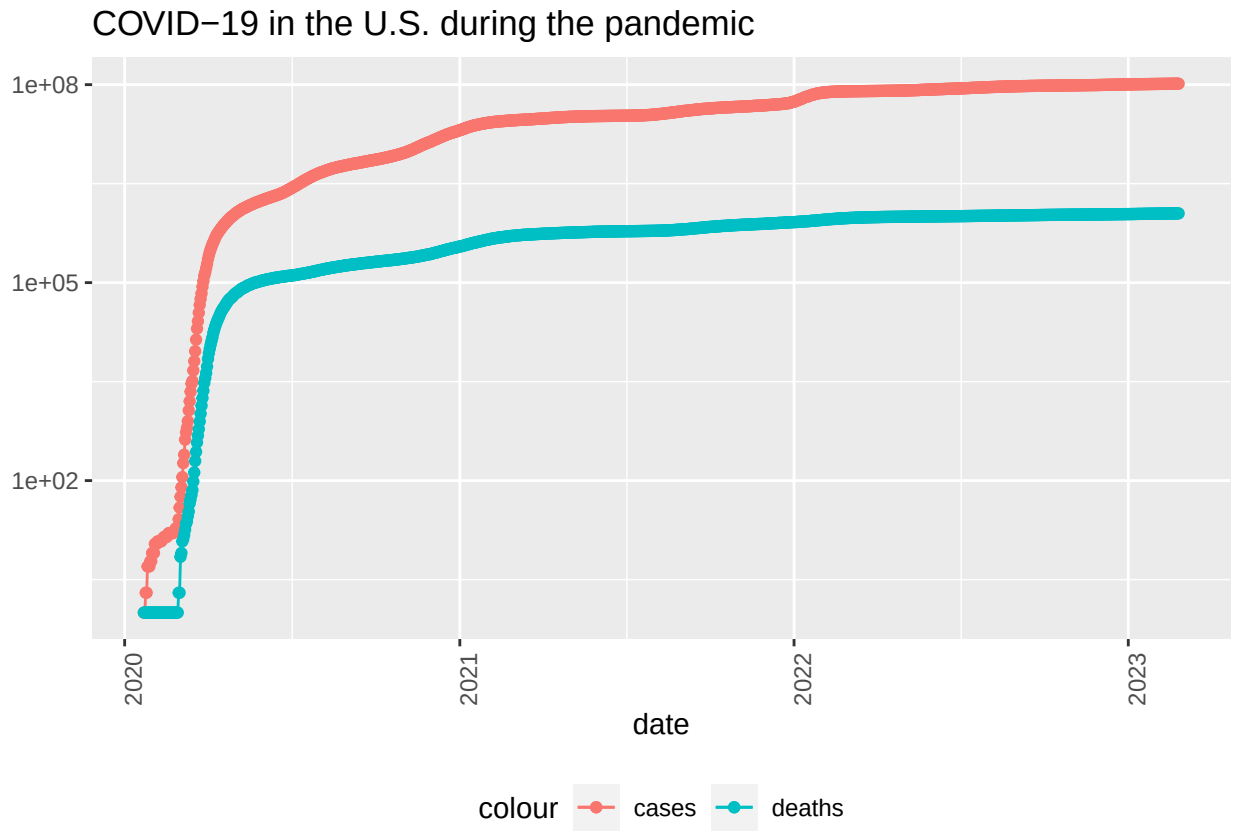
```
## `summarise()` has grouped output by 'Province_State', 'Country_Region'. You can
## override using the `.groups` argument.
```

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

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

```
US_totals %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
```

```
  labs(title = "COVID-19 in the U.S. during the pandemic", y=NULL)
```



Determining how infection spread over time

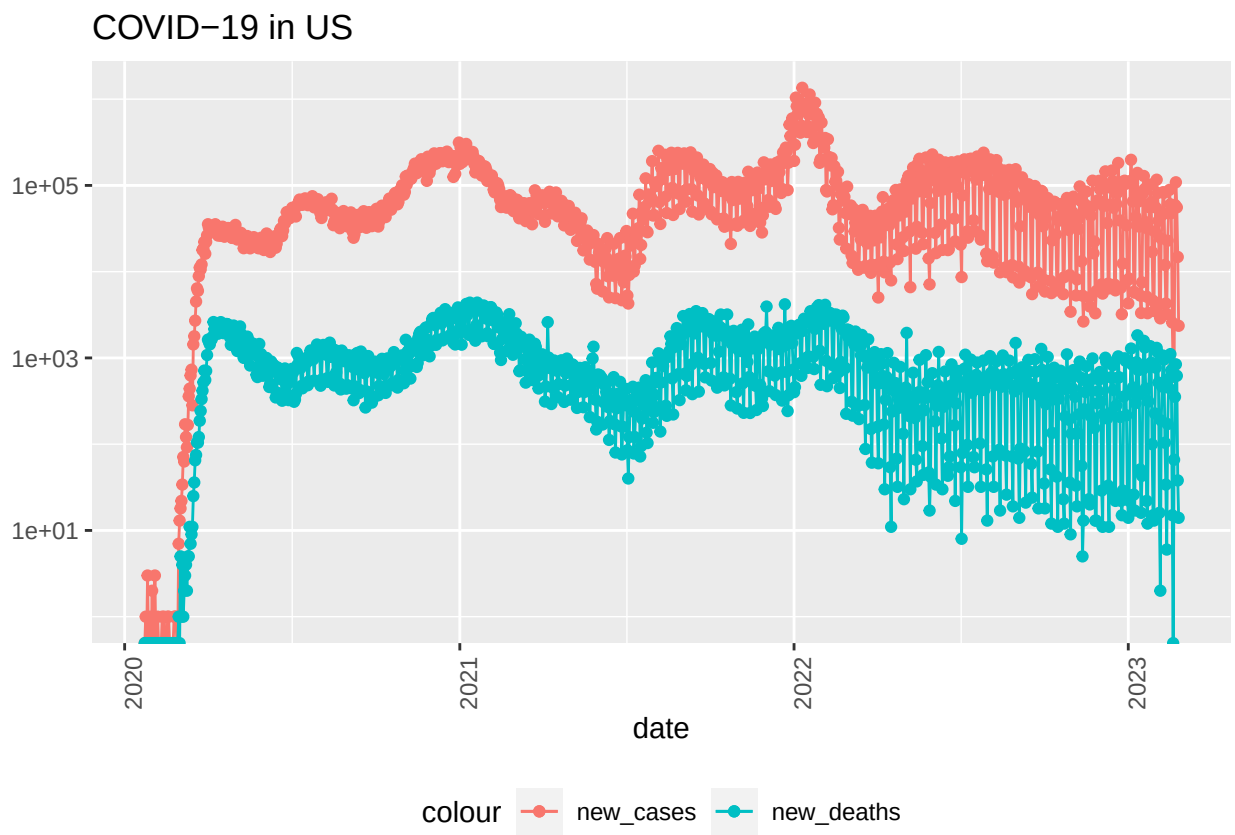
Building out data by U.S. state

```
 US_by_state <- US_by_state %>%
   mutate(new_cases = cases - lag(cases),
          new_deaths = deaths - lag(deaths))
US_totals <- US_totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))
tail(US_totals %>% select(new_cases, new_deaths, everything()))
```

```
## # A tibble: 6 x 8
##   new_cases new_deaths Country_Region date          cases deaths death~1 Popul~2
##       <dbl>      <dbl> <chr>          <date>        <dbl>  <dbl>   <dbl>   <dbl>
## 1       602         66 US             2023-02-20   1.03e8 1.12e6   3358.  3.33e8
## 2     59278        354 US             2023-02-21   1.03e8 1.12e6   3359.  3.33e8
## 3    108723        843 US             2023-02-22   1.03e8 1.12e6   3361.  3.33e8
## 4     55659        621 US             2023-02-23   1.03e8 1.12e6   3363.  3.33e8
## 5     14800         38 US             2023-02-24   1.03e8 1.12e6   3363.  3.33e8
## 6      2359         14 US             2023-02-25   1.03e8 1.12e6   3363.  3.33e8
## # ... with abbreviated variable names 1: deaths_per_mil, 2: Population
```

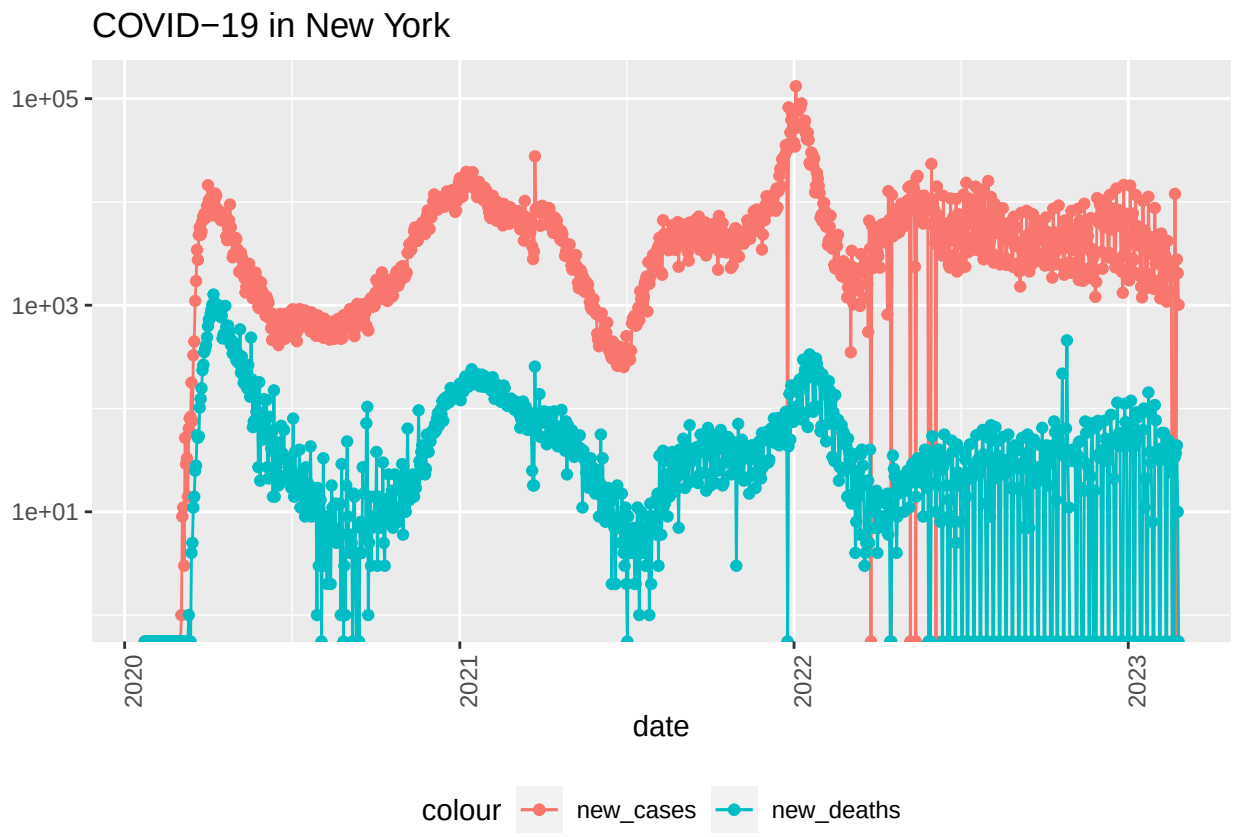
```
US_totals %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "COVID-19 in US", y=NULL)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
## Warning in self$trans$transform(x): NaNs produced
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning in self$trans$transform(x): NaNs produced
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 1 row containing missing values (`geom_line()`).
## Warning: Removed 1 rows containing missing values (`geom_point()`).
## Warning: Removed 1 row containing missing values (`geom_line()`).
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



```
state <- "New York"
US_by_state %>%
  filter(Province_State == state) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = str_c("COVID-19 in ", state), y=NULL)
```

```
## Warning in self$trans$transform(x): NaNs produced
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning in self$trans$transform(x): NaNs produced
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning in self$trans$transform(x): NaNs produced
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning in self$trans$transform(x): NaNs produced
## Warning: Transformation introduced infinite values in continuous y-axis
## Warning: Removed 1 row containing missing values (`geom_line()`).
## Warning: Removed 1 rows containing missing values (`geom_point()`).
## Warning: Removed 1 row containing missing values (`geom_line()`).
## Warning: Removed 8 rows containing missing values (`geom_point()`).
```



Normalizing cases and deaths by population

```
US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thou = 1000 * cases / population,
            deaths_per_thou = 1000 * deaths / population) %>%
  filter(cases > 0, population > 0)
```

The following is a view of the 10 lowest states and 10 highest states as ranked by number of deaths per thousand.

```
US_state_totals %>% slice_min(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State          deaths  cases popul~1
##              <dbl>          <dbl> <chr>                    <dbl>  <dbl>   <dbl>
##  1           0.611           150. American Samoa              34 8.32e3   55641
##  2           0.744           247. Northern Mariana Islands    41 1.36e4   55144
##  3           1.21            231. Virgin Islands             130 2.48e4  107268
##  4           1.29            268. Hawaii                    1822 3.79e5 1415872
##  5           1.46            243. Vermont                    910 1.51e5  623989
##  6           1.54            292. Puerto Rico               5791 1.10e6 3754939
##  7           1.65            339. Utah                      5276 1.09e6 3205958
```

```
##  8           2.01            414. Alaska                     1486 3.07e5  740995
##  9           2.02            252. District of Columbia       1427 1.78e5  705749
## 10           2.05            253. Washington                15622 1.92e6 7614893
## # ... with abbreviated variable name 1: population
```

```
US_state_totals %>% slice_max(deaths_per_thou, n=10) %>%
  select(deaths_per_thou, cases_per_thou, everything())
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State deaths   cases population
##              <dbl>          <dbl> <chr>           <dbl>   <dbl>      <dbl>
##  1            4.54           334. Arizona         33042 2434631    7278717
##  2            4.52           325. Oklahoma        17887 1284450    3956971
##  3            4.48           332. Mississippi     13320  987105    2976149
##  4            4.42           357. West Virginia    7926  639246    1792147
##  5            4.30           319. New Mexico       9020  668677    2096829
##  6            4.30           333. Arkansas        12975 1004294    3017804
##  7            4.27           367. Tennessee       29169 2503667    6829174
##  8            4.27           334. Alabama         20932 1638348    4903185
##  9            4.20           305. Michigan        41957 3049739    9986857
## 10            4.05           342. New Jersey      35950 3038713    8882190
```

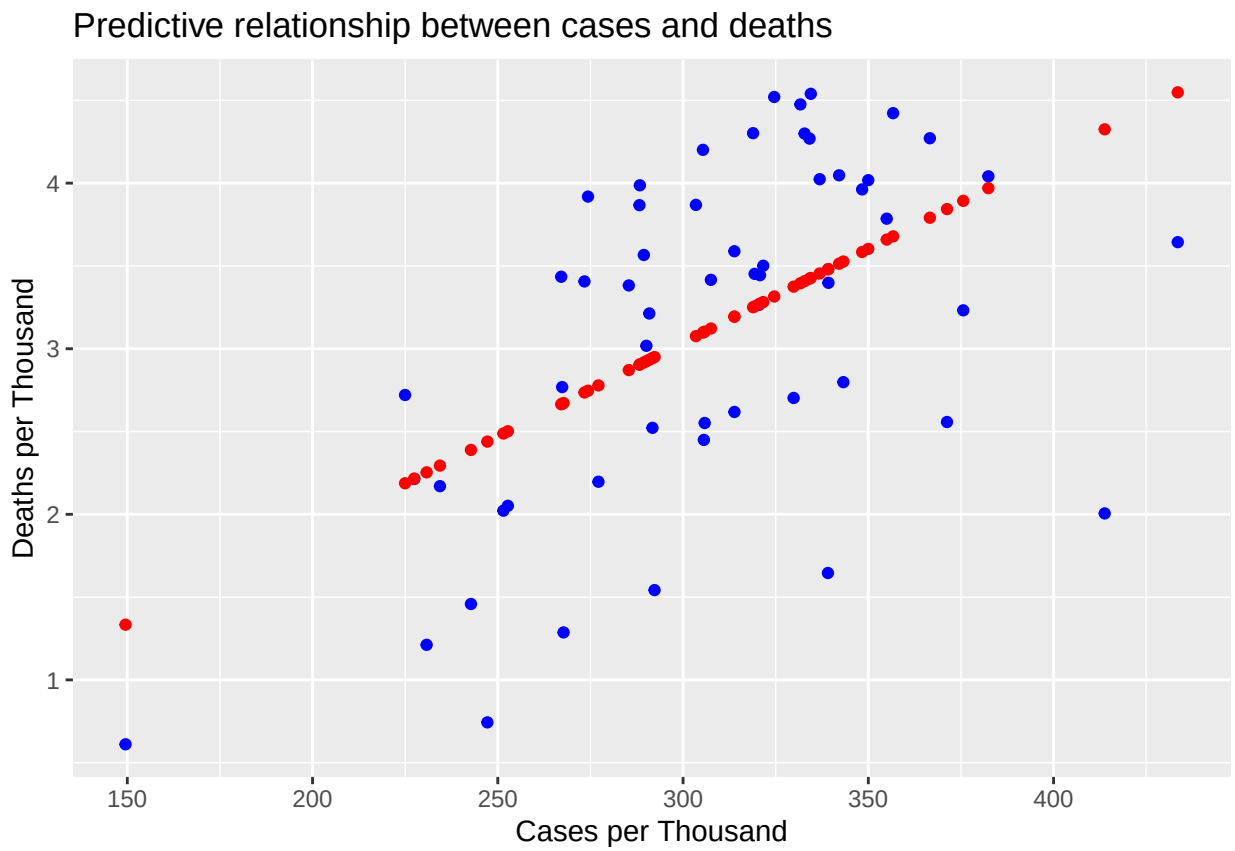
Predictive model for relationship between cases and deaths

The following visual shows the relationship between the number of deaths per thousand and cases per thousand.

```
mod <- lm(deaths_per_thou ~ cases_per_thou, data = US_state_totals)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3198 -0.5990  0.1492  0.6545  1.2049
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.359290   0.722884  -0.497    0.621
## cases_per_thou  0.011321   0.002323   4.873    1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8594 on 54 degrees of freedom
## Multiple R-squared:  0.3054, Adjusted R-squared:  0.2925
## F-statistic: 23.74 on 1 and 54 DF,  p-value: 1.004e-05
```

```
US_tot_w_pred <- US_state_totals %>% mutate(pred = predict(mod))
US_tot_w_pred %>% ggplot() +
  geom_point(aes(x = cases_per_thou, y = deaths_per_thou), color = "blue") +
  geom_point(aes(x = cases_per_thou, y = pred), color = "red") +
  labs(title = "Predictive relationship between cases and deaths",
       x = "Cases per Thousand", y = "Deaths per Thousand")
```



U.S. States by Region and Division

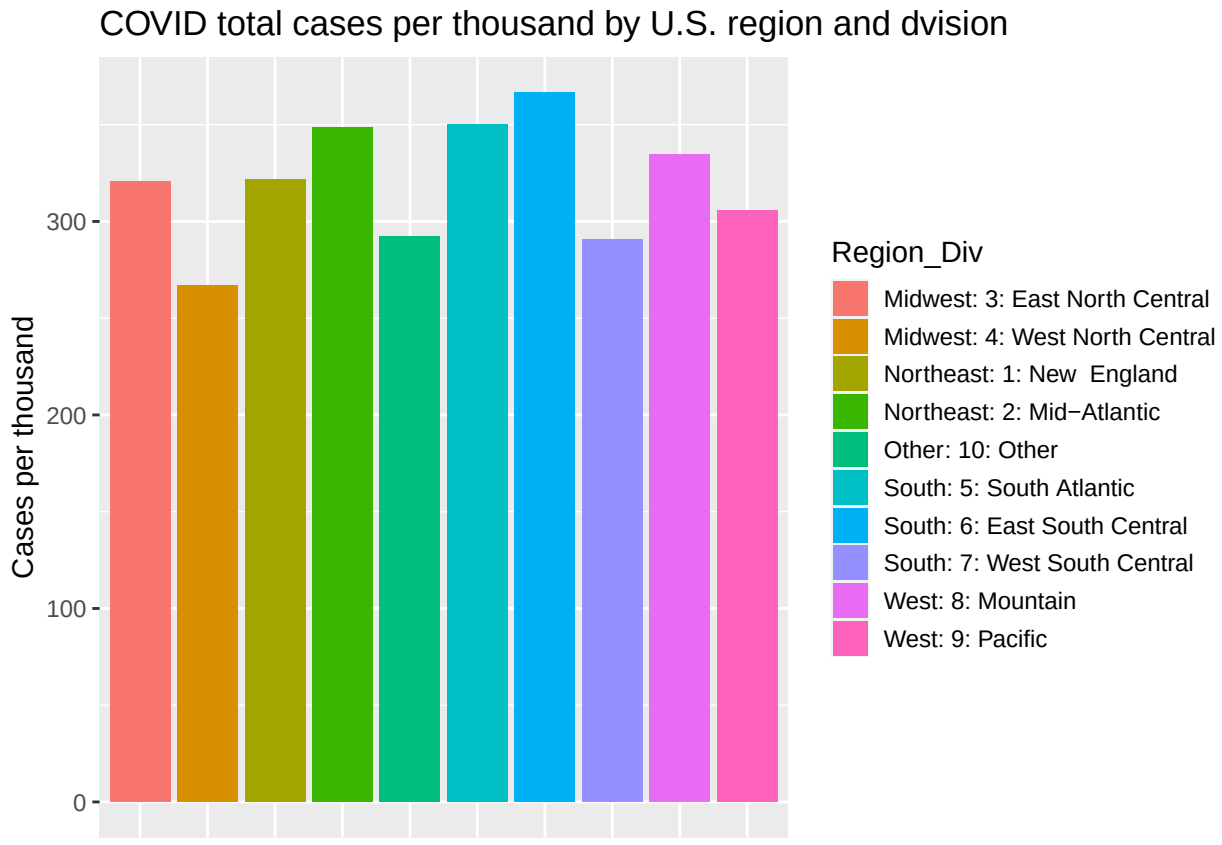
The U.S. Census Bureau divides the U.S. geographically into four regions and nine divisions. Here, the data is grouped and summarized into the cases and deaths per thousand per geographic region and division.

```
# Merge in the region / division data
US_states_by_region_div <- read_csv("US_regions_and_divisions.csv")
```

```
## Rows: 56 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Region, Division, State, ABV
## dbl (2): RegionNum, DivisionNum
##
## 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_by_state <- US_by_state %>% rename("State" = "Province_State")
US_by_state <- merge(US_by_state, US_states_by_region_div, by="State")

# Combine the region and division columns
US_by_state <- US_by_state %>%
  #mutate(region = get_US_region[Province_State]) %>%
  #mutate(division = get_US_division[Province_State]) %>%
  #mutate(ABV = get_abv[Province_State]) %>%
  unite("Region_Div",
```

```
        Region:Division,
        sep = ": ",
        na.rm = TRUE,
        remove = FALSE)

US_by_state$Region_Div[US_by_state$Region_Div == ""] <- "Other"

US_regionDiv_totals <- US_by_state %>%
  group_by(Region_Div) %>%
  summarize(deaths = max(deaths), cases = max(cases),
            population = max(Population),
            cases_per_thou = 1000 * cases / population,
            deaths_per_thou = 1000 * deaths / population) %>%
  filter(cases > 0, population > 0)

US_regionDiv_totals %>%
  ggplot(aes(x = Region_Div, y = cases_per_thou, fill = Region_Div)) +
    geom_bar(stat = "identity") +
    labs(title = "COVID total cases per thousand by U.S. region and dvision",
         x = NULL, y = "Cases per thousand") +
    theme(axis.text.x=element_blank(), axis.ticks.x=element_blank(),
          legend.key.height = unit(0.5, 'cm'),
          legend.key.width = unit(0.5, 'cm'))
```



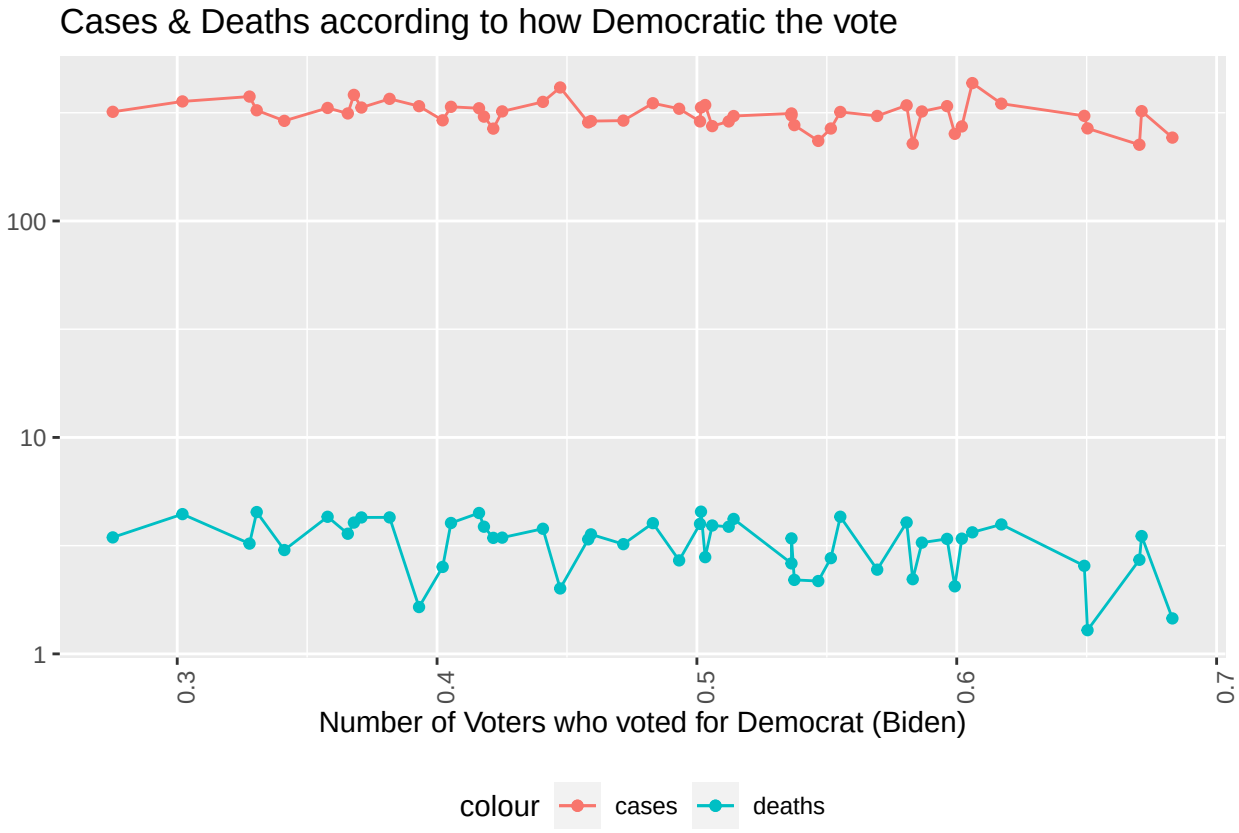
The chart above shows that, when normalized by population, there was not a significant amount of variation across geographic regions and divisions. # Using 2020 Presidential Election results
The results of the 2020 presidential election are available through many sources online. One example is here,

https://www.cookpolitical.com/2020-national-popular-vote-tracker. The following examines the relationship between case and deaths rates and the popular vote results. Specifically, it looks at how states fared that voted more or less Democratic (voted for Joe Biden).

```
US_state_totals_w_geopol <- US_state_totals %>% rename("State" = "Province_State")
US_state_totals_w_geopol <- merge(US_state_totals_w_geopol, US_states_by_region_div, by="State")
US_state_totals_w_geopol <- US_state_totals_w_geopol %>%
  unite("Region_Div",
        Region:Division,
        sep = ": ",
        na.rm = TRUE,
        remove = FALSE)

election_results <- read_csv("2020presgeresults.csv")
```

```
## Rows: 57 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ABV
## num (3): BIDEN, TRUMP, TotalVotes
##
## 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.
```

```
election_results <- election_results %>% drop_na()
US_state_totals_w_geopol <- merge(US_state_totals_w_geopol, election_results, by="ABV")

US_state_totals_w_geopol <- US_state_totals_w_geopol %>%
  mutate(DemVotes = BIDEN/TotalVotes) %>%
  drop_na()

US_state_totals_w_geopol %>%
  ggplot(aes(x = DemVotes, y = cases_per_thou)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths_per_thou, color = "deaths")) +
  geom_point(aes(y = deaths_per_thou, color = "deaths")) +
  scale_y_log10() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Cases & Deaths according to how Democratic the vote",
       x = "Number of Voters who voted for Democrat (Biden)", y=NULL)
```



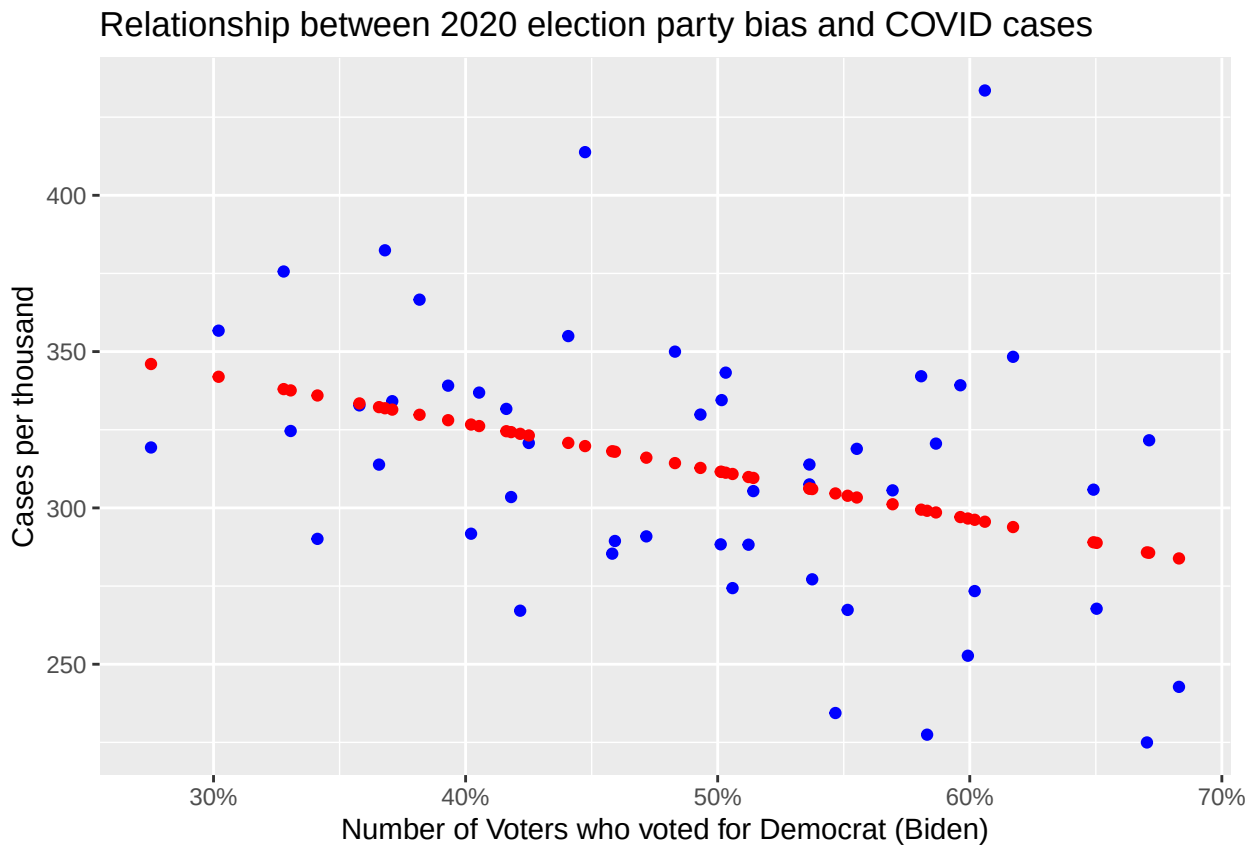
Model of Election Data

```
case_model <- lm(cases_per_thou ~ DemVotes, data = US_state_totals_w_geopol)
summary(case_model)
```

```
##
## Call:
## lm(formula = cases_per_thou ~ DemVotes, data = US_state_totals_w_geopol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.614 -28.091   0.341  22.738 137.937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   388.01      27.68  14.016  < 2e-16 ***
## DemVotes     -152.51      55.20  -2.763  0.00809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.95 on 48 degrees of freedom
## Multiple R-squared:  0.1372, Adjusted R-squared:  0.1193
## F-statistic: 7.635 on 1 and 48 DF,  p-value: 0.008095
```

```
case_pred <- US_state_totals_w_geopol %>% mutate(pred = predict(case_model))
case_pred %>% ggplot() +
  geom_point(aes(x = DemVotes, y = cases_per_thou), color = "blue") +
  geom_point(aes(x = DemVotes, y = pred), color = "red") +
```

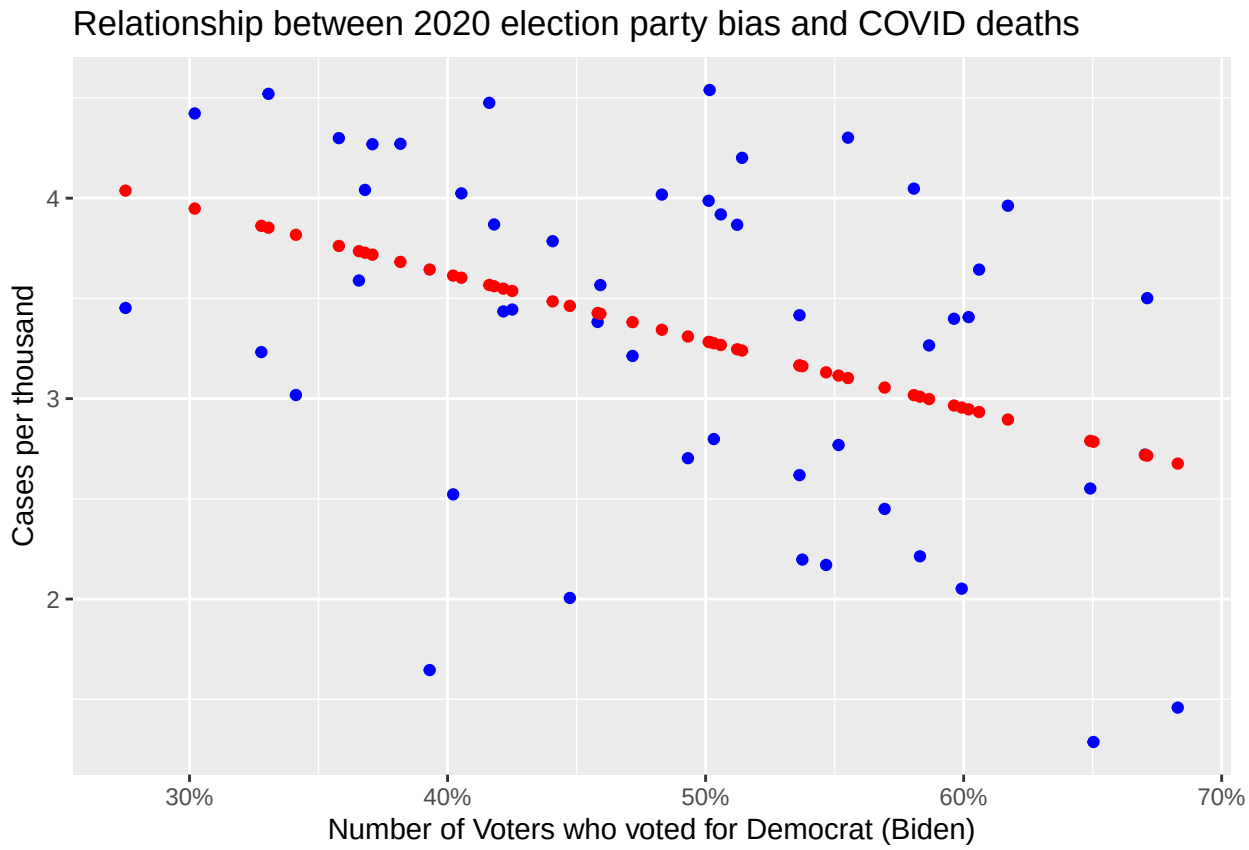
```
  scale_x_continuous(labels = scales::percent) +
  labs(title = "Relationship between 2020 election party bias and COVID cases",
       x = "Number of Voters who voted for Democrat (Biden)", y = "Cases per thousand") +
  theme(legend.position = "bottom")
```




```
death_model <- lm(deaths_per_thou ~ DemVotes, data = US_state_totals_w_geopol)
summary(death_model)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ DemVotes, data = US_state_totals_w_geopol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9984 -0.6006  0.1969  0.6130  1.2579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.9567     0.5318   9.320 2.41e-12 ***
## DemVotes     -3.3396     1.0604  -3.149  0.00281 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7867 on 48 degrees of freedom
## Multiple R-squared:  0.1713, Adjusted R-squared:  0.154
## F-statistic: 9.919 on 1 and 48 DF,  p-value: 0.002814
```

```
death_pred <- US_state_totals_w_geopol %>% mutate(pred = predict(death_model))
death_pred %>% ggplot() +
  geom_point(aes(x = DemVotes, y = deaths_per_thou), color = "blue") +
  geom_point(aes(x = DemVotes, y = pred), color = "red") +
  scale_x_continuous(labels = scales::percent) +
  labs(title = "Relationship between 2020 election party bias and COVID deaths",
       x = "Number of Voters who voted for Democrat (Biden)", y = "Cases per thousand") +
  theme(legend.position = "bottom")
```

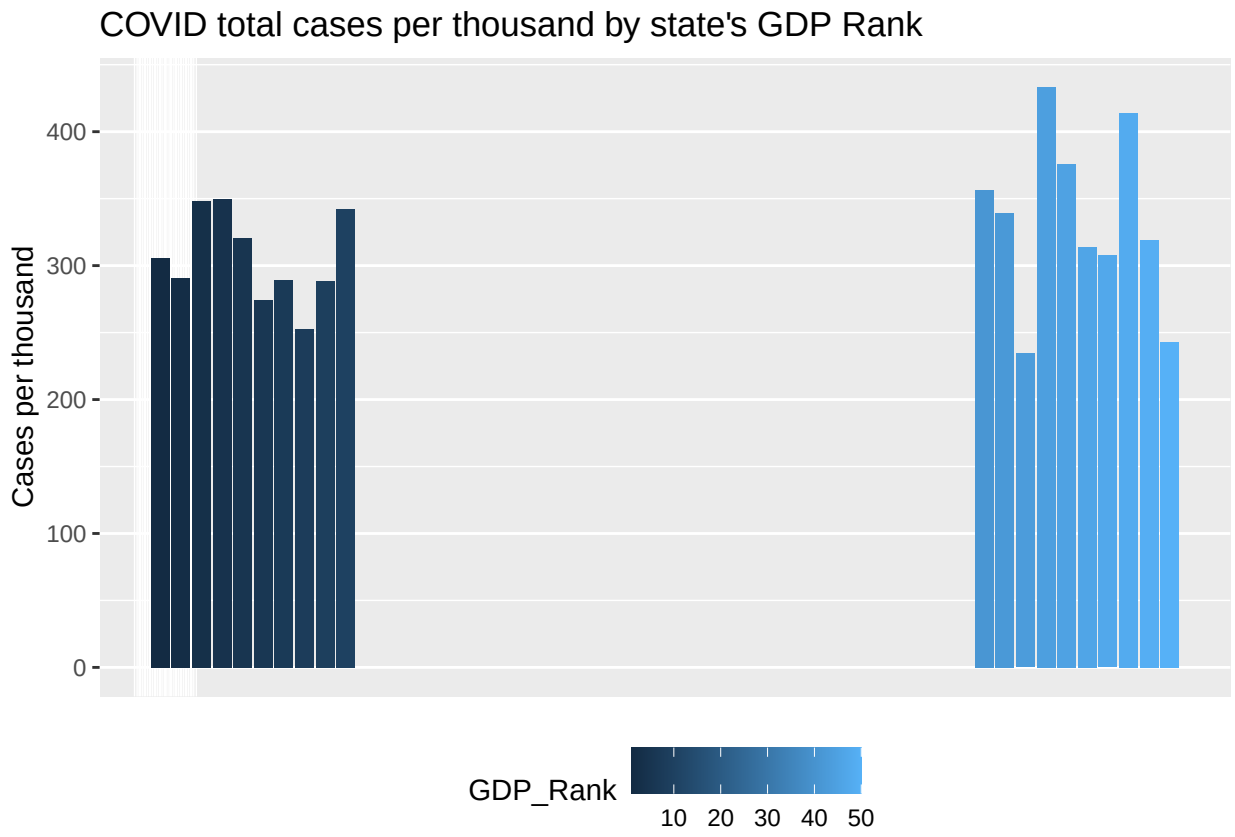


Economic base

The GDP per state and GDP per capita per state is available in many places online such as at: https://www.statista.com/statistics/248063/per-capita-us-real-gross-domestic-product-gdp-by-state/. Here, the analysis looks at the COVID case and death rates in relation to state GDP. The top 10 and lowest 10 states are examined to see if richer or poorer states fared better or worse through the pandemic.

```
top_states_by_gdp <- as.factor(
  c("California", "Texas", "New York", "Florida", "Illinois",
    "Pennsylvania", "Ohio", "Washington", "Georgia", "New Jersey"))
bottom_states_by_gdp <- as.factor(
  c("West Virginia", "Delaware", "Maine", "Rhode Island", "North Dakota",
    "South Dakota", "Montana", "Alaska", "Wyoming", "Vermont"))
top_states_gdp_per_capita <- as.factor(
  c("New York", "Massachusetts", "Washington", "Connecticut", "California",
    "Delaware", "Illinois", "Aalska", "Maryland", "North Dakota"))
```

```
bottom_states_gdp_per_capita <- as.factor(
  c("Maine", "Montana", "New Mexico", "Kentucky", "South Carolina",
    "Idaho", "Alabama", "Arkansas", "West Virginia", "Mississippi"))
top_states_df <- data.frame(GDP_Rank = c(1:10), State = top_states_by_gdp)
bottom_states_df <- data.frame(GDP_Rank = c(41:50), State = bottom_states_by_gdp)
states_by_gdp <- rbind(top_states_df, bottom_states_df)

US_state_totals_w_econ <- US_state_totals %>% rename("State" = "Province_State")
US_state_totals_w_econ <- merge(US_state_totals_w_econ, states_by_gdp, by="State")

US_state_totals_w_econ %>%
  ggplot(aes(x = GDP_Rank, y = cases_per_thou, fill = GDP_Rank)) +
    geom_bar(stat = "identity") +
    labs(title = "COVID total cases per thousand by state's GDP Rank",
         x = NULL, y = "Cases per thousand") +
     scale_x_continuous(breaks = seq(from = 0, to = 2.5, by = 0.25)) +
    theme(legend.position = "bottom",
          axis.text.x=element_blank(), axis.ticks.x=element_blank())
```



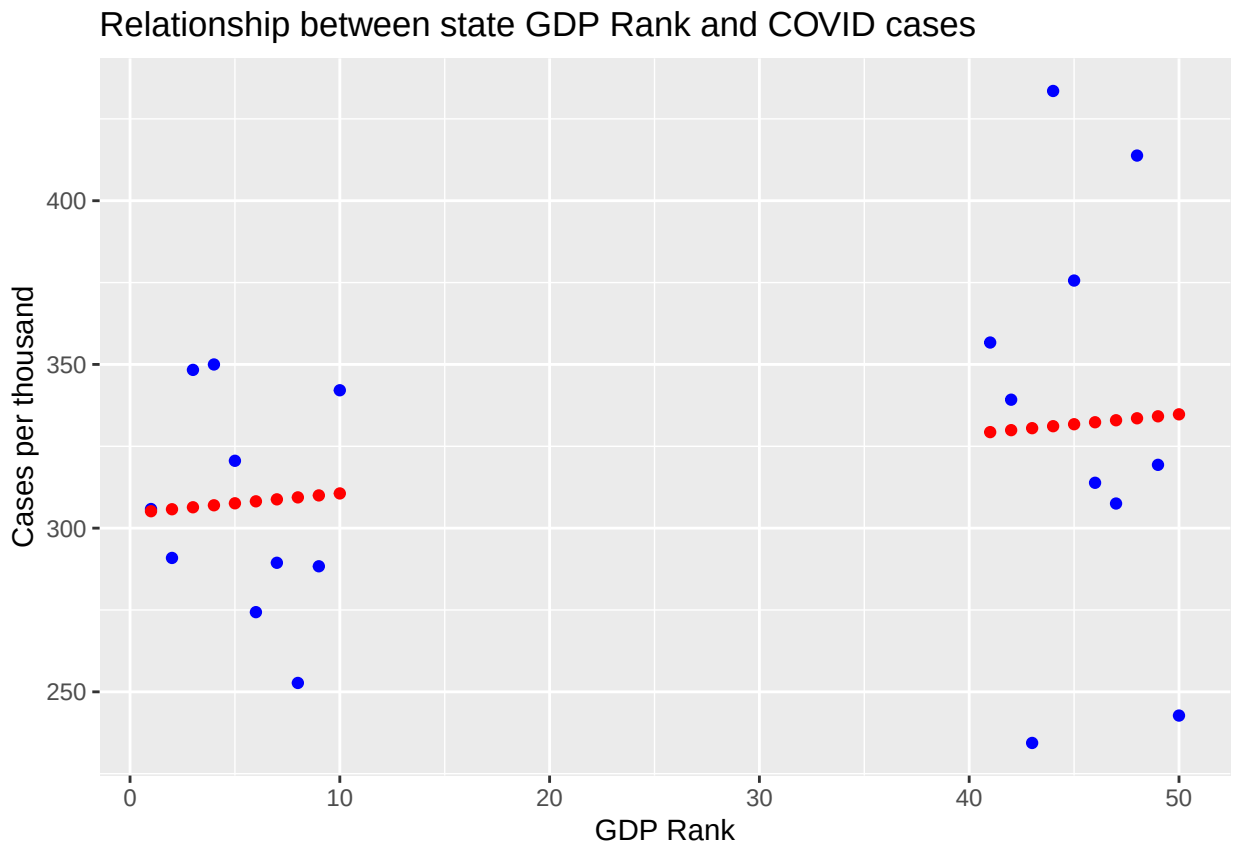
Predictive model for relationship between GDP and COVID case rate

```
econ_model <- lm(cases_per_thou ~ GDP_Rank, data = US_state_totals_w_econ)
summary(econ_model)
```

```
##
```

```
## Call:
## lm(formula = cases_per_thou ~ GDP_Rank, data = US_state_totals_w_econ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.143 -22.598  -7.077  34.122 102.387
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 304.5645    18.7466  16.246 3.37e-12 ***
## GDP_Rank      0.6039     0.5762   1.048    0.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.07 on 18 degrees of freedom
## Multiple R-squared:  0.05751,    Adjusted R-squared:  0.00515
## F-statistic: 1.098 on 1 and 18 DF,  p-value: 0.3085
```

```
econ_pred <- US_state_totals_w_econ %>% mutate(pred = predict(econ_model))
econ_pred %>% ggplot() +
  geom_point(aes(x = GDP_Rank, y = cases_per_thou), color = "blue") +
  geom_point(aes(x = GDP_Rank, y = pred), color = "red") +
  #scale_x_continuous(breaks = c(1:10, 41:50)) +
  labs(title = "Relationship between state GDP Rank and COVID cases",
       x = "GDP Rank", y = "Cases per thousand") +
  theme(legend.position = "bottom")
```



The above chart and model show that there is essentially no relationship between case and deaths rates with respect to state GDP.

Conclusion

Summary

This analysis has found that the number of cases of COVID initially skyrocketed but has largely flattened out over the past year. The rates in New York are very similar to the rest of the country. There is a loose relationship between the number of confirmed cases and the death rate. Looking at the data geographically, the cases per thousand were slightly higher in some the South Region's divisions, notably the division containing Kentucky, Tennessee, Mississippi, and Alabama. Using data from the 2020 Presidential election, a weak political bias can be seen in which states that lean more Democratic and less Republican were likely to have somewhat lower confirmed cases and deaths. Finally, the analysis has shown that a state's GDP is a weak indicator of case rates. Specifically, lower GDPs also have lower case rates.

Sources of Bias

Sources of bias in COVID reporting data have been heavily discussed in the media. There is inconsistency across states and hospitals as to when and how COVID is reported. For example, listing a COVID death when there is a co-morbidity has been a key controversy.