Final_Part_2

Imran Gosla

11/21/2021

Statement of Question of Interest:

Which out of the three states: California, Texas, and New York had the overall worst experience with handling covid?

Get the current data

The source we are getting the data from is from the center for Systems Science and Engineering (CSSE) at Johns Hopkins University. The data sources they used are scattered across the country based on county, and state health department. I am deciding to bring in both the US and Global datasets just for referencial purposes.

```
url_in_1 <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
url_in_2 <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
url_in_3 <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
url_in_4 <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
```

Read in the data set and show summary

```
covid_us_deaths <- read_csv(url_in_1)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Combined_Key = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
covid_global_deaths <- read_csv(url_in_2)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ‘Province/State‘ = col_character(),
##   ‘Country/Region‘ = col_character()
## )
## See spec(...) for full column specifications.
```

```
covid_us_cases <- read_csv(url_in_3)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   iso2 = col_character(),
##   iso3 = col_character(),
##   Admin2 = col_character(),
##   Province_State = col_character(),
##   Country_Region = col_character(),
##   Combined_Key = col_character()
## )
## See spec(...) for full column specifications.
```

```
covid_global_cases <- read_csv(url_in_4)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ‘Province/State‘ = col_character(),
##   ‘Country/Region‘ = col_character()
## )
## See spec(...) for full column specifications.
```

View the first couple rows of the tables

```
head(covid_us_deaths)
```

```
## # A tibble: 6 x 681
##       UID iso2  iso3  code3  FIPS Admin2 Province_State Country_Region   Lat
##     <dbl> <chr> <chr> <dbl> <dbl> <chr>  <chr>          <chr>          <dbl>
## 1 8.40e7 US    USA     840  1001 Autau~ Alabama        US              32.5
## 2 8.40e7 US    USA     840  1003 Baldw~ Alabama        US              30.7
## 3 8.40e7 US    USA     840  1005 Barbo~ Alabama        US              31.9
## 4 8.40e7 US    USA     840  1007 Bibb   Alabama        US              33.0
## 5 8.40e7 US    USA     840  1009 Blount Alabama        US              34.0
## 6 8.40e7 US    USA     840  1011 Bullo~ Alabama        US              32.1
## # ... with 672 more variables: Long_ <dbl>, Combined_Key <chr>,
## #   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>, ‘2/16/20‘ <dbl>, ‘2/17/20‘ <dbl>,
## #   ‘2/18/20‘ <dbl>, ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ‘2/21/20‘ <dbl>,
## #   ‘2/22/20‘ <dbl>, ‘2/23/20‘ <dbl>, ‘2/24/20‘ <dbl>, ‘2/25/20‘ <dbl>,
## #   ‘2/26/20‘ <dbl>, ‘2/27/20‘ <dbl>, ‘2/28/20‘ <dbl>, ‘2/29/20‘ <dbl>,
## #   ‘3/1/20‘ <dbl>, ‘3/2/20‘ <dbl>, ‘3/3/20‘ <dbl>, ‘3/4/20‘ <dbl>,
## #   ‘3/5/20‘ <dbl>, ‘3/6/20‘ <dbl>, ‘3/7/20‘ <dbl>, ‘3/8/20‘ <dbl>,
## #   ‘3/9/20‘ <dbl>, ‘3/10/20‘ <dbl>, ‘3/11/20‘ <dbl>, ‘3/12/20‘ <dbl>,
## #   ‘3/13/20‘ <dbl>, ‘3/14/20‘ <dbl>, ‘3/15/20‘ <dbl>, ‘3/16/20‘ <dbl>,
## #   ‘3/17/20‘ <dbl>, ‘3/18/20‘ <dbl>, ‘3/19/20‘ <dbl>, ‘3/20/20‘ <dbl>,
## #   ‘3/21/20‘ <dbl>, ‘3/22/20‘ <dbl>, ‘3/23/20‘ <dbl>, ‘3/24/20‘ <dbl>,
## #   ‘3/25/20‘ <dbl>, ‘3/26/20‘ <dbl>, ‘3/27/20‘ <dbl>, ‘3/28/20‘ <dbl>,
## #   ‘3/29/20‘ <dbl>, ‘3/30/20‘ <dbl>, ‘3/31/20‘ <dbl>, ‘4/1/20‘ <dbl>,
## #   ‘4/2/20‘ <dbl>, ‘4/3/20‘ <dbl>, ‘4/4/20‘ <dbl>, ‘4/5/20‘ <dbl>,
## #   ‘4/6/20‘ <dbl>, ‘4/7/20‘ <dbl>, ‘4/8/20‘ <dbl>, ‘4/9/20‘ <dbl>,
## #   ‘4/10/20‘ <dbl>, ‘4/11/20‘ <dbl>, ‘4/12/20‘ <dbl>, ‘4/13/20‘ <dbl>,
## #   ‘4/14/20‘ <dbl>, ‘4/15/20‘ <dbl>, ‘4/16/20‘ <dbl>, ‘4/17/20‘ <dbl>,
## #   ‘4/18/20‘ <dbl>, ‘4/19/20‘ <dbl>, ‘4/20/20‘ <dbl>, ‘4/21/20‘ <dbl>,
## #   ‘4/22/20‘ <dbl>, ‘4/23/20‘ <dbl>, ‘4/24/20‘ <dbl>, ‘4/25/20‘ <dbl>,
## #   ‘4/26/20‘ <dbl>, ‘4/27/20‘ <dbl>, ...
```

```
head(covid_global_deaths)
```

```
## # A tibble: 6 x 673
##   ‘Province/State‘ ‘Country/Region‘   Lat   Long ‘1/22/20‘ ‘1/23/20‘ ‘1/24/20‘
##   <chr>            <chr>            <dbl>  <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9  67.7          0         0         0
## 2 <NA>             Albania           41.2  20.2          0         0         0
## 3 <NA>             Algeria           28.0   1.66         0         0         0
## 4 <NA>             Andorra           42.5   1.52         0         0         0
## 5 <NA>             Angola           -11.2  17.9          0         0         0
## 6 <NA>             Antigua and Bar~  17.1 -61.8          0         0         0
## # ... with 666 more variables: ‘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>, ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>, ‘2/19/20‘ <dbl>,
## #   ‘2/20/20‘ <dbl>, ‘2/21/20‘ <dbl>, ‘2/22/20‘ <dbl>, ‘2/23/20‘ <dbl>,
## #   ‘2/24/20‘ <dbl>, ‘2/25/20‘ <dbl>, ‘2/26/20‘ <dbl>, ‘2/27/20‘ <dbl>,
## #   ‘2/28/20‘ <dbl>, ‘2/29/20‘ <dbl>, ‘3/1/20‘ <dbl>, ‘3/2/20‘ <dbl>,
## #   ‘3/3/20‘ <dbl>, ‘3/4/20‘ <dbl>, ‘3/5/20‘ <dbl>, ‘3/6/20‘ <dbl>,
## #   ‘3/7/20‘ <dbl>, ‘3/8/20‘ <dbl>, ‘3/9/20‘ <dbl>, ‘3/10/20‘ <dbl>,
## #   ‘3/11/20‘ <dbl>, ‘3/12/20‘ <dbl>, ‘3/13/20‘ <dbl>, ‘3/14/20‘ <dbl>,
## #   ‘3/15/20‘ <dbl>, ‘3/16/20‘ <dbl>, ‘3/17/20‘ <dbl>, ‘3/18/20‘ <dbl>,
## #   ‘3/19/20‘ <dbl>, ‘3/20/20‘ <dbl>, ‘3/21/20‘ <dbl>, ‘3/22/20‘ <dbl>,
## #   ‘3/23/20‘ <dbl>, ‘3/24/20‘ <dbl>, ‘3/25/20‘ <dbl>, ‘3/26/20‘ <dbl>,
## #   ‘3/27/20‘ <dbl>, ‘3/28/20‘ <dbl>, ‘3/29/20‘ <dbl>, ‘3/30/20‘ <dbl>,
## #   ‘3/31/20‘ <dbl>, ‘4/1/20‘ <dbl>, ‘4/2/20‘ <dbl>, ‘4/3/20‘ <dbl>,
## #   ‘4/4/20‘ <dbl>, ‘4/5/20‘ <dbl>, ‘4/6/20‘ <dbl>, ‘4/7/20‘ <dbl>,
## #   ‘4/8/20‘ <dbl>, ‘4/9/20‘ <dbl>, ‘4/10/20‘ <dbl>, ‘4/11/20‘ <dbl>,
## #   ‘4/12/20‘ <dbl>, ‘4/13/20‘ <dbl>, ‘4/14/20‘ <dbl>, ‘4/15/20‘ <dbl>,
## #   ‘4/16/20‘ <dbl>, ‘4/17/20‘ <dbl>, ‘4/18/20‘ <dbl>, ‘4/19/20‘ <dbl>,
## #   ‘4/20/20‘ <dbl>, ‘4/21/20‘ <dbl>, ‘4/22/20‘ <dbl>, ‘4/23/20‘ <dbl>,
## #   ‘4/24/20‘ <dbl>, ‘4/25/20‘ <dbl>, ‘4/26/20‘ <dbl>, ‘4/27/20‘ <dbl>,
## #   ‘4/28/20‘ <dbl>, ‘4/29/20‘ <dbl>, ‘4/30/20‘ <dbl>, ‘5/1/20‘ <dbl>,
## #   ‘5/2/20‘ <dbl>, ‘5/3/20‘ <dbl>, ...
```

```
head(covid_us_cases)
```

```
## # A tibble: 6 x 680
##        UID iso2  iso3  code3  FIPS Admin2 Province_State Country_Region   Lat
##      <dbl> <chr> <chr> <dbl> <dbl> <chr>  <chr>          <chr>          <dbl>
## 1 8.40e7 US    USA     840  1001 Autau~ Alabama        US              32.5
## 2 8.40e7 US    USA     840  1003 Baldw~ Alabama        US              30.7
## 3 8.40e7 US    USA     840  1005 Barbo~ Alabama        US              31.9
## 4 8.40e7 US    USA     840  1007 Bibb   Alabama        US              33.0
## 5 8.40e7 US    USA     840  1009 Blount Alabama        US              34.0
## 6 8.40e7 US    USA     840  1011 Bullo~ Alabama        US              32.1
## # ... with 671 more variables: Long_ <dbl>, Combined_Key <chr>,
## #   ‘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>, ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>,
## #   ‘2/19/20‘ <dbl>, ‘2/20/20‘ <dbl>, ‘2/21/20‘ <dbl>, ‘2/22/20‘ <dbl>,
## #   ‘2/23/20‘ <dbl>, ‘2/24/20‘ <dbl>, ‘2/25/20‘ <dbl>, ‘2/26/20‘ <dbl>,
## #   ‘2/27/20‘ <dbl>, ‘2/28/20‘ <dbl>, ‘2/29/20‘ <dbl>, ‘3/1/20‘ <dbl>,
## #   ‘3/2/20‘ <dbl>, ‘3/3/20‘ <dbl>, ‘3/4/20‘ <dbl>, ‘3/5/20‘ <dbl>,
## #   ‘3/6/20‘ <dbl>, ‘3/7/20‘ <dbl>, ‘3/8/20‘ <dbl>, ‘3/9/20‘ <dbl>,
## #   ‘3/10/20‘ <dbl>, ‘3/11/20‘ <dbl>, ‘3/12/20‘ <dbl>, ‘3/13/20‘ <dbl>,
## #   ‘3/14/20‘ <dbl>, ‘3/15/20‘ <dbl>, ‘3/16/20‘ <dbl>, ‘3/17/20‘ <dbl>,
## #   ‘3/18/20‘ <dbl>, ‘3/19/20‘ <dbl>, ‘3/20/20‘ <dbl>, ‘3/21/20‘ <dbl>,
## #   ‘3/22/20‘ <dbl>, ‘3/23/20‘ <dbl>, ‘3/24/20‘ <dbl>, ‘3/25/20‘ <dbl>,
## #   ‘3/26/20‘ <dbl>, ‘3/27/20‘ <dbl>, ‘3/28/20‘ <dbl>, ‘3/29/20‘ <dbl>,
## #   ‘3/30/20‘ <dbl>, ‘3/31/20‘ <dbl>, ‘4/1/20‘ <dbl>, ‘4/2/20‘ <dbl>,
## #   ‘4/3/20‘ <dbl>, ‘4/4/20‘ <dbl>, ‘4/5/20‘ <dbl>, ‘4/6/20‘ <dbl>,
## #   ‘4/7/20‘ <dbl>, ‘4/8/20‘ <dbl>, ‘4/9/20‘ <dbl>, ‘4/10/20‘ <dbl>,
## #   ‘4/11/20‘ <dbl>, ‘4/12/20‘ <dbl>, ‘4/13/20‘ <dbl>, ‘4/14/20‘ <dbl>,
## #   ‘4/15/20‘ <dbl>, ‘4/16/20‘ <dbl>, ‘4/17/20‘ <dbl>, ‘4/18/20‘ <dbl>,
## #   ‘4/19/20‘ <dbl>, ‘4/20/20‘ <dbl>, ‘4/21/20‘ <dbl>, ‘4/22/20‘ <dbl>,
## #   ‘4/23/20‘ <dbl>, ‘4/24/20‘ <dbl>, ‘4/25/20‘ <dbl>, ‘4/26/20‘ <dbl>,
## #   ‘4/27/20‘ <dbl>, ‘4/28/20‘ <dbl>, ...
```

```
head(covid_global_cases)
```

```
## # A tibble: 6 x 673
##   ‘Province/State‘ ‘Country/Region‘   Lat   Long ‘1/22/20‘ ‘1/23/20‘ ‘1/24/20‘
##   <chr>            <chr>            <dbl>  <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9  67.7          0         0         0
## 2 <NA>             Albania           41.2  20.2          0         0         0
## 3 <NA>             Algeria           28.0   1.66         0         0         0
## 4 <NA>             Andorra           42.5   1.52         0         0         0
## 5 <NA>             Angola           -11.2  17.9          0         0         0
## 6 <NA>             Antigua and Bar~  17.1 -61.8          0         0         0
## # ... with 666 more variables: ‘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>, ‘2/17/20‘ <dbl>, ‘2/18/20‘ <dbl>, ‘2/19/20‘ <dbl>,
## #   ‘2/20/20‘ <dbl>, ‘2/21/20‘ <dbl>, ‘2/22/20‘ <dbl>, ‘2/23/20‘ <dbl>,
## #   ‘2/24/20‘ <dbl>, ‘2/25/20‘ <dbl>, ‘2/26/20‘ <dbl>, ‘2/27/20‘ <dbl>,
## #   ‘2/28/20‘ <dbl>, ‘2/29/20‘ <dbl>, ‘3/1/20‘ <dbl>, ‘3/2/20‘ <dbl>,
## #   ‘3/3/20‘ <dbl>, ‘3/4/20‘ <dbl>, ‘3/5/20‘ <dbl>, ‘3/6/20‘ <dbl>,
## #   ‘3/7/20‘ <dbl>, ‘3/8/20‘ <dbl>, ‘3/9/20‘ <dbl>, ‘3/10/20‘ <dbl>,
## #   ‘3/11/20‘ <dbl>, ‘3/12/20‘ <dbl>, ‘3/13/20‘ <dbl>, ‘3/14/20‘ <dbl>,
## #   ‘3/15/20‘ <dbl>, ‘3/16/20‘ <dbl>, ‘3/17/20‘ <dbl>, ‘3/18/20‘ <dbl>,
## #   ‘3/19/20‘ <dbl>, ‘3/20/20‘ <dbl>, ‘3/21/20‘ <dbl>, ‘3/22/20‘ <dbl>,
## #   ‘3/23/20‘ <dbl>, ‘3/24/20‘ <dbl>, ‘3/25/20‘ <dbl>, ‘3/26/20‘ <dbl>,
## #   ‘3/27/20‘ <dbl>, ‘3/28/20‘ <dbl>, ‘3/29/20‘ <dbl>, ‘3/30/20‘ <dbl>,
## #   ‘3/31/20‘ <dbl>, ‘4/1/20‘ <dbl>, ‘4/2/20‘ <dbl>, ‘4/3/20‘ <dbl>,
## #   ‘4/4/20‘ <dbl>, ‘4/5/20‘ <dbl>, ‘4/6/20‘ <dbl>, ‘4/7/20‘ <dbl>,
## #   ‘4/8/20‘ <dbl>, ‘4/9/20‘ <dbl>, ‘4/10/20‘ <dbl>, ‘4/11/20‘ <dbl>,
## #   ‘4/12/20‘ <dbl>, ‘4/13/20‘ <dbl>, ‘4/14/20‘ <dbl>, ‘4/15/20‘ <dbl>,
## #   ‘4/16/20‘ <dbl>, ‘4/17/20‘ <dbl>, ‘4/18/20‘ <dbl>, ‘4/19/20‘ <dbl>,
## #   ‘4/20/20‘ <dbl>, ‘4/21/20‘ <dbl>, ‘4/22/20‘ <dbl>, ‘4/23/20‘ <dbl>,
## #   ‘4/24/20‘ <dbl>, ‘4/25/20‘ <dbl>, ‘4/26/20‘ <dbl>, ‘4/27/20‘ <dbl>,
## #   ‘4/28/20‘ <dbl>, ‘4/29/20‘ <dbl>, ‘4/30/20‘ <dbl>, ‘5/1/20‘ <dbl>,
## #   ‘5/2/20‘ <dbl>, ‘5/3/20‘ <dbl>, ...
```

Clean the data

```
covid_gd_cleaned <- covid_global_deaths %>%
  pivot_longer(cols = -c(‘Province/State‘, ‘Country/Region‘, Lat, Long),
               names_to = "date",
               values_to = "deaths") %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long)) %>%
  rename(Province_State = ‘Province/State‘,
         Country_Region = ‘Country/Region‘)

covid_gc_cleaned <- covid_global_cases %>%
  pivot_longer(cols = -c(‘Province/State‘, ‘Country/Region‘, Lat, Long),
               names_to = "date",
               values_to = "cases") %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat, Long)) %>%
  rename(Province_State = ‘Province/State‘,
         Country_Region = ‘Country/Region‘)

covid_gd_cleaned$Country_Region <- factor(covid_gd_cleaned$Country_Region)
covid_gd_cleaned$Province_State <- factor(covid_gd_cleaned$Province_State)
```

```r
covid_gc_cleaned$Country_Region <- factor(covid_gc_cleaned$Country_Region)
covid_gc_cleaned$Province_State <- factor(covid_gc_cleaned$Province_State)


covid_usd_cleaned <- covid_us_deaths %>%
  pivot_longer(cols = -c(Province_State, Country_Region, UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long
               names_to = "date",
               values_to = "deaths") %>%
  mutate(date = mdy(date)) %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long_))



covid_usc_cleaned <- covid_us_cases %>%
  pivot_longer(cols = -c(Province_State, Country_Region, UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long
               names_to = "date",
               values_to = "cases") %>%
  mutate(date = mdy(date)) %>%
  select(-c(UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long_))

covid_usd_cleaned$Province_State <- factor(covid_usd_cleaned$Province_State)
covid_usc_cleaned$Province_State <- factor(covid_usc_cleaned$Province_State)

summary(covid_gd_cleaned)
```

```
##                      Province_State          Country_Region
##  Alberta                     :   669   China         : 22746
##  Anguilla                    :   669   Canada        : 10704
##  Anhui                       :   669   France        :  8028
##  Aruba                       :   669   United Kingdom:  8028
##  Australian Capital Territory:   669   Australia     :  5352
##  (Other)                     : 54858   Netherlands   :  3345
##  NA's                        :129117   (Other)       :129117
##       date                 deaths
##  Min.   :2020-01-22   Min.   :     0
##  1st Qu.:2020-07-07   1st Qu.:     1
##  Median :2020-12-21   Median :    47
##  Mean   :2020-12-21   Mean   :  7601
##  3rd Qu.:2021-06-06   3rd Qu.:  1134
##  Max.   :2021-11-20   Max.   :771013
##
```

```r
summary(covid_usd_cleaned)
```

```
##   Province_State    Country_Region     Combined_Key         Population
##  Texas    : 171264   Length:2235798     Length:2235798     Min.   :       0
##  Georgia  : 107709   Class :character   Class :character   1st Qu.:    9917
##  Virginia:  90315   Mode  :character   Mode  :character   Median :   24892
##  Kentucky:  81618                                          Mean   :   99604
##  Missouri:  78942                                          3rd Qu.:   64979
##  Kansas  :  71583                                          Max.   :10039107
##  (Other) :1634367
##       date                 deaths
##  Min.   :2020-01-22   Min.   :    0.0
##  1st Qu.:2020-07-07   1st Qu.:    0.0
##  Median :2020-12-21   Median :   13.0
##  Mean   :2020-12-21   Mean   :  106.7
##  3rd Qu.:2021-06-06   3rd Qu.:   60.0
##  Max.   :2021-11-20   Max.   :26999.0
##
```

```
summary(covid_gc_cleaned)
```

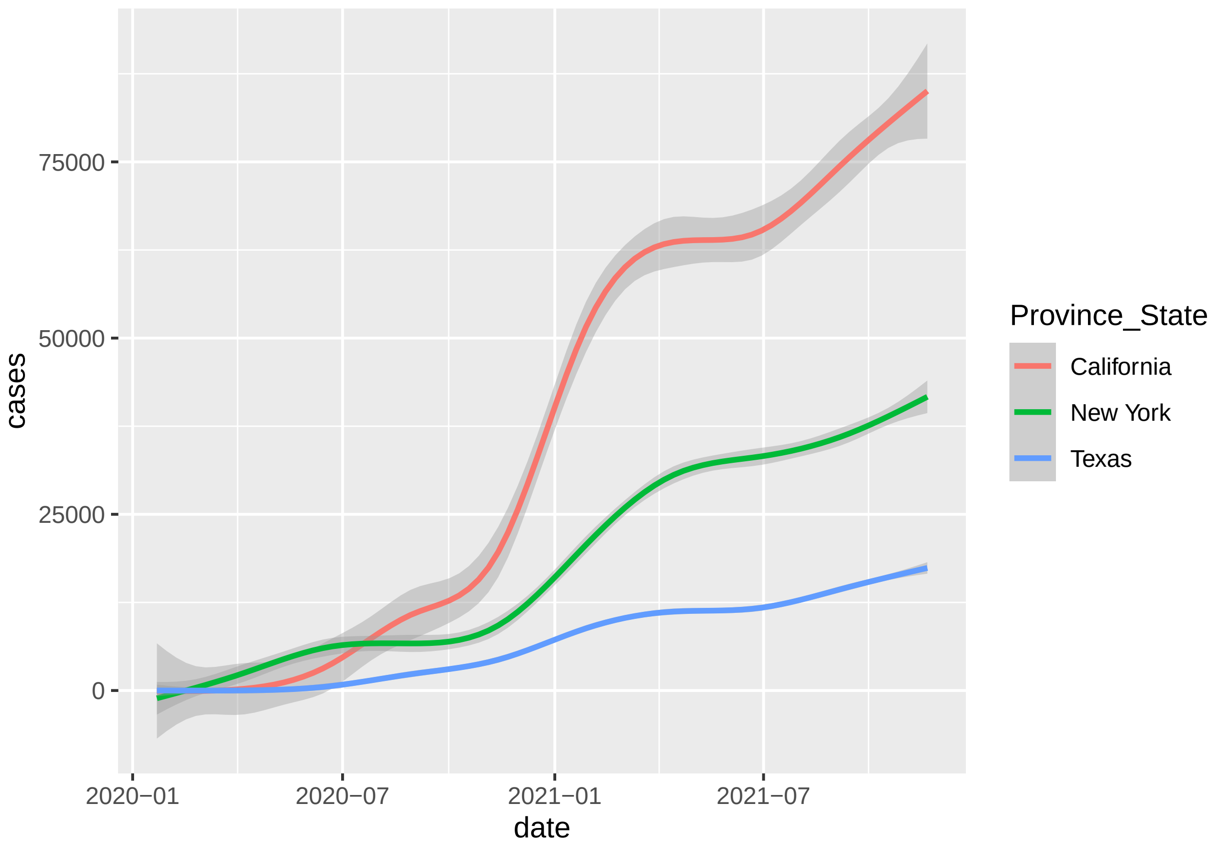
```
##                      Province_State          Country_Region
##  Alberta                     :   669   China         : 22746
##  Anguilla                    :   669   Canada        : 10704
##  Anhui                       :   669   France        :  8028
##  Aruba                       :   669   United Kingdom:  8028
##  Australian Capital Territory:   669   Australia     :  5352
##  (Other)                     : 54858   Netherlands   :  3345
##  NA’s                        :129117   (Other)       :129117
##       date                 cases
##  Min.   :2020-01-22   Min.   :       0
##  1st Qu.:2020-07-07   1st Qu.:     170
##  Median :2020-12-21   Median :    3016
##  Mean   :2020-12-21   Mean   :  339796
##  3rd Qu.:2021-06-06   3rd Qu.:   66890
##  Max.   :2021-11-20   Max.   :47701872
##
```

```
summary(covid_usc_cleaned)
```

```
##   Province_State    Country_Region     Combined_Key            date
##  Texas    : 171264   Length:2235798     Length:2235798     Min.   :2020-01-22
##  Georgia : 107709   Class :character   Class :character   1st Qu.:2020-07-07
##  Virginia:  90315   Mode  :character   Mode  :character   Median :2020-12-21
##  Kentucky:  81618                                         Mean   :2020-12-21
##  Missouri:  78942                                         3rd Qu.:2021-06-06
##  Kansas  :  71583                                         Max.   :2021-11-20
##  (Other) :1634367
##      cases
##  Min.   :      0
##  1st Qu.:     47
##  Median :    771
##  Mean   :   5780
##  3rd Qu.:   3215
##  Max.   :1518732
##
```

```
ggplot(subset(covid_usc_cleaned, Province_State %in% c("California", "Texas", "New York")),
       aes(x = date, y = cases)) +
  labs(tite = "Cases in 3 of the Biggest States in the US") +
geom_smooth(aes(y = cases, color = Province_State))
```

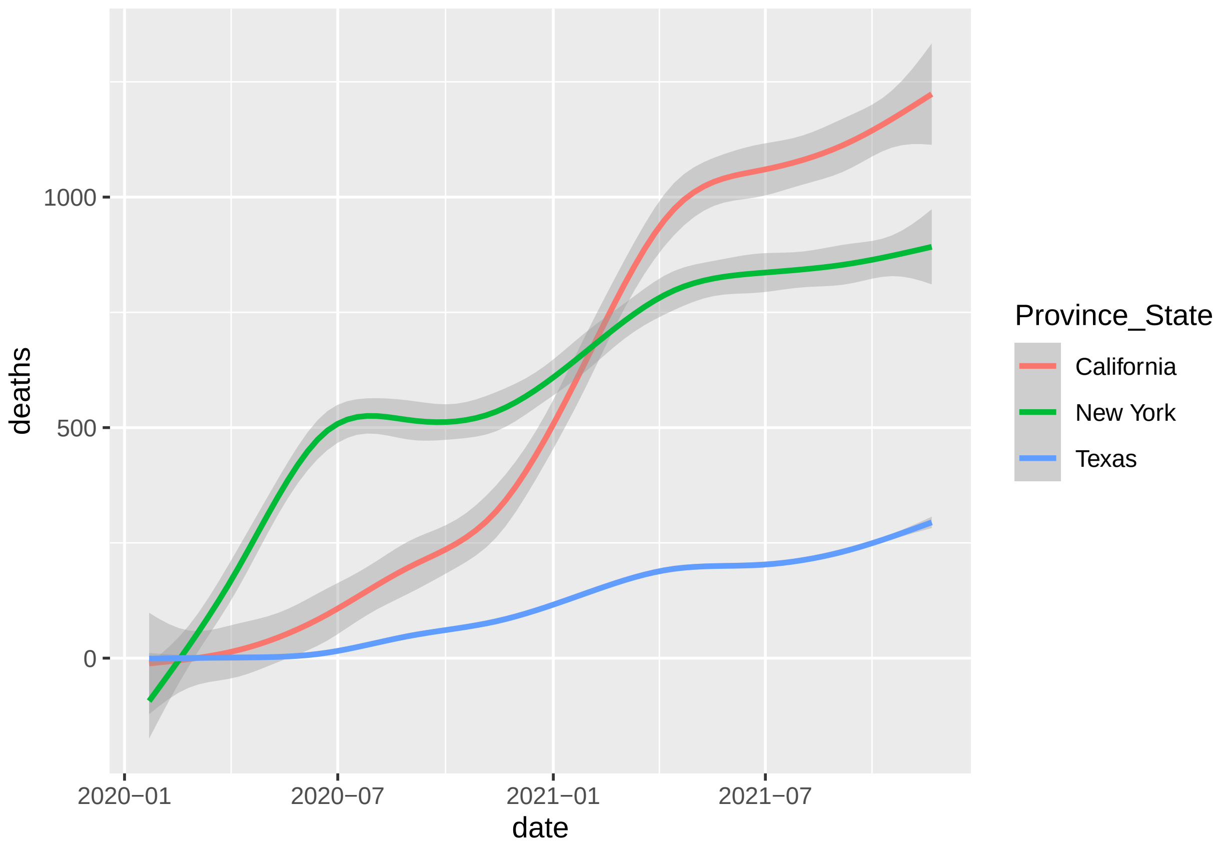
```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



## Observation: The number of cases in California reflects the large population number that California has. According to the data there was a very big spike in cases in California near November 2020 to April 2021. This is interesting because according to the graph, Texas and New York also spiked but not nearly as much. Once again, I am assuming this is due to the population density of California.

```
ggplot(subset(covid_usd_cleaned, Province_State %in% c("California", "Texas", "New York")),
       aes(x = date, y = deaths)) +
geom_smooth(aes(y = deaths, color = Province_State))
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



## Observation: According to the dataset, it looks like New York spiked in covid related deaths early on from January 2020 to July 2020. Whereas California also was rising in deaths, but right around the same time frame the covid cases spiked in California, the deaths also did. Texas seems to be constanty rising in covid deaths.

Creating a model

If you are trying to run this, it will take a while. It is not broken.

```
lm_cstates = lm(cases~Province_State, data = covid_usc_cleaned)
lm_dstates = lm(deaths~Province_State, data = covid_usd_cleaned)
```

```
summary(lm_cstates)
```

```
##
## Call:
## lm(formula = cases ~ Province_State, data = covid_usc_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -36742   -4730   -2199    -561 1481990
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                4874.1      122.9  39.648  < 2e-16 ***
## Province_StateAlaska                    -3570.1      216.1 -16.518  < 2e-16 ***
## Province_StateAmerican Samoa            -4873.9     1028.6  -4.739 2.15e-06 ***
## Province_StateArizona                   25352.5      276.5  91.690  < 2e-16 ***
## Province_StateArkansas                  -2135.1      169.3 -12.613  < 2e-16 ***
## Province_StateCalifornia                31867.9      180.3 176.790  < 2e-16 ***
## Province_StateColorado                   -334.6      175.8  -1.903  0.05701 .
## Province_StateConnecticut               13928.8      345.5  40.311  < 2e-16 ***
## Province_StateDelaware                   7104.8      472.9  15.023  < 2e-16 ***
## Province_StateDiamond Princess          -4829.2     1028.6  -4.695 2.66e-06 ***
## Province_StateDistrict of Columbia       4907.9      602.3   8.149 3.66e-16 ***
## Province_StateFlorida                   15989.8      173.9  91.971  < 2e-16 ***
## Province_StateGeorgia                    -670.7      146.9  -4.565 5.00e-06 ***
## Province_StateGrand Princess            -4780.8     1028.6  -4.648 3.35e-06 ***
## Province_StateGuam                        910.7     1028.6   0.885  0.37593
## Province_StateHawaii                    -1232.3      405.1  -3.042  0.00235 **
## Province_StateIdaho                     -2371.5      194.4 -12.200  < 2e-16 ***
## Province_StateIllinois                   2744.3      158.6  17.308  < 2e-16 ***
## Province_StateIndiana                    -311.7      161.9  -1.925  0.05421 .
## Province_StateIowa                      -2697.3      159.5 -16.912  < 2e-16 ***
## Province_StateKansas                    -3135.0      157.7 -19.883  < 2e-16 ***
## Province_StateKentucky                  -2662.4      153.8 -17.308  < 2e-16 ***
## Province_StateLouisiana                  -143.9      175.8  -0.818  0.41326
## Province_StateMaine                     -2956.6      270.3 -10.939  < 2e-16 ***
## Province_StateMaryland                   5225.0      235.0  22.235  < 2e-16 ***
## Province_StateMassachusetts             18039.2      276.5  65.240  < 2e-16 ***
## Province_StateMichigan                   1063.0      164.6   6.458 1.06e-10 ***
## Province_StateMinnesota                 -1111.3      163.8  -6.785 1.16e-11 ***
## Province_StateMississippi               -2453.3      165.9 -14.787  < 2e-16 ***
## Province_StateMissouri                  -1760.7      154.8 -11.377  < 2e-16 ***
## Province_StateMontana                   -3743.6      181.9 -20.579  < 2e-16 ***
## Province_StateNebraska                  -3505.0      161.5 -21.700  < 2e-16 ***
## Province_StateNevada                     5301.0      264.6  20.036  < 2e-16 ***
## Province_StateNew Hampshire              -614.1      319.4  -1.923  0.05452 .
## Province_StateNew Jersey                19571.4      245.9  79.600  < 2e-16 ***
## Province_StateNew Mexico                -1534.0      211.9  -7.239 4.53e-13 ***
## Province_StateNew York                  13425.4      177.2  75.755  < 2e-16 ***
## Province_StateNorth Carolina              803.5      159.2   5.048 4.47e-07 ***
## Province_StateNorth Dakota              -3690.8      184.6 -19.995  < 2e-16 ***
## Province_StateNorthern Mariana Islands  -4755.1     1028.6  -4.623 3.78e-06 ***
## Province_StateOhio                       2001.1      163.4  12.247  < 2e-16 ***
## Province_StateOklahoma                  -1509.3      168.3  -8.970  < 2e-16 ***
## Province_StateOregon                    -1701.3      206.3  -8.247  < 2e-16 ***
## Province_StatePennsylvania               4484.5      173.9  25.794  < 2e-16 ***
## Province_StatePuerto Rico               -3939.7      167.8 -23.482  < 2e-16 ***
## Province_StateRhode Island               6951.1      405.1  17.160  < 2e-16 ***
## Province_StateSouth Carolina             2393.6      191.9  12.471  < 2e-16 ***
## Province_StateSouth Dakota              -3816.6      174.5 -21.872  < 2e-16 ***
## Province_StateTennessee                   444.9      160.8   2.767  0.00566 **
## Province_StateTexas                      2000.1      138.5  14.439  < 2e-16 ***
## Province_StateUtah                       1492.8      208.1   7.174 7.28e-13 ***
## Province_StateVermont                   -4088.0      283.4 -14.427  < 2e-16 ***
## Province_StateVirgin Islands            -2488.2     1028.6  -2.419  0.01556 *
## Province_StateVirginia                  -2045.8      151.1 -13.537  < 2e-16 ***
## Province_StateWashington                 1451.9      201.4   7.210 5.59e-13 ***
## Province_StateWest Virginia             -3283.3      182.8 -17.963  < 2e-16 ***
## Province_StateWisconsin                   428.6      170.9   2.508  0.01214 *
## Province_StateWyoming                   -3411.6      238.4 -14.312  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 26410 on 2235740 degrees of freedom
## Multiple R-squared:  0.05388,    Adjusted R-squared:  0.05386
## F-statistic:  2234 on 57 and 2235740 DF,  p-value: < 2.2e-16
```

```
summary(lm_dstates)
```

```
##
## Call:
## lm(formula = deaths ~ Province_State, data = covid_usd_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -791.5   -73.5   -30.5    -5.0 26443.8
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              90.7742     2.3697  38.307  < 2e-16 ***
## Province_StateAlaska                    -84.1583     4.1661 -20.201  < 2e-16 ***
## Province_StateAmerican Samoa            -90.7742    19.8260  -4.579 4.68e-06 ***
## Province_StateArizona                   497.7551     5.3298  93.391  < 2e-16 ***
## Province_StateArkansas                  -46.3072     3.2630 -14.192  < 2e-16 ***
## Province_StateCalifornia                464.3964     3.4746 133.655  < 2e-16 ***
## Province_StateColorado                  -29.4583     3.3891  -8.692  < 2e-16 ***
## Province_StateConnecticut               466.5182     6.6604  70.044  < 2e-16 ***
## Province_StateDelaware                  119.5711     9.1163  13.116  < 2e-16 ***
## Province_StateDiamond Princess          -90.7742    19.8260  -4.579 4.68e-06 ***
## Province_StateDistrict of Columbia      158.6114    11.6089  13.663  < 2e-16 ***
## Province_StateFlorida                   243.3815     3.3512  72.625  < 2e-16 ***
## Province_StateGeorgia                   -14.7161     2.8323  -5.196 2.04e-07 ***
## Province_StateGrand Princess            -88.1031    19.8260  -4.444 8.84e-06 ***
## Province_StateGuam                       -0.6651    19.8260  -0.034 0.973238
## Province_StateHawaii                    -46.3316     7.8081  -5.934 2.96e-09 ***
## Province_StateIdaho                     -63.2876     3.7468 -16.891  < 2e-16 ***
## Province_StateIllinois                   56.5771     3.0563  18.512  < 2e-16 ***
## Province_StateIndiana                    -4.6790     3.1204  -1.499 0.133751
## Province_StateIowa                      -57.2958     3.0743 -18.637  < 2e-16 ***
## Province_StateKansas                    -64.9211     3.0391 -21.362  < 2e-16 ***
## Province_StateKentucky                  -60.6754     2.9650 -20.464  < 2e-16 ***
## Province_StateLouisiana                  17.7354     3.3891   5.233 1.67e-07 ***
## Province_StateMaine                     -65.2367     5.2097 -12.522  < 2e-16 ***
## Province_StateMaryland                  131.7196     4.5296  29.080  < 2e-16 ***
## Province_StateMassachusetts             598.8963     5.3298 112.368  < 2e-16 ***
## Province_StateMichigan                   50.3333     3.1731  15.862  < 2e-16 ***
## Province_StateMinnesota                 -41.3378     3.1573 -13.093  < 2e-16 ***
## Province_StateMississippi               -36.4426     3.1981 -11.395  < 2e-16 ***
## Province_StateMissouri                  -44.2219     2.9831 -14.824  < 2e-16 ***
## Province_StateMontana                   -75.3972     3.5065 -21.502  < 2e-16 ***
## Province_StateNebraska                  -76.9741     3.1135 -24.723  < 2e-16 ***
## Province_StateNevada                    81.1543     5.0998  15.913  < 2e-16 ***
## Province_StateNew Hampshire            -24.8536     6.1565  -4.037 5.42e-05 ***
## Province_StateNew Jersey               700.6805     4.7393 147.844  < 2e-16 ***
## Province_StateNew Mexico               -23.6415     4.0848  -5.788 7.14e-09 ***
## Province_StateNew York                 496.2788     3.4160 145.280  < 2e-16 ***
## Province_StateNorth Carolina           -17.1221     3.0682  -5.581 2.40e-08 ***
## Province_StateNorth Dakota             -74.6990     3.5581 -20.994  < 2e-16 ***
## Province_StateNorthern Mariana Islands -88.9357    19.8260  -4.486 7.26e-06 ***
## Province_StateOhio                      38.8824     3.1497  12.345  < 2e-16 ***
## Province_StateOklahoma                 -44.3003     3.2434 -13.659  < 2e-16 ***
## Province_StateOregon                   -49.2842     3.9764 -12.394  < 2e-16 ***
## Province_StatePennsylvania             138.5926     3.3512  41.356  < 2e-16 ***
## Province_StatePuerto Rico              -73.1961     3.2340 -22.634  < 2e-16 ***
## Province_StateRhode Island             152.0018     7.8081  19.467  < 2e-16 ***
## Province_StateSouth Carolina            28.5584     3.6996   7.719 1.17e-14 ***
## Province_StateSouth Dakota             -74.6708     3.3635 -22.200  < 2e-16 ***
## Province_StateTennessee                -19.4082     3.0999  -6.261 3.83e-10 ***
## Province_StateTexas                     25.7133     2.6700   9.631  < 2e-16 ***
## Province_StateUtah                     -55.4792     4.0109 -13.832  < 2e-16 ***
## Province_StateVermont                  -81.4756     5.4618 -14.917  < 2e-16 ***
## Province_StateVirgin Islands           -65.9566    19.8260  -3.327 0.000879 ***
## Province_StateVirginia                 -45.0107     2.9130 -15.452  < 2e-16 ***
## Province_StateWashington                -1.9098     3.8814  -0.492 0.622698
## Province_StateWest Virginia            -63.7281     3.5232 -18.088  < 2e-16 ***
## Province_StateWisconsin                -30.7793     3.2941  -9.344  < 2e-16 ***
## Province_StateWyoming                  -74.1984     4.5949 -16.148  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 509.1 on 2235740 degrees of freedom
## Multiple R-squared:  0.06946,    Adjusted R-squared:  0.06944
## F-statistic:  2928 on 57 and 2235740 DF,  p-value: < 2.2e-16
```

```
covid_us <- covid_usd_cleaned %>% inner_join(covid_usc_cleaned, by=c("Province_State", "date", "Country
head(covid_us)
```

```
## # A tibble: 6 x 7
##   Province_State Country_Region Combined_Key  Population date       deaths cases
##   <fct>          <chr>          <chr>              <dbl> <date>      <dbl> <dbl>
## 1 Alabama        US             Autauga, Ala~      55869 2020-01-22      0     0
## 2 Alabama        US             Autauga, Ala~      55869 2020-01-23      0     0
## 3 Alabama        US             Autauga, Ala~      55869 2020-01-24      0     0
## 4 Alabama        US             Autauga, Ala~      55869 2020-01-25      0     0
## 5 Alabama        US             Autauga, Ala~      55869 2020-01-26      0     0
## 6 Alabama        US             Autauga, Ala~      55869 2020-01-27      0     0
```

```
max(covid_us$cases[covid_us$Province_State == ’California’]) / max(covid_us$Population[covid_us$Provinc
```

```
## [1] 0.1512816
```

```
max(covid_us$cases[covid_us$Province_State == ’Texas’]) / max(covid_us$Population[covid_us$Province_Sta
```

```
## [1] 0.123865
```

```
max(covid_us$cases[covid_us$Province_State == ’New York’]) / max(covid_us$Population[covid_us$Province_S
```

```
## [1] 0.1347371
```

```
max(covid_us$deaths[covid_us$Province_State == ’California’]) / max(covid_us$cases[covid_us$Province_Sta
```

```
## [1] 0.01777733
```

```
max(covid_us$deaths[covid_us$Province_State == ’Texas’]) / max(covid_us$cases[covid_us$Province_State ==
```

```
## [1] 0.01621915
```

```
max(covid_us$deaths[covid_us$Province_State == ’New York’]) / max(covid_us$cases[covid_us$Province_State
```

```
## [1] 0.03170935
```

Analysis and bias:

According the data, there seems to be somewhat of a correlation between state and covid cases/deaths. The r-squared value of both the models are around 0.05, which is not the best for a model, but it does give us some insight into if population or just the state in general handled covid well or not. Instead, I will use a simple statistic to find the percentage of cases based on the maximum population. California has a 15% covid case rate, Texas has a 12% covid case rate, and New York has a 13% covid case rate. However when we look at the mortality rates of individuals who are infected with covid, we see that California has a 1.7% rate of death based on the covid case rate. What’s interesting here is that New York has the highest out of the 3, but it also has a smaller case rate than California.

There is definitely some bias that is involved because I am assuming that handling of covid is directly correlated with covid cases and deaths in the state. We do not know if the data has people that were from out of state that acquired covid, or died in a state with covid. In conclusion, the data points to Texas actually having a pretty decent covid case and death rate, which is something I personally would not have guessed. It seems like more dense populations are infected with covid at a higher rate.