

COVID-19 Analysis

Peter Tarara

2022-10-08

Data Summary and Tidying

```
url_in <-  
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_ti  
  
file_names <-  
  c("time_series_covid19_confirmed_global.csv",  
    "time_series_covid19_deaths_global.csv",  
    "time_series_covid19_confirmed_US.csv",  
    "time_series_covid19_deaths_US.csv")  
  
urls <- str_c(url_in, file_names)  
  
global_cases <- read_csv(urls[1])
```

```
## Rows: 289 Columns: 994  
## -- Column specification -----  
## Delimiter: ","  
## chr   (2): Province/State, Country/Region  
## dbl (992): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...  
##  
## 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.
```

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

```
## Rows: 289 Columns: 994  
## -- Column specification -----  
## Delimiter: ","  
## chr   (2): Province/State, Country/Region  
## dbl (992): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20, ...  
##  
## 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 <- read_csv(urls[3])
```

```
## Rows: 3342 Columns: 1001
```

```
## -- Column specification -----
## Delimiter: ","
## chr (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (995): 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_deaths <- read_csv(urls[4])
```

```
## Rows: 3342 Columns: 1002
## -- Column specification -----
## Delimiter: ","
## chr (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (996): 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.
```

```
#global_cases <- global_cases %>% select(-c(UID,iso2,iso3,code3,FIPS,Admin2,Combined_Key))
```

```
global_cases <- global_cases %>%
  pivot_longer(cols = -c(`Province/State`,
                        `Country/Region`, Lat, Long),
              names_to = "date",
              values_to = "cases") %>%
  select(-c(Lat,Long))
```

```
global_deaths <- global_deaths %>%
  pivot_longer(cols = -c(`Province/State`, `Country/Region`, Lat, Long),
              names_to = "date",
              values_to = "deaths") %>%
  select(-c(Lat, Long))
```

```
#global_deaths <- global_deaths %>% rename(`Country_Region` = `Country/Region`)
#global_deaths <- global_deaths %>% rename(`Province_State` = `Province/State`)
```

```
global <- global_cases %>%
  full_join(global_deaths) %>%
  mutate(date = mdy(date))
```

```
## Joining, by = c("Province/State", "Country/Region", "date")
```

```
global <- global %>% filter(cases > 0) #issue with this filter
```

```
summary(global)
```

```
## Province/State    Country/Region      date      cases
## Length:262916     Length:262916    Min.   :2020-01-22  Min.   :      1
## Class :character   Class :character 1st Qu.:2020-11-02  1st Qu.:    959
## Mode  :character   Mode  :character Median :2021-06-29  Median :   14833
```

```
##                               Mean   :2021-06-25   Mean   : 825013
##                               3rd Qu.:2022-02-19   3rd Qu.: 214798
##                               Max.    :2022-10-07   Max.    :96686904
##      deaths
## Min.      :      0
## 1st Qu.:      6
## Median :    168
## Mean    :   12905
## 3rd Qu.:   3056
## Max.    : 1062513
```

```
us_cases <- 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 <- 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, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key",
## "date")
```

```
global <- global %>%
  unite("Combined_Key",
        c(`Province/State`, `Country/Region`),
        sep = ", ",
        na.rm = TRUE,
        remove = FALSE)
```

```
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/"
```

```
uid <- read_csv(uid_lookup_url) %>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))
```

```
## Rows: 4321 Columns: 12
## -- Column specification -----
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population
##
## 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.
```

```

global <- global %>% rename(`Country_Region` = `Country/Region`)
global <- global %>% rename(`Province_State` = `Province/State`)

global <- global %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>%
  select(-c(UID, FIPS)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population, Combined_Key)

```

Visuals

```

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

```

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

```

US_state_totals <- US_by_state %>%
  group_by(Province_State) %>%
  summarize(deaths = max(deaths), cases = max(cases), Population = max(Population)) %>%
  filter(cases > 0, Population > 0)

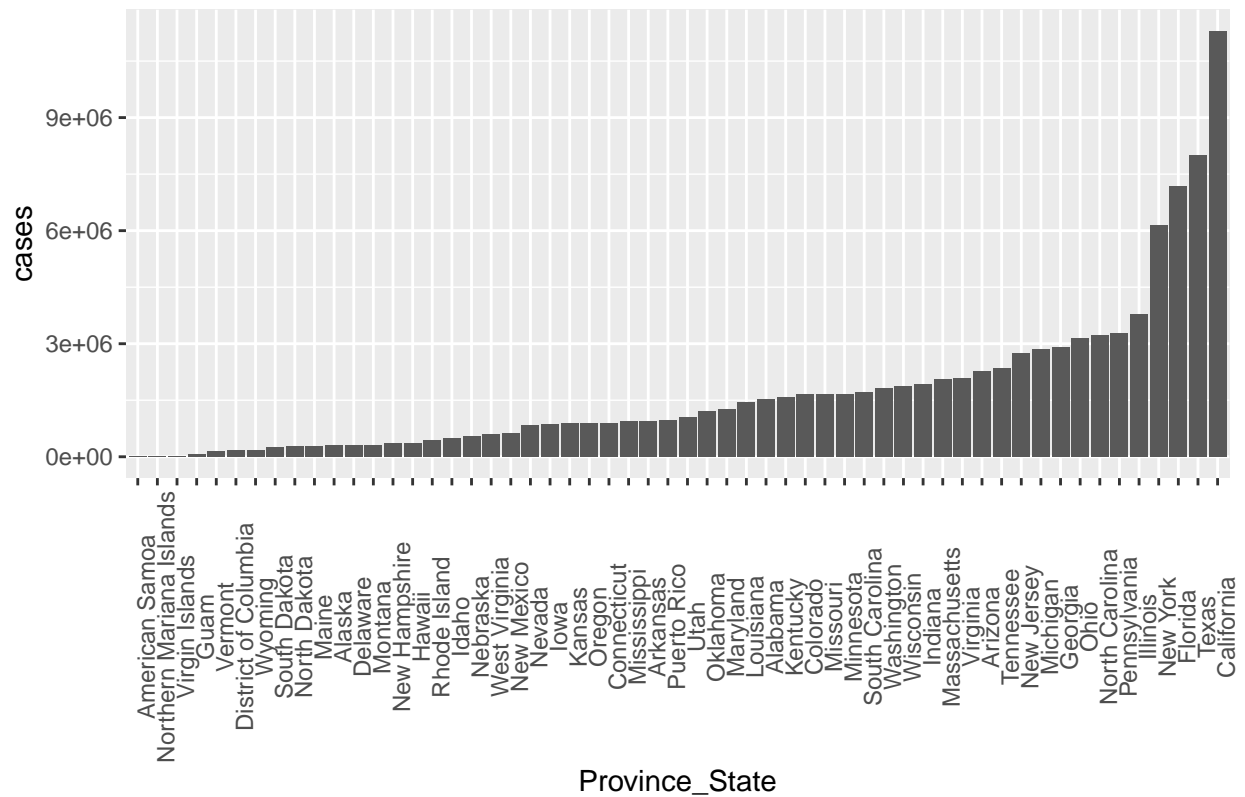
```

```

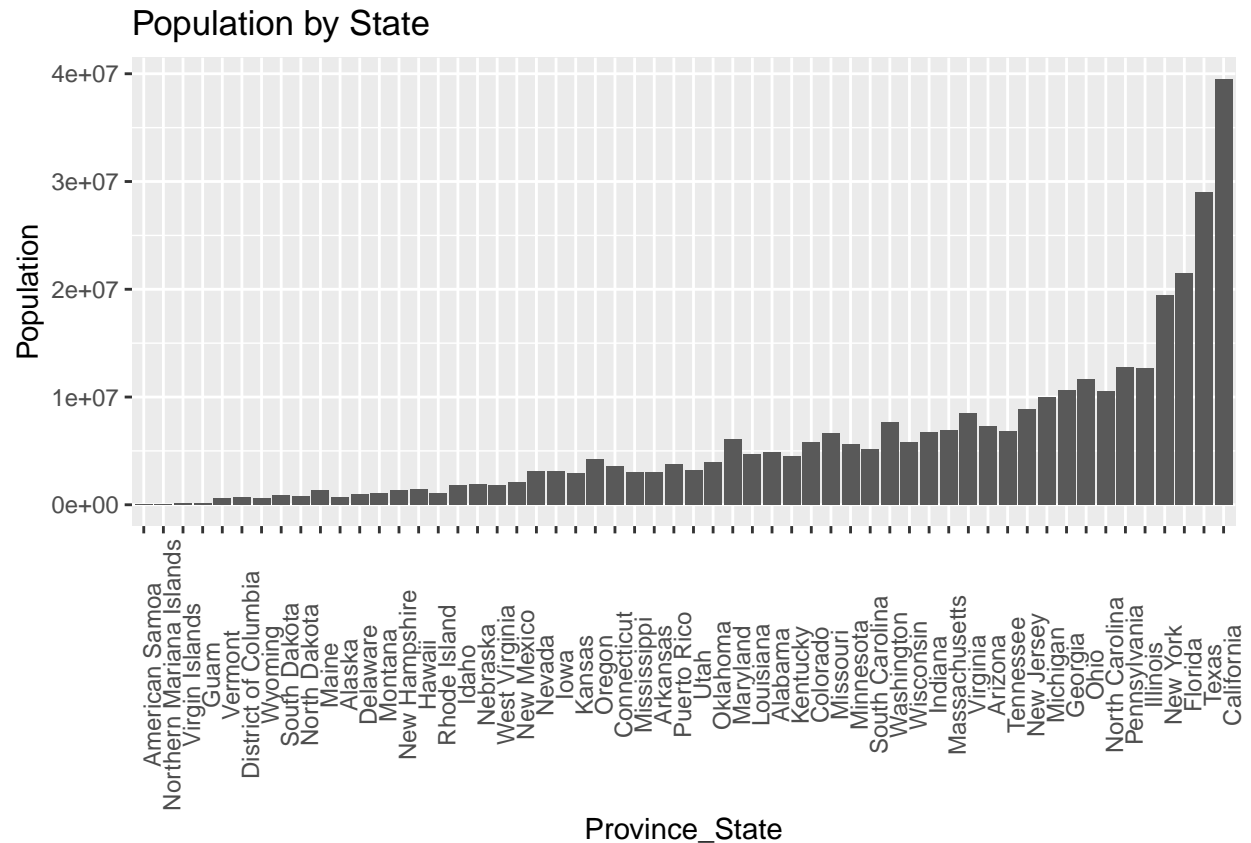
US_state_totals %>%
  mutate(Province_State = fct_reorder(Province_State, cases)) %>%
  filter(cases > 0) %>%
  ggplot(aes(x=Province_State, y=cases)) +
  geom_bar(stat="identity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "COVID19 Cases by State")

```

COVID19 Cases by State



```
US_state_totals %>%
  mutate(Province_State = fct_reorder(Province_State, cases)) %>%
  filter(cases > 0) %>%
  ggplot(aes(x=Province_State, y=Population)) +
  geom_bar(stat="identity") +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Population by State")
```



Analysis

When comparing the above two charts, it appears that the population count of a state is useful for predicting the total number of COVID cases. Although this is an obvious observation, I was curious to see if any outliers could raise more questions, such as a clear difference between states with various approaches to virus containment. To test this theory, I used a linear model to visualize the relationship between population and the number of COVID cases.

Model

```
mod <- lm(cases ~ Population, data = US_state_totals)
```

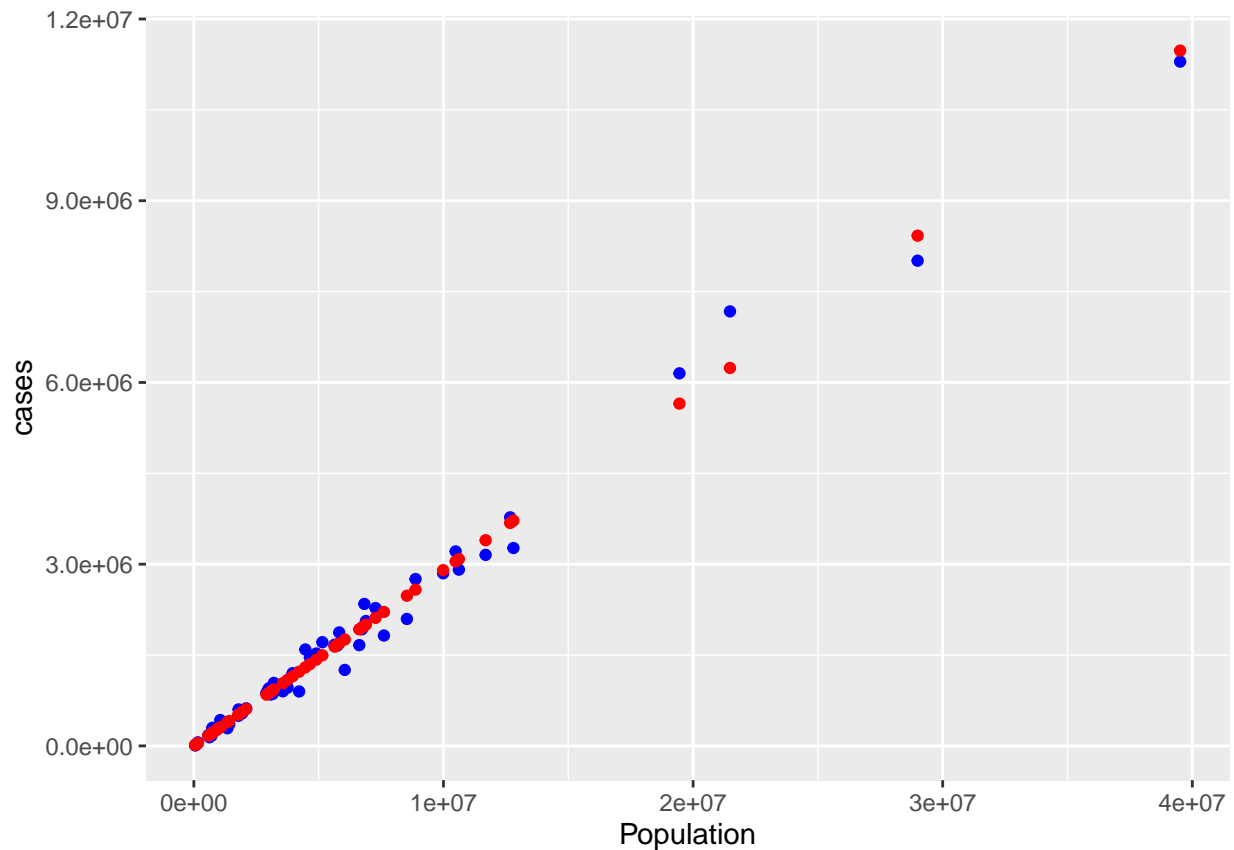
```
US_state_totals %>% mutate(pred = predict(mod))
```

```
## # A tibble: 56 x 5
##   Province_State   deaths    cases Population    pred
##   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
## 1 Alabama         20473  1525724  4903185  1424132.
## 2 Alaska           1393   298869   740995   215003.
## 3 American Samoa     34     8250    55641    15906.
## 4 Arizona          31406  2275235  7278717  2114231.
## 5 Arkansas          12276   953681  3017804   876423.
## 6 California        96217 11296777 39512223 11478160.
```

```
## 7 Colorado          13334 1658928  5758736 1672672.
## 8 Connecticut       11385  901180  3565287 1035468.
## 9 Delaware          3112  309804   973764  282623.
## 10 District of Columbia 1392  168678   705749  204764.
## # ... with 46 more rows
```

```
US_state_totals_w_pred <- US_state_totals %>% mutate(pred = predict(mod))

US_state_totals_w_pred %>% ggplot() +
  geom_point(aes(x= Population, y = cases), color = "blue") +
  geom_point(aes(x= Population, y = pred), color = "red")
```



Conclusion and Bias

In the above chart, the red dots are the predictions, and the blue dots are the actual values. Based on this model, I believe it is fair to conclude that population is a valuable predictor of the total number of COVID cases per state.

My bias for this analysis would be my expectation that a greater population would lead to a higher number of COVID cases.