COVID19_Data

2024-10-14

COVID-19 Data Analysis Project

This data analysis examines COVID-19 case and death data for US, and globally from January of 2020 to March of 2023. The original raw data sets include information about the US county where the case was recorded, population of the county, and the date of the case. This project aims to answer the question, what were the US COVID cases that resulted in death compared to the overall case, and how does Colorado compare to the other states overall?

Possible Bias

In any data science project, it is important to note the potential sources of bias and ensure they are identified to others consuming the information, as well as any appropriate mitigation steps are taken in analysis, if possible. Some possible sources of bias in this COVID-19 data set are;

1. Under-Reporting: At the height of the pandemic, many communities did not have the resources to support adequate testing. There were also periods where testing was either inaccessible or not free, which may have prevented some people from testing. This could cause the number of cases to be lower than reality.

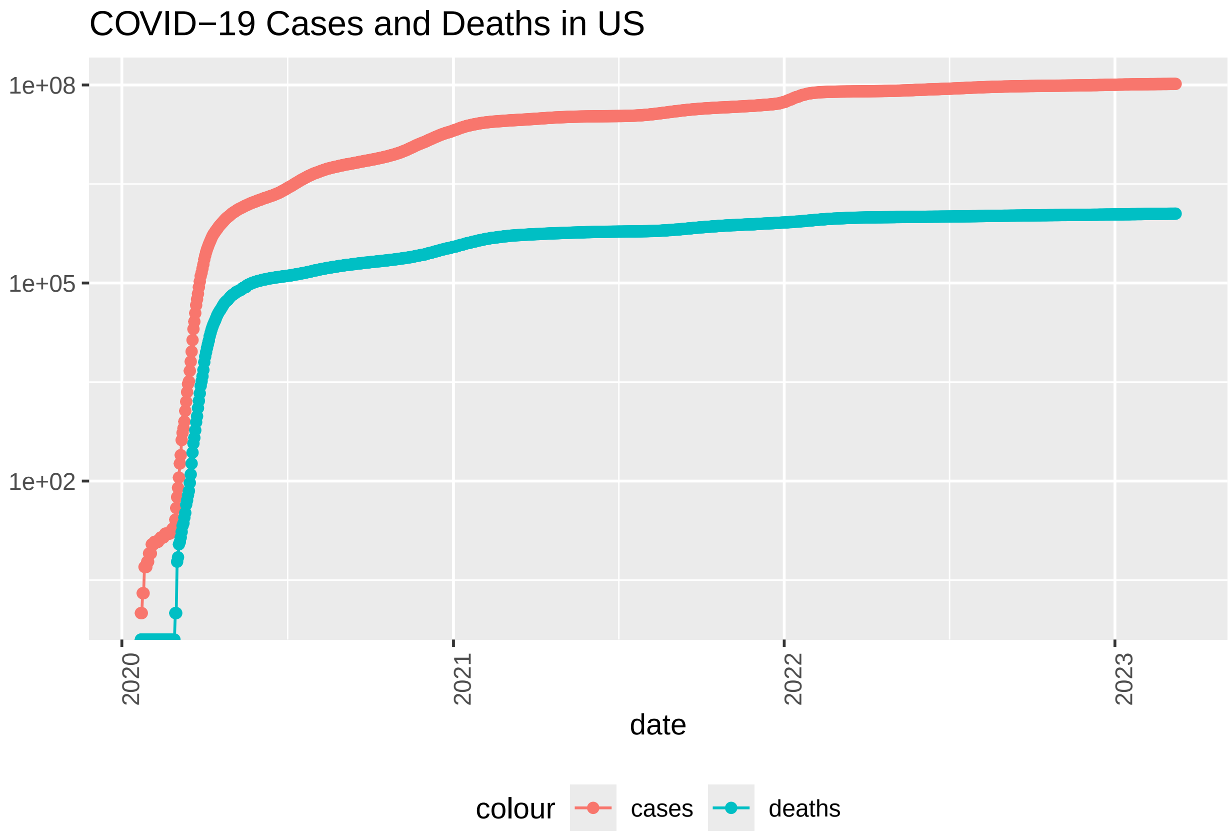
2. Asymptomatic Cases: For some individuals, COVID-19 symptoms were lessened or not noticeable at all, so those individuals may have not gotten tested but still carried or had the potential to spread the virus. This could cause the number of cases to be lower than reality.

3. Healthcare Funding: During the pandemic, some hospital's resourcing and federal aid was dependent on the number of COVID-19 cases being treated at the hospital. This could cause an inflation in the number of reported COVID-19 deaths or cases in order for hospitals to get essential resources to support patients.

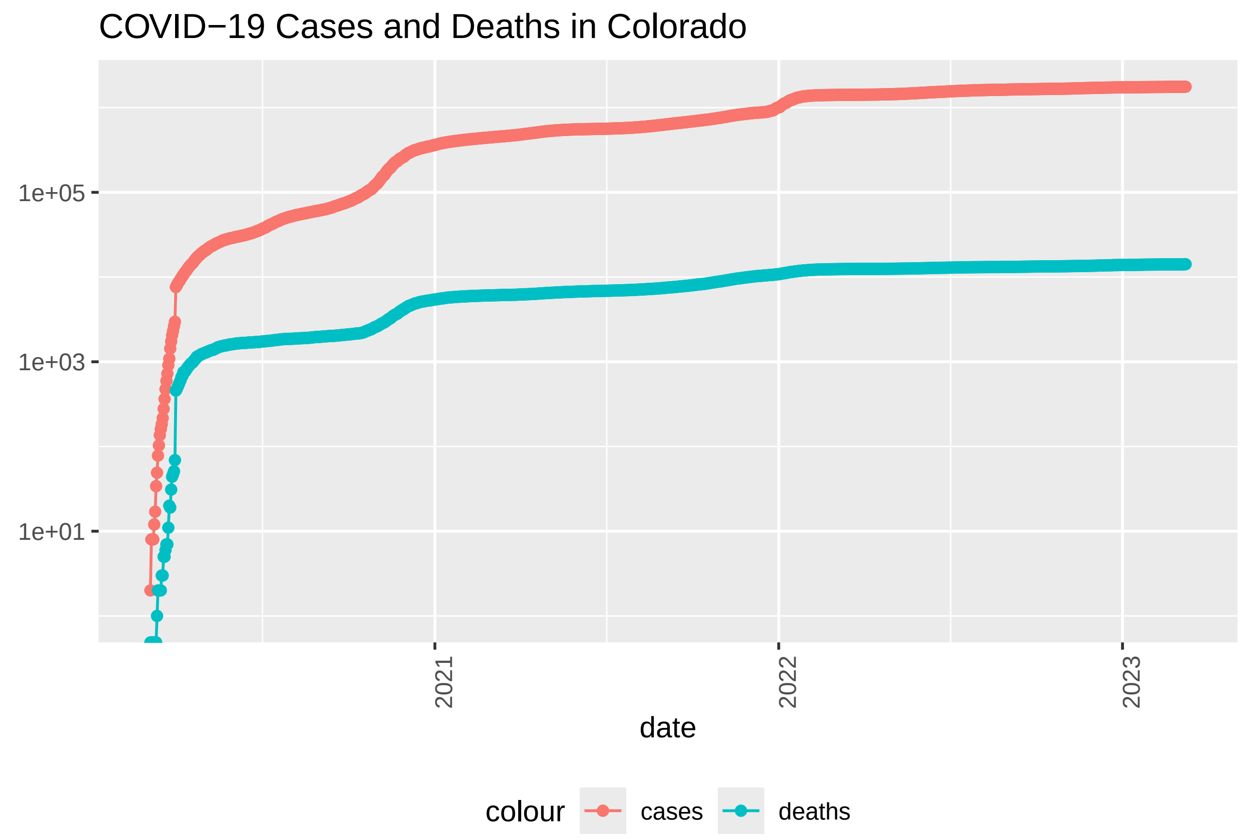
```
#Importing Data
library(tidyverse)
library(lubridate)
library(forecast)
library(mgcv)

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",  "time_series_covid19_confirmed_global.csv", "t

urls <- str_c(url_in,file_names)

global_cases <- read.csv(urls[2])
global_deaths <- read.csv(urls[4])
US_cases <- read.csv(urls[1])
US_deaths <- read.csv(urls[3])
```

```
#Looking at Global Cases
global_cases <- global_cases[, !(names(global_cases) %in% c("Admin2", "Lat", "Long_","UID", "iso2", "is
global_cases <- global_cases %>% rename(Country_Region = Country.Region, Province_State = Province.Stat
global_deaths <- global_deaths %>% rename(Country_Region = Country.Region, Province_State = Province.St

global_deaths <- global_deaths %>% pivot_longer(cols = -c('Province_State','Country_Region', 'Lat', 'Lo
global_cases <- global_cases %>% pivot_longer(cols = -c('Province_State','Country_Region'), names_to =

global_cases <- global_cases %>% mutate(date = gsub("^X", "", date)) %>% mutate(date = gsub("^(\\d)\\."
global_deaths <- global_deaths %>% mutate(date = gsub("^X", "", date)) %>% mutate(date = gsub("^(\\d)\\
global <- global_cases %>% full_join(global_deaths) %>% mutate(date = mdy(date))
global <- global %>% filter(cases > 0)

#Examining US Cases
US_cases <- US_cases[, !(names(US_cases) %in% c("UID", "Lat", "Long_","UID", "iso2", "iso3","code3", "F
US_cases <- US_cases %>% pivot_longer(cols = -c('Province_State','Country_Region','Combined_Key','Admin
US_cases <- US_cases %>% mutate(date = mdy(date))

US_deaths <- US_deaths[, !(names(US_deaths) %in% c("UID", "Lat", "Long_","UID", "iso2", "iso3","code3",
US_deaths <- US_deaths %>% pivot_longer(cols = -c('Province_State','Country_Region','Combined_Key','Adm
US_deaths <- US_deaths %>% mutate(date = mdy(date))

US <- US_cases %>% full_join(US_deaths)
US <- US %>% filter(cases > 0)
```

```
#Plotting US Data
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()

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()

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 Cases and Deaths in US", y= NULL)
```



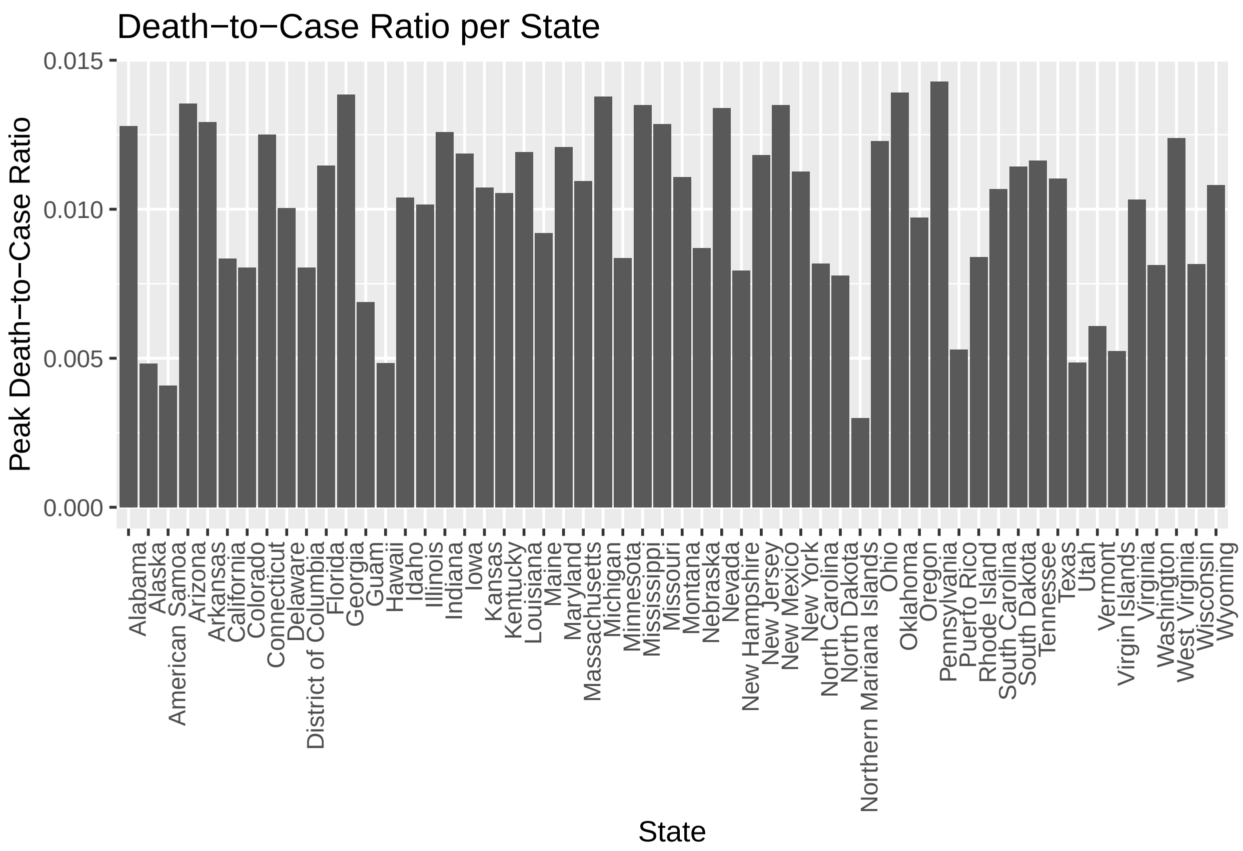
```
#Plotting Colorado Data
US_by_state %>%
     filter(Province_State == "Colorado") %>%
     filter(cases > 0) %>%
     ggplot (aes(x = date, y = cases)) +
     geom_line (aes (color = "cases")) +
     geom_point(aes(color = "cases")) +
     geom_line(aes (y = deaths, color = "deaths")) +
     geom_point (aes(y = deaths, color = "deaths")) +
     scale_y_log10() +
     theme(legend.position="bottom",
             axis.text.x = element_text(angle = 90)) +
     labs(title = "COVID-19 Cases and Deaths in Colorado", y= NULL)
```



Cases and Deaths Over Time

The trends in the national cases and deaths over the collection period are reflected in the Colorado trends as well. This makes sense, because largely when there were spikes in cases, those trends would quickly be reflected nationally.

```r
#Examining Peak Ratios for Death to Case variables
 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)

 #Plotting the Peak Death to Case Ratio Per State
 US_state_totals <- US_state_totals %>% mutate(death_case_ratio = deaths_per_thou / cases_per_thou)

 ggplot(US_state_totals, aes(x = Province_State, y = death_case_ratio)) + geom_col() +
    labs(title = "Death-to-Case Ratio per State", x = "State", y = "Peak Death-to-Case Ratio") +
    theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```r
#Examining Colorado Data
colorado_data <- US_cases %>% filter(Province_State == "Colorado")
colorado_data_clean <- colorado_data %>% group_by(date) %>% summarize(cases = sum(cases))
colorado_data_clean$date <- as.Date(colorado_data_clean$date)
US_state_totals %>%  slice_min(deaths_per_thou, n = 15)
```

```
## # A tibble: 15 x 7
##    Province_State        deaths   cases population cases_per_thou deaths_per_thou
##    <chr>                  <int>   <int>      <int>          <dbl>           <dbl>
##  1 American Samoa            34 8.32e3      55641           150.            0.611
##  2 Northern Mariana Isl~     41 1.37e4      55144           248.            0.744
##  3 Virgin Islands           130 2.48e4     107268           231.            1.21
##  4 Hawaii                  1841 3.81e5    1415872           269.            1.30
##  5 Vermont                  929 1.53e5     623989           245.            1.49
##  6 Puerto Rico             5823 1.10e6    3754939           293.            1.55
##  7 Utah                    5298 1.09e6    2785478           391.            1.90
##  8 District of Columbia    1432 1.78e5     705749           252.            2.03
##  9 Alaska                  1486 3.08e5     728809           422.            2.04
## 10 Washington             15683 1.93e6    7614893           253.            2.06
## 11 Maine                   2928 3.18e5    1344212           237.            2.18
## 12 New Hampshire           3003 3.78e5    1359711           278.            2.21
## 13 Oregon                  9373 9.64e5    4217737           228.            2.22
## 14 Colorado               14181 1.76e6    5758736           306.            2.46
## 15 Nebraska                4936 5.67e5    1934408           293.            2.55
## # i 1 more variable: death_case_ratio <dbl>
```

Peak Colorado COVID-19 Cases Compared to US Overall

As we can see in the peak death-to-case ratio plot, Colorado ranked amongst the lower states in terms of death-to-case ratio over the course of the pandemic. Colorado had the 14th lowest peak death-to-case ratio out of the US states and territories. Also of note, only one state (Washington) had a larger population but lower death-to-case ratio.

```
#Modeling US Data
mod = lm(deaths_per_thou ~ cases_per_thou, data = US_state_totals)

US_state_totals %>%  slice_max(cases_per_thou, n = 5)
```

```
## # A tibble: 5 x 7
##   Province_State deaths   cases population cases_per_thou deaths_per_thou
##   <chr>           <int>   <int>      <int>          <dbl>           <dbl>
## 1 Rhode Island     3870  460697    1059361           435.            3.65
## 2 Alaska           1486  307655     728809           422.            2.04
## 3 Utah             5298 1090346    2785478           391.            1.90
## 4 Kentucky        18130 1718471    4467673           385.            4.06
## 5 North Dakota     2232  286950     762062           377.            2.93
## # i 1 more variable: death_case_ratio <dbl>
```

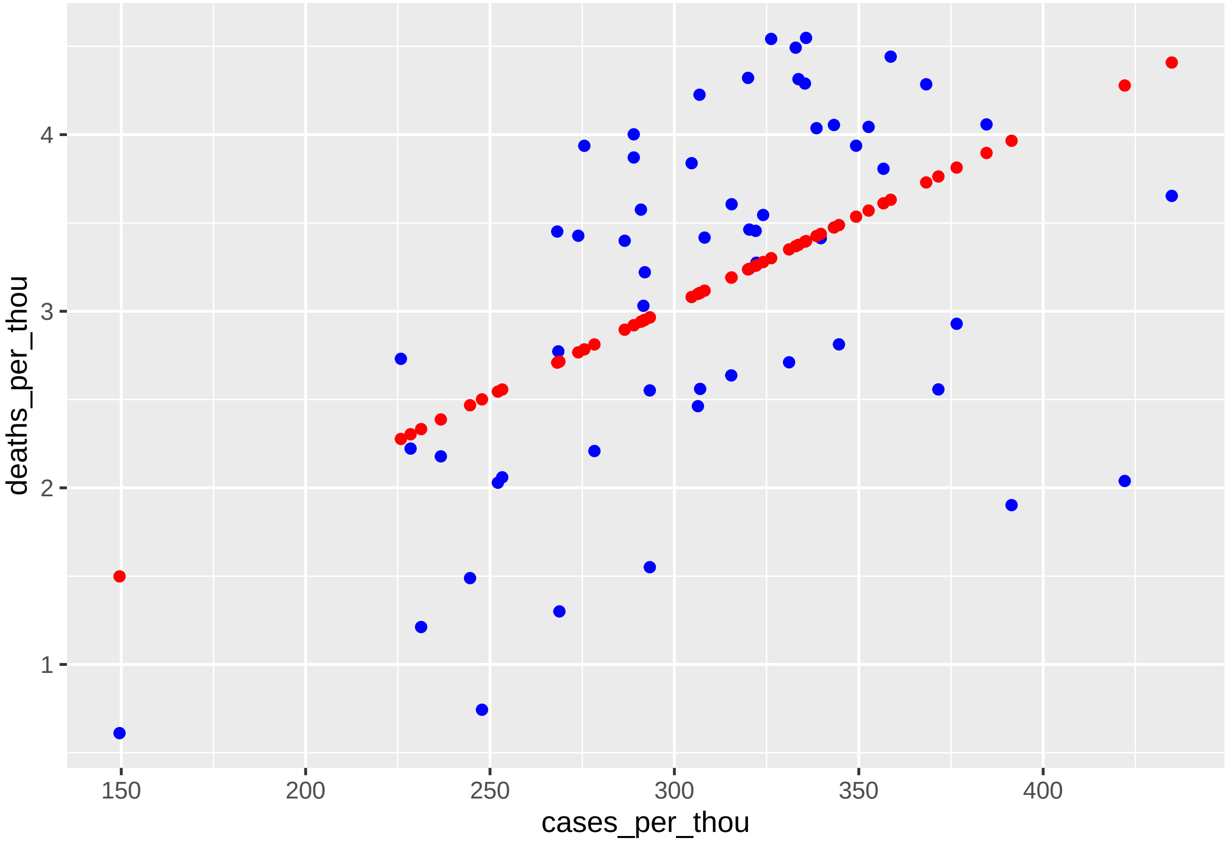
```
US_state_totals %>%  slice_max(deaths_per_thou, n = 5)
```

```
## # A tibble: 5 x 7
##   Province_State deaths   cases population cases_per_thou deaths_per_thou
##   <chr>           <int>   <int>      <int>          <dbl>           <dbl>
## 1 Arizona         33102 2443514    7278717           336.            4.55
## 2 Oklahoma        17972 1290929    3956971           326.            4.54
## 3 Mississippi     13370  990756    2976149           333.            4.49
## 4 West Virginia    7960  642760    1792147           359.            4.44
## 5 New Mexico       9061  670929    2096829           320.            4.32
## # i 1 more variable: death_case_ratio <dbl>
```

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

```
## # A tibble: 56 x 8
##    Province_State       deaths   cases population cases_per_thou deaths_per_thou
##    <chr>                 <int>   <int>      <int>          <dbl>           <dbl>
##  1 Alabama               21032 1.64e6    4903185           335.            4.29
##  2 Alaska                 1486 3.08e5     728809           422.            2.04
##  3 American Samoa           34 8.32e3      55641           150.            0.611
##  4 Arizona               33102 2.44e6    7278717           336.            4.55
##  5 Arkansas              13020 1.01e6    3017804           334.            4.31
##  6 California           101159 1.21e7   39512223           307.            2.56
##  7 Colorado              14181 1.76e6    5758736           306.            2.46
##  8 Connecticut           12220 9.77e5    3565287           274.            3.43
##  9 Delaware               3324 3.31e5     973764           340.            3.41
## 10 District of Columbia   1432 1.78e5     705749           252.            2.03
## # i 46 more rows
## # i 2 more variables: death_case_ratio <dbl>, pred <dbl>
```

```
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") +
```



Linear Model

As we can see in the model prediction above, the deaths per thousand and the cases per thousand don't show a strong linear relationship overall with respect to US data, but given the small range in deaths per thousand, there is a decent accuracy overall. The fact that this relationship isn't perfectly linear makes sense, because one would assume the number of deaths would decrease over time, as prevention methods and vaccines were more widely accessible and understood. While this graph isn't time series, it does include data from the entire time span (2020-2023), so one could draw the conclusion that because the deaths per case average went down over time, there may not be a noticeable trend of death/case ratio when looking over the entire duration of the pandemic.

Conclusion

In conclusion, I was able to analyze the COVID-19 dataset to understand the trends in cases, deaths, and the ratio of the two on a global scale, as well as a comparison between the US as a whole, and Colorado as an individual state.