Covid-19 Analysis

Morgan I. MacKay

2024-09-21

Setup and Goal

The data for this project consists of Covid-19 confirmed cases and confirmed deaths information provided by John Hopkins University. The data can be found at the following link:https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series). The general structure of the data is that the first few columns identify what US county is being discussed, then the following columns each represent a date and how many deaths/confirmed cases were recorded.

The goal of my analysis is to get a better understanding of how the Covid-19 pandemic effected different communities. I hope to do this by graphing the confirmed deaths by each state over time. I also want to plot the death data against population, because I think that evauating the correlation could lead to valuable insights. Finally, I will attempt to find the correlation between per capita GDP and state mortality rate. I will plot the results on a scatter plot.

For this project I expect to need the following libraries: Lubridate, Readr, and Tidyverse. Lubridate to manipulate different date formats, Readr to import different data types, and the Tidyverse for a variety of data manipulation purposes.

```
library(lubridate)
library(readr)
library(tidyverse)

# Loading messages have been hidden for simplicity
```

###Loading Data

I downloaded the data I wanted to work with onto my personal computer. I then used the “Import Dataset” feature to locate, preview, and import the data. The code shown below represents the commands used to import that data.

After importing each data set, I renamed them to something more understandable. Then I dropped some of the columns I knew I wouldn’t need.

```
time_series_covid19_deaths_US <- read_csv("~/Downloads/time_series_covid19_deaths_US.csv")
usDeaths <- time_series_covid19_deaths_US

usDeaths <- usDeaths %>%
  select(-Admin2, -Lat, -Long_, -iso2, -iso3, -FIPS, -Country_Region, -code3)

# Loading messages have been hidden for simplicity
```

###Cleaning Data

To clean the data I wanted to perform a couple of checks. First, I wanted to make sure that none of the rows or columns contained null values. Additionally, there should be no negative values in any of the areas that are meant to record population.

The second main check I did was to make sure that the deaths in each county never totaled more than the full population for that county.

I combined the “error” rows into a variable called “rows_to_drop”. I then removed them from the data set I was working with. With so many different observations, dropping 67 rows wouldn’t make much of a difference.

```
# First check: Rows with negative or null values in columns 4 onwards
negative_or_null_rows <- usDeaths %>%
  filter(if_any(4:ncol(.), ~ . < 0 | is.na(.)))

# Second check: Rows where values in columns 5 onwards are greater than column 4
violating_rows <- usDeaths %>%
  rowwise() %>%
  filter(any(c_across(5:ncol(.)) > `Population`)) %>%
  ungroup()

# Combine the two sets of rows to drop (remove duplicates if needed)
rows_to_drop <- bind_rows(negative_or_null_rows, violating_rows) %>%
  distinct()

# Drop the rows from the original dataset
cleaned_usDeaths <- usDeaths %>%
  anti_join(rows_to_drop)

# View the cleaned dataset
head(cleaned_usDeaths)
```

###Further Data Manipulation

To make the data easier to work with, I converted it from a “wide” data set to a “tall” set. I followed the instructions from class pretty closely for this. I also realized that the data set and the computing platform I was using weren’t the best match. Each operation was taking a considerable amount of time with so many oberservations. My first step to fix this was to drop the US territories involved in the data. Finally, I called the “head” function to take a look at what I had created.

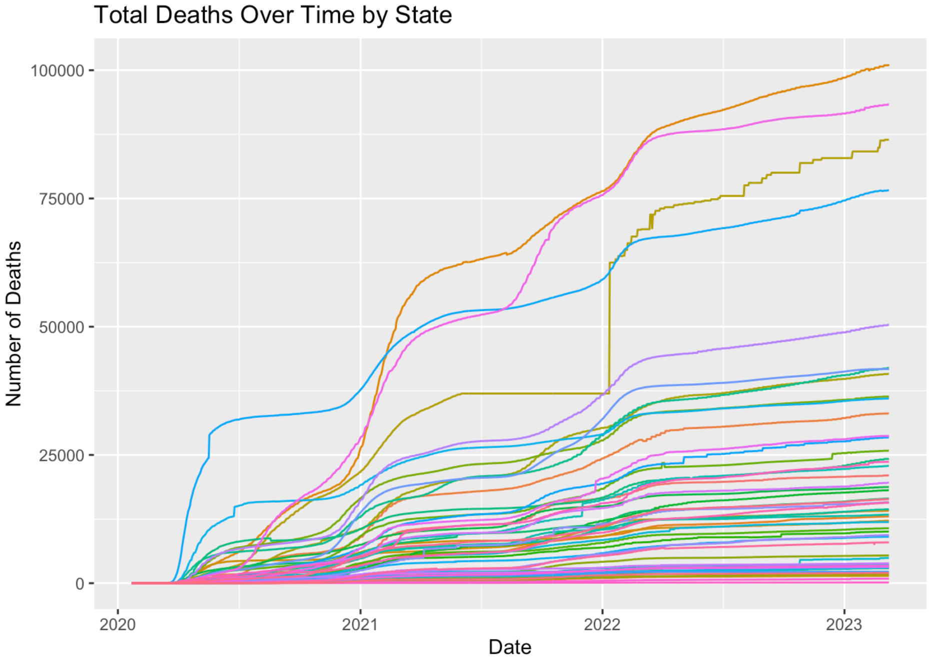
```
long_death_data <- cleaned_usDeaths %>%
  pivot_longer(cols = 5:ncol(.),  # Columns 5 and onward are the dates
               names_to = "Date",  # Create a 'Date' column
               values_to = "Deaths")
long_death_data$Date <- mdy(long_death_data$Date)

long_death_data <- long_death_data %>%
  filter(!(Province_State %in% c("Diamond Princess", "Northern Mariana Islands", "American Samoa", "Guam", "Puert
o Rico")))


head(long_death_data)
```

```
## # A tibble: 6 × 6
##        UID Province_State Combined_Key         Population Date       Deaths
##      <dbl> <chr>          <chr>                     <dbl> <date>      <dbl>
## 1 84001001 Alabama        Autauga, Alabama, US      55869 2020-01-22      0
## 2 84001001 Alabama        Autauga, Alabama, US      55869 2020-01-23      0
## 3 84001001 Alabama        Autauga, Alabama, US      55869 2020-01-24      0
## 4 84001001 Alabama        Autauga, Alabama, US      55869 2020-01-25      0
## 5 84001001 Alabama        Autauga, Alabama, US      55869 2020-01-26      0
## 6 84001001 Alabama        Autauga, Alabama, US      55869 2020-01-27      0
```

###Graphing Deaths by State Over Time

My first visualization is a graph of deaths over time, grouped by state. I think this is a good exercise to do with this data set because it gives the user a high level overview of the data and its trends. When looking at the graph produced by the code below, we can see distinct surges in covid deaths across states. You can also see distinct outlier states that had much sharper death rates than others. Another interesting trend is that certain states seemed to have reporting issues. We know that deaths don’t follow rigid patterns, yet we see distinct “steps” in the amount of covid deaths being reported by different states.

```
# Summarizing by state (or county)
state_deaths <- long_death_data %>%
  group_by(Province_State, Date) %>%  # Group by state and date
  summarise(Total_Deaths = sum(Deaths, na.rm = TRUE)) %>%  # Summarize total deaths per date
  ungroup()

# Plotting the total deaths over time for each state
ggplot(state_deaths, aes(x = Date, y = Total_Deaths, color = Province_State, group = Province_State)) +
  geom_line() +
  labs(title = "Total Deaths Over Time by State",
       x = "Date",
       y = "Number of Deaths") +
  theme(legend.position = "none")
```



###Focusing on the Rockies

I was starting to lose my temper with how long my computer was taking to process the commands I was giving it for this data set. I decided to focus just on the broader set of Rocky Mountain states. The code below does that, and then calls the “head” function to review.

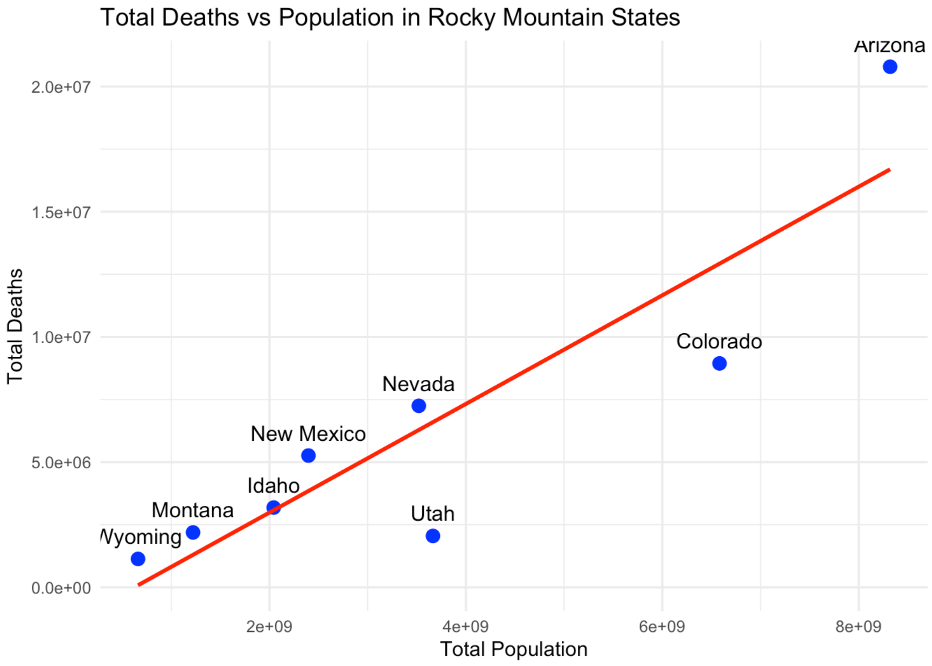
```
# Data set is too large, if it's not rocky, it's got to go.

rocky_mountain_states <- c("Colorado", "Idaho", "Montana", "Nevada", "New Mexico", "Utah", "Wyoming", "Arizona")

rocky_mountain_data <- long_death_data %>%
  filter(Province_State %in% rocky_mountain_states)

state_summary <- rocky_mountain_data %>%
  group_by(Province_State) %>%
  summarise(total_population = sum(Population, na.rm = TRUE),
            total_deaths = sum(Deaths, na.rm = TRUE))

head(rocky_mountain_data)
```

```
## # A tibble: 6 × 6
##        UID Province_State Combined_Key        Population Date       Deaths
##      <dbl> <chr>          <chr>                    <dbl> <date>      <dbl>
## 1 84004001 Arizona        Apache, Arizona, US      71887 2020-01-22      0
## 2 84004001 Arizona        Apache, Arizona, US      71887 2020-01-23      0
## 3 84004001 Arizona        Apache, Arizona, US      71887 2020-01-24      0
## 4 84004001 Arizona        Apache, Arizona, US      71887 2020-01-25      0
## 5 84004001 Arizona        Apache, Arizona, US      71887 2020-01-26      0
## 6 84004001 Arizona        Apache, Arizona, US      71887 2020-01-27      0
```

###Total Deaths VS. Population

I plotted the total amount of covid deaths for each state with their total population to form a scatter plot. If every state handled things in the exact same way, then we should see a straight line or perfect correlation between population size and death size. This graph doesn’t show that, so we can assume that there were differences across the states that lead to more or less deaths. The P value in this case was considered to be statistically significant, so we can say that there is correlation between higher population and higher deaths.

```
ggplot(state_summary, aes(x = total_population, y = total_deaths)) +
  geom_point(color = "blue", size = 3) +
  geom_text(aes(label = Province_State), vjust = -1, hjust = 0.5, size = 4) +  # Add state labels
  geom_smooth(method = "lm", se = FALSE, color = "red") +  # Add trend line
  labs(title = "Total Deaths vs Population in Rocky Mountain States",
       x = "Total Population",
       y = "Total Deaths") +
  theme_minimal()
```



```
# Fit a linear model
lm_model <- lm(total_deaths ~ total_population, data = state_summary)
summary(lm_model)$coefficients
```

```
##                       Estimate   Std. Error    t value    Pr(>|t|)
## (Intercept)      -1.351641e+06 1.920887e+06 -0.7036545 0.508021006
## total_population  2.168891e-03 4.432995e-04  4.8926078 0.002731195
```

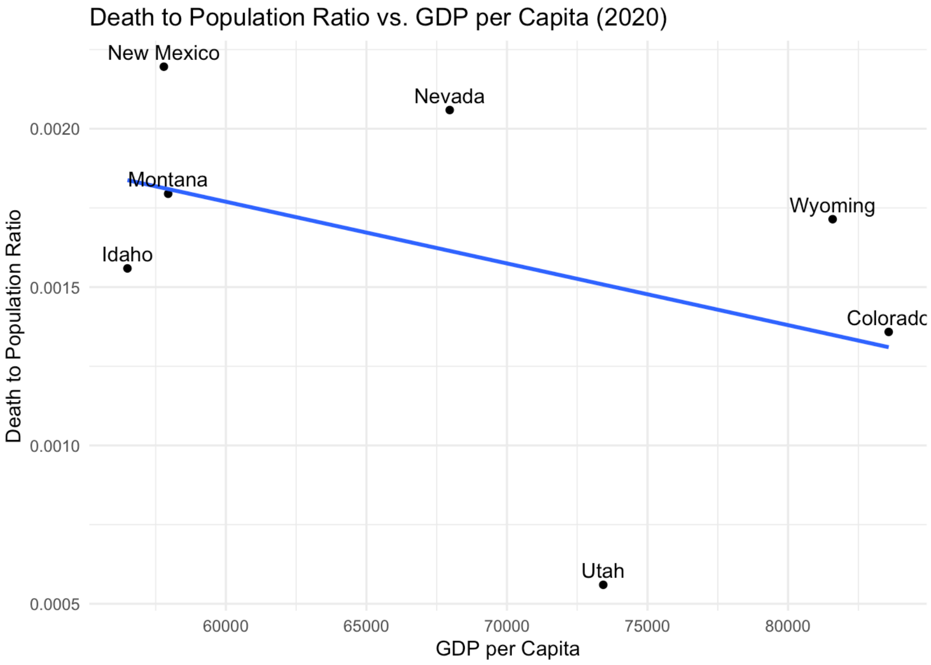
###Death Ratio VS. GDP per Capita

Finally I wanted to see what the relation was like between total deaths/ population and GDP per captia. I used the existing information on state covid deaths and populations, as well as economic data from the Bureau of Economic Analysis.

The analysis turned out to not be the best. Although there were poorer states with higher death rations, and richer states with lower death ratios, the data was too scattered to make a very good model or prediction. Instead what we’re left with is a general notion that as wealth increases deaths tend to fall relatively, likely because of additional access to healthcare resources or other benefits wealth brings. Unfortunately the model is not good enough to elaborate further.

```
# Covid deaths ratio compared to GDP per capita.

rocky_data <- long_death_data %>%
  filter(Province_State %in% c("Colorado", "Idaho", "Montana", "Nevada", "New Mexico", "Utah", "Wyoming")) %>%
  group_by(Province_State) %>%
  summarize(total_deaths = sum(Deaths, na.rm = TRUE),
            total_population = sum(Population, na.rm = TRUE)) %>%
  mutate(death_to_population_ratio = total_deaths / total_population)

gdp_data <- data.frame(
  Province_State = c("Colorado", "Idaho", "Montana", "Nevada", "New Mexico", "Utah", "Wyoming", "Arizona"),
  GDP_percap = c(83580, 56496, 57945, 67962, 57792, 73424, 81586, 64010)
)

rocky_data <- rocky_data %>%
  left_join(gdp_data, by = "Province_State")

ggplot(rocky_data, aes(x = GDP_percap, y = death_to_population_ratio, label = Province_State)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_text(vjust = -0.5) +
  labs(title = "Death to Population Ratio vs. GDP per Capita (2020)",
       x = "GDP per Capita",
       y = "Death to Population Ratio") +
  theme_minimal()
```

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



###Bias and Conclusion

The analysis performed above provided a good overview of the data set and some of the major trends contained within. Clear spikes in covid deaths were identified, discrepancies in how states reacted to the covid pandemic were supported, and a brief look into the effects of income were seen.

The data and analysis are subject to a few different sources of bias. Some of the main issues are the following:

- Issues in Reporting - as discussed earlier, there seem to be reporting issues in the data. Anytime deaths are involved the specifics of those deaths can lead to reporting errors. This is due to a number of reasons. Limited resources in certain areas of the country could have also contributed to less than ideal reporting practices.
- Personal Bias - I myself am someone who believes that wealth and income can buy better healthcare outcomes in the US. This, as well as other personal biases, could have impacted how I reviewed and judged the analysis.