US Covid-19 Analysis

Brandon Fry

2/10/2022

Overview

Covid-19 has made a significant impact on the world, and has affected almost every person in every corner of the planet. Even after 2 years, it still continues to cause thousands of new cases and deaths. Understanding it's behavior and patterns is critical in determining a response and reducing its impact. We can use data science to help understand it's behavior and determine an appropriate response that best limits its impact.

This analysis aims to look at the comparison between cases and deaths, and how the cases and deaths in the state of Washington compare to the country as a whole. Since Washington employs some of the strictest mandates, we can compare how those mandates are working compared to the country and understand how effective they are.

Data

The data we used for this analysis was provide by Johns Hopkins University of Medicine, and provides daily case and death records since the beginning of 2020. There is alot of unnecessary data for this analysis that will be removed when cleaning the data. The next step to preparing the data was to reorganize the long rows of dates into individual rows that could be more easily plotted. The final stage of preparation was to combine all of the state totals for the US dataset, and filter Washington stats for the Washington state dataset. We also filtered all of the data to remove any dates that had 0 cases.

To determine how Washington state compares with the total US, we simply look at the percentage of cases in the US that belong to Washington over time. To get this data, we calculate the percentage across all rows between the US and Washington case and death columns, and create a new column in the Washington dataset that shows the percentage of cases.

```
library("tidyverse")
library(plyr)

# Gather data from source
raw_github_url = "https://raw.githubusercontent.com/"
case_data_url = "CSSEGISandData/COVID-19/master/csse_covid_19_data/"
file_url = "csse_covid_19_time_series/time_series_covid19_confirmed_US.csv"
case_data <- paste0(raw_github_url, case_data_url, file_url) %>%
    read_csv()
raw_github_url = "https://github.com/"
death_data_url = "CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/"
file_url = "csse_covid_19_time_series/time_series_covid19_deaths_US.csv"
death_data <- paste0(raw_github_url, death_data_url, file_url) %>%
    read_csv()
```

```
library(lubridate)
state_case_data <- subset(case_data, select = -c(UID, FIPS, iso2,
    iso3, code3, Admin2, Lat, Long_, Combined_Key))
state_death_data <- subset(death_data, select = -c(UID, FIPS,
    iso2, iso3, code3, Admin2, Lat, Long_, Combined_Key, Population))

state_case_data <- state_case_data %>%
    pivot_longer(cols = -c("Province_State", "Country_Region"),
        names_to = "date", values_to = "cases")
state_case_data_agg <- aggregate(cases ~ ., data = state_case_data,
    FUN = sum)

state_death_data <- state_death_data %>%
    pivot_longer(cols = -c("Province_State", "Country_Region"),
        names_to = "date", values_to = "deaths")
state_death_data_agg <- aggregate(deaths ~ ., data = state_death_data,
    FUN = sum)

state_comb_data <- state_case_data_agg %>%
    full_join(state_death_data_agg) %>%
    mutate(date = mdy(date))

country_case_data <- state_case_data %>%
    select(-c(Province_State))

country_case_data_agg <- aggregate(cases ~ ., data = country_case_data,
    FUN = sum)

country_death_data <- state_death_data %>%
    select(-c(Province_State))

country_death_data_agg <- aggregate(deaths ~ ., data = country_death_data,
    FUN = sum)

country_comb_data <- country_case_data_agg %>%
    full_join(country_death_data_agg) %>%
    mutate(date = mdy(date))
```

```
wa_cases <- state_comb_data %>%
    filter(Province_State == "Washington")
```

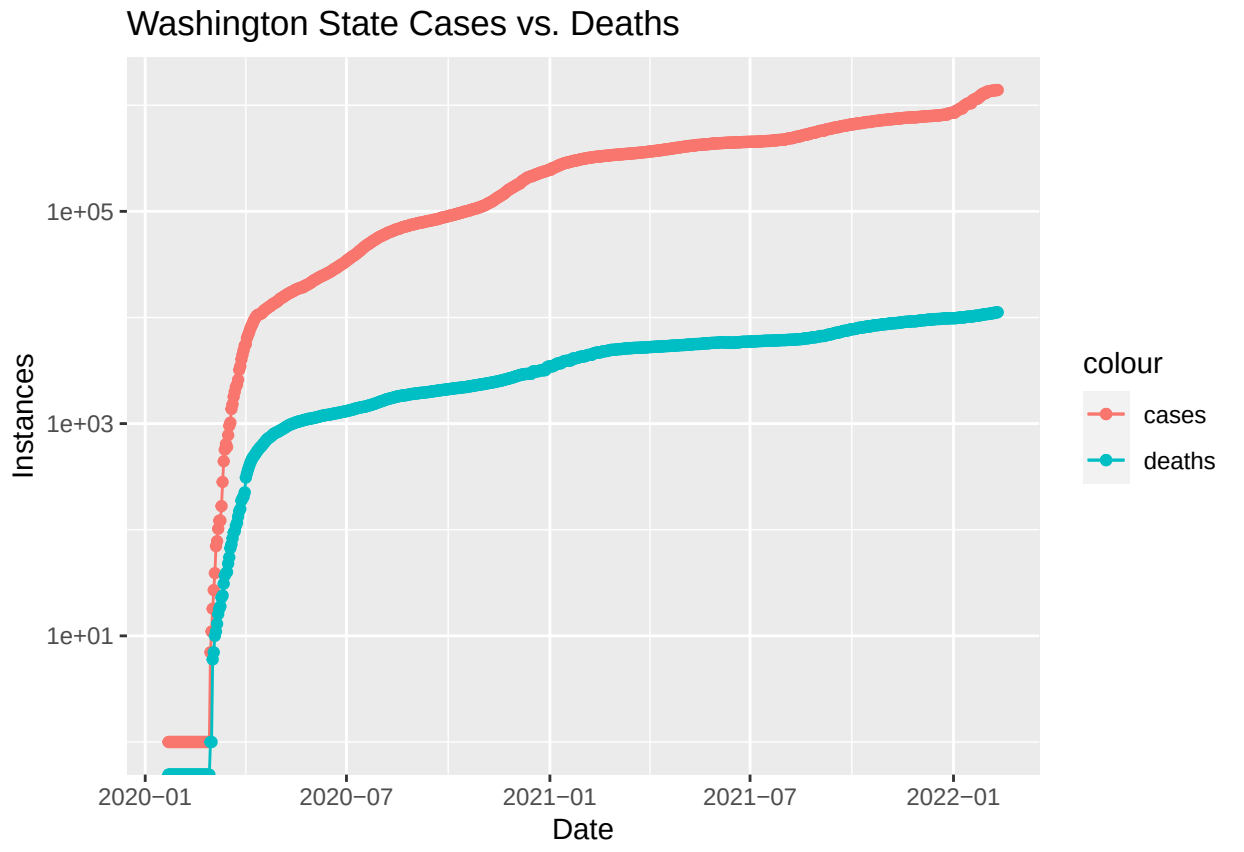
```
wa_cases <- mutate(wa_cases, case_percent = (wa_cases$cases/country_comb_data$cases) *
    100)
wa_cases <- mutate(wa_cases, death_percent = (wa_cases$deaths/country_comb_data$deaths) *
    100)
```

```
# filter out cases > 0
wa_cases <- wa_cases %>%
    filter(cases > 0)
country_comb_data <- country_comb_data %>%
    filter(cases > 0)
```

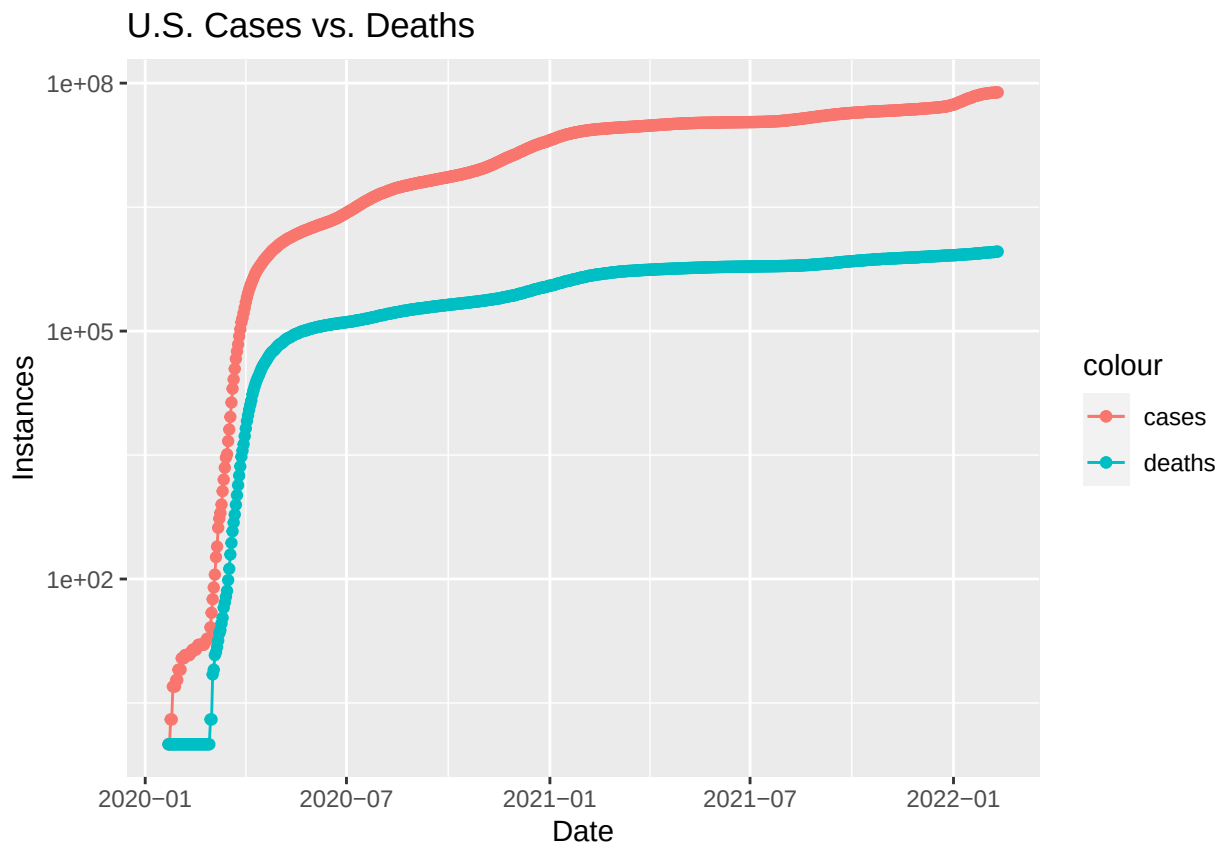
Data Analysis

To being our Analysis, lets first take a look at the Total US cases and deaths as well as Washington's cases and deaths. It's important for us to get a baseline of how the virus has moved through the country as well as the state of Washington, and how deaths have followed a similar pattern. To better visualize how each compare to eachother, we will view the data on a log scale, which makes the relationship more obvious. What we see from these graphs is that total deaths follow a similar pattern to cases. This observation is to be expected, as the virus has an expected mortality rate.

```
wa_cases %>%
    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() + labs(title = "Washington State Cases vs. Deaths",
    x = "Date", y = "Instances")
```

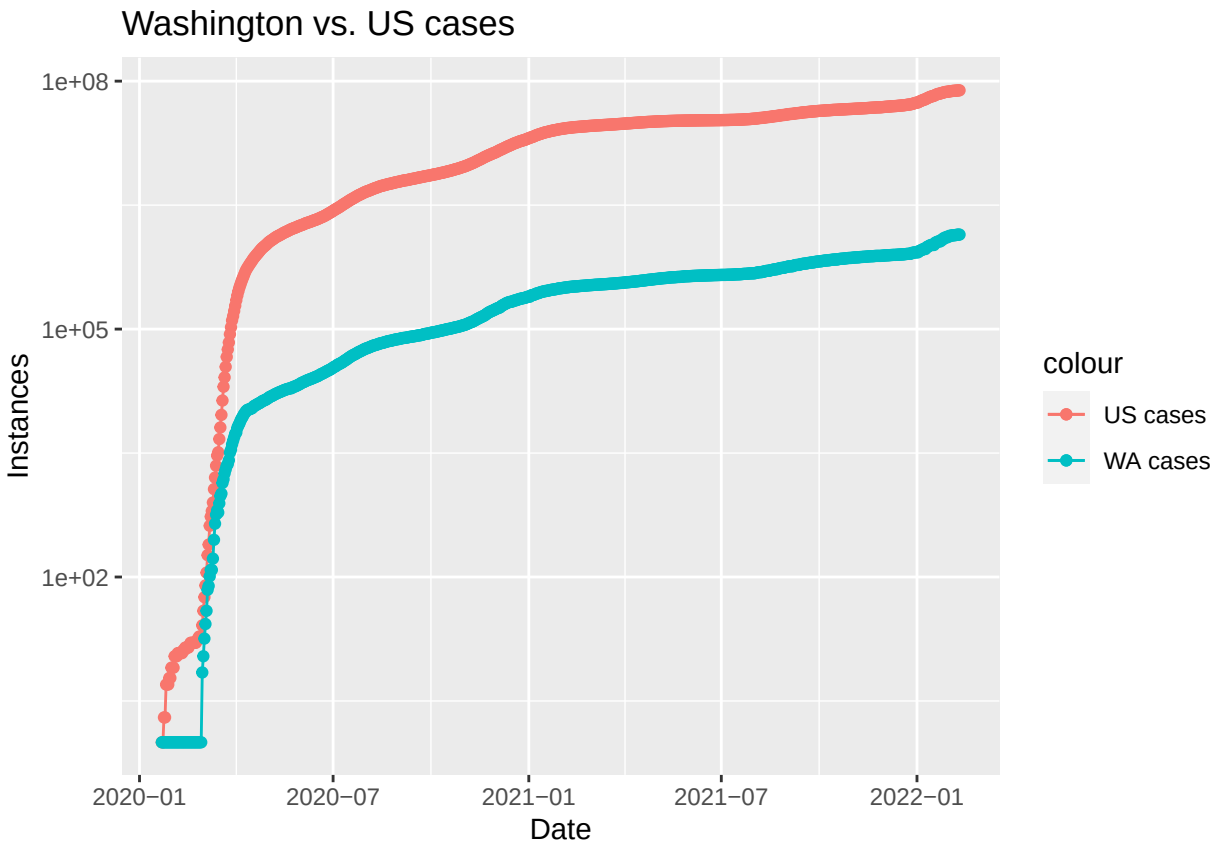



```
country_comb_data %>%
    ggplot(aes(x = date, y = cases)) + geom_line(aes(color = "cases")) +
    geom_point(aes(color = "cases")) + geom_line(aes(y = deaths,
    color = "deaths")) + geom_point(aes(y = deaths, color = "deaths")) +
    scale_y_log10() + labs(title = "U.S. Cases vs. Deaths", x = "Date",
    y = "Instances")
```



To continue the analysis, we will split the two graphs and compare US cases to Washington cases, and US deaths to Washington deaths. From this view, its somewhat difficult to see how they compare, as Washington cases and deaths follow a very similar pattern to the US cases and deaths. To better understand how Washington compares with the US, we can look at the percentage of US cases and deaths that Washington contributes.

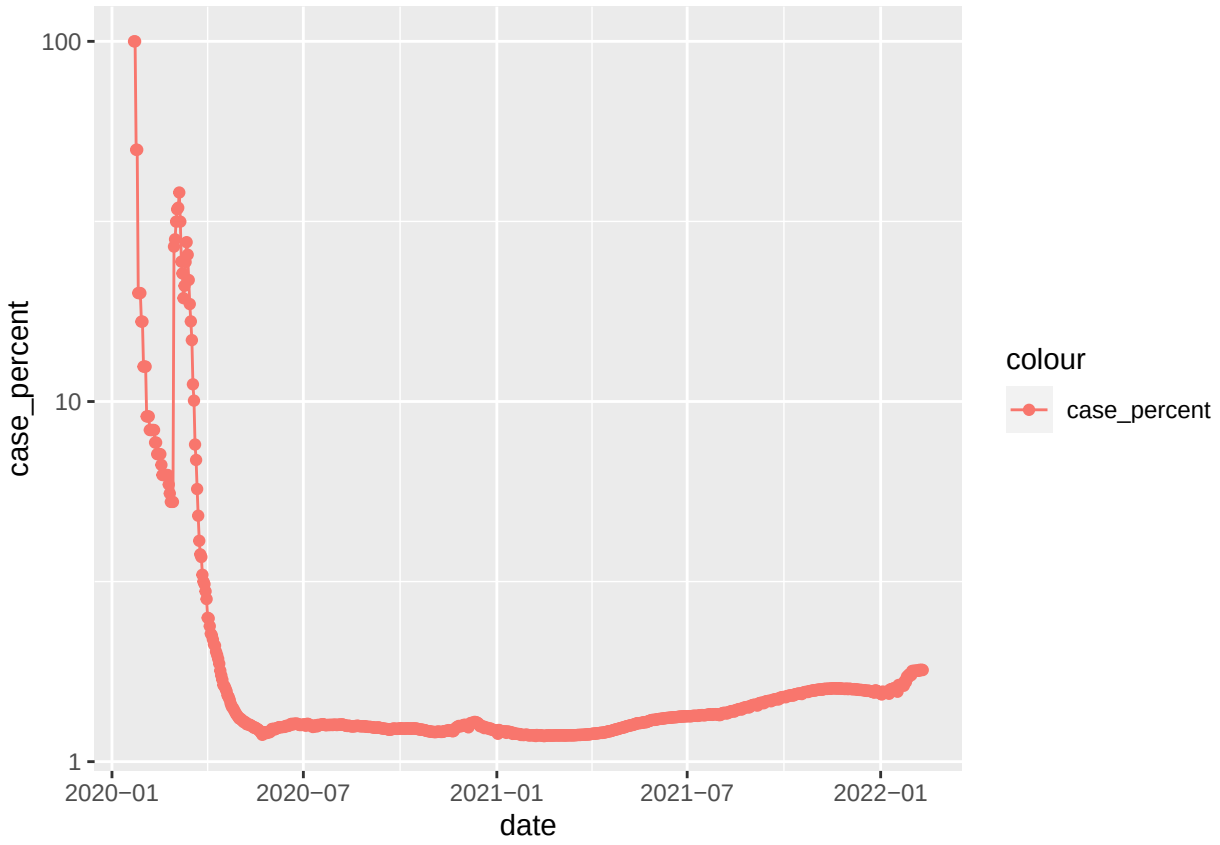
```
country_comb_data %>%
    ggplot(aes(x = date, y = cases)) + geom_line(aes(color = "US cases")) +
    geom_point(aes(color = "US cases")) + geom_line(data = wa_cases,
    aes(y = cases, color = "WA cases")) + geom_point(data = wa_cases,
    aes(y = cases, color = "WA cases")) + scale_y_log10() + labs(title = "Washington vs. US cases",
    x = "Date", y = "Instances")
```



Washington Cases vs US Cases

To understand how Washington compares to the United States in total cases, we look at the percentage of US cases that are from Washington. What we see from the data below, is that Washington cases are increasing at a faster rate than the US as a whole. This data is somewhat surprising, as Washington has some of the strictest mandates to help reduce the spread of COVID. There are some potential reasons as to why this maybe that will be discussed in the Error and Bias section below.

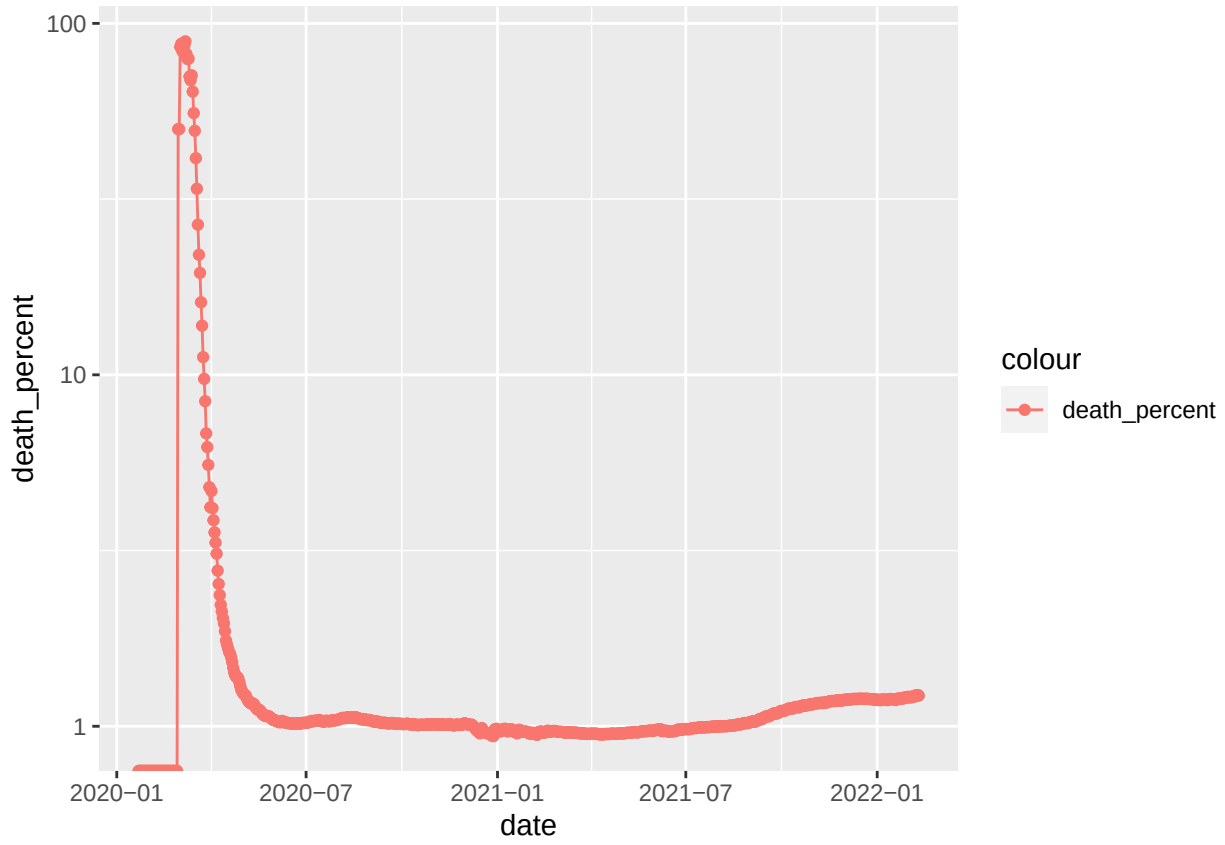
```
wa_cases %>%
    ggplot(aes(x = date, y = case_percent)) + geom_line(aes(color = "case_percent")) +
    geom_point(aes(color = "case_percent")) + scale_y_log10()
```



Washington Deaths vs US Deaths

Another key comparison we want to analyze is how the deaths in Washington compare to the deaths in the US. What we see is a very similar observation as we did with the cases, which is to be expected. We see a continuous increase in the deaths relative to the country, which again is suprising considering the mandates.

```
wa_cases %>%
    ggplot(aes(x = date, y = death_percent)) + geom_line(aes(color = "death_percent")) +
    geom_point(aes(color = "death_percent")) + scale_y_log10()
```



Error and Bias

There are a few different aspects to consider that may contribute to error or misleading results. The first point to consider is the spikes in the comparison charts at the beginning of the pandemic. The reason for this spike is due to the fact that Washington got the first cases in the US. We also expect alot more variablity in the comparison at the beginning of the pandemic since there were so few cases in the country. Every new case in Washington would result in a much larger change.

A source of bias to consider in this analysis is my understanding on observation of the mandates implemented in the state of Washington. As a citizen of the state of Washington, I have experienced first hand the impact the mandates have, and how they can compare with the rest of the country. However, this analysis does not use a data driven approach to understand how the mandates have an affect.

Summary

Overall, we can get a good understanding of how Washington is doing relative to the rest of the country in thier fight against the pandemic, and how Washington has been able to mitigate cases and deaths. From this analysis, we can see that Washington maybe slipping behind the rest of the country. This data can provide useful insight for the health department so that they can make more informed decisions into how to continue the fight again COVID-19.