COVID-19 Cases

Introduction

The following analysis is taking the US / Global Cases of COVID19 and using the data to draw some insights of the data. Throughout this analysis we are curious on understanding these cases throughout the last couple years to provide context and to help identify patterns within the data that could help provide legislators or health professionals the opportunity to make changes. Much of the work was done by following what Dr Wall has done for us. So in addition to the questions Dr Wall asked, how does the cases / 1000 differ between States? We will be focusing on Tennessee and New York as an interesting comparison between the states.

Importing data

First course of action is to import the data necessary for this analysis. The data in this analysis comes from the John Hopkins Github site which holds daily numbers from across the globe on COVID. Much of the data include details down to the county in later months of the pandemic.

```
link <- "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_se

files <- c("time_series_covid19_confirmed_US.csv", "time_series_covid19_confirmed_global.csv", "time_se

US_cases <- read_csv(paste0(link,files[1]))
global_cases <- read_csv(paste0(link,files[2]))
US_deaths <-  read_csv(paste0(link,files[3]))
global_deaths <- read_csv(paste0(link,files[4]))
# covid_data <- read_csv(link)

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

Data Processing

Currently there are 5 files that contain all of the information that we need in order to do the proper analysis of the data. Therefore we are going to use a number of techniques to meld the data into a version that we can analyze.

Cleaning / Joining Global Cases

Taking the global cases, we have a number of columns that we are not going to be using in the analysis.

Pro-Tip on large datasets One think to remember when working with larger datasets, the data is loaded into memory. Thus when you have 8 GB of memory and 9 gb of data, R is going to crash and will nto be able to perform the analysis. Therefore it is good to get into habit of considering it.

```
global_cases  <- global_cases %>% pivot_longer(cols = -c(`Province/State`, `Country/Region`, Lat, Long)

global_deaths  <- global_deaths %>% pivot_longer(cols = -c(`Province/State`, `Country/Region`, Lat, Long

global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = `Country/Region`, Province_State = `Province/State`) %>%
  mutate(date = mdy(date))
```

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

Cleaning / Joining US Cases

Again we are going to do the same process as above, however just for the US data.

```
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")
```

Join Global and US Cases

Here we are going to take both datasets (US_cases and global) and `unite` two files to create a dataset that we can analyze the data together.

```
global <- global %>% unite("Combined_Key",
                           c(Province_State, Country_Region),
                           sep = ", ",
                           na.rm = TRUE,
                           remove = FALSE)
```

One thing that is missing from the Global dataset is population. Therefore to add population to the global dataset we need to find the data (can see the import above) and then join the two datasets together.

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

Now that all of the information is in one dataset, we can now start asking questions of the data.

```
head(global)
```

```
## # A tibble: 6 x 7
##   Province_State Country_Region date       cases deaths Population Combined_Key
##   <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl> <chr>
## 1 <NA>           Afghanistan    2020-01-22     0      0   38928341 Afghanistan
## 2 <NA>           Afghanistan    2020-01-23     0      0   38928341 Afghanistan
## 3 <NA>           Afghanistan    2020-01-24     0      0   38928341 Afghanistan
## 4 <NA>           Afghanistan    2020-01-25     0      0   38928341 Afghanistan
## 5 <NA>           Afghanistan    2020-01-26     0      0   38928341 Afghanistan
## 6 <NA>           Afghanistan    2020-01-27     0      0   38928341 Afghanistan
```

Data Analysis

How can we measure the rate of death between US and US States?

First we need to create a new variable that will give us a rate in order for us to better understand death rates by the US as a whole and states

```
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.
```

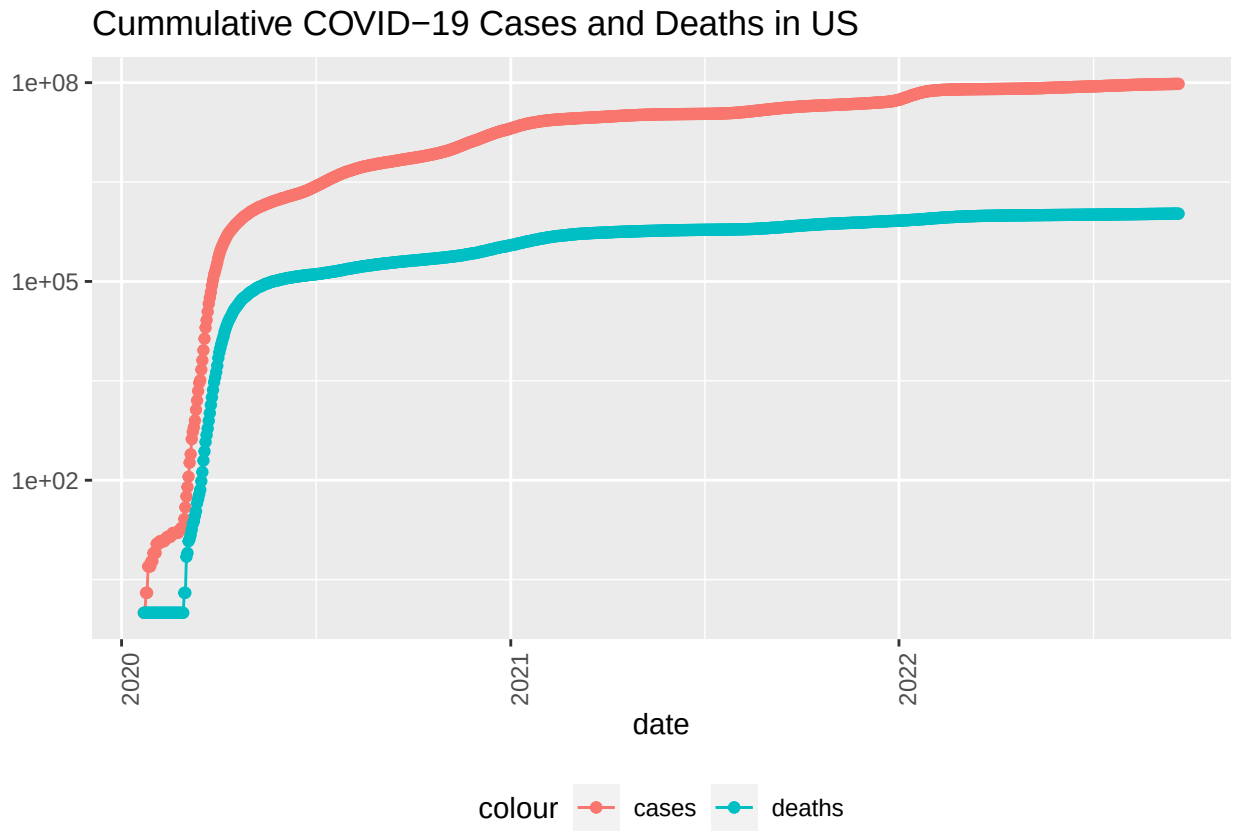
```
US_totals <- US_by_state %>%
  group_by(Country_Region, date ) %>%
  summarize(cases = sum(cases), deaths = sum(deaths),
            Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
  select(Country_Region, date, cases, deaths, deaths_per_mill, Population)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

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

Visualizing COVID cases in the US

First we wanted to see how the cumulative cases stack up to the deaths of COVID in the US. This gives us a LOG graph of the data rendering the information to create an almost straight line due to the compressing nature of a log graph.

```
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 = "Cummulative COVID-19 Cases and Deaths in US", y = NULL)
```



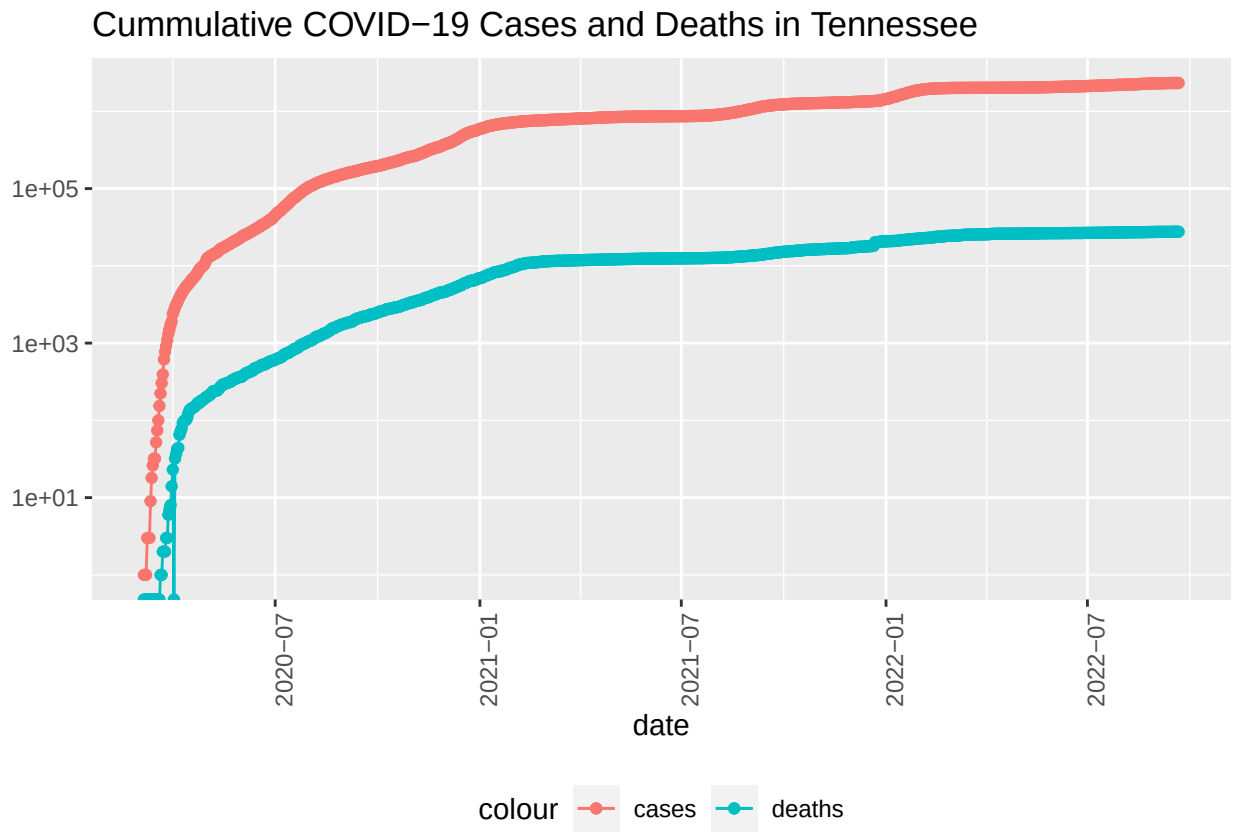
Visualizing COVID cases on a particular state

Now we create another graph of the same analysis as we had on the US data, but this time we will look at the state of Tennessee. Again we see that past 2022, the data is compressed that it gives us a flat line and is would be probably difficult to draw any insight from the data.

```
state <- "Tennessee"
US_by_state %>%
  filter(Province_State == state) %>%
  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=str_c("Cummulative COVID-19 Cases and Deaths in ", state), y = NULL)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```



Above we observed the cumulative cases day to day, as the magnitude grew, as we see in the above graphs it what does it look like when we only observe the new cases day to day. We need to add a few more details to our dataset in order to further analyze.

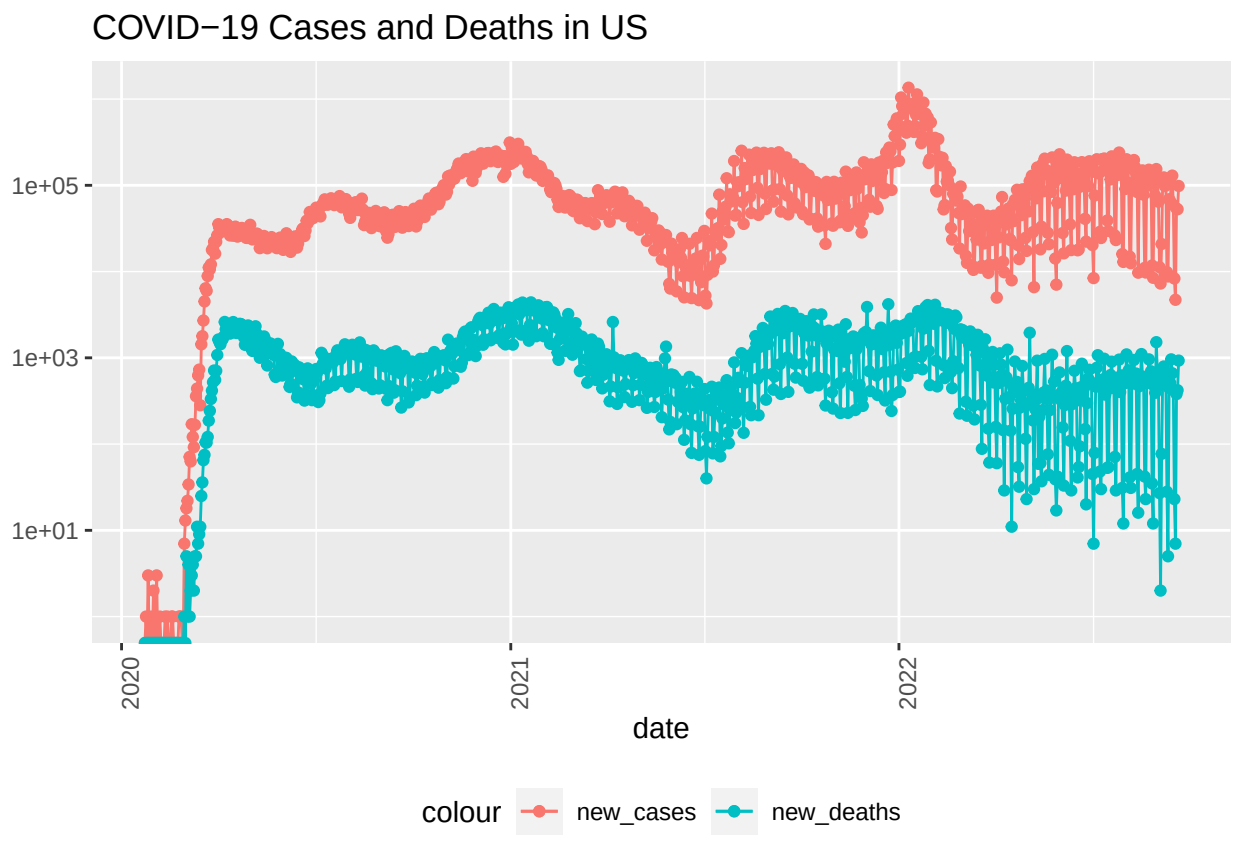
```
US_by_state <- US_by_state %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))

US_totals <- US_totals %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths))
```

We have added two columns `new_cases` and `new_deaths` to give us a better understanding of the new cases by day.

```
tail(US_totals %>% select(new_cases, new_deaths, everything()))
```

```
## # A tibble: 6 x 8
##   new_cases new_deaths Country_Region date          cases deaths deaths_per_mill
##       <dbl>      <dbl> <chr>          <date>        <dbl>  <dbl>           <dbl>
## 1     59018        423 US             2022-09-16 95656684 1.05e6           3165.
## 2      8314         23 US             2022-09-17 95664998 1.05e6           3165.
## 3      4710          7 US             2022-09-18 95669708 1.05e6           3165.
## 4     53789        380 US             2022-09-19 95723497 1.05e6           3166.
## 5     52812        421 US             2022-09-20 95776309 1.05e6           3167.
## 6     98282        929 US             2022-09-21 95874591 1.06e6           3170.
## # ... with 1 more variable: Population <dbl>
```

```
US_totals %>%
  ggplot(aes(x=date, y=new_cases)) +
  geom_line(aes(color="new_cases")) +
  geom_point(aes(color="new_cases")) +
  geom_line(aes(y=new_deaths, color="new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_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)
```

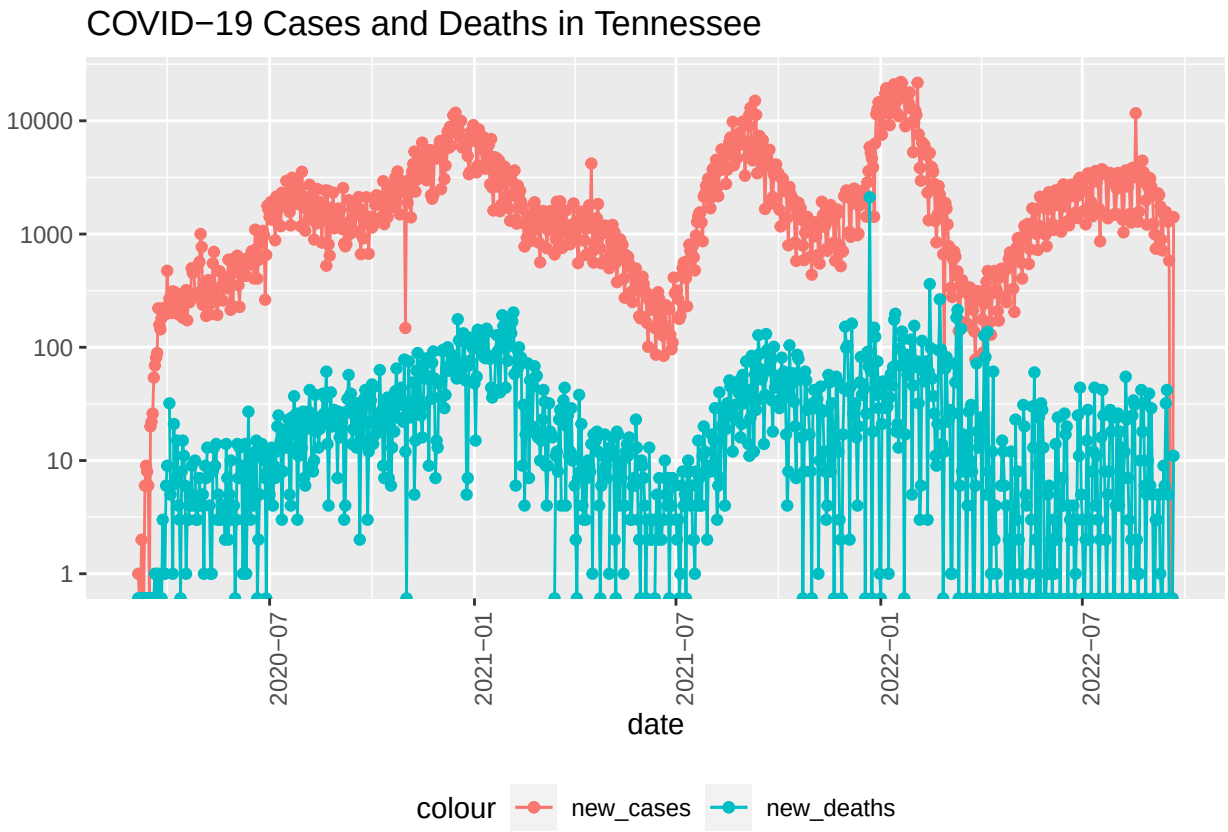


This gives us a little more granularity of the data. We see a huge upswing in new cases occur in Feb 2022 as those were due to the spreading of COVID-19 strain of Delta. We also see that it was the first time where

the death trended differently from the cases of COVID. Of course there are many views that could explain this.

- Mortality rate of COVID-19 variant not as "successful"
- Herd immunity according to some news platforms
- Inconsistent data collection between states and centers (though this is going to have a margin of error)

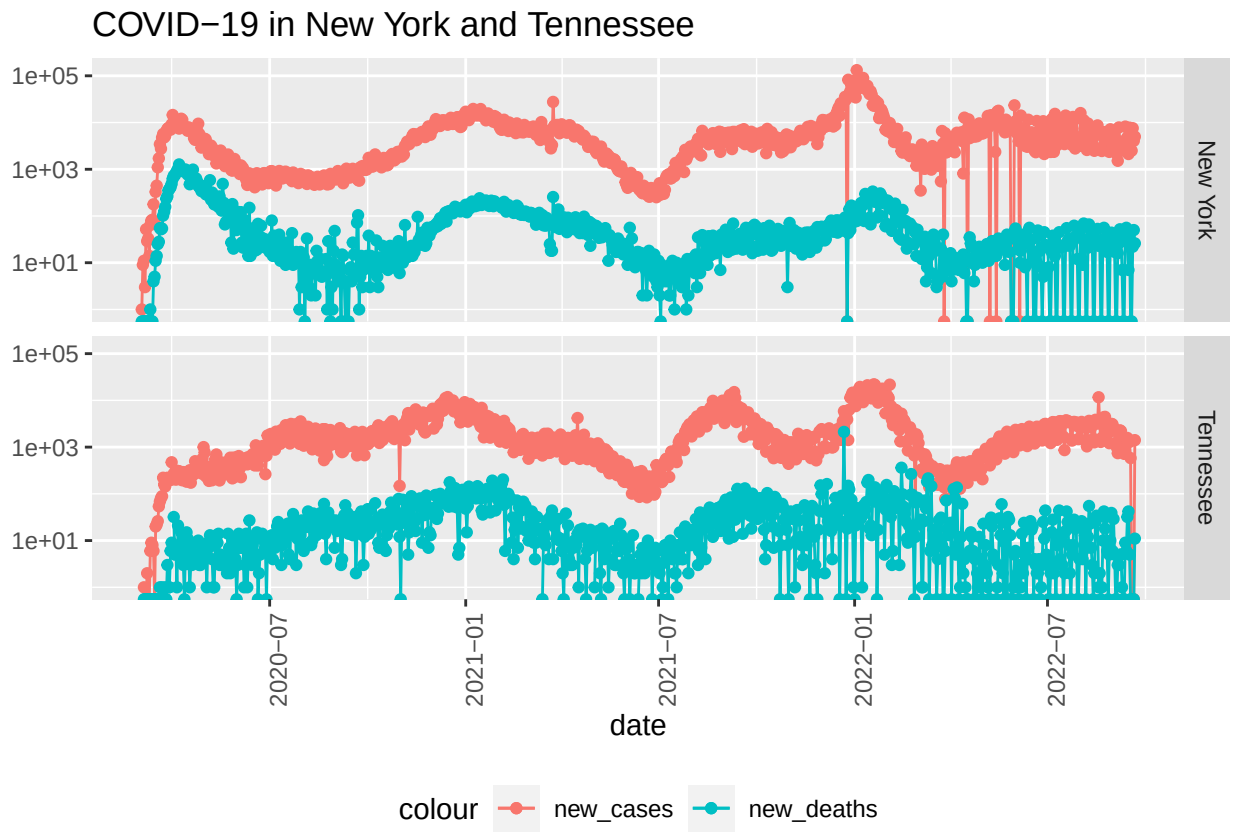
```
state <- "Tennessee"
US_by_state %>%
  filter(cases > 0) %>%
filter(Province_State == state) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  scale_y_log10() +
  theme(legend.position  = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title=str_c("COVID-19 Cases and Deaths in ", state), y = NULL)
```



The following is an interesting trend that keeps oscillating due to the "less strict" policies in TN. However there isn't much to be compared to here. Let's take NY and TN for example as a case and point.

```
state <- c("Tennessee","New York")
US_by_state %>%
```

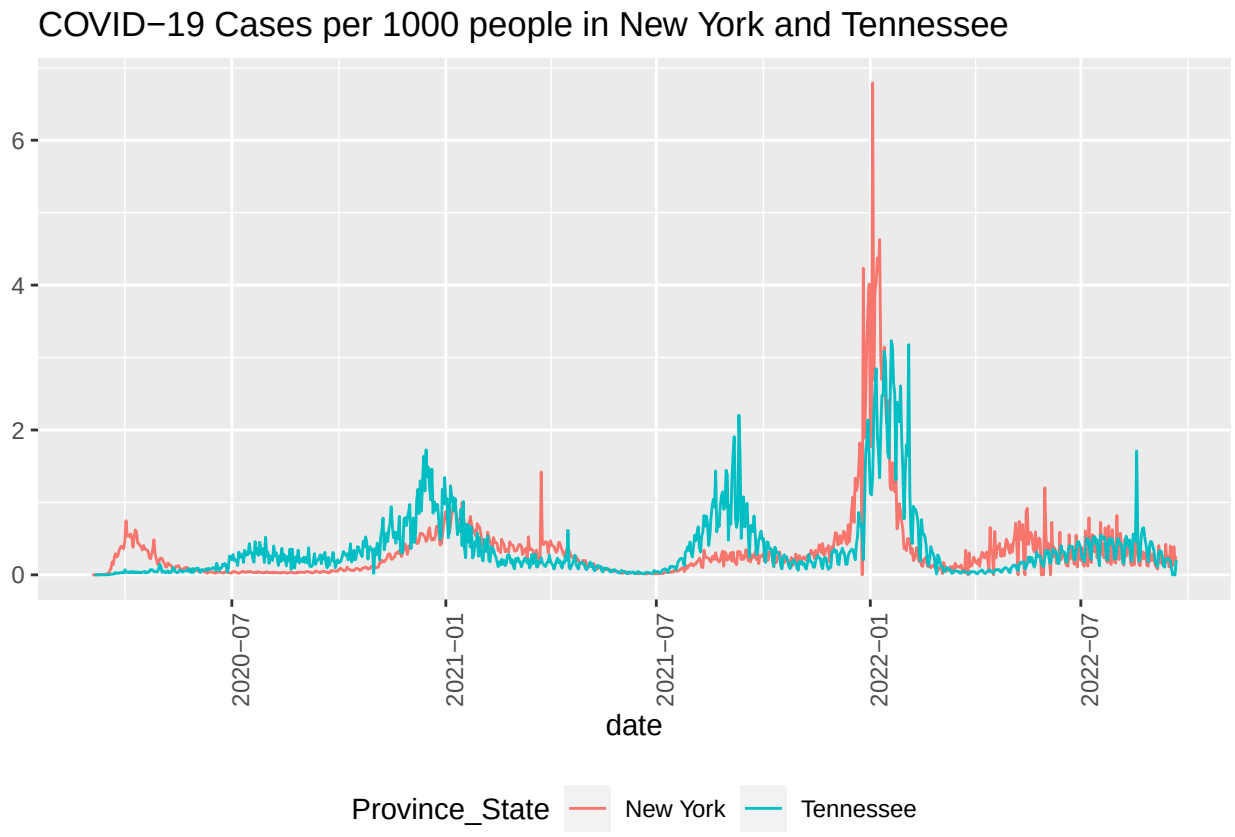
```
  filter(cases > 0) %>%
filter(Province_State %in% state) %>%
  ggplot(aes(x = date, y = new_cases)) +
  geom_line(aes(color = "new_cases")) +
  geom_point(aes(color = "new_cases")) +
  geom_line(aes(y = new_deaths, color = "new_deaths")) +
  geom_point(aes(y = new_deaths, color = "new_deaths")) +
  facet_grid("Province_State~.") +
  scale_y_log10() +
  theme(legend.position  = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title=str_c("COVID-19 in ", state[2], " and ", state[1]), y = NULL)
```



There is definitely some usefulness when observing the data like so... We see that the September 2021 upswing in cases was higher... but really to make the data comparable we need to look at the information from a per 1000 incidence in order to understand what might be going on with the data.

```
state <- c("Tennessee","New York")
US_by_state %>%
   filter(cases > 0 & Province_State %in% state) %>%
  mutate( cases_per_thou = 1000*new_cases/Population,
          deaths_per_thou = 1000*new_deaths/Population) %>%
  ggplot(aes(x = date, y = cases_per_thou, color = Province_State)) +
  geom_line(aes()) +
  theme(legend.position  = "bottom",
```

```
      axis.text.x = element_text(angle = 90)) +
 labs(title=str_c("COVID-19 Cases per 1000 people in ", state[2], " and ", state[1]), y = NULL)
```



This is actually a fascinating story to look at, in this graph. We see that New York started out the pandemic with a much high rate than TN due to it being ground 0 for COVID and a central hub with a much higher contact factor (essentially the number of people one might have contact with in a day).

But the cycle of the first wave of COVID started subsiding due to the lockdowns within the first 5 months of the year.

We see a rise start in August (when school begins in TN) and we see that it seems to hover pretty consistently till the holidays and has a spike in Nov / Dec and drops off. Look at the rates at which those drop offs occur... we see TN have a significant spike and drop off, but in NY we see that the drop off is extended. This is most likely due to the masking mandates, homeschool and lockdowns. This provides some insight into how for health care workers the hospital resources would not be overwhelmed but rather find themselves able to manage and provide better care for the sick.

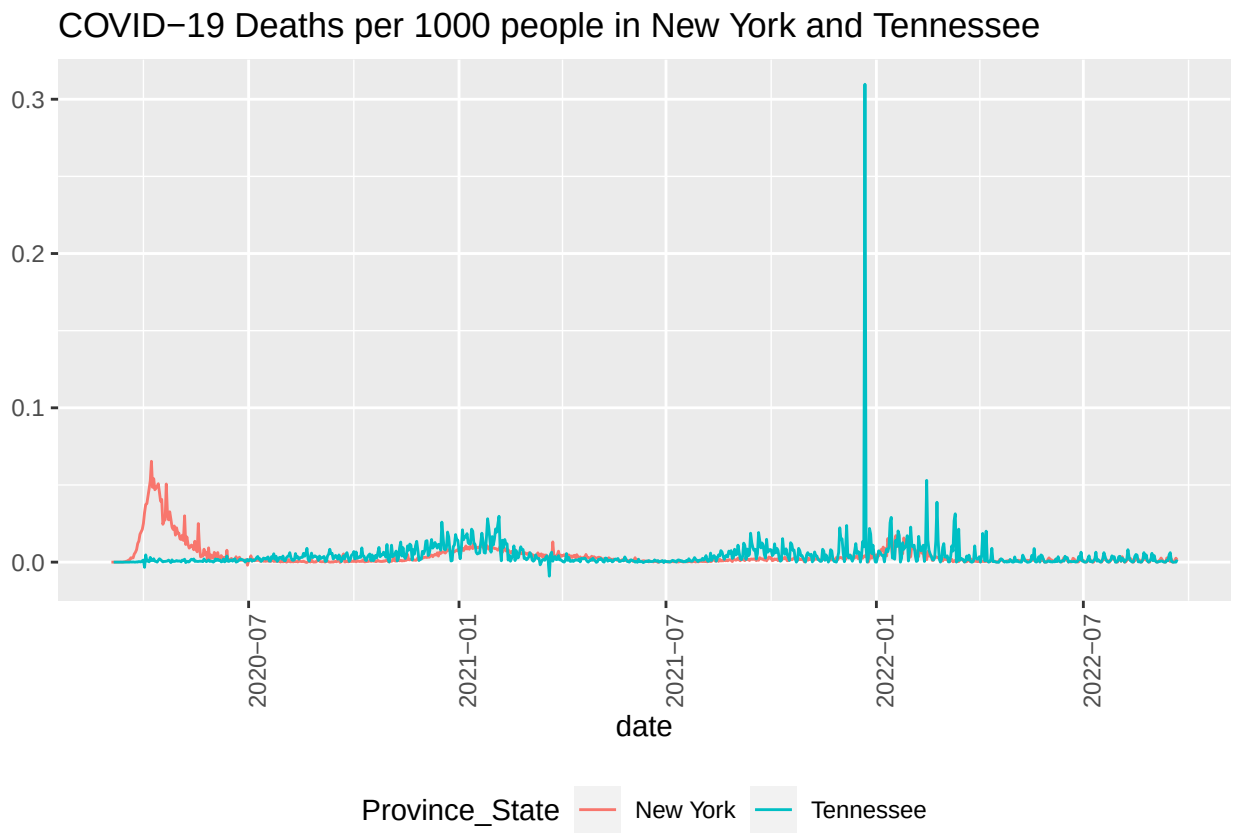
Up until the school year of 2021, we have an identical spike in TN as we had previously with the same tapering off of cases... but almost no increase until the holidays of 2021 where we see a much larger spike in NY then we had seen in TN.

There are a number of reasons for that... the delta started started becoming the predominant strain in Nov 2021 and it was known to be more contagious. So we see a much high case load but it wouldn't surprise me in the least that there was some COVID fatigue going on as well and people were more willing to take the risk as it wasn't an unknown disease at this point anymore.

This analysis is definitely a viewpoint that would need much more verification.

What about the rate of deaths between TN and NY?

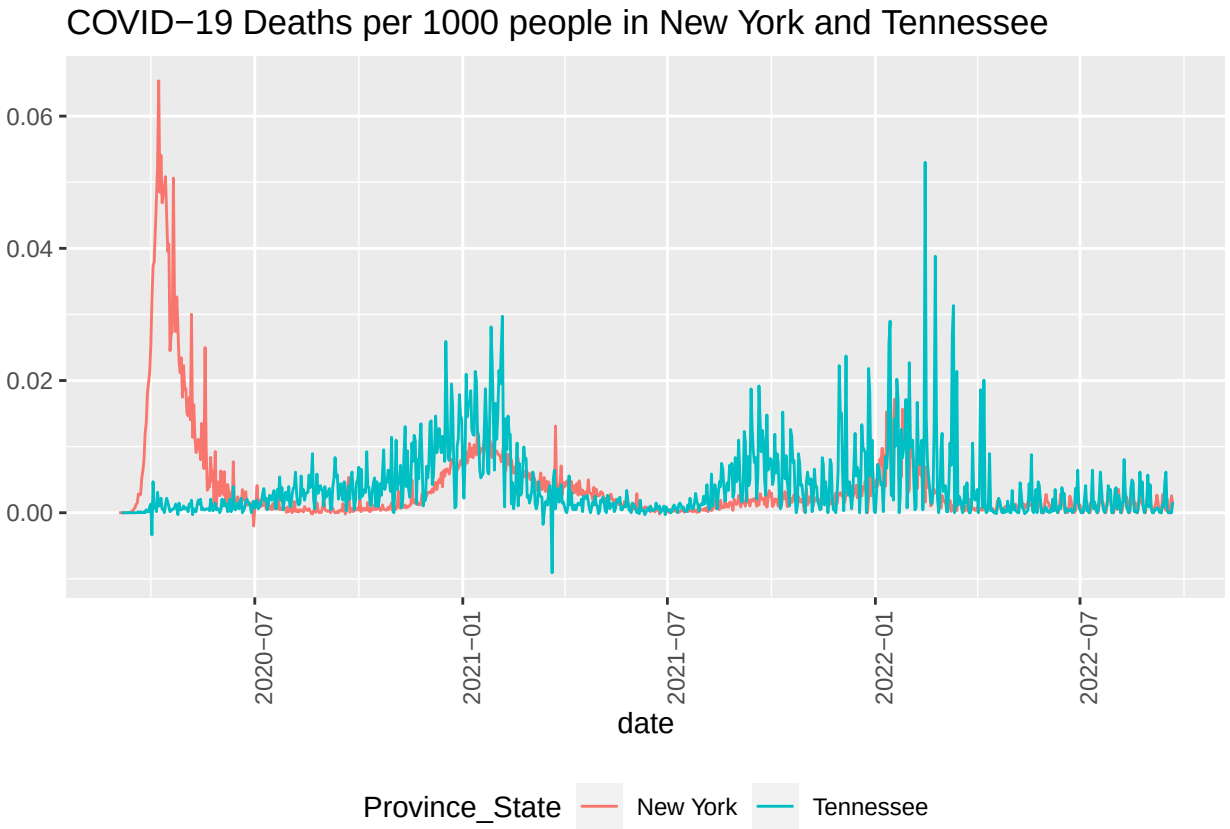
```
state <- c("Tennessee","New York")
US_by_state %>%
   filter(cases > 0 & Province_State %in% state) %>%
  mutate( cases_per_thou = 1000*new_cases/Population,
          deaths_per_thou = 1000*new_deaths/Population) %>%
  ggplot(aes(x = date, y = deaths_per_thou, color = Province_State)) +
  geom_line(aes()) +
  theme(legend.position  = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title=str_c("COVID-19 Deaths per 1000 people in ", state[2], " and ", state[1]), y = NULL)
```



Interesting enough, we see the .3 see the rate at the end of 2021 skyrocket making it difficult to simply visulaize the data. This could have been a correction where data hadn't been kept in earlier days due to holiday backlog. So just for the sake of clarity, lets see what it looks like if we were to take the outlier out.

```
state <- c("Tennessee","New York")
US_by_state %>%
   filter(cases > 0 & Province_State %in% state) %>%
  mutate( cases_per_thou = 1000*new_cases/Population,
          deaths_per_thou = 1000*new_deaths/Population) %>%
  filter(deaths_per_thou < 0.1) %>%
  ggplot(aes(x = date, y = deaths_per_thou, color = Province_State)) +
  geom_line(aes()) +
```

```
  theme(legend.position  = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title=str_c("COVID-19 Deaths per 1000 people in ", state[2], " and ", state[1]), y = NULL)
```



Outside of the first spike in NY, we see that TN though alot of spikes, we could smoothen out the data and we'd find that the rate is for the majority of the last 2.5 years are higher in TN then they are in NY. The scale here might seem small.. but even at 0.04 per 1000 is 4 people per 100,000 or 40 people per million can be quite telling as that rate never even came close in NY.

How do we compare data between countries or states?

```
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)
```

After summarizing the data by state, we can start comparing the worst day throughout the entire 2.5 years that COVID has been around. We see that for the 10 lowest states:

```
US_state_totals %>% slice_min(deaths_per_thou,n = 10)%>% select(deaths_per_thou, cases_per_thou, everyth
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State        deaths  cases population
##              <dbl>          <dbl> <chr>                 <dbl>  <dbl>      <dbl>
##  1           0.611           148. American Samoa           34 8.22e3      55641
##  2           0.725           239. Northern Mariana Isl~    40 1.32e4      55144
##  3           1.15            215. Virgin Islands          123 2.31e4     107268
##  4           1.15            227. Vermont                 719 1.42e5     623989
##  5           1.19            242. Hawaii                 1679 3.43e5    1415872
##  6           1.35            252. Puerto Rico            5085 9.48e5    3754939
##  7           1.56            323. Utah                   5001 1.03e6    3205958
##  8           1.84            402. Alaska                 1364 2.98e5     740995
##  9           1.87            237. Washington            14245 1.81e6    7614893
## 10           1.91            214. Maine                  2569 2.87e5    1344212
```

And then the 10 highest states with a bunch of those states being in the in the southern states. This can be due to again a few different things.

- legislative policies - the north in general was much stricter in masking mandates and lockdown initiatives.
- health - there could be a case made that the southern states is more obese than some of the other states
- Population demographics - Age of the population could also cause this data to be higher in the southern states than in the northern states.

All these reasons could be due to a number of things but without external data and confirmation it would be difficult to know for certain if there is any validity in reasonings given.

```
US_state_totals %>% slice_max(deaths_per_thou,n = 10)%>% select(deaths_per_thou, cases_per_thou, everyth
```

```
## # A tibble: 10 x 6
##    deaths_per_thou cases_per_thou Province_State deaths   cases population
##              <dbl>          <dbl> <chr>           <dbl>   <dbl>      <dbl>
##  1            4.33           311. Mississippi     12879  925596    2976149
##  2            4.30           312. Arizona         31326 2268158    7278717
##  3            4.25           301. Oklahoma        16798 1192446    3956971
##  4            4.14           308. Alabama         20322 1512134    4903185
##  5            4.12           333. West Virginia    7379  597303    1792147
##  6            4.07           294. New Mexico       8534  616648    2096829
##  7            4.05           341. Tennessee       27642 2329632    6829174
##  8            3.99           314. Arkansas        12028  947967    3017804
##  9            3.91           306. New Jersey      34694 2719358    8882190
## 10            3.87           312. Louisiana       17993 1451304    4648794
```

Modelling the data

Modelling data never gives us exact information, however it gives us an idea of what possibly might happen. For this part of class, we created a linear model.

```
mod <- lm(deaths_per_thou ~ cases_per_thou, data = US_state_totals)
summary(mod)
```

```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = US_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3796 -0.6000  0.1199  0.6667  1.1463
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.273447   0.702941  -0.389    0.699
## cases_per_thou  0.011191   0.002425   4.616 2.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8407 on 54 degrees of freedom
## Multiple R-squared:  0.2829, Adjusted R-squared:  0.2696
## F-statistic:  21.3 on 1 and 54 DF,  p-value: 2.459e-05
```

Here are are simply drawing a linear model $y = mx + b$.. we can see that there is an intercept of -0.273 and a slope of 0.011 which says that for every for every additional 1000 cases, we see the deaths_per_thou increase 0.011 or 11 people based on on the model. But lets see what that might look like in action.

Do a little data transformation to add the predictions to the dataset.

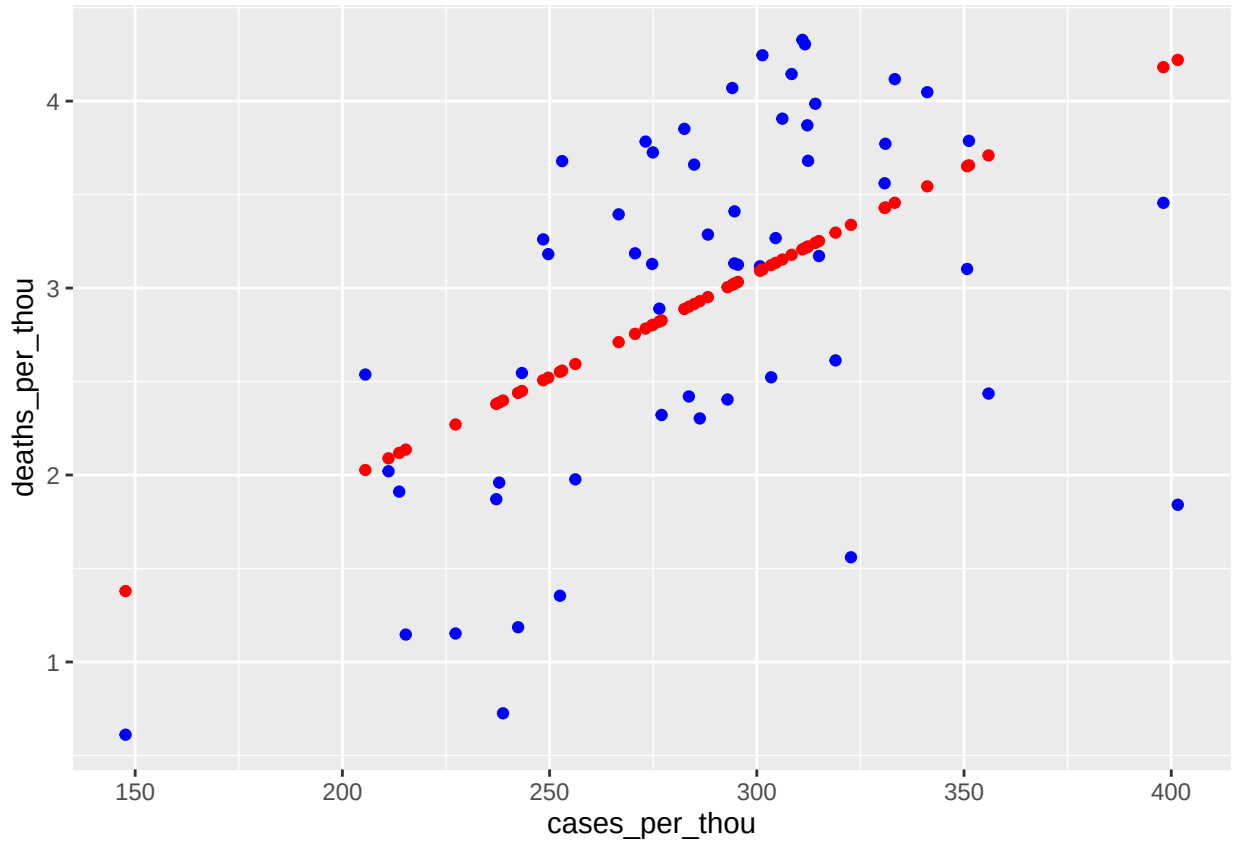
```
x_grid <- seq(1,151)
new_df <- tibble(cases_per_thou = x_grid)
US_state_totals %>% mutate(pred = predict(mod))
```

```
## # A tibble: 56 x 7
##    Province_State  deaths  cases population cases_per_thou deaths_per_thou  pred
##    <chr>            <dbl>  <dbl>      <dbl>          <dbl>           <dbl> <dbl>
##  1 Alabama          20322 1.51e6    4903185           308.            4.14   3.18
##  2 Alaska            1364 2.98e5     740995           402.            1.84   4.22
##  3 American Samoa      34 8.22e3      55641           148.            0.611  1.38
##  4 Arizona          31326 2.27e6    7278717           312.            4.30   3.21
##  5 Arkansas         12028 9.48e5    3017804           314.            3.99   3.24
##  6 California       95632 1.12e7   39512223           284.            2.42   2.90
##  7 Colorado         13260 1.65e6    5758736           286.            2.30   2.93
##  8 Connecticut      11343 8.90e5    3565287           250.            3.18   2.52
##  9 Delaware          3088 3.07e5     973764           315.            3.17   3.25
## 10 District of Co~   1383 1.68e5     705749           238.            1.96   2.39
## # ... with 46 more rows
```

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

And graph them...

```
US_tot_w_pred %>% ggplot() +
  geom_point(aes(x = cases_per_thou, y = deaths_per_thou), color = "blue") +
  geom_point(aes(x = cases_per_thou, y = pred), color = "red")
```



This model, when you take into consideration has an r^2 of 0.84 which makes it to be a poor predictor of what is actually going on. With additional information either in rows of data or in additional types of data we probably would be able to make a more accurate model of the data. One can also use a polynomial model that might provide a little more accuracy to the data.

Discussion on Bias

There is nothing as wonderful as being able to have so much data in one sheet. The upstream work that goes into providing this data is a rather big undertaking that we really don't comprehend. However there are still many ways bias' can get into the data.

Data accuracy based on technology

There are going to be varying levels of accuracy of the data between countries and states across the world. Technology is a wonderful thing, it helps us coordinate to have this information but there are still many places even in the United States that are hand counting this information and using paper documentation and not EMR systems.

Data accuracy based on nefarious reporting

There are also rumours that COVID cases might be high due to the payout a hospital receives from an insurance company. It would not surprise me that there will be a hand full of cases across the countless of hospitals and emergency care facilities that serve 330 million people. (just to rant... insurance companies will quickly get onto figure this out... they aren't in the market for losing money).

Early symptoms were confusing

In the early days, it could be said that PCR primer production has to be ramped up for all hospitals to be able to do PCR tests for COVID-19.

Discussion / Conclusion

To conclude this analysis, we find that looking at the rate of cases / 1000 can tell alot of interesting stories about the data that we are looking at. We looked at NY and TN to give us some insight and an avenue to explore possible reasons for these differences. This could provide a rather lengthy project to further analyze the differences in the data and coming up with asystematic analysis to evaluate government decisions made at a local level to better understand their effectiveness.