Analysis of Covid Deaths, Cases, and Administered Vaccine Doses in the USA

2022-10-07

Covid Cases, Deaths, and total Vaccine Doses within the United States

For the purpose of this analysis current date will be considered October 7th, 2022– the date at which this analysis was completed and finalized.

This analysis was performed using Covid-19 data recorded by John Hopkins University. The data collection began in 2019 and it is still tracking current data which is updated on the GitHub page from which it was obtained. The US death and Covid-19 case data was attained from: John Hopkins Time series summary (csse_covid_19_time_series).
Vaccine dose administration data in this analysis was obtained from: John Hopkins U.S. States vaccine data. The Vaccination data tracks from late 2020 and is still being updated to track current vaccine dosing data through its GitHub page.

First, data from the US Cases and Deaths CSVs was obtained and tidied. The data was oriented so that the Data Frame rows followed the date and unnecessary informational columns were not retained.

The death and case data was then verified for possible typos or issues by examining a summary of its statistics and determining if any values may have been typos or incorrectly recorded.

In the data summary it was noticed that the maximum values for cases and deaths seemed large and could have possibly been typo errors, so data was filtered by size to see if there were similar large values. Since this data frame contains the total cases and deaths over time, it would follow that there should be similar values recorded before or after the date that the value in question was recorded for. It appeared that the values were not outlying typos or errors.

```
summary(US_data)
```

```
##     Admin2          Province_State     Country_Region          date           
##  Length:2941358     Length:2941358     Length:2941358     Min.   :2020-01-22  
##  Class :character   Class :character   Class :character   1st Qu.:2020-11-19  
##  Mode  :character   Mode  :character   Mode  :character   Median :2021-07-07  
##                                                           Mean   :2021-07-06  
##                                                           3rd Qu.:2022-02-22  
##                                                           Max.   :2022-10-09  
##      cases              Population           deaths        
##  Min.   :      1   Min.   :      86   Min.   :    0.0  
##  1st Qu.:    539   1st Qu.:   11758   1st Qu.:    8.0  
##  Median :   2334   Median :   26868   Median :   40.0  
##  Mean   :  13017   Mean   :  106395   Mean   :  182.7  
##  3rd Qu.:   7694   3rd Qu.:   69922   3rd Qu.:  120.0  
##  Max.   :3464157   Max.   :10039107   Max.   :33740.0
```

```
Uscheck<-US_data %>% filter(cases>3400000) %>% filter(deaths> 33000)
Uscheck
```

```
## # A tibble: 42 x 7
##    Admin2      Province_State Country_Region date        cases Population deaths
##    <chr>       <chr>          <chr>          <date>      <int>      <int>  <int>
##  1 Los Angeles California     US             2022-08-29 3.40e6   10039107  33124
##  2 Los Angeles California     US             2022-08-30 3.41e6   10039107  33138
##  3 Los Angeles California     US             2022-08-31 3.41e6   10039107  33155
##  4 Los Angeles California     US             2022-09-01 3.41e6   10039107  33171
##  5 Los Angeles California     US             2022-09-02 3.41e6   10039107  33187
##  6 Los Angeles California     US             2022-09-03 3.41e6   10039107  33187
##  7 Los Angeles California     US             2022-09-04 3.41e6   10039107  33187
##  8 Los Angeles California     US             2022-09-05 3.42e6   10039107  33209
##  9 Los Angeles California     US             2022-09-06 3.42e6   10039107  33217
## 10 Los Angeles California     US             2022-09-07 3.42e6   10039107  33227
## # ... with 32 more rows
```

An additional data frame was prepared in order to allow for observing trends in regions of the United states (The West, Southwest, North East, MidWest, and South East). Also, locations not directly located in the 50 States of the United States were removed from the data set.

Regions were categorized as:

West: Washington, Oregon, Idaho, Montana, Wyoming, Colorado, Utah, California, Nevada, Alaska, Hawaii

South West: Arizona, New Mexico, Texas, Oklahoma

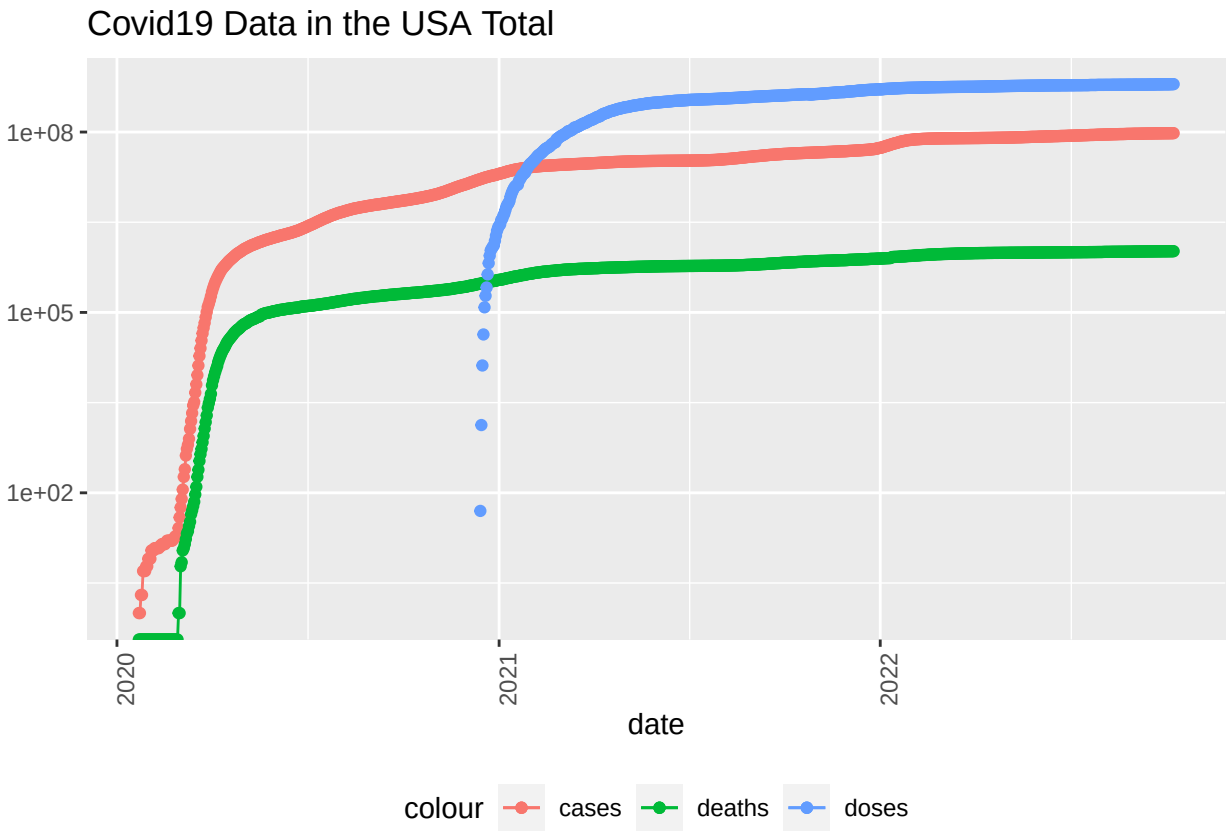
North East: Pennsylvania, Massachusetts, New York, New Jersey, Maine, Vermont, Rhode Island, Connecticut, New Hampshire

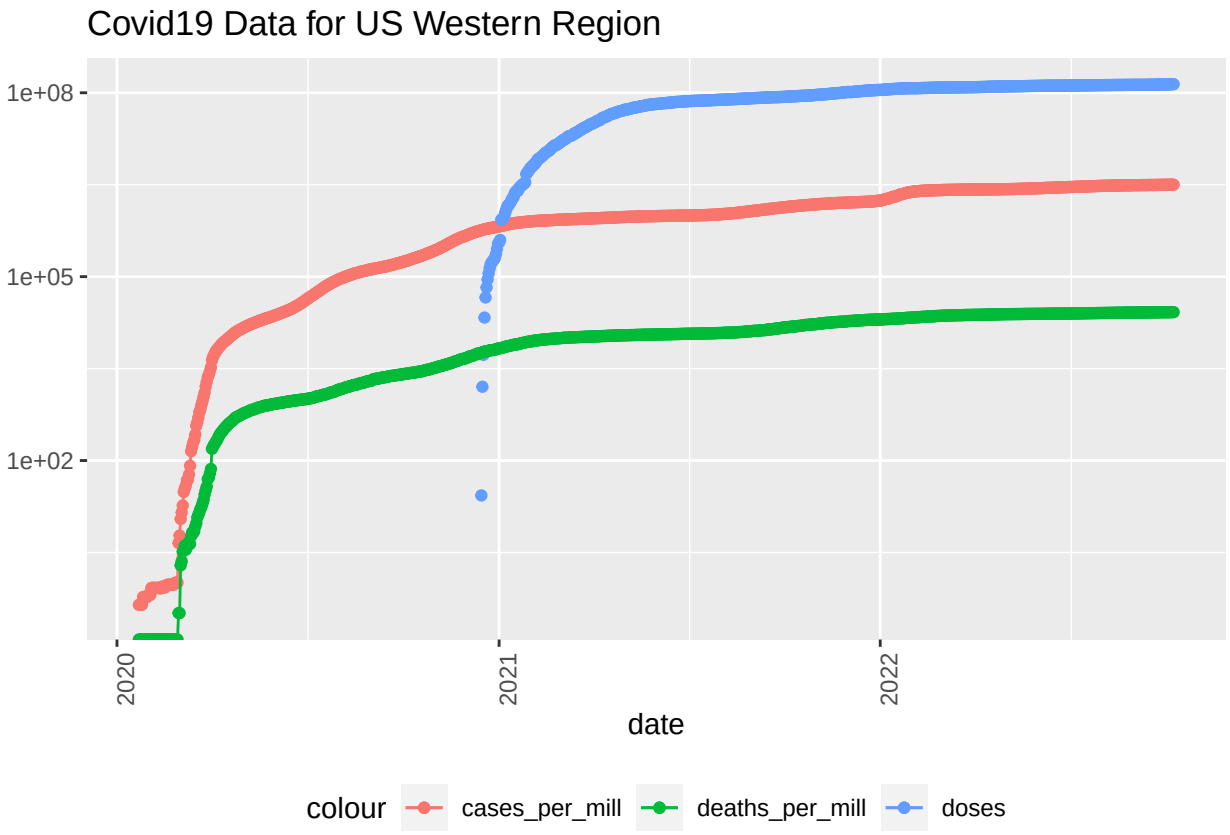
Midwest: North Dakota, Minnesota, South Dakota, Iowa, Nebraska, Kansas, Missouri, Wisconsin, Illinois, Indiana, Michigan, Ohio

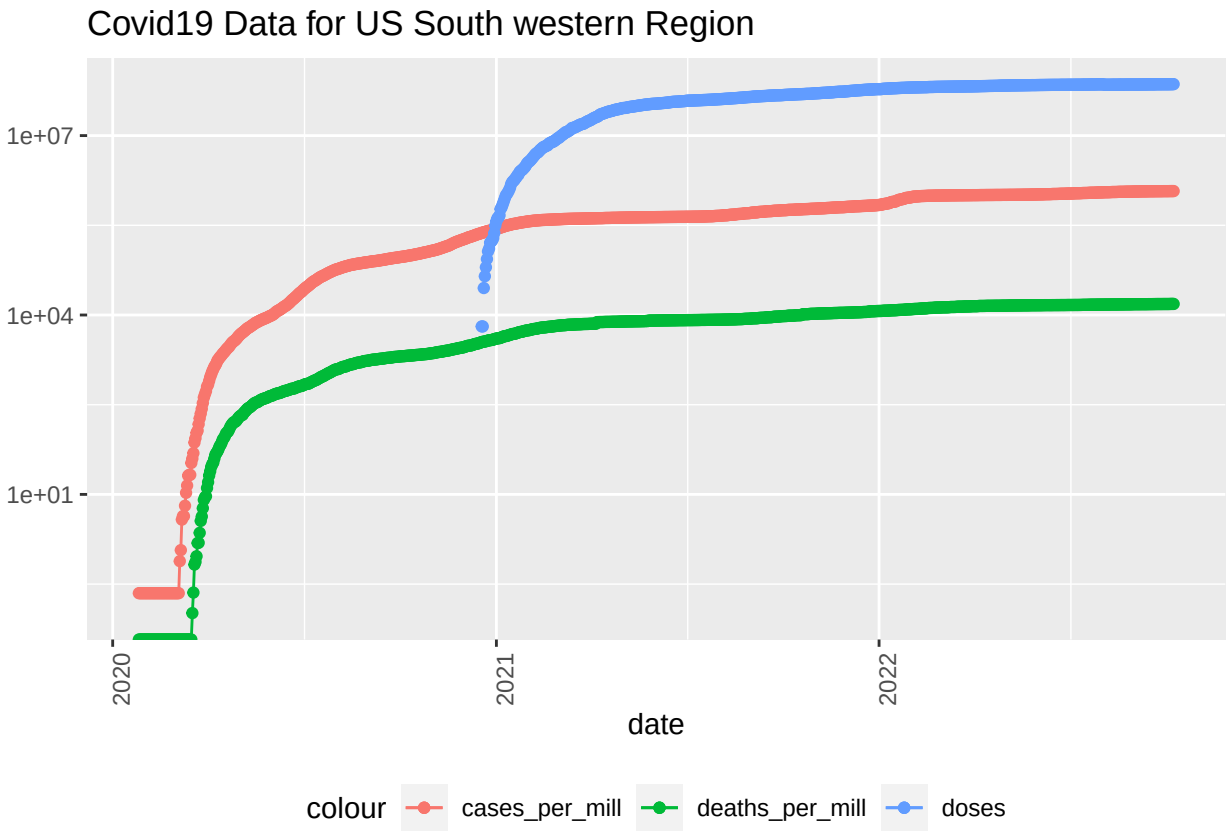
South East: West Virginia, Delaware, Virginia, Kentucky, Tennessee, Maryland, North Carolina, South Carolina, Georgia, Florida, Alabama, Mississippi, Louisiana, Arkansas

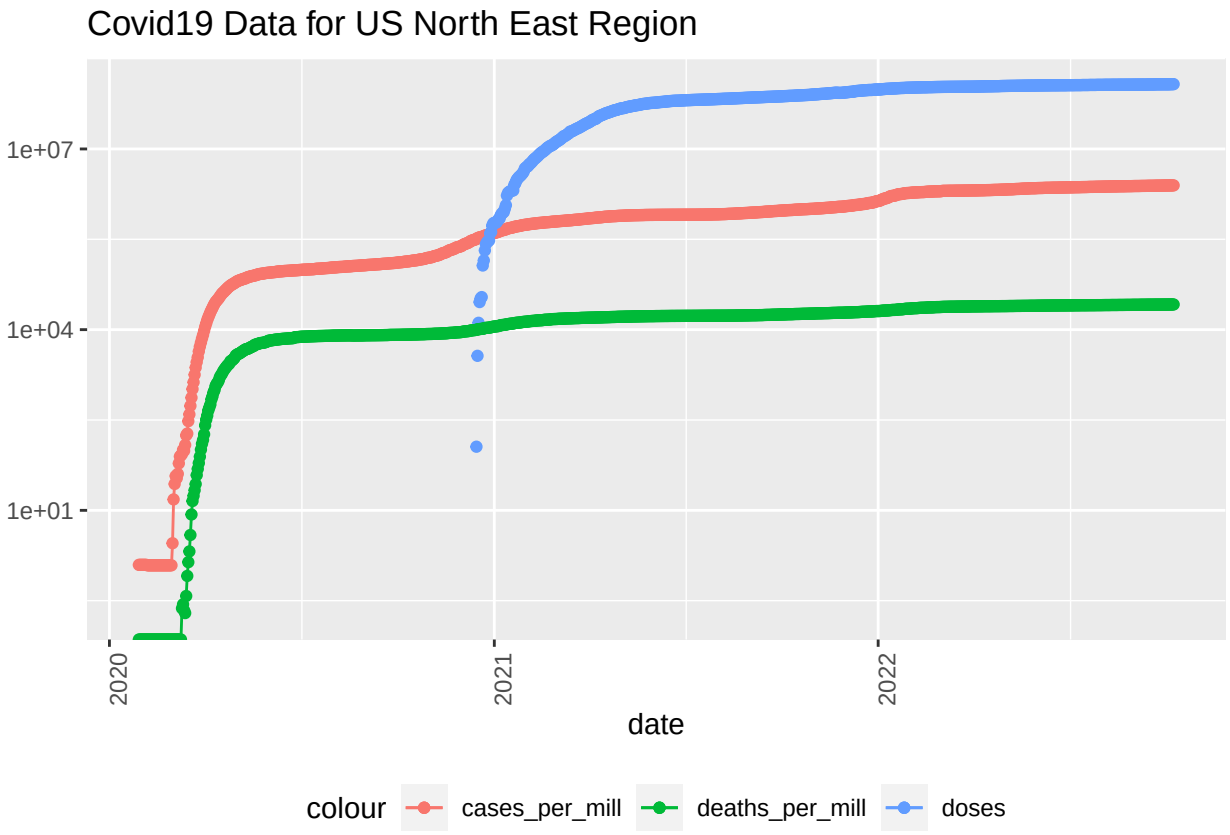
Next, columns were also added to both the region dependent and total USA data sets to include “Deaths per Million” and “Cases per Million” metrics, to create data that allows comparisons for areas with different population totals. The vaccination data was also joined with both of these data frames. The Vaccination data specifically describes the number of doses administered per date in different locations of the United States. It does not included information for how many individuals were fully vaccinated nor vaccine brand.

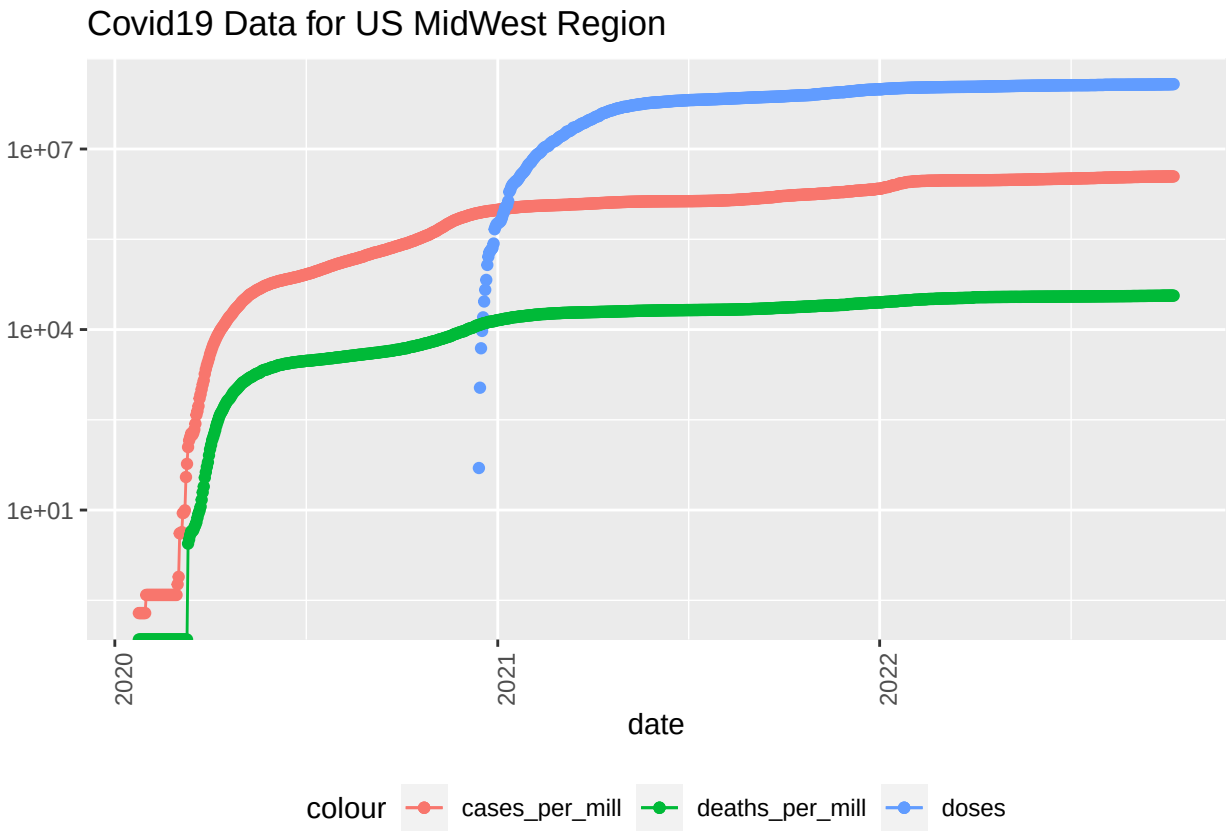
#Data Visualizations for U.S. Covid Cases, Deaths, and total Vaccine Doses

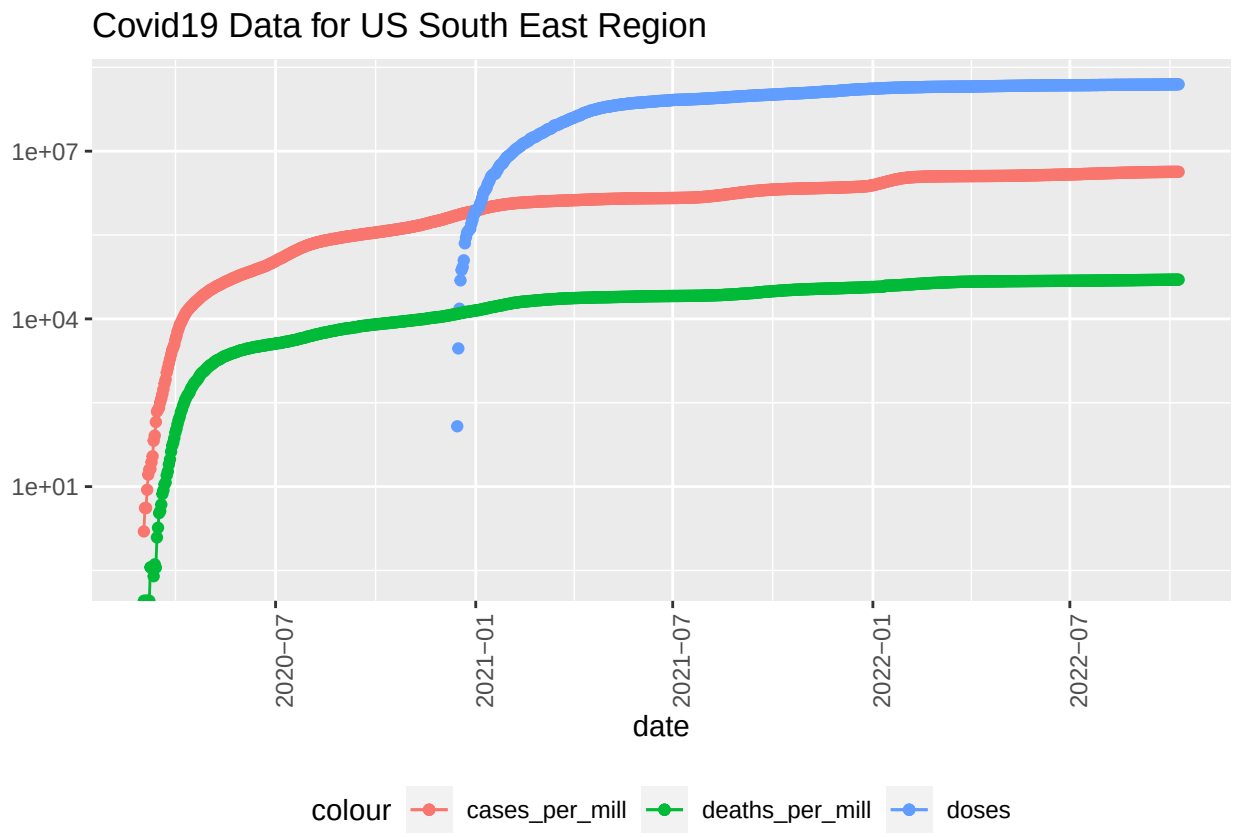

Covid19 Data in the USA Total
1e+08
1e+05
1e+02
2020
2021
2022
date
colour
cases
deaths
doses


Covid19 Data for US Western Region
1e+08
1e+05
1e+02
2020
2021
2022
date
colour
cases_per_mill
deaths_per_mill
doses


Covid19 Data for US South western Region
1e+07
1e+04
1e+01
2020
2021
2022
date
colour
cases_per_mill
deaths_per_mill
doses


Covid19 Data for US North East Region
1e+07
1e+04
1e+01
2020
2021
2022
date
colour
cases_per_mill
deaths_per_mill
doses


Covid19 Data for US MidWest Region
1e+07
1e+04
1e+01
2020
2021
2022
date
colour
cases_per_mill
deaths_per_mill
doses



As can be seen, the rate of infection can be quantified by the number of cases. In the above plots, the cases deaths and number of administered doses are shown. The plots were generated for the U.S. in total as well as each of the five earlier specified regions of the United States.

In the total United States plot, it is seen that in early 2020 cases quickly increased. This sharp increase in cases was closely followed by a sharp increase in deaths. Later, in mid-2020, the total cases and total deaths begin to level off and the rate at which the case and death totals increase is slows down. In late 2021, we see the first recorded data for vaccine doses that were administered. In conjunction with vaccine, doses the cases and deaths continue to slow their rate of growth. However, from this visual alone it cannot be concluded whether the vaccine doses contributed to the slow down in these variables. It can also be seen that the increase in deaths per million and cases per million began to slow down before the administration of vaccine doses.

It can also be seen that the number of total vaccine doses administered quickly increased from late 2020 to mid 2021. Later the number of doses tapers off and very gradually increases.

These trends are mirrored for each of the five regions in the United States. There were no particular visual anomalies or major differences noticed in their plots that indicate significant variations in the way that cases, deaths, and vaccine doses trended over the same time periods in the five different regions.

United States Regional Summary for Cases and Deaths per Million

The current total deaths, cases, and ratio of deaths to cases were generated for each of the fifty states and for the United states overall.

```
## # A tibble: 5 x 4
##   region    Total_Deaths_Per_Million Total_Cases_Per_Million Deaths_per_Cases
##   <chr>                        <dbl>                   <dbl>            <dbl>
## 1 Midwest                     36633.                3498902.          0.0105
## 2 Northeast                   26201.                2483233.          0.0106
## 3 Southeast                   50286.                4267135.          0.0118
## 4 Southwest                   15313.                1181247.          0.0130
## 5 West                        26418.                3164838.          0.00835
```

United States by-State Summary for Cases and Deaths per Million

```
## # A tibble: 50 x 4
##    Province_State Total_Deaths_Per_Million Total_Cases_Per_Mil~ Deaths_per_Cases
##    <chr>                             <dbl>                <dbl>            <dbl>
##  1 Alabama                           4175.              311170.          0.0134
##  2 Alaska                            1911.              407579.          0.00469
##  3 Arizona                           4315.              312587.          0.0138
##  4 Arkansas                          4076.              305756.          0.0133
##  5 California                        2435.              285796.          0.00852
##  6 Colorado                          2315.              288041.          0.00804
##  7 Connecticut                       3191.              252198.          0.0127
##  8 Delaware                          3196.              316945.          0.0101
##  9 Florida                           3726.              333812.          0.0112
## 10 Georgia                           3654.              263901.          0.0138
## 11 Hawaii                            1188.              246202.          0.00483
## 12 Idaho                             2903.              278114.          0.0104
## 13 Illinois                          2768.              297792.          0.00929
## 14 Indiana                           3656.              286034.          0.0128
## 15 Iowa                              3200.              271874.          0.0118
## 16 Kansas                            3286.              302279.          0.0109
## 17 Kentucky                          3812.              354393.          0.0108
## 18 Louisiana                         3894.              313242.          0.0124
## 19 Maine                             1942.              217210.          0.00894
## 20 Maryland                          2547.              207531.          0.0123
## 21 Massachusetts                     3168.              273557.          0.0116
## 22 Michigan                          3858.              280311.          0.0138
## 23 Minnesota                         2363.              296112.          0.00798
## 24 Mississippi                       4344.              312175.          0.0139
## 25 Missouri                          3279.              251050.          0.0131
## 26 Montana                           3326.              291763.          0.0114
## 27 Nebraska                          1820.              270753.          0.00672
## 28 Nevada                            3739.              272861.          0.0137
## 29 New Hampshire                     1989.              257269.          0.00773
```

```
## 30 New Jersey                3917.              310343.          0.0126
## 31 New Mexico                4099.              292477.          0.0140
## 32 New York                  3674.              316403.          0.0116
## 33 North Carolina            2560.              306141.          0.00836
## 34 North Dakota              2929.              354680.          0.00826
## 35 Ohio                      3417.              269802.          0.0127
## 36 Oklahoma                  3761.              303156.          0.0124
## 37 Oregon                    2037.              213150.          0.00955
## 38 Pennsylvania              3697.              255190.          0.0145
## 39 Rhode Island              3464.              375947.          0.00921
## 40 South Carolina            3572.              332565.          0.0107
## 41 South Dakota              3428.              296400.          0.0116
## 42 Tennessee                 4001.              327788.          0.0122
## 43 Texas                     3138.              273027.          0.0115
## 44 Utah                      1393.              336316.          0.00414
## 45 Vermont                   1159.              225116.          0.00515
## 46 Virginia                  2572.              245548.          0.0105
## 47 Washington                1890.              238905.          0.00791
## 48 West Virginia             4157.              336167.          0.0124
## 49 Wisconsin                 2628.              321815.          0.00817
## 50 Wyoming                   3281.              306112.          0.0107
```

Total United States Summary for Cases and Deaths per Million

```
## # A tibble: 1 x 3
##   Total_Deaths_Per_Million Total_Cases_Per_Million Deaths_Per_Cases
##                      <dbl>                   <dbl>            <dbl>
## 1                    3140.                 288392.           0.0109
```

From the summary tables, we can see the total deaths per million and cases per million as well as a ratio of deaths to cases. The highest cases per million is 407579 in Alaska. Whereas lowest cases per million is in the state of Maryland at 207531. This shows there was significant variance between states in terms of how many citizens got Covid-19 per million people.

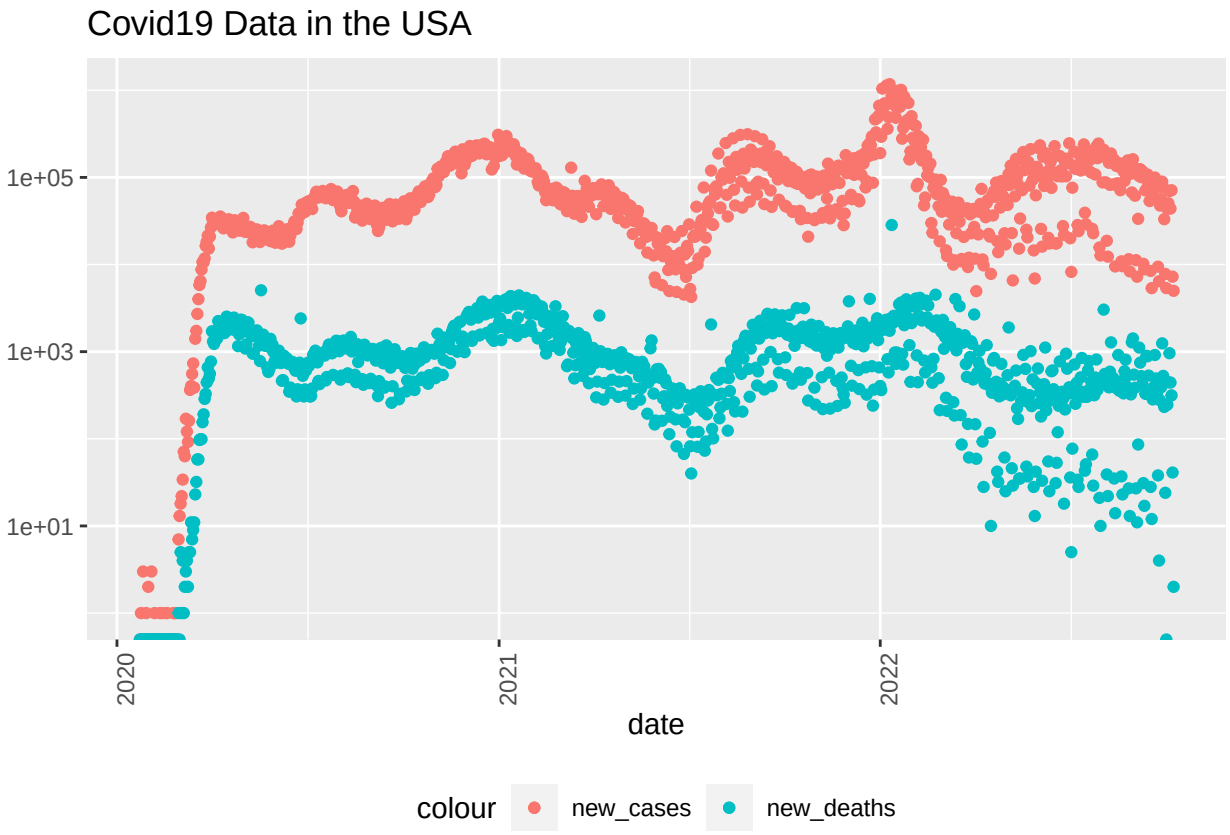
On the other hand, it can be seen that the lowest deaths per million is in Vermont at 1159. The state with the highest deaths per million was discovered to be Mississippi.

In regards to the deaths per cases ratio that was calculated, Pennsylvania has the highest ratio at 0.145. The state with the lowest deaths to cases ratio is Utah with a ratio of 0.00414.

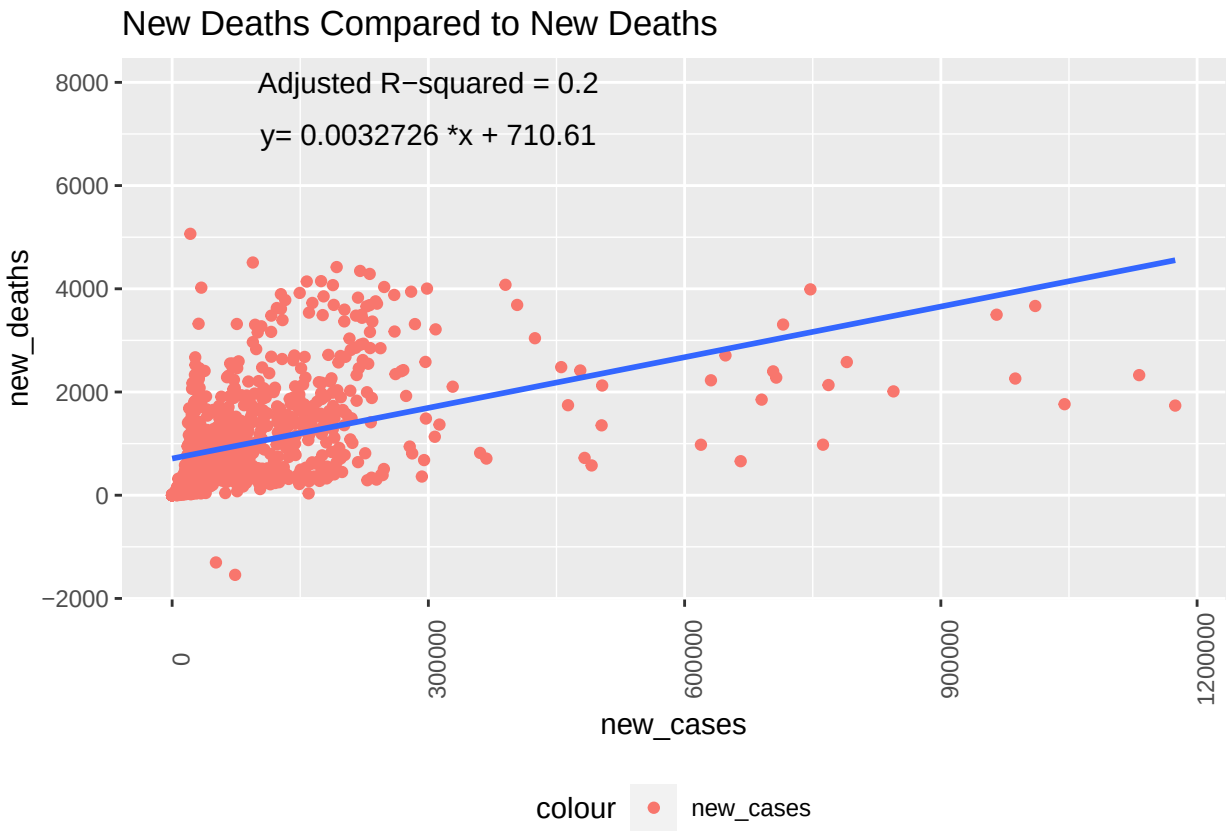
The deaths per cases ratio was caculated by dividing the total number of cases per million by the total number of deaths per million. This calculation effectively shows which states faired best in terms of least deaths per cases of Covid-19 infection.

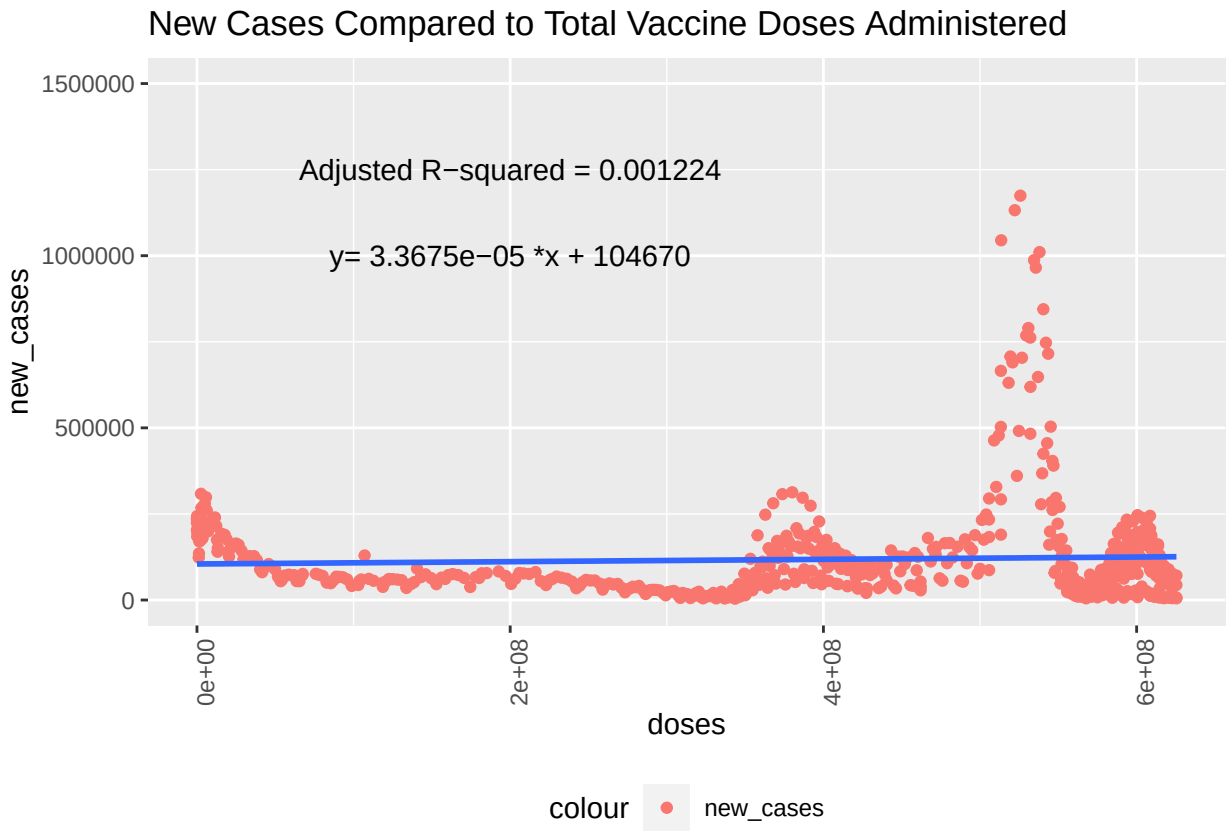
##Number of New Cases and Doses per Day

A new data frame was generated and plotted to indicate the new number of deaths and cases for each day.



Furthermore, a plot for new deaths versus doses and a plot for new cases versus doses was created to examine the data for general trends between the number of doses administered and total Covid-19 cases and Covid-19 caused deaths. Along with these two plots, a linear regression line was fit over the data and an adjusted R-squared value was calculated to indicate the regressions' modeling strengths.


New Deaths Compared to New Deaths
Adjusted R−squared = 0.2
y= 0.0032726 *x + 710.61
8000
6000
4000
2000
0
−2000
new_deaths
0
300000
600000
900000
1200000
new_cases
colour
new_cases



From the above plots it can be seen that there is somewhat of a positive slope in the regression line created for the "New Cases Compared to Total Vaccine Doses Administered plot". The slope of this regression line was 3.9328e-05. This is very small number and indicates a slope of almost zero. Furthermore, the adjusted R-squared value for this regression line is 0.001224, indicating a very weak relationship between number of doses administered and the number of new Covid-19 cases. Therefore, this plot indicates that there is not a linear correlation in the United States between the total number of doses administered and the number of new cases of Covid-19 that arose.

The plot of new Covid-19 related deaths versus total vaccine doses administered indicates that a negative relationship between new deaths and vaccine doses. The fitted regression line, visually indicates an inverse relationship between these two variables. However, the adjusted R-squared value identified for this regression is 0.1818. This r-squared value does not indicate a strong linear correlation and therefore this data can not exclusively prove a correlation between new Covid-19 related deaths and total vaccine doses administered.

A third plot was generated for Covid-related deaths per million versus Covid-19 cases per million. This plot was also fitted with a linear regression and showed a positive correlation between the number of cases and number of Covid-19 related deaths. However, this linear model may not be a strong predictor in estimating how many deaths would be caused given a number of Covid-19 cases. This is because the model's calculated adjusted-squared value was only 0.2.

#Major Conclusions Drawn

From the analysis of Covid-19 cases, Covid-19 related deaths, and total vaccine doses administerd in the United States– several things were discovered. First, from plotting the deaths per million, cases per million, and total vaccine number for the United States and for different regions of the U.S. it was discovered that rise in cases and deaths followed similar trends of increasing rapidly and then tapering off for all of the U.S. regions. No major anomalies in trends were discovered.

Next, a closer look at individual state data was taken. This revealed that The highest cases per million is in

Alaska and lowest cases per million is in Maryland. Also the lowest deaths per million is in Vermont at and highest deaths per million was discovered to be Mississippi. The deaths per cases ratio that was calculated showed that Pennsylvania has the highest ratio and Utah has the lowest with a ratio. This indicates that people who caught Covid-19 in Pennsylvania were most likely to die of it and least likely to in the state of Utah.

An analysis was also completed to observe the number of new cases per day and new deaths per day over time. Similarly to the earlier plots showing total deaths per million and total cases per million there is a large rise in early 2020 followed by a leveling off in early 2020. The number of new cases per day then oscillates around 10000 and number of new deaths per day oscillates around 1000.

Lastly, the data was modeled against different variables. The first model compared the number of new Covid-19 cases against the number of vaccine doses administered. The second modeled the number of new deaths against the number of vaccine doses administered. Both models displayed low adjusted R-squared values so they may be ineffective in predicting accurate estimations for new cases or new deaths given by the number of doses of Covid-19 vaccine administered. From the analysis, the number of Covid-19 related deaths did decrease with number of doses, The number of Covid-19 cases seemed not to be affected by number of doses as the regression line indicated a slope of nearly zero for this relationship.

#Possible Bias and Analysis Short-Comings

As mentioned earlier, one shortcoming of the data utilized in this analysis was within the Vaccine Dose data. This data specifically looked at the total number of doses administered this does not include any information regarding how many of doses were the first or second dose in the cases of vaccines such as Moderna or Pfizer. It also does not account for booster shots. Therefore, the data cannot be used for assumptions regarding efficacy in reducing Covid-19 cases and Covid-related deaths in regards to full or partial vaccination. Efficacy by vaccine brand also cannot be described.

Furthermore, it is worth mentioning that there could be possible typos or inaccuracy in the data recorded by the John Hopkins Github sources. Through this analysis it is not possible to identify all possible errors within the data, especially minute discrepancies.

In observing the new cases and new deaths against the number of vaccine doses administered, a regression line was created. However, the regression showed low adjusted R-squared values indicating that the regression line does not have a strong fit to the data. The regression does show a general trend of decreased new cases and new deaths, but it cannot be considered a strong correlation due to the low adjusted R-squared. Additionally, there may be unaccounted factors that caused changes to deaths and Covid-19 cases. These may include factors like loosening of quarantine policies and seasonal changes altering peoples' behavior and making them more susceptible to disease exposure and infection.

R Code Used to Perform this Analysis has been Listed Below

```
library("tidyverse")
library("lubridate")

url_begin<-("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c

#obtain the beginning portion of url for files needed

file_names<-c("time_series_covid19_confirmed_US.csv", "time_series_covid19_deaths_US.csv")
#grab files from github and create vector with second half of url adress

urls_combined<-str_c(url_begin, file_names)
#create a list with the entire urls for all the csv files we need using string
```

```r
#concatenate


US_Cases<-(read.csv(urls_combined[1], check.names=FALSE,
                    stringsAsFactors = FALSE))

US_Deaths<-(read.csv(urls_combined[2], check.names=FALSE,
                     stringsAsFactors = FALSE))

#Now let's do the same for the US data that we have available

US_Cases<-US_Cases %>% pivot_longer(cols=-c(UID, iso2,iso3,code3, FIPS, Admin2,
                  Province_State, Country_Region, Lat, Long_,Combined_Key),
                  names_to = "date", values_to = "cases") %>%
                  select(-c(UID, iso2,iso3,code3, FIPS,Lat,Long_,Combined_Key))

#tidy the data so it goes by date and remove unnecessary columns
US_Deaths<- US_Deaths %>% pivot_longer(cols=-(UID:Population),
                          names_to = "date", values_to = "deaths") %>%
                          select(Admin2:deaths) %>%
                          select(-c(Lat, Long_, Combined_Key))

US_data<- US_Cases %>% full_join(US_Deaths) %>% mutate(date = mdy(date)) %>%
  filter(cases>0) %>% filter(!Population==0)

summary(US_data)

Uscheck<-US_data %>% filter(cases>3400000) %>% filter(deaths> 33000)
Uscheck

#create vector for rach region of the united states to use and also remove
#locations from DF that are not a part of the 50 states

West<-c("Washington", "Oregon", "Idaho", "Montana", "Wyoming", "Colorado",
        "Utah", "California", "Nevada", "Alaska", "Hawaii")
Southwest<-c("Arizona","New Mexico","Texas","Oklahoma")
Northeast<-c("Pennsylvania","Massachusetts","New York","New Jersey","Maine",
             "Vermont","Rhode Island","Connecticut","New Hampshire")
Midwest<-c("North Dakota","Minnesota","South Dakota","Iowa","Nebraska","Kansas",
           "Missouri","Wisconsin","Illinois","Indiana","Michigan","Ohio")
Southeast<-c("West Virginia","Delaware","Virginia","Kentucky","Tennessee",
             "Maryland","North Carolina","South Carolina","Georgia","Florida",
             "Alabama","Mississippi","Lousiana","Arkansas")

notstatelist<-c("District of Columbia","Grand Princess", "Diamond Princess",
                "Guam", "Northern Mariana Islands", "Puerto Rico",
                "American Samoa", "Virgin Islands")

#now we take the data set and we group the data to get deaths per million,
#total cases and total deaths per state

US_state_data<-US_data %>% group_by(Province_State,Country_Region, date) %>%
  summarize(cases=sum(cases), deaths= sum(deaths),
```

```
  Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths*1000000/ Population) %>%
  mutate(cases_per_mill= cases*1000000/Population) %>%
  select(Province_State, Country_Region, date, cases, deaths,deaths_per_mill,
  Population, cases_per_mill) %>% ungroup()


#create a column in the data frame that includes region and combine occurances
#for that single region on a single date
US_regions<-US_state_data %>% mutate(region=case_when(Province_State %in%
               West~"West", Province_State %in% Southwest~"Southwest",
               Province_State %in% Northeast~"Northeast", Province_State %in%
               Midwest~"Midwest",Province_State %in% Southeast~"Southeast")) %>%
               aggregate(cbind(cases,deaths,deaths_per_mill,
               cases_per_mill)~date+region, FUN=sum)

US_regions<-US_regions %>% filter

#now we take the data set and we group the data to get deaths per million,
#total cases and total deaths for the entire USA
Total_US_data<- US_data %>% group_by(Country_Region, date) %>%
  summarize(cases=sum(cases), deaths=sum(deaths),
  Population=sum(Population)) %>%
  mutate(deaths_per_mill = deaths*1000000/ Population) %>%
  mutate(cases_per_mill= cases*1000000/Population) %>%
  select(Country_Region, date, cases, deaths,deaths_per_mill, cases_per_mill,
  Population)%>% ungroup()

#let's include vaccine dose data so we can model vaccine dose impact
vaccine_data<-(read.csv("https://raw.githubusercontent.com/govex/COVID-19/master/data_tables/vaccine_da

vaccine_data<- vaccine_data %>% pivot_longer(cols=-c(UID:Population),
  names_to = "date", values_to = "doses") %>%
  select(c(Province_State,Country_Region,Combined_Key,date,doses)) %>%
  filter(doses>0) %>% filter(!Country_Region=="NA") %>%
  filter(!Province_State %in% notstatelist)

#let's check for typos or wrong values in vaccine data
summary(vaccine_data)
# this gives us a max of 587903405, which seems too high

vac_data_check<- vaccine_data %>% filter(doses>300000000)
vac_data_check
#the vac data check indicates only three values (all 3 are identically
#587903405) well above 300000000
#so they are most likely typos and will be removed from the DF

vaccine_data<- vaccine_data %>% filter(doses<300000000)

#for US total Vaccine data, let's combine rows that occur on the same date
vaccine_data_tot<-vaccine_data %>% mutate(date = ymd(date)) %>%
  aggregate(doses~date,  FUN=sum)
```

```
#let's create a separate DF for vaccine data which includes region and also
#combines by date and region

vaccine_data_regional<-vaccine_data %>% mutate(date = ymd(date))%>%
  mutate(region=case_when(Province_State %in% West~"West",
       Province_State %in% Southwest~"Southwest",
       Province_State %in% Northeast~"Northeast",
       Province_State %in% Midwest~"Midwest",
       Province_State %in% Southeast~"Southeast")) %>%
  aggregate(cbind(doses)~date+region, FUN=sum)

#combine vaccine data with case/death data
Total_US_data<- Total_US_data %>% full_join(vaccine_data_tot)
US_regions<- US_regions %>% full_join(vaccine_data_regional)

#so now we can visualize the data
US_total<-Total_US_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")) +
  geom_point(aes(y=doses, color="doses"))+
  scale_y_log10()+ theme(legend.position = "bottom",
  axis.text.x = element_text(angle= 90))+
  labs(title= "Covid19 Data in the USA Total", y=NULL)
US_total
```

```
Covid_West<- US_regions %>% filter(region== "West") %>%
  ggplot(aes(x=date,y=cases_per_mill))+geom_line(aes(color="cases_per_mill"))+
  geom_point(aes(color="cases_per_mill"))+
  geom_line(aes(y=deaths_per_mill,color="deaths_per_mill"))+
  geom_point(aes(y=doses, color="doses"))+
  geom_point(aes(y=deaths_per_mill, color="deaths_per_mill"))+scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle= 90))+
  labs(title= "Covid19 Data for US Western Region", y=NULL)
Covid_West
```

```
Covid_Southwest<- US_regions %>% filter(region=="Southwest") %>%
  ggplot(aes(x=date,y=cases_per_mill))+geom_line(aes(color="cases_per_mill"))+
  geom_point(aes(color="cases_per_mill"))+
  geom_line(aes(y=deaths_per_mill,color="deaths_per_mill"))+
  geom_point(aes(y=doses, color="doses"))+
  geom_point(aes(y=deaths_per_mill, color="deaths_per_mill"))+scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle= 90))+
  labs(title= "Covid19 Data for US South western Region", y=NULL)
Covid_Southwest
```

```
Covid_Northeast<- US_regions %>% filter(region=="Northeast") %>%
  ggplot(aes(x=date,y=cases_per_mill))+geom_line(aes(color="cases_per_mill"))+
  geom_point(aes(color="cases_per_mill"))+
  geom_line(aes(y=deaths_per_mill,color="deaths_per_mill"))+
  geom_point(aes(y=doses, color="doses"))+
  geom_point(aes(y=deaths_per_mill, color="deaths_per_mill"))+scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle= 90))+
```

```
  labs(title= "Covid19 Data for US North East Region", y=NULL)
Covid_Northeast
```

```
Covid_Midwest<- US_regions %>% filter(region=="Midwest") %>%
  ggplot(aes(x=date,y=cases_per_mill))+geom_line(aes(color="cases_per_mill"))+
  geom_point(aes(color="cases_per_mill"))+
  geom_line(aes(y=deaths_per_mill,color="deaths_per_mill"))+
  geom_point(aes(y=doses, color="doses"))+
  geom_point(aes(y=deaths_per_mill, color="deaths_per_mill"))+scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle= 90))+
  labs(title= "Covid19 Data for US MidWest Region", y=NULL)
Covid_Midwest
```

```
Covid_Southeast<- US_regions %>% filter(region=="Southeast") %>%
  ggplot(aes(x=date,y=cases_per_mill))+geom_line(aes(color="cases_per_mill"))+
  geom_point(aes(color="cases_per_mill"))+
  geom_line(aes(y=deaths_per_mill,color="deaths_per_mill"))+
  geom_point(aes(y=doses, color="doses"))+
  geom_point(aes(y=deaths_per_mill, color="deaths_per_mill"))+scale_y_log10()+
  theme(legend.position = "bottom", axis.text.x = element_text(angle= 90))+
  labs(title= "Covid19 Data for US South East Region", y=NULL)
Covid_Southeast
```

```
#Look at total deaths and cases per region and also ratio of
#deaths/cases for the total USA and each state
US_regions_summ<-US_regions %>% group_by(region) %>%
  summarize(Total_Deaths_Per_Million=max(deaths_per_mill),
            Total_Cases_Per_Million=max(cases_per_mill),
            Deaths_per_Cases= max(deaths_per_mill)/max(cases_per_mill))
US_regions_summ

US_states_summ<-US_state_data %>% group_by(Province_State) %>%
  filter(!Province_State %in% notstatelist ) %>%
  summarize(Total_Deaths_Per_Million=max(deaths_per_mill),
  Total_Cases_Per_Million=max(cases_per_mill),
  Deaths_per_Cases= max(deaths_per_mill)/max(cases_per_mill))

print(US_states_summ, n=nrow(US_states_summ))


US_tot_summ<- Total_US_data %>%
  summarize(Total_Deaths_Per_Million=max(deaths_per_mill),
            Total_Cases_Per_Million=max(cases_per_mill),
            Deaths_Per_Cases= max(deaths_per_mill)/max(cases_per_mill))
US_tot_summ


#Now let's take a look at new case trendings

Total_US_datanew<- Total_US_data %>% mutate(new_cases= cases-lag(cases),
                                            new_deaths= deaths-lag(deaths))

US_totalnew<-Total_US_datanew %>% ggplot(aes(x=date,y=new_cases)) +
```

```
  geom_point(aes(color="new_cases")) +
  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= "Covid19 Data in the USA", y=NULL)
US_totalnew
```

```
us_new_with_doses<-Total_US_datanew %>% full_join(vaccine_data_tot)
us_new_with_doses<-us_new_with_doses %>% filter(!new_deaths > 25000)


c_v_de<-lm(new_deaths ~ new_cases, us_new_with_doses)
summary(c_v_de)
c_V_dec<-coef(c_v_de)
c_V_dec_eq<-paste("y=",signif(c_V_dec[[2]],digits=5),"*x","+",
                  signif(c_V_dec[[1]], digits=5))
c_V_dec_eq


new_cases_vs_new_deaths<- us_new_with_doses %>% ggplot(aes(x=new_cases,
  y=new_deaths))+ geom_point(aes(color="new_cases"))+
  theme(legend.position = "bottom", axis.text.x = element_text(angle= 90)) +
  annotate("text",x=300000,y=8000, label="Adjusted R-squared = 0.2")+
  annotate("text", x= 300000, y=7000, label= c_V_dec_eq)+
  labs(title= "New Deaths Compared to New Deaths")+
    geom_smooth(method="lm", se=0)
new_cases_vs_new_deaths
```

```
c_v_d<-lm(new_cases ~ doses, us_new_with_doses)
c_V_dc<-coef(c_v_d)
c_V_dc_eq<-paste("y=",signif(c_V_dc[[2]],digits=5),"*x",
  "+",signif(c_V_dc[[1]], digits=5))


new_cases_vs_doses<- us_new_with_doses %>% ggplot(aes(x=doses, y=new_cases))+
  geom_point(aes(color="new_cases"))+
  scale_y_continuous(limits=c(1000,1500000))+
  theme(legend.position = "bottom", axis.text.x = element_text(angle= 90))+
  annotate("text",x=200000000,y=1250000, label="Adjusted R-squared = 0.001224")+
  annotate("text", x= 200000000, y=1000000, label= c_V_dc_eq)+
  labs(title= "New Cases Compared to Total Vaccine Doses Administered")+
  geom_smooth(method="lm", se=0)

new_cases_vs_doses
```

```
d_v_d<-lm(new_deaths ~ doses, us_new_with_doses)
d_V_dc<-coef(c_v_d)
d_V_dc_eq<-paste("y=",signif(d_V_dc[[2]],digits=5),"*x","+",signif(d_V_dc[[1]],
  digits=5))


new_deaths_vs_doses<- us_new_with_doses %>%
    ggplot(aes(x=doses, y=new_deaths))+ geom_point(aes(color="new_deaths"))+
```

```
   geom_line(aes(color="new_deaths"))+ scale_y_log10()+
   theme(legend.position = "bottom", axis.text.x = element_text(angle= 90))+
   annotate("text",x=200000000,y=30, label="Adjusted R-squared = 0.1818")+
   annotate("text", x= 200000000, y=10, label= d_V_dc_eq)+
   labs(title= "New Deaths Compared to Total Vaccine Doses Administered")+
   geom_smooth(method="lm", se=0)
new_deaths_vs_doses
```