COVID-19 Data Analysis Report

October 1, 2021

```
options(digits = 5)
```

Load and Read Files

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_cov
file_names <- c("time_series_covid19_confirmed_US.csv", "time_series_covid19_confirmed_global.csv", "ti
urls <- str_c(url_in,file_names)
urls
```

```
[1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_
[2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_
[3] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_
[4] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_
[5] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_
```

```
US_Cases <- read_csv(urls[1])
Global_Cases <- read_csv(urls[2])
US_Deaths <- read_csv(urls[3])
Global_Deaths <- read_csv(urls[4])
```

Clean Data

```
Global_Cases <- Global_Cases%>%
    pivot_longer(cols= -c('Province/State', 'Country/Region', Lat, Long),
                 names_to ="date",
                 values_to = "cases") %>%
    select(-c(Lat,Long))

Global_Deaths <- Global_Deaths%>%
    pivot_longer(cols= -c('Province/State', 'Country/Region', Lat, Long),
                 names_to ="date",
                 values_to = "deaths") %>%
    select(-c(Lat,Long))


US_Cases <- US_Cases%>%
    pivot_longer(cols= -(UID:Combined_Key),
                  names_to = "date",
                  values_to = "cases") %>%
```

```
    select(Admin2:cases) %>%
    select(-c(Lat,Long_))

US_Deaths <- US_Deaths%>%
    pivot_longer(cols= -(UID:Population),
                 names_to = "date",
                 values_to = "deaths") %>%
    select(Admin2:deaths) %>%
    select(-c(Lat,Long_))


Global_Cases$date <- date(mdy(Global_Cases$date))
Global_Deaths$date <- date(mdy(Global_Deaths$date))
US_Cases$date <- date(mdy(US_Cases$date))
US_Deaths$date <- date(mdy(US_Deaths$date))
```

Explore some of the values of the columns in Global Cases

```
# proportion of NA's to the number of rows in the data set
sum(is.na(Global_Cases$date))/nrow(Global_Cases)
```

```
[1] 0
```

```
#Explore some of the values of the columns Global Cases
min(Global_Cases$date)
```

```
[1] "2020-01-22"
```

```
max(Global_Cases$date)
```

```
[1] "2021-10-11"
```

```
min(Global_Cases$cases)
```

```
[1] 0
```

```
max(Global_Cases$cases)
```

```
[1] 44455949
```

Explore some of the values of the columns in US Cases

```
# proportion of NA's to the number of rows in the data set
sum(is.na(US_Cases$date))/nrow(US_Cases)
```

```
[1] 0
```

```
#Explore some of the values of the columns in US Cases
min(US_Cases$date)
```

```
[1] "2020-01-22"
```

```
max(US_Cases$date)
```

```
[1] "2021-10-11"
```

```
min(US_Cases$cases)
```

```
[1] 0
```

```
max(US_Cases$cases)
```

```
[1] 1471645
```

Explore some of the values of the columns in Global Deaths

```
# proportion of NA's to the number of rows in the data set
sum(is.na(Global_Deaths$date))/nrow(Global_Deaths)
```

```
[1] 0
```

```
#Explore some of the values of the columns Global Deaths
min(Global_Deaths$date)
```

```
[1] "2020-01-22"
```

```
max(Global_Deaths$date)
```

```
[1] "2021-10-11"
```

```
min(Global_Deaths$deaths)
```

```
[1] 0
```

```
max(Global_Deaths$deaths)
```

```
[1] 714055
```

Explore some of the values of the columns in US Deaths

```
# proportion of NA's to the number of rows in the data set
sum(is.na(US_Deaths$date))/nrow(US_Deaths)
```

```
[1] 0
```

```
#Explore some of the values of the columns US Deaths
min(US_Deaths$date)
```

```
[1] "2020-01-22"
```

```
max(US_Deaths$date)
```

```
[1] "2021-10-11"
```

```
min(US_Deaths$Population)
```

```
[1] 0
```

```
max(US_Deaths$Population)
```

```
[1] 10039107
```

```
min(US_Deaths$deaths)
```

```
[1] 0
```

```
max(US_Deaths$deaths)
```

```
[1] 26338
```

Joining data sets and Transforming

```
Global <- Global_Cases %>%
    full_join(Global_Deaths) %>%
    rename(Country_Region = 'Country/Region',
           Province_State = 'Province/State')
# Filter to cases > 0
Global <- Global %>% filter(cases > 0)

#Checking data after the join that it matches the data before we joined it
summary(Global)
```

```
 Province_State     Country_Region          date                cases
 Length:159434      Length:159434      Min.   :2020-01-22   Min.   :       1
 Class :character   Class :character   1st Qu.:2020-07-29   1st Qu.:     382
 Mode  :character   Mode  :character   Median :2020-12-25   Median :    4622
```

```
                                    Mean   :2020-12-22   Mean   :  336438
                                    3rd Qu.:2021-05-20   3rd Qu.:   75351
                                    Max.   :2021-10-11   Max.   :44455949
     deaths
 Min.   :     0
 1st Qu.:     3
 Median :    69
 Mean   :  7675
 3rd Qu.:  1335
 Max.   :714055
```

```
US <- US_Cases %>%
  full_join(US_Deaths)

#Checking data after the join that it matches the data before we joined it
summary(US)
```

```
    Admin2          Province_State     Country_Region     Combined_Key
 Length:2102118     Length:2102118     Length:2102118     Length:2102118
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



      date                cases           Population           deaths
 Min.   :2020-01-22   Min.   :      0   Min.   :       0   Min.   :    0.0
 1st Qu.:2020-06-27   1st Qu.:     36   1st Qu.:    9917   1st Qu.:    0.0
 Median :2020-12-01   Median :    663   Median :   24892   Median :   11.0
 Mean   :2020-12-01   Mean   :   5267   Mean   :   99604   Mean   :   99.4
 3rd Qu.:2021-05-07   3rd Qu.:   2876   3rd Qu.:   64979   3rd Qu.:   54.0
 Max.   :2021-10-11   Max.   :1471645   Max.   :10039107   Max.   :26338.0
```

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

```
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/U
```

```
uid <- read_csv(uid_lookup_url) %>%
    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)
```

```
Global_totals<- Global %>%
    group_by(Province_State,Country_Region,date) %>%
    summarize(cases = sum(cases), deaths = sum(deaths), Population = sum(Population)) %>%
    mutate(new_cases = cases - lag(cases) ,
           new_deaths = deaths - lag(deaths))
```

```
Global_totals
```

```
# A tibble: 159,434 x 8
# Groups:   Province_State, Country_Region [279]
   Province_State Country_Region date       cases deaths Population new_cases
   <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl>     <dbl>
 1 Alberta        Canada         2020-03-06     1      0    4413146        NA
 2 Alberta        Canada         2020-03-07     2      0    4413146         1
 3 Alberta        Canada         2020-03-08     4      0    4413146         2
 4 Alberta        Canada         2020-03-09     7      0    4413146         3
 5 Alberta        Canada         2020-03-10     7      0    4413146         0
 6 Alberta        Canada         2020-03-11    19      0    4413146        12
 7 Alberta        Canada         2020-03-12    19      0    4413146         0
 8 Alberta        Canada         2020-03-13    29      0    4413146        10
 9 Alberta        Canada         2020-03-14    29      0    4413146         0
10 Alberta        Canada         2020-03-15    39      0    4413146        10
# ... with 159,424 more rows, and 1 more variable: new_deaths <dbl>
```

```
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_by_State
```

```
# A tibble: 36,482 x 7
   Province_State Country_Region date       cases deaths deaths_per_mill
   <chr>          <chr>          <date>     <dbl>  <dbl>           <dbl>
 1 Alabama        US             2020-01-22     0      0               0
 2 Alabama        US             2020-01-23     0      0               0
 3 Alabama        US             2020-01-24     0      0               0
 4 Alabama        US             2020-01-25     0      0               0
 5 Alabama        US             2020-01-26     0      0               0
 6 Alabama        US             2020-01-27     0      0               0
 7 Alabama        US             2020-01-28     0      0               0
 8 Alabama        US             2020-01-29     0      0               0
 9 Alabama        US             2020-01-30     0      0               0
10 Alabama        US             2020-01-31     0      0               0
# ... with 36,472 more rows, and 1 more variable: Population <dbl>
```

```
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()

tail(US_totals)
```

```
# A tibble: 6 x 6
  Country_Region date          cases deaths deaths_per_mill Population
```

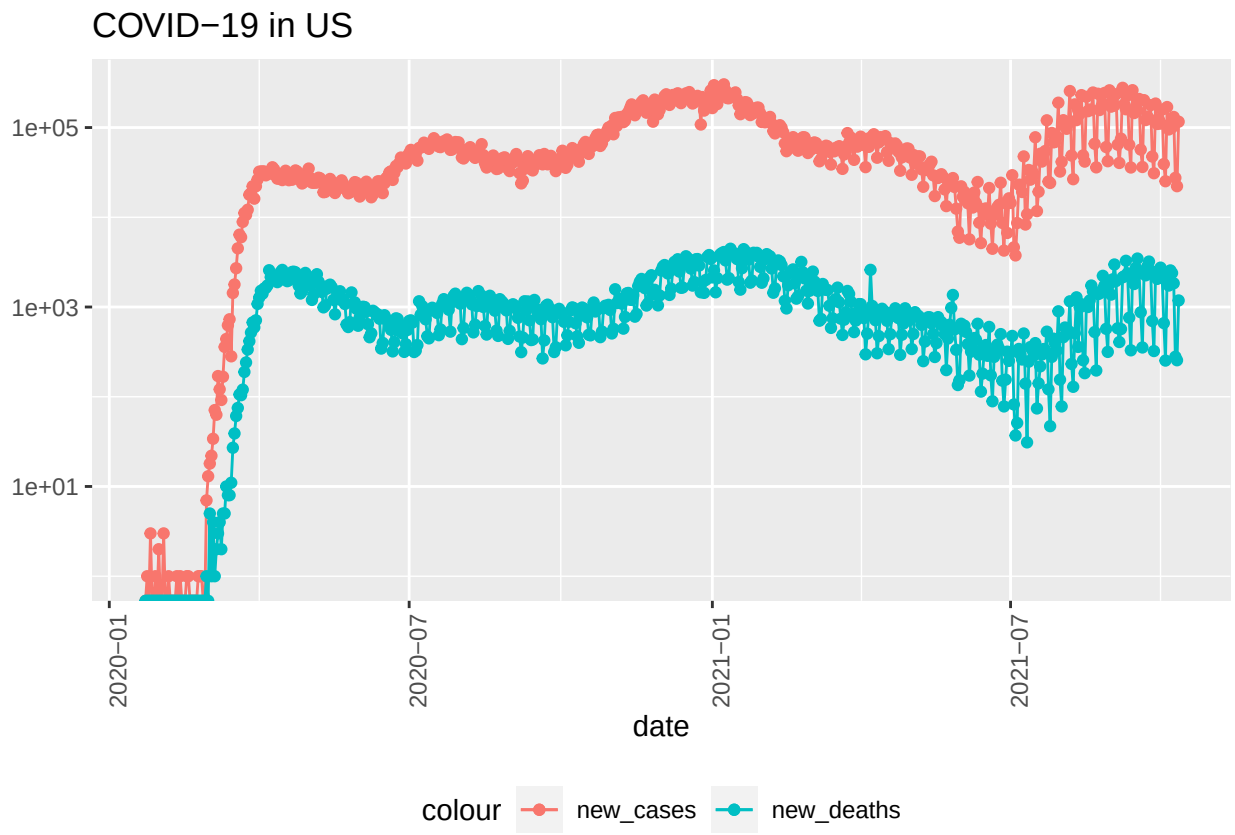
```
  <chr>          <date>        <dbl>  <dbl>           <dbl>      <dbl>
1 US             2021-10-06 44058827 708110          2127.  332875137
2 US             2021-10-07 44158910 710502          2134.  332875137
3 US             2021-10-08 44290052 712339          2140.  332875137
4 US             2021-10-09 44317553 712618          2141.  332875137
5 US             2021-10-10 44339747 712873          2142.  332875137
6 US             2021-10-11 44455957 714056          2145.  332875137
```

Analyzing and Visualizations

```
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))

tail(US_totals)
```

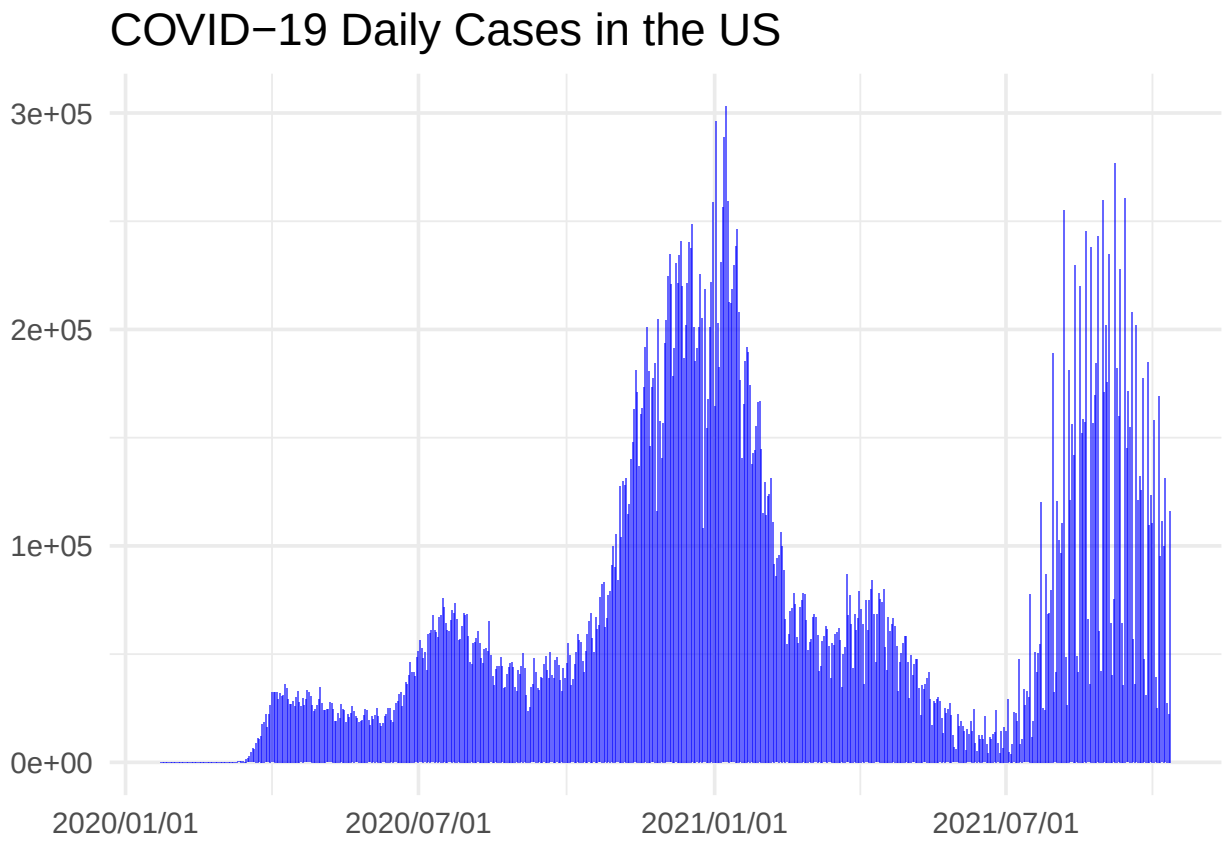
```
# A tibble: 6 x 8
  Country_Region date          cases deaths deaths_per_mill Population new_cases
  <chr>          <date>        <dbl>  <dbl>           <dbl>      <dbl>     <dbl>
1 US             2021-10-06 44058827 708110          2127.  332875137    111338
2 US             2021-10-07 44158910 710502          2134.  332875137    100083
3 US             2021-10-08 44290052 712339          2140.  332875137    131142
4 US             2021-10-09 44317553 712618          2141.  332875137     27501
5 US             2021-10-10 44339747 712873          2142.  332875137     22194
6 US             2021-10-11 44455957 714056          2145.  332875137    116210
# ... with 1 more variable: new_deaths <dbl>
```

```
#with new_cases added trends to COVID-19 in US
US_totals %>%
    filter(cases > 0) %>%
    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 in US" , y = NULL)
```

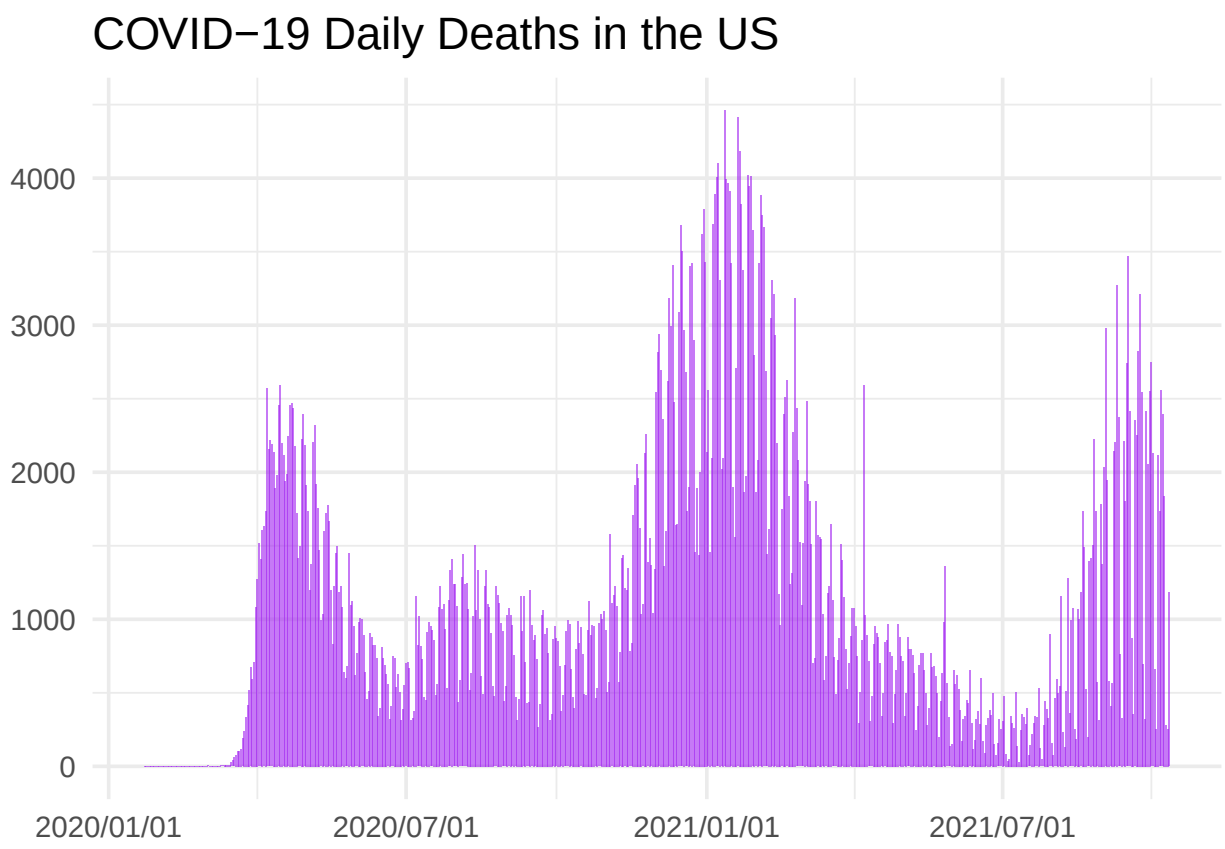



```
US_cases <- ggplot(US_totals,
 aes(date, as.numeric(new_cases))) +
 geom_col(fill = "blue", alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")
```

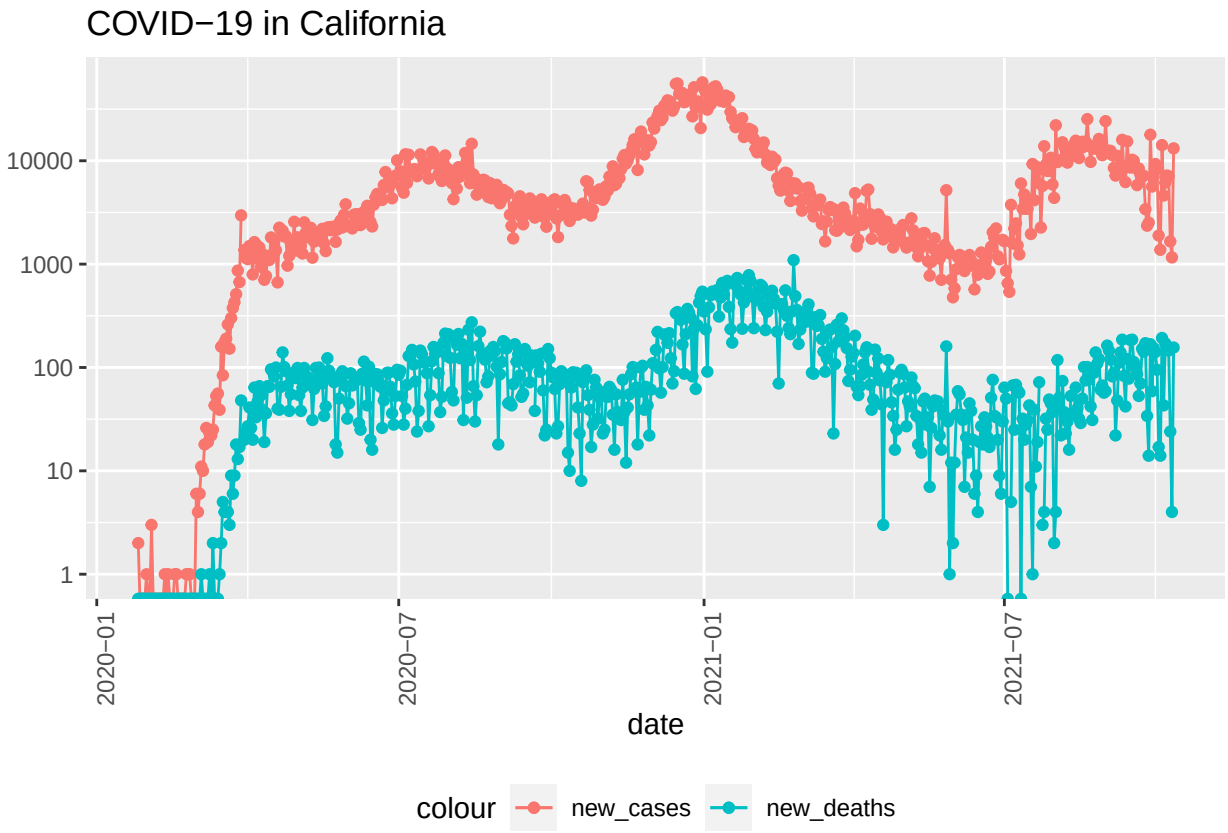
```
US_cases + labs(title = "COVID-19 Daily Cases in the US" , y = NULL)
```




```
US_deaths <- ggplot(US_totals,
 aes(date, as.numeric(new_deaths))) +
 geom_col(fill = "purple", alpha = 0.6) +
 theme_minimal(base_size = 14) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")

 US_deaths + labs(title = "COVID-19 Daily Deaths in the US" , y = NULL)
```




```
#with new_cases added trends to COVID-19 by State
state <- "California"
US_by_State %>%
    filter(Province_State == state) %>%
    filter(cases > 0) %>%
    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 in ", state), y = NULL)
```




```
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)

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

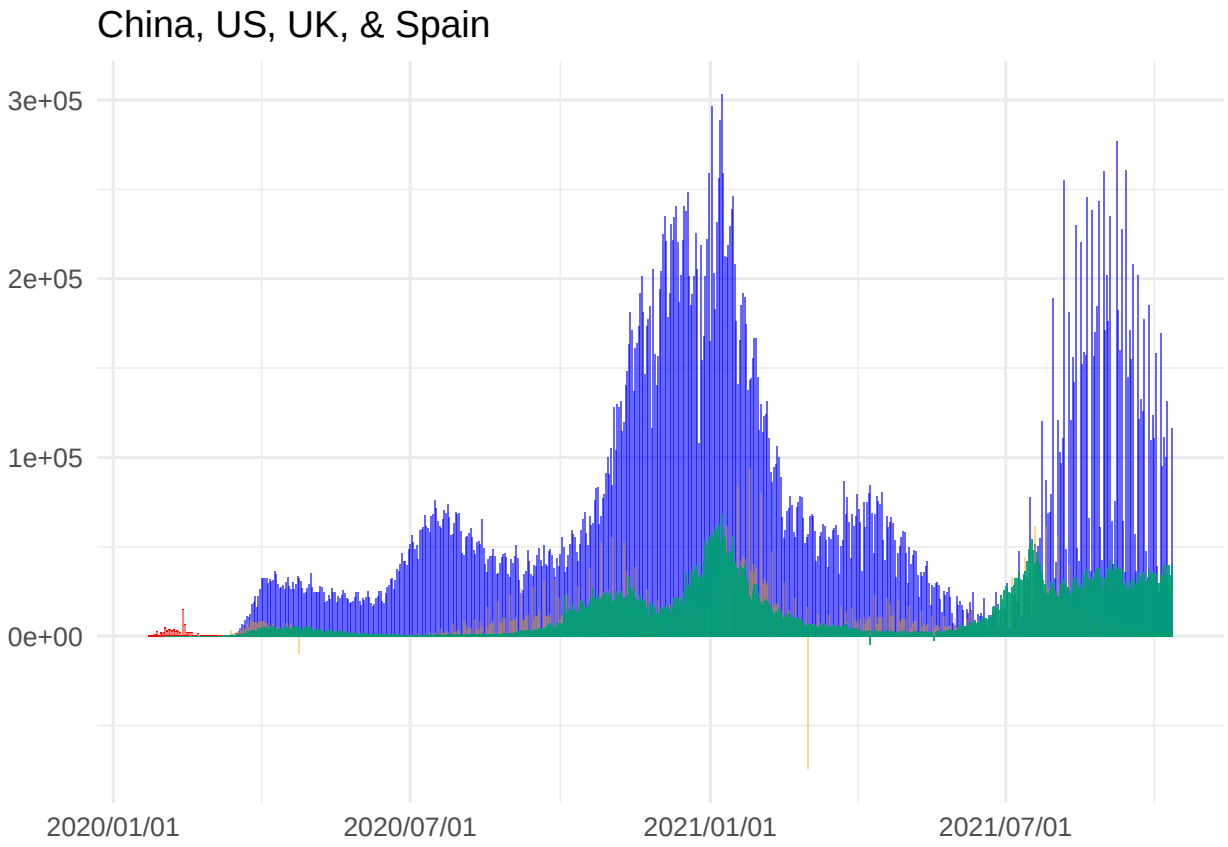
```
# A tibble: 10 x 6
   Province_State           deaths  cases population cases_per_thou deaths_per_thou
   <chr>                     <dbl>  <dbl>      <dbl>          <dbl>           <dbl>
 1 American Samoa                0      3      55641         0.0539          0
 2 Northern Mariana Islands      3    281      55144         5.10            0.0544
 3 Vermont                     335  35892     623989        57.5             0.537
 4 Hawaii                      845  81614    1415872        57.6             0.597
 5 Virgin Islands               73   6950     107268        64.8             0.681
 6 Maine                      1075  94948    1344212        70.6             0.800
 7 Alaska                      596 124123     740995       168.              0.804
 8 Puerto Rico                3192 183117    3754939        48.8             0.850
 9 Utah                       2994 520190    3205958       162.              0.934
10 Oregon                     4002 343993    4217737        81.6             0.949
```

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

```
# A tibble: 10 x 6
   Province_State deaths   cases population cases_per_thou deaths_per_thou
   <chr>           <dbl>   <dbl>      <dbl>          <dbl>           <dbl>
 1 Mississippi      9833  495312    2976149           166.            3.30
 2 New Jersey      27603 1172527    8882190           132.            3.11
 3 Louisiana       14204  748793    4648794           161.            3.06
 4 Alabama         14859  808599    4903185           165.            3.03
 5 New York        55749 2480082   19453561           127.            2.87
 6 Arizona         20382 1120361    7278717           154.            2.80
 7 Massachusetts   18746  823485    6892503           119.            2.72
 8 Arkansas         8132  503089    3017804           167.            2.69
 9 Rhode Island     2854  174570    1059361           165.            2.69
10 Florida         56667 3645290   21477737           170.            2.64
```

Comparing Multiple Countries

```
# Now lets add in a few more countries

china <- Global_totals[Global_totals$Country_Region == 'China',]
spain <- Global_totals[Global_totals$Country_Region == 'Spain',]
UK <- Global_totals[Global_totals$Country_Region == 'United Kingdom',]

USplot <- ggplot(US_totals,
 aes(date, as.numeric(new_cases))) +
 geom_col(fill = 'blue', alpha = 0.6) +
 theme_minimal(base_size =12) +
 xlab(NULL) + ylab(NULL) +
 scale_x_date(date_labels = "%Y/%m/%d")

China_US <- USplot + geom_col(data = china,
 aes(date, as.numeric(new_cases)),
 fill='red',
 alpha = 0.5)

Ch_US_Sp <- China_US + geom_col(data = spain,
 aes(date, as.numeric(new_cases)),
 fill='#E69F00',
 alpha = 0.4)

Chn_US_Sp_UK <- Ch_US_Sp + geom_col(data = UK,
 aes(date, as.numeric(new_cases)),
 fill='#009E73',
 alpha = 0.9)

Chn_US_Sp_UK + labs(title= "China, US, UK, & Spain")
```



Modeling

SIR Model

```
state <- "California"
California <- US_by_State %>%
    filter(Province_State == state) %>%
    filter(cases > 0)
```

```
#SIR Model

SIR <- function(time, state, parameters) {
  par <- as.list(c(state, parameters))
  with(par, {
    dS <- -beta * I * S/N
    dI <- beta * I * S/N - gamma * I
    dR <- gamma * I
    list(c(dS, dI, dR))
  })
}
```

```
#create a vector of cumulative cases
```

```r
infected <- California %>%
    filter(cases >0) %>%
    pull(new_cases)
```

```r
# Create an incrementing Day vector the same length as our cases vector
day <- 1:(length(infected))
N <- 14446515
#specify initial values for S, I, R
init <- c(S = N - infected[1], I = infected[1], R = 0)

RSS <- function(parameters) {
  names(parameters) <- c("beta", "gamma")
  out <- ode(y = init, times = day, func = SIR, parms = parameters)
  fit <- out[, 3]
  sum((infected - fit)^2)
}
```

```r
# now find the values of beta and gamma that give the
# smallest RSS, which represents the best fit to the data.
# Start with values of 0.5 for each, and constrain them to
# the interval 0 to 1.0
library(deSolve)
optimization <- optim(c(0.5, 0.5), RSS, method = "L-BFGS-B", lower = c(0,0), upper = c(1, 1))

# check for convergence
optimization$message
```

```
[1] "ERROR: ABNORMAL_TERMINATION_IN_LNSRCH"
```

```r
# Optimization Parameters
opt_par <- setNames(optimization$par, c("beta", "gamma"))
opt_par
```

```
   beta   gamma
0.50872 0.49128
```

```r
# Reproduction Number
R0 <- opt_par[1]/opt_par[2]
R0
```

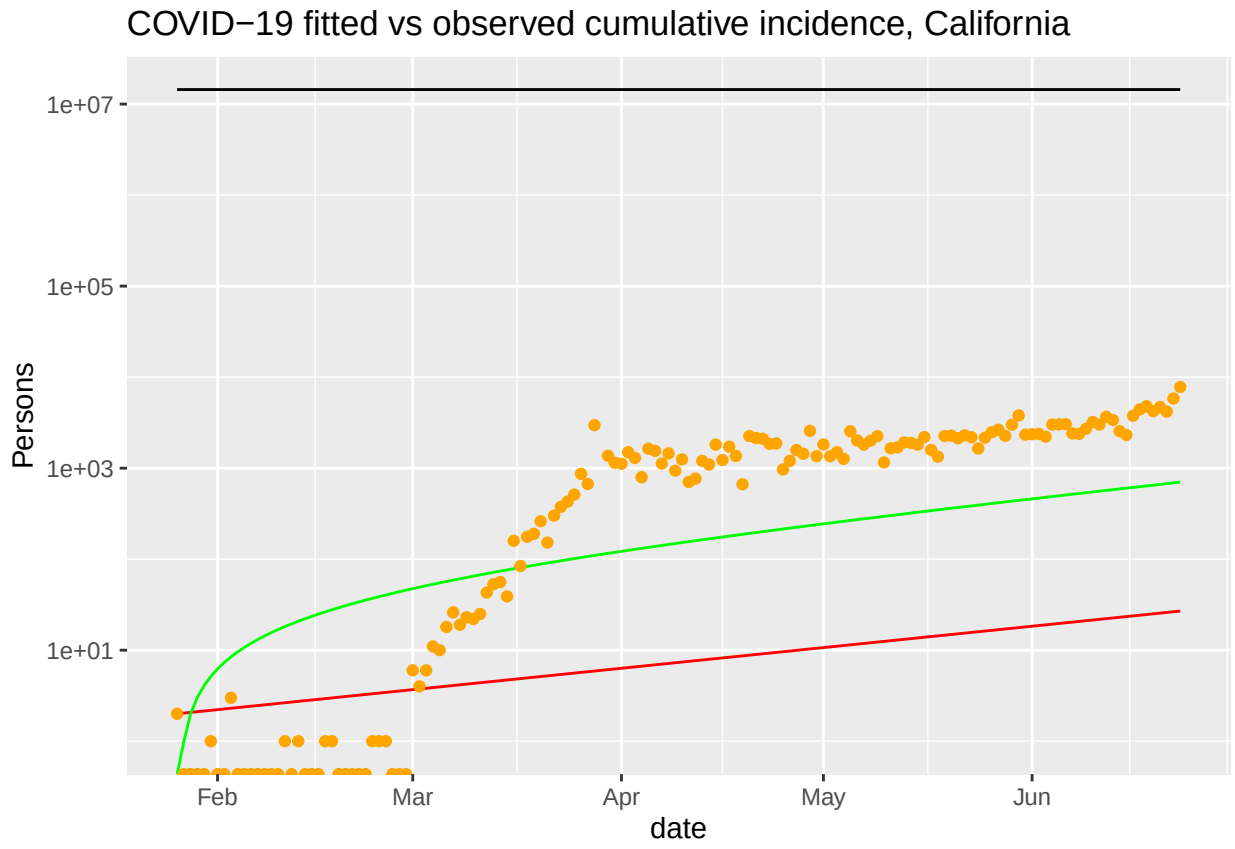
```
  beta
1.0355
```

Prediction

```r
# time in days for predictions
startdate <- min(California$date)
t <- 1:150
```

```r
# get the fitted values from our SIR model
fittedcum <- data.frame(ode(y = init, times = t, func = SIR, parms = opt_par))
# add a Date column and join the observed incidence data
fittedcum <- fittedcum %>%
    mutate(date = as.Date(startdate) + t-1)%>%
              left_join(California %>% select(date,new_cases))
```

```r
# plot the data
ggplot(fittedcum, aes(x = date)) +
  geom_line(aes(y = I), colour = "red") +
  geom_line(aes(y = S), colour = "black") +
  geom_line(aes(y = R), colour = "green") +
  geom_point(aes(y = new_cases), colour = "orange") +
  scale_y_continuous(labels = scales::comma) +
  labs(y = "Persons", title = "COVID-19 fitted vs observed cumulative incidence, California") +
  scale_colour_manual(name = "",
                      values = c(red = "red", black = "black", green = "green", orange = "orange"),
                      labels = c("Susceptible", "Recovered", "Observed incidence", "Infectious")) +
  scale_y_continuous(trans="log10")
```



Bias

Some Bias is that I decided to analyze the state of California in more detail since I reside there. As for bias that may exist in the data would be in regards to how accurately cases are reported Globally and even

state-wide. Thus, that is why I decided to focus on the state I currently reside in as it was of great interest to me.