Project Part 4: COVID19 Analysis

Group22 (asr4, janagel2, jasonjc3, sk17, vmyadam2)

April 15, 2020

Contents

Introduction

Considering current world events and how we could have never imagined a situation like this 6-7 months ago, it was prudent to do an exploration of a COVID19 dataset.

Choosing a Dataset

The Dataset we have chosen is the **COVID19 Dataset** by **Devakumar** and can be found here.

The **COVID19 Dataset** tracks the number of **Confirmed**,**Recovered**, and **Death** cases across the globe as a resut of the COVID19 pandemic. This is a great source and we was attracted to it due to its simplicity as well as the methods used to compile it, and to those interested in the compilation process, this will be of interest.

Objectives

The objetives of this analysis are simple and will be as follows:

- Create data visualizations that explain the dataset to a random audience.
- Model number of Deaths due to coronavirus.
- Compare overall model against models froms data limited by country.
  - Compare accuracy of said models against one another.

Methods

To explain this dataset to everyone, we decided to make two visuals.

1. First, simply table that shows the number of Confirmed Cases by Country and is sorted by the number of Deaths per Country. This lays out what this dataset is all about and is easy to understand.
2. We noticed the United States has the largest number of cases in the word, but since we do not know many people in the US who are affected, we wanted to explore the US further. We will create a map of the US that reflects a continous scale showing how many cases are in each state.

For the modeling process we will make three models.

1. A world model which uses data from the entire world to try and predict the number of deaths.
    - This model will be used to predit number of deaths for countries with extreme conditions like Italy and the US.
2. A US model which uses data from the US only.
    - We will compare how this model fares against the rest of the world.
3. An Italian model which uses data from Italy only
    - This too wil be compared against the rest of the world
4. The comparison will be made by checing the RMSE of the actual data against the predicted from the models as outlined above. The world mode will be tested again the entire world minus the US and Italy, to see which model performs the best under which circumstances.

Results

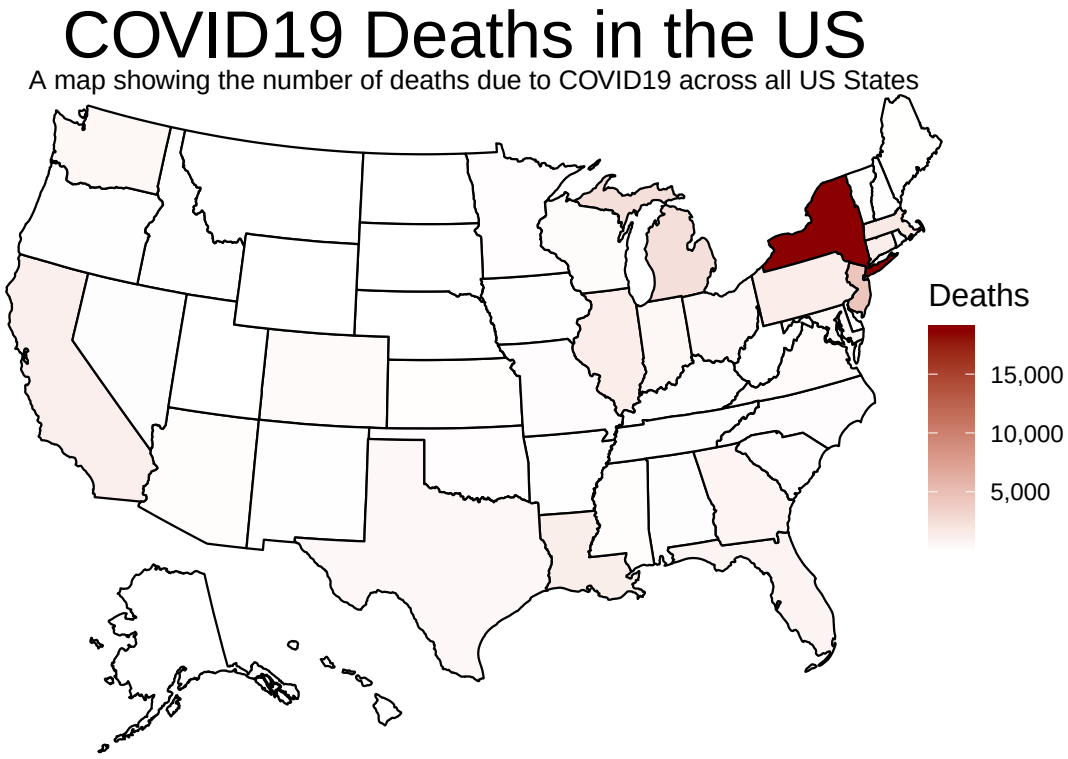
The results are as follow

Visuals

Here is a death toll by country from the COVID19 Pandemic

```
## # A tibble: 10 x 3
##    Country        `Total Deaths` `Confirmed Cases`
##    <fct>                   <int>             <int>
##  1 US                      42094            784326
##  2 Italy                   24114            181228
##  3 Spain                   20852            200210
##  4 France                  20292            156480
##  5 United Kingdom          16550            125856
##  6 Belgium                  5828             39983
##  7 Iran                     5209             83505
##  8 Germany                  4862            147065
##  9 China                    4636             83817
## 10 Netherlands              3764             33583
```

Since the United States have the highest confirmed cases, let us have a look at how the virus has spread throughout the country itself.



- This map shows that while the US itself is not doing terribly, the state of New York is in dire trouble.

Models

We made three MLR Models:

- An overall model where data from the entire world is used.
- A model where data from the US is used (the country with the most cases and deaths).
- A model where data from Italy is used (the country with the second most deaths).
- Variabes Used:
  - `Long`: Longitude
  - `Lat`: Latitude
  - `Confirmed`: Number of Confirmed Cases
  - `Recovered`: Number of Recovered Cases
  - `Days`: Number of Days elsapsed since 1/21/20

The world model was used to predict both Italian and US conditions, while the Italian and US models were used to predict the world conditions to see which was the most useful. The RMSE are shown below

- **World Model against US Data** : 3213.44
- **World Model against Italian Data**: 5572.7

- **US Model against World Data**: 2119.36
- **Italian Model against World Data**: 1.0435826×10^5
- **World Model against World Data minus US and Italy**: 379.7

Whilst the world model performs badly against both US and Itlaian sets, it completely outperforms the the US and Italian models when evluating for the entire world.

We can also run some tests on the world model to see if some of the assumptions that are made when making models are held.

We can conclude from the small p-value that the constant variance assumption has been violated.

By looking at the graphs of the data[**Appendix: Plots**], we can visually see that there are some sections that are more spread out when compared to other regions. Although this is not optimal for building a MLR model, this is the data that has been collected and it must be considered.

Conclusions

After having looked at the RMSE values for all sets, We have to conclude that the world model, which looks as follows, is the most suitable for predicting the number of **Deaths** from any **random** set of data.

```
summary(world_model)
```

```
##
## Call:
## lm(formula = Deaths ~ Lat + Long + Confirmed + Recovered + Days,
##     data = covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8387.7    -7.3     7.6    16.6 14227.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.2235245  4.9544794  -1.862   0.0627 .
## Lat          0.0435939  0.1454026   0.300   0.7643
## Long        -0.2202995  0.0509815  -4.321 1.56e-05 ***
## Confirmed    0.0511945  0.0002443 209.525  < 2e-16 ***
## Recovered    0.0195804  0.0011756  16.655  < 2e-16 ***
## Days         0.0328072  0.0007188  45.639  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 548.7 on 23836 degrees of freedom
## Multiple R-squared:  0.8068, Adjusted R-squared:  0.8067
## F-statistic: 1.991e+04 on 5 and 23836 DF,  p-value: < 2.2e-16
```

Appendix

Source: https://www.kaggle.com/imdevskp/corona-virus-report

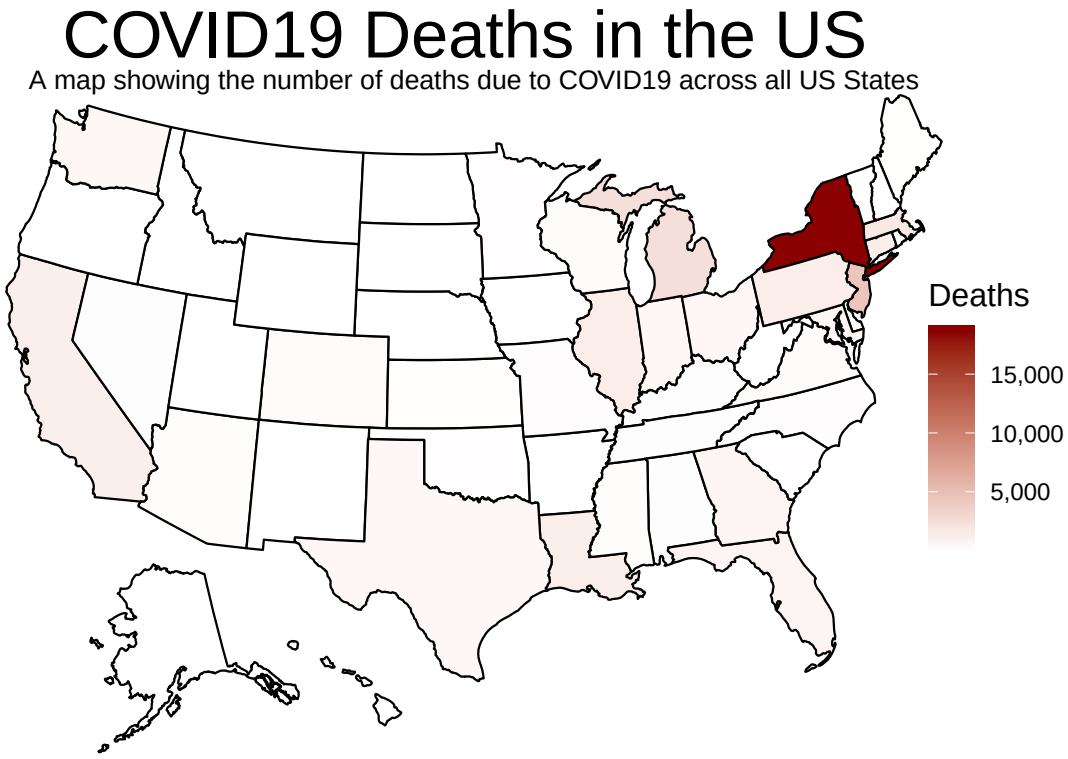
COVID-19 Death toll by country

```
names(covid)[2] <- "Country"
total_deaths = group_by(filter(covid, Date=="4/20/20"), Country)
count = arrange(summarise(total_deaths, `Total Deaths` = sum(Deaths), `Confirmed Cases` = sum(Confirmed
count
```

```
## # A tibble: 10 x 3
##    Country        `Total Deaths` `Confirmed Cases`
##    <fct>                   <int>             <int>
##  1 US                      42094            784326
##  2 Italy                   24114            181228
##  3 Spain                   20852            200210
##  4 France                  20292            156480
##  5 United Kingdom          16550            125856
##  6 Belgium                  5828             39983
##  7 Iran                     5209             83505
##  8 Germany                  4862            147065
##  9 China                    4636             83817
## 10 Netherlands              3764             33583
```

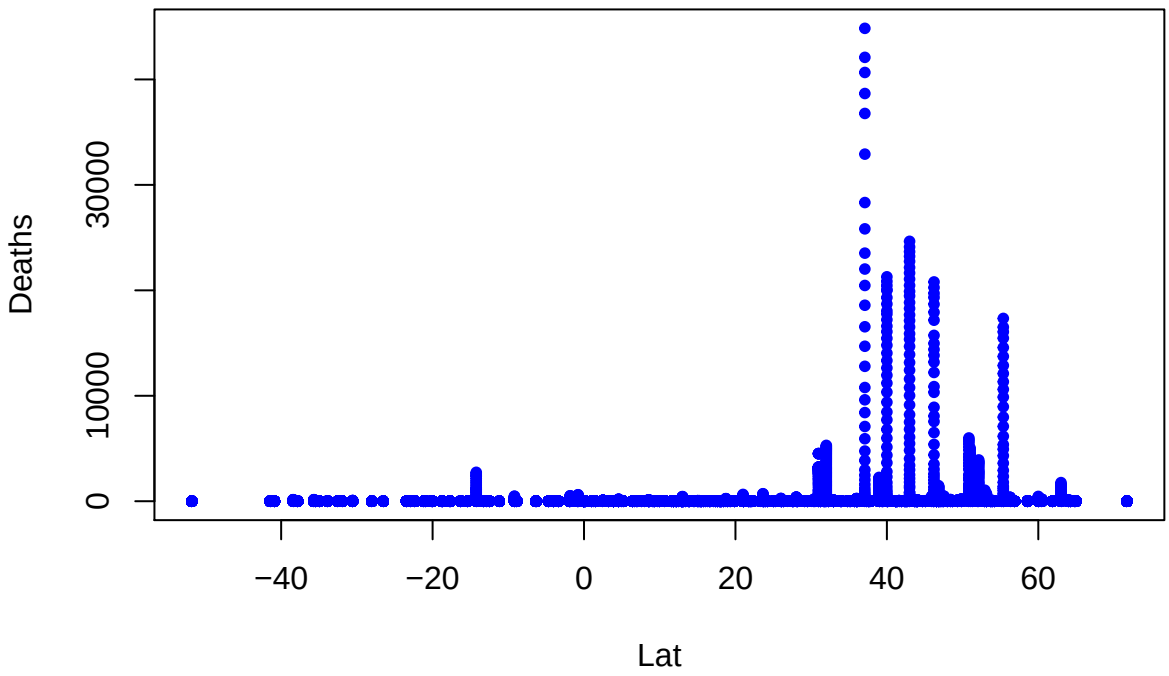
COVID-19 US Deaths Map

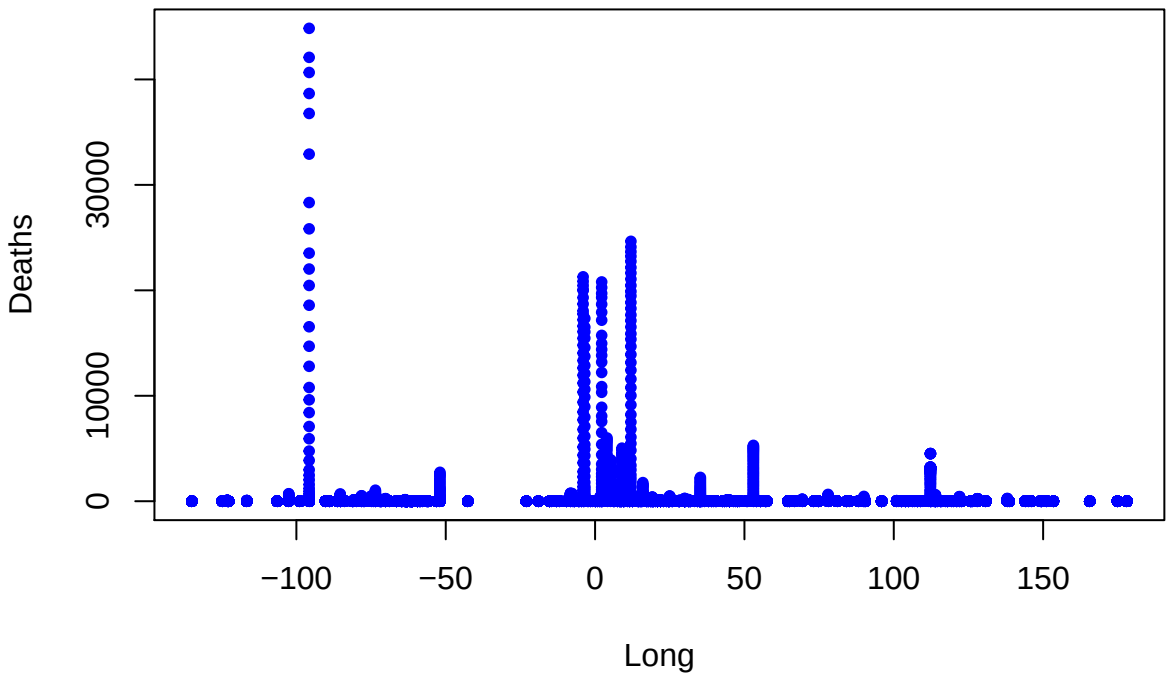
```
# create summary of deaths by state
usdeaths = group_by(filter(uscovid, Date=="4/20/20"), Province_State)
count2 = summarise(usdeaths, `Total Deaths` = sum(Deaths))
g2 <- statepop
g2$deaths = g2$abbr
for (i in g2$full) {
  g2$deaths[g2$full==i]=count2$`Total Deaths`[count2$Province_State==i]
}

g2$deaths = as.numeric(g2$deaths)

plot_usmap(data = g2, values = "deaths", color = "black") +
  labs(title = "COVID19 Deaths in the US ",
       subtitle = "A map showing the number of deaths due to COVID19 across all US States") +
  scale_fill_continuous(name = "Deaths", low="white",high="darkred", label=scales::comma) +
  theme(legend.position = c(0.93,0.3), legend.title = element_text(size=12),legend.text = element_text(
```

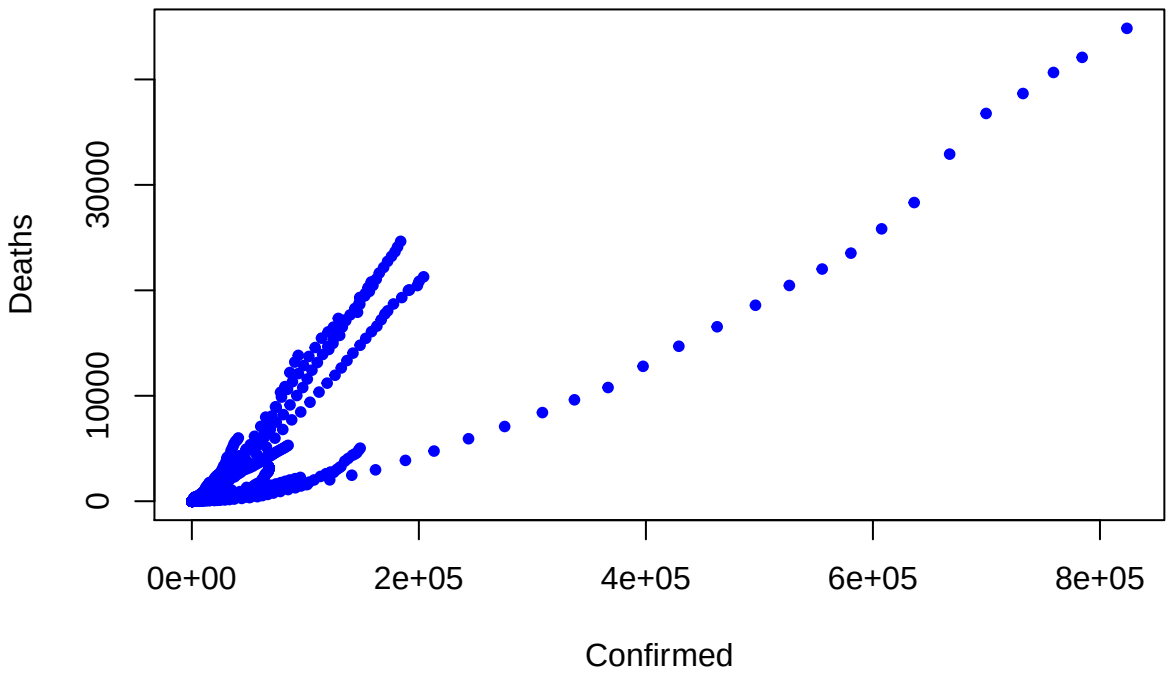


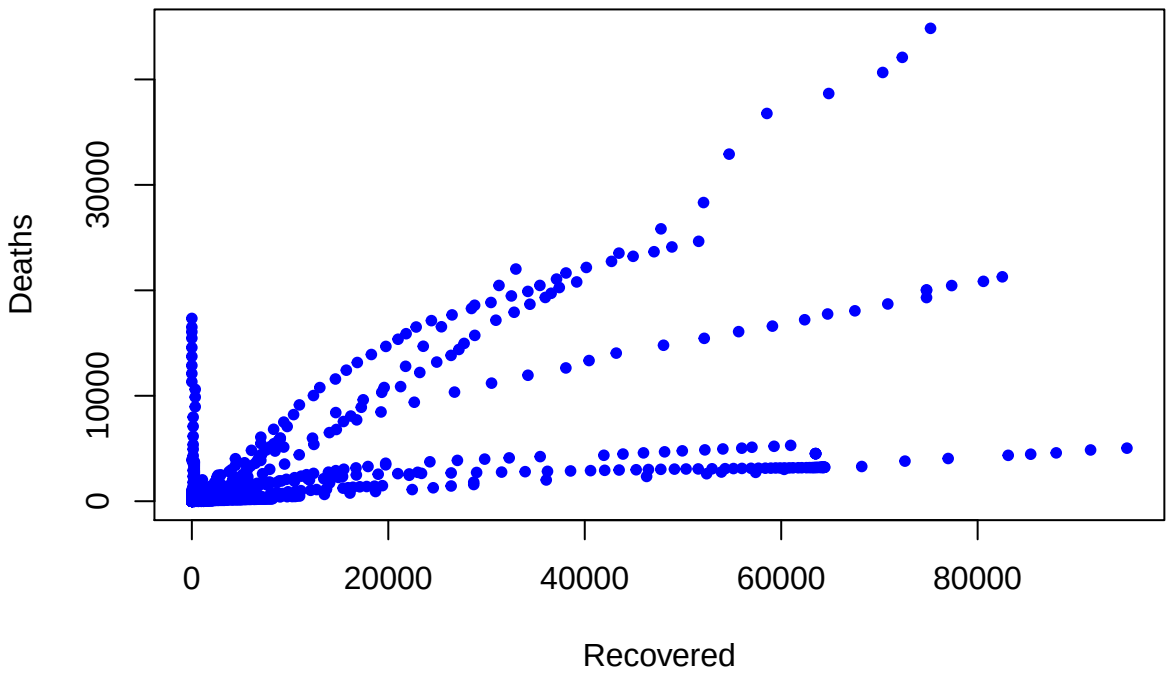
Plots

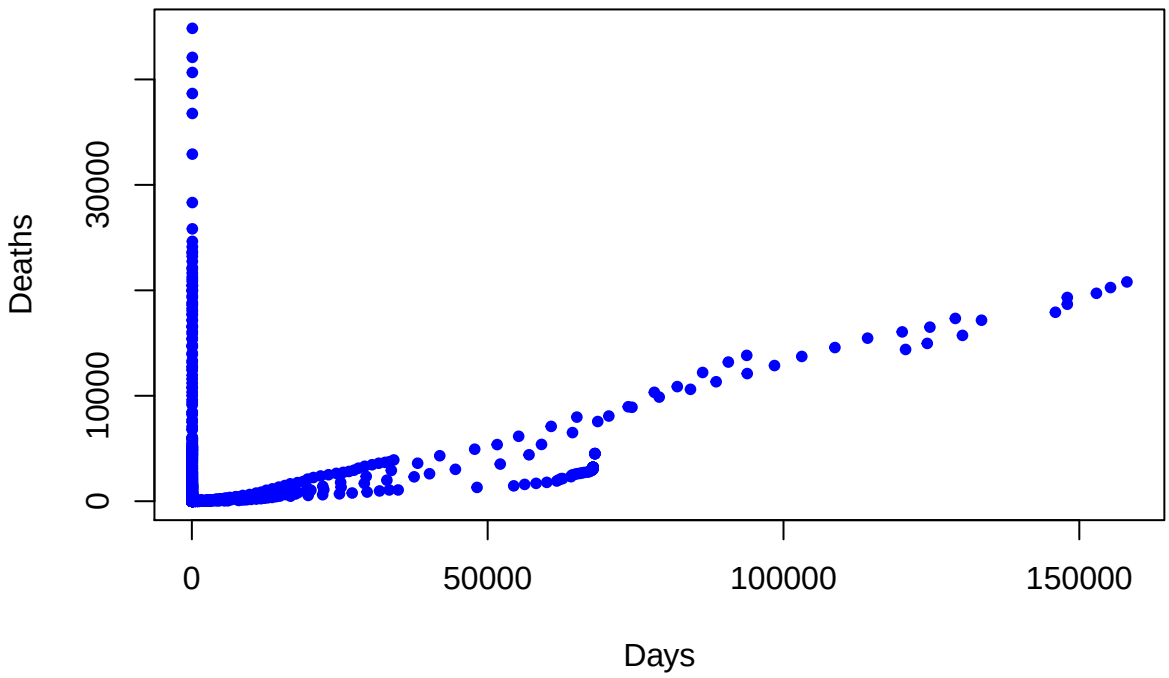
```
plot(Deaths ~ Lat + Long + Confirmed + Recovered+ Days, data = covid,
     col = "blue",
     pch = 20) # plot data
```


Deaths
0
10000
30000
−40
−20
0
20
40
60
Lat


Deaths
0
10000
30000
−100
−50
0
50
100
150
Long


Deaths
0
10000
30000
0e+00
2e+05
4e+05
6e+05
8e+05
Confirmed


Deaths
0
10000
30000
0
20000
40000
60000
80000
Recovered


Deaths
0
10000
30000
0
50000
100000
150000
Days