Final Project R

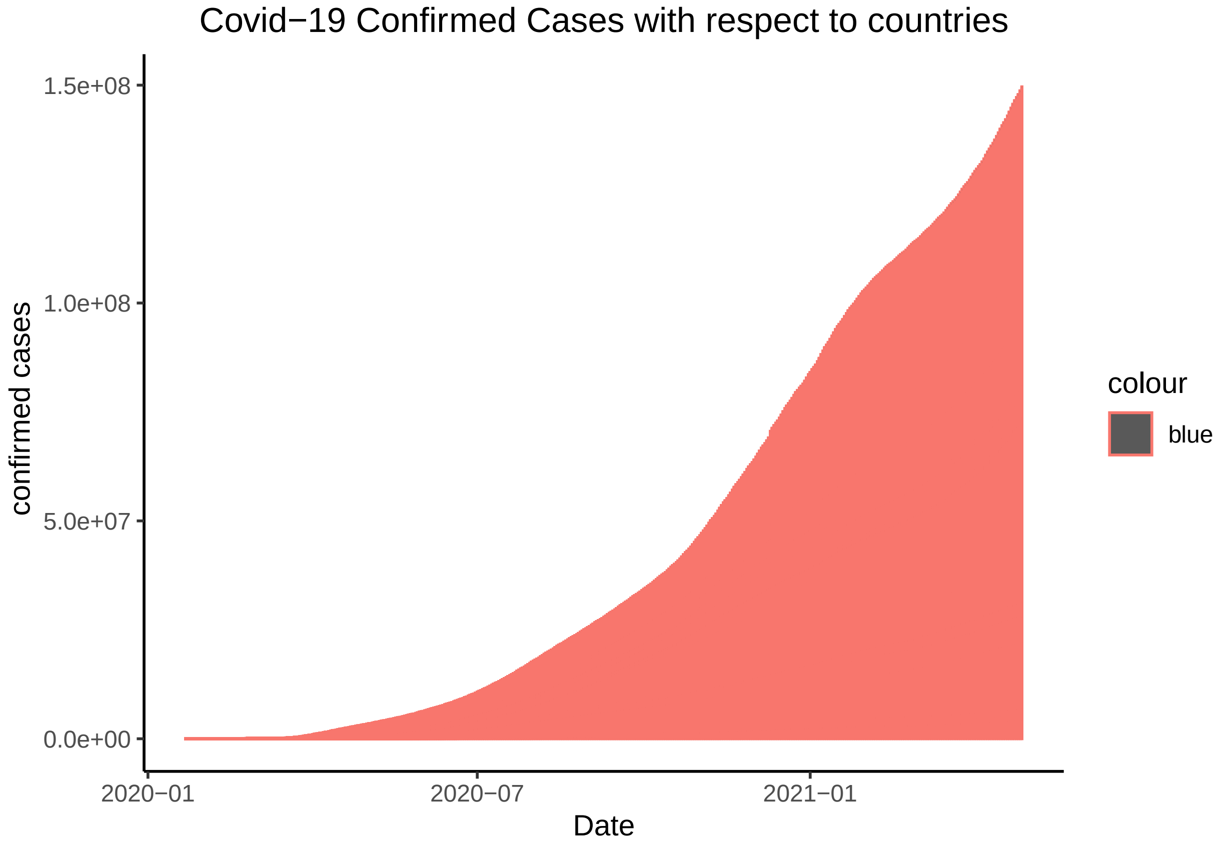
Surbhi Rathore

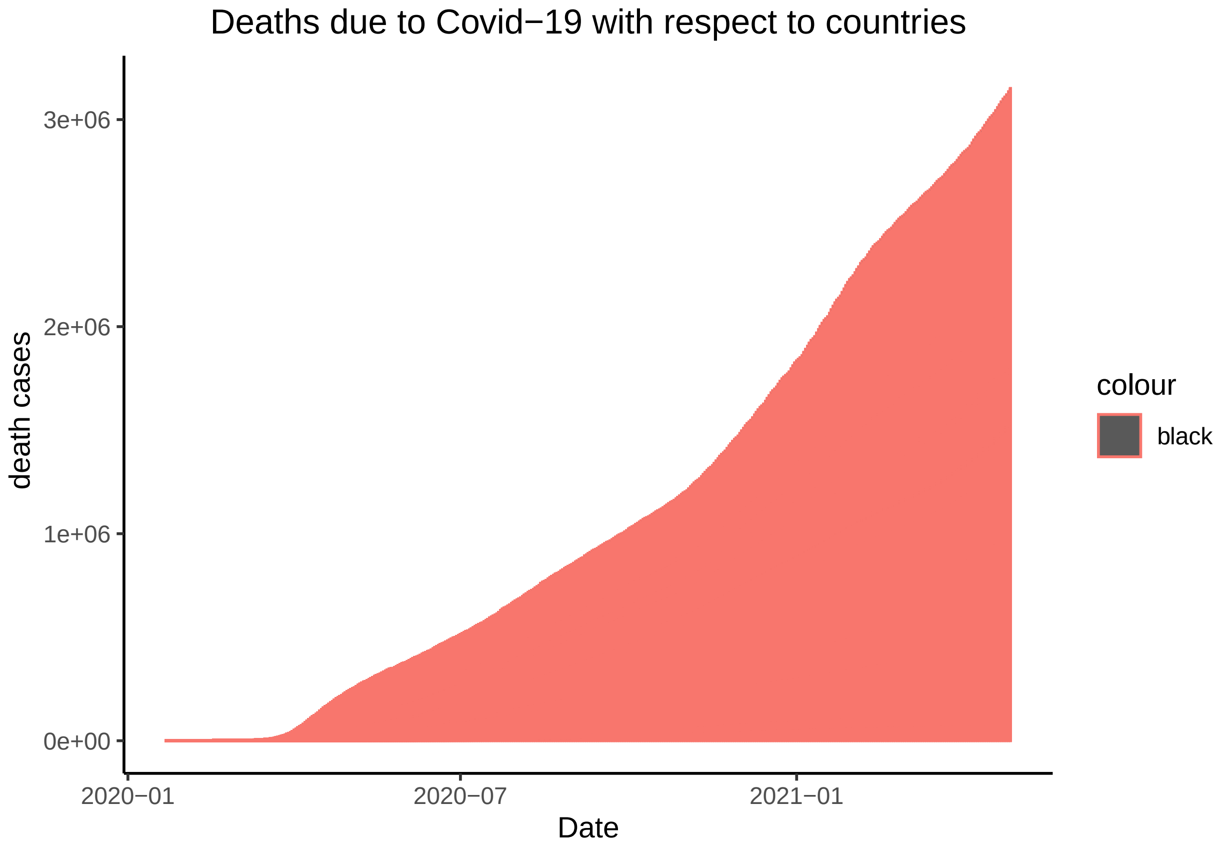
4/26/2021

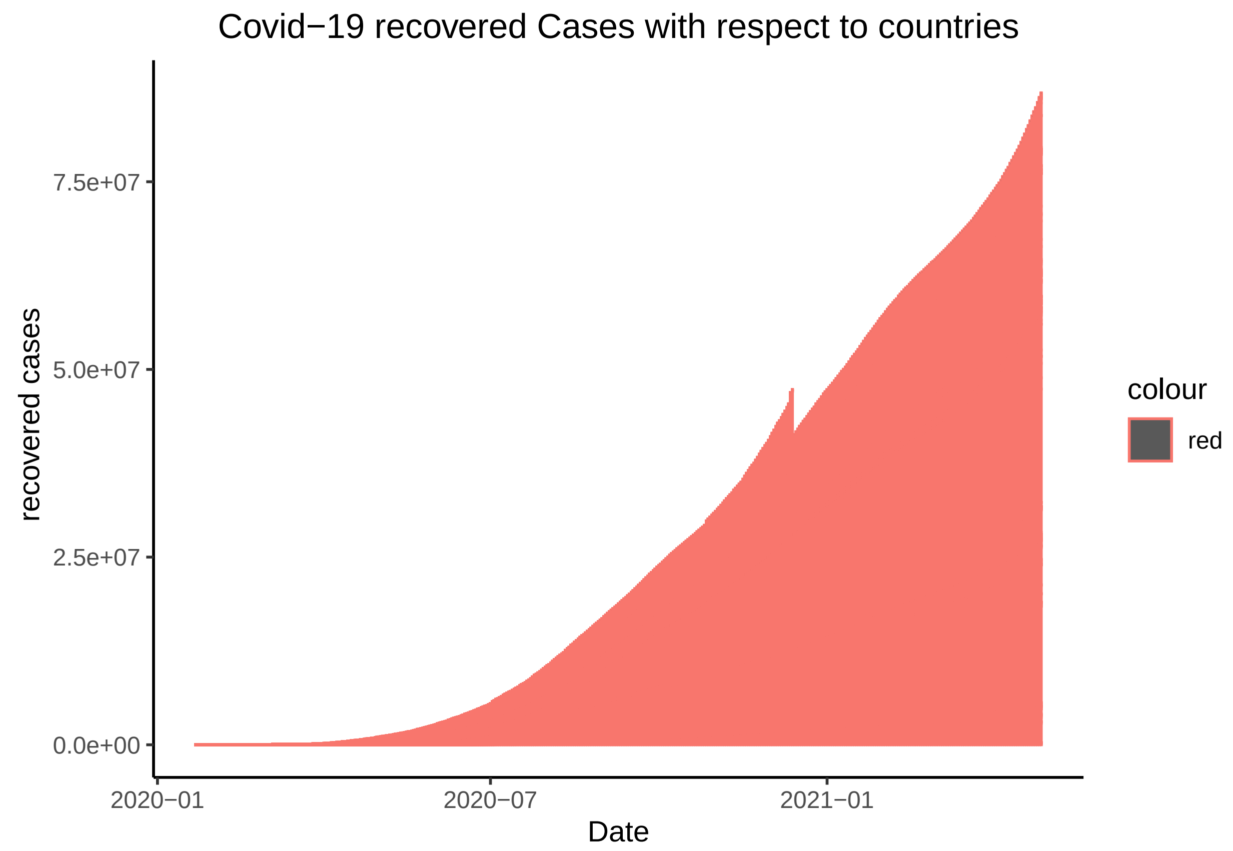
##DATASET Data sources: Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

```
## ‘summarise()‘ has grouped output by ’Country.Region’. You can override using the ‘.groups‘ argument.
## ‘summarise()‘ has grouped output by ’Country.Region’. You can override using the ‘.groups‘ argument.
## ‘summarise()‘ has grouped output by ’Country.Region’. You can override using the ‘.groups‘ argument.
```

```
## Joining, by = c("Country.Region", "date")
## Joining, by = c("Country.Region", "date")
```




Deaths due to Covid−19 with respect to countries
3e+06
2e+06
1e+06
0e+00
death cases
2020−01
2020−07
2021−01
Date
colour
black



##Coronavirus disease 2019 (COVID-19), also known as the coronavirus or COVID, is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first known case was identified in Wuhan, China, in December 2019. The disease has since spread worldwide, leading to an ongoing pandemic. Thus, It became more important to analyze the disease and understand the statistics to spread awareness among people. Today when there is no immunity against the disease, we can probably understand the statistics of how precarious the disease is and how much death it has caused and can cause in future.

##Observations: First plot of confirmed cases shows that the spread took over more that 140 million people till January 2021. The wave stated in the month of December 2019 and led to several symptoms that existed for 14 or more days followed by testing results.

Second plot shows the number of death cases due to covid 19 virus. the numbers started ranging from zero to 100, 1000 and reached up to a million by july 2020. currently number of deaths have reached upto 3 million.

The third plot shows the recovered cases. It account for people who were infected but recovered and the current number is 75 million, 2021.

```
## Registered S3 method overwritten by 'cli':
##   method     from
##   print.tree tree
```

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

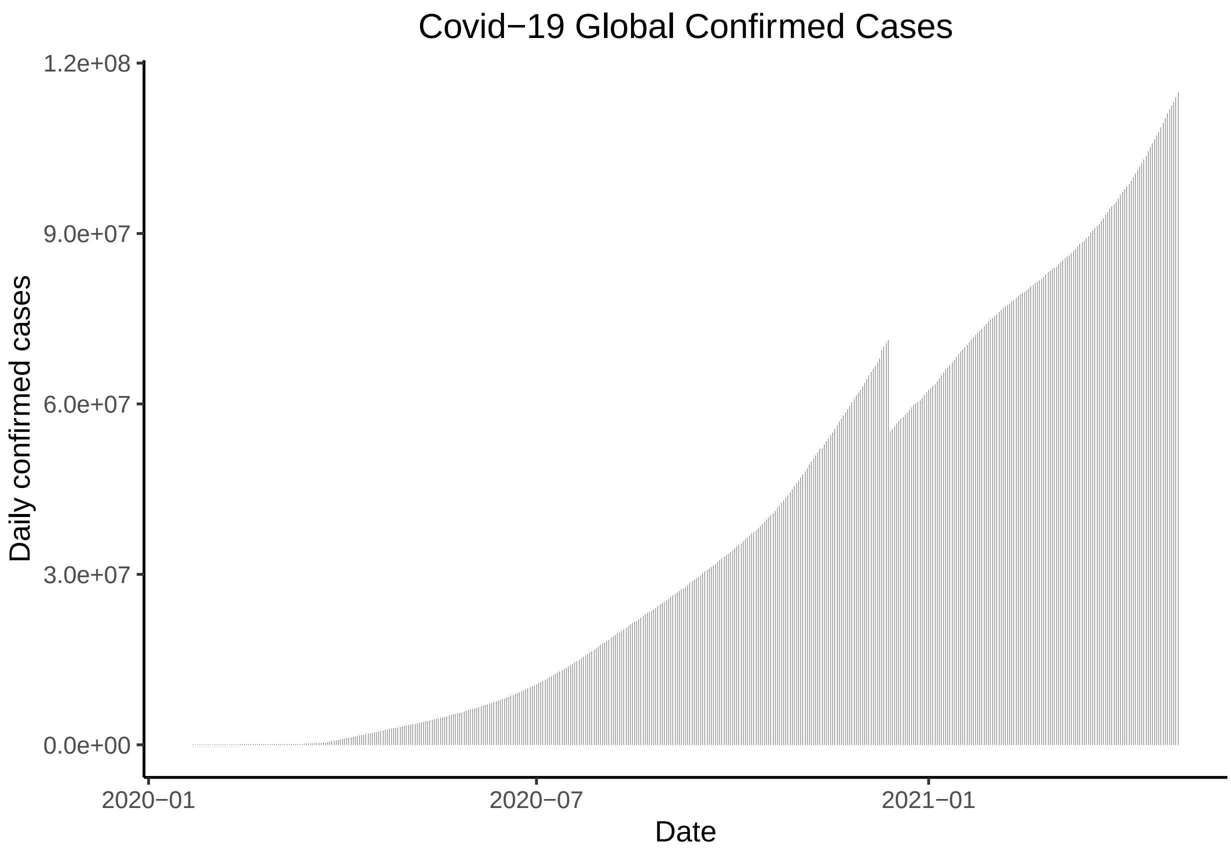
```
##  Country.Region          date                  confirmed             deaths
##  Length:414         Min.   :2020-03-11   Min.   :       62   Min.   :      1
##  Class :character   1st Qu.:2020-06-22   1st Qu.:   444207   1st Qu.: 14127
##  Mode  :character   Median :2020-10-03   Median : 6586594   Median :102234
##                     Mean   :2020-10-03   Mean   : 6000751   Mean   : 87338
##                     3rd Qu.:2021-01-14   3rd Qu.:10539052   3rd Qu.:152049
##                     Max.   :2021-04-28   Max.   :18376421   Max.   :204832
##
##    recovered         cumconfirmed          days
##  Min.   :       4   Min.   :1.029e+07   Length:414
##  1st Qu.:  250814   1st Qu.:5.703e+08   Class :difftime
##  Median : 5548334   Median :1.167e+09   Mode  :numeric
##  Mean   : 5473280   Mean   :1.135e+09
##  3rd Qu.:10175471   3rd Qu.:1.669e+09
##  Max.   :15086740   Max.   :2.138e+09
##                     NA's   :162
```

```
##  Country.Region          date                confirmed            deaths
##  Length:289         Min.   :2020-02-29   Min.   :      25   Min.   :     1
##  Class :character   1st Qu.:2020-05-11   1st Qu.: 1358293   1st Qu.: 84176
##  Mode  :character   Median :2020-07-22   Median : 3974630   Median :144171
##                     Mean   :2020-07-22   Mean   : 4832489   Mean   :141910
##                     3rd Qu.:2020-10-02   3rd Qu.: 7336043   3rd Qu.:208934
##                     Max.   :2020-12-13   Max.   :16432729   Max.   :303605
##    recovered        cumconfirmed         days
##  Min.   :      7   Min.   :7.281e+06   Length:289
##  1st Qu.: 230287   1st Qu.:7.725e+08   Class :difftime
##  Median :1210849   Median :8.216e+08   Mode  :numeric
##  Mean   :1719625   Mean   :8.008e+08
##  3rd Qu.:2873369   3rd Qu.:9.925e+08
##  Max.   :6298082   Max.   :1.397e+09
```

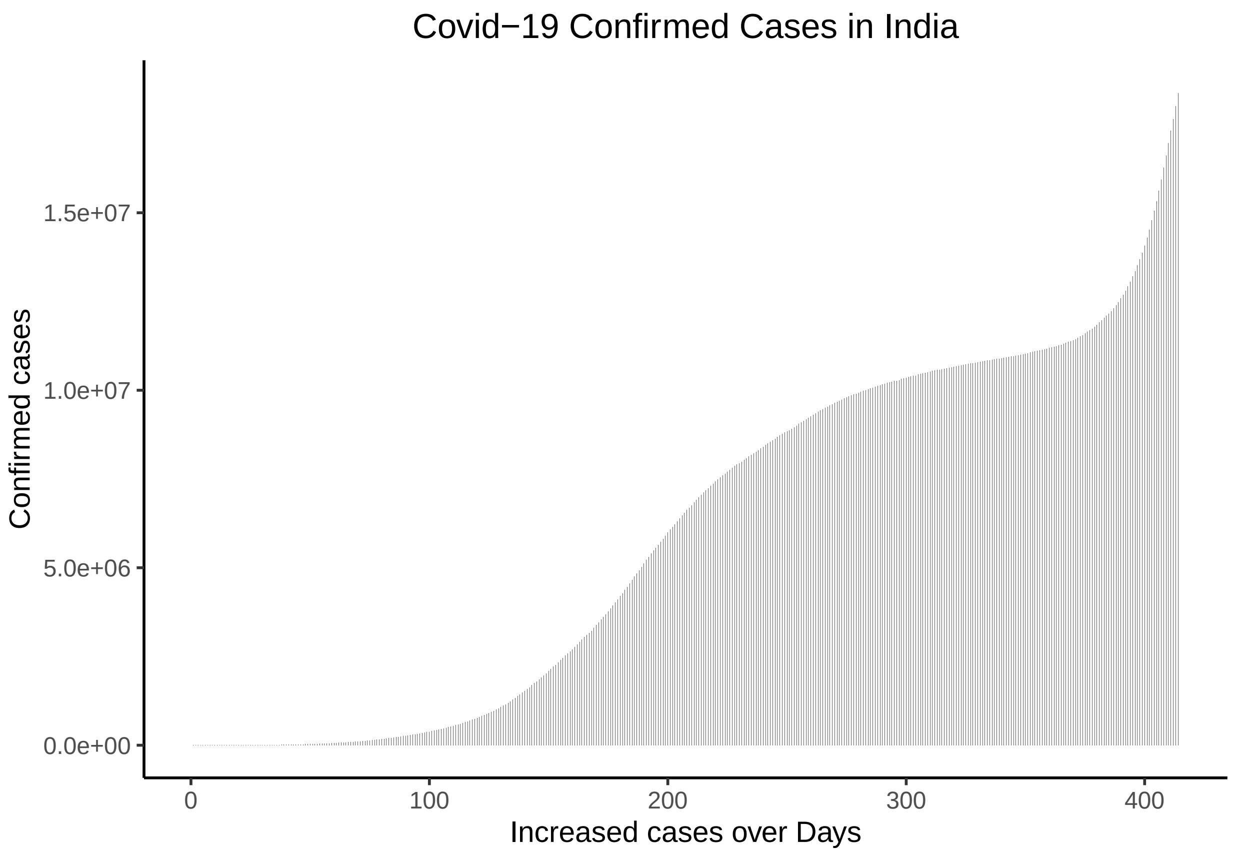
##Summary of India Mean of the confirmed cases for India was recorded as 5310060 by september 2020 while the mean of death cases was recorded as 77553 followed by recovered cases around 4850399.

##Summary of USA Mean of the confirmed cases for USA was recorded as 10770420 by september 2020 while the mean of death cases was recorded as 226267 followed by recovered cases around 1078030.

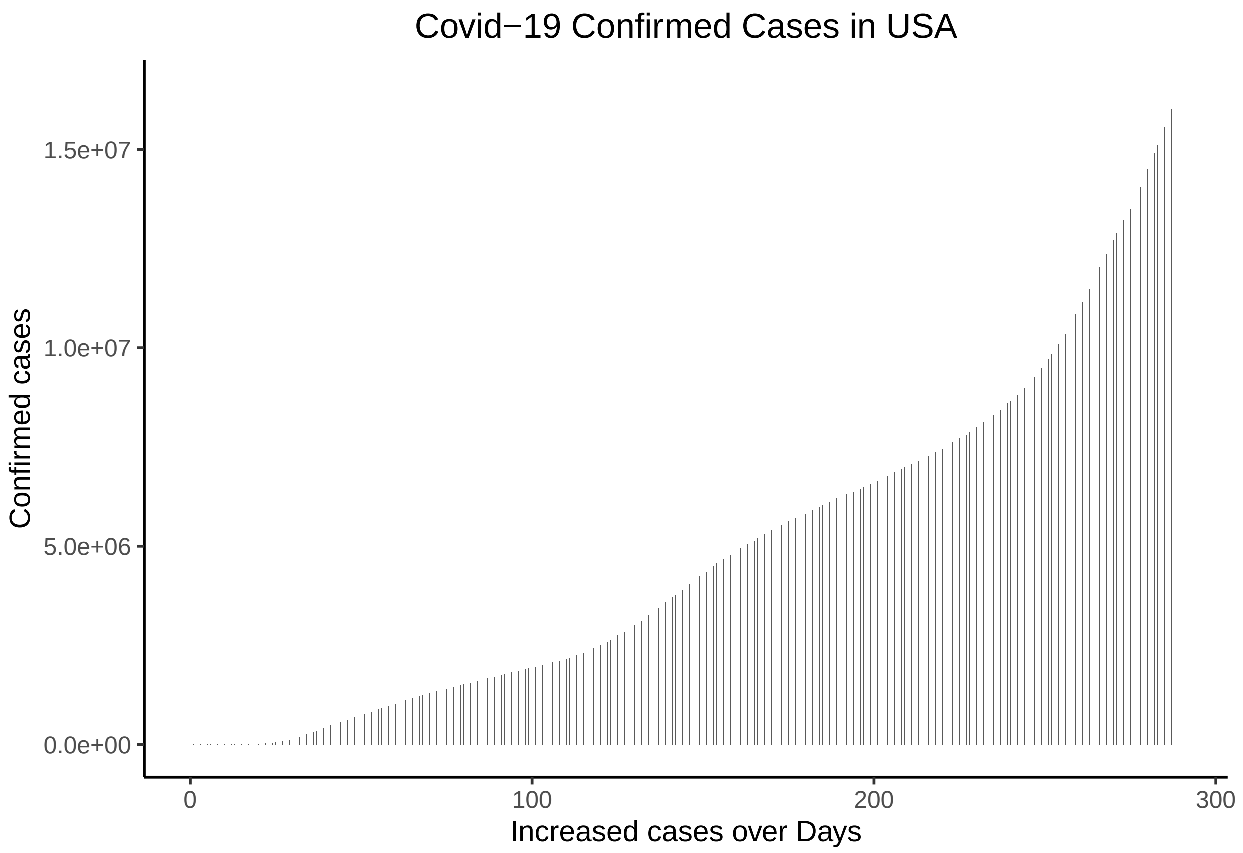
Indian population and American population has huge difference, no conclusion can be drawn with the analysis but we can get and an idea of the current numbers.

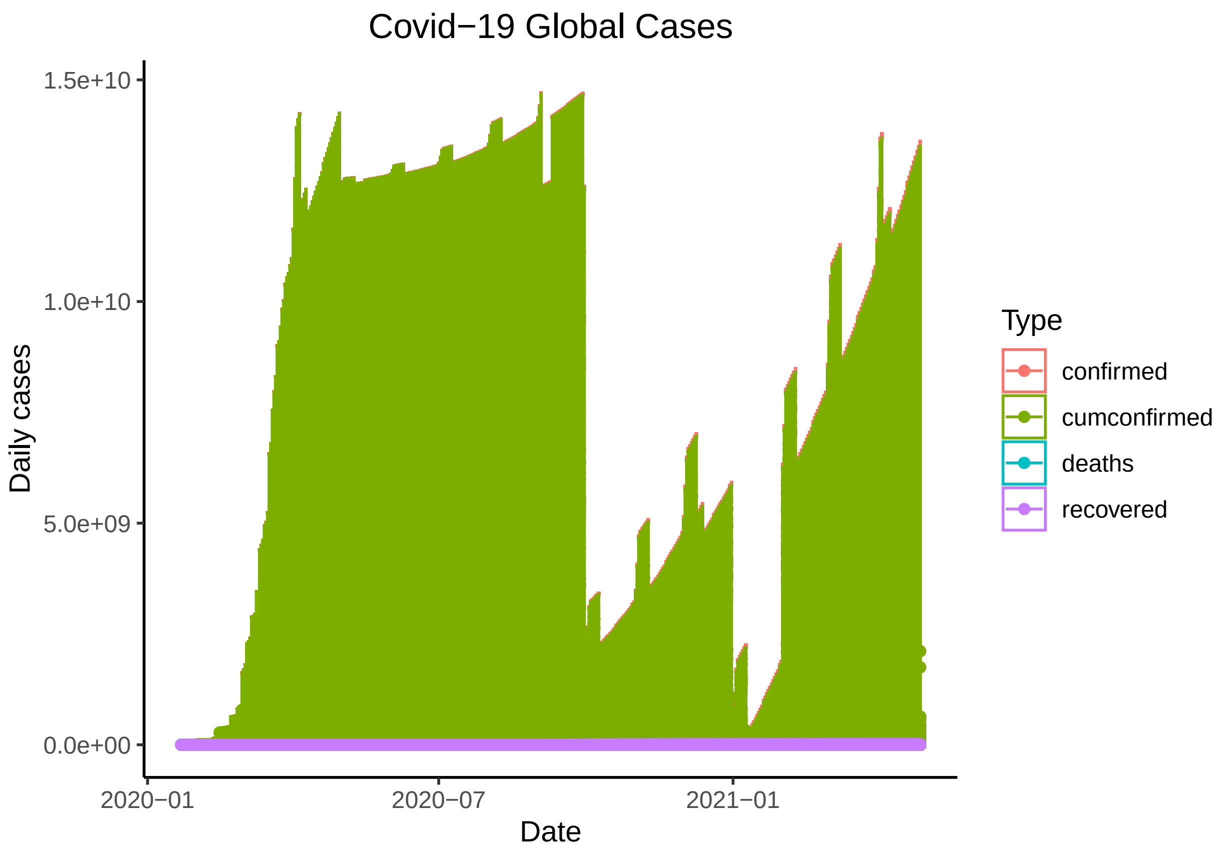


```
## Don’t know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



```
## Don’t know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```


Covid−19 Confirmed Cases in USA
Confirmed cases
1.5e+07
1.0e+07
5.0e+06
0.0e+00
0
100
200
300
Increased cases over Days



##Observations:

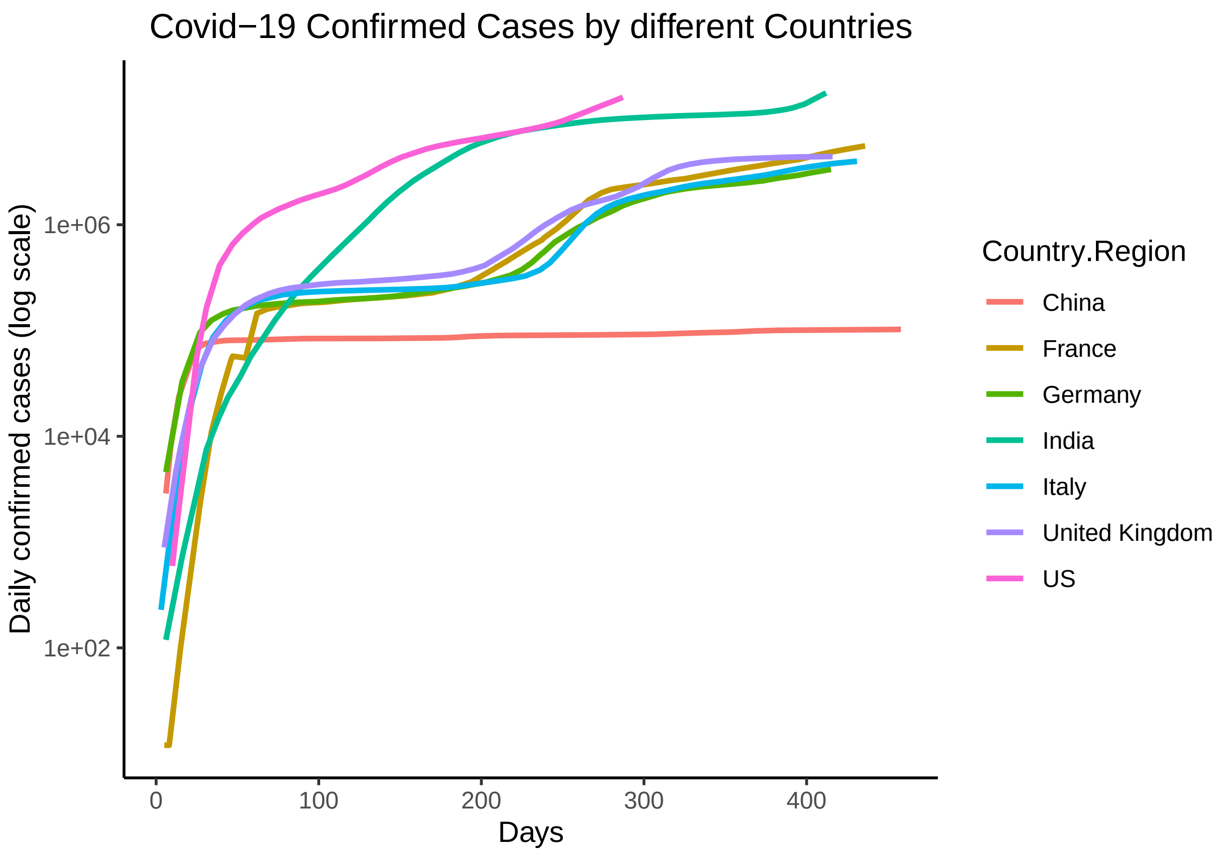
Looking at the plots for data set globally, we can make following relevant observations:

*The sum of data is growing with time. This lets us view the total contribution so far of a given measure against time.

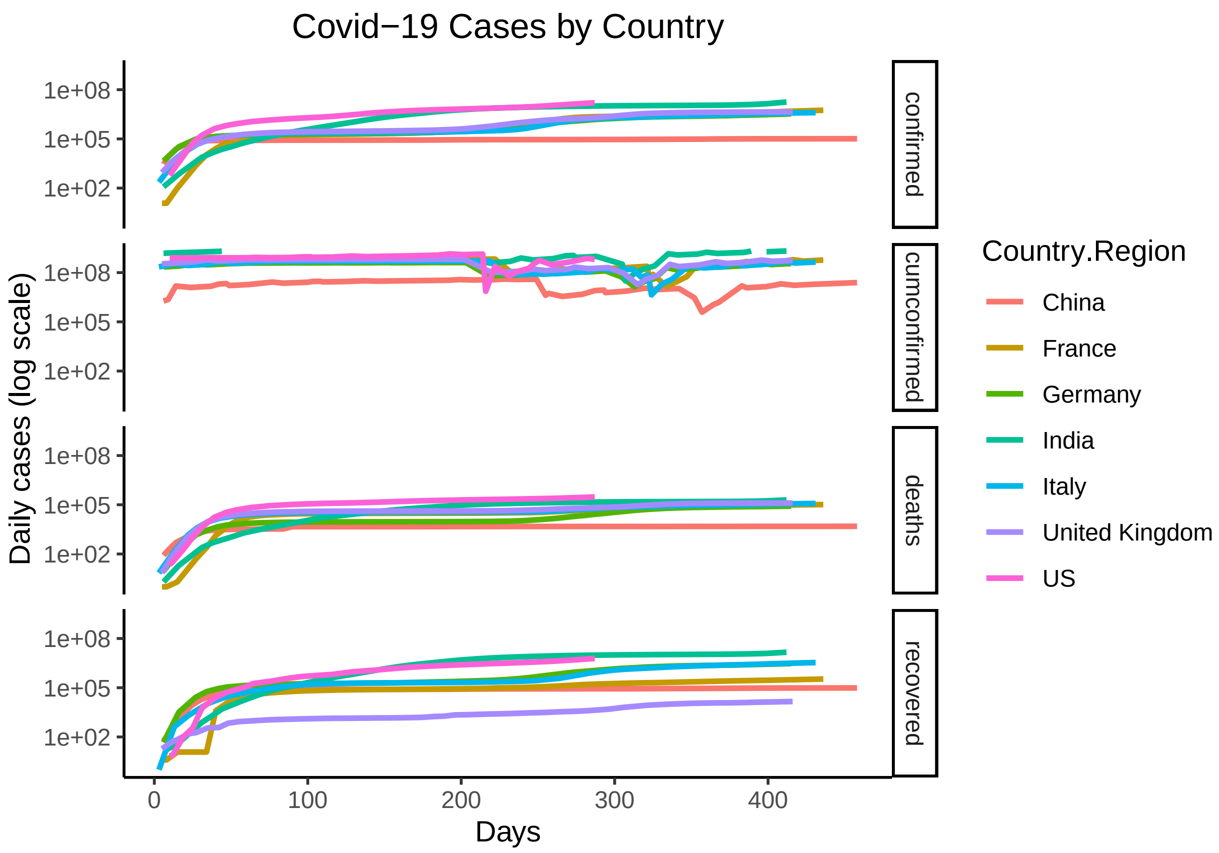
*Number figures also shows us that deaths are increasing as well, which notifies of how harmful effects of the virus is.

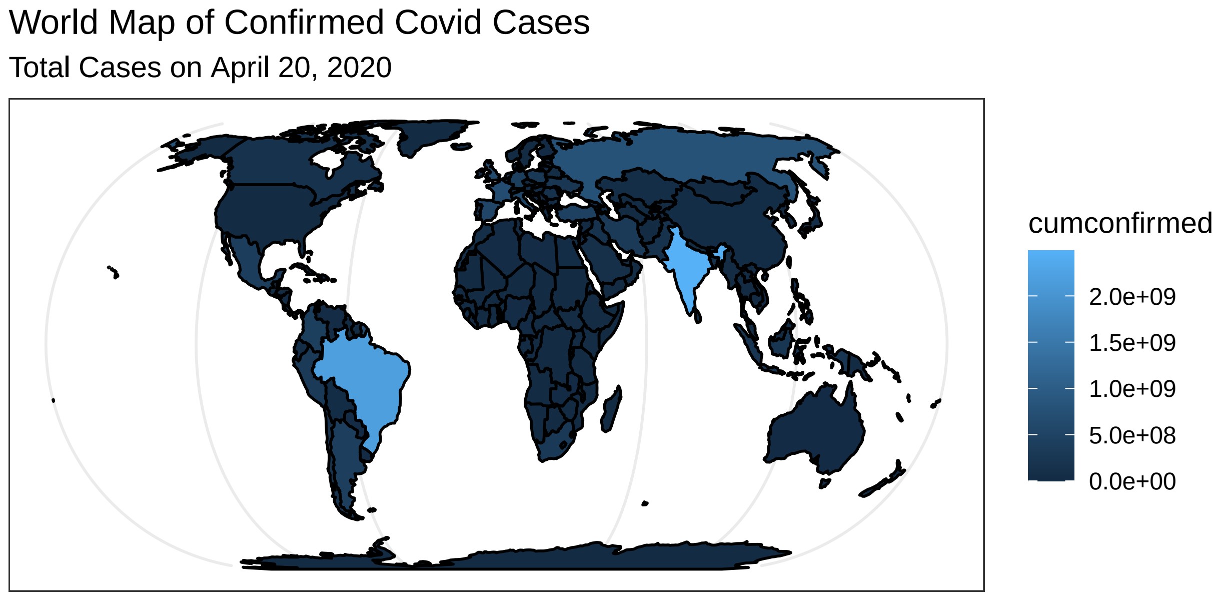
*Though we have recovered case numbers yet it does not rule out the death numbers and demands people to follow safety measures to cope up with the pandemic.

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```



```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```


Covid−19 Cases by Country
Daily cases (log scale)
1e+08
1e+05
1e+02
confirmed
cumconfirmed
deaths
recovered
0
100
200
300
400
Days
Country.Region
China
France
Germany
India
Italy
United Kingdom
US


World Map of Confirmed Covid Cases
Total Cases on April 20, 2020
cumconfirmed
2.0e+09
1.5e+09
1.0e+09
5.0e+08
0.0e+00

First Graph shows us the numbers from different countries like: China, France, Germany, India, Italy, UK, and US.

Observations:

The disease is wide spreading, though it started from china, but china is stabilizing the numbers by taking strict measures. whereas countries like India and US still have no control over the disease and is taking over uninfected people thus increasing the numbers. Other countries, France, Italy UK, Germany showing increasing number but little less than numbers of India and US.

Second Graph shows us collaborative reports for confirmed, deaths, recovered and cumulative confirmed cases for countries specified for graph 1.

Observations:

The most promising results in the graph for all the 3 cases is for China, even though the country led to spread of this deadly virus, it played better role in controlling the infection and bending down the death rates as well. Another promising result is the recovered cases from India, even though the cumulative confirmed cases are higher but some how the virus is leading to less deaths compared to other countries. Reason may be better immunity or health care infrastructure (It's just speculated idea). Hence India shows highest number of recovered cases. In addition, US shows least number in the recovered cases, where as the confirmed cases are increasing by each passing days, even the death cases have higher number compared to other countries.

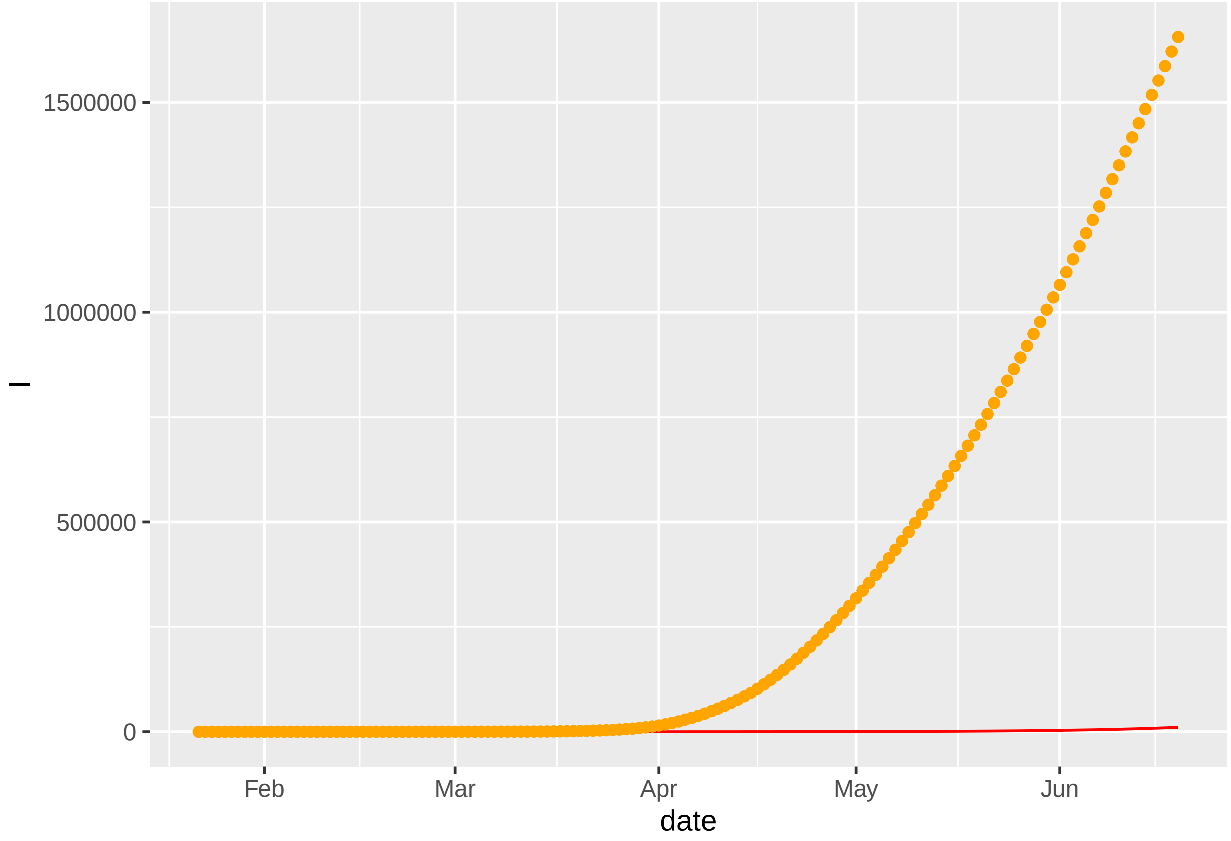
Third Graph is a World map which shows cumulative confirmed cases all over the world

Observations:

It shows the numbers in shades of blue and has recognized least to increasing number of cases. Brazil covers the lightest shade of blue with most of the cumulative confirmed cases.

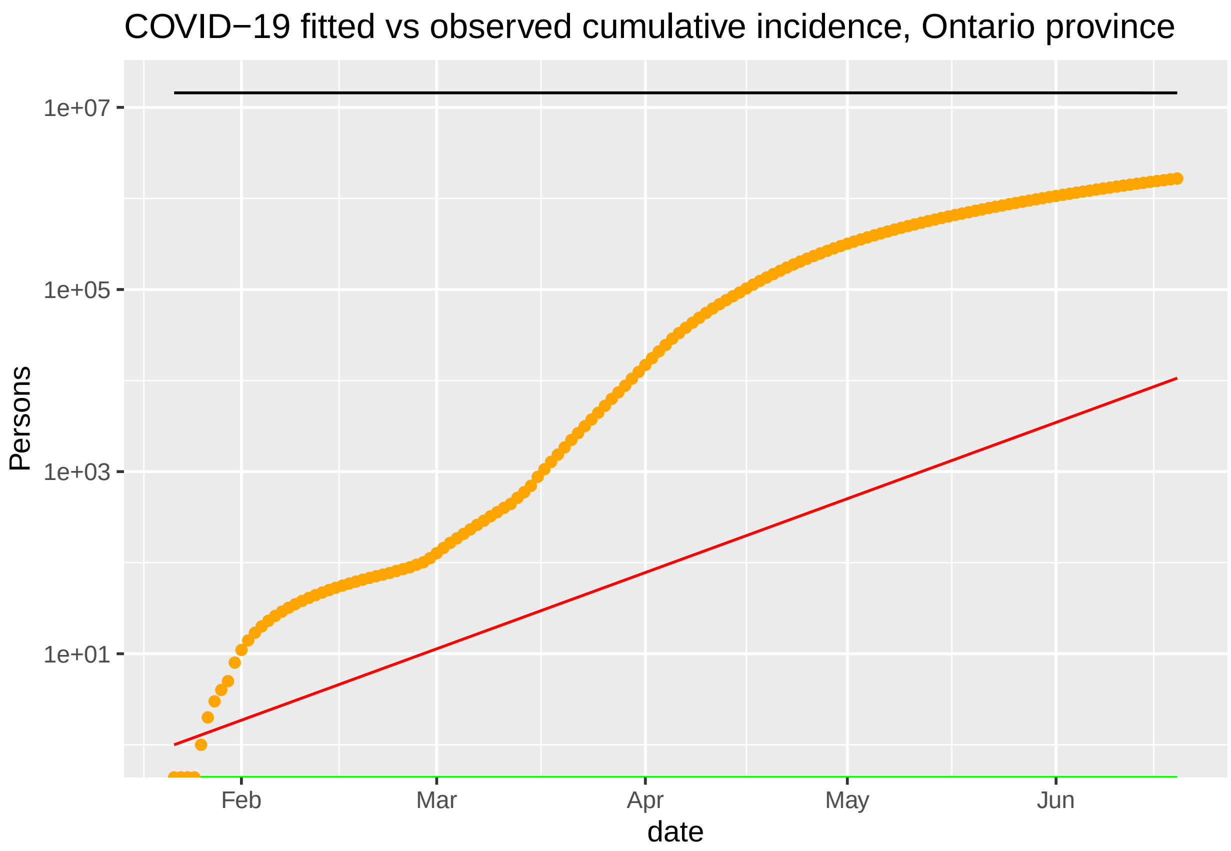
```
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
##       beta      gamma
## 0.06223188 0.00000000
```

```
## beta
##  Inf
```



```
## $y
## [1] "Cumulative incidence"
##
## $x
## [1] "Date"
##
## $title
## [1] "COVID-19 fitted vs observed cumulative incidence, Ontario"
##
## $subtitle
## [1] "(red=fitted incidence from SIR model, orange=observed incidence)"
##
## attr(,"class")
## [1] "labels"
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```



For the first graph, I plotted infected(red line) along with cumulative confirmed cases(orange line)

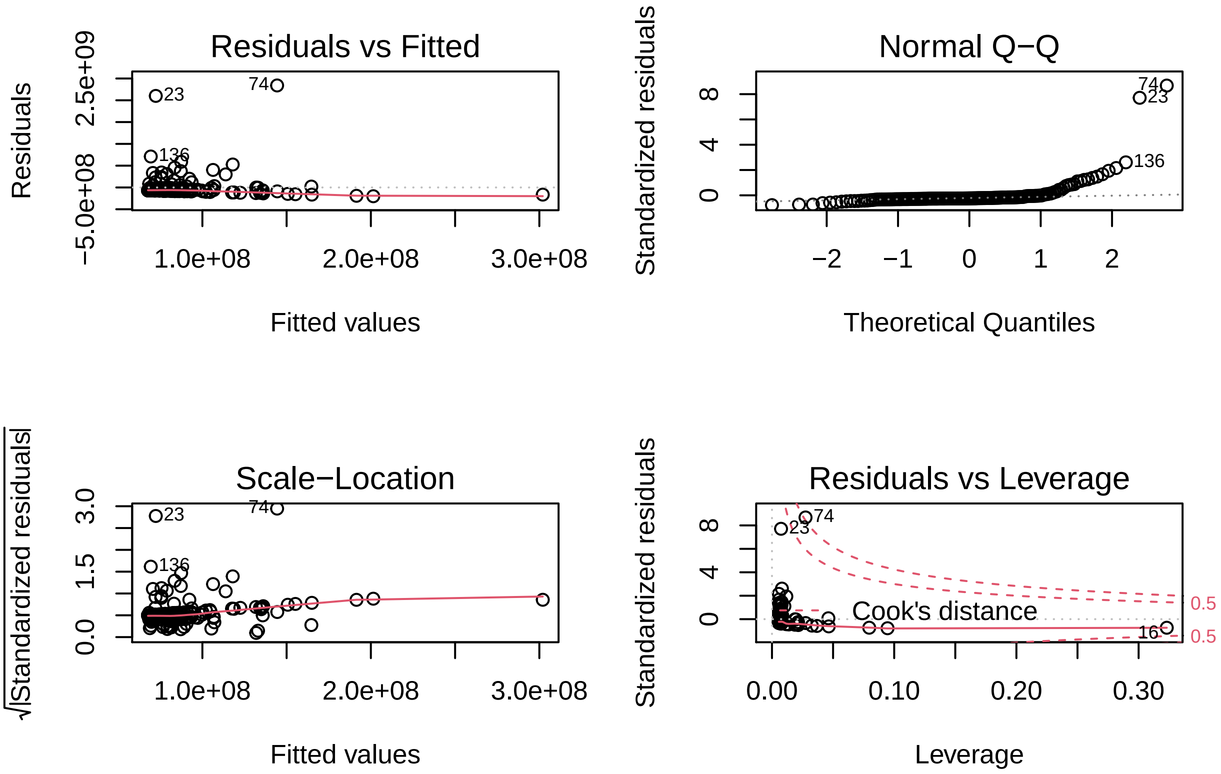
In general infection leads to confirmed cases hence the number should have gone higher for infected as well, but my model does not perform better in fetching results for infected numbers(red line). I plan on improving results with better approach.

Second graph shows the SIR graph along with cumulative confirmed numbers The graph does show rising infection numbers but shows flat reading for both susceptible and recovered, which mean the infection is rising and the whole population is susceptible but there is no recovery rate, this output is not acceptable, considering the current scenario. This prediction model is basically not accurate as it should show figures and numbers based on current scenario, where people are getting infected and few have been recovered too along with few deaths.

```
##
## Call:
## lm(formula = cumconfirmed ~ pop_est_dens, data = worldmap)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -201449565  -75565843  -65542519  -41397540 2339943691
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  67706569   25368136   2.669  0.00832 **
## pop_est_dens   195464     140898   1.387  0.16712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

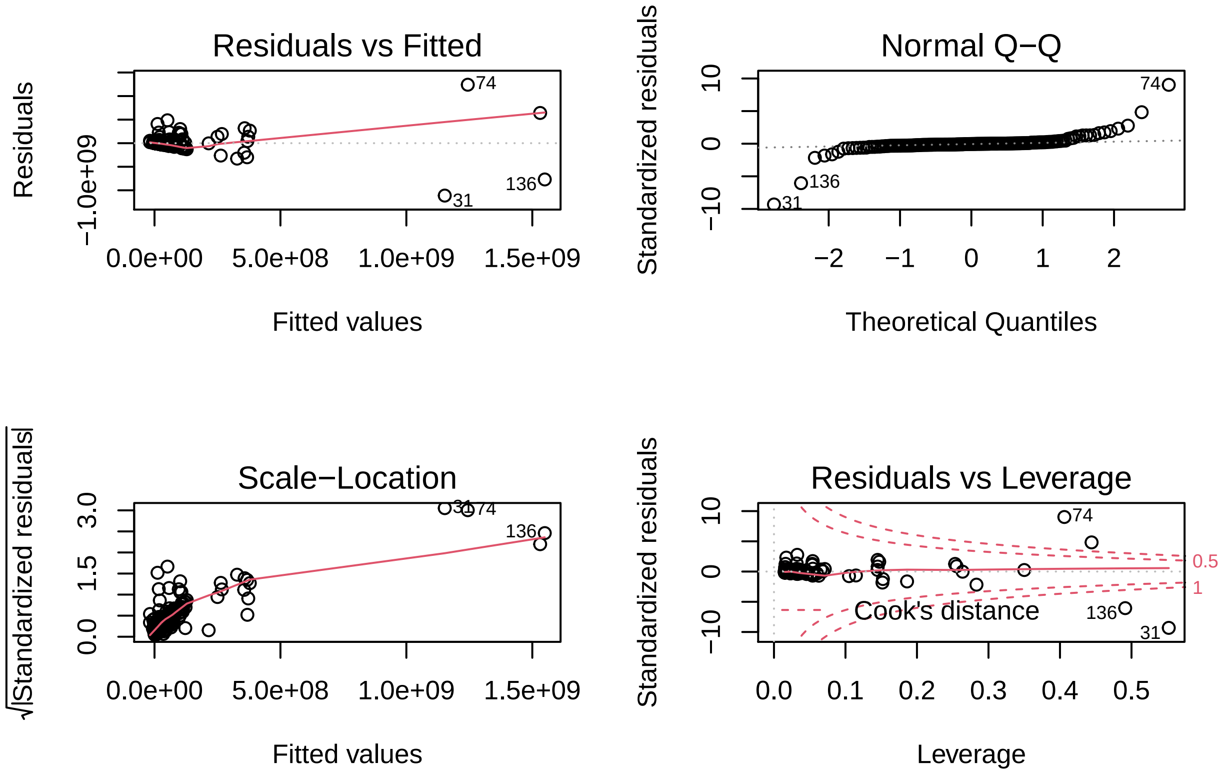
```
## Residual standard error: 273400000 on 175 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.005226
## F-statistic: 1.925 on 1 and 175 DF,  p-value: 0.1671
```

```
## # A tibble: 2 x 7
##   term          estimate std_error statistic p_value  lower_ci   upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>     <dbl>      <dbl>
## 1 intercept    67706569. 25368136.      2.67   0.008 17639699. 117773439.
## 2 pop_est_dens   195464.   140898.      1.39   0.167   -82614.    473542.
```



```
##
## Call:
## lm(formula = cumconfirmed ~ pop_est_dens + pop_est + economy,
##     data = worldmap)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.112e+09 -3.275e+07 -7.905e+06  1.113e+07  1.240e+09
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.672e+08  7.050e+07   5.208 5.54e-07 ***
## pop_est_dens                  1.275e+05  9.481e+04   1.345  0.18033
## pop_est                      -3.456e-01  1.562e-01  -2.213  0.02827 *
## economy2. Developed region: nonG7 -3.120e+08  7.588e+07  -4.112 6.13e-05 ***
## economy3. Emerging region: BRIC    1.230e+09  1.467e+08   8.382 2.01e-14 ***
## economy4. Emerging region: MIKT   -8.595e+07  1.118e+08  -0.769  0.44324
## economy5. Emerging region: G20    -2.530e+08  7.952e+07  -3.182  0.00174 **
## economy6. Developing region       -3.445e+08  7.256e+07  -4.748 4.38e-06 ***
## economy7. Least developed region  -3.653e+08  7.380e+07  -4.950 1.79e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 178400000 on 168 degrees of freedom
## Multiple R-squared:  0.5959, Adjusted R-squared:  0.5767
## F-statistic: 30.97 on 8 and 168 DF,  p-value: < 2.2e-16
```

```
## # A tibble: 9 x 7
##   term                    estimate   std_error statistic p_value lower_ci upper_ci
##   <chr>                      <dbl>       <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                3.67e+8     7.05e+7     5.21    0        2.28e+8  5.06e+8
## 2 pop_est_dens             1.28e+5     9.48e+4     1.34    0.18    -5.96e+4  3.15e+5
## 3 pop_est                 -3.46e-1     1.56e-1    -2.21    0.028   -6.54e-1 -3.7 e-2
## 4 economy2. Developed ~ -3.12e+8     7.59e+7    -4.11    0       -4.62e+8 -1.62e+8
## 5 economy3. Emerging r~  1.23e+9     1.47e+8     8.38    0        9.40e+8  1.52e+9
## 6 economy4. Emerging r~ -8.60e+7     1.12e+8    -0.769   0.443   -3.07e+8  1.35e+8
## 7 economy5. Emerging r~ -2.53e+8     7.95e+7    -3.18    0.002   -4.10e+8 -9.60e+7
## 8 economy6. Developing~ -3.44e+8     7.26e+7    -4.75    0       -4.88e+8 -2.01e+8
## 9 economy7. Least deve~ -3.65e+8     7.38e+7    -4.95    0       -5.11e+8 -2.20e+8
```



Linear Model The Pr(>t) acronym found in the model output relates to the probability of observing any value equal or larger than t. A small p-value indicates that it is unlikely we will observe a relationship between the predictor (pop_est_dens) and response (cumconfirmed) variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. In our model, the p-values are very close to zero. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between pop_est_dens and cumconfirmed cases.

Multiple Linear Regression Model In multiple linear regression, the R2 represents the correlation coefficient between the observed values of the cumconfirmed and the fitted i.e.pop_est_dens + pop_est + economy values of cumconfirmed, For this reason, the value of R will always be positive and will range from zero to one. A problem with the R2, is that, it will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. A solution is to adjust the R2 by taking into account the number of predictor variables.

The adjustment in the "Adjusted R Square" value in the summary output is a correction for the number of predictor variables included in the prediction model. The lower the RSE, the more accurate the model, In our case RSE is 0.4902.

A small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between pop_est_dens + pop_est + economy and cumconfirmed cases.

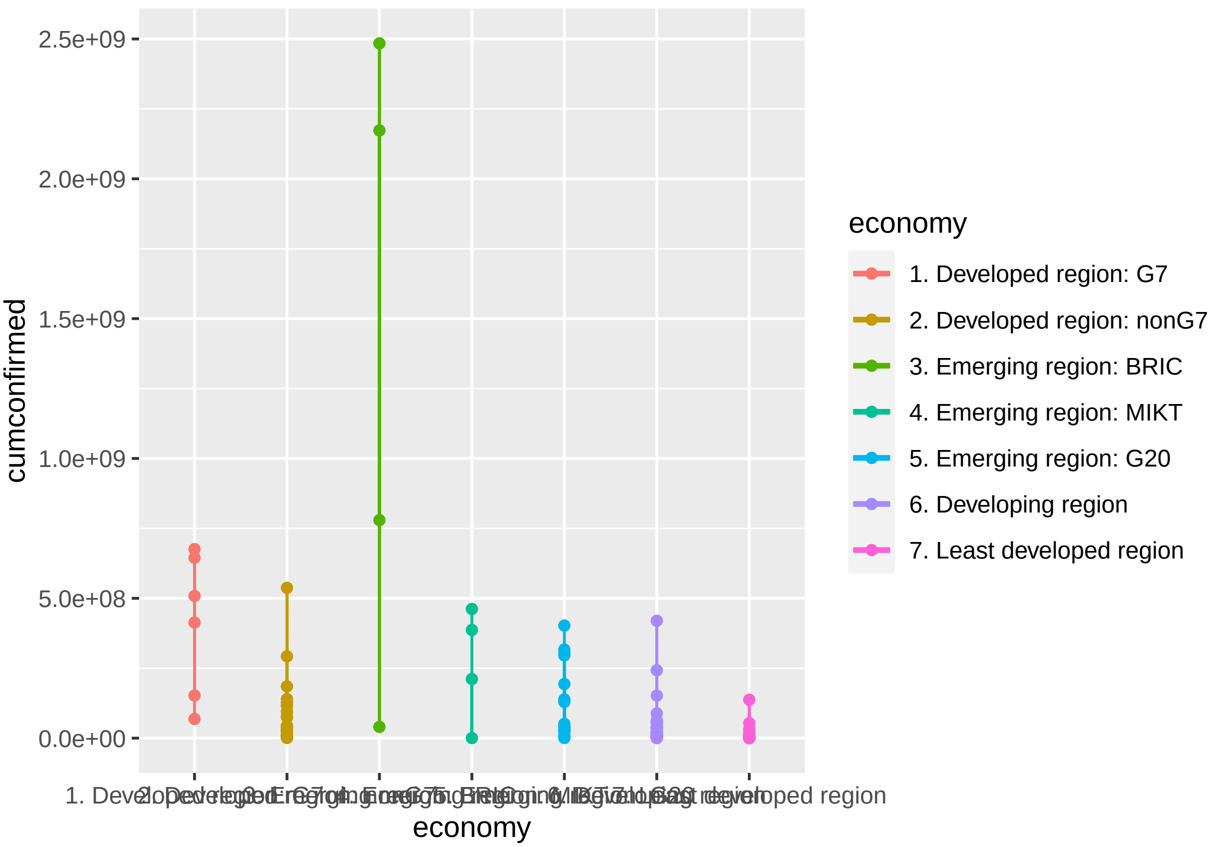
Plot1 The results show a strong evidence that estimated population density is the major influencing cumulative confirmed cases.

Shown residual plots also adds strength for the same. There are few outliers found while performing the linear fit.

Approximately 90+% of variation in value variable can be explained by this model with these two independent variables (pop_est_dens and cumconfirmed). Very low P-value also strengthens the assumption. The residual standard error also shows there is not much distance between our observed value from the predicted Value. We can see from the plots that there exist leverage points but, they are not much influential.

Plot2 The results show a strong evidence that estimated population density, total population estimate and economy collectively are the majorly influencing cumulative confirmed cases.

```
## `geom_smooth()` using formula 'y ~ x'
```



```
## # A tibble: 7 x 7
##   term                   estimate std_error statistic p_value  lower_ci  upper_ci
##   <chr>                     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>     <dbl>
## 1 intercept                4.11e8    8.43e7      4.87   0        2.44e8    5.78e8
## 2 economy2. Develope~     -3.36e8    9.32e7     -3.60   0       -5.20e8   -1.51e8
## 3 economy3. Emerging~      9.59e8    1.33e8      7.19   0        6.95e8    1.22e9
## 4 economy4. Emerging~     -1.46e8    1.33e8     -1.09   0.277   -4.09e8    1.18e8
## 5 economy5. Emerging~     -2.97e8    9.73e7     -3.06   0.003   -4.90e8   -1.05e8
## 6 economy6. Developi~     -3.71e8    9.00e7     -4.12   0       -5.49e8   -1.93e8
## 7 economy7. Least de~     -4.00e8    9.23e7     -4.33   0       -5.83e8   -2.17e8
```

```
## # A tibble: 2 x 7
##   term          estimate std_error statistic p_value  lower_ci   upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>     <dbl>      <dbl>
## 1 intercept    88326971. 34525232.      2.56   0.012 20022929. 156631014.
## 2 pop_est_dens   231102.   187721.      1.23   0.221  -140283.    602486.
```

Plot x: economy distribution and y: cumulative confirmed cases

Well, the plot does not explains the relation between the spread of virus among people belonging to developing or emerging economy. It shows variable results. We would agree with confirmed numbers shown for emerging region: MIKT for probable reasons like, they may not have better access to infrastructures or may have to meet people in order to get their work done which would have led to transmission of virus or many more. But if we consider, Least developed region, they have the least number of cumulative confirmed cases, which is great considering their economical conditions. In contrast to above note, even developed region has second highest number of cases.

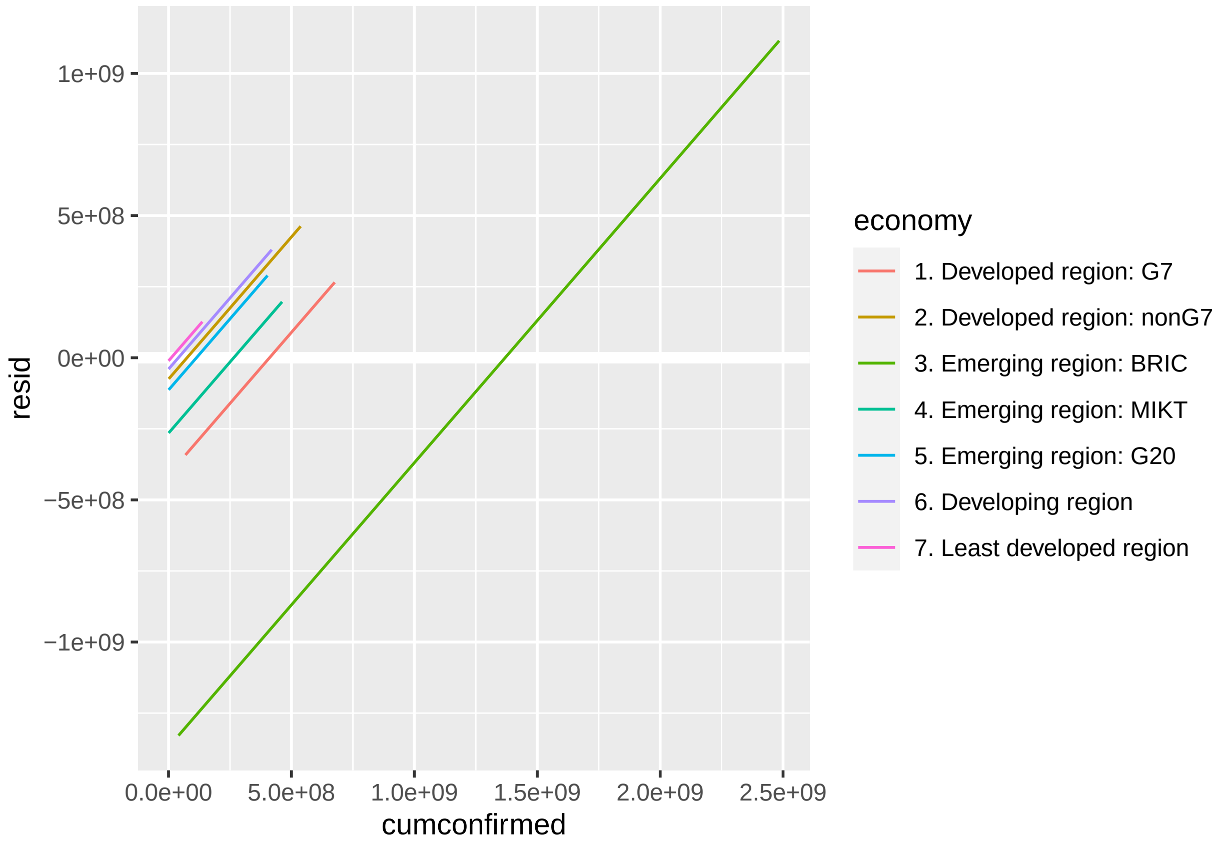
This this graph does not conclude any thing in particular.

Linear Model, Predictor- Economy

To understand the relationship better, we did explore the linear model with economy as predictor and cumconfirmed as target, and to our surprise the model does work better. the P value and R squared both are close to 0, which explains the assumption.

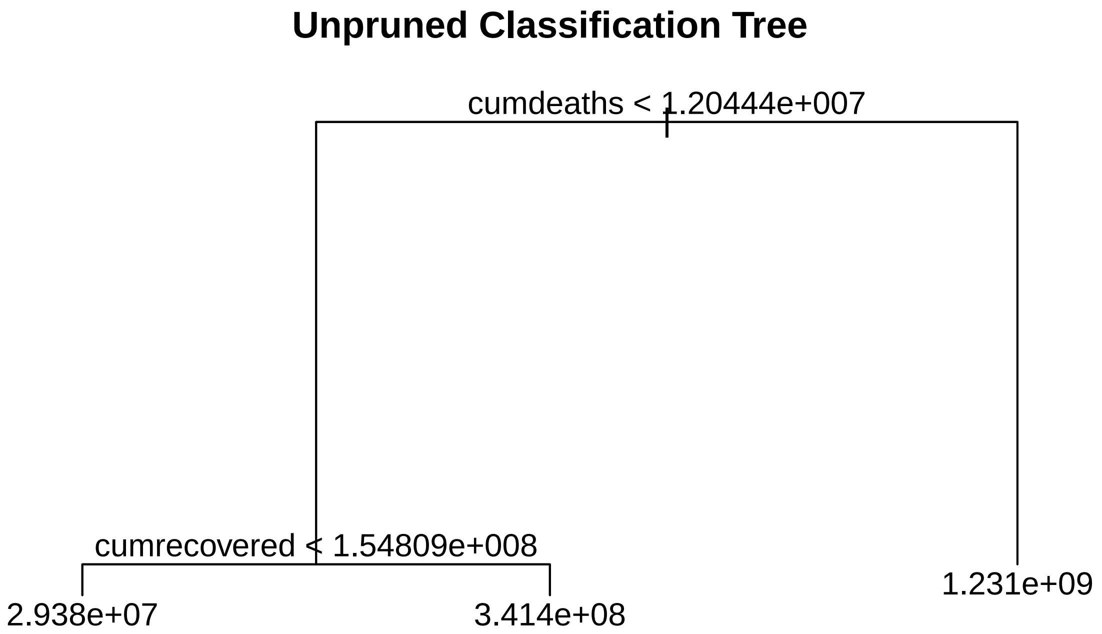
Linear Model, Predictor- population density estimate

I had an inclination towards this particular model, as I believed the spread will increase if people assemble in closed area or are around others(Just an idea which is far away from the concept of population density but does explains the concept of density in a closed area). And this model explains it all, it gives us p-value: 0.2883 and Adjusted R-squared: 0.001011. It does explain that our model fits with the data provided and spread of the disease can be related to higher population density estimated.



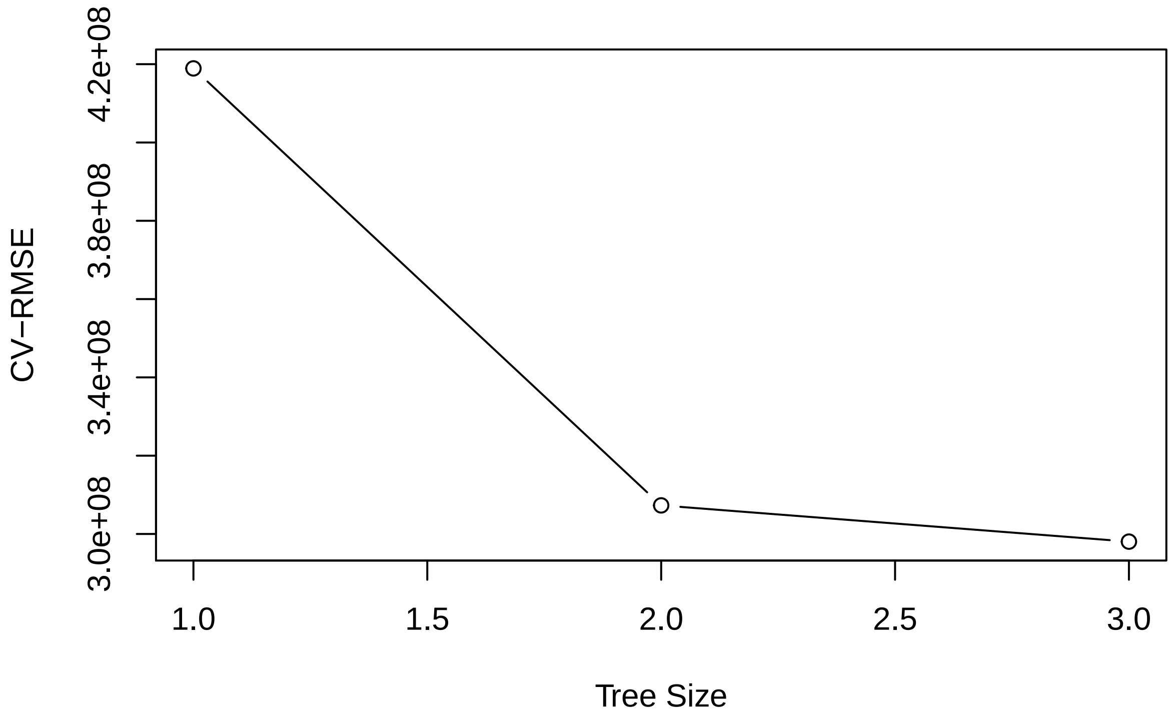
Residuals: The basic idea behind finding residuals is that we can get the difference between predicted value and the measured values and find the line that fits our data set.

```
##
## Regression tree:
## tree(formula = cumconfirmed ~ ., data = worldmap_b_no_geometry_trn)
## Variables actually used in tree construction:
## [1] "cumdeaths"    "cumrecovered"
## Number of terminal nodes:  3
## Residual mean deviance:  6.738e+16 = 4.245e+18 / 63
## Distribution of residuals:
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -914500000  -27650000  -19490000          0   10610000 1253000000
```



##Regression Trees analysis Performed another type of statistical model called Regression Trees, because it has a good graphical representation. *Basic Regression trees Covid 19 data set.* The regression tree building methodology allows input variables to be a mixture of continuous and categorical variables. *A decision tree is generated when each decision node in the tree contains a test on some input variable's value.* The terminal nodes of the tree contain the predicted output variable values.

As with classification trees, we can use cross-validation to select a good pruning of the tree.

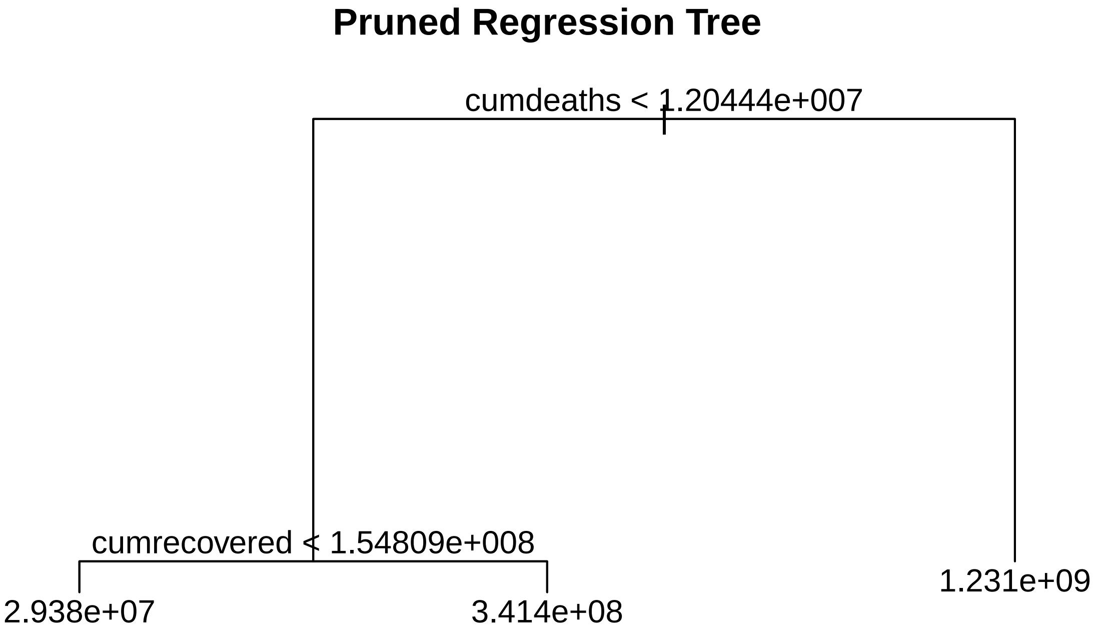


*Cross-validation refers a methods for measuring the performance of a given predictive model on new test data sets.

The basic idea, behind cross-validation techniques, consists of dividing the data into two sets: 1.The training set, used to train (i.e. build) the model; 2. The testing set (or validation set), used to test (i.e. validate) the model by estimating the prediction error.

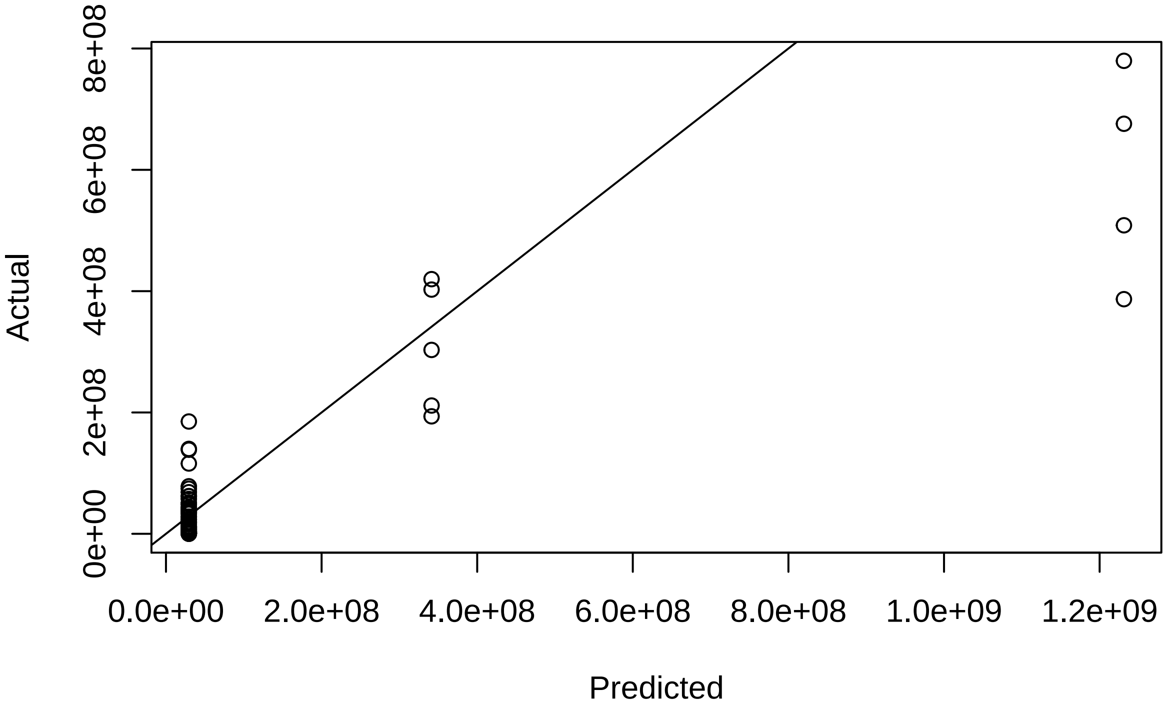
While the tree of size 380000000 does have the lowest RMSE, we'll prune to perform just as well. The pruned tree is, as expected, smaller and easier to interpret.

```
##
## Regression tree:
## tree(formula = cumconfirmed ~ ., data = worldmap_b_no_geometry_trn)
## Variables actually used in tree construction:
## [1] "cumdeaths"    "cumrecovered"
## Number of terminal nodes:  3
## Residual mean deviance:  6.738e+16 = 4.245e+18 / 63
## Distribution of residuals:
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -914500000  -27650000  -19490000          0   10610000 1253000000
```

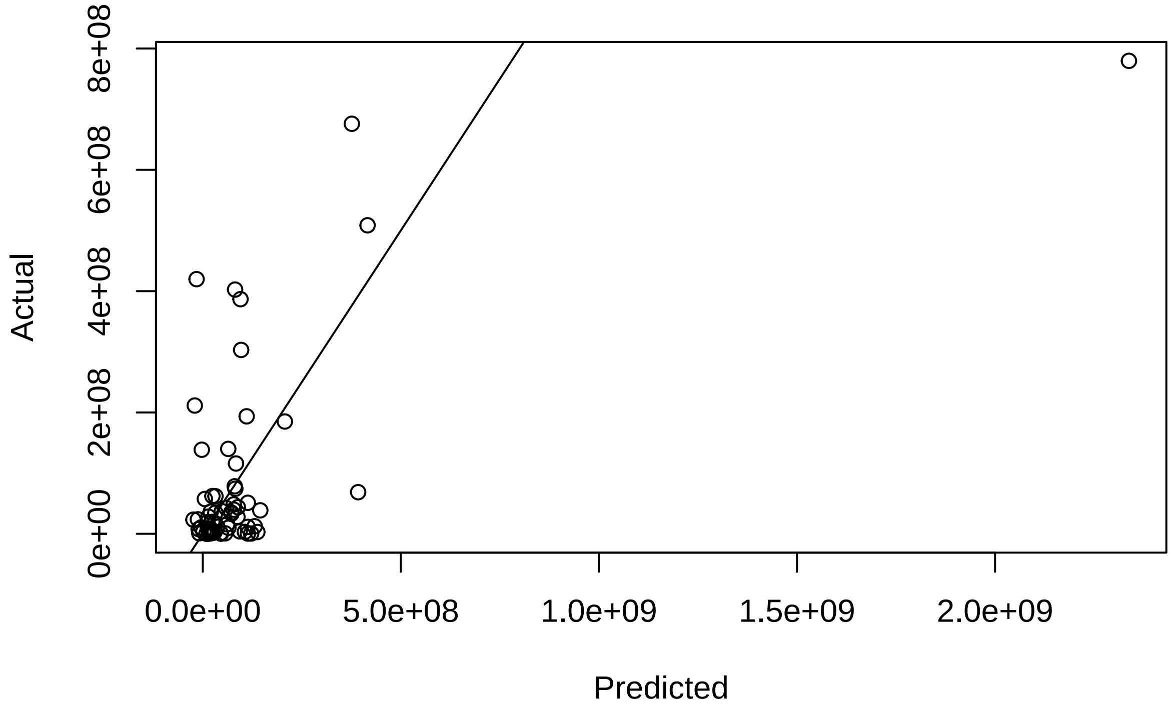


Regressing tree does not provide us better model view to understand the distribution of data, let's compare this regression tree to an additive linear model and use RMSE as our metric.

We obtain predictions on the train and test sets from the pruned tree. We also plot actual vs predicted. This plot may look odd. We'll compare it to a plot for linear regression below.



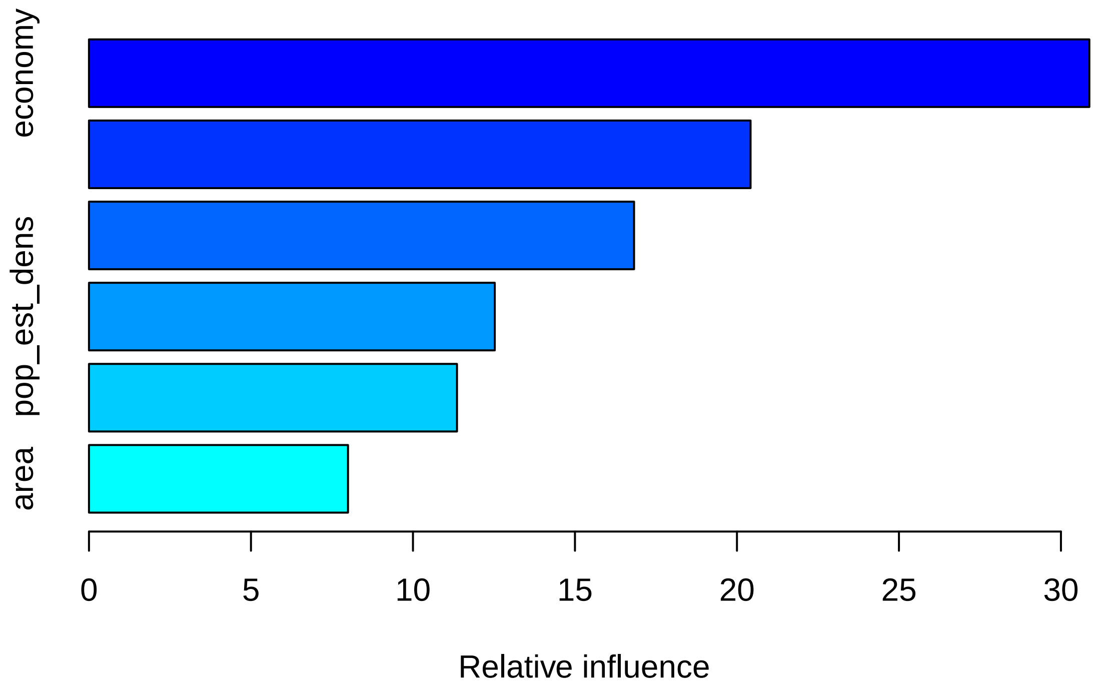
Here, using an additive linear regression the actual vs predicted looks much more like what we are used to.



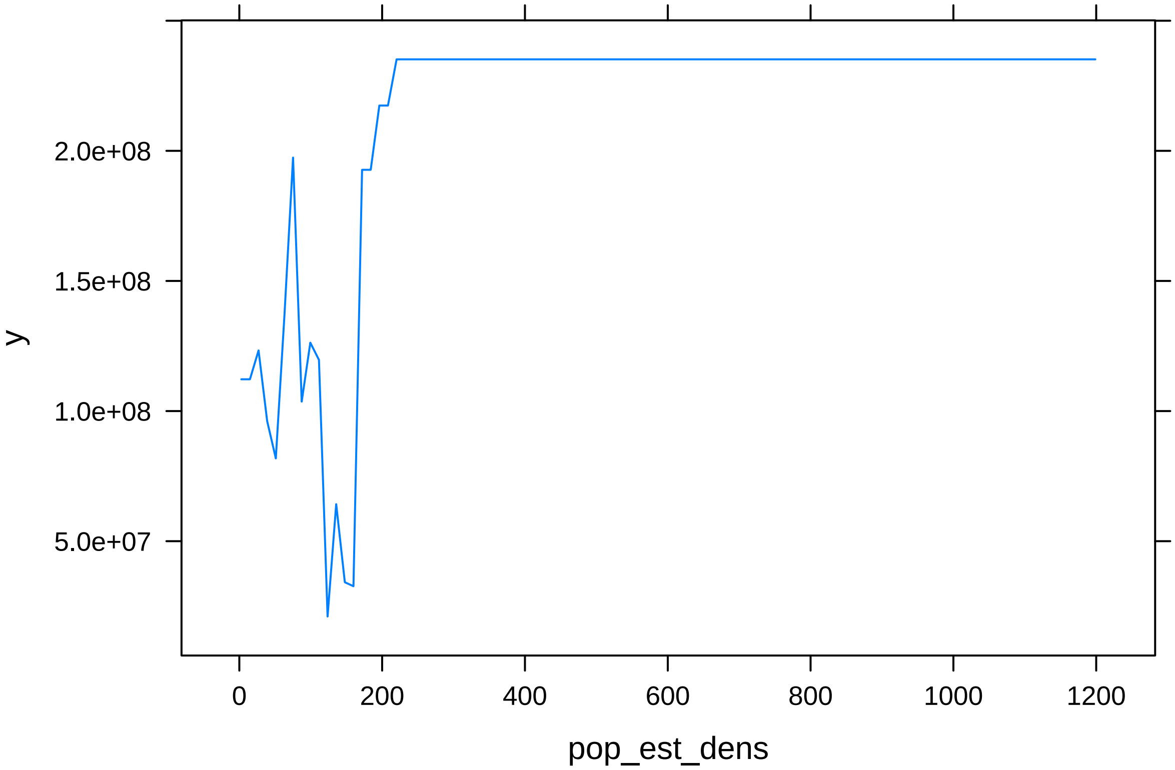
The tree after pruning seems to be easier to interpret and has a better graphical representation.

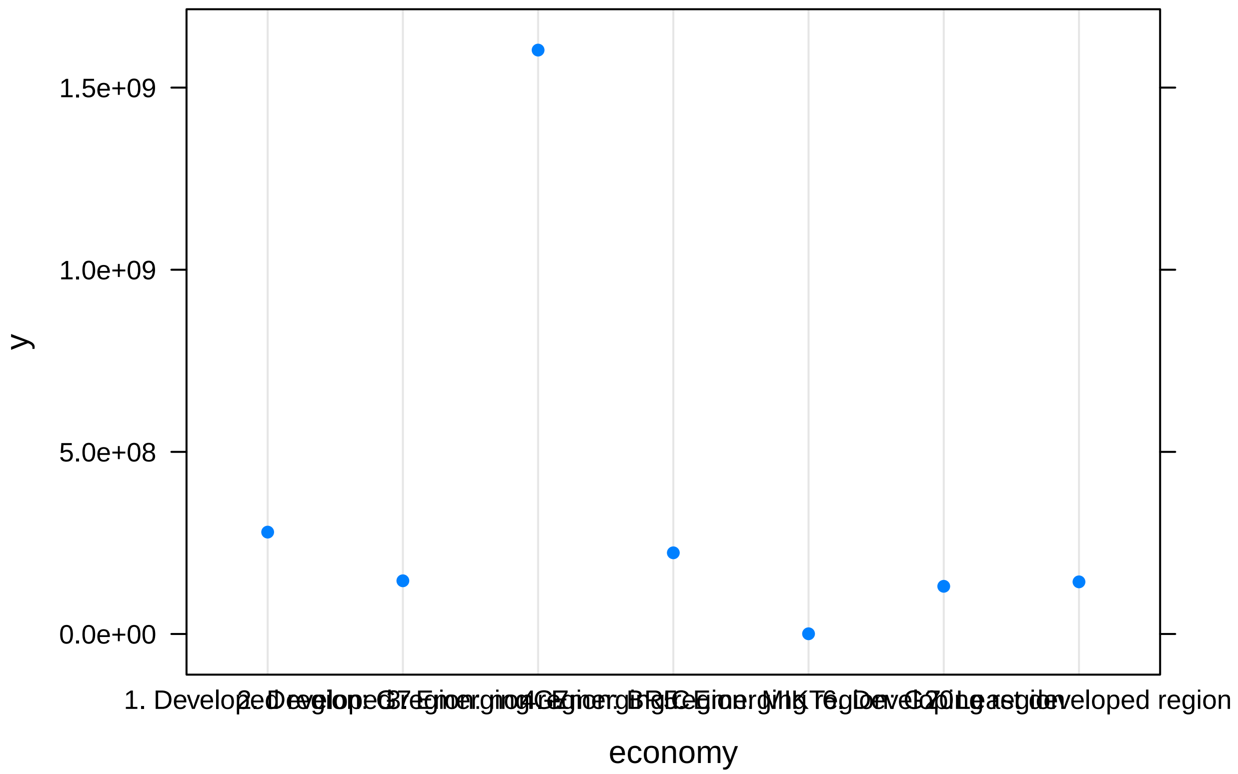
```
## Loading required package: gbm
```

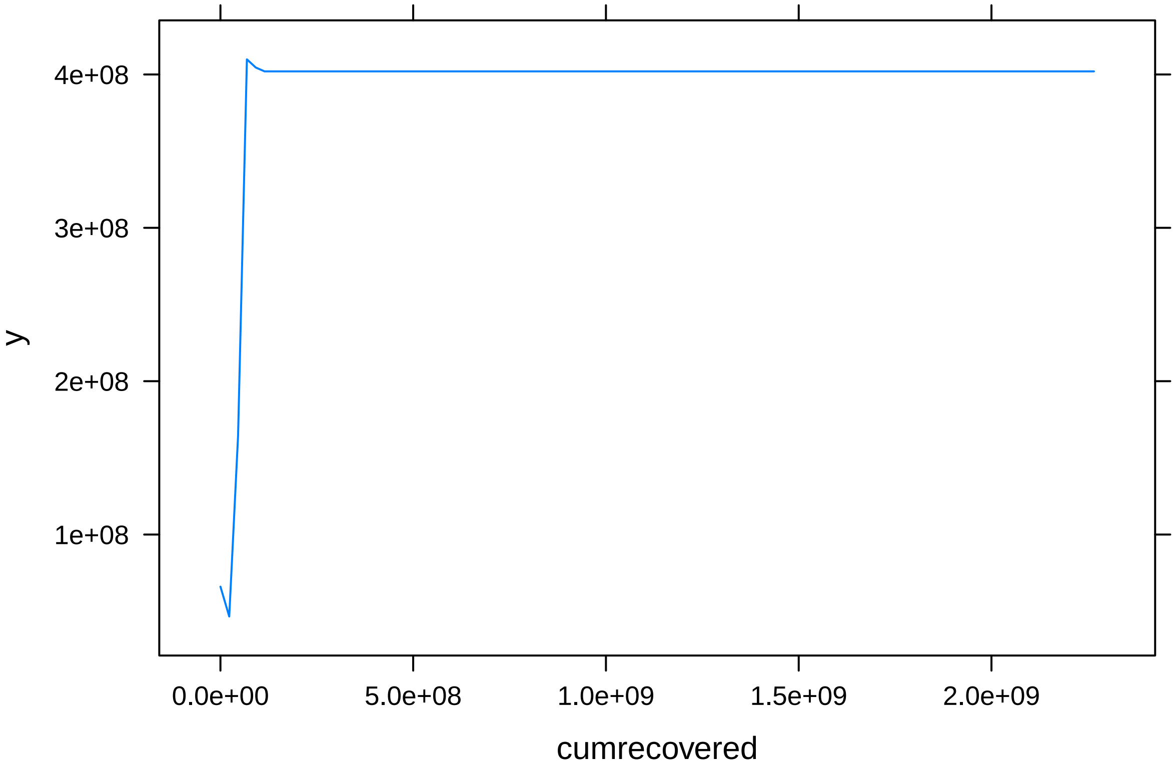
```
## Loaded gbm 2.1.8
```

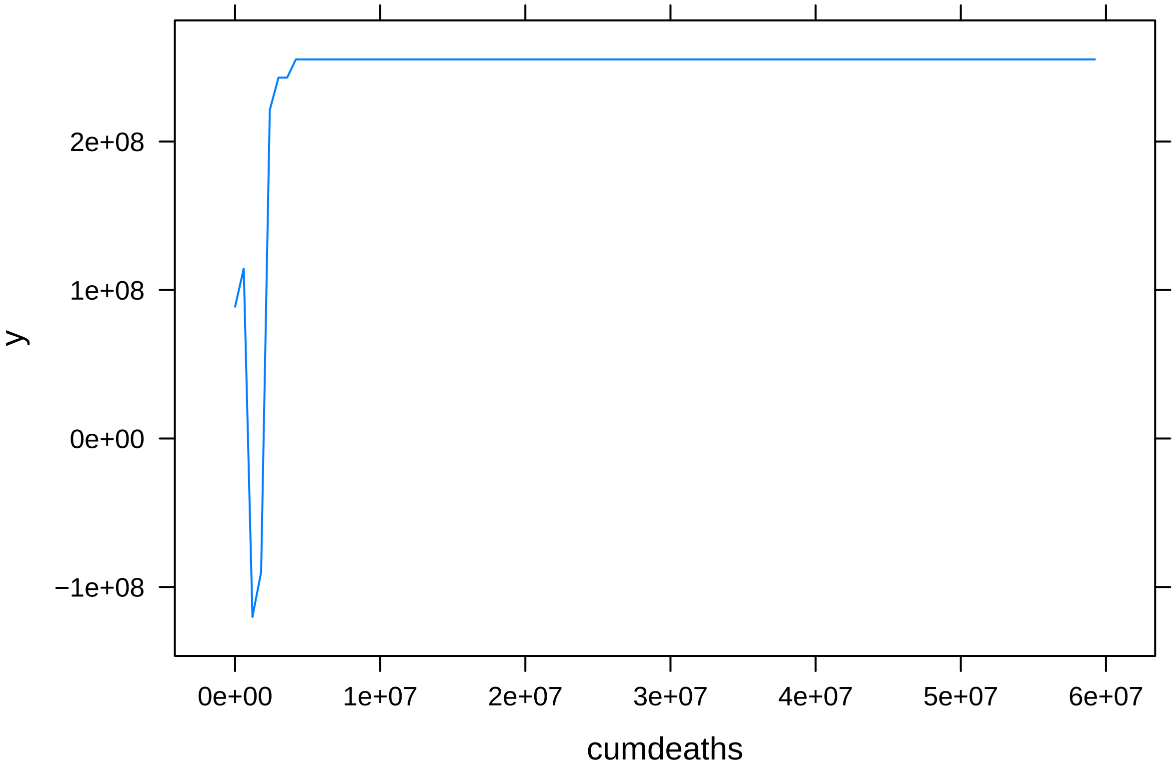


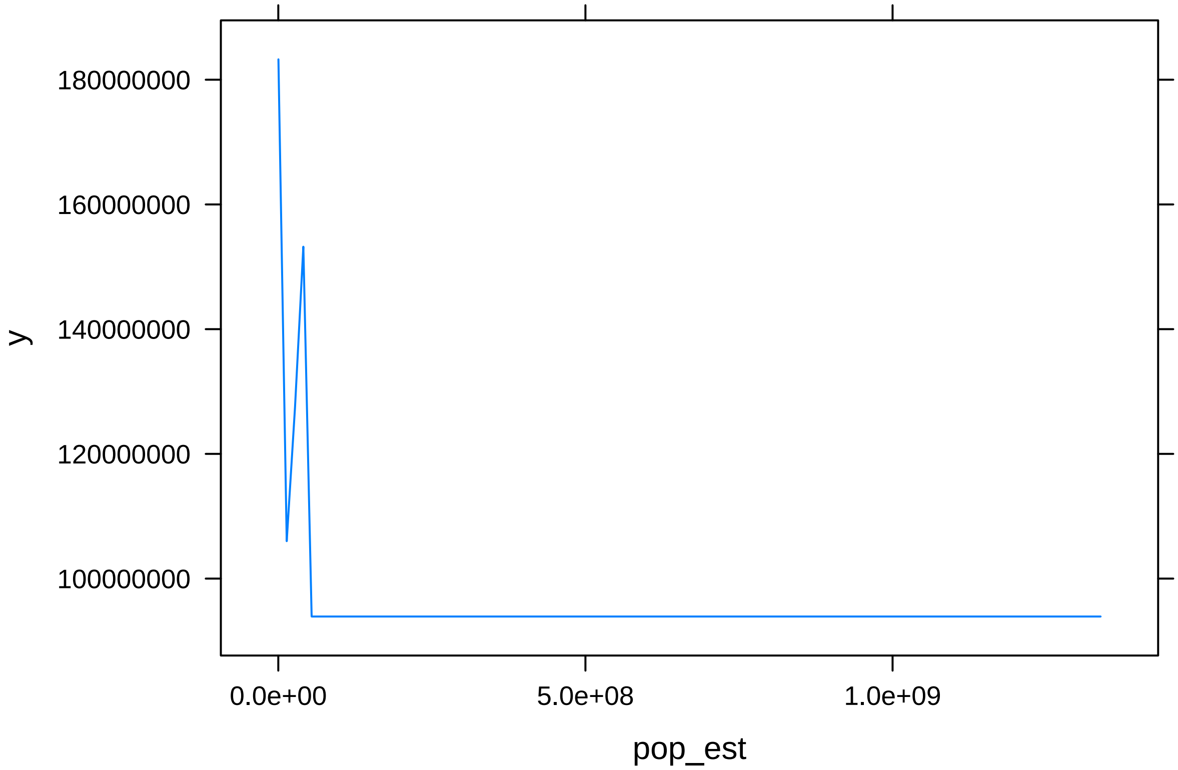
```
##                       var   rel.inf
## economy           economy 30.876848
## cumrecovered cumrecovered 20.420271
## cumdeaths       cumdeaths 16.824129
## pop_est_dens pop_est_dens 12.524883
## pop_est           pop_est 11.358863
## area                 area  7.995007
```


y
2.0e+08
1.5e+08
1.0e+08
5.0e+07
0
200
400
600
800
1000
1200
pop_est_dens


y
1.5e+09
1.0e+09
5.0e+08
0.0e+00
1. Developed region: G7
2. Developed region: nonG7
3. Emerging region: BRIC
4. Emerging region: MIKT
5. Emerging region: G20
6. Developing region
7. Least developed region
economy


4e+08
3e+08
2e+08
1e+08
y
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
cumrecovered


2e+08
1e+08
0e+00
−1e+08
y
0e+00
1e+07
2e+07
3e+07
4e+07
5e+07
6e+07
cumdeaths



The best predictor seems to be economy.

cumrecovered estimate is the next best predictor used followed by cumdeaths, pop_est_dens, area and pop_est.

The tree() function has used pop_est and economy. for building the Regression tree.