indi_asm

Chan Hou Long, Guyver

10/17/2021

1

1a, b

```
set.seed(5312)
x = rnorm(100)
eps = rnorm(100, 0, sqrt(0.25))
```

1c

B0 is -1 and B1 is 0.5. The length of y is 100.

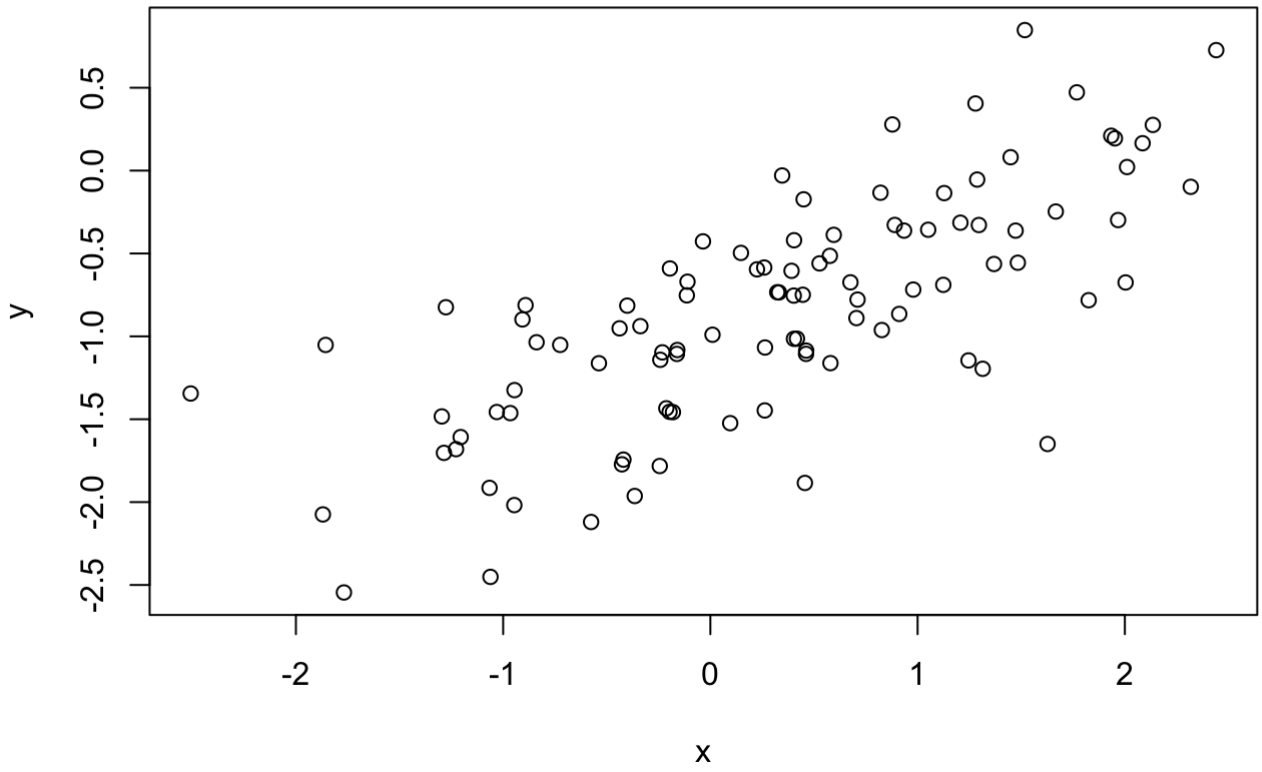
```
y = -1 + 0.5 * x + eps
length(y)
```

```
## [1] 100
```

1d

The relationship betweena x and y seems linear with some outliers by eps.

```
plot(y ~ x)
```



1e

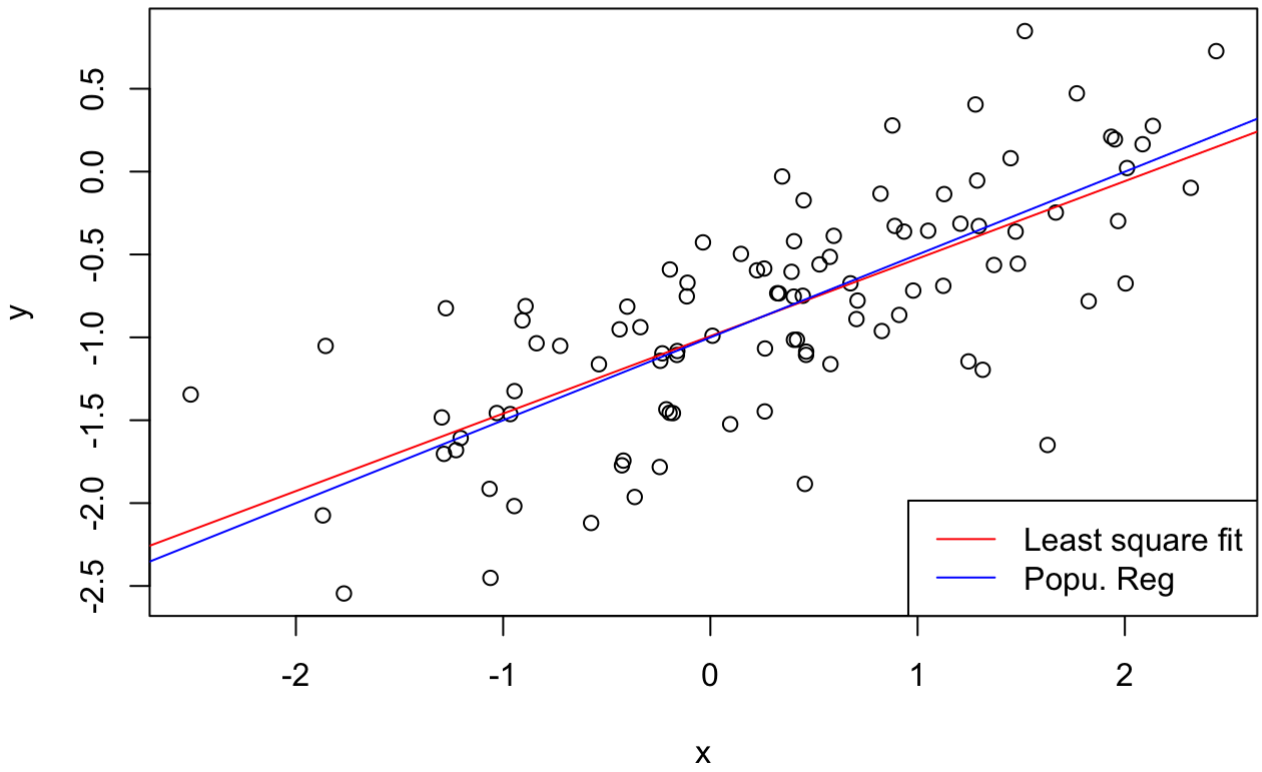
Since the p-value of the model is very small (less than 0.05) with a large F-statistic, there is significant evidence to show the relationship between B0 and B1. As the values of both samples B0 and B1 are closed to B0 and B1, the H0 can be rejected.

```
fit = lm(y ~ x)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41808 -0.26769  0.02606  0.29404  1.13044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.9922     0.0494  -20.09   <2e-16 ***
## x             0.4676     0.0448   10.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4729 on 98 degrees of freedom
## Multiple R-squared:  0.5264, Adjusted R-squared:  0.5216
## F-statistic: 108.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

1f

```
plot(x, y)
abline(fit, col = "red")
abline(-1, 0.5, col = "blue")
legend("bottomright", c("Least square fit", "Popu. Reg"), col = c("red", "blue"), lty
= c(1, 1))
```



1g

Although there is a slight improvement in RSE and R^2, since the p-value of X^2 is 0.281, which is higher than 0.05, so there is not sufficient evidence to show this fit2 can improve the fitness of the model.

```
fit2 <- lm(y ~ poly(x,2))
summary(fit2)
```

```
##
## Call:
## lm(formula = y ~ poly(x, 2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4488 -0.2913  0.0376  0.3310  1.1107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84313    0.04725 -17.845   <2e-16 ***
## poly(x, 2)1  4.93565    0.47246  10.447   <2e-16 ***
## poly(x, 2)2  0.51268    0.47246   1.085    0.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4725 on 97 degrees of freedom
## Multiple R-squared:  0.5321, Adjusted R-squared:  0.5225
## F-statistic: 55.15 on 2 and 97 DF,  p-value: < 2.2e-16
```

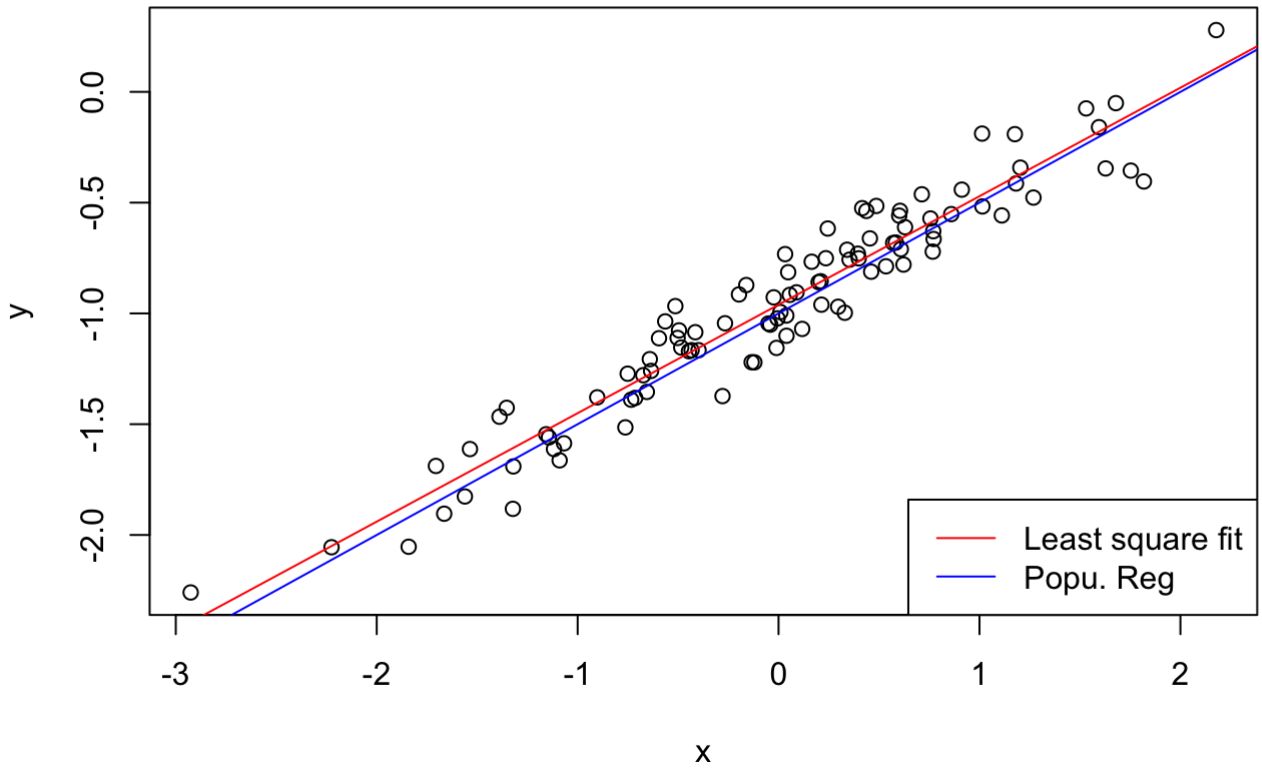
1h

To reduce the noise in the data, I would decrease the variance of eps' normal distribution. Both intercept and x have small p-value which smaller than 0.05, but the value of R^2 and RSE is much higher and lower respectively. So the relationship between them is linear with little noise.

```
set.seed(5312)
eps <- rnorm(100, sd = 0.125)
x <- rnorm(100)
y <- -1 + 0.5 * x + eps
plot(x, y)
fit3 <- lm(y ~ x)
summary(fit3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33465 -0.08515  0.00704  0.09672  0.27570
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.96021    0.01329  -72.24   <2e-16 ***
## x            0.48976    0.01416   34.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1329 on 98 degrees of freedom
## Multiple R-squared:  0.9243, Adjusted R-squared:  0.9235
## F-statistic:  1196 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
abline(fit3, col = "red")
abline(-1, 0.5, col = "blue")
legend("bottomright", c("Least square fit", "Popu. Reg"), col = c("red", "blue"), lty
= c(1, 1))
```



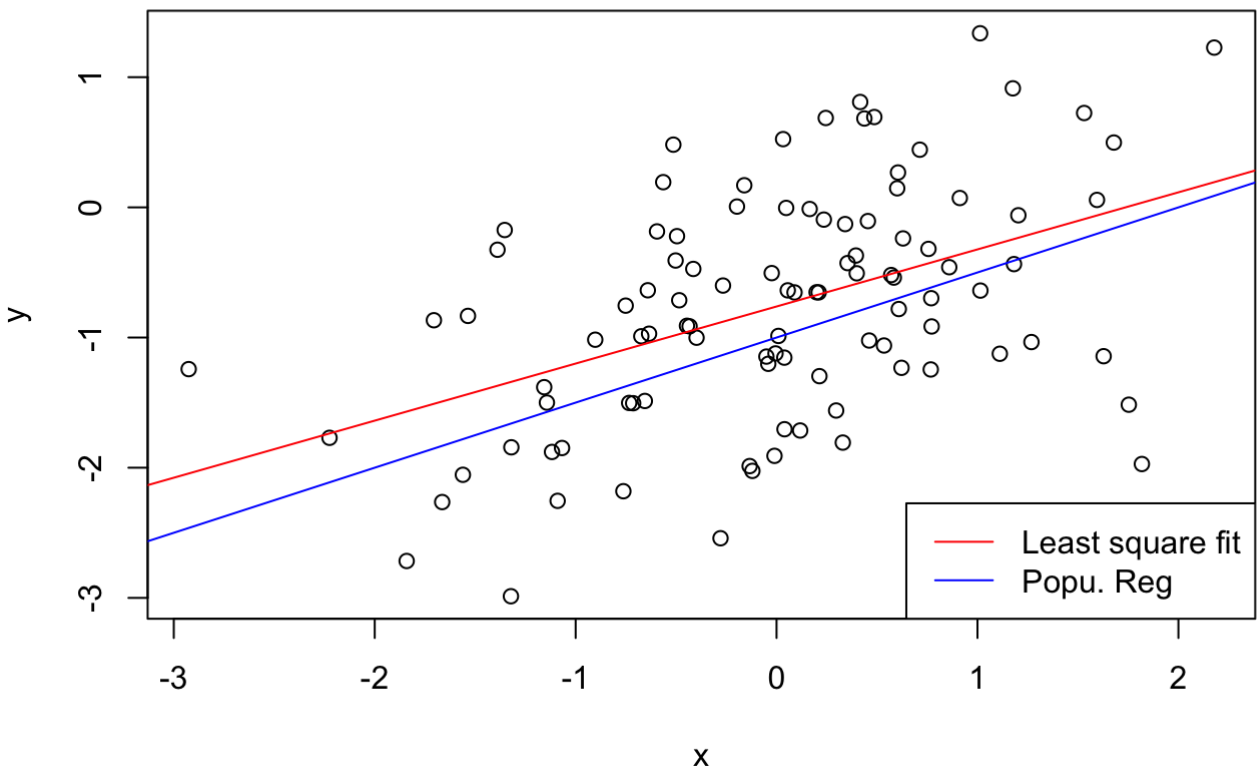
1i

I increase the eps' variance to add more noise. Although the variables have a small p-value where smaller than 0.05, the RSE and R^2 is much higher and lower respectively. Thus the relationship is not quite linear with a wider space within 2 lines.

```
set.seed(5312)
eps <- rnorm(100, sd = 0.75)
x <- rnorm(100)
y <- -1 + 0.5 * x + eps
plot(x, y)
fit4 <- lm(y ~ x)
summary(fit4)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00788 -0.51091  0.04224  0.58030  1.65418
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.76127    0.07976  -9.545 1.18e-15 ***
## x            0.43858    0.08496   5.162 1.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7975 on 98 degrees of freedom
## Multiple R-squared:  0.2138, Adjusted R-squared:  0.2058
## F-statistic: 26.65 on 1 and 98 DF,  p-value: 1.278e-06
```

```
abline(fit4, col = "red")
abline(-1, 0.5, col = "blue")
legend("bottomright", c("Least square fit", "Popu. Reg"), col = c("red", "blue"), lty
= c(1, 1))
```



1j

0.5 seems to be centred in all 3 intervals. The number of noise is proportional to the width of the confidence intervals. For example, there is more noise with wider intervals.

```
confint(fit)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0901912 -0.8941339
## x            0.3786996  0.5565165
```

```
confint(fit3)
```

```
##                  2.5 %     97.5 %
## (Intercept) -0.9865896 -0.9338321
## x            0.4616639  0.5178616
```

```
confint(fit4)
```

```
##                  2.5 %     97.5 %
## (Intercept) -0.9195374 -0.6029927
## x            0.2699835  0.6071697
```

2

This report provides a brief analysis of the daily and weekly COVID-19 case numbers of different countries. The data set concludes the observes and variables from daily and weekly dataframe.

Required library

```
library(naniar)
library(ggplot2)
library(readr)
library(tidyverse)
library(hrbrthemes)
library(reshape2)
library(plotly)
library(randomForest)
library(psych)
```

Data structure overview

The dataframe daily is used in the report. Firstly, the original data contains 56487 rows with 25 columns. It includes features such as `Country` , `Datetime` and policies. However, there are 9583 empty values in total which `H3_Contact tracing` contains the most empty values. Moreover, there are 170 countries in daily. From the summary, we can discover that the time period is from `2020-01-03` to `2021-01-31` . The distribution of `New_cases` and `New_deaths` is skewed right. From the boxplot containing all policies, `H1_Public information campaigns` and `H7_Vaccination policy` have more outliers. Most of the policies have a centre between 1-3.

```
load("/Users/guyverchan/Documents/HKU/SOWK3136/covid.RData")

dim(daily)
```

```
## [1] 56487    24
```

```
names(daily)
```

```
##  [1] "Country"
##  [2] "CountryCode"
##  [3] "Datetime"
##  [4] "New_cases"
##  [5] "Cumulative_cases"
##  [6] "New_deaths"
##  [7] "Cumulative_deaths"
##  [8] "C1_School closing"
##  [9] "C2_Workplace closing"
## [10] "C3_Cancel public events"
## [11] "C4_Restrictions on gatherings"
## [12] "C5_Close public transport"
## [13] "C6_Stay at home requirements"
## [14] "C7_Restrictions on internal movement"
## [15] "C8_International travel controls"
## [16] "E1_Income support"
## [17] "E2_Debt/contract relief"
## [18] "H1_Public information campaigns"
## [19] "H2_Testing policy"
## [20] "H3_Contact tracing"
## [21] "H6_Facial Coverings"
## [22] "H7_Vaccination policy"
## [23] "GEI"
## [24] "GHS"
```

```
# Empty value
sum(is.na(daily))
```

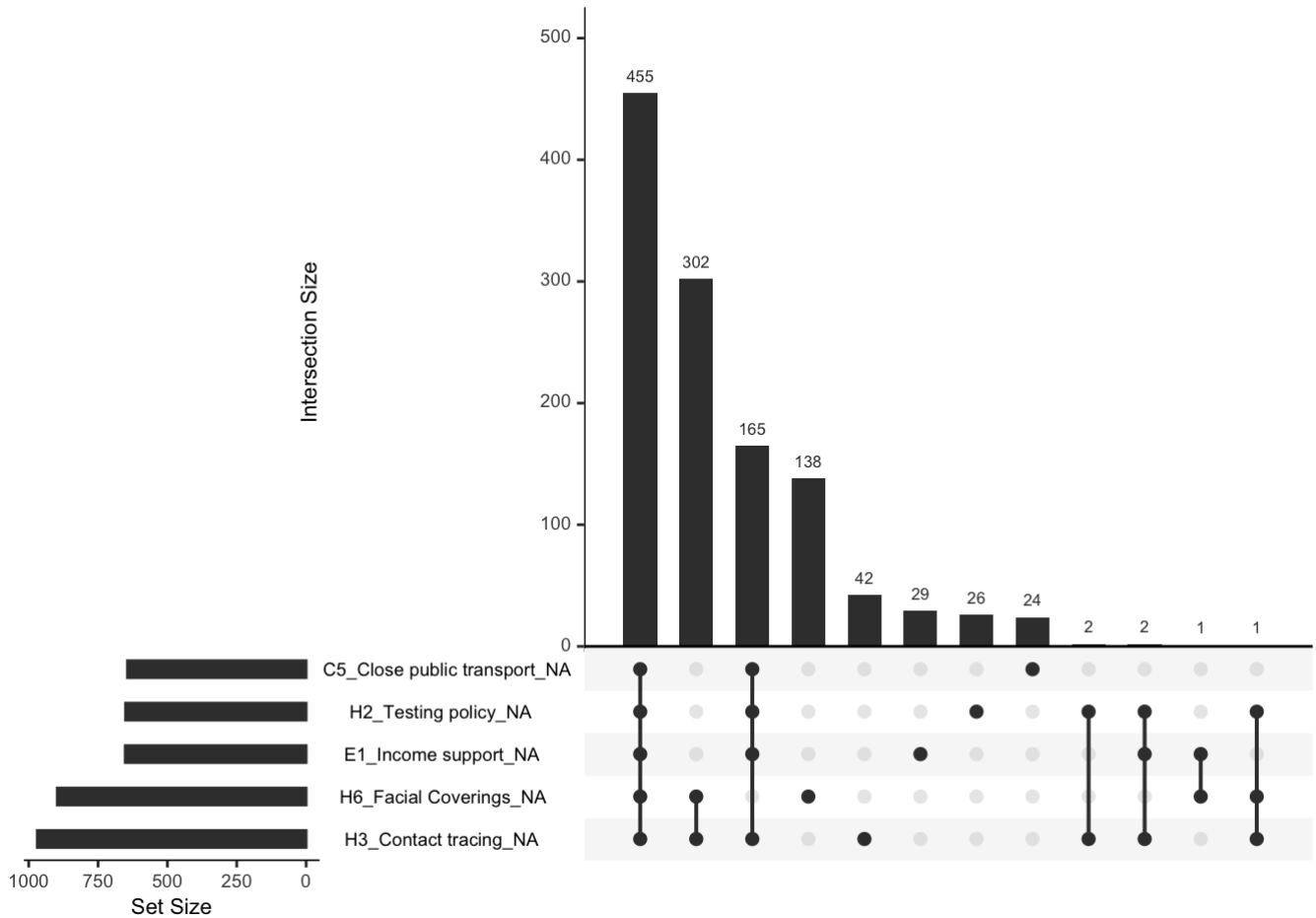
```
## [1] 9583
```

```
apply(apply(daily,2,is.na),2,sum) ; nrow(daily)
```

```
##                        Country                      CountryCode
##                              0                                0
##                       Datetime                       New_cases
##                              0                                1
##              Cumulative_cases                      New_deaths
##                              1                                0
##             Cumulative_deaths                C1_School closing
##                              0                              551
##           C2_Workplace closing          C3_Cancel public events
##                            632                              555
##  C4_Restrictions on gatherings          C5_Close public transport
##                            557                              644
##   C6_Stay at home requirements C7_Restrictions on internal movement
##                            632                              632
##  C8_International travel controls                E1_Income support
##                            559                              652
##         E2_Debt/contract relief      H1_Public information campaigns
##                            634                              557
##              H2_Testing policy                H3_Contact tracing
##                            651                              969
##             H6_Facial Coverings             H7_Vaccination policy
##                            897                              459
##                            GEI                              GHS
##                              0                                0
```

```
## [1] 56487
```

```
gg_miss_upset(daily)
```



```
# Sum of Contries
countCountry = daily[!duplicated(daily$Country), ]
dim(countCountry)
```

```
## [1] 170  24
```

```
summary(daily)
```

```
##    Country           CountryCode           Datetime            New_cases
##  Length:56487       Length:56487       Min.   :2020-01-03   Min.   : -8956
##  Class :character   Class :character   1st Qu.:2020-05-26   1st Qu.:     1
##  Mode  :character   Mode  :character   Median :2020-08-18   Median :    53
##                                        Mean   :2020-08-16   Mean   :  1780
##                                        3rd Qu.:2020-11-09   3rd Qu.:   567
##                                        Max.   :2021-01-31   Max.   :402270
##                                                             NA's   :1
##  Cumulative_cases     New_deaths      Cumulative_deaths C1_School closing
##  Min.   :       0   Min.   :-514.00   Min.   :     0    Min.   :0.000
##  1st Qu.:     599   1st Qu.:   0.00   1st Qu.:    10    1st Qu.:1.000
##  Median :    6200   Median :   1.00   Median :   113    Median :2.000
##  Mean   :  182365   Mean   :  38.75   Mean   :  4947    Mean   :2.062
##  3rd Qu.:   62956   3rd Qu.:   9.00   3rd Qu.:  1160    3rd Qu.:3.000
##  Max.   :25676612   Max.   :6409.00   Max.   :433173    Max.   :3.000
##  NA's   :1                                              NA's   :551
##  C2_Workplace closing C3_Cancel public events C4_Restrictions on gatherings
##  Min.   :0.000        Min.   :0.000           Min.   :0.000
##  1st Qu.:1.000        1st Qu.:1.000           1st Qu.:2.000
##  Median :2.000        Median :2.000           Median :3.000
##  Mean   :1.561        Mean   :1.545           Mean   :2.747
##  3rd Qu.:2.000        3rd Qu.:2.000           3rd Qu.:4.000
##  Max.   :3.000        Max.   :2.000           Max.   :4.000
##  NA's   :632          NA's   :555             NA's   :557
##  C5_Close public transport C6_Stay at home requirements
##  Min.   :0.000             Min.   :0.000
##  1st Qu.:0.000             1st Qu.:0.000
##  Median :0.000             Median :1.000
##  Mean   :0.652             Mean   :1.134
##  3rd Qu.:1.000             3rd Qu.:2.000
##  Max.   :2.000             Max.   :3.000
##  NA's   :644               NA's   :632
##  C7_Restrictions on internal movement C8_International travel controls
##  Min.   :0.000                        Min.   :0.000
##  1st Qu.:0.000                        1st Qu.:2.000
##  Median :1.000                        Median :3.000
##  Mean   :1.049                        Mean   :2.821
##  3rd Qu.:2.000                        3rd Qu.:4.000
##  Max.   :2.000                        Max.   :4.000
##  NA's   :632                          NA's   :559
##  E1_Income support E2_Debt/contract relief H1_Public information campaigns
##  Min.   :0.0000    Min.   :0.000           Min.   :0.000
##  1st Qu.:0.0000    1st Qu.:0.000           1st Qu.:2.000
##  Median :1.0000    Median :1.000           Median :2.000
##  Mean   :0.8856    Mean   :1.075           Mean   :1.905
##  3rd Qu.:2.0000    3rd Qu.:2.000           3rd Qu.:2.000
##  Max.   :2.0000    Max.   :2.000           Max.   :2.000
##  NA's   :652       NA's   :634             NA's   :557
##  H2_Testing policy H3_Contact tracing H6_Facial Coverings H7_Vaccination policy
##  Min.   :0.000     Min.   :0.000      Min.   :0.000       Min.   :0.0000
##  1st Qu.:1.000     1st Qu.:1.000      1st Qu.:1.000       1st Qu.:0.0000
##  Median :2.000     Median :2.000      Median :3.000       Median :0.0000
##  Mean   :1.799     Mean   :1.452      Mean   :2.197       Mean   :0.0754
##  3rd Qu.:2.000     3rd Qu.:2.000      3rd Qu.:3.000       3rd Qu.:0.0000
##  Max.   :3.000     Max.   :2.000      Max.   :4.000       Max.   :5.0000
##  NA's   :651       NA's   :969        NA's   :897         NA's   :459
##       GEI               GHS
```

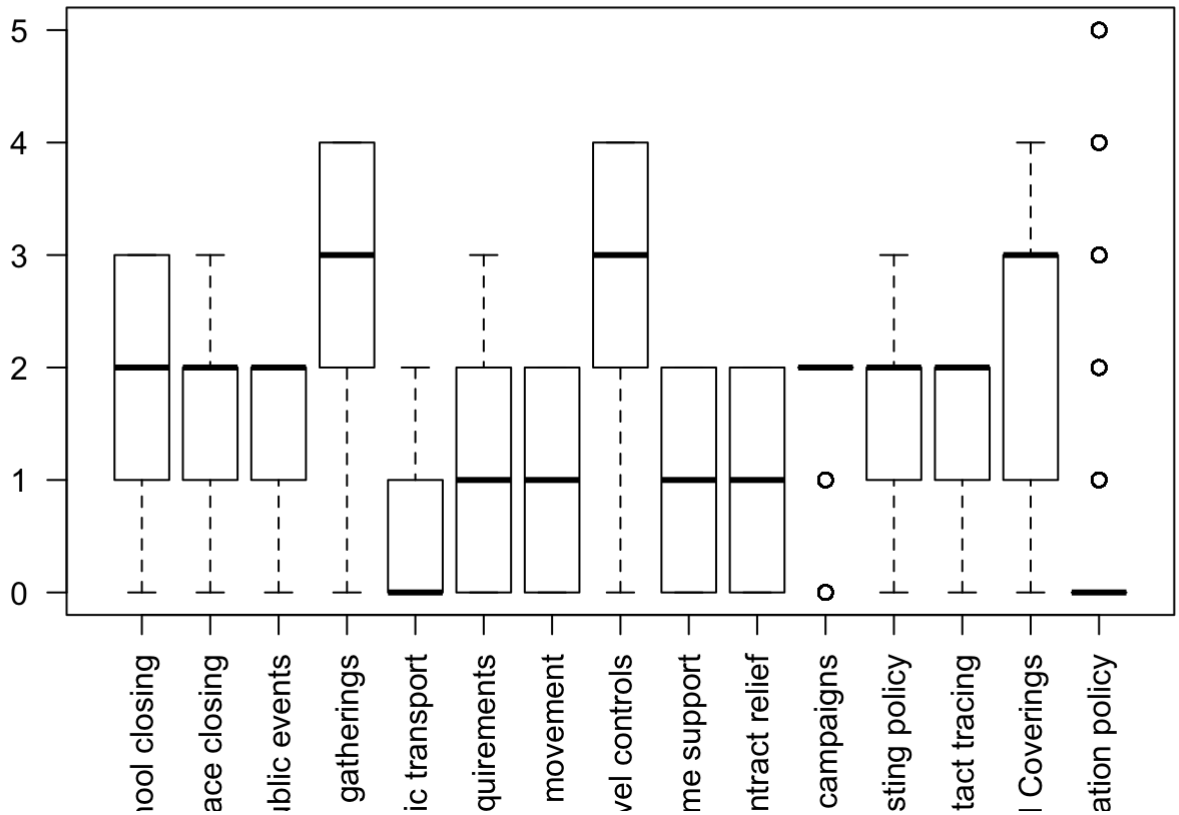
```
##  Min.   :-2.45000   Min.   :16.60
##  1st Qu.:-0.68000   1st Qu.:31.80
##  Median :-0.21000   Median :40.10
##  Mean   :-0.02194   Mean   :42.36
##  3rd Qu.: 0.61000   3rd Qu.:52.00
##  Max.   : 2.23000   Max.   :83.50
##
```

```
describe(daily[4:22])
```

	v... <int>	n <dbl>	mean <dbl>	sd <dbl>	med... <dbl>
New_cases	1	56486	1.779690e+03	9.583167e+03	53
Cumulative_cases	2	56486	1.823645e+05	9.810252e+05	6200
New_deaths	3	56487	3.874911e+01	1.739163e+02	1
Cumulative_deaths	4	56487	4.946922e+03	2.148163e+04	113
C1_School closing	5	55936	2.061570e+00	1.034335e+00	2
C2_Workplace closing	6	55855	1.561382e+00	9.811605e-01	2
C3_Cancel public events	7	55932	1.544554e+00	7.166066e-01	2
C4_Restrictions on gatherings	8	55930	2.746952e+00	1.420682e+00	3
C5_Close public transport	9	55843	6.520065e-01	7.472649e-01	0
C6_Stay at home requirements	10	55855	1.134133e+00	9.281785e-01	1

1-10 of 19 rows | 1-7 of 14 columns Previous **1** 2 Next

```
boxplot(daily[8:22], las=2)
```



Data preparation

It is interesting to observe that some policies are correlated to each other. For example, `C3_Cancel public events` and `C4_Restrictions on gatherings` share `0.65` correlate coefficient, `C6_Stay at home requirements` and `C7_Restrictions on internal movement` share `0.59` correlate coefficient. It is reasonable in reality because gatherings are also one of the public events and people would stay at home more while restricting the internal movement.

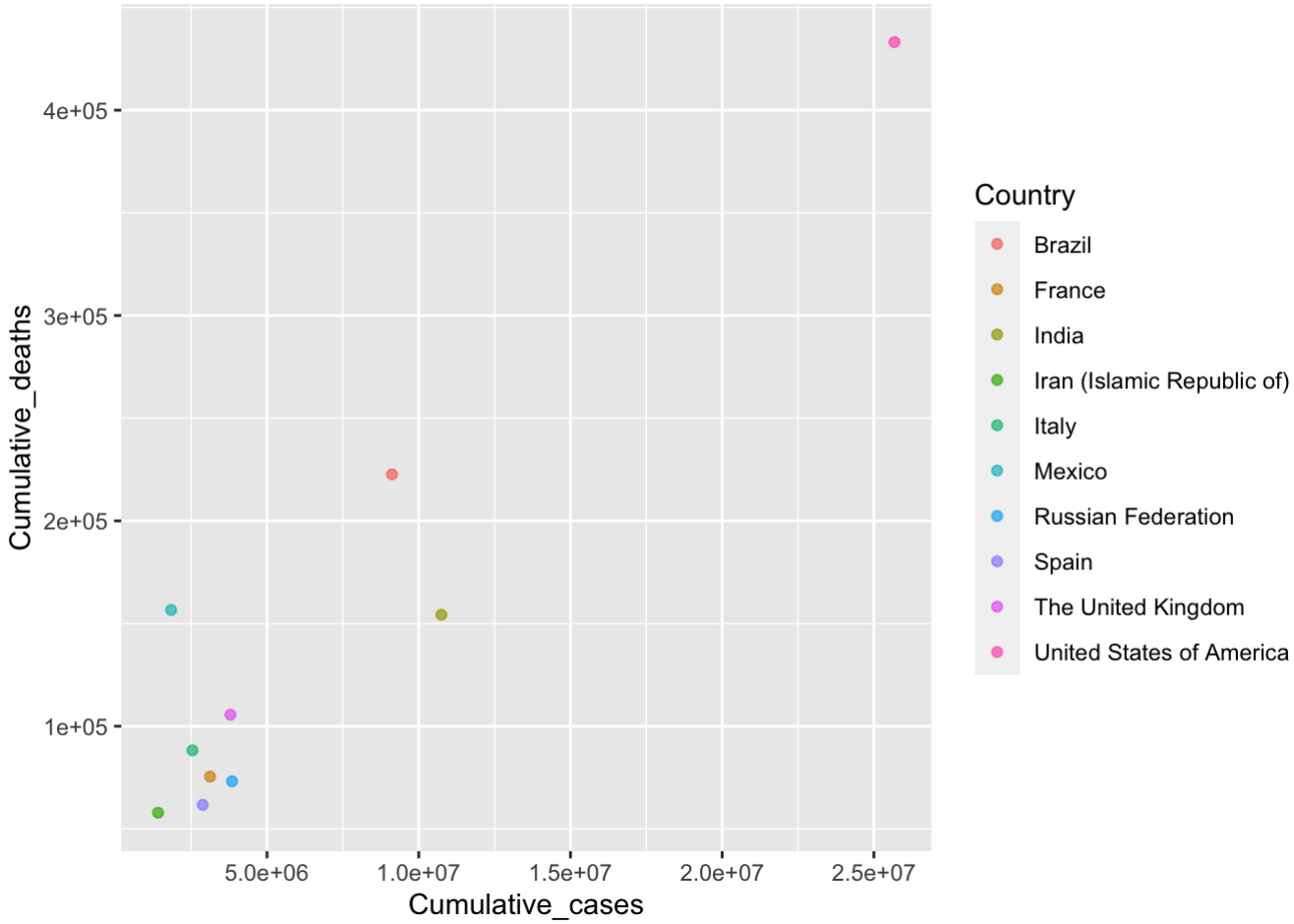
```
df = daily[complete.cases(daily$New_cases), ]
df$Day = as.Date(df$Datetime, "&b")
df$New_cases[df$New_cases<0] = 0
df$New_deaths[df$New_deaths<0] = 0

end = df[df$Day=="2021-01-31", ]

corDf = select(df,-Day,-Datetime, -Country, -CountryCode)
corM = cor(model.matrix(~0+., data=corDf), use="pairwise.complete.obs")
zdf = as.data.frame(as.table(corM))
zdf = zdf[zdf$Freq != 1, ]
zdf = arrange(zdf,desc(Freq))
```

2a

What is the top 10 most cases and deaths countries?

From the scattered plot of the top 10 most cases and deaths countries, it can be shown that `USA` is the most cumulative cases and deaths country. `Brazil` is the country with the second most cumulative death and followed by `India` and `France`. Therefore, I would select the four countries `USA`, `Brazil`, `France` and `Mexico` for the research.

```
end = end %>% arrange(desc(Cumulative_deaths, Cumulative_cases))
ggplot(end[1:10, ], aes(x=Cumulative_cases, y=Cumulative_deaths, color=Country)) +
    geom_point(alpha=0.7)
```



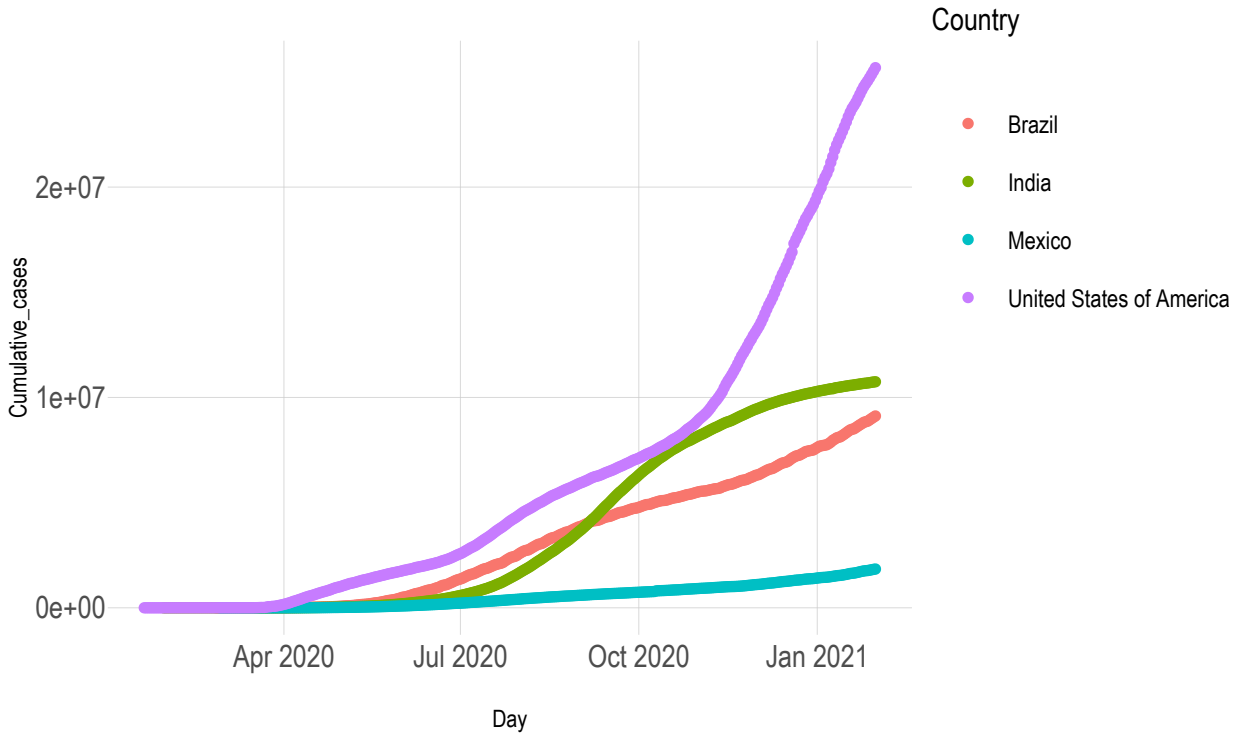

```
df_selected = df[df$Country == "Brazil" | df$CountryCode == "IND" |
                   df$CountryCode == "USA" | df$Country == "Mexico", ]
```

2b

From this line graph, the cumulative cases of the USA increased explosively in Q4. Both India and Brazil still rose steadily during the whole year. Only Mexico grow slowly among all four countries.

```
cCases = ggplot(df_selected, aes(x=Day, y=Cumulative_cases, col=Country)) +
  geom_point(size=1) +
  theme_ipsum() +
  ggtitle("Cumulative cases in 4 countries")

ggplotly(cCases)
```

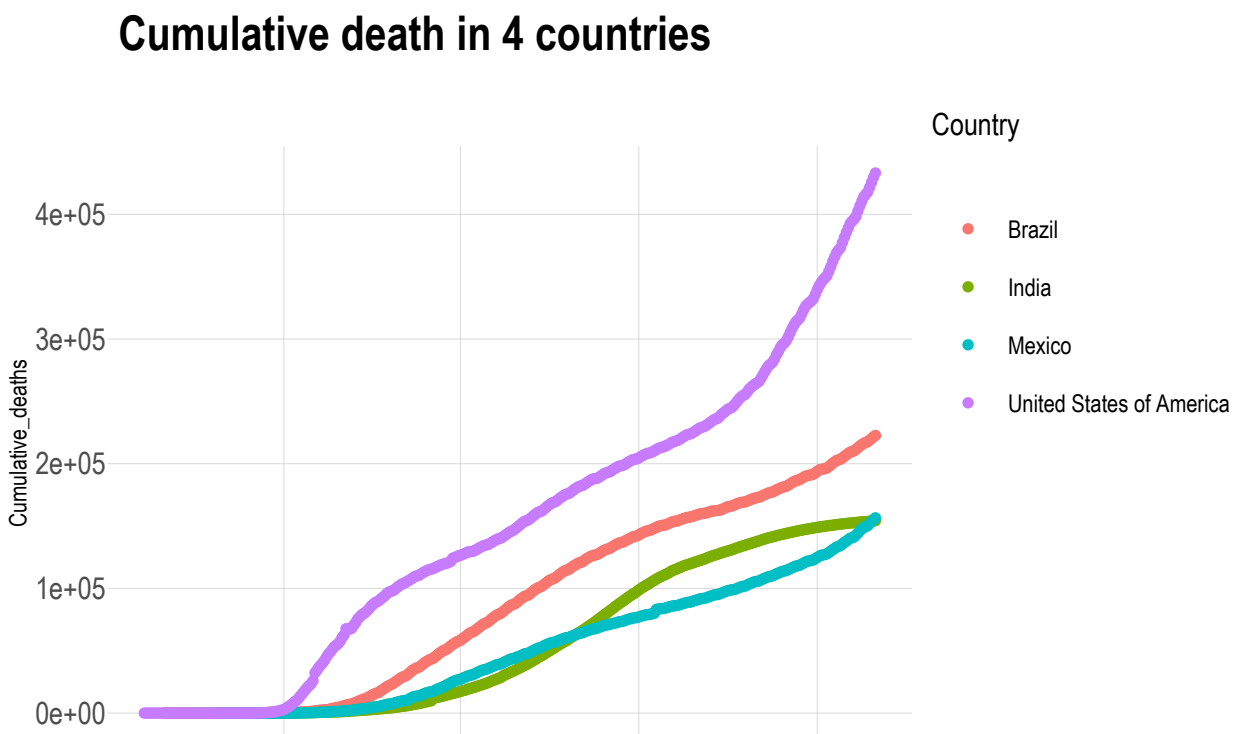
Cumulative cases in 4 countries

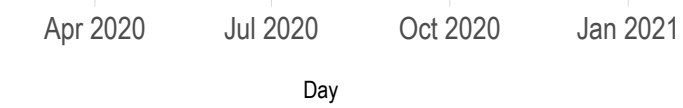


2c

The graph shows that the cumulative deaths had an up-rising number starting from April 2020 inside the USA. It led for a year among four countries. In contrast, the other three countries kept increasing steadily.

```
cDeaths = ggplot(df_selected, aes(x=Day, y=Cumulative_deaths, col=Country)) +
  geom_point(size=1) +
  theme_ipsum() +
  ggtitle("Cumulative death in 4 countries")

ggplotly(cDeaths)
```

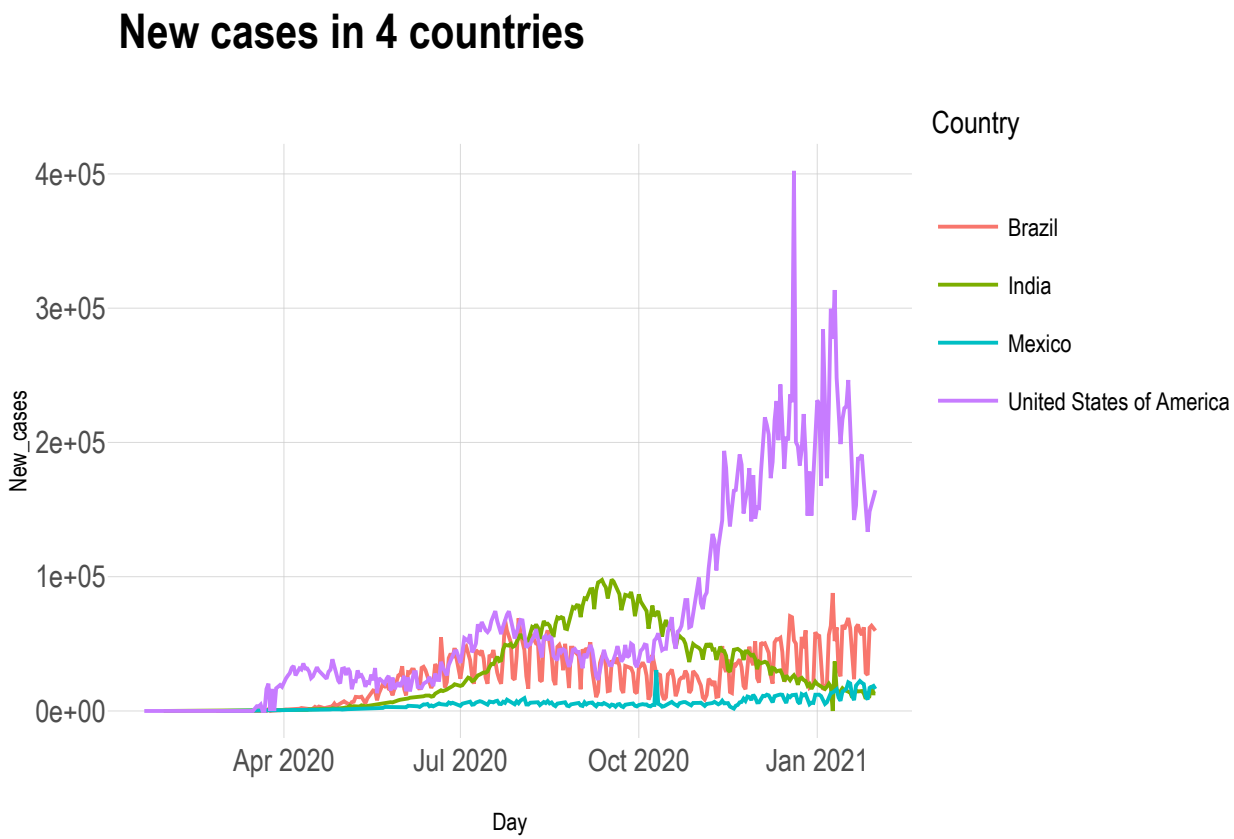




2d

Entering November of 2020, the USA faced a new wave of infection which led to the number of new cases increasing and reaching the peak in the Christmas holiday. In contrast, the other three countries had a steady number of cases during the year although India faced a small wave in September 2020.

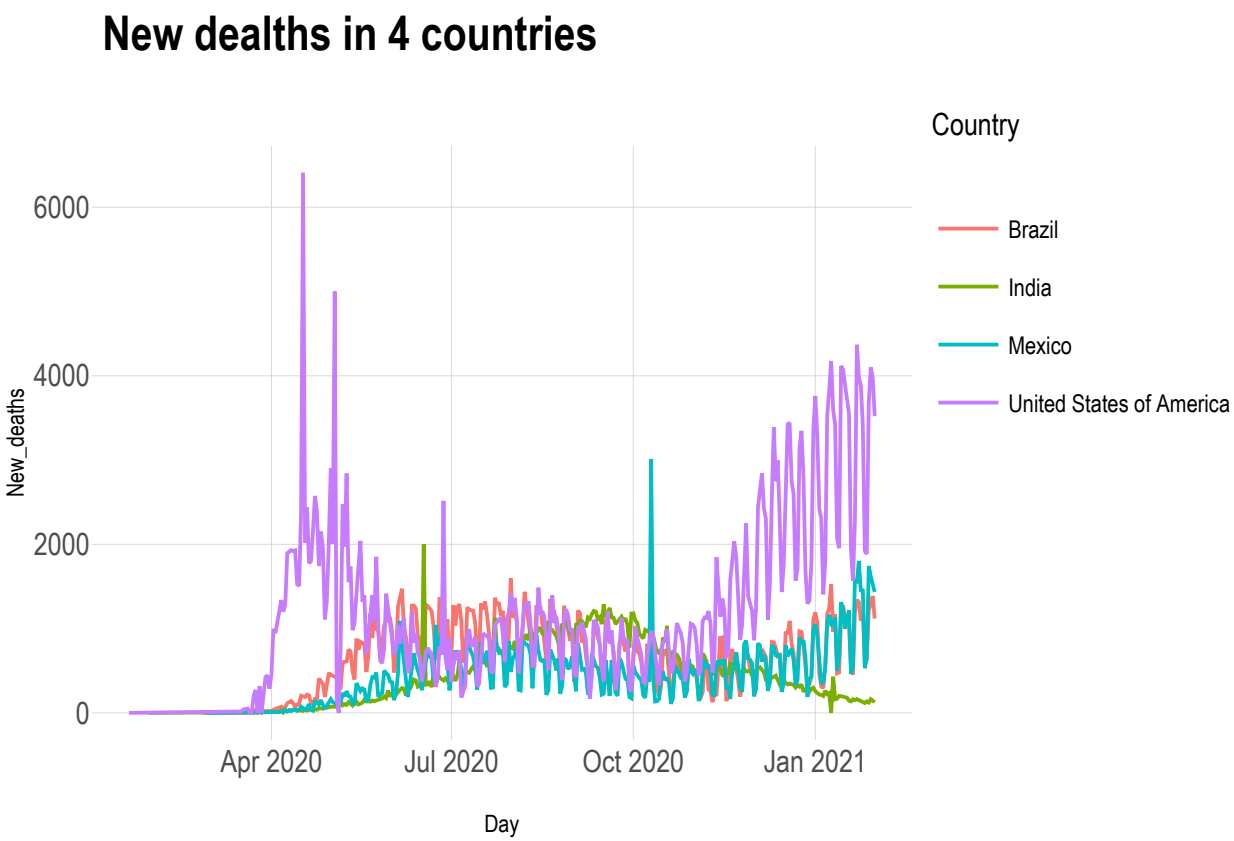
```
cases = ggplot(df_selected, aes(x=Day, y=New_cases, col=Country)) +
  geom_line(size=.51) +
  theme_ipsum() +
  ggtitle("New cases in 4 countries")

ggplotly(cases)
```



2e

From this line chart, the USA had a sudden boost in deaths number in April 2020. Although the number dropped for the third quarter of 2020, it rebounded starting from October and kept increasing. Although Mexico also had a sudden rise on 10 October 2020, both Mexico and Brazil regularly increased the deaths from December. However, India was the only country that dropped the number.

```
dealths = ggplot(df_selected, aes(x=Day, y=New_deaths, col=Country)) +
  geom_line(size=.5) +
  theme_ipsum() +
  ggtitle("New dealths in 4 countries")

ggplotly(dealths)
```

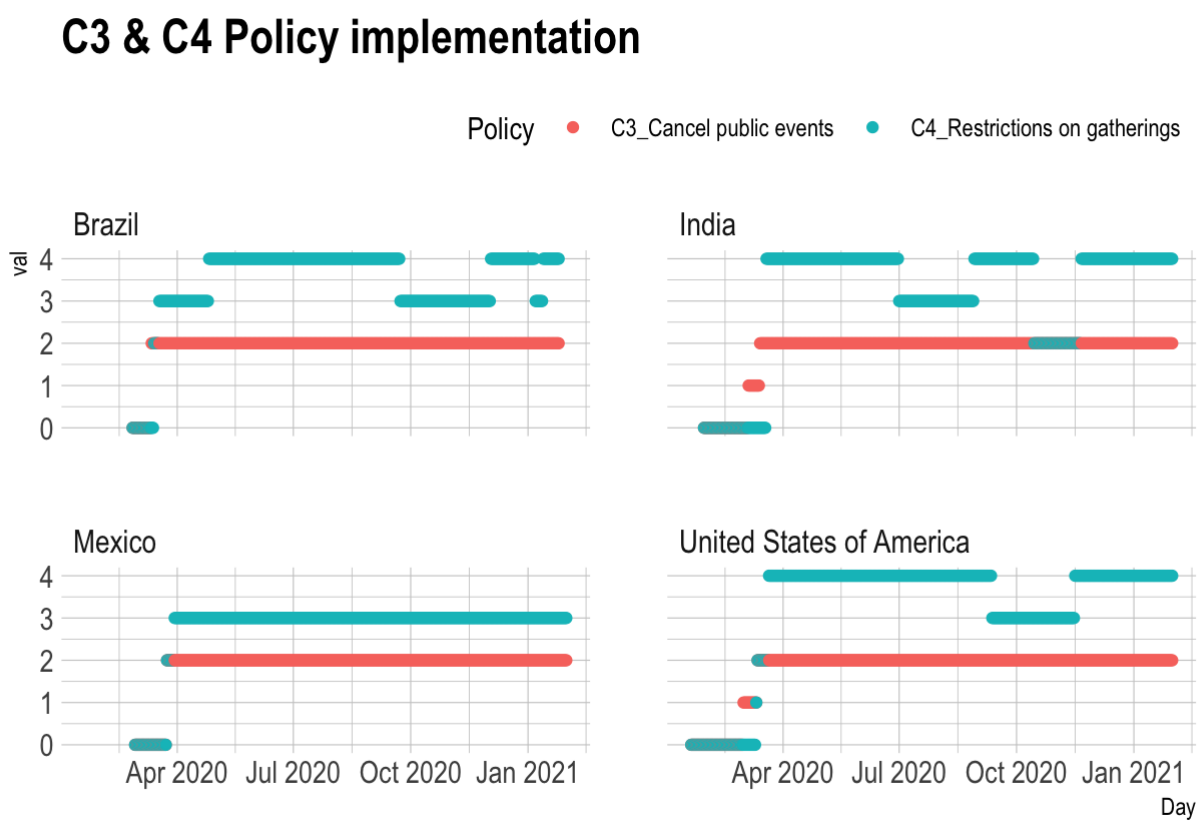


2f

What is the change after implementing C3 & C4?

From the policy implementation of C3 and C4, the pattern of all four countries seems similar which may have a relationship between them. Starting from April 2020, all four countries had implemented restrictions on public events and gatherings. C4 should be an effective policy in India as the new cases increased while the government lowered the level during September. But in other countries, there were still having new cases every day, especially in the USA. The number still grow while banning people from having gatherings.

```
df_selected %>% pivot_longer(cols = c(`C3_Cancel public events`, `C4_Restrictions on
gatherings`), values_to = "val", names_to = "Policy") %>%
  ggplot(aes(x = Day, y = val, col = Policy))+
  geom_point()+
  theme_ipsum() +
  facet_wrap( ~ Country, ncol=2) +
  ggtitle("C3 & C4 Policy implementation") +
  theme(legend.position = "top",
        legend.justification='right')
```

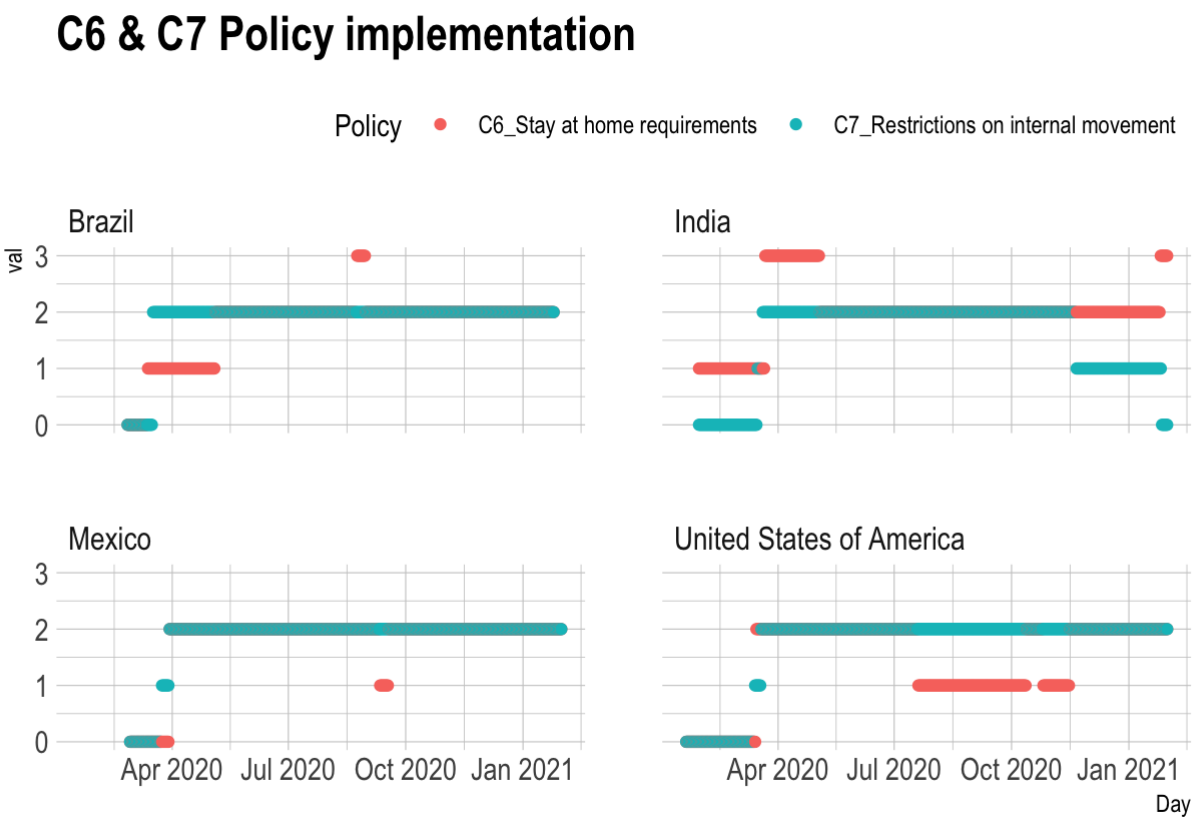


2e

What is the change after implementing C6 & C7?

From the policy implementation of C6 and C7, the pattern of all four countries seems similar which may have a relationship between them. Starting from April 2020, all four countries had implemented restrictions on internal movement and stay at home policy. However, the two policies should have a smaller weight affecting the number of new cases as they kept increasing while implementing these rules.

```
df_selected %>% pivot_longer(cols = c(`C6_Stay at home requirements`, `C7_Restriction
s on internal movement`), values_to = "val", names_to = "Policy") %>%
  ggplot(aes(x = Day, y = val, col = Policy))+
  geom_point()+
  theme_ipsum() +
  facet_wrap( ~ Country, ncol=2) +
  ggtitle("C6 & C7 Policy implementation") +
  theme(legend.position = "top",
        legend.justification='right')
```

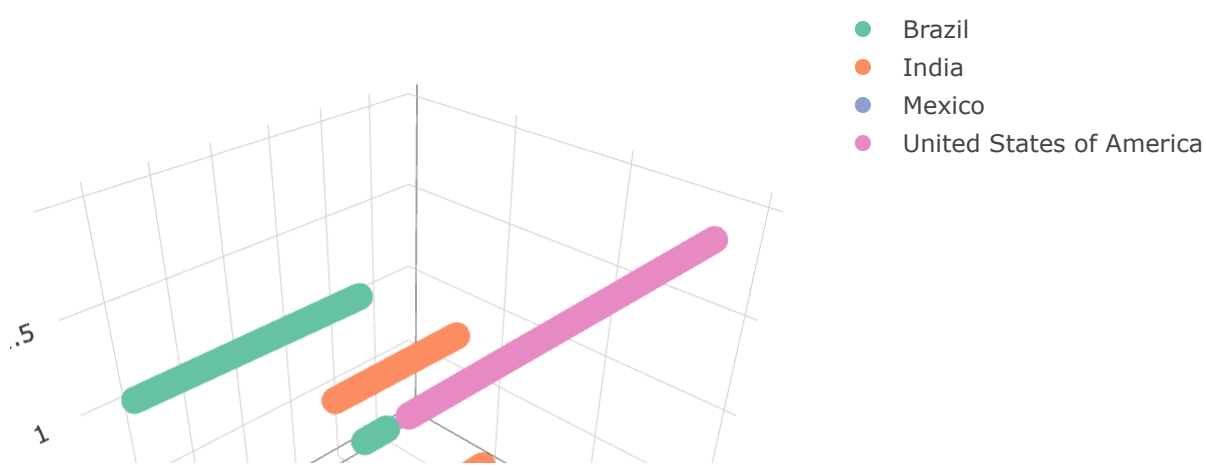


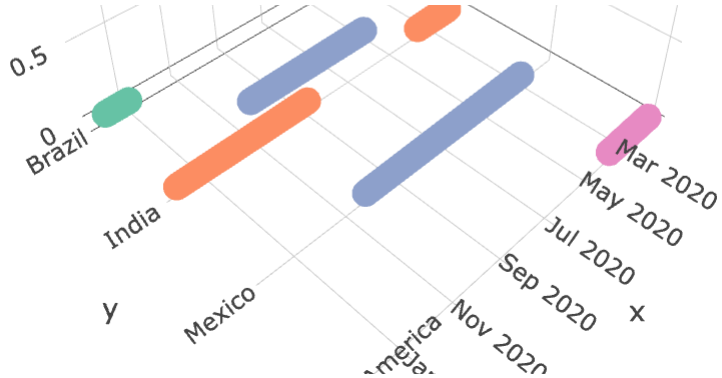
2f

Did the countries help affected citizens in economic aspects?

Among the four countries, the USA provided the most financial assistance to their citizens. The number should also affect by the most number of new cases and deaths. In contrast, India decreased the level of income support while the number of new cases dropped and fewer citizens were affected. However, India still promoted its debt relief plan for the same with Mexico. Thus, all four countries had provided various planning assisting their people.

```
plot_ly(x=df_selected$Day, y=df_selected$Country, z=df_selected$`E1_Income support`,
type="scatter3d", mode="markers", color=df_selected$Country)
```






```
plot_ly(x=df_selected$Day, y=df_selected$Country, z=df_selected$`E2_Debt/contract rel
ief`, type="scatter3d", mode="markers", color=df_selected$Country)
```