Covid-19 Analysis

Code

INTRODUCTION

The most crucial thing in analyzing a pandemic is to ensure that all the factors that are responsible for the outbreak are well considered. Missing out on any of such factors, however trivial it might seem, will not give an accurate analysis. What makes this type of analysis challenging is the number of factors that keep varying across different countries. Considering the common and important ones which can be relied on for accurate analysis is what we get to learn from this project.

Sometimes drilling down to root cause for analysis can turn out to be a futile effort. But identifying the right factors to do right analysis is equally important for analysing this kind of varying data. To name a few factors in the dataset that play a key role to help us analyse the outbreak are total cases, total deaths, total tests, population, recommendation delay, lockdown delay, health care index and many more. Data analysis of this dataset is needed to identify the direct and indirect relation between these variables. Using the existing statistics to identify such correlation can help countries in future to mitigate the widespread in case of such recurring pandemic. This has been the main interest behind analysing the dataset.

Analysing data for the COVID-19 pandemic is very important. There are few countries who were successful in controlling and few who were not. Despite a good health care index, some countries still failed miserably in controlling the widespread. Some, despite having more population, had fewer cases as compared to other countries with a smaller population size but a lot of positive corona infected cases. At this point it becomes difficult to consider which factors can be assumed to help countries from preventing the spread of this virus. This analysis will tell us that it is not just one or two factors that are essential but a combination of multiple factors that needs to be addressed and focused up on equally, which will help us to control the widespread and also on how to put an end to this pandemic or any such pandemic in future.

Data

Analysis of such data is challenging. As expected, there was no such particular data that had covered all the variables which play an essential role to help in analysing data for COVID-19. In this project multiple datasets were merged into one to represent a meaningful meaningful analysis that helps to make an accurate analysis.

Our dataset consists of 198 countries with 26 variables. There are few variables which were used to create data sets for accurate analysis.

Sr. No	Variable	Description
1	Country	Will contain all the country names
2	TotalCases	Total number of COVID-19 cases in each country
3	TotalDeaths	Total number of deaths due to COVID-19 in each country
4	TotalRecovered	Total patients who recovered from COVID-19
5	Population2020	Population of each country as of year 2020
6	TotalTests	Total number of tests conducted

7	Median_age	Canada
8	Recommendation_delay	Number of days delayed to issue recommendation since first corona case
9	Lockdown_delay	Number of days delayed to impose lockdown since first corona case
10	Health_care_index	Representing overall quality of the health care system on scale of 100
11	Density	Represents population per square kilometer for each country
12	Total Cases_per_million	Represents total cases per million

$$Preloading Libraries$$

Hide

```
library(tidyverse)
require(knitr)
require(lattice)
library(dplyr)
require(ggplot2)
install.packages("corrplot")
require(corrplot)
```

$$Loading Data$$

Hide

```
COVID <- read_csv("COVID19.csv")
```

Checking Data for NA values and analyzing the all the Variables

Hide

```
COVID  %>% is.na() %>% colSums()
```

```
                Country               TotalCases               TotalDeaths           TotalRecove
red              ActiveCases
                      0                        0                         0
0                       0
         SeriousCritical           Population2020   TotalCases_per_million       Deaths_per_mill
ion               TotalTests
                      0                        0                         0
0                       0
 Total tests_per_million Net_change_in_population    Yearly_change_percent                  Dens
ity            Land Area_km2
                      0                        0                         0
0                       0
            Migrants_net           Fertility_rate               Median_age Urban_Population_perc
ent      World_share_percent
                      0                        0                         0
0                       0
       First_corona_case      Recommendation_date            Lockdown_date    Recomemendation_de
lay           Lockdown_delay
                      0                        0                        19
0                      19
       Health_care_index
                      0
```

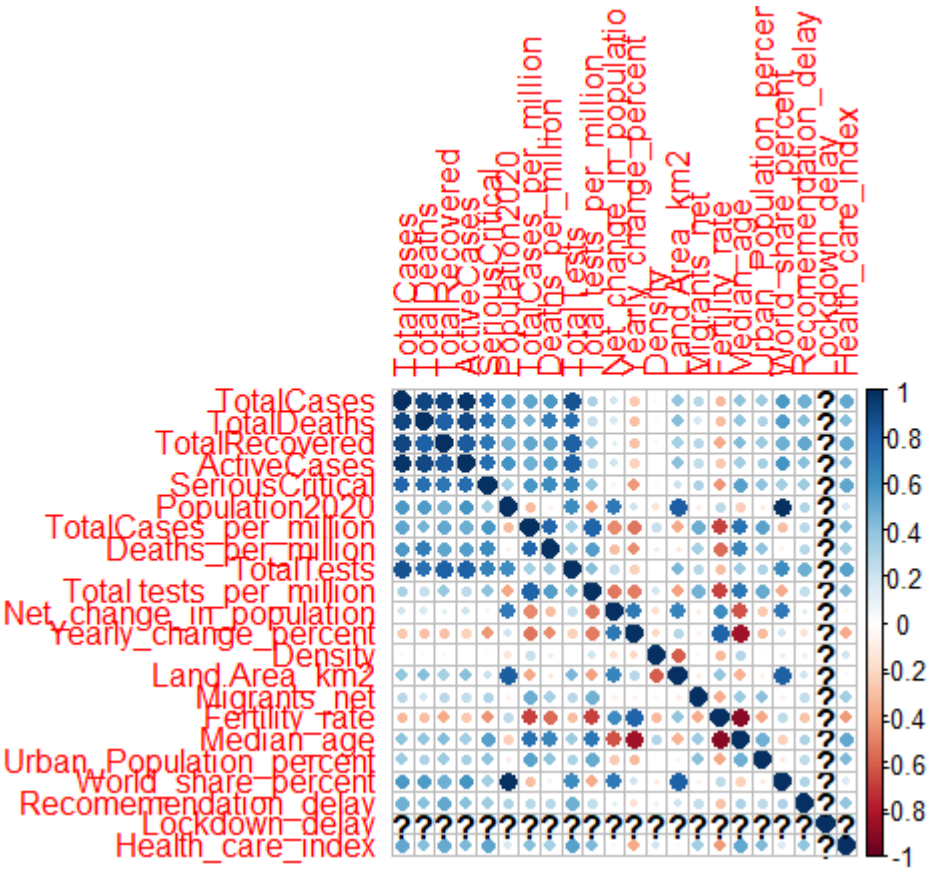
Creating a numeric table and performing corelation

Hide

```
nums <- unlist(lapply(COVID, is.numeric))
numeric <- COVID[ , nums]

numeric.rcorr = rcorr(as.matrix(numeric))
numeric.rcorr
```

Hide

```
numeric.cor = cor(numeric, method = c("spearman"))
corrplot(numeric.cor)
```



DiscriptiveAnalysis

Knowing how ADVERSELY is the world affected due to the new Pendamic COVID-19

Hide

```
sum(numeric$Population2020)
```

```
[1] 7631471065
```

Hide

```
sum(numeric$TotalCases)
```

```
[1] 2990756
```

Hide

```
sum(numeric$TotalDeaths)
```

```
[1] 206884
```

Hide

```
sum(numeric$TotalRecovered)
```

```
[1] 875643
```

Hide

```
sum(numeric$ActiveCases)
```

```
[1] 1907885
```

Hide

```
mean(numeric$Recomemendation_delay)
```

```
[1] 16.09596
```

Hide

```
range(numeric$Median_age)
```

```
[1] 15 52
```

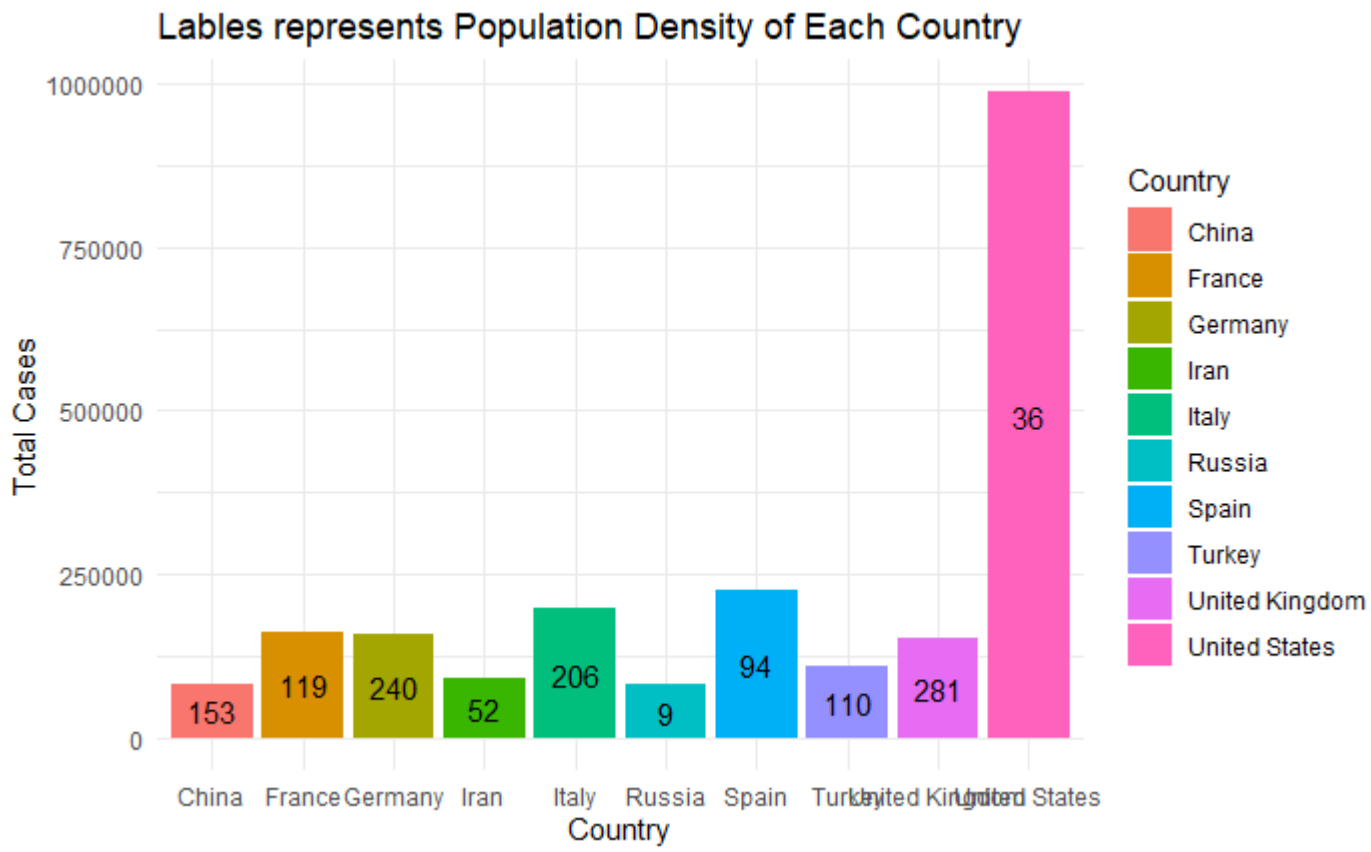
Now as we know our data well, its time to explore some insites

$$Exploratory Analysis$$

1. Which are the top 10 countries affected by COVID-19?

Hide

```
few<-COVID%>% select(Country,TotalCases,ActiveCases,Population2020,Density,TotalCases_per_millio
n,Recomemendation_delay,World_share_percent,Lockdown_delay) %>% arrange(desc(COVID$`TotalCases
`)) %>% head(10)

ggplot(few, aes(x=Country, y=TotalCases, fill=Country)) +
geom_bar(stat="identity")+theme_minimal() + geom_text(aes(label = paste0(round(Density))), posit
ion = position_stack(vjust = 0.5))+labs(title='Lables represents Population Density of Each Coun
try', x = "Country", y = "Total Cases")
```



Hide

```
NA
NA
```

Hide

```
df<-COVID%>% select(Country,TotalCases,ActiveCases,Population2020,Recomemendation_delay,Health_care_index,Lockdown_delay) %>% arrange(desc(COVID$`TotalCases`)) %>% head(10)

df
```

Country <chr>	TotalCases <dbl>	ActiveCases <dbl>	Population2020 <dbl>	Recomemendation_delay <dbl>	Health_
United States	987160	812966	331002651	56	
Spain	226629	85712	46754778	38	
Italy	197675	106103	60461826	40	
France	162100	94341	65273511	44	
Germany	157770	39794	83783942	51	
United Kingdom	152840	131764	67886011	51	
Turkey	110130	78185	84339067	44	
Iran	90481	15114	83992949	15	

Country <chr>	TotalCases <dbl>	ActiveCases <dbl>	Population2020 <dbl>	Recomemendation_delay <dbl>	Health_
China	82827	801	1408526449	67	
Russia	80949	73435	145934462	34	

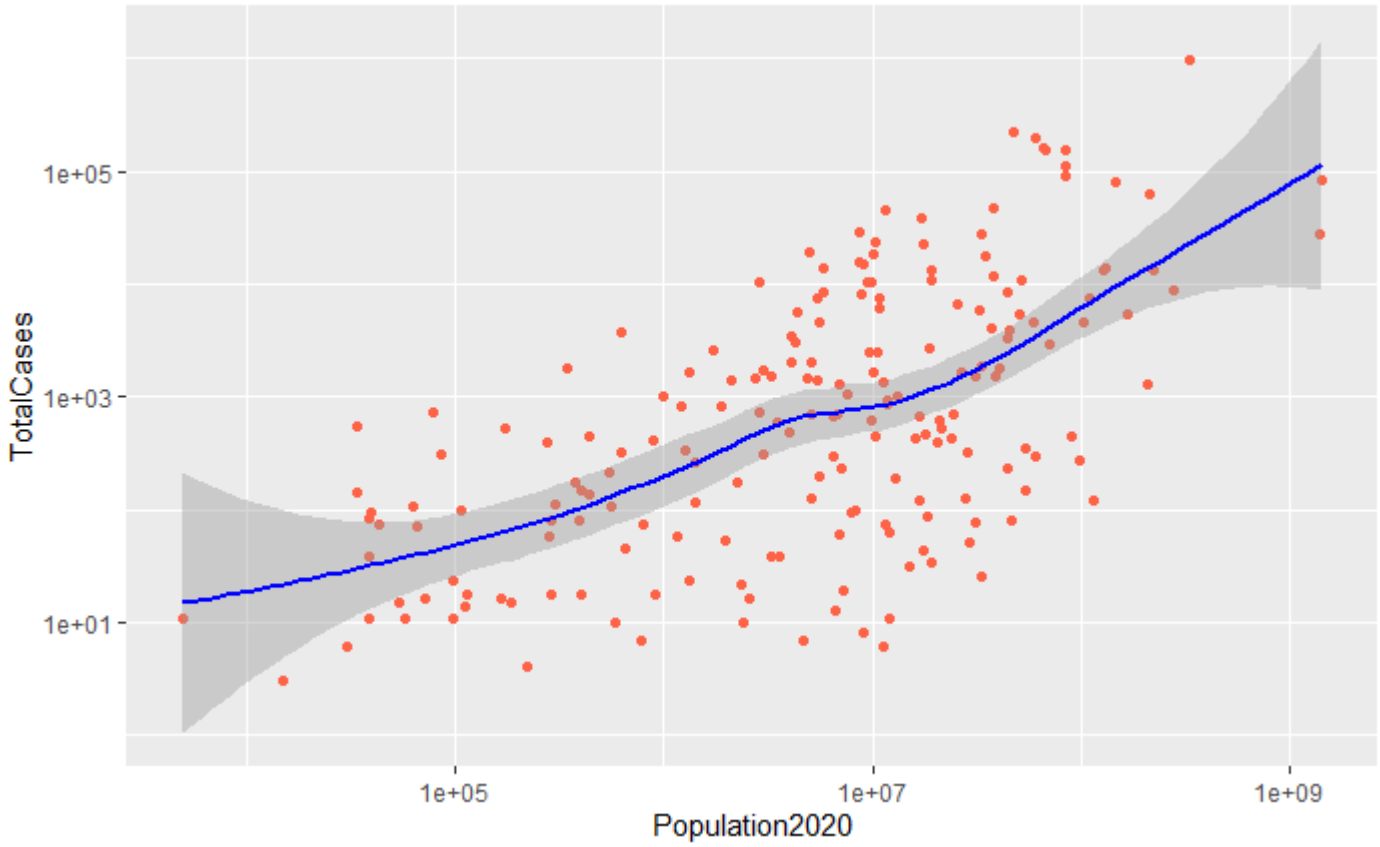
1-10 of 10 rows | 1-6 of 7 columns

-> These are the top 10 most severely affected countries.

2.How does population affect Total Number of Positive COVID-19 Cases?

Hide

```
 ggplot(data = COVID, mapping = aes(x = Population2020, y = TotalCases)) +
    geom_point(alpha = 3, color = "blue") +
    geom_jitter(alpha = 9, color = "tomato") + scale_x_log10() + scale_y_log10() + geom_smooth(c
olor ="blue")
```



-> It seems that the population of a country is not much related to this outbreak. There are some countries with higher populations which have lower cases compared to countries with lesser populations and more cases, which means that we need to explore other factors for further analysis.

3. Are there any countries that have not yet Imposed Lockdown? Which are they and How are they affected?

Hide

```
nolock<-
  COVID %>%
  filter(is.na(Lockdown_delay))
```

Hide

```
nolock %>% select(Country,Recomemendation_delay,TotalCases,ActiveCases,Median_age,TotalDeaths,Po
pulation2020)
```

Country <chr>	Recomemendation_delay <dbl>	TotalCases <dbl>	ActiveCases <dbl>	Median_a… <dbl>
Belarus	53	10463	8696	40
Cameroon	12	1621	779	19
Channel Islands	15	525	158	43
Equatorial Guinea	9	258	249	22
Hong Kong	17	1038	262	45
Iceland	16	1792	174	37
Latvia	11	812	533	44
Macao	11	45	14	39
Maldives	11	214	197	30
Réunion	18	417	117	36

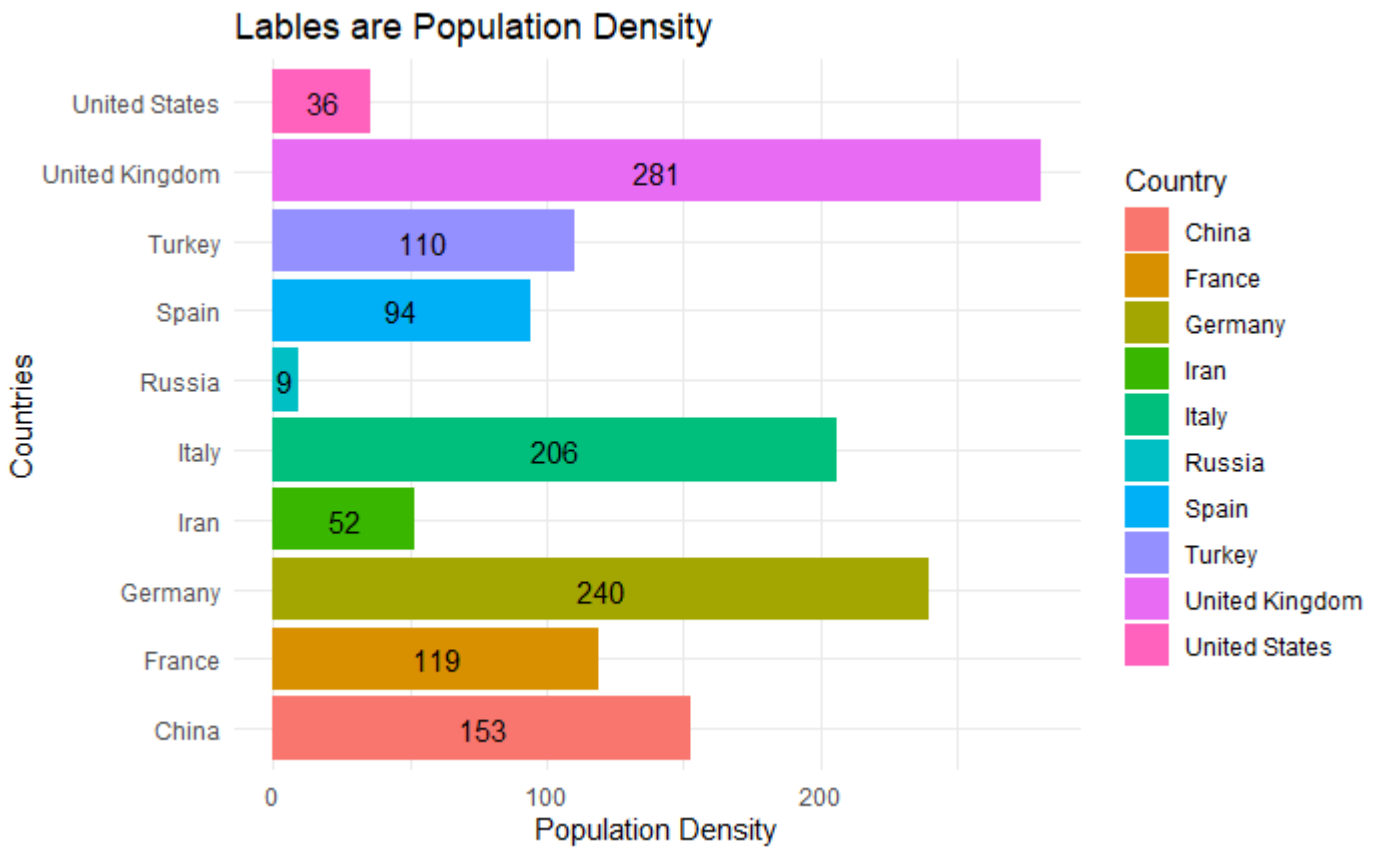
1-10 of 19 rows | 1-6 of 7 columns Previous **1** 2 Next

->There are 19 Countries that did not implement Lockdown. We get a vague observation that there are countries with less cases and also more cases which didn't implement lockdown. This points out to the fact that we need to explore other variables that might be responsible for the outbreak.

4. Is there any relation between Total cases per million and density?

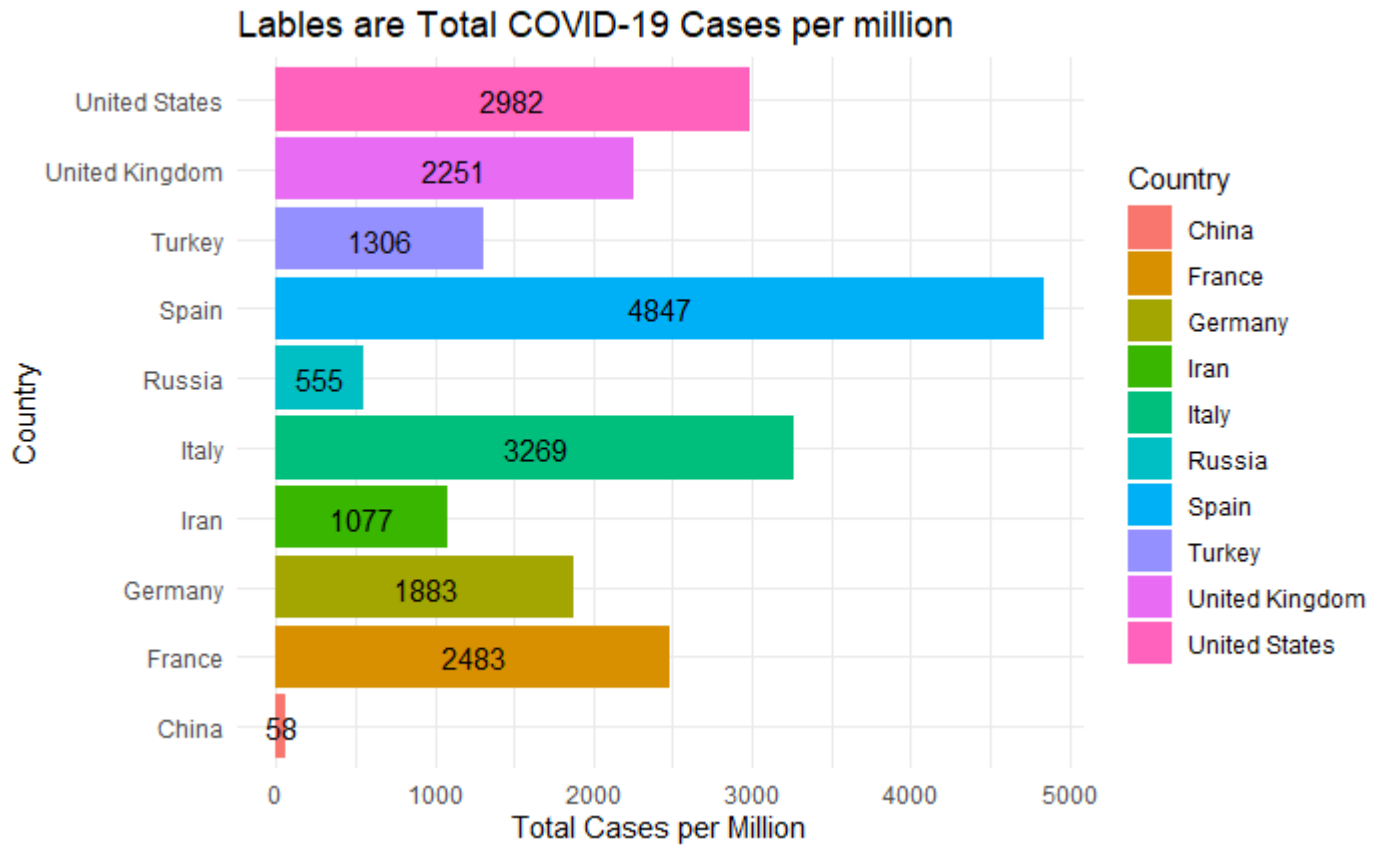
Hide

```
ggplot(few, aes(x=Density, y=Country, fill=Country)) +
geom_bar(stat="identity")+theme_minimal() + geom_text(aes(label = paste0(round(Density))), posit
ion = position_stack(vjust = 0.5))+labs(title='Lables are Population Density', x = "Population D
ensity", y = "Countries")
```



Hide

```
ggplot(few, aes(x=TotalCases_per_million, y=Country, fill=Country)) +
geom_bar(stat="identity")+theme_minimal() + geom_text(aes(label = paste0(round(TotalCases_per_mi
llion))), position = position_stack(vjust = 0.5))+labs(title='Lables are Total COVID-19 Cases pe
r million', x = "Total Cases per Million", y = "Country")
```

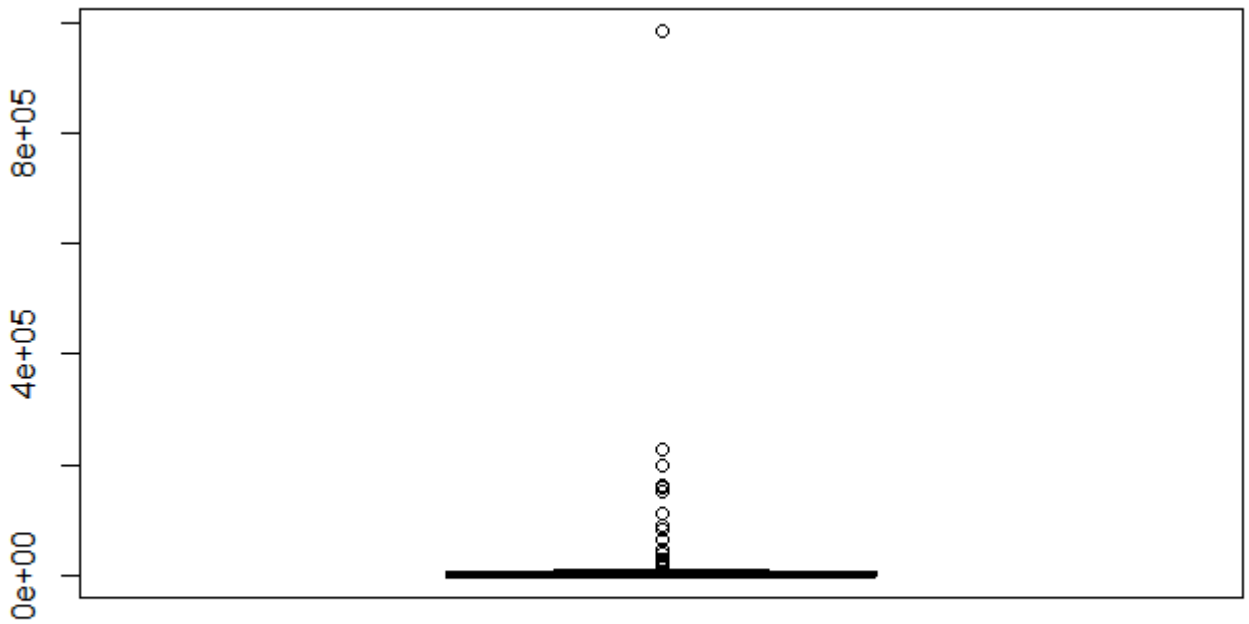


->We can see that from the graph, there is no relation between density and total cases per million. Some countries with higher population density but have lesser total cases per million and vice versa. This leads us to analyse the delay in actions against the COVID-19 outbreak which has increased the total cases in each country.

5.Are there any outliers in the Impact of COVID-19 Cases and Lockdown Delay?

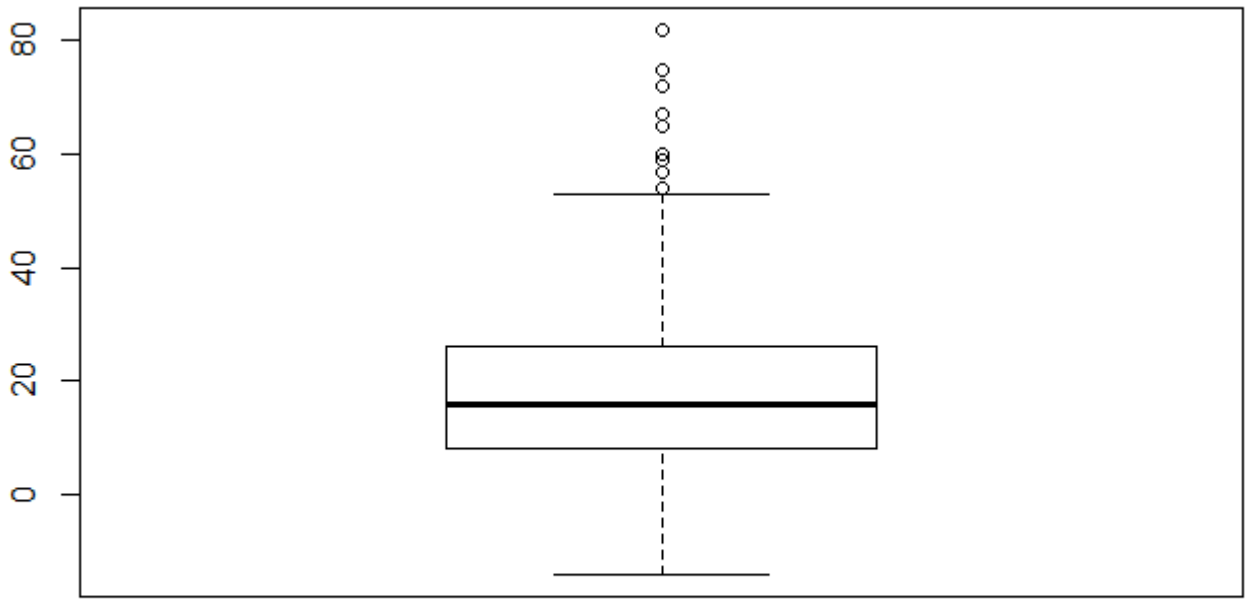
Hide

```
boxplot(numeric$TotalCases)
```



Hide

```
boxplot(numeric$Lockdown_delay)
```

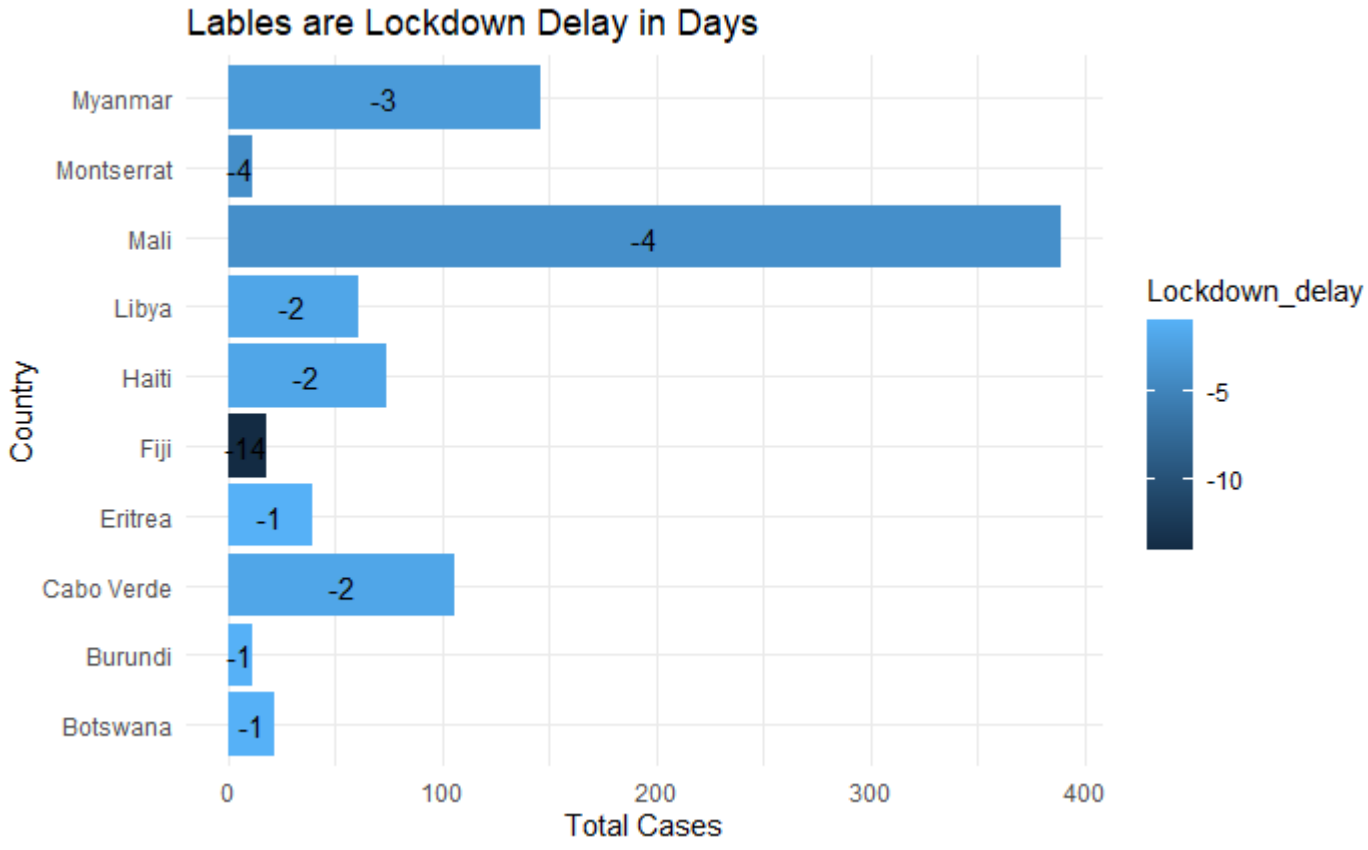


-> There are many outliers in the Lockdown delay compared to TotalCases hence we need to check the impact of some more variables.

6.How were the countries based on delay in lockdown affected by COVID-19? Is there any difference in impact?

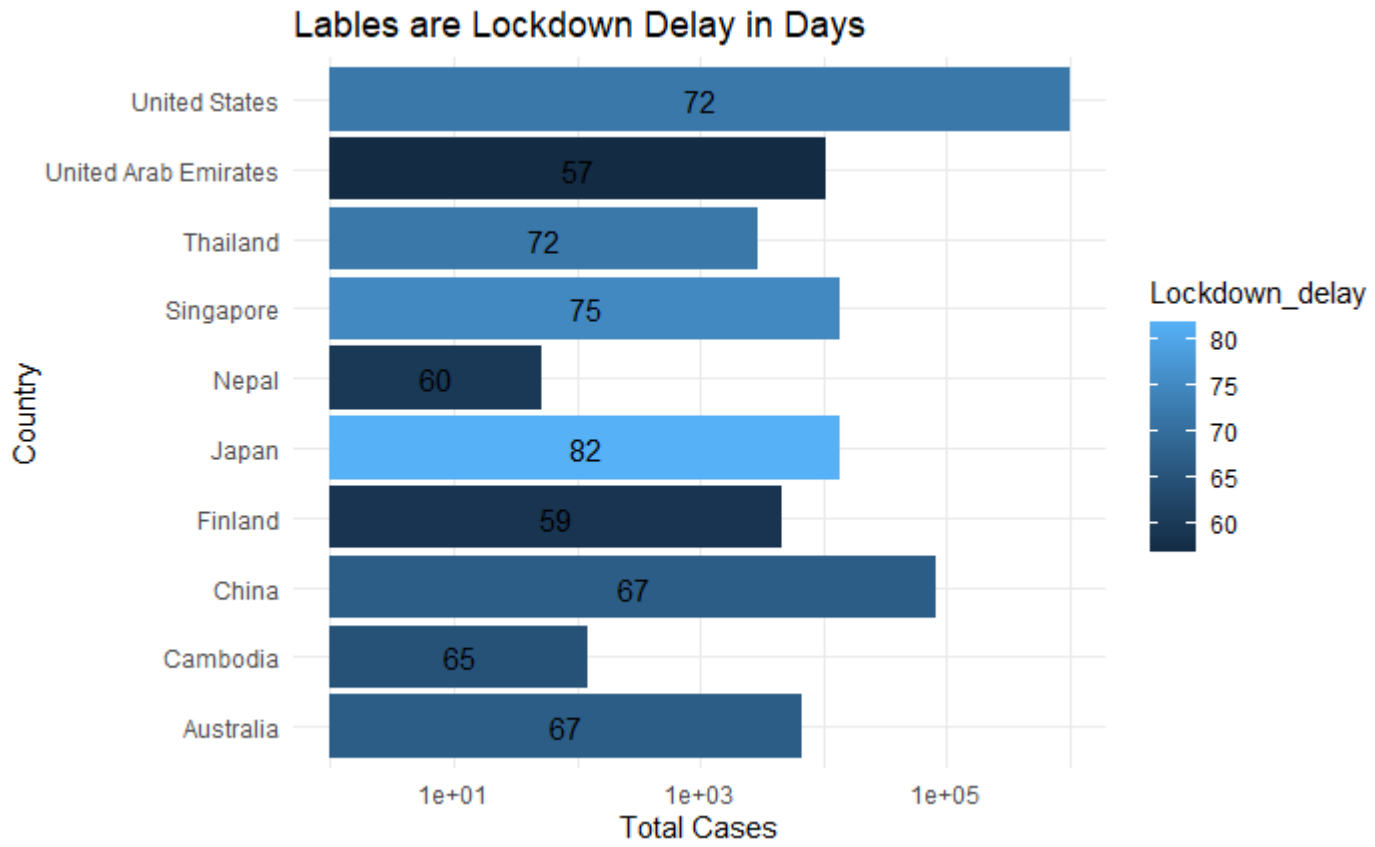
Hide

```
nld <-COVID%>% select(Country,TotalCases,ActiveCases,Population2020,Recomemendation_delay,Health
_care_index,Lockdown_delay) %>%   arrange(COVID$`Lockdown_delay`) %>% head(10)

ld<-COVID%>% select(Country,TotalCases,ActiveCases,Population2020,Recomemendation_delay,Health_c
are_index,Lockdown_delay) %>%   arrange(desc(COVID$`Lockdown_delay`)) %>% head(10)

ggplot(nld, aes(x=TotalCases , y=Country, fill=Lockdown_delay)) +
geom_bar(stat="identity")+theme_minimal() + geom_text(aes(label = paste0(round(Lockdown_dela
y))), position = position_stack(vjust = 0.5))+labs(title='Lables are Lockdown Delay in Days', x
 = "Total Cases", y = "Country")
```



Hide

```
ggplot(ld, aes(x=TotalCases, y=Country, fill=Lockdown_delay)) +
geom_bar(stat="identity")+theme_minimal() + geom_text(aes(label = paste0(round(Lockdown_dela
y))), position = position_stack(vjust = 0.5)) + scale_x_log10()+labs(title='Lables are Lockdown
 Delay in Days', x = "Total Cases", y = "Country")
```



Hide

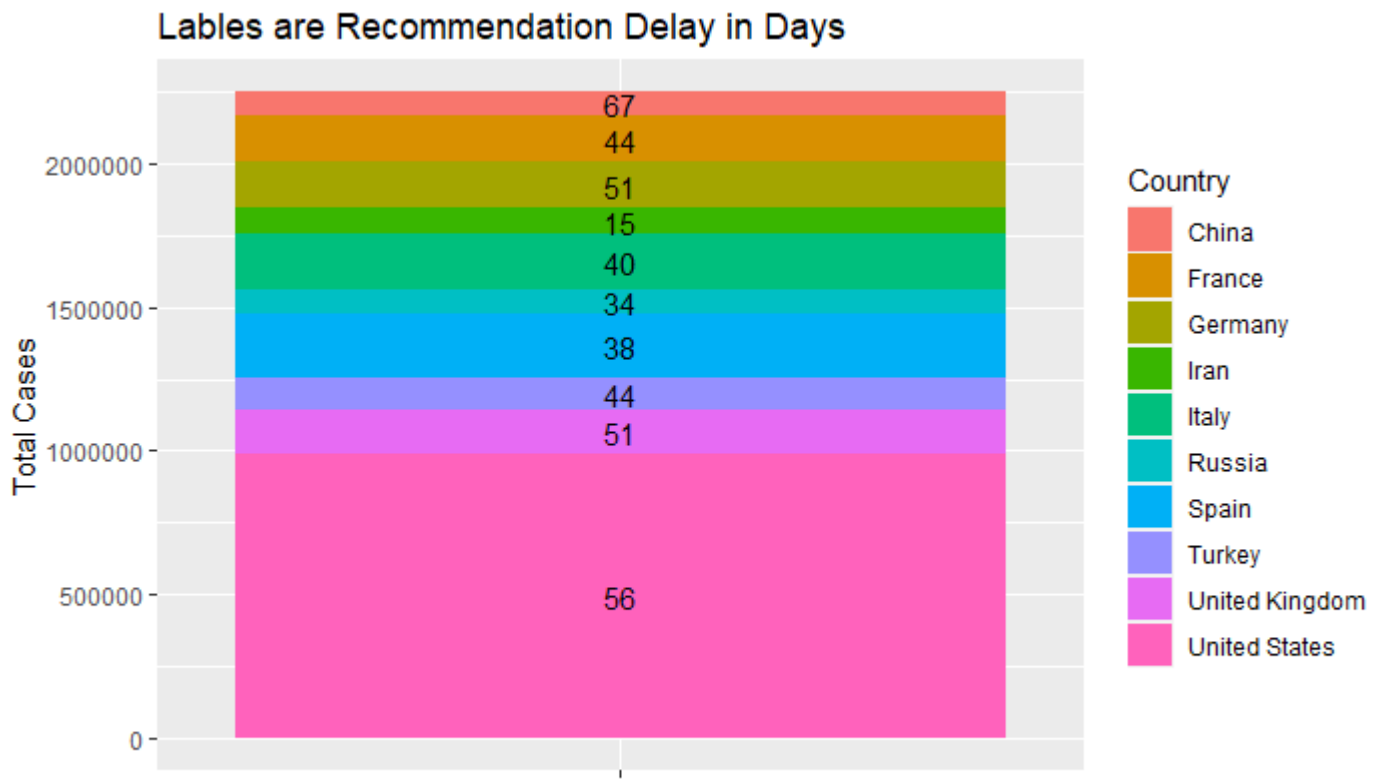
```
NA
NA
NA
```

->From the first graph, we can see that some of the countries that had imposed lockdown even before the first case was detected, the total number of cases did not increase drastically. On the contrary, from the second graph, the countries that delayed in imposing lockdown since the first corona case was detected, have a significantly higher number of total cases. This supports the analysis that the delay in lockdown can be one of the many reasons for the COVID-19 widespread.

7. Does Total Cases has any Impact due to Delay in Recommending the citizen “ways to protect themselves”?

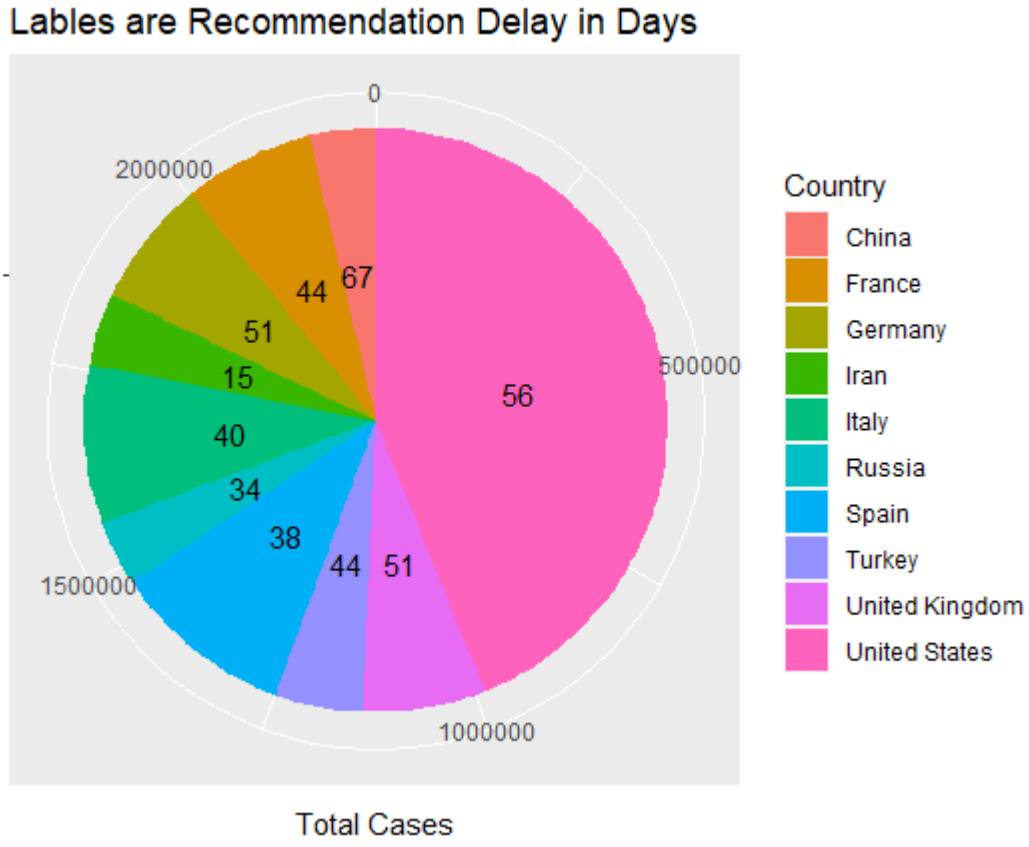
Hide

```
bp<- ggplot(few, aes(x="", y=TotalCases, fill=Country))+
geom_bar(width = 1, stat = "identity") + geom_text(aes(label = paste0(round(Recomemendation_dela
y))), position = position_stack(vjust = 0.5)) + labs(title='Lables are Recommendation Delay in D
ays', x = "", y = "Total Cases")
bp
```



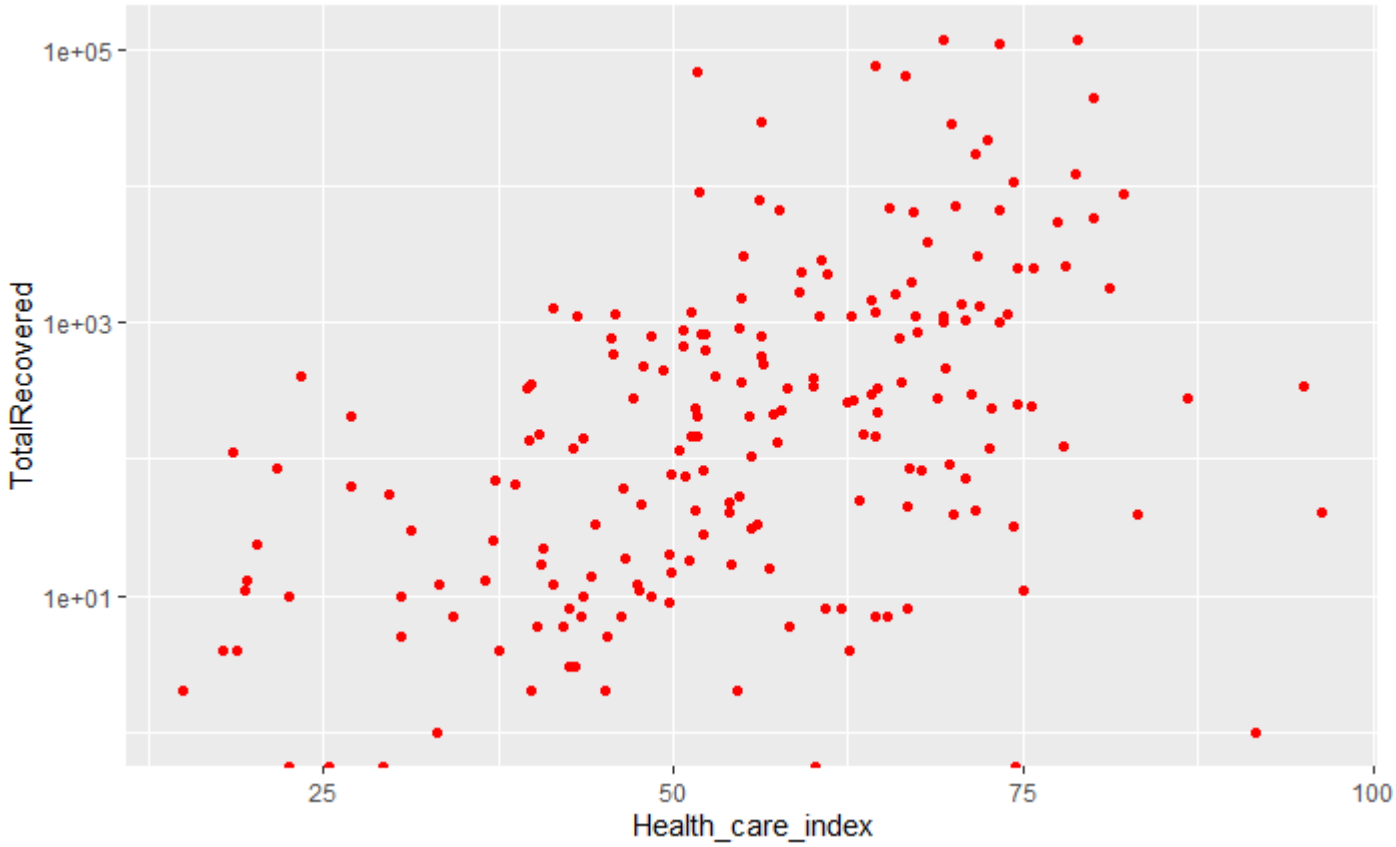
Hide

```
pie <- bp + coord_polar("y",start=0)
pie
```



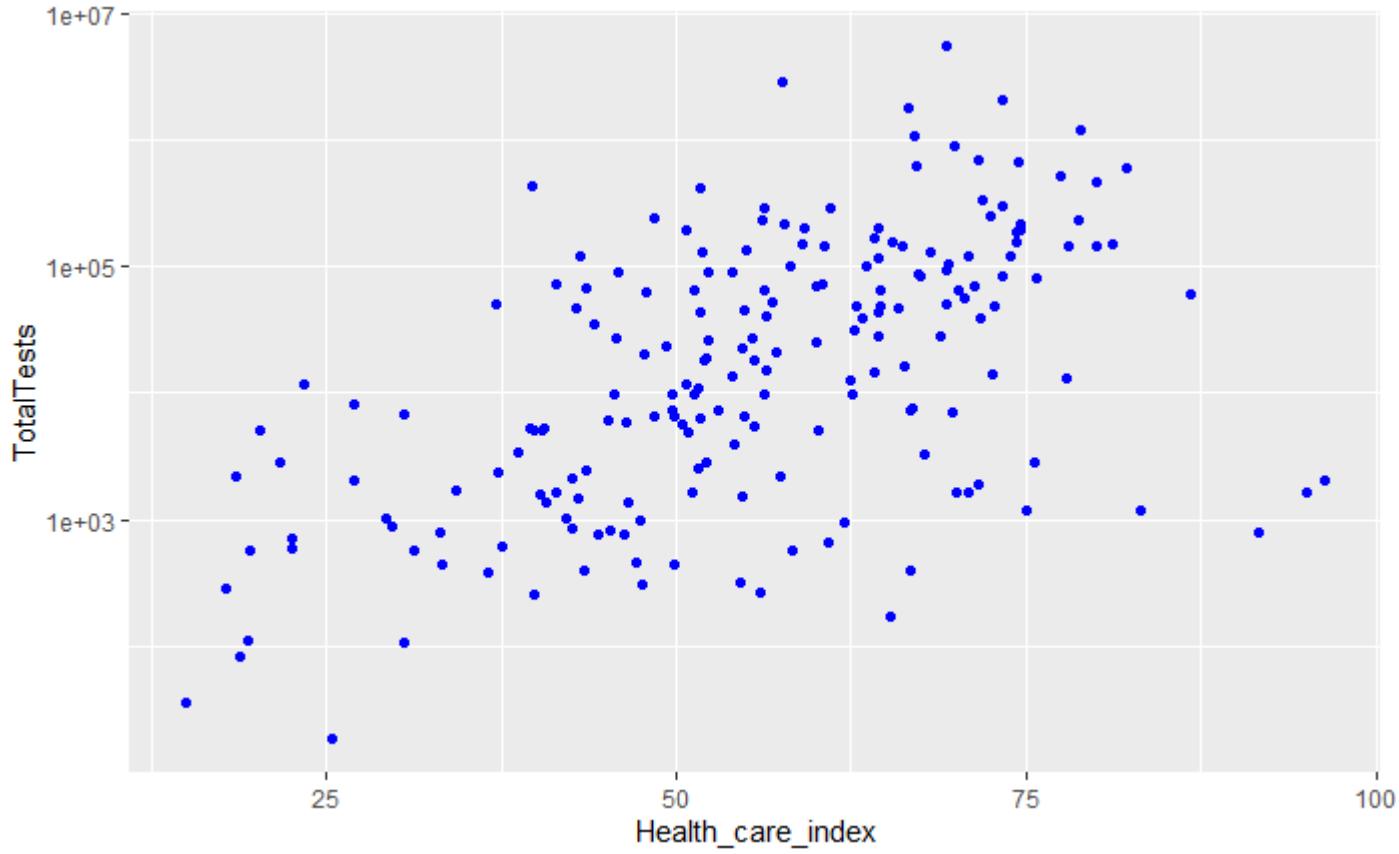
->Recommendation Delay surely has impacted the number of positive COVID-19 cases. We see some countries with very few cases having negative recommendation delay, which means before the occurence of First case, warning and preventive ways were issued in the country.

8.Does Health care Index of the countries affect their Total test taken and Recoveries



Hide

```
m <- ggplot(data = numeric, mapping = aes(x = Health_care_index, y = TotalTests)) +
        geom_point(alpha = 4, color = "blue")

m + scale_y_log10()
```

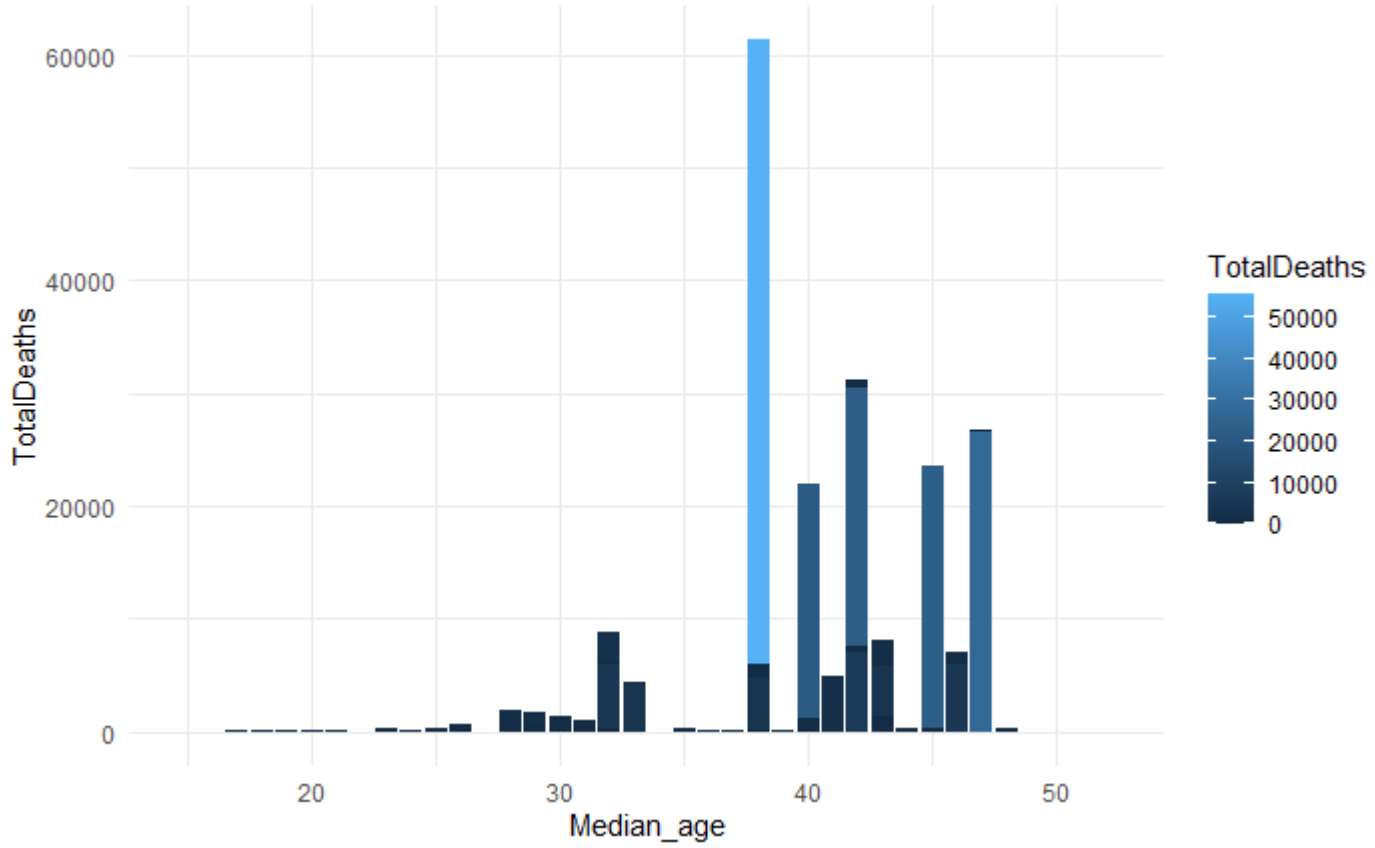


Hide

NA

-> Health care Index plays an important role for a country. Number of tests taken and total recoveries are higher in countries with higher Health care Index.

9.Does certain median age group of a country has any relation to total deaths?



->It appears that the countries who are affected most by the COVID-19 have a median age group mostly between 37 to 47. This might also be a possibility that COVID-19 affects countries with a higher age group.

Conclusion

Control of COVID-19 outbreak depends on various factors. Since the first case was detected in each country, it was supposed that they took the first line of action recommending ways for preventing the spread. If that was still not enough, further actions like national lockdown should have been imposed to implement social distancing. However, few countries who were quick in their actions were able to control the widespread as compared to the countries who delayed it.

This explains why some countries despite having a good health care index failed to control the spread as compared to other countries who had lower health care index. The total number of tests performed varied from country to country. Countries with higher health care index were able to perform more tests, which resulted in finding out more positive cases, as compared to countries with lower health care index where the total number of cases couldn't be considered accurate. With that, it can be concluded that the total number of cases detected were not entirely an accurate count for each country.

Total cases and total deaths are directly proportional to the recommendation and lockdown delays. More the delay, more COVID-19 cases were observed. COVID-19 cases were also found to be more in countries who have a median age of 37 to 47 which tells us that median age of the country might also be one of the factors for the widespread. A surprising observation can be made that the scale of the outbreak remains independent of the population of the country. As countries with higher populations but who quickly came with actions against this outbreak were able to control widespread.