Project Report

Name – Suruchi Sharma

SJSU ID-015900794

Introduction

The data set mainly focuses on other health complications with Covid-19 that led to a patient's death from the disease, that too in every state in USA. The data is classified on the basis of age group and state of residence of a person. By analyzing the above data we can deduce the most common contributing condition which worsens the health of the patient if he is suffering from Covid-19. It will also help to deduce that people in which age group suffered the most because of Covid-19.

The main reason for selecting the above data set is that Covid-19 had bad effects all over the world as far as health domain is concerned. This is why it makes me curious to study the above data and draw certain conclusions out of it that might be useful.

Data

Data source - https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group

Data collection: The data was collected between the year 2020 and 2021 when the pandemic was at it's peak. As the data tells us about causes and contributing condition that led to patient's death from covid19. It is an observational data.

Units of observations: The main units of observation in the following data set are: 1)COVID-19-DEATHS: Which gives us a count of number of people died because of Covid-19 in a particular state of residence , in a given tenure , based on various complications .

2) Number of Mentions: This attribute gives the number of mentions of people who suffered from Covid-19.

Variables: The data set includes following set of variables:

1) Start Date- Date - Start date to collect information
2) End Date- Date - End date of information collected.
3) Group - Categorical - which states data is collected by year or by month
4) Condition.Group- Categorical - Which describes different condition groups such as respiratory diseases .
5) Condition - Categorical - which describes a particular condition in a condition group
6) Covid . 19 . Deaths- Continuous - States number of people died because of Covid 19
7) Age Groups - Categorical- Describes the various age groups on which the death count is calculated
8) Number of Mentions - Continuous - Number of people had covid 19 .

9) State - Categorical - State of residence
10) Year - Categorical - Specifies the particular year in which covid 19 death count value was taken .
11) Month - Categorical- Specifies the particular month in which covid 19 death count value was taken .
12) ICD10_codes - These are just coded clinical entry for Condition Group and Condition . It will not be used in analysis .
13) Data As of- Date on which data was entered

The number of rows in data set are approximately 298000.

From the above variables the ones which I will be using for my study are :

1) Condition.Group- Categorical - Which describes different condition groups such as respiratory diseases .
2) Condition - Categorical - which describes a particular condition in a condition group
3) Covid . 19 . Deaths- Continuous - States number of people died because of Covid 19
4) Age Groups - Categorical- Describes the various age groups on which the death count is calculated
5) Number of Mentions - Continuous - Number of people had covid 19 .
6) State - Categorical - State of residence

As per the aim I mentioned in the introduction for studying the above data set the selected 6 variables will be the part of my analysis .

Data cleanup :

```
#importing libraries
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(ggplot2)
library(ggfortify)
library(corrplot)

## corrplot 0.90 loaded

library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.1.2

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

`

Initializing the working directory

```
# Get the working directory
getwd()

## [1] "C:/Users/abc/OneDrive/Documents"

data<-as_tibble(read.csv("Covid19_conditions.csv"))

head(data)

## # A tibble: 6 x 14
##   Data.As.Of Start.Date End.Date   Group     Year Month State
Condition.Group
##   <chr>      <chr>      <chr>      <chr>    <int> <int> <chr>
<chr>
## 1 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## 2 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## 3 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## 4 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## 5 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## 6 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## # ... with 6 more variables: Condition <chr>, ICD10_codes <chr>,
## #   Age.Group <chr>, COVID.19.Deaths <int>, Number.of.Mentions <int>,
## #   Flag <chr>
```

There are six more variables in the above table: Condition, ICD10_codes, Age.Group,Covid-19-deaths , Number-of-mentions, Flag.

Finding missing values

```
# Finding missing values
colSums(is.na(data))
```

```
##         Data.As.Of         Start.Date           End.Date
Group
##                  0                  0                  0
0
##               Year              Month              State
Condition.Group
##              12420              37260                  0
0
##          Condition        ICD10_codes          Age.Group
COVID.19.Deaths
##                  0                  0                  0
78295
## Number.of.Mentions               Flag
##              76087                  0
```

Removing rows where number of mentions and number of deaths are not defined because they cannot contribute in the analysis .

```
data<-filter(data , !is.na(COVID.19.Deaths) | !is.na(Number.of.Mentions))

print(data)
```

```
## # A tibble: 209,573 x 14
##    Data.As.Of Start.Date End.Date   Group     Year Month State
Condition.Group
##    <chr>      <chr>      <chr>      <chr>    <int> <int> <chr>
<chr>
##  1 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  2 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  3 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  4 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  5 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  6 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  7 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  8 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
##  9 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## 10 08/29/2021 01-01-2020 08/28/2021 By Total    NA    NA United States
Respiratory di~
## # ... with 209,563 more rows, and 6 more variables: Condition <chr>,
## #   ICD10_codes <chr>, Age.Group <chr>, COVID.19.Deaths <int>,
## #   Number.of.Mentions <int>, Flag <chr>
```

```
#Checking if any column has missing data
colSums(is.na(data))
```

```
##         Data.As.Of         Start.Date           End.Date
Group
##                  0                  0                  0
0
##               Year              Month              State
Condition.Group
##              10259              29908                  0
0
##          Condition        ICD10_codes          Age.Group
COVID.19.Deaths
##                  0                  0                  0
2208
## Number.of.Mentions               Flag
##                  0                  0
```

Removing Data As of , Start Date , End Date , Group , Year , month , flag attributes as they do not contribute in the analysis . Removing rows with missing count of Covid 19 Deaths

```
data<-dplyr::select(data,-(Data.As.Of:Month))
data<-dplyr::select(data,-(Flag))
data<-filter(data,!is.na(COVID.19.Deaths))

head(data)
```

```
## # A tibble: 6 x 7
##   State         Condition.Group  Condition ICD10_codes Age.Group
COVID.19.Deaths
##   <chr>         <chr>            <chr>     <chr>       <chr>
<int>
## 1 United States Respiratory dis~ Influenz~ J09-J18     0-24
520
## 2 United States Respiratory dis~ Influenz~ J09-J18     25-34
2348
## 3 United States Respiratory dis~ Influenz~ J09-J18     35-44
6191
## 4 United States Respiratory dis~ Influenz~ J09-J18     45-54
17515
## 5 United States Respiratory dis~ Influenz~ J09-J18     55-64
42471
## 6 United States Respiratory dis~ Influenz~ J09-J18     65-74
71361
## # ... with 1 more variable: Number.of.Mentions <int>
```

The above table has one more column Number-of-Mentions

Checking for missing values

```
#checking for missing values
colSums(is.na(data))
```

```
##              State    Condition.Group          Condition
ICD10_codes
##                  0                  0                  0
0
##          Age.Group    COVID.19.Deaths Number.of.Mentions
##                  0                  0                  0
```

Removing column of IDC10_codes as it will not contribute in the study .

```
data<-dplyr::select(data,-(ICD10_codes))
```

Removing rows where age group not stated and count for all ages

```
data<-filter(data,!(Age.Group=="Not stated"|Age.Group=='All Ages'))

print(data)
```

```
## # A tibble: 156,465 x 6
##    State         Condition.Group Condition Age.Group COVID.19.Deaths
Number.of.Menti~
##    <chr>         <chr>           <chr>     <chr>                <int>
<int>
##  1 United States Respiratory di~ Influenz~ 0-24                   520
540
##  2 United States Respiratory di~ Influenz~ 25-34                 2348
2405
##  3 United States Respiratory di~ Influenz~ 35-44                 6191
6367
##  4 United States Respiratory di~ Influenz~ 45-54                17515
18048
##  5 United States Respiratory di~ Influenz~ 55-64                42471
43677
##  6 United States Respiratory di~ Influenz~ 65-74                71361
73125
##  7 United States Respiratory di~ Influenz~ 75-84                80612
82103
##  8 United States Respiratory di~ Influenz~ 85+                  72238
73003
##  9 United States Respiratory di~ Chronic ~ 0-24                    72
74
## 10 United States Respiratory di~ Chronic ~ 25-34                  210
210
## # ... with 156,455 more rows
```

Exploratory Data Analysis

Part 1) Consider two sets of age group <65 and >=65, in the considered two sets(<=65 and >65) how may people died purely because of Covid 19 , without any other complication involved ?

Summarizing data set based on State ,Age group and Condition Group .

```
#Summarizing data based on State and Age.Group and summing Covid 19 deaths
data2<-dplyr::select(data,State,Age.Group,COVID.19.Deaths,Condition.Group)
data3<-data2 %>% group_by(State,Age.Group,Condition.Group) %>%
summarise(freq=sum(COVID.19.Deaths))

## `summarise()` has grouped output by 'State', 'Age.Group'. You can override
using the `.groups` argument.
```

Population 1 - Observations having age group above 65

```
population_1<-filter(data3,(Age.Group=='65-74'|Age.Group=='75-
84'|Age.Group=='85+'))

print(population_1)

## # A tibble: 1,944 x 4
## # Groups:   State, Age.Group [162]
##    State   Age.Group Condition.Group
freq
##    <chr>   <chr>     <chr>
<int>
##  1 Alabama 65-74     All other conditions and causes (residual)
3339
##  2 Alabama 65-74     Alzheimer disease
47
##  3 Alabama 65-74     Circulatory diseases
4924
##  4 Alabama 65-74     COVID-19
9204
##  5 Alabama 65-74     Diabetes
1067
##  6 Alabama 65-74     Intentional and unintentional injury, poisoning, and~
70
##  7 Alabama 65-74     Malignant neoplasms
366
##  8 Alabama 65-74     Obesity
171
##  9 Alabama 65-74     Renal failure
1101
## 10 Alabama 65-74     Respiratory diseases
9473
## # ... with 1,934 more rows
```

Sample 1- Considering sample from population 1 .

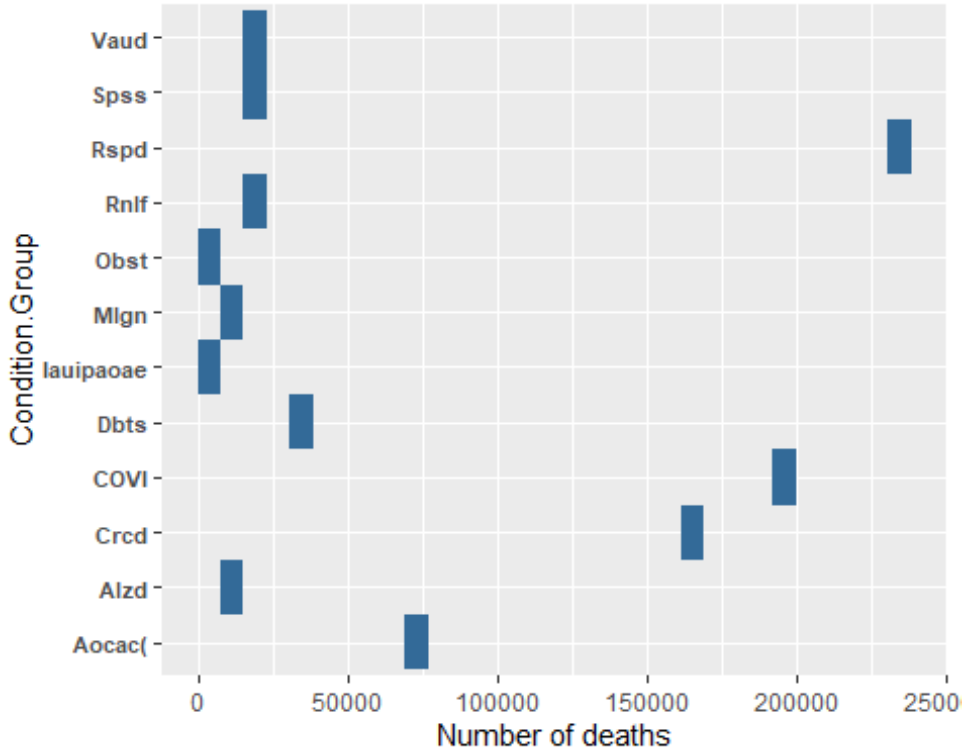
```
dataset1<-
filter(population_1,State=="Alabama"|State=="Alaska"|State=="Delaware"|State=
="California"|State=="Colorado")
```

Summarizing Number of deaths based on various states , age group , Condition Groups and calculating total deaths and deaths only because of COVID19 .

```
dataset1<-dataset1 %>% group_by(Condition.Group) %>%
summarise(freq=sum(freq))

print(abbreviate(unique(dataset1$Condition.Group)))
```

```
##                                 All other conditions and causes (residual)
##                                                                  "Aocac("
##                                                     Alzheimer disease
##                                                                   "Alzd"
##                                                 Circulatory diseases
##                                                                   "Crcd"
##                                                               COVID-19
##                                                                   "COVI"
##                                                               Diabetes
##                                                                   "Dbts"
## Intentional and unintentional injury, poisoning, and other adverse events
##                                                                "Iauipaoae"
##                                                  Malignant neoplasms
##                                                                   "Mlgn"
##                                                                Obesity
##                                                                   "Obst"
##                                                         Renal failure
##                                                                   "Rnlf"
##                                                 Respiratory diseases
##                                                                   "Rspd"
##                                                                  Sepsis
##                                                                   "Spss"
##                                          Vascular and unspecified dementia
##                                                                   "Vaud"
```

```
options(scipen=9999)
ggplot(data=dataset1)+
  geom_bin2d(mapping=aes(x=freq,y=Condition.Group))+
  theme(legend.position = "None")+theme(axis.text.y
=element_text(size=8,face="bold"))+
scale_y_discrete(label=abbreviate)+
  labs(x="Number of deaths")
```



Total Number of deaths in the given sample

```
print(sum(dataset1$freq))

## [1] 790717
```

Total Number of deaths due to covid 19 in the sample

```
d1<-filter(dataset1,Condition.Group=="COVID-19")
print(d1$freq)

## [1] 196579
```

Population 2 - Observations having age group below 65

```
population_2<-filter(data3,(Age.Group=='0-24'|Age.Group=='25-
34'|Age.Group=='35-44'|Age.Group=='45-54'|
                  Age.Group=='55-64'))

print(population_2)

## # A tibble: 3,240 x 4
## # Groups:   State, Age.Group [270]
##    State   Age.Group Condition.Group
freq
##    <chr>   <chr>     <chr>
<int>
##  1 Alabama 0-24      All other conditions and causes (residual)
```

```
0
##  2 Alabama 0-24      Alzheimer disease
0
##  3 Alabama 0-24      Circulatory diseases
0
##  4 Alabama 0-24      COVID-19
28
##  5 Alabama 0-24      Diabetes
0
##  6 Alabama 0-24      Intentional and unintentional injury, poisoning, and~
0
##  7 Alabama 0-24      Malignant neoplasms
0
##  8 Alabama 0-24      Obesity
0
##  9 Alabama 0-24      Renal failure
0
## 10 Alabama 0-24      Respiratory diseases
0
## # ... with 3,230 more rows
```

Sample 2 - Considering sample from population 2 .

```
dataset2<-
filter(population_2,State=="Alabama"|State=="Alaska"|State=="Delaware"|State=
="California"|State=="Colorado")
```

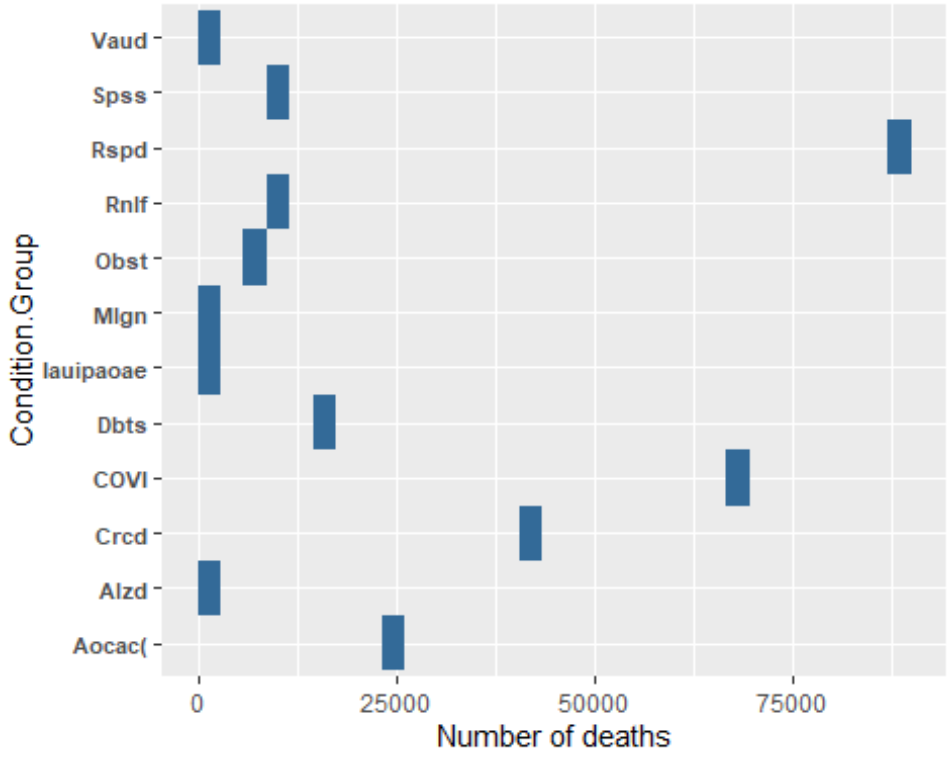
Summarizing Number of deaths based on various states , age group , Condition Groups and calculating total deaths and deaths only because of COVID19.

```
dataset2<-dataset2 %>% group_by(Condition.Group) %>%
summarise(freq=sum(freq))

print(abbreviate(unique(dataset1$Condition.Group)))
```

```
##                                All other conditions and causes (residual)
##                                                                 "Aocac("
##                                                    Alzheimer disease
##                                                                   "Alzd"
##                                                 Circulatory diseases
##                                                                   "Crcd"
##                                                              COVID-19
##                                                                   "COVI"
##                                                              Diabetes
##                                                                   "Dbts"
## Intentional and unintentional injury, poisoning, and other adverse events
##                                                                "Iauipaoae"
##                                                  Malignant neoplasms
##                                                                   "Mlgn"
##                                                               Obesity
##                                                                   "Obst"
```

```
##                                                                 Renal failure
##                                                                        "Rnlf"
##                                                           Respiratory diseases
##                                                                        "Rspd"
##                                                                       Sepsis
##                                                                        "Spss"
##                                                 Vascular and unspecified dementia
##                                                                        "Vaud"
```

```
options(scipen=9999)
ggplot(data=dataset2)+
  geom_bin2d(mapping=aes(x=freq,y=Condition.Group))+
  theme(legend.position = "None")+theme(axis.text.y
=element_text(size=8,face="bold"))+
scale_y_discrete(label=abbreviate)+
  labs(x="Number of deaths")
```



Total Number of deaths in the given sample

```
print(sum(dataset2$freq))
```

```
## [1] 271451
```

Total Number of deaths due to covid 19 in the sample

```
d2<-filter(dataset2,Condition.Group=="COVID-19")
print(d2$freq)
```

```
## [1] 69290
```

2)Out of all the people who lost their lives due to Covid 19 , most of them died only because of Covid-19 or was their any other complication involved?

```
# Summarising data set based on Condition.Group and Covid.19.Deaths
dataset1<- dplyr::select(data,Condition.Group,COVID.19.Deaths)
dataset2<-dataset1 %>%group_by(Condition.Group) %>%
summarise(total=sum(COVID.19.Deaths))
print(dataset2)
```

```
## # A tibble: 12 x 2
##    Condition.Group
total
##    <chr>
<int>
##  1 All other conditions and causes (residual)
1.46e6
##  2 Alzheimer disease
1.26e5
##  3 Circulatory diseases
2.53e6
##  4 COVID-19
3.78e6
##  5 Diabetes
5.86e5
##  6 Intentional and unintentional injury, poisoning, and other adverse ev~
6.90e4
##  7 Malignant neoplasms
1.73e5
##  8 Obesity
1.48e5
##  9 Renal failure
3.73e5
## 10 Respiratory diseases
4.12e6
## 11 Sepsis
3.56e5
## 12 Vascular and unspecified dementia
3.46e5
```

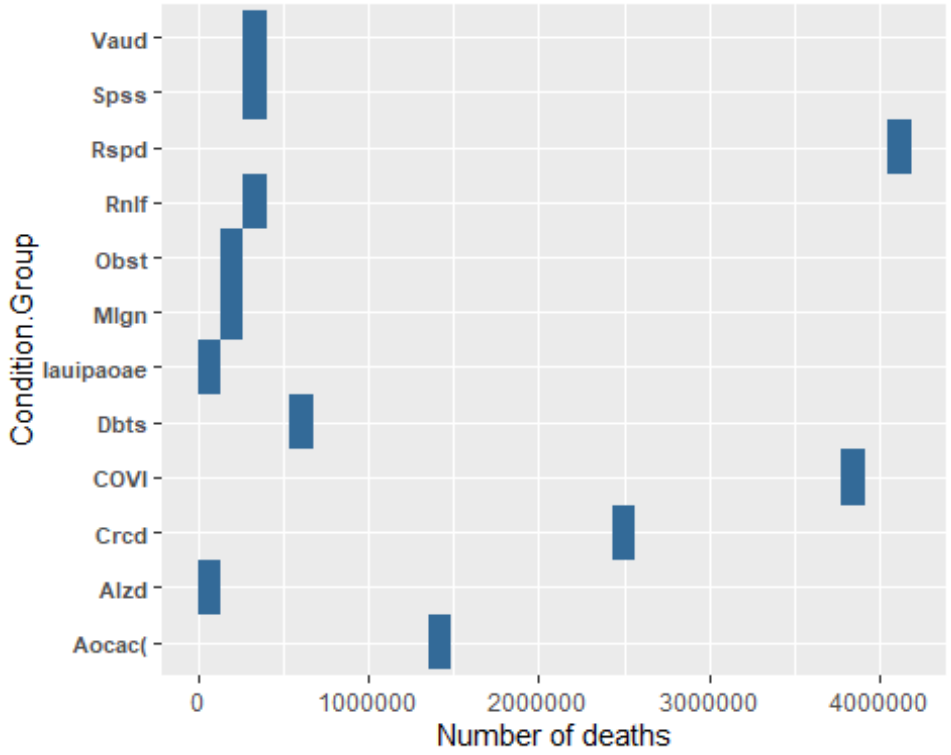
```
print(sum(dataset2$total))
```

```
## [1] 14064982
```

```
print(dataset2$Condition.Group)
```

```
##  [1] "All other conditions and causes (residual)"
##  [2] "Alzheimer disease"
##  [3] "Circulatory diseases"
##  [4] "COVID-19"
```

```
##  [5] "Diabetes"
##  [6] "Intentional and unintentional injury, poisoning, and other adverse
events"
##  [7] "Malignant neoplasms"
##  [8] "Obesity"
##  [9] "Renal failure"
## [10] "Respiratory diseases"
## [11] "Sepsis"
## [12] "Vascular and unspecified dementia"

print(abbreviate( dataset2$Condition.Group ))

##                                All other conditions and causes (residual)
##                                                                 "Aocac("
##                                                    Alzheimer disease
##                                                                  "Alzd"
##                                                 Circulatory diseases
##                                                                  "Crcd"
##                                                               COVID-19
##                                                                  "COVI"
##                                                               Diabetes
##                                                                  "Dbts"
## Intentional and unintentional injury, poisoning, and other adverse events
##                                                                "Iauipaoae"
##                                                  Malignant neoplasms
##                                                                  "Mlgn"
##                                                                 Obesity
##                                                                  "Obst"
##                                                           Renal failure
##                                                                  "Rnlf"
##                                                 Respiratory diseases
##                                                                  "Rspd"
##                                                                 Sepsis
##                                                                  "Spss"
##                                          Vascular and unspecified dementia
##                                                                  "Vaud"

#Plotting dataset2
options(scipen=9999)
ggplot(data=dataset2)+
  geom_bin2d(mapping=aes(x=total,y=Condition.Group))+
  theme(legend.position = "None")+theme(axis.text.y
=element_text(size=8,face="bold"))+
scale_y_discrete(label=abbreviate)+
  labs(x="Number of deaths")
```



Summary After summarizing the above plot we can see that the highest contributing complication with Covid 19 was respiratory diseases.

Selecting data based on two condition.group 1)Respiratory diseases

```
set1<-dplyr::select(data,Age.Group, COVID.19.Deaths,Condition.Group)
set2<-filter(set1,(Condition.Group=="Respiratory diseases" ))

print(set2)
```

```
## # A tibble: 40,490 x 3
##    Age.Group COVID.19.Deaths Condition.Group
##    <chr>               <int> <chr>
##  1 0-24                  520 Respiratory diseases
##  2 25-34                2348 Respiratory diseases
##  3 35-44                6191 Respiratory diseases
##  4 45-54               17515 Respiratory diseases
##  5 55-64               42471 Respiratory diseases
##  6 65-74               71361 Respiratory diseases
##  7 75-84               80612 Respiratory diseases
##  8 85+                 72238 Respiratory diseases
##  9 0-24                   72 Respiratory diseases
## 10 25-34                 210 Respiratory diseases
## # ... with 40,480 more rows
```

```
print(sum(set2$COVID.19.Deaths))
```

```
## [1] 4115886
```

Hypothesis Testing

Considering Statistics 1

Parameter of interest

p1 = Proportion of people in sample 1 , who died due to Covid 19 without any other complication

p2 = Proportion of people in sample 2 who died due to Covid 19 without any other complication

Null Hypothesis H0: p1 = p2

Alternate Hypothesis H1: p1 != p2

Test Statistic z0= (p1-p2)/((p* (1-p) * (1/n1+1/n2))^0.5)

x1= 196579 x2= 69290

n1= 790717

n2= 271451

p1=x1/n1 p2=x2/n2

p= (x1+x2)/(n1+n2)

Computation

```
x1<-196579
x2<-69290

n1<-790717
n2<-271451

p1<-x1/n1
p2<-x2/n2

p<-(x1+x2)/(n1+n2)

n<-(1/n1)+(1/n2)

N<-(p2-p1)
D<-(p*(1-p)*n)**0.5

print("Z0")

## [1] "Z0"

print(N/D)

## [1] 6.900138
```

```
z_al<-pnorm(0.025)
print(z_al)

## [1] 0.5099725
```

As z0 > zα/2

```
options(scipen=9999)
print("P value")

## [1] "P value"

pval<-pnorm(2*(1-6.900138))
print(pval)

## [1] 0.00000000000000000000000000001945167
```

As p-value<0.05

Conclusion

Based on two outcomes: 1) p-value<0.05 2) z0 > zα/2

We can reject the null hypothesis and conclude that p1!=p2

Thus proportion of people in sample 1, who died purely due to Covid 19 without any other complication are not same as proportion of people in sample 2 who died purely due to Covid-19 (without any other complication). In fact from the above hypothesis test we can conclude that p1> p2, so we can conclude that irrespective of having any complication, more number of people above the age of 65 died because of Covid-19 as compared to people below the age of 65. So age group plays an important role when it comes to immunity requirements against Covid-19.

2) As per calculations in question 2 of the report :

Total Number of deaths due to Covid 19 are 14064982 = n

Total Number of deaths due to Covid 19 with respiratory disease as complication= 4115886= X

As per CDC website : Number of people in USA died because of respiratory disorder in 2017-18 = 61000 Total number of deaths in US = 2,813,503

Therefore proportion of people died from respiratory infection - 61000/2813503 = 0.0217

p0=0.0217

Parameter of Interest phat= proportion of people died from Covid 19 due to respiratory disease as a complication.

Null Hypothesis H0: phat=0.0217

Alternate Hypothesis H1: phat >0.0217

Test Statistic z0= (X-np0)/((np0(1-p0))^0.5) X= 4115886 n= 14064982 p0=0.0217

Reject H0 if p-value<0.05 reject null hypothesis

Computation

Computing the test statistic.

```
 n<-14064982
 X<-4115886
 p0<-0.0217

 N<-(X-(n*p0))
 D<-((n*p0*(1-p0))**0.5)
 z_0<-N/D
 print("Z0")
```

```
## [1] "Z0"
```

```
 print(z_0)
```

```
## [1] 6973.752
```

```
zal<-qnorm(0.05)
print(z_al)
```

```
## [1] 0.5099725
```

Hence zo>zα

Computing Pnorm

```
 # for alternate hypothesis  phat>p0 pvalue = 1-pnorm(z0)

pval<-pnorm(z_0,lower.tail = FALSE)
print("p-value")
```

```
## [1] "p-value"
```

```
print(pval)
```

```
## [1] 0
```

p-value <0.05

Conclusion Based on two outcomes: 1) zo>zα 2) p-value <0.05 *we reject the null hypothesis and conclude that phat>0.0217. Hence we can deduce that, death rate due to COVID -19 with respiratory disease as a complication is greater than death rate due to simple respiratory infection. We can say that respiratory disease comorbid with COVID-19 can be harmful.*

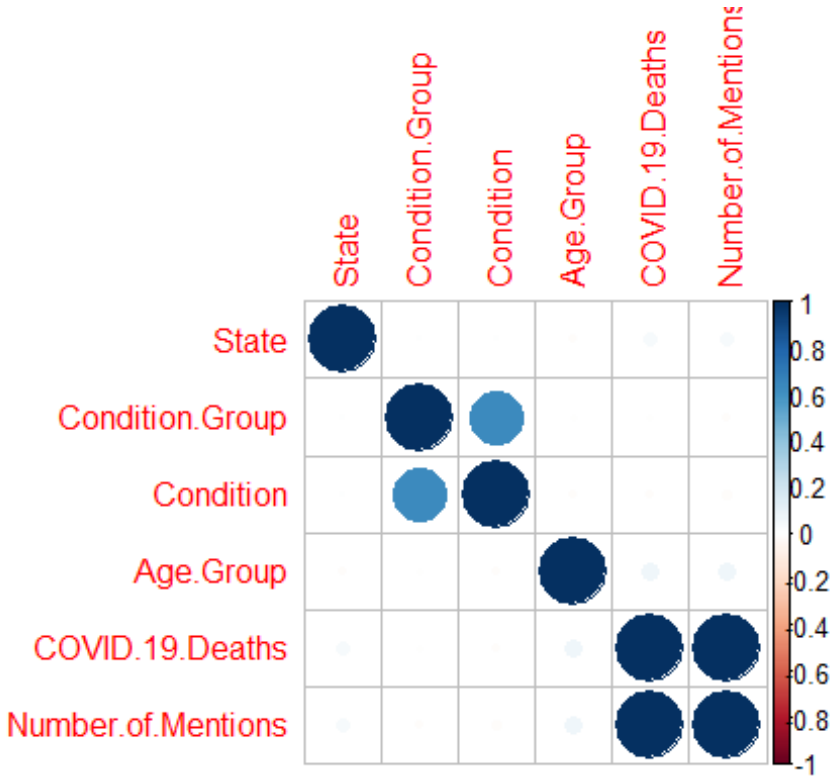
Linear Regression Analysis

```
print(data)
```

```
## # A tibble: 156,465 x 6
##    State         Condition.Group Condition Age.Group COVID.19.Deaths
Number.of.Menti~
##    <chr>         <chr>           <chr>     <chr>                <int>
<int>
##  1 United States Respiratory di~ Influenz~ 0-24                   520
540
##  2 United States Respiratory di~ Influenz~ 25-34                 2348
2405
##  3 United States Respiratory di~ Influenz~ 35-44                 6191
6367
##  4 United States Respiratory di~ Influenz~ 45-54                17515
18048
##  5 United States Respiratory di~ Influenz~ 55-64                42471
43677
##  6 United States Respiratory di~ Influenz~ 65-74                71361
73125
##  7 United States Respiratory di~ Influenz~ 75-84                80612
82103
##  8 United States Respiratory di~ Influenz~ 85+                  72238
73003
##  9 United States Respiratory di~ Chronic ~ 0-24                    72
74
## 10 United States Respiratory di~ Chronic ~ 25-34                  210
210
## # ... with 156,455 more rows
```

Out of the 6 variables that are considered for analysis , 4 are categorical and 2 are continuous . So we need to convert these categorical values into numeric values to visualize the correlation plot.

```
x<-data
x$State<-factor(x$State)
x$State<-as.numeric((x$State))

x$Condition.Group<-factor(x$Condition.Group)
x$Condition.Group<-as.numeric((x$Condition.Group))

x$Condition<-factor(x$Condition)
x$Condition<-as.numeric((x$Condition))

x$Age.Group<-factor(x$Age.Group)
x$Age.Group<-as.numeric((x$Age.Group))

corrplot(cor(x))
```



We can see from the correlation plot that attributes Condition and Condition.Group as well as Covid 19 Deaths and Number of Mentions are highly correlated. That is why I am not considering Condition and Number of Mentions attribute in regression analysis.

For applying regression technique on the data set . I will use 90 percent of observations for training the model and 10 % of observations for testing the model .

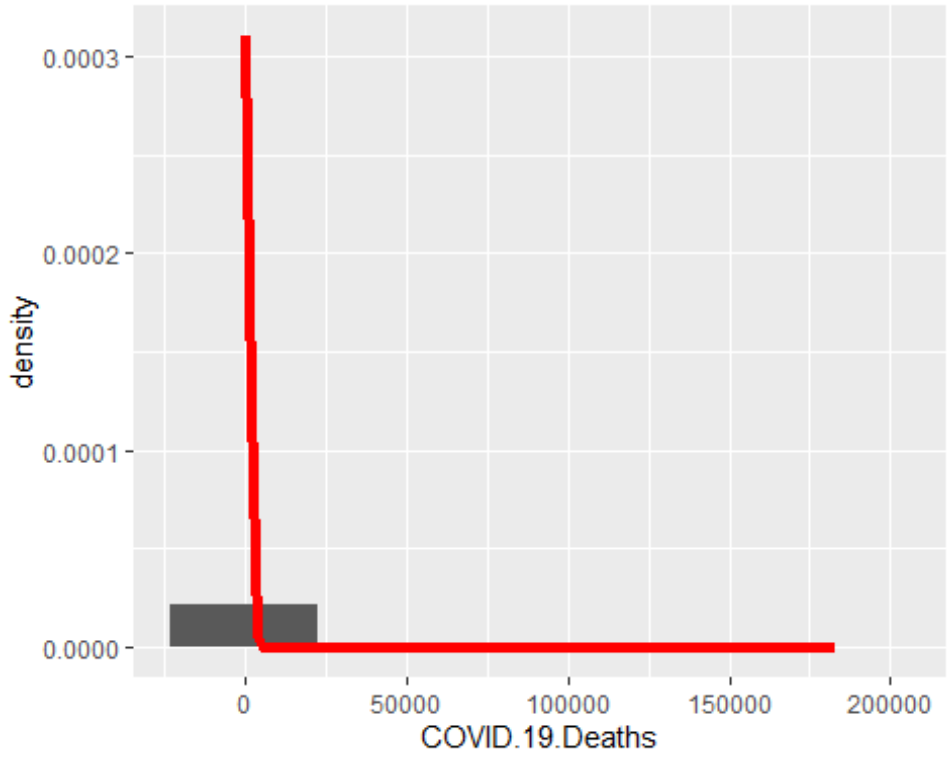
```
data<-dplyr::select(data,(-Number.of.Mentions))
data<-dplyr::select(data,(-Condition))
dt = sort(sample(nrow(data), nrow(data)*.9))
train_data<-data[dt,]
pred_data<-data[-dt,]

print(train_data)
```

```
## # A tibble: 140,818 x 4
##    State         Condition.Group      Age.Group COVID.19.Deaths
##    <chr>         <chr>                <chr>               <int>
##  1 United States Respiratory diseases 0-24                  520
##  2 United States Respiratory diseases 25-34                2348
##  3 United States Respiratory diseases 35-44                6191
##  4 United States Respiratory diseases 45-54               17515
##  5 United States Respiratory diseases 65-74               71361
##  6 United States Respiratory diseases 85+                 72238
##  7 United States Respiratory diseases 0-24                   72
##  8 United States Respiratory diseases 25-34                 210
##  9 United States Respiratory diseases 35-44                 480
```

```
## 10 United States Respiratory diseases 45-54                  1480
## # ... with 140,808 more rows
```

The prediction variable is Covid.19.Deaths - Checking if it is normally distributed or not .

```
ggplot(data =train_data, aes(COVID.19.Deaths)) + geom_histogram(mapping =
aes(x = COVID.19.Deaths, y = stat(density)), bins=5) +
   stat_function(
    fun = dnorm,
    args = list(mean = mean(data$COVID.19.Deaths), sd =
sd(data$COVID.19.Deaths)),
    lwd = 2,
    col = 'red')
```



As we can see the data is not normal , so with the help of box-cox transformation we will make the data symmetric.But before that we will try to fit the data into regression model without transfromation and see the difference.

Fitting the training data into a regression model

```
Lr_model1<-
lm(COVID.19.Deaths~as.factor(State)+as.factor(Condition.Group)+as.factor(Age.
Group),data=train_data)
```

Calculating square sum of errors and model of sum of errors

```
sse<-sum((fitted(Lr_model1)-train_data$COVID.19.Deaths)^2)
sse

## [1] 213157087260

ssr<-sum((fitted(Lr_model1)-mean(train_data$COVID.19.Deaths))^2)
ssr

## [1] 11652317181
```

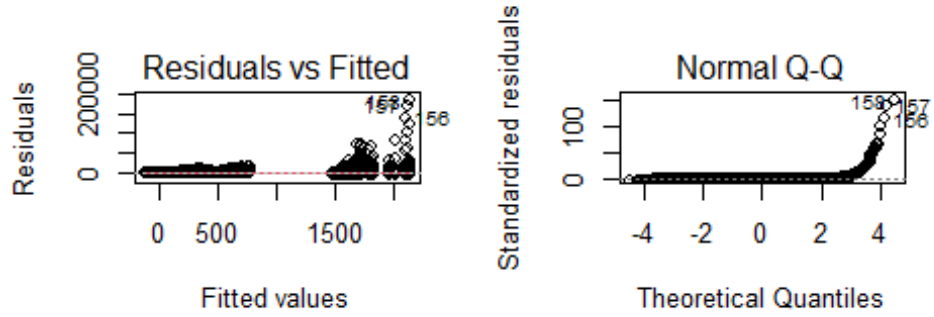
total sum of errors is sse+ssr

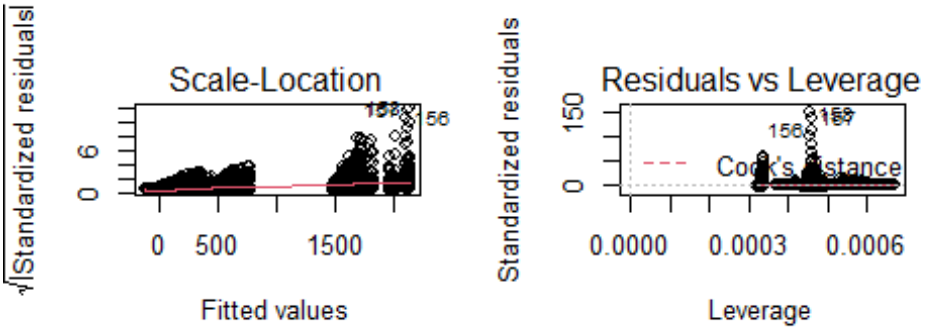
```
sst<-sse+ssr

print("R squared value for model")

## [1] "R squared value for model"

print(ssr/sst)

## [1] 0.05183198

print("Number of observations")

## [1] "Number of observations"

print(nrow(train_data))

## [1] 140818

n<-nrow(train_data)
k<-3
R_sq<-ssr/sst
Adj_Rsq<-1-(((1-R_sq)*(n-1))/(n-k-1))
print("Adjusted R square")

## [1] "Adjusted R square"

print(Adj_Rsq)

## [1] 0.05181178
```

Now from the summarized results we can see that the adjusted R squared value is 0.05 which is very less and hence the model is predictive .

Plotting the residuals for model adequacy

```
par(mfrow=c(2,2))
plot(Lr_model1)
```





The inadequacy of the regression model can also be tested using the above residual plots:

1) In plot 1-The errors do not have constant variance.

2) In the QQ plot many points are not aligned properly with the regression line hence we can conclude that the error is not normally distributed.

3) In plot 3- Errors do not have constant variance.

4) In plot 4- Many points are beyond cook's distance, in order for model to be adequate most of points should remain within cook's distance.

Let just try to fit test data into the model

```
predicted<-predict(Lr_model1,pred_data)
```

Calculating the root mean square error

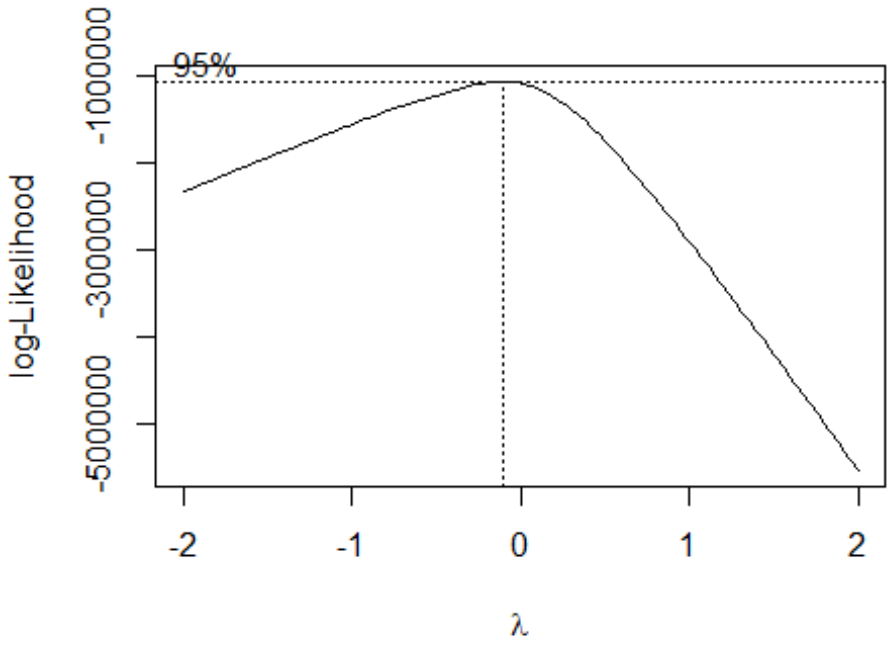
```
print(sqrt(mean((pred_data$COVID.19.Deaths - predicted)^2)))

## [1] 1387.491
```

Replacing zero values with small value as box-cox transformation does not consider zero values. The reason for considering observations having zero Covid.19.Deaths is that they are important as far as regression analysis is considered , because there are some states where pandemic has less severe effects as compared to other states.

```
train_data$COVID.19.Deaths<-
replace(train_data$COVID.19.Deaths,train_data$COVID.19.Deaths==0,0.000001)
```

Applying box cox transformation

```
box_cox<-
boxcox(COVID.19.Deaths~State+Condition.Group+Age.Group,data=train_data)
```



After applying box-cox transformation we can see data being normally distributed.

Computing the lambda value

```
lambda <- box_cox$x[which.max(box_cox$y)]

print(lambda)

## [1] -0.1010101
```

Fitting the data into regression model using the lambda value.

```
Lr_model<-lm(((COVID.19.Deaths^lambda-
1)/lambda)~as.factor(State)+as.factor(Condition.Group)+as.factor(Age.Group),d
ata=train_data)
```

```
anova(Lr_model)
```

```
## Analysis of Variance Table
##
## Response: ((COVID.19.Deaths^lambda - 1)/lambda)
##                                Df   Sum Sq Mean Sq F value
## as.factor(State)               53  6639820  125280  979.16
## as.factor(Condition.Group)     11  2537105  230646 1802.67
## as.factor(Age.Group)            7  8252064 1178866 9213.74
## Residuals                  140746 18007971     128
##                                              Pr(>F)
## as.factor(State)           < 0.00000000000000022 ***
## as.factor(Condition.Group) < 0.00000000000000022 ***
## as.factor(Age.Group)       < 0.00000000000000022 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calculating square sum of errors and model of sum of errors

```
sse1<-sum((fitted(Lr_model)-((train_data$COVID.19.Deaths^lambda-
1)/lambda))^2)
ssr1<-sum((fitted(Lr_model)-mean((train_data$COVID.19.Deaths^lambda-
1)/lambda))^2)
```

total sum of errors is sse+ssr

```
sst1<-sse1+ssr1

print("R squared value for mode")

## [1] "R squared value for mode"

print(ssr1/sst1)

## [1] 0.4918308

print("Number of observations")

## [1] "Number of observations"

print(nrow(train_data))

## [1] 140818

n<-nrow(train_data)
k<-3
R_sq<-ssr1/sst1
Adj_Rsq<-1-(((1-R_sq)*(n-1))/(n-k-1))
print("Adjusted R square")

## [1] "Adjusted R square"
```

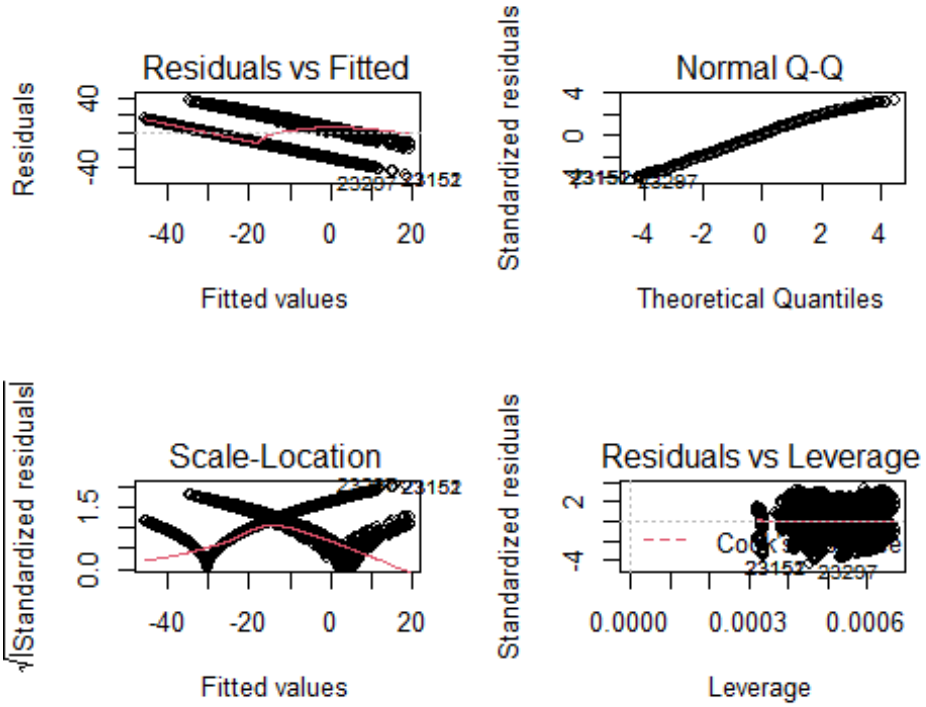
```
print(Adj_Rsq)

## [1] 0.49182
```

We can deduce that:

1) As P-value<0.05 the model is statistically different from zero

2) Adjusted R-squared value is 0.49 which is not good so the model is not predictive

```
par(mfrow=c(2,2))
plot(Lr_model)
```



From the above residual plots we can summarize that:

1)From plot 1 - variance of errors is not constant at all hence model is not adequate as per this plot

2) From plot 2 - qq plot - the quantile plot is not a straight line so the errors are not normally distributed.

3) From plot 3-Variance is not constant

4) In graph 4 - Many points are not within cook's distance .

Based on all the 4 plots one can say that the model is not adequate but still better than the previous one.

Predicting values for the test set

```
pred_data$COVID.19.Deaths<-((pred_data$COVID.19.Deaths^lambda-1)/lambda)
pred_data$COVID.19.Deaths<-
replace(pred_data$COVID.19.Deaths,pred_data$COVID.19.Deaths<0,0)
predict1<-predict(Lr_model,pred_data)
```

Calculating the root mean square error

```
print(sqrt(mean((pred_data$COVID.19.Deaths - predict1)^2)))

## [1] 21.92758
```

Finally after finishing with the regression analysis I can say that , before applying box-cox transformation the adjusted R-squared value of the model was 0.05 and after applying the box cox transformation the adjusted R-squared value increased to 0.5 , although the model became more predictive , but it is not adequate. It can be seen from the residual plots also. The performance of the model is improved which can also be seen through root mean squared error value but still there is scope for improvement. In order to improve the adequacy of the model one can try to add more attributes in the data set that are relevant for regression analysis, so that the adequacy of the model is increased. Secondly the data set consist of information related to Covid-19 deaths up to august 2021. One can try to gather information after August 2021, so that there is more data for training and prediction.

Conclusion:

The main findings , that I would like to summarize after analyzing the above set are :
a) Hypothesis Testing:

1)Age plays an important role when it comes to immunity required to fight the COVID 19 virus , because by looking at the above plots we can see that more people in aged 65 and above died because of the diseases , rather than people below the age of 65 irrespective of any complication.

2) The biggest contributing condition that led to death from covid 19 is respiratory disease. The proof for this finding is the hypothesis test which states that proportion of people who lost their lives due to covid 19 having respiratory disease as complication factor were more than people who usually die due to normal respiratory infection.

b)Regression: In regression analysis it can be observed that even after applying box-cox transformation the model was not adequate. But the adequacy can be improved by adding more relevant features or increasing the number of observations. But another question that arises is that, the data set is an observed data set, so can we extrapolate the knowledge to the observations which are not in the scope of this data set or not?

Limitations and Future Scope:

The limitation in the above study was getting a good regression model to fit the data set and predict number of Covid-19 deaths. In future scope one can try to add more attributes and observations or try to fill in the missing values, in the data set which are relevant as far as regression is concerned.

References:

1)https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_09-508.pdf

2) https://usafacts.org/articles/how-many-people-die-flu/