Data Visisualization on Covid-19 in USA

CHARLES WIREDU

5/05/2022

1. Introduction

With over 40 million confirmed cases of sever acute respiratory syndrome Coronavirus-2(SARS COV-2) infected disease called Coronavirus disease-2019 been reported by the Center for Disease Control(CDC) in the United States,there has been a significant high mortality rate recorded across age and sex.However regardless of these high numbers recorded in the United States, almost every single country in the world has recorded the existence of the infectious disease in their borders as revolving, making prevention of its spread a top priority for most governments, medical and political communities around the world.

The SARS-COVID was first detected in Wuhan, China in 2019, with the CDS subsequently confirming the case of its kind in the United states in January,2020. However since the outbreak of Covid-19, there have been several studies conducting visualization on rise of cases coupled with the high mortality rates associated with it. Whiles other studies tends to engage how to evident a scientific approach and how to develop significant research findings to flatten the curve.

With the evolving spread of the virus across states in the United States, some studies have predicted that a patients sex , age and pre-underlying health conditions might render patients vulnerable to either increase mortality or infection.The increasing mortality rate discovered in the era of the spread of COVID-19 on the entire population of the United States have been extensively researched in United Sates leading to key reports from CDC,Financial Times and Our World in Data.

In United States, Covid-19 posed a severe threat to the public health system, resulting in hospitals and health experts in hospitals been stretched to their maximum capacity early days of its discovery and spread.This however saw the introduction of string measures such as social distancing and face mask and consistent vaccination to curb the spread. Moreover, there is consideration given to age , sex and prior health conditions as acting triggers for the infection and mortality associated with Covid-19.This reports aims to employ various tools to visualize the various key variables in the data set with the aim of enhancing proper communication to the target audience on covid-19 deaths with regards to whether age or sex matters.

This report is organised as follows;

DATA OVERVIEW ========> DATA ANALYSIS/VISUALIZATION ========> CONCLUSION

2. Data Overview

The data used for visualization in this report is basically provisional and relates to covid-19 death by sex and age reported to the National Center for Health Statistics in the United States.The initial data obtained for this report contained 85375 observations and 16 total variables. The R statistical software is applied to the dataset which makes provision for biographic information such as age and sex coupled state and country level information.Below is the reading and glimpse of the intial data for this report computed using the R statistical software.

READING INTIAL DATASET

```
library(ggplot2)
library(ggpubr)
library(tidyverse)
```

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

## v tibble  3.1.6     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4

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

```
library(broom)
library(AICcmodavg)
library(colorspace)
library(pastecs)
```

```
##
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':
##
##     first, last

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(summarytools)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## library/tcltk/libs//tcltk.so'' had status 1

##
## Attaching package: 'summarytools'

## The following object is masked from 'package:tibble':
##
##     view
```

```
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
## The following object is masked from 'package:purrr':
##
##     compact
```

```
## The following object is masked from 'package:ggpubr':
##
##     mutate
```

```
library(dplyr)
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.
```

```
##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and
```

```
##       if Arial Narrow is not on your system, please see https://bit.ly/arialnarrow
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(gganimate)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
extrafont::loadfonts()
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, mutate, rename, summarise
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
Provisional_COVID_19_Deaths_by_Sex_and_Age_new_ <- read_csv("~/Desktop/Provisional_COVID-19_Deaths_by_S
```

```
## Rows: 88128 Columns: 16
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): Data As Of, Start Date, End Date, Group, State, Sex, Age Group, Foo...
## dbl (8): Year, Month, COVID-19 Deaths, Total Deaths, Pneumonia Deaths, Pneum...
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

For the purpose of this report and visualization, there are exclusion of some columns from the data set made available for this report which were are however not considered under this report. The data is however subjected to deep cleaning using the R statistical software to remove all missing values before analysis and visualization are conducted.The original data file downloaded had a size of 14.6MB in csv format. Also the initial number of observations of the data set before cleaning was implemented was 88128 coupled with 16 variables. The years recorded in the data set is 2020(start year) to 2022 (end year) Below is an overview of how the data looked before cleaning is initiated of the data file.

```
summary(Provisional_COVID_19_Deaths_by_Sex_and_Age_new_)
```

```
##   Data As Of         Start Date          End Date            Group          
##  Length:88128       Length:88128       Length:88128       Length:88128      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##       Year          Month           State               Sex           
##  Min.   :2020   Min.   : 1.000   Length:88128       Length:88128      
##  1st Qu.:2020   1st Qu.: 3.000   Class :character   Class :character  
##  Median :2021   Median : 5.500   Mode  :character   Mode  :character  
##  Mean   :2021   Mean   : 5.929                                        
##  3rd Qu.:2021   3rd Qu.: 9.000                                        
##  Max.   :2022   Max.   :12.000                                        
##  NA's   :2754   NA's   :11016                                         
##   Age Group         COVID-19 Deaths   Total Deaths     Pneumonia Deaths 
##  Length:88128       Min.   :     0   Min.   :      0   Min.   :     0.0  
##  Class :character   1st Qu.:     0   1st Qu.:     42   1st Qu.:     0.0  
##  Mode  :character   Median :    12   Median :    154   Median :    19.0  
##                     Mean   :   405   Mean   :   2799   Mean   :   380.9  
##                     3rd Qu.:    75   3rd Qu.:    682   3rd Qu.:    88.0  
##                     Max.   :987630   Max.   :7702058   Max.   :870936.0  
##                     NA's   :22149    NA's   :12928     NA's   :26375     
##  Pneumonia and COVID-19 Deaths Influenza Deaths  
##  Min.   :     0.0              Min.   :    0.000  
##  1st Qu.:     0.0              1st Qu.:    0.000  
##  Median :     0.0              Median :    0.000  
##  Mean   :   209.5              Mean   :    3.728  
##  3rd Qu.:    37.0              3rd Qu.:    0.000  
##  Max.   :511312.0              Max.   :10894.000  
##  NA's   :21762                 NA's   :15031      
##  Pneumonia, Influenza, or COVID-19 Deaths   Footnote        
##  Min.   :      0.0                        Length:88128      
##  1st Qu.:      0.0                        Class :character  
##  Median :     27.0                        Mode  :character  
##  Mean   :    584.8                                          
##  3rd Qu.:    131.0                                          
##  Max.   :1356536.0                                          
##  NA's   :25542                                              
```

However for the purpose of the project and relevance(visualization), the initial data set is subjected to rigorous cleaning, in preparation of the data set for visualization and analysis. The initial data set downloaded contained missing values which are omitted in the process of generating a clean data set coupled with the dropping of other variables which are deem to be of no interest to the purpose this project. The cleaned data set however reports 40418 observations with 10 variables. Below is an overview of the cleaned data set that will be used for data analysis and visualization.

```
COVID_19_new <- subset(Provisional_COVID_19_Deaths_by_Sex_and_Age_new_, select = -c(`Pneumonia Deaths`,
covid.data<-COVID_19_new

covid19_data<-na.omit(covid.data)

covid19_data$Sex.code <- revalue(covid19_data$Sex, c("Male"="1", "Female"="2", "All Sexes"="3"))
covid19_data$Sex.code <- mapvalues(covid19_data$Sex, from = c("Male", "Female", "All Sexes"), to = c("1

summary(covid19_data)
```

```
##   Data As Of         Start Date          End Date            State          
##  Length:40418       Length:40418       Length:40418       Length:40418      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##      Sex             Age Group         COVID-19 Deaths     Total Deaths    
##  Length:40418       Length:40418       Min.   :     0.0   Min.   :      0  
##  Class :character   Class :character   1st Qu.:     0.0   1st Qu.:     92  
##  Mode  :character   Mode  :character   Median :    49.0   Median :    537  
##                                        Mean   :   658.5   Mean   :   5079  
##                                        3rd Qu.:   184.0   3rd Qu.:   1678  
##                                        Max.   :987630.0   Max.   :7702058  
##  Pneumonia and COVID-19 Deaths Pneumonia, Influenza, or COVID-19 Deaths
##  Min.   :     0                Min.   :      0.0                       
##  1st Qu.:     0                1st Qu.:     12.0                       
##  Median :    24                Median :     78.0                       
##  Mean   :   344                Mean   :    897.9                       
##  3rd Qu.:    94                3rd Qu.:    266.0                       
##  Max.   :511312                Max.   :1356536.0                       
##    Sex.code        
##  Length:40418      
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

The above reports the processes of cleaning the data before commencing data analysis and visualization. Also there is a summary descriptive statistics of the all the key variables that will be used for visualization.There is a report of the mean, standard deviation, median, mean, skewness and kurtosis of these variables. For instance the mean for covid-19 deaths, Pneumonia and covid-19 deaths and Total death is 658.46, 344.02 and 5079.05 respectively

DATA SOURCE

The data for this project is publicly made available and can be accessed freely without a cost on the Center for Disease Control and Prevention website through this link; https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku. This data set is explored with plots for descriptive, distributions and correlations. For each plot , there is a further interpretation after writing the code. The data set comes as a result of the detection of the first covid-19 case by the CDC in the United States and in efforts to find comprehensive data on covid-19 deaths happenings in the United States in relation to age and sex.

DATA VARIABLES AND DEFINITONS

The final cleaned data set makes provision for 11 variables which will be implemented for data analysis and visualization in this project. the variables include the following;

```
names(covid19_data)
```

```
##  [1] "Data As Of"
##  [2] "Start Date"
##  [3] "End Date"
##  [4] "State"
##  [5] "Sex"
##  [6] "Age Group"
##  [7] "COVID-19 Deaths"
##  [8] "Total Deaths"
##  [9] "Pneumonia and COVID-19 Deaths"
## [10] "Pneumonia, Influenza, or COVID-19 Deaths"
## [11] "Sex.code"
```

The following details a brief description and definition of the variables mentioned above;

STATE

The states under review in this report includes all the 50 states in the united states the occurrence of covid-19 moralities accross the various hospitals in the country level.

COVID-19 DEATHS

This variable reports the moralities recorded across the united states as a result of sever acute respiratory syndrome Corona virus-2(SARS COV-2) infectious disease called Corona virus disease-2019 in the United States

SEX

This variable also defines the gender(male or female) of individual death proportions resulting from Covid-19.

TOTAL DEATH

This variable reports the total death resulting from covid-19 in the United States.

AGE GROUP

This variable defines the various age groups and the related proportions affected by the covid-19 deaths in the united states.The data set has age group ranging from under 1 to 85 years and over.

START DATE/ END DATE

These two variables defines the start and end dates of the Covid-19 as reported by the CDC in the United States.

PNUEMONIA, INFLUENZA OR COVID-19 DEATH

This variable defines the deaths reported that resulted from pnuemonia, influenza or covid-19 in the United states

DATA ANALYSIS AND VISUALIZATION

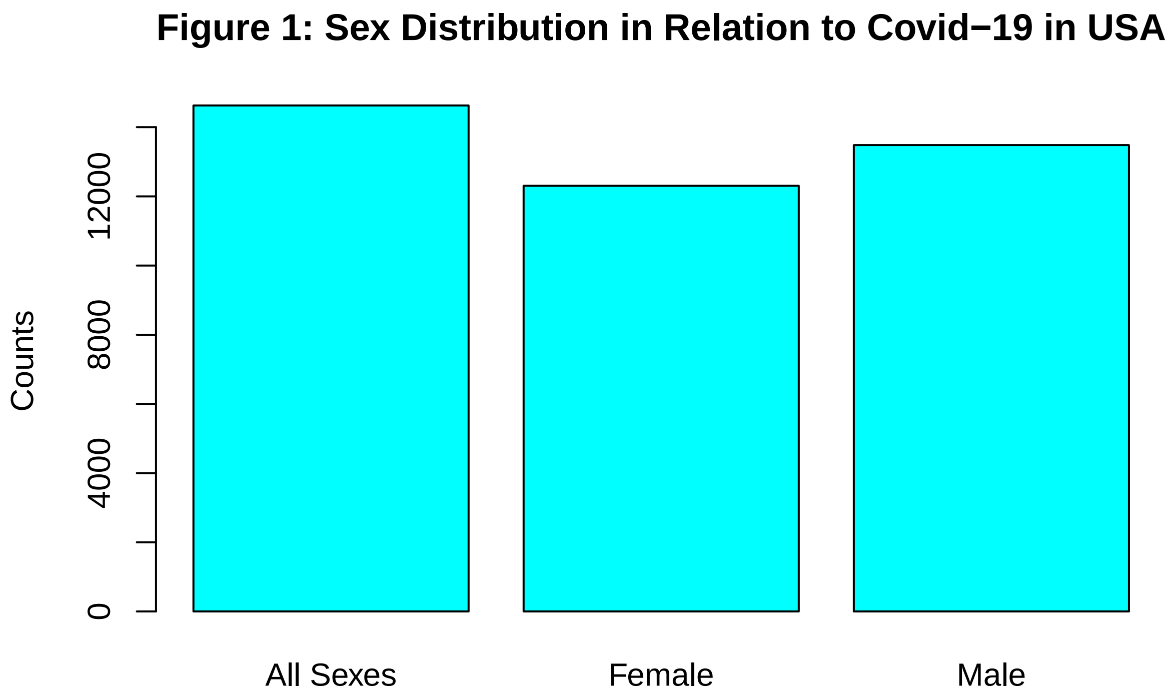
The project intend to find answers relating to Covid-19 deaths and as to whether age, sex and other factors are triggering factors contributing to the high mortality rates in the United States. As such analysis are conducted on the various key variables in the quest to find key relationships and patterns. These analysis coupled with visualizations intends to help audience understand the meaning behind the data set used for the project and also to informed their understanding on the Covid-19 death happening in the United States. Below are some analysis and visualization of the various variables used in this project;

SEX

The variable sex involves all sex, male and female.However there is proportion death relating to covid-19 death.The reports of 14629, 12308 and 13481 respectively for all sexes , female and male.

```
counts<-table(covid19_data$Sex)
```

```
barplot(counts, main="Figure 1: Sex Distribution in Relation to Covid-19 in USA", col = "cyan",ylab= "C
```



The plot above shows that males have a higher representation than females, which can be attributed to the elimination of missing data values during the cleaning stage with the R statistical software

AGE GROUPS

The project reports age groups <1 years, 1-4 years, 5-14years, 15-24 years, 25-34years, 35-64years, 65-74years, 75-84years and 85 and above years in relation to covid -19 deaths occurrence in the United States. This is however explored graphically below;

```
age_group <- c("All Ages","85 years and over","75-84 years","65-74 years","50-64 years","55-64 years","
count<- c("4562","3628","3778","3786","3540","3329","2475","2002","1670","1449","1317","1301","1342","
covid_19deaths<- c("5975708","1528042","1539635","1373203","1119107","876796","395427","253602","161737

DF<- data.frame(age_group,count,covid_19deaths)
knitr::kable(DF,caption= "Age Groups with Covid-19 Deaths.")
```

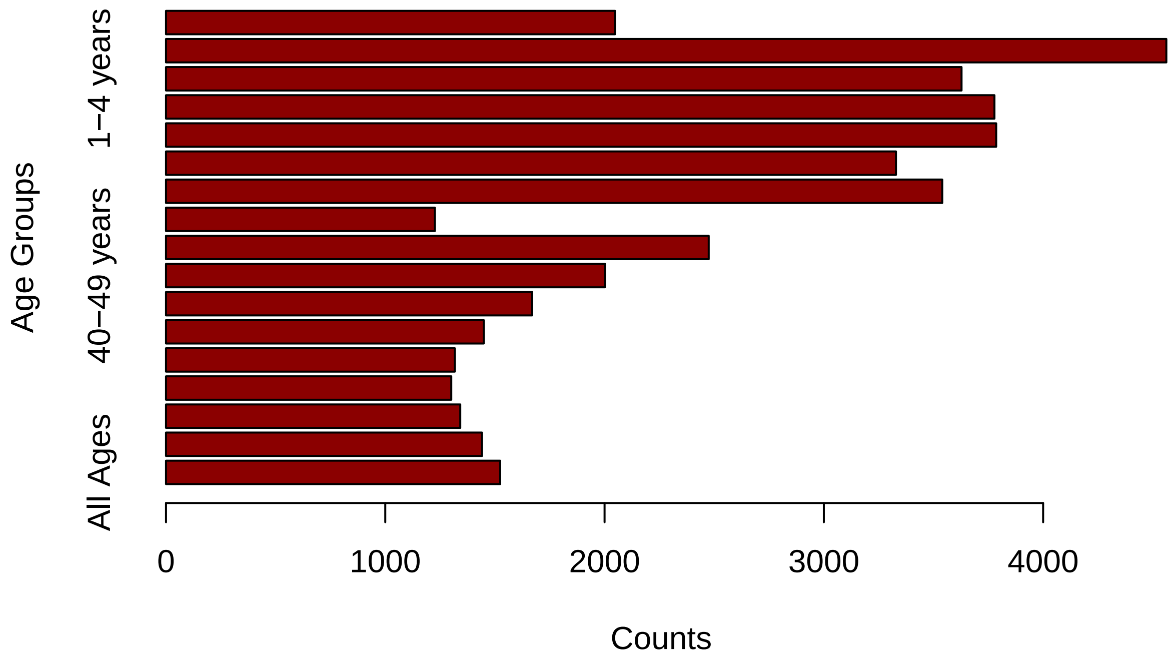
Table 1: Age Groups with Covid-19 Deaths.

age_group	count	covid_19deaths
All Ages	4562	5975708
85 years and over	3628	1528042

age_group	count	covid_19deaths
75-84 years	3778	1539635
65-74 years	3786	1373203
50-64 years	3540	1119107
55-64 years	3329	876796
45-54 years	2475	395427
40-49 years	2002	253602
35-44 years	1670	161737
30-39 years	1449	104436
25-34 years	1317	62831
18-29 years	1301	33992
15-24 years	1342	1326
5-4 years	1226	561
1-4 years	1441	13362
0-17 years	1524	6241
Under 1 year	2048	935

```
barplot(table(covid19_data$`Age Group`),
main = "Figure 2: Age Group Distribution in Relation to Covid-19 in USA",
xlab = "Counts",
ylab = "Age Groups",
names.arg = c("All Ages","85 years and over","75-84 years","65-74 years","50-64 years","55-64 years","4
col = "darkred",
horiz = TRUE)
```

Figure 2: Age Group Distribution in Relation to Covid−19 in USA



The table plot above reports the age group distribution of the data used for this reports coupled with covid 19 death in the united states. The tables further reports the covid-19 death associated with each age group affected by the pandemic. From the table it can be asserted that more deaths in the united states were reported among the aged ranging from 65 years and upwards. This however confirms high mortality rate among this age group during the peak time of the pandemic.

COVID-19 DEATH

This variable details moralities associated with the infectious disease. The United States happens to be one of the countries deeply hurt by this disease, with the country seeing high rate of moralities accross all states.This resulted in the stretch of pressure on the health care system in the United States during its peak time. The provisional death count is based on death certificate data recieved by the Center for Health Statistics occurring within all the 50 States. This is visually reported coupled with other control variables such age and sex using the R statistical software.

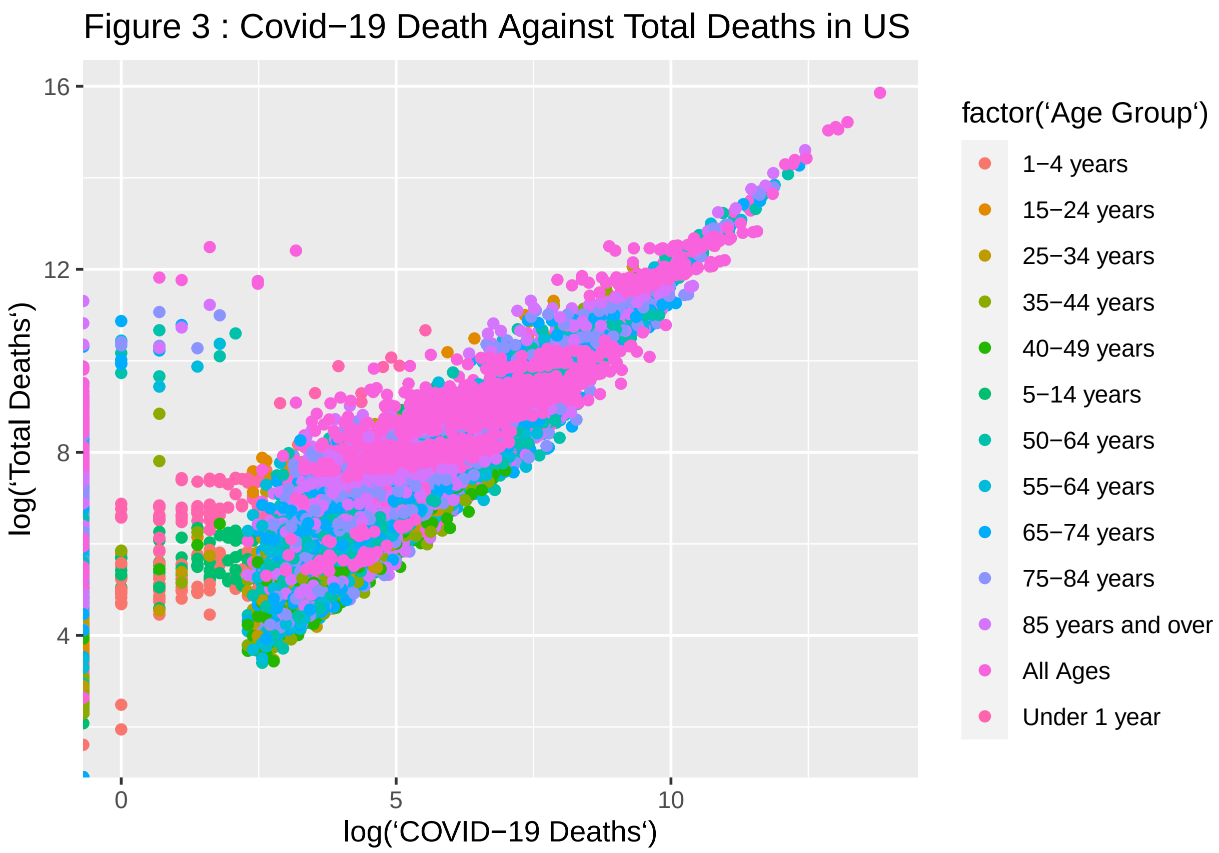
```
age_data <- unique(covid19_data$`Age Group`)
age_data <- age_data[-c(3, 7, 9, 12)]
gender_data <- unique(covid19_data$Sex)
View(gender_data)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/Resources/
## modules/R_de.so'' had status 1
```

```
covid19_data <- covid19_data %>%
  filter(
    State!= 'Puerto Rico',
    `Age Group` %in% age_data
  )
total_deaths <- covid19_data %>%
  filter(`Total Deaths`  == 'All Ages')
#view(`Total Deaths` )
covid_new <- na.omit(covid19_data)
#names(covid_new)
#head(covid_new)


P<-ggplot(covid_new, aes(y=log(`Total Deaths`), x= log(`COVID-19 Deaths`)) ) +
  geom_point(aes(color= factor(`Age Group`)))

P+ labs(title= "Figure 3 : Covid-19 Death Against Total Deaths in US")
```



From the above visual illustration, the data shows an uphill trend towards the right portion of the graph which indicates a positive trend. That's there is a positive linear relationship between Covid-19 deaths and total death. This can however be asserted that an increase in covid-19 deaths makes impacts or affects the total deaths happenings in the United States with regards to the respective age groups.

TOTAL DEATH BY AGE GROUP AND SEX

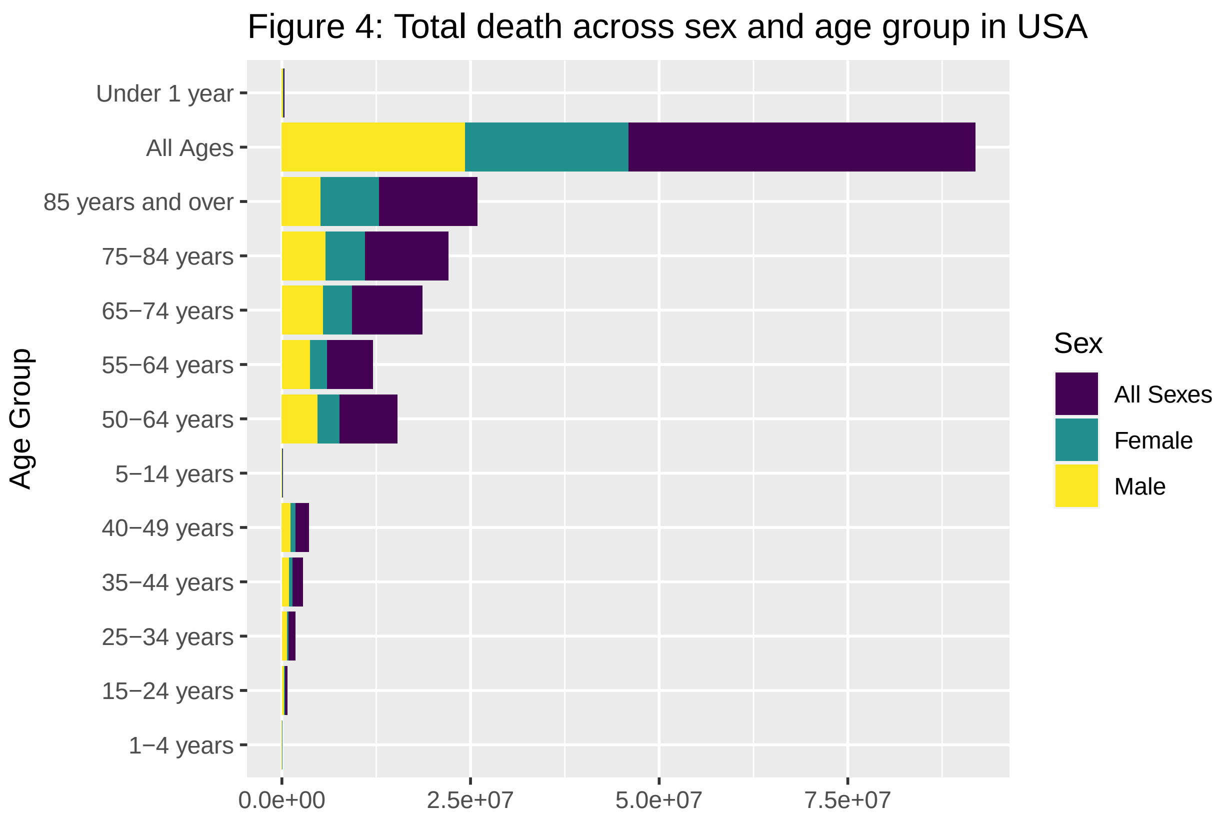
The CDC prior to its(covid-19) first detection in the United States had sounded a warning call that there will be more deaths recorded across the country, resulting in State mandates across the country to curb the spread. The reported mean from the data set is 5079.05 and a standard deviation of 63841.65. Below is graphical illustration of total death from covid-19 in relation to age group and sex.

```
library(rlang)
```

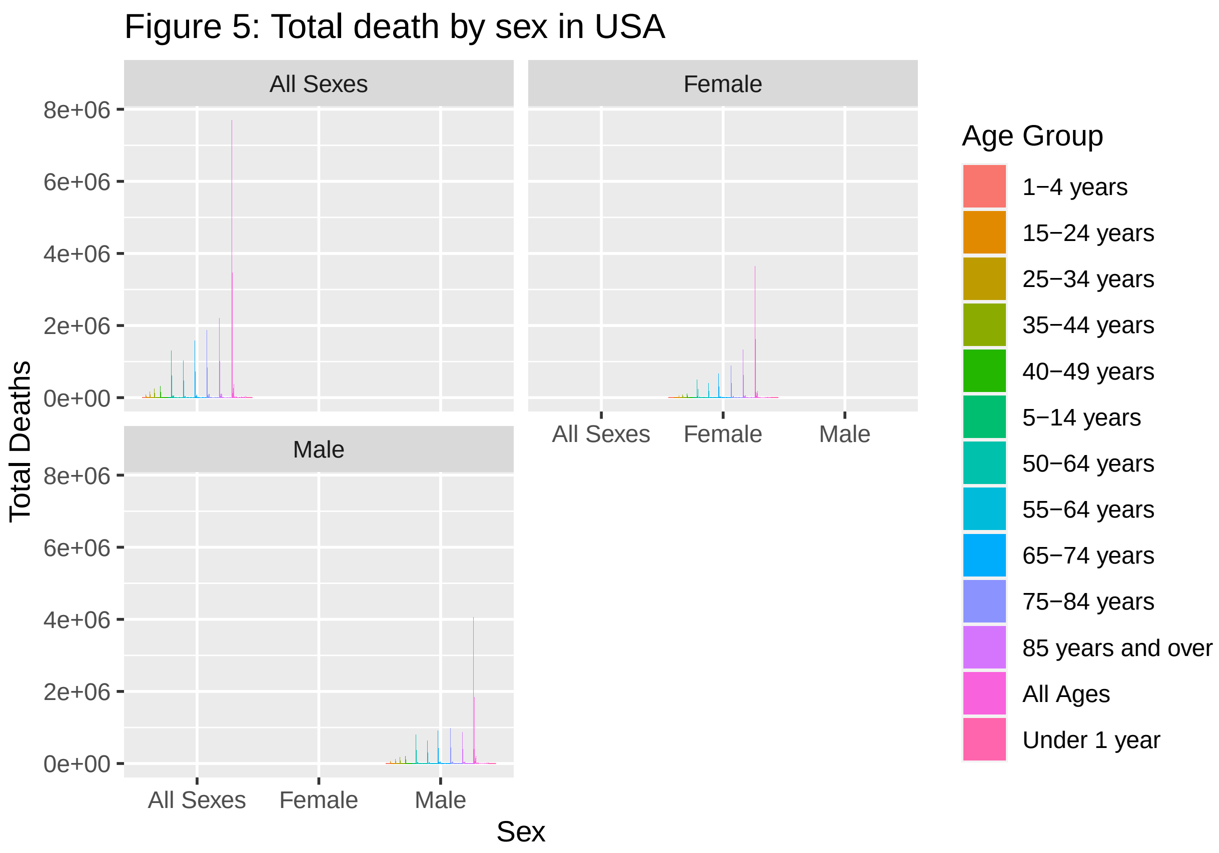
```
##
## Attaching package: 'rlang'

## The following objects are masked from 'package:purrr':
##
##     %@%, as_function, flatten, flatten_chr, flatten_dbl, flatten_int,
##     flatten_lgl, flatten_raw, invoke, splice
```

```
ggplot(covid_new, aes(fill=Sex, y=`Age Group`, x=`Total Deaths`)) +
  geom_bar(position="stack", stat="identity") +
  scale_fill_viridis(discrete = T) +
  ggtitle("Figure 4: Total death across sex and age group in USA") +
  xlab("")
```



```
ggplot(covid_new,aes(x = Sex, y = `Total Deaths`,fill=`Age Group`)) +
  geom_bar(position="dodge2", stat="identity") +
  ggtitle("Figure 5: Total death by sex in USA") +
  facet_wrap(~Sex, nrow = 2)
```



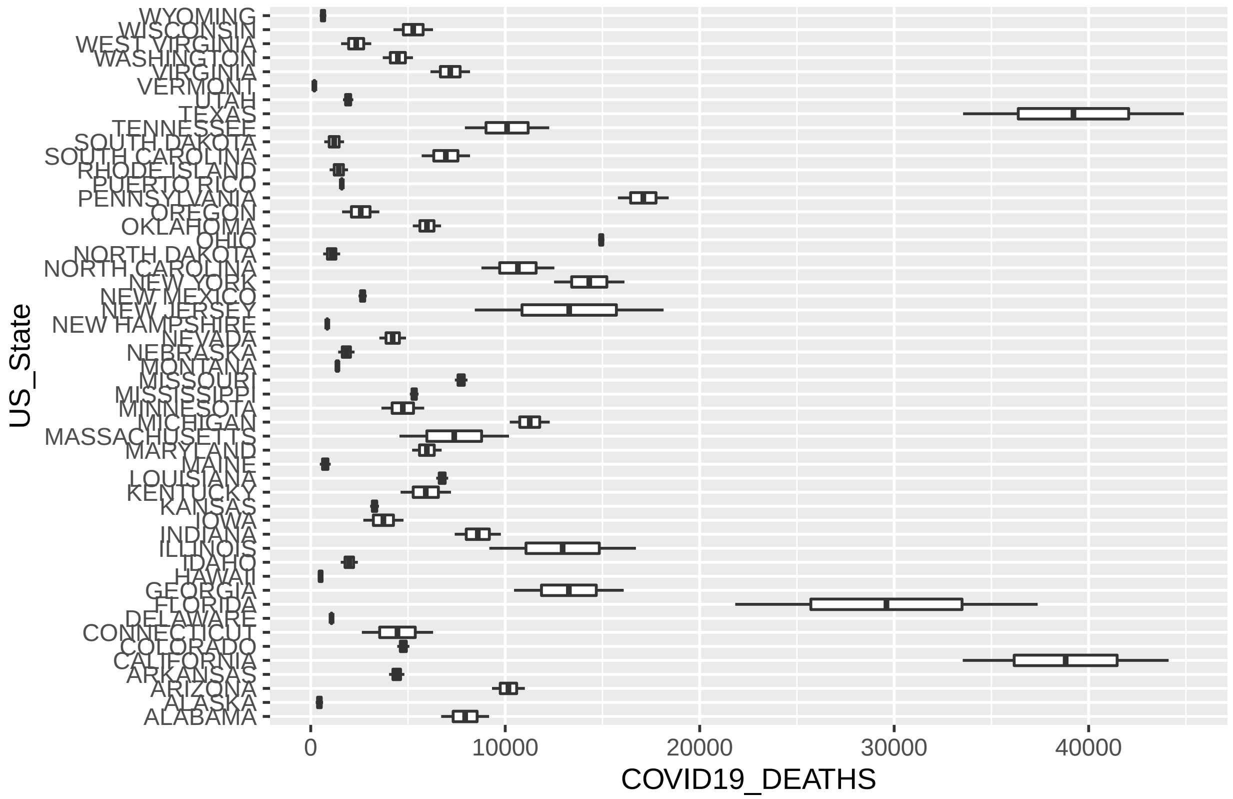
The plots above shows the total death happenings across age group and sex(male and female).However the second also serves as visualization of total death resulting from the infectious disease(COVID-19) in a group plot in relation to sex and age group.

STATES AND COVID-19 DEATHS

All states within the United States reported fatalities occurring because of the severe acute respiratory syndrome Corona virus-2(SARS-COV-2) infected disease called Coronavirus-19. This was documented ranging from the hospitals to senior assisted living homes within each and every state. This however resulted to most states such as New York imposing lock-down restrictions and state emergency order on its residents. Below is plot to visualize states and covid-19 deaths occurrences.

```
library(readxl)
 COVID19_STATES <- read_excel("~/Desktop/COVID19_STATES.xlsx")
ggplot(COVID19_STATES, aes(x=COVID19_DEATHS, y=US_State, fill=YEAR, stat = "identity")) +
ggtitle("Figure 6: US STATES PER COVID-19 DEATHS OCCURRENCE") +
geom_boxplot()
```

Figure 6: US STATES PER COVID−19 DEATHS OCCURRENC



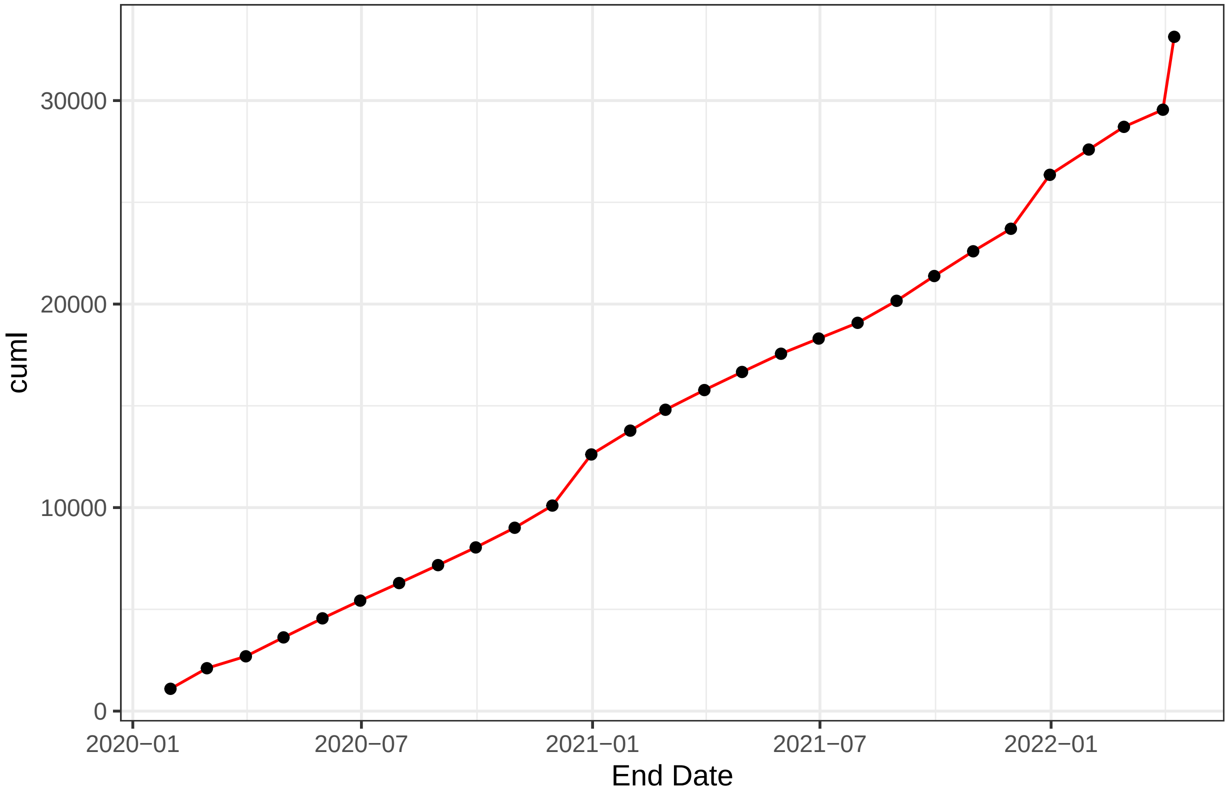
CUMMULATIVE COVID-19 CASES BY YEAR END DATES

The Covid-19 pandemic as projected by the CDC in the united states had its peak sessions with high record of cases across all the 50 states in the United States. The following plot reports the graphical illustration of the cumulative cases of covid-19 by the year end dates since it was discovered in the United States.

```
datanew<- covid_new %>%
  mutate(`End Date`= mdy(`End Date`))
require(dplyr)
require(plyr)

##Animated time series
datanew %>% group_by(`End Date`) %>%
  dplyr::summarise(count= n()) %>%
  mutate(cuml = cumsum(count)) %>%
  ggplot(aes(x=`End Date`, y= cuml)) +
  geom_line(color= "red") +
  geom_point(size= 1.5) +
  theme_bw()+
  ggtitle('Figure 7: Cummulative Cases In United States')
```

Figure 7: Cummulative Cases In United States



From the above plot there is a report on the cumulative number of cases with regards to the year 2021 and beyond. The plot tends to move uphill positively indicating the high record of numbers and moralities that happens afterwards.

TABLE SUMMARY OF DATASET

Below is a table(first five) of the dataset used for the visualization in this report

```
knitr::kable(head(covid_new[,3: 8]), caption= "Covid-19 Death by Age and Sex.")
```

Table 2: Covid-19 Death by Age and Sex.

End Date	State	Sex	Age Group	COVID-19 Deaths	Total Deaths
04/09/2022	United States	All Sexes	All Ages	987630	7702058
04/09/2022	United States	All Sexes	Under 1 year	253	42943
04/09/2022	United States	All Sexes	1-4 years	123	7997
04/09/2022	United States	All Sexes	5-14 years	312	12750
04/09/2022	United States	All Sexes	15-24 years	2605	81690
04/09/2022	United States	All Sexes	25-34 years	10988	172691

CONCLUSION

From the above plots, there has been exhibition of visualization of the Provisional dataset on Covid-19 death by age and sex made available by the Center for Disease Control.These visualization ranges from bar graph to line graph to explore the dataset.In relation to figure(1), it explains the representation of sex groups used in this report coupled with mortalities associated. Also from figure(3), it can also be asserted that there is an uphill trend towards the right signifying a positive relationship between the two variables. Moreover, it can also be observed that there has been a rapid increase in the spread of the infectious disease(Covid-19) as reported by figure(7). That's there is a steep rise in the line plot towards which shows more cases recorded in the United States since its first detection. In the nutshell, the above visualization explore all the necessary techniques with explanations that communicates to readers of this report to enhance the understanding of covid-19 deaths in relation to sex and age.