# Final Project - Covid-19 Analysis

Kesav Adithya Venkidusamy

11/16/2021

## *Introduction*

Coronavirus disease or COVID-19 is a global pandemic infectious disease caused by virus called sars-cov-2. Most people infected with the virus will experience mild to moderate respiratory illness and recover without requiring special treatment. However, some will become seriously ill and require medical attention. Older people and those with underlying medical conditions like cardiovascular disease, diabetes, chronic respiratory disease, or cancer are most likely to develop serious complications from COVID-19 illness.

The center for disease control and prevention (CDC) is the national public health agency of the United States. The agency's main goal is the protection of public health and safety through control and prevention of disease, injury, and disability in US and worldwide. CDC plays an essential role in the response to COVID-19. The agency collects the data on regular basis and provide for public use. Among numerous datasets available in CDC, below are the ones considered for analysis

1. Provisional COVID-19 deaths by sex and age
2. Provisional COVID-19 deaths by week, sex and age
3. Conditions contributing to COVID-19 deaths, by state and age, provisional 2020-21

The COVID-19 pandemic also pushed many companies to develop new vaccines to minimize the severity of symptoms. These vaccines were developed rapidly and underwent clinical trials rigorous enough to meet FDA (Food and Drug Administration) requirement for emergency use. The government played a role in monitoring the adverse reactions of these newly developed vaccines with Vaccine Adverse Event Report System (VAERS). VARES is co-managed by the Central for Disease Control and Prevention (CDC) and U.S Food and Drug Administration (FDA).

VARES accepts reports from people who have received vaccines and experienced adverse effects or from heathcare providers who are required by law to report:

1. Any adverse event listed in the VARES table of reportable events following vaccination that occurs within the specified time period after vaccinations
2. An adverse event listed by the vaccine manufacturer as a contradiction to further doses of vaccine

VARES data is accessible by two mechanisms: by downloading raw data in comma-separated values (CSV) files for import into a database, spreadsheet or, by use of CDC WONDER online search tool. For this project, below datasets from VARES is considered.

1. VARESDATA.csv
2. VARESVAX.csv
3. VARESSYMPTOMS.csv

## *The problem statement you addressed*

The problem that I addressed is Covid-19 impact by age and people with underlying conditions. I analyzed if the deaths caused by Covid-19 virus is high for the older people having age greater than 55 compared to the young people whose age is less than 55. I also analyzed the impact of Covid-19 and deaths caused to the people with underlying condition like diabetes, blood pressure and stroke compared to the people who don't have any underlying conditions. Then, I analyzed Covid-19 vaccine data to see if it has any impact in controlling deaths.

## *How you addressed this problem statement*

I addressed the problem statement as follows:

- Loading CDC (Covid-19 deaths) and VAERS (vaccines) data sets for analysis
- Cleaning the data sets
- Splitting and merging the data sets
- Adding additional variables derived from existing variables to the data sets
- Slicing and dicing the data sets
- Viewing various metrics and graphs to perform the analysis

## *Analysis*

### *Loading the r libraries required for the analysis*

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(broom)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(coefplot)
library(knitr)
library(tidyr)
```

## *Loading data sets for analysis*

### CDC Datasets

```
Numerous Covid-19 related datasets are available for public use in CDC website. Those datasets feature c
```

Provisional COVID-19 deaths by week, sex and age + Data as of – Date of Analysis + State - Jurisdiction of occurrence + MMWR Week – MMWR week number + End Week - Last week-ending date of data period + Sex - Sex + Age Group - Age group + Total Deaths – Deaths from all causes of deaths + COVID-19 Deaths - Deaths Involving COVID-19

Conditions contributing to COVID-19 deaths, by state and age, provisional 2020-21 + Start Date - First week-ending date of data period + End Date - Last week-ending date of data period + Group - Time-period Indicator for record: by Month, by Year, Total + State - Jurisdiction of occurrence + Condition - Condition contributing to deaths involving COVID-19 + Age Group - Age group + COVID-19 Deaths - COVID 19 Deaths

### VARES Dataset

VARES data are distributed in three data sets, VARESVAX, VARESDATA and VARESSYMPTOMS. Data sets belong to year 2020 and 2021 will be used for this project. The code book for this data set is available in the below link

Code Book

```
covid19_week <- read.csv("Provisional_COVID-19_Deaths_by_Week_Sex_and_Age.csv")
covid19_condition <- read.csv("COVID-19_Deaths_by_State_and_Age.csv")
data20 <- read.csv("2020VAERSDATA.csv")
data21 <- read.csv("2021VAERSDATA.csv")
symptoms20 <- read.csv("2020VAERSSYMPTOMS.csv")
symptoms21 <- read.csv("2021VAERSSYMPTOMS.csv")
vaccine20 <- read.csv("2020VAERSVAX.csv")
vaccine21 <- read.csv("2021VAERSVAX.csv")
```

## *Cleaning the data sets*

```
#Columns present in weekly covid-19 death data set
colnames(covid19_week)
```

```
## [1] "ï..Data.as.of"   "State"           "MMWR.Week"       "End.Week"
## [5] "Sex"             "Age.Group"       "Total.Deaths"    "COVID.19.Deaths"
```

```
print("Total number of rows and columns present in covid-19 weekly death data set")
```

```
## [1] "Total number of rows and columns present in covid-19 weekly death data set"
```

```
dim(covid19_week)
```

```
## [1] 3276    8
```

```
#Columns present in covid-19 death occurred due to underlying condition
colnames(covid19_condition)
```

```
##  [1] "ï..Data.As.Of"      "Start.Date"         "End.Date"
##  [4] "Group"              "Year"               "Month"
##  [7] "State"              "Condition.Group"    "Condition"
## [10] "ICD10_codes"        "Age.Group"          "COVID.19.Deaths"
## [13] "Number.of.Mentions" "Flag"
```

```
print("Total number of rows and columns present in covid-19 death data set with underlying conditon")
```

```
## [1] "Total number of rows and columns present in covid-19 death data set with underlying conditon"
```

```
dim(covid19_condition)
```

```
## [1] 310500     14
```

```
#Applying filter to the data sets as defined above
covid19_week_filter <- covid19_week %>% filter(State == "United States" & Sex == "All Sex" & Age.Group
dim(covid19_week_filter)
```

```
## [1] 1001    8
```

```
covid19_cond_filter <- covid19_condition %>% filter(State == "United States" & Group == "By Month" & Ag
dim(covid19_cond_filter)
```

```
## [1] 4048   14
```

```
#Removing unwanted columns that are not required for the analysis
covid19_week_cols <- c(2,4,5,6,8)
covid19_week_final <- covid19_week_filter[,covid19_week_cols]
colnames(covid19_week_final)
```

```
## [1] "State"           "End.Week"        "Sex"             "Age.Group"
## [5] "COVID.19.Deaths"
```

```
covid19_condition_cols <- c(2,3,5,6,7,8,11,12)
covid19_condition_final <- covid19_cond_filter[,covid19_condition_cols]
colnames(covid19_condition_final)
```

```
## [1] "Start.Date"      "End.Date"        "Year"            "Month"
## [5] "State"           "Condition.Group" "Age.Group"       "COVID.19.Deaths"
```

```r
#Merge data sets by year for VARES
merged_vaccine_20 <- merge(data20, symptoms20)
merged_vaccine_20 <- merge(merged_vaccine_20, vaccine20)
dim(merged_vaccine_20)
```

```
## [1] 74253    52
```

```r
colnames(merged_vaccine_20)
```

```
##  [1] "VAERS_ID"        "RECVDATE"        "STATE"           "AGE_YRS"
##  [5] "CAGE_YR"         "CAGE_MO"         "SEX"             "RPT_DATE"
##  [9] "SYMPTOM_TEXT"    "DIED"            "DATEDIED"        "L_THREAT"
## [13] "ER_VISIT"        "HOSPITAL"        "HOSPDAYS"        "X_STAY"
## [17] "DISABLE"         "RECOVD"          "VAX_DATE"        "ONSET_DATE"
## [21] "NUMDAYS"         "LAB_DATA"        "V_ADMINBY"       "V_FUNDBY"
## [25] "OTHER_MEDS"      "CUR_ILL"         "HISTORY"         "PRIOR_VAX"
## [29] "SPLTTYPE"        "FORM_VERS"       "TODAYS_DATE"     "BIRTH_DEFECT"
## [33] "OFC_VISIT"       "ER_ED_VISIT"     "ALLERGIES"       "SYMPTOM1"
## [37] "SYMPTOMVERSION1" "SYMPTOM2"        "SYMPTOMVERSION2" "SYMPTOM3"
## [41] "SYMPTOMVERSION3" "SYMPTOM4"        "SYMPTOMVERSION4" "SYMPTOM5"
## [45] "SYMPTOMVERSION5" "VAX_TYPE"        "VAX_MANU"        "VAX_LOT"
## [49] "VAX_DOSE_SERIES" "VAX_ROUTE"       "VAX_SITE"        "VAX_NAME"
```

```r
merged_vaccine_21 <- merge(data21, symptoms21)
merged_vaccine_21 <- merge(merged_vaccine_21, vaccine21)
dim(merged_vaccine_21)
```

```
## [1] 881205     52
```

```r
colnames(merged_vaccine_21)
```

```
##  [1] "VAERS_ID"        "RECVDATE"        "STATE"           "AGE_YRS"
##  [5] "CAGE_YR"         "CAGE_MO"         "SEX"             "RPT_DATE"
##  [9] "SYMPTOM_TEXT"    "DIED"            "DATEDIED"        "L_THREAT"
## [13] "ER_VISIT"        "HOSPITAL"        "HOSPDAYS"        "X_STAY"
## [17] "DISABLE"         "RECOVD"          "VAX_DATE"        "ONSET_DATE"
## [21] "NUMDAYS"         "LAB_DATA"        "V_ADMINBY"       "V_FUNDBY"
## [25] "OTHER_MEDS"      "CUR_ILL"         "HISTORY"         "PRIOR_VAX"
## [29] "SPLTTYPE"        "FORM_VERS"       "TODAYS_DATE"     "BIRTH_DEFECT"
## [33] "OFC_VISIT"       "ER_ED_VISIT"     "ALLERGIES"       "SYMPTOM1"
## [37] "SYMPTOMVERSION1" "SYMPTOM2"        "SYMPTOMVERSION2" "SYMPTOM3"
## [41] "SYMPTOMVERSION3" "SYMPTOM4"        "SYMPTOMVERSION4" "SYMPTOM5"
## [45] "SYMPTOMVERSION5" "VAX_TYPE"        "VAX_MANU"        "VAX_LOT"
## [49] "VAX_DOSE_SERIES" "VAX_ROUTE"       "VAX_SITE"        "VAX_NAME"
```

```r
#Cleaning VARES data set. From the entire data set, We have to choose vaccines given for COVID-19 only.

filter_vaccine_20 <- filter(merged_vaccine_20, grepl("COVID19", merged_vaccine_20$VAX_TYPE))
filter_vaccine_21 <- filter(merged_vaccine_21, grepl("COVID19", merged_vaccine_21$VAX_TYPE))
```

```
#Removing unwanted columns from the data set
vaccine_cols <- c(1,3,4,7,9,10,12,21,23,28,35,36,38,40,42,44,46,47,48,49,52)

vaccine_20_final <- filter_vaccine_20[,vaccine_cols]
vaccine_21_final <- filter_vaccine_21[,vaccine_cols]

#The columns present in vaccine 2020 data set
colnames(vaccine_20_final)
```

```
##  [1] "VAERS_ID"        "STATE"           "AGE_YRS"         "SEX"
##  [5] "SYMPTOM_TEXT"    "DIED"            "L_THREAT"        "NUMDAYS"
##  [9] "V_ADMINBY"       "PRIOR_VAX"       "ALLERGIES"       "SYMPTOM1"
## [13] "SYMPTOM2"        "SYMPTOM3"        "SYMPTOM4"        "SYMPTOM5"
## [17] "VAX_TYPE"        "VAX_MANU"        "VAX_LOT"         "VAX_DOSE_SERIES"
## [21] "VAX_NAME"
```

```
#Total number of rows and columns present in the data set
dim(vaccine_20_final)
```

```
## [1] 14116    21
```

```
#The columns present in vaccine 2021 data set
colnames(vaccine_21_final)
```

```
##  [1] "VAERS_ID"        "STATE"           "AGE_YRS"         "SEX"
##  [5] "SYMPTOM_TEXT"    "DIED"            "L_THREAT"        "NUMDAYS"
##  [9] "V_ADMINBY"       "PRIOR_VAX"       "ALLERGIES"       "SYMPTOM1"
## [13] "SYMPTOM2"        "SYMPTOM3"        "SYMPTOM4"        "SYMPTOM5"
## [17] "VAX_TYPE"        "VAX_MANU"        "VAX_LOT"         "VAX_DOSE_SERIES"
## [21] "VAX_NAME"
```

```
#Total number of rows and columns present in the data set
dim(vaccine_21_final)
```

```
## [1] 843061     21
```

### *Final data sets*

```
#The final data sets after cleaning and before slicing and dicing
#covid-19 weekly death count by Age
print(str(covid19_week_final))
```

```
## 'data.frame':    1001 obs. of  5 variables:
##  $ State          : chr  "United States" "United States" "United States" "United States" ...
##  $ End.Week       : chr  "01/04/2020" "01/04/2020" "01/04/2020" "01/04/2020" ...
##  $ Sex            : chr  "All Sex" "All Sex" "All Sex" "All Sex" ...
##  $ Age.Group      : chr  "Under 1 year" "1-4 Years" "5-14 Years" "15-24 Years" ...
##  $ COVID.19.Deaths: chr  "0" "0" "0" "0" ...
## NULL
```

```r
#Covid-19 monthly deaths by age with underlying condition
print(str(covid19_condition_final))
```

```
## 'data.frame':    4048 obs. of  8 variables:
##  $ Start.Date     : chr  "01/01/2020" "02/01/2020" "03/01/2020" "04/01/2020" ...
##  $ End.Date       : chr  "01/31/2020" "02/29/2020" "03/31/2020" "04/30/2020" ...
##  $ Year           : chr  "2,020" "2,020" "2,020" "2,020" ...
##  $ Month          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ State          : chr  "United States" "United States" "United States" "United States" ...
##  $ Condition.Group: chr  "Respiratory diseases" "Respiratory diseases" "Respiratory diseases" "Respi
##  $ Age.Group      : chr  "0-24" "0-24" "0-24" "0-24" ...
##  $ COVID.19.Deaths: chr  "0" "0" "9" "27" ...
## NULL
```

```r
#Covid-19 Vaccine data for 2020 and 2021
print(str(vaccine_20_final))
```

```
## 'data.frame':    14116 obs. of  21 variables:
##  $ VAERS_ID       : int  902418 902440 902446 902464 902465 902465 902468 902468 902479 902490 ...
##  $ STATE          : chr  "NJ" "AZ" "WV" "LA" ...
##  $ AGE_YRS        : num  56 35 55 42 60 60 59 59 46 37 ...
##  $ SEX            : chr  "F" "F" "F" "M" ...
##  $ SYMPTOM_TEXT   : chr  "Patient experienced mild numbness traveling from injection site up and dow
##  $ DIED           : chr  "" "" "" "" ...
##  $ L_THREAT       : chr  "" "" "" "" ...
##  $ NUMDAYS        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ V_ADMINBY      : chr  "PVT" "PVT" "OTH" "PVT" ...
##  $ PRIOR_VAX      : chr  "" "" "" "" ...
##  $ ALLERGIES      : chr  "none" "" "Contrast Dye IV contrast, shellfish, strawberry" "none" ...
##  $ SYMPTOM1       : chr  "Hypoaesthesia" "Headache" "Erythema" "Dizziness" ...
##  $ SYMPTOM2       : chr  "Injection site hypoaesthesia" "" "Feeling hot" "Electrocardiogram normal"
##  $ SYMPTOM3       : chr  "" "" "Flushing" "Hyperhidrosis" ...
##  $ SYMPTOM4       : chr  "" "" "" "Laboratory test normal" ...
##  $ SYMPTOM5       : chr  "" "" "" "Presyncope" ...
##  $ VAX_TYPE       : chr  "COVID19" "COVID19" "COVID19" "COVID19" ...
##  $ VAX_MANU       : chr  "PFIZER\\BIONTECH" "PFIZER\\BIONTECH" "PFIZER\\BIONTECH" "PFIZER\\BIONTECH"
##  $ VAX_LOT        : chr  "EH9899" "EH 9899" "EH9899" "EH9899" ...
##  $ VAX_DOSE_SERIES: chr  "1" "1" "1" "UNK" ...
##  $ VAX_NAME       : chr  "COVID19 (COVID19 (PFIZER-BIONTECH))" "COVID19 (COVID19 (PFIZER-BIONTECH))"
## NULL
```

```r
print(str(vaccine_21_final))
```

```
## 'data.frame':    843061 obs. of  21 variables:
##  $ VAERS_ID       : int  916600 916601 916602 916603 916604 916606 916607 916608 916609 916610 ...
##  $ STATE          : chr  "TX" "CA" "WA" "WA" ...
##  $ AGE_YRS        : num  33 73 23 58 47 44 50 33 71 18 ...
##  $ SEX            : chr  "F" "F" "F" "F" ...
##  $ SYMPTOM_TEXT   : chr  "Right side of epiglottis swelled up and hinder swallowing pictures taken Be
##  $ DIED           : chr  "" "" "" "" ...
##  $ L_THREAT       : chr  "" "" "" "" ...
##  $ NUMDAYS        : int  2 0 0 0 7 0 1 2 8 1 ...
```

```
##  $ V_ADMINBY      : chr  "PVT" "SEN" "SEN" "WRK" ...
##  $ PRIOR_VAX      : chr  "" "" "" "got measles from measel shot, mums from mumps shot, headaches and
##  $ ALLERGIES      : chr  "Pcn and bee venom" "\"Dairy\"" "Shellfish" "Diclofenac, novacaine, lidocai
##  $ SYMPTOM1       : chr  "Dysphagia" "Anxiety" "Chest discomfort" "Dizziness" ...
##  $ SYMPTOM2       : chr  "Epiglottitis" "Dyspnoea" "Dysphagia" "Fatigue" ...
##  $ SYMPTOM3       : chr  "" "" "Pain in extremity" "Mobility decreased" ...
##  $ SYMPTOM4       : chr  "" "" "Visual impairment" "" ...
##  $ SYMPTOM5       : chr  "" "" "" "" ...
##  $ VAX_TYPE       : chr  "COVID19" "COVID19" "COVID19" "COVID19" ...
##  $ VAX_MANU       : chr  "MODERNA" "MODERNA" "PFIZER\\BIONTECH" "MODERNA" ...
##  $ VAX_LOT        : chr  "037K20A" "025L20A" "EL1284" "unknown" ...
##  $ VAX_DOSE_SERIES: chr  "1" "1" "1" "UNK" ...
##  $ VAX_NAME       : chr  "COVID19 (COVID19 (MODERNA))" "COVID19 (COVID19 (MODERNA))" "COVID19 (COVID
## NULL
```

### *Adding additional variable to final data sets*

I will be adding a variable called people to covid19_weekly data set which tells if the people is young or old based on the age. In addition, I will be adding a variable called "condition_flag" to covid19_condition data set which tells if the people had underlying conditions.

```
old <- c("55-64 Years","65-74 Years","75-84 Years","85 Years and Over")
covid19_week_final$people <- ifelse(covid19_week_final$Age.Group %in% old, "Old", "Young")
colnames(covid19_week_final)
```

```
## [1] "State"           "End.Week"        "Sex"             "Age.Group"
## [5] "COVID.19.Deaths" "people"
```

```
print(str(covid19_week_final))
```

```
## 'data.frame':    1001 obs. of  6 variables:
##  $ State          : chr  "United States" "United States" "United States" "United States" ...
##  $ End.Week       : chr  "01/04/2020" "01/04/2020" "01/04/2020" "01/04/2020" ...
##  $ Sex            : chr  "All Sex" "All Sex" "All Sex" "All Sex" ...
##  $ Age.Group      : chr  "Under 1 year" "1-4 Years" "5-14 Years" "15-24 Years" ...
##  $ COVID.19.Deaths: chr  "0" "0" "0" "0" ...
##  $ people         : chr  "Young" "Young" "Young" "Young" ...
## NULL
```

```
covid19_condition_final$condition_flag <- ifelse(covid19_condition_final$Condition.Group == "COVID-19",
print(str(covid19_condition_final))
```

```
## 'data.frame':    4048 obs. of  9 variables:
##  $ Start.Date     : chr  "01/01/2020" "02/01/2020" "03/01/2020" "04/01/2020" ...
##  $ End.Date       : chr  "01/31/2020" "02/29/2020" "03/31/2020" "04/30/2020" ...
##  $ Year           : chr  "2,020" "2,020" "2,020" "2,020" ...
##  $ Month          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ State          : chr  "United States" "United States" "United States" "United States" ...
##  $ Condition.Group: chr  "Respiratory diseases" "Respiratory diseases" "Respiratory diseases" "Respi
##  $ Age.Group      : chr  "0-24" "0-24" "0-24" "0-24" ...
##  $ COVID.19.Deaths: chr  "0" "0" "9" "27" ...
##  $ condition_flag : chr  "Yes" "Yes" "Yes" "Yes" ...
## NULL
```

### *Different ways to view the data*

Some of the different ways to look at the data set

- Provisional COVID-19 deaths by week, sex and age
  - Age
  - Week
  - Covid-19 Deaths
  - People (derived variable based on age of the people)
- Conditions contributing to COVID-19 deaths, by state and age, provisional 2020-21
  - Age
  - Condition.Group
  - Covid-19 Deaths
  - condition_flag (derived variable based on condition.group)
- Vaccine data sets
  - Age
  - Died
  - VAERS_ID

### *Slicing and Dicing the data sets*

#### Covid-19 Weekly death data set

```
covid19_week_final$COVID.19.Deaths <- as.numeric(covid19_week_final$COVID.19.Deaths)
```

```
## Warning: NAs introduced by coercion
```

```
print(str(covid19_week_final,10))
```

```
## 'data.frame':    1001 obs. of  6 variables:
##  $ State          : chr  "United States" "United States" "United States" "United States" ...
##  $ End.Week       : chr  "01/04/2020" "01/04/2020" "01/04/2020" "01/04/2020" ...
##  $ Sex            : chr  "All Sex" "All Sex" "All Sex" "All Sex" ...
##  $ Age.Group      : chr  "Under 1 year" "1-4 Years" "5-14 Years" "15-24 Years" ...
##  $ COVID.19.Deaths: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ people         : chr  "Young" "Young" "Young" "Young" ...
## NULL
```

```
#Total deaths by Covid-19 for Young and old People
covid19_week_final %>% group_by(people) %>% summarise(COVID19_Deaths=sum(COVID.19.Deaths, na.rm = TRUE))
```

```
## # A tibble: 2 x 2
##   people COVID19_Deaths
##   <chr>           <dbl>
## 1 Old             67898
## 2 Young           52829
```

```r
#Slicing the data set based on People (Young and Old)
covid19_week_young <- filter(covid19_week_final, people=="Young" & COVID.19.Deaths>0)
dim(covid19_week_young)
```

```
## [1] 482   6
```

```r
covid19_week_old <- filter(covid19_week_final, people=="Old" & COVID.19.Deaths>0)
dim(covid19_week_old)
```

```
## [1] 132   6
```

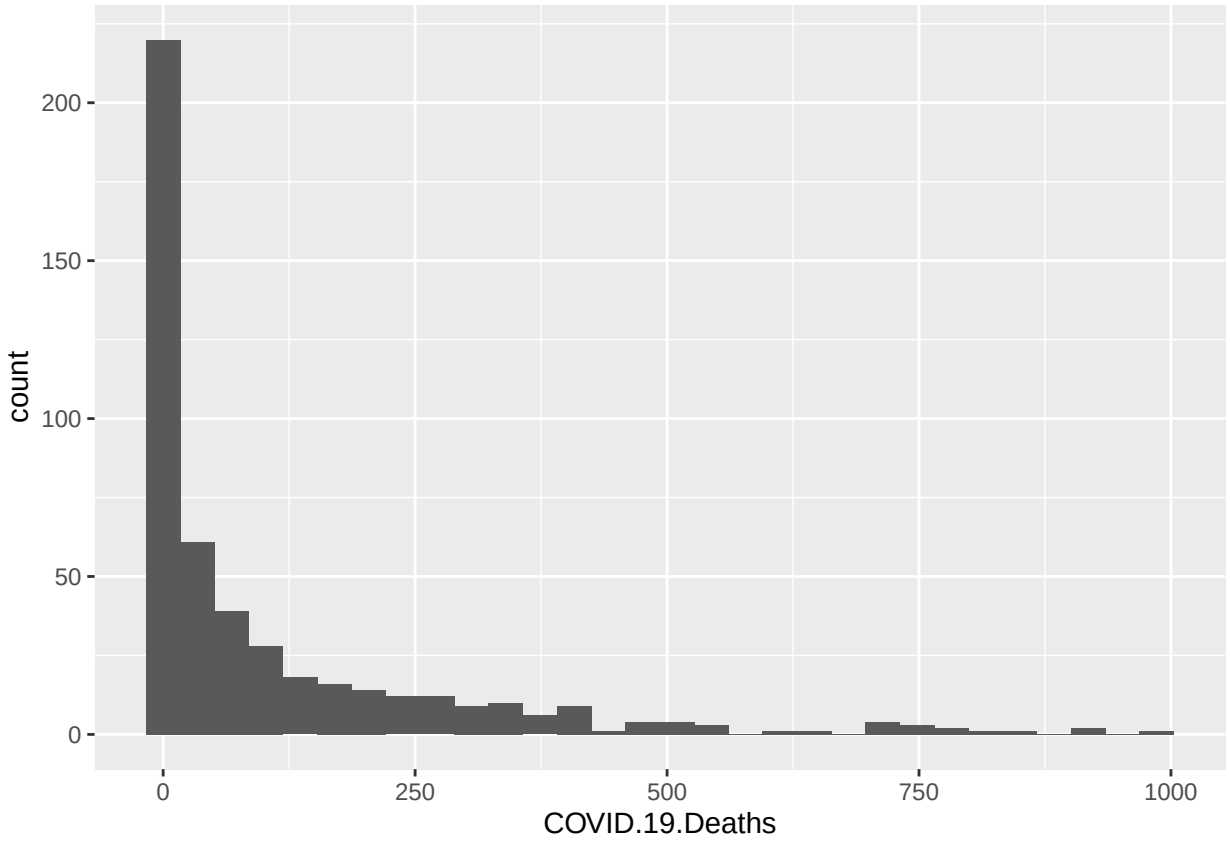
```r
#Printing the total number of deaths for young and old people
cat("Total number of covid-19 deaths for young people: ",sum(covid19_week_young$COVID.19.Deaths))
```

```
## Total number of covid-19 deaths for young people:  52829
```

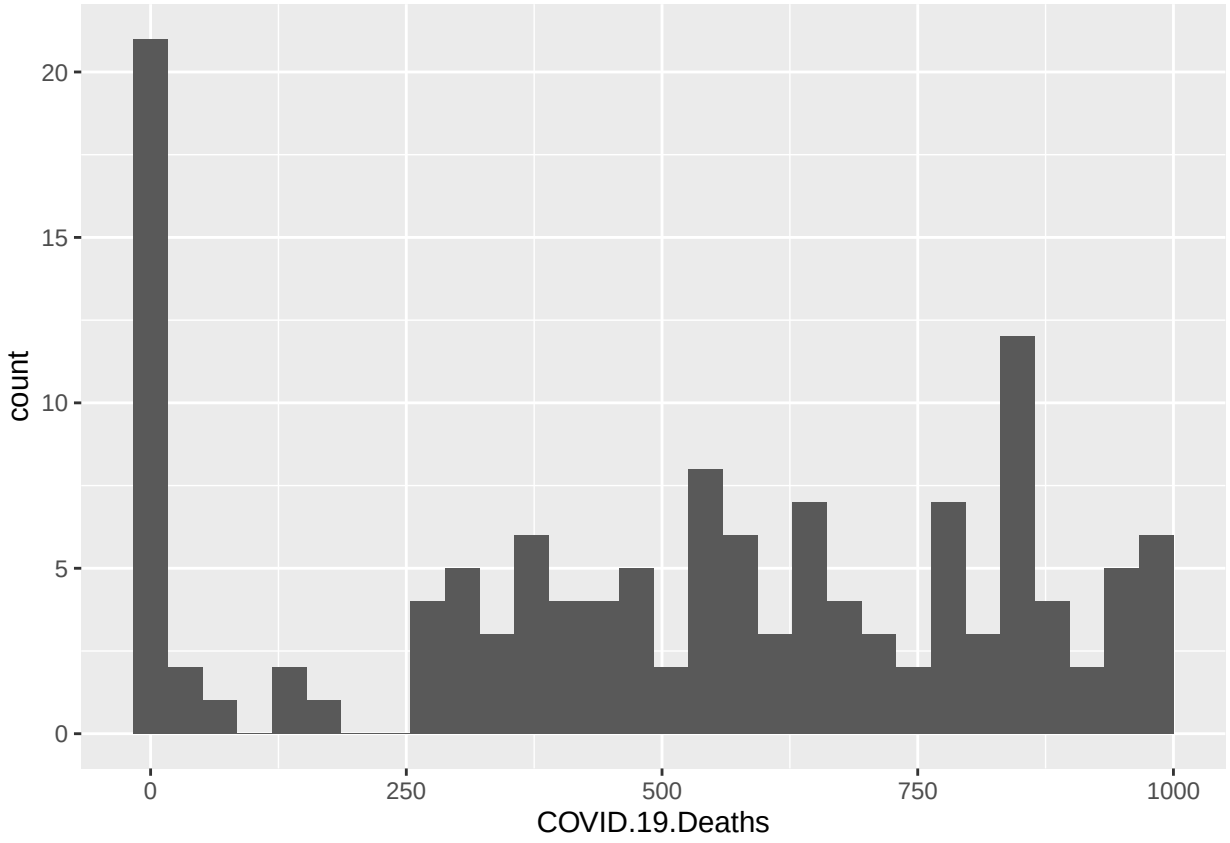
```r
cat("Total number of covid-19 deaths for old people: ",sum(covid19_week_old$COVID.19.Deaths))
```

```
## Total number of covid-19 deaths for old people:  67898
```

```r
#Histograms on covid-19 death for young and old people
ggplot(covid19_week_young, aes(COVID.19.Deaths)) + geom_histogram(bins=30)
```



```
ggplot(covid19_week_old, aes(COVID.19.Deaths)) + geom_histogram(bins=30)
```




```
#Summary of weekly covid-19 deaths data set
summary(covid19_week_young)
```

```
##     State              End.Week            Sex              Age.Group
##  Length:482         Length:482         Length:482         Length:482
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  COVID.19.Deaths    people
##  Min.   :  1.0   Length:482
##  1st Qu.:  2.0   Class :character
##  Median : 25.5   Mode  :character
##  Mean   :109.6
##  3rd Qu.:141.5
##  Max.   :987.0
```

```
summary(covid19_week_old)
```

```
##     State              End.Week            Sex              Age.Group
##  Length:132         Length:132         Length:132         Length:132
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  COVID.19.Deaths    people
##  Min.   :  1.0   Length:132
##  1st Qu.:310.8   Class :character
##  Median :552.5   Mode  :character
##  Mean   :514.4
##  3rd Qu.:795.2
##  Max.   :984.0
```

```
cat("The variance of death count for young people: ", var(covid19_week_young$COVID.19.Deaths))
```

```
## The variance of death count for young people:  30973.93
```

```
cat("The standard deviation of death count for young people: ", sd(covid19_week_young$COVID.19.Deaths))
```

```
## The standard deviation of death count for young people:  175.9941
```

```
cat("The variance of death count for old people: ", var(covid19_week_old$COVID.19.Deaths))
```

```
## The variance of death count for old people:  99185.96
```

```
cat("The standard deviation of death count for old people: ", sd(covid19_week_old$COVID.19.Deaths))
```

```
## The standard deviation of death count for old people:  314.938
```

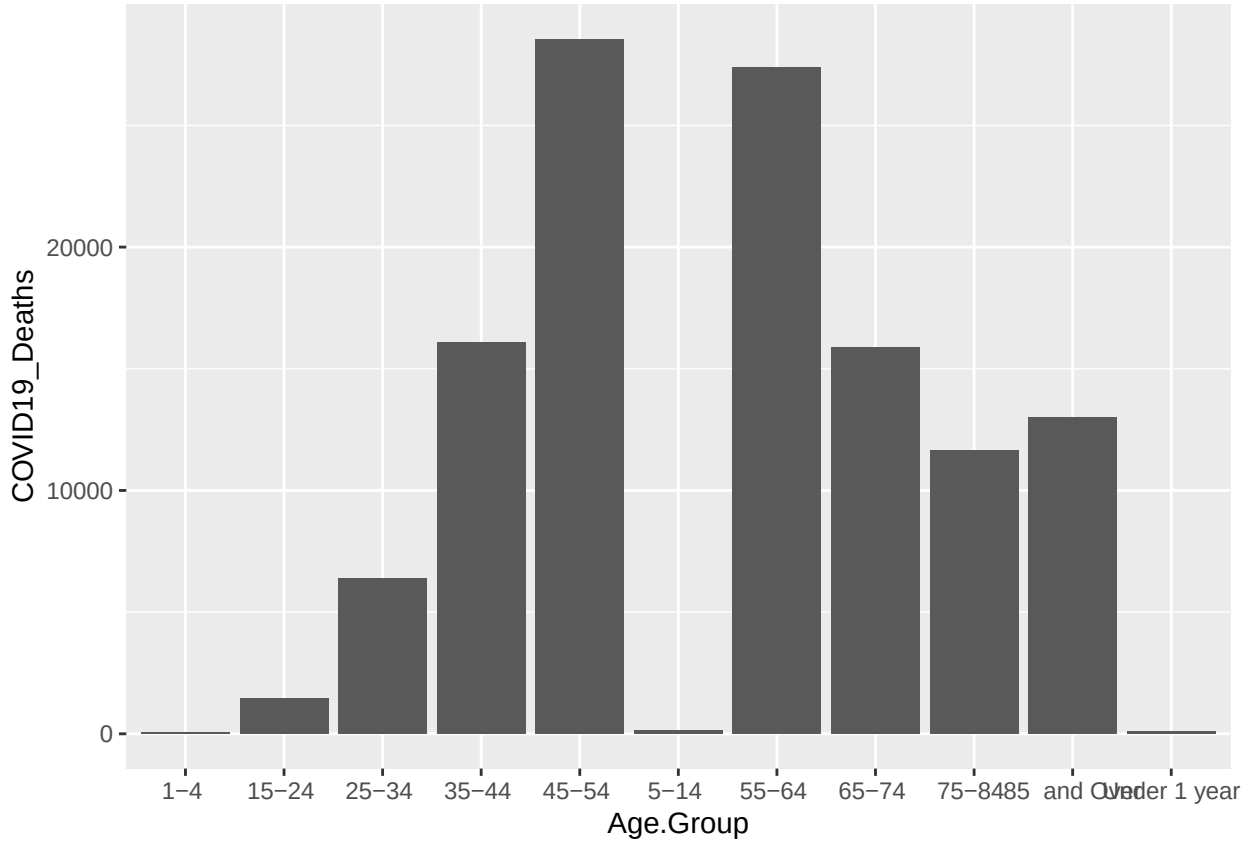
```
covid19_deaths_by_people <- covid19_week_final %>% group_by(Age.Group) %>% summarise(COVID19_Deaths=sum

#Remove Years from the Age.Group
covid19_deaths_by_people$Age.Group <- gsub("Years","", covid19_deaths_by_people$Age.Group)

covid19_deaths_by_people
```

```
## # A tibble: 11 x 2
##    Age.Group      COVID19_Deaths
##    <chr>                   <dbl>
##  1 "1-4 "                     59
##  2 "15-24 "                 1463
##  3 "25-34 "                 6394
##  4 "35-44 "                16094
##  5 "45-54 "                28550
##  6 "5-14 "                   154
##  7 "55-64 "                27367
##  8 "65-74 "                15873
##  9 "75-84 "                11661
## 10 "85  and Over"          12997
## 11 "Under 1 year"            115
```

```
ggplot(covid19_deaths_by_people, aes(x=Age.Group, y=COVID19_Deaths)) + geom_bar(stat = "identity")
```




**Covid-19 death underlying condition**

```
#Converting datatype to numeric
covid19_condition_final$COVID.19.Deaths <- as.numeric(covid19_condition_final$COVID.19.Deaths)
```

```
## Warning: NAs introduced by coercion
```

```
print(str(covid19_condition_final))
```

```
## 'data.frame':    4048 obs. of  9 variables:
##  $ Start.Date     : chr  "01/01/2020" "02/01/2020" "03/01/2020" "04/01/2020" ...
##  $ End.Date       : chr  "01/31/2020" "02/29/2020" "03/31/2020" "04/30/2020" ...
##  $ Year           : chr  "2,020" "2,020" "2,020" "2,020" ...
##  $ Month          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ State          : chr  "United States" "United States" "United States" "United States" ...
##  $ Condition.Group: chr  "Respiratory diseases" "Respiratory diseases" "Respiratory diseases" "Respi
##  $ Age.Group      : chr  "0-24" "0-24" "0-24" "0-24" ...
##  $ COVID.19.Deaths: num  0 0 9 27 19 17 38 32 13 9 ...
##  $ condition_flag : chr  "Yes" "Yes" "Yes" "Yes" ...
## NULL
```

```r
#Slicing the data set based on People with and without condition
covid19_condition_no <- filter(covid19_condition_final, condition_flag=="No" & COVID.19.Deaths>0)
dim(covid19_condition_no)
```

```
## [1] 67  9
```

```r
covid19_condition_yes <- filter(covid19_condition_final, condition_flag=="Yes" & COVID.19.Deaths>0)
dim(covid19_condition_yes)
```

```
## [1] 2755    9
```

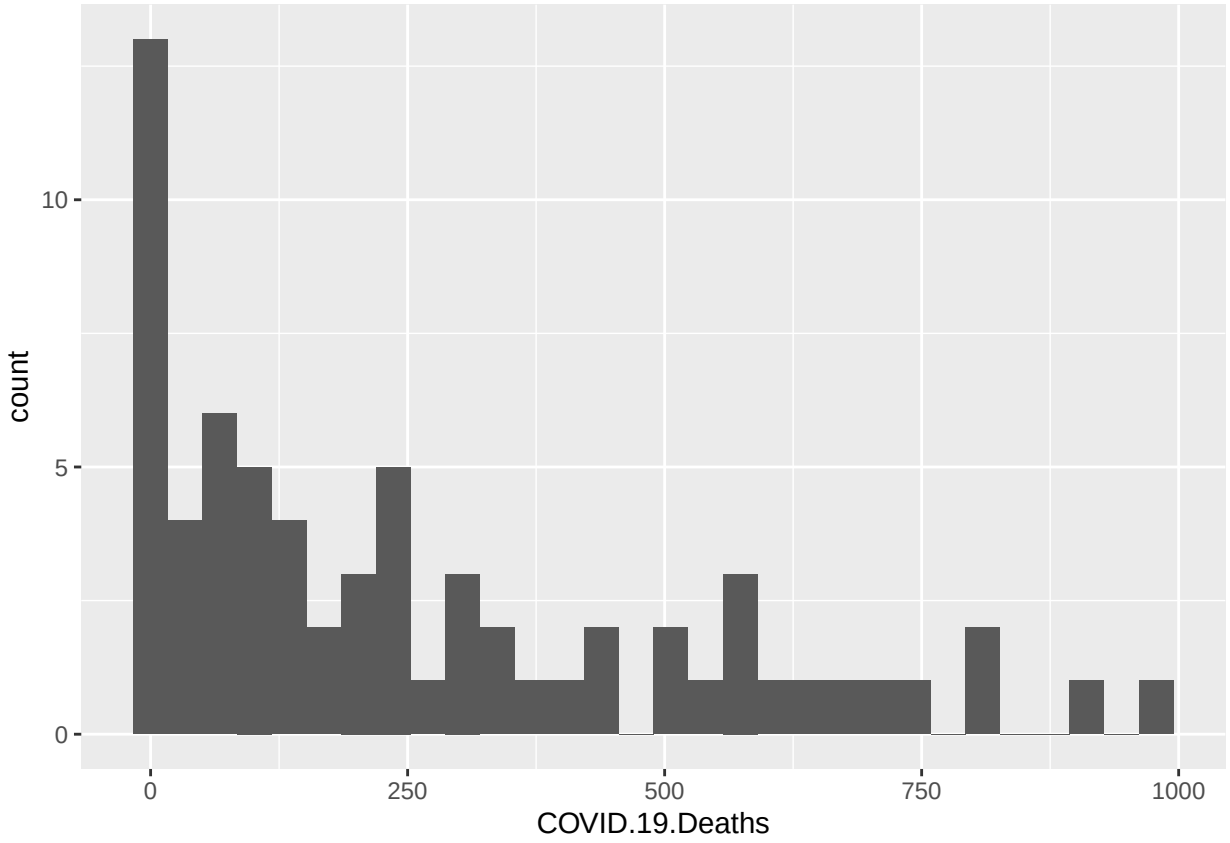
```r
#Printing the total number of deaths for young and old people
cat("Total number of covid-19 deaths for the people without underlying condition: ",sum(covid19_conditi
```

```
## Total number of covid-19 deaths for the people without underlying condition:  17250
```

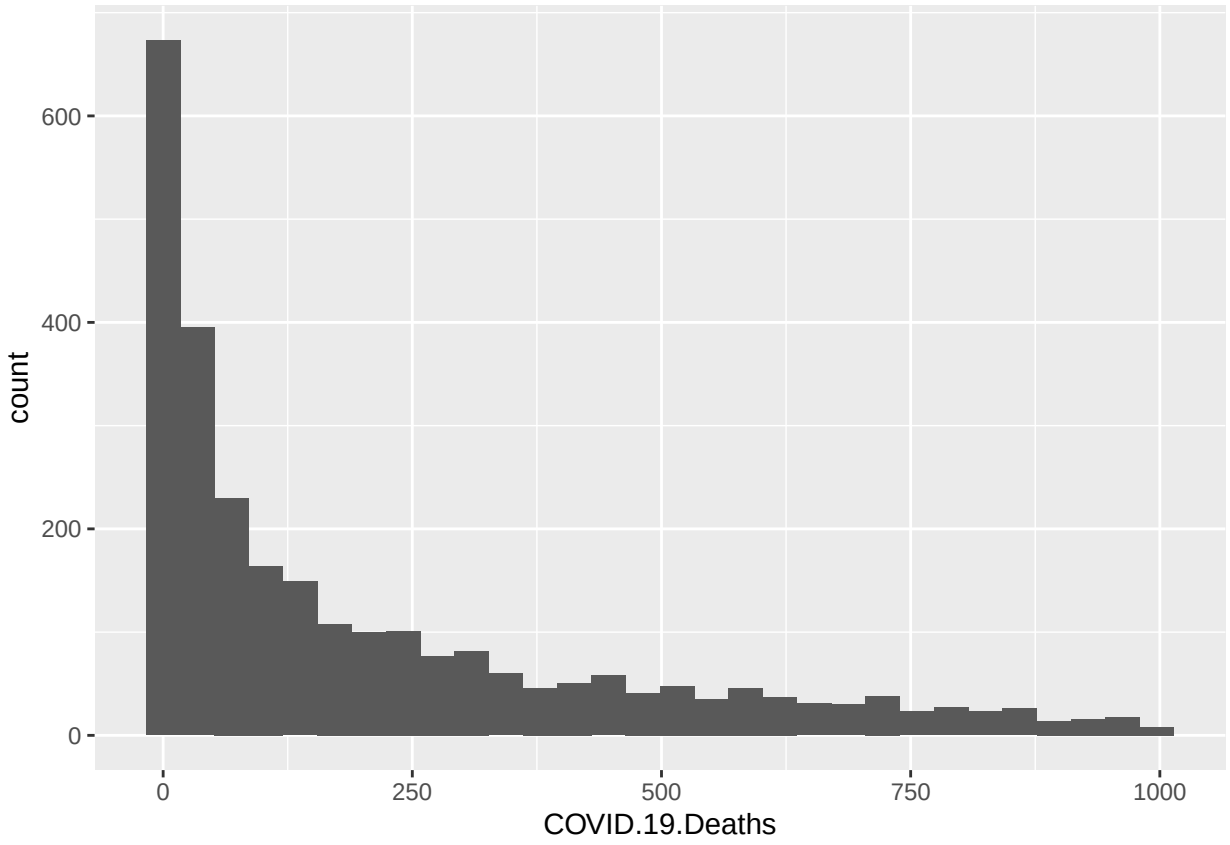
```r
cat("Total number of covid-19 deaths for the people with underlying condition: ",sum(covid19_condition_
```

```
## Total number of covid-19 deaths for the people with underlying condition:  579299
```

```r
#Histograms on covid-19 death for young and old people
ggplot(covid19_condition_no, aes(COVID.19.Deaths)) + geom_histogram(bins=30)
```



```
ggplot(covid19_condition_yes, aes(COVID.19.Deaths)) + geom_histogram(bins=30)
```




```
#Summary of covid-19 deaths
summary(covid19_condition_no)
```

```
##   Start.Date          End.Date              Year              Month
##  Length:67          Length:67          Length:67          Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 3.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.000
##                                                           Mean   : 5.701
##                                                           3rd Qu.: 8.500
##                                                           Max.   :12.000
##     State           Condition.Group     Age.Group          COVID.19.Deaths
##  Length:67          Length:67          Length:67          Min.   :  1.0
##  Class :character   Class :character   Class :character   1st Qu.: 53.0
##  Mode  :character   Mode  :character   Mode  :character   Median :164.0
##                                                           Mean   :257.5
##                                                           3rd Qu.:413.0
##                                                           Max.   :979.0
##  condition_flag
##  Length:67
##  Class :character
##  Mode  :character
##
##
```

```
##
```

```
summary(covid19_condition_yes)
```

```
##   Start.Date          End.Date             Year               Month
##  Length:2755        Length:2755        Length:2755        Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 4.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 6.000
##                                                           Mean   : 6.216
##                                                           3rd Qu.: 8.000
##                                                           Max.   :12.000
##     State           Condition.Group     Age.Group          COVID.19.Deaths
##  Length:2755        Length:2755        Length:2755        Min.   :  1.0
##  Class :character   Class :character   Class :character   1st Qu.: 19.0
##  Mode  :character   Mode  :character   Mode  :character   Median : 99.0
##                                                           Mean   :210.3
##                                                           3rd Qu.:322.0
##                                                           Max.   :998.0
##  condition_flag
##  Length:2755
##  Class :character
##  Mode  :character
##
##
##
```

```
cat("The variance of death count for the people without underlying condition: ", var(covid19_condition_
```

```
## The variance of death count for the people without underlying condition:  68582.8
```

```
cat("The standard deviation of death count for the people without underlying condition: ", sd(covid19_c
```

```
## The standard deviation of death count for the people without underlying condition:  261.8832
```

```
cat("The variance of death count for the people with underlying condition: ", var(covid19_condition_yes
```

```
## The variance of death count for the people with underlying condition:  61526.05
```

```
cat("The standard deviation of death count for the people with underlying condition: ", sd(covid19_cond
```

```
## The standard deviation of death count for the people with underlying condition:  248.0445
```

```
head(covid19_condition_final)
```

```
##   Start.Date   End.Date  Year Month         State      Condition.Group
## 1 01/01/2020 01/31/2020 2,020     1 United States Respiratory diseases
## 2 02/01/2020 02/29/2020 2,020     2 United States Respiratory diseases
## 3 03/01/2020 03/31/2020 2,020     3 United States Respiratory diseases
## 4 04/01/2020 04/30/2020 2,020     4 United States Respiratory diseases
```

```
## 5 05/01/2020 05/31/2020 2,020     5 United States Respiratory diseases
## 6 06/01/2020 06/30/2020 2,020     6 United States Respiratory diseases
##   Age.Group COVID.19.Deaths condition_flag
## 1      0-24               0            Yes
## 2      0-24               0            Yes
## 3      0-24               9            Yes
## 4      0-24              27            Yes
## 5      0-24              19            Yes
## 6      0-24              17            Yes
```

```
covid19_condition_final$COVID.19.Deaths <- as.numeric(covid19_condition_final$COVID.19.Deaths)

covid19_deaths_by_condition <- covid19_condition_final %>% group_by(condition_flag,Age.Group) %>% summa
```

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

```
covid19_deaths_by_condition
```

```
## # A tibble: 16 x 3
## # Groups:   condition_flag [2]
##    condition_flag Age.Group COVID19_Deaths
##    <chr>          <chr>              <dbl>
##  1 No             0-24                1834
##  2 No             25-34               5397
##  3 No             35-44               7687
##  4 No             45-54               2308
##  5 No             55-64                  6
##  6 No             65-74                  6
##  7 No             75-84                  7
##  8 No             85+                    5
##  9 Yes            0-24                4408
## 10 Yes            25-34              16617
## 11 Yes            35-44              39760
## 12 Yes            45-54              84184
## 13 Yes            55-64             100762
## 14 Yes            65-74             111232
## 15 Yes            75-84             115510
## 16 Yes            85+               106826
```

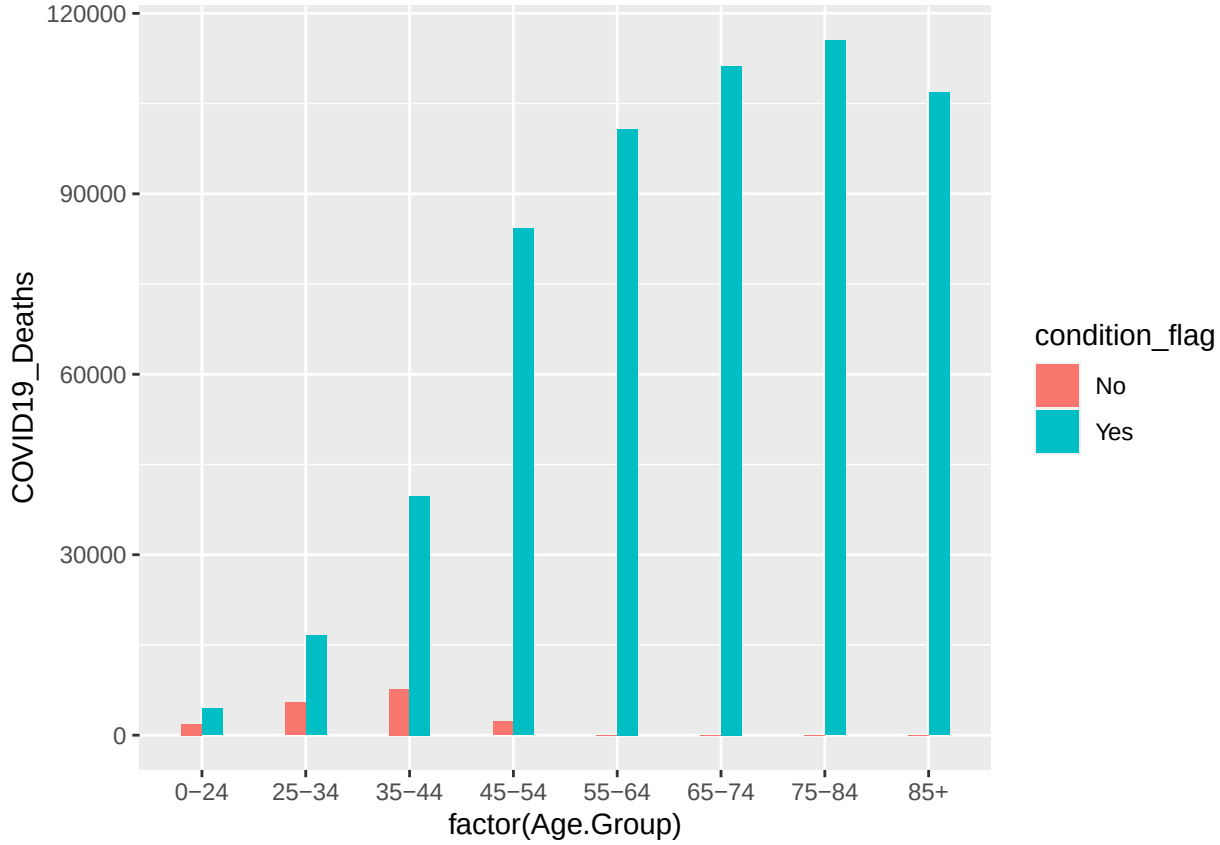
```
covid19_deaths_by_condition.long <- covid19_deaths_by_condition %>% gather("Stat","Value",-Age.Group)

covid19_deaths_by_condition
```

```
## # A tibble: 16 x 3
## # Groups:   condition_flag [2]
##    condition_flag Age.Group COVID19_Deaths
##    <chr>          <chr>              <dbl>
##  1 No             0-24                1834
##  2 No             25-34               5397
##  3 No             35-44               7687
##  4 No             45-54               2308
##  5 No             55-64                  6
```

```
##  6 No             65-74                6
##  7 No             75-84                7
##  8 No             85+                  5
##  9 Yes            0-24              4408
## 10 Yes            25-34            16617
## 11 Yes            35-44            39760
## 12 Yes            45-54            84184
## 13 Yes            55-64           100762
## 14 Yes            65-74           111232
## 15 Yes            75-84           115510
## 16 Yes            85+             106826
```

```
ggplot(covid19_deaths_by_condition, aes(x=factor(Age.Group), y=COVID19_Deaths, fill = condition_flag)) +
```

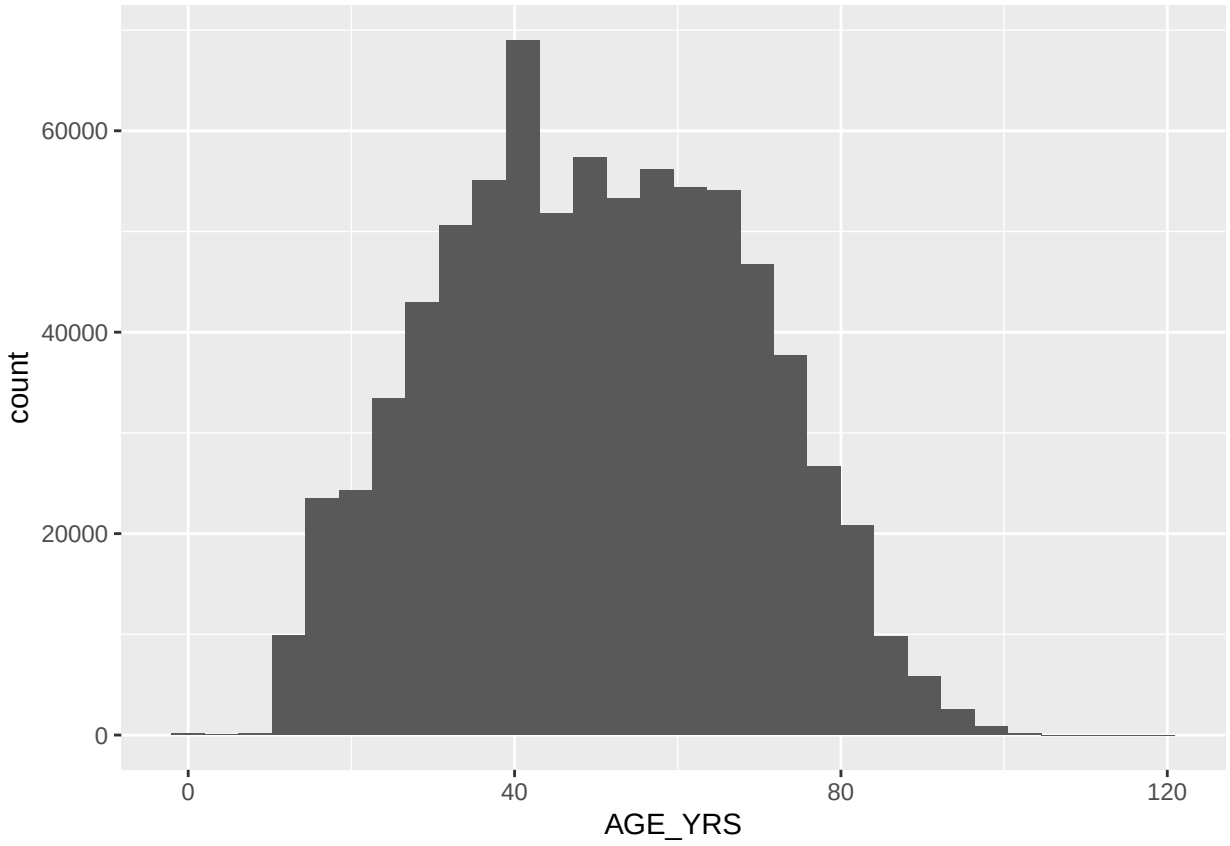



**Vaccine data sets**

```
#Combining vaccine data for both 2020 and 2021
vaccine_df <- union(vaccine_20_final, vaccine_21_final)

#Age Analysis
age_variable <- vaccine_df[-is.na(vaccine_df$AGE_YRS)]
age_hist <- ggplot(vaccine_df, aes(AGE_YRS)) + geom_histogram(bins=30)
age_hist
```

```
## Warning: Removed 68171 rows containing non-finite values (stat_bin).
```




```
#Death data analysis

vaccine_died <- dplyr::filter(vaccine_df, grepl("Y",vaccine_df$DIED))
colnames(vaccine_died)
```

```
##  [1] "VAERS_ID"        "STATE"           "AGE_YRS"         "SEX"
##  [5] "SYMPTOM_TEXT"    "DIED"            "L_THREAT"        "NUMDAYS"
##  [9] "V_ADMINBY"       "PRIOR_VAX"       "ALLERGIES"       "SYMPTOM1"
## [13] "SYMPTOM2"        "SYMPTOM3"        "SYMPTOM4"        "SYMPTOM5"
## [17] "VAX_TYPE"        "VAX_MANU"        "VAX_LOT"         "VAX_DOSE_SERIES"
## [21] "VAX_NAME"
```

```
vaccine_died_nodup <- vaccine_died |> dplyr::distinct(VAERS_ID, .keep_all = TRUE)
dim(vaccine_died_nodup)
```

```
## [1] 7848   21
```

```
cat("Total number of people died after taking vaccine: ", length(unique(vaccine_died$VAERS_ID)))
```

```
## Total number of people died after taking vaccine:  7848
```

```r
#Splitting the data set into young and old based on age.
vaccine_died_young <- filter(vaccine_died, AGE_YRS<55)
dim(vaccine_died_young)
```

```
## [1] 1617   21
```

```r
cat("Total number of young people died after taking vaccine: ", length(unique(vaccine_died_young$VAERS_
```

```
## Total number of young people died after taking vaccine:  856
```

```r
vaccine_died_old <- filter(vaccine_died, AGE_YRS>=55)
dim(vaccine_died_old)
```

```
## [1] 12049    21
```

```r
cat("Total number of old people died after taking vaccine: ", length(unique(vaccine_died_old$VAERS_ID))
```

```
## Total number of old people died after taking vaccine:  6281
```

# *Implications*

## CDC weekly death count data set

From this data set, I could see the death count of young people (age $< 55$) is less compared to old people (age $>= 55$). The death count during the initial months were less as Covid-19 infection started spreading and peaked in the later months on 2020 and initial months of 2021, and again started going down from middle of 2021 due to vaccinations.

The histogram for the Covid-19 deaths for young people is positively skewed distribution whereas the histogram for the Covid-19 deaths for old people is also positively skewed distribution but shows some pattern for multiple distribution as well.

Bar chart also depicts the same where the count of covid-19 deaths is higher for old people compared to young people.

## CDC's Covid-19 death underlying condition data set

This data set tells that the death count of the people without any underlying condition is less compared to those people with underlying condition.

The histograms for the Covid-19 deaths for the people with and without underlying conditions are positively skewed distribution. This is because the covid-19 death count is high during 2020 and 1st quarter of 2021. From 2nd quarter of 2021, the count started decreasing.

## VAERS Vaccine data set

The vaccine data set also shows the death count of the people having young age (less than 54) is less compared to the old people having age greater than 55.

### Average death due to covid-19 for young and old people

- Average covid-19 death count for young people: 109.6
- Average covid-19 death count for old people: 514.4

### Variance and standard deviation of covid-19 death for young and old people

- The variance and standard deviation of covid-19 death count for young people: 30973.93 and 175.9941
- The variance and standard deviation of covid-19 death count for old people: 99185.96 and 314.938

### Average death due to covid-19 for the people with and without underlying condition

- Average covid-19 death count for the people without underlying condition: 164
- Average covid-19 death count for the people with underlying condition: 210

### Variance and standard deviation of covid-19 death for the people with and without underlying condition?

- The variance and standard deviation of covid-19 death for the people without underlying condition: 68582.8 and 261.8832
- The variance and standard deviation of covid-19 death for the people with underlying condition: 61526.05 and 248.0445

### Role age played in covid-19 deaths

```
#Calculate total deaths by age
cat("Number of covid-19 deaths by age group:\n")
```

```
## Number of covid-19 deaths by age group:
```

```
covid19_week_final %>% group_by(Age.Group) %>% summarise(COVID19_Deaths=sum(COVID.19.Deaths, na.rm = TRU
```

```
## # A tibble: 11 x 2
##    Age.Group         COVID19_Deaths
##    <chr>                      <dbl>
##  1 1-4 Years                     59
##  2 15-24 Years                 1463
##  3 25-34 Years                 6394
##  4 35-44 Years                16094
##  5 45-54 Years                28550
##  6 5-14 Years                   154
##  7 55-64 Years                27367
##  8 65-74 Years                15873
##  9 75-84 Years                11661
## 10 85 Years and Over          12997
## 11 Under 1 year                 115
```

```
#Filtering the data till Aug 2021 and applying group by to calculate total deaths by end week
cat("Number of covid-19 deaths by week:\n")
```

```
## Number of covid-19 deaths by week:
```

```
covid19_week_final %>% filter(as.Date(End.Week, format= "%m/%d/%Y") < "2021-09-01") %>% group_by(as.Dat
```

```
## # A tibble: 87 x 2
##    `as.Date(End.Week, format = "%m/%d/%Y")` COVID19_Deaths
##    <date>                                          <dbl>
##  1 2020-01-04                                          0
##  2 2020-01-11                                          1
##  3 2020-01-18                                          2
##  4 2020-01-25                                          2
##  5 2020-02-01                                          0
##  6 2020-02-08                                          2
##  7 2020-02-15                                          2
##  8 2020-02-22                                          6
##  9 2020-02-29                                          9
## 10 2020-03-07                                         37
## # ... with 77 more rows
```

### Role underlying condition played in covid-19 deaths

```
#Total deaths by Covid-19 for the people with and without underlying condition
cat("Number of deaths by underlying condition: \n")
```

```
## Number of deaths by underlying condition:
```

```
covid19_condition_final %>% group_by(condition_flag) %>% summarise(COVID19_Deaths=sum(COVID.19.Deaths, r
```

```
## # A tibble: 2 x 2
##   condition_flag COVID19_Deaths
##   <chr>                   <dbl>
## 1 No                      17250
## 2 Yes                    579299
```

```
#Death count by underlying condition
cat("Number of covid-19 deaths by underlying condition")
```

```
## Number of covid-19 deaths by underlying condition
```

```
covid19_condition_yes %>% group_by(Condition.Group) %>% summarise(COVID19_Deaths=sum(COVID.19.Deaths, na
```

```
## # A tibble: 11 x 2
##    Condition.Group                                                COVID19_Deaths
##    <chr>                                                                   <dbl>
##  1 All other conditions and causes (residual)                              26357
```

```
##  2 Alzheimer disease                                                    11834
##  3 Circulatory diseases                                                196742
##  4 Diabetes                                                             32593
##  5 Intentional and unintentional injury, poisoning, and other ad~       14424
##  6 Malignant neoplasms                                                  25116
##  7 Obesity                                                              28579
##  8 Renal failure                                                        34375
##  9 Respiratory diseases                                                155191
## 10 Sepsis                                                               38029
## 11 Vascular and unspecified dementia                                    16059
```

### Number of covid-19 deaths after taking vaccine by age and manufacture

```
cat("Number of covid-19 deaths after taking vaccination by age")
```

```
## Number of covid-19 deaths after taking vaccination by age
```

```
death_age <- table(vaccine_died$AGE_YRS)
print(death_age)
```

```
##
## 0.42    1 1.08   11   12   13   15   16   17   18   19   20   21   22   23   24
##    2    1    1    1    1   10    8   10    7   29    9   19   23    8    9   13
##   25   26   27   28   29   30   31   32   33   34   35   36   37   38   39   40
##    9   30   23   22   20   21   15   19   18   16   51   74   60   58   42   35
##   41   42   43   44   45   46   47   48   49   50   51   52   53   54   55   56
##   41   35   32   63   68   44   77   81   64  104  108   69   75   92  122  106
##   57   58   59   60   61   62   63   64   65   66   67   68   69   70   71   72
##  135  195  173  167  282  229  248  250  298  280  344  253  320  323  270  357
##   73   74   75   76   77   78   79   80   81   82   83   84   85   86   87   88
##  439  388  314  336  459  352  372  358  354  380  413  368  339  346  321  299
##   89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104
##  307  326  214  176  177  184   95  128  103   53   21   29   24   10    7    3
##  105  106
##    1    1
```

```
cat("Number of covid-19 deaths after taking vaccination by manufacture")
```

```
## Number of covid-19 deaths after taking vaccination by manufacture
```

```
death_type <- table(vaccine_died$VAX_MANU)
print(death_type)
```

```
##
##              JANSSEN              MODERNA     PFIZER\\BIONTECH
##                 1368                 5940                 7074
## UNKNOWN MANUFACTURER
##                   54
```

```
cat("Number of covid-19 deaths after taking vaccination by state")
```

```
## Number of covid-19 deaths after taking vaccination by state
```

```
death_state <- table(vaccine_died$STATE)
print(death_state)
```

```
##
##        AK   AL   AR   AS   AZ   CA   CO   CT   DC   DE   FL   GA   GU   HI   IA
## 2383   36  117  149    3  156  864  157   83   23   31  819  513    5   47  114
##   ID   IL   IN   KS   KY   LA   MA   MD   ME   MI   MN   MO   MP   MS   MT   NC
##   31  459  224   95  891   91  166  150   59  652  393  258   15   72   67  209
##   ND   NE   NH   NJ   NM   NV   NY   OH   OK   OR   PA   PR   RI   SC   SD   TN
##   70  111  126  266   98   47  530  351   81  127  413  155   23  112   69  440
##   TX   UT   VA   VT   WA   WI   WV   WY   XB
## 1038   45  193   16  369  320   63   39    2
```

## *Limitations*

Some of the limitations are below

- The number of deaths reported by CDC may not be accurate. Only the deaths occurred in hospital certified by doctors are reported in the data set.
- Number of reports may increase in response to media attention
- I would want to analyze the percentage of vaccines given across the states and check for correlations between number of vaccine and deaths. However, I am unsure of how far I will get due to limitations in data.
- Moreover, as part of initial phase, Covid-19 vaccines are given only to the people who are 18 years older. So, the vaccine data sets used is not complete.
- The analysis on Covid-19 deaths by age and underlying condition has been done for Unites States as a whole. The same can be extended to state level analysis but not done due to limitation with data set.
- It is generally not possible to find out deaths from VARES data if a vaccine caused the adverse effect.

## *Concluding Remarks*

Based on analysis of the data sets extracted from CDC and VAERS, I conclude that adverse events caused by Covid-19 is high for the old people having age greater than 55 years compared to young people whose age is less than 55. In addition, the adverse effect caused by Covid-19 is high for the people having underlying condition compared to those who are healthy. The metrics and graphs generated out of this data sets also proving similar information. However, we need to keep in mind that the total number of death reported by CDC may not have complete information.