R Notebook

AI-A B3 G4 56-Somesh Kamnapure 57-Saurav Kamtalwar 59-Kartik Rupauliha 60-Kartik Rajput

Refer to covid_death_age_gender. 1. Set up a hypothesis about 'age group' and 'covid-19 deaths'. Undergo the process of hypothesis testing.Also 2. Set up a hypothesis about 'sex' and 'covid-19 deaths'

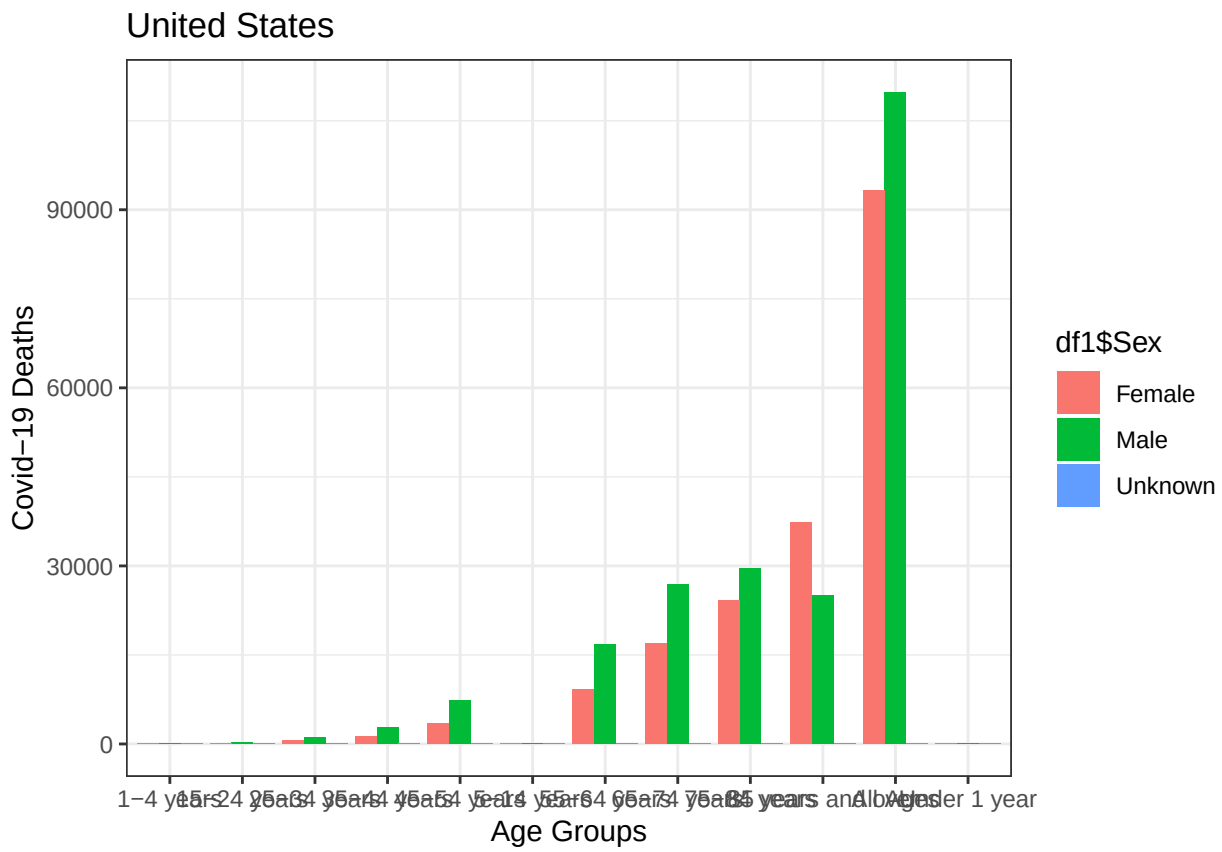
```
data<-read.csv(file.choose(), header = TRUE)
```

```
df <- data
df1 <- subset( df, State == "United States" )
df1 <- subset(df1, Sex!="All Sexes")
df1 <- subset(df1, Age.group!="0-17 years")
df1 <- subset(df1, Age.group!="18-29 years")
df1 <- subset(df1, Age.group!="30-49 years")
df1 <- subset(df1, Age.group!="50-64 years")
df2 = data.frame(df1$ Age.group, df1$ COVID.19.Deaths, df1$ Sex)

print(df2)
```

```
##        df1.Age.group df1.COVID.19.Deaths df1.Sex
## 1           All Ages               109838    Male
## 2        Under 1 year                   15    Male
## 3           1-4 years                    7    Male
## 4          5-14 years                   24    Male
## 5         15-24 years                  230    Male
## 6         25-34 years                 1047    Male
## 7         35-44 years                 2803    Male
## 8         45-54 years                 7377    Male
## 9         55-64 years                16809    Male
## 10        65-74 years                26924    Male
## 11        75-84 years                29620    Male
## 12 85 years and over                 24982    Male
## 13          All Ages                 93198  Female
## 14       Under 1 year                     7  Female
## 15          1-4 years                     8  Female
## 16         5-14 years                    13  Female
## 17        15-24 years                   144  Female
## 18        25-34 years                   541  Female
## 19        35-44 years                  1316  Female
## 20        45-54 years                  3458  Female
## 21        55-64 years                  9162  Female
## 22        65-74 years                 17002  Female
## 23        75-84 years                 24175  Female
## 24 85 years and over                  37372  Female
## 25          All Ages                     7 Unknown
## 26       Under 1 year                    0 Unknown
```

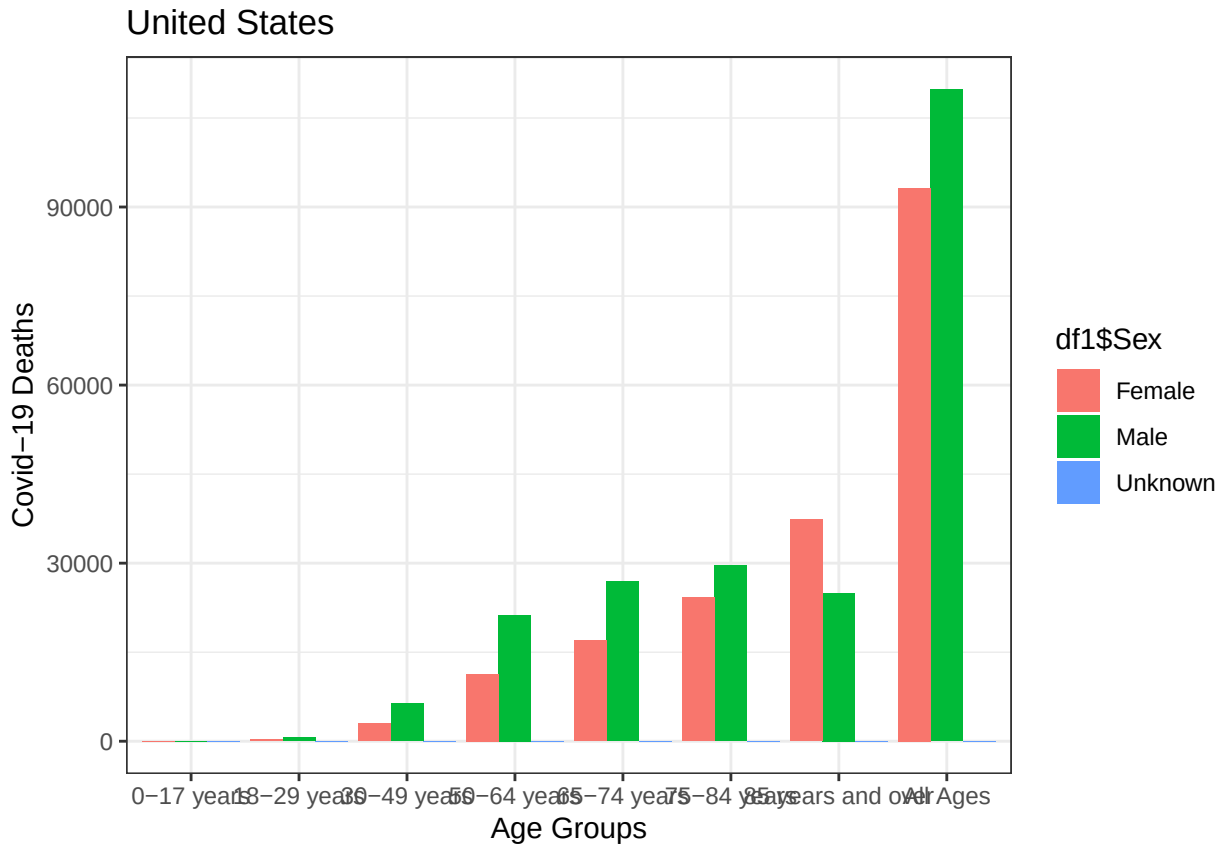
```
## 27         1-4 years                   0 Unknown
## 28        5-14 years                   0 Unknown
## 29       15-24 years                   0 Unknown
## 30       25-34 years                   0 Unknown
## 31       35-44 years                   0 Unknown
## 32       45-54 years                   2 Unknown
## 33       55-64 years                   0 Unknown
## 34       65-74 years                   1 Unknown
## 35       75-84 years                   1 Unknown
## 36 85 years and over                   3 Unknown
```

```
library(ggplot2)

ggplot(df2, aes(x=df1$ Age.group, y= df1$ COVID.19.Deaths, fill=df1$ Sex)) + geom_bar(stat="identity", p
```




```
df1 <- subset( df, State == "United States" )
df1 <- subset(df1, Sex!="All Sexes")
df1 <- subset(df1, Age.group!="Under 1 year")
df1 <- subset(df1, Age.group!="1-4 years")
df1 <- subset(df1, Age.group!="5-14 years")
df1 <- subset(df1, Age.group!="15-24 years")
df1 <- subset(df1, Age.group!="25-34 years")
df1 <- subset(df1, Age.group!="35-44 years")
df1 <- subset(df1, Age.group!="45-54 years")
df1 <- subset(df1, Age.group!="55-64 years")
```

```
df2 = data.frame(df1$ Age.group, df1$ COVID.19.Deaths, df1$ Sex)

#print(df2)


library(ggplot2)

ggplot(df2, aes(x=df1$ Age.group, y= df1$ COVID.19.Deaths, fill=df1$ Sex)) + geom_bar(stat="identity", p
```

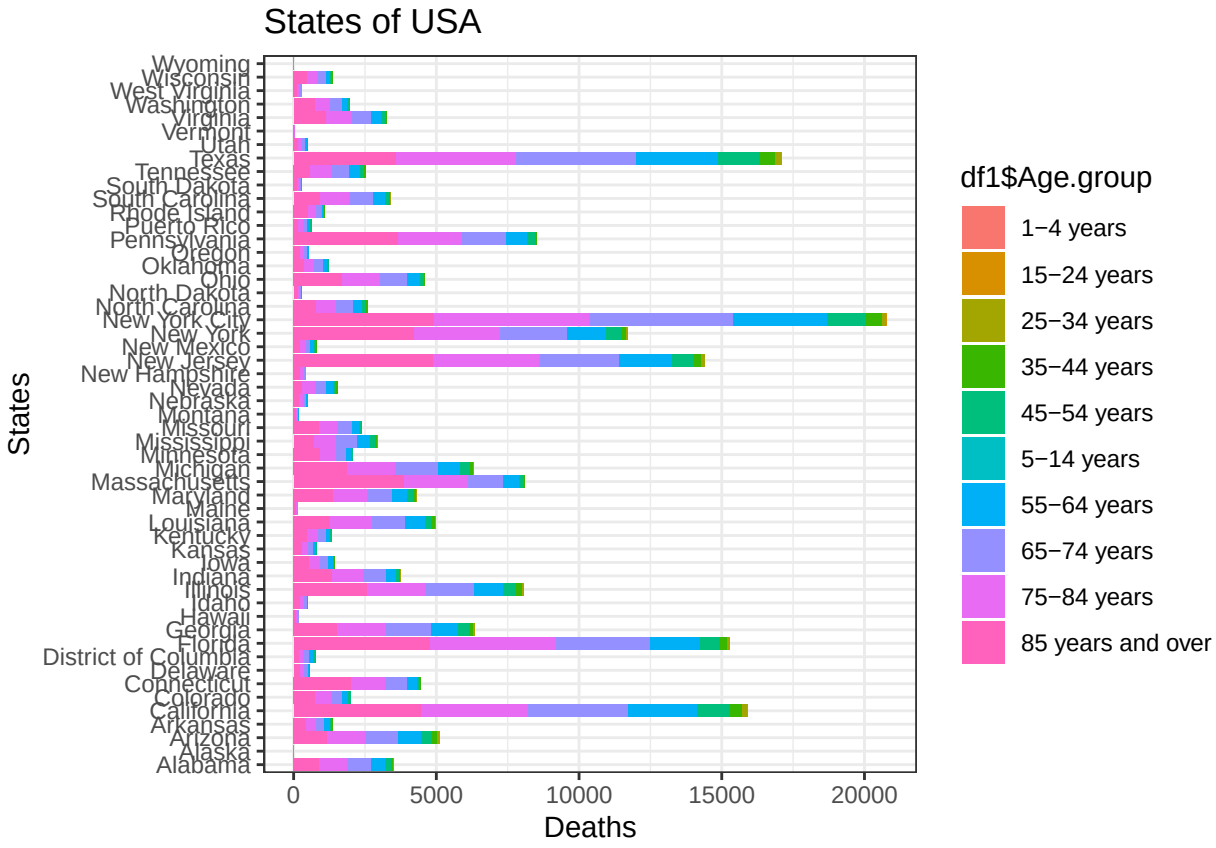


1a.Statement: H0-The means of Aged people and non aged people dying due to covid 19 is equal. Ha-The means of Aged people and non aged people dying due to covid 19 is not equal.

```
df1 <- subset( df, Sex!="All Sexes" )
df1 <- subset( df1, Age.group!="All Ages" )
df1 <- subset(df1, Age.group!="Under 1 year")
df1 <- subset(df1, Age.group!="0-17 years")
df1 <- subset(df1, Age.group!="18-29 years")
df1 <- subset(df1, Age.group!="30-49 years")
df1 <- subset(df1, Age.group!="50-64 years")
df1 <- subset( df1, State!="United States" )

df2 = data.frame(df1$State, df1$ COVID.19.Deaths, df1$Age.group )


library(ggplot2)
```

```
ggplot(df2, aes(x=df1$ COVID.19.Deaths, y= df1$ State, fill=df1$ Age.group)) + geom_bar(stat="identity")
```

```
## Warning: Removed 246 rows containing missing values (position_stack).
```




```
aged <- df1[df1$Age.group == "65-74 years" | df1$Age.group == "75-84 years" | df1$Age.group == "85 years
nonaged <- df1[df1$Age.group != "65-74 years" & df1$Age.group != "75-84 years" & df1$Age.group != "85 y
#head(aged)
#head(nonaged)

t.test(nonaged$COVID.19.Deaths, aged$COVID.19.Deaths)
```

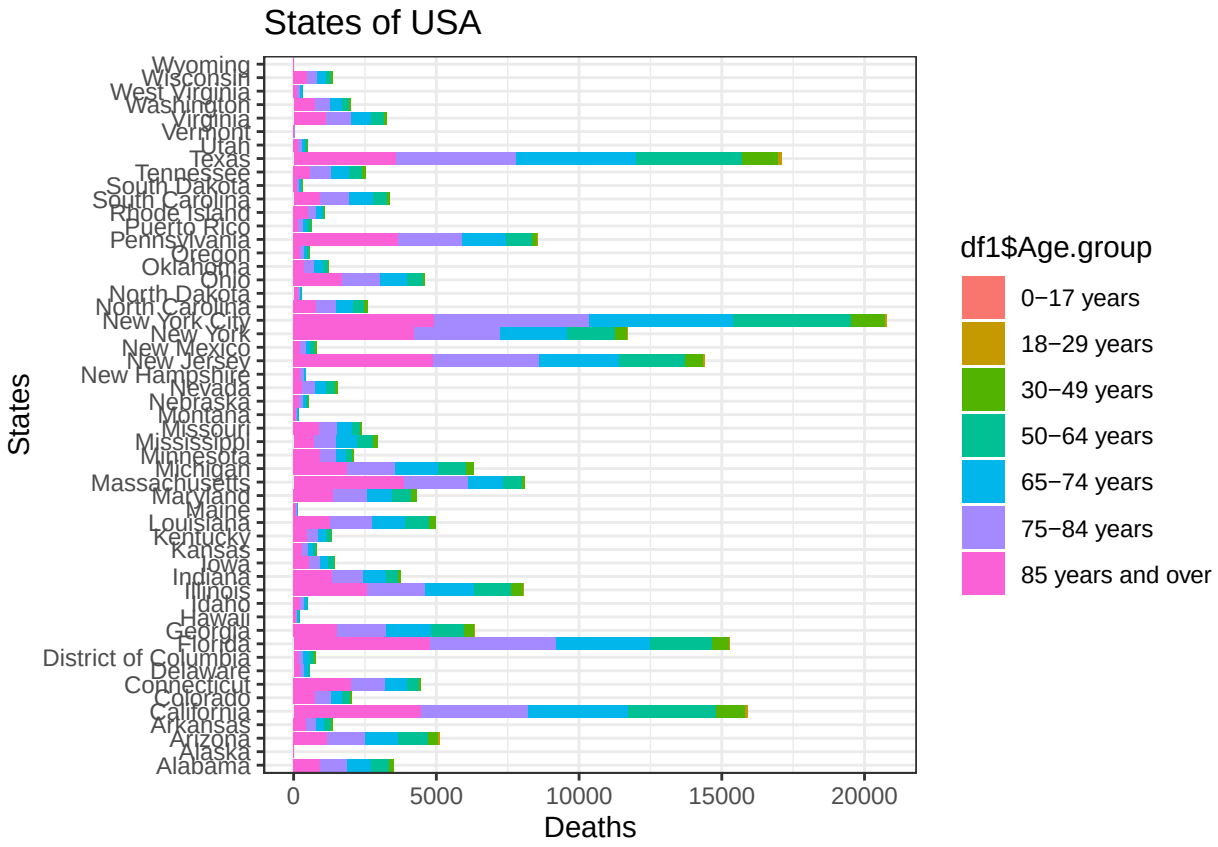
```
##
##  Welch Two Sample t-test
##
## data:  nonaged$COVID.19.Deaths and aged$COVID.19.Deaths
## t = -10.597, df = 493.84, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -359.6443 -247.1377
## sample estimates:
## mean of x mean of y
##  47.72717 351.11816
```

Null hypothesis for alpha = 0.05 can be rejected

1b.Statement: H0-The means of Aged people and non aged people dying due to covid 19 is equal. Ha-The means of Aged people and non aged people dying due to covid 19 is not equal.

```
df1 <- subset( df, Sex!="All Sexes" )
df1 <- subset( df1, Age.group!="All Ages" )
df1 <- subset(df1, Age.group!="Under 1 year")
df1 <- subset(df1, Age.group!="1-4 years")
df1 <- subset(df1, Age.group!="5-14 years")
df1 <- subset(df1, Age.group!="15-24 years")
df1 <- subset(df1, Age.group!="25-34 years")
df1 <- subset(df1, Age.group!="35-44 years")
df1 <- subset(df1, Age.group!="45-54 years")
df1 <- subset(df1, Age.group!="55-64 years")
df1 <- subset( df1, State!="United States" )

df2 = data.frame(df1$State, df1$ COVID.19.Deaths, df1$Age.group )


library(ggplot2)

ggplot(df2, aes(x=df1$ COVID.19.Deaths, y= df1$ State, fill=df1$ Age.group)) + geom_bar(stat="identity")
```

```
## Warning: Removed 156 rows containing missing values (position_stack).
```

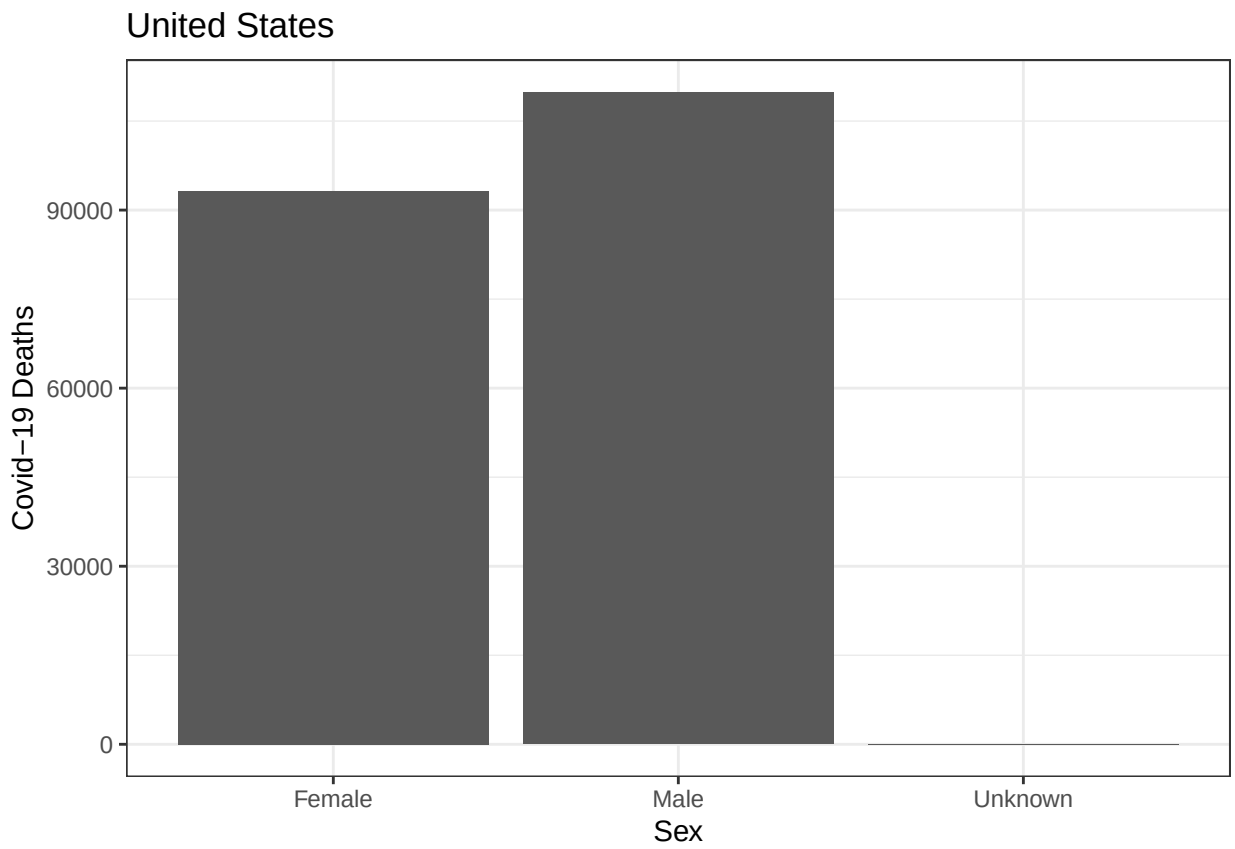


```r
aged <- df1[df1$Age.group == "65-74 years" | df1$Age.group == "75-84 years" | df1$Age.group == "85 years
nonaged <- df1[df1$Age.group != "65-74 years" & df1$Age.group != "75-84 years" & df1$Age.group != "85 y
#head(aged)
#head(nonaged)

t.test(nonaged$COVID.19.Deaths, aged$COVID.19.Deaths)
```

```
##
##  Welch Two Sample t-test
##
## data:  nonaged$COVID.19.Deaths and aged$COVID.19.Deaths
## t = -8.7143, df = 617.79, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -325.6579 -205.8744
## sample estimates:
## mean of x mean of y
##   85.3520  351.1182
```

Null hypothesis for alpha = 0.05 can be rejected

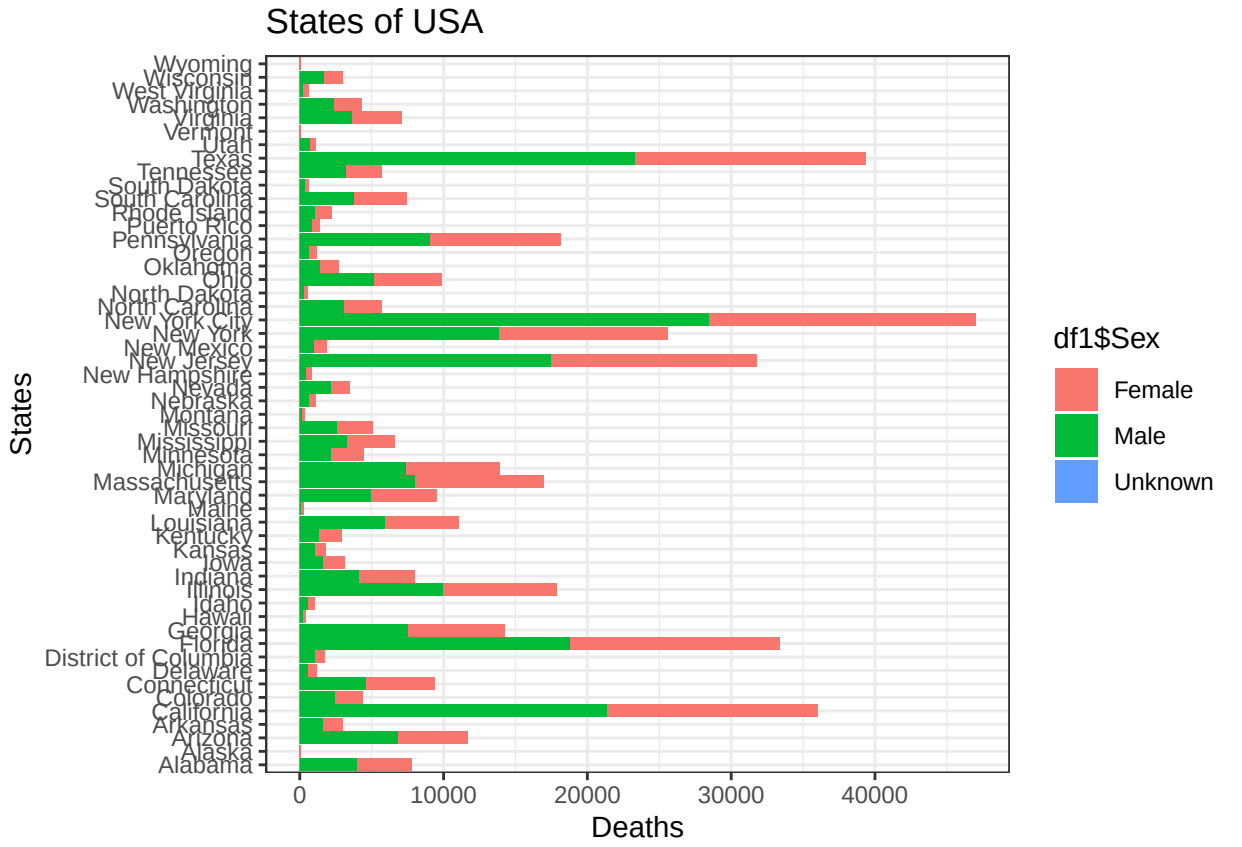
```r
df1 <- subset( df, State == "United States" )
df1 <- subset(df1, Age.group == "All Ages")
df1 <- subset( df1, Sex!= "All Sexes" )

df2 = data.frame(df1$ COVID.19.Deaths, df1$ Sex)
#print(df2)


library(ggplot2)

ggplot(df2, aes(x=df1$ Sex, y= df1$ COVID.19.Deaths)) + geom_bar(stat="identity") + theme_bw() + labs(t
```



2.Statement: H0-The means of Males and Females dying due to covid 19 is equal. Ha-The means of Males and Females dying due to covid 19 is not equal.

```
df1 <- subset( df, Age.group == "All Ages" )
df1 <- subset( df, State!= "United States" )
df1 <- subset( df1, Sex!= "All Sexes" )


df2 = data.frame(df1$State, df1$ COVID.19.Deaths, df1$Age.group, df1$Sex )
#print(df2)

library(ggplot2)

ggplot(df2, aes(x=df1$ COVID.19.Deaths, y= df1$ State, fill=df1$ Sex)) + geom_bar(stat="identity") + th
```

```
## Warning: Removed 407 rows containing missing values (position_stack).
```




```
male <- subset(df1, Sex== "Male")
female <- subset(df1, Sex== "Female")
t.test(male$COVID.19.Deaths, female$COVID.19.Deaths, conf.level = 0.95, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  male$COVID.19.Deaths and female$COVID.19.Deaths
## t = 1.0858, df = 1303, p-value = 0.2778
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -46.12001 160.45393
## sample estimates:
## mean of x mean of y
##  372.1199  314.9530
```

Null hypothesis for alpha = 0.05 can not be rejected

3.Statement: H0 - Average no. Of covid 19 deaths per day in age group 45-54 is 400 Ha - Average no. Of covid 19 deaths per day in age group 45-54 is less than what it is claimed

```
d1<-data.frame(data$Age.group[data$Age.group=="45-54 years"], data$COVID.19.Deaths[data$Age.group=="45-5
colnames(d1) <- c("age_group", "covid19deaths")
d1<-na.omit(d1)
```

```r
x_bar <- mean(d1$covid19deaths, na.rm=TRUE)
cat("Sample mean xbar",x_bar,"\n")
```

```
## Sample mean xbar 236.8613
```

```r
std <- sd(d1$covid19deaths, na.rm=TRUE)
cat("Standard deviationr",std,"\n")
```

```
## Standard deviationr 1152.387
```

```r
n<-nrow(d1)
mue <- 400

SE <- std/sqrt(n)
cat("Standard Error",SE,"\n")
```

```
## Standard Error 98.4551
```

```r
Z <- (x_bar-mue)/SE
cat("Zee Scorer",Z,"\n")
```

```
## Zee Scorer -1.656986
```

```r
p<-pnorm(Z,0,1,lower.tail = TRUE)
cat("value of distribution",p,"\n")
```

```
## value of distribution 0.04876119
```

Null hypothesis for alpha = 0.05 can be rejected

4.Statement: H0 - Average no. Of covid 19 deaths per day in Males is 700 Ha - Average no. Of covid 19 deaths per day in Males is less than what it is claimed

```r
d2<-data.frame(data$Sex[data$Sex=="Male"], data$COVID.19.Deaths[data$Sex=="Male"])
colnames(d2) <- c("Sex", "covid19deaths")
d2<-na.omit(d2)
```

```r
x_bar <- mean(d2$covid19deaths, na.rm=TRUE)
cat("Sample mean xbar",x_bar,"\n")
```

```
## Sample mean xbar 726.489
```

```r
std <- sd(d2$covid19deaths, na.rm=TRUE)
cat("Standard deviationr",std,"\n")
```

```
## Standard deviationr 4778.771
```

```r
n<-nrow(d2)
mue <- 700

SE <- std/sqrt(n)
cat("Standard Error",SE,"\n")
```

```
## Standard Error 182.8546
```

```r
Z <- (x_bar-mue)/SE
cat("Zee Scorer",Z,"\n")
```

```
## Zee Scorer 0.1448639
```

```r
p<-pnorm(Z,0,1,lower.tail = TRUE)
cat("value of distribution",p,"\n")
```

```
## value of distribution 0.5575908
```

Null hypothesis for alpha = 0.05 can not be rejected