BRFSS-Healthcare Data Analytics using R

Presented By,

Harsh Raizada

Presentation Path……

- **Rationale for selection and description of dataset.**
- **Description of analysis procedure and objectives.**
- **Objectives**
- **R coding and output for the objective analysis.**

Rationale for selection and description of dataset.

- Because of my past education and working experience in Healthcare I have selected The Behavioral Risk Factor Surveillance System (BRFSS) open data set, which is the United states premier system of health-related telephone surveys that collect data about U.S. residents (18 year or older) regarding their health-related risk behaviours, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

- What is risk factor surveillance?

Keeping track of the rates of risk factors which are the things or states in our daily lives that confers risk to our health is defined as a Risk Factor Surveillance.

- By collecting behavioural health risk data at the federal and state level, BRFSS has become a powerful tool for targeting and designing health promotion activities for the US population.

Description of analysis procedure and objectives:

- I have tried to done Descriptive analytics which usually aims at developing population based rates (or percentage) with this data set.
- Originally BRFSS data set were having sample of 486303 US residents (observations) and 275 variables in the data set but then I have removed unnecessary columns from the data by sub setting the data set after which I have left with 421192 observation with 13 variables (which were required for my analysis). Data is pretty cleaned so only one variable (General Health) needs data cleaning for the analysis

- Objective-1 To find out the perception of the people about their existing general health status among US residents.
- Objective-2 To find out the existing health care coverage rate among US residents.
- Objective-3 To find out the existing exercise or physical activity rates among US residents.
- Objective-4 To find out the existing smoking rate of respondents who have ever smoked in their life.
- Objective-5 To find out the current smoking status of the respondent who ever smoked in their life.
- Objective -6 To find out the existing coverage rate of flu vaccine among US residents.
- Objective-7 To find out the existing seat belt usage practices (or rate) among US residents.

Preliminary R coding before objective analysis

```
getwd()
setwd("/Users/harsh/Desktop/R-project")
library(SASxport) #installed package to read .XPT file
library(ggplot2) # installed package for plots
library(dplyr) # For grouping of data and other functions
library(car) # To recode the variable responses
library(ggthemes) # installed theme package for plots
brfss=read.xport("LLCP2016.XPT ") # reading of LLCP2016.XPT file into brfss reading file
cbrfss=brfss #copy of data set brfss
cbrfss[] <- lapply(cbrfss, unclass) # l apply to change from labelled integer to integer
brfssci=subset(cbrfss,DISPCODE==1100) #Data frame having those respondent data who has completed interview
brfssvarList <- c("GENHLTH","HLTHPLN1","EXERANY2","SEX","MARITAL","EDUCA",
                  "VETERAN3","EMPLOY1","SMOKE100","SMOKDAY2","FLUSHOT6","SEATBELT"
                  ,"ADDEPEV2") # working variable list by removing unecessary variables
brfssciwd=brfssci[brfssvarList]
str(brfssciwd)
```

Objective-1 To find out the perception of the people about their existing general health status among US residents.

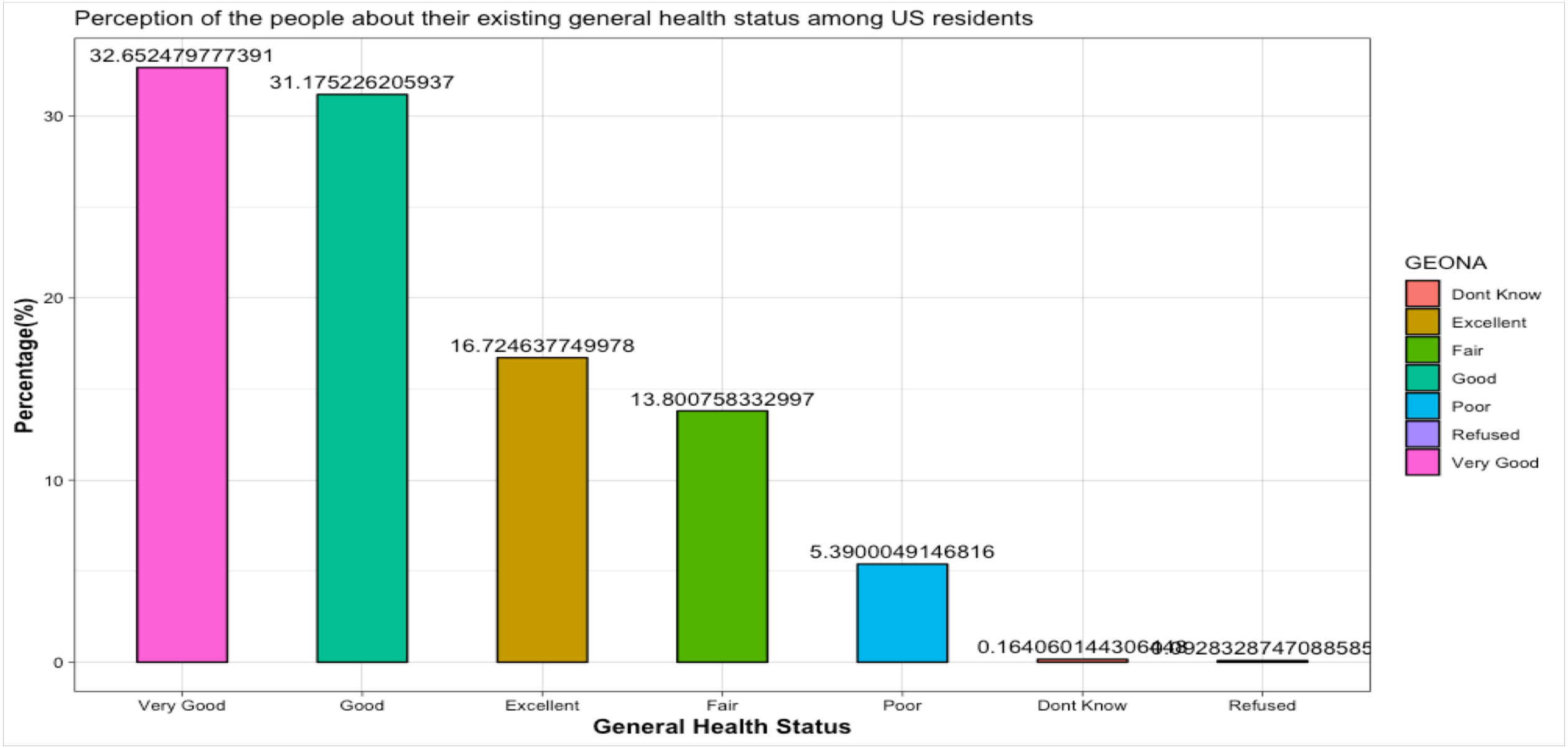
R coding-

```
GEONA=na.omit(brfssciwd$GENHLTH) # Data cleaning by removing NA
GEONADF=as.data.frame(GEONA) # converted to data frame
GE=group_by(GEONADF,GEONA) # group by function
GE1=dplyr::summarise(GE,Frequency=n())%>% mutate(Percentage=Frequency/sum(Frequency)*100) #summarise function

GE1$GEONA=recode(GE1$GEONA,"1='Excellent';2='Very Good';3='Good';4='Fair';5='Poor';7='Dont Know';9='Refused'")

ggplot(aes(x=reorder(GEONA,-Percentage),y=Percentage),data = GE1)+
  geom_bar(stat="identity",width = 0.5,aes(fill=GEONA),colour="black")+
  ggtitle('Perception of the people about their existing general health status among US residents')+
  geom_text(aes(label=Percentage,vjust=-0.5))+
  xlab('General Health Status')+
  ylab("Percentage(%)")+
  theme_linedraw()+
  theme(plot.title = element_text(size = 12),
        axis.title = element_text(size = 12,face="bold"))
```

Output-


Perception of the people about their existing general health status among US residents
32.652479777391
31.175226205937
16.724637749978
13.800758332997
5.3900049146816
0.16406014430644
0.08928328747088585
30
20
10
0
Percentage(%)
Very Good
Good
Excellent
Fair
Poor
Dont Know
Refused
General Health Status
GEONA
Dont Know
Excellent
Fair
Good
Poor
Refused
Very Good

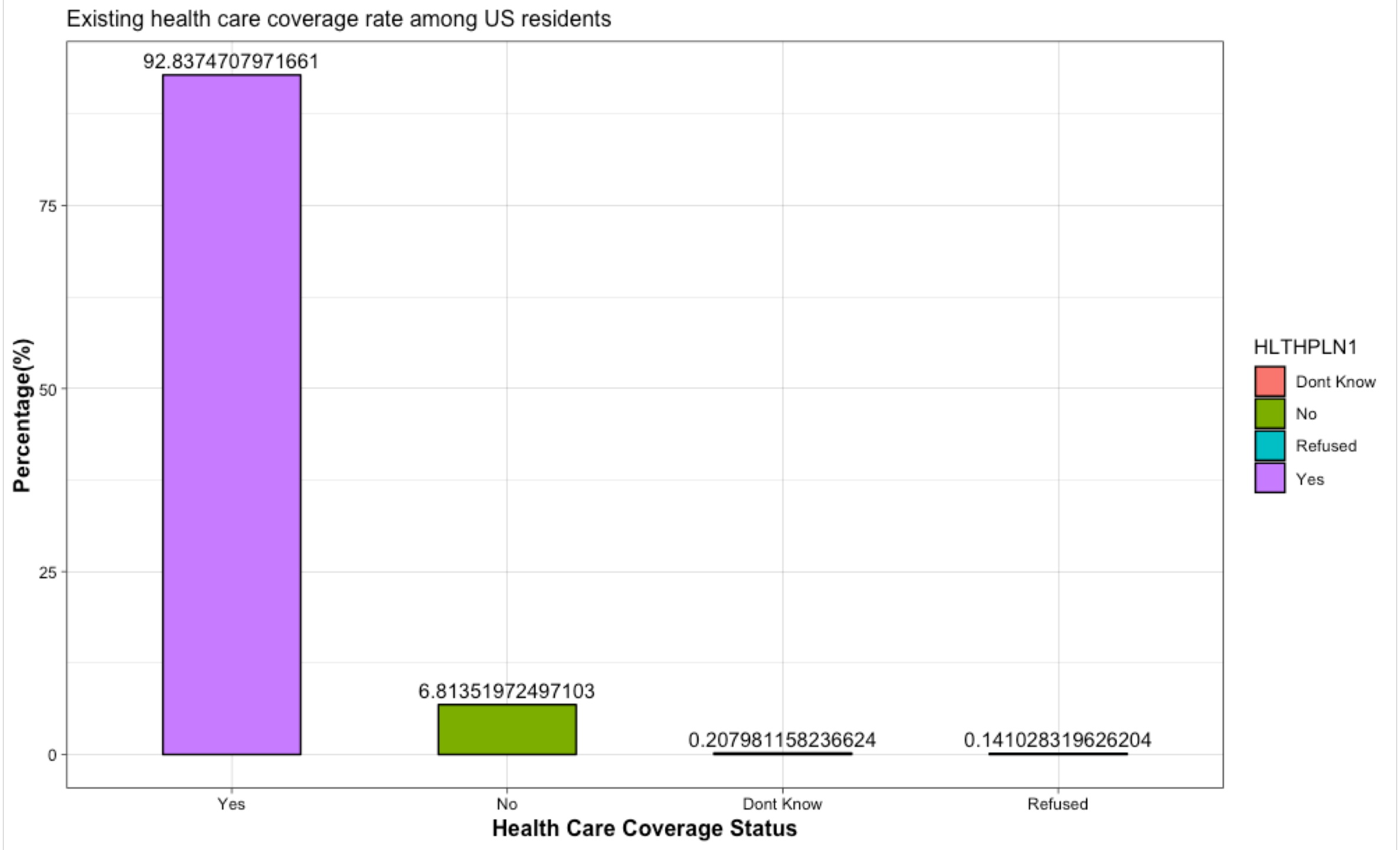
Conclusion-

- Around 80% of the respondents perceives that there health conditions are very good.
- 5 % perceives that there health condition is not good.

Objective-2 To find out the existing health care coverage rate among US residents.

R coding-

```
HLTP=group_by(brfssciwd,HLTHPLN1)
HLTP1=dplyr::summarise(HLTP,Frequency=n())%>%
mutate(Percentage=Frequency/sum(Frequency)*100)
HLTP1$HLTHPLN1=recode(HLTP1$HLTHPLN1,"1='Yes';2='No';7='Dont Know';9='Refused'")
ggplot(aes(x=reorder(HLTHPLN1,-Percentage),y=Percentage),data = HLTP1)+
  geom_bar(stat="identity",width = 0.5,aes(fill=HLTHPLN1),colour="black")+
  ggtitle('Existing health care coverage rate among US residents')+
  geom_text(aes(label=Percentage,vjust=-0.5))+
  xlab('Health Care Coverage Status')+
  ylab("Percentage(%)")+
  theme_linedraw()+
  theme(plot.title = element_text(size = 12),
        axis.title = element_text(size = 12,face="bold"))
```


Existing health care coverage rate among US residents
92.8374707971661
6.81351972497103
0.207981158236624
0.141028319626204
75
50
25
0
Percentage(%)
Yes
No
Dont Know
Refused
Health Care Coverage Status
HLTHPLN1
Dont Know
No
Refused
Yes

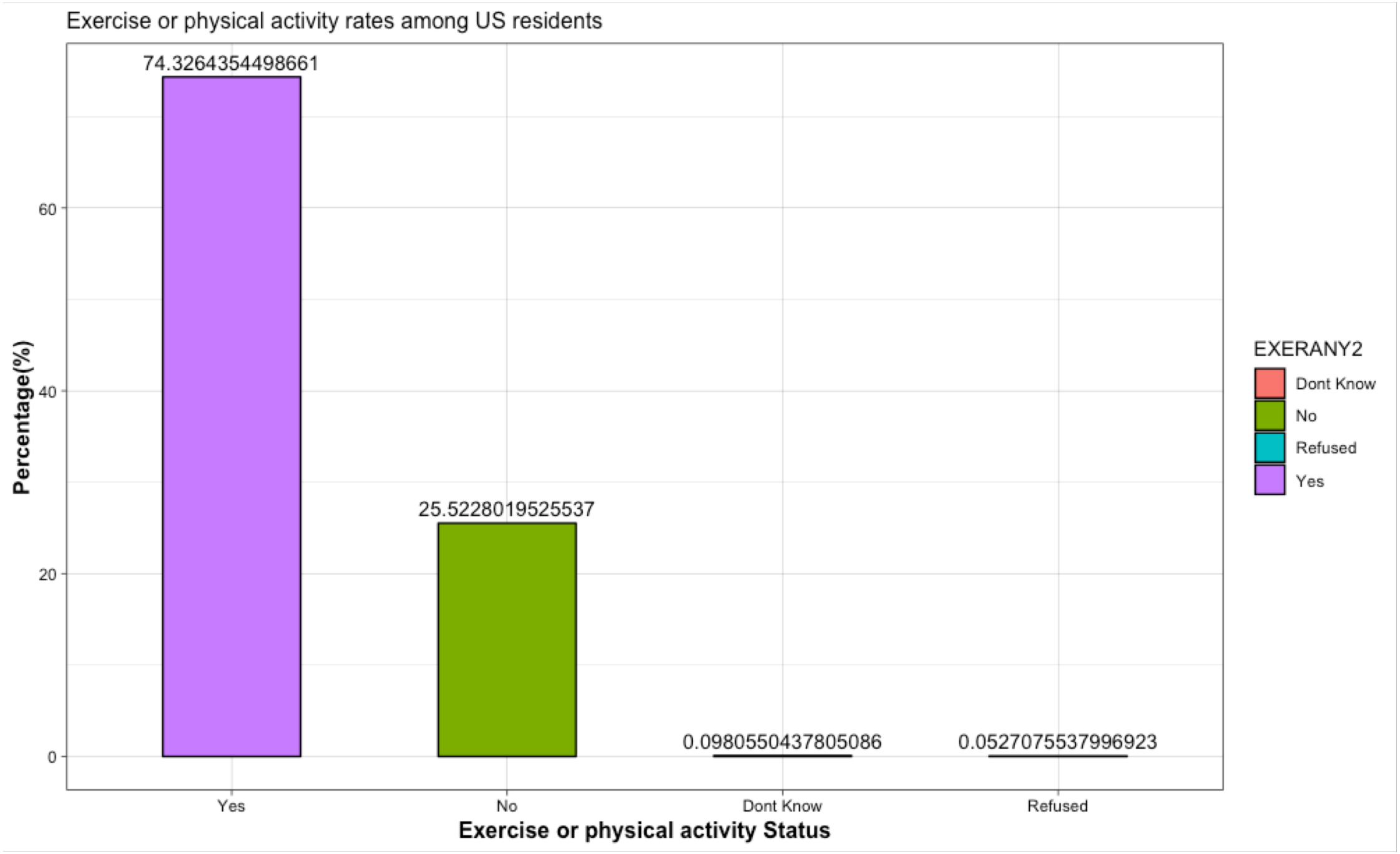
Conclusion-

- 92.83% of total respondents have their health coverage but still 7 % do not have it.

Objective-3 To find out the existing exercise or physical activity rates among US residents.

R coding-

```
EXE=group_by(brfsciwd,EXERANY2)
EXE1=dplyr::summarise(EXE,Frequency=n())%>%
mutate(Percentage=Frequency/sum(Frequency)*100)
EXE1$EXERANY2=recode(EXE1$EXERANY2,"1='Yes';2='No';7='Dont Know';9='Refused'")
ggplot(aes(x=reorder(EXERANY2,-Percentage),y=Percentage),data = EXE1)+
  geom_bar(stat="identity",width = 0.5,aes(fill=EXERANY2),colour="black")+
  ggtitle('Exercise or physical activity rates among US residents')+
  geom_text(aes(label=Percentage,vjust=-0.5))+
  xlab('Exercise or physical activity Status')+
  ylab("Percentage(%)")+
  theme_linedraw()+
  theme(plot.title = element_text(size = 12),
        axis.title = element_text(size = 12,face="bold"))
```


Exercise or physical activity rates among US residents
74.3264354498661
25.5228019525537
0.0980550437805086
0.0527075537996923
Percentage(%)
0
20
40
60
Yes
No
Dont Know
Refused
Exercise or physical activity Status
EXERANY2
Dont Know
No
Refused
Yes

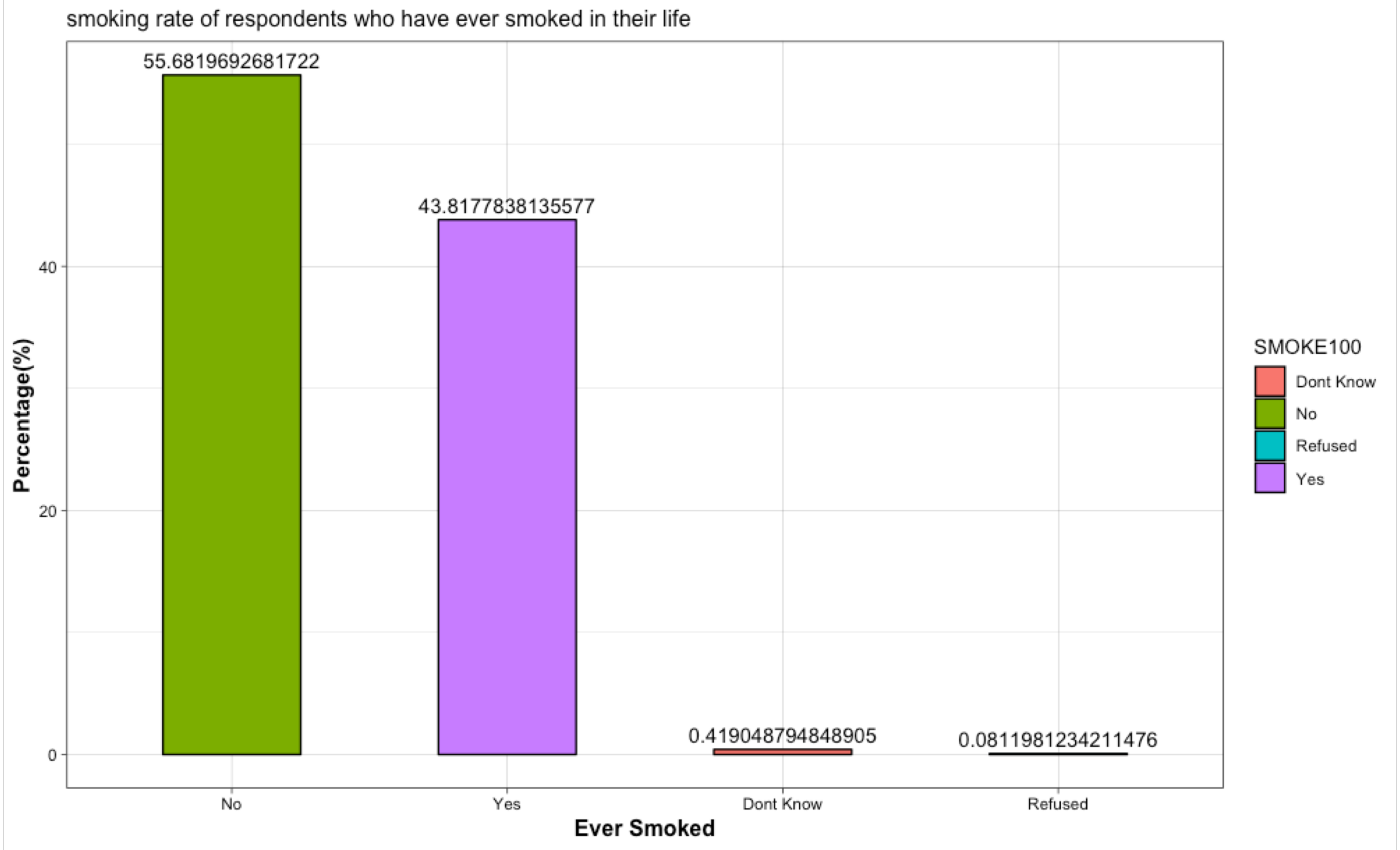
Conclusion-

- 75 % of total respondents are involved in any kind of exercise or physical activity which shows that they are conscious about their health. This we can relate with their perception towards their health status (objective-1) where 80 % perceives that they are healthy.
- But still 25.50% of people are not involved in any kind of physical activity.

Objective-4 To find out the existing smoking rate of respondents who have ever smoked in their life.

R coding-

```
SMO=group_by(brfsciwd,SMOKE100)
SMO1=dplyr::summarise(SMO,Frequency=n())%>%
mutate(Percentage=Frequency/sum(Frequency)*100)
SMO1$SMOKE100=recode(SMO1$SMOKE100,"1='Yes';2='No';7='Dont Know';9='Refused'")
ggplot(aes(x=reorder(SMOKE100,-Percentage),y=Percentage),data = SMO1)+
  geom_bar(stat="identity",width = 0.5,aes(fill=SMOKE100),colour="black")+
  ggtitle('smoking rate of respondents who have ever smoked in their life')+
  geom_text(aes(label=Percentage,vjust=-0.5))+
  xlab('Ever Smoked')+
  ylab("Percentage(%)")+
  theme_linedraw()+
  theme(plot.title = element_text(size = 12),
        axis.title = element_text(size = 12,face="bold"))
```


smoking rate of respondents who have ever smoked in their life
55.6819692681722
43.8177838135577
0.419048794848905
0.0811981234211476
40
20
0
Percentage(%)
No
Yes
Dont Know
Refused
Ever Smoked
SMOKE100
Dont Know
No
Refused
Yes

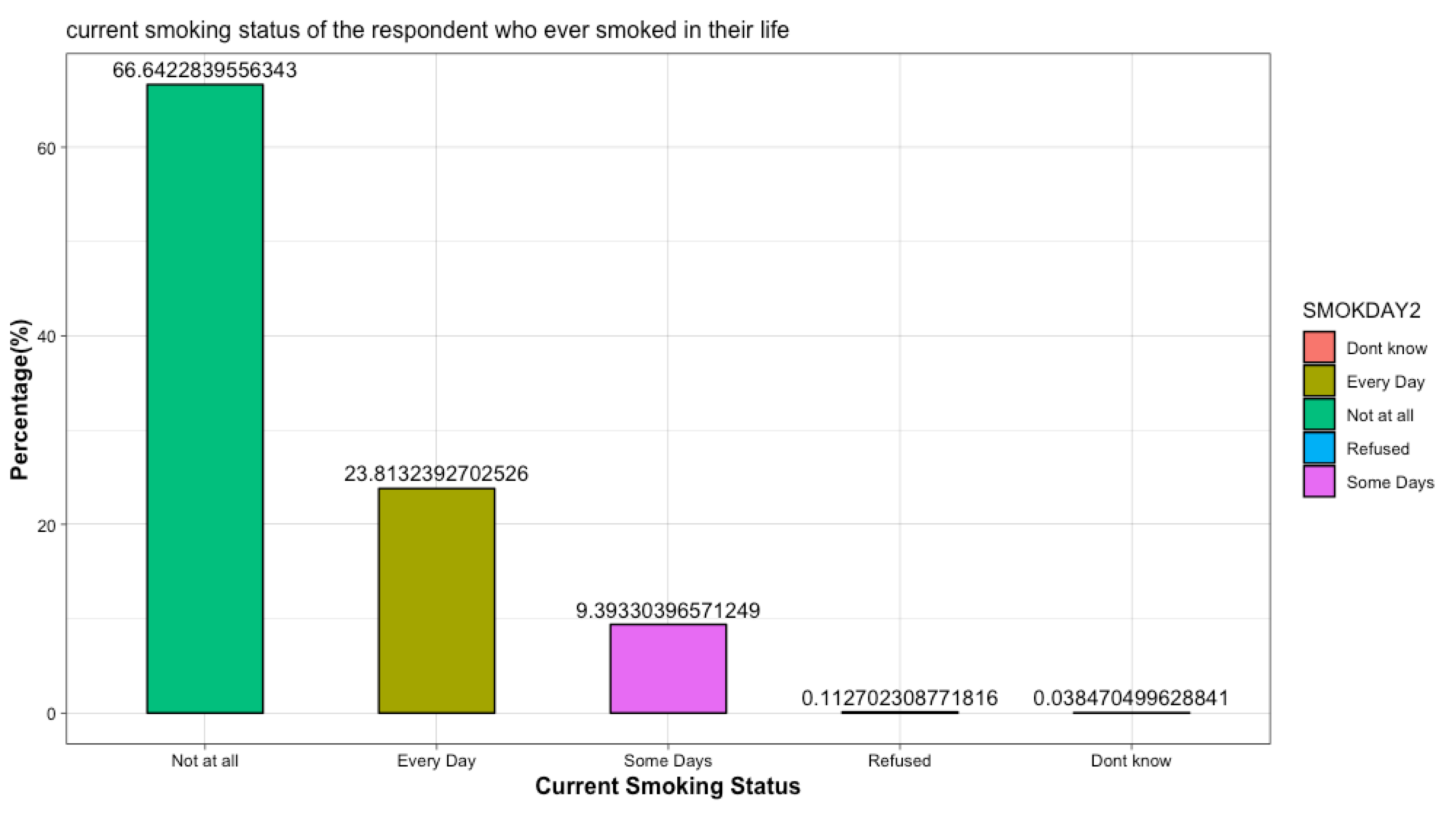
Conclusion-

- Among total respondents 43.81% have ever smoked in their life.
- But 55.68 never smoked which shows awareness regarding health hazards of tobacco or smoking.

Objective-5 To find out the current smoking status of the respondent who ever smoked in their life.

R coding-

```
SS=subset(brfssciwd,SMOKE100==1)
SS1=group_by(SS,SMOKDAY2)
SS2=dplyr::summarise(SS1,Frequency=n())%>% mutate(Percentage=Frequency/sum(Frequency)*100)
SS2$SMOKDAY2=recode(SS2$SMOKDAY2,"1='Every Day';2='Some Days';3='Not at all';7='Dont know';9='Refused'")
ggplot(aes(x=reorder(SMOKDAY2,-Percentage),y=Percentage),data = SS2)+
  geom_bar(stat="identity",width = 0.5,aes(fill=SMOKDAY2),colour="black")+
  ggtitle('current smoking status of the respondent who ever smoked in their life')+
  geom_text(aes(label=Percentage,vjust=-0.5))+
  xlab('Current Smoking Status')+
  ylab("Percentage(%)")+
  theme_linedraw()+
  theme(plot.title = element_text(size = 12),
        axis.title = element_text(size = 12,face="bold"))
```


current smoking status of the respondent who ever smoked in their life
66.6422839556343
23.8132392702526
9.39330396571249
0.112702308771816
0.038470499628841
60
40
20
0
Percentage(%)
Not at all
Every Day
Some Days
Refused
Dont know
Current Smoking Status
SMOKDAY2
Dont know
Every Day
Not at all
Refused
Some Days

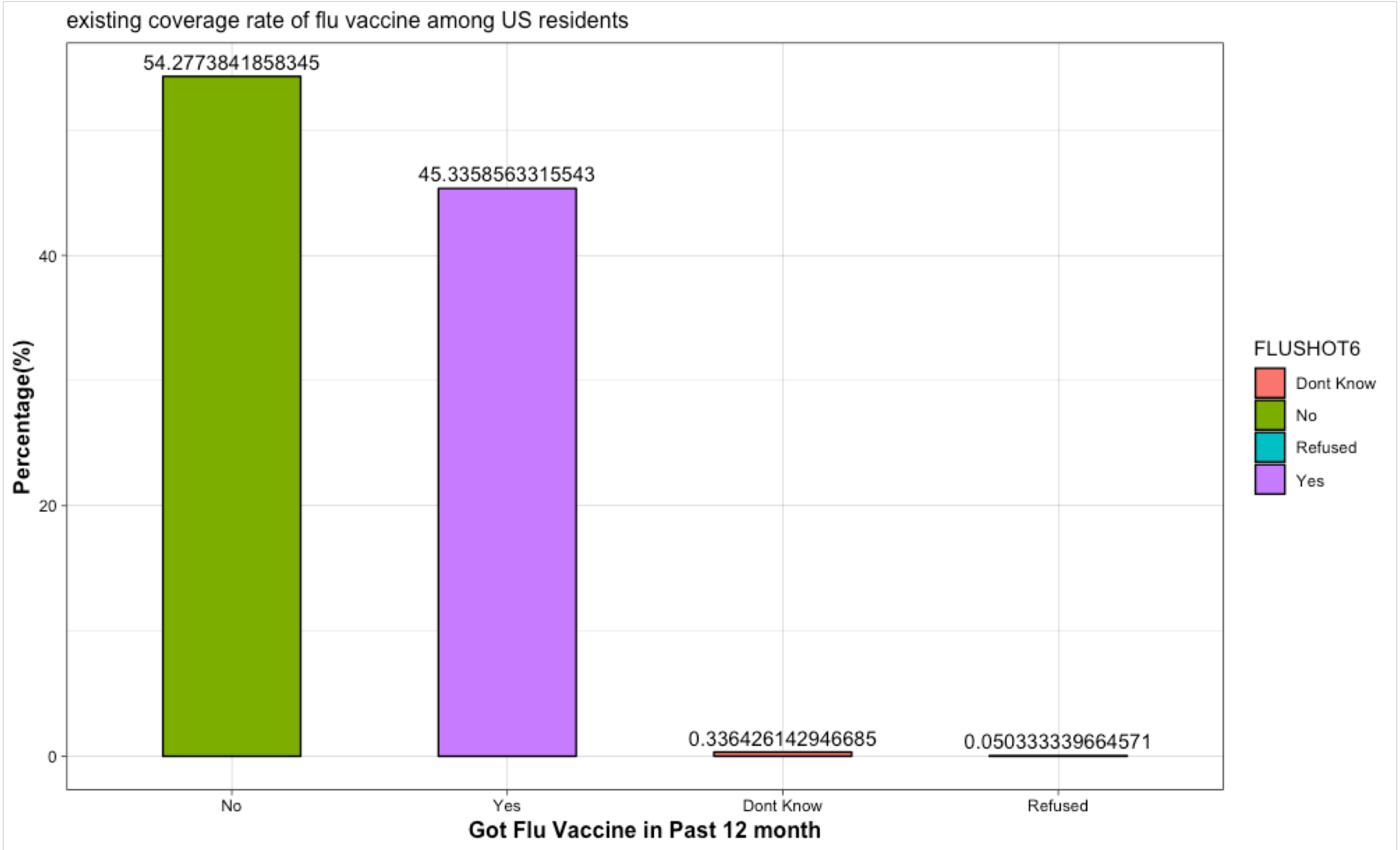
Conclusion-

- Among total respondents who have ever smoked in their life, 23.81% are still smoking everyday and 9 % are smoking somedays.
- But 66.64% quitted smoking which shows their awareness regarding health hazards of tobacco or smoking.

Objective -6 To find out the existing coverage rate of flu vaccine among US residents.

R coding-

```
FLU=group_by(brfssciwd,FLUSHOT6)
FLU1=dplyr::summarise(FLU,Frequency=n())%>%
mutate(Percentage=Frequency/sum(Frequency)*100)
FLU1$FLUSHOT6=recode(FLU1$FLUSHOT6,"1='Yes';2='No';7='Dont Know';9='Refused'")
ggplot(aes(x=reorder(FLUSHOT6,-Percentage),y=Percentage),data = FLU1)+
  geom_bar(stat="identity",width = 0.5,aes(fill=FLUSHOT6),colour="black")+
  ggtitle('existing coverage rate of flu vaccine among US residents')+
  geom_text(aes(label=Percentage,vjust=-0.5))+
  xlab('Got Flu Vaccine in Past 12 month')+
  ylab("Percentage(%)")+
  theme_linedraw()+
  theme(plot.title = element_text(size = 12),
        axis.title = element_text(size = 12,face="bold"))
```


existing coverage rate of flu vaccine among US residents
54.2773841858345
45.3358563315543
0.336426142946685
0.050333339664571
40
20
0
Percentage(%)
No
Yes
Dont Know
Refused
Got Flu Vaccine in Past 12 month
FLUSHOT6
Dont Know
No
Refused
Yes

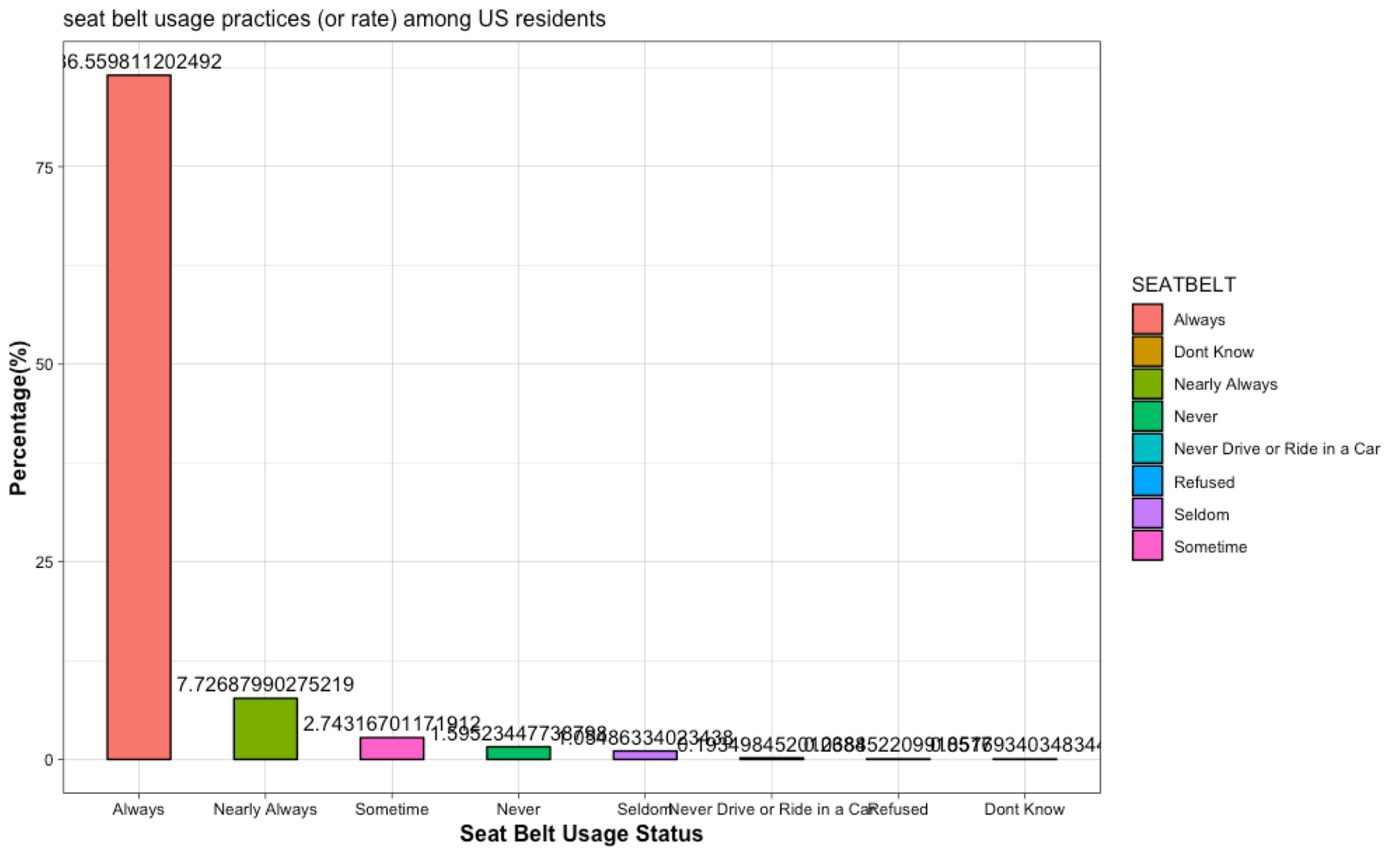
Conclusion-

- Among total respondents only 45% of respondents have flu shot in past 12 month which makes rest of the population (54.27%) more prone to flu infection.
- Public Health department should intervene for this situation.

Objective-7 To find out the existing seat belt usage practices (or rate) among US residents.

R coding-

```
SB=group_by(brfssciwd,SEATBELT)
SB1=dplyr::summarise(SB,Frequency=n())%>% mutate(Percentage=Frequency/sum(Frequency)*100)
SB1$SEATBELT=recode(SB1$SEATBELT,"1='Always';2='Nearly Always';3='Sometime';4='Seldom';5='Never';7='Dont Know';8='Never Drive or Ride in a Car';9='Refused'")
ggplot(aes(x=reorder(SEATBELT,-Percentage),y=Percentage),data = SB1)+
  geom_bar(stat="identity",width = 0.5,aes(fill=SEATBELT),colour="black")+
  ggtitle('seat belt usage practices (or rate) among US residents')+
  geom_text(aes(label=Percentage,vjust=-0.5))+
  xlab('Seat Belt Usage Status')+
  ylab("Percentage(%)")+
  theme_linedraw()+
  theme(plot.title = element_text(size = 12),
        axis.title = element_text(size = 12,face="bold"))
```



seat belt usage practices (or rate) among US residents
7.72687990275219
Percentage(%)
75
50
25
0
Always
Nearly Always
Sometime
Never
Seldom
Never Drive or Ride in a Car
Refused
Dont Know
Seat Belt Usage Status
SEATBELT
Always
Dont Know
Nearly Always
Never
Never Drive or Ride in a Car
Refused
Seldom
Sometime

Conclusion-

- Among total respondents 94.27 % of them are using seatbelts while driving in a car which depicts high awareness rate about road safety.

THANK YOU!