```{r}

library(tidyverse)
data <- read.csv(file.choose(),header = T)
head(data)

```

Data Cleaning
```{r}
a<- sum(is.na(data))
a
# There is total 158 rows which consist of NA values
```
```{r}
Ndata <- na.omit(data)
Ndata
#b<- sum(is.na(Ndata))
#b
#  Removed all the Row containing "NA" values
#summary(Ndata)
#Ndata
```
Data Visulization

```{R}
library(tidyverse)
# create a new data frame that will help to build the pie chart
```

```
# for married varieble
pie_df <- data.frame(value = c(5060, 2522),
                     group = c("Married", "Not_Married"))
```

pie chart

```{r}
# install.packages("ggplot2")
library(ggplot2)

# pie chart plot for married variable
pie_df %>%
  group_by(group) %>%
  summarise(sum_values= sum(value)) %>%
  mutate(mean_values=sum_values/sum(sum_values)) %>%
  ggplot(aes(x="", y= mean_values,
             fill=reorder(group, sum_values))) +
  geom_col() + geom_text(aes(label = scales::percent(round(mean_values,2))),
                         position = position_stack(vjust = 0.5))+
  coord_polar(theta = "y") +
  labs(title = "Married and Not Married People") +
  guides(fill = guide_legend(title = "Marrital Status")) +
  theme_void()
```

Histogram

```{r}
## No children based distribution

Cost <- Ndata$cost
```

BMI distribution

BMI <- Ndata$bmi

hist(BMI)

Age distribution

Age <- Ndata$age

hist(Age)

```

##Bar plot

```{R}

## Gender

Gender <- Ndata$gender

ggplot(Ndata, aes(x =Gender)) +

  geom_bar(color = 4,

           fill = 4,

           alpha = 0.25,width = 0.5)

## Smoker

Smoker <- Ndata$smoker

ggplot(Ndata, aes(x =Smoker)) +

  geom_bar(color = 4,

           fill = 4,

```
            alpha = 0.25,width = 0.5)

Location type
Location <- Ndata$location_type

ggplot(Ndata, aes(x =Location)) +
  geom_bar(color = 4,
            fill = 4,
            alpha = 0.25,width = 0.5)

Education level
Education <- Ndata$education_level

ggplot(Ndata, aes(x =Education)) +
  geom_bar(color = 4,
            fill = 4,
            alpha = 0.25,width = 0.5)
No_of_children
Children <- Ndata$children

ggplot(Ndata, aes(x =Children)) +
  geom_bar(color = 4,
            fill = 4,
            alpha = 0.25,width = 0.5)

Yearly Physical

Yearly_visit <- Ndata$yearly_physical
```

```
ggplot(Ndata, aes(x =Yearly_visit)) +
  geom_bar(color = 4,
           fill = 4,
           alpha = 0.25,width = 0.5)

## Marriage status


Marraige_status <- Ndata$married

ggplot(Ndata, aes(x =Marraige_status)) +
  geom_bar(color = 4,
           fill = 4,
           alpha = 0.25,width = 0.5)

## hypertension

Hypertension <- Ndata$hypertension

ggplot(Ndata, aes(x =Hypertension)) +
  geom_bar(color = 4,
           fill = 4,
           alpha = 0.25,width = 0.5)

## Exercise

Exercise <- Ndata$exercise
ggplot(Ndata, aes(x =Exercise)) +
  geom_bar(color = 4,
```

```
        fill = 4,
        alpha = 0.25,width = 0.5)


```
```{R}
## Plot of Cost vs Age
ggplot(Ndata,aes(x=Age, y=cost))+geom_bar(stat="identity")

```
scatter plot
```{R}
# cost and BMI based on smoking habit
smoker_status<- Ndata$smoker

scatt1<-
ggplot(Ndata)+geom_point(aes(x=Ndata$bmi,y=Ndata$cost,color=smoker_status))+
ylab('cost')+xlab('BMI')+ggtitle("cost and BMI based on smoking habit")
scatt1

#cost and BMI based on exercising habit
Exercise_status <- Ndata$exercise

scatt2<-
ggplot(Ndata)+geom_point(aes(x=Ndata$bmi,y=Ndata$cost,color=Exercise_status))+
ylab('cost')+xlab('BMI')+ggtitle("cost and BMI based on smoking habit")
scatt2

#cost and BMI based on location type
```

```
Location_type <- Ndata$location_type

scatt3<-
ggplot(Ndata)+geom_point(aes(x=Ndata$bmi,y=Ndata$cost,color=Location_type))+
ylab('cost')+xlab('BMI')+ggtitle("cost and BMI based on location type")
scatt3

#cost and BMI based on Marriage_status

Marraige_status <- Ndata$married

scatt4<-
ggplot(Ndata)+geom_point(aes(x=Ndata$bmi,y=Ndata$cost,color=Marraige_status))+
ylab('cost')+xlab('BMI')+ggtitle("cost and BMI based on Marriage_status")
scatt4

# cost and BMI based on hypertension

#Hypertension <- Ndata$hypertension

#scatt5<-
ggplot(Ndata)+geom_point(aes(x=Ndata$bmi,y=Ndata$cost,color=Hypertension))+ylab('cost')+xlab('BMI
')+ggtitle("cost and BMI based on hypertension")
#scatt5

# cost and BMI based on Yearly Physical

Yearly_physician_visit <- Ndata$yearly_physical
```

```
scatt6<-
ggplot(Ndata)+geom_point(aes(x=Ndata$bmi,y=Ndata$cost,color=Yearly_physician_visit))+ylab('cost')+x
lab('BMI')+ggtitle(" cost and BMI based on Yearly Physical")

scatt6



# cost and BMI based on gender


gender <- Ndata$gender


scatt6<-

ggplot(Ndata)+geom_point(aes(x=Ndata$bmi,y=Ndata$cost,color=gender))+

ylab('cost')+xlab('BMI')+ggtitle("cost and BMI based on gender")

scatt6
```

Map

```{r}

#install.packages("maps")

#install.packages("ggmap")

install.packages("mapproj")

library(ggplot2)

library(maps)

library(ggmap)

library(mapproj)


```{r}

#install.packages("caret")
```
```

#install.packages("kernlab")

#install.packages('mapproj')

```

```{R}

library(tidyverse)

#source("some_functions.R")

us <- map_data("state")

myMap <- ggplot(us) +

  geom_polygon(color="black", fill="white",

    aes(x=long, y=lat, group=group)) +

  coord_map()

myMap

us %>% filter(region=="pennsylvania")

Ndata$region <- tolower(Ndata$location)

Merge_data<- merge(us,Ndata, by = "region",sort=FALSE)

Merge_data

Merge_data <- Merge_data %>% arrange(order)

ggplot(Merge_data) +

 geom_polygon(color="black",

            aes(x=long,y=lat, group=group,

                fill=cost)) + coord_map()

```

```{R}

us <- map_data("state")

```
us$state_name = tolower(us$region)

Ndata$State <- tolower(Ndata$location)

dfMerged <- merge(Ndata, us, all.y = TRUE, by.x="State", by.y = "region")

dfMerged <- dfMerged %>% arrange(order)

map <- ggplot(dfMerged)

map <- map + aes(x=long, y=lat, group=group,fill=dfMerged$cost) + geom_polygon(color = "black")

map <- map + expand_limits(x=dfMerged$long, y=dfMerged$lat)

map <- map + coord_map() + ggtitle("Cost per state")

map
```

```

#LM Module Training

```{r}
S1 <- sample(Ndata, size = 100,replace = TRUE)

lmOut1<-lm(formula=cost~age+children+bmi+smoker+exercise+hypertension,data=S1)

#lmOut1

summary(lmOut1)
```

#Predict

```{r}
#plot(lmOut1)

Tdata <- read.csv(file.choose(),header = T)

Tdata$Expectedcost <- predict(lmOut1,Tdata)

Tdata$expensive <- ifelse(Tdata$Expectedcost>5000,TRUE, FALSE)

Tdata
```

SVM model

```{r}
#HMO_df_new <- HMO_df_new %>%mutate(expensive = ifelse(cost > 5000,TRUE,FALSE))
```
```

```
#HMO_df_new$expensive <- as.factor(as.logical(HMO_df_new$expensive))

set.seed(111)
# randomly sample for training dataset elements

# install.packages("caret")
#install.packages("kernlab")
library(caret)
library(kernlab)

# randomly sample for training dataset elements
trainlist <- createDataPartition(
  y = Ndata$cost, p=.70, list=FALSE)



# create training and testing datasets
trainData <- Ndata[trainlist,]
#testData <- read_csv(file = 'Test data.csv')
testData <- Ndata[-trainlist,]

svm.model <- train(expensive ~ smoker+bmi+children+hypertension+age+exercise,
            data = Ndata, method = "svmRadial",
            trControl = trainControl(method = "none"),preProcess = c("center", "scale"))
svm.model

predictValues <- predict(svm.model, newdata = Tdata2)
Tdata2$expensive <- predictValues
confusionMatrix(predictValues, Tdata$expensive)
```

table(predictValues)

table(testData$expensive)

str(HMO_df_new)

view(testData)

view(trainData)

```

## Tree Model

```{r}

#install.packages('e1071')

#install.packages("rpart")

#install.packages("rpart.plot")

#install.packages("rio")

library(e1071)

library(rpart)

library(rpart.plot)

library(caret)

library(rio)

Ndata$expensive <- ifelse(Ndata$cost >5000,'TRUE','FALSE')

Ndata$expensive <- as.factor(Ndata$expensive)

Tree<- train(expensive ~
X+age+bmi+children+smoker+location+location_type+education_level+yearly_physical+exercise+marrie
d+hypertension+gender, method = "rpart",data = Ndata)

rpart.plot(Tree$finalModel)

```
```

```{r}
Tdata1 <- read.csv(file.choose(),header = T)
#Tdata1
library(caret)
pred <- predict(Tree,Tdata1)
```

```{r}
Tdata2<- read.csv(file.choose(),header = T)
confusionMatrix(pred,as.factor(Tdata2$expensive))

```

```{r}
Tdata2
```