IST 687 Final project Group 5

2023-04-29

```
#required libaries
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(ggplot2)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ lubridate 1.9.2     ✓ tibble    3.2.1
## ✓ purrr     1.0.1     ✓ tidyr     1.3.0
## ✓ readr     2.1.4
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
library(rsample)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(kernlab)
```

```
##
## Attaching package: 'kernlab'
##
## The following object is masked from 'package:purrr':
##
##     cross
##
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
library(e1071)
```

```
##
## Attaching package: 'e1071'
##
## The following object is masked from 'package:rsample':
##
##     permutations
```

```
library(arules)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
## Attaching package: 'arules'
##
## The following object is masked from 'package:kernlab':
##
##     size
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
library(imputeTS)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(rio)
library(rpart)
library(rpart.plot)

library(caret)
```

```
# Loading the given dataset
MyData <- read_csv("https://intro-datascience.s3.us-east-2.amazonaws.com/HMO_data.csv")
```

```
## Rows: 7582 Columns: 14
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (8): smoker, location, location_type, education_level, yearly_physical, ...
## dbl (6): X, age, bmi, children, hypertension, cost
##
## 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.
```

```
#checking dataset
head(MyData)
```

```
## # A tibble: 6 × 14
##       X   age   bmi children smoker location      location_type education_level
##   <dbl> <dbl> <dbl>    <dbl> <chr>  <chr>         <chr>         <chr>
## 1     1    18  27.9        0 yes    CONNECTICUT   Urban         Bachelor
## 2     2    19  33.8        1 no     RHODE ISLAND  Urban         Bachelor
## 3     3    27  33          3 no     MASSACHUSETTS Urban         Master
## 4     4    34  22.7        0 no     PENNSYLVANIA  Country       Master
## 5     5    32  28.9        0 no     PENNSYLVANIA  Country       PhD
## 6     7    47  33.4        1 no     PENNSYLVANIA  Urban         Bachelor
## # ℹ 6 more variables: yearly_physical <chr>, exercise <chr>, married <chr>,
## #   hypertension <dbl>, gender <chr>, cost <dbl>
```

```
# Viewing the dataframe
#view(data)
str(MyData)
```

```
## spc_tbl_ [7,582 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ X              : num [1:7582] 1 2 3 4 5 7 9 10 11 12 ...
##  $ age            : num [1:7582] 18 19 27 34 32 47 36 59 24 61 ...
##  $ bmi            : num [1:7582] 27.9 33.8 33 22.7 28.9 ...
##  $ children       : num [1:7582] 0 1 3 0 0 1 2 0 0 0 ...
##  $ smoker         : chr [1:7582] "yes" "no" "no" "no" ...
##  $ location       : chr [1:7582] "CONNECTICUT" "RHODE ISLAND" "MASSACHUSETTS" "PENNSYLVANIA"
...
##  $ location_type  : chr [1:7582] "Urban" "Urban" "Urban" "Country" ...
##  $ education_level: chr [1:7582] "Bachelor" "Bachelor" "Master" "Master" ...
##  $ yearly_physical: chr [1:7582] "No" "No" "No" "No" ...
##  $ exercise       : chr [1:7582] "Active" "Not-Active" "Active" "Not-Active" ...
##  $ married        : chr [1:7582] "Married" "Married" "Married" "Married" ...
##  $ hypertension   : num [1:7582] 0 0 0 1 0 0 0 1 0 0 ...
##  $ gender         : chr [1:7582] "female" "male" "male" "male" ...
##  $ cost           : num [1:7582] 1746 602 576 5562 836 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X = col_double(),
##   ..   age = col_double(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   location = col_character(),
##   ..   location_type = col_character(),
##   ..   education_level = col_character(),
##   ..   yearly_physical = col_character(),
##   ..   exercise = col_character(),
##   ..   married = col_character(),
##   ..   hypertension = col_double(),
##   ..   gender = col_character(),
##   ..   cost = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#summary of the dataset
summary(MyData)
```

```
##        X                  age             bmi           children
##  Min.   :        1   Min.   :18.00   Min.   :15.96   Min.   :0.000
##  1st Qu.:     5635   1st Qu.:26.00   1st Qu.:26.60   1st Qu.:0.000
##  Median :    24916   Median :39.00   Median :30.50   Median :1.000
##  Mean   :   712602   Mean   :38.89   Mean   :30.80   Mean   :1.109
##  3rd Qu.:   118486   3rd Qu.:51.00   3rd Qu.:34.77   3rd Qu.:2.000
##  Max.   :131101111   Max.   :66.00   Max.   :53.13   Max.   :5.000
##                                      NA's   :78
##     smoker            location         location_type      education_level
##  Length:7582        Length:7582        Length:7582        Length:7582
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  yearly_physical      exercise           married           hypertension
##  Length:7582        Length:7582        Length:7582        Min.   :0.0000
##  Class :character   Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Mode  :character   Median :0.0000
##                                                           Mean   :0.2005
##                                                           3rd Qu.:0.0000
##                                                           Max.   :1.0000
##                                                           NA's   :80
##     gender               cost
##  Length:7582        Min.   :    2
##  Class :character   1st Qu.:  970
##  Mode  :character   Median : 2500
##                     Mean   : 4043
##                     3rd Qu.: 4775
##                     Max.   :55715
##
```

```
# cleaning the dataframe and Checking for Null values
colSums(is.na(MyData))
```

```
##              X             age             bmi        children          smoker
##              0               0              78               0               0
##        location   location_type education_level yearly_physical        exercise
##              0               0               0               0               0
##         married    hypertension          gender            cost
##              0              80               0               0
```

```
# Removing Null values
MyData$bmi<- na_interpolation(MyData$bmi)
MyData$hypertension <- na_interpolation(MyData$hypertension)
```

```
#Checking third quantile of cost to set threshold for cost as expensive or inexpensive variable
quantile(MyData$cost, probs = c(0.75))
```

```
##  75%
## 4775
```

```
#creating expensive column
MyData$expensive <- MyData$cost>4775
head(MyData)
```

```
## # A tibble: 6 × 15
##       X   age   bmi children smoker location      location_type education_level
##   <dbl> <dbl> <dbl>    <dbl> <chr>  <chr>         <chr>         <chr>
## 1     1    18  27.9        0 yes    CONNECTICUT   Urban         Bachelor
## 2     2    19  33.8        1 no     RHODE ISLAND  Urban         Bachelor
## 3     3    27  33          3 no     MASSACHUSETTS Urban         Master
## 4     4    34  22.7        0 no     PENNSYLVANIA  Country       Master
## 5     5    32  28.9        0 no     PENNSYLVANIA  Country       PhD
## 6     7    47  33.4        1 no     PENNSYLVANIA  Urban         Bachelor
## # ℹ 7 more variables: yearly_physical <chr>, exercise <chr>, married <chr>,
## #   hypertension <dbl>, gender <chr>, cost <dbl>, expensive <lgl>
```

```
#replacing all true values with 1 and false with 0's
MyData <- MyData %>%mutate(expensive=str_replace_all(string=expensive,pattern="TRUE","1"))

MyData <- MyData %>%mutate(expensive=str_replace_all(string=expensive,pattern="FALSE","0"))

head(MyData)
```

```
## # A tibble: 6 × 15
##       X   age   bmi children smoker location      location_type education_level
##   <dbl> <dbl> <dbl>    <dbl> <chr>  <chr>         <chr>         <chr>
## 1     1    18  27.9        0 yes    CONNECTICUT   Urban         Bachelor
## 2     2    19  33.8        1 no     RHODE ISLAND  Urban         Bachelor
## 3     3    27  33          3 no     MASSACHUSETTS Urban         Master
## 4     4    34  22.7        0 no     PENNSYLVANIA  Country       Master
## 5     5    32  28.9        0 no     PENNSYLVANIA  Country       PhD
## 6     7    47  33.4        1 no     PENNSYLVANIA  Urban         Bachelor
## # ℹ 7 more variables: yearly_physical <chr>, exercise <chr>, married <chr>,
## #   hypertension <dbl>, gender <chr>, cost <dbl>, expensive <chr>
```

```
 #dividing expensive and inexpensive people into 2 categories
expensivePeople <- subset(MyData,expensive=="1")
inexpensivePeople <- subset(MyData,expensive=="0")
head(expensivePeople)
```

```
## # A tibble: 6 × 15
##       X   age   bmi children smoker location     location_type education_level
##   <dbl> <dbl> <dbl>    <dbl> <chr>  <chr>        <chr>         <chr>
## 1     4    34  22.7        0 no     PENNSYLVANIA Country       Master
## 2    10    59  25.8        0 no     PENNSYLVANIA Country       Bachelor
## 3    15    26  42.1        0 yes    PENNSYLVANIA Urban         Bachelor
## 4    20    31  35.3        0 yes    PENNSYLVANIA Urban         PhD
## 5    24    32  31.9        1 yes    NEW JERSEY   Urban         No College Degree
## 6    30    31  36.3        2 yes    PENNSYLVANIA Urban         Bachelor
## # ℹ 7 more variables: yearly_physical <chr>, exercise <chr>, married <chr>,
## #   hypertension <dbl>, gender <chr>, cost <dbl>, expensive <chr>
```

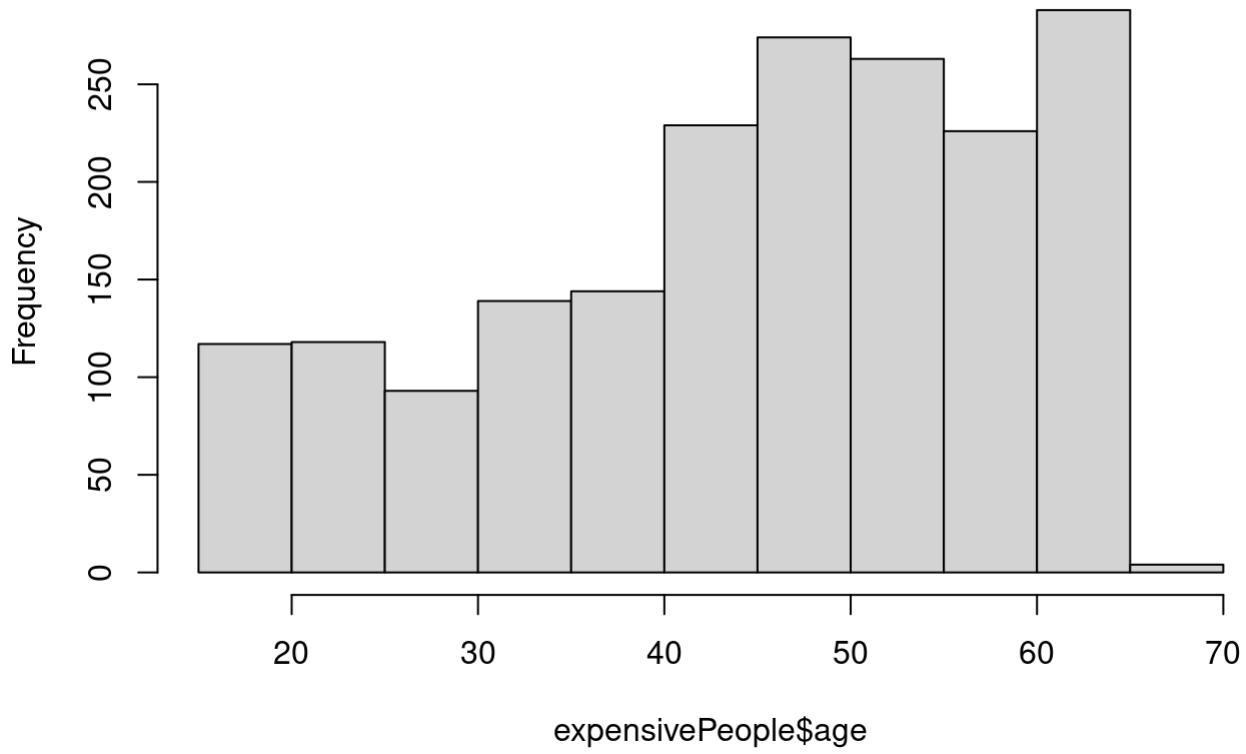
```
head(inexpensivePeople)
```

```
## # A tibble: 6 × 15
##       X   age   bmi children smoker location      location_type education_level
##   <dbl> <dbl> <dbl>    <dbl> <chr>  <chr>         <chr>         <chr>
## 1     1    18  27.9        0 yes    CONNECTICUT   Urban         Bachelor
## 2     2    19  33.8        1 no     RHODE ISLAND  Urban         Bachelor
## 3     3    27  33          3 no     MASSACHUSETTS Urban         Master
## 4     5    32  28.9        0 no     PENNSYLVANIA  Country       PhD
## 5     7    47  33.4        1 no     PENNSYLVANIA  Urban         Bachelor
## 6     9    36  29.8        2 no     PENNSYLVANIA  Urban         Bachelor
## # ℹ 7 more variables: yearly_physical <chr>, exercise <chr>, married <chr>,
## #   hypertension <dbl>, gender <chr>, cost <dbl>, expensive <chr>
```

```
smokerPeople <- subset(MyData,smoker=="yes")
head(smokerPeople)
```

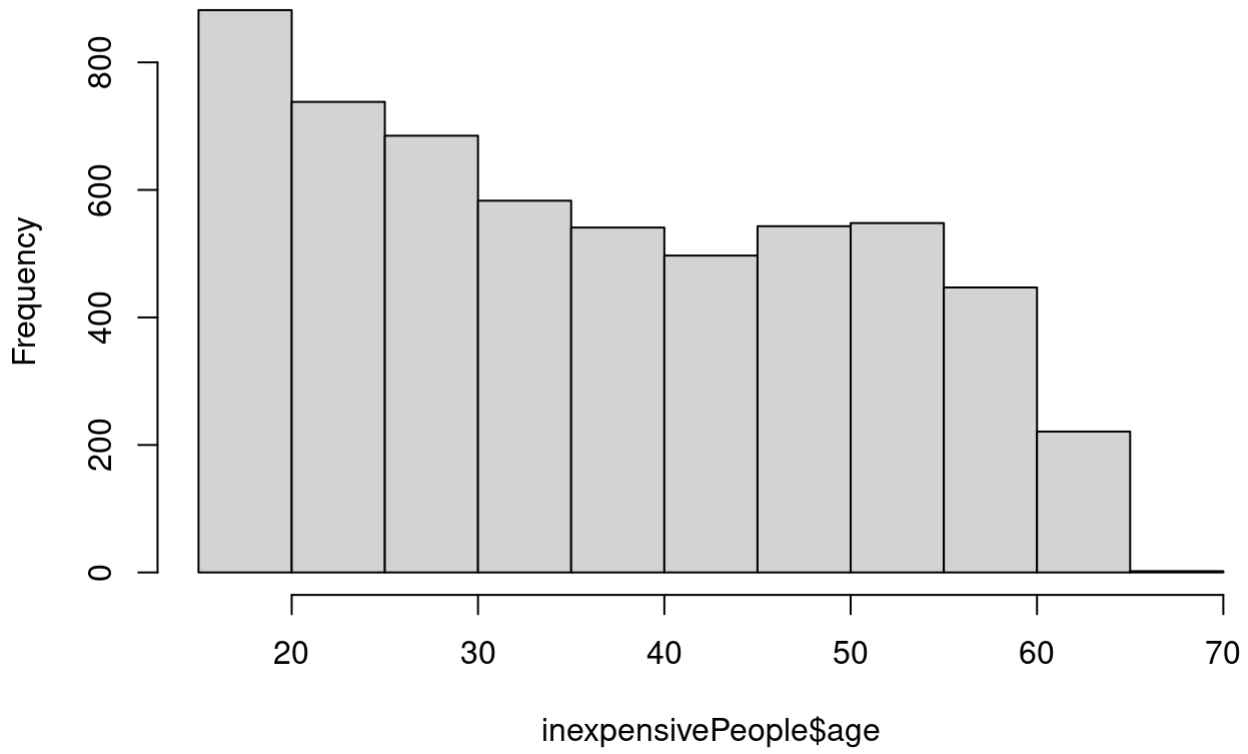
```
## # A tibble: 6 × 15
##       X   age   bmi children smoker location     location_type education_level
##   <dbl> <dbl> <dbl>    <dbl> <chr>  <chr>        <chr>         <chr>
## 1     1    18  27.9        0 yes    CONNECTICUT  Urban         Bachelor
## 2    12    61  26.3        0 yes    CONNECTICUT  Urban         No College Degree
## 3    15    26  42.1        0 yes    PENNSYLVANIA Urban         Bachelor
## 4    20    31  35.3        0 yes    PENNSYLVANIA Urban         PhD
## 5    24    32  31.9        1 yes    NEW JERSEY   Urban         No College Degree
## 6    30    31  36.3        2 yes    PENNSYLVANIA Urban         Bachelor
## # ℹ 7 more variables: yearly_physical <chr>, exercise <chr>, married <chr>,
## #   hypertension <dbl>, gender <chr>, cost <dbl>, expensive <chr>
```

```
# exploratory analysis
# creating histogram for people's age under expensive category
hist(expensivePeople$age)
```



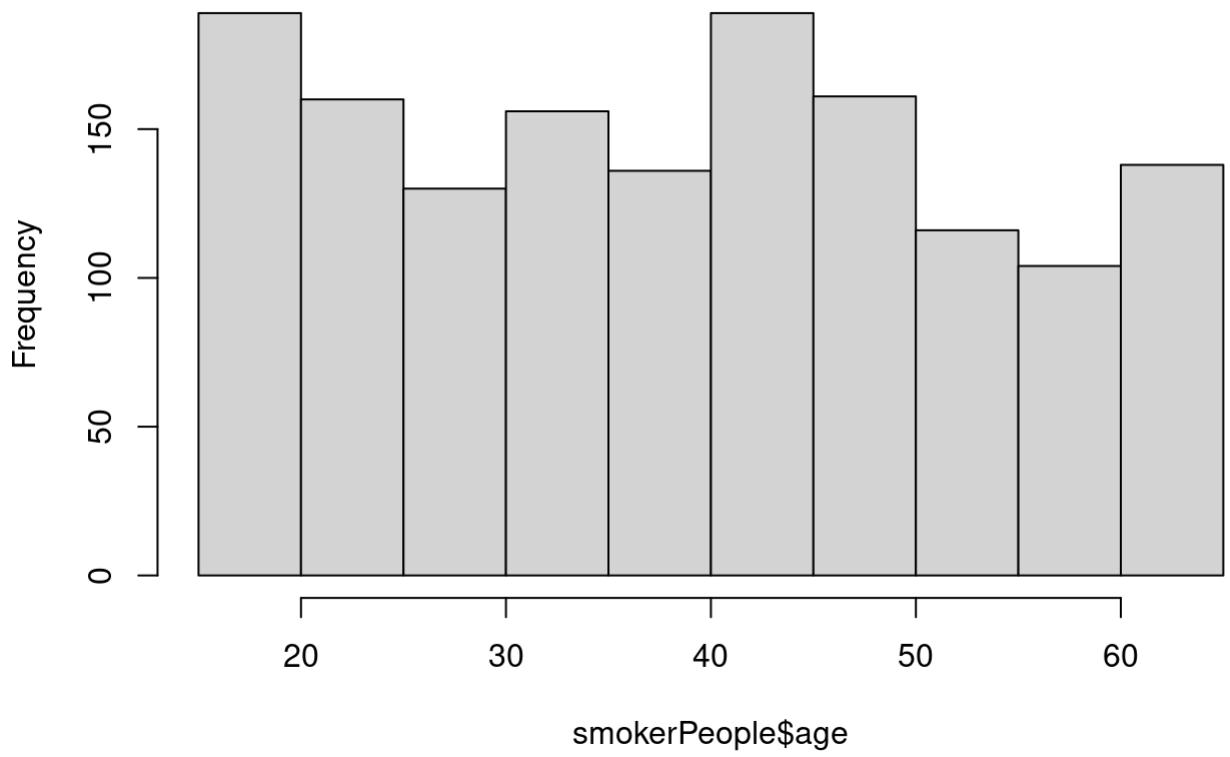

```
#people with age between 40 to 65 most likely to pay more for their healthcare cost
```

```
#creating histogram for analyzing age group of inexpensive healthcare group
hist(inexpensivePeople$age)
```



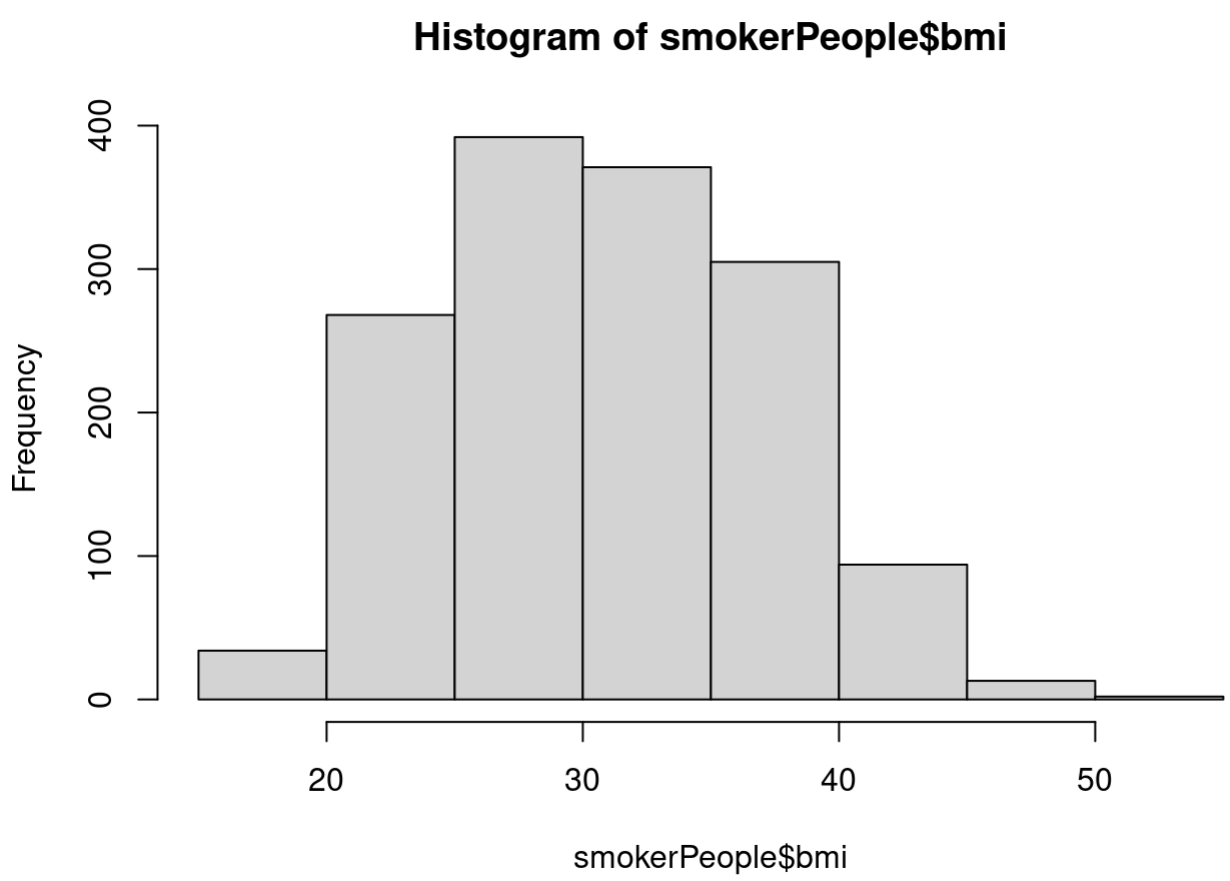

```
#younger group starting from age 18 to 40 comes under inexpensive healthcare cost group
```

```
#creating histogram for analyzing age group of people who smoke
hist(smokerPeople$age)
```



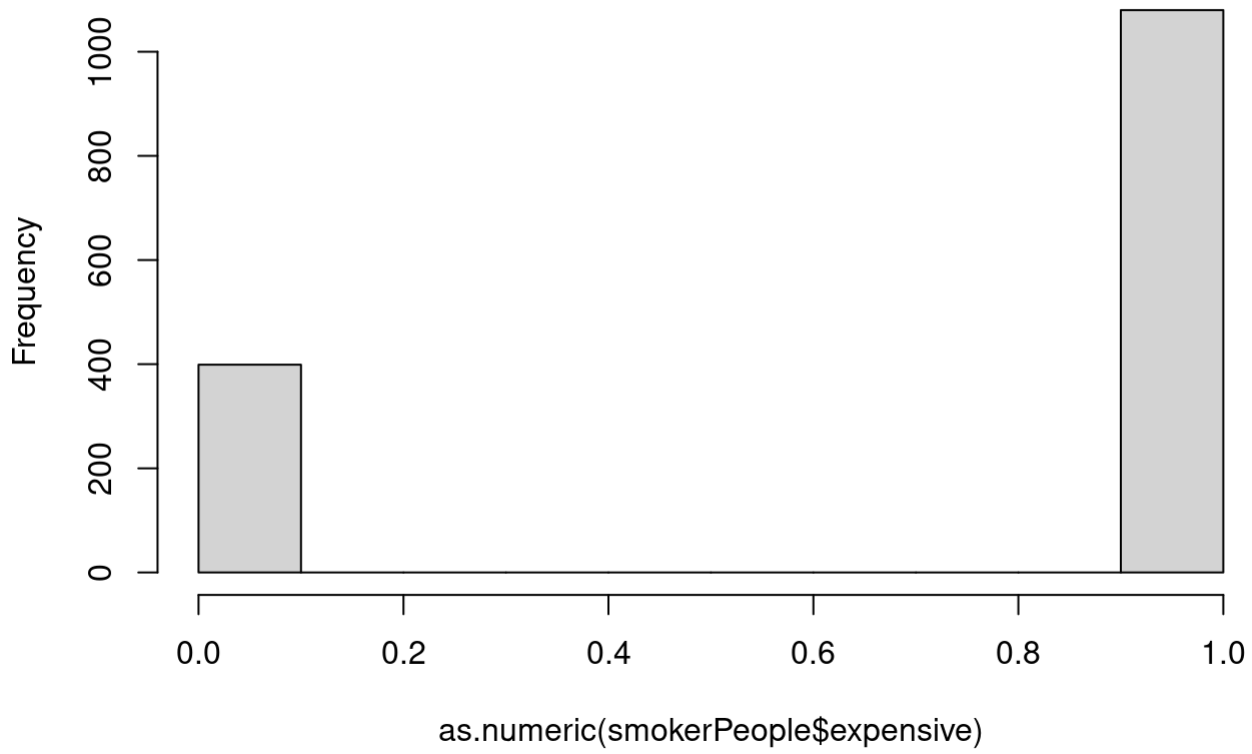

```
#peole of age between 18 to 25 and 40 to 45 tends to smoke more than other age groups
```

```
#creating histogram to analyze bmi for people who fall under smoking category
hist(smokerPeople$bmi)
```



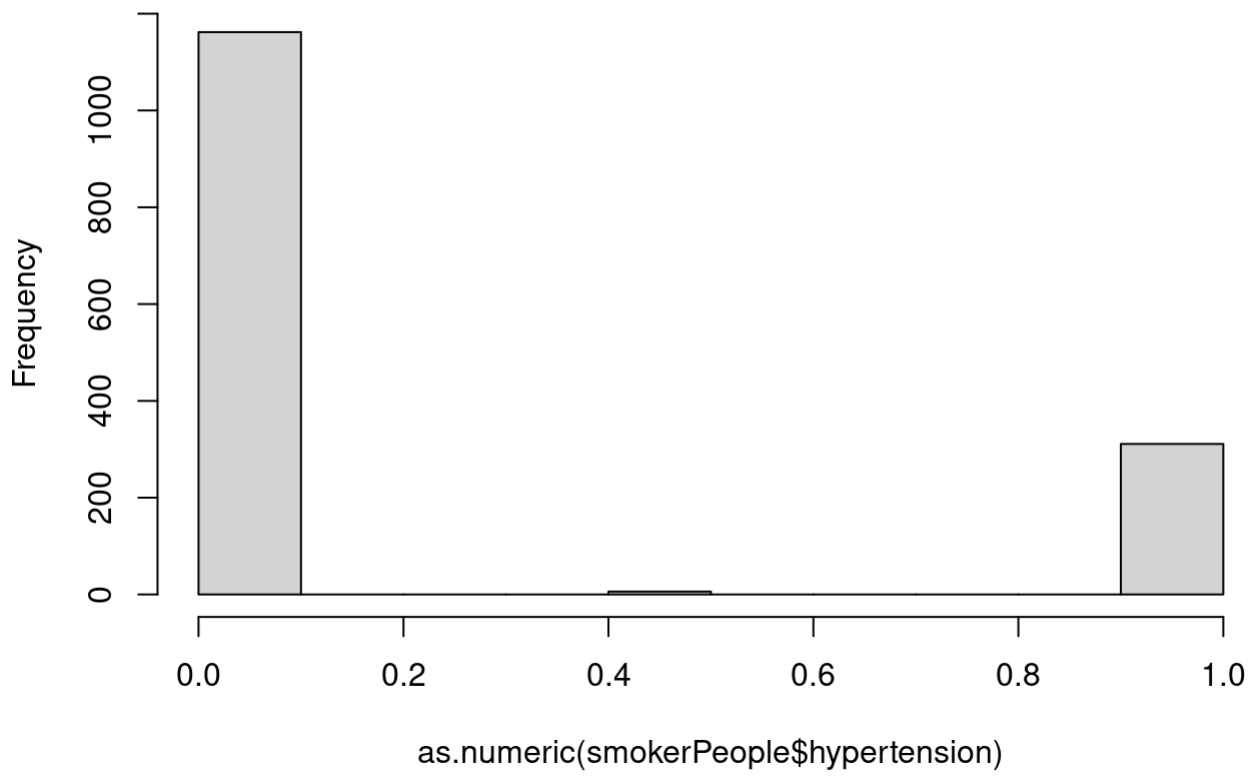

```
#majority people who smoke do not fall under healthy bmi range of 18 to 25
```

```
#creating histogram to analyze whether smokers pay more for their healthcare
hist(as.numeric(smokerPeople$expensive))
```



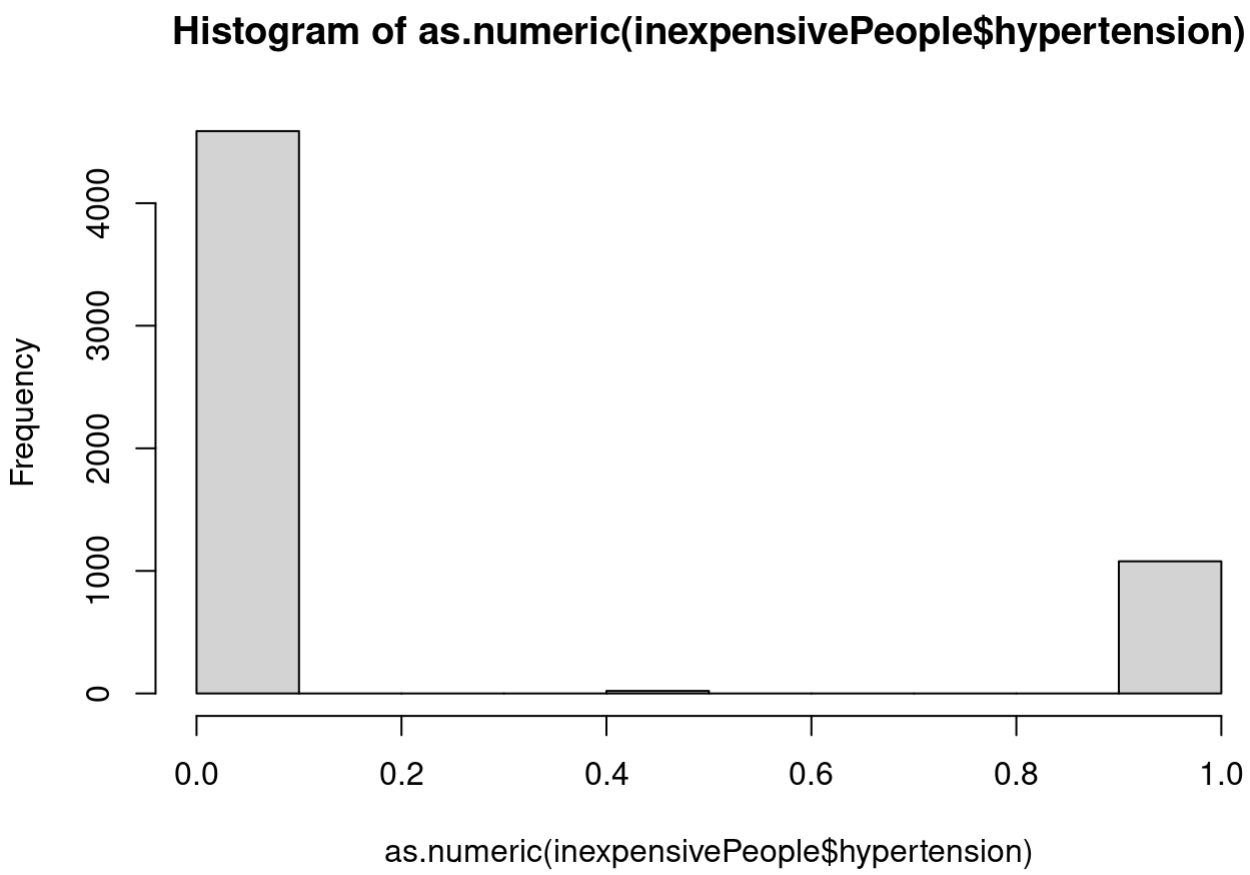

```
# majority of people who smoke tends to pay more
```

```
#creating histogram to analyze relation between smokers and hypertension
hist(as.numeric(smokerPeople$hypertension))
```




```
# majority of people who smoke are less likely to have hypertension. hence we can say that hyper
tension does not directly effect on healthcare cost for people who smoke
```

```
#creating histogram to analyze relation between people who pay less for health care cost and hav
e hypertension
hist(as.numeric(inexpensivePeople$hypertension))
```




```
# majority of people who pays less for healthcare are less likely to have hypertension. hence we
can say that peopel who dont have hypertension tend to pay less for there healthcare cost
```

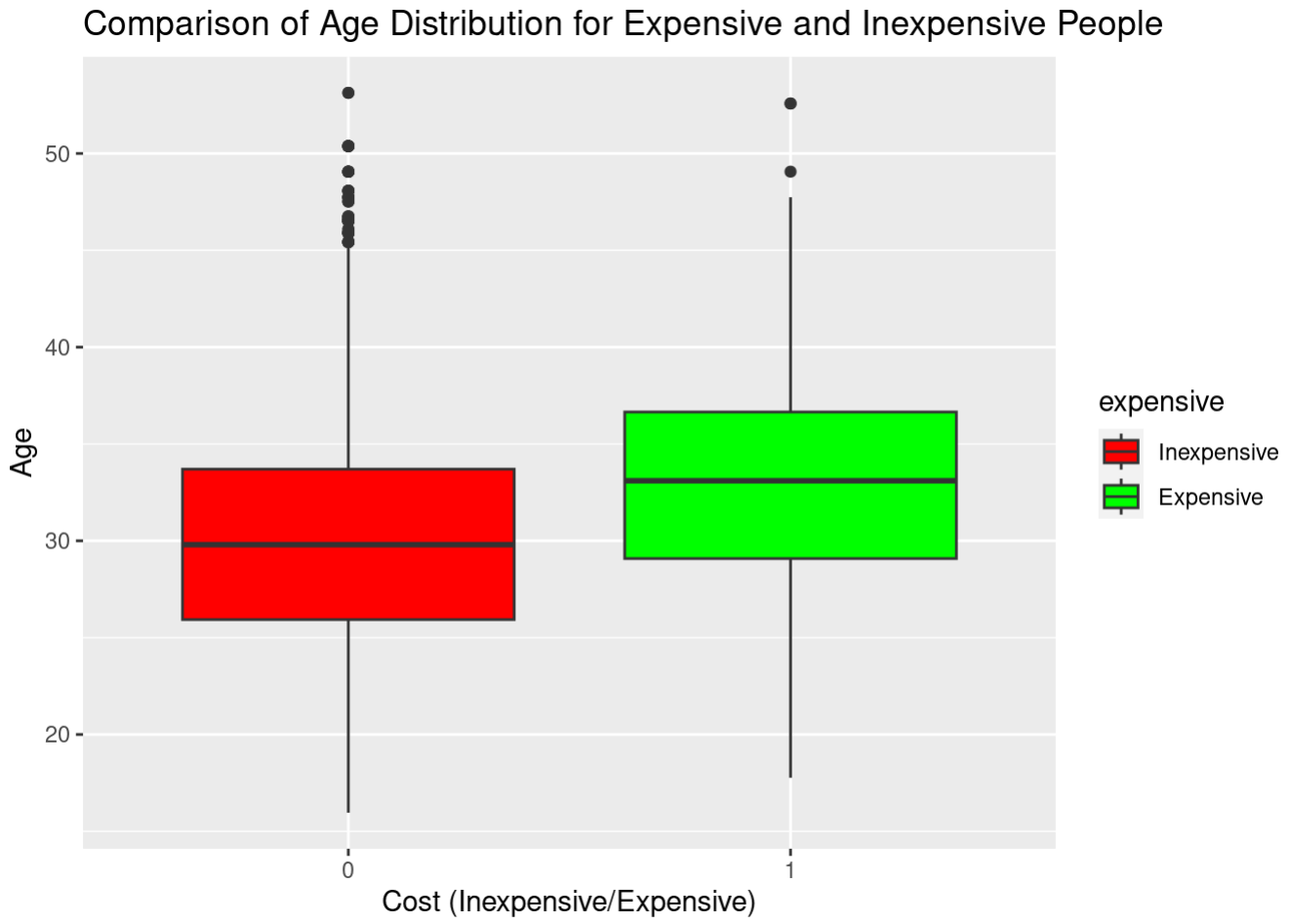
```
library(ggplot2)

# Create a subset of MyData for expensive and inexpensive people
expensivePeople <- subset(MyData, expensive == "1")
inexpensivePeople <- subset(MyData, expensive == "0")

# Create a box plot to compare age distribution for expensive and inexpensive people
ggplot(data = rbind(expensivePeople, inexpensivePeople), aes(x = expensive, y = age, fill = expe
nsive)) +
  geom_boxplot() +
  labs(title = "Comparison of Age Distribution for Expensive and Inexpensive People", x = "Cost
(Inexpensive/Expensive)", y = "Age") +
  scale_fill_manual(values = c("red", "green"), labels = c("Inexpensive", "Expensive"))
```

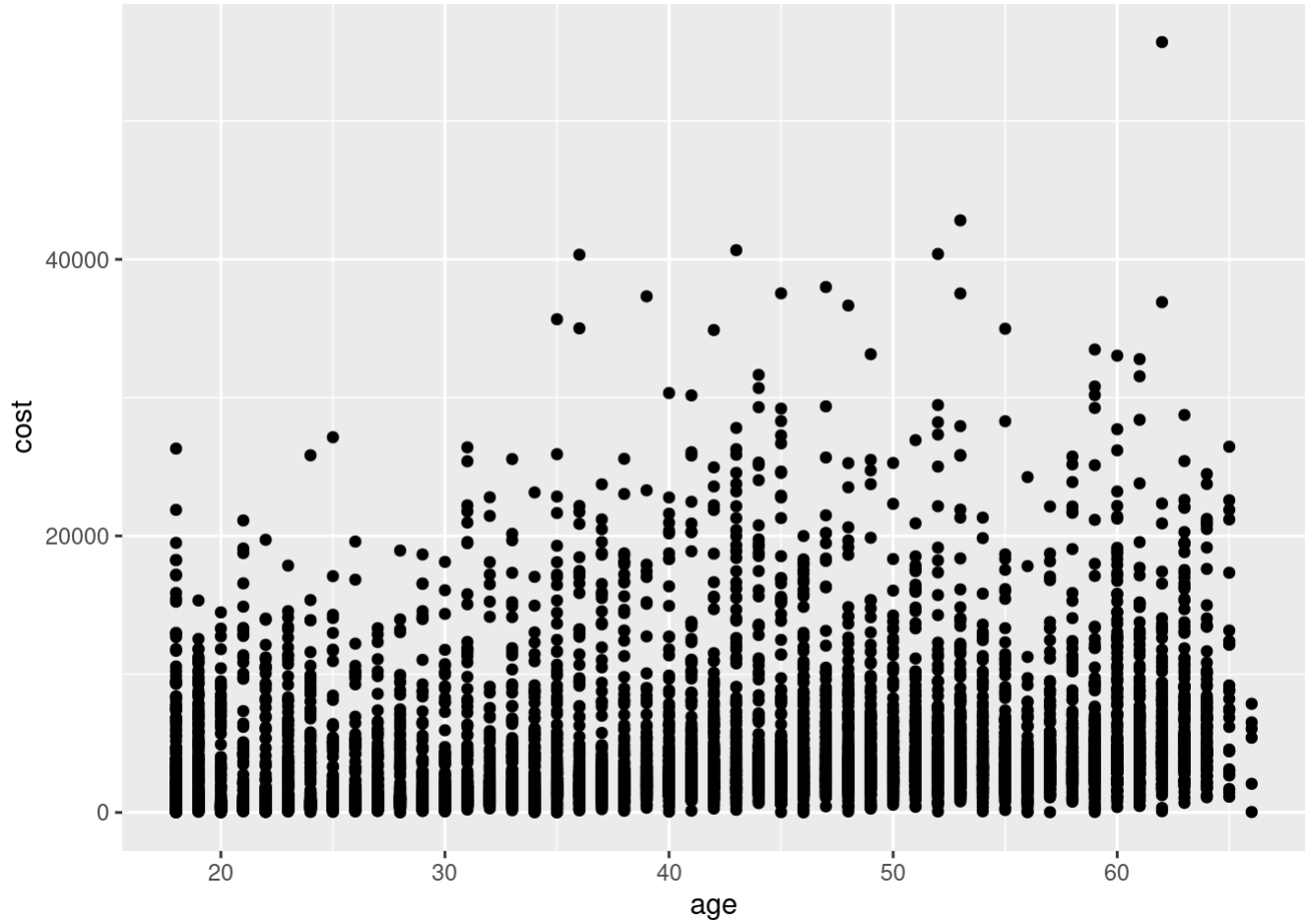
#We used a box plot to compare the age distribution of people categorized as "expensive" and "in expensive" in terms of their healthcare cost. The results showed that the cost by age is general ly higher for those in the "expensive" category compared to those in the "inexpensive" category. This is indicated by the median age of approximately 47 for the "expensive" group, which is high er than the median age of 35 for the "inexpensive" group.

```
ggplot(data = rbind(expensivePeople, inexpensivePeople), aes(x = expensive, y = bmi, fill = expe
nsive)) +
  geom_boxplot() +
  labs(title = "Comparison of Age Distribution for Expensive and Inexpensive People", x = "Cost
(Inexpensive/Expensive)", y = "Age") +
  scale_fill_manual(values = c("red", "green"), labels = c("Inexpensive", "Expensive"))
```



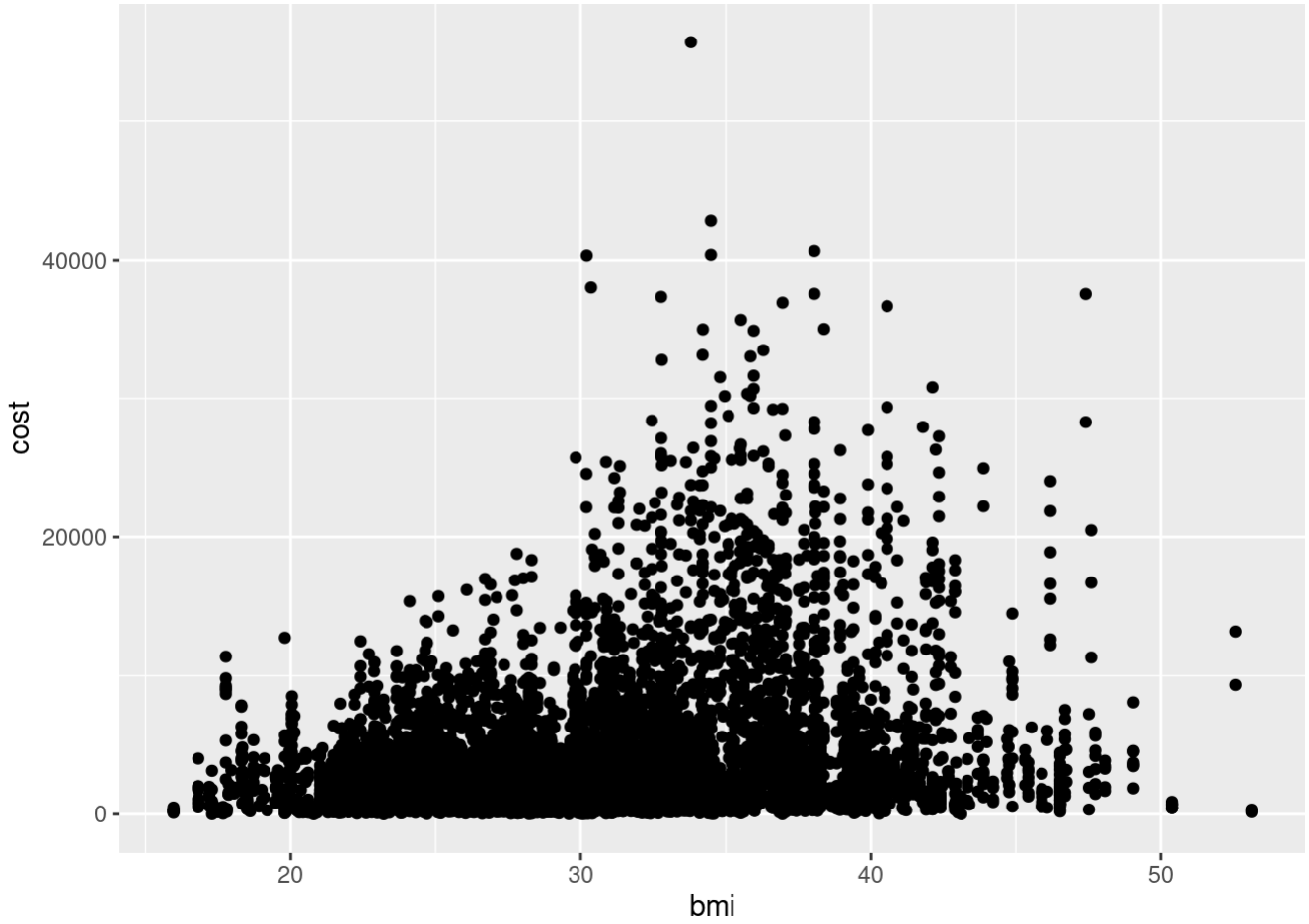

```
#We also generated a box plot to compare the BMI distribution between "expensive" and "inexpensive" groups. The results revealed that the median BMI for "expensive" group is greater than that of "inexpensive" group, indicating that the healthcare cost tends to be higher for people with higher BMI.
```

```
#Scatterplot for Age Vs Cost
AgeCost <- ggplot(MyData,aes(x=age, y=cost)) + geom_point()
AgeCost
```



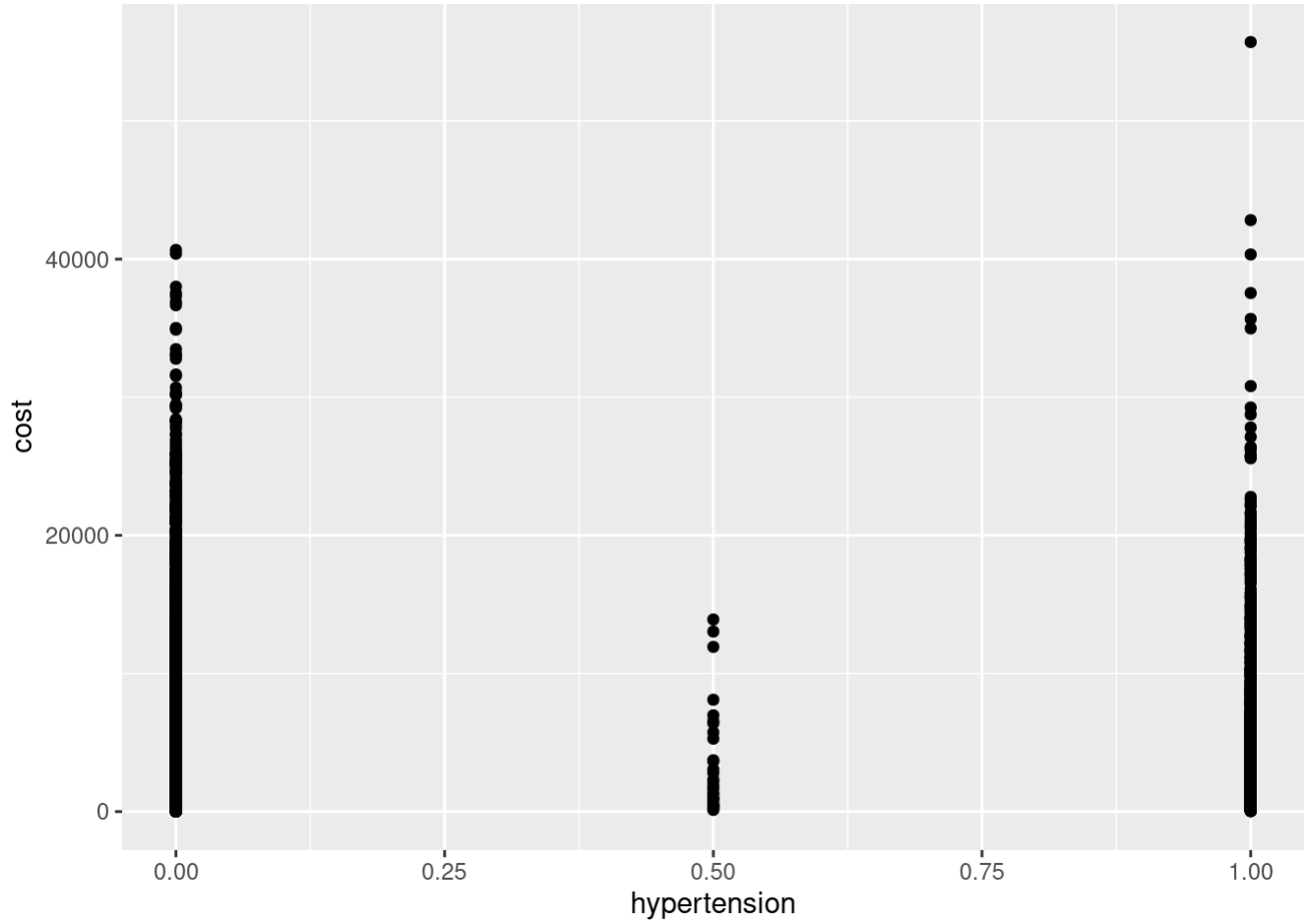

```
#cost vs age : age has a positive correlation to increasing healthcare costs.
```

```
#Scatterplot for Bmi Vs Cost
BmiCost <- ggplot(MyData,aes(x=bmi, y=cost)) + geom_point()
BmiCost
```



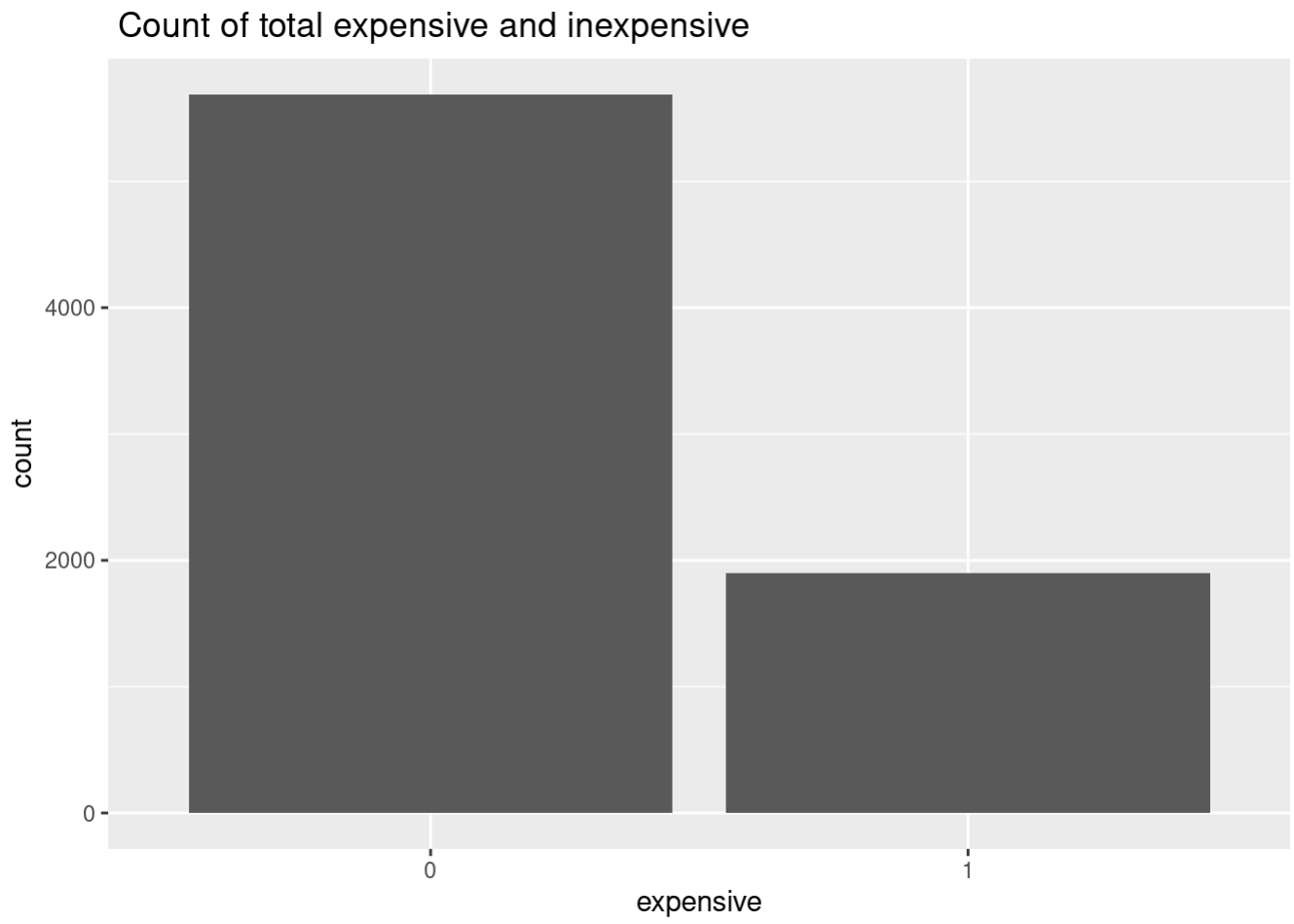

```
#cost vs BMI: BMI has a positive correlation with cost , Peopel with BMI in the range 30-40 tend
s to pay higher health care cost
```

```
#Scatterplot for Hypertension Vs Cost
HyperTensionCost <- ggplot(MyData,aes(x=hypertension, y=cost)) + geom_point()
HyperTensionCost
```



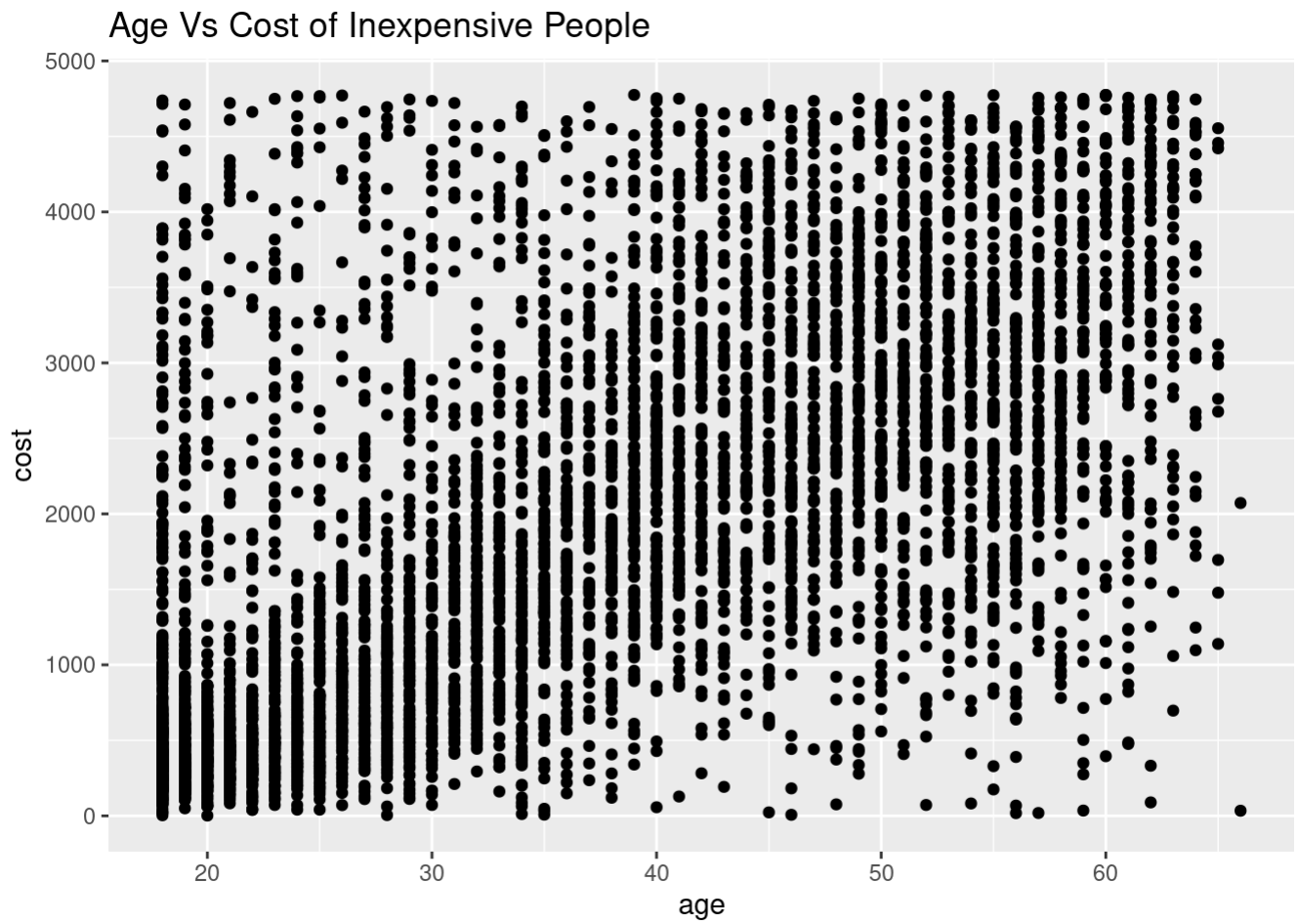

```
#from the above visualization we can say that hypertension is not one of the significant factor to determine healthcare cost
```

```
#barplot for expensive count
ExpensivePlot <- ggplot(MyData,aes(x=expensive)) + geom_bar() + ggtitle(" Count of total expensive and inexpensive")
ExpensivePlot
```



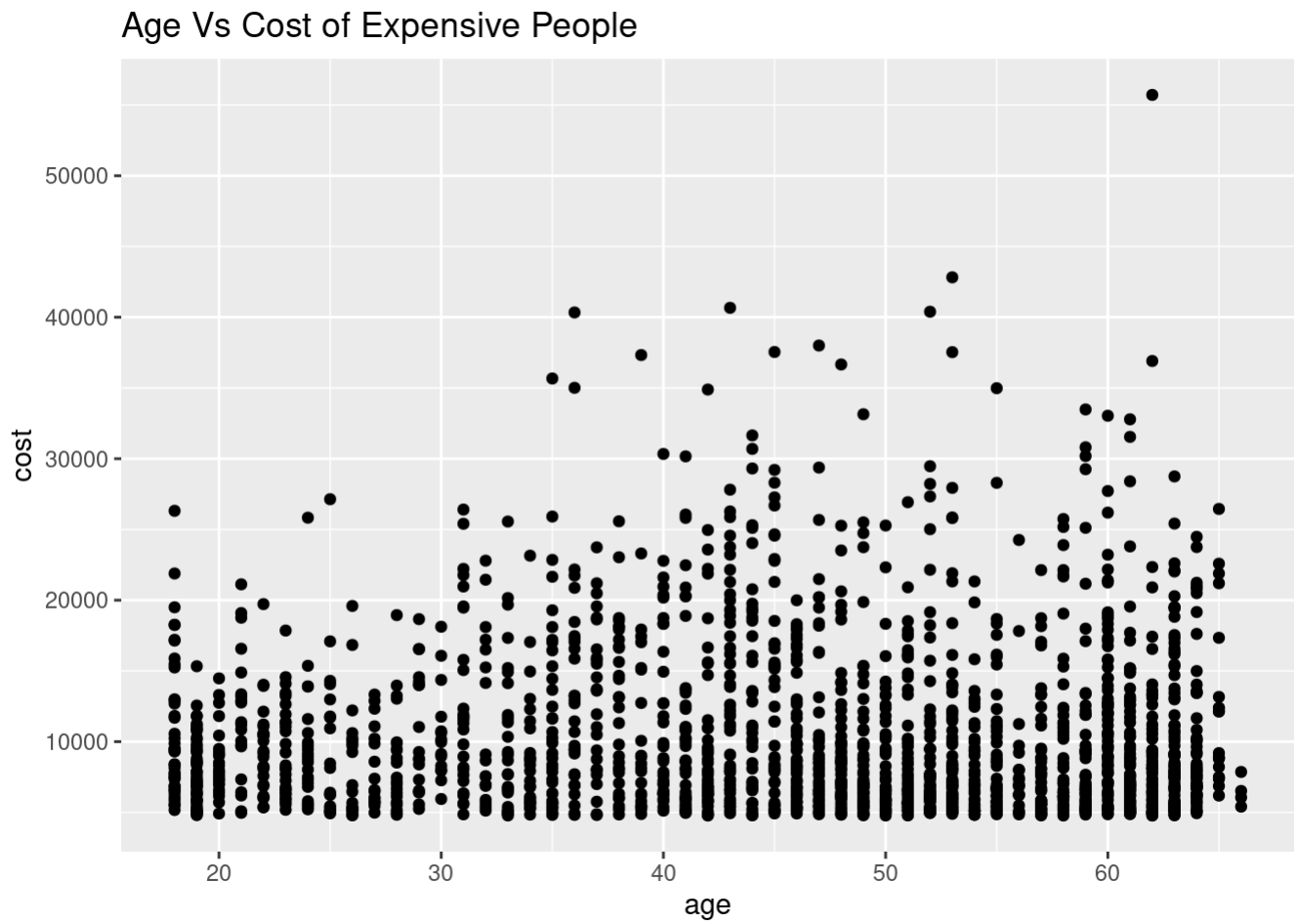

```
#Majority of people from sample data falls under the inexpensive category
```

```
#Bar plot for Age vs Expensive and Inexpensive people
ggplot(inexpensivePeople, aes(x=age,y=cost))+geom_point() + ggtitle("Age Vs Cost of Inexpensive People")
```




```
#there is a positive correlation between age and cost refering to the dense area of the graph
```

```
ggplot(expensivePeople, aes(x=age,y=cost))+geom_point() + ggtitle("Age Vs Cost of Expensive Peop
le")
```




```
#creating a new data frame
HMOData <- data.frame(age = MyData$age,
bmi = MyData$bmi,
                      smoker= MyData$smoker,
                      yearly_physical= MyData$yearly_physical,
                      children = MyData$children,
                      exercise =MyData$exercise,
                      hypertension = MyData$hypertension,
                      expensive=as.factor(MyData$expensive))
# replacing TRUE with 1 and FALSE with 0
HMOData <- HMOData %>% mutate( expensive = str_replace_all( string = expensive, pattern = "TRU
E","1"))
HMOData <- HMOData %>% mutate( expensive = str_replace_all( string = expensive, pattern = "FALS
E","0"))
HMOData$expensive <- as.factor(HMOData$expensive)
str(HMOData)
```

```
## 'data.frame':    7582 obs. of  8 variables:
##  $ age            : num  18 19 27 34 32 47 36 59 24 61 ...
##  $ bmi            : num  27.9 33.8 33 22.7 28.9 ...
##  $ smoker         : chr  "yes" "no" "no" "no" ...
##  $ yearly_physical: chr  "No" "No" "No" "No" ...
##  $ children       : num  0 1 3 0 0 1 2 0 0 0 ...
##  $ exercise       : chr  "Active" "Not-Active" "Active" "Not-Active" ...
##  $ hypertension   : num  0 0 0 1 0 0 0 1 0 0 ...
##  $ expensive      : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 2 1 1 ...
```

```
library(caret)
# Spliting data into training and testing sets for svm
trainListS <- createDataPartition(y=HMOData$expensive,p=0.80,list=FALSE)
trainSetS <- HMOData[trainListS,]
testSetS <- HMOData[-trainListS,]
dim(trainSetS)
```

```
## [1] 6066    8
```

```
summary(trainSetS)
```

```
##       age             bmi           smoker          yearly_physical
##  Min.   :18.00   Min.   :15.96   Length:6066        Length:6066
##  1st Qu.:26.00   1st Qu.:26.60   Class :character   Class :character
##  Median :39.00   Median :30.50   Mode  :character   Mode  :character
##  Mean   :39.04   Mean   :30.77
##  3rd Qu.:51.00   3rd Qu.:34.60
##  Max.   :66.00   Max.   :53.13
##     children       exercise          hypertension   expensive
##  Min.   :0.000   Length:6066        Min.   :0.000   0:4550
##  1st Qu.:0.000   Class :character   1st Qu.:0.000   1:1516
##  Median :1.000   Mode  :character   Median :0.000
##  Mean   :1.105                      Mean   :0.201
##  3rd Qu.:2.000                      3rd Qu.:0.000
##  Max.   :5.000                      Max.   :1.000
```

```
# Building SVM model
set.seed(123)
library(e1071)
ksvm_model <- svm(data= trainSetS, expensive~.,C=5, CV=3, prob.model= TRUE)
svmPred<- predict(ksvm_model,newdata= testSetS, type= "response")
head(svmPred)
```

```
##  2  7  8 17 26 31
##  0  0  0  1  1  0
## Levels: 0 1
```

```
# Checking accuracy of svm model using confusion matrix
confusionMatrix(svmPred,as.factor(testSetS$expensive))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1116  173
##          1   21  206
##
##                Accuracy : 0.872
##                  95% CI : (0.8542, 0.8884)
##     No Information Rate : 0.75
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6061
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9815
##             Specificity : 0.5435
##          Pos Pred Value : 0.8658
##          Neg Pred Value : 0.9075
##              Prevalence : 0.7500
##          Detection Rate : 0.7361
##    Detection Prevalence : 0.8503
##       Balanced Accuracy : 0.7625
##
##        'Positive' Class : 0
##
```

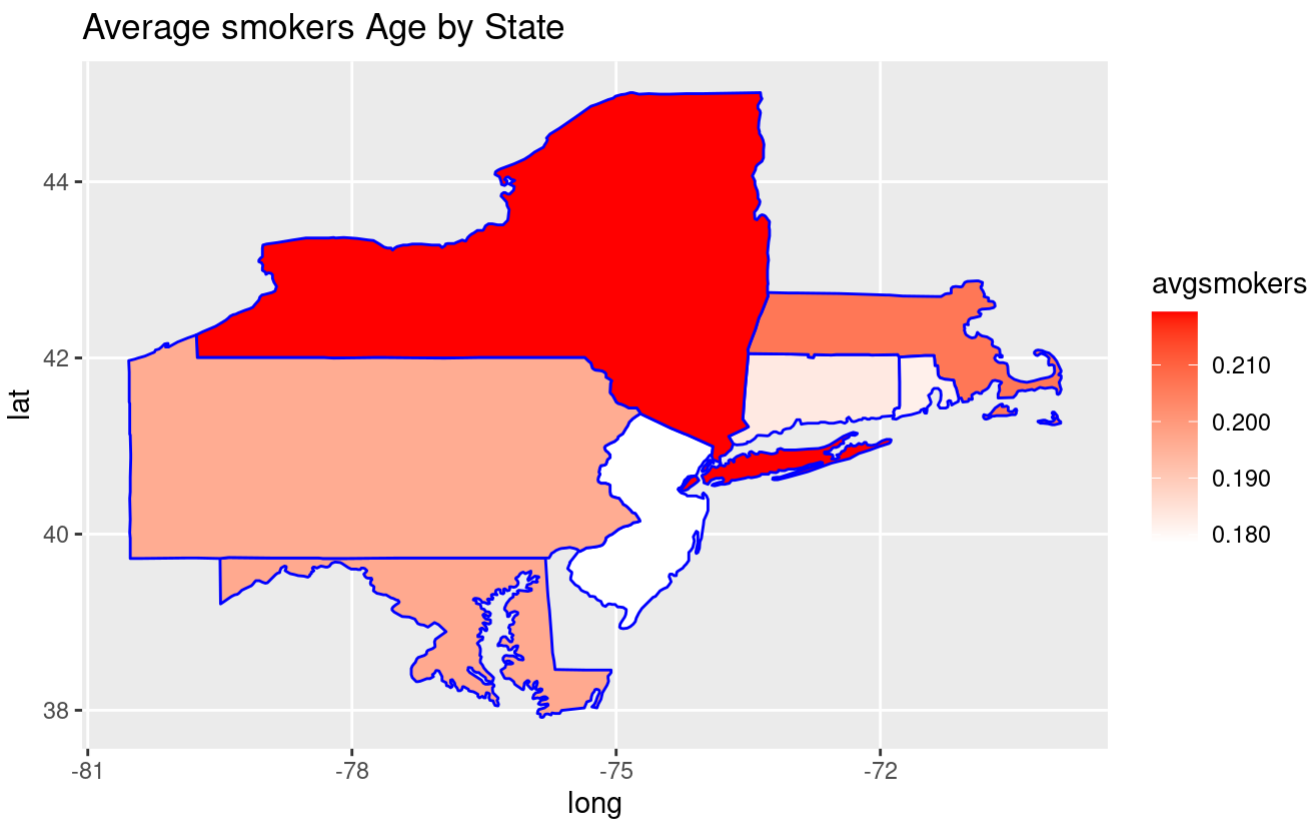
```
# Building a tree model
rpart_model <- rpart(expensive ~ age+bmi+children+smoker+hypertension+exercise+yearly_physical,
data = trainSetS, method = "class")
rpartPred <- predict(rpart_model, newdata= testSetS, type= "class")
confusionMatrix(rpartPred, as.factor(testSetS$expensive))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1114  162
##          1   23  217
##
##                Accuracy : 0.878
##                  95% CI : (0.8604, 0.894)
##     No Information Rate : 0.75
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6293
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9798
##             Specificity : 0.5726
##          Pos Pred Value : 0.8730
##          Neg Pred Value : 0.9042
##              Prevalence : 0.7500
##          Detection Rate : 0.7348
##    Detection Prevalence : 0.8417
##       Balanced Accuracy : 0.7762
##
##        'Positive' Class : 0
##
```

```
# Linear model
trainSetS$expensive<-as.numeric(trainSetS$expensive)
testSetS$expensive<-as.numeric(testSetS$expensive)
lmOut <- lm(expensive~age+bmi+children+smoker+hypertension+exercise+yearly_physical,data=trainSe
tS)
summary(lmOut)
```

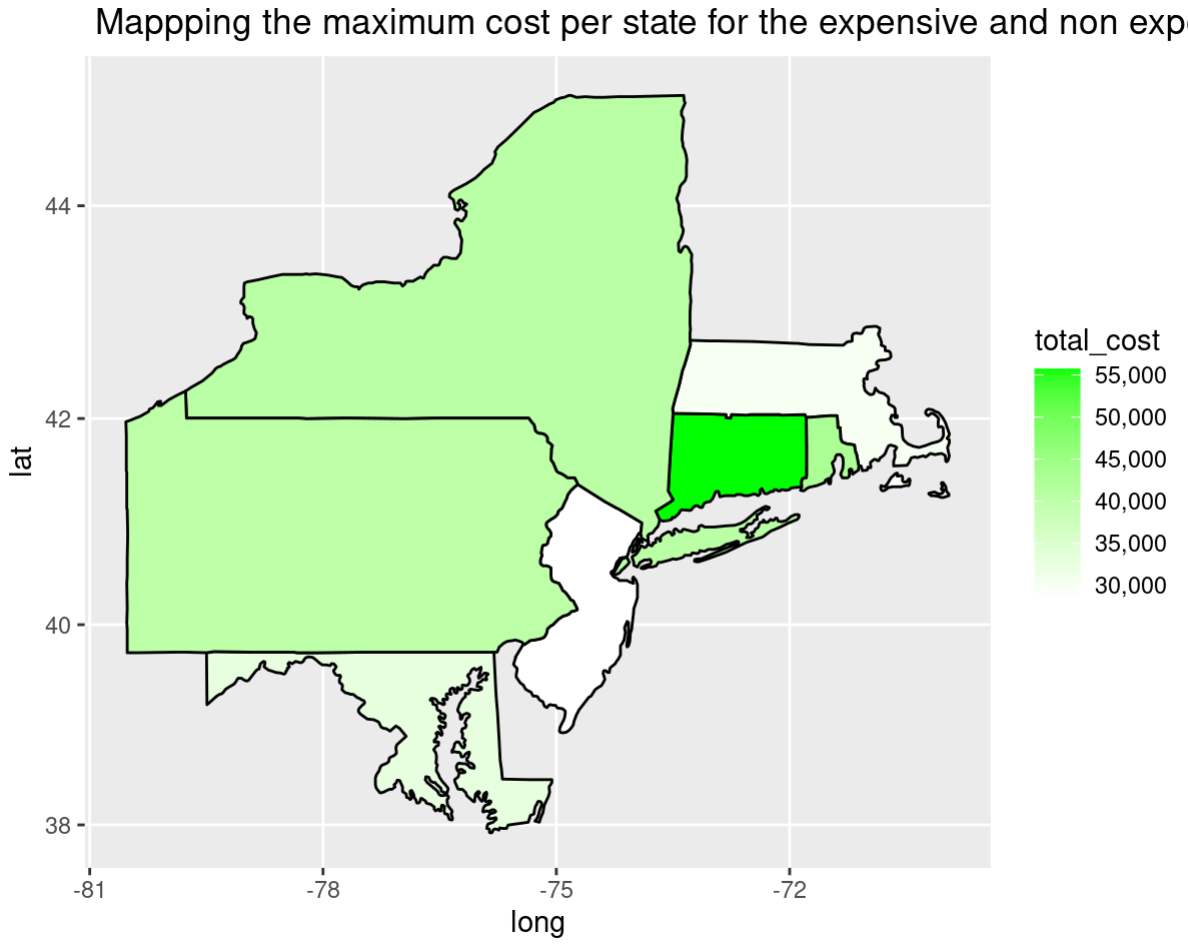
```
##
## Call:
## lm(formula = expensive ~ age + bmi + children + smoker + hypertension +
##     exercise + yearly_physical, data = trainSetS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95783 -0.20712 -0.05892  0.12967  1.14956
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         0.2968854  0.0258540  11.483  < 2e-16 ***
## age                 0.0074299  0.0003013  24.658  < 2e-16 ***
## bmi                 0.0127934  0.0007083  18.062  < 2e-16 ***
## children            0.0105171  0.0034877   3.015  0.00258 **
## smokeryes           0.5936228  0.0106632  55.670  < 2e-16 ***
## hypertension        0.0461545  0.0105684   4.367 1.28e-05 ***
## exerciseNot-Active 0.1667487  0.0097815  17.047  < 2e-16 ***
## yearly_physicalYes 0.0284944  0.0097478   2.923  0.00348 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3289 on 6058 degrees of freedom
## Multiple R-squared:  0.4238, Adjusted R-squared:  0.4231
## F-statistic: 636.5 on 7 and 6058 DF,  p-value: < 2.2e-16
```

```
#Maps(Avg age based on location)
MyData <- MyData %>% mutate( smoker = str_replace_all( string = smoker, pattern = "yes", "1"))
MyData <- MyData %>% mutate( smoker = str_replace_all( string = smoker, pattern = "no", "0"))
MyData$smoker <- as.numeric(MyData$smoker)
dfAgg <- MyData %>% group_by(location) %>% summarise(avgsmokers = mean(smoker))
dfAgg$state <- tolower(dfAgg$location)
us <- map_data("state")
us$state <- us$region
mergedNew <- merge(dfAgg,us,on = "state")
mergedNew <- mergedNew[order(mergedNew$order),]
map <- ggplot(mergedNew) + geom_polygon(aes(x = long, y = lat, group = group,fill = avgsmokers),
color= "Blue" )
map + scale_fill_continuous(low = "white", high = "red", name = "avgsmokers", label = scales::co
mma) +
  coord_map("albers", lat0 = 110, lat1 = 110) +
  labs(title = "Average smokers Age by State")
```



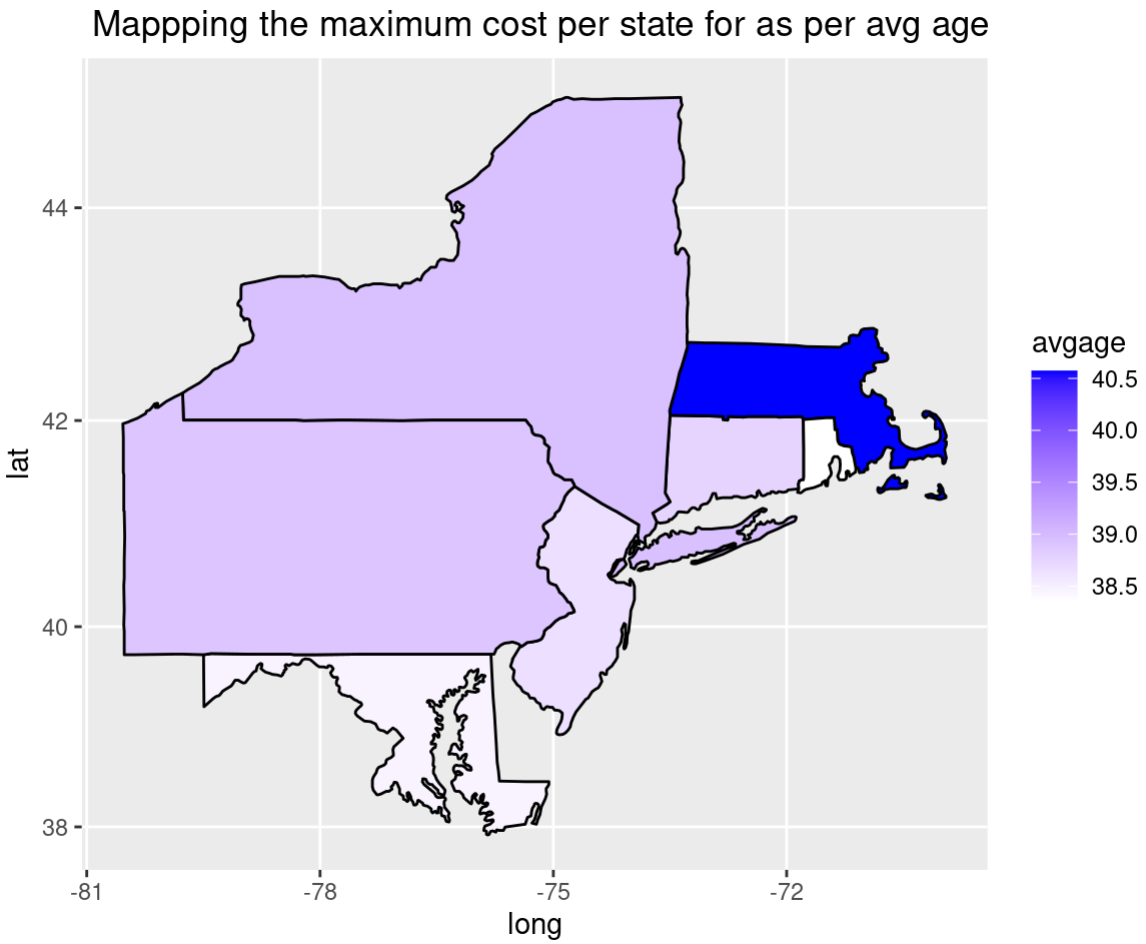

```
#the average number of smokers in the states, and based on the analysis NY has the avg smokers w
ith a frequency of 0.210
```

```
#Maps(Cost based on location)
dfAgg <- MyData %>% group_by(location) %>% summarise(total_cost = max(cost))
dfAgg$state <- tolower(dfAgg$location)
us <- map_data("state")
us$state <- us$region
mergedNew <- merge(dfAgg,us,on = "state")
mergedNew <- mergedNew[order(mergedNew$order),]
map <- ggplot(mergedNew) + geom_polygon(aes(x = long, y = lat, group = group,fill = total_cost),
color = "black")
map + scale_fill_continuous(low = "white", high = "green", name = "total_cost", label = scales::
comma) + coord_map() +ggtitle(" Mappping the maximum cost per state for the expensive and non ex
pensive people")
```



#this maps shows out where healthcare has been utilised max, that is the maximum cost per state for expensive and non expensive people. Based on the map, we can see that the Highest cost is in CT(Connecticut), with a frequency of 55,000$

```
#Maps(Avg age based on location)
dfAgg <- MyData %>% group_by(location) %>% summarise(avgage = mean(age))
dfAgg$state <- tolower(dfAgg$location)
us <- map_data("state")
us$state <- us$region
mergedNew <- merge(dfAgg,us,on = "state")
mergedNew <- mergedNew[order(mergedNew$order),]
map <- ggplot(mergedNew) + geom_polygon(aes(x = long, y = lat, group = group,fill = avgage), col
or = "black")
map + scale_fill_continuous(low = "white", high = "Blue", name = "avgage", label = scales::comm
a) + coord_map() +ggtitle(" Mappping the maximum cost per state for as per avg age")
```




```
# the average age of people using healthcare and their location, based on the map, the Avg age i
s found in MA(Massachusetts) with a frequency of 40.5
```