Final Project

Achraf Safsafi

5/28/2020

According to the Centers for Disease Control and Prevention(CDC), heart disease is the major cause of death in the United States. Around 647,000 Americans die from heart disease each year. In 2017, 365,914 people died because of coronary heart disease (CHD) only. CHD is considered as the most common type of heart disease.

therefore, the project goal is to identify the factor risks of coronary heart disease so it can help reduce those rates using a logistic regression algorithm. The dataset used in this project is from the Framingham Heart study dataset. It is an ongoing heart study on residents of the town of Framingham, Massachusetts published on the Kagle website, https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data. The dataset contains 4238 observations of 16 variables, which are described below.

male : 0 = Female; 1 = Male

age : Age at exam time

education :1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = college

currentSmoker:0 = nonsmoker; 1 = smoker

cigsPerDay:number of cigarettes smoked per day

BPMeds: 0 = not on Blood Pressure medications; 1 = Is on Blood Pressure medications

prevalentStroke : 0= no Prevalent Stroke , 1 = Prevalent Stroke

prevalentHyp :0 = no prevalent hypertension , 1 = has prevalent hypertension

diabetes : 0 = no diabetes ; 1 = has diabetes

totChol: total cholesterol level (mg/dL)

sysBP :systolic blood pressure (mmHg)

dia BP:diastolic blood pressure (mmHg)

BMI :Body Mass Index calculated as: Weight (kg) / Height(meter-squared)

heartRate :Heart Rate

glucose :glucose level (mg/dL)

TenYearCHD :10 year risk of coronary heart disease CHD1 = yes0 = no

The research questions that focus on the problem statement are cited below.

What is the leading medical history risk factor for coronary heart disease? What is the leading medical current risk factor for coronary heart disease? how strong is the evidence linking smoking consumption to coronary heart disease? What gender is most likely to have coronary heart disease? How does age affect the risk of coronary heart disease? What are the major risk factors, including demographic, behavioral, and medical, for coronary heart disease?

```
##Understanding the structure of the data
# set working directory
path_loc <- "C:/Users/asafs/Desktop/DSC520-final project"
setwd(path_loc)

# reading in the data
mydata <- read.csv("FraminghamHeartstudydataset.csv")

#view the dimensions and the class of the dataset
class(mydata)

## [1] "data.frame"

dim(mydata)

## [1] 4238   16

library(tidyverse)

## -- Attaching packages ---------------------------------------------------
----------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------
------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

#look at the variable names and types
glimpse(mydata)

## Rows: 4,238
## Columns: 16
## $ male            <int> 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1,
...
## $ age             <int> 39, 46, 48, 61, 46, 43, 63, 45, 52, 43, 50, 43,
46,...
## $ education       <int> 4, 2, 1, 3, 3, 2, 1, 2, 1, 1, 1, 2, 1, 3, 2, 2, 3,
...
```

```
## $ currentSmoker   <int> 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1,
...
## $ cigsPerDay      <int> 0, 0, 20, 30, 23, 0, 0, 20, 0, 30, 0, 0, 15, 0, 9,
...
## $ BPMeds          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
...
## $ prevalentStroke <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
...
## $ prevalentHyp    <int> 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1,
...
## $ diabetes        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
...
## $ totChol         <int> 195, 250, 245, 225, 285, 228, 205, 313, 260, 225,
2...
## $ sysBP           <dbl> 106.0, 121.0, 127.5, 150.0, 130.0, 180.0, 138.0,
10...
## $ diaBP           <dbl> 70.0, 81.0, 80.0, 95.0, 84.0, 110.0, 71.0, 71.0,
89...
## $ BMI             <dbl> 26.97, 28.73, 25.34, 28.58, 23.10, 30.30, 33.11,
21...
## $ heartRate       <int> 80, 95, 75, 65, 85, 77, 60, 79, 76, 93, 75, 72,
98,...
## $ glucose         <int> 77, 76, 70, 103, 85, 99, 85, 78, 79, 88, 76, 61,
64...
## $ TenYearCHD      <int> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
...
```

```
#summary of the data
summary(mydata)
```

```
##       male             age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.003   Mean   :0.02963   Mean   :0.005899   Mean   :0.3105
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP            diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.00
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.00
##  Mean   :0.02572   Mean   :236.7   Mean   :132.4   Mean   : 82.89
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 89.88
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.50
##                    NA's   :50
##       BMI          heartRate         glucose          TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.000
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.000
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.97   Mean   :0.152
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.000
##  NA's   :19      NA's   :1        NA's   :388
```

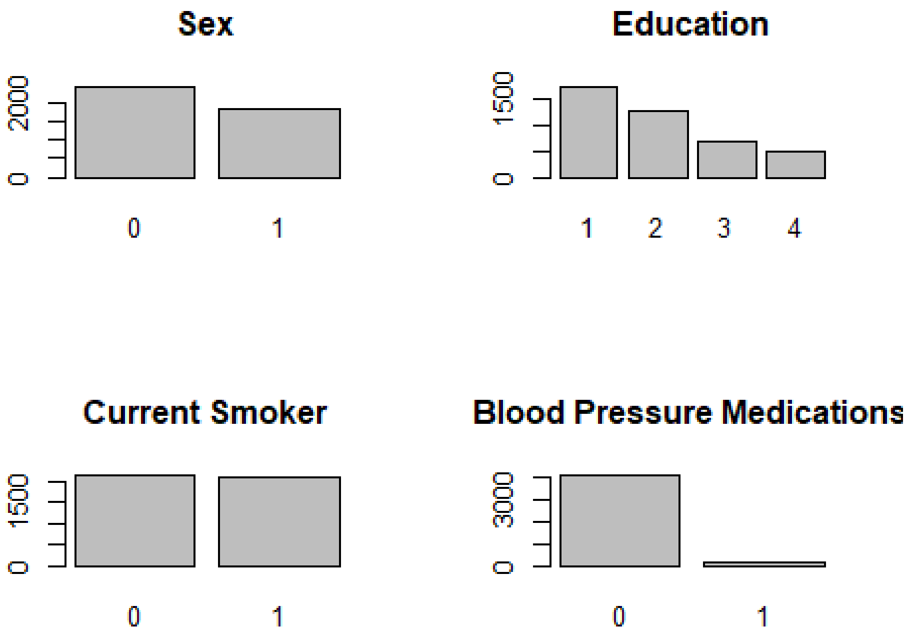
```
# Looking at and visualizing data
head(mydata)
```

```
##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1  39         4             0          0      0               0
## 2    0  46         2             0          0      0               0
## 3    1  48         1             1         20      0               0
## 4    0  61         3             1         30      0               0
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
TenYearCHD
## 1            0        0     195 106.0    70 26.97        80      77
0
## 2            0        0     250 121.0    81 28.73        95      76
0
## 3            0        0     245 127.5    80 25.34        75      70
0
## 4            1        0     225 150.0    95 28.58        65     103
1
## 5            0        0     285 130.0    84 23.10        85      85
0
## 6            1        0     228 180.0   110 30.30        77      99
0
```

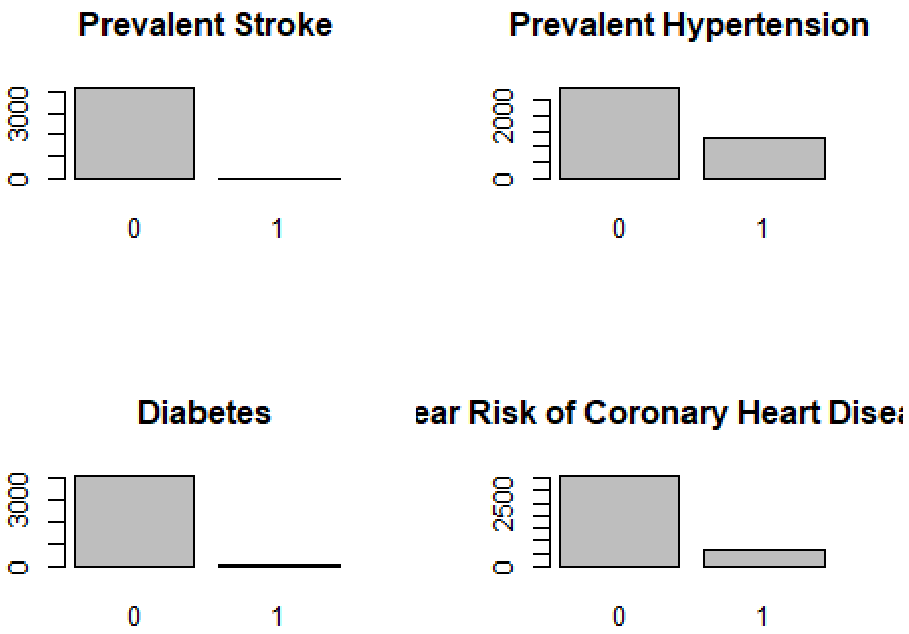
```
tail(mydata)
```

```
##      male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 4233    1  68         1             0          0      0               0
## 4234    1  50         1             1          1      0               0
## 4235    1  51         3             1         43      0               0
## 4236    0  48         2             1         20     NA               0
## 4237    0  44         1             1         15      0               0
## 4238    0  52         2             0          0      0               0
##      prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
## 4233            1        0     176 168.0    97 23.14        60      79
## 4234            1        0     313 179.0    92 25.97        66      86
## 4235            0        0     207 126.5    80 19.71        65      68
## 4236            0        0     248 131.0    72 22.00        84      86
## 4237            0        0     210 126.5    87 19.16        86      NA
## 4238            0        0     269 133.5    83 21.47        80     107
##      TenYearCHD
## 4233          1
## 4234          1
## 4235          0
## 4236          0
## 4237          0
## 4238          0
```

```
##visualizing the data
# categorical variables
# Basic barplot:
par(mfrow=c(2,2))
barplot(table(mydata$male), main="Sex")
barplot(table(mydata$education), main="Education")
barplot(table(mydata$currentSmoker), main="Current Smoker")
barplot(table(mydata$BPMeds), main="Blood Pressure Medications")
```

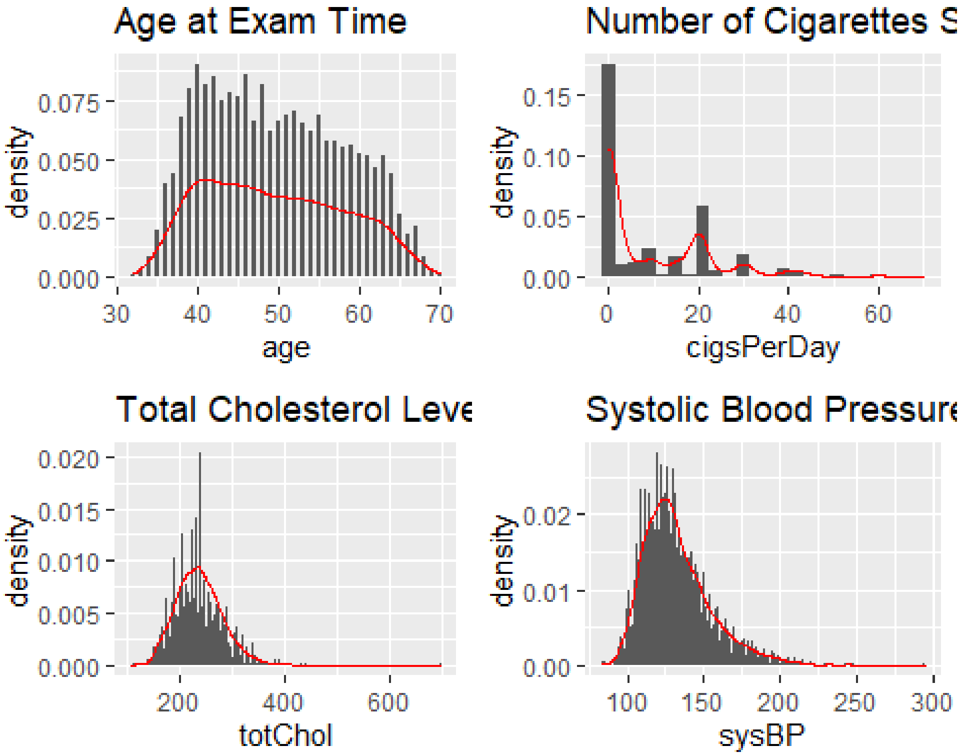


```
barplot(table(mydata$prevalentStroke), main="Prevalent Stroke")
barplot(table(mydata$prevalentHyp), main="Prevalent Hypertension")
barplot(table(mydata$diabetes), main="Diabetes")
barplot(table(mydata$TenYearCHD), main="10 Year Risk of Coronary Heart
Disease CHD")
```

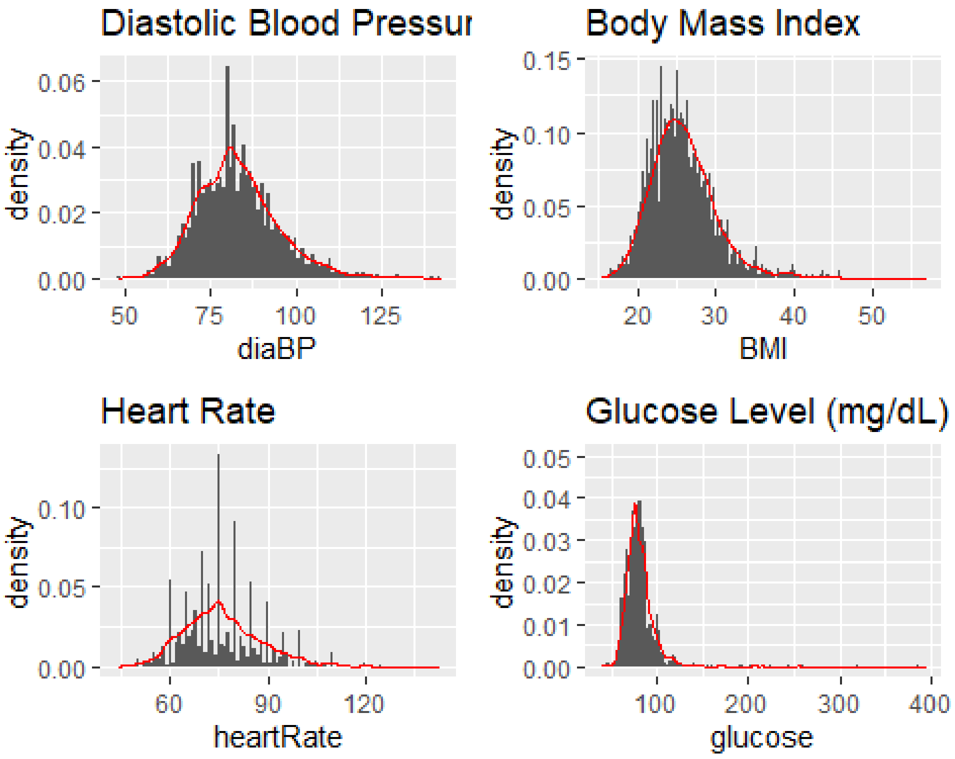


```
##visualizing the data
#numerical variables
library(gridExtra)

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

plot1 <- ggplot(mydata, aes(x=age)) +
  geom_histogram(binwidth= 0.5,aes(y = ..density..))+
  labs(title="Age at Exam Time")+geom_density(col="red")

plot2<- ggplot(mydata, aes(x=cigsPerDay)) +
  geom_histogram(binwidth= 3,aes(y = ..density..))+
  labs(title="Number of Cigarettes Smoked per Day")+geom_density(col="red")

plot3<- ggplot(mydata, aes(x=totChol)) +
  geom_histogram(binwidth= 1,aes(y = ..density..))+
  labs(title="Total Cholesterol Level (mg/dL)")+geom_density(col="red")

plot4<- ggplot(mydata, aes(x=sysBP)) +
  geom_histogram(binwidth= 1,aes(y = ..density..))+
  labs(title="Systolic Blood Pressure (mmHg)")+geom_density(col="red")
grid.arrange(plot1, plot2,plot3,plot4 ,ncol=2, nrow = 2)
```

```
## Warning: Removed 29 rows containing non-finite values (stat_bin).

## Warning: Removed 29 rows containing non-finite values (stat_density).

## Warning: Removed 50 rows containing non-finite values (stat_bin).

## Warning: Removed 50 rows containing non-finite values (stat_density).
```



```
plot5<-ggplot(mydata, aes(x=diaBP)) +
  geom_histogram(binwidth= 1,aes(y = ..density..))+
  labs(title="Diastolic Blood Pressure (mmHg)")+geom_density(col="red")

plot6<- ggplot(mydata, aes(x=BMI)) +
  geom_histogram(binwidth= 0.1,aes(y = ..density..))+
  labs(title="Body Mass Index")+geom_density(col="red")

plot7 <- ggplot(mydata, aes(x=heartRate)) +
  geom_histogram(binwidth= 1,aes(y = ..density..))+
  labs(title="Heart Rate")+geom_density(col="red")

plot8 <- ggplot(mydata, aes(x=glucose)) +
  geom_histogram(binwidth= 1,aes(y = ..density..))+
  labs(title="Glucose Level (mg/dL)")+geom_density(col="red")
grid.arrange(plot5, plot6,plot7,plot8 ,ncol=2, nrow = 2)

## Warning: Removed 19 rows containing non-finite values (stat_bin).

## Warning: Removed 19 rows containing non-finite values (stat_density).
## Warning: Removed 1 rows containing non-finite values (stat_bin).

## Warning: Removed 1 rows containing non-finite values (stat_density).

## Warning: Removed 388 rows containing non-finite values (stat_bin).

## Warning: Removed 388 rows containing non-finite values (stat_density).
```



Finding and replacing missing values

when we run summary function, we see that there are some columns have missing values : (education , 105 NA’s) (cigsPerDay , 29 NA’s) (BPMeds , 53 NA’s) (totChol, 50 NA’s) (BMI , 19 NA’s) (heartRate , 1 NA’s) (glucose, 388 NA’s)

We will use Mean value for missing values replacement.However, if there are many outliers,we will use Median value,and Mode value for categorical variables.

```
# Replacing missing values
names(table(mydata$education))[table(mydata$education)==max(table(mydata$educ
ation))]

## [1] "1"

Mode <- 1
mydata$education=ifelse(is.na(mydata$education),Mode,mydata$education)
mydata$cigsPerDay=ifelse(is.na(mydata$cigsPerDay),median(mydata$cigsPerDay,na
.rm=T),mydata$cigsPerDay)
mydata$BPMeds=ifelse(is.na(mydata$BPMeds),median(mydata$BPMeds,na.rm=T),mydat
a$BPMeds)
mydata$totChol=ifelse(is.na(mydata$totChol),median(mydata$totChol,na.rm=T),my
data$totChol)
mydata$BMI=ifelse(is.na(mydata$BMI),mean(mydata$BMI,na.rm=T),mydata$BMI)
mydata$heartRate=ifelse(is.na(mydata$heartRate),mean(mydata$heartRate,na.rm=T
),mydata$heartRate)
mydata$glucose=ifelse(is.na(mydata$glucose),median(mydata$glucose,na.rm=T),my
data$glucose)
summary(mydata)
```

```
##       male             age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.955   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 8.941   Mean   :0.02926   Mean   :0.005899   Mean   :0.3105
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##     diabetes         totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.00
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.00
##  Mean   :0.02572   Mean   :236.7   Mean   :132.4   Mean   : 82.89
##  3rd Qu.:0.00000   3rd Qu.:262.0   3rd Qu.:144.0   3rd Qu.: 89.88
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.0   Min.   :0.000
##  1st Qu.:23.08   1st Qu.: 68.00   1st Qu.: 72.0   1st Qu.:0.000
##  Median :25.41   Median : 75.00   Median : 78.0   Median :0.000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.6   Mean   :0.152
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 85.0   3rd Qu.:0.000
##  Max.   :56.80   Max.   :143.00   Max.   :394.0   Max.   :1.000
```

```
head(mydata)
```

```
##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1  39         4             0          0      0               0
## 2    0  46         2             0          0      0               0
## 3    1  48         1             1         20      0               0
## 4    0  61         3             1         30      0               0
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
TenYearCHD
```

```
## 1            0        0     195 106.0    70 26.97        80      77
0
## 2            0        0     250 121.0    81 28.73        95      76
0
## 3            0        0     245 127.5    80 25.34        75      70
0
## 4            1        0     225 150.0    95 28.58        65     103
1
## 5            0        0     285 130.0    84 23.10        85      85
0
## 6            1        0     228 180.0   110 30.30        77      99
0
```

```
tail(mydata)
```

```
##      male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 4233    1  68         1             0          0      0               0
## 4234    1  50         1             1          1      0               0
## 4235    1  51         3             1         43      0               0
## 4236    0  48         2             1         20      0               0
## 4237    0  44         1             1         15      0               0
## 4238    0  52         2             0          0      0               0
##      prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
## 4233            1        0     176 168.0    97 23.14        60      79
## 4234            1        0     313 179.0    92 25.97        66      86
## 4235            0        0     207 126.5    80 19.71        65      68
## 4236            0        0     248 131.0    72 22.00        84      86
## 4237            0        0     210 126.5    87 19.16        86      78
## 4238            0        0     269 133.5    83 21.47        80     107
##      TenYearCHD
## 4233          1
## 4234          1
## 4235          0
## 4236          0
## 4237          0
## 4238          0
```

Detect multicollinearity

```
# Computing Variance Inflation Factor VIF
library(usdm)

## Loading required package: sp

## Loading required package: raster

## 
## Attaching package: 'raster'

## The following object is masked from 'package:dplyr':
## 
##     select
## The following object is masked from 'package:tidyr':
## 
##     extract
```

```r
vif(mydata)
```

```
##          Variables      VIF
## 1             male 1.199226
## 2              age 1.403305
## 3        education 1.054576
## 4    currentSmoker 2.454244
## 5       cigsPerDay 2.582759
## 6           BPMeds 1.101648
## 7  prevalentStroke 1.020942
## 8     prevalentHyp 2.054857
## 9         diabetes 1.589788
## 10         totChol 1.106977
## 11           sysBP 3.758255
## 12           diaBP 2.964899
## 13             BMI 1.236628
## 14       heartRate 1.095462
## 15         glucose 1.617096
## 16      TenYearCHD 1.107694
```

The results above shows that there is no collinearity ,all variables are moderately correlated. All values of VIF below 5.

```r
##splitting the data set into training(80%) and testing(20%)data set
set.seed(123)
training <- sample(1:nrow(mydata),size=nrow(mydata)*0.8,replace = FALSE)
train.mydata <- mydata[training,]
test.mydata <- mydata[-training,]
head(test.mydata)
```

```
##    male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 6     0  43         2             0          0      0               0
## 14    0  41         3             0          0      1               0
## 22    0  43         1             0          0      0               0
## 47    0  65         1             0          0      0               0
## 50    1  36         3             1         20      0               0
## 53    0  47         2             1         20      0               0
##    prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose
TenYearCHD
## 6             1        0     228 180.0 110.0 30.30        77      99
0
## 14            1        0     332 124.0  88.0 31.31        65      84
0
## 22            0        0     185 123.5  77.5 29.89        70      78
0
## 47            1        0     252 179.5 114.0 30.47        90      87
0
```

```
## 50            1        0     194 139.0  93.0 24.33        80      62
0
## 53            0        0     237 130.0  78.0 19.66        80      75
0
```

```
# building model
glm_model <- glm(TenYearCHD ~ ., data = train.mydata, family=binomial)
summary(glm_model)
```

```
## 
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = train.mydata)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9662  -0.5881  -0.4210  -0.2817   2.8543  
## 
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)     -8.399059   0.758160 -11.078  < 2e-16 ***
## male             0.503588   0.113159   4.450 8.58e-06 ***
## age              0.068211   0.007072   9.645  < 2e-16 ***
## education       -0.012938   0.052287  -0.247 0.804571    
## currentSmoker   -0.008670   0.163030  -0.053 0.957590    
## cigsPerDay       0.023080   0.006505   3.548 0.000388 ***
## BPMeds           0.547105   0.249288   2.195 0.028187 *  
## prevalentStroke  0.798849   0.519041   1.539 0.123783    
## prevalentHyp     0.155124   0.145070   1.069 0.284931    
## diabetes         0.060734   0.333566   0.182 0.855524    
## totChol          0.001684   0.001179   1.428 0.153318    
## sysBP            0.015236   0.003939   3.868 0.000110 ***
## diaBP           -0.002418   0.006783  -0.356 0.721473    
## BMI             -0.001561   0.013379  -0.117 0.907137    
## heartRate       -0.002605   0.004397  -0.592 0.553556    
## glucose          0.007012   0.002506   2.799 0.005131 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2520.6  on 3374  degrees of freedom
## AIC: 2552.6
## 
## Number of Fisher Scoring iterations: 5
```

According to the results above the variables,male,age,cigsPerDay,BPMeds,sysBP , and glucose are connecting in a statistically way to the dependent variable,TenYearCHD .

```
# Accuracy
predictTrain <- predict(glm_model,newdata= test.mydata,type="response")
```

```
Table <- table(test.mydata$TenYearCHD,predictTrain>0.5)
Table

## 
##     FALSE TRUE
##   0   699   14
##   1   128    7

Accuracy <- sum(diag(Table))/sum(Table)
Accuracy

## [1] 0.8325472
```

As the accuracy more than 80 % ,The model is not bad

let's rank the variables according to their importance.

```
library(caret)

## Loading required package: lattice

## 
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
## 
##     lift

# ranking  the variables according to importance
imp <- as.data.frame(varImp(glm_model, scale = FALSE))
imp <- data.frame(overall = imp$Overall,
           names   = rownames(imp))

imp[order(imp$overall,decreasing = T),]

##       overall           names
## 2  9.64485087             age
## 1  4.45026719            male
## 11 3.86815006           sysBP
## 5  3.54820330      cigsPerDay
## 15 2.79872064         glucose
## 6  2.19466754          BPMeds
## 7  1.53908583 prevalentStroke
## 10 1.42790974         totChol
## 8  1.06930863    prevalentHyp
## 14 0.59243930       heartRate
## 12 0.35649074           diaBP
## 3  0.24743611       education
## 9  0.18207436        diabetes
## 13 0.11665109             BMI
## 4  0.05317785    currentSmoker
```

According to the results abovw,we will remove the variables that less important and we keep only the importance ones .

```
# create new dataframe
library(dplyr)

important <-
select(mydata,male,age,sysBP,cigsPerDay,glucose,BPMeds,TenYearCHD)
head(important)

##   male age sysBP cigsPerDay glucose BPMeds TenYearCHD
## 1    1  39 106.0          0      77      0          0
## 2    0  46 121.0          0      76      0          0
## 3    1  48 127.5         20      70      0          0
## 4    0  61 150.0         30     103      0          1
## 5    0  46 130.0         23      85      0          0
## 6    0  43 180.0          0      99      0          0

##splitting the data set into training(80%) and testing(20%)data set
set.seed(123)
training1 <- sample(1:nrow(important),size=nrow(important)*0.8,replace =
FALSE)
train.important <- important[training1,]
test.important <- important[-training1,]
dim(test.important)

## [1] 848   7

head(test.important)

##    male age sysBP cigsPerDay glucose BPMeds TenYearCHD
## 6     0  43 180.0          0      99      0          0
## 14    0  41 124.0          0      84      1          0
## 22    0  43 123.5          0      78      0          0
## 47    0  65 179.5          0      87      0          0
## 50    1  36 139.0         20      62      0          0
## 53    0  47 130.0         20      75      0          0

# building a new model
glm_model1 <- glm(TenYearCHD ~ ., data = train.important, family=binomial)
summary(glm_model1)

## 
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = train.important)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.0216  -0.5866  -0.4208  -0.2880   2.8334  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.714819   0.449247 -19.399  < 2e-16 ***
## male         0.489014   0.109705   4.458 8.29e-06 ***
## age          0.071027   0.006729  10.556  < 2e-16 ***
## sysBP        0.016394   0.002277   7.201 5.98e-13 ***
## cigsPerDay   0.022786   0.004388   5.193 2.07e-07 ***
## glucose      0.007296   0.001886   3.869 0.000109 ***
## BPMeds       0.620694   0.246943   2.514 0.011954 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2526.7  on 3383  degrees of freedom
## AIC: 2540.7
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(summary(glm_model1)))
```

```
##                 Estimate Std. Error      z value Pr(>|z|)
## (Intercept) 0.0001641354   1.567132 3.760498e-09 1.000000
## male        1.6307075109   1.115948 8.627621e+01 1.000008
## age         1.0736101400   1.006751 3.840083e+04 1.000000
## sysBP       1.0165287159   1.002279 1.340608e+03 1.000000
## cigsPerDay  1.0230476739   1.004398 1.799844e+02 1.000000
## glucose     1.0073226554   1.001888 4.787434e+01 1.000109
## BPMeds      1.8602180795   1.280106 1.234821e+01 1.012025
```

```
# Accuracy
predictTrain1 <- predict(glm_model1,newdata= test.important,type="response")
Table1 <- table(test.important$TenYearCHD,predictTrain1 >0.5)
Table1
```

```
##
##     FALSE TRUE
##   0   700   13
##   1   128    7
```

```
Accuracy1 <- sum(diag(Table1))/sum(Table1)
Accuracy1
```

```
## [1] 0.8337264
```

The accuracy increases negligibly.

the next model will remove BPMeds variable and we will check the new accuracy.

```
# building a new model
glm_model2 <- glm(TenYearCHD ~ .-BPMeds, data = train.important,
family=binomial)
summary(glm_model2)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ . - BPMeds, family = binomial, data =
train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0632  -0.5888  -0.4223  -0.2875   2.8421
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.876123   0.444338 -19.976  < 2e-16 ***
## male         0.482794   0.109536   4.408 1.05e-05 ***
## age          0.071238   0.006720  10.600  < 2e-16 ***
## sysBP        0.017647   0.002216   7.964 1.67e-15 ***
## cigsPerDay   0.022622   0.004386   5.158 2.49e-07 ***
## glucose      0.007378   0.001867   3.951 7.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2532.8  on 3384  degrees of freedom
## AIC: 2544.8
##
## Number of Fisher Scoring iterations: 5
```

```
# Accuracy
predictTrain2 <- predict(glm_model2,newdata= test.important,type="response")
Table2 <- table(test.important$TenYearCHD,predictTrain2 >0.5)
Table2
```

```
##
##     FALSE TRUE
##   0   703   10
##   1   128    7
```

```
Accuracy2 <- sum(diag(Table2))/sum(Table2)
Accuracy2
```

```
## [1] 0.8372642
```

The accuracy increases slightly.

the next model will remove the predictor,glucose, and we will check the new model accuracy.

```
# building new model
glm_model3 <- glm(TenYearCHD ~ .-BPMeds-glucose, data = train.important,
family=binomial)
summary(glm_model3)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ . - BPMeds - glucose, family = binomial,
##     data = train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5193  -0.5952  -0.4259  -0.2890   2.8329
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.436931   0.423852 -19.905  < 2e-16 ***
## male         0.491951   0.109098   4.509 6.51e-06 ***
## age          0.072531   0.006691  10.839  < 2e-16 ***
## sysBP        0.018466   0.002198   8.401  < 2e-16 ***
## cigsPerDay   0.021906   0.004360   5.024 5.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2548.5  on 3385  degrees of freedom
## AIC: 2558.5
##
## Number of Fisher Scoring iterations: 5
```

```
# Accuracy
predictTrain3 <- predict(glm_model3,newdata= test.important,type="response")
Table3 <- table(test.important$TenYearCHD,predictTrain3 >0.5)
Table3
```

```
##
##     FALSE TRUE
##   0   705    8
##   1   131    4
```

```
Accuracy3 <- sum(diag(Table3))/sum(Table3)
Accuracy3
```

```
## [1] 0.8360849
```

the model accuracy slightly decreases

Removing the predictor, glucose, affects the model accuracy, So the next model will keep the predictor, glucose, and we remove the predictor, cigsPerDay.Then we check the new model accuracy.

```
# building new model
glm_model4 <- glm(TenYearCHD ~ .-BPMeds-cigsPerDay , data = train.important,
family=binomial)
summary(glm_model4)

##
## Call:
## glm(formula = TenYearCHD ~ . - BPMeds - cigsPerDay, family = binomial,
##     data = train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1071  -0.5898  -0.4317  -0.3007   2.7587
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.308123   0.421967 -19.689  < 2e-16 ***
## male         0.651357   0.103485   6.294 3.09e-10 ***
## age          0.063395   0.006438   9.846  < 2e-16 ***
## sysBP        0.017664   0.002209   7.997 1.28e-15 ***
## glucose      0.007015   0.001854   3.783 0.000155 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2558.8  on 3385  degrees of freedom
## AIC: 2568.8
##
## Number of Fisher Scoring iterations: 5

# Accuracy
predictTrain4 <- predict(glm_model4,newdata= test.important,type="response")
Table4 <- table(test.important$TenYearCHD,predictTrain4 >0.5)
Table4

##
##     FALSE TRUE
##   0   707    6
##   1   129    6

Accuracy4 <- sum(diag(Table4))/sum(Table4)
Accuracy4

## [1] 0.8408019
```

The model accuracy has little improvement. the next model will remove the predictor,sysBP, and we will check the new model accuracy.

```
# building new model
glm_model5 <- glm(TenYearCHD ~ .-BPMeds-cigsPerDay-sysBP  , data =
train.important, family=binomial)
summary(glm_model5)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ . - BPMeds - cigsPerDay - sysBP, family =
binomial,
##     data = train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0294  -0.6035  -0.4423  -0.3304   2.5836
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.749096   0.358972 -18.801  < 2e-16 ***
## male         0.545890   0.100583   5.427 5.72e-08 ***
## age          0.078923   0.006071  12.999  < 2e-16 ***
## glucose      0.008310   0.001810   4.592 4.40e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2623.0  on 3386  degrees of freedom
## AIC: 2631
##
## Number of Fisher Scoring iterations: 5
```

```
# Accuracy
predictTrain5 <- predict(glm_model5,newdata= test.important,type="response")
Table5 <- table(test.important$TenYearCHD,predictTrain5 >0.5)
Table5
```

```
##
##     FALSE TRUE
##   0   712    1
##   1   130    5
```

```
Accuracy5 <- sum(diag(Table5))/sum(Table5)
Accuracy5
```

```
## [1] 0.8455189
```

The model accuracy has little improvement. the next model will remove the predictor,age, and we will check the new model accuracy.

```
# building new model
glm_model6 <- glm(TenYearCHD ~ male + glucose , data = train.important,
family=binomial)
summary(glm_model6)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ male + glucose, family = binomial,
##     data = train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8994  -0.6061  -0.5141  -0.4835   2.2401
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.858503   0.164650 -17.361  < 2e-16 ***
## male         0.464613   0.097304   4.775 1.80e-06 ***
## glucose      0.010859   0.001761   6.166 7.01e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2805.3  on 3387  degrees of freedom
## AIC: 2811.3
##
## Number of Fisher Scoring iterations: 4
```

```
# Accuracy
predictTrain6 <- predict(glm_model6,newdata= test.important,type="response")
Table6 <- table(test.important$TenYearCHD,predictTrain6 >0.5)
Table6
```

```
##
##     FALSE TRUE
##   0   712    1
##   1   132    3
```

```
Accuracy6 <- sum(diag(Table6))/sum(Table6)
Accuracy6
```

```
## [1] 0.8431604
```

the model accuracy slightly decreases Removing the predictor,age, affects the model accuracy, So the next model will keep the predictor,age, and we remove the predictor,male.Then we check the new model accuracy.

```
# building new model
glm_model7 <- glm(TenYearCHD ~ age + glucose   , data = train.important,
family=binomial)
summary(glm_model7)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ age + glucose, family = binomial,
##     data = train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8905  -0.6078  -0.4486  -0.3492   2.5168
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.390911   0.347838 -18.373  < 2e-16 ***
## age          0.076992   0.006022  12.785  < 2e-16 ***
## glucose      0.008289   0.001795   4.619 3.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2652.6  on 3387  degrees of freedom
## AIC: 2658.6
##
## Number of Fisher Scoring iterations: 5
```

```
# Accuracy
predictTrain7 <- predict(glm_model7,newdata= test.important,type="response")
Table7 <- table(test.important$TenYearCHD,predictTrain7 >0.5)
Table7
```

```
##
##     FALSE TRUE
##   0   712    1
##   1   134    1
```

```
Accuracy7 <- sum(diag(Table7))/sum(Table7)
Accuracy7
```

```
## [1] 0.8408019
```

```
# building new model
glm_model8 <- glm(TenYearCHD ~ glucose  , data = train.important,
family=binomial)
summary(glm_model8)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ glucose, family = binomial, data =
train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7923  -0.5691  -0.5522  -0.5249   2.1497
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.641141   0.156144 -16.915  < 2e-16 ***
## glucose      0.010874   0.001757   6.188 6.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2828.1  on 3388  degrees of freedom
## AIC: 2832.1
##
## Number of Fisher Scoring iterations: 4
```

```
# Accuracy
predictTrain8 <- predict(glm_model8,newdata= test.important,type="response")
Table8 <- table(test.important$TenYearCHD,predictTrain8 >0.5)
Table8
```

```
##
##     FALSE TRUE
##   0   712    1
##   1   132    3
```

```
Accuracy8 <- sum(diag(Table8))/sum(Table8)
Accuracy8
```

```
## [1] 0.8431604
```

the model accuracy slightly decreases we will run The next model only with the predictor ,age.

```
# building new model
glm_model9 <- glm(TenYearCHD ~ age  , data = train.important,
family=binomial)
summary(glm_model9)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ age, family = binomial, data = train.important)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0704  -0.6250  -0.4472  -0.3553   2.4885
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.843241   0.322687  -18.11   <2e-16 ***
## age          0.079803   0.005972   13.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2867.7  on 3389  degrees of freedom
## Residual deviance: 2674.4  on 3388  degrees of freedom
## AIC: 2678.4
##
## Number of Fisher Scoring iterations: 5
```

```
# Accuracy
predictTrain9 <- predict(glm_model9,newdata= test.important,type="response")
Table9 <- table(test.important$TenYearCHD,predictTrain9 >0.5)
Table9
```

```
##
##     FALSE
##   0   713
##   1   135
```

```
Accuracy9 <- sum(diag(Table9))/sum(Table9)
Accuracy9
```

```
## [1] 0.8408019
```

we got same result as glm_model7 (~ age + glucose) model.

```
# contingency table
table(train.important$male,train.important$TenYearCHD)
```

```
##
##         0    1
##   0 1690  240
##   1 1191  269
```

```
#male odds ratio
OR <- (1690/1191)*(269/240)
OR
```

```
## [1] 1.590435
```

Men likely have a higher risk to develop coronary heart disease than women.

Conclusion :

blood pressure and having stroke are considered the most medical history risk factors for coronary heart disease. However, glucose level, systolic blood pressure, and total cholesterol level are the leading medical current risk factors. the number of cigarettes that the person smoked on average in one day can be a strong predictor for being experienced CHD. And the age remains the major risk factor for coronary heart disease where the men are more likely to experience a CHD more than women.