customer_prediction.R

sayal

Sat Apr 20 10:53:57 2019

```
#Data Analysis and Desicion-making
#Title- Predicting potential customers for a Health Insurance Company
```

```
#Introduction
```

```
#1.1 Problem Description
#Our analysis is focussed on the problem which insurance providers are facing today
#to define their target market nd plan their sale strategies which helps them increase
#their market share and thereby, maximize their profitability.
```

```
#1 Load Libraries


library(e1071) #Package for Skewness function used for data analysis
```

```
## Warning: package 'e1071' was built under R version 3.5.2
```

```
library(stats) #Package for finding cook's distance
library(ggplot2)#Package for visualisation of data
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
library(Amelia)#package to visually display the missing values
```

```
## Warning: package 'Amelia' was built under R version 3.5.3
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.5, built: 2018-05-07)
## ## Copyright (C) 2005-2019 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(gridExtra)#Package for arranging different plots in a single grid
library(caTools)# Package for validation of models
```

```
## Warning: package 'caTools' was built under R version 3.5.2
```

```
library(ROCR)#Package for ROC graphs
```

```
## Warning: package 'ROCR' was built under R version 3.5.2
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.5.2
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(AER)
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 3.5.2
```

```
## Loading required package: carData
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
#2 Data Pre-Processing
```

```
#Description

#Cross-section data originating from the Medical Expenditure Panel Survey
#survey conducted in 1996.

data(HealthInsurance)

#Format
#A data frame containing 8,802 observations on 11 variables.
```

```
#2.1 Load the data
```

```
library("AER")
data("HealthInsurance")
View(HealthInsurance)
summary(HealthInsurance)
```

```
##  health          age        limit         gender     insurance  married
##  no : 629   Min.   :18.00   no :7571   female:4169   no :1750   no :3369
##  yes:8173   1st Qu.:30.00   yes:1231   male  :4633   yes:7052   yes:5433
##             Median :39.00
##             Mean   :38.94
##             3rd Qu.:48.00
##             Max.   :62.00
##
##  selfemp        family             region     ethnicity
##  no :7731   Min.   : 1.000   northeast:1682   other: 365
##  yes:1071   1st Qu.: 2.000   midwest  :2023   afam :1083
##             Median : 3.000   south    :3075   cauc :7354
##             Mean   : 3.094   west     :2022
##             3rd Qu.: 4.000
##             Max.   :14.000
##
##       education
##  none      :1119
##  ged       : 374
##  highschool:4434
##  bachelor  :1549
##  master    : 524
##  phd       : 135
##  other     : 667
```

```
#Initially we will process the data by discovering and labeling the missing data with NA;
#and converting categorical variable(s) to proper factors with meaningful labels.
```

```
#2.2 FIle Structue and Content
```

```
head(HealthInsurance)
```

```
##   health age limit gender insurance married selfemp family region
## 1    yes  31    no   male       yes     yes     yes      4  south
## 2    yes  31    no female       yes     yes      no      4  south
## 3    yes  54    no   male       yes     yes      no      5   west
## 4    yes  27    no   male       yes      no      no      5   west
## 5    yes  39    no   male       yes     yes      no      5   west
## 6    yes  32    no female        no      no      no      3  south
##   ethnicity  education
## 1      cauc   bachelor
## 2      cauc highschool
## 3      cauc        ged
## 4      cauc highschool
## 5      cauc       none
## 6      afam   bachelor
```

```
str(HealthInsurance)
```

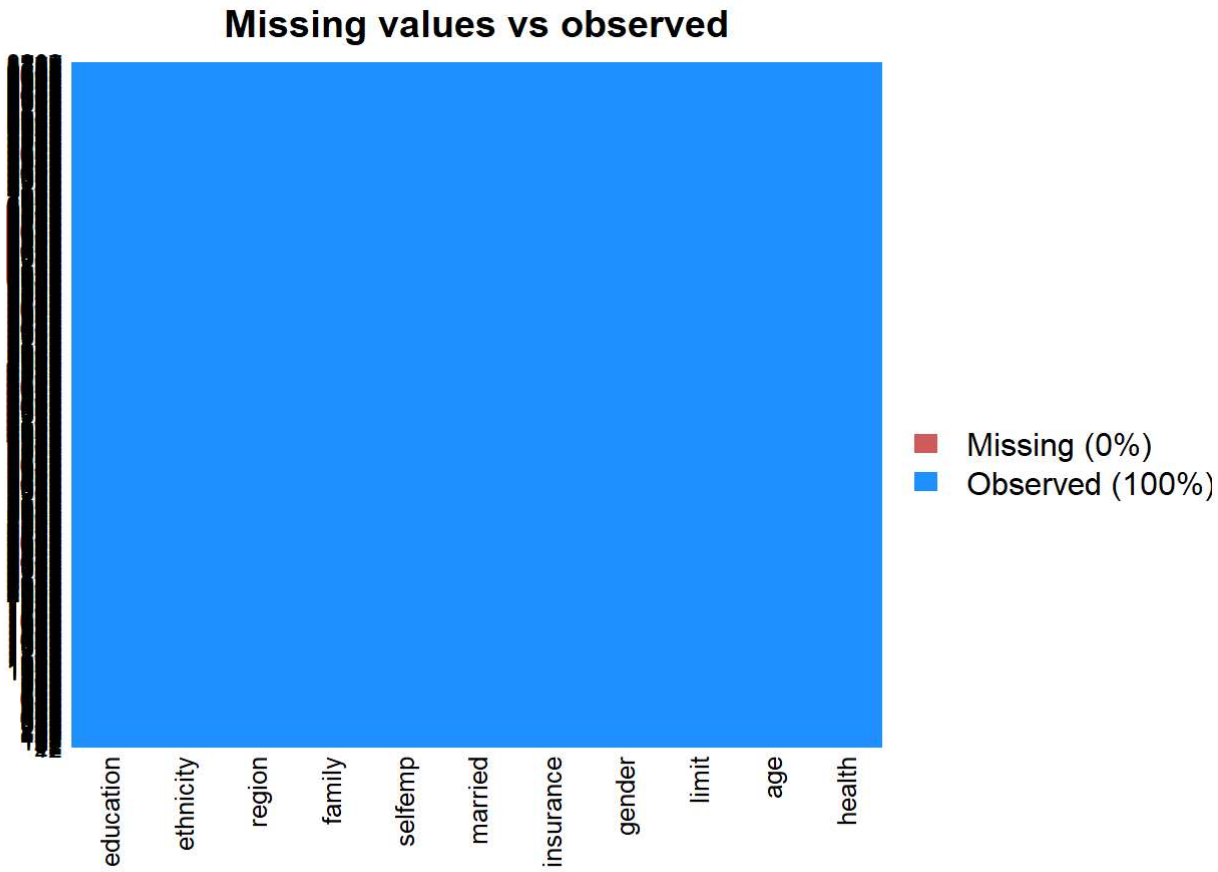
```
## 'data.frame':    8802 obs. of  11 variables:
##  $ health   : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 1 2 2 2 ...
##  $ age      : num  31 31 54 27 39 32 56 60 62 52 ...
##  $ limit    : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 2 1 1 1 ...
##  $ gender   : Factor w/ 2 levels "female","male": 2 1 2 2 2 1 1 1 2 1 ...
##  $ insurance: Factor w/ 2 levels "no","yes": 2 2 2 2 2 1 2 2 2 1 ...
##  $ married  : Factor w/ 2 levels "no","yes": 2 2 2 1 2 1 2 2 2 2 ...
##  $ selfemp  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
##  $ family   : num  4 4 5 5 5 3 2 2 2 2 ...
##  $ region   : Factor w/ 4 levels "northeast","midwest",..: 3 3 4 4 4 3 4 3 3 1 ...
##  $ ethnicity: Factor w/ 3 levels "other","afam",..: 3 3 3 3 3 2 3 3 3 2 ...
##  $ education: Factor w/ 7 levels "none","ged","highschool",..: 4 3 2 3 1 4 3 3 3 3 ...
```

```
#2.3 Missing Values
```

```
#Replace the possible missing values with NA

summary(HealthInsurance)
```

```
##  health          age        limit         gender     insurance  married
##  no : 629   Min.   :18.00   no :7571   female:4169   no :1750   no :3369
##  yes:8173   1st Qu.:30.00   yes:1231   male  :4633   yes:7052   yes:5433
##             Median :39.00
##             Mean   :38.94
##             3rd Qu.:48.00
##             Max.   :62.00
##
##  selfemp        family             region     ethnicity
##  no :7731   Min.   : 1.000   northeast:1682   other: 365
##  yes:1071   1st Qu.: 2.000   midwest  :2023   afam :1083
##             Median : 3.000   south    :3075   cauc :7354
##             Mean   : 3.094   west     :2022
##             3rd Qu.: 4.000
##             Max.   :14.000
##
##       education
##  none      :1119
##  ged       : 374
##  highschool:4434
##  bachelor  :1549
##  master    : 524
##  phd       : 135
##  other     : 667
```

```
HealthInsurance$age[HealthInsurance$age==0]<- NA
HealthInsurance$family[HealthInsurance$family==0]<- NA

#Function missmap is used to check if the data is complete.

missmap(HealthInsurance, main ="Missing values vs observed")
```




```
#2.4 Validating Total factor variable in dataset.
```

```
is.factor(HealthInsurance$ethnicity)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$health)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$limit)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$gender)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$age)
```

```
## [1] FALSE
```

```
is.factor(HealthInsurance$insurance)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$selfemp)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$family)
```

```
## [1] FALSE
```

```
is.factor(HealthInsurance$region)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$married)
```

```
## [1] TRUE
```

```
is.factor(HealthInsurance$education)
```

```
## [1] TRUE
```

```
#We can see that all the variables are factor except age and family variables.
#For Better understanding, of how R is going to deal with the categorical variables,
#We can use the contrasts() function for the factors.


contrasts(HealthInsurance$health)
```

```
##     yes
## no    0
## yes   1
```

```
contrasts(HealthInsurance$limit)
```

```
##     yes
## no    0
## yes   1
```

```
contrasts(HealthInsurance$gender)
```

```
##        male
## female    0
## male      1
```

```
contrasts(HealthInsurance$insurance)
```

```
##     yes
## no    0
## yes   1
```

```
contrasts(HealthInsurance$married)
```

```
##     yes
## no    0
## yes   1
```

```
contrasts(HealthInsurance$selfemp)
```

```
##     yes
## no    0
## yes   1
```

```
contrasts(HealthInsurance$region)
```

```
##           midwest south west
## northeast       0     0    0
## midwest         1     0    0
## south           0     1    0
## west            0     0    1
```

```
contrasts(HealthInsurance$ethnicity)
```

```
##       afam cauc
## other    0    0
## afam     1    0
## cauc     0    1
```

```
contrasts(HealthInsurance$education)
```

```
##            ged highschool bachelor master phd other
## none         0          0        0      0   0     0
## ged          1          0        0      0   0     0
## highschool   0          1        0      0   0     0
## bachelor     0          0        1      0   0     0
## master       0          0        0      1   0     0
## phd          0          0        0      0   1     0
## other        0          0        0      0   0     1
```

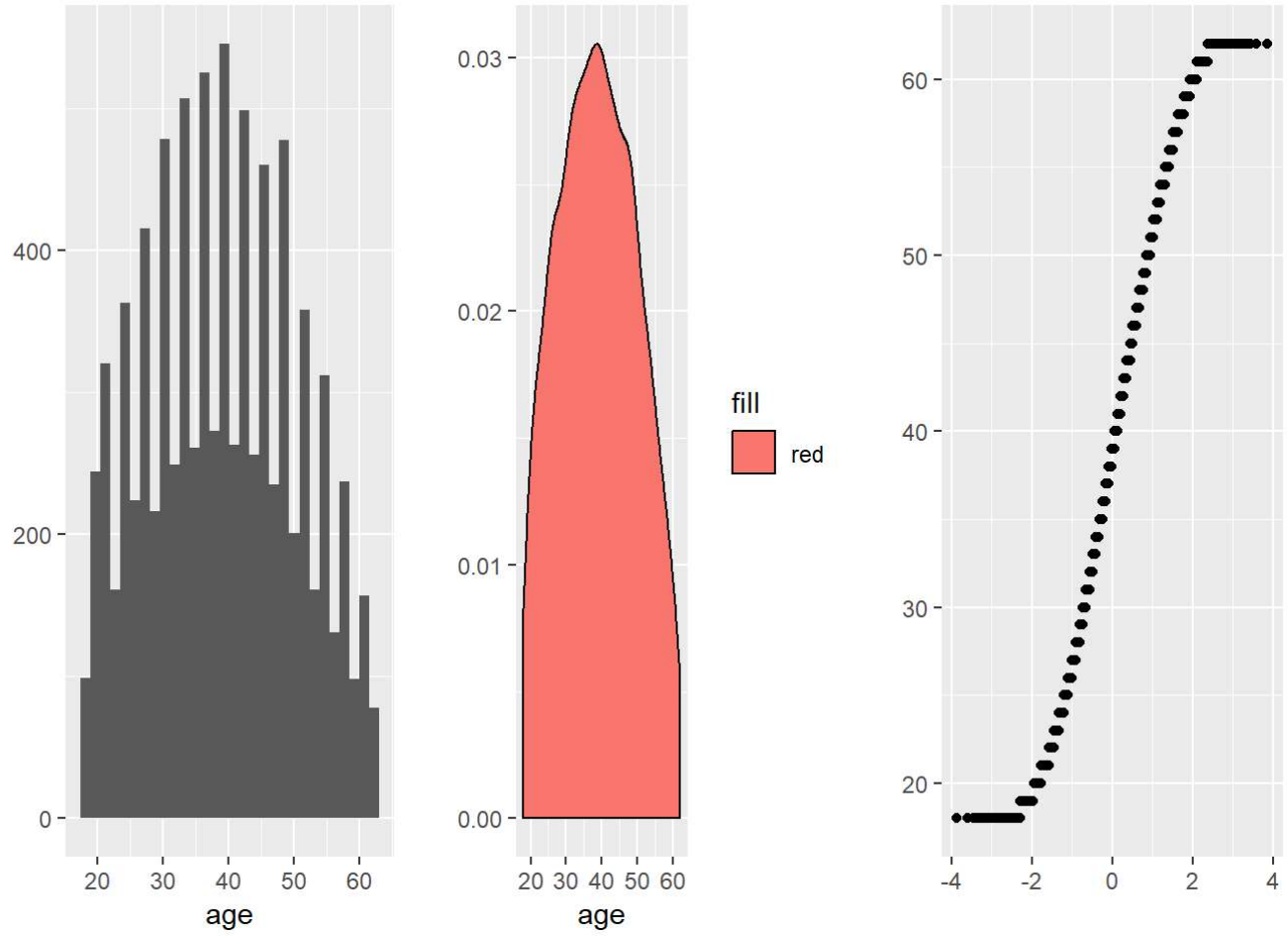
```
#It can be said that the raw data taken is a processed data and
#does not need any cleaning or formatting.
```

```
#3 Exploratory Data Visualisation
```

```
#3.1 Scatter plots for Continous Variable
```

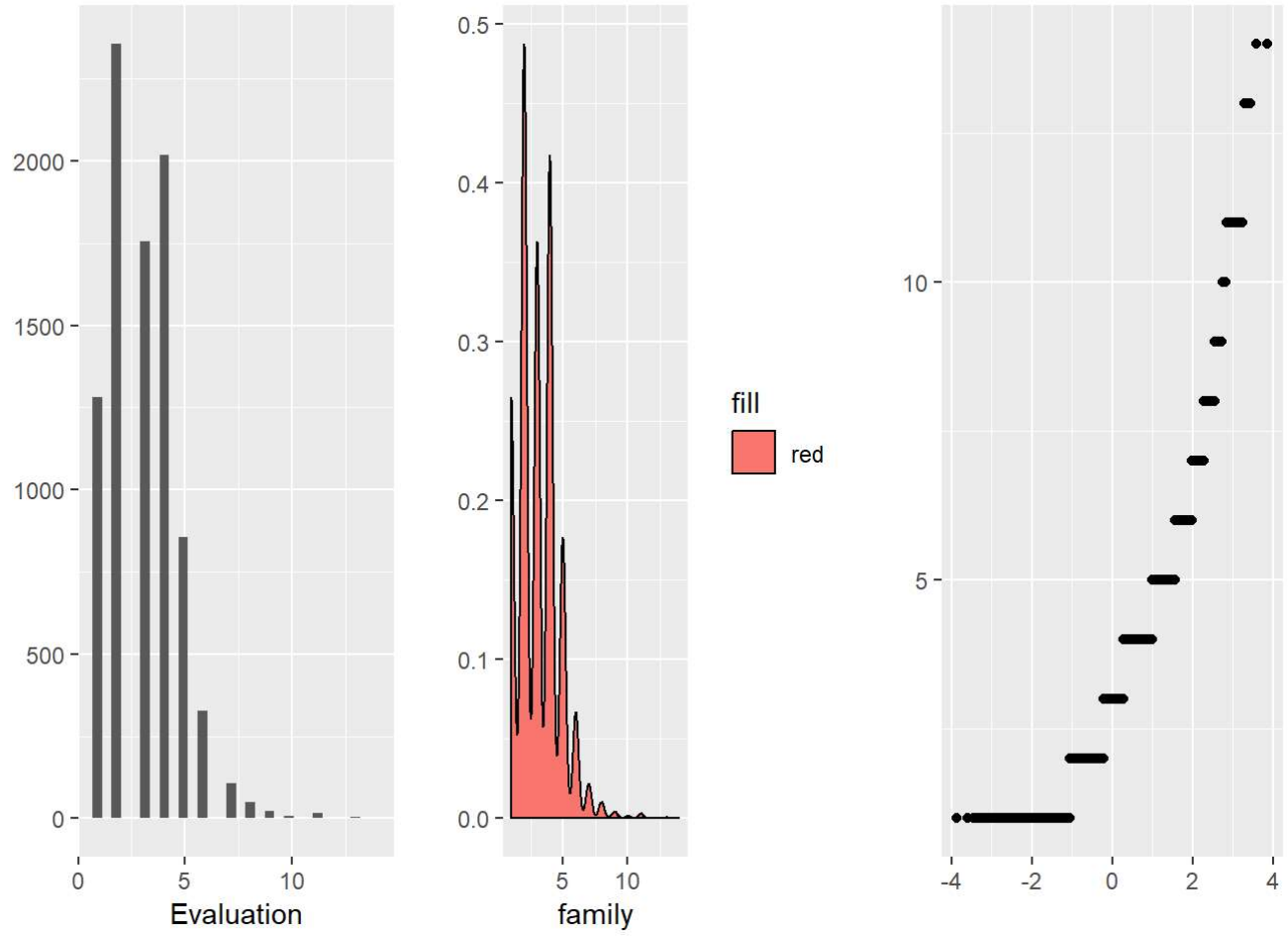
```
#pairs(~health,age,limit,gender,insurance,married,selfemp,family,region,tehnicity,education ,dat
a=HealthInsurance)
#Unlike pairs(), ggpairs() works with non-numeric and predictors in addition to numeric ones.
#Hence we use simple plot for the output.

plot1 = qplot(age, data = HealthInsurance, xlab = "age")
plot2 = qplot(age, data = HealthInsurance, geom = "density", fill = "red")
plot3 = qplot(sample = age, data = HealthInsurance)
grid.arrange(plot1, plot2, plot3, ncol = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
plot4 = qplot(family, data = HealthInsurance, xlab = "Evaluation")
plot5 = qplot(family, data = HealthInsurance, geom = "density", fill = "red")
plot6 = qplot(sample = family, data = HealthInsurance)
grid.arrange(plot4, plot5, plot6, ncol = 3)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Analysis
#On observing the density plot graph , we can conclude that variable age is normally distributed
and
#not skewed as the normal distribution graph is neither left nor right skewed.
#Since the variable is not skewed so we need not use any tranformations like log or sqrt
#to make age variable normally distributed.
#We can observe that the age is uniformly clustered around qauantile[-4:4]
```

```
#3.2 Skewness
```

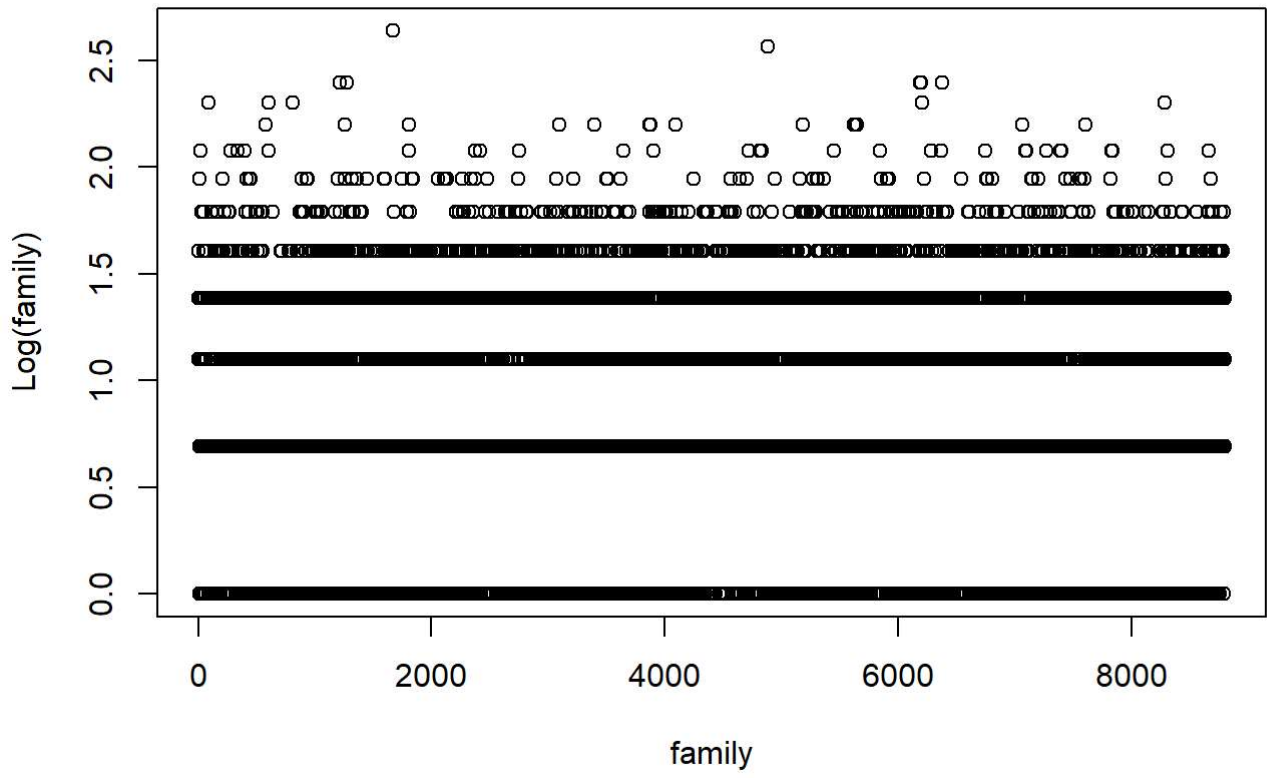
```
#We can also check the skewness using skewness() function:

skewness(HealthInsurance$age)
```

```
## [1] 0.07341805
```

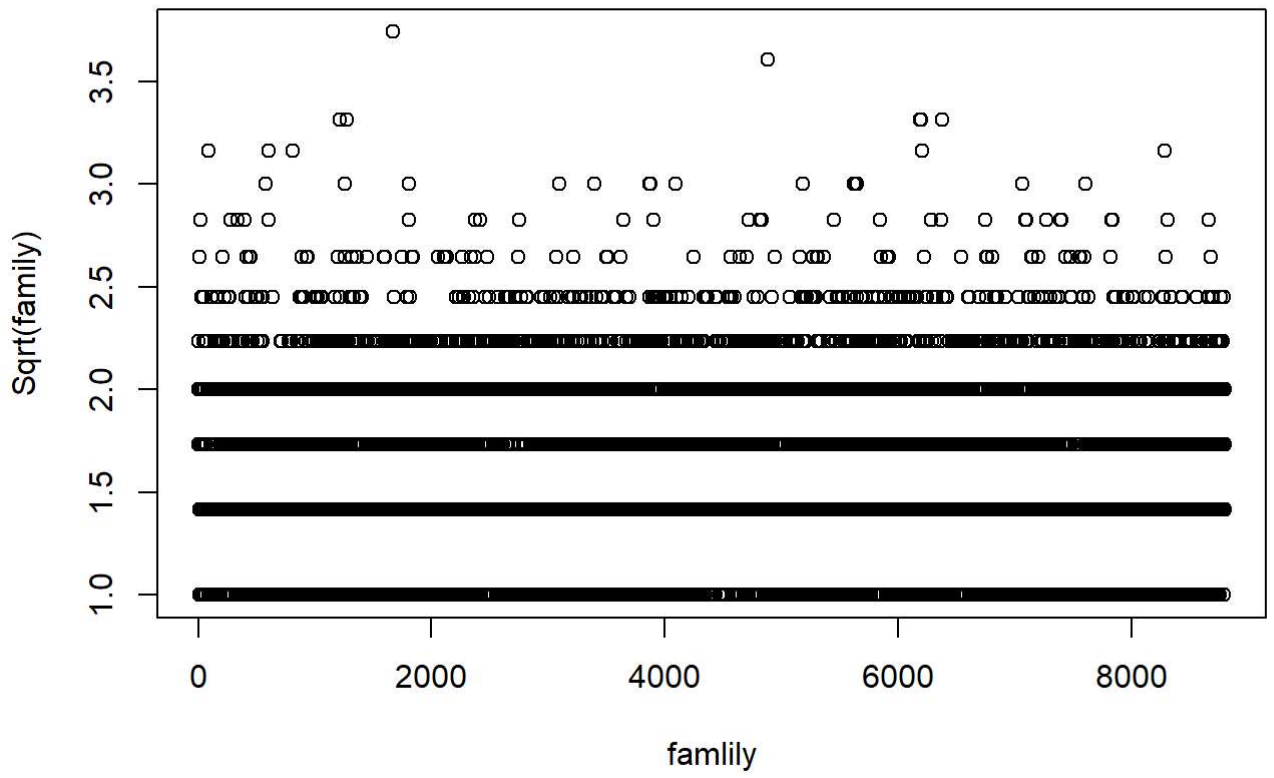
```
skewness(HealthInsurance$family)
```

```
## [1] 0.9832373
```

```
#It can be seen that skewness factor is close to zero for age so
#we can say that age is not skewed. However, family variable exhibit right skewness as the value
s is positive
#hence we can use log or sqrt tranformation for the same.

log.y<-log(HealthInsurance$family)
plot7 <-plot(log.y,xlab="family",ylab="Log(family)")
```

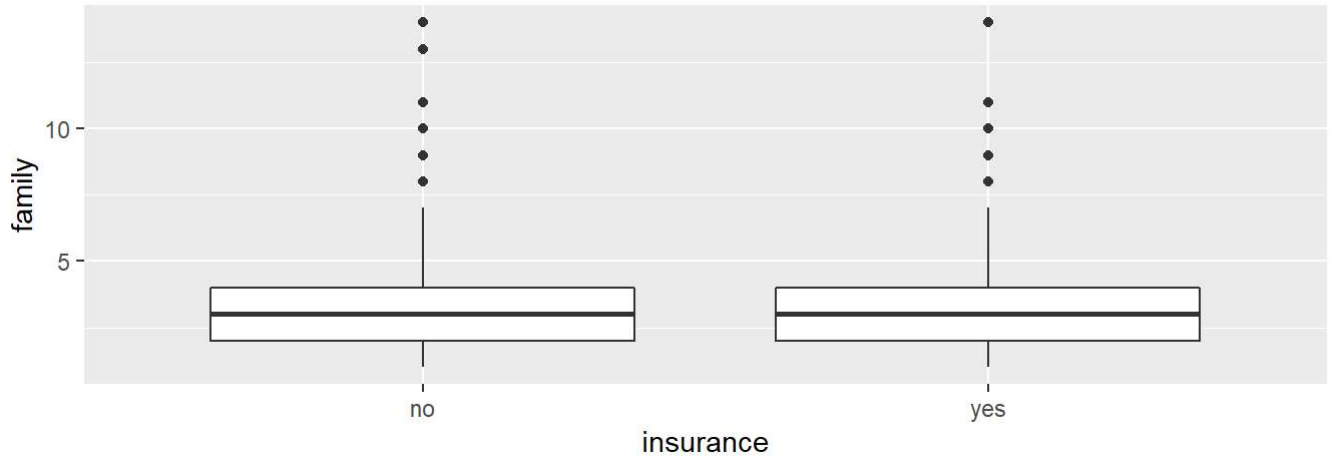



```
sqrt.y<-sqrt(HealthInsurance$family)
plot8 <-plot(sqrt.y,xlab="famlily",ylab="Sqrt(family)")
```




```
#3.3 BoxPlots and Stripcharts
```

```
#Scatterplot matrix of all the continous variables while
#viewing the insurance variable as the output variable.

plot4<- qplot(insurance,family, data=HealthInsurance, geom=c("boxplot"))

plot5<- qplot(insurance,age, data=HealthInsurance, geom=c("boxplot"))

grid.arrange(plot4, plot5)
```



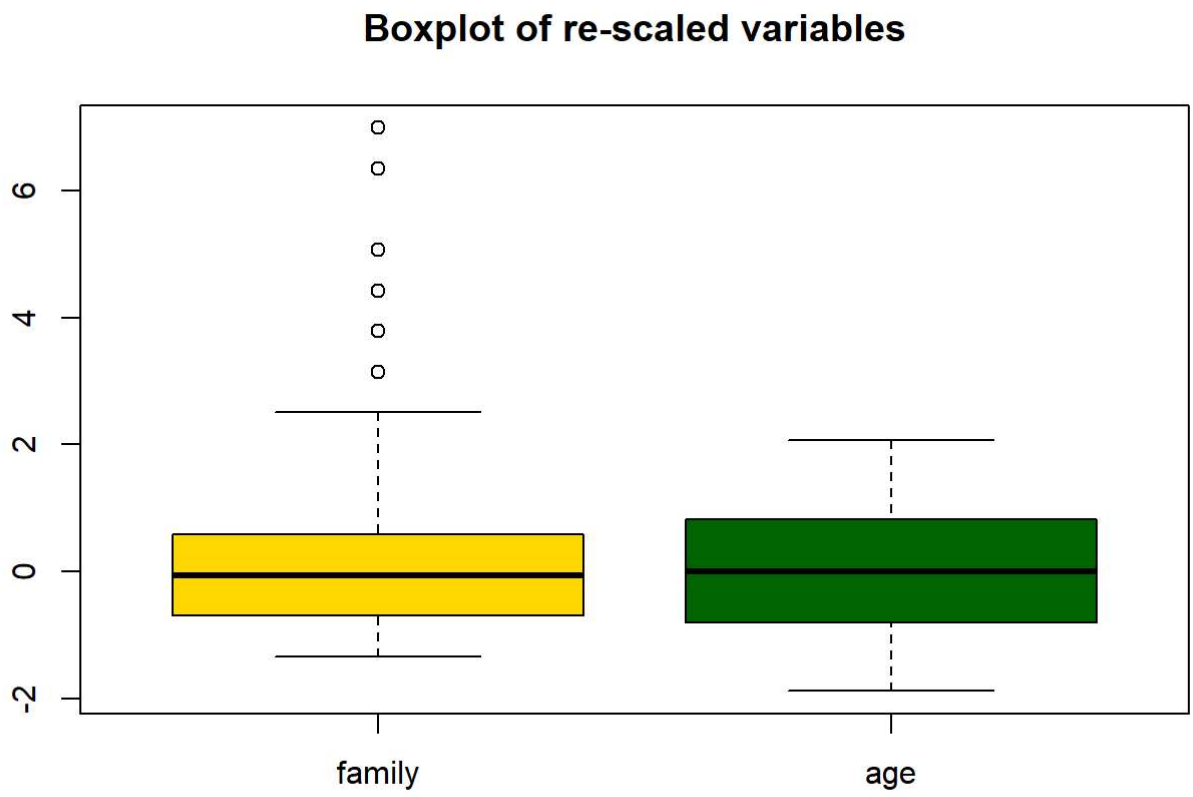

```
#Visual Inspection between the 2 continous variables: Family vs age box plots tells us
#that there is considerable amount of potential outliers for family variable
```

```
#3.4 Z-Scores Scaling
```

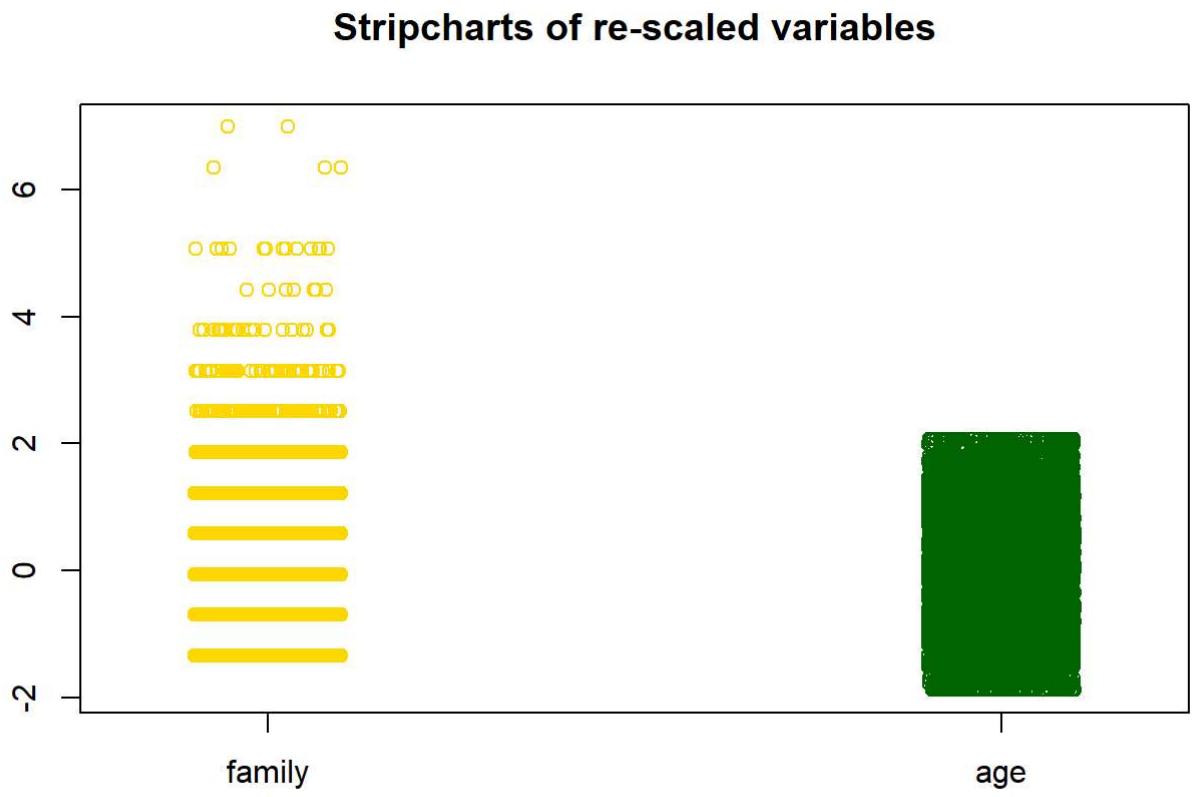
```
numeric_data <- HealthInsurance[,c("family","age")]
numeric_data <- data.frame(scale(numeric_data ))
healthinsurance_r = data.frame(scale(numeric_data))
summary(numeric_data)
```

```
##      family              age          
##  Min.   :-1.34230   Min.   :-1.884364  
##  1st Qu.:-0.70113   1st Qu.:-0.804336  
##  Median :-0.05995   Median : 0.005685  
##  Mean   : 0.00000   Mean   : 0.000000  
##  3rd Qu.: 0.58122   3rd Qu.: 0.815706  
##  Max.   : 6.99299   Max.   : 2.075739
```

```
#Mean after rescaling the variables is 0 for both the variables.
#Checking the 1st and 3rd quantiles for both the variables we can infer
#that they lie between -2 and +2 with few exceptions.


#We will now plot boxplot and strip charts on the basis of z-score

boxplot(numeric_data, main = "Boxplot of re-scaled variables",col = (c("gold","darkgreen")))
```




```
stripchart(numeric_data, vertical = TRUE, method = "jitter", col = (c("gold","darkgreen")), pch
 = 1, main = "Stripcharts of re-scaled variables")
```




```
#It provides confirmation of the variable transformations as all the variables now have mean 0.
#Also , the number of potential outliers are distinctly visible for family after z-score tranfor
mations.
```

```
#3.5 Correlation Matrix
```

```
cor(numeric_data)
```

```
##            family        age
## family  1.0000000 -0.1314596
## age    -0.1314596  1.0000000
```

```
#Family and age have inverse relation which should not be the case as family
#is directly related with age of the person.
```

```
#3.6 CHI Square Test of Independence
```

```
#The Chi Square test of independence is used to determine if there is a
#significant relationship between two categorical variables.

a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$health)
a.data = table(HealthInsurance$insurance, HealthInsurance$health)
print(a.data)
```

```
##
##         no  yes
##   no   171 1579
##   yes  458 6594
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$health)))
```

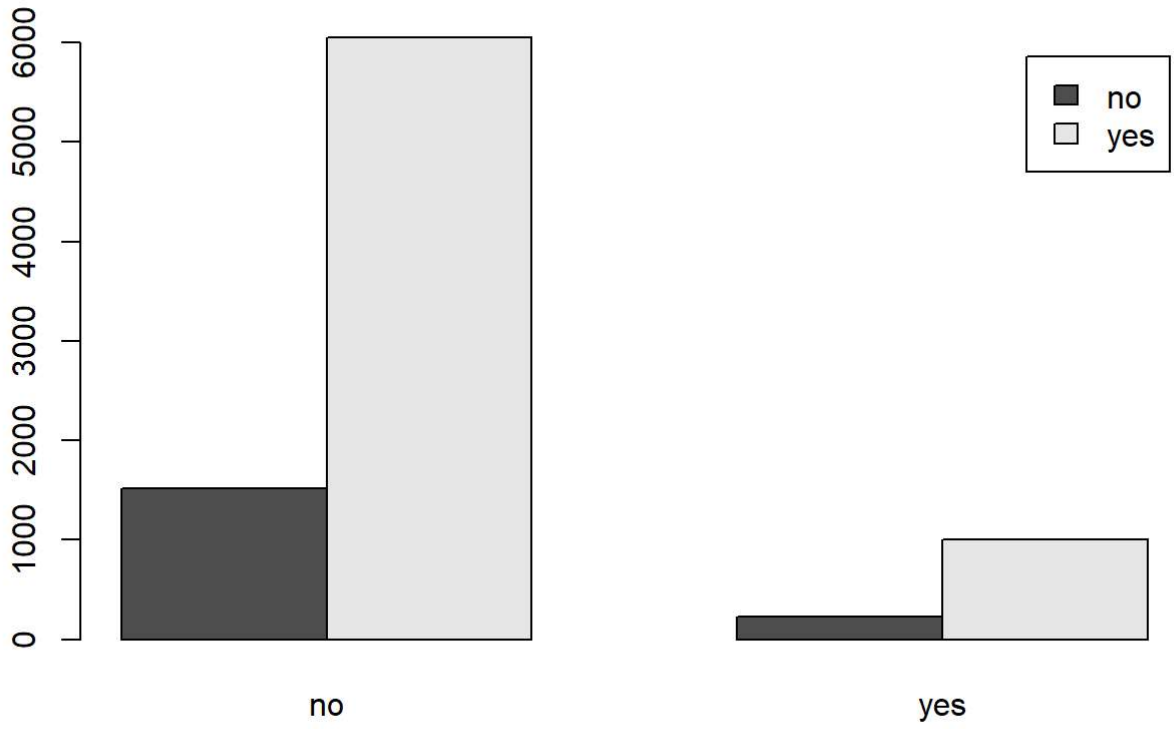

```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  a.data
## X-squared = 22.197, df = 1, p-value = 2.46e-06
```

```
a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$limit)
a.data = table(HealthInsurance$insurance, HealthInsurance$limit)
print(a.data)
```

```
##
##         no  yes
##   no  1519  231
##   yes 6052 1000
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$limit)))
```



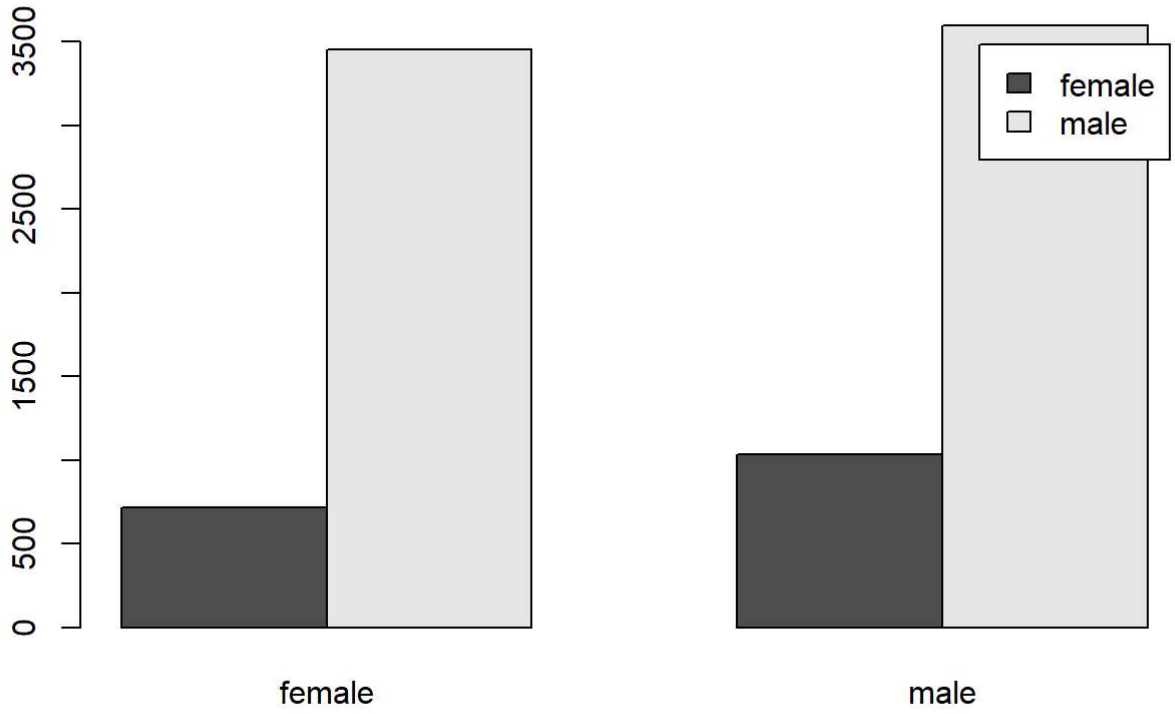

```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  a.data
## X-squared = 1.0402, df = 1, p-value = 0.3078
```

```
a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$gender)
a.data = table(HealthInsurance$insurance, HealthInsurance$gender)
print(a.data)
```

```
##
##       female male
##   no     715 1035
##   yes   3454 3598
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$gender)))
```



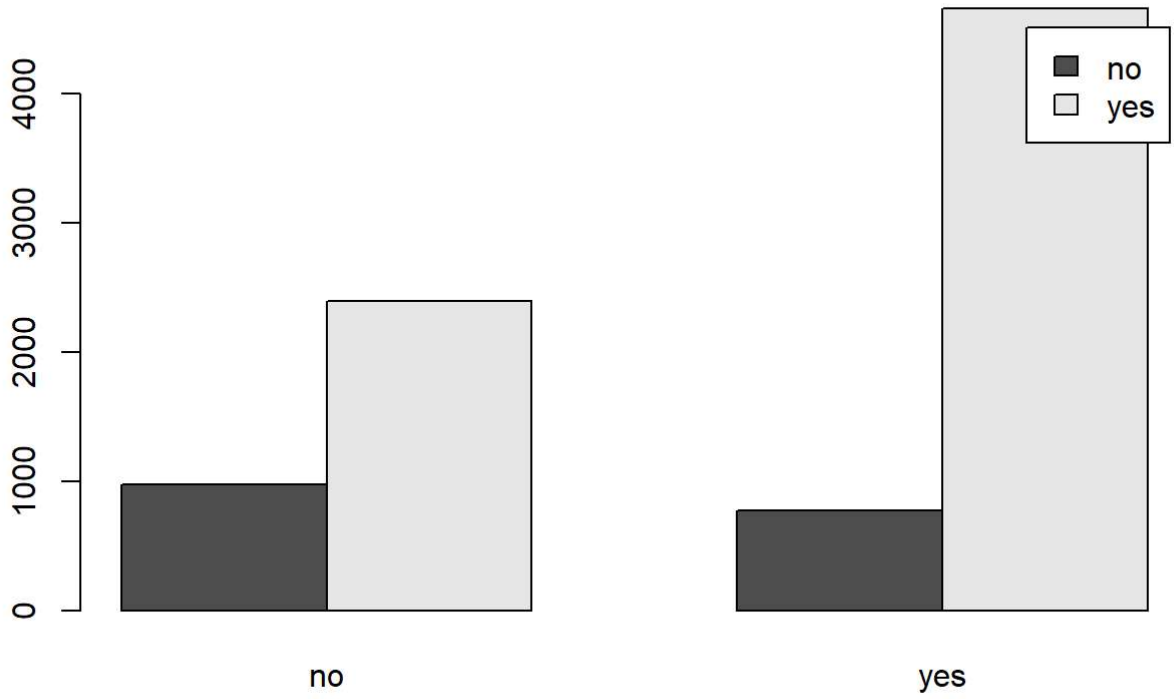

```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  a.data
## X-squared = 36.773, df = 1, p-value = 1.327e-09
```

```
a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$married)
a.data = table(HealthInsurance$insurance, HealthInsurance$married)
print(a.data)
```

```
##
##         no  yes
##   no   976  774
##   yes 2393 4659
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$married)))
```



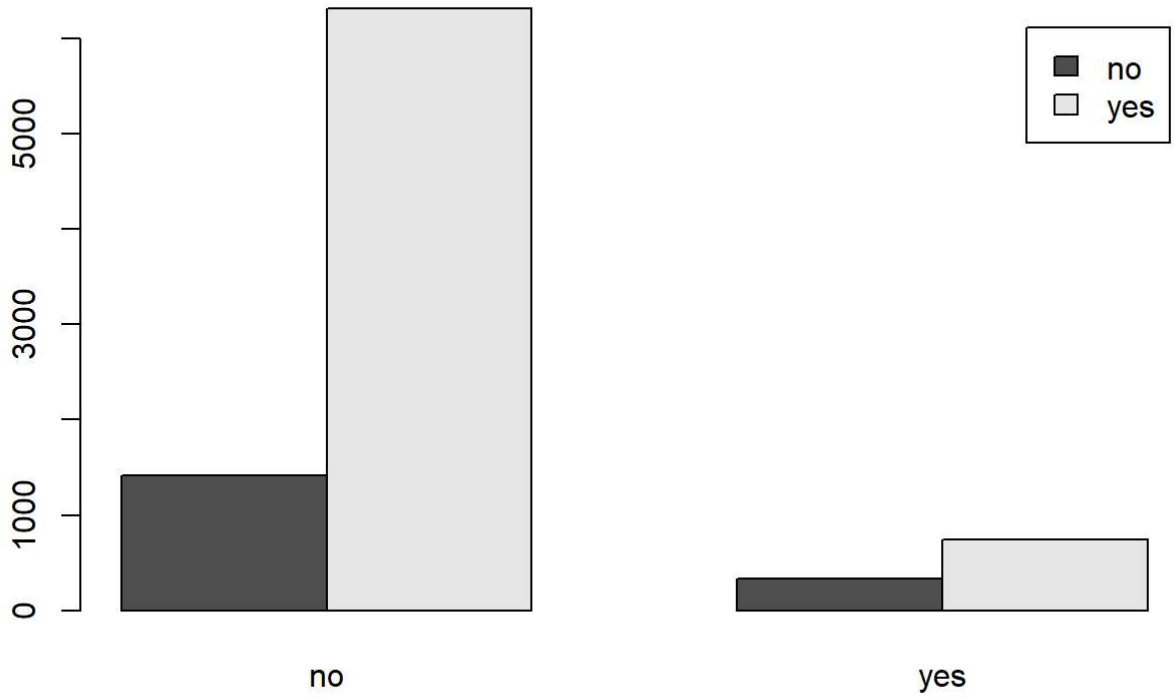

```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  a.data
## X-squared = 282.09, df = 1, p-value < 2.2e-16
```

```
a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$selfemp)
a.data = table(HealthInsurance$insurance, HealthInsurance$selfemp)
print(a.data)
```

```
##
##         no  yes
##   no  1417  333
##   yes 6314  738
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$selfemp)))
```



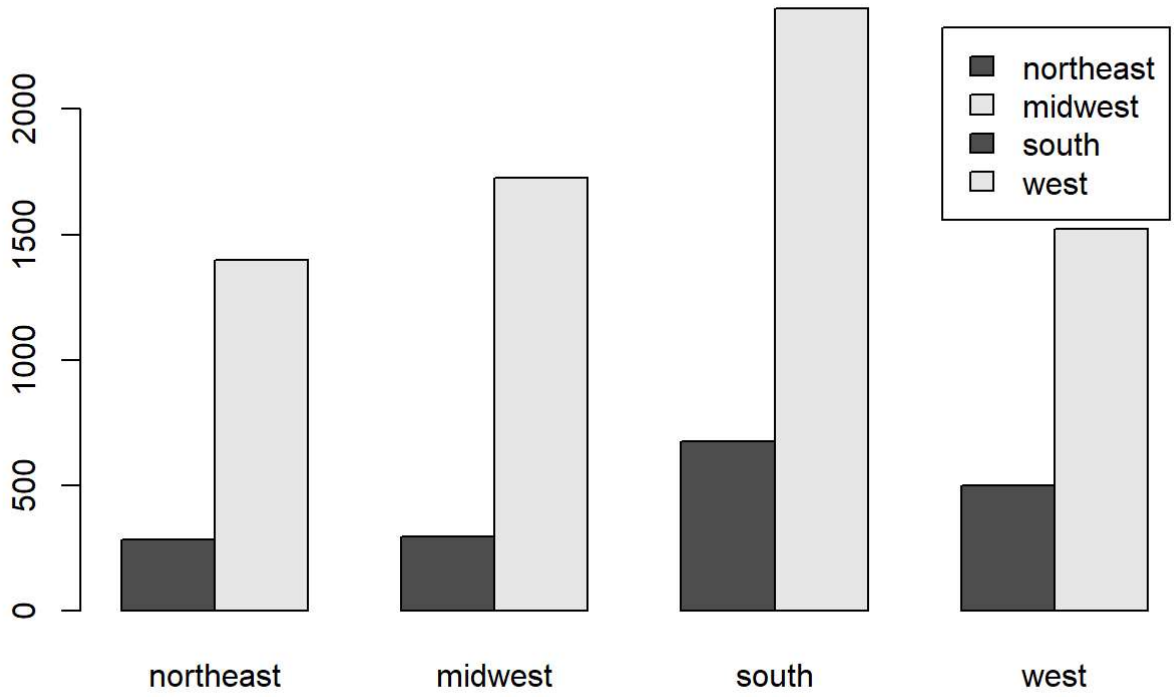

```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  a.data
## X-squared = 95.407, df = 1, p-value < 2.2e-16
```

```
a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$region)
a.data = table(HealthInsurance$insurance, HealthInsurance$region)
print(a.data)
```

```
##
##       northeast midwest south west
##   no        283     296   673  498
##   yes      1399    1727  2402 1524
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$region)))
```



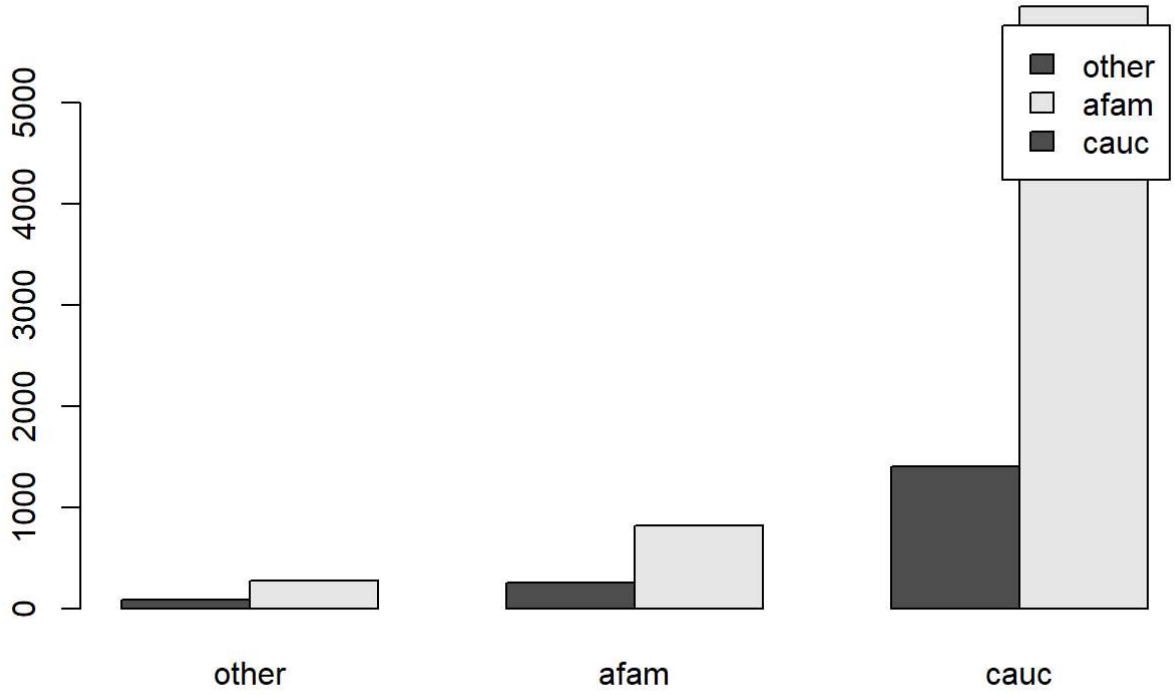

```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test
##
## data:  a.data
## X-squared = 81.234, df = 3, p-value < 2.2e-16
```

```
a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$ethnicity)
a.data = table(HealthInsurance$insurance, HealthInsurance$ethnicity)
print(a.data)
```

```
##
##       other afam cauc
##   no     90  259 1401
##   yes   275  824 5953
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$ethnicity)))
```



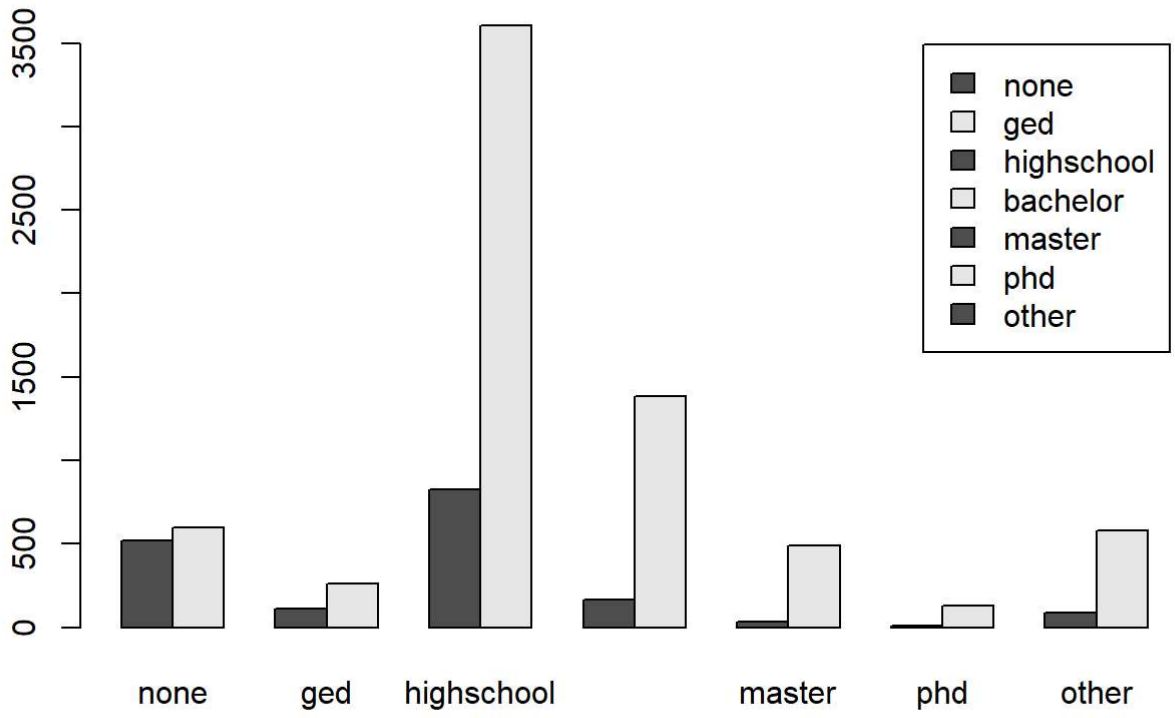

```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test
##
## data:  a.data
## X-squared = 19.474, df = 2, p-value = 5.906e-05
```

```
a.data <- data.frame(HealthInsurance$insurance, HealthInsurance$education)
a.data = table(HealthInsurance$insurance, HealthInsurance$education)
print(a.data)
```

```
##
##       none  ged highschool bachelor master  phd other
##   no   518  109        828      167     33    8    87
##   yes  601  265       3606     1382    491  127   580
```

```
barplot(a.data, beside = TRUE, legend = levels(unique(HealthInsurance$education)))
```




```
print(chisq.test(a.data))
```

```
##
##  Pearson's Chi-squared test
##
## data:  a.data
## X-squared = 691.5, df = 6, p-value < 2.2e-16
```

```
#We can infer by Chi Square test that the variable limit is relatively insignificant
#as compared to other variables.
```

```
#4 Predictive Modelling
```

```
#4.1 Convert the data into training and Test in 80-20 ratio :
set.seed(100)

split<- sample.split(HealthInsurance,SplitRatio = 0.8)
training<- subset(HealthInsurance, split=="TRUE")
testing<- subset(HealthInsurance, split=="FALSE")

dim(training)
```

```
## [1] 6402   11
```

```
dim(testing)
```

```
## [1] 2400   11
```

```
#4.2 Generalized linear Model
```

```
#We would not be implementing linear or polynomial model as the response is not a continous vari
able.
#We need the prediction in Yes or No. Hence we would implement glm model with Binomial distribut
ion.
#We cannot implement Poisson Distribution as we have the factors as "Yes" and "No" which would b
e
#taken as missing values by Poisson.

#Big Model
model1 <- glm (insurance ~. , data = training, family = binomial(link='logit'))
summary(model1)
```

```
##
## Call:
## glm(formula = insurance ~ ., family = binomial(link = "logit"),
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6840   0.2999   0.4558   0.6550   1.9276
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.061622   0.275890  -3.848 0.000119 ***
## healthyes            0.358093   0.123580   2.898 0.003759 **
## age                  0.027872   0.003454   8.070 7.04e-16 ***
## limityes            -0.089587   0.101807  -0.880 0.378877
## gendermale          -0.289364   0.069592  -4.158 3.21e-05 ***
## marriedyes           0.914308   0.078752  11.610  < 2e-16 ***
## selfempyes          -1.171439   0.096223 -12.174  < 2e-16 ***
## family              -0.090280   0.022003  -4.103 4.08e-05 ***
## regionmidwest        0.060715   0.114906   0.528 0.597228
## regionsouth         -0.283113   0.100933  -2.805 0.005032 **
## regionwest          -0.400093   0.107403  -3.725 0.000195 ***
## ethnicityafam        0.051255   0.183890   0.279 0.780455
## ethnicitycauc        0.321003   0.161458   1.988 0.046795 *
## educationged         0.599099   0.158918   3.770 0.000163 ***
## educationhighschool  1.265054   0.091355  13.848  < 2e-16 ***
## educationbachelor    1.769671   0.125885  14.058  < 2e-16 ***
## educationmaster      2.123218   0.225458   9.417  < 2e-16 ***
## educationphd         2.104473   0.384968   5.467 4.59e-08 ***
## educationother       1.435576   0.155235   9.248  < 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: 6367.0  on 6401  degrees of freedom
## Residual deviance: 5489.1  on 6383  degrees of freedom
## AIC: 5527.1
##
## Number of Fisher Scoring iterations: 5
```

```
#We can see that this model does a good job on deciding the goodness of the training model.
#If the p-value is less than or equal to the alpha (i.e p < .05),
#the result is statistically significant. If the p-value is greater than alpha (p > .05),
#the result is statistically insignificant.
```

```
#4.3 Hypothesis Testing for Stepwise Regression
```

```
#No matter how significant a model can be we can still make it better by using
#Hypothesis testing so that all the co-efficients of variable are significant.
#If The below hypothesis holds true as the p-value(ethnicityafam):.78 > .05
#Null Hypothesis for X(Limit): H(1): coef(ethnicityafam)=0

#We can remove the ethnicity to construct a new model
#which would have a much better significane

#Remove Ethnicity to construct the model

model2 <- glm(insurance ~ health+age+limit+gender+married+selfemp+family+region+education, famil
y =binomial (link='logit'),data=training)
summary(model2)
```

```
##
## Call:
## glm(formula = insurance ~ health + age + limit + gender + married +
##     selfemp + family + region + education, family = binomial(link = "logit"),
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6665   0.3018   0.4561   0.6572   1.8491
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.785304   0.224694  -3.495 0.000474 ***
## healthyes            0.373118   0.123496   3.021 0.002517 **
## age                  0.027364   0.003442   7.950 1.87e-15 ***
## limityes            -0.076733   0.101584  -0.755 0.450032
## gendermale          -0.281956   0.069437  -4.061 4.89e-05 ***
## marriedyes           0.942834   0.077926  12.099  < 2e-16 ***
## selfempyes          -1.155650   0.095914 -12.049  < 2e-16 ***
## family              -0.094578   0.021946  -4.310 1.64e-05 ***
## regionmidwest        0.074262   0.114647   0.648 0.517153
## regionsouth         -0.295369   0.100689  -2.933 0.003352 **
## regionwest          -0.397591   0.106031  -3.750 0.000177 ***
## educationged         0.594830   0.158807   3.746 0.000180 ***
## educationhighschool  1.249389   0.091113  13.712  < 2e-16 ***
## educationbachelor    1.762238   0.125728  14.016  < 2e-16 ***
## educationmaster      2.108079   0.225065   9.367  < 2e-16 ***
## educationphd         2.074333   0.384607   5.393 6.91e-08 ***
## educationother       1.423410   0.155134   9.175  < 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: 6367.0  on 6401  degrees of freedom
## Residual deviance: 5499.5  on 6385  degrees of freedom
## AIC: 5533.5
##
## Number of Fisher Scoring iterations: 5
```

```
#We can still observe insignificant variables whose
#values are greater than .05. We'll apply null hypothesis testing again for model 2 and
#see that P-value(regionmidwest)>.05.

#Hence we remove region variable as well.

#Model after removing region variable

model3 <- glm(insurance ~ health+age+limit+gender+married+selfemp+family+education, family =bino
mial (link='logit'),data=training)
summary(model3)
```

```
##
## Call:
## glm(formula = insurance ~ health + age + limit + gender + married +
##     selfemp + family + education, family = binomial(link = "logit"),
##     data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6620   0.3140   0.4606   0.6595   1.8364
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.013380   0.209517  -4.837 1.32e-06 ***
## healthyes            0.382655   0.123099   3.109 0.001880 **
## age                  0.027532   0.003434   8.017 1.08e-15 ***
## limityes            -0.089198   0.101146  -0.882 0.377847
## gendermale          -0.283490   0.069218  -4.096 4.21e-05 ***
## marriedyes           0.940812   0.077608  12.123  < 2e-16 ***
## selfempyes          -1.156233   0.095473 -12.111  < 2e-16 ***
## family              -0.099113   0.021877  -4.530 5.89e-06 ***
## educationged         0.603730   0.158016   3.821 0.000133 ***
## educationhighschool  1.296635   0.090598  14.312  < 2e-16 ***
## educationbachelor    1.805073   0.125199  14.418  < 2e-16 ***
## educationmaster      2.152587   0.224409   9.592  < 2e-16 ***
## educationphd         2.104738   0.383622   5.486 4.10e-08 ***
## educationother       1.490643   0.154195   9.667  < 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: 6367.0  on 6401  degrees of freedom
## Residual deviance: 5529.7  on 6388  degrees of freedom
## AIC: 5557.7
##
## Number of Fisher Scoring iterations: 5
```

```
#We remove limit variable as well.

#Model after removing limit variable

model4 <- glm(insurance ~ health+age+gender+married+selfemp+family+education, family =binomial
 (link='logit'),data=training)
summary(model4)
```

```
##
## Call:
## glm(formula = insurance ~ health + age + gender + married + selfemp +
##     family + education, family = binomial(link = "logit"), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6538   0.3144   0.4614   0.6589   1.8305
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.031598   0.208526  -4.947 7.53e-07 ***
## healthyes            0.400621   0.121427   3.299 0.000969 ***
## age                  0.027051   0.003388   7.984 1.42e-15 ***
## gendermale          -0.281960   0.069189  -4.075 4.60e-05 ***
## marriedyes           0.941053   0.077600  12.127  < 2e-16 ***
## selfempyes          -1.155132   0.095451 -12.102  < 2e-16 ***
## family              -0.097188   0.021767  -4.465 8.01e-06 ***
## educationged         0.596891   0.157887   3.780 0.000157 ***
## educationhighschool  1.295902   0.090594  14.304  < 2e-16 ***
## educationbachelor    1.806640   0.125193  14.431  < 2e-16 ***
## educationmaster      2.158476   0.224313   9.623  < 2e-16 ***
## educationphd         2.109208   0.383653   5.498 3.85e-08 ***
## educationother       1.492153   0.154152   9.680  < 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: 6367.0  on 6401  degrees of freedom
## Residual deviance: 5530.5  on 6389  degrees of freedom
## AIC: 5556.5
##
## Number of Fisher Scoring iterations: 5
```

```
#We can see that all the variables are significant, so model 4 is the potential model to
#predcit the response variable.
```

```
#4.4 Partial F-test for Confidence Intervals
```

```
#For Alpha=.05
#We find the 95% probable interval from the 0.0 and 0.95 quantiles of the F distribution for
#(6388,6401) degree of freedom for model 4

anova(model4)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: insurance
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev
## NULL                       6401     6367.0
## health     1    19.31      6400     6347.7
## age        1   153.56      6399     6194.2
## gender     1    31.45      6398     6162.7
## married    1   122.15      6397     6040.6
## selfemp    1   119.38      6396     5921.2
## family     1    64.58      6395     5856.6
## education  6   326.11      6389     5530.5
```

```
lwr <- qf(0, 6388, model4$df.residual)
upr <- qf(0.95, 6388, model4$df.residual)
c(lwr, upr)
```

```
## [1] 0.00000 1.04202
```

```
#So if the z-value falls outside this interval, we could decide that null Hypothesis as false
#and use alternative hypothesis.

#Here are two Decision Rules based on the F distribution for our case using the pima dataset:
#Risk: a=0.05 Rule: If Z-Value falls within |1.042| Accept Null Hypothesis
#Risk: a=0.05 Rule: If Z-Value greater than |1.042| accept Alternative Hypothesis

#To see the z-value we use the summary() function:

summary(model4)
```

```
##
## Call:
## glm(formula = insurance ~ health + age + gender + married + selfemp +
##     family + education, family = binomial(link = "logit"), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6538   0.3144   0.4614   0.6589   1.8305
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.031598   0.208526  -4.947 7.53e-07 ***
## healthyes            0.400621   0.121427   3.299 0.000969 ***
## age                  0.027051   0.003388   7.984 1.42e-15 ***
## gendermale          -0.281960   0.069189  -4.075 4.60e-05 ***
## marriedyes           0.941053   0.077600  12.127  < 2e-16 ***
## selfempyes          -1.155132   0.095451 -12.102  < 2e-16 ***
## family              -0.097188   0.021767  -4.465 8.01e-06 ***
## educationged         0.596891   0.157887   3.780 0.000157 ***
## educationhighschool  1.295902   0.090594  14.304  < 2e-16 ***
## educationbachelor    1.806640   0.125193  14.431  < 2e-16 ***
## educationmaster      2.158476   0.224313   9.623  < 2e-16 ***
## educationphd         2.109208   0.383653   5.498 3.85e-08 ***
## educationother       1.492153   0.154152   9.680  < 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: 6367.0  on 6401  degrees of freedom
## Residual deviance: 5530.5  on 6389  degrees of freedom
## AIC: 5556.5
##
## Number of Fisher Scoring iterations: 5
```

```
#It can be inferred that even in 5% of probable interval all the variables are significant to pr
edict
#the outcome as the z-value is not lying in the interval. Hence, alternative hypothesis is accep
ted which makes it a significant predictor.
#Conclusion of the test with risk a=0.05 using the P-value


lwrpf <- pf(0, 6388, model4$df.residual)
uprpf <- pf(0.95, 6388, model4$df.residual)
c(lwrpf, uprpf)
```

```
## [1] 0.0000000 0.0201972
```

```
#So if the P-value falls outside this interval, we could decide
#that null Hypothesis H(1) and H(2) is false.
#Here are two Decision Rules based on the F distribution
#for our case using the pima dataset:


#Risk: a=0.05 Rule: If P-value> .0209 Accept Null Hypothesis
#Risk: a=0.05 Rule: If P-value< .0209 Accept Alternate Hypothesis

#To see the p-value we use the summary() function:

summary(model4)
```

```
##
## Call:
## glm(formula = insurance ~ health + age + gender + married + selfemp +
##     family + education, family = binomial(link = "logit"), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6538   0.3144   0.4614   0.6589   1.8305
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.031598   0.208526  -4.947 7.53e-07 ***
## healthyes            0.400621   0.121427   3.299 0.000969 ***
## age                  0.027051   0.003388   7.984 1.42e-15 ***
## gendermale          -0.281960   0.069189  -4.075 4.60e-05 ***
## marriedyes           0.941053   0.077600  12.127  < 2e-16 ***
## selfempyes          -1.155132   0.095451 -12.102  < 2e-16 ***
## family              -0.097188   0.021767  -4.465 8.01e-06 ***
## educationged         0.596891   0.157887   3.780 0.000157 ***
## educationhighschool  1.295902   0.090594  14.304  < 2e-16 ***
## educationbachelor    1.806640   0.125193  14.431  < 2e-16 ***
## educationmaster      2.158476   0.224313   9.623  < 2e-16 ***
## educationphd         2.109208   0.383653   5.498 3.85e-08 ***
## educationother       1.492153   0.154152   9.680  < 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: 6367.0  on 6401  degrees of freedom
## Residual deviance: 5530.5  on 6389  degrees of freedom
## AIC: 5556.5
##
## Number of Fisher Scoring iterations: 5
```

```
#Now It can be inferred that in 5% of probable interval all
#the variable are still significant predictor.
```

```
#4.5 Test for Validating Models Significance
```

```
#Anova Test

anova(model1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: insurance
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       6401     6367.0
## health     1   19.307      6400     6347.7 1.113e-05 ***
## age        1  153.559      6399     6194.2 < 2.2e-16 ***
## limit      1    0.961      6398     6193.2   0.32696
## gender     1   31.668      6397     6161.5 1.829e-08 ***
## married    1  121.353      6396     6040.2 < 2.2e-16 ***
## selfemp    1  119.497      6395     5920.7 < 2.2e-16 ***
## family     1   66.027      6394     5854.7 4.448e-16 ***
## region     3   51.978      6391     5802.7 3.027e-11 ***
## ethnicity  2    6.511      6389     5796.2   0.03856 *
## education  6  307.027      6383     5489.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: insurance
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       6401     6367.0
## health     1   19.307      6400     6347.7 1.113e-05 ***
## age        1  153.559      6399     6194.2 < 2.2e-16 ***
## limit      1    0.961      6398     6193.2     0.327
## gender     1   31.668      6397     6161.5 1.829e-08 ***
## married    1  121.353      6396     6040.2 < 2.2e-16 ***
## selfemp    1  119.497      6395     5920.7 < 2.2e-16 ***
## family     1   66.027      6394     5854.7 4.448e-16 ***
## region     3   51.978      6391     5802.7 3.027e-11 ***
## education  6  303.191      6385     5499.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model3, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: insurance
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       6401     6367.0
## health     1    19.31      6400     6347.7 1.113e-05 ***
## age        1   153.56      6399     6194.2 < 2.2e-16 ***
## limit      1     0.96      6398     6193.2     0.327
## gender     1    31.67      6397     6161.5 1.829e-08 ***
## married    1   121.35      6396     6040.2 < 2.2e-16 ***
## selfemp    1   119.50      6395     5920.7 < 2.2e-16 ***
## family     1    66.03      6394     5854.7 4.448e-16 ***
## education  6   324.93      6388     5529.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model4, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: insurance
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       6401     6367.0
## health     1    19.31      6400     6347.7 1.113e-05 ***
## age        1   153.56      6399     6194.2 < 2.2e-16 ***
## gender     1    31.45      6398     6162.7 2.047e-08 ***
## married    1   122.15      6397     6040.6 < 2.2e-16 ***
## selfemp    1   119.38      6396     5921.2 < 2.2e-16 ***
## family     1    64.58      6395     5856.6 9.279e-16 ***
## education  6   326.11      6389     5530.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

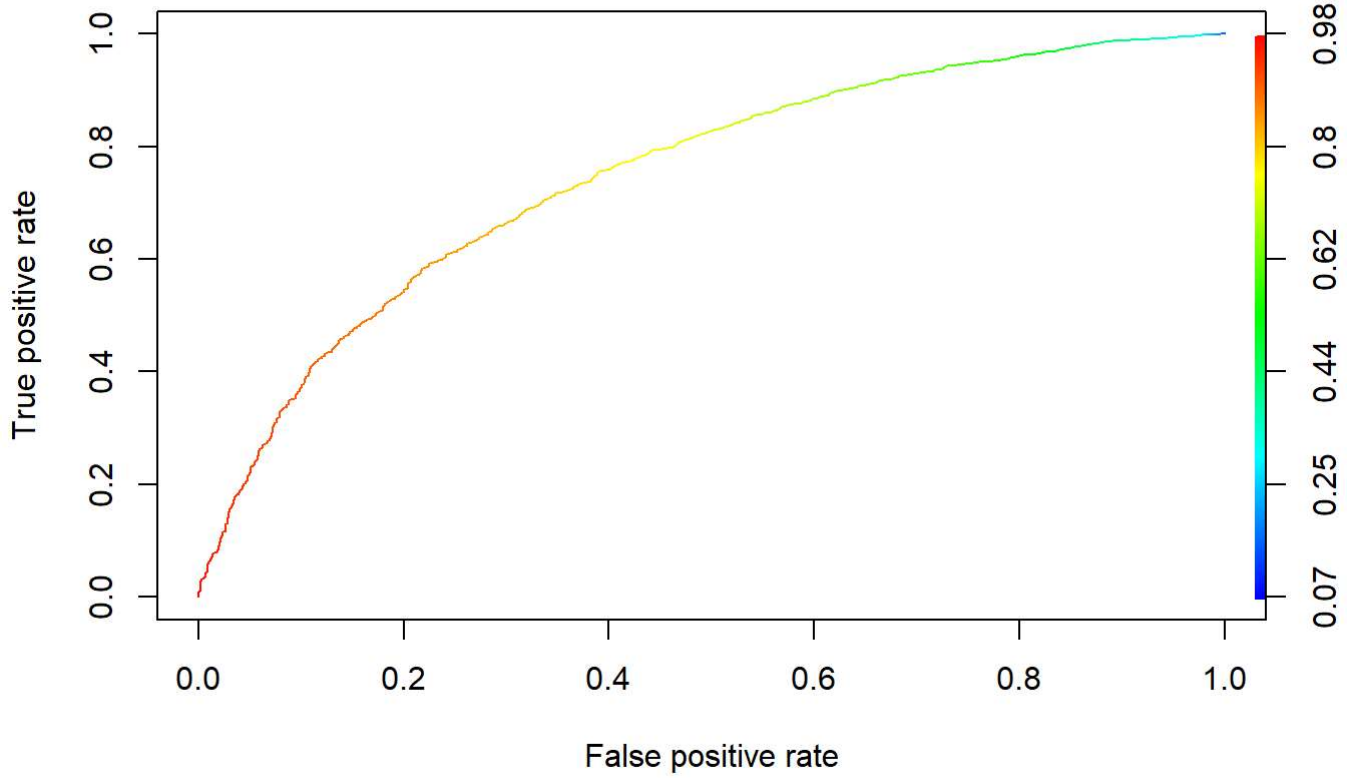
```
#The difference between the null deviance and the residual deviance shows how significant model
#is doing against the null model (a model with only the intercept).
#The wider this gap, the better which is the max for model 4.
#Analyzing the table we can see the increase in deviance when removing
#each variable one at a time.
#We can see that it is a significant increase in deviance and
#the AIC as we go from model 1 to model 4.
```

```
#4.5 Potential Outliers
```

```
plot1 <- qplot(insurance, model4$fitted.values, geom = "boxplot", data=training)+labs(y="Fitted
 Values")+ggtitle("Residuals vs Test Plot")

#The Box Plot for insurance factor variable reveals a lot of outliers
#when compared with fitted values.
```

```
#4.6 ROC Curve
```

```
#To assess the predictive ability of the model we use ROC curve and calculate the AUC(Area under
curve)
#which are typical performance measurementsfor a binary classifier.

pred<- predict(model4, training, type= 'response')
pred<- prediction(pred, training$insurance)
eval<- performance(pred,'tpr','fpr')
plot(eval, colorize = TRUE)
```




```
#The ROC is a curve generated by plotting the true positive rate (TPR) against the false positiv
e rate (FPR) at various threshold settings while the AUC is the area under the ROC curve.
```

```
#4.7 Assesing the predictive ability of the model
```

```
model.probs=predict(model4,training,type="response")

# misclassification error:train data
pred1<- ifelse(model.probs>0.5, 1, 0)
#Confusion Matrix
tab1<- table(Predicted= pred1, Actual= training$insurance)

tab1
```

```
##          Actual
## Predicted   no  yes
##         0  210  158
##         1 1056 4978
```

```
# misclassification error:train data
trainerror<- 1- sum(diag(tab1))/ sum(tab1)
trainerror
```

```
## [1] 0.1896282
```

```
#Accuracy of Training data
print(paste('Accuracy',1-trainerror))
```

```
## [1] "Accuracy 0.810371758825367"
```

```
# Test Error
model.test=predict(model4,testing,type="response")

# misclassification error:test data
pred_test<- ifelse(model.test>0.5, 1, 0)
#Confusion Matrix
tab_test<- table(Predicted= pred_test, Actual = testing$insurance)
tab_test
```

```
##          Actual
## Predicted   no  yes
##         0   79   55
##         1  405 1861
```

```
# misclassification error:test data
testerror<- 1- sum(diag(tab_test))/ sum(tab_test)
testerror
```

```
## [1] 0.1916667
```

```
#Accuracy of Training data
print(paste('Accuracy',1-testerror))
```

```
## [1] "Accuracy 0.808333333333333"
```

```
#The error rate for training is roughly 19% and accuracy is 81.03% which is very high
#compared to real time predictions.
#After fitting the model with the testing data we can observe that the
#acccuracy is 80.833 % which indicates satisfactory goodness of fit of the model.
```

```
#4.8 Area Under the Curve
```

```
auc<- performance(pred,"auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.7486046
```

```
#As a rule of thumb, a model with good predictive ability should have an AUC closer to 1
#(1 is ideal) than to 0.5 which is the case with model 4 having area under the curve as .748604
6.


#4.9 To Find the actual movement and predict whether person takes insurance or not

#To predict Direction for new values of Insurance we simply use the predict() function and
#feed in a data frame of new values. We want to predict Direction on a day when Lag1 and Lag2 eq
ual 1.2 and 1.1, respectively,
#and on a day when they equal 1.5 and -0.8.


predict(model4,newdata=data.frame(health= "yes",age=20,family=3,gender="female",education="bache
lor",married="yes",selfemp="no"),
data=testing,type="response")
```

```
##         1
## 0.9142104
```

```
#As can be seen we can see the actual movement of whether the person has insurance or not by cre
ating a
#new dataframe. Suppose a person walks into hospital with

#health = yes

#age=20

#family=3

#gender=female

#education=bachelor

#married=yes

#selfemp=no

#Then ther are 92.38% prediction chance that she has insurance .
```

```
#5 Conclusion

#This feature can be widely used by the insurance companies to predict whether
#the customer has health insurance or not.This would in turn help to infer the
#potential insurance buyers and help the companies to target the right audience to get maximum h
ealth insurance sales
```