titanic.R

Wed Jan 31 21:11:52 2018

```
library(ggplot2)
library(data.table)
library(dplyr)
library(rpart)
#First take a look at the over structure of the dataset. There are 1310 observations and 14 variables.
titanic <- fread('titanic.csv')
data <- titanic
str(data)
```

```
## Classes 'data.table' and 'data.frame':   1310 obs. of  14 variables:
##  $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived : int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name     : chr  "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson Trevor" "Allison, Miss. H
##  $ sex      : chr  "female" "male" "female" "male" ...
##  $ age      : num  29 0.917 2 30 25 ...
##  $ sibsp    : int  0 1 1 1 1 0 1 0 2 0 ...
##  $ parch    : int  0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket   : chr  "24160" "113781" "113781" "113781" ...
##  $ fare     : num  211 152 152 152 152 ...
##  $ cabin    : chr  "B5" "C22 C26" "C22 C26" "C22 C26" ...
##  $ embarked : chr  "S" "S" "S" "S" ...
##  $ boat     : chr  "2" "11" "" "" ...
##  $ body     : int  NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest: chr  "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
#Remove last 3 variables because they are unlikely related to survival.
data <- data[,-c("name","ticket",12:14)]
```

```
## Warning in `[.data.table`(data, , -c("name", "ticket", 12:14)): column(s)
## not removed because not found: 12,13,14
```

```
attach(data)
```

```
## The following objects are masked from data (pos = 23):
##
##     age, boat, body, cabin, embarked, fare, home.dest, parch,
##     pclass, sex, sibsp, survived
```

```
#Since there are missing values and the dataset is not large, we should try not to delete the rows or c
#that contains NA. Rather, try to predict the missing values.

sapply(data,function(x){sum(is.na(x)|x=="")})
```

```
##    pclass  survived       sex       age     sibsp     parch      fare
##         1         1         1       264         1         1         2
##     cabin  embarked      boat      body home.dest
##      1015         3       824      1189       565
```

```
data <- data[,-"cabin"]
which(is.na(data$survived))
```

```
## [1] 1310
```

```
data <- data[-1310,]
which(is.na(data$fare))
```

```
## [1] 1226
```

```
data[1226]
```

```
##    pclass survived  sex  age sibsp parch fare embarked boat body home.dest
## 1:      3        0 male 60.5     0     0   NA        S       261
```

```
#The only one missing fare is a male in class 3. It is reasonable to predict fare by pclass and embarke
data[1226]$fare <- median(data[data$pclass=='3' & data$embarked=='S']$fare,na.rm = TRUE)

#There are 263 missing age, I will use mice.
library(mice)
library(randomForest)
set.seed(1234)
mice_mod <- mice(data,method='rf')
```

```
##
##  iter imp variable
##   1   1  age  body
##   1   2  age  body
##   1   3  age  body
##   1   4  age  body
##   1   5  age  body
##   2   1  age  body
##   2   2  age  body
##   2   3  age  body
##   2   4  age  body
##   2   5  age  body
##   3   1  age  body
##   3   2  age  body
##   3   3  age  body
##   3   4  age  body
##   3   5  age  body
##   4   1  age  body
##   4   2  age  body
##   4   3  age  body
##   4   4  age  body
##   4   5  age  body
##   5   1  age  body
##   5   2  age  body
##   5   3  age  body
##   5   4  age  body
##   5   5  age  body
```

```
md.pattern(data)
```

```
## Warning in data.matrix(x): NAs introduced by coercion
```

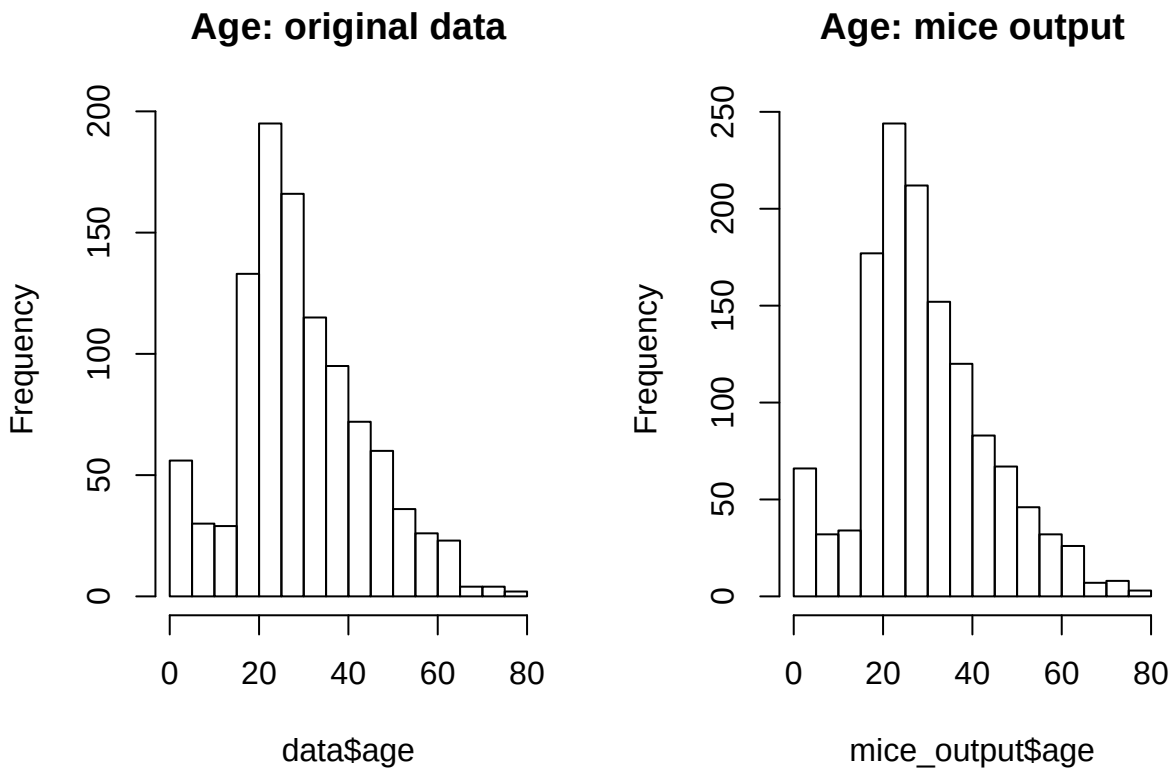
```
## Warning in data.matrix(x): NAs introduced by coercion
```

```
## Warning in data.matrix(x): NAs introduced by coercion
```

```
## Warning in data.matrix(x): NAs introduced by coercion
```

```
##     pclass survived sibsp parch fare age boat body  sex embarked home.dest
## 120      1        1     1     1    1   1    0    1    0        0         0
## 352      1        1     1     1    1   1    1    0    0        0         0
##   1      1        1     1     1    1   0    0    1    0        0         0
##  46      1        1     1     1    1   0    1    0    0        0         0
## 574      1        1     1     1    1   1    0    0    0        0         0
## 216      1        1     1     1    1   0    0    0    0        0         0
##          0        0     0     0    0 263  911 1188 1309     1309      1309
##
## 120    4
## 352    4
##   1    5
##  46    5
## 574    5
## 216    6
##     6289
```

```
mice_output <- complete(mice_mod)
#From the plots we can see that mice prediction is pretty good. So we can put mice predicted age to ori
par(mfrow=c(1,2))
hist(data$age,main = 'Age: original data')
hist(mice_output$age,main = 'Age: mice output')
```




```
data$age <- mice_output$age

#2 missing embarked
which(data$embarked=='')
```

```
## [1] 169 285
```

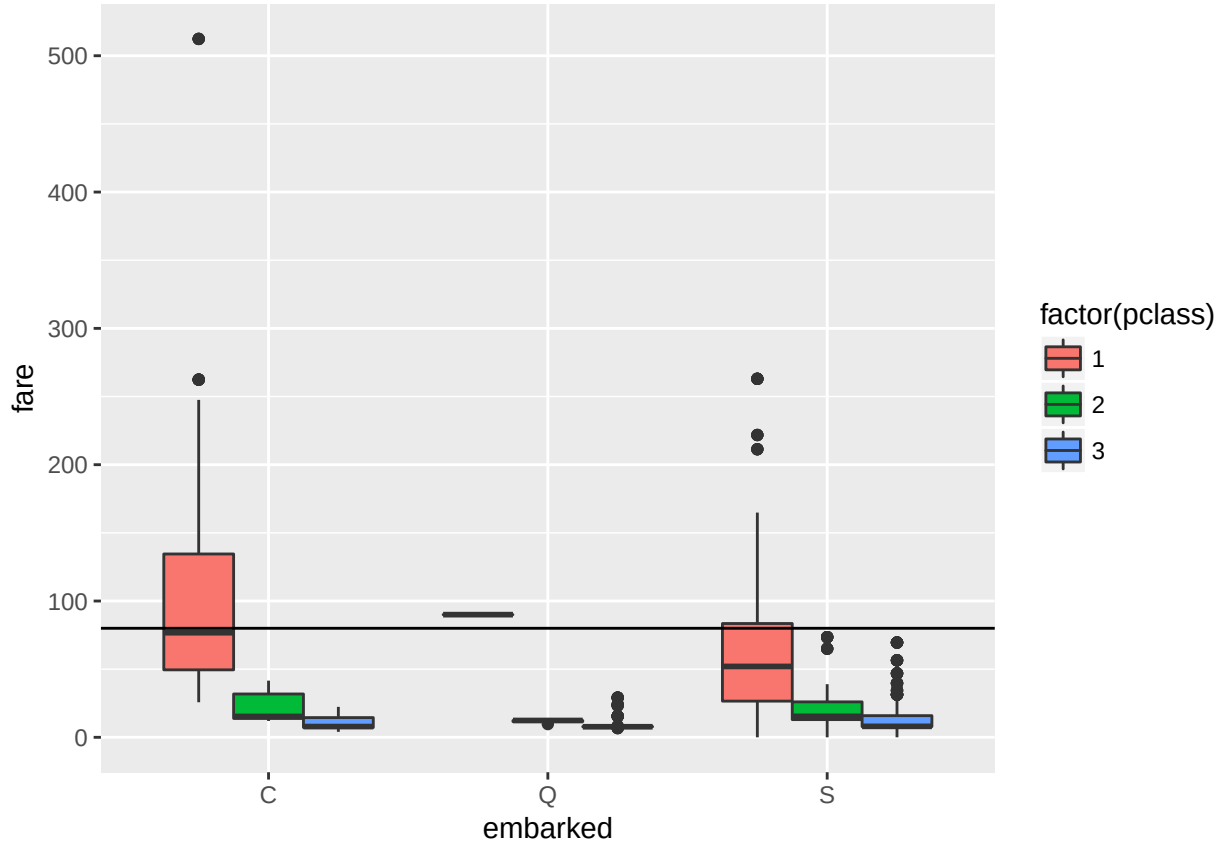
```
data[169]
```

```
##    pclass survived    sex age sibsp parch fare embarked boat body
## 1:      1        1 female  38     0     0   80             6   NA
##    home.dest
## 1:
```

```
data[285]
```

```
##    pclass survived    sex age sibsp parch fare embarked boat body
## 1:      1        1 female  62     0     0   80             6   NA
##         home.dest
## 1: Cincinatti, OH
```

```
#We notice that two passengers are from first class with the same ticket, they may be friends
#and embarked together. I will use pclass anf fare to predict embarked.
ggplot(data[c(-169,-285)],aes(x=embarked,y=fare,fill=factor(pclass)))+
  geom_boxplot()+
  geom_hline(aes(yintercept=80))
```




```
#From the graph we can see that first class with fare of $80 was most likely embarked at C.
data$embarked[c(169,285)] <- "C"

#Now we have no missing value.
sapply(data,function(x){sum(is.na(x)|x=="")})
```

```
##    pclass  survived       sex       age     sibsp     parch      fare
```

```
##         0         0         0         0         0         0         0
##  embarked      boat      body home.dest
##         0       823      1188       564
```

```
#Create new variable
data$familysize <- data$sibsp + data$parch +1

#Basic idea of overall data
par(mfrow=c(1,1))
library(evtree)
model <- factor(survived)~pclass+sex+age+familysize+embarked+fare
data <- as.data.frame(data)
fitEv   <-evtree(model,data=data)
```

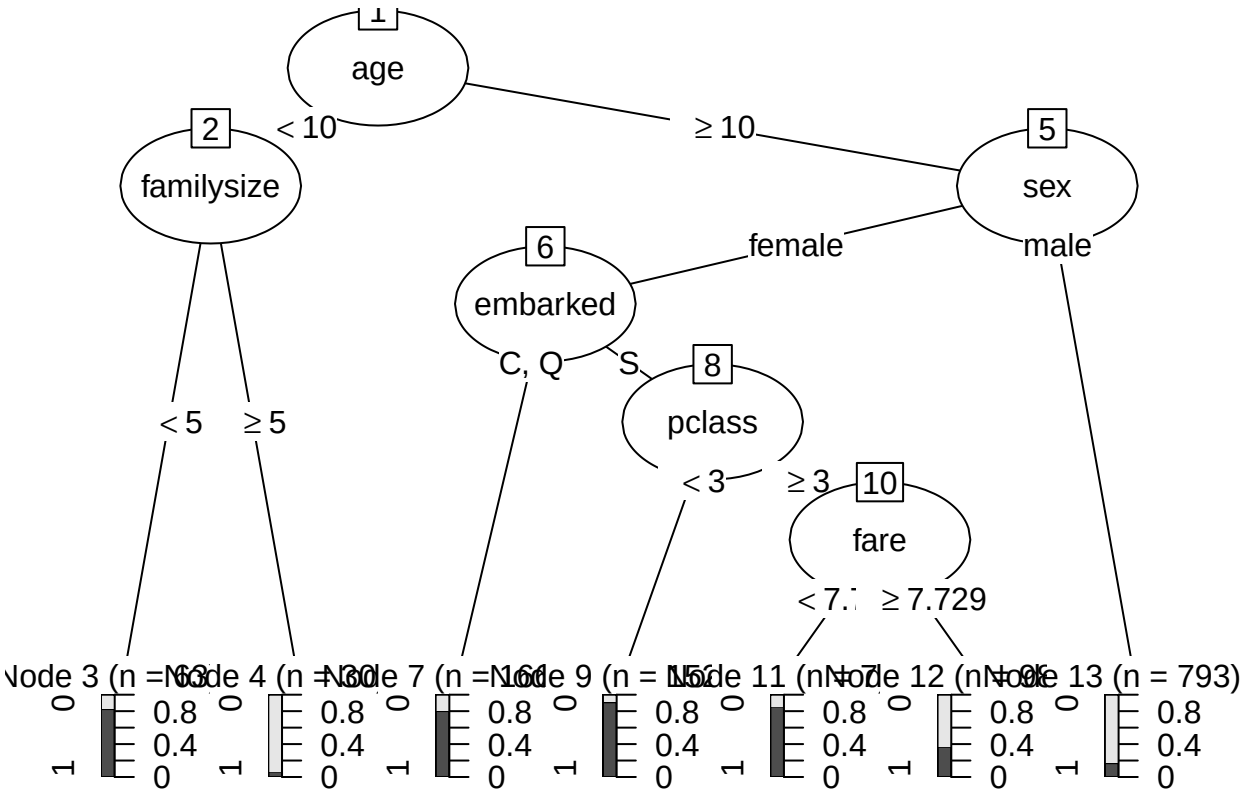
```
## Warning in evtree(model, data = data): character variable sex was converted
## to a factor
```

```
## Warning in evtree(model, data = data): character variable embarked was
## converted to a factor
```

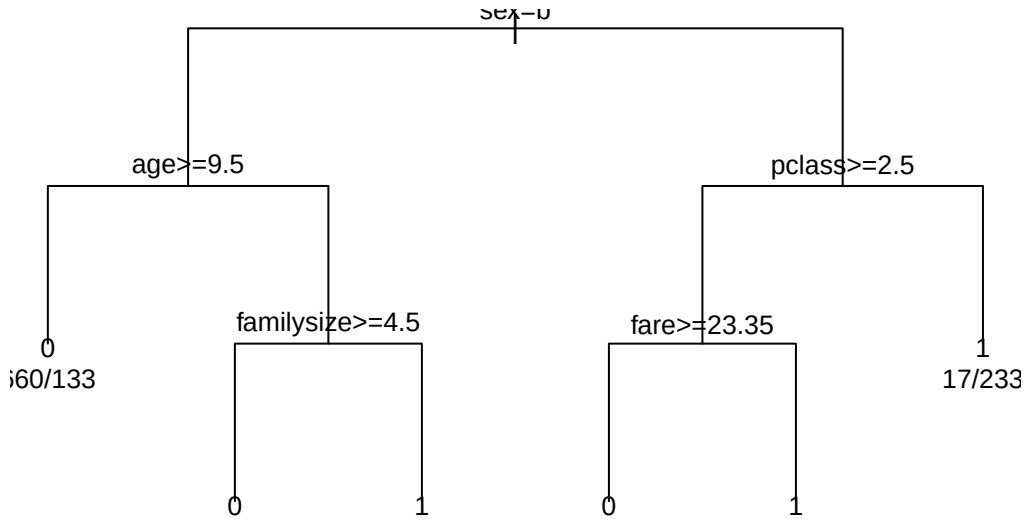
```
fitEv
```

```
##
## Model formula:
## factor(survived) ~ pclass + sex + age + familysize + embarked +
##     fare
##
## Fitted party:
## [1] root
## |   [2] age < 10
## |   |   [3] familysize < 5: 1 (n = 63, err = 17.5%)
## |   |   [4] familysize >= 5: 0 (n = 30, err = 6.7%)
## |   [5] age >= 10
## |   |   [6] sex in female
## |   |   |   [7] embarked in C, Q: 1 (n = 166, err = 19.9%)
## |   |   |   [8] embarked in S
## |   |   |   |   [9] pclass < 3: 1 (n = 152, err = 9.2%)
## |   |   |   |   [10] pclass >= 3
## |   |   |   |   |   [11] fare < 7.7292: 1 (n = 7, err = 14.3%)
## |   |   |   |   |   [12] fare >= 7.7292: 0 (n = 98, err = 36.7%)
## |   |   [13] sex in male: 0 (n = 793, err = 16.8%)
##
## Number of inner nodes:    6
## Number of terminal nodes: 7
```

```
plot(fitEv)
```

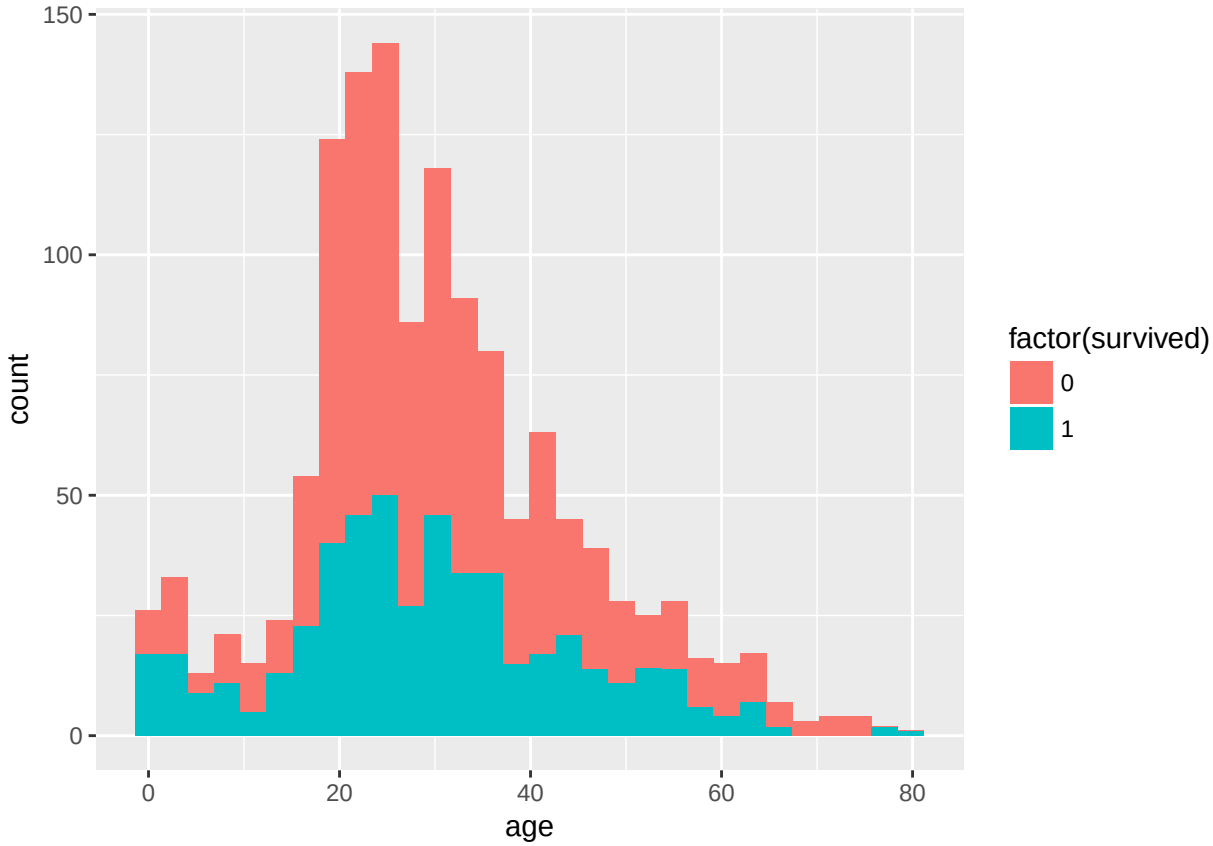



```
fitRpart <- rpart(model,method="class",data=data)
plot(fitRpart, uniform=TRUE,main="Regression Tree (rpart)")
text(fitRpart, use.n=TRUE, cex=0.8)
```

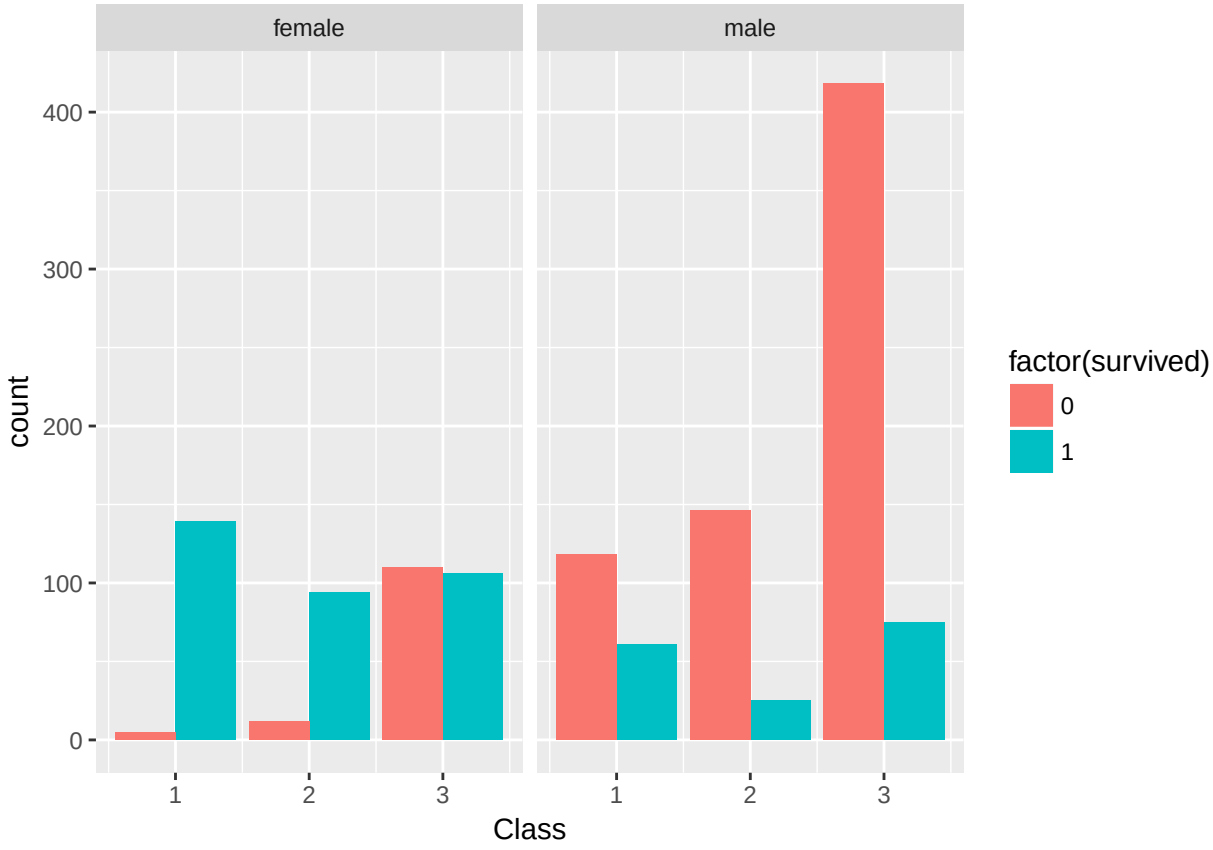



```
ggplot(data,aes(x=age,fill=factor(survived)))+
  geom_histogram()
```

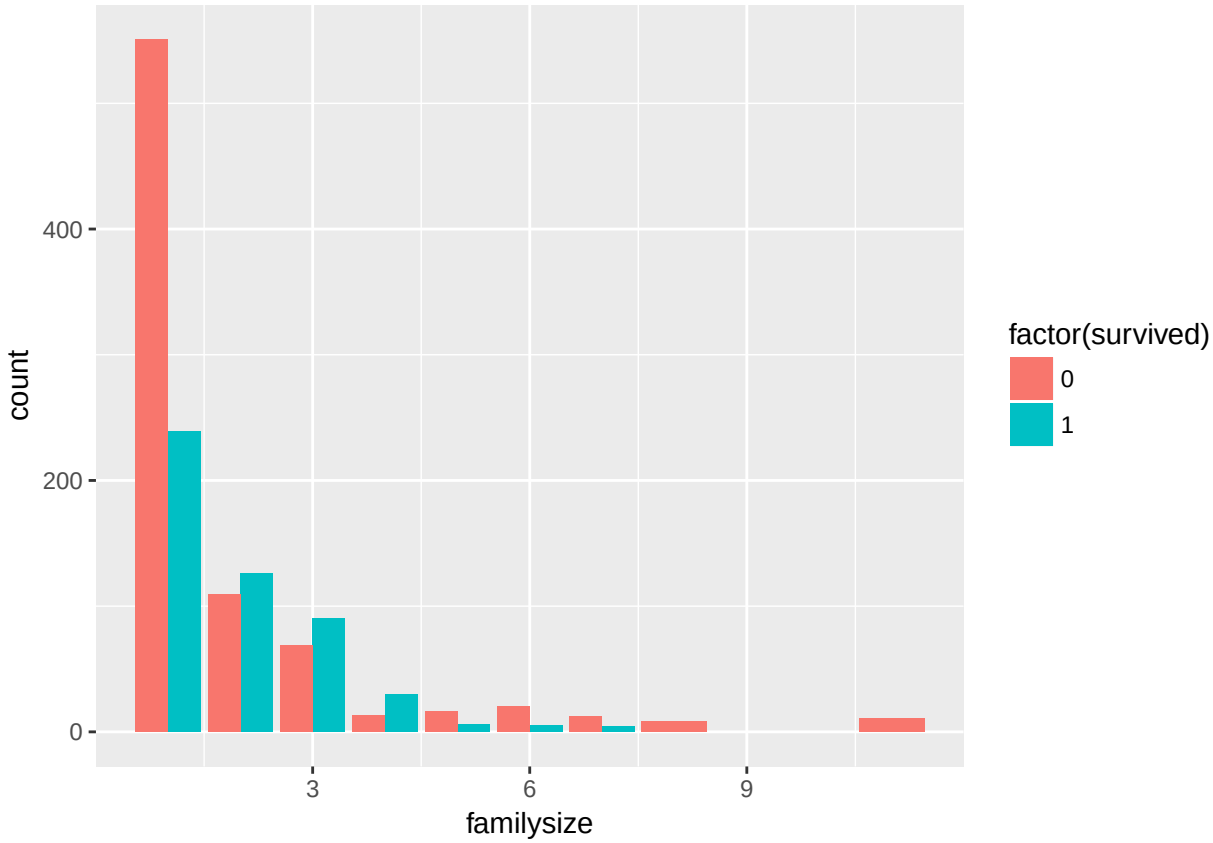
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(data,aes(x=pclass,fill=factor(survived)))+
  geom_bar(stat = 'count',position = position_dodge())+
  labs(x="Class")+
  facet_grid(.~sex)
```

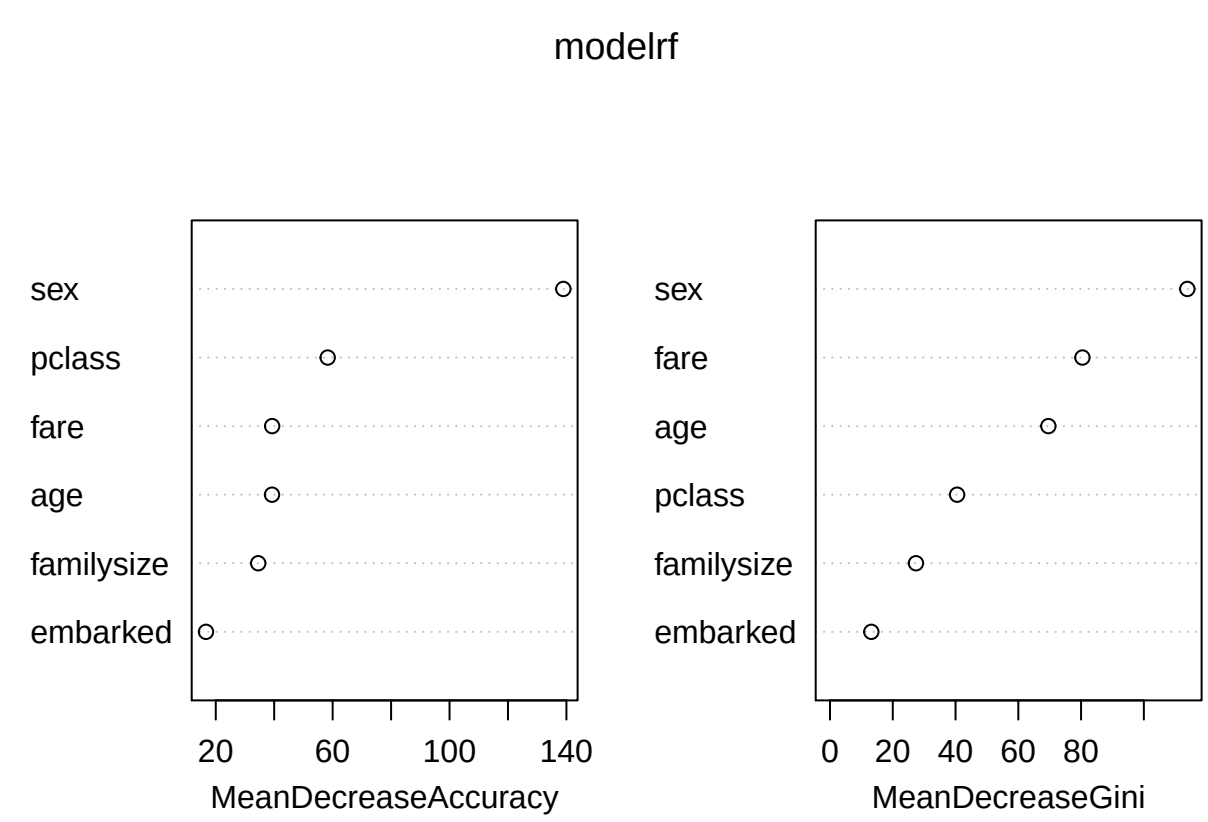



```
ggplot(data,aes(x=familysize,fill=factor(survived)))+
  geom_bar(stat='count',position=position_dodge())
```




```
#data partition
data$survived <- as.factor(data$survived)
data$sex <- as.factor(data$sex)
data$embarked <- as.factor(data$embarked)
set.seed(1234)
ind <- sample(2,nrow(data),replace = T,prob = c(0.8,0.2))
train <- data[ind==1,]
test <- data[ind==2,]

#use random forest model
modelrf <- randomForest(survived~pclass+sex+age+familysize+embarked+fare,data=train,importance=TRUE,ntr
#variables impotance.
varImpPlot(modelrf)
```




```
#prediction
#compare the results of predicted survived with original survived in train data.
p1 <- predict(modelrf,train)
table(train$survived,p1)
```

```
##    p1
##       0   1
##   0 632  16
##   1  85 311
```

```
#the incorrect prediction is about 10% of data
mean(p1!=train$survived)
```

```
## [1] 0.0967433
```

```
#use the model for test data.
p2 <- predict(modelrf,test)
table(test$survived,p2)
```

```
##    p2
##       0   1
##   0 147  14
##   1  41  63
```

```
#it incorrectly predict 18% of data
mean(p2!=test$survived)
```

```
## [1] 0.2075472
```