Problem Set 6

Donggyun Kim
27008257

5/4/2018

Exploratory Data Analysis

```r
training <- read.csv("data/training.csv", row.names = 1)
test <- read.csv("data/test.csv", row.names = 1)
test <- test[, -1]

# structure
str(training)
```

```
## 'data.frame':    30155 obs. of  13 variables:
##  $ income        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age           : int  39 50 38 53 28 37 49 23 32 34 ...
##  $ capital_gain  : Factor w/ 3 levels "High","Low","None": 2 3 3 3 3 3 3 3 3 3 ...
##  $ capital_loss  : Factor w/ 3 levels "High","Low","None": 3 3 3 3 3 3 3 3 3 3 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 30 50 45 ...
##  $ workclass     : Factor w/ 5 levels " Federal-gov",..: 3 5 4 4 4 4 4 4 4 4 ...
##  $ education     : Factor w/ 7 levels " Associates",..: 2 2 5 4 2 6 4 2 1 4 ...
##  $ marital.status: Factor w/ 4 levels "Married","Never-Married",..: 2 1 3 1 1 1 3 2 2 1 ...
##  $ occupation    : Factor w/ 6 levels " Administration",..: 1 4 2 2 3 4 6 1 5 2 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 4 2 1 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian",..: 5 5 5 3 3 5 3 5 3 1 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 2 1 1 1 1 2 2 ...
##  $ native.country: Factor w/ 11 levels "Asia-Developed",..: 8 8 8 8 7 8 6 8 8 6 ...
```

```r
str(test)
```

```
## 'data.frame':    15060 obs. of  13 variables:
##  $ income        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age           : num  -1.029 -0.0574 -0.3564 -1.1037 1.2131 ...
##  $ capital_gain  : Factor w/ 3 levels "High","Low","None": 3 3 3 3 3 3 3 3 3 3 ...
##  $ capital_loss  : Factor w/ 3 levels "High","Low","None": 3 3 3 3 3 3 3 3 3 3 ...
##  $ hours_per_week: num  -0.0789 0.7501 -0.9079 -0.0789 -2.5659 ...
##  $ workclass     : Factor w/ 5 levels " Federal-gov",..: 4 4 4 4 4 1 4 3 4 4 ...
##  $ education     : Factor w/ 7 levels " Associates",..: 4 5 4 5 4 2 5 5 5 5 ...
##  $ marital.status: Factor w/ 4 levels "Married","Never-Married",..: 2 1 2 2 1 1 2 2 1 4 ...
##  $ occupation    : Factor w/ 6 levels " Administration",..: 3 2 6 6 2 1 1 6 1 3 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 4 1 2 5 1 1 2 4 6 5 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian",..: 3 5 5 5 5 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 1 2 2 1 2 1 1 ...
##  $ native.country: Factor w/ 11 levels "Asia-Developed",..: 8 8 8 8 8 8 8 8 8 8 ...
```

```r
# summary statistics
summary(training)
```

```
##      income             age         capital_gain capital_loss
##  Min.   :0.0000   Min.   :17.00   High: 1090   High:  686
```

```
##  1st Qu.:0.0000   1st Qu.:28.00   Low : 1448   Low :  734  
##  Median :0.0000   Median :37.00   None:27617   None:28735  
##  Mean   :0.2489   Mean   :38.43                            
##  3rd Qu.:0.0000   3rd Qu.:47.00                            
##  Max.   :1.0000   Max.   :90.00                            
##                                                            
##  hours_per_week          workclass           education    
##  Min.   : 1.00   Federal-gov  :  942   Associates : 2315  
##  1st Qu.:40.00   Not-Working  :   14   Bachelors  : 5044  
##  Median :40.00   Other-gov    : 3345   Doctorate  :  374  
##  Mean   :40.93   Private      :22281   Dropout    : 3739  
##  3rd Qu.:45.00   Self-Employed: 3573   HS-Graduate:16514  
##  Max.   :99.00                         Masters    : 1627  
##                                        Prof-School:  542  
##       marital.status            occupation          relationship  
##  Married      :14086   Administration:3720   Husband       :12463  
##  Never-Married: 9725   Blue-Collar   :9906   Not-in-family : 7724  
##  Not-Married  : 5518   High-Service  :4035   Other-relative:  888  
##  Widowed      :  826   Management    :3991   Own-child     : 4465  
##                        Sales         :3584   Unmarried     : 3209  
##                        Service       :4919   Wife          : 1406  
##                                                                   
##           race           sex                        native.country 
##  Amer-Indian:  286   Female: 9776   United-States          :27497  
##  Asian      :  895   Male  :20379   South-America-Emerging:  970  
##  Black      : 2817                  Western-Developed      :  466  
##  Other      :  231                  Asia-Emerging          :  273  
##  White      :25926                  South-America-Frontier:  242  
##                                     Asia-Frontier          :  199  
##                                     (Other)                :  508  
```

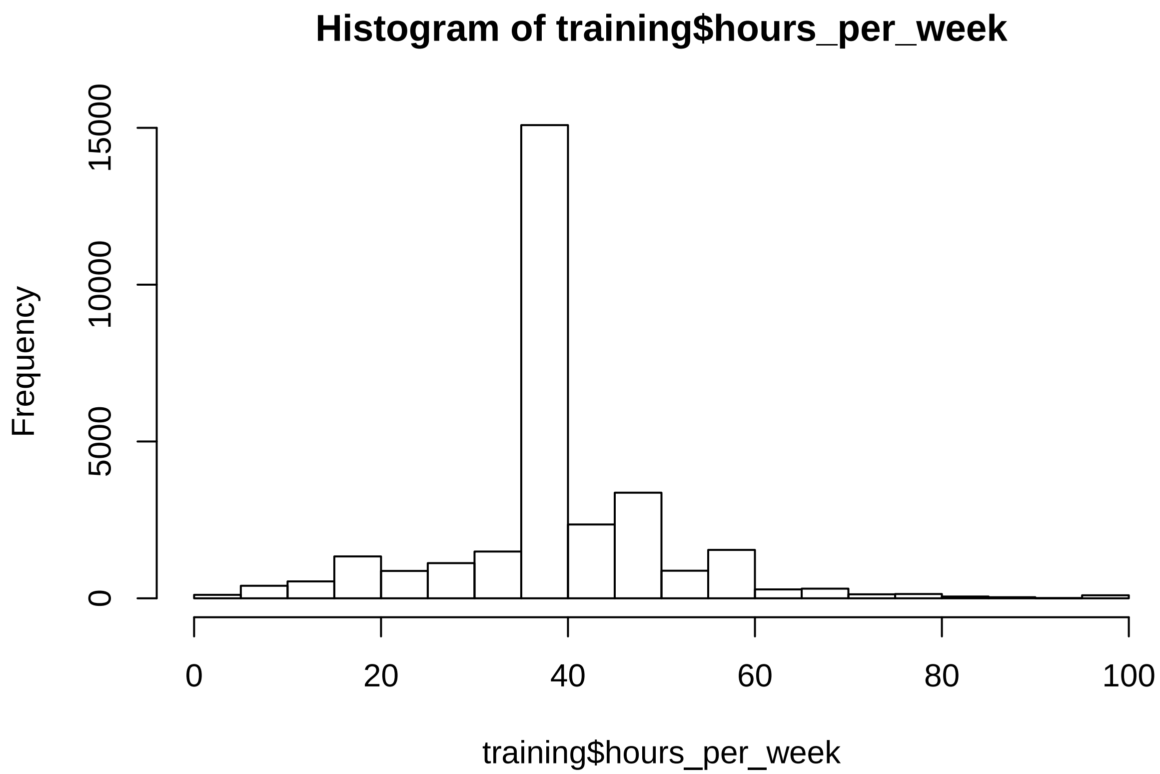
```r
summary(test)
```

```
##      income            age           capital_gain capital_loss
##  Min.   :0.0000   Min.   :-1.6268   High:  519   High:  264  
##  1st Qu.:0.0000   1st Qu.:-0.8048   Low :  733   Low :  449  
##  Median :0.0000   Median :-0.1322   None:13808   None:14347  
##  Mean   :0.2457   Mean   : 0.0000                            
##  3rd Qu.:0.0000   3rd Qu.: 0.6899                            
##  Max.   :1.0000   Max.   : 3.8288                            
##                                                              
##  hours_per_week           workclass           education   
##  Min.   :-3.31196   Federal-gov  :  463   Associates :1151  
##  1st Qu.:-0.07889   Not-Working  :    7   Bachelors  :2526  
##  Median :-0.07889   Other-gov    : 1700   Doctorate  : 169  
##  Mean   : 0.00000   Private      :11021   Dropout    :1920  
##  3rd Qu.: 0.33561   Self-Employed: 1869   HS-Graduate:8164  
##  Max.   : 4.81217                         Masters    : 887  
##                                           Prof-School: 243  
##       marital.status           occupation          relationship 
##  Married      :7001   Administration:1819   Husband       :6203  
##  Never-Married:4872   Blue-Collar   :3225   Not-in-family :3976  
##  Not-Married  :2737   High-Service  :3063   Other-relative: 460  
##  Widowed      : 450   Management    :1992   Own-child     :2160  
##                       Sales         :1824   Unmarried     :1576  
```

```
##                         Service       :3137    Wife          : 685
##
##            race            sex                        native.country
##  Amer-Indian:  149   Female: 4913   United-States           :13788
##  Asian      :  408   Male  :10147   South-America-Emerging:  450
##  Black      : 1411                  Western-Developed     :  224
##  Other      :  122                  Asia-Emerging         :  152
##  White      :12970                  Western-Emerging      :  114
##                                     South-America-Frontier:  113
##                                     (Other)               :  219
```

```r
# histograms
hist(training$age)
```


```r
hist(training$hours_per_week)
```




```r
hist(training$income)
```

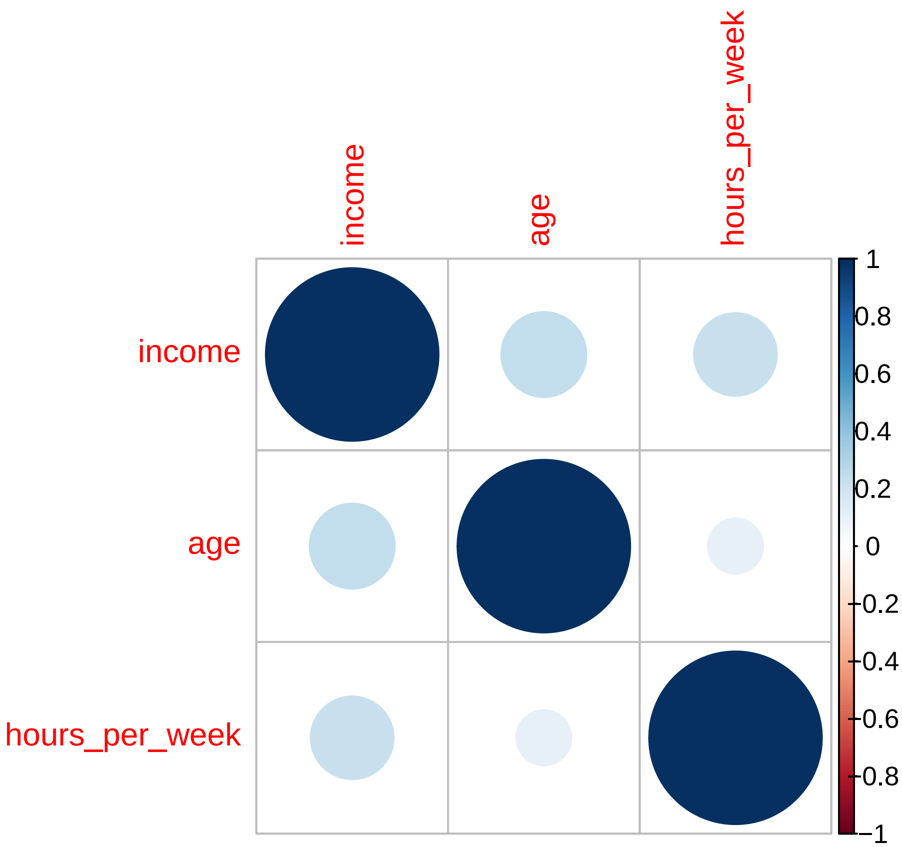

```r
# proportion table
names_ct <- names(training)[c(-1, -2, -5)]  # categorical variable names
list_ct <- as.list(names_ct)
```

```r
names(list_ct) <- names_ct
sapply(list_ct, function (x) {
  round(prop.table(table(training[x])) * 100, 3)
})
```

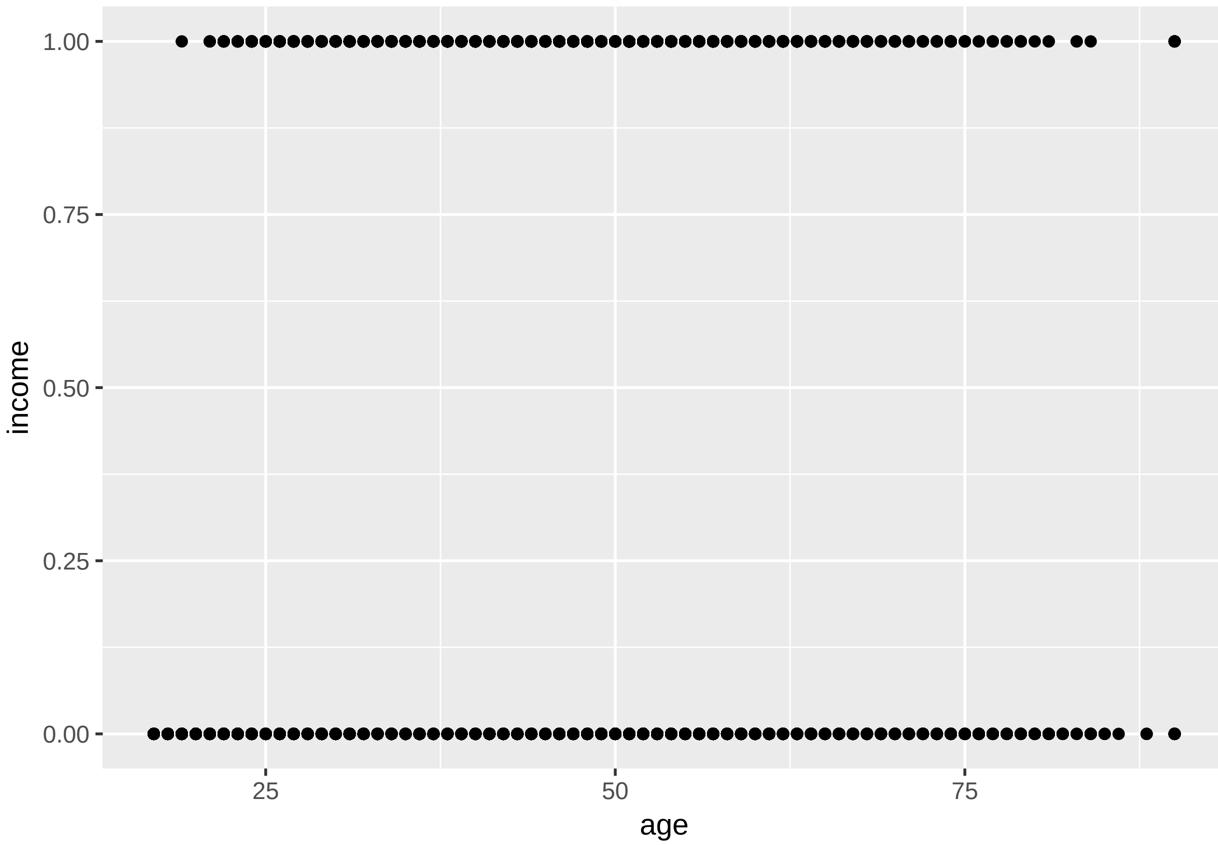
```
## $capital_gain
##
##   High    Low   None
##  3.615  4.802 91.583
##
## $capital_loss
##
##   High    Low   None
##  2.275  2.434 95.291
##
## $workclass
##
##    Federal-gov    Not-Working      Other-gov        Private  Self-Employed
##          3.124          0.046         11.093         73.888         11.849
##
## $education
##
##   Associates    Bachelors    Doctorate      Dropout  HS-Graduate
##        7.677       16.727        1.240       12.399       54.764
##      Masters  Prof-School
##        5.395        1.797
##
## $marital.status
##
##       Married Never-Married   Not-Married       Widowed
##        46.712        32.250        18.299         2.739
##
## $occupation
##
##  Administration     Blue-Collar    High-Service      Management
##          12.336          32.850          13.381          13.235
##           Sales         Service
##          11.885          16.312
##
## $relationship
##
##         Husband   Not-in-family  Other-relative       Own-child
##          41.330          25.614           2.945          14.807
##       Unmarried            Wife
##          10.642           4.663
##
## $race
##
## Amer-Indian       Asian       Black       Other       White
##       0.948       2.968       9.342       0.766      85.976
##
## $sex
##
##  Female    Male
```

```
##   32.419   67.581
##
## $native.country
##
##          Asia-Developed          Asia-Emerging          Asia-Frontier
##                   0.398                  0.905                  0.660
##                   Other South-America-Developed  South-America-Emerging
##                   0.139                  0.060                  3.217
##   South-America-Frontier          United-States      Western-Developed
##                   0.803                 91.186                  1.545
##        Western-Emerging       Western-Frontier
##                   0.570                  0.517
```

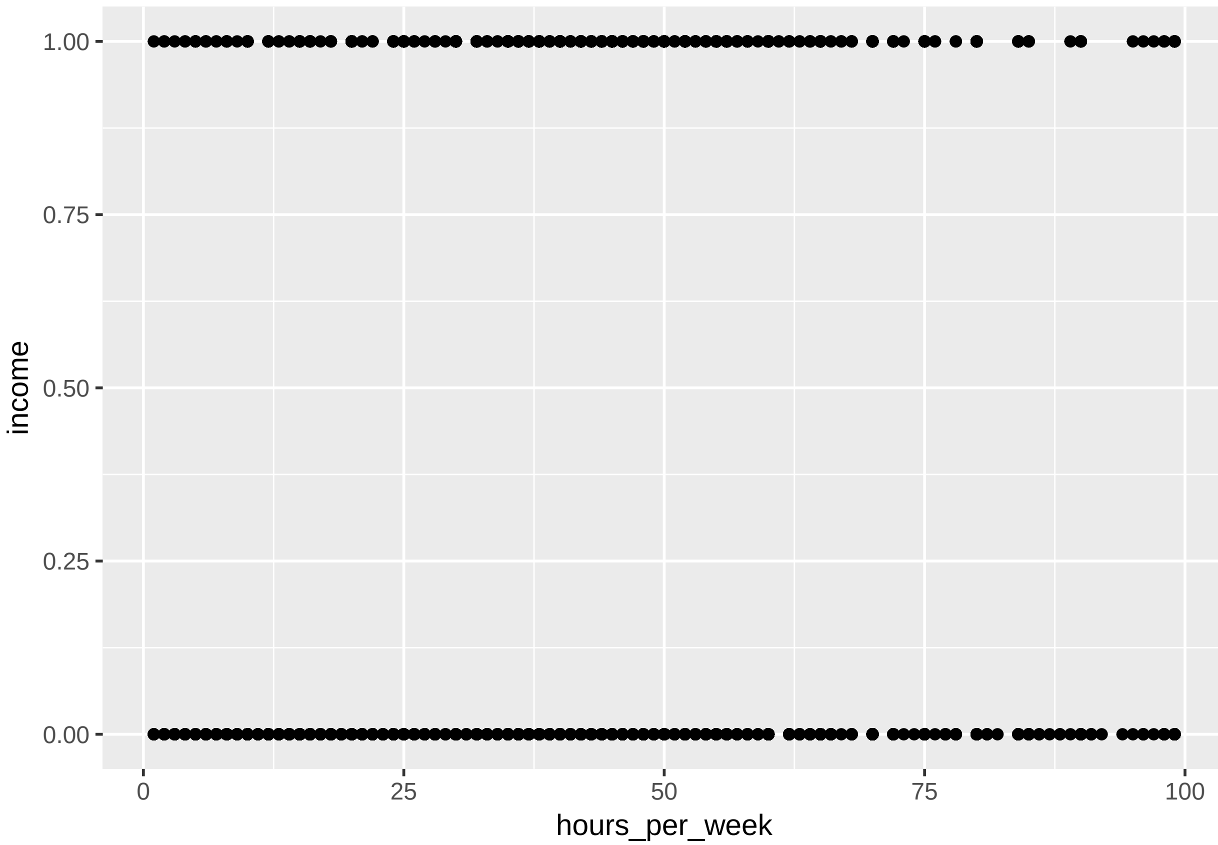
```r
# correlation matrix plot
library(corrplot)
corrplot(cor(training[, c(1, 2, 5)]), method = 'circle')

# barplots and scatterplots
library(ggplot2)
```




```r
ggplot(training, aes(x = age, y = income)) +
  geom_point()
```

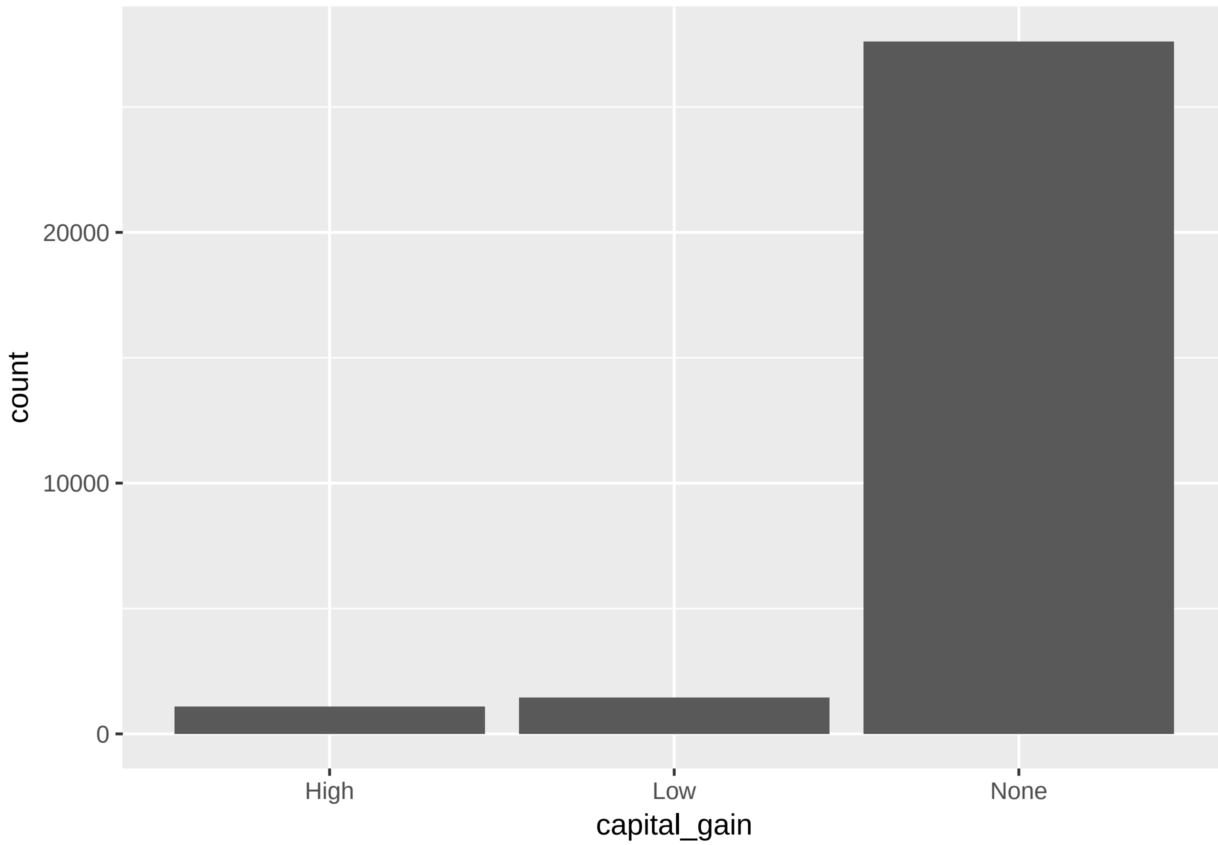



```r
ggplot(training, aes(x = hours_per_week, y = income)) +
  geom_point()
```

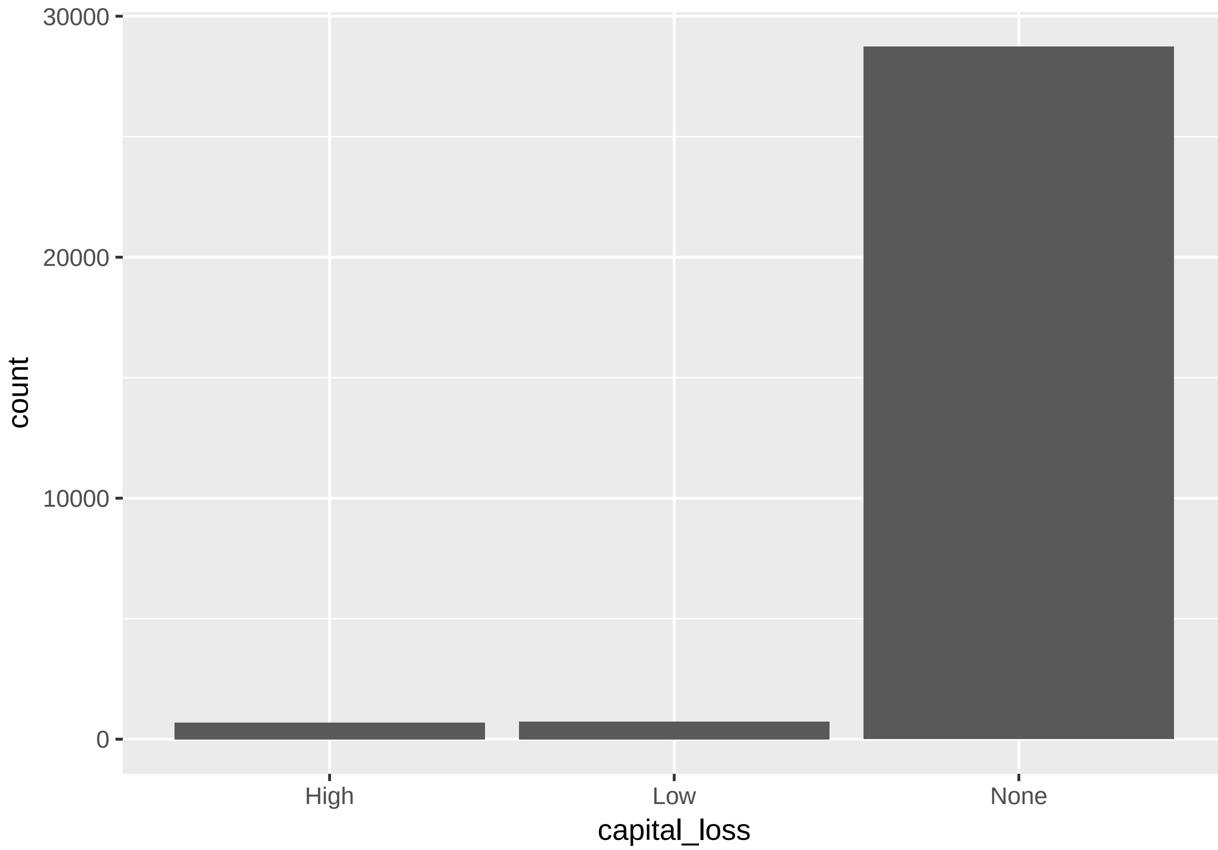



```r
lapply(names_ct, function (x) {
  ggplot(training, aes_string(x)) +
    geom_bar()
})
```

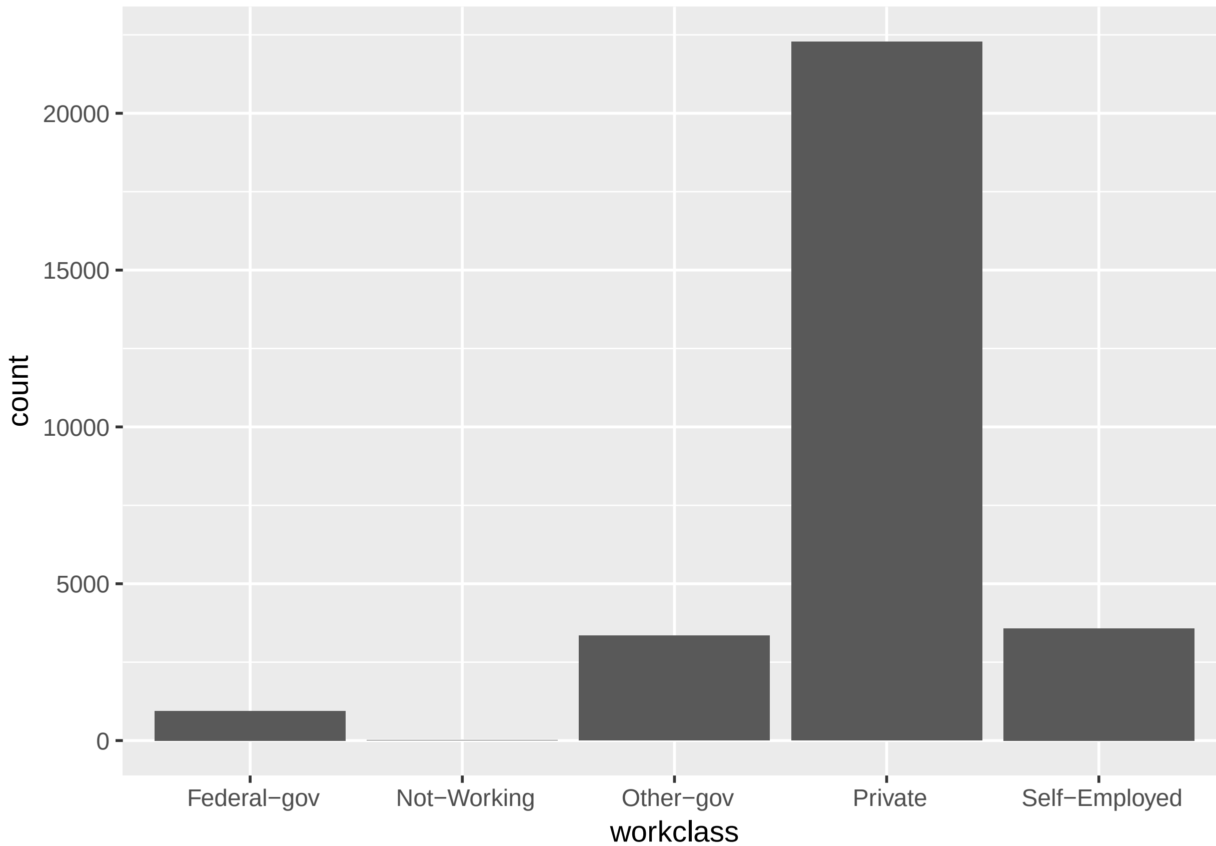
```
## [[1]]
```



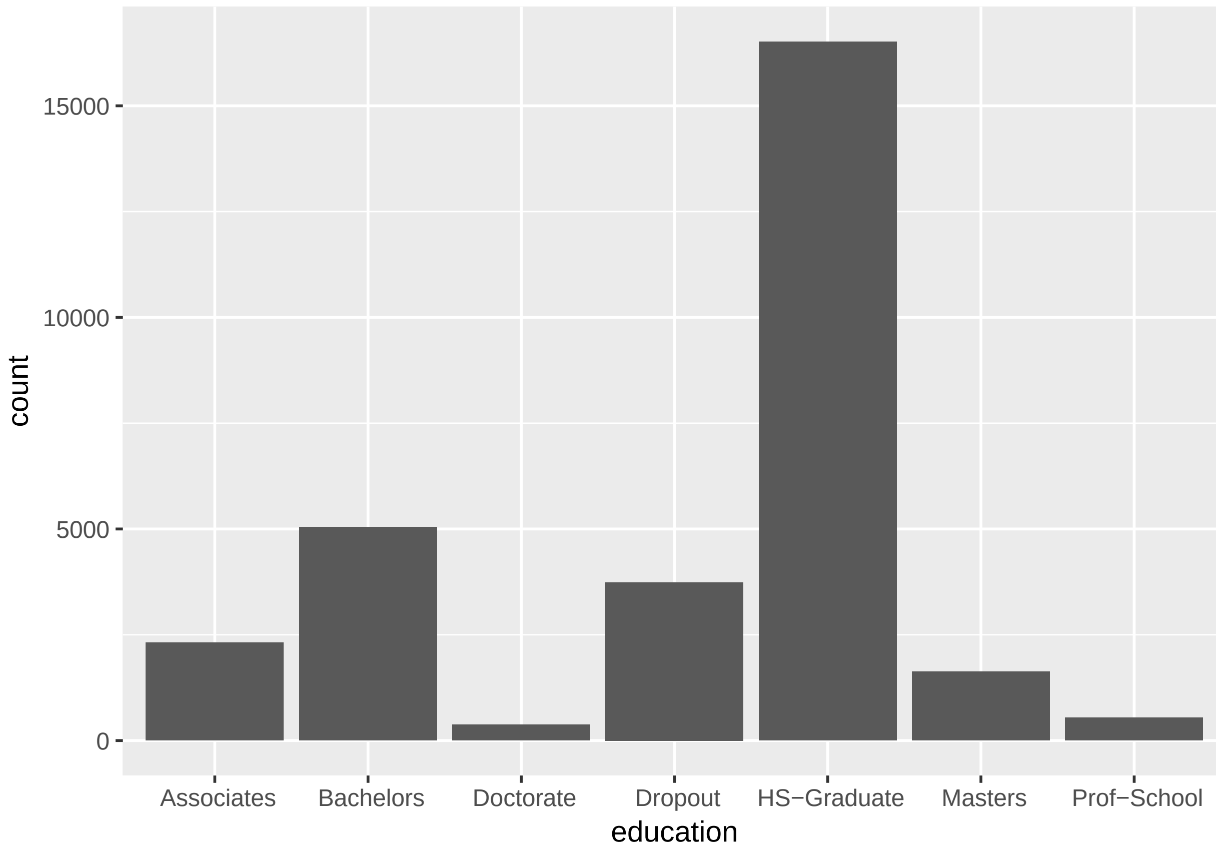

```
##
## [[2]]
```



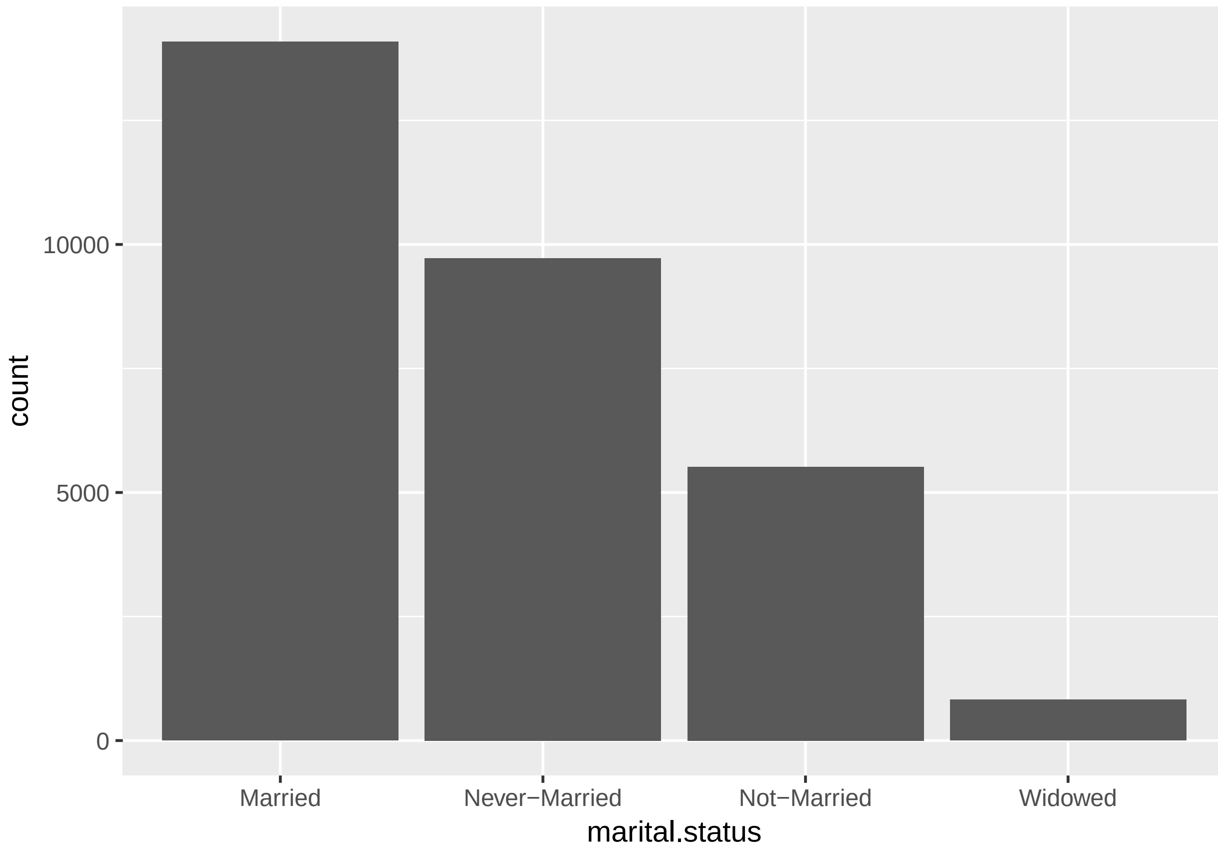

```
##
## [[3]]
```


20000
15000
10000
5000
0
count
Federal−gov
Not−Working
Other−gov
Private
Self−Employed
workclass

```
##
## [[4]]
```

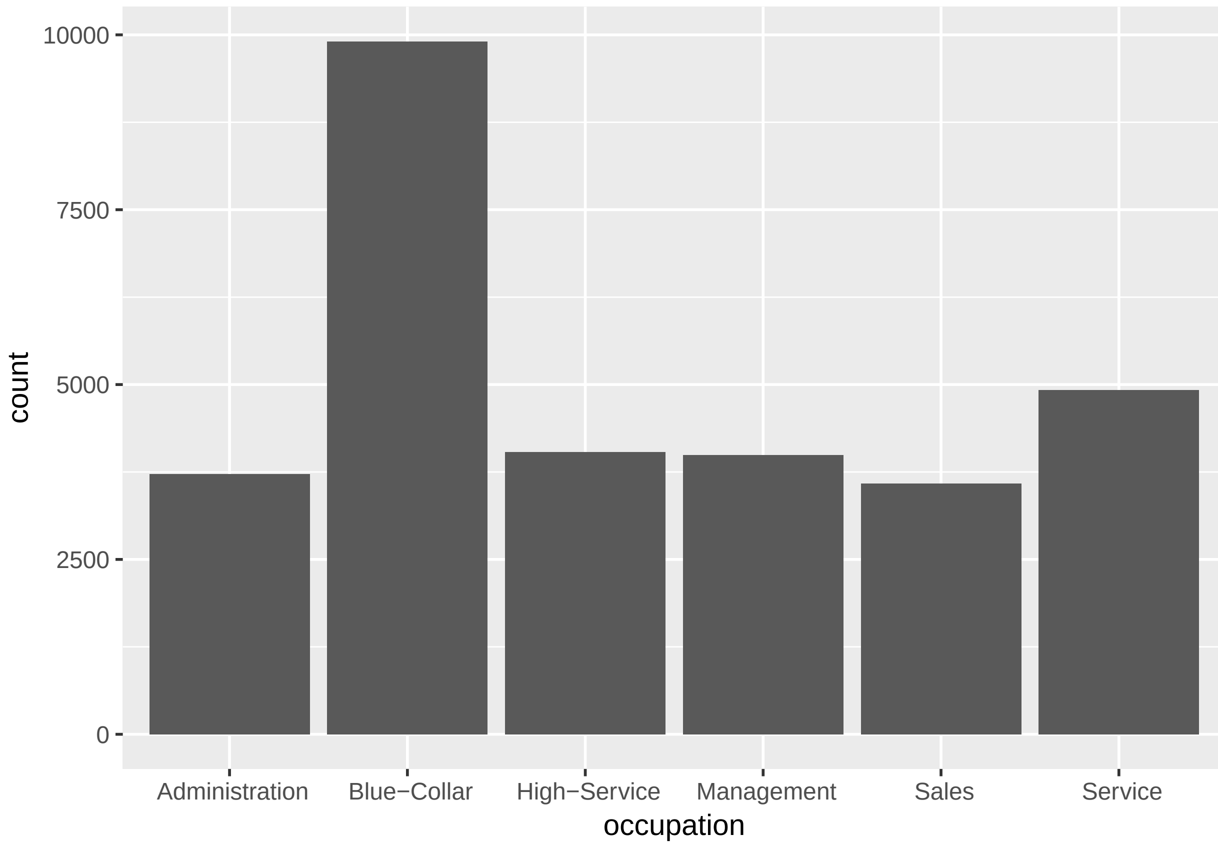

15000
10000
5000
0
count
Associates
Bachelors
Doctorate
Dropout
HS−Graduate
Masters
Prof−School
education

```
##
## [[5]]
```

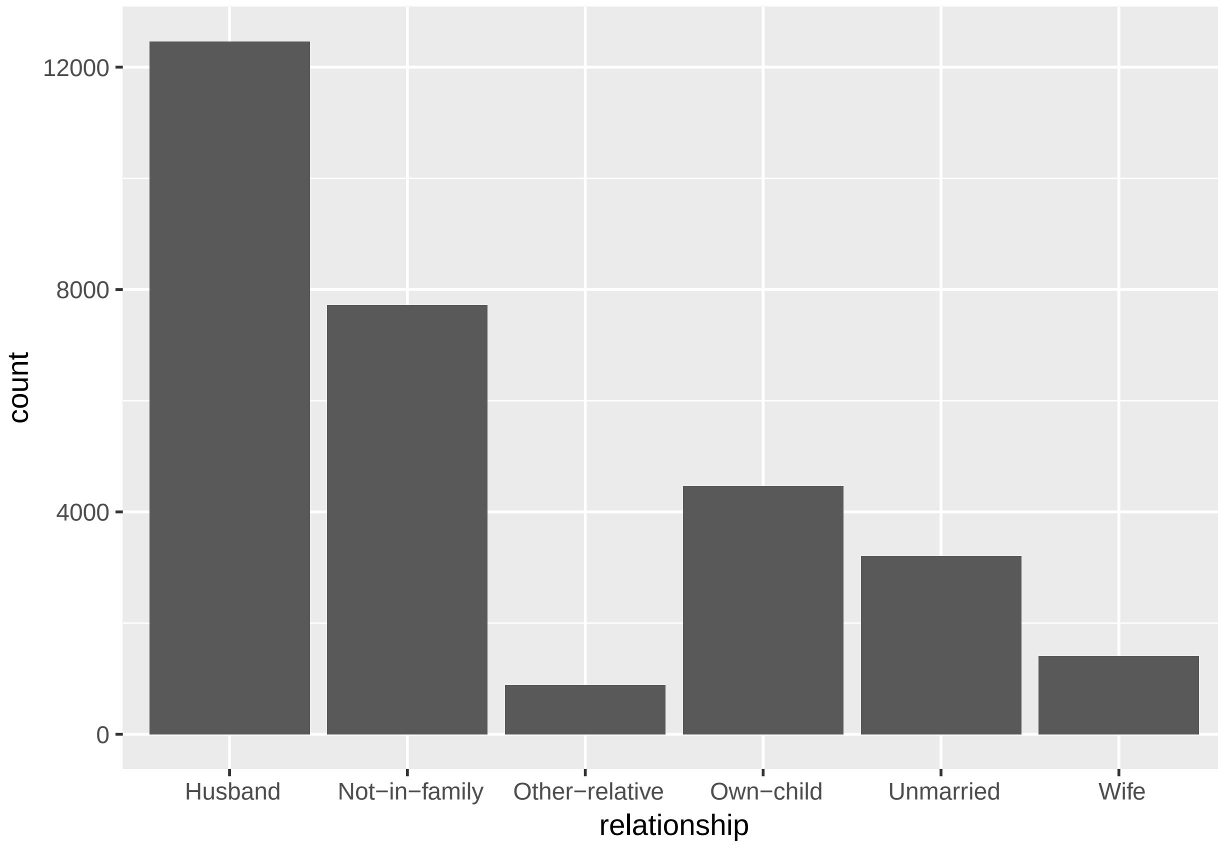


count
10000
5000
0
Married
Never−Married
Not−Married
Widowed
marital.status


```
##
## [[6]]
```

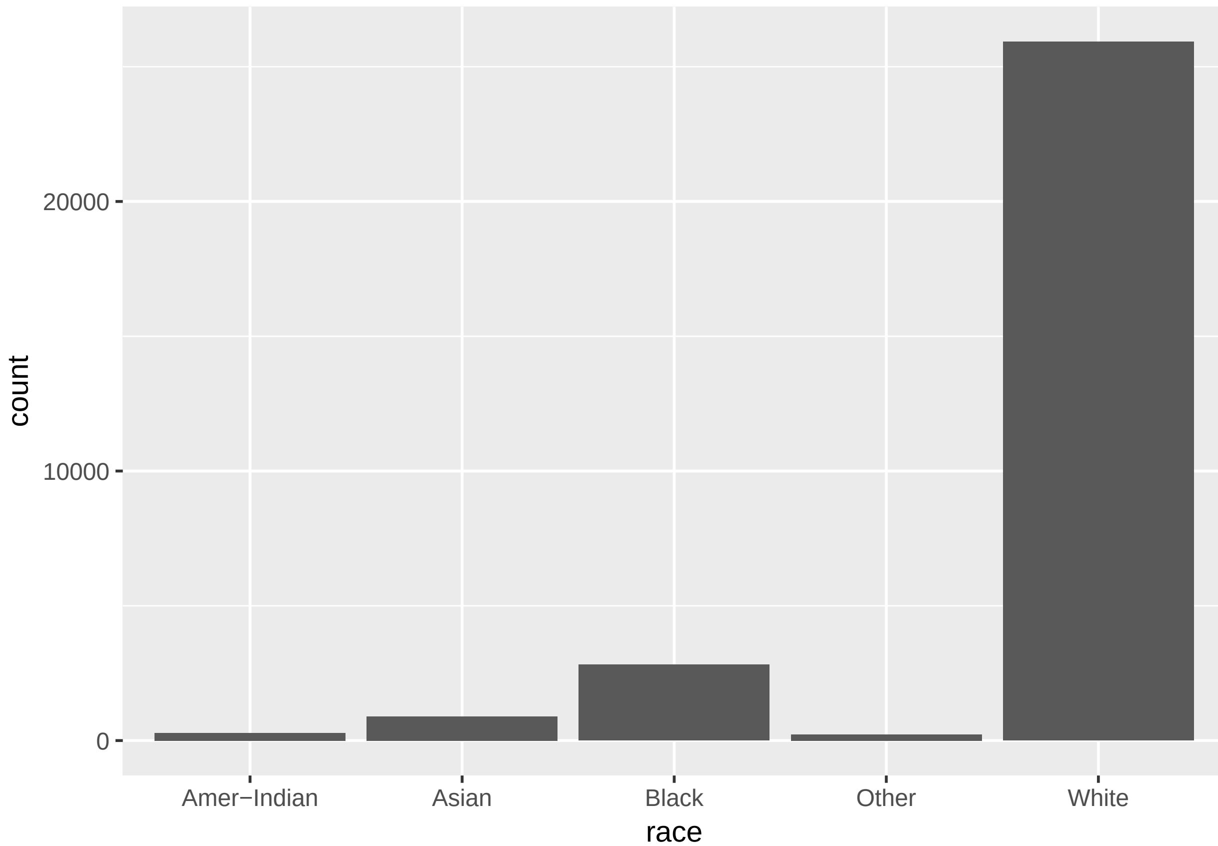

10000
7500
5000
2500
0
count
Administration
Blue−Collar
High−Service
Management
Sales
Service
occupation

```
##
## [[7]]
```

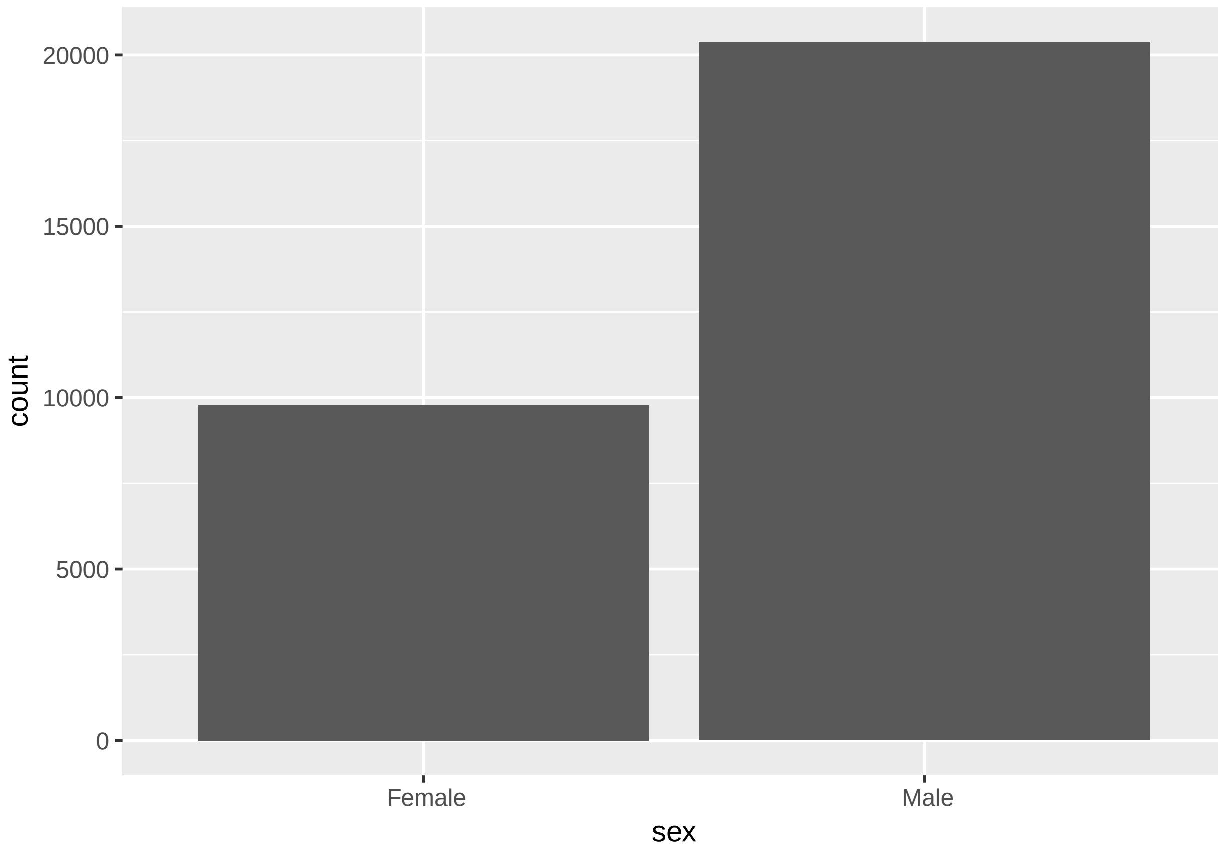


12000
8000
4000
0
count
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife
relationship

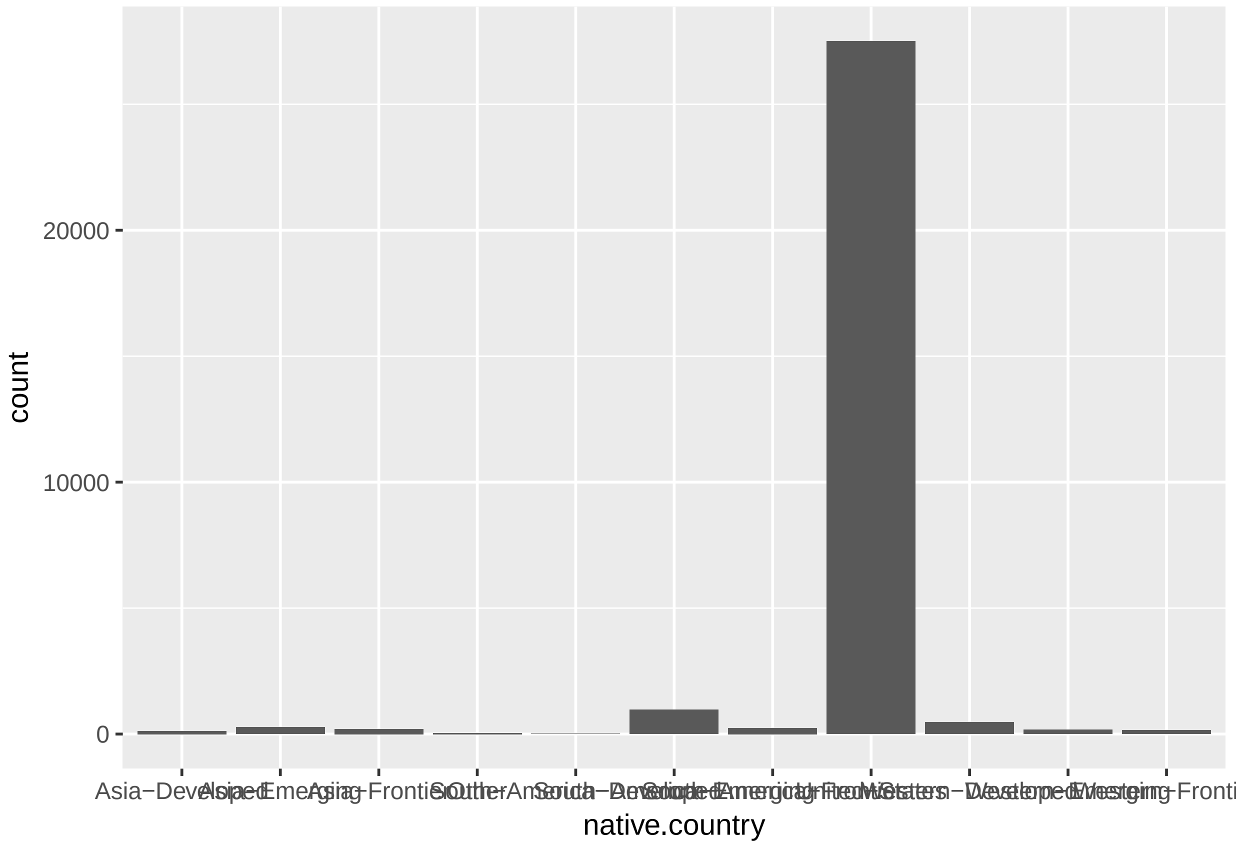

```
##
## [[8]]
```


20000
10000
0
count
Amer−Indian
Asian
Black
Other
White
race

```
##
## [[9]]
```




```
##
## [[10]]
```

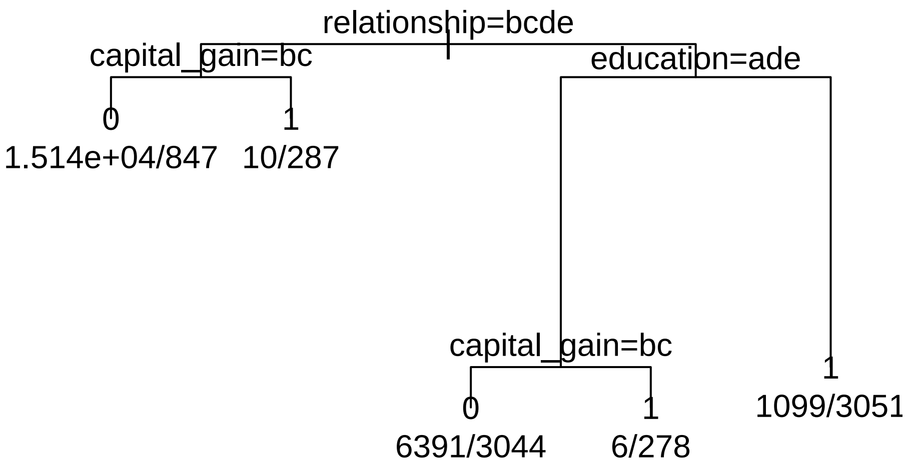


Build a Classification Tree

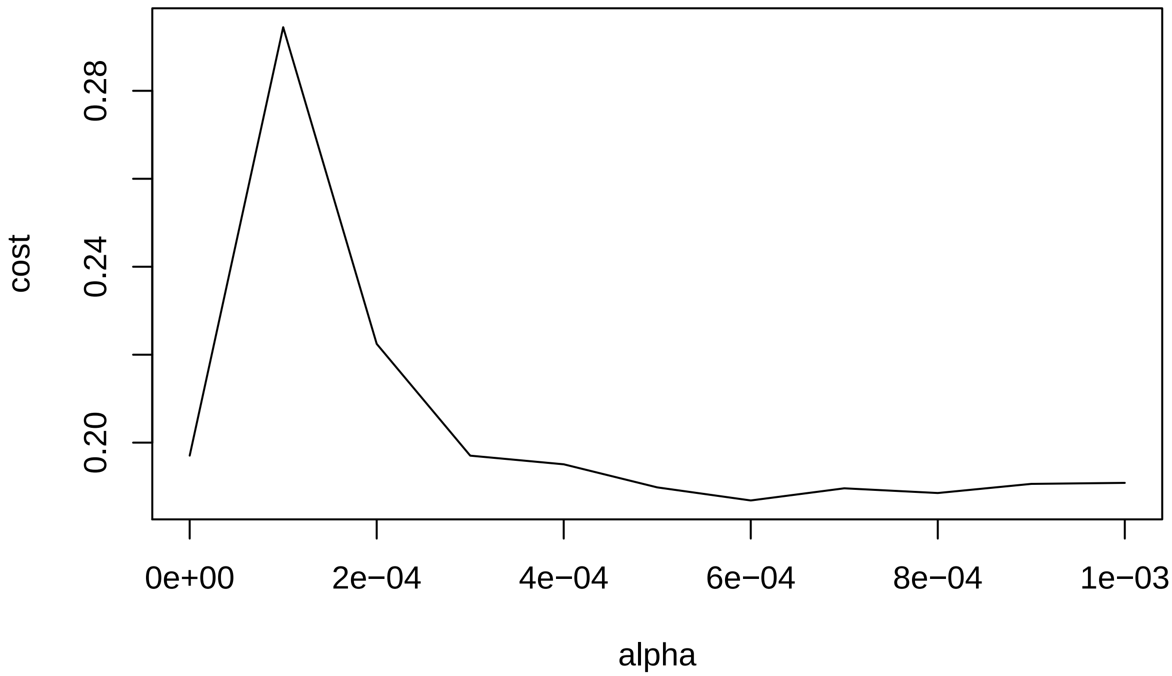
```
library(rpart)
library(caret)
library(pROC)

# fit a tree with minsplit = 20 and cp = 0.01
tree_training <- rpart(income ~ ., training, method = "class",
                        control = rpart.control(minsplit = 20, cp = 0.01))
tree_training
```

```
## n= 30155
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 30155 7507 0 (0.75105289 0.24894711)
##    2) relationship= Not-in-family, Other-relative, Own-child, Unmarried 16286 1134 0 (0.93036964 0.06
##      4) capital_gain=Low,None 15989  847 0 (0.94702608 0.05297392) *
##      5) capital_gain=High 297   10 1 (0.03367003 0.96632997) *
##    3) relationship= Husband, Wife 13869 6373 0 (0.54048598 0.45951402)
##      6) education= Associates, Dropout, HS-Graduate 9719 3322 0 (0.65819529 0.34180471)
##       12) capital_gain=Low,None 9435 3044 0 (0.67737149 0.32262851) *
##       13) capital_gain=High 284    6 1 (0.02112676 0.97887324) *
##      7) education= Bachelors, Doctorate, Masters, Prof-School 4150 1099 1 (0.26481928 0.73518072) *
```

```r
plot(tree_training, margin = 0.15)
text(tree_training, use.n = TRUE)
```




```r
alpha <- seq(from = 0, to = 0.001, by = 0.0001)
set.seed(1991)
folds <- createFolds(1:nrow(training))
cost_mat <- matrix(0, nrow = length(alpha), ncol = length(folds))
rownames(cost_mat) <- paste("alpha", seq_along(alpha))
colnames(cost_mat) <- paste("folds", seq_along(folds))
# 10-fold cross validation
for (k in seq_along(folds)) {
  for (i in seq_along(alpha)) {
    # fit the largest tree with minsplit = 0 and cp = 0
    tree_obj <- rpart(income ~ ., training[-folds[[k]], ], method = "class",
                      control = rpart.control(minsplit = 0, cp = 0))
    y <- training[folds[[k]], ]$income
    # prune tree with each alpha
    tree_prune <- prune.rpart(tree_obj, cp = alpha[i])
    # number of leafs (terminal nodes)
    size <- sum(tree_prune$frame$var == "<leaf>")
    # predicted value
    tree_predict <- predict(tree_prune, newdata = training[folds[[k]], ],
                            type = "class")
    # confusion matrix
    tbl <- table(y, tree_predict)
    # test error rate
    error <- 1 - sum(diag(tbl)) / sum(tbl)
    # cost
    cost_mat[i, k] <- error + alpha[i] * size
  }
}
cost <- apply(cost_mat, 1, mean)
plot(alpha, cost, type = "l")
```




```
# optimal tuning parameter
optpar_tree <-  alpha[which.min(cost)]
optpar_tree
```

```
## [1] 6e-04
```

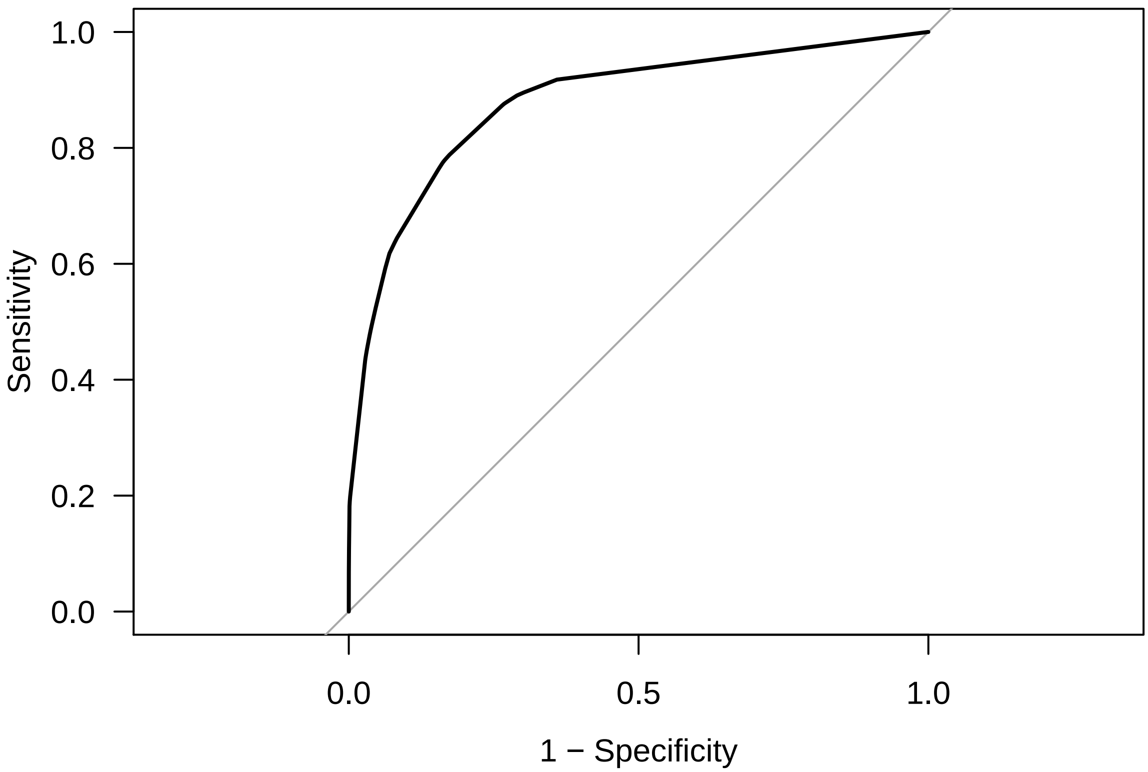
```
tree_obj <- rpart(income ~ ., training, method = "class",
                   control = rpart.control(minsplit = 0, cp = optpar_tree))

# variable importance statistics
tree_obj$variable.importance
```

```
##   relationship marital.status      education   capital_gain            sex
##    2294.929944    2243.167072    1143.203490     877.232839     756.468646
##     occupation            age hours_per_week   capital_loss native.country
##     705.662171     695.226806     448.565013      91.027163      40.260931
##      workclass           race
##      19.878244       7.830616
```

```
# training accuracy rate
tree_predict <- predict(tree_obj, type = "class")
y <- training$income
tbl <- table(y, tree_predict)  # confusion matrix
sum(diag(tbl)) / sum(tbl)
```

```
## [1] 0.852197
```

```
# ROC curve
y <- training$income
prb <- predict(tree_obj, type = "prob")[, 2]
tree_roc <- roc(
  response = y,
  predictor = prb)
plot(tree_roc, las = 1, legacy.axes = TRUE)
```



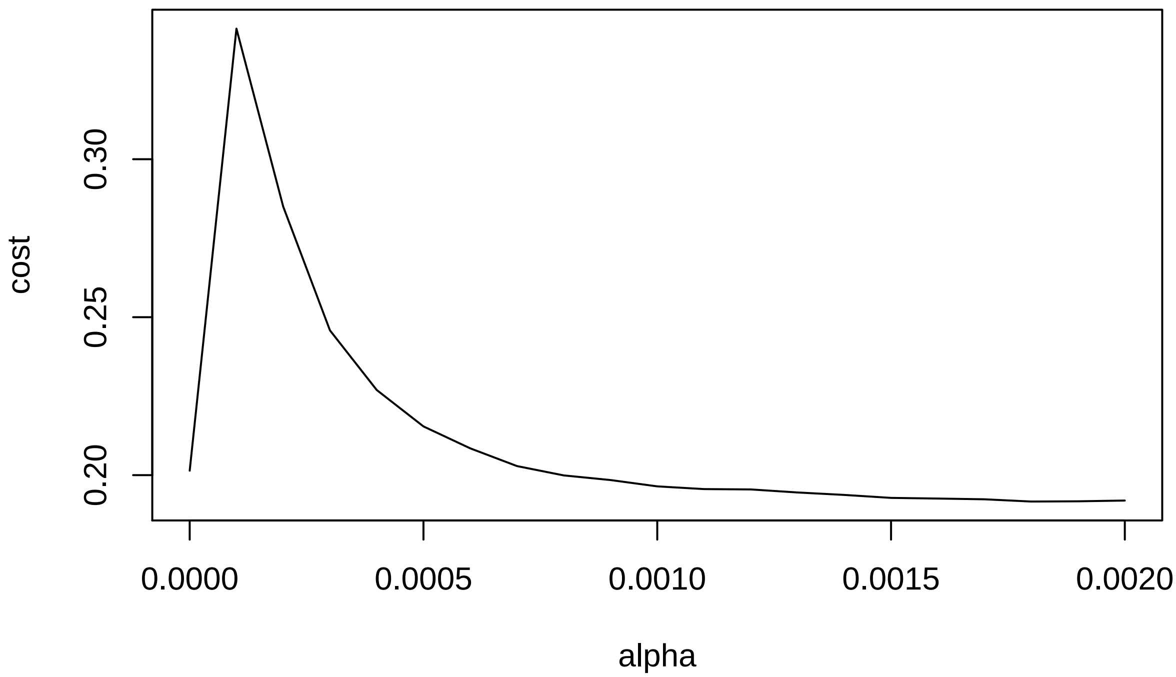

```
# AUC
auc(tree_roc)
```

```
## Area under the curve: 0.8792
```

Build a Bagged Tree

```
B <- 100
set.seed(1991)
resample <- createResample(training$income, B)
tree_bag <- list(0)
for (b in 1:B) {
  tree_bag[[b]] <- rpart(income ~ ., training[resample[[b]], ],
                         method = "class",
                         control = rpart.control(minsplit = 0, cp = 0))
}

# aggregation
alpha <- seq(from = 0, to = 0.002, by = 0.0001)
cost_mat <- matrix(0, nrow = length(alpha), ncol = B)
rownames(cost_mat) <- paste("alpha", seq_along(alpha))
colnames(cost_mat) <- paste("bag", 1:B)
for (b in 1:B) {
  tree_obj <- tree_bag[[b]]
  y <- training[-resample[[b]], ]$income
  for (i in seq_along(alpha)) {
      tree_prune <- prune.rpart(tree_obj, cp = alpha[i])
```

```r
      size <- sum(tree_prune$frame$var == "<leaf>")
      tree_predict <- predict(tree_prune, newdata = training[-resample[[b]], ],
                              type = "class")
      tbl <- table(y, tree_predict)
      error <- 1 - sum(diag(tbl)) / sum(tbl)
      cost_mat[i, b] <- error + alpha[i] * size
  }
}
cost <- apply(cost_mat, 1, mean)
plot(alpha, cost, type = "l")
```




```r
# optimal tuning parameter
optpar_bag <-  alpha[which.min(cost)]
optpar_bag
```

```
## [1] 0.0018
```

```r
tree_obj <- rpart(income ~ ., training, method = "class",
                  control = rpart.control(minsplit = 0, cp = optpar_bag))

# variable importance statistics
tree_obj$variable.importance
```

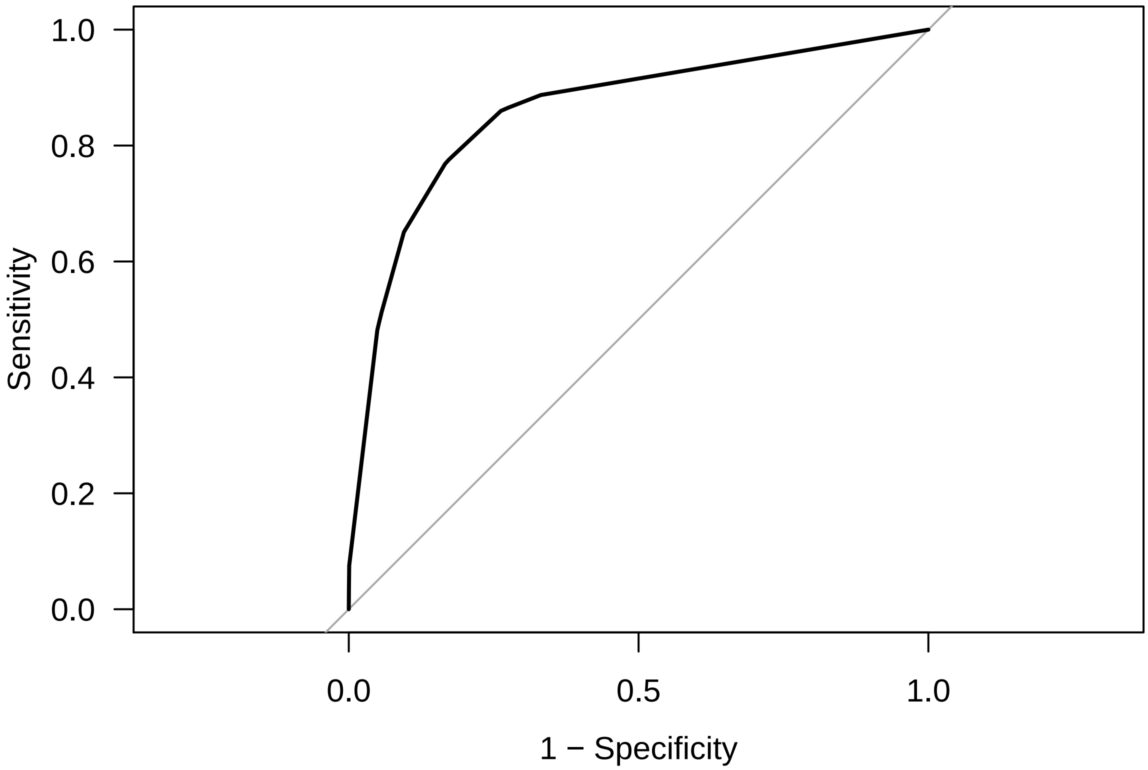
```
##   relationship marital.status      education   capital_gain            sex
##    2283.076064    2241.565618    1049.683595     791.260491     752.467019
##            age     occupation hours_per_week   capital_loss native.country
##     645.120889     631.259666     381.516551      32.519564      21.237913
##           race      workclass
##       5.868274       2.003081
```

```r
# training accuracy rate
tree_predict <- predict(tree_obj, type = "class")
y <- training$income
tbl <- table(y, tree_predict)  # confusion matrix
```

```r
sum(diag(tbl)) / sum(tbl)
```

```
## [1] 0.8415851
```

```r
# ROC curve
y <- training$income
prb <- predict(tree_obj, type = "prob")[, 2]
tree_roc <- roc(
  response = y,
  predictor = prb)
plot(tree_roc, las = 1, legacy.axes = TRUE)
```

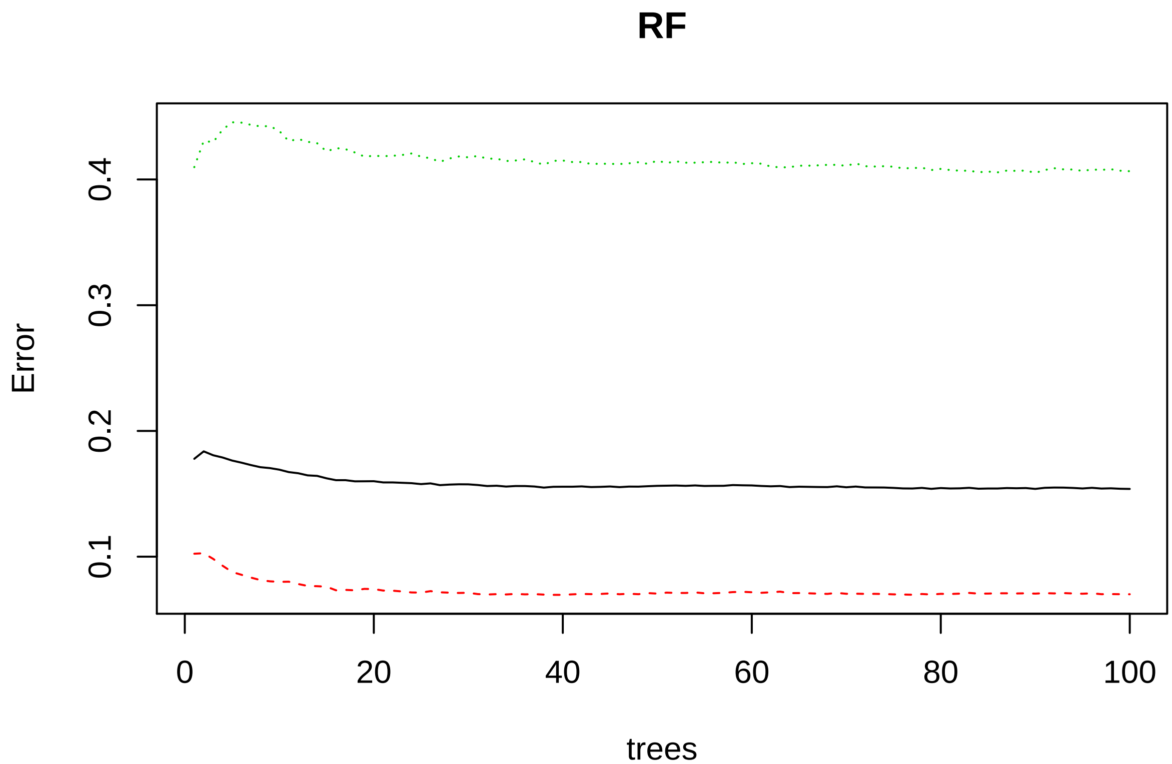



```r
# AUC
auc(tree_roc)
```

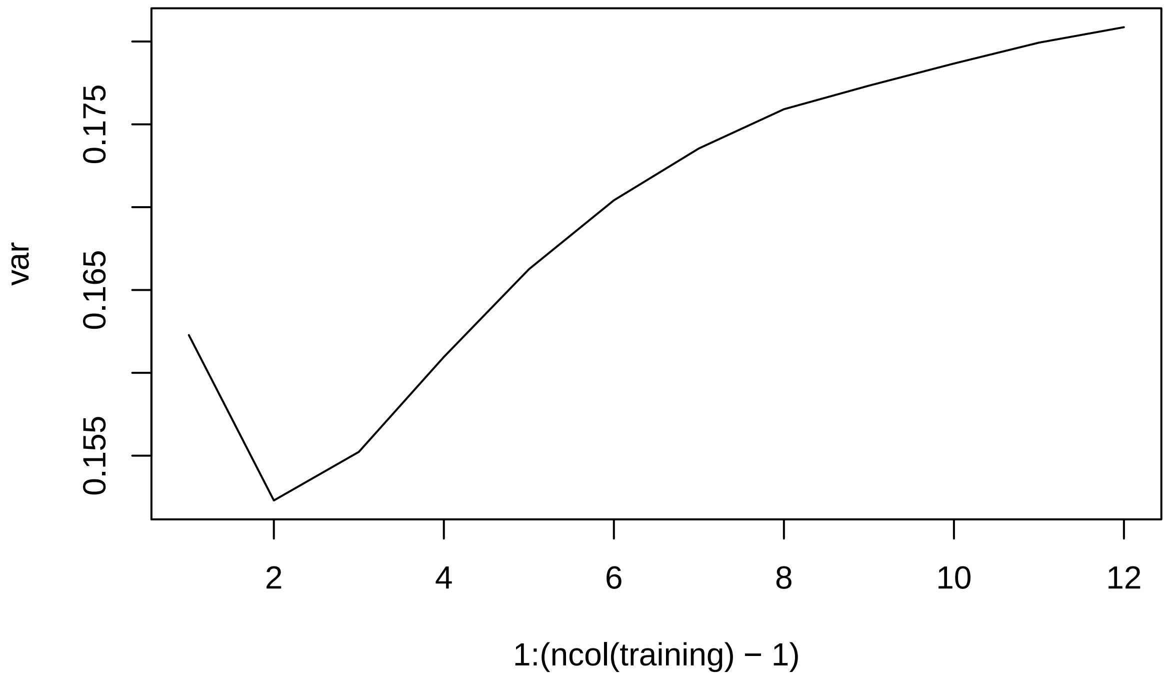
```
## Area under the curve: 0.8596
```

Build a Random Forest

```r
library(randomForest)
training$income <- factor(training$income)
set.seed(1991)
RF <- randomForest(income ~ ., training, ntree = 100, importance = TRUE)
plot(RF)
```




```
B <- 100
set.seed(1991)
resample <- createResample(training$income, B)
err_mat <- matrix(0, nrow = ncol(training) - 1, ncol = B)
rownames(err_mat) <- paste(1:12, "variables")
colnames(err_mat) <- paste("bag", 1:B)
for (b in 1:B) {
  for (i in 1:(ncol(training)-1)) {
    rf <- randomForest(income ~ ., training[resample[[b]], ],
                       ntree = 100, mtry = i)
    y <- training[-resample[[b]], ]$income
    rf_predict <- predict(rf, newdata = training[-resample[[b]], ])
    tbl <- table(y, rf_predict)
    err_mat[i, b] <- 1 - sum(diag(tbl)) / sum(tbl)
  }
}
var <- apply(err_mat, 1, mean)
plot(1:(ncol(training)-1), var, type = "l")
```



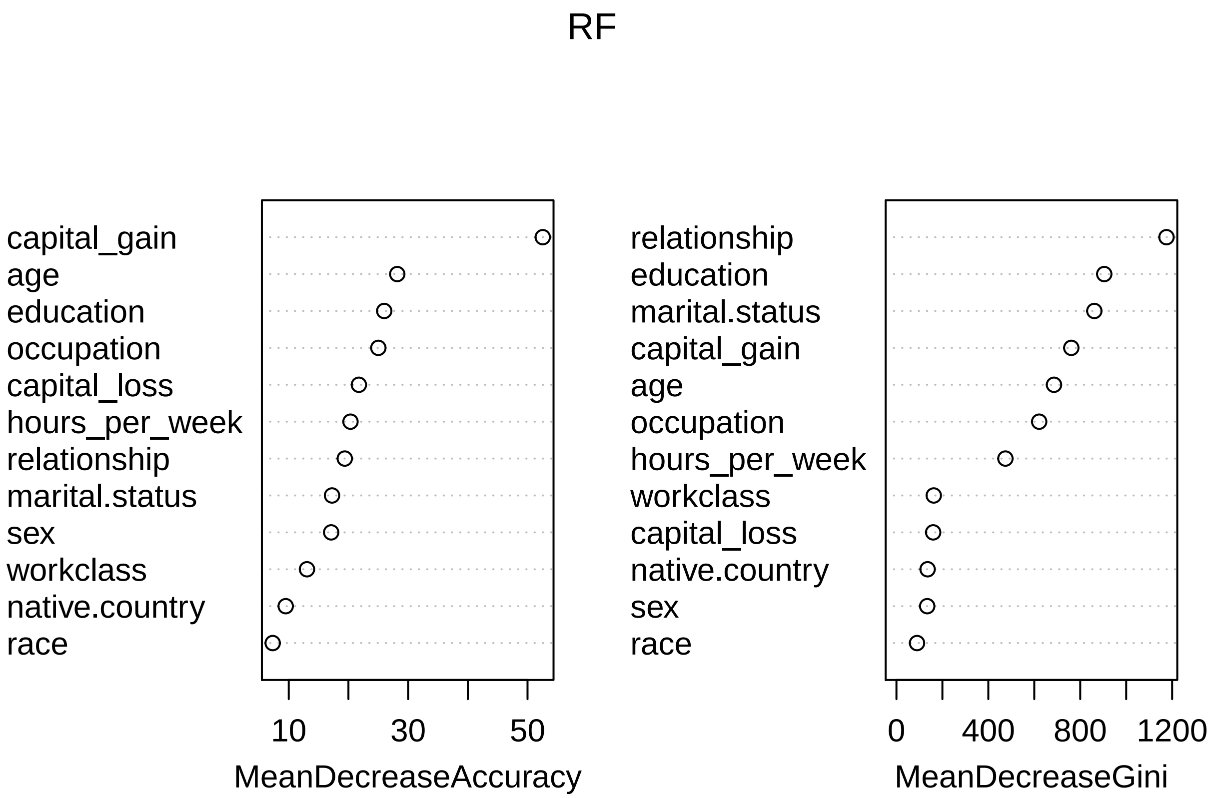

```r
# optimal tuning parameter
opt_var <- which.min(var)
opt_var
```

```
## 2 variables
##           2
```

```r
RF <- randomForest(income ~., training, ntree = 100, importance = TRUE,
                   mtry = opt_var)

# variable importance statistics
importance(RF)
```

```
##                        0         1 MeanDecreaseAccuracy MeanDecreaseGini
## age            -2.644640 27.526418            28.177284        685.78520
## capital_gain   50.921207 35.636374            52.547219        761.06304
## capital_loss   17.507497 17.872101            21.748708        159.88491
## hours_per_week  2.188473 19.776134            20.342991        474.34121
## workclass      11.200586  2.940117            13.055451        162.41762
## education      13.334597 23.093608            25.994977        904.45746
## marital.status 15.633425  9.777223            17.253938        861.35378
## occupation     15.974073 18.674522            25.009730        621.26213
## relationship   11.530489 17.434287            19.386471       1175.37690
## race            5.129806  2.618156             7.312750         89.62496
## sex             8.664511  4.833491            17.111009        133.94295
## native.country  8.077225  4.831371             9.494777        135.64343
```

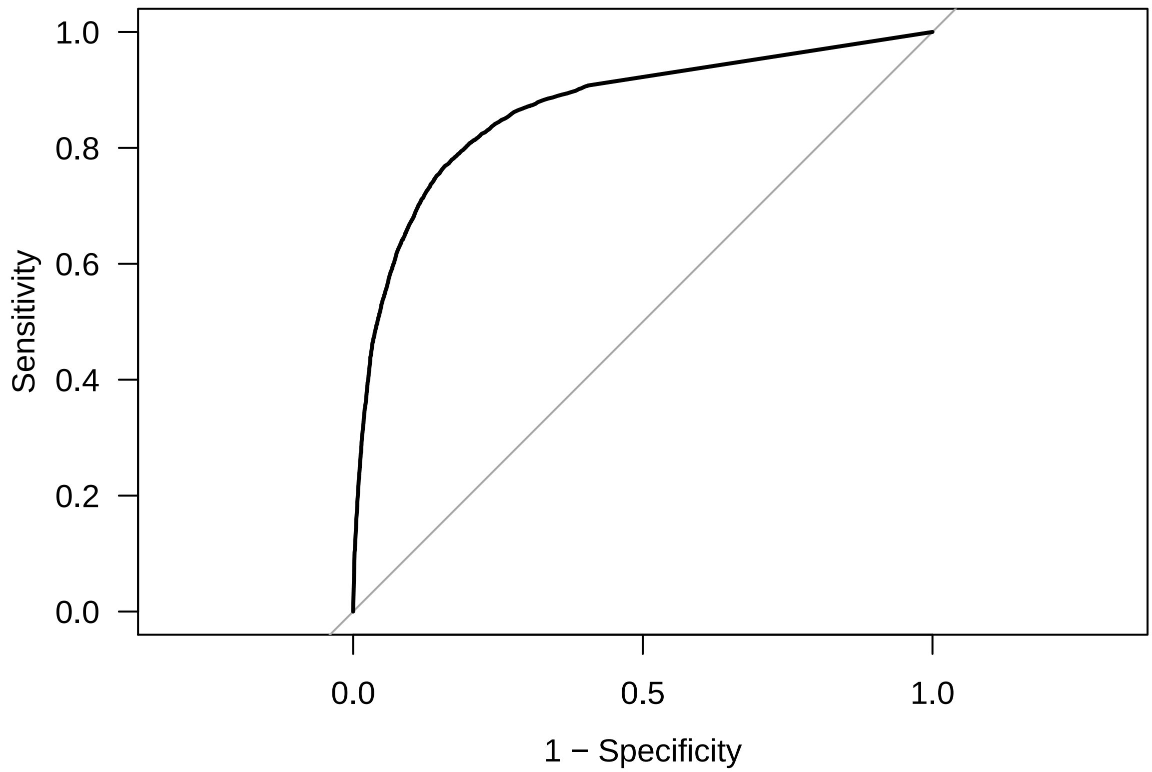
```r
# variable importance plot
varImpPlot(RF)
```




```r
# training accuracy rate
RF_predict <- predict(RF, type = "class")
y <- training$income
tbl <- table(y, tree_predict)  # confusion matrix
sum(diag(tbl)) / sum(tbl)
```

```
## [1] 0.8415851
```

```r
# ROC curve
y <- training$income
prb <- predict(RF, type = "prob")[, 2]
tree_roc <- roc(
  response = y,
  predictor = prb)
plot(tree_roc, las = 1, legacy.axes = TRUE)
```




```
# AUC
auc(tree_roc)
```

```
## Area under the curve: 0.8689
```

Model Selection

```
y <- test$income
# classification tree
tree_cl <- rpart(income ~ ., test, method = "class",
                 control = rpart.control(minsplit = 0, cp = optpar_tree))
y_hat_cl <- predict(tree_cl, type = "class")

# confusion matrix
tbl_cl <- table(y, y_hat_cl)
tbl_cl
```

```
##    y_hat_cl
## y       0     1
##   0 10738   622
##   1  1485  2215
```

```
# Sensitivity(TPR) and Specificity(TNR)
TPR_cl <- tbl_cl[2, 2] / sum(tbl_cl[2, ])
TPR_cl
```

```
## [1] 0.5986486
```

```r
TNR_cl <- tbl_cl[1, 1] / sum(tbl_cl[1, ])
TNR_cl
```

```
## [1] 0.9452465
```

```r
# bagged tree
tree_bag <- rpart(income ~ ., test, method = "class",
                  control = rpart.control(minsplit = 0, cp = optpar_bag))
y_hat_bag <- predict(tree_bag, type = "class")

# confusion matrix
tbl_bag <- table(y, y_hat_bag)
tbl_bag
```

```
##    y_hat_bag
## y       0     1
##   0 10523   837
##   1  1416  2284
```

```r
# Sensitivity(TPR) and Specificity(TNR)
TPR_bag <- tbl_bag[2, 2] / sum(tbl_bag[2, ])
TPR_bag
```

```
## [1] 0.6172973
```

```r
TNR_bag <- tbl_bag[1, 1] / sum(tbl_bag[1, ])
TNR_bag
```

```
## [1] 0.9263204
```

```r
# random forest
test$income <- factor(test$income)
rf <- randomForest(income ~ ., test, ntree = 100, importance = TRUE,
                   mtry = opt_var)
y_hat_rf <- predict(rf, type = "class")

# confusion matrix
tbl_rf <- table(y, y_hat_rf)
tbl_rf
```

```
##    y_hat_rf
## y       0     1
##   0 10695   665
##   1  1632  2068
```

```r
# Sensitivity(TPR) and Specificity(TNR)
TPR_rf <- tbl_rf[2, 2] / sum(tbl_rf[2, ])
TPR_rf
```

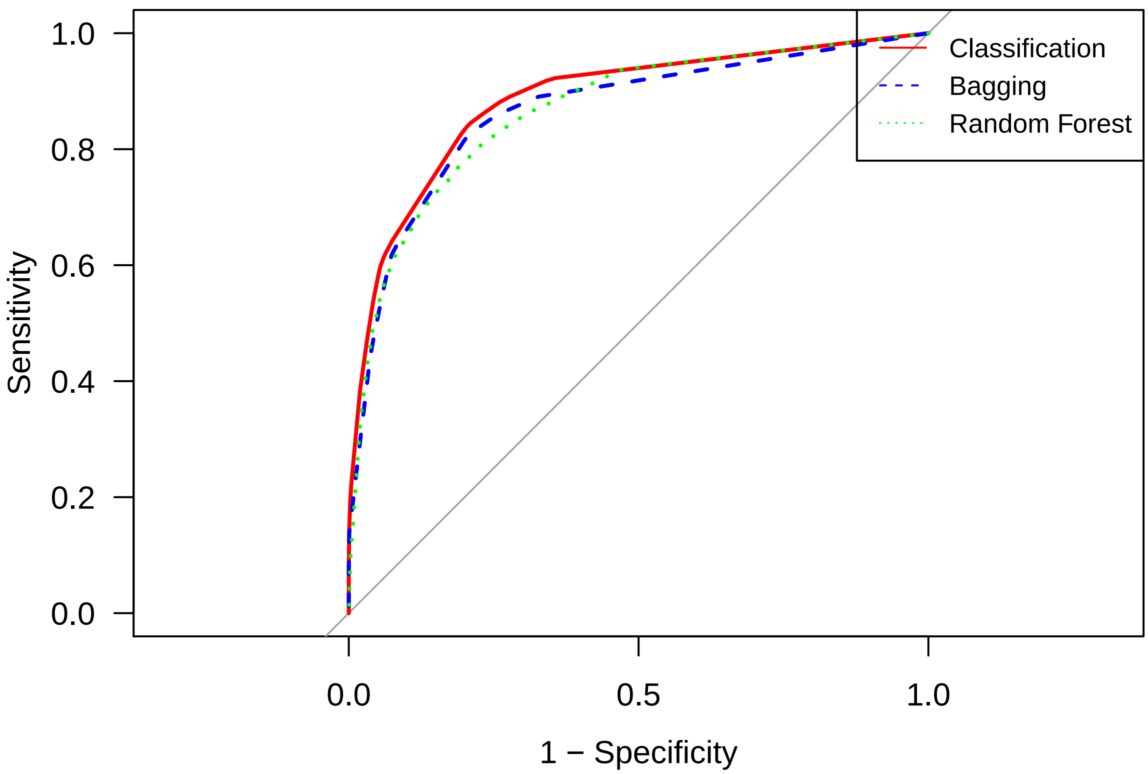
```
## [1] 0.5589189
```

```r
TNR_rf <- tbl_rf[1, 1] / sum(tbl_rf[1, ])
TNR_rf
```

```
## [1] 0.9414613
```

```r
# ROC curve
prb_cl <- predict(tree_cl, type = "prob")[, 2]
cl_roc <- roc(
```

```r
  response = y,
  predictor = prb_cl)


prb_bag <- predict(tree_bag, type = "prob")[, 2]
bag_roc <- roc(
  response = y,
  predictor = prb_bag)


prb_rf <- predict(rf, type = "prob")[, 2]
rf_roc <- roc(
  response = y,
  predictor = prb_rf)


plot(cl_roc, las = 1, legacy.axes = TRUE, col = "red", lwd = 2)
plot(bag_roc, las = 1, legacy.axes = TRUE, col = "blue", lwd = 2, add = TRUE,
     lty = 2)
plot(rf_roc, las = 1, legacy.axes = TRUE, col = "green", lwd = 2, add = TRUE,
     lty = 3)
legend("topright", legend=c("Classification", "Bagging", "Random Forest"),
       col=c("red", "blue", "green"), lty=1:3, cex=0.8)
```




```r
# AUC of cl, bag, and rf respectively
c(auc(cl_roc), auc(bag_roc), auc(rf_roc))
```

```
## [1] 0.8854217 0.8662442 0.8671155
```