Adult Census Income

Michihito Mitsuyasu

2019/5/12

Contents

0.1 library

```r
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.1.0       v purrr   0.2.5
## v tibble  2.0.0       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.3.1       v forcats 0.3.0
## Warning: package 'ggplot2' was built under R version 3.4.4
## Warning: package 'tidyr' was built under R version 3.4.4
## Warning: package 'readr' was built under R version 3.4.4
## Warning: package 'purrr' was built under R version 3.4.4
## Warning: package 'dplyr' was built under R version 3.4.4
## Warning: package 'stringr' was built under R version 3.4.4
## -- Conflicts -------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
## Loading required package: lattice
## Warning: package 'lattice' was built under R version 3.4.4
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.4.4
## randomForest 4.6-14
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:dplyr':
##
##     combine
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
library(ggridges)
```

```
## Warning: package 'ggridges' was built under R version 3.4.4
##
## Attaching package: 'ggridges'
## The following object is masked from 'package:ggplot2':
##
##     scale_discrete_manual
```

```r
library(rpart)
library(partykit)
```

```
## Warning: package 'partykit' was built under R version 3.4.4
## Loading required package: grid
## Loading required package: libcoin
## Warning: package 'libcoin' was built under R version 3.4.4
## Loading required package: mvtnorm
## Warning: package 'mvtnorm' was built under R version 3.4.4
```

0.2 Executive summary section

0.2.1 describes the dataset

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)).

The prediction task is to determine whether a person makes over $50K a year.

https://www.kaggle.com/uciml/adult-census-income/activity

0.2.1.1 Description of fnlwgt (final weight)

The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:

A single cell estimate of the population 16+ for each state.

Controls for Hispanic Origin by age and sex.

Controls by Race, age and sex.

We use all three sets of controls in our weighting program and "rake" through them 6 times so that by the end we come back to all the controls we used. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.

0.2.1.2 About this file

About this file Attributes:

50K, <=50K

age: continuous

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

fnlwgt: continuous

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool

education-num: continuous

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black

sex: Female, Male

capital-gain: continuous

capital-loss: continuous

hours-per-week: continuous

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

0.2.2 train set and test set

I will create the trani set and the test set.

The train set is used to develop my algorithm.

The test set is used to evaluate how close your predictions are to the true values.

0.2.3 The goal of the project

The prediction task is to use a test set to determine whether a person makes over $50K a year.

The predictions will be compared to the true ratings in the test set using accuracy.

0.2.4 Key steps that were performed

1. Describes the adult dataset
2. Data exploration and visualization
3. Generate the train set and the test set
4. Comparison of CART and Random Forests
    1. CART
    2. Random Forests

0.3 Methods section

0.3.1 describes the adult dataset

I use the following code to generate the adult set.

```
#create adult dataset
adult <- read_csv("https://github.com/mitti1210/edx-Choose_Your_Own/blob/master/adult.csv?raw=true")
```

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   workclass = col_character(),
##   fnlwgt = col_double(),
##   education = col_character(),
##   education.num = col_double(),
##   marital.status = col_character(),
##   occupation = col_character(),
##   relationship = col_character(),
```

```
##   race = col_character(),
##   sex = col_character(),
##   capital.gain = col_double(),
##   capital.loss = col_double(),
##   hours.per.week = col_double(),
##   native.country = col_character(),
##   income = col_character()
## )
```

```
adult %>%
  select_if(is.character) %>%
  map(., ~ levels(factor(.x)))
```

```
## $workclass
## [1] "?"                "Federal-gov"      "Local-gov"
## [4] "Never-worked"     "Private"          "Self-emp-inc"
## [7] "Self-emp-not-inc" "State-gov"        "Without-pay"
##
## $education
##  [1] "10th"         "11th"         "12th"         "1st-4th"
##  [5] "5th-6th"      "7th-8th"      "9th"          "Assoc-acdm"
##  [9] "Assoc-voc"    "Bachelors"    "Doctorate"    "HS-grad"
## [13] "Masters"      "Preschool"    "Prof-school"  "Some-college"
##
## $marital.status
## [1] "Divorced"              "Married-AF-spouse"     "Married-civ-spouse"
## [4] "Married-spouse-absent" "Never-married"         "Separated"
## [7] "Widowed"
##
## $occupation
##  [1] "?"                 "Adm-clerical"      "Armed-Forces"
##  [4] "Craft-repair"      "Exec-managerial"   "Farming-fishing"
##  [7] "Handlers-cleaners" "Machine-op-inspct" "Other-service"
## [10] "Priv-house-serv"   "Prof-specialty"    "Protective-serv"
## [13] "Sales"             "Tech-support"      "Transport-moving"
##
## $relationship
## [1] "Husband"        "Not-in-family"  "Other-relative" "Own-child"
## [5] "Unmarried"      "Wife"
##
## $race
## [1] "Amer-Indian-Eskimo" "Asian-Pac-Islander" "Black"
## [4] "Other"              "White"
##
## $sex
## [1] "Female" "Male"
##
## $native.country
##  [1] "?"                          "Cambodia"
##  [3] "Canada"                     "China"
##  [5] "Columbia"                   "Cuba"
##  [7] "Dominican-Republic"         "Ecuador"
##  [9] "El-Salvador"                "England"
## [11] "France"                     "Germany"
## [13] "Greece"                     "Guatemala"
```

```
## [15] "Haiti"                      "Holand-Netherlands"
## [17] "Honduras"                   "Hong"
## [19] "Hungary"                    "India"
## [21] "Iran"                       "Ireland"
## [23] "Italy"                      "Jamaica"
## [25] "Japan"                      "Laos"
## [27] "Mexico"                     "Nicaragua"
## [29] "Outlying-US(Guam-USVI-etc)" "Peru"
## [31] "Philippines"                "Poland"
## [33] "Portugal"                   "Puerto-Rico"
## [35] "Scotland"                   "South"
## [37] "Taiwan"                     "Thailand"
## [39] "Trinadad&Tobago"            "United-States"
## [41] "Vietnam"                    "Yugoslavia"
##
## $income
## [1] "<=50K" ">50K"
```

```r
#String processing was performed because "?", "NA", and "-" were used. I changed income to 0,1.

adult <-
  adult %>%
  mutate_if(is_character, funs(str_replace_all(., pattern = "\\-", "_"))) %>%
  mutate_if(is_character, funs(str_replace_all(., pattern = "\\&", "_"))) %>%
  mutate_if(is_character, funs(str_replace_all(., pattern = "\\?", "NA"))) %>%
  mutate_if(is_character, as.factor) %>%
  mutate(income = as.factor(ifelse(income %in% ">50K", "1", "0")))
```

```
## Warning: funs() is soft deprecated as of dplyr 0.8.0
## please use list() instead
##
## # Before:
## funs(name = f(.)
##
## # After:
## list(name = ~f(.))
## This warning is displayed once per session.
```

```r
head(adult)
```

```
## # A tibble: 6 x 15
##     age workclass fnlwgt education education.num marital.status occupation
##   <dbl> <fct>      <dbl> <fct>             <dbl> <fct>          <fct>
## 1    90 NA         77053 HS_grad               9 Widowed        NA
## 2    82 Private   132870 HS_grad               9 Widowed        Exec_mana~
## 3    66 NA        186061 Some_col~            10 Widowed        NA
## 4    54 Private   140359 7th_8th               4 Divorced       Machine_o~
## 5    41 Private   264663 Some_col~            10 Separated      Prof_spec~
## 6    34 Private   216864 HS_grad               9 Divorced       Other_ser~
## # ... with 8 more variables: relationship <fct>, race <fct>, sex <fct>,
## #   capital.gain <dbl>, capital.loss <dbl>, hours.per.week <dbl>,
## #   native.country <fct>, income <fct>
```

0.3.2 Data exploration and visualization

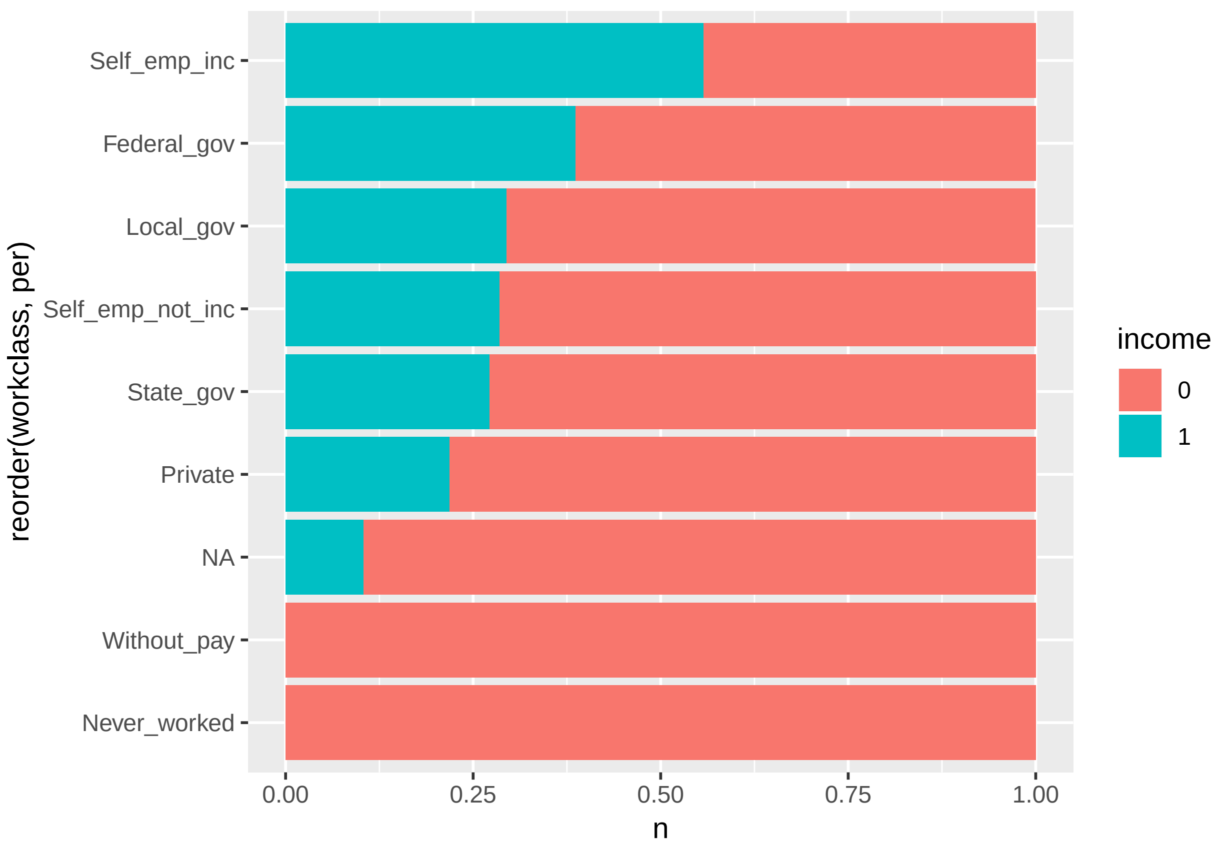
0.3.2.1 Exploration

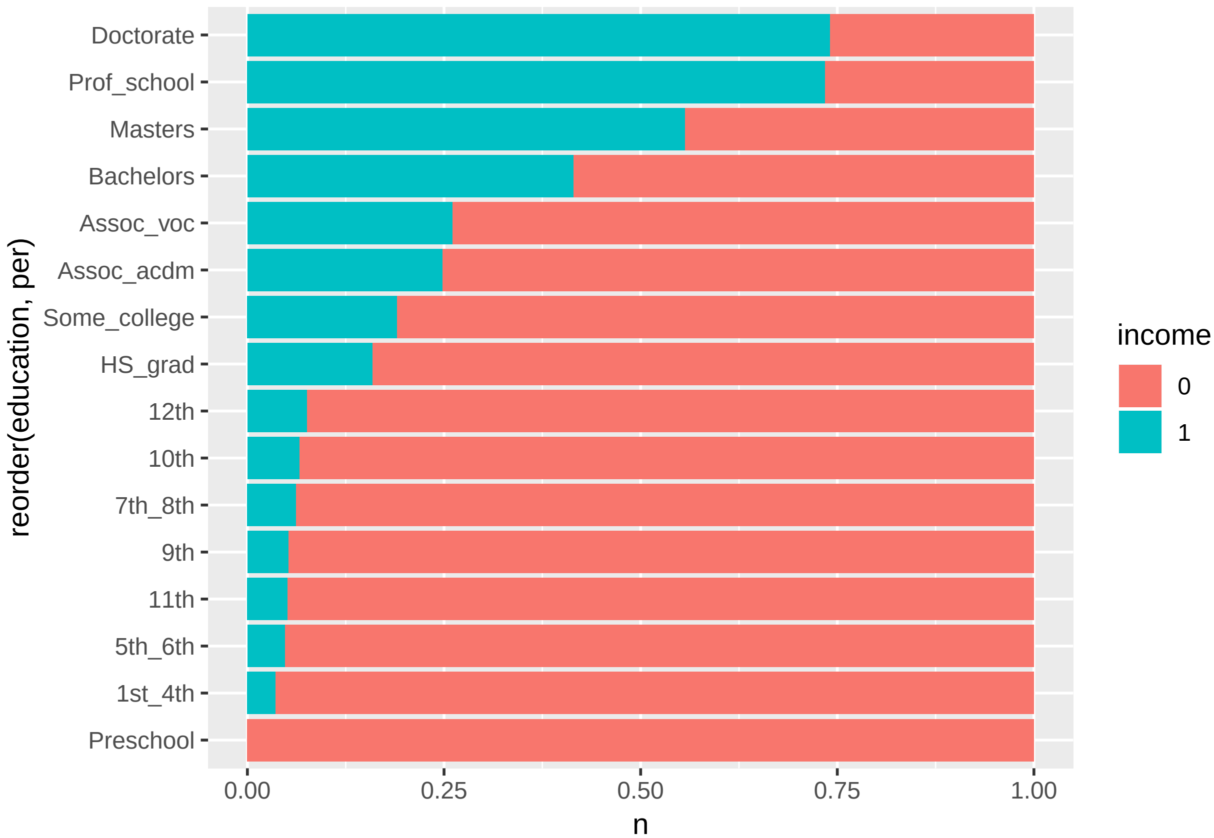
```r
str(adult)
```

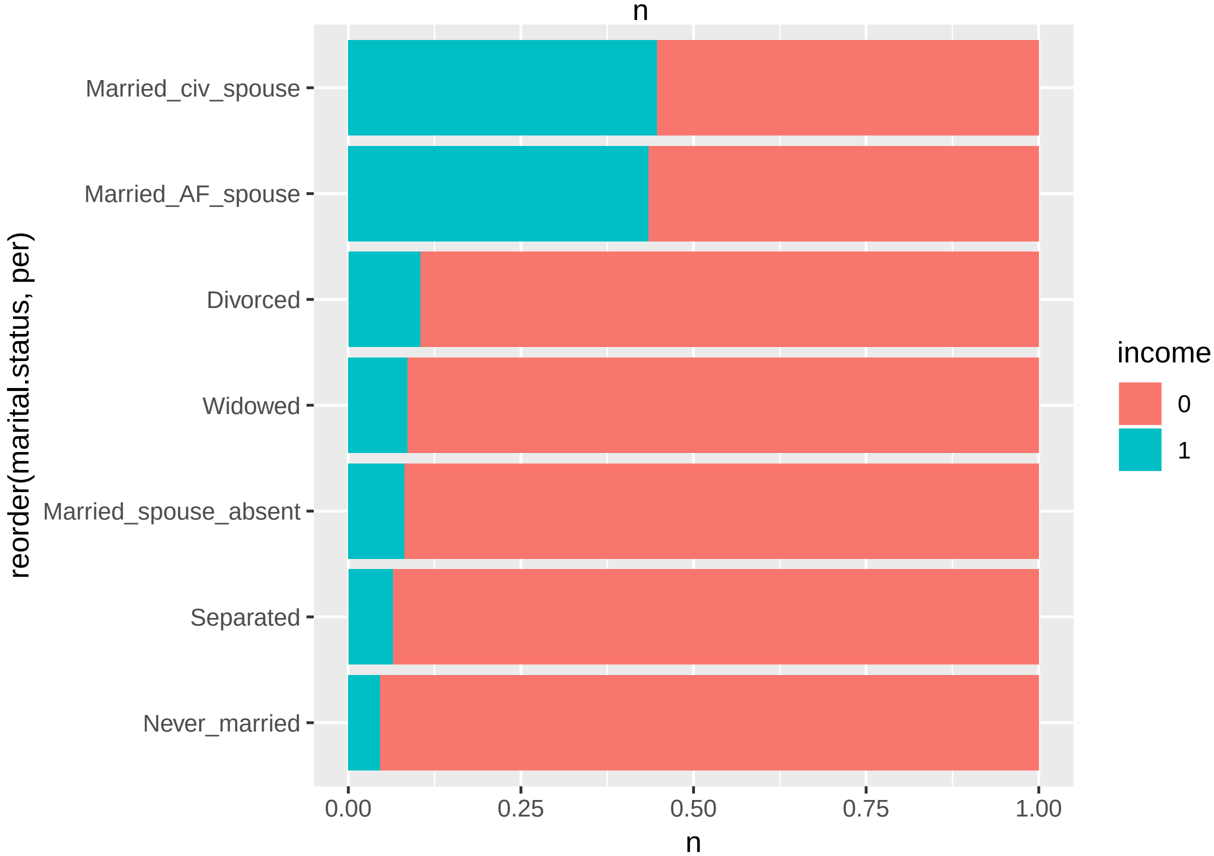
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 32561 obs. of  15 variables:
##  $ age           : num  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : Factor w/ 9 levels "Federal_gov",..: 3 5 3 5 5 5 5 8 1 5 ...
##  $ fnlwgt        : num  77053 132870 186061 140359 264663 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16 ...
##  $ education.num : num  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married_AF_spouse",..: 7 7 7 1 6 1 6 5 1 5 ...
##  $ occupation    : Factor w/ 15 levels "Adm_clerical",..: 8 4 8 7 11 9 1 11 11 3 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not_in_family",..: 2 2 5 5 4 5 5 3 2 5 ...
##  $ race          : Factor w/ 5 levels "Amer_Indian_Eskimo",..: 5 5 3 5 5 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ capital.gain  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : num  4356 4356 4356 3900 3900 ...
##  $ hours.per.week: num  40 18 40 40 40 45 40 20 40 60 ...
##  $ native.country: Factor w/ 42 levels "Cambodia","Canada",..: 40 40 40 40 40 40 40 40 40 27 ...
##  $ income        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 2 1 2 ...
```

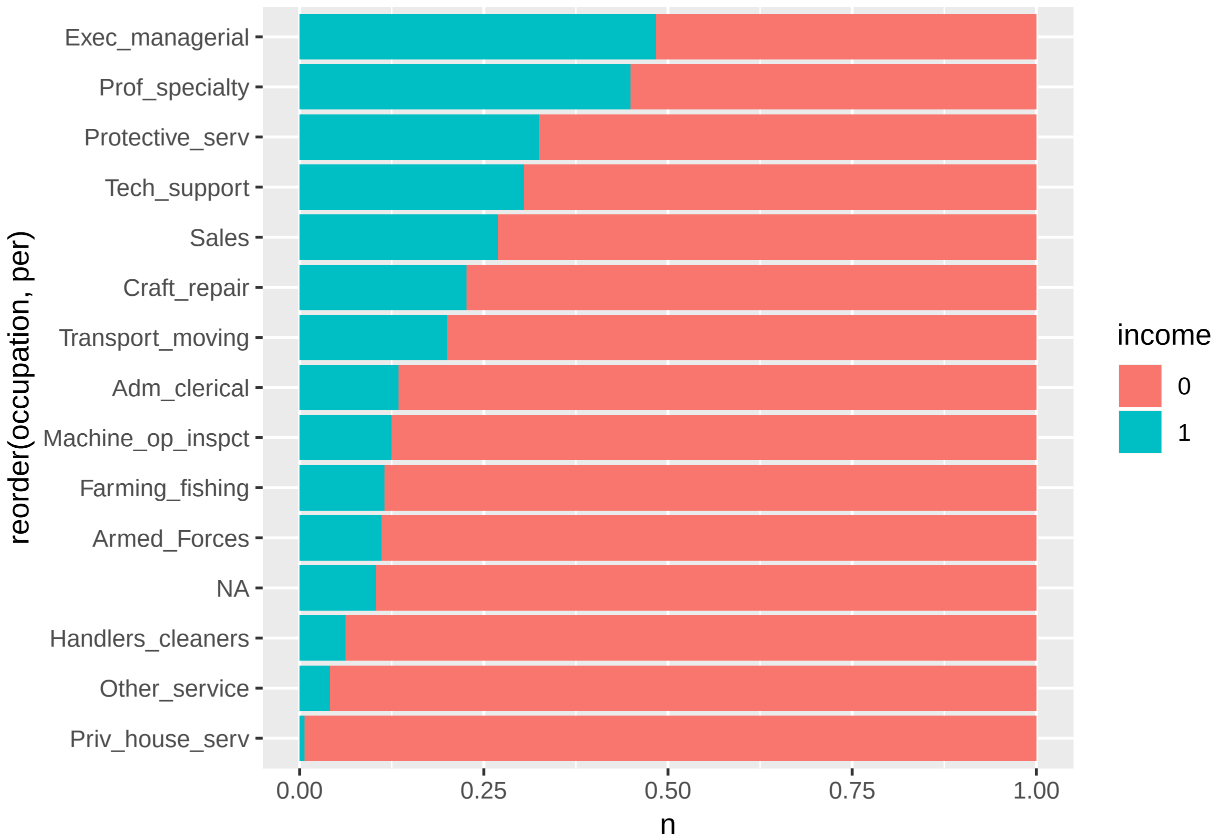
0.3.2.2 Visualization

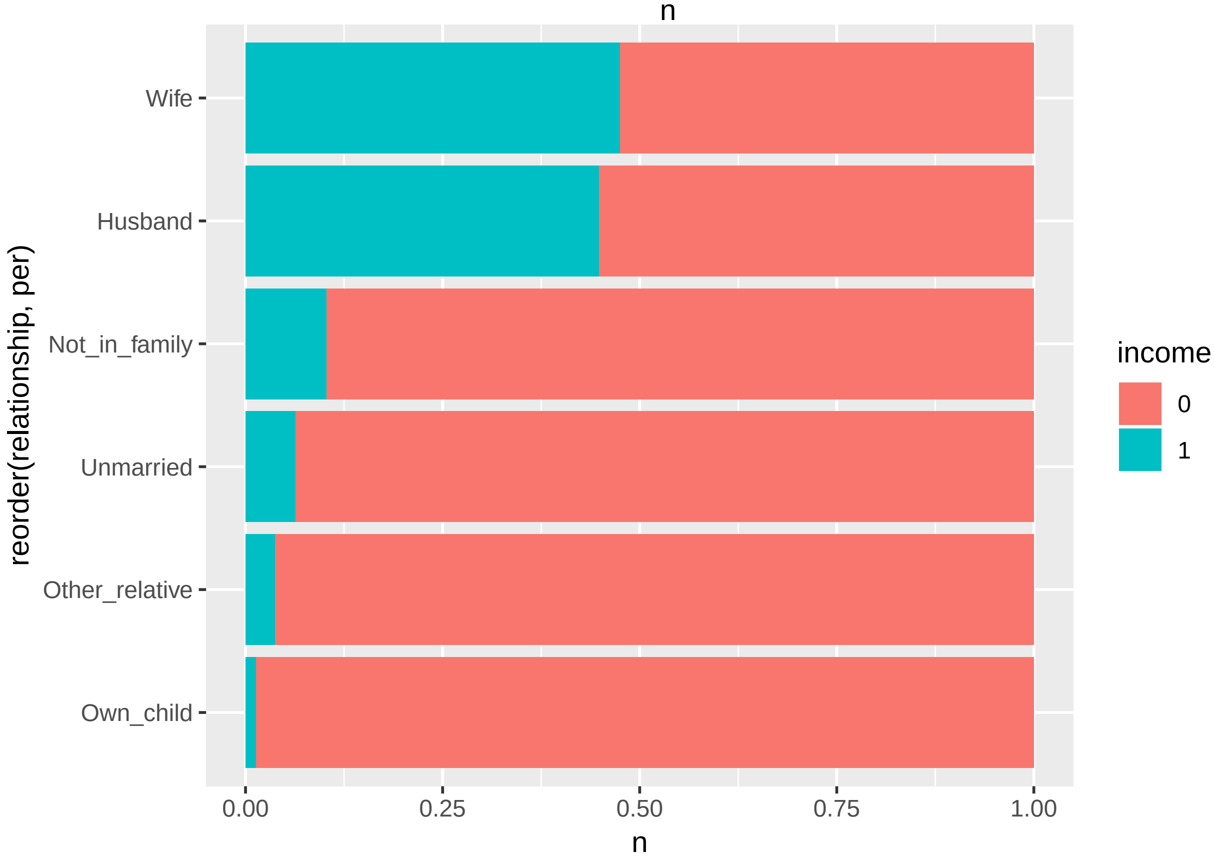
0.3.2.2.1 Bar chart and density chart for each attribute

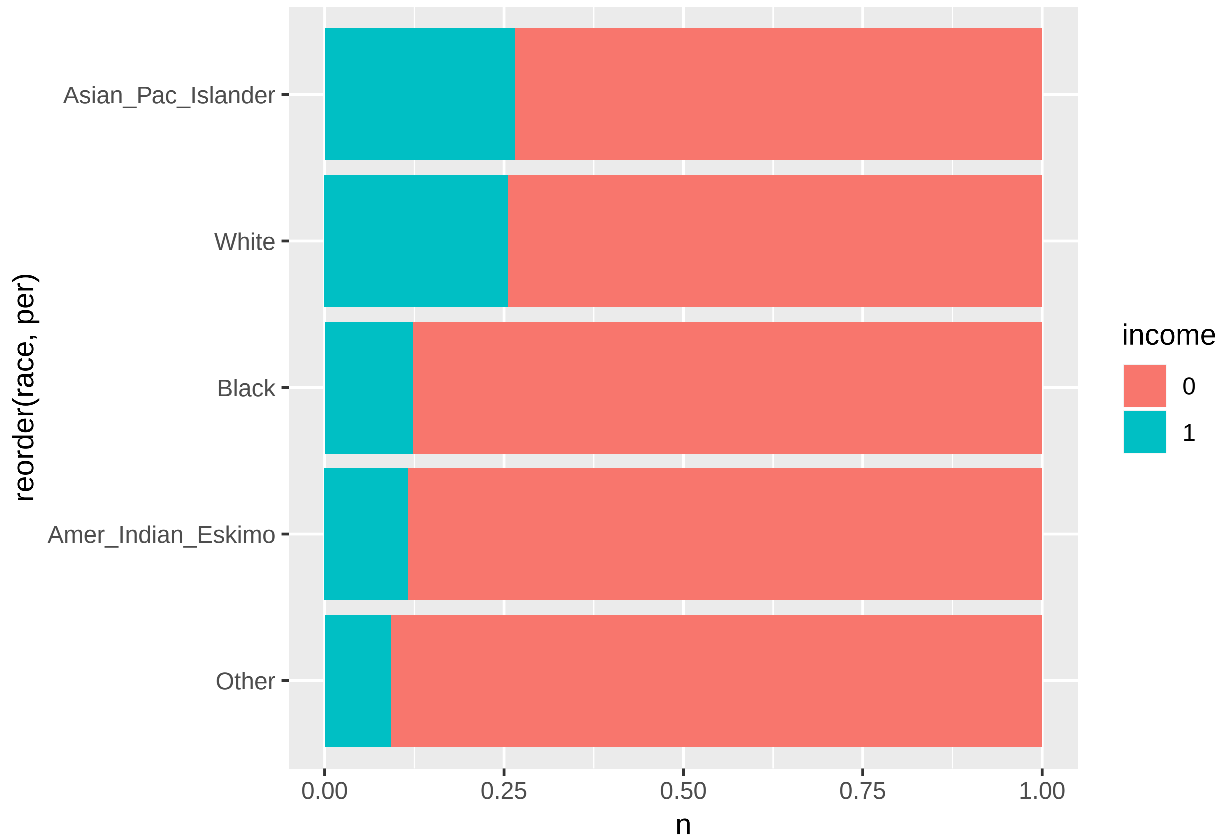


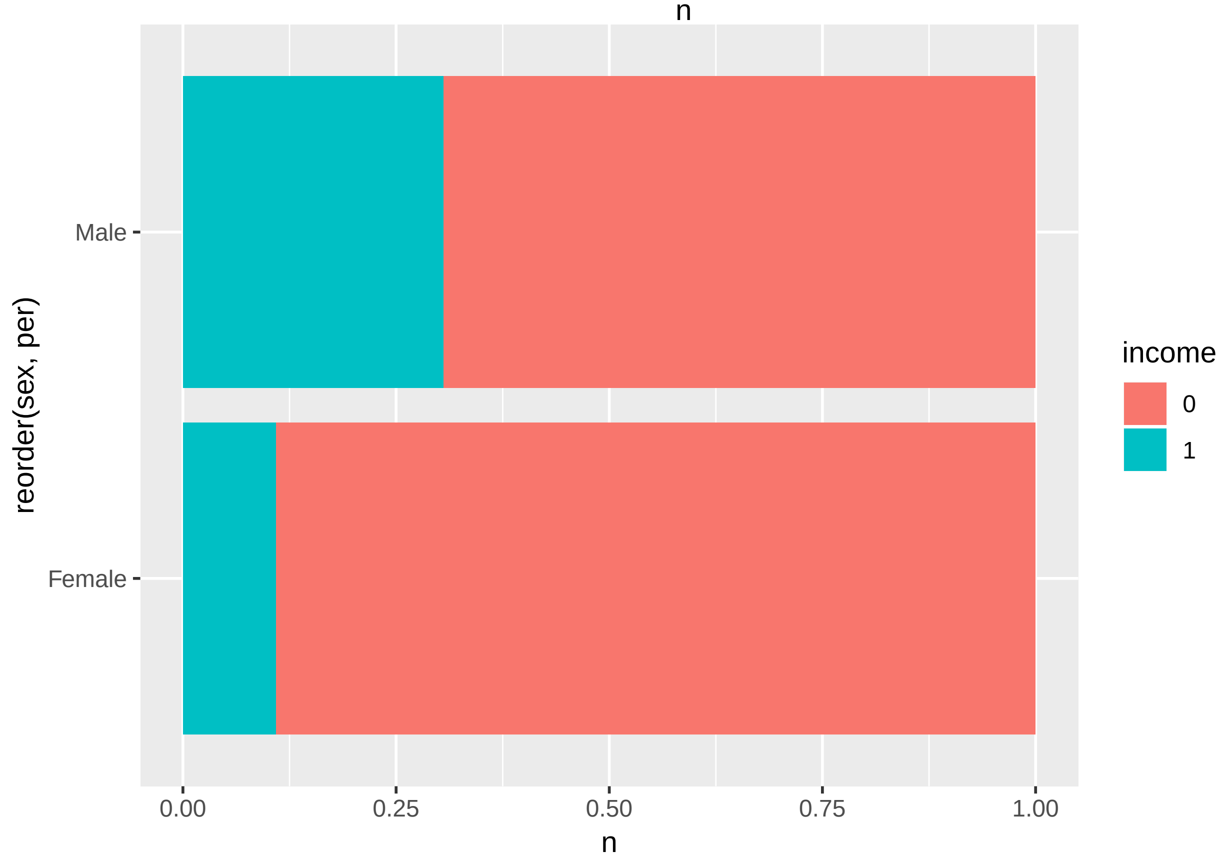

Doctorate
Prof_school
Masters
Bachelors
Assoc_voc
Assoc_acdm
Some_college
HS_grad
12th
10th
7th_8th
9th
11th
5th_6th
1st_4th
Preschool
reorder(education, per)
0.00
0.25
0.50
0.75
1.00
n
income
0
1

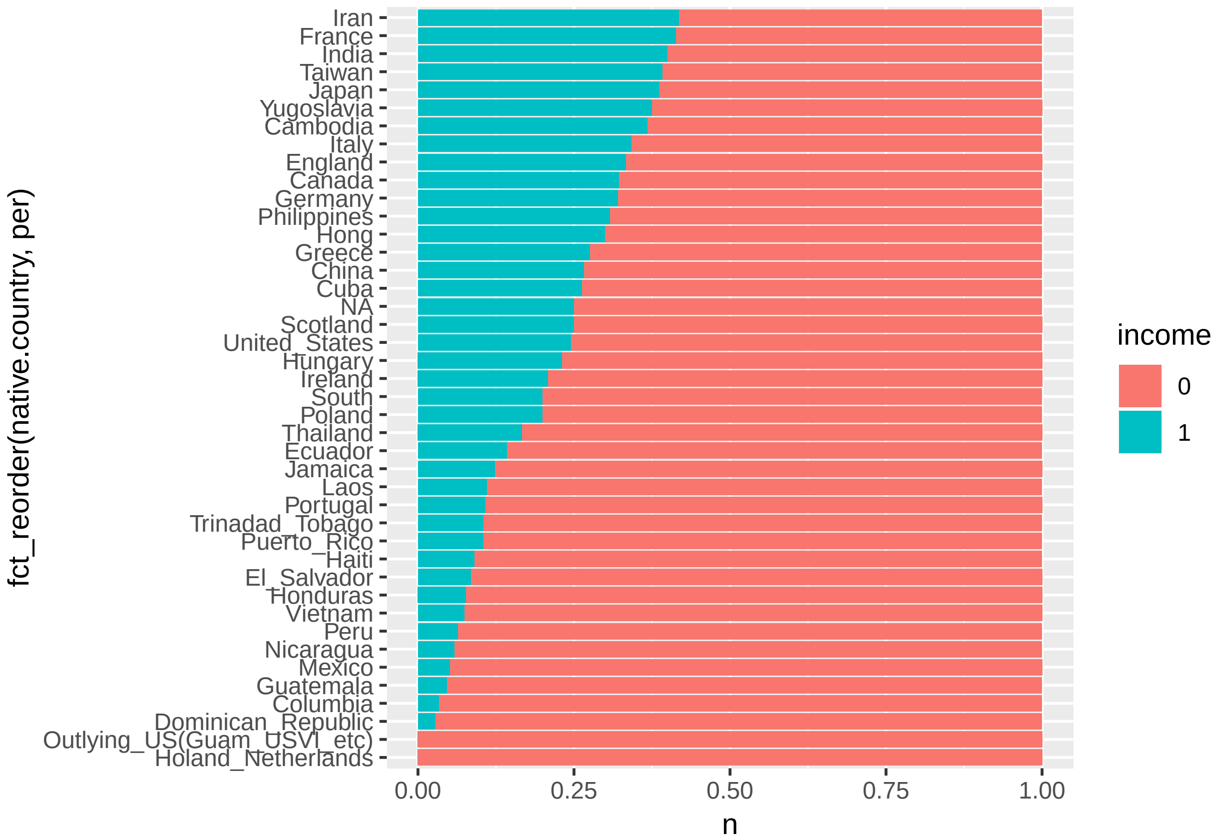

Married_civ_spouse
Married_AF_spouse
Divorced
Widowed
Married_spouse_absent
Separated
Never_married
reorder(marital.status, per)
0.00
0.25
0.50
0.75
1.00
n
income
0
1


Exec_managerial
Prof_specialty
Protective_serv
Tech_support
Sales
Craft_repair
Transport_moving
Adm_clerical
Machine_op_inspct
Farming_fishing
Armed_Forces
NA
Handlers_cleaners
Other_service
Priv_house_serv
reorder(occupation, per)
0.00
0.25
0.50
0.75
1.00
n
income
0
1

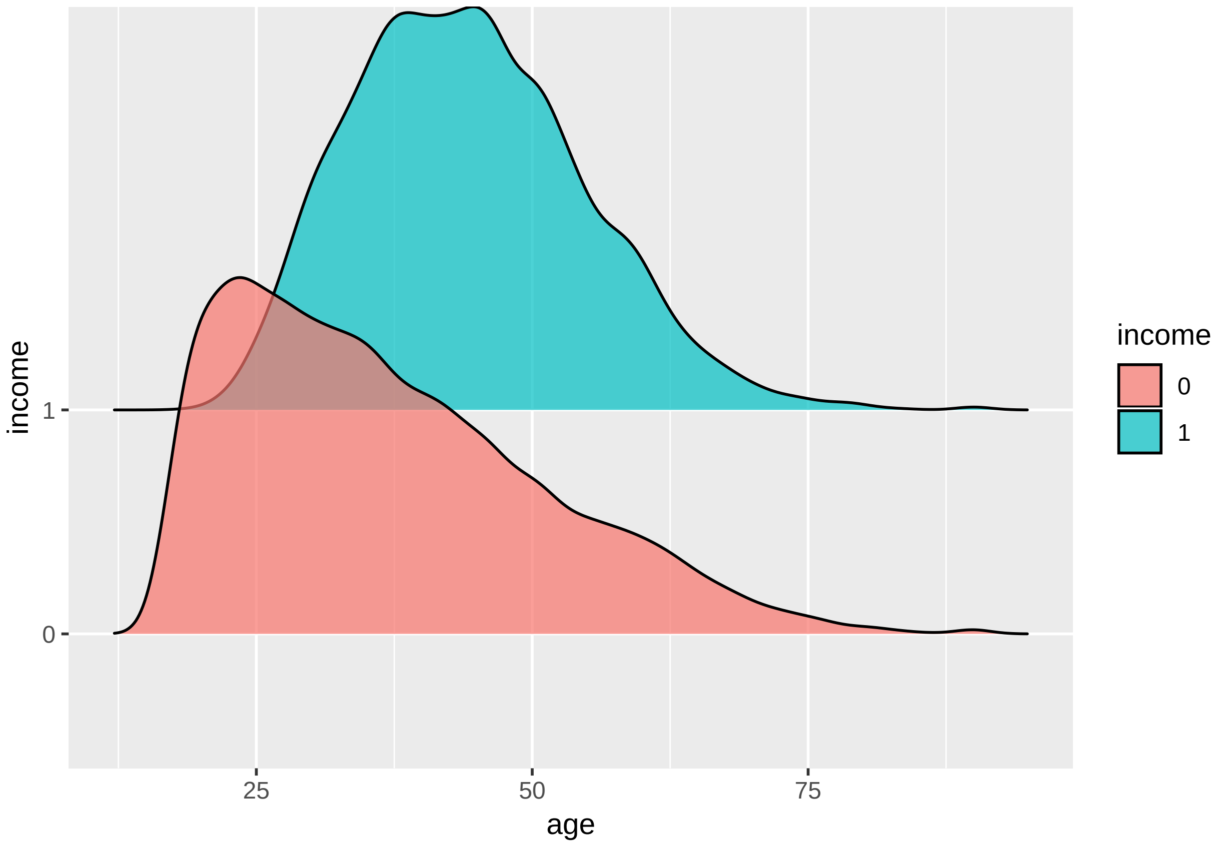

Wife
Husband
Not_in_family
Unmarried
Other_relative
Own_child
reorder(relationship, per)
0.00
0.25
0.50
0.75
1.00
n
income
0
1


Asian_Pac_Islander
White
Black
Amer_Indian_Eskimo
Other
reorder(race, per)
0.00
0.25
0.50
0.75
1.00
n
income
0
1

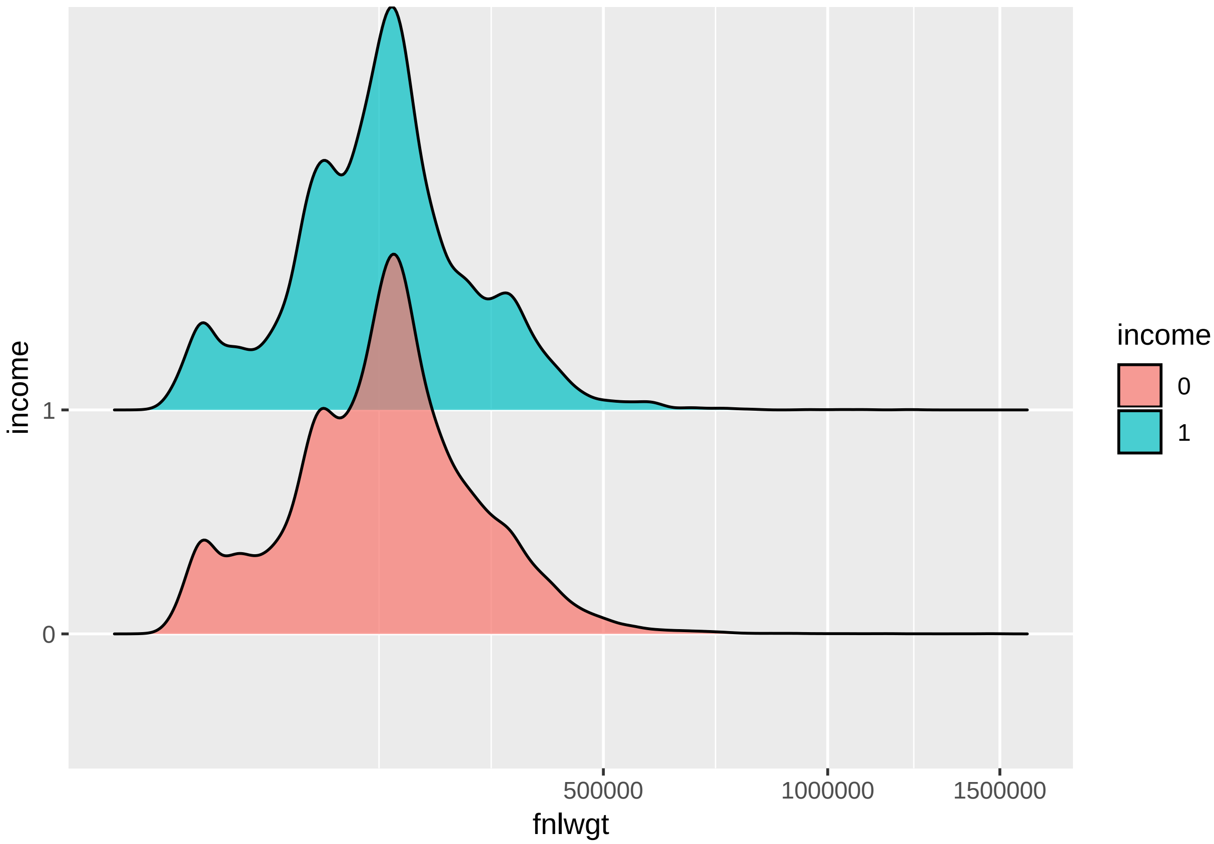

Male
Female
reorder(sex, per)
0.00
0.25
0.50
0.75
1.00
n
income
0
1

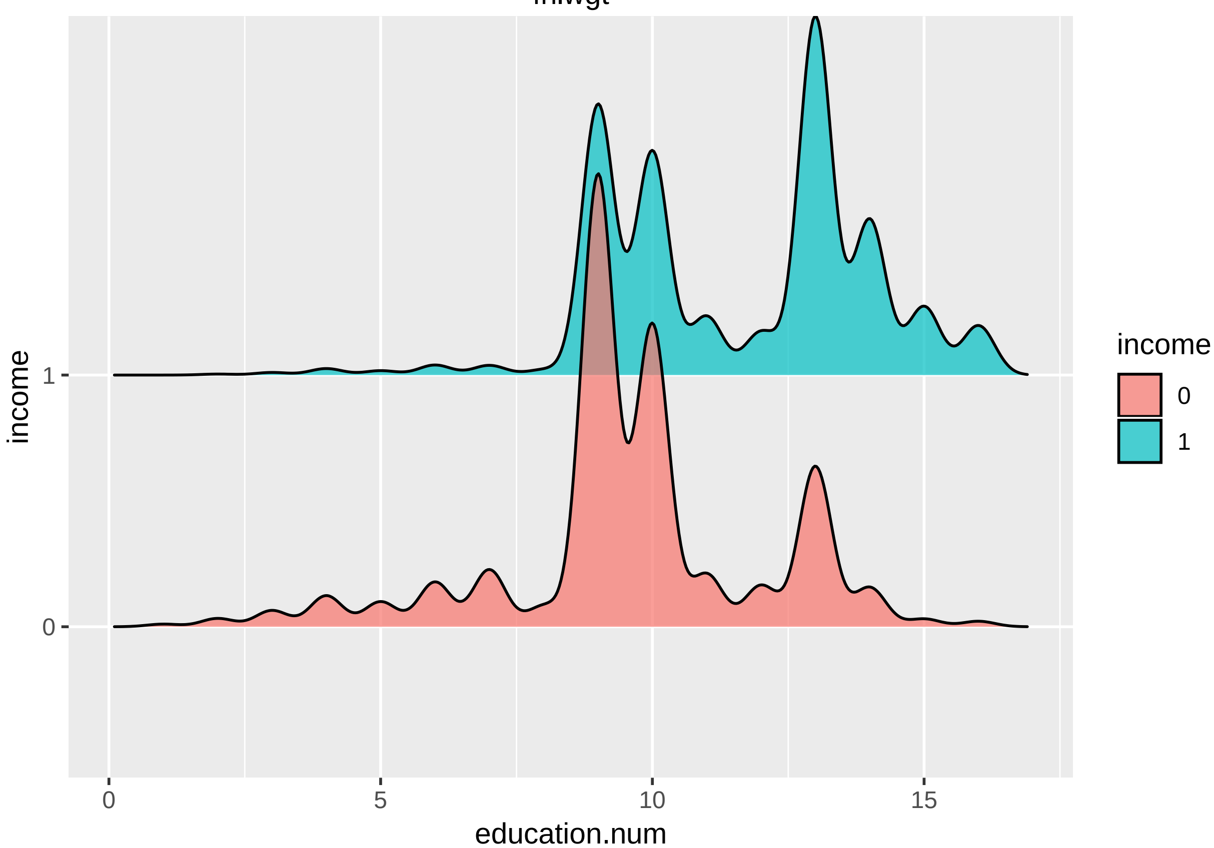



```
## Picking joint bandwidth of 1.62
```

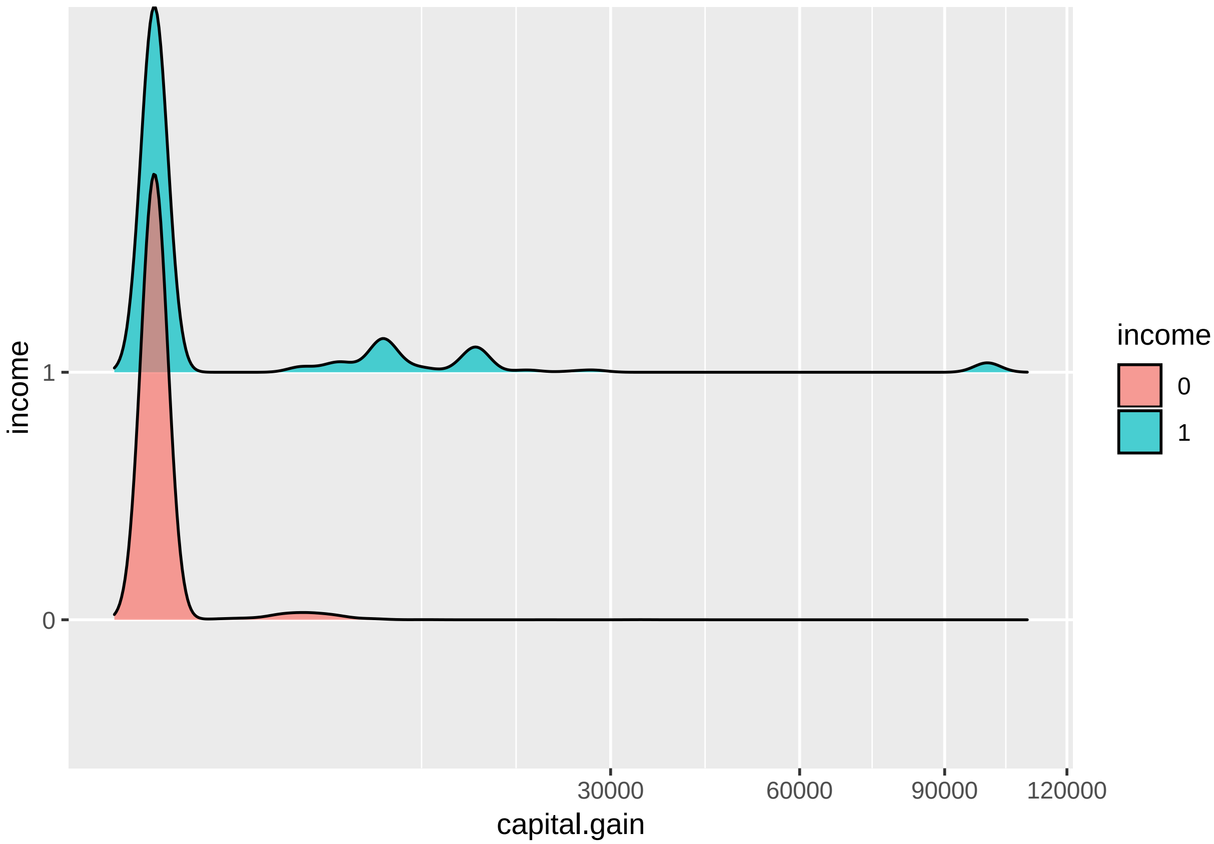



```
## Picking joint bandwidth of 14.1
```

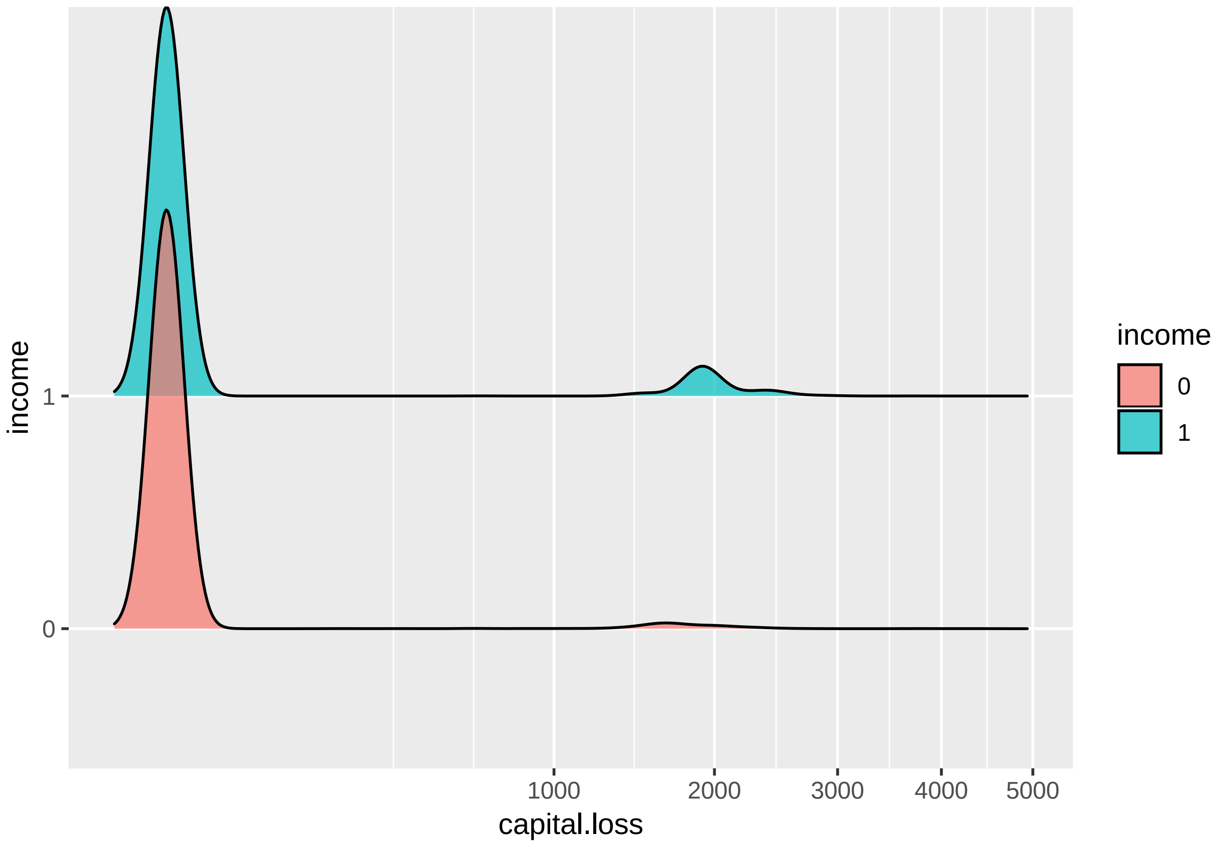


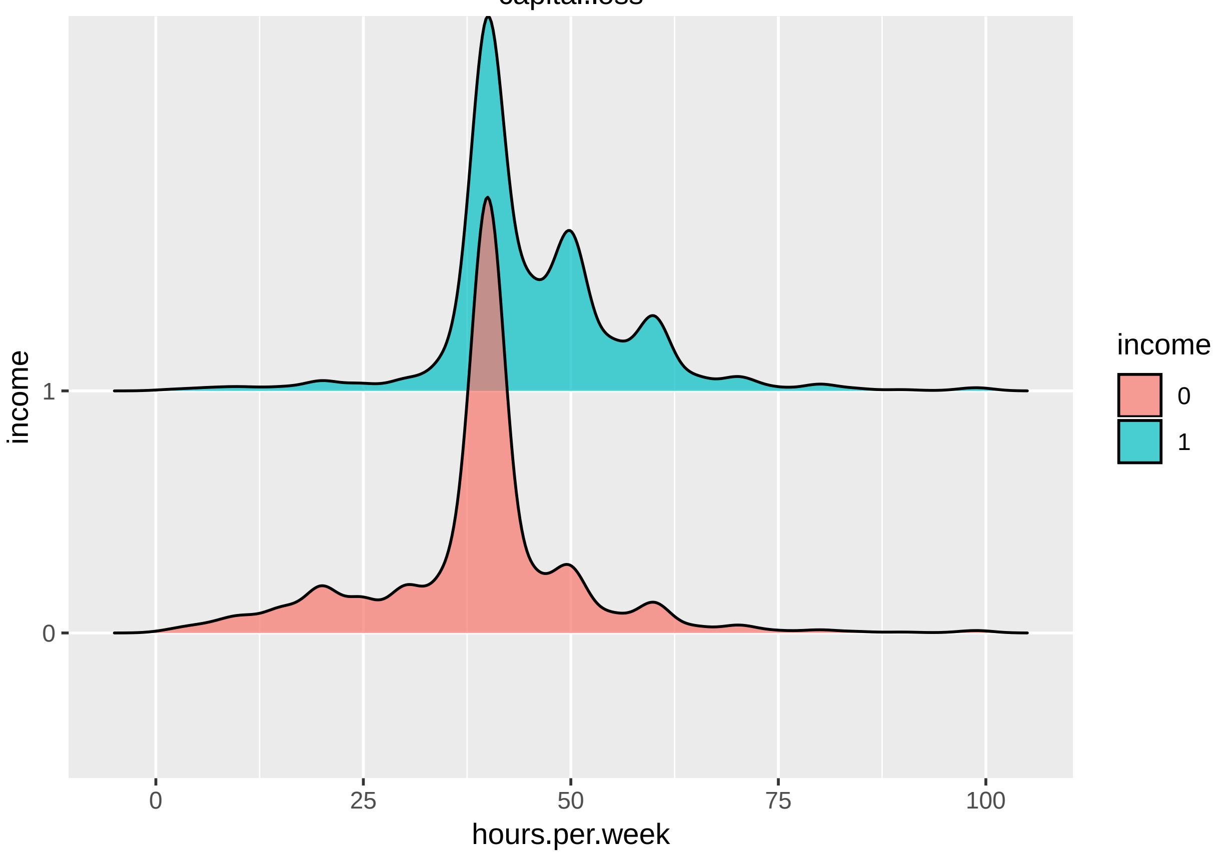



```
## Picking joint bandwidth of 5.05
```




```
## Picking joint bandwidth of 1.42
```

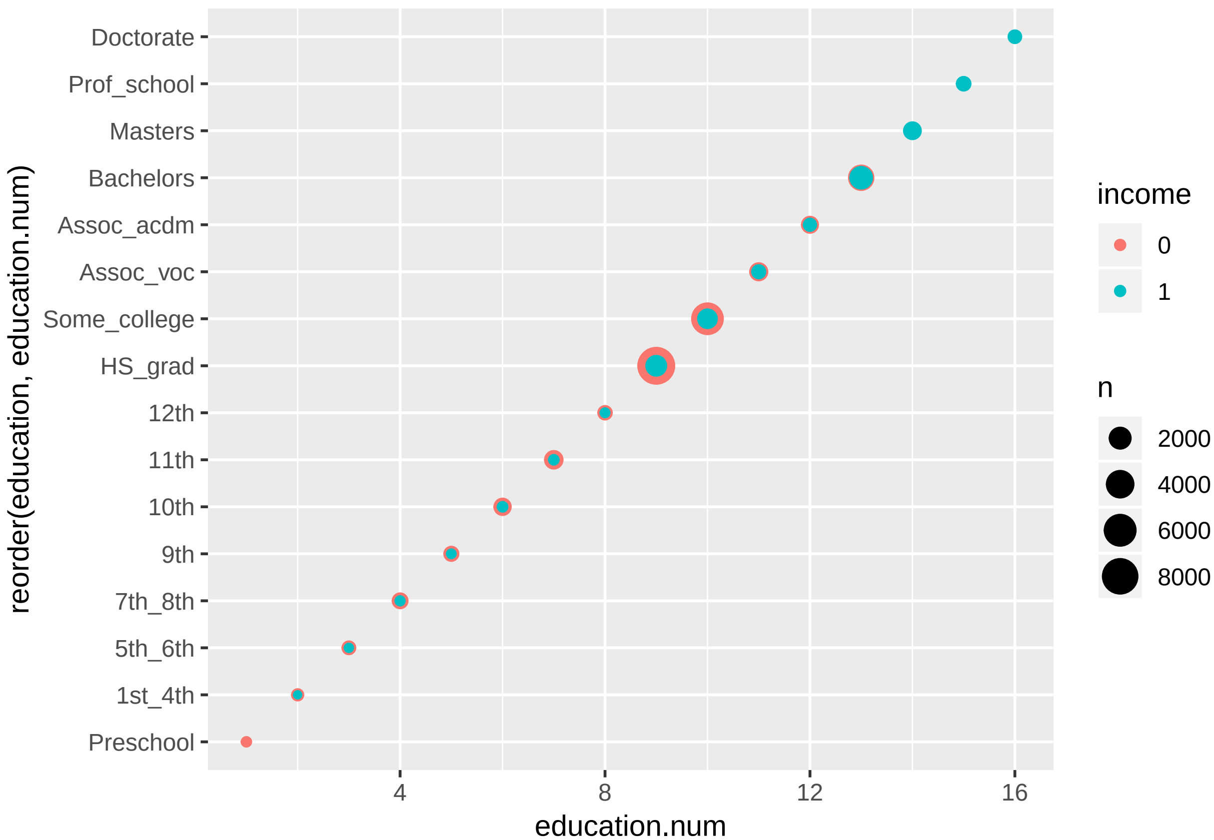




The capital gain could be divided into categories.

0.3.2.2.2 Create scatter plots of the relationship between education and education num.

```
adult%>%
  group_by(education, education.num, income) %>%
  summarize(n=n()) %>%
  arrange(education.num) %>%
  ggplot(aes(x = education.num, y = reorder(education, education.num), color = income))+
  geom_point(aes(size = n))
```



Education was excluded from variables because education and education.num were in a linear relationship.

0.3.3 Generate the train set and the test set

```
# Education was excluded from variables.
adult <-
  adult %>%
  select(-education)

# Test set will be 10% of adult data.
set.seed(1)
test_index <- createDataPartition(y = adult$income, times = 1, p = 0.1, list = FALSE)
train <- adult[-test_index,]
test <- adult[test_index,]
```

0.3.4 Comparison of CART and Random Forests

0.3.4.1 CART

CART used rpart and partykit package.

```
set.seed(1)
fit_rpart <- rpart(income ~ ., data = train)
plot(as.party(fit_rpart))

yhat_rpart = factor(predict(fit_rpart, test, type = "class"), levels = levels(test$income))

confusionMatrix(yhat_rpart,test$income)
```

0.3.4.2 Random Forests

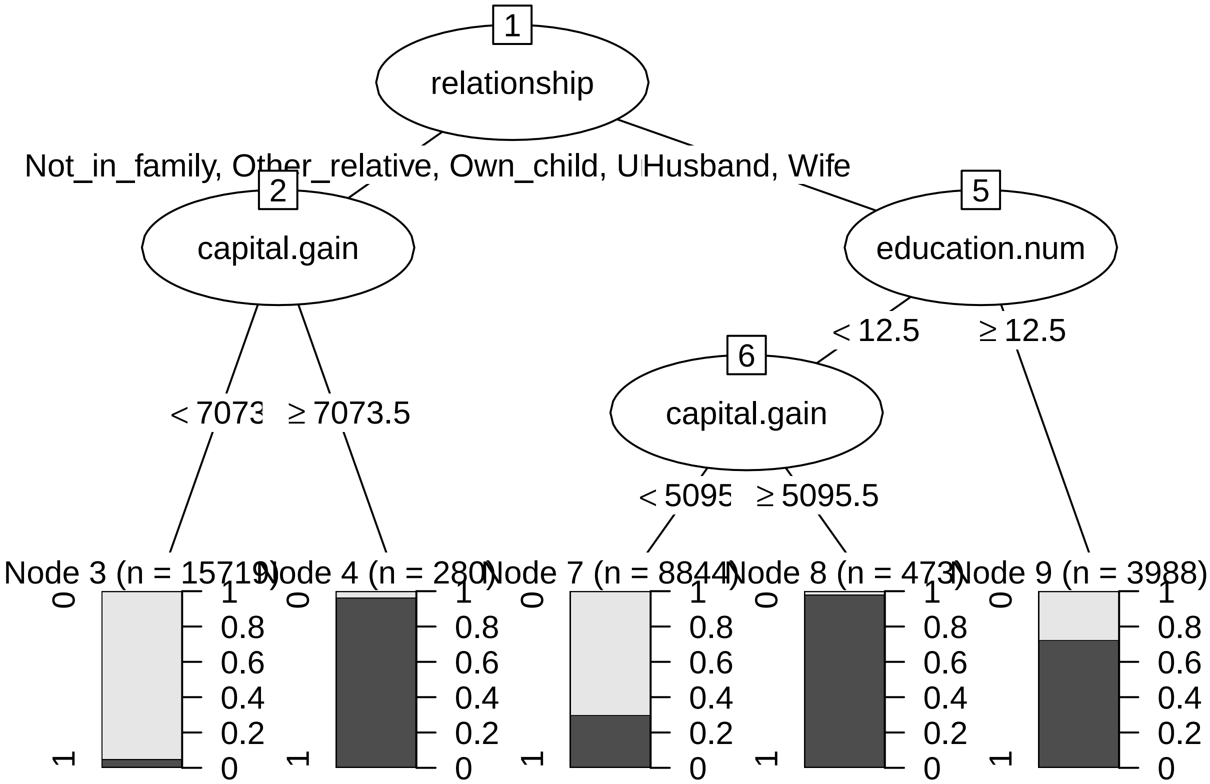
Random Forests used the randomForest package.

```
set.seed(1)
fit_rf <- randomForest(income ~ ., data = train)
fit_rf
varImpPlot(fit_rf)

yhat_rf = factor(predict(fit_rf, test), levels = levels(test$income))

confusionMatrix(yhat_rf, test$income)
```

0.4 Results section

0.4.1 CART

```
set.seed(1)
fit_rpart <- rpart(income ~ ., data = train)
plot(as.party(fit_rpart))
```




```r
yhat_rpart = factor(predict(fit_rpart, test, type = "class"), levels = levels(test$income))

confusionMatrix(yhat_rpart,test$income)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2340  386
##          1  132  399
##
##                Accuracy : 0.841
##                  95% CI : (0.8279, 0.8534)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5113
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9466
##             Specificity : 0.5083
##          Pos Pred Value : 0.8584
##          Neg Pred Value : 0.7514
##              Prevalence : 0.7590
##          Detection Rate : 0.7185
##    Detection Prevalence : 0.8370
##       Balanced Accuracy : 0.7274
##
##        'Positive' Class : 0
```

```
##
```

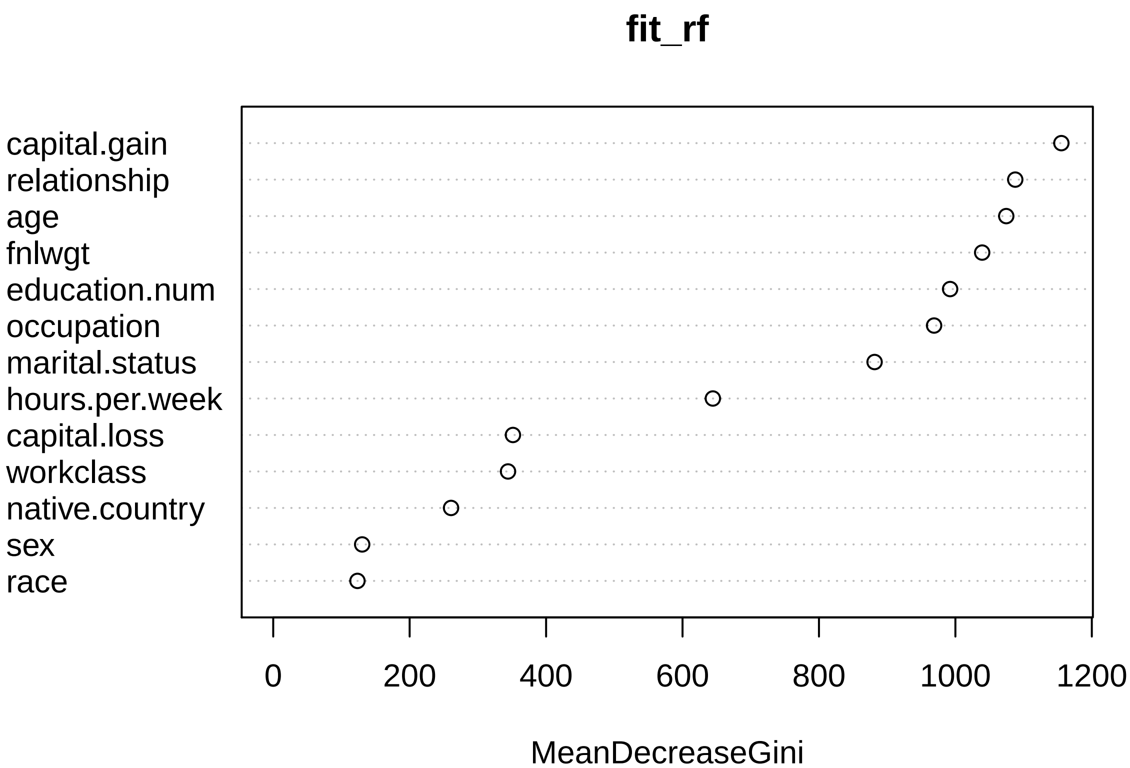
In CART, relationship, capital gain and education number were selected.

0.4.2 Random Forests

```r
set.seed(1)
fit_rf <- randomForest(income ~ ., data = train)
fit_rf
```

```
##
## Call:
##  randomForest(formula = income ~ ., data = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 13.37%
## Confusion matrix:
##       0    1 class.error
## 0 20842 1406  0.06319669
## 1  2512 4544  0.35600907
```

```r
varImpPlot(fit_rf)
```




```r
yhat_rf = factor(predict(fit_rf, test), levels = levels(test$income))

confusionMatrix(yhat_rf, test$income)
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction    0    1
##          0 2308  268
##          1  164  517
##
##                Accuracy : 0.8674
##                  95% CI : (0.8552, 0.8788)
##     No Information Rate : 0.759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6203
##  Mcnemar's Test P-Value : 7.211e-07
##
##             Sensitivity : 0.9337
##             Specificity : 0.6586
##          Pos Pred Value : 0.8960
##          Neg Pred Value : 0.7592
##              Prevalence : 0.7590
##          Detection Rate : 0.7086
##    Detection Prevalence : 0.7909
##       Balanced Accuracy : 0.7961
##
##        'Positive' Class : 0
##
```

In addition to the CART results, age, fnlwgt, occupation, and marital.

0.5 Conclusion section

0.5.1 Why CART and Random Forests

This data set contains both numerical data and categorical data, so we selected a method that can support both.

CART can be interpreted, and Random Forests can be visualized using varImpPlot.

0.5.2 About Accuracy, Sensitivity, and Specificity

Random Forests is more accurate, but CART is more sensitive. The accuracy may be due to differences in specificity.

method	accuracy	Sensitivity	Specificity
CART	0.8409579	0.9466019	0.5082803
Random Forests	0.8673626	0.9336570	0.6585987

0.5.3 Results

The accuracy of CART was 0.8409579, and that of Random Forests was 0.8673626.

In CART, relationship, capital gain and education number were selected.

In addition to the CART results, age, fnlwgt, occupation, and marital.