

DatasetProcessingCode.R

rharidas

2021-11-26

```
#####
# Create Census pay train set, and validation set
#####

# Note: this process could take a couple of minutes

if (!require(tidyverse))
  install.packages("tidyverse", repos = "http://cran.us.r-project.org")

## Loading required package: tidyverse

## -- Attaching packages ----- tidyverse 1.3.1 --

## v ggplot2 3.3.5      v purrr  0.3.4
## v tibble  3.1.2      v dplyr  1.0.7
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1

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

if (!require(caret))
  install.packages("caret", repos = "http://cran.us.r-project.org")

## Loading required package: caret

## Loading required package: lattice

##
## Attaching package: 'caret'

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

if (!require(data.table))
  install.packages("data.table", repos = "http://cran.us.r-project.org")
```

```

## Loading required package: data.table

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

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

if (!require(dplyr))
  install.packages("dplyr", repos = "http://cran.us.r-project.org")

if (!require(gridExtra))
  install.packages("gridExtra", repos = "http://cran.us.r-project.org")

## Loading required package: gridExtra

##
## Attaching package: 'gridExtra'

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

if (!require(kableExtra))
  install.packages("kableExtra", repos = "http://cran.us.r-project.org")

## Loading required package: kableExtra

##
## Attaching package: 'kableExtra'

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

if (!require(epiDisplay))
  install.packages("epiDisplay")

## Loading required package: epiDisplay

## Loading required package: foreign

## Loading required package: survival

```

```
##
## Attaching package: 'survival'

## The following object is masked from 'package:caret':
##
##      cluster

## Loading required package: MASS

##
## Attaching package: 'MASS'

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

## Loading required package: nnet

##
## Attaching package: 'epiDisplay'

## The following object is masked from 'package:lattice':
##
##      dotplot

## The following object is masked from 'package:ggplot2':
##
##      alpha
```

```
library(tidyverse)
library(caret)
library(data.table)
library(dplyr)
library(gridExtra)
library(kableExtra)
library(epiDisplay)

# Adult Census Income
# https://www.kaggle.com/uciml/adult-census-income

#download the dataset from the staging github location
dl <- tempfile()
download.file("https://github.com/rajeshharidas/havardxwork2/raw/main/adult.csv.zip",
             dl)

#read all the data into R dataset
adultpay <-
  fread(
    text = gsub(",", "\t", readLines(unzip(dl, "adult.csv"))),
    col.names = c(
      "age",
```

```

    "workclass",
    "fnlwgt",
    "education",
    "education.num",
    "marital.status",
    "occupation",
    "relationship",
    "race",
    "sex",
    "capital.gain",
    "capital.loss",
    "hours.per.week",
    "native.country",
    "income"
  )
)

#Keep only USA data
#Remove '?' from the work class and rename it to class, and finally remove workclass
#Rename all columns with a '.' in it
#Remove capital gain and loss column
#remove non-alphanumeric character from column data
#rename the label for below and above 50K income
adultpayclean <-
  adultpay %>% filter (native.country == 'United-States') %>%
  mutate (class = ifelse(workclass == '?', 'Unknown', str_replace_all(workclass, "-", ""))) %>%
  dplyr::select(-workclass, -capital.gain, -capital.loss) %>%
  rename(
    c(
      eduyears = education.num,
      maritalstatus = marital.status,
      hoursperweek = hours.per.week,
      native = native.country
    )
  ) %>%
  mutate (maritalstatus = ifelse(
    maritalstatus == '?',
    'Unknown',
    str_replace_all(maritalstatus, "-", "")
  )) %>%
  mutate (occupation = ifelse(
    occupation == '?',
    'Unknown',
    str_replace_all(occupation, "-", "")
  )) %>%
  mutate (education = ifelse(education == '?', 'Unknown', str_replace_all(education, "-", ""))) %>%
  mutate (relationship = ifelse(
    relationship == '?',
    'Unknown',
    str_replace_all(relationship, "-", "")
  )) %>%
  mutate (native = ifelse(native == '?', 'Unknown', str_replace_all(native, "-", ""))) %>%
  mutate (income = ifelse(

```

```

    income == '?',
    'Unknown',
    str_replace_all(income, "<=50K", "AtBelow50K")
  )) %>%
mutate (income = ifelse(
  income == '?',
  'Unknown',
  str_replace_all(income, ">50K", "Above50K")
))

# R 4.0 or later:
#convert all the character labels to factors
adultpayclean <-
  as.data.frame(adultpayclean) %>% mutate(
    education = as.factor(education),
    maritalstatus = as.factor(maritalstatus),
    occupation = as.factor(occupation),
    relationship = as.factor(relationship),
    race = as.factor(race),
    sex = as.factor(sex),
    class = as.factor(class),
    income = as.factor(income)
  )

# Validation set will be 10% of adultpay data
set.seed(1, sample.kind = "Rounding") # if using R 3.5 or earlier, use `set.seed(1)`

```

```

## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used

```

```

test_index <-
  createDataPartition(
    y = adultpayclean$income,
    times = 1,
    p = 0.1,
    list = FALSE
  )
adultpayclean_train <- adultpayclean[-test_index, ]
adultpayclean_validation <- adultpayclean[test_index, ]

glimpse(adultpay)

```

```

## Rows: 32,561
## Columns: 15
## $ age          <int> 90, 82, 66, 54, 41, 34, 38, 74, 68, 41, 45, 38, 52, 32, ~
## $ workclass    <chr> "?", "Private", "?", "Private", "Private", "Private", "~
## $ fnlwgt       <int> 77053, 132870, 186061, 140359, 264663, 216864, 150601, ~
## $ education    <chr> "HS-grad", "HS-grad", "Some-college", "7th-8th", "Some--
## $ education.num <int> 9, 9, 10, 4, 10, 9, 6, 16, 9, 10, 16, 15, 13, 14, 16, 1~
## $ marital.status <chr> "Widowed", "Widowed", "Widowed", "Divorced", "Separated~
## $ occupation   <chr> "?", "Exec-managerial", "?", "Machine-op-inspct", "Prof~
## $ relationship <chr> "Not-in-family", "Not-in-family", "Unmarried", "Unmarri~

```

```
## $ race      <chr> "White", "White", "Black", "White", "White", "White", "~
## $ sex       <chr> "Female", "Female", "Female", "Female", "Female", "Fema~
## $ capital.gain <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ capital.loss <int> 4356, 4356, 4356, 3900, 3900, 3770, 3770, 3683, 3683, 3~
## $ hours.per.week <int> 40, 18, 40, 40, 40, 45, 40, 20, 40, 60, 35, 45, 20, 55,~
## $ native.country <chr> "United-States", "United-States", "United-States", "Uni~
## $ income     <chr> "<=50K", "<=50K", "<=50K", "<=50K", "<=50K", "<=50K", "~
```

```
glimpse(adultpayclean)
```

```
## Rows: 29,170
## Columns: 13
## $ age      <int> 90, 82, 66, 54, 41, 34, 38, 74, 68, 45, 38, 52, 32, 51, ~
## $ fnlwgt   <int> 77053, 132870, 186061, 140359, 264663, 216864, 150601, 8~
## $ education <fct> HSgrad, HSgrad, Somecollege, 7th8th, Somecollege, HSgrad~
## $ eduyears <int> 9, 9, 10, 4, 10, 9, 6, 16, 9, 16, 15, 13, 14, 16, 15, 7,~
## $ maritalstatus <fct> Widowed, Widowed, Widowed, Divorced, Separated, Divorced~
## $ occupation <fct> Unknown, Execmanagerial, Unknown, Machineopinspct, Profs~
## $ relationship <fct> Notinfamily, Notinfamily, Unmarried, Unmarried, Ownchild~
## $ race      <fct> White, White, Black, White, White, White, White, White, ~
## $ sex       <fct> Female, Female, Female, Female, Female, Female, Male, Fe~
## $ hoursperweek <int> 40, 18, 40, 40, 40, 45, 40, 20, 40, 35, 45, 20, 55, 40, ~
## $ native    <chr> "UnitedStates", "UnitedStates", "UnitedStates", "UnitedS~
## $ income    <fct> AtBelow50K, AtBelow50K, AtBelow50K, AtBelow50K, AtBelow5~
## $ class     <fct> Unknown, Private, Unknown, Private, Private, Private, Pr~
```

```
dim(adultpayclean)
```

```
## [1] 29170    13
```

```
dim(adultpayclean_train)
```

```
## [1] 26252    13
```

```
dim(adultpayclean_validation)
```

```
## [1] 2918    13
```

```
summary(adultpayclean)
```

```
##      age      fnlwgt      education      eduyears
## Min.   :17.00   Min.   : 12285   HSgrad      :9702   Min.   : 1.00
## 1st Qu.:28.00   1st Qu.: 115895   Somecollege:6740   1st Qu.: 9.00
## Median :37.00   Median : 176730   Bachelors  :4766   Median :10.00
## Mean   :38.66   Mean   : 187069   Masters    :1527   Mean   :10.17
## 3rd Qu.:48.00   3rd Qu.: 234139   Assocvoc   :1289   3rd Qu.:12.00
## Max.   :90.00   Max.   :1484705   11th       :1067   Max.   :16.00
##                               (Other)    :4079
##      maritalstatus      occupation      relationship
## Divorced      : 4162   Execmanagerial:3735   Husband      :11861
```

```

## MarriedAFspouse      : 23   Profspecialty :3693   Notinfamily : 7528
## Marriedcivspouse     :13368  Craftrepair  :3685   Otherrelative: 696
## Marriedspouseabsent: 253   Admclerical  :3449   Ownchild     : 4691
## Nevermarried         : 9579   Sales        :3364   Unmarried    : 3033
## Separated            : 883   Otherservice  :2777   Wife         : 1361
## Widowed              : 902   (Other)      :8467
##
##           race           sex           hoursperweek       native
## Amer-Indian-Eskimo: 296   Female: 9682   Min. : 1.00   Length:29170
## Asian-Pac-Islander: 292   Male :19488   1st Qu.:40.00   Class :character
## Black                : 2832               Median :40.00   Mode  :character
## Other                : 129               Mean   :40.45
## White                :25621               3rd Qu.:45.00
##                               Max.   :99.00
##
##           income           class
## Above50K : 7171   Private :20135
## AtBelow50K:21999   Selfempnotinc: 2313
##                               Localgov : 1956
##                               Unknown  : 1659
##                               Stategov : 1210
##                               Selfempinc : 991
##                               (Other)  : 906

```

```
summary(adultpayclean_train)
```

```

##           age           fnlwgt           education           eduyears
## Min. :17.00   Min. : 12285   HSgrad :8716   Min. : 1.00
## 1st Qu.:28.00   1st Qu.: 116052   Somecollege:6071   1st Qu.: 9.00
## Median :37.00   Median : 176904   Bachelors :4318   Median :10.00
## Mean :38.66   Mean : 187117   Masters :1366   Mean :10.17
## 3rd Qu.:48.00   3rd Qu.: 234099   Assocvoc :1164   3rd Qu.:12.00
## Max. :90.00   Max. :1484705   11th : 948   Max. :16.00
##                               (Other) :3669
##
##           maritalstatus           occupation           relationship
## Divorced : 3757   Execmanagerial:3382   Husband :10674
## MarriedAFspouse : 21   Profspecialty :3318   Notinfamily : 6803
## Marriedcivspouse :12033   Craftrepair :3300   Otherrelative: 629
## Marriedspouseabsent: 229   Admclerical :3095   Ownchild : 4213
## Nevermarried : 8616   Sales :3035   Unmarried : 2707
## Separated : 792   Otherservice :2490   Wife : 1226
## Widowed : 804   (Other) :7632
##
##           race           sex           hoursperweek       native
## Amer-Indian-Eskimo: 261   Female: 8708   Min. : 1.00   Length:26252
## Asian-Pac-Islander: 265   Male :17544   1st Qu.:40.00   Class :character
## Black                : 2537               Median :40.00   Mode  :character
## Other                : 119               Mean   :40.47
## White                :23070               3rd Qu.:45.00
##                               Max.   :99.00
##
##           income           class
## Above50K : 6453   Private :18093
## AtBelow50K:19799   Selfempnotinc: 2087
##                               Localgov : 1777
##                               Unknown : 1494

```

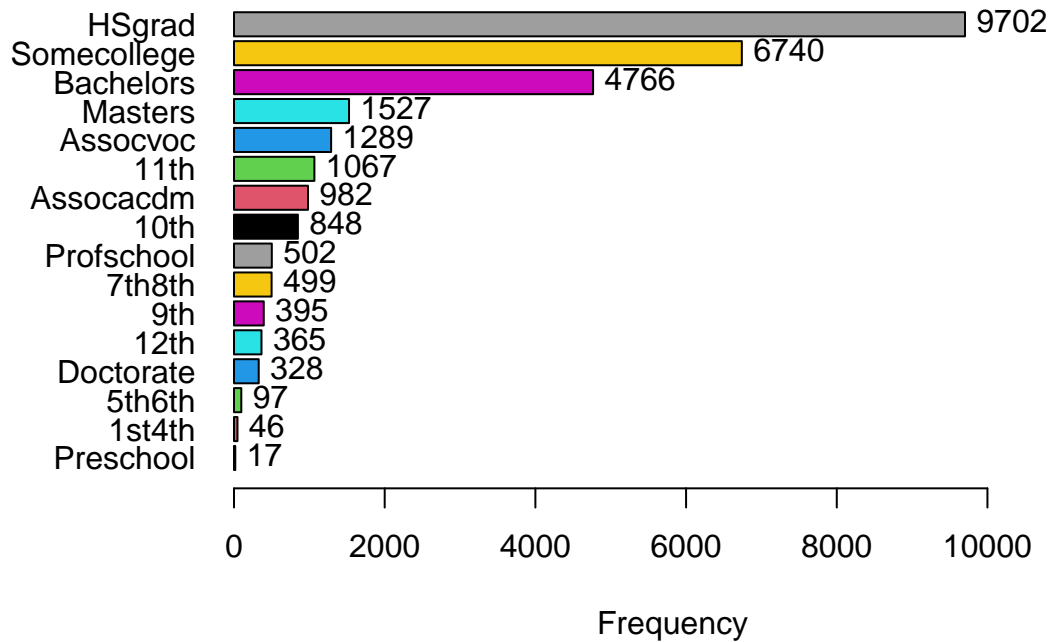
Dataset	Number of Rows	Number of Columns
train	26252	13
validation	2918	13

```
##          Stategov      : 1081
##          Selfempinc    :  910
##          (Other)       :  810
```

```
tribble(
  ~"Dataset",      ~"Number of Rows",    ~"Number of Columns",
  #--              /--                  /----
  "train",         nrow(adultpayclean_train), ncol(adultpayclean_train),
  "validation",    nrow(adultpayclean_validation), ncol(adultpayclean_validation)
) %>% knitr::kable() %>% kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

```
tab1(adultpayclean$education, sort.group = "decreasing", cum.percent = TRUE)
```

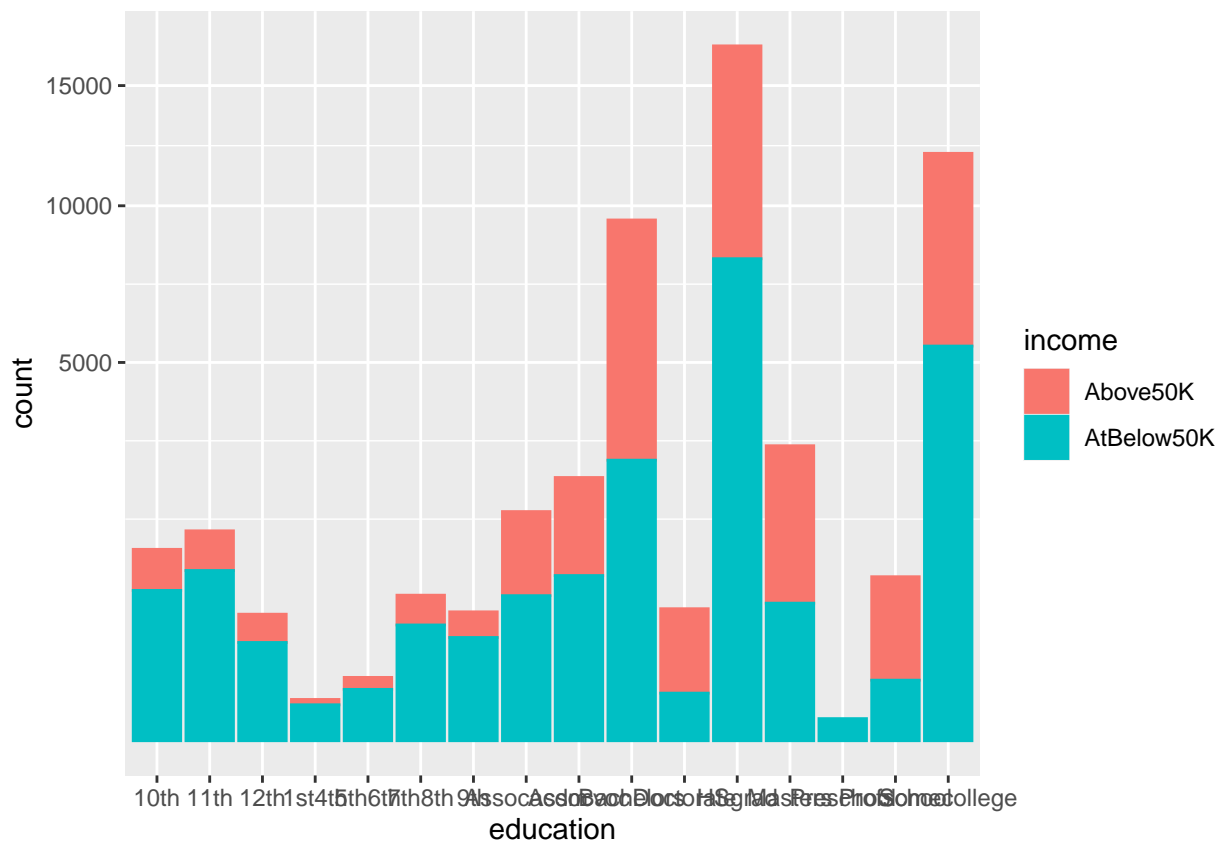
Distribution of adultpayclean\$education



```
## adultpayclean$education :
##          Frequency Percent Cum. percent
## HSgrad          9702    33.3         33.3
## Somecollege      6740    23.1         56.4
## Bachelors        4766    16.3         72.7
## Masters          1527     5.2         77.9
```

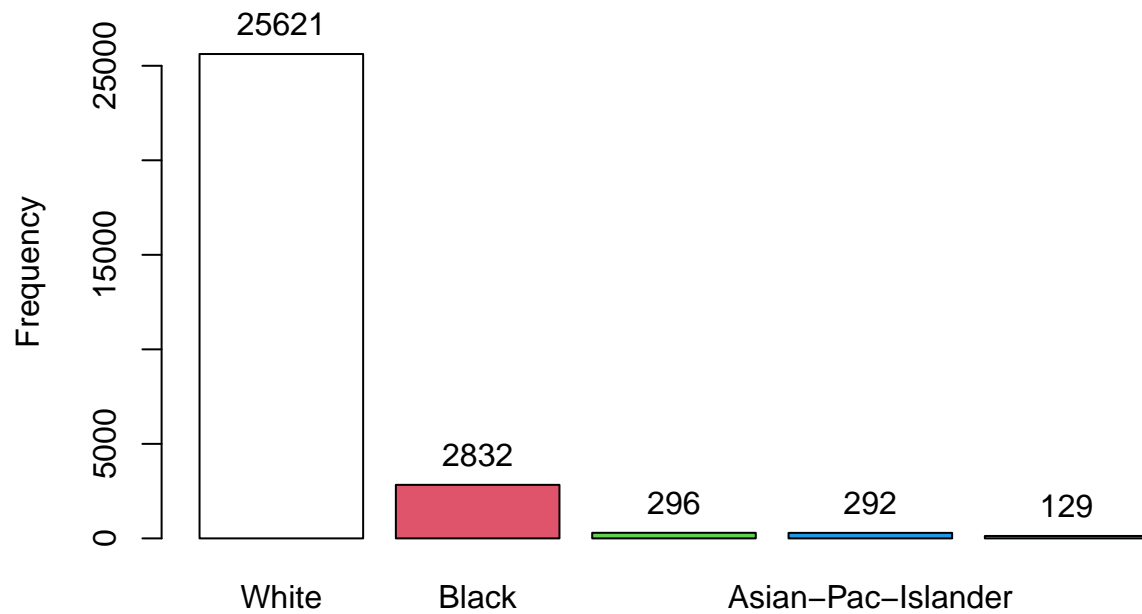

## Assocvoc	1289	4.4	82.4
## 11th	1067	3.7	86.0
## Assocacdm	982	3.4	89.4
## 10th	848	2.9	92.3
## Profschool	502	1.7	94.0
## 7th8th	499	1.7	95.7
## 9th	395	1.4	97.1
## 12th	365	1.3	98.3
## Doctorate	328	1.1	99.5
## 5th6th	97	0.3	99.8
## 1st4th	46	0.2	99.9
## Preschool	17	0.1	100.0
## Total	29170	100.0	100.0

```
adultpayclean %>% group_by(education) %>%
  mutate(n=n()) %>% ggplot() +
  geom_bar(aes(education,col=income,fill=income)) + scale_y_sqrt()
```



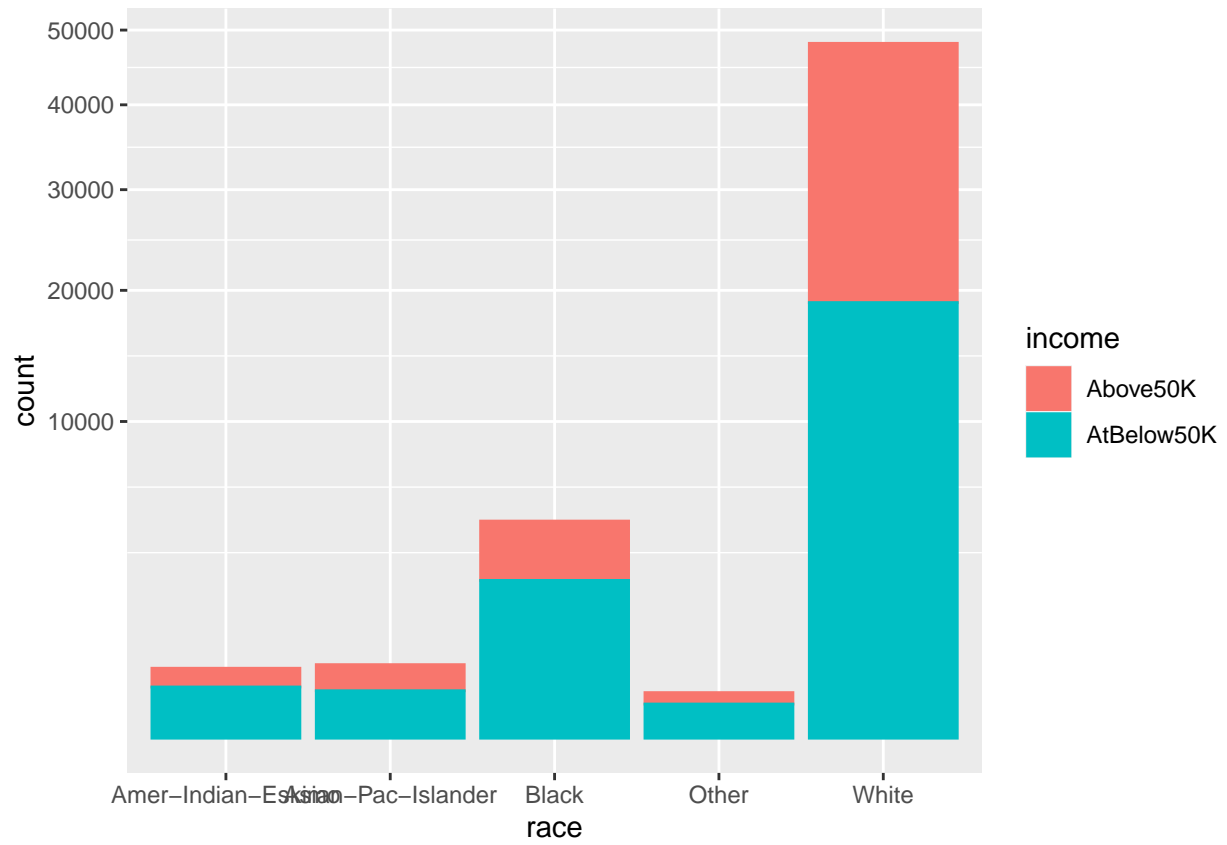
```
tab1(adultpayclean$race, sort.group = "decreasing", cum.percent = TRUE)
```

Distribution of adultpayclean\$race



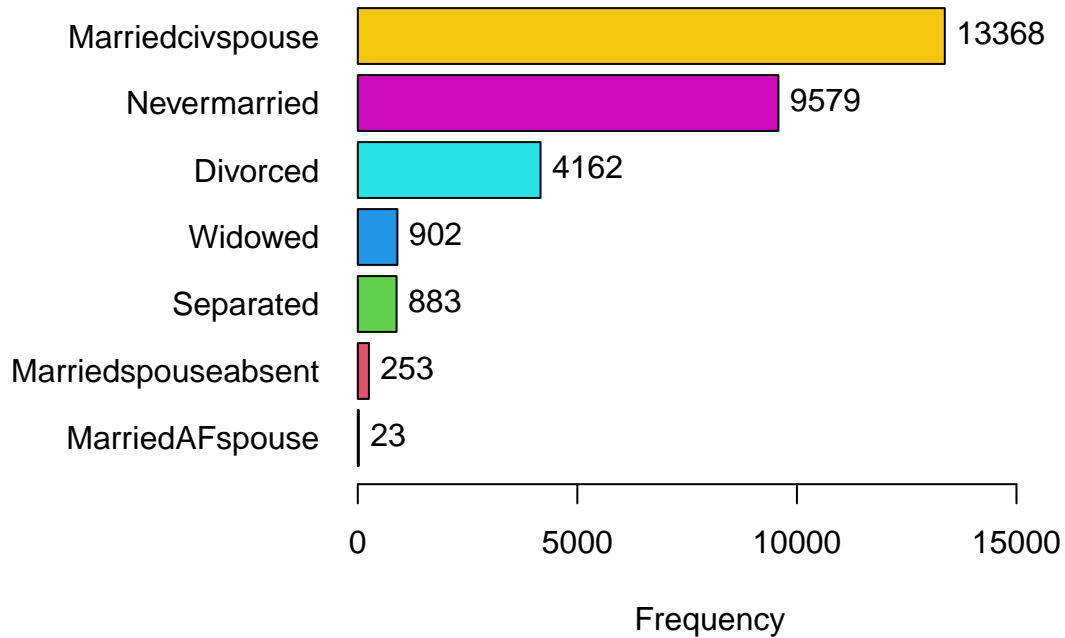
```
## adultpayclean$race :
##               Frequency Percent Cum. percent
## White             25621      87.8         87.8
## Black              2832       9.7         97.5
## Amer-Indian-Eskimo   296       1.0         98.6
## Asian-Pac-Islander   292       1.0         99.6
## Other                129       0.4        100.0
## Total              29170      100.0        100.0
```

```
adultpayclean %>% group_by(race) %>%
  mutate(n=n()) %>% ggplot() +
  geom_bar(aes(race,col=income,fill=income)) + scale_y_sqrt()
```



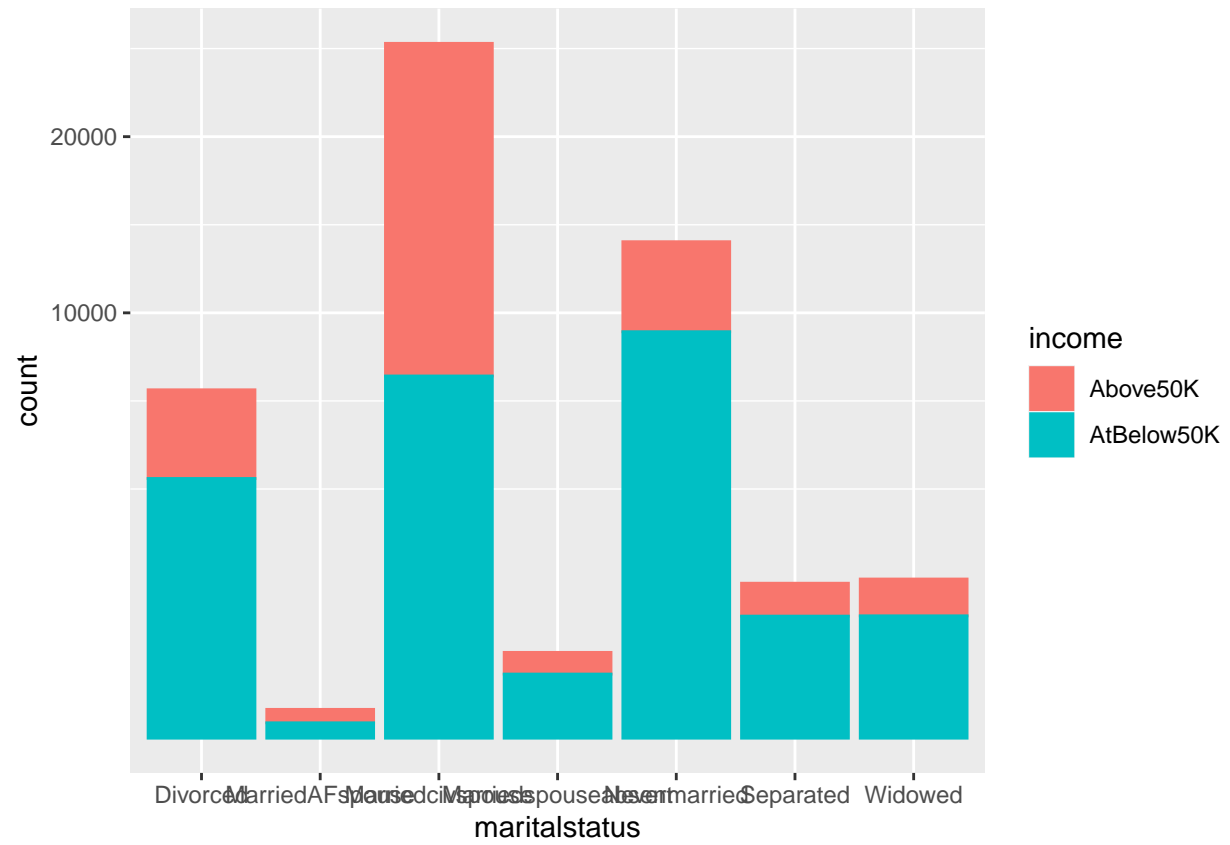
```
tab1(adultpayclean$maritalstatus, sort.group = "decreasing", cum.percent = TRUE)
```

Distribution of adultpayclean\$maritalstatus



```
## adultpayclean$maritalstatus :
##               Frequency Percent Cum. percent
## Marriedcivspouse      13368    45.8         45.8
## Nevermarried          9579    32.8         78.7
## Divorced              4162    14.3         92.9
## Widowed               902     3.1         96.0
## Separated             883     3.0         99.1
## Marriedspouseabsent    253     0.9         99.9
## MarriedAFspouse        23     0.1        100.0
## Total                 29170   100.0        100.0
```

```
adultpayclean %>% group_by(maritalstatus) %>%
  mutate(n=n()) %>% ggplot() +
  geom_bar(aes(maritalstatus,col=income,fill=income)) + scale_y_sqrt()
```



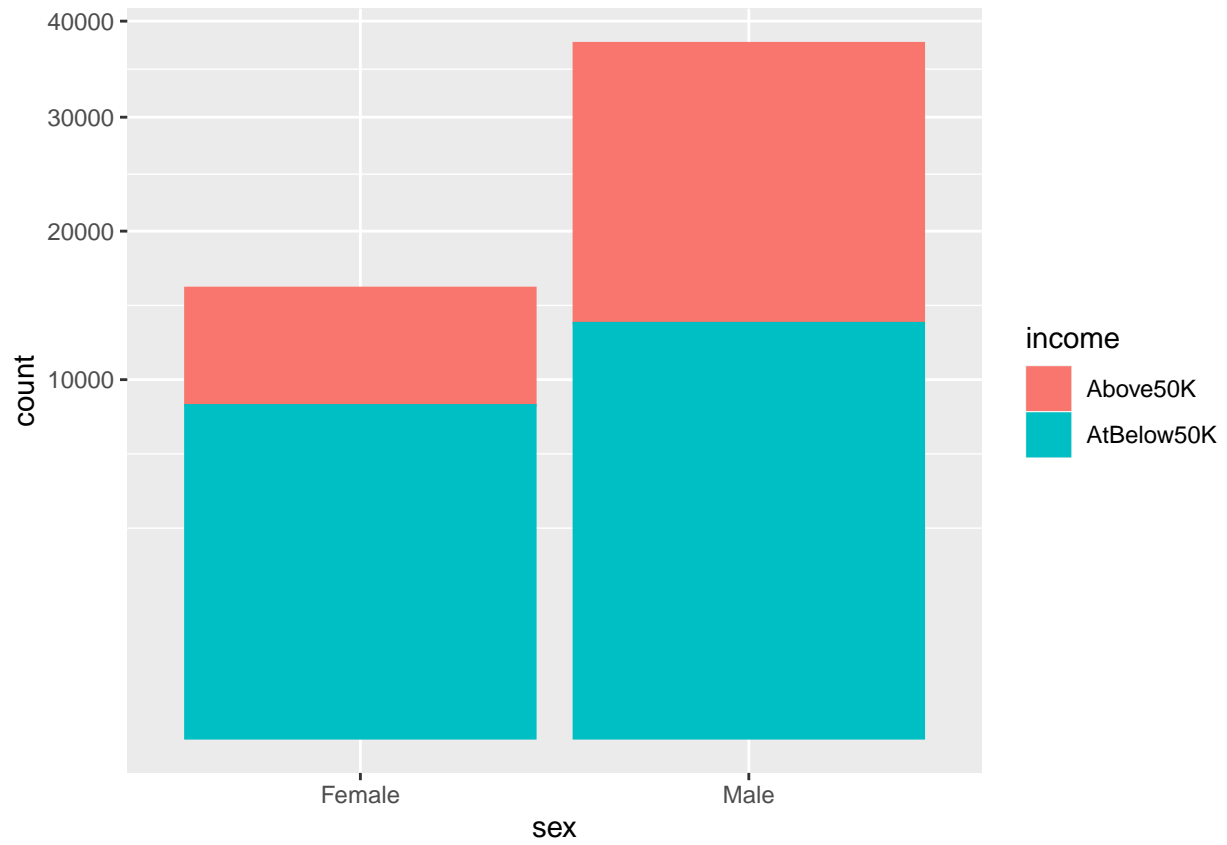
```
tab1(adultpayclean$sex, sort.group = "decreasing", cum.percent = TRUE)
```

Distribution of adultpayclean\$sex



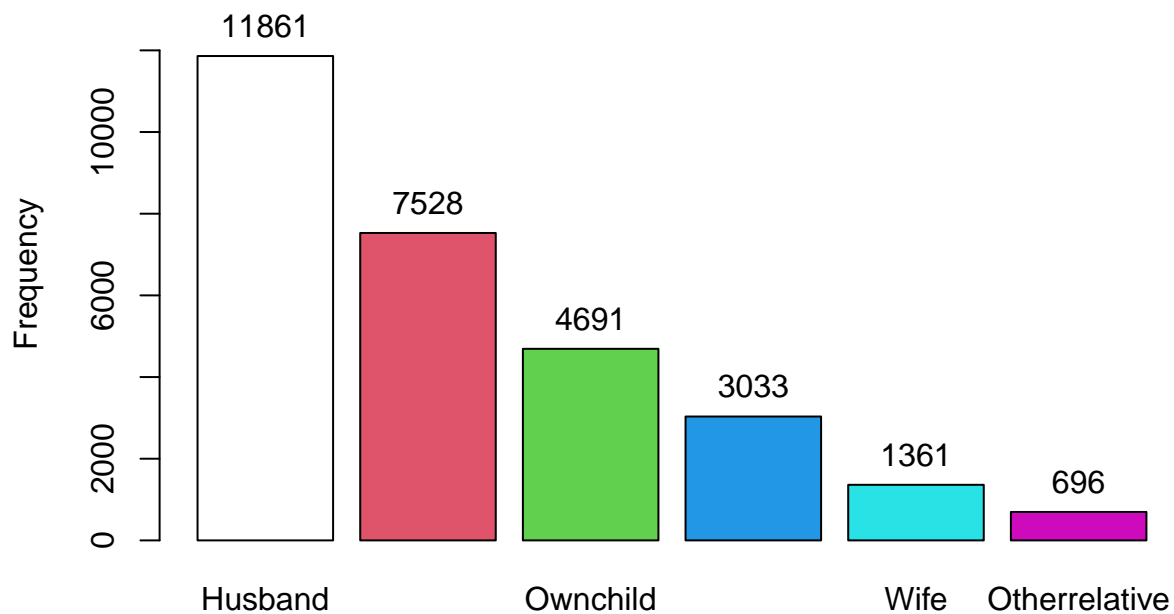
```
## adultpayclean$sex :  
##      Frequency Percent Cum. percent  
## Male      19488     66.8         66.8  
## Female     9682     33.2        100.0  
##   Total     29170    100.0        100.0
```

```
adultpayclean %>% group_by(sex) %>%  
  mutate(n=n()) %>% ggplot() +  
  geom_bar(aes(sex,col=income,fill=income)) + scale_y_sqrt()
```



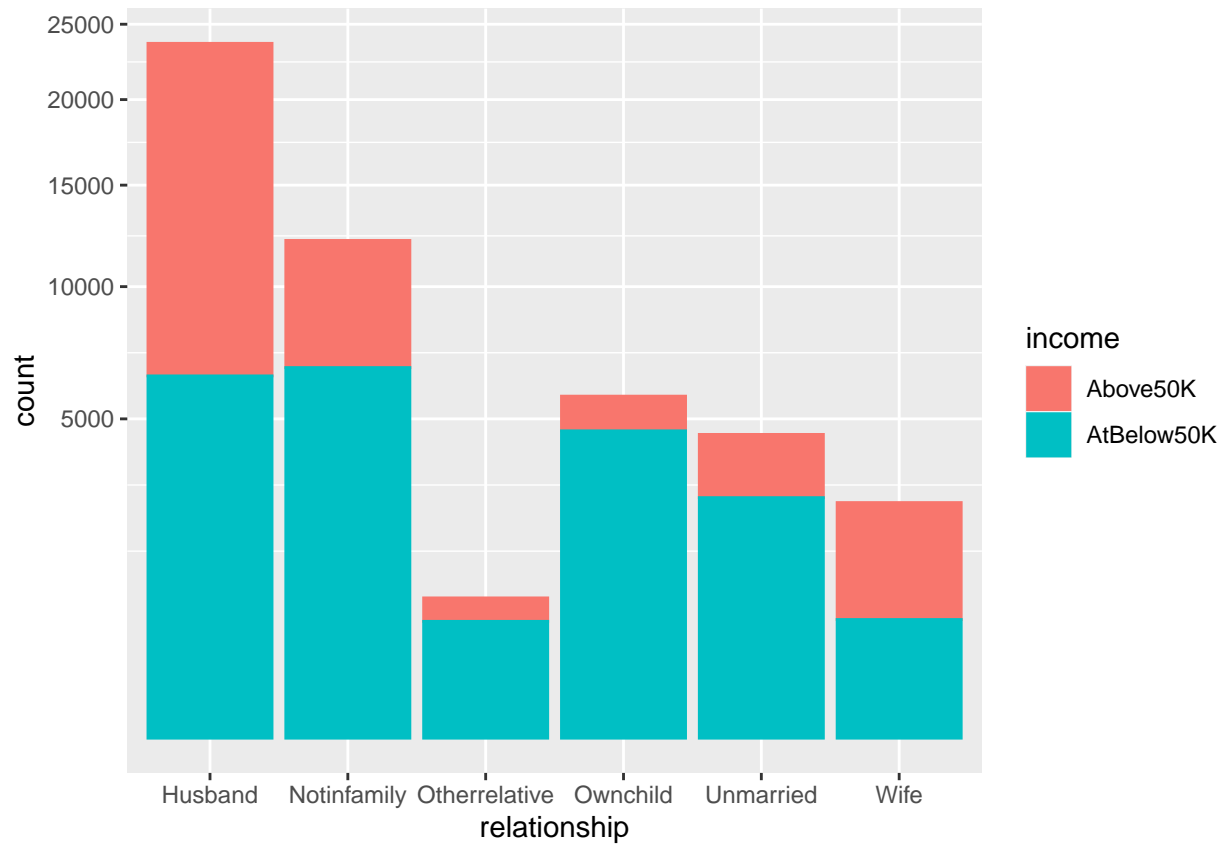
```
tab1(adultpayclean$relationship, sort.group = "decreasing", cum.percent = TRUE)
```

Distribution of adultpayclean\$relationship



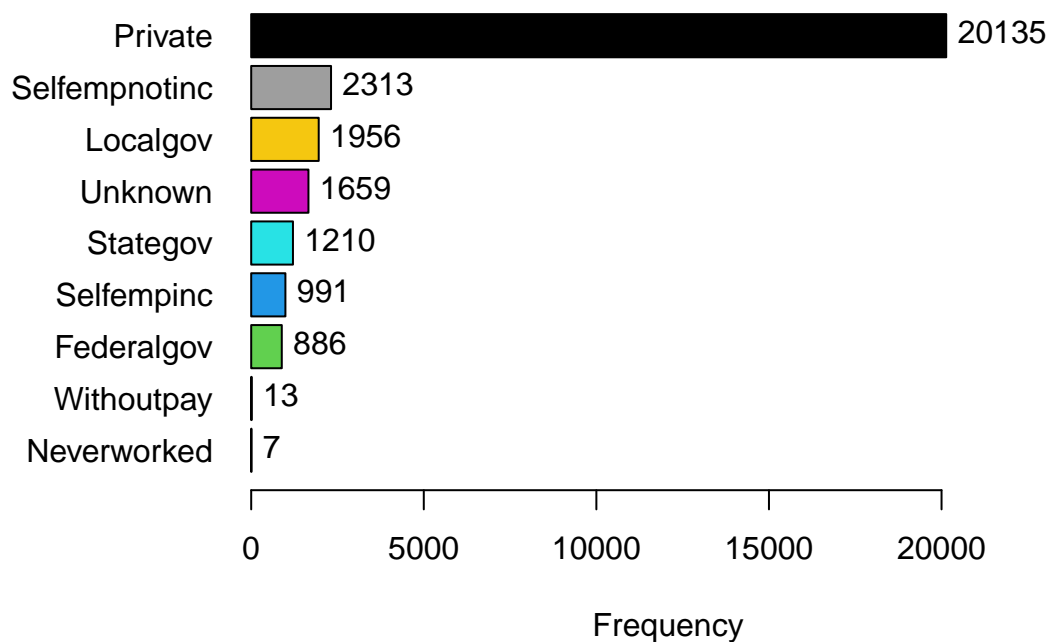
```
## adultpayclean$relationship :
##           Frequency Percent Cum. percent
## Husband           11861     40.7         40.7
## Notinfamily         7528     25.8         66.5
## Ownchild            4691     16.1         82.6
## Unmarried           3033     10.4         92.9
## Wife                1361      4.7         97.6
## Otherrelative         696      2.4        100.0
## Total              29170    100.0        100.0
```

```
adultpayclean %>% group_by(relationship) %>%
  mutate(n=n()) %>% ggplot() +
  geom_bar(aes(relationship,col=income,fill=income)) + scale_y_sqrt()
```

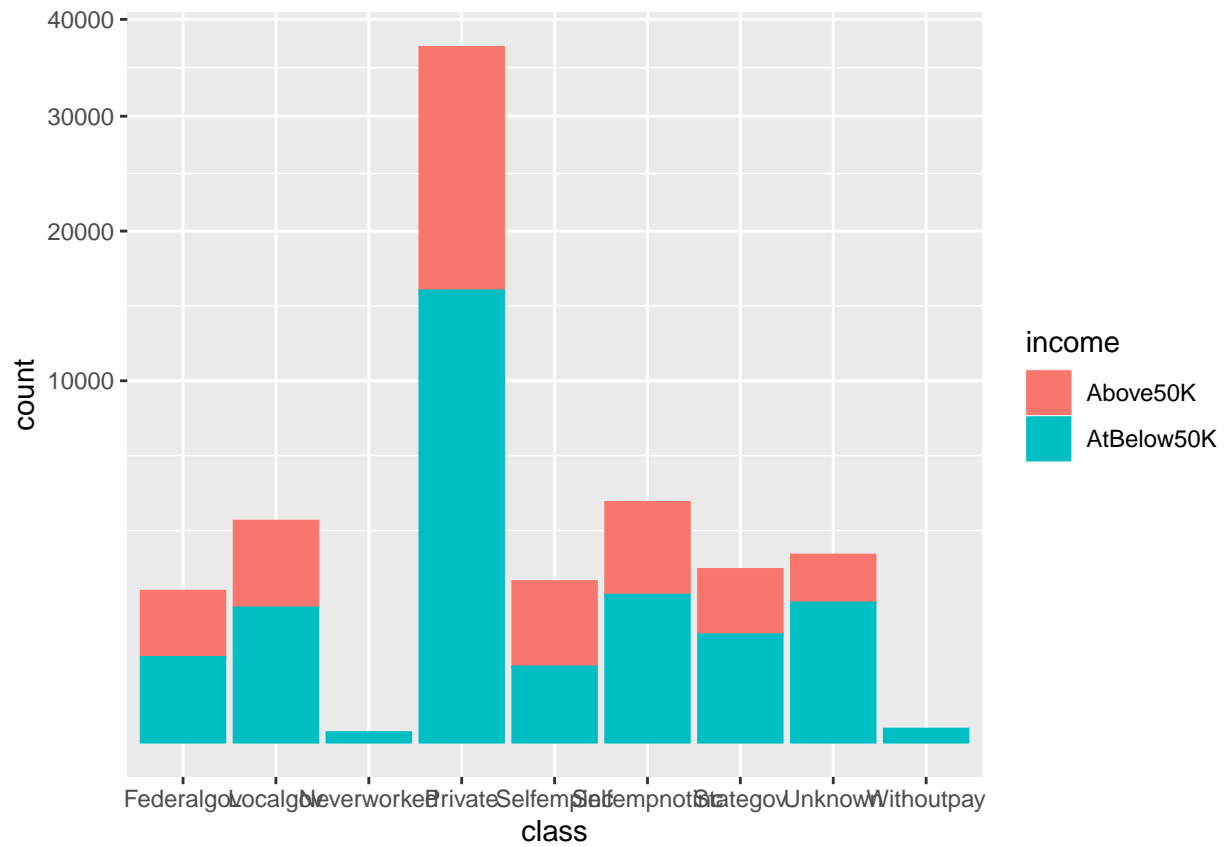
```
tab1(adultpayclean$class, sort.group = "decreasing", cum.percent = TRUE)
```

Distribution of adultpayclean\$class



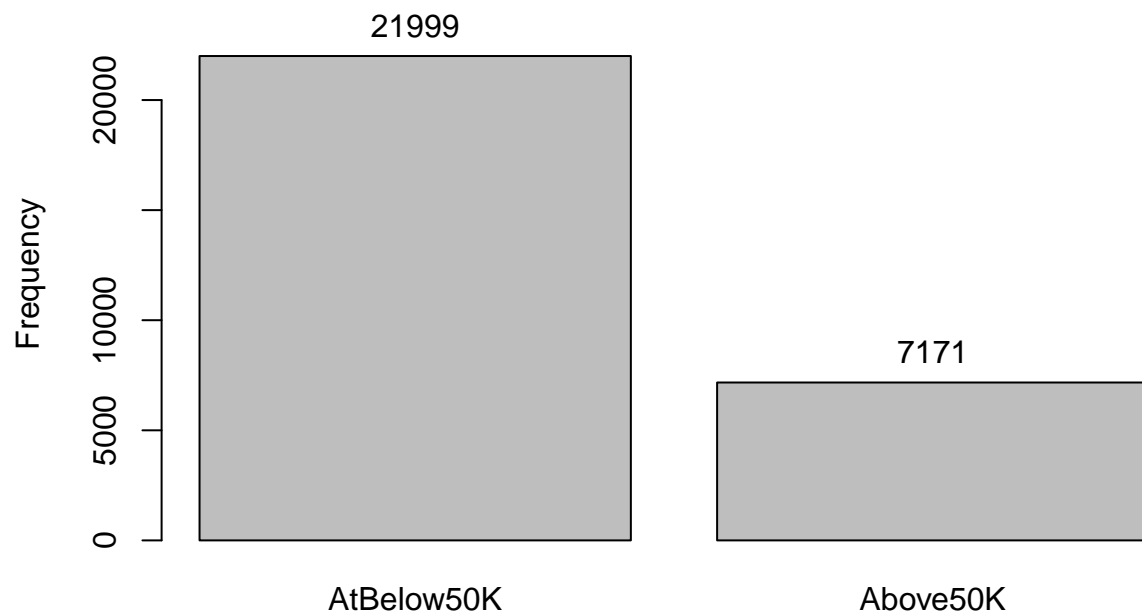
```
## adultpayclean$class :
##           Frequency Percent Cum. percent
## Private           20135      69.0         69.0
## Selfempnotinc       2313       7.9         77.0
## Localgov            1956       6.7         83.7
## Unknown             1659       5.7         89.3
## Stategov            1210       4.1         93.5
## Selfempinc           991       3.4         96.9
## Federalgov           886       3.0         99.9
## Withoutpay           13       0.0        100.0
## Neverworked           7       0.0        100.0
## Total              29170     100.0        100.0
```

```
adultpayclean %>% group_by(class) %>%
  mutate(n=n()) %>% ggplot() +
  geom_bar(aes(class,col=income,fill=income)) + scale_y_sqrt()
```



```
tab1(adultpayclean$income, sort.group = "decreasing", cum.percent = TRUE)
```

Distribution of adultpayclean\$income



```
## adultpayclean$income :  
##           Frequency Percent Cum. percent  
## AtBelow50K      21999      75.4         75.4  
## Above50K        7171      24.6         100.0  
## Total           29170     100.0         100.0
```

```
adultpayclean %>% group_by(age) %>%  
  mutate(n=n()) %>% ggplot() +  
  geom_bar(aes(age,col=income,fill=income)) + scale_y_sqrt()
```

