Bank.R

TG

```
library(ggplot2)

bankdata <- read.csv("bank-full.csv", sep = ";")

str(bankdata)
```

```
## 'data.frame':    45211 obs. of  17 variables:
##  $ age      : int  58 44 33 47 33 35 28 42 58 43 ...
##  $ job      : Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
##  $ marital  : Factor w/ 3 levels "divorced","married",..: 2 3 2 2 3 2 3 1 2 3 ...
##  $ education: Factor w/ 4 levels "primary","secondary",..: 3 2 2 4 4 3 3 3 1 2 ...
##  $ default  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ balance  : int  2143 29 2 1506 1 231 447 2 121 593 ...
##  $ housing  : Factor w/ 2 levels "no","yes": 2 2 2 2 1 2 2 2 2 2 ...
##  $ loan     : Factor w/ 2 levels "no","yes": 1 1 2 1 1 1 2 1 1 1 ...
##  $ contact  : Factor w/ 3 levels "cellular","telephone",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ day      : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ month    : Factor w/ 12 levels "apr","aug","dec",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ duration : int  261 151 76 92 198 139 217 380 50 55 ...
##  $ campaign : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pdays    : int  -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
##  $ previous : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poutcome : Factor w/ 4 levels "failure","other",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ y        : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(bankdata)
```

```
##       age                 job           marital          education
##  Min.   :18.00   blue-collar:9732   divorced: 5207   primary  : 6851
##  1st Qu.:33.00   management :9458   married :27214   secondary:23202
##  Median :39.00   technician :7597   single  :12790   tertiary :13301
##  Mean   :40.94   admin.     :5171                    unknown  : 1857
##  3rd Qu.:48.00   services   :4154
##  Max.   :95.00   retired    :2264
##                  (Other)    :6835
##  default        balance       housing      loan            contact
##  no :44396   Min.   : -8019   no :20081   no :37967   cellular :29285
##  yes:  815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906
##              Median :   448                           unknown  :13020
##              Mean   :  1362
##              3rd Qu.:  1428
##              Max.   :102127
##
##       day            month          duration         campaign
##  Min.   : 1.00   may    :13766   Min.   :   0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul    : 6895   1st Qu.: 103.0   1st Qu.: 1.000
##  Median :16.00   aug    : 6247   Median : 180.0   Median : 2.000
##  Mean   :15.81   jun    : 5341   Mean   : 258.2   Mean   : 2.764
##  3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000
##  Max.   :31.00   apr    : 2932   Max.   :4918.0   Max.   :63.000
##                  (Other): 6060
##      pdays          previous           poutcome       y
##  Min.   : -1.0   Min.   :  0.0000   failure: 4901   no :39922
##  1st Qu.: -1.0   1st Qu.:  0.0000   other  : 1840   yes: 5289
##  Median : -1.0   Median :  0.0000   success: 1511
##  Mean   : 40.2   Mean   :  0.5803   unknown:36959
##  3rd Qu.: -1.0   3rd Qu.:  0.0000
##  Max.   :871.0   Max.   :275.0000
##
```

```
head(bankdata)
```

```
##   age          job marital education default balance housing loan contact
## 1  58   management married  tertiary      no    2143     yes   no unknown
## 2  44   technician  single secondary      no      29     yes   no unknown
## 3  33 entrepreneur married secondary      no       2     yes  yes unknown
## 4  47  blue-collar married   unknown      no    1506     yes   no unknown
## 5  33      unknown  single   unknown      no       1      no   no unknown
## 6  35   management married  tertiary      no     231     yes   no unknown
##   day month duration campaign pdays previous poutcome  y
## 1   5   may      261        1    -1        0  unknown no
## 2   5   may      151        1    -1        0  unknown no
## 3   5   may       76        1    -1        0  unknown no
## 4   5   may       92        1    -1        0  unknown no
## 5   5   may      198        1    -1        0  unknown no
## 6   5   may      139        1    -1        0  unknown no
```

```
sample_dataset <- sample(1:nrow(bankdata), 20, replace = FALSE)
sample_dataset
```

```
##  [1]  9905 15219 43068  8021 20951  9443 24645 31137 29421 26432 11892
## [12]  2043 11572 19539  4262  1688 13031  9225  2286 26186
```

```
bankdata[sample_dataset,]
```

```
##       age           job marital education default balance housing loan
## 9905   37        admin. married secondary      no    1093      no   no
## 15219  45      services married secondary      no     307      no  yes
## 43068  66       retired married secondary      no    2326      no  yes
## 8021   32    management  single secondary      no     249     yes   no
## 20951  35    management married  tertiary      no     178      no   no
## 9443   28    technician  single secondary     yes   -1042     yes   no
## 24645  31 self-employed  single  tertiary      no     917      no  yes
## 31137  28       student  single secondary      no       0      no   no
## 29421  33    management married  tertiary      no     644     yes  yes
## 26432  31   blue-collar  single  tertiary      no     328     yes   no
## 11892  37      services married secondary      no       0      no   no
## 2043   28   blue-collar married   primary      no      26      no   no
## 11572  46   blue-collar married   unknown      no     626     yes   no
## 19539  45   blue-collar married   primary      no    -485     yes   no
## 4262   35    technician married   unknown      no     208     yes   no
## 1688   30    management married  tertiary      no    1825     yes   no
## 13031  33    management married  tertiary      no    -139      no   no
## 9225   58       retired married secondary      no     150     yes   no
## 2286   38    management married  tertiary      no     551     yes   no
## 26186  39  entrepreneur married secondary      no    5562      no   no
##        contact day month duration campaign pdays previous poutcome   y
## 9905   unknown   9   jun      160        1    -1        0  unknown  no
## 15219 cellular  17   jul      147        3    -1        0  unknown  no
## 43068 cellular  18   feb      232        1   181        1  success yes
## 8021   unknown   2   jun      258        1    -1        0  unknown  no
## 20951 cellular  14   aug       76        2    -1        0  unknown  no
## 9443   unknown   6   jun      712        2    -1        0  unknown yes
## 24645 cellular  17   nov      367        3    -1        0  unknown  no
## 31137 cellular  18   feb      209       20    -1        0  unknown yes
## 29421 cellular   3   feb      124        2   271        2  failure  no
## 26432 cellular  20   nov       89        1    -1        0  unknown  no
## 11892  unknown  20   jun       12       14    -1        0  unknown  no
## 2043   unknown  12   may       40        3    -1        0  unknown  no
## 11572  unknown  19   jun      177        2    -1        0  unknown  no
## 19539 cellular   7   aug      122        2    -1        0  unknown  no
## 4262   unknown  19   may      214        1    -1        0  unknown  no
## 1688   unknown   9   may      164        1    -1        0  unknown  no
## 13031 cellular   8   jul      290        1    -1        0  unknown  no
## 9225   unknown   5   jun      420        4    -1        0  unknown  no
## 2286   unknown  12   may     1875        4    -1        0  unknown  no
## 26186 cellular  20   nov      149        3    -1        0  unknown  no
```

```
str(bankdata$job)
```

```
##  Factor w/ 12 levels "admin.","blue-collar",..: 5 10 3 2 12 5 5 3 6 10 ...
```

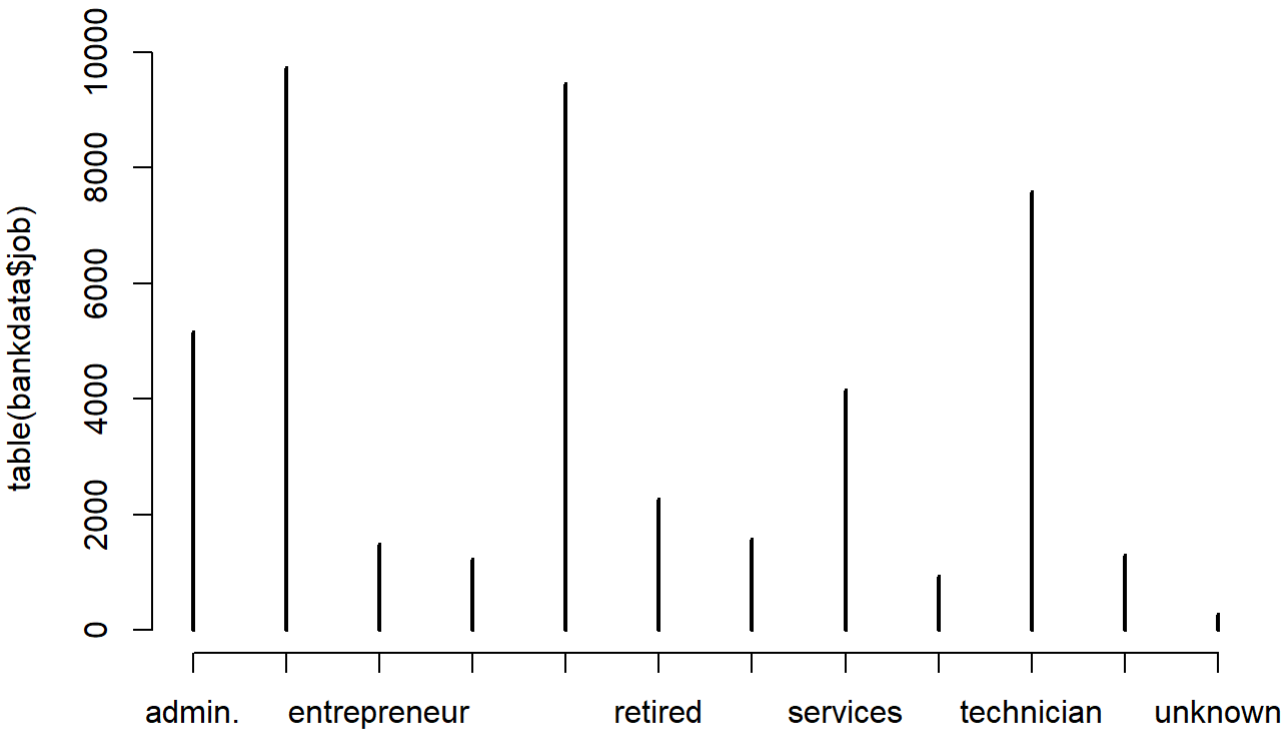
```
levels(bankdata$job)
```

```
##  [1] "admin."        "blue-collar"   "entrepreneur"  "housemaid"
##  [5] "management"    "retired"       "self-employed" "services"
##  [9] "student"       "technician"    "unemployed"    "unknown"
```

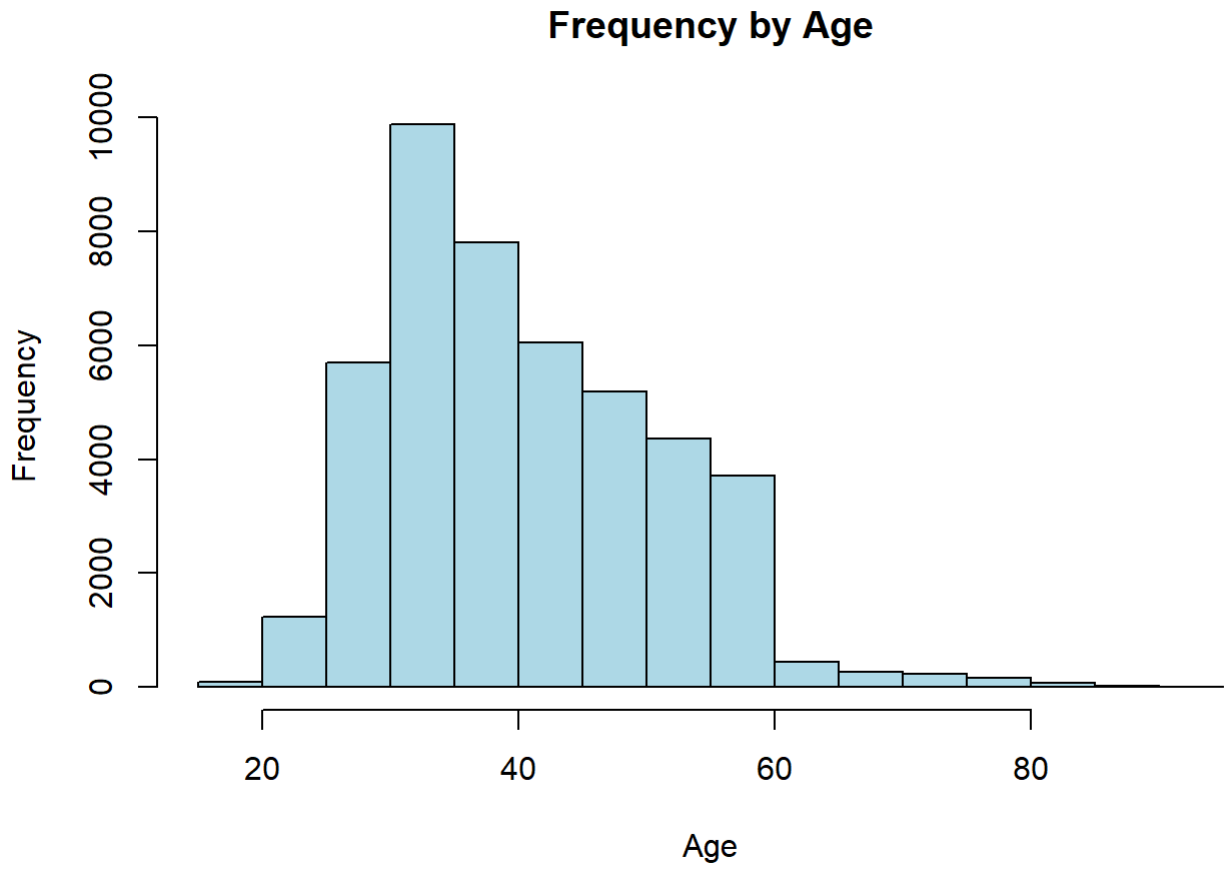
```
table(bankdata$job)
```

```
##
##        admin.   blue-collar  entrepreneur     housemaid    management
##          5171          9732          1487          1240          9458
##       retired self-employed      services       student    technician
##          2264          1579          4154           938          7597
##    unemployed       unknown
##          1303           288
```

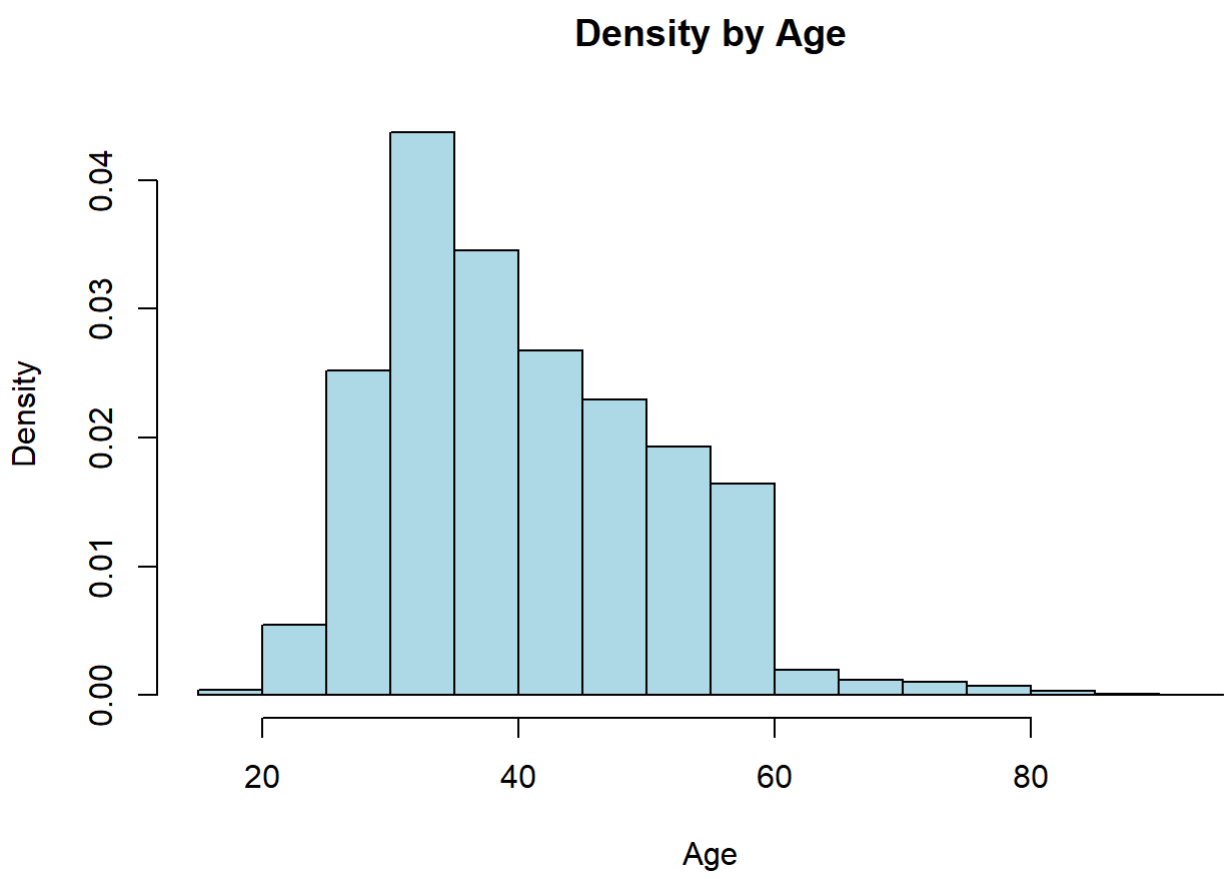
```
plot(table(bankdata$job))
```



```
hist(bankdata$age, col = "light blue", main="Frequency by Age",
     xlab="Age", ylab="Frequency")
```

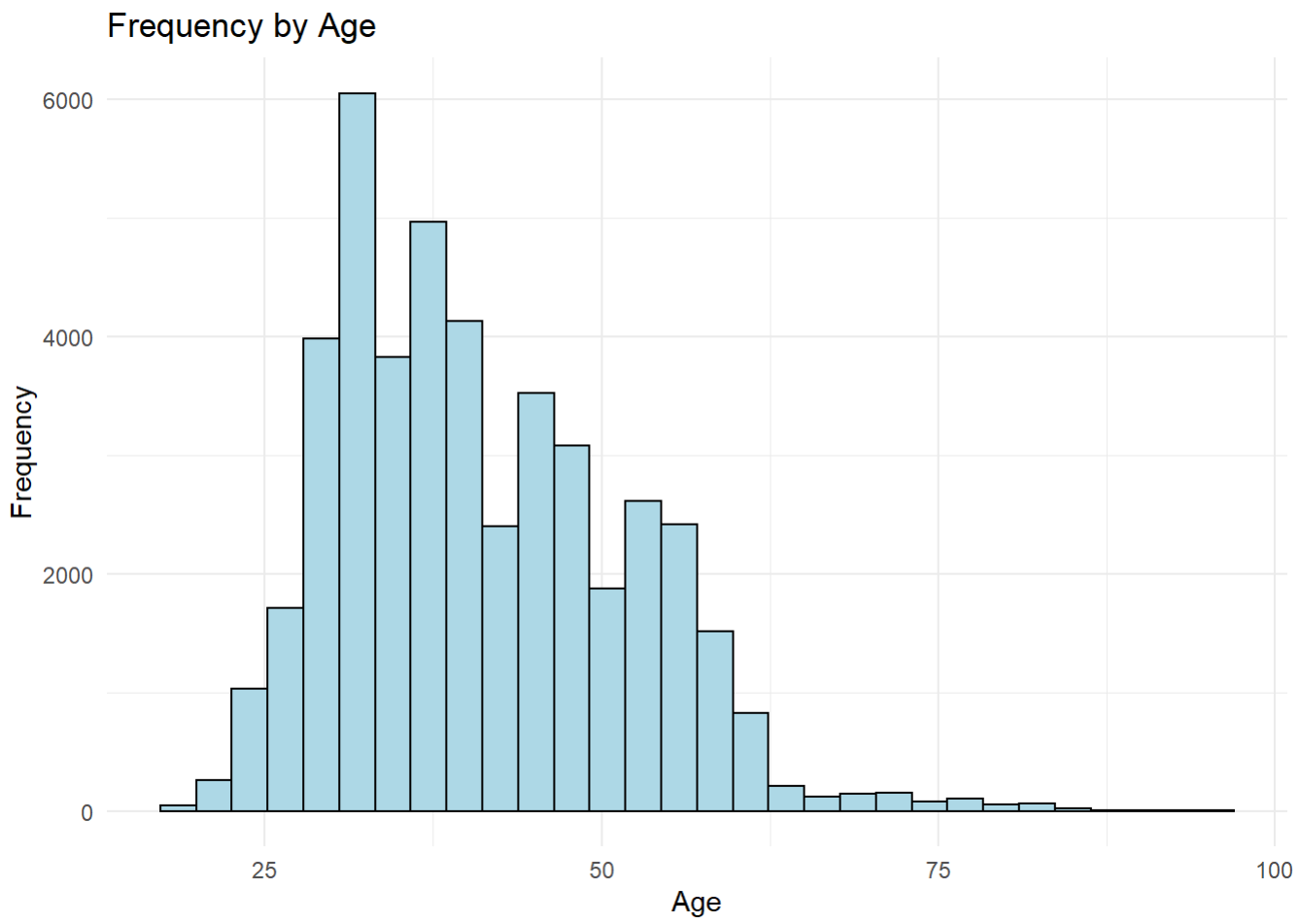



```
hist(bankdata$age, col = "light blue", freq = FALSE, main="Density by Age",
     xlab="Age", ylab="Density")
```




```
ggplot(data = bankdata, aes(x = age),) +
  geom_histogram( fill = "light blue", color = "black") +
  labs(x = "Age", y = "Frequency", title = "Frequency by Age") +
  theme_minimal()
```

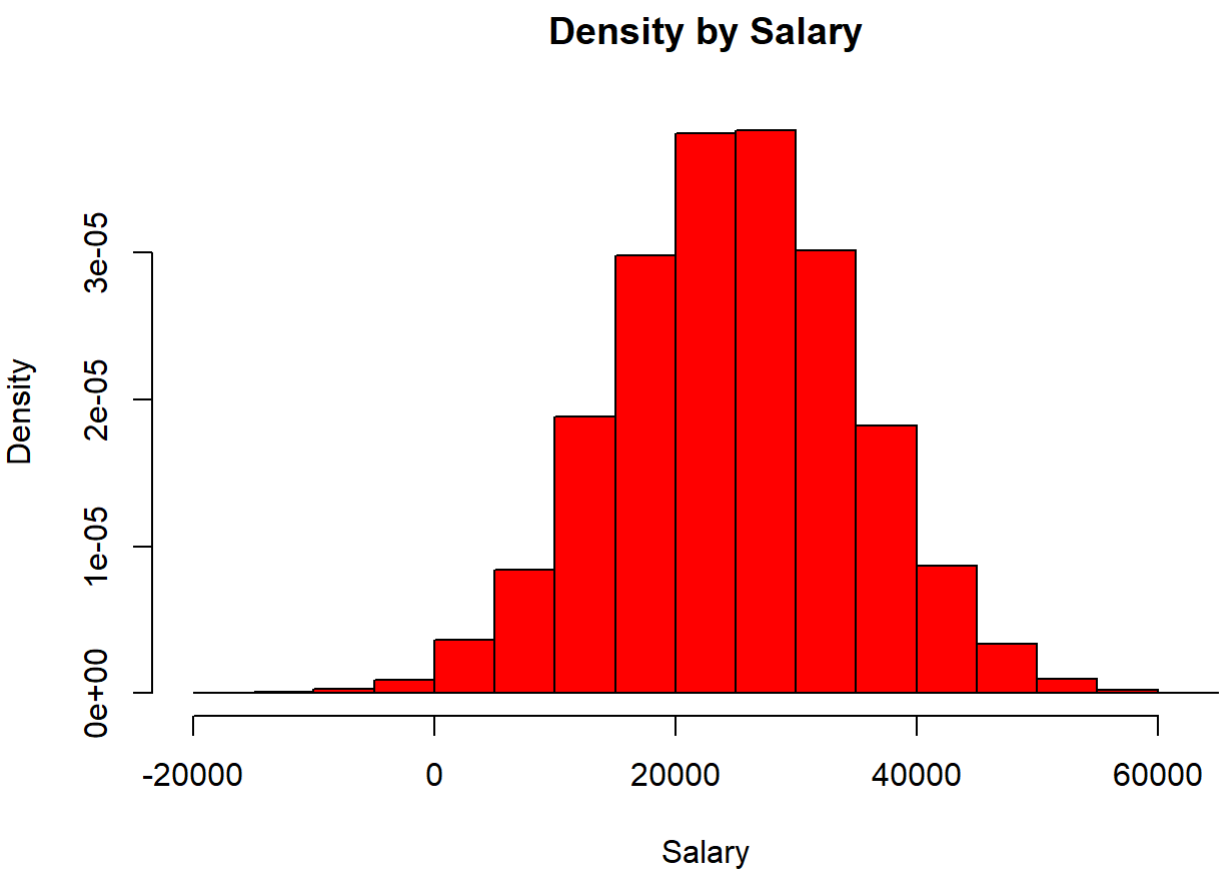
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
bankdata$salary = rnorm(nrow(bankdata), mean = 25000, sd = 10000)


hist(bankdata$salary, col = "Red", main="Frequency by Salary",
     xlab="Salary", ylab="Frequency")
```


```
hist(bankdata$salary, col = "Red", freq = FALSE, main="Density by Salary",
     xlab="Salary", ylab="Density")
```




```
yeardata = ifelse(bankdata$month %in% c("jan", "feb", "mar", "apr", "may", "jun"), yes = "1st Ha
lf",no = "2nd Half")
bankdata$yeardata = yeardata


tabEDMar = table(bankdata$education, bankdata$marital)
tabEDMar
```

```
##
##             divorced married single
##   primary        752    5246    853
##   secondary     2815   13770   6617
##   tertiary      1471    7038   4792
##   unknown        169    1160    528
```

```
round(prop.table(tabEDMar)*100,3)
```

```
##
##             divorced married single
##   primary      1.663  11.603  1.887
##   secondary    6.226  30.457 14.636
##   tertiary     3.254  15.567 10.599
##   unknown      0.374   2.566  1.168
```

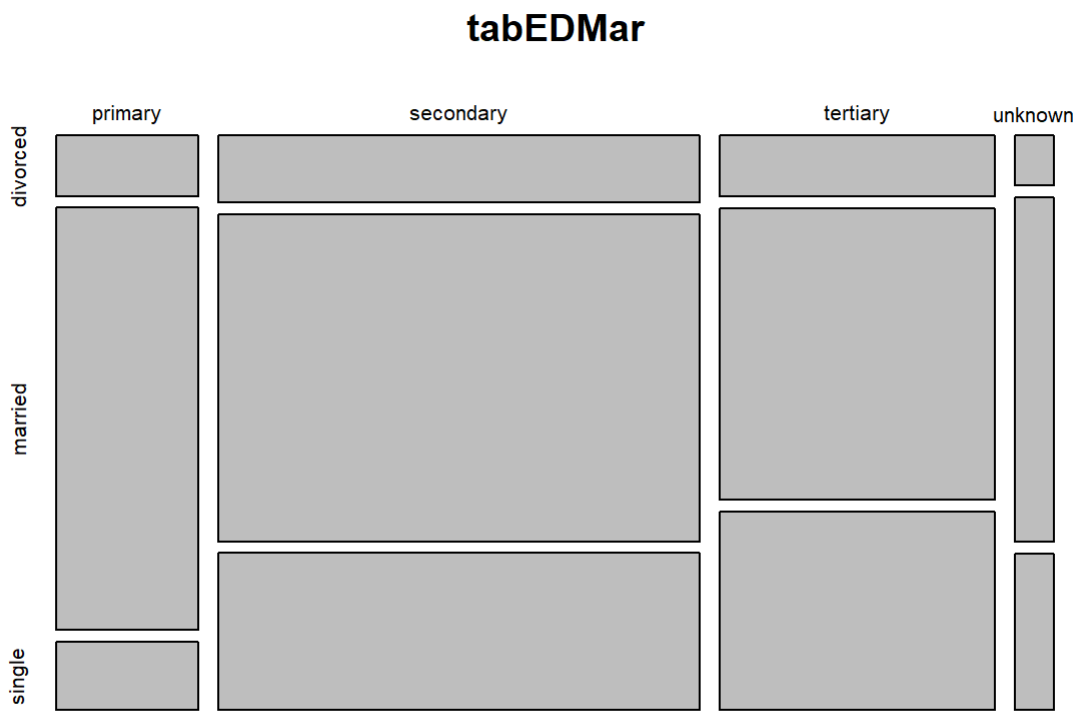
```
round(prop.table(tabEDMar,1)*100,3)
```

```
##
##             divorced married single
##   primary     10.976  76.573 12.451
##   secondary   12.133  59.348 28.519
##   tertiary    11.059  52.913 36.027
##   unknown      9.101  62.466 28.433
```

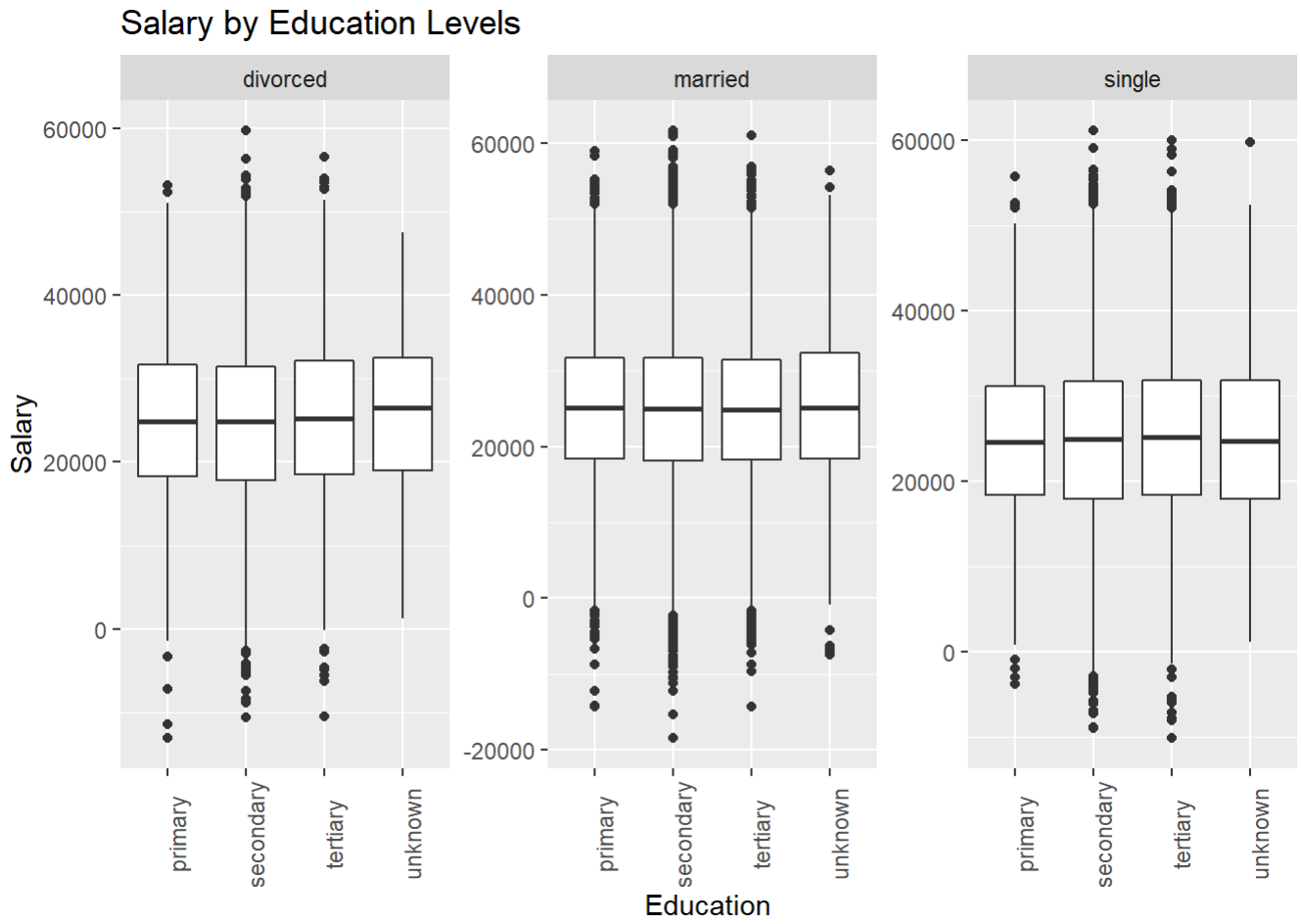
```
round(prop.table(tabEDMar,2)*100,3)
```

```
##
##             divorced married single
##   primary     14.442  19.277  6.669
##   secondary   54.062  50.599 51.736
##   tertiary    28.250  25.862 37.467
##   unknown      3.246   4.263  4.128
```

```
mosaicplot(tabEDMar)
```



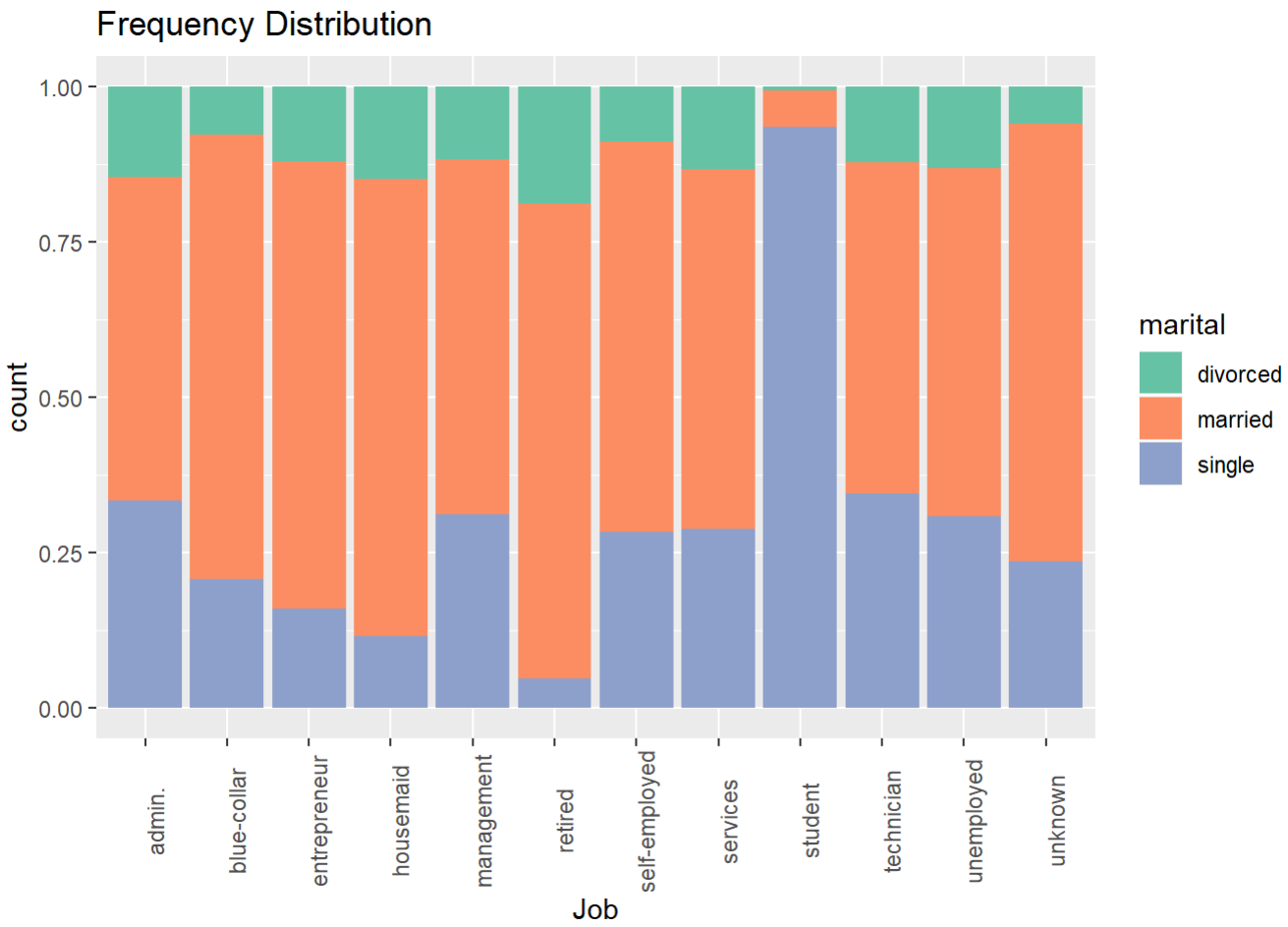
```
ggplot(data = bankdata, aes(x = education, y = salary)) +
  geom_boxplot() +
  facet_wrap(~ marital, scales = "free_y") +
  labs(x = "Education", y = "Salary", title = "Salary by Education Levels") +
  theme(axis.text.x=element_text(angle=90))
```



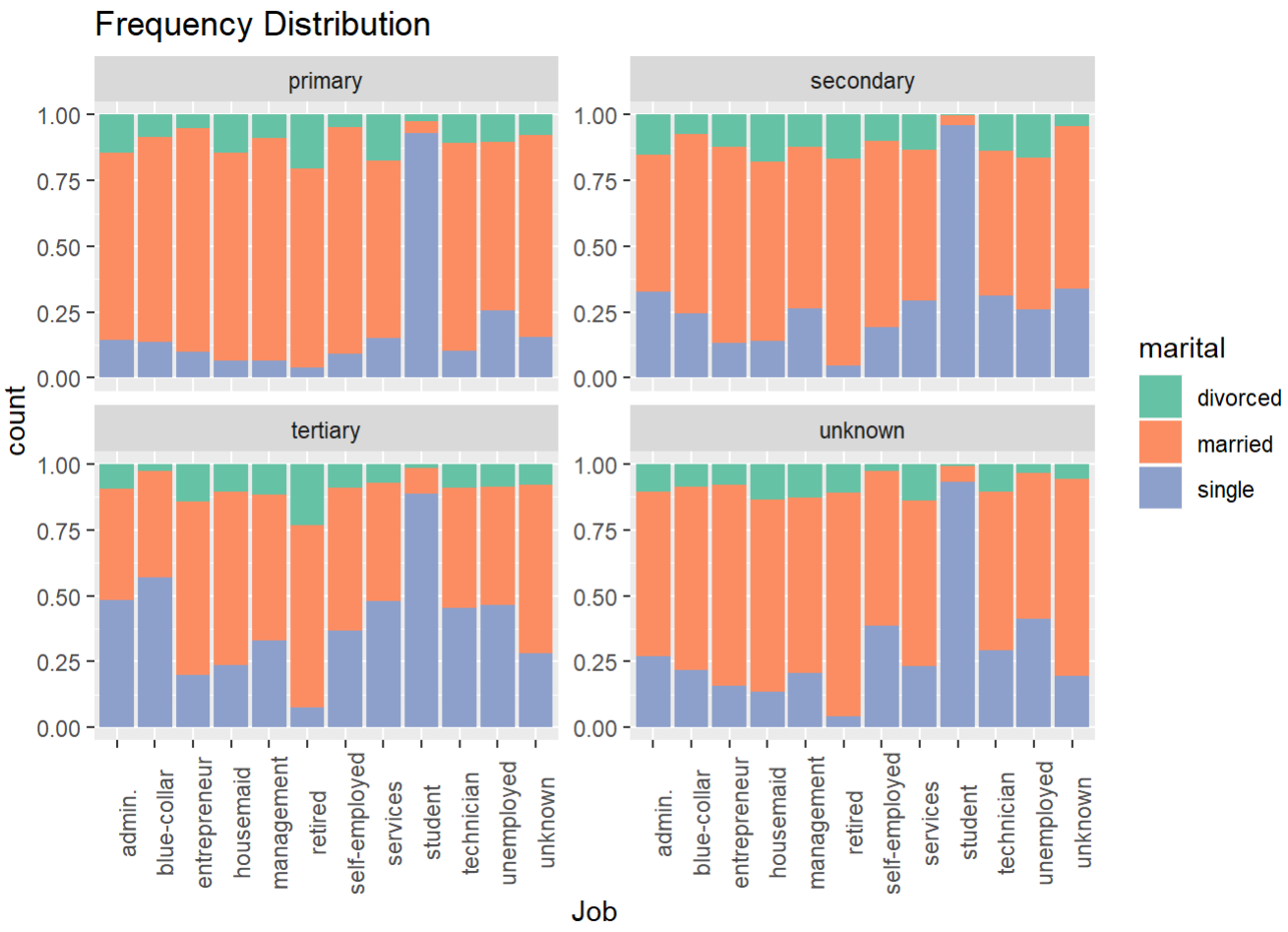

```
ggplot(bankdata, aes(x = age, y = salary, color = y)) +
  geom_point() +
  facet_wrap(~ job, scales = "free_y") +
  labs(x = "Age", y = "Salary", title = "Salary by Age Group") +
  theme(axis.text.x=element_text(angle=90))
```


```
ggplot(bankdata, aes(x = age, y = salary, color = y)) +
  geom_point() +
  geom_smooth(method=lm) +
  geom_smooth(method=lm, aes(group = 1)) +
  facet_wrap(~ job, scales = "free_y") +
  labs(x = "Age", y = "Salary", title = "Salary by Age Group") +
  theme(axis.text.x=element_text(angle=90))
```


```
ggplot(bankdata, aes(x = job, fill = marital)) +
  geom_bar(position = "fill") +
  scale_fill_brewer(palette = "Set2") +
  labs(x = "Job", title = "Frequency Distribution") +
  theme(axis.text.x=element_text(angle=90))
```




```
ggplot(bankdata, aes(x = job, fill = marital)) +
  geom_bar(position = "fill") +
  scale_fill_brewer(palette = "Set2") +
  facet_wrap(~ education, scales = "free_y") +
  labs(x = "Job", title = "Frequency Distribution") +
  theme(axis.text.x=element_text(angle=90))
```


Frequency Distribution
primary
secondary
tertiary
unknown
count
1.00
0.75
0.50
0.25
0.00
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown
Job
marital
divorced
married
single