data mining1

kiyana mirbaghestan

2022-11-10

#Question-1

:داده ها را به صورت زیر فراخوانی میکنیم

```
adult<-read.csv("C:/Users/asus/Documents/adult_income_data.txt")
```

:تشخیص نوع داده با دستور زیر می باشد که خوشبختانه آر بدرستی تشخیص داده است

```
#View(adult)
str(adult)

## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" " Priv
ate" ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209
642 45781 159449 ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital.status: chr  " Never-married" " Married-civ-spouse" " Divorced"
" Married-civ-spouse" ...
##  $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-clea
ners" " Handlers-cleaners" ...
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" " Hus
band" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ sex           : chr  " Male" " Male" " Male" " Male" ...
##  $ capital.gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital.loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours.per.week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native.country: chr  " United-States" " United-States" " United-States"
" United-States" ...
##  $ income        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...

attach(adult)
```

:بعد داده بصورت زیر می باشد

```
dim(adult)

## [1] 32561    15
```

#Question-2

برای دید بهتر نسبت به داده ها ده سطر اول را فراخوانی می کنیم:

```
adult[1:10,]

##    age         workclass fnlwgt  education education.num         marital.s
tatus
## 1   39         State-gov  77516  Bachelors            13          Never-ma
rried
## 2   50  Self-emp-not-inc  83311  Bachelors            13     Married-civ-s
pouse
## 3   38           Private 215646    HS-grad             9               Div
orced
## 4   53           Private 234721       11th             7     Married-civ-s
pouse
## 5   28           Private 338409  Bachelors            13     Married-civ-s
pouse
## 6   37           Private 284582    Masters            14     Married-civ-s
pouse
## 7   49           Private 160187        9th             5  Married-spouse-a
bsent
## 8   52  Self-emp-not-inc 209642    HS-grad             9     Married-civ-s
pouse
## 9   31           Private  45781    Masters            14          Never-ma
rried
## 10  42           Private 159449  Bachelors            13     Married-civ-s
pouse
##            occupation   relationship   race     sex capital.gain capital.l
oss
## 1        Adm-clerical  Not-in-family  White    Male         2174
0
## 2     Exec-managerial        Husband  White    Male            0
0
## 3   Handlers-cleaners  Not-in-family  White    Male            0
0
## 4   Handlers-cleaners        Husband  Black    Male            0
0
## 5      Prof-specialty           Wife  Black  Female            0
0
## 6     Exec-managerial           Wife  White  Female            0
0
## 7       Other-service  Not-in-family  Black  Female            0
0
## 8     Exec-managerial        Husband  White    Male            0
0
## 9      Prof-specialty  Not-in-family  White  Female        14084
0
## 10    Exec-managerial        Husband  White    Male         5178
0
```

```
##    hours.per.week native.country income
## 1              40  United-States  <=50K
## 2              13  United-States  <=50K
## 3              40  United-States  <=50K
## 4              40  United-States  <=50K
## 5              40           Cuba  <=50K
## 6              40  United-States  <=50K
## 7              16        Jamaica  <=50K
## 8              45  United-States   >50K
## 9              50  United-States   >50K
## 10             40  United-States   >50K

#head(adult,10)
```

#Question-3

چون درصد داده های گمشده کم است روش حذف آنها میتواند راه حل مناسبی باشد:

```
adult[adult==" ?"]=NA
k=is.na(adult)
sum(is.na(adult))

## [1] 4262

adult2<-na.omit(adult)
dim(adult2)

## [1] 30162    15

attach(adult2)

## The following objects are masked from adult:
##
##     age, capital.gain, capital.loss, education, education.num, fnlwgt,
##     hours.per.week, income, marital.status, native.country, occupation,
##     race, relationship, sex, workclass
```

نسبت داده های گمشده:

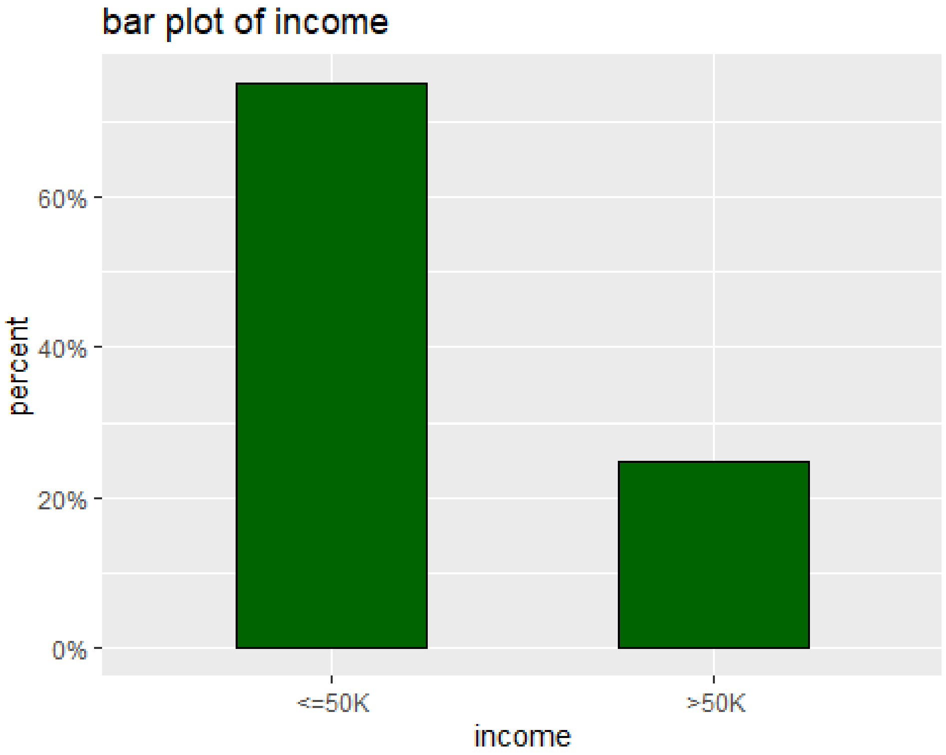
```
missingpercent<-4262/32561
```

#Question-4to7(categoricals) #response #Frequency #mosaicplot #barplot #ggplot

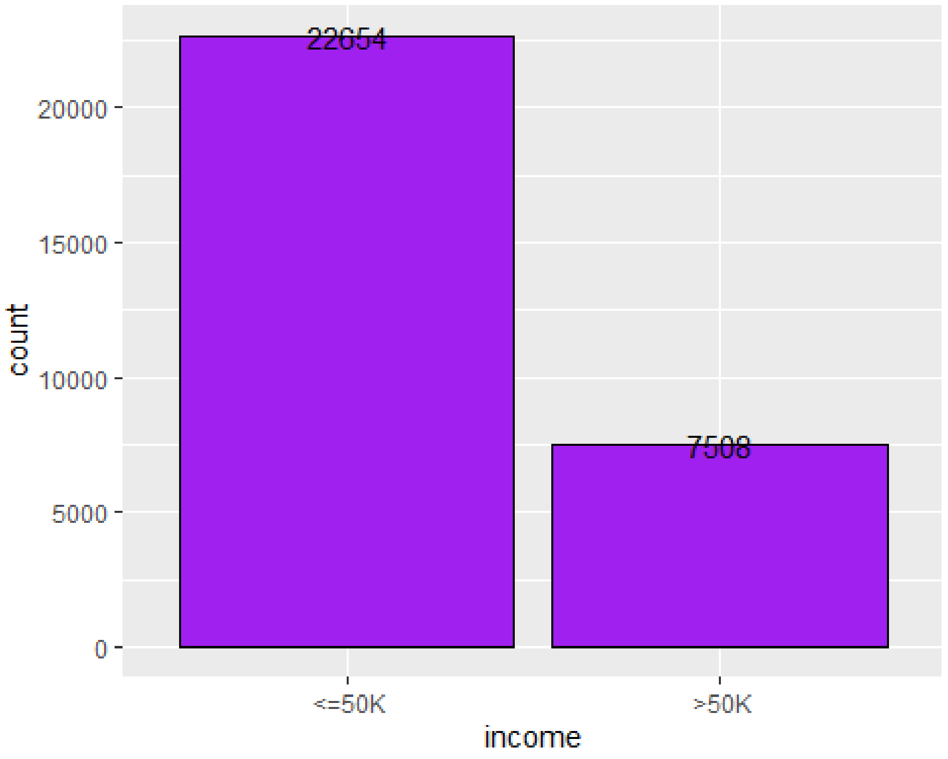
نسبت کسانی که درآمد بیشتر از ۵۰ دارند تقریبا ۲۵ درصد و کسانی که درآمد کمتر از ۵۰ دارند ۷۵ درصد می باشد:

```
i<-table(adult2$income)
proportions(i)

##
##     <=50K      >50K
## 0.7510775 0.2489225
```

```
library(ggplot2)
ggplot(data=adult2,aes(x=income))+
  geom_bar(fill="darkgreen",colour="black",width=0.5,aes(y=..prop..,group=1))
+
  scale_y_continuous(labels=scales::percent_format())+
  labs(y="percent",title="bar plot of income")
```



```
ggplot(adult2) +
  aes(x = income) +
  geom_bar(fill="purple",colour="black")+
  geom_text(stat="count",aes(label=..count..))
```

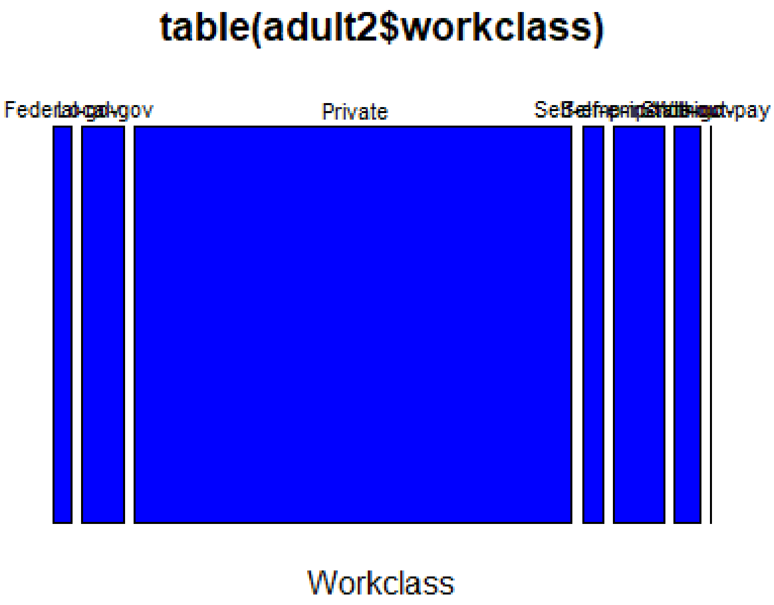


ازنمودار بالا در می یابیم کسانیکه حقوق بیش از ۵۰ دارند ۷۵۰۸نفر و کسانیکه حقوق کمتر دارند۲۲۰۵٤ هستند لذا
نیازمند افزایش حقوقها هستیم.

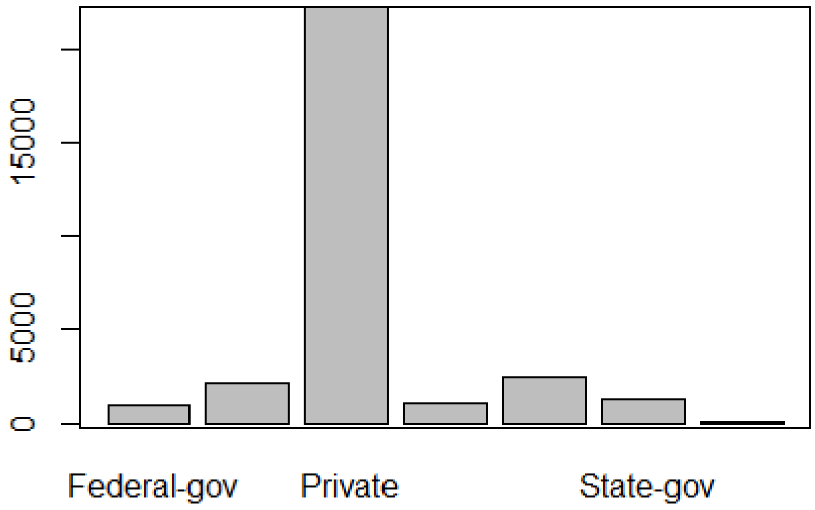
از جدول ونمودارهای زیر میفهمیم که قسمت خصوصی مد این متغیر است چون فراوانی بیشتری دارد:

#Workclass

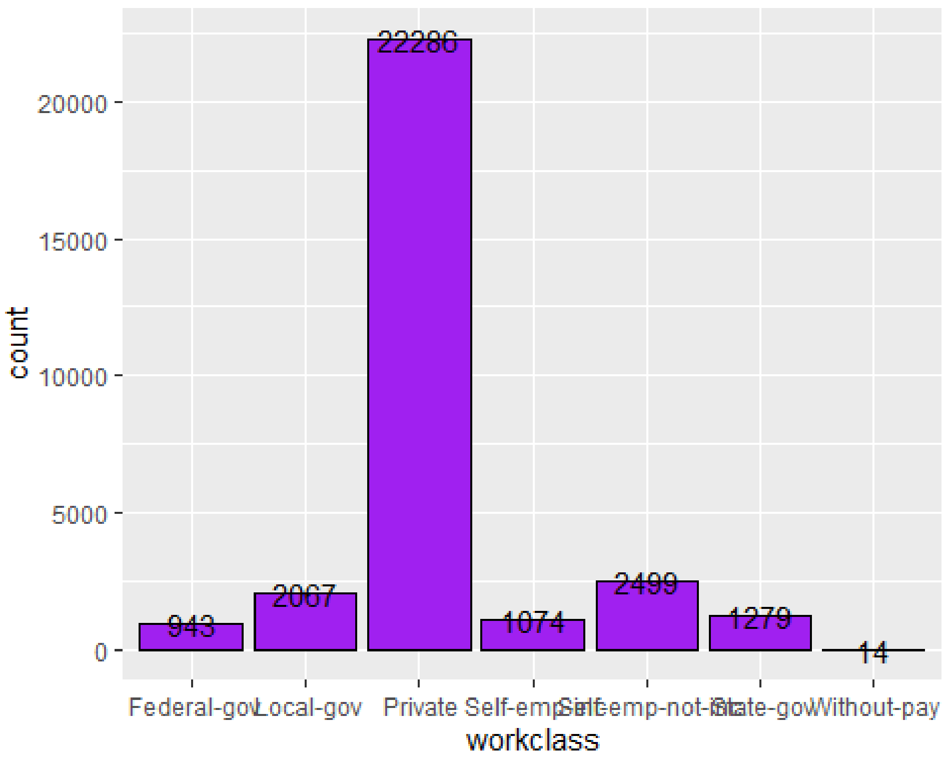
```
w<-table(adult2$workclass)
proportions(w)

##
##       Federal-gov         Local-gov           Private      Self-emp-inc
##      0.0312645050      0.0685299383      0.7388767323      0.0356077183
##  Self-emp-not-inc         State-gov       Without-pay
##      0.0828525960      0.0424043498      0.0004641602

mosaicplot(table(adult2$workclass),
           color = "Blue",
           xlab = "Workclass", # label for x-axis
)
```



```
barplot(table(adult2$workclass))
box(which = "plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```

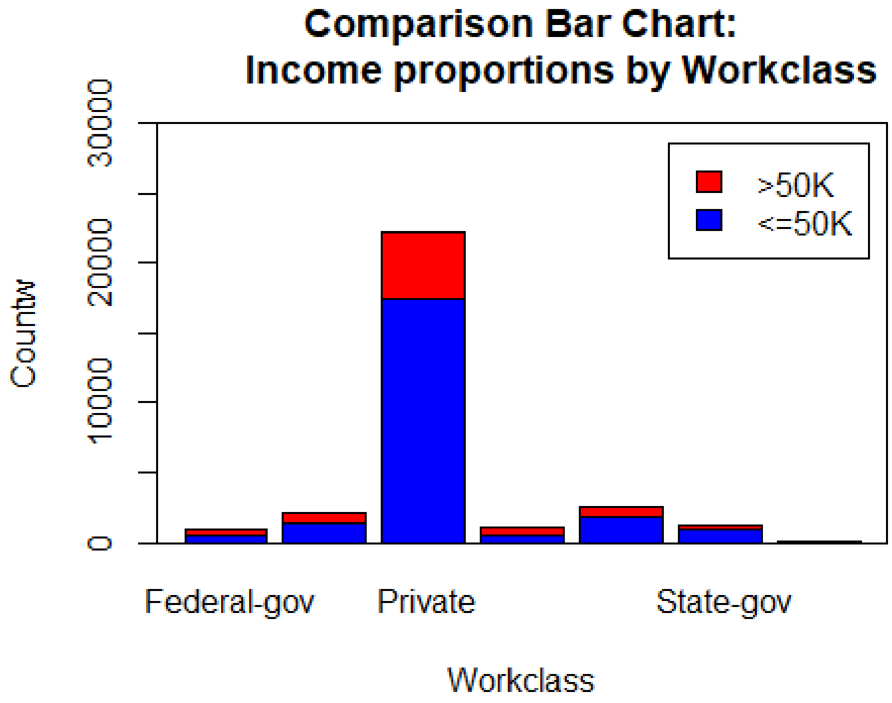


```
ggplot(adult2) +
  aes(x = workclass) +
  geom_bar(fill="purple",colour="black")+
  geom_text(stat="count",aes(label=..count..))
```

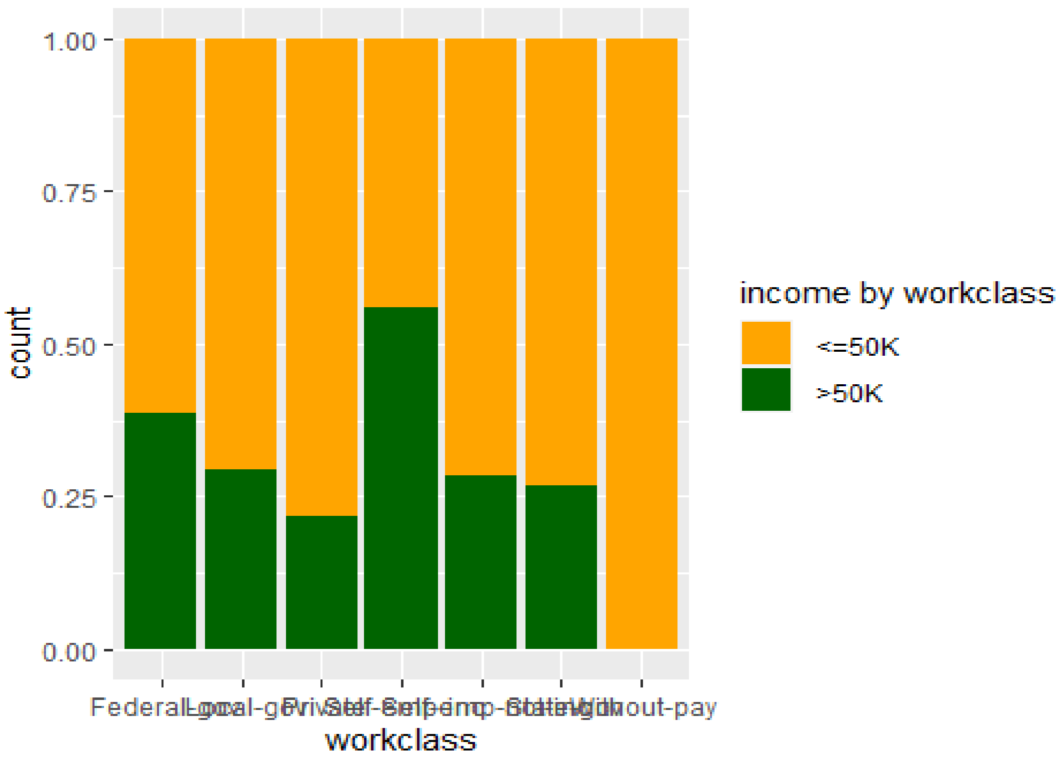


```
#with response
xtabs(~income+workclass,data=adult2)

##          workclass
## income    Federal-gov  Local-gov  Private  Self-emp-inc  Self-emp-not-inc
##    <=50K          578       1458    17410           474              1785
##    >50K           365        609     4876           600               714
##          workclass
## income    State-gov  Without-pay
##    <=50K        935           14
##    >50K         344            0

prop.table(xtabs(~income+workclass,data=adult2))

##          workclass
## income    Federal-gov    Local-gov      Private  Self-emp-inc  Self-emp-no
t-inc
##    <=50K 0.0191631855 0.0483389696 0.5772163650  0.0157151383      0.05918
04257
##    >50K  0.0121013195 0.0201909688 0.1616603673  0.0198925801      0.02367
21703
##          workclass
## income      State-gov  Without-pay
##    <=50K 0.0309992706 0.0004641602
##    >50K  0.0114050792 0.0000000000
```

```
countw<-table(adult2$income,adult2$workclass,
              dnn=c("Income","Workclass"))
countw

##          Workclass
## Income    Federal-gov Local-gov Private Self-emp-inc Self-emp-not-inc
##    <=50K          578      1458   17410          474             1785
##    >50K           365       609    4876          600              714
##          Workclass
## Income    State-gov Without-pay
##    <=50K        935          14
##    >50K         344           0

sumtable<-addmargins(countw,FUN=sum)

## Margins computed over dimensions
## in the following order:
## 1: Income
## 2: Workclass

barplot(countw,
        legend=rownames(countw),
        col=c("blue","red"),
        ylim=c(0,30000),
        ylab="Countw",
        xlab="Workclass",
        main="Comparison Bar Chart:
        Income proportions by Workclass")
box(which="plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



```r
#with ggplot
ggplot(adult2,aes(x=workclass,group=income,fill=income))+
  geom_bar(position="fill")+
  scale_fill_manual(values=c("orange","darkgreen"),name="income by workclass"
)
```

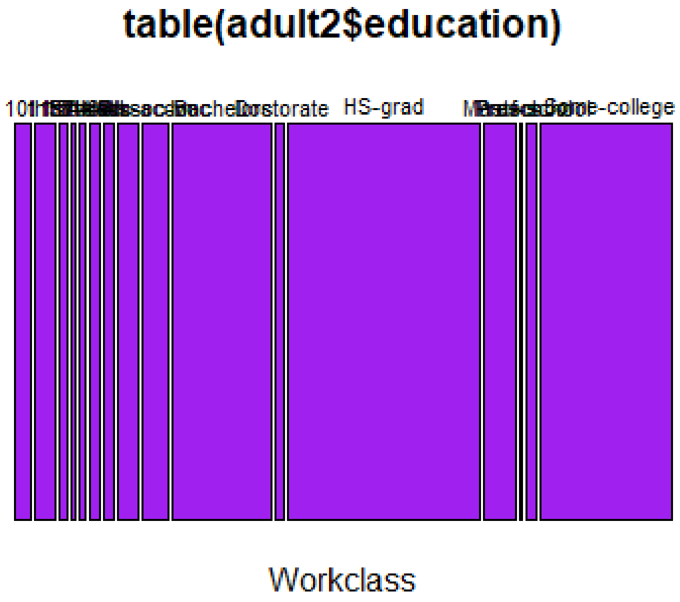


از نمودار زرد و سبز بالا در می یابیم در قسمت سلف امپ اینس نسبت بالاتری حقوق بیش از ۵۰ میگیرند.

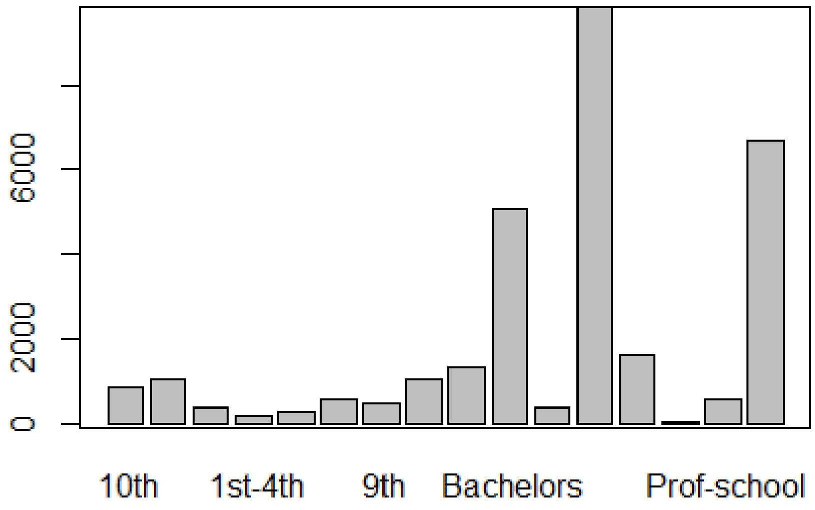
برای تحصیلات :قسمت اچ اس افراد بیشتری را تشکیل می دهند(مد) اما اگر بخواهیم بیشتر بودن میزان حقوق را در نظز بگیریم باتوجه به نمودارها دکتراها و پروف اسکول ها نسبت بیشتری حقوق بیش از ۵۰ دارند.

#Education

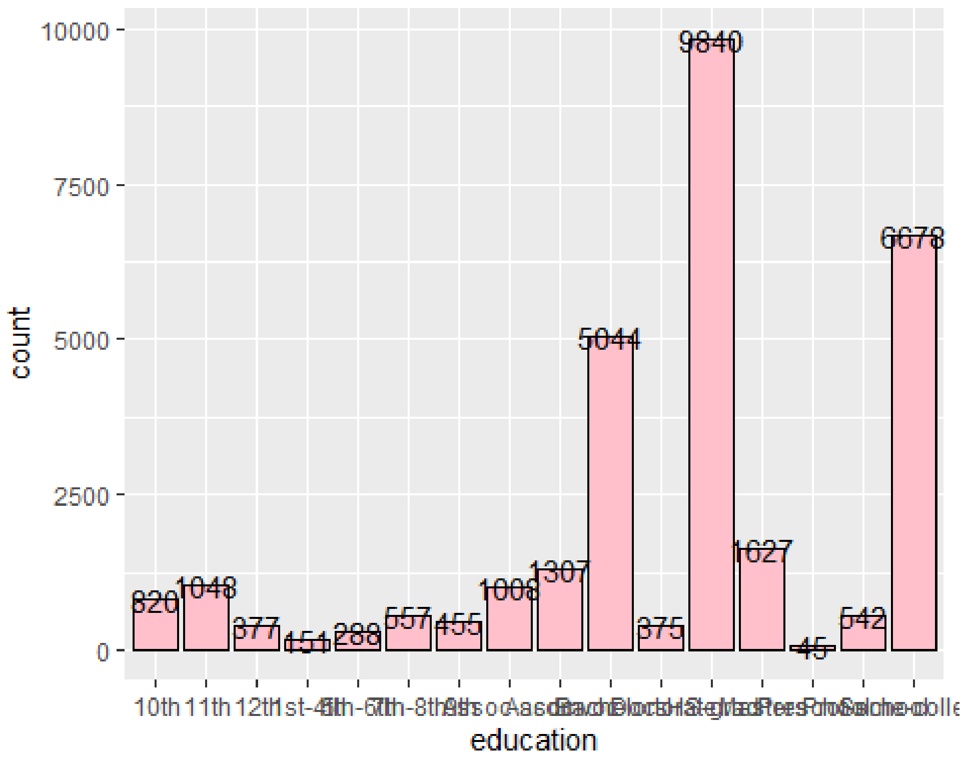
```
e<-table(adult2$education)
proportions(e)

## 
##         10th        11th        12th     1st-4th     5th-6th 
##  0.027186526 0.034745707 0.012499171 0.005006299 0.009548438 
##      7th-8th         9th  Assoc-acdm   Assoc-voc   Bachelors 
##  0.018466945 0.015085207 0.033419535 0.043332670 0.167230290 
##    Doctorate     HS-grad     Masters   Preschool Prof-school 
##  0.012432863 0.326238313 0.053942046 0.001491944 0.017969631 
## Some-college 
##  0.221404416

mosaicplot(table(adult2$education),
           color = "purple",
           xlab = "Workclass", # label for x-axis
)
```



```
barplot(table(adult2$education))
box(which = "plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



```
ggplot(adult2) +
  aes(x = education) +
  geom_bar(fill="pink",colour="black")+
  geom_text(stat="count",aes(label=..count..))
```



```
#with response
xtabs(~income+education,data=adult2)

##          education
## income    10th  11th  12th  1st-4th  5th-6th  7th-8th  9th  Assoc-acdm
##    <=50K   761   989   348      145      276      522  430         752
##    >50K     59    59    29        6       12       35   25         256
##          education
## income    Assoc-voc  Bachelors  Doctorate  HS-grad  Masters  Preschool
##    <=50K        963       2918         95     8223      709         45
##    >50K         344       2126        280     1617      918          0
##          education
## income    Prof-school  Some-college
##    <=50K          136          5342
##    >50K           406          1336

prop.table(xtabs(~income+education,data=adult2))

##          education
## income            10th         11th         12th      1st-4th      5th-6th
##    <=50K 0.0252304224 0.0327896028 0.0115376964 0.0048073735 0.0091505868
##    >50K  0.0019561037 0.0019561037 0.0009614747 0.0001989258 0.0003978516
##          education
## income         7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
##    <=50K 0.0173065447 0.0142563490 0.0249320337 0.0319275910 0.0967442477
##    >50K  0.0011604005 0.0008288575 0.0084875008 0.0114050792 0.0704860420
##          education
## income      Doctorate       HS-grad      Masters     Preschool  Prof-school
##    <=50K 0.0031496585 0.2726278098 0.0235063988 0.0014919435 0.0045089848
##    >50K  0.0092832040 0.0536105033 0.0304356475 0.0000000000 0.0134606458
##          education
## income    Some-college
##    <=50K  0.1771102712
##    >50K   0.0442941450
```

```
counte<-table(adult2$income,adult2$education,
              dnn=c("Income","Education"))
counte
```

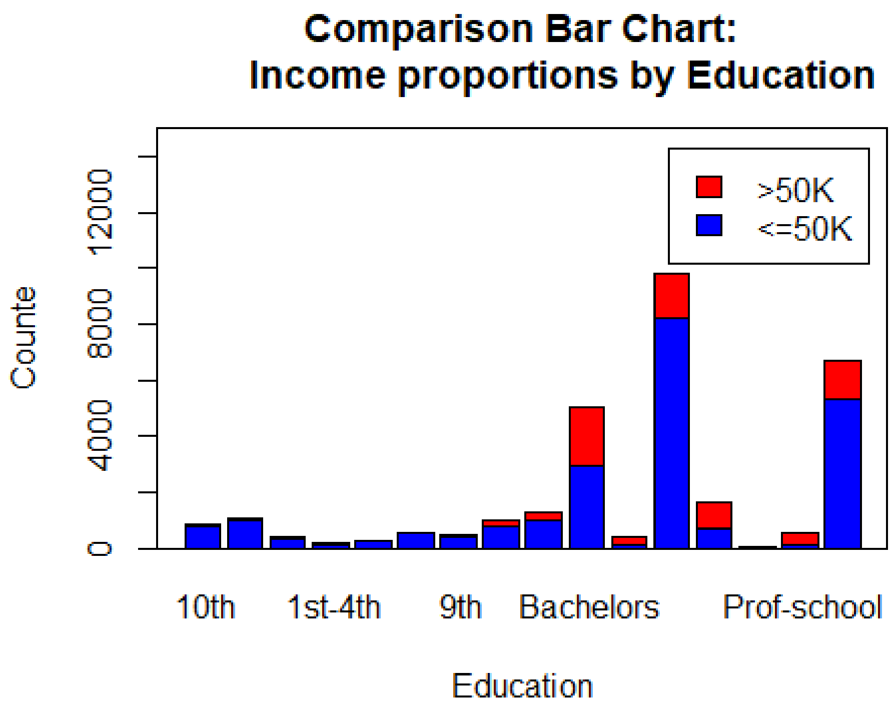
```
##          Education
## Income    10th 11th 12th 1st-4th 5th-6th 7th-8th 9th  Assoc-acdm
##    <=50K   761  989  348     145     276     522  430        752
##    >50K     59   59   29       6      12      35   25        256
##          Education
## Income    Assoc-voc Bachelors Doctorate HS-grad Masters Preschool
##    <=50K        963      2918        95    8223     709        45
##    >50K         344      2126       280    1617     918         0
##          Education
## Income    Prof-school Some-college
##    <=50K          136         5342
##    >50K           406         1336
```

```
sumtable<-addmargins(counte,FUN=sum)
```

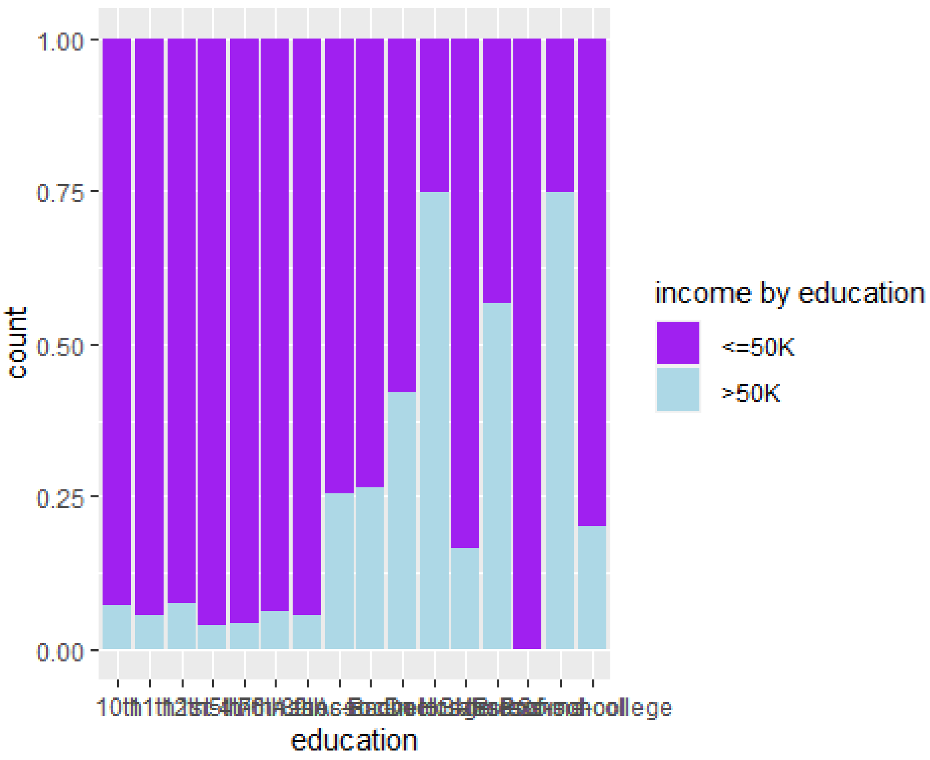
```
## Margins computed over dimensions
## in the following order:
## 1: Income
## 2: Education
```

```
barplot(counte,
        legend=rownames(counte),
        col=c("blue","red"),
        ylim=c(0,15000),
        ylab="Counte",
        xlab="Education",
        main="Comparison Bar Chart:
        Income proportions by Education")
box(which="plot",
    Ity="solid",
    col="black")
```

```
## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



```
#with ggplot
ggplot(adult2,aes(x=education,group=income,fill=income))+
  geom_bar(position="fill")+
  scale_fill_manual(values=c("purple","lightblue"),name="income by education"
)
```



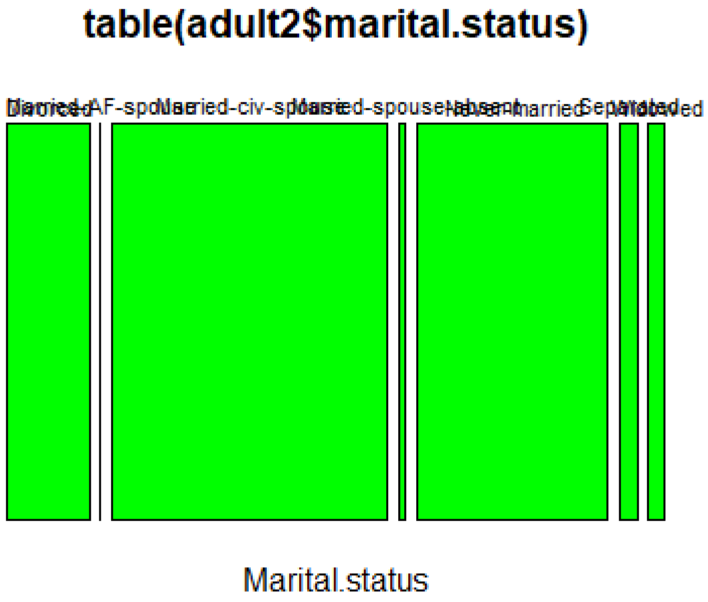
درمورد وضعیت تاهل قسمت مرید سیو اسپوز بیشترین فراوانی را دارذ بعبارتی مد مورد نظرماست.

#Marital.status

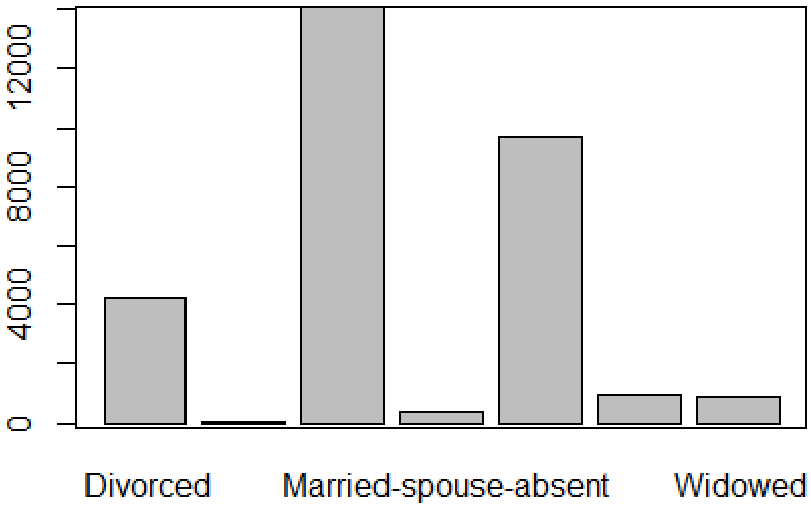
```
m<-table(adult2$marital.status)
proportions(m)
```

```
## 
##              Divorced     Married-AF-spouse     Married-civ-spouse 
##          0.1397122207          0.0006962403           0.4663152311 
##  Married-spouse-absent         Never-married              Separated 
##          0.0122670910          0.3224587229           0.0311318878 
##               Widowed 
##          0.0274186062
```

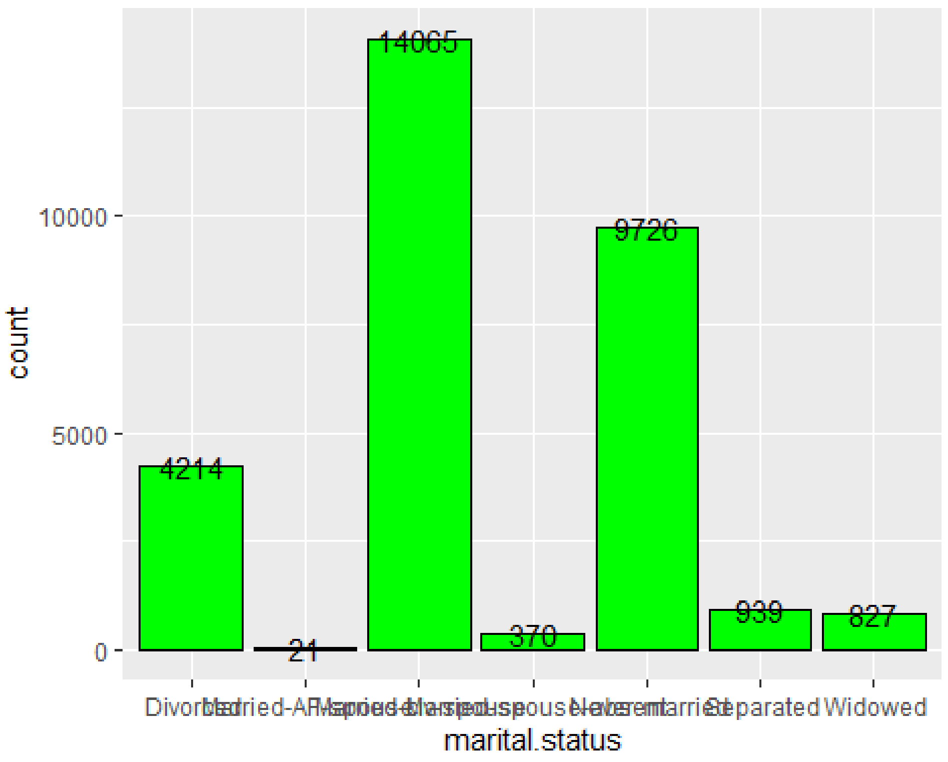
```
mosaicplot(table(adult2$marital.status),
           color = "Green",
           xlab = "Marital.status", # label for x-axis
)
```



```
barplot(table(adult2$marital.status))
box(which = "plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



```
ggplot(adult2) +
  aes(x = marital.status) +
  geom_bar(fill="green",colour="black")+
  geom_text(stat="count",aes(label=..count..))
```



```
#with response
xtabs(~income+marital.status,data=adult2)

##         marital.status
## income    Divorced  Married-AF-spouse  Married-civ-spouse
##    <=50K      3762                 11                7666
##    >50K        452                 10                6399
##         marital.status
## income    Married-spouse-absent  Never-married  Separated  Widowed
##    <=50K                    339           9256        873      747
##    >50K                      31            470         66       80

prop.table(xtabs(~income+marital.status,data=adult2))

##         marital.status
## income       Divorced  Married-AF-spouse  Married-civ-spouse
##    <=50K 0.1247264770       0.0003646973        0.2541608647
##    >50K  0.0149857437       0.0003315430        0.2121543664
##         marital.status
## income    Married-spouse-absent  Never-married    Separated      Widowed
##    <=50K           0.0112393077   0.3068762018 0.0289437040 0.0247662622
##    >50K            0.0010277833   0.0155825211 0.0021881838 0.0026523440

countm<-table(adult2$income,adult2$marital.status,
              dnn=c("Income","Artial.status"))
countm
```

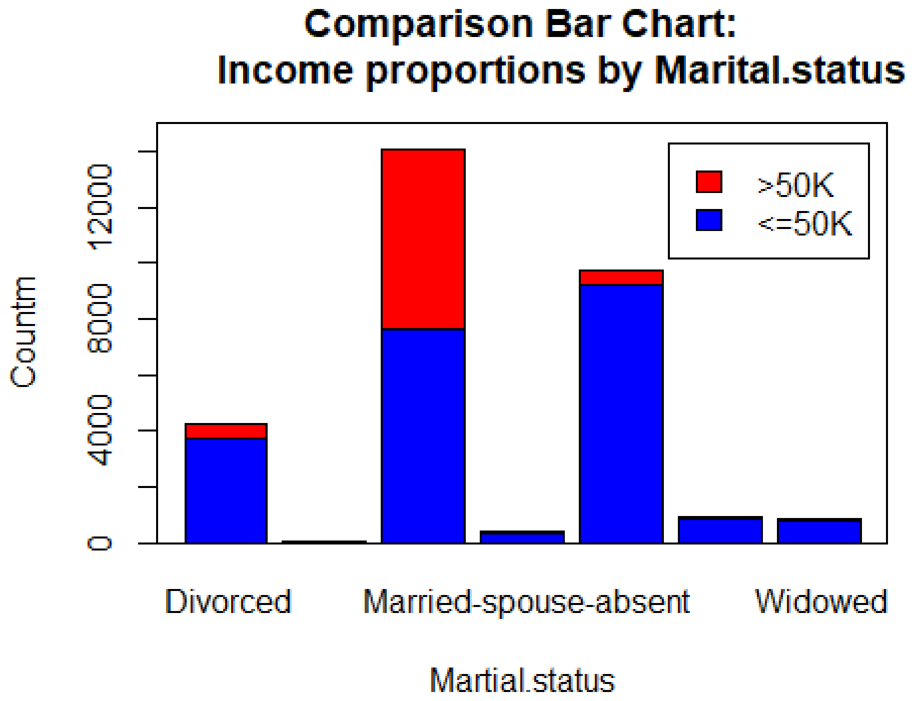
```
##          Artial.status
## Income    Divorced  Married-AF-spouse  Married-civ-spouse
##    <=50K      3762                 11                7666
##    >50K        452                 10                6399
##          Artial.status
## Income    Married-spouse-absent  Never-married  Separated  Widowed
##    <=50K                    339           9256        873      747
##    >50K                      31            470         66       80
```

```
sumtable<-addmargins(countm,FUN=sum)
```

```
## Margins computed over dimensions
## in the following order:
## 1: Income
## 2: Artial.status
```

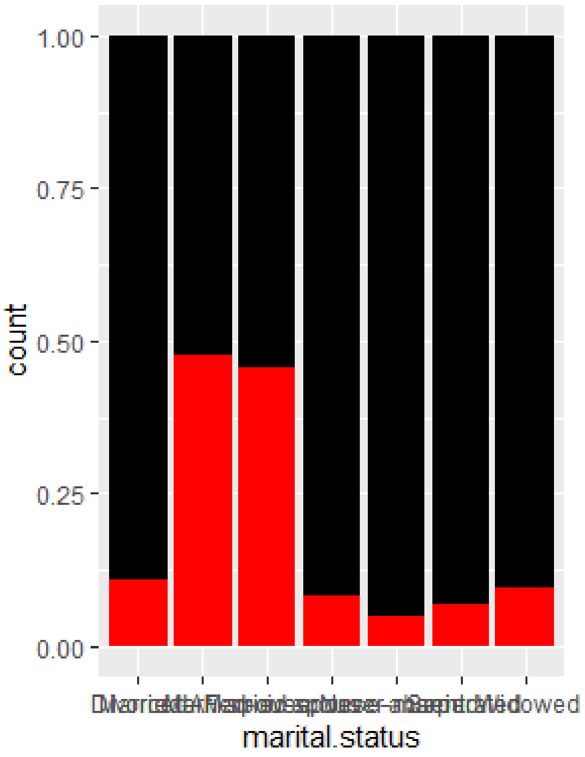
```
barplot(countm,
        legend=rownames(countm),
        col=c("blue","red"),
        ylim=c(0,15000),
        ylab="Countm",
        xlab="Martial.status",
        main="Comparison Bar Chart:
        Income proportions by Marital.status")
box(which="plot",
    Ity="solid",
    col="black")
```

```
## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



از نمودار آبی قرمز و نمودار پایین در می یابیم که در قسمت مریداف اسپوز نسبت کسانی که درامد بیشتراز ۵۰ دارند از بقیه قسمت ها بیشتر است.

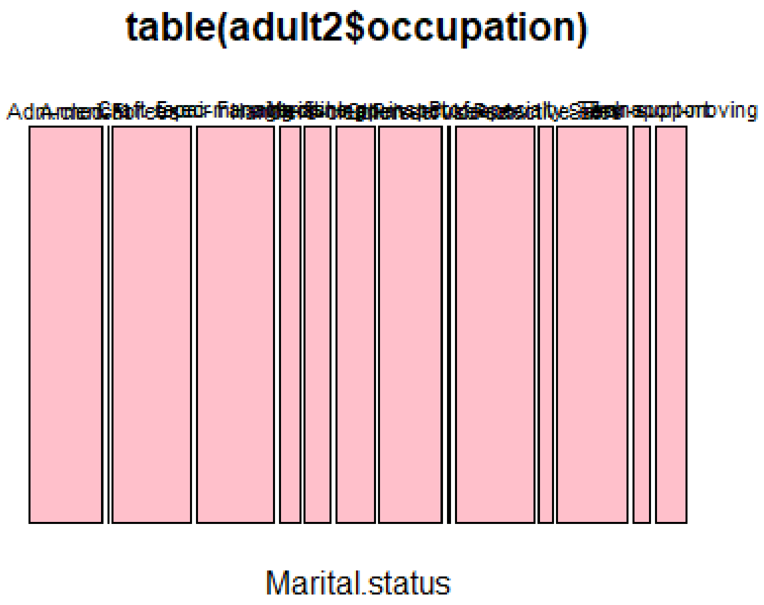
```
#with ggplot
ggplot(adult2,aes(x=marital.status,group=income,fill=income))+
  geom_bar(position="fill")+
  scale_fill_manual(values=c("black","red"),name="income by Marital.status")
```



در قسمت شغل کسانی که کاردستی تعمیر میکنند فراوانی بیشتری دارندو مد هستند.

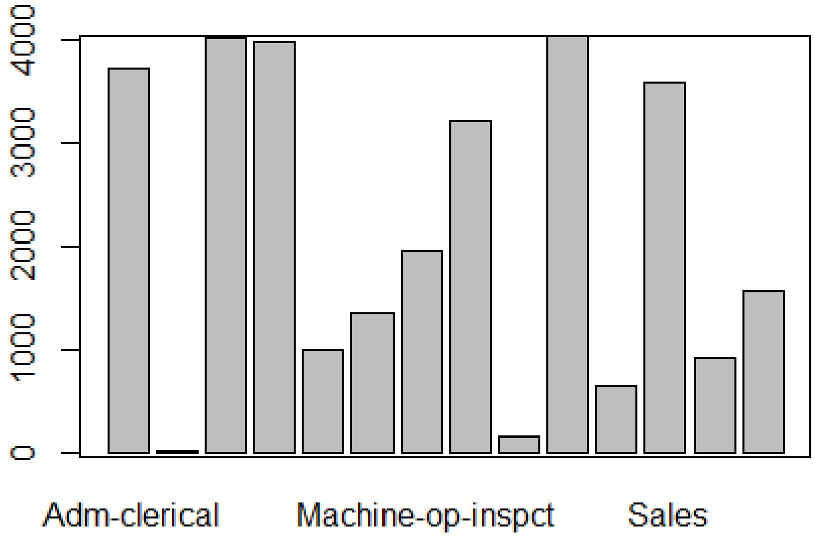
#Occupation

```
o<-table(adult2$occupation)
proportions(o)

## 
##       Adm-clerical       Armed-Forces       Craft-repair    Exec-manageria
l 
##       0.1233671507       0.0002983887       0.1336118295       0.132351966
0 
##    Farming-fishing  Handlers-cleaners  Machine-op-inspct      Other-servic
e 
##       0.0327896028       0.0447583052       0.0651813540       0.106491612
0 
##    Priv-house-serv     Prof-specialty    Protective-serv              Sale
s 
##       0.0047410649       0.1338770639       0.0213513693       0.118825011
6 
##       Tech-support   Transport-moving 
##       0.0302367217       0.0521185598

mosaicplot(table(adult2$occupation),
           color = "pink",
           xlab = "Marital.status", # label for x-axis
)
```

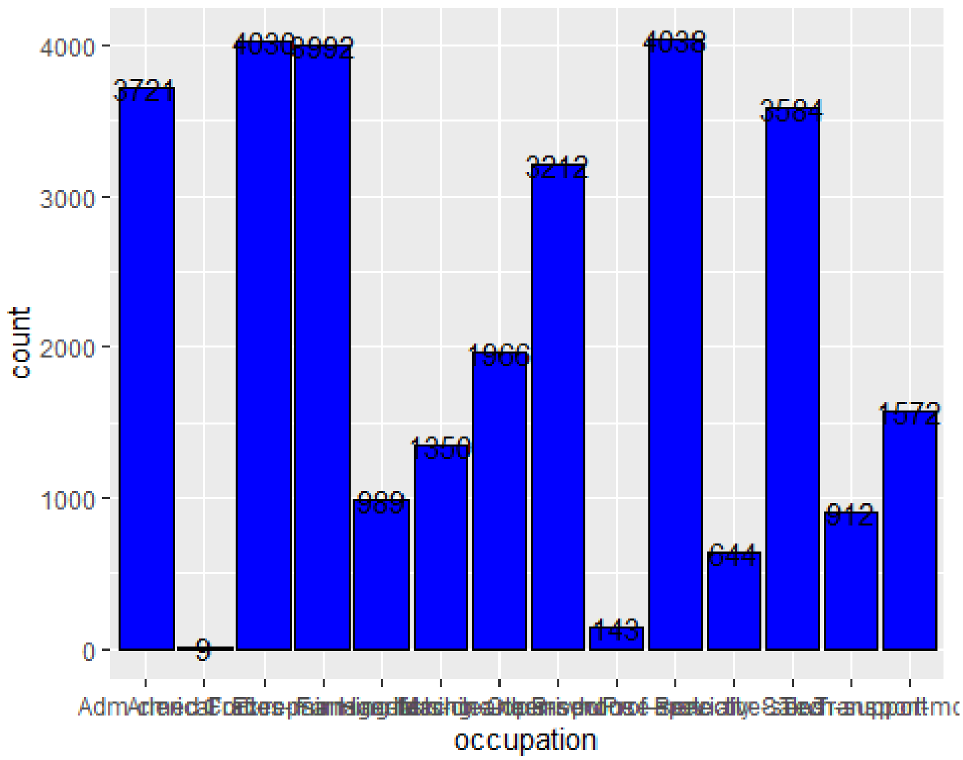


```
barplot(table(adult2$occupation))
box(which = "plot",
    Ity="solid",
    col="black")
```

```
## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



```
ggplot(adult2) +
  aes(x = occupation) +
  geom_bar(fill="blue",colour="black")+
  geom_text(stat="count",aes(label=..count..))
```



```
#with response
xtabs(~income+occupation,data=adult2)

##         occupation
## income    Adm-clerical  Armed-Forces  Craft-repair  Exec-managerial
##    <=50K          3223             8          3122             2055
##    >50K            498             1           908             1937
##         occupation
## income    Farming-fishing  Handlers-cleaners  Machine-op-inspct  Other-ser
vice
##    <=50K              874               1267               1721
3080
##    >50K               115                 83                245
132
##         occupation
## income    Priv-house-serv  Prof-specialty  Protective-serv  Sales  Tech-su
pport
##    <=50K              142            2227              434   2614
634
##    >50K                 1            1811              210    970
278
##         occupation
## income    Transport-moving
##    <=50K              1253
##    >50K                319

prop.table(xtabs(~income+occupation,data=adult2))
```

```
##         occupation
## income    Adm-clerical  Armed-Forces  Craft-repair  Exec-managerial
##    <=50K  0.1068563093  0.0002652344  0.1035077250     0.0681320867
##    >50K   0.0165108415  0.0000331543  0.0301041045     0.0642198793
##         occupation
## income    Farming-fishing  Handlers-cleaners  Machine-op-inspct  Other-ser
vice
##    <=50K     0.0289768583       0.0420064982       0.0570585505   0.102115
2443
##    >50K      0.0038127445       0.0027518069       0.0081228035   0.004376
3676
##         occupation
## income    Priv-house-serv  Prof-specialty  Protective-serv        Sales
##    <=50K     0.0047079106    0.0738346264     0.0143889662 0.0866653405
##    >50K      0.0000331543    0.0600424375     0.0069624030 0.0321596711
##         occupation
## income    Tech-support  Transport-moving
##    <=50K  0.0210198263      0.0415423380
##    >50K   0.0092168954      0.0105762217
```

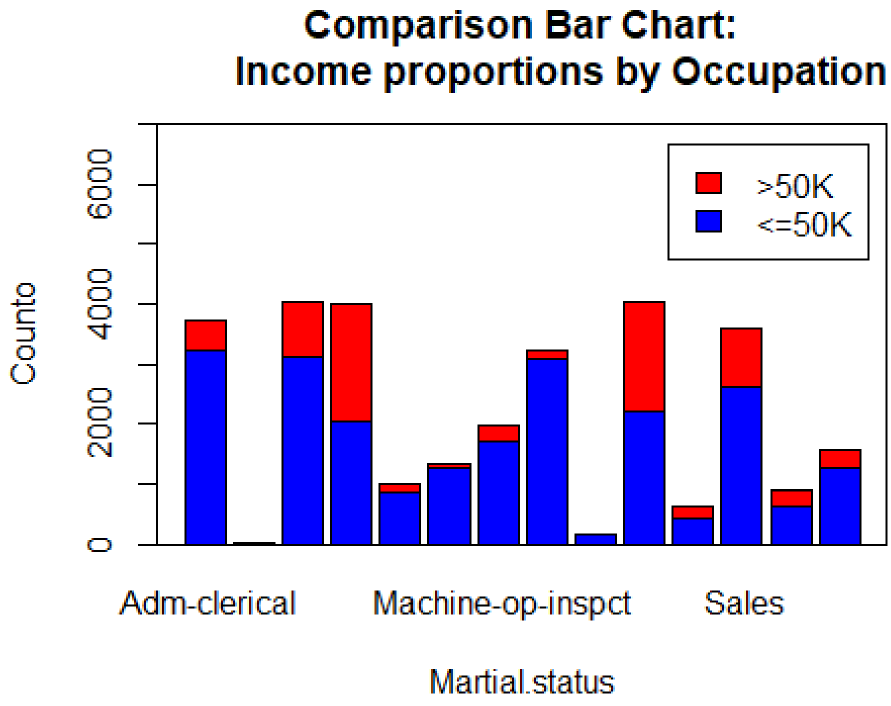
```
counto<-table(adult2$income,adult2$occupation,
              dnn=c("Income","Occupation"))
counto
```

```
##         Occupation
## Income    Adm-clerical  Armed-Forces  Craft-repair  Exec-managerial
##    <=50K          3223             8          3122             2055
##    >50K            498             1           908             1937
##         Occupation
## Income    Farming-fishing  Handlers-cleaners  Machine-op-inspct  Other-ser
vice
##    <=50K              874               1267               1721
3080
##    >50K               115                 83                245
132
##         Occupation
## Income    Priv-house-serv  Prof-specialty  Protective-serv  Sales  Tech-su
pport
##    <=50K              142            2227              434   2614
634
##    >50K                 1            1811              210    970
278
##         Occupation
## Income    Transport-moving
##    <=50K              1253
##    >50K                319
```

```
sumtable<-addmargins(counto,FUN=sum)
```

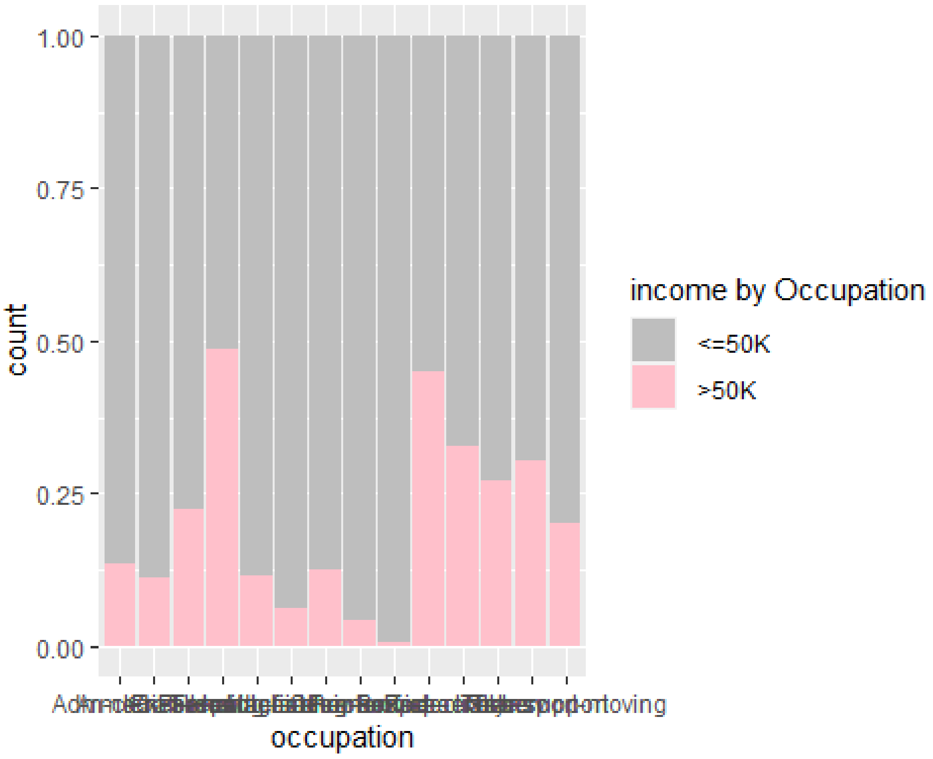
```
## Margins computed over dimensions
## in the following order:
```

```
## 1: Income
## 2: Occupation

barplot(counto,
        legend=rownames(counto),
        col=c("blue","red"),
        ylim=c(0,7000),
        ylab="Counto",
        xlab="Martial.status",
        main="Comparison Bar Chart:
        Income proportions by Occupation")
box(which="plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



از نمودار بالا و پایین متوجه میشویم که افرادی که اکست منیجر هستند نسبت درامد بالای۵۰شان بیشتراست.

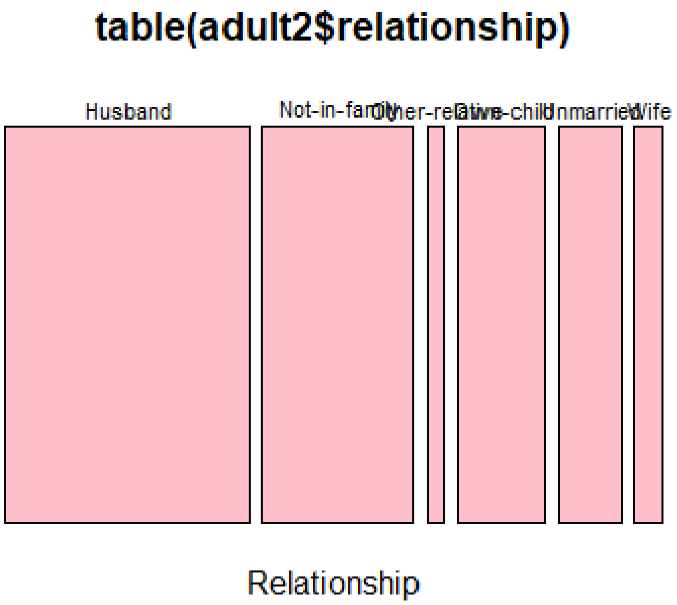
```
#with ggplot
ggplot(adult2,aes(x=occupation,group=income,fill=income))+
  geom_bar(position="fill")+
  scale_fill_manual(values=c("gray","pink"),name="income by Occupation")
```



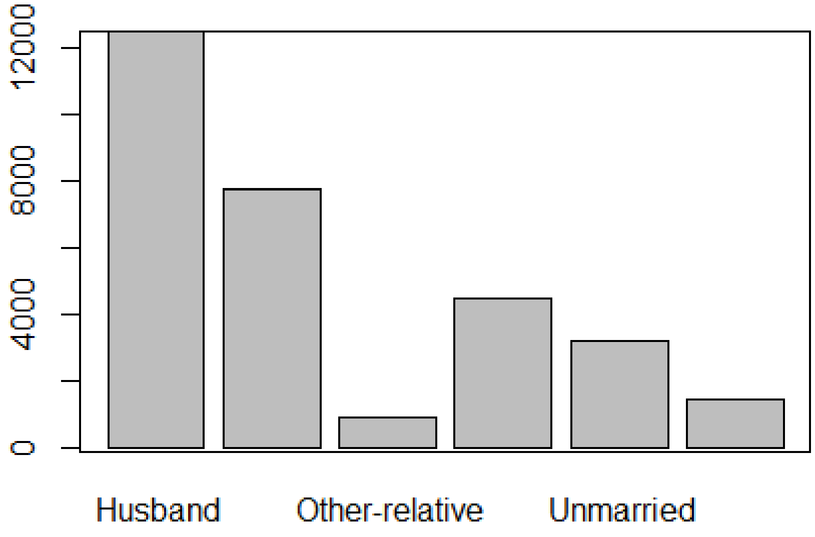
در روابط افرادی که شوهر هستند مد می باشند.

#Relationship

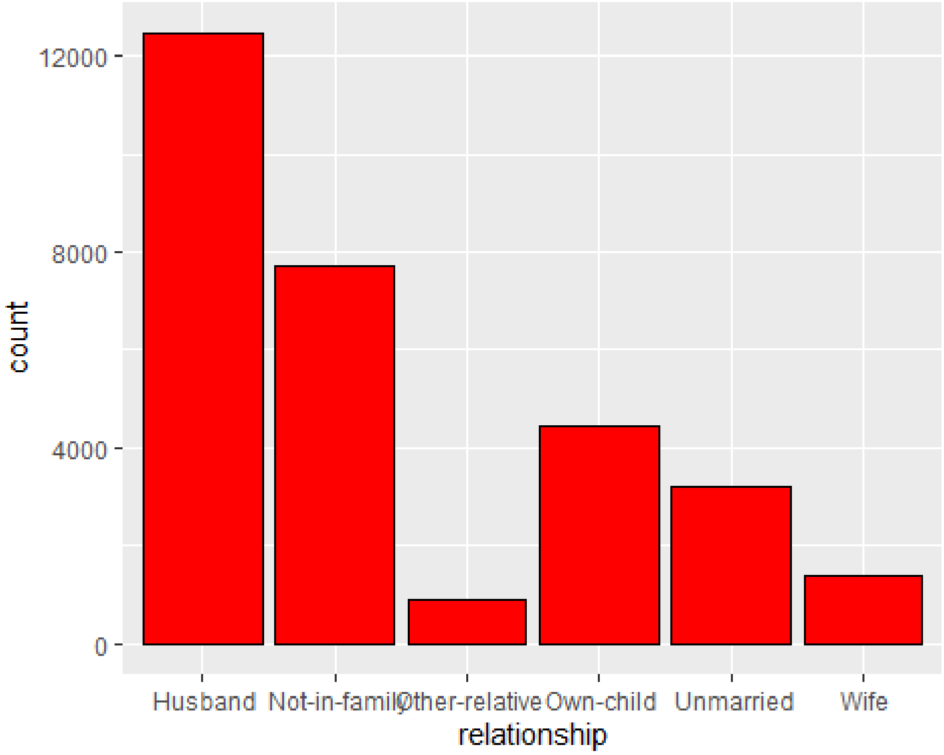
```
re<-table(adult2$relationship)
proportions(re)

##
##        Husband   Not-in-family  Other-relative       Own-child       Unma
rried
##     0.41320204      0.25615012      0.02947417      0.14806710      0.106
49161
##           Wife
##     0.04661495

mosaicplot(table(adult2$relationship),
           color = "pink",
           xlab = "Relationship", # label for x-axis
)
```



```
barplot(table(adult2$relationship))
box(which = "plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```

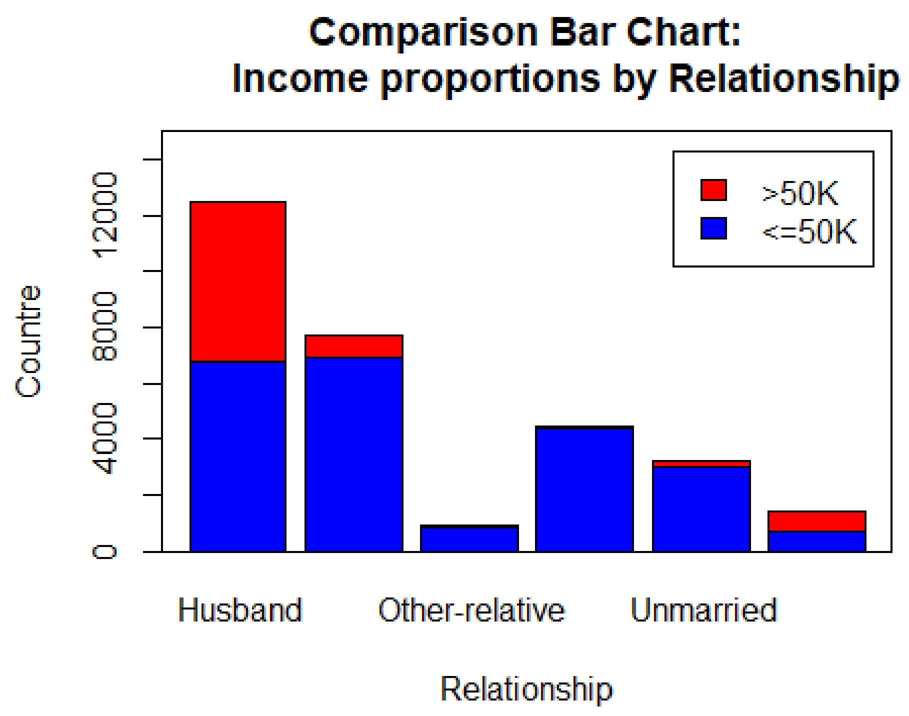


```
ggplot(adult2) +
  aes(x = relationship) +
  geom_bar(fill="red",colour="black")
```



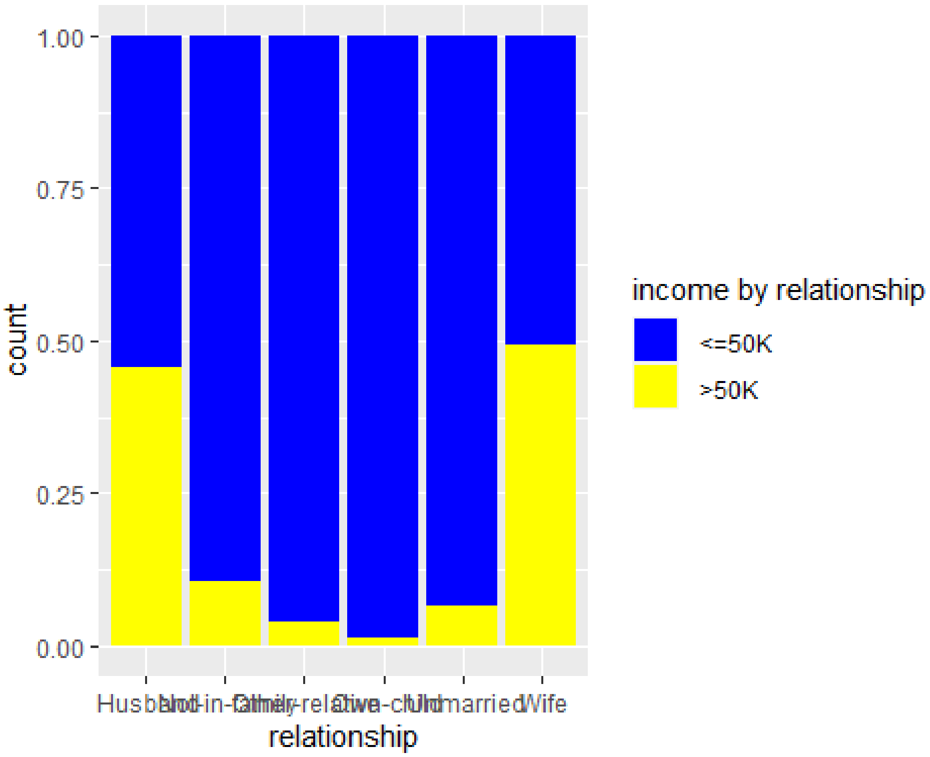
```
#with response
xtabs(~income+relationship,data=adult2)

##         relationship
## income    Husband  Not-in-family  Other-relative  Own-child  Unmarried  Wi
fe
##    <=50K     6784           6903               854       4402        2999   7
12
##    >50K      5679            823                35         64         213   6
94

prop.table(xtabs(~income+relationship,data=adult2))

##         relationship
## income       Husband  Not-in-family  Other-relative   Own-child   Unmarrie
d
##    <=50K 0.224918772    0.228864134     0.028313772 0.145945229 0.09942974
6
##    >50K  0.188283270    0.027285989     0.001160401 0.002121875 0.00706186
6
##         relationship
## income          Wife
##    <=50K 0.023605862
##    >50K  0.023009084

countre<-table(adult2$income,adult2$relationship,
               dnn=c("Income","Relationship"))
countre

##         Relationship
## Income    Husband  Not-in-family  Other-relative  Own-child  Unmarried  Wi
fe
##    <=50K     6784           6903               854       4402        2999   7
12
##    >50K      5679            823                35         64         213   6
94

sumtable<-addmargins(countre,FUN=sum)

## Margins computed over dimensions
## in the following order:
## 1: Income
## 2: Relationship

barplot(countre,
        legend=rownames(countre),
        col=c("blue","red"),
        ylim=c(0,15000),
        ylab="Countre",
        xlab="Relationship",
        main="Comparison Bar Chart:
        Income proportions by Relationship")
```

```
box(which="plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



افرادی که نسبت زن را دارند نسبت حقوق بالای ۵۰ بیشتر از کمتر از ۵۰ است.(باتواجه به جداول بالا و پایین)

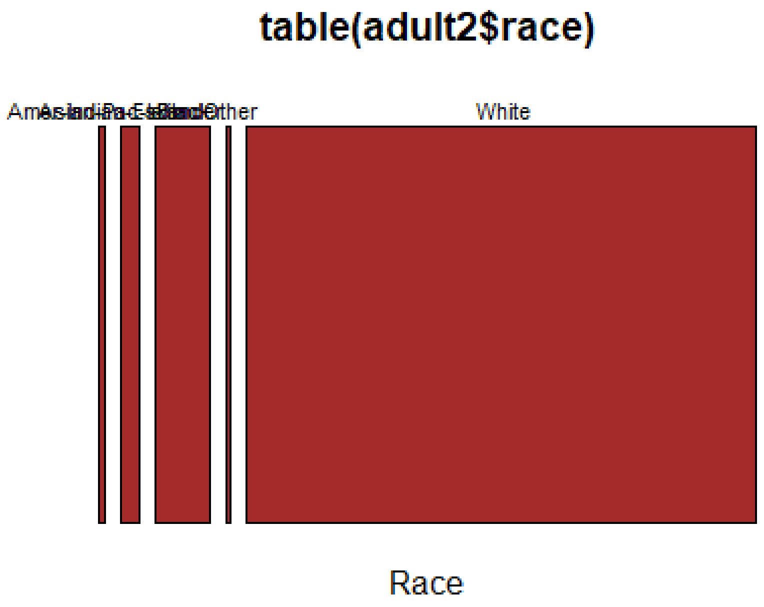
```
#with ggplot
ggplot(adult2,aes(x=relationship,group=income,fill=income))+
  geom_bar(position="fill")+
  scale_fill_manual(values=c("blue","yellow"),name="income by relationship")
```



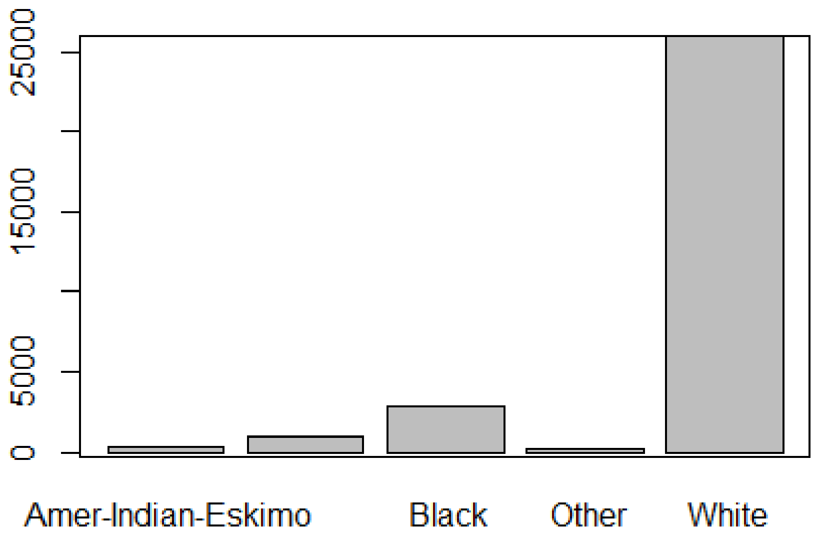
سفیدپوستها فراوانی بیشتری نسبت به سایرین دارد که دراین مساله باید تجدید نظر شود.(چه با حقوق بالا چه پایین!)

#Race

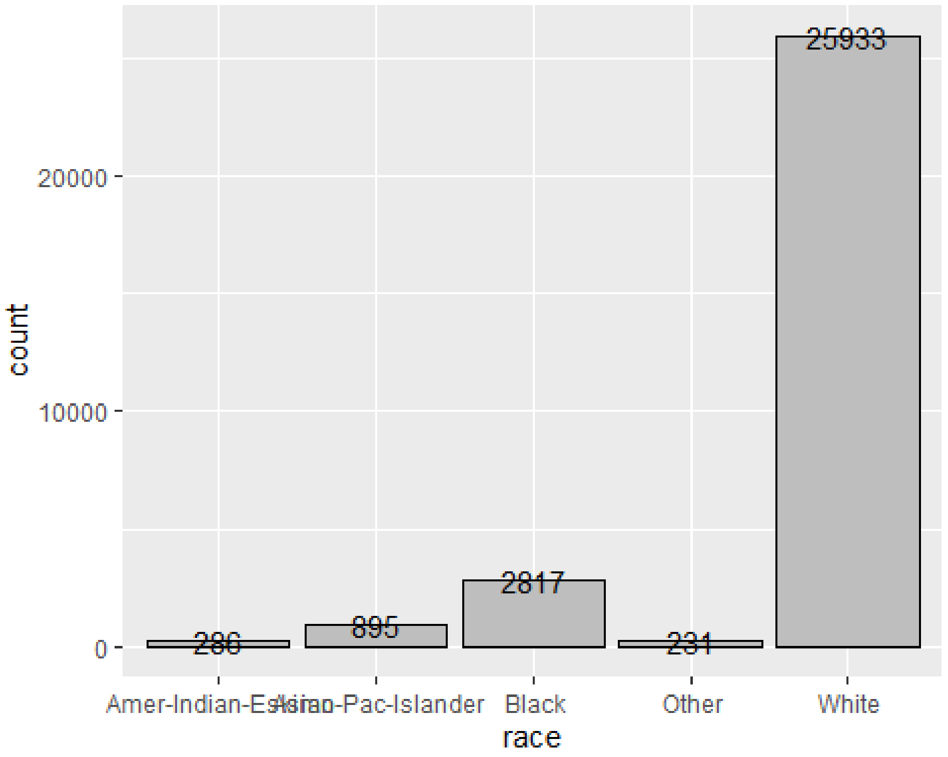
```
ra<-table(adult2$race)
proportions(ra)

##
##  Amer-Indian-Eskimo  Asian-Pac-Islander               Black 
Other
##         0.009482130         0.029673099         0.093395663         0.0076
58643
##               White
##         0.859790465

mosaicplot(table(adult2$race),
           color = "brown",
           xlab = "Race", # label for x-axis
)
```



```
barplot(table(adult2$race))
box(which = "plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```

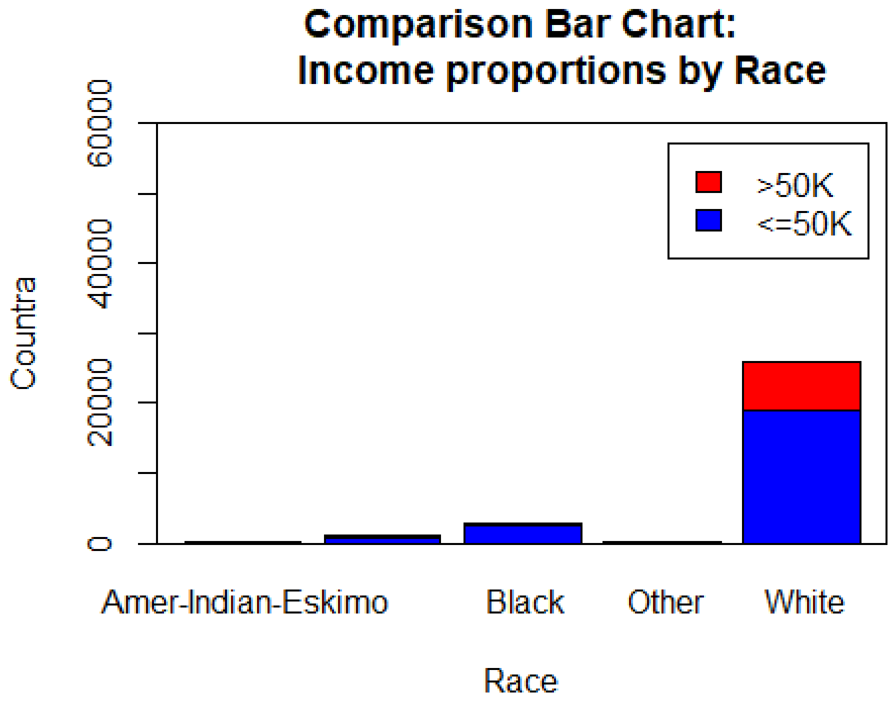


```
ggplot(adult2) +
  aes(x = race) +
  geom_bar(fill="gray",colour="black")+
  geom_text(stat="count",aes(label=..count..))
```



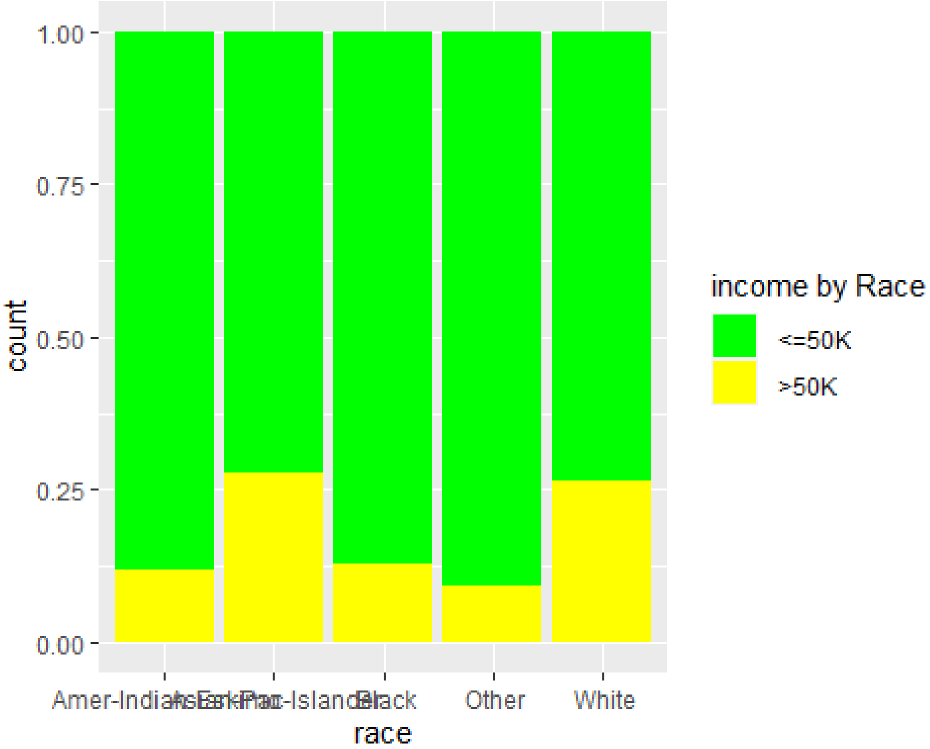
```
#with response
xtabs(~income+race,data=adult2)

##         race
## income   Amer-Indian-Eskimo Asian-Pac-Islander Black Other White
##    <=50K                252                647  2451   210 19094
##    >50K                  34                248   366    21  6839

prop.table(xtabs(~income+race,data=adult2))

##         race
## income   Amer-Indian-Eskimo Asian-Pac-Islander        Black        Other
##    <=50K       0.0083548836       0.0214508322 0.0812611896 0.0069624030
##    >50K        0.0011272462       0.0082222664 0.0121344738 0.0006962403
##         race
## income          White
##    <=50K 0.6330482064
##    >50K  0.2267422585

countra<-table(adult2$income,adult2$race,
               dnn=c("Income","Race"))
countra

##         Race
## Income   Amer-Indian-Eskimo Asian-Pac-Islander Black Other White
##    <=50K                252                647  2451   210 19094
##    >50K                  34                248   366    21  6839
```

```
sumtable<-addmargins(countra,FUN=sum)

## Margins computed over dimensions
## in the following order:
## 1: Income
## 2: Race

barplot(countra,
        legend=rownames(countra),
        col=c("blue","red"),
        ylim=c(0,60000),
        ylab="Countra",
        xlab="Race",
        main="Comparison Bar Chart:
        Income proportions by Race")
box(which="plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



نژاد آسیایی نسبت درآمد بالای ۵۰ش بیشتر از پایین ۵۰ است و به نسبت بیشتر از سایر نژادها حقوق میگیرند.

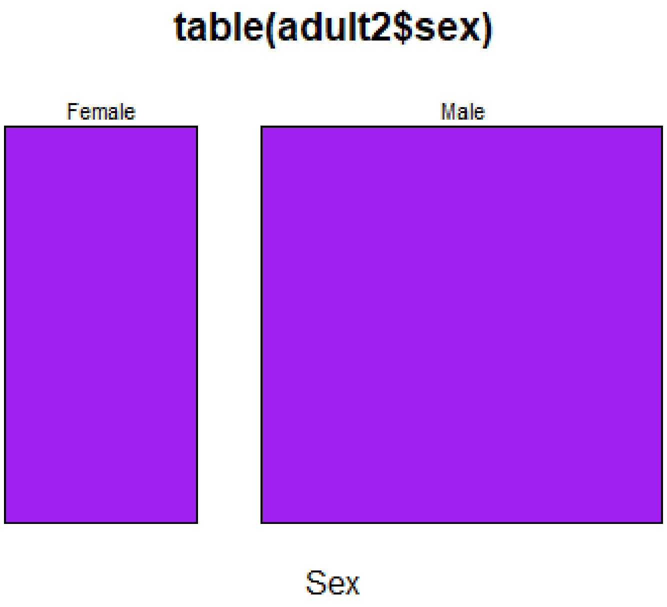
```
#with ggplot
ggplot(adult2,aes(x=race,group=income,fill=income))+
  geom_bar(position="fill")+
  scale_fill_manual(values=c("green","yellow"),name="income by Race")
```



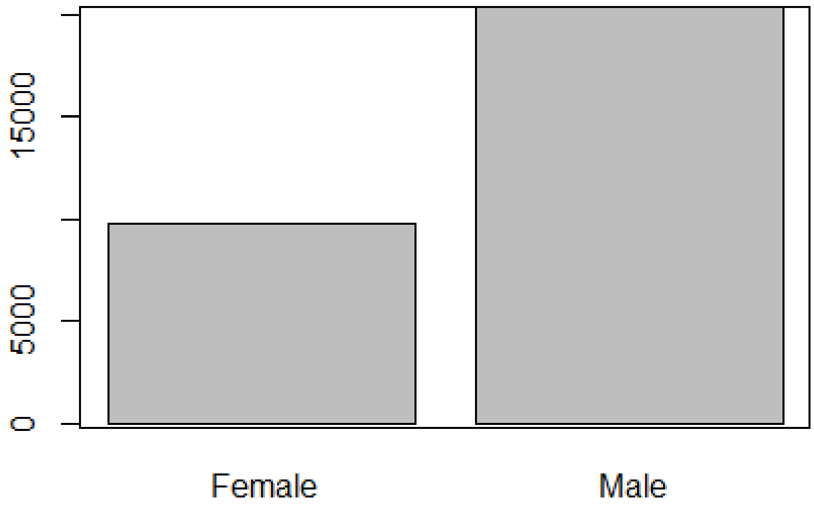
مردها فراوانی بیشتری نسبت به زنها دارند که باز باید تجدید نطر شود.

#Sex

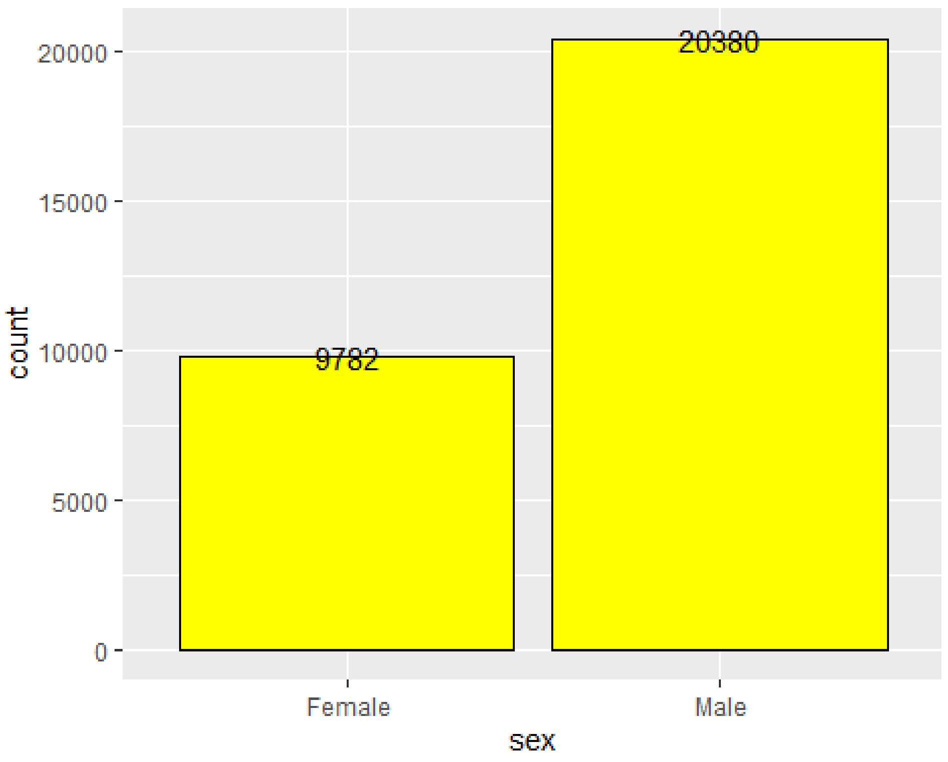
```
s<-table(adult2$sex)
proportions(s)

## 
##    Female      Male 
## 0.3243154 0.6756846

mosaicplot(table(adult2$sex),
           color = "Purple",
           xlab = "Sex", # label for x-axis
)
```



```
barplot(table(adult2$sex))
box(which = "plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```

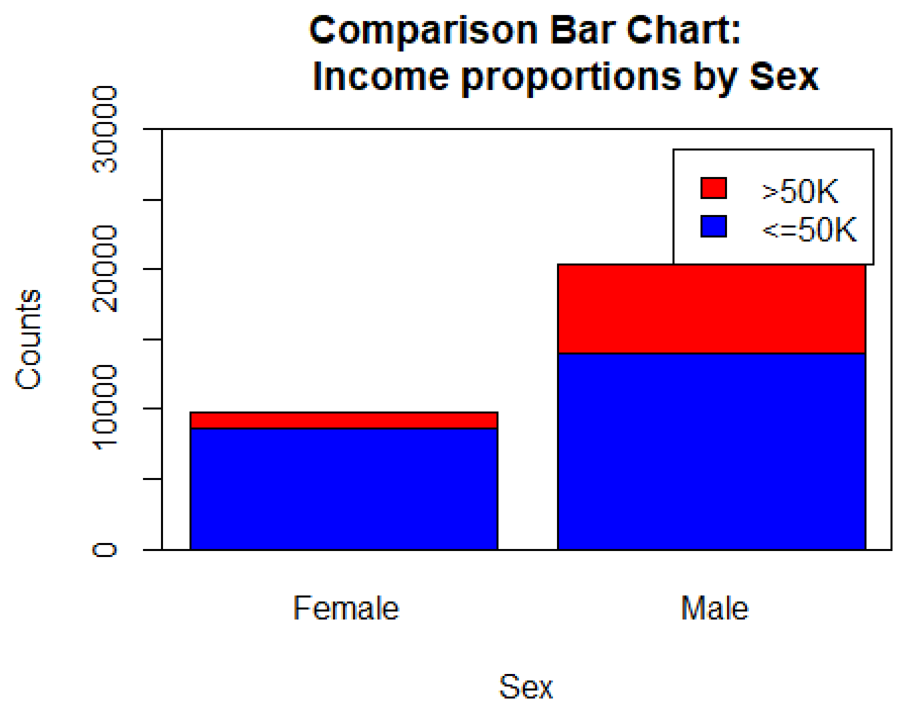


```
ggplot(adult2) +
  aes(x = sex) +
  geom_bar(fill="yellow",colour="black")+
  geom_text(stat="count",aes(label=..count..))
```



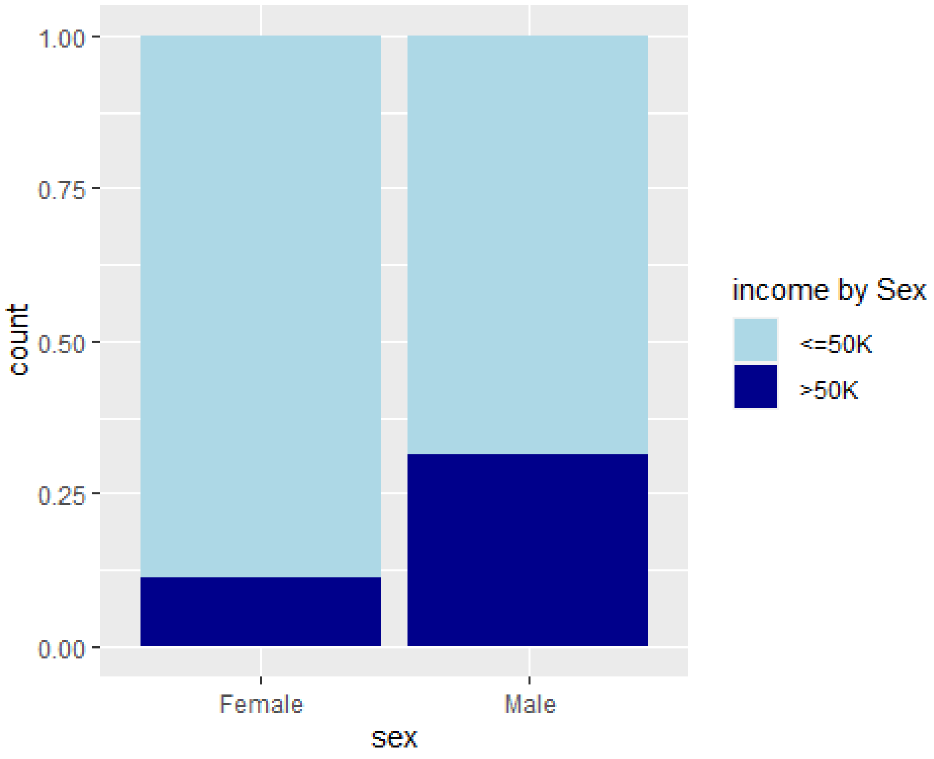
```
#with response
xtabs(~income+sex,data=adult2)

##         sex
## income    Female  Male
##    <=50K    8670 13984
##    >50K     1112  6396

prop.table(xtabs(~income+sex,data=adult2))

##         sex
## income       Female       Male
##    <=50K 0.28744778 0.46362973
##    >50K  0.03686758 0.21205490

counts<-table(adult2$income,adult2$sex,
              dnn=c("Income","Sex"))
counts

##         Sex
## Income    Female  Male
##    <=50K    8670 13984
##    >50K     1112  6396

sumtable<-addmargins(counts,FUN=sum)

## Margins computed over dimensions
## in the following order:
```

```
## 1: Income
## 2: Sex

barplot(counts,
        legend=rownames(counts),
        col=c("blue","red"),
        ylim=c(0,30000),
        ylab="Counts",
        xlab="Sex",
        main="Comparison Bar Chart:
        Income proportions by Sex")
box(which="plot",
    Ity="solid",
    col="black")

## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



مردها نسبت به زنها حقوق بیشتری میگیرند پس چه از نظر اشتغال و چه از نظر درامد باید اصلاح شود.

```
#with ggplot
ggplot(adult2,aes(x=sex,group=income,fill=income))+
  geom_bar(position="fill")+
  scale_fill_manual(values=c("lightblue","darkblue"),name="income by Sex")
```



از متغیر زیر تنها میتوان تبعیض ملیتی را نتیجه گرفت که ایالت متحده تعداد افراد بیشتری را در این شرکتها جا داده است
و نمودار به وضوح نشان میداد از طرفی کشورها تعداد زیادی بودند اما با افراد بسیار کمتر!

#Native.country

```
n<-table(adult2$native.country)
proportions(n)

##
##                    Cambodia                      Canada
##                0.0005967774                0.0035475101
##                       China                    Columbia
##                0.0022544924                0.0018566408
##                        Cuba          Dominican-Republic
##                0.0030501956                0.0022213381
##                     Ecuador                 El-Salvador
##                0.0008951661                0.0033154300
##                     England                      France
##                0.0028512698                0.0008951661
##                     Germany                      Greece
##                0.0042437504                0.0009614747
##                   Guatemala                       Haiti
##                0.0020887209                0.0013924806
##          Holand-Netherlands                    Honduras
##                0.0000331543                0.0003978516
##                        Hong                     Hungary
##                0.0006299317                0.0004310059
##                     India                        Iran
##              0.0033154300                0.0013924806
##                   Ireland                       Italy
##              0.0007957032                0.0022544924
##                   Jamaica                       Japan
##              0.0026523440                0.0019561037
##                      Laos                      Mexico
##              0.0005636231                0.0202241231
##                 Nicaragua  Outlying-US(Guam-USVI-etc)
##              0.0010940919                0.0004641602
##                      Peru                 Philippines
##              0.0009946290                0.0062330084
##                    Poland                    Portugal
##              0.0018566408                0.0011272462
##               Puerto-Rico                    Scotland
##              0.0036138187                0.0003646973
##                     South                      Taiwan
##              0.0023539553                0.0013924806
##                  Thailand             Trinadad&Tobago
##              0.0005636231                0.0005967774
##             United-States                     Vietnam
##              0.9118758703                0.0021218752
##                Yugoslavia
##              0.0005304688
```

```
#with response:
xtabs(~income+native.country,data=adult2)
```

```
##         native.country
## income    Cambodia  Canada  China  Columbia  Cuba  Dominican-Republic  Ecu
ador
##    <=50K        11      71     48        54    67                  65
23
##    >50K          7      36     20         2    25                   2
4
##         native.country
## income    El-Salvador  England  France  Germany  Greece  Guatemala  Haiti
##    <=50K           91       56      15       84      21         60     38
##    >50K             9       30      12       44       8          3      4
##         native.country
## income    Holand-Netherlands  Honduras  Hong  Hungary  India  Iran  Irelan
d
##    <=50K                   1        11    13       10     60    24       1
9
##    >50K                    0         1     6        3     40    18
5
##         native.country
## income    Italy  Jamaica  Japan  Laos  Mexico  Nicaragua
##    <=50K     44       70     36    15     577         31
##    >50K      24       10     23     2      33          2
##         native.country
## income    Outlying-US(Guam-USVI-etc)  Peru  Philippines  Poland  Portugal
##    <=50K                          14    28          128      45        30
##    >50K                            0     2           60      11         4
##         native.country
## income    Puerto-Rico  Scotland  South  Taiwan  Thailand  Trinadad&Tobago
##    <=50K           97         9     57      23        14               16
##    >50K            12         2     14      19         3                2
##         native.country
## income    United-States  Vietnam  Yugoslavia
##    <=50K          20509       59          10
##    >50K            6995        5           6
```

```
prop.table(xtabs(~income+native.country,data=adult2))
```

```
##         native.country
## income       Cambodia       Canada        China     Columbia         Cuba
##    <=50K 0.0003646973 0.0023539553 0.0015914064 0.0017903322 0.0022213381
##    >50K  0.0002320801 0.0011935548 0.0006630860 0.0000663086 0.0008288575
##         native.country
## income    Dominican-Republic      Ecuador  El-Salvador      England
##    <=50K        0.0021550295 0.0007625489 0.0030170413 0.0018566408
##    >50K         0.0000663086 0.0001326172 0.0002983887 0.0009946290
##         native.country
## income         France      Germany       Greece    Guatemala        Haiti
##    <=50K 0.0004973145 0.0027849612 0.0006962403 0.0019892580 0.0012598634
##    >50K  0.0003978516 0.0014587892 0.0002652344 0.0000994629 0.0001326172
##         native.country
## income    Holand-Netherlands     Honduras         Hong      Hungary
##    <=50K        0.0000331543 0.0003646973 0.0004310059 0.0003315430
##    >50K         0.0000000000 0.0000331543 0.0001989258 0.0000994629
##         native.country
## income          India         Iran      Ireland        Italy      Jamaica
##    <=50K 0.0019892580 0.0007957032 0.0006299317 0.0014587892 0.0023208010
##    >50K  0.0013261720 0.0005967774 0.0001657715 0.0007957032 0.0003315430
##         native.country
## income          Japan         Laos       Mexico    Nicaragua
##    <=50K 0.0011935548 0.0004973145 0.0191300312 0.0010277833
##    >50K  0.0007625489 0.0000663086 0.0010940919 0.0000663086
##         native.country
## income    Outlying-US(Guam-USVI-etc)         Peru  Philippines        Polan
d
##    <=50K                0.0004641602 0.0009283204 0.0042437504 0.001491943
5
##    >50K                 0.0000000000 0.0000663086 0.0019892580 0.000364697
3
##         native.country
## income       Portugal  Puerto-Rico     Scotland        South       Taiwan
##    <=50K 0.0009946290 0.0032159671 0.0002983887 0.0018897951 0.0007625489
##    >50K  0.0001326172 0.0003978516 0.0000663086 0.0004641602 0.0006299317
##          native.country
## income       Thailand  Trinadad&Tobago  United-States      Vietnam   Yugos
lavia
##    <=50K 0.0004641602     0.0005304688   0.6799615410 0.0019561037 0.00033
15430
##    >50K  0.0000994629     0.0000663086   0.2319143293 0.0001657715 0.00019
89258
```

```
countn<-table(adult2$income,adult2$native.country,
              dnn=c("Income","Native.country"))
counts
```

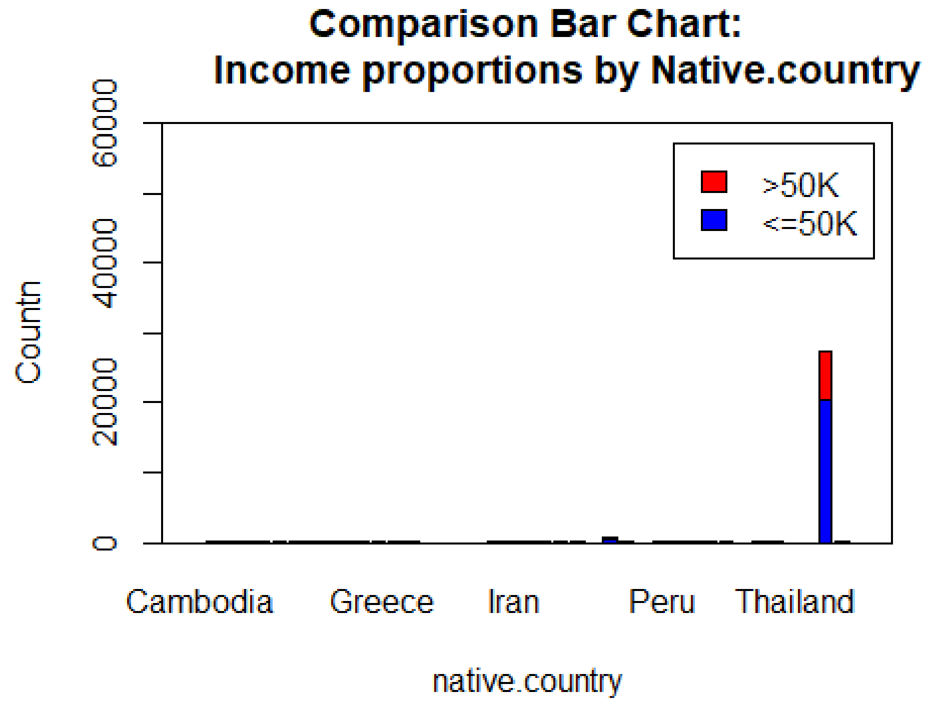
```
##          Sex
## Income    Female  Male
##    <=50K    8670 13984
##    >50K     1112  6396
```

```
sumtable<-addmargins(countn,FUN=sum)
```

```
## Margins computed over dimensions
## in the following order:
## 1: Income
## 2: Native.country
```

```
barplot(countn,
        legend=rownames(countn),
        col=c("blue","red"),
        ylim=c(0,60000),
        ylab="Countn",
        xlab="native.country",
        main="Comparison Bar Chart:
        Income proportions by Native.country")
box(which="plot",
    Ity="solid",
    col="black")
```

```
## Warning in box(which = "plot", Ity = "solid", col = "black"): "Ity" is not
a
## graphical parameter
```



در این قسمت دو متغیر را با متغیر مورد علاقه بررسی میکنیم

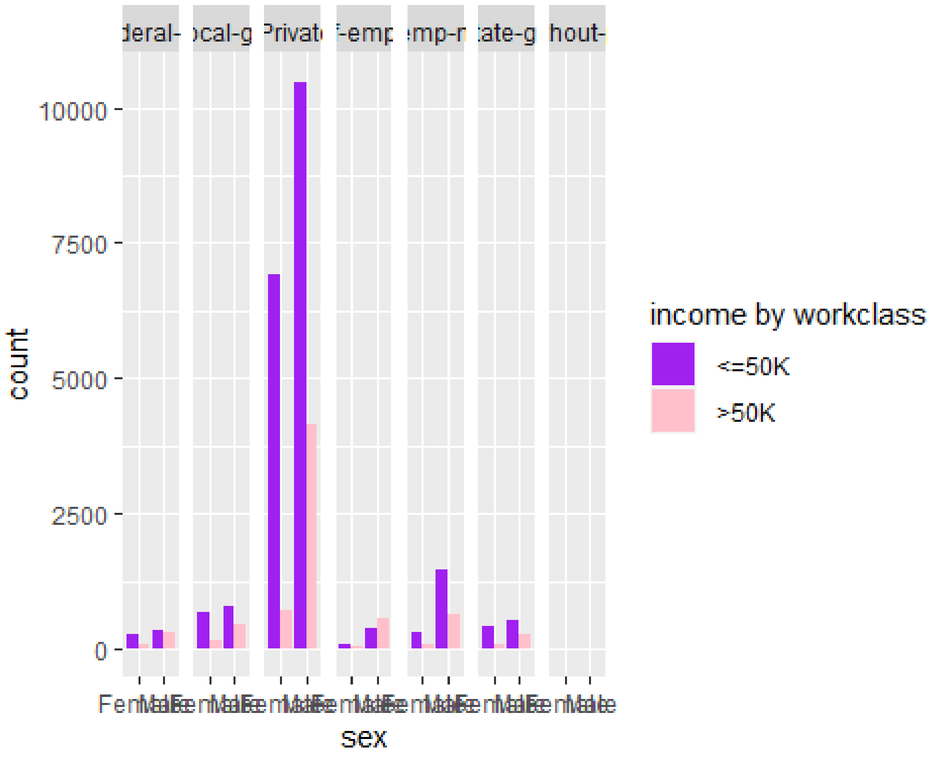
#More than 2 variables

```
xtabs(~income+sex+workclass,data=adult2)

## , , workclass =  Federal-gov
##
##         sex
## income    Female  Male
##    <=50K     254   324
##    >50K       55   310
##
## , , workclass =  Local-gov
##
##         sex
## income    Female  Male
##    <=50K     672   786
##    >50K      152   457
##
## , , workclass =  Private
##
##         sex
## income    Female  Male
##    <=50K    6921 10489
##    >50K      721  4155
##
```

```
## , , workclass =  Self-emp-inc
## 
##         sex
## income    Female  Male
##    <=50K      88   386
##    >50K       38   562
## 
## , , workclass =  Self-emp-not-inc
## 
##         sex
## income    Female  Male
##    <=50K     312  1473
##    >50K       80   634
## 
## , , workclass =  State-gov
## 
##         sex
## income    Female  Male
##    <=50K     418   517
##    >50K       66   278
## 
## , , workclass =  Without-pay
## 
##         sex
## income    Female  Male
##    <=50K       5     9
##    >50K        0     0
```

```
ggplot(adult2,aes(x=sex,group=income,fill=income))+
  geom_bar(position=position_dodge())+
  scale_fill_manual(values=c("purple","pink"),name="income by workclass")+fac
et_grid(~workclass)
```

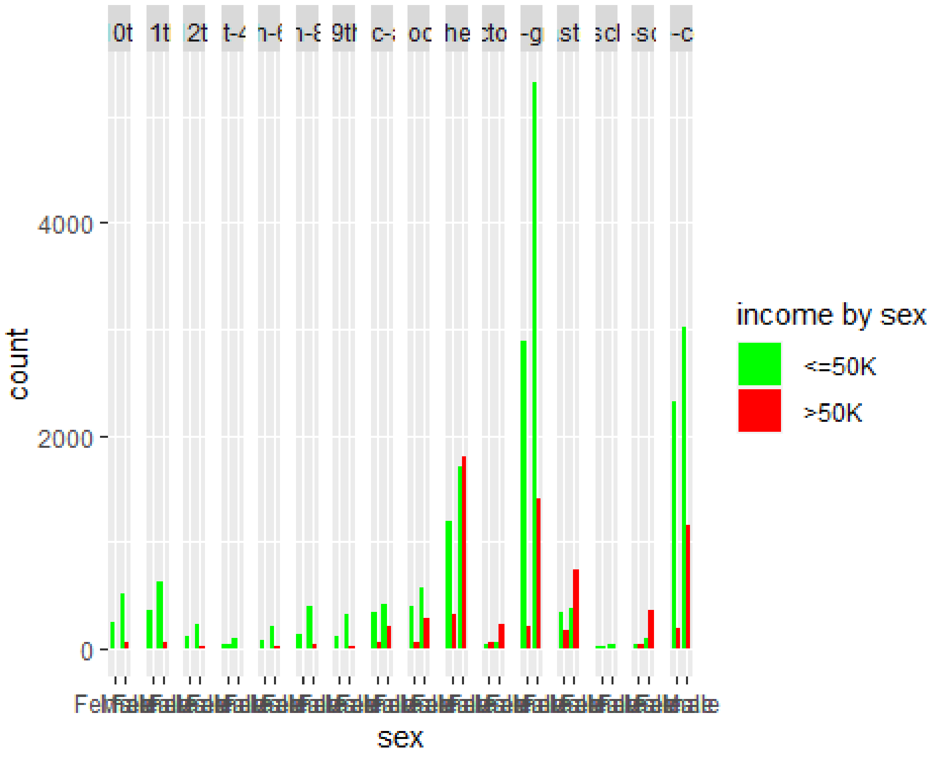
از نمودار پایین نتیجه میگیریم که در انواع ادارات مردها نسبت حقوق بیشتری نسبت به زن ها دارند.



```
xtabs(~income+sex+education,data=adult2)

## , , education =  10th
##
##          sex
## income    Female  Male
##    <=50K     248   513
##    >50K        2    57
##
## , , education =  11th
##
##          sex
## income    Female  Male
##    <=50K     363   626
##    >50K        8    51
##
## , , education =  12th
##
##          sex
## income    Female  Male
##    <=50K     120   228
##    >50K        2    27
##
## , , education =  1st-4th
##
##          sex
## income    Female  Male
##    <=50K      43   102
##    >50K        0     6
## 
## , , education =  5th-6th
## 
##          sex
## income    Female  Male
##    <=50K      67   209
##    >50K        2    10
## 
## , , education =  7th-8th
## 
##          sex
## income    Female  Male
##    <=50K     131   391
##    >50K        1    34
## 
## , , education =  9th
## 
##          sex
## income    Female  Male
##    <=50K     114   316
##    >50K        5    20
## 
## , , education =  Assoc-acdm
## 
##          sex
## income    Female  Male
##    <=50K     342   410
##    >50K       53   203
## 
## , , education =  Assoc-voc
## 
##          sex
## income    Female  Male
##    <=50K     394   569
##    >50K       61   283
## 
## , , education =  Bachelors
## 
##          sex
## income    Female  Male
##    <=50K    1205  1713
##    >50K      317  1809
## 
## , , education =  Doctorate
## 
##          sex
## income    Female  Male
##    <=50K      32    63
##    >50K       49   231
## 
## , , education =  HS-grad
## 
##         sex
## income    Female  Male
##    <=50K    2893  5330
##    >50K      213  1404
## 
## , , education =  Masters
## 
##         sex
## income    Female  Male
##    <=50K     337   372
##    >50K      172   746
## 
## , , education =  Preschool
## 
##         sex
## income    Female  Male
##    <=50K      14    31
##    >50K        0     0
## 
## , , education =  Prof-school
## 
##         sex
## income    Female  Male
##    <=50K      45    91
##    >50K       42   364
## 
## , , education =  Some-college
## 
##         sex
## income    Female  Male
##    <=50K    2322  3020
##    >50K      185  1151
```

```
ggplot(adult2,aes(x=sex,group=income,fill=income))+
  geom_bar(position=position_dodge())+
  scale_fill_manual(values=c("green","red"),name="income by sex")+facet_grid(
~education)
```

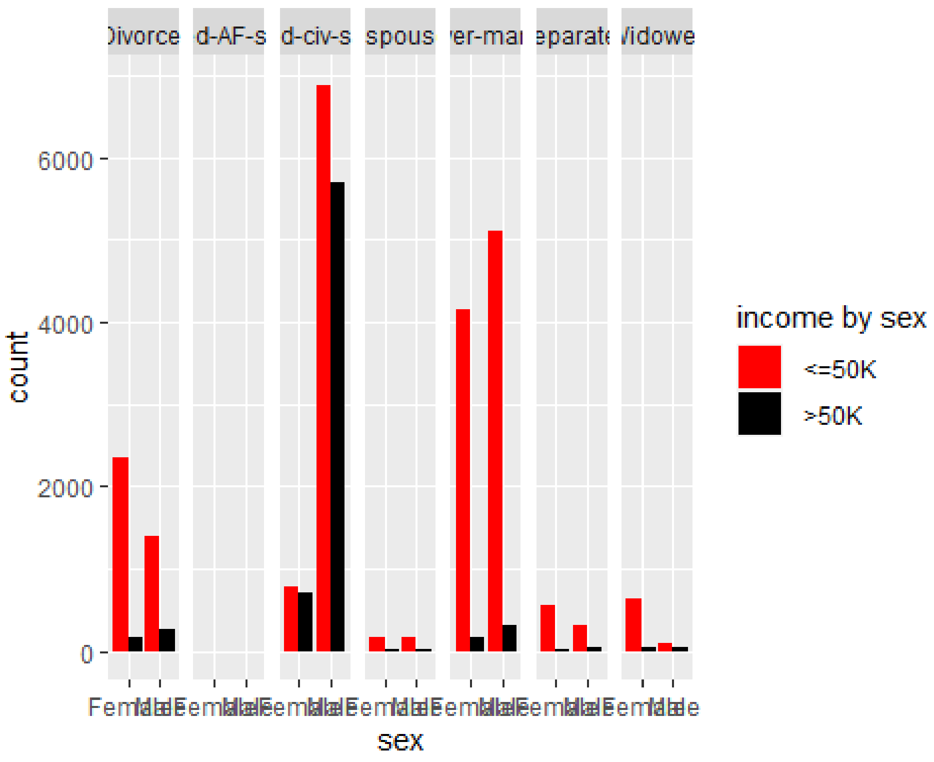
در مورد تحصیلات نمی توان درامد را برای مرد و زن تعمیم داد اما باز هم نسبت حقوق بالای۵۰در مردها در اکثر موارد بیش از زنان است.



```
xtabs(~income+sex+marital.status,data=adult2)

## , , marital.status =  Divorced
## 
##         sex
## income    Female  Male
##    <=50K    2355  1407
##    >50K      174   278
## 
## , , marital.status =  Married-AF-spouse
## 
##         sex
## income    Female  Male
##    <=50K       6     5
##    >50K        6     4
## 
## , , marital.status =  Married-civ-spouse
## 
##         sex
## income    Female  Male
##    <=50K     780  6886
##    >50K      700  5699
## 
## , , marital.status =  Married-spouse-absent
## 
##         sex
## income    Female  Male
##    <=50K      178   161
##    >50K        11    20
##
## , , marital.status =  Never-married
##
##         sex
## income    Female  Male
##    <=50K    4149  5107
##    >50K      163   307
##
## , , marital.status =  Separated
##
##         sex
## income    Female  Male
##    <=50K     557   316
##    >50K       17    49
##
## , , marital.status =  Widowed
##
##         sex
## income    Female  Male
##    <=50K     645   102
##    >50K       41    39
```

```
ggplot(adult2,aes(x=sex,group=income,fill=income))+
  geom_bar(position=position_dodge())+
  scale_fill_manual(values=c("red","black"),name="income by sex")+facet_grid(
~marital.status)
```

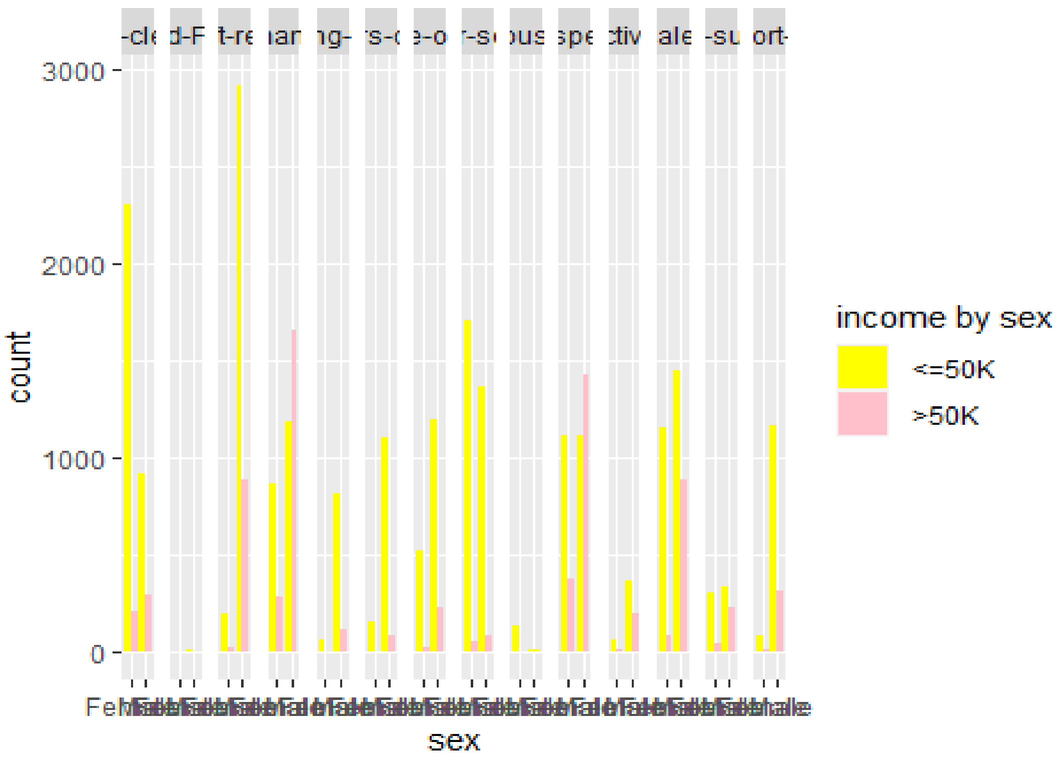
نسبت درامد در مرد و زن براساس وضعیت تاهل تقریبا یکسان می باشد.



```
xtabs(~income+sex+occupation,data=adult2)

## , , occupation =  Adm-clerical
##
##          sex
## income    Female  Male
##    <=50K    2303   920
##    >50K      209   289
##
## , , occupation =  Armed-Forces
##
##          sex
## income    Female  Male
##    <=50K       0     8
##    >50K        0     1
##
## , , occupation =  Craft-repair
##
##          sex
## income    Female  Male
##    <=50K     197  2925
##    >50K       19   889
##
## , , occupation =  Exec-managerial
##
##          sex
## income    Female  Male
##    <=50K     866  1189
##    >50K      277  1660
## 
## , , occupation =  Farming-fishing
## 
##         sex
## income    Female  Male
##    <=50K      63   811
##    >50K        2   113
## 
## , , occupation =  Handlers-cleaners
## 
##         sex
## income    Female  Male
##    <=50K     160  1107
##    >50K        4    79
## 
## , , occupation =  Machine-op-inspct
## 
##         sex
## income    Female  Male
##    <=50K     523  1198
##    >50K       20   225
## 
## , , occupation =  Other-service
## 
##         sex
## income    Female  Male
##    <=50K    1709  1371
##    >50K       49    83
## 
## , , occupation =  Priv-house-serv
## 
##         sex
## income    Female  Male
##    <=50K     134     8
##    >50K        1     0
## 
## , , occupation =  Prof-specialty
## 
##         sex
## income    Female  Male
##    <=50K    1111  1116
##    >50K      380  1431
## 
## , , occupation =  Protective-serv
## 
##         sex
## income    Female  Male
##    <=50K      66   368
##    >50K       10   200
## 
## , , occupation =  Sales
## 
##         sex
## income    Female  Male
##    <=50K    1160  1454
##    >50K       88   882
## 
## , , occupation =  Tech-support
## 
##         sex
## income    Female  Male
##    <=50K     297   337
##    >50K       44   234
## 
## , , occupation =  Transport-moving
## 
##         sex
## income    Female  Male
##    <=50K      81  1172
##    >50K        9   310
```

```
ggplot(adult2,aes(x=sex,group=income,fill=income))+
  geom_bar(position=position_dodge())+
  scale_fill_manual(values=c("yellow","pink"),name="income by sex")+facet_gri
d(~occupation)
```

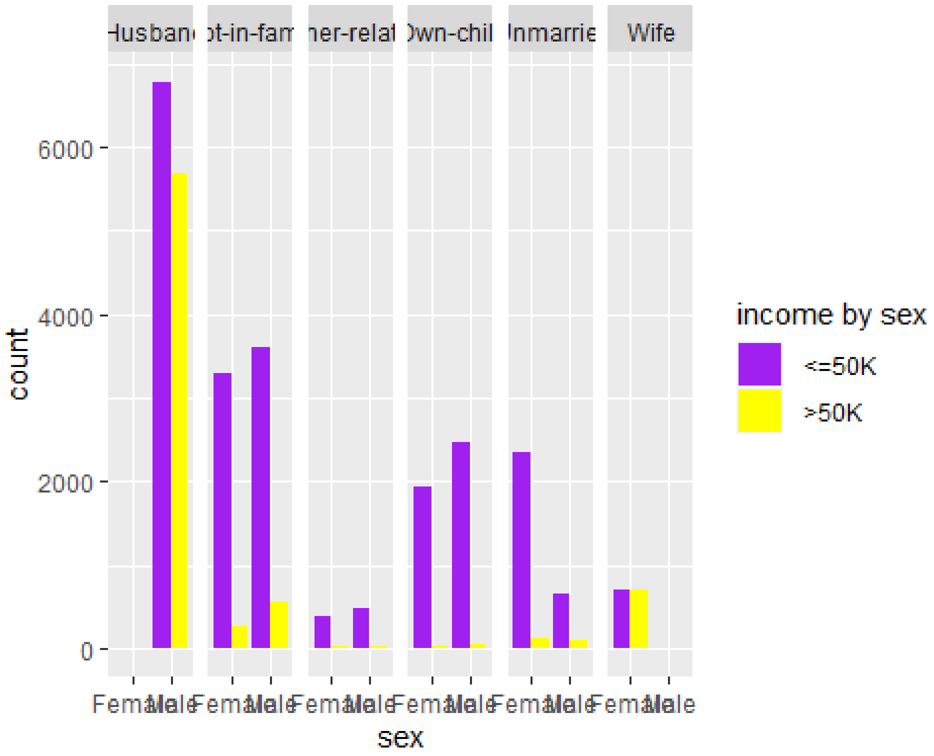
در قسمت شغل هم باز مردها نسبت حقوق بالای ۵۰ شان نسبت به زنها بیشتر است.



```
xtabs(~income+sex+relationship,data=adult2)

## , , relationship =  Husband
## 
##         sex
## income    Female  Male
##    <=50K       1  6783
##    >50K        0  5679
## 
## , , relationship =  Not-in-family
## 
##         sex
## income    Female  Male
##    <=50K    3291  3612
##    >50K      275   548
## 
## , , relationship =  Other-relative
## 
##         sex
## income    Female  Male
##    <=50K     374   480
##    >50K       12    23
## 
## , , relationship =  Own-child
## 
##         sex
## income    Female  Male
##     <=50K     1938  2464
##     >50K        23    41
##
## , , relationship =  Unmarried
##
##          sex
## income    Female  Male
##     <=50K     2354   645
##     >50K       109   104
##
## , , relationship =  Wife
##
##          sex
## income    Female  Male
##     <=50K      712     0
##     >50K       693     1
```

```
ggplot(adult2,aes(x=sex,group=income,fill=income))+
  geom_bar(position=position_dodge())+
  scale_fill_manual(values=c("purple","yellow"),name="income by sex")+facet_g
rid(~relationship)
```

در مورد روابط هم با قاطعیت نمیتوان بین زن و مرد اظهار نظر کرد چون برخی مواقع زنان نسبت افرادی که بالای ۵۰ حقوق می گیرند بیشتر است و در برخی مواقع مردان!

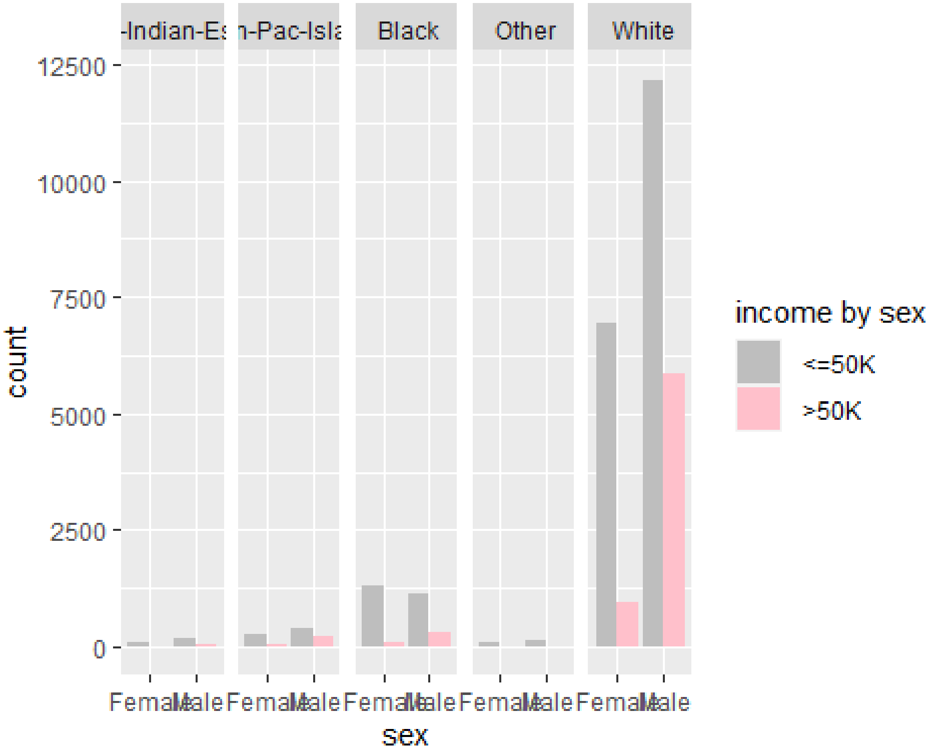


```
xtabs(~income+sex+race,data=adult2)
```

```
## , , race =  Amer-Indian-Eskimo
## 
##          sex
## income    Female  Male
##    <=50K      96   156
##    >50K       11    23
## 
## , , race =  Asian-Pac-Islander
## 
##          sex
## income    Female  Male
##    <=50K     253   394
##    >50K       41   207
## 
## , , race =  Black
## 
##          sex
## income    Female  Male
##    <=50K    1314  1137
##    >50K       85   281
## 
## , , race =  Other
## 
##          sex
## income    Female  Male
##    <=50K      83   127
##    >50K        4    17
## 
## , , race =  White
## 
##          sex
## income    Female  Male
##    <=50K    6924 12170
##    >50K      971  5868
```

```
ggplot(adult2,aes(x=sex,group=income,fill=income))+
  geom_bar(position=position_dodge())+
  scale_fill_manual(values=c("gray","pink"),name="income by sex")+facet_grid(
~race)
```

در بخش نژاد هم باز زنان حقوق کمتری به نسبت مردان دریافت می کنند(ممکن است زنان آسیایی وضعیت به نسبت بهتری از سایر زنان داشته باشند.)



#Q8

طبق خواسته سوال میانگین،میانه،مینیمم،ماکزیمم و استاندارد ارور را حساب میکنیم:

```
##Mean
summary(adult2)

##       age         workclass             fnlwgt          education        
##  Min.   :17.00   Length:30162       Min.   :  13769   Length:30162      
##  1st Qu.:28.00   Class :character   1st Qu.: 117627   Class :character  
##  Median :37.00   Mode  :character   Median : 178425   Mode  :character  
##  Mean   :38.44                      Mean   : 189794                     
##  3rd Qu.:47.00                      3rd Qu.: 237629                     
##  Max.   :90.00                      Max.   :1484705                     
##  education.num   marital.status      occupation        relationship      
##  Min.   : 1.00   Length:30162       Length:30162       Length:30162      
##  1st Qu.: 9.00   Class :character   Class :character   Class :character  
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :10.12                                                           
##  3rd Qu.:13.00                                                           
##  Max.   :16.00                                                           
##      race               sex             capital.gain    capital.loss    
##  Length:30162       Length:30162       Min.   :    0   Min.   :   0.00  
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.00  
##  Mode  :character   Mode  :character   Median :    0   Median :   0.00  
##                                        Mean   : 1092   Mean   :  88.37  
##                                        3rd Qu.:    0   3rd Qu.:   0.00  
##                                         Max.   :99999   Max.   :4356.00  
##  hours.per.week  native.country        income         
##  Min.   : 1.00   Length:30162       Length:30162      
##  1st Qu.:40.00   Class :character   Class :character  
##  Median :40.00   Mode  :character   Mode  :character  
##  Mean   :40.93                                        
##  3rd Qu.:45.00                                        
##  Max.   :99.00
```

```
mean(adult2$age)
```

```
## [1] 38.4379
```

```
mean(adult2$fnlwgt)
```

```
## [1] 189793.8
```

```
mean(adult2$capital.gain)
```

```
## [1] 1092.008
```

```
mean(adult2$capital.loss)
```

```
## [1] 88.37249
```

```
mean(adult2$hours.per.week)
```

```
## [1] 40.93124
```

```
##Median
median(adult2$age)
```

```
## [1] 37
```

```
median(adult2$fnlwgt)
```

```
## [1] 178425
```

```
median(adult2$capital.gain)
```

```
## [1] 0
```

```
median(adult2$capital.loss)
```

```
## [1] 0
```

```
median(adult2$hours.per.week)
```

```
## [1] 40
```

```
##Minimum
min(adult2$age)
```

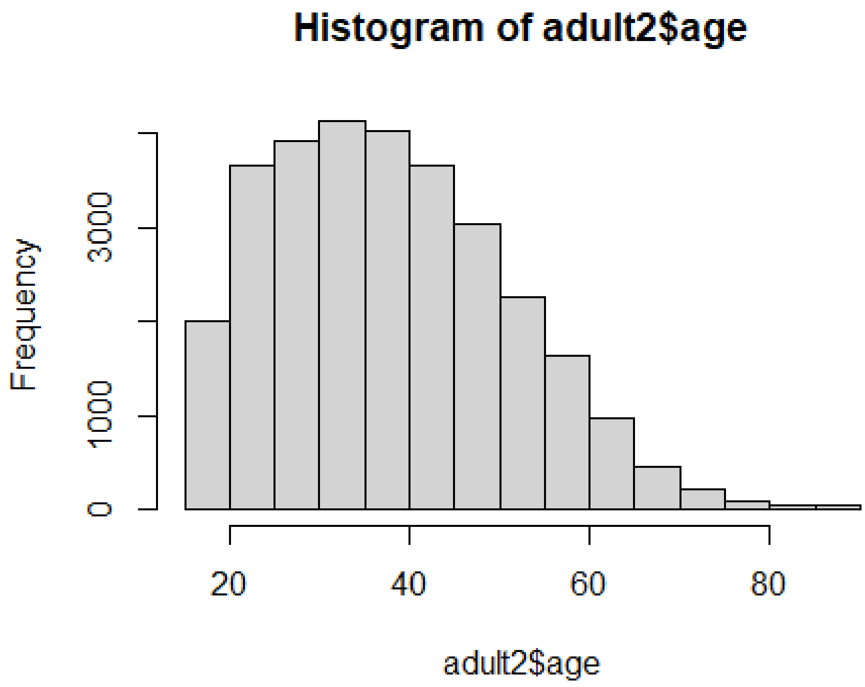
```
## [1] 17
```

```
min(adult2$fnlwgt)

## [1] 13769

min(adult2$capital.gain)

## [1] 0

min(adult2$capital.loss)

## [1] 0

min(adult2$hours.per.week)

## [1] 1

##Maximum
max(adult2$age)

## [1] 90

max(adult2$fnlwgt)

## [1] 1484705

max(adult2$capital.gain)

## [1] 99999

max(adult2$capital.loss)

## [1] 4356

max(adult2$hours.per.week)

## [1] 99

##SD
sd(adult2$age)

## [1] 13.13466

sd(adult2$fnlwgt)

## [1] 105653

sd(adult2$capital.gain)

## [1] 7406.346

sd(adult2$capital.loss)

## [1] 404.2984

sd(adult2$hours.per.week)
```

```
## [1] 11.97998
```

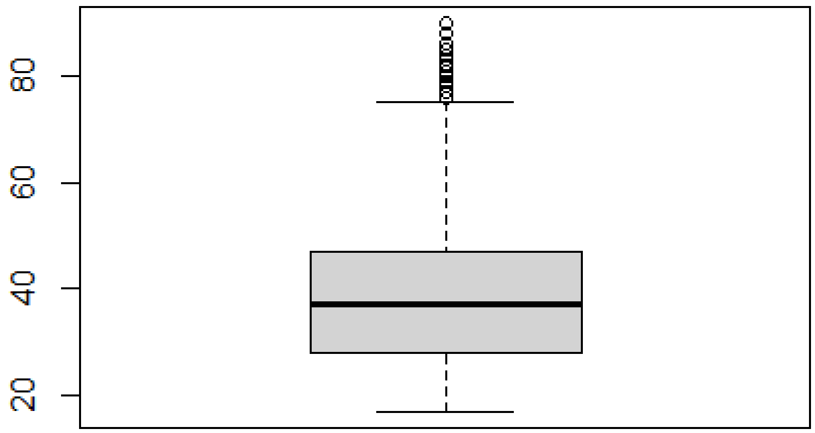
#Q9to11 #Age

```
library(ggplot2)
#table(adult2$age)
hist(adult2$age)
```

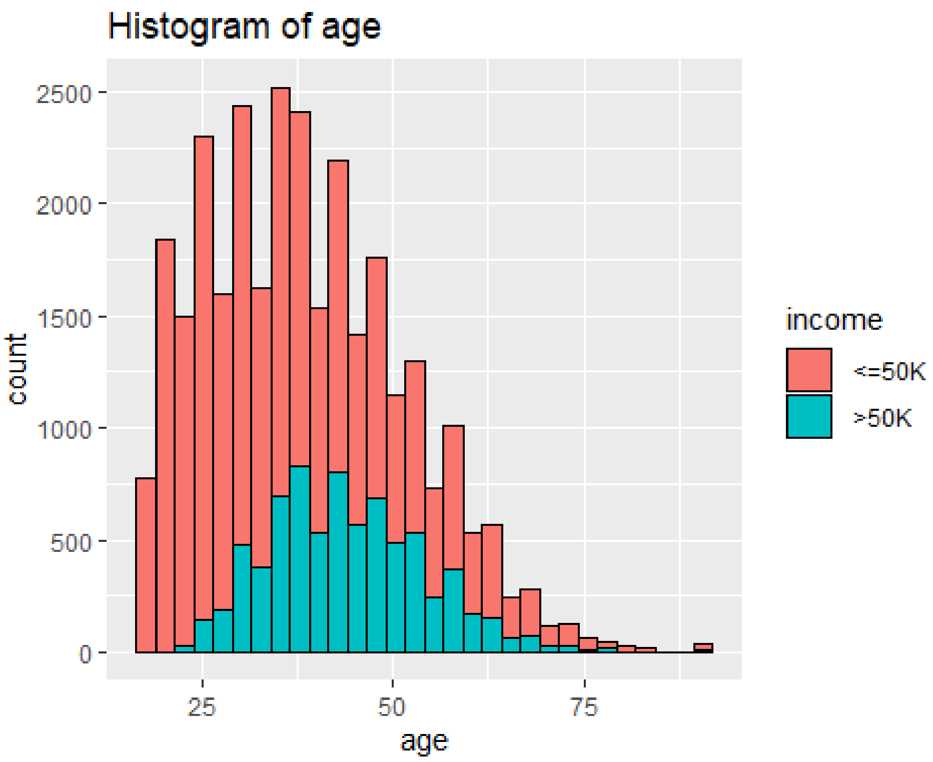


تعداد افرااد میانسال و بعد از آن جوانان بیشتر است.

```
boxplot(adult2$age)
```

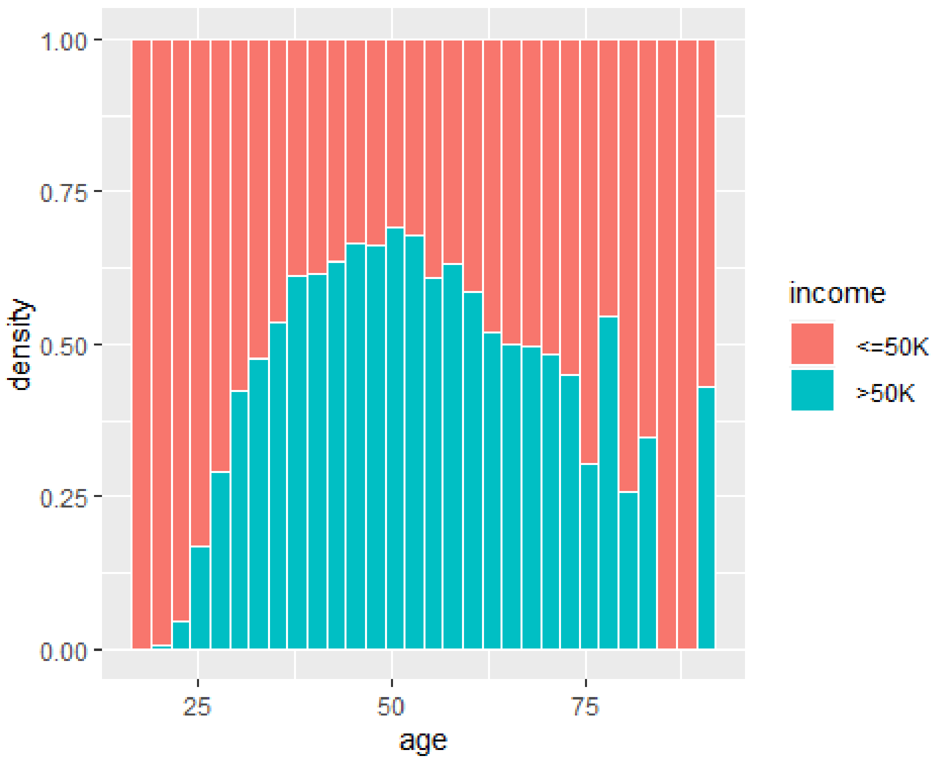


```
ggplot(data=adult2,aes(x=age,fill=income)) +
  geom_histogram(col="black")+
  labs(title="Histogram of age",x="age")

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



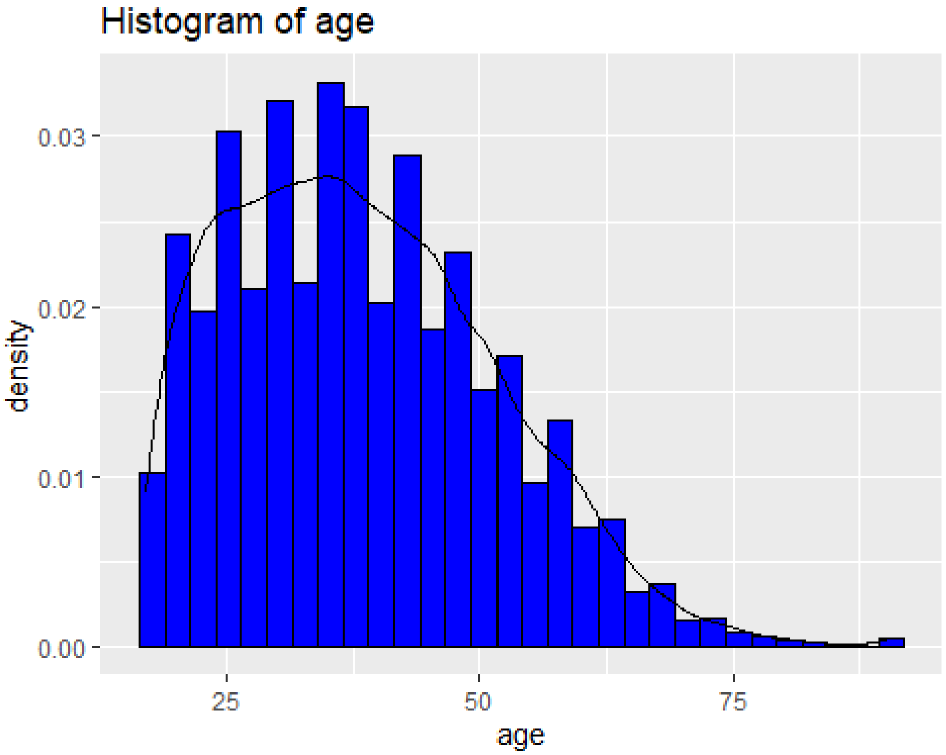
```
summary(adult2$age[adult2$income==" >50K"]);summary(adult2$age[adult2$income=
=" >50K"])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   36.00   43.00   43.96   51.00   90.00

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   36.00   43.00   43.96   51.00   90.00

ggplot(data=adult2,aes(x=age,fill=income))+
  geom_histogram(aes(y=..density..),col="white",position="fill")+
  labs("Histogram of Age",x="age")

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

از نمودار بالا و پایین می توان نتیجه گرفت که افراد میانسال درامد بیشتری دارند. واضح است که درآمد بازنشسته ها باید افزایش پیدا کند.



```
ggplot(data=adult2,aes(x=age))+
  geom_histogram(aes(y=..density..),fill="blue",col="black")+
  labs(title="Histogram of age",x="age")+
  geom_density()

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

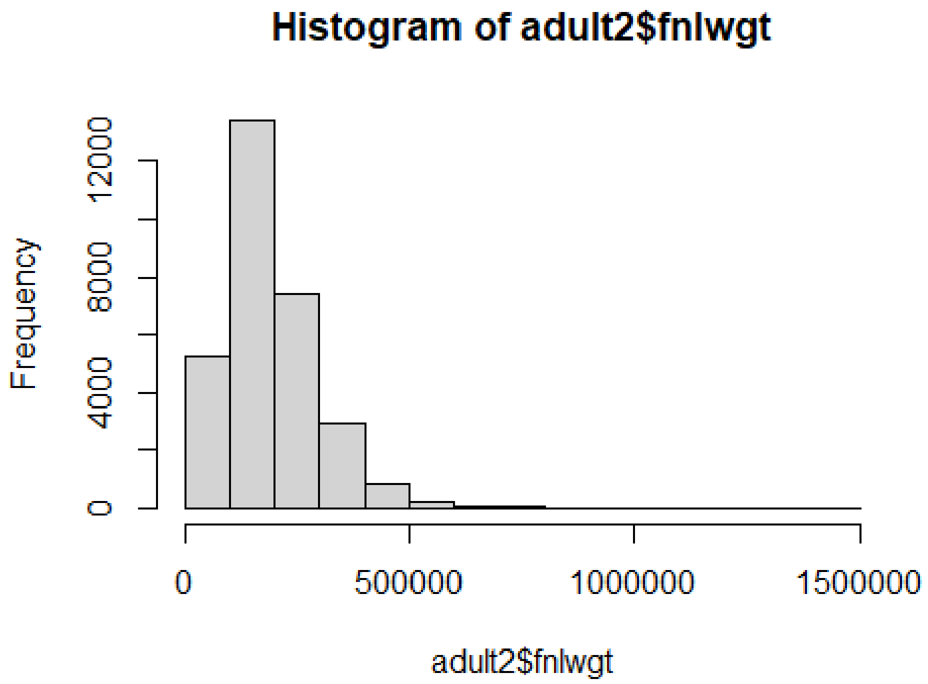
نمودار زیر چوله به راست است.



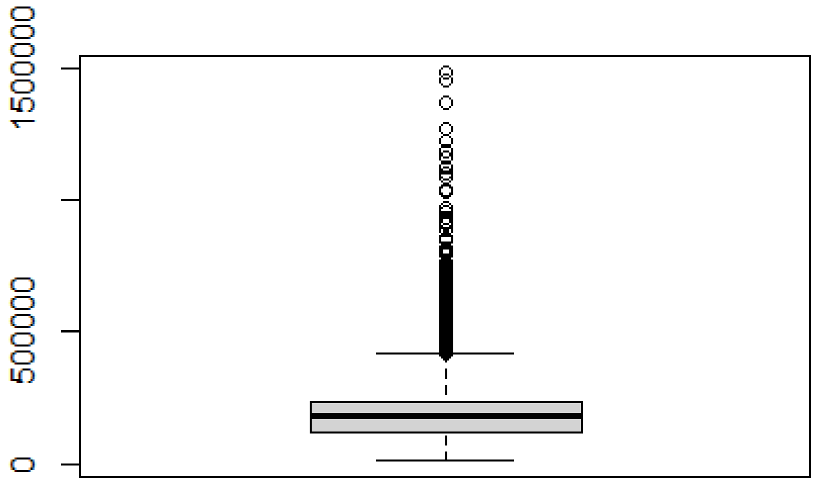
از متغیر پیوسته زیر نتیجه ای دریافت نکردم

#fnlwgt

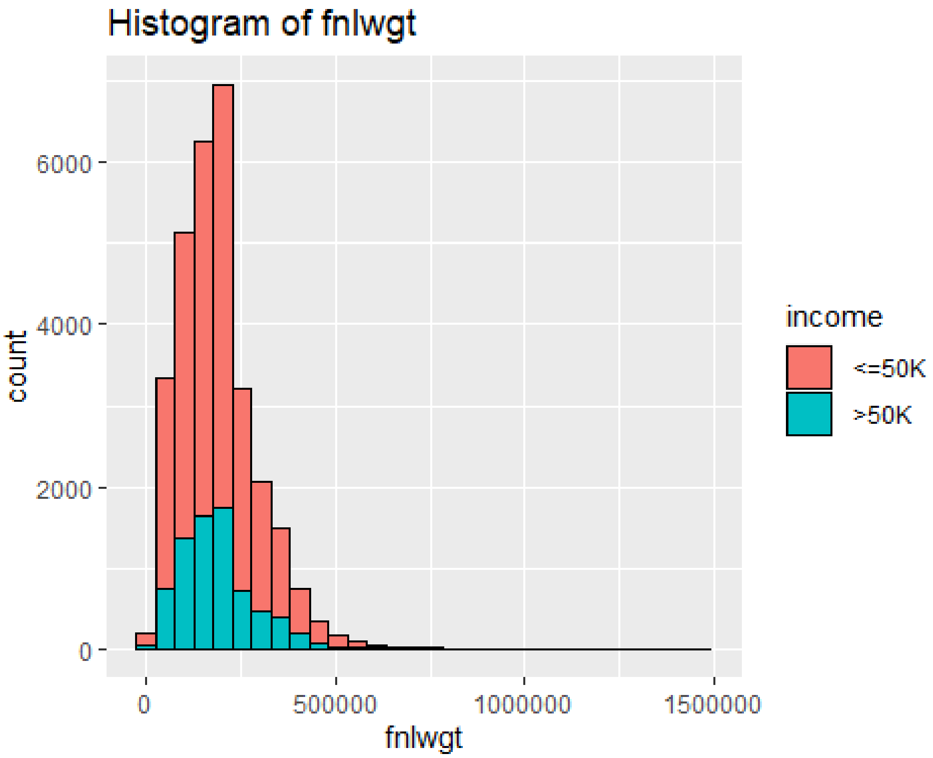
```
#table(adult2$fnlwgt)
hist(adult2$fnlwgt)
```



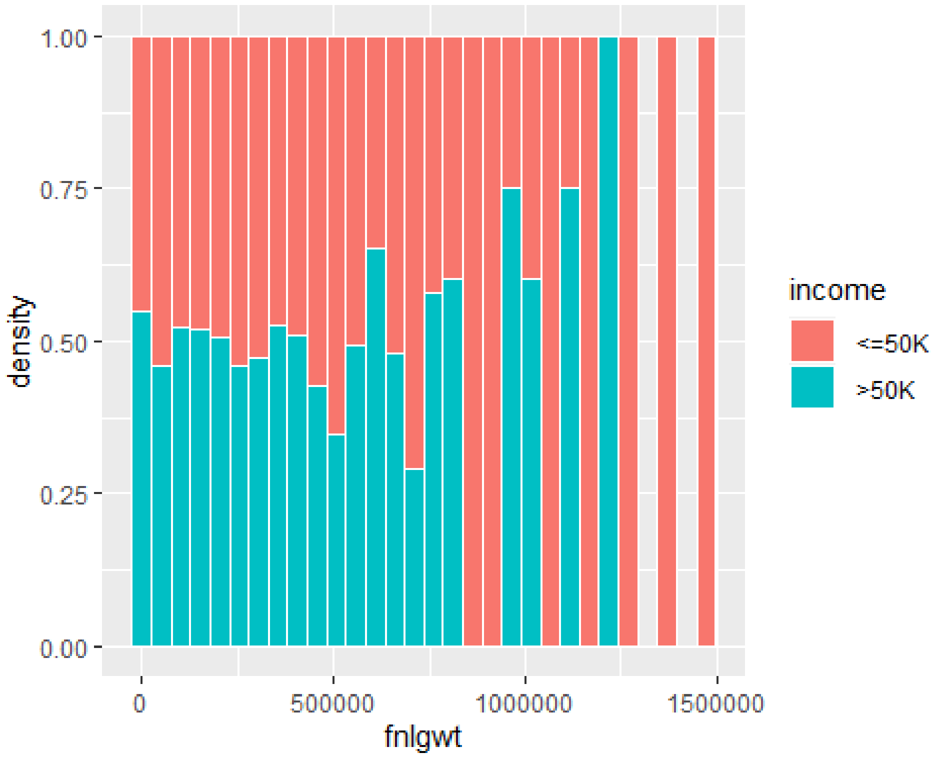
```
boxplot(adult2$fnlwgt)
```

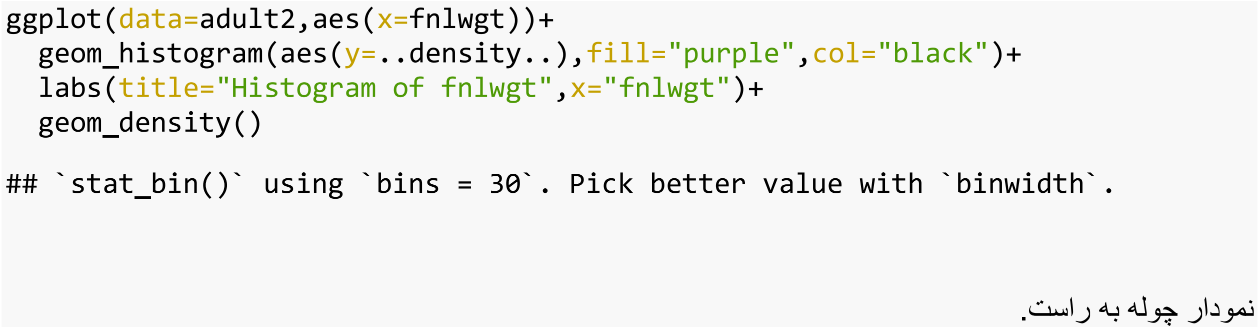


```
ggplot(data=adult2,aes(x=fnlwgt,fill=income)) +
  geom_histogram(col="black")+
  labs(title="Histogram of fnlwgt",x="fnlwgt")

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



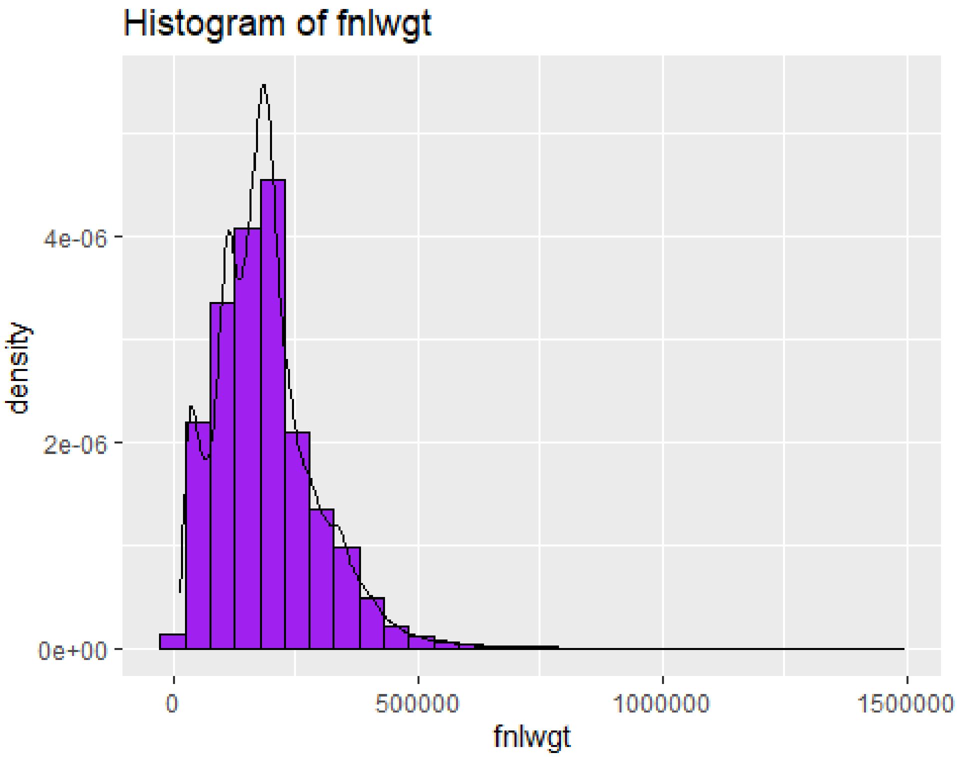
```
summary(adult2$fnlwgt[adult2$income==" >50K"]);summary(adult2$fnlwgt[adult2$i
ncome==" >50K"])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14878  119101  176185  188150  231066 1226583

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14878  119101  176185  188150  231066 1226583

ggplot(data=adult2,aes(x=fnlwgt,fill=income))+
  geom_histogram(aes(y=..density..),col="white",position="fill")+
  labs("Histogram of Fnlgwt",x="fnlgwt")

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

## Warning: Removed 4 rows containing missing values (geom_bar).
```





```
ggplot(data=adult2,aes(x=fnlwgt))+
  geom_histogram(aes(y=..density..),fill="purple",col="black")+
  labs(title="Histogram of fnlwgt",x="fnlwgt")+
  geom_density()

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

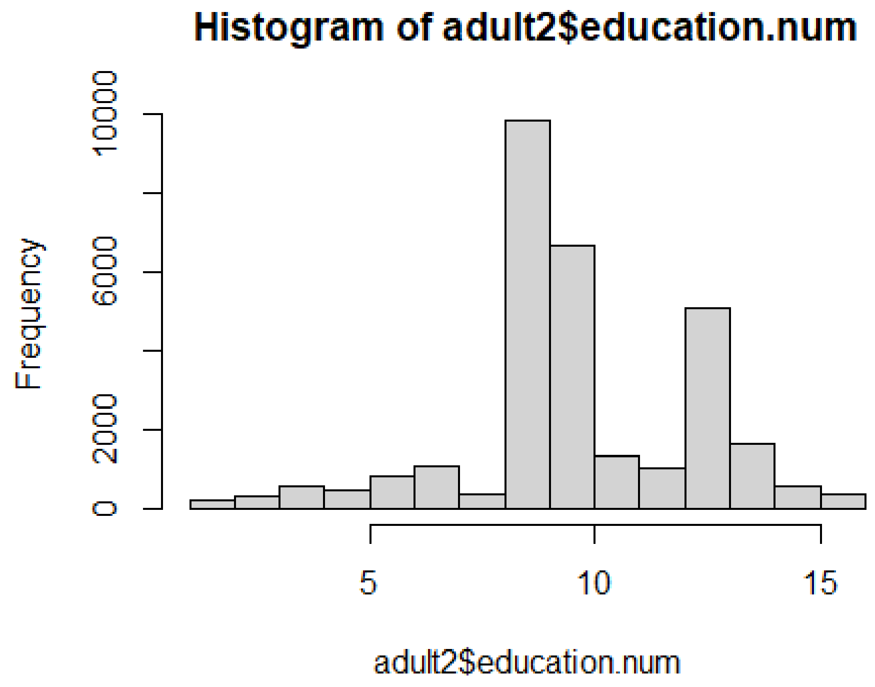
نمودار چوله به راست.



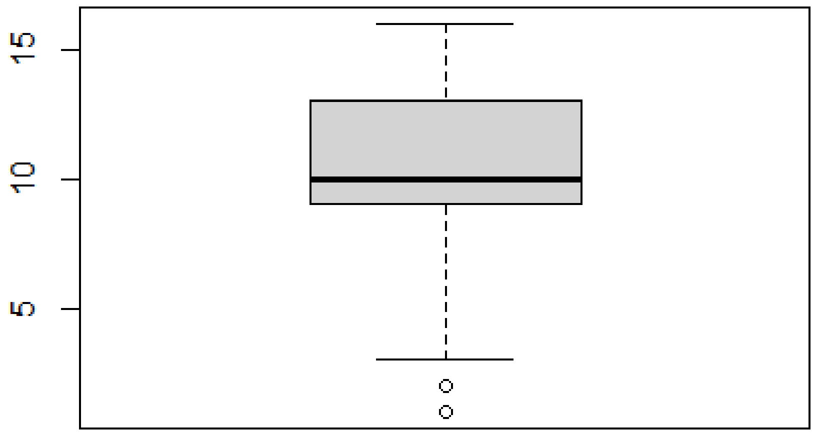
کسانی که اجوکیشن نامبر بین ۸تا ۱۰ دارند فراوانی بیشتری دارند.

#Education.num

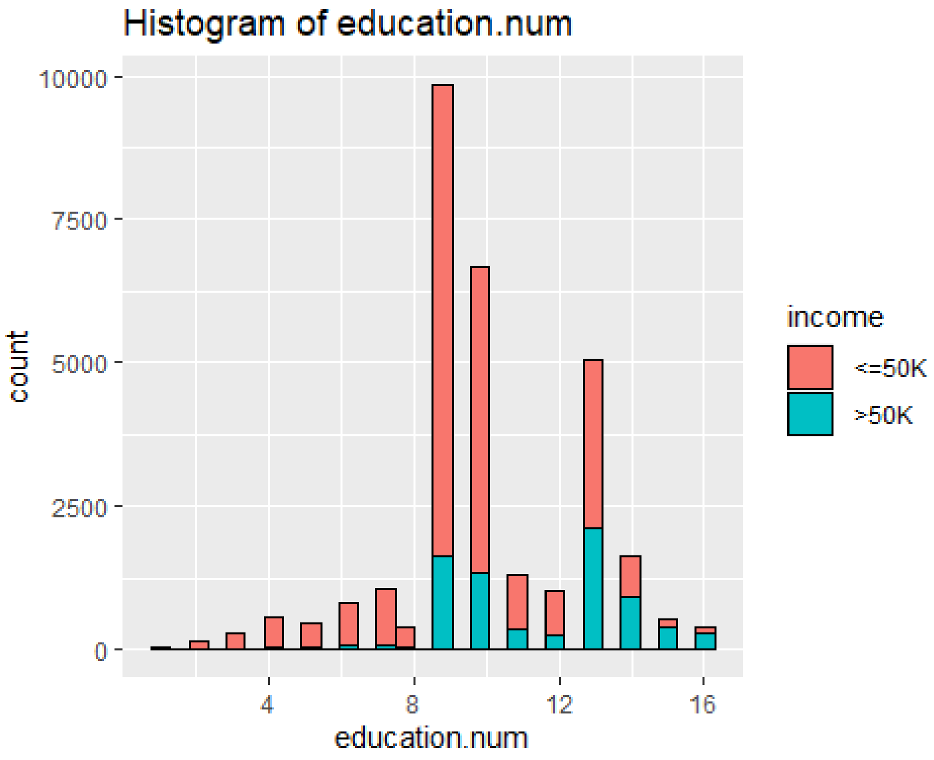
```
#table(adult2$education.num)
hist(adult2$education.num)
```



```
boxplot(adult2$education.num)
```

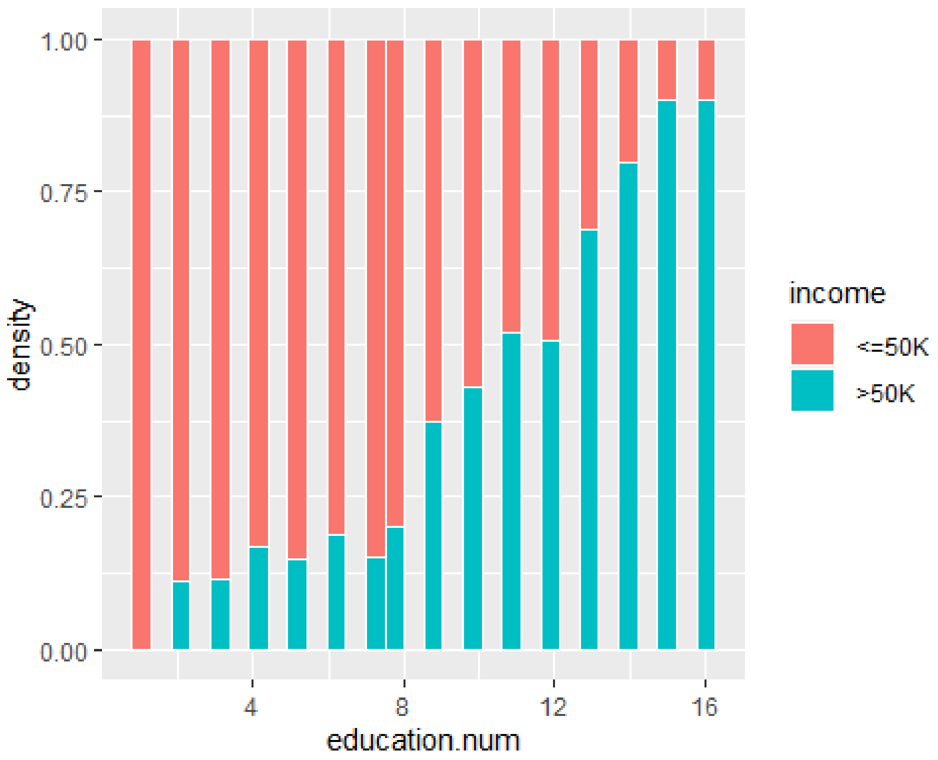


```
ggplot(data=adult2,aes(x=education.num,fill=income)) +
  geom_histogram(col="black")+
  labs(title="Histogram of education.num",x="education.num")

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



```
summary(adult2$education.num[adult2$income==" >50K"]);summary(adult2$educatio
n.num[adult2$income==" >50K"])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   10.00   12.00   11.61   13.00   16.00

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00   10.00   12.00   11.61   13.00   16.00

ggplot(data=adult2,aes(x=education.num,fill=income))+
  geom_histogram(aes(y=..density..),col="white",position="fill")+
  labs("Histogram of Education.num",x="education.num")

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

## Warning: Removed 28 rows containing missing values (geom_bar).
```

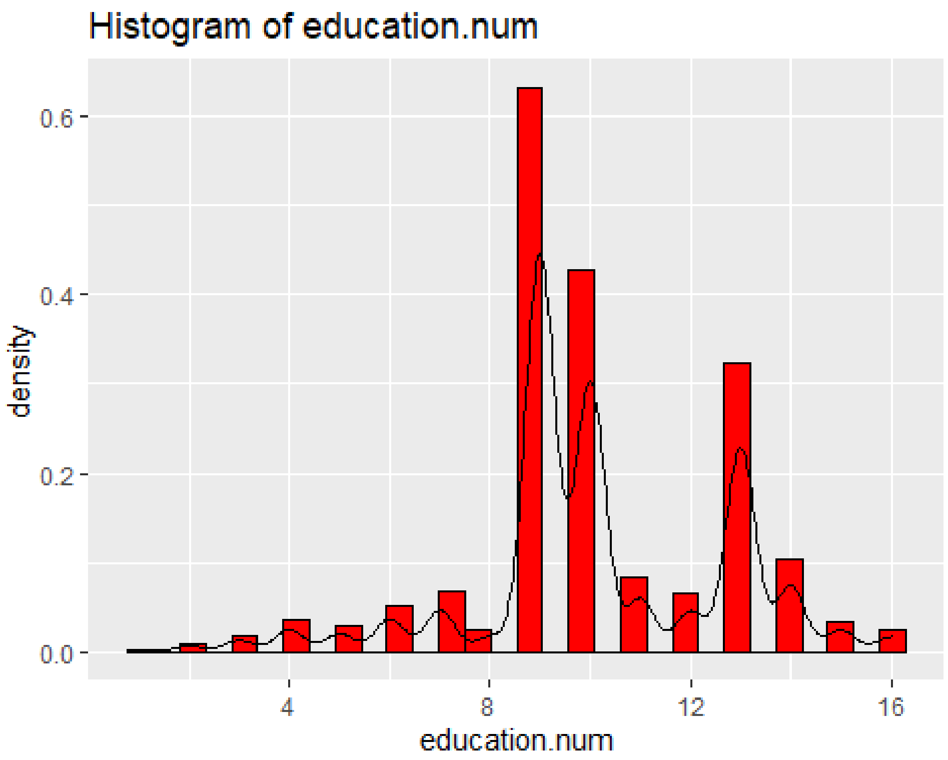
در برخی موارد کسانی که اجوکیشن نامبر زوج و در همسایگی آن داشتند درامد بیشتر و برخی مواقع کسانی که فرد و در همسایگی آن بودند لذا نتیجه خاصی نمیتوان گرفت.



```
ggplot(data=adult2,aes(x=education.num))+
  geom_histogram(aes(y=..density..),fill="red",col="black")+
  labs(title="Histogram of education.num",x="education.num")+
  geom_density()
```

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

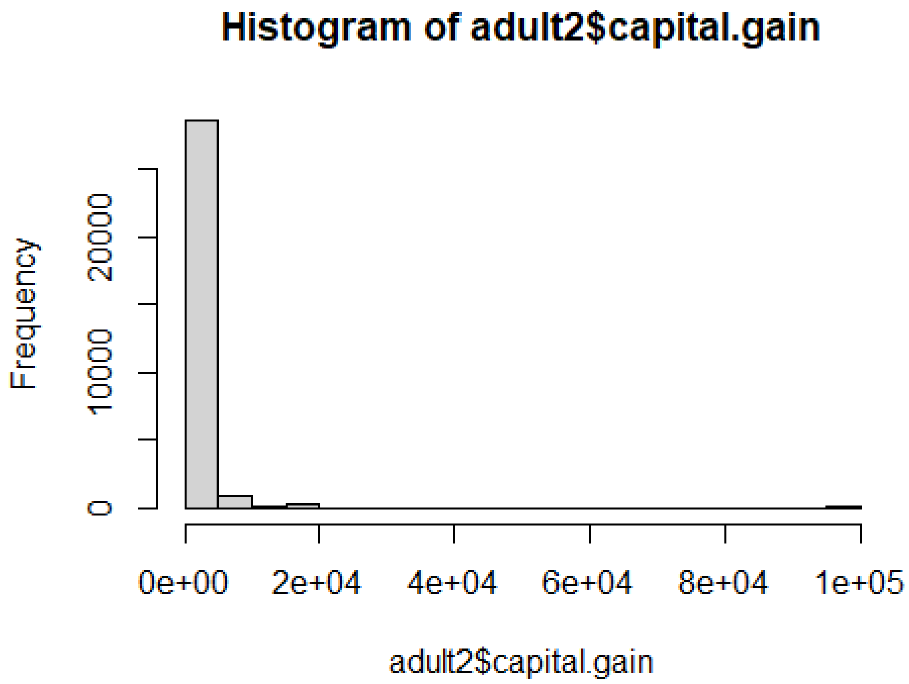
نمودار تقریبا چوله به چپ است.



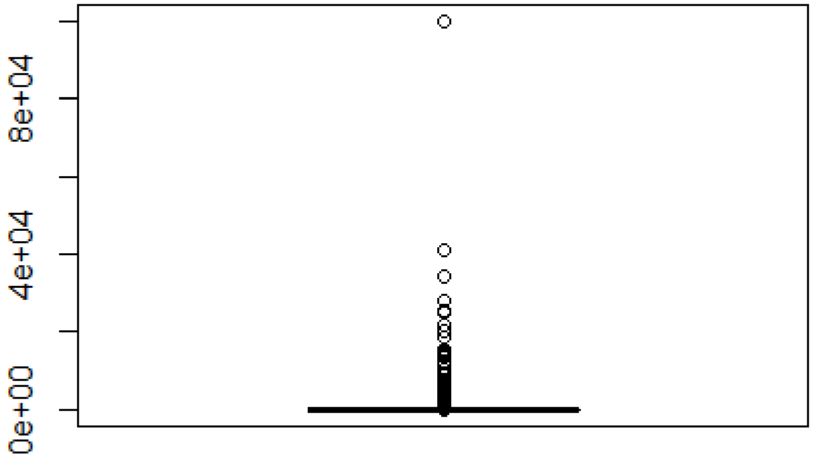
باز نتیجه خاصی نگرفتم. فقط میتوان گفت کمترین مقدار بیشترین فراوانی را به خود اختصاص داده است.

#capital.gain

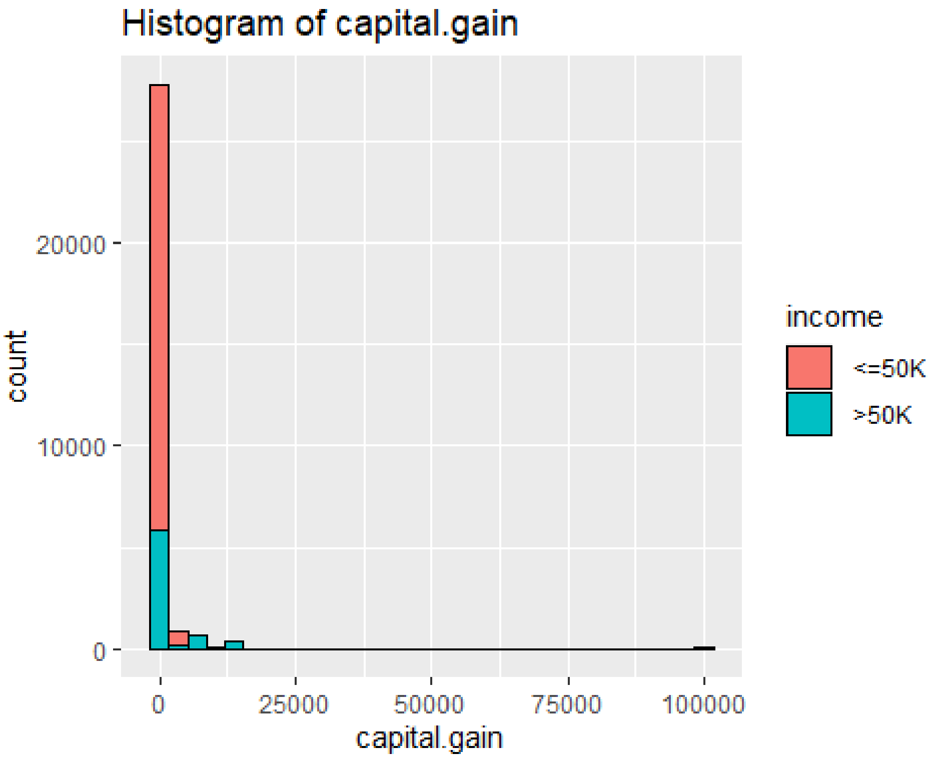
```
#table(adult2$capital.gain)
hist(adult2$capital.gain)
```



```
boxplot(adult2$capital.gain)
```

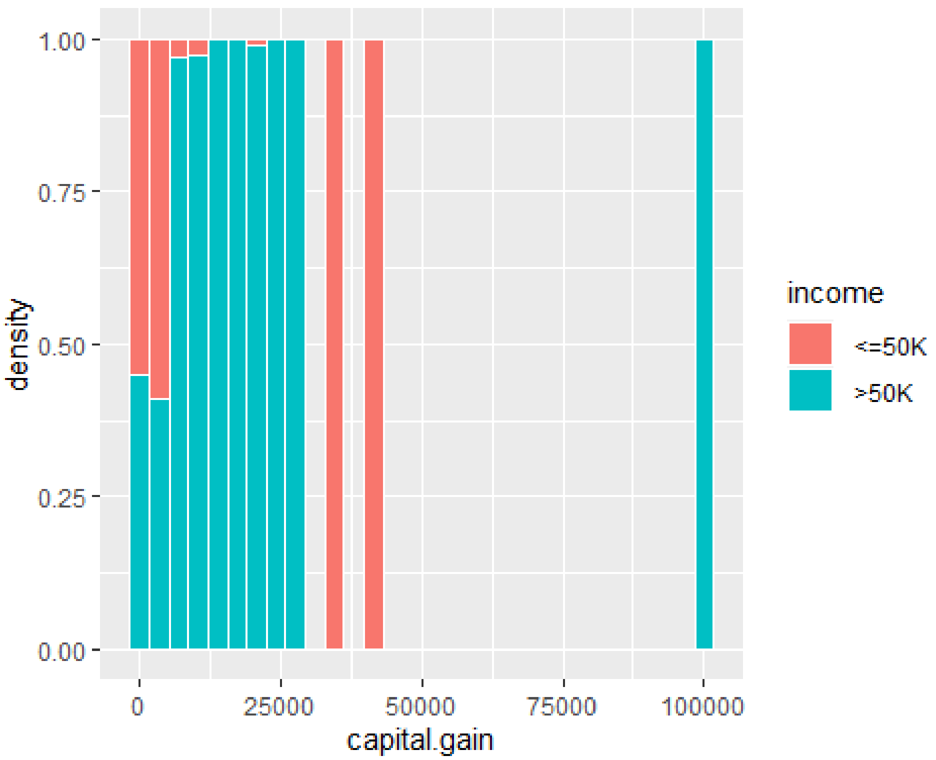


```
ggplot(data=adult2,aes(x=capital.gain,fill=income)) +
  geom_histogram(col="black")+
  labs(title="Histogram of capital.gain",x="capital.gain")

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



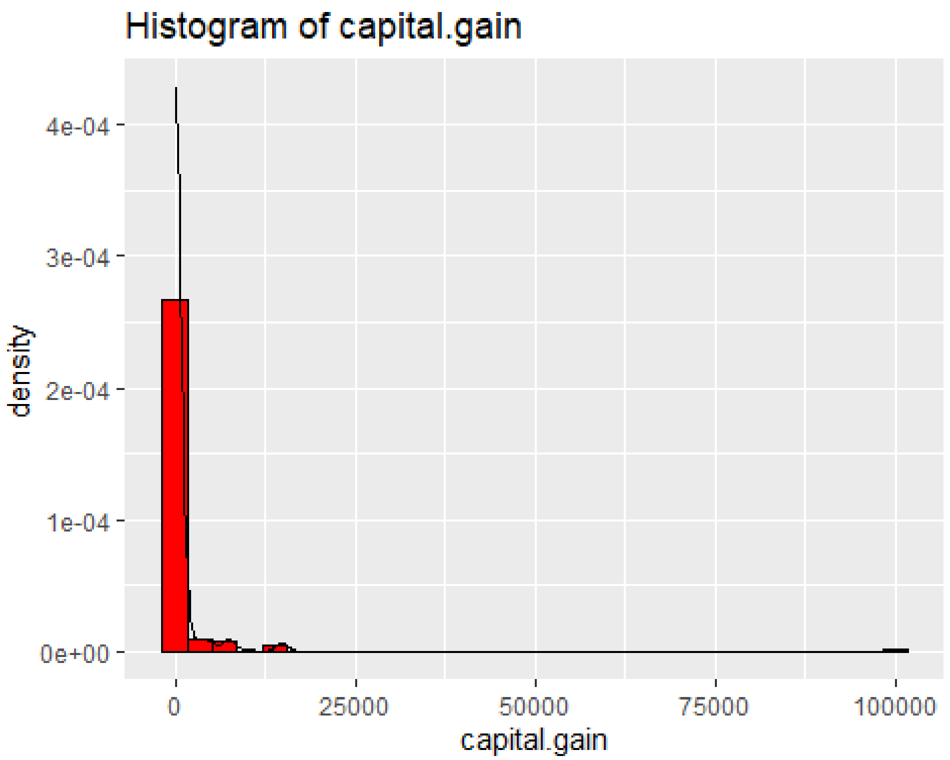
```
summary(adult2$capital.gain[adult2$income==" >50K"]);summary(adult2$capital.g
ain[adult2$income==" >50K"])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    3938       0   99999

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    3938       0   99999

ggplot(data=adult2,aes(x=capital.gain,fill=income))+
  geom_histogram(aes(y=..density..),col="white",position="fill")+
  labs("Histogram of Capital.gain",x="capital.gain")

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

## Warning: Removed 36 rows containing missing values (geom_bar).
```

مقادیر کمتر درآمدهای های بالای۵۰ شان نسبت بالاتری دارد.



```
ggplot(data=adult2,aes(x=capital.gain))+
  geom_histogram(aes(y=..density..),fill="red",col="black")+
  labs(title="Histogram of capital.gain",x="capital.gain")+
  geom_density()

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

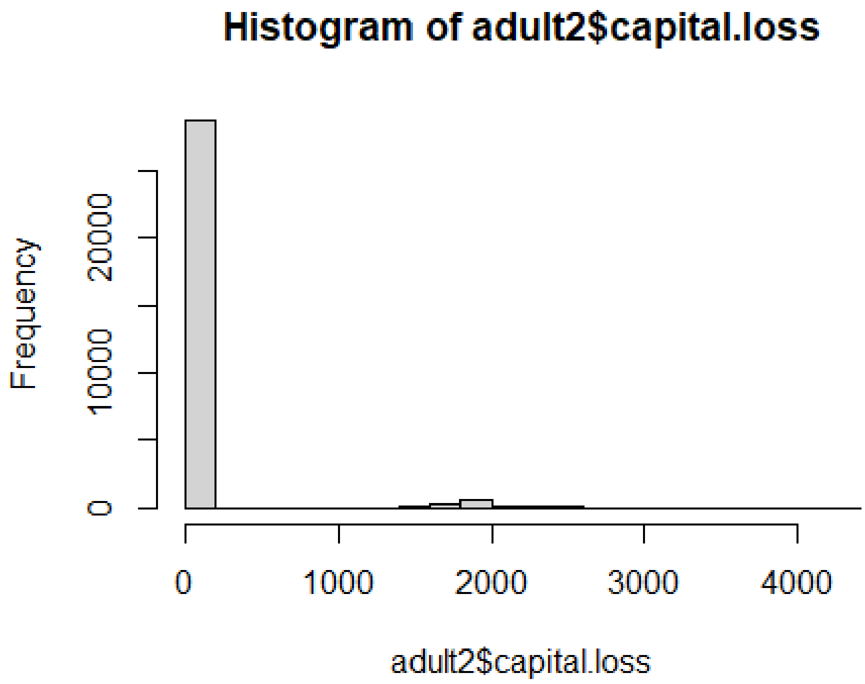
نمودار چوله به راست است.



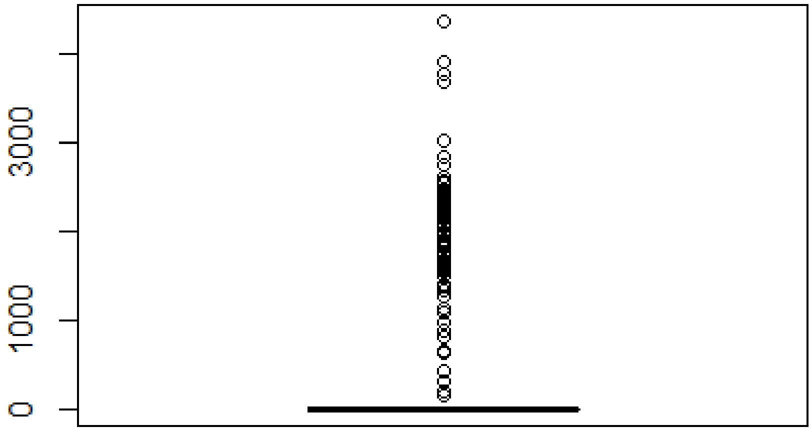
#capital.loss

باز هم افراد با دارای مقادیر کمتر فراوانی بیشتر دارند.

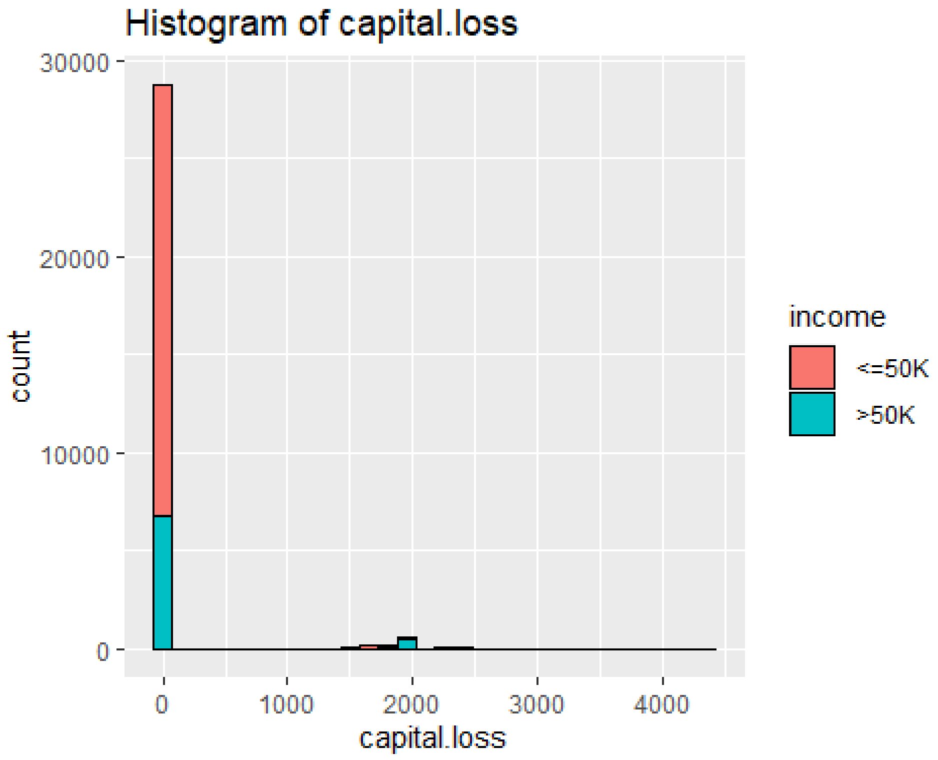
```
#table(adult2$capital.loss)
hist(adult2$capital.loss)
```



```
boxplot(adult2$capital.loss)
```

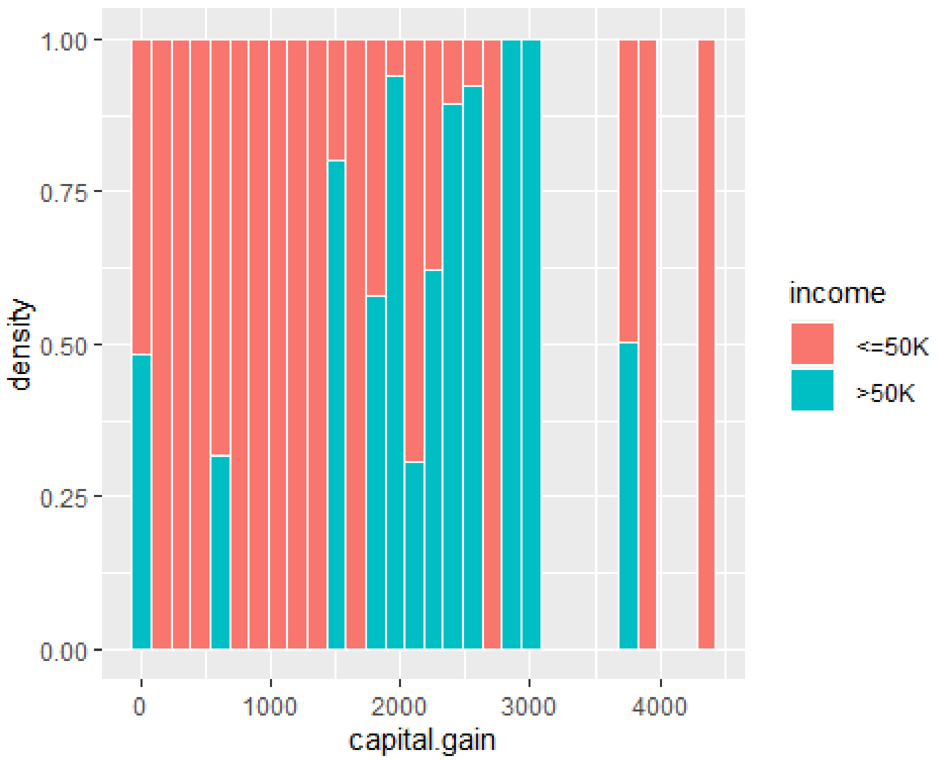


```
ggplot(data=adult2,aes(x=capital.loss,fill=income)) +
  geom_histogram(col="black")+
  labs(title="Histogram of capital.loss",x="capital.loss")

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



```
summary(adult2$capital.loss[adult2$income==" >50K"]);summary(adult2$capital.l
oss[adult2$income==" >50K"])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   193.8     0.0  3683.0

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   193.8     0.0  3683.0

ggplot(data=adult2,aes(x=capital.loss,fill=income))+
  geom_histogram(aes(y=..density..),col="white",position="fill")+
  labs("Histogram of Capital.gain",x="capital.gain")

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

## Warning: Removed 12 rows containing missing values (geom_bar).
```

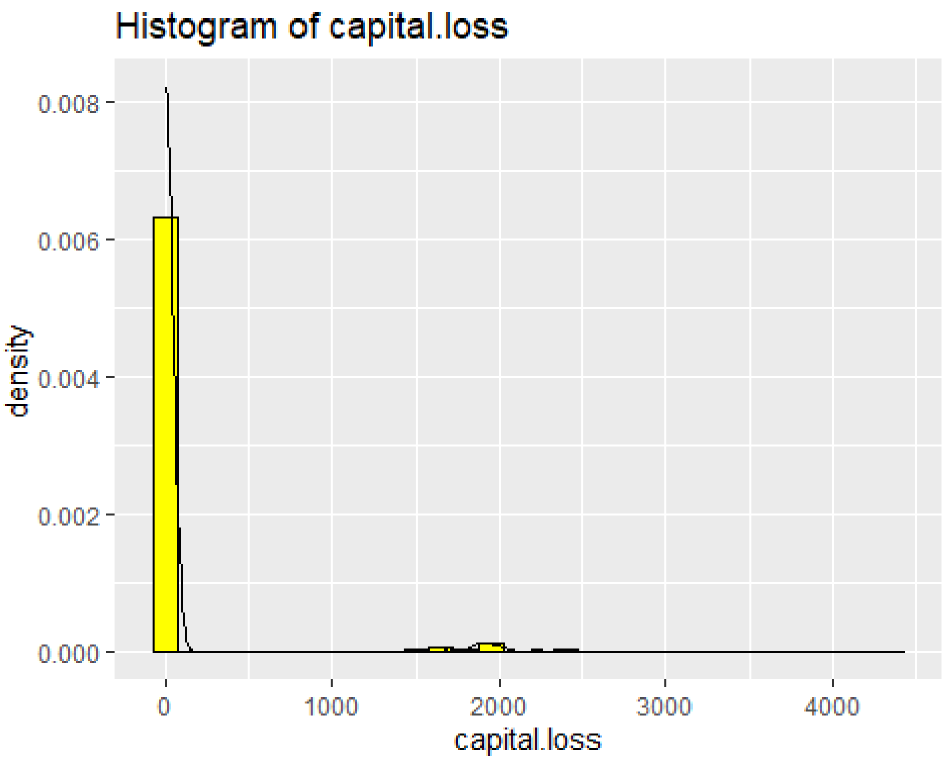
نمیتوان نتیجه خاصی از این دو نمودار برداشت کرد.



```
ggplot(data=adult2,aes(x=capital.loss))+
  geom_histogram(aes(y=..density..),fill="yellow",col="black")+
  labs(title="Histogram of capital.loss",x="capital.loss")+
  geom_density()
```

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

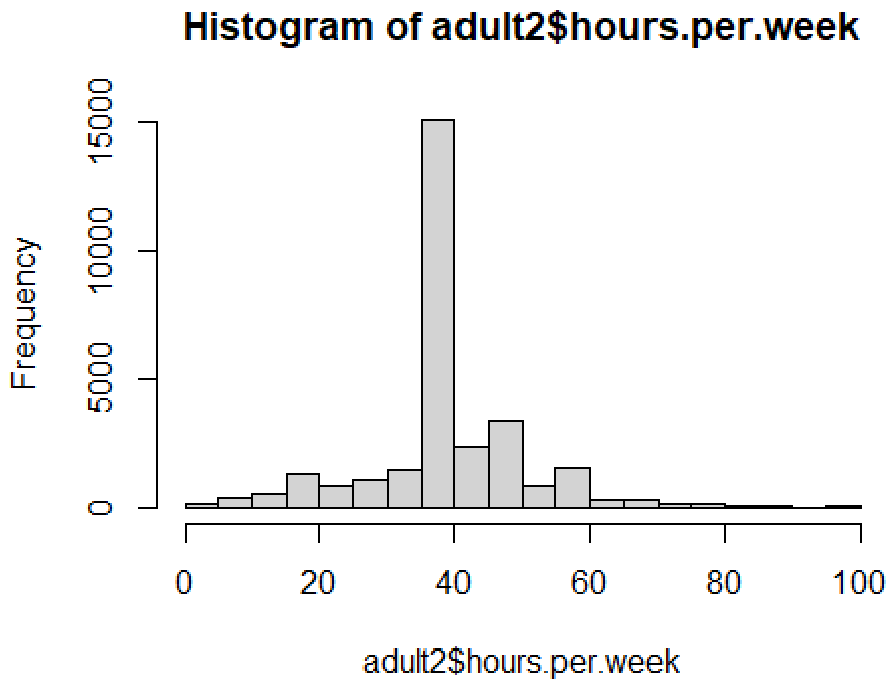
نمودار چوله به راست است.



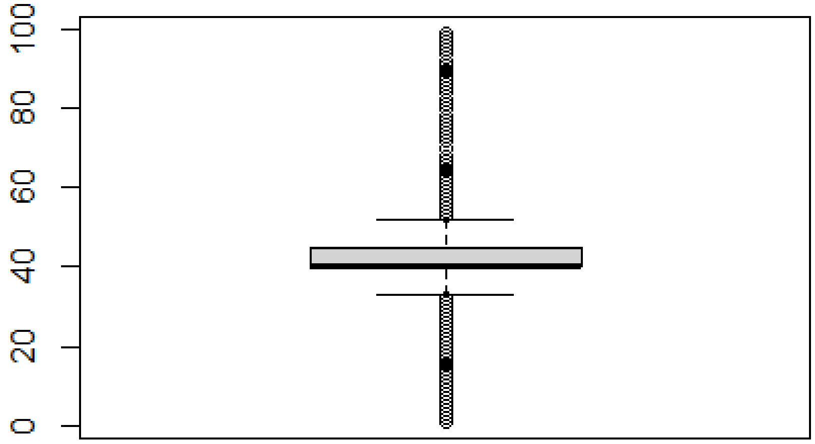
کسانیکه بین ۳۵ تا۴۰ ساعت در هفته کار میکنند فراوانی بیشتر دارند.

#hours.per.week

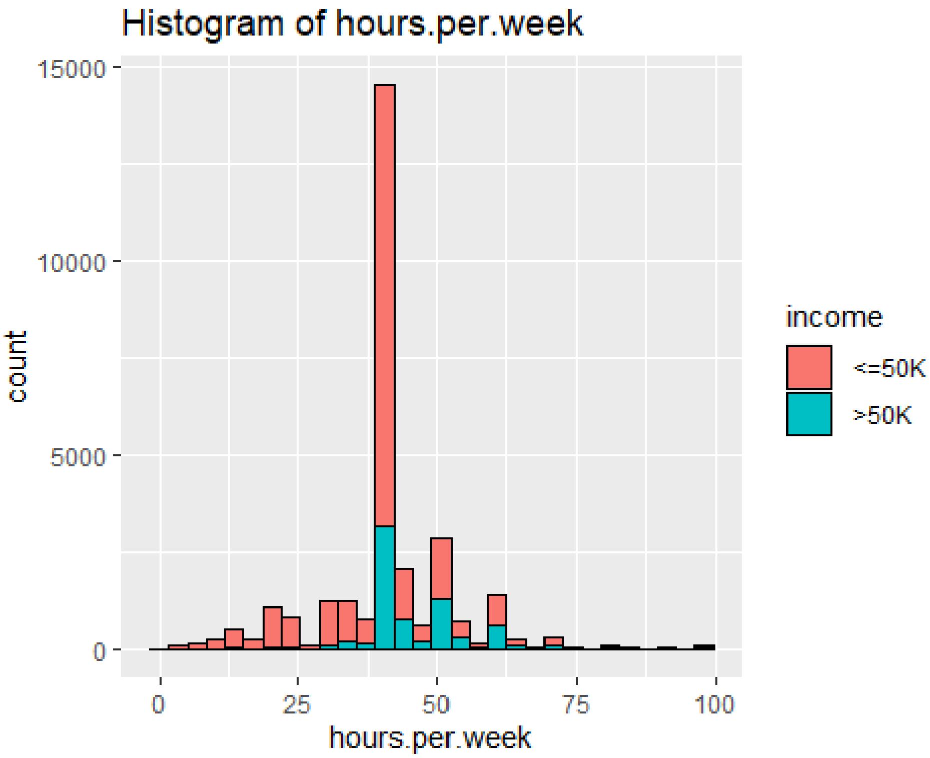
```
#table(adult2$hours.per.week)
hist(adult2$hours.per.week)
```



```
boxplot(adult2$hours.per.week)
```

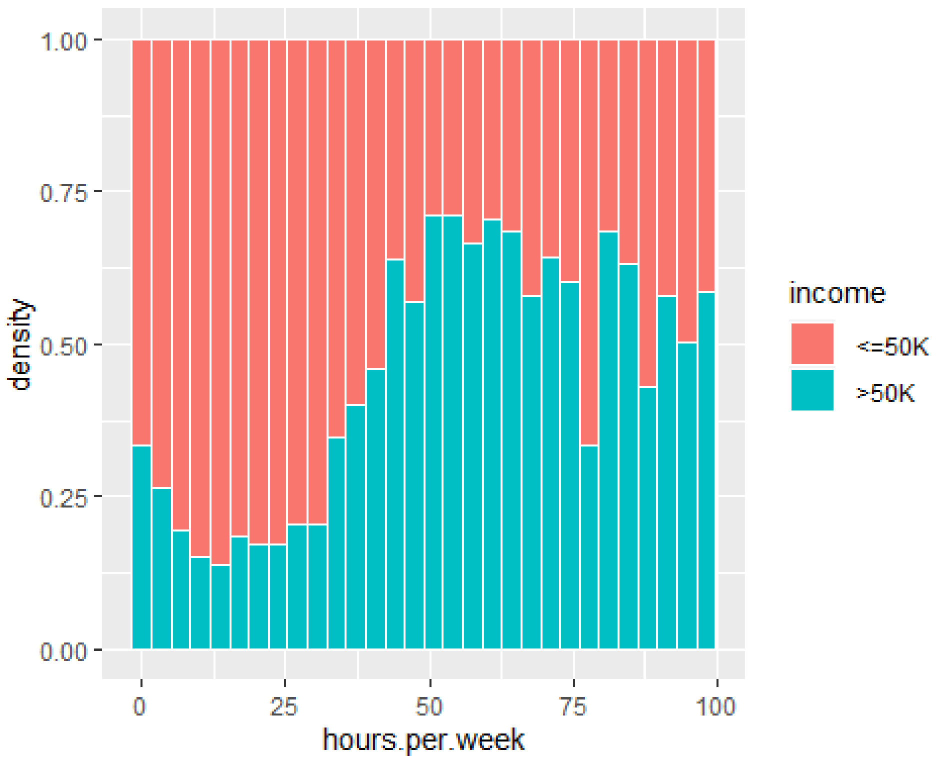


```
ggplot(data=adult2,aes(x=hours.per.week,fill=income)) +
  geom_histogram(col="black")+
  labs(title="Histogram of hours.per.week",x="hours.per.week")

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



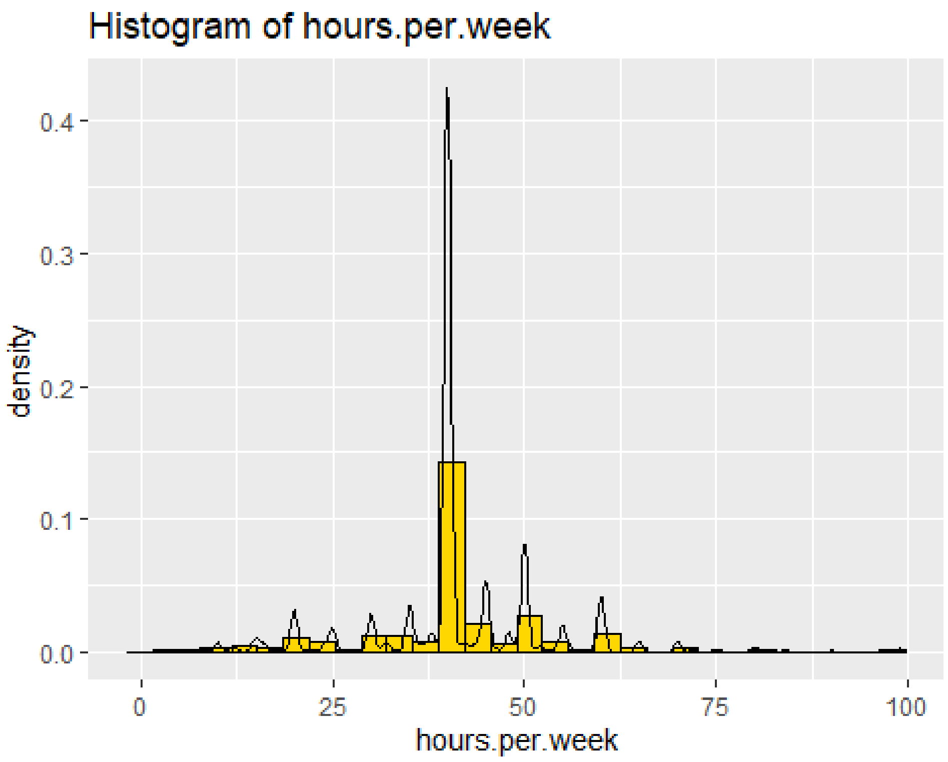
```
summary(adult2$hours.per.week[adult2$income==" >50K"]);summary(adult2$hours.p
er.week[adult2$income==" >50K"])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   40.00   40.00   45.71   50.00   99.00

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   40.00   40.00   45.71   50.00   99.00

ggplot(data=adult2,aes(x=hours.per.week,fill=income))+
  geom_histogram(aes(y=..density..),col="white",position="fill")+
  labs("Histogram of Hours.per.week",x="hours.per.week")

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

ساعت های کاری متوسط و بالاتر درهفته درآمد بیشتری دارند که نقطه قوت است.



```
ggplot(data=adult2,aes(x=hours.per.week))+
  geom_histogram(aes(y=..density..),fill="gold",col="black")+
  labs(title="Histogram of hours.per.week",x="hours.per.week")+
  geom_density()

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

نمودار تقریبا نرمال داریم.

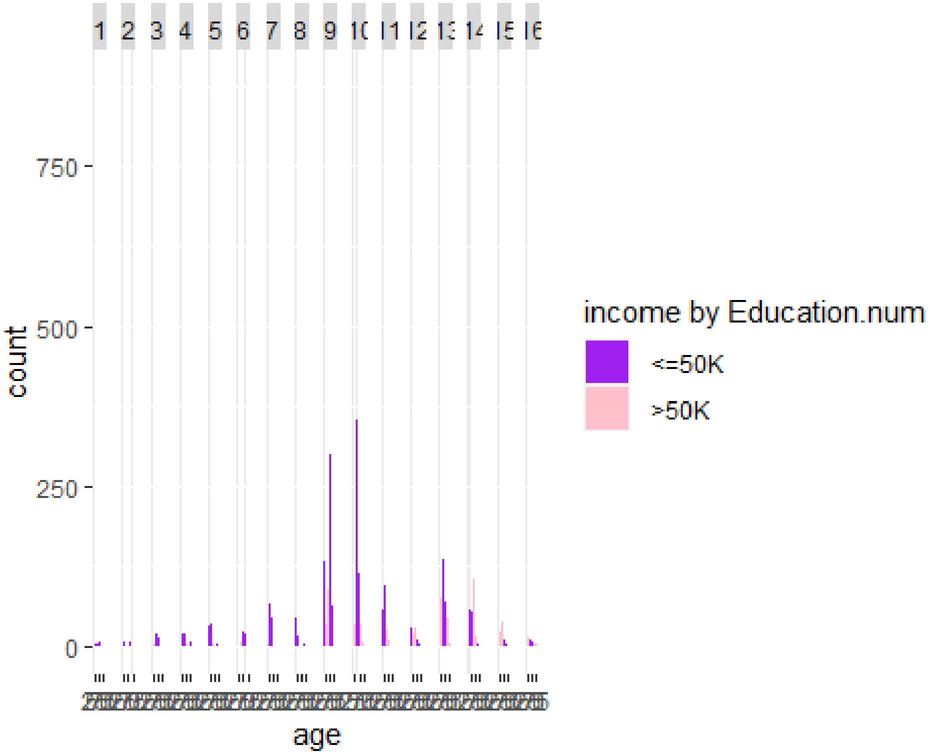


#more than 2 variables

در زیر نمودارهای چوله به راست زیادی به چشم میخورد.در اجوکیشن نامبرهای گوناگون هرچه سن بیشتر شود درامد
کمتر خواهد بود.

```
library(ggplot2)
#xtabs(~income+age+fnlwgt,data=adult2)
#xtabs(~income+age+education.num,data=adult2)
ggplot(adult2,aes(x=age,group=income,fill=income))+
  geom_histogram(position=position_dodge())+
  scale_fill_manual(values=c("purple","pink"),name="income by Education.num")
+facet_grid(~education.num)
```

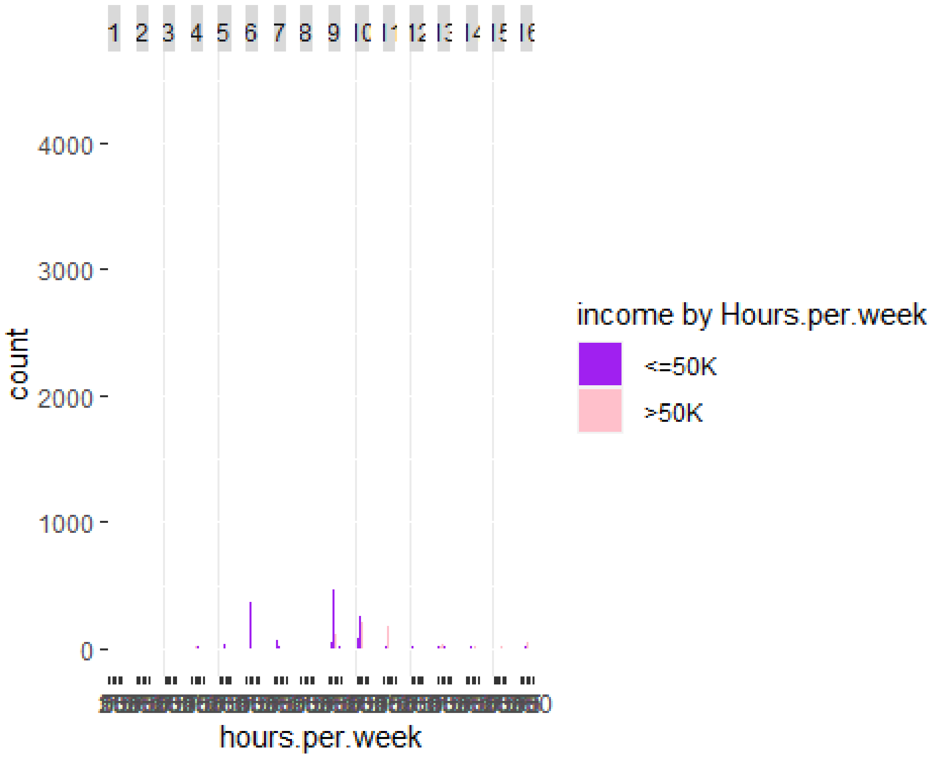
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



در نهمین اجوکیشن تعداد به میزان قابل توجهی بیشتر است.(بخصوص ساعت کاری متوسط)اما نسبت حقوق های بالای ۵
۰ پایین می باشد.در اجوکیشن نامبر ۱۳ نسبت افرادی که درامد بالای ۵۰ دارند بیشتر است.

```
#xtabs(~income+hours.per.week+education.num,data=adult2)
ggplot(adult2,aes(x=hours.per.week,group=income,fill=income))+
  geom_histogram(position=position_dodge())+
  scale_fill_manual(values=c("purple","pink"),name="income by Hours.per.week"
)+facet_grid(~education.num)
```

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


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
4000
3000
2000
1000
0
count
hours.per.week
income by Hours.per.week
<=50K
>50K