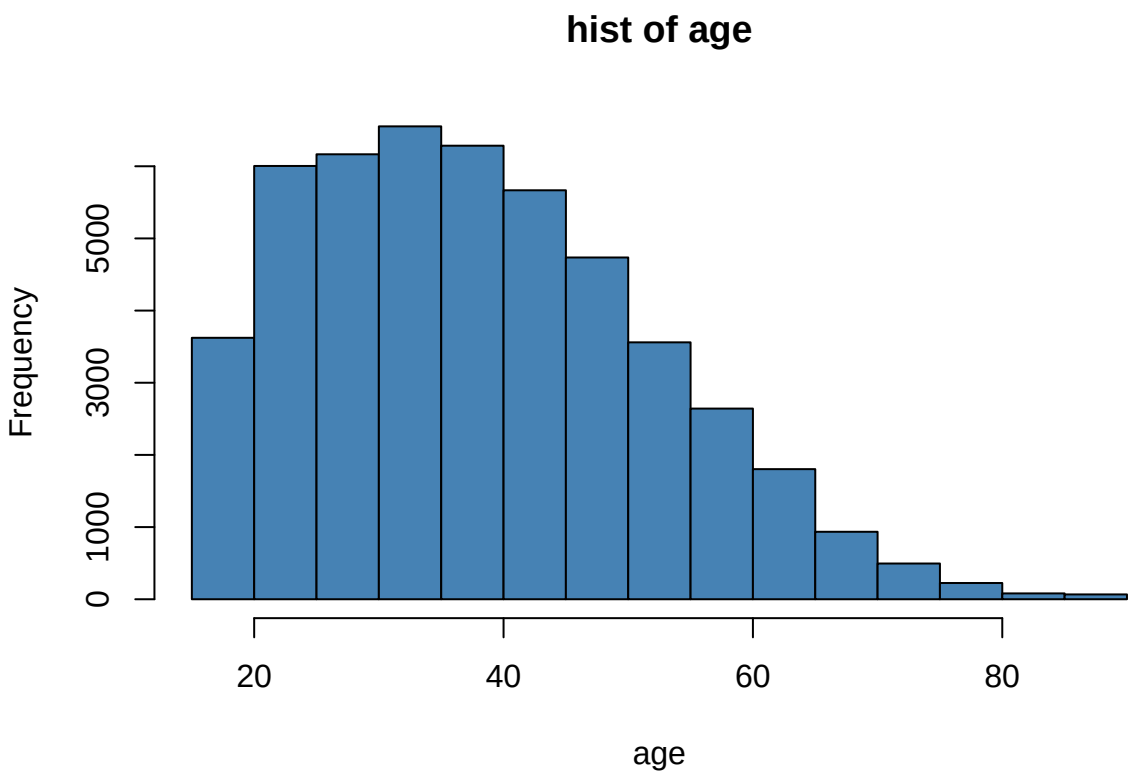
Basic descriptive (prior to preprocessing)

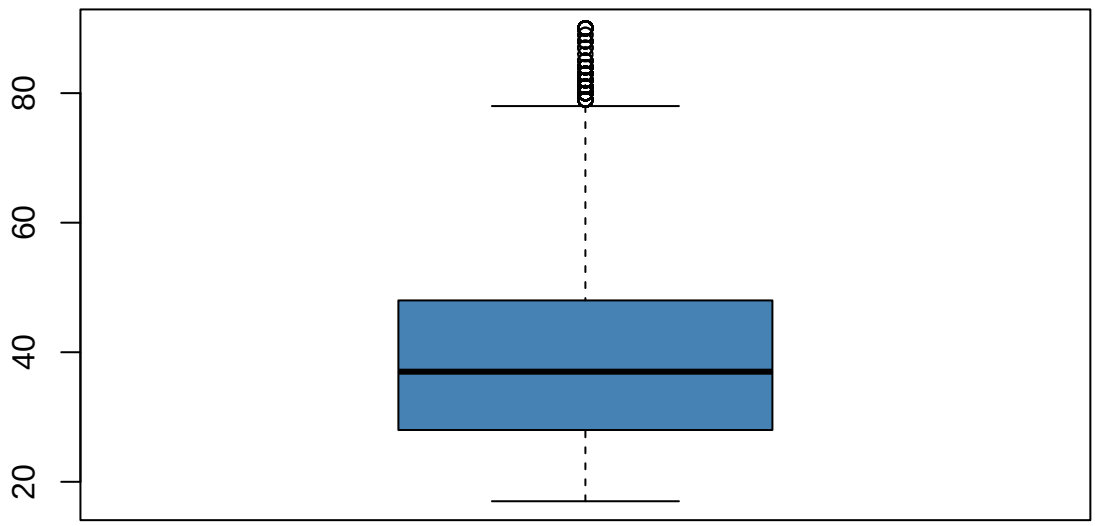
2025-02-20

```
setwd("/home/adpaoj/Documents/ADEI/Project")
dd <- read.csv("~/Documents/ADEI/Project/adult.csv")
library(RColorBrewer)

#Numerical (prior to preprocessing)

class(dd[,1])
```

```
## [1] "integer"
```

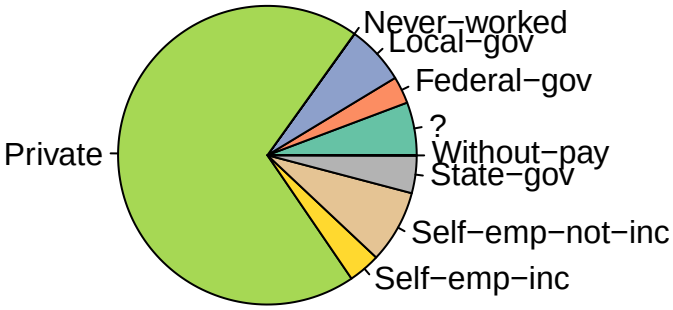
```
for ( i in 1:15){
  if(is.numeric(dd[,i])){
    hist(dd[,i],main=paste("hist of", names(dd)[i]), col = "steelblue", xlab=names(dd)[i])
    boxplot(dd[,i],main=paste("boxplot of", names(dd)[i]), col = "steelblue")
    cat("Summary of", names(dd)[i], ":\n")
    print(summary(dd[,i]))
  } else{
    par(mar = c(8, 4, 4, 2))
    barplot(table(dd[,i]),main=paste("barplot of", names(dd)[i]), col = "tomato", las = 2)
    pie(table(dd[,i]),main=paste("pie of", names(dd)[i]), col = brewer.pal(min(length(table(dd[, i])),
  }
}
```

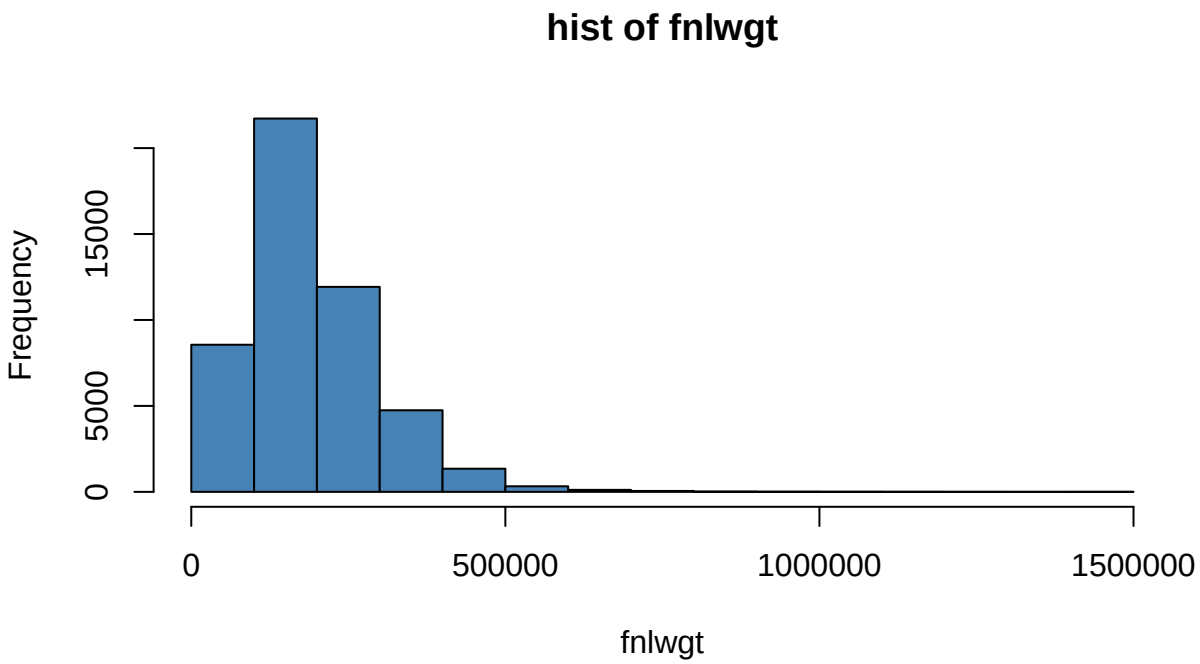

hist of age
Frequency
0
1000
3000
5000
20
40
60
80
age

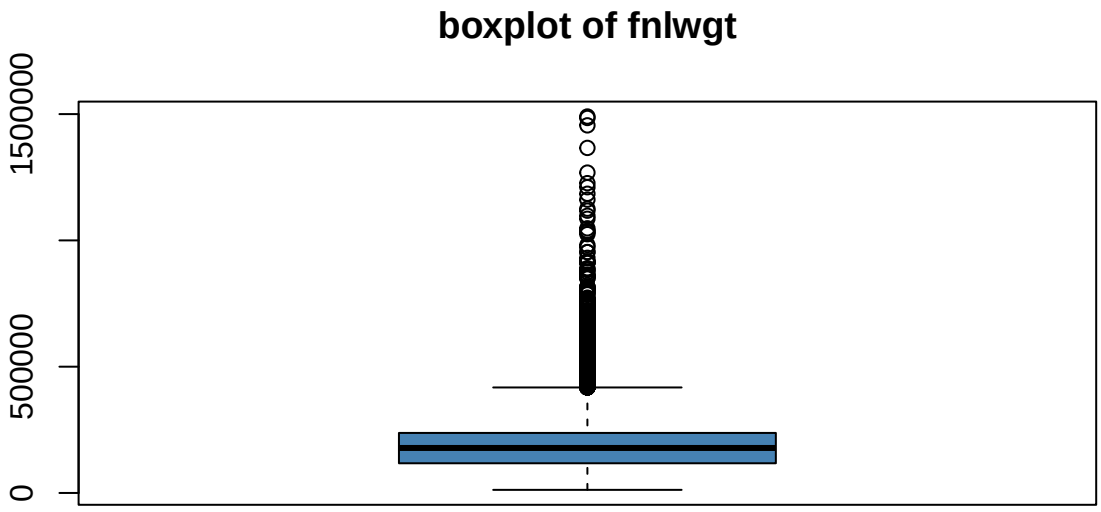



```
## Summary of age :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.64   48.00   90.00
```

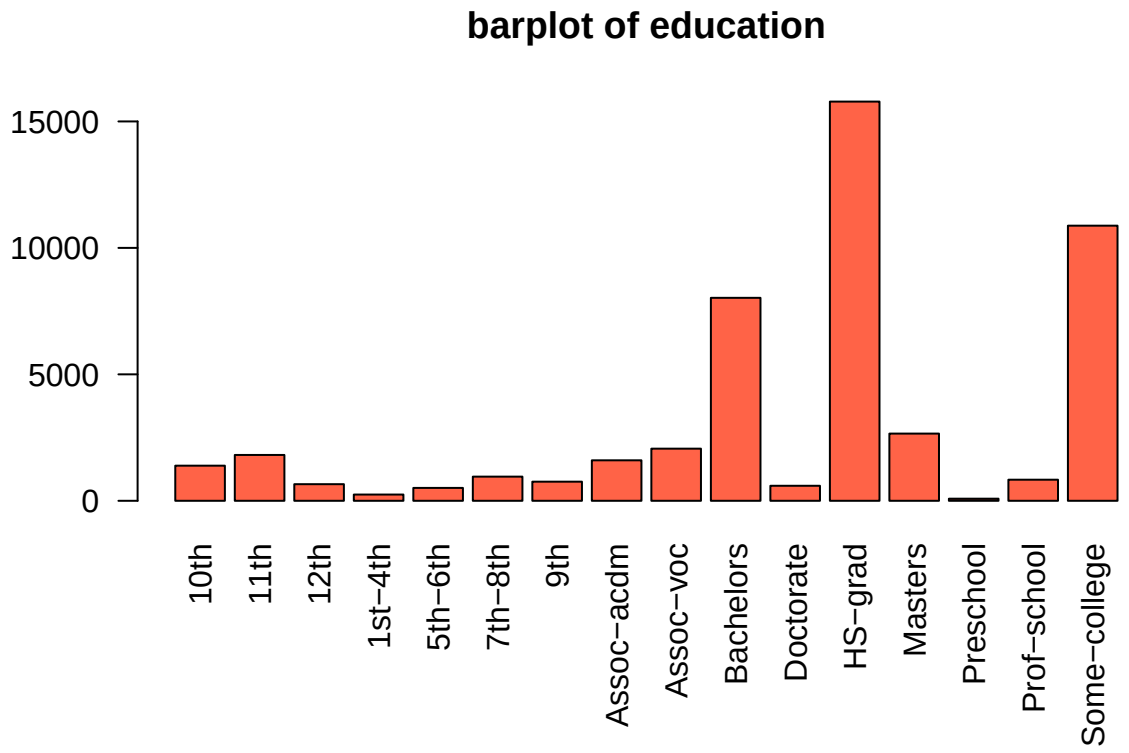

barplot of workclass
30000
25000
20000
15000
10000
5000
0
?
Federal−gov
Local−gov
Never−worked
Private
Self−emp−inc
Self−emp−not−inc
State−gov
Without−pay

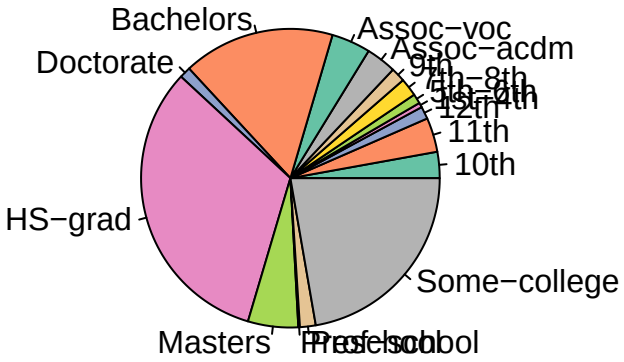

pie of workclass
Never−worked
Local−gov
Federal−gov
?
Without−pay
State−gov
Self−emp−not−inc
Self−emp−inc
Private

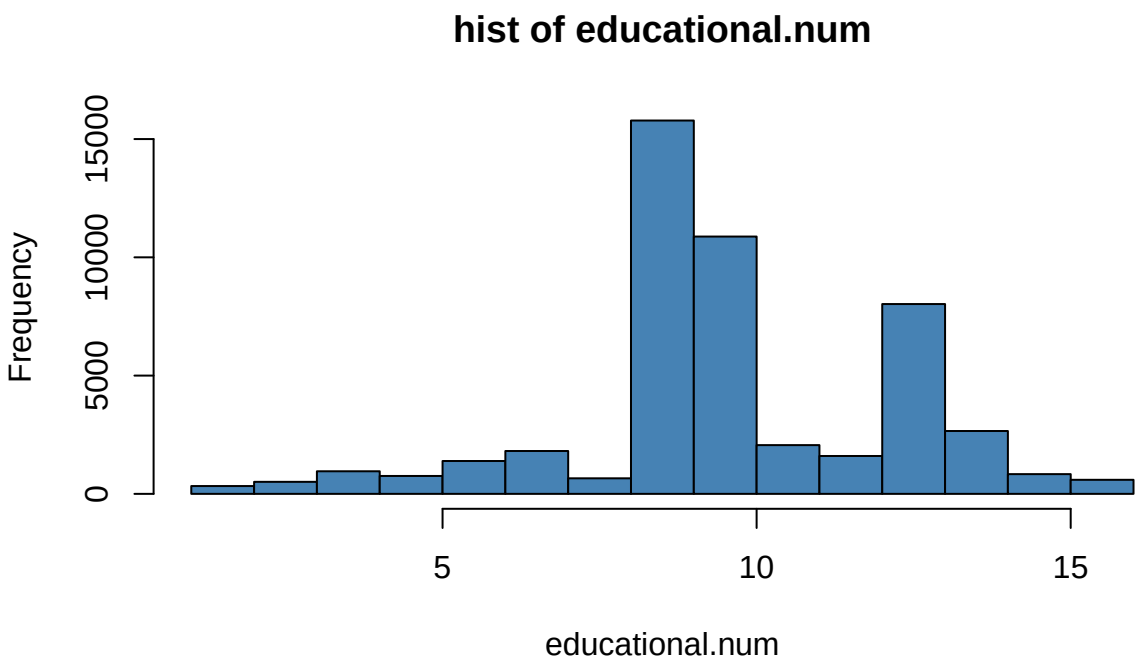

hist of fnlwgt
Frequency
0
5000
15000
0
500000
1000000
1500000
fnlwgt

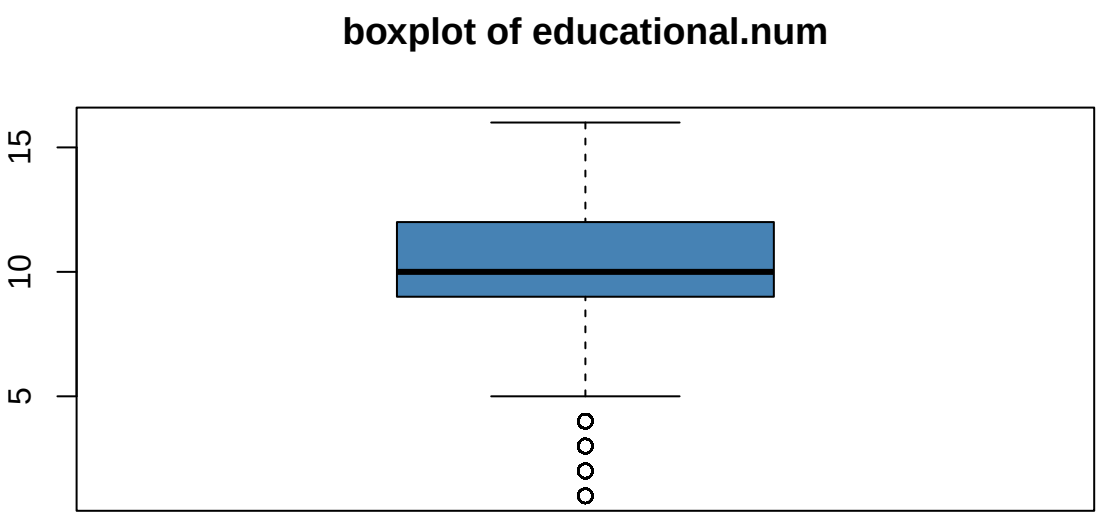



```
## Summary of fnlwgt :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12285  117550  178144  189664  237642 1490400
```

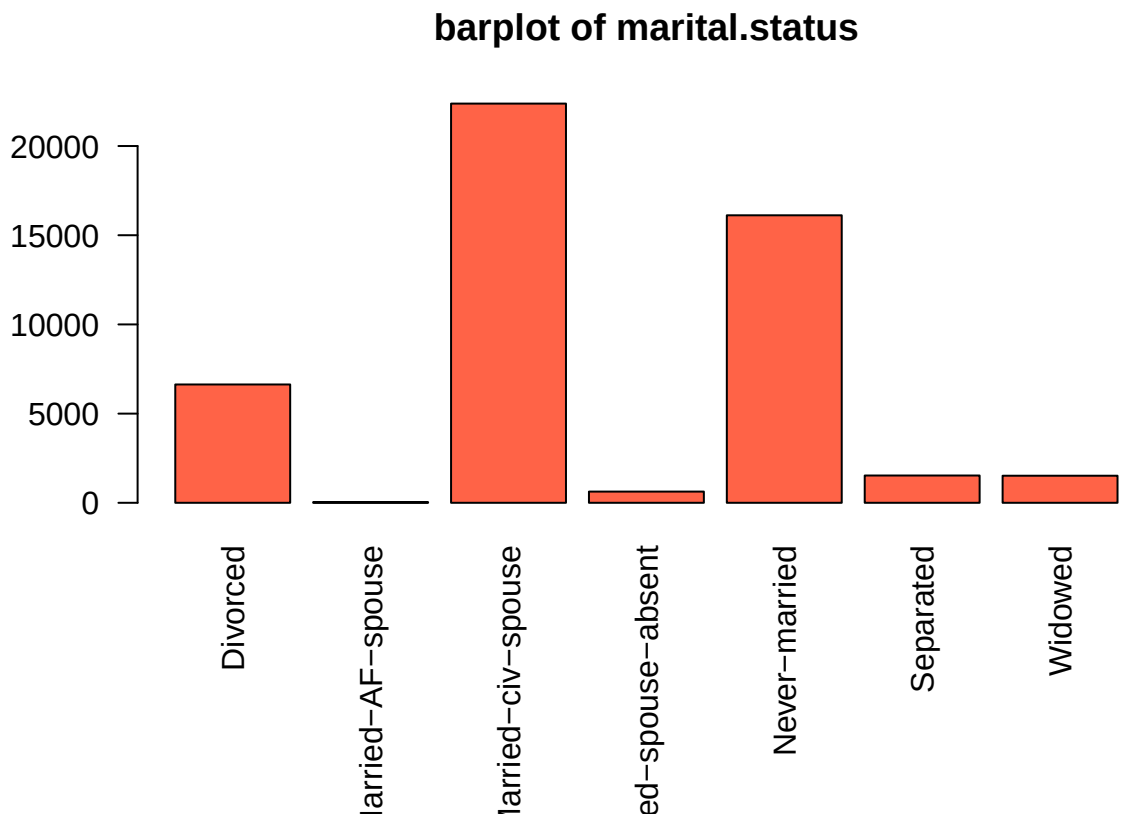

barplot of education
15000
10000
5000
0
10th
11th
12th
1st−4th
5th−6th
7th−8th
9th
Assoc−acdm
Assoc−voc
Bachelors
Doctorate
HS−grad
Masters
Preschool
Prof−school
Some−college

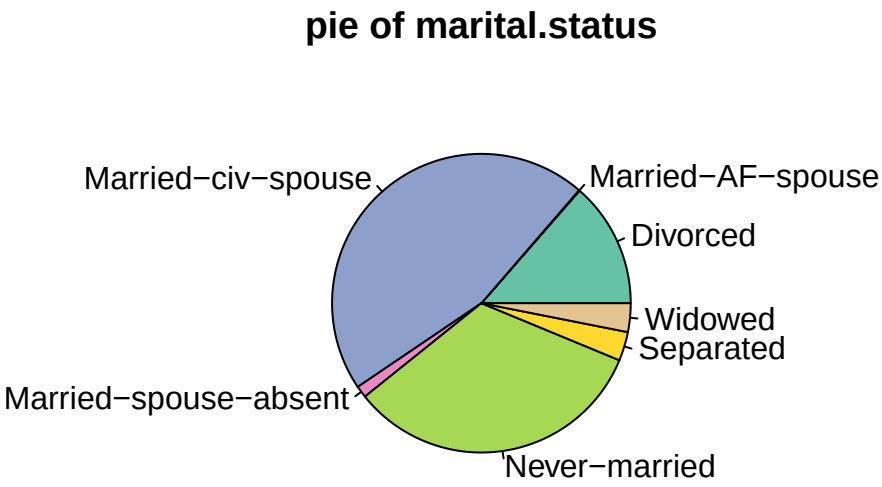

pie of education
Bachelors
Assoc−voc
Assoc−acdm
Doctorate
11th
10th
HS−grad
Some−college
Masters

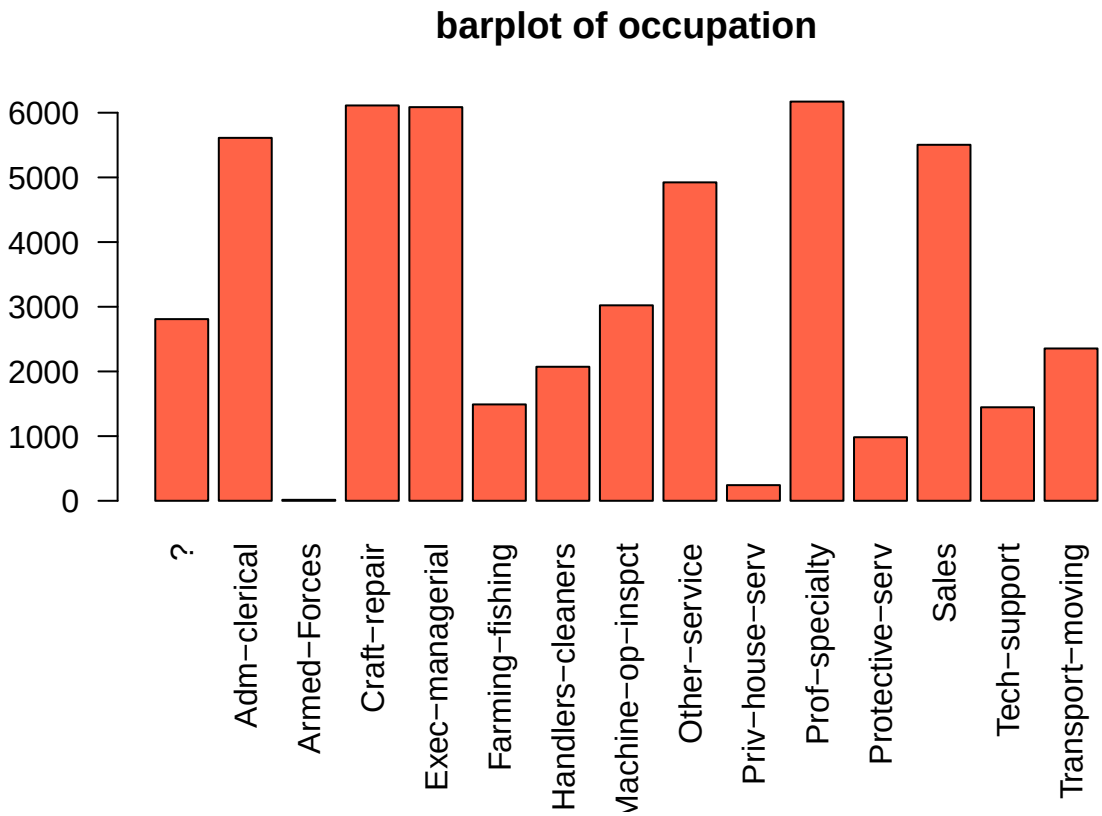

hist of educational.num
Frequency
0
5000
10000
15000
5
10
15
educational.num

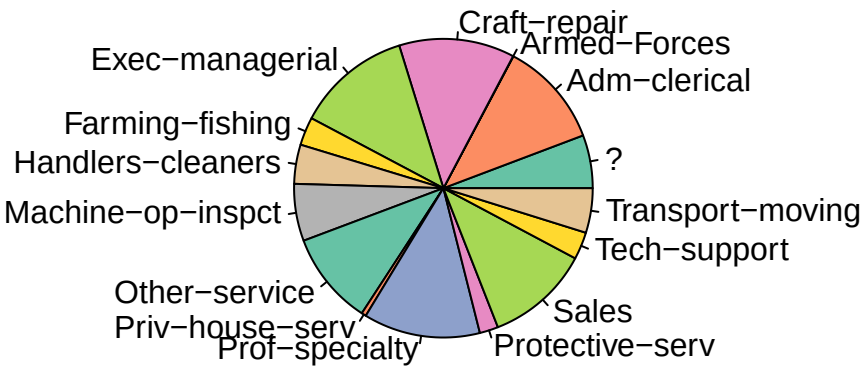


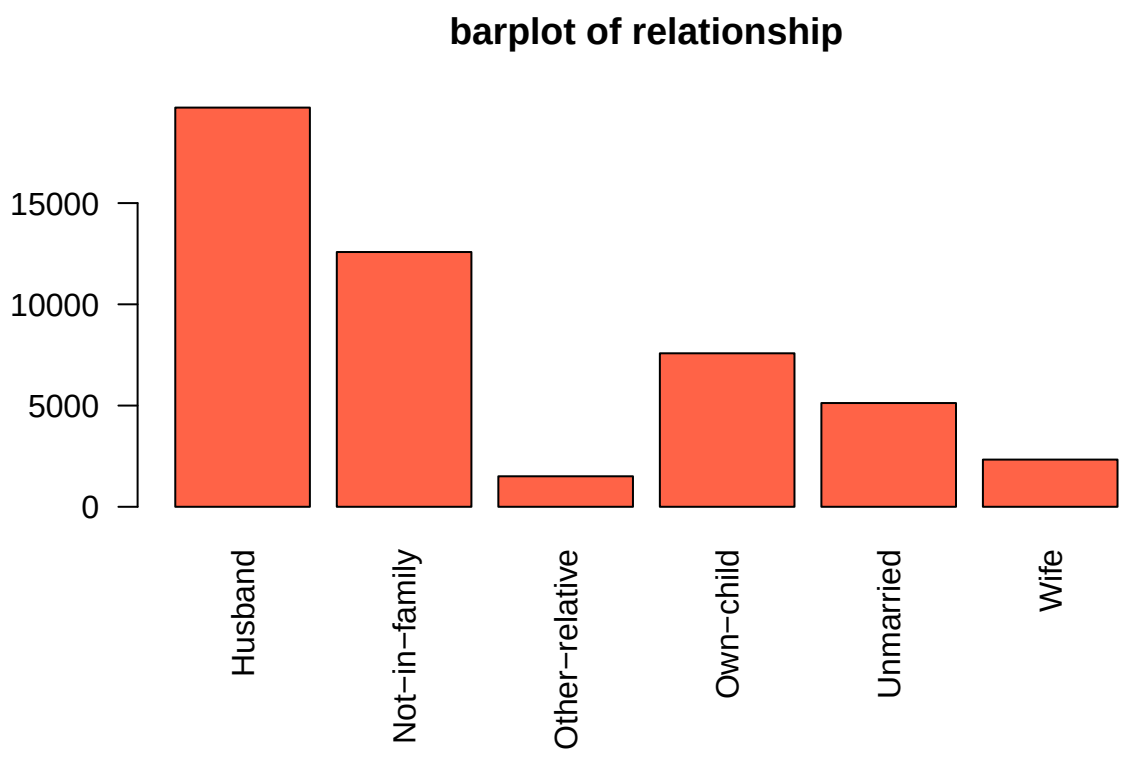

```
## Summary of educational.num :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.08   12.00   16.00
```

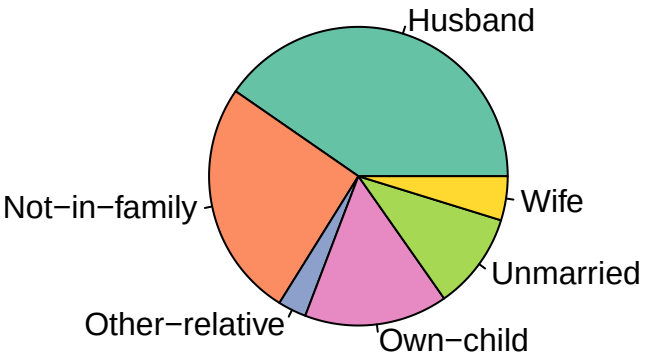

barplot of marital.status
20000
15000
10000
5000
0
Divorced
larried−AF−spouse
larried−civ−spouse
ed−spouse−absent
Never−married
Separated
Widowed

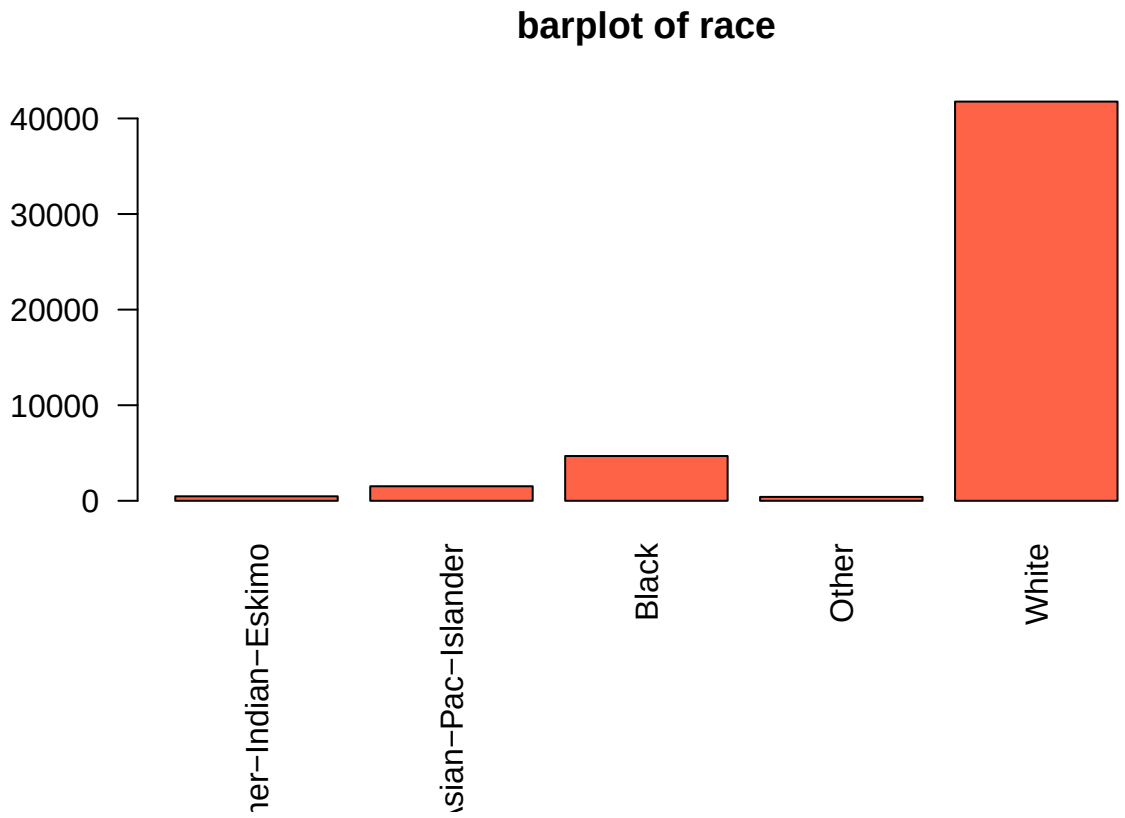

pie of marital.status
Married−civ−spouse
Married−AF−spouse
Divorced
Widowed
Separated
Married−spouse−absent
Never−married

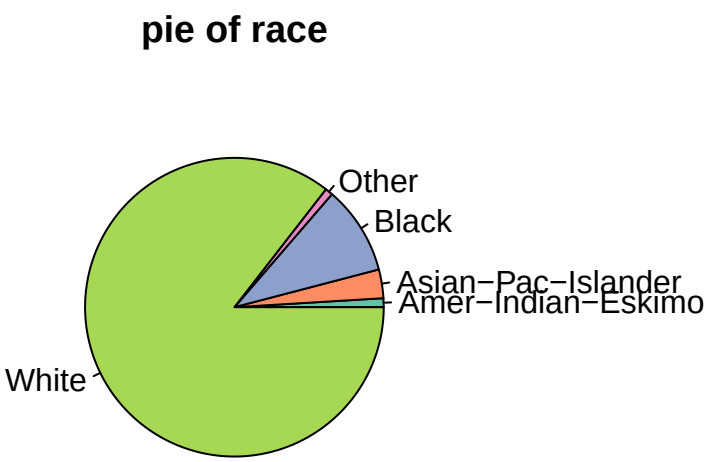

barplot of occupation
6000
5000
4000
3000
2000
1000
0
?
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving

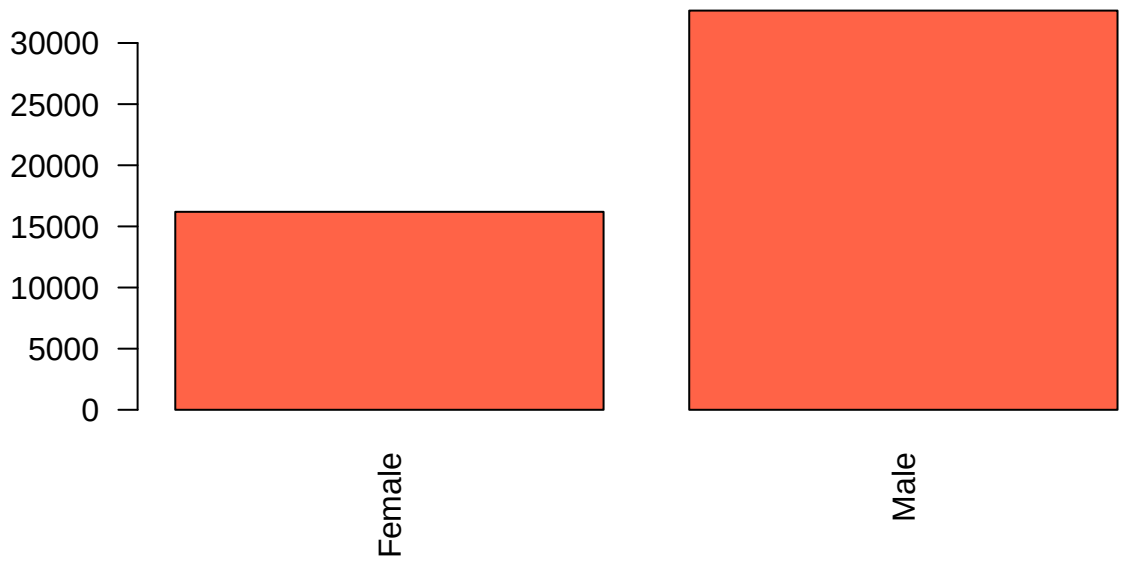

pie of occupation
Craft−repair
Armed−Forces
Adm−clerical
?
Transport−moving
Tech−support
Sales
Protective−serv
Prof−specialty
Priv−house−serv
Other−service
Machine−op−inspct
Handlers−cleaners
Farming−fishing
Exec−managerial


barplot of relationship
15000
10000
5000
0
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife


pie of relationship
Husband
Wife
Unmarried
Own−child
Other−relative
Not−in−family

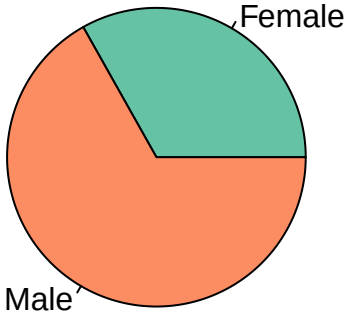

barplot of race
40000
30000
20000
10000
0
ner−Indian−Eskimo
sian−Pac−Islander
Black
Other
White

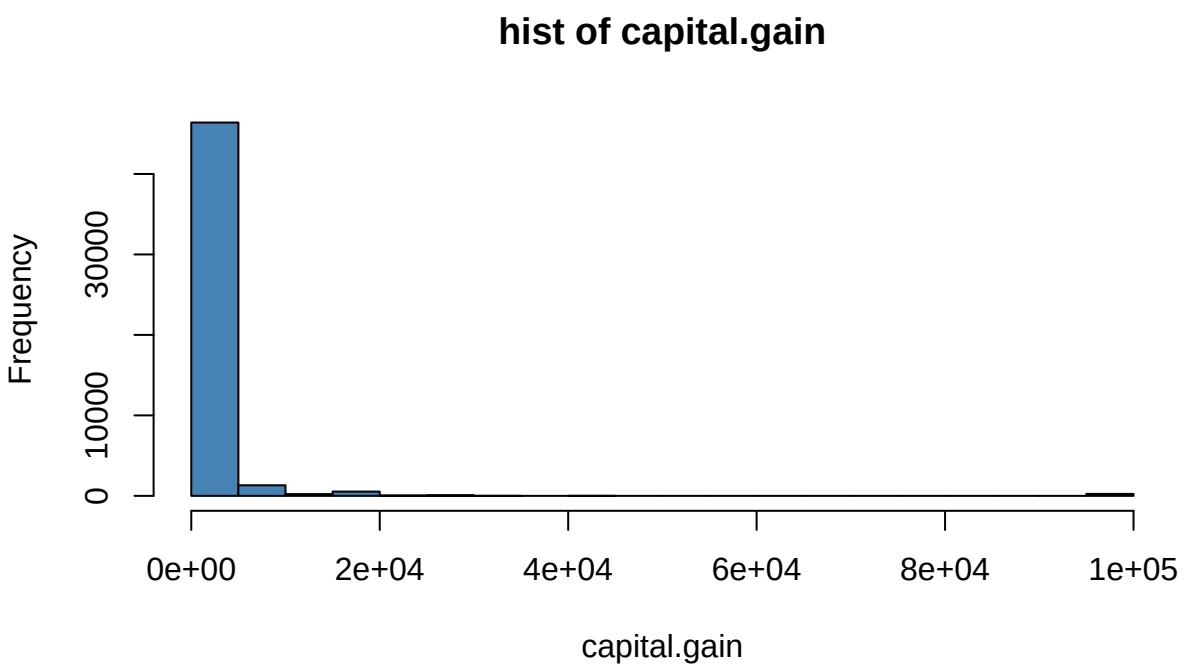

pie of race
Other
Black
Asian−Pac−Islander
Amer−Indian−Eskimo
White

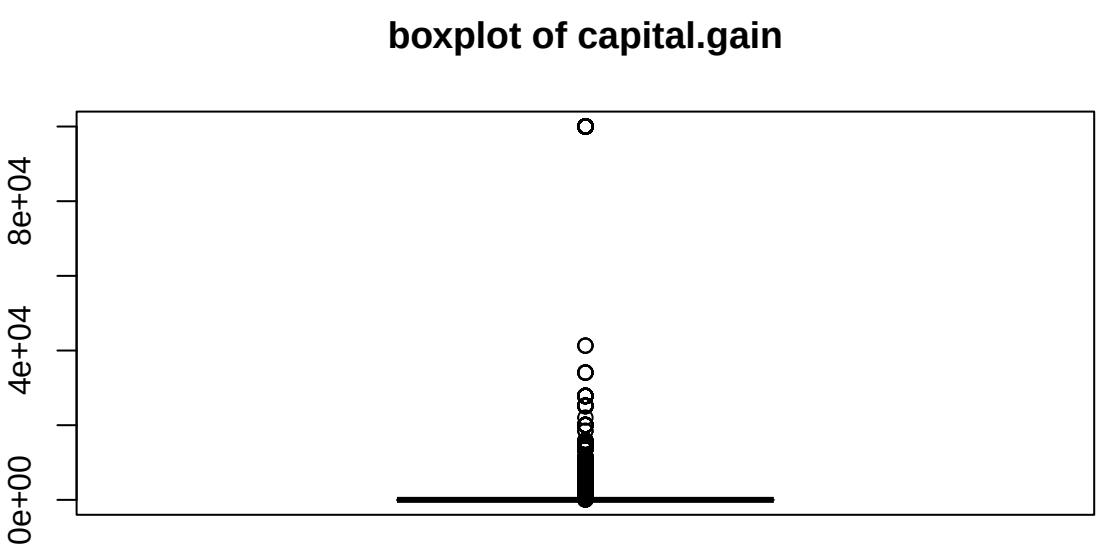



```
## Warning in brewer.pal(min(length(table(dd[, i])), 8), "Set2"): minimal value for n is 3, returning re
```

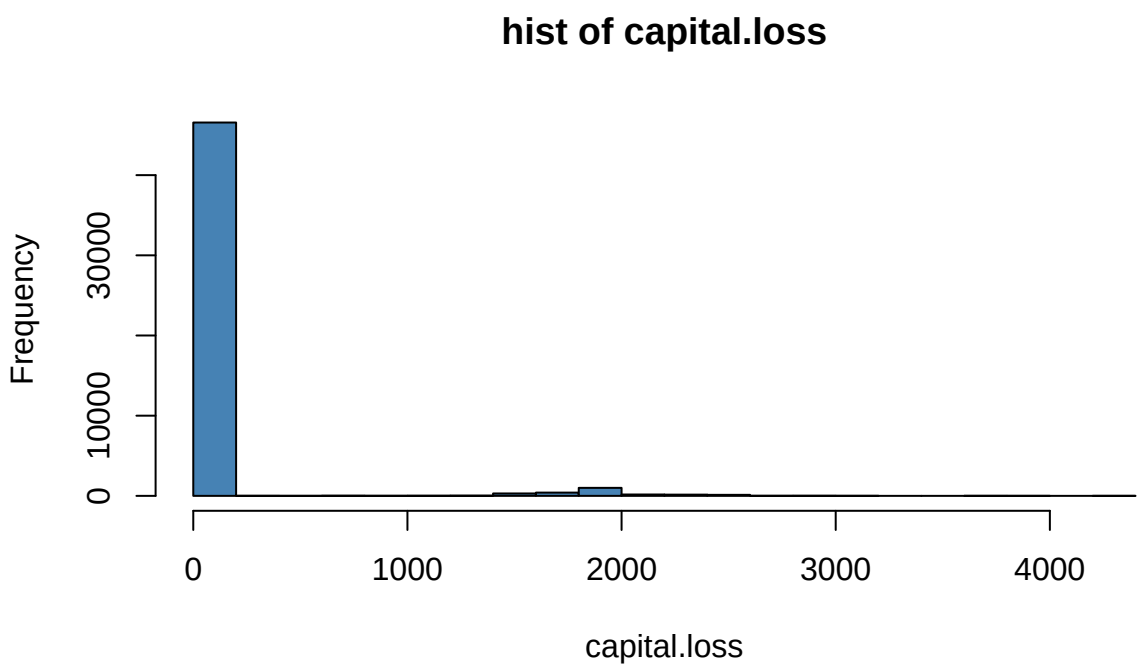
pie of gender

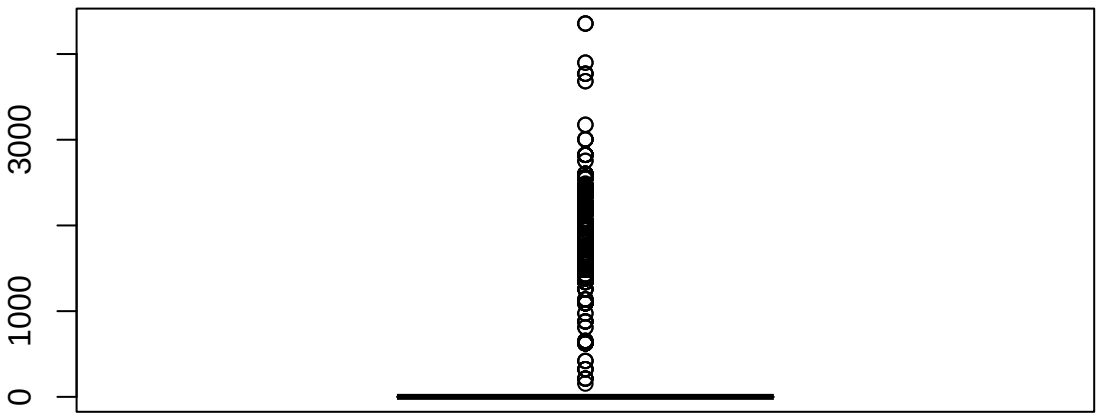



hist of capital.gain
Frequency
0
10000
30000
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
capital.gain

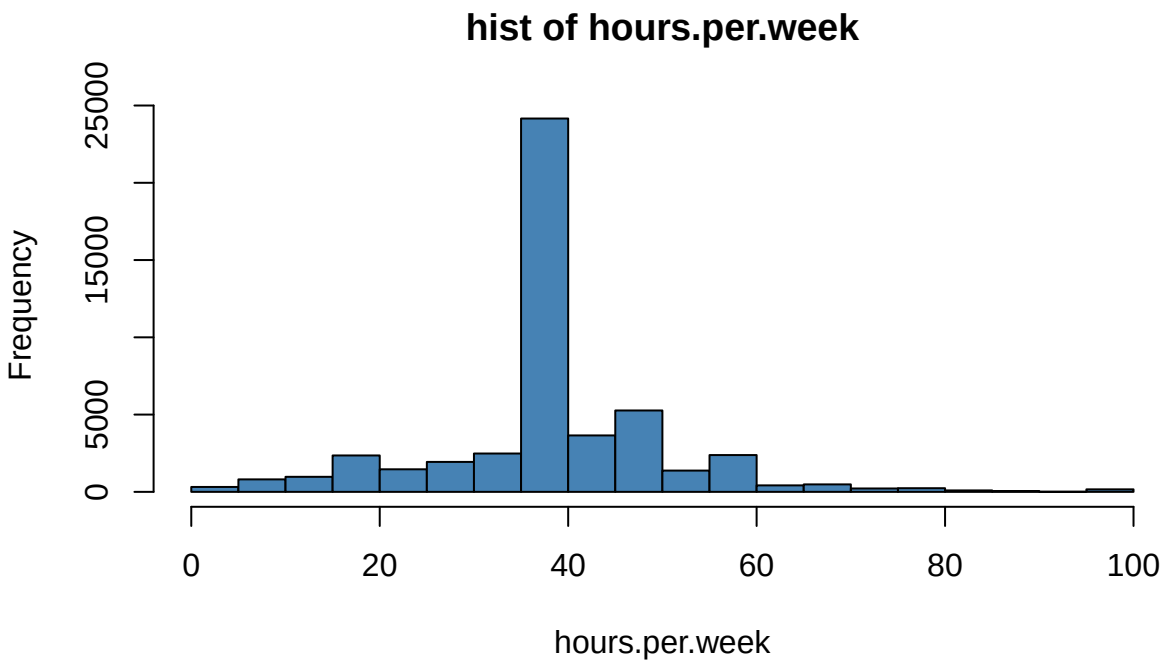


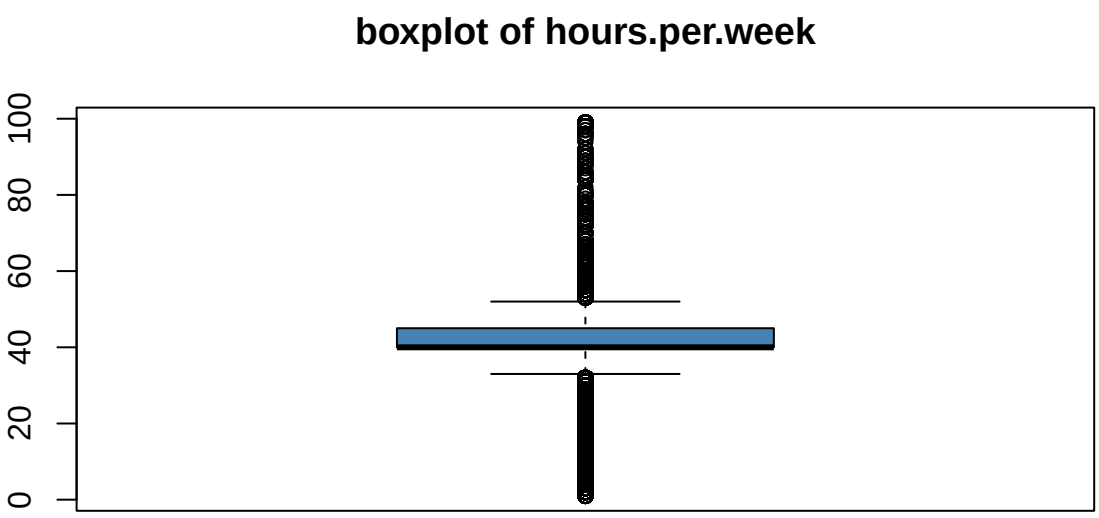

```
## Summary of capital.gain :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1079       0   99999
```


hist of capital.loss
Frequency
0
10000
30000
0
1000
2000
3000
4000
capital.loss

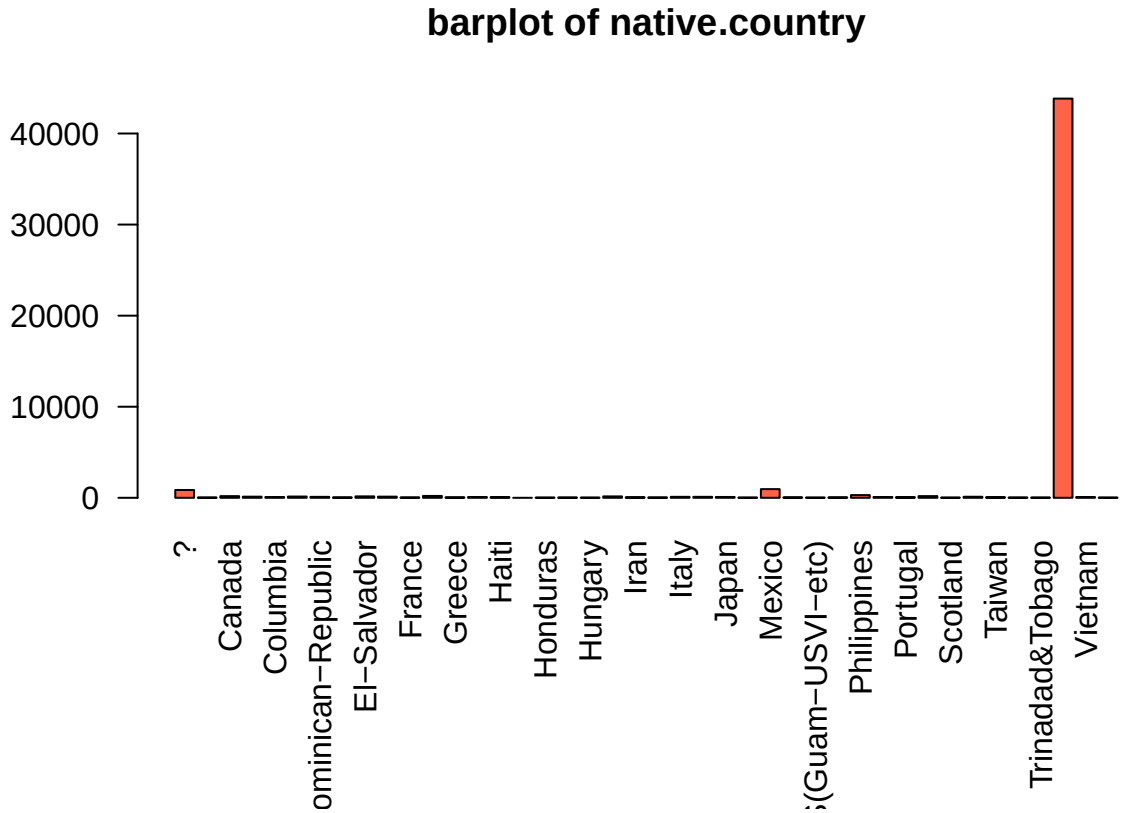


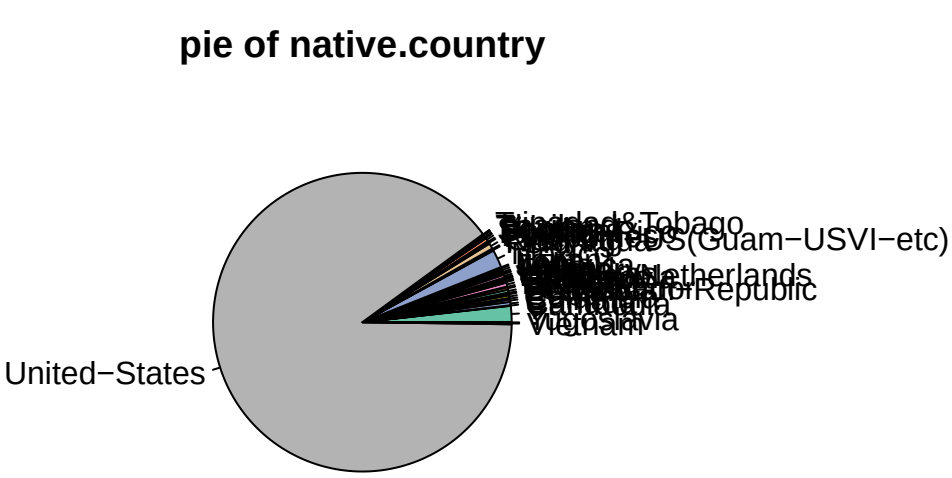

```
## Summary of capital.loss :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0    87.5     0.0  4356.0
```

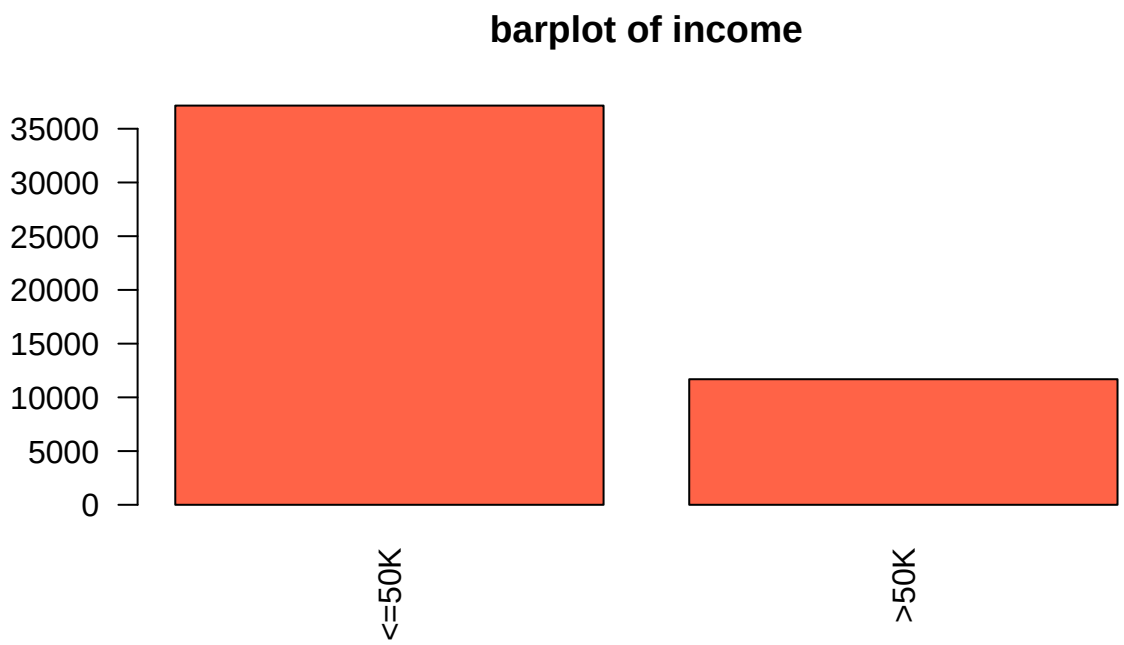

hist of hours.per.week
Frequency
0
5000
15000
25000
0
20
40
60
80
100
hours.per.week



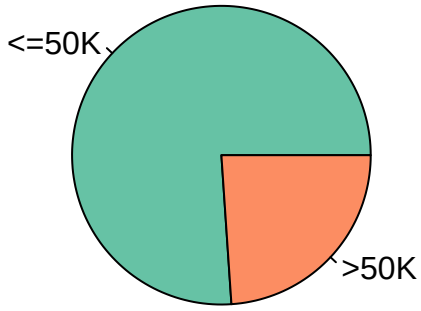

```
## Summary of hours.per.week :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.42   45.00   99.00
```


barplot of native.country
40000
30000
20000
10000
0
?
Canada
Columbia
ominican−Republic
El−Salvador
France
Greece
Haiti
Honduras
Hungary
Iran
Italy
Japan
Mexico
s(Guam−USVI−etc)
Philippines
Portugal
Scotland
Taiwan
Trinadad&Tobago
Vietnam


pie of native.country
United−States
Vietnam




```
## Warning in brewer.pal(min(length(table(dd[, i])), 8), "Set2"): minimal value for n is 3, returning r
```


pie of income
<=50K
>50K