Basic_descriptive_ap

2025-02-24

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

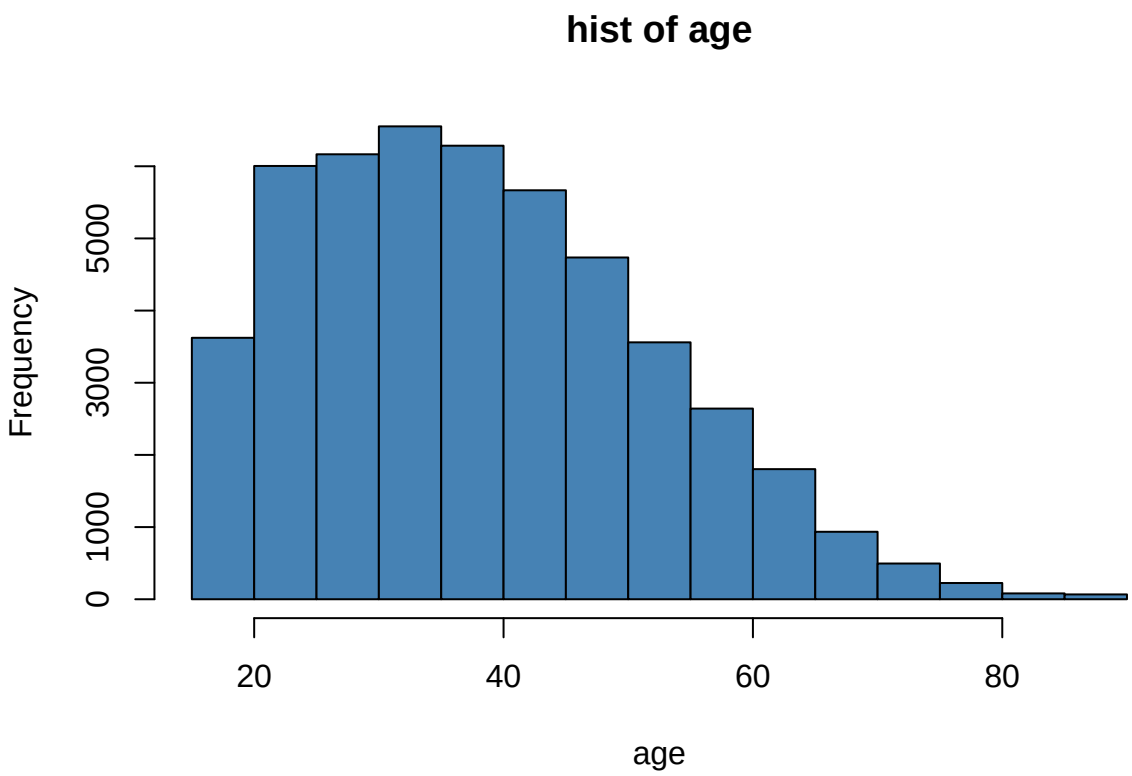
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

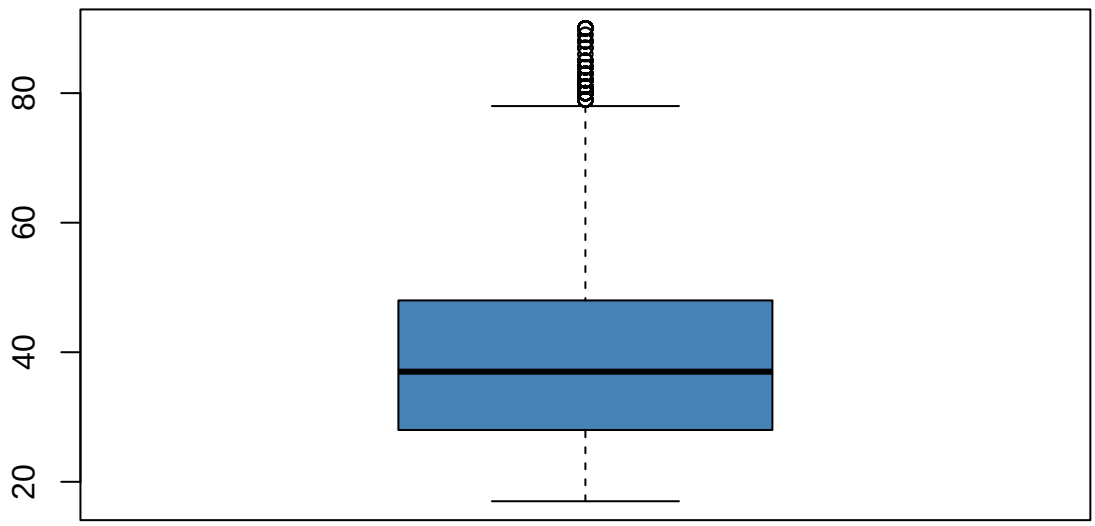
```
library(RColorBrewer)
dd <- read.csv("~/Documents/ADEI/Project/adult.csv")  # Read data

#Rename columns
colnames(dd) <- c("age", "workclass", "fnlwgt", "education", "edu_num", "marital", "occupation",
                  "relationship", "race", "sex", "cap_gain", "cap_loss",
                  "hours_week", "native_country", "income")


#Rename modalities
dd$workclass <- recode(dd$workclass,
                       "Private" = "Priv",
                       "Self-emp-not-inc" = "SelfN",
                       "Self-emp-inc" = "SelfI",
                       "Federal-gov" = "Fed",
                       "Local-gov" = "Loc",
                       "State-gov" = "State",
                       "Without-pay" = "NoPay",
                       "Never-worked" = "NoPay")

dd$marital <- recode(dd$marital,
                     "Never-married" = "NevMarr",
                     "Married-civ-spouse" = "Married",
                     "Married-AF-spouse" = "Married",
                     "Married-spouse-absent" = "Sep",
                     "Separated" = "Sep",
                     "Divorced" = "Div",
                     "Widowed" = "Widow")

dd$occupation <- recode(dd$occupation,
                        "Exec-managerial" = "ExecMan",
                        "Prof-specialty" = "Prof",
                        "Adm-clerical" = "AdminCler",
```

```
                          "Sales" = "Sales",
                          "Craft-repair" = "CraftRep",
                          "Transport-moving" = "Trans",
                          "Handlers-cleaners" = "HandlCl",
                          "Machine-op-inspct" = "MachOp",
                          "Tech-support" = "Tech",
                          "Protective-serv" = "ProtServ",
                          "Armed-Forces" = "Army",
                          "Farming-fishing" = "FarmFish",
                          "Priv-house-serv" = "House",
                          "Other-service" = "Other")

dd$native_country <- recode(dd$native_country,
                            "United-States" = "USA",
                            .default = "Other")  # Group all other countries as "Other"


#a function to find the mode (most frequent value)
fill_mode <- function(x) {
  mode_value <- names(sort(table(x), decreasing=TRUE))[1]  # Get the most frequent value
  x[is.na(x)| x == "?"] <- mode_value  # Replace NA with mode
  return(x)
}


#we apply it to columns with missing values
dd$workclass <- fill_mode(dd$workclass)
dd$occupation <- fill_mode(dd$occupation)
dd$native_country <- fill_mode(dd$native_country)

#we drop education (the same information is found in the column educational.num (in numbers))
dd <- subset(dd, select = -c(education))

class(dd[,1])
```

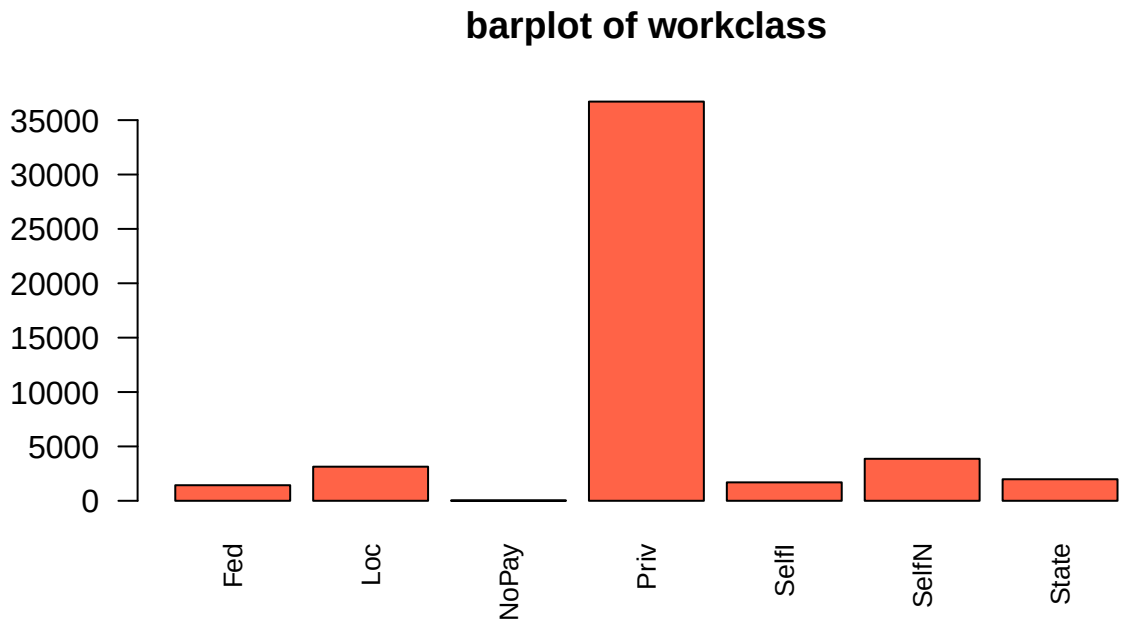
```
## [1] "integer"
```

```
for ( i in 1:14){
  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, cex.names =
    pie(table(dd[,i]),main=paste("pie of", names(dd)[i]), col = brewer.pal(min(length(table(dd[, i])),
  }
}
```

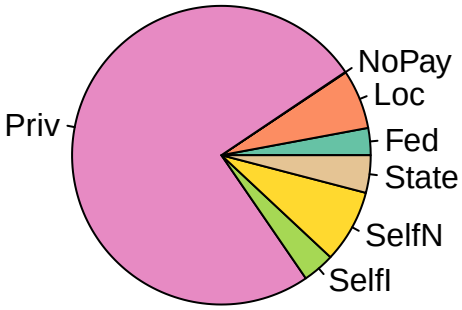

hist of age
Frequency
0
1000
3000
5000
20
40
60
80
age

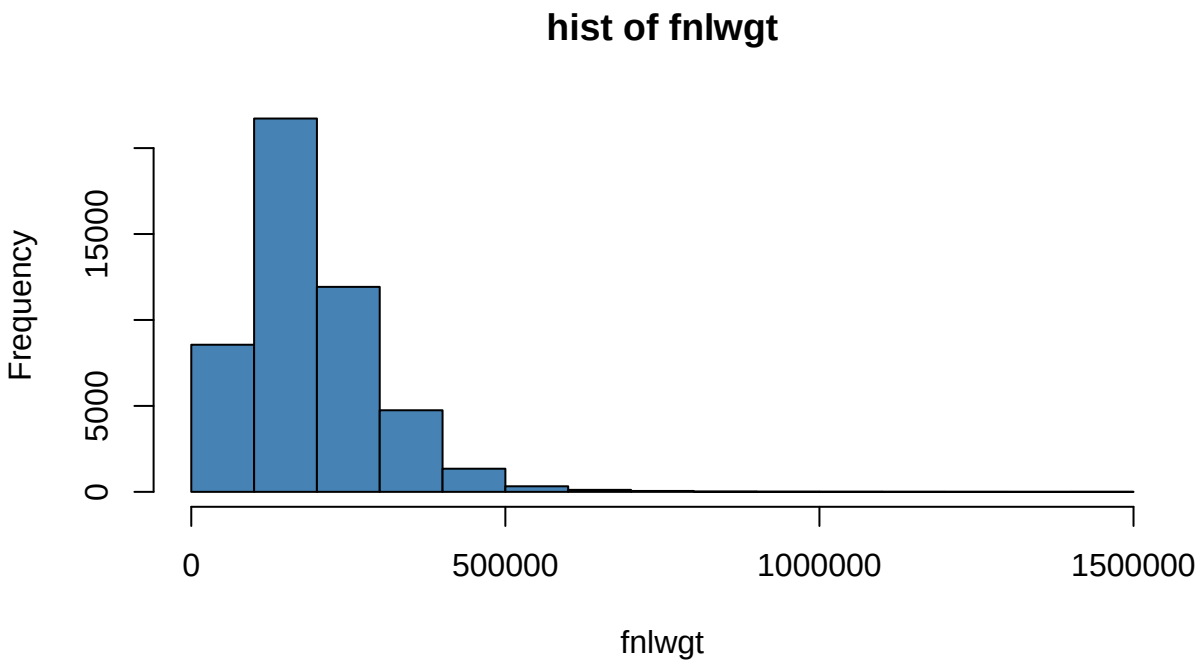


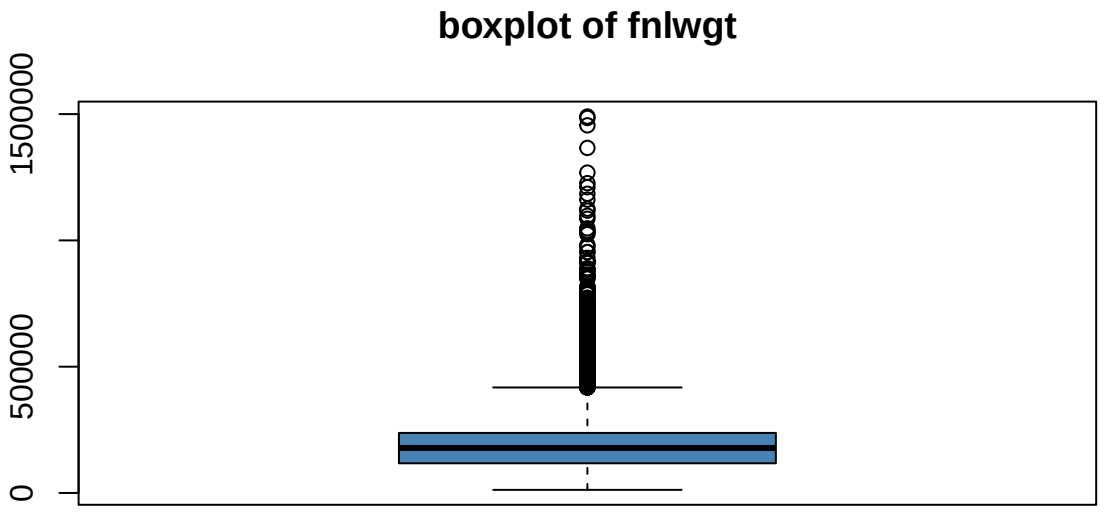

```
## Summary of age :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.64   48.00   90.00
```


barplot of workclass
35000
30000
25000
20000
15000
10000
5000
0
Fed
Loc
NoPay
Priv
Selfi
SelfN
State

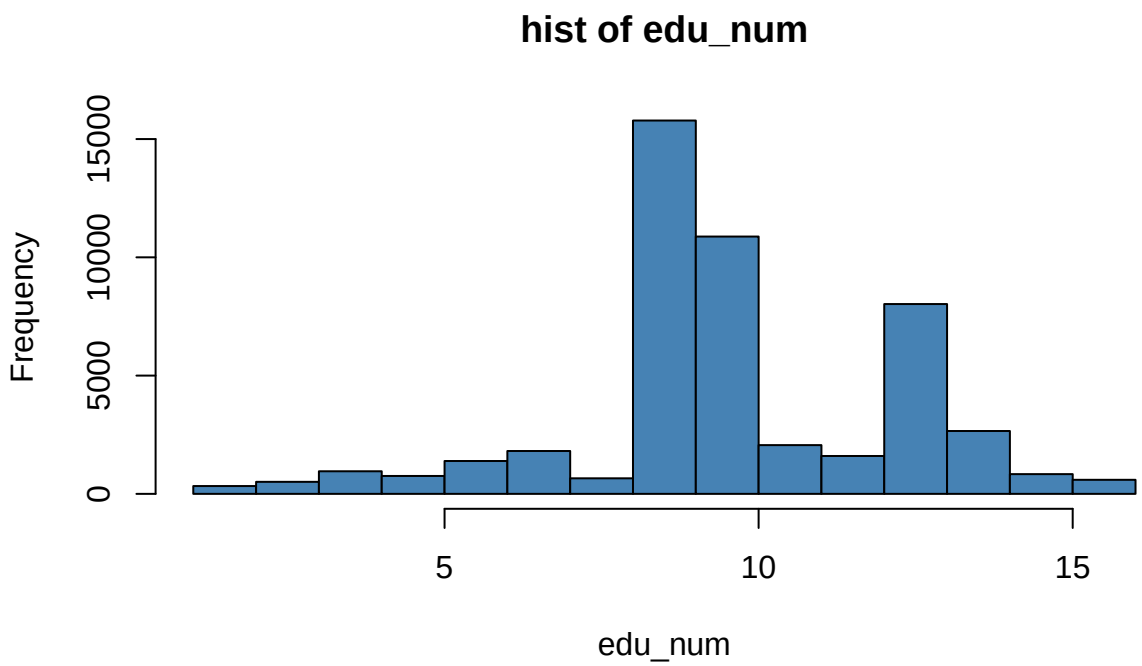
pie of workclass

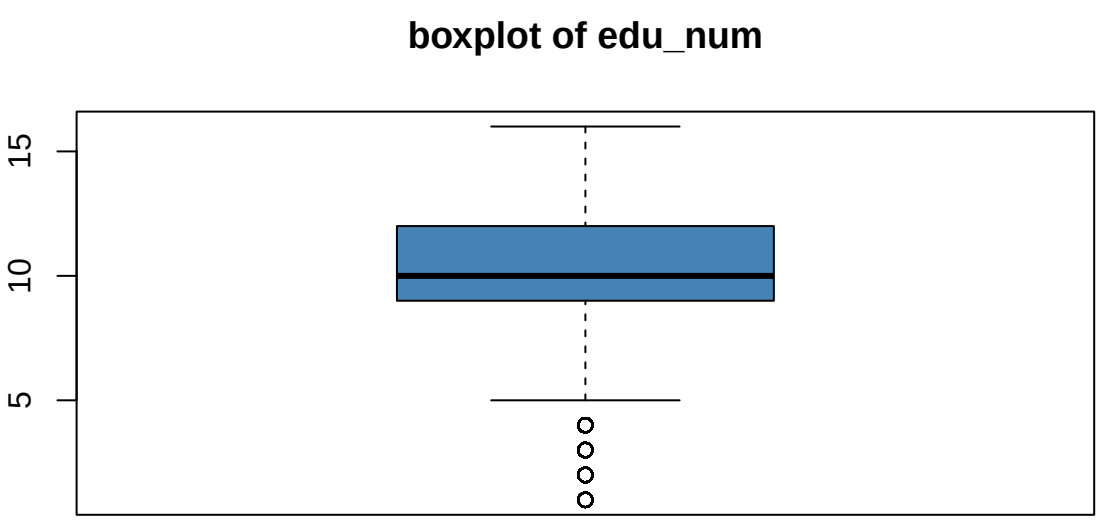



hist of fnlwgt
Frequency
0
5000
15000
0
500000
1000000
1500000
fnlwgt

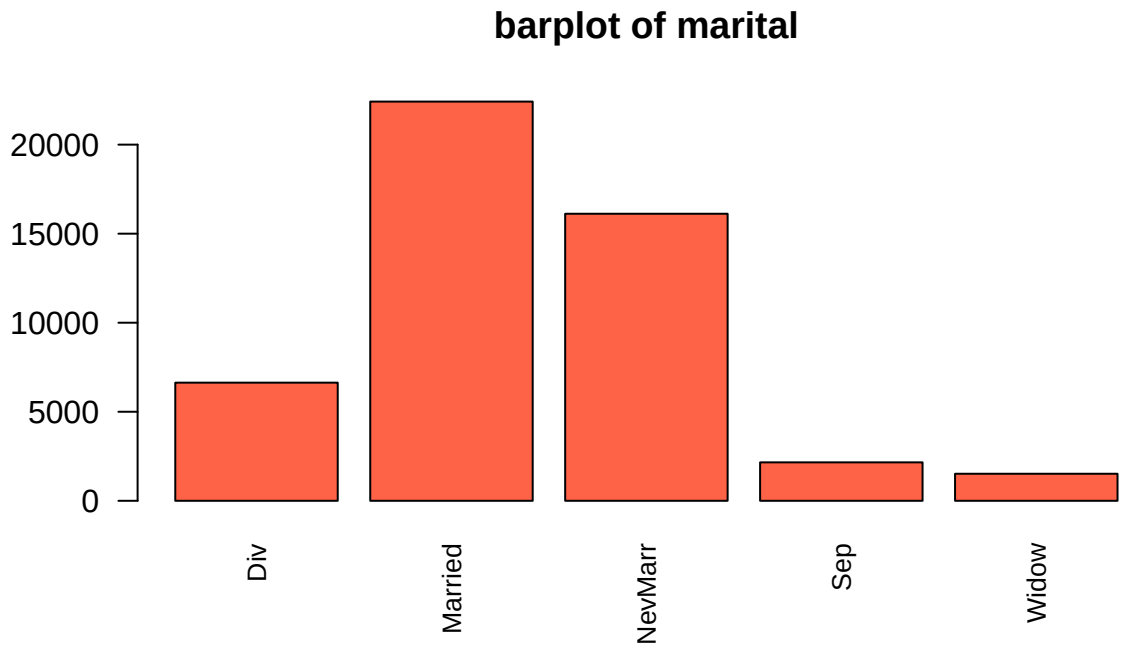


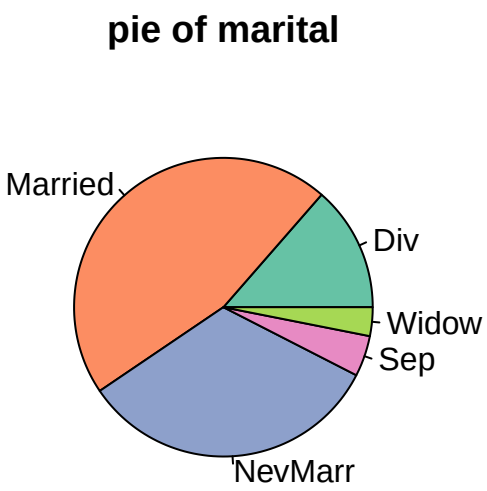

```
## Summary of fnlwgt :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12285  117550  178144  189664  237642 1490400
```

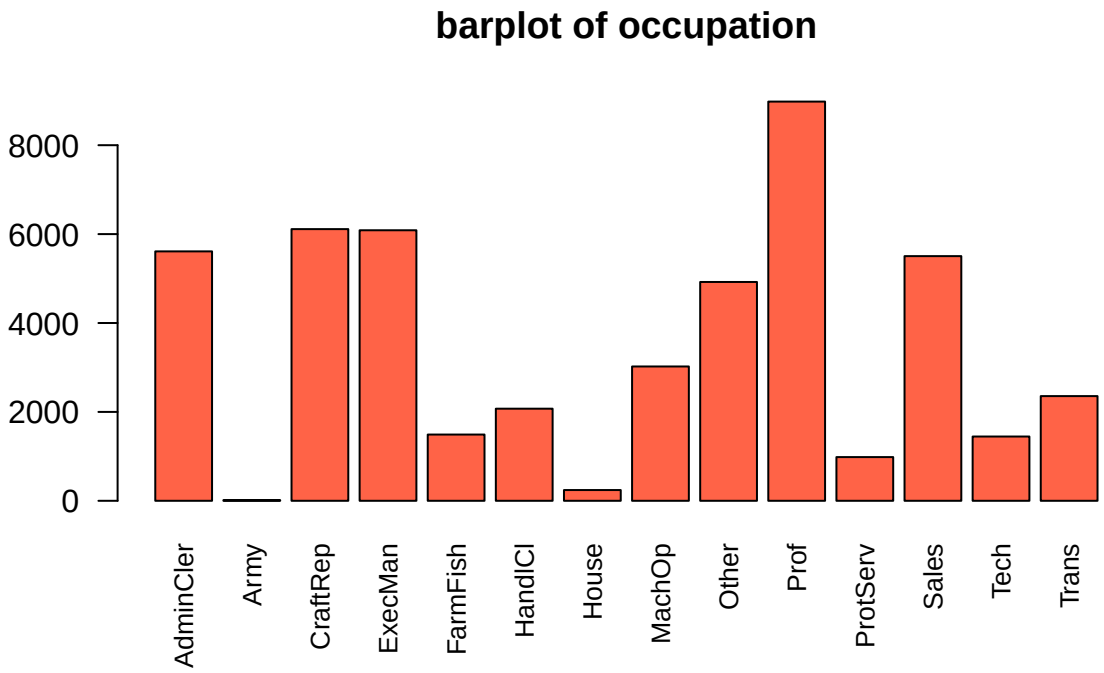

hist of edu_num
Frequency
0
5000
10000
15000
5
10
15
edu_num



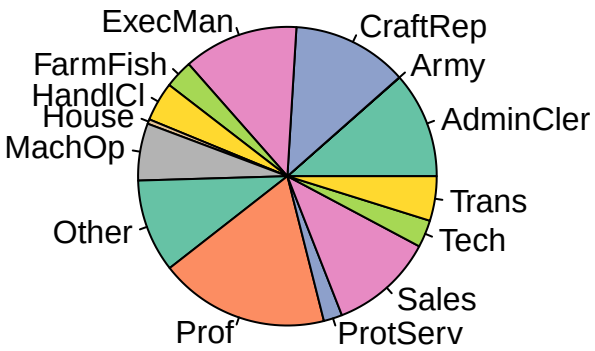

```
## Summary of edu_num :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   10.00   10.08   12.00   16.00
```

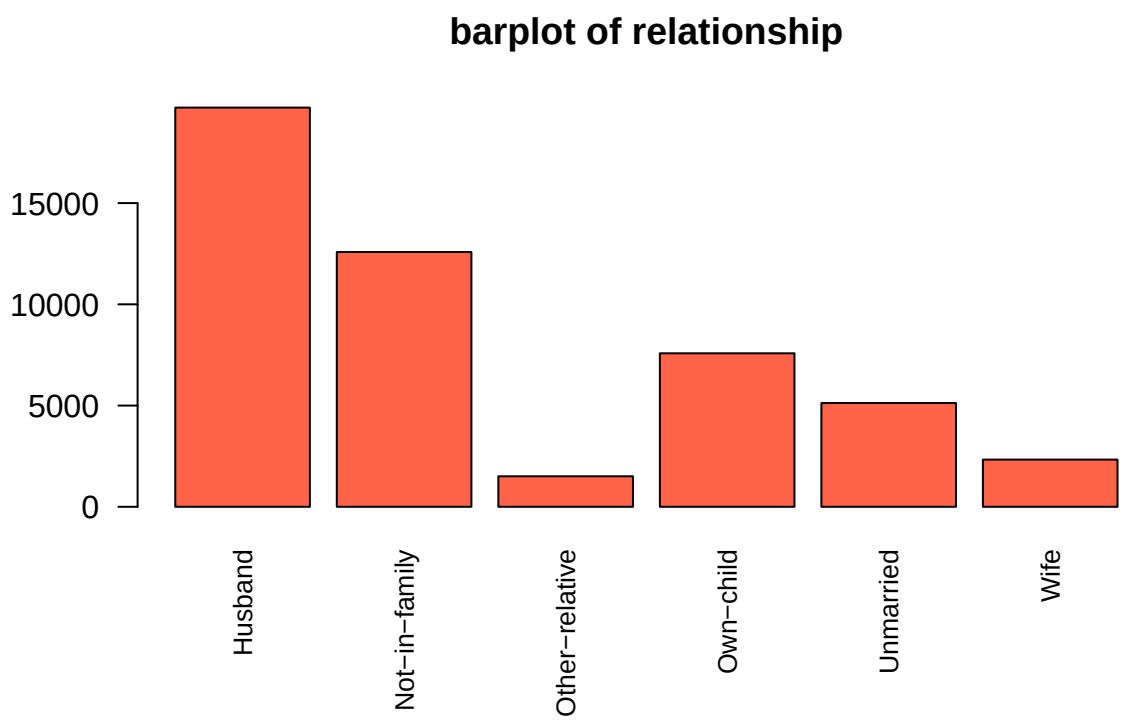

barplot of marital
20000
15000
10000
5000
0
Div
Married
NevMarr
Sep
Widow

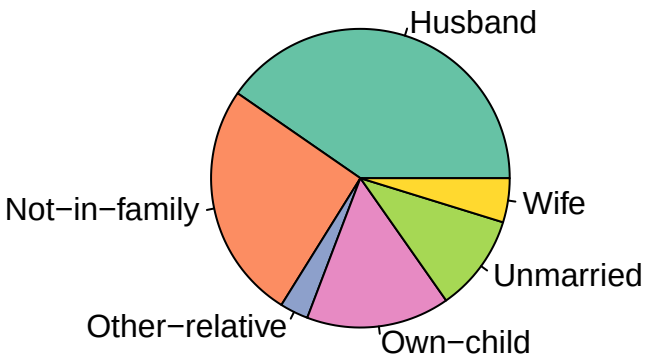

pie of marital
Married
Div
Widow
Sep
NevMarr

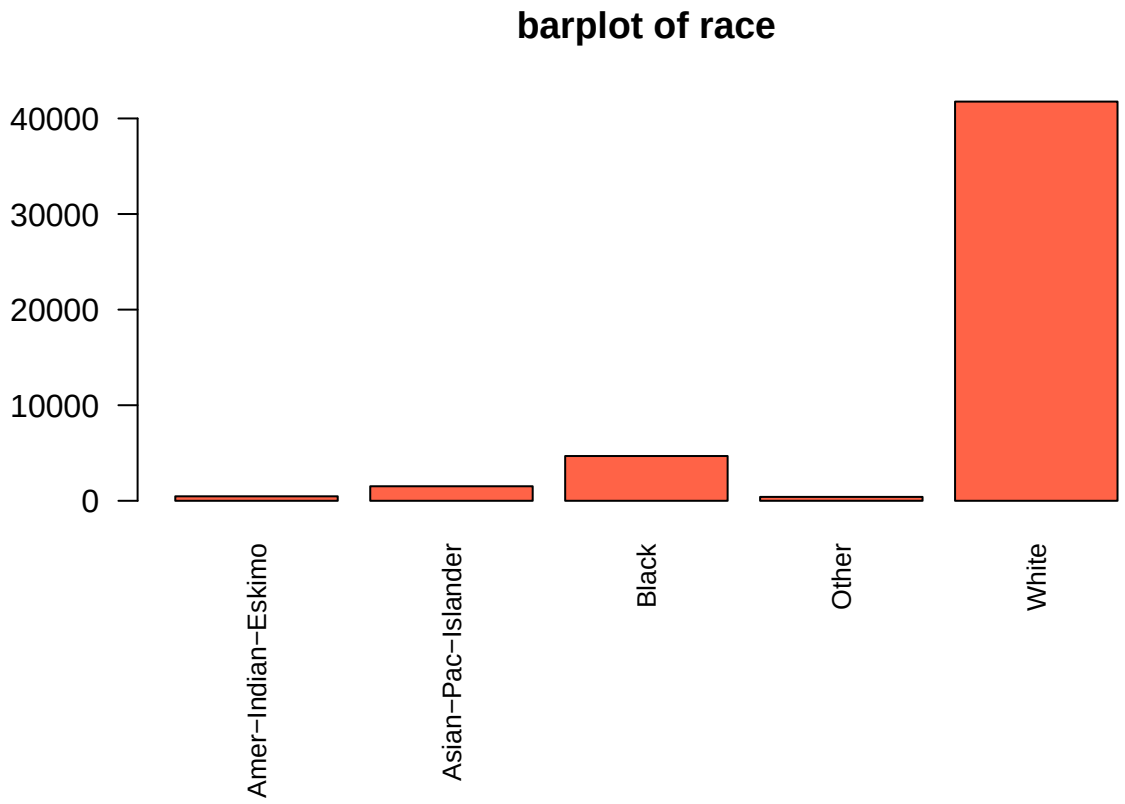

barplot of occupation
8000
6000
4000
2000
0
AdminCler
Army
CraftRep
ExecMan
FarmFish
HandlCl
House
MachOp
Other
Prof
ProtServ
Sales
Tech
Trans

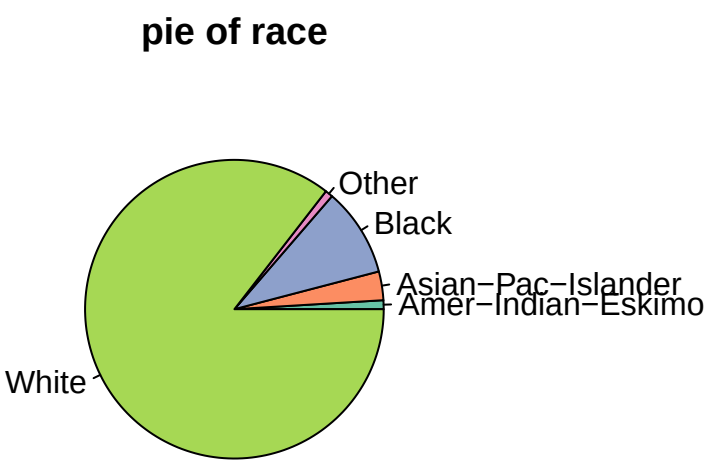
pie of occupation

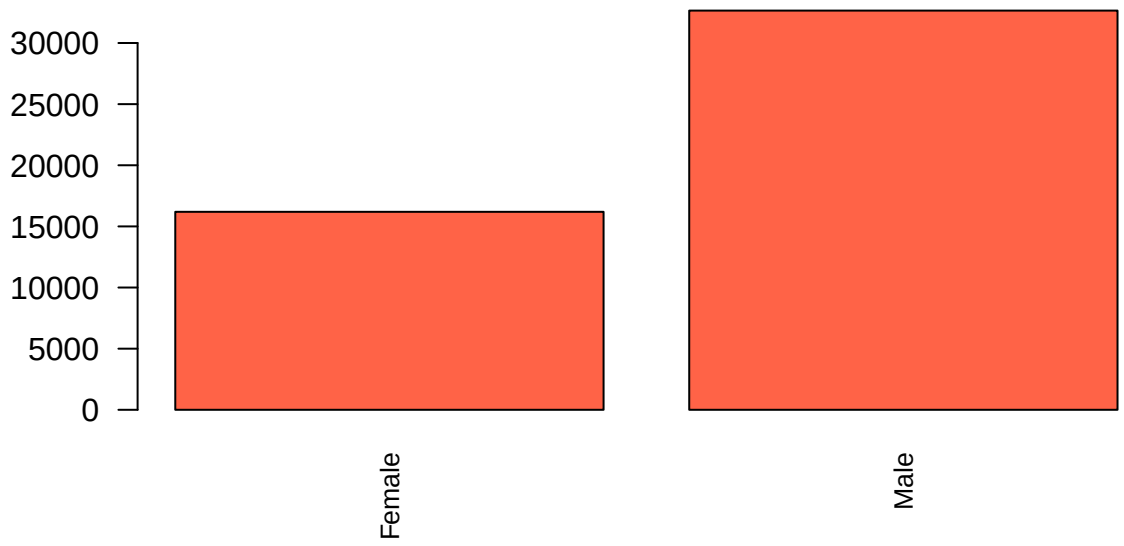



barplot of relationship
15000
10000
5000
0
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife

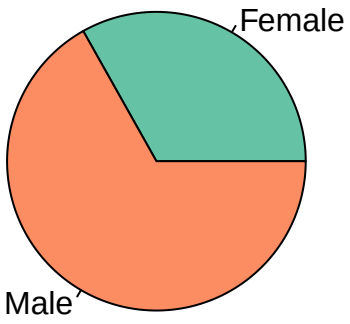

pie of relationship
Husband
Wife
Unmarried
Own−child
Other−relative
Not−in−family

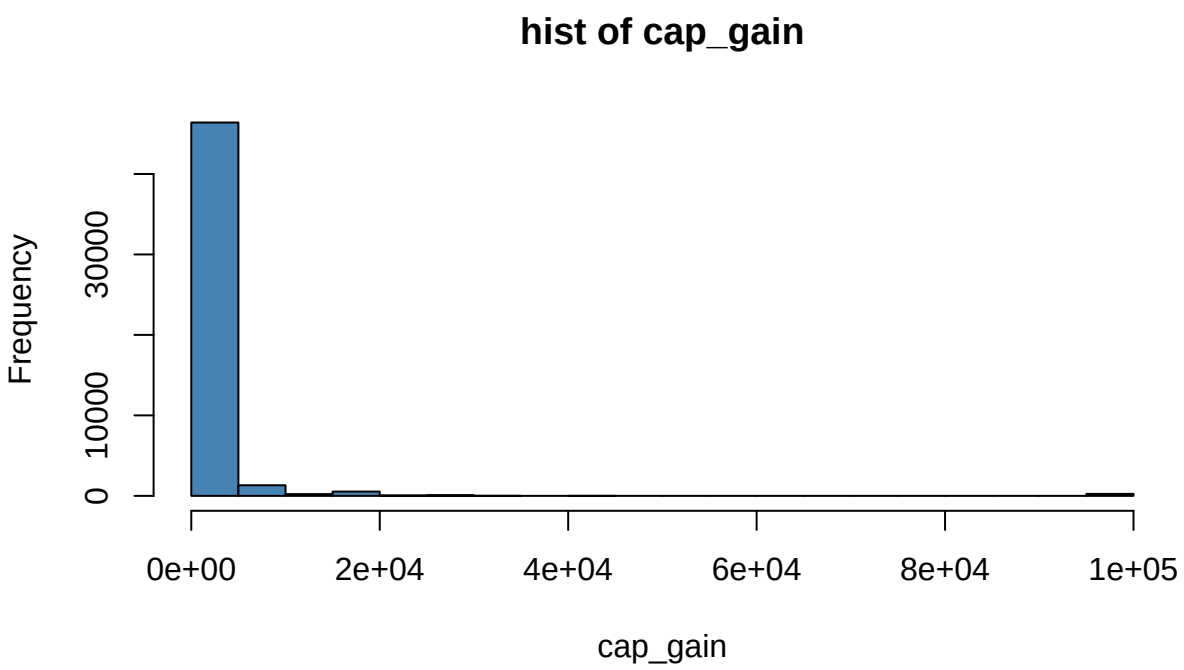

barplot of race
40000
30000
20000
10000
0
Amer−Indian−Eskimo
Asian−Pac−Islander
Black
Other
White

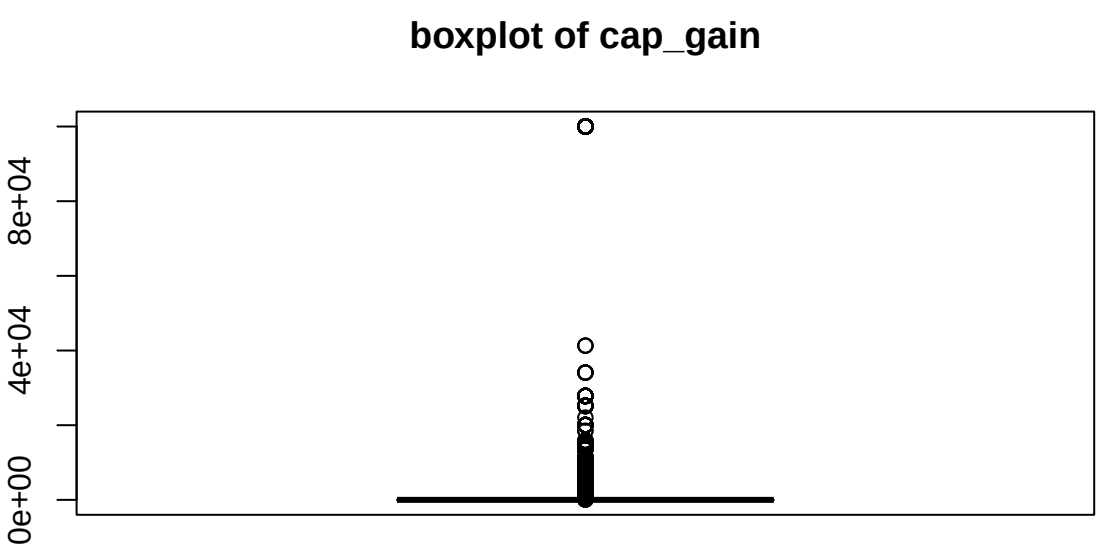

pie of race
Other
Black
Asian−Pac−Islander
Amer−Indian−Eskimo
White



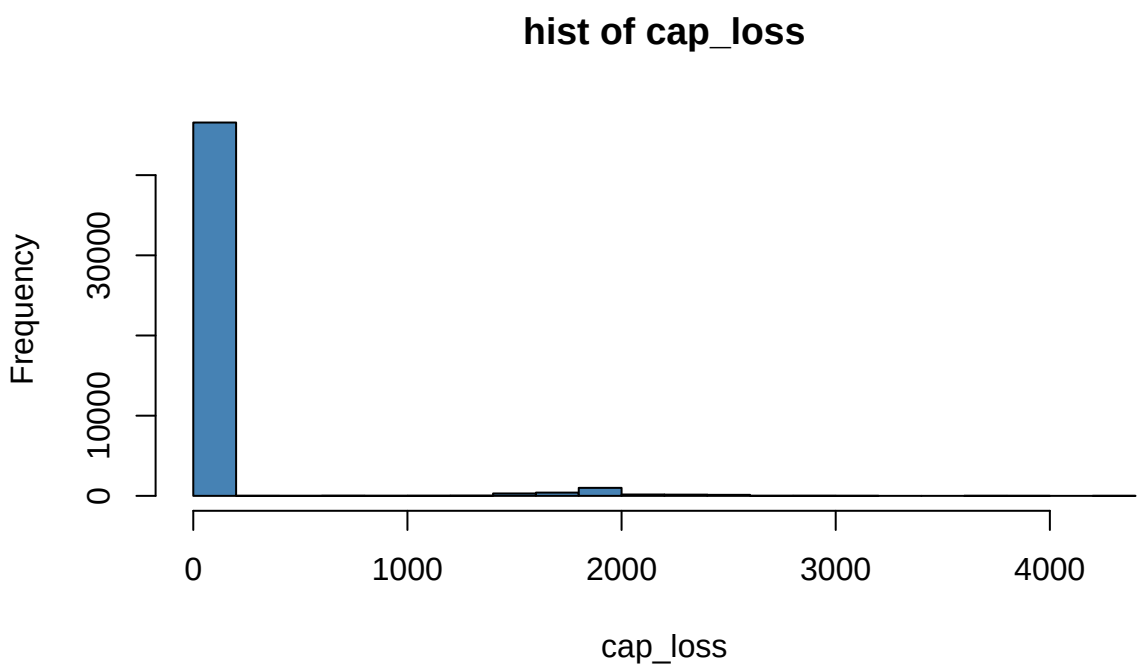

```
## Warning in brewer.pal(min(length(table(dd[, i])), 8), "Set2"): minimal value for n is 3, returning re
```

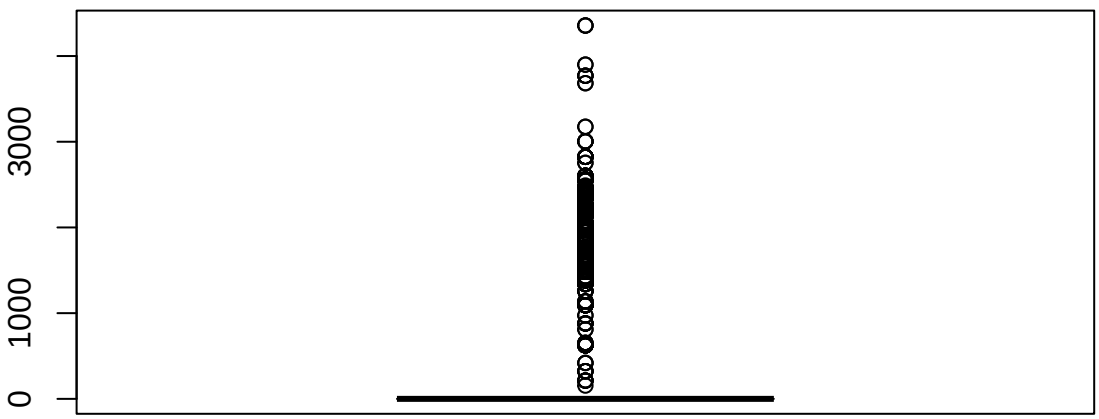

pie of sex
Female
Male


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

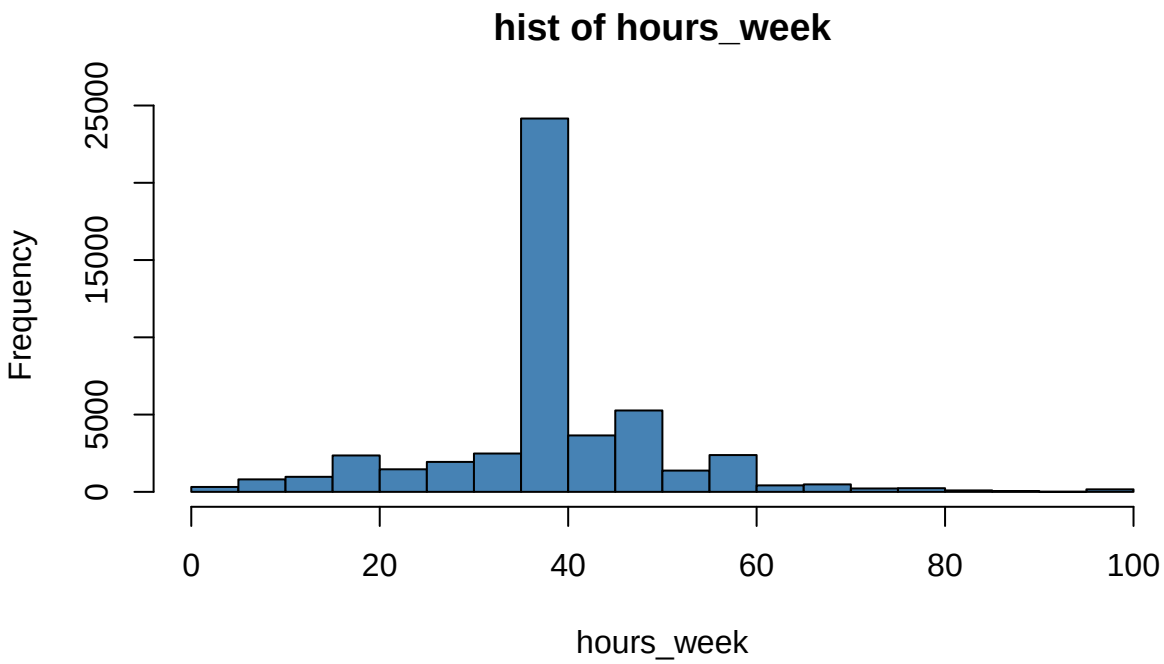


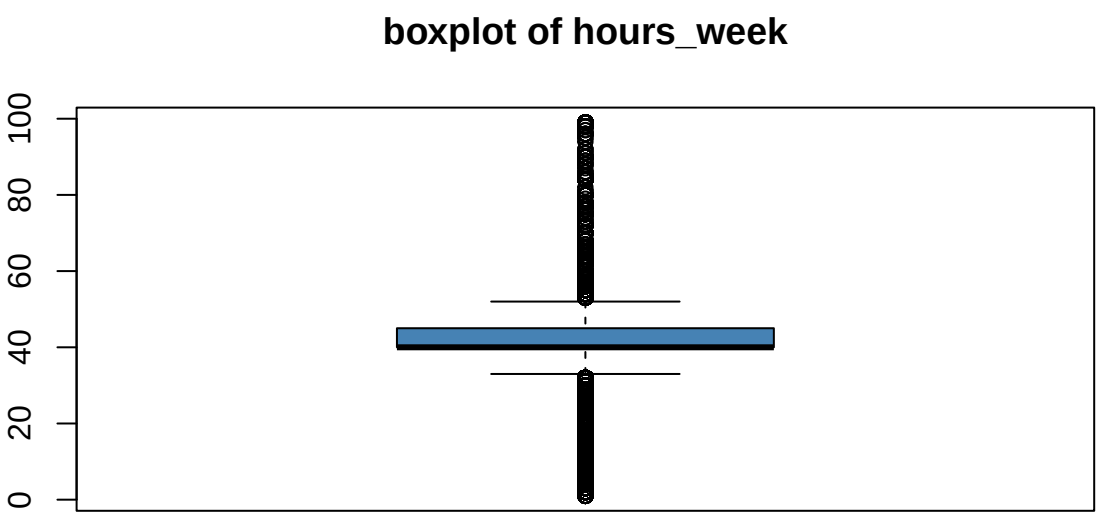

```
## Summary of cap_gain :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0    1079       0   99999
```


hist of cap_loss
Frequency
0
10000
30000
0
1000
2000
3000
4000
cap_loss

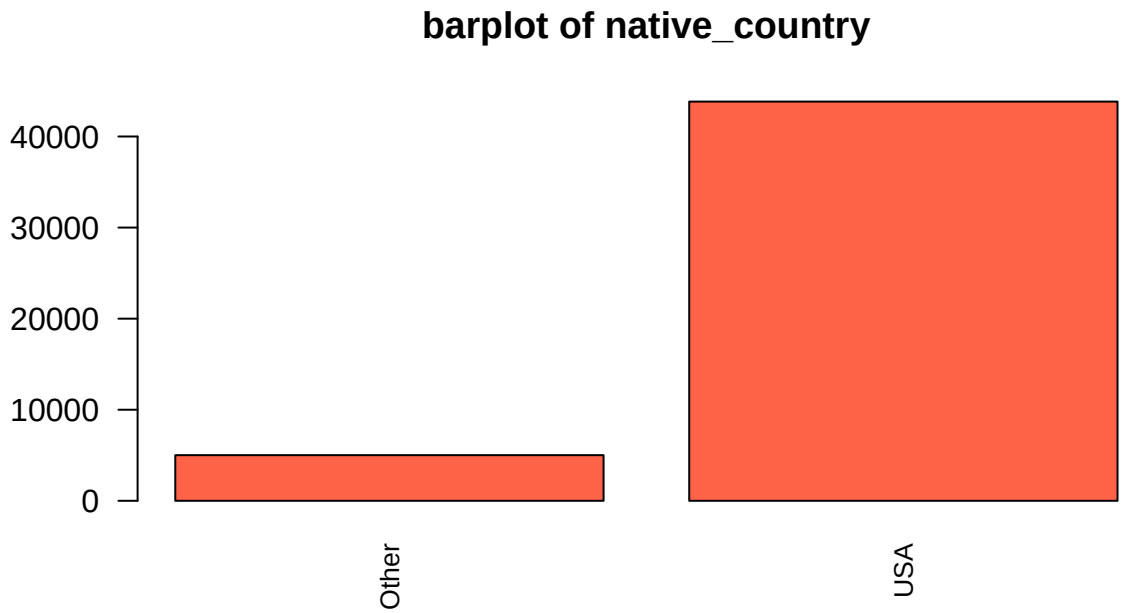



```
## Summary of cap_loss :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0    87.5     0.0  4356.0
```

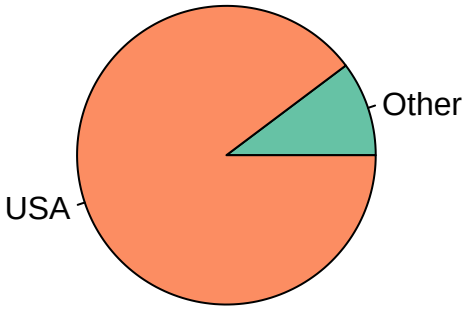

hist of hours_week
Frequency
0
5000
15000
25000
0
20
40
60
80
100
hours_week

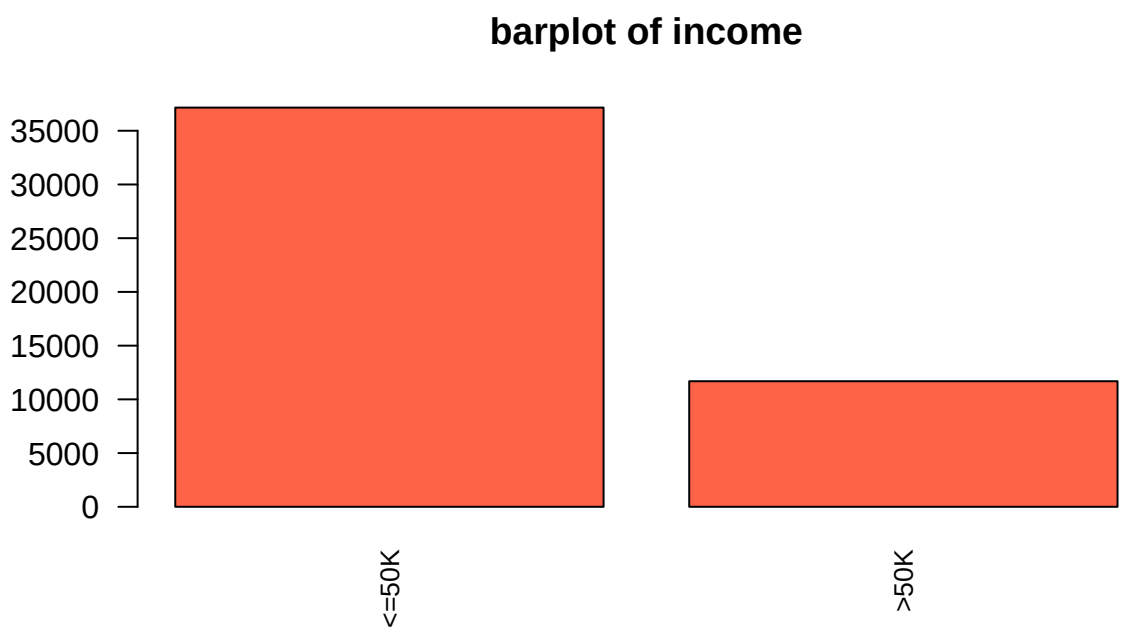



```
## Summary of hours_week :
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.42   45.00   99.00
```

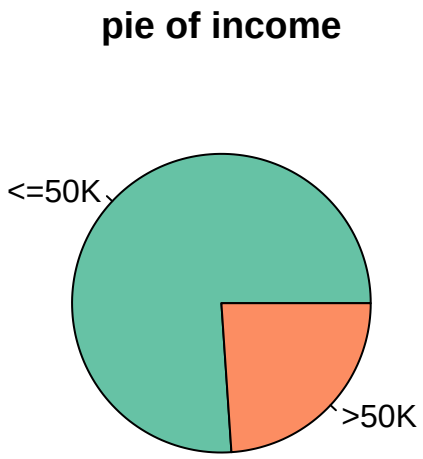



```
## Warning in brewer.pal(min(length(table(dd[, i])), 8), "Set2"): minimal value for n is 3, returning r
```


pie of native_country
Other
USA




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


pie of income
<=50K
>50K