initialAnalysis.R

Ameya Nagnur

Kavya Kotian Bhaskar

Ketan Balbhim Kokane

Siddarth Sargunaraj

2019-04-07

```
# Introduction to Big Data
# Phase 3
#
# Data used:
#     The data is from a census bureau database.
#
# This script file reads the data, cleans it of missing values and visualizes
the data by plotting histograms of each feature
#
# Dependencies (Libraries used):
#     1. corrplot (used to display the correlation matrix of the dataset)
#
#
# Installing and loading dependencies
#
# Install the corrplot library
install.packages("corrplot")

#Load required libraries
library(corrplot)

## corrplot 0.84 loaded

#
# Preprocessing and cleaning the data
#


# Read the data into a data frame
adultDataset = read.table("adult.data", header= TRUE, sep = ",")
# Convert the dataset to integer format
adultDataset[] <- lapply(adultDataset,as.integer)
# Removing null values
na.omit(adultDataset)

# Attach the database to the R search path
attach(adultDataset)
```

```r
#
# Printing details of the dataset
#

# Print the summary of the dataset
summary(adultDataset)
```

```
##       age          workclass         fnlwgt          education
##  Min.   :17.00   Min.   :1.000   Min.   :  12285   Min.   : 1.0
##  1st Qu.:28.00   1st Qu.:5.000   1st Qu.: 117827   1st Qu.:10.0
##  Median :37.00   Median :5.000   Median : 178356   Median :12.0
##  Mean   :38.58   Mean   :4.869   Mean   : 189778   Mean   :11.3
##  3rd Qu.:48.00   3rd Qu.:5.000   3rd Qu.: 237051   3rd Qu.:13.0
##  Max.   :90.00   Max.   :9.000   Max.   :1484705   Max.   :16.0
##  education.num   marital.status    occupation      relationship
##  Min.   : 1.00   Min.   :1.000   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:3.000   1st Qu.: 4.000   1st Qu.:1.000
##  Median :10.00   Median :3.000   Median : 8.000   Median :2.000
##  Mean   :10.08   Mean   :3.612   Mean   : 7.573   Mean   :2.446
##  3rd Qu.:12.00   3rd Qu.:5.000   3rd Qu.:11.000   3rd Qu.:4.000
##  Max.   :16.00   Max.   :7.000   Max.   :15.000   Max.   :6.000
##       race            sex         capital.gain    capital.loss
##  Min.   :1.000   Min.   :1.000   Min.   :    0   Min.   :   0.0
##  1st Qu.:5.000   1st Qu.:1.000   1st Qu.:    0   1st Qu.:   0.0
##  Median :5.000   Median :2.000   Median :    0   Median :   0.0
##  Mean   :4.666   Mean   :1.669   Mean   : 1078   Mean   :  87.3
##  3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:    0   3rd Qu.:   0.0
##  Max.   :5.000   Max.   :2.000   Max.   :99999   Max.   :4356.0
##  hours.per.week  native.country    prediction
##  Min.   : 1.00   Min.   : 1.00   Min.   :1.000
##  1st Qu.:40.00   1st Qu.:40.00   1st Qu.:1.000
##  Median :40.00   Median :40.00   Median :1.000
##  Mean   :40.44   Mean   :37.72   Mean   :1.241
##  3rd Qu.:45.00   3rd Qu.:40.00   3rd Qu.:1.000
##  Max.   :99.00   Max.   :42.00   Max.   :2.000
```

```r
# Display internal structure of dataset
str(adultDataset)
```

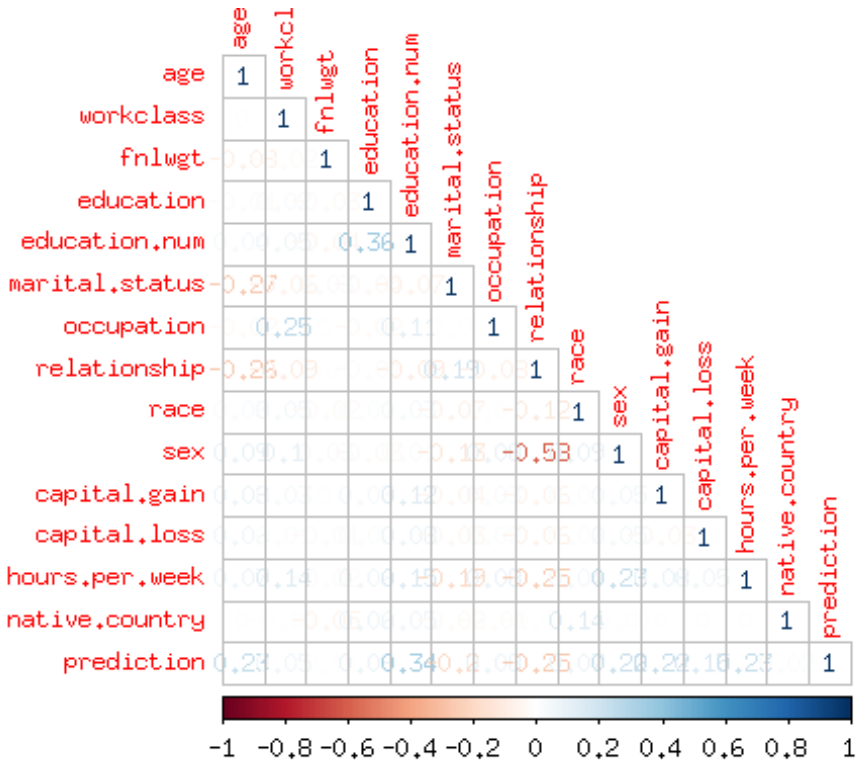
```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : int  8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187
209642 45781 159449 ...
##  $ education     : int  10 10 12 2 10 13 7 12 13 10 ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital.status: int  5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : int  2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship  : int  2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : int  5 5 5 3 3 5 3 5 5 5 ...
```

```
##  $ sex           : int  2 2 2 2 1 1 1 2 1 2 ...
##  $ 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: int  40 40 40 40 6 40 24 40 40 40 ...
##  $ prediction    : int  1 1 1 1 1 1 1 2 2 2 ...
```

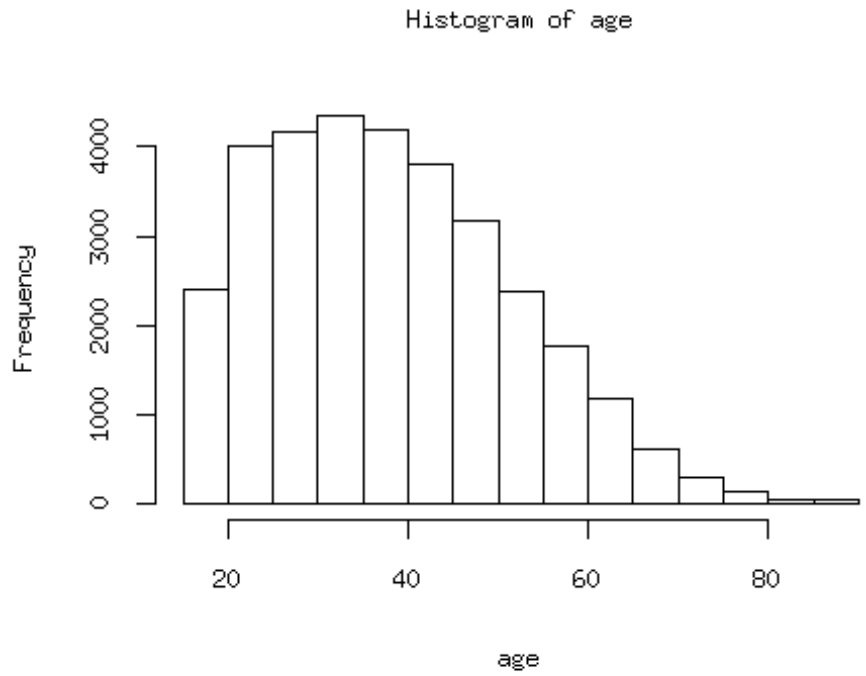
```
# Print the feature names
colnames(adultDataset)
```

```
##  [1] "age"            "workclass"      "fnlwgt"         "education"
##  [5] "education.num"  "marital.status" "occupation"     "relationship"
##  [9] "race"           "sex"            "capital.gain"   "capital.loss"
## [13] "hours.per.week" "native.country" "prediction"
```

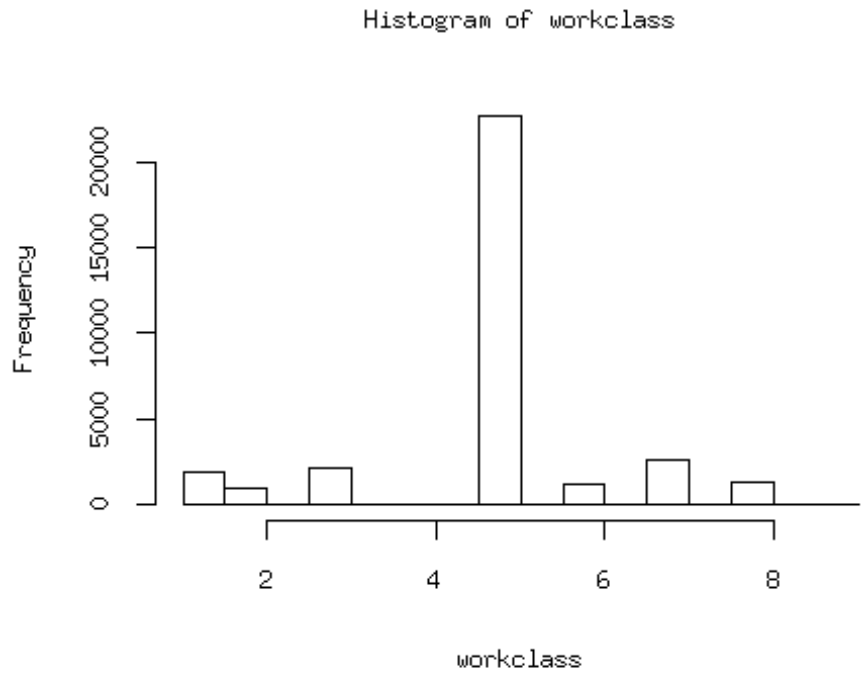
```
#
# Visualization
#

# Display the lower correlation plot of the dataset
corrplot(cor(adultDataset), method="number", type = "lower")
```



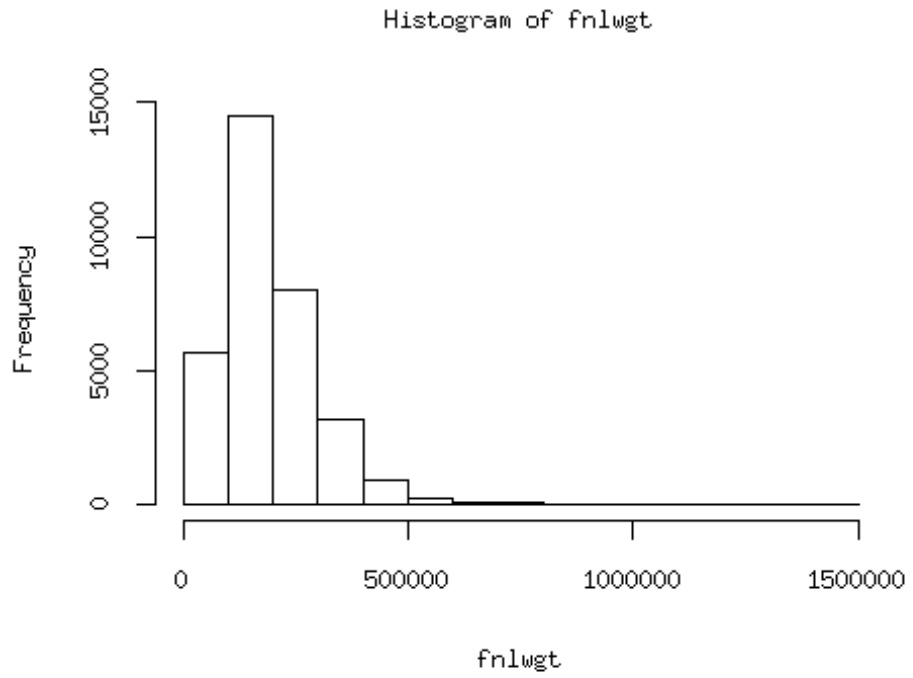

```
# Display histogram of feature "age"
hist(age)
```



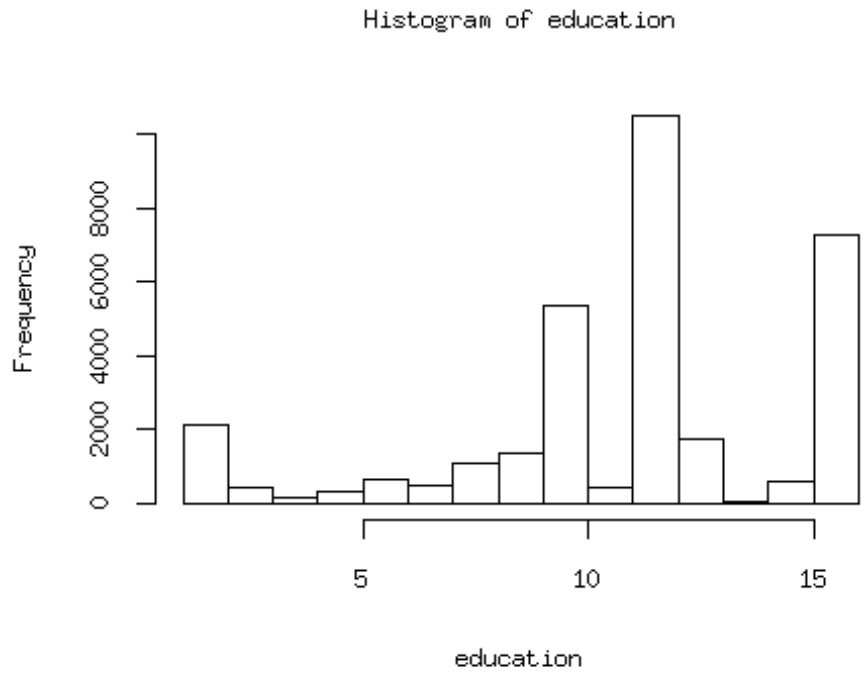

```r
# Display histogram of feature "workclass"
hist(workclass)
```



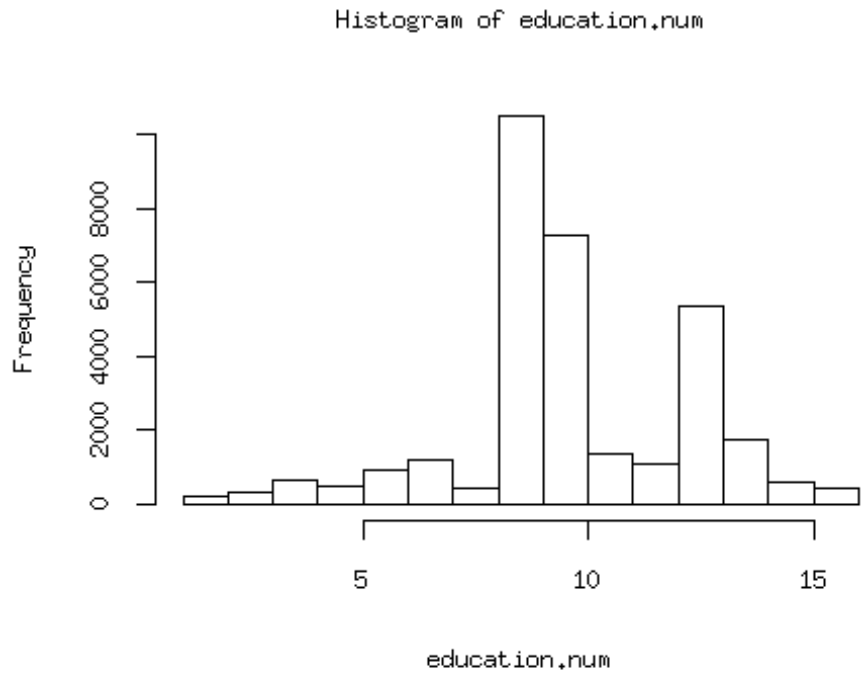
```r
# Display histogram of feature "fnlwgt"
hist(fnlwgt)
```



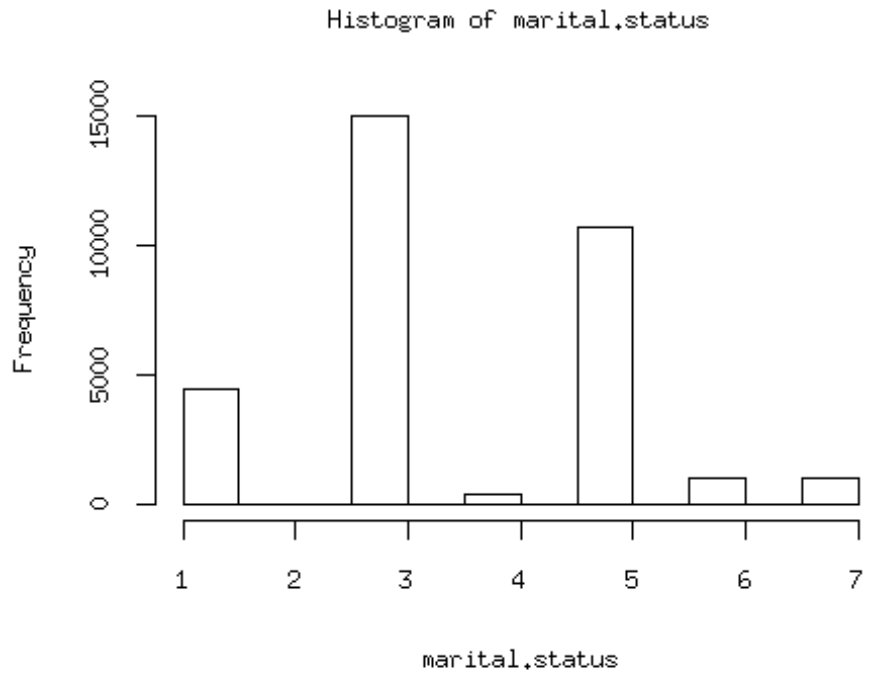

```r
# Display histogram of feature "education"
hist(education)
```



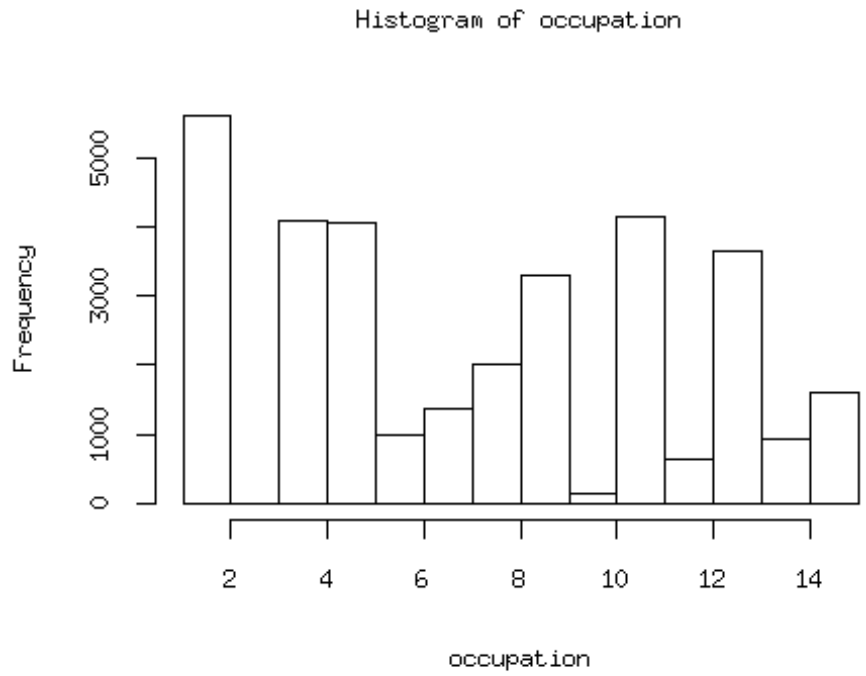

```r
# Display histogram of feature "education.num"
hist(education.num)
```



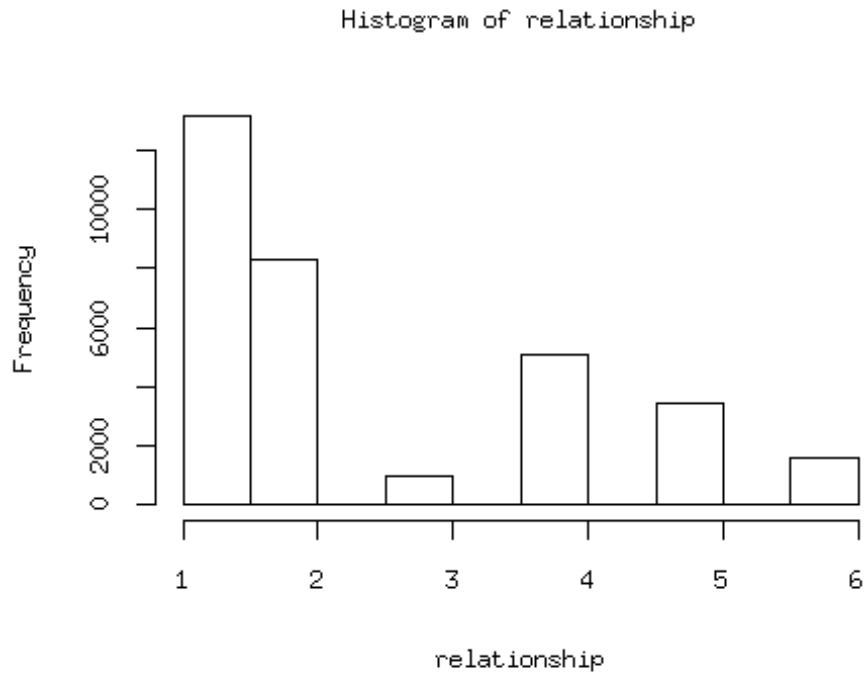
```r
# Display histogram of feature "marital.status"
hist(marital.status)
```



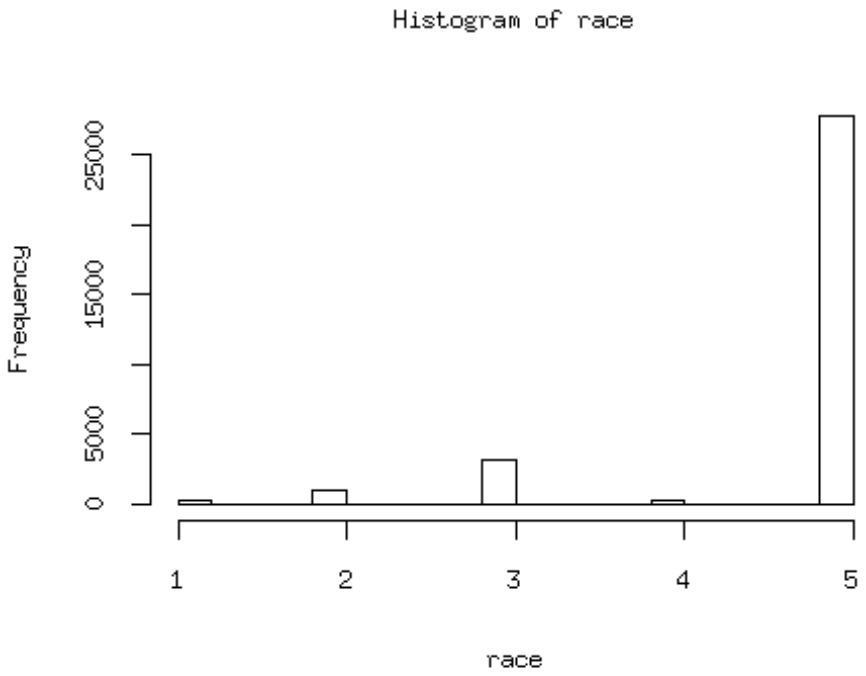

```r
# Display histogram of feature "occupation"
hist(occupation)
```



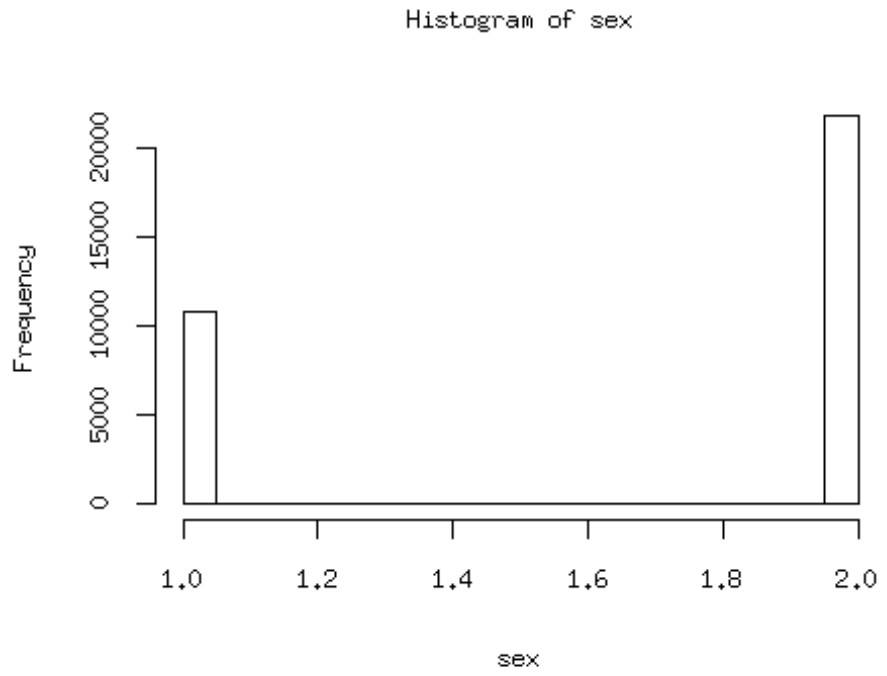

```r
# Display histogram of feature "relationship"
hist(relationship)
```



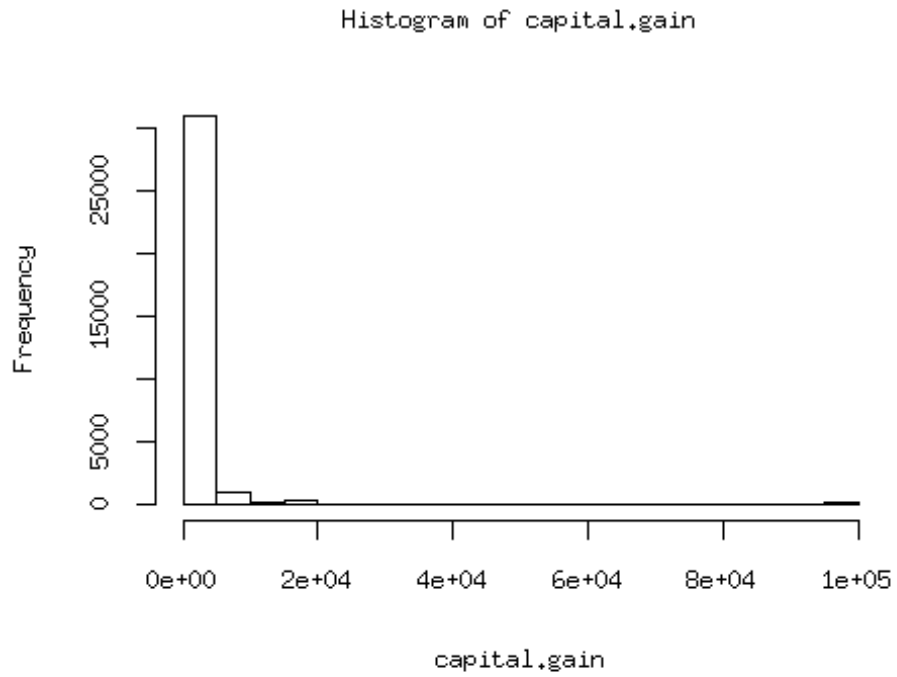
```r
# Display histogram of feature "race"
hist(race)
```



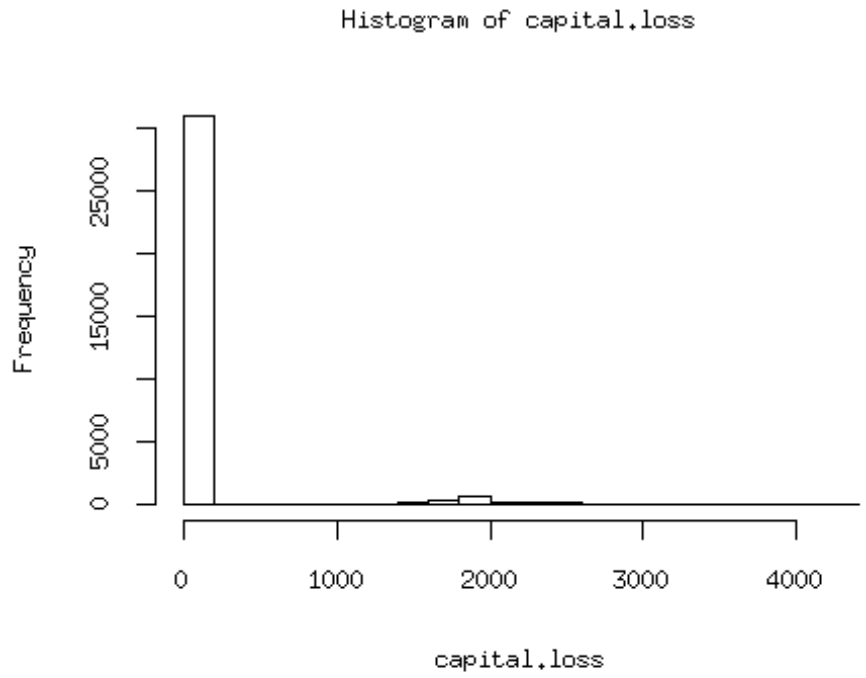

```r
# Display histogram of feature "sex"
hist(sex)
```



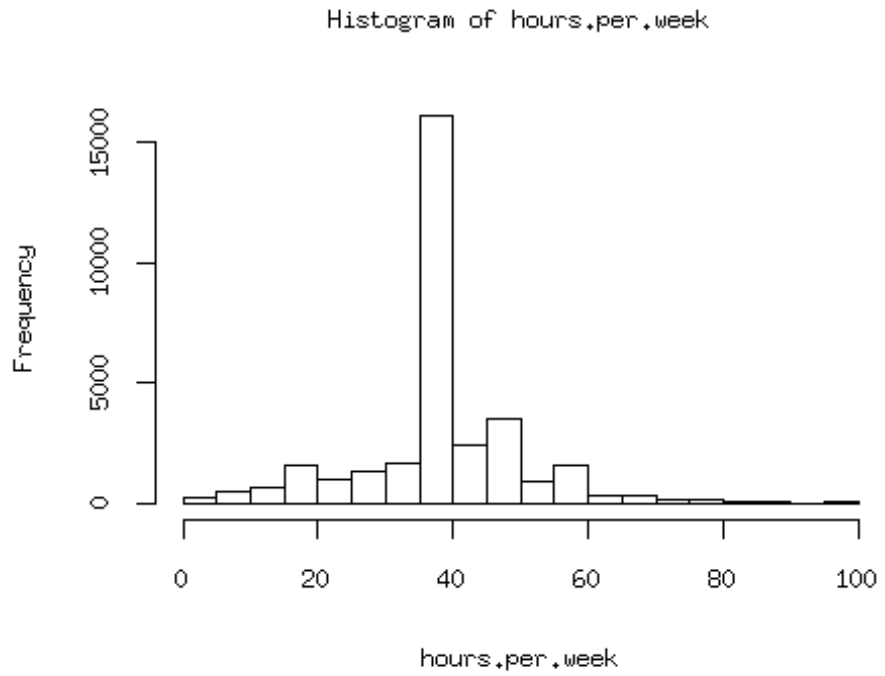

```r
# Display histogram of feature "capital.gain"
hist(capital.gain)
```



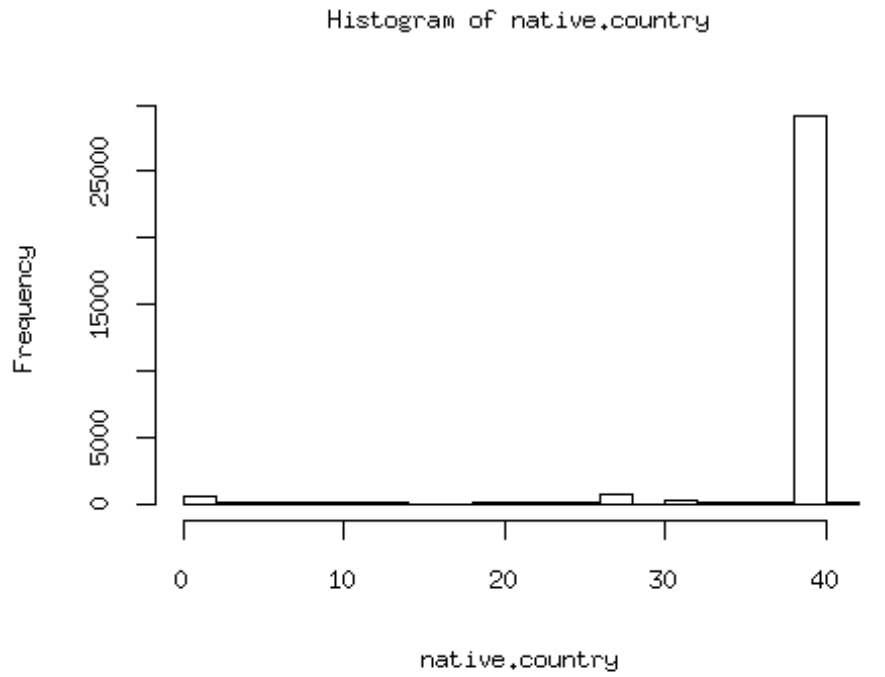
```r
# Display histogram of feature "capital.loss"
hist(capital.loss)
```




```r
# Display histogram of feature "hours.per.week"
hist(hours.per.week)
```




```r
# Display histogram of feature "native.country"
hist(native.country)
```



```r
# Display histogram of feature "prediction"
hist(prediction)
```