initialAnalysis.R

ketankokane

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 hist
#
# 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
dataset = read.table("adult.data", header= TRUE, sep = ",")
# Print the feature names
colnames(dataset)
```

```
##  [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"
```

```
# Dimensions of the raw data
dim(dataset)
```

```
## [1] 32561    15
```

```
# Convert the dataset to integer format
#dataset[] <- lapply(dataset,as integer)

# Removing null values
#na.omit(dataset)
#Checking dimensions after getting rid of null values
dim(dataset)
```

```
## [1] 32561    15
```

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

#
# Printing details of the dataset
#

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

```
##       age                     workclass         fnlwgt
##  Min.   :17.00    Private         :22696   Min.   :  12285
##  1st Qu.:28.00    Self-emp-not-inc: 2541   1st Qu.: 117827
##  Median :37.00    Local-gov       : 2093   Median : 178356
##  Mean   :38.58    ?               : 1836   Mean   : 189778
##  3rd Qu.:48.00    State-gov       : 1298   3rd Qu.: 237051
##  Max.   :90.00    Self-emp-inc    : 1116   Max.   :1484705
##                  (Other)          :  981
##          education     education.num                  marital.status
##   HS-grad     :10501   Min.   : 1.00    Divorced             : 4443
##   Some-college: 7291   1st Qu.: 9.00    Married-AF-spouse    :   23
##   Bachelors   : 5355   Median :10.00    Married-civ-spouse   :14976
##   Masters     : 1723   Mean   :10.08    Married-spouse-absent:  418
##   Assoc-voc   : 1382   3rd Qu.:12.00    Never-married        :10683
##   11th        : 1175   Max.   :16.00    Separated            : 1025
##  (Other)      : 5134                    Widowed              :  993
##             occupation           relationship
##   Prof-specialty :4140    Husband       :13193
##   Craft-repair   :4099    Not-in-family : 8305
##   Exec-managerial:4066    Other-relative:  981
##   Adm-clerical   :3770    Own-child     : 5068
##   Sales          :3650    Unmarried     : 3446
##   Other-service  :3295    Wife          : 1568
##  (Other)         :9541
##                    race            sex         capital.gain
##   Amer-Indian-Eskimo:  311    Female:10771   Min.   :    0
##   Asian-Pac-Islander: 1039    Male  :21790   1st Qu.:    0
##   Black             : 3124                   Median :    0
##   Other             :  271                   Mean   : 1078
##   White             :27816                   3rd Qu.:    0
##                                              Max.   :99999
##
##   capital.loss    hours.per.week         native.country   prediction
##  Min.   :   0.0   Min.   : 1.00    United-States:29170    <=50K:24720
##  1st Qu.:   0.0   1st Qu.:40.00    Mexico       :  643    >50K : 7841
##  Median :   0.0   Median :40.00    ?            :  583
##  Mean   :  87.3   Mean   :40.44    Philippines  :  198
##  3rd Qu.:   0.0   3rd Qu.:45.00    Germany      :  137
##  Max.   :4356.0   Max.   :99.00    Canada       :  121
##                                   (Other)      : 1709
```

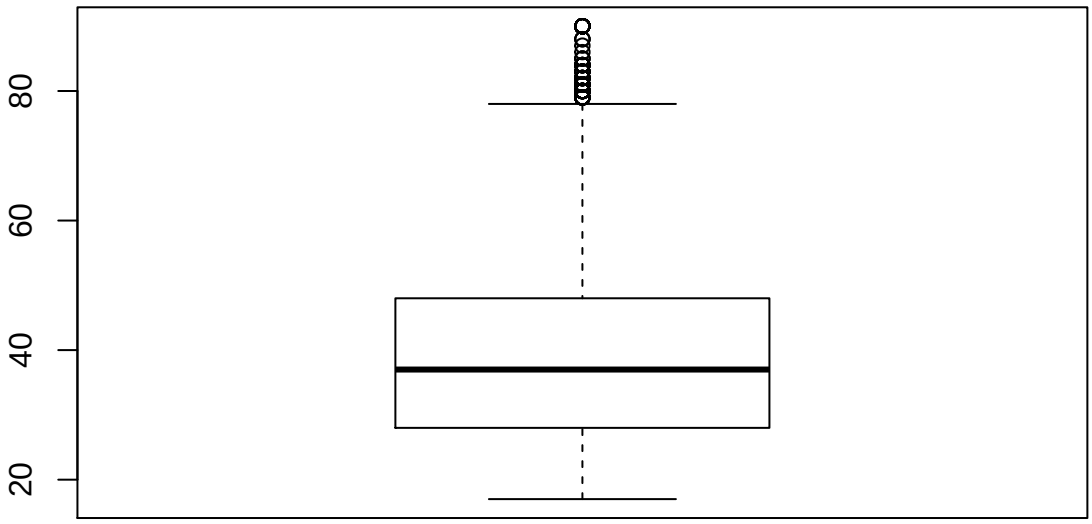
```
# Display internal structure of dataset
str(dataset)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 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: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 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: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ prediction    : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
```

```
#
# Visualization
#
# Our dataset inclused people ranging from 17-90years of age.
summary(age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.58   48.00   90.00
```

```
boxplot(age)
```




```
# Display histogram of feature "age" . Our dataset is concentrated
# in the 28-38(first quartile-second quartile) year range i.e which is expected as that would
# categorize the working age group
## Frequency table
counts <- table(age)
counts
```

```
## age
##  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
## 395 550 712 753 720 765 877 798 841 785 835 867 813 861 888 828 875 886
##  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52
```

```
## 876 898 858 827 816 794 808 780 770 724 734 737 708 543 577 602 595 478
##  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70
## 464 415 419 366 358 366 355 312 300 258 230 208 178 150 151 120 108  89
##  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88
##  72  67  64  51  45  46  29  23  22  22  20  12   6  10   3   1   1   3
##  90
##  43
```

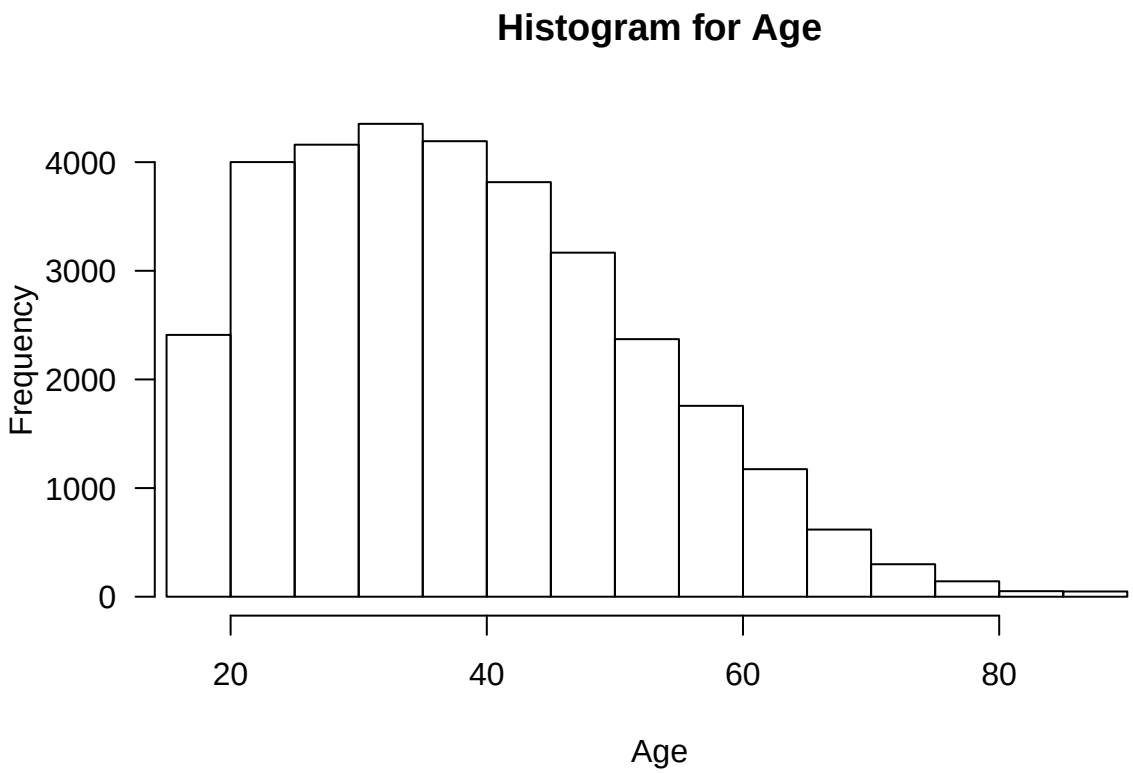
```
## The most frequent and least frequent values.
# Most frequently occuring value is of the 36year olds.
# Least frequent values for age 86 and 87.
counts[which.max(counts)]
```

```
##  36
## 898
```

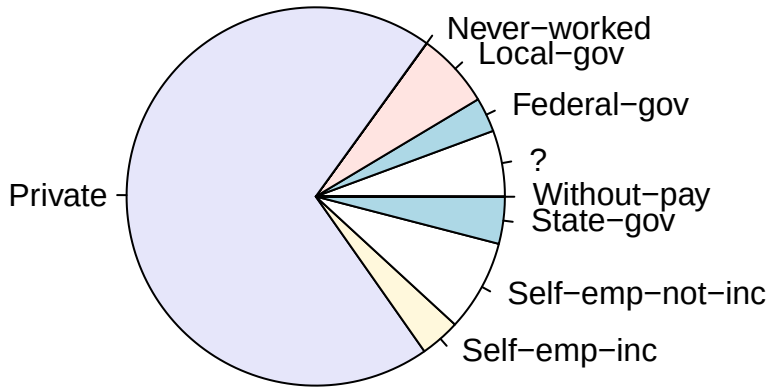
```
counts[which.min(counts)]
```

```
## 86
##  1
```

```
hist(age,main="Histogram for Age",xlab="Age", xlim=c(17,90),las=1,
     breaks=20)
```



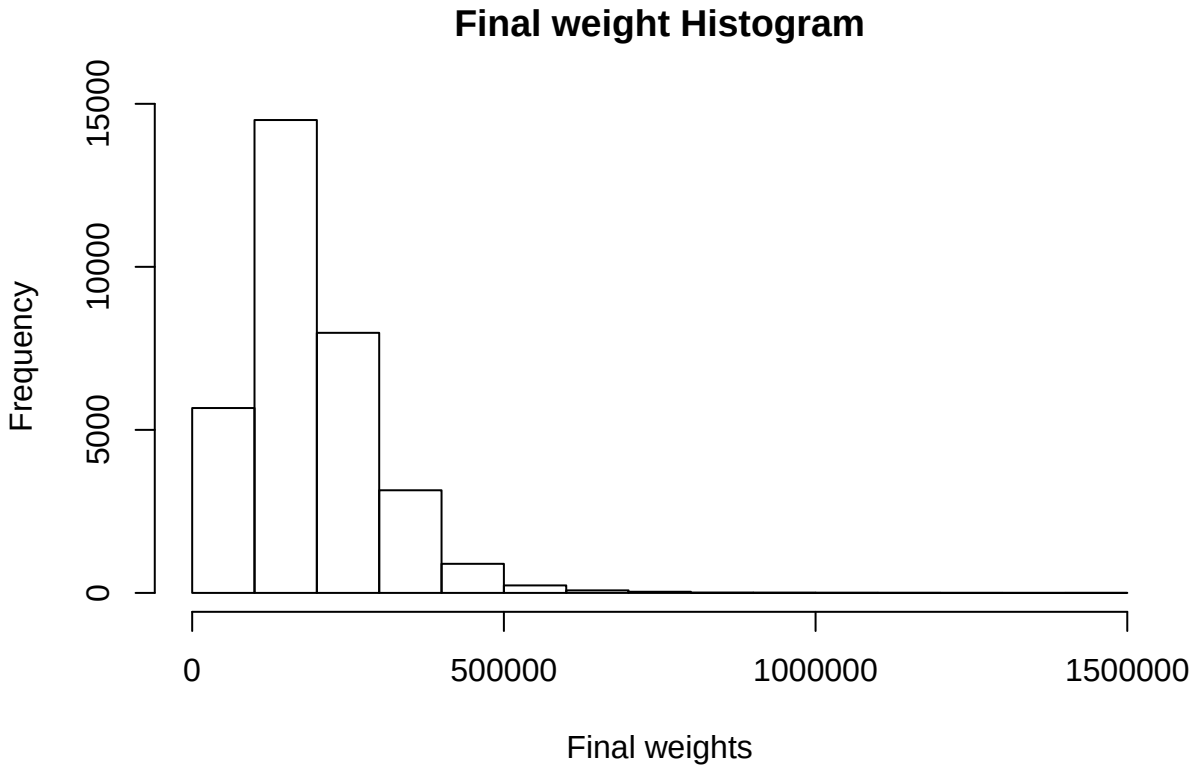

```
# Display histogram of feature "workclass". Majority of the dataset
# are employed in the private sector
pie(table(workclass))
```



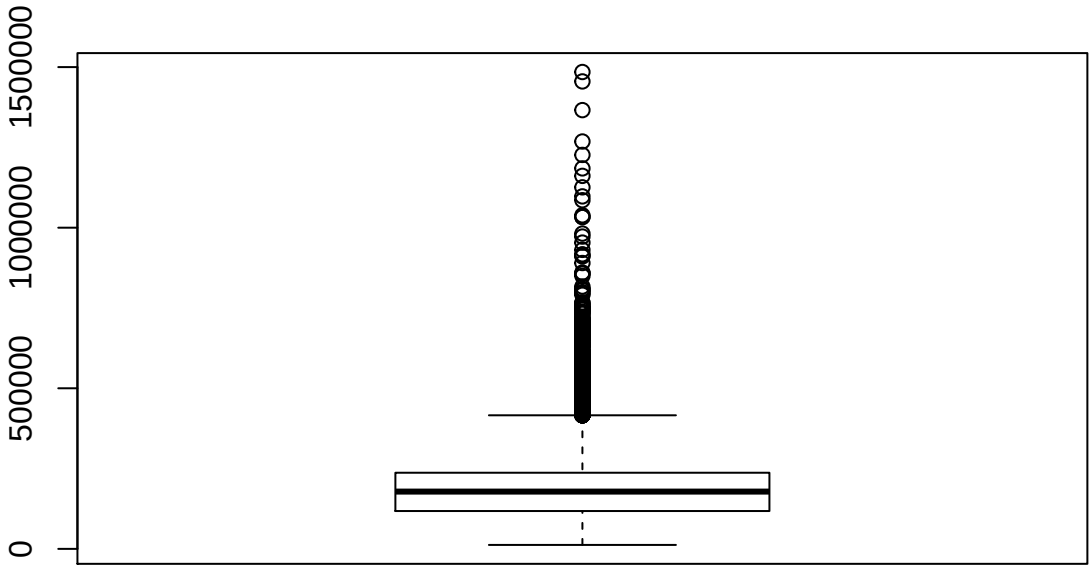

```
cor(as.numeric(workclass),age)
```

```
## [1] 0.003787353
```

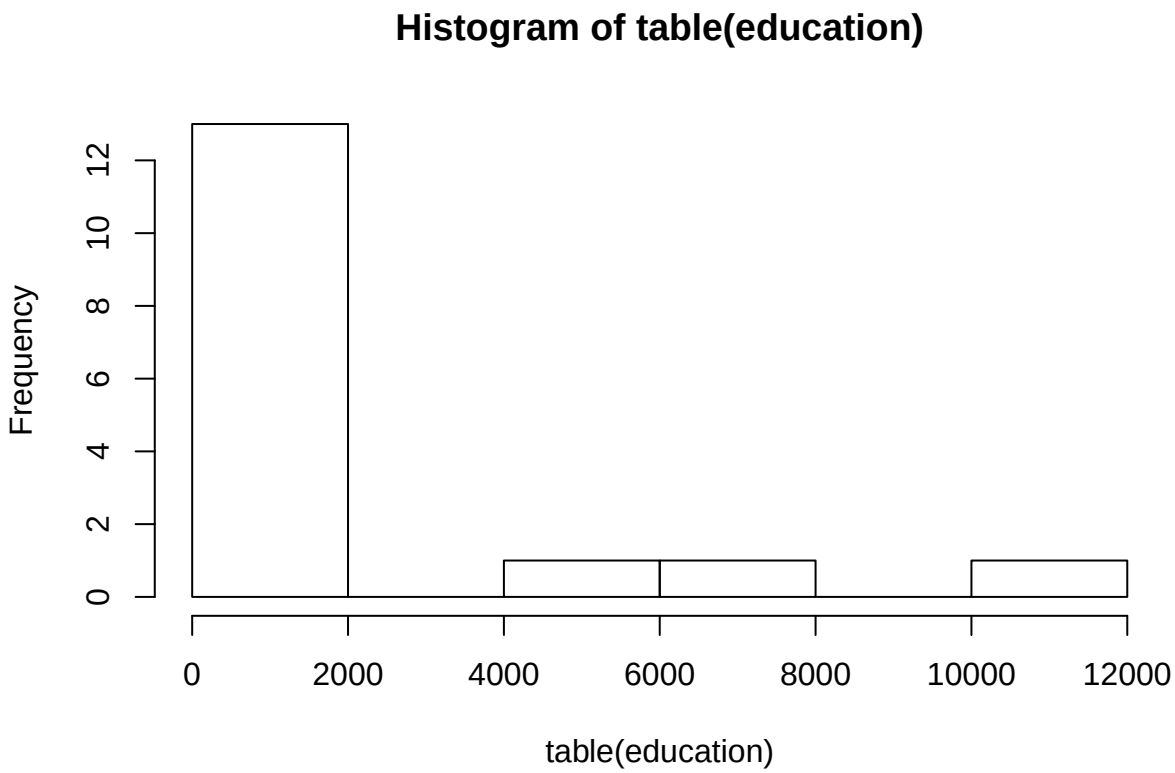
```
#hist(as.factor(workclass)~age)
# Display histogram of feature "fnlwgt".
hist(fnlwgt, main = "Final weight Histogram", xlab = "Final weights")
```



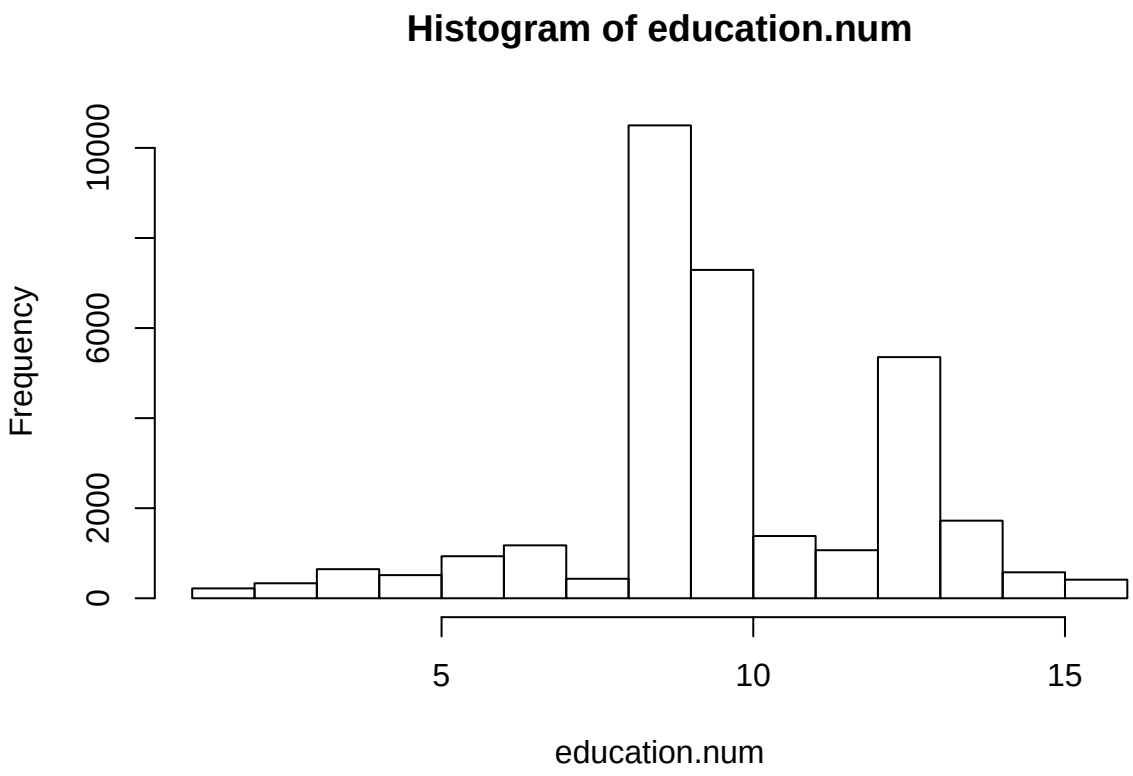

```
boxplot(fnlwgt)
```



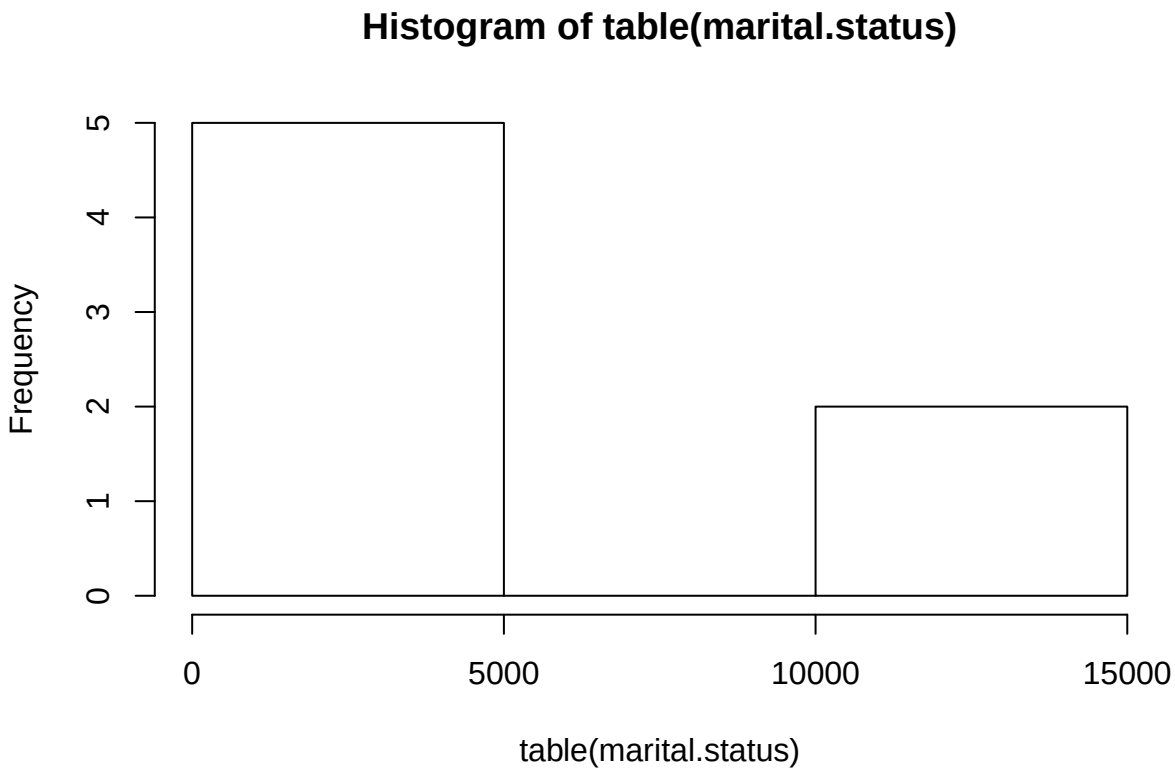

```
#Final weight attribute consists of
# continuous values. On analysing the corrplot, final weight doesn't seem to be
# correlated to any of the other values. Hence no functional dependencies.
# fnlwgt doesn't seem very relevant in this datset. And so we might choose to drop
# this attribute.

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




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

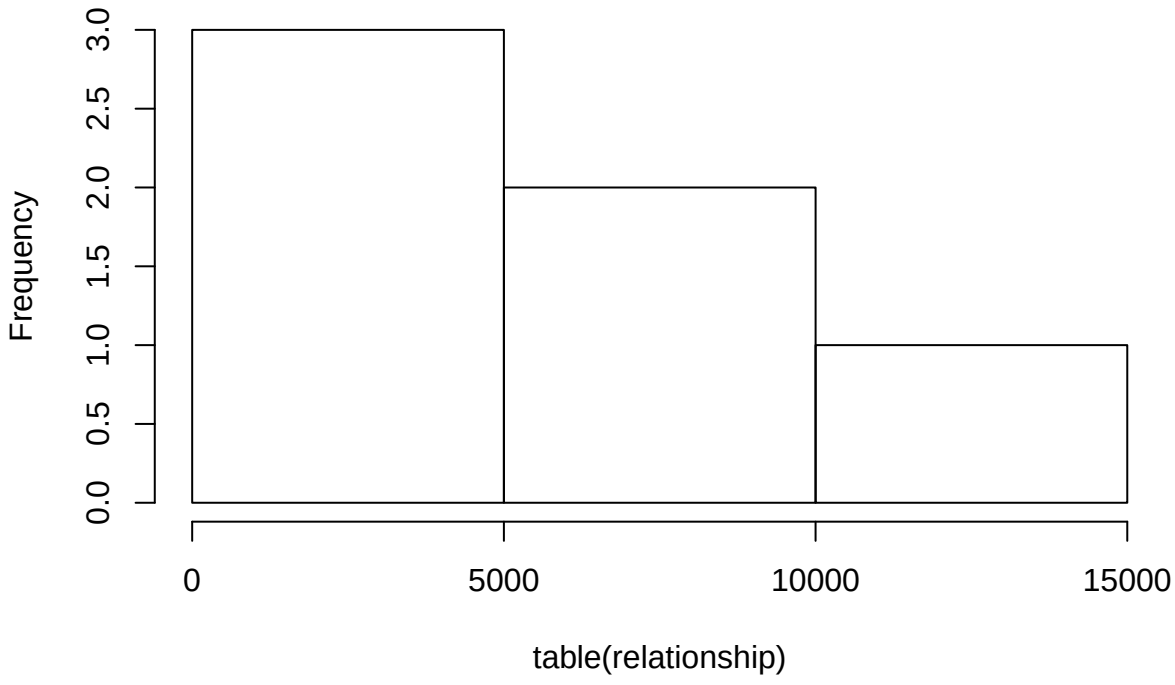



```r
# Display histogram of feature "marital.status"
hist(table(marital.status))
```

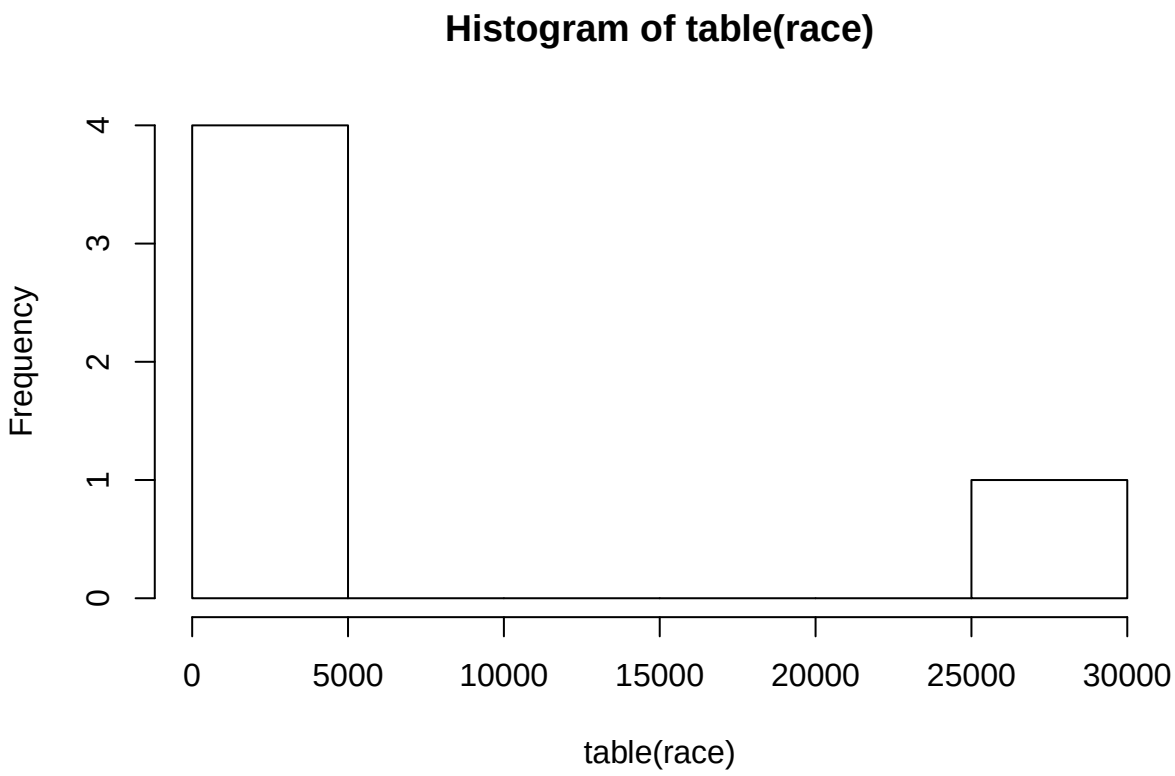



```r
# Display histogram of feature "occupation"
hist(table(occupation))
```

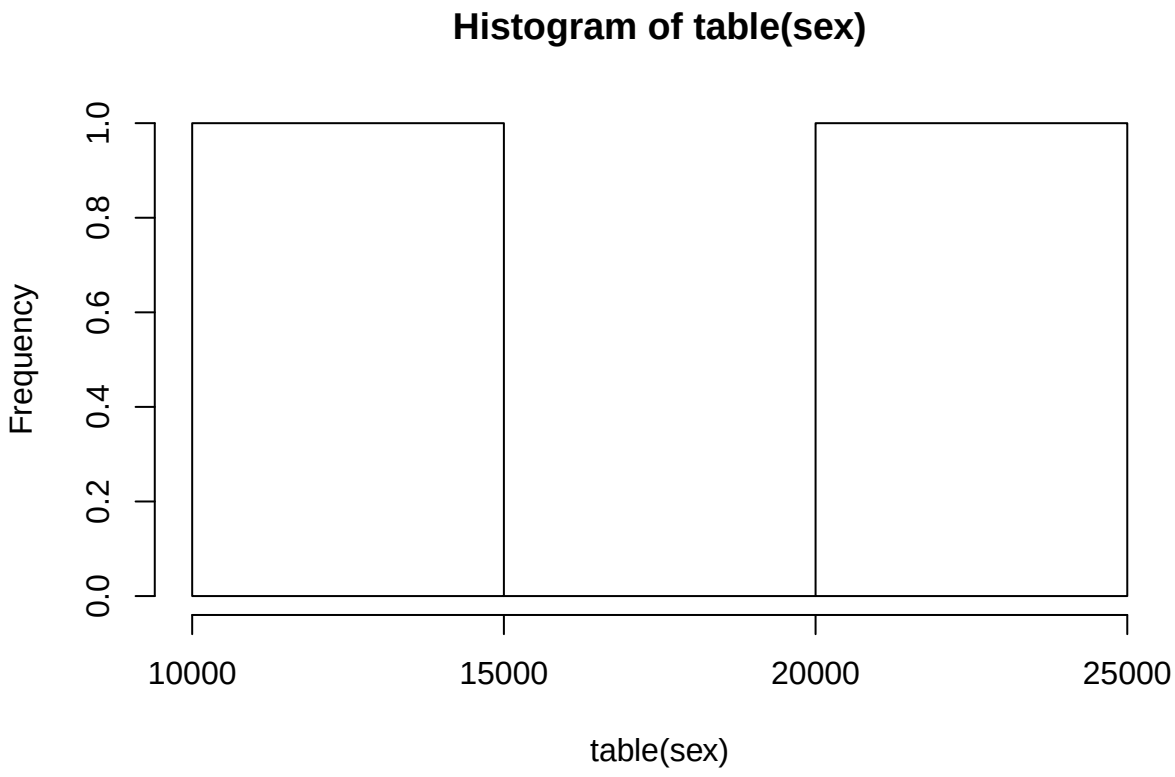

```r
# Display histogram of feature "relationship"
hist(table(relationship))
```



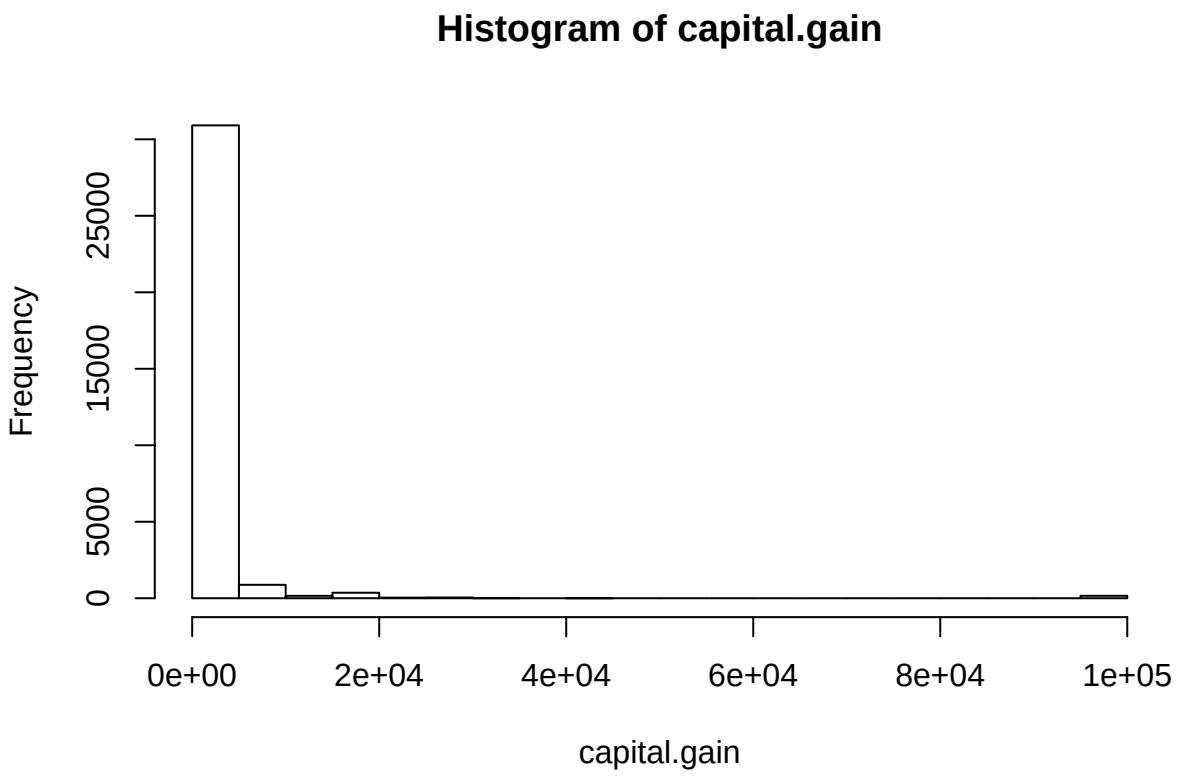

```r
# Display histogram of feature "race"
hist(table(race))
```



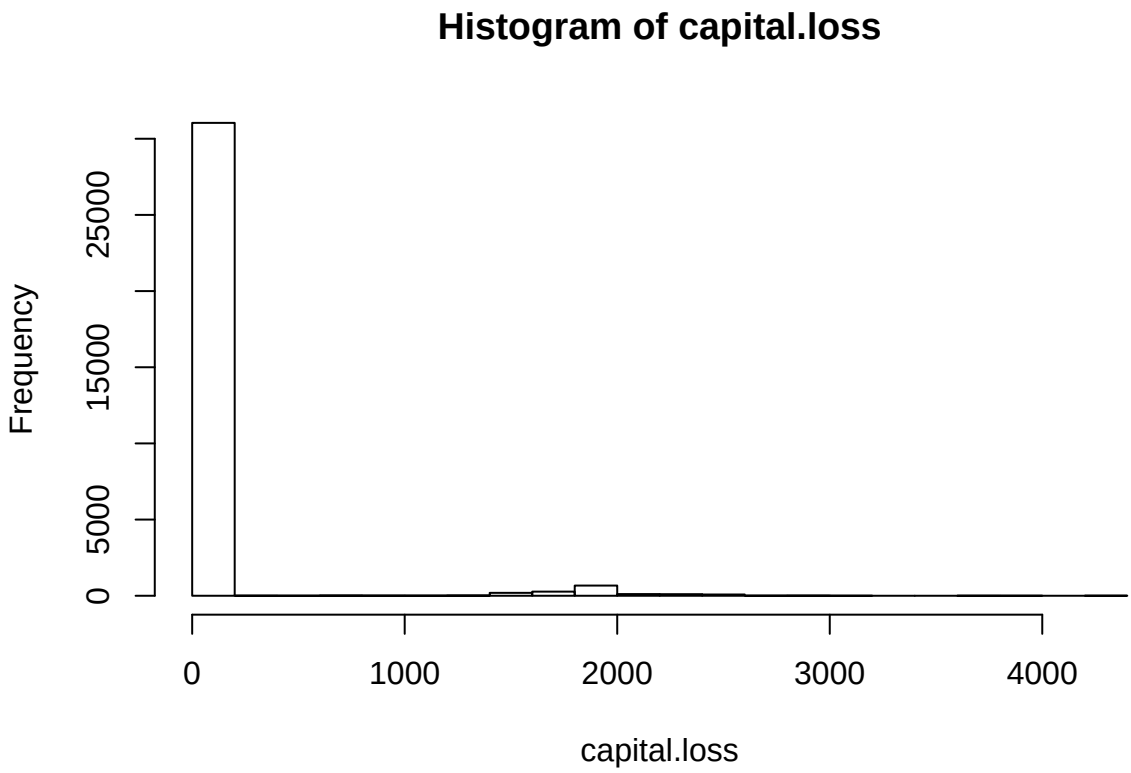

```r
# Display histogram of feature "sex"
hist(table(sex))
```



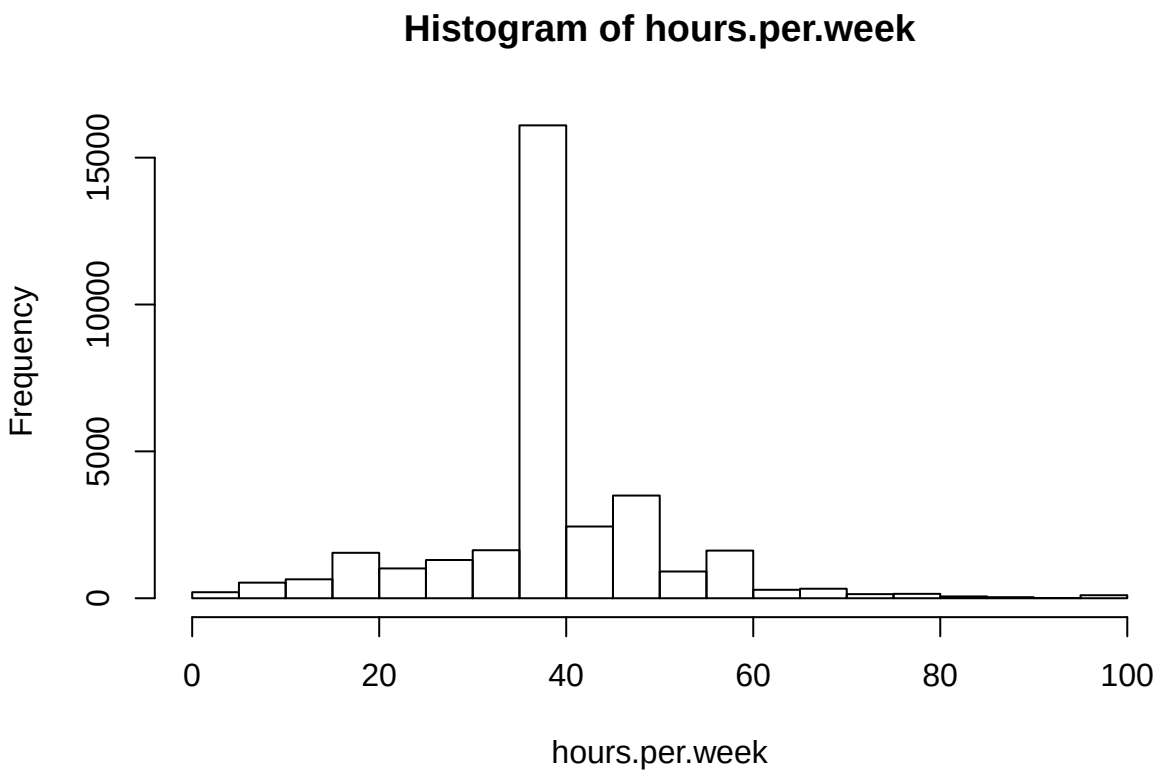

```r
# Display histogram of feature "capital.gain"
hist(capital.gain)
```



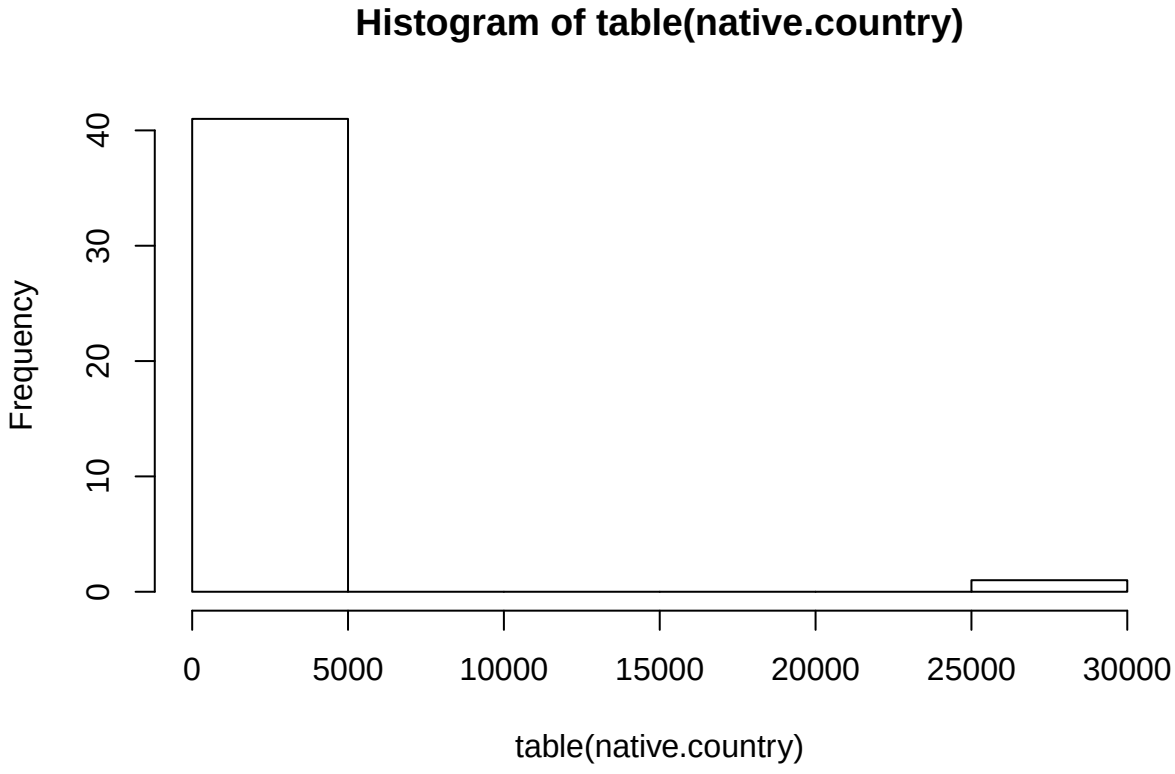

```r
# Display histogram of feature "capital.loss"
hist(capital.loss)
```



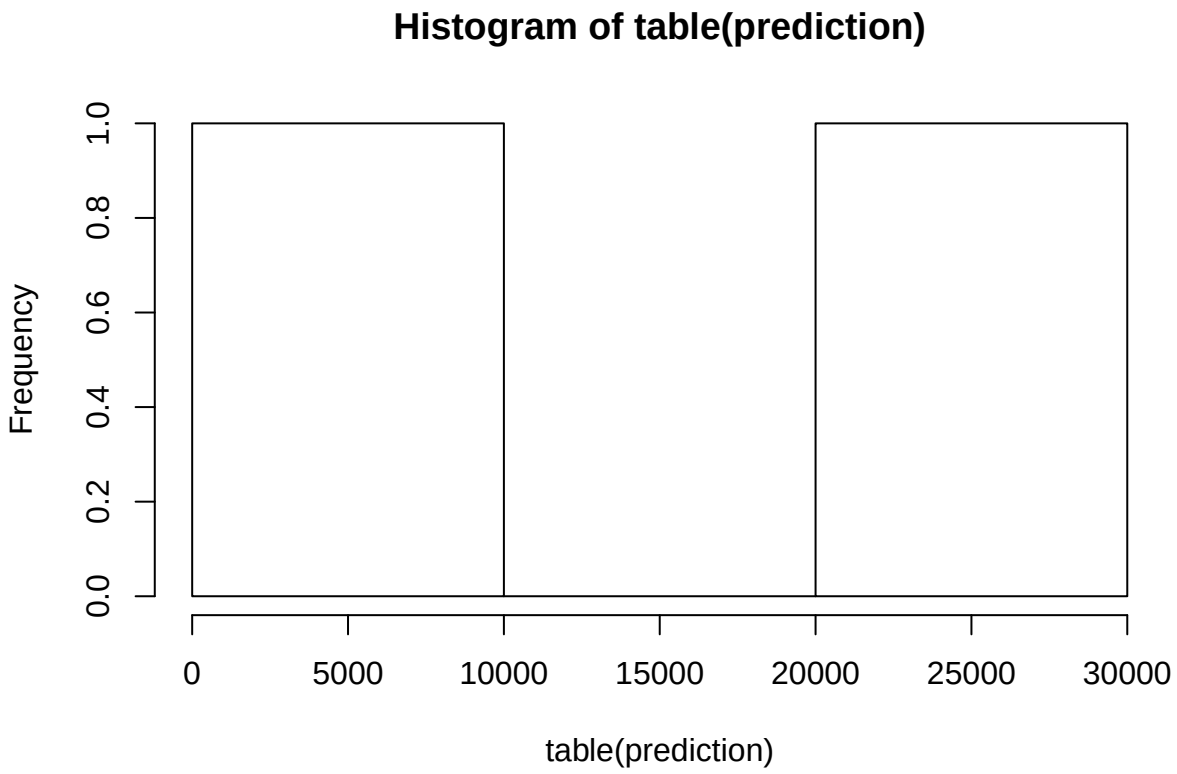

```r
# Display histogram of feature "hours.per.week"
hist(hours.per.week)
```




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




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




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