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,
# visualizes the data by plotting histograms of each feature
# Finds and states outlier of every attribute

#
# Analyzing the dataset
#

# 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
```

```
# 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, which tells what are the different values of every attribute a
#levels
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 includes people ranging from 17-90 years of age which seems appropriate in census dataset
summary(age)
```

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

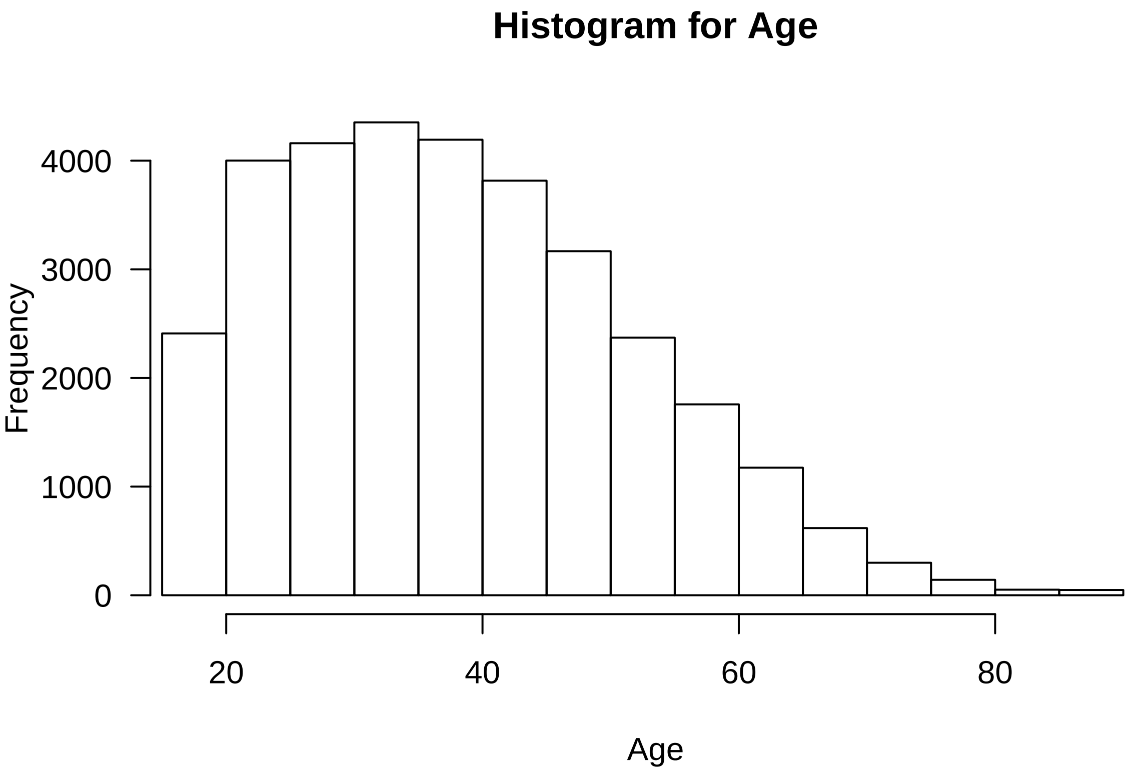
```
# Display histogram of feature "age" . Our dataset is concentrated
# in the 28-38 (first quartile~second quartile) year range which is expected as that would
# categorize the working age group
## Frequency table
counts <- table(age)
## The most frequent and least frequent values.
# Most frequently occuring value is of the 36year olds.
# Least frequent values for age 86 and 87.
counts[which.max(counts)]
```

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

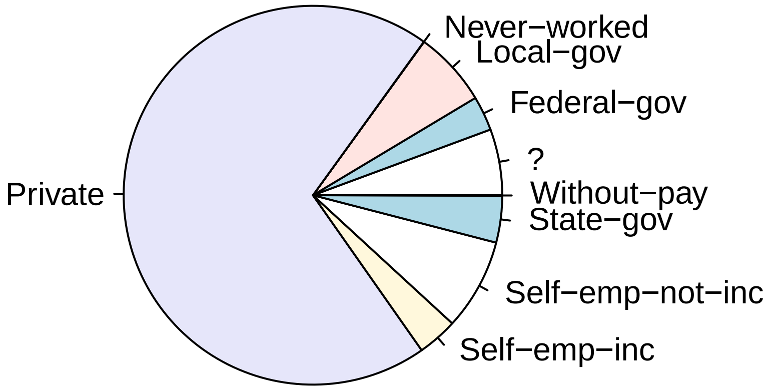
```
counts[which.min(counts)]
```

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

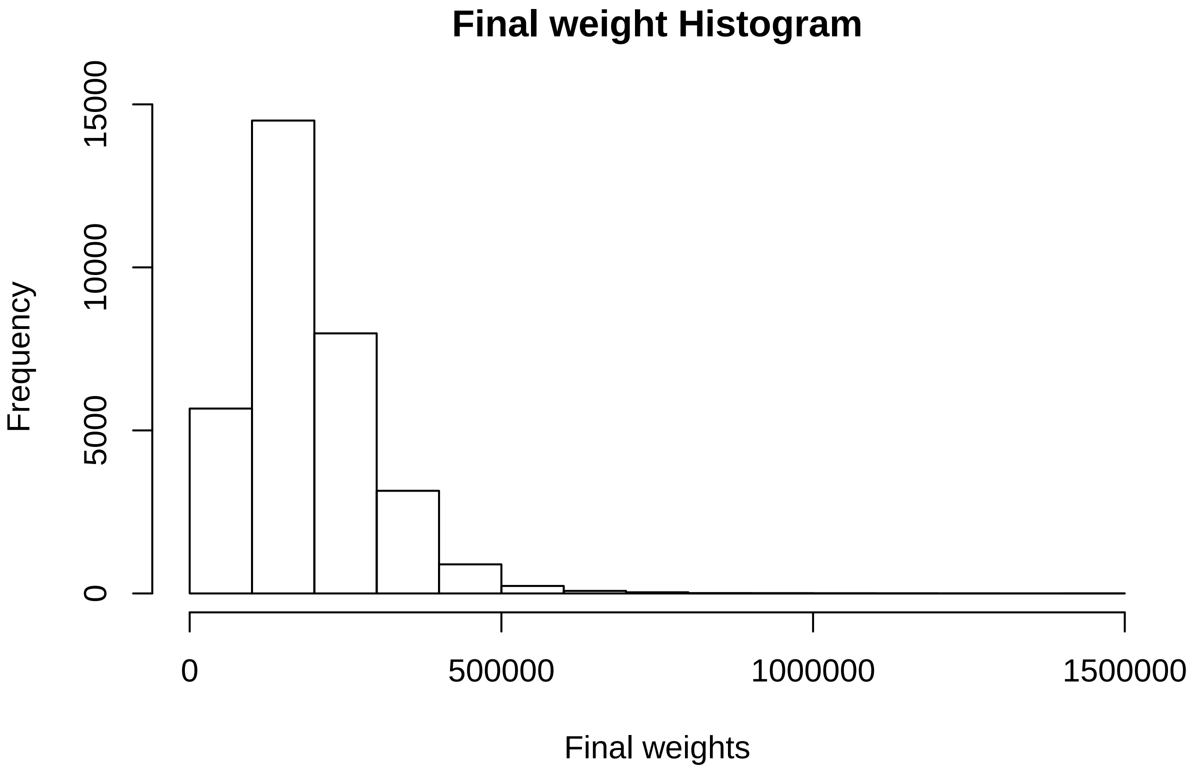
```
hist(age,main="Histogram for Age",xlab="Age", xlim=c(17,90),las=1,
     breaks=20)
```



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



```r
# Display histogram of feature "fnlwgt".
hist(fnlwgt, main = "Final weight Histogram", xlab = "Final weights")
```



```r
#Final weight attribute consists of
# continuous values. final weight doesn't seem to be
# correlated to any of the other values.
# fnlwgt doesn't seem very relevant in this datset. And so we might choose to drop
# this attribute.

#
# Display table of feature "education"
educationTable <-data.frame(count=sort(table(education), decreasing=TRUE))
educationTable
```

```
##    count.education count.Freq
## 1          HS-grad      10501
## 2     Some-college       7291
## 3        Bachelors       5355
## 4          Masters       1723
## 5         Assoc-voc      1382
## 6              11th      1175
## 7        Assoc-acdm      1067
## 8              10th       933
## 9           7th-8th       646
## 10      Prof-school       576
## 11              9th       514
## 12             12th       433
## 13        Doctorate       413
## 14          5th-6th       333
## 15          1st-4th       168
## 16        Preschool        51
```

```
#We have a hypothesis that the higher the education, the higher the income. We would emphasise this usi
under20yearsAge <- dataset[ which(age<20), ]
dim(under20yearsAge)
```

```
## [1] 1657   15
```

```
table(under20yearsAge$education)
```

```
##
##          10th         11th         12th      1st-4th      5th-6th
##           192          391          126            3            7
##       7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
##            17           39            1            3            2
##     Doctorate      HS-grad      Masters    Preschool  Prof-school
##             0          426            1            1            0
##  Some-college
##           448
```

```
#demonstrates the education qualification frequency of people under the age of 20

# Display table of feature "education.num"
summary(education.num)
```

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

```
table(education.num)
```

```
## education.num
##     1     2     3     4     5     6     7     8     9    10    11    12
##    51   168   333   646   514   933  1175   433 10501  7291  1382  1067
##    13    14    15    16
##  5355  1723   576   413
```

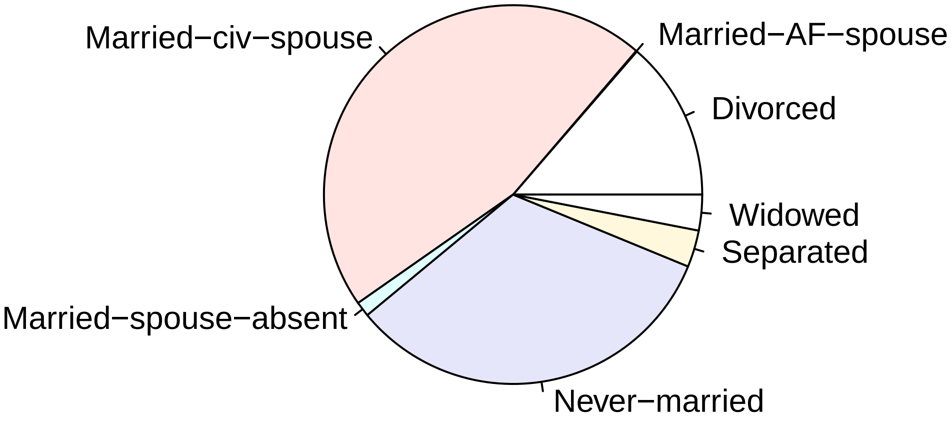
```
dim(educationTable)
```

```
## [1] 16  2
```

```
#the quantity education.num ranges from 1 to 16. Majority values concentrated between 9 and 12.
# Number of distinct values for education attribute is 16. There seems to be some correlation between t
# education.num seems to be certain measure of the education attribute

# Display pie chart of feature "marital.status". Majority of our dataset fall under the
# Married-civ-spouse or the never married category
```

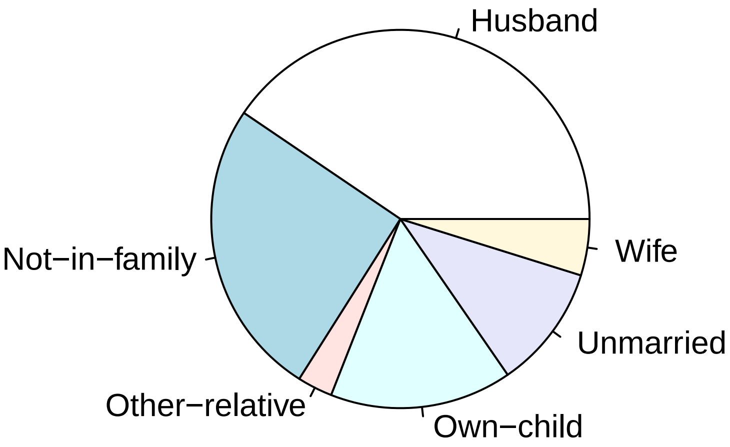
```r
pie(table(marital.status))
```



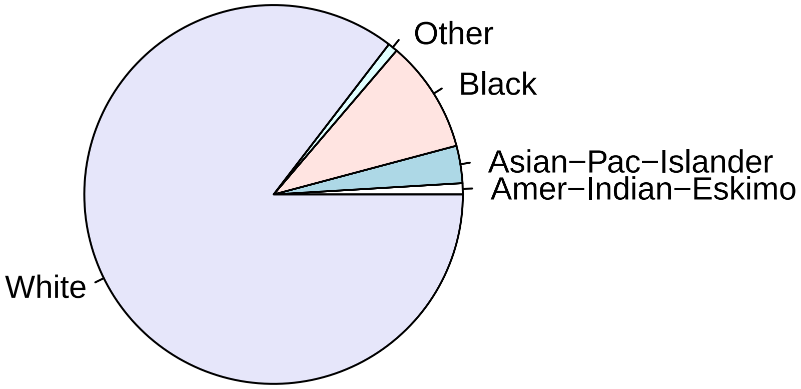
```r
# Display feature "occupation". "?" represent null values
occupationTable <-data.frame(count=sort(table(occupation), decreasing=TRUE))
occupationTable
```

```
##      count.occupation count.Freq
## 1      Prof-specialty       4140
## 2        Craft-repair       4099
## 3     Exec-managerial       4066
## 4        Adm-clerical       3770
## 5               Sales       3650
## 6       Other-service       3295
## 7   Machine-op-inspct       2002
## 8                   ?       1843
## 9    Transport-moving       1597
## 10  Handlers-cleaners       1370
## 11    Farming-fishing        994
## 12       Tech-support        928
## 13    Protective-serv        649
## 14    Priv-house-serv        149
## 15       Armed-Forces          9
```

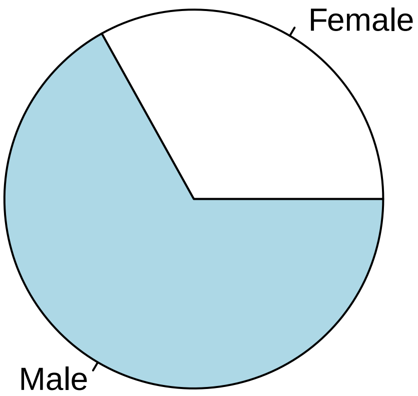
```r
# Display pie chart of feature "relationship"
pie(table(relationship))
```



```
# Display pie chart of feature "race". More than 75% of the dataset are white people. This column would
pie(table(race))
```



```
# Display plot of feature "sex". Almost 3/4th of the dataset are male
pie(table(sex))
```

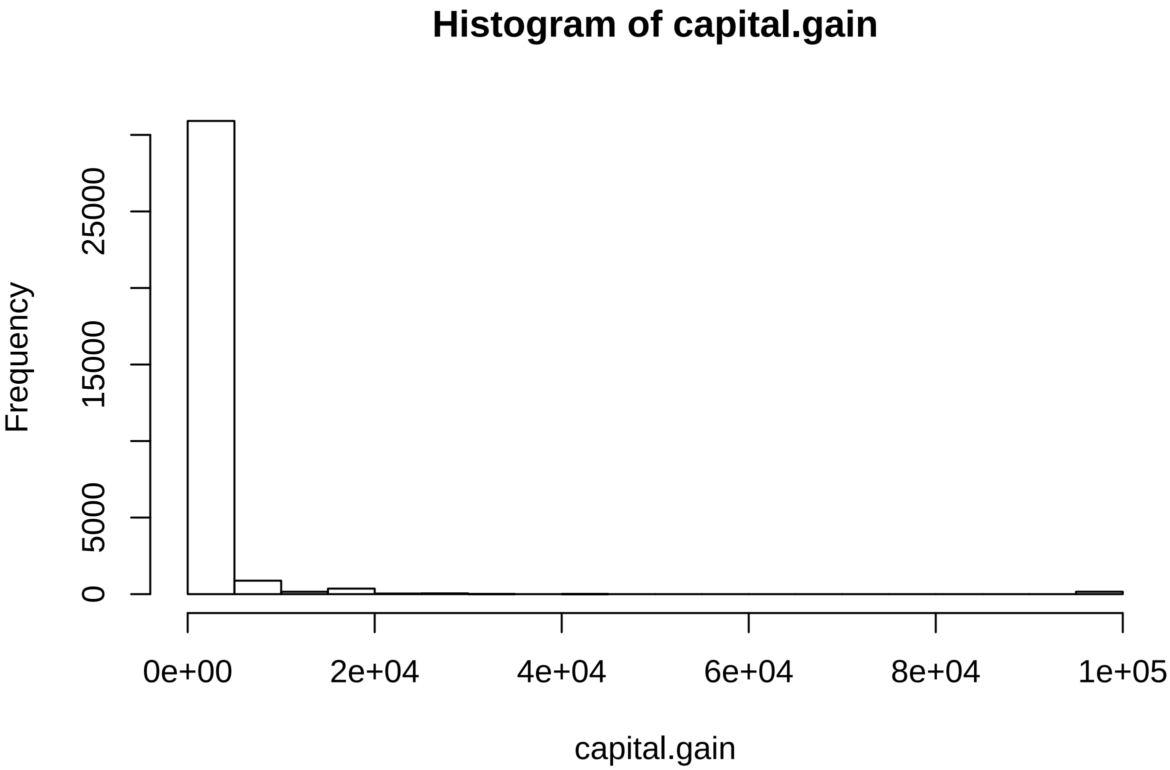


```
husbandData <- dataset[ which(sex == " Female" & relationship==" Husband"), ]
dim(husbandData)
```

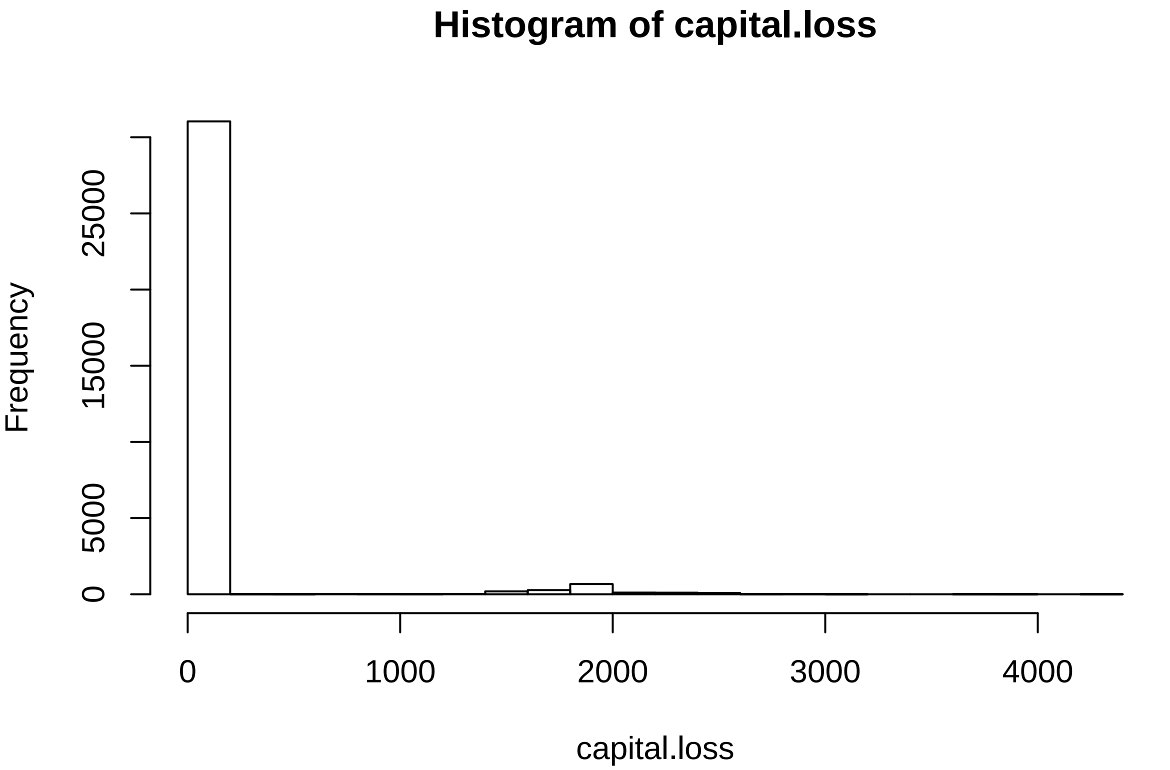
```
## [1]  1 15
```

```
#noisy data like the above state that an entry with relationship as Husband, has sex as Female exists.
#data need to be identified

# Display histogram of feature "capital.gain".Most values have value zero. Hence the column will be dro
hist(capital.gain)
```

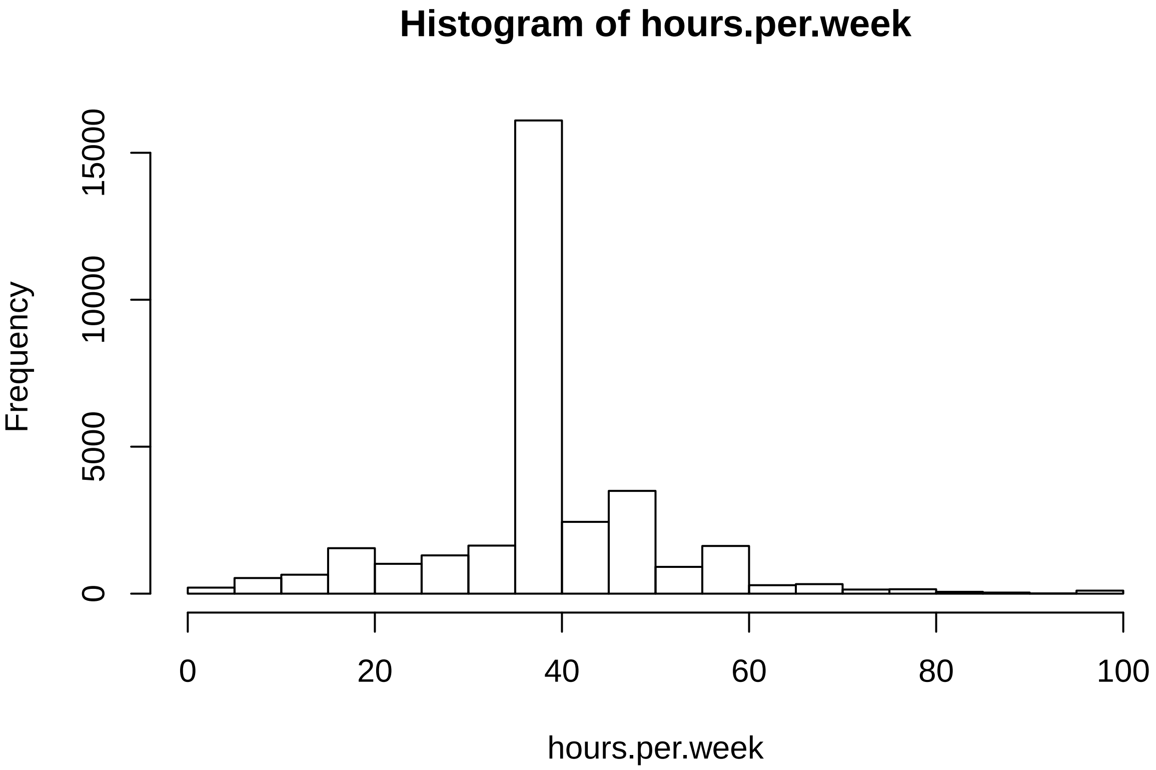


```
# Display histogram of feature "capital.loss".Most values have value zero. Hence the column will be dro
hist(capital.loss)
```

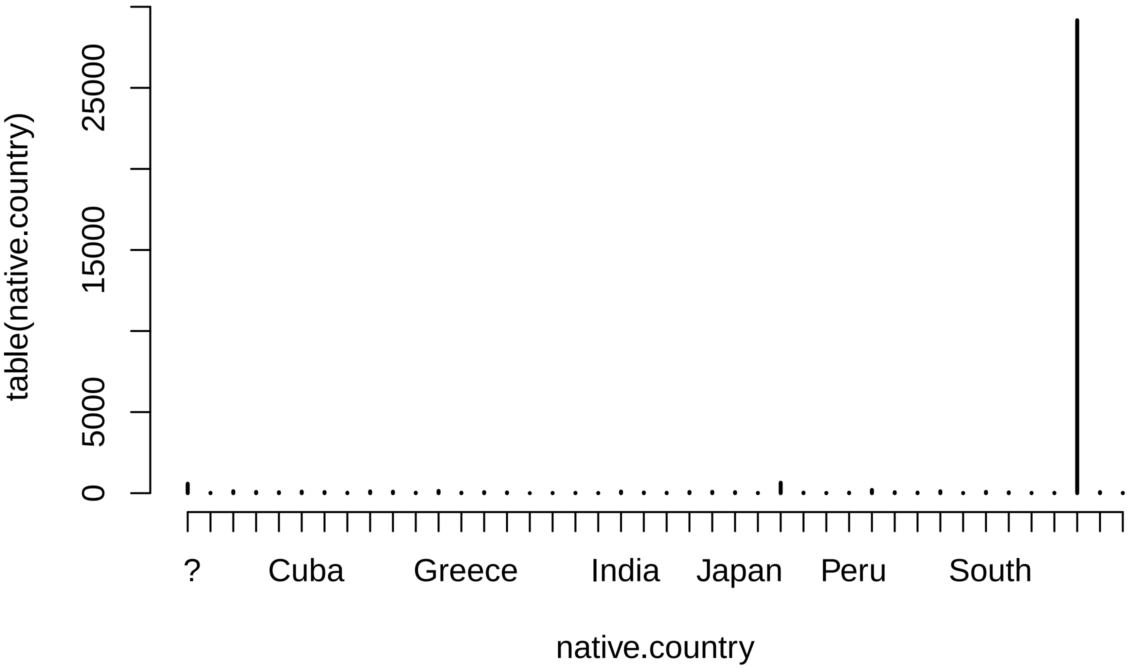


```
# Display histogram of feature "hours.per.week". As the working class is expected to work 40 hours a we
# appropriate
```

```r
hist(hours.per.week)
```



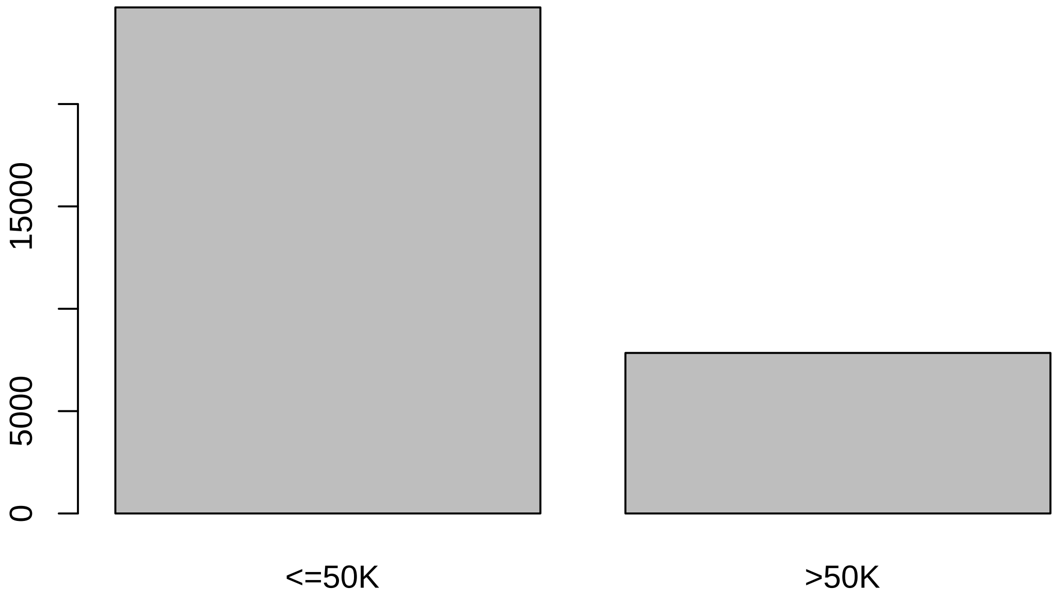
```r
# Display plot of feature "native.country". The dataset consists of values from people in the
#United States. Thus this column would be dropped
plot(table(native.country))
```



```r
countries<-table(native.country)
countries[which.max(countries)]
```

```
## United-States
##         29170
```

```
# Display plot of feature "prediction"
plot(prediction)
```


0
5000
15000
<=50K
>50K