**CS6220: Data Mining Homework 01**

**Due: Feb. 17th**

**Name: Mingzhe Xu**

**Problem 1:**

# i) Compute mean, min, 1st Quartile, median, 3rd Quartile, Max, mode

## sum

sumfun <- function(col){

sum = 0

for(i in 1 : length(col))

{

if(!(is.na(col[i])))

{

sum = sum + col[i]

}

}

return(sum)

}

## Missing

missingfun <- function(col){

num = 0

for(i in 1 : length(col)){

if(is.na(col[i])){

num = num + 1

}

}

return(num)

}

## Mean:

meanfun <- function(col){

return(sumfun(col)/(length(col) - missingfun(col)))

}

## Median

medianfun <- function(col){

col <- sort(col)

len = length(col) - missingfun(col)

#print(col[1])

#print(col[length(col)])

if(len%%2==1)

{

return(col[len%/%2 + 1])

}

return((col[len/2] + col[len/2 + 1])/2)

}

## max

maxfun <- function(col){

max = 0

for (i in 1 : length(col))

{

if(!(is.na(col[i])))

if(col[i] > max)

{

max = col[i]

}

max = max

}

return(max)

}

## min

minfun <- function(c){

min = .Machine$integer.max

for (i in 1 : length(c))

{

if(!(is.na(c[i])))

if(c[i] < min)

{

min = c[i]

}

min = min

}

return(min)

}

## mode: the number that appear most frequently

modefun <- function(col){

return(names(sort(-table(col)))[1])

}

## Q1:

q1fun <- function (col){

col <- sort(col)

pos <- (length(col)+1)%/%4

rem <- (length(col)+1)%%4

if((length(col)+1) %% 4 == 0)

{

return(col[pos])

}

return(col[pos]+(rem\*(col[pos+1] - col[pos])))

}

## Q3:

q3fun <- function (col){

col <- sort(col)

pos <- 3\*((length(col)+1)%/%4)

rem <- (3\*(length(col)+1))%%4

if(3\*((length(col)+1)%%4) == 0)

{

return(col[pos])

}

return(col[pos] + rem \* (col[pos + 1] - col[pos]))

}

ii) Table for summary statistics for continuous variables

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Attribute | Min | Q1 | Median | Q3 | Max | Mean | Mode | Missing |
| Age | 17 | 28 | 37 | 48 | 90 | 38.65 | 36 | 470 |
| Fnlwgt | 12285 | 117549 | 178145 | 237630 | 1490400 | 189664 | 203488 | 0 |
| Edu.num | 1 | 9 | 10 | 12 | 16 | 10 | 9 | 254 |
| Cap.gain | 0 | 0 | 0 | 0 | 99999 | 1079 | 0 | 0 |
| Cap.loss | 0 | 0 | 0 | 0 | 4356 | 87.5 | 0 | 0 |
| Hours | 1 | 40 | 40 | 45 | 99 | 40 | 40 | 737 |

(b) Visualizing Data

ii) Box plot and class-conditional box plot

- Based on the box plot of age, we can derive several arguments:

# - the median is 37 so half of the people are below 37, half above.

# - half of the people are from 28 to 48

# - the conditional boxplot:

# -- the median age for income <= 50k is less than 37 and median age for income > 50k is more than 40

- Based on the box plot of education number, we can derive several arguments:

# - the median is 10 and half of the sample is within 9 - 12

# - 25% of the sample is 9 and 25% are 11 or 12

- median of education number for those income <= 50k is 9, while for income more than 50k, the median education number is obviously more than 10

- Based on the box plot of hours per month, we can derive several arguments:

# - the median is 40 and around 25% of the sample is 40

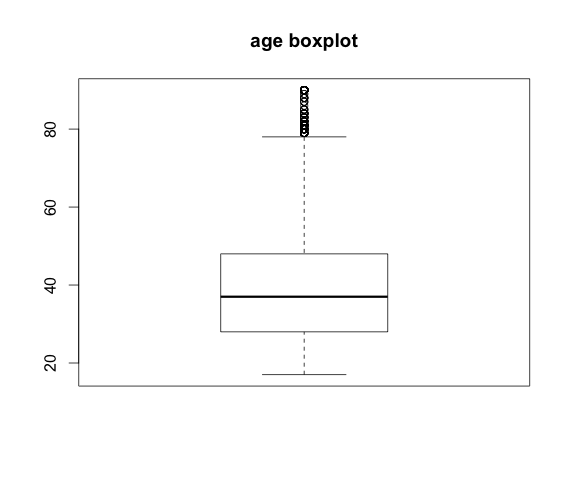
# - 25% of the sample is between 40 and 45

# - the conditional boxplot shows that the median hours from people who has income less or equal to 50k is 40, the same as those who earn more than 50k

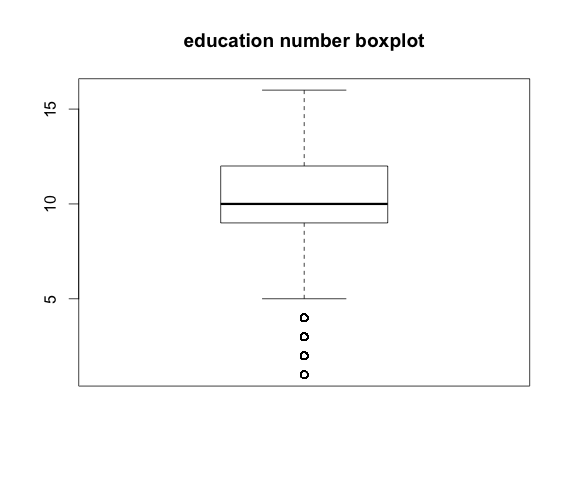
- 75% people whose income is less or equal to 50k work less than or equal to 40 hour

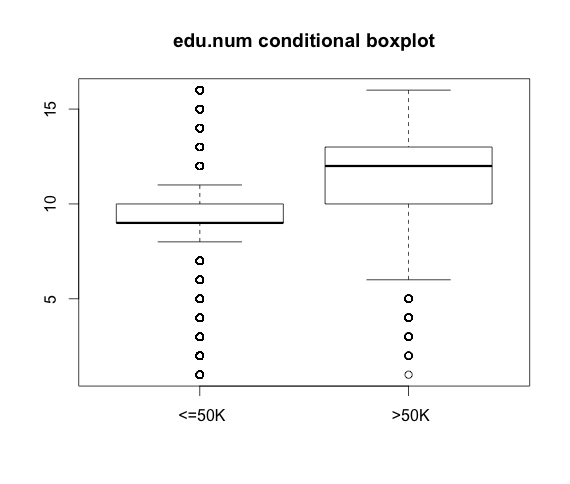
- 75% people whose income is more than 50k work more than or equal to 40 hours

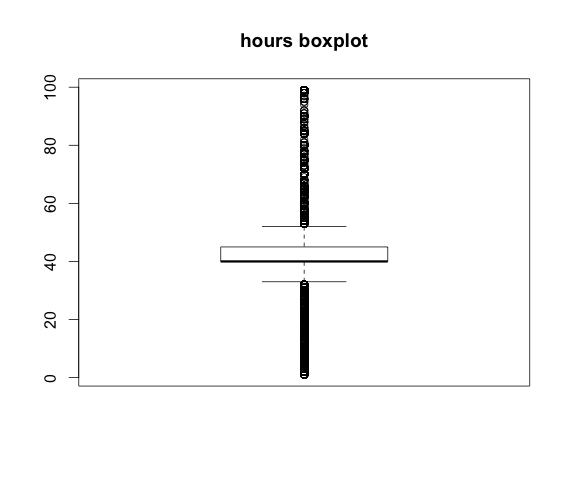
For capital.gain or capital.loss or capital gain, the value is either 0 or a large number and mostly is 0, so we couldn't tell much from the boxplot with median q1 q3 being 0.

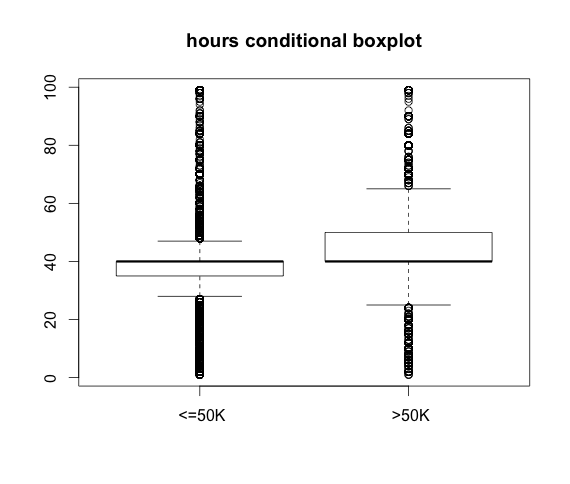




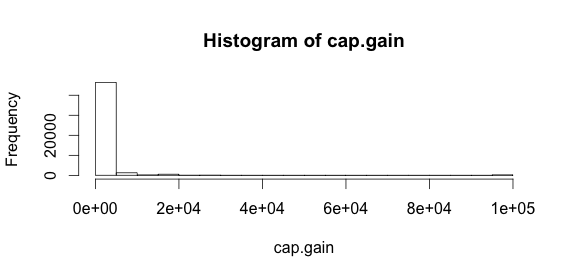


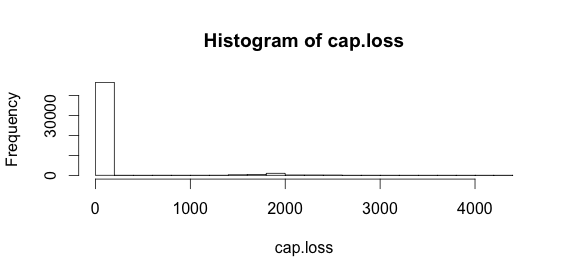






iii) Histograms for age, cap.loss, cap.gain





hist(age)

hist(cap.gain)

hist(cap.loss)

# Interpret:

# The histogram tells more than boxplots.

# For age, it is a little bit skewed to the left.

# And there are more people range from age 20 to 45 than other ranges

# cap.gain and cap.loss have similar characteristics - both are heavily skewed with most values being 0.

iv) barplots and class conditional bar plots for categorical variables



