Manipulate Data and Plot in R

week3

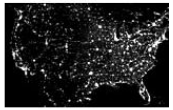
Import Adult data to R

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Adult Data Set

Download: Data Folder, Data Set Description



Abstract: Predict whether income exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	619300

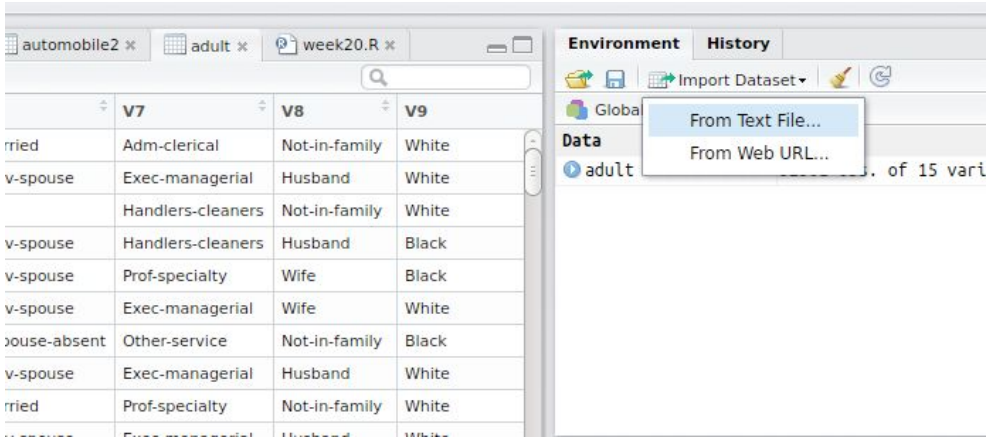
Source:

Donor:

Ronny Kohavi and Barry Becker
Data Mining and Visualization

1. Download adult data set from UCI ML repository.
2. remove question marks in by replace

Import Adult data to R



1. Import the data file you downloaded, set the variable name ‘adult’
2. check the data content, left click adult variable.
3. The data does not contain any header information

Add headers in R

1.headers are also called column names, first create a vector containing column names(header information is in .name file on the uci website)

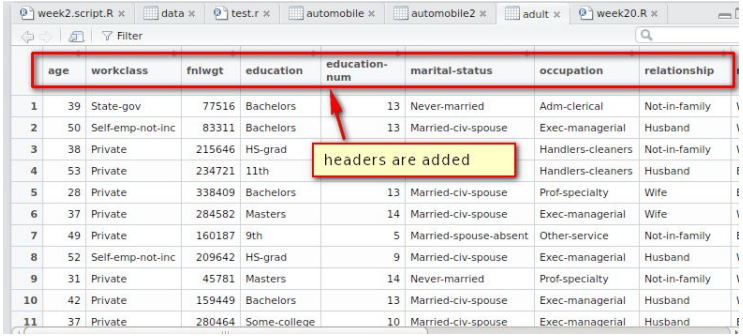
```
adult.header <- c('age','workclass','fnlwgt','education','education-num','marital-status','occupation','relationship','race','sex','captical-gain','capital-loss','hours-per-week','native-country','salary')
```

2.set the column names vector as the data column names(vector)

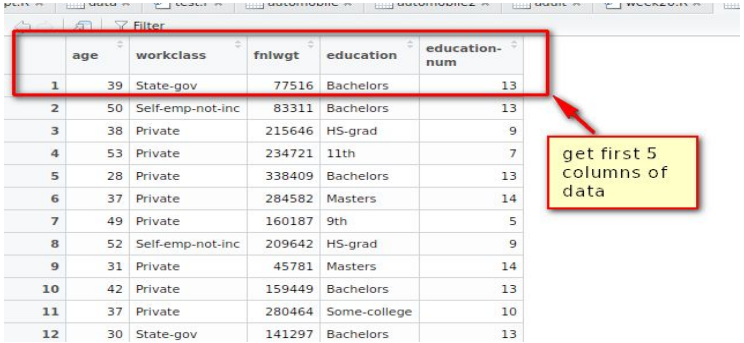
```
colnames(adult) <- adult.header
```

3. click the adult variable again and check if it is successfully added.

Data subset - columns



	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband
3	38	Private	215646	HS-grad			Handlers-cleaners	Not-in-family
4	53	Private	234721	11th			Handlers-cleaners	Husband
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife
7	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family
8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband
11	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband



	age	workclass	fnlwgt	education	education-num
1	39	State-gov	77516	Bachelors	13
2	50	Self-emp-not-inc	83311	Bachelors	13
3	38	Private	215646	HS-grad	9
4	53	Private	234721	11th	7
5	28	Private	338409	Bachelors	13
6	37	Private	284582	Masters	14
7	49	Private	160187	9th	5
8	52	Self-emp-not-inc	209642	HS-grad	9
9	31	Private	45781	Masters	14
10	42	Private	159449	Bachelors	13
11	37	Private	280464	Some-college	10
12	30	State-gov	141297	Bachelors	13

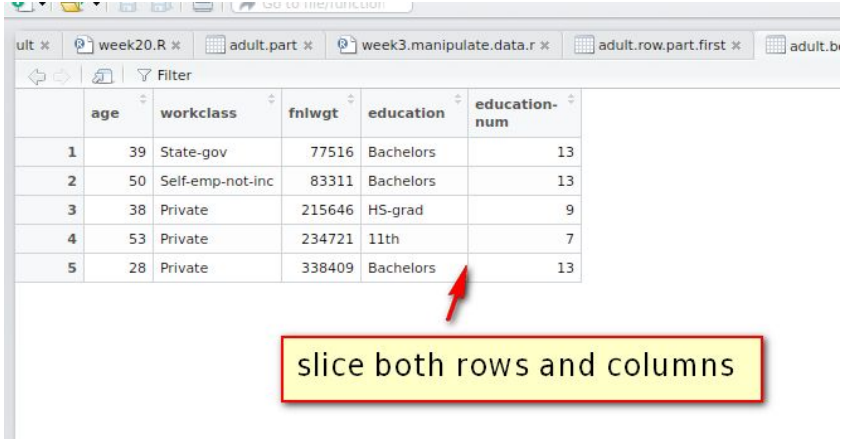
We just want to keep subset of the data based on columns because other columns are useless for us. Keep first 5 columns.

```
adult.part <- adult[1:5]
```

Keep specific columns(1st,3rd,5th,9th)

```
adult.part <- adult[c(1,3,5,9)]
```

Data subset - rows



	age	workclass	fnlwgt	education	education-num
1	39	State-gov	77516	Bachelors	13
2	50	Self-emp-not-inc	83311	Bachelors	13
3	38	Private	215646	HS-grad	9
4	53	Private	234721	11th	7
5	28	Private	338409	Bachelors	13

get first 5 rows data, comma is used to separate two dimension, we leave column dimension blank, which means keep all columns.

```
adult.row.part.first <- adult[1:5,]
```

Keep specific row s(1st,3rd,5th,9th)

```
adult.row.part.spec <- adult[c(1,3,5,9),]
```

slice rows and columns

```
adult.both.part <- adult[1:5,1:5]
```

Data subset - values

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	captical-gain
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	21
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	
3	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	
4	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	
6	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	
7	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	
8	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	
9	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	140
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	51
11	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	
14	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	
15	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	
16	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	
18	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	
19	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	
20	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	
21	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	

Filter

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male
2	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male
5	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female
10	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male
12	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male
13	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female
26	56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male
33	45	Private	386940	Bachelors	13	Divorced	Exec-managerial	Own-child	White	Male
42	53	Self-emp-not-inc	88506	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	White	Male
43	24	Private	172987	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male
46	57	Federal-gov	337895	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Black	Male
54	50	Federal-gov	251585	Bachelors	13	Divorced	Exec-managerial	Not-in-family	White	Male
61	30	Private	59496	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male
72	31	Private	309974	Bachelors	13	Separated	Sales	Own-child	Black	Female
73	29	Self-emp-not-inc	162298	Bachelors	13	Married-civ-spouse	Sales	Husband	White	Male
82	52	Private	276515	Bachelors	13	Married-civ-spouse	Other-service	Husband	White	Male
95	34	Local-gov	226296	Bachelors	13	Married-civ-spouse	Protective-serv	Husband	White	Male
102	44	Private	198282	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male
111	24	Private	432376	Bachelors	13	Never-married	Sales	Other-relative	White	Male

find person who is older than or equal to 30.

```
adult.age.part <- subset(adult,age>=30)
```

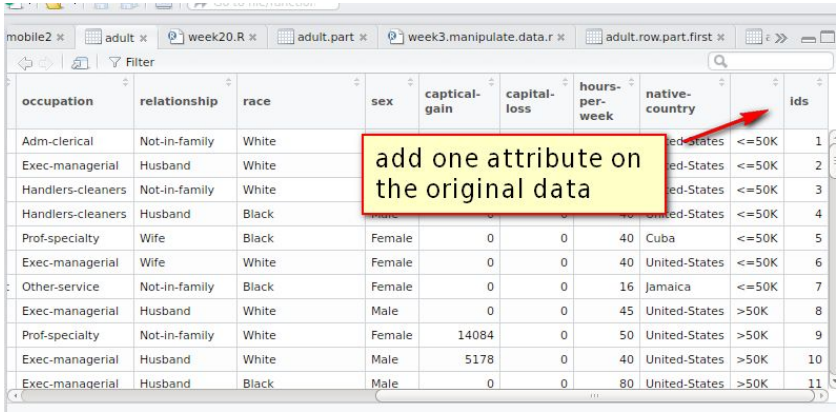
find all female

```
adult.women <- subset(adult,sex == ' Female')
```

find all who has bachelor degree adult.

```
bachelors <- subset(adult,education == ' Bachelors')
```

Add attribute to data



occupation	relationship	race	sex	captical-gain	capital-loss	hours-per-week	native-country		ids
Adm-clerical	Not-in-family	White					ed-States	<=50K	1
Exec-managerial	Husband	White					ed-States	<=50K	2
Handlers-cleaners	Not-in-family	White					ed-States	<=50K	3
Handlers-cleaners	Husband	Black					ited-States	<=50K	4
Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K	5
Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K	6
Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K	7
Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K	8
Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K	9
Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K	10
Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K	11

let's add an id for each person in the data for organization

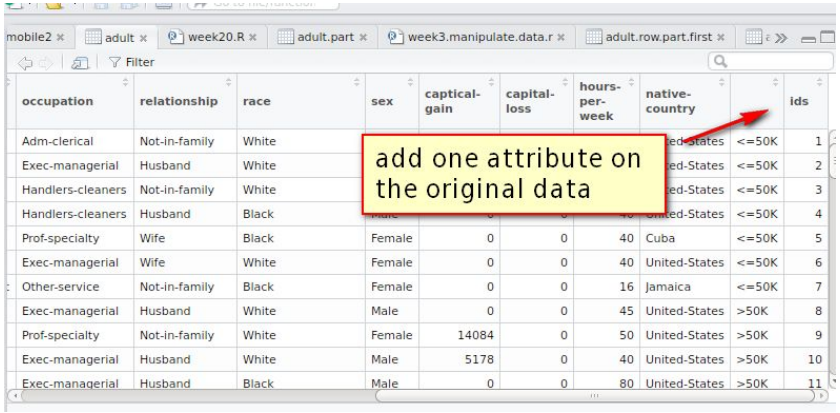
First, create a vector containing ids, generate id from 1 to number of rows in the data.

```
ids <- c(1:nrow(adult))
```

add ids in the data table.

```
adult$ids <- ids
```

Add attribute to data



occupation	relationship	race	sex	captical-gain	capital-loss	hours-per-week	native-country		ids
Adm-clerical	Not-in-family	White					ited-States	<=50K	1
Exec-managerial	Husband	White					ted-States	<=50K	2
Handlers-cleaners	Not-in-family	White					ted-States	<=50K	3
Handlers-cleaners	Husband	Black					United-States	<=50K	4
Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K	5
Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K	6
Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K	7
Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K	8
Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K	9
Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K	10
Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K	11

let's add an id for each person in the data for organization

First, create a vector containing ids, generate id from 1 to number of rows in the data.

```
ids <- c(1:nrow(adult))
```

add ids in the data table.

```
adult$ids <- ids
```

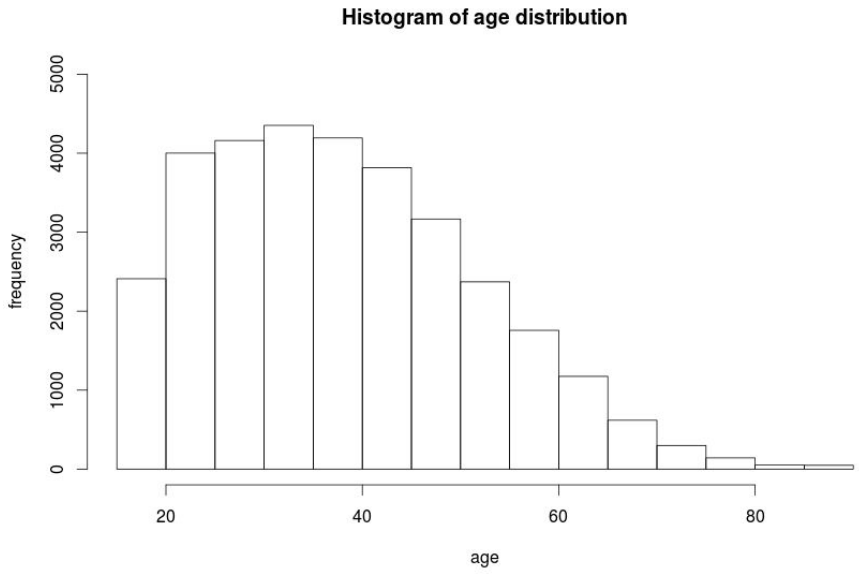
Data summary information

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

To get data summary information, we use

s <- summary(adult)

Plot - Histograms



We can use the histogram to show the distribution of ONE specific attribute.

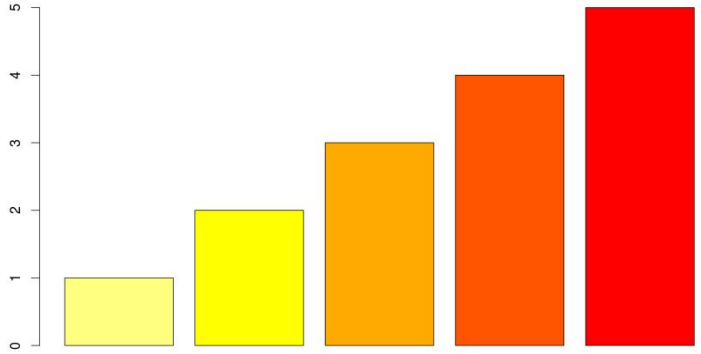
If we check the distribution of age, then

```
hist(adult$age)
```

Do more on Histogram, add Title, xlabel, ylabel and set the range of y value:

```
hist(adult$age,main=paste("Histogram of age distribution"),
xlab = 'age', ylab = 'frequency',ylim = c(0,5000))
```

Plot - barplot

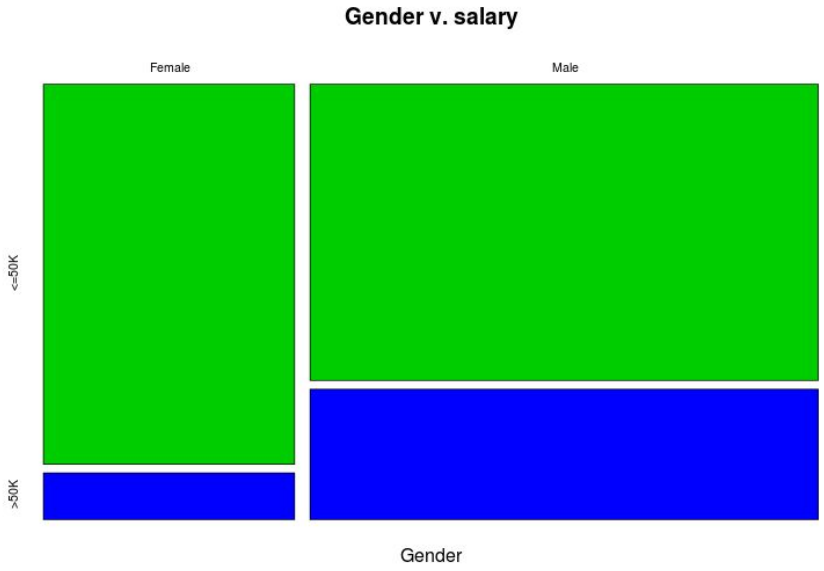


The difference between barplot and histogram plot is
hist(x) plots the frequency of x and barplot (x) plots the value of x.

If we check the distribution of age, then

```
barplot(c(1,2,3,4,5),col = rev(heat.colors(5)))
```

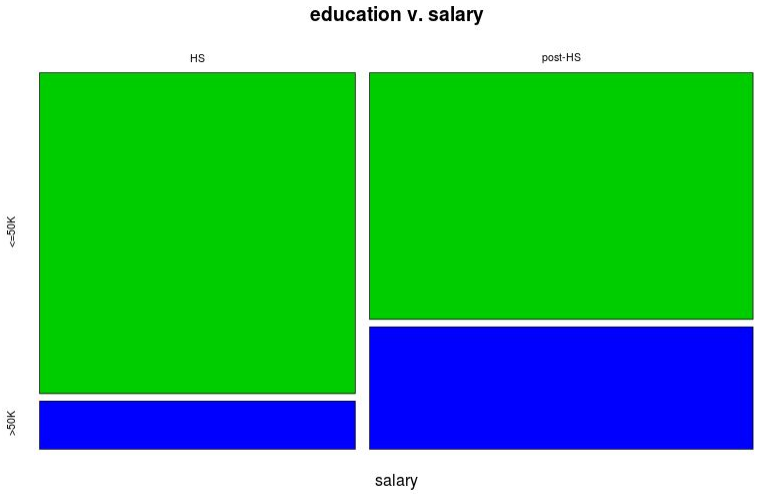
Plot - mosaicplot



Find the relation between gender and salary

```
mosaicplot(table(adult$sex,adult$salary),color=3:4,
xlab="Gender",main="Gender v. salary")
```

Transform values



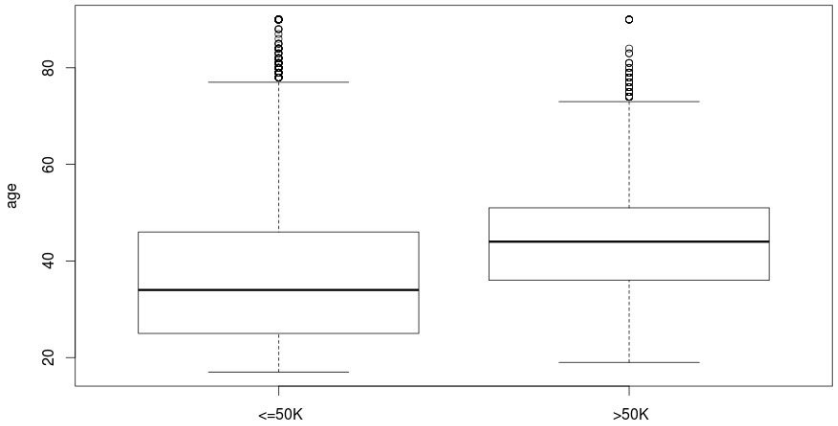
Find the relation between gender and salary

```
education <- rep("post-HS",nrow(adult))

education[adult$education %in% c(" 1st-4th", " 5th-6th", " 7th-8th", " 9th", " 10th", " 11th", " 12th", " HS-grad")] <- "HS"

mosaicplot(table(education,adult$salary),color=3:4, xlab="salary",main="education v. salary")
```

Plot - box plot



Find the relation between age and salary

```
boxplot(adult$age~adult$salary,ylab="age")
```