Exercise 5

[Number of the rows: 1](#_Toc506239792)

[quantitative variables: 1](#_Toc506239793)

[age: 1](#_Toc506239794)

[fnlwgt: 3](#_Toc506239795)

[education-num: 4](#_Toc506239796)

[capital-gain: 5](#_Toc506239797)

[capital-loss: 7](#_Toc506239798)

[hours-per-week: 8](#_Toc506239799)

[qualitative variables: 9](#_Toc506239800)

[workclass: 9](#_Toc506239801)

[education: 10](#_Toc506239802)

[marital-status: 12](#_Toc506239803)

[occupation: 13](#_Toc506239804)

[relationship: 15](#_Toc506239805)

[race: 16](#_Toc506239806)

[sex: 17](#_Toc506239807)

[native-country: 18](#_Toc506239808)

[Income: 20](#_Toc506239809)

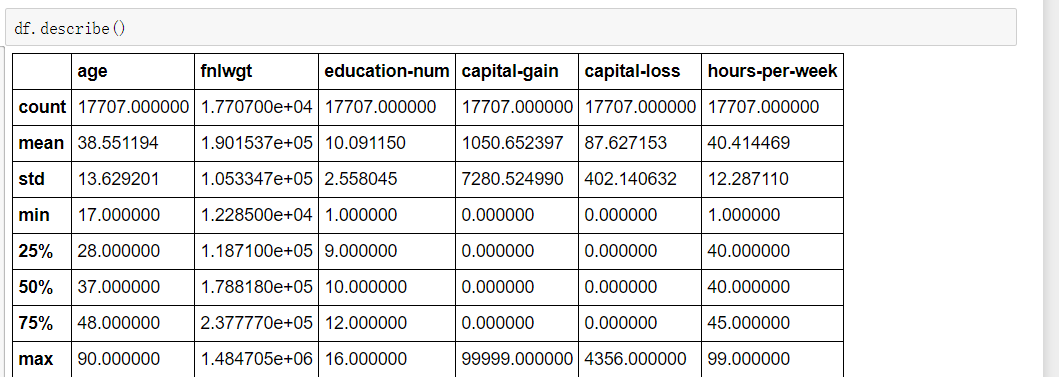
# Number of the rows:

48842 (there are 48842 instances)

I get this data from UCI data base: <http://archive.ics.uci.edu/ml/datasets/Census+Income>. It is an extraction was done by Barry Becker from the 1994 Census database. For this data set, prediction task is to determine whether a person makes over 50K a year.

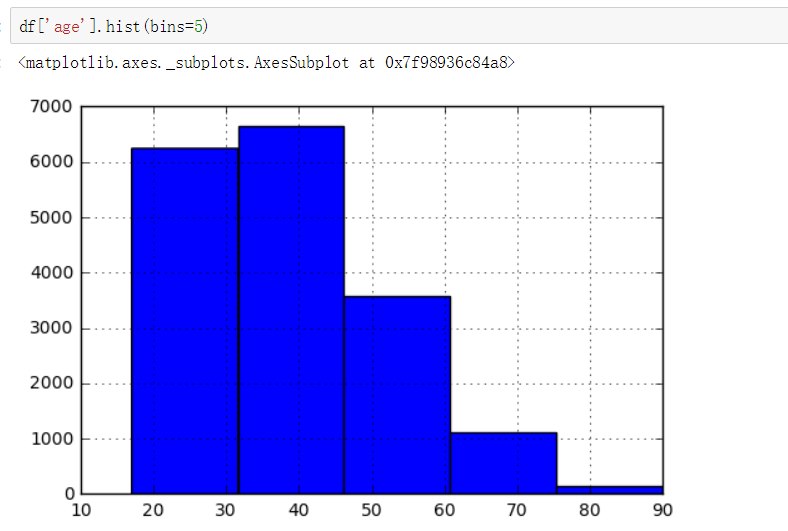
# quantitative variables:

there are 6 attributes

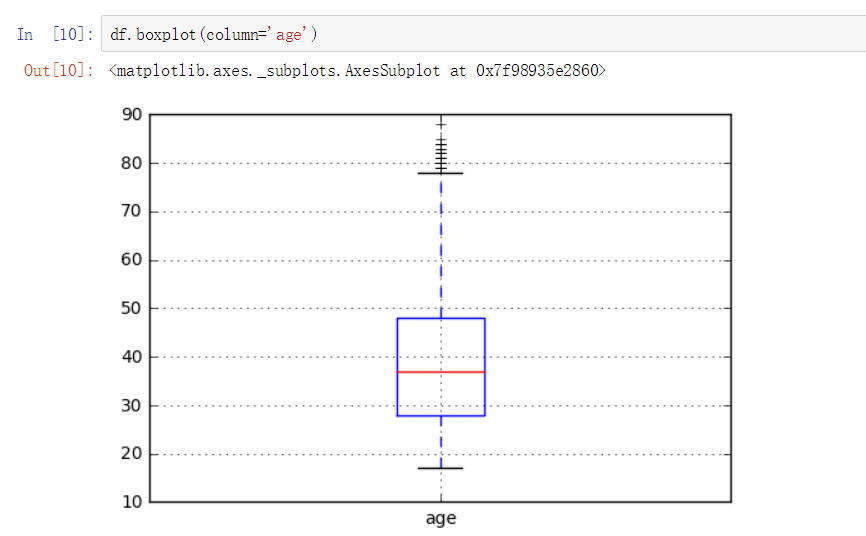


## age:

we can find that people are mainly distributed on the age from 18 to 45

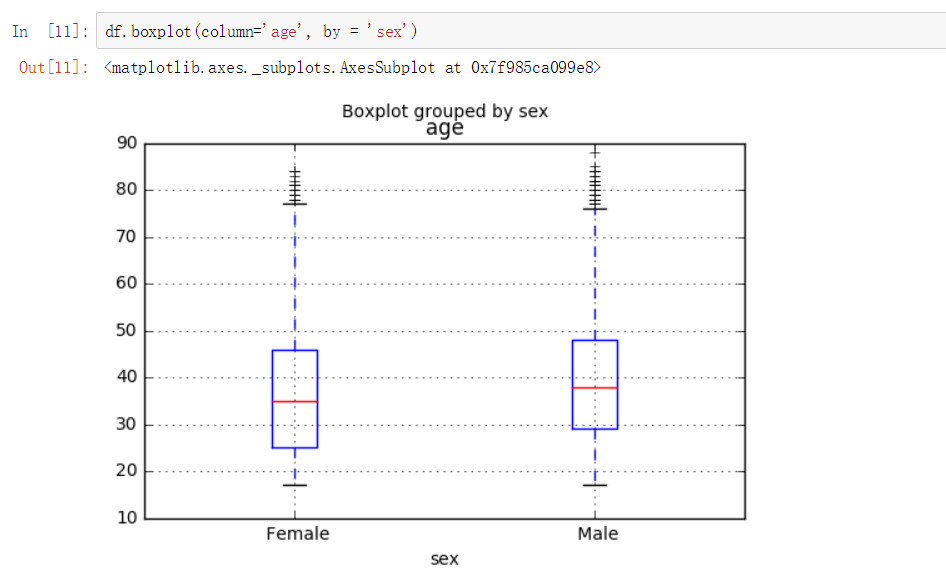


In the boxplot, we can find :



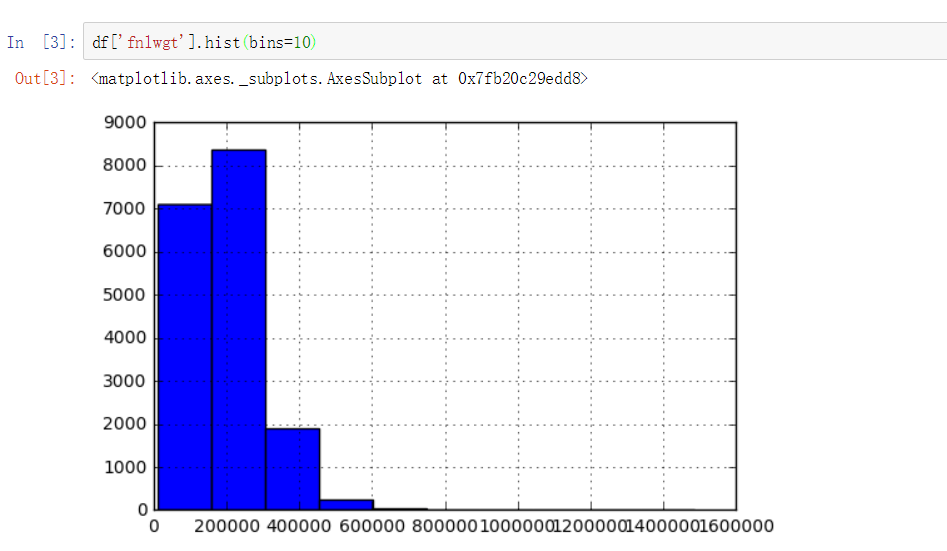
We can find there are some extreme values where the age is from 78 to 88, most of the age are from 28 to 48

If we separate them by gender, we can get: there is no clear difference on age between female and male.



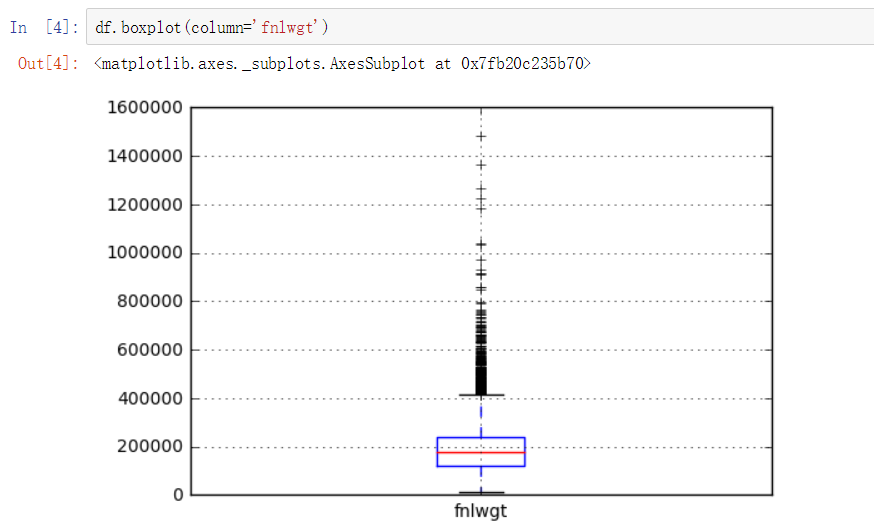
## fnlwgt:

The weights are controlled to independent estimates of the civilian noninstitutionalized population of the US. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population.

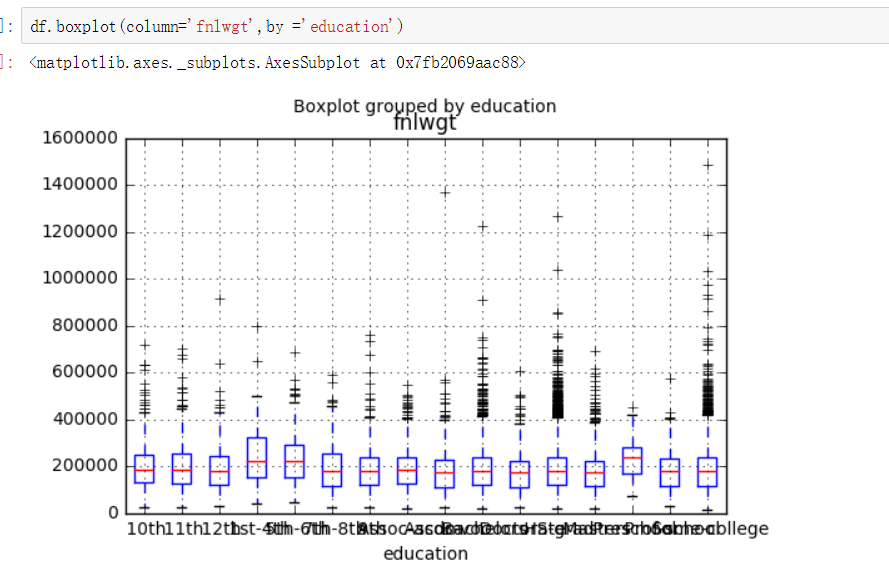


We can find that the ‘fnlwgt’ weight is distributed mainly in 0 to 500000, the value from 200000 to 300000 takes the highest frequency.

Then we get this box plot:



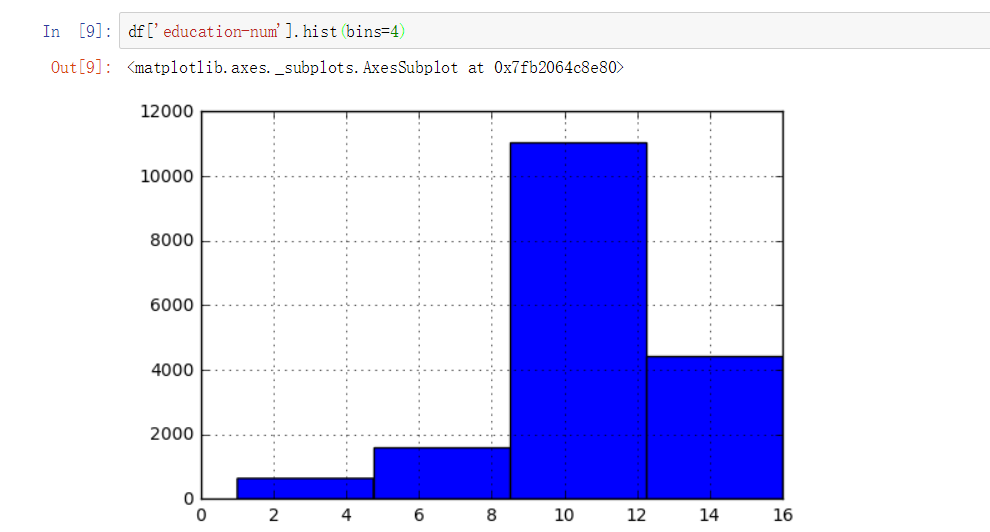
There are many extreme values that are distributed on 400000 to 800000, but the majority is from 100000 to 300000.



We can find there is no clear difference for this weight on different education

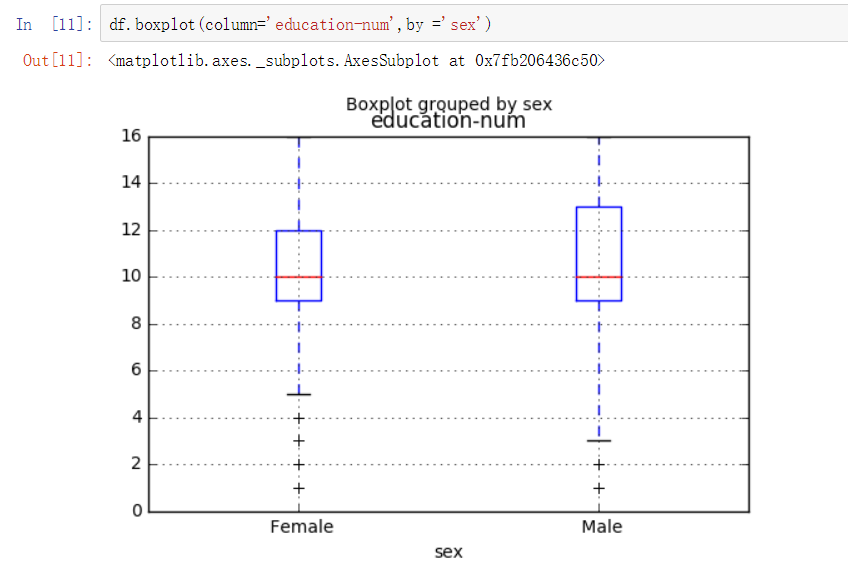
## education-num:

this variable is not clearly described in the web page, in my opinion, it is the number of years for getting education



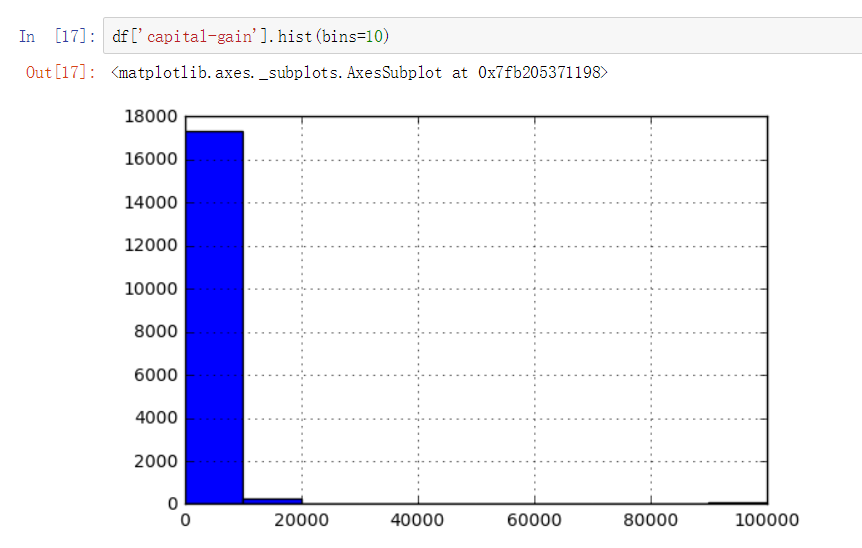
We can find most of the people have 9 to 12 years of education, Then is 12 years to 16 years, the least goes to 1 to 5 years.

In the box plot, we can find that if we separate the values by gender, for females, most of them are distributed on 9 to 12 years, for males, most of them have 9 to 13 years of educations. Also there are some extreme value from 0 to 5 years.

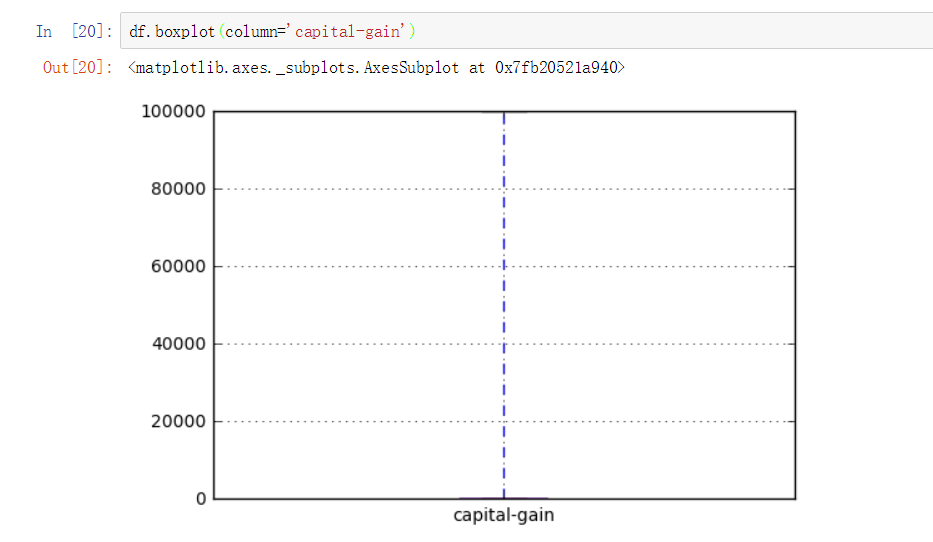


## capital-gain:

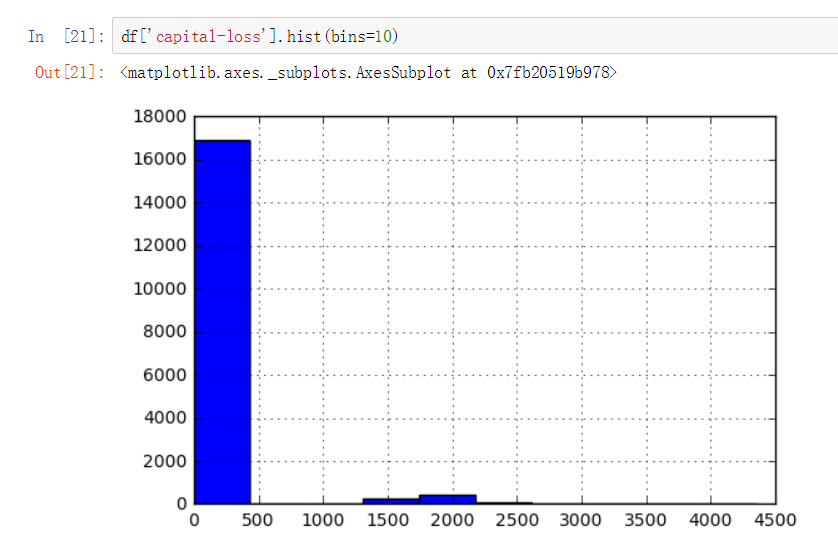
we can find this variable is mainly distributed on 0 to 20000, which means most of people have 0 to 20000 capital-gain



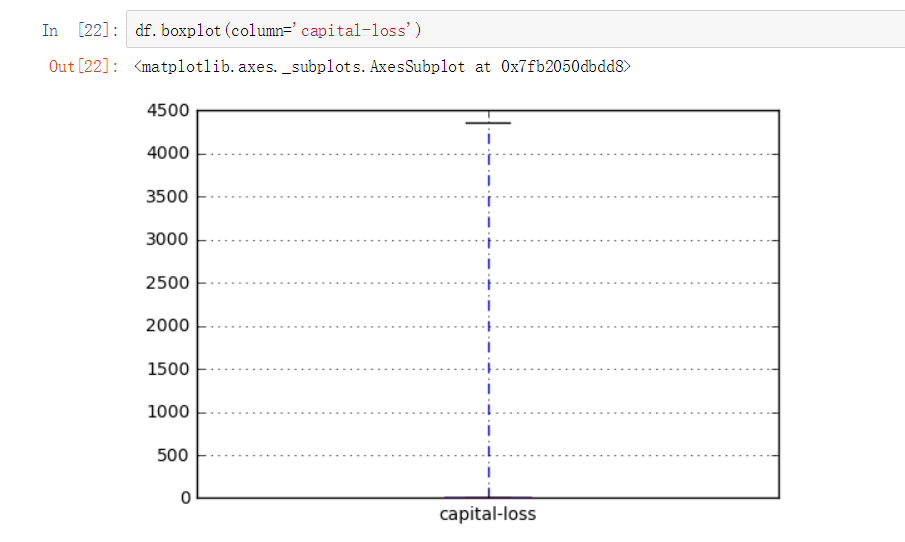
In the box plot, we can find there is no clear distribution for this variable



## capital-loss:

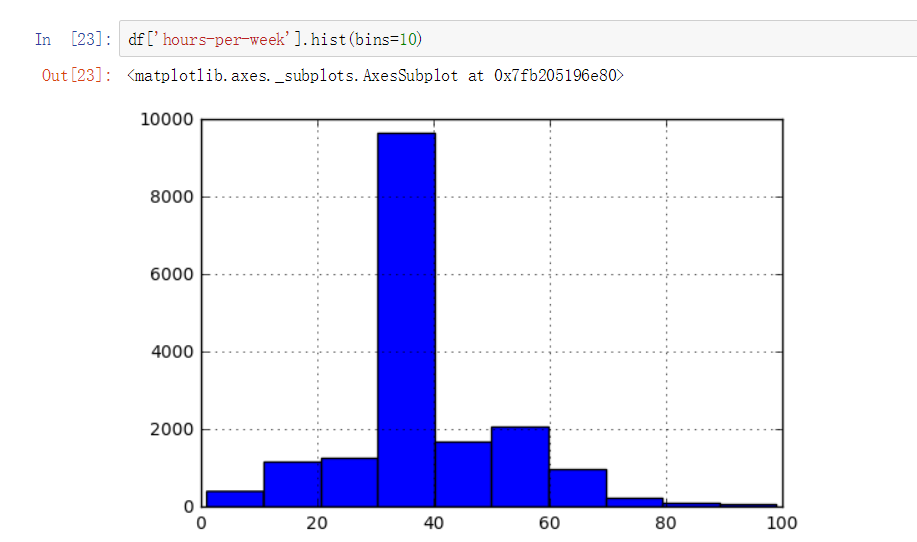


We can find most of people have 0 to 400 capital-loss



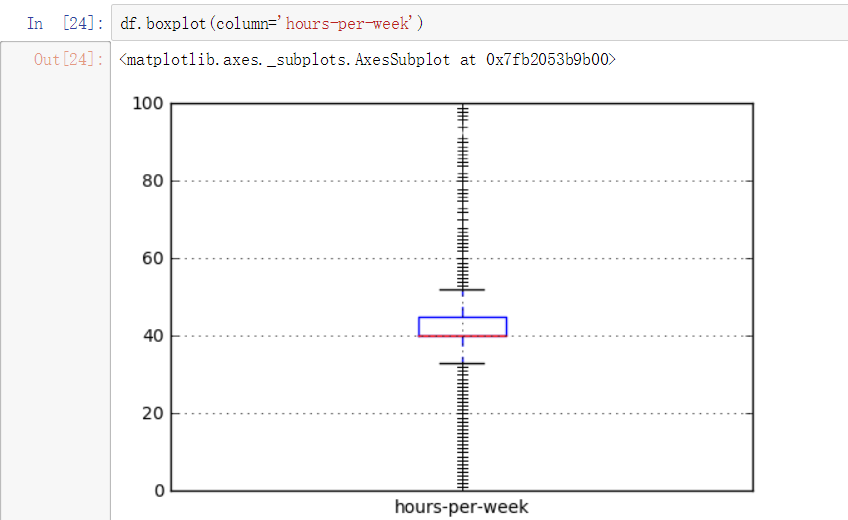
We can see the distribution is on 0 to 4500

## hours-per-week:



We can see that most of the people work 30 to 40 per week. The range is from 0 to 100

In the box plot, we can find that :



We can find the majority of the hours-per-week is from 30 to 50.

# qualitative variables:

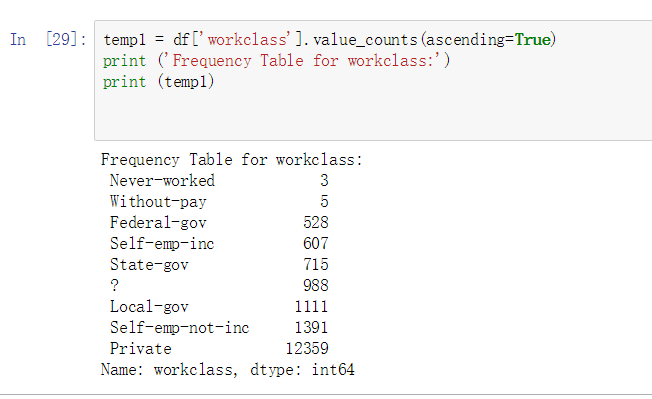
there are 9 variables

we can see those are the qualitative variables:

## workclass:

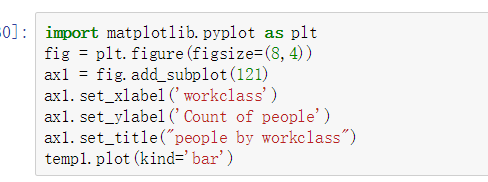
Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

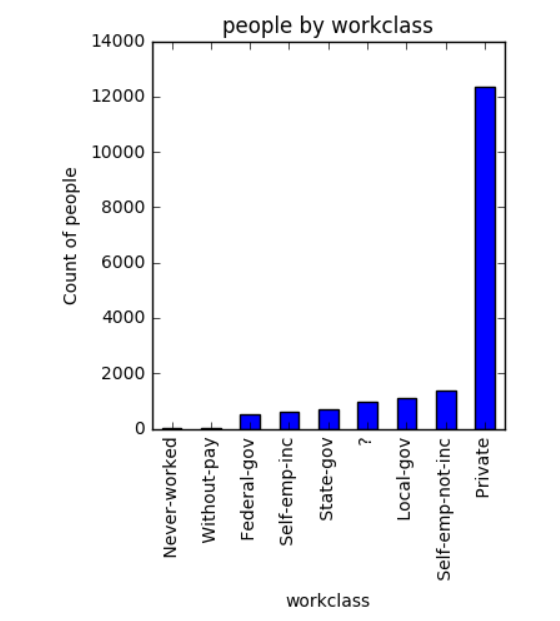
We can see the frequency of this variable in an ascending way:



‘Private’ work class take the highest proportion,

By those commands below, we can visualize the distribution of this variable:

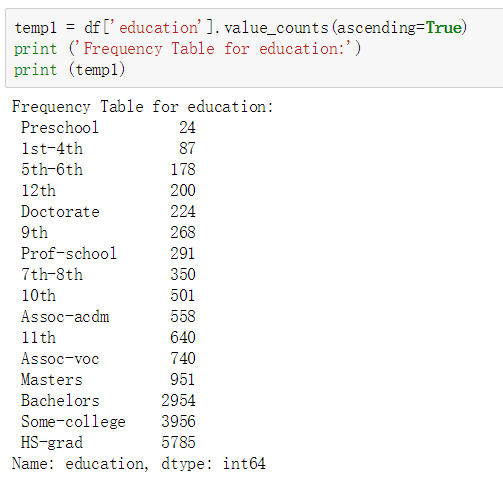




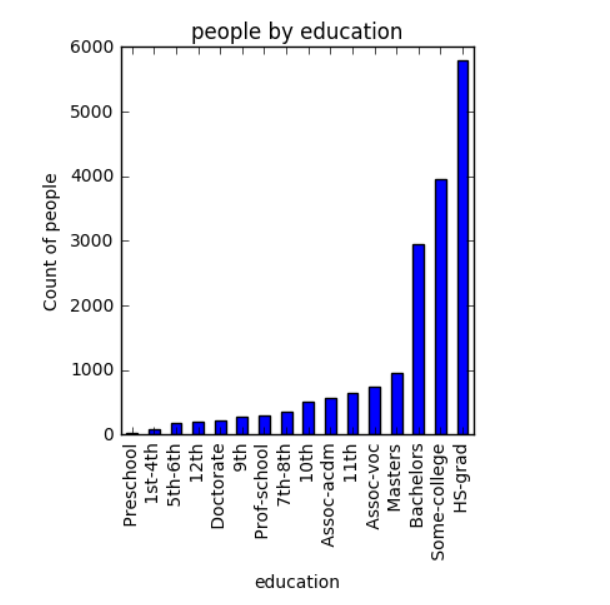
We can see that least people never worked, most people have private work class.

## education:

Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.



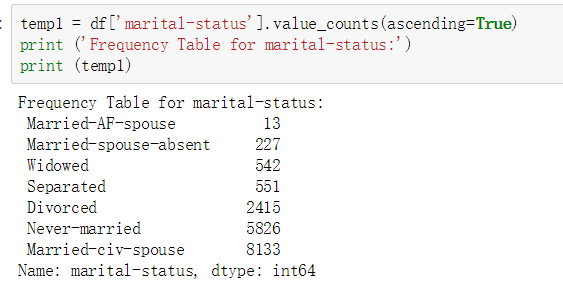
We can see that most of the US people at that time have high school graduating education, the frequency of them is 5785. Least of people have only preschool education.



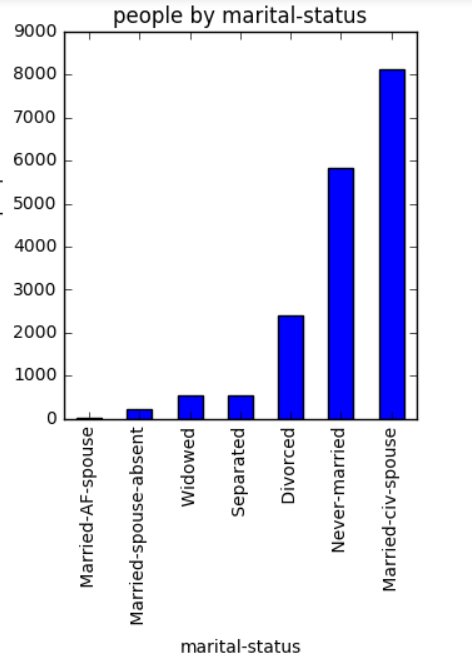
We can find that most people have high school education.

## marital-status:

Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.



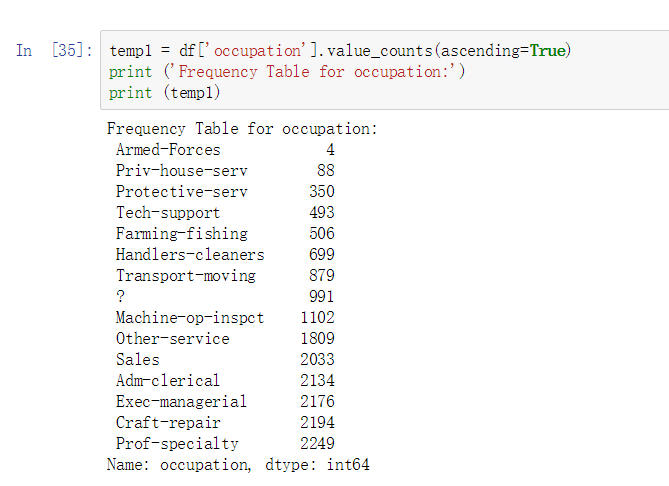
For marital status, we can find most people have married to civil spouse, least people have married to army forced spouse.



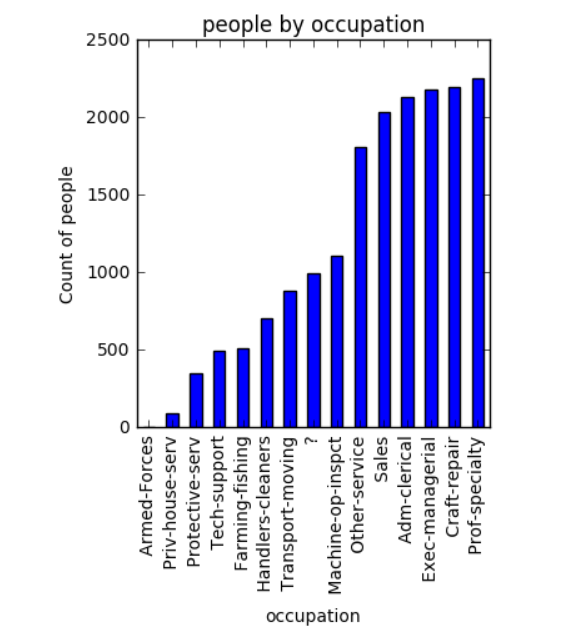
We can see that the mainly distribution are: Divorced, Never married, and married-civ-spouse.

## occupation:

Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.



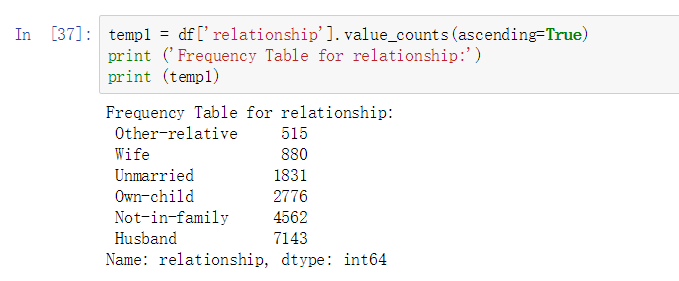
We can see that most people have Prof-specialty occupations, the frequency is 2249, and least people have armed-forces occupations.



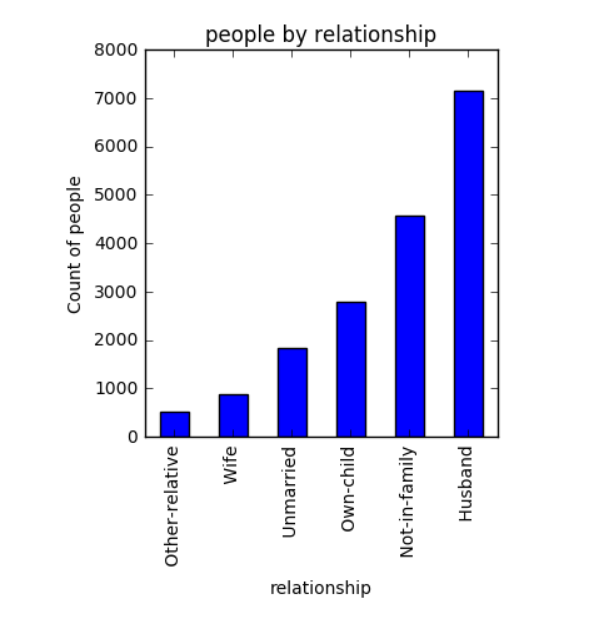
We can find main distribution of occupation are: Other service, Sales, Adm-clerical, Exec-managerial, Craft-repair, Prof-specialty.

## relationship:

Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.



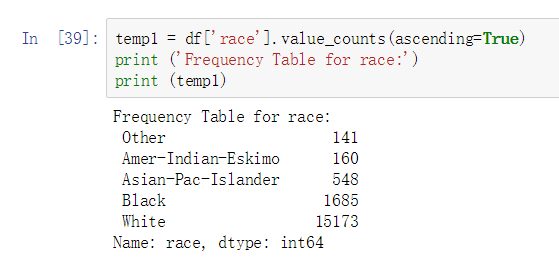
We can see most people take the role of ‘Husband’, least people take the ‘ Other-relative’ relationship.



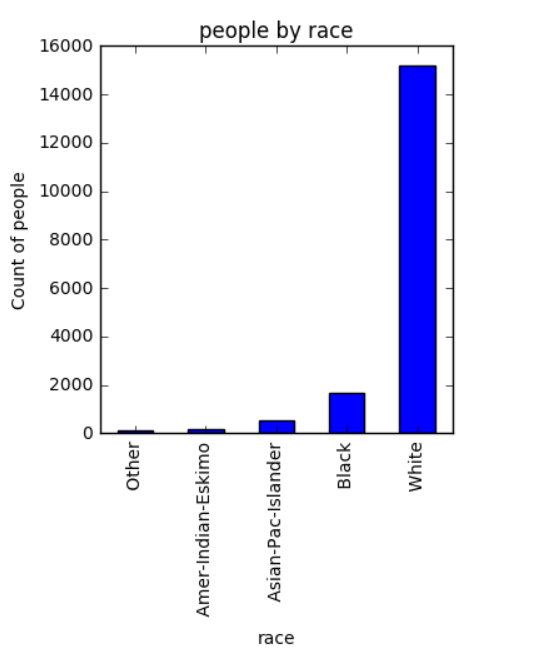
We can see the ‘Husband’ takes highest proportion.

## race:

White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.



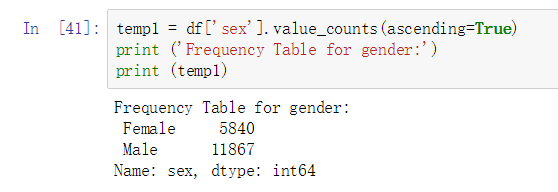
We can see that white people take the highest proportion, frequency is 15173, ‘Other’ race takes least number.



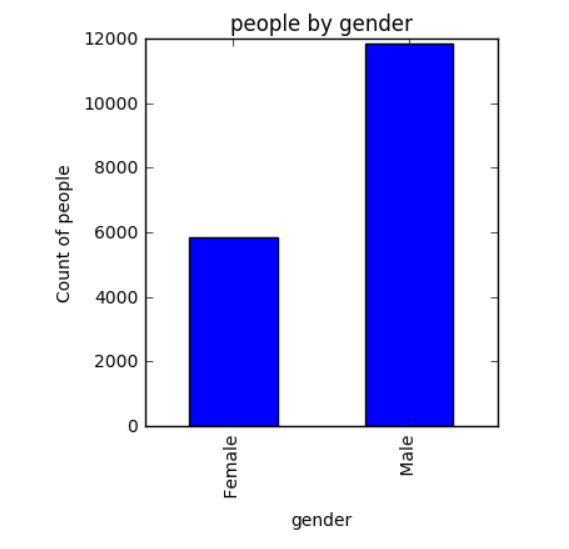
We can see there is a clear difference between white race and other race, white race has most people.

## sex:

Female, Male.



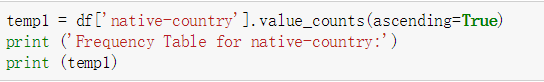
We can see males are almost twice of females, the frequency is 11867

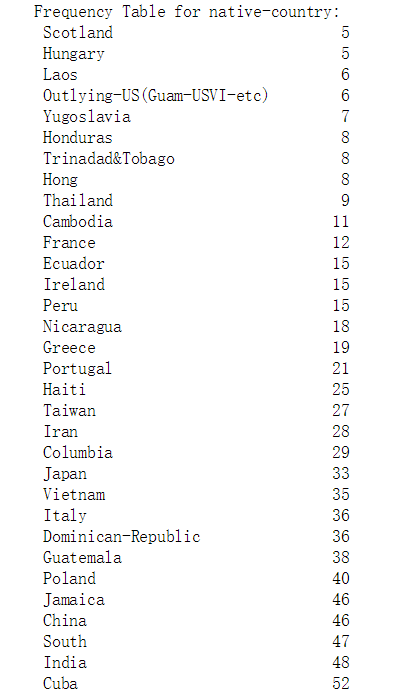


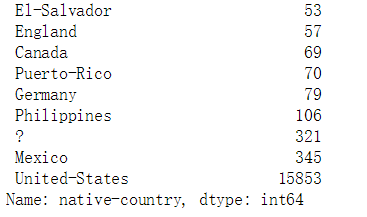
We can see that males take the majority.

## native-country:

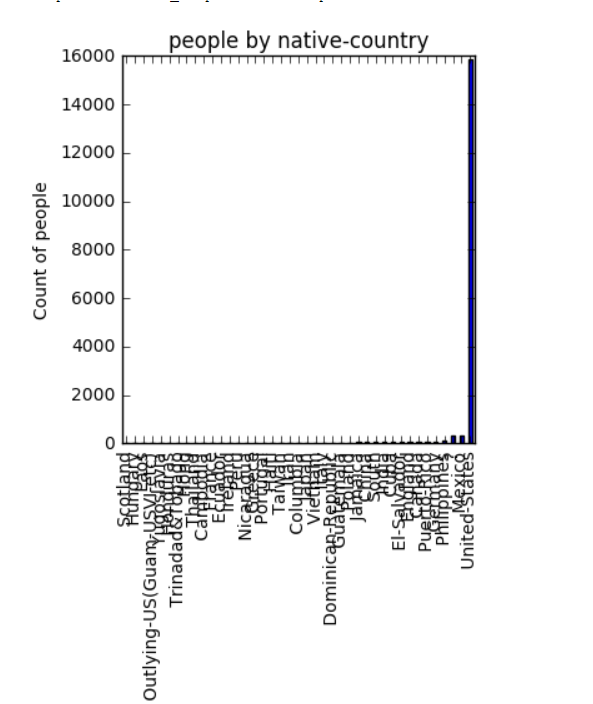
United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal. Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.







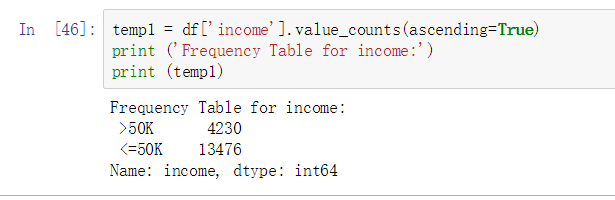
We can find most US people’s native-country is United-states, least people’s native-country is Scotland.



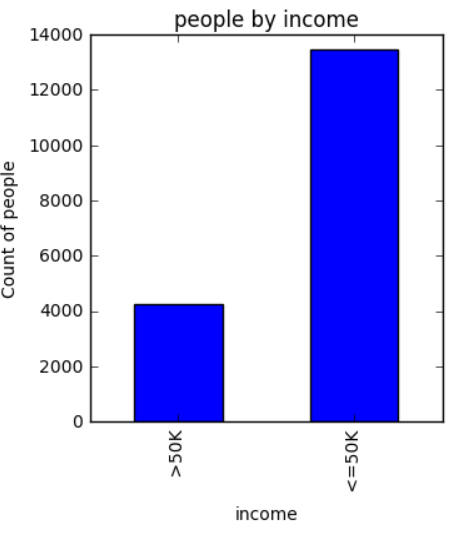
We can see most people’s native-country is the United-States

## Income:

>50K, <=50K.



We can see that most people incomes per year are lower than 50K dollars, which are the 3 times of people who have incomes per year which are higher than 50K dollars.



We can see most of people incomes <= 50K per year.