20122065 _ R Lab 12

Code

Importing Dataset

Hide

```
df <- read.csv("adult.csv", stringsAsFactors = T)
head(df)
```

	... <int>	workclass <fctr>	fnlwgt <int>	education <fctr>	education.num <int>	marital.status <fctr>	occupation <fctr>
1	90	?	77053	HS-grad	9	Widowed	?
2	82	Private	132870	HS-grad	9	Widowed	Exec-managerial
3	66	?	186061	Some-college	10	Widowed	?
4	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct
5	41	Private	264663	Some-college	10	Separated	Prof-specialty
6	34	Private	216864	HS-grad	9	Divorced	Other-service

6 rows | 1-9 of 15 columns

Summary of the dataset

Hide

```
summary(df)
```

```
      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    HS-grad     :10501
Min.   : 1.00
 Some-college: 7291   1st Qu.: 9.00
 Bachelors   : 5355   Median :10.00
 Masters     : 1723   Mean   :10.08
 Assoc-voc   : 1382   3rd Qu.:12.00
 11th        : 1175   Max.   :16.00    (Other)
: 5134
marital.status             occupation    Divorced
: 4443   Prof-specialty :4140
 Married-AF-spouse    :   23   Craft-repair   :4099
 Married-civ-spouse   :14976   Exec-managerial:4066
 Married-spouse-absent:  418   Adm-clerical   :3770
 Never-married        :10683   Sales          :3650
 Separated            : 1025   Other-service  :3295
Widowed              :  993   (Other)        :9541
relationship                    race           sex
```

```
Husband       :13193   Amer-Indian-Eskimo:  311
Female:10771
 Not-in-family : 8305   Asian-Pac-Islander: 1039   Male  :21790
 Other-relative:  981   Black             : 3124
 Own-child     : 5068   Other             :  271
 Unmarried     : 3446   White             :27816
 Wife          : 1568
capital.gain    capital.loss    hours.per.week   Min.   :    0
Min.   :   0.0   Min.   : 1.00
 1st Qu.:    0   1st Qu.:   0.0   1st Qu.:40.00
 Median :    0   Median :   0.0   Median :40.00
 Mean   : 1078   Mean   :  87.3   Mean   :40.44
 3rd Qu.:    0   3rd Qu.:   0.0   3rd Qu.:45.00
 Max.   :99999   Max.   :4356.0   Max.   :99.00
native.country    income      United-States:29170
<=50K:24720
 Mexico       :  643   >50K : 7841
 ?            :  583
 Philippines  :  198
 Germany      :  137
 Canada       :  121
 (Other)      : 1709
```

Structure of the dataset

Hide

```
str(df)
```

```
'data.frame':   32561 obs. of  15 variables:
 $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
 $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 1 5 1 5 5 5 5 8 2 5 ...
 $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037
...
 $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16
...  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
 $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 7 7 1 6 1 6 5 1 5
...
 $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 1 5 1 8 11 9 2 11 11 4 ...
 $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 2 5 5 4 5 5 3 2 5 ...
 $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 5 5 5 5 5 5 5 ...
 $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2
...  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ capital.loss  : int  4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
 $ hours.per.week: int  40 18 40 40 40 45 40 20 40 60 ...
 $ native.country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 1 ...
$ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 1 2 ...
```

WE USE THE LIBRARY AMELIA FOR THE VISULAISATION OF MISSING VALUES FOR ANALYSIS OF DATASET

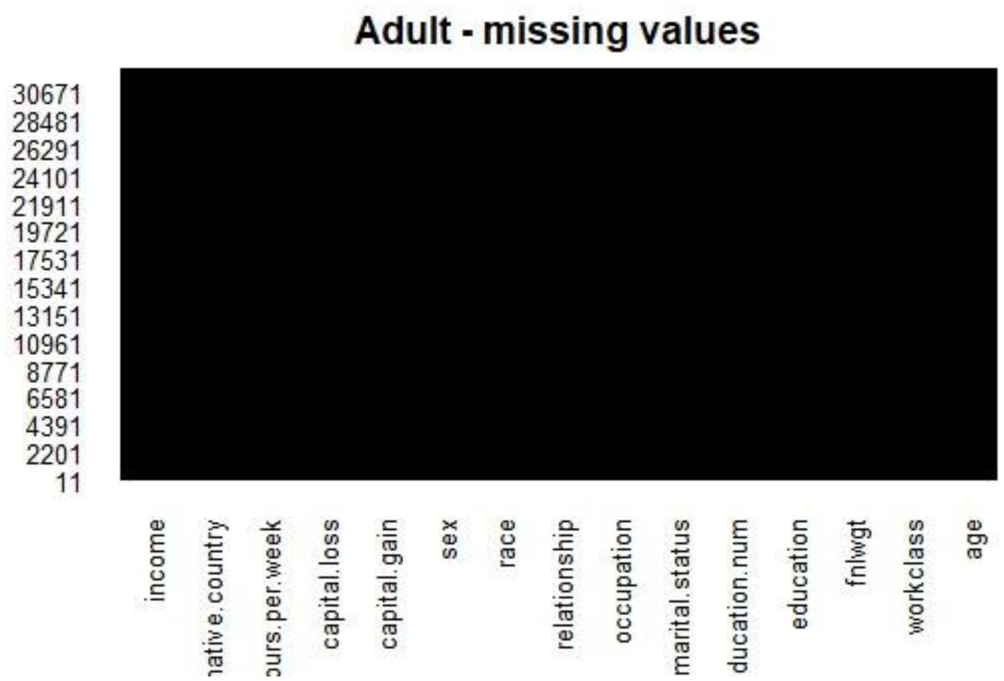
Hide

```
library(Amelia)
```

```
package 恊牠Amelia恊牫 was built under R version 4.0.4Loading required package: Rcpp
package 恊牠Rcpp恊牫 was built under R version 4.0.3##
## Amelia II: Multiple Imputation
## (Version 1.7.6, built: 2019-11-24)
## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell
## Refer to http://gking.harvard.edu/amelia/ for more information
##
```

Hide

```
missmap(df,main="Adult - missing values",col = c("yellow","black"), legend = FALSE)
```



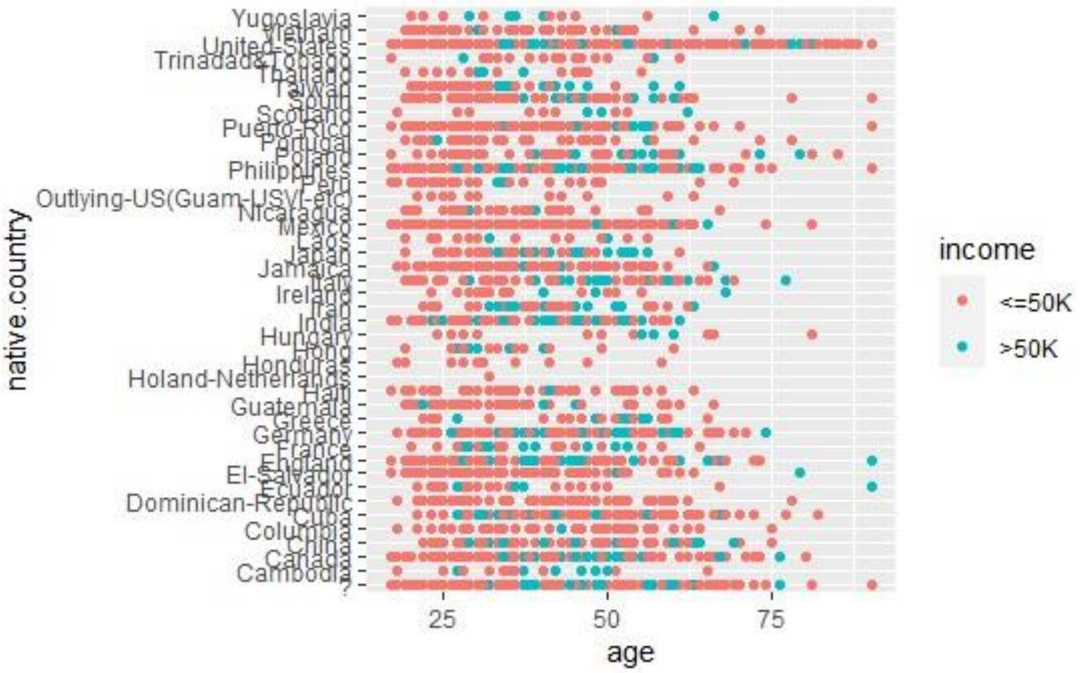
There are no missing values.

EDA

WE USE GGPLOT2 FOR VISUALIZATION OF RELATIONSHIP BETWEEN THE VARIABLE FOR ANALYSIS IN OUR DATASET

Hide

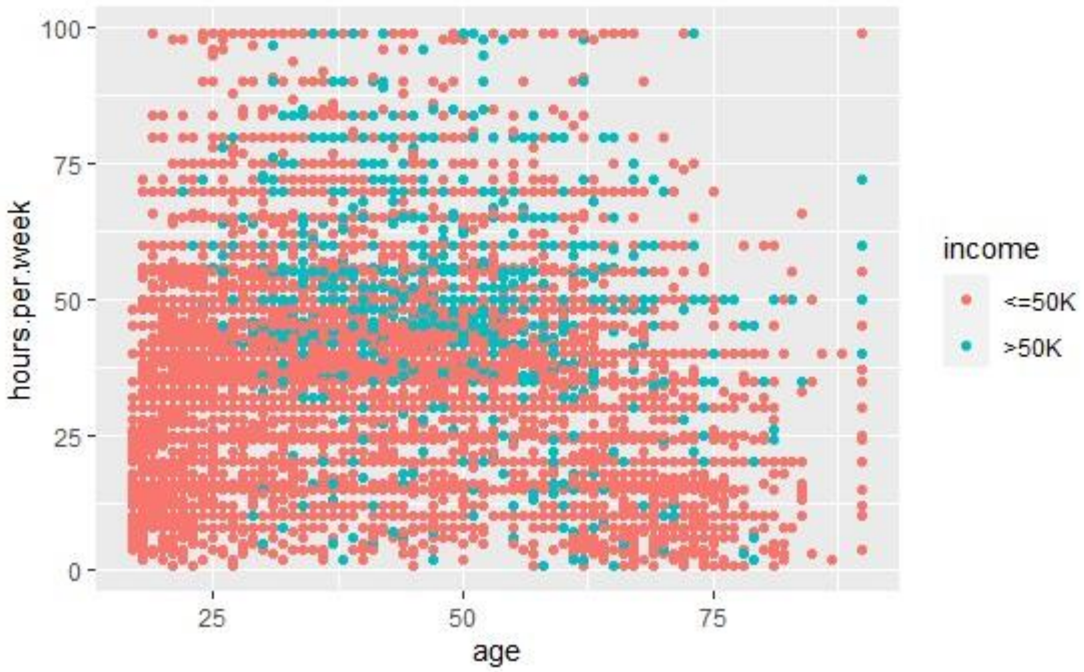
```
library(ggplot2)
ggplot(df,aes(age, native.country))+geom_point(aes(color=income))
```



People under the ageof 25 have income <=50K. United states has more citizens with income <=50K.Every country has more citizens with income <=50K

Hide

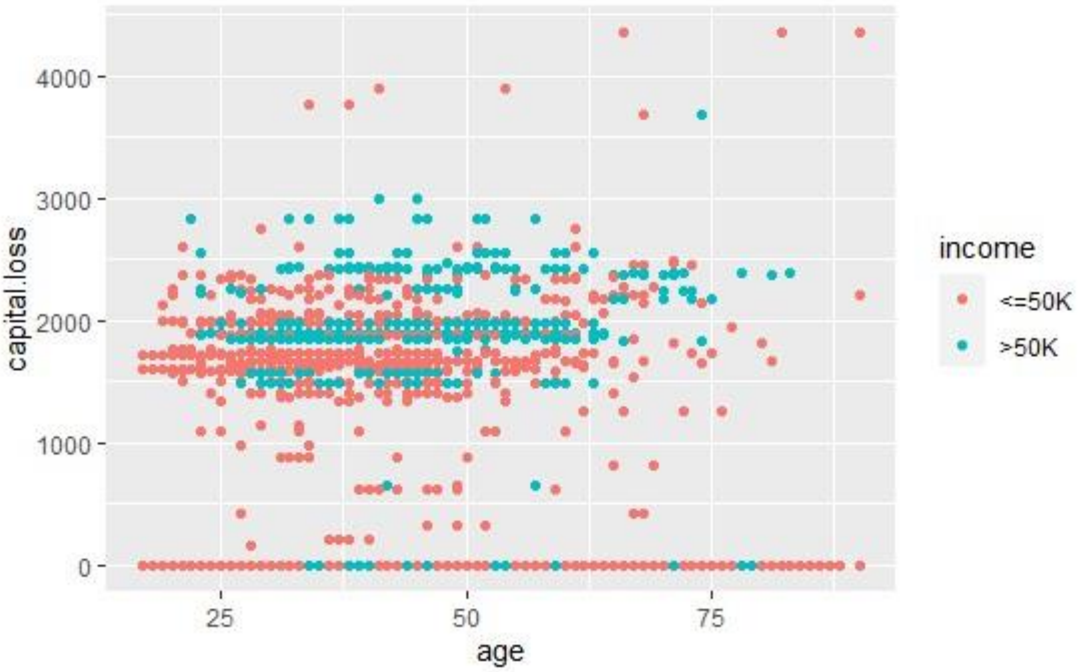
```
ggplot(df,aes(age, hours.per.week))+geom_point(aes(color=income))
```



Majority of people with working hours <50 per week, earn <=50K. People who earn >50K, work >=35hrs per week.

Hide

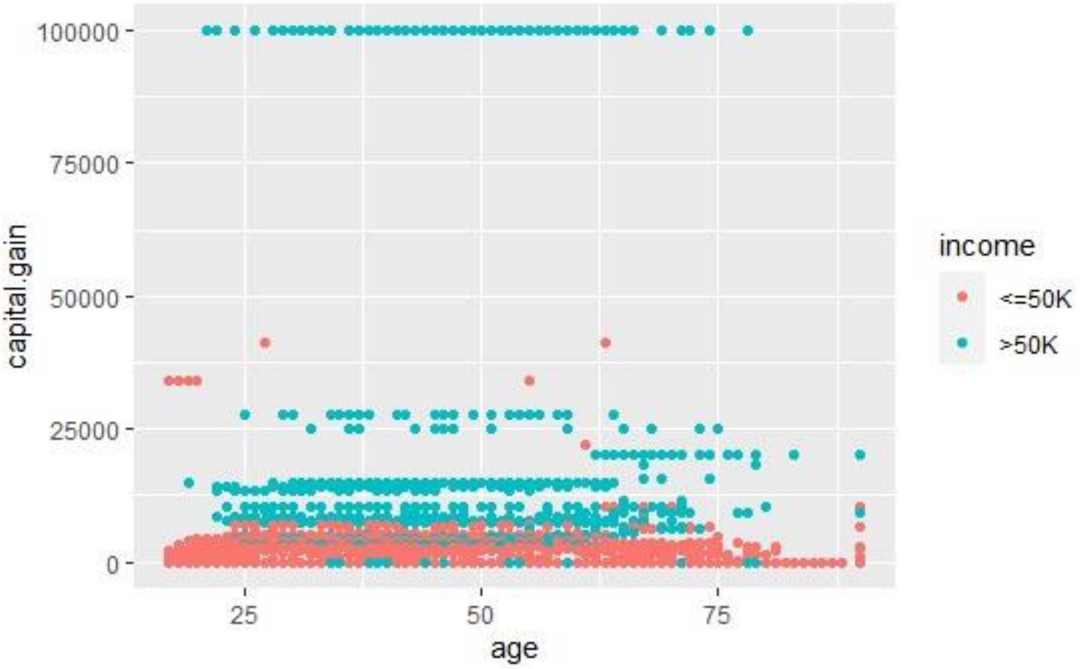
```
ggplot(df,aes(age, capital.loss))+geom_point(aes(color=income))
```



There are outliers. Citizens with income >50K income have more captial loss

Hide

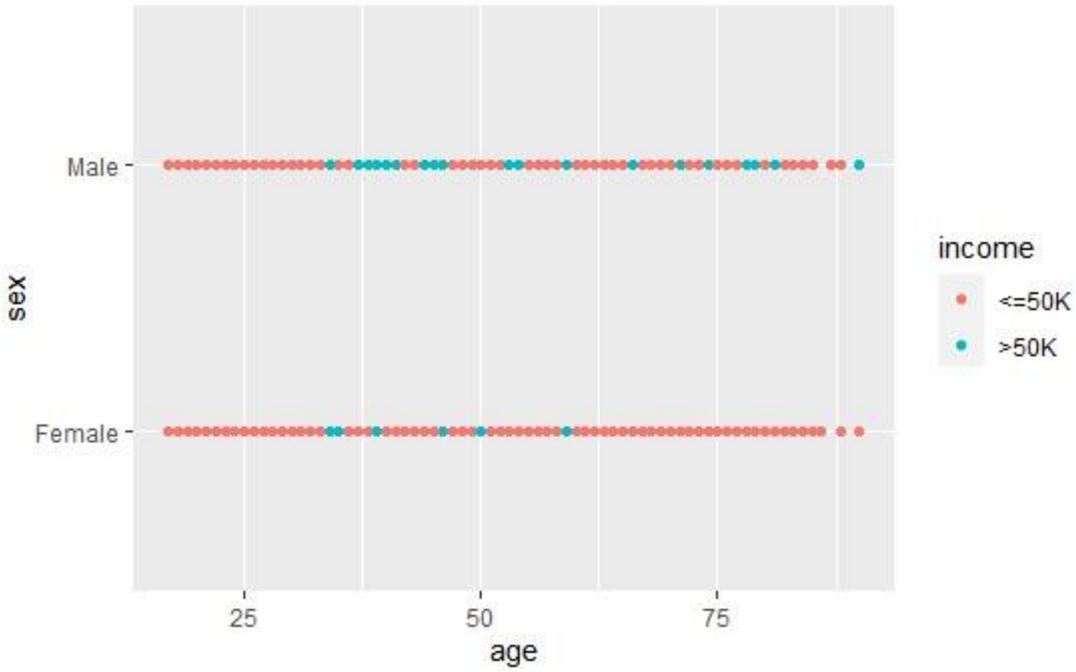
```
ggplot(df,aes(age, capital.gain))+geom_point(aes(color=income))
```



Citizens with income <=50K have nerly no capital gain

Hide

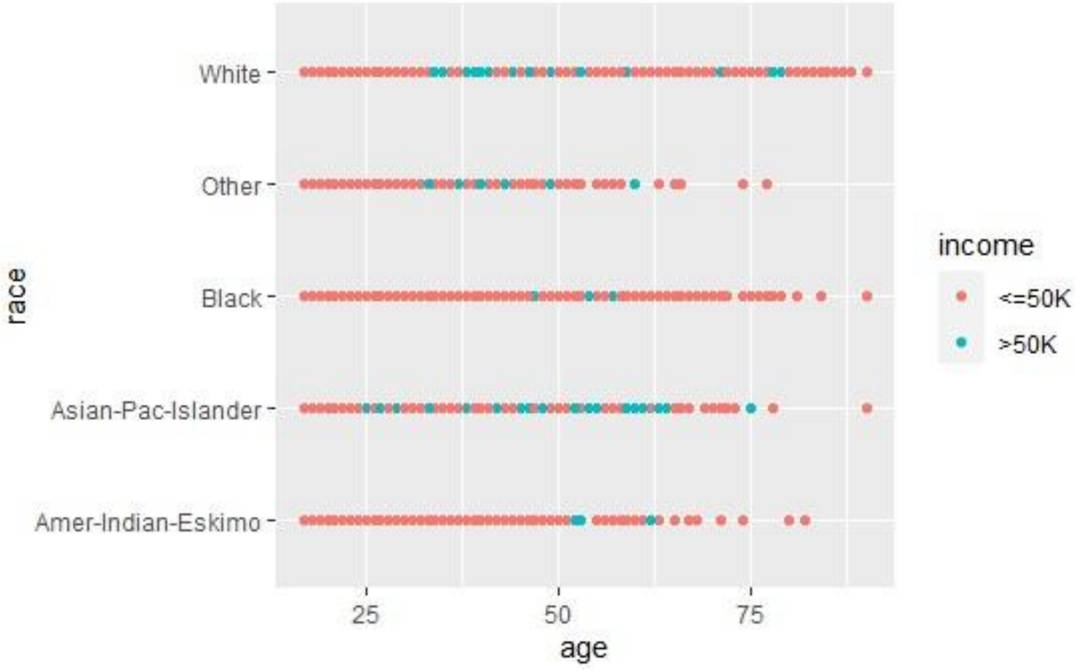
```
ggplot(df,aes(age, sex))+geom_point(aes(color=income))
```



Men with age >25 have income >50K. Most of the female citizens have income<=50K. Men earn more in all countries

Hide

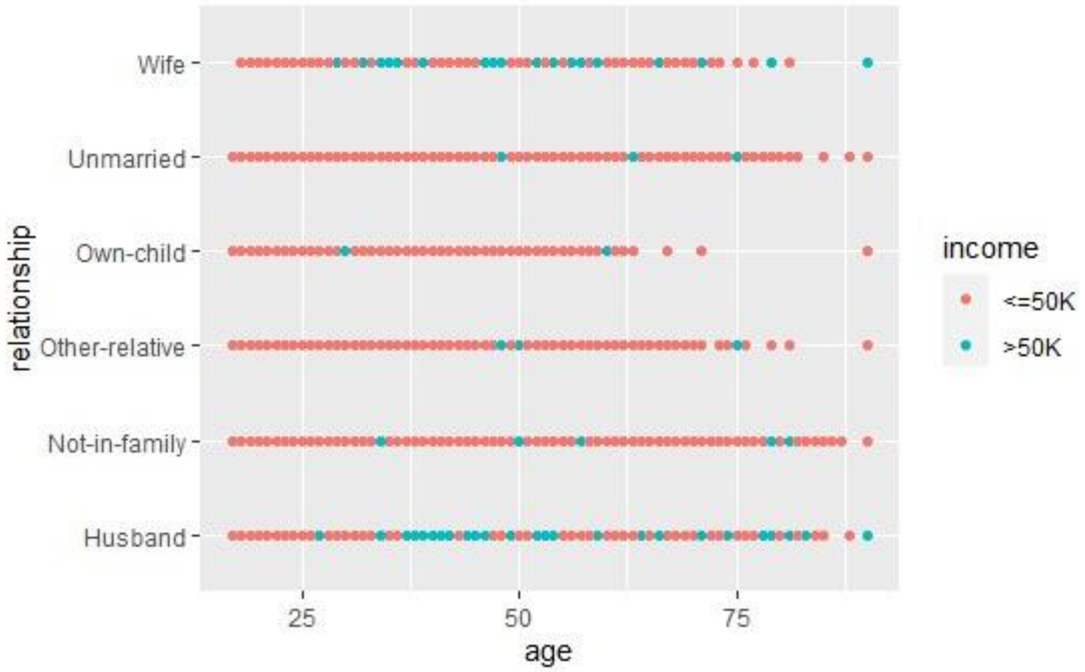
```
ggplot(df,aes(age, race))+geom_point(aes(color=income))
```



Only few black citizens earn >50K. White and Asian-Pac-Islander have some citizens out of all other races who earb >50K

Hide

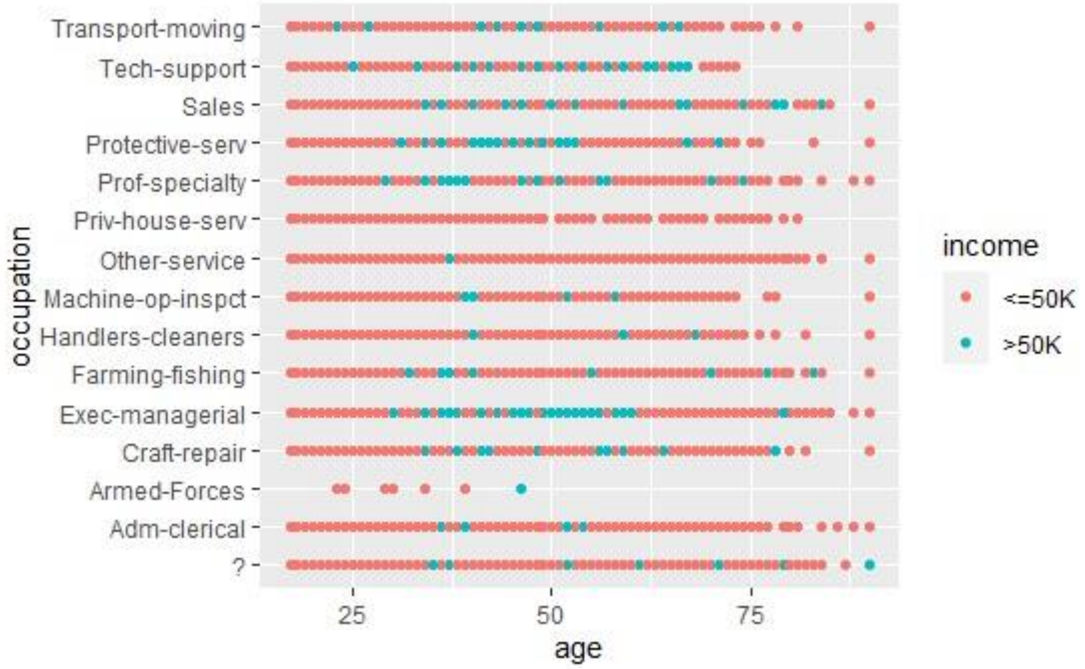
```
ggplot(df,aes(age, relationship))+geom_point(aes(color=income))
```



Mostly Husbands and Wives earn >50K. Others mostley earn <=50K.

Hide

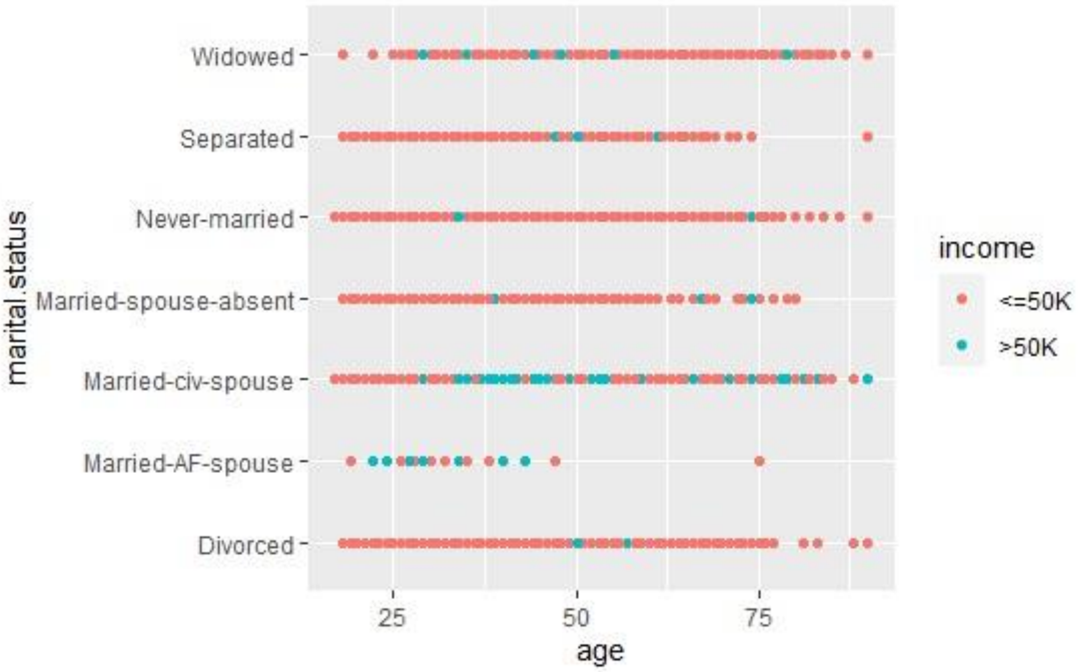
```
ggplot(df,aes(age, occupation))+geom_point(aes(color=income))
```



Citizens with more experience in occupations earn >50K. Services and Armed forces have an income <=50K

Hide

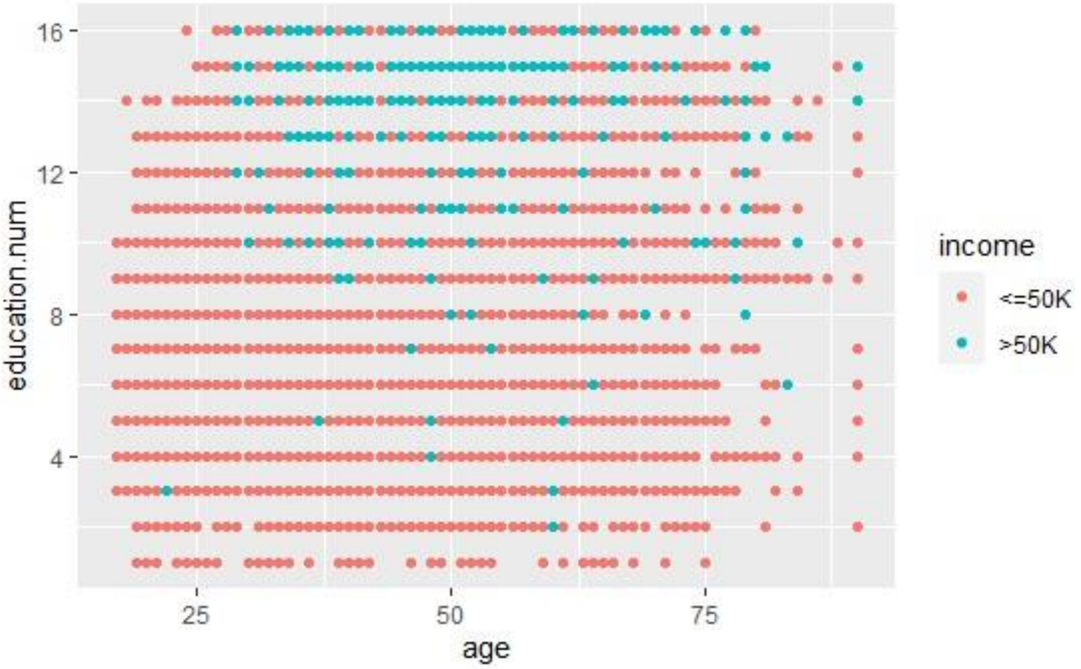
```
ggplot(df,aes(age, marital.status))+geom_point(aes(color=income))
```



Married citizens have income >50K. Divorced, separated and Windowed citizens earn <=50K.

Hide

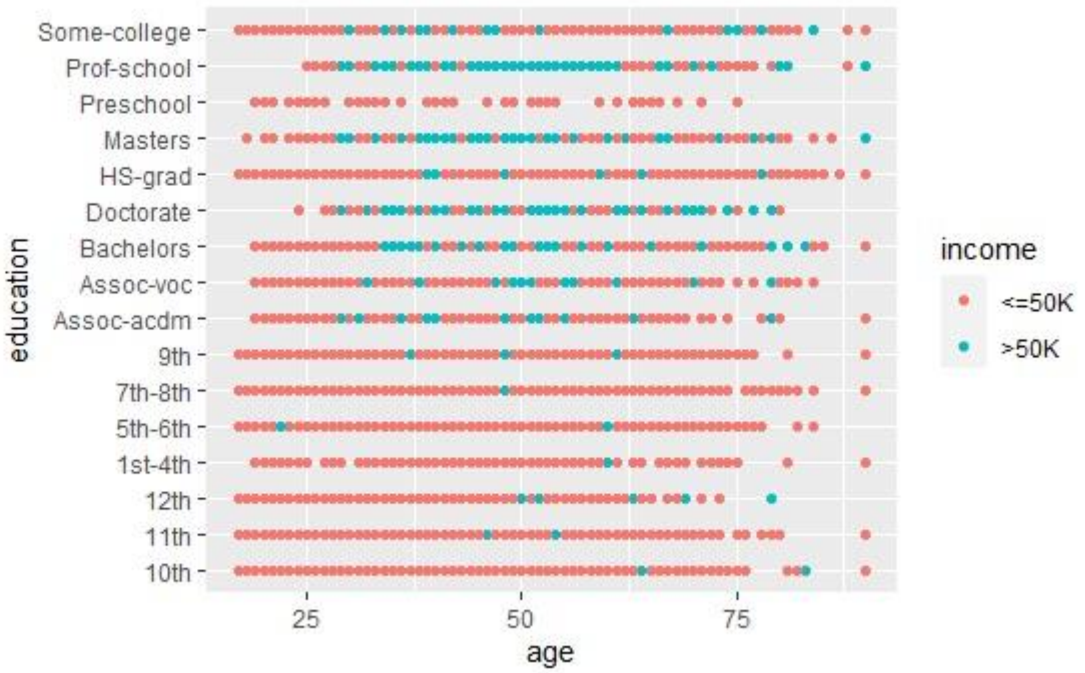
```
ggplot(df,aes(age, education.num))+geom_point(aes(color=income))
```



People with more years of education have income >50K

Hide

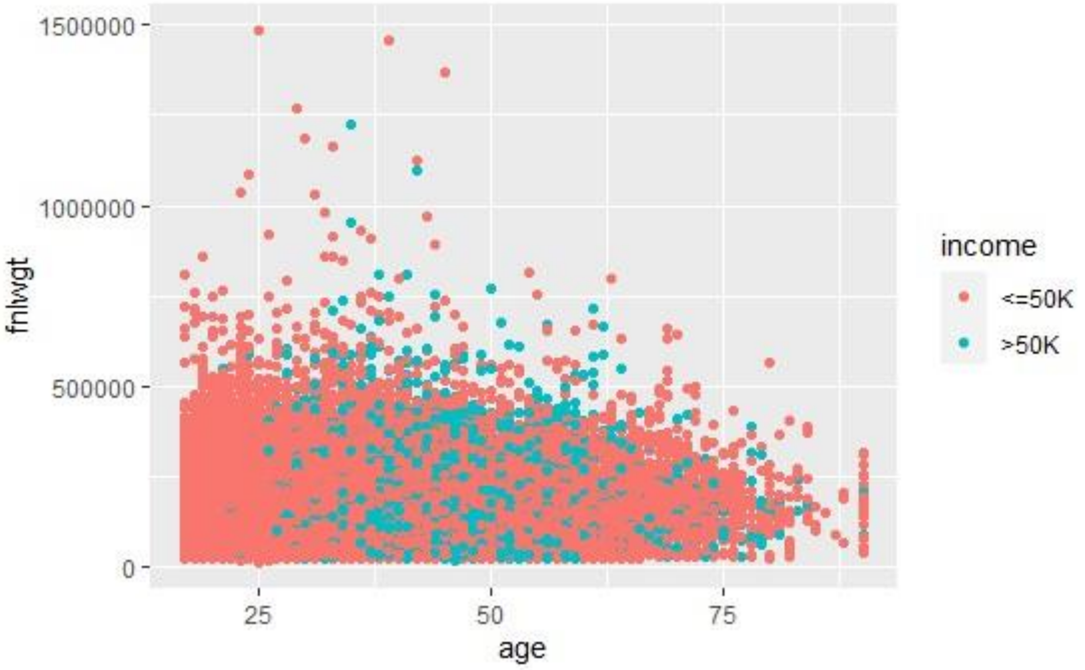
```
ggplot(df,aes(age, education))+geom_point(aes(color=income))
```



Citizens with high qualifications have high income.

Hide

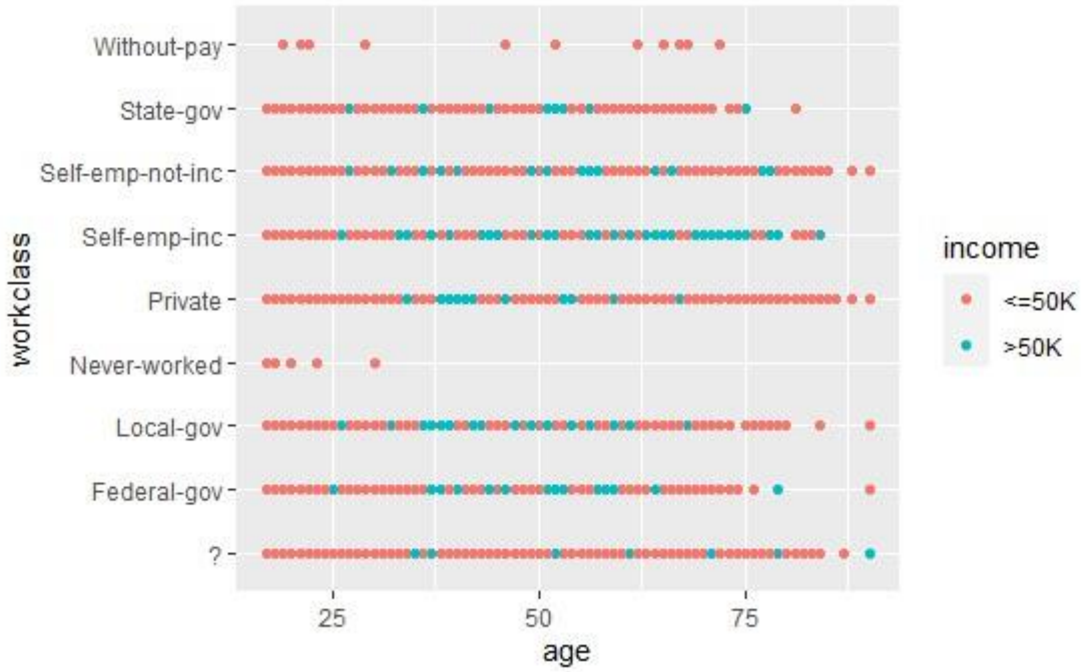
```
ggplot(df,aes(age, fnlwgt))+geom_point(aes(color=income))
```



There re outliers .

Hide

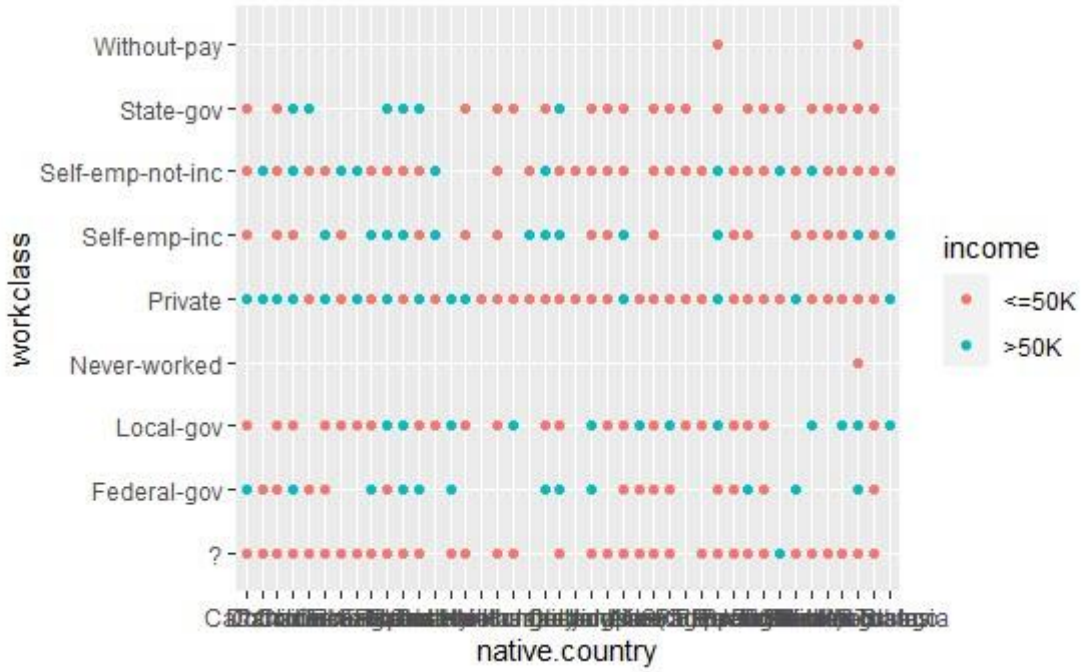
```
ggplot(df,aes(age, workclass))+geom_point(aes(color=income))
```



Citizens with workclass of Local gov, self employed and Private earn >=50K

Hide

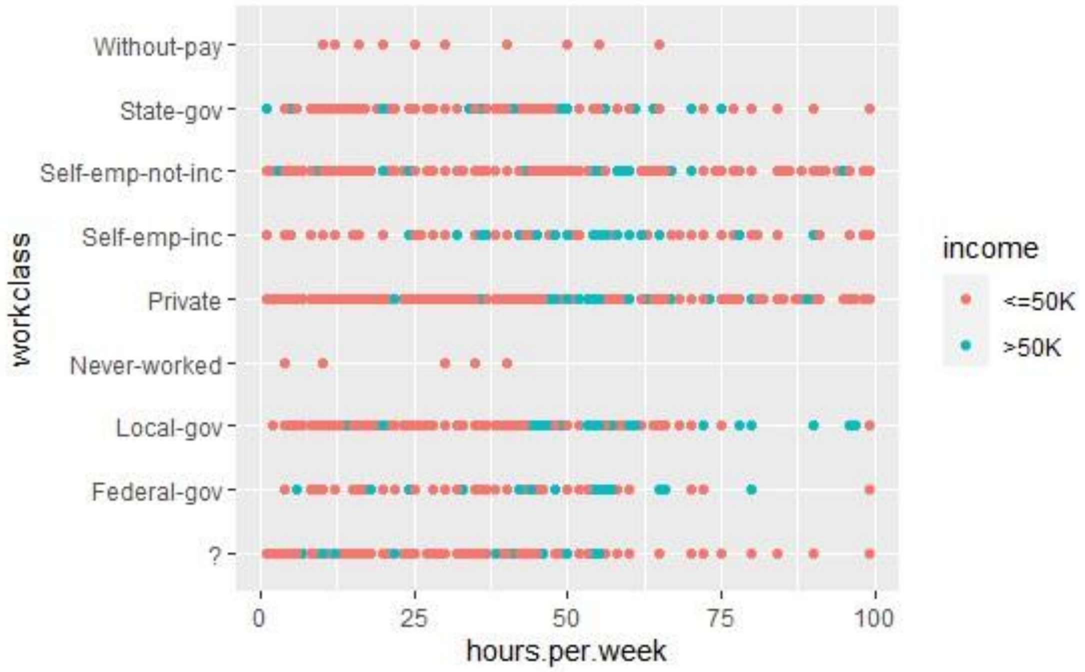
```
ggplot(df,aes( native.country, workclass ))+geom_point(aes(color=income))
```



citizens who have never worked don’t have income. Citizens with Private workclass have income >50K.

Hide

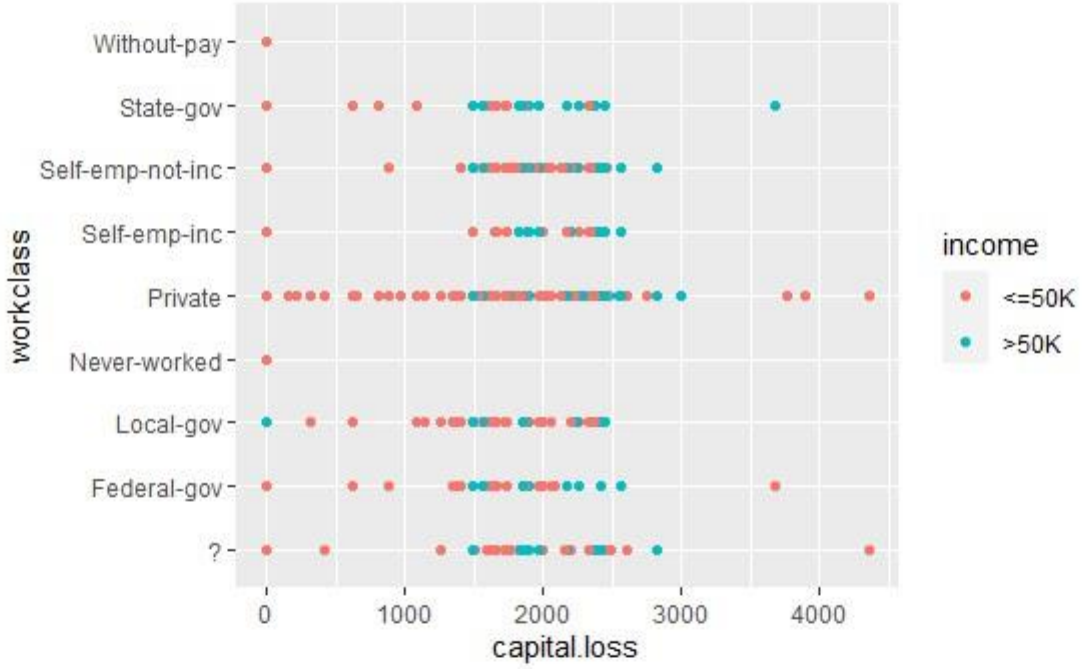
```
ggplot(df,aes( hours.per.week, workclass )) + geom_point(aes(color=income))
```



Private workclss, self employees and local gov workclass citizes with more hours of work per week have income >50K

Hide

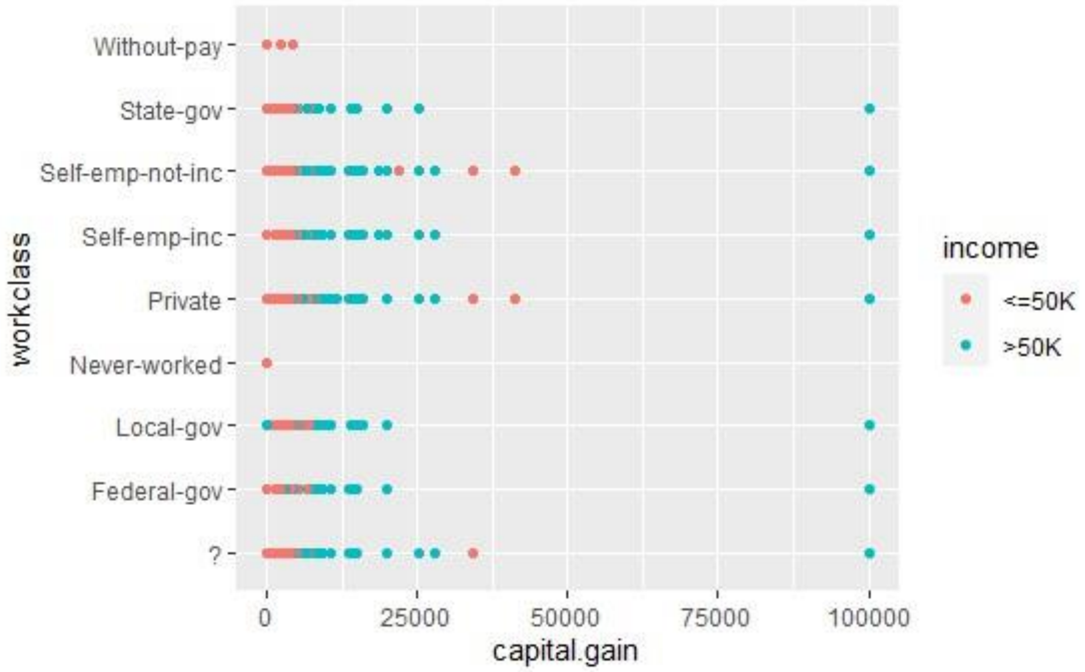
```
ggplot(df,aes(capital.loss, workclass))+geom_point(aes(color=income))
```



Private workclass had more capital loss.

Hide

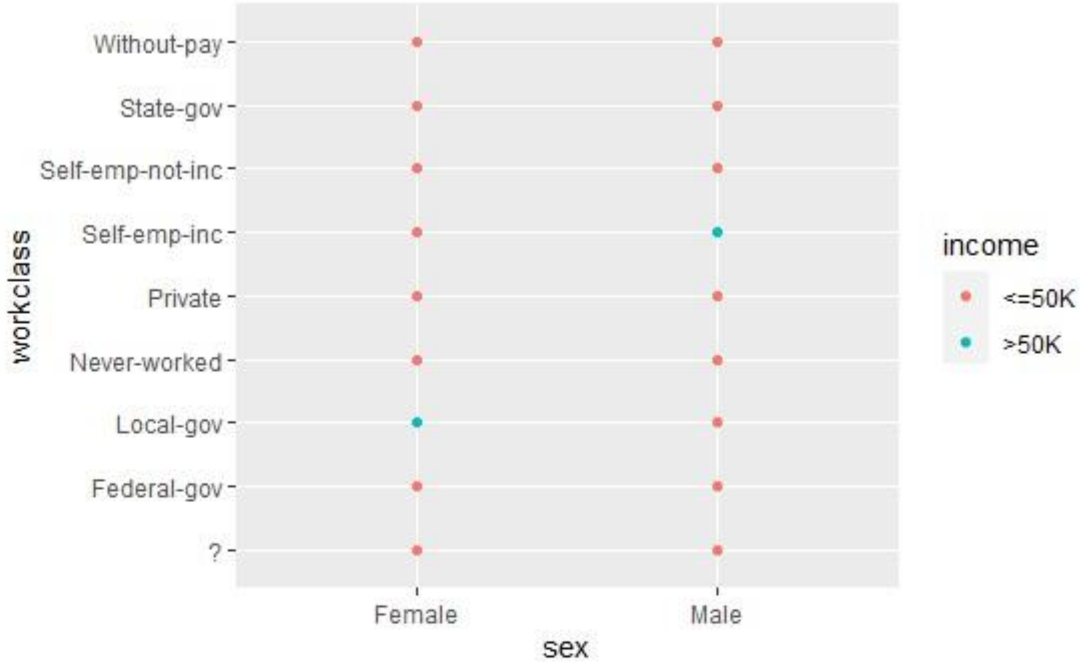
```
ggplot(df,aes(capital.gain, workclass))+geom_point(aes(color=income))
```



Almost all workclass has very less capital gain for citizens with income <=50K and upto 25000 for citizens with income>50K

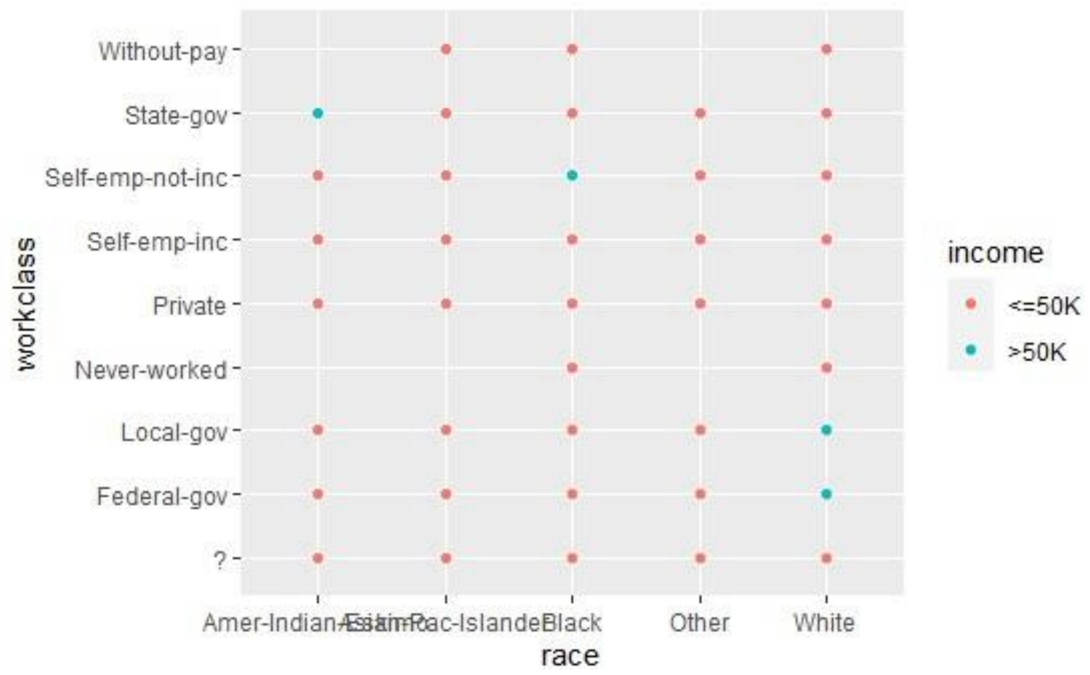
Hide

```
ggplot(df,aes( sex, workclass))+geom_point(aes(color=income))
```



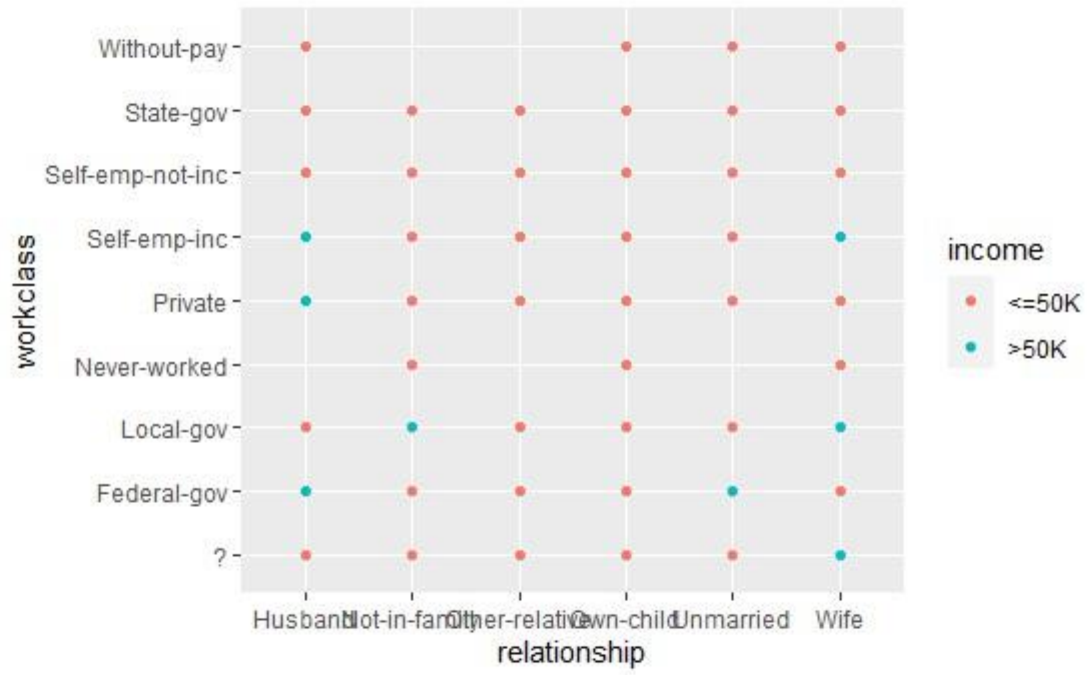
Hide

```
ggplot(df,aes(race, workclass))+geom_point(aes(color=income))
```



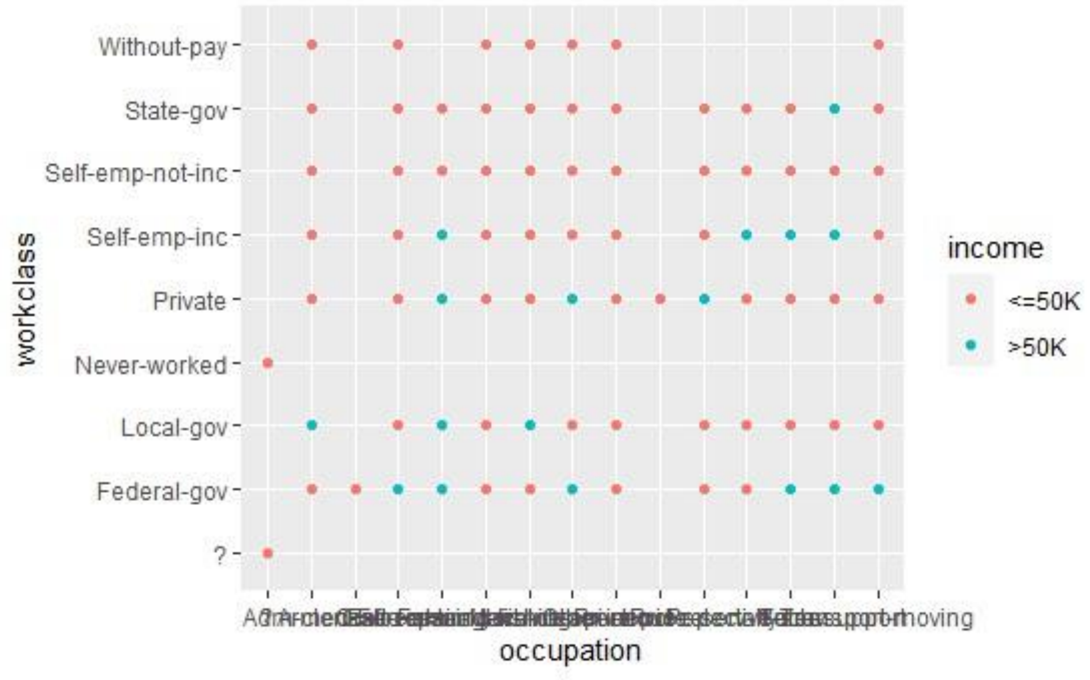
Hide

```
ggplot(df,aes(relationship, workclass))+geom_point(aes(color=income))
```



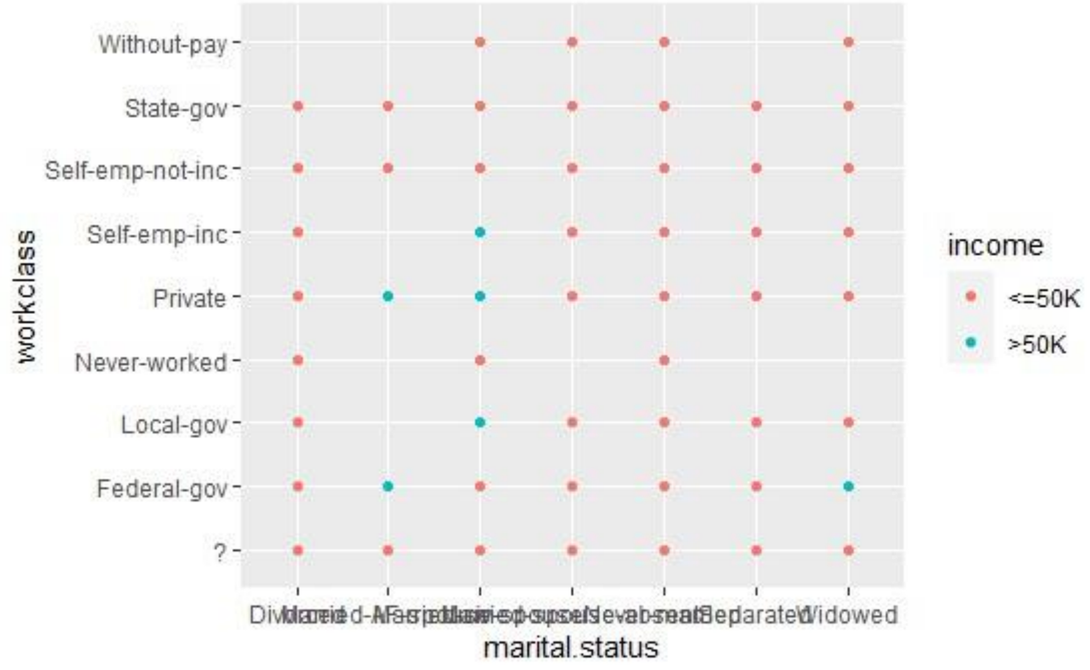
Hide

```
ggplot(df,aes(occupation, workclass))+geom_point(aes(color=income))
```



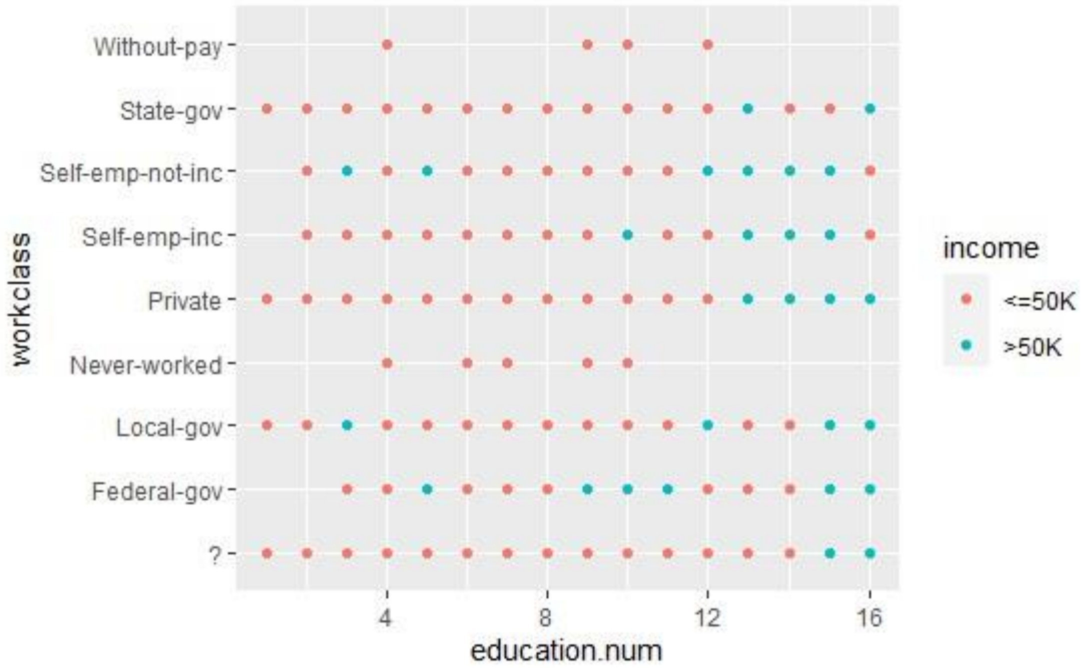
Hide

```
ggplot(df,aes(marital.status, workclass))+geom_point(aes(color=income))
```



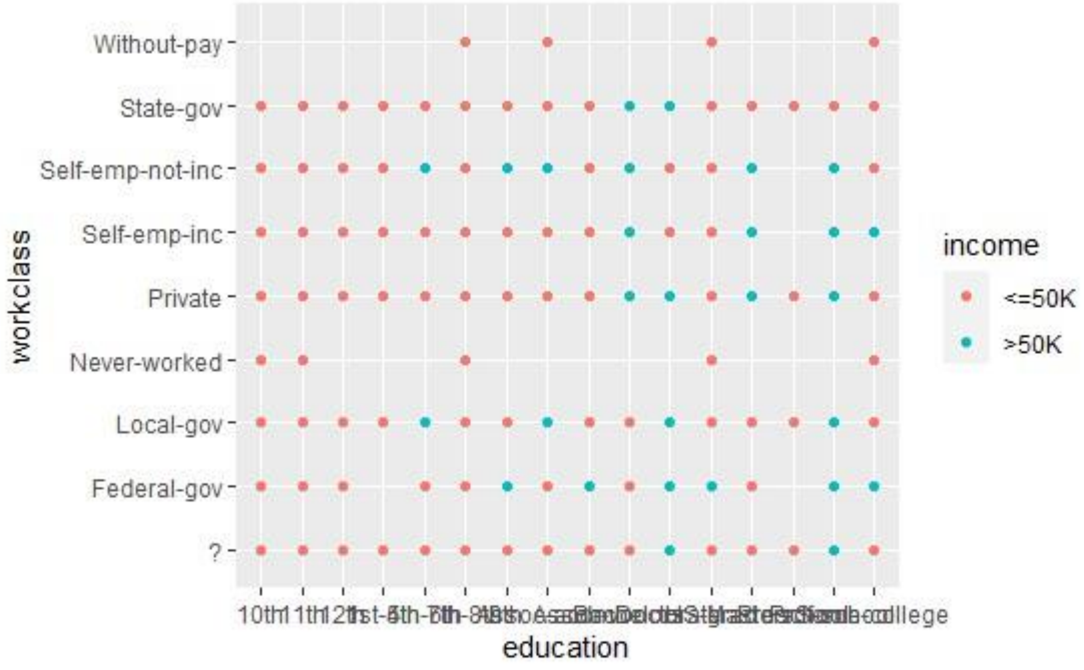
Hide

```
ggplot(df,aes(education.num, workclass))+geom_point(aes(color=income))
```



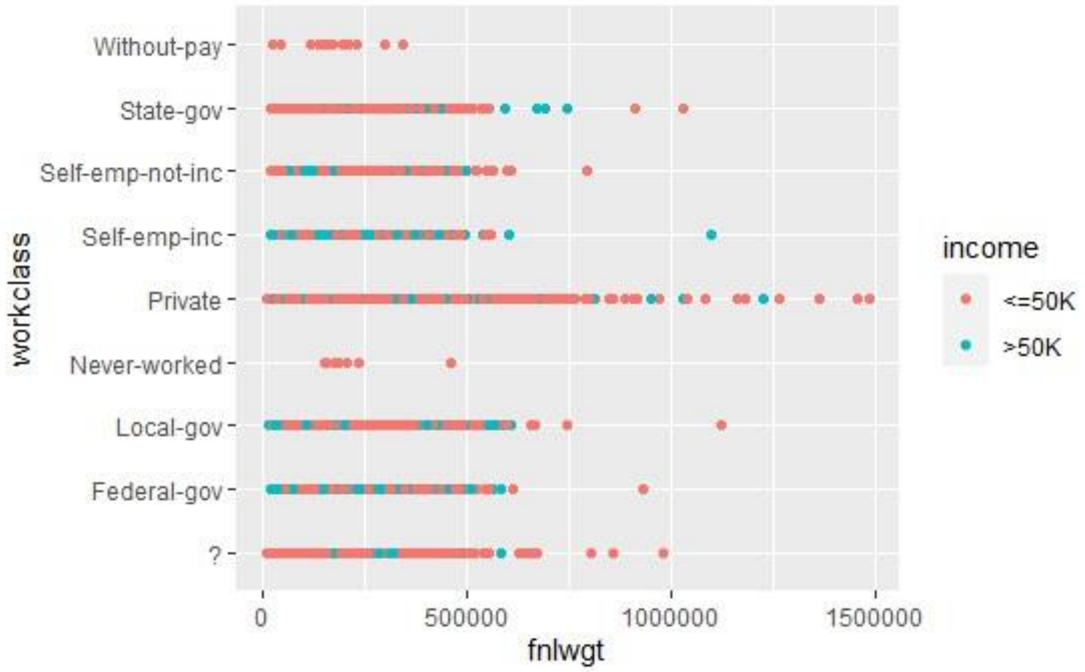
Hide

```
ggplot(df,aes(education, workclass))+geom_point(aes(color=income))
```



Hide

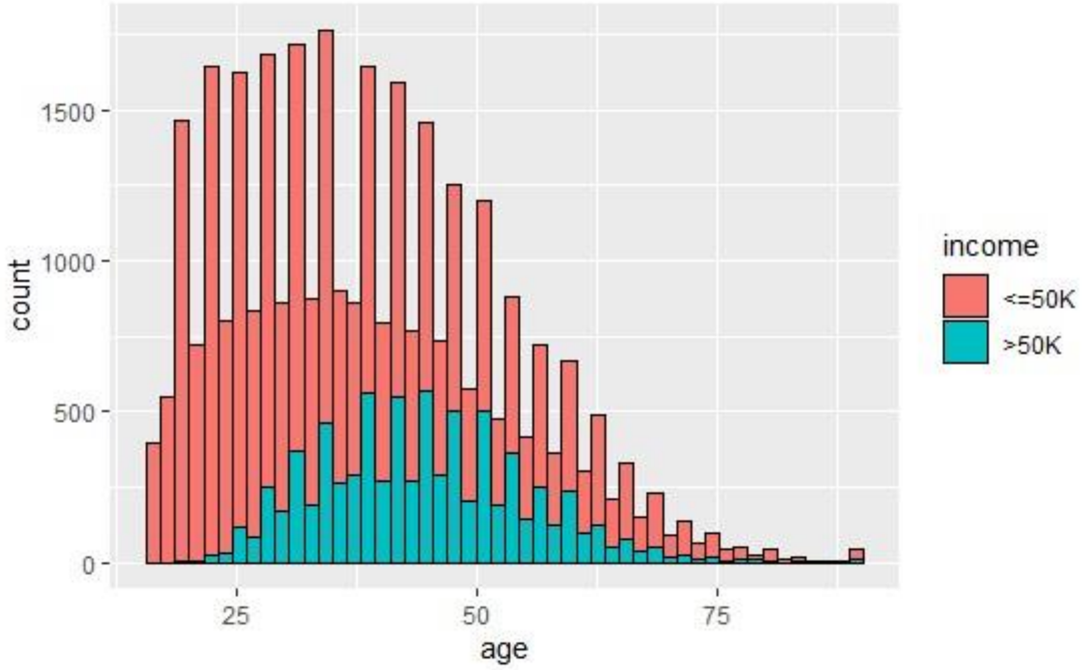
```
ggplot(df,aes( fnlwgt, workclass))+geom_point(aes(color=income))
```



Private workclass has more fnlwgt

Hide

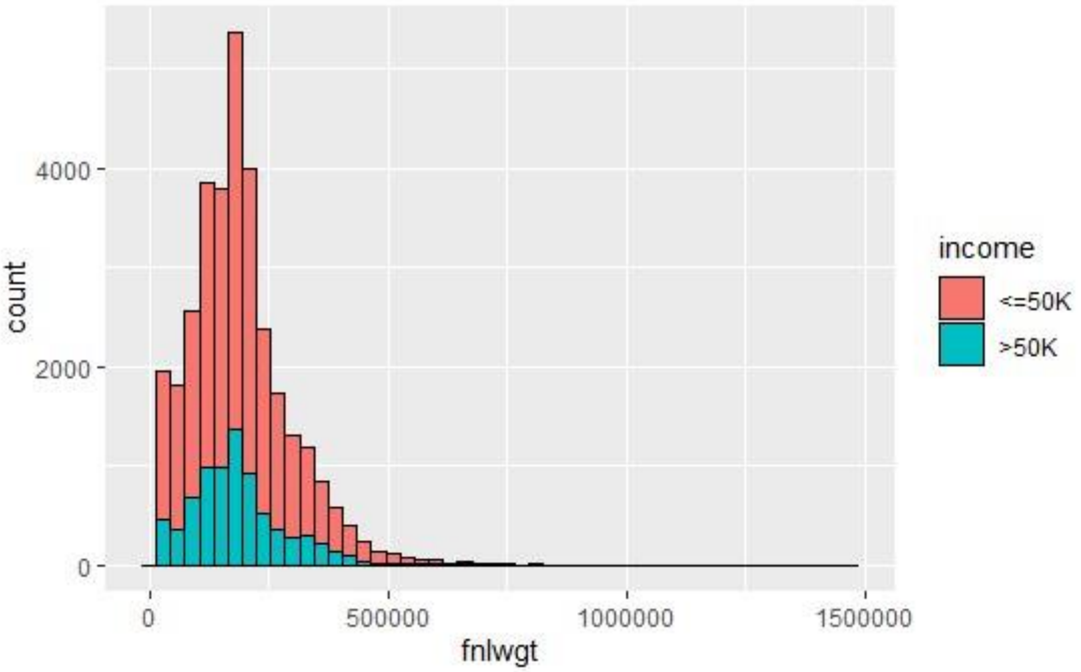
```
ggplot(df,aes(age))+geom_histogram(aes(fill=income),color ='black',bins = 50)
```



Majority of people earn <=50K.

Hide

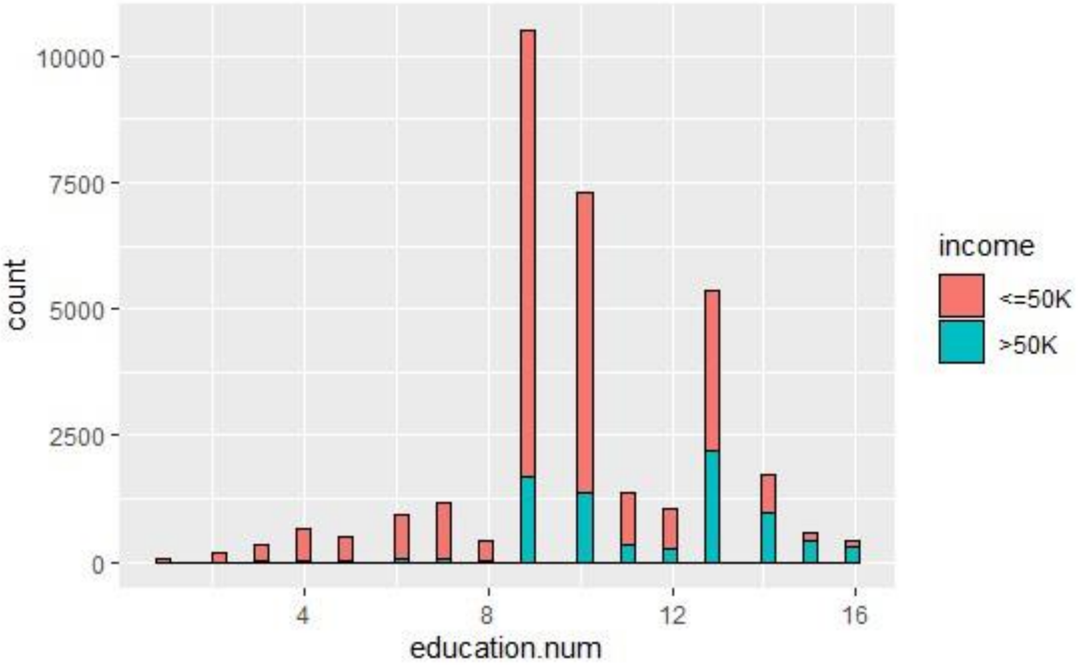
```
ggplot(df,aes(fnlwgt))+geom_histogram(aes(fill=income),color ='black',bins = 50)
```



Final weight determined by census org, ranges 0-500000. Citizens with income<=50K are more

Hide

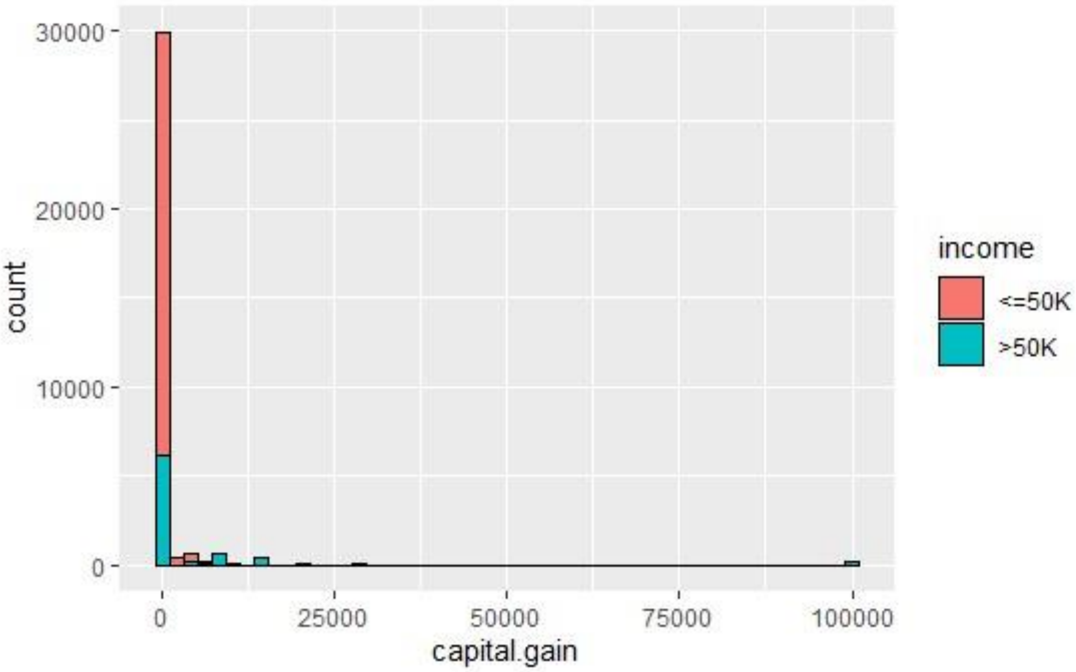
```
ggplot(df,aes(education.num))+geom_histogram(aes(fill=income),color ='black',bins = 50)
```



Majority of people who have 9-10 years of education earn <=50K

Hide

```
ggplot(df,aes(capital.gain))+geom_histogram(aes(fill=income),color ='black',bins = 50)
```



Majority of citizens who have income <=50K don’t have a capital gain

TRAIN AND TEST OF MODEL

Hide

```
library(caTools)
```

```
package 㤘牠caTools㤘牸 was built under R version 4.0.4
```

Hide

```
set.seed(100)
sample = sample.split(df$income, SplitRatio =0.70)
train = subset(df,sample ==TRUE)
test = subset(df,sample ==FALSE)
```

Hide

```
library(rpart)
library(rpart.plot)
tree <- rpart(income~., method ='class',data=train)
```

Hide

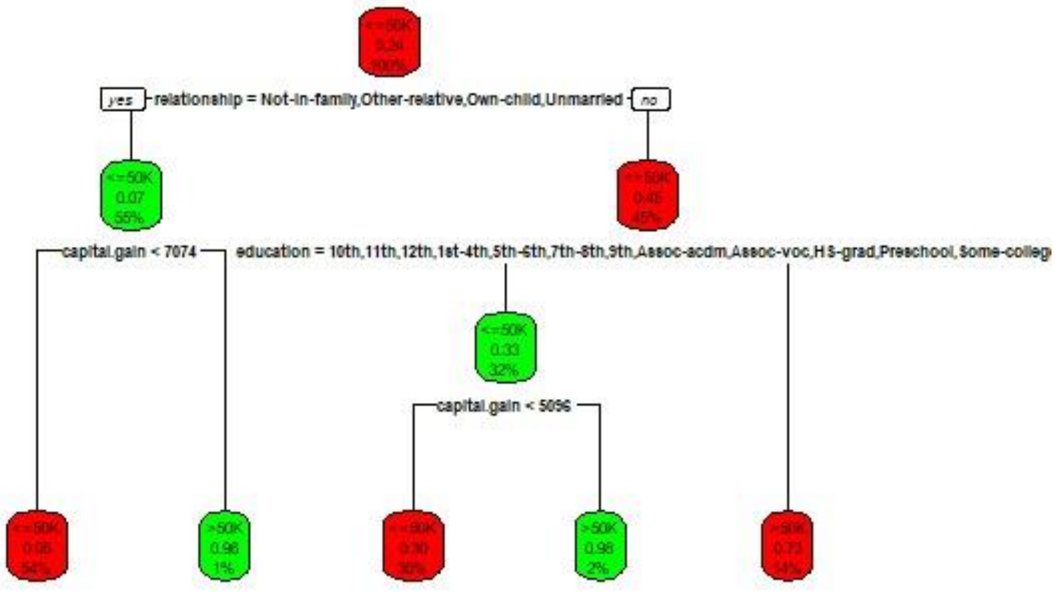
```
tree.preds <- predict(tree,test)
head(tree.preds)
```

```
       <=50K       >50K
7  0.9509564 0.04904365
10 0.9509564 0.04904365
11 0.9509564 0.04904365
13 0.9509564 0.04904365
19 0.9509564 0.04904365
21 0.9509564 0.04904365
```

Plotting

Hide

```
rpart.plot(tree,box.col=c('red','green'))
```


<=50K
0.24
100%
yes
relationship = Not-in-family,Other-relative,Own-child,Unmarried
no
<=50K
0.07
55%
<=50K
0.46
45%
capital.gain < 7074
education = 10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,HS-grad,Preschool,Some-colleg
<=50K
0.33
32%
capital.gain < 5096
<=50K
0.05
54%
>50K
0.98
1%
<=50K
0.30
30%
>50K
0.98
2%
>50K
0.73
14%

Hide

```
tree.preds <- as.data.frame(tree.preds)
joiner <- function(x){
  if(x>0.5){ #threshold value
    return('>50K')
  } else{
    return('<=50K')
    }
}
```

If the values in X is > 0.5, then >50K, else <=50K

Hide

```
tree.preds
```

	<=50K	>50K
	<dbl>	<dbl>
7	0.95095635	0.04904365
10	0.95095635	0.04904365
11	0.95095635	0.04904365
13	0.95095635	0.04904365
19	0.95095635	0.04904365
21	0.95095635	0.04904365
25	0.69915501	0.30084499
28	0.69915501	0.30084499
30	0.95095635	0.04904365
31	0.95095635	0.04904365

1-10 of 9,768 rows Previous **1** 2 3 4 5 6 ... 100 Next

Hide

```
tree.preds$income <- sapply(tree.preds$`>50K`, joiner)
head(tree.preds)
```

	<=50K	>50K	income
	<dbl>	<dbl>	<chr>
7	0.9509564	0.04904365	<=50K
10	0.9509564	0.04904365	<=50K
11	0.9509564	0.04904365	<=50K
13	0.9509564	0.04904365	<=50K
19	0.9509564	0.04904365	<=50K
21	0.9509564	0.04904365	<=50K

6 rows

FOR VALIDATION, WE USE
CONFUSION MATRIX

Hide

```
library(caret)
```

```
package 惐栬caret惐砟 was built under R version 4.0.5Loading required package: lattice
Registered S3 method overwritten by 'data.table':
  method           from
  print.data.table
```

Hide

```
cf <- table(tree.preds$income, test$income)
confusionMatrix(cf,positive='>50K')
```

```
Confusion Matrix and Statistics


        <=50K >50K
  <=50K  7040 1151
  >50K    376 1201

               Accuracy : 0.8437
                 95% CI : (0.8363, 0.8508)
    No Information Rate : 0.7592
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5182

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5106
            Specificity : 0.9493
         Pos Pred Value : 0.7616
         Neg Pred Value : 0.8595
             Prevalence : 0.2408
         Detection Rate : 0.1230
   Detection Prevalence : 0.1614
      Balanced Accuracy : 0.7300

       'Positive' Class : >50K
```

Accuracy is 84.37%. The confidence interval at 95% is (0.8363, 0.8508)