Q1.Classify the nine attributes as one or more of nominal, binary, ordinal, interval, and ratio. Provide justification.
Age : Ratio because age can be 0.
Work : Nominal because there various unordered types of it.
Edu : Nominal because it has various unordered types
Marital : Nominal because it has various unordered types
Occupation : Nominal because it has various unordered types
Race : Nominal because it has various unordered types
Sex : Binary because it has various unordered types
Hrs_per_week : Ratio because no of hrs per week can be zero.
Income : Binary because it has two categories ie >50k and <=50k

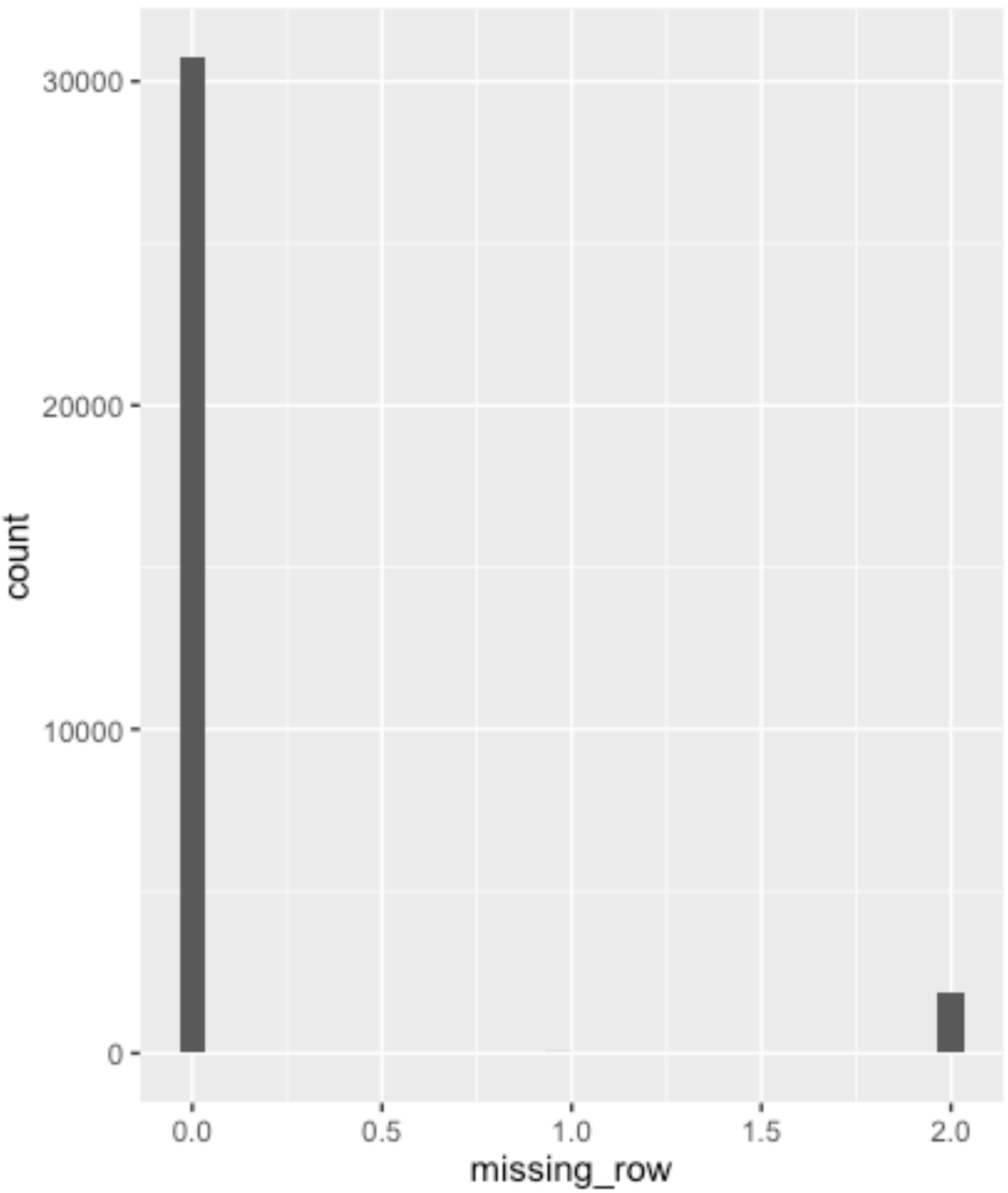
Q2.

1. Missing values in the data are marked with "?"

(a) For each attribute, compute the percentage missing values.

```
$age
[1] 0

$work
[1] 5.638647

$edu
[1] 0

$marital
[1] 0

$occupation
[1] 5.660146

$race
[1] 0

$sex
[1] 0

$hrs_per_week
[1] 0

$income
[1] 0
```

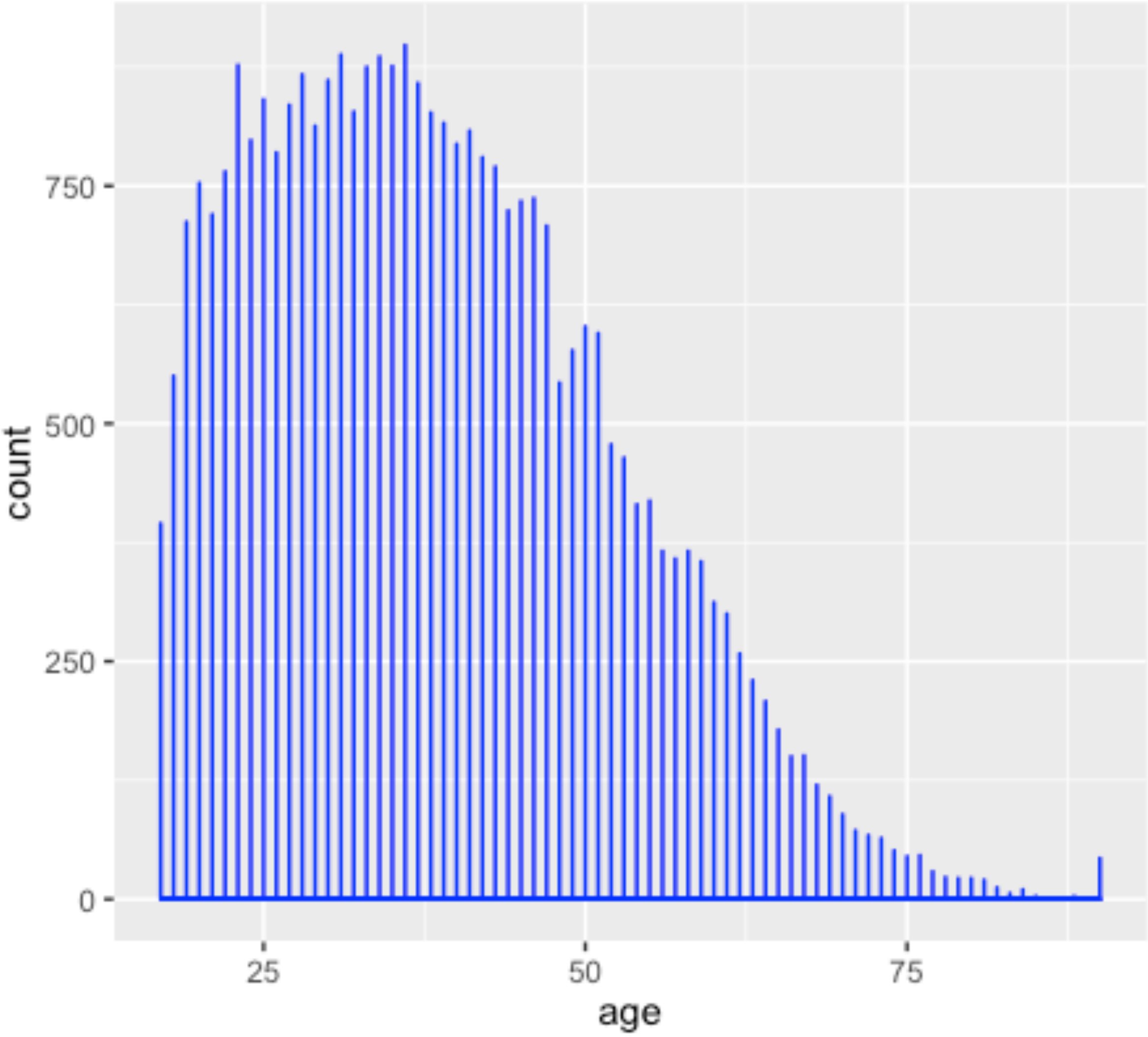
(b) Plot a histogram for the number of missing values per data point (row). (That is, plot the frequency of points (rows) with 1,2,3,4... missing values)



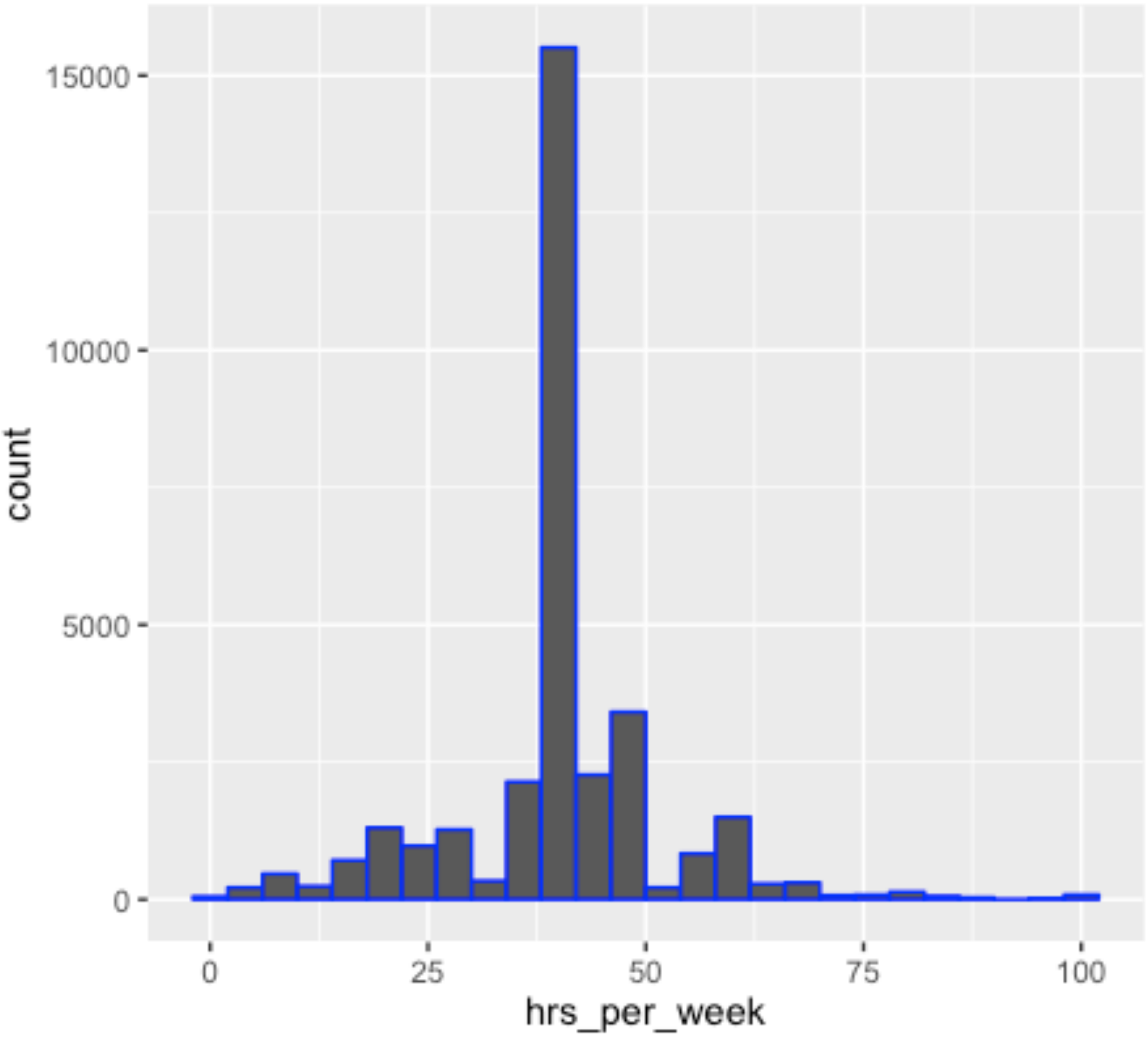
Q3 a)

For each numeric variable, plot a histogram

```
cd + geom_histogram(aes(x=age),colour = "blue",binwidth = 0.01)
```



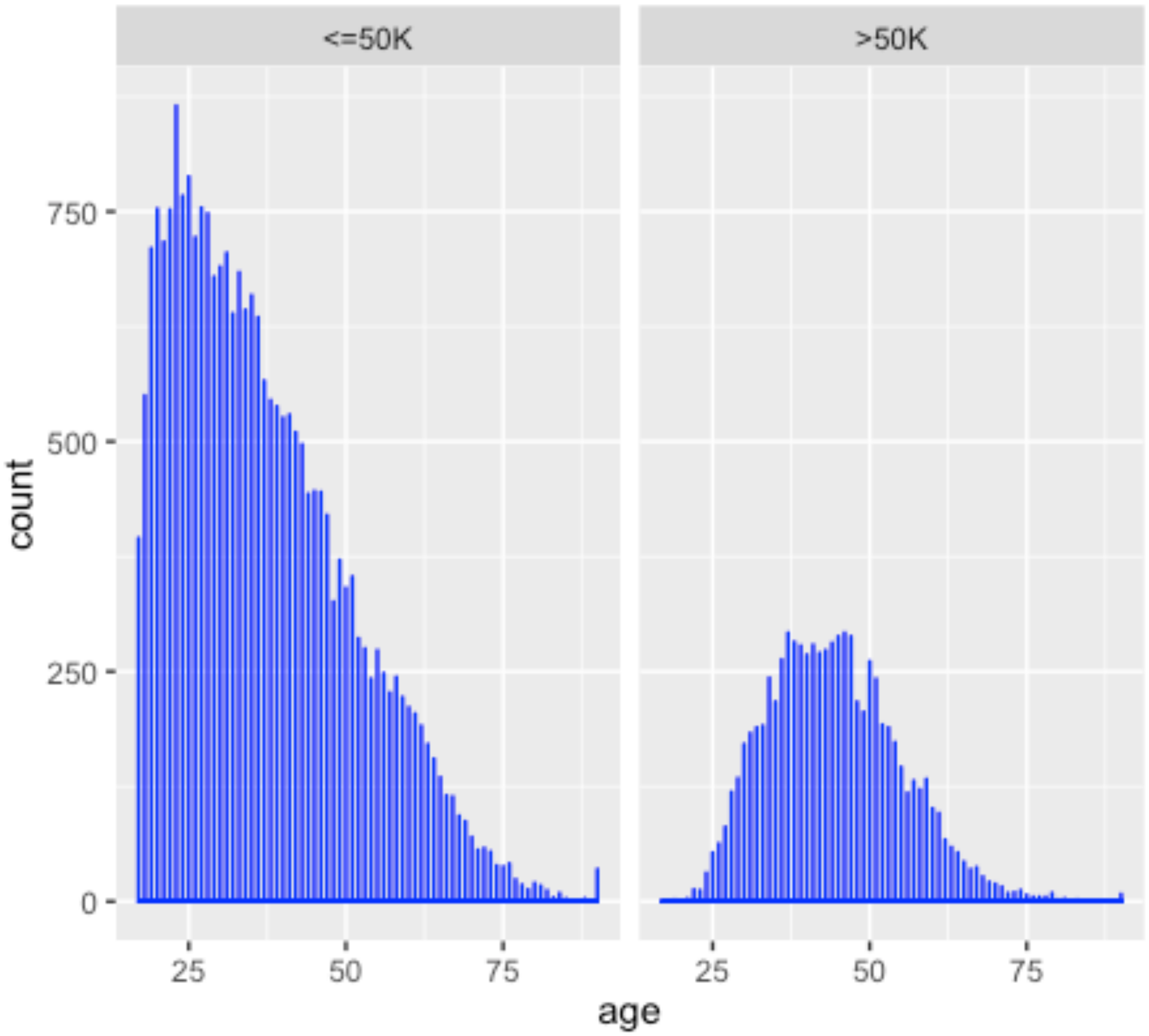
Age vs count plot is positively skewed. We can see that maximum number of people are between 20 to 55 age group and the number of employees decreases after the age of 55.



Count vs hrs_per_week plot shows that maximum people work for 25-50 hrs per week and most of them between 30-35 hrs per week.

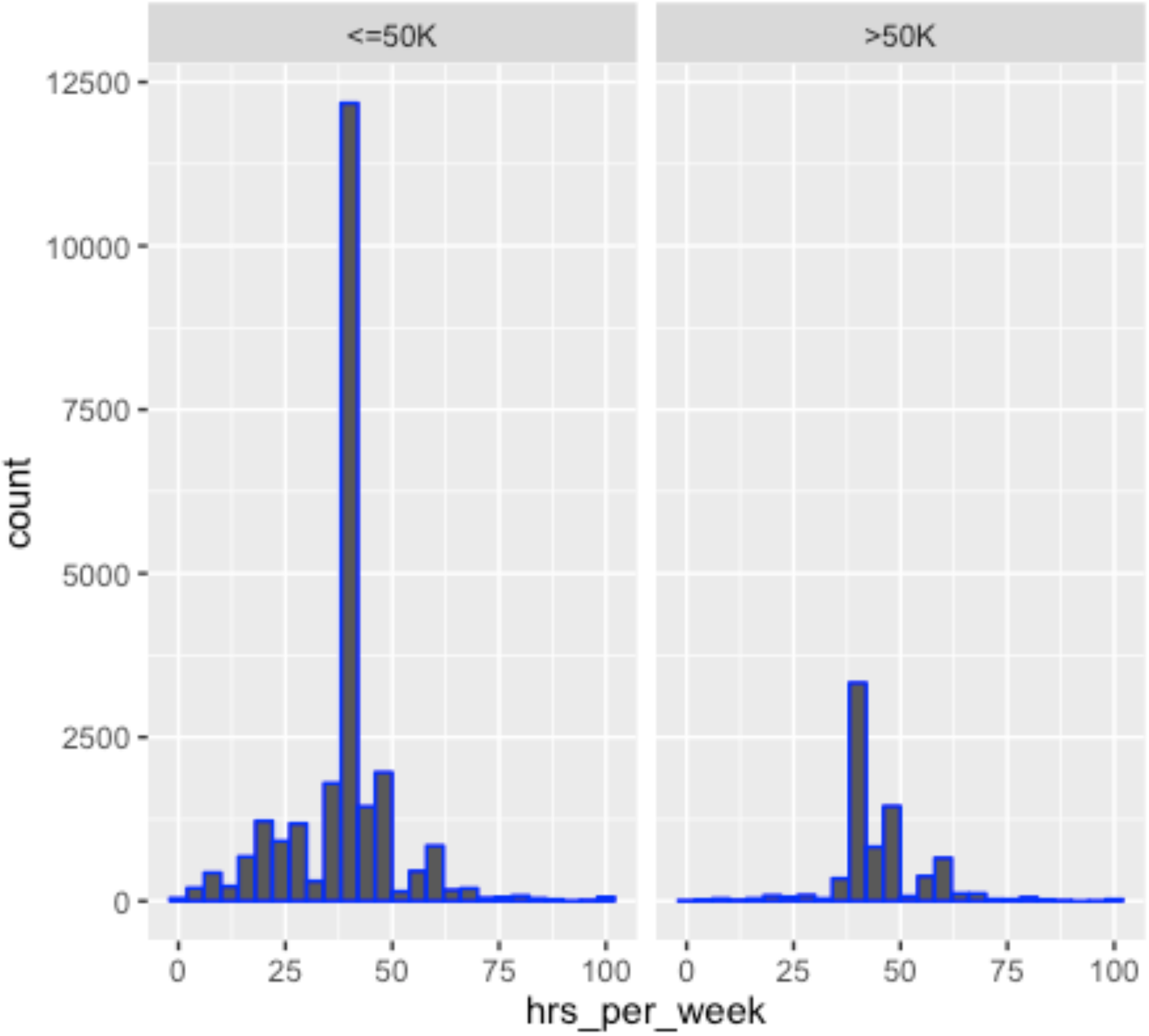
For each numeric variable, plot two histograms (with the same scales, so they can be compared) for the two values of income (<= 50K and > 50K)

```
cd + geom_histogram(aes(x=age),binwidth = 0.01,color="blue")+facet_wrap(~income)
```



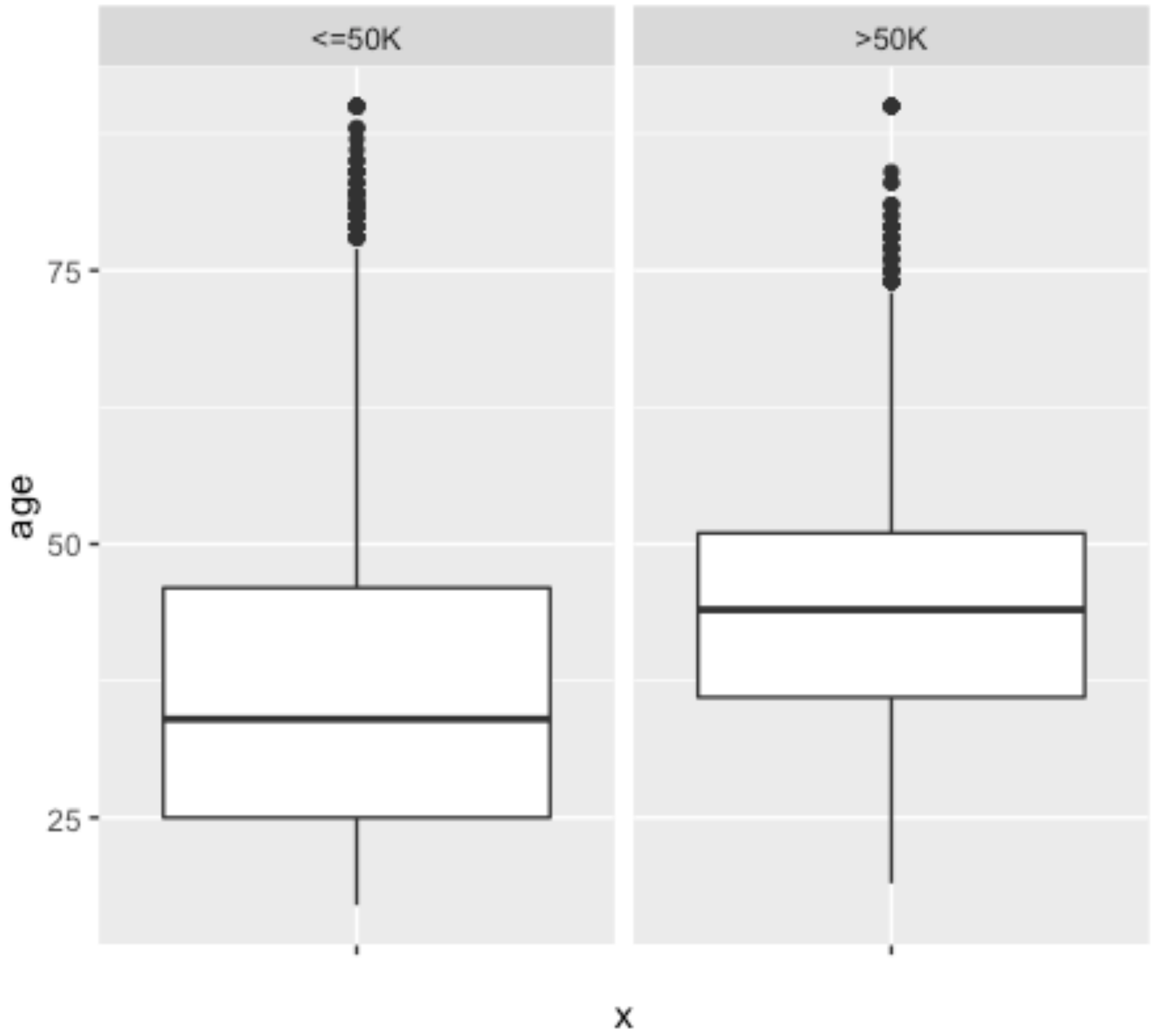
cd + geom_histogram(aes(x=hrs_per_week),binwidth = 4,color="blue")+facet_wrap(~income)

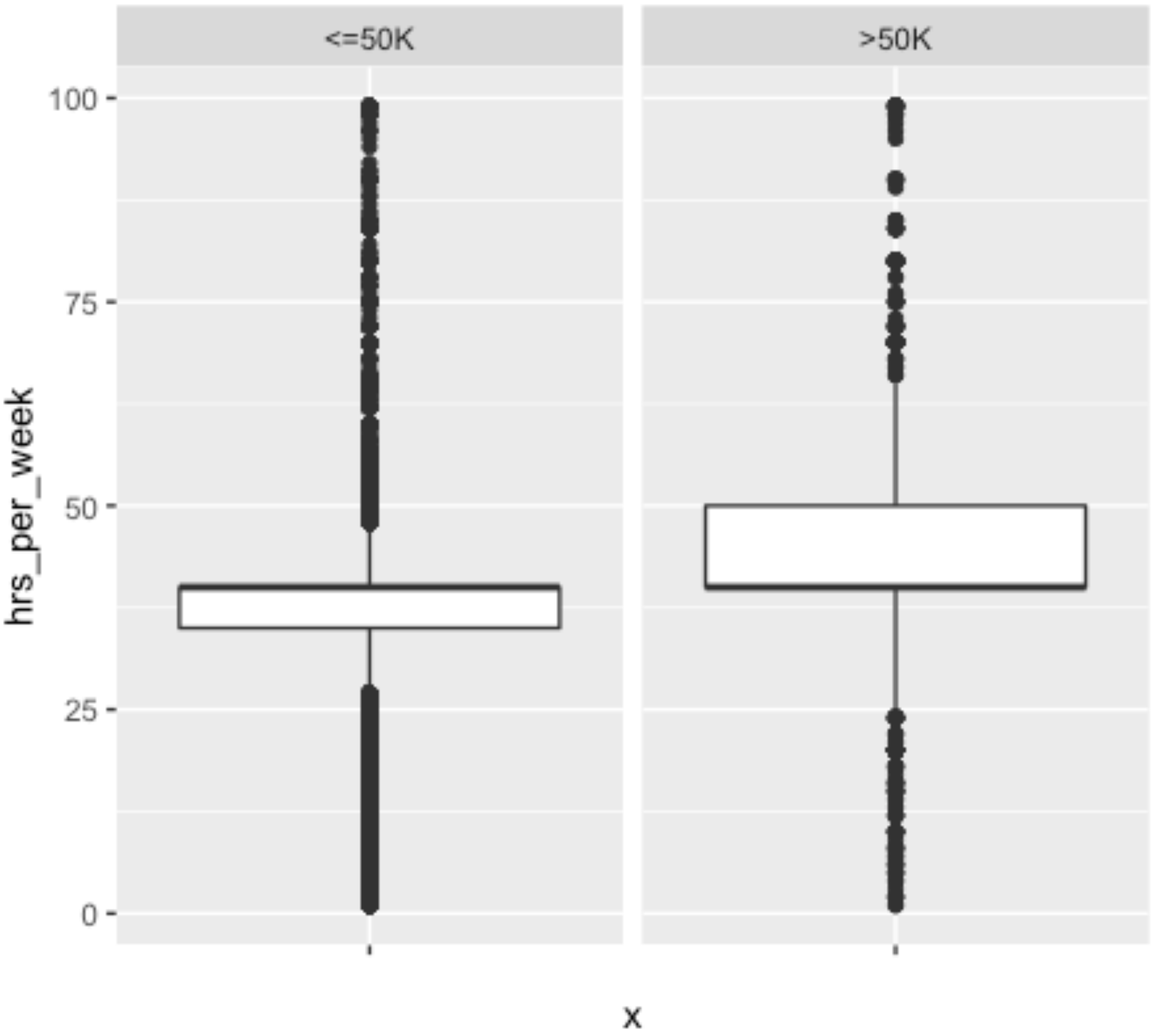
The histogram of age vs count for income of <=50 k is positively skewed whereas the histogram of age vs count for income >=50k is normally distributed.



Hrs_per_week vs count for income >=50k and < 50 k is normally distributed.

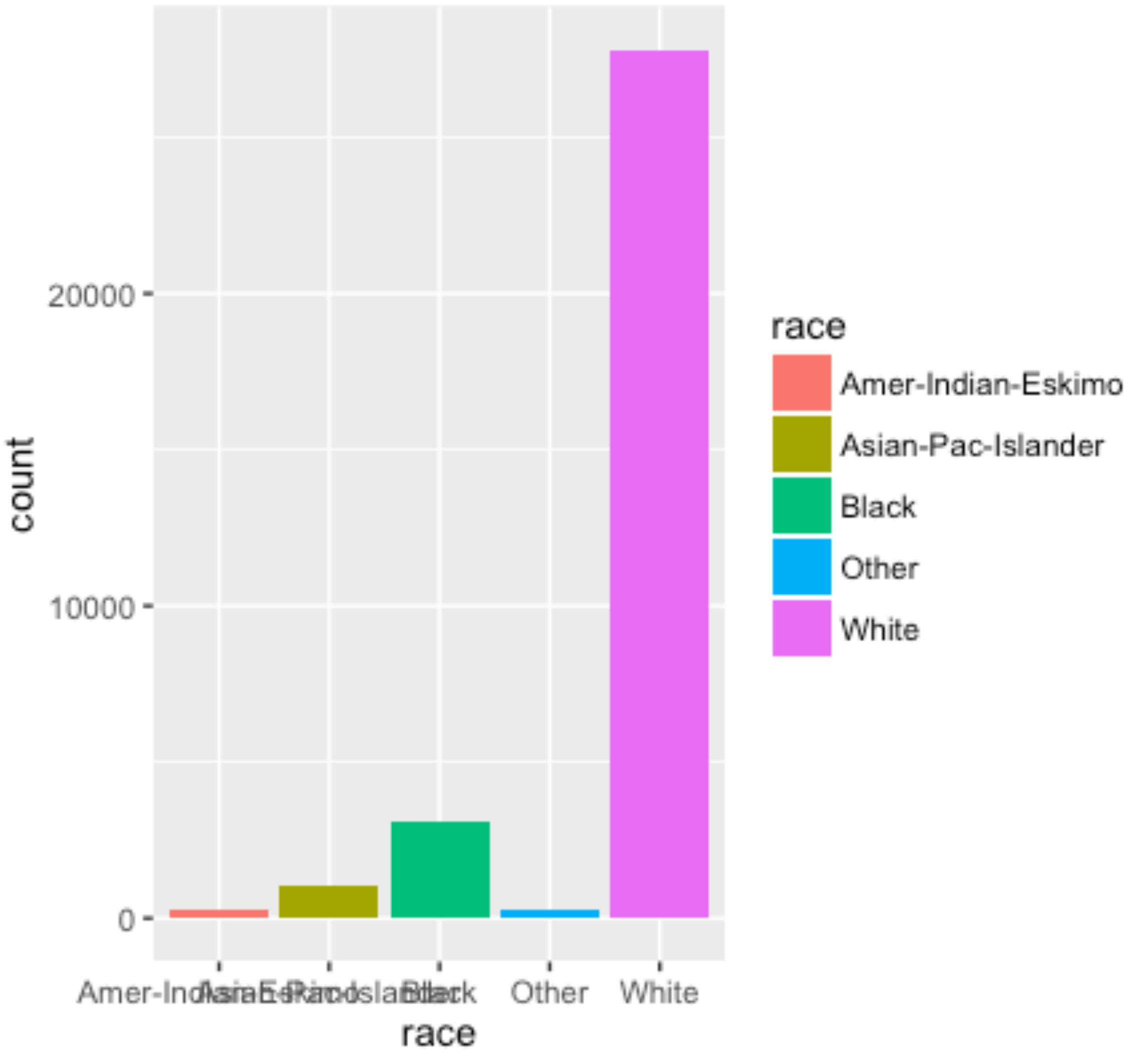
(c) For each numeric variable, plot two boxplots side-by-side in the same plot for the two values of income.


<=50K
>50K
age
75
50
25
x

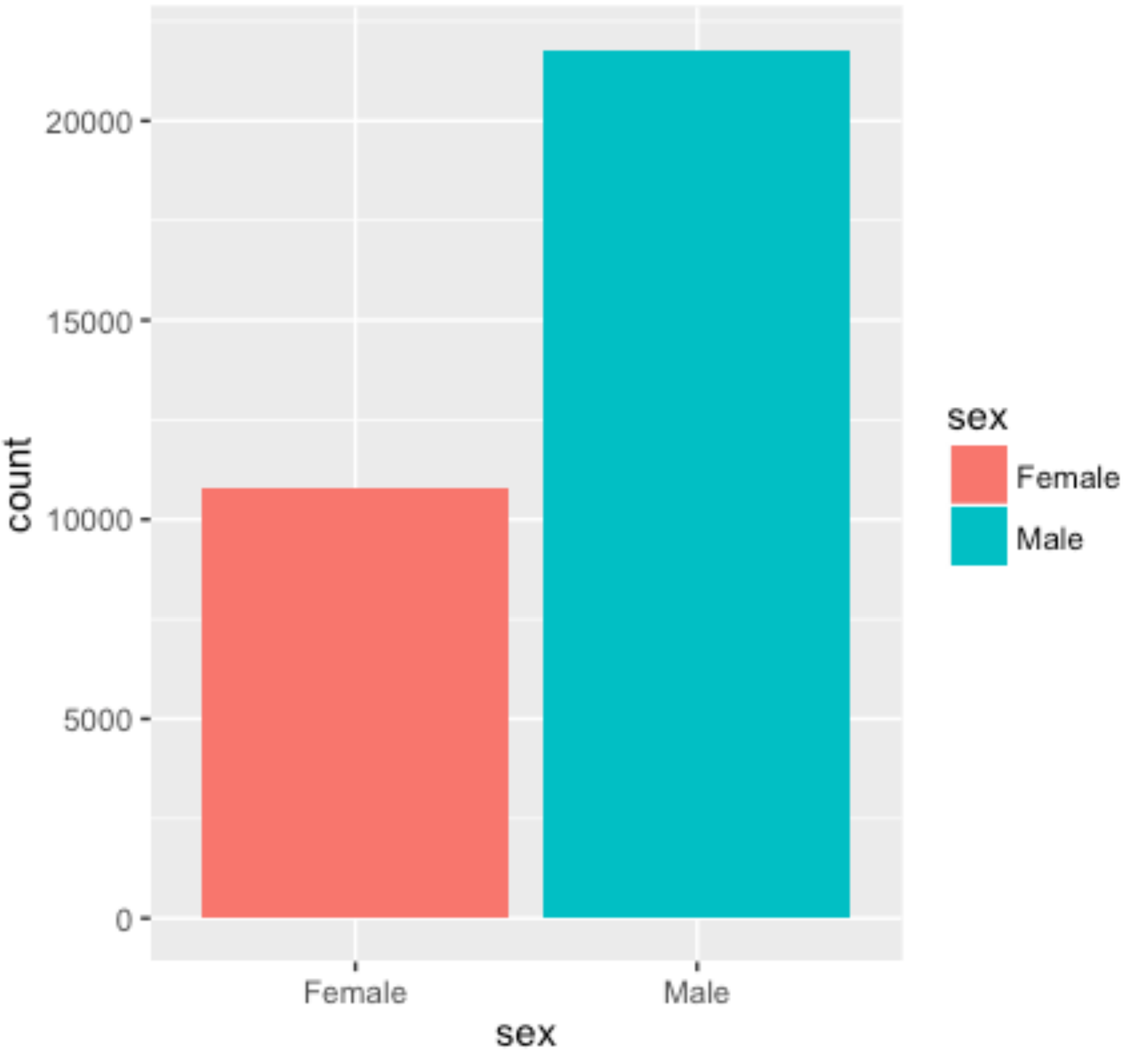


Q4 a. (a) For each categorical variable, plot a bar plot with frequencies. How many unique values does each variable have?

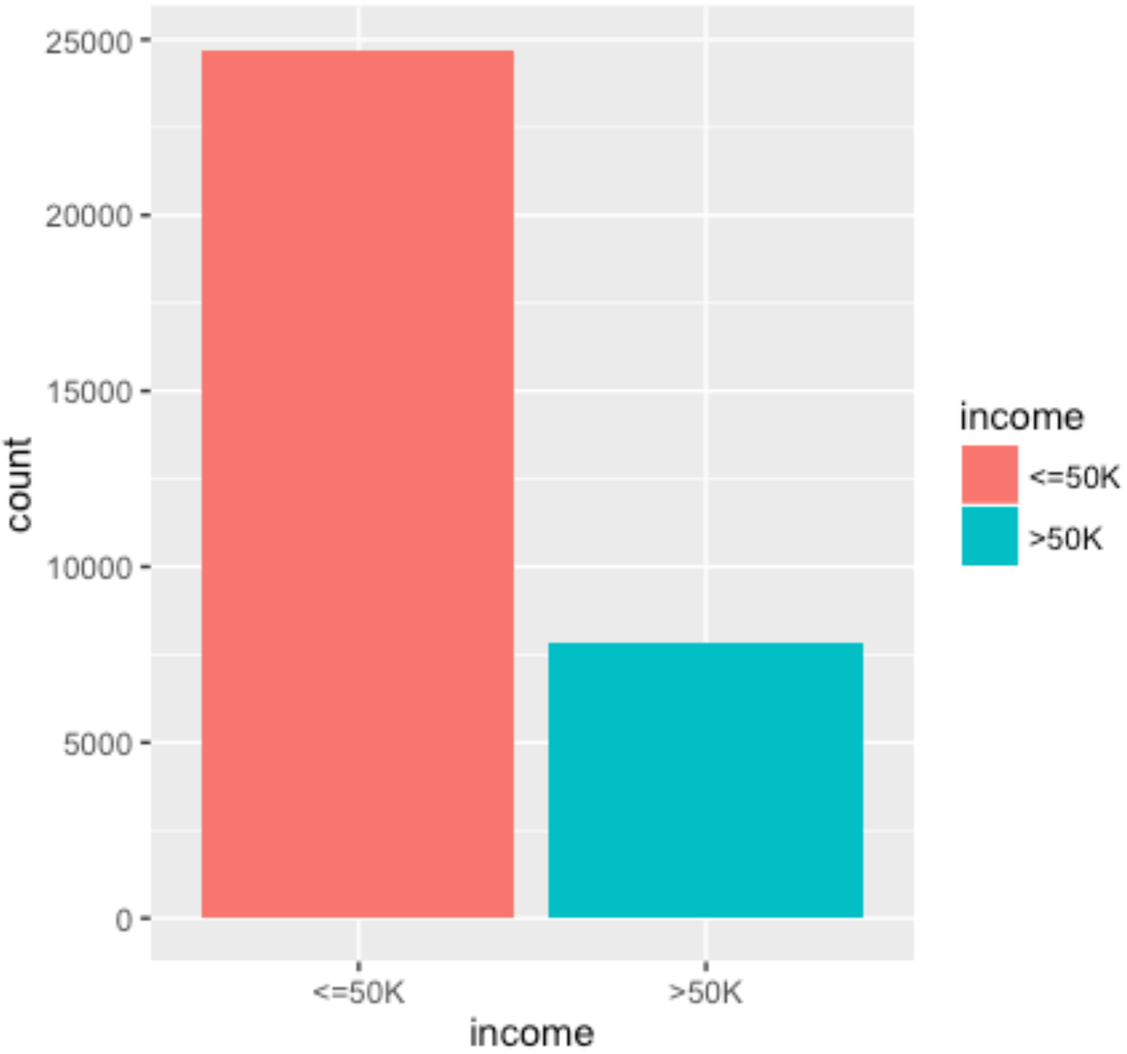
```
cd + geom_bar(aes(x=race,fill=race),size =0.5)
```



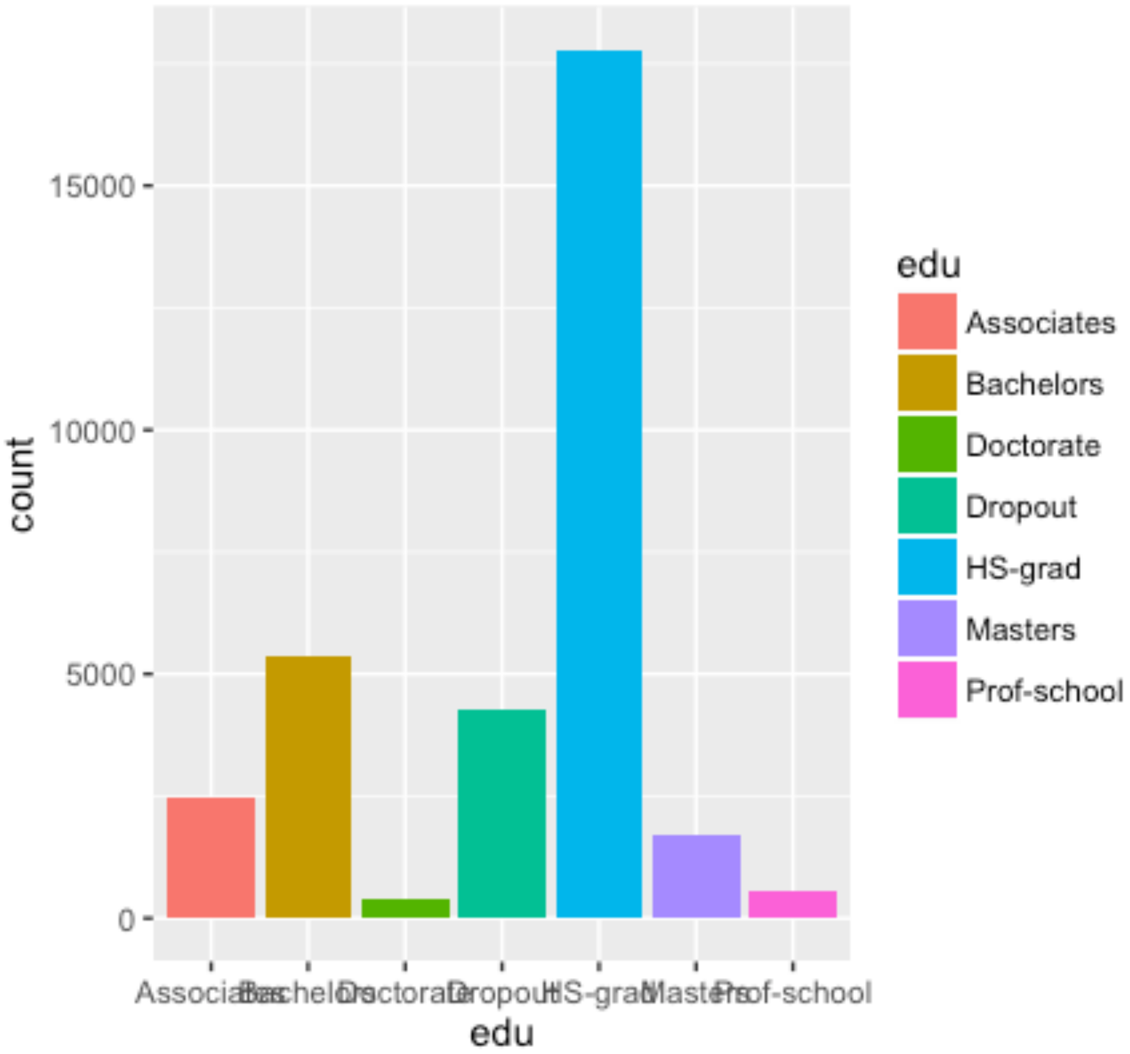
Race vs count plot shows that the number of white people are maximum.



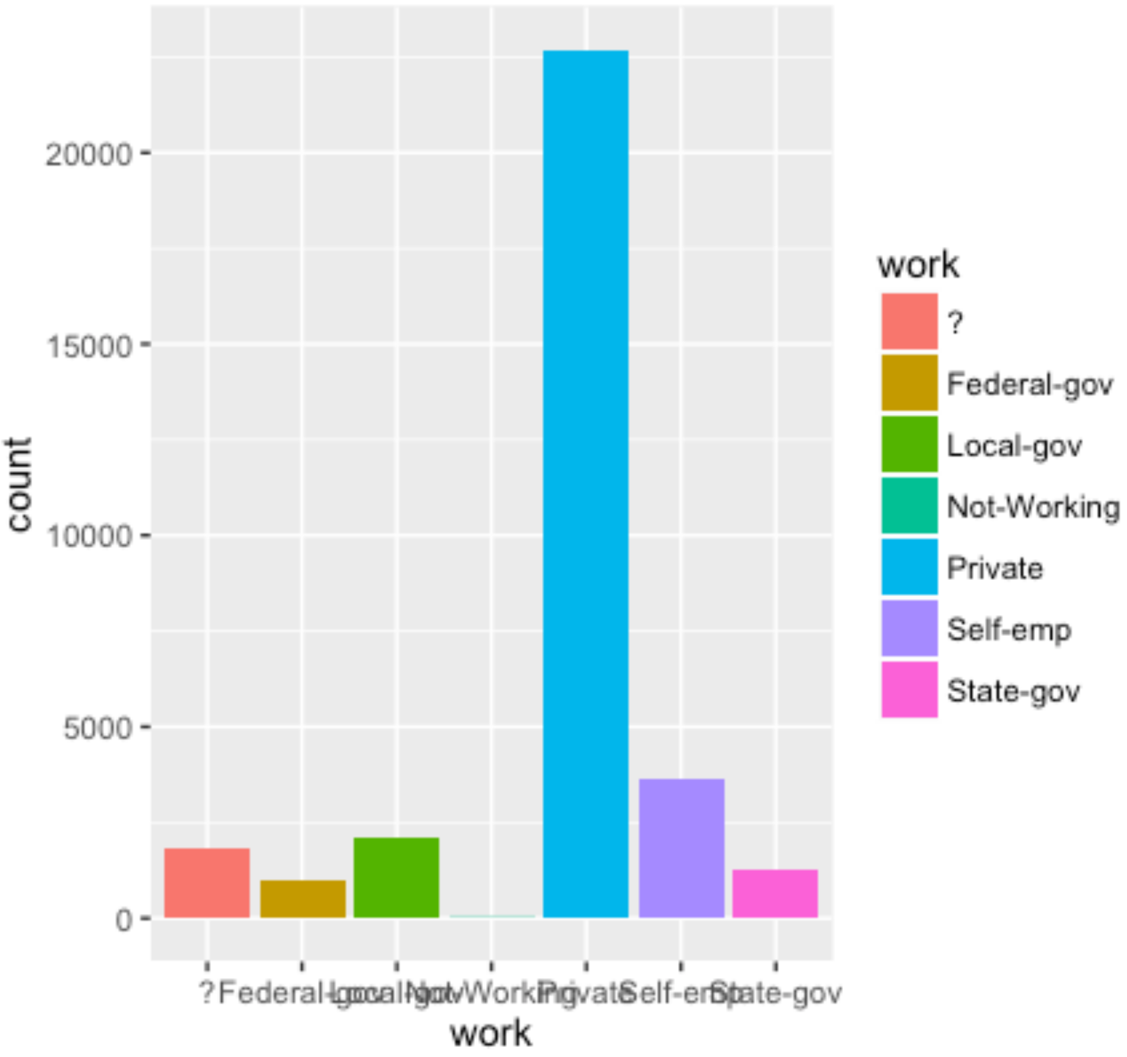
Sex vs Count shows that Male population is more than female population.



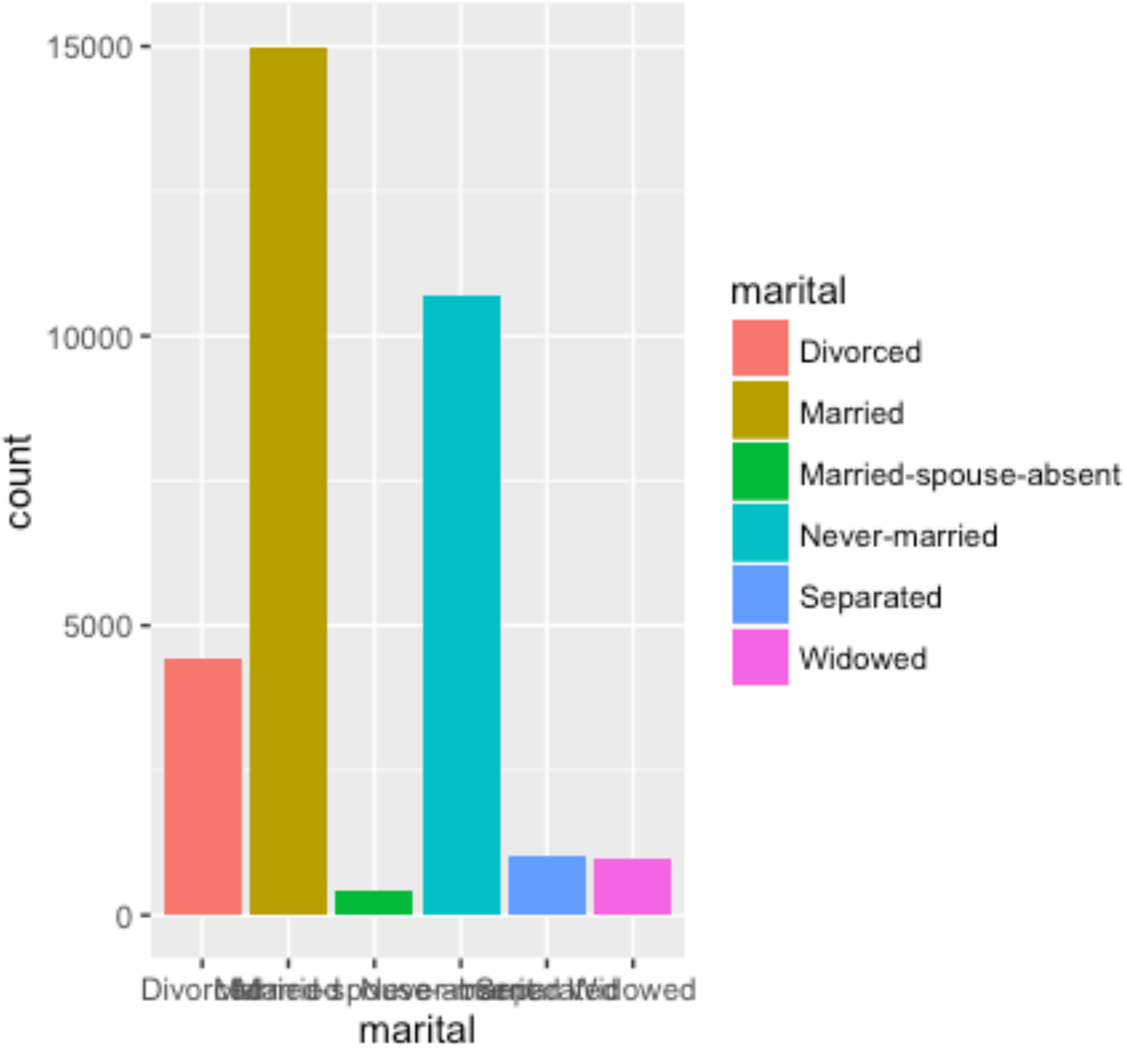
Income vs count shows that number of people with income of <=50k is more than number of people with income >50k



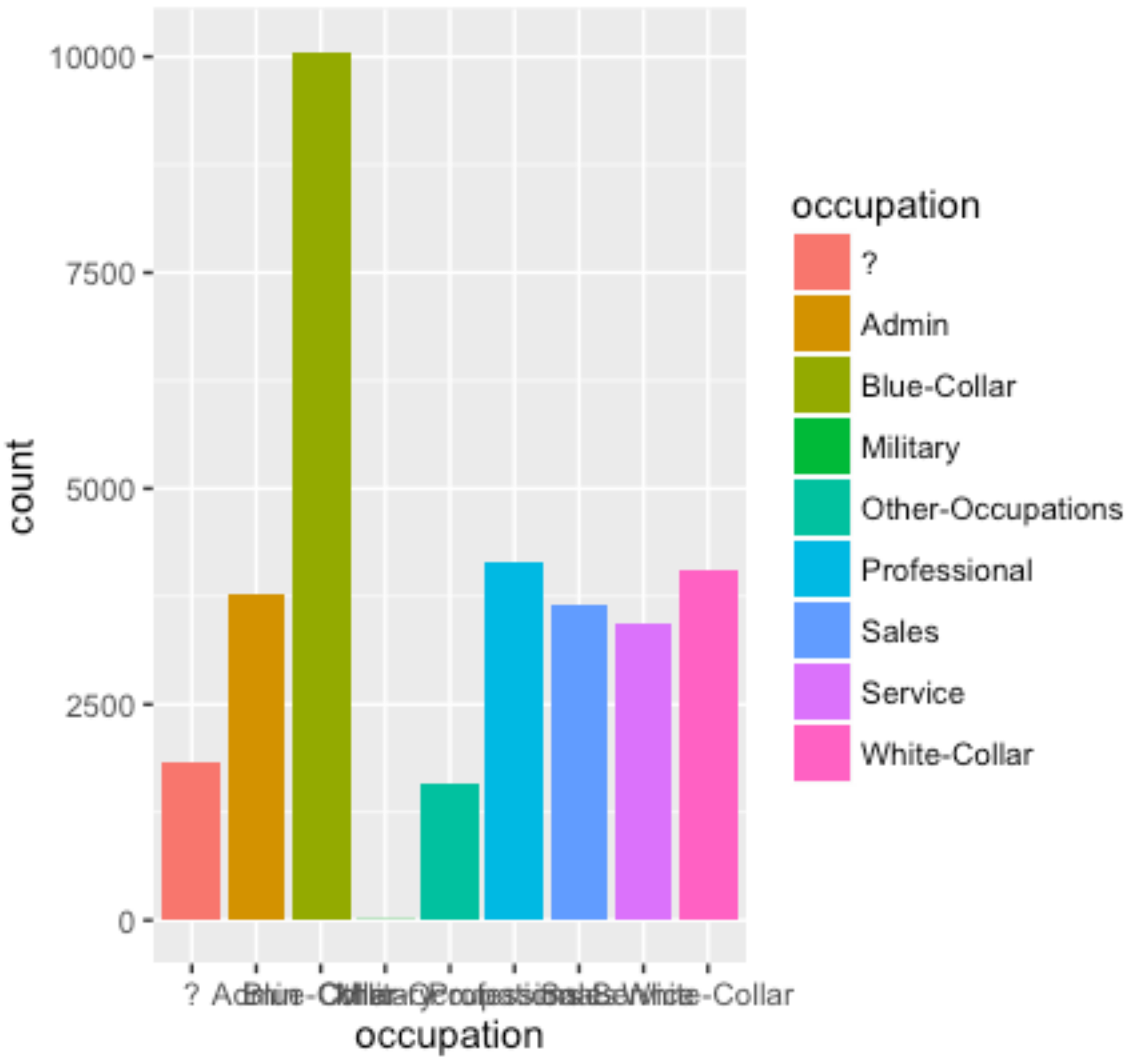
Edu vs count shows that max number of people are High school grads



Work vs count shows that maximum people work in a private firm.



Marital vs count shows that maximum people are Married with married spouse absent being the second largest.



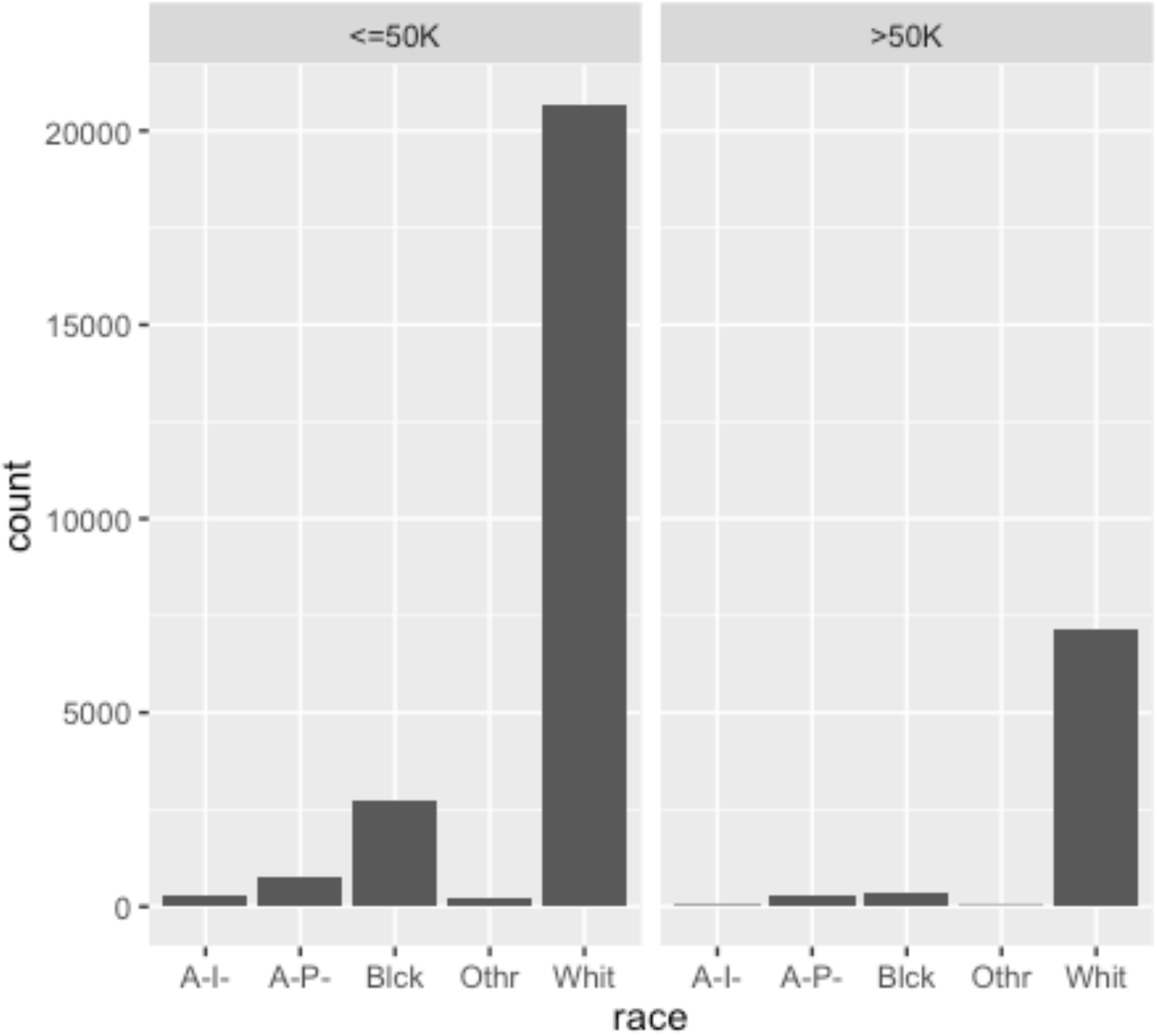
occupation vs count shows that maximum number of people work in blue collar industry .

```
$age
[1] 73

$work
[1] 7

$edu
[1] 7

$marital
[1] 6
```

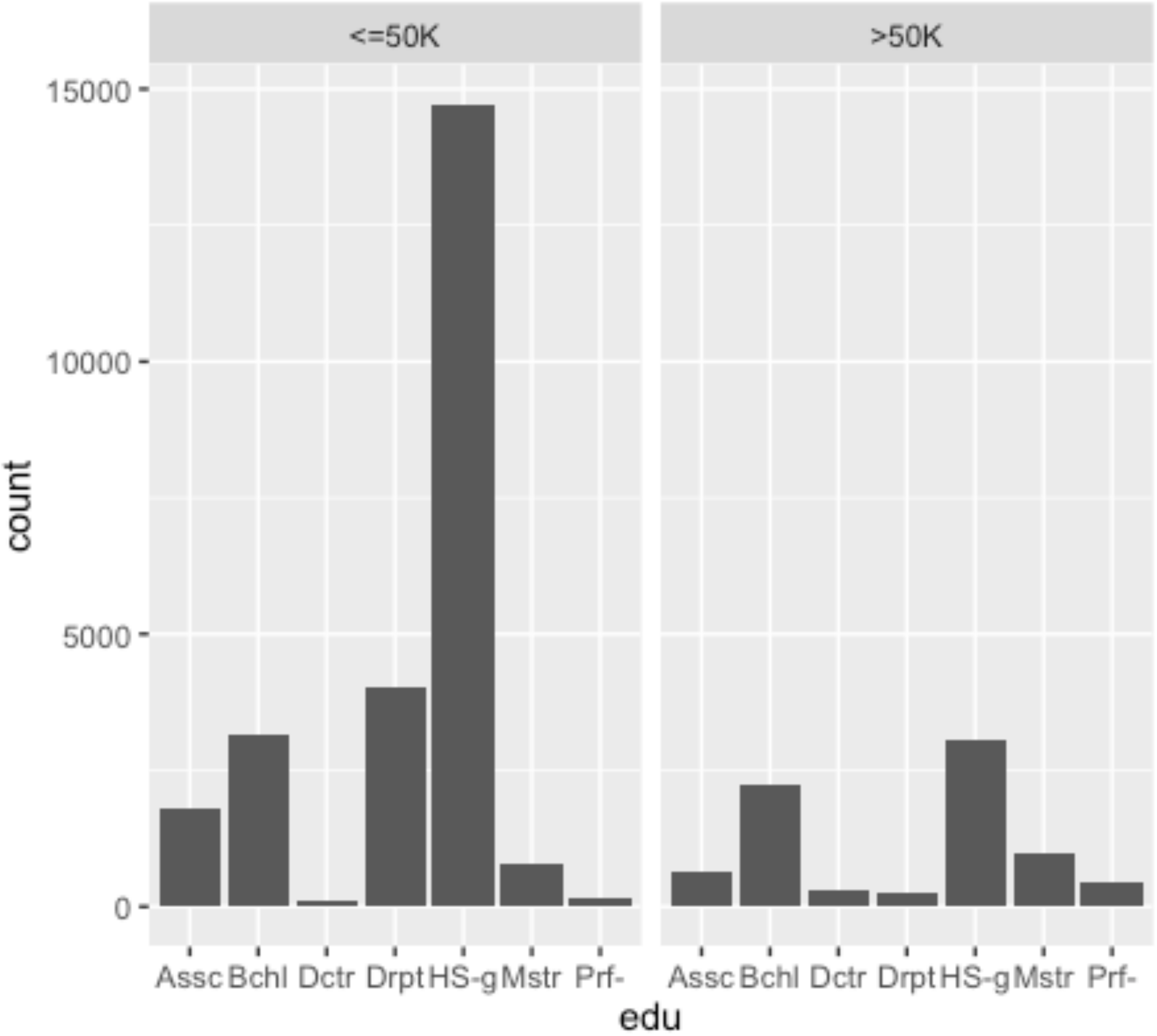
```
$occupation
[1] 9

$race
[1] 5

$sex
[1] 2

$hrs_per_week
[1] 94

$income
[1] 2
```

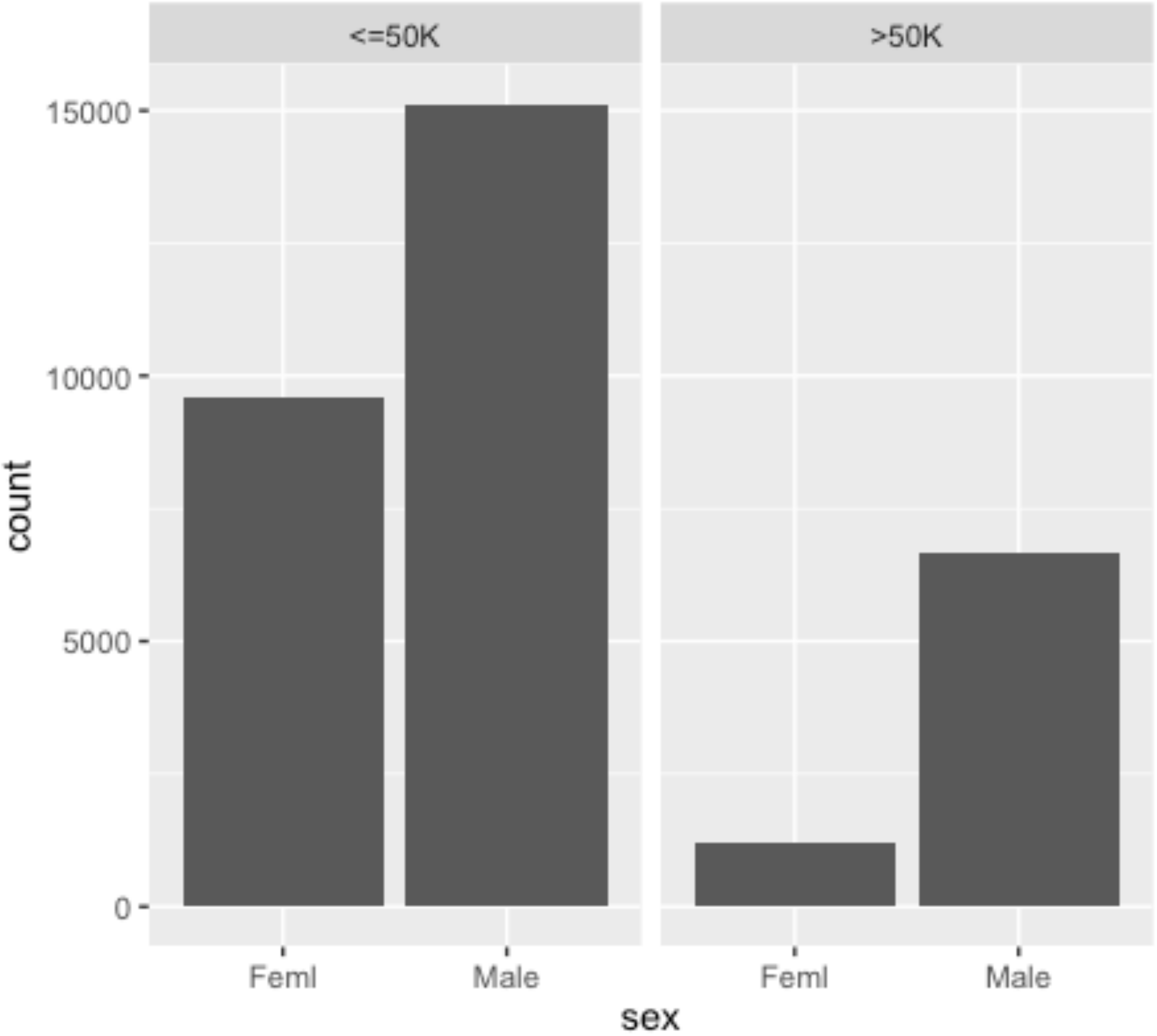
Q4 b . (b) For each categorical variable, plot two bar plots (with the same scales) for the two values of income.

(c) For (a) and (b) above, briefly describe what you learned from the plots.

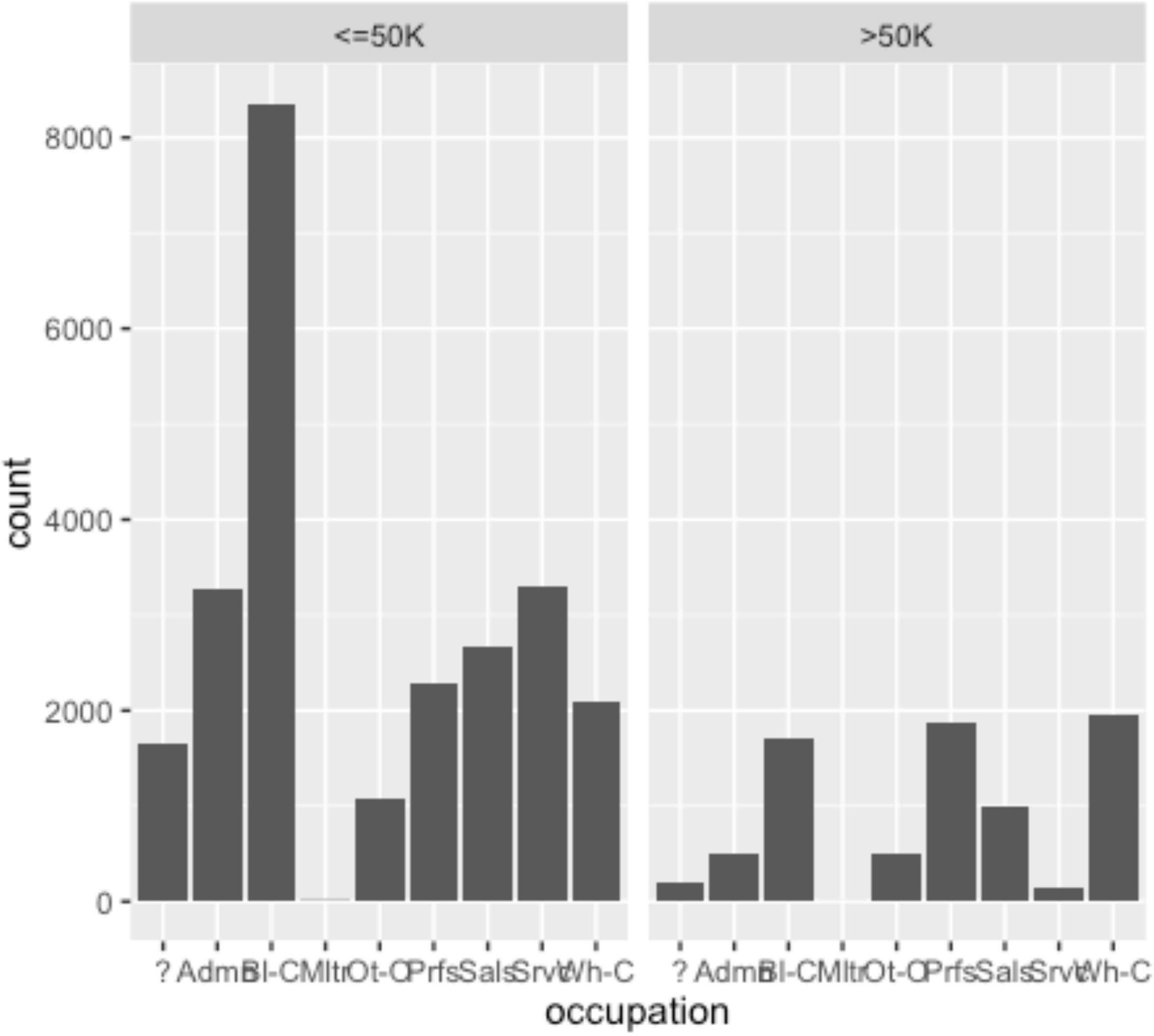


As we can see white people earn the maximum in both income groups.

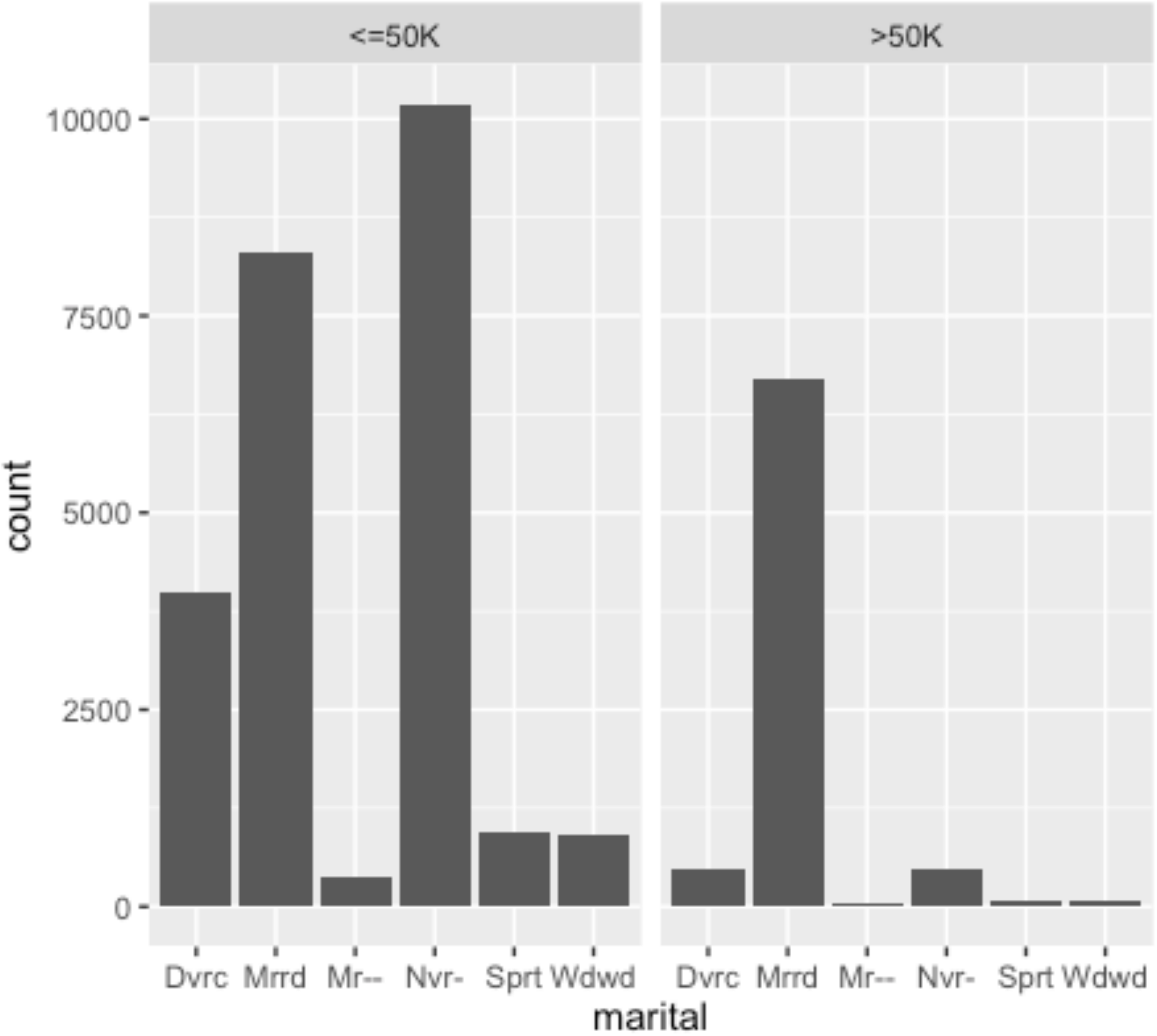

<=50K
>50K
15000
10000
5000
0
count
Assc
Bchl
Dctr
Drpt
HS-g
Mstr
Prf-
Assc
Bchl
Dctr
Drpt
HS-g
Mstr
Prf-
edu



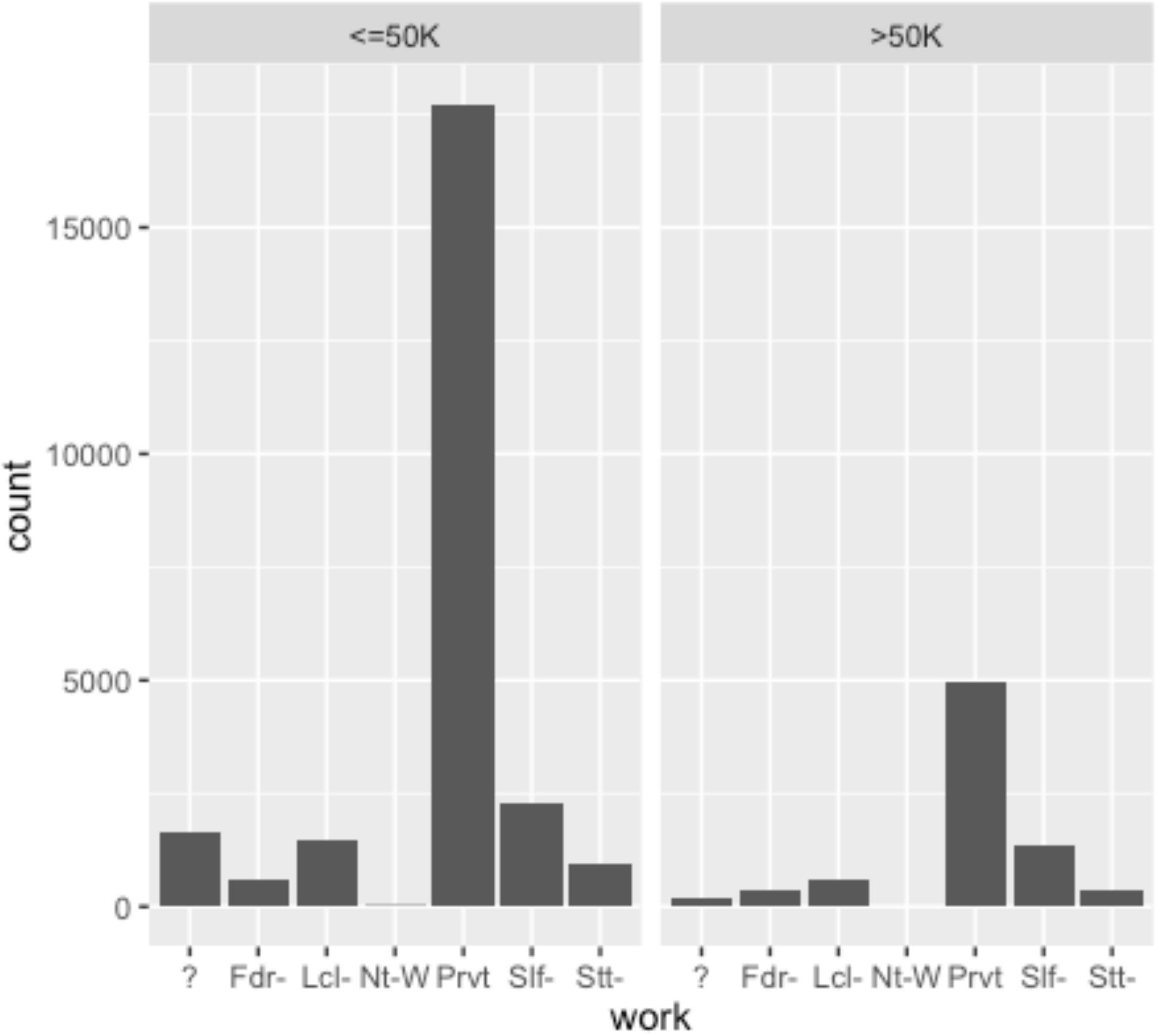
As we can the number of males are more than females in both income groups.



There are more people in blue collar industry with income of <=50k , however in income groups >50k Blue collar , White collar and professional earn near about same.

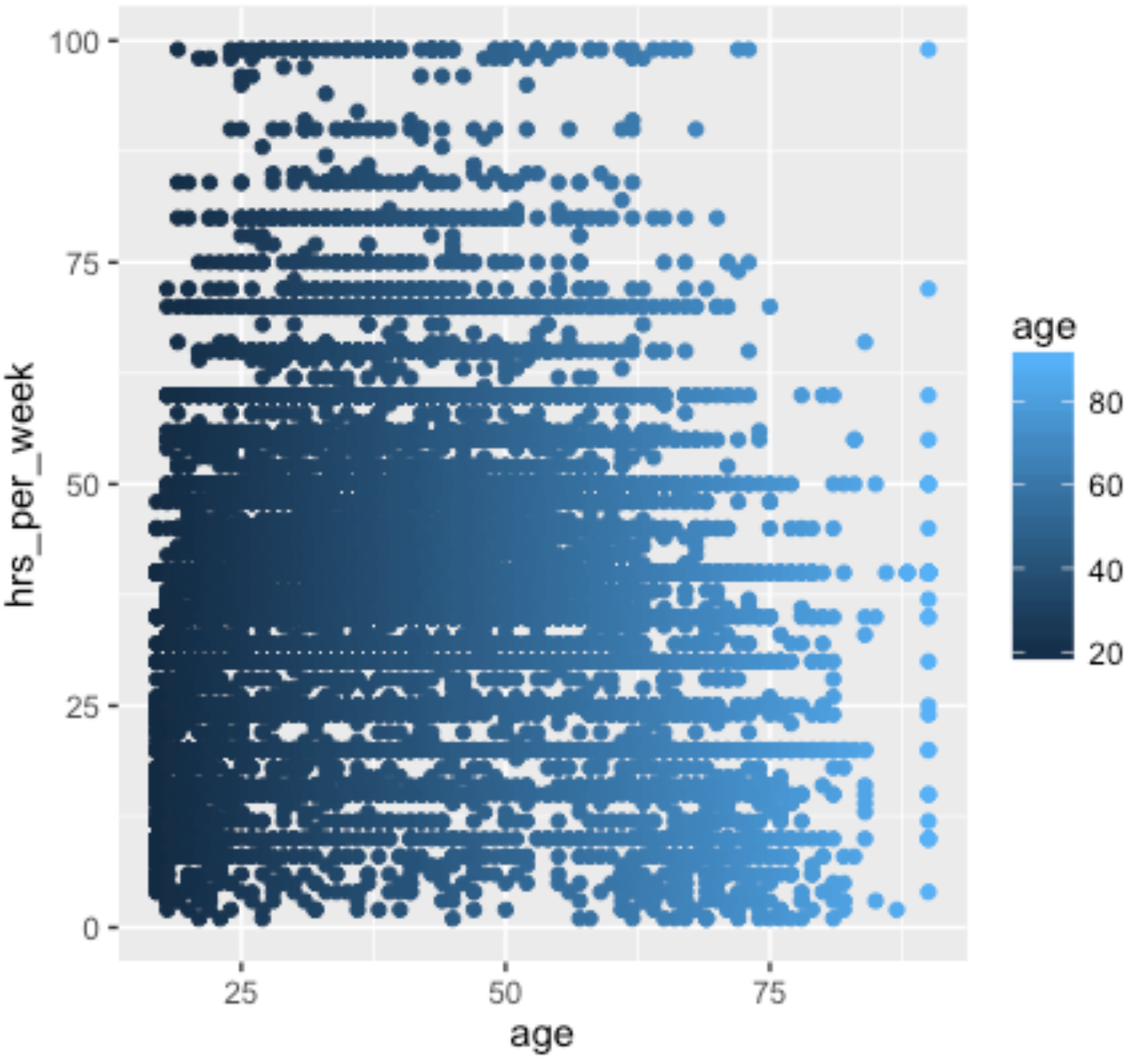


As we can see the number of never-married in income group of <=50k are maximum whereas in income group >50k married people are more.

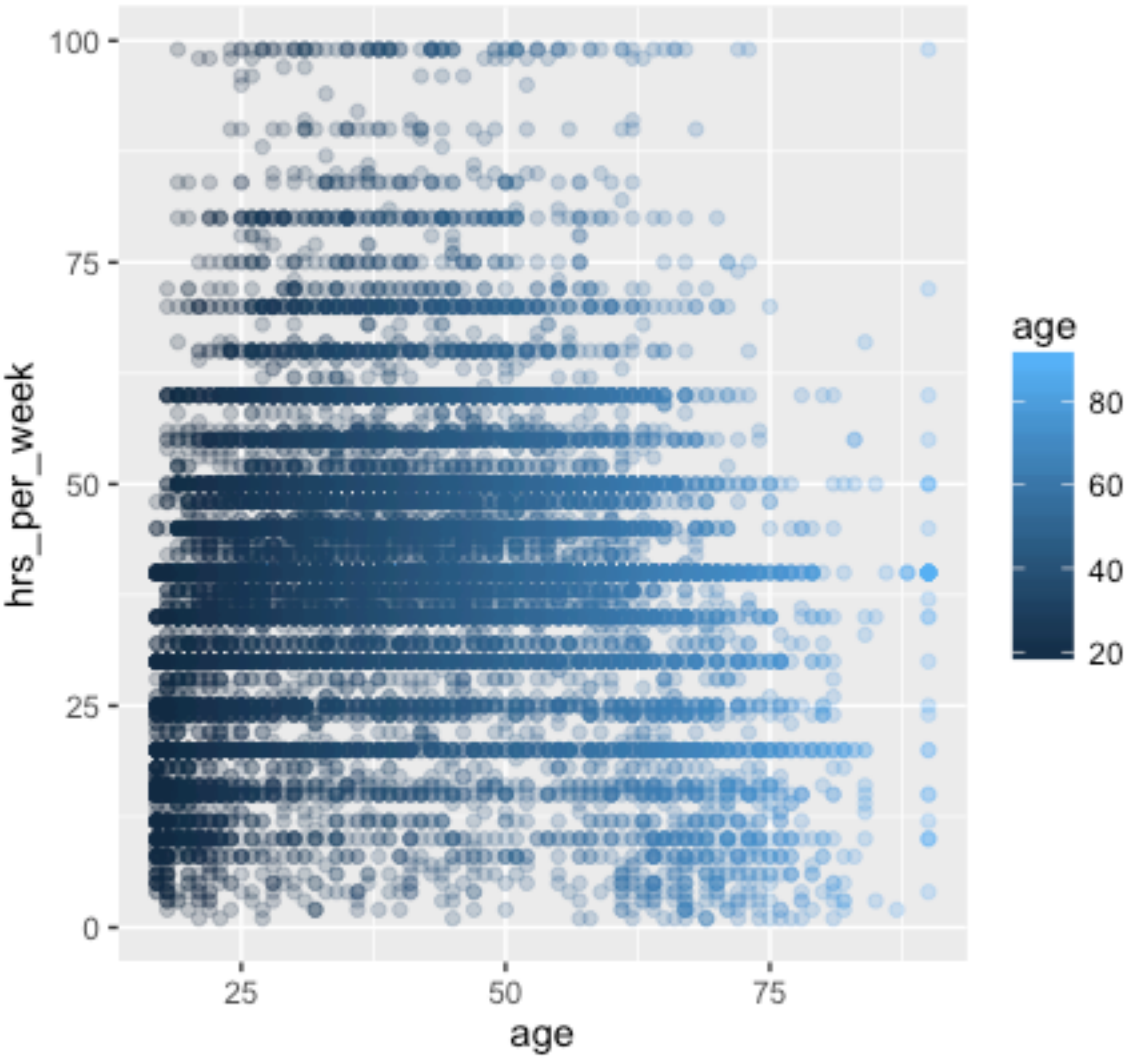


People working in private firms are more in both the income groups.

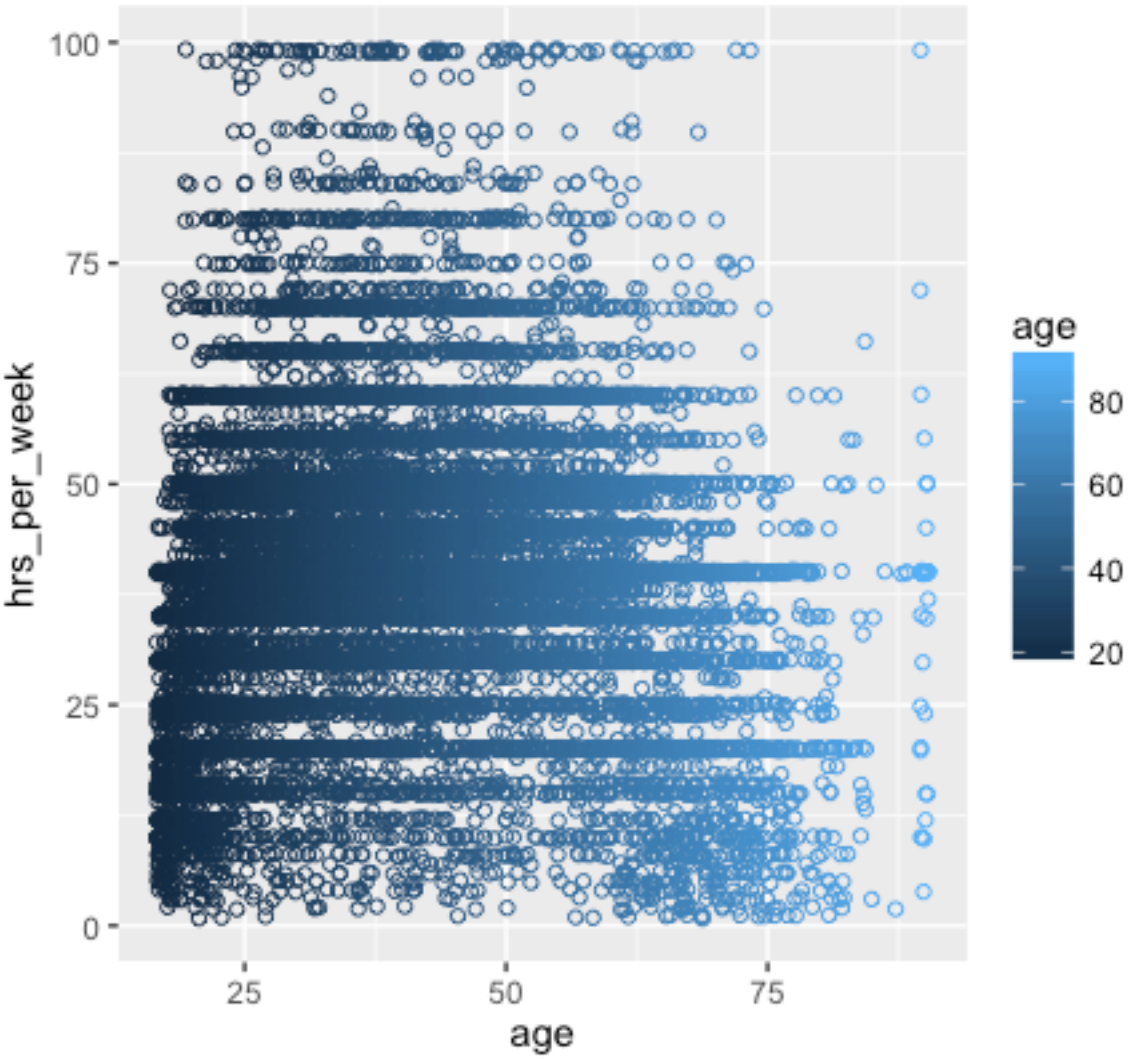
Q5) Plot a scatter plot between "age" and "hrs_per_week". Is there any overplotting? If yes, try to fix it. Briefly Describe what you learned from these plots. Compute the correlation coefficient between these two variables.



For overplotting : transparency



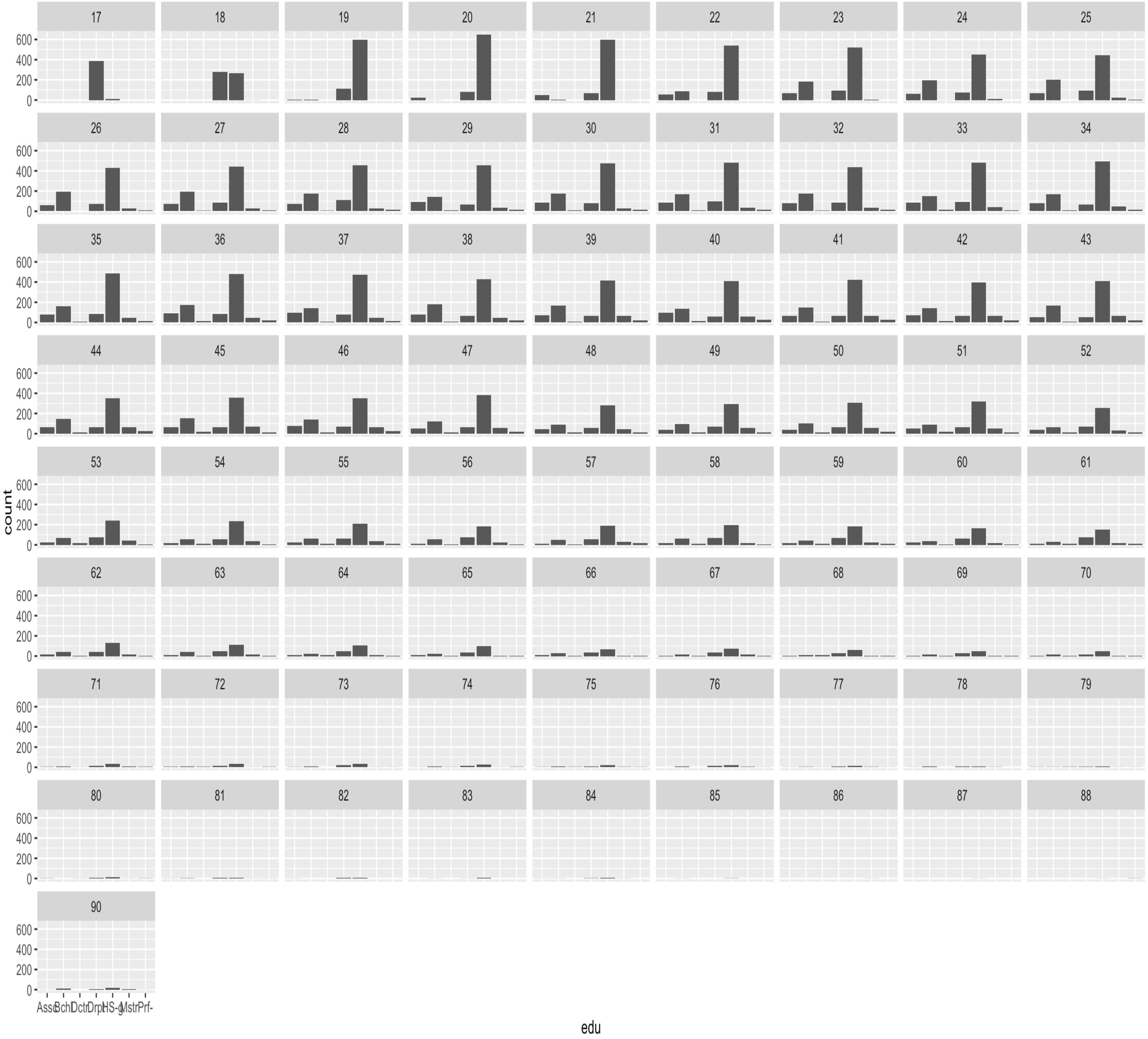
We can also jitter the points



.
Correlation coefficient :

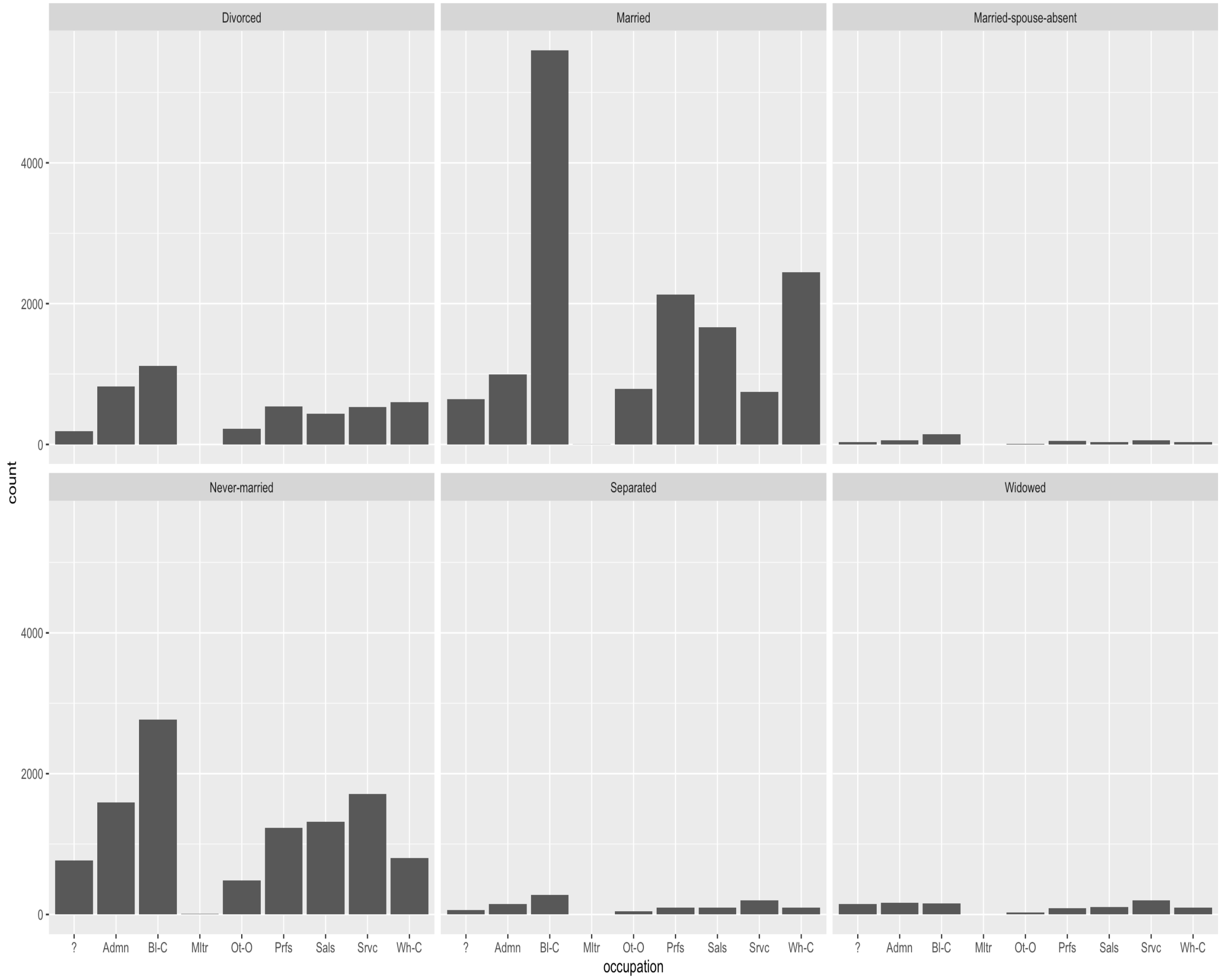
```
 [1] 0.06875571
```

6. Find any additional insightful relationship / plot in this data set.



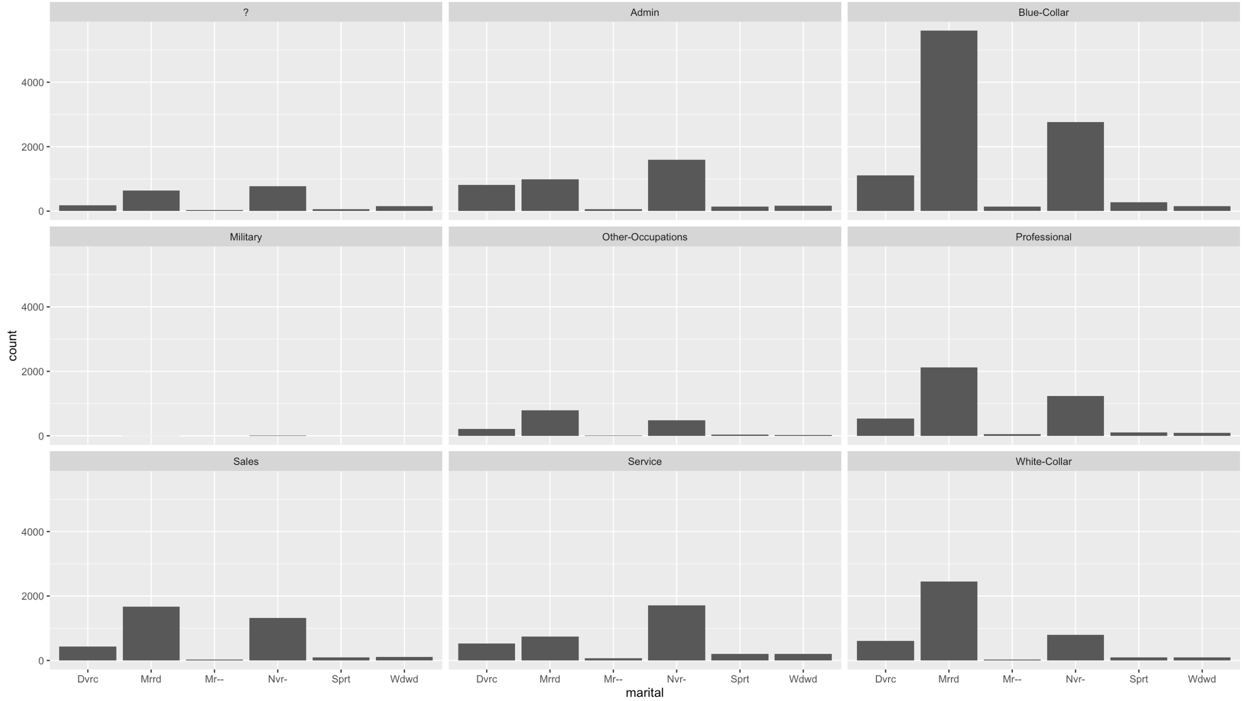
From the plot above we can see that maximum drop outs are of the age 17.

2.



Married people are maximum in Blue collar industry.

3.



Very few people are involved in sales industry.