Home Work 1

Write-up for the plots generated from censusData.csv

Balaji Narayanaswami (W1151287)

10.12.2015

COEN 281 Pattern Recognition & Data Mining

Question 1:

The attributes and the corresponding data types are:

age : Ratio

Justification: Age has an inherent zero point (birth day). Also we can say that a 40 year old person is twice as old as 20 year old guy.

work : Nominal

Justification: work has values that describe the type of employer that each person holds. Since it has no order amongst it but it denotes classes, it is clearly Nominal.

edu: Ordinal

Justification: Education status follows an order(rank) that is Bachelors is ranked below Masters. Hence edu is Ordinal.

marital: Nominal

Justification: Since Marital denotes the relationship status (states) of a person i.e. Married, Widowed etc., it is considered Nominal.

occupation: Nominal

Justification: Since occupation denotes the type of work (Blue-collar, Sales etc.,) a person does and there is no order among the values, it is Nominal.

race: Nominal

Justification: Since a person can be a member or not a member of a particular race, this is classified as Nominal.

sex: Binary

Justification: Since sex is a nominal attribute with two distinct values(Male and Female), it is a binary attribute.

hrs_per_week: Ratio

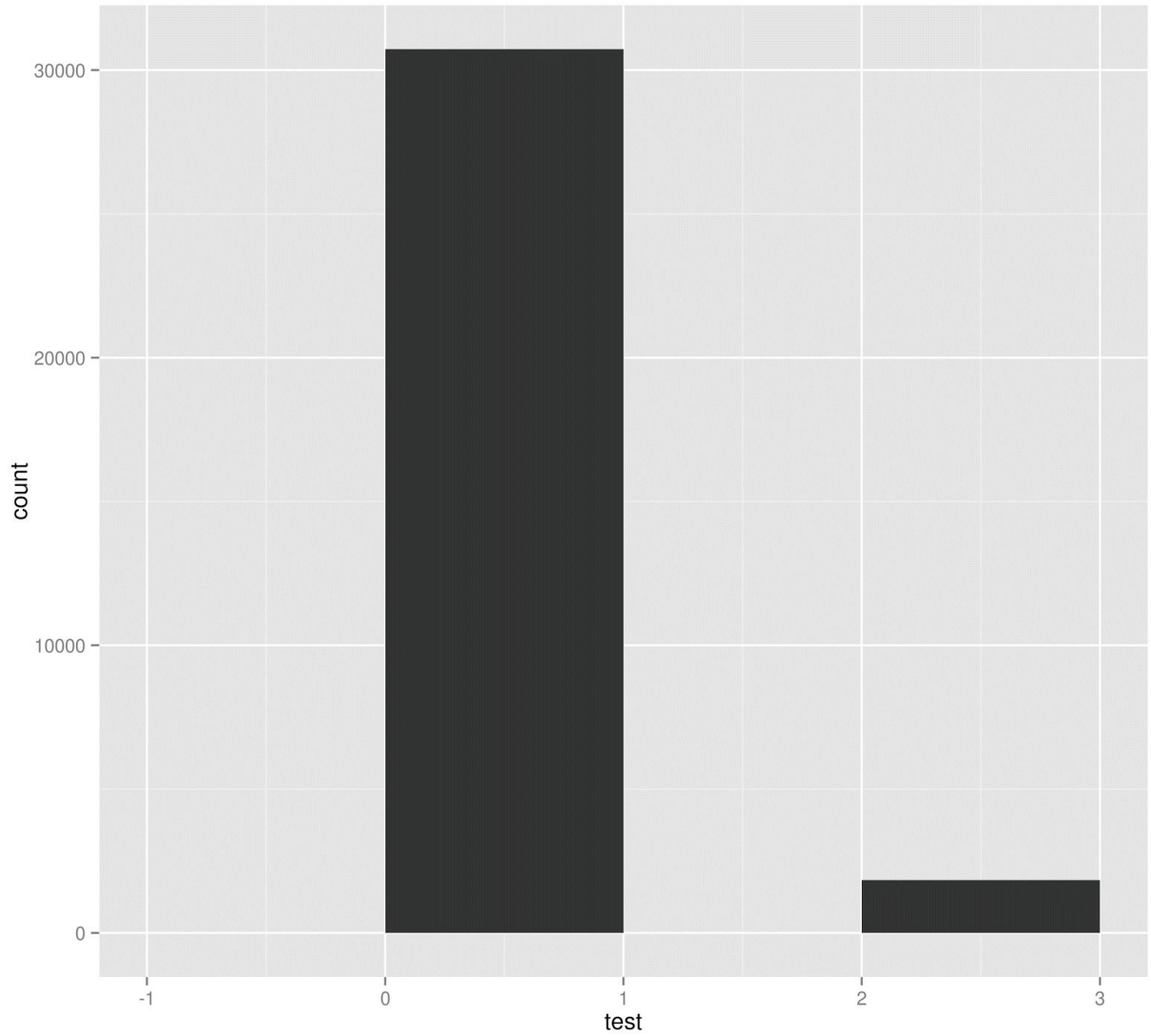
Justification: Since hours per week a person works is strictly >=0 and hence it can be compared with another person's work hours, it is Ratio attribute.

income: Binary

Justification: Since income has two possible outcome >50k and <=50k in the census data, this is a binary attribute.

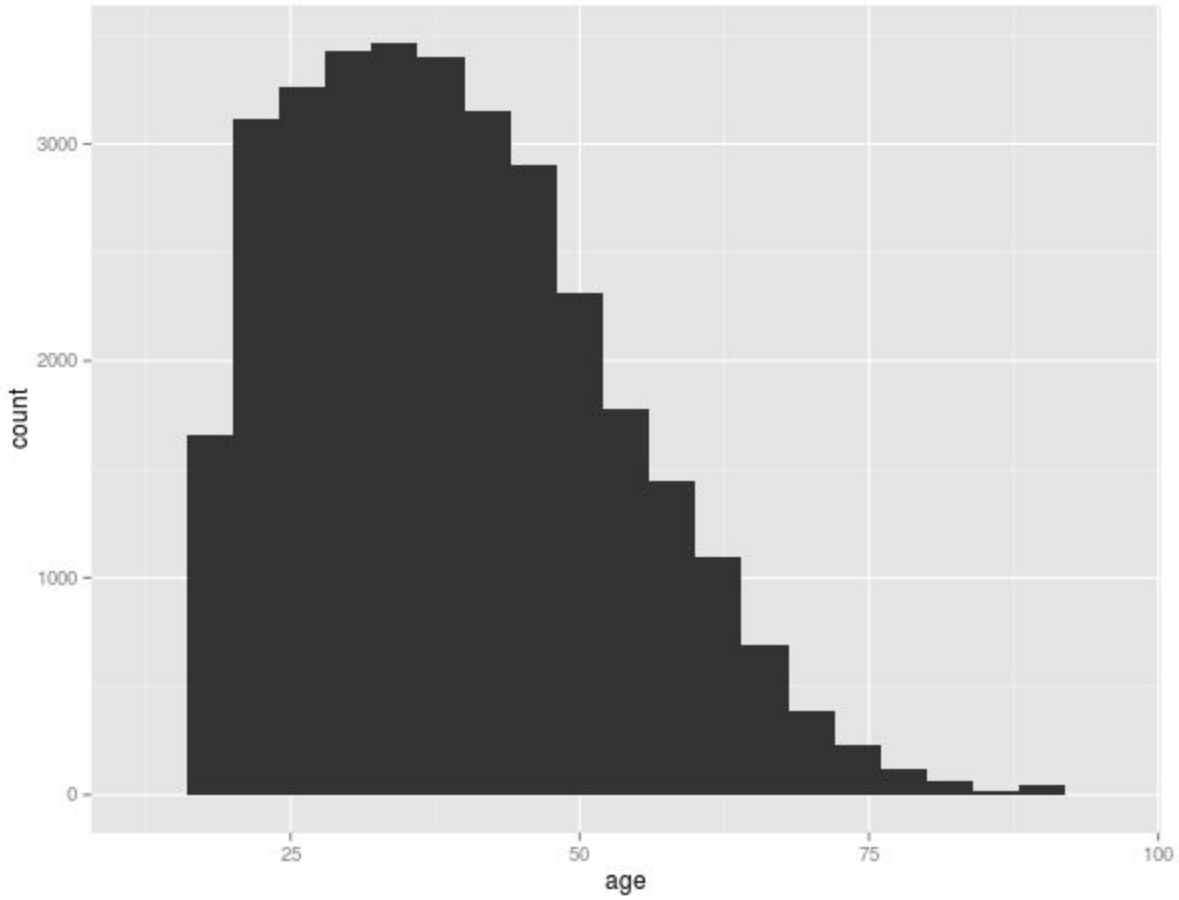
Question 2:

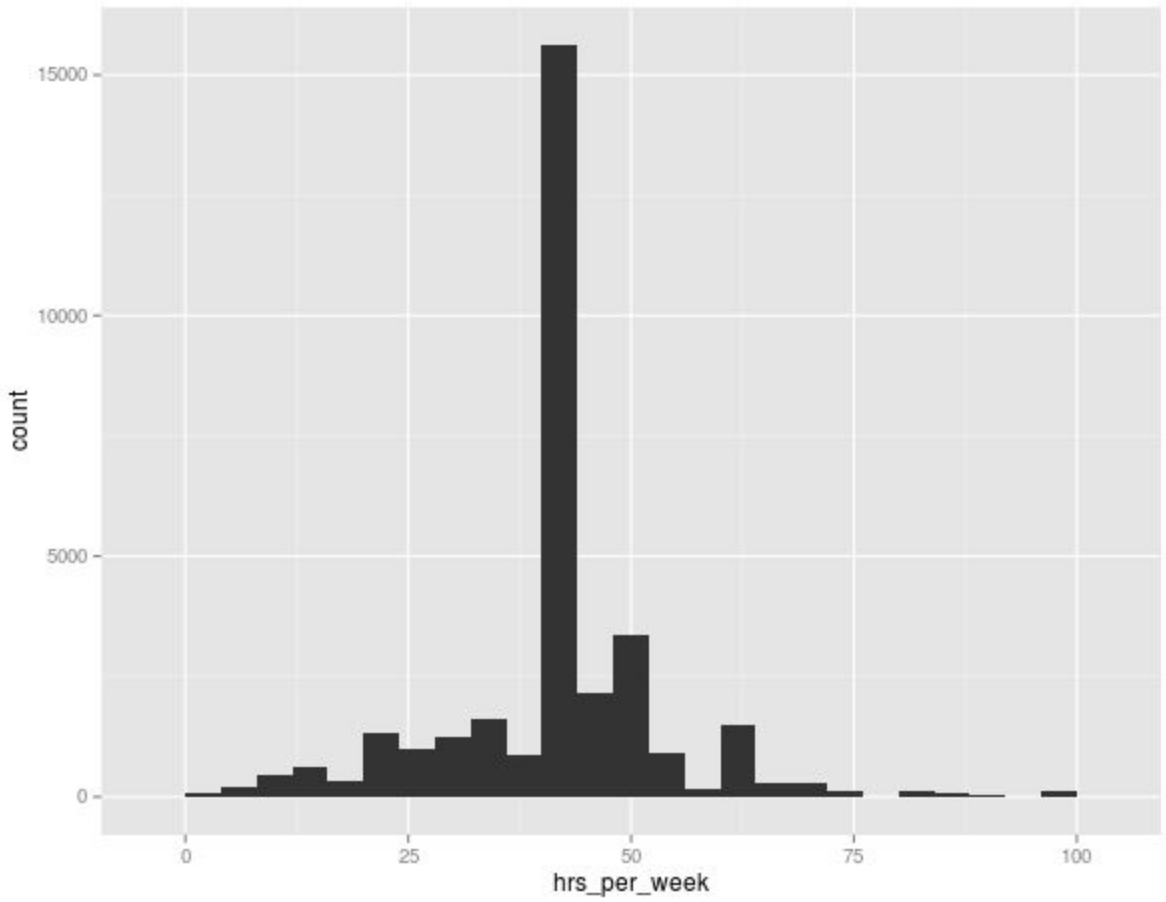
(B):



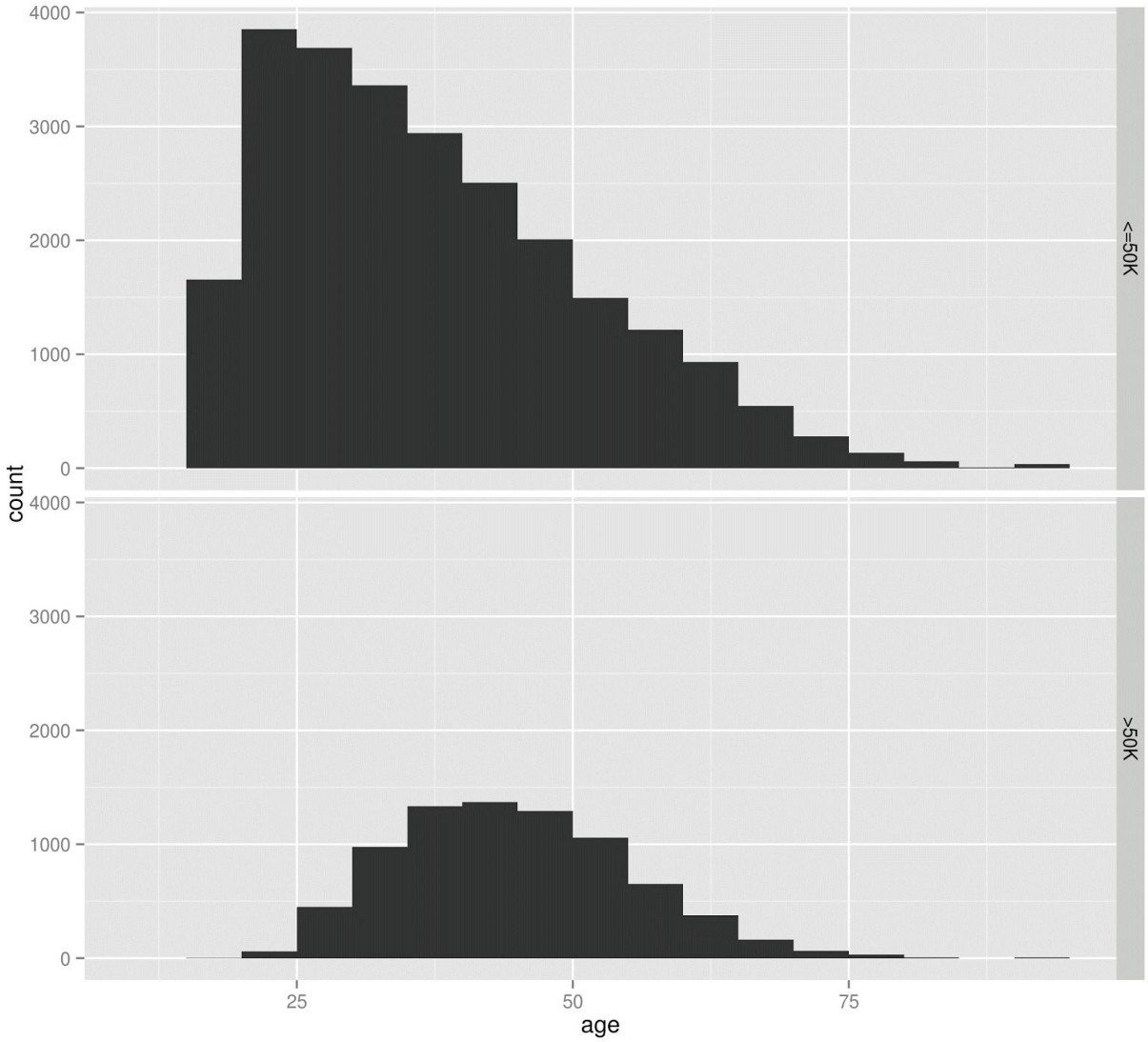
Question 3

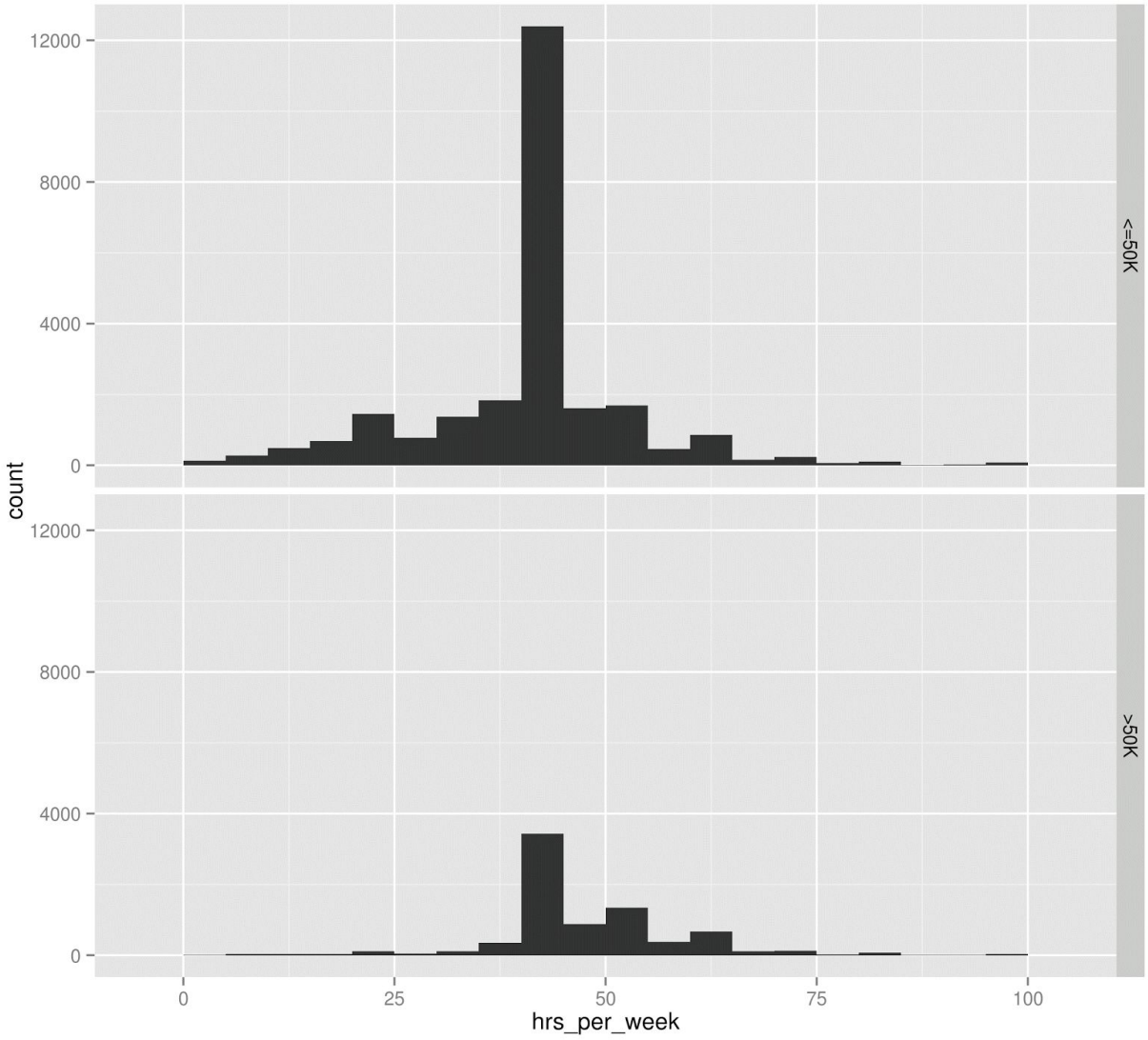
(A):



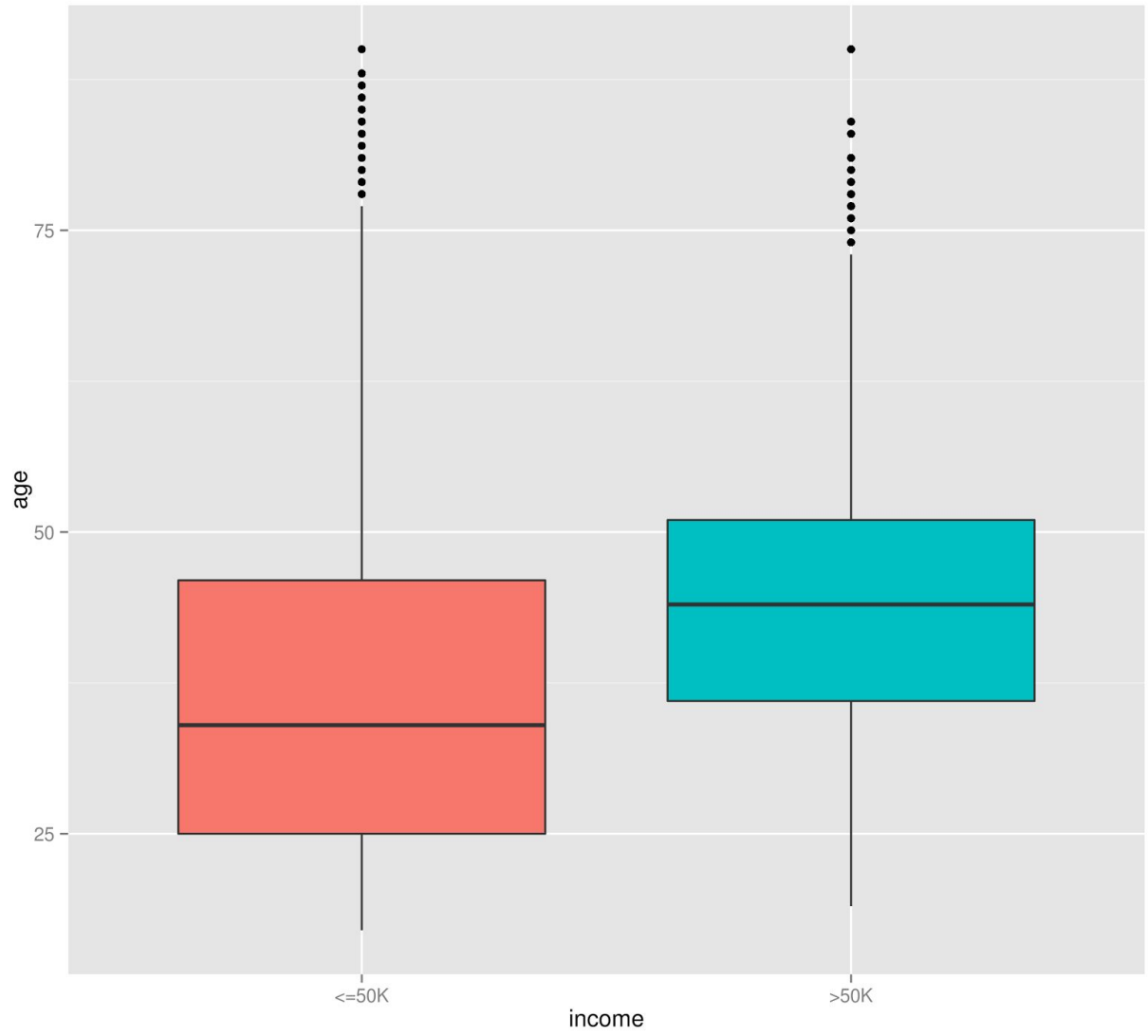

15000
10000
5000
0
count
0
25
50
75
100
hrs_per_week

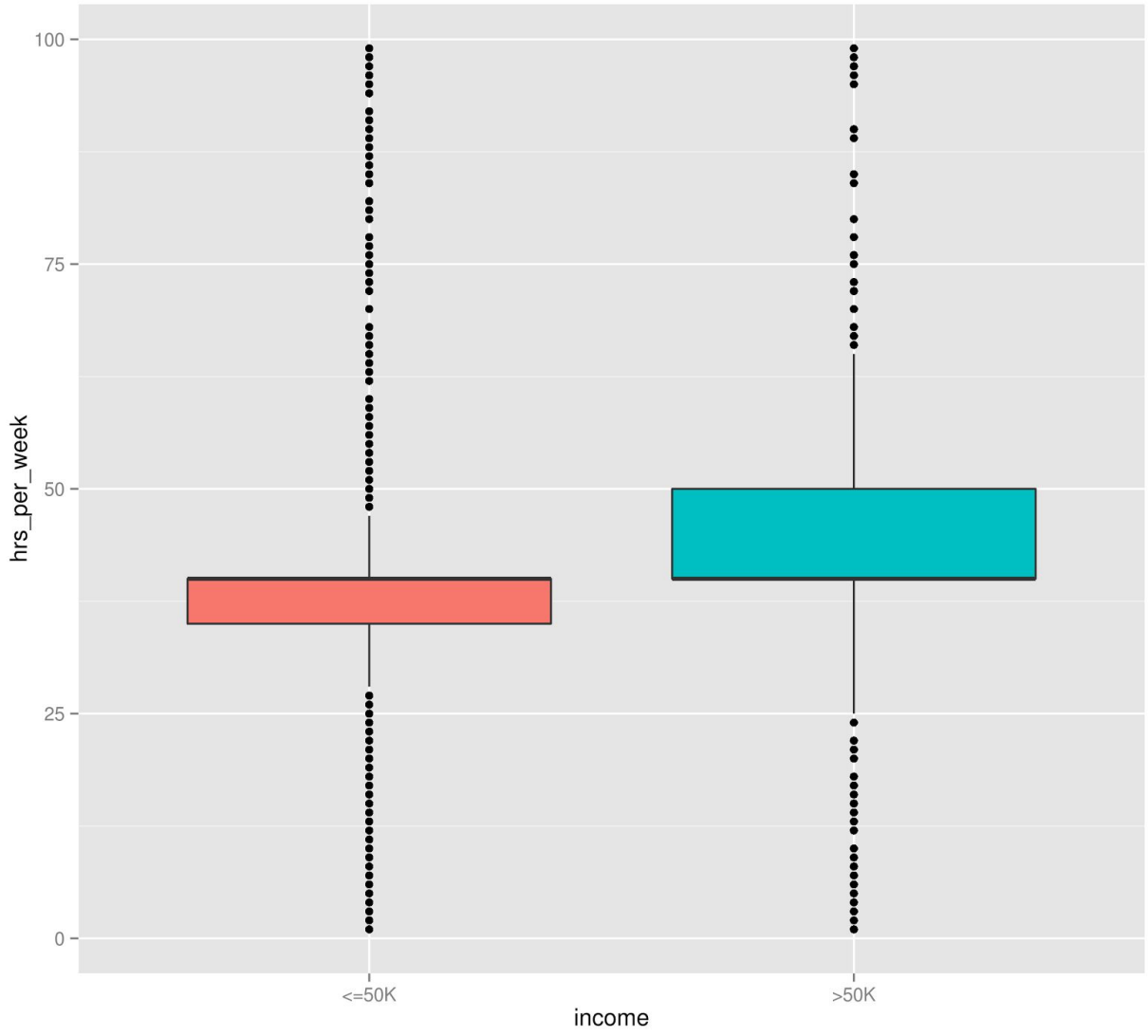
(B):




12000
8000
4000
0
count
<=50K
>50K
0
25
50
75
100
hrs_per_week

(C):



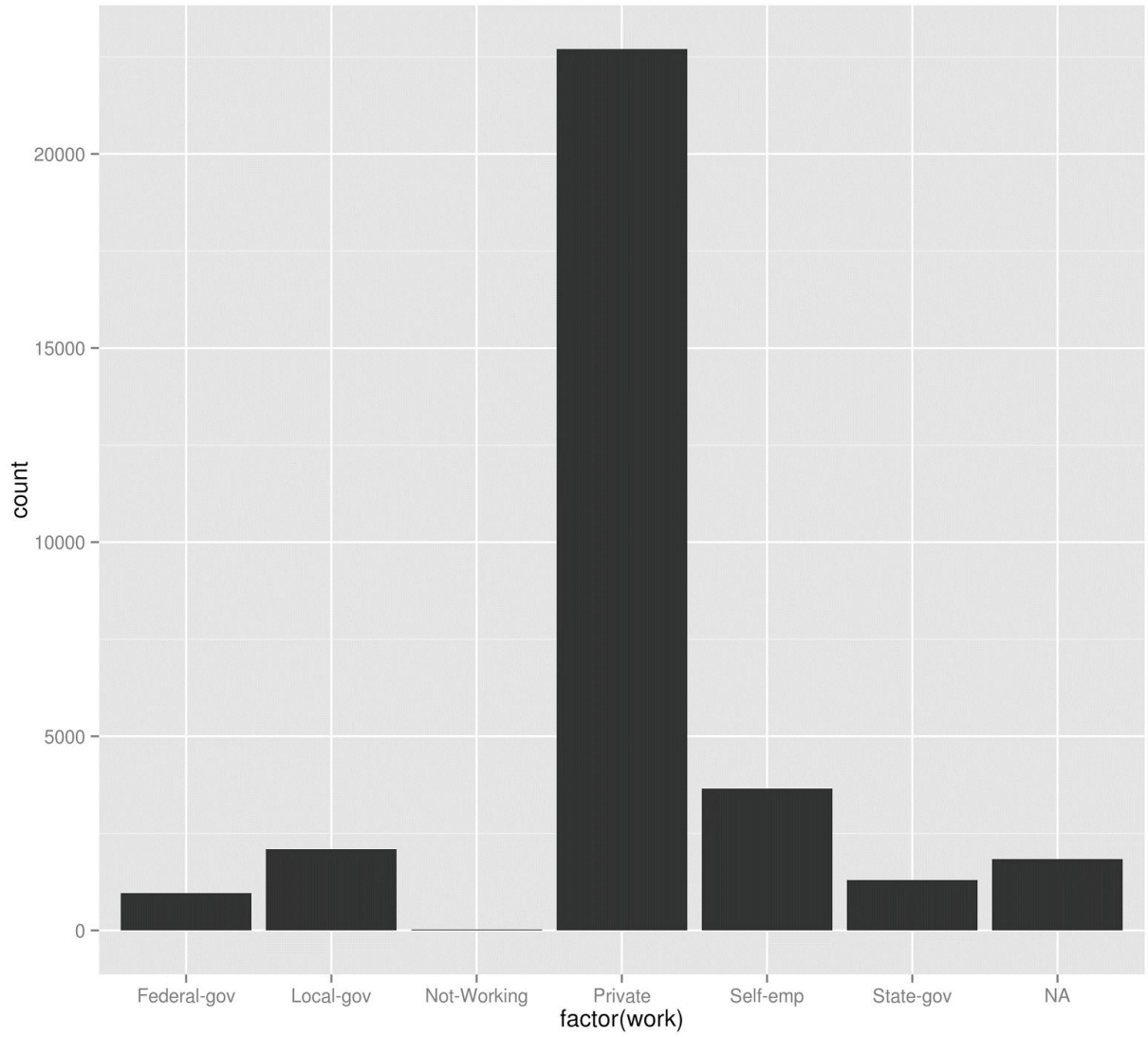


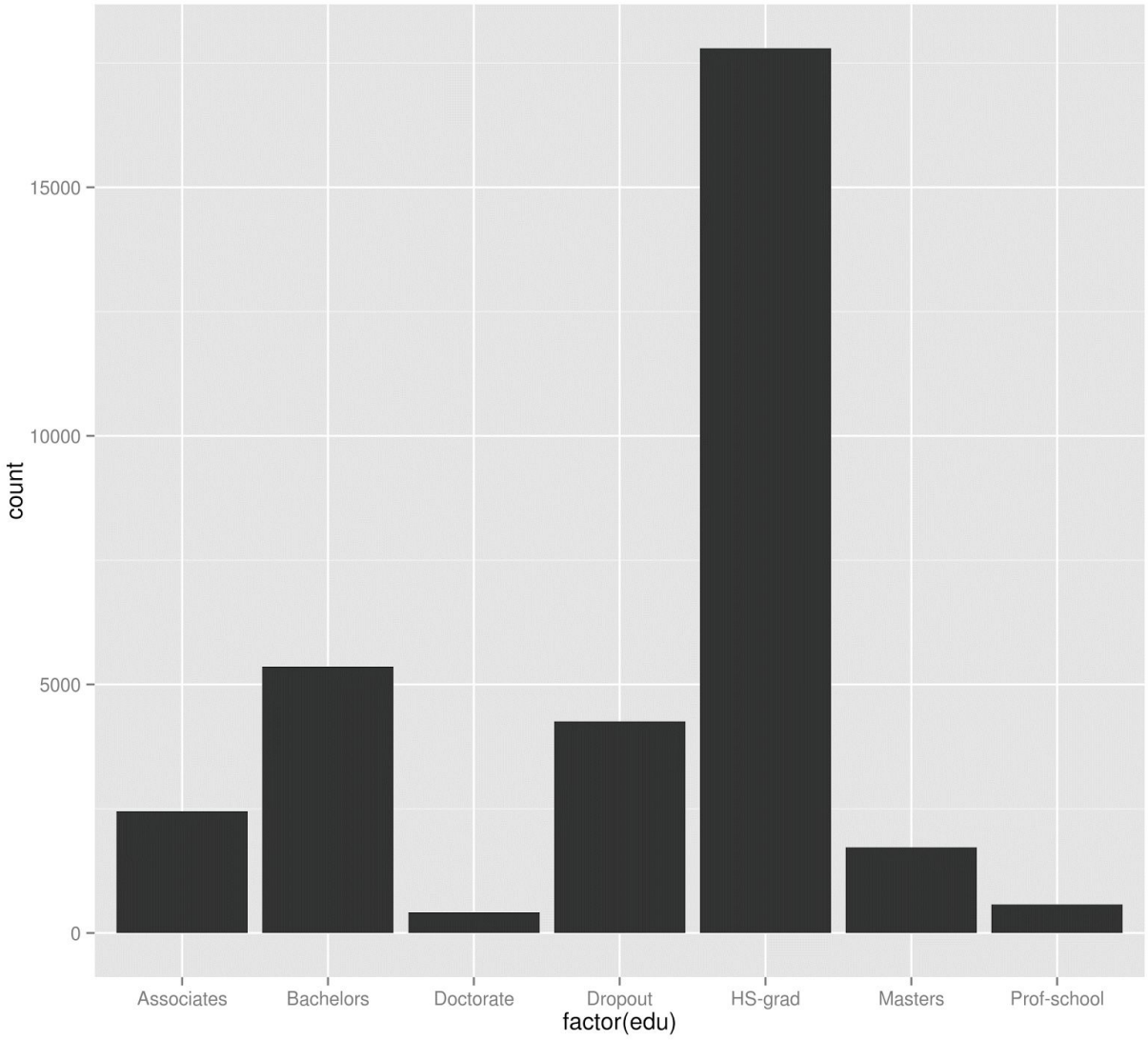
(D):

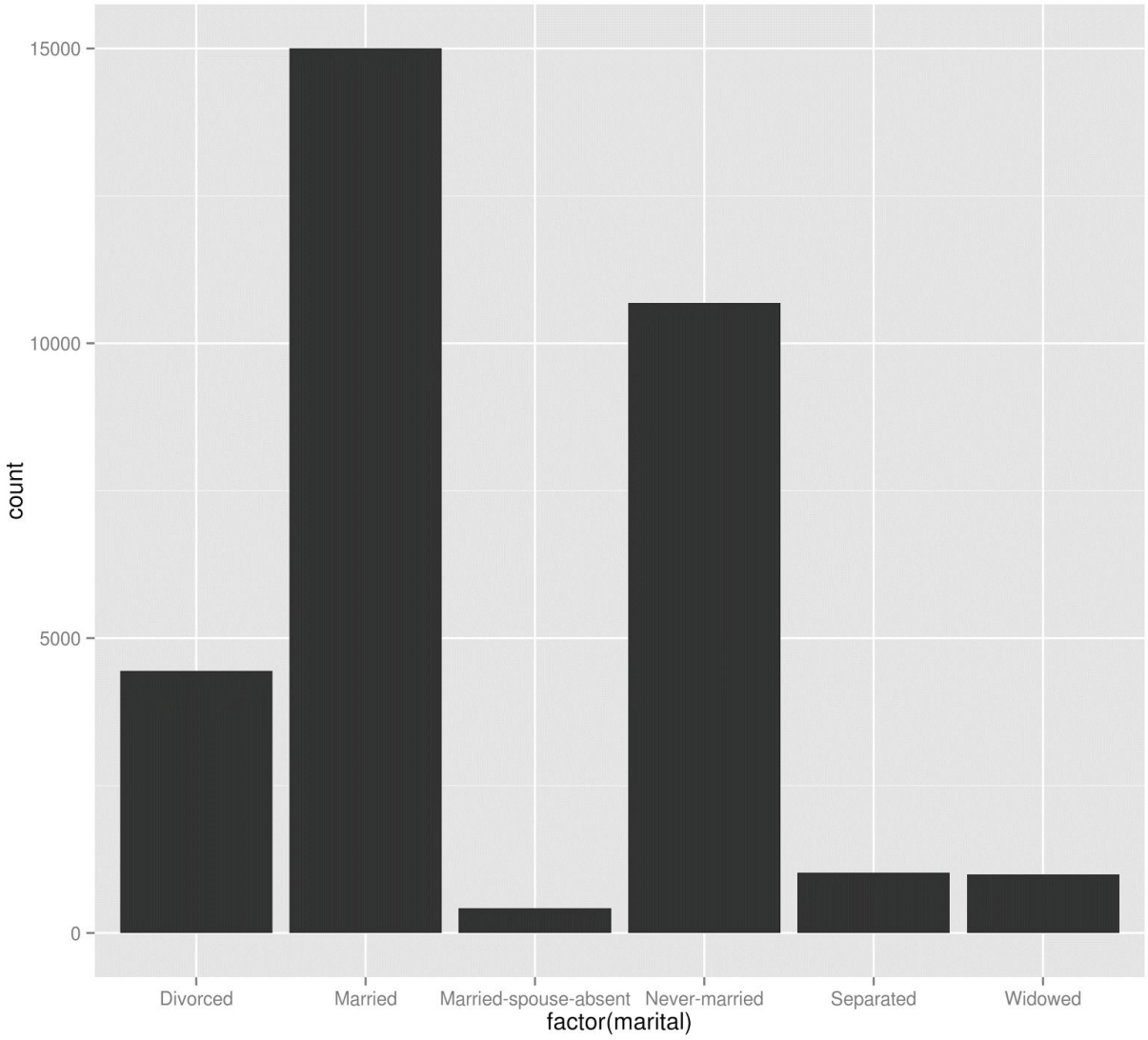
While plotting age against income, we have a right skewed plot (non-symmetric) for income <= 50k. But for income > 50k we have a symmetric plot where mean = median = mode. While plotting hrs_per_week against income, we learn that most people work around 40 hrs per week. From the boxplots, we derive the fact that the median age of people earning > 50k is higher than people earning <=50k. Also, the median hours of work per week is the same irrespective of the income.

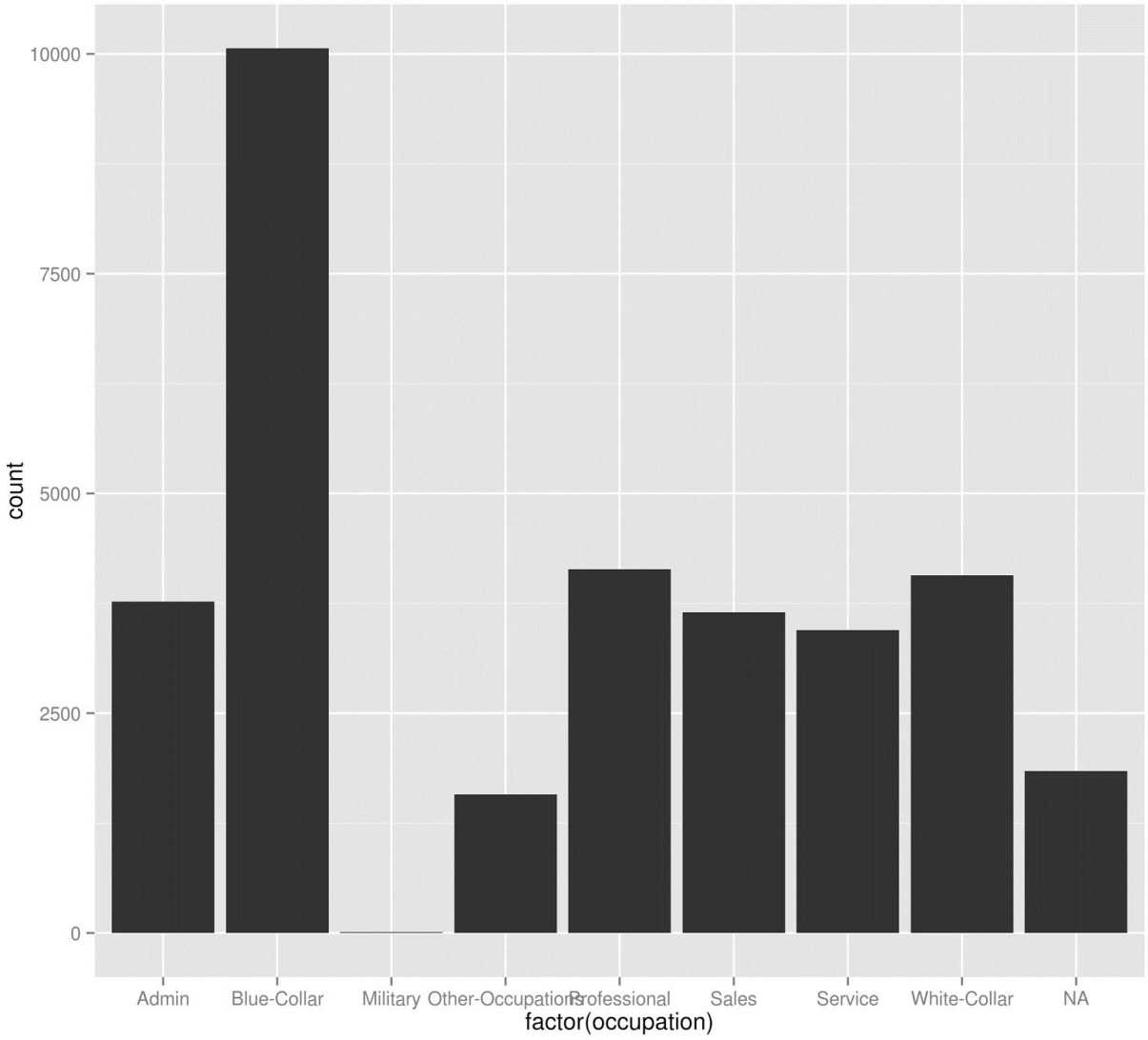
Question 4:

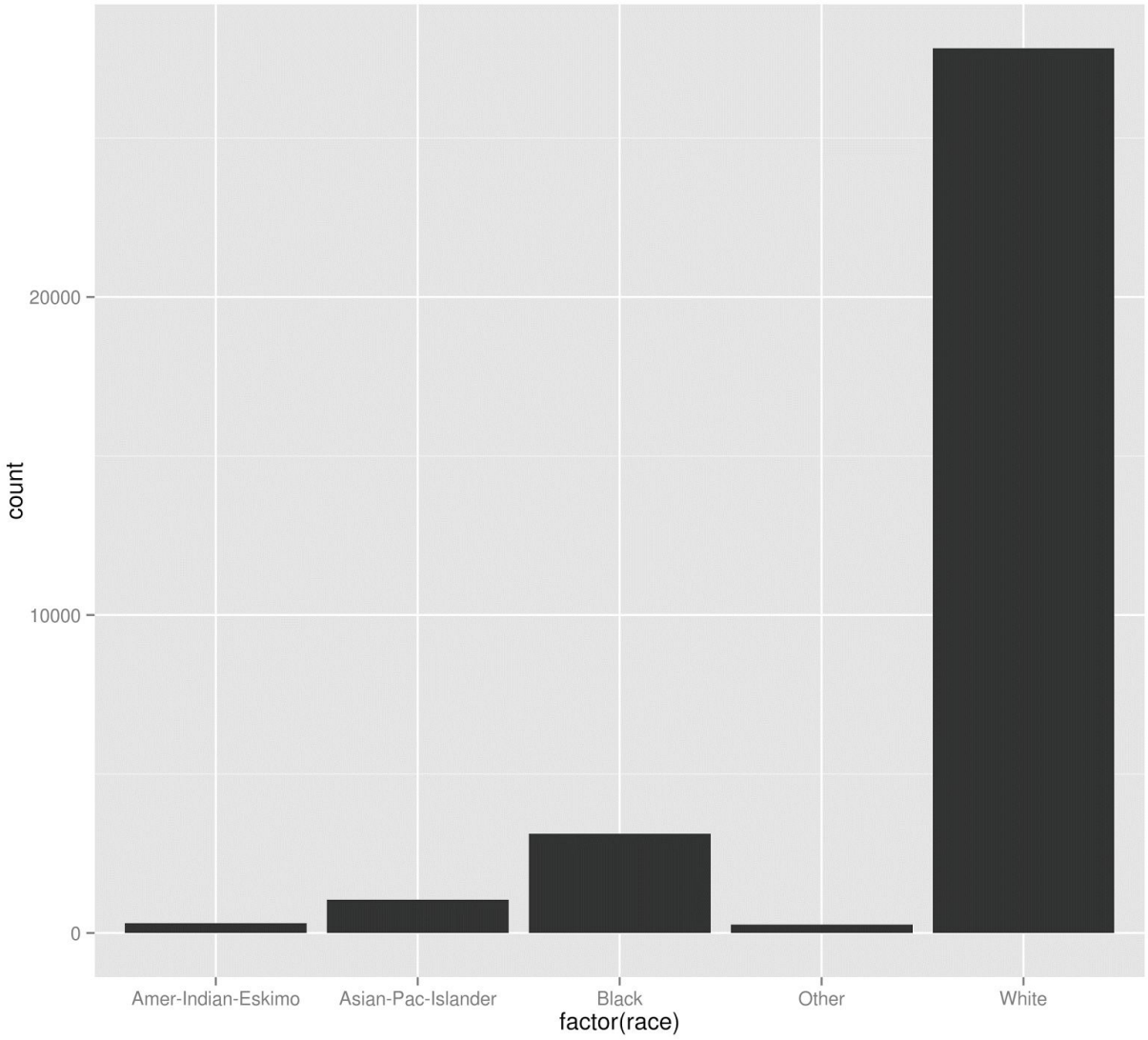
(A):

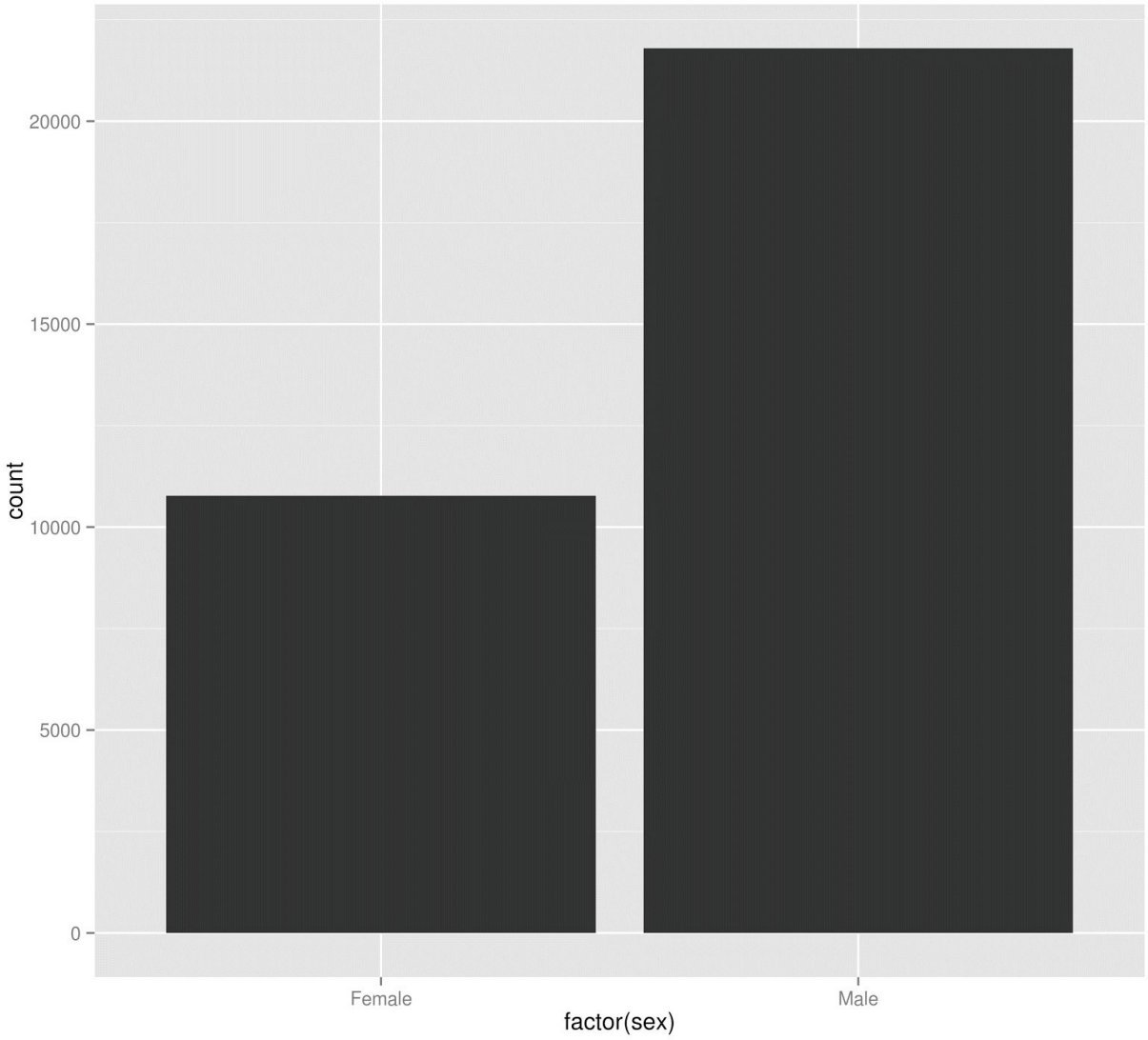

20000
15000
10000
5000
0
count
Federal-gov
Local-gov
Not-Working
Private
Self-emp
State-gov
NA
factor(work)

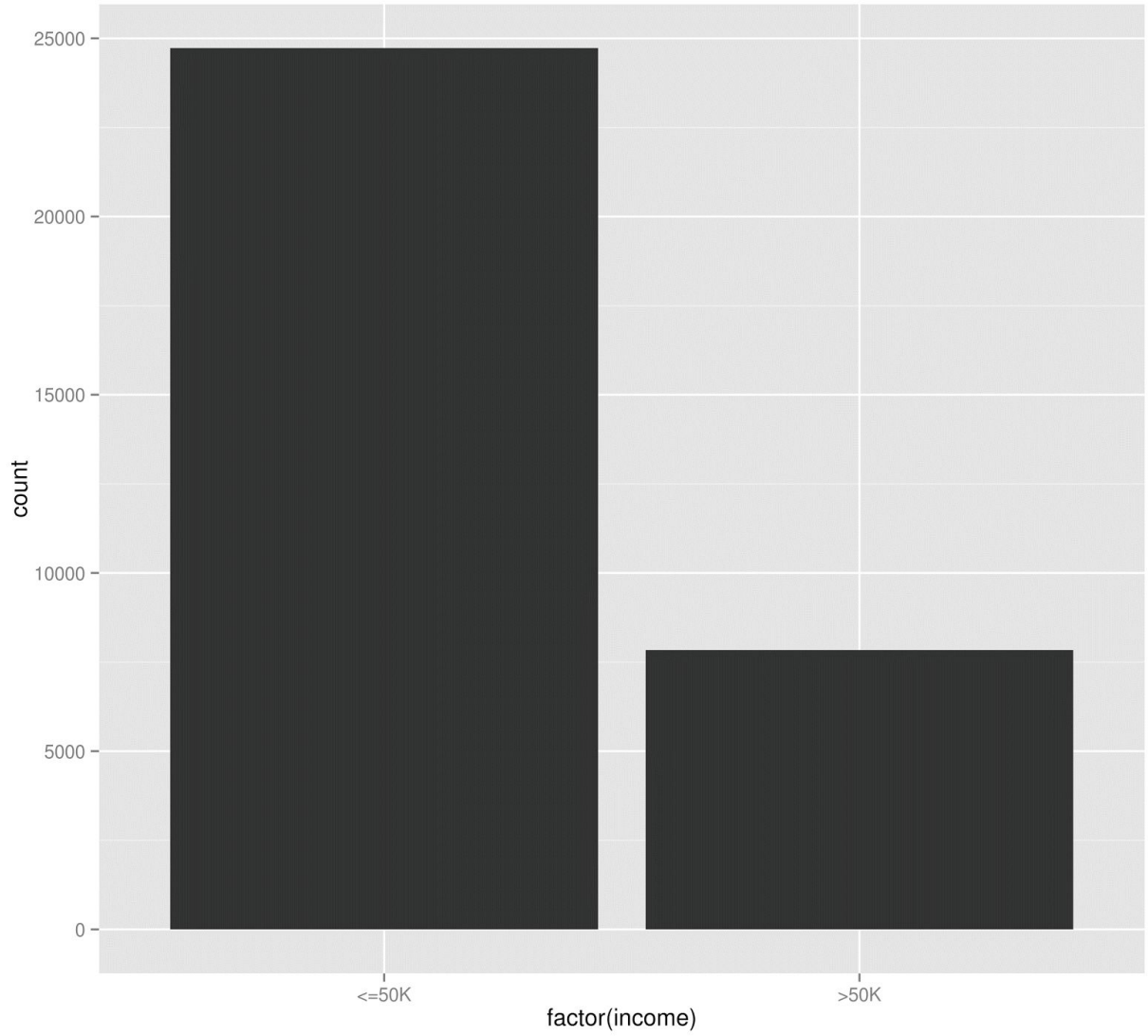

15000
10000
5000
0
count
Associates
Bachelors
Doctorate
Dropout
HS-grad
Masters
Prof-school
factor(edu)


15000
10000
5000
0
count
Divorced
Married
Married-spouse-absent
Never-married
Separated
Widowed
factor(marital)


10000
7500
5000
2500
0
count
Admin
Blue-Collar
Military
Other-Occupations
Professional
Sales
Service
White-Collar
NA
factor(occupation)


20000
10000
0
count
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White
factor(race)

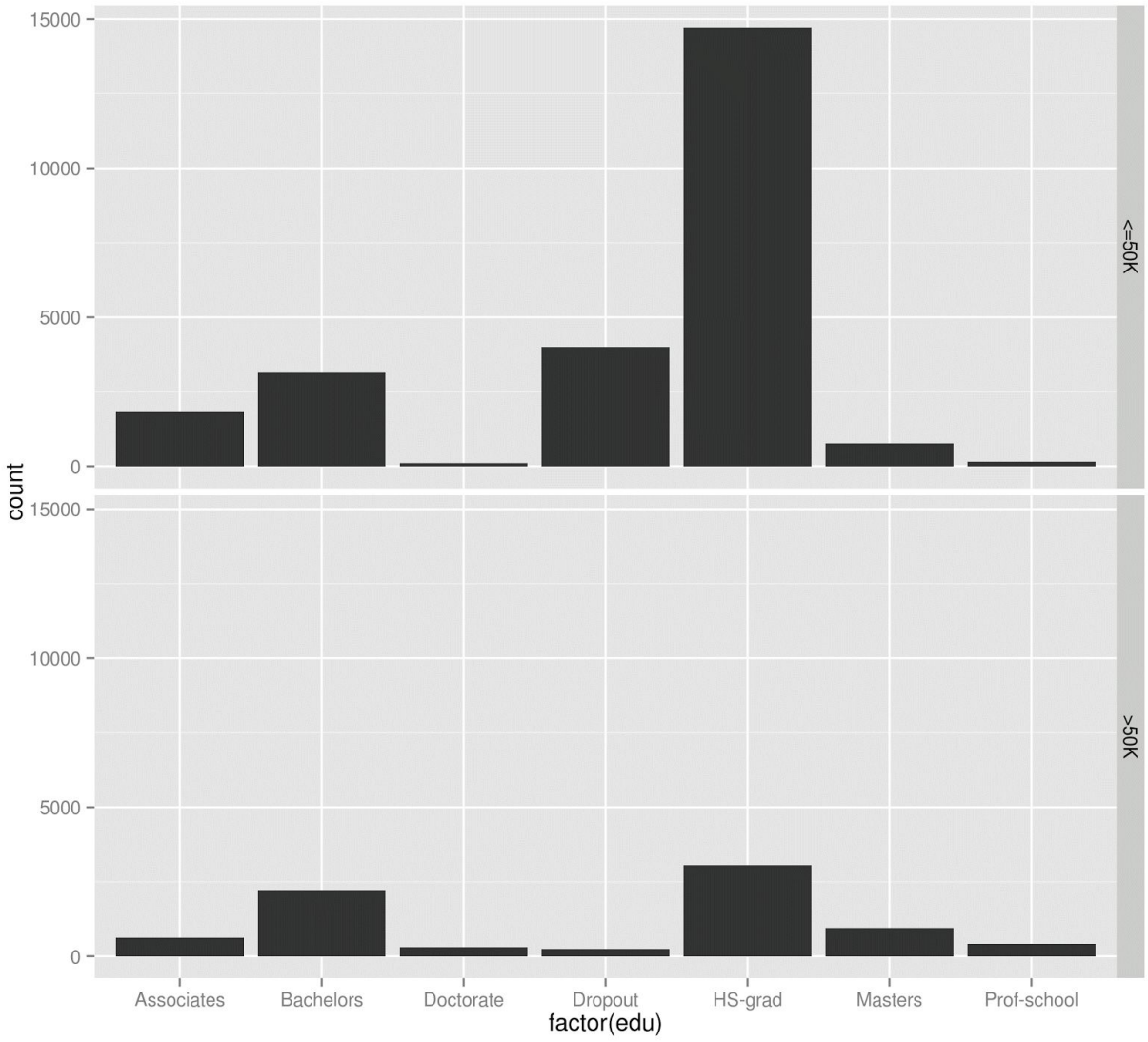

20000
15000
10000
5000
0
count
Female
Male
factor(sex)

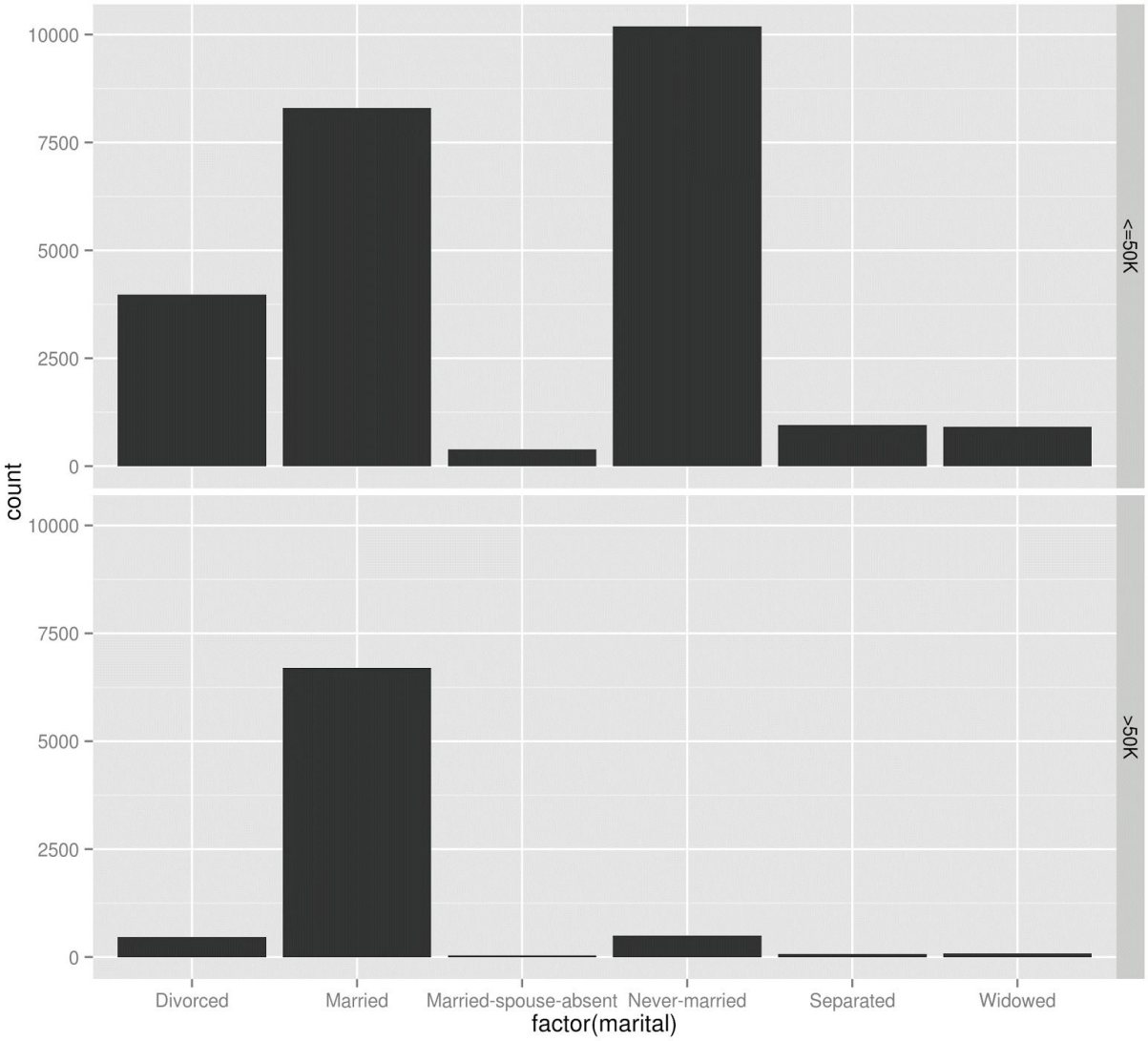


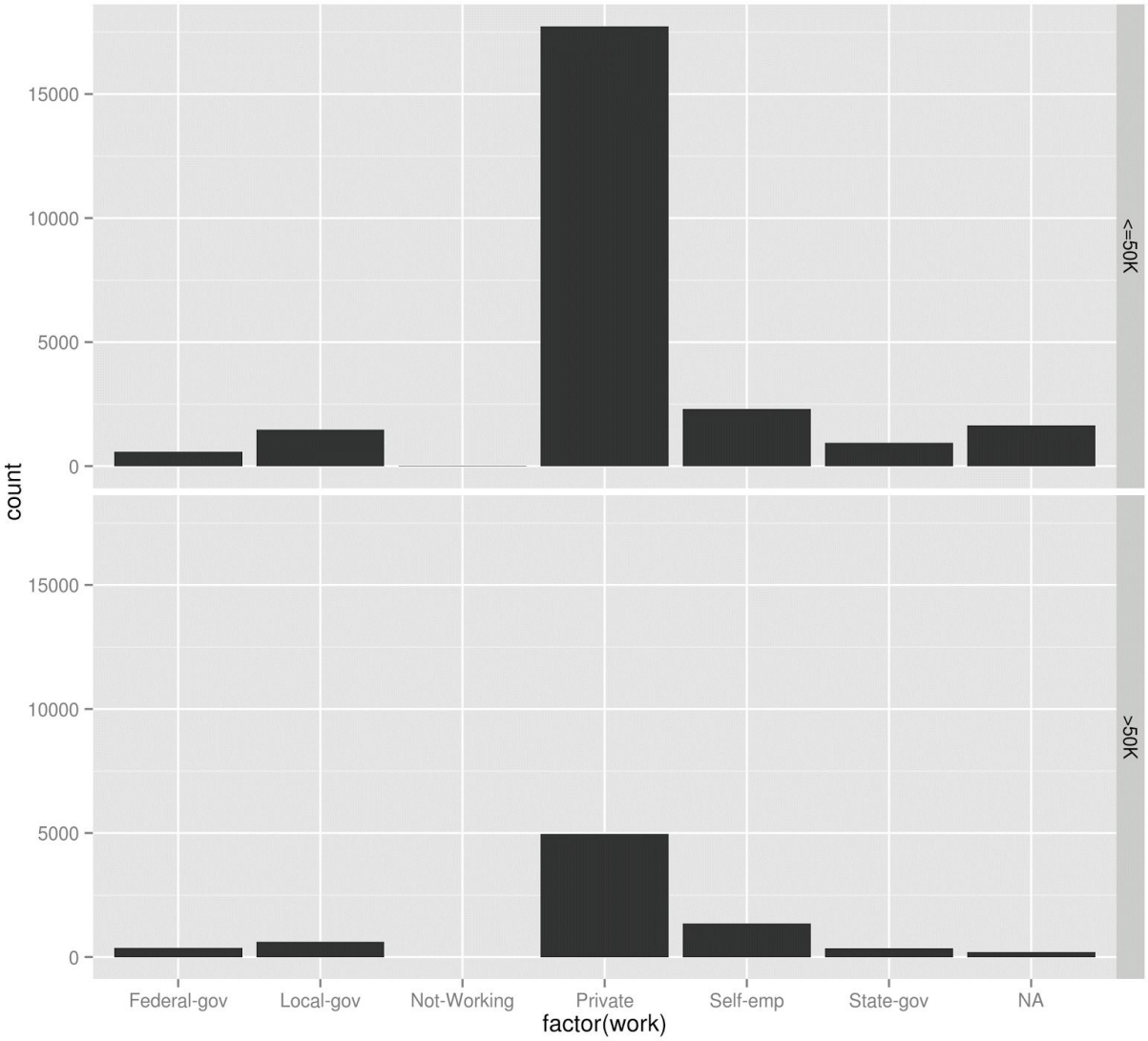
While computing unique values, we exclude NA from the calculation, then we get:

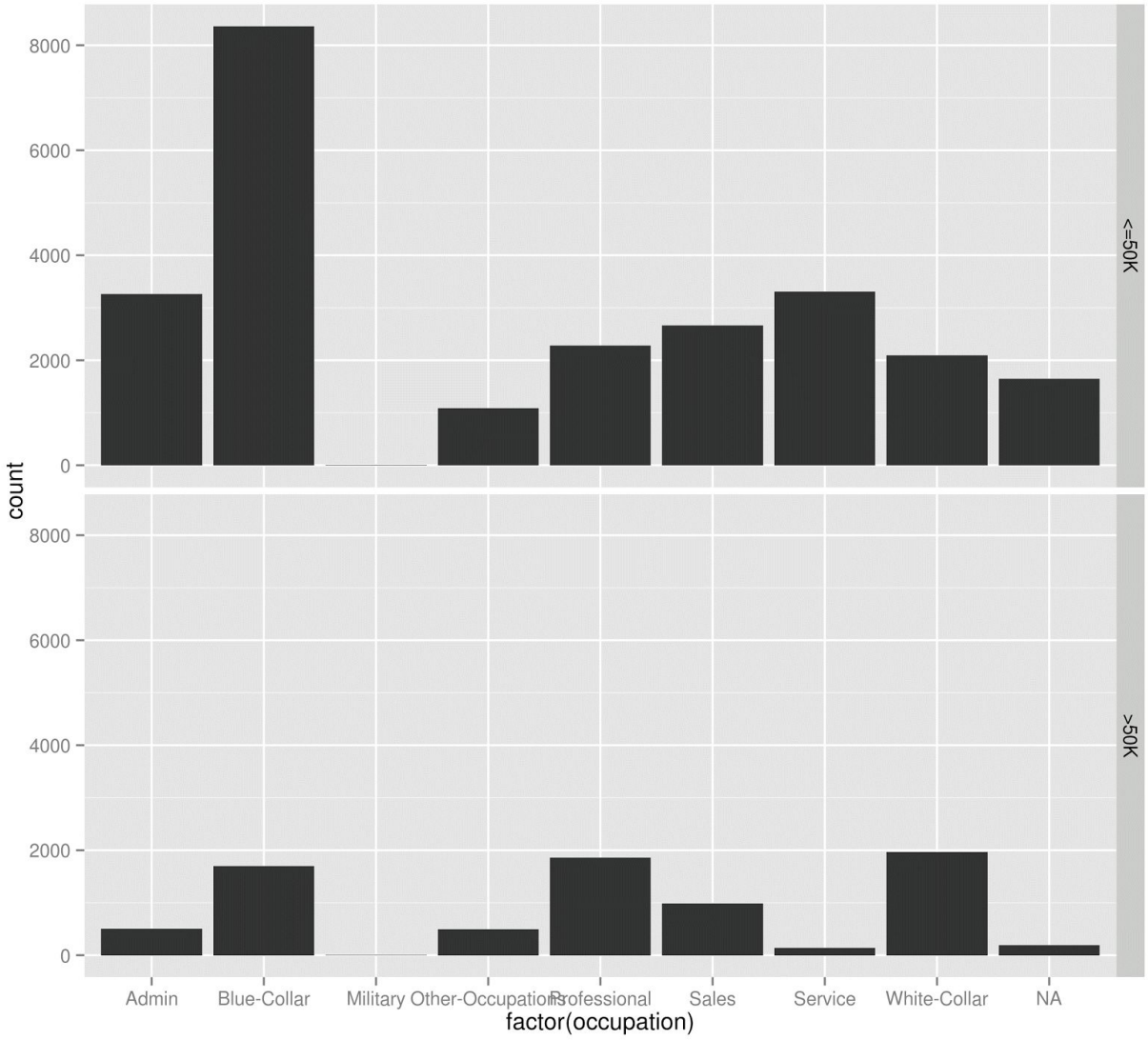
the number of unique values in work : 6
the number of unique values in edu : 7
the number of unique values in marital : 6
the number of unique values in occupation : 8
the number of unique values in race : 5
the number of unique values in sex : 2
the number of unique values in income : 2

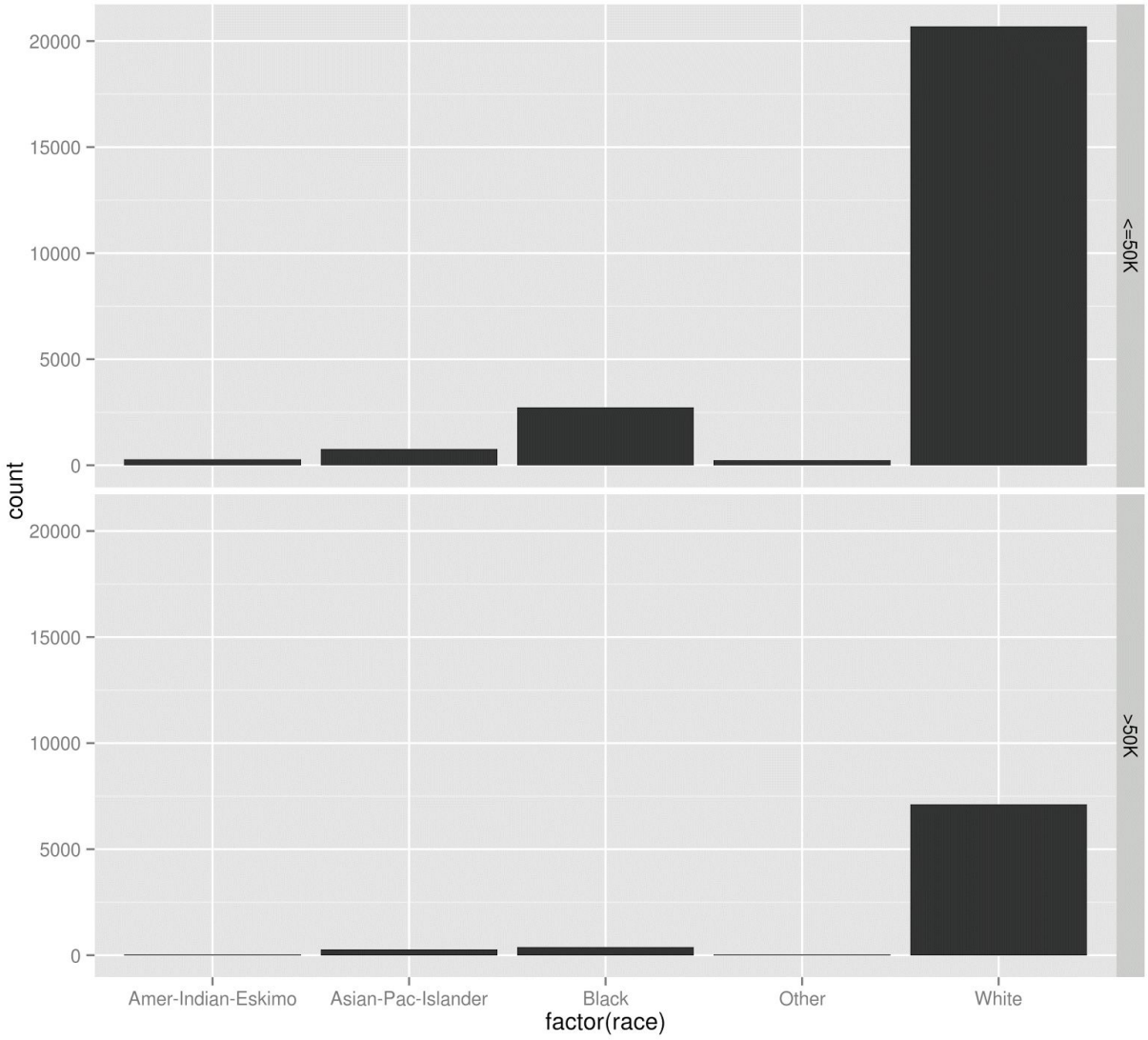
(B):

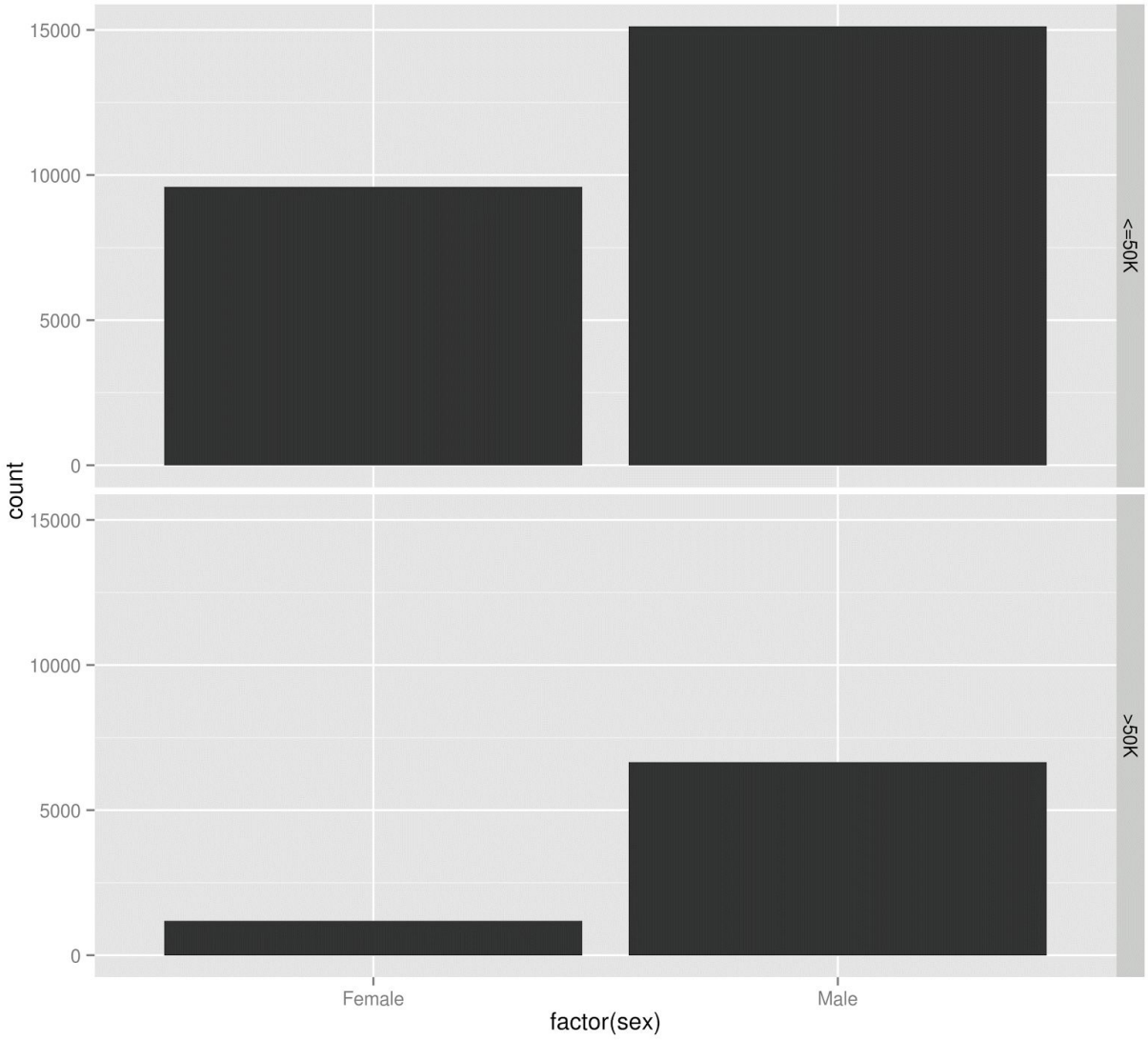

15000
10000
5000
0
count
<=50K
>50K
Associates
Bachelors
Doctorate
Dropout
HS-grad
Masters
Prof-school
factor(edu)

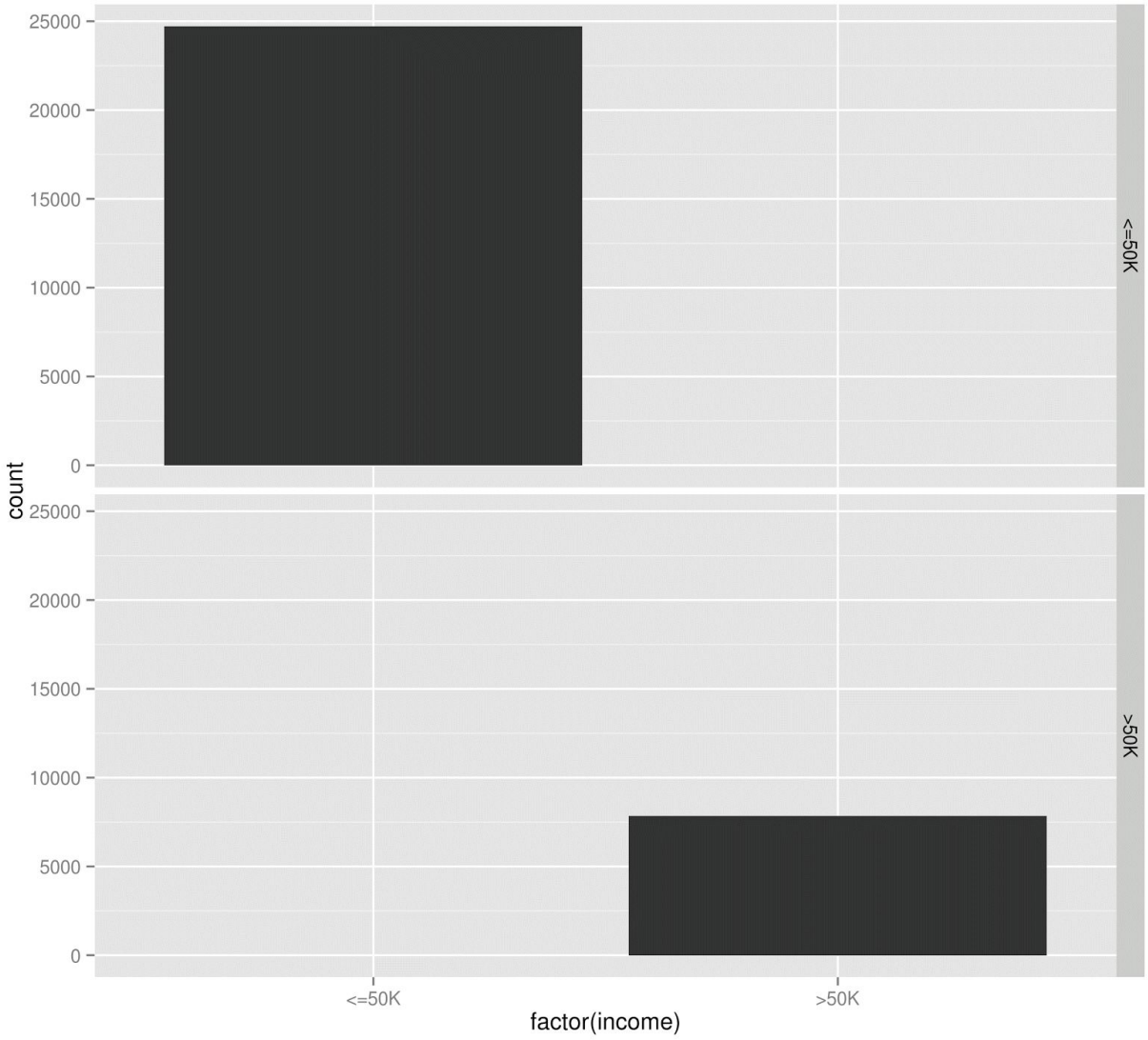

10000
7500
5000
2500
0
count
<=50K
>50K
Divorced
Married
Married-spouse-absent
Never-married
Separated
Widowed
factor(marital)


15000
10000
5000
0
count
<=50K
>50K
Federal-gov
Local-gov
Not-Working
Private
Self-emp
State-gov
NA
factor(work)


8000
6000
4000
2000
0
count
<=50K
>50K
Admin
Blue-Collar
Military
Other-Occupations
Professional
Sales
Service
White-Collar
NA
factor(occupation)


20000
15000
10000
5000
0
count
<=50K
>50K
Amer-Indian-Eskimo
Asian-Pac-Islander
Black
Other
White
factor(race)


15000
10000
5000
0
count
<=50K
>50K
Female
Male
factor(sex)



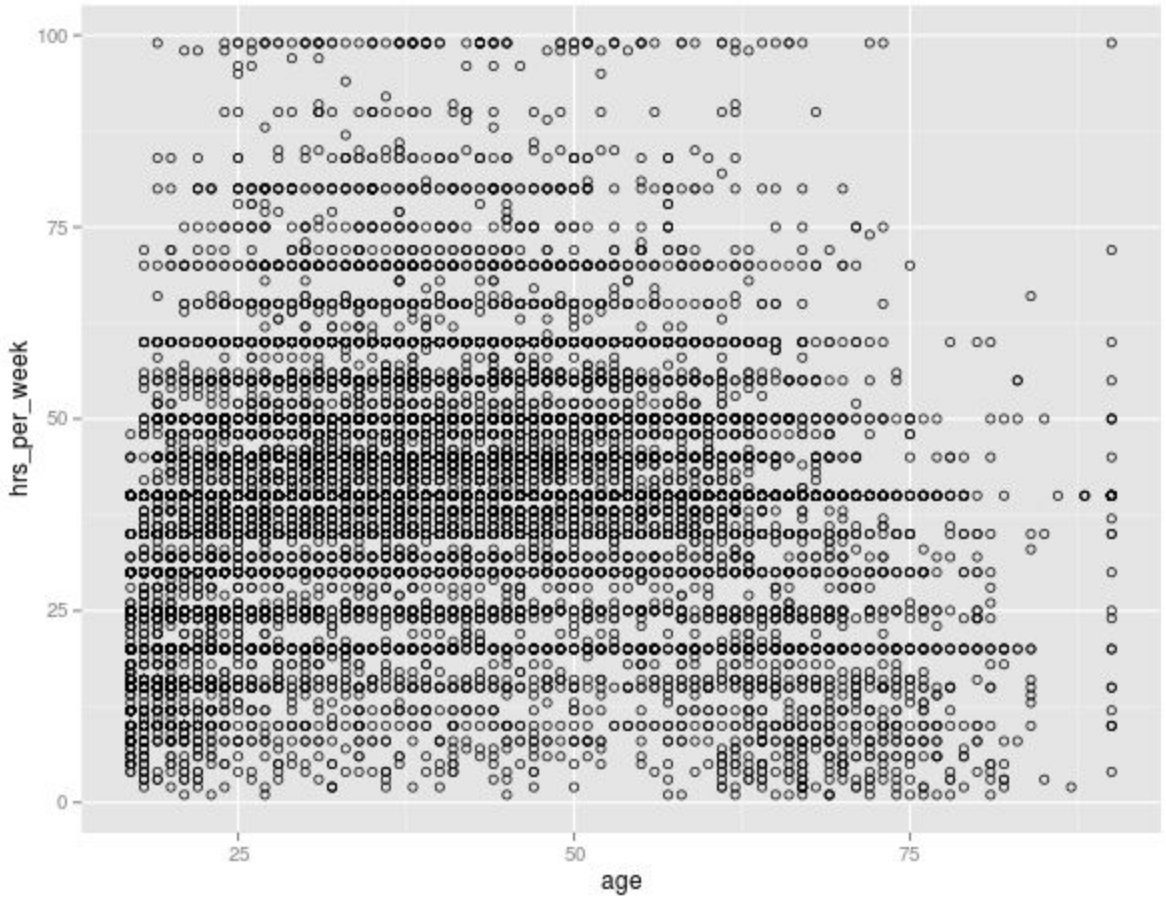
(C):

From 4 (A), we learn that majority of the people surveyed seem to be educated upto atleast High School, employed by a Private concern or do some blue-collar job. Also the sampled data has more married people and don't have a liking for Military service.

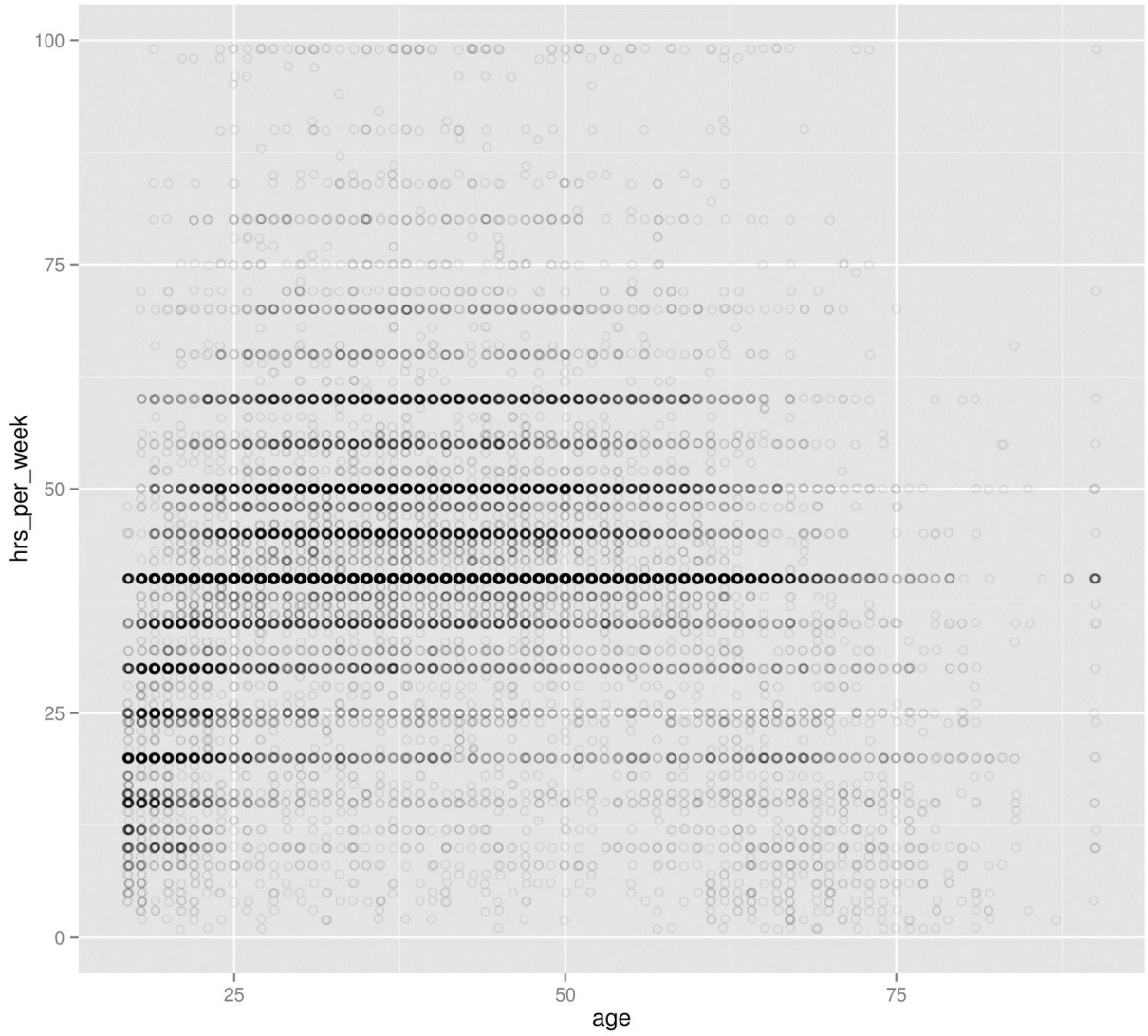
From 4 (B), we learn the economic stratification of the society. This gives us an insight that people who never married earn less than married people in the >50k bracket. However this trend reverses in the <=50k bracket. We also learn that people don't join Military service for the paycheck as people earning >50k is under-represented.

Question 5:

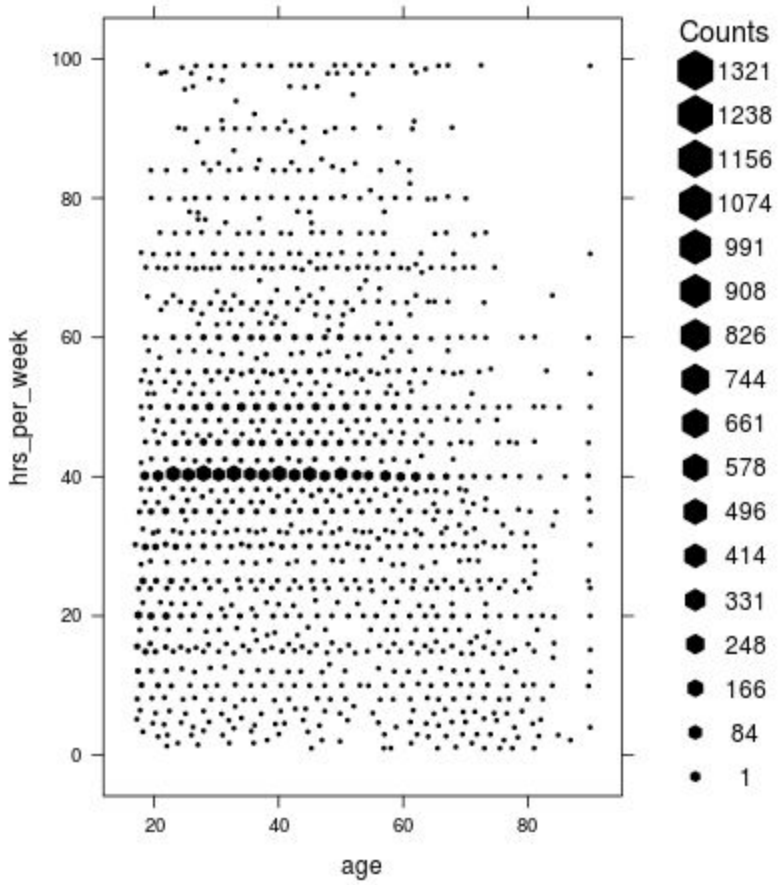
(A):



Yes, there is overplotting in this plot since many points overlap. We can handle this using alpha(opacity) and adding jitter.



Using hexbin package, we can display hexagonally binned plot, which is another method of handling overplotting. This is shown below:



These plots describe the fact that the productivity of the people of even the same age group is not uniform but there are many people who work the same hours per week on average hence leading to overplotting.

The correlation coefficient between age and hrs_per_week has been calculated as 0.06875571.