Income Prediction. Classification Predictive Modeling

by Anupama r.k, Queenie Tsang, Crystal (Yunan) Zhu

12/02/2021

Modeling and Evaluation

Describe the data

Data Dictionary

```
## The dimension of the dataset is 32561 by 15 .
```

There are 32,561 records and 15 columns in the original data set.

There are 6 numeric and 9 categorical variables shown as follows:

Column Name	Data Type	Column Description
age	Integer	The age of the adult (e.g., 39, 50, 38, etc.)
workclass	Factor	The work class of the adult (e.g., Private, Self-emp-not-inc, Federal-gov, etc.)
fnl_wgt	Integer	The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US (e.g., 77516, 83311, etc.)
education	Factor	The education of the adult (e.g., Bachelors, Some-college, 10th, etc.)
education_num	Integer	The number years of the adult's education (e.g., 13, 9, 7, etc.)
marital_status	Factor	The marital status of the adult (e.g., Divorced, Never-married, Separated, etc.)
occupation	Factor	The occupation of the adult (e.g., Tech-support, Craft-repair, Sales, etc.)
relationship	Factor	The relationship of the adult in a family (e.g., Wife, Own-child, Husband, etc.)
race	Factor	The race of the adult (e.g., White, Asian-Pac-Islander, Amer-Indian-Eskimo, etc.)
sex	Factor	The gender of the adult.(Female, Male)
capital_gain	Integer	The capital gain of the adult (e.g., 0, 2174, 14084, etc.)

Column Name	Data Type	Column Description
capital_loss	Integer	The capital loss of the adult (e.g., 0, 1408,2042, etc.)
hours_per_week	Integer	The number of working hours each week for the adult (e.g. 40, 13, 16, etc.)
native_country	Factor	The native country of the adult (e.g. Cambodia, Canada, Mexico, etc.)
income	Factor	The yearly income of the adult at 2 levels: <=50K and >50K.

Data Description

First, let's check whether there are duplicates in the dataset.

```
## The number of duplicated records in the dataset is 24 .
```

For the benefit of this report's length, let's look at a sample of duplicated records:

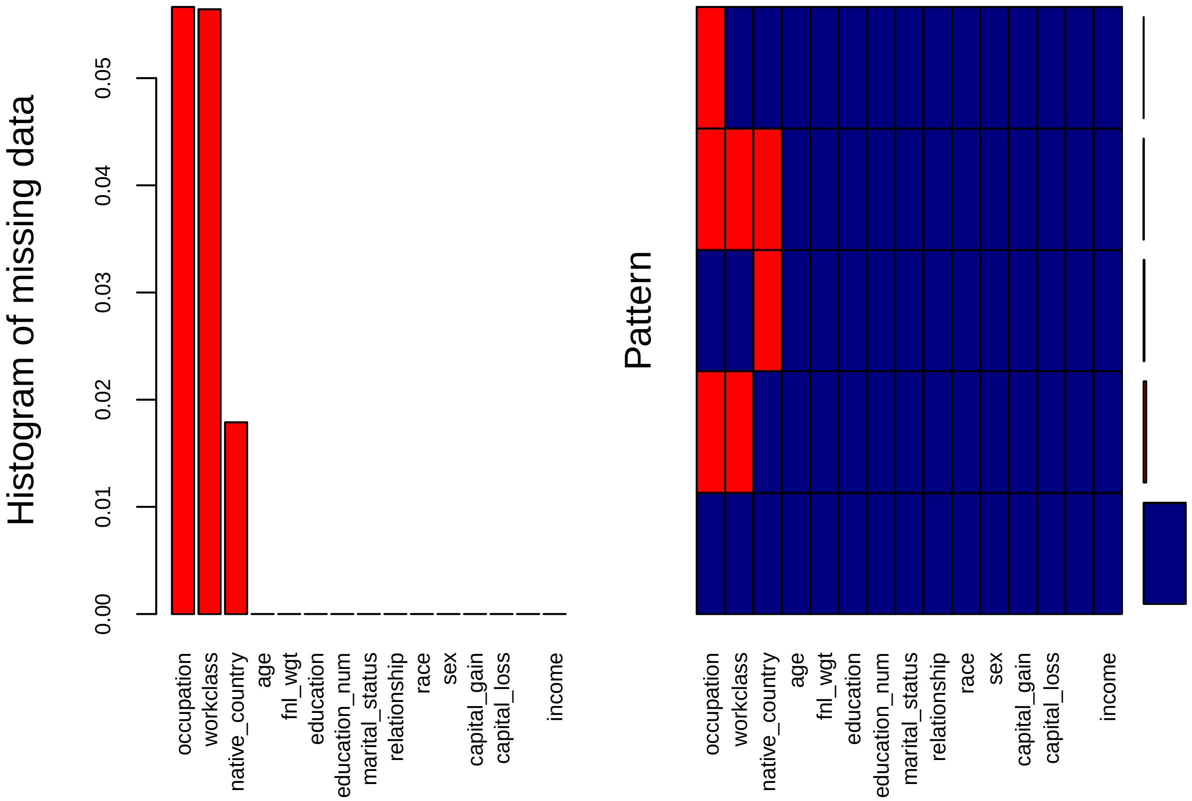
	age	workclass	fnl_wgt	education	education_num	marital_status	occupation	relationship
4768	21	Private	250051	Some-college	10	Never-married	Prof-specialty	Own-child
9172	21	Private	250051	Some-college	10	Never-married	Prof-specialty	Own-child
4326	25	Private	308144	Bachelors	13	Never-married	Craft-repair	Not-in-family
4882	25	Private	308144	Bachelors	13	Never-married	Craft-repair	Not-in-family

	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
4768	White	Female	0	0	10	United-States	<=50K
9172	White	Female	0	0	10	United-States	<=50K
4326	White	Male	0	0	40	Mexico	<=50K
4882	White	Male	0	0	40	Mexico	<=50K

The 24 duplicated rows will be removed from all later analysis.

Then let's check whether there are any missing values in the dataset.

```
## Warning in plot.aggr(res, ...): not enough horizontal space to display
## frequencies
```

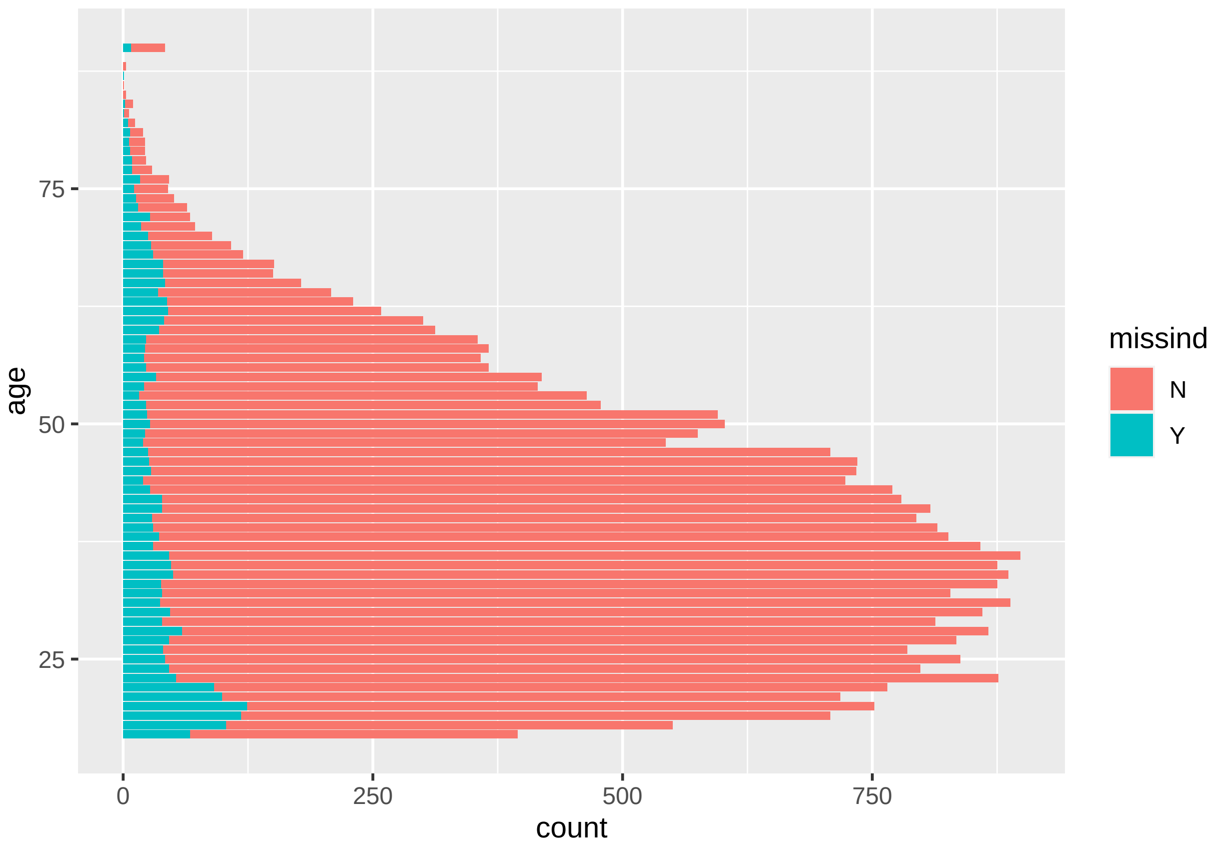


```
##
##  Variables sorted by number of missings:
##        Variable      Count
##      occupation 0.05664321
##       workclass 0.05642807
##  native_country 0.01788733
##             age 0.00000000
##         fnl_wgt 0.00000000
##       education 0.00000000
##   education_num 0.00000000
##  marital_status 0.00000000
##    relationship 0.00000000
##            race 0.00000000
##             sex 0.00000000
##    capital_gain 0.00000000
##    capital_loss 0.00000000
##  hours_per_week 0.00000000
##          income 0.00000000
```

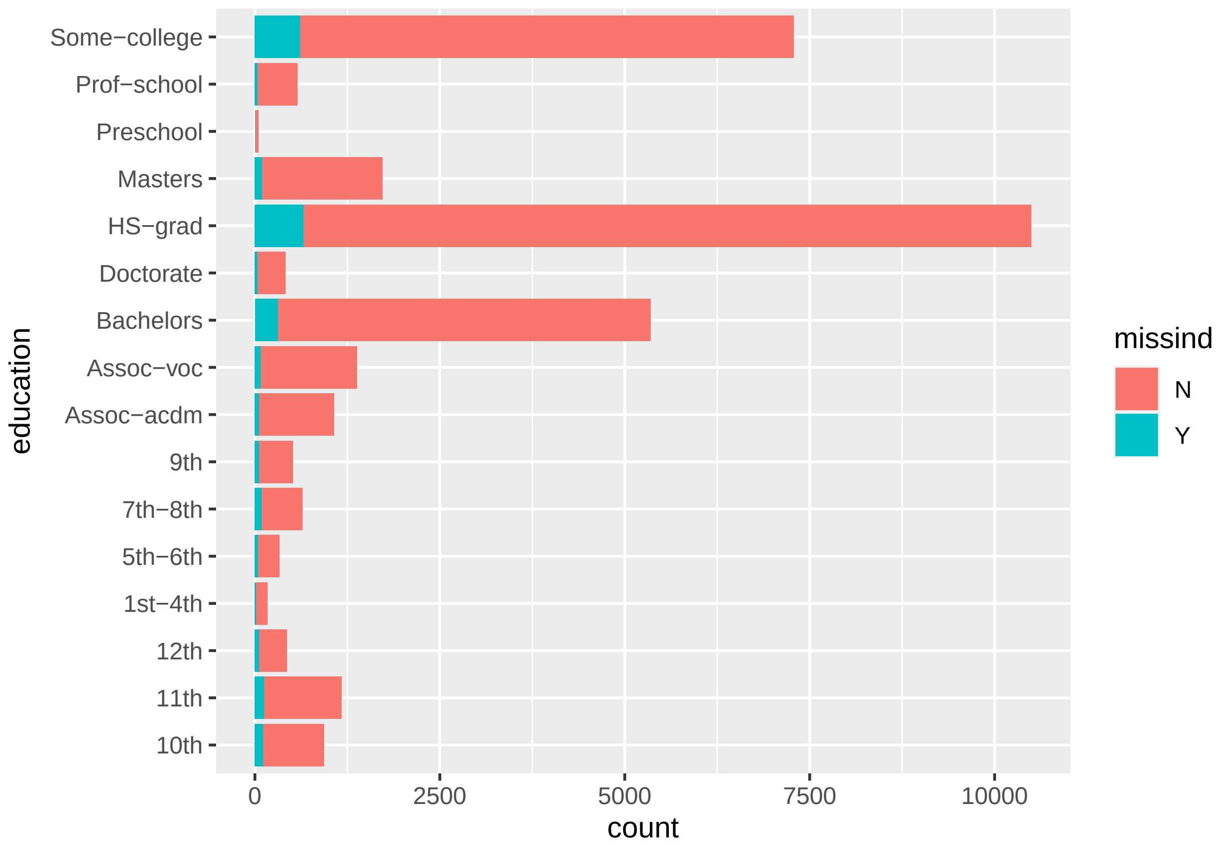
From the above, there are missing values in this data set and all the missing values are from categorical variables.

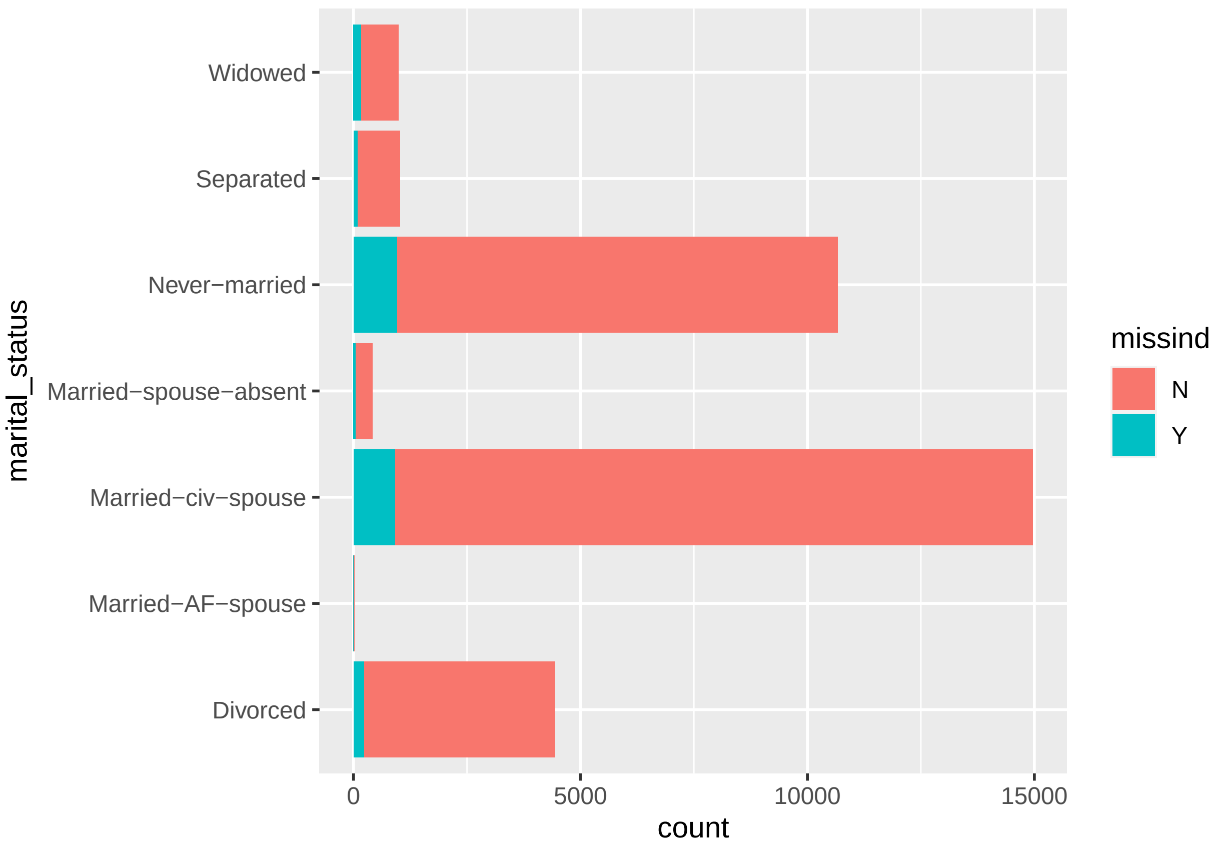
Comparing records with at least one missing value to those without any missing values.

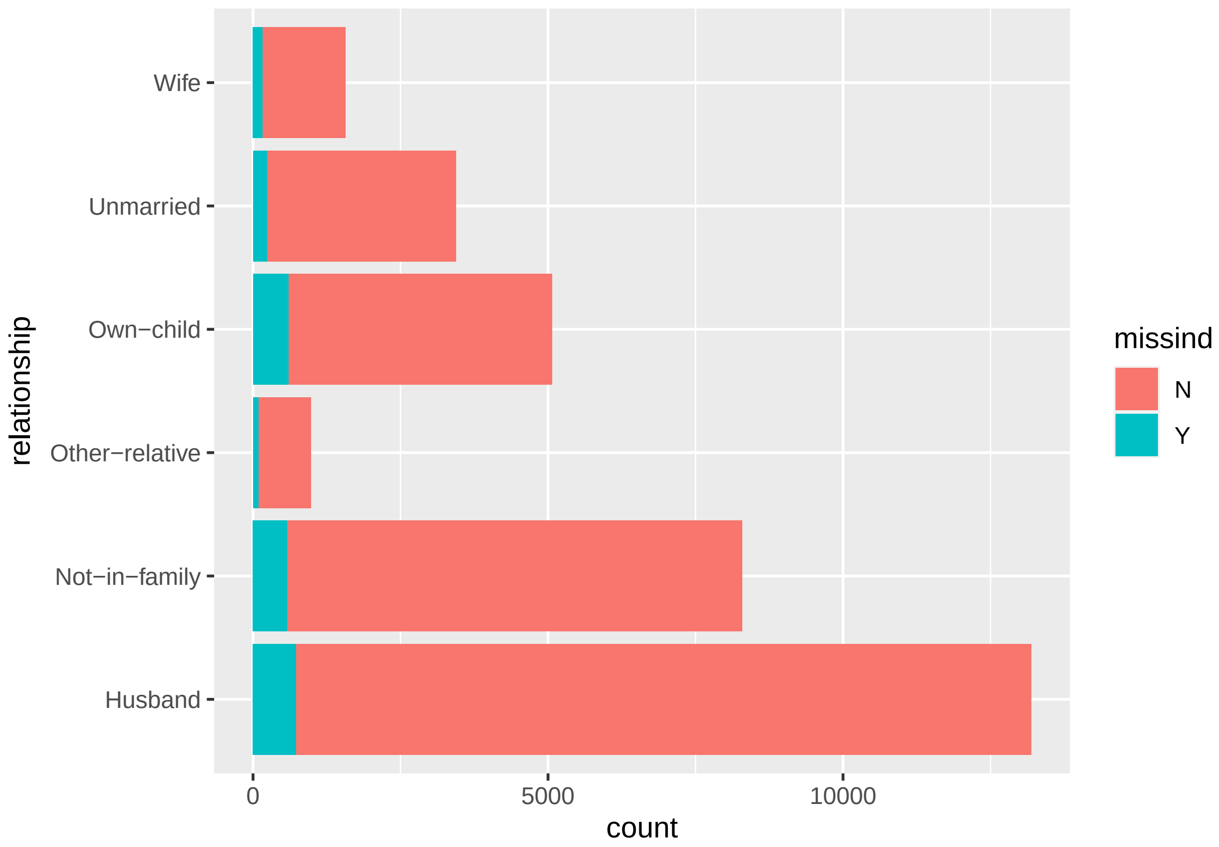
In order to better understand the patterns of the missing values, let's look at some descriptions of the records with missing values.

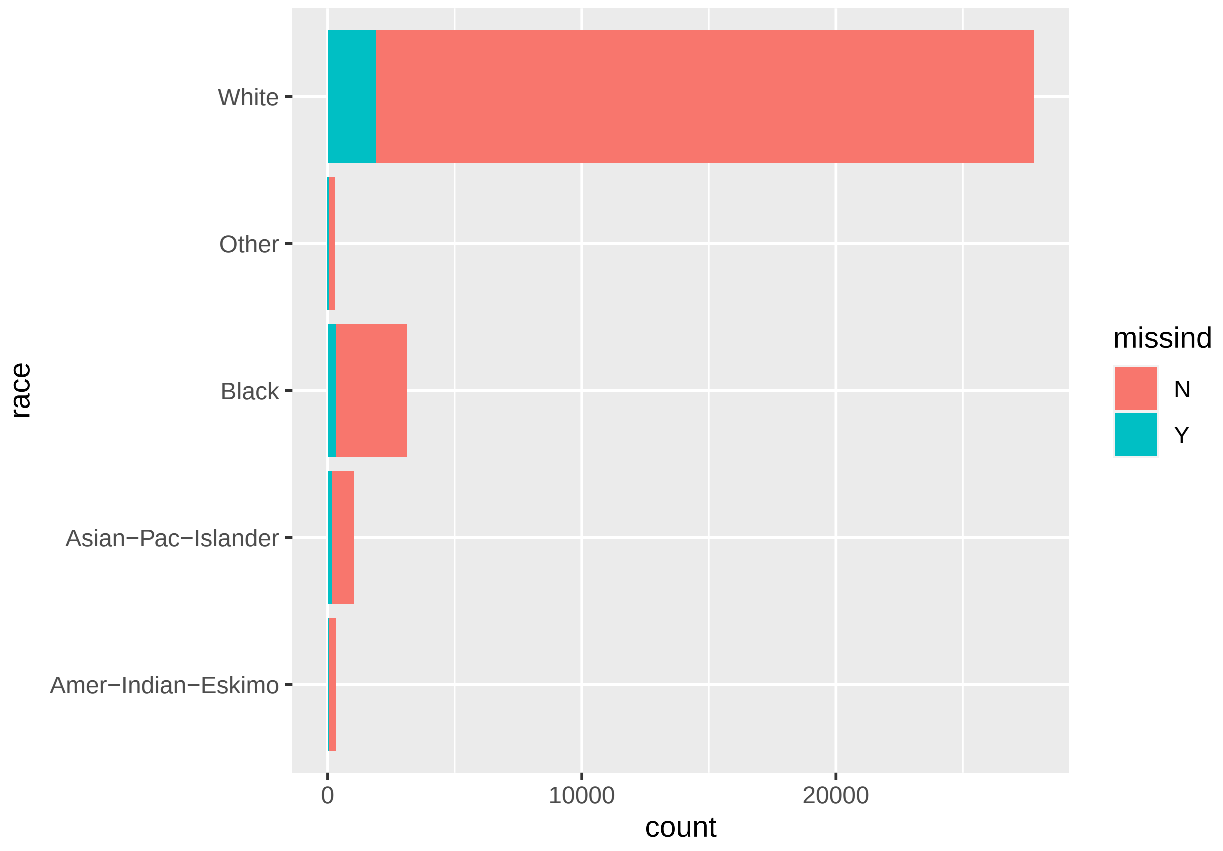


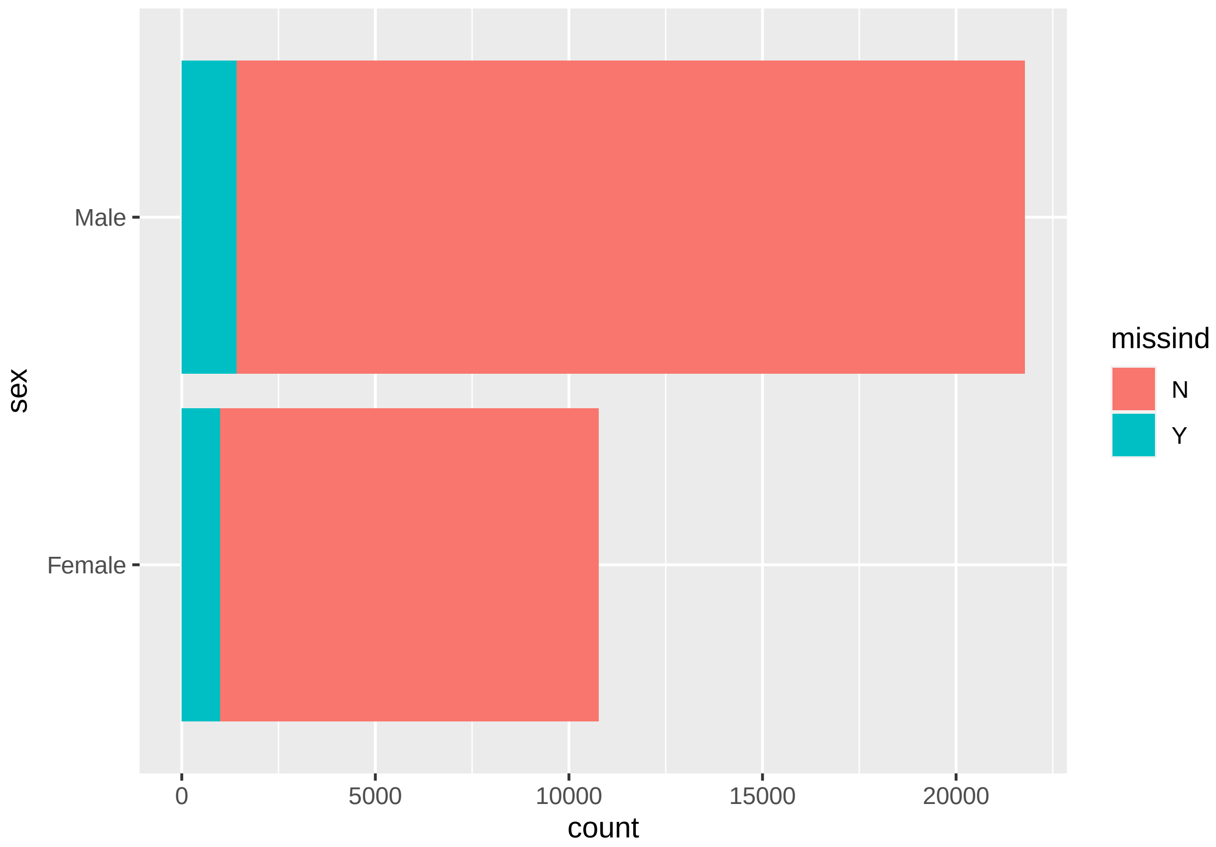
age
75
50
25
0
250
500
750
count
missind
N
Y

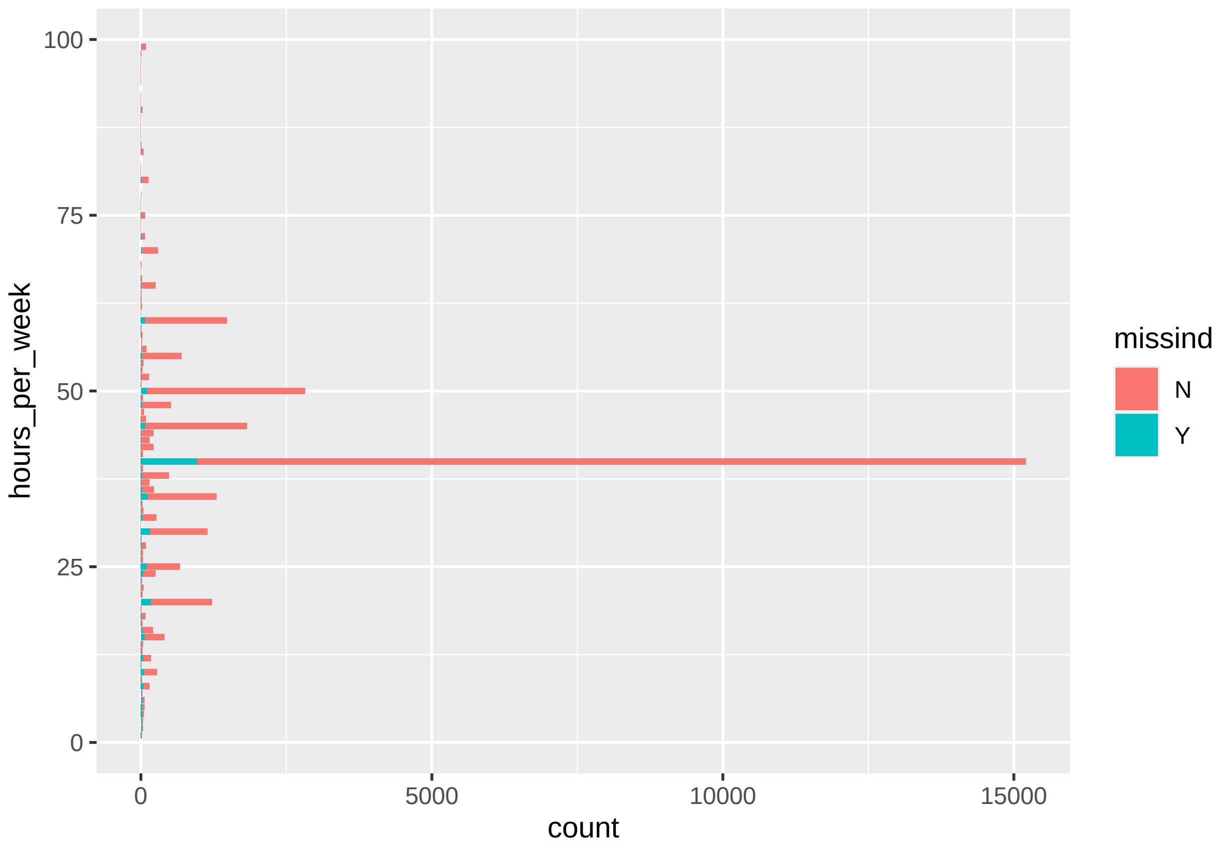

Some−college
Prof−school
Preschool
Masters
HS−grad
Doctorate
Bachelors
Assoc−voc
Assoc−acdm
9th
7th−8th
5th−6th
1st−4th
12th
11th
10th
education
0
2500
5000
7500
10000
count
missind
N
Y

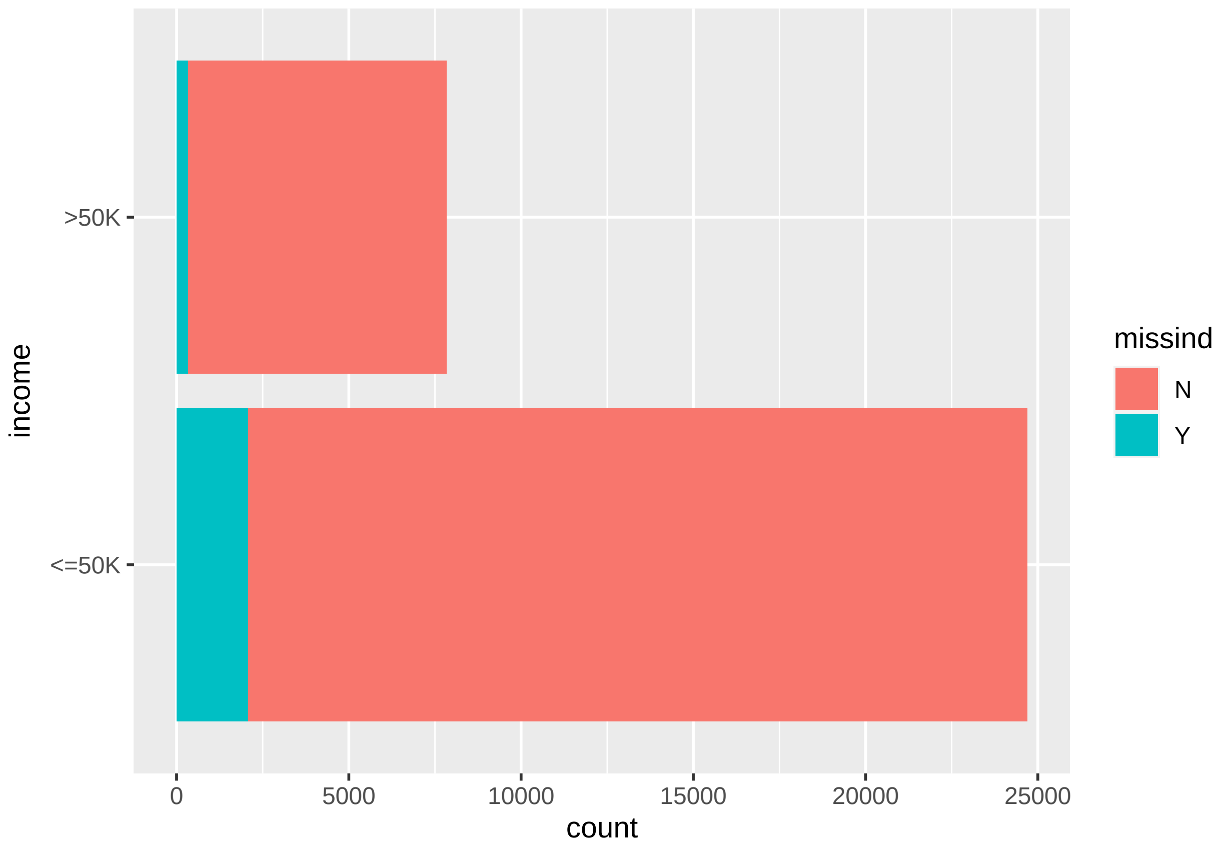

Widowed
Separated
Never−married
Married−spouse−absent
Married−civ−spouse
Married−AF−spouse
Divorced
marital_status
0
5000
10000
15000
count
missind
N
Y


Wife
Unmarried
Own−child
Other−relative
Not−in−family
Husband
relationship
0
5000
10000
count
missind
N
Y


White
Other
Black
Asian−Pac−Islander
Amer−Indian−Eskimo
race
0
10000
20000
count
missind
N
Y


sex
Male
Female
0
5000
10000
15000
20000
count
missind
N
Y


hours_per_week
100
75
50
25
0
count
0
5000
10000
15000
missind
N
Y



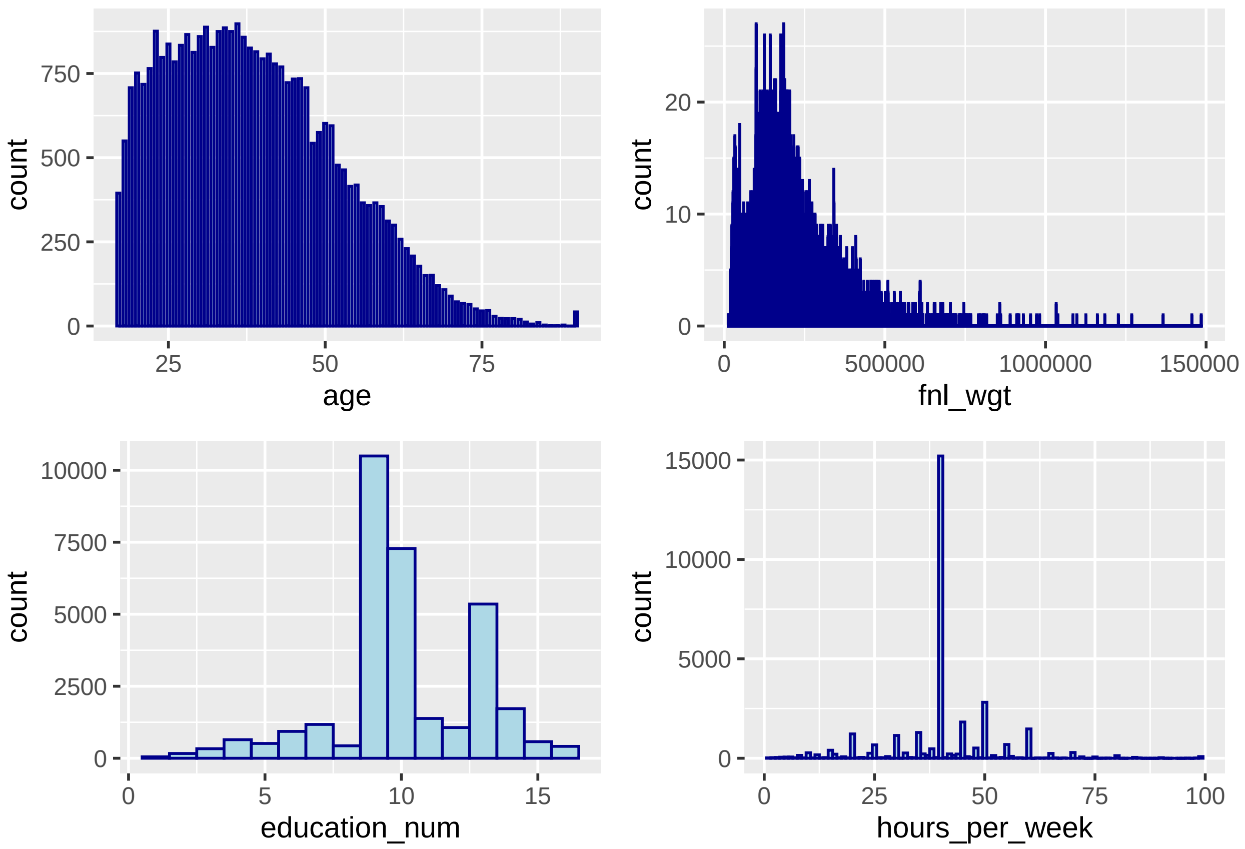
From the above bar charts comparing the distributions of 7 variables of the group that do not have missing values and the group that have at least one missing records, we can see that the missing records are generally evenly distributed across all ages, education level, marital status, family relationship, race, working hours per week and the target variable income. When compared with the whole population in the census, the percentages of records with missing values are having slightly lower percentages in the age group between 20-50, Married civ spouse marital status, husband, and slightly higher percentages for 60-70 years old, never-married. Males tend to have fewer missing records than females.

Since the proportion of missing values is relatively small (7%) where we would have 30K records left, and it's generally the same for people with income higher and lower than 50K USD, we think it would be reasonable to remove the records for our analysis in this report. If we had more time, we'd recommend fitting models separately for female and male since they have different willingness to answer occupation, work class or native country related questions, which could be strong predictors for adult income.

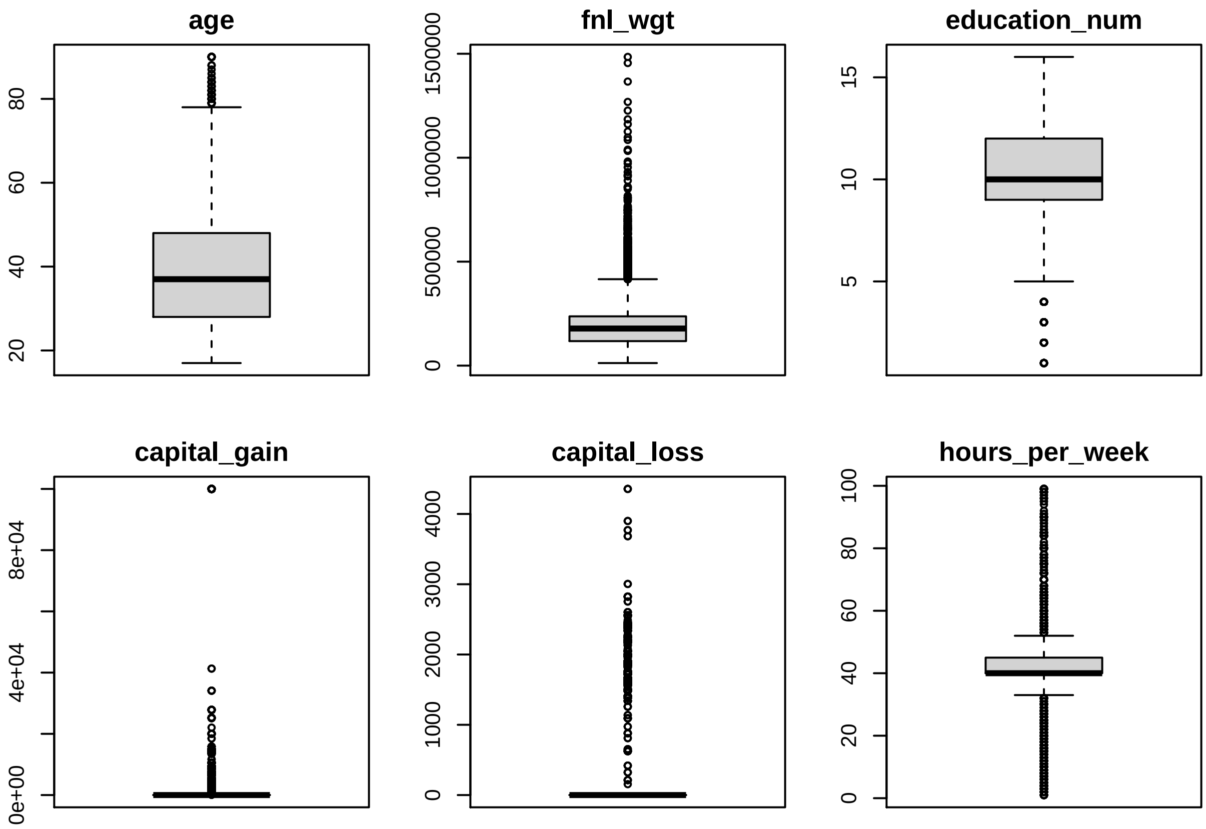
Now let's view the summary of the 6 numeric columns:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
age	17.00	28.00	37.00	38.59	48.00	90.00
fnl_wgt	12285.00	117827.00	178356.00	189780.85	236993.00	1484705.00
education_num	1.00	9.00	10.00	10.08	12.00	16.00
capital_gain	0.00	0.00	0.00	1078.44	0.00	99999.00
capital_loss	0.00	0.00	0.00	87.37	0.00	4356.00
hours_per_week	1.00	40.00	40.00	40.44	45.00	99.00

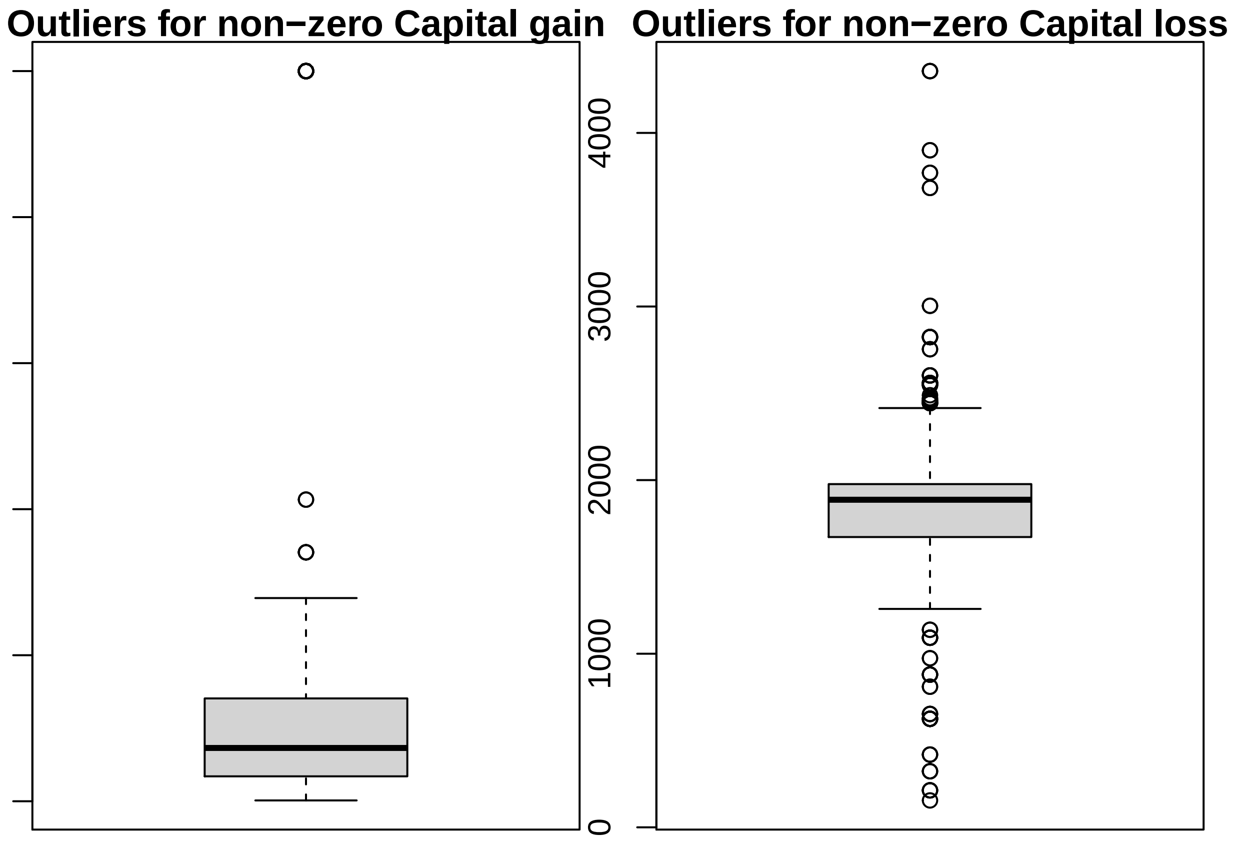
Let's take a clearer look at the numeric values by visualizing their distributions using histograms, except for capital gain and capital loss.



Let’s use boxplots to see whether there are outliers for each numeric variable.

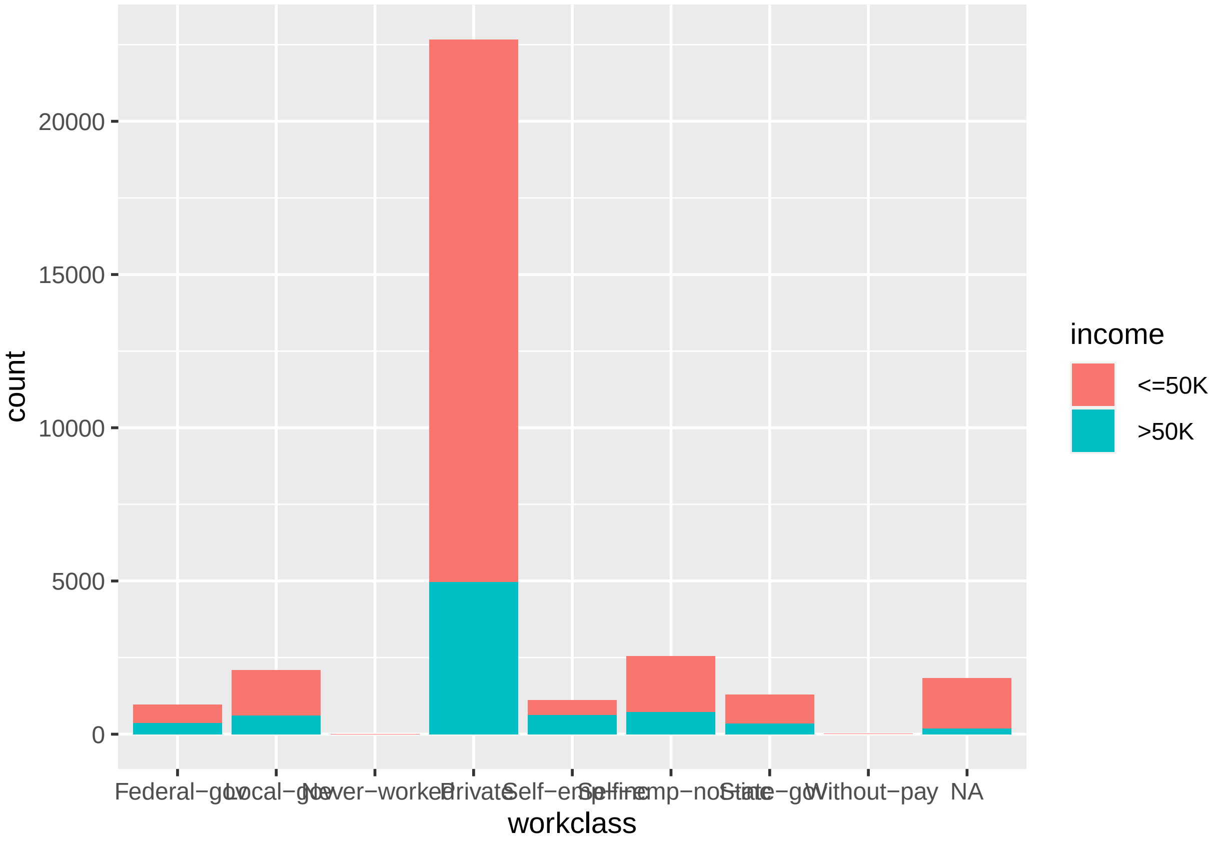


Since there are large number of zeros in capitalgain & capitalloss variables, let's check if there are outliers for non-zero values.

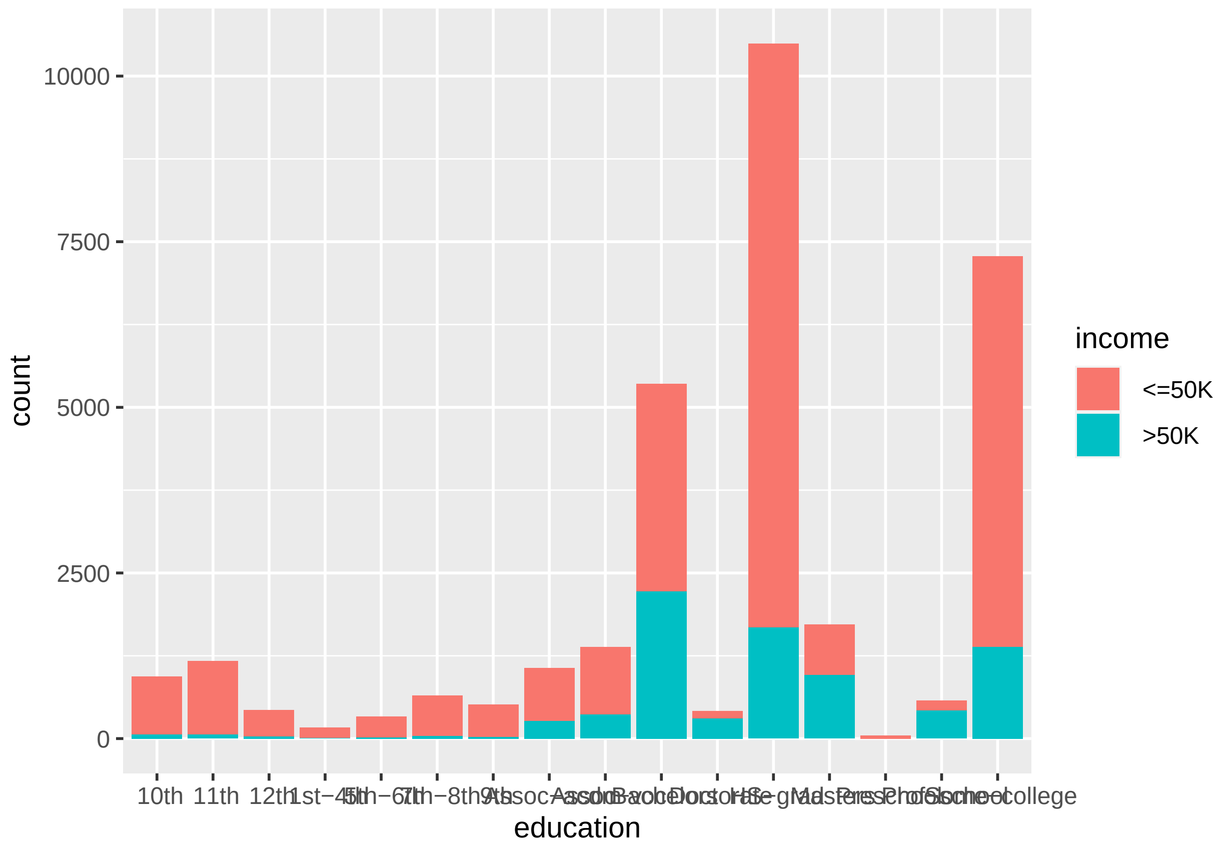


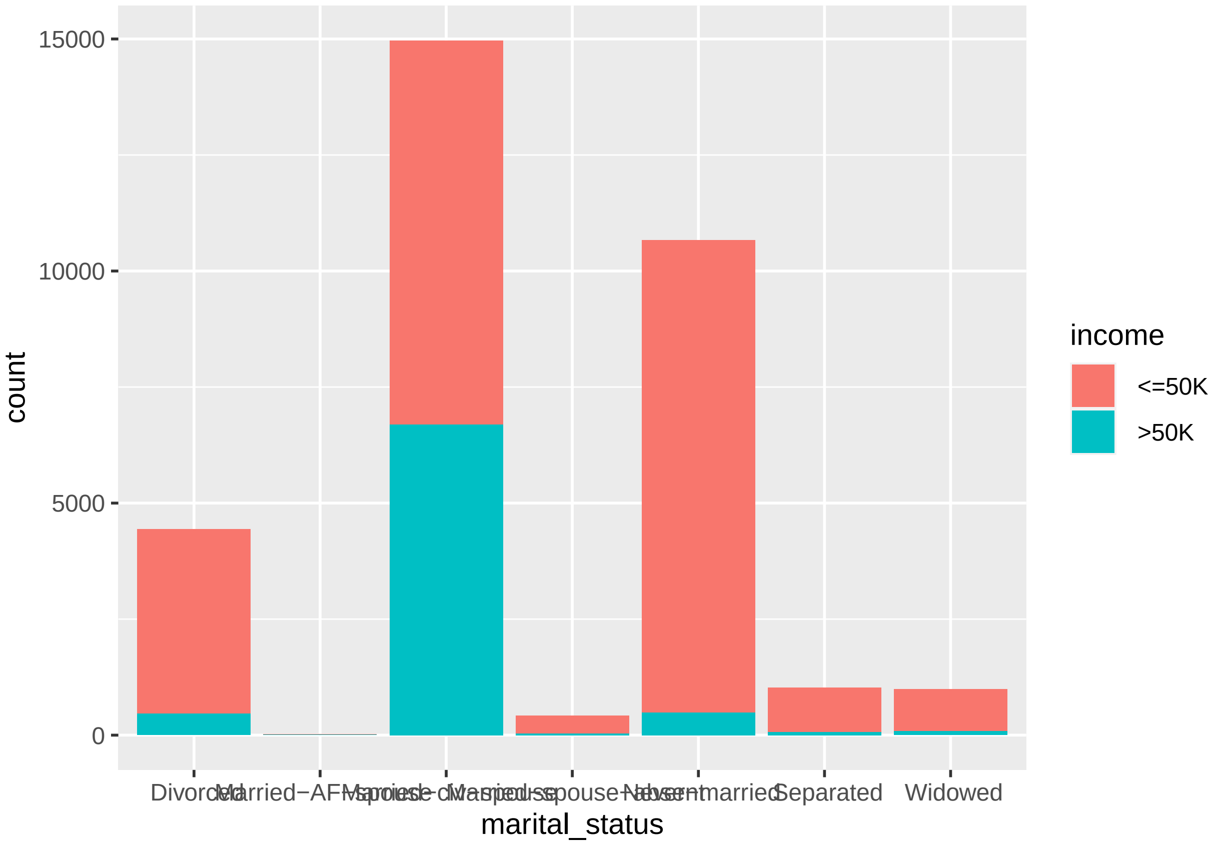
We can see there are still outliers even excluding zeros for capital gain and capital loss variables.

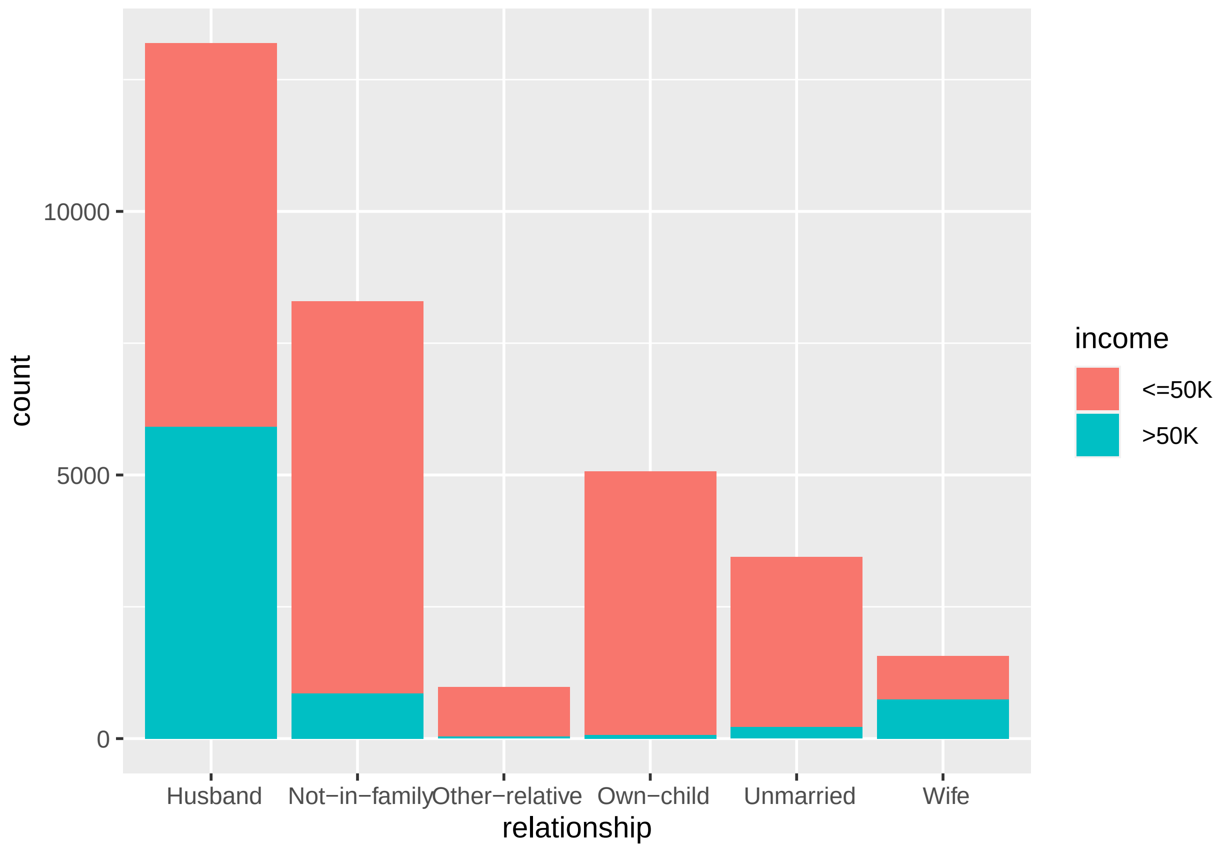
Distributions of categorical variables by target variable.

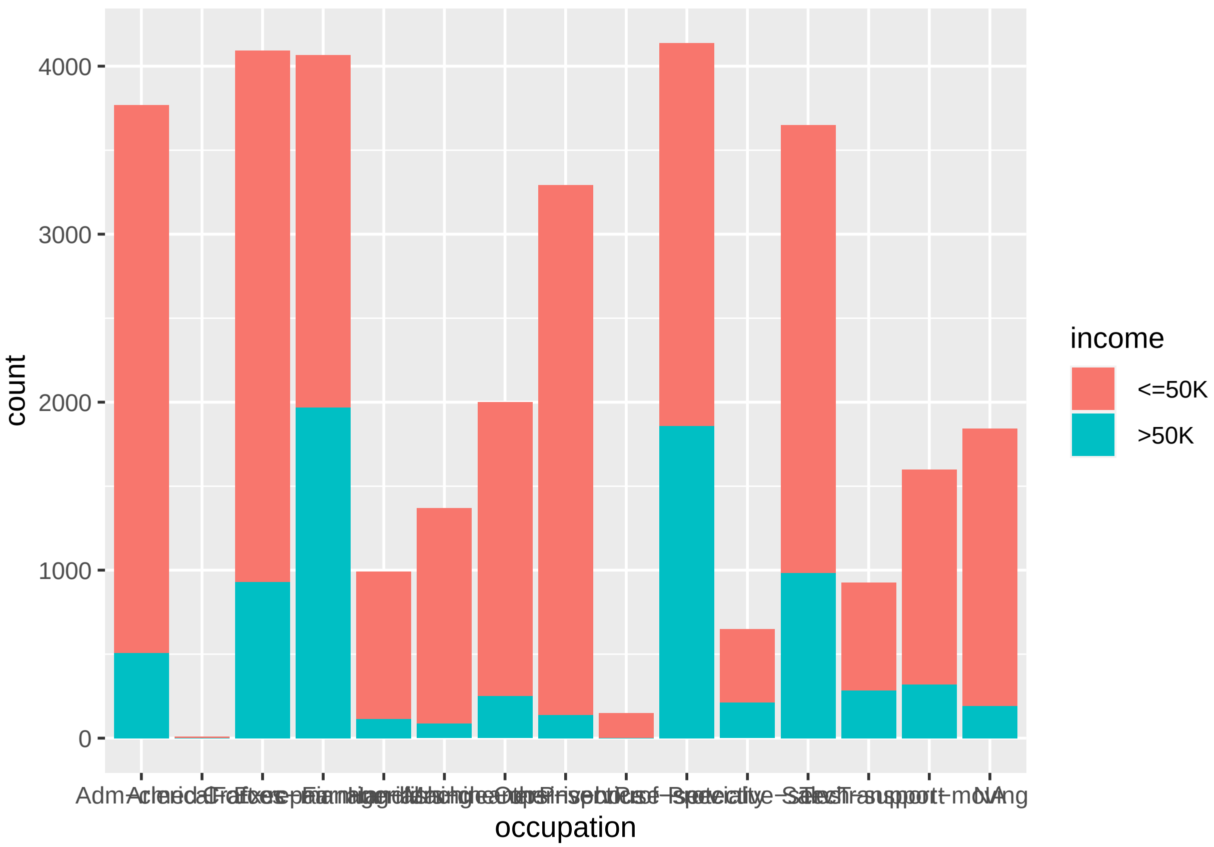


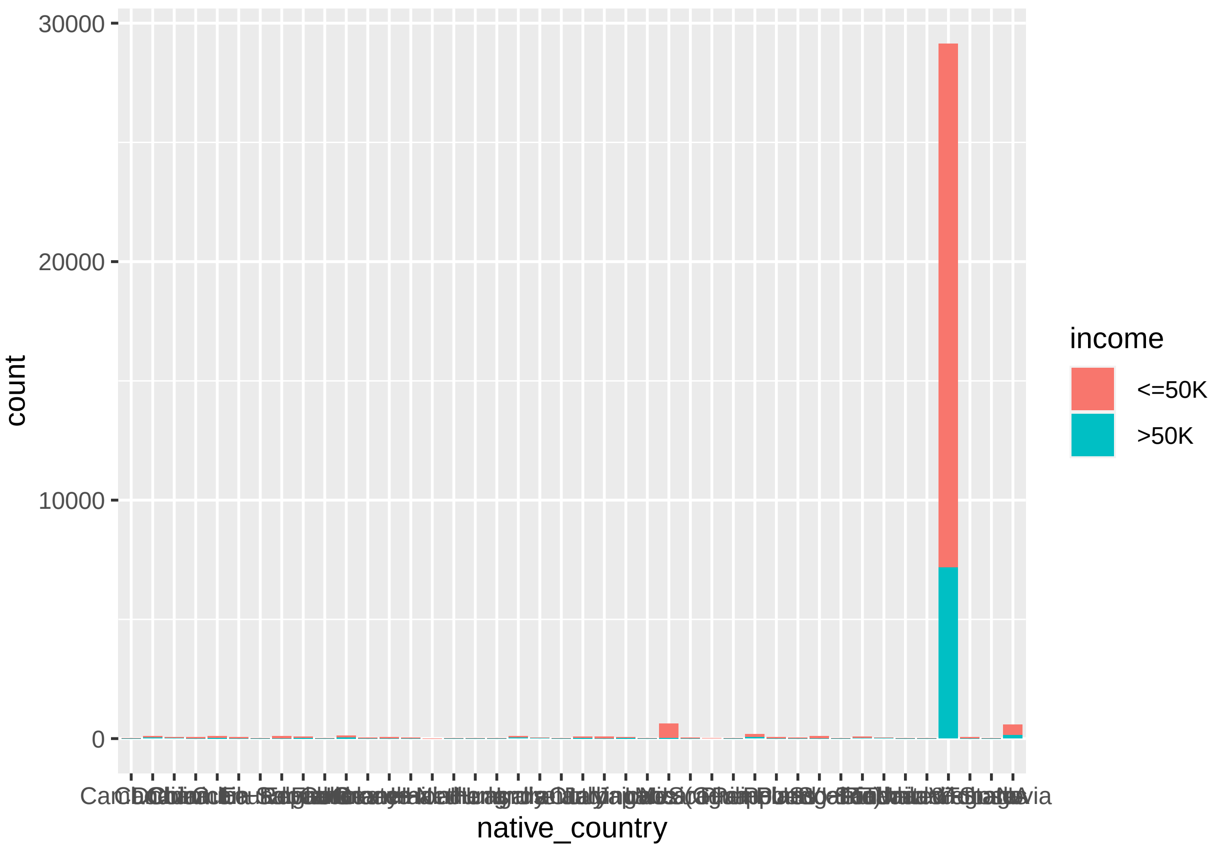
From the above bar chart we can see the majority of adults in the census were working in private sectors.

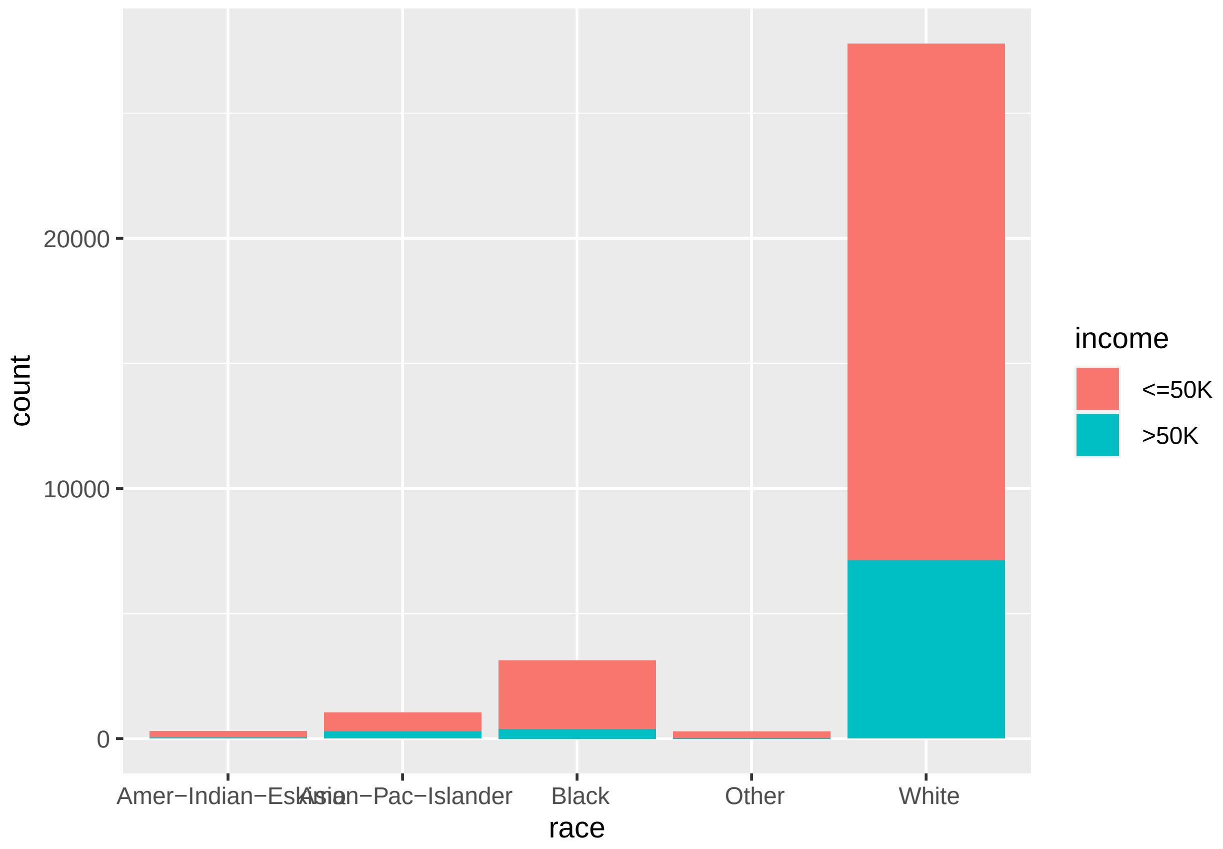

count
10000
7500
5000
2500
0
10th
11th
12th
1st−4th
5th−6th
7th−8th
9th
Assoc−acdm
Assoc−voc
Bachelors
Doctorate
HS−grad
Masters
Preschool
Prof−school
Some−college
education
income
<=50K
>50K

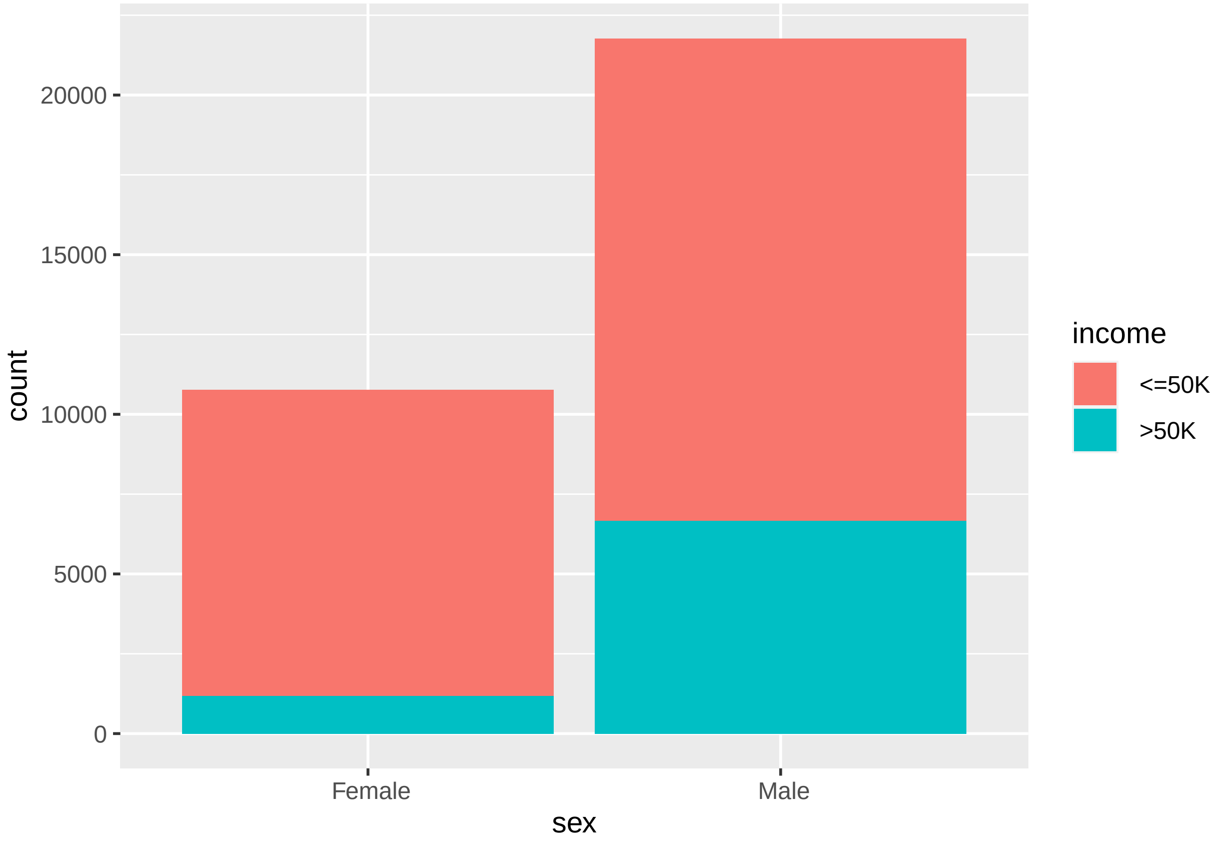

15000
10000
5000
0
count
Divorced
Married−AF−spouse
Married−civ−spouse
Married−spouse−absent
Never−married
Separated
Widowed
marital_status
income
<=50K
>50K


10000
5000
0
count
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife
relationship
income
<=50K
>50K


count
4000
3000
2000
1000
0
Adm-clerical
Armed-Forces
Craft-repair
Exec-managerial
Farming-fishing
Handlers-cleaners
Machine-op-inspct
Other-service
Priv-house-serv
Prof-specialty
Protective-serv
Sales
Tech-support
Transport-moving
NA
occupation
income
🟥 <=50K
🟦 >50K


30000
20000
10000
0
count
native_country
income
<=50K
>50K


count
20000
10000
0
Amer−Indian−Eskimo
Asian−Pac−Islander
Black
Other
White
race
income
<=50K
>50K


count
20000
15000
10000
5000
0
Female
Male
sex
income
<=50K
>50K