Income Prediction. Classification Predictive Modeling

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

12/02/2021

Business and Data Understanding

The data we are using comes from the US Census data collected in 1994. The dataset can be obtained at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/census+income). The donor of the dataset is Ronny Kohavi and Barry Becker, Data Mining and Visualization, Silicon Graphics. The current dataset was extracted by Barry Becker from the 1994 Census database.

Reformulate a problem statement as an analytics problem

The current business problem we are trying to solve is how to predict the income of a given customer into 2 classes: less than or equal to $50 thousand USD, or greater than $50 thousand USD. This is a business problem, because given some demographic information such as age, sex, education, marital status, occupation, we want to be able to predict the customer's income into the =<50K category or >50K category. If we can predict this income accurately, companies can use this information to determine whether they should allocate resources to market some premium grade products to the customer.

Develop a proposed set of drivers and relationships to inputs

The output function is the predictiong of income, and whether it belongs to the =< $50K class or to the >$50K class. The input variables are the age, sex, occupation, workclass, education level, education number, relationship, marital status, final weight(refering to the weight of that demographic class within the current population survey), the capital gain, capital loss, hours per week (of work) and the native country.
- How does age affect the income class of a customer? - How does education level affect the income class of a customer? - What types of occupation is associated with income greater than $50K or with income less than or equal to $50K?

State the set of assumptions related to the problem

One assumption related to this problem is that the relationships between the input variables (such as age, occupation, workclass, marital status) to the target variable income obtained through the 1994 census data will hold true to what is observed today in 2021.

Define key metrics of success

One key metric of success is that the prediction model can accurately predict the income class, given the input information.

Describe how you have applied the ethical ML framework

Identify and prioritize means of data acquisition

The means of data acquisition is through downloading the US census adult data set.

Data Preparation

Describe the purpose of the data set you selected (i.e., why was this data collected in the first place?).

Describe how you would define and measure the outcomes from the dataset.

##How would you measure the effectiveness of a good prediction algorithm or clustering algorithm?

Define and prepare your target variables. Use proper variable representations (int, float, one-hot, etc.).

The target variable is income.

Use pre-processing methods (as needed) for dimensionality reduction, scaling, etc. Remove variables that are not needed/useful for the analysis

Describe the final dataset that is used for classification (include a description of any newly formed variables you created).

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.)

Column Name	Data Type	Column Description
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.)
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:

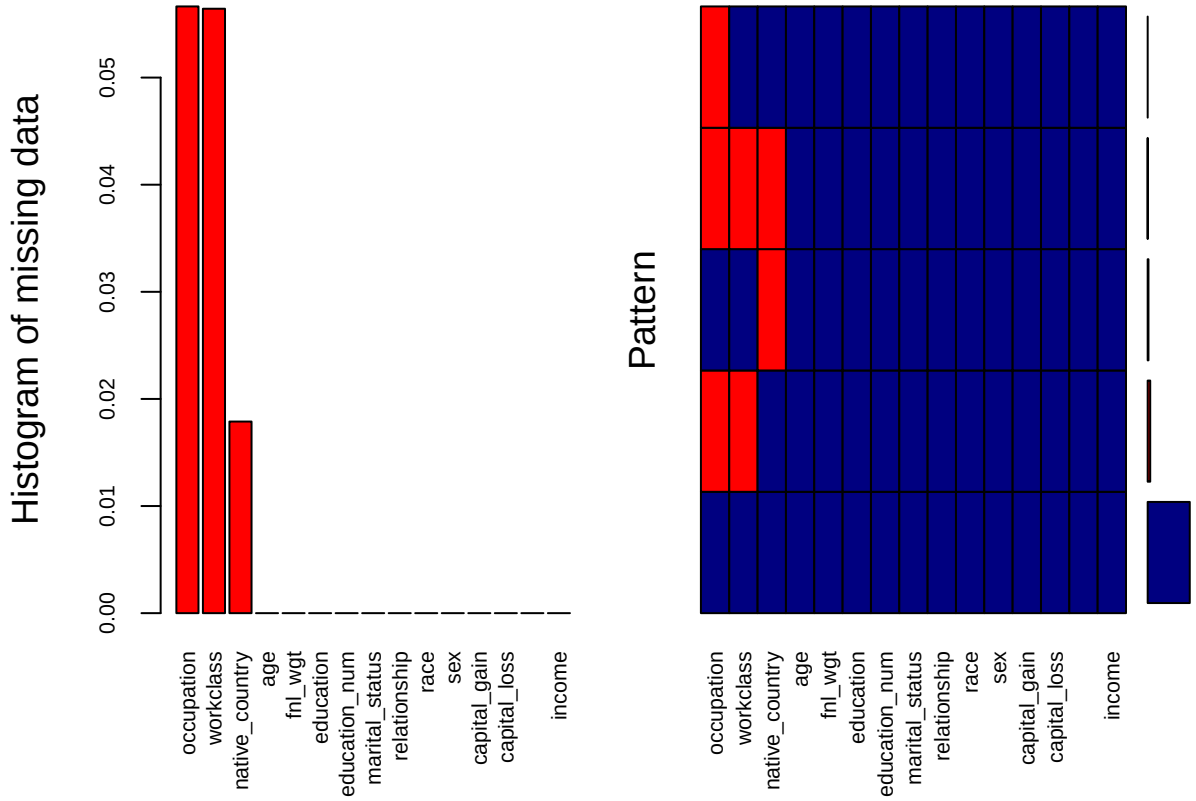
	age	workclass	fnl_wgt	education	education_num	marital_status	occupation	relationship
4768	21	Private	250051	Some-college	10	Never-married	Prof-specialty	Own-child
9172	21	Private	250051	Some-college	10	Never-married	Prof-specialty	Own-child
4326	25	Private	308144	Bachelors	13	Never-married	Craft-repair	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
```

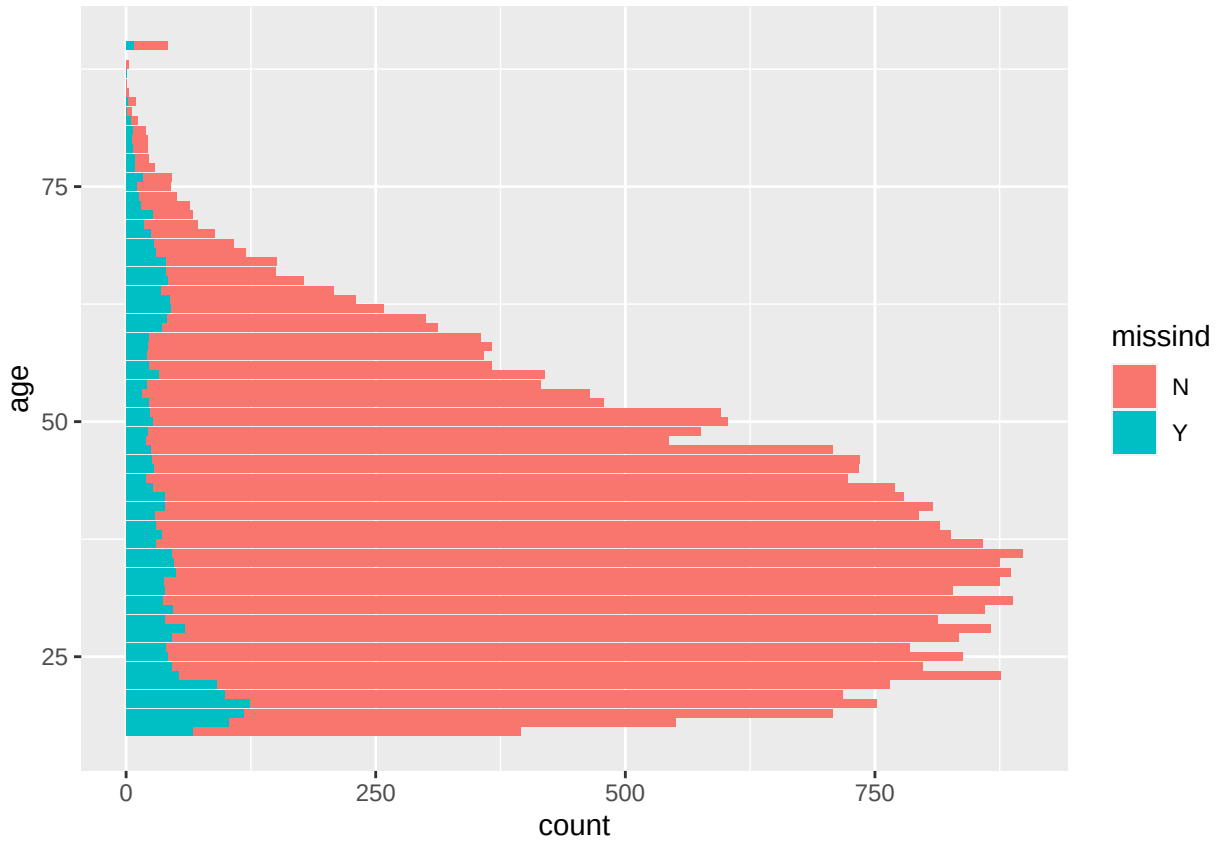


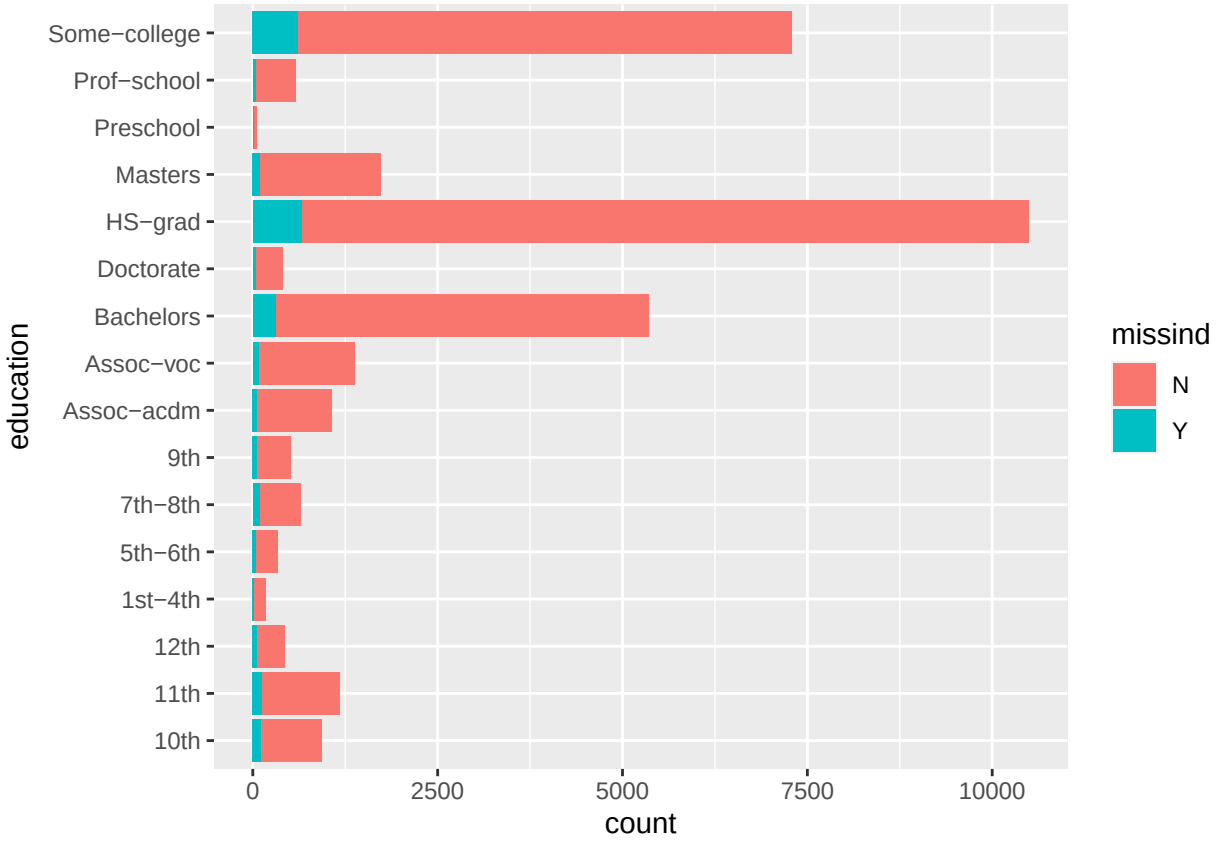

```
##
##  Variables sorted by number of missings:
##        Variable      Count
##      occupation 0.05664321
##       workclass 0.05642807
##  native_country 0.01788733
##             age 0.00000000
##         fnl_wgt 0.00000000
##       education 0.00000000
##   education_num 0.00000000
##  marital_status 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
```

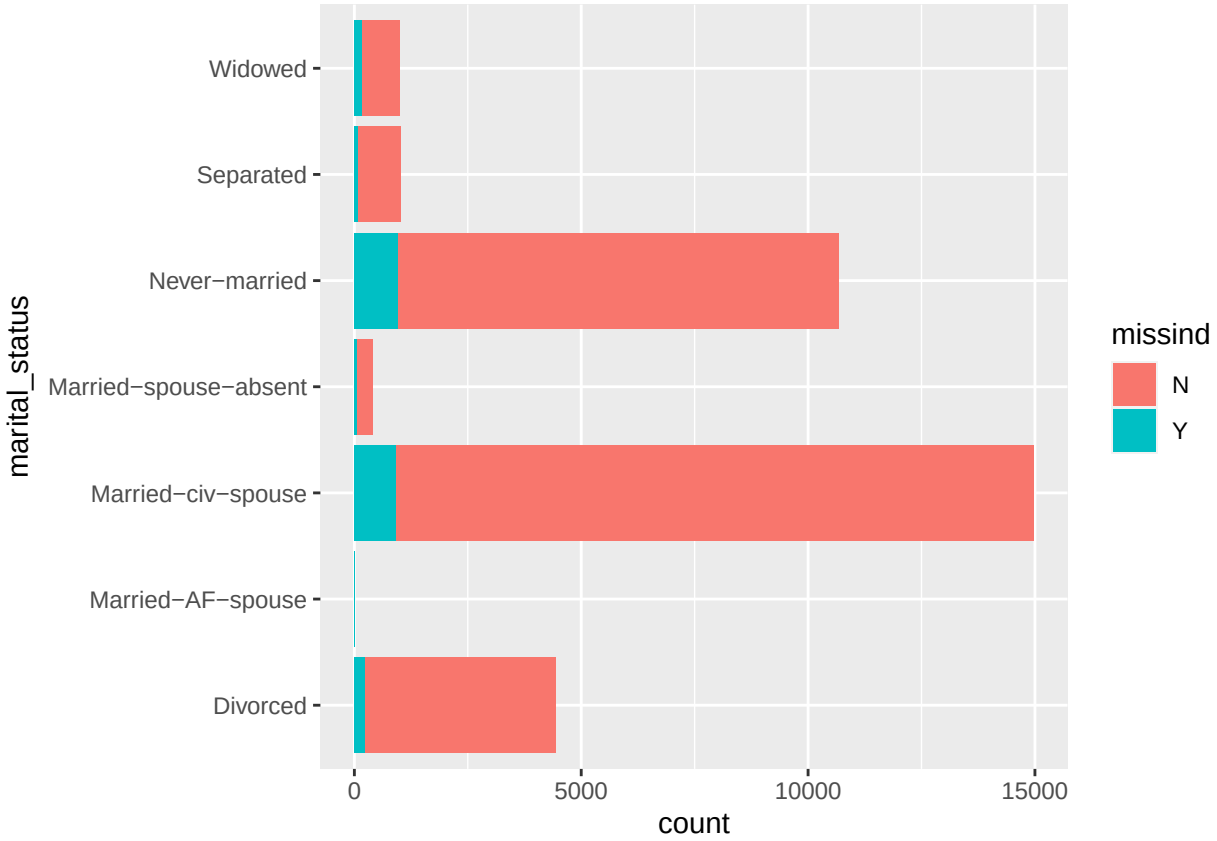
From the above, there are missing values in this data set and all the missing values are from categorical variables.

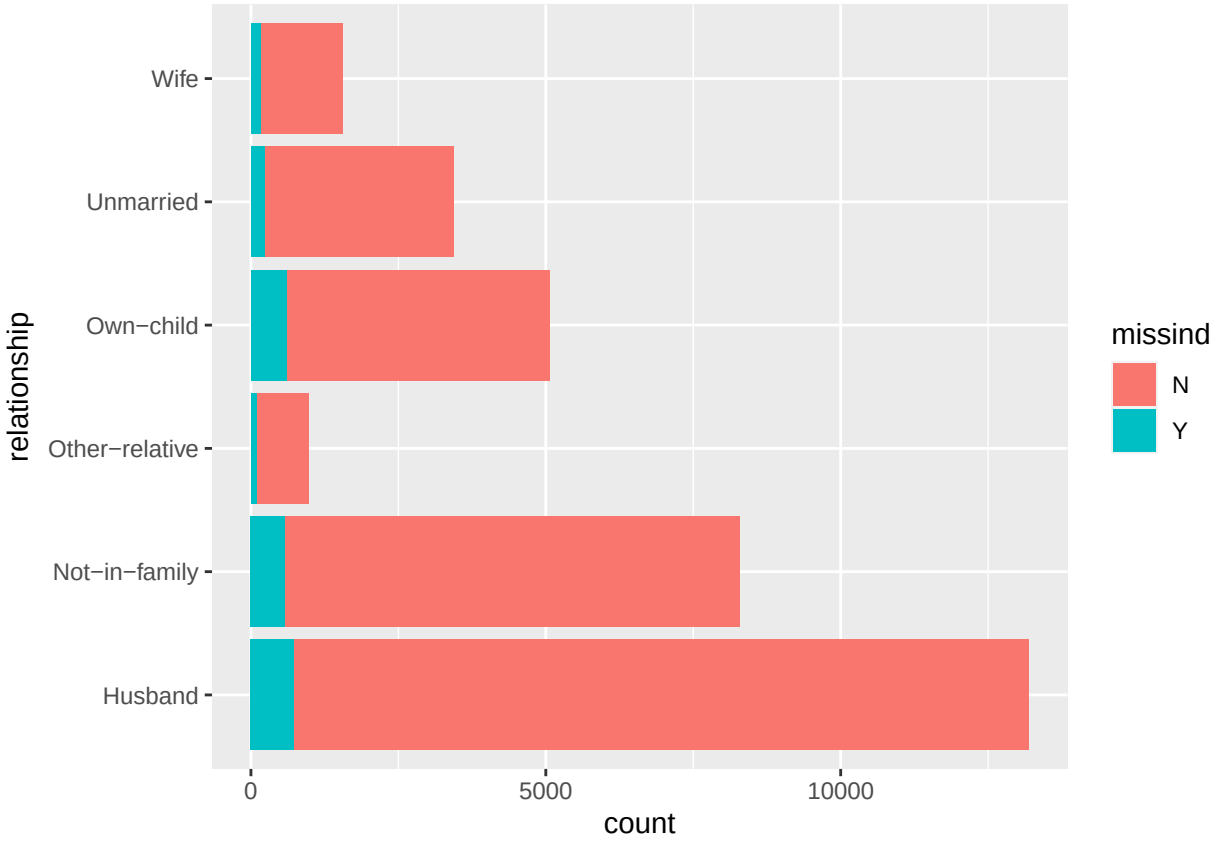
Comparing records with at least one missing value to those without any missing values.

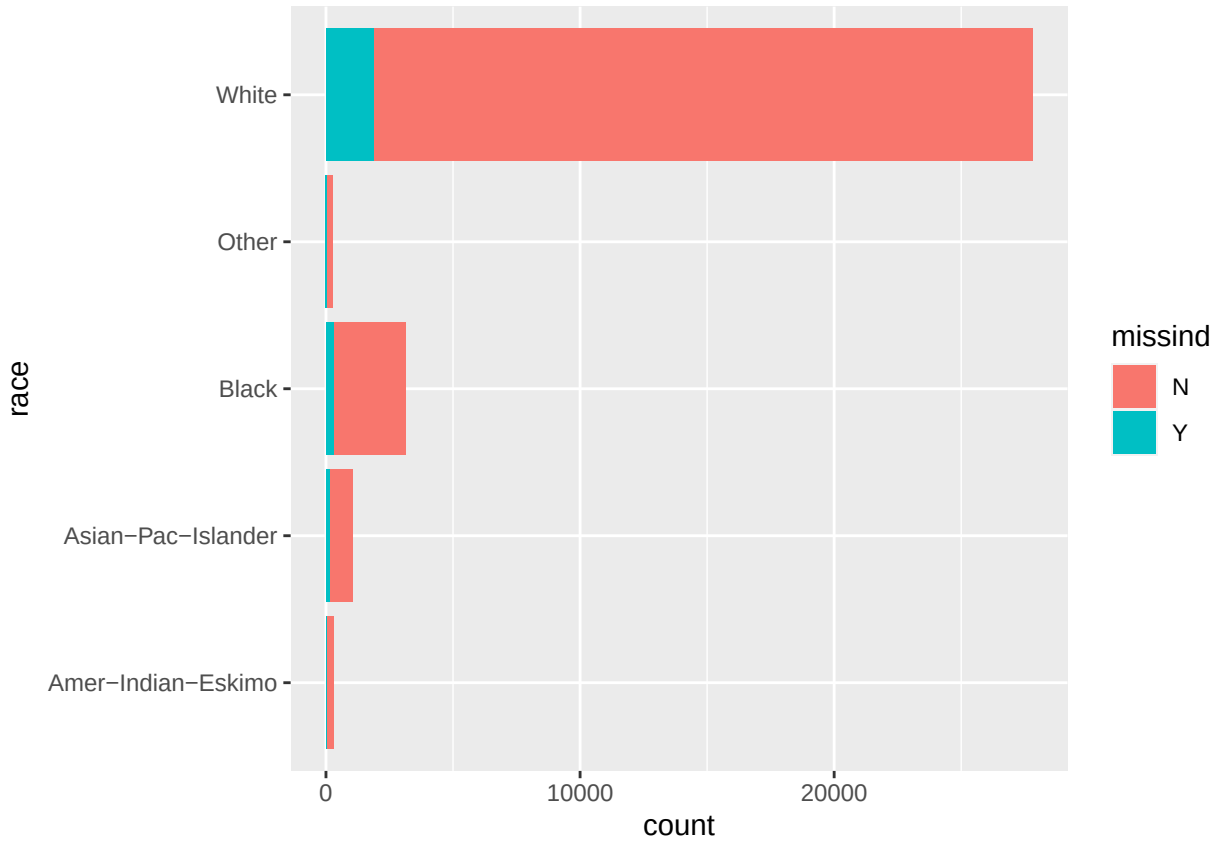
In order to better understand the patterns of the missing values, let's look at some descriptions of the records with missing values.

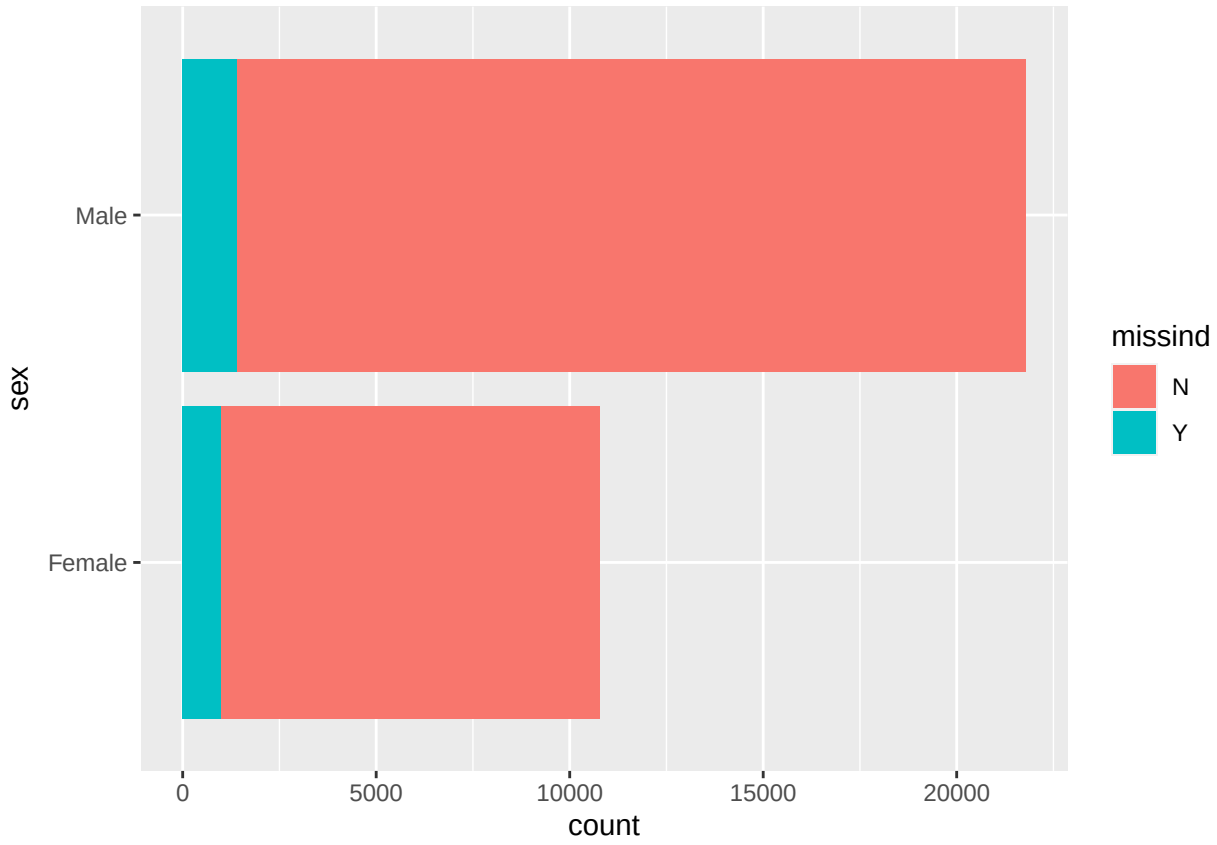

age
75
50
25
0
250
500
750
count
missind
N
Y

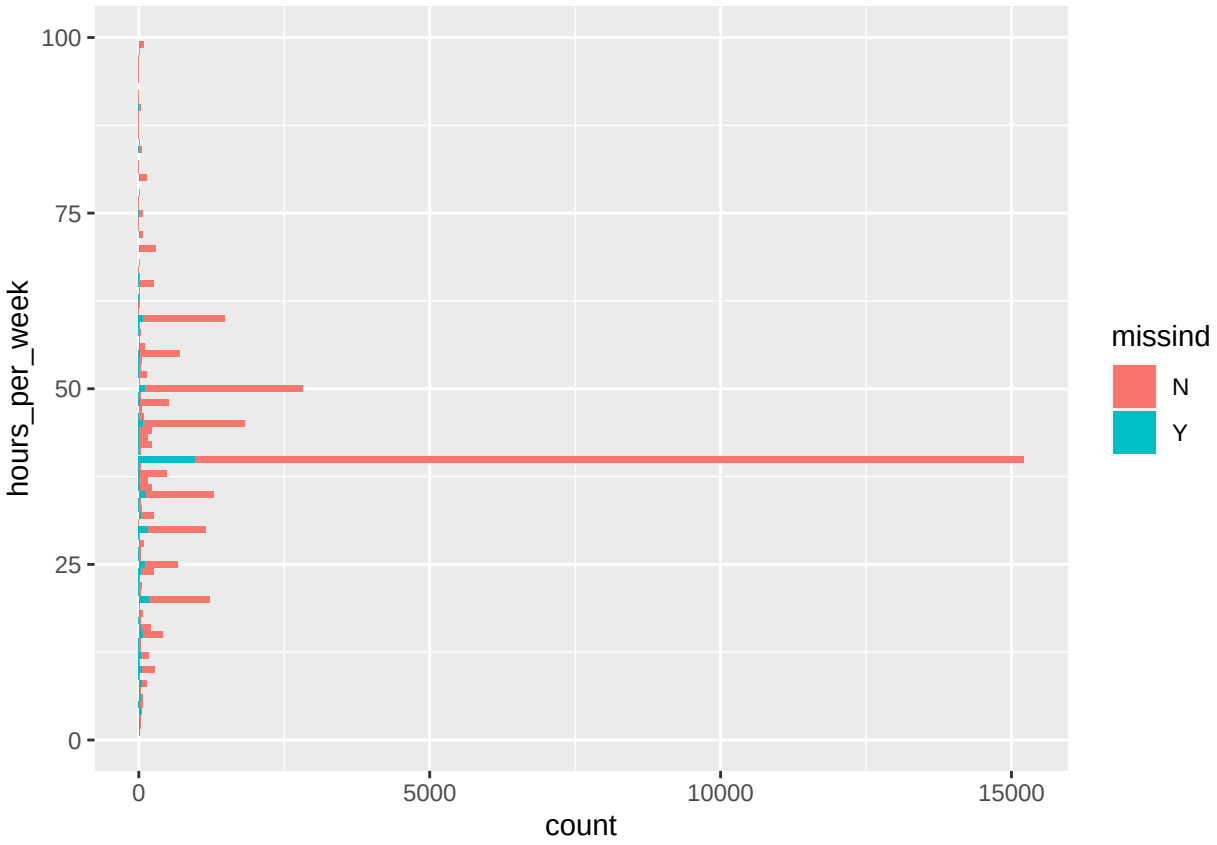

education
Some−college
Prof−school
Preschool
Masters
HS−grad
Doctorate
Bachelors
Assoc−voc
Assoc−acdm
9th
7th−8th
5th−6th
1st−4th
12th
11th
10th
0
2500
5000
7500
10000
count
missind
N
Y

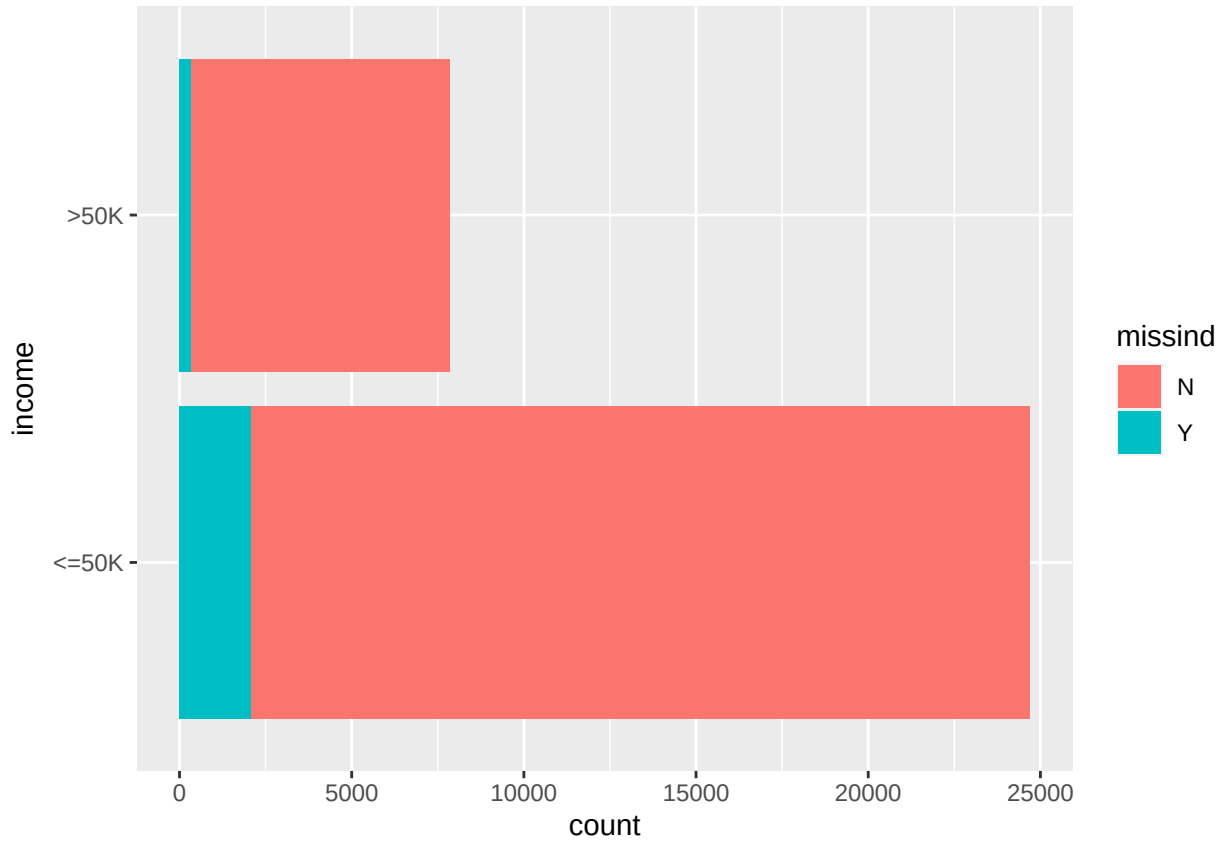

marital_status
Widowed
Separated
Never−married
Married−spouse−absent
Married−civ−spouse
Married−AF−spouse
Divorced
0
5000
10000
15000
count
missind
N
Y


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


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


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


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



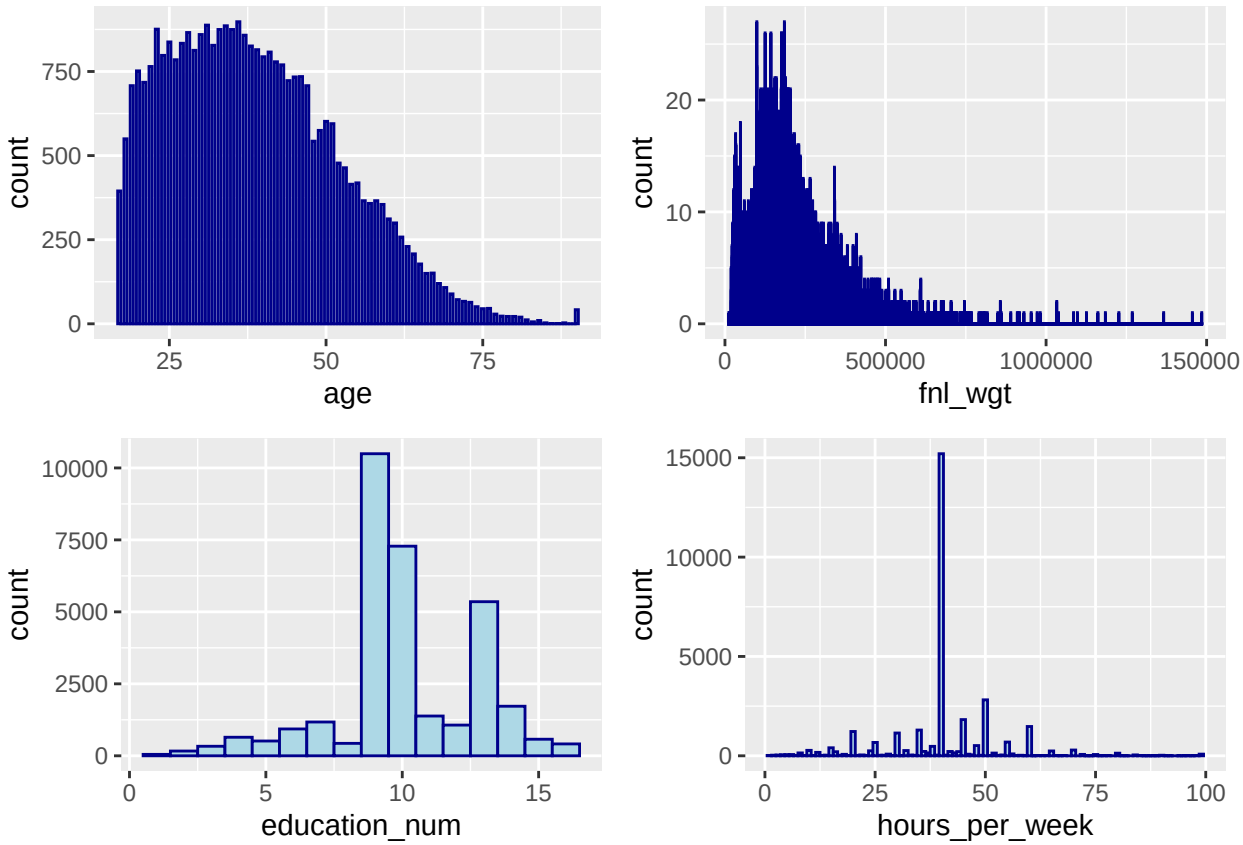
From the above bar charts comparing the distributions of 7 variables of the group that do not have missing values and the group that have at least one missing records, we can see that the missing records are generally evenly distributed across all ages, education level, marital status, family relationship, 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:

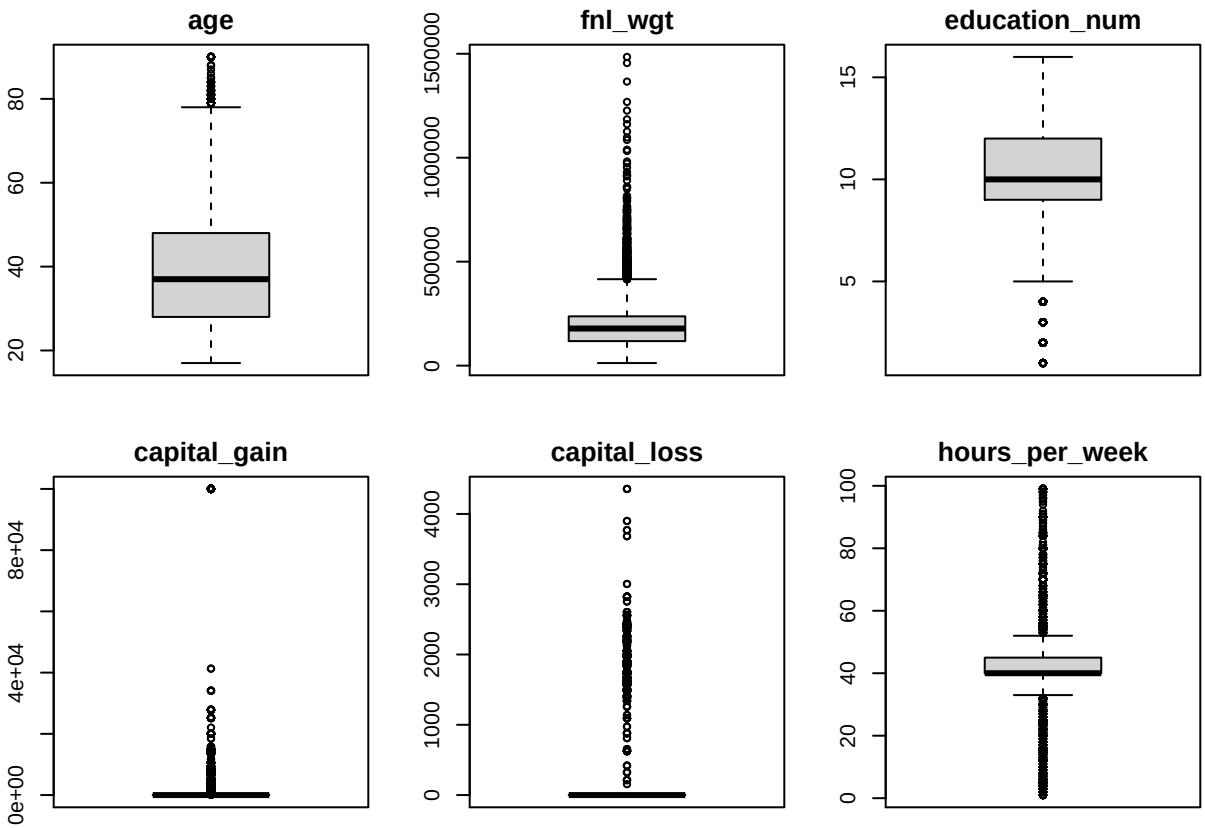
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
age	17.00	28.00	37.00	38.59	48.00	90.00
fnl_wgt	12285.00	117827.00	178356.00	189780.85	236993.00	1484705.00
education_num	1.00	9.00	10.00	10.08	12.00	16.00
capital_gain	0.00	0.00	0.00	1078.44	0.00	99999.00
capital_loss	0.00	0.00	0.00	87.37	0.00	4356.00
hours_per_week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



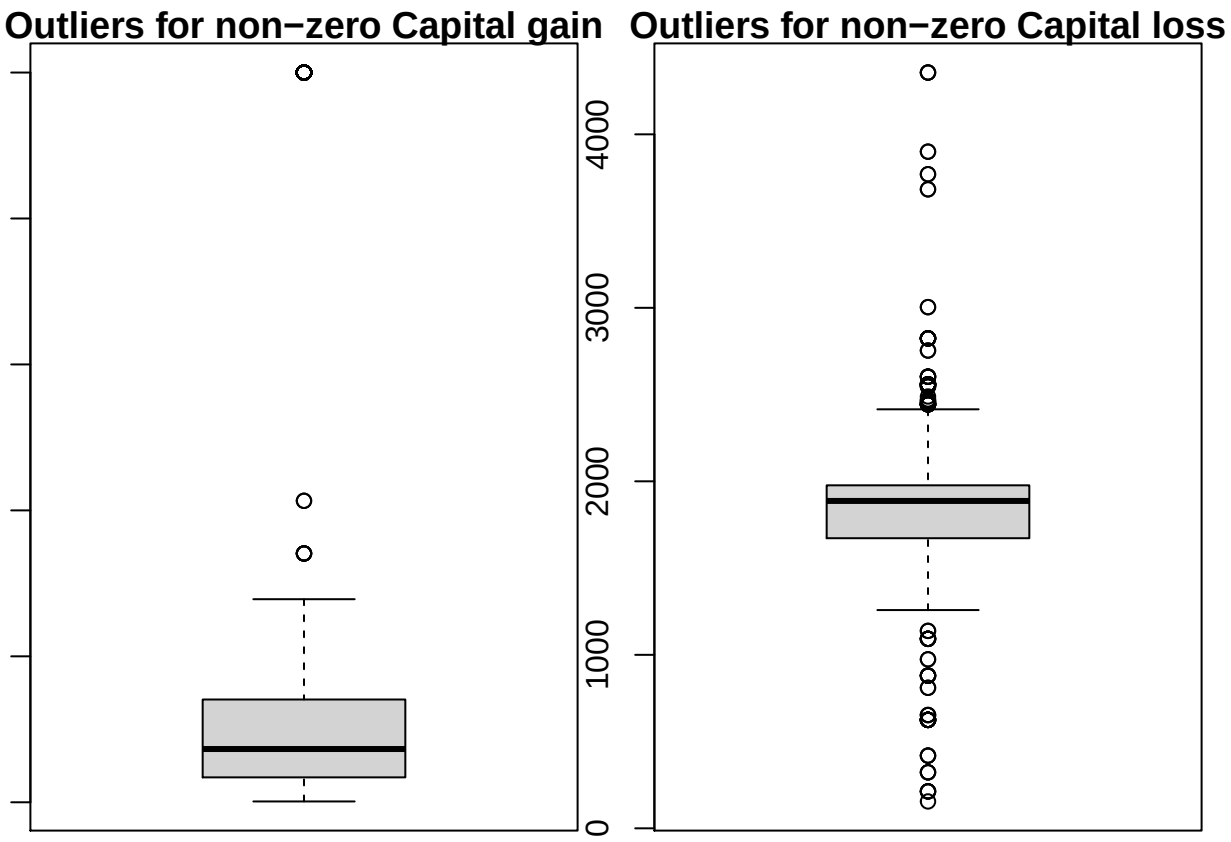
Both age and fnl_wght variables are skewed to the right, where log-normal transformation could help if modeling techniques with normality assumptions were to be used. The working hours per week variable is heavy-tailed distributed, meaning there are still quite a number of people working extremely small or large number of hours each week.

Boxplots to discover outliers for each numeric variable.



There are outliers for all numeric variables. There are large amount of records with ages and fnl_wght larger than their upper quartiles, which are consistent with the histograms showing their distributions are skewed to the right.

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

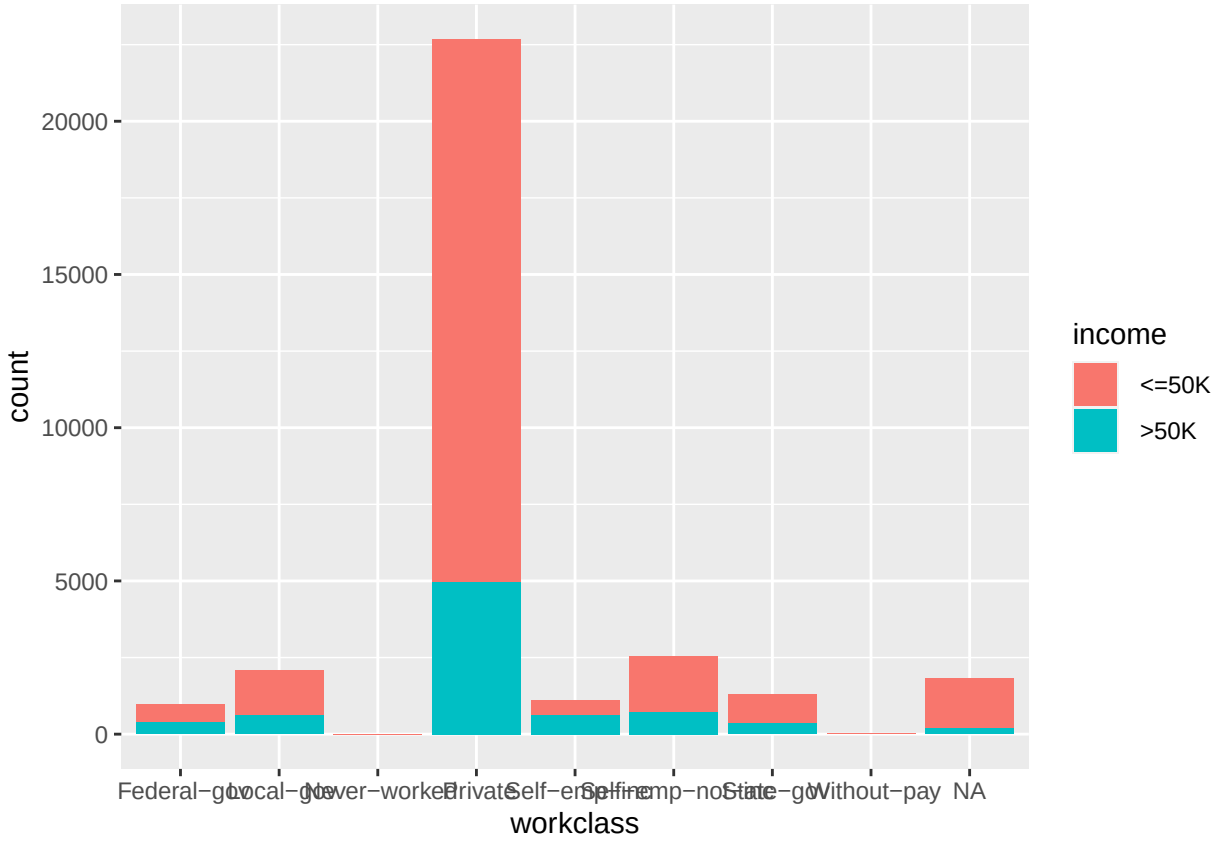


We can see there are still outliers even excluding zeros for capital gain and capital loss variables.

Distributions of categorical variables by target variable.

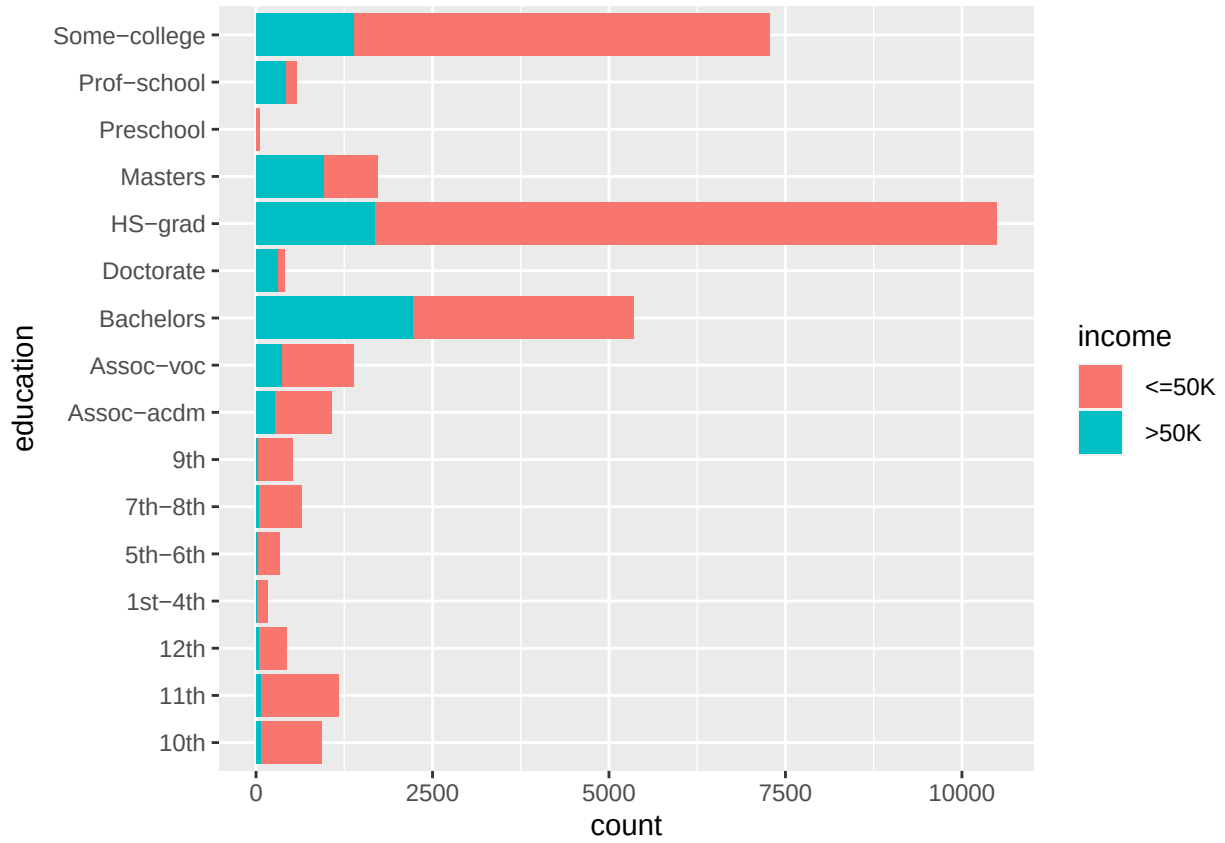
```
par(mar=c(1,1,1,1))

ggplot(X, aes(workclass, fill = income)) + geom_bar()
```



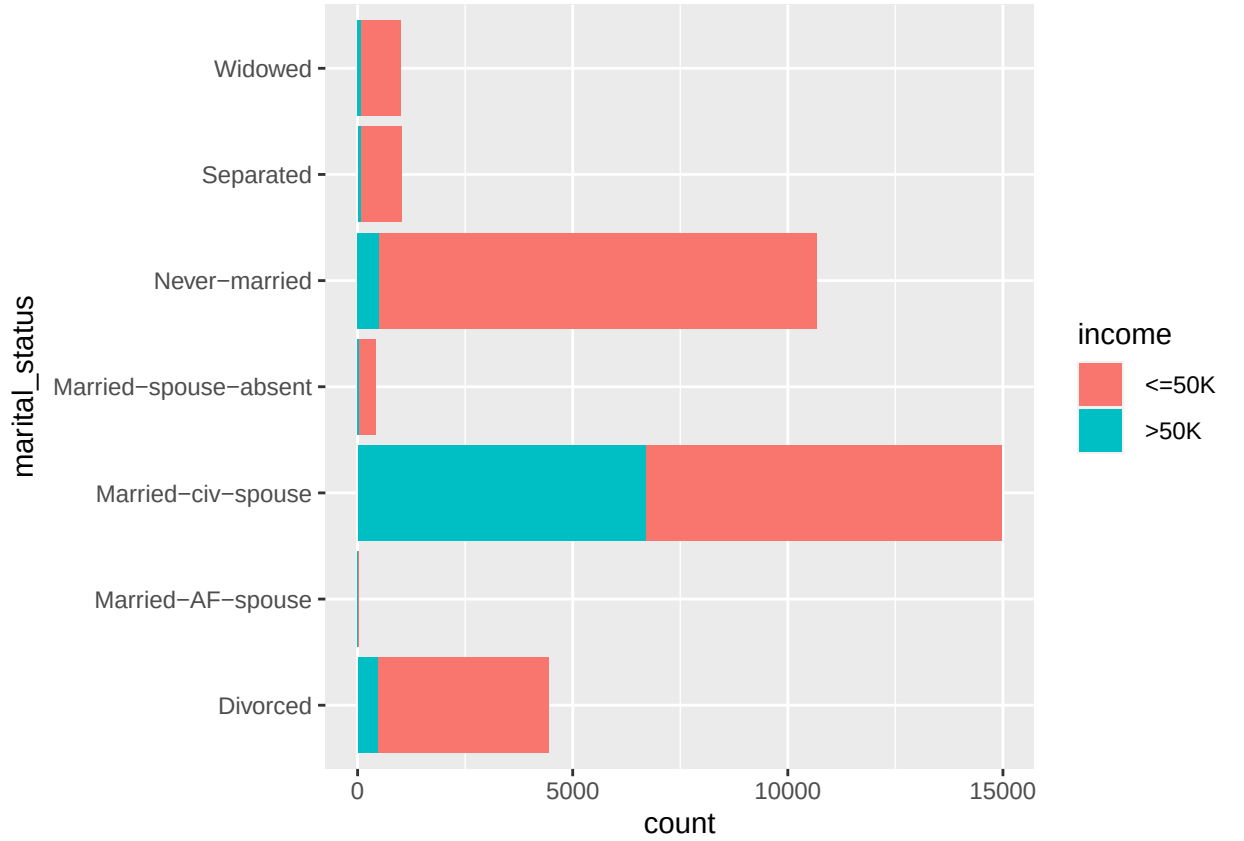
From the above bar chart we can see the majority of adults in the census were working in private sectors.

```
#plotting education vs income
ggplot(data = X, aes(y = education, fill = income)) +
  geom_bar(position = "stack")   #different bars stacked together
```



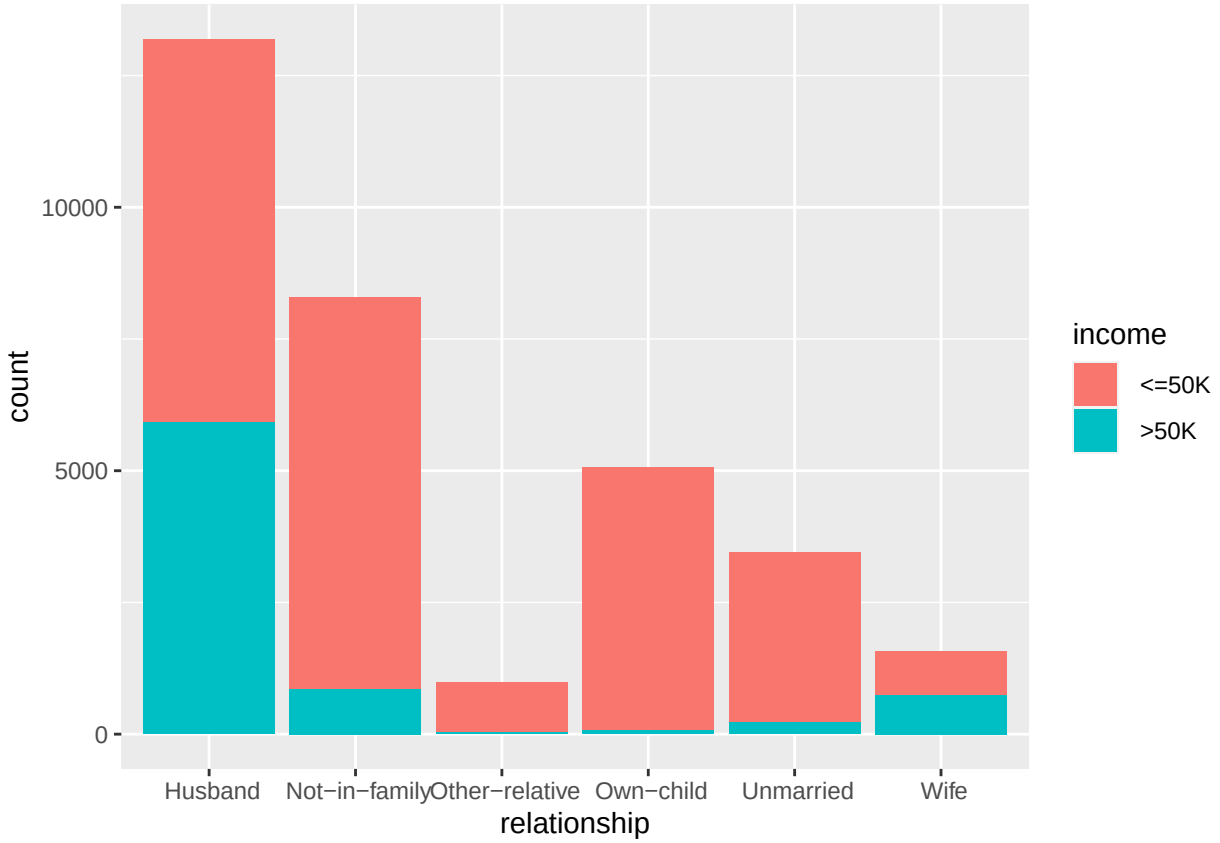
The majority of people earning less than $50K are high school graduates. The next largest education group is some college, and the third largest education group is Bachelors.

```
#plotting marital status vs. income
ggplot(data = X, aes(y = marital_status, fill = income))+
  geom_bar(position = "stack")
```



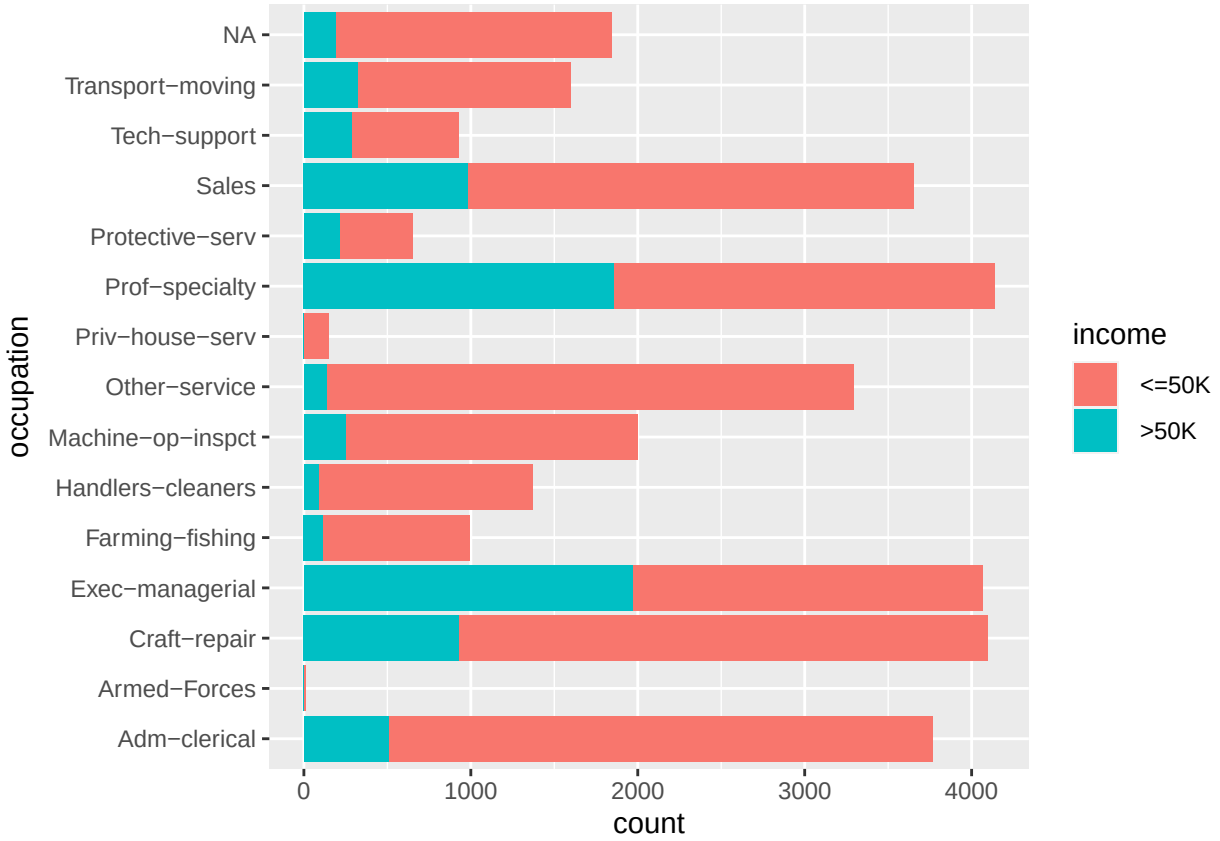
The majority of people surveyed are Married civ spouse, and in this marital status category, the income is roughly equally divided between <=50K or >50K. The second largest category is Never-married, with the majority of people earning <=50K.

```
ggplot(X, aes(relationship, fill = income)) + geom_bar()
```



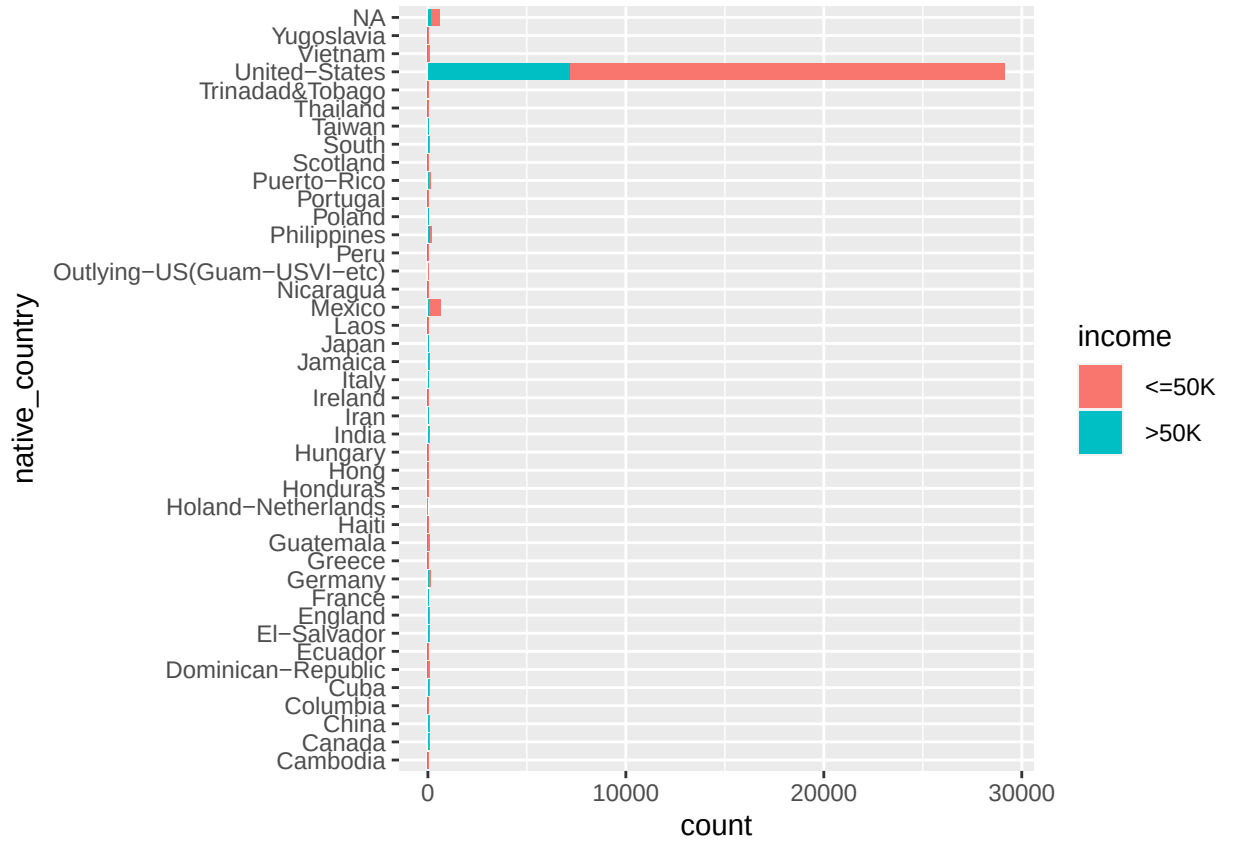
Most people surveyed in the census belong to the Husband category of relationships, with slightly more people earning less than or equal to 50K. However, in the Husband category, there is almost an even split between the 2 target income classes. Not-in-family is the second largest category for relationships and the majority people in this category have income <=50K.

```
#plotting occupation vs income
ggplot(data = X, aes(y = occupation, fill = income))+
  geom_bar(position = "stack")
```



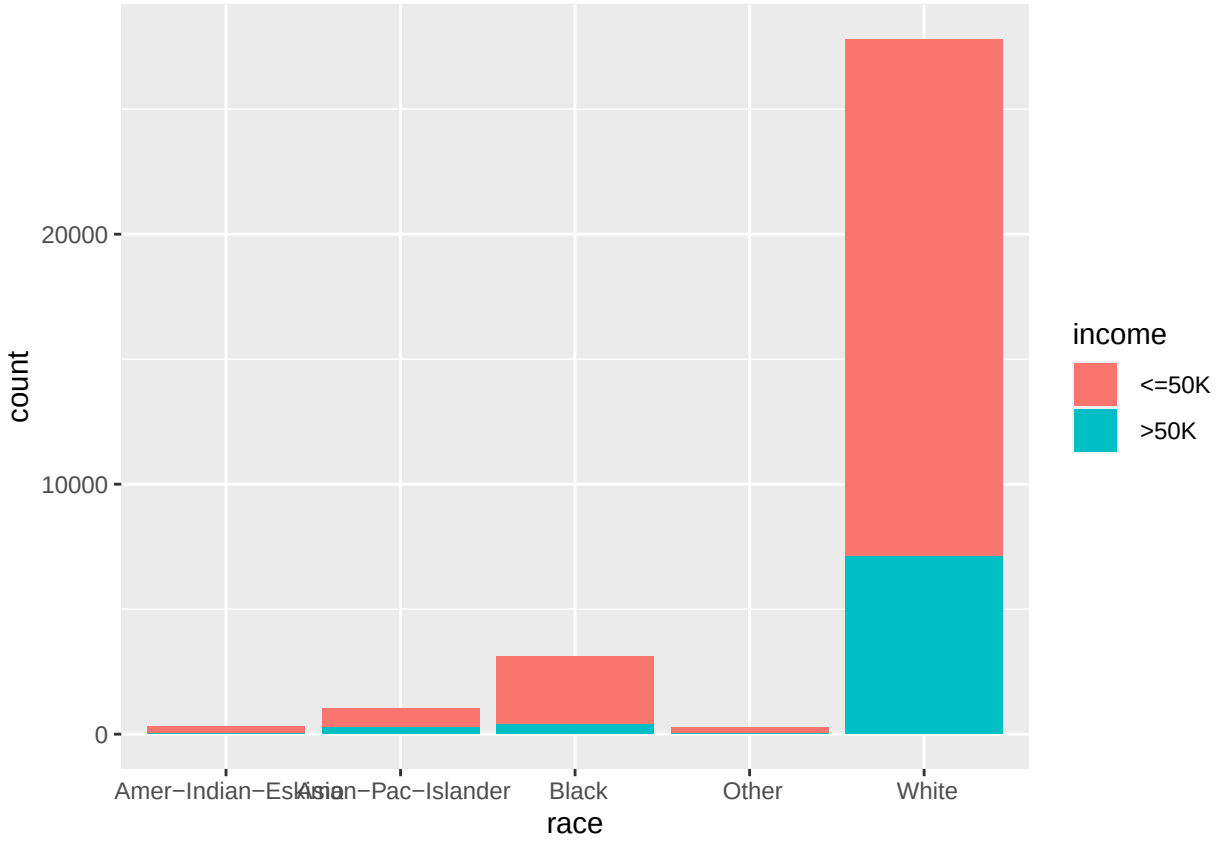
Most common occupations are Prof-specialty, Exec-managerial, Craft-repair, Sales, and Adm-clerical. For Exec-managerial, and Prof-specialty, there is an even number of people earning <=50K and >50K. For Craft-repair, Adm-clerical, and Sales, the majority of people earn <=50K.

```
#plotting native country vs. income
ggplot(data = X, aes(y = native_country, fill = income))+
  geom_bar(position = "stack")
```



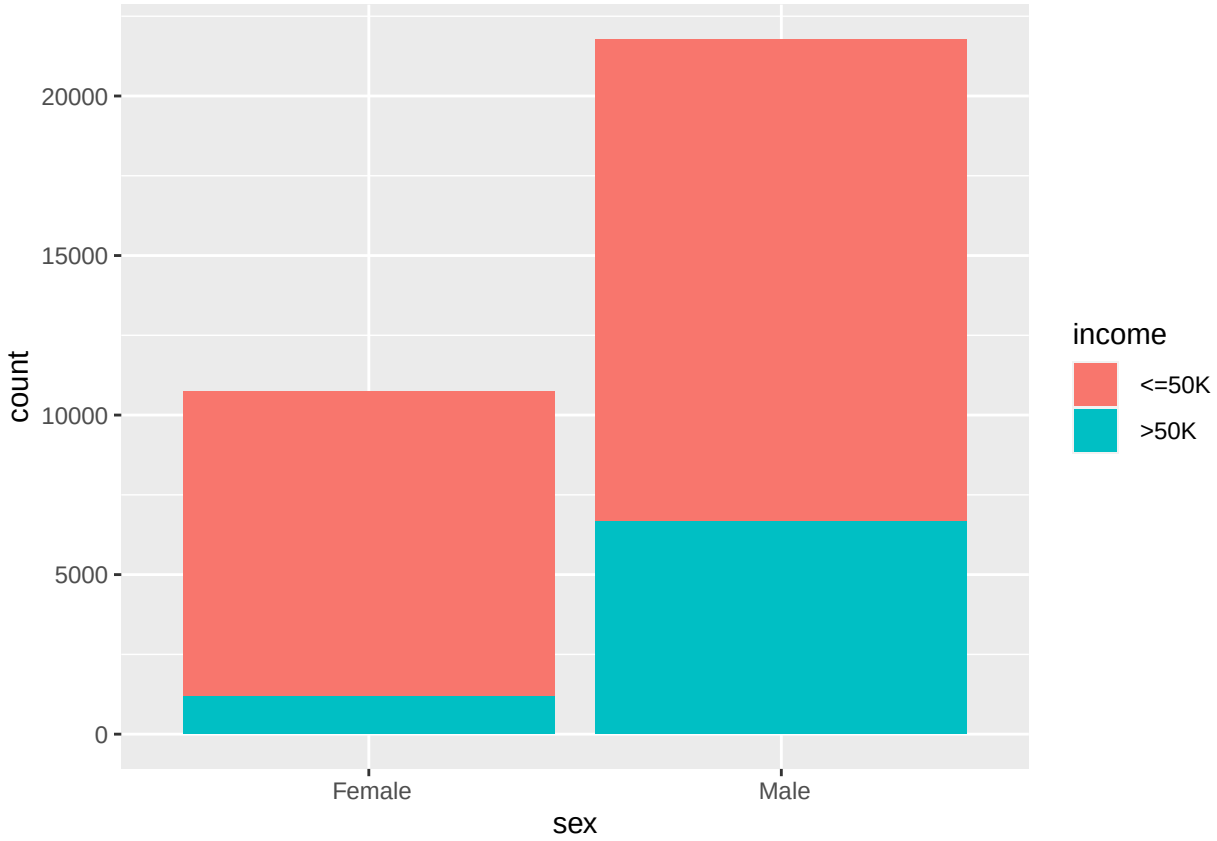
Most people surveyed come from the United States. This makes sense as the census was conducted in the US. Other than the United States, the second highest number of people come from Mexico.

```
ggplot(X, aes(race, fill = income)) + geom_bar()
```



Most people surveyed are White, and earn <=50K. The second highest race category is Black.

```
ggplot(X, aes(sex, fill = income)) + geom_bar()
```



There are more than twice as many males surveyed in this census compared to females.