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

Our client is looking to open a business in a new location. The client is looking to open a store that sells products of one of their luxury brands. The luxury brand is trying to target people with income of above 50K. 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, the company can use this information to determine whether they should allocate resources to market some premium grade products to the customer. The marketing team can use this tool to find the audience for our marketing pitch in anticipation of the branch opening and improve targeted advertising to people who have income above 50K. The tool allows a true/false output against each demographic item.

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.

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.

Modeling and Evaluation

Describe the data

Data Dictionary

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
## Warning: package 'VIM' was built under R version 4.0.3
```

```
## Warning: package 'colorspace' was built under R version 4.0.3
```

```
## 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.)

Column Name	Data Type	Column Description
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.)
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

looking at some statistics for the dataset

```
summary(X)
```

```
##       age                    workclass        fnl_wgt
##  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                  marital_status
##  HS-grad     :10501   Min.   : 1.00   Divorced             : 4443
##  Some-college: 7291   1st Qu.: 9.00   Married-AF-spouse    :   23
##  Bachelors   : 5355   Median :10.00   Married-civ-spouse   :14976
##  Masters     : 1723   Mean   :10.08   Married-spouse-absent:  418
##  Assoc-voc   : 1382   3rd Qu.:12.00   Never-married        :10683
```

```
##  11th        : 1175   Max.   :16.00   Separated            : 1025
##  (Other)     : 5134                   Widowed              :  993
##            occupation           relationship                    race
##  Prof-specialty :4140   Husband       :13193   Amer-Indian-Eskimo:  311
##  Craft-repair   :4099   Not-in-family : 8305   Asian-Pac-Islander: 1039
##  Exec-managerial:4066   Other-relative:  981   Black             : 3124
##  Adm-clerical   :3770   Own-child     : 5068   Other             :  271
##  Sales          :3650   Unmarried     : 3446   White             :27816
##  Other-service  :3295   Wife          : 1568
##  (Other)        :9541
##      sex         capital_gain    capital_loss    hours_per_week
##  Female:10771   Min.   :    0   Min.   :   0.0   Min.   : 1.00
##  Male  :21790   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
```

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

```
## Warning: package 'tibble' was built under R version 4.0.3
```

```
## Warning: package 'tidyr' was built under R version 4.0.3
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.3
```

```
## Warning: package 'forcats' was built under R version 4.0.3
```

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

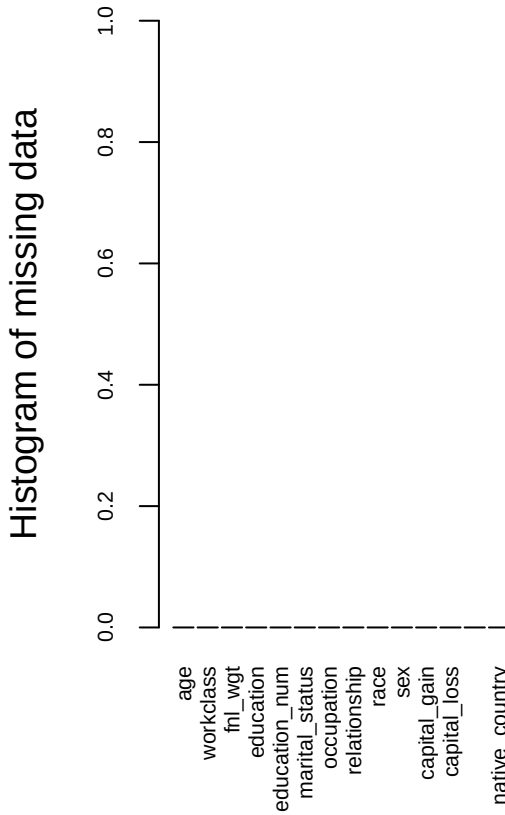
```
## Let's look at several examples of the duplicated records:
```

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.



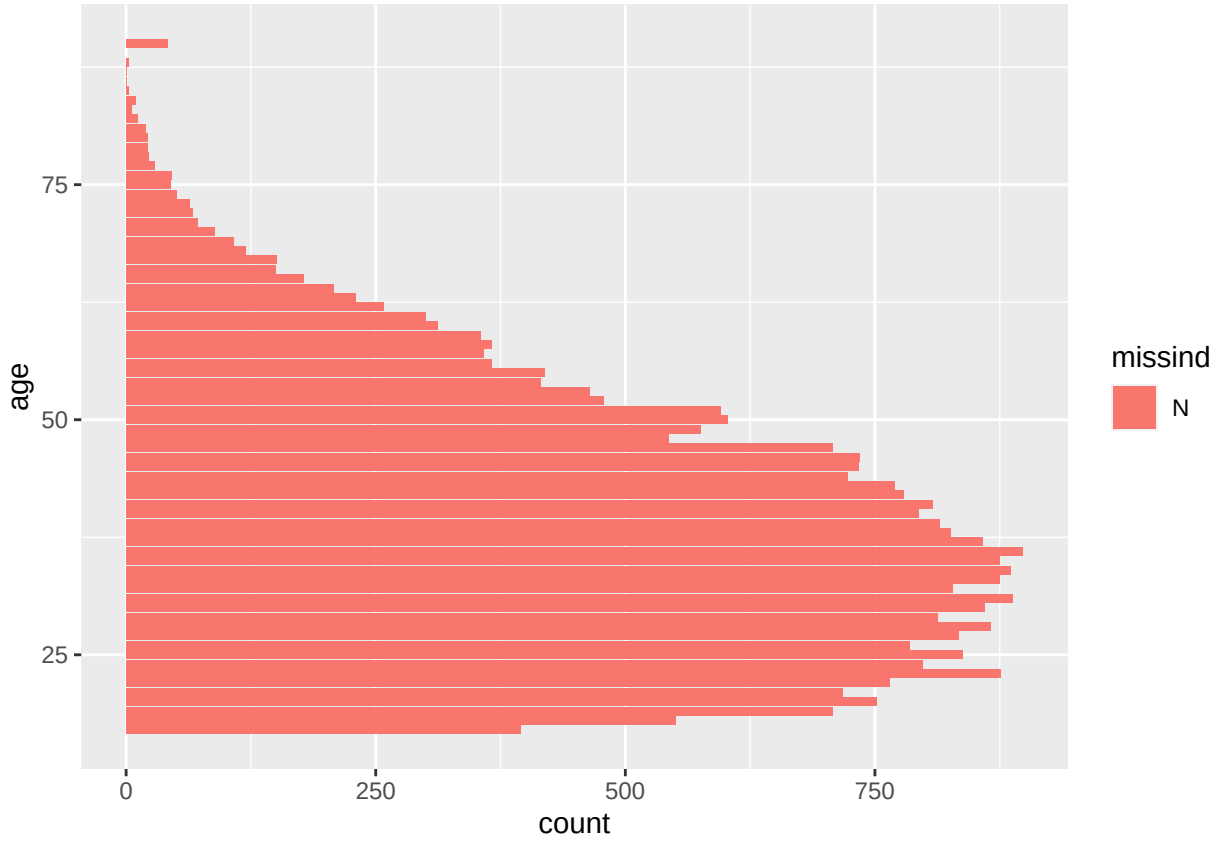
Then let's check whether there are any missing values in the dataset.

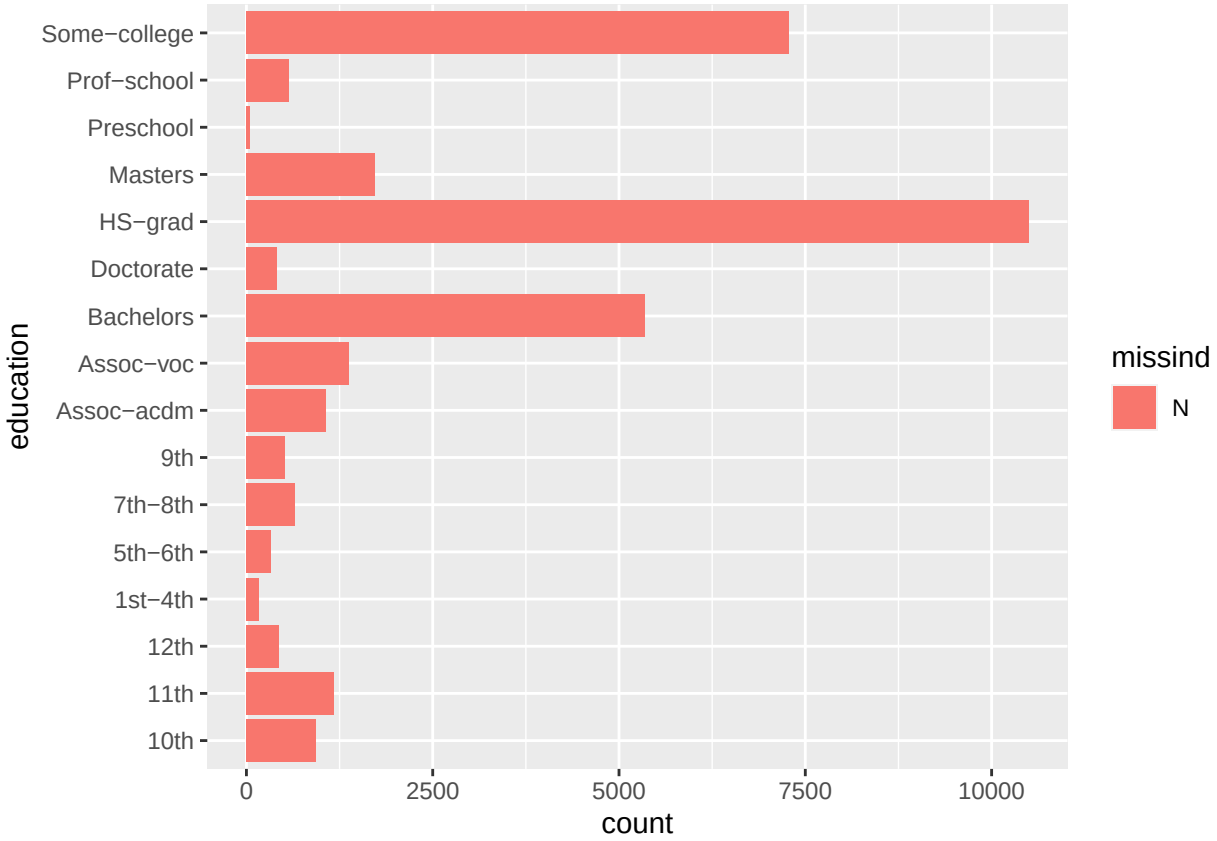
```
##
##  Variables sorted by number of missings:
##        Variable Count
##             age     0
##       workclass     0
##         fnl_wgt     0
##       education     0
##   education_num     0
##  marital_status     0
##      occupation     0
##    relationship     0
##            race     0
##             sex     0
##    capital_gain     0
##    capital_loss     0
##  hours_per_week     0
##  native_country     0
##          income     0
```

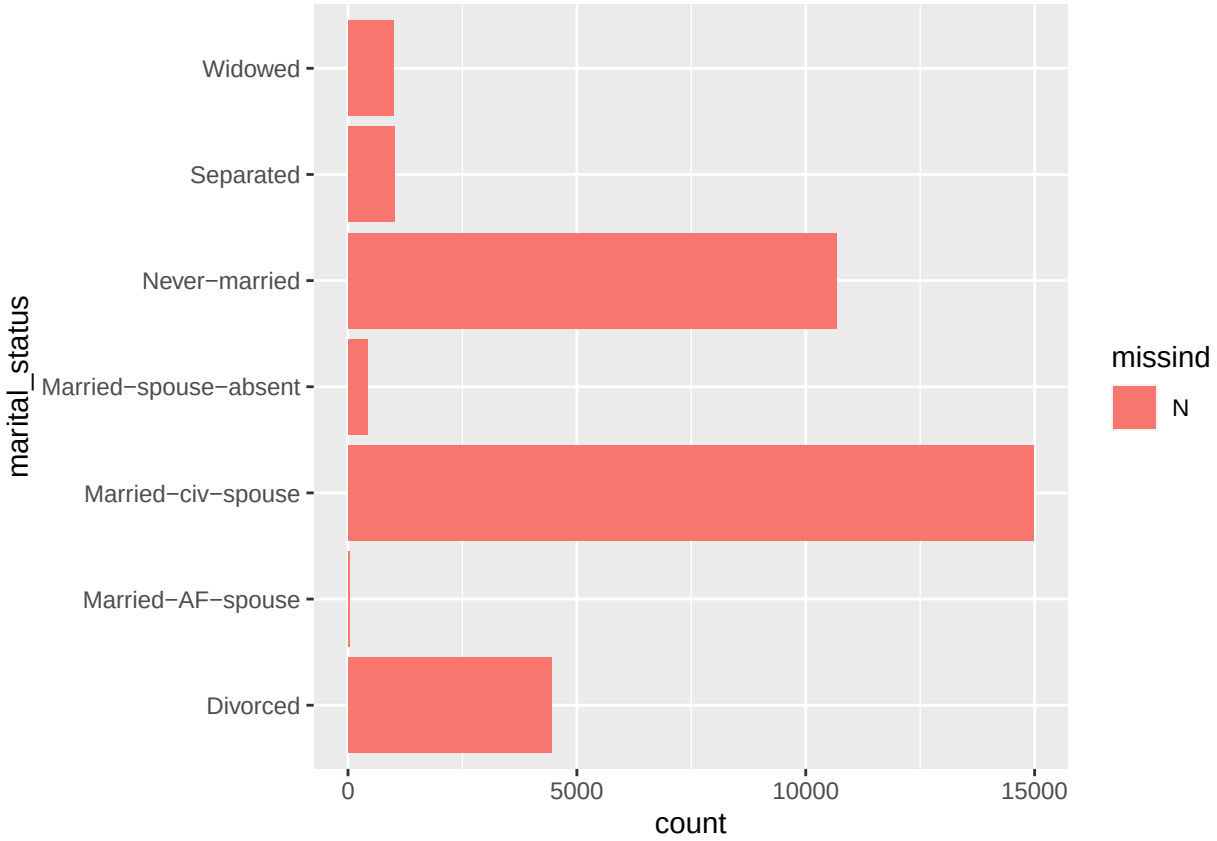
From the above, there are missing values in the data and all the missing values are from categorical variables. Thus we decide to remove the records with missing values.

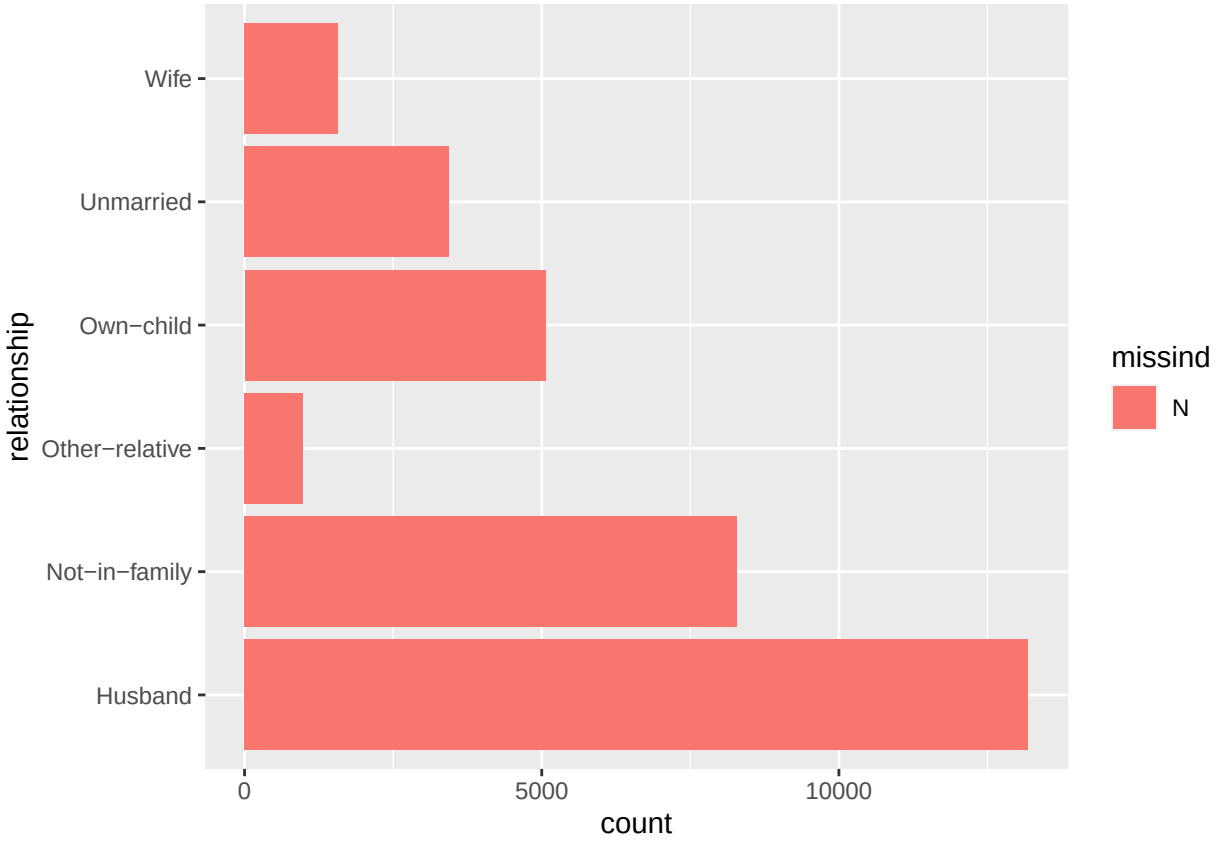
Comparing records with at least one missing value to those without any missing values.

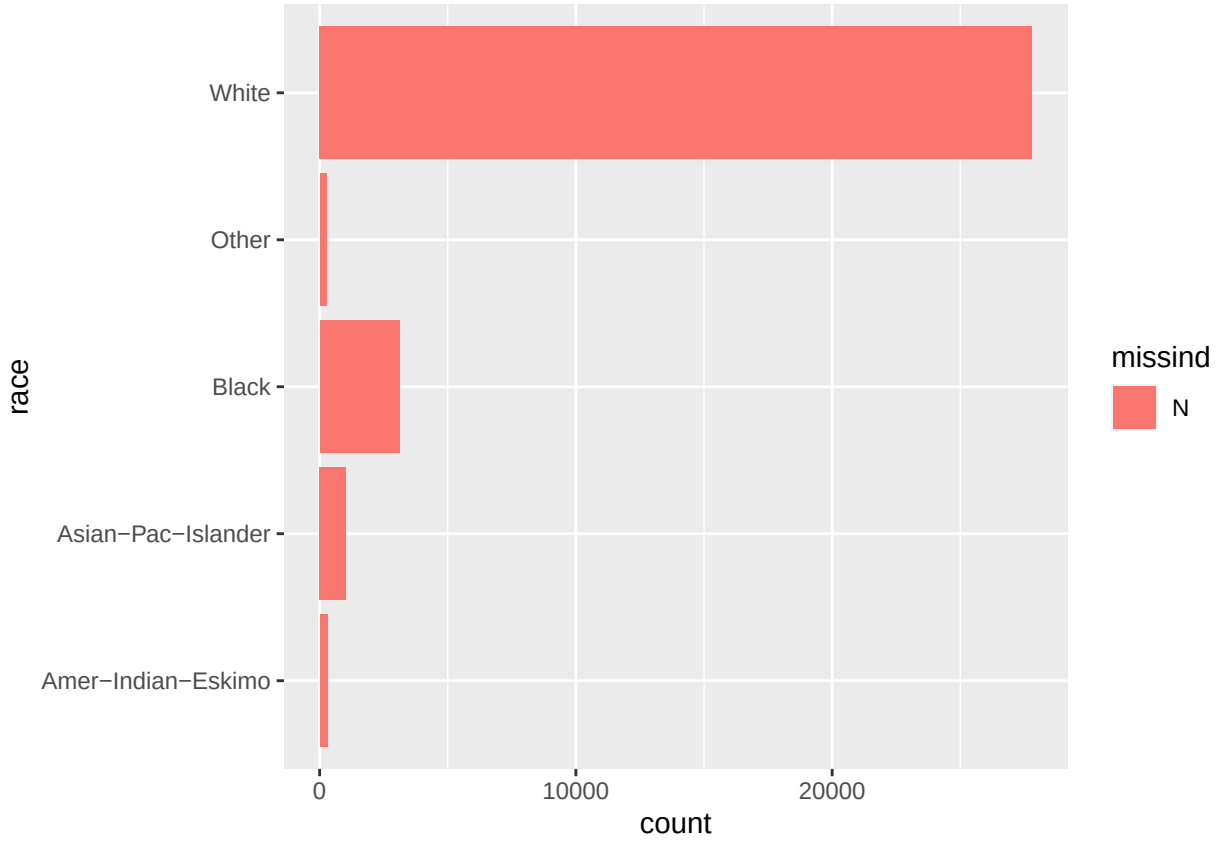
In order to better understand the patterns of the missing values, let's look at some descriptions of the records with missing values.

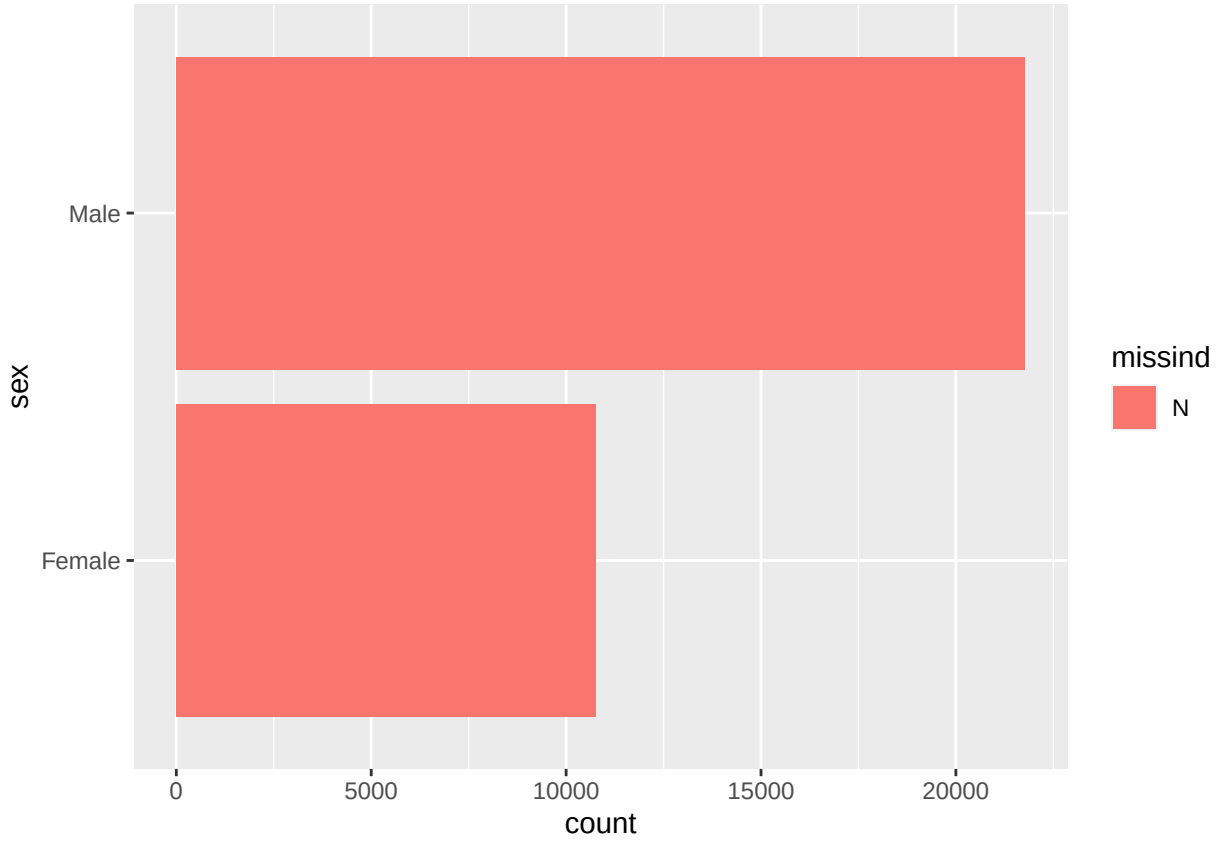


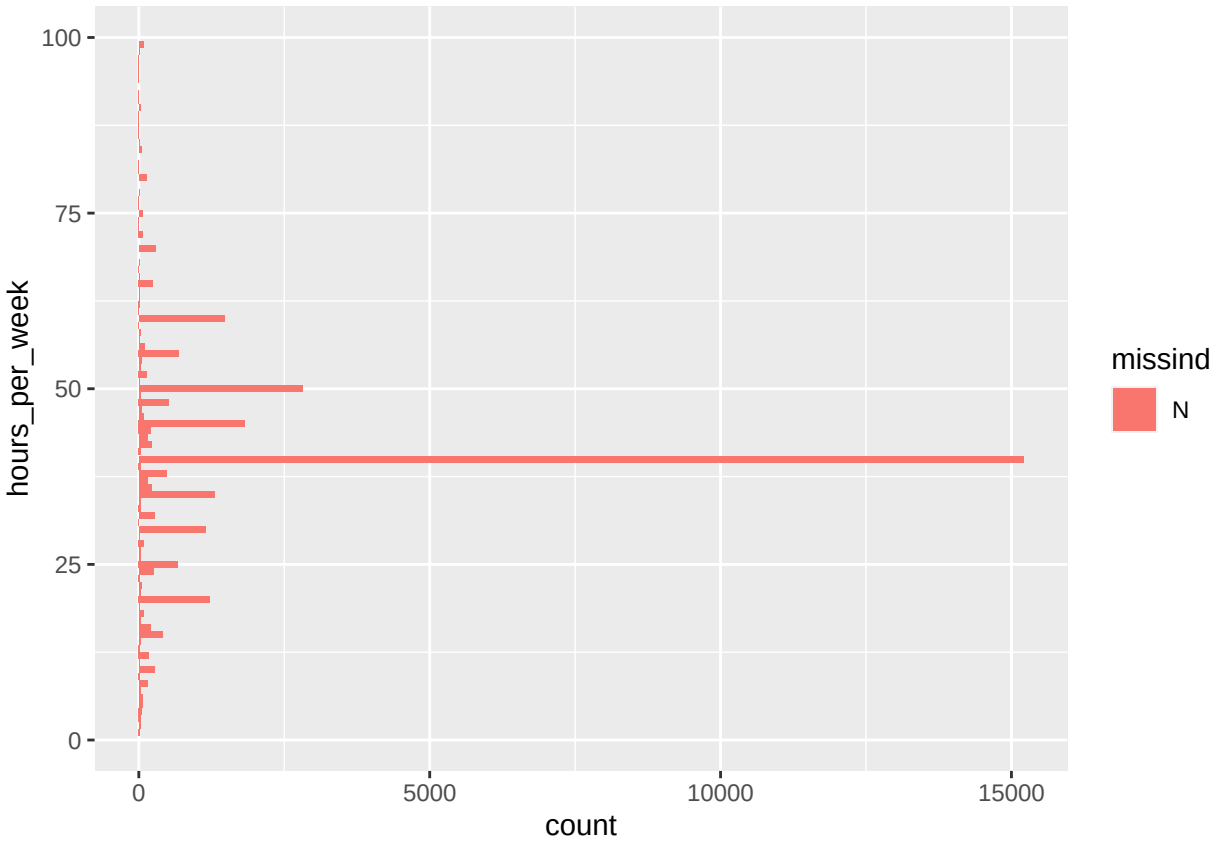

education
Some−college
Prof−school
Preschool
Masters
HS−grad
Doctorate
Bachelors
Assoc−voc
Assoc−acdm
9th
7th−8th
5th−6th
1st−4th
12th
11th
10th
0
2500
5000
7500
10000
count
missind
N

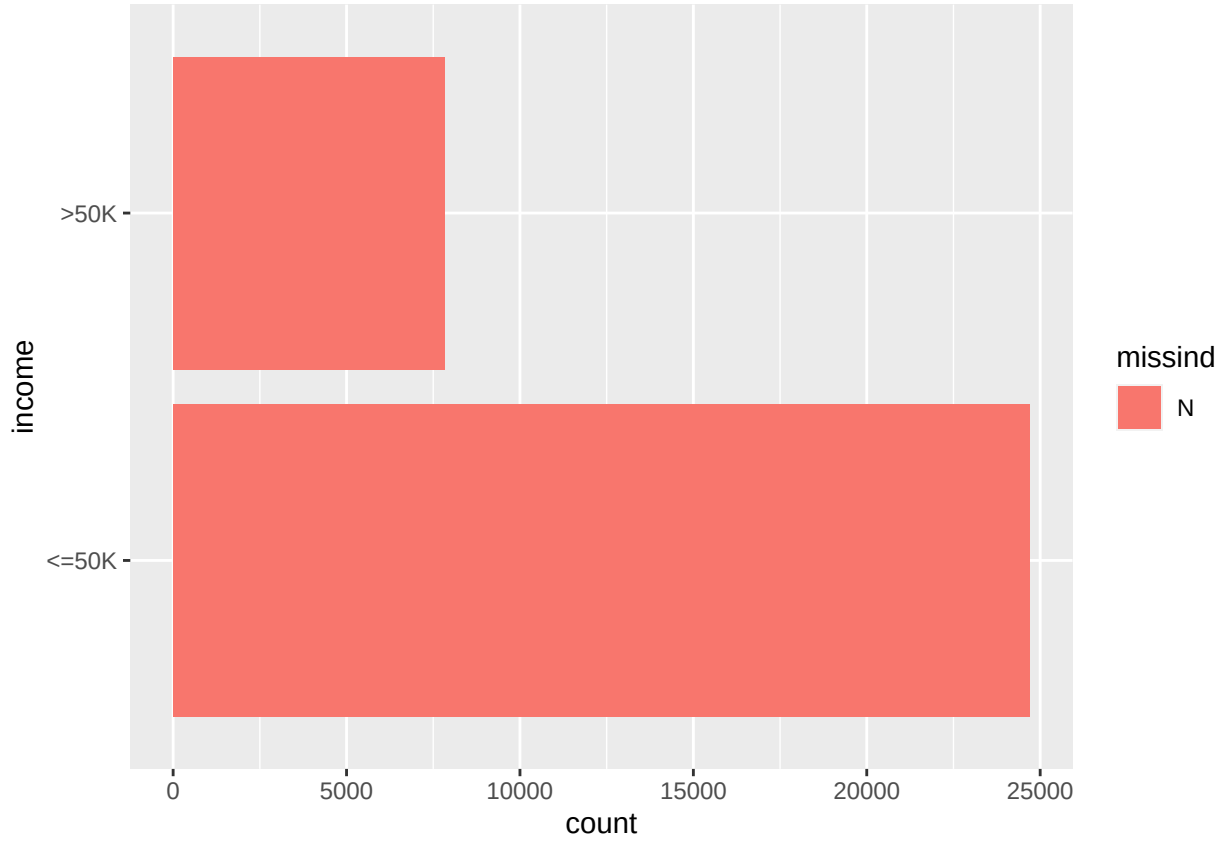

Widowed
Separated
Never−married
Married−spouse−absent
Married−civ−spouse
Married−AF−spouse
Divorced
marital_status
0
5000
10000
15000
count
missind
N


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


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


Male
Female
sex
0
5000
10000
15000
20000
count
missind
N


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



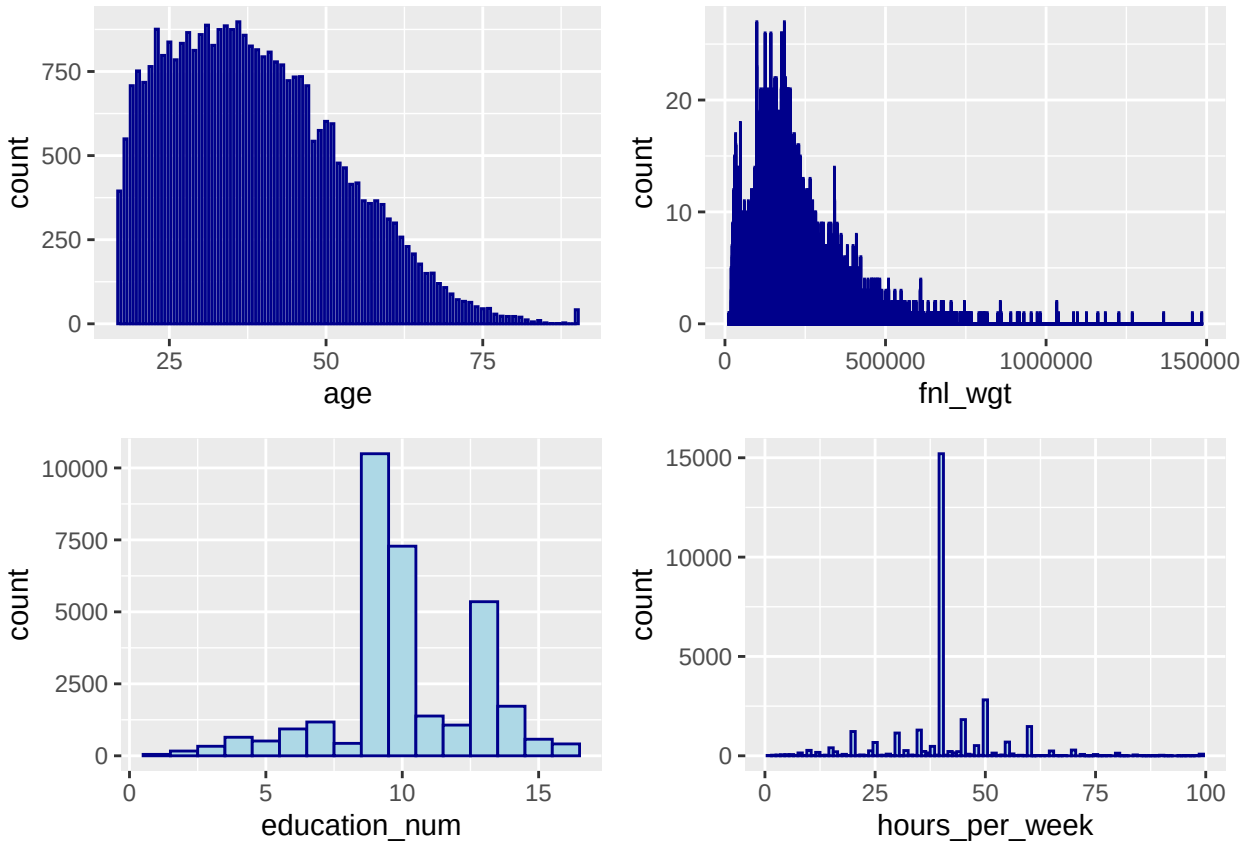
From the above bar charts comparing the distributions of 7 variables of the group that do not have missing values and the group that have at least one missing records, we can see that the missing records are generally evenly distributed across all ages, education level, marital status, family relationship,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

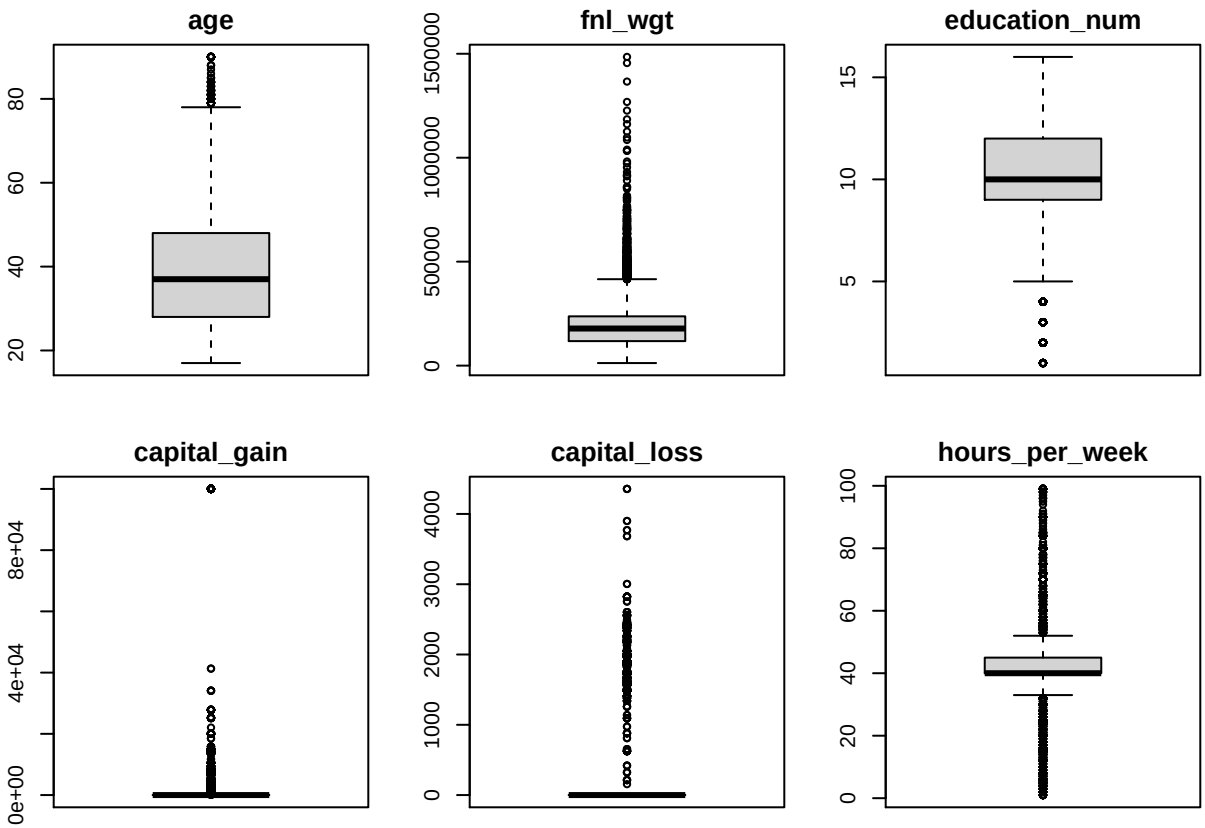
Now let's view the summary of the 6 numeric columns:

	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

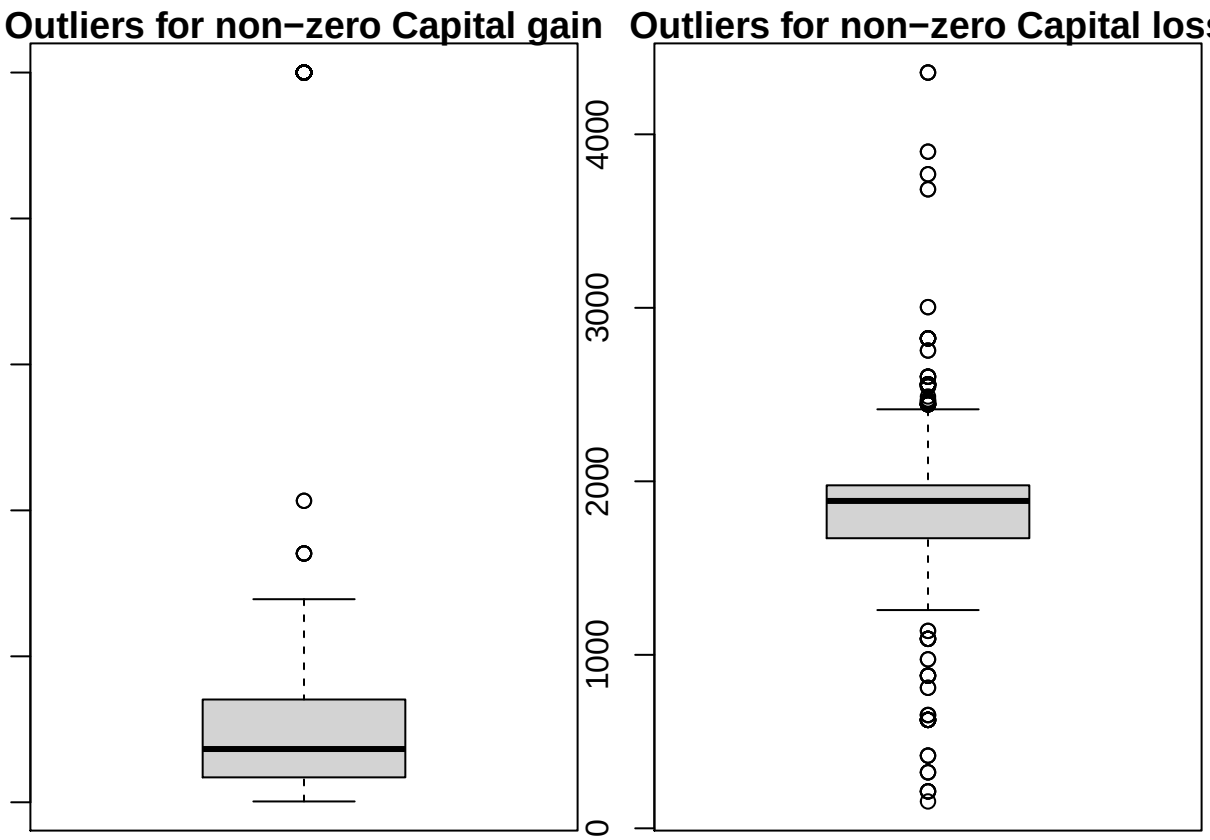
Let's take a clearer look at the numeric values by visualizing their distributions using histograms, except for capital gain and capital loss.



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.

```
# for some reason the missing values are still remaining in this version of the script,
# so do the conversion of ? to NA again and remove the NA values before plotting:

#convert ? values to NA in dataset and save into new dataframe:
X <- as.data.frame(X %>%
                           mutate_if(is.factor, list(~na_if(., "?"))))
```

```
#check that ? values have been converted to NA:
str(X)
```

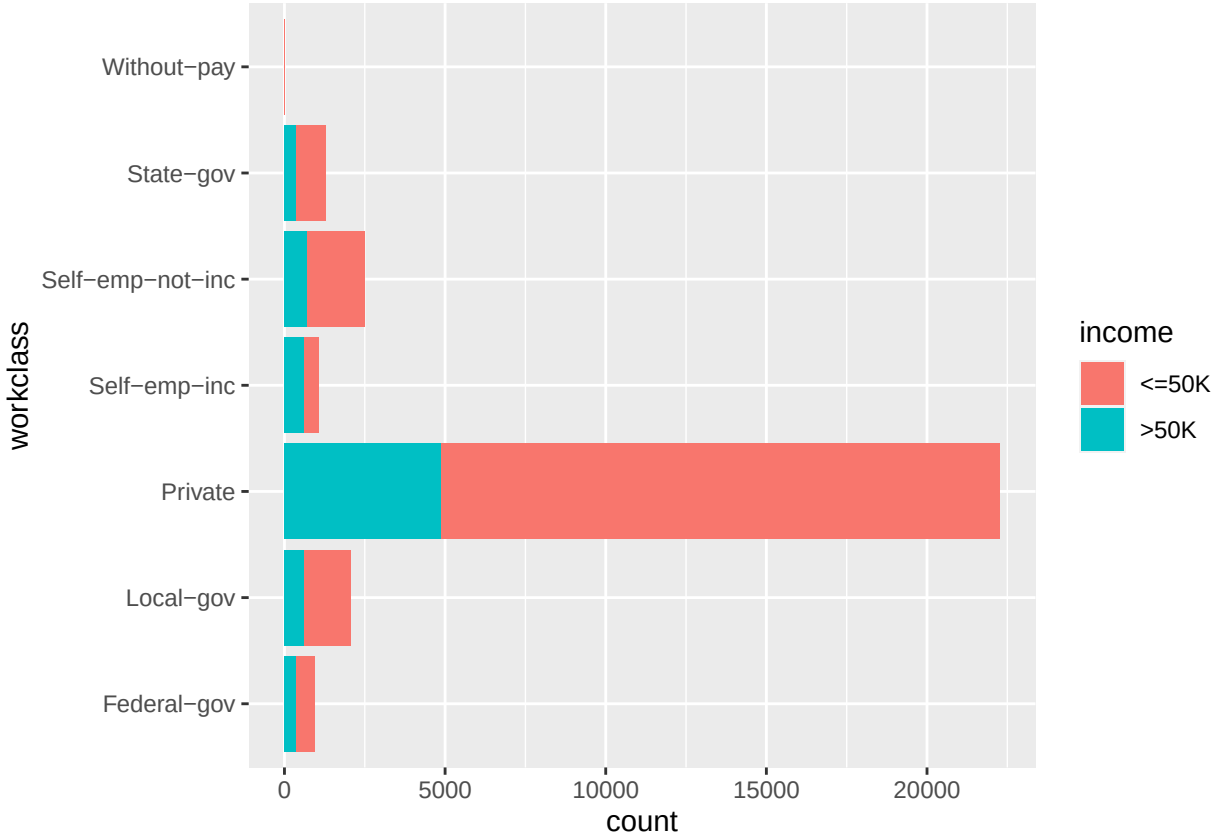
```
## 'data.frame':    32537 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...
##  $ fnl_wgt       : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
```

```
##  $ native_country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
```

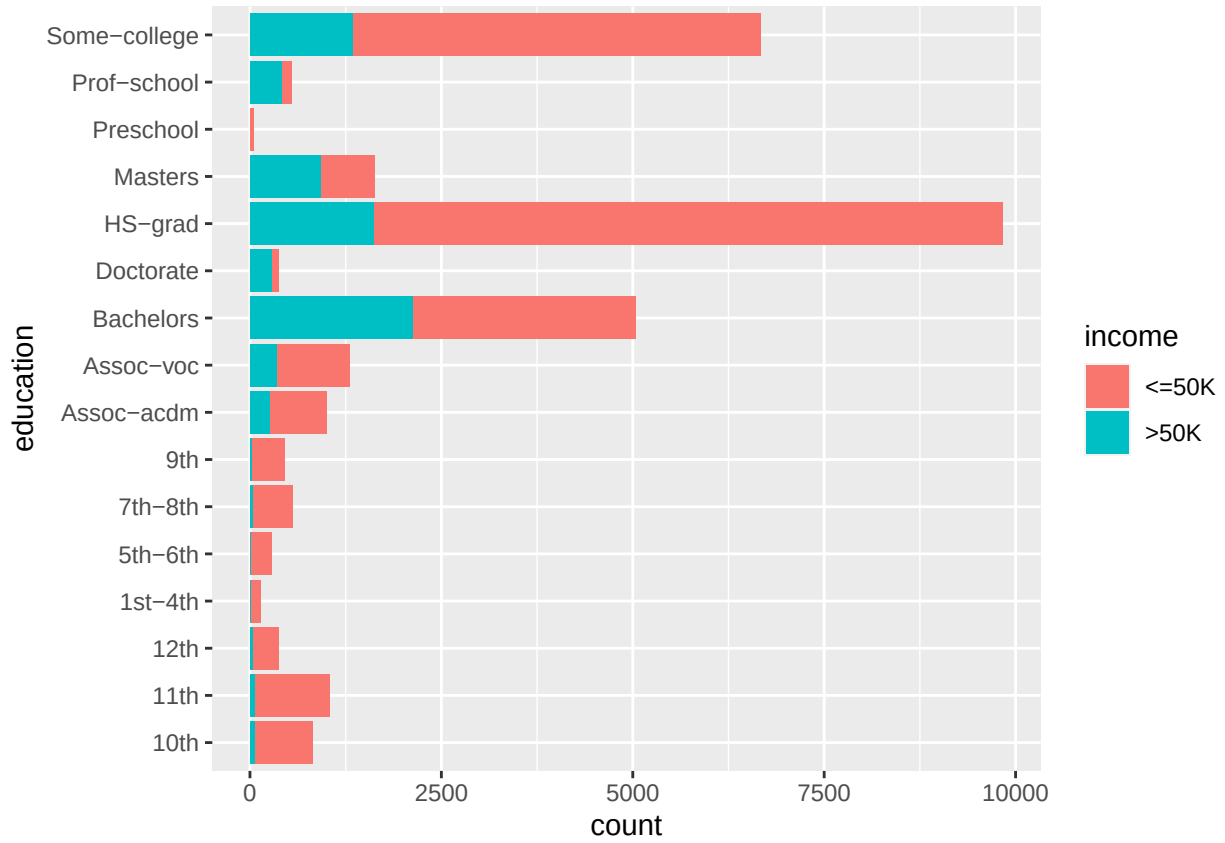
Remove the NA values from the dataset before plotting:

```
X <- na.omit(X)
```

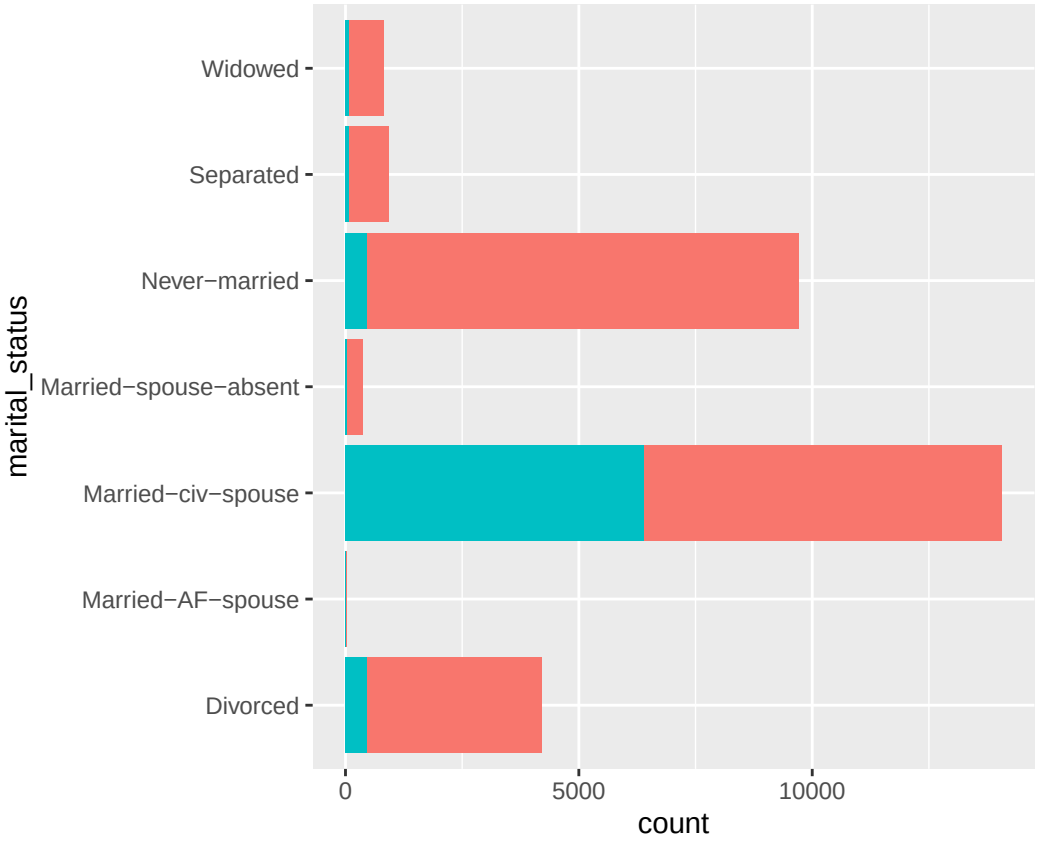
Distributions of categorical variables by target variable



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



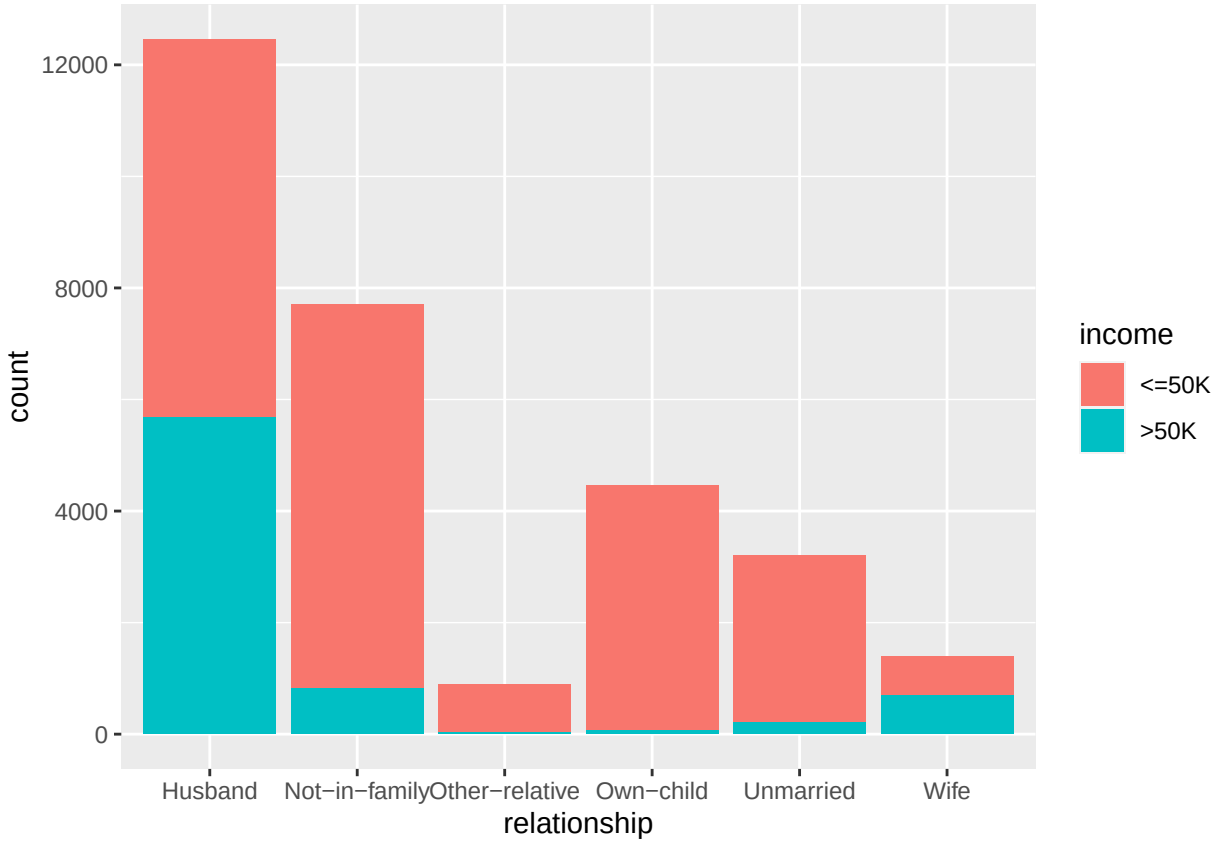
The majority of people earning less than $50K are high school graduates. The next largest education group is some college, and the third largest education group is Bachelors.



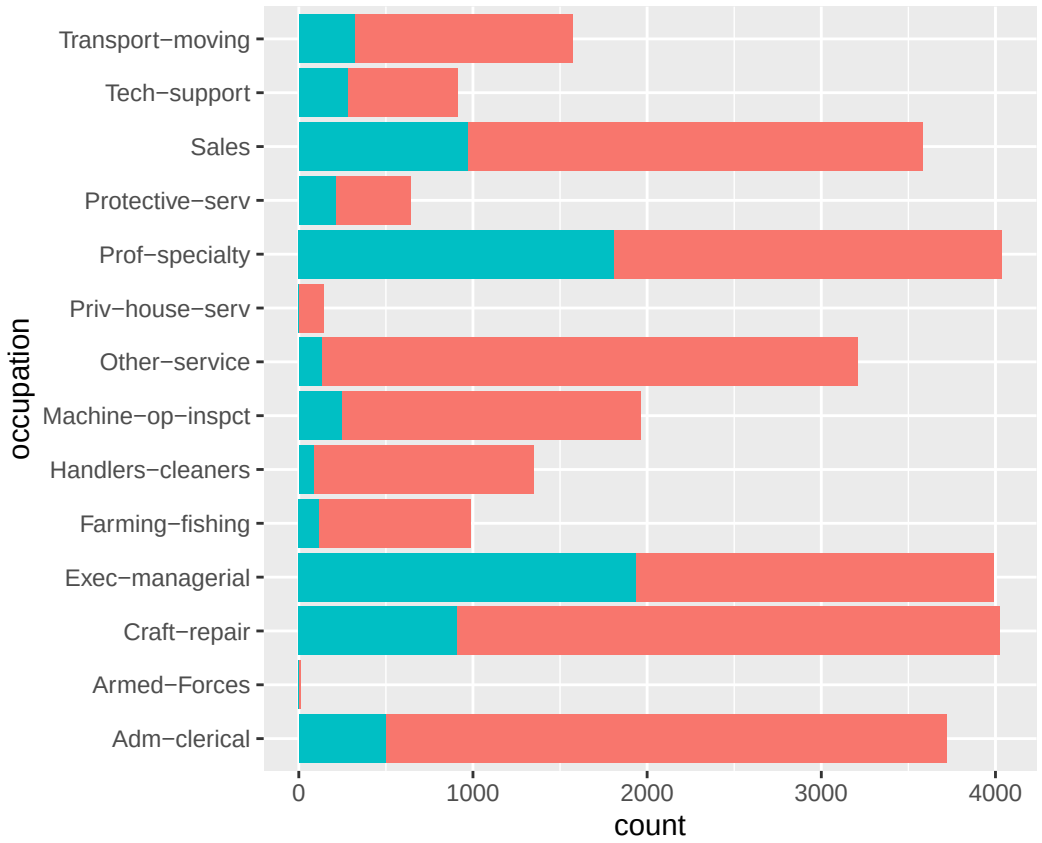
plotting marital status vs. income

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()
```

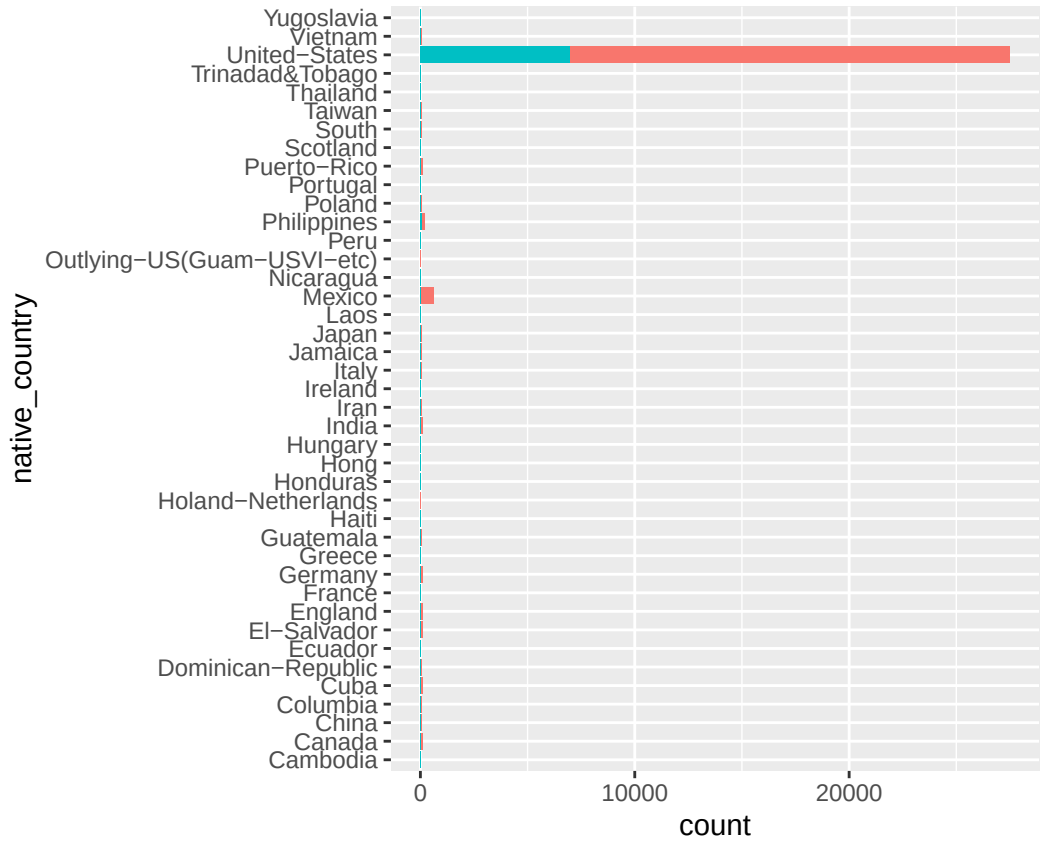


Most people surveyed in the census belong to the Husband category of relationships, with slightly more people earning less than or equal to 50K. However, in the Husband category, there is almost an even split between the 2 target income classes. Not-in-family is the second largest category for relationships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



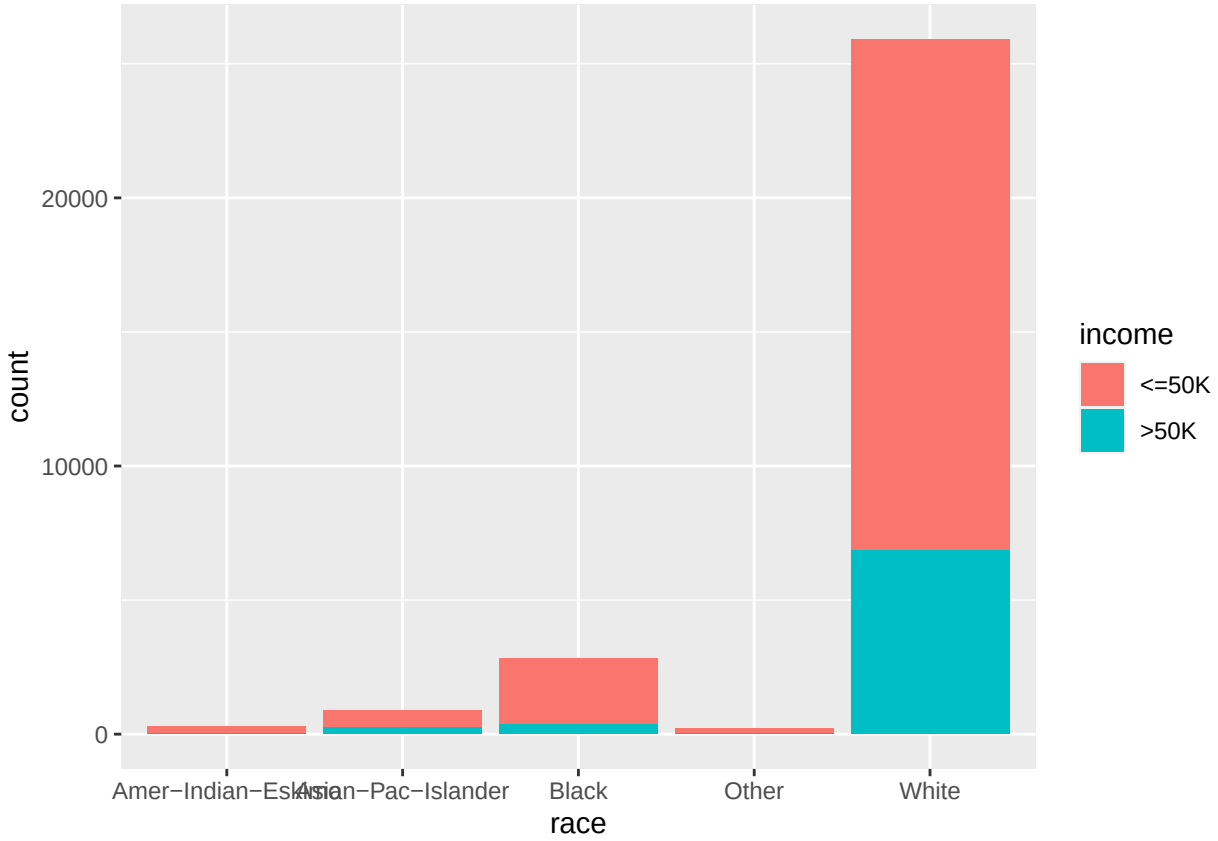
plotting occupation vs income

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.

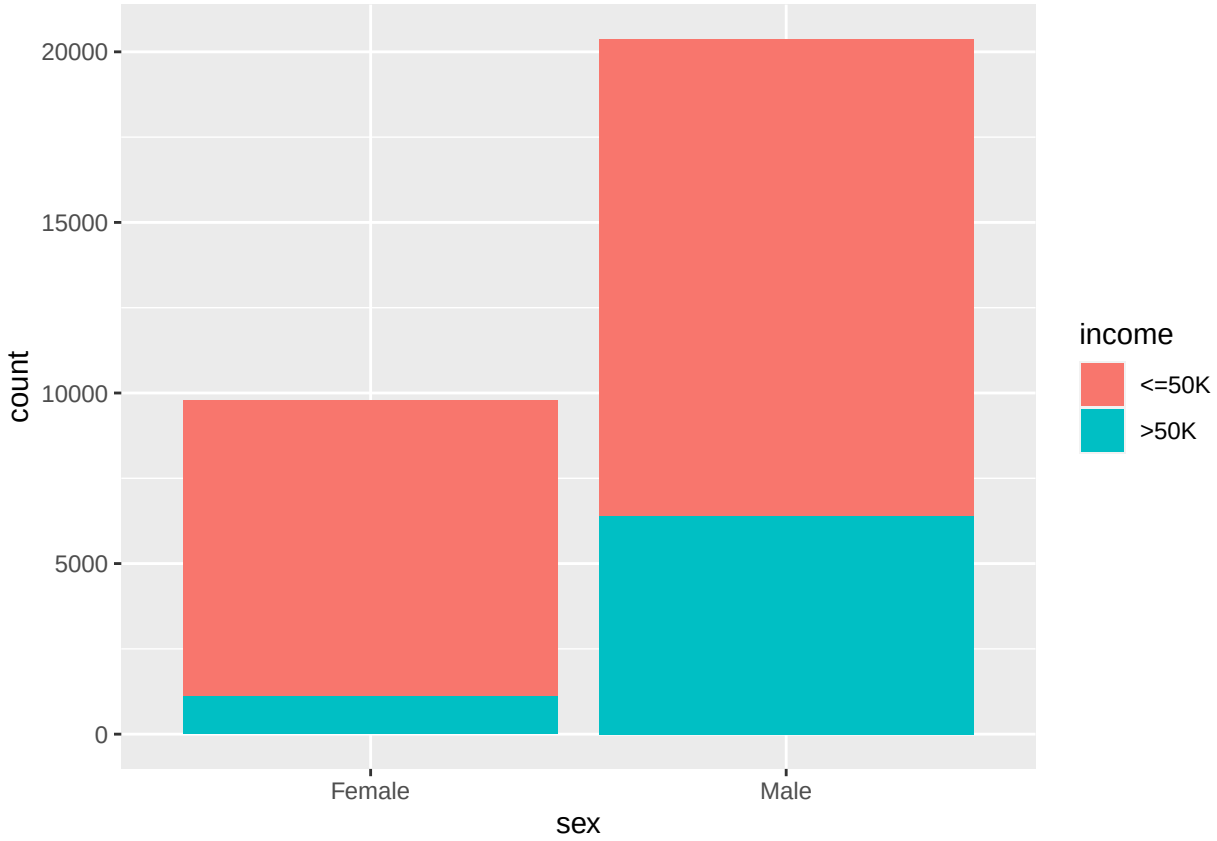


plotting native country vs. income

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.



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



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

The missing ? levels in the X dataset still remain, can drop them:

? Adm-clerical Armed-Forces Craft-repair

0 3719 9 4025

Exec-managerial Farming-fishing Handlers-cleaners Machine-op-inspct

3991 987 1349 1964

Other-service Priv-house-serv Prof-specialty Protective-serv

3209 141 4034 644

Sales Tech-support Transport-moving

3584 911 1572

drop the unused ? level in occupation, workclass and native_country variables

```r
X <- droplevels(X)
```

```r
str(X)
```

```
## 'data.frame':    30139 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : Factor w/ 7 levels "Federal-gov",..: 6 5 3 3 3 3 3 5 3 3 ...
##  $ fnl_wgt       : int  77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...
##  $ education_num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital_status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...
##  $ occupation    : Factor w/ 14 levels "Adm-clerical",..: 1 4 6 6 10 4 8 4 10 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...
##  $ capital_gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital_loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours_per_week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native_country: Factor w/ 41 levels "Cambodia","Canada",..: 39 39 39 39 5 39 23 39 39 39 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...
##  - attr(*, "na.action")= 'omit' Named int [1:2398] 15 28 39 52 62 70 78 94 107 129 ...
##   ..- attr(*, "names")= chr [1:2398] "15" "28" "39" "52" ...
```

#Can plot a Correlation Matrix - default one in R to see the relationship between numeric variables

```
##                      age      fnl_wgt education_num capital_gain capital_loss
## age           1.00000000 -0.076278672    0.04320291  0.080162360   0.06014079
## fnl_wgt      -0.07627867  1.000000000   -0.04519925  0.000420045  -0.00975537
## education_num 0.04320291 -0.045199246    1.00000000  0.124455206   0.07961280
## capital_gain  0.08016236  0.000420045    0.12445521  1.000000000  -0.03225478
## capital_loss  0.06014079 -0.009755370    0.07961280 -0.032254777   1.00000000
## hours_per_week 0.10134839 -0.023033322    0.15284194  0.080428761   0.05238014
##                hours_per_week
## age                 0.10134839
## fnl_wgt            -0.02303332
## education_num       0.15284194
## capital_gain        0.08042876
## capital_loss        0.05238014
## hours_per_week      1.00000000
```

```
## Warning: package 'Hmisc' was built under R version 4.0.3
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 4.0.3
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:xtable':
##
##     label, label<-

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units

##                       age      fnl_wgt education_num capital_gain capital_loss
## age             1.00000000 -0.076278672    0.04320291  0.080162360   0.06014079
## fnl_wgt        -0.07627867  1.000000000   -0.04519925  0.000420045  -0.00975537
## education_num   0.04320291 -0.045199246    1.00000000  0.124455206   0.07961280
## capital_gain    0.08016236  0.000420045    0.12445521  1.000000000  -0.03225478
## capital_loss    0.06014079 -0.009755370    0.07961280 -0.032254777   1.00000000
## hours_per_week  0.10134839 -0.023033322    0.15284194  0.080428761   0.05238014
##                hours_per_week
## age                 0.10134839
## fnl_wgt            -0.02303332
## education_num       0.15284194
## capital_gain        0.08042876
## capital_loss        0.05238014
## hours_per_week      1.00000000
```

```
## look at income with respect to age:
income_by_age =
  X %>%
  group_by(income) %>%
  summarise(mean_age = mean(age, na.rm = TRUE))
```

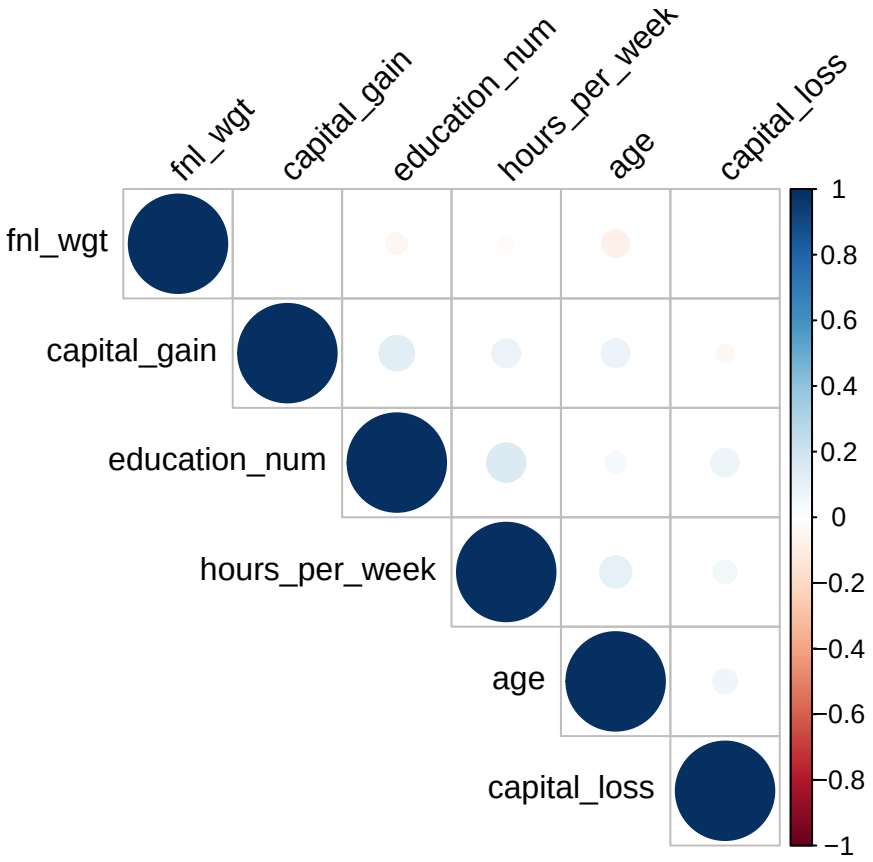
```
#flattening correlation plot
flattenCorrMatrix <- function(cormat, pmat) {
  ut <- upper.tri(cormat)
  data.frame(
    row = rownames(cormat)[row(cormat)[ut]],
    column = rownames(cormat)[col(cormat)[ut]],
    cor  =(cormat)[ut],
    p = pmat[ut]
  )
}
```

```
#flatten correlation matrix:
cor_result_flat = flattenCorrMatrix(cor_result$r, cor_result$P)
head(cor_result_flat)
```

```
##             row        column          cor            p
```

```
## 1           age       fnl_wgt -0.076278672 0.000000e+00
## 2           age education_num  0.043202909 6.217249e-14
## 3       fnl_wgt education_num -0.045199246 3.996803e-15
## 4           age  capital_gain  0.080162360 0.000000e+00
## 5       fnl_wgt  capital_gain  0.000420045 9.418704e-01
## 6 education_num  capital_gain  0.124455206 0.000000e+00
```

Correlation values are low, but some of the P values indicate for example age and education_num are significantly correlated. Fnl_wgt and education num are also significantly correlated.

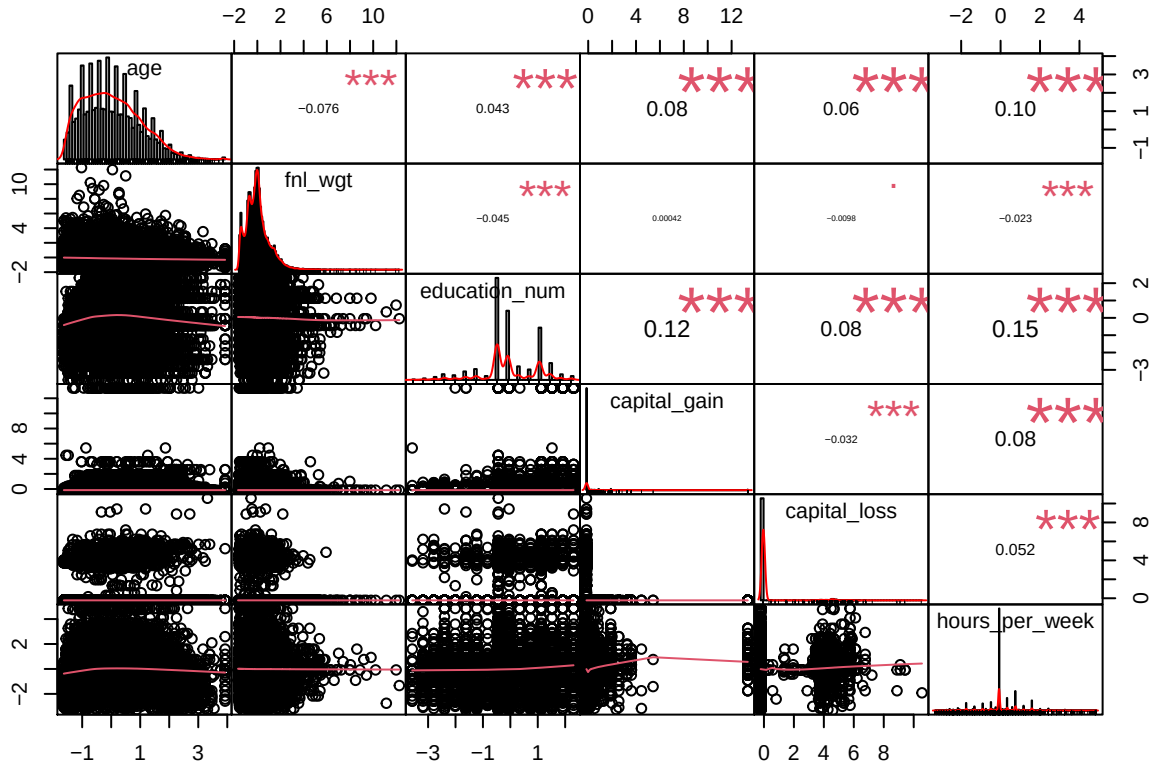


The numeric variables have very small correlation with each other.

Explore relationships between attributes

Pearson's correlation between numeric variables.

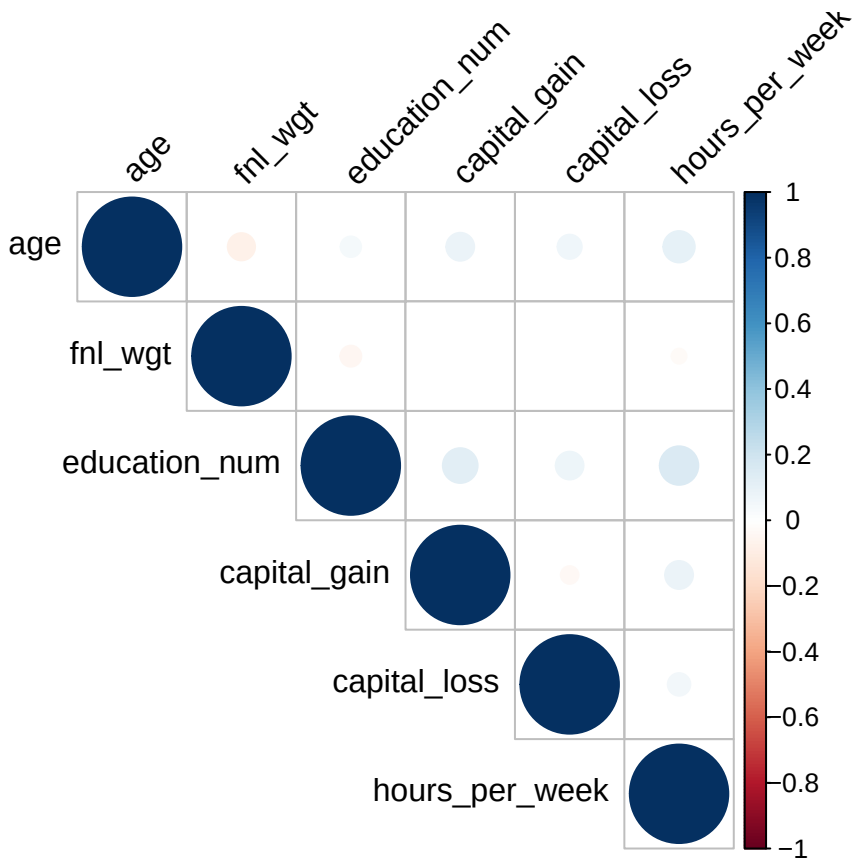
```
#Display the chart of a correlation matrix
library(PerformanceAnalytics)
numindex=datatype=="integer"
chart.Correlation(scale(X[,numindex]), histogram=TRUE, pch=19)
```



From the chart of the correlation matrix, we can see that while the magnitude of the correlations are small, all of them are statistically significant.

The following correlogram confirms that the correlations between numeric variables are very small, yet they are significantly different from zero, probably due to large sample size.

```
library("Hmisc")
cormat <- rcorr(as.matrix(X[,numindex]))
#Draw a correlogram
library(corrplot)
corrplot(cormat$r, type = "upper",
         tl.col = "black", tl.srt = 45, p.mat = cormat$P, sig.level = 0.01, insig = "blank")
```



Chi-square test & Cramer's V to show associations between categorical variables

```
## Warning: package 'DescTools' was built under R version 4.0.3
```

The test statistics from Chiq-Square Test between each pair of the categorical variables.

	workclass	education	marital_status	occupation	relationship	race	sex	native_country	income
workclass		2445.29	1720534.10	9314.40	9314.40	14712.97	4214.70	12388.39	357.31
education			127730.72	2182.09	2182.09	1078.26	8534.11	1198.45	398.14
marital_status				321142.94	321142.94	137835.19	269434.63	115466.27	111620.41
occupation					452085.00	1364.78	15299.12	2088.27	689.73
relationship						1364.78	15299.12	2088.27	689.73
race							3154.45	35767.89	844.68
sex								4814.17	846.78
native_country									1136.52
income									

The p-values from the Chi-Square Test between each pair of the categorical variables.

	workclass	education	marital_status	occupation	relationship	race	sex	native_country	income
workclass		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
education			0.00	0.00	0.00	0.00	0.00	0.00	0.00
marital_status				0.00	0.00	0.00	0.00	0.00	0.00
occupation					0.00	0.00	0.00	0.00	0.00
relationship						0.00	0.00	0.00	0.00
race							0.00	0.00	0.00
sex								0.00	0.00
native_country									0.00
income									

The Cramer's V statistics between each pair of categorical variables to measure their associations.

	workclass	education	marital_status	occupation	relationship	race	sex	native_country	income
workclass		0.12	0.90	0.14	0.14	0.29	0.10	0.29	0.05
education			0.84	0.11	0.11	0.08	0.22	0.09	0.06
marital_status				0.84	0.84	0.87	0.83	0.88	0.96
occupation					1.00	0.09	0.20	0.12	0.08
relationship						0.09	0.20	0.12	0.08
race							0.13	0.49	0.08
sex								0.18	0.08
native_country									0.10
income									

Feature Engineering

From the above exploratory analysis on the numeric and categorical variables, we think the following transformations can be adopted to help with building predictive models. **1.Education and education number are redundant.**

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
10th	0	0	0	0	0	820	0	0	0	0	0	0	0	0	0	0
11th	0	0	0	0	0	0	1048	0	0	0	0	0	0	0	0	0
12th	0	0	0	0	0	0	0	377	0	0	0	0	0	0	0	0
1st-4th	0	149	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5th-6th	0	0	287	0	0	0	0	0	0	0	0	0	0	0	0	0
7th-8th	0	0	0	556	0	0	0	0	0	0	0	0	0	0	0	0
9th	0	0	0	0	455	0	0	0	0	0	0	0	0	0	0	0
Assoc-acdm	0	0	0	0	0	0	0	0	0	0	0	1008	0	0	0	0
Assoc-voc	0	0	0	0	0	0	0	0	0	0	1307	0	0	0	0	0
Bachelors	0	0	0	0	0	0	0	0	0	0	0	0	5042	0	0	0
Doctorate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	375
HS-grad	0	0	0	0	0	0	0	0	9834	0	0	0	0	0	0	0
Masters	0	0	0	0	0	0	0	0	0	0	0	0	0	1626	0	0
Preschool	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Prof-school	0	0	0	0	0	0	0	0	0	0	0	0	0	0	542	0
Some-college	0	0	0	0	0	0	0	0	0	6669	0	0	0	0	0	0

From the above perfectly 1-1 relationship, we can see these two variables are essentially exact the same. So we decide to remove the educationnum variable.

Modeling

For the simplicity in coding, let's recode the values of target variable to be "Y", meaning yearly income is higher than 50K USA, and "N", indicating the income is no more than 50K.

```
## Now the levels of the target variable are:
```

```
##     N     Y
## 22633  7506
```

currently the country variable has 41 levels, so we will group countries together by region to simplify the number of levels:

```
levels(X$native_country)
```

```
##  [1] "Cambodia"                    "Canada"
##  [3] "China"                       "Columbia"
##  [5] "Cuba"                        "Dominican-Republic"
##  [7] "Ecuador"                     "El-Salvador"
```

```
##  [9] "England"                   "France"
## [11] "Germany"                   "Greece"
## [13] "Guatemala"                 "Haiti"
## [15] "Holand-Netherlands"        "Honduras"
## [17] "Hong"                      "Hungary"
## [19] "India"                     "Iran"
## [21] "Ireland"                   "Italy"
## [23] "Jamaica"                   "Japan"
## [25] "Laos"                      "Mexico"
## [27] "Nicaragua"                 "Outlying-US(Guam-USVI-etc)"
## [29] "Peru"                      "Philippines"
## [31] "Poland"                    "Portugal"
## [33] "Puerto-Rico"               "Scotland"
## [35] "South"                     "Taiwan"
## [37] "Thailand"                  "Trinadad&Tobago"
## [39] "United-States"             "Vietnam"
## [41] "Yugoslavia"
```

Adapted from (https://rpubs.com/vassitar/us_census_preprocessing)

Convert Hong to Hong Kong in the country region factor level for clarity:

```
#replace "Hong" to "Hong Kong" in native_country column for clarity:
X$native_country <-  str_replace(as.character(X$native_country), "Hong", "Hong Kong")

#convert it back to a factor
X$native_country <- as.factor(X$native_country)
```

We will group the countries into country regions.

```
Asia_East <- c("Cambodia", "China", "Hong Kong", "Laos", "Thailand", "Japan", "Taiwan", "Vietnam", "Phi

Asia_Central <-c("India", "Iran")

Central_America <- c("Cuba", "Guatemala", "Jamaica", "Nicaragua", "Puerto-Rico", "Dominican-Republic",

South_America <-c("Ecuador", "Peru", "Columbia")

Europe_West <- c("England", "Scotland", "Germany", "Holand-Netherlands", "Ireland", "France", "Greece",

Europe_East <- c("Poland", "Yugoslavia", "Hungary")

North_America <- c("Canada", "Mexico", "United-States", "Outlying-US(Guam-USVI-etc)")
```

Then we will create a new variable called native region which will contain the country region values as categorised above.

```
X <- mutate(X, native_region = ifelse(native_country %in% Asia_East, "East Asia",
                               ifelse(native_country %in% Asia_Central, "Central Asia",
                               ifelse(native_country %in% Central_America, "Central America",
                               ifelse(native_country %in% South_America, "South America",
                               ifelse(native_country %in% Europe_West, "Europe West",
                               ifelse(native_country %in% Europe_East, "Europe East",
                               ifelse(native_country %in% North_America, "North America","North-America"
```

```
{r. echo= FALSE} summary(X$capital_gain)
```

What is the mean for capital gain for non zero values?

What is the mean of capital loss for non zero values?

```
library(knitr)
```

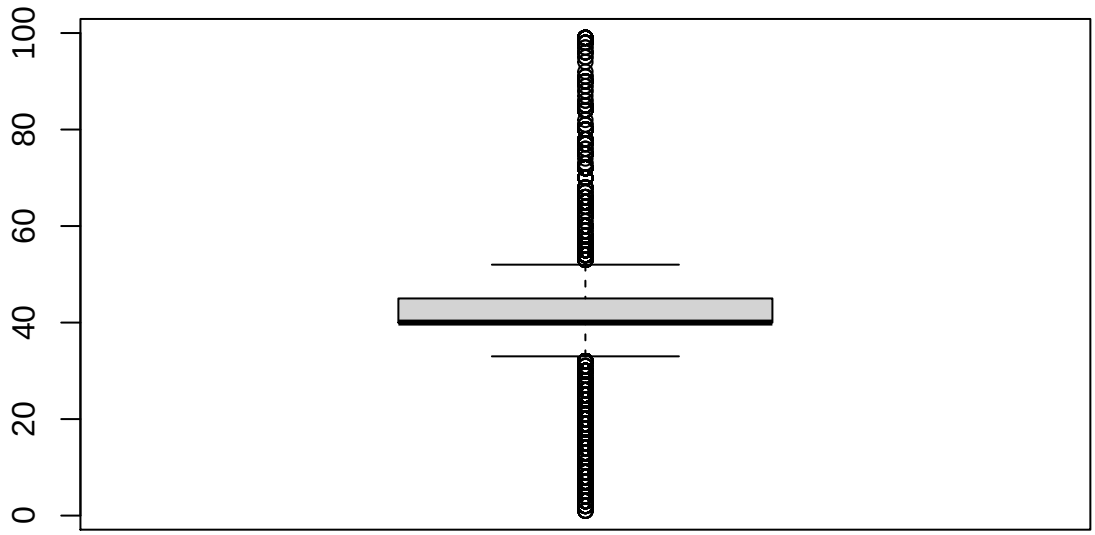
```
## Warning: package 'knitr' was built under R version 4.0.3
```

```
quantile_nonzero_cap_gain<-quantile(x = subset(X$capital_gain, X$capital_gain > 0),
                    probs = seq(0, 1, 0.25))
quantile_nonzero_cap_loss<-quantile(x = subset(X$capital_loss, X$capital_loss > 0),
                    probs = seq(0, 1, 0.25))
kable(x = data.frame(CapitalGain = quantile_nonzero_cap_gain, CapitalLoss = quantile_nonzero_cap_loss),
      caption = "Quantiles of the Nonzero Capital")
```

Table 2: Quantiles of the Nonzero Capital

	CapitalGain	CapitalLoss
0%	114	155
25%	3464	1672
50%	7298	1887
75%	14084	1977
100%	99999	4356

most values in capital gain is zero, let's put 0 values in their own category.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   40.00   40.00   40.93   45.00   99.00
```

We decide to group this variable in the following way: if the value is lower than the 1st quantile (40), it's called "less_than_40". If a value is between the 1st and 3rd quantitle (45), it's called "between_40_and_45". If the value is higher than 3rd quantile, it's called "higher_than_45".

Variables that we will remove for the analysis are education num, finalweight, capital_gain, capital_loss, hours per week, and native_country.

We are going to remove final weight for the model because it has to do with a weighted value for certain demographics when the census was conducted. However, it would not make sense as an input for individual users of the Shiny app.

Education num is redundant with education. Capital_gain and capital_loss, and hours_per_week, native_country can be removed because they are replaced with cap_gain, cap_loss, hours_w and native_region.