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.

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.)

Column Name	Data Type	Column Description
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

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:

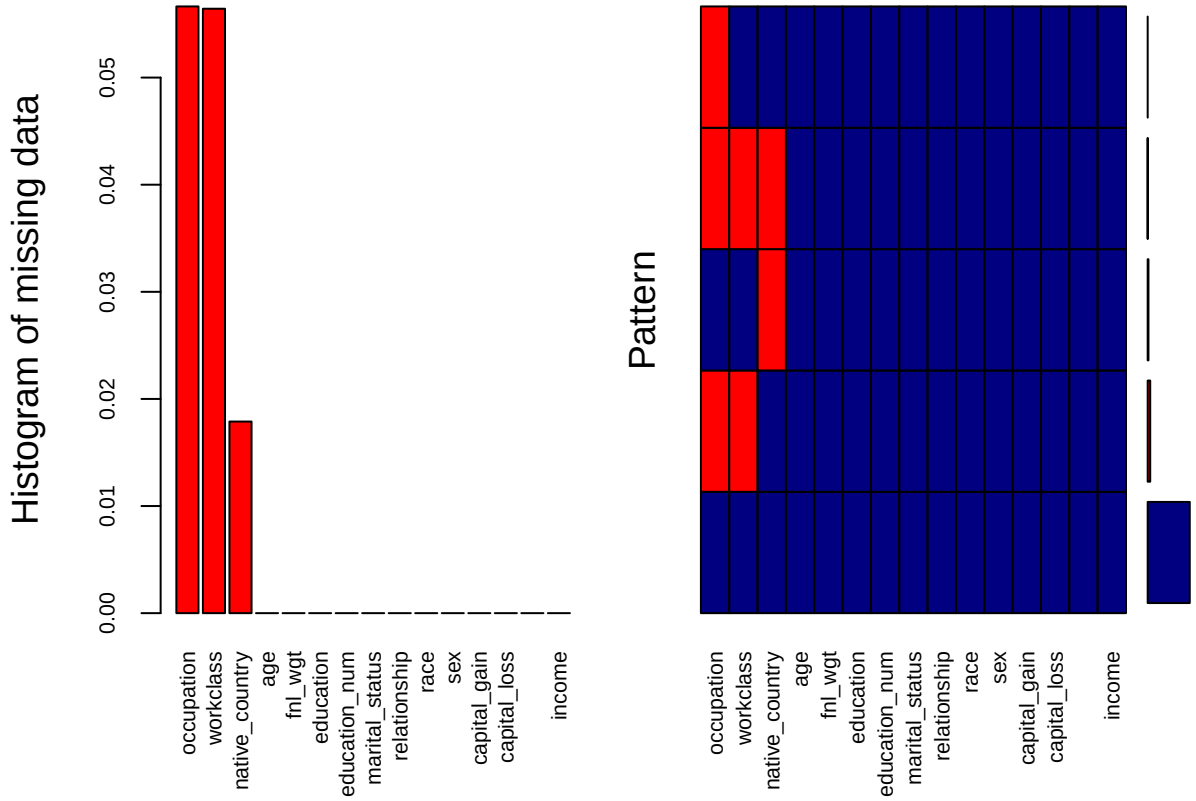
	age	workclass	fnl_wgt	education	education_num	marital_status	occupation	relationship
4768	21	Private	250051	Some-college	10	Never-married	Prof-specialty	Own-child
9172	21	Private	250051	Some-college	10	Never-married	Prof-specialty	Own-child
4326	25	Private	308144	Bachelors	13	Never-married	Craft-repair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

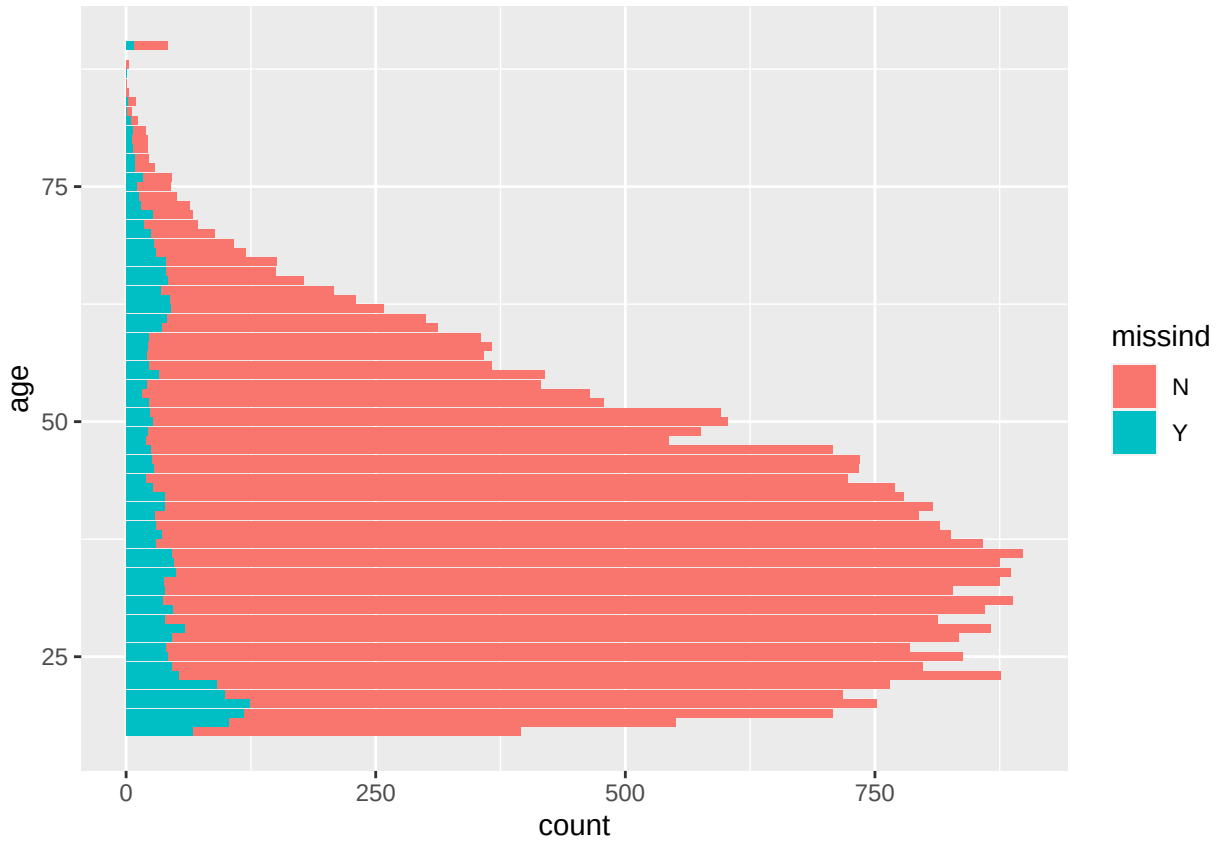


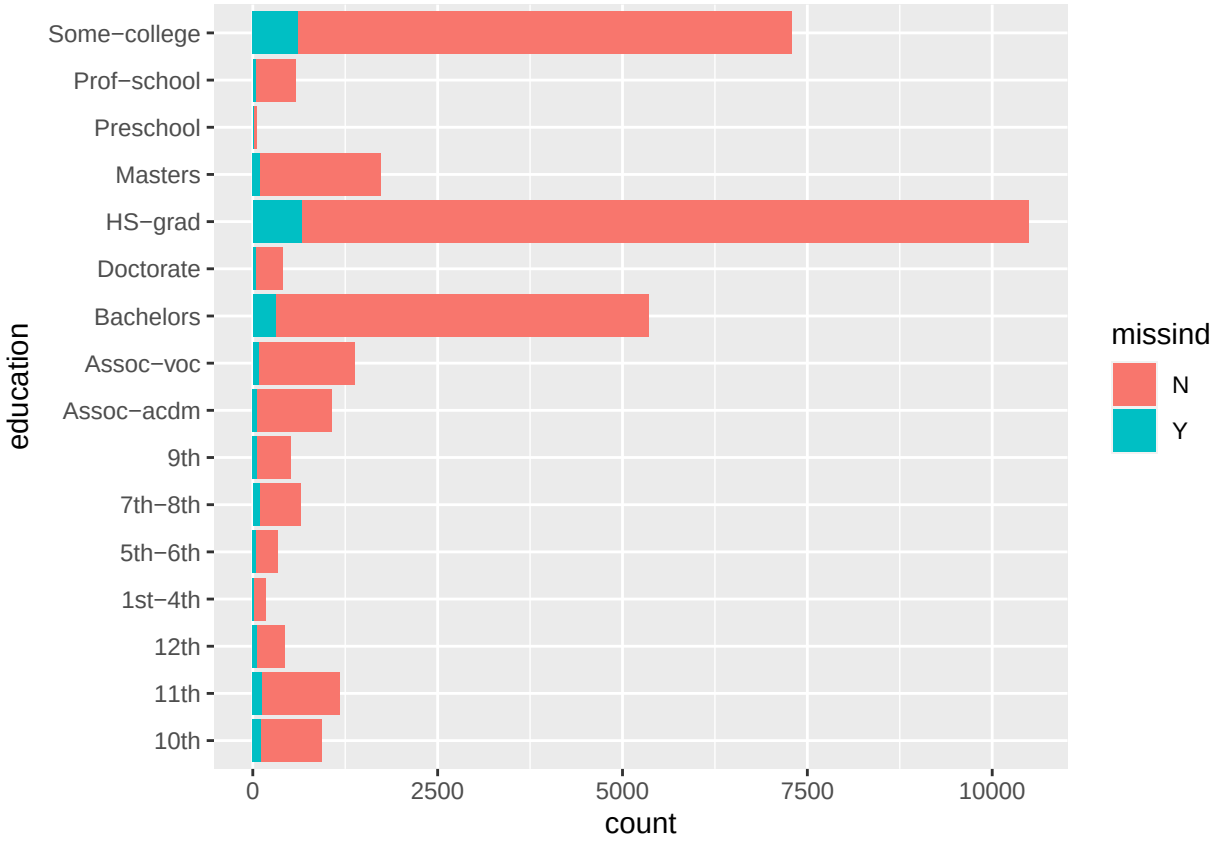

```
##
##  Variables sorted by number of missings:
##        Variable      Count
##      occupation 0.05664321
##       workclass 0.05642807
##  native_country 0.01788733
##             age 0.00000000
##         fnl_wgt 0.00000000
##       education 0.00000000
##   education_num 0.00000000
##  marital_status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

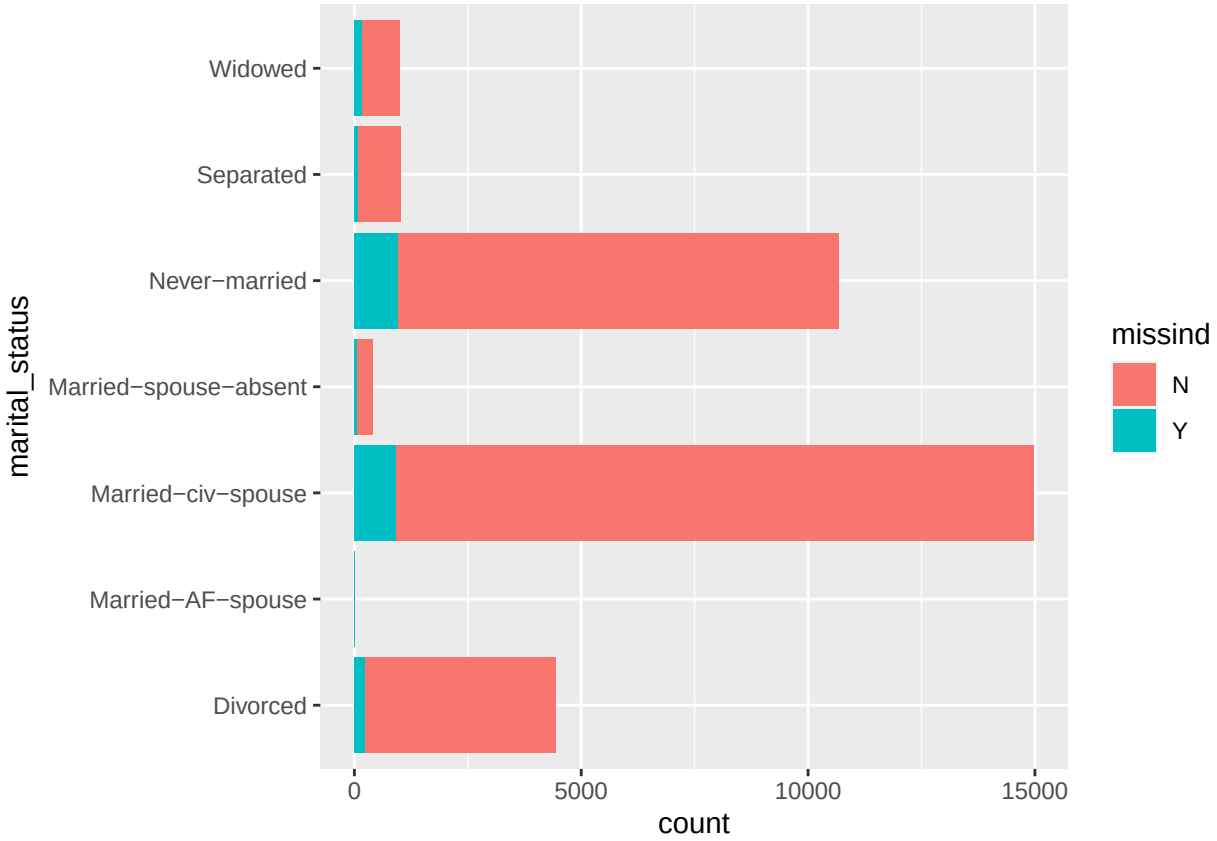
From the above, there are missing values in this data set and all the missing values are from categorical variables.

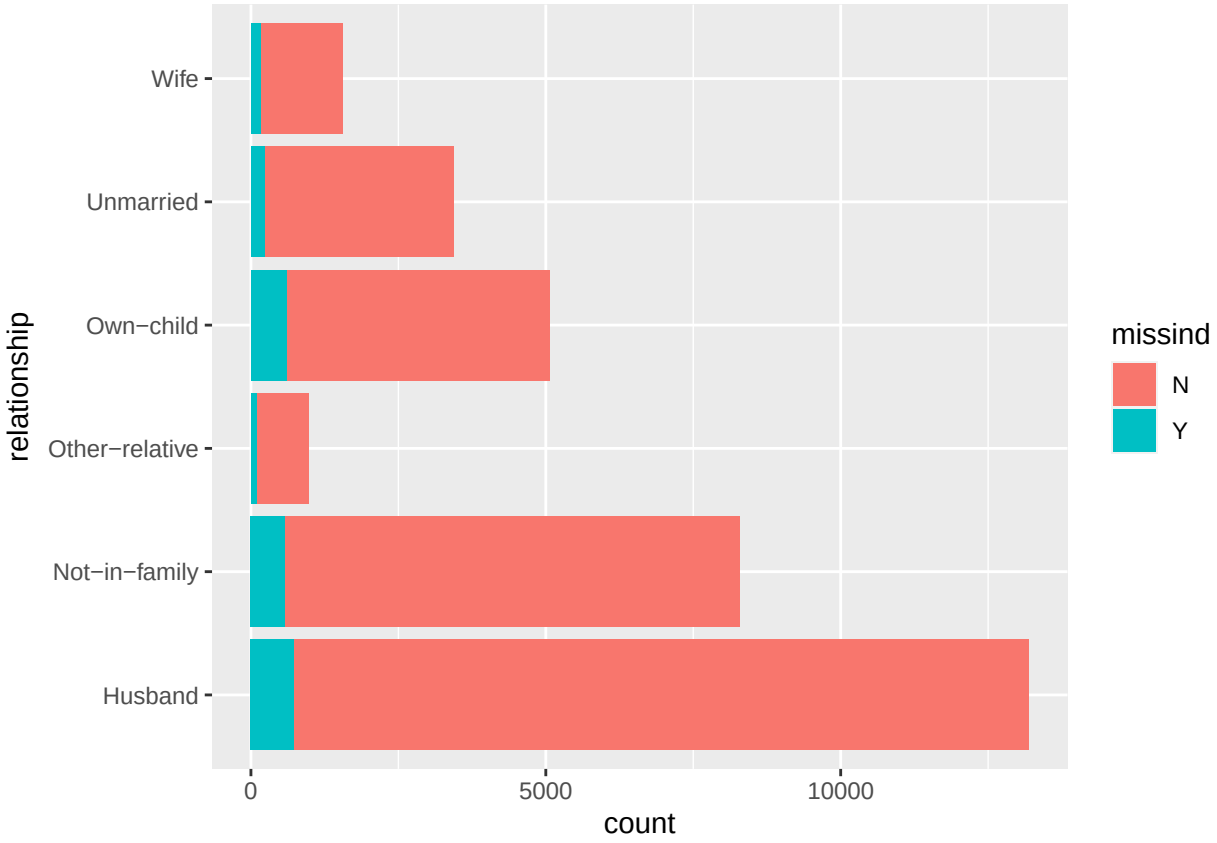
Comparing records with at least one missing value to those without any missing values.

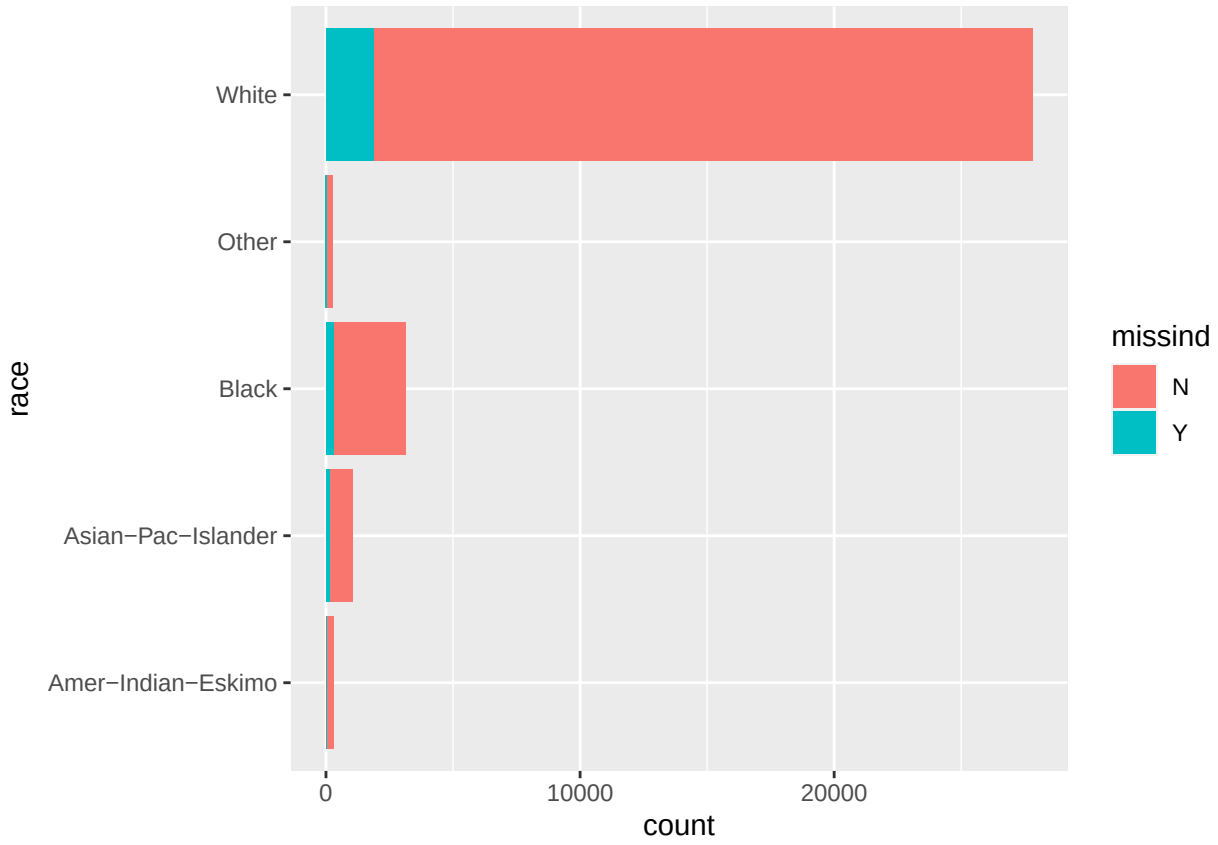
In order to better understand the patterns of the missing values, let's look at some descriptions of the records with missing values.

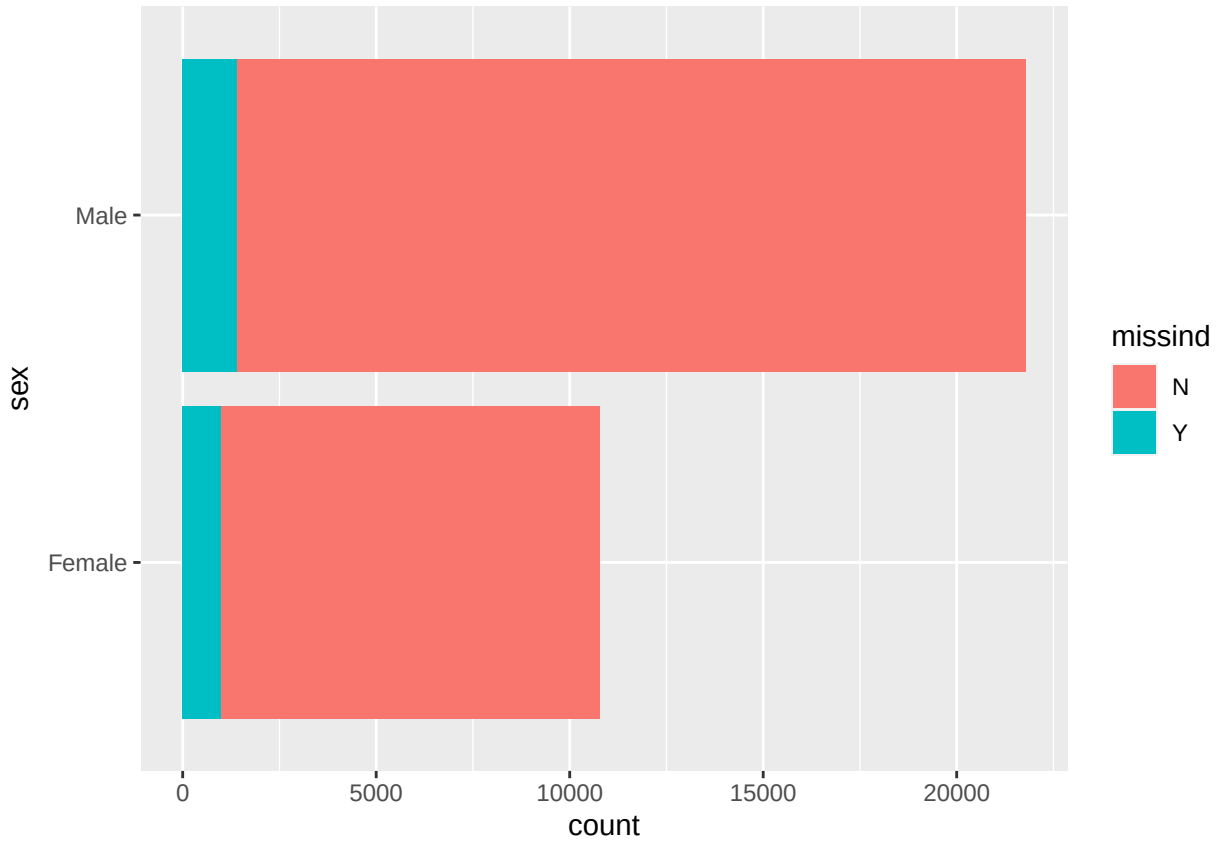

age
75
50
25
0
250
500
750
count
missind
N
Y

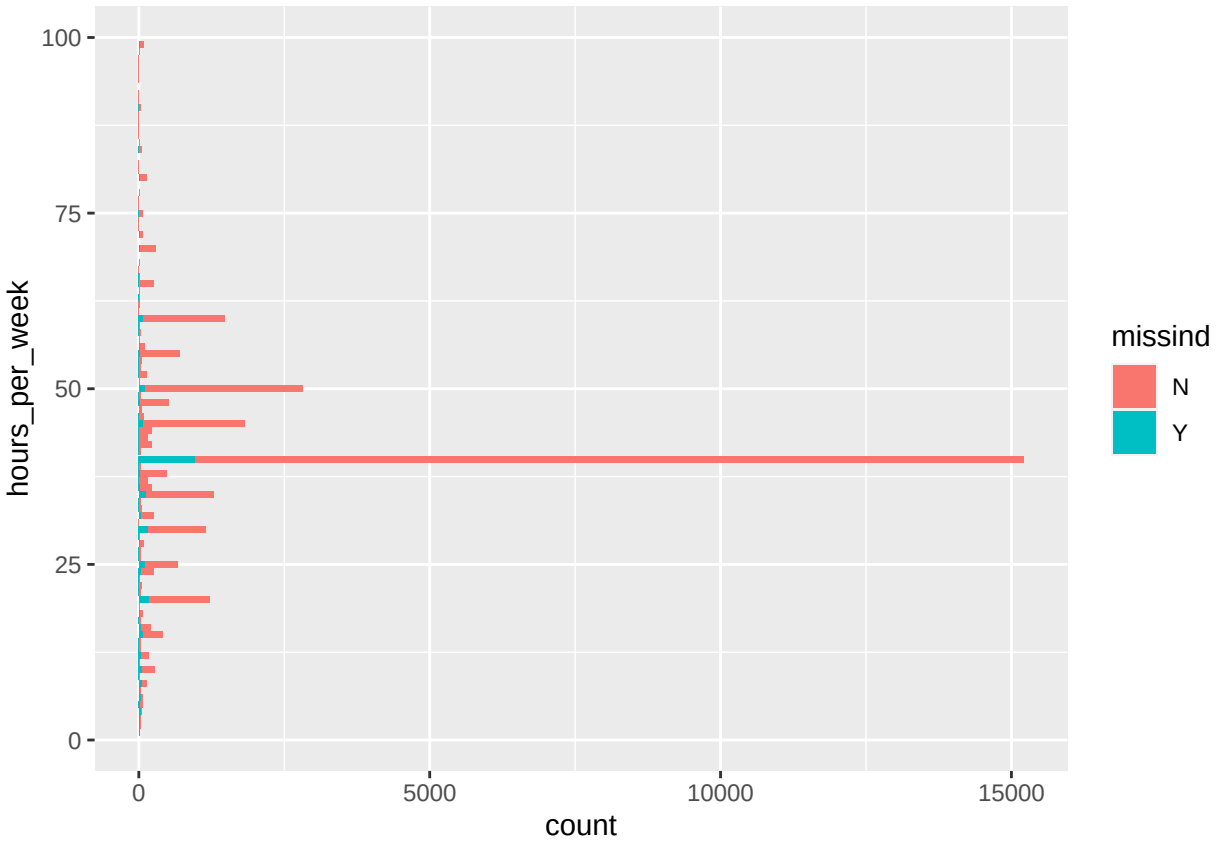

education
Some−college
Prof−school
Preschool
Masters
HS−grad
Doctorate
Bachelors
Assoc−voc
Assoc−acdm
9th
7th−8th
5th−6th
1st−4th
12th
11th
10th
0
2500
5000
7500
10000
count
missind
N
Y

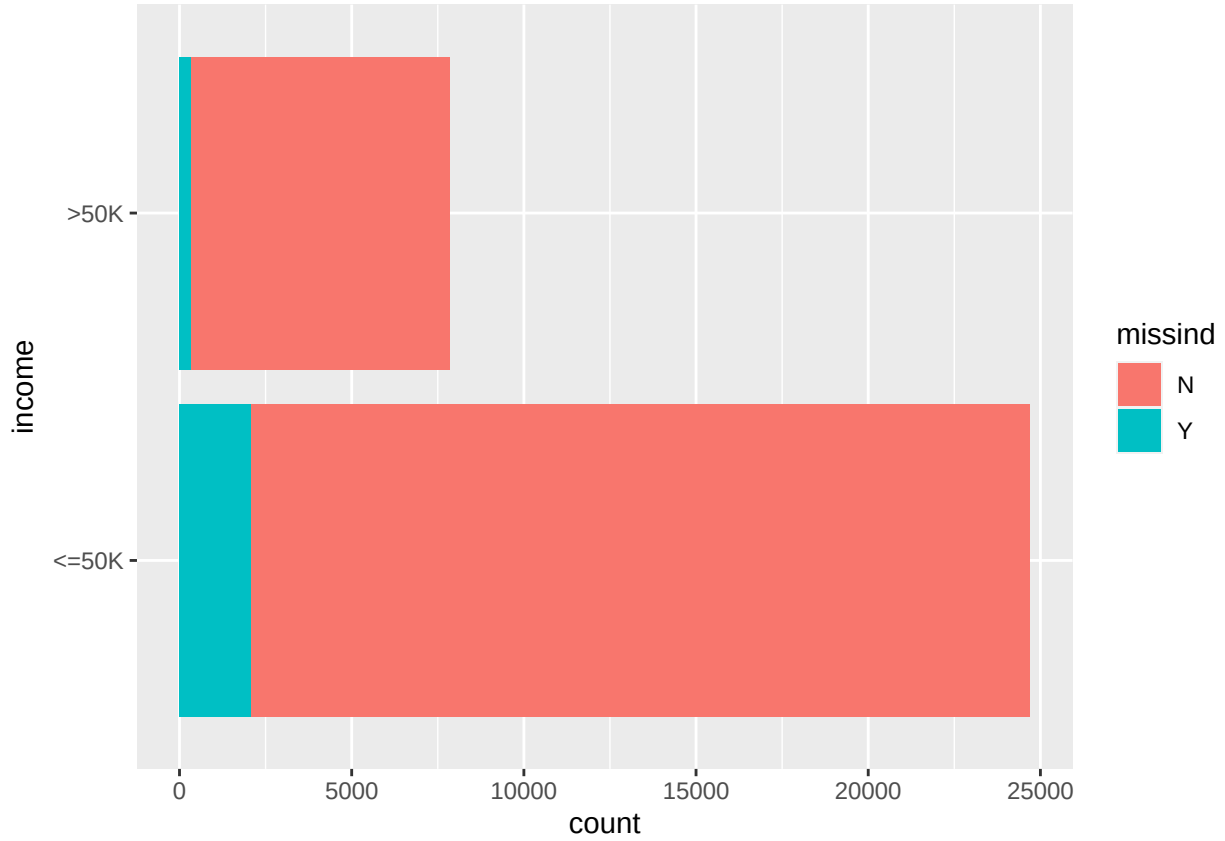

marital_status
Widowed
Separated
Never−married
Married−spouse−absent
Married−civ−spouse
Married−AF−spouse
Divorced
0
5000
10000
15000
count
missind
N
Y


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.

Remove rows with missing values - THIS IS OPTIONAL!!

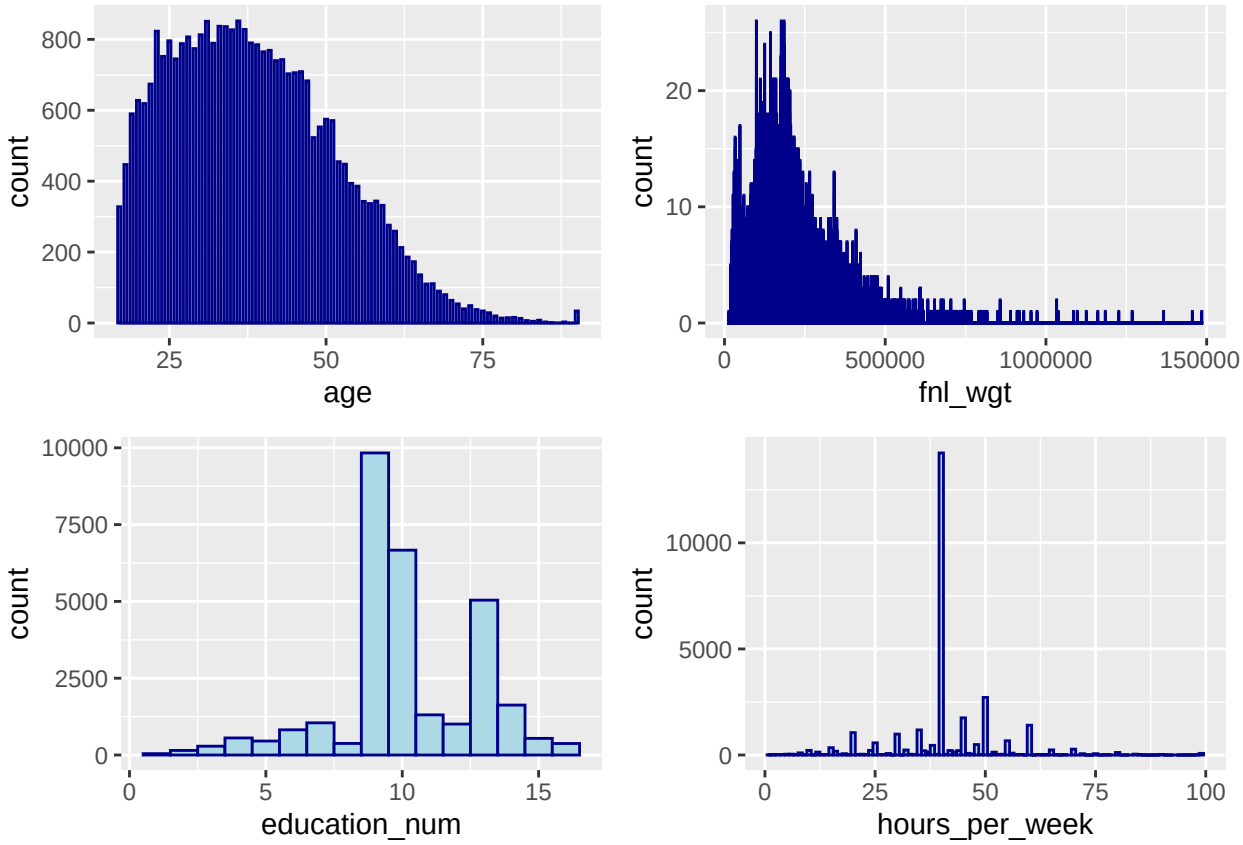
```
## Now the dimension of the dataset becomes:
```

```
## [1] 30139    16
```

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

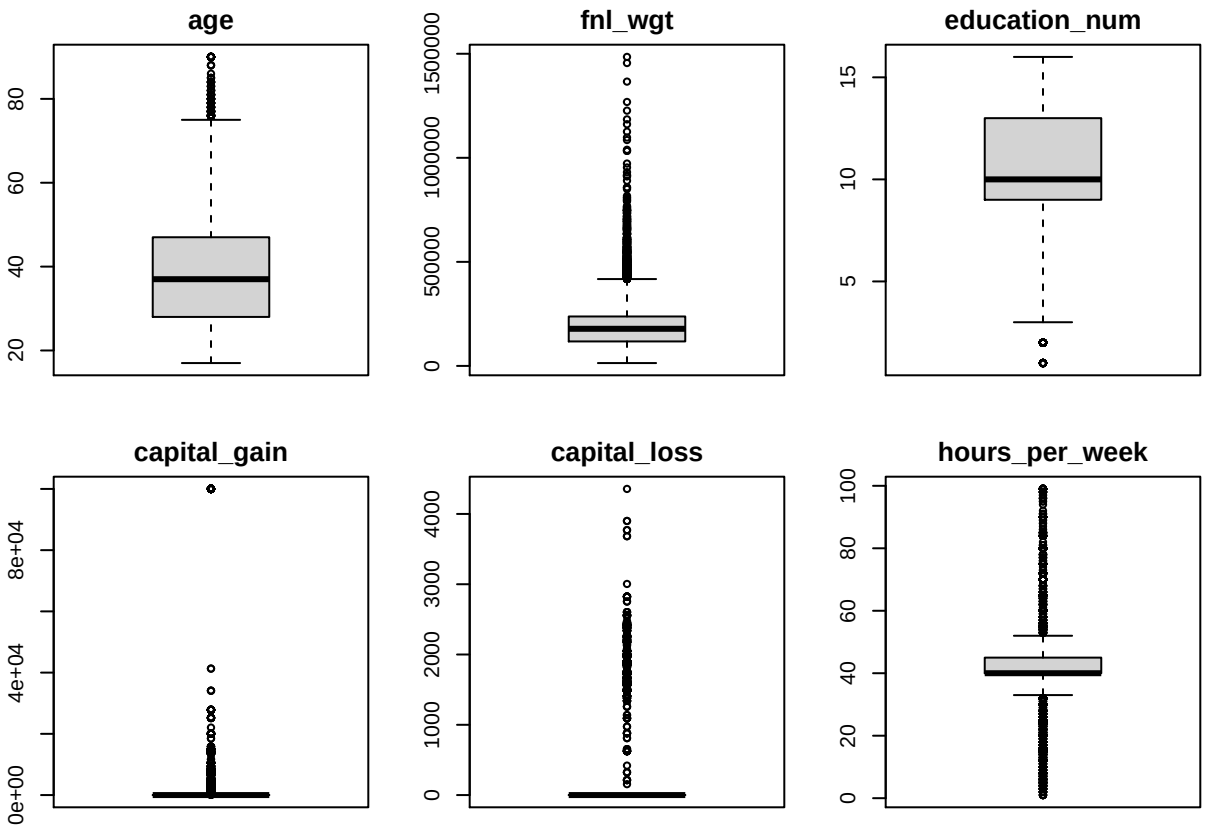
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
age	17.00	28.00	37.00	38.44	47.00	90.00
fnl_wgt	13769.00	117627.50	178417.00	189795.03	237604.50	1484705.00
education_num	1.00	9.00	10.00	10.12	13.00	16.00
capital_gain	0.00	0.00	0.00	1092.84	0.00	99999.00
capital_loss	0.00	0.00	0.00	88.44	0.00	4356.00
hours_per_week	1.00	40.00	40.00	40.93	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.



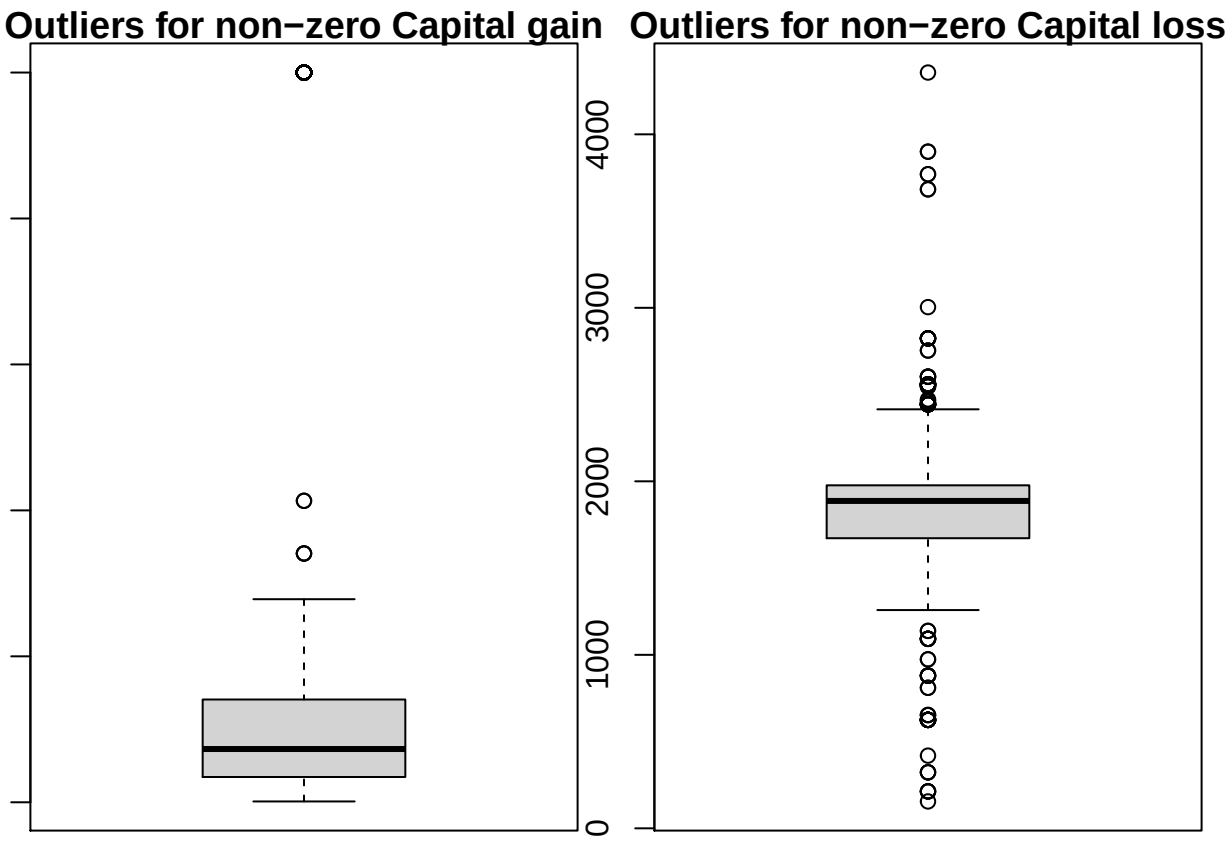
Both age and fnl_wght variables are skewed to the right, where log-normal transformation could help if modeling techniques with normality assumptions were to be used. The working hours per week variable is heavy-tailed distributed, meaning there are still quite a number of people working extremely small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

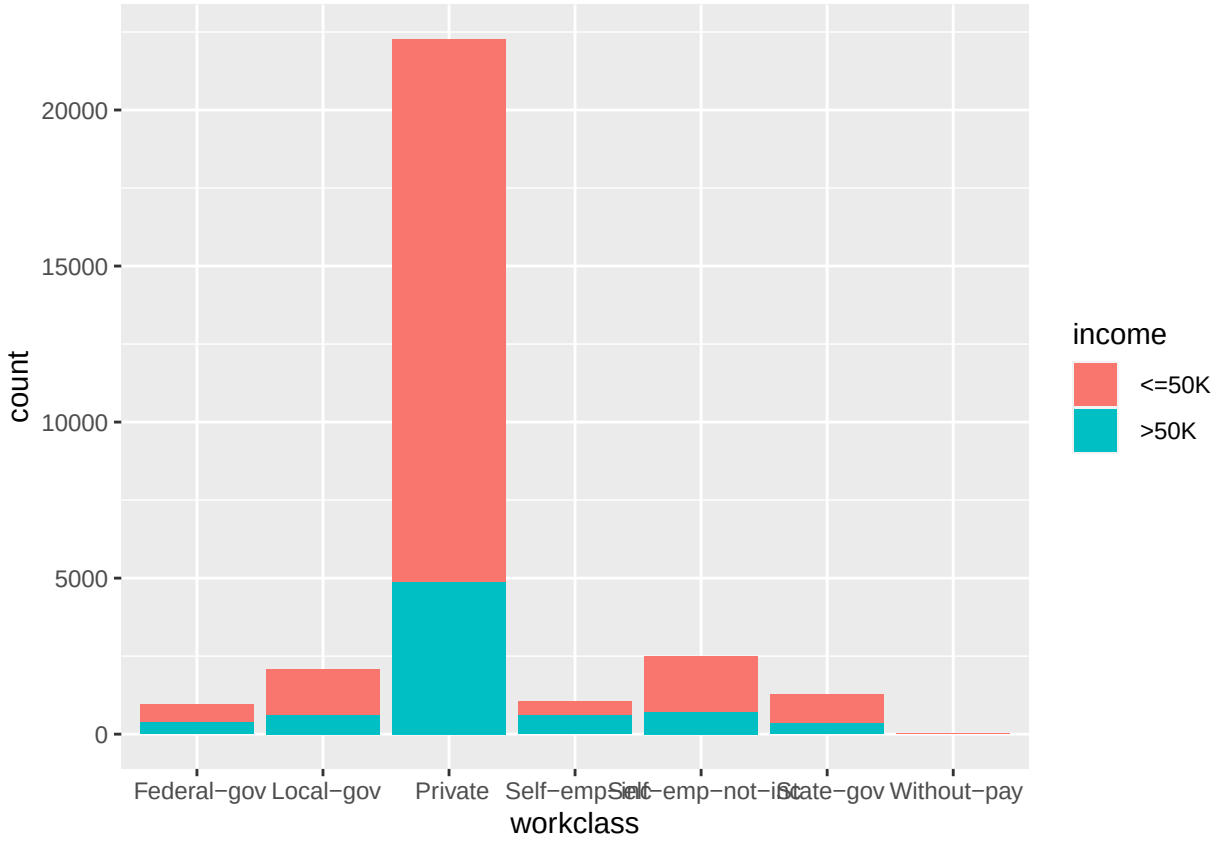


We can see there are still outliers even excluding zeros for capital gain and capital loss variables.

Distributions of categorical variables by target variable.

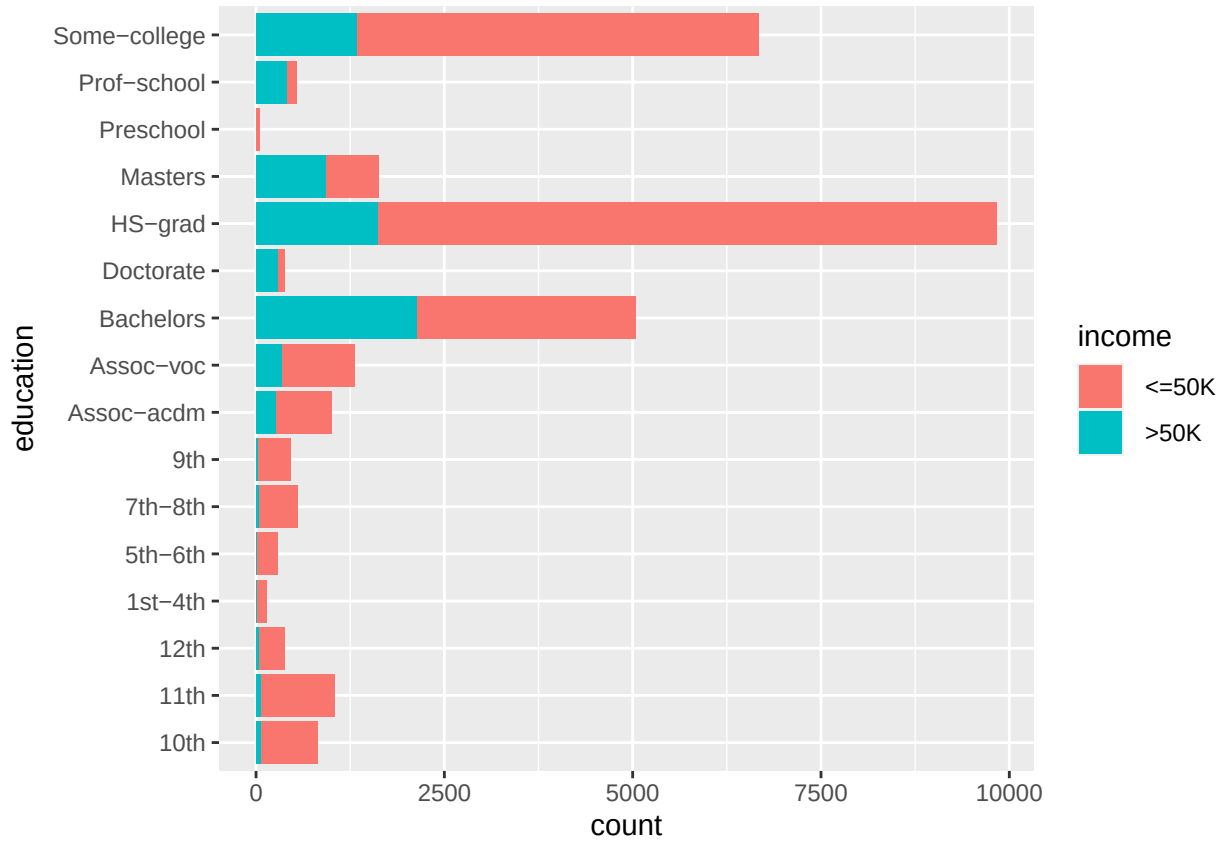
```
par(mar=c(1,1,1,1))

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



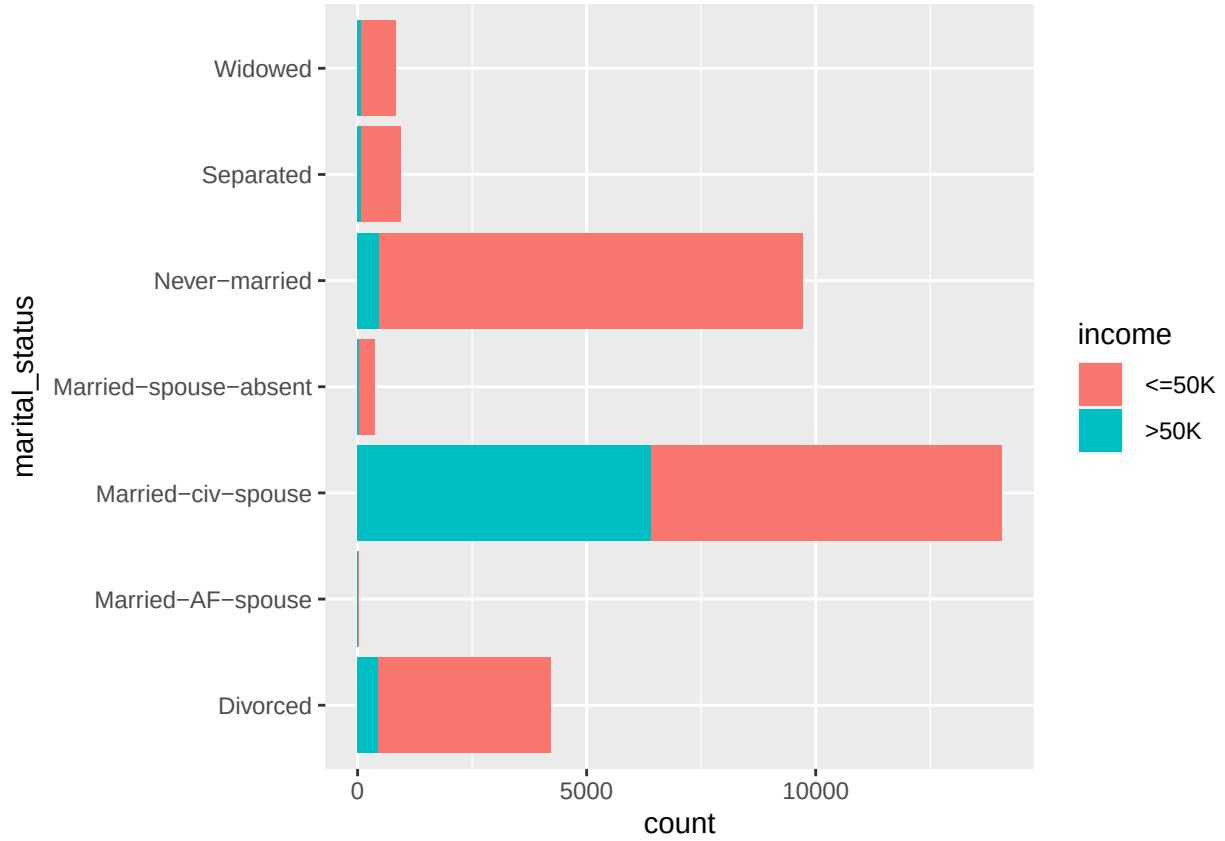
From the above bar chart we can see the majority of adults in the census were working in private sectors.

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



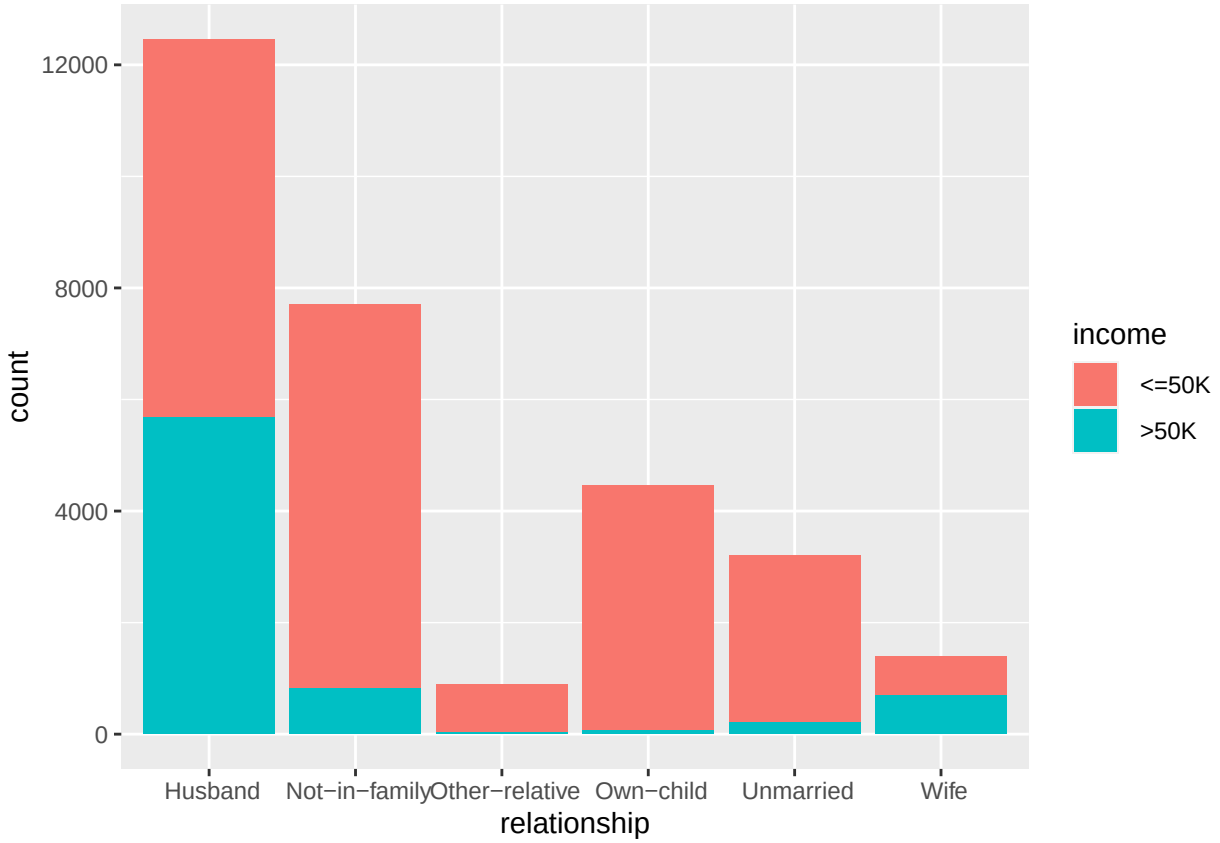
The majority of people earning less than $50K are high school graduates. The next largest education group is some college, and the third largest education group is Bachelors.

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



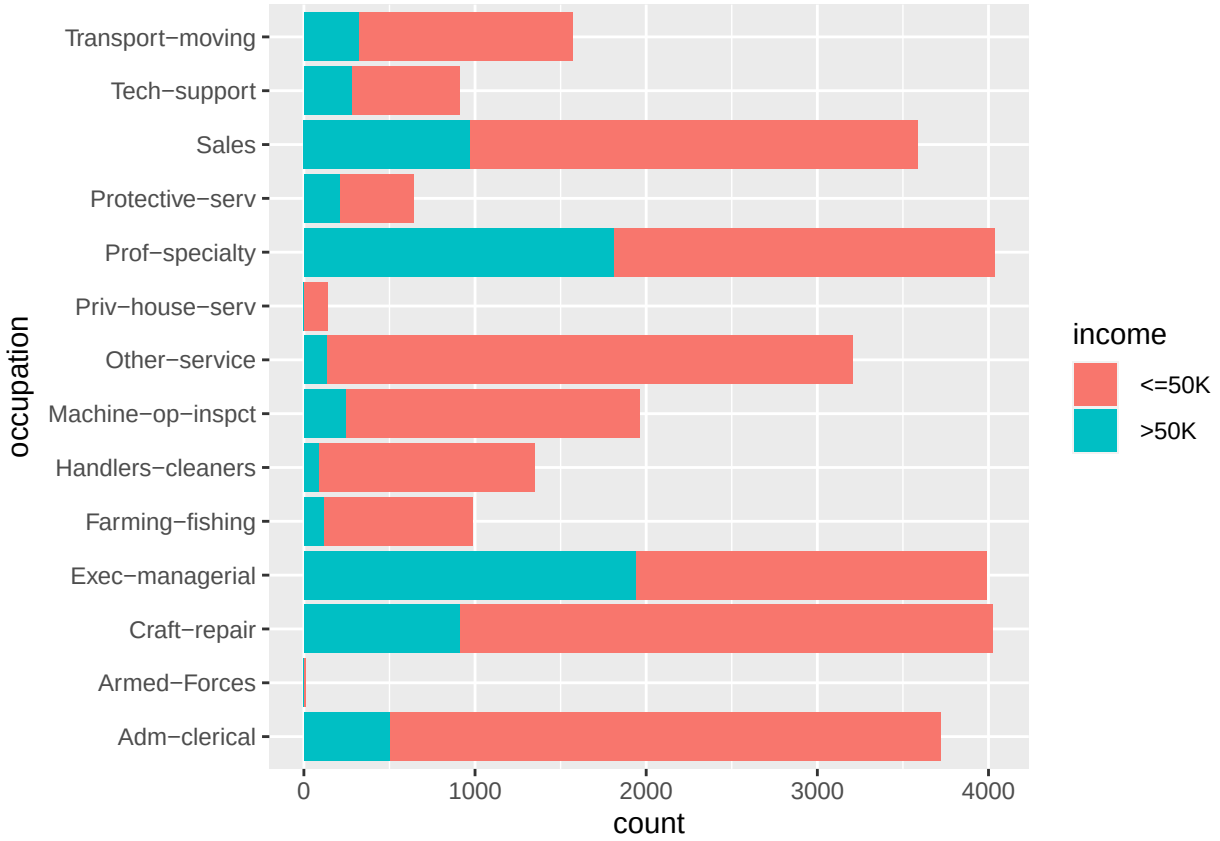
The majority of people surveyed are Married civ spouse, and in this marital status category, the income is roughly equally divided between <=50K or >50K. The second largest category is Never-married, with the majority of people earning <=50K.

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



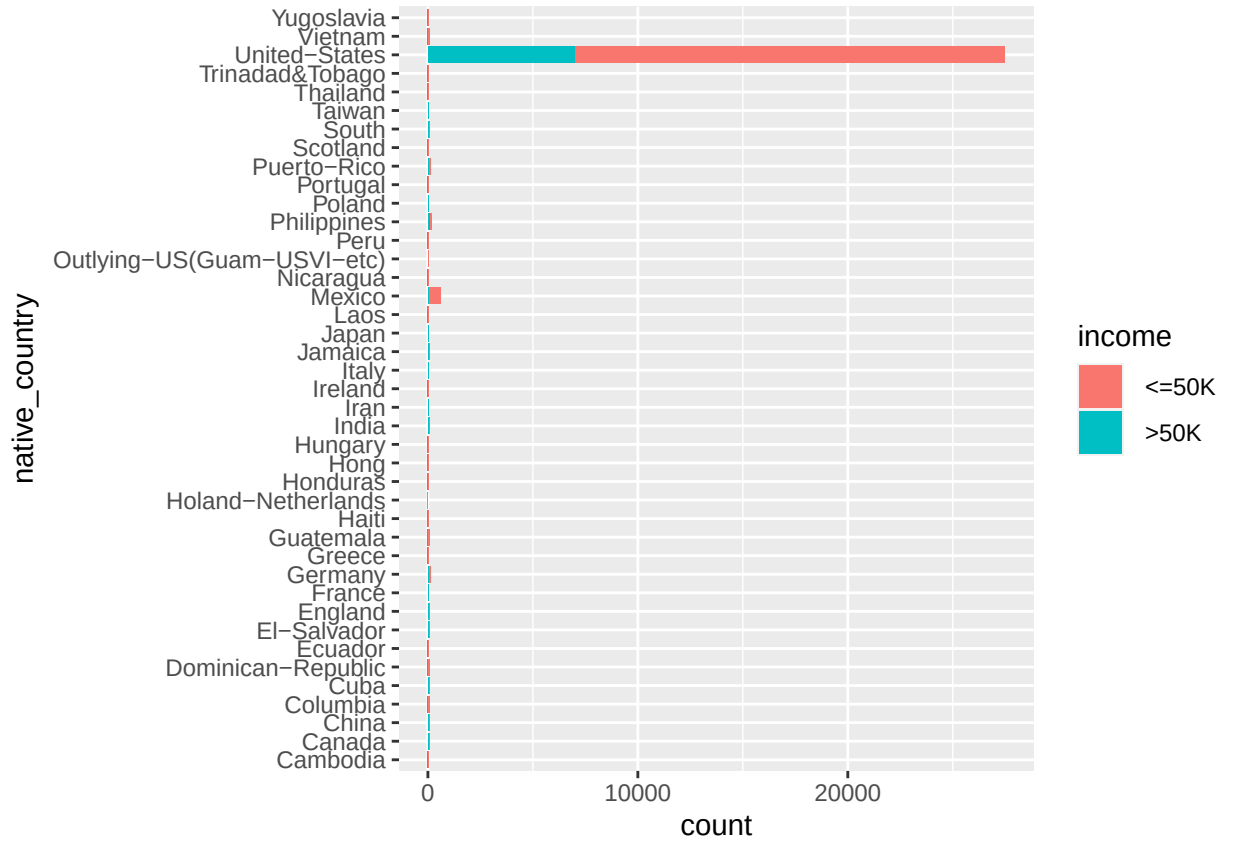
Most people surveyed in the census belong to the Husband category of relationships, with slightly more people earning less than or equal to 50K. However, in the Husband category, there is almost an even split between the 2 target income classes. Not-in-family is the second largest category for relationships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



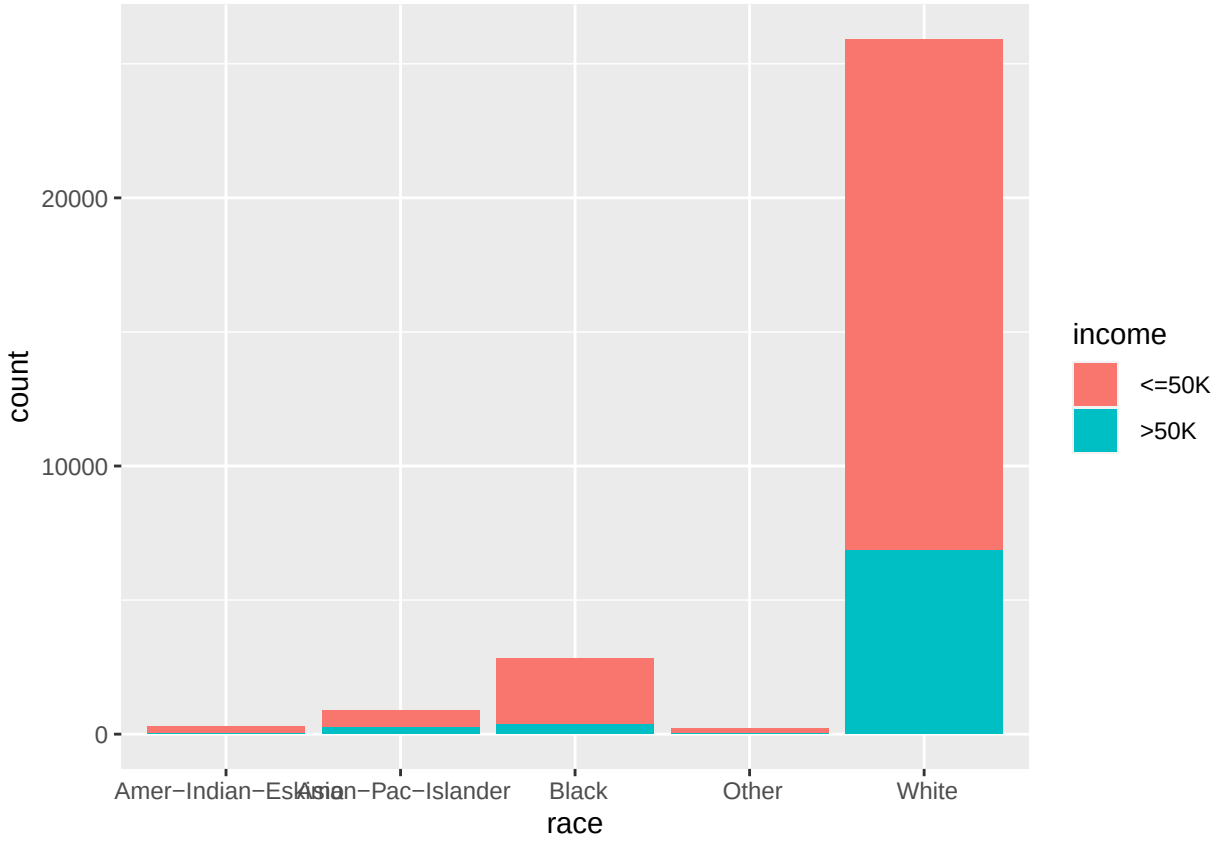
Most common occupations are Prof-specialty, Exec-managerial, Craft-repair, Sales, and Adm-clerical. For Exec-managerial, and Prof-specialty, there is an even number of people earning <=50K and >50K. For Craft-repair, Adm-clerical, and Sales, the majority of people earn <=50K.

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



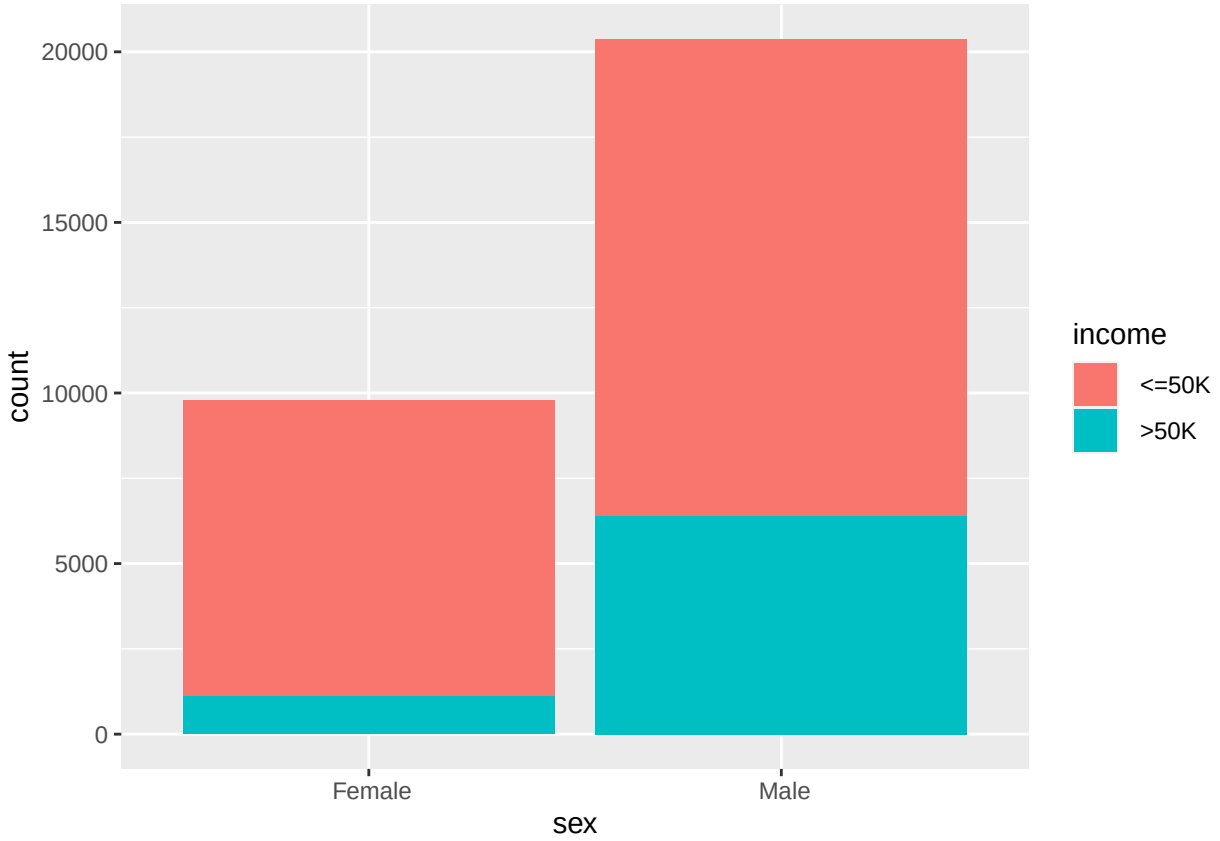
Most people surveyed come from the United States. This makes sense as the census was conducted in the US. Other than the United States, the second highest number of people come from Mexico.

```
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()
```

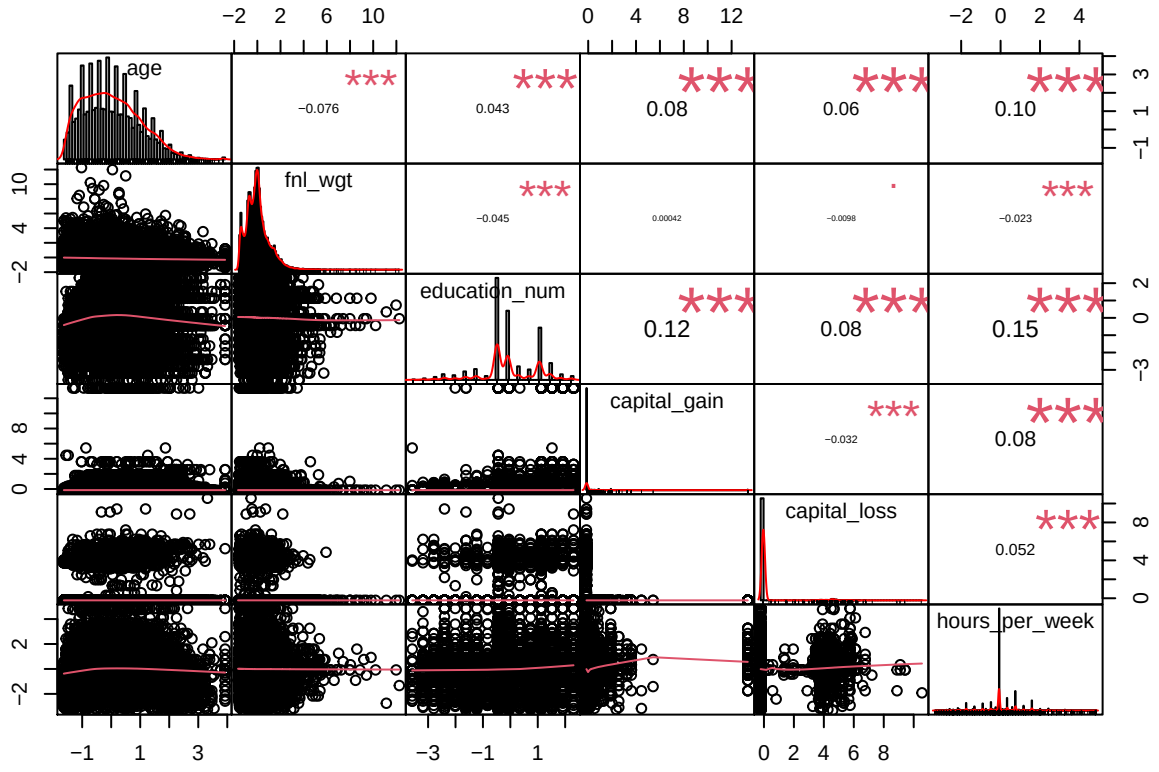


There are more than twice as many males surveyed in this census compared to females.

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)
```



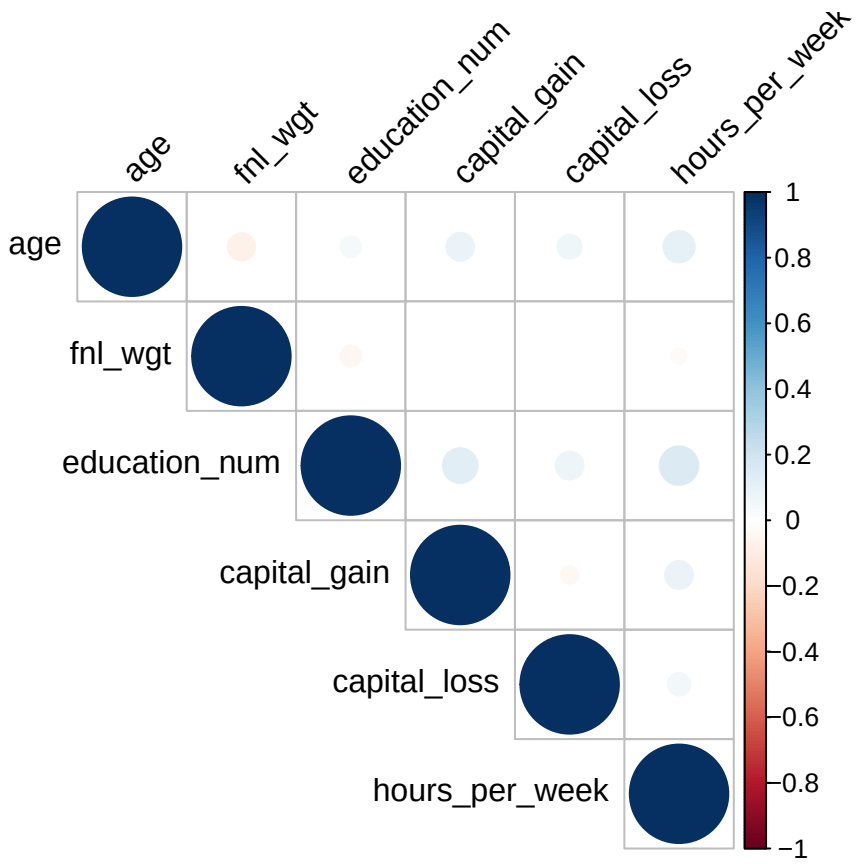
From the chart of the correlation matrix, we can see that while the magnitude of the correlations are small, all of them are statistically significant.

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

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			0.90	0.14	0.14	0.29	0.10	0.29	0.05
education									
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 variabls are essentially exact the same. So we decide to remove the educationnum variable.

```
#drop educationnum variable
X=subset(X,select = -education_num)
```

2. Re-group Native countries.

There are 41 levels, namely 41 different countries, in the dataset. Since too many levels for a categorical variable could lead to overfitting, we decide to regroup the native countries into regions.

```
##
##  Amer-Indian-Eskimo  Asian-Pac-Islander              Black              Other
##                   2                  68                  0                  0
##              White
##                   1
```

Since the race of almost all records with Native country "South" is "Asian-Pac-Islander", we think the country "South" is very likely to be South Korea. So we decided to group the country "South" into Asia_East

```
Asia_East <- c(" Cambodia", " China", " Hong", " Laos", " Thailand",
               " Japan", " Taiwan", " Vietnam", " South", " Philippines")

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

```
Central_America <- c(" Cuba", " Guatemala", " Jamaica", " Nicaragua",
                     " Puerto-Rico",  " Dominican-Republic", " El-Salvador",
                     " Haiti", " Honduras",  " Trinadad&Tobago")

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

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

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

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


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",                    "O

#convert native_region to a factor
#7 regions now
X$native_region=as.factor(X$native_region)
```

3. Re-group Capital Gain and Capital Loss.

```
## the proportion of zeros in Capital Gain is 91.57902%
## the proportion of zeros in Capital Loss is 95.26527%
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     114    3464    7298   12978   14084   99999
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     155    1672    1887    1868    1977    4356
```

Based the boxplots and summary of Captital Gain and Capital Loss variables in the Data Description section, and the situation that over 90% of the two variables are 0, we decided to categorize these two variables into groups in the following way: - the first group is for the zeros; - the other groups based on quantiles of non-zeros. More specifically, if a value is larger than zero and lower than the 1st quantile, it's grouped as "Low". - The values between 1st and 3rd quantitles are grouped into "Medium" and those higher than the 3rd quantile are categorized as "High".

```
X=mutate(X, cap_gain=ifelse(X$capital_gain==0, "Zero",
                     ifelse(X$capital_gain>0 & X$capital_gain<3464,"Low",
                     ifelse(X$capital_gain>=3464  &      X$capital_gain<14084,"Medium","High"))),

         cap_loss=ifelse(X$capital_loss==0,"Zero",
                  ifelse(X$capital_loss>0  & X$capital_loss<1672,"Low",
                  ifelse(X$capital_loss>=1672 & X$capital_loss<1977,"Medium","High")))

           )

X$cap_gain=as.factor(X$cap_gain)
```

```
X$cap_loss=as.factor(X$cap_loss)
```

4. Hours per week

From the boxplot and histogram in the Data Description section before, we've known there are large number of outliers in the hours_per_week variable.

So 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 hours";
- if a value is between the 1st and 3rd quantitle (40 to 45), it's called "40 to 45 hours";
- if the value is higher than 3rd quantile but lower than 50 , it's called "45 to 50 hours";
- if the value is between 50 and 60, it's called "50 to 60 hours";
- if the value is between 60 and 70, it's called "60 to 70 hours";
- if the value is between 70 and 80, it's called "70 to 80 hours";
- if the value is more than 80, it's called "greater than 80 hours";

```
summary(X$hours_per_week)
```

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

```
X=mutate(X,
         weekly_hours=ifelse(X$hours_per_week<40,"less than 40 hrs",
                         ifelse(X$hours_per_week<=45, "40 to 45 hrs",
                         ifelse(X$hours_per_week<=50, "45 to 50 hrs",
                         ifelse(X$hours_per_week<=60, "50 to 60 hrs",
                         ifelse(X$hours_per_week<=70, "60 to 70 hrs",
                         ifelse(X$hours_per_week<=80, "70 to 80 hours",
                                "greater than 80 hrs")))))))


X$weekly_hours=as.factor(X$weekly_hours)

table4=as.matrix(table(X$weekly_hours))
print(xtable(table4, caption = "Frequency table of weekly_hours" ) )
```

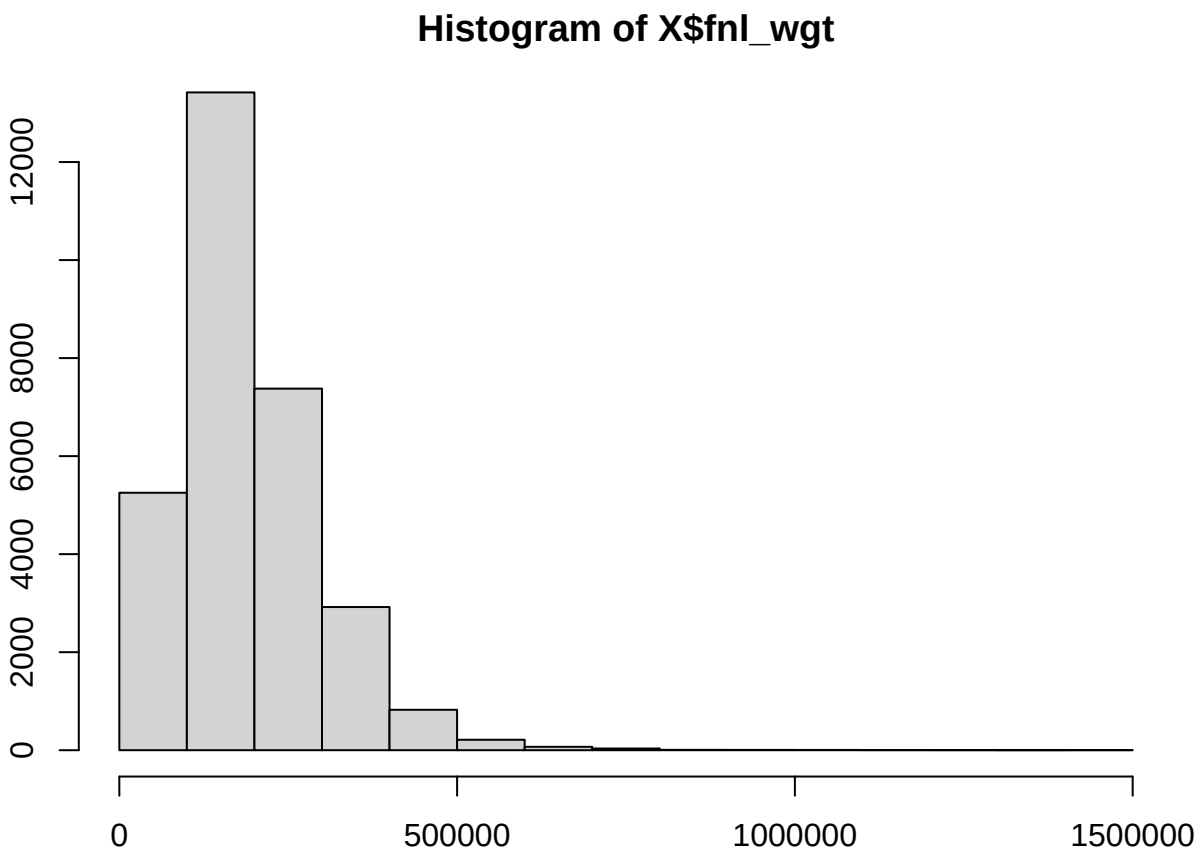
	x
40 to 45 hrs	16593
45 to 50 hrs	3364
50 to 60 hrs	2422
60 to 70 hrs	592
70 to 80 hours	265
greater than 80 hrs	195
less than 40 hrs	6708

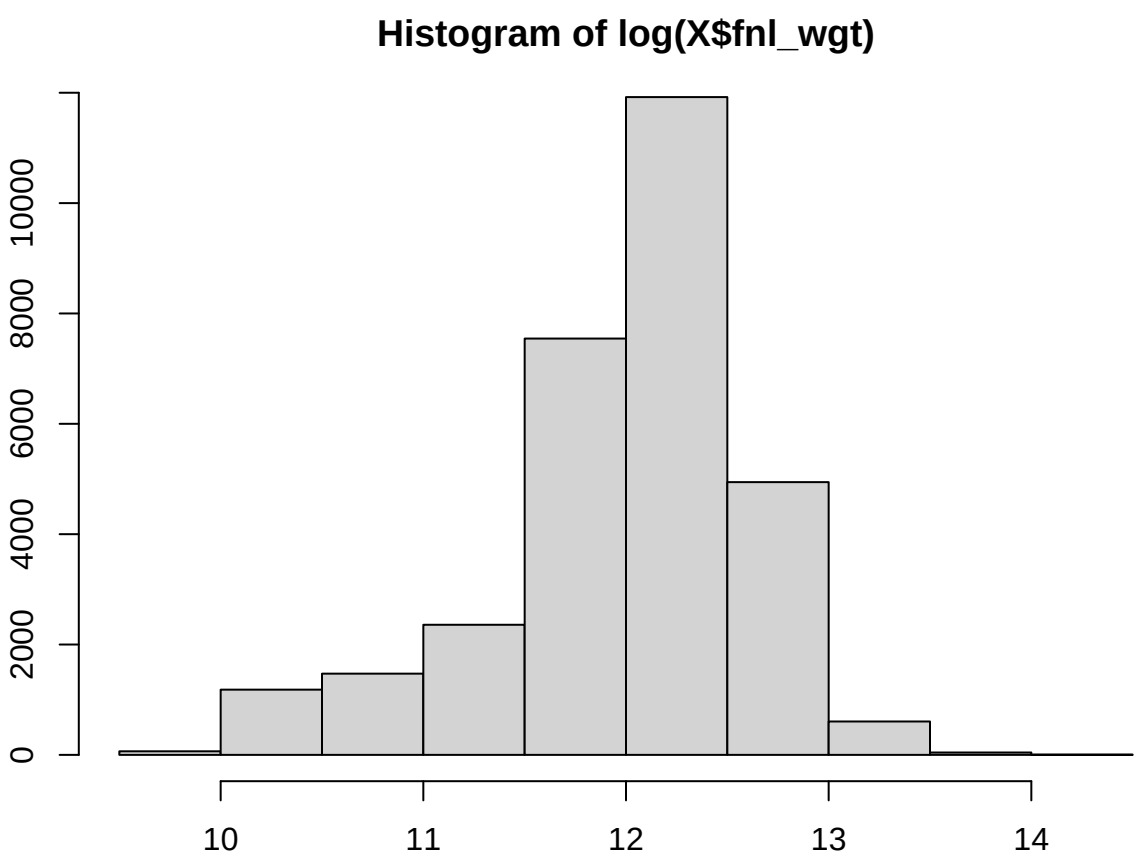
5. Drop empty level of work class.

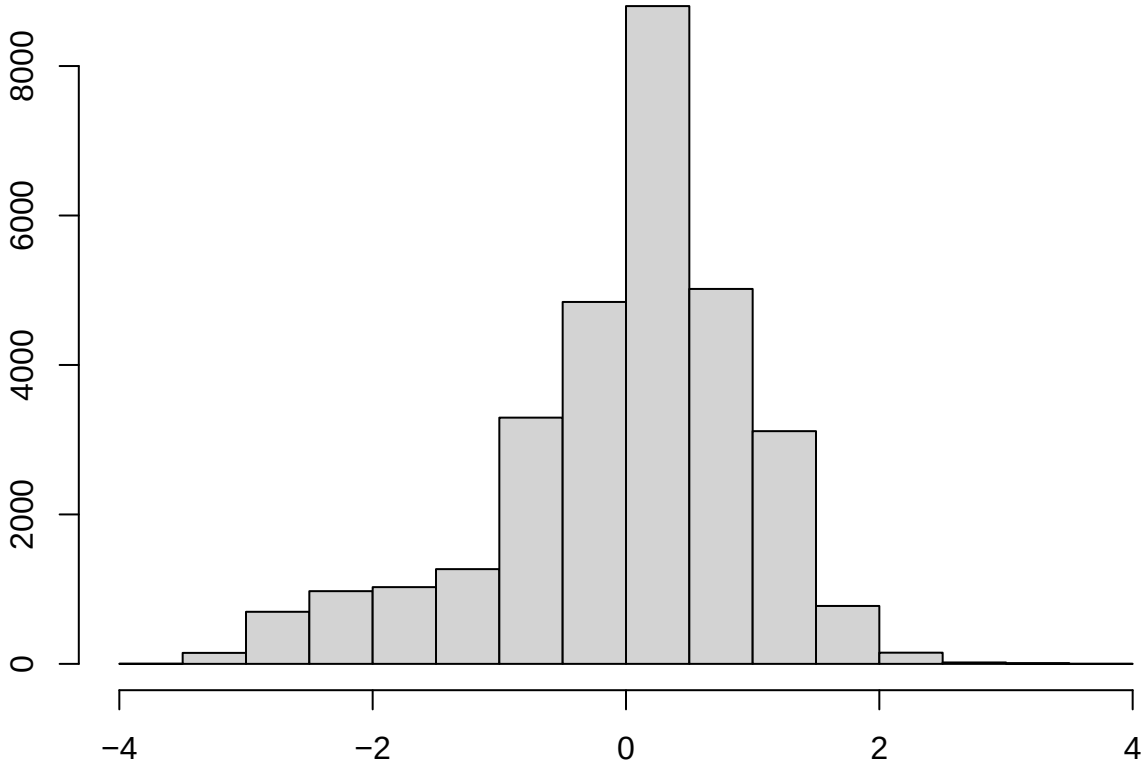
We noticed one of the levels of the work class variable does not have any records in it, so we decide to drop that level to avoid potential issues caused by empty cells.

	x
Federal-gov	943
Local-gov	2067
Never-worked	0
Private	22264
Self-emp-inc	1074
Self-emp-not-inc	2498
State-gov	1279
Without-pay	14

6. Log transformation and standardization of fnl_wgt - OPTIONAL




Histogram of log(X$fnl_wgt)
0
2000
4000
6000
8000
10000
10
11
12
13
14



From the above histograms, we can see that the fnl_wgt generally follows a log-normal distribution and the values are generally large. So we decide to perform a log transformation and then a standardization on it, after which the variable is generally normally distributed with small values.

```
X=mutate(X,
         fnlwgt_logstand=scale(log(X$fnl_wgt)))
```

6. Age standardization.

```
X=mutate(X,
        age_stand=scale(X$age)
         )
```

To make age in the same range of the standardized fnl_wgt, we decide to standardize age as well.

7. Group marital status.

```
X=mutate(X,
         marital_status_group=ifelse(marital_status %in% c(" Married-AF-spouse"," Married-civ-spouse","
         )
X$marital_status_group=as.factor(X$marital_status_group)
```

Drop the variables that would not be used for building predictive models.

```
X=subset(X, select = -c(capital_gain, capital_loss,hours_per_week,native_country,marital_status))
cat("The current structure of the dataset is:")
```

```
## The current structure of the dataset is:
```

```
str(X)
```

```
## 'data.frame':    30139 obs. of  16 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 ...
##  $ 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 ...
##  $ income              : Factor w/ 2 levels " <=50K"," >50K": 1 1 1 1 1 1 1 2 2 2 ...
##  $ native_region       : Factor w/ 7 levels " Central-America",..: 7 7 7 7 1 7 1 7 7 7 ...
##  $ cap_gain            : Factor w/ 4 levels "High","Low","Medium",..: 2 4 4 4 4 4 4 4 1 3 ...
##  $ cap_loss            : Factor w/ 4 levels "High","Low","Medium",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ weekly_hours        : Factor w/ 7 levels "40 to 45 hrs",..: 1 7 1 1 1 1 7 1 2 1 ...
##  $ fnlwgt_logstand     : num [1:30139, 1] -1.151 -1.036 0.472 0.606 1.186 ...
##  $ age_stand           : num [1:30139, 1] 0.0425 0.8802 -0.0336 1.1087 -0.7952 ...
##  $ marital_status_group: Factor w/ 5 levels " Divorced"," Never-married",..: 2 5 1 5 5 5 5 5 2 5 ...
```

Modeling

Recode target variable.

First, for the simplicity in coding, let's re-code 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.

```
#
X$income=ifelse(X$income==" <=50K","N","Y")
X$income=as.factor(X$income)
cat("Now the levels of the target variable are:")
```

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

```
summary(X$income)
```

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

1. Train-test split.

Before training supervised learning models, we first split the dataset into training and testing sets.

```
## The number of the two levels of the target variable are:
```

```
##     N     Y
## 18107  6005
```

2. Down-sample the majority group to balance the training data.

We've seen from Data Exploratory analysis that there are about 3 times of people making no more than 50K annually (the majority group) than those who make over 50K a year (the minority group), so the dataset is imbalanced, which could lead to over-fitting issues. So we decide to down-sample the majority group. More specifically, we will randomly select 34% of the records that have annual income no more than 50K, and then combine them with all the records in the minority group.

- First, randomly select 34% of records from the majority group.

```
## The number of the two levels of the target variable at the randomly selected sample are:
```

```
##    N    Y
## 6156    0
```

- Then, combine the randomly down sampled majority group with the original minority group.

```
## The number of the two levels of the target variable at the combined balanced sample are:
```

```
##    N    Y
## 6156 6005
```

- Next, randomly shuffle the rows so that not all "N" records are on the top and "Y"s are in the end.

```
## The number of the two levels of the target variable at the final training dataset are not changed!
```

```
##    N    Y
## 6156 6005
```

Create different sets of features to train the models.

Based on different ways of dealing with the variables, we have different sets of features to train supervised learning models.

Training scenario 1: apply Random Forest, KNN and Logistic Regresstion to transformed variables. Random Forest Model.

We will use 5-fold cross-validation for the re-sampling to tune the model parameters.

```
cv_5=trainControl(method = "cv", number = 5,
                  allowParallel = TRUE,
                  summaryFunction = twoClassSummary ,
                  classProbs = TRUE)
```

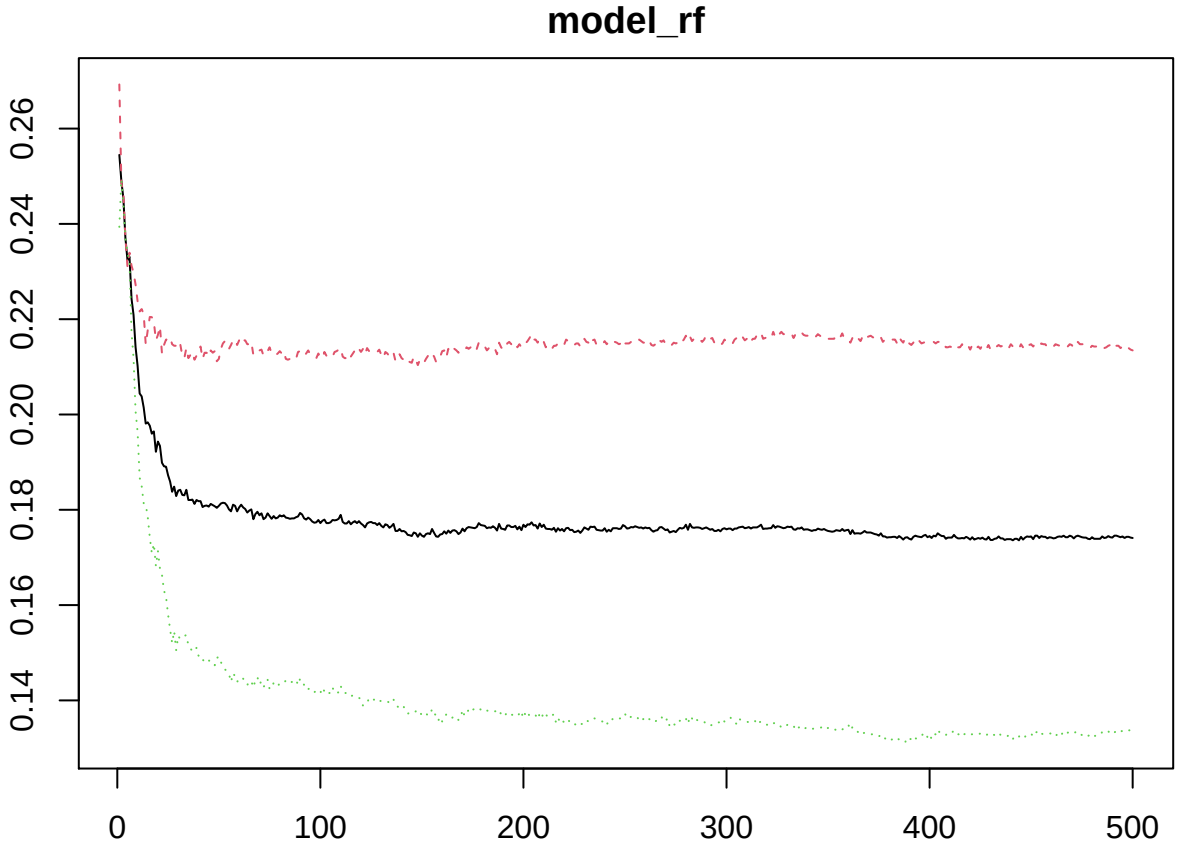
```
#random forest
library(randomForest)
set.seed(123)
t1_rf=proc.time()
model_rf <- randomForest(income~age_stand+workclass+fnlwgt_logstand+education+marital_status_group+occu
                          metric="ROC",
                         data = X_train_bal,
                         trControl=cv_5 )
t2_rf=proc.time()
cat("The computational time for training a random forest model is",(t2_rf-t1_rf)[3],"s.","\n")
```

```
## The computational time for training a random forest model is 7.17 s.
```

```
## The fitted random forest model:
```

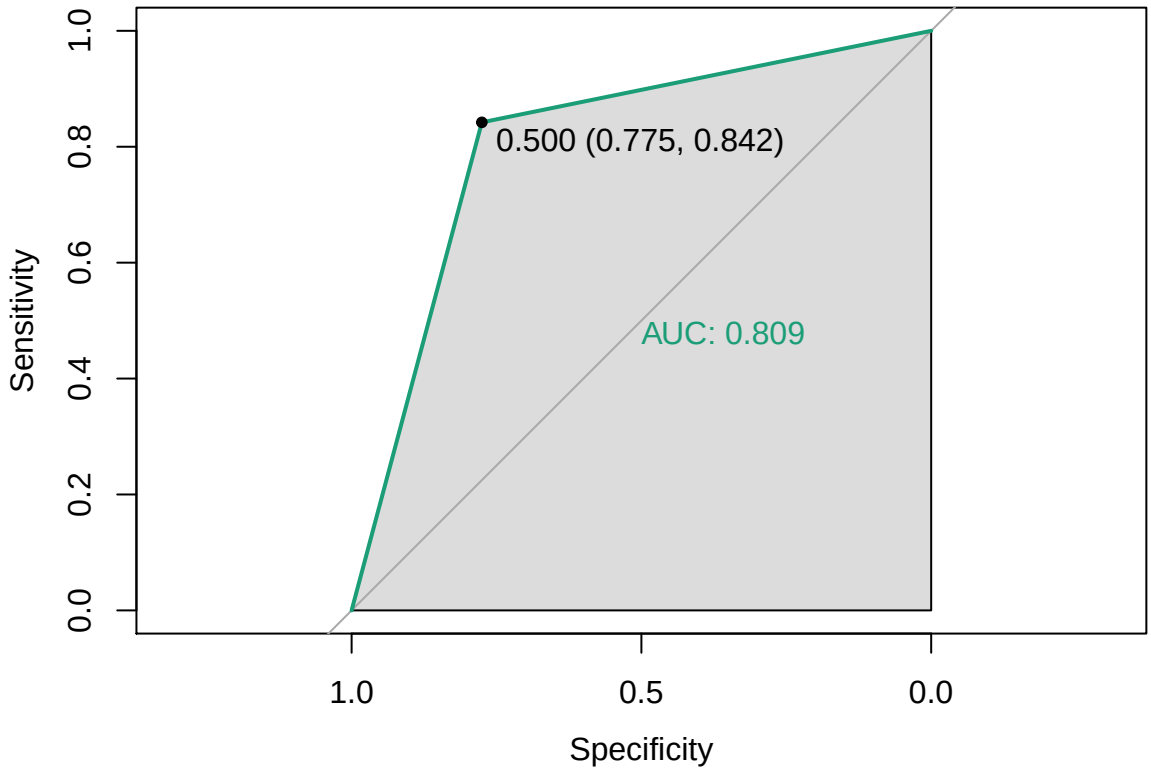
```
##
## Call:
##  randomForest(formula = income ~ age_stand + workclass + fnlwgt_logstand +      education + marital_
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 17.41%
## Confusion matrix:
##      N    Y class.error
## N 4842 1314   0.2134503
## Y  803 5202   0.1337219
```

```
## Importance of features based on the random forest model:
##                       MeanDecreaseGini
## age_stand                     780.51237
## workclass                     202.16609
## fnlwgt_logstand               555.73061
## education                     586.81450
## marital_status_group          671.41456
## occupation                    658.79789
## relationship                  820.14554
## race                           72.90622
## sex                           101.63749
## native_region                  63.96797
## cap_loss                       95.95992
## cap_gain                      342.74626
## weekly_hours                  284.87383
```




```
## The confusion matrix of random forest model on test data:
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 3504  237
##          Y 1022 1264
##
##                Accuracy : 0.7911
```

```
##                  95% CI : (0.7806, 0.8013)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : 1.139e-13
##
##                   Kappa : 0.5246
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8421
##             Specificity : 0.7742
##          Pos Pred Value : 0.5529
##          Neg Pred Value : 0.9366
##              Prevalence : 0.2490
##          Detection Rate : 0.2097
##    Detection Prevalence : 0.3793
##       Balanced Accuracy : 0.8081
##
##        'Positive' Class : Y
##
```



KNN Model.

```
#knn
set.seed(123)
t1_knn=proc.time()
model_knn=train(income ~ age_stand+workclass+fnlwgt_logstand+education+marital_status_group+occupation+
                data = X_train_bal,
```

```
                 metric="ROC",
                trControl=cv_5, method="knn")
t2_knn=proc.time()
cat("The computational time of training a knn model is",(t2_knn-t1_knn)[3],"s.","\n")
```

```
## The computational time of training a knn model is 88.33 s.
```

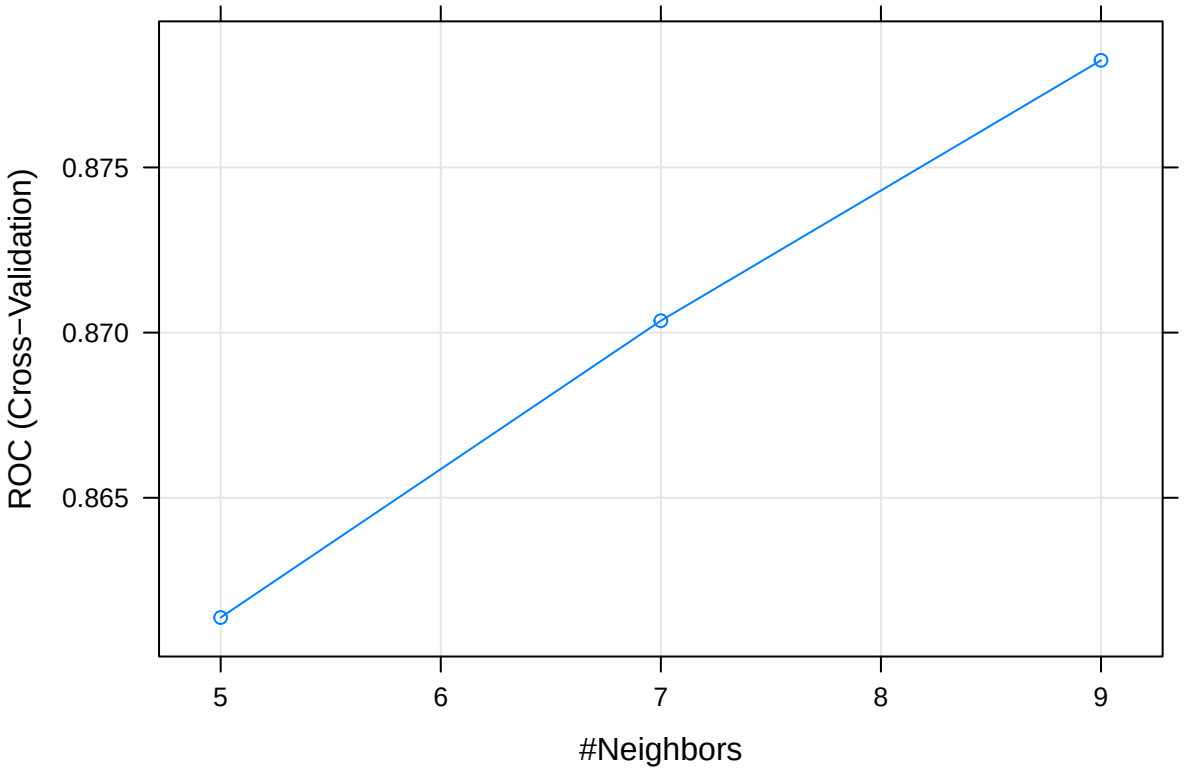
```
cat("The fitted knn model:","\n")
```

```
## The fitted knn model:
```

```
model_knn
```

```
## k-Nearest Neighbors
##
## 12161 samples
##    13 predictor
##     2 classes: 'N', 'Y'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 9729, 9729, 9728, 9729, 9729
## Resampling results across tuning parameters:
##
##   k  ROC        Sens       Spec
##   5  0.8613749  0.7621836  0.8231474
##   7  0.8703603  0.7600706  0.8353039
##   9  0.8782422  0.7668922  0.8404663
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was k = 9.
```

```
plot(model_knn)
```

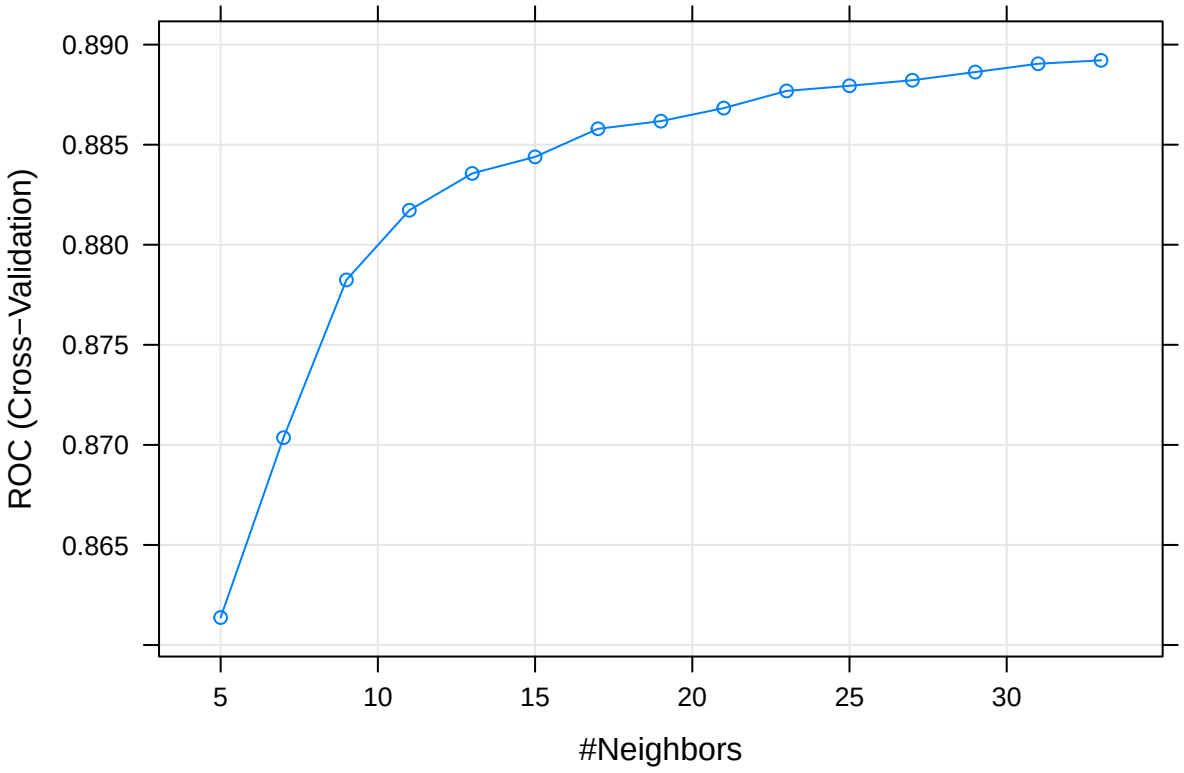


From the above plot of accuracy and number of neighbors, we can see that the accuracy still increases after training on the default number of ks. So we decide to increasing the number of k values to 15 to try to locate a better model.

```
set.seed(123)
t1_knn=proc.time()
model_knn=train(income ~ age_stand+workclass+fnlwgt_logstand+education+marital_status_group+occupation+
                data = X_train_bal,
                metric="ROC",
                trControl=cv_5,
                method="knn",
                tuneLength = 15)
t2_knn=proc.time()
cat("The computational time of training a knn model is",(t2_knn-t1_knn)[3],"s.","\n")
```

```
## The computational time of training a knn model is 462.86 s.
```

```
plot(model_knn)
```



Now we can see the accuracy does not increase much once the number of neighbors reaches about 15.

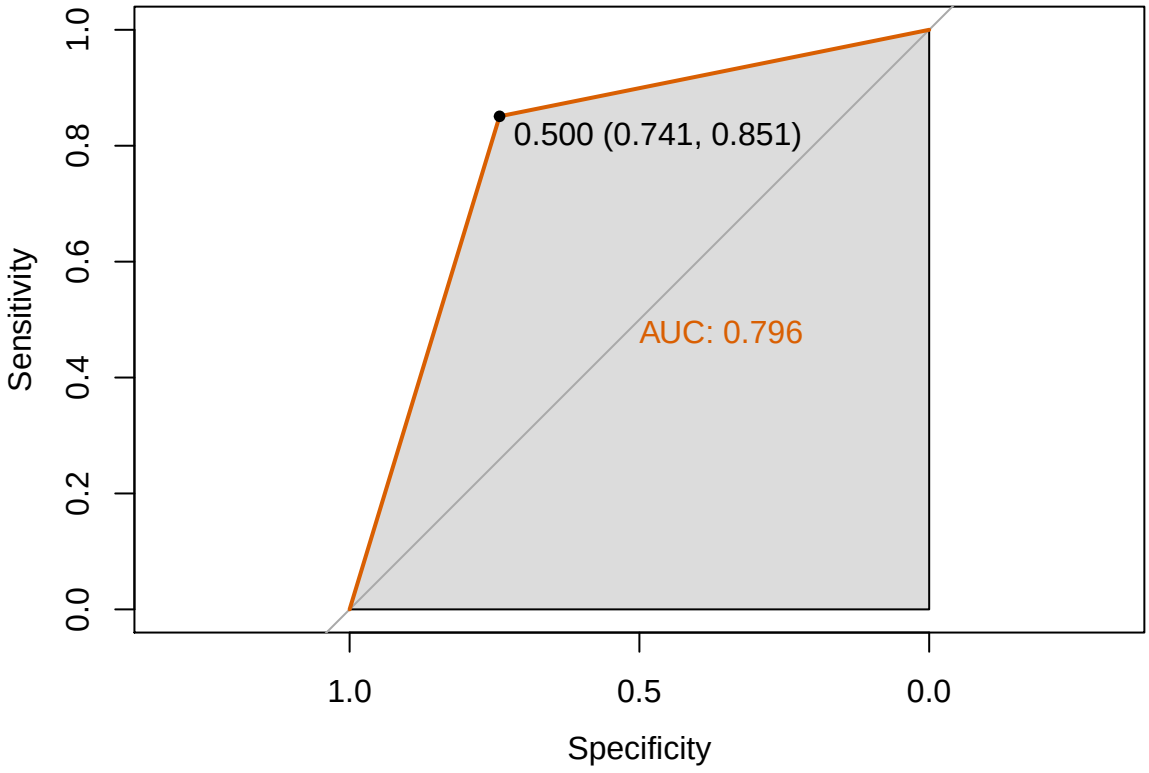
```
## The current fitted KNN models trying more tuning parameters:
```

```
## k-Nearest Neighbors
##
## 12161 samples
##    13 predictor
##     2 classes: 'N', 'Y'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 9729, 9729, 9728, 9729, 9729
## Resampling results across tuning parameters:
##
##   k   ROC        Sens       Spec
##    5  0.8613749  0.7623461  0.8231474
##    7  0.8703603  0.7600707  0.8353039
##    9  0.8782422  0.7668922  0.8404663
##   11  0.8817225  0.7616937  0.8459617
##   13  0.8835633  0.7628319  0.8459617
##   15  0.8843907  0.7597458  0.8504580
##   17  0.8857936  0.7589346  0.8544546
##   19  0.8861756  0.7540612  0.8551207
##   21  0.8868277  0.7548743  0.8564530
##   23  0.8876855  0.7530869  0.8559534
##   25  0.8879434  0.7511372  0.8566195
```

```
##   27  0.8882202  0.7501625  0.8569525
##   29  0.8886291  0.7514625  0.8607827
##   31  0.8890439  0.7498378  0.8591174
##   33  0.8892134  0.7483756  0.8574521
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was k = 33.
```

```
## The confusion matrix of knn model on test data:
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 3355  224
##          Y 1171 1277
##
##                Accuracy : 0.7685
##                  95% CI : (0.7577, 0.7791)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : 0.0007745
##
##                   Kappa : 0.4889
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8508
##             Specificity : 0.7413
##          Pos Pred Value : 0.5217
##          Neg Pred Value : 0.9374
##              Prevalence : 0.2490
##          Detection Rate : 0.2119
##    Detection Prevalence : 0.4062
##       Balanced Accuracy : 0.7960
##
##        'Positive' Class : Y
##
```



As there are mixed types of data in our dataset, we are supposed to use the gower distance metric in the KNN method. However, we are not able to specify "gower" in the train() function and the knngow() function which can run knn using the gower metric is not available to our personal version of R, so for future discussion, if time allows, please try to run the KNN method using gower distance metric on this data set.

Logistic Regression Model.

```
#logreg
set.seed(123)
t1_log=proc.time()
model_logreg=train(income ~age_stand+workclass+fnlwgt_logstand+education+marital_status_group+occupatio
                   data = X_train_bal,
                   metric="ROC",
                   trControl=cv_5, method="regLogistic")

t2_log=proc.time()
cat("The computational time of training a logistic regression on transformed variables is",(t2_log-t1_l
```

```
## The computational time of training a logistic regression on transformed variables is 33.55 s.
```

```
#cat("The fitted logistic regression model:","\n")
#model_logreg

#confusion matrix on test data
conf_logreg=confusionMatrix( predict(model_logreg, newdata = X_test),
                         reference=X_test$income,
                         positive="Y")
cat("The confusion matrix of logistic regression model:","\n")
```

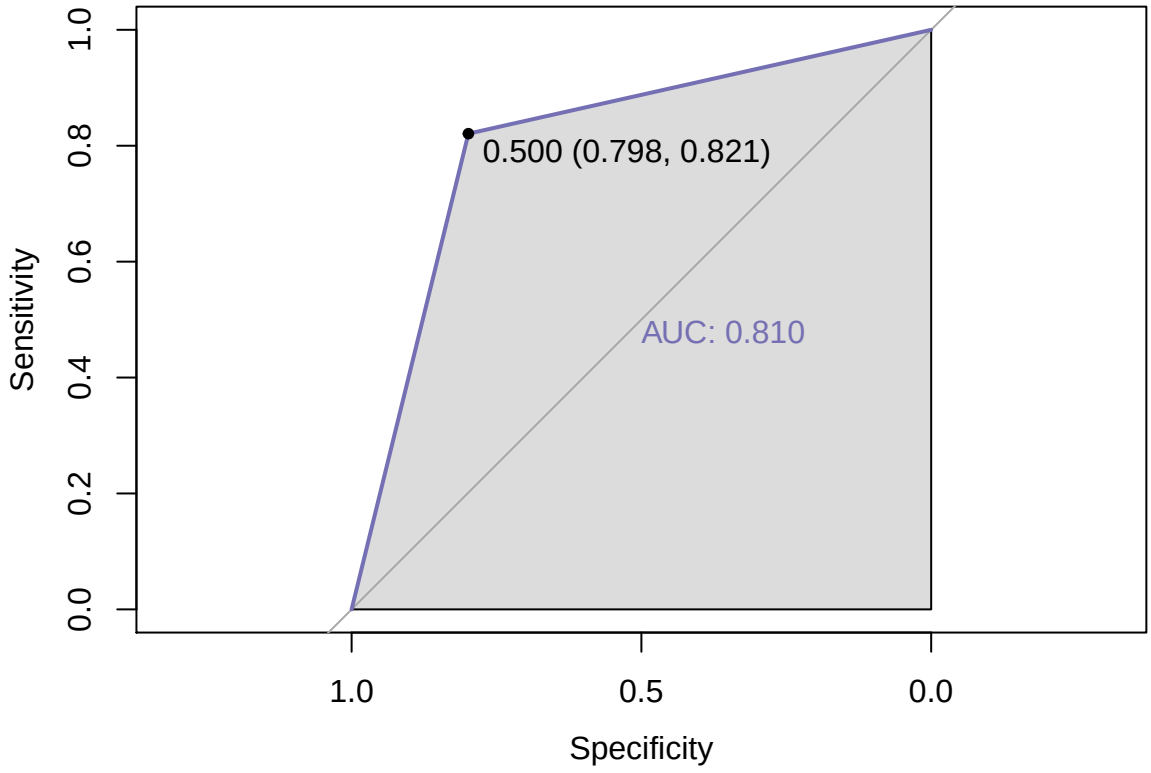
```
## The confusion matrix of logistic regression model:
conf_logreg
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 3614  269
##          Y  912 1232
##
##                Accuracy : 0.804
##                  95% CI : (0.7938, 0.814)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5417
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8208
##             Specificity : 0.7985
##          Pos Pred Value : 0.5746
##          Neg Pred Value : 0.9307
##              Prevalence : 0.2490
##          Detection Rate : 0.2044
##    Detection Prevalence : 0.3557
##       Balanced Accuracy : 0.8096
##
##        'Positive' Class : Y
##
```

```
#accuracy 80%, sens 82%, spec 80%, NIR 75%

#ROC & AUC
roc.logreg = roc(X_test$income,                          as.vector(ifelse(predict(model_logreg, newdata

auc.logreg = auc(roc.logreg)
```



Training scenario 2: Logistic Regression Model on unstandardized age and fnl_wgt.

```
#logreg
set.seed(123)
t1_log_2=proc.time()
model_logreg_ori=train(income ~age+workclass+fnl_wgt+education+marital_status_group+occupation+relations
                   data = X_train_bal,
                   metric="ROC",
                   trControl=cv_5,
                   method="regLogistic")
t2_log_2=proc.time()
cat("The computational time of training a logistic regression model using unstandardized variables is",
```

```
## The computational time of training a logistic regression model using unstandardized variables is 452
```

```
#cat("The fitted logistic regression model:","\n")
#model_logreg_ori

#confusion matrix on test data
conf_logreg_ori=confusionMatrix( predict(model_logreg_ori, newdata = X_test),
                         reference=X_test$income,
                         positive="Y")
cat("The confusion matrix of logistic regression model:","\n")
```

```
## The confusion matrix of logistic regression model:
```

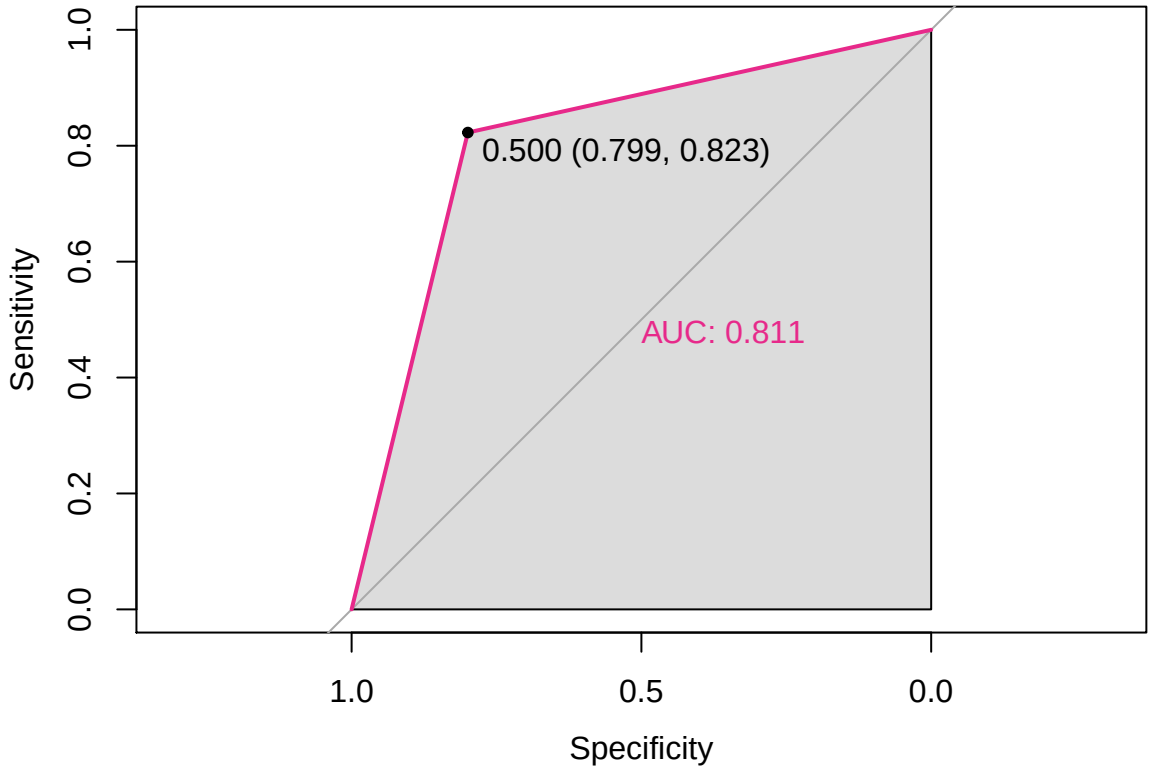
```
conf_logreg_ori
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 3618  266
##          Y  908 1235
##
##                Accuracy : 0.8052
##                  95% CI : (0.795, 0.8151)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5444
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8228
##             Specificity : 0.7994
##          Pos Pred Value : 0.5763
##          Neg Pred Value : 0.9315
##              Prevalence : 0.2490
##          Detection Rate : 0.2049
##    Detection Prevalence : 0.3556
##       Balanced Accuracy : 0.8111
##
##        'Positive' Class : Y
##
```

```
#accuracy 80%, sens 82%, spec 80%, NIR 75%

#ROC & AUC
roc.logreg_ori = roc(X_test$income,                          as.vector(ifelse(predict(model_logreg_ori,

auc.logreg_ori = auc(roc.logreg_ori)
```



Training scenario 3: Logistic Regression Model on unstandardized age without fnl_wgt.

```
#logreg
set.seed(123)
t1_log_3=proc.time()
model_logreg_nofnlwgt=train(income ~age+workclass+education+marital_status_group+occupation+relationshi
                    data = X_train_bal,
                    trControl=cv_5,
                    metric="ROC",
                    method="regLogistic")
t2_log_3=proc.time()
cat("The computational time of training a logistic regression model is",(t2_log_3-t1_log_3)[3],"s.","\n
```

```
## The computational time of training a logistic regression model is 473 s.
```

```
#cat("The fitted logistic regression model:","\n")
#model_logreg_nofnlwgt

#confusion matrix on test data
conf_logreg_nofnlwgt=confusionMatrix( predict(model_logreg_nofnlwgt, newdata = X_test),
                          reference=X_test$income,
                          positive="Y")
cat("The confusion matrix of logistic regression model:","\n")
```

```
## The confusion matrix of logistic regression model:
```

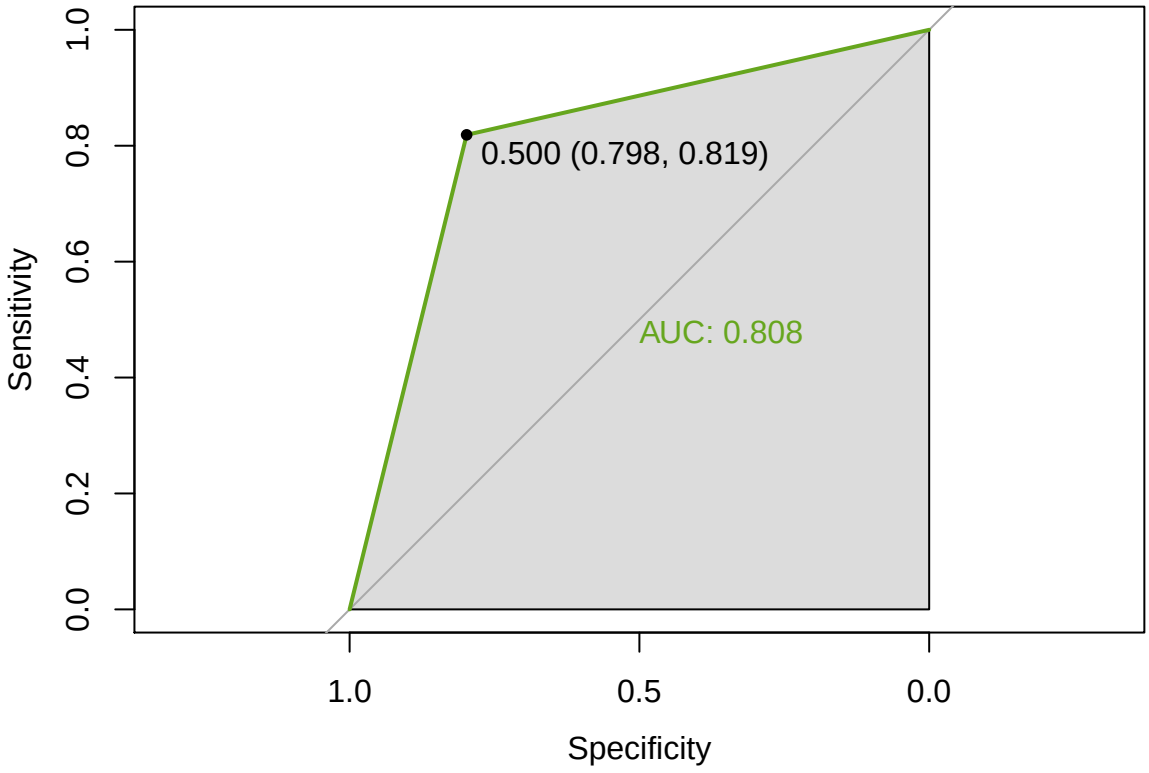
```
conf_logreg_nofnlwgt
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    N    Y
##          N 3612  272
##          Y  914 1229
##
##                Accuracy : 0.8032
##                  95% CI : (0.793, 0.8132)
##     No Information Rate : 0.751
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5397
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8188
##             Specificity : 0.7981
##          Pos Pred Value : 0.5735
##          Neg Pred Value : 0.9300
##              Prevalence : 0.2490
##          Detection Rate : 0.2039
##    Detection Prevalence : 0.3556
##       Balanced Accuracy : 0.8084
##
##        'Positive' Class : Y
##
```

```
#accuracy 80%, sens 82%, spec 80%, NIR 75%

#ROC & AUC
roc.logreg_nofnlwgt= roc(X_test$income,                          as.vector(ifelse(predict(model_logreg_

auc.logreg_nofnlwgt = auc(roc.logreg_nofnlwgt)
```




```
save(model_logreg_nofnlwgt , file = 'RegularizedLogisticRegression.rda')
saveRDS(model_logreg_nofnlwgt, "RegularizedLogisticRegression.rds")
```

The models give similar accuracy level, so for simplicity, we use the logistic model with un-transformed age and with fnl_wgt removed for deploying the shiny App.