DATA 606 Data Project Proposal

Adult Census Income

Ramnivas Singh

04/11/2021

Contents

Data Preparation

```
library(ggplot2)
library(dplyr)
library(plotly)
library(tidyverse)
library (readr)
library(DT)
library(lares)
library(ggthemes)
library(data.table)
```

```
adult.data <- read.csv("https://raw.githubusercontent.com/rnivas2028/MSDS/Data606/Final-Project/adult-al
                       stringsAsFactors = FALSE, header = TRUE, strip.white = TRUE)
# strip.white = TRUE to keep out the latest rows (empty rows)
# header = TRUE to retain header information
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

The Adult dataset is from the Census Bureau is picked for this project. The task is to find whether a given adult makes more than $50,000 a year based on the attributes such as education, age, Occupation, hours of work per week

Cases

What are the cases, and how many are there?

There are two class values '>50K' and '<=50K', meaning it is a binary classification task. The classes are imbalanced, with a skew toward the '<=50K' class label.

- '>50K': majority class, approximately 25%.
- '<=50K': minority class, approximately 75%.

There are total 48842 rows (cases) and 15 variables in this dataset

Data analysis will be performed for following categories: * Income by Education * Income by workclass & Occupation * Income by Marital status and relationship * Income by Age * Income by Gender * Income by Native country

Data collection

Describe the method of data collection.

The United States Census Bureau, officially the Bureau of the Census, is a principal agency of the U.S. Federal Statistical System, responsible for producing data about the American people and economy. Every year, the U.S. Census Bureau contacts households across the country to participate in the American Community Survey (ACS).

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

Type of study

What type of study is this (observational/experiment)?

This study is observational

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

- A data file for this project is downloaded from this link (https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data)
- A copy of this dataset and naming is retained at this link (https://raw.githubusercontent.com/rnivas2028/MSDS/Data606/Final-Project/adult-all.csv)

Response

What is the response variable, and what type is it (numerical/categorical)?

Response variable is categorical (income <=50K or >50K)

Explanatory

What is the explanatory variable, and what type is it (numerical/categorical)?

The explanatory variable is median income and is numerical. Other variables such as marital status, relationship and education level are categorical.

Relevant summary statistics

Provide summary statistics relevant to your research question. For example, if you're comparing means across groups provide means, SDs, sample sizes of each group. This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.'

Add a new field `education.segment` to show income by education

```
adult.data$education.segment <- cut(adult.data$education.num, breaks = c(0,4,8,12,17),
                              labels = c("0 to 4", "5 to 8", "9 to 12", ">= 13"))
```

Some of the variables are not self-explanatory. The variable education_num stands for the number of years of education in total, which is a continuous representation of the discrete variable education. The variable relationship represents the responding unit's role in the family. For simplicity of this analysis, the weighting factor is discarded

```
# Print header of this dataset
names(adult.data)
```

```
##  [1] "age"               "workclass"         "fnlwgt"
##  [4] "education"         "education.num"     "marital.status"
##  [7] "occupation"        "relationship"      "race"
## [10] "sex"               "capital.gain"      "capital.loss"
## [13] "hours.per.week"    "native.country"    "income"
## [16] "education.segment"
```

```
# Print first few rows
head(adult.data)
```

```
##   age        workclass fnlwgt education education.num     marital.status
## 1  39        State-gov  77516 Bachelors            13      Never-married
## 2  50 Self-emp-not-inc  83311 Bachelors            13 Married-civ-spouse
## 3  38          Private 215646   HS-grad             9           Divorced
## 4  53          Private 234721      11th             7 Married-civ-spouse
## 5  28          Private 338409 Bachelors            13 Married-civ-spouse
## 6  37          Private 284582   Masters            14 Married-civ-spouse
##          occupation  relationship  race    sex capital.gain capital.loss
## 1      Adm-clerical Not-in-family White   Male         2174            0
## 2   Exec-managerial       Husband White   Male            0            0
## 3 Handlers-cleaners Not-in-family White   Male            0            0
## 4 Handlers-cleaners       Husband Black   Male            0            0
## 5    Prof-specialty          Wife Black Female            0            0
## 6   Exec-managerial          Wife White Female            0            0
##   hours.per.week native.country income education.segment
## 1             40  United-States  <=50K             >= 13
## 2             13  United-States  <=50K             >= 13
## 3             40  United-States  <=50K           9 to 12
## 4             40  United-States  <=50K            5 to 8
## 5             40           Cuba  <=50K             >= 13
## 6             40  United-States  <=50K             >= 13
```

```r
# Print columns in the dataset with ? (value not available)
colSums(adult.data=="?")
```

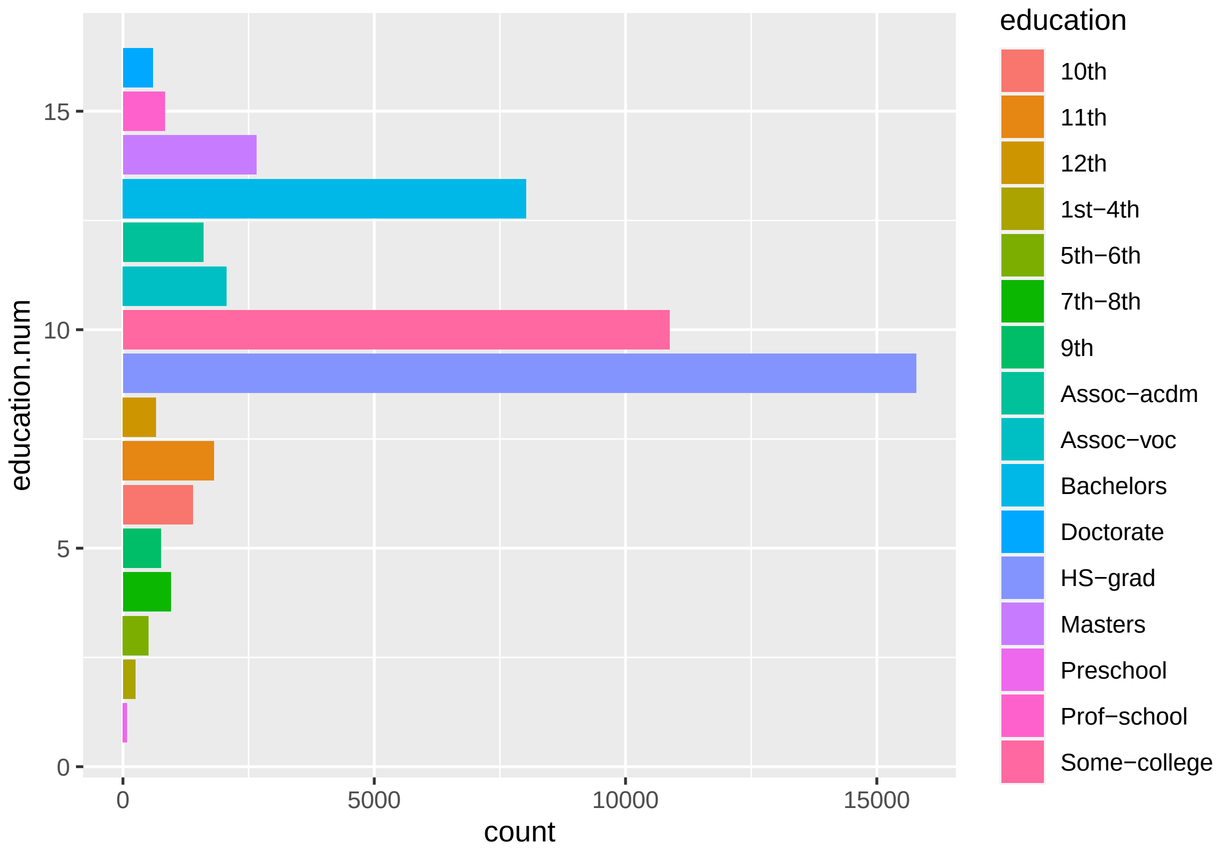
```
##               age         workclass            fnlwgt         education
##                 0              2799                 0                 0
##     education.num    marital.status        occupation      relationship
##                 0                 0              2809                 0
##              race               sex      capital.gain      capital.loss
##                 0                 0                 0                 0
##    hours.per.week    native.country            income education.segment
##                 0               857                 0                 0
```

```r
# Print summary of date set
summary(adult.data)
```

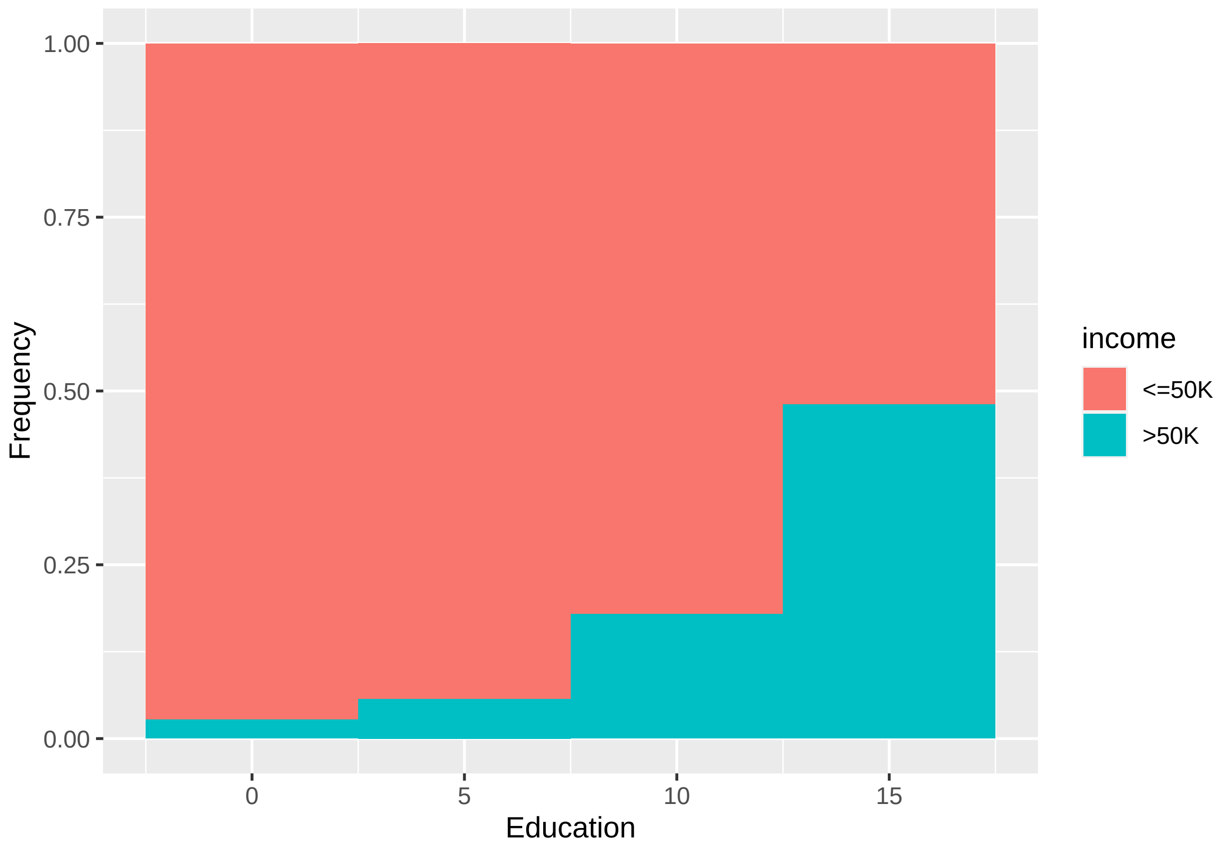
```
##       age         workclass             fnlwgt         education
##  Min.   :17.00   Length:48842       Min.   :  12285   Length:48842
##  1st Qu.:28.00   Class :character   1st Qu.: 117551   Class :character
##  Median :37.00   Mode  :character   Median : 178145   Mode  :character
##  Mean   :38.64                      Mean   : 189664
##  3rd Qu.:48.00                      3rd Qu.: 237642
##  Max.   :90.00                      Max.   :1490400
##  education.num   marital.status     occupation         relationship
##  Min.   : 1.00   Length:48842       Length:48842       Length:48842
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      race               sex             capital.gain    capital.loss
##  Length:48842       Length:48842       Min.   :    0   Min.   :   0.0
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.0
##  Mode  :character   Mode  :character   Median :    0   Median :   0.0
##                                        Mean   : 1079   Mean   :  87.5
##                                        3rd Qu.:    0   3rd Qu.:   0.0
##                                        Max.   :99999   Max.   :4356.0
##  hours.per.week  native.country      income          education.segment
##  Min.   : 1.00   Length:48842       Length:48842       0 to 4 : 1794
##  1st Qu.:40.00   Class :character   Class :character   5 to 8 : 4614
##  Median :40.00   Mode  :character   Mode  :character   9 to 12:30324
##  Mean   :40.42                                         >= 13  :12110
##  3rd Qu.:45.00
##  Max.   :99.00
```

1.0 Income by Education Lets print various education level for reference purposes

```r
ggplot(data = adult.data) +
  aes(y=education.num, fill=education) +
  geom_bar()
```

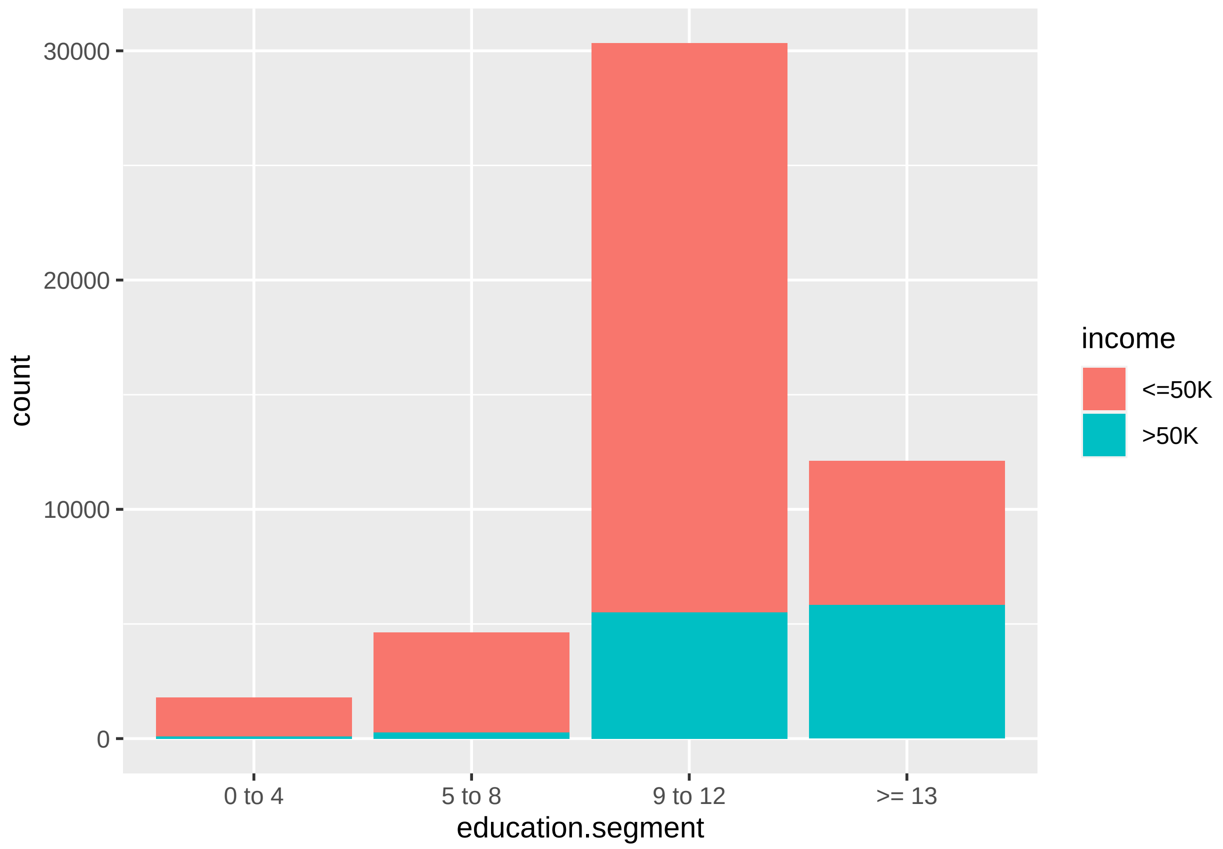


```
ggplot(data = adult.data) +
  aes(x = education.num, fill = income) +
  geom_histogram(binwidth=5, position="fill") +
  labs(x="Education", y="Frequency")
```

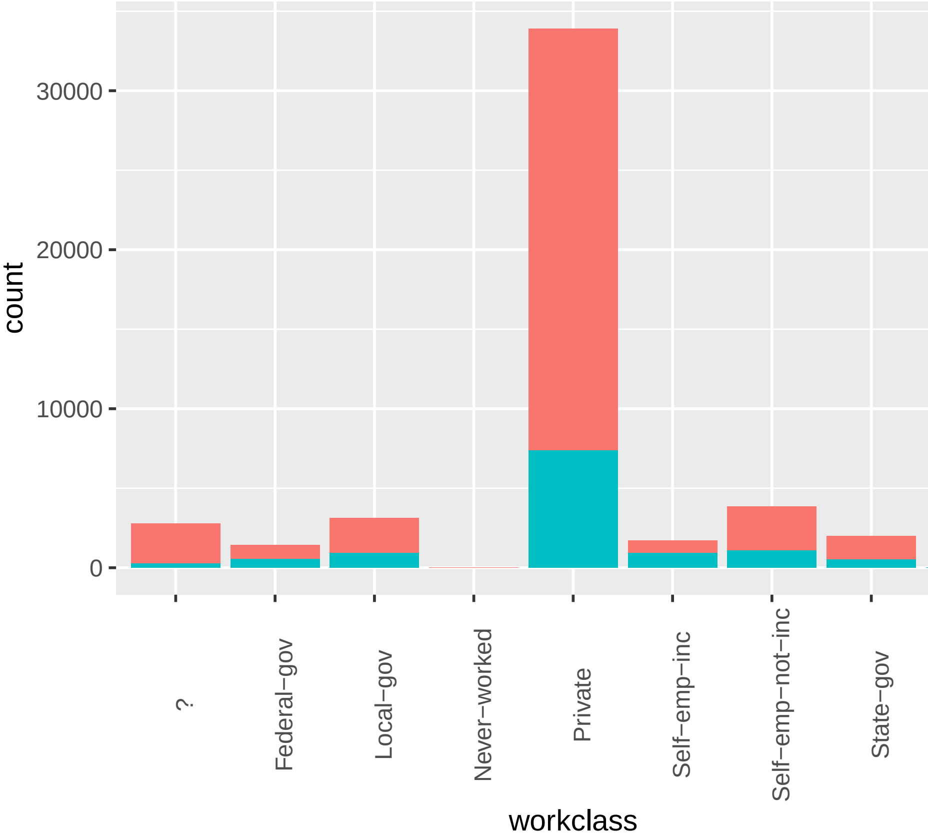


Relationship between education and income

```
ggplot(data = adult.data) +
  aes(x=education.segment ,fill=income) +
  geom_bar()
```

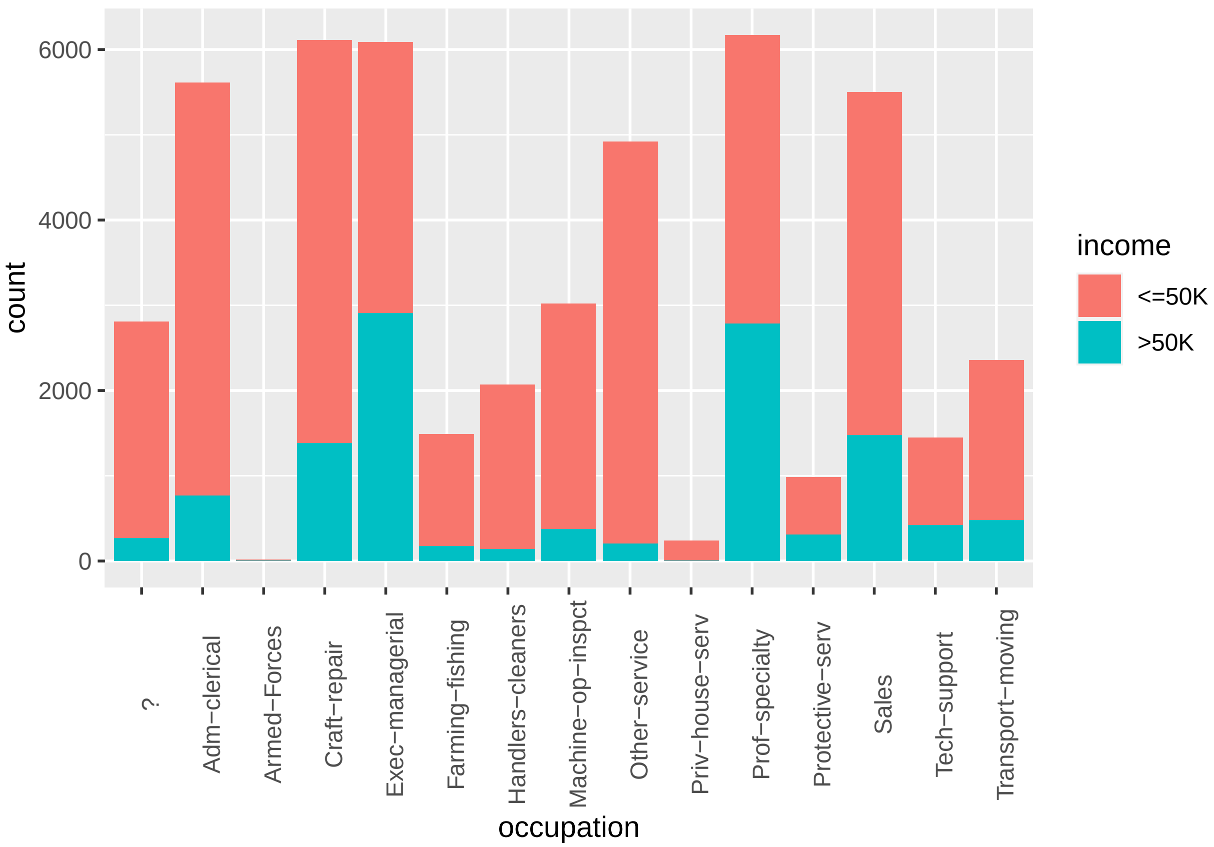

30000
20000
10000
0
count
0 to 4
5 to 8
9 to 12
>= 13
education.segment
income
<=50K
>50K

```
ggplot(data = adult.data) +
  aes(x=workclass,fill=income) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90))
```

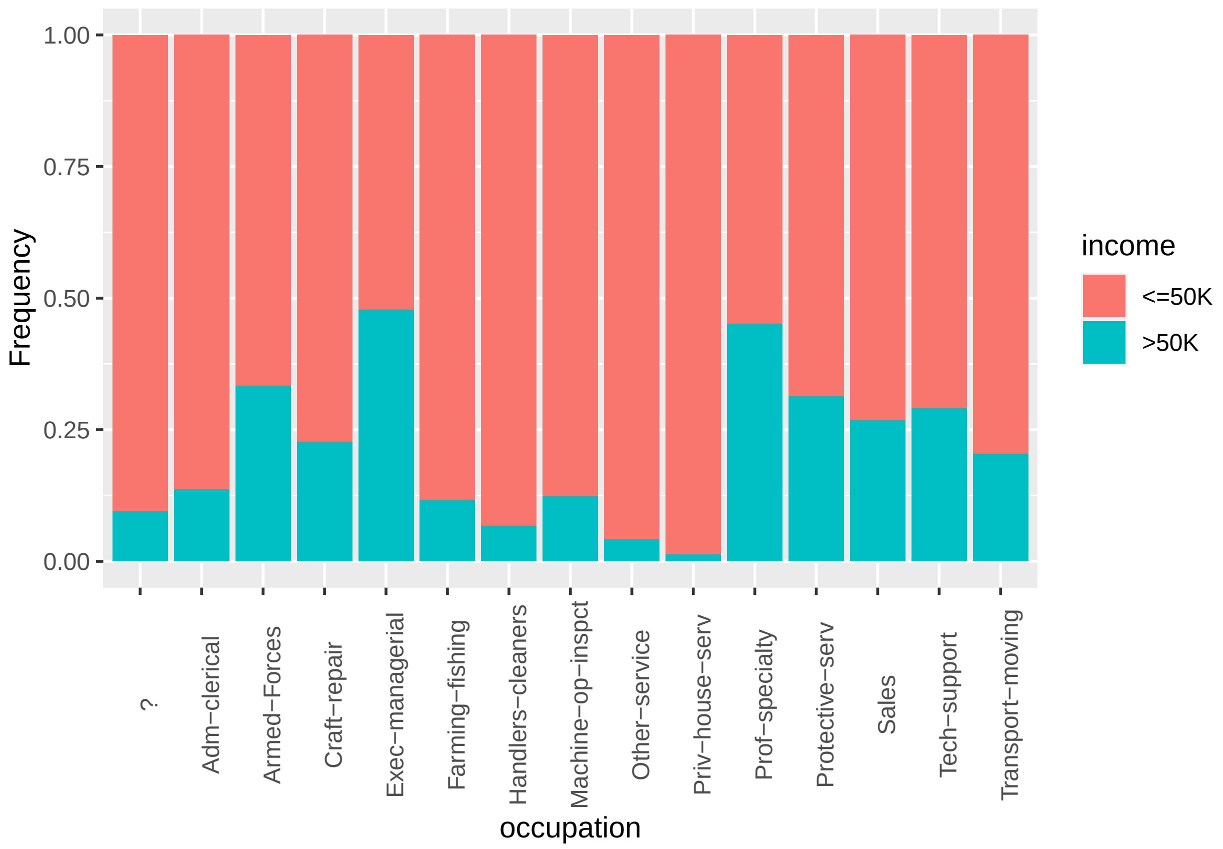


2.0 Income workclass & Occupation:
Private sector workers are the most likely to have an income of over 50K.

```
ggplot(data = adult.data) +
  aes(x=occupation,fill=income) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90))
```

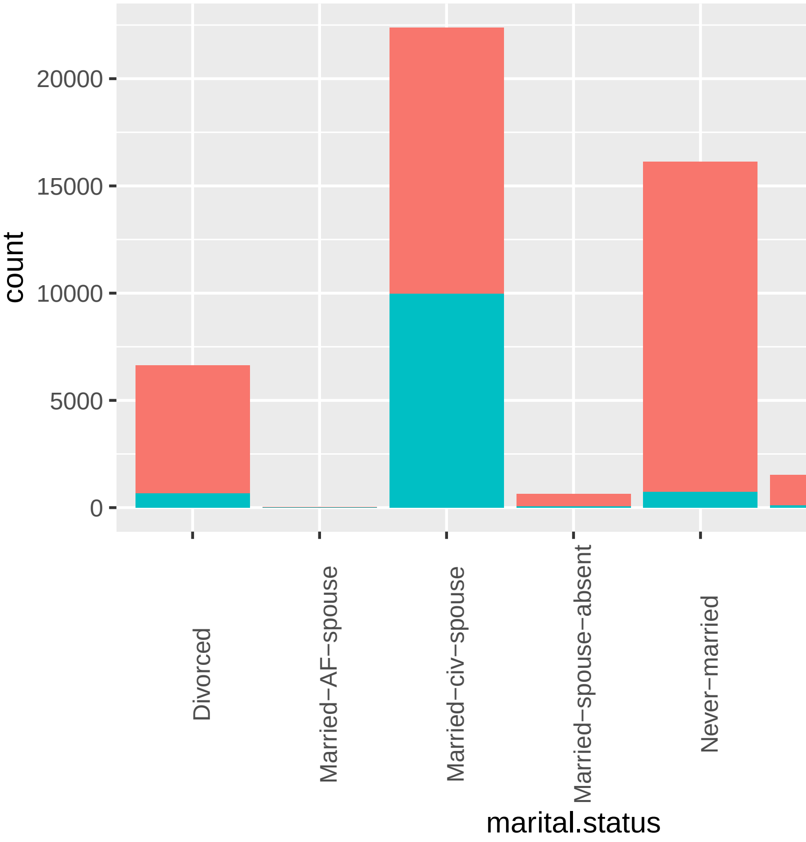


```
ggplot(data = adult.data) +
  aes(x=occupation,fill=income) +
  geom_bar(position="fill") +
  ylab("Frequency") +
  theme(axis.text.x = element_text(angle = 90))
```



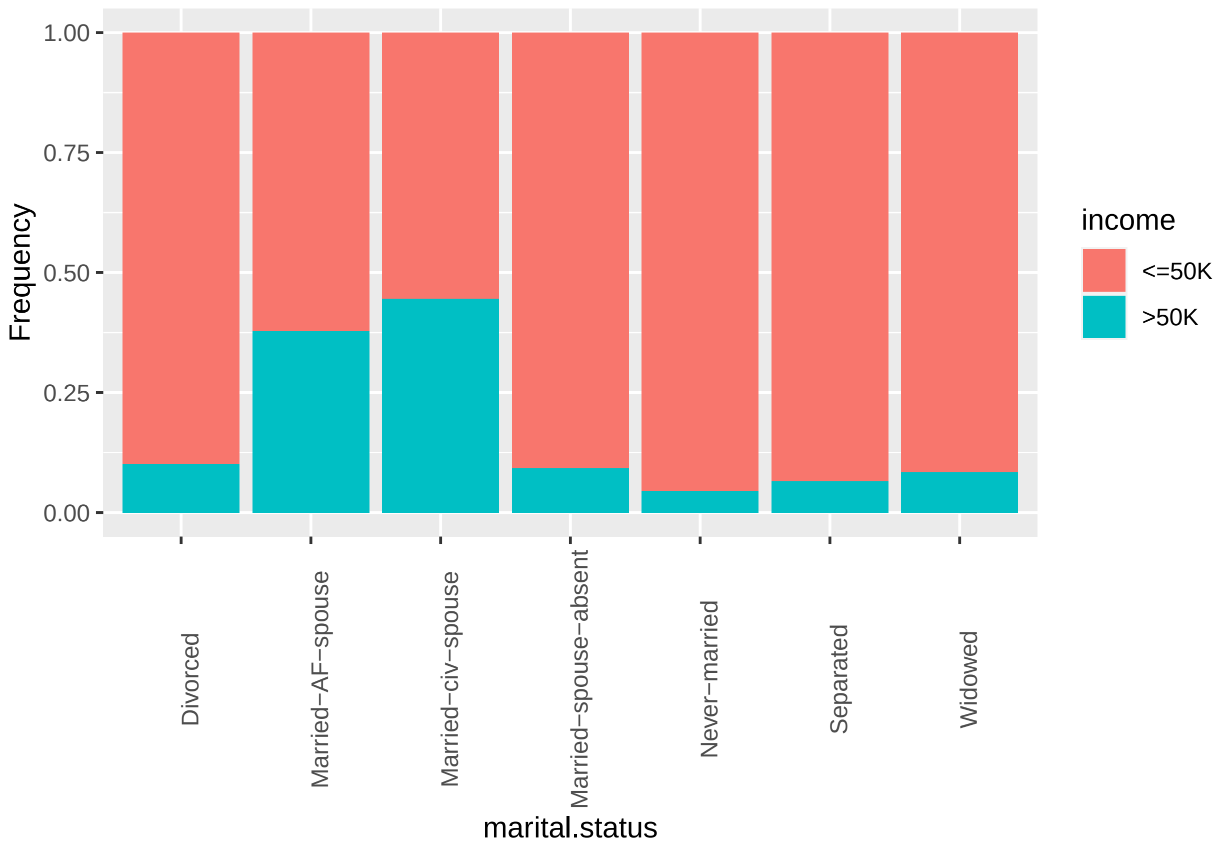
Frequency
1.00
0.75
0.50
0.25
0.00
?
Adm−clerical
Armed−Forces
Craft−repair
Exec−managerial
Farming−fishing
Handlers−cleaners
Machine−op−inspct
Other−service
Priv−house−serv
Prof−specialty
Protective−serv
Sales
Tech−support
Transport−moving
occupation
income
<=50K
>50K

```
ggplot(data = adult.data) +
  aes(x=marital.status,fill=income) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90))
```



3.0 Income by Marital status and relationship

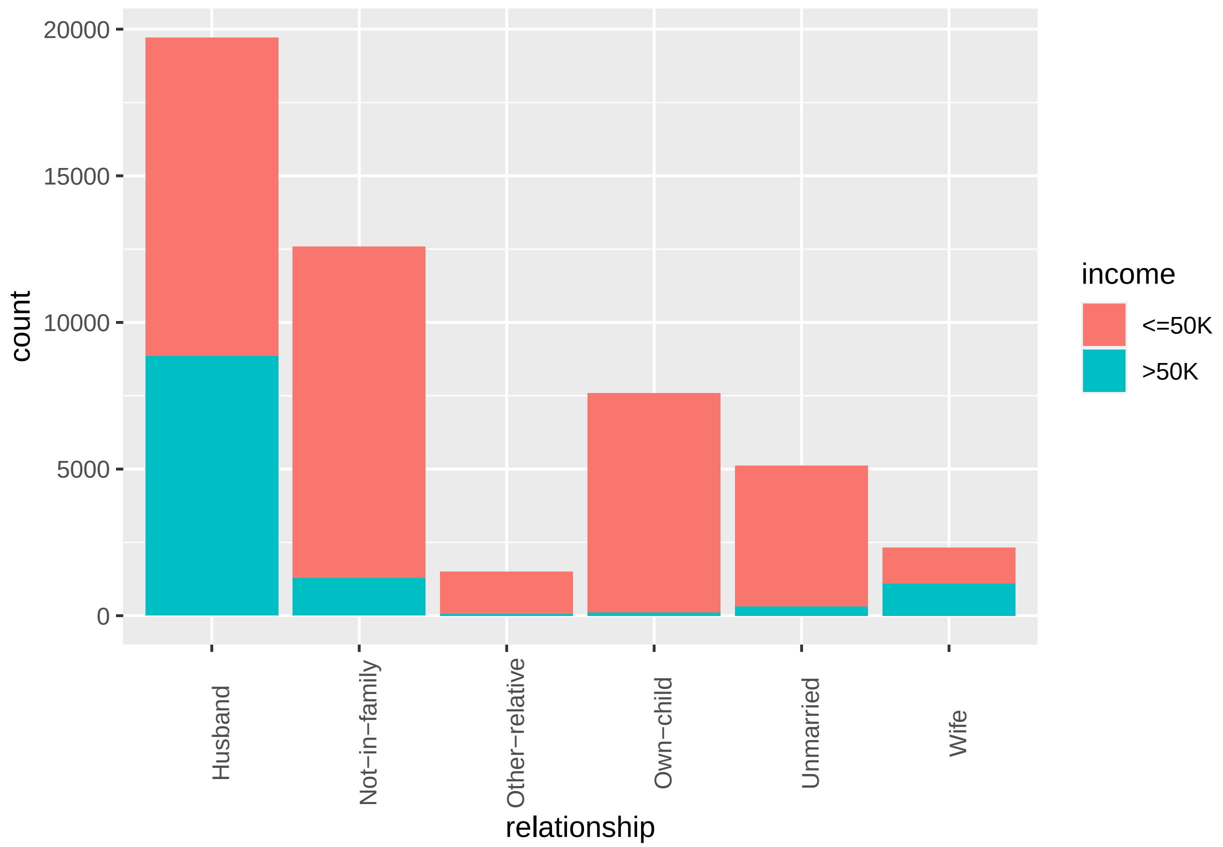
```
ggplot(data = adult.data) +
  aes(x=marital.status,fill=income) +
  geom_bar(position="fill") +
  ylab("Frequency") +
  theme(axis.text.x = element_text(angle = 90))
```



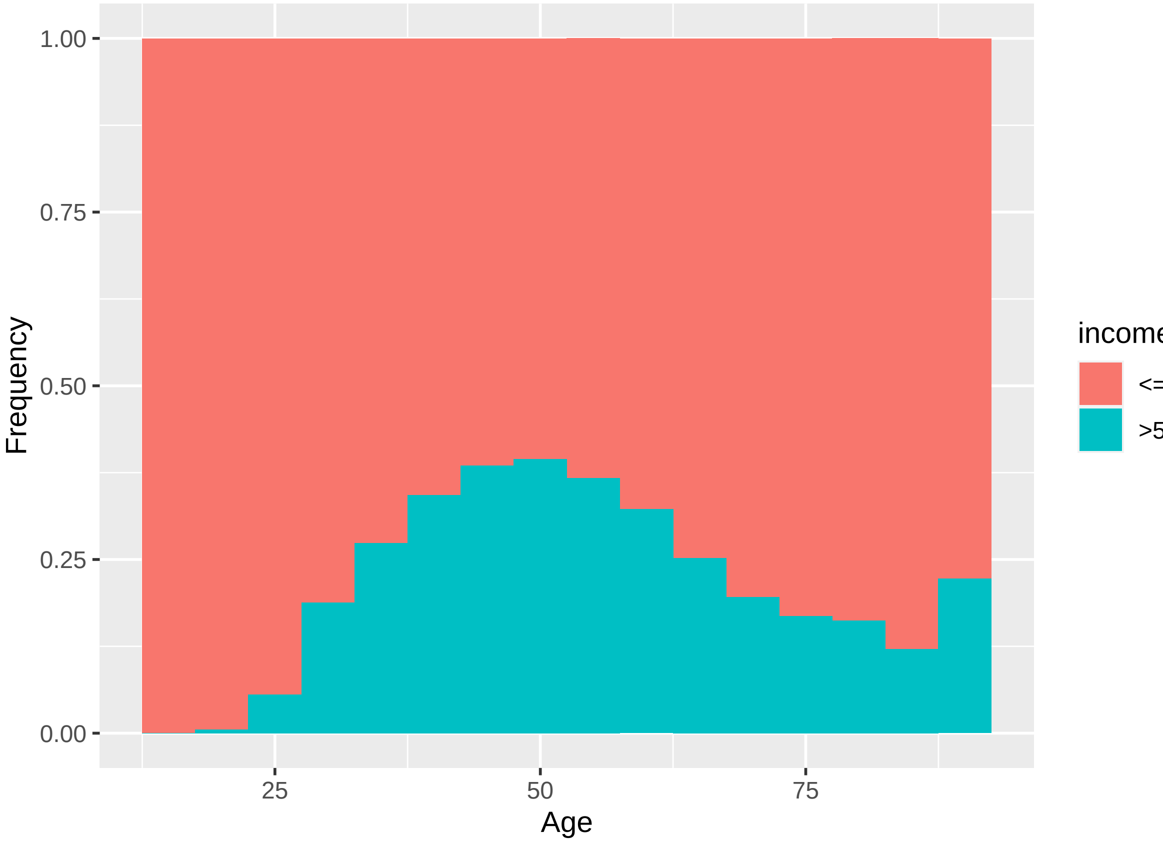
Looks like higher percentage of people with incomes above 50K among married people

By relationship:

```
ggplot(data = adult.data) +
  aes(x=relationship,fill=income) +
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90))
```


count
20000
15000
10000
5000
0
Husband
Not−in−family
Other−relative
Own−child
Unmarried
Wife
relationship
income
<=50K
>50K

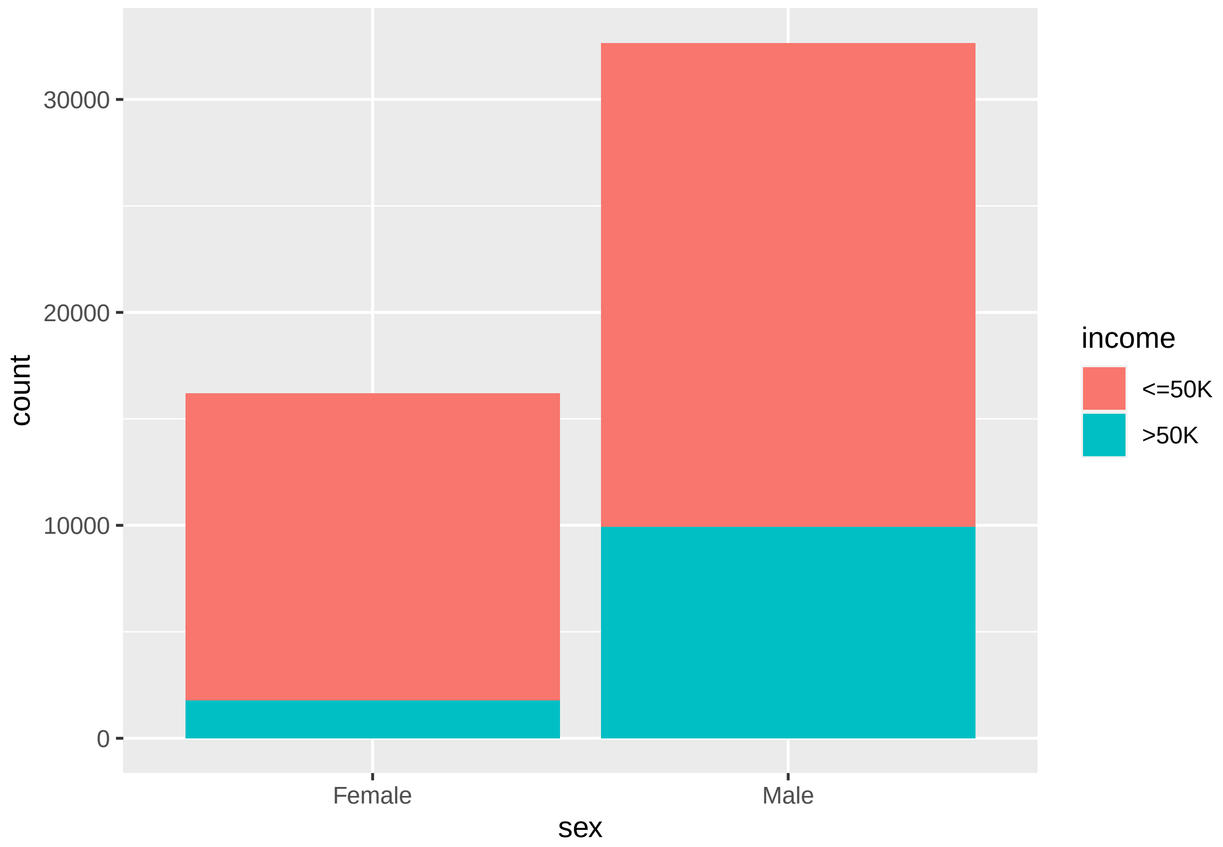
```
ggplot(data = adult.data) +
  aes(x = age, fill = income) +
  geom_histogram(binwidth=5, position="fill") +
  labs(x="Age", y="Frequency")
```



4.0 Income by Age

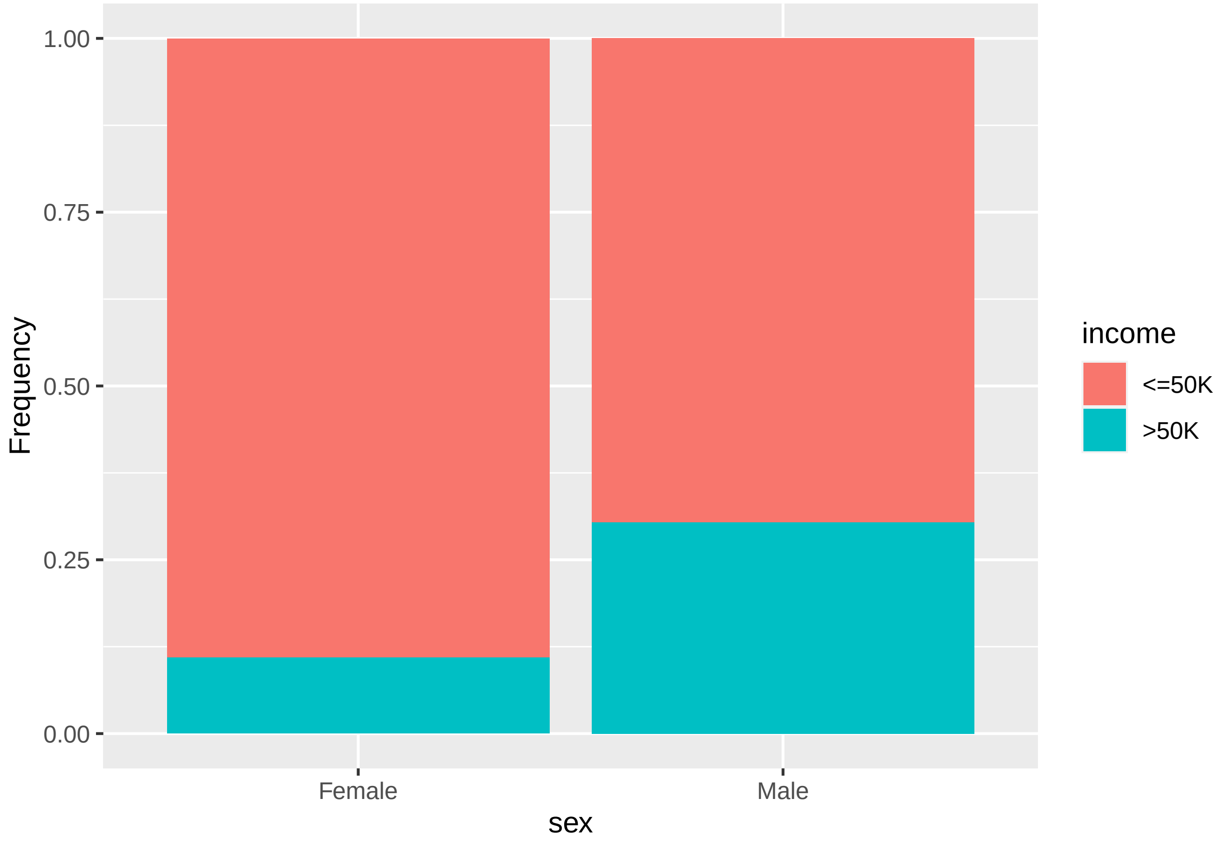
5.0 Income by Gender We see the distribution by gender.

```
ggplot(data = adult.data) +
  aes(x=sex,fill=income) +
  geom_bar()
```



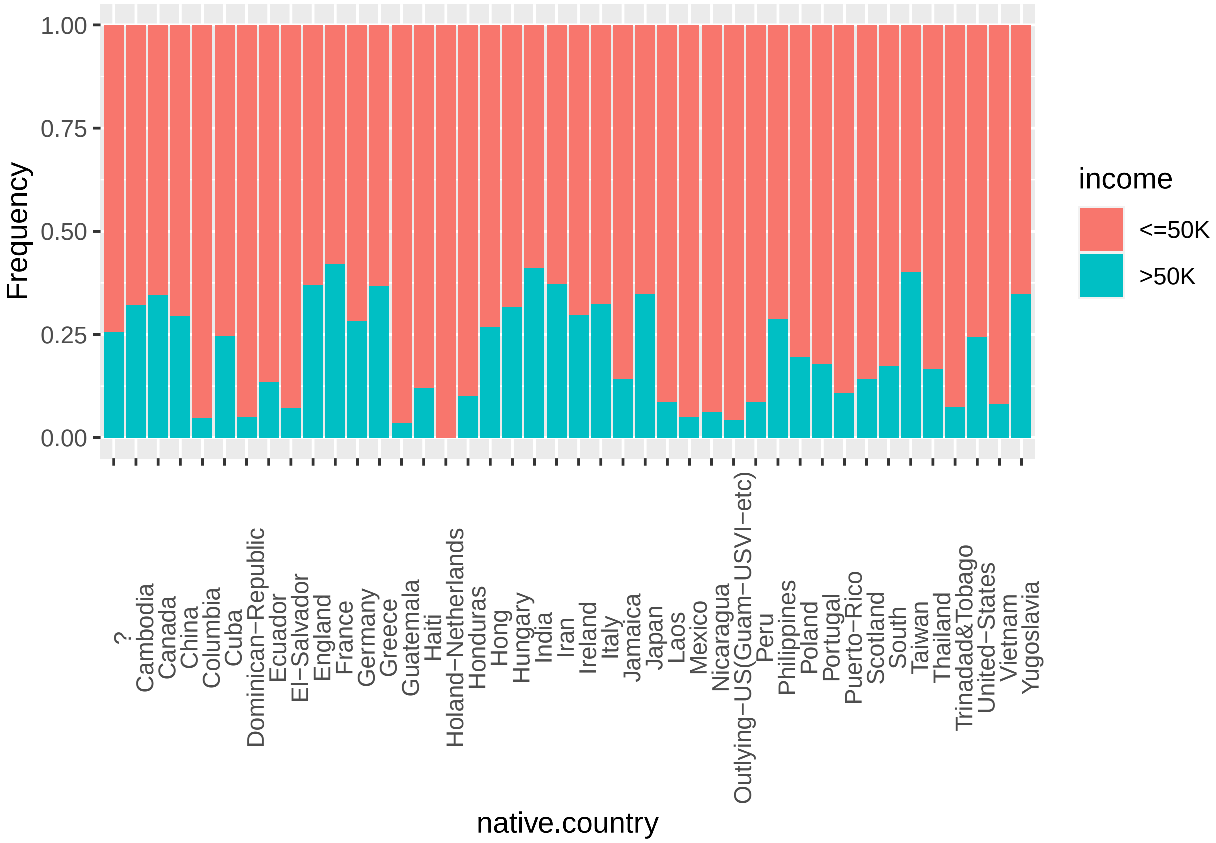
At first glance, it can be seen that the proportion of women with income above 50K is lower than the proportion of men with income above 50K. Still, we show the frequency distribution.

```
ggplot(data = adult.data) +
  aes(x=sex,fill=income) +
  geom_bar(position="fill") +
  ylab("Frequency")
```


1.00
0.75
0.50
0.25
0.00
Frequency
Female
Male
sex
income
<=50K
>50K

6.0 Income by Native country We show income by country of origin.

```
ggplot(data = adult.data) +
  aes(x=native.country,fill=income) +
  geom_bar(position="fill") +
  ylab("Frequency") +
  theme(axis.text.x = element_text(angle = 90))
```



There are important differences in income depending on the country of origin.

Conclusion

- Based on this analysis on Census bureau database - its clear that no many individuals and families income is more than 50K

- Professional specialty and managerial occupations clearly earn more than 50K
- Around 75% people earn less than 50K where as around 25% earn more than 50K
- People with native from Taiwan, India, France reported 50K or more income

References

Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics)
Imbalanced Classification with the Adult Income Dataset by Jason Brownlee on March 6, 2020 in Imbalanced Classification
Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.