P02 Model Planning and Building

Kris Walker

11/27/2019

Introduction

Our final goal is to build a model which can predict whether the income of a random adult American citizen is less than or greater than $50,000 a year based on given features such as age, education, occupation, gender, race, etc.

Here we perform some initial data analysis, investigating the potential predictor variables as well as their relationship with the response variable `income` .

First we load the necessary packages:

```
library(ggplot2)
library(plyr)
library(gridExtra)
library(gmodels)
library(grid)
library(vcd)
library(scales)
library(ggthemes)
library(tinytex)
```

Reading the Preprocessed Data

Below we set the working directory and we then read the preprocessed training dataset into the `adult_data` dataframe:

```
setwd("/home/taudin/MiscFiles/Fall19/CSCI385/DSProject/CensusData")
adult_data <- read.csv("adult_df.csv")
```

Analysis of the Variables and Their Correlation with Income

The Variable `income`

We will start by taking a look at the response variable `income` . Our goal will be to build a model that predicts if a person earns more than $50,000 a year. Let us recall that `income` is a factor variable that has two levels:

```
class(adult_data$income)
```

```
## [1] "factor"
```

```
levels(adult_data$income)
```

```
## [1] " <=50K" " >50K"
```

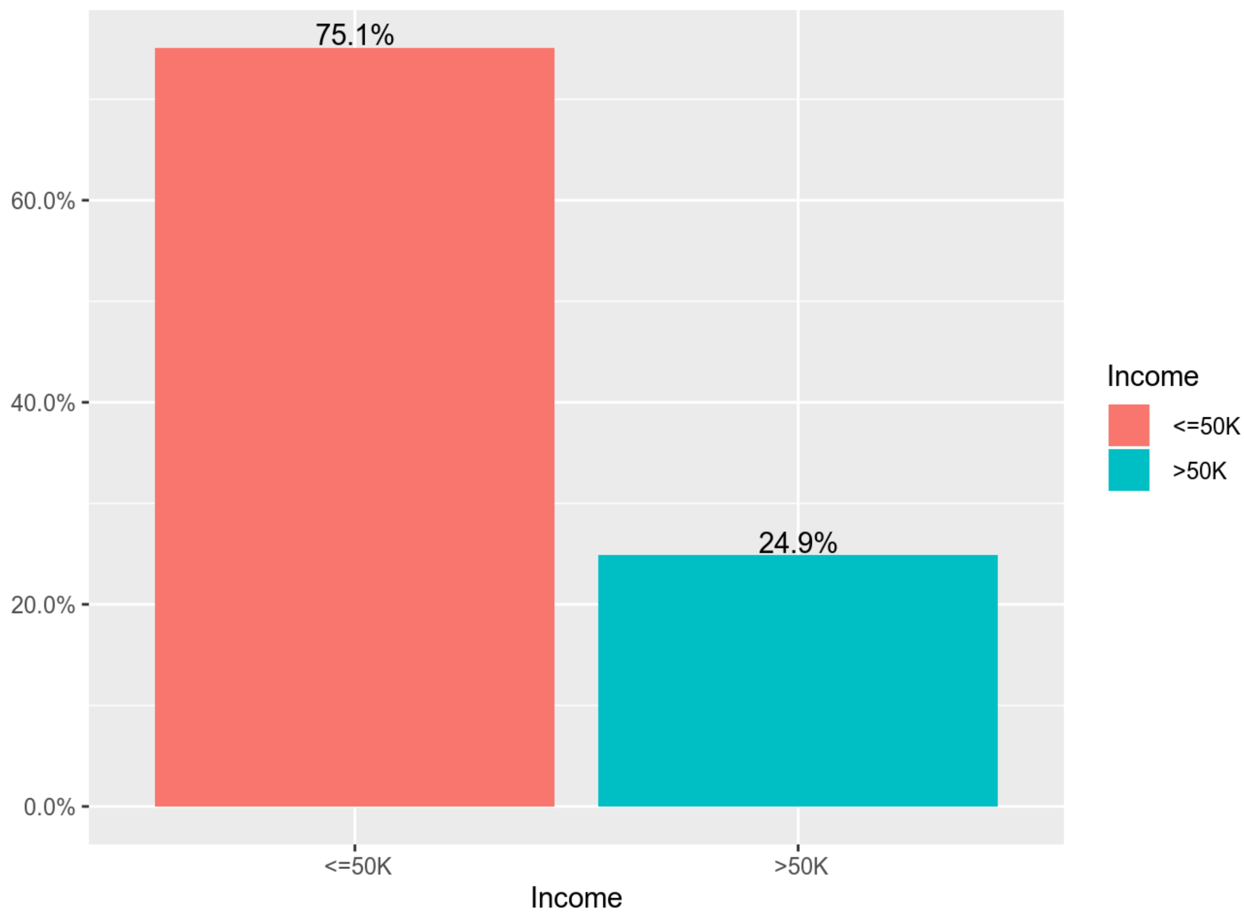
Let's look at the summary statistic…

```
summary(adult_data$income)
```

```
##  <=50K   >50K
##  22654   7508
```

…and visualize the above results with a bar plot:

```
ggplot(data = adult_data, mapping = aes(x = adult_data$income, fill = adult_data$income))
+
  geom_bar(mapping = aes(y = (..count..) / sum(..count..))) +
  geom_text(mapping = aes(label = scales::percent((..count..) / sum(..count..)),
                         y = (..count..) / sum(..count..)), stat = "count", vjust = -.1)
 +
  labs(x = "Income", y = "", fill = "Income") +
  scale_y_continuous(labels = percent)
```



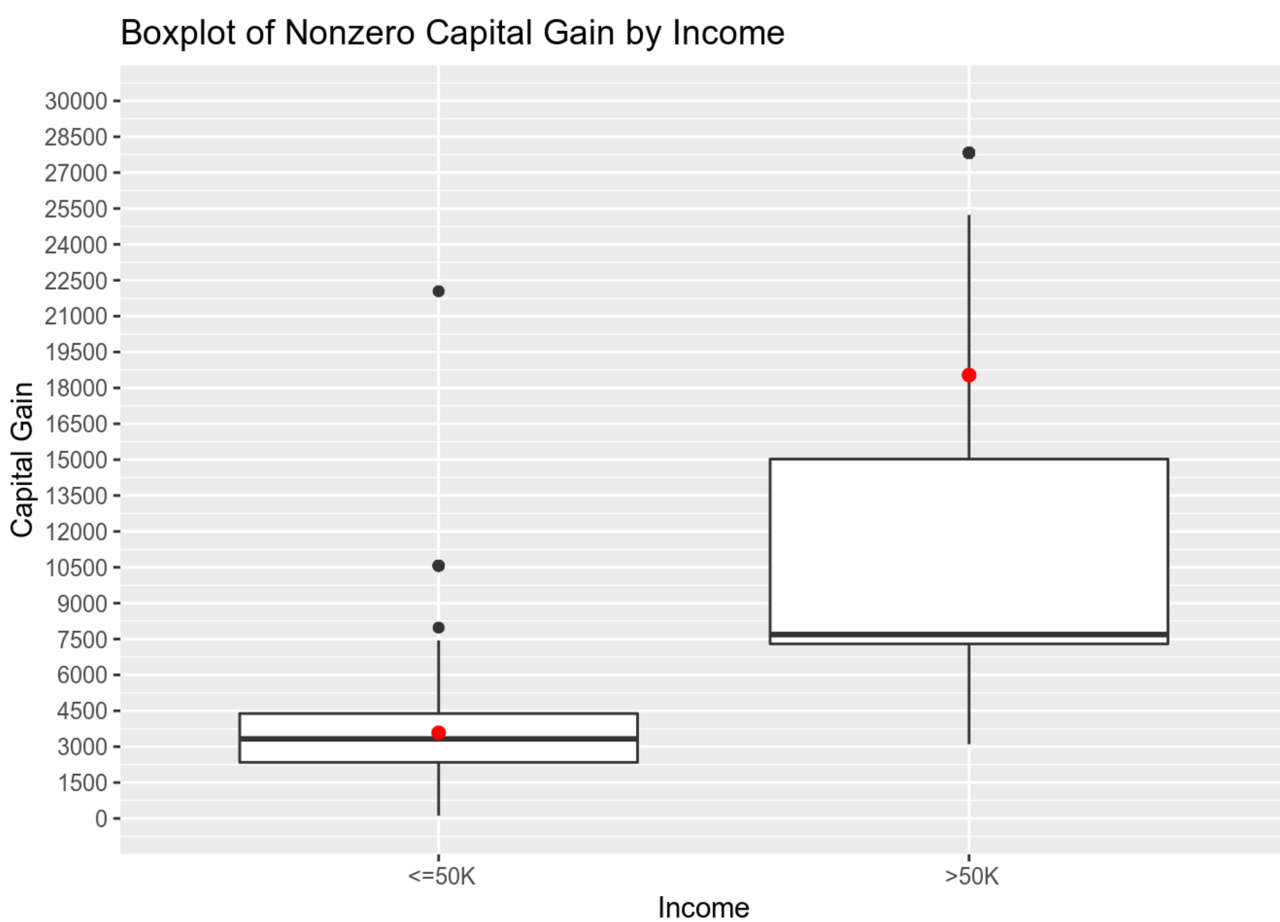
The graph shows us the percentage of people that earn less than or more than 50K year. We see that 75.1% of the participants are paid less than 50K a year and 24.9% are paid more than 50K a year.

The Variables `capital_gain`, `cap_gain`, `capital_loss`, and `cap_loss`

Nonzero `capital_gain` and `capital_loss`

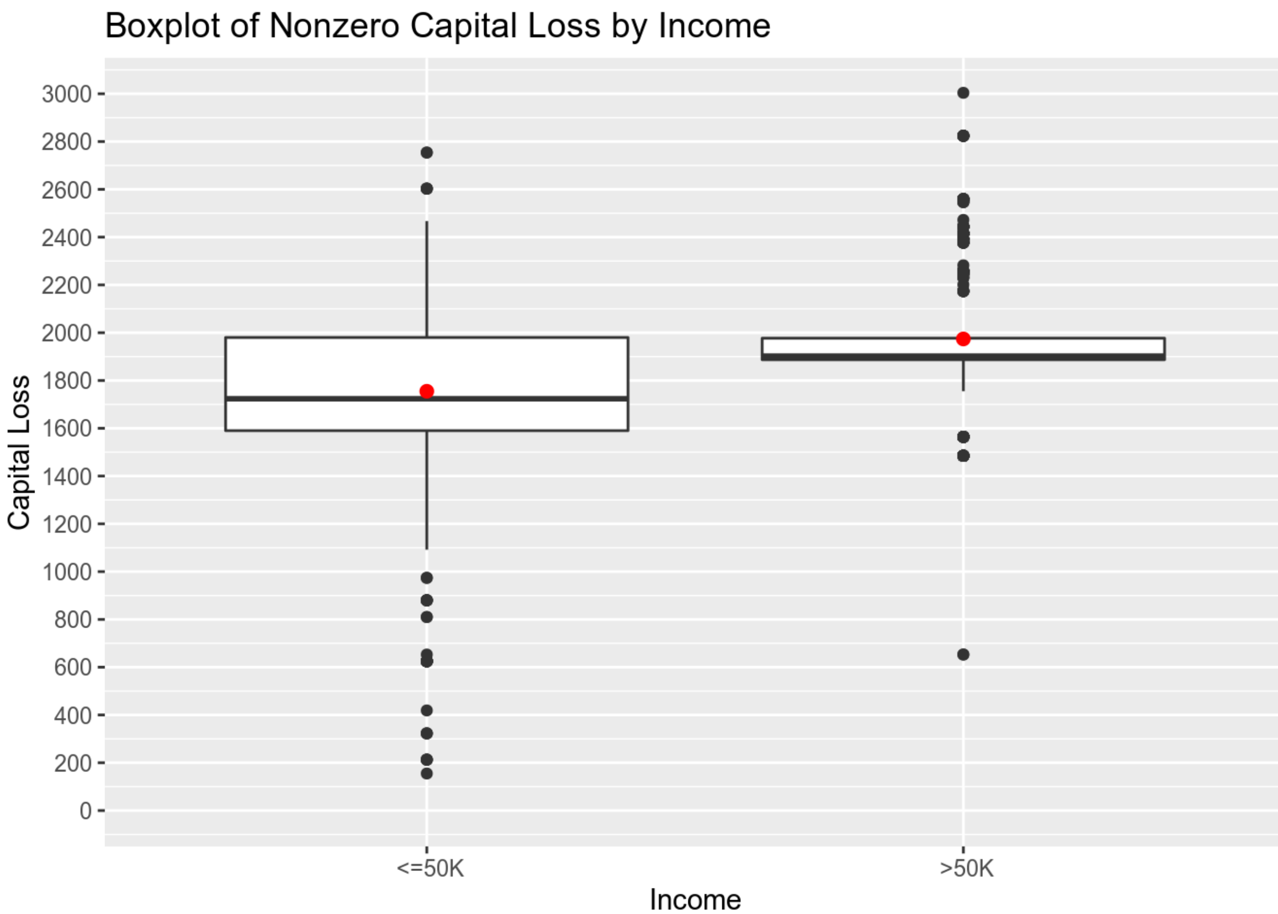
We will show the boxplots of the nonzero capital gain and loss grouped by income. The mean is depicted with a red dot. Considering the variable `capital_gain` , a major portion of the values (50% of the data points), as well as the median and the mean are significantly greater for those earning more than 50K a year than that data is for those that earn less than 50K a year:

```
ggplot(mapping = aes(x = income, y = capital_gain), data = subset(adult_data, adult_data$c
apital_gain > 0)) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 19, color = "red", cex = 2) +
  coord_cartesian(ylim = c(0, 30000)) +
  scale_y_continuous(breaks = seq(0, 30000, 1500)) +
  labs(x = "Income", y = "Capital Gain") +
  ggtitle("Boxplot of Nonzero Capital Gain by Income")
```



Now we will show a boxplot of nonzero capital loss grouped by income and we will see the same trend as we observed for capital gain: The mean and median for those earning more than 50K a year is greater than the mean and median for those earning less than 50K a year. This could possibly be due to people with higher incomes are likely to invest more of their money more often which then leads to higher chances of not only rewards through good investments, but also losses from bad investments.

```
ggplot(mapping = aes(x = income, y = capital_loss), data = subset(adult_data, adult_data$c
apital_loss > 0)) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 19, color = "red", cex = 2) +
  coord_cartesian(ylim = c(0, 3000)) +
  scale_y_continuous(breaks = seq(0, 3000, 200)) +
  labs(x = "Income", y = "Capital Loss") +
  ggtitle("Boxplot of Nonzero Capital Loss by Income")
```

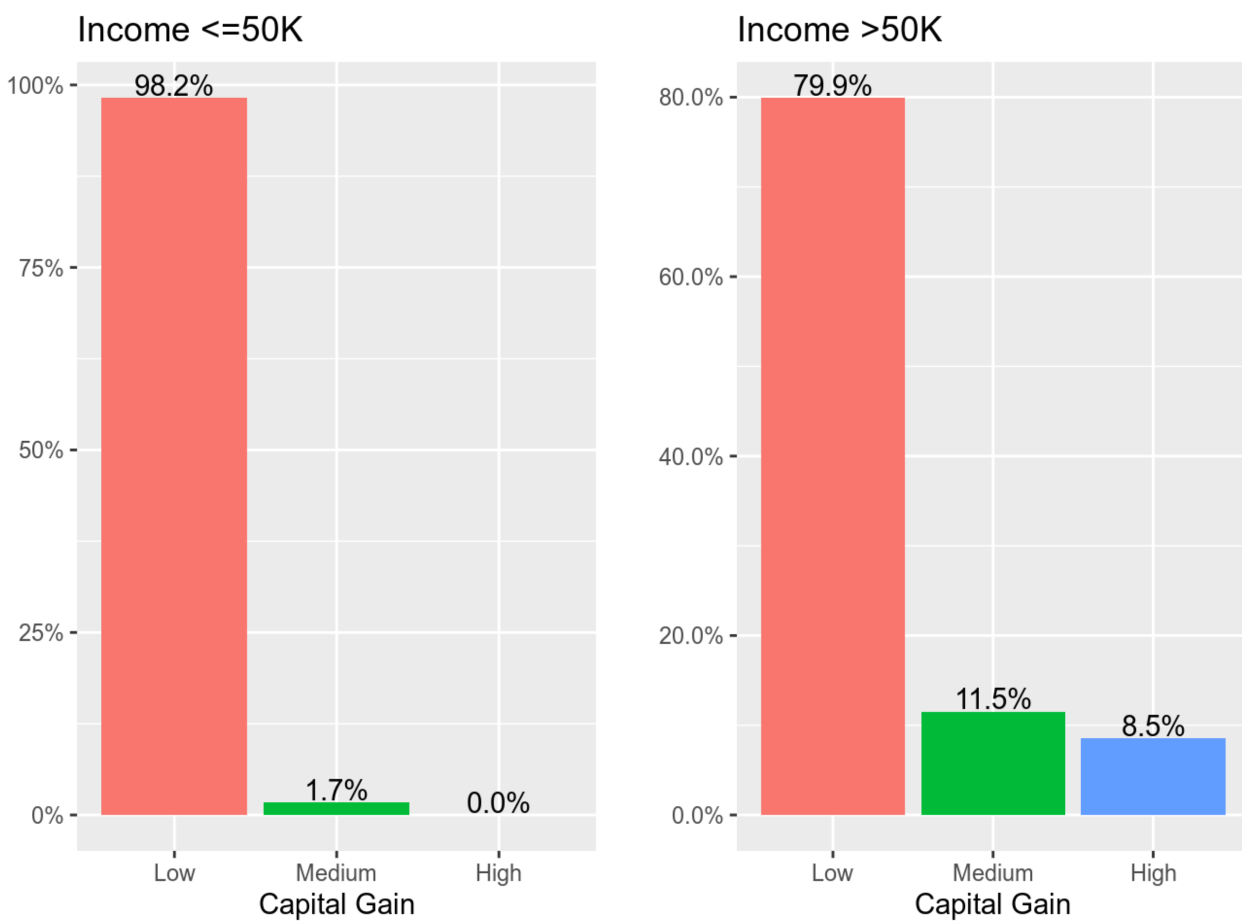


We can say that there is evidence for a strong relationship between the nonzero values of `capital_gain`, `capital_loss`, and `income`, but we will not include these variables in the predictive model we're building because of the high number of zeros among the values of these variables. Also, less than 10% of the participants make investments.

`cap_gain` and `cap_loss`

We will explore the relationship between the factor variables `cap_gain`, `cap_loss` and the categorical variable `income`. Let's take a look at bar plots of the two variables grouped by income:

```
lg_cap_gain <- lapply(X = levels(adult_data$income), FUN = function(v){
  df <- subset(adult_data, adult_data$income == v)
  df <- within(df, cap_gain <- factor(cap_gain, levels = names(sort(table(cap_gain), decre
asing = TRUE))))
  ggplot(data = df, aes(x = cap_gain, fill = cap_gain)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = -.1) +
    labs(x = "Capital Gain", y = "", fill = "Capital Gain") +
    theme(legend.position = "none") +
    ggtitle(paste("Income", v, sep = "")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_cap_gain, ncol = 2)
```



We see that 0% of the people who earn less than 50K a year have a high capital gain. To check to see if this result is due to some rounding error, we display the number of individuals with income less than 50K a year and high capital gain:

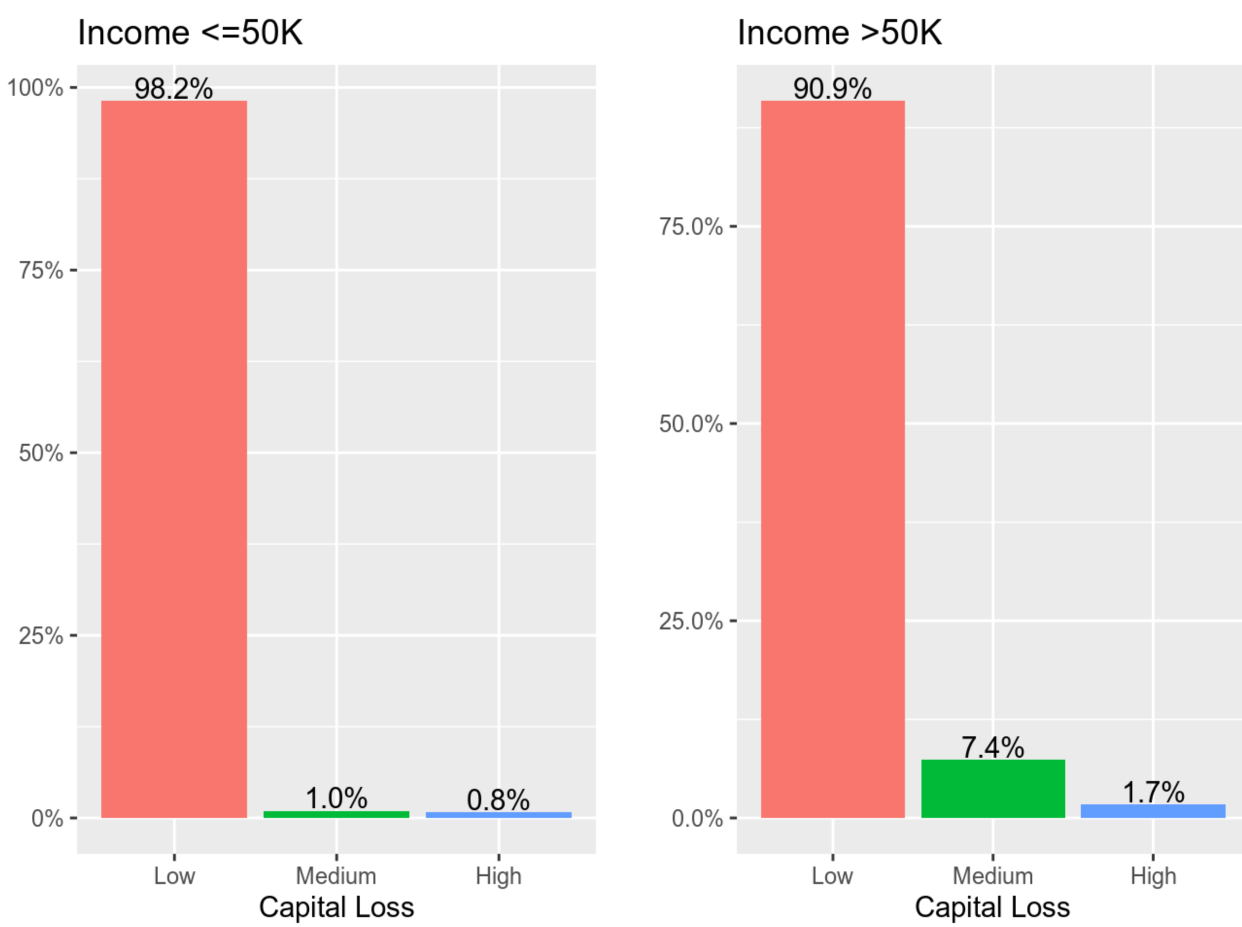
```
nrow(subset(adult_data, adult_data$cap_gain == "High" & adult_data$income == " <=50K"))
```

```
## [1] 6
```

There are indeed people with income less than 50K and high capital gain. The bar plot also tells us that the proportion of people who have medium and high capital gain is larger within the group of people with income of more than 50K a year compared to the respective proportion within the group of people with income less than 50K yearly. We can safely conclude that there is a relationship between `cap_gain` and `income`.

We shall consider the variable `cap_loss`:

```
lg_cap_loss <- lapply(levels(adult_data$income), function(v){
  df <- subset(adult_data, adult_data$income == v)
  df <- within(df, cap_loss <- factor(cap_loss, levels = names(sort(table(cap_loss), decre
asing = TRUE))))
  ggplot(data = df, aes(x = cap_loss, fill = cap_loss)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = -.1) +
    labs(x = "Capital Loss", y = "", fill = "Capital Loss") +
    theme(legend.position = "none") +
    ggtitle(paste("Income", v, sep = "")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_cap_loss, ncol = 2)
```



We observe the same trend as in the case of the variable `cap_gain`.

The Variable `age`

Let’s take a look at the the summary of age and its IQR:

```
summary(adult_data$age)
```

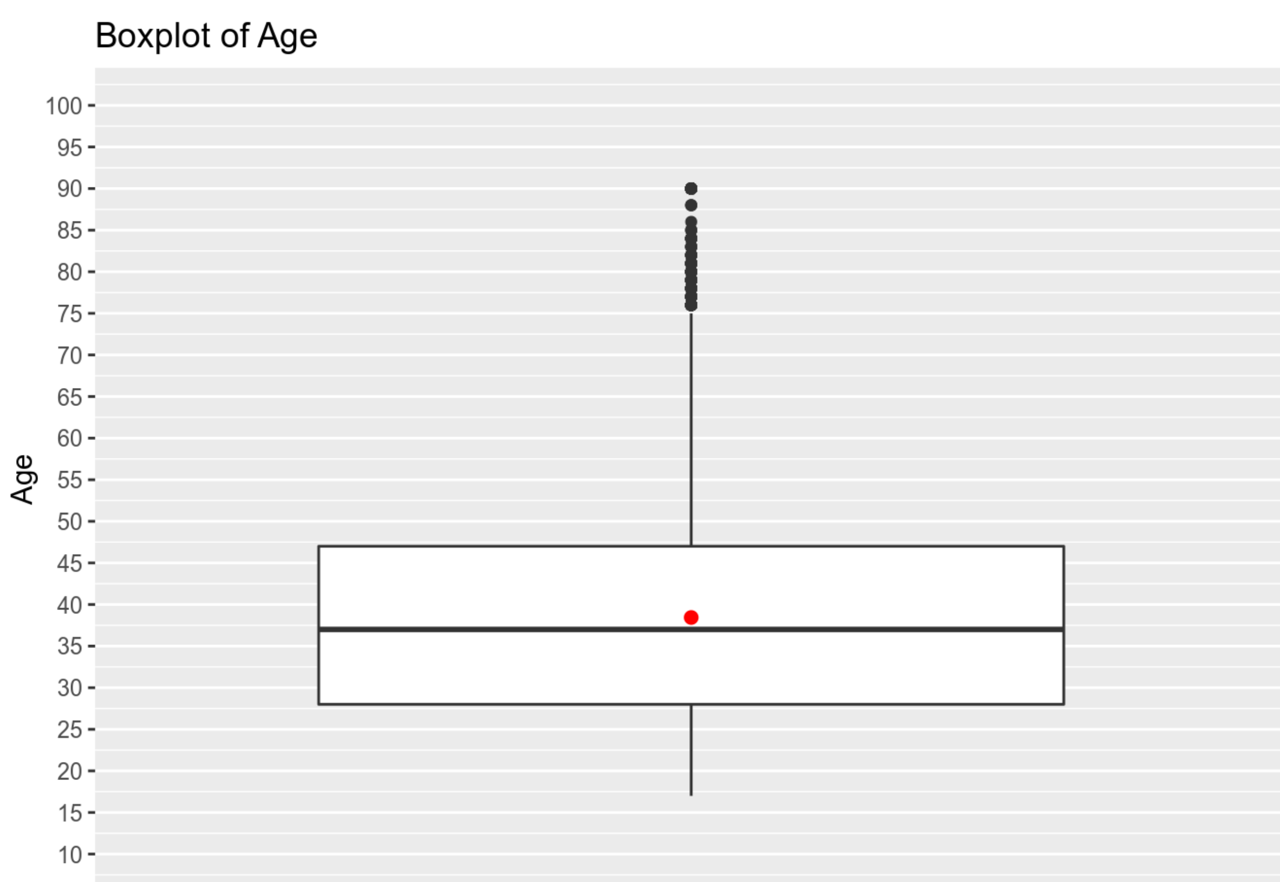
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   28.00   37.00   38.44   47.00   90.00
```

```
IQR(adult_data$age)
```

```
## [1] 19
```

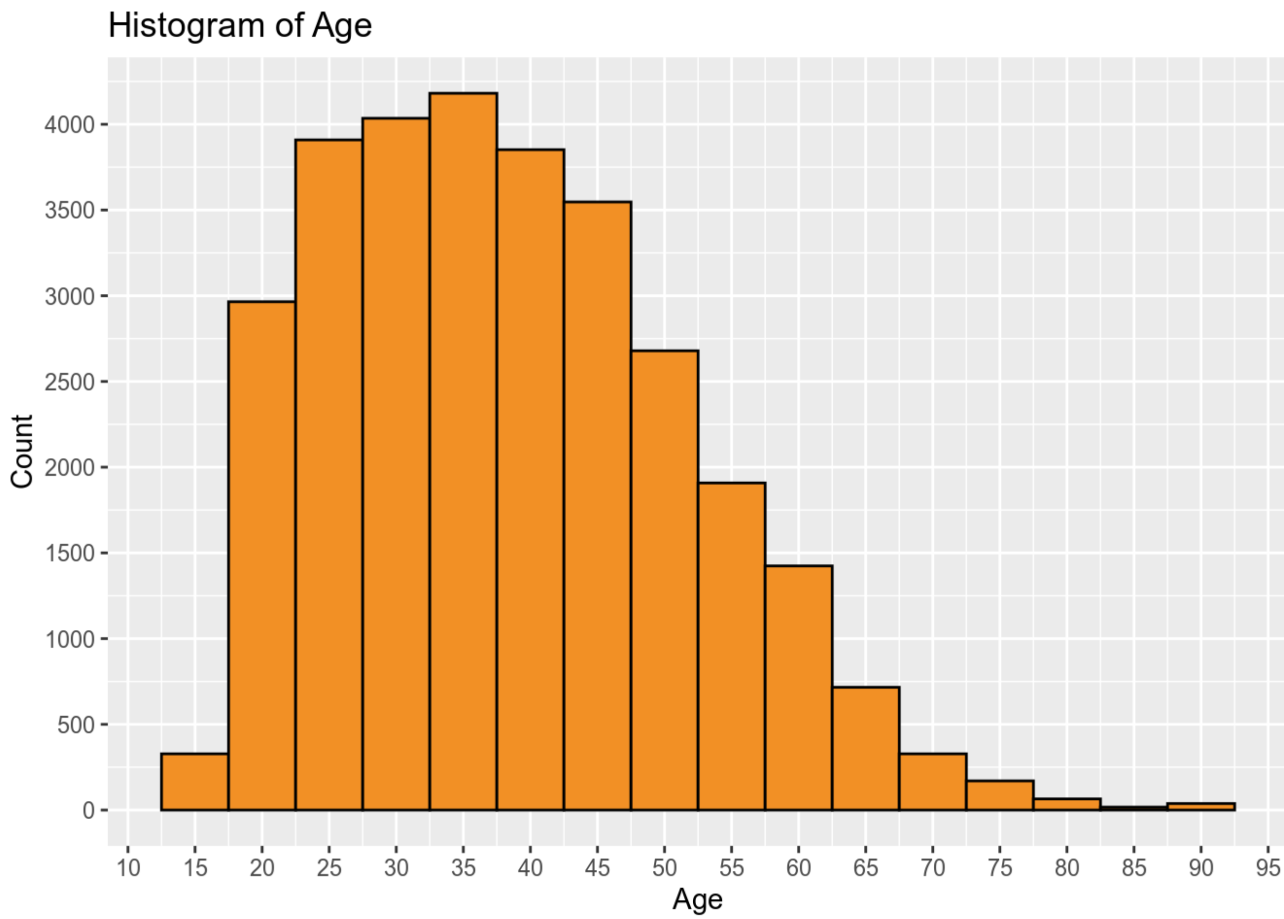
The summary shows us that at least 50% of the people in the study are between 28 and 47 years old with median age 37 and the mean age 38. We see that there are some outliers where some people are between 75 and 90 years old. We will display a boxplot of the variable `age` to visualize our summary statistics:

```
ggplot(mapping = aes(x = factor(0), y = age), data = adult_data) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 19, color = "red", cex = 2) +
  coord_cartesian(ylim = c(10, 100)) +
  scale_y_continuous(breaks = seq(10, 100, 5)) +
  ylab("Age") +
  xlab("") +
  ggtitle("Boxplot of Age") +
  scale_x_discrete(breaks = NULL)
```



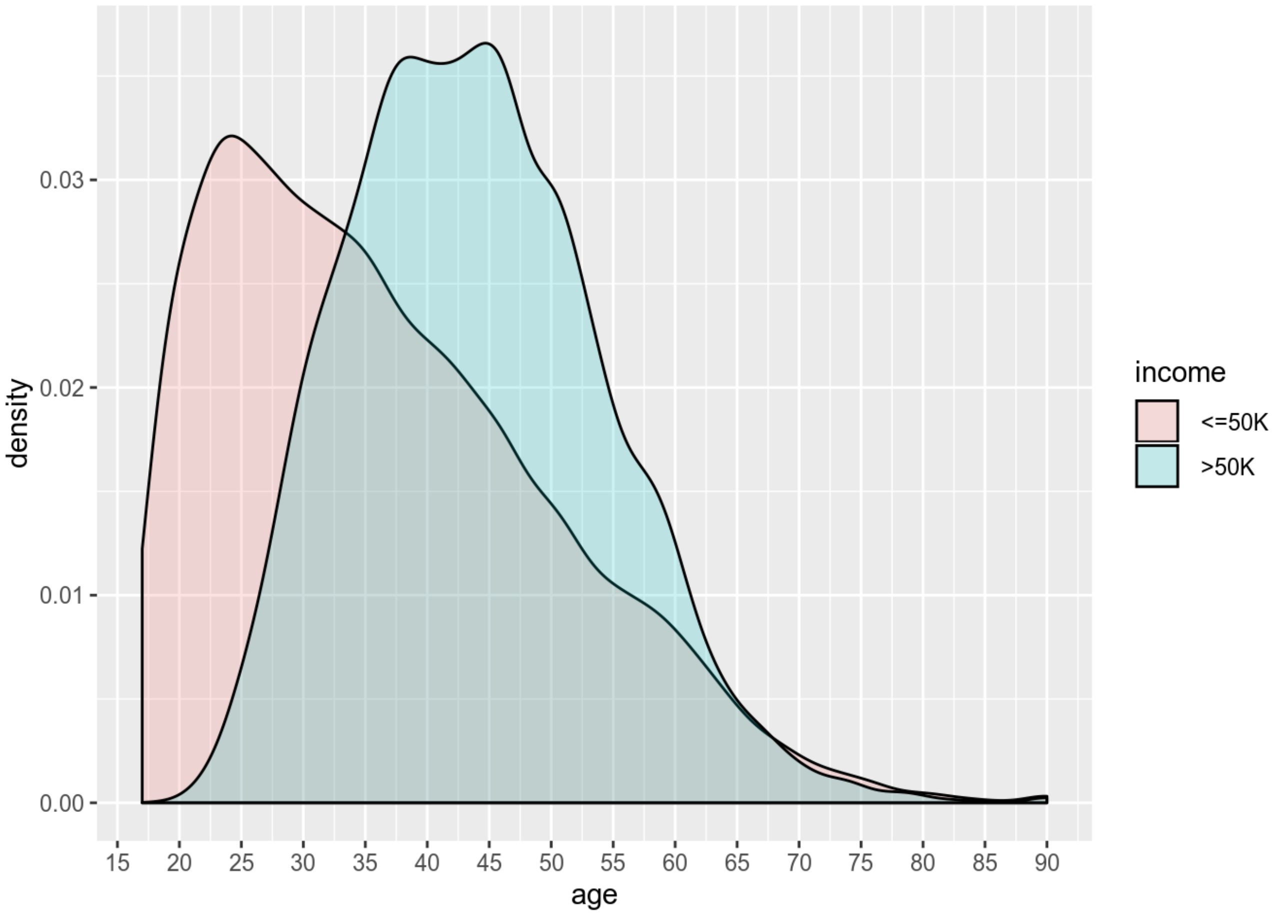
From the histogram displayed below we see that the majority of individuals are between 20 and 50 years old:

```
qplot(x = adult_data$age, data = adult_data, binwidth = 5, color = I("black"), fill = I("#
F29025"),
      xlab = "Age", ylab = "Count", main = "Histogram of Age") +
  scale_x_continuous(breaks = seq(0, 95, 5)) +
  scale_y_continuous(breaks = seq(0, 4500, 500))
```



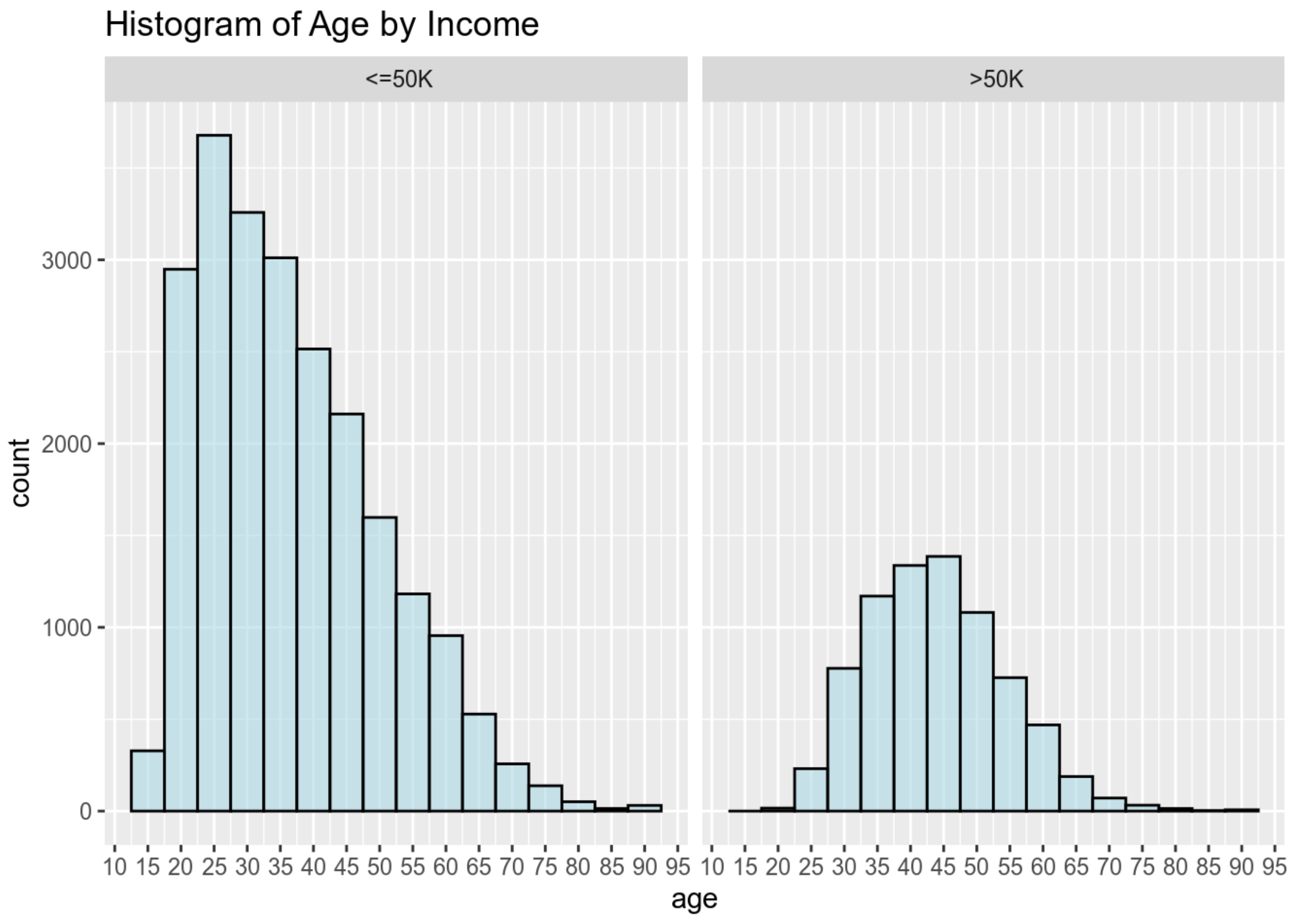
From an empirical density of age grouped by income, we see that the majority of people earning more than 50K a year are between 33 and 55 years old, whereas the greater number of people who earn less are between 18 and 45:

```
ggplot(data = adult_data, aes(age, fill = income)) +
  geom_density(alpha = 0.2) +
  scale_x_continuous(breaks = seq(0, 95, 5))
```



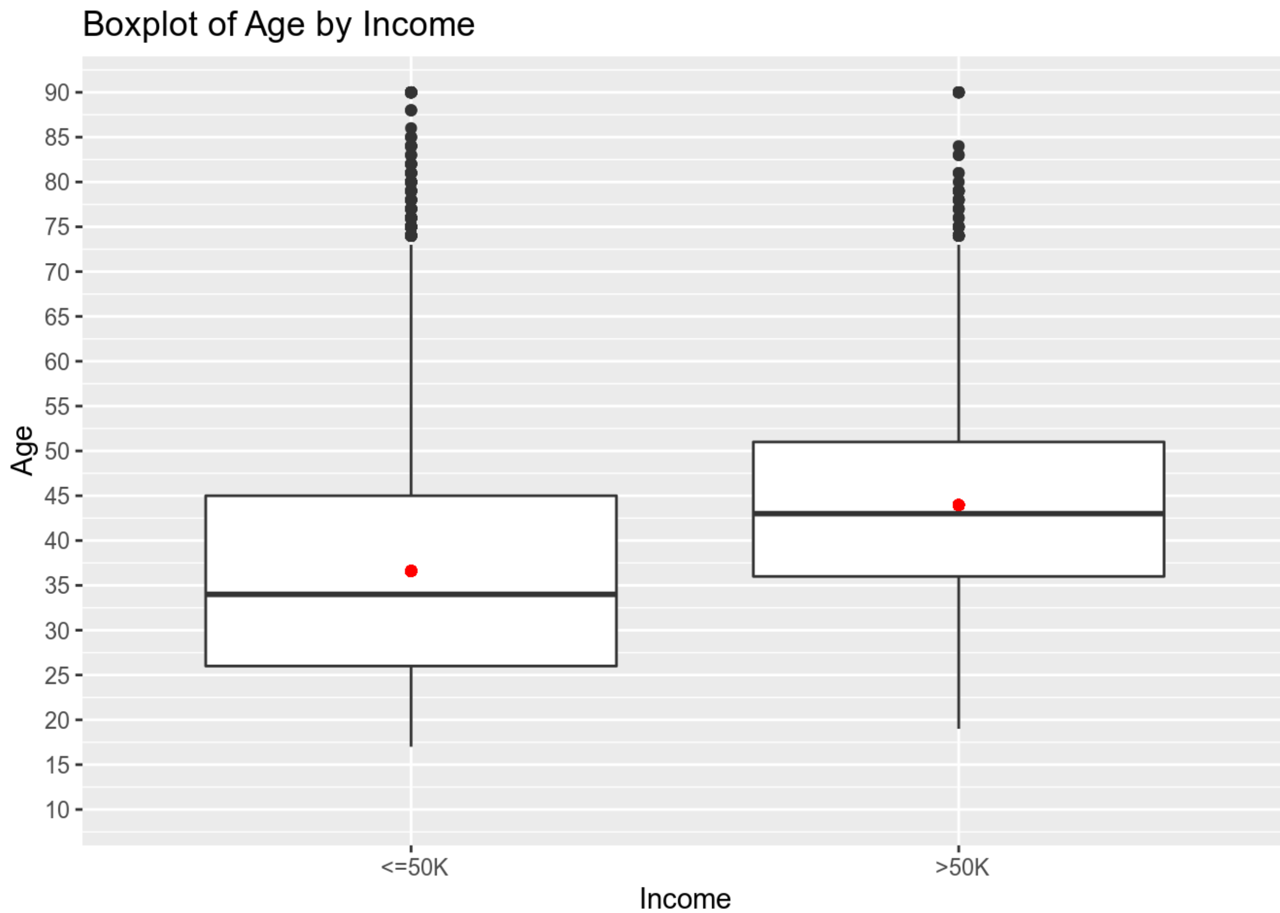
The above shows us that income and age are definitely correlated– older people have higher incomes. Further evidence of this is provided from the following histograms of age by income:

```
ggplot(data = adult_data, mapping = aes(x = age)) +
  geom_histogram(binwidth = 5, color = "black", fill = "lightblue", alpha = 0.6) +
  scale_x_continuous(breaks = seq(0, 95, 5)) +
  facet_wrap(~income) +
  ggtitle("Histogram of Age by Income")
```



The boxplot of age grouped by income:

```
ggplot(aes(x = income, y = age), data = adult_data) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 16, cex = 2, col = "red") +
  coord_cartesian(ylim = c(10, 90)) +
  scale_y_continuous(breaks = seq(10, 90, 5)) +
  ylab("Age") +
  xlab("Income") +
  ggtitle("Boxplot of Age by Income")
```



Again, we see the relationship between age and income. Check the summary statistic below:

```
summary(subset(adult_data$age, adult_data$income == " <=50K"))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   26.00   34.00   36.61   45.00   90.00
```

```
summary(subset(adult_data$age, adult_data$income == " >50K"))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   36.00   43.00   43.96   51.00   90.00
```

We notice that the first quartiles for both groups differ significantly. The first quartile for people who have an income of more than 50K is equal to 36 whereas the first quartile for people earning less than 50K equals 26. This indicates that the elder a person is, the bigger the chance of them having a higher income.

Let's take a look at one more plot to demonstrate the correlation between age and income:

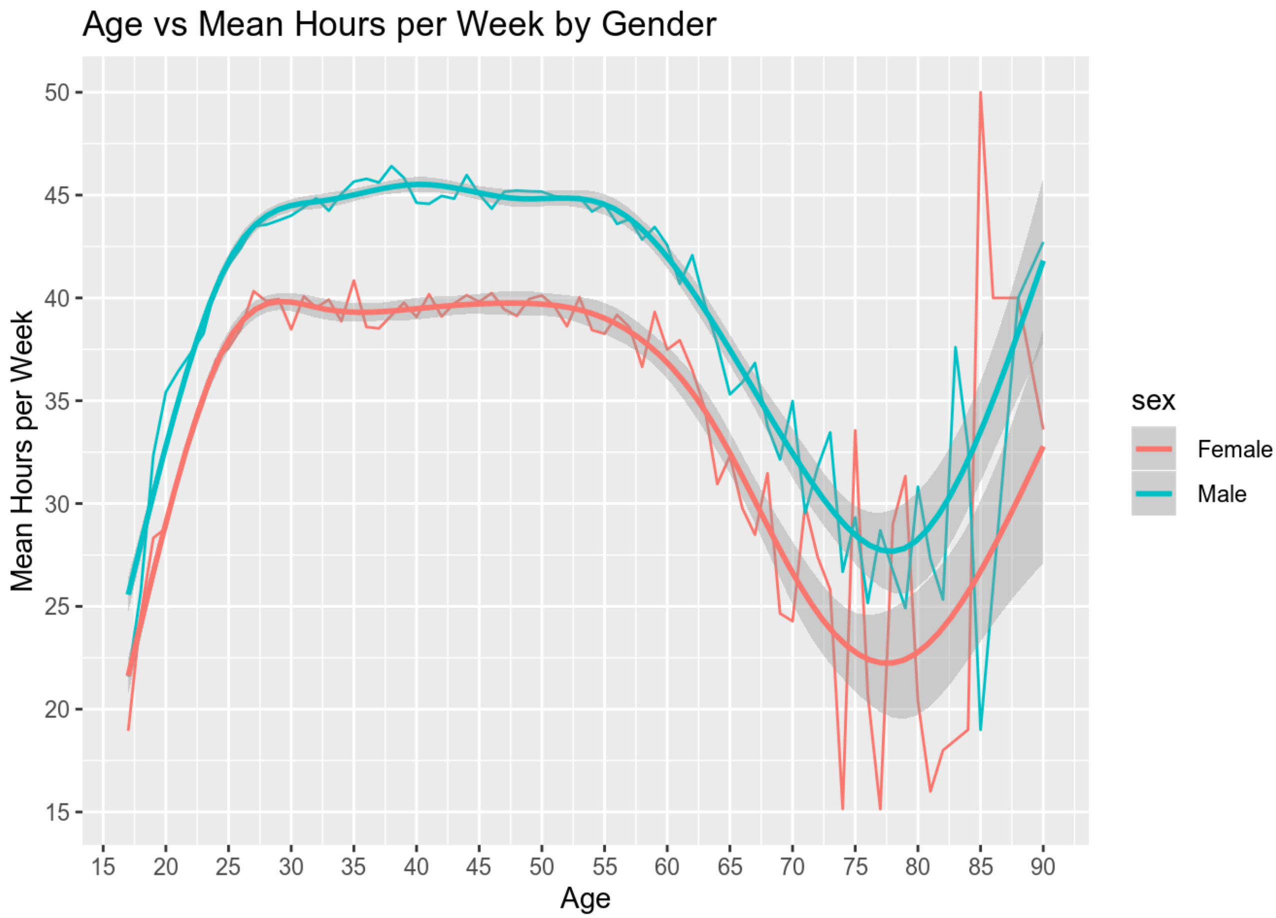
```
cd_plot(x = adult_data$age, y = adult_data$income, xlab = "Age", ylab = "Income",
        main = "Conditional Density Plot of Income versus Age")
```

The probability of having an income greater than 50K is highest for individuals in their 50s and smallest for people in their 20s.

In our plot below of mean working hours per week versus age, grouped by gender, we see that on average, men work more hours per week than women at almost all ages, an exception being for people between 77 and 80 and also between 85 and 90, where you'll find women having more average working hours per week. The mean working hours per week for women between 25 and 60 years old is 40 hours, and 45 hours for men in the same age range.

```
ggplot(aes(x = age, y = hours_per_week), data = adult_data) +
  geom_line(mapping = aes(color = sex), stat = "summary", fun.y = mean) +
  geom_smooth(mapping = aes(color = sex)) +
  scale_x_continuous(breaks = seq(10, 100, 5)) +
  scale_y_continuous(breaks = seq(0, 55, 5)) +
  labs(x = "Age", y = "Mean Hours per Week") +
  ggtitle("Age vs Mean Hours per Week by Gender")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The Variables `hours_per_week` and `hours_worked`

`hours_per_week`

Below we display a summary statistic for the variable `hours_per_week` :

```
summary(adult_data$hours_per_week)
```

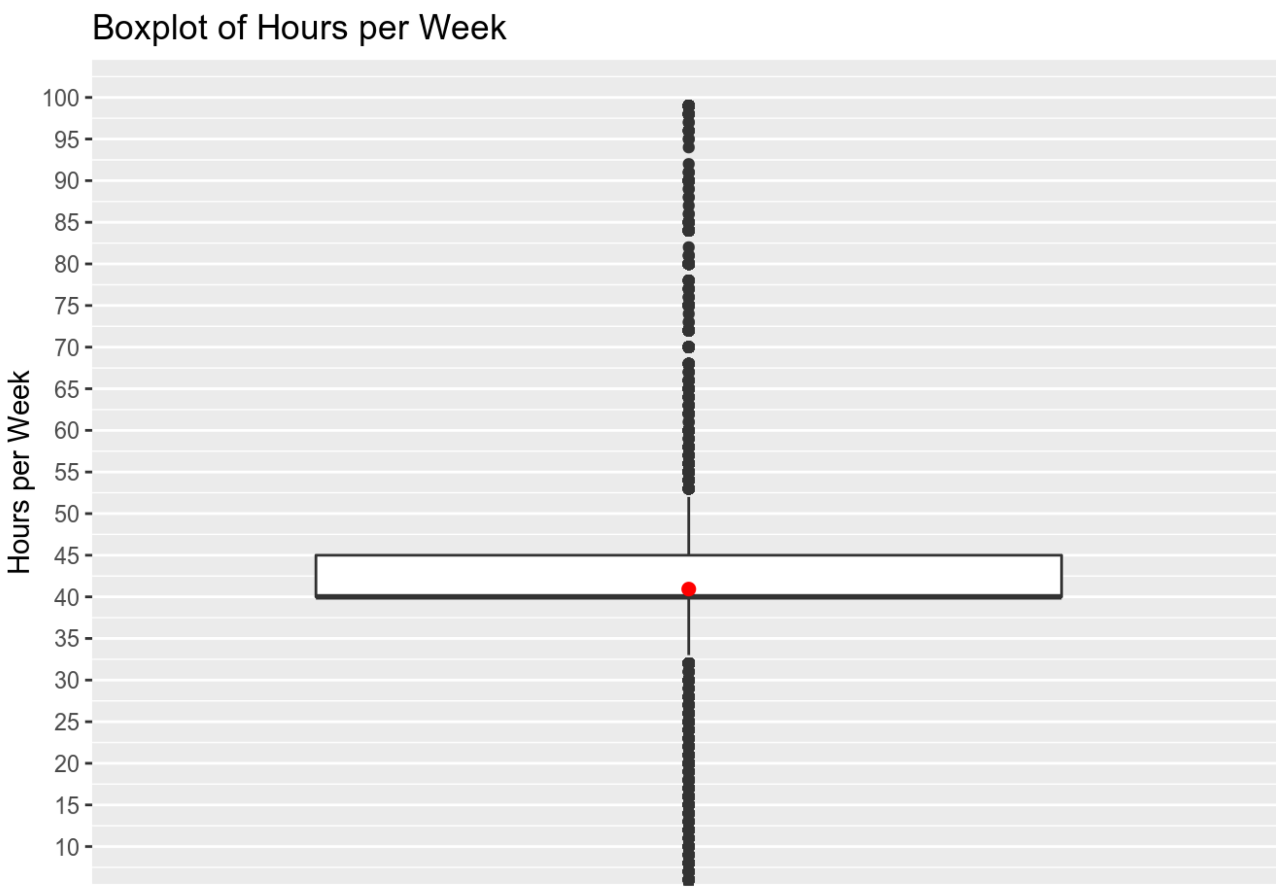
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   40.00   40.00   40.93   45.00   99.00
```

```
IQR(adult_data$hours_per_week)
```

```
## [1] 5
```

Next, we'll show a boxplot of `hours_per_week` which visualizes the summary statistic. In it, we'll see that there are many outliers:

```
ggplot(aes(x = factor(0), y = hours_per_week), data = adult_data) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 19, color = "red", cex = 2) +
  coord_cartesian(ylim = c(10, 100)) +
  scale_x_discrete(breaks = NULL) +
  scale_y_continuous(breaks = seq(10, 100, 5)) +
  ylab("Hours per Week") +
  xlab("") +
  ggtitle("Boxplot of Hours per Week")
```



Being that we're interested in the relationship between income and working hours per week, we'll go ahead and display the boxplot of hours per week grouped by income:

```
ggplot(aes(x = income, y = hours_per_week), data = adult_data) +
  geom_boxplot() +
  stat_summary(fun.y = mean, geom = "point", shape = 19, color = "red", cex = 2) +
  coord_cartesian(ylim = c(10, 100)) +
  scale_y_continuous(breaks = seq(10, 100, 10)) +
  ylab("Hours per Week") +
  xlab("Income") +
  ggtitle("Boxplot of Hours per Week by Income")
```

We can see what we would expect– the mean working hours per week is higher for people who earn more than 50K a year. Although the two medians are equal, the median coincides with the third quartile for earners under 50K a year, while the median corresponds with the first quartile for earners over 50K annually. Those exact numbers can be seen better with a summary:

```
summary(subset(adult_data$hours_per_week, adult_data$income == " <=50K"))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   38.00   40.00   39.35   40.00   99.00
```

```
summary(subset(adult_data$hours_per_week, adult_data$income == " >50K"))
```

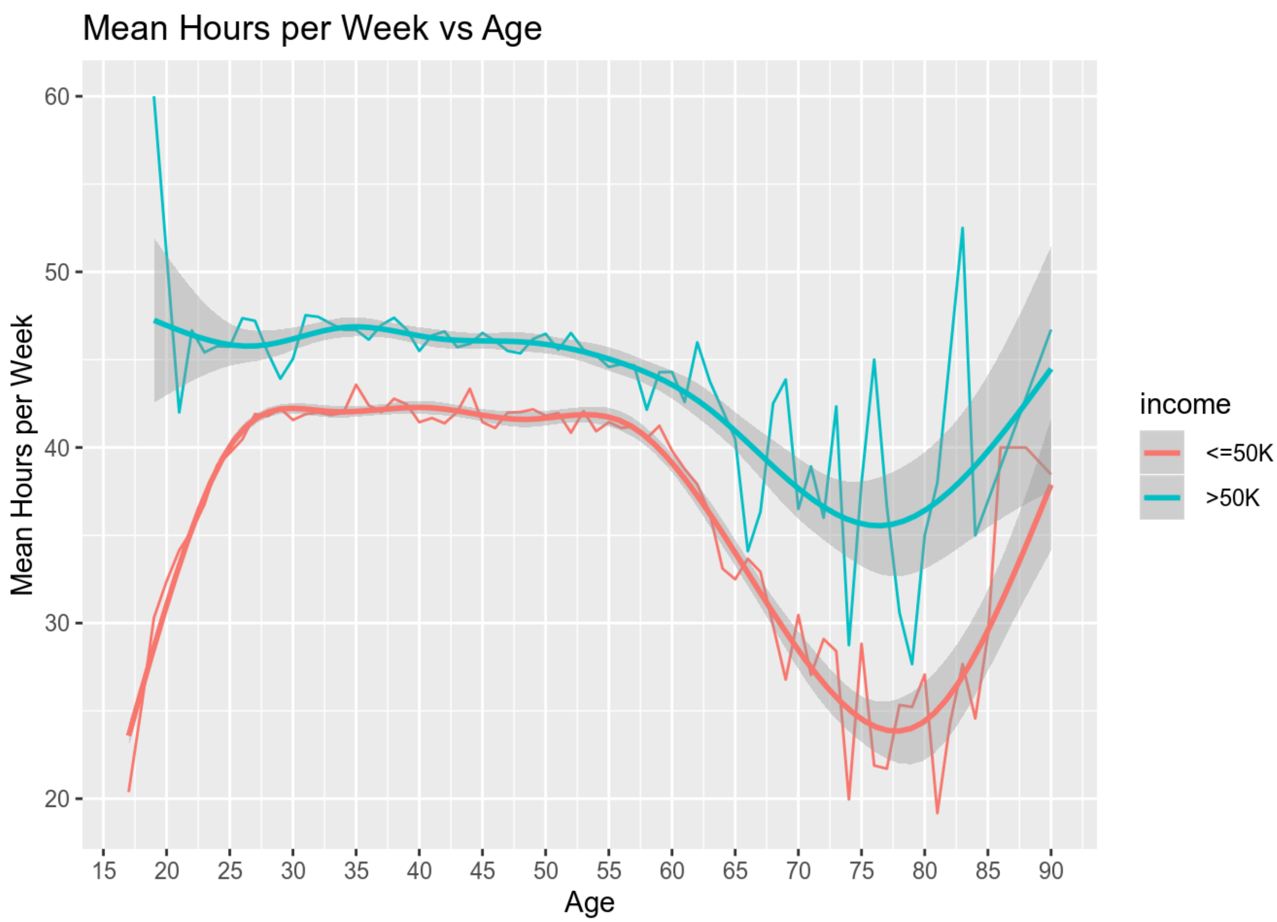
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00   40.00   40.00   45.71   50.00   99.00
```

We can see that there is a definite correlation between income and working hours per week.

Let's take a look at a graph that shows the mean working hours per week versus age, but grouped by income:

```
ggplot(mapping = aes(x = age, y = hours_per_week), data = adult_data) +
  geom_line(mapping = aes(color = income), stat = "summary", fun.y = mean) +
  geom_smooth(mapping = aes(color = income)) +
  scale_x_continuous(breaks = seq(10, 100, 5)) +
  labs(x = "Age", y = "Mean Hours per Week") +
  ggtitle("Mean Hours per Week vs Age")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



For all age groups, the mean number of working hours per week is greater for people with income greater than 50K a year.

`hours_worked`

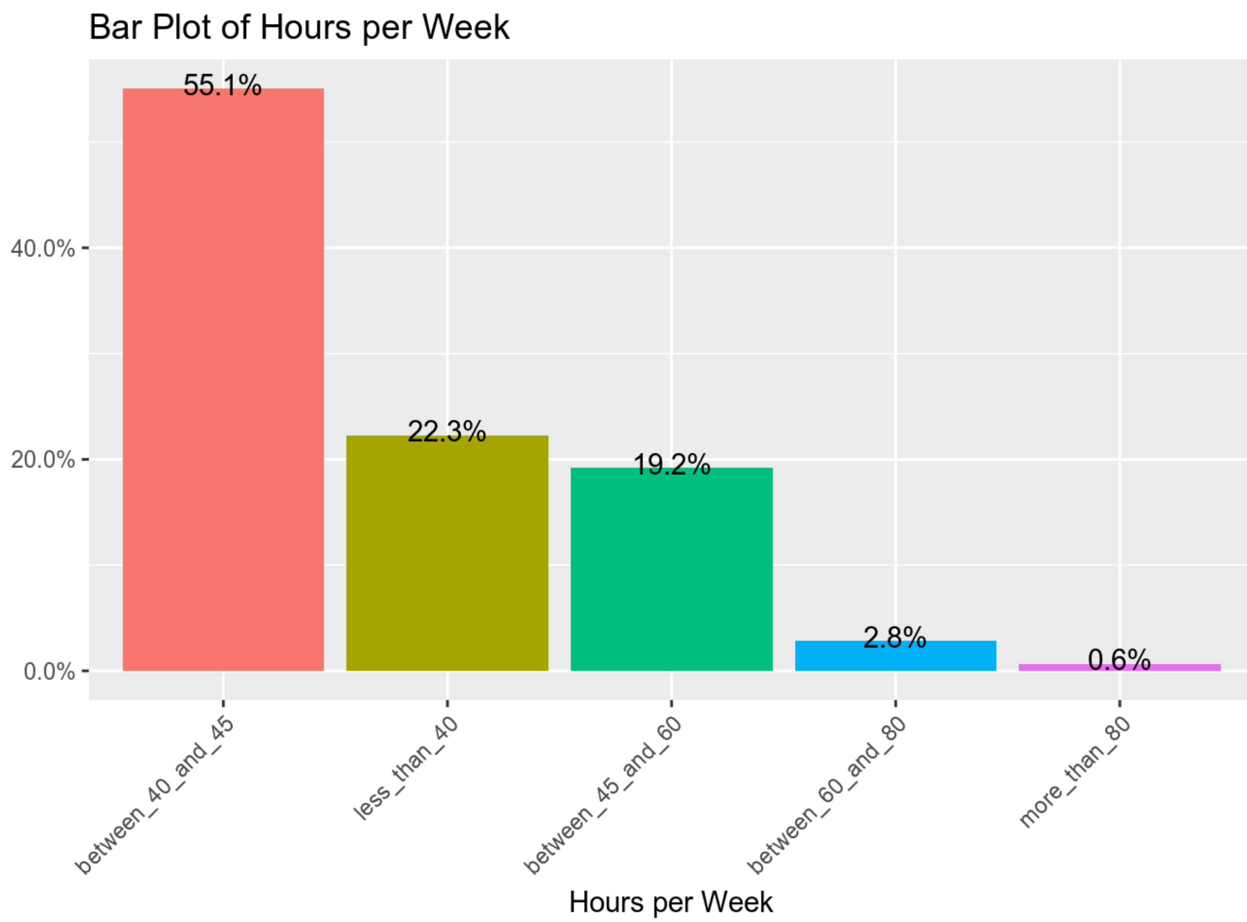
We've already determined that the majority of people work between 40 and 45 hours a week which is confirmed below:

```
summary(adult_data$hours_worked)
```

```
##  between_40_and_45  between_45_and_60  between_60_and_80 
##              16606               5790                857 
##       less_than_40       more_than_80 
##               6714                195
```

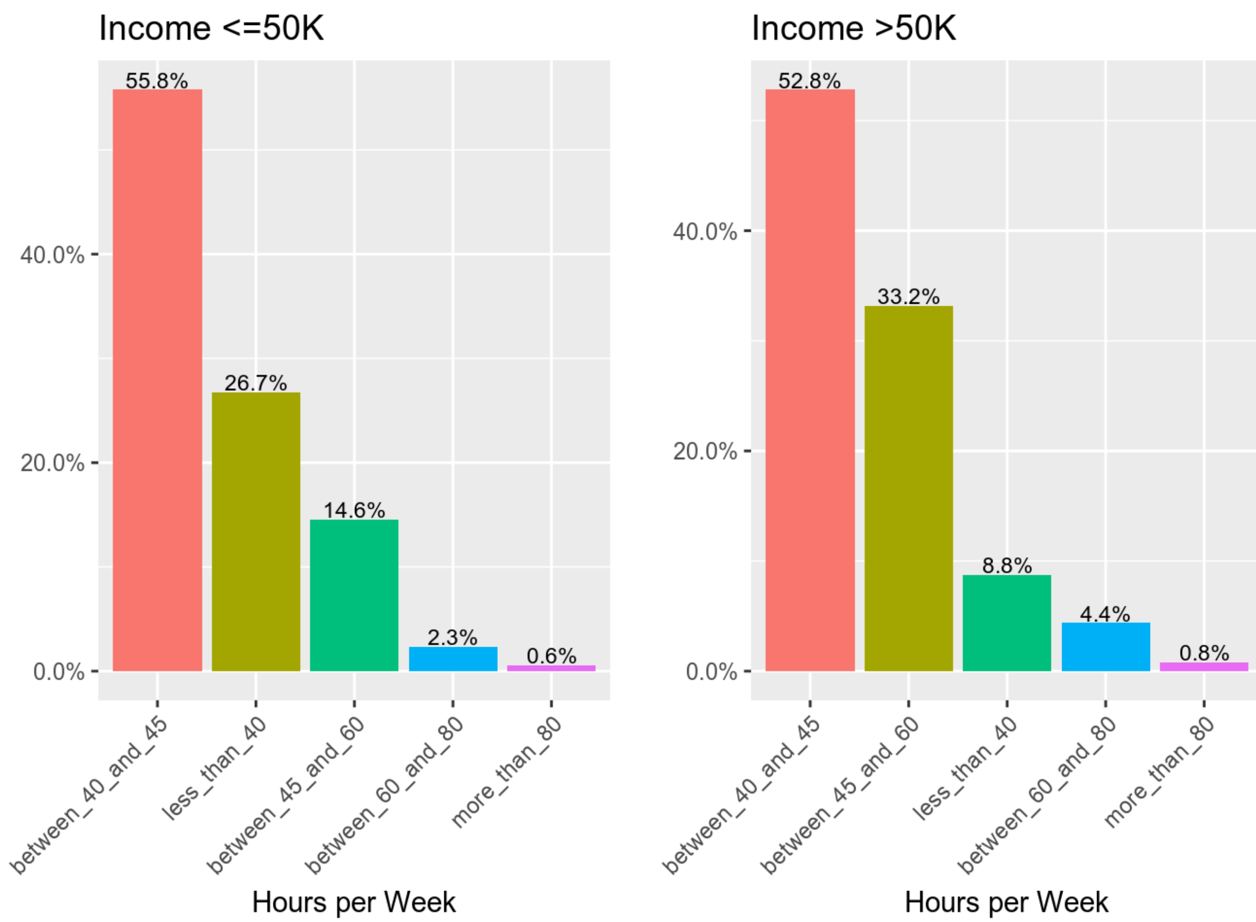
Showing the percentage of people belonging to each category of the factor variable, we give a barplot of `hours_worked`:

```
adult_data <- within(adult_data, hours_worked <- factor(hours_worked, levels = names(sort
(table(hours_worked),
                                                                                    decre
asing = TRUE))))

ggplot(adult_data, aes(x = adult_data$hours_worked, fill = adult_data$hours_worked)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = 0.3) +
  labs(x = "Hours per Week", y = "", fill = "Hours per Week") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  ggtitle("Bar Plot of Hours per Week") +
  scale_y_continuous(labels = percent)
```



Hours worked grouped by income:

```
lg_hpw <- lapply(levels(adult_data$income), function(v){
  df <- subset(adult_data, adult_data$income == v)
  df <- within(df, hours_worked <- factor(hours_worked, levels = names(sort(table(hours_wo
rked),
                                                                          decreasing = T
RUE))))
  ggplot(data = df, aes(x = hours_worked, fill = hours_worked)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = -.1, size = 3) +
    labs(x = "Hours per Week", y = "", fill = "Hours per Week") +
    theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
    ggtitle(paste("Income", v, sep = "")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_hpw, ncol = 2)
```



The proportion of people with income greater than 50K a year who work between 45 and 60 hours a week is 33.2% compared to 14.6% for that of people with income less than 50K a year.

The Variable `native_region`

We start with a summary statistic as usual:

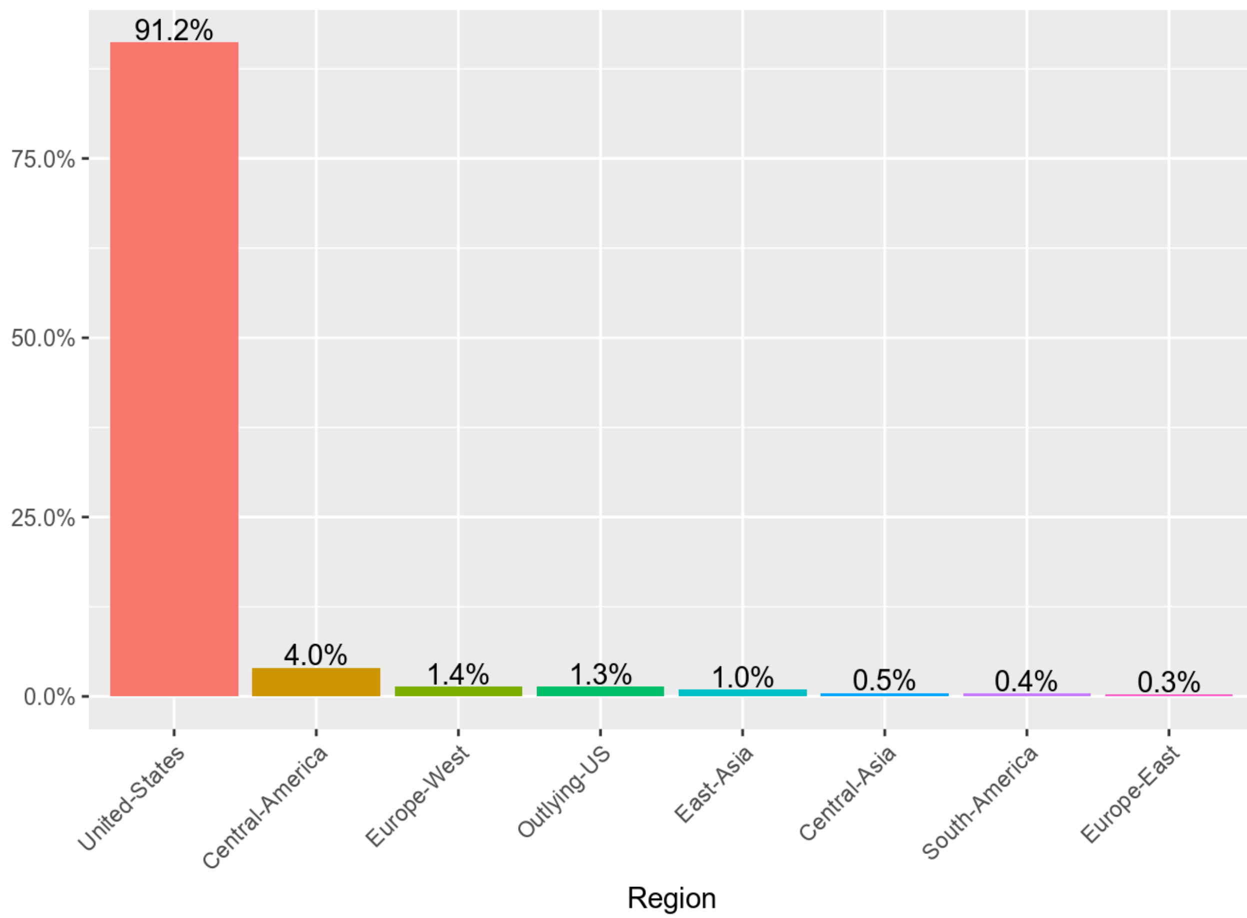
```
summary(adult_data$native_region)
```

```
##  Central-America     Central-Asia        East-Asia      Europe-East
##             1208              142              304               85
##      Europe-West    South-America    United-States      Outlying-US
##              408              113            27504              398
```

The majority of the people come from the US and Central America.

What's the percentage of people belonging to each region?

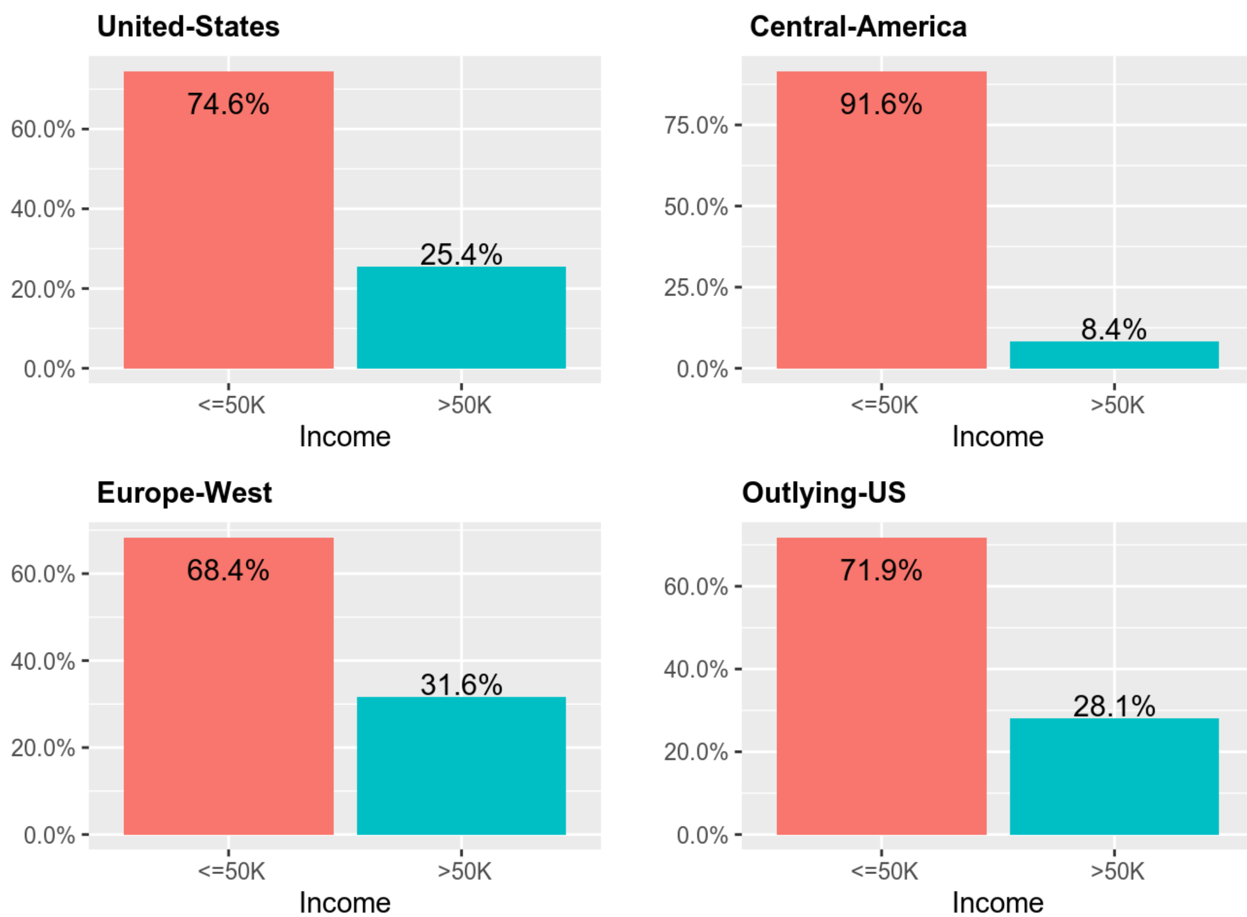
```
adult_data$native_region <- factor(adult_data$native_region,
                                   levels = names(sort(table(adult_data$native_region), de
creasing = TRUE)))

ggplot(adult_data, aes(x = adult_data$native_region, fill = adult_data$native_region)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1) +
  labs(x = "Region", y = "", fill = "Regions") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



A majority of the participants of the study come from the US. We have a small number of people from each of the other native regions leading to random samples which might not be representative for the respective population, any further analysis must be carried out with caution.

We display the percentage of people earning less than 50K and more than 50K annually among all individuals belonging to a given native region:

```
lp_region <- lapply(levels(adult_data$native_region), function(v){
  df <- subset(adult_data, adult_data$native_region == v)
  ggplot(data = df, aes(x = income, fill = income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / 
sum(..count..)),
              stat = "count", vjust = c(2, -0.1), size = 4) +
    labs(x = "Income", y = "", fill = "Income" ) +
    ggtitle(v) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lp_region[1:4], ncol = 2)
```




```
grid.arrange(grobs = lp_region[5:8], ncol = 2)
```

The highest percentage of people earning greater than 50K annually among all native regions are those with Central Asian origin at 40.8%, however, we can't completely rely on this inference since this is coming from only 142 observations compared to 27,504 people with US origin. If we disregard the small number of observations from the rest of of the non US regions, these results indicate that `income` is dependent on `native_region`.

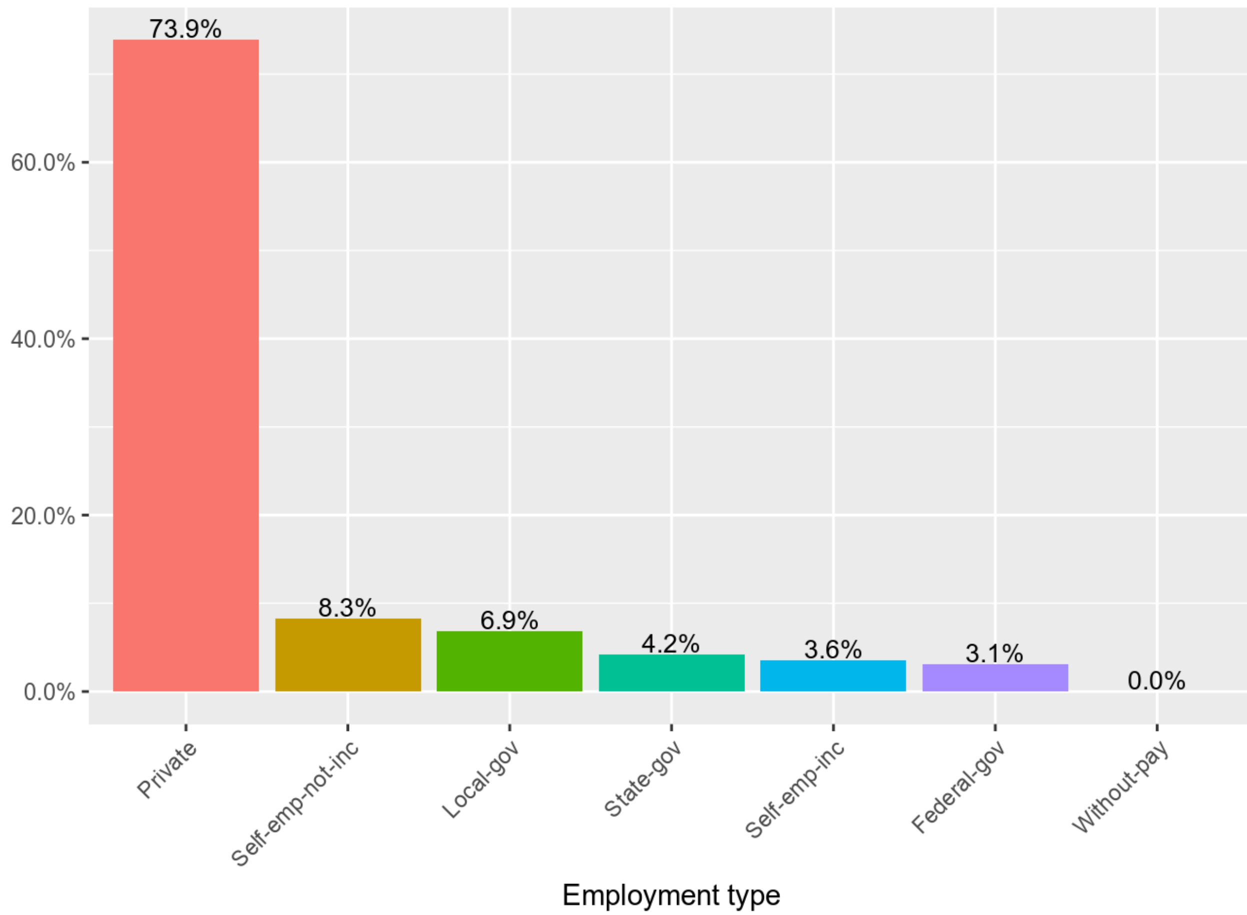
The Variable `workclass`

```
table(adult_data$workclass)
```

```
##
##      Federal-gov        Local-gov          Private     Self-emp-inc
##              943             2067            22286             1074
## Self-emp-not-inc        State-gov      Without-pay
##             2499             1279               14
```

The majority of people are employed in the private sector. We show a graph displaying the percentage of people belonging to each category of `workclass`:

```
adult_data$workclass <- factor(adult_data$workclass, levels = names(sort(table(adult_data
$workclass),
                                                                          decreasing = TRUE
)))

ggplot(adult_data, aes(x = adult_data$workclass, fill = adult_data$workclass)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1, size = 3.5) +
  labs(x = "Employment type", y = "", fill = "Employment type") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



We observe that there are no people in the category “Never-worked”:

```
nrow(subset(adult_data, adult_data$workclass == "Never-worked"))
```

```
## [1] 0
```

Obviously there are no people in the category “Without-pay” who earn more than 50K a year…

```
nrow(subset(adult_data, adult_data$workclass == "Without-pay" & adult_data$income == " >50
K"))
```

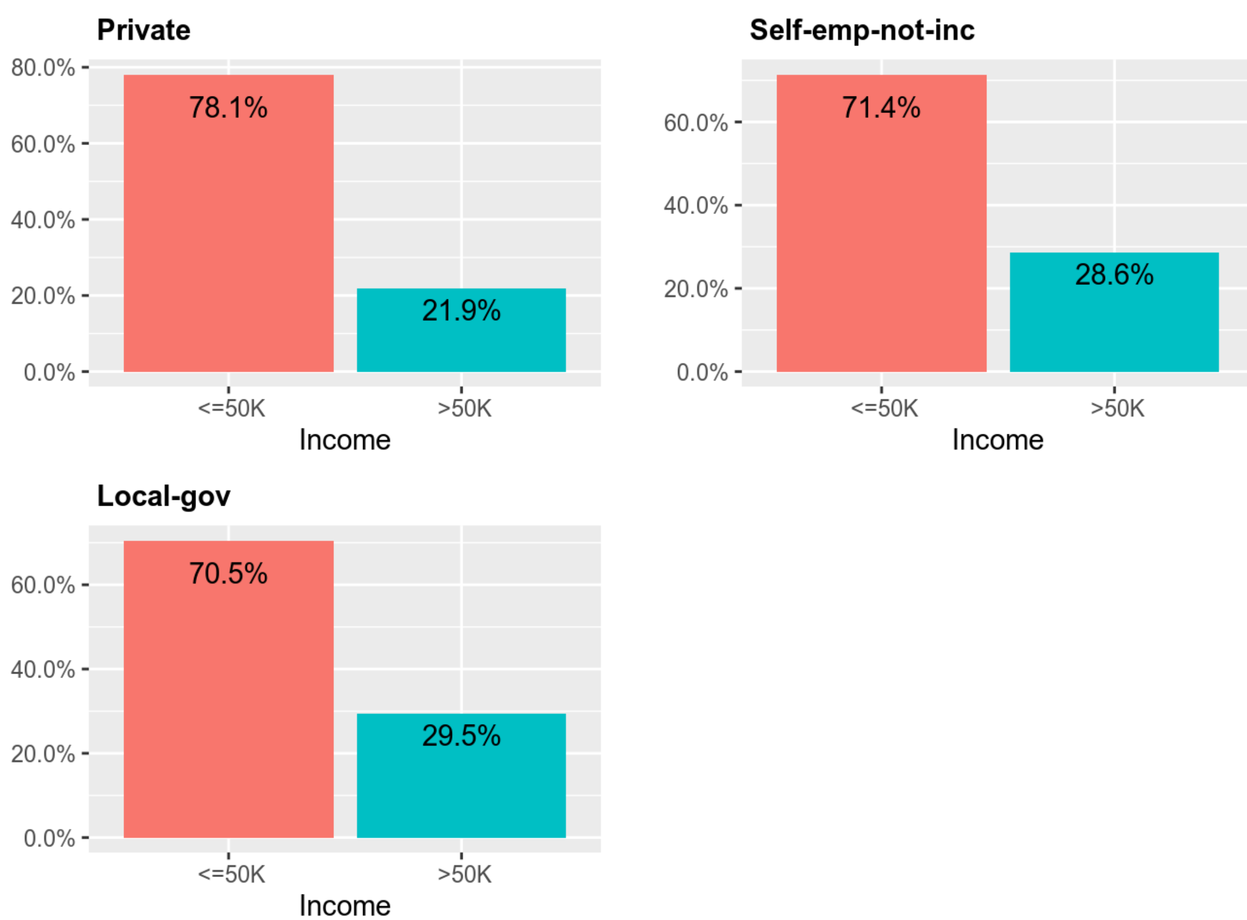
```
## [1] 0
```

…so making the plots more readable and meaningful, we'll exclude those factor levels…

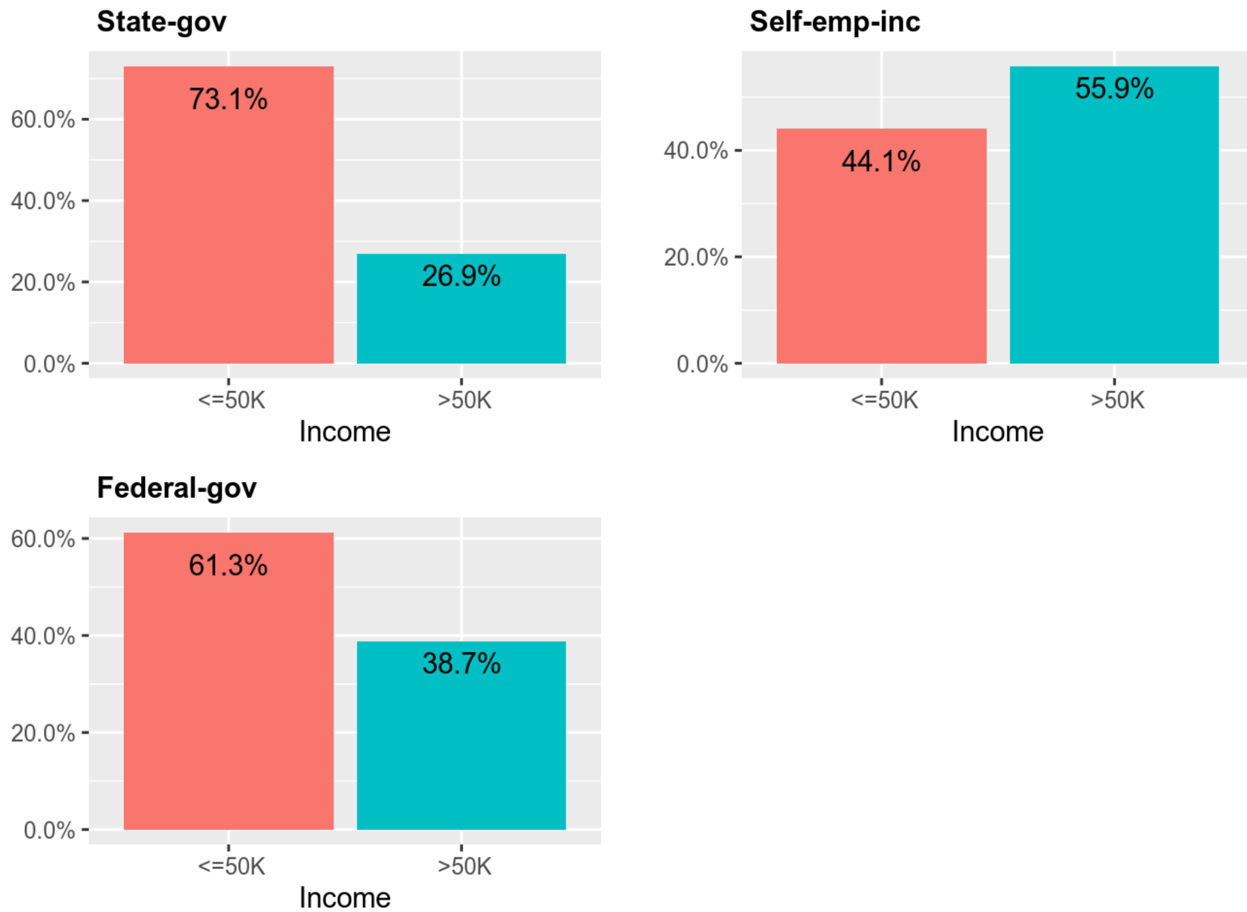
```
modified_work <- levels(adult_data$workclass)
modified_work <- modified_work[!is.element(modified_work, c("Never-worked", "Without-pay"
))]
```

… and plot the percentage of people earning less than and more than 50K annually based on their employment status:

```
lg_workclass_mod <- lapply(modified_work, function(v){
  ggplot(data = subset(adult_data, adult_data$workclass == v),
         aes(x = subset(adult_data, adult_data$workclass == v)$income,
             fill = subset(adult_data, adult_data$workclass == v)$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = c(2, 1.5)) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(v) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_workclass_mod[1:3], ncol = 2)
```



```
grid.arrange(grobs = lg_workclass_mod[4:6], ncol = 2)
```



We see that the percentage of individuals having an income of more than 50K is highest for the category "Self-emp-inc" (Self employed with income) at 55.9% followed by federal government employees. There is a relationship between the variables `income` and `workclass`.

The Variable `education`

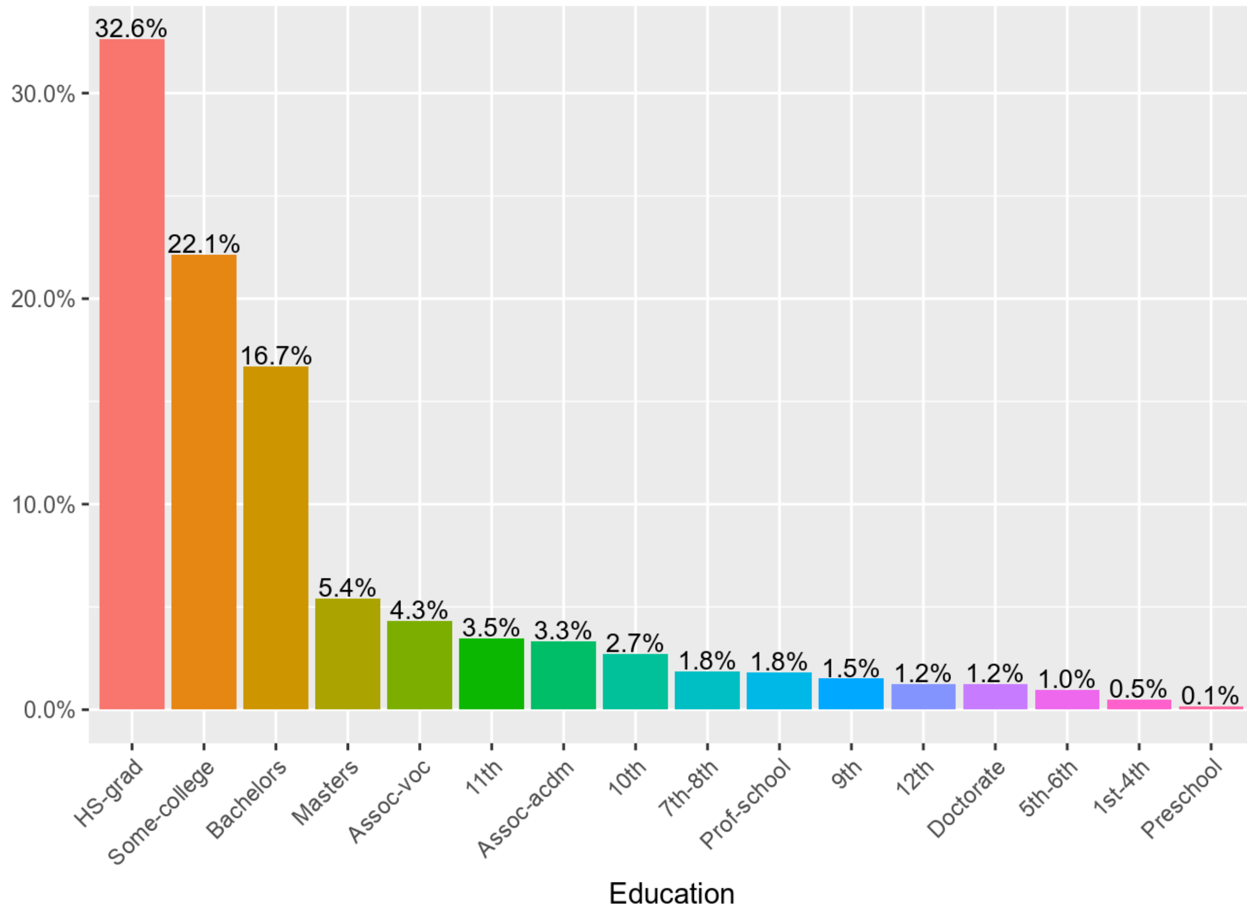
We start with a summary of the `education` variable:

```
summary(adult_data$education)
```

```
##         10th         11th         12th      1st-4th      5th-6th 
##          820         1048          377          151          288 
##      7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors 
##          557          455         1008         1307         5044 
##    Doctorate      HS-grad      Masters    Preschool  Prof-school 
##          375         9840         1627           45          542 
## Some-college 
##         6678
```

The percentage of people belonging to each category of education is displayed below:

```
adult_data$education <- factor(adult_data$education,
                               levels = names(sort(table(adult_data$education), decreasing
= TRUE)))

ggplot(adult_data, aes(x = adult_data$education, fill = adult_data$education)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1, size = 3.5) +
  labs(x = "Education", y = "", fill = "Education") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



What!? No people with only a Preschool education earning more than 50K a year?

```
nrow(subset(adult_data, adult_data$education == " Preschool" & adult_data$income == " >=50
K"))
```

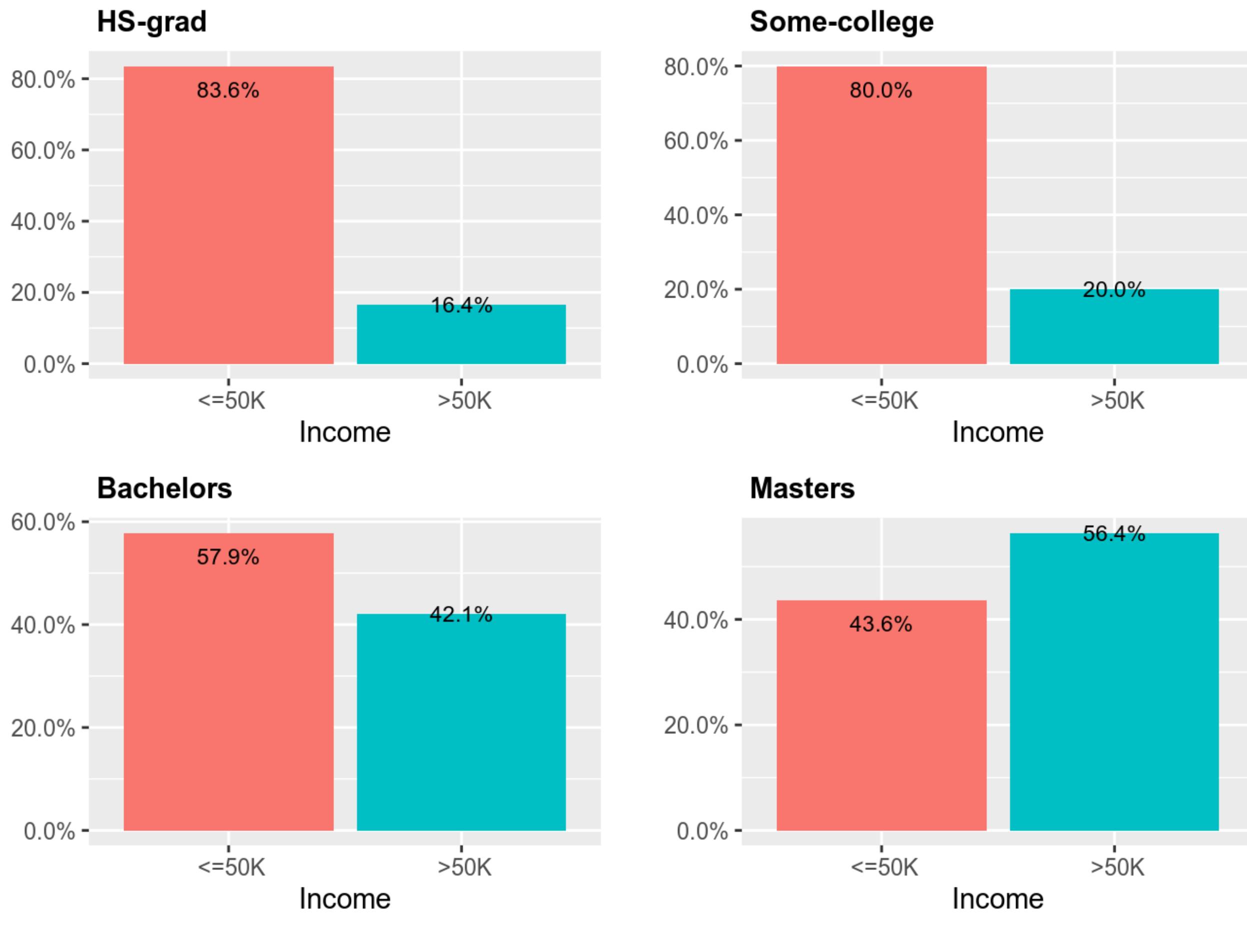
```
## [1] 0
```

Then we shall reluctantly remove the the factor level “Preschool” before we continue further:

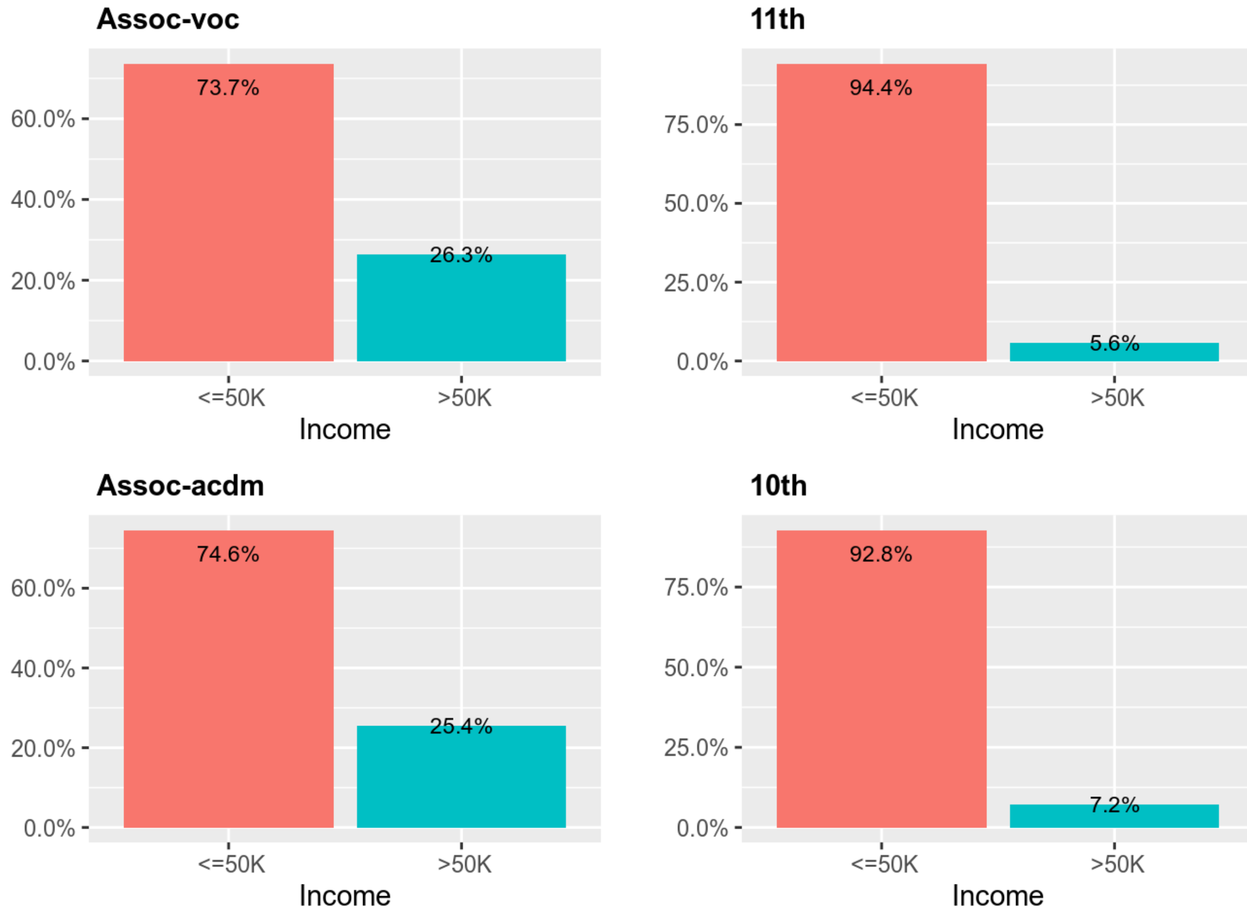
```
modified_edu <- levels(adult_data$education)
modified_edu <- modified_edu[!is.element(modified_edu, " Preschool")]
```

We display the barplot of each education category grouped by income:

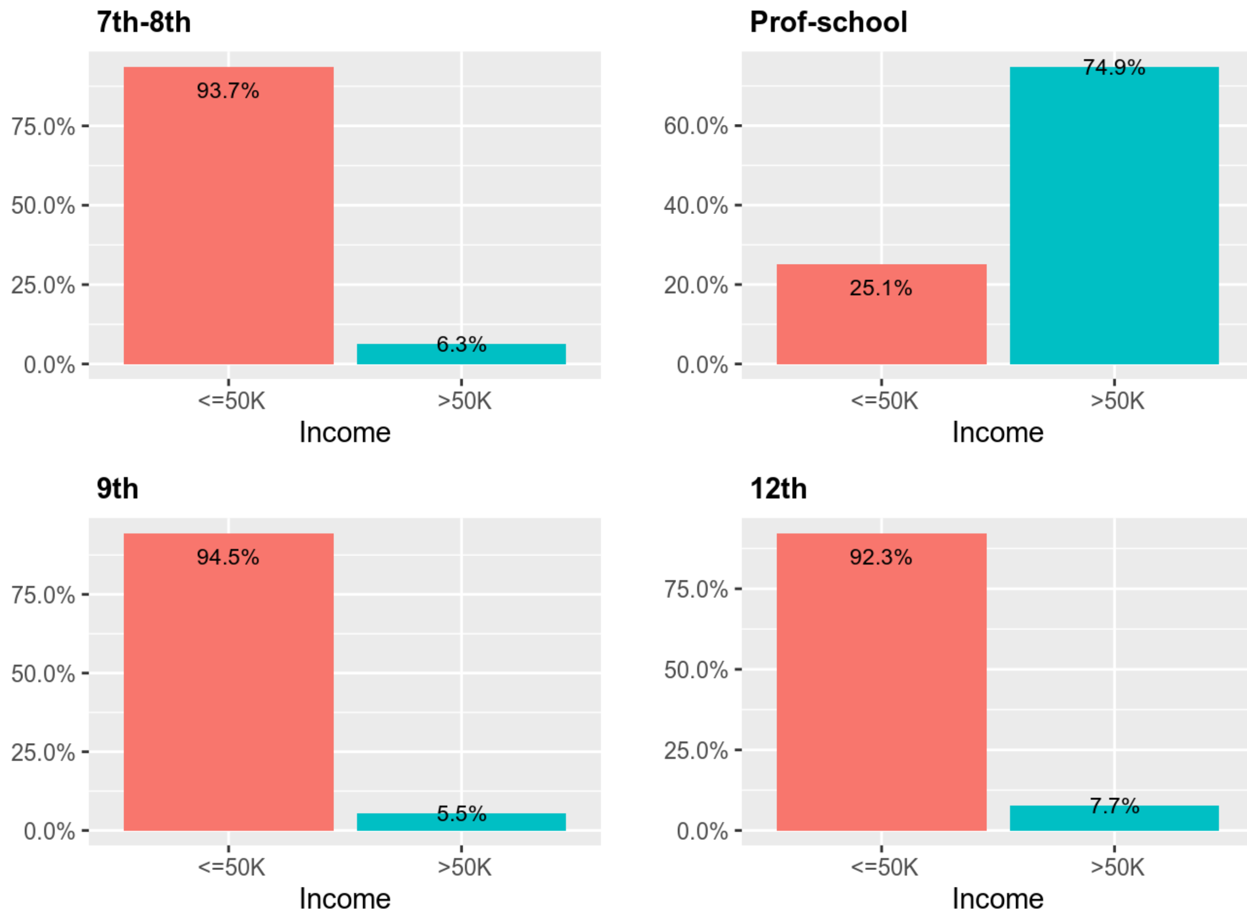
```
lg_mod_edu <- lapply(modified_edu, function(v){
  ggplot(data = subset(adult_data, adult_data$education == v),
         aes(x = subset(adult_data, adult_data$education == v)$income,
             fill = subset(adult_data, adult_data$education == v)$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = c(2, 0.5), size = 3) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(v)+
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_mod_edu[1:4], ncol = 2)
```



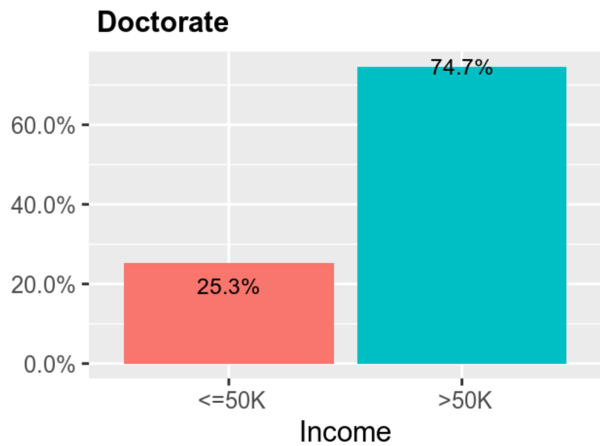

```
grid.arrange(grobs = lg_mod_edu[5:8], ncol = 2)
```

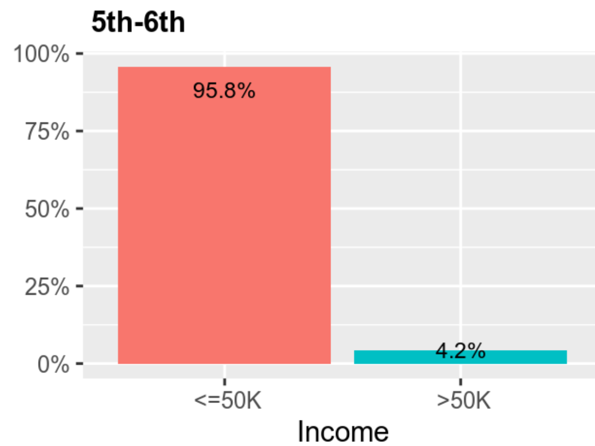


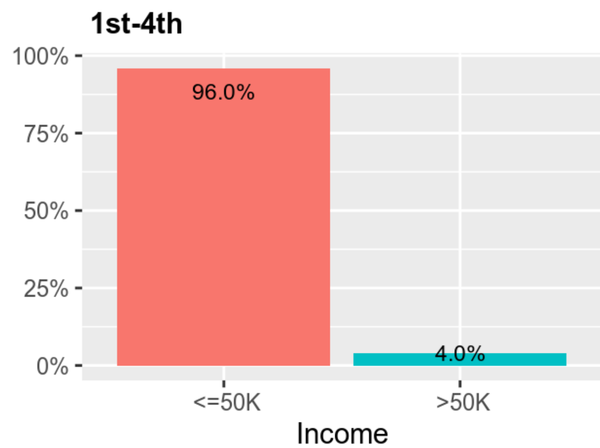

```
grid.arrange(grobs = lg_mod_edu[9:12], ncol = 2)
```




```
grid.arrange(grobs = lg_mod_edu[13:15], ncol = 2)
```

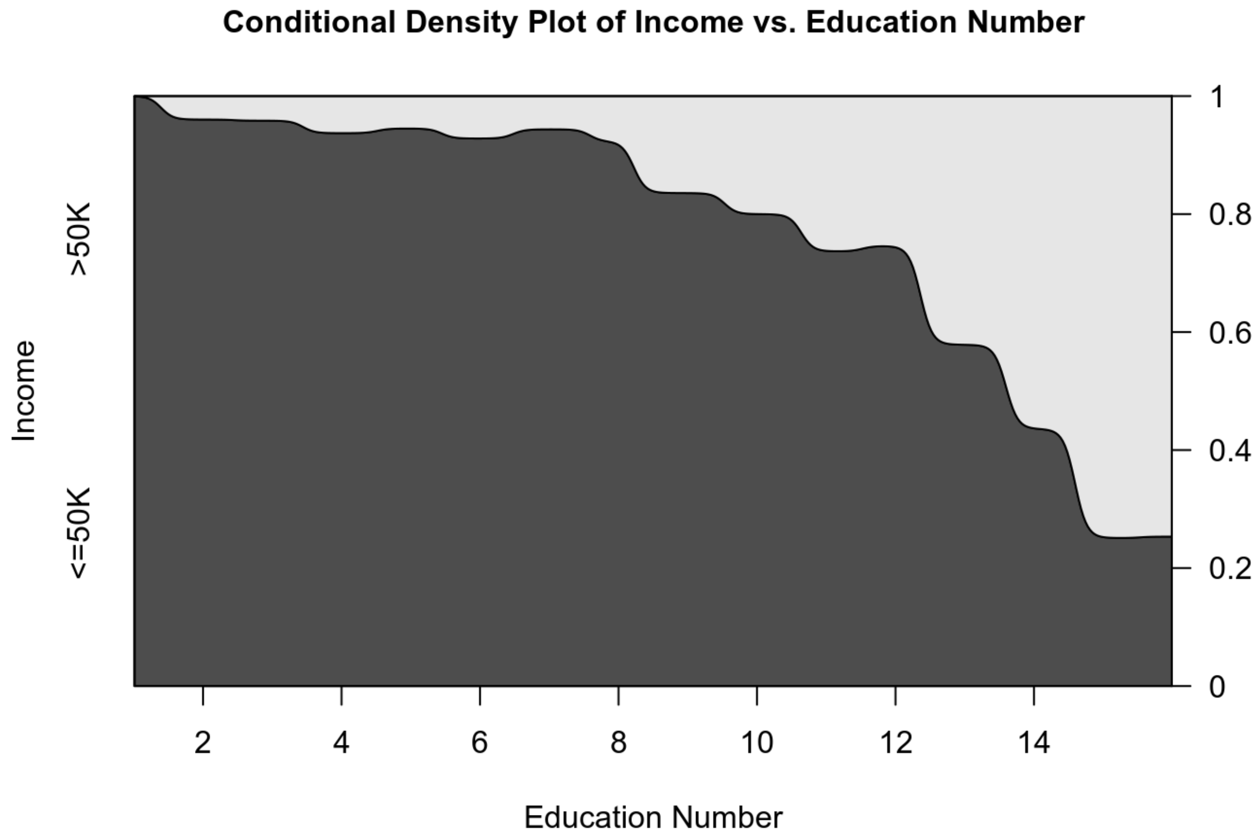






Folks that have obtained only a primary education have a very small percentage of people earning incomes greater than 50K annually. The biggest percentage of employees who have an annual income higher than 50K is 74.9% belonging to those who've attained "Prof-school" educations. They are followed by "Doctorates", "Masters", and "Bachelors". We can see that there is a relationship between `education` and `income`. To demonstrate this correlation visually, we'll display a conditional density plot of income versus education number:

```
cd_plot(x = adult_data$education_num, y = adult_data$income, xlab = "Education Number", yl
ab = "Income",
        main = "Conditional Density Plot of Income vs. Education Number")
```



Each number (1-16) in this integer variable corresponds to an education level from the factor variable `education`, starting from the lowest level ("Preschool") and reaching the highest education level ("Doctorate"). The higher the education level, the greater the probability of earning more than 50K annually.

The Variable `marital_status`

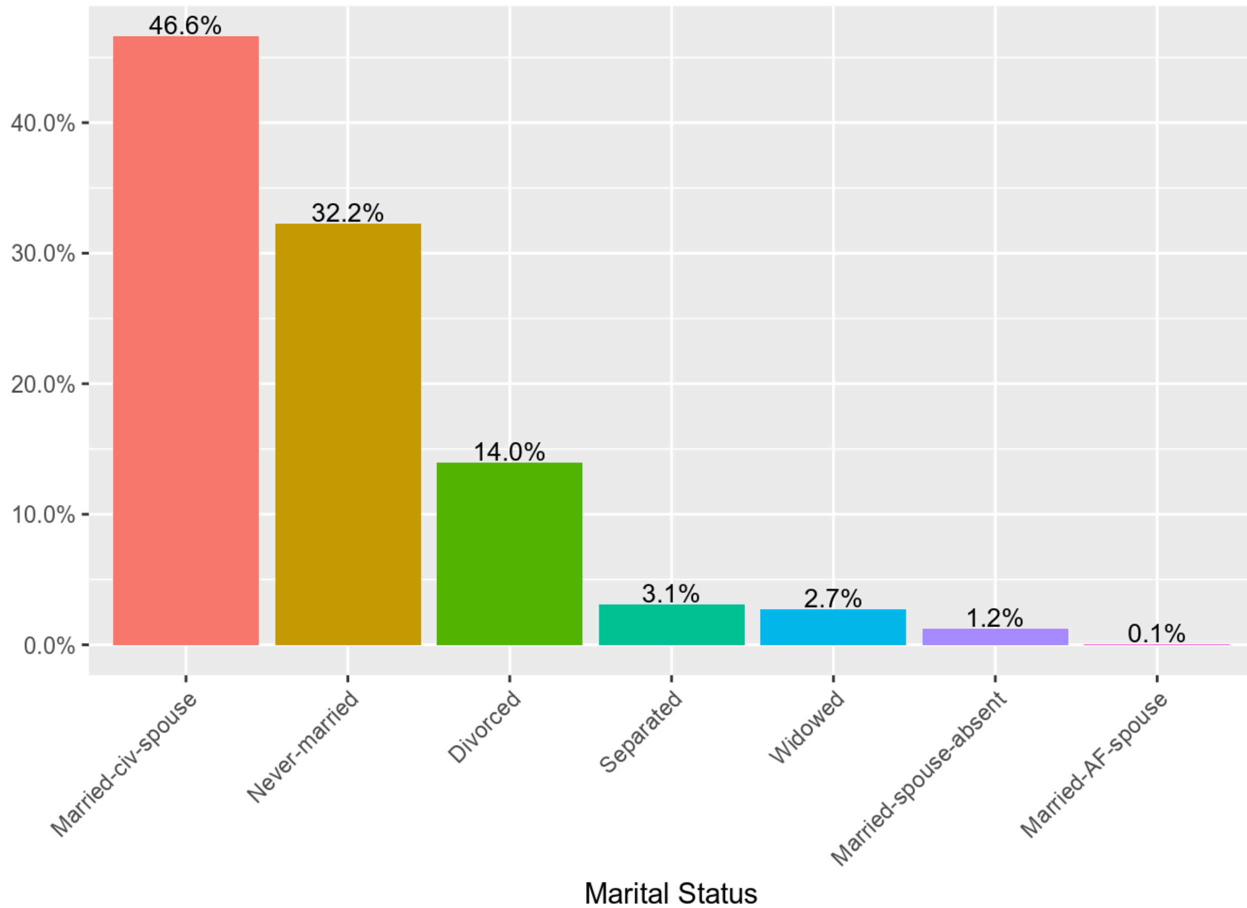
How many people belong to each category of the variable `marital_status`?

```
summary(adult_data$marital_status)
```

```
##              Divorced     Married-AF-spouse    Married-civ-spouse
##                  4214                    21                 14065
##  Married-spouse-absent         Never-married             Separated
##                   370                  9726                   939
##               Widowed
##                   827
```

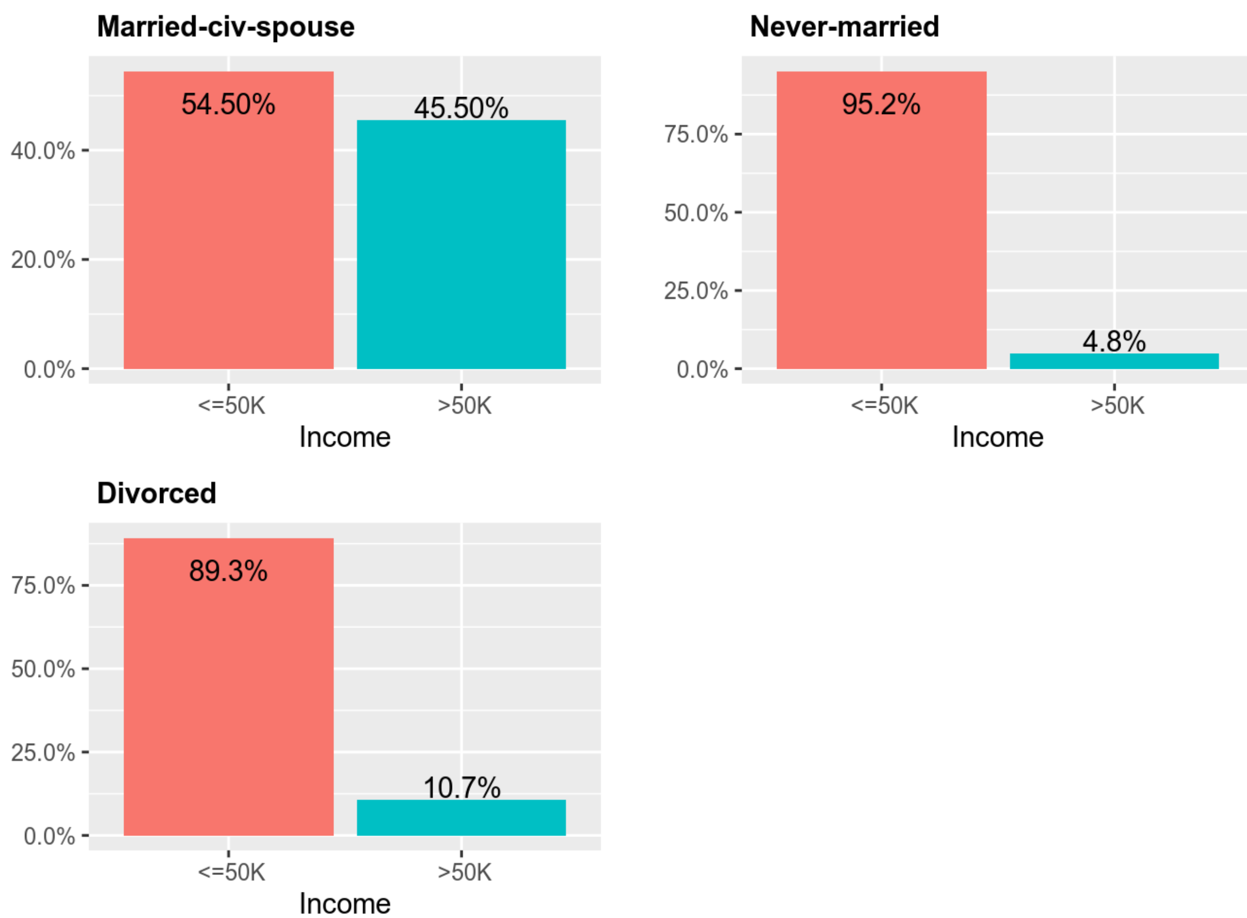
Let's visualize the percentage of people belonging to each category…

```
adult_data$marital_status <- factor(adult_data$marital_status,
                                    levels = names(sort(table(adult_data$marital_status),
decreasing = TRUE)))

ggplot(adult_data, aes(x = adult_data$marital_status, fill = adult_data$marital_status)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1, size = 3.5) +
  labs(x = "Marital Status", y = "", fill = "Marital Status") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



…and give barplots of income grouped by marital status:

```
lp_marital <- lapply(levels(adult_data$marital_status), function(v){
  ggplot(data = subset(adult_data, adult_data$marital_status == v),
         aes(x = subset(adult_data, adult_data$marital_status == v)$income,
             fill = subset(adult_data, adult_data$marital_status == v)$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = c(2, -0.1)) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(v) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lp_marital[1:3], ncol = 2)
```




```
grid.arrange(grobs = lp_marital[4:7], ncol = 2)
```

The largest percentage of people with income greater than 50K annually come from those belonging to the category “Married-AF-spouse”, but there are only 21 observations from this category so we cannot draw any trustworthy conclusions from this data. However, the category “Married-civ-spouse” has 14,065 observations, so it can be considered representative and the percentage of people within this category with income greater than 50K annually is relatively high at 45.5%. The same cannot be said for categories like “Divorced”, “Never-married”, “Married-spouse-absent”, “Separated”, and “Widowed”. One explanation as to why people who never got married earn less than married people is that those that belong to the category “Never-married” probably are young individuals who work part-time, including younger people as a whole who are in the beginning of their professional careers. This conclusion seems to be in agreement with the variable `age`, where we saw that the older an individual is, the higher the likelihood of earning over 50K annually. There is a correlation between income and marital status and it cannot be explanined only with the confounding `age` variable.

The Variable `occupation`

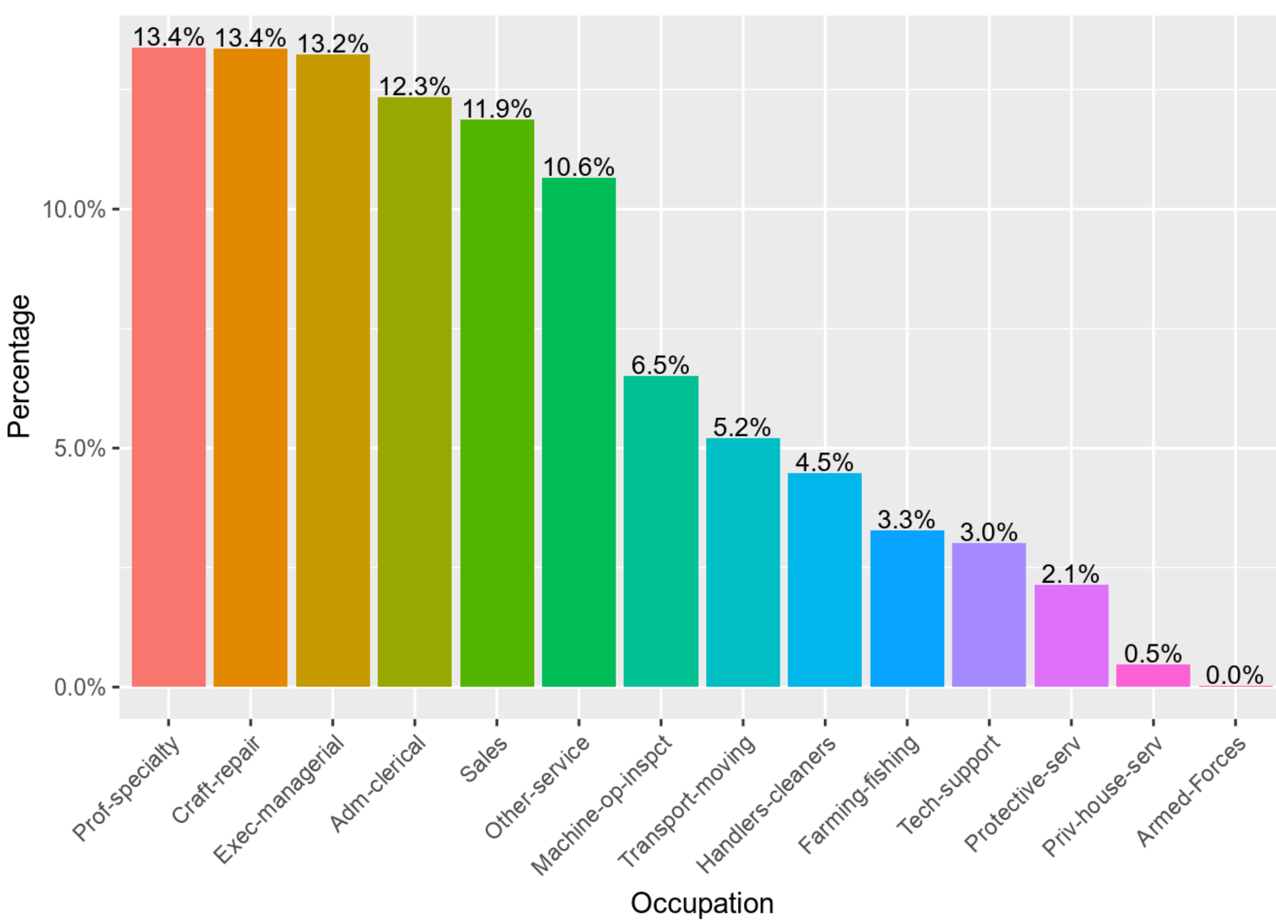
First, we show a summary statistic of `occupation`:

```
summary(adult_data$occupation)
```

```
##       Adm-clerical       Armed-Forces       Craft-repair
##               3721                  9               4030
##    Exec-managerial    Farming-fishing  Handlers-cleaners
##               3992                989               1350
##  Machine-op-inspct      Other-service    Priv-house-serv
##               1966               3212                143
##     Prof-specialty    Protective-serv              Sales
##               4038                644               3584
##       Tech-support   Transport-moving
##                912               1572
```

Then, we visualize the percentage of people belonging to each category of the factor variable `occupation` :

```
adult_data$occupation <- factor(adult_data$occupation, levels = names(sort(table(adult_dat
a$occupation),
                                                                          decreasing = TR
UE)))

ggplot(adult_data, aes(x = adult_data$occupation, fill = adult_data$occupation)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1, size = 3.5) +
  labs(x = "Occupation", y = "Percentage", fill = "Occupation") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



There are no women working in the category “Armed-Forces” and there’s no men working in the “Priv-house-serv” sector making more than 50K annually:

```
nrow(subset(adult_data, adult_data$sex == " Female" & adult_data$occupation == " Armed-For
ces"))
```

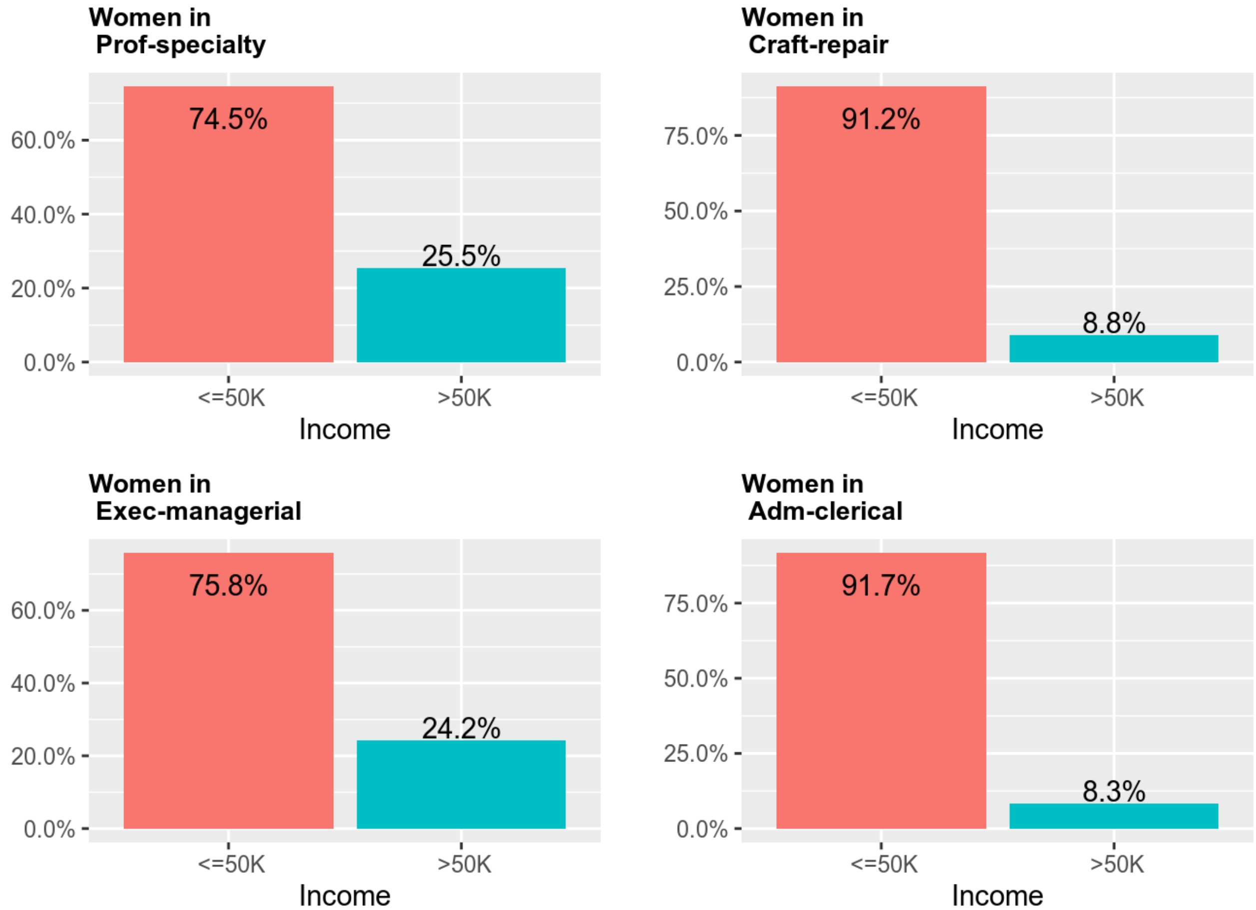
```
## [1] 0
```

```
nrow(subset(adult_data, adult_data$sex == " Male" & adult_data$occupation == " Priv-house-
serv" &
              adult_data$income == " >50K"))
```

```
## [1] 0
```

We exclude “Armed-Forces” for the following barplots with percentages of women earning less than 50K and more than 50K annually for each type of occupation:

```
modified_occup_f <- levels(adult_data$occupation)
modified_occup_f <- modified_occup_f[!is.element(modified_occup_f, c(" Armed-Forces"))]

lp_occupation_f <- lapply(modified_occup_f, function(v){
  ggplot(data = subset(adult_data, adult_data$occupation == v & adult_data$sex == " Femal
e"),
         aes(x = subset(adult_data, adult_data$occupation == v & adult_data$sex == " Femal
e")$income,
             fill = subset(adult_data, adult_data$occupation == v & adult_data$sex == " Fe
male")$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = c(2, -0.1)) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(paste("Women in \n", v, sep = "")) +
    theme(legend.position = "none", plot.title = element_text(size = 10, face = "bold")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lp_occupation_f[1:4], ncol = 2)
```




```
grid.arrange(grobs = lp_occupation_f[5:9], ncol = 3)
```


```
grid.arrange(grobs = lp_occupation_f[10:13], ncol = 3)
```

The category “Prof-specialty” provides the largest percentage of women earning over 50K annually, followed by the category “Exec-managerial”, 25.5% and 24.2% respectively. This percentage is less than 10% in the rest of the categories with the exception of “Protective-serv” and “Tech-support”. We’ll give a summary statistic for the number of women belonging to each category of `occupation` :

```
summary(adult_data[adult_data$sex == " Female",]$occupation)
```

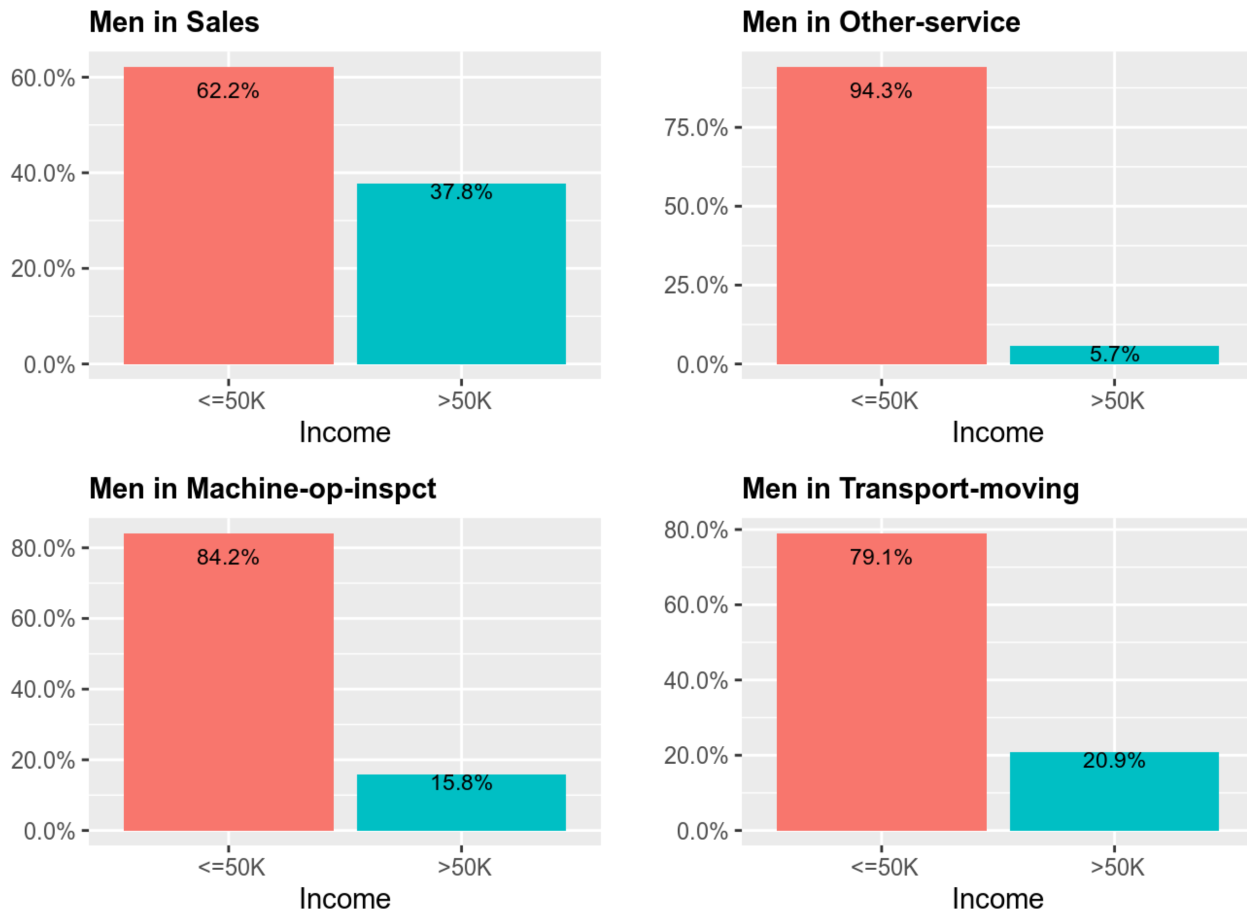
```
##     Prof-specialty       Craft-repair    Exec-managerial
##               1491                216               1143
##       Adm-clerical              Sales      Other-service
##               2512               1248               1758
##  Machine-op-inspct   Transport-moving  Handlers-cleaners
##                543                 90                164
##    Farming-fishing       Tech-support    Protective-serv
##                 65                341                 76
##    Priv-house-serv       Armed-Forces
##                135                  0
```

The categories “Adm-clerical”, “Exec-managerial”, “Machine-op-inspct”, “Other-service”, “Prof-specialty”, “Sales”, and “Tech-support” can be considered representative random samples while any inferences drawn from the remaining categories should be viewed cautiously.

Since no men working in the “Priv-house-serv” sector are earning more than 50K annually, we leave out that category when we display the barplots of income for men grouped by occupation:

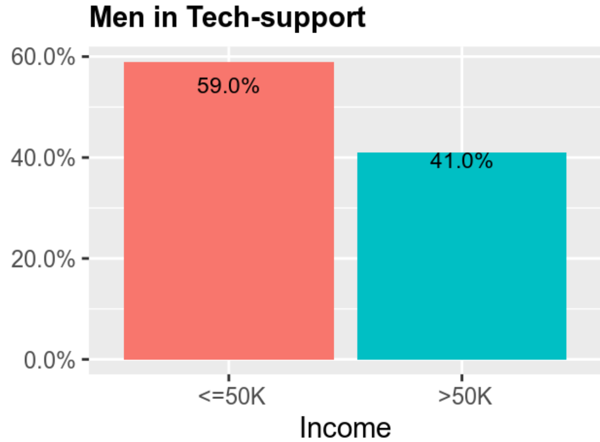
```
modified_occup_m <- levels(adult_data$occupation)
modified_occup_m <- modified_occup_m[!is.element(modified_occup_m, " Priv-house-serv")]

lp_occupation_m <- lapply(modified_occup_m, function(v){
  ggplot(data = subset(adult_data, adult_data$occupation == v & adult_data$sex == " Male"
),
         aes(x = subset(adult_data, adult_data$occupation == v & adult_data$sex == " Male"
)$income,
             fill = subset(adult_data, adult_data$occupation == v & adult_data$sex == " Ma
le")$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = c(2, 1), size = 3) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(paste("Men in", v, sep = "")) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lp_occupation_m[1:4], ncol = 2)
```

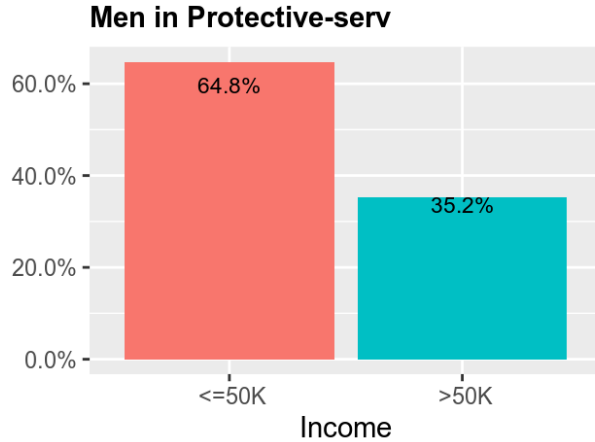

```
grid.arrange(grobs = lp_occupation_m[5:8], ncol = 2)
```

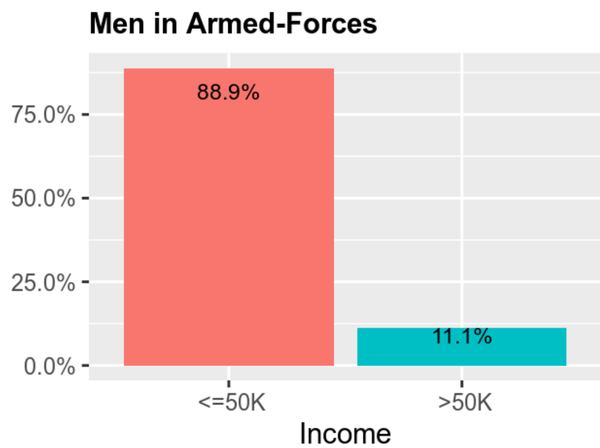



```
grid.arrange(grobs = lp_occupation_m[9:10], ncol = 2)
```


```
grid.arrange(grobs = lp_occupation_m[11:13], ncol = 2)
```







From the summary statistic we see the number of men in each category of the variable `occupation` :

```
summary(adult_data[adult_data$sex == " Male",]$occupation)
```

```
##      Prof-specialty       Craft-repair    Exec-managerial
##                2547               3814               2849
##        Adm-clerical              Sales      Other-service
##                1209               2336               1454
##   Machine-op-inspct   Transport-moving  Handlers-cleaners
##                1423               1482               1186
##     Farming-fishing       Tech-support    Protective-serv
##                 924                571                568
##     Priv-house-serv       Armed-Forces
##                   8                  9
```

Overall, we see the tendency that work requiring highly qualified specialists with college degrees compensates higher in terms of income, a reasonable observation as it reflects the actual real job market.

The Variable `relationship`

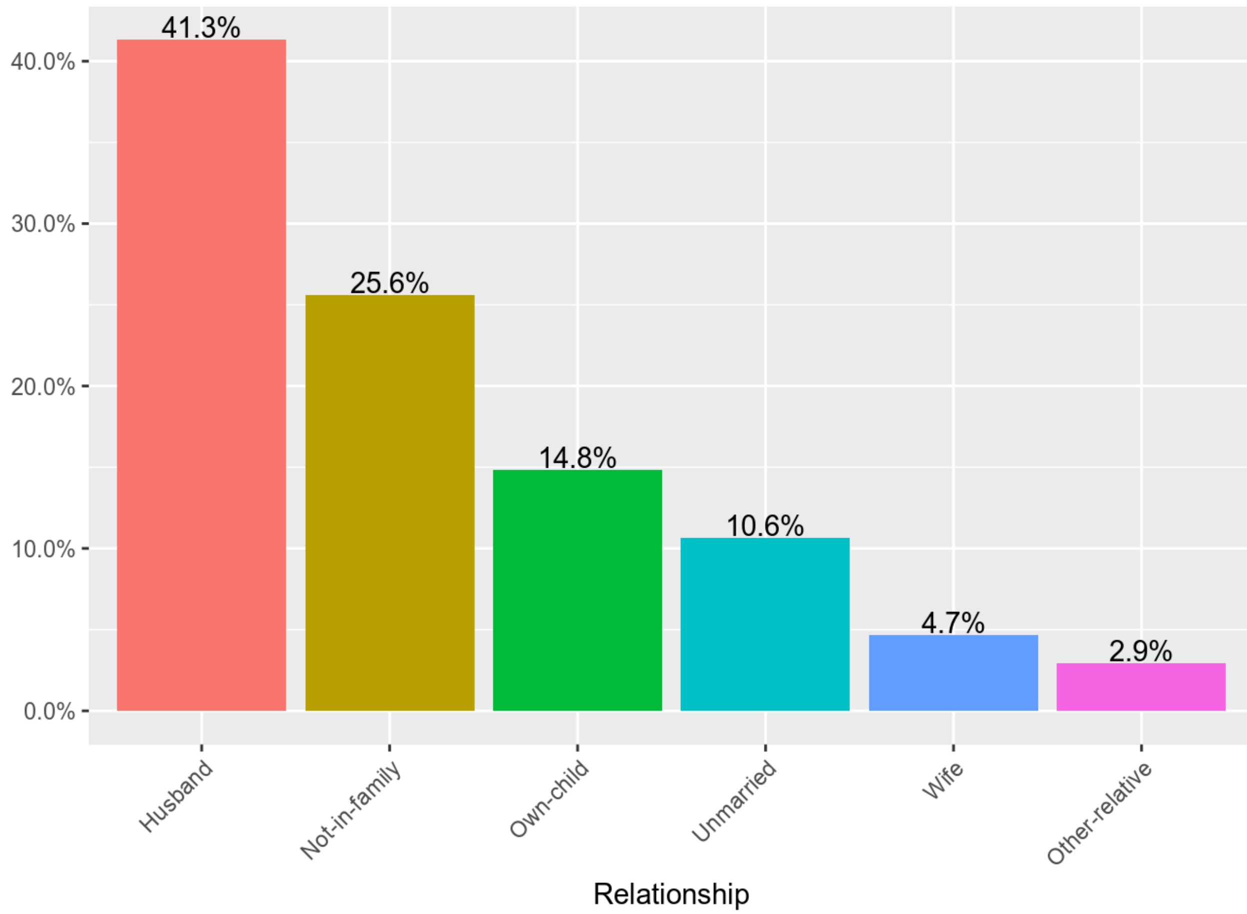
This variable is closely related to the variable `marital_status` and together they should be considered. We notice from the summary below, that the majority of people are married because they identified themselves as “Husband” or “Wife”, and this is in agreement with the summary statistic of `marital_status` .

```
summary(adult_data$relationship)
```

```
##         Husband   Not-in-family  Other-relative       Own-child
##           12463            7726             889            4466
##       Unmarried            Wife
##            3212            1406
```

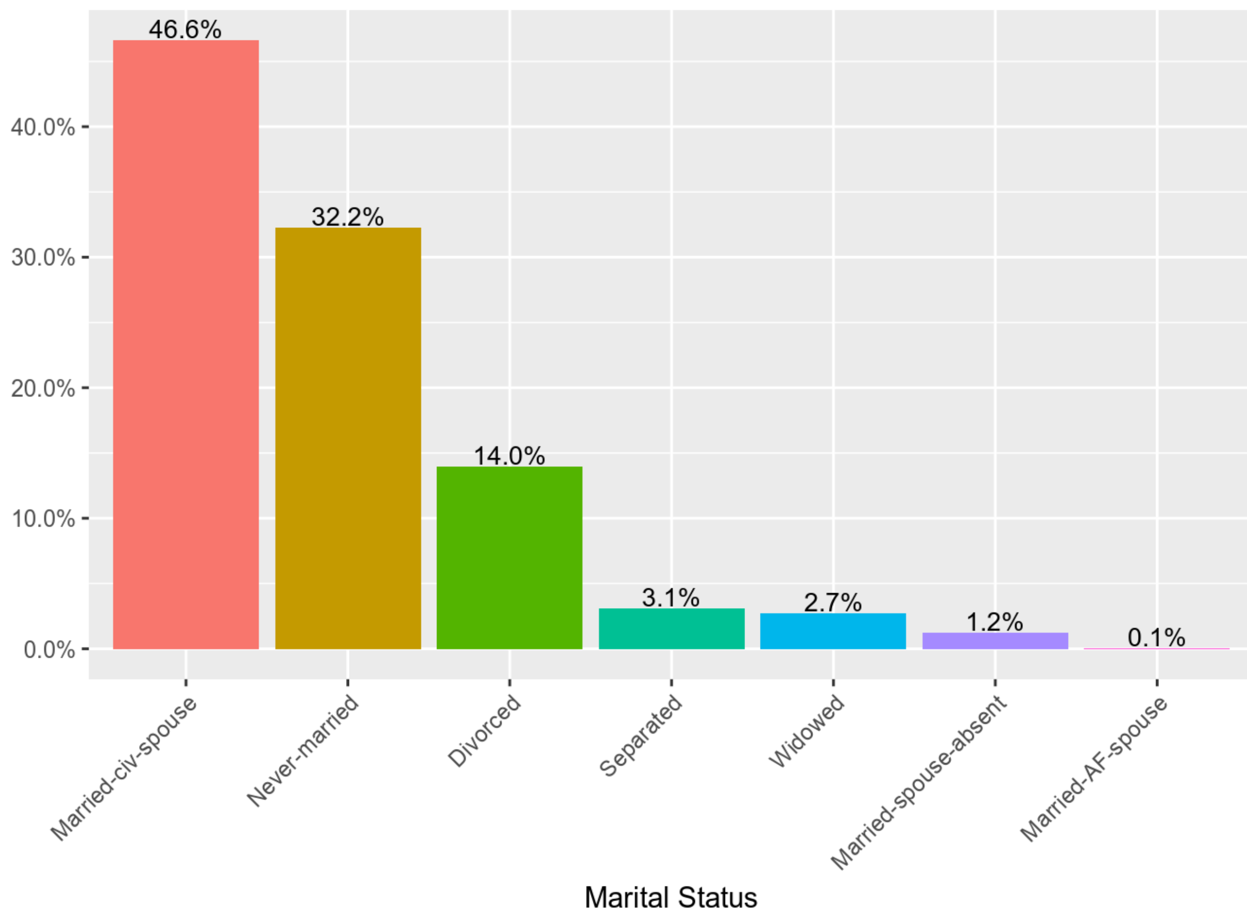
We show the percentage of people belonging to each category of `relationship` :

```
adult_data$relationship <- factor(adult_data$relationship, levels = names(sort(table(adult
_data$relationship),
                                                                               decreasing
= TRUE)))

ggplot(adult_data, aes(x = adult_data$relationship, fill = adult_data$relationship)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1) +
  labs(x = "Relationship", y = "", fill = "Relationship") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



The distribution of people in each category of `relationship` is connected to that of `marital_status` :

```
ggplot(adult_data, aes(x = adult_data$marital_status, fill = adult_data$marital_status)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1, size = 3.5) +
  labs(x = "Marital Status", y = "", fill = "Marital Status") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



Let’s give a summary statistic of the marital status of each level of the factor variable `relationship` :

```
summary(adult_data[adult_data$relationship == " Not-in-family",]$marital_status)
```

```
##     Married-civ-spouse          Never-married               Divorced
##                     14                   4448                   2268
##              Separated                Widowed  Married-spouse-absent
##                    383                    432                    181
##      Married-AF-spouse
##                      0
```

```
summary(adult_data[adult_data$relationship == " Husband",]$marital_status)
```

```
##     Married-civ-spouse          Never-married              Divorced
##                  12454                      0                     0
##              Separated                Widowed  Married-spouse-absent
##                      0                      0                     0
##      Married-AF-spouse
##                      9
```

```
summary(adult_data[adult_data$relationship == " Other-relative",]$marital_status)
```

```
##     Married-civ-spouse          Never-married              Divorced
##                    118                    548                   103
##              Separated                Widowed  Married-spouse-absent
##                     53                     40                    26
##      Married-AF-spouse
##                      1
```

```
summary(adult_data[adult_data$relationship == " Own-child",]$marital_status)
```

```
##     Married-civ-spouse          Never-married              Divorced
##                     83                   3929                   308
##              Separated                Widowed  Married-spouse-absent
##                     90                     12                    43
##      Married-AF-spouse
##                      1
```

```
summary(adult_data[adult_data$relationship == " Unmarried",]$marital_status)
```

```
##     Married-civ-spouse          Never-married              Divorced
##                      0                    801                  1535
##              Separated                Widowed  Married-spouse-absent
##                    413                    343                   120
##      Married-AF-spouse
##                      0
```

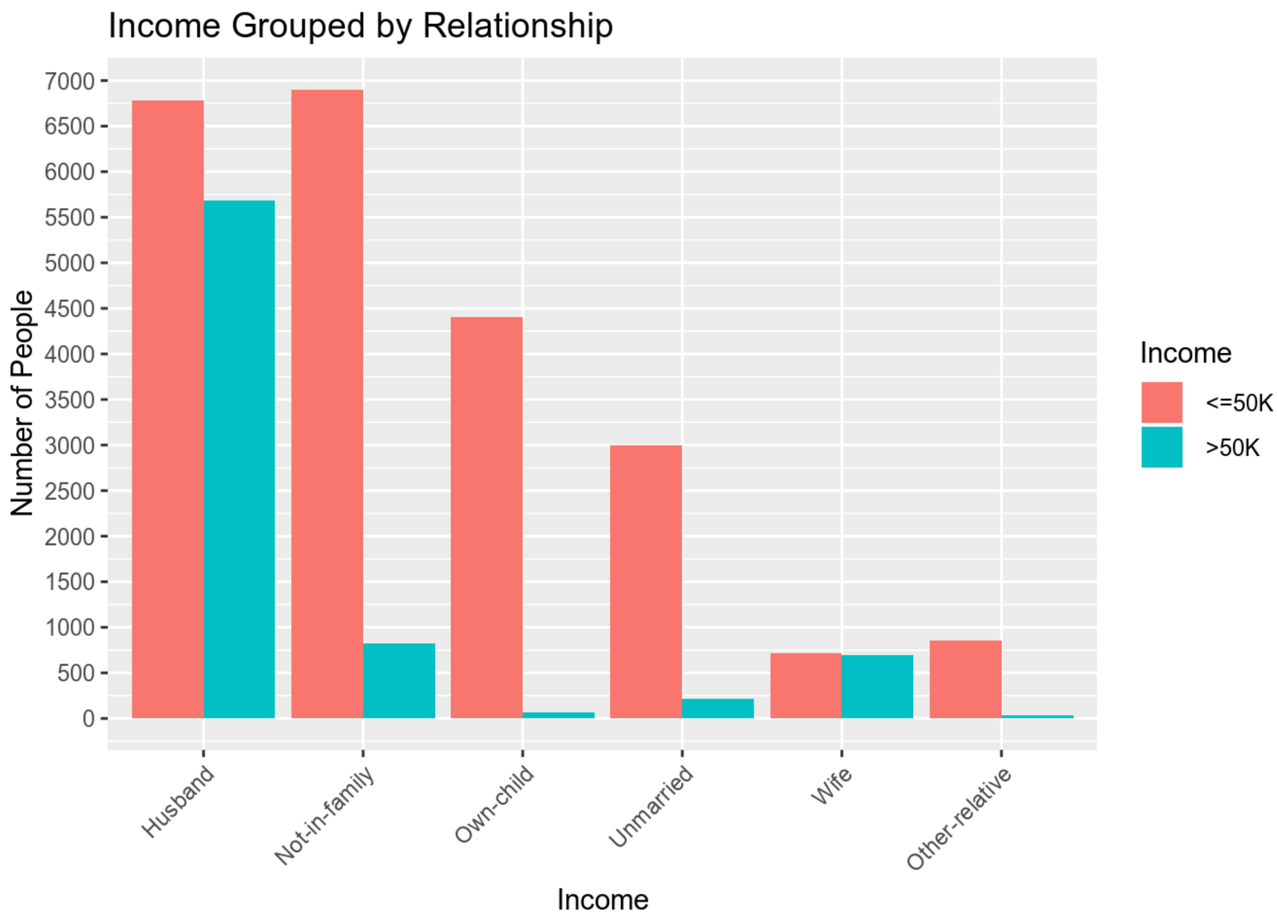
```
summary(adult_data[adult_data$relationship == " Wife",]$marital_status)
```

```
##     Married-civ-spouse          Never-married              Divorced
##                   1396                      0                     0
##              Separated                Widowed  Married-spouse-absent
##                      0                      0                     0
##      Married-AF-spouse
##                     10
```

Most of these results are in accordance with the variable `marital_status` .

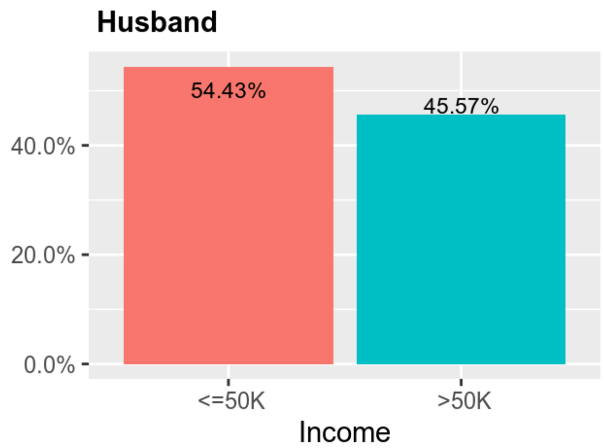
We show barplots of income by relationship status:

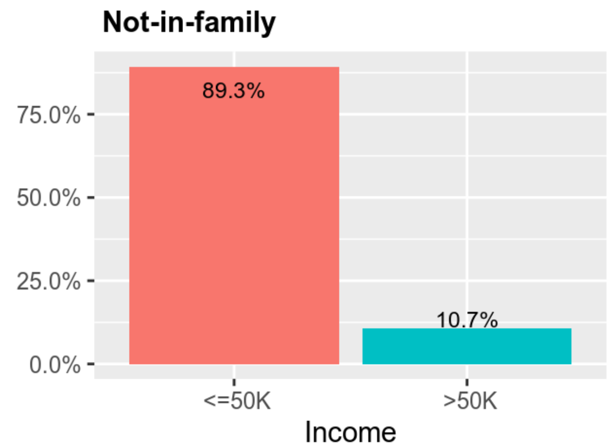
```
ggplot(adult_data, aes(x = adult_data$relationship, fill = adult_data$income)) +
  geom_bar(position = position_dodge()) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Income", y = "Number of People", fill = "Income") +
  ggtitle("Income Grouped by Relationship") +
  scale_y_continuous(breaks = seq(0, 7000, 500))
```

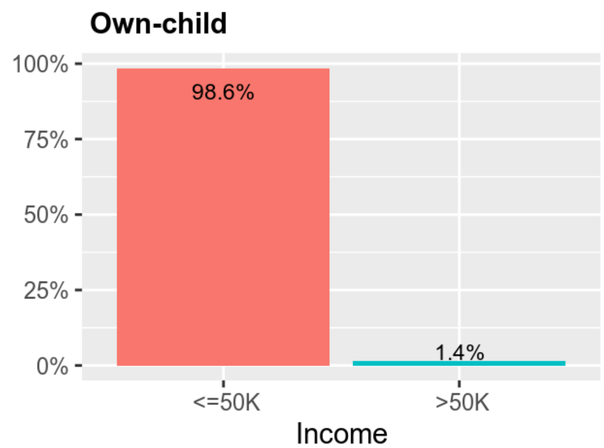


And we give barplots with the percentages of people having an income lower and higher than 50K annually for each relationship level:

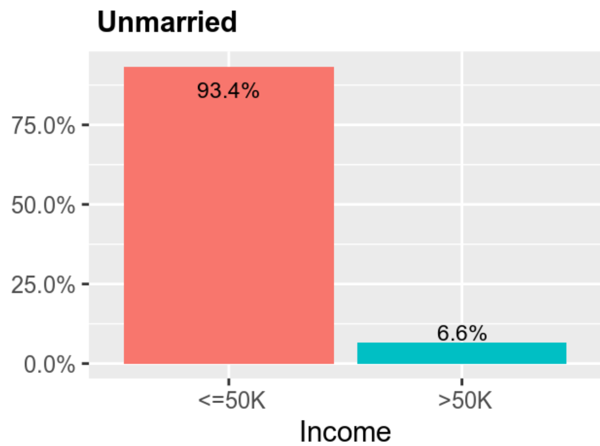
```
lg_relationship <- lapply(levels(adult_data$relationship), function(v){
  ggplot(data = subset(adult_data, adult_data$relationship == v), aes(x = subset(adult_dat
a, adult_data$relationship == v)$income, fill = subset(adult_data, adult_data$relationship
== v)$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = c(2, -0.1), size = 3) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(paste(v)) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold")) +
    scale_y_continuous(labels = percent)})

grid.arrange(grobs = lg_relationship[1:3], ncol = 2)
```

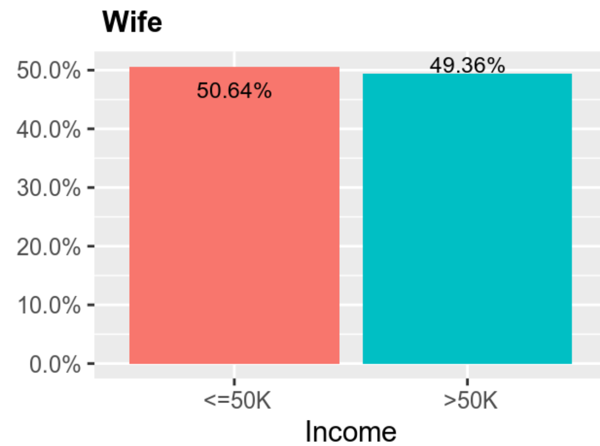


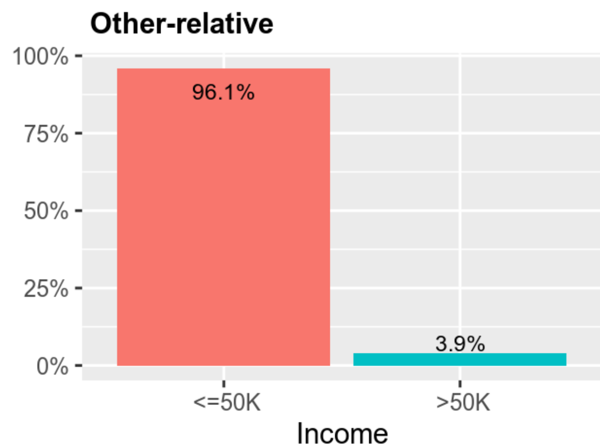





```
grid.arrange(grobs = lg_relationship[4:6], ncol = 2)
```







As was the case for `marital_status` , we can observe a correlation between `income` and `relationship` .

The Variable `race`

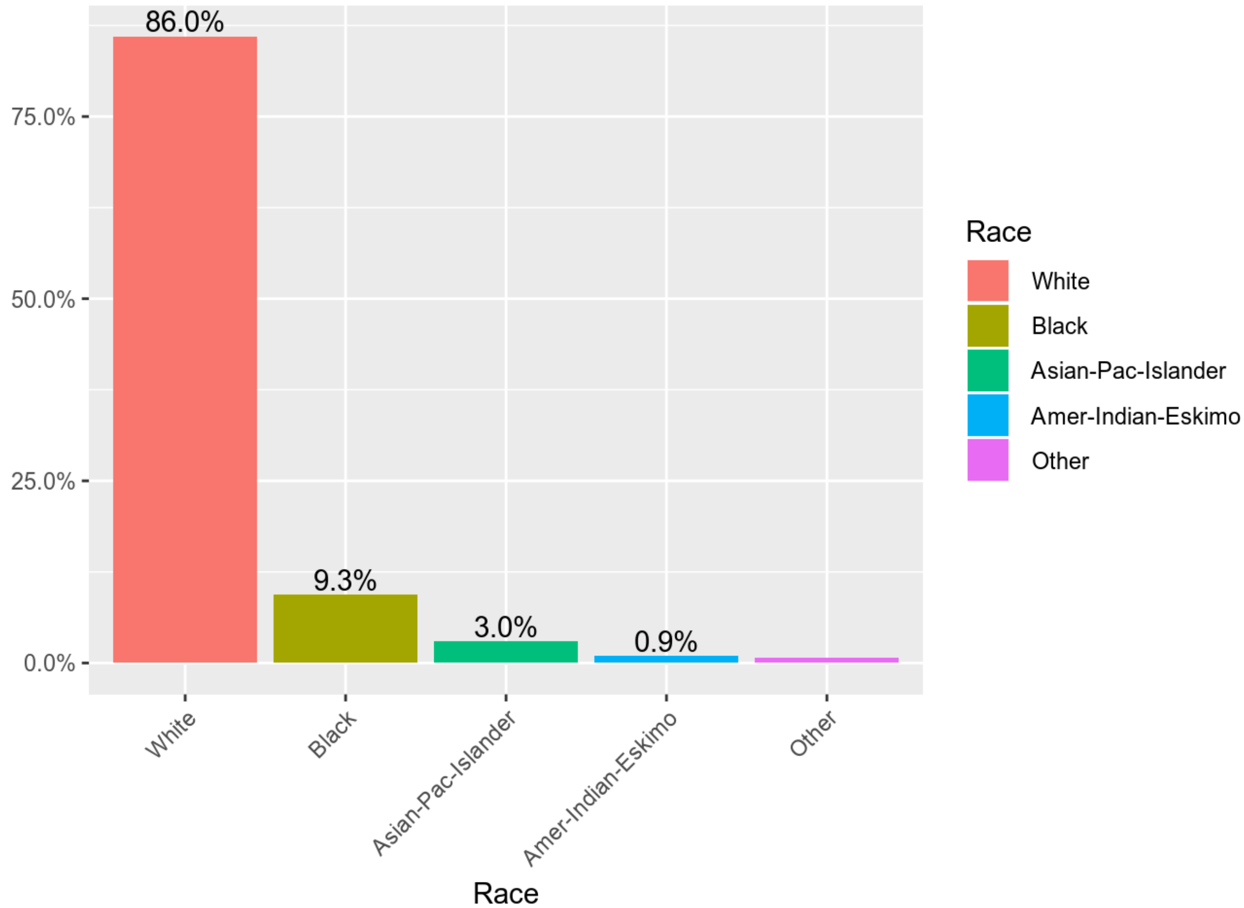
We start with a summary statistic:

```
summary(adult_data$race)
```

```
##  Amer-Indian-Eskimo  Asian-Pac-Islander               Black
##                 286                 895                2817
##               Other               White
##                 231               25933
```

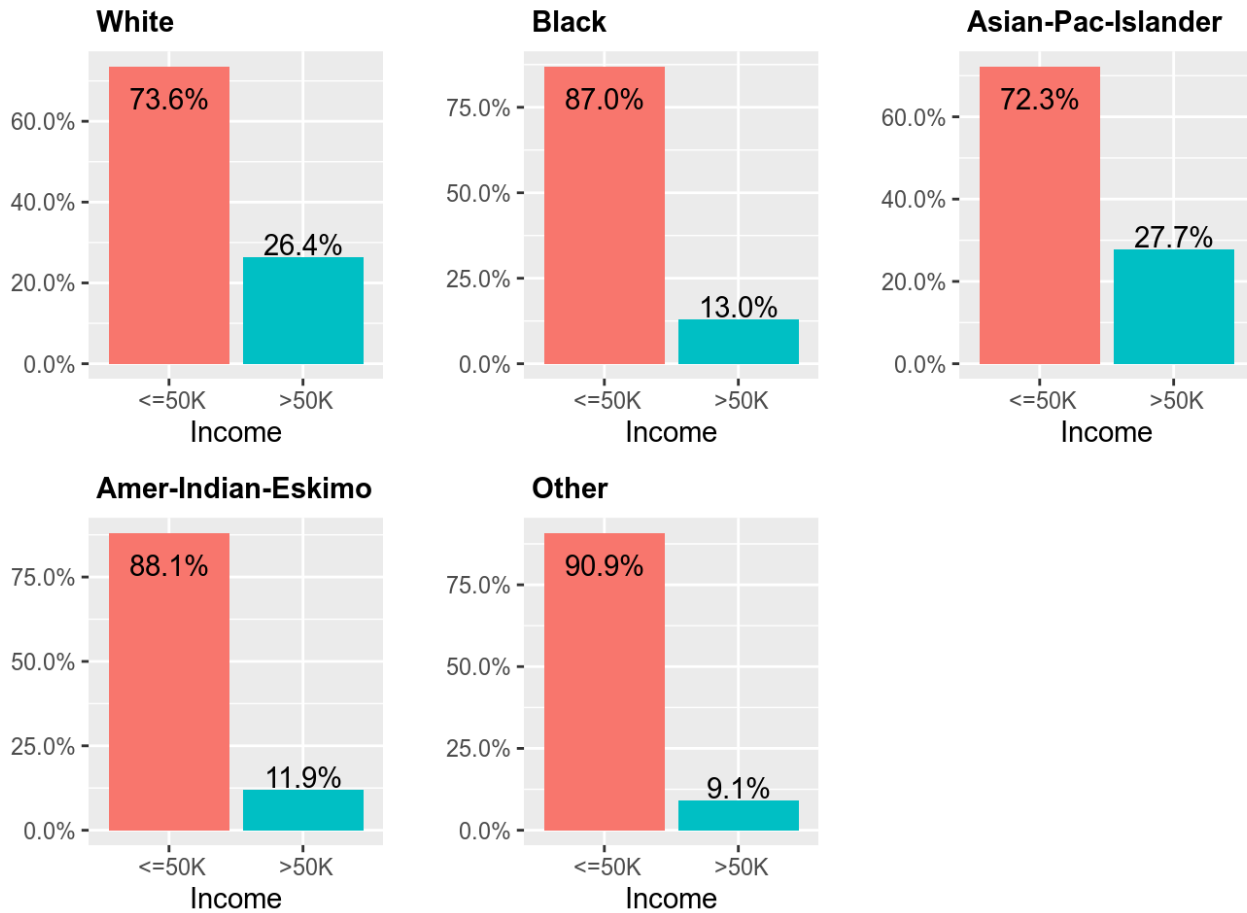
Most of the individuals belong to the category “White”, followed by the category “Black”.

```
adult_data$race <- factor(adult_data$race, levels = names(sort(table(adult_data$race), dec
reasing = TRUE)))

ggplot(adult_data, aes(x = adult_data$race, fill = adult_data$race)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = c(-0.2, -0.2, -0.2, -0.2, 3)) +
  labs(x = "Race", y = "", fill = "Race") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(labels = percent)
```



We show the bar plots of income by race:

```
lg_race <- lapply(levels(adult_data$race), function(v){
  ggplot(data = subset(adult_data, adult_data$race == v), aes(x = subset(adult_data, adult
_data$race == v)$income,
                                                              fill = subset(adult_data, ad
ult_data$race == v)$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = c(2, -0.1)) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(paste(v)) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold")) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_race, ncol = 3)
```



The Variable `sex`

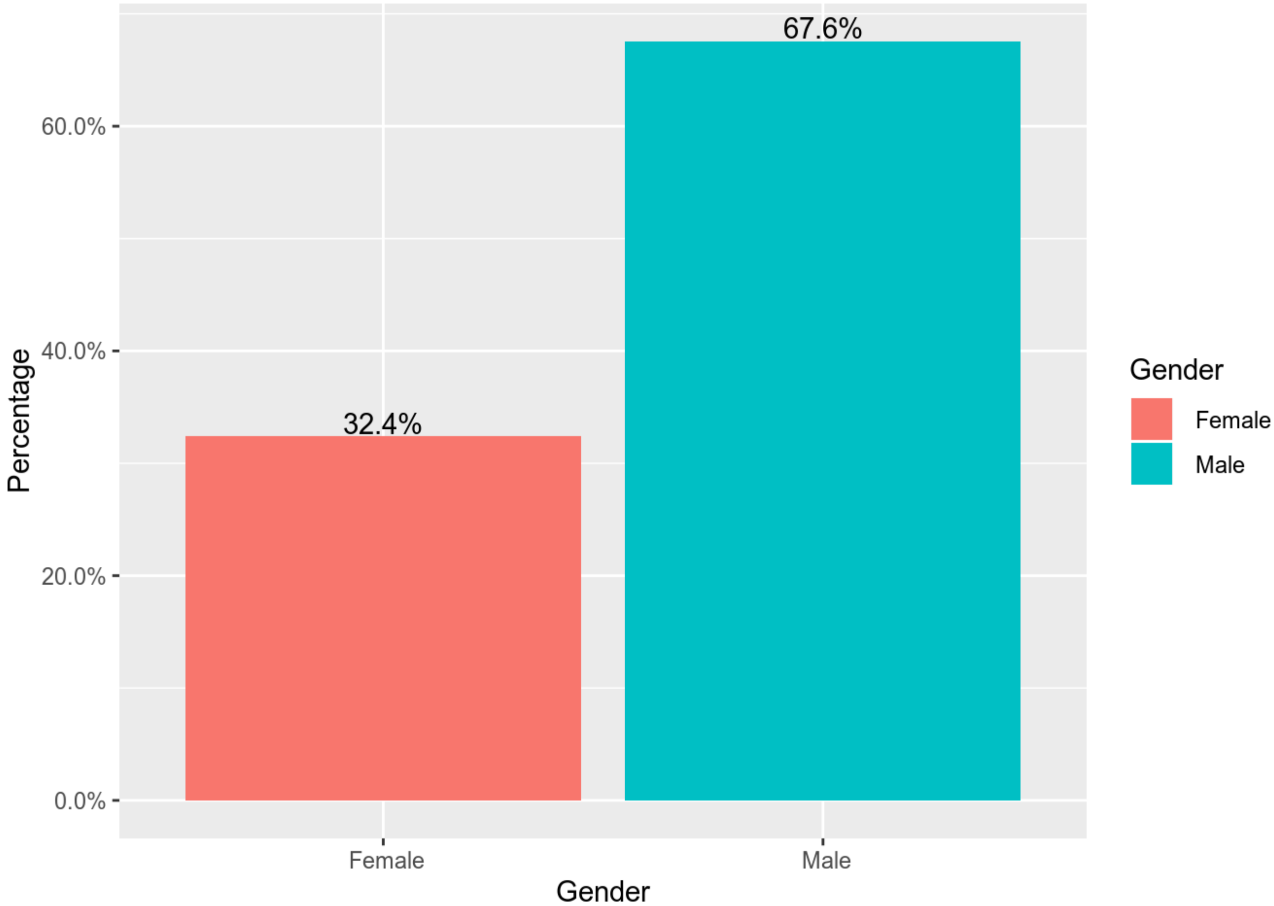
We can see that there were 20,380 men and 9,782 women who took part in the survey:

```
summary(adult_data$sex)
```

```
##  Female    Male
##    9782   20380
```

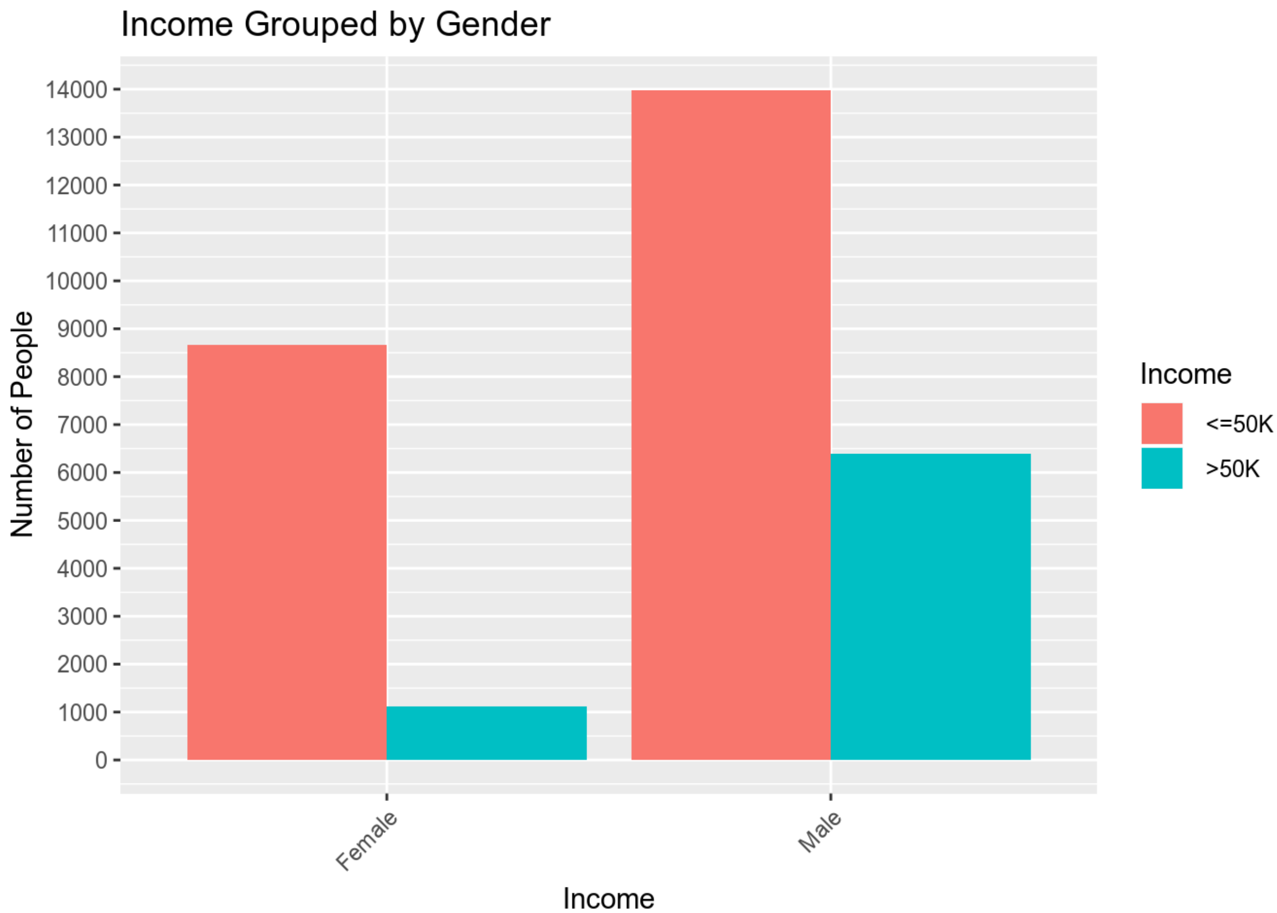
Percentage-wise, this is 67.6% male and 32.4% female:

```
ggplot(adult_data, aes(x = adult_data$sex, fill = adult_data$sex)) +
  geom_bar(aes(y = (..count..) / sum(..count..))) +
  geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) / s
um(..count..)),
            stat = "count", vjust = -.1) +
  labs(x = "Gender", y = "Percentage", fill = "Gender") +
  scale_y_continuous(labels = percent)
```



Here is the number of men and women who earn less than and more than 50K annually:

```
ggplot(adult_data, aes(x = adult_data$sex, fill = adult_data$income)) +
  geom_bar(position = position_dodge()) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Income", y = "Number of People", fill = "Income") +
  ggtitle("Income Grouped by Gender") +
  scale_y_continuous(breaks = seq(0, 14500, 1000))
```

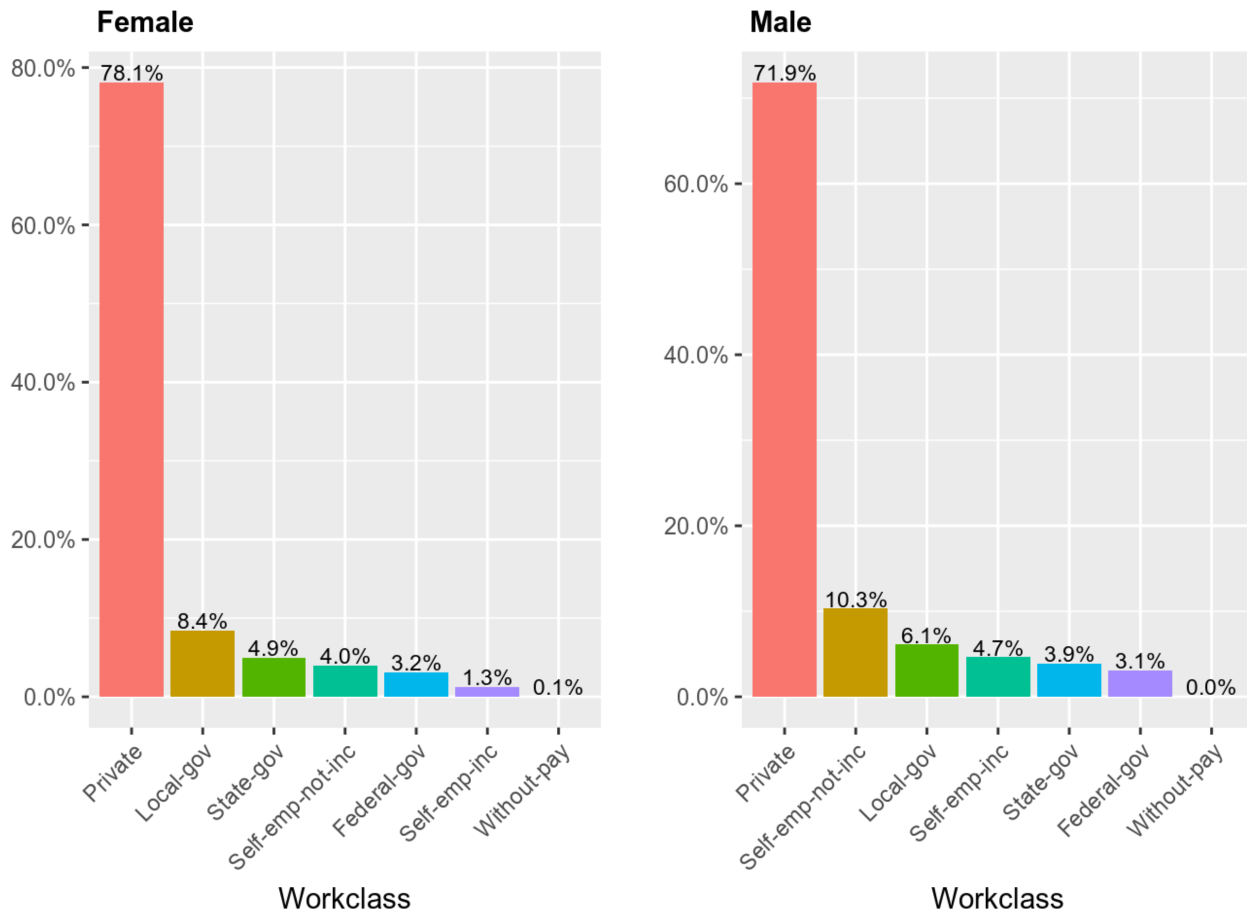


Here is the same information expressed proportionally:

```
gender_income <- lapply(levels(adult_data$sex), function(v){
  ggplot(data = subset(adult_data, adult_data$sex == v),
         aes(x = subset(adult_data, adult_data$sex == v)$income,
         fill = subset(adult_data, adult_data$sex == v)$income)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = -0.1, size = 3) +
    labs(x = "Income", y = "", fill = "Income") +
    ggtitle(paste(v)) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold"),
          axis.text.x = element_text(hjust = 1)) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = gender_income, ncol = 2)
```

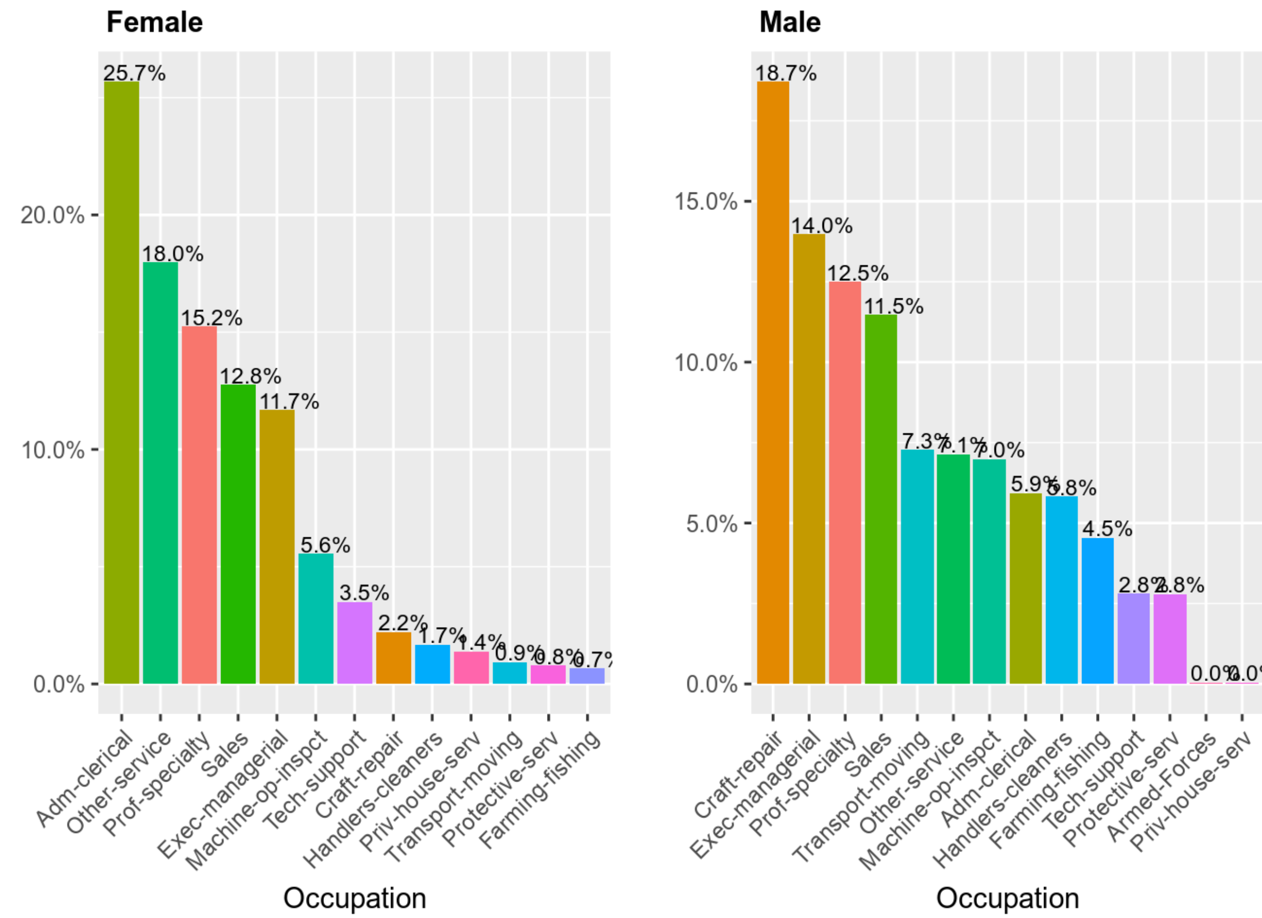
Barplots of workclass grouped by gender:

```
lg_gender_workclass <- lapply(levels(adult_data$sex), function(v){
  df <- subset(adult_data, adult_data$sex == v)
  df <- within(df, workclass <- factor(workclass, levels = names(sort(table(workclass), de
creasing = TRUE))))
  ggplot(data = df, aes(x = df$workclass, fill = df$workclass)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = -0.1, size = 3) +
    labs(x = "Workclass", y = "", fill = "Workclass") +
    ggtitle(paste(v)) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold"),
          axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_gender_workclass, ncol = 2)
```



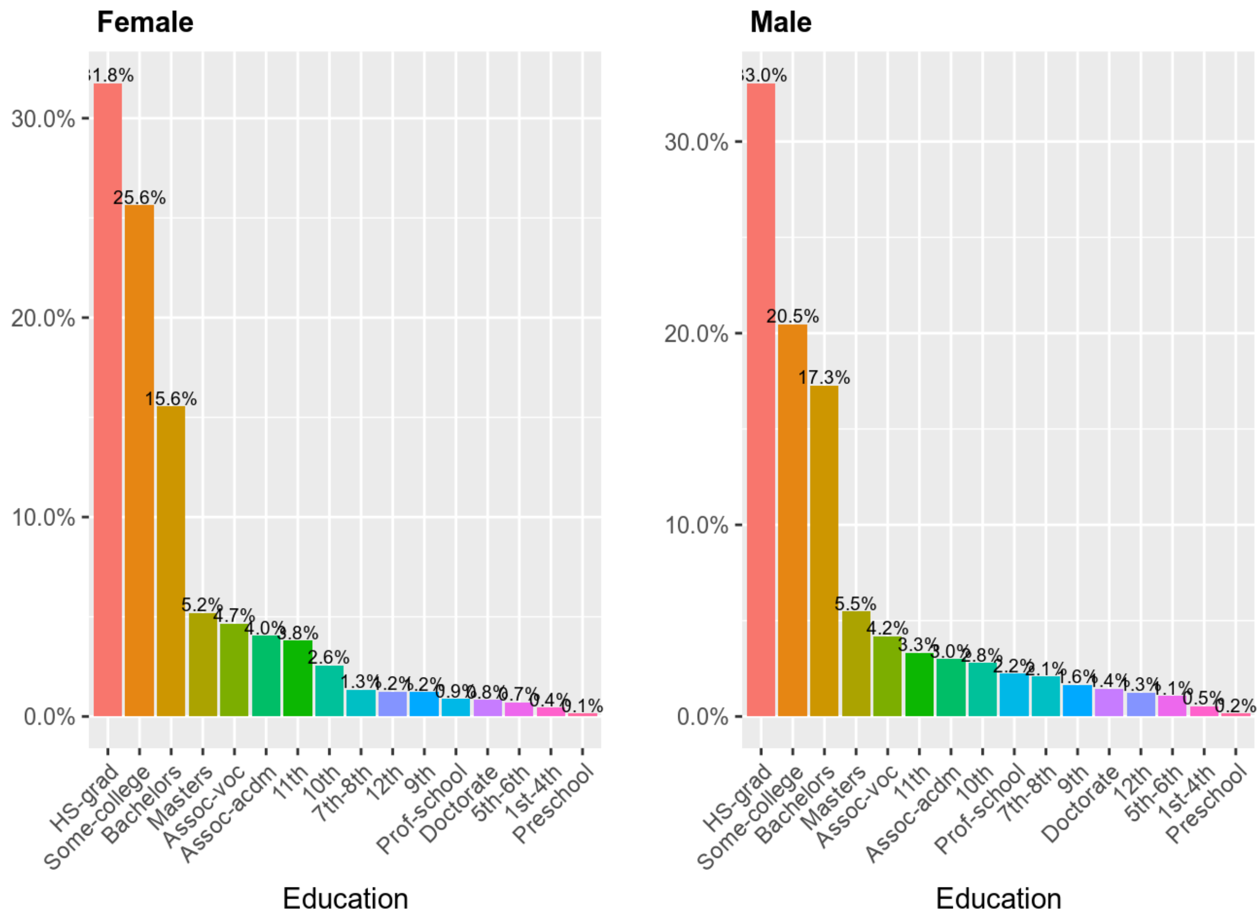
The private sector employs the largest proportion of both male and female individuals. Next, we display barplots of occupation by gender:

```
lg_gender_occupation <- lapply(levels(adult_data$sex), function(v){
  df <- subset(adult_data, adult_data$sex == v)
  df <- within(df, occupation <- factor(occupation, levels = names(sort(table(occupation),
decreasing = TRUE))))
  ggplot(data = df, aes(x = df$occupation, fill = subset(adult_data, adult_data$sex == v)
$occupation)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = -0.1, hjust = 0.3, size = 3) +
    labs(x = "Occupation", y = "", fill = "Occupation") +
    ggtitle(paste(v)) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold"),
          axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_gender_occupation, ncol = 2)
```



The categories “Prof-specialty”, “Exec-managerial”, and “Sales” indicate that there is some overlap for men and women in regards to the most popular occupation levels. We will show barplots of education grouped by gender:

```
lg_gender_education <- lapply(levels(adult_data$sex), function(v){
  df <- subset(adult_data, adult_data$sex == v)
  df <- within(df, education <- factor(education, levels = names(sort(table(education), de
creasing = TRUE))))
  ggplot(data = df, aes(x = df$education, fill = subset(adult_data, adult_data$sex == v)$e
ducation)) +
    geom_bar(aes(y = (..count..) / sum(..count..))) +
    geom_text(aes(label = scales::percent((..count..) / sum(..count..)), y = (..count..) /
sum(..count..)),
              stat = "count", vjust = -0.1, size = 2.5) +
    labs(x = "Education", y = "", fill = "Education") +
    ggtitle(paste(v)) +
    theme(legend.position = "none", plot.title = element_text(size = 11, face = "bold"),
          axis.text.x = element_text(angle = 45, hjust = 1)) +
    scale_y_continuous(labels = percent)
})

grid.arrange(grobs = lg_gender_education, ncol = 2)
```



We see that the percentages of men and women belonging to each level of `education` are very similar, with the exception being the category “Doctorate”, where we observe that there are almost two times more men than women.

Tests for Independence of the Variables

We will test the independence of the categorical variables two-by-two with the Pearson’s Chi Square Test of Independence. The test checks the following null hypothesis,

$$H_0 : \qquad \text{The two categorical variables are independent in the considered population}$$

against the alternative hypothesis,

$$H_A : \qquad \text{The two categorical variables are dependent (and thus, related) in the considered population.}$$

Using the Pearson’s chi-square test, we will check whether the categorical variable `income` is related to some of the other categorical variables.

The Variables `sex` and `income`

```
#Pearson's chi-square test of independence for the variables sex and income:
chisq.test(adult_data$sex, adult_data$income)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  adult_data$sex and adult_data$income
## X-squared = 1415.3, df = 1, p-value < 2.2e-16
```

The p-value is less than 0.05, so we fail to accept the null hypothesis that the categorical variables `sex` and `income` are independent.

The Variables `race` and `income`

```
chisq.test(adult_data$race, adult_data$income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  adult_data$race and adult_data$income
## X-squared = 304.24, df = 4, p-value < 2.2e-16
```

We reject the null hypothesis at the 0.05 significance level. There is a strong indication that `race` and `income` are correlated.

The Variables `workclass` and `income`

```
chisq.test(table(adult_data$workclass, adult_data$income))
```

```
## Warning in chisq.test(table(adult_data$workclass, adult_data$income)): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(adult_data$workclass, adult_data$income)
## X-squared = 804.16, df = 6, p-value < 2.2e-16
```

```
chisq.test(table(adult_data$workclass, adult_data$income))$expected
```

```
## Warning in chisq.test(table(adult_data$workclass, adult_data$income)): Chi-
## squared approximation may be incorrect
```

```
##
##                           <=50K        >50K
##    Private           16738.51349 5547.486506
##    Self-emp-not-inc   1876.94271  622.057291
##    Local-gov          1552.47722  514.522777
##    State-gov           960.62814  318.371859
##    Self-emp-inc        806.65725  267.342749
##    Federal-gov         708.26610  234.733904
##    Without-pay          10.51509    3.484915
```

There are two cells ([“Never-worked”, “<=50K”] and [“Never-worked”, “>50K”]) with expected cell counts equal to 0 and one cell ([“Without-pay”, “>50K”]) with expected cell count equal to 3.5 < 5. If we look at the observed counts of the levels of `workclass` :

```
table(adult_data$workclass)
```

```
##
##          Private  Self-emp-not-inc         Local-gov         State-gov
##            22286              2499              2067              1279
##      Self-emp-inc       Federal-gov       Without-pay
##             1074               943                14
```

There are no participants in the study who identify themselves as belonging to the category “Never-worked”. We will remove this unused factor level from the categorical variable `workclass` :

```
adult_data$workclass <- droplevels(adult_data$workclass)
levels(adult_data$workclass)
```

```
## [1] " Private"          " Self-emp-not-inc" " Local-gov"
## [4] " State-gov"        " Self-emp-inc"     " Federal-gov"
## [7] " Without-pay"
```

```
summary(adult_data$workclass)
```

```
##          Private  Self-emp-not-inc         Local-gov         State-gov
##            22286              2499              2067              1279
##      Self-emp-inc       Federal-gov       Without-pay
##             1074               943                14
```

And we will perform the Pearson’s chi-square test again:

```
CrossTable(adult_data$workclass, adult_data$income, prop.chisq = TRUE, chisq = TRUE)
```

```
## Warning in chisq.test(t, correct = FALSE, ...): Chi-squared approximation
## may be incorrect
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  30162
##
##
##                      | adult_data$income
## adult_data$workclass |     <=50K |      >50K | Row Total |
## ---------------------|-----------|-----------|-----------|
##              Private |     17410 |      4876 |     22286 |
##                      |    26.938 |    81.279 |           |
##                      |     0.781 |     0.219 |     0.739 |
##                      |     0.769 |     0.649 |           |
##                      |     0.577 |     0.162 |           |
## ---------------------|-----------|-----------|-----------|
##     Self-emp-not-inc |      1785 |       714 |      2499 |
##                      |     4.504 |    13.590 |           |
##                      |     0.714 |     0.286 |     0.083 |
##                      |     0.079 |     0.095 |           |
##                      |     0.059 |     0.024 |           |
## ---------------------|-----------|-----------|-----------|
##            Local-gov |      1458 |       609 |      2067 |
##                      |     5.749 |    17.348 |           |
##                      |     0.705 |     0.295 |     0.069 |
##                      |     0.064 |     0.081 |           |
##                      |     0.048 |     0.020 |           |
## ---------------------|-----------|-----------|-----------|
##            State-gov |       935 |       344 |      1279 |
##                      |     0.684 |     2.063 |           |
##                      |     0.731 |     0.269 |     0.042 |
##                      |     0.041 |     0.046 |           |
##                      |     0.031 |     0.011 |           |
## ---------------------|-----------|-----------|-----------|
##         Self-emp-inc |       474 |       600 |      1074 |
##                      |   137.184 |   413.929 |           |
##                      |     0.441 |     0.559 |     0.036 |
##                      |     0.021 |     0.080 |           |
##                      |     0.016 |     0.020 |           |
## ---------------------|-----------|-----------|-----------|
##          Federal-gov |       578 |       365 |       943 |
##                      |    23.959 |    72.291 |           |
##                      |     0.613 |     0.387 |     0.031 |
##                      |     0.026 |     0.049 |           |
##                      |     0.019 |     0.012 |           |
## ---------------------|-----------|-----------|-----------|
##          Without-pay |        14 |         0 |        14 |
```

```
##                      |     1.155 |     3.485 |           |
##                      |     1.000 |     0.000 |     0.000 |
##                      |     0.001 |     0.000 |           |
##                      |     0.000 |     0.000 |           |
## ---------------------|-----------|-----------|-----------|
##         Column Total |     22654 |      7508 |     30162 |
##                      |     0.751 |     0.249 |           |
## ---------------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  804.1575     d.f. =  6     p =  1.946096e-170
##
##
##
```

Here we see a small p-value, which means we will reject the null hypothesis at the 0.05 significance level.

The Variables `occupation` and `income`

```
chisq.test(adult_data$occupation, adult_data$income)
```

```
## Warning in chisq.test(adult_data$occupation, adult_data$income): Chi-
## squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  adult_data$occupation and adult_data$income
## X-squared = 3687.6, df = 13, p-value < 2.2e-16
```

We check if there are any cells with expected count less than 5:

```
chisq.test(adult_data$occupation, adult_data$income)$expected
```

```
## Warning in chisq.test(adult_data$occupation, adult_data$income): Chi-
## squared approximation may be incorrect
```

```
##                        adult_data$income
## adult_data$occupation       <=50K        >50K
##     Prof-specialty    3032.851005 1005.148995
##     Craft-repair      3026.842384 1003.157616
##     Exec-managerial   2998.301439  993.698561
##     Adm-clerical      2794.759432  926.240568
##     Sales             2691.861813  892.138187
##     Other-service     2412.460977  799.539023
##     Machine-op-inspct 1476.618394  489.381606
##     Transport-moving  1180.693853  391.306147
##     Handlers-cleaners 1013.954645  336.045355
##     Farming-fishing    742.815662  246.184338
##     Tech-support       684.982693  227.017307
##     Protective-serv    483.693920  160.306080
##     Priv-house-serv    107.404085   35.595915
##     Armed-Forces         6.759698    2.240302
```

We see that there is one problematic cell, so we'll consider this Pearson's test untrustworthy. However, the p-value is quite small, so we'll go ahead and take that as a strong indication that the variables `occupation` and `income` are dependent.

We will go ahead and summarize the rest of the tests rather briefly…

The Variables `education` and `income`

```
chisq.test(adult_data$education, adult_data$income)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  adult_data$education and adult_data$income
## X-squared = 4070.4, df = 15, p-value < 2.2e-16
```

The Variables `marital_status` and `income`

```
chisq.test(adult_data$marital_status, adult_data$income)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  adult_data$marital_status and adult_data$income
## X-squared = 6061.7, df = 6, p-value < 2.2e-16
```

#3 The Variables `relationship` and `income`

```
chisq.test(adult_data$relationship, adult_data$income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  adult_data$relationship and adult_data$income
## X-squared = 6233.8, df = 5, p-value < 2.2e-16
```

The Variables `native_region` and `income`

```
chisq.test(adult_data$native_region, adult_data$income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  adult_data$native_region and adult_data$income
## X-squared = 233.11, df = 7, p-value < 2.2e-16
```

The Variables `hours_worked` and `income`

```
chisq.test(adult_data$hours_worked, adult_data$income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  adult_data$hours_worked and adult_data$income
## X-squared = 1940.4, df = 4, p-value < 2.2e-16
```

The rest of the Pearson's tests yield very small p-values meaning that it is very unlikely that the categorical variables are not related to `income` .

Secondary Data

For secondary data, we will be using the IncomeESL data set. Originating from an example in the book 'The Elements of Statistical Learning'. The data set is an extract from this survey. It consists of 8993 instances (obtained from the original data set with 9409 instances, by removing those observations with the annual income missing) with 14 demographic attributes. The data set is a good mixture of categorical and continuous variables with a lot of missing data. This dataset has many of the same variables as our census data such as `income` , `sex` , `marital status` , `age` , `occupation` , and `ethnicity` including some other interesting ones like `number of children` and whether the participant rents or owns. Perhaps these additional variables may provide us with useful insights on additional factors that may predict whether or not a person makes 50K annually.

```
library(arules)
```

Let's take a look at the first few lines of this data set

```
data("IncomeESL")
IncomeESL[1:3, ]
```

```
##   income    sex marital status   age           education
## 1    75+ female        married 45-54 college (1-3 years)
## 2    75+   male        married 45-54    college graduate
## 3    75+ female        married 25-34    college graduate
##                occupation years in bay area dual incomes
## 1               homemaker               >10           no
## 2               homemaker               >10           no
## 3 professional/managerial               >10          yes
##   number in household number of children householder status type of home
## 1                   3                  0                own        house
## 2                   5                  2                own        house
## 3                   3                  1               rent    apartment
##   ethnic classification language in home
## 1                 white             <NA>
## 2                 white          english
## 3                 white          english
```

Remove the incomplete cases:

```
## remove incomplete cases
IncomeESL <- IncomeESL[complete.cases(IncomeESL), ]
```

Do some light preparation on the data:

```
IncomeESL[["income"]] <- factor((as.numeric(IncomeESL[["income"]]) > 6) +1,
  levels = 1 : 2 , labels = c("$0-$40,000", "$40,000+"))

IncomeESL[["age"]] <- factor((as.numeric(IncomeESL[["age"]]) > 3) +1,
  levels = 1 : 2 , labels = c("14-34", "35+"))

IncomeESL[["education"]] <- factor((as.numeric(IncomeESL[["education"]]) > 4) +1,
  levels = 1 : 2 , labels = c("no college graduate", "college graduate"))

IncomeESL[["years in bay area"]] <- factor(
  (as.numeric(IncomeESL[["years in bay area"]]) > 4) +1,
  levels = 1 : 2 , labels = c("1-9", "10+"))

IncomeESL[["number in household"]] <- factor(
  (as.numeric(IncomeESL[["number in household"]]) > 3) +1,
  levels = 1 : 2 , labels = c("1", "2+"))

IncomeESL[["number of children"]] <- factor(
  (as.numeric(IncomeESL[["number of children"]]) > 1) +0,
  levels = 0 : 1 , labels = c("0", "1+"))
```

We first notice that the factors for the variable `income` don’t quite match what we are looking for when it comes to our prediction model, but maybe some of the same patterns can be noticed within this dataset.

```
levels(IncomeESL$income)
```

```
## [1] "$0-$40,000" "$40,000+"
```