Exploring the BRFSS data

February 7, 2018

Objective

This document serves the purpose of a final evaluation of the 5 week Introduction to Probability and Data course by Duke University. The data of interest is Behavioral Risk Factor Surveillance System (**BRFSS**) dataset. It can be downloaded from here (http://d3c33hcgiwev3.cloudfront.net/_384b2d9eda4b29131fb681b243a7767d_brfss2013.RData?Expires=1518134400&Signature=ZRx5eAKCMHWjiE7CSQUjNa7xDlGSWU5vARrbu1gEcqAVIZ4vyCm6X6KULK7-UtvhqEQZBssQJJIgs8Pzi9TFXmpfu9MN5F1fGsGLjDMXkspZUypI1vSMJpiZ3c25TDn3MpvnVj9IpvHCicD0uIxeLi6M8iw4MKizjwZDjyLKcUA_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A):

Part 1: Data

The Behavioral Risk Factor Surveillance System (BRFSS) is a collaborative project between all of the states in the United States (US) and participating US territories and the Centers for Disease Control and Prevention (CDC).

- The BRFSS objective is to collect uniform, state-specific data on preventive health practices and risk behaviors that are linked to chronic diseases, injuries, and preventable infectious diseases that affect the adult population. Factors assessed by the BRFSS in 2013 include tobacco use, HIV/AIDS knowledge and prevention, exercise, immunization, health status, healthy days - health-related quality of life, health care access, inadequate sleep, hypertension awareness, cholesterol awareness, chronic health conditions, alcohol consumption, fruits and vegetables consumption, arthritis burden, and seatbelt use.
- Since 2011, BRFSS conducts both landline telephone- and cellular telephone-based surveys. In conducting the BRFSS landline telephone survey, interviewers collect data from a randomly selected adult in a household. In conducting the cellular telephone version of the BRFSS questionnaire, interviewers collect data from an adult who participates by using a cellular telephone and resides in a private residence or college housing. Health characteristics estimated from the BRFSS pertain to the non-institutionalized adult population, aged 18 years or older, who reside in the US. In 2013, additional question sets were included as optional modules to provide a measure for several childhood health and wellness indicators, including asthma prevalence for people aged 17 years or younger.

More detailed information of the data can be found here (https://www.cdc.gov/brfss/annual_data/2013/pdf/Overview_2013.pdf):

More detailed information of the dataset can be found here (http://d3c33hcgiwev3.cloudfront.net/_e34476fda339107329fc316d1f98e042_brfss_codebook.html?Expires=1518134400&Signature=HmIKBR9uyX1IsUJIukFEaboihXbifTqMSBiu-exY2wk9NQeRWs5Dxs2P0BhJ2d8WqGWBJWzMIFJlZrAkdi59eHxZd26OL9NFGZYKToJ6f~j1W-KxRlkUGEUsn-LyS3LGJaqul13YPxG1limXnkoqhUwhYgSpJDfOITzVINHq26k_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A):

Required packages

```
library(ggplot2)
library(dplyr)
```

Load data

```
load("brfss2013.RData")
```

First we'll select and extract the variables we want to work with and clean them from NA's

```
# make a small data frame with the variables of interest
t1 <- select(brfss2013, sex, income2, seatbelt, educa, scntwrk1, employ1, lsatisfy, X_bmi5) %>%
      filter(!is.na(sex), !is.na(income2), !is.na(seatbelt), !is.na(educa), !is.na(scntwrk1), !is.na(lsatisfy),!
is.na(X_bmi5),!is.na(employ1))
```

Part 2: Research questions

Research quesion 1:
What is the body mass index between different incomes of men and women? Does it change when people have higher income, consequently they can afford a better lifestyle and nutrition? Does employment status reflect the BMI between genders?

Research quesion 2:
What type of education brings you what amount of salary? What kind of level of education have men vs women?And are Men or Women better educated? Does education reflects putting a seatbelt on?

Research quesion 3:
What is the overall satisfaction level in different education levels among genders?

Part 3: Exploratory data analysis

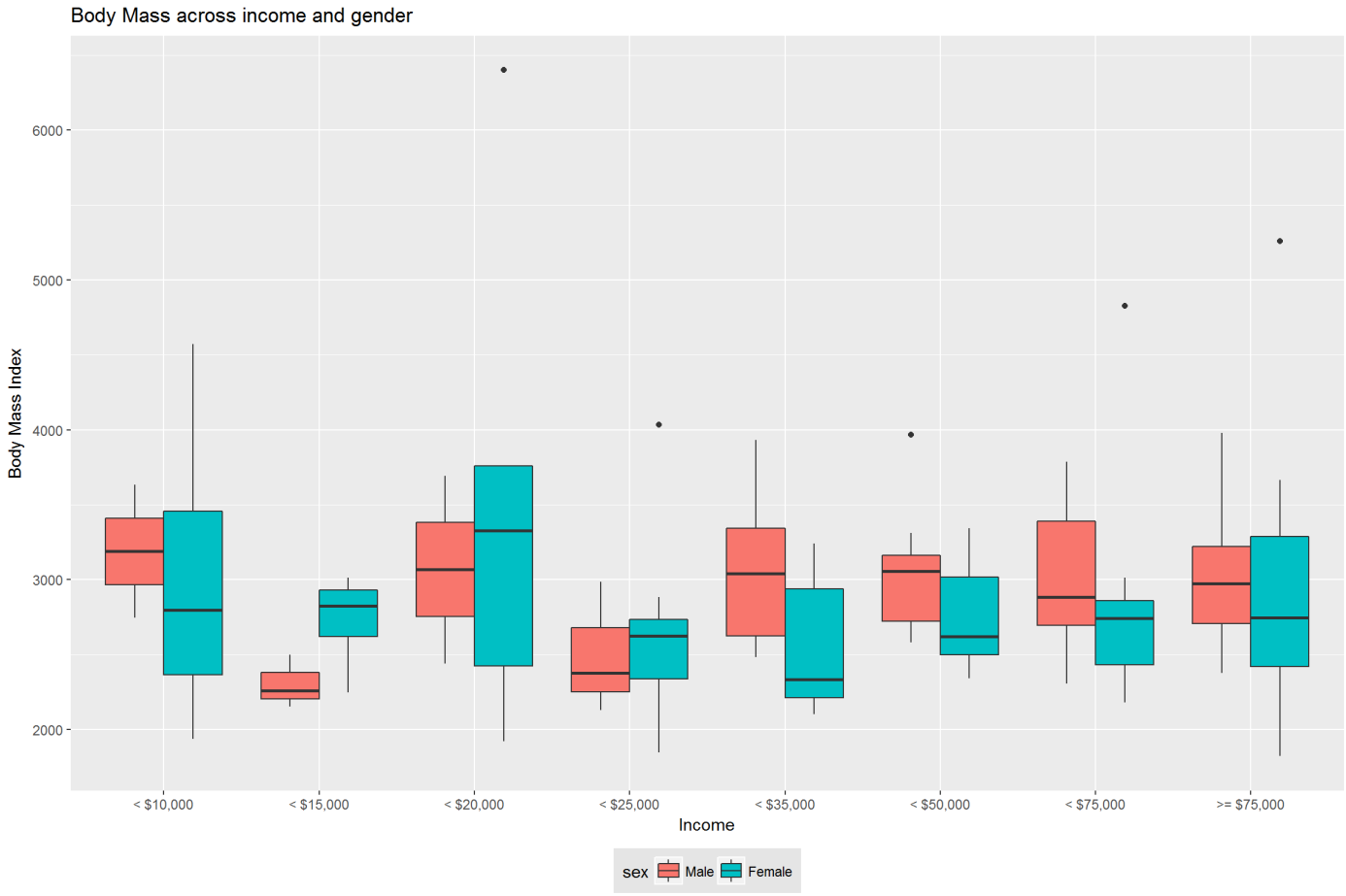
Research quesion 1:

```
# Let's first see the median BMI across gender
t1 %>%
      group_by(sex) %>%
      summarise(medBMI = median(X_bmi5))
```

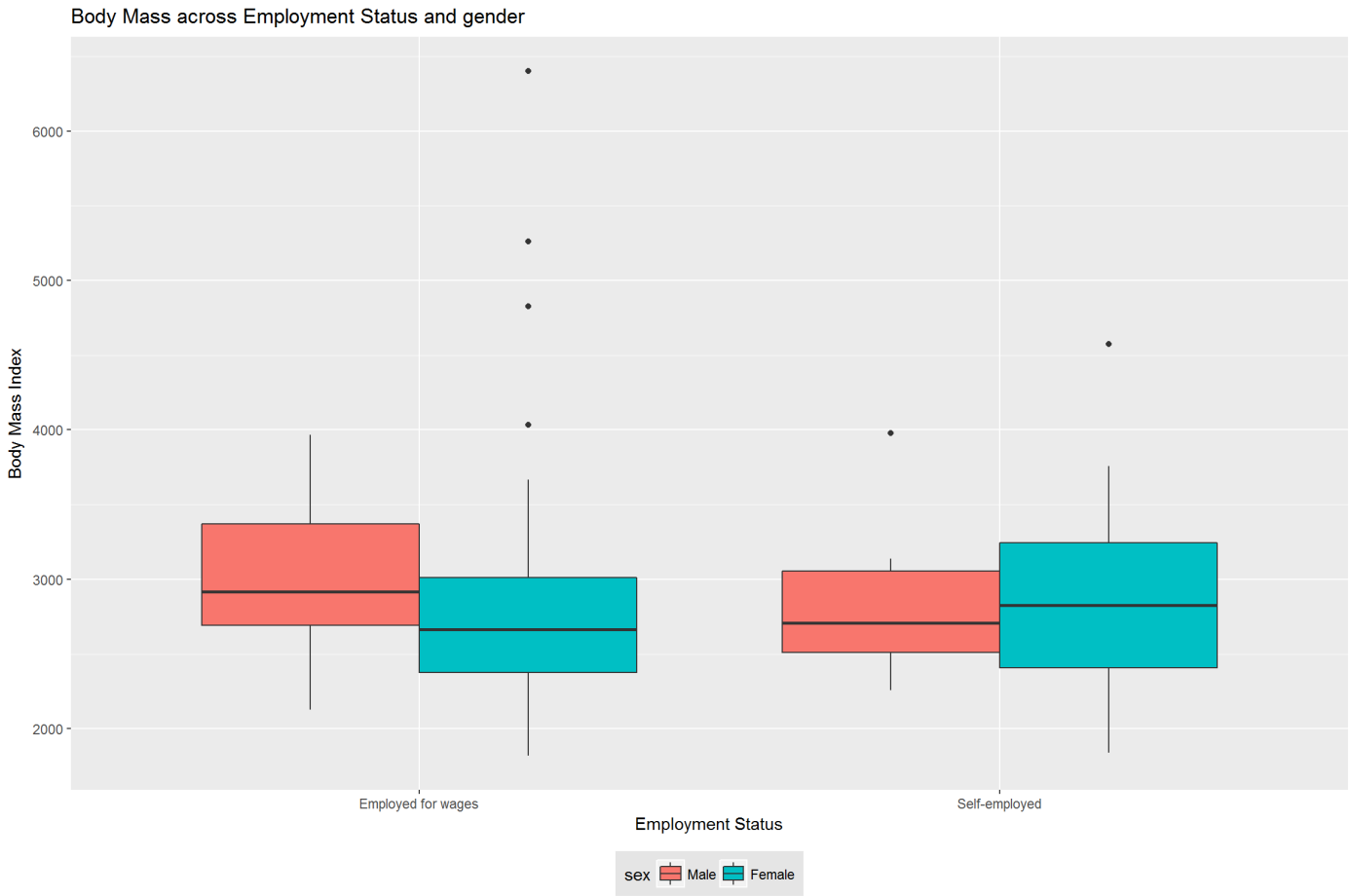
```
## # A tibble: 2 x 2
##      sex medBMI
##   <fctr>  <int>
## 1   Male   2889
## 2 Female   2694
```

We see that men have higher BMI than women when looking at their median.

```
### BODY MASS, INCOME, GENDER
ggplot(data = t1, aes(x = income2, y = X_bmi5, fill = sex)) +
      geom_boxplot() +
      scale_x_discrete(name = "Income",
                       labels = c("Less than $10,000" = "< $10,000",
                                  "Less than $15,000" = "< $15,000",
                                  "Less than $20,000" = "< $20,000",
                                  "Less than $25,000" = "< $25,000",
                                  "Less than $35,000" = "< $35,000",
                                  "Less than $50,000" = "< $50,000",
                                  "Less than $75,000" = "< $75,000",
                                  "$75,000 or more" = ">= $75,000")) +
      scale_y_continuous(name = "Body Mass Index") +
      ggtitle("Body Mass across income and gender") +
      theme(legend.position = "bottom",
      legend.background = element_rect(fill = "gray90", size = .5, linetype = "dotted"))
```



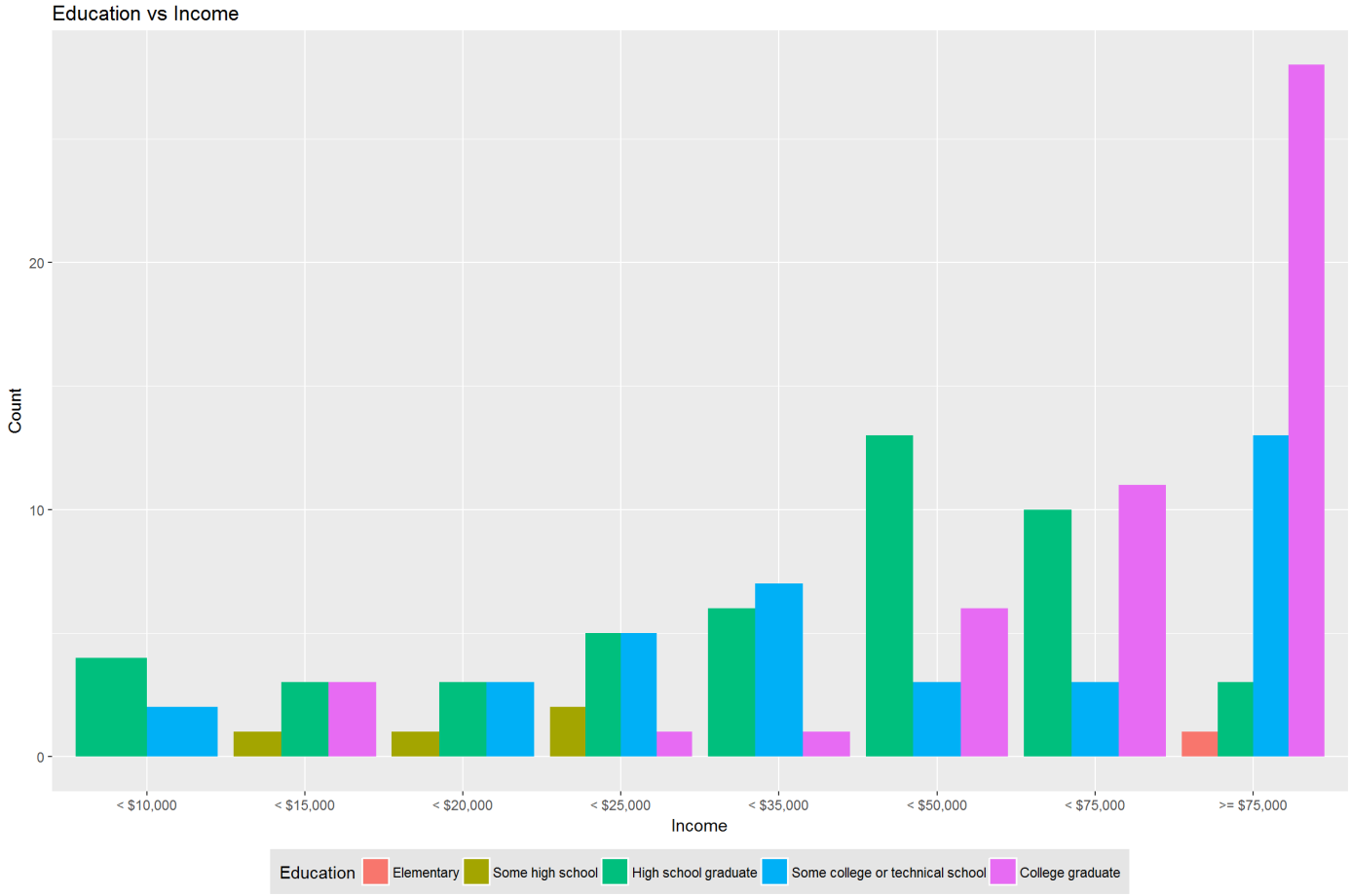
```
### BMI, emplyment status, gender
ggplot(data = t1, aes(x = employ1, y = X_bmi5, fill = sex)) +
      geom_boxplot() +
      labs(title = "Body Mass across Employment Status and gender",
           x = "Employment Status", y = "Body Mass Index") +
      theme(legend.position = "bottom",
      legend.background = element_rect(fill = "gray90", size = .5, linetype = "dotted"))
```



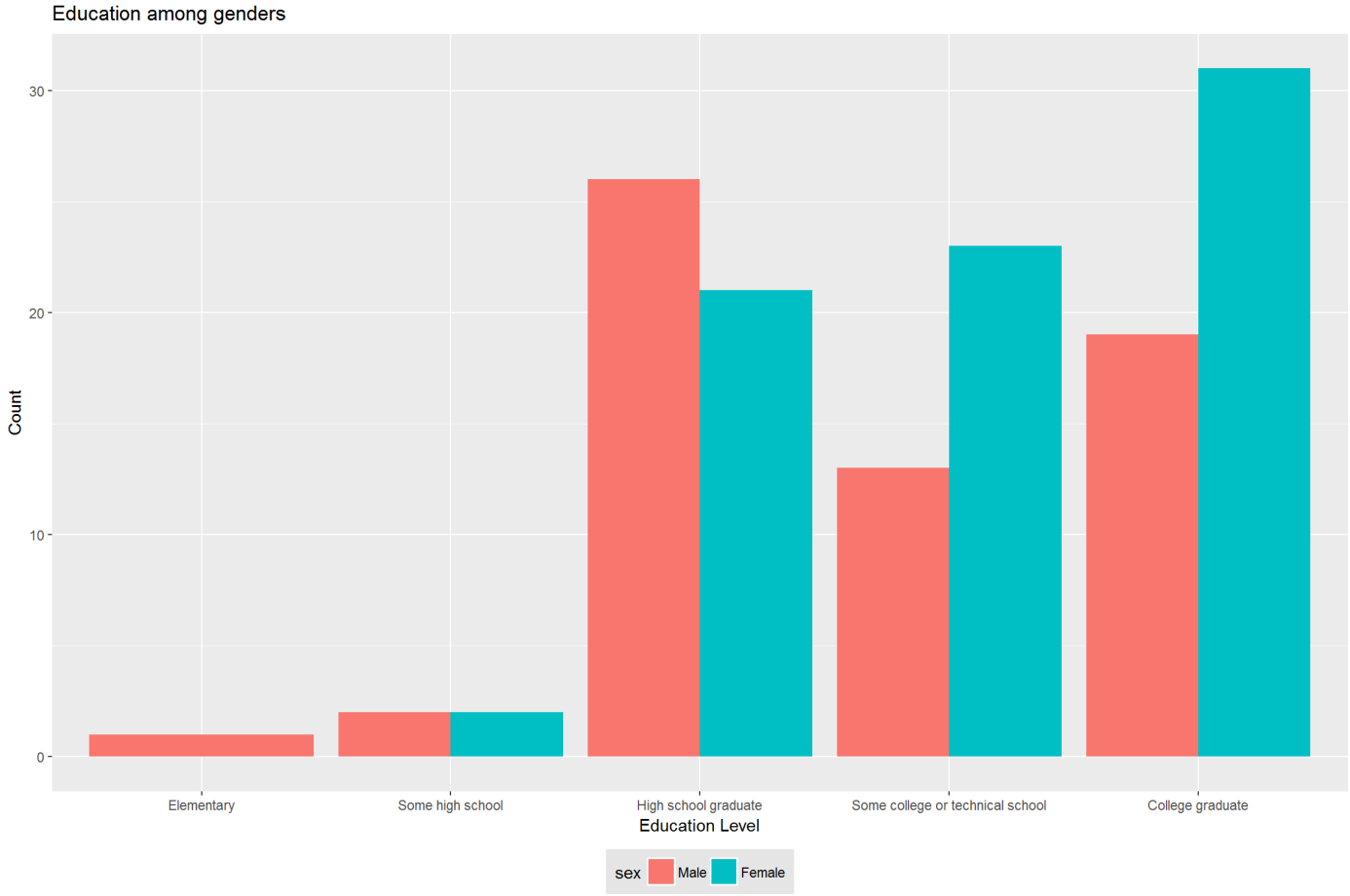
What we can observe here is that women who earn less than $25.000 but more then $10.000 tend to be with lower BMI than men. In any other income level women have BMI higher than men. When women are selfemplyed have lower BMI than the employed for wages women. With men is vice versa.

Research quesion 2:

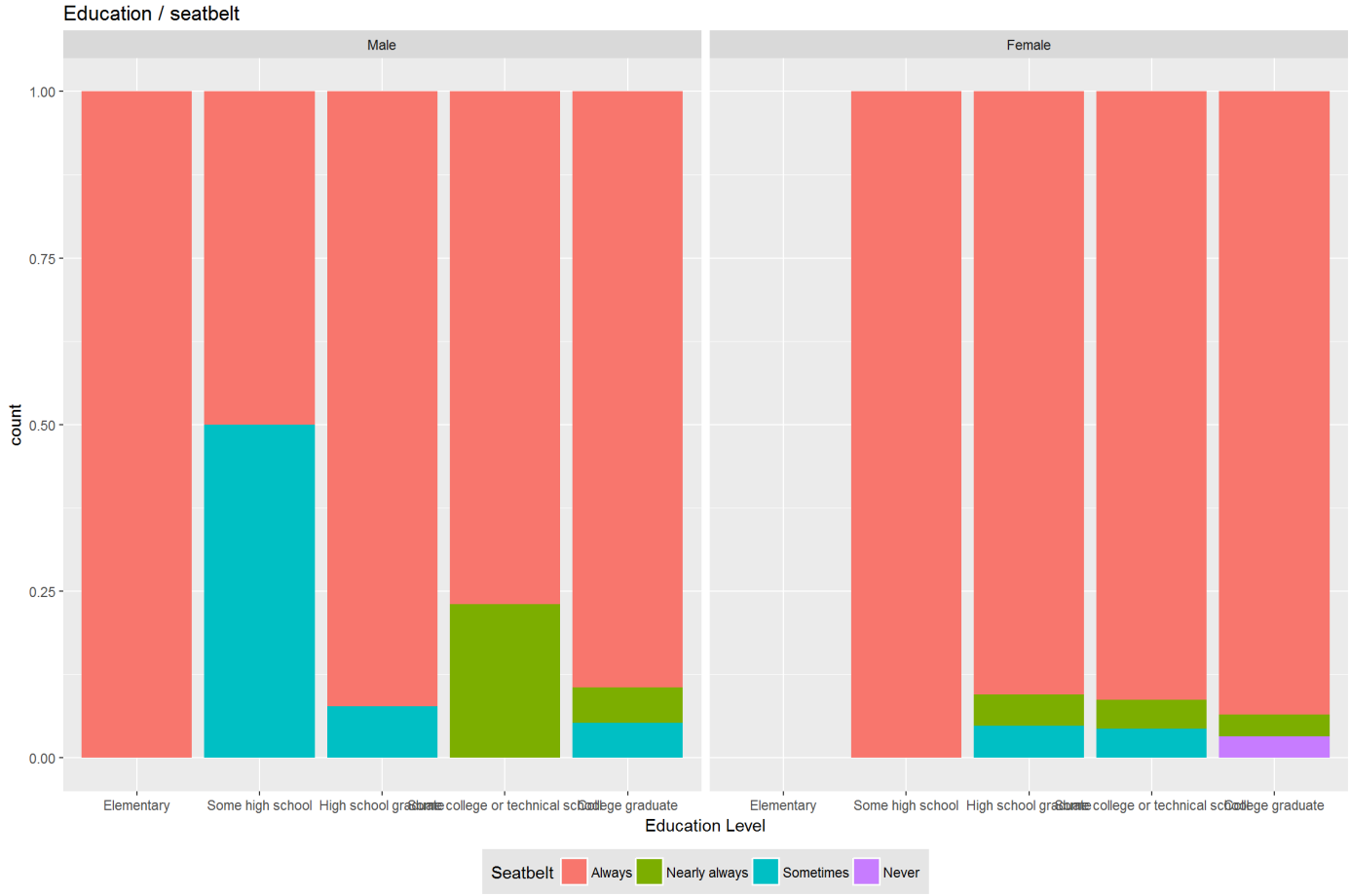
```
# education, income
ggplot(t1, aes(income2)) +
      geom_bar(aes(fill = educa), position = "dodge") +
      scale_x_discrete(name = "Income",
                       labels = c("Less than $10,000" = "< $10,000",
                                  "Less than $15,000" = "< $15,000",
                                  "Less than $20,000" = "< $20,000",
                                  "Less than $25,000" = "< $25,000",
                                  "Less than $35,000" = "< $35,000",
                                  "Less than $50,000" = "< $50,000",
                                  "Less than $75,000" = "< $75,000",
                                  "$75,000 or more" = ">= $75,000")) +
      scale_y_continuous(name = "Count") +
      ggtitle("Education vs Income") +
      labs(fill = "Education") +
      scale_fill_discrete(labels = c("Never attended school or only kindergarten" = "Never attended school",
                         "Grades 1 through 8 (Elementary)" = "Elementary",
                         "Grades 9 though 11 (Some high school)" = "Some high school",
                         "Grade 12 or GED (High school graduate)" = "High school graduate",
                         "College 1 year to 3 years (Some college or technical school)" = "Some college or tech
nical school",
                         "College 4 years or more (College graduate)" = "College graduate")) +
      theme(legend.position = "bottom",
      legend.background = element_rect(fill = "gray90", size = .5, linetype = "dotted"))
```




```
# education and sex
ggplot(t1, aes(educa)) +
      geom_bar(aes(fill = sex), position = "dodge") +
      labs(title = "Education among genders",
           x = "Education Level", y = "Count") +
      scale_x_discrete(labels = c("Never attended school or only kindergarten" = "Never attended school",
                                  "Grades 1 through 8 (Elementary)" = "Elementary",
                                  "Grades 9 though 11 (Some high school)" = "Some high school",
                                  "Grade 12 or GED (High school graduate)" = "High school graduate",
                                  "College 1 year to 3 years (Some college or technical school)" = "Some coll
ege or technical school",
                                  "College 4 years or more (College graduate)" = "College graduate")) +
      theme(legend.position = "bottom",
      legend.background = element_rect(fill = "gray90", size = .5, linetype = "dotted"))
```



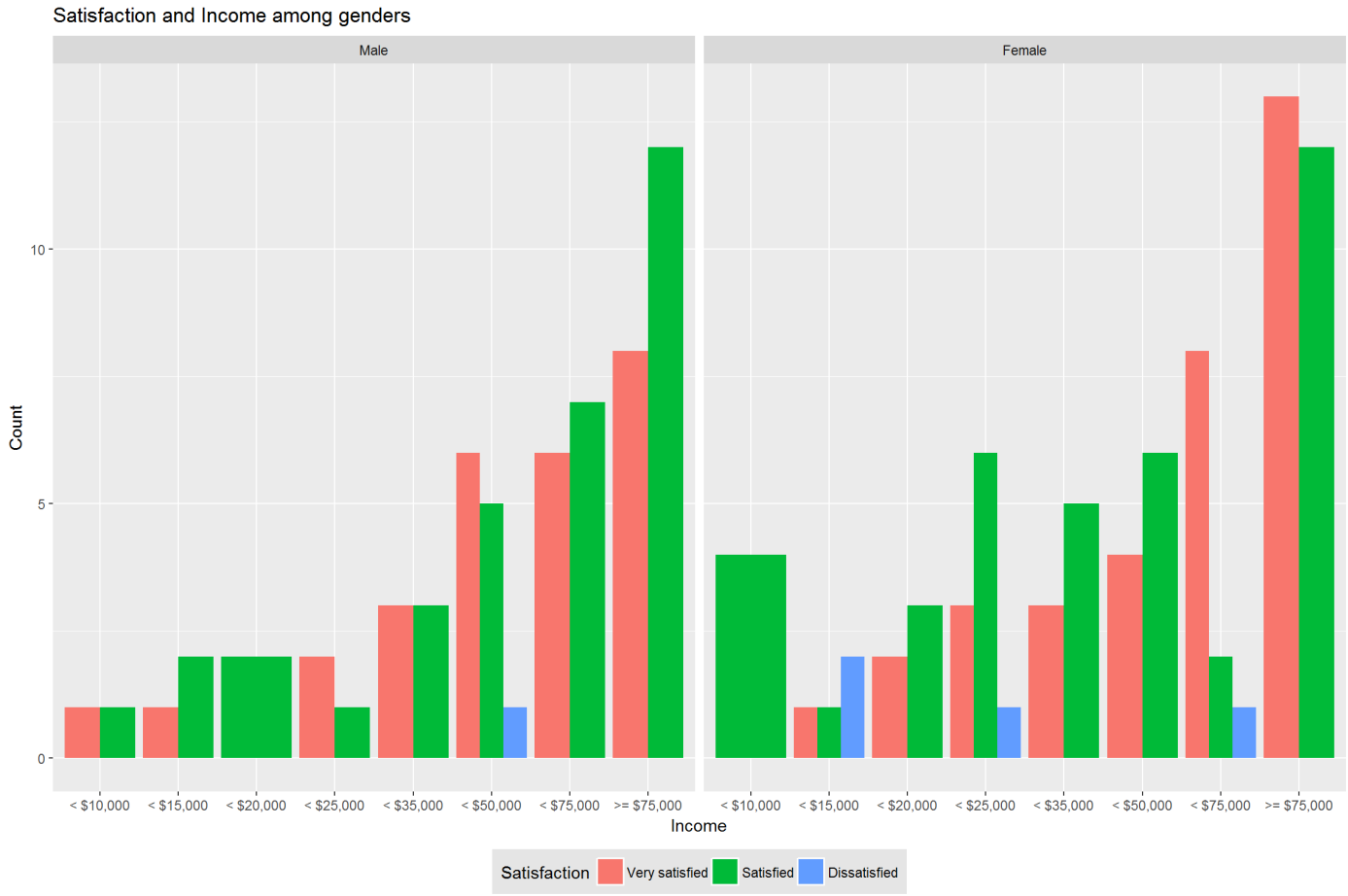

```
# education and seatbelt
ggplot(t1, aes(educa)) +
      geom_bar(aes(fill = seatbelt), position = "fill") +
      facet_grid(.~sex) +
      labs(title = "Education / seatbelt",
           x = "Education Level", fill = "Seatbelt") +
      scale_x_discrete(labels = c("Never attended school or only kindergarten" = "Never attended school",
                                  "Grades 1 through 8 (Elementary)" = "Elementary",
                                  "Grades 9 though 11 (Some high school)" = "Some high school",
                                  "Grade 12 or GED (High school graduate)" = "High school graduate",
                                  "College 1 year to 3 years (Some college or technical school)" = "Some college
or technical school",
                                  "College 4 years or more (College graduate)" = "College graduate")) +
      theme(legend.position = "bottom",
      legend.background = element_rect(fill = "gray90", size = .5, linetype = "dotted"))
```



Of course, the old concept “the more you learn, richer you’ll become” is also supported by this data, but what’s interesting here is that with a high school degree you can also receive up to $75.000, which level is dominant by the college graduates, who are women (shown by the second graph) and college graduated women are the only ones among all other educated levels women or men who never put a seatbealt on (shown by the third graph)? The question whether a person puts a seatbelt on was not stated as wheter the person puts a seatbelt on when driving, so we don’t know whether women drivers put their seatbelt on or not? But we do know that a small procent of college educated women don’t put a seatbelt on regardles of whether they are driving or not!

Research quesion 3:

```
# income and satisfaction
ggplot(t1, aes(income2)) +
      geom_bar(aes(fill = lsatisfy), position = "dodge") +
      facet_grid(.~sex) +
      scale_x_discrete(name = "Income",
                       labels = c("Less than $10,000" = "< $10,000",
                                  "Less than $15,000" = "< $15,000",
                                  "Less than $20,000" = "< $20,000",
                                  "Less than $25,000" = "< $25,000",
                                  "Less than $35,000" = "< $35,000",
                                  "Less than $50,000" = "< $50,000",
                                  "Less than $75,000" = "< $75,000",
                                  "$75,000 or more" = ">= $75,000")) +
      labs(title = "Satisfaction and Income among genders",
           x = "Income", y = "Count", fill = "Satisfaction")+
      theme(legend.position = "bottom",
      legend.background = element_rect(fill = "gray90", size = .5, linetype = "dotted"))
```



What's interesting in this graph is that women experience more often an overall dissatisfaction, regardless of their payment than men do.