Title: Assignment 2
Subtitle: Crim 250: Statistics for the Social Sciences
Name: Aisha Njie
Date: 09/23/2021

Problem 1: Load data

# Set your working directory to the folder where you downloaded the data.
setwd("/Users/isatounjie/Documents/Assignment 2")

# Read the data
dat <- read.csv("dat.nsduh.small.csv")

# What are the dimensions of the dataset?
names(dat)
#Answer: The dimensions of the dataset are mjage, cigever, alecever, AGE2, sexatract, speakengl, and irsex.

Problem 2: Variables
class(dat$mjage)
class(dat$cigever)
class(dat$alcever)
class(dat$AGE2)
class(dat$sexatract)
class(dat$speakengl)
class(dat$irsex)
# Describe the variables in the dataset.
#Answer: It appears that mjage, cigever, alcever, AGE2, sexatract, and speakengl are all ordinal variables and irsex is a categorical variable. It could also be regarded as an ordinal variable. In terms of r type, they are all intergers.

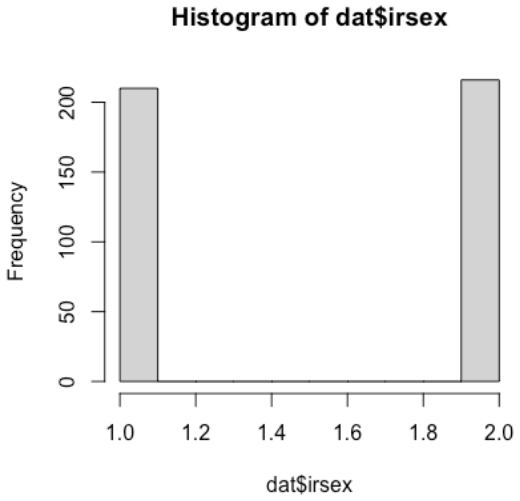
What is this dataset about? Who collected the data, what kind of sample is it, and what was the purpose of generating the data?
#Answer: This dataset observes the age at which participants first tried marijuana, first started smoking cigarettes every day, first tried alcohol, what they identify as in terms of gender, their sexual attraction, how well they speak English as well as the final recorded age of the participants. The data was collected from The National Survey on Drug Use and Health, specifically RTI International. Even though participants are selected and then interviewed, I believe this is an example of simple random sampling. Participants aren't chosen just because they fit a certain criterion.

Problem 3: Age and gender

What is the age distribution of the sample like? Make sure you read the codebook to know what the variable values mean.
#Answer: Ranges 1-10 signify a participant at one, specific age, however, ranges 11-12 signify an option between two ages, and ranges 13-16 signify a range of ages. 17 signifies the largest range with participants that are 65 or older.

Do you think this age distribution representative of the US population? Why or why not?
#Answer: Yes, I believe this is representative of the US population in the context of this study. Children start experimenting (in terms of drugs, sex, and alcohol) around the age of 12. Also, it is unlikely to receive parental consent for a study like this for children that are too young.

# Is the sample balanced in terms of gender? If not, are there more females or males?
hist(dat$irsex)

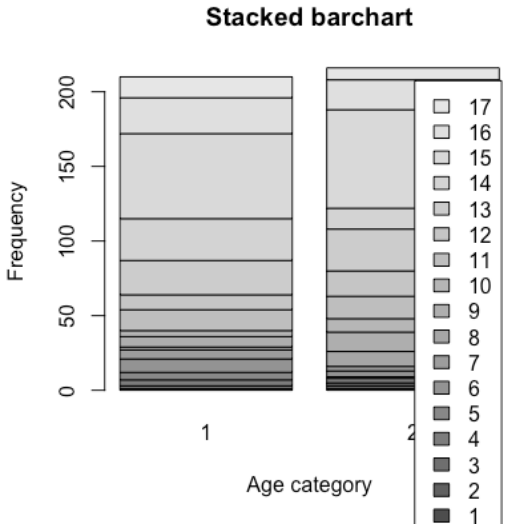


#Answer: This sample is balanced, there are 200 participants that identify as female and 200 that identify as male.

# Use this code to draw a stacked bar plot to view the relationship between sex and age. What can you conclude from this plot?
table(dat$AGE2,dat$irsex)

```
tab.agesex <- table(dat$AGE2,dat$irsex)
barplot(tab.agesex,
      main = "Stacked barchart",
      xlab = "Age category", ylab = "Frequency",
      legend.text = rownames(tab.agesex),
      beside = FALSE) # Stacked bars (default)
```

```
    1  2
1   1  1
2   0  2
3   2  2
4   4  3
5   5  1
6   9  4
7   6  3
8   2 10
9   7 13
10  4  9
11 14 15
12 10 17
13 23 28
14 28 14
15 57 66
16 24 20
17 14  8
```

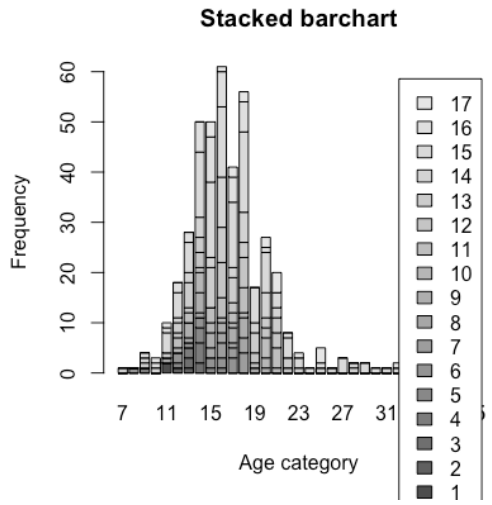


#Answer: There seems to be an equal spread in age regardless of gender. However, the frequency of older-aged women is slightly higher than that of men.

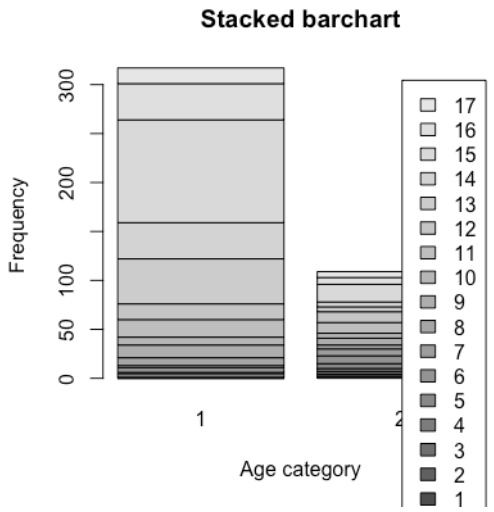
Problem 4: Substance use
For which of the three substances included in the dataset (marijuana, alcohol, and cigarettes) do individuals tend to use the substance earlier?
#Answer: Individuals tend to use alcohol earliest.

```
table(dat$AGE2,dat$mjage)
tab.agemjage <- table(dat$AGE2,dat$mjage)
```

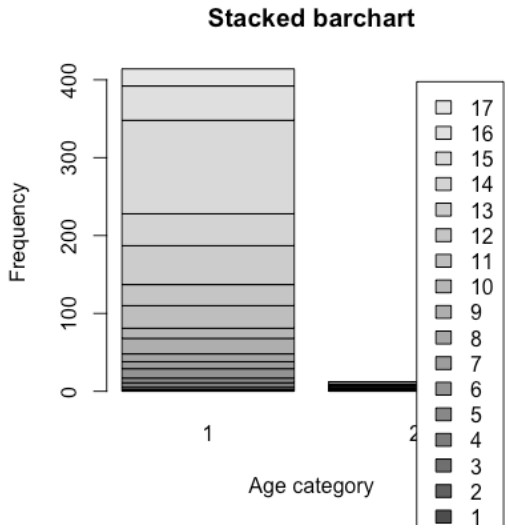
```
barplot(tab.agemjage,
        main = "Stacked barchart",
        xlab = "Age category", ylab = "Frequency",
        legend.text = rownames(tab.agemjage),
        beside = FALSE) # Stacked bars (default)
#Earliest age range is 7.
```




```
table(dat$AGE2,dat$cigever)
tab.agecigever <- table(dat$AGE2,dat$cigever)
barplot(tab.agecigever,
        main = "Stacked barchart",
        xlab = "Age category", ylab = "Frequency",
        legend.text = rownames(tab.agecigever),
        beside = FALSE) # Stacked bars (default)
#Earliest age range is 6.
```



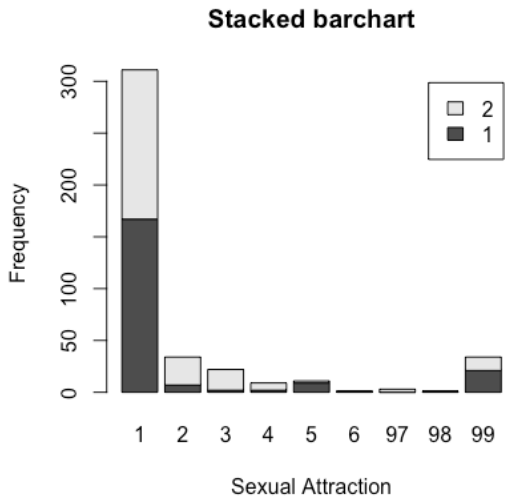

```
table(dat$AGE2,dat$alcever)
tab.agealcever <- table(dat$AGE2,dat$alcever)
barplot(tab.agealcever,
      main = "Stacked barchart",
      xlab = "Age category", ylab = "Frequency",
      legend.text = rownames(tab.agealcever),
      beside = FALSE) # Stacked bars (default)
#Earliest age range is 6.
```



#Alcohol and cigarettes start at the same age range, but the interval in alcohol's barchart is bigger, so *more* individuals start alcohol earlier.

Problem 5: Sexual attraction

```
# What does the distribution of sexual attraction look like? Is this what you expected?
table(dat$irsex,dat$sexatract)
tab.irsexatract <- table(dat$irsex,dat$sexatract)
barplot(tab.irsexatract,
      main = "Stacked barchart",
      xlab = "Sexual Attraction", ylab = "Frequency",
      legend.text = rownames(tab.irsexatract),
      beside = FALSE) # Stacked bars (default)
```

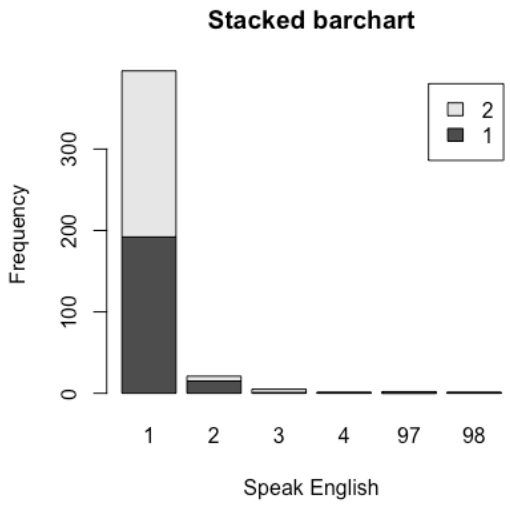


#Answer: There seems to be an equal distribution among men and women for "I am only attracted to the opposite sex." From there, the numbers are very low with individuals in 97, 98, 99 not even answering the question. I am not surprised. There's a great deal of non-binary, intersex, transgender, and other sexual identities outside of male and female that aren't accounted for. Furthermore, these individuals left out would most likely not be in the first category, thus evening out the distribution of the data.

What is the distribution of sexual attraction by gender?
#Answer: There's an equal number of men and women who are only attracted to the same sex. More women stated that there are mostly attracted to the same sex, equally attracted to the opposite sex, and mostly attracted to same sex while more men stated that there are only attracted to the same sex and not sure. More women refused to answer, but more men left the question blank and or skipped it.

Problem 6: English speaking

# What does the distribution of English speaking look like in the sample? Is this what you might expect for a random sample of the US population?
table(dat$irsex,dat$speakengl)
tab.irsexspeakengl <- table(dat$irsex,dat$speakengl)
barplot(tab.irsexspeakengl,
    main = "Stacked barchart",
    xlab = "Speak English", ylab = "Frequency",
    legend.text = rownames(tab.irsexspeakengl),
    beside = FALSE) # Stacked bars (default)



#Answer: There is an extremely high frequency of individuals that can speak English very well. A small percentage of individuals refused or left this question blank. Another small group of individuals stated they spoke English well.

Are there more English speaker females or males?
#Answer: There are more female English speakers.