Assignments

This page will contain all the assignments you submit for the class.

Instructions for all assignments

I want you to submit your assignment as a PDF, so I can keep a record of what the code looked like that day. I also want you to include your answers on your personal GitHub website. This will be good practice for editing your website and it will help you produce something you can keep after the class is over.

1. Download the Assignment1.Rmd file from Canvas. You can use this as a template for writing your answers. It's the same as what you can see on my website in the Assignments tab. Once we're done with this I'll edit the text on the website to include the solutions.
2. On RStudio, open a new R script in RStudio (File > New File > R Script). This is where you can test out your R code. You'll write your R commands and draw plots here.
3. Once you have finalized your code, copy and paste your results into this template (Assignment 1.Rmd). For example, if you produced a plot as the solution to one of the problems, you can copy and paste the R code in R markdown by using the ```` ```{r} ``` ```` command. Answer the questions in full sentences and Save.
4. Produce a PDF file with your answers. To do this, knit to PDF (use Knit button at the top of RStudio), locate the PDF file in your docs folder (it's in the same folder as the Rproj), and submit that on on Canvas in Assignment 1.
5. Build Website, go to GitHub desktop, commit and push. Now your solutions should be on your website as well.

Assignment 1

Collaborators: Lorem Ipsum.

This assignment is due on Canvas on Monday 9/20 before class, at 10:15 am. Include the name of anyone with whom you collaborated at the top of the assignment.

Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

```
options(repos = list(CRAN="http://cran.rstudio.com/"))
#install.packages ("dataset_load")
```

Load the USArrests dataset and rename it `dat`. Note that this dataset comes with R, in the package datasets, so there's no need to load data from your computer. Why is it useful to rename the dataset?

```
dat <- USArrests
```

Answer: It is useful to rename datasets because it is good practice and it is more convenient to use instead of the full names of data sets which are usually longer. Also, might reduce errors in using a short and simple name such as dat.

Problem 2

Use this command to make the state names into a new variable called State.

```
dat$state <- tolower(rownames(USArrests))
```

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset `USArrests`.

```
names(dat)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"     "state"
```

The five variables are Murder, Assault, UrbanPop, Rape, and State.

Problem 3

What type of variable (from the DVB chapter) is `Murder`?

Answer: Murder is a quantitative variable.

What R Type of variable is it?

```
class(dat$Murder)
```

```
## [1] "numeric"
```

Answer: Murder is a numeric value.

Problem 4

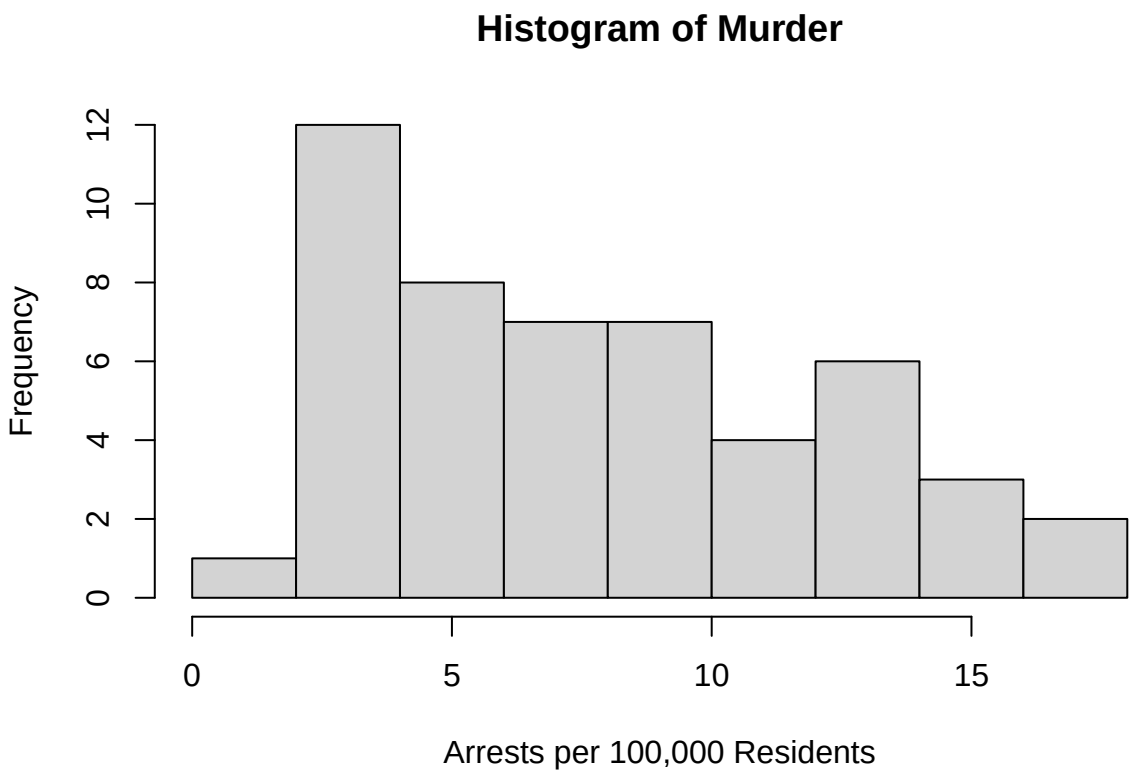
What information is contained in this dataset, in general? What do the numbers mean?

Answer: This dataset includes the number of murder, assault, urbanpop, and rape cases throughout 50 states. The datasets relies on the recorded cases of crimes that offenders/criminals commit. It was most likely collected from the series of reported crime statisitcs on the internet or perhaps even the Federal Bureau of Justice Statistics. The numbers represent the frequency of that crime for each state. I assume that reserchers of crime rates,statiticians in the field of law enforcement or legal justice created this dataset to compare crime rates across the U.S. but also the frequency of the different crimes against each other.

Problem 5

Draw a histogram of `Murder` with proper labels and title.

```
hist(dat$Murder, main="Histogram of Murder", xlab="Arrests per 100,000 Residents", ylab="Frequency")
```



Problem 6

Please summarize `Murder` quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

```
summary(dat$Murder)
```

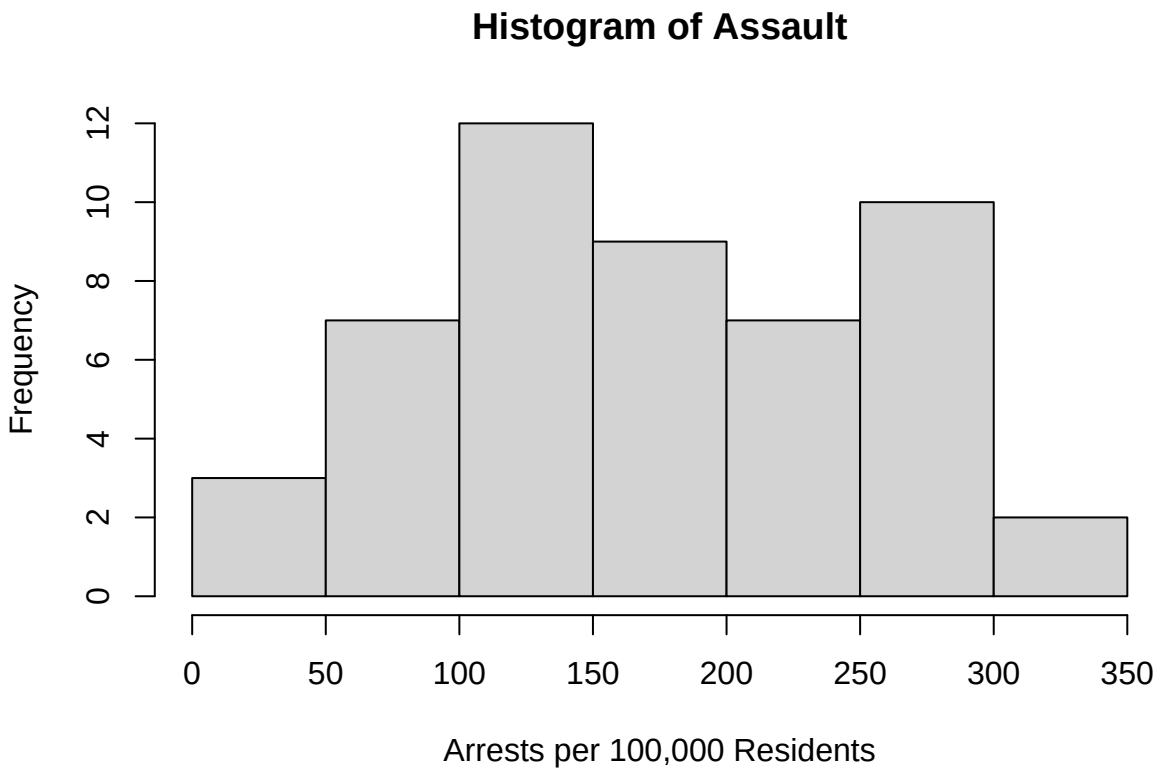
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

The mean is 7.788 and the median is 7.250. Mean is the average of the data set. It is found by adding all the numbers in the data set and then dividing by the number of values in the set. The median is the middle value when a data set is ordered from least to greatest. A quartile is a type of quantile which divides the data set into four parts. You can deduce the interquartile range (IQR) from Q1 and Q3 and this is significant because the IQR, also known as the midspread/middle 50%/H spread is a measure of statistical dispersion or the variability in a data set.

Problem 7 (a)

Repeat the same steps you followed for `Murder`, for the variables `Assault` and `Rape`.

```
hist(dat$Assault, main="Histogram of Assault", xlab="Arrests per 100,000 Residents", ylab="Frequency")
```

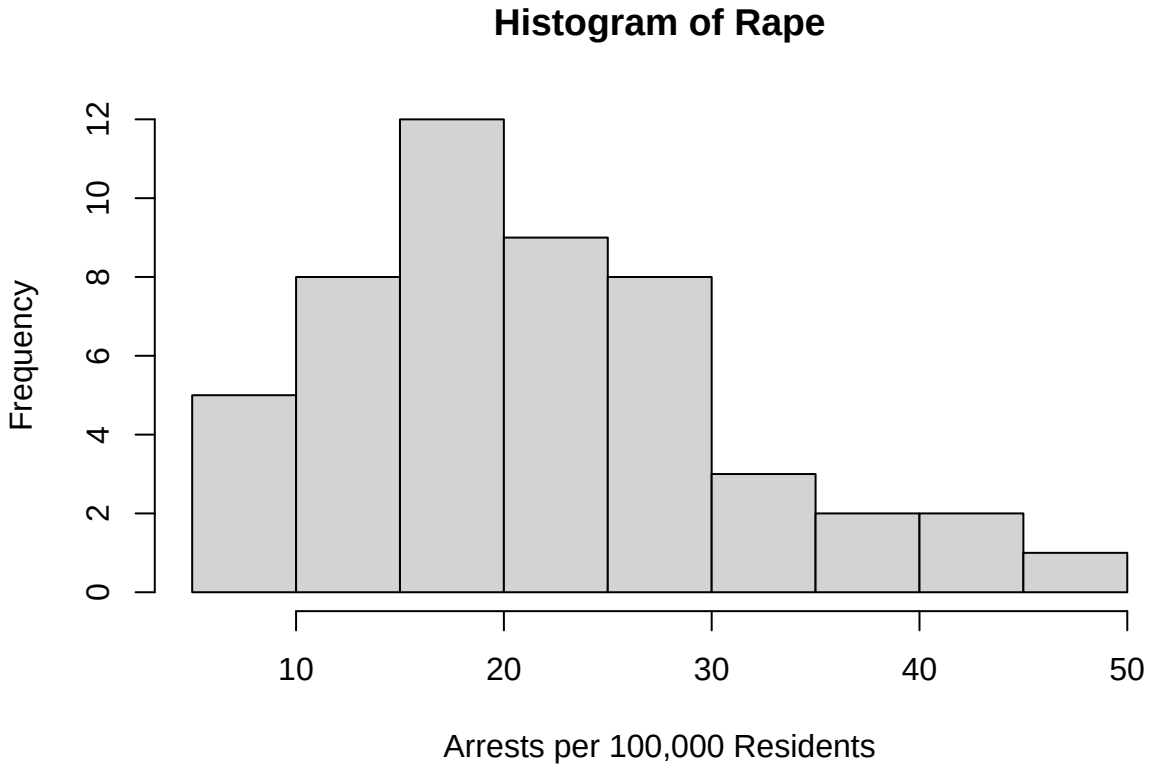



```
summary(dat$Assault)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

The mean is 170.8 and the median is 159.0.

```
hist(dat$Rape, main="Histogram of Rape", xlab="Arrests per 100,000 Residents", ylab="Frequency")
```




```
summary(dat$Rape)
```

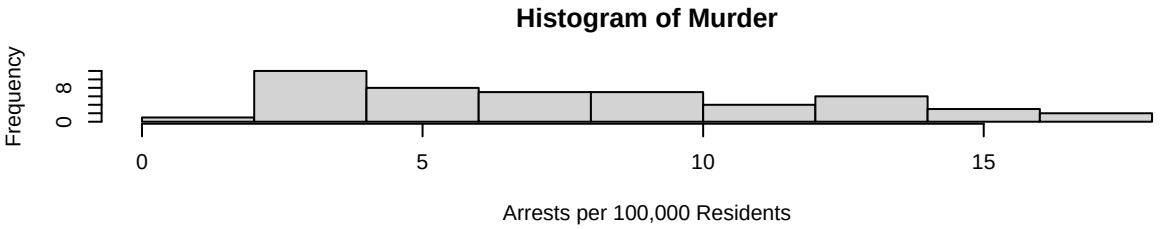
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

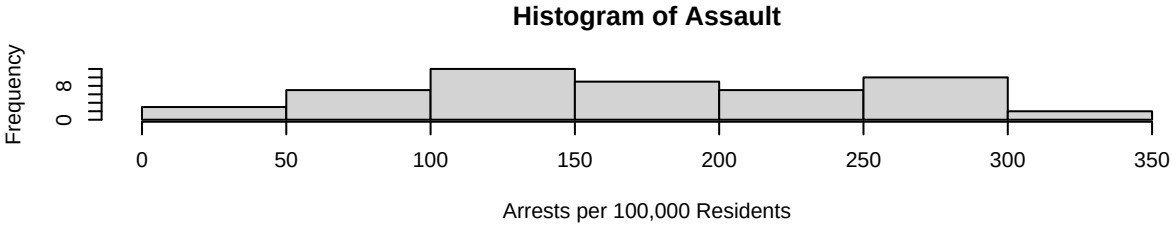
The mean is 21.23 and the median is 20.10.

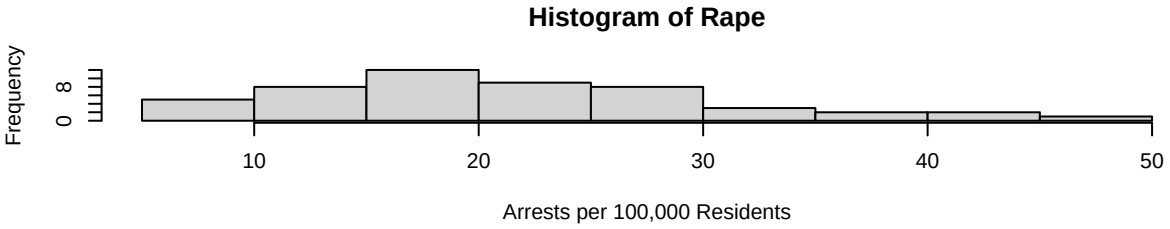
Problem 7 (b)

Now plot all three histograms together. You can do this by using the command `par(mfrow=c(3,1))` and then plotting each of the three.

```
par(mfrow=c(3,1))
hist(dat$Murder, main="Histogram of Murder", xlab="Arrests per 100,000 Residents", ylab="Frequency")
hist(dat$Assault, main="Histogram of Assault", xlab="Arrests per 100,000 Residents", ylab="Frequency")
hist(dat$Rape, main="Histogram of Rape", xlab="Arrests per 100,000 Residents", ylab="Frequency")
```







What does the command par do, in your own words (you can look this up by asking R `?par`)?

```
?par
```

Answer: par can be used to set either give you information about graphs and/or let you set parameters for graphs.

What can you learn from plotting the histograms together?

Answer: By plotting the histogams together, we can observe the scale at which the different crimes occurred. You can compare the frequencies across the different crimes too.

Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
install.packages("maps")
install.packages("ggplot2")

library(maps)
library(ggplot2)

ggplot(dat, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```

What does this code do? Explain what each line is doing.

Answer: The first line determines the dimensions of the graph. The second and third line installs the package necessary to make the graph, specifically a map. The fourth and fifth lines load the library of the two

packages necessary to construct a map. The sixth lines tells the map to only include states and the frequency of Murder in each state. The last three lines serves as the data frame that contains the map coordinates.

Assignment 2

Problem 1: Load data

Set your working directory to the folder where you downloaded the data.

```
setwd("/Users/isatounjie/Documents/GitHub/Aishas-Website/Assignment 2")
```

Read the data

```
dat <- read.csv("dat.nsduh.small.1.csv")
```

What are the dimensions of the dataset?

```
names(dat)
```

```
## [1] "mjage"     "cigage"    "iralcage"  "age2"      "sexatract" "speakengl"
## [7] "irsex"
```

Answer: The dimensions of the dataset are mjage, cigever, alecever, AGE2, sexatract, speakengl, and irsex.

Problem 2: Variables

```
class(dat$mjage)
```

```
## [1] "integer"
```

```
class(dat$cigage)
```

```
## [1] "integer"
```

```
class(dat$iralcage)
```

```
## [1] "integer"
```

```
class(dat$age2)
```

```
## [1] "integer"
```

```
class(dat$sexatract)
```

```
## [1] "integer"
```

```
class(dat$speakengl)
```

```
## [1] "integer"
```

```
class(dat$irsex)
```

```
## [1] "integer"
```

Describe the variables in the dataset.

Answer: It appears that mjage, cigage, iralcage, 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. Paticipants aren't chosen just because they fit a certain criteria.

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.

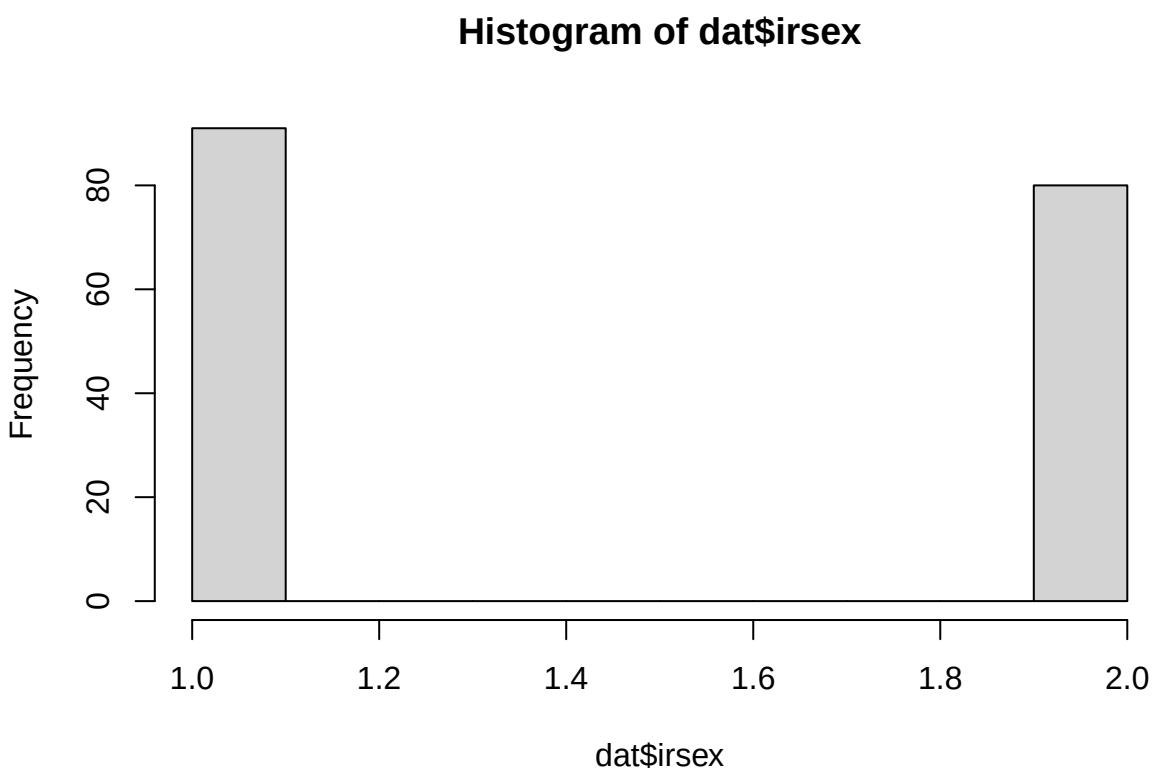
Answer: Ranges 1-10 signify a participant at one, specific age, however ranges 11-12 signify an option between two ages, and ranges 13-16 signify a range of ages. 17 signifies the largest range with participants that are 65 or older.

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.

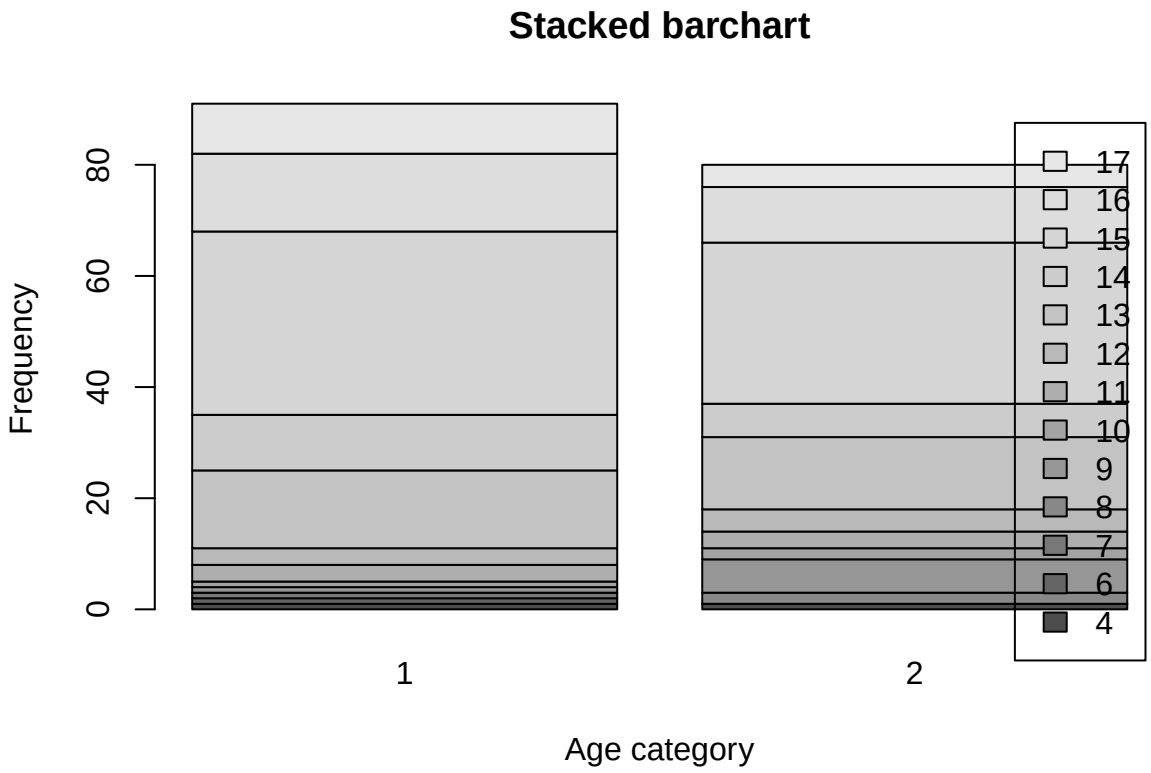
Use this code to draw a stacked bar plot to view the relationship between sex and age. What can you conclude from this plot?

```
table(dat$age2,dat$irsex)
```

```
##
```

```
##        1  2
##   4    1  1
##   6    1  0
##   7    1  0
##   8    0  2
##   9    1  6
##   10   1  2
##   11   3  3
##   12   3  4
##   13  14 13
##   14  10  6
##   15  33 29
##   16  14 10
##   17   9  4
```

```
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)
```



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

```
table(dat$age2,dat$mjage)
```

```
##
```

```
##       7 9 10 11 12 13 14 15 16 17 18 19 20 21 22 25 27 30 32 33 35
##   4   0 1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   6   0 0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   7   0 0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
##   8   0 0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
##   9   0 0  0  0  0  2  2  1  0  2  0  0  0  0  0  0  0  0  0  0  0
##   10  0 0  0  0  0  0  1  1  0  1  0  0  0  0  0  0  0  0  0  0  0
##   11  0 0  0  0  1  1  0  1  1  1  1  0  0  0  0  0  0  0  0  0  0
##   12  0 0  0  0  1  1  0  0  3  1  1  0  0  0  0  0  0  0  0  0  0
##   13  1 1  0  1  2  4  2  2  5  3  0  2  0  2  1  1  0  0  0  0  0
##   14  0 1  1  0  0  1  3  0  5  1  1  0  2  0  0  0  0  0  0  1  0
##   15  0 0  0  3  5  5  9  8  7  5  8  2  4  1  1  0  2  1  0  0  1
##   16  0 1  1  1  1  1  4  6  6  1  2  0  0  0  0  0  0  0  0  0  0
##   17  0 0  0  1  0  0  0  3  0  1  2  0  1  3  0  1  0  0  1  0  0
```

```
table(dat$age2,dat$cigage)
```

```
##
##       10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 35 45 50
##   4    0  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   6    0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   7    0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   8    0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0
##   9    0  0  0  0  1  2  0  2  2  0  0  0  0  0  0  0  0  0  0  0
##   10   0  0  0  0  0  0  1  1  1  0  0  0  0  0  0  0  0  0  0  0
##   11   0  0  0  0  0  1  1  0  3  0  1  0  0  0  0  0  0  0  0  0
##   12   0  0  1  0  1  0  0  1  2  2  0  0  0  0  0  0  0  0  0  0
##   13   0  0  1  0  2  6  4  4  4  2  2  1  0  0  1  0  0  0  0  0
##   14   0  0  0  0  0  0  3  3  4  1  0  1  1  2  0  1  0  0  0  0
##   15   1  0  0  5  3 10  8  2 12  6  4  4  3  1  0  1  1  1  0  0
##   16   0  0  1  3  1  4  7  2  2  0  2  0  0  1  0  0  0  0  0  1
##   17   0  1  0  1  0  1  1  3  1  0  1  0  1  0  0  2  0  0  1  0
```

```
table(dat$age2,dat$iralcage)
```

```
##
##       5  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 23
##   4   0  0  0  0  0  1  0  1  0  0  0  0  0  0  0  0  0
##   6   0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
##   7   0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0
##   8   0  0  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0
##   9   0  0  0  0  0  0  1  2  1  2  1  0  0  0  0  0  0
##   10  0  0  0  0  0  0  1  0  1  0  0  0  1  0  0  0  0
##   11  0  0  0  0  0  0  0  0  1  1  0  0  1  1  1  1  0
##   12  0  0  0  0  0  0  2  0  1  1  1  1  1  0  0  0  0
##   13  1  0  0  1  0  0  0  5  4  7  2  2  1  1  0  3  0
##   14  0  0  0  0  0  0  1  1  2  1  6  0  3  0  0  2  0
##   15  0  0  2  0  2  2 10  9  6  4  7  7 10  2  0  0  1
##   16  1  1  0  0  1  0  3  1  5  3  5  2  2  0  0  0  0
##   17  0  0  0  0  0  0  1  2  1  0  2  0  4  2  1  0  0
```

For which of the three substances included in the dataset (marijuana, alcohol, and cigarettes) do individuals tend to use the substance earlier?

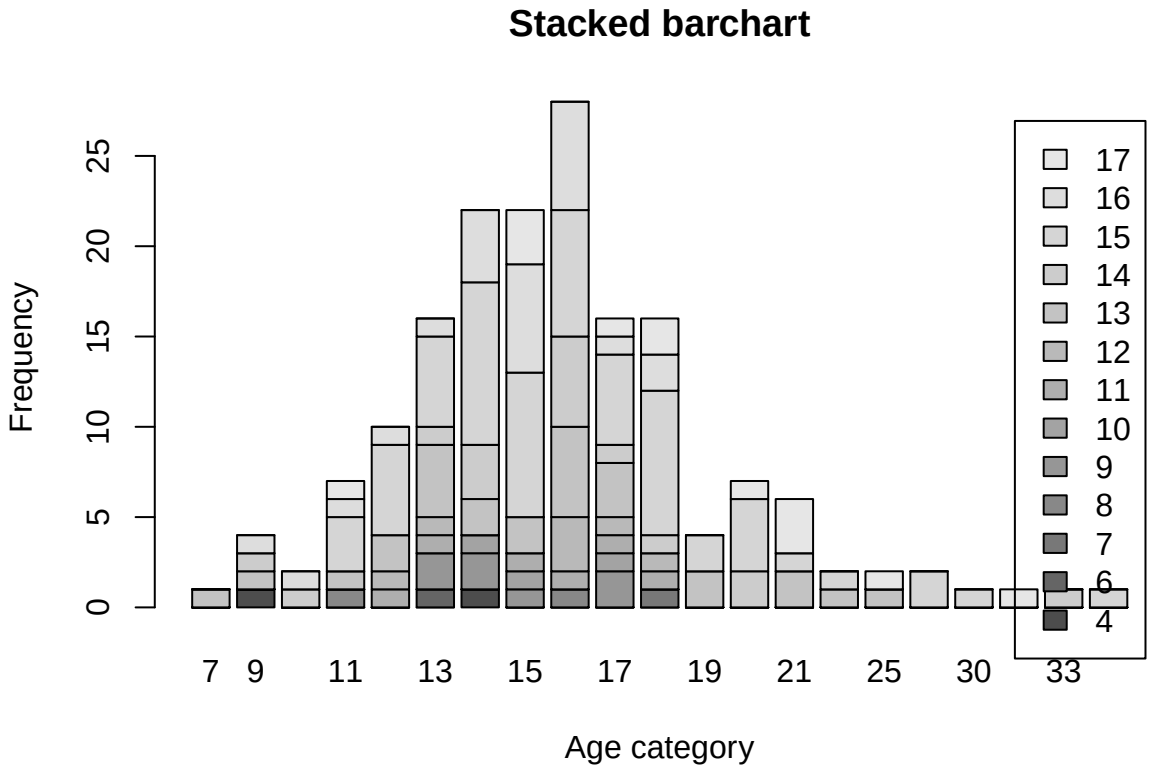
Individuals tend to use alcohol earliest.

CHECKING ANSWER

```
table(dat$age2,dat$mjage)
```

```
##
##      7 9 10 11 12 13 14 15 16 17 18 19 20 21 22 25 27 30 32 33 35
##   4  0 1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   6  0 0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   7  0 0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
##   8  0 0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
##   9  0 0  0  0  0  2  2  1  0  2  0  0  0  0  0  0  0  0  0  0  0
##   10 0 0  0  0  0  0  1  1  0  1  0  0  0  0  0  0  0  0  0  0  0
##   11 0 0  0  0  1  1  0  1  1  1  1  0  0  0  0  0  0  0  0  0  0
##   12 0 0  0  0  1  1  0  0  3  1  1  0  0  0  0  0  0  0  0  0  0
##   13 1 1  0  1  2  4  2  2  5  3  0  2  0  2  1  1  0  0  0  0  0
##   14 0 1  1  0  0  1  3  0  5  1  1  0  2  0  0  0  0  0  0  1  0
##   15 0 0  0  3  5  5  9  8  7  5  8  2  4  1  1  0  2  1  0  0  1
##   16 0 1  1  1  1  1  4  6  6  1  2  0  0  0  0  0  0  0  0  0  0
##   17 0 0  0  1  0  0  0  3  0  1  2  0  1  3  0  1  0  0  1  0  0
```

```
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)
```



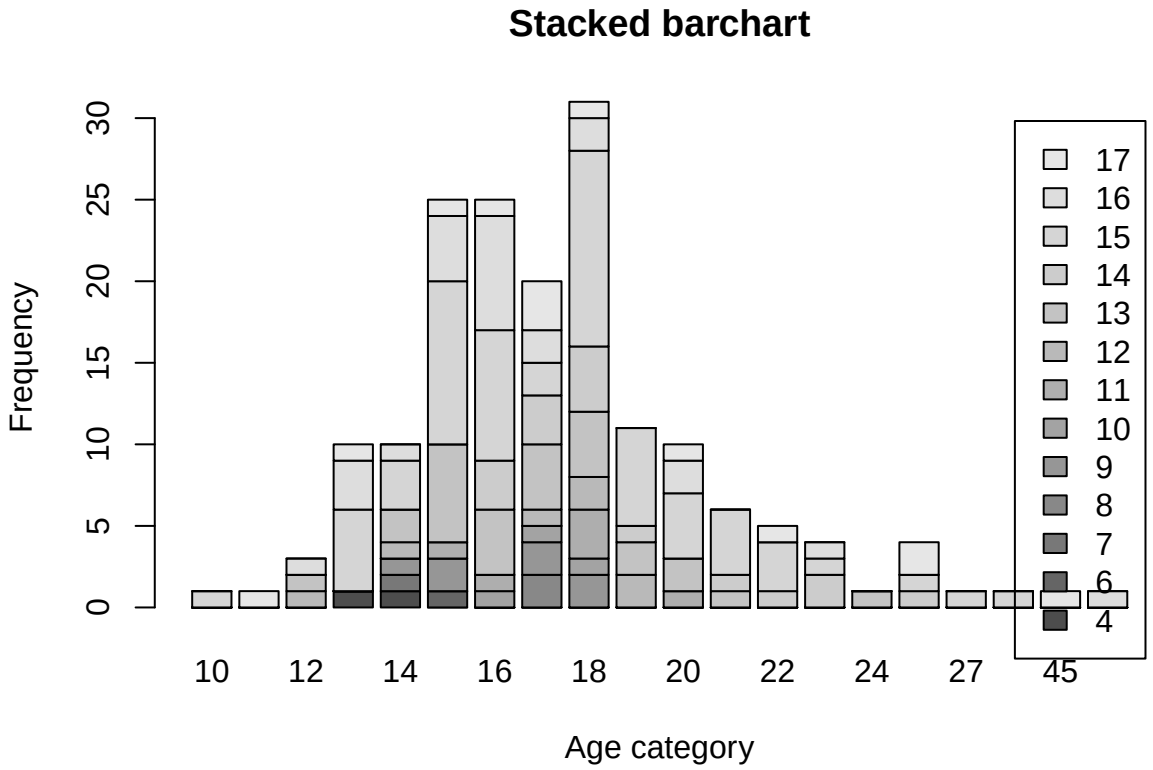
Earliest age range is 7.

```
table(dat$age2,dat$cigage)
```

```
##
##      10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 35 45 50
```

```
##    4   0  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##    6   0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##    7   0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##    8   0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0
##    9   0  0  0  0  1  2  0  2  2  0  0  0  0  0  0  0  0  0  0  0
##    10  0  0  0  0  0  0  1  1  1  0  0  0  0  0  0  0  0  0  0  0
##    11  0  0  0  0  0  1  1  0  3  0  1  0  0  0  0  0  0  0  0  0
##    12  0  0  1  0  1  0  0  1  2  2  0  0  0  0  0  0  0  0  0  0
##    13  0  0  1  0  2  6  4  4  4  2  2  1  0  0  1  0  0  0  0  0
##    14  0  0  0  0  0  0  3  3  4  1  0  1  1  2  0  1  0  0  0  0
##    15  1  0  0  5  3 10  8  2 12  6  4  4  3  1  0  1  1  1  0  0
##    16  0  0  1  3  1  4  7  2  2  0  2  0  0  1  0  0  0  0  0  1
##    17  0  1  0  1  0  1  1  3  1  0  1  0  1  0  0  2  0  0  1  0
```

```
tab.agecigage <- table(dat$age2,dat$cigage)
barplot(tab.agecigage,
        main = "Stacked barchart",
        xlab = "Age category", ylab = "Frequency",
        legend.text = rownames(tab.agecigage),
        beside = FALSE) # Stacked bars (default)
```



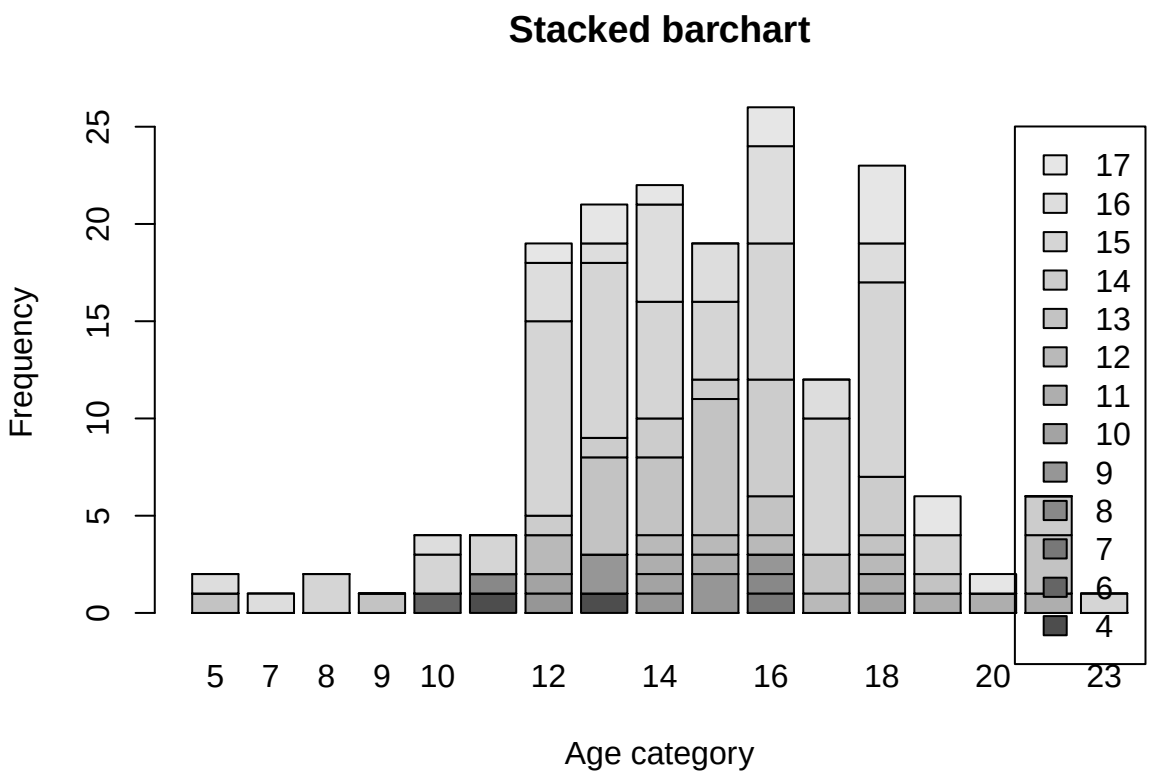
Earliest age range is 6.

```
table(dat$age2,dat$iralcage)
```

```
##
##       5  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 23
##    4  0  0  0  0  0  1  0  1  0  0  0  0  0  0  0  0  0
##    6  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
##    7  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0
##    8  0  0  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0
```

```
##    9   0  0  0  0  0  0  1  2  1  2  1  0  0  0  0  0  0
##    10  0  0  0  0  0  0  1  0  1  0  0  0  1  0  0  0  0
##    11  0  0  0  0  0  0  0  0  1  1  0  0  1  1  1  1  0
##    12  0  0  0  0  0  0  2  0  1  1  1  1  1  0  0  0  0
##    13  1  0  0  1  0  0  0  5  4  7  2  2  1  1  0  3  0
##    14  0  0  0  0  0  0  1  1  2  1  6  0  3  0  0  2  0
##    15  0  0  2  0  2  2 10  9  6  4  7  7 10  2  0  0  1
##    16  1  1  0  0  1  0  3  1  5  3  5  2  2  0  0  0  0
##    17  0  0  0  0  0  0  1  2  1  0  2  0  4  2  1  0  0
```

```
tab.ageiralcage <- table(dat$age2,dat$iralcage)
barplot(tab.ageiralcage,
        main = "Stacked barchart",
        xlab = "Age category", ylab = "Frequency",
        legend.text = rownames(tab.ageiralcage),
        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 barchat is bigger, so more individuals start alcohol earlier.

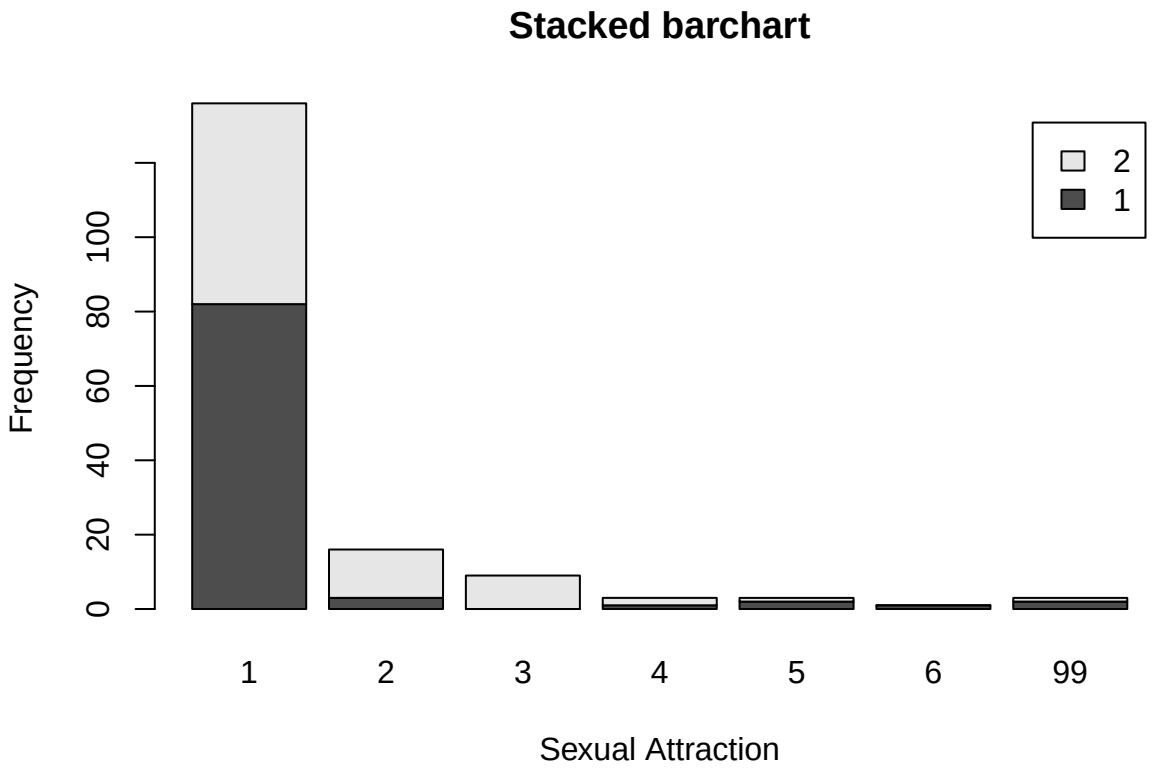
Problem 5: Sexual attraction

What does the distribution of sexual attraction look like? Is this what you expected?

```
table(dat$irsex,dat$sexatract)
```

```
##
##      1  2  3  4  5  6 99
##    1 82  3  0  1  2  1  2
##    2 54 13  9  2  1  0  1
```

```
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 oppisote 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 oppisote 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.

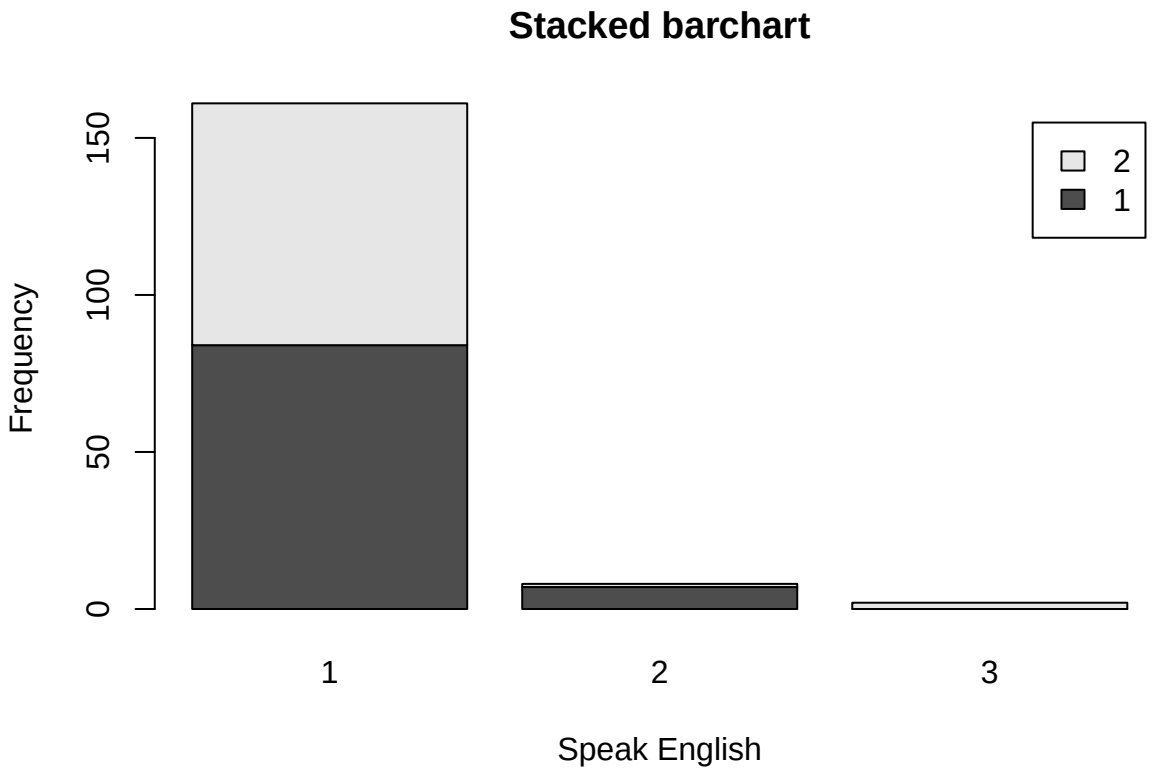
Problem 6: English speaking

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)
```

```
##
##      1  2  3
##   1 84  7  0
##   2 77  1  2
```

```r
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.