Assignments

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

TEST

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

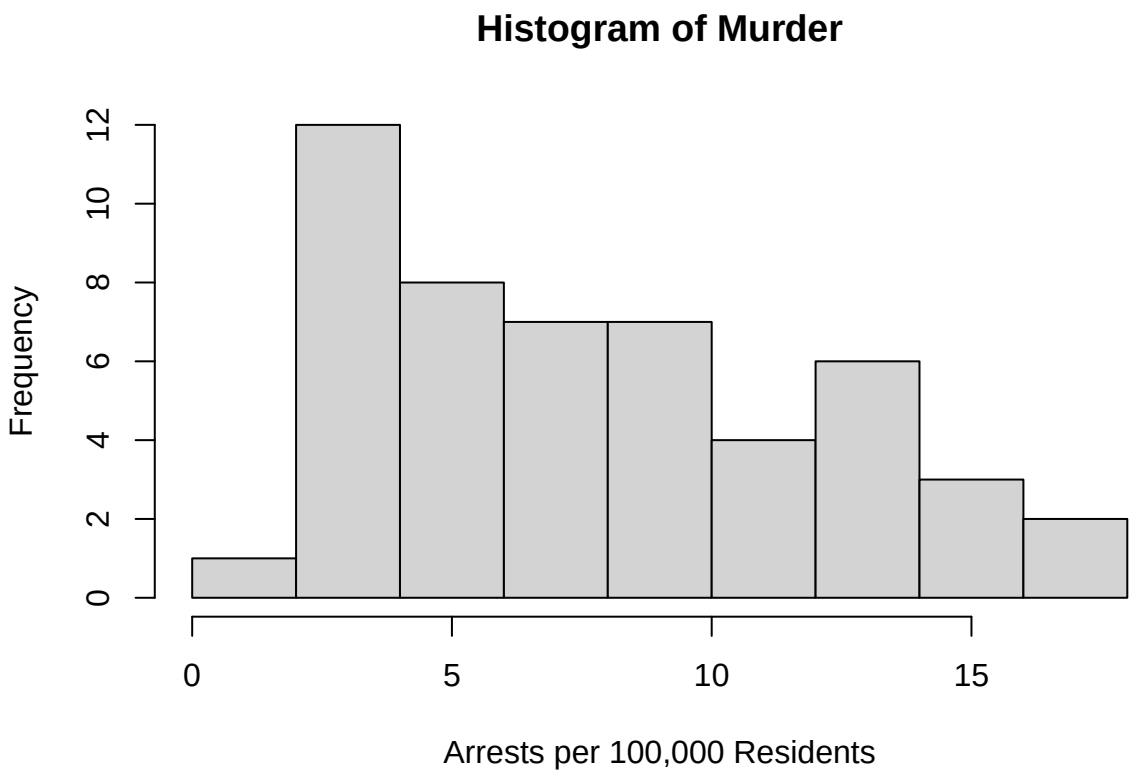
What information is contained in this dataset, in general? What do the numbers mean?

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)
```

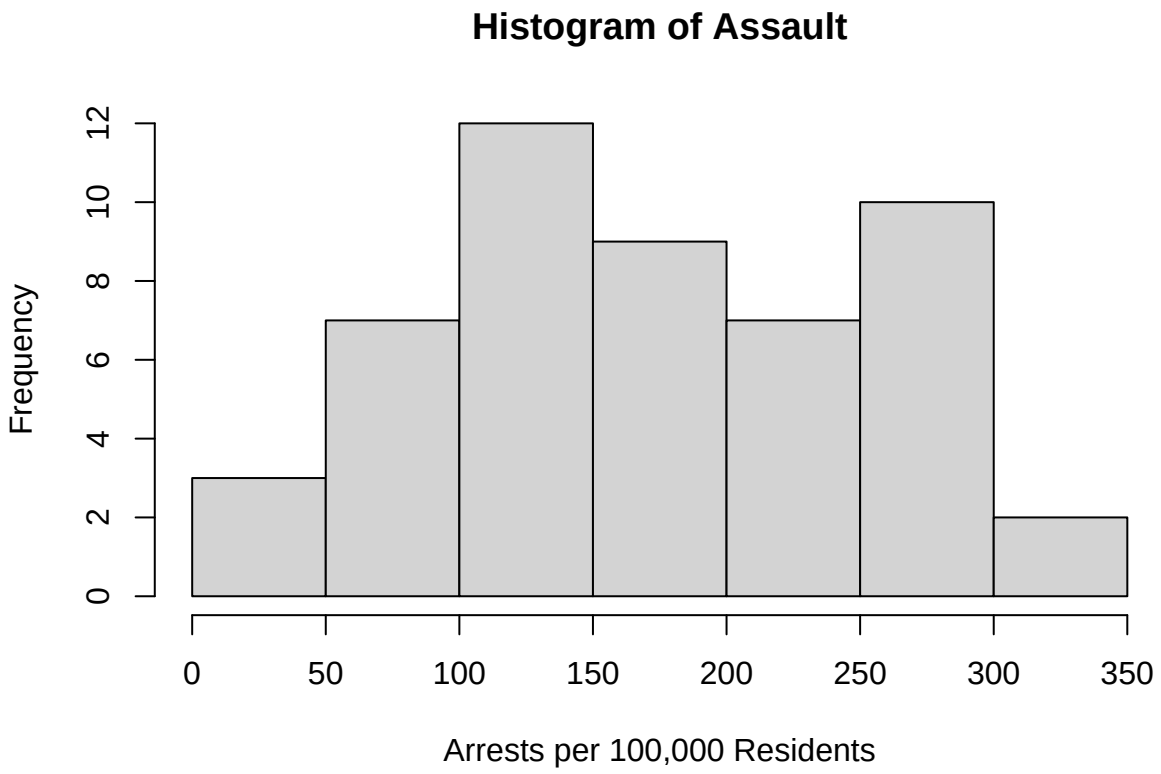
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

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")
```

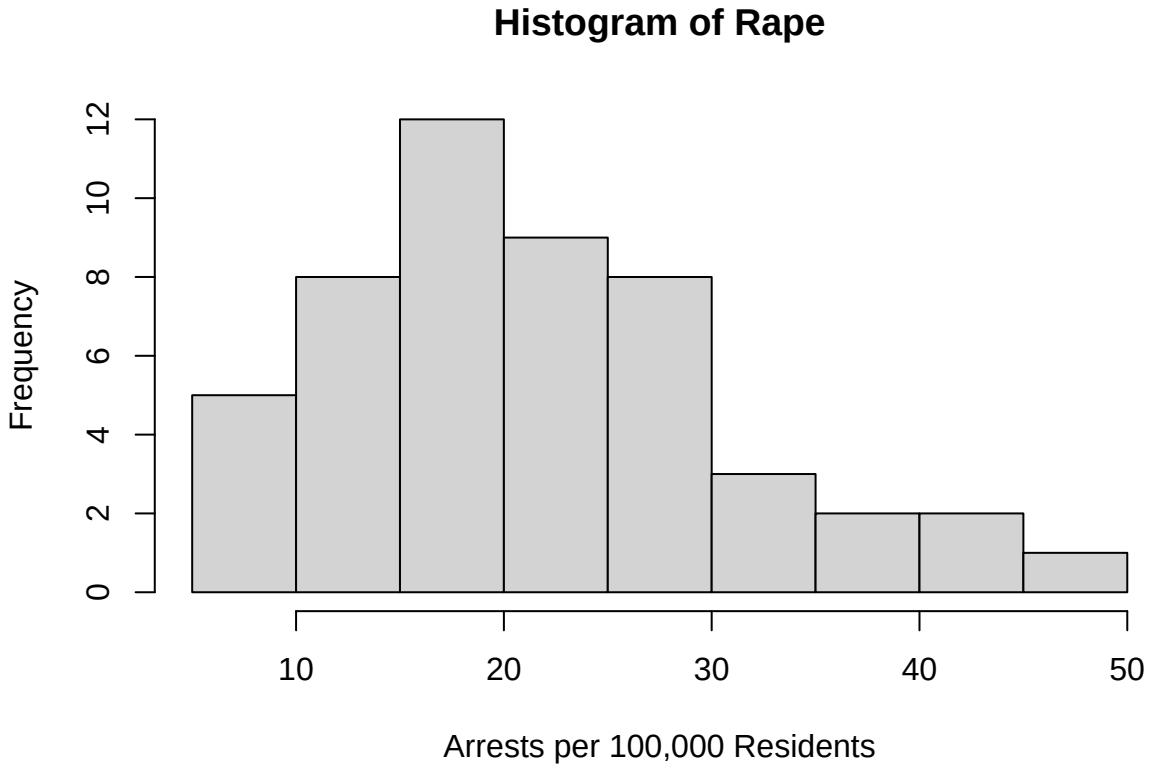



```
summary(dat$Assault)
```

```
##    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)
```

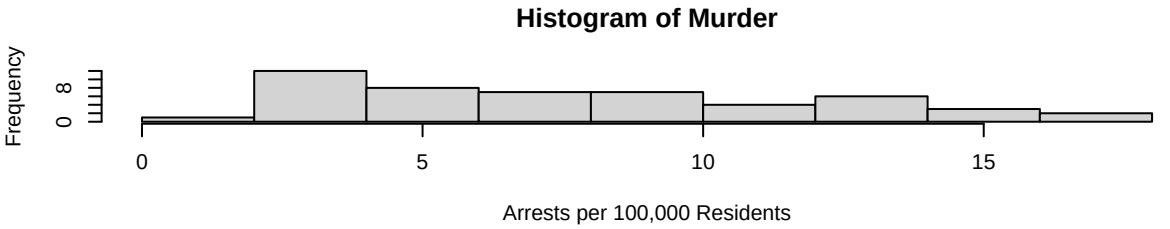
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

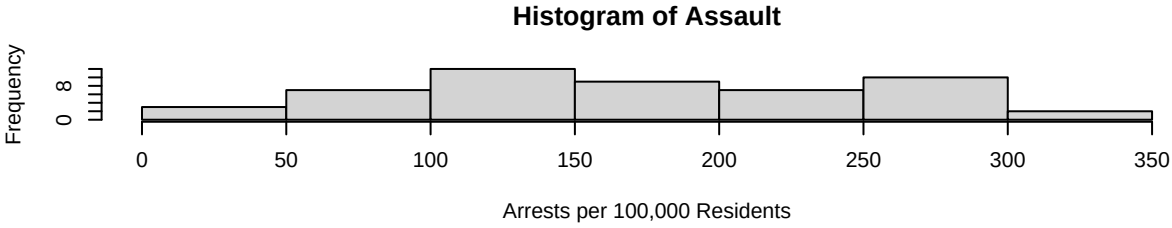
The mean is 21.23 and the median is 20.10.

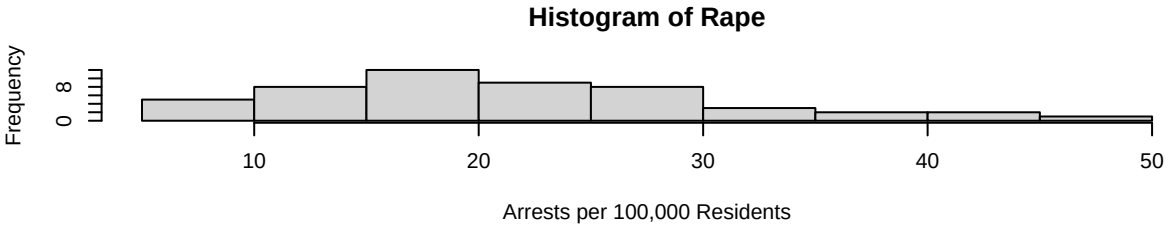
Problem 7 (b)

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, cigage, iralcage, 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. Participants 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.

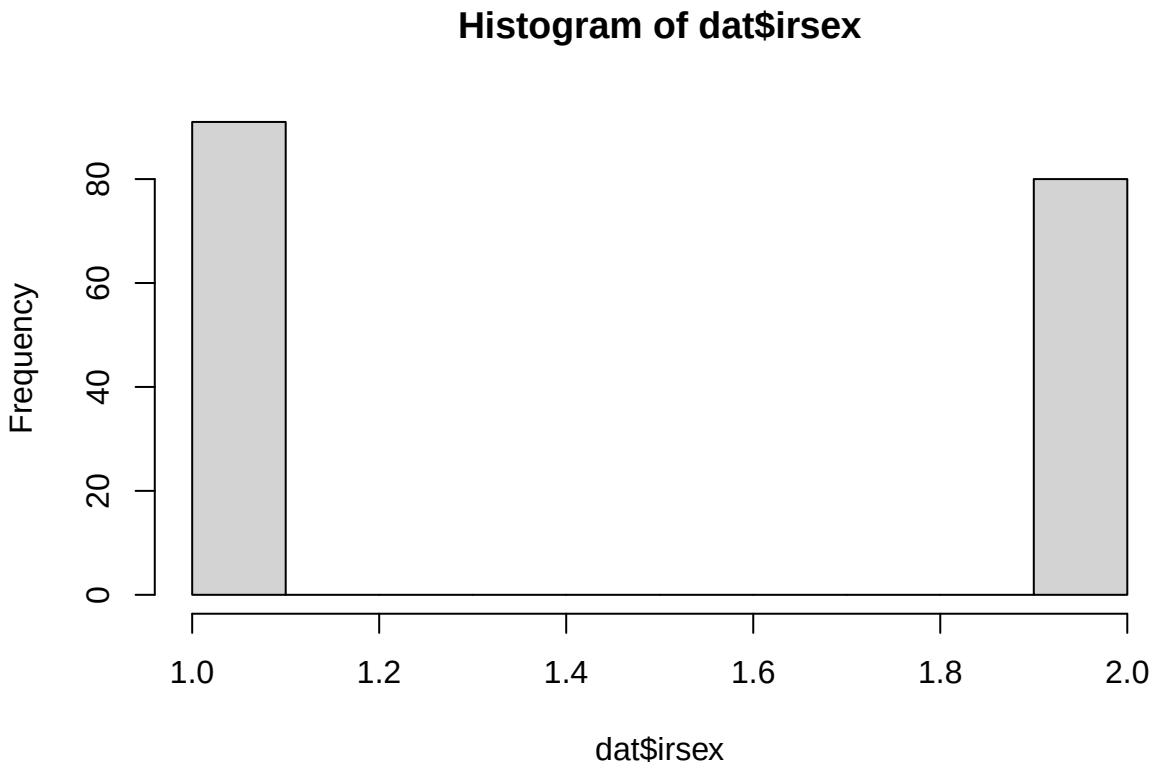
Do you think this age distribution representative of the US population? Why or why not?

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.

*Correction: They cut off the survey at age 12 by design, not because they found people used drugs less below that age (although that is probably also true). There are other surveys of children's use of drugs.

Is the sample balanced in terms of gender? If not, are there more females or males?

```
hist(dat$irsex)
```



Answer: This sample is nearly balanced, there are 91 participants that identify as male and 80 that identify as female.

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

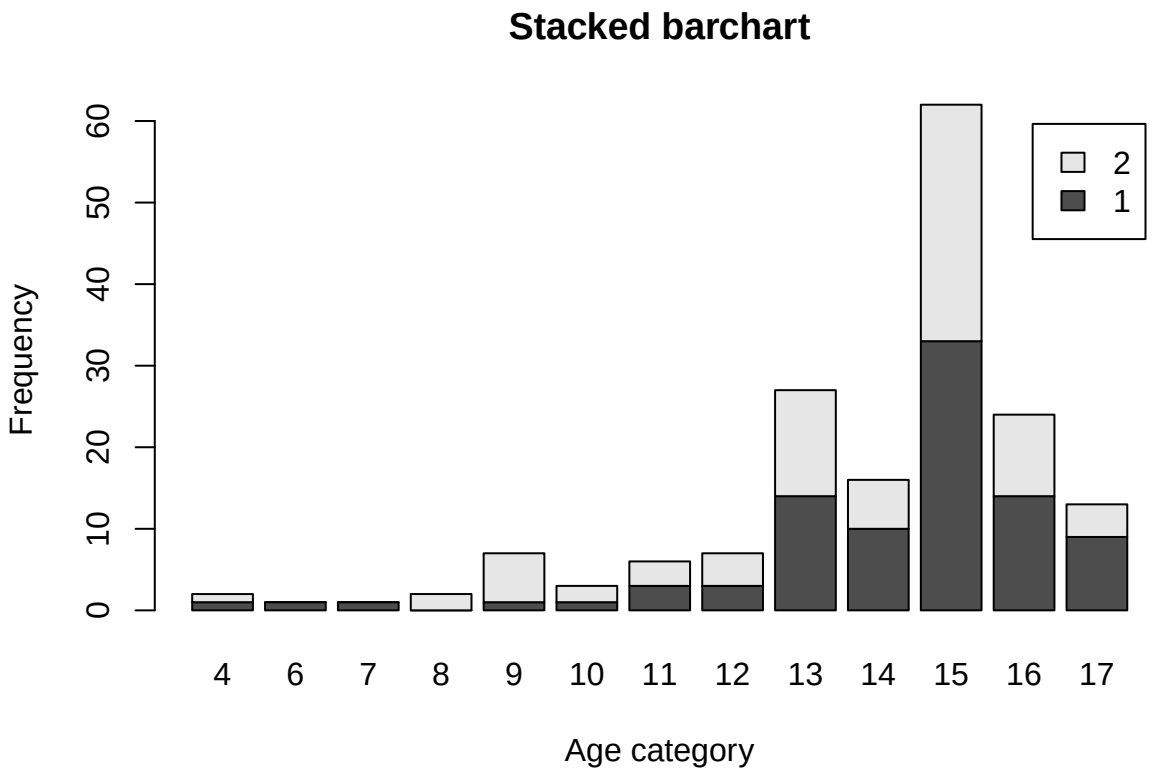
```
table(dat$irsex,dat$age2)
```

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

```
head(dat)
```

```
##   mjage cigage iralcage age2 sexatract speakengl irsex
## 1    14     50       14   16         1         1     1
## 2    11     14        5   13         2         1     2
## 3    12     35       12   15         2         1     2
## 4    16     18       18   14         1         1     1
## 5    14     16       14   16         4         1     1
## 6    12     16       18   15         4         1     2
```

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



Answer: In the youngest age group (4) and group 11, there is an even split. In age groups 6, 7, 13, 14, 15, 16, and 17 there are more men than female, however in age groups 8, 9,10, and 12 there are more females. Men dominate the majority of all the age groups.

Problem 4: Substance use

*correction: table is not needed. just look at lowest value for all the different groups.

why do the age values go past 17 for mjage, cigage, and iralcage. barplots don't actually help to see this, but I don't know how to get a breakdown of the data points within each group.

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.

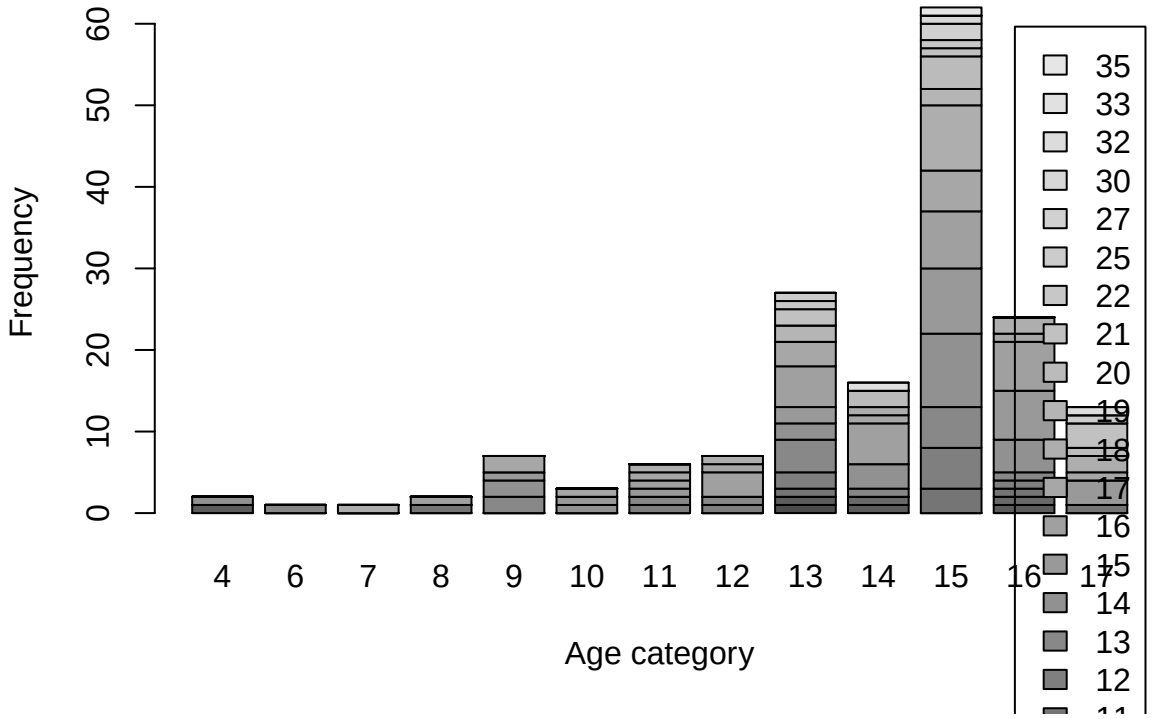
```
table(dat$mjage,dat$age2)
```

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

```
head(dat)
```

```
##   mjage cigage iralcage age2 sexatract speakengl irsex
## 1    14     50       14   16         1         1     1
## 2    11     14        5   13         2         1     2
## 3    12     35       12   15         2         1     2
## 4    16     18       18   14         1         1     1
## 5    14     16       14   16         4         1     1
## 6    12     16       18   15         4         1     2
```

```
tab.agemjage <- table(dat$mjage,dat$age2)
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 8.

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

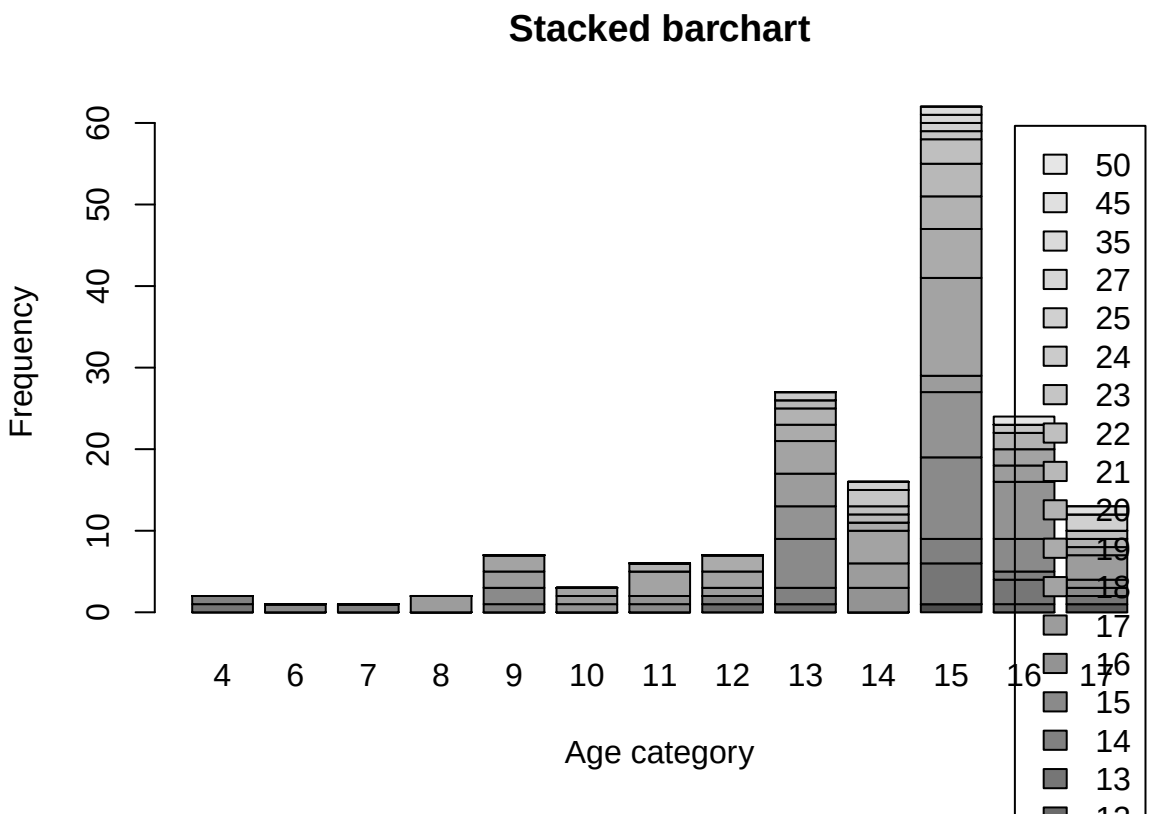
```
##
##       4  6  7  8  9 10 11 12 13 14 15 16 17
##   10  0  0  0  0  0  0  0  0  0  0  1  0  0
##   11  0  0  0  0  0  0  0  0  0  0  0  0  1
##   12  0  0  0  0  0  0  0  1  1  0  0  1  0
##   13  1  0  0  0  0  0  0  0  0  0  5  3  1
##   14  1  0  1  0  1  0  0  1  2  0  3  1  0
##   15  0 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

```
head(dat)
```

```
##   mjage cigage iralcage age2 sexatract speakengl irsex
```

```
## 1    14     50       14   16         1         1     1
## 2    11     14        5   13         2         1     2
## 3    12     35       12   15         2         1     2
## 4    16     18       18   14         1         1     1
## 5    14     16       14   16         4         1     1
## 6    12     16       18   15         4         1     2
```

```
tab.agecigage <- table(dat$cigage,dat$age2)
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 10.

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

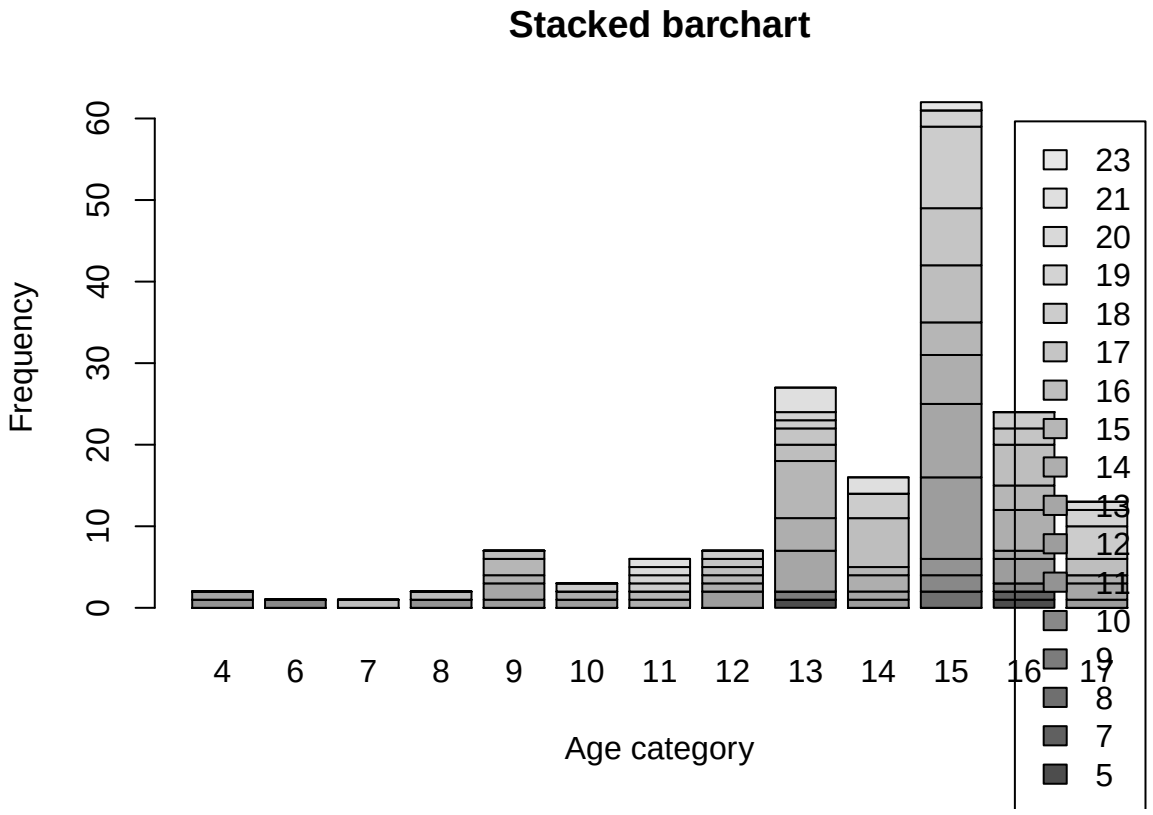
```
##
##       4  6  7  8  9 10 11 12 13 14 15 16 17
##   5   0  0  0  0  0  0  0  0  1  0  0  1  0
##   7   0  0  0  0  0  0  0  0  0  0  0  1  0
##   8   0  0  0  0  0  0  0  0  0  0  2  0  0
##   9   0  0  0  0  0  0  0  0  1  0  0  0  0
##   10  0  1  0  0  0  0  0  0  0  0  2  1  0
##   11  1 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

```
##   16  0  0  1  1  1  0  0  1  2  6  7  5  2
##   17  0  0  0  0  0  0  0  1  2  0  7  2  0
##   18  0  0  0  0  0  1  1  1  1  3 10  2  4
##   19  0  0  0  0  0  0  1  0  1  0  2  0  2
##   20  0  0  0  0  0  0  1  0  0  0  0  0  1
##   21  0  0  0  0  0  0  1  0  3  2  0  0  0
##   23  0  0  0  0  0  0  0  0  0  0  1  0  0
```

```
head(dat)
```

```
##   mjage cigage iralcage age2 sexatract speakengl irsex
## 1    14     50       14   16         1         1     1
## 2    11     14        5   13         2         1     2
## 3    12     35       12   15         2         1     2
## 4    16     18       18   14         1         1     1
## 5    14     16       14   16         4         1     1
## 6    12     16       18   15         4         1     2
```

```
tab.ageiralcage <- table(dat$iralcage,dat$age2)
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 5.

Problem 5: Sexual attraction

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

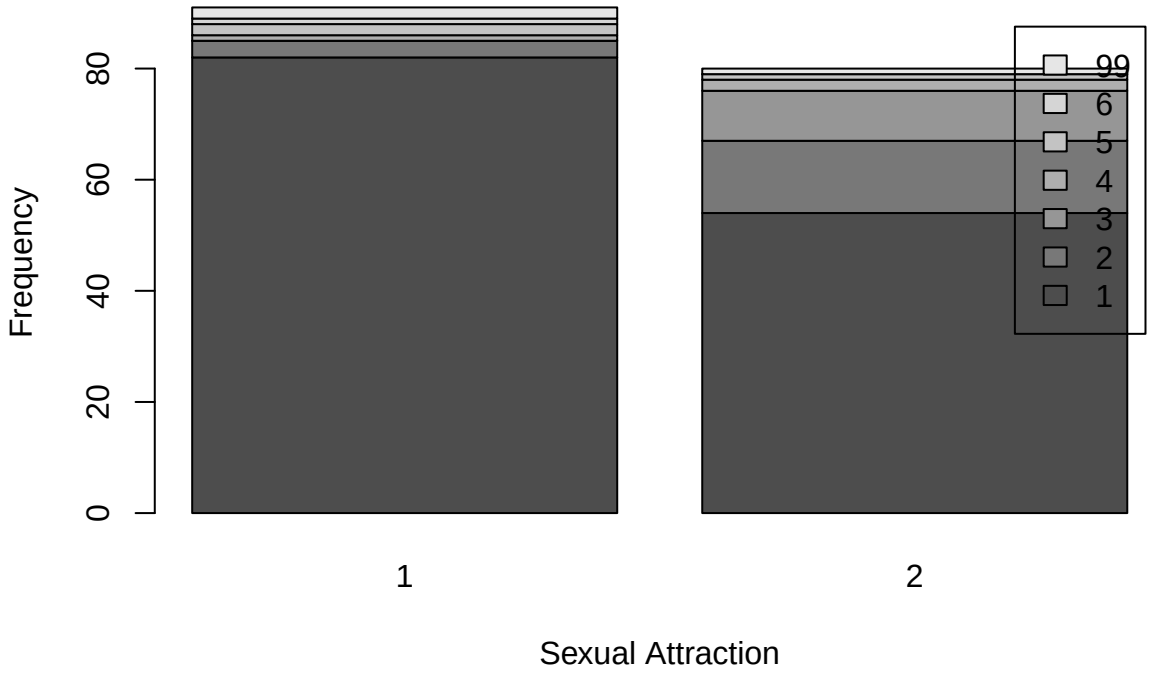
```
table(dat$sexatract,dat$irsex)
```

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

```
head(dat)
```

```
##   mjage cigage iralcage age2 sexatract speakengl irsex
## 1    14     50       14   16         1         1     1
## 2    11     14        5   13         2         1     2
## 3    12     35       12   15         2         1     2
## 4    16     18       18   14         1         1     1
## 5    14     16       14   16         4         1     1
## 6    12     16       18   15         4         1     2
```

```
tab.irsexatract <- table(dat$sexatract,dat$irsex)
barplot(tab.irsexatract,
        main = "Stacked barchart",
        xlab = "Sexual Attraction", ylab = "Frequency",
        legend.text = rownames(tab.irsexatract),
        beside = FALSE) # Stacked bars (default)
```



Answer: "I am only attracted to the oppisote sex" accounts for the majority of both males and females, however it accounts for nearly all men. From there, the numbers are very low. More women state they are as equally attracted to men as they are to other women. The difference is very minimal for those who selected "I am mostly attracted to same sex," "I am only attracted to same sex," I am not sure," and those who skipped 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 varying the data and evening out the distribution of the data.

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?

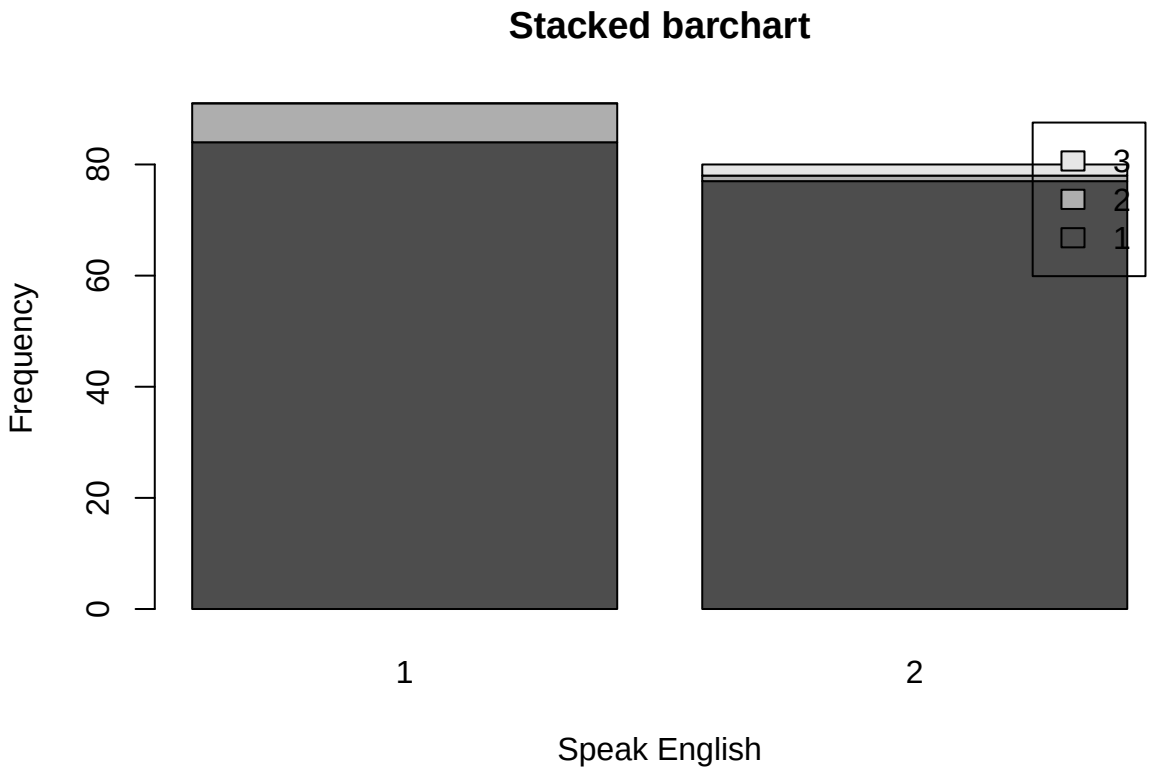
```
table(dat$speakengl,dat$irsex)
```

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

```
head(dat)
```

```
##   mjage cigage iralcage age2 sexatract speakengl irsex
## 1    14     50       14   16         1         1     1
## 2    11     14        5   13         2         1     2
## 3    12     35       12   15         2         1     2
## 4    16     18       18   14         1         1     1
## 5    14     16       14   16         4         1     1
## 6    12     16       18   15         4         1     2
```

```
tab.irsexspeakengl <- table(dat$speakengl,dat$irsex)
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. It accounts for nearly all men and women respectively. A small group of individuals stated they spoke english well (7 men and 1 woman) and 2 women stated they spoke english not well.

Are there more English speaker females or males?

Answer: There are more male English speakers.

#EXAM 1

Instructions

a. Create a folder in your computer (a good place would be under Crim 250, Exams).

b. Download the dataset from the Canvas website (fatal-police-shootings-data.csv) onto that folder, and save your Exam 1.Rmd file in the same folder.

c. Download the README.md file. This is the codebook.

d. Load the data into an R data frame.

```
dat <- read.csv("fatal-police-shootings-data.csv")
head(dat)
```

```
##   id              name       date  manner_of_death      armed age gender race
```

```
## 1  3         Tim Elliot 2015-01-02             shot        gun  53      M    A
## 2  4   Lewis Lee Lembke 2015-01-02             shot        gun  47      M    W
## 3  5 John Paul Quintero 2015-01-03 shot and Tasered    unarmed  23      M    H
## 4  8    Matthew Hoffman 2015-01-04             shot toy weapon  32      M    W
## 5  9  Michael Rodriguez 2015-01-04             shot   nail gun  39      M    H
## 6 11  Kenneth Joe Brown 2015-01-04             shot        gun  18      M    W
##            city state signs_of_mental_illness threat_level        flee
## 1       Shelton    WA                    True       attack Not fleeing
## 2         Aloha    OR                   False       attack Not fleeing
## 3       Wichita    KS                   False        other Not fleeing
## 4 San Francisco    CA                    True       attack Not fleeing
## 5         Evans    CO                   False       attack Not fleeing
## 6       Guthrie    OK                   False       attack Not fleeing
##   body_camera longitude latitude is_geocoding_exact
## 1       False  -123.122   47.247               True
## 2       False  -122.892   45.487               True
## 3       False   -97.281   37.695               True
## 4       False  -122.422   37.763               True
## 5       False  -104.692   40.384               True
## 6       False   -97.423   35.877               True
```

Problem 1 (10 points)

a. Describe the dataset. This is the source: https://github.com/washingtonpost/data-police-shootings . Write two sentences (max.) about this.

The dataset looks at the records of every fatal shooting in the US by a police officer in the line of duty since January 1, 2015. Variables include a unique identifier for each victim, the name of the victim, the date of the fatal shooting in YYYY-MM-DD format, the manner of death (shot or shot and tasered), whether the victim was armed (armed, undetermined, unknown, or unarmed), the age, gender, and race of the victim, the city and state where the shooting took place, if the victim had any signs of mental illness, threat level, whether the victim fled (by foot, by car, or did not flee), whether there was a body camera, the latitude and longitude of the locattion, and finally whether the coordinates were accurate to the location of the shooting. .

b. How many observations are there in the data frame?

By opening the data frame (not using r) I can see that there are 6,694 observations. (excel gives the number of rows, but I subtracted 1 because I didn't count the column titles as observations.)

c. Look at the names of the variables in the data frame. Describe what "body_camera", "flee", and "armed" represent, according to the codebook. Again, only write one sentence (max) per variable.

"body_camera" indicates whether the officer was wearing a body camera (represented in data frame as true or false). "flee" indicates whether the victim fled and if so how they fled (represented in data frame as foot, car, or not fleeing). "armed" indicates whether the victim was armed with some sort of implement that a police officer believed could inflict harm (represented in data frame as the weapon that qualified the victim as armed, undetermined, unknown, or unarmed).

d. What are three weapons that you are surprised to find in the "armed" variable? Make a table of the values in "armed" to see the options.

```
table(dat$armed)
```

```
##                                                  air conditioner
##                              207                                1
##                       air pistol                  Airsoft pistol
##                                1                                3
##                               ax                         barstool
##                               24                                1
##                     baseball bat           baseball bat and bottle
##                               20                                1
## baseball bat and fireplace poker            baseball bat and knife
##                                1                                1
##                            baton                          BB gun
##                                6                               15
##                BB gun and vehicle                     bean-bag gun
##                                1                                1
##                      beer bottle                      binoculars
##                                3                                1
##                     blunt object                          bottle
##                                5                                1
##                    bow and arrow                      box cutter
##                                1                               13
##                            brick              car, knife and mace
##                                2                                1
##                          carjack                           chain
##                                1                                3
##                        chain saw                        chainsaw
##                                2                                1
##                            chair              claimed to be armed
##                                4                                1
##                contractor's level                   cordless drill
##                                1                                1
##                         crossbow                         crowbar
##                                9                                5
##                        fireworks                        flagpole
##                                1                                1
##                       flashlight                     garden tool
##                                2                                2
##                      glass shard                          grenade
##                                4                                1
##                              gun                      gun and car
##                             3798                               12
##                     gun and knife                  gun and machete
##                               22                                3
##                     gun and sword                  gun and vehicle
##                                1                               17
##                guns and explosives                         hammer
##                                3                               18
##                       hand torch                         hatchet
##                                1                               14
##                   hatchet and gun                       ice pick
##                                2                                1
##                  incendiary device                          knife
##                                2                              955
##                  knife and vehicle                 lawn mower blade
```

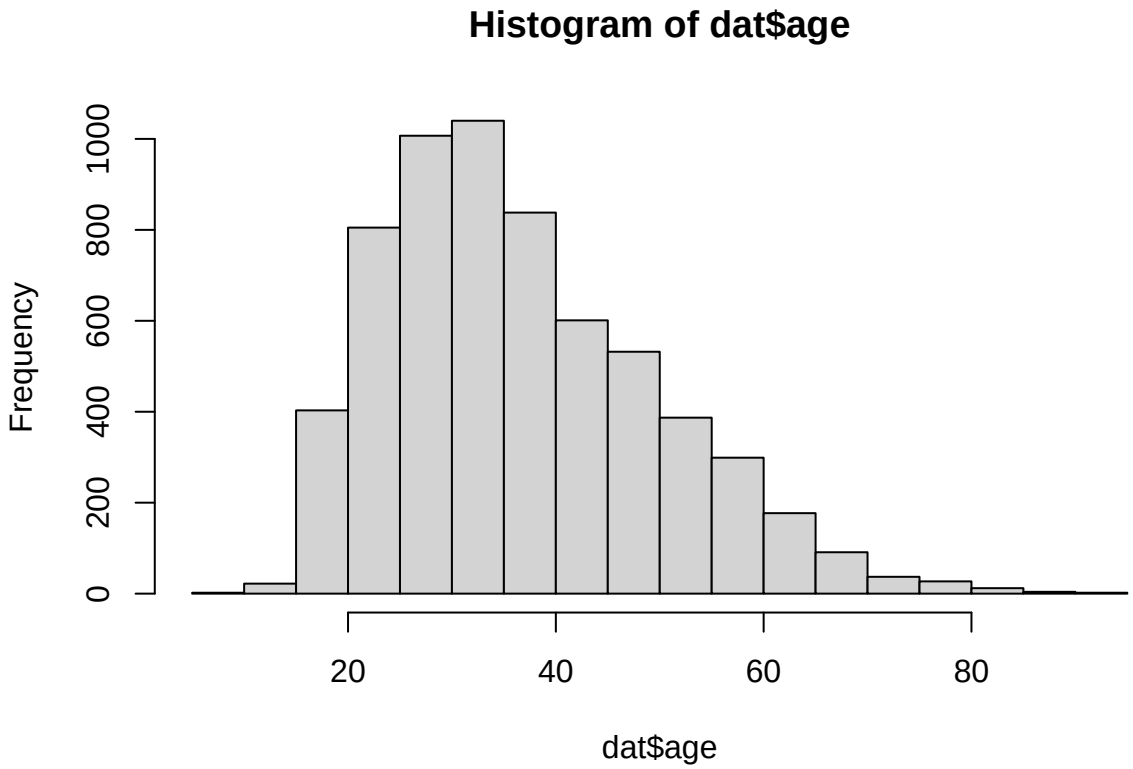
```
##                          1                          2
##                   machete         machete and gun
##                         51                          1
##              meat cleaver         metal hand tool
##                         6                          2
##              metal object              metal pipe
##                         5                         16
##                metal pole              metal rake
##                         4                          1
##               metal stick              microphone
##                         3                          1
##                motorcycle                nail gun
##                         1                          1
##                       oar              pellet gun
##                         1                          3
##                       pen            pepper spray
##                         1                          2
##                  pick-axe           piece of wood
##                         4                          7
##                      pipe               pitchfork
##                         7                          2
##                      pole           pole and knife
##                         3                          2
##            railroad spikes                      rock
##                         1                          7
##             samurai sword                  scissors
##                         4                          9
##                screwdriver            sharp object
##                        16                         14
##                     shovel                      spear
##                         7                          2
##                   stapler      straight edge razor
##                         1                          5
##                     sword                     Taser
##                        23                         34
##                tire iron              toy weapon
##                         4                        226
##                   unarmed           undetermined
##                       421                        188
##            unknown weapon                  vehicle
##                        82                        213
##           vehicle and gun      vehicle and machete
##                         8                          1
##             walking stick              wasp spray
##                         1                          1
##                     wrench
##                         1
```

I am surprised to see that toy weapon, binoculars, and microphone qualify a victim as "armed." None of these items are intended for harm, and even if attempted would probably not do much damage. Of course, the codebook does state that any item that the officer "believes" could do harm to him is included. His belief won't always be valid, as was the case for several of these cases.

Problem 2 (10 points)

a. Describe the age distribution of the sample. Is this what you would expect to see?

```
hist(dat$age)
```



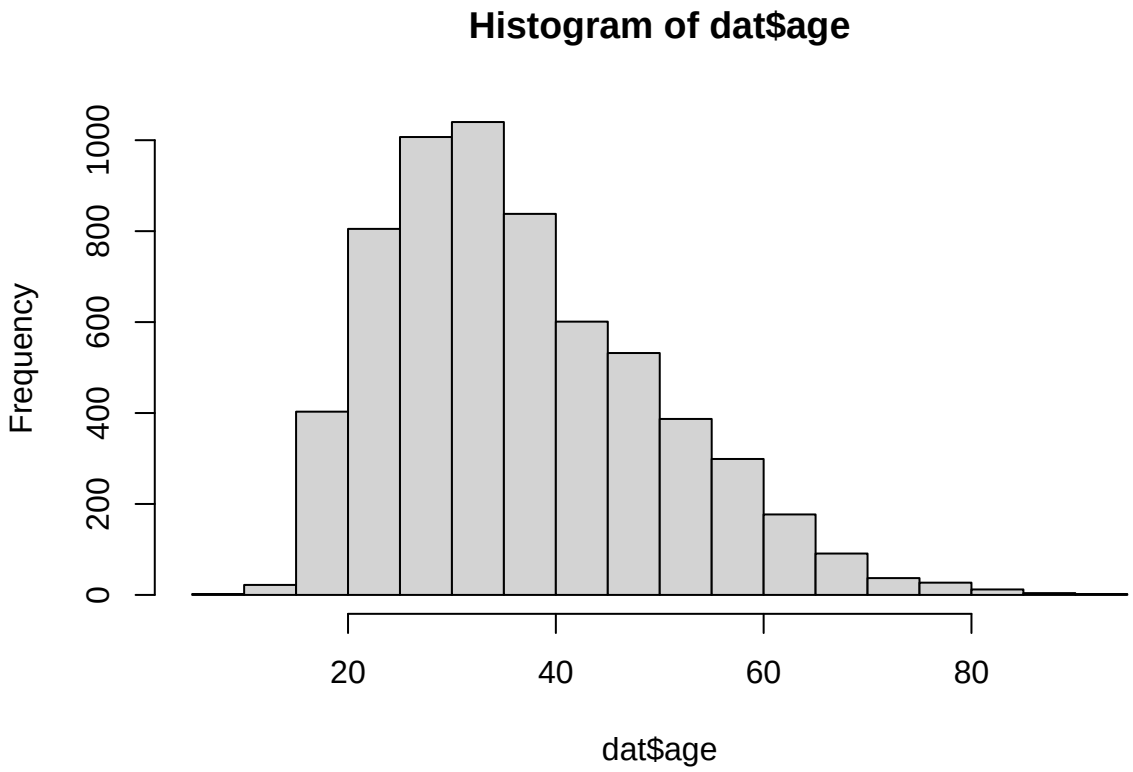

```
summary(dat$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   27.00   35.00   37.12   45.00   91.00     308
```

The data is skewed to the right, meaning that most of the victims were towards the younger side. There are less cases of older victim and the peak age is between between 30 and 40. I expected the spread of data, however the minimum and maximum values were very surprising.

b. To understand the center of the age distribution, would you use a mean or a median, and why? Find the one you picked.

```
hist(dat$age)
```




```
summary(dat$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   27.00   35.00   37.12   45.00   91.00     308
```

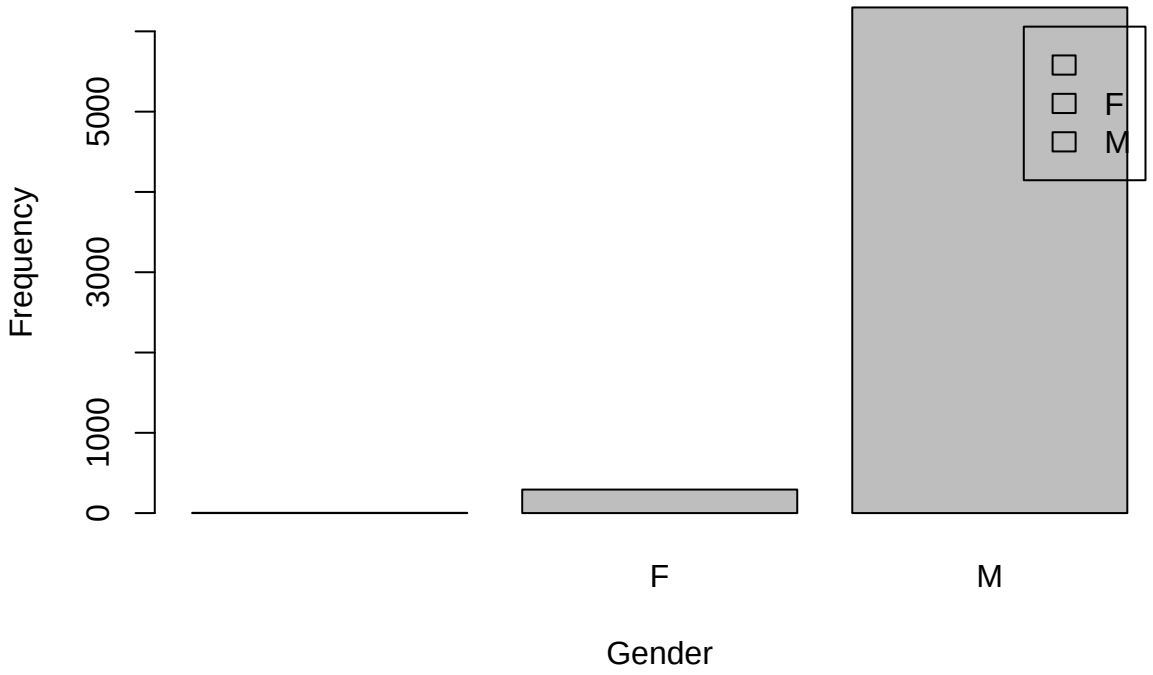
I would use the median because the center of distribution is the median whereas the mean is the average of all the data points. The median is 35.00. There are some missing values, but since they are not assigned numeric values, they don't affect any analysis of my data, such as finding the median.

c. Describe the gender distribution of the sample. Do you find this surprising?

```
table(dat$gender)
```

```
##
##         F    M
##    3  293 6298
```

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



There is about 21 times as much male victims as they are female victims. It is important to note that there are also 3 missing values, however not having these values do not impact the data because of how many more male victims there are in comparison to female victims. I am not surprised. I knew, before analyzing this data set, that men in the US are shot to death by the police more than women.

Problem 3 (10 points)

a. How many police officers had a body camera, according to news reports? What proportion is this of all the incidents in the data? Are you surprised that it is so high or low?

```
table(dat$body_camera)
```

```
##
## False  True
##  5684   910
```

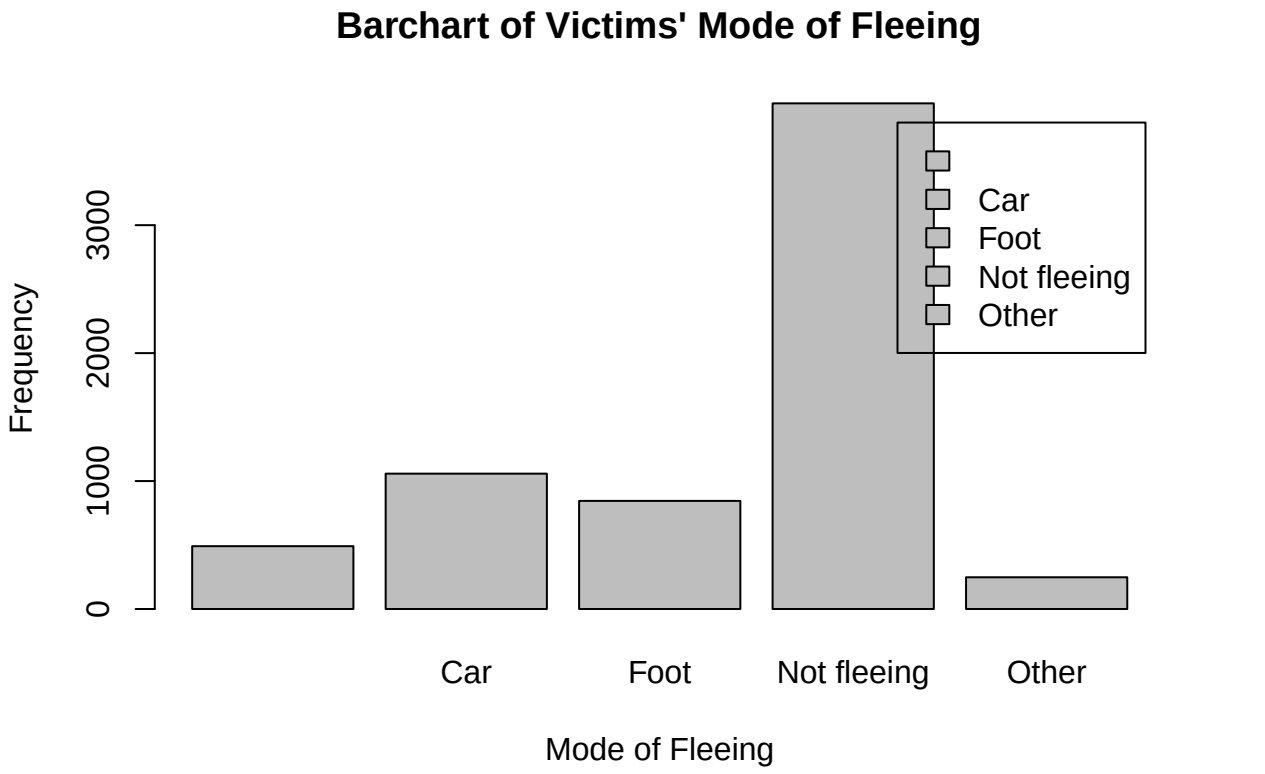
910 police officers had a body camera, according to news reports.This is 13.6 % of all police officers. That is really surprising! It seems that having a body camera would be a measure of precaution for the officer that should be required, unless it is widely acknowledged that police often kill civilians for no just reason and so evidence on the body cameras would be damaging for the officers and that's the reason why they are not required or at least not worn.

b. In how many of the incidents was the victim fleeing? What proportion is this of the total number of incidents in the data? Is this what you would expect?

```
table(dat$flee)
```

```
##
##                  Car         Foot Not fleeing        Other
##         491        1058         845        3952         248
```

```r
tab.flee <- table(dat$flee)
barplot(tab.flee,
        main = "Barchart of Victims' Mode of Fleeing",
        xlab = "Mode of Fleeing", ylab = "Frequency",
        legend.text = rownames(tab.flee),
        beside = FALSE) # Stacked bars (default)
```



The data is not very clear with the different modes of fleeing. There are 248 cases that are categorized as "other" in terms of fleeing. There are also 491 cases that have missing values. Therefore, I will only consider those who fled by car or foot as fleeing and disregard the cases categorized as missing or other. There are 1903 victims that fled and this is 28.9% of all victims. I expected more victims to have fled, but because there is such a great number of values (the other and the missing) that aren't included, I don't really trust this data's records of the number of victims who fled.

Problem 4 (10 points) - Answer only one of these (a or b).

a. Describe the relationship between the variables "body camera" and "flee" using a stacked barplot. What can you conclude from this relationship?

Hint 1: The categories along the x-axis are the options for "flee", each bar contains information about whether the police officer had a body camera (vertically), and the height along the y-axis shows the frequency of that category).

Hint 2: Also, if you are unsure about the syntax for barplot, run ?barplot in R and see some examples at the bottom of the documentation. This is usually a good way to look up the syntax of R code. You can also Google it.

Your answer here.

b. Describe the relationship between age and race by using a boxplot. What can you conclude from this relationship?

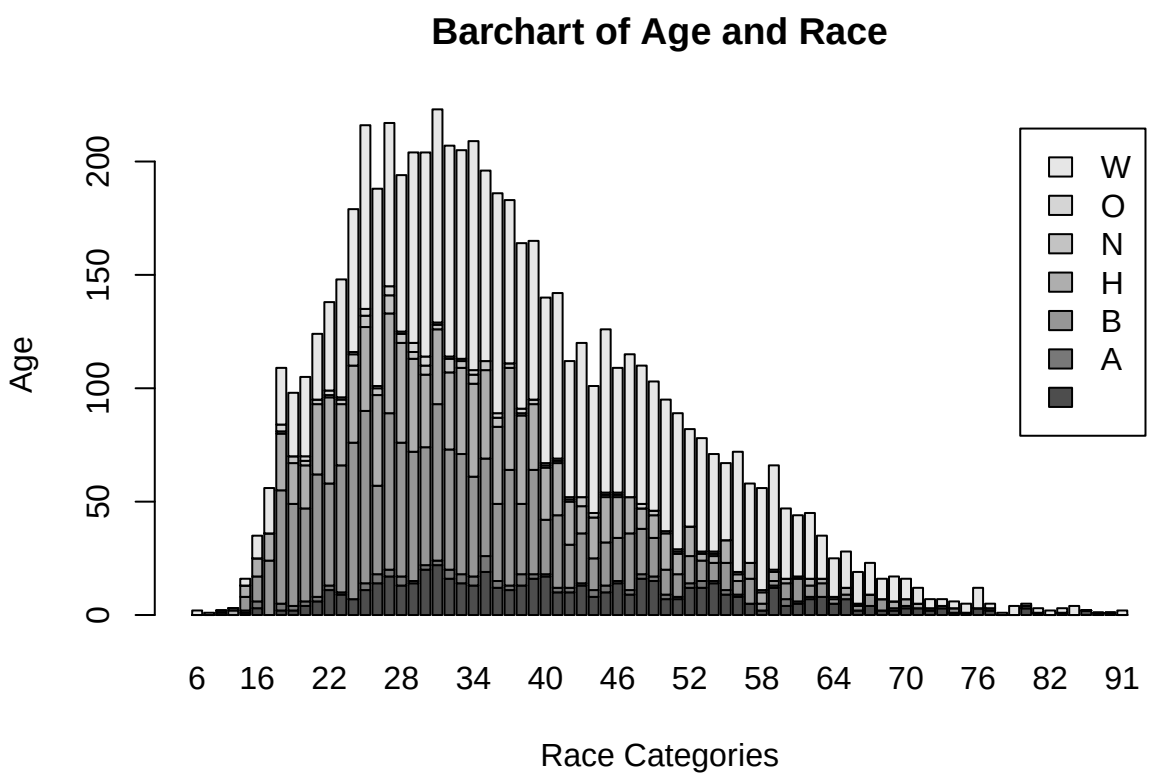
Hint 1: The categories along the x-axis are the race categories and the height along the y-axis is age.

Hint 2: Also, if you are unsure about the syntax for boxplot, run ?boxplot in R and see some examples at the bottom of the documentation. This is usually a good way to look up the syntax of R code. You can also Google it.

```
table(dat$race,dat$age)
```

```
##
##       6  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
##       0   0   0   0   1   3   0   2   2   4   6  11   9   7  11  14  17  13  14
##   A   0   0   0   0   1   3   0   3   2   2   2   2   1   0   3   4   3   4   1
##   B   0   0   1   0   6  11  24  50  45  41  54  45  56  69  76  39  69  59  57
##   H   0   0   1   2   5   8  12  25  18  19  31  38  27  34  37  40  44  44  41
##   N   0   0   0   1   0   0   0   1   3   2   2   1   2   5   5   3   8   4   3
##   O   0   0   0   0   0   0   0   3   0   2   0   2   1   1   3   1   4   1   4
##   W   2   1   0   0   3  10  20  25  28  35  29  39  52  63  81  87  72  69  84
##
##      30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
##      20  22  16  14  13  19  12  11  13  16  17  10  10  13   8  10  14   9  16
##   A   2   2   4   4   4   7   3   2   5   2   1   2   2   1   3   3   1   2   2
##   B  52  69  53  53  44  43  34  51  31  46  24  32  19  22  14  19  19  25  20
##   H  32  33  34  38  41  39  34  45  39  29  23  23  19  12  18  20  18  16   9
##   N   4   2   6   3   4   4   4   2   1   2   1   1   1   4   2   1   1   0   0
##   O   4   1   1   1   2   0   2   0   2   0   1   1   1   0   0   1   1   0   2
##   W  90  94  93  92 101  84  97  72  73  70  73  73  60  68  56  72  55  63  61
##
##      49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67
##      15   7   7  12  12  14   9   8   5   2  12   4   5   7   8   5   7   2   4
##   A   2   2   1   2   3   1   2   1   0   0   1   3   1   1   0   0   0   0   0
##   B  17  11  10  12   9   8  12   6  11   3   2   7  10   5   6   2   2   2   5
##   H  10  16   9  13   3   3  10   3   7   5   4   2   1   3   2   1   3   1   0
##   N   2   1   1   0   1   1   0   0   0   1   0   0   0   0   0   0   0   0   0
##   O   0   0   1   0   0   1   0   1   0   0   1   0   0   0   0   0   0   0   0
##   W  57  58  60  43  50  43  34  53  35  45  46  31  27  29  19  17  16  14  14
##
##      68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  86  88
##       2   2   3   3   2   3   1   1   3   2   0   0   3   1   0   1   0   2   0
##   A   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   B   5   1   1   0   1   0   2   0   0   1   0   0   0   0   0   0   0   0   1
##   H   0   3   3   2   0   1   0   0   0   0   0   0   1   0   0   0   0   0   0
##   N   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   O   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   W   9  11   9   7   4   3   3   4   9   2   1   4   1   2   2   2   4   0   0
##
##      89  91
##       1   0
##   A   0   0
##   B   0   0
##   H   0   0
##   N   0   0
##   O   0   0
##   W   0   2
```

```
tab.raceage <- table(dat$race,dat$age)
barplot(tab.raceage,
        main = "Barchart of Age and Race",
        xlab = "Race Categories", ylab = "Age",
        legend.text = rownames(tab.raceage),
        beside = FALSE) # Stacked bars (default)
```



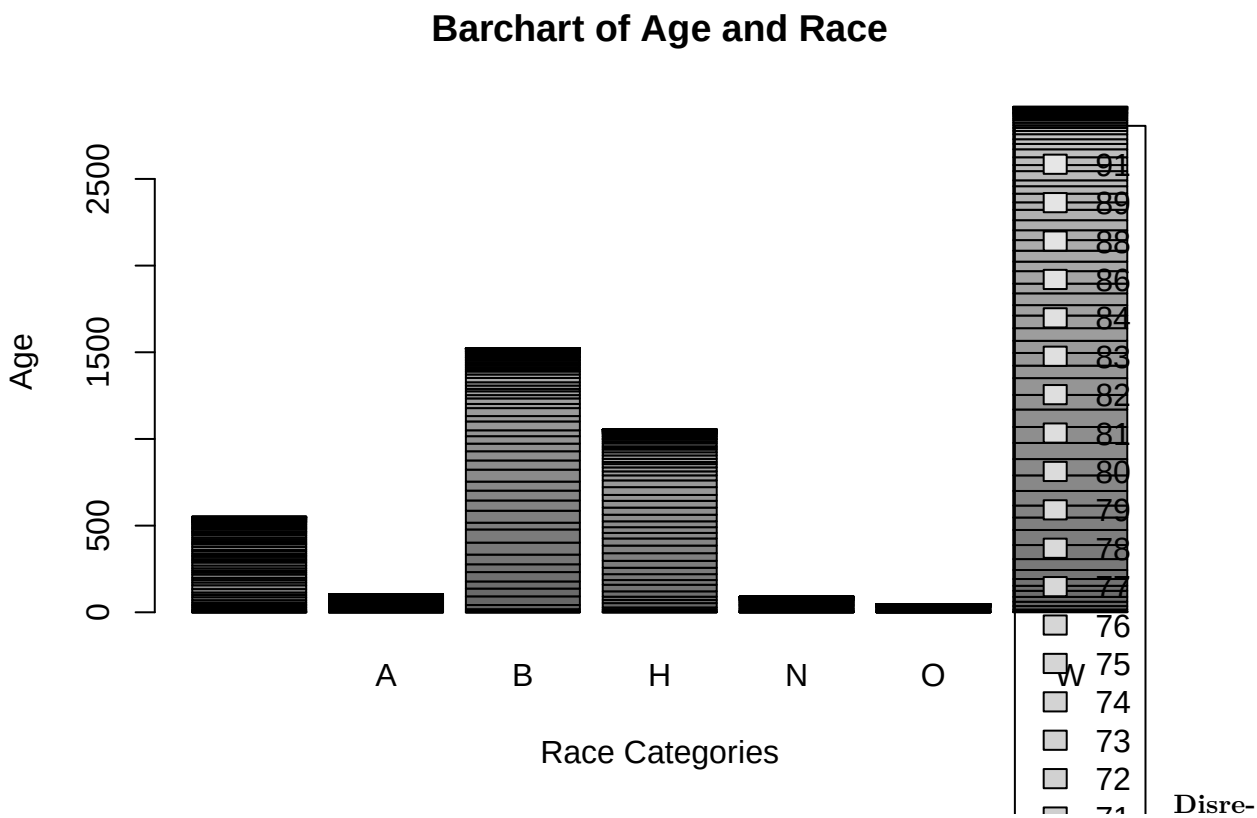

```
table(dat$age,dat$race)
```

```
##
##          A   B   H   N   O   W
##   6    0   0   0   0   0   0   2
##   12   0   0   0   0   0   0   1
##   13   0   0   1   1   0   0   0
##   14   0   0   0   2   1   0   0
##   15   1   1   6   5   0   0   3
##   16   3   3  11   8   0   0  10
##   17   0   0  24  12   0   0  20
##   18   2   3  50  25   1   3  25
##   19   2   2  45  18   3   0  28
##   20   4   2  41  19   2   2  35
##   21   6   2  54  31   2   0  29
##   22  11   2  45  38   1   2  39
##   23   9   1  56  27   2   1  52
##   24   7   0  69  34   5   1  63
##   25  11   3  76  37   5   3  81
##   26  14   4  39  40   3   1  87
##   27  17   3  69  44   8   4  72
##   28  13   4  59  44   4   1  69
```

```
##   29  14   1  57  41   3   4  84
##   30  20   2  52  32   4   4  90
##   31  22   2  69  33   2   1  94
##   32  16   4  53  34   6   1  93
##   33  14   4  53  38   3   1  92
##   34  13   4  44  41   4   2 101
##   35  19   7  43  39   4   0  84
##   36  12   3  34  34   4   2  97
##   37  11   2  51  45   2   0  72
##   38  13   5  31  39   1   2  73
##   39  16   2  46  29   2   0  70
##   40  17   1  24  23   1   1  73
##   41  10   2  32  23   1   1  73
##   42  10   2  19  19   1   1  60
##   43  13   1  22  12   4   0  68
##   44   8   3  14  18   2   0  56
##   45  10   3  19  20   1   1  72
##   46  14   1  19  18   1   1  55
##   47   9   2  25  16   0   0  63
##   48  16   2  20   9   0   2  61
##   49  15   2  17  10   2   0  57
##   50   7   2  11  16   1   0  58
##   51   7   1  10   9   1   1  60
##   52  12   2  12  13   0   0  43
##   53  12   3   9   3   1   0  50
##   54  14   1   8   3   1   1  43
##   55   9   2  12  10   0   0  34
##   56   8   1   6   3   0   1  53
##   57   5   0  11   7   0   0  35
##   58   2   0   3   5   1   0  45
##   59  12   1   2   4   0   1  46
##   60   4   3   7   2   0   0  31
##   61   5   1  10   1   0   0  27
##   62   7   1   5   3   0   0  29
##   63   8   0   6   2   0   0  19
##   64   5   0   2   1   0   0  17
##   65   7   0   2   3   0   0  16
##   66   2   0   2   1   0   0  14
##   67   4   0   5   0   0   0  14
##   68   2   0   5   0   0   0   9
##   69   2   0   1   3   0   0  11
##   70   3   0   1   3   0   0   9
##   71   3   0   0   2   0   0   7
##   72   2   0   1   0   0   0   4
##   73   3   0   0   1   0   0   3
##   74   1   0   2   0   0   0   3
##   75   1   0   0   0   0   0   4
##   76   3   0   0   0   0   0   9
##   77   2   0   1   0   0   0   2
##   78   0   0   0   0   0   0   1
##   79   0   0   0   0   0   0   4
##   80   3   0   0   1   0   0   1
##   81   1   0   0   0   0   0   2
##   82   0   0   0   0   0   0   2
```

```
##    83   1   0   0   0   0   0   2
##    84   0   0   0   0   0   0   4
##    86   2   0   0   0   0   0   0
##    88   0   0   1   0   0   0   0
##    89   1   0   0   0   0   0   0
##    91   0   0   0   0   0   0   2
```

```
tab.raceage <- table(dat$age,dat$race)
barplot(tab.raceage,
        main = "Barchart of Age and Race",
        xlab = "Race Categories", ylab = "Age",
        legend.text = rownames(tab.raceage),
        beside = FALSE) # Stacked bars (default)
```



Disregard first histogram. THere is a great number of older black and hispanic victims. There is also a great deal of victims whose races are not clarified, which is problematic in making a conclusion of the types of victims that are killed at the hands of police. White males however represent the group with the oldest victims killed by police.

Extra credit (10 points)

a. What does this code tell us?

```
mydates <- as.Date(dat$date)
head(mydates)
(mydates[length(mydates)] - mydates[1])
```

b. On Friday, a new report was published that was described as follows by The Guardian: "More than half of US police killings are mislabelled or not reported, study finds." Without reading this article now (due to limited time), why do you think police killings might be mislabelled or underreported?

c. Regarding missing values in problem 4, do you see any? If so, do you think that's all that's missing from the data?

This code arranges my data into the dates during which the cases occured.The last line of the code tells us the time difference. They might be mislabelled or underreported because so much of the data in this data set are either missing or unclear so that you can't make clear conclusions of the rate of police killings. There are missing data points in problem 4, however I don't think that's all that is missing, if the rest of the data frame is any indication.