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: Natalie Yang.

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.

```
library(datasets)
```

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
```

It's useful to rename the dataset so you know can keep track of which dataset is which - it can stick in your mind more if you rename it. Also, when you're altering data, it's best to keep track of the different versions are which (ie which datasets are untouched, etc). Lastly, I wouldn't want to alter a Base R package dataset, and then save it to my workspace, and have that alter things in the future.

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`.

```
colnames(dat)
```

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

The output is: [1] "Murder" "Assault" "UrbanPop" "Rape" "state"

Problem 3

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

Answer: A quantitative variable.

What R Type of variable is it?

Answer: Numeric (class(dat$Murder) returns numeric)

Problem 4

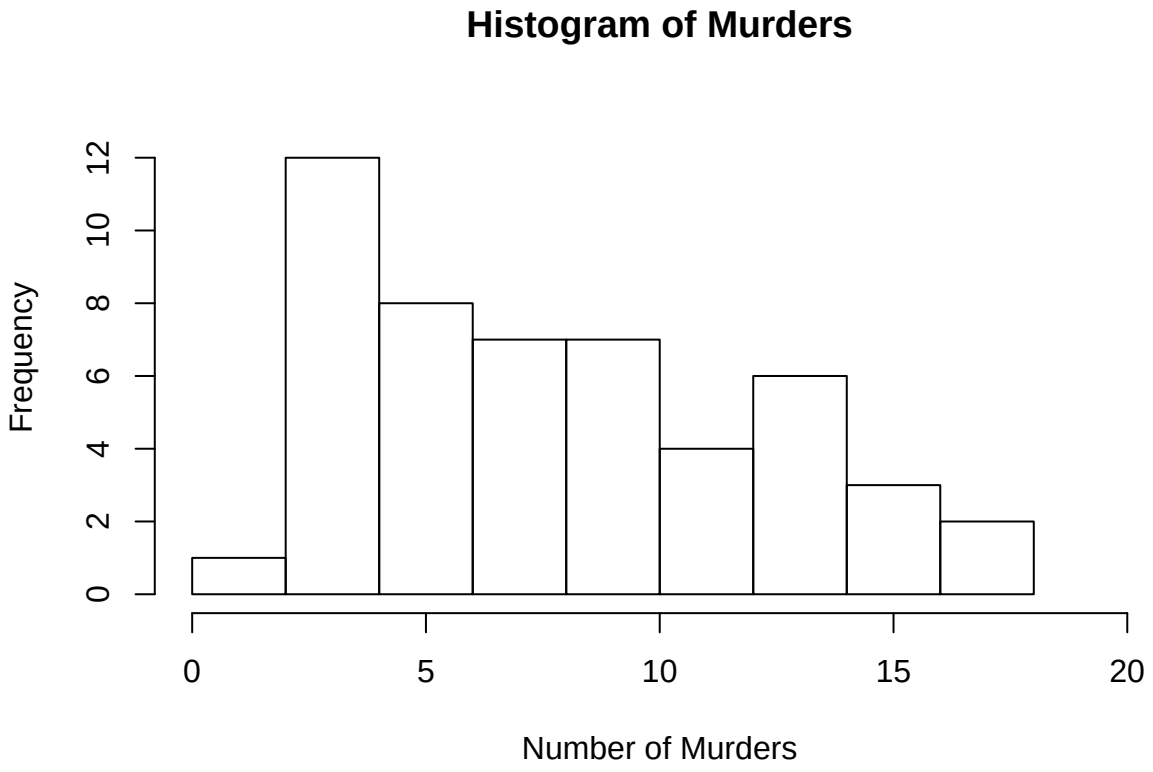
What information is contained in this dataset, in general? What do the numbers mean?

Answer: This dataset contains information on Murder, Assault, and Rape per 100,000 residents, by state for all 50 states in the United States.

Problem 5

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

```
hist(dat$Murder,
     main = "Histogram of Murders",
     xlim = c(0, 20),
     ylim = c(0, 13),
     xlab = "Number of Murders",
     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.?

```
mean(dat$Murder)
```

```
## [1] 7.788
```

```
median(dat$Murder)
```

```
## [1] 7.25
```

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

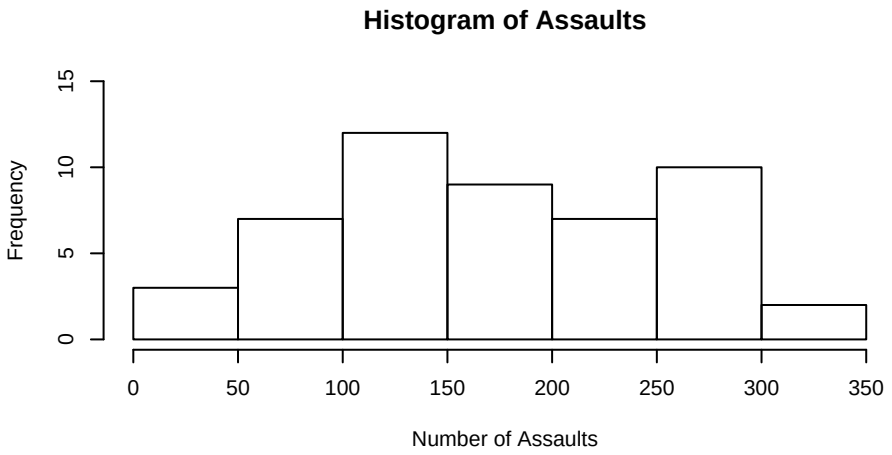
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

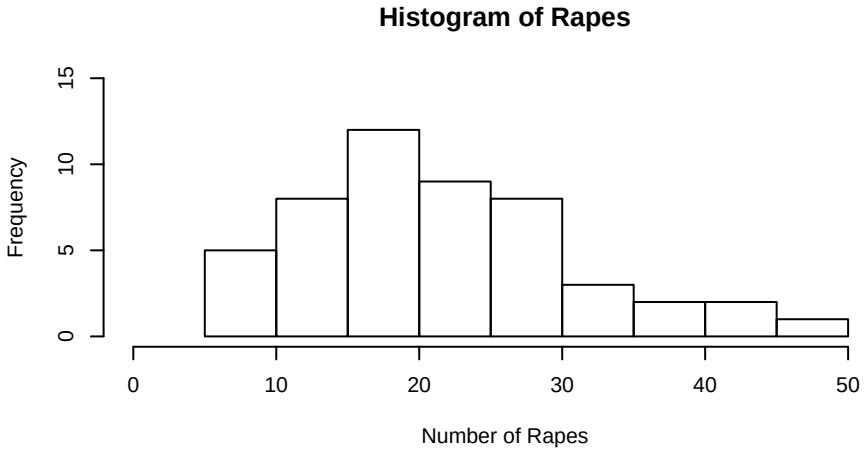
Mean is the average amount of Murders in all 50 states, while median is the middle state Murder level in the dataset. The 1st quartile shows us that the lowest 25% state Murder rates are under 4.075. The 3rd quartile shows us that for the 50.1-75% of states, which is the block right above the median. We don't need the 2nd quartile because we have the Median to know that the 25.1-50% of states fall between that number and the 1st quartile. We are given the max as well, which is technically the 4th quartile, and that alongside the 3rd quartile amount gives us all the information we need about distribution of Murder rate in the dataset. The mean is also greater than the median, so the dataset is positively skewed.

Problem 7

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

```r
par(mfrow=c(3, 1))

hist(dat$Assault,
     main = "Histogram of Assaults",
     xlim = c(0, 355),
     ylim = c(0, 15),
     xlab = "Number of Assaults",
     ylab = "Frequency")

hist(dat$Murder,
     main = "Histogram of Murders",
     xlim = c(0, 20),
     ylim = c(0, 13),
     xlab = "Number of Murders",
     ylab = "Frequency")

hist(dat$Rape,
     main = "Histogram of Rapes",
     xlim = c(0, 52),
     ylim = c(0, 15),
     xlab = "Number of Rapes",
     ylab = "Frequency")
```





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

Answer: Par splits up our screen so we can see more than one figure in one viewing. In the example above, we used par to split the screen into three horizontal chunks. By changing the second number in the par function, we could also bisect the screen vertically.

What can you learn from plotting the histograms together?

Answer: You can compare frequencies, and you can also see how some crimes (such as Assaults) are committed

more than others (e.g. Murders).

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:

```
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: This code shows the murder rates graphically. With the entire contiguous US being shown on a map, the code shades states on a light-dark scale based on how high their Murder rate (per 100,000 residents) are. States with lower murder rates are shaded darker, and vice versa for higher ones. We can see states in the PNW and NE regions have lower Murder rates than states in the SE visually with this map.

Assignment 2

Collaborators: Natalie Yang.

Instructions: Copy your code, paste it into a Word document, and turn it into Canvas. You can turn in a .docx or .pdf file. Show any EDA (graphical or non-graphical) you have used to come to this conclusion.

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

Read the data

```
library(readr)
dat <- read_csv("dat.nsduh.small.1.csv")
```

```
## Parsed with column specification:
## cols(
##   mjage = col_double(),
##   cigage = col_double(),
##   iralcage = col_double(),
```

```
##   age2 = col_double(),
##   sexatract = col_double(),
##   speakengl = col_double(),
##   irsex = col_double()
## )
```

```
View(dat)
```

```
## Warning in system2("/usr/bin/otool", c("-L", shQuote(DSO)), stdout = TRUE):
## running command ''/usr/bin/otool' -L '/Library/Frameworks/R.framework/
## Resources/modules/R_de.so'' had status 1
```

What are the dimensions of the dataset?

```
dim(dat)
```

```
## [1] 171   7
```

The dimensions are 171 rows by 7 columns

Problem 2: Variables

Describe the variables in the dataset.

The variables are all numeric. They represent different number ranges, which account for different meanings.For example, the age2 variable represents the respondents age. In some buckets, only the exact age is contain (ie 1 represents the age 12), but in others it represents a range (eg 13 represents ages 26-29).

What is this dataset about? Who collected the data, what kind of sample is it, and what was the purpose of generating the data?

This data set is from the National Survey of Drug Use in 2019. It looks into when respondents first started using drugs/alcohol/nicotine, even if it's before the legal age. The NSDUH collected the data, and it seems to be a random sample from a large range of ages. This data is meant to reflect drug/alcohol/etc use across the whole country.

```
names(dat)
```

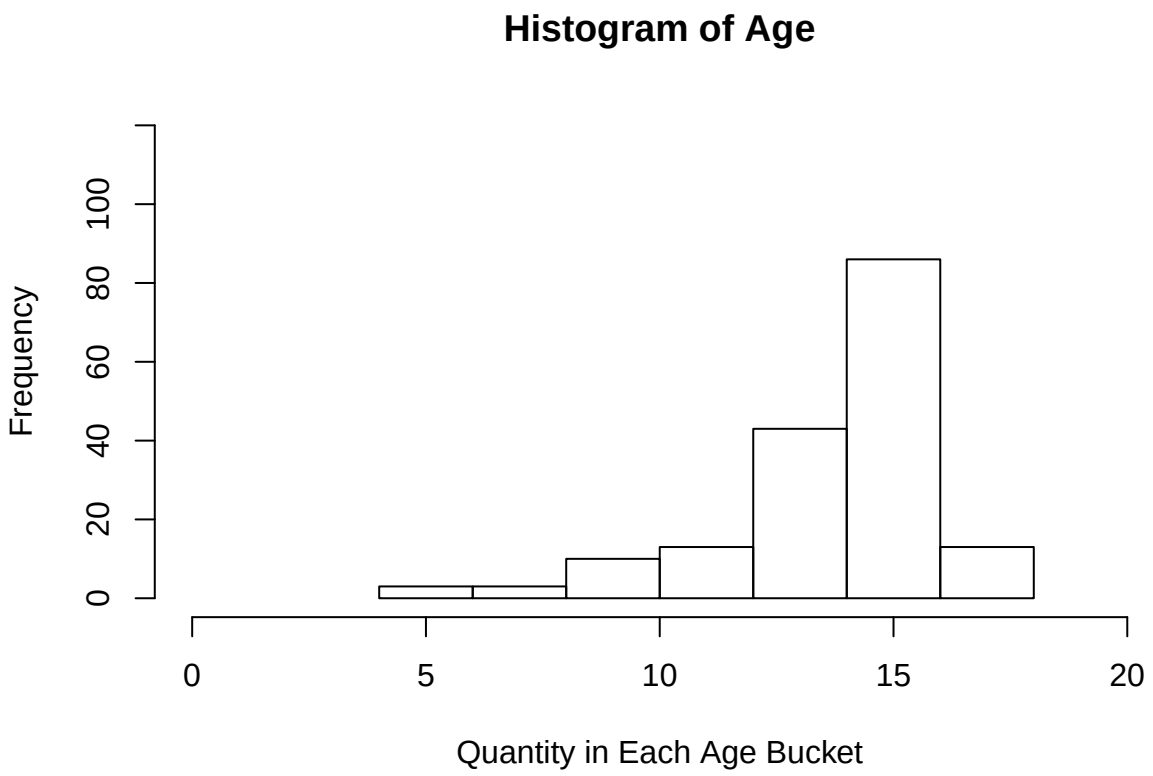
```
## [1] "mjage"     "cigage"    "iralcage"  "age2"      "sexatract" "speakengl"
## [7] "irsex"
```

The variables are: mjage, cigage, iralcage, age2, sexatract, speakengl, and irsex

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.

```
hist(dat$age2,
     main = "Histogram of Age",
     xlim = c(0, 20),
     ylim = c(0, 120),
     xlab = "Quantity in Each Age Bucket",
     ylab = "Frequency")
```



The age frequency obciously isn't fully representative of the US population, because it doesn't have respondents below the age of 12.

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

```
sum(dat$irsex[dat$irsex == 2])
```

```
## [1] 160
```

```
sum(dat$irsex)
```

```
## [1] 251
```

160 females, 251 respondents in total

This sample isn't balanced in terms of gender because there are 69 more females than males. This is not representative of the US population, where gender is mostly balanced.

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?

```
dat$iralcage[dat$iralcage == 5]
```

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

```
dat$cigage[dat$cigage == 10]
```

```
## [1] 10
```

```
dat$mjage[dat$mjage == 7]
```

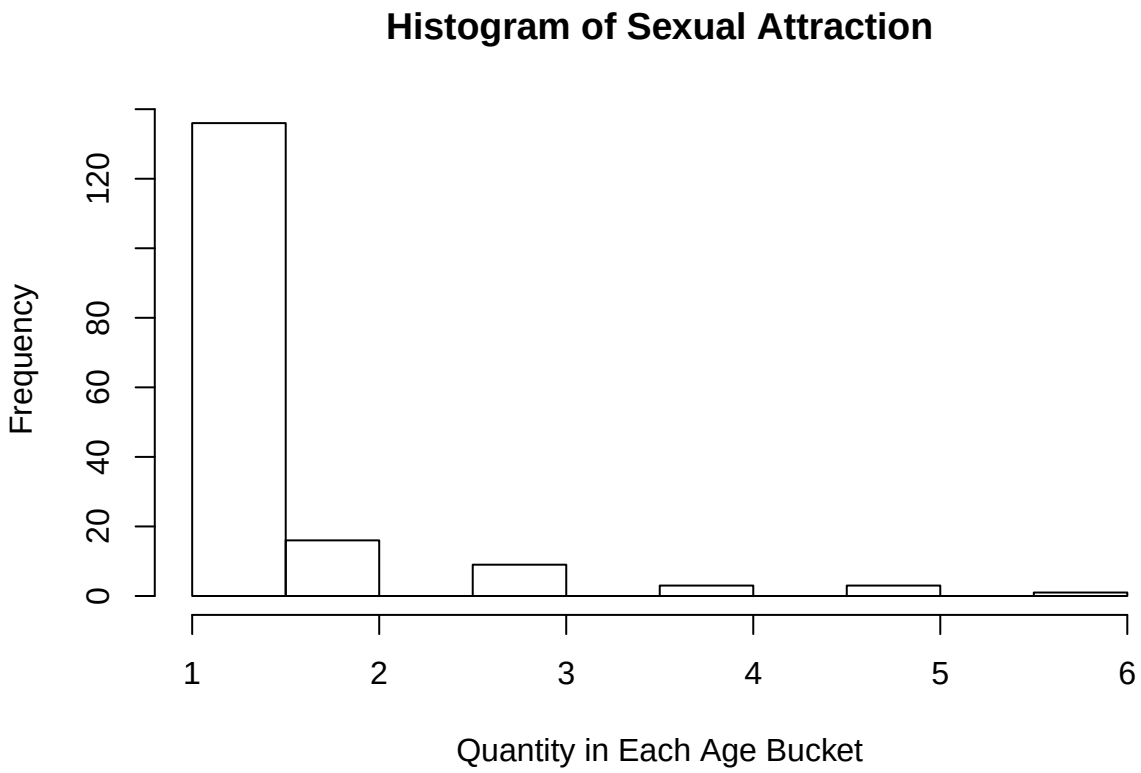
```
## [1] 7
```

There is only one user in the earliest bucket for marijuana and cigarette use. There are two for alcohol use. I am deducing that individuals tend to use alcohol earlier.

Problem 5: Sexual attraction

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

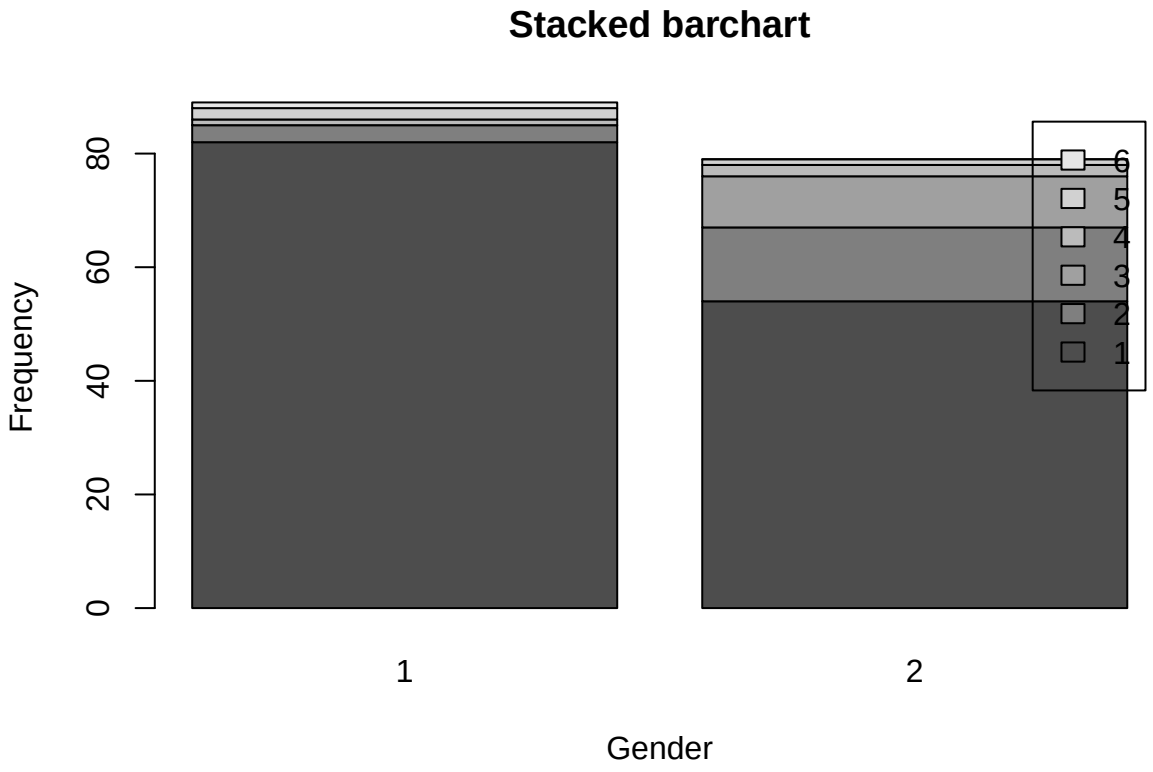
```
dat$sex.attract <- dat$sexatract
dat$sex.attract[dat$sex.attract == 85 | dat$sex.attract == 94 | dat$sex.attract == 97 |
                  dat$sex.attract == 98 | dat$sex.attract == 99 ] <- NA

hist(dat$sex.attract,
     main = "Histogram of Sexual Attraction",
     xlab = "Quantity in Each Age Bucket",
     ylab = "Frequency")
```



Most people identify as heterosexual, which is what I expected.

What is the distribution of sexual attraction by gender?

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

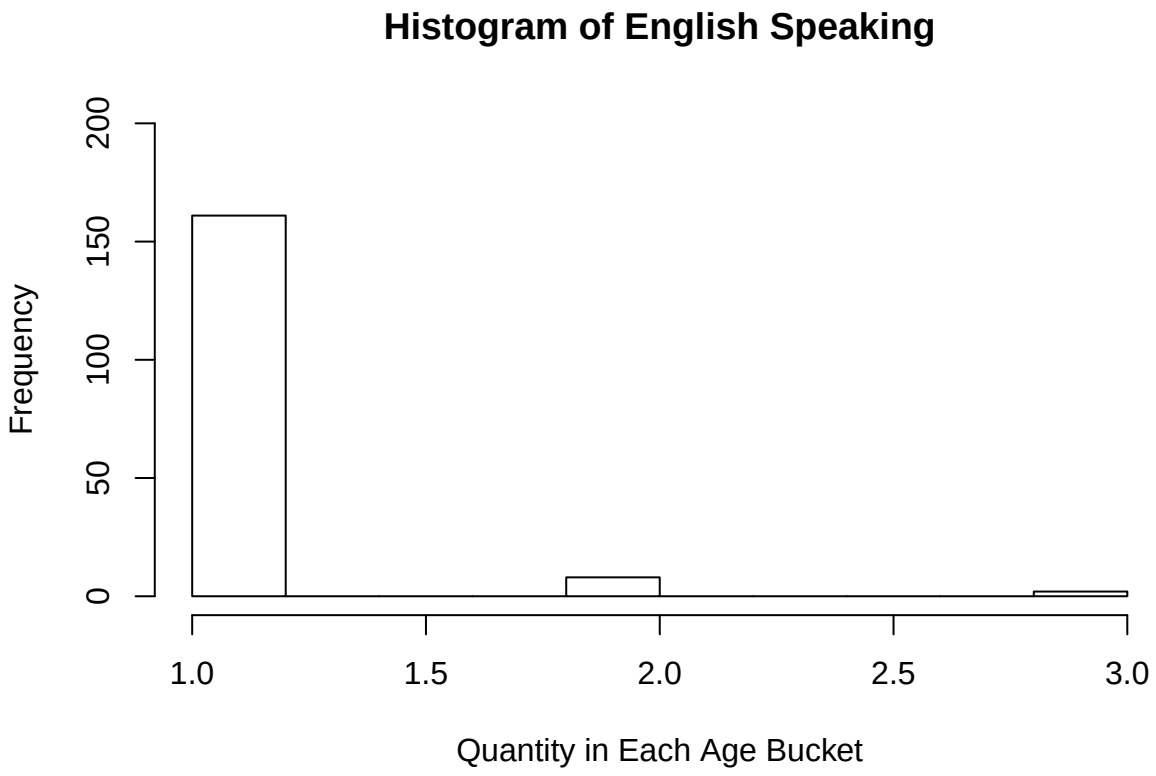


More males identify as straight than females.

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?

```
hist(dat$speakengl,
     main = "Histogram of English Speaking",
     xlim = c(1,3),
     ylim = c(0, 200),
     xlab = "Quantity in Each Age Bucket",
     ylab = "Frequency")
```



Mostly everyone is an English speaker, and speaks very well/well, which is what I expected, as this is a national sample coming from the United States, and the official language is English. It would be interesting to look at how many people are native English speakers in the future.

Are there more English speaker females or males?

(in this, I'm using values 1-3 to count as speaking English)

```
sum(dat$irsex[dat$irsex == 1 & (dat$speakengl == 1 | dat$speakengl == 2 | dat$speakengl == 3)])
```

```
## [1] 91
```

```
sum(dat$irsex[dat$irsex == 2 & (dat$speakengl == 1 | dat$speakengl == 2 | dat$speakengl == 3)])
```

```
## [1] 160
```

91 English speakers are male, and 160 are females. All respondents fall into a English-speaking bucket, and so with the unbalanced gender we saw earlier, there are more female English speakers than male.