Title: Assignment 2

Subtitle: Crim 250: Statistics for the Social Sciences

Name: Marcus Ramirez

Date: 09/23/2021

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.

Problem 1: Load data

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

setwd("/Users/cruzllano/Documents/R/") dat2 <- read_csv("~/Downloads/dat.nsduh.small.csv") # Read the data dat2 <- read_csv("~/Downloads/dat.nsduh.small.csv")

What are the dimensions of the dataset?

dim(dat2) Answer: The dimensions of the dataset are 171 by 7. names(dat)

Problem 2: Variables

Describe the variables in the dataset.

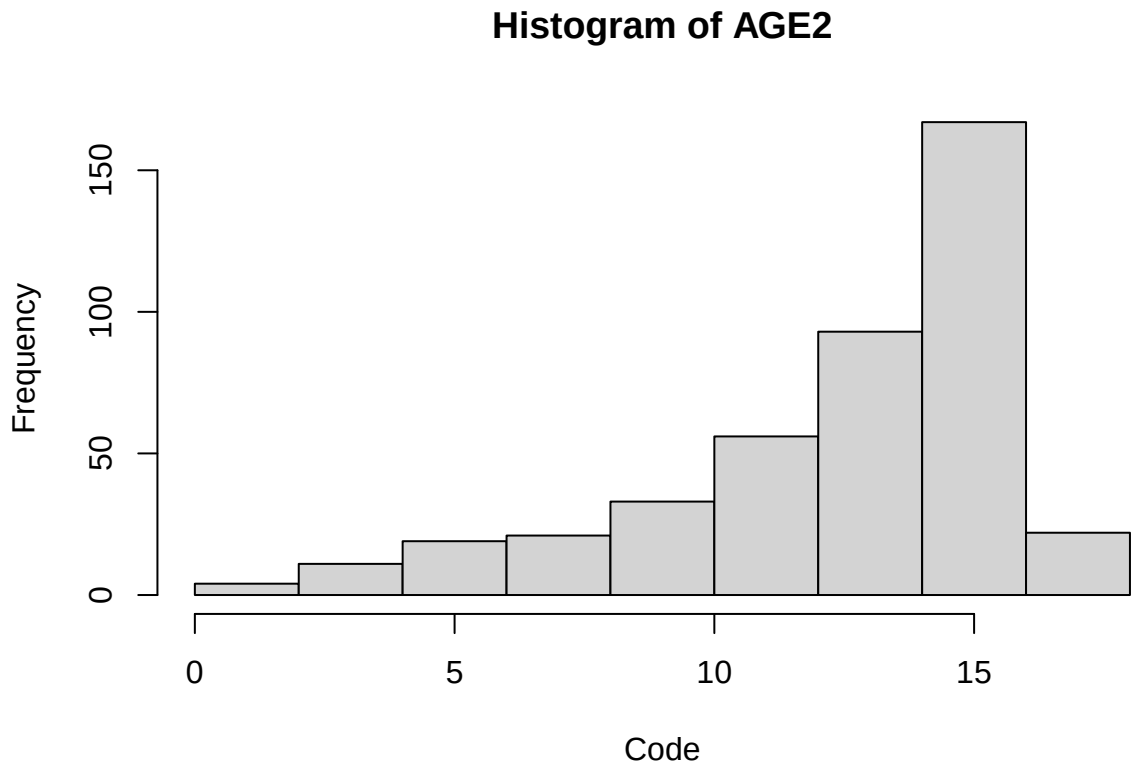
The varibles in this dataset are mjage, cigever, alcever, AGE2, sexatract, speakengl, irsex. Mjage describes how old someone was when they first used marijuana or hashish. Cigever describes how old someone was when they first started smoking cigarettes everyday. Alcever describes how old someone was when they first tried alcochol. AGE2 describes the final edited age of the respondant. Sexatract describes a respondant's sexual orientation. Speakengl describes how well someone speaks English. Irsex describes someone's gender. # 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 dataset is about the age at which individuals of varying ages, sexual orientations, genders, and English proficiencies first began experimenting or using particular drugs (including but not limited to marijuana, cigarettes, and alcohol). The data was collected by the National Survey on Drug Use and Health. This is a simple random sample as the first 1000 cases were chooseCRANmirror(The purpose of this generating this data was to form more general conclusions about the population from the sample, pertaining to drug use. This way, the NSDUH can better predict where to provide support prevention and monitor substance use trends.

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.

summary(dat1)

```
hist(dat1$AGE2, main="Histogram of AGE2", xlab="Code", ylab="Frequency")
```



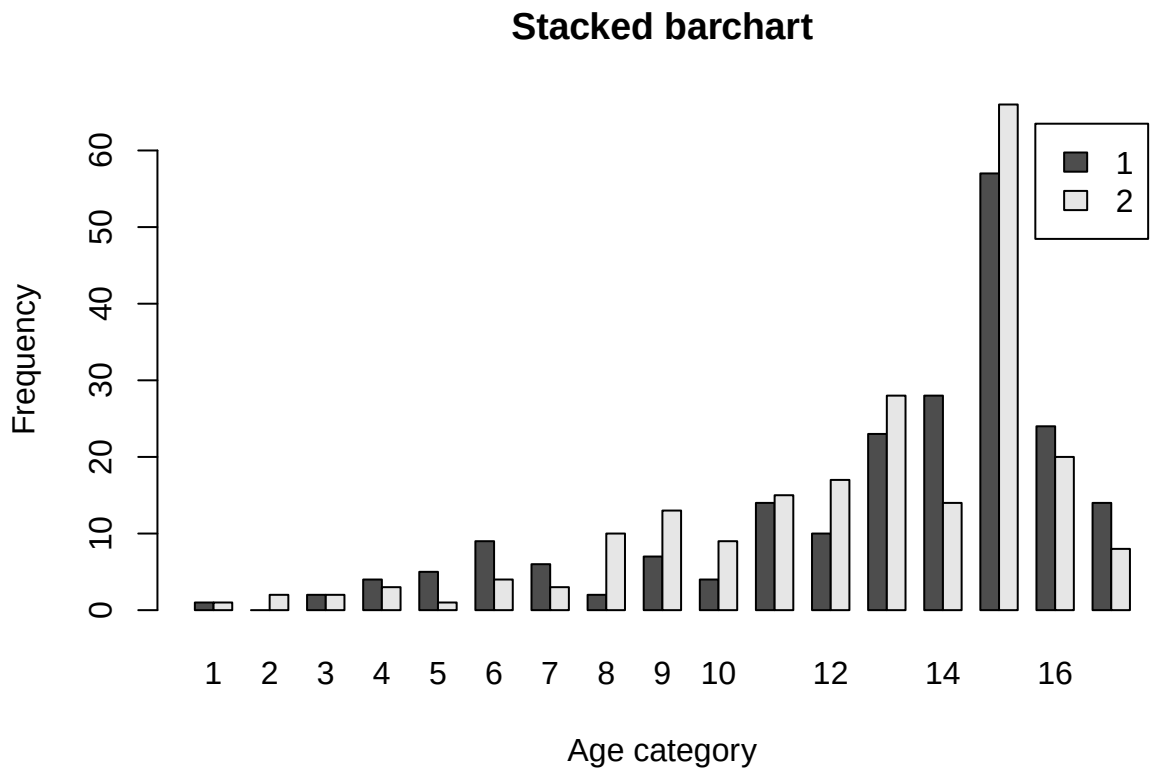
The age distribution is skewed left with older respondents tending to be more common. The median of the distribution is 14 which indicates respondents between 30 and 34 years old in the codebook. The mean is about 13 which indicates respondants between 26 and 29 years old in the codebook. # Do you think this age distribution representative of the US population? Why or why not? This age distribution seems representative of the US population because the median would indicate about half of the respondents are under 30 years old and half are above 34 years old (30 is typically defined as middle-aged). Also, the extremes of the data (12 years old and 65 years or older) are the lowest represented in the histogram which would make sense. # Is the sample balanced in terms of gender? If not, are there more females or males?

```
table(dat1$irsex)
```

```
##
##   1   2
## 210 216
```

The sample is nearly balanced in terms of gender, there are 6 more females than males in the sample, however. # Use this code to draw a stacked bar plot to view the relationship between sex and age. What can you conclude from this plot?

```
tab.agesex <- table(dat1$irsex, dat1$AGE2)
barplot(tab.agesex,
        main = "Stacked barchart",
        xlab = "Age category", ylab = "Frequency",
        legend.text = rownames(tab.agesex),
        beside = TRUE) # Stacked bars (default)
```



From this plot I can conclude that older respondants, particularly those 19 years and older were typically more likely to be female than male. In comparison, those younger than 19 years old were more likely to be male then female.

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?

Problem 5: Sexual attraction

```
summary(dat1)
```

```
##      mjage           cigever          alcever           AGE2
##  Min.   : 7.00   Min.   :1.000   Min.   :1.000   Min.   : 1.00
##  1st Qu.:14.00   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:11.00
##  Median :16.00   Median :1.000   Median :1.000   Median :14.00
##  Mean   :17.08   Mean   :1.256   Mean   :1.028   Mean   :12.77
##  3rd Qu.:18.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:15.00
##  Max.   :45.00   Max.   :2.000   Max.   :2.000   Max.   :17.00
##    sexatract       speakengl          irsex
##  Min.   : 1.00   Min.   : 1.000   Min.   :1.000
##  1st Qu.: 1.00   1st Qu.: 1.000   1st Qu.:1.000
##  Median : 1.00   Median : 1.000   Median :2.000
##  Mean   :10.09   Mean   : 1.758   Mean   :1.507
##  3rd Qu.: 2.00   3rd Qu.: 1.000   3rd Qu.:2.000
##  Max.   :99.00   Max.   :98.000   Max.   :2.000
```

According to the data, individuals tend to use marijuana the earliest.

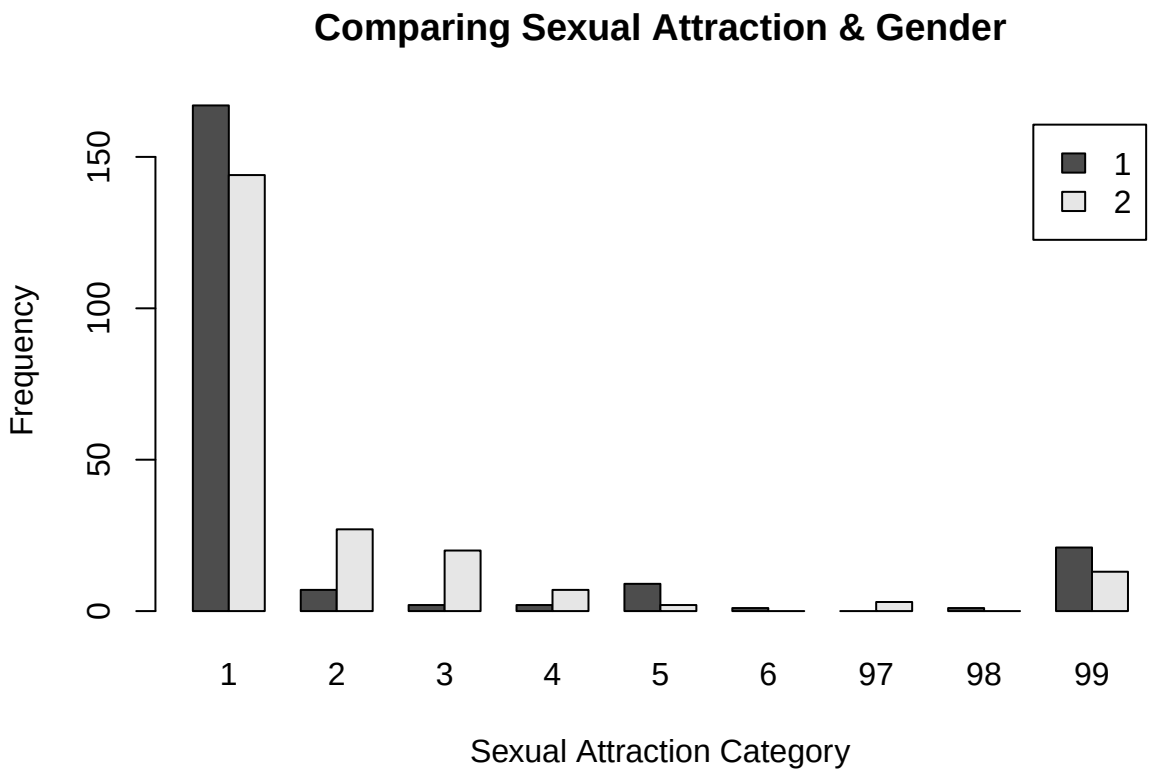
What does the distribution of sexual attraction look like? Is this what you expected?

```
table(dat1$sexatract)
```

```
##
##   1   2   3   4   5   6  97  98  99
## 311  34  22   9  11   1   3   1  34
```

The distribution is skewed right which is what I expected since only being attracted to the opposite sex (code 1) is most common. As the code numbers increased from 1 to 6, there was less strict of an attraction to the opposite sex and more openness to attraction to the same sex, so this distribution is consistent with my expectations of bisexuality and homosexuality not being as common. # What is the distribution of sexual attraction by gender?

```
tab.agesex <- table(dat1$irsex, dat1$sexatract)
barplot(tab.agesex,
        main = "Comparing Sexual Attraction & Gender",
        xlab = "Sexual Attraction Category", ylab = "Frequency",
        legend.text = rownames(tab.agesex),
        beside = TRUE) # Stacked bars (default)
```



It looks like the highest distribution is associated with respondents only attracted to the opposite sex. Females tend to show more variablity in sexual attraction than males.

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?

Are there more English speaker females or males?