```
#Carolina Herrera Figueroa
# CRIM250- Statistics for the Social Sciences
# Assignment 2
# Collaborators: Eliza Epstein

setwd("/Users/carolinaherrera/Documents/Assignment 2 - CRIM 250")
dat <- read.csv(file = 'dat.nsduh.small.1.csv')

summary(dat)
head(dat)

names(dat)

dim(dat)
# The data consists of 171 rows and 7 columns.

#Problem 2: Variables
# ANSWER: The variables in this data set are: mjage, cigage, iralcage, age2, sexatract, speakengl, irsex. (mjage)
looks at how old the individual was the first time they used marijuana or hashish, (cigage) is have they ever smoked
part of a cigarette, (iralcage) looks at how old the individual was when they first tried alcohol, (age2) represents their
coded age, (irsex) represents their gender, (sexatract) looks at what best describes their sexual atractment and
(speakengl) represents how well the individual speaks English. These are all categorical variables.
names(dat)

# 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 drug and alcohol use. It was collected using the National Survey of Drug Use and Health, and
this is a random sample of the first 1000 cases of the survey. The purpose of generating the data was to perhaps be
able to study the relationship between drug use and other variables such as age or gender. I think the results could
perhaps be used to allocate resources to individuals or groups at highest risk of an early onset of drug use.

-----------------

#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*: The age distribution of the sample is skewed with most individuals being between the ages of 35-49
years old, where the peak and the mean of the data is located, as shown by Figure 1. The lowest age of a respondent
was 15 years old. I think those over 65 years old are also underrepresented in the sample.

summary(dat)

counts <- table(dat$age2)
barplot(counts, main = "Fig 1. Age of Participants", xlab="Code for Age")

# Do you think this age distribution representative of the US population? Why or why  not?
# ANSWER: I think this age distribution is not representative of the US population because it suggests that the
population is mostly comprised of individuals between the age of 35-49 years old.

# Is the sample balanced in terms of gender? If not, are there more females or males?
# ANSWER: The same seems to be relatively balanced in terms of gender, but there are slightly more males than
females. The bar chart (Figure 2) shows us that the bar for code "1" which represents males in the codebook is taller,
meaning our sample has more males.


counts <- table(dat$irsex)
barplot(counts, main = "Fig 2. Gender of Participants", xlab="Code for Gender")

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

# ANSWER: This plot shows us that for most age categories an equal number of males and females were surveyed, with the exception of a few. It seems that the younger the participants were the more likely it was that the sample consisted of mostly males or mostly females, but it was not even. For example, for age categories 6 and 7, which translates to the respondents being 17 or 18 years old, the data is made up of males only. Starting at age category 11 we see a sample more evenly split among males and females.

```
tab.agesex <- table(dat$irsex, dat$age2)
barplot(tab.agesex,
    main = "Fig 3. Relationship between Sex and Age",
    xlab = "Age category", ylab = "Frequency",
    legend.text = rownames(tab.agesex),
    beside = FALSE) # Stacked bars (default)
```

## Problem 4: Substance use

# For which of the three substances included in the dataset (marijuana, alcohol, and cigarettes) do individuals tend to use the substance earlier?
# Answer: Individuals tends to use alcohol earlier, it was reported being consumed as early as 16 years old, compared with 18 years old for marijuana/hashnish and 21 years old for cigarettes. We can also see that 25% of the respondents were under the age of 29 when they tried alcohol for the first time, compared to 25% being below 34 years old the first time they tried marijuana/hashish and 25% of those who respondents who were under 49 years old when theey started smoking cigarettes. Alcohol is also the only substance in our sample that respondents tried for the first time under the age of 18.

```
counts <- table(dat$mjage)
barplot(counts, main = "Fig 4. Age when First Used Marijuana/Hashish", xlab="Age")

counts <- table(dat$cigage)
barplot(counts, main = "Fig 5. Age when first started Smoking Cigarettes Every day", xlab="Age")

counts <- table(dat$iralcage)
barplot(counts, main = "Fig 6. Age When First tried Alcohol", xlab="Age")
```

## Problem 5: Sexual attraction

# What does the distribution of sexual attraction look like? Is this what you expected?
# ANSWER: The distribution of sexual attraction is skewed, with a great majority of individuals reporting as being attracted only to the opposite sex. I was not surprised to see such a high number of individuals reporting to be heterosexual, but rather the difference between those who reported to be heterosexual and those who reported to be homosexual. I think we should keep in mind that it's possible that some individuals may have chosen to not report accurately out of fear of embarassment.

```
counts <- table(dat$sexatract)
barplot(counts, main = " Fig 7:Sexual Attraction", xlab="Code for Sexual Attraction")
```

# What is the distribution of sexual attraction by gender?
# Answer: The distribution of sexual attraction by gender is interesting because those who identify as strongly being heterosexual (1 on the codebook) are mostly males, while those who identify as being "mostly heterosexual" are mostly female. I thought this was interesting to see because it may be showing us the importance of the wording of the question since males were more likley to choose the option that strongly asserted their sexuality. In our sample of 1,000 individuals, those who identified as bisexual were only female. There was an even split between genders among those who identified as being attracted only to the same sex.

```
tab.sexatract <- table(dat$irsex, dat$sexatract)
barplot(tab.sexatract,
    main = "Fig 8: Relationship between Gender and Sexual Attraction",
    xlab = "Code for Sexual Attraction", ylab = "Frequency",
    legend.text = rownames(tab.sexatract),
    beside = FALSE) # Stacked bars (default)
```

```
## 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?
# ANSWER: The distribution of English speaking looks very skewed in the sample, with the majority of individuals speaking English very well. The barplot shows us that over 150 individuals speak English very well while about 10 speak English not well. This is not what I would expect of a random sample of the US population, I expected to see a greater number of individuals who spoke English "well" and "not well" considering that English is not the nativie language for a great portion of the population. I didn't expect the data to be as skewed at it is.

# Are there more English speaker females or males?
# ANSWER: There are more English speaking males, and in our sample none fall under the category of speaking English "not well", represented by the bar labeled "3" on our plot.

counts <- table(dat$speakengl)
barplot(counts, main = "Figure 9: How Well do Participants Speak English", xlab = "Categories")



tab.sexengl <- table(dat$irsex, dat$speakengl)
barplot(tab.sexengl,
        main = "Stacked barchart",
        xlab = "Code for English Speaking", ylab = "Frequency",
        legend.text = rownames(tab.sexengl),
        beside = FALSE) # Stacked bars (default)
```