# Title: Assignment 2

# Subtitle: Crim 250: Statistics for the Social Sciences

# Name: Elizabeth Stoner

# Date: 09/27/2021

# 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("C:/Users/Elizabeth/Desktop/CRIM 250")

# Read the data

library(readr)

dat\_nsduh\_small\_1 <- read\_csv("Assignment 2/dat.nsduh.small.1.csv")

View(dat\_nsduh\_small\_1)

dat <- read\_csv("Assignment 2/dat.nsduh.small.1.csv")

# What are the dimensions of the dataset?

dim(dat)

## The dimensions of the dataset are 171 observations of 7 variables.

names(dat)

## The names of the variables are mjage, cigage, iralcage, age2, sexatract, speakengl, and irsex.

## Problem 2: Variables

# Describe the variables in the dataset.

class(dat$mjage)

class(dat$cigage)

class(dat$iralcage)

class(dat$age2)

class(dat$sexatract)

class(dat$speakengl)

class(dat$irsex)

## There are seven variables in the dataset. In R, each variable is numeric. Mjage stands for the age at which respondents used marijuana or hashish for the first time. Cigage shows the ages at which respondents first started smoking cigarettes every day. Iralcage is the ages at which respondents first tried alcohol. Age2 is the most recent age of the respondents. Irsex is the gender of the respondents and is a categorical variable. Sexatract lists the sexual attraction of the respondents. Speakengl lists how well the respondents speak English.

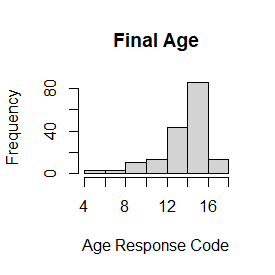
# What is this dataset about? Who collected the data, what kind of sample is it, and what was the purpose of generating the data?

## The dataset is from the National Survey of Drug Use and Health. The dataset is used to understand the extent of drug use and other health issues throughout the United States. The data were collected by the Substance Abuse and Mental Health Services Administration, part of the U.S. Department of Health and Human Services. The NSDUH uses a stratified sample within the United States population. The purpose of gathering this data is to provide an understanding of the health issues pervasive in America in order to establish helpful treatment programs and inform public policy.

## 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="Final Age", xlab="Age Response Code", ylab="Frequency")



## The majority of the age variables were listed between 12 and 16. According to the codebook, this indicates that most respondents were between 24 and 64 years old. The most frequent ages reported were from 30 to 64 years old. The least frequent ages reported were between 15 to 19 years old. In this dataset there were no ages younger than 15.

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

## I would expect to see a higher frequency of 18 to 30 year olds, rather than have the peak start at 30. Additionally, according to the codebook, the ages of respondents when the questionnaire began were different from when it ended. As a result, I expect that this distribution is not entirely reflective of the age distribution of the current U.S. population; however, it is likely not too skewed either.

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

table(dat$irsex)

## According to the table generated by R, there are 91 males and 80 females in the dataset. Therefore, it is not balanced as there are more males than females.

# 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(dat$irsex, dat$age2)

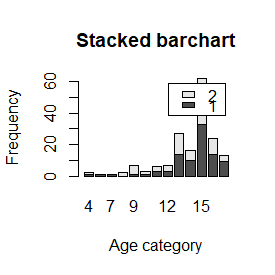
barplot(tab.agesex,

main = "Stacked barchart",

xlab = "Age category", ylab = "Frequency",

legend.text = rownames(tab.agesex),

beside = FALSE) # Stacked bars (default)



## This plot shows, as would be expected, that there were more men in each age group, aside from ages 19 to 20, than women who participated. However, the most frequent age range of both males and females was ages 35 to 49. It is also worth noting that the only participants ages 16 to 17 were male. It is at these younger ages that we see more discrepancies between male and female participants. But starting at age 21, the male-female distribution evens out a bit.

## 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?

par("mar")

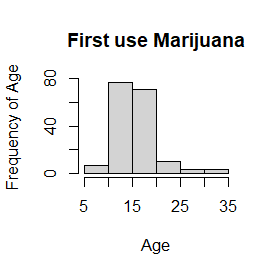
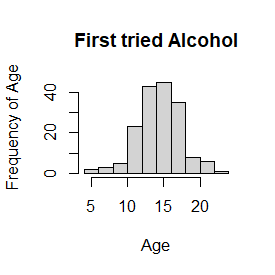
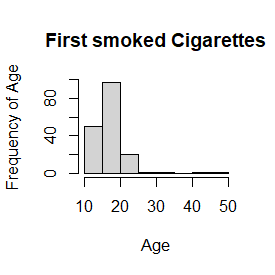
par(mar=c(1,1,1,1))

par(mfrow=c(3,1))

hist(dat$mjage, main="First use Marijuana", xlab="Age", ylab="Frequency of Age")

hist(dat$cigage, main="First smoked Cigarettes", xlab="Age", ylab="Frequency of Age")

hist(dat$iralcage, main="First tried Alcohol", xlab="Age", ylab="Frequency of Age")



## Based on the histograms, individuals tend to use alcohol earlier. It had the most frequent values of ages 5 to 10 compared to marijuana and cigarettes.

## Problem 5: Sexual attraction

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

install.packages("magrittr") # package installations are only needed the first time you use it

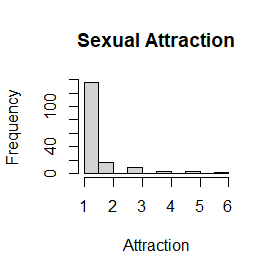
install.packages("dplyr") # alternative installation of the %>%

library(magrittr) # needs to be run every time you start R and want to use %>%

library(dplyr) # alternatively, this also loads %>%

dat$sexatract <- dat$sexatract %>% na\_if(., "99")

hist(dat$sexatract, main="Sexual Attraction", xlab="Attraction", ylab="Frequency")



## After removing the outlying values of 99, which indicates a skip of data, the histogram peaks at a value of 1. This means that the majority of respondents reported that they are only attracted to the opposite sex. Very few respondents answered that they were attracted, or mostly attracted, to the same sex. At least one respondent reported that they were unsure. I would have expected more variation in attraction with higher frequencies of people attracted to the same sex. This could be explained by the fact that the sample size is relatively small, and therefore may not be entirely representative of the U.S. population in terms of sexual attraction.

# What is the distribution of sexual attraction by gender?

tab.sexatract <- table(dat$irsex, dat$sexatract)

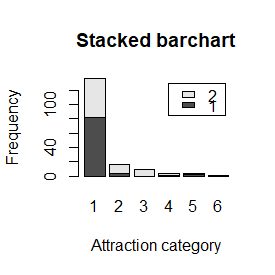
barplot(tab.sexatract,

main = "Stacked barchart",

xlab = "Attraction category", ylab = "Frequency",

legend.text = rownames(tab.sexatract),

beside = FALSE) # Stacked bars (default)

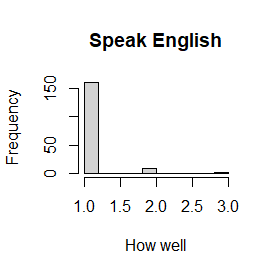


##According to the barchart more males than females reported that they were only attracted to the opposite sex. Additionally, the majority of males and females reported that they were only attracted to the opposite sex. More females than males reported that they were mostly attracted to the opposite sex or were equally attracted to both sexes. More males than females reported that they were mostly or only attracted to members of the same sex. Additionally, the respondent who reported they were unsure was male.

## 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="Speak English", xlab="How well", ylab="Frequency")



## Most of the respondents, over 150, reported that they speak English very well, with few simply answering well, and even fewer reporting that they do not speak it well. In a random sample of the U.S. population, I would expect that more people would answer that they do not speak English very well, creating more variation in the histogram. Due to the large number of immigrants from all over the world it would make sense for more people within the whole population to not feel as confident in their English speaking.

# Are there more English speaker females or males?

tab.sexengl <- table(dat$irsex, dat$speakengl)

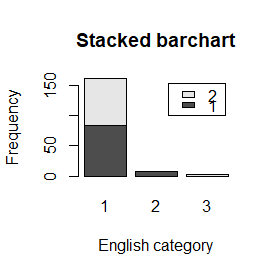
barplot(tab.sexengl,

main = "Stacked barchart",

xlab = "English category", ylab = "Frequency",

legend.text = rownames(tab.sexengl),

beside = FALSE) # Stacked bars (default)



## According to the barchart, those who speak English very well are almost evenly distributed between males and females. However, more males than females reported that they speak English well, indicating that overall, more males than females speak English within the sample. Additionally, more females than males reported that they do not speak English well.