Assignments

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

Assignment 1

Collaborators: Theodora Athanitis and Halle Wasser

Problem 1

Install the datasets package on the console below using `install.packages("datasets")`. Now load the library.

```
# install.packages('datasets')
library(datasets)
```

Answer: Installed!

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: Renaming data sets ensures that you do not edit any of your original data on accident, and it is also easier to type.

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"
```

Answer: The variables are Murder, Assault, Urban Population, and Rape, in addition to the new variable that we just created, state.

Problem 3

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

Answer: Murder is a quantitative variable according to the DVB chapter because we are measuring the amount of murder arrests per 100,000.

What R Type of variable is it?

```
class(dat$Murder)
```

```
## [1] "numeric"
```

Answer: In R, the data points in the Murder column are numeric variables according to the class() function.

Problem 4

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

```
head(dat)
```

```
##            Murder Assault UrbanPop Rape       state
## Alabama      13.2     236       58 21.2     alabama
## Alaska       10.0     263       48 44.5      alaska
## Arizona       8.1     294       80 31.0     arizona
## Arkansas      8.8     190       50 19.5    arkansas
## California    9.0     276       91 40.6  california
## Colorado      7.9     204       78 38.7    colorado
```

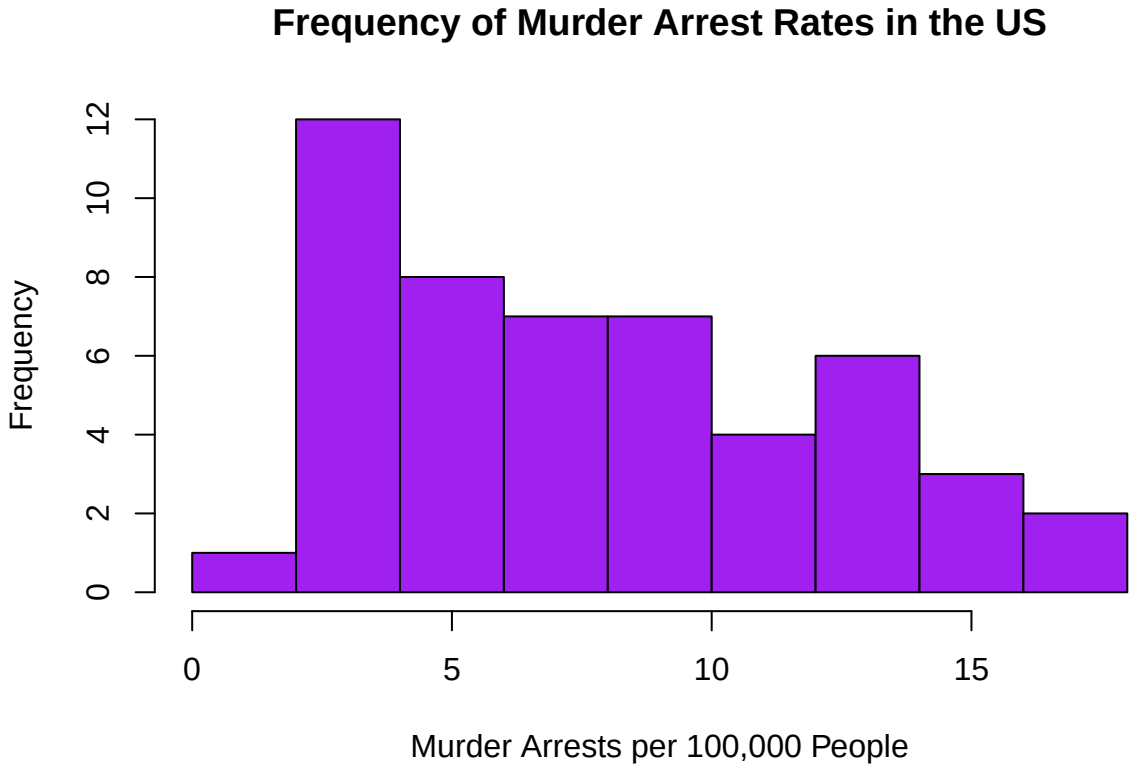
```
`?`(USArrests)
```

Answer: Generally, this data set contains rates for specific crime arrests per 100,000 people collected by D.R. McNeil, as well as percentages of urban population per state in 1973. The numbers in the Murder, Assault, and Rape columns are equivalent to the calculated arrest rates for Murder, Assault, and Rape. The Urban Population numbers are the percentages of the population in each state that live in an urban area.

Problem 5

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

```
hist(dat$Murder, xlab = "Murder Arrests per 100,000 People",
    main = "Frequency of Murder Arrest Rates in the US", col = "purple")
```



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)
```

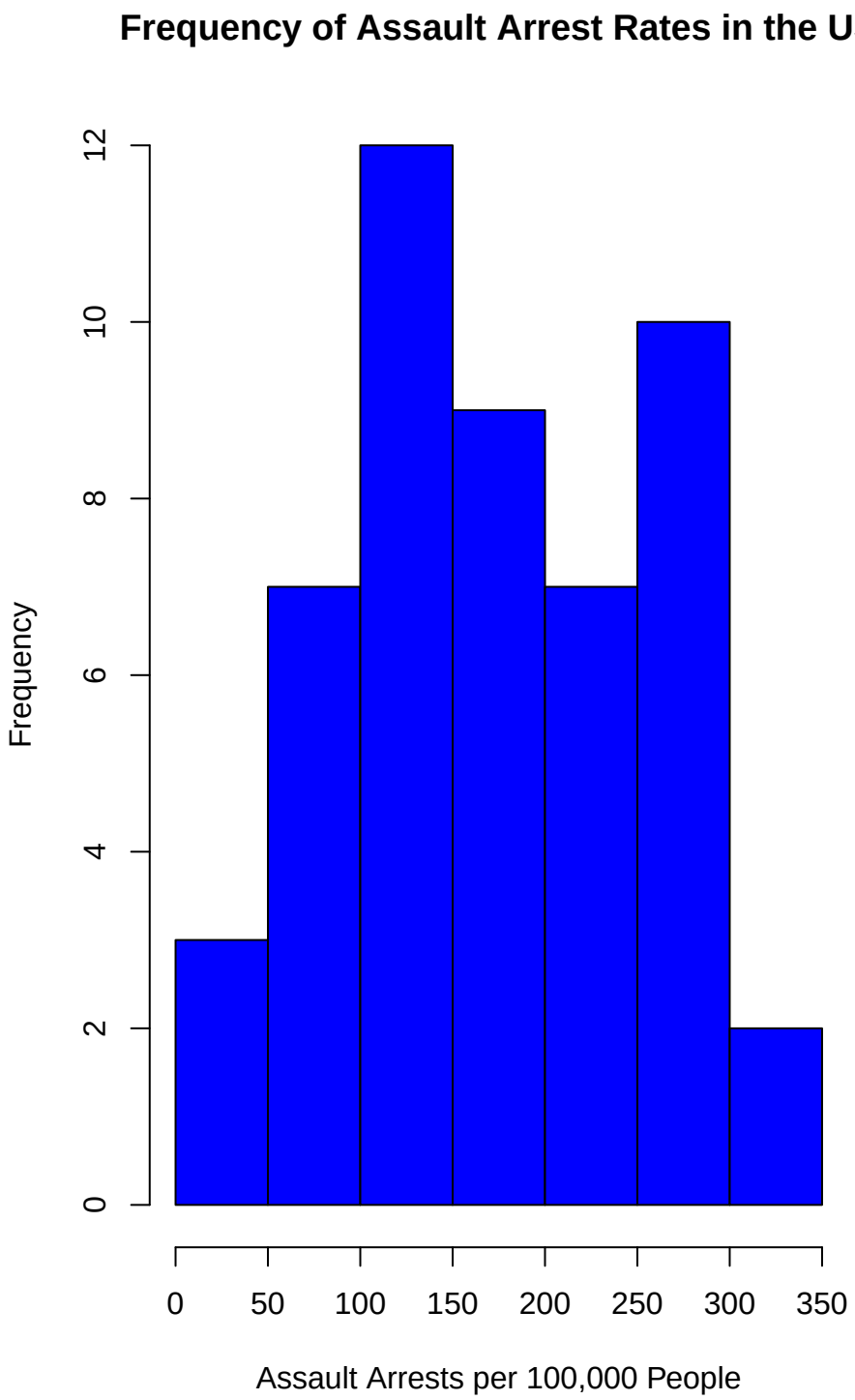
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   4.075   7.250   7.788  11.250  17.400
```

Answer: The mean is 7.788. The median is 7.250. Q1 is 4.075 and Q3 is 11.250. The minimum value is .800 and the maximum value is 17.400. Mean and median are both measures of center; however, the mean is the average value of a data set, which is calculated by adding all of the values and dividing by the number of total data points, and the mean is calculated by finding the middle value when the data set is arranged in order. A quartile is essentially the same as a median, but with 25% and 75% of the data, rather than 50%. This means that 25% of the data in a data set is below the 1st quartile, 50% is below the second quartile (the median), and 75% is below the 3rd quartile. R gives you quartiles in order to help visualize the data and see if it is distributed symmetrically, and only gives you quartiles 1 and 3 because the median is the 2nd quartile.

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.

```
hist(dat$Assault, xlab = "Assault Arrests per 100,000 People",
    main = "Frequency of Assault Arrest Rates in the US", col = "Blue")
```



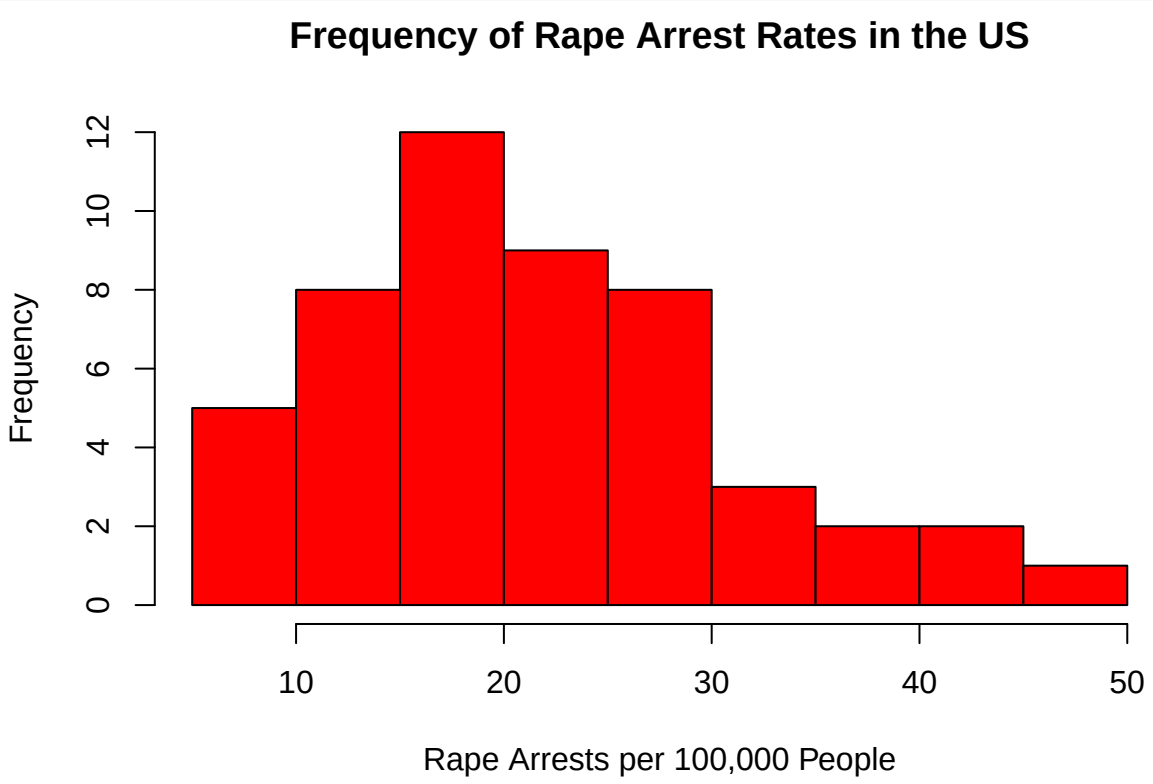

```
summary(dat$Assault)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

Answer: The mean is 170.8. The median is 159.0. Q1 is 109.0 and Q3 is 249.0. The minimum value is 45.0 and the maximum value is 337.0.

```
hist(dat$Rape, xlab = "Rape Arrests per 100,000 People", main = "Frequency of Rape Arrest Rates in the U
```

```
    col = "Red")
```

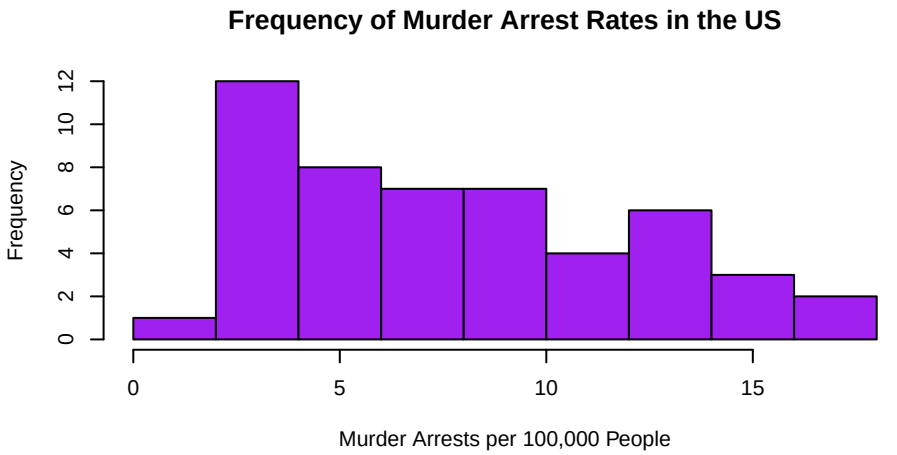


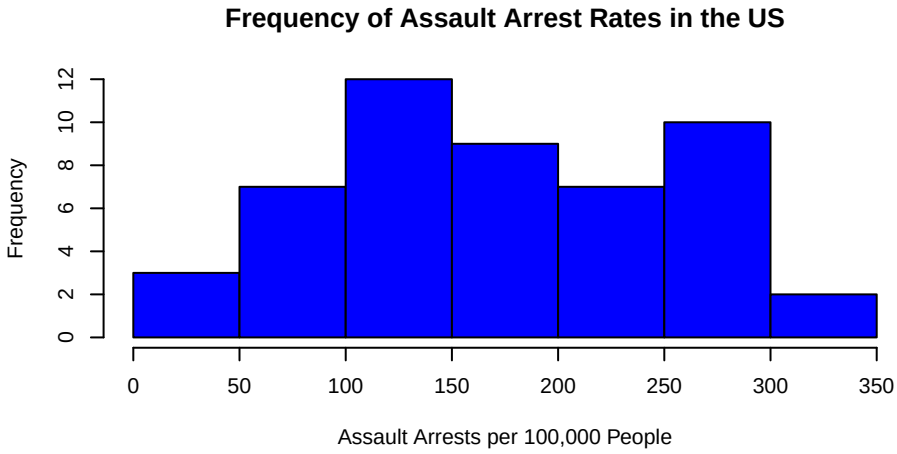

```
summary(dat$Rape)
```

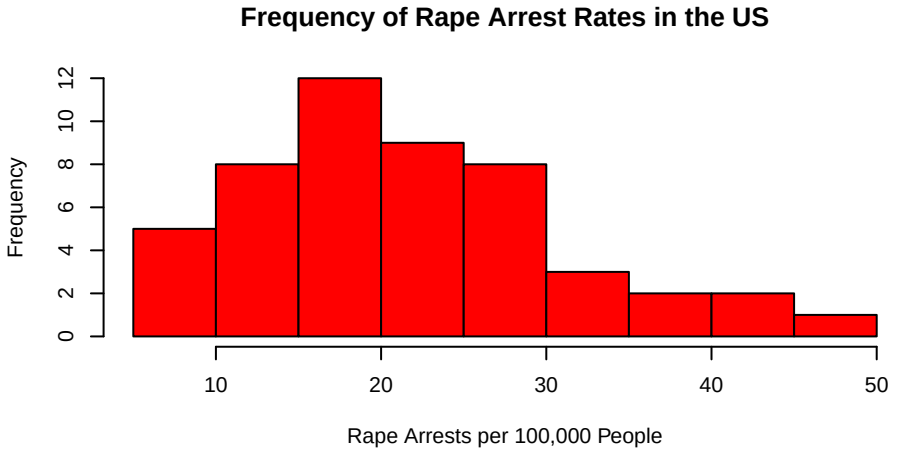
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.30   15.07   20.10   21.23   26.18   46.00
```

Answer: The mean is 21.23. The median is 20.10. Q1 is 15.07 and Q3 is 26.18. The minimum value is 7.30 and the maximum value is 46.00.

```
par(mfrow = c(3, 1))
hist(dat$Murder, xlab = "Murder Arrests per 100,000 People",
    main = "Frequency of Murder Arrest Rates in the US", col = "purple")
hist(dat$Assault, xlab = "Assault Arrests per 100,000 People",
    main = "Frequency of Assault Arrest Rates in the US", col = "Blue")
hist(dat$Rape, xlab = "Rape Arrests per 100,000 People", main = "Frequency of Rape Arrest Rates in the U
    col = "Red")
```







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

Answer: The 'par' command sets parameters for graphs, changes how they are displayed, and allows us to tell R to compare additional variables within the data set. 'mfrow' puts things in an array, and the parameters of (1,3) tell it to display them stacked on top of each other, rather than side by side or otherwise.

What can you learn from plotting the histograms together?

Answer: We can learn that assault arrest rates per 100,000 people are far higher than the arrest rates for

murder and rape by comparing the x axis of all three histograms. Additionally, while the frequency of arrest rates for rape and murder skew relatively right and are unimodal, the frequency of assault arrest rates is bimodal. Comparing and plotting them together can generally help us visually compare distributions overall.

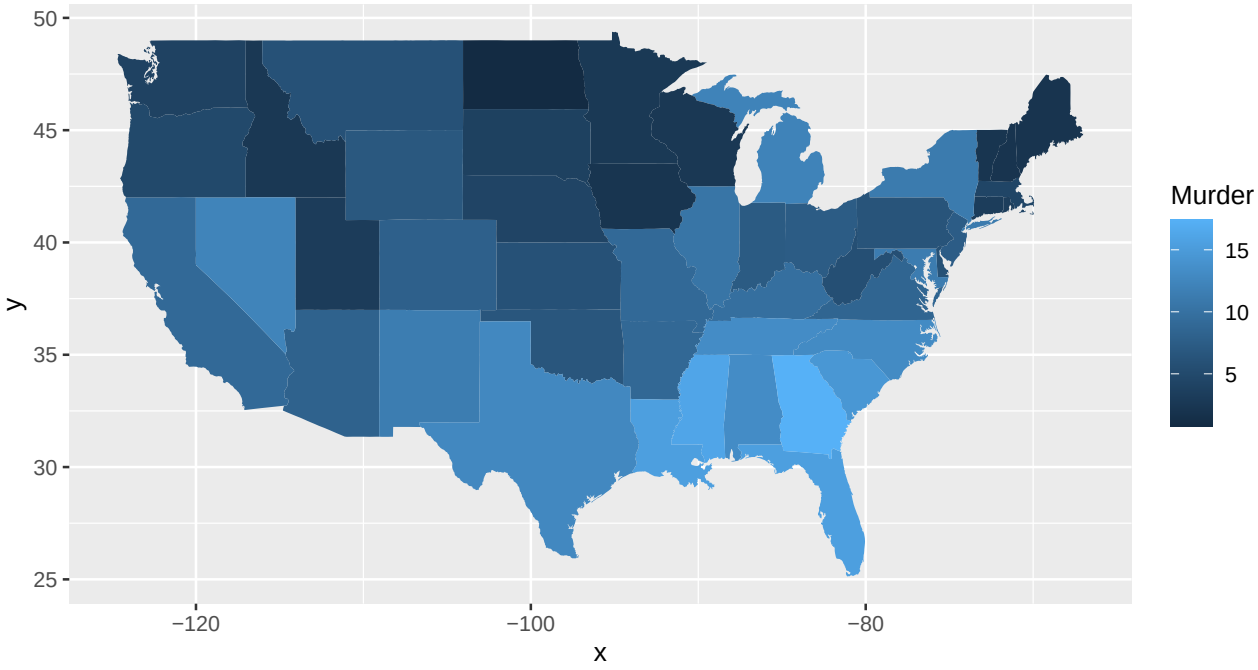
Problem 8

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:

```
# install.packages('maps')
#- This line installs a package that allows for map-drawing based on state names provided within the da
# install.packages('ggplot2')
#- This line installs a package that makes improves the data visualization so that the map is easier to
# library('maps')
#- This line tells R where to find the package to help draw the map
# library('ggplot2')
#- This line tells R where to find the package to help draw the map
# 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)
#- This line tells R to use ggplot and to take information from dat variable 'state', and then to draw
```

Assignment 2

Title: Assignment 2

Subtitle: Crim 250: Statistics for the Social Sciences

Name: Tori Borlase

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("/Users/toriborlase/Desktop/University of Pennsylvania/Fall 2021/CRIM 250/Tori-Borlase-Crim-250")
```

Read the data

```
dat <- read.csv(file = "dat.nsduh.small.1.csv")
```

What are the dimensions of the dataset?

dim(dat)

#Answer: 171 by 7. 171 rows and 7 columns.

Problem 2: Variables

Describe the variables in the dataset.

names(dat)

Answer: The variables in the dataset are mjage, cigage, iralcage, age2, sexatract, speakengl, and irsex. According to the code book, these correspond to how old someone was when they first tried marijuana or hashish, how old someone was when they started smoking cigarettes every day, how old someone was when they first tried alcohol, the final age that someone was determined to be at the time of taking the survey (which was asked multiple times as a consistency check), which statement about sexual orientation best described the respondent's feelings, how well they speak English, and imputation revised 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?

Answer: This dataset is from the 2019 National Survey on Drug Use and Health (NSDUH). Individuals within the Center for Behavioral Health Statistics and Quality collected this data in order to measure the "prevalence and correlates of substance use and mental health issues" in the US. This was a stratified random sample, as they collected data in all 50 states for civilian and non-institutionalized populations.

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.

counts <- table(dat$age2) barplot(counts, main="Histogram of Ages of Participants", xlab="Age Category", ylab="Frequency")

Answer: The age distribution of this sample is skewed left. However, because the numbers within the data set are not actual ages, but rather a range of ages or individual ages that are coded as numbers 1 through 17, we cannot base our true age distribution off of the numbers that form the labels of the histogram. While it appears that the Final Edited Age is skewed left, this may just be because the upper coded ages contain wider ranges of ages, and the lower coded ages contain only one or two ages.

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

No, I do not believe this age distribution is representative of the US population. According to Census data, and other population pyramids that display the distribution of population, there are many more younger people than are represented in this graph, as well as many individuals outside of the scope of the study, such as those under 12, etc. Even though (as I mentioned before) there are multiple ages within certain categories, this data shows many more older individuals compared to teenagers and those in their early twenties, and Census data shows large numbers of those populations.

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

table(dat$irsex)

1 (Male) 2 (Female)

91 80

counts <- table(dat$irsex) barplot(counts, main="Gender Distribution", xlab="Gender", ylab="Frequency", names=c("Male", "Female"))

The sample is not gender balanced, as there are more males (91) than females (80). Using a bar chart clearly demonstrates this disparity.

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 = "Age and Gender Comparison", xlab = "Age Category", ylab = "Frequency", legend.text = c("Male", "Female"), xlim = c(0,17), beside = FALSE) # Stacked bars (default)

We can conclude from this plot that for most age categories, there were more males than females. However, for age categories 8, 9, 13, and 15, there appear to be more females or the same number of males and females.

Problem 4: Substance use

For which of the three substances included in the data set (marijuana, alcohol, and cigarettes) do individuals tend to use the substance earlier?

par(mfrow = c(3,1)) hist(dat$mjage, main = "Histogram of Age of First Marijuana Use", xlab = "Age Categories", ylab = "Frequency", xlim = c(0, 50), ylim = c(0, 50), breaks = 20) hist(dat$iralcage, main="Histogram of Age of First Alcohol Use", xlab="Age Categories", ylab="Frequency", xlim=c(0,50), ylim=c(0,50), breaks = 20) hist(dat$cigage, main="Histogram of Age of Starting to Smoke Cigarettes Daily", xlab="Age Categories", ylab="Frequency", xlim=c(0,50), ylim=c(0,50), breaks = 20)

Individuals tend to use alcohol earlier, as seen on the histograms.

Problem 5: Sexual attraction

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

dat1 <- dat[dat$sexatract !=99,] counts1 <- table(dat1$sexatract) barplot(counts1, main="Sexual Attraction Distribution", xlab="Sexual Attraction Categories", ylab="Frequency", ylim = c(0,150))

The distribution of sexual attraction is skewed right, which is what I expected, as most people identify as only being attracted to the opposite sex, with fewer people identifying as being attracted to the same sex in any way. This is what I expected because I believe that LGBT+ populations are still a relatively small minority within the US compared to those who are only attracted to the opposite sex.

What is the distribution of sexual attraction by gender?

dat1 <- dat[dat$sexatract !=99,] tab.sexor <- table(dat1$irsex, dat1sexatract) barplot(tab.sexor, main = "Sexual Attraction and Gender Comparison", xlab = "Sexual Attraction Category", ylab = "Frequency", legend.text = c("Male", "Female"), xlim = c(0,7), ylim = c(0,140), beside = FALSE) # Stacked bars (default)

The distribution by gender is also skewed right, but there are more females that identify with statements about being attracted to the same sex in some way, and more males that identify with statements about being attracted to the opposite sex, as can be seen in the stacked bar chart.

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?

counts2 <- table(dat$speakengl) barplot(counts2, main="English Language Category Frequency", xlab="English Speaking Categories", ylab="Frequency", ylim = c(0,200))

This distribution is also skewed right, with most individuals responding that they speak English very well, and less than 50 individuals total saying that the speak English well or not well, and nobody saying they did not speak English at all. This is similar to the distribution that I would expect in the United States because even though the US has no official language, most people need to speak some English in order to work and live here. However, there are probably more people in the US that speak no English at all, but they just were not able to participate in the survey because it was conducted in English.

Are there more English speaker females or males?

table(dat$irsex, dat$speakengl) tab.sexor <- table(dat$irsex, dat$speakengl) barplot(tab.sexor, main = "English Skill and Gender Comparison", xlab = "English Ability Category", ylab = "Frequency", legend.text = c("Male", "Female"), xlim = c(0,4), ylim = c(0,200), beside = FALSE) # Stacked bars (default)

There are more male English speakers than female English speakers, but that might also be due to the fact that there are more males than females within the data set to begin with.