Assignments

Table of Contents

# Assignment 1

**Collaborators: Carolina Herrera Figueroa, Niko Amber**

### Problem 1

Install the datasets package on the console below and load the data

dat<-USArrests

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?

Answer: It is useful to rename the data set for convenience and easy accessibility.

USArrests

## Murder Assault UrbanPop Rape  
## Alabama 13.2 236 58 21.2  
## Alaska 10.0 263 48 44.5  
## Arizona 8.1 294 80 31.0  
## Arkansas 8.8 190 50 19.5  
## California 9.0 276 91 40.6  
## Colorado 7.9 204 78 38.7  
## Connecticut 3.3 110 77 11.1  
## Delaware 5.9 238 72 15.8  
## Florida 15.4 335 80 31.9  
## Georgia 17.4 211 60 25.8  
## Hawaii 5.3 46 83 20.2  
## Idaho 2.6 120 54 14.2  
## Illinois 10.4 249 83 24.0  
## Indiana 7.2 113 65 21.0  
## Iowa 2.2 56 57 11.3  
## Kansas 6.0 115 66 18.0  
## Kentucky 9.7 109 52 16.3  
## Louisiana 15.4 249 66 22.2  
## Maine 2.1 83 51 7.8  
## Maryland 11.3 300 67 27.8  
## Massachusetts 4.4 149 85 16.3  
## Michigan 12.1 255 74 35.1  
## Minnesota 2.7 72 66 14.9  
## Mississippi 16.1 259 44 17.1  
## Missouri 9.0 178 70 28.2  
## Montana 6.0 109 53 16.4  
## Nebraska 4.3 102 62 16.5  
## Nevada 12.2 252 81 46.0  
## New Hampshire 2.1 57 56 9.5  
## New Jersey 7.4 159 89 18.8  
## New Mexico 11.4 285 70 32.1  
## New York 11.1 254 86 26.1  
## North Carolina 13.0 337 45 16.1  
## North Dakota 0.8 45 44 7.3  
## Ohio 7.3 120 75 21.4  
## Oklahoma 6.6 151 68 20.0  
## Oregon 4.9 159 67 29.3  
## Pennsylvania 6.3 106 72 14.9  
## Rhode Island 3.4 174 87 8.3  
## South Carolina 14.4 279 48 22.5  
## South Dakota 3.8 86 45 12.8  
## Tennessee 13.2 188 59 26.9  
## Texas 12.7 201 80 25.5  
## Utah 3.2 120 80 22.9  
## Vermont 2.2 48 32 11.2  
## Virginia 8.5 156 63 20.7  
## Washington 4.0 145 73 26.2  
## West Virginia 5.7 81 39 9.3  
## Wisconsin 2.6 53 66 10.8  
## Wyoming 6.8 161 60 15.6

dat<-USArrests

### Problem 2

List the variables contained in the dataset:

The four variables within the dataset are Murder, Assault, Urbanpop and Rape

### Problem 3

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

Answer: categorical

What R Type of variable is it?

Answer: character

### Problem 4

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

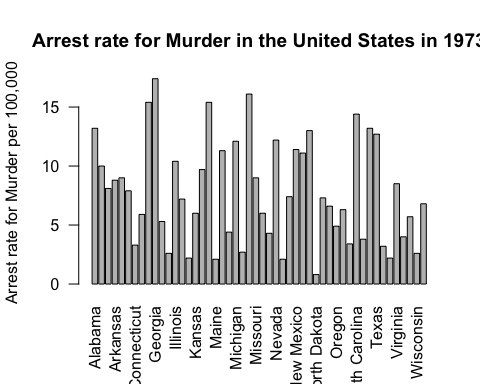
Answer: The data set shows the arrest rate per 100,000 in the US in 1973. The rows show each state and the columns show the type of arrest. Each number shows the amount of arrests of that type within that state per 100,000.

### Problem 5

Draw a histogram of Murder with proper labels and title.

I chose to do a bar chart instead due to the categorical nature of the values.

state.names = row.names(USArrests)  
barplot(USArrests$Murder, names.arg = state.names, las = 2, ylab = "Arrest rate for Murder per 100,000",   
 main = "Arrest rate for Murder in the United States in 1973")



### Problem 6

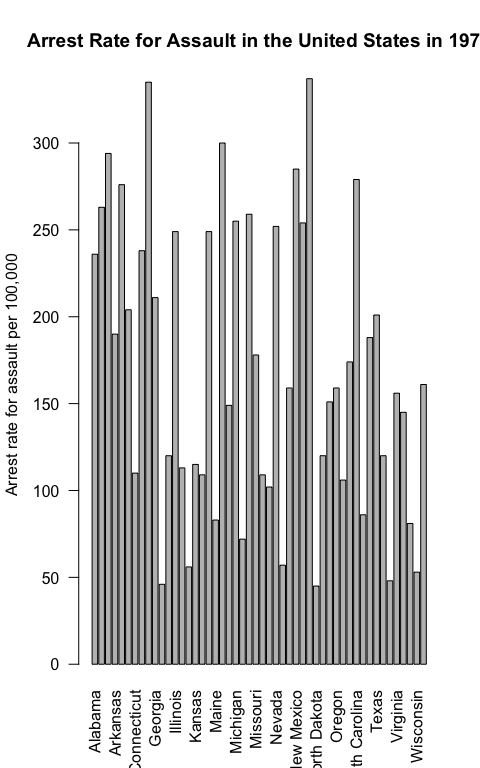
Please summarize Murder quantitatively. What are its mean and median? What is the difference between mean and median? What is a quartile, and why do you think R gives you the 1st Qu. and 3rd Qu.?

Answer: The mean for murder is 7.788 and the median for murder is 7.250. Mean is the average of the data (that is: if you were to add up all of the values then divide by the amount of values present). Median is the middle value (half of the values are above and half are below). If the data is well distributed, mean and median will be similar or the same, but the major differences occur when there are outliers: mean is impacted by outliers whereas median is not. In this data seet, mean and median appear rather similar. Quartiles are the data broken up into 4 parts. R gives the 1st quartile to give a sense of the values up until 25% of the data and the 3rd quartile to give values up until 75% of the data.

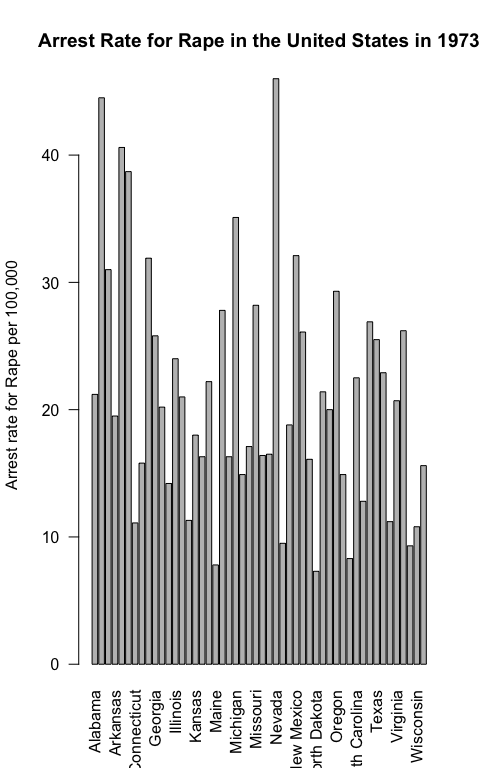
### Problem 7

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.

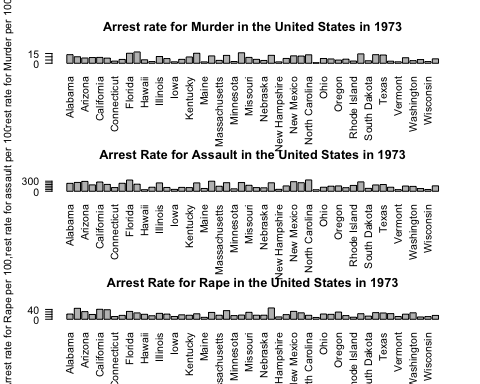
state.names = row.names(USArrests)  
barplot(USArrests$Assault, names.arg = state.names, las = 2, ylab = "Arrest rate for assault per 100,000",   
 main = "Arrest Rate for Assault in the United States in 1973")



state.names = row.names(USArrests)  
barplot(USArrests$Rape, names.arg = state.names, las = 2, ylab = "Arrest rate for Rape per 100,000",   
 main = "Arrest Rate for Rape in the United States in 1973")



par(mfrow=c(3,1))  
 state.names = row.names(USArrests)  
barplot(USArrests$Murder, names.arg = state.names, las = 2, ylab = "Arrest rate for Murder per 100,000",   
 main = "Arrest rate for Murder in the United States in 1973")  
  
state.names = row.names(USArrests)  
barplot(USArrests$Assault, names.arg = state.names, las = 2, ylab = "Arrest rate for assault per 100,000",   
 main = "Arrest Rate for Assault in the United States in 1973")  
  
 state.names = row.names(USArrests)  
barplot(USArrests$Rape, names.arg = state.names, las = 2, ylab = "Arrest rate for Rape per 100,000",   
 main = "Arrest Rate for Rape in the United States in 1973")



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

Answer: Command par allows for multiple graphs to be plotted together. This command makes this possible by defining parameters.

What can you learn from plotting the histograms together?

Answer: By plotting the histograms together it allows us to visually compare the rates for each. While it is useful to compare the numbers when looking at the table, plotting the values together allows for visual comparison.

### Problem 8

What does the given code do? Explain what each line is doing.

Answer: This code creates a heat map of murders in the US. That is, this is a map of the country with darker shades of blue signifying higher murder rates and lighter shades of blue portraying lower murder rates. The first line tells R that you will be making a heat map, arranging by state and filling with arrests for murder rates. The next line sets up the map itself separating by state and the final line expands the range

$$\\[2in]$$

# Assignment 2

**Collaborators: Carolina Herrera Figueroa, Niko Amber**

## Problem 1: Load data

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

setwd("/Users/elizaepstein/Desktop/testgithub/ElizalearnR")

Read the data

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

What are the dimensions of the dataset?

names(dat)

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

nrow(dat)

## [1] 171

ncol(dat)

## [1] 7

Answer: There are 7 columns and 171 rows.

## Problem 2: Variables

Describe the variables in the dataset.

Answer: The variables in this dataset are the age that an individual began smoking cigarettes, the age they began drinking alcohol, their age, their sexual attraction, whether they speak english and their sex.

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 data set shows the trends in substance abuse by age, sex, sexual attraction and their english language proficiency. This data was collected through the National Survey of Drug Use and Health; it is a sample of the first 1,000 individuals of the survey. The purpose of this data was to analyze trends in substance abuse. While it is not stated, these trends may be used to further understand which individuals to target for treatment.

## 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 starts from 12 and goes to 65+. The codebook explains how each age is labelled a number, so “17” represents individuals who are 65 or older.

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

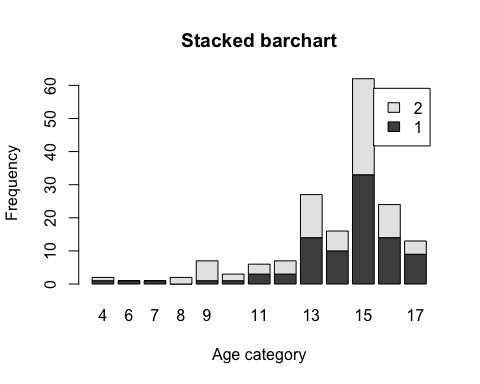
Answer: I do believe that the age distribution is representative of the US population because it covers all ages twelve and above. Due to the fact that this survey is questioning the ages someone began using substances, it is reasonable that it begins at 12.

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

Answer: I believe the sample is mostly balanced in terms of gender, but there are slightly more females. As seen in the codebook, the sample consists of 47.72% male and 52.28% 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)



Answer: This plot shows the frequency of sex of the respondent by age. For most of the younger age groups there seems to be more female respondents then the group between eight to twelve show more male respondents. Group 15 had the most respondents and seems to show a pretty even split between gendeers.

## Problem 4: Substance use

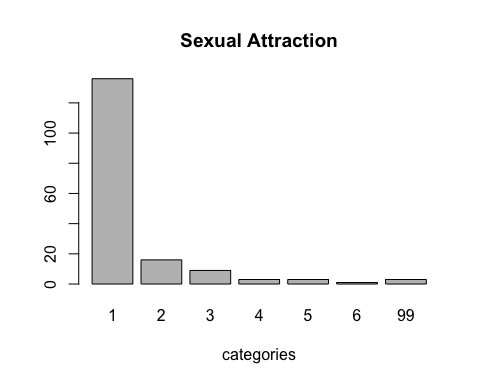
For which of the three substances included in the dataset (marijuana, alcohol, and cigarettes) do individuals tend to use the substance earlier?

Answer: Of the three substances included individuals tend to use alcohol the youngest.

## Problem 5: Sexual attraction

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

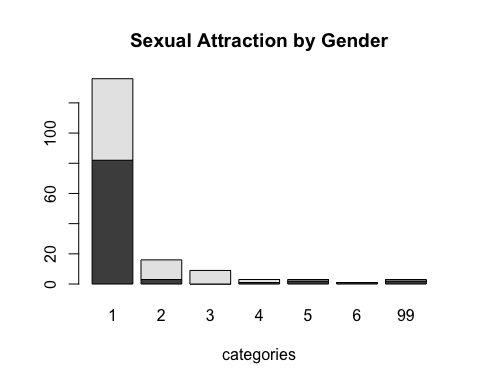
counts<- table(dat$sexatract)  
barplot(counts,main= "Sexual Attraction", xlab="categories")



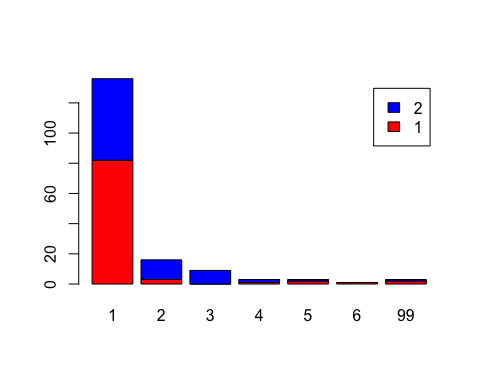
Answer: The distribution of sexual attraction is heavily weighted towards number 1 which indicated heterosexuality. I was not surprised by there being the greatest weight on heterosexuality, but I was slightly surprised quite how extreme the difference is.

What is the distribution of sexual attraction by gender?

counts <- table(dat$irsex, dat$sexatract)  
barplot(counts,main= "Sexual Attraction by Gender", xlab="categories")



barplot(counts, col=c("red", "blue"), legend=TRUE)

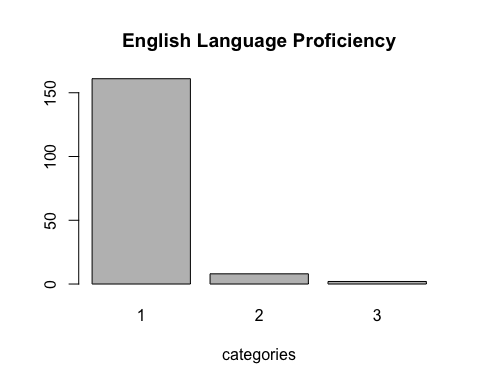


Answer: The distribution of sexual attraction by gender is intriguing to me because the first group (strongly heterosexual) appers mostly weighted towards males. The group showing bisexuality is heavily weighted towards females. The group identifying as homosexual is mostly male. The groups not knowing or leaving blank are all female and the group skipping the question is mostly 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?

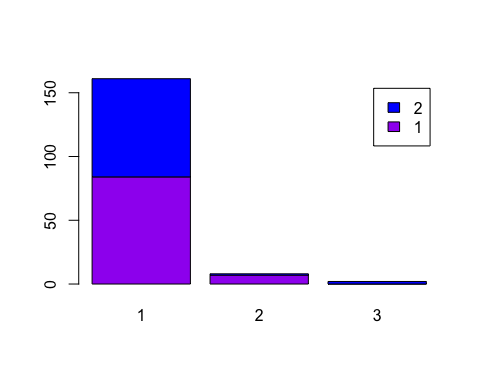
counts<- table(dat$speakengl)  
barplot(counts,main= "English Language Proficiency", xlab="categories")



Answer: The distribution of english speaking proficiency is heavily weighted towards the number 1 which indicates being able to speak english very well. I was not surpsied by this within the sample

Are there more English speaker females or males?

counts <- table(dat$irsex, dat$speakengl)  
barplot(counts, col=c("purple", "blue"), legend=TRUE)



Answer: For those that speak english very well there is a somewhat even split between the sexes, but there are overall more females. That said, there are more males than females who answered that they speak english “well”