Assignment Journal

Contents

This is where all of my assignments and exams for CRIM250 are located.

Assignment 1

Problem 1

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

```
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
```

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
dat.USArrests <- dat
```

Answer: It is useful to rename the dataset for two reasons. First, it will help you keep track of your work and not confuse it with other generic-looking names of other datasets. Second, it will allow you to keep an original copy of the file while creating a new file with all of the changes you are currently making on it.

Problem 2

Use this command to make the state names into a new variable called State.

```
dat.USArrests$state <- tolower(rownames(USArrests))
dat.USArrests
```

```
##               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
## Connecticut      3.3     110       77 11.1  connecticut
## Delaware         5.9     238       72 15.8     delaware
## Florida         15.4     335       80 31.9      florida
## Georgia         17.4     211       60 25.8      georgia
## Hawaii           5.3      46       83 20.2       hawaii
## Idaho            2.6     120       54 14.2        idaho
## Illinois        10.4     249       83 24.0     illinois
## Indiana          7.2     113       65 21.0      indiana
## Iowa             2.2      56       57 11.3         iowa
## Kansas           6.0     115       66 18.0       kansas
```

```
## Kentucky            9.7    109    52 16.3      kentucky
## Louisiana          15.4    249    66 22.2     louisiana
## Maine               2.1     83    51  7.8         maine
## Maryland           11.3    300    67 27.8      maryland
## Massachusetts       4.4    149    85 16.3 massachusetts
## Michigan           12.1    255    74 35.1      michigan
## Minnesota           2.7     72    66 14.9     minnesota
## Mississippi        16.1    259    44 17.1   mississippi
## Missouri            9.0    178    70 28.2      missouri
## Montana             6.0    109    53 16.4       montana
## Nebraska            4.3    102    62 16.5      nebraska
## Nevada             12.2    252    81 46.0        nevada
## New Hampshire       2.1     57    56  9.5 new hampshire
## New Jersey          7.4    159    89 18.8    new jersey
## New Mexico         11.4    285    70 32.1    new mexico
## New York           11.1    254    86 26.1      new york
## North Carolina     13.0    337    45 16.1 north carolina
## North Dakota        0.8     45    44  7.3  north dakota
## Ohio                7.3    120    75 21.4          ohio
## Oklahoma            6.6    151    68 20.0      oklahoma
## Oregon              4.9    159    67 29.3        oregon
## Pennsylvania        6.3    106    72 14.9  pennsylvania
## Rhode Island        3.4    174    87  8.3   rhode island
## South Carolina     14.4    279    48 22.5 south carolina
## South Dakota        3.8     86    45 12.8  south dakota
## Tennessee          13.2    188    59 26.9     tennessee
## Texas              12.7    201    80 25.5         texas
## Utah                3.2    120    80 22.9          utah
## Vermont             2.2     48    32 11.2       vermont
## Virginia            8.5    156    63 20.7      virginia
## Washington          4.0    145    73 26.2    washington
## West Virginia       5.7     81    39  9.3 west virginia
## Wisconsin           2.6     53    66 10.8     wisconsin
## Wyoming             6.8    161    60 15.6       wyoming
```

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.USArrests)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"      "state"
```

Answer: The variables include Murder, Assault, Rape, Urban Population, and State.

Problem 3

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

Answer: In the DVB chapter, "Murder" would be considered a qualitative, or categorical, variable.

What R Type of variable is it?

*Answer:* "Murder" is considered a character in R.

3

Problem 4

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

Answer: The dataset includes the arrest rates for murder, assault, and rape per 100,000 residents in each of the US's 50 states. Additionally, the percent of the population living in urban areas is given. Here, then, the numbers mean either the arrest rates for a crime per 100k residents in a state or the percent of residents living in urban spaces in a state.

Problem 5

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

```
hist(dat.USArrests$Murder, main="Histogram of Murder Rates", xlab="Murder Rates per State", ylab="Freque
```

Histogram of Murder Rates
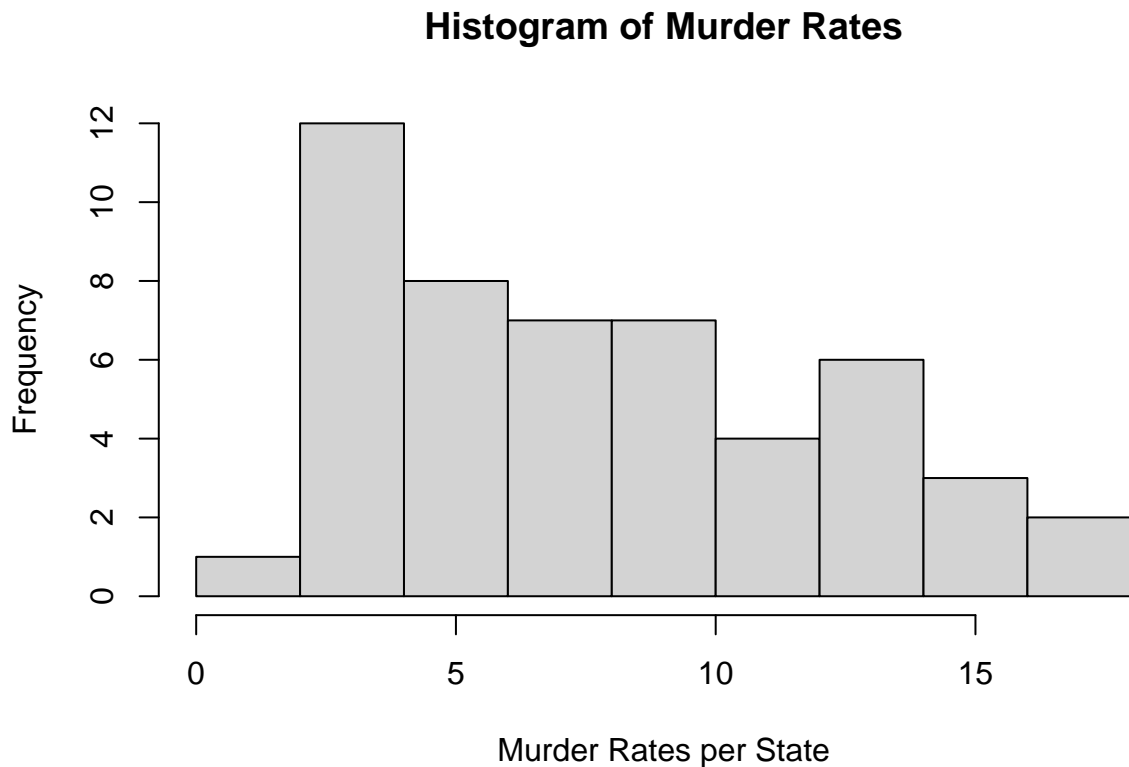
Problem 6

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.USArrests$Murder)
```

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

Answer: The mean of "Murder" is 7.788, while its median is 7.250. Generally, mean signifies the solution of all of the values added together and then divided by the number of values, while median signifies the middle value when all values are lined up in ascending order. A quartile constitutes one of three values that divides a data distribution into fourths. Lastly, R would provide the first and third quartile in order to help the statistician understand where the majority of values lie (in between the first and third quartile) or what values might be considered outliers (before the first and after the third).

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.USArrests$Assault, main="Histogram of Assault Rates", xlab="Assault Rates per State", ylab="Fr
```


**Histogram of Assault Rates**
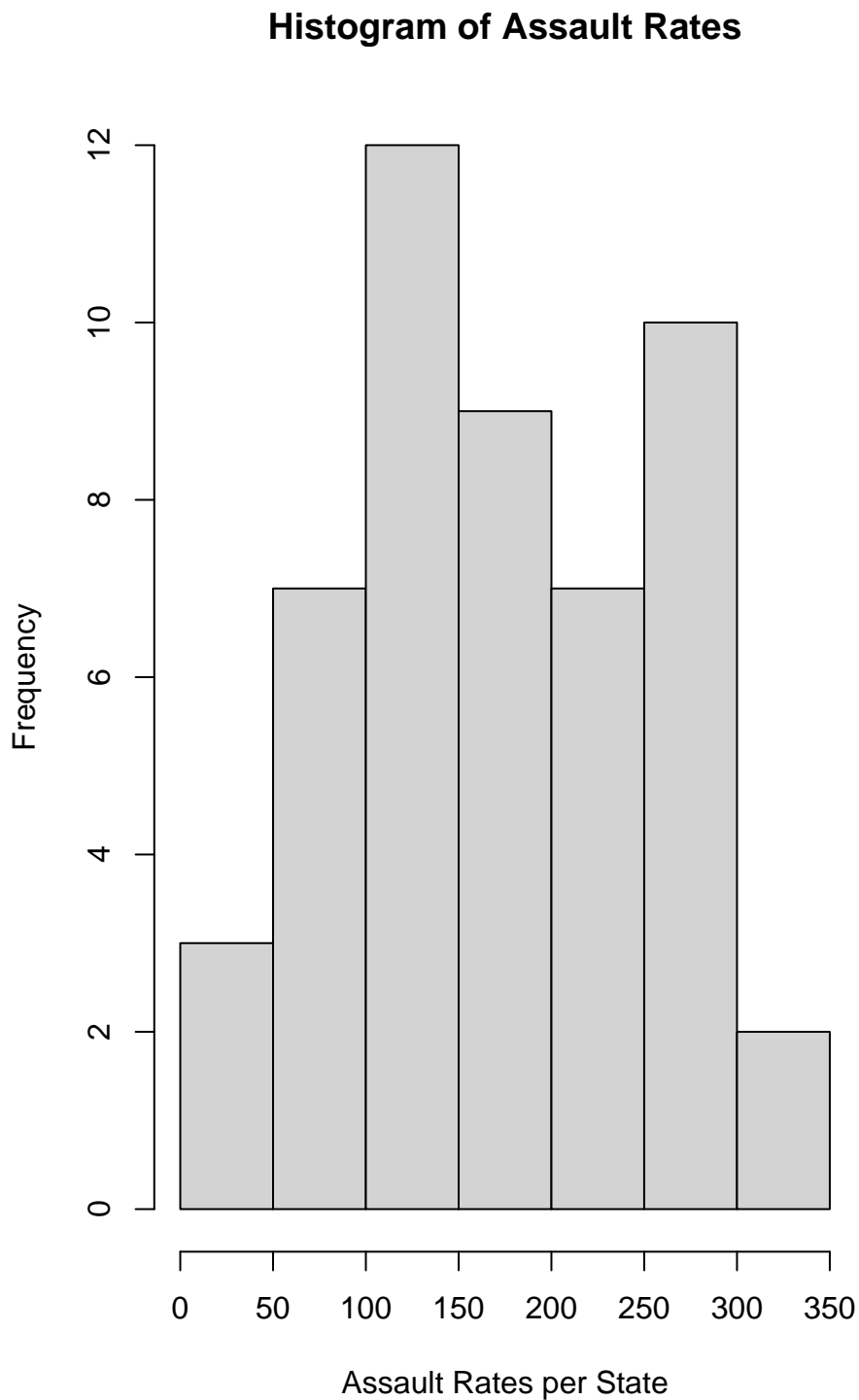
```
hist(dat.USArrests$Rape, main="Histogram of Rape Rates", xlab="Rape Rates per State", ylab="Frequency")
```

Histogram of Rape Rates
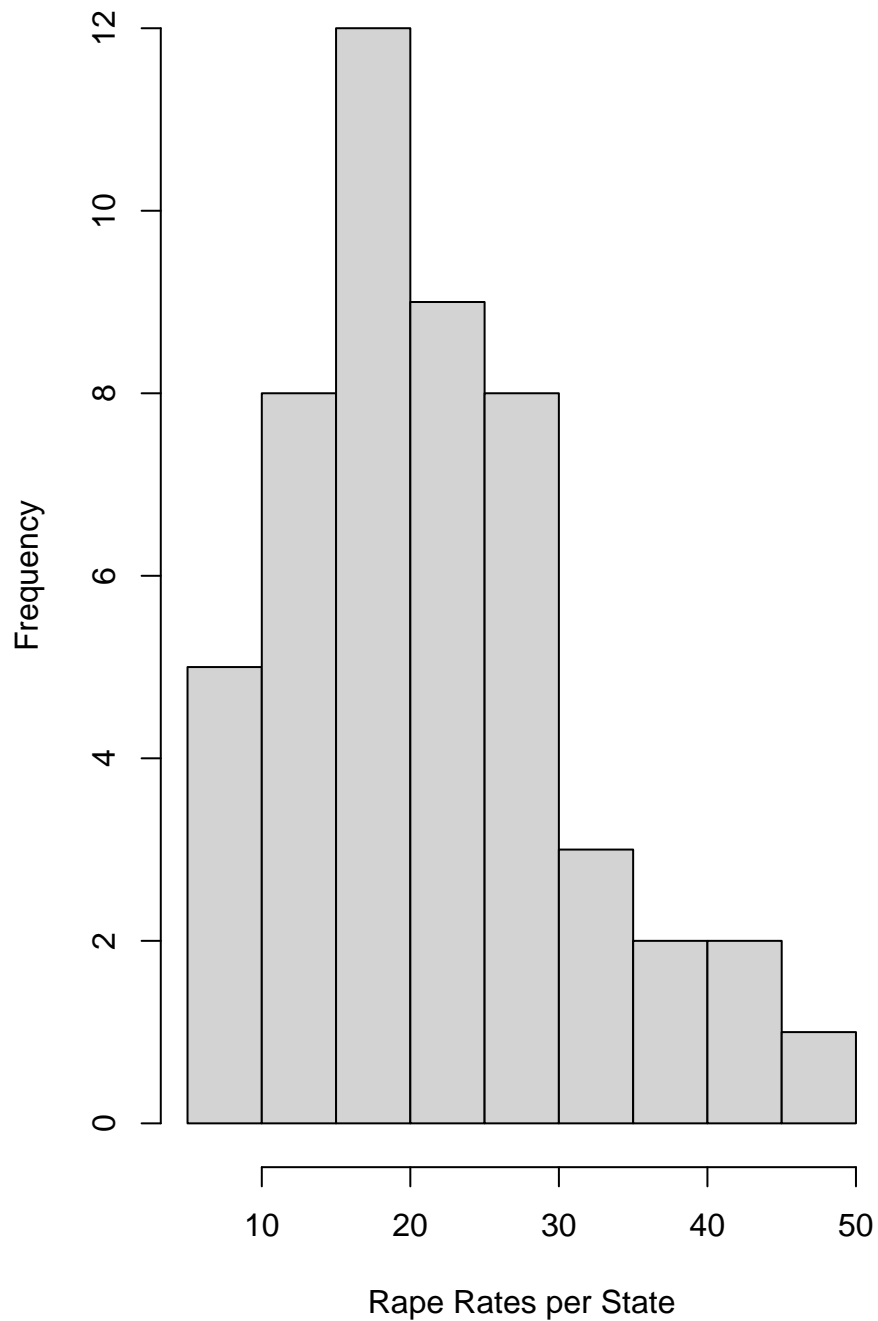

```
par(mfrow=c(3,1))
hist(dat.USArrests$Murder, main="Histogram of Murder Rates", xlab="Murder Rates per State", ylab="Freque
hist(dat.USArrests$Assault, main="Histogram of Assault Rates", xlab="Assault Rates per State", ylab="Fr
hist(dat.USArrests$Rape, main="Histogram of Rape Rates", xlab="Rape Rates per State", ylab="Frequency")
```

Histogram of Murder Rates

Histogram of Assault Rates

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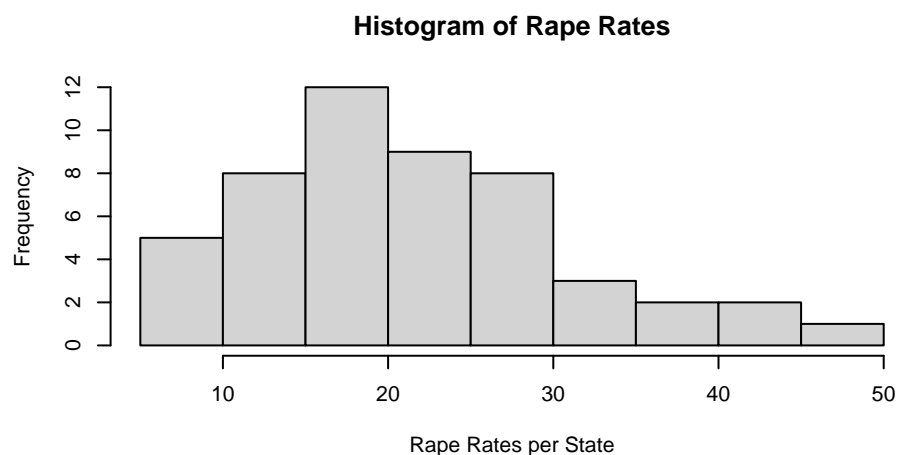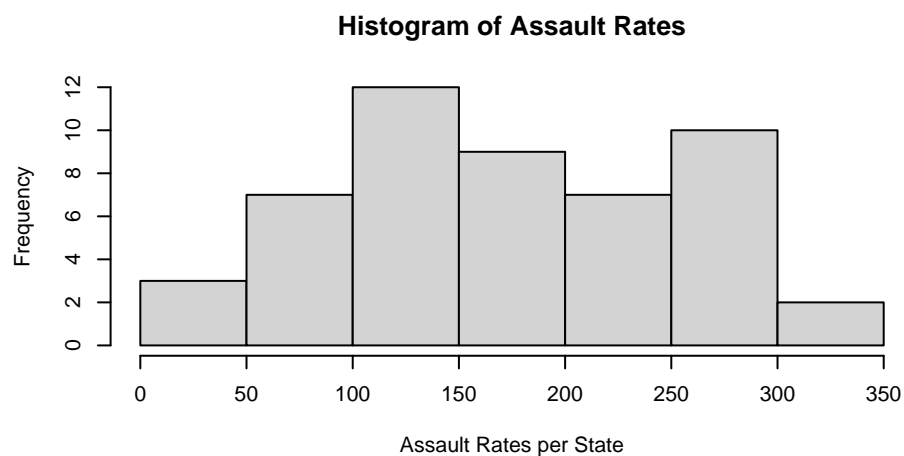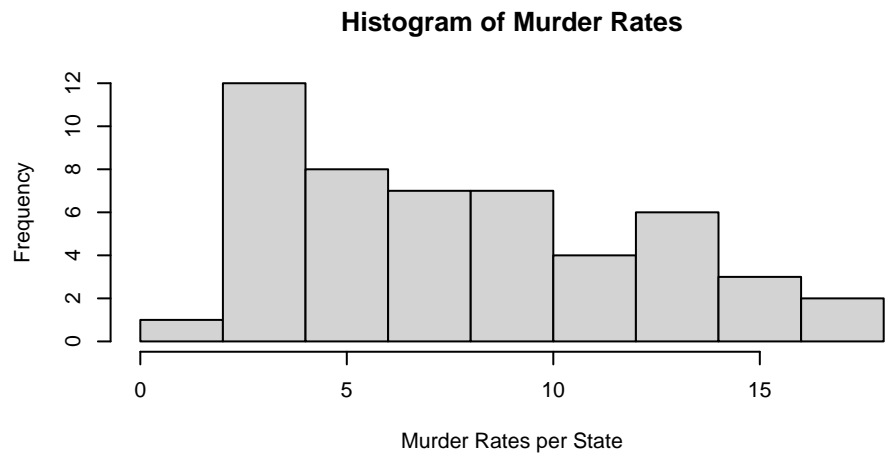
What does the command par do, in your own words (you can look this up by asking R ?par)?

Answer: The command par enables the statistician to set graphical parameters for data in either a singular graph or multiple graphs.

What can you learn from plotting the histograms together?

Answer: By plotting histograms together, you are able to compare the data between different categories – in this case, comparing the differences in assault and murder rates per state, for example. Additionally, you can

gain a better understand of the data overall by looking at it holistically instead of piece-by-piece.

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:

```r
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: This code is mapping the arrest rates of murder per 100,000 citizens per state. With this, we are able to see the salience and prominence of arrest rates through a colored map of the United States. The first line is using the data groups of "state" and "Murder" to construct aesthetic mapping in a ggplot, filling the map with "Murder" rates. Next, the second line is the direction to map the states, while the third line is expanding the x and y axes, i.e. longitude and latitude, in the graph.

Assignment 2

Problem 1: Load Data

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

```
##   mjage cigage iralcage age2 sexatract speakengl irsex
## 1    14     50       14   16         1         1     1
## 2    11     14        5   13         2         1     2
## 3    12     35       12   15         2         1     2
## 4    16     18       18   14         1         1     1
## 5    14     16       14   16         4         1     1
## 6    12     16       18   15         4         1     2
```

What are the dimensions of the data set?

```r
names(dat)
```

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

Problem 2: Variables

Describe the variables in the dataset

The variables are forms of quantitative variables, and all the datasets are described as integers within R. Here, these variables are labeled as mjage (age of first use of marijuana), cigage (age of first daily use of cigarettes), iralcage (age that first tried alcohol), age2 (age recorded the second time), sexatract (sexual attraction/action towards different genders), speakengl (proficiency in English), and irsex (sex of individual).

```r
class(dat$mjage)
```

```
## [1] "integer"
```

```
class(dat$cigage)
```

```
## [1] "integer"
```
```
class(dat$iralcage)
```

```
## [1] "integer"
```
```
class(dat$age2)
```

```
## [1] "integer"
```
```
class(dat$sexatract)
```

```
## [1] "integer"
```
```
class(dat$speakengl)
```

```
## [1] "integer"
```
```
class(dat$irsex)
```

```
## [1] "integer"
```

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 data describe a respondent's first use of a variety of substances, including alcohol, marijuana, and cigarettes, as well as describing the age, language, sex, and sexual orientation of the respondents. The data presented in this set are a small sample from the entirety of data collected by the National Survey of Drug Use and Health. Here, only the first 1,000 responses are used out of the entire representative sample taken by the NSDUH. By generating this data, the NSDUH is able to greater understand the use of particular substances around the country, as well as the sexual tendencies of individuals, based on age, gender, and language proficiency.

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="Age of Respondents", xlab = "Code for Age Group", ylab = "Frequency")
```

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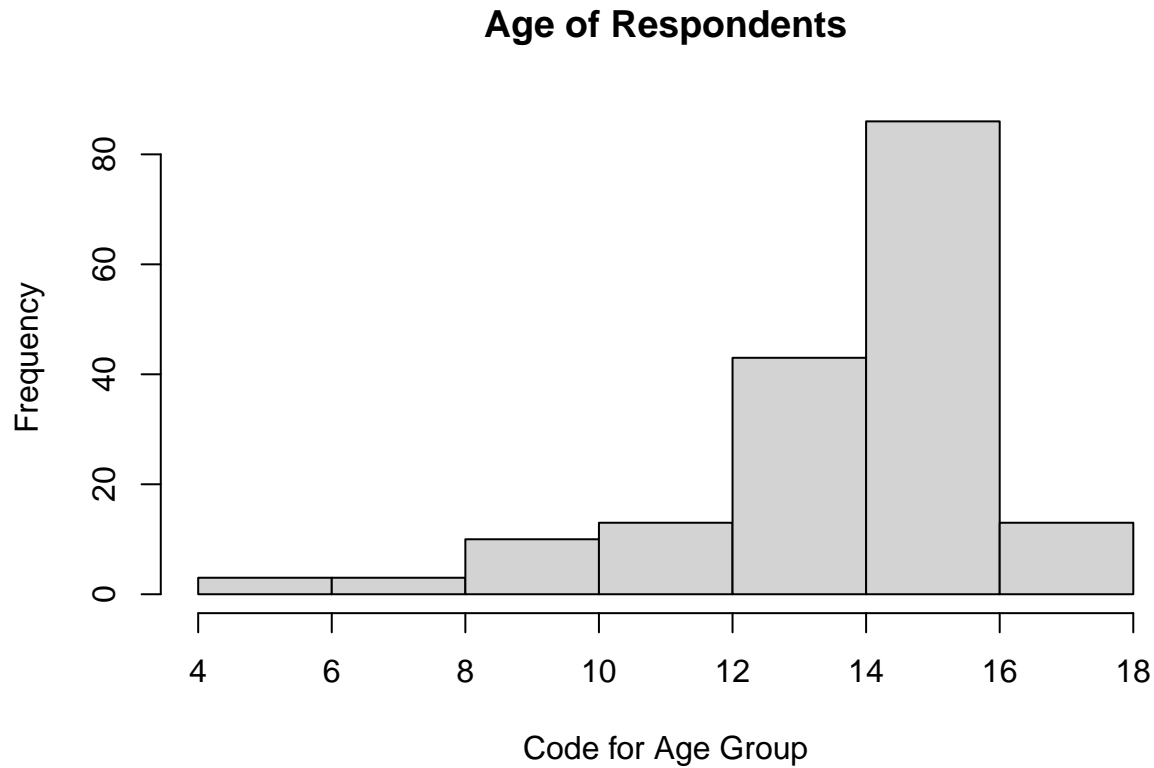
Do you think this age distribution representative of the US population? Why or why not?

I do not believe this distribution is representative of the US population, because, more than anything, the way the ages are distributed in these coded groups are completely skewed, with codes 1-12 solely for individuals at or under 25 years of age, while the codes 12-17 are responsible for the rest of the population's ages. Here, nearly 2/3 of the codes account for solely over 1/3 of the United States population (that under 25 years of age). In this data set, there are no patterns to age distribution, with some codes representing one age, others representing four, and others representing fourteen.

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

```
counts <- table(dat$irsex)
counts
```

```
##
##  1  2
## 91 80
```

```
barplot(counts, main="Gender Distribution", xlab="Gender", names=c("Male","Female"))
```

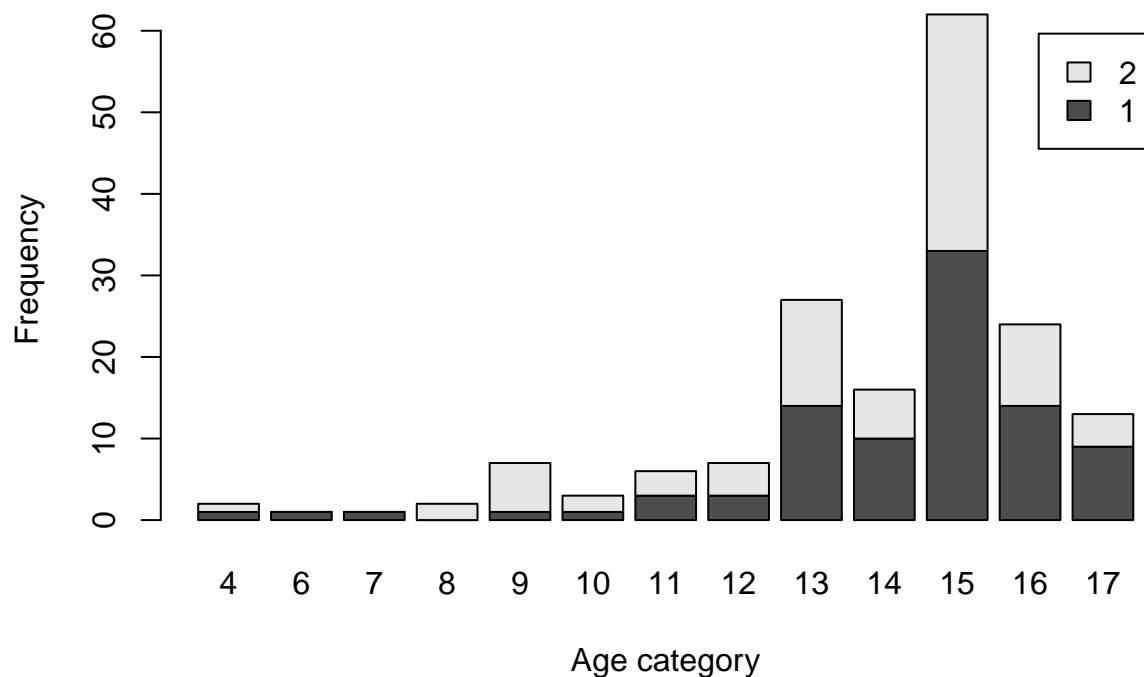**Gender Distribution**

Gender

The sample is nearly balanced by gender, yet based on this sample of the data, there are 11 more male respondents (totaling 91) than female respondents (totaling 80).

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 =
```

Stacked barchart

From this plot, it is shown that the majority of individuals, both male and female, are within the 15 age category, with gender being nearly evenly distributed in this area. Moving to other age categories, the outlier categories on both sides tend to be more male, while women are more centered along the 8-16 age categories.

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(mfrow=c(3,1))
hist(dat$mjage, main="Histogram of Marijuana Use", xlab="Age of First Marijuana Use", ylab="Frequency")
hist(dat$cigage, main="Histogram of Cigarette Use", xlab="Age of First Cigarette Use", ylab="Frequency")
hist(dat$iralcage, main="Histogram of Alcohol Use", xlab="Age of First Alcohol Use", ylab="Frequency")
```

Histogram of Marijuana Use

Histogram of Cigarette Use

Histogram of Alcohol Use
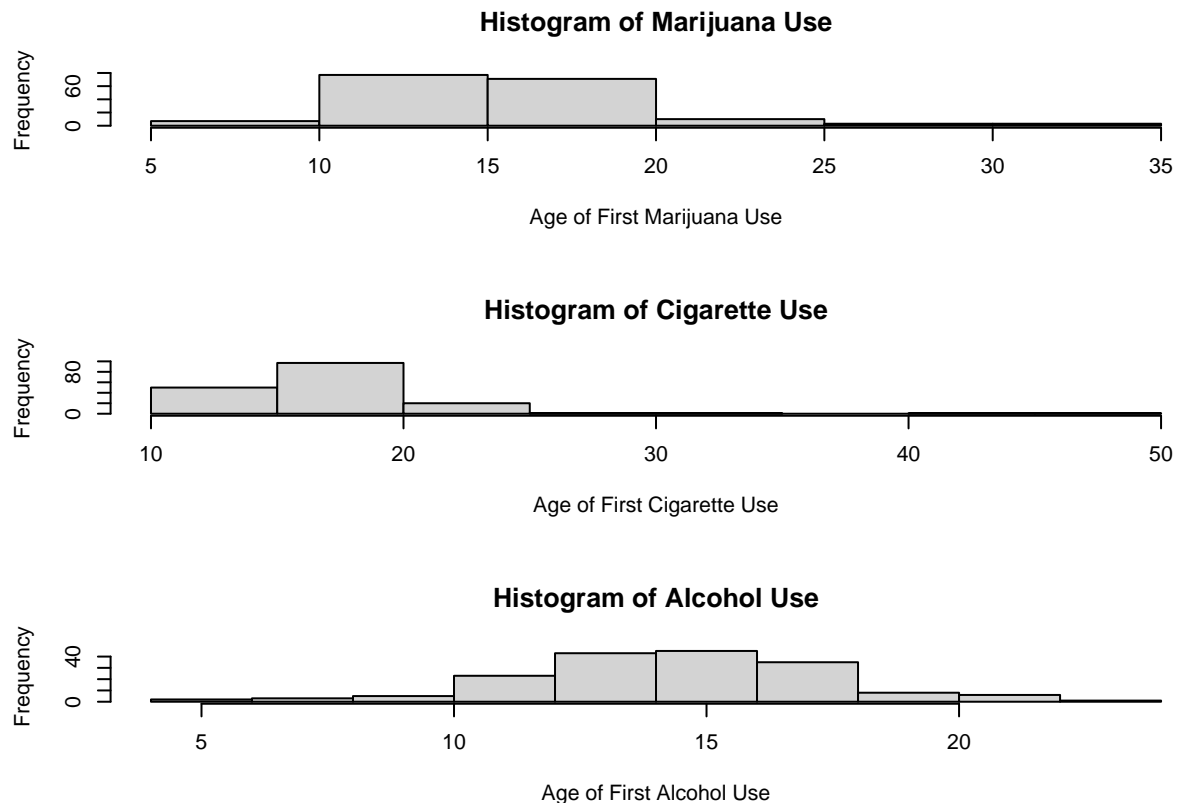
In this dataset, marijuana is tended to be used the youngest, with a large frequency of first use between 10-15 years of age in comparison to the entire histogram. In the other histograms, the dispersion of first use tends to be highest around the 15 or 15-20 age ranges.

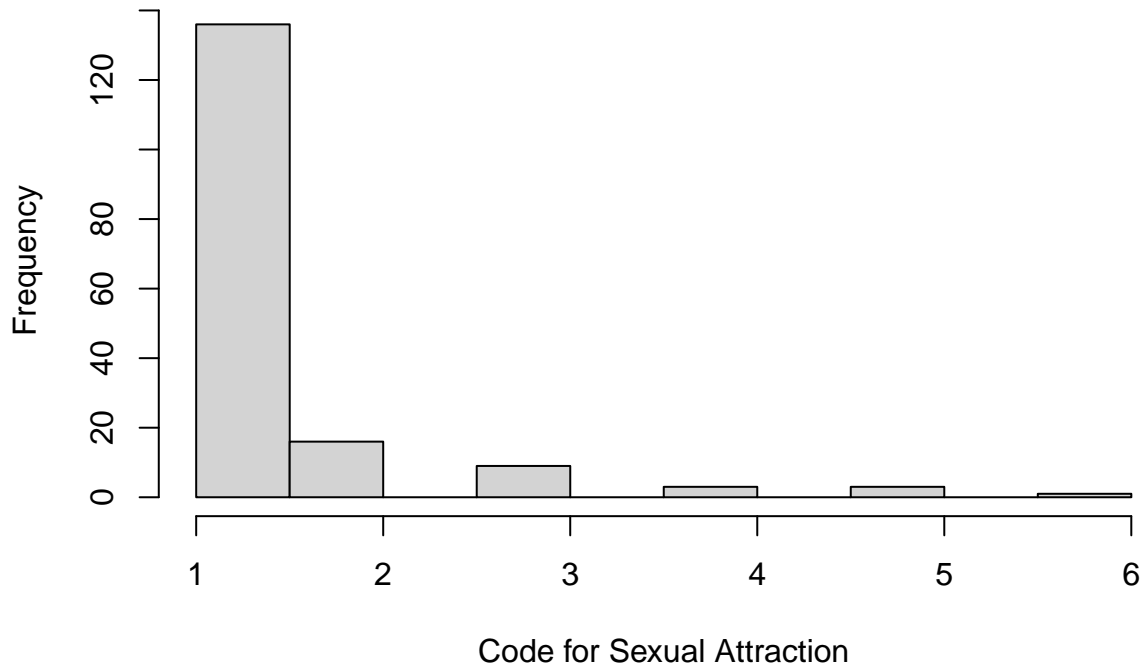### Problem 5: Sexual Attraction

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

```
as.numeric(dat$sexatract)
```

```
##    [1]  1  2  2  1  4  4  1  1  1  1  1  1  1  1  1  1  1  1  1  1  5  1  1  5  2
##   [26]  1  1  1  1 99  1  1  1  2 99  1  1  1  1  2  1  1  1  1  2  1  1  3  1  1
##   [51]  2  1  1  1  1  1  1  1  1  1  1  3  2  1  1  3  1  1  1  1  1  1  1  1  1
##   [76]  1  1  5  1  1  1  1  1  4  1  1  2  1  1  1  1  2  2  1  1  1  6  1  1  1
##  [101]  1  1  1  1  1  1  3  1  1  2  3  1  2  1  1  1  1  1  1  1  3  1  1  1  1
##  [126]  1  2  3  1  1  3  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##  [151]  1  1  1  1  1  1  1  1  1  1  2  1  1  2  1  1  1  1  3  1 99
```

```
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
dat$sexatract <- dat$sexatract %>% na_if(., "99")
hist(dat$sexatract, main="Sexual Attraction Histogram", xlab = "Code for Sexual Attraction", ylab = "Fre
```

13

Sexual Attraction Histogram

The distribution of sexual attraction is heavily skewed to opposite-sex attraction, under the code 1. As expected in surveying the majority of Americans, this plot is in line with a general representation of the population and sexuality.

*What is the distribution of sexual attraction by gender?*

```
as.numeric(dat$sexatract)
```

```
##   [1]   1  2  2  1  4  4  1  1  1  1  1  1  1  1  1  1  1  1  1  5  1  1  5  2
##  [26]   1  1  1  1 NA  1  1  1  2 NA  1  1  1  1  2  1  1  1  1  2  1  1  3  1  1
##  [51]   2  1  1  1  1  1  1  1  1  1  1  3  2  1  1  3  1  1  1  1  1  1  1  1  1
##  [76]   1  1  5  1  1  1  1  1  4  1  1  2  1  1  1  1  2  2  1  1  1  6  1  1  1
## [101]   1  1  1  1  1  1  3  1  1  2  3  1  2  1  1  1  1  1  1  1  3  1  1  1  1
## [126]   1  2  3  1  1  3  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## [151]   1  1  1  1  1  1  1  1  1  1  2  1  1  2  1  1  1  1  3  1 NA
```

```
tab.atractsex <- table(dat$irsex, dat$sexatract)
barplot(tab.atractsex, main = "Stacked Barchart of Sexual Attraction and Sex", xlab = "Sexual Attraction
```

## Stacked Barchart of Sexual Attraction and Sex
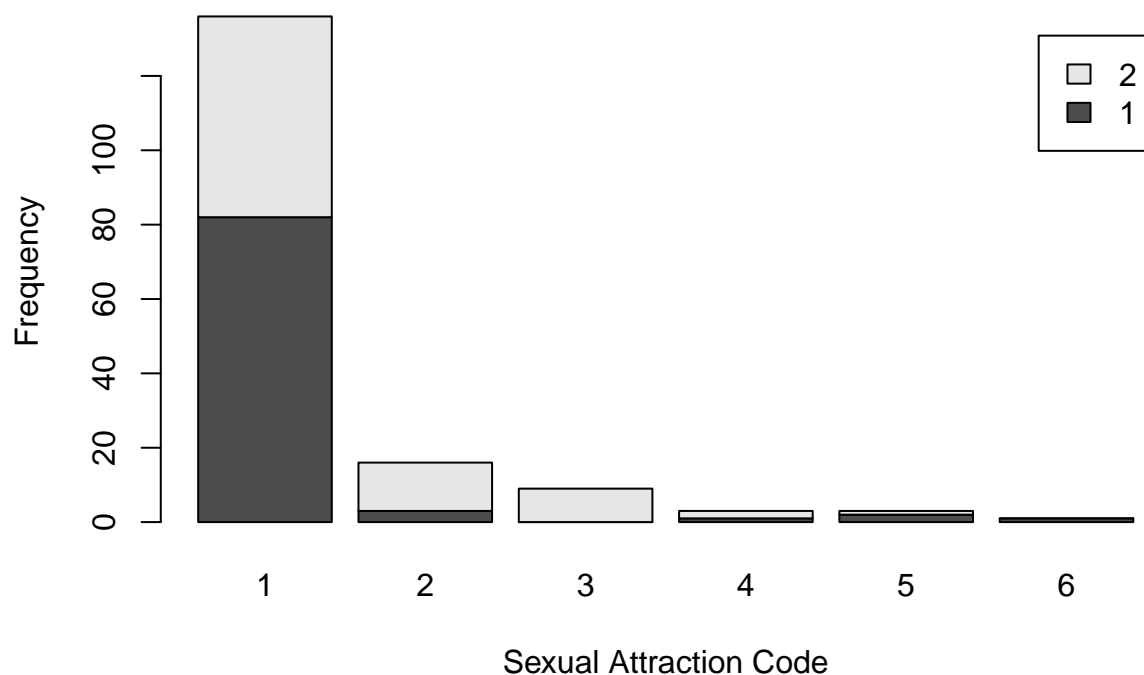
Based on gender, men are more likely to fall under the category 1 (opposite-sex attraction), with a small proportion being strictly homosexual. On the other hand, women tend to have more experiences in sexual fluidity, and although the majority falls under the category 1 as well, there is much more representation in the categories of 2, 3, and 4, signifying sexual experimentation and/or variety.

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?

```
tab.langsex <- table(dat$irsex, dat$speakengl)
barplot(tab.langsex, main = "Barchart of Sex and Language Ability", xlab = "Language Ability", ylab = "
```

Barchart of Sex and Language Ability
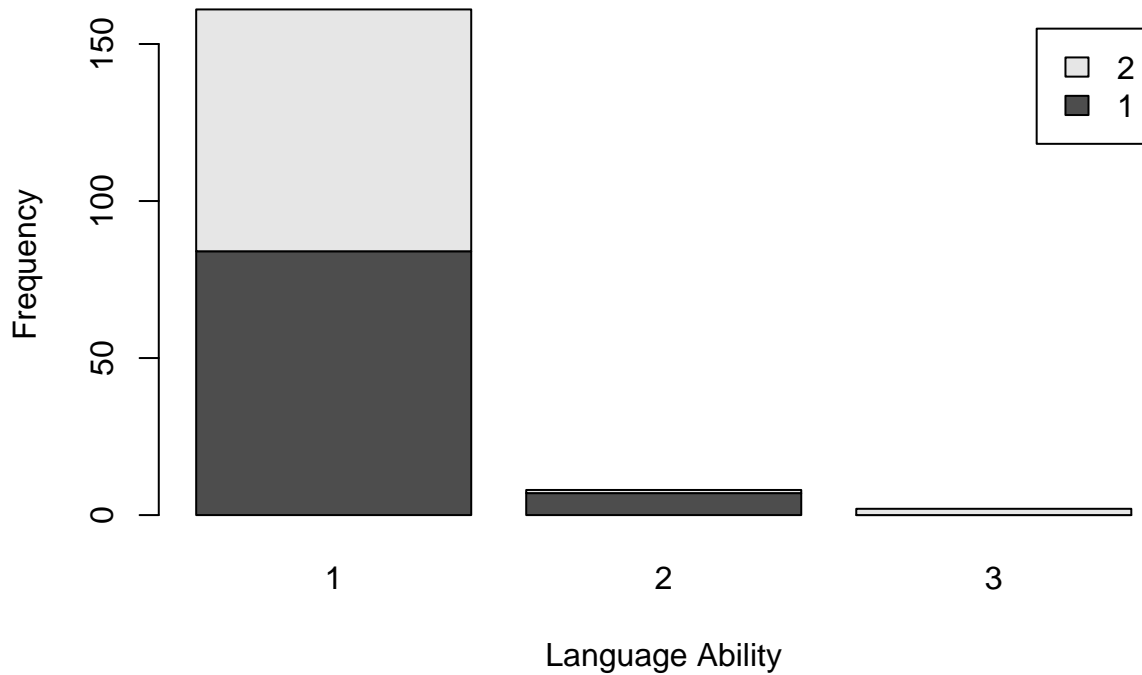
The distribution of English ability in this sample is heavily skewed towards 1, that of speaking English "very well," nearly equally for both males and females. However, based on the diversity of the US population, I believe a random sample should need to include less of the first code of perfect English and more of the codes 2, 3, and 4 in English ability

Are there more English speaker females or males?

```
tab.langsex
```

```
##
##      1  2  3
##   1 84  7  0
##   2 77  1  2
```

Here, there are a few more male English speakers than female English speakers.

Exam 1

Instructions

a. Create a folder in your computer (a good place would be under Crim 250, Exams).

b. Download the dataset from the Canvas website (fatal-police-shootings-data.csv) onto that folder, and save your Exam 1.Rmd file in the same folder.

c. Download the README.md file. This is the codebook.

d. Load the data into an R data frame.

Problem 1 (10 points)

a. Describe the dataset. This is the source: https://github.com/washingtonpost/data-police-shootings .Write two sentences (max.) about this.

The dataset describes every instance since 01/01/15 where an individual has been fatally shot by a police officer. With this, every individual is logged, along with the date of their death, the manner of death, if they were armed, their age, their gender, their race, their city and state, if there was any history of mental illness, the threat level, if they fled, if the officer had a body camera, and the details of distance during the shooting.

b. How many observations are there in the data frame?

There are 6,594 total observations within the data frame.

c. Look at the names of the variables in the data frame. Describe what "body_camera", "flee", and "armed" represent, according to the codebook. Again, only write one sentence (max) per variable.

Within the codebook, "body_camera" signifies that a news report stated that the officer may have had a body camera on, which could have a recording of the incident. Here, "flee" signifies if the individual was moving away from the officer, either on foot, in a car, or none of the above. Additionally, "armed" signifies that the victim was in possession of an item that may have been seen as harmful by the officer.

d. What are three weapons that you are surprised to find in the "armed" variable? Make a table of the values in "armed" to see the options.

Three items that I was very surprised to see that constituted as weapons by the officer were a flashlight, a beer bottle, and wasp spray.

Problem 2 (10 points)

a. Describe the age distribution of the sample. Is this what you would expect to see?

Usually, yes, one would expect for the majority of incidents to happen to younger individuals, somewhere in their 20's, as that reflects the age group of individuals that may be involved in crimes the most. However, I was surprised to see how many fatal shootings stretched into the later years, such as those in their late 30's and into their 40's and 60's.

b. To understand the center of the age distribution, would you use a mean or a median, and why? Find the one you picked.

I would use the median, because the grouping of ages is unsymmetric, and thus, the data might be skewed towards one end or the other if using the mean. Instead of the data being skewed through the mean (plus those unknown values skewing it more), the median would offer a better representation of the true center of the data.

c. Describe the gender distribution of the sample. Do you find this surprising?

The gender distribution of the data is heavily skewed towards male, which is mostly aligned with what would be predicted, as the majority of men are charged with committing crimes within the United States, and thus, men in criminal, or fatal, situations may also be aligned with this prediction. Although there is a small portion which seems to be of missing values, the barplot is still undeniably skewed towards male.

Problem 3 (10 points)

a. How many police officers had a body camera, according to news reports? What proportion is this of all the incidents in the data? Are you surprised that it is so high or low?

According to news reports, 910 police officers had a body camera on them. In relation to all of the incidents, though, it is surprising that only 13.8% of the officers who were involved in a fatal shooting had a body camera on. Based on the implementation of body cameras nationwide, one would have expected that percentage to be higher, yet it also begs the question of the correlation between higher rates of fatal shootings and no body cameras.

b. In how many of the incidents was the victim fleeing? What proportion is this of the total number of incidents in the data? Is this what you would expect?

Not including the "other" values, 1,903 times, or in about 28.85% of the instances, the victim was attempting to flee. Consequently, I would have expected more instances of fleeing because that then implies that while the victim was complying (at least in not fleeing) the police officer still founds grounds to fatally shoot the individual.

Problem 4 (10 points)

a. Describe the relationship between the variables "body camera" and "flee" using a stacked barplot. What can you conclude from this relationship?

(Hint 1: The categories along the x-axis are the options for "flee", each bar contains information about whether the police officer had a body camera (vertically), and the height along the y-axis shows the frequency of that category.)

(Hint 2: Also, if you are unsure about the syntax for barplot, run ?barplot in R and see some examples at the bottom of the documentation. This is usually a good way to look up the syntax of R code. You can also Google it.)

Here, fleeing is plotted on the x axis, while the frequency of fleeing is plotted on the y axis, while "true" and "false" on the stacked barplots signify if a body camera was used in that exact situation. With this relationship, one may conclude that out of all of the instances of fleeing or not fleeing, there was more body camera usage covering that an individual was not fleeing.

b. Describe the relationship between age and race by using a boxplot. What can you conclude from this relationship?

(Hint 1: The categories along the x-axis are the race categories and the height along the y-axis is age.)

(Hint 2: Also, if you are unsure about the syntax for boxplot, run ?boxplot in R and see some examples at the bottom of the documentation. This is usually a good way to look up the syntax of R code. You can also Google it.)

Plotting age and race both ways, the first graph demonstrates ages along the x axis and frequency along the y axis, with the key representing different races. The second graph demonstrates race along the x axis, frequency along the y axis, and ages within the key. Here, it is shown that the majority of fatal shooting victims are indeed white, with African Americans coming as the second highest grouping for victims, with the majority of both being within the age ranges of 26-36 years old.

Extra credit (10 points)

a. What does this code tell us?

The code explains how much of a time difference there was between all occurances of fatal shootings, with the difference being 2,458 days total.

b. On Friday, a new report was published that was described as follows by The Guardian: "More than half of US police killings are mislabelled or not reported, study finds." Without reading this article now (due to limited time), why do you think police killings might be mislabelled or underreported?

Police killings might be mislabeled or underreported first due to processing issues, where an officer might not process the report properly, the report may not have all necessary information, or different processors may not have investigated the event thoroughly. Second, officers who may be liable for a wrongly death may not report the incident, or even do so correctly. Third, with a wrongly death, an officer may report that the victim had a weapon in his hand, or that he was fleeing, when he truly was not, in order to grant the officer clemency for the situation.

c. Regarding missing values in problem 4, do you see any? If so, do you think that's all that's missing from the data?

In section a of problem 4, there are many missing values within the "fleeing" variable, off to the left side of the graph, where there is no data if the victim fled or not. There is simply a bar determining if there

was a body camera during the situation or not. Additionally, the "other" portion is unspecified as to what "other" may be, thus causing more vagueness in the results of the data. With this data, I believe not only are there missing values to it, but the vagueness of the "other" section leads to more missing values and less understanding of the situation based on if the victim was fleeing or not (and how).