Assignment 2

Niko Amber Assignments!

Assignment 1

Collaborators: Eliza Epstein.

Problem 1

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

Answer: I've loaded the library!

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: It is beneficial to rename the data set, so we can replicate analyses without disturbing the original data set. Additionally, it is nice to rename your data set to know exactly it is called

Problem 2

First I am making states lowercarse, to be used as variables later.

```
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`.

The variables contained in the dataset 'USArrests' are Murder, Assault, and Rape. Additionally, the data set shows us what percentage of people live in urban areas.

```
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
```

Problem 3

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

Answer: Murder is a categorical variable, it is one of many catagories that crime falls into. It is not ordinal, there is no ordering of crime.

What R Type of variable is it?

Answer: 'Murder' is a character variable, it contains information that isn't numeric.
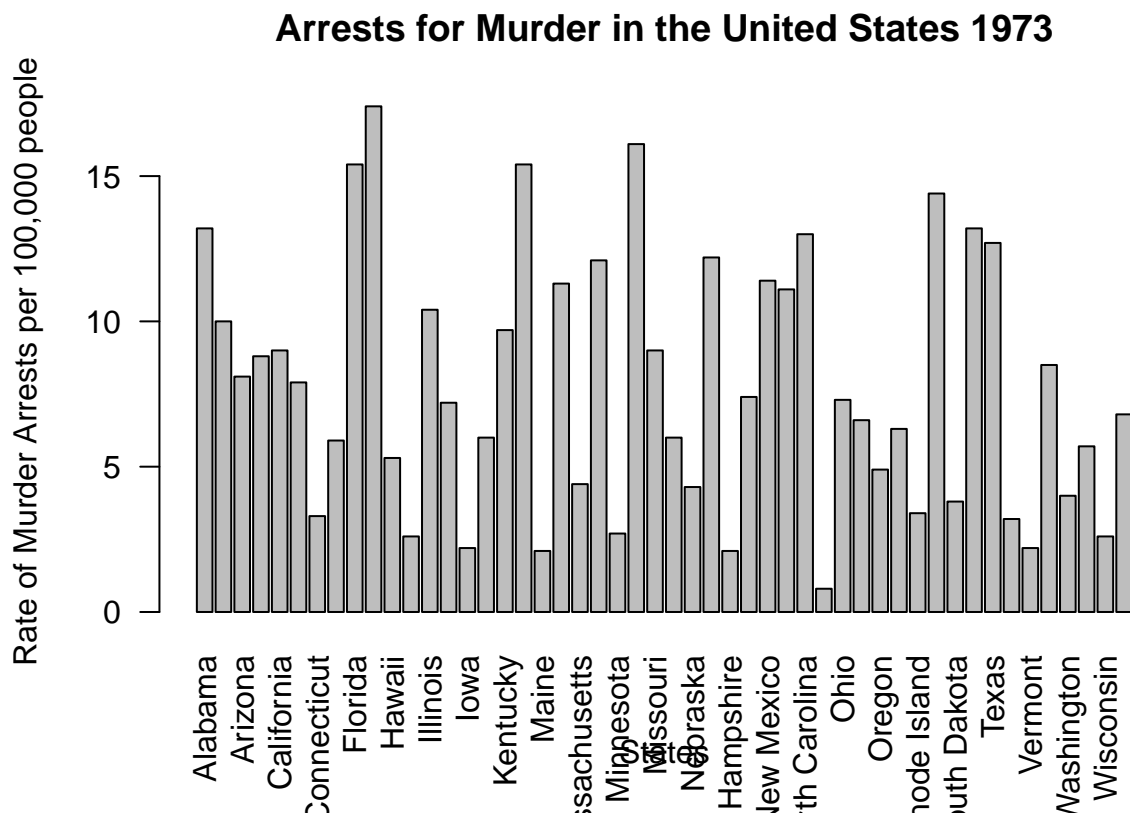
Problem 4

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

Answer: The Data set USArrests contains data about the rate of arrests for murder, rape and assault per 100,000 residents in each US state in 1973. The data set also includes the percent of the population living in urban cities in each state. The collums represent each type or crime (and urban population percentage) The numbers in each row are the rate of arrests per state (per 100,000).

Problem 5

Draw a (histogram) bar graph of `Murder` with proper labels and title. I used a bar graph instead of a histogram because I feel that is a better way to represent this data.

```
barplot(USArrests$Murder , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Murder Arrest
        main = "Arrests for Murder in the United States 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.?

I used the summary function to gather this information

3

```
summary(USArrests$Murder)
```

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

The mean for murder is 7.788 meaning this is the average rate of people (per 100,000) murdered in each US state in 1973. The median for murder is 7.250 meaning in the United States in 1973 half of the states had a rate per 100,000 more that 7.250 and half had fewer. Median is the middle of numbers in a data set, while mean is the average of all numbers in said set. If the data is evenly distributed the median will equal the mean.

A quartile is when the data is divided into four equal parts: the 1st, 2nd, 3rd, and 4th quartile. R gives the 1st and 3rd quartile because it represents the middle of the data. 1st quartile is lowest 25% and 3rd quartile is highest 25%.

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.

Note: I used bar graphs

```
barplot(USArrests$Murder , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Murder Arrests
        main = "Arrests for Murder in the United States 1973")
```
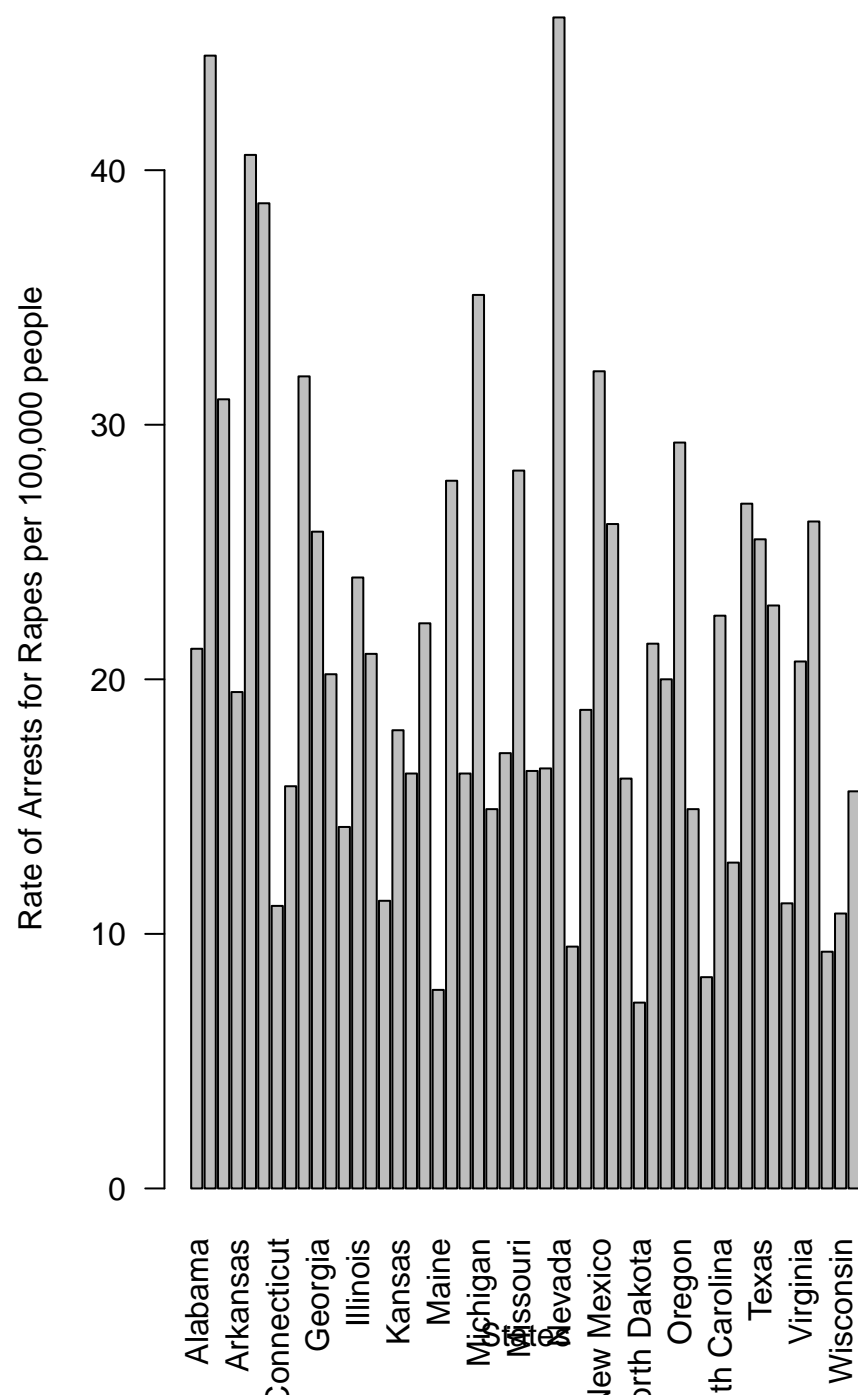
Arrests for Murder in the United States 1973
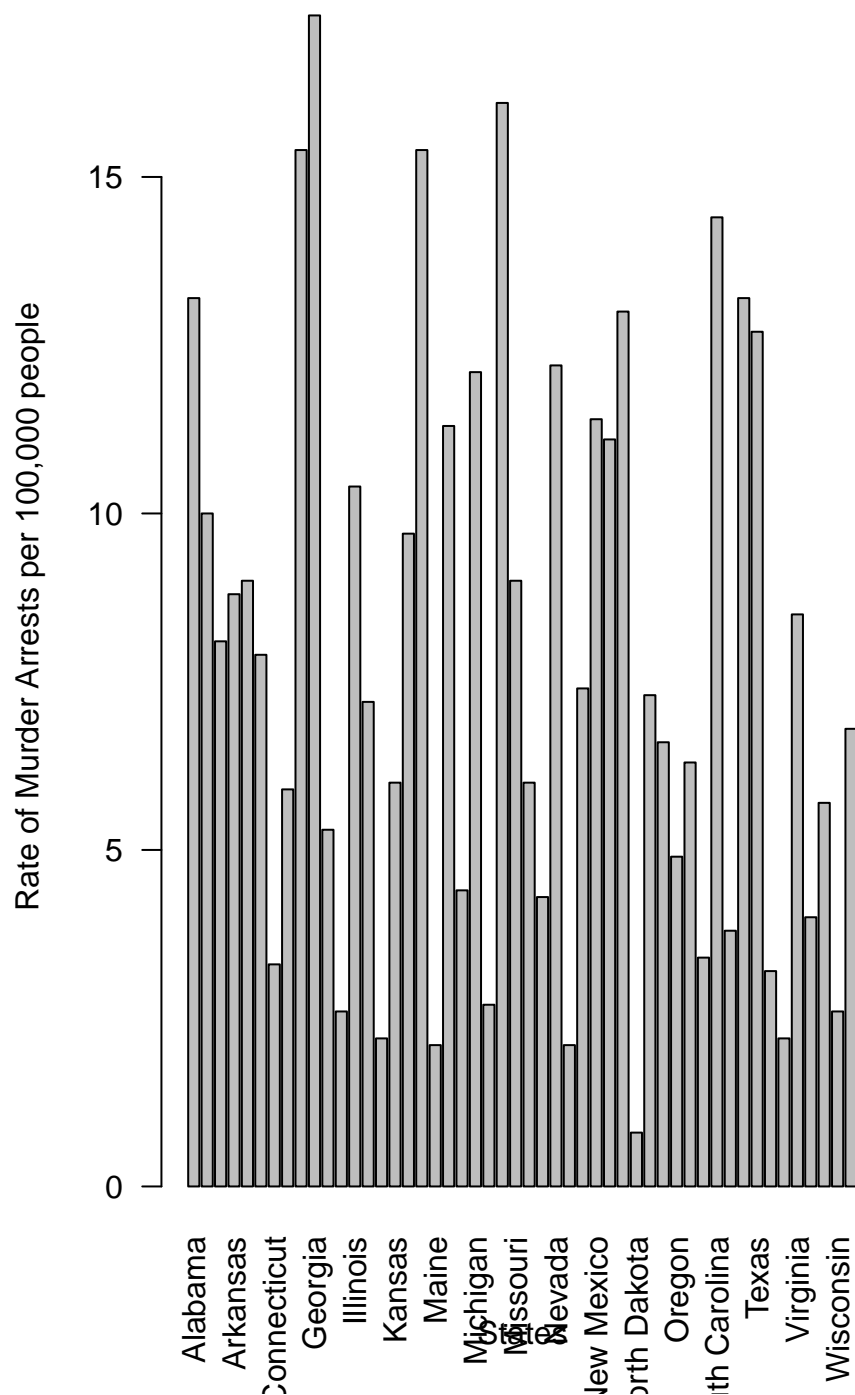

```
#Bar plot for rape arrests
barplot(USArrests$Rape , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for Rap
        main = "Rate of Arrests for Rape in the United States in 1973")
```
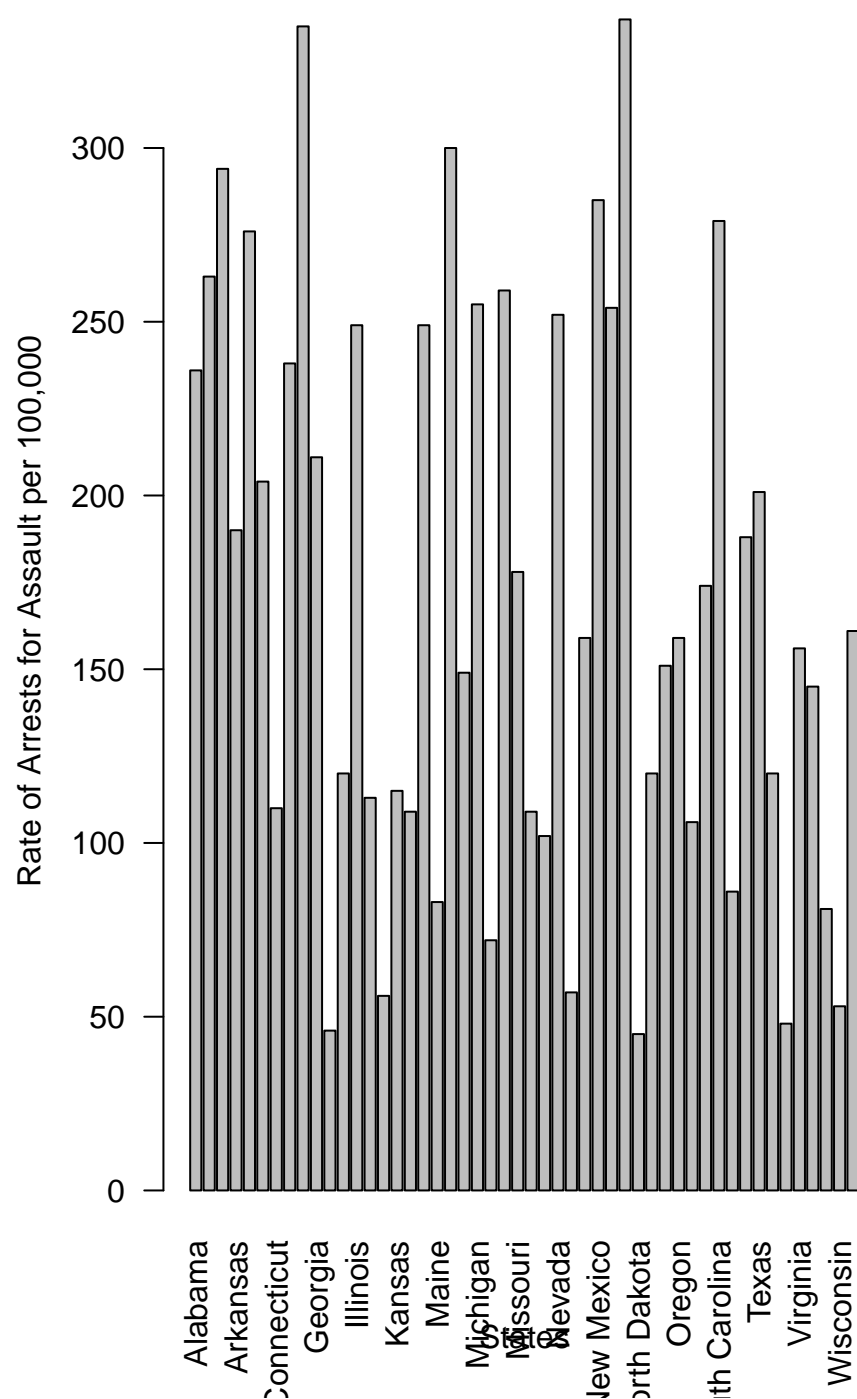
Rate of Arrests for Rape in the United States in 197


```
#Bar plot for assault arrests
barplot(USArrests$Assault , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for A
        main = "Rate of Arrests for Assaults in the United States")
```

Rate of Arrests for Assaults in the United States


```
#plotting the graphs together

par(mfrow=c(3,1))
barplot(USArrests$Murder , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Murder Arrests
        main = "Arrests for Murder in the United States 1973")
```

```r
#Bar plot for rape arrests
barplot(USArrests$Rape , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for Rap
        main = "Rate of Arrests for Rape in the United States in 1973")

#Bar plot for assault arrests
barplot(USArrests$Assault , names.arg = state.name,las=2, xlab = "States", ylab = "Rate of Arrests for A
        main = "Rate of Arrests for Assaults in the United States")
```

Arrests for Murder in the United States 1973

Rate of Arrests for Rape in the United States in 1973

Rate of Arrests for Assaults in the United States
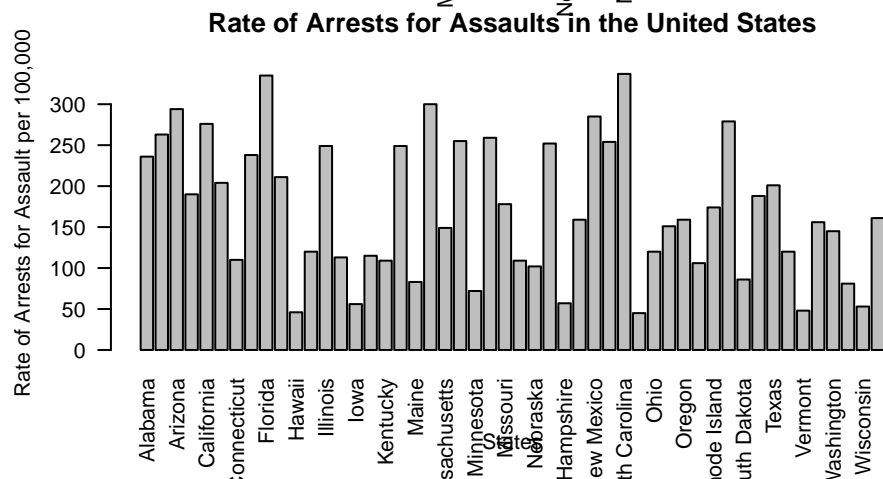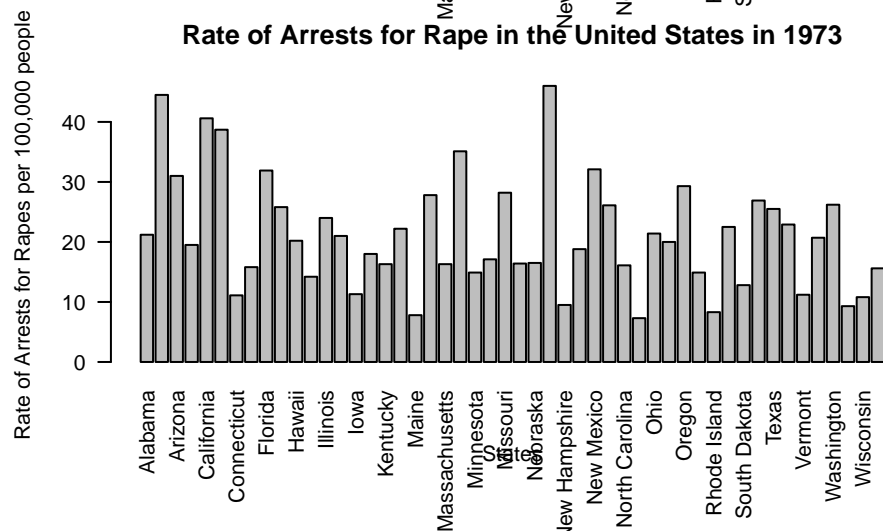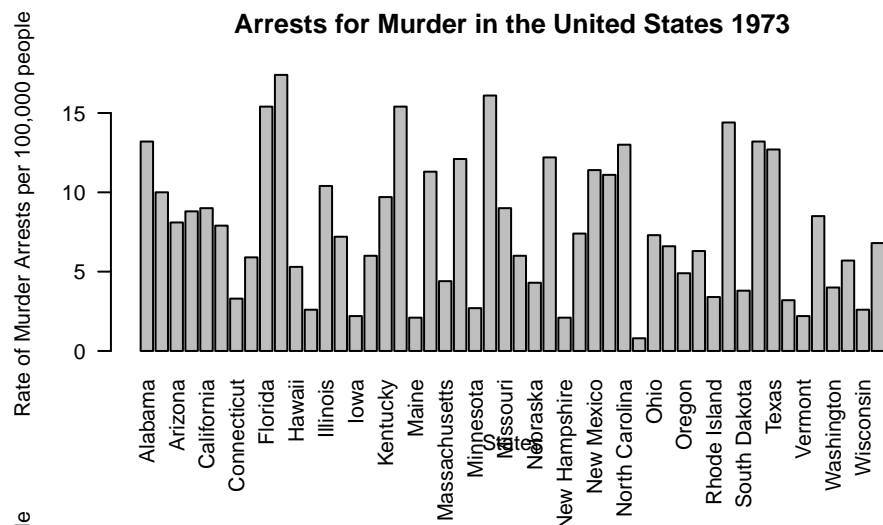
```
par
```

```
## function (..., no.readonly = FALSE)
## {
##     .Pars.readonly <- c("cin", "cra", "csi", "cxy", "din", "page")
```

```
##     single <- FALSE
##     args <- list(...)
##     if (!length(args))
##         args <- as.list(if (no.readonly)
##             .Pars[-match(.Pars.readonly, .Pars)]
##         else .Pars)
##     else {
##         if (all(unlist(lapply(args, is.character))))
##             args <- as.list(unlist(args))
##         if (length(args) == 1) {
##             if (is.list(args[[1L]]) || is.null(args[[1L]]))
##                 args <- args[[1L]]
##             else if (is.null(names(args)))
##                 single <- TRUE
##         }
##     }
##     value <- .External2(C_par, args)
##     if (single)
##         value <- value[[1L]]
##     if (!is.null(names(args)))
##         invisible(value)
##     else value
## }
## <bytecode: 0x7fd079af1e80>
## <environment: namespace:graphics>
```

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

Answer: This command allows R to set paramaters, this way multiple data sets can be graphed together.

What can you learn from bar graphs the histograms together?

Answer: When we plot these bar graphs together we can compare each state's arrest rates for different crimes. For example, when looking at the bar graphs it is easy to see that North Carolina's arrest rate for assault is much higher than the arrest rate for rape. This can lead researchers to ask questions: why were there more arrests for assaults? Were there possible sexually based assault that should have been charged as rape?


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:

```
library('maps')
library('ggplot2')

#this code creates a 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)
```
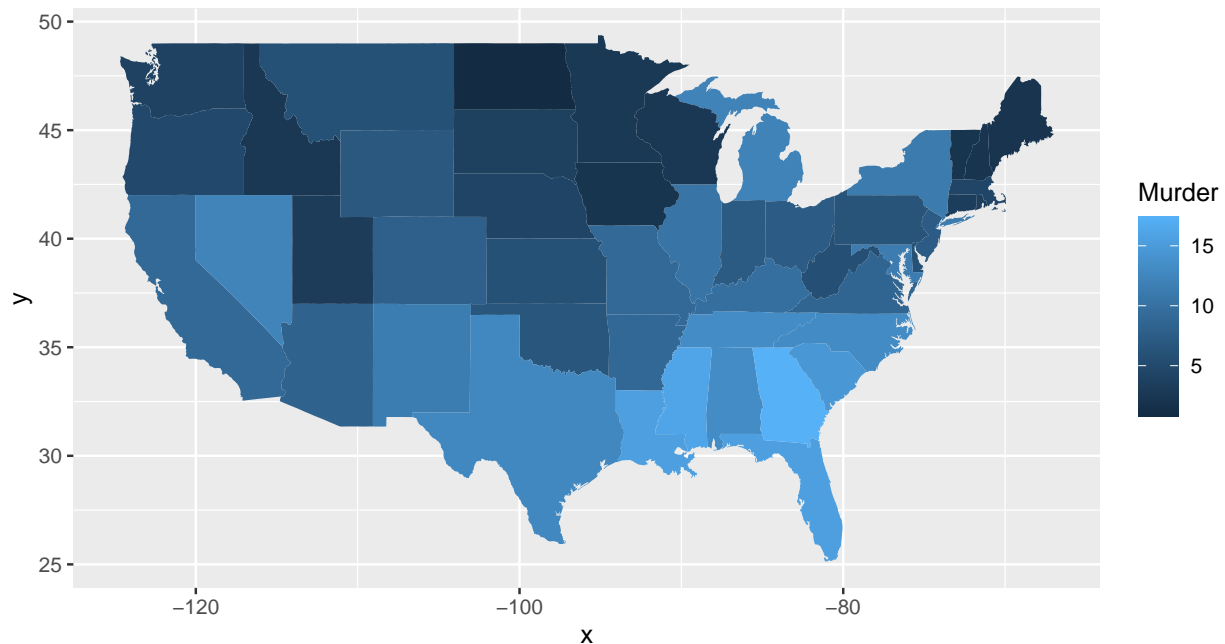
What does this code do? Explain what each line is doing.

Answer: This code creates a colored map that uses our data set to show murder rates. The lighter the blue, the higher the murder rate.

The first line of code creates a map that is divided by states. TYhe first line also applies the data set to the map, so the color of the state will be lighter if murder rates are higher. The second line of code designs the map by breaking it up by state. The last line of code expands the limits of the map

Here is my histogram

```
hist(dat$Murder, main = "Murder Rate in The United States in 1973", xlab = "Murder Rate", ylab = "Frequ
```

Murder Rate in The United States in 1973
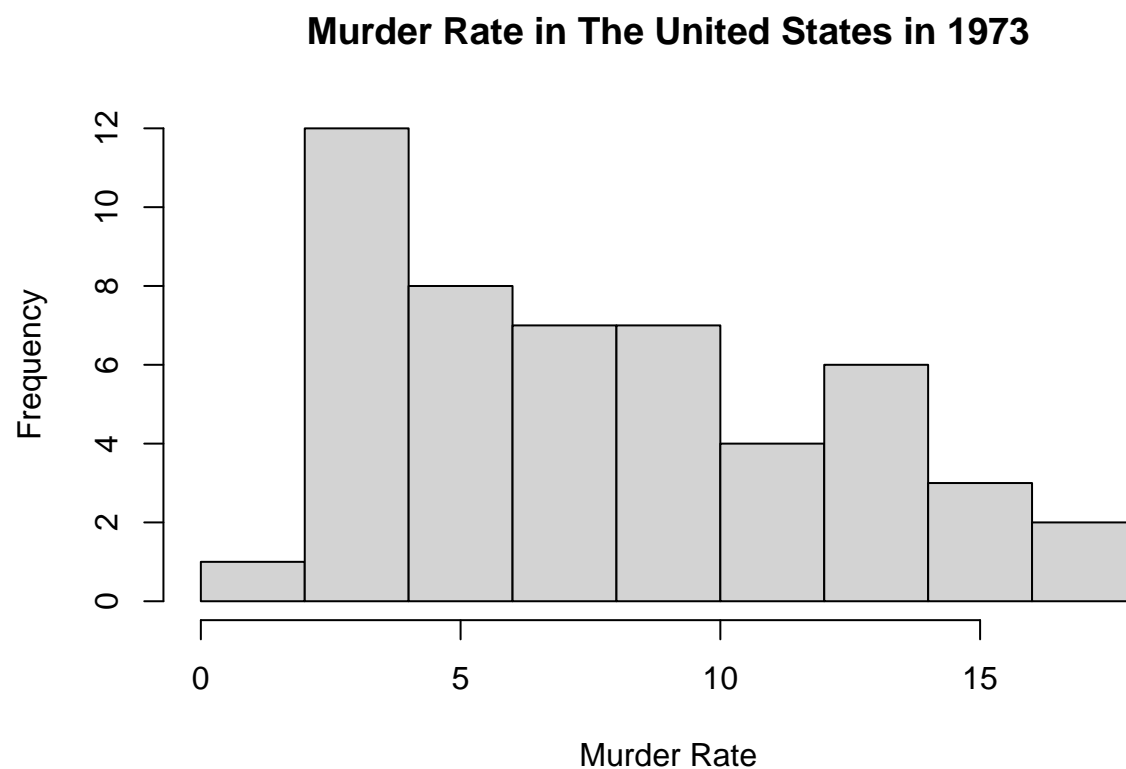
Assignment 2

Assignment 2

Problem 1

```
knitr::opts_chunk$set(echo = TRUE)
```

```
setwd("/Users/nikoamber/Library/Mobile Documents/com~apple~CloudDocs/Penn/Freshman Fall/Criminology/Ass:

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

names(dat)
```

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

```
nrow(dat)
```

```
## [1] 171
```

```
ncol(dat)
```

```
## [1] 7
```

There are seven columns and 171 rows

Problem 2

Describe the variables in the dataset.

The variables in the data are:

MJAGE which is how old the participants are the first time they used marijuana/hashish. It is a quantitative continuous variable.

CIGAGE which is how old the participants were when they first started smoking cigarettes everyday. It is a quantitative continuous variable.

IRALCAGE which is how old the participants were when they first tried alcohol. it is a quantitative continuous variable.

AGE2S which is the age of the respondents of the survey, categorized into groups. It is a categorical nominal variable.

IRSEX which is which sex participants identify as. It is a categorical, nominal variable. Participants can only identify as male or female. which is coded as 1 or 2.

SEXATRACT which is who participants of the study are attracted to. It is a categorical, nominal variable. Participants identified their sexual preference on a catagory scale of 1-6, but there is no ranking.

SPEAKENGL which is how well do participants speak English. It is a categorical, ordinal variable. Participants rate on a scale of 1-4 where 1 is speaking English well and 4 is not at all.

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 data set is a sample form a 2019 survey taken by the National Survey of Drug Use and Health. This data is only a sample of the survey and has the first 1000 values, and does not include missing values. The data is a random sampling of the US population, and it can be used to tell us more about the US population and their relationship with drugs.

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.
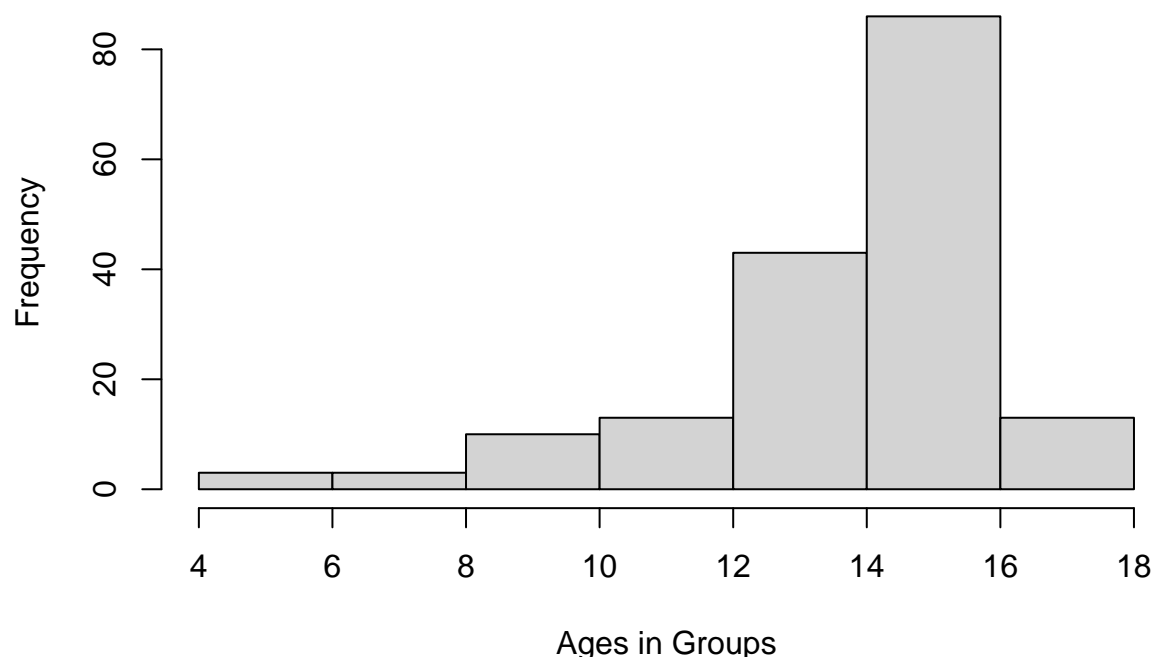
```
hist(dat$age2, main= "Histogram of ages of participants of 2019 NSDUH survey", xlab= "Ages in Groups",
```

Histogram of ages of participants of 2019 NSDUH survey

Ages in Groups

Histogram of ages of participants of 2019 NSDUH survey

The age distribution is from ages 12-65+. As mentioned in the code book, it is grouped in sections. but not every group holds the same number of years. Because of this it is difficult to see the age distribution properly. While it appears that most participants are in group 15, group 15 has ages 35-49 (14 years) while groups 1-9 only have one year.

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

Because of the way the ages are grouped together, this histogram is left skewed. However, there are no respondents below the age of 12, which makes sense given the content of this survey. There are also very few older participants. Therefore this data is probably a good representation of the US population for the purposes of this study, but obviously is not representative of the entire US population.

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

```
table(dat$irsex)
```

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

There are more males than females in this sample. It is not representative of the US population, because in the population there are slightly more women then men. This scale is also on a bianary, when there are many people in the United States who may not identify with one of these two genders. According to the code book, the entire data set has more females then males, but in these 1000 data points, there are more males then 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)
```

Stacked barchart
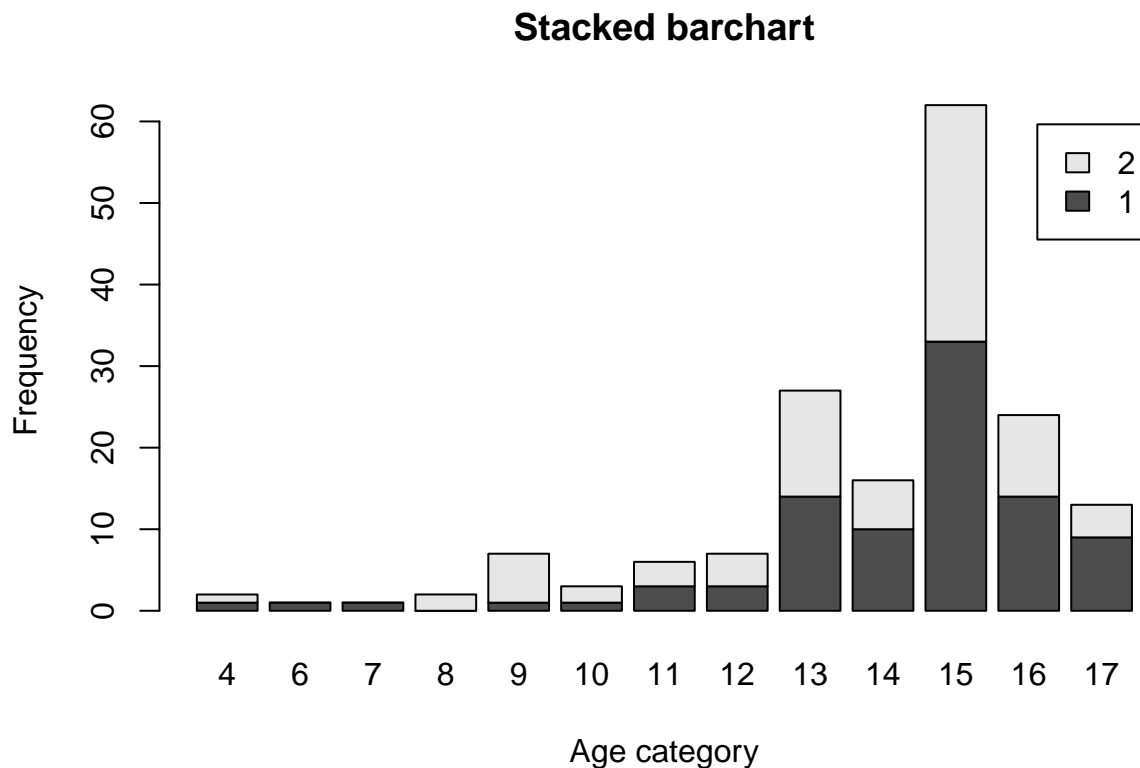
In the age category 15, that represents respondents between 35 and 49 and 26-29, there is about an even distribution between men and women. However, the younger groups tend to have more female participants then male (group 8 does not have any male respondents) and the groups above 15 seerm to have a larger percentage of men.

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 (1,1))
plot(dat$age2, dat$mjage)
```
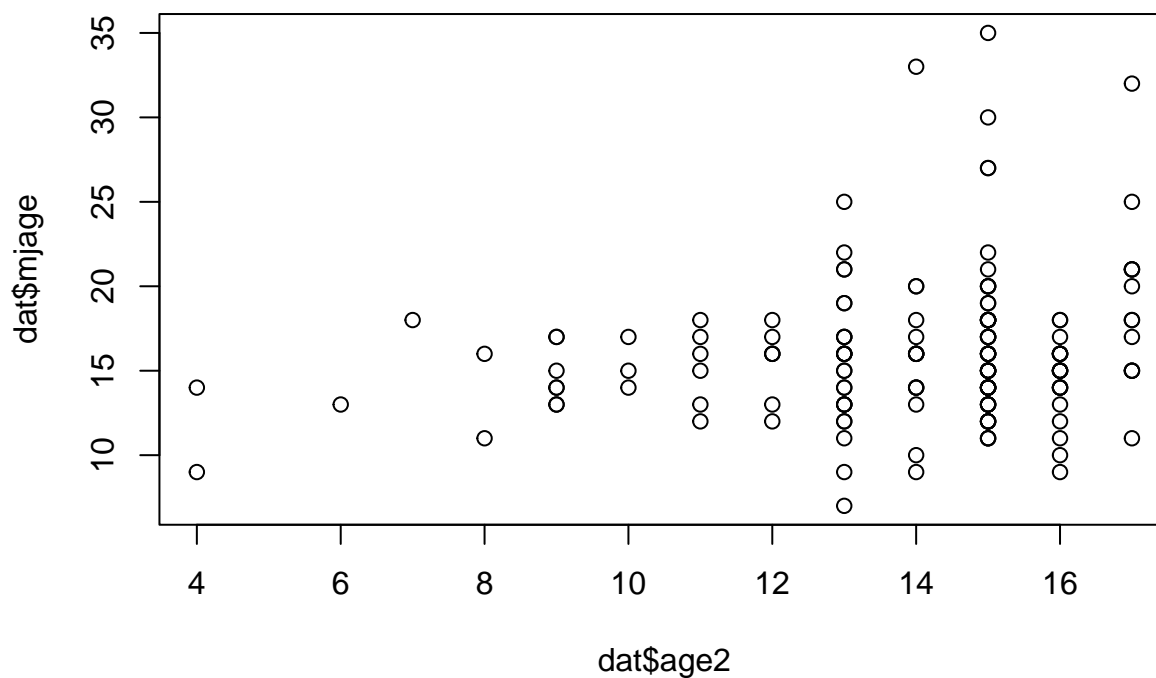


```
cor(dat$age2, dat$mjage)
```

```
## [1] 0.1811713
```

```
cor(dat$age2, dat$cigage)
```

```
## [1] 0.1697763
```

```
plot(dat$age2, dat$cigage)
```
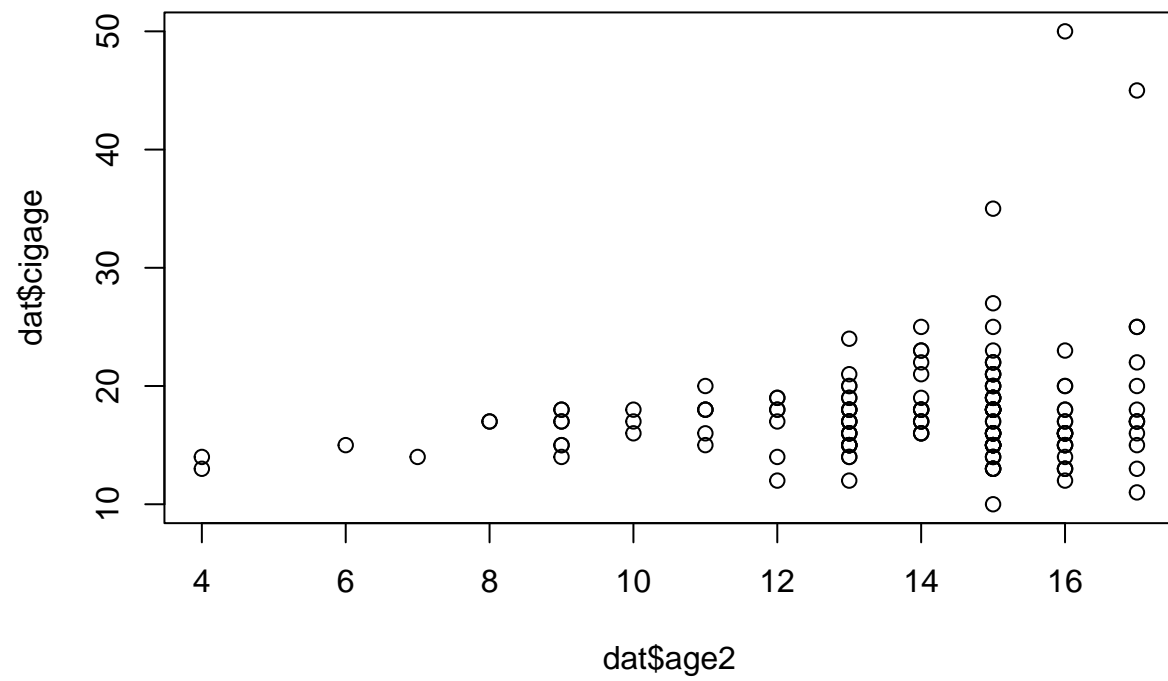
```
cor(dat$age2, dat$iralcage)
```

```
## [1] 0.07253557
```

```
plot(dat$age2, dat$iralcage)
```
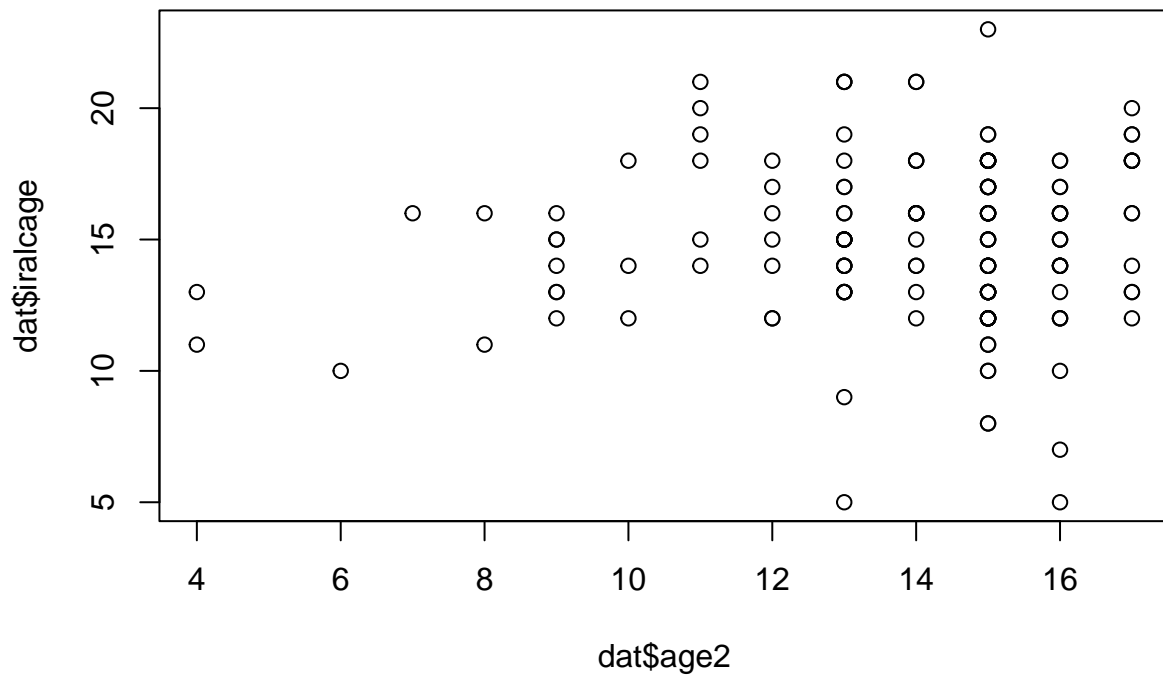
People in the survey tended to use alcohol earlier than other substances. This can be seen by the scatter plots. Additionally, alcohol has the lowest correlation coefficient.

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 Preferences Based on Small Sample from NSDUH 2019", xlab= "Cat
```

Sexual Attraction Preferences Based on Small Sample from NSDUH 2
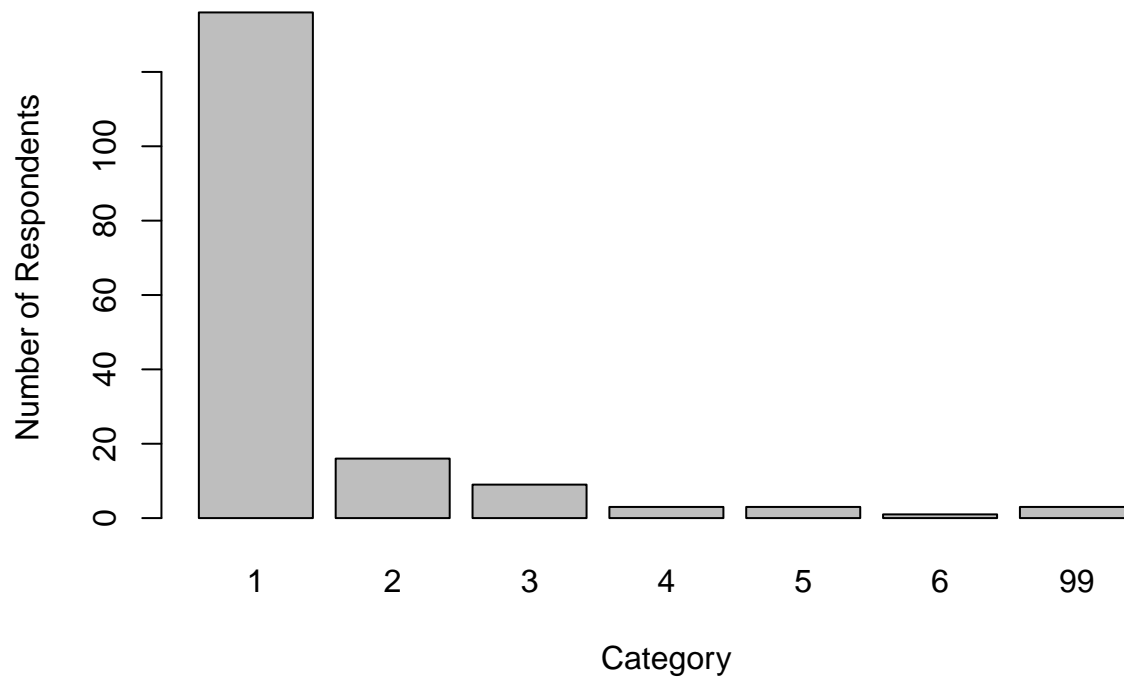
According to this bargraph, most people answered 1, which indicates preference for the opposite sex.This does not surprise me, as most people in the United States are heterosexual, however I would expect more people to answer 3 or 5.

What is the distribution of sexual attraction by gender?

```
counts <- table(dat$irsex, dat$sexatract)
barplot(counts, main= "Sexual Attraction Preferences Seperated by Gender ", xlab="Categories", ylab= "Nu
```

Sexual Attraction Preferences Seperated by Gender
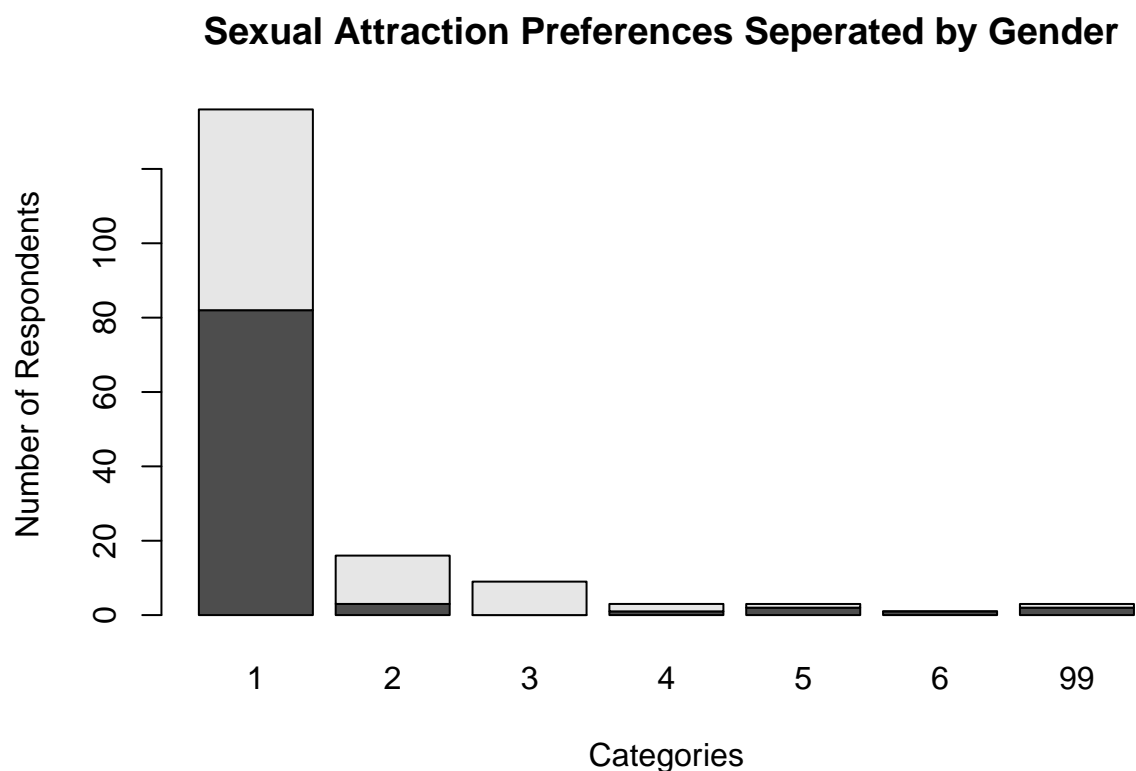
Group 1, which is heterosexual, is mainly composed of males. While group 3, which is bisexual is completely female. Group 6, which is homosexual, is also completely male.

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= "How Well Sample of Respondents of NSDUH 2019 Speak English", xlab="Categories", )
```

How Well Sample of Respondents of NSDUH 2019 Speak English
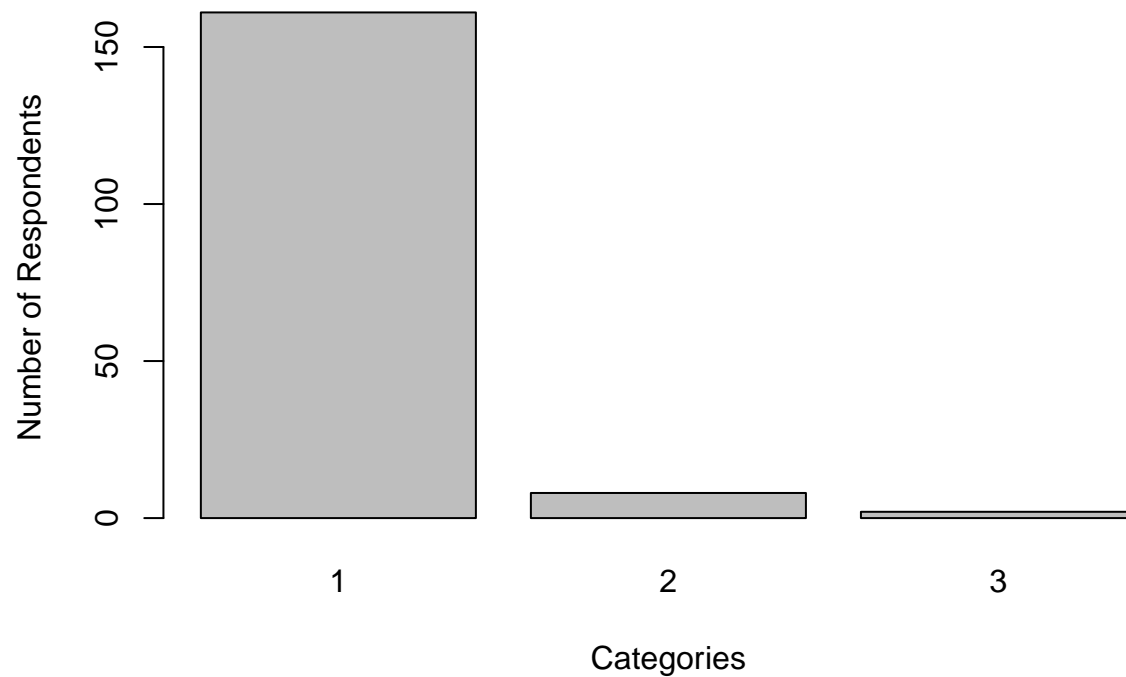
Most participants speak English very well, very few speak well and almost none speak not well or none at all. If this survey was only given out in English, this distribution makes sense. But in the US population, while the majority of people speak English, many do not speak it "very well."

#Are there more English speaker female or males?

```
counts <- table(dat$irsex, dat$speakengl)
barplot(counts, main= "How Well Sample of Respondents of NSDUH 2019 Speak English Seperated by Gender",
```

Well Sample of Respondents of NSDUH 2019 Speak English Seperated
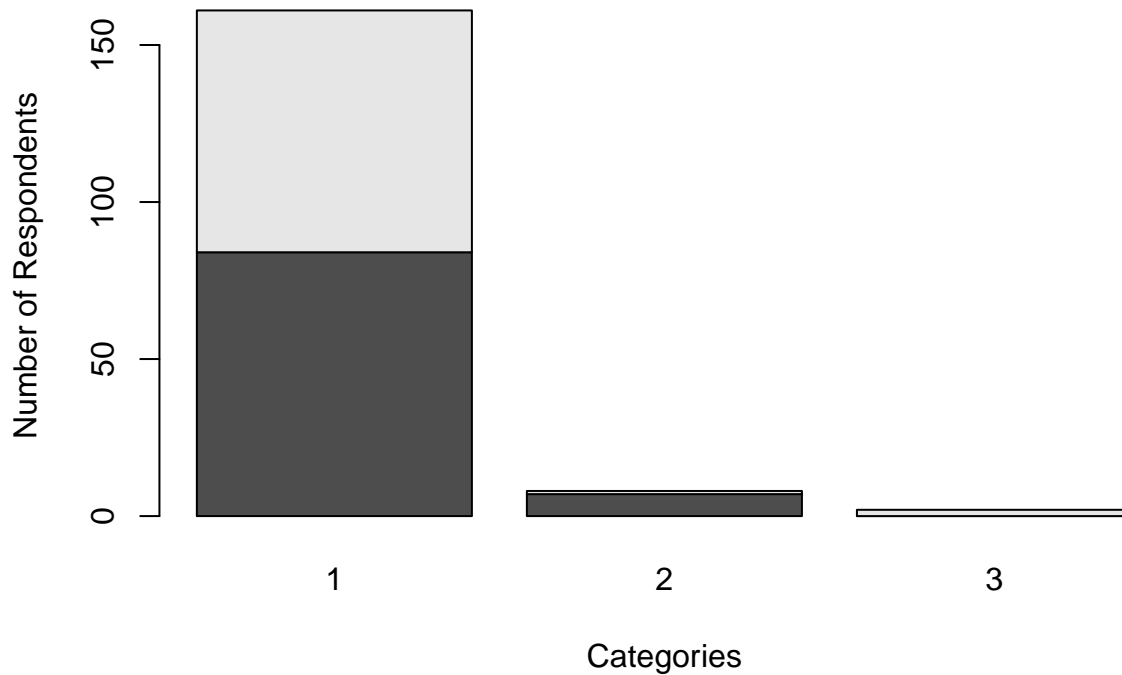
There are a similar number of males and females who speak English "very well" There are many more females who speak English "Well," so overall there are more females then males who speak English.

#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.

```
dat <- read.csv(file = 'fatal-police-shootings-data.csv')

head(dat)
```

```
##   id              name       date  manner_of_death      armed age gender race
## 1 3        Tim Elliot 2015-01-02             shot        gun  53      M    A
## 2 4  Lewis Lee Lembke 2015-01-02             shot        gun  47      M    W
## 3 5 John Paul Quintero 2015-01-03 shot and Tasered    unarmed  23      M    H
## 4 8    Matthew Hoffman 2015-01-04             shot toy weapon  32      M    W
## 5 9 Michael Rodriguez 2015-01-04             shot   nail gun  39      M    H
```

```
## 6 11  Kenneth Joe Brown 2015-01-04                shot        gun  18      M    W
##            city state signs_of_mental_illness threat_level        flee
## 1      Shelton   WA                     True       attack Not fleeing
## 2        Aloha   OR                    False       attack Not fleeing
## 3      Wichita   KS                    False        other Not fleeing
## 4 San Francisco  CA                     True       attack Not fleeing
## 5        Evans   CO                    False       attack Not fleeing
## 6      Guthrie   OK                    False       attack Not fleeing
##   body_camera longitude latitude is_geocoding_exact
## 1       False  -123.122   47.247               True
## 2       False  -122.892   45.487               True
## 3       False   -97.281   37.695               True
## 4       False  -122.422   37.763               True
## 5       False  -104.692   40.384               True
## 6       False   -97.423   35.877               True
```

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.

This data set is data collected by the Washington Post. It gives information about how many fatal shootings there were by an active duty police officer, and gives details about the victim (example: armed or unarmed.)

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

```
nrow(dat)
```

```
## [1] 6594
```

There are 6594 observations in 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.

```
names(dat)
```

```
##  [1] "id"                    "name"
##  [3] "date"                  "manner_of_death"
##  [5] "armed"                 "age"
##  [7] "gender"                "race"
##  [9] "city"                  "state"
## [11] "signs_of_mental_illness" "threat_level"
## [13] "flee"                  "body_camera"
## [15] "longitude"             "latitude"
## [17] "is_geocoding_exact"
```

"Body_camera" indicates whether or not the officer was wearing a body camera that may have recorded part of the incident. "flee" represents whether or not the victim was moving away from the officers and how they fled. "armed" represents whether or not the victim was armed with something the officers belived could inflict harm and specifies what, if anything, the victims were carrying.

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.

```
table(dat$armed)
```

```
##
##                                          air conditioner
##                              207                        1
##                        air pistol           Airsoft pistol
##                                1                        3
##                               ax                 barstool
##                               24                        1
##                      baseball bat   baseball bat and bottle
##                               20                        1
## baseball bat and fireplace poker    baseball bat and knife
##                                1                        1
##                            baton                   BB gun
##                                6                       15
##                BB gun and vehicle             bean-bag gun
##                                1                        1
##                      beer bottle              binoculars
##                                3                        1
##                     blunt object                  bottle
##                                5                        1
##                    bow and arrow               box cutter
##                                1                       13
##                            brick      car, knife and mace
##                                2                        1
##                          carjack                    chain
##                                1                        3
##                        chain saw                 chainsaw
##                                2                        1
##                            chair        claimed to be armed
##                                4                        1
##                contractor's level            cordless drill
##                                1                        1
##                         crossbow                  crowbar
##                                9                        5
##                        fireworks                 flagpole
##                                1                        1
##                       flashlight              garden tool
##                                2                        2
##                      glass shard                  grenade
##                                4                        1
##                              gun              gun and car
##                             3798                       12
##                    gun and knife          gun and machete
##                               22                        3
##                    gun and sword           gun and vehicle
##                                1                       17
##              guns and explosives                   hammer
##                                3                       18
##                       hand torch                  hatchet
##                                1                       14
```

```
##            hatchet and gun                 ice pick
##                        2                          1
##         incendiary device                    knife
##                        2                        955
##           knife and vehicle      lawn mower blade
##                        1                          2
##                  machete         machete and gun
##                       51                          1
##             meat cleaver          metal hand tool
##                        6                          2
##              metal object               metal pipe
##                        5                         16
##                metal pole               metal rake
##                        4                          1
##               metal stick              microphone
##                        3                          1
##                motorcycle                 nail gun
##                        1                          1
##                       oar               pellet gun
##                        1                          3
##                       pen             pepper spray
##                        1                          2
##                  pick-axe            piece of wood
##                        4                          7
##                      pipe                pitchfork
##                        7                          2
##                      pole           pole and knife
##                        3                          2
##            railroad spikes                    rock
##                        1                          7
##              samurai sword                 scissors
##                        4                          9
##               screwdriver             sharp object
##                       16                         14
##                    shovel                    spear
##                        7                          2
##                   stapler        straight edge razor
##                        1                          5
##                     sword                    Taser
##                       23                         34
##                  tire iron              toy weapon
##                        4                        226
##                   unarmed            undetermined
##                      421                        188
##            unknown weapon                  vehicle
##                       82                        213
##            vehicle and gun     vehicle and machete
##                        8                          1
##             walking stick               wasp spray
##                        1                          1
##                    wrench
##                        1
```

I was surprised to find binoculars, air conditioners and a motorcycle, because these are not things that I

would typically think of as weapons. I wonder who determined that these items were a threat.

Problem 2 (10 points)

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

```
hist(dat$age, main= "Frequency of ages of police shooting victims from Washington Post Data", xlab = "A
```

Frequency of ages of police shooting victims from Washington Post D
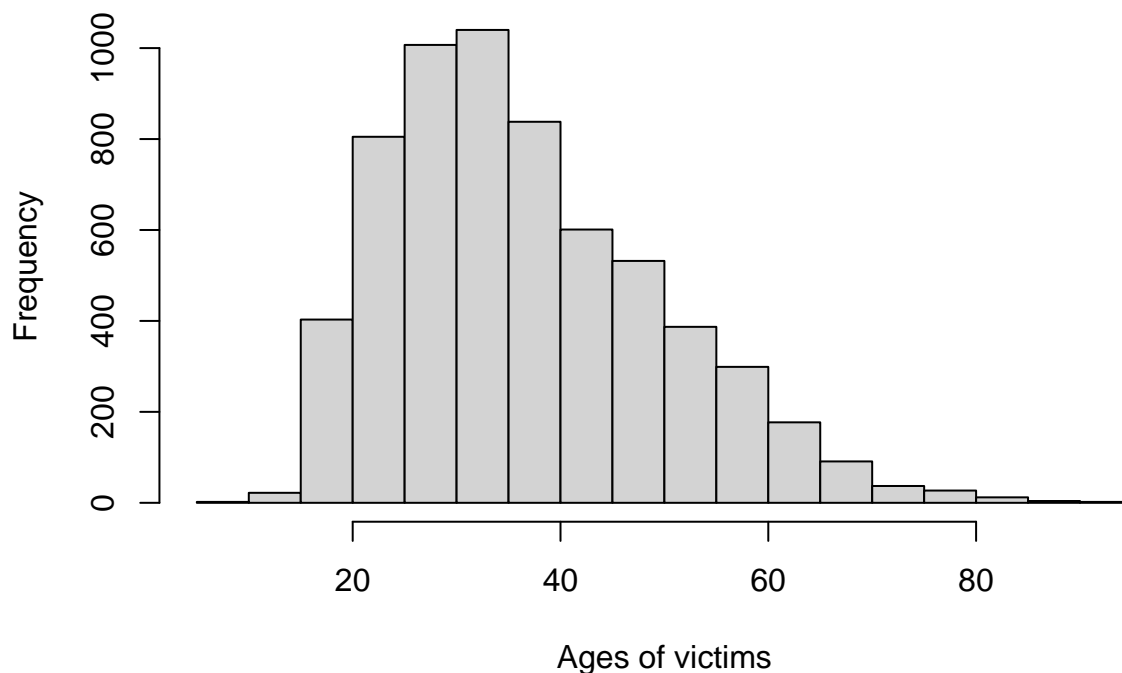
The age distribution is right skewed. This means that most of the victims were younger (20-40) This is what I would expect to see, because police shootings typically happen to people who police "perceive as threats" and very old people typically wont be seen as threats.

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

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

median(counts)
```

```
## [1] 66.5
```

Because this data is right skewed I used a median to find the center of age distribution. Outliers can have more of an impact on the mean, which would not make it as useful as a statistician.

27

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

```
table(dat$gender)
```

```
##
##          F    M
##    3  293 6298
```

The gender distribution is not even. There are far more men then women. This makes sense because police usually perceive men as more of a threat than women. Additionally, men are arrested more than women overall according to the FBI.

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?

```
table(dat$body_camera)
```

```
##
## False   True
##  5684    910
```

```
sum(table(dat$body_camera))
```

```
## [1] 6594
```

```
5684/6594
```

```
## [1] 0.8619958
```

There are many more people who were not wearing body cams then people who were. There were 5684 people not wearing body cams out of the 6594 recorded, which means 86.2 percent of officers did not have their cameras on. I would think that there would be more people wearing body cams in data recorded since 2015, because there has been a big push for more body cams in the recent years. However, many police officers do not turn on their 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?

```
table(dat$flee)
```

```
##
##                   Car       Foot Not fleeing      Other
##          491     1058        845        3952        248
```

```
1058 + 845 + 3952 + 248
```

```
## [1] 6103
```

```
(1058 + 845 + 248)/6103
```

```
## [1] 0.3524496
```

There are 491 pieces of missing data I excluded these from my calculations. I assume this means that the people who collected the data do not know whether the victims fled, but they did not make it clear. Additionally, I am assuming that "other" means people who fled in a different way then car or foot. There are a total of 6103 pieces of data (minus the data that was missing) and there are 845 people who fled on foot, 1058 by car and 248 by other. This means that only 35 percent of victims attempted to flee. I would expect it to be more, because police officers should not have to shoot people who are not fleeing. At the same time, if a victim is fleeing they may not be an imminent danger to a police officer (because they are running, not attacking), so it is possible the officers wouldn't shoot.

Problem 4 (10 points) - Answer only one of these (a or b).

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.

Your answer here.

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.

```
table(dat$race)
```

```
##
##      A     B     H     N     O     W
##    752   106  1553  1083    91    47  2962
```

```
plot(factor(dat$race), dat$age, main= "Barplot of Age of Victims and Race",xlab= "Race in Catagories", y
      names=c("unkown","Asian","Black","Hispanic","NA ","Other","White"))
```

Barplot of Age of Victims and Race
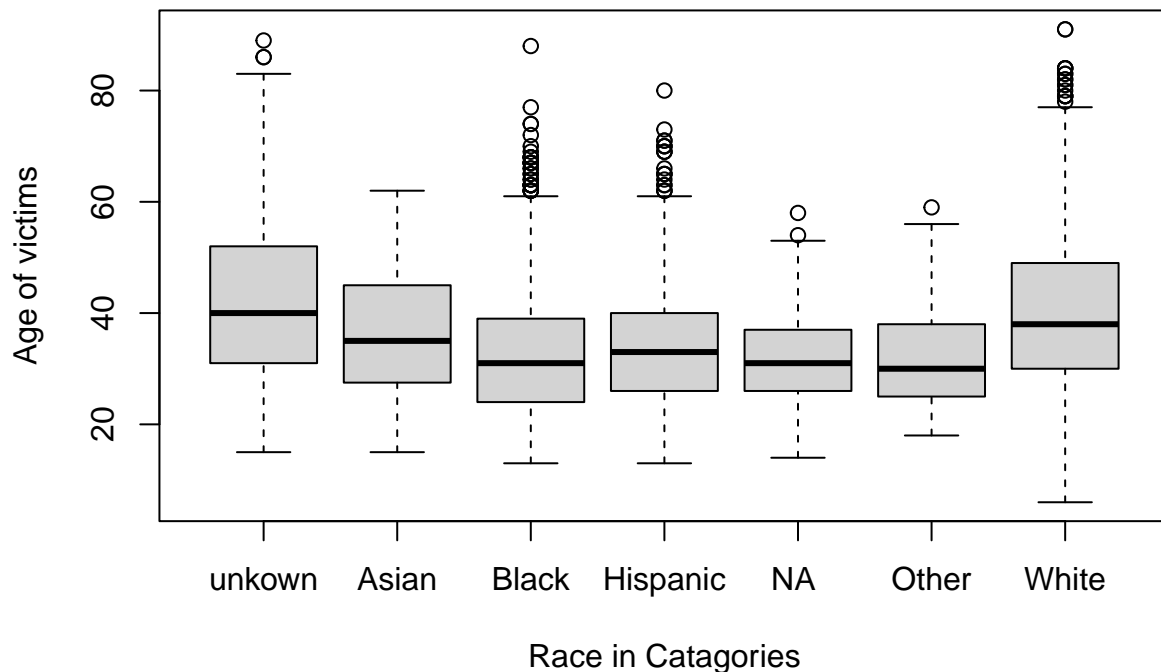
NOTE there is missing data. The first bar plot represents "unknown" data. Black= Black Non-Hispanic, White= White, non-Hispanic, NA= Native American

There is a clear distinction between the median age victims who are white and every other race (excluding unknown). The median age of victims who are white are much higher than the median age of victims of the other races. Additionally, the ages of white victims are much more evenly dispered. The box is close to the middle which shows a more even dispersion, and the box is longer which shows more dispersed data within Q2 and Q3. Native Americans and black victims have the lowest median age, although there are many outliers who are black. This data shows that victims or color tend to be younger than white victims of police shootings.

Extra credit (10 points)

a. What does this code tell us?

```
mydates <- as.Date(dat$date)
head(mydates)
(mydates[length(mydates)] - mydates[1])

2458/365
```

This code tells us the number of days of data we have. The data has been being kept since 2015, and it is now 2021. So it makes sense that the data is of 2458 says (6.7 years)

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?

It seems that police killings are underreported in federal data bases. According to the Washington Post Study, the FBI and CDC does not log as many fatal shootings as there actually are. Police killings are most likely under reported because it is a liability to police forces when one of their officers shoots and kills a victim. Additionally, as this data shows, minority victims tend to be younger, which demonstrates possible racial prejudice. Police killings are most likely misreported because people may blame death on other factors than a gun wound. If the cause of death is reported differently, then a victim would not be counted towards a police shooting death.

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 problem 4 the missing value is an unknown race. This differs from the race catagory of 'other,' which shows victims who do not identify as one of the specific racial categories laid out by the Washington Post. I assume that there is much more information missing in general. There is missing data about whether victims fled. There are probably many missing values of signs of mental illness, because that is not easy data to collect. Additionally, the data that the Washington Post collected is from many different sources, none of which have all the pieces data they wish to collect. Overall, there is probably a lot of data and variables missing that are important to understanding the full scope of the circumstances of the shooting. This data just shows us a very small lens of information.