Assignments

Contents

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

Instructions for all assignments

I want you to submit your assignment as a PDF, so I can keep a record of what the code looked like that day. I also want you to include your answers on your personal GitHub website. This will be good practice for editing your website and it will help you produce something you can keep after the class is over.

1. Download the Assignment1.Rmd file from Canvas. You can use this as a template for writing your answers. It's the same as what you can see on my website in the Assignments tab. Once we're done with this I'll edit the text on the website to include the solutions.
2. On RStudio, open a new R script in RStudio (File > New File > R Script). This is where you can test out your R code. You'll write your R commands and draw plots here.
3. Once you have finalized your code, copy and paste your results into this template (Assignment 1.Rmd). For example, if you produced a plot as the solution to one of the problems, you can copy and paste the R code in R markdown by using the ``` ```{r} ``` ``` command. Answer the questions in full sentences and Save.
4. Produce a PDF file with your answers. To do this, knit to PDF (use Knit button at the top of RStudio), locate the PDF file in your docs folder (it's in the same folder as the Rproj), and submit that on on Canvas in Assignment 1.
5. Build Website, go to GitHub desktop, commit and push. Now your solutions should be on your website as well.

Assignment 1

Collaborators: Lorem Ipsum.

This assignment is due on Canvas on Monday 9/20 before class, at 10:15 am. Include the name of anyone with whom you collaborated at the top of the assignment.

Problem 1

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

```
library(datasets)
```

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?

It is useful to rename datasets because it gives us a shorthand to work with. So in this case, instead of referring to the data with "USArrests" we can ref to it with dat.

```
dat <- USArrests
```

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`.

```
summary(dat)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
names(dat)
```

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

The four variables are "Murder", "Assault", "UrbanPop", and "Rape" (and the state variable which we created).

Problem 3

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

Answer: It is a quantitative variable because this variable is representing some numerical value in relation to a state.

What R Type of variable is it?

Answer: This variable is a character because the word murder itself is represented in a string format.

Problem 4

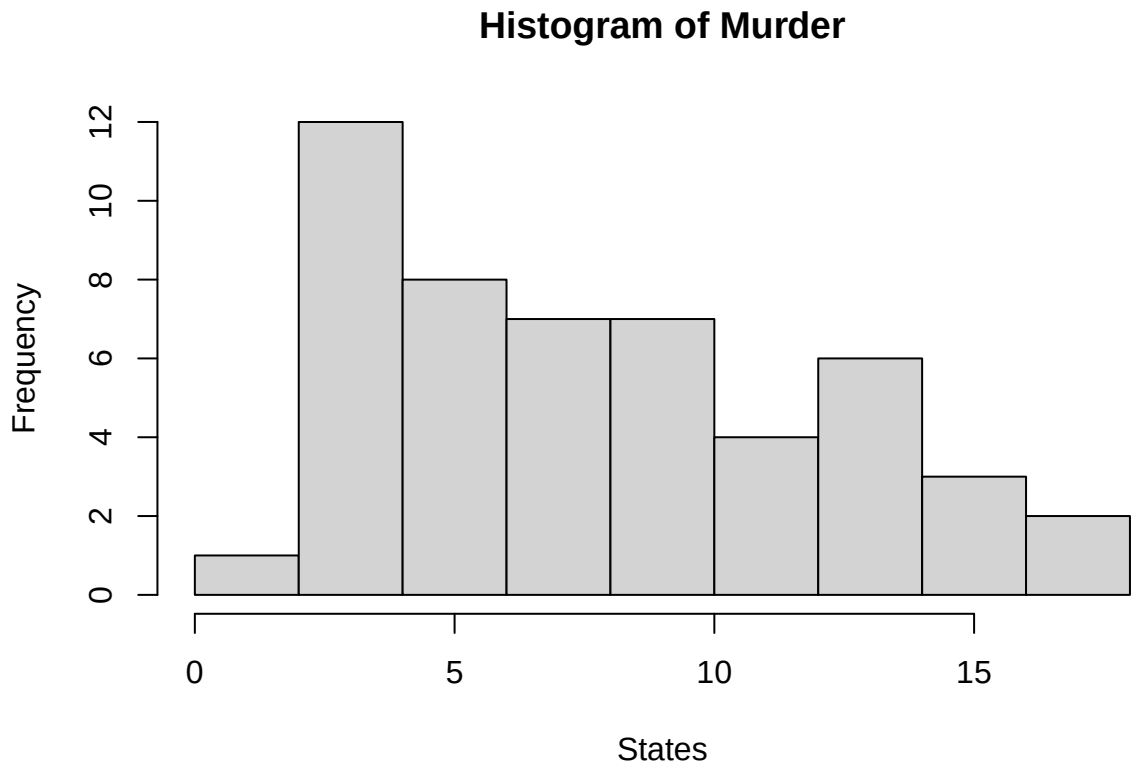
What information is contained in this dataset, in general? What do the numbers mean?

Answer: The dataset contains information about murder, assault, and rape. Additionally, it seems to give us some numbers for a states urban population to help see the relation aswell. These numbers show us the relationship with often they are occuring) of these different variables in different states. For example a number for murder is telling it there was some amount of murders within this state (and we can compare this to other states by seeing how much more or less these crimes occur in other states).

Problem 5

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

```
hist(dat$Murder, main = "Histogram of Murder", xlab = "States")
```



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
```

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

The mean is 7.788 and the median and 7.250. The mean is the average of the dataset while the median gives us a central value of our dataset. A quartile tells us the variability around the median. So the 1st and 3rd quartiles show us the variability before the median is reached and after the median is reached. R gives us this data to show us where it might be more or less skewed.

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.

Answer (for data on the other two variables) : For assaults, the mean is 170.8 and the median and 159.0.

For rapes, the mean is 21.23 and the median and 20.10.

```
summary(dat$Murder)
```

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

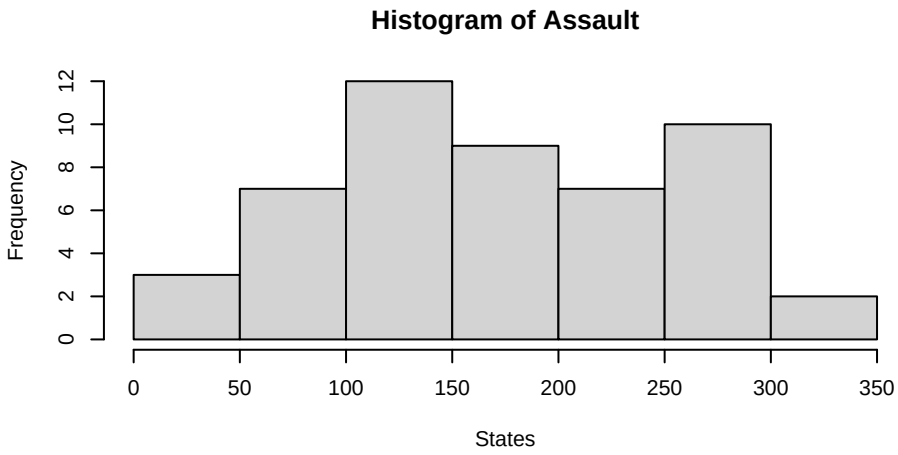
```
summary(dat$Assault)
```

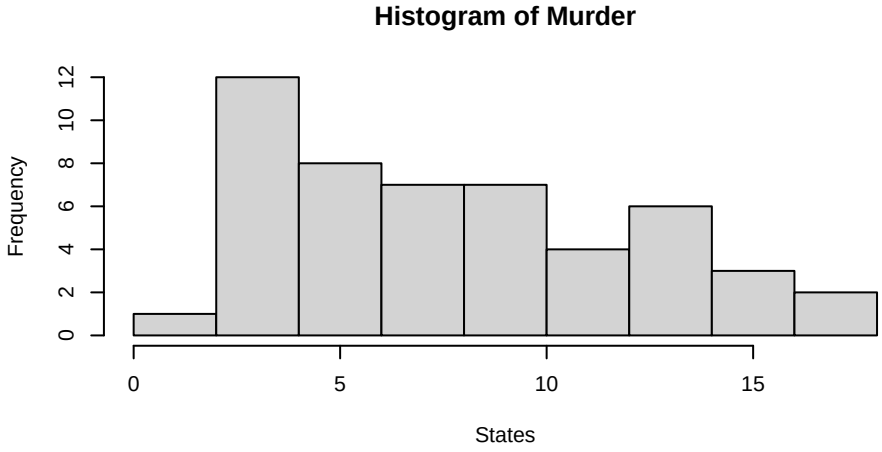
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    45.0   109.0   159.0   170.8   249.0   337.0
```

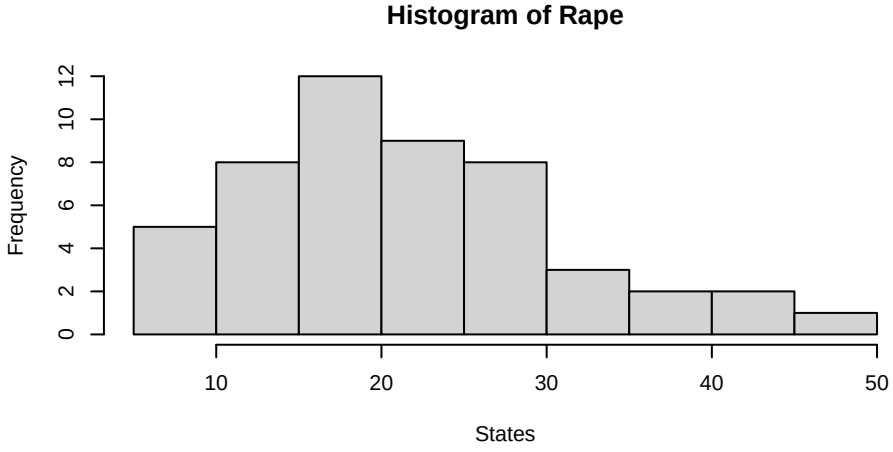
```
summary(dat$Rape)
```

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

```
par(mfrow=c(3,1))
hist(dat$Assault, main = "Histogram of Assault", xlab = "States")
hist(dat$Murder, main = "Histogram of Murder", xlab = "States")
hist(dat$Rape, main = "Histogram of Rape", xlab = "States")
```







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

Answer: It helps us combine multiple plots that we created into one big vertical plot.

What can you learn from plotting the histograms together?

Answer: We can see the correlation between the data. If there are some points where there is a peak at the

same time then we can generalize and say that state could be more dangerous than others. This could work likewise for the converse situation.

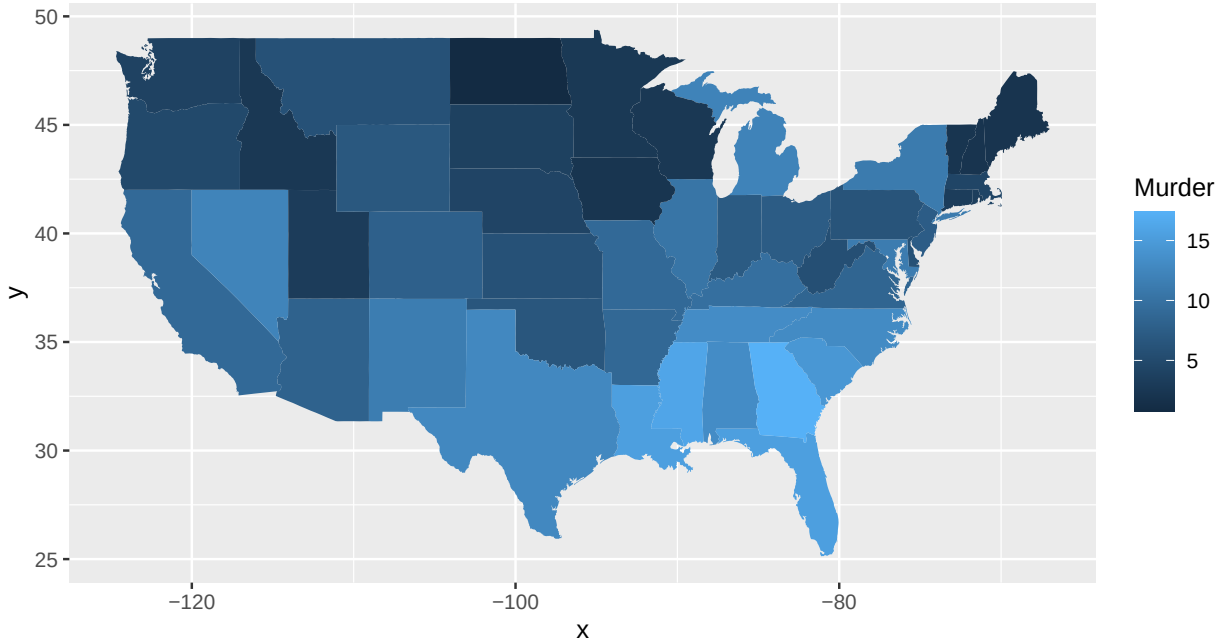
Problem 8

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')

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: The first two lines import the libraries map and ggplot2. Line 154 imports dat wants to define the map id of states with murder. The second line is filling out map with state data.The third line maps it out in a x and y axis so we can see the data in states.

Assignment 2

Problem 1: Load data

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

```
setwd("/Users/akshatshah/Desktop/upenn/crim250/LabJournal")
```

Read the data

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

What are the dimensions of the dataset?

```
dim(dat)
```

```
## [1] 171   7
```

Answer: There are 171 rows and 7 columns in this dataset.

```
names(dat)
```

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

Problem 2: Variables

Describe the variables in the dataset.

The variables are mjage, ciage, iralcage, age2, sexatract, speakengl, and irsex. They each tell us different things. Mjage tells how old someone was when they first starting using marijuana. Ciage tells how old someone was when they first starting smoking cigs every day. iralcage tells how old someone was when they tried alcohol. age2 tells us hpw old the person currently is. However, this gives us a range because a person could have changes their choice based on previous responses and the questions they say. Irsex tells us their gender. sexatract tells us describes their attraction/sexuality. speakengl tells how well the individual speaks english.

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

This dataset is a survey about national drug use and health. The data is sponsored by the United States Health and Human Services. The sample is a scientific random sample of household addresses. This data gives us a state and nationwide statistics on drug use. This information is used to help with prevention, trend studies, and inform public health policy.

Problem 3: Age and gender

What is the age distribution of the sample like? Make sure you read the codebook to know what the variable values mean.

The age distribution is more towards the 13-17 year old range. However, when we look at this data we should understand that it can be more of a range since participants could change their answers based on the decisions on they made or what they see fit.

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

I believe this data not a good representative of the US population. We are looking at a younger population that makes up around 35% of the age distribution. So we are leaving out a large majority that can help us see more trends.

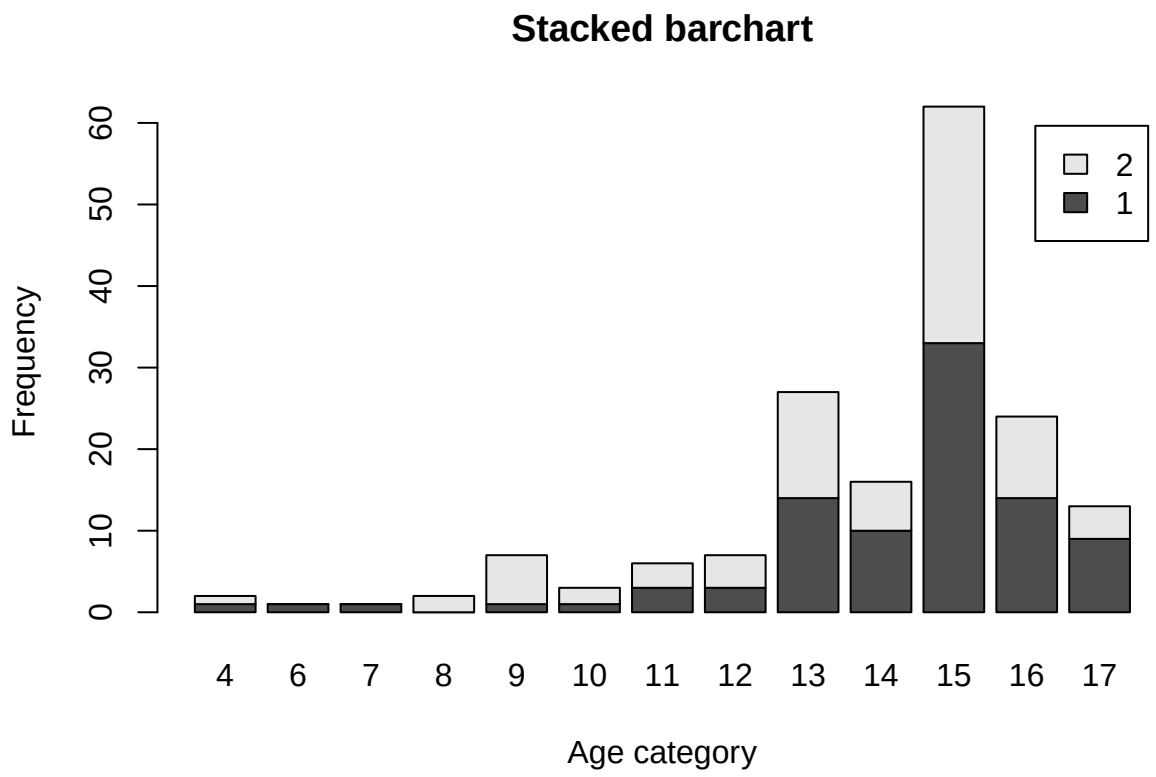
Is the sample balanced in terms of gender? If not, are there more females or males?

I believe this data is pretty balanced. There is a pretty even distribution but for some of the ages we can see that there is a clear majority like for 17.

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)
```



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?

```
summary(dat)
```

```
##      mjage           cigage         iralcage          age2
##  Min.   : 7.00   Min.   :10.00   Min.   : 5.00   Min.   : 4.00
##  1st Qu.:14.00   1st Qu.:15.00   1st Qu.:13.00   1st Qu.:13.00
##  Median :16.00   Median :17.00   Median :15.00   Median :15.00
##  Mean   :15.99   Mean   :17.65   Mean   :14.95   Mean   :13.98
##  3rd Qu.:17.50   3rd Qu.:19.00   3rd Qu.:17.00   3rd Qu.:15.00
##  Max.   :35.00   Max.   :50.00   Max.   :23.00   Max.   :17.00
##    sexatract       speakengl        irsex
##  Min.   : 1.00   Min.   :1.00   Min.   :1.000
##  1st Qu.: 1.00   1st Qu.:1.00   1st Qu.:1.000
##  Median : 1.00   Median :1.00   Median :1.000
##  Mean   : 3.07   Mean   :1.07   Mean   :1.468
##  3rd Qu.: 1.00   3rd Qu.:1.00   3rd Qu.:2.000
##  Max.   :99.00   Max.   :3.00   Max.   :2.000
```

Looking at the summary of the data, we can see on average what is used earlier on. Alcohol seems to be used the earliest at an average of 14.95. Then it is mjage at 15.99 Finally, it is cigage at 17.65.

Problem 5: Sexual attraction

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

```
table(dat$sexatract)
```

```
##
##   1   2   3   4   5   6  99
## 136  16   9   3   3   1   3
```

We see that the largest amount of people (136) chose 1. Yes, this is what I expected since it is the norm to be heterosexual.

What is the distribution of sexual attraction by gender?

```
table(dat$sexatract, dat$irsex)
```

```
##
##       1  2
##   1  82 54
##   2   3 13
##   3   0  9
##   4   1  2
##   5   2  1
##   6   1  0
##   99  2  1
```

By gender, the distribution is the same in that the majority of both gender are attracted to the opposite gender. However, more females chose other options.

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?

```
table(dat$speakengl)
```

```
##
##   1   2   3
## 161   8   2
```

The majority of people chose that they speak english very well. And there were very few who chose well and not well (10 total). This is probably what I would expect since it is the dominant language.

Are there more English speaker females or males?

```
table(dat$speakengl, dat$irsex)
```

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

There are more English speakers who are male than female.

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 = "Crim 250 - Exam 1/fatal-police-shootings-data.csv")
```

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 is a dataset that is compiled by the Washington Post of victims of fatal police shootings. At each row, we are given a victim's name and data on the situation that was at hand.

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

```
dim((dat))
```

```
## [1] 6594   17
```

We know there are 6594 rows and 17 columns. We know that the number of observations are the number of rows. Therefore there are 6593 observations (not including the first row since this is the title of the columns).

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 is a variable that is telling us if an officer was wearing a body camera and if it was recording what happened.

Flee is a variable that was indicating if the victim was moving away, and if they were fleeing this tells us by what method.

Armed is a variable that tells if the officer believe they had some tool that could inflict damage.

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
```

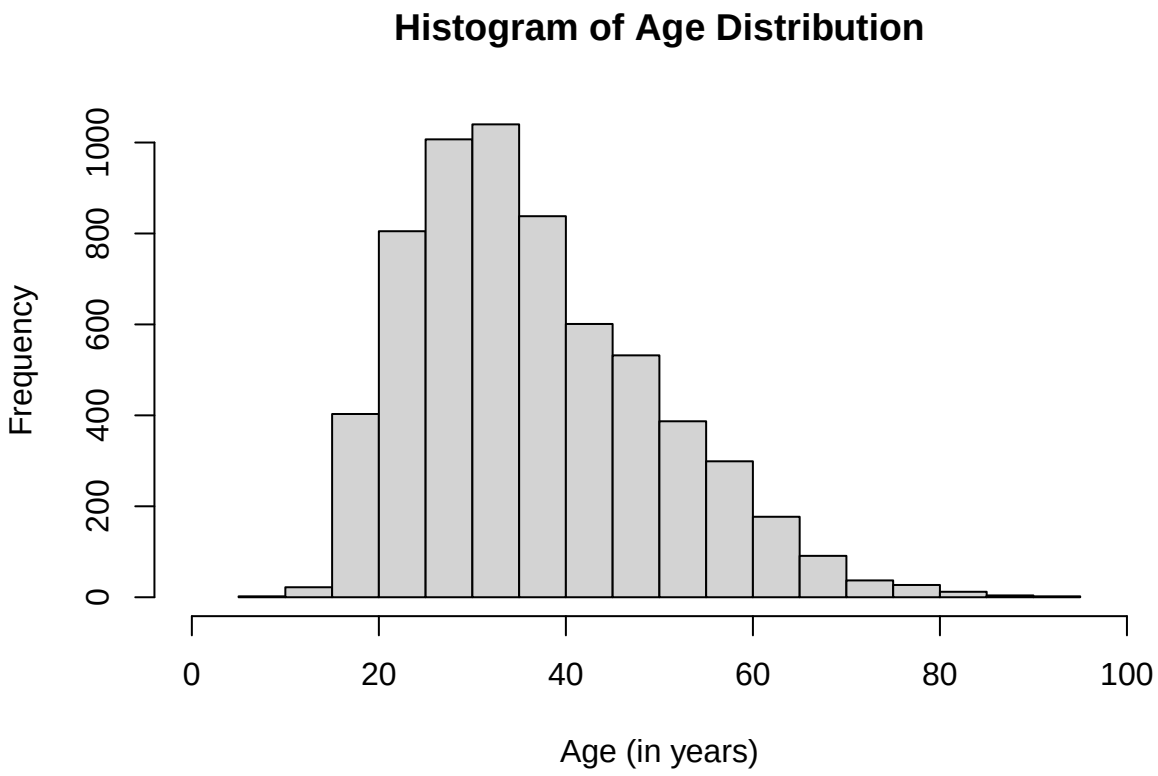
```
##                      pick-axe                   piece of wood
##                             4                               7
##                          pipe                       pitchfork
##                             7                               2
##                          pole                  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 am suprised to see pen, binoculars, and contractor's level.

Problem 2 (10 points)

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

```
hist(dat$age, main = "Histogram of Age Distribution", xlab = "Age (in years)", xlim = c(0, 100))
```



This distribution is skewed to the right. This isn't exactly what I expected, (I thought it would be skewed even more to the right.) because I believed that victims would be a lot younger.

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

```
summary(dat$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   27.00   35.00   37.12   45.00   91.00     308
```

I would use mean to understand the center of the age distribution because I don't believe there are enough outliers/extremes that would skew this result substantially. The mean gives us an average over our data. We can see that the average age of victims for police fatal shootings were 37.12 years old.

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

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

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

Within this dataset, there are significantly more men than women (6005 more). This is not surprising because statistics have shown that men have been higher arrest rate than women. Since men have a higher arrest rate than women and more encounters with police, I expected there to be more men than women within this dataset. Additonally, there were 3 blank indexes for this data and it was not taken into account for the calculations above since it is inconclusive.

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
```

Only 910 officers had a body camera according to this dataset. This means that only 14% of incidents had a body camera. This is extremely suprising that is so low because people are losing their lives and a proportion of officers have no concrete evidence of the situation due to no body camera.

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
```

Out of 6103 incidents that recorded something in the fleeing category, 2151 victims were fleeing. This means about 35% of people who were a victim of a fatal police shooting were fleeing. I suspected a larger proportion of people were fleeing, but this dataset refutes that idea. Additionally, there is 491 indexes of data that are blank for fleeing, so this was removed from the total number of incidents and the proportion since it is inconclusive.

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.

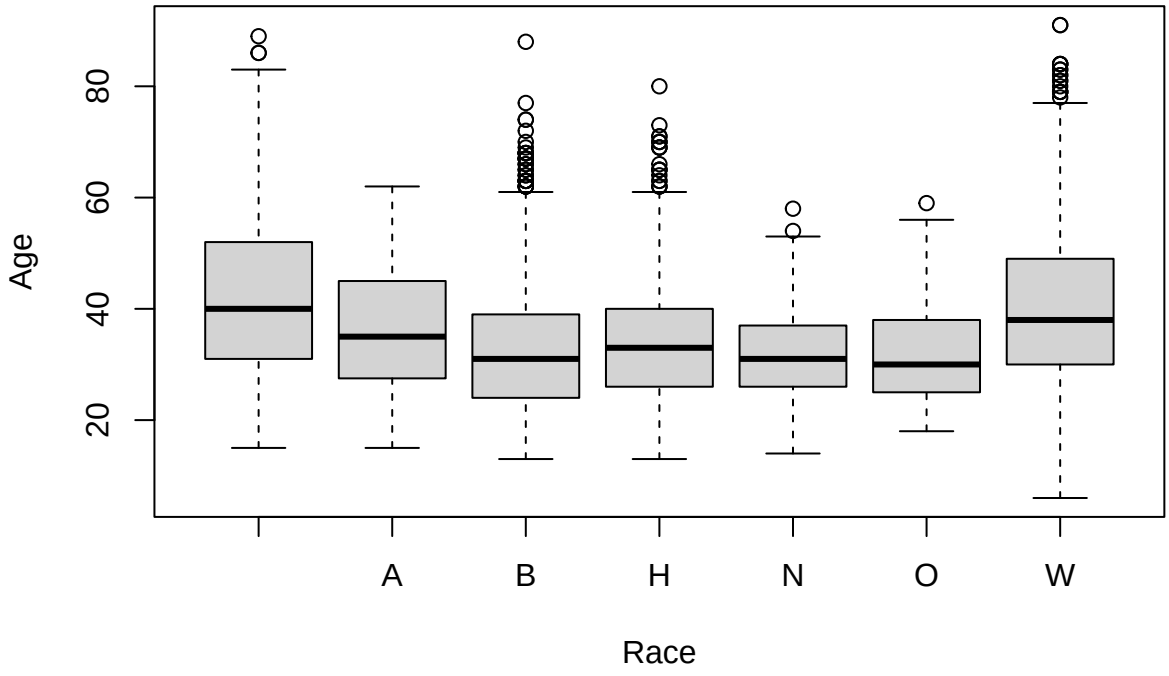
Your answer here.

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.

```
boxplot(dat$age~factor(dat$race), ylab = "Age", xlab = "Race")
```



We see from the data that some races have a slightly lower mean age than others. For examples, the mean age for white people looks to be around 40, while the mean age for black people seems to be about 35. Additionally, some races have a lot more outliers than others. For example, asian people have no outliers while black people have a lot of outliers. It is hard to make a concrete conclusion based on this relationship, but we can observe that there are clear small differences that we can observe (as stated above) from the different races and their ages. Furthermore, the first category simple represent people who did not have a specified age (left blank in the dataset). I did not omit this data because I think it could still be important to see this in relation to the other data.

Extra credit (10 points)

a. What does this code tell us?

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

This data tells us how long it has been since the first entry within this dataset to the most recent entry within the dataset. We are taking the first index because the 0th index states the name of each column, and we are taking the last index to get the last entry. The difference is 2458 days which is about 6.7 years. This makes sense because this dataset was created in 2015 and we are about 6 and 3/4 years from this time.

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?

I believe this is because of bias. Police who have been apart of police killings are going to defend themselves. In order to do so, they might underreport or mislabel the incident in order to save face and justify the actions they took. Additionally, co workers of police who have been apart of such an incident may look to defend each other building even more of a bias.

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, there were clear missing values (race was not defined). I believe there is more missing from the data. If we were to look closer into the data we can see that sometimes, for example, gender isn't specified. Additionally, another data that has missing values is fleeing (491 are blank).