Assignments

Contents

Assignment 1

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?

```
dat <- USArrests
```

Problem 2

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

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

This dataset has the state names as row names, so we just want to make them into a new variable. We also make them all lower case, because that will help us draw a map later - the map function requires the states to be lower case.

List the variables contained in the dataset `USArrests`.

```
names(dat)
```

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

Problem 3

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

Answer: According to the DVB reading, murder would be a quantitative variable within this particular context.

What R Type of variable is it?

Answer: Murder is a UNIVARIATE NON-GRAPHICAL variable in R.

Problem 4

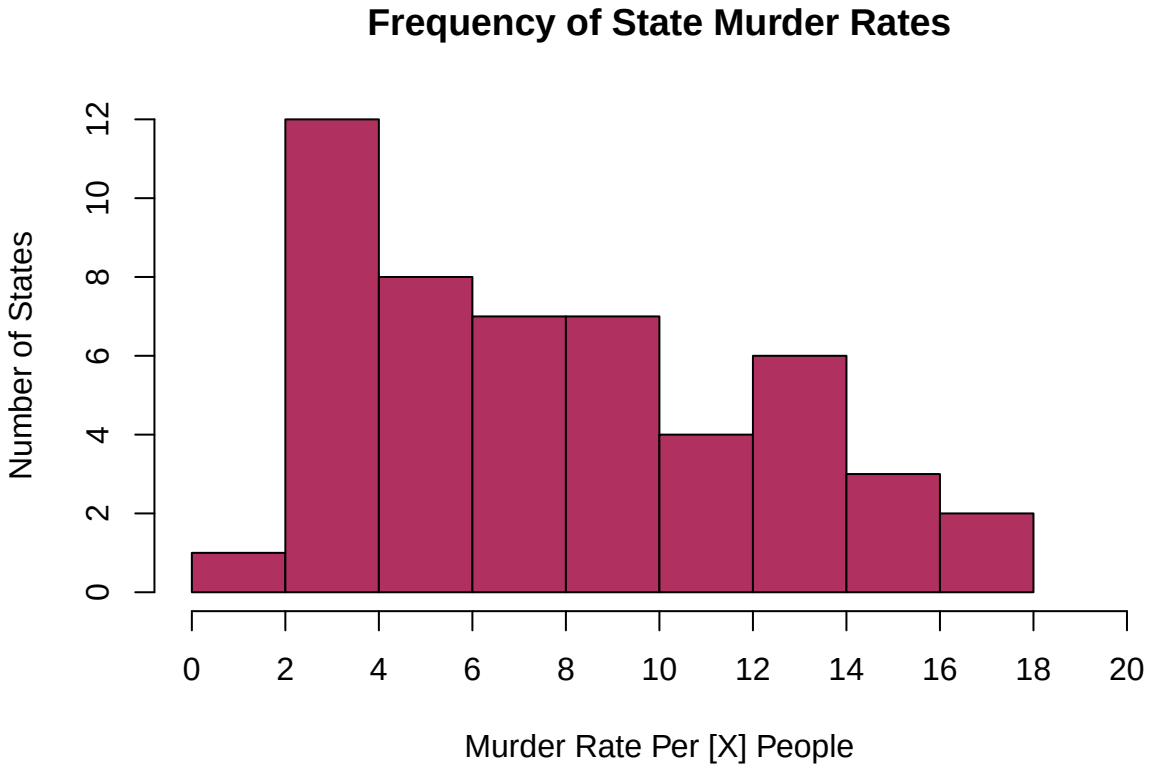
What information is contained in this dataset, in general? What do the numbers mean?

Answer: The dataset's variables compare the incidents or rates of murder, assault, and rape along state lines, as well as those of urban population sizes. Given the tenths decimal places featured in the categories of murder and rape, it is safe to assume those numbers present rates of reported incidents as neither murder nor rape can occur to a tenth degree. Assault, on the contrary, appears to be reported as raw, reported incidents given the data is presented as whole integers. The state category lists the various American states, adjacent to each states' urban population. As a note, the urban population data is presented seemingly as either proportional integers or on a scale; to illustrate, any one state's urban population data may identify what percentage of the state's population lives in an urban center.

Problem 5

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

```
hist(dat$Murder, main="Frequency of State Murder Rates",
     xlab="Murder Rate Per [X] People", ylab="Number of States",
     xlim=c(0,20), col="maroon", breaks=10, xaxp=c(0,20,10))
```



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          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
##     state
##  Length:50
##  Class :character
##  Mode  :character
##
##
##
```

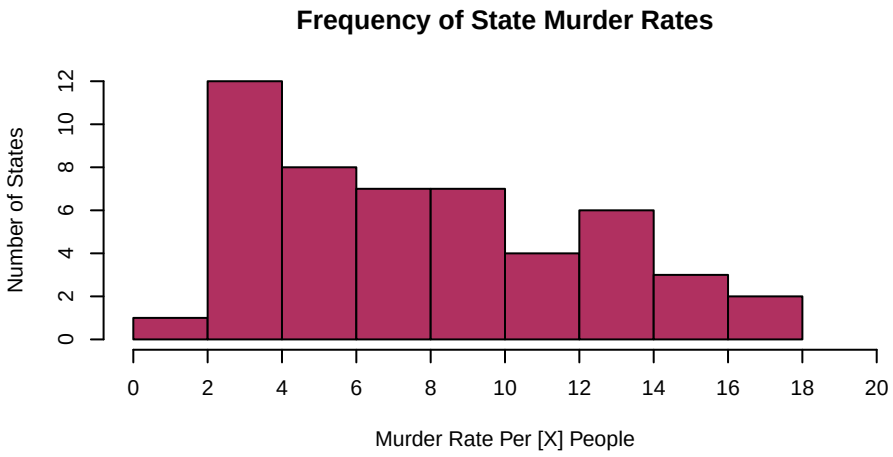
The mean for murder is 7.788, which is the average of all state murder rates; the median of 7.25 is in contrast the figure at the middle of the string of variables if sequenced. The quartiles are the percentiles of the data. Combined, both the 1st and 3rd quartile depict where half the data lies betwee. That is, 25% of states have murder rates below 4.075, the 1st Quartile, while 75% of states have murder rates are less than 11.25. Likewise, 75% of states have murder rates higher than 4.075 and 25% of states have a murder rate higher than 11.25; 50% of states would thus have a murder rate between 4.075 and 11.25. R must provide the

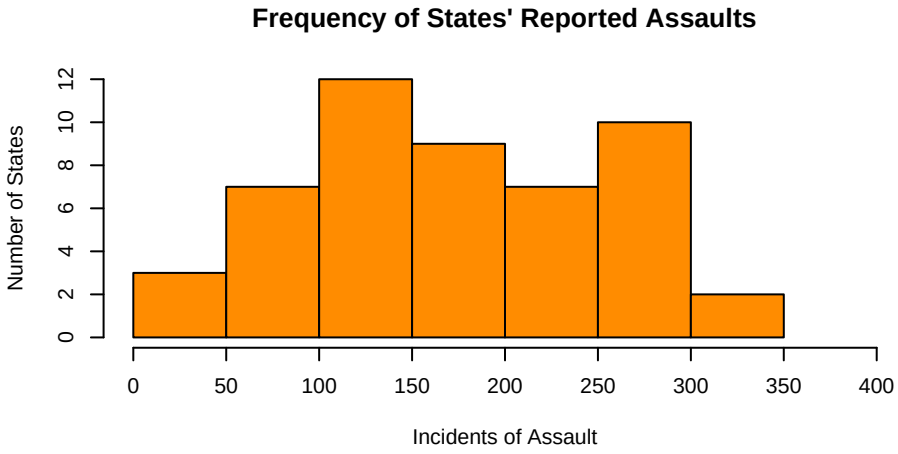
percentiles or quartiles to depict where half of the data lies between, presenting a numerical bell curve of sorts. *The .rmd file notably produces a summary of the state variable as well, with a length of 50 and a class and mode of character. The calculation, however, is cut out in the knitted pdf document.

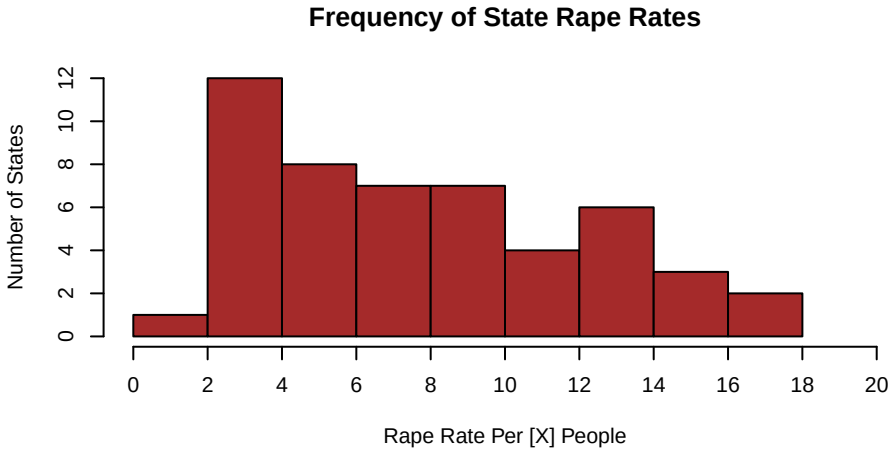
Problem 7

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.

```
par(mfrow=c(3,1))
hist(dat$Murder, main="Frequency of State Murder Rates",
     xlab="Murder Rate Per [X] People", ylab="Number of States",
     xlim=c(0,20), col="maroon", breaks=10, xaxp=c(0,20,10))
hist(dat$Assault, main="Frequency of States' Reported Assaults",
     xlab="Incidents of Assault", ylab="Number of States",
     xlim=c(0,400), col="darkorange", breaks=10, xaxp=c(0,400,8))
hist(dat$Murder, main="Frequency of State Rape Rates",
     xlab="Rape Rate Per [X] People", ylab="Number of States",
     xlim=c(0,20), col="brown", breaks=10, xaxp=c(0,20,10))
```







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

Answer: Command par allows you define the parameters of a graph or computation. In this specific case, we set the parameters to the entirety of three graphs, which are all included in one plot.

What can you learn from plotting the histograms together?

Answer: By plotting all three plots adjacent of one another, the combined graphic allows for a quick assessment of the number of states with varying assault, murder, and rape rates. The graph thus allows you for a

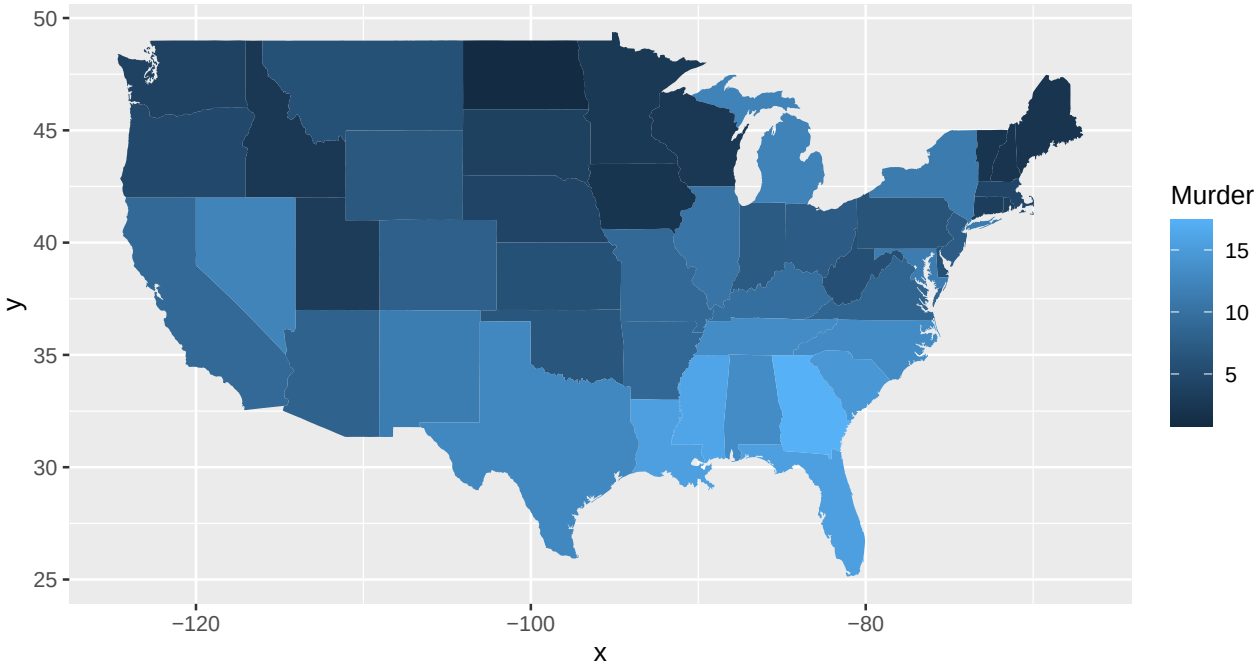
quick comparison of central tendency, spread, skewness, and kurtosis for the various crimes.

Problem 8

In the console below (not in text), type `install.packages("maps")` and press Enter, and then type `install.packages("ggplot2")` and press Enter. This will install the packages so you can load the libraries.

Run this code:

```
# install.packages("maps")
# install.packages("ggplot2")
library('maps')
library('ggplot2')


ggplot(dat, aes(map_id=state, fill=Murder)) +
  geom_map(map=map_data("state")) +
  expand_limits(x=map_data("state")$long, y=map_data("state")$lat)
```



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

Answer: The code generates a map of the United States where each state is shaded according to their murder rate. States with darker shadings have lower murder rates as those with ligher shades have higher murder rates. The install commands download the graphed data, that is they download a map and hold the manipulated data. The ggplot command then creates the graph by loading the "dat" variable, which was predefined as the US Arrest data. The map is then divided along state lines and shaded according to the murder rate, per the map_id and fill command respectively. The geom_map command allows for additional manipulation to the states' borders, defining each state alongside existing state boundaries. Finally, the expand_limits command expands the graph by manipulating the x and y axes.

Assignment 2

Problem 1: Load data

```
dat <- read.csv(file = 'assignment 2/dat.nsduh.small.1.csv')
```

What are the dimensions of the dataset?

```
names(dat)
```

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

Problem 2: Variables

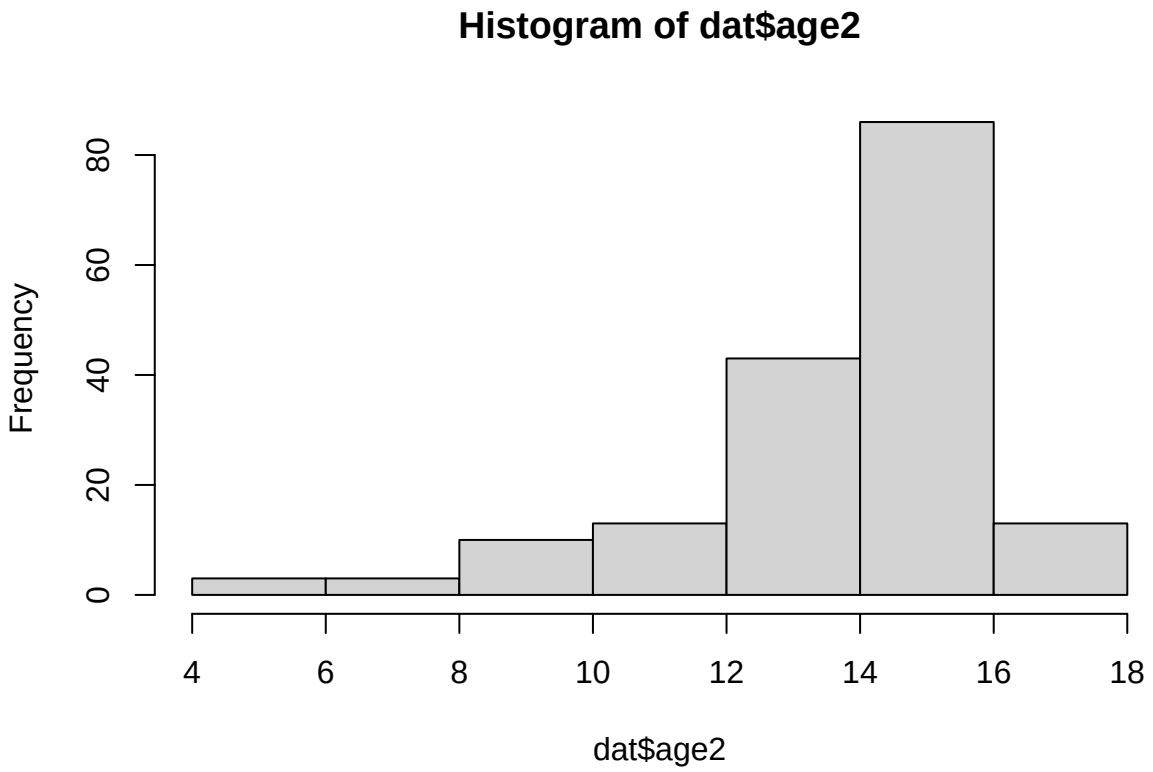
Describe the variables in the dataset. "mjage": age of when participants first tried marijuana "cigage": age of when participants first tried cigarettes "iralcage": age of when participants first tried alcohol "age2": calculated ages of participants given multiple responses "sexatract":sexual attraction of participants "speakengl": english proficiency level of participants "irsex": sex of participants

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 dataset is collected from the National Study of Drug Use and Health, which seeks to identify the ages at which the study's participants first began consuming alcohol, cigarettes, and marijuana. The study further collects data on its participants' demographic data, such as age, gender, and sexual orientation, as well as English proficiency. The study's variables are a mixture of categorical and quantitative variables. An example of the former, categorical, includes the sex, English proficiency, and sexual orientation variables. Others, such as the age variables, would be categorized as quantitative.

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)
```



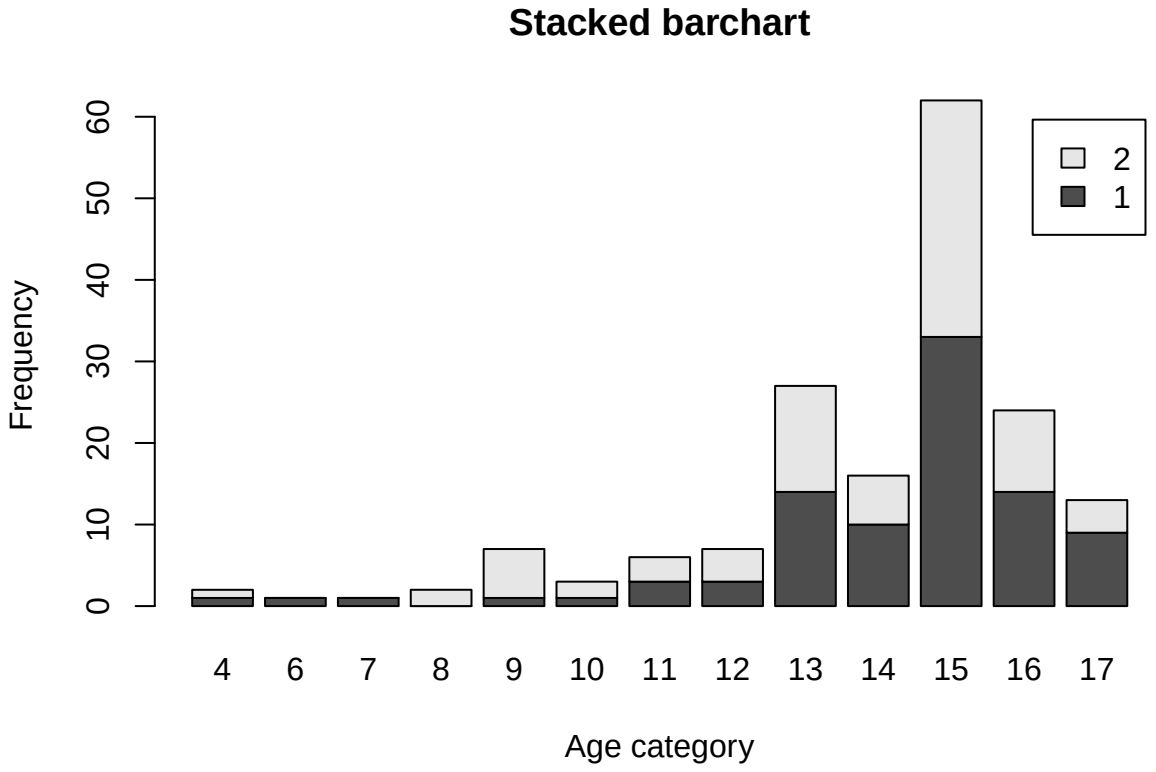
The age distribution of the sample size is very heavily skewed towards individuals 35 to 49 years old, or rather the category associated with the age group is the largest proportionately. This may be because the younger age groups are split and divided along single years and later grouped in categories of growing margins. For example, ages 12 to 21 are all categories along single-year lines whereas ages 30 to 64 are all divided in three categories. The sum of the earlier younger ages may result in a more proportional balance of the different ages.

Do you think this age distribution representative of the US population? Why or why not? No, the distribution does not represent an accurate representation of the US population. This is due to two primary reasons, with the study's intent being the most important. Because the study seeks to identify the ages at which its participants first began utilizing substances, its participants are not entirely random. The study was conducted for a purpose other than studying the US population, which may present sample size issues. An example of this influence is the lack of participants ages 12 and younger. The absence of this age group presents the second reason why the study cannot produce accurate figures on the US population.

Is the sample balanced in terms of gender? If not, are there more females or males? The sample size is nearly balanced; however, women do slightly outnumber men in a ratio of approximately 52 to 48.

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) The stacked bar plot above allows us to view the gender ratio of each individual age group. Most categories appear to be relatively balances, although there appear to be some notable exceptions. Columns 6, 7, and 17 are disproportionately male, whereas columns 8 and 9 are mostly comprised of women.

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? As can be seen below, the substance with the earliest apparent use is alcohol, which has the lowest mean, median, and first quarter figure, with the lowest reported minimum age.

```
summary(dat)
```

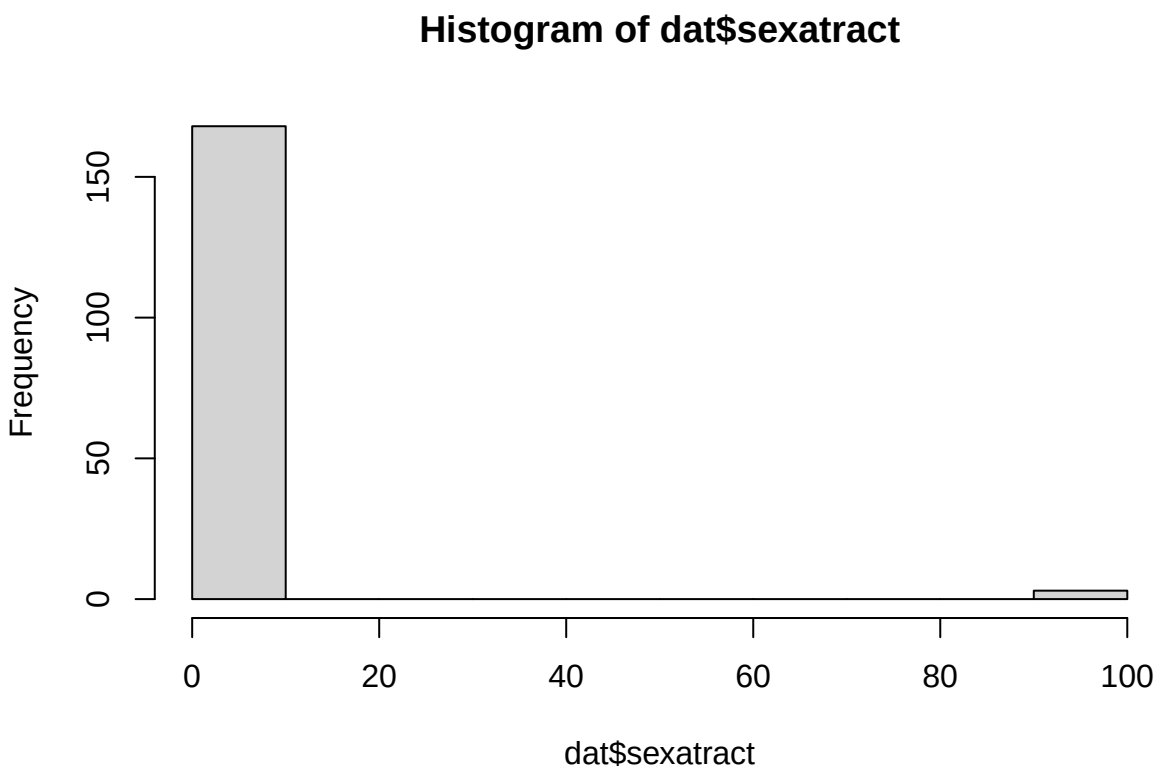
```
##      mjage           cigage         iralcage          age2
##  Min.   : 7.00   Min.   :10.00   Min.   : 5.00   Min.   : 4.00
##  1st Qu.:14.00   1st Qu.:15.00   1st Qu.:13.00   1st Qu.:13.00
##  Median :16.00   Median :17.00   Median :15.00   Median :15.00
##  Mean   :15.99   Mean   :17.65   Mean   :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
```

mjage: minimum age of 7 with a first quarter of 14 and a mean of 15.99. cigage: minimum age of 10 with a first quarter of 15 and a mean of 17.65. iralcage: minimum age of 5 with a first quarter of 13 and a mean of 14.95.

Problem 5: Sexual attraction

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

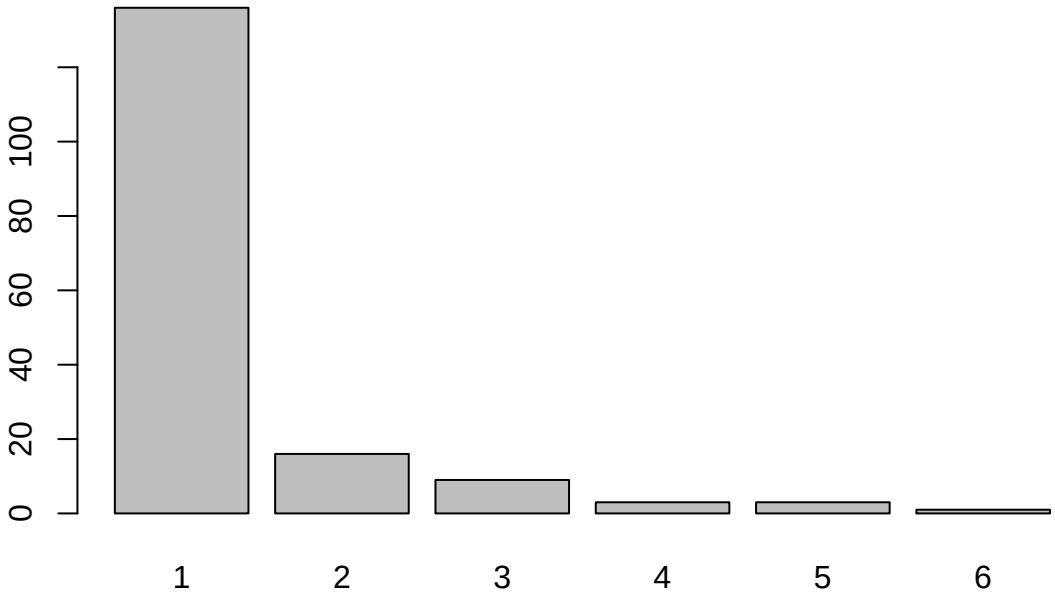
```
hist(dat$sexatract)
```




```
dat$sexatract.nas <- ifelse(dat$sexatract==99, NA, dat$sexatract)
dat$sexatract.nas
```

```
##   [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
```

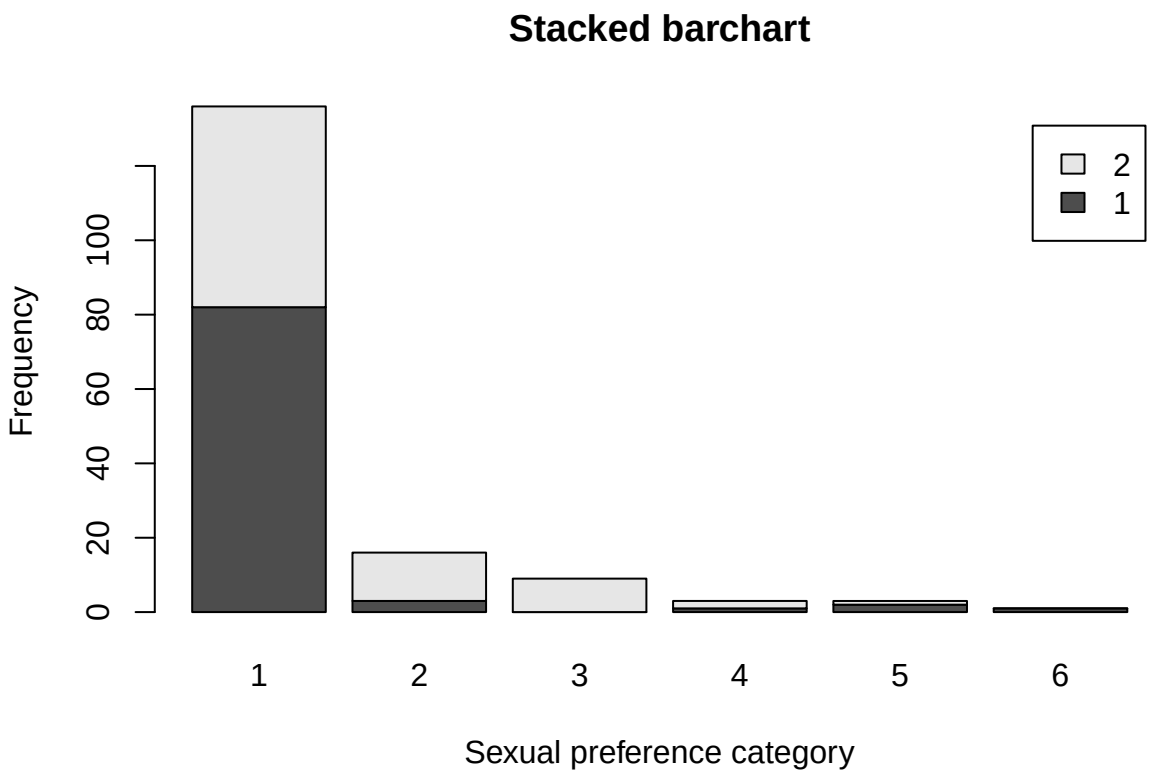
```
sexatract.no.nas <- na.omit(dat$sexatract.nas)
barplot(table(sexatract.no.nas))
```



Yes, the unaltered histogram presents two notable columns, which represent the variables' options of 1-6 and 85, 94, and 97-99. Most participants answered the survey, however nearly a quarter of participants were categorized in the latter portion of answers as they either skipped the question, left the answer blank, or refused. Respondents who participated and answered the question were overwhelmingly attracted to and only to members of the opposite sex; however, small, single-digit percentages did document being attracted to the same sex or both the opposite and same sex. The figures, which may not accurately represent the United States, are expected.

What is the distribution of sexual attraction by gender?

```
tab.prefsex <- table(dat$irsex, dat$sexatract.nas)
barplot(tab.prefsex,
        main = "Stacked barchart",
        xlab = "Sexual preference category", ylab = "Frequency",
        legend.text = rownames(tab.prefsex),
        beside = FALSE)
```

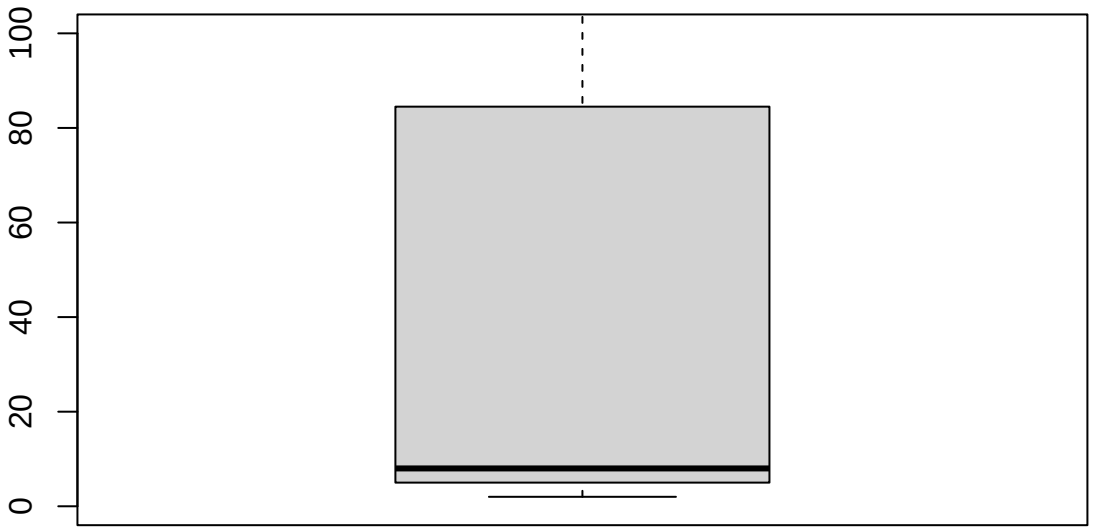


Each category has its own sex ratio. The first, 1, is nearly 2 to 3 men to women, while later options such as 2 and three are disproportionately comprised of women. Later options, such as 5 and 6 are mostly male.

Problem 6: English speaking

What does the distribution of English speaking look like in the sample? Is this what you might expect for a random sample of the US population?

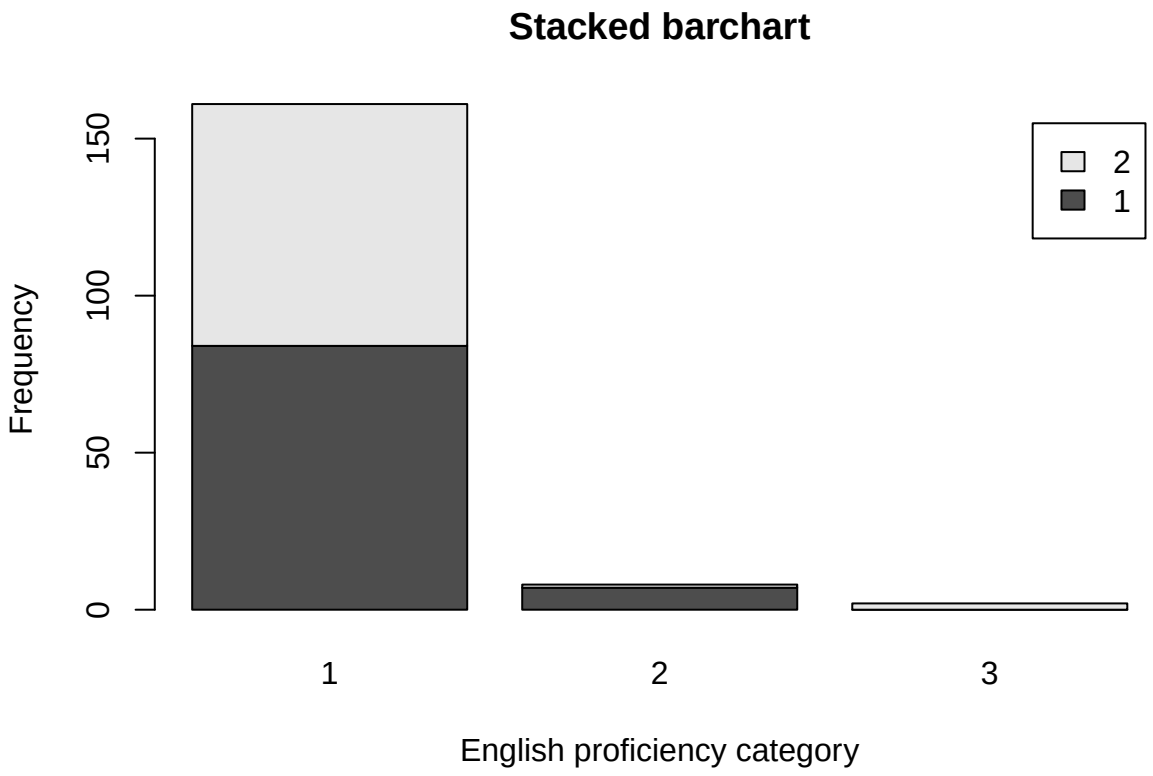
```
boxplot(table(dat$speakengl), ylim = c(0,100))
```



Yes, the vast majority of participants are English proficient with only a very small minority expressing hesitance with the language. The participants may not reflect the US population, however, as communities of particular groups, such as migrants, may be hesitant to participant in a study.

Are there more English speaker females or males?

```
tab.engprof <- table(dat$irsex, dat$speakengl)
barplot(tab.engprof,
        main = "Stacked barchart",
        xlab = "English proficiency category", ylab = "Frequency",
        legend.text = rownames(tab.engprof),
        beside = FALSE)
```



Most participants who answered they were English proficient were male.

Midterm 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("midterm 1/fatal-police-shootings-data.csv")
```

Problem 1

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

The dataset, compiled by the Washington Post, seeks to maintain a record of all fatal police shootings since January 1, 2015. The dataset further elaborates on shooting incidents by tracking various variables, including the race of the deceased, the deceased's mental health statuses, whether the deceased had a gun, and the circumstances surrounding their death.

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

```
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
```

By simply running "dat", we can see that there are 6,594 rows of data with 17 columns. Said columns, which can be identified using the "names(dat)" command are as follows: "id", "name", "date", "manner_of_death", "armed", "age", "gender", "race", "city", "state", "signs_of_mental_illness", "threat_level", "flee", "body_camera", "longitude", "latitude", and"is_geocoding_exact"

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"
```

```
#dat$body_camera
#dat$flee
#dat$armed
```

The prior three listed variables all present various conditions on the shooting's circumstances. "body_camera" explains whether the police officer was wearing a body camera during the interaction with the deceased. "flee" explains whether the deceased was moving away or fleeting from the police officers during the interaction and by what means, as reported by news reports. Lastly, "armed" details whether the victim was armed during the interaction with the officer and what object was believed to be the armed weapon.

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
```

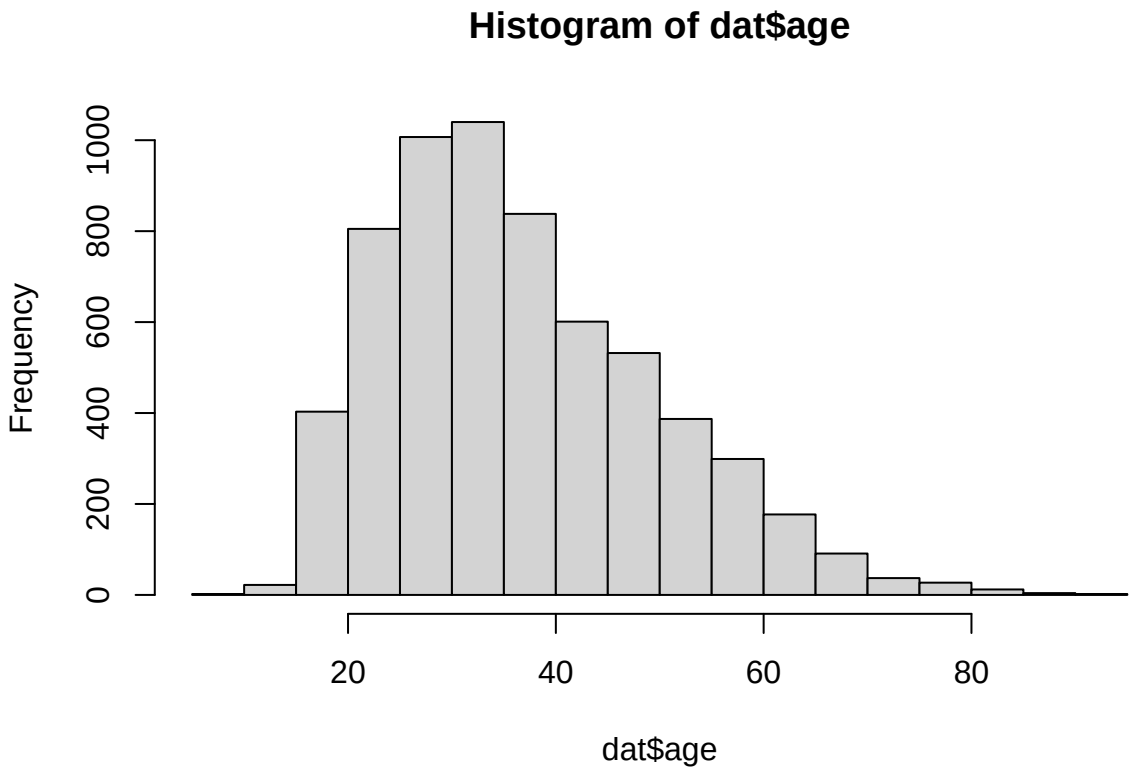
```
##                             1                             2
##                      pick-axe                 piece of wood
##                             4                             7
##                          pipe                     pitchfork
##                             7                             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
```

The objects that most surprised me were a flagpole, binoculars, and air conditioner.

Problem 2

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

```
hist(dat$age)
```



The distribution of dat$age is right-skewed with a median of 35 but a mean of 37.12. This indicates that most victims of police fatality shootings were younger individuals between the ages of 27 and 45, per the first and third percentiles The findings are expected as individuals of this age group are more likely to be out committing crimes and then fleeing or objecting to arrest with weapons, relative to the elderly or very young children. There are also no unrealistic outliers in the dataset.

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

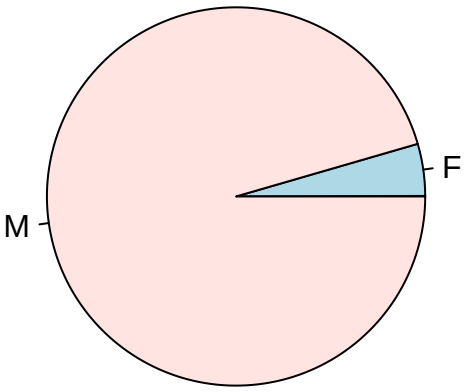
```
summary(dat$age)
```

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

Due to the lack of outliers and the seemingly representative, normal distribution of the data, I would opt for the mean as it would allow for better calculations with age as a quantitative variable. While the distribution is slightly skewed, there are little concerns about the tail variables throwing off the study. In fact, their inclusion would be ideal.

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

```
# table(dat$gender)
pie(table(dat$gender))
```




```
#barplot(table(dat$gender))
```

The distribution of gender is very heavily male-dominate. There are 6298 male incidents of fatal police shootings to 293 for female victims, with 3 cases being outliners. As can be seen in the pie chart, males are the vast majority of victims in police shootings. Women are approximately 4.65% of cases while men are 95. 35% of cases. I opted for a pie chart over a barplot to highlight the overwhelming ratio; however, both graphs can be used for gender/sex comparisons.

Problem 3

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
```

As reported by R, 5684 police officers were not wearing body cameras while only 910 were; that is, approximately 86.2% of police officers in all of the Washington Post's documented fatal shootings were not wearing body cameras, while only 13.8% were indeed wearing body cameras. The report is surprising given how widespread body camera use has become nationwide. As an individual, I am also conflicted that such a large percentage of police shootings went undocumented by the police officers' account. That presents issues for accountability assuming police officers were acting in ill standards. The percentage may be slim but for 86.2% of all documented fatal shootings to be unrecorded from the police officer's point of view is troubling.

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
```

The table command indicates that there were 3952 cases of victims not fleeing, with 1058 incidents of victims fleeing in a car, 845 via foot, or 248 by other means for a total of 2151 fleeing individuals. The final proportions are thus 59.9% of victims not fleeing and 32.6% of victims fleeing either via car, foot, or other means. As a note, there are 491 cases of missing variables. The findings are therefore puzzling as only approximately a third of fatal police shooting victims were fleeing. The approximate percent of 59.9% of victims who were not fleeing is troubling from the civilian's perspective.

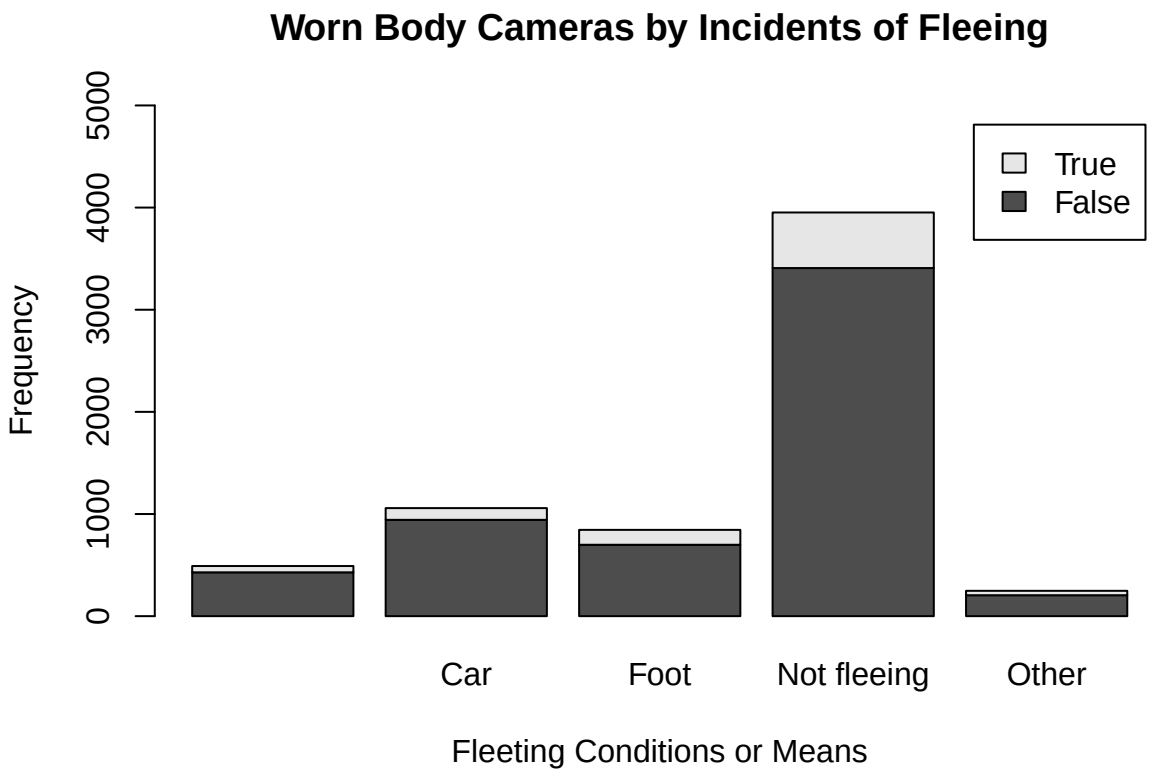
Problem 4

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.

```
# dat$body_camera
# dat$flee
tab.cameraflee <- table(dat$body_camera, dat$flee)
barplot(tab.cameraflee,
        main = "Worn Body Cameras by Incidents of Fleeing",
        xlab = "Fleeting Conditions or Means", ylab = "Frequency",
        legend.text = rownames(tab.cameraflee),
        ylim = c(0, 5000),
        beside = FALSE)
```



The relationship between variables body_camera and fleeing detail what proportion of incidents involving victims fleeing form the police were recorded by police body cameras. As can be noted, police body cameras were not present across all incidents of fleeing or the lack thereof as well. Most cases involving police body cameras were those where the victim was not fleeing. As a note, the column furthest on the left includes incidents of missing variables where the fleeing condition is unknown. We still see here that police body cameras were still not widely used.

Extra credit

a. What does this code tell us?

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

The code presents an ordinal sequence for the dates in the dataset, during which all incidents of police shootings were captured. "mydates <- as.Date(dat$date)" reformates the "date" variable as dates in R, with "head(mydates)" presenting a brief summary of the first few entires. The command "(mydates[length(mydates)] - mydates[1])" then subtracts the final date entry from the first in order to conclude how many dates are included overall in the dataset. There are 2458 days for a total of 6,594 fatal police shootings.

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 shootings may be mislabelled or underreported due to the presence of missing variables. For values such as fleeing, armed, and race, there are various missing variables. The result creates some uncertainty about specific statistical findings. The main reason, however, for why cases may be underreported or mislabelled has to do with the lack of a concrete dataset. Various institutions and organizations nationwide document fatal police shootings but, as stated, the figures are non-exhaustive. The Washington Post has alone documented more than twice the recorded fatal police shooting incidents than the FBI and the Centers for Disease Control and Prevention. Existing datasets are further jeopardized by the inconsistency of local police office in incident reporting as each department and office can report cases distinctly.

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?

Yes, there are approximately 500 missing fleeing values in problem 4. I am unable to determine, however, if the missing variables are solely absent from the fleeing variable or if they have broader absent values as well. I will note, however, that the variable "body_camera" has 6594 cases and "gender" has 6591 complete variables. Their near completion indicates that some variables are indeed exhaustive or missing just 3 entries. Thus, I am unable to determine whether the missing fleeing variables are an anomaly or whether data is consistently absent.

```
#table(dat$gender)
#table(dat$body_camera)
```

Assignment 3

Load the data.

```
library(readr)
library(knitr)
dat <- read.csv(file = "assignment 3/crime_simple.txt")
```

This is a dataset from a textbook by Brian S. Everitt about crime in the US in 1960. The data originate from the Uniform Crime Report of the FBI and other government sources. The data for 47 states of the USA are given.

Codebook:

R: Crime rate: # of offenses reported to police per million population

Age: The number of males of age 14-24 per 1000 population

S: Indicator variable for Southern states (0 = No, 1 = Yes)

Ed: Mean of years of schooling x 10 for persons of age 25 or older

Ex0: 1960 per capita expenditure on police by state and local government

Ex1: 1959 per capita expenditure on police by state and local government

LF: Labor force participation rate per 1000 civilian urban males age 14-24

M: The number of males per 1000 females

N: State population size in hundred thousands

NW: The number of non-whites per 1000 population

U1: Unemployment rate of urban males per 1000 of age 14-24

U2: Unemployment rate of urban males per 1000 of age 35-39

W: Median value of transferable goods and assets or family income in tens of $

X: The number of families per 1000 earning below 1/2 the median income

We are interested in checking whether the reported crime rate (# of offenses reported to police per million population) and the average education (mean number of years of schooling for persons of age 25 or older) are related.

Problem 1.

How many observations are there in the dataset? To what does each observation correspond?

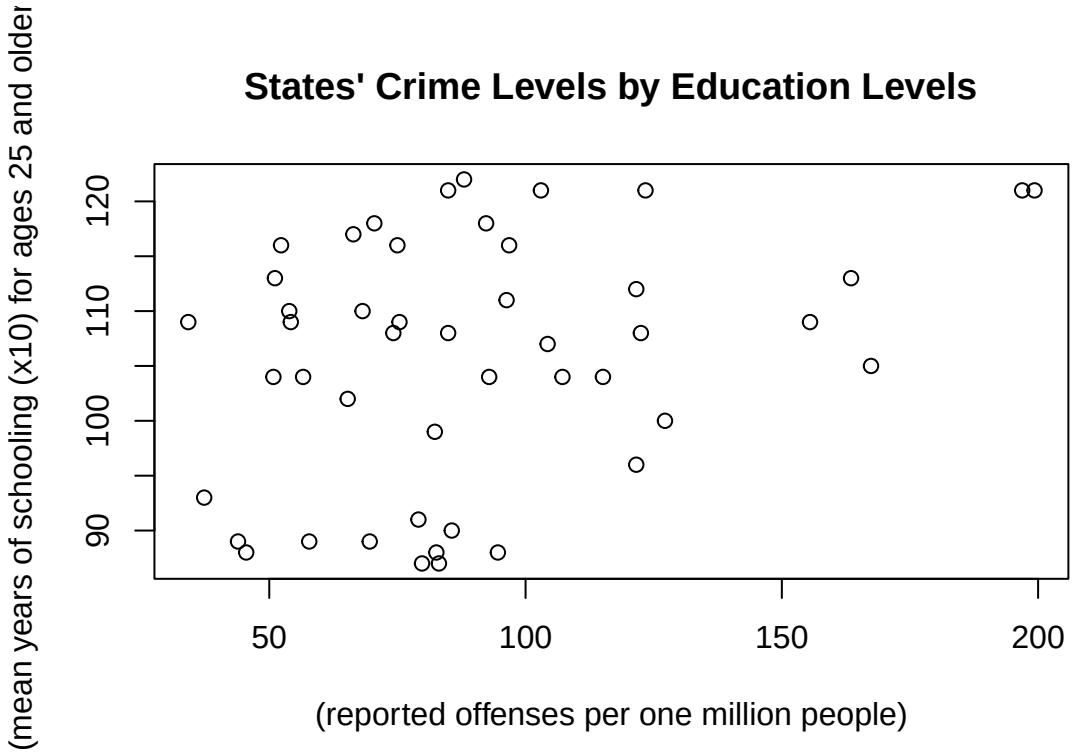
```
dat.crime
```

The overall "dat" dataset has fifty rows of data, with each row corresponding to a state. With the four variables of murder, assault, urban population, and rape, as well as an identifying state name value, the dataset lists rape and murder rates, as well as total cases of assault, in relation to states' urban populations. In contrast, the "dat.crime" dataset has 47 rows of data with 14 variables, or columns. Each row once again corresponds to a state's reported crime rate with their respective demographic information.

Problem 2.

Draw a scatterplot of the two variables. Calculate the correlation between the two variables. Can you come up with an explanation for this relationship?

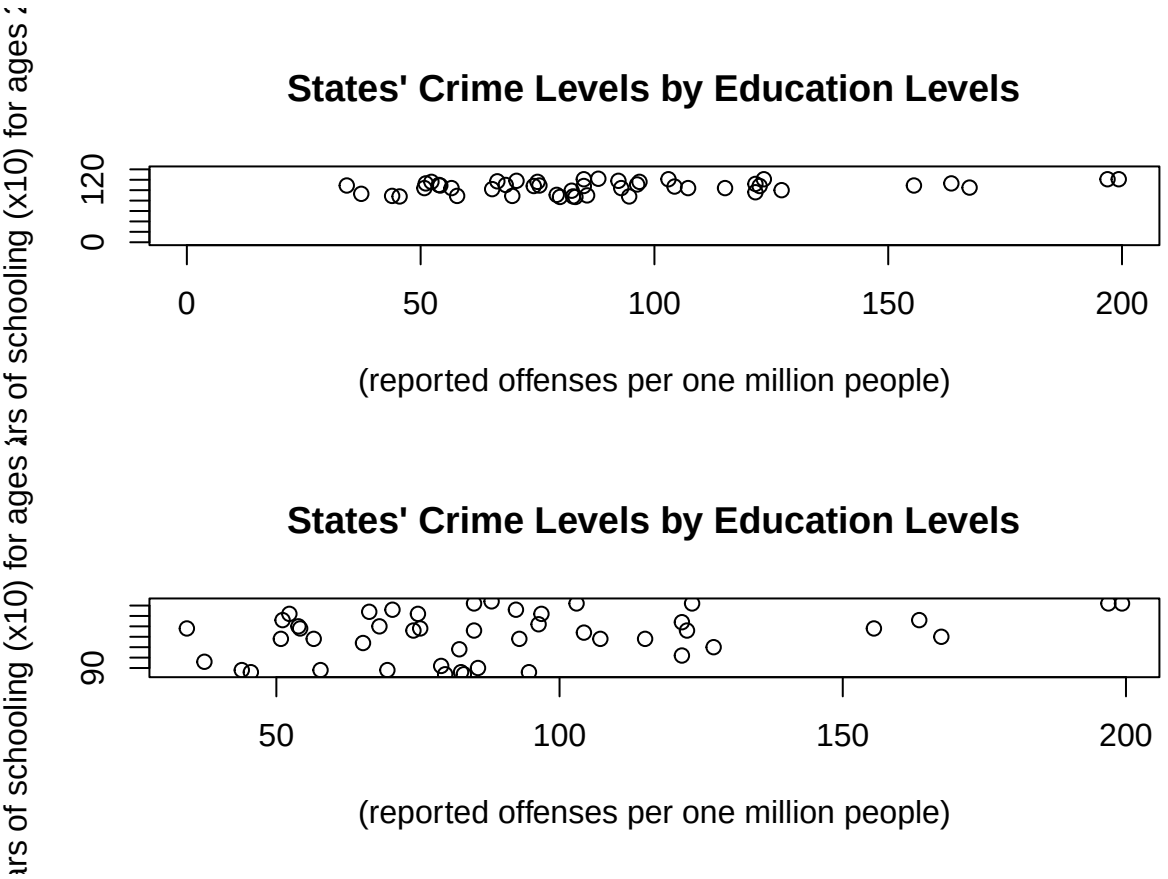
```
crime.lm <- plot(dat.crime$R, dat.crime$Ed,  main="States' Crime Levels by Education Levels",
    xlab="(reported offenses per one million people)", ylab="(mean years of schooling (x10) for ages 25
```




```
cor(dat.crime$R, dat.crime$Ed)
```

```
## [1] 0.3228349
```

```
par(mfrow=c(2,1))
plot(dat.crime$R, dat.crime$Ed,  main="States' Crime Levels by Education Levels",
    xlab="(reported offenses per one million people)", ylab="(mean years of schooling (x10) for ages 25
plot(dat.crime$R, dat.crime$Ed,  main="States' Crime Levels by Education Levels",
    xlab="(reported offenses per one million people)", ylab="(mean years of schooling (x10) for ages 25
```



The calculated correlation between years of education and reported crime is 0.3228349. The relationship between the two variables thus appears to be low and statistically insignificant as the calculated value is a mere 0.3228349. The existing correlation may be a consequence of external forces. Disparities in schooling requirements may for example skew both variables as residents of particularly large or small states may be required to continue attending or deterred from remaining in school. The differences in social action may contribute to the formulation of an insignificant correlation. Visually, as indicated in the first graph, most plots appear to be scattered randomly across the determined scales. Fitting a regression line would be difficult for such a distribution. While the distribution does appear to become more narrowed once the scale is made to include 0 as the minimum value for the x- and y-axis, the distribution appears to be a consequence of the graph's manipulation rather than the variable's relationship.

Problem 3.

Regress reported crime rate (y) on average education (x) and call this linear model `crime.lm` and write the summary of the regression by using this code, which makes it look a little nicer `{r, eval=FALSE} kable(summary(crime.lm)$coef, digits = 2)`.

```
dat.crime$R.c = scale(dat.crime$R, center=TRUE, scale=FALSE) # scaling the data
crime.lm <- lm(R ~ Ed, data = dat.crime) # running regression
summary(crime.lm) # calling the summary of the fitted model
```

```
##
## Call:
## lm(formula = R ~ Ed, data = dat.crime)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.061 -27.125  -4.654  17.133  91.646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.3967    51.8104  -0.529   0.5996
## Ed            1.1161     0.4878   2.288   0.0269 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.01 on 45 degrees of freedom
## Multiple R-squared:  0.1042, Adjusted R-squared:  0.08432
## F-statistic: 5.236 on 1 and 45 DF,  p-value: 0.02688
```

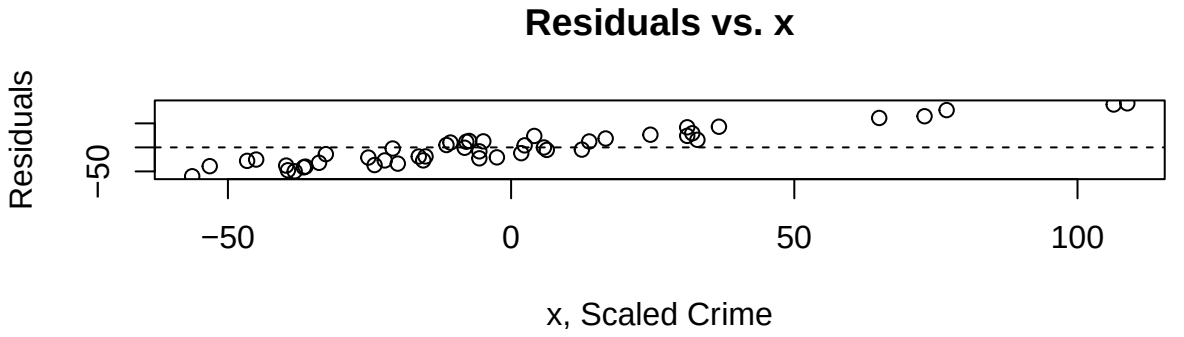
```
kable(summary(crime.lm)$coef, digits = 2) # making the summary look nicer
```

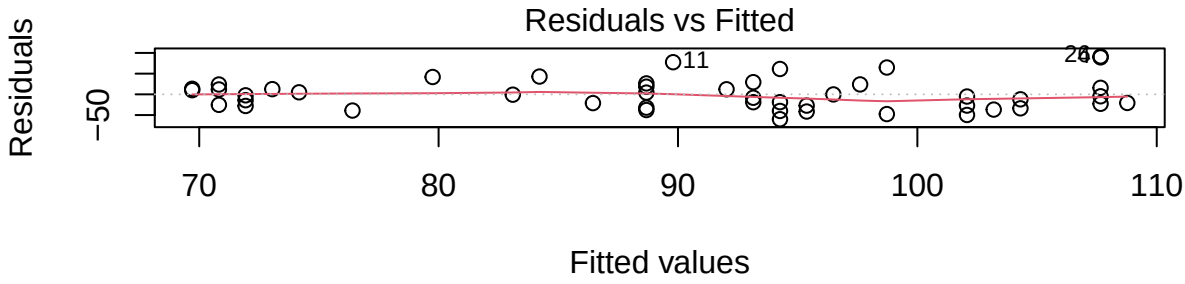
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.40	51.81	-0.53	0.60
Ed	1.12	0.49	2.29	0.03

Problem 4.

Are the four assumptions of linear regression satisfied? To answer this, draw the relevant plots. (Write a maximum of one sentence per assumption.) 1. The linearity assumption is not satisfied as the residuals versus x graph depicts a patterned upward distribution, which should be flat and roughly evenly distributed; the residuals versus fitted plot also does not follow a straight line.

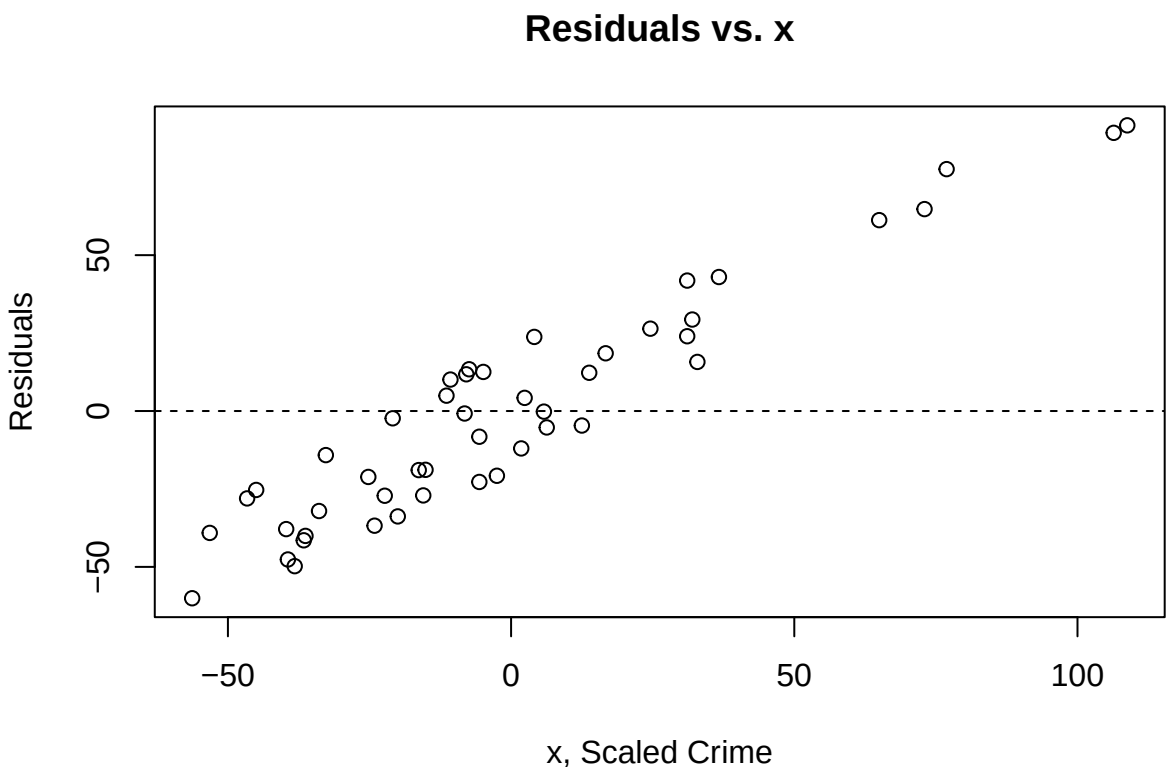
```
par(mfrow=c(2,1))
plot(dat.crime$R.c, crime.lm$residuals, main="Residuals vs. x", xlab="x, Scaled Crime", ylab="Residuals
abline(h = 0, lty="dashed")
plot(crime.lm, which=1)
```





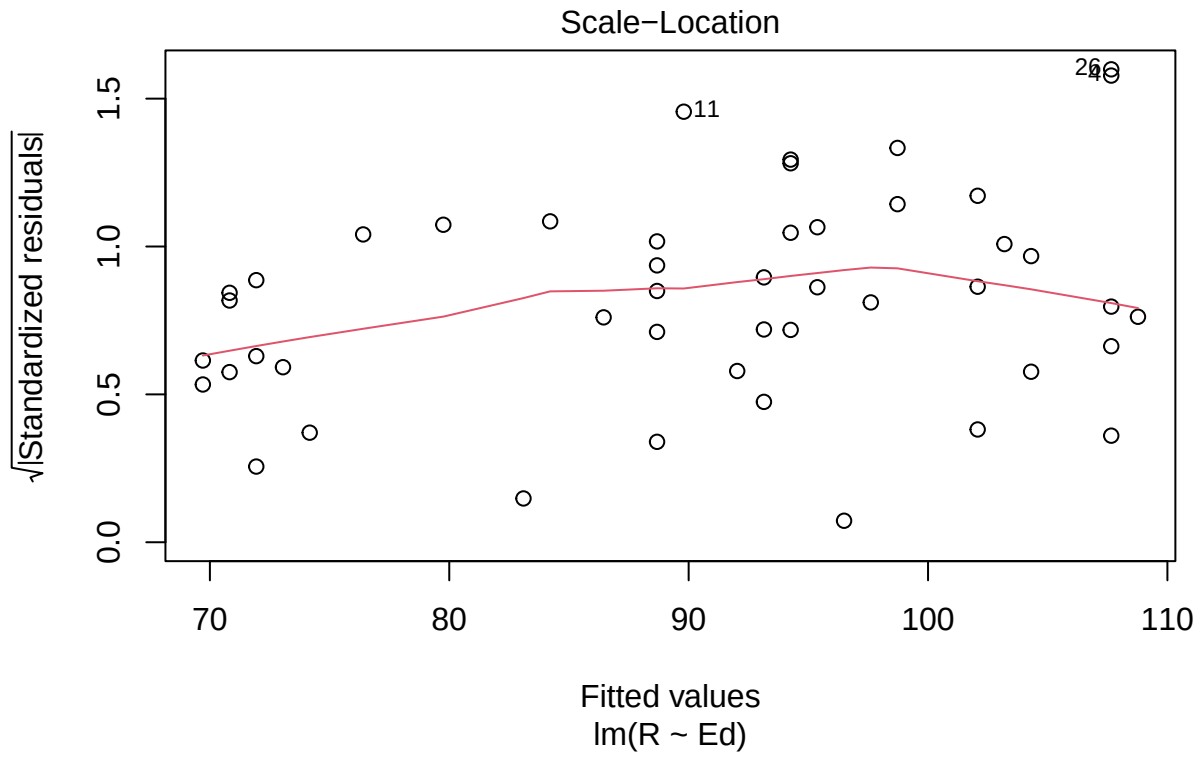
2. The independence assumption is not satisfied as, once again, a pattern can be noted in the distribution for residuals versus x.

```
plot(dat.crime$R.c, crime.lm$residuals, main="Residuals vs. x", xlab="x, Scaled Crime", ylab="Residuals"
abline(h = 0, lty="dashed")
```



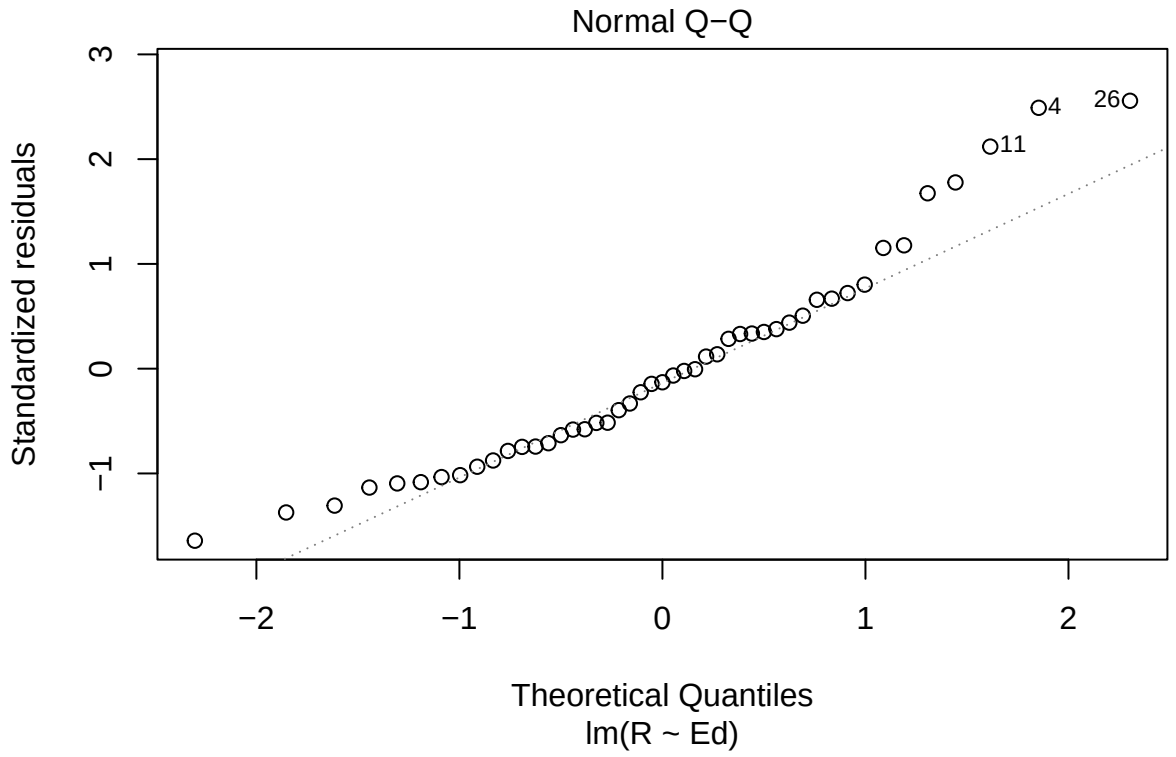
3. The equal variance assumption/homoscedasticity is not satisfied as the line is curved in the scale location plot and appears to fan out from a few relatively constant variables near the x value of 80 to a wide distribution where x>85.

```
plot(crime.lm, which=3)
```

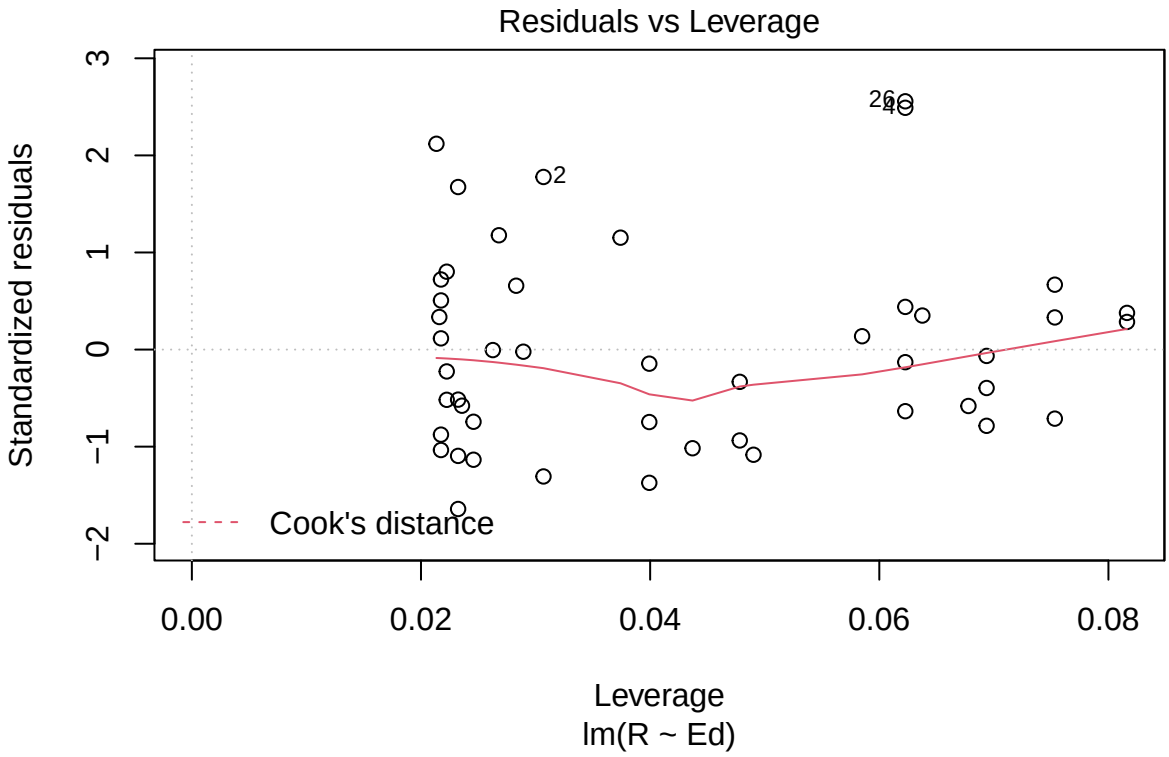


4. The normal population assumption is not satisfied as both ends of the QQ plot appear to be heavily skewed leftward of the fitted line.

```
plot(crime.lm, which=2)
```




```
plot(crime.lm, which=5)
```



Problem 5.

Is the relationship between reported crime and average education statistically significant? Report the estimated coefficient of the slope, the standard error, and the p-value. What does it mean for the relationship to be statistically significant?

The relationship between reported crime and average education is not statistically significant, with a reported estimated coefficient of 1.1161 for the slope, 0.4878 for the standard error, and 0.0269 for a one star p-value. In order for the relationship to be significant, the error value would have to be lower with a smaller decimal for the p-value, as well as a greater slope marking a stronger or more drastic relationship between the two variables. Had the variables had appropriate coefficients, then education level and reported crime rate may have had a correlatory relationship; however, such assessment would require additional information and studies.

Problem 6.

How are reported crime and average education related? In other words, for every unit increase in average education, how does reported crime rate change (per million) per state?

The calculated correlation between 1 year of education, expressed in values of ten, is 1.1161 per 100 million people in a given state. The finding would suggest 1 marginal year of education would equate to a 8.96 million increase in additional reported crimes.

Problem 7.

Can you conclude that if individuals were to receive more education, then crime will be reported more often? Why or why not?

As stated, the relationship between the two variables appears to be low and reinstates the importance of differentiating correlation and causation, as well as the importance of assessing findings. The estimated coefficients not only indicate a weak statistical relationship but the derived weak correlation could be affected by external social forces. Determining any positive or negative relationship would thus be illogical. You cannot conclude that if individuals were to receive more education, then crime would be reported at higher rates.

Midterm 2

Problem 1: EDA

Describe the dataset and variables. Perform exploratory data analysis for the two variables of interest: funds and po.brut.

```
dat <- read.csv(file = 'midterm 2/sim.data.csv')
head(dat)
```

```
##   po.dept.code funds po.brut
## 1            1  48.1      23
## 2            2  81.4      10
## 3            3  41.8      25
## 4            4  61.7      19
## 5            5  86.4       8
## 6            6  51.6      22
```

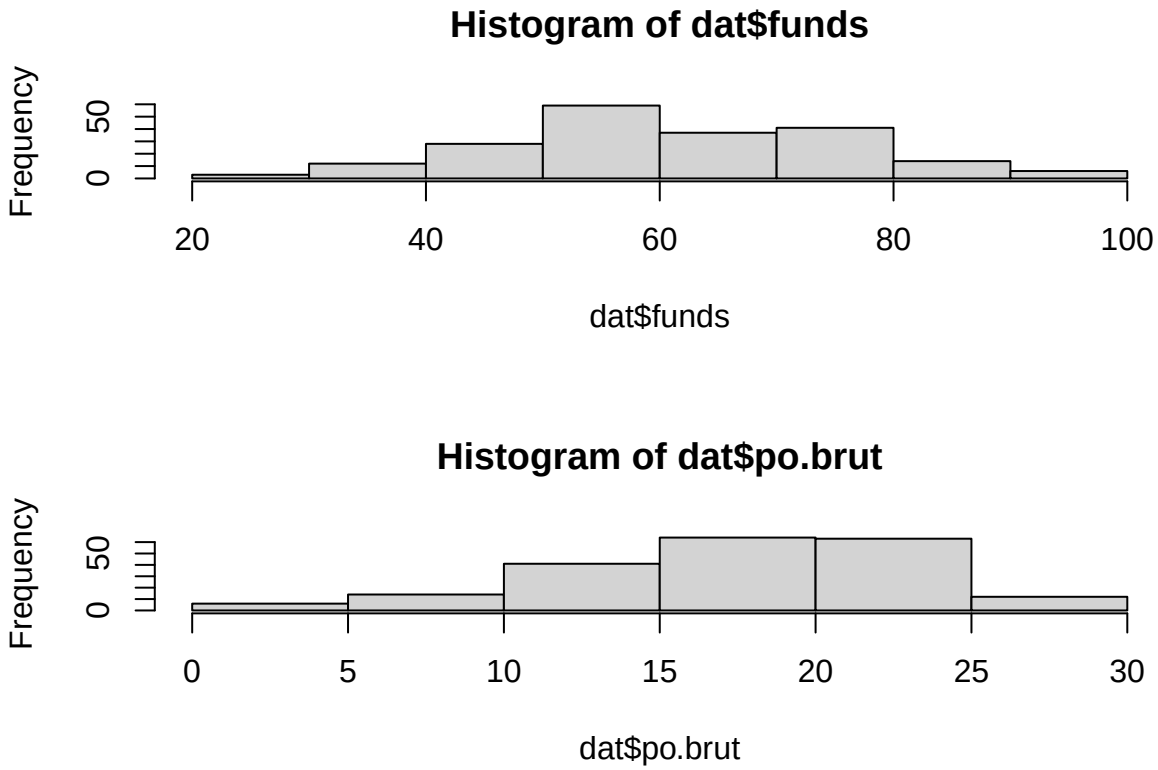
```
summary(dat$funds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.40   51.67   59.75   61.04   72.17   99.70
```

```
summary(dat$po.brut)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   14.00   19.00   18.14   22.00   29.00
```

```
par(mfrow=c(2,1))
hist(dat$funds)
hist(dat$po.brut)
```



The dataset contains 200 rows of data with three columns: po.dept.code, funds, and po.brut. The first variable serves as an identifer, providing each studied police department with a unique variable for later reference. The second deliniates how much funding the respective police department recieves, with the last being the incidents of police brutality. The early data analysis indicates there are an average of 61 million dollars provided in funds to police departments and 18 cases of police brutality. Further, as indicated by the histograms, both variables appear to have normal distributions with no significant outliers.

Problem 2: Linear regression

a. Perform a simple linear regression to answer the question of interest. To do this, name your linear model "reg.output" and write the summary of the regression by using "summary(reg.output)".

```
reg.output <- lm(po.brut ~ funds, data = dat)
summary(reg.output)
```

```
##
## Call:
## lm(formula = po.brut ~ funds, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9433 -0.2233  0.2544  0.5952  1.1803
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.543069   0.282503  143.51   <2e-16 ***
## funds       -0.367099   0.004496  -81.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9464 on 198 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.971
## F-statistic:  6666 on 1 and 198 DF,  p-value: < 2.2e-16
```

b. Report the estimated coefficient, standard error, and p-value of the slope. Is the relationship between funds and incidents statistically significant? Explain.

Estimated coefficient: -0.367099

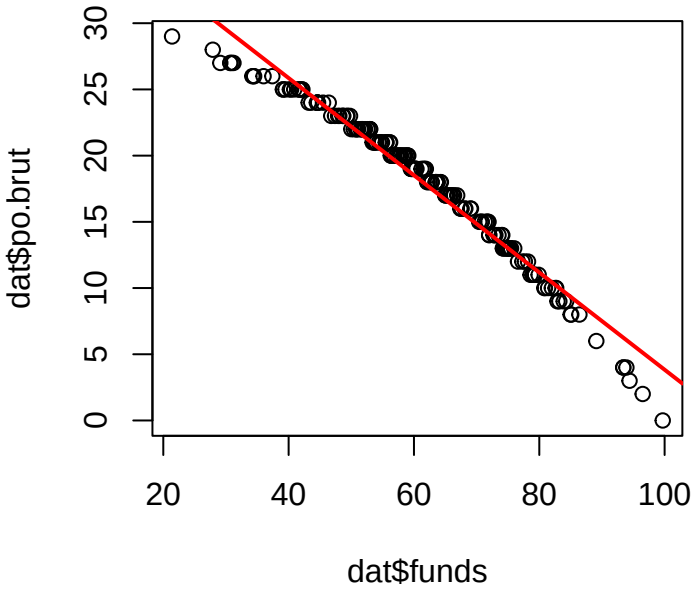
Estimated standard error: 0.004496

Estimated p-value of the slope: <2e-16 ***

According to the estimated results for the linear regression, it appears we can reject the null hypothesis as there does indeed seem to be a significant correlation between increased marginal funding and decreasing incidents of police brutality. If noted, the estimated t-value is -81.64, which is significantly removed from the value of 0, thus allowing us to reject the null hypothesis. Further, with an estimated t value less than 2e-16, the relationship was generated a three star significance code, the highest of significant codes for correlation between two variables. The value is additionally complimented by a very low estimated standard error value, indicating a significant relationship. However, the difference between correlation and causation must be highlighted once again. It may be possible that the marginal funds provided are not causing the decrease in incidents, but rather the funding of possible programs that is enabled by additional funding. That is to say, there does appear to be a relationship between funding and incidents; yet, the relationship may not be causational and, thus, the correlation must be assessed critically.

c. Draw a scatterplot of po.brut (y-axis) and funds (x-axis). Right below your plot command, use abline to draw the fitted regression line, like this:

```
# Remember to remove eval=FALSE!!
plot(dat$funds, dat$po.brut)
abline(reg.output, col = "red", lwd=2)
```



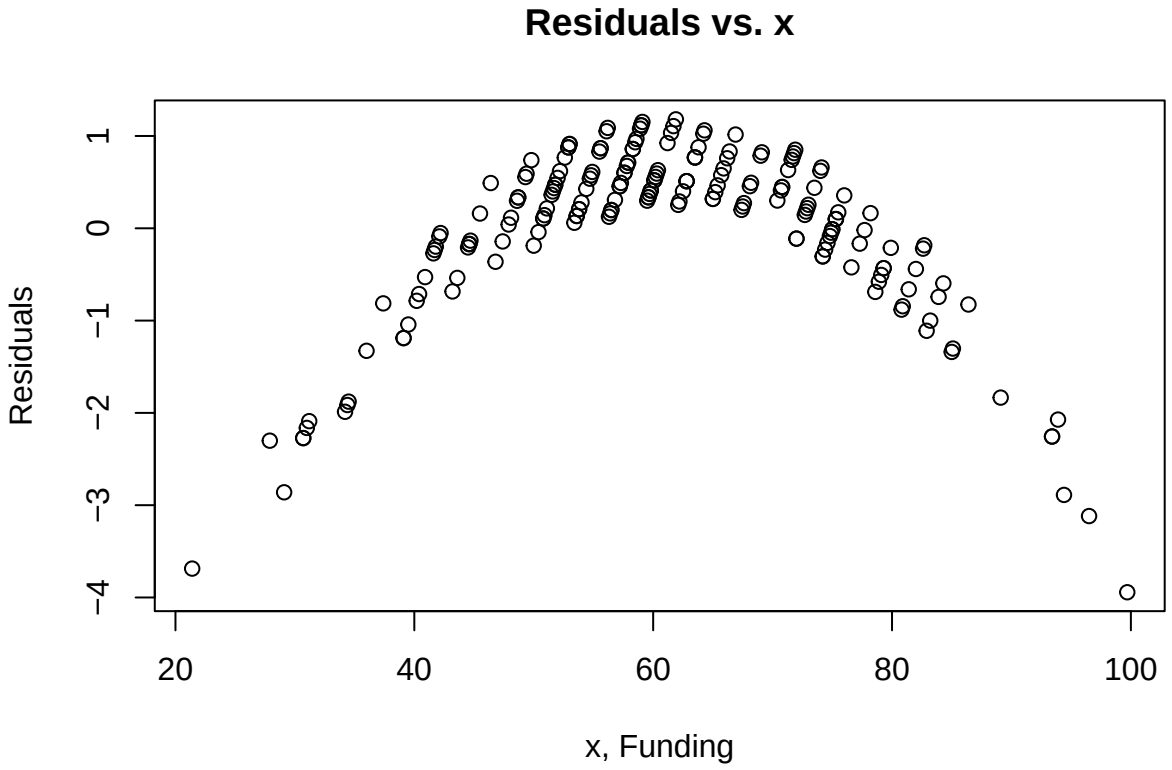
Does the line look like a good fit? Why or why not?

The fitted line does appear to be a good fit. While there are a few outliers, particularly in the data's rightward and leftward values, most entries generally fall close near the produced line. The slope of the fitted line additionally follows the general distribution of the dataset.

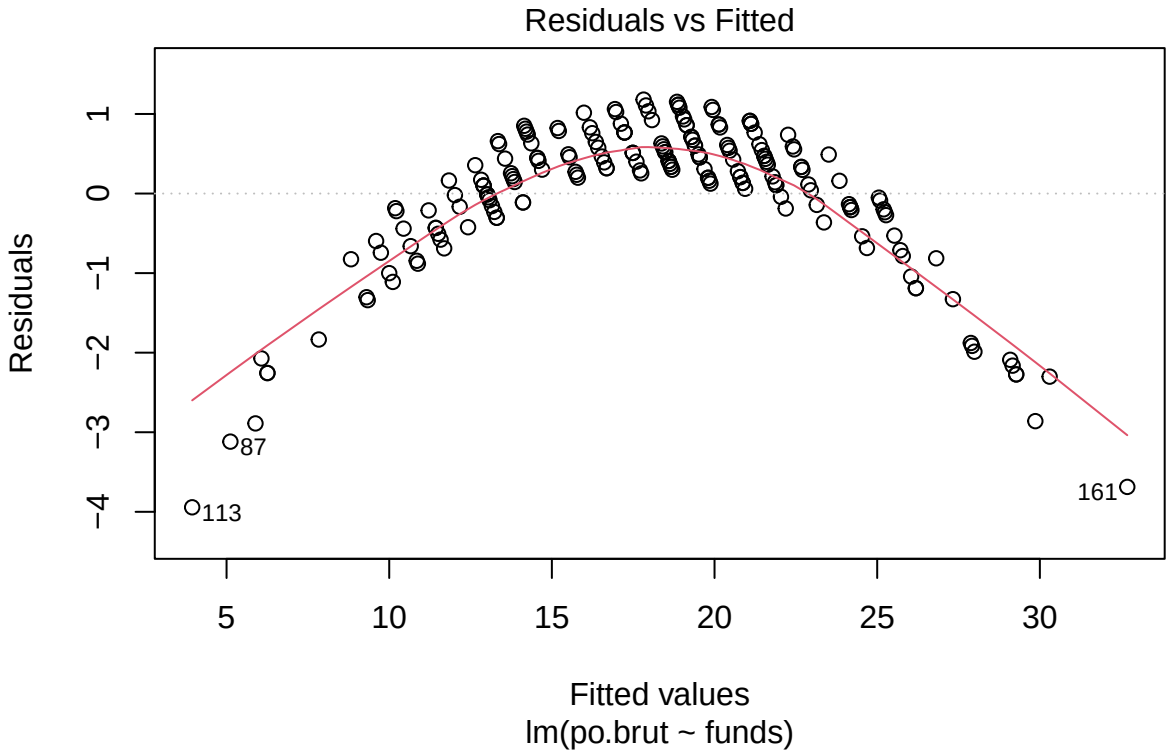
d. Are the four assumptions of linear regression satisfied? To answer this, draw the relevant plots. (Write a maximum of one sentence per assumption.) If not, what might you try to do to improve this (if you had more time)?

1) Linearity assumption is not satisfied as the first graph, residuals vs x, depict a patterned distribution when it should be flat and the second graph, residuals vs fitted, has a significantly curved line that should be flat.

```
plot(dat$funds, reg.output$residuals, main="Residuals vs. x", xlab="x, Funding", ylab="Residuals")
```

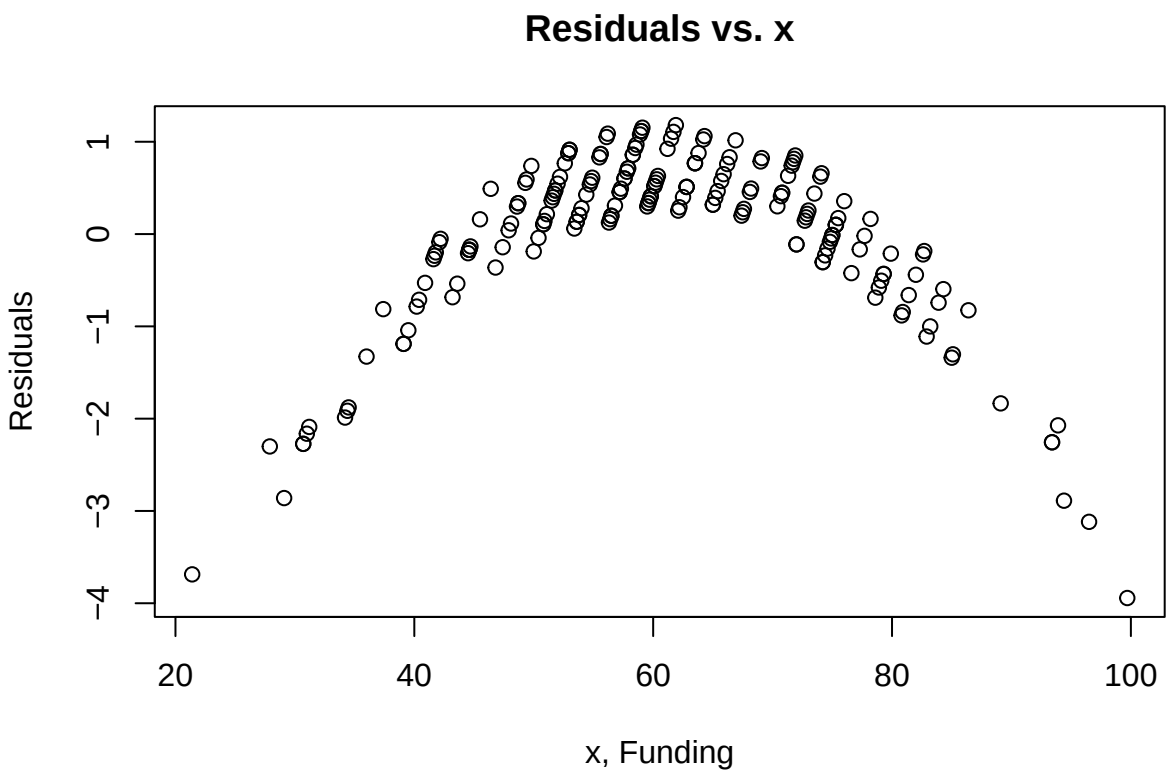



```
plot(reg.output, which = 1)
```



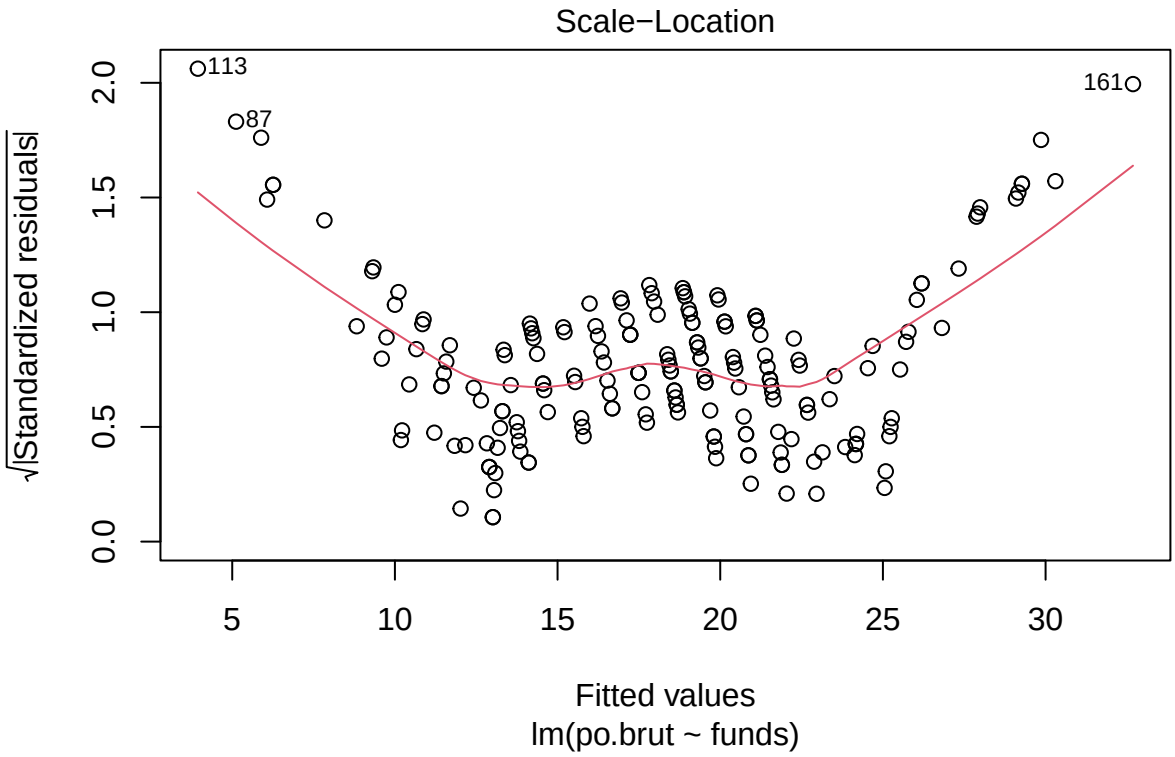
2) The independence assumption is not satisfied as the produced graph, residuals vs x, displays a pattern between the two variables, suggesting a failure of independence.

```
plot(dat$funds, reg.output$residuals, main="Residuals vs. x", xlab="x, Funding", ylab="Residuals")
```



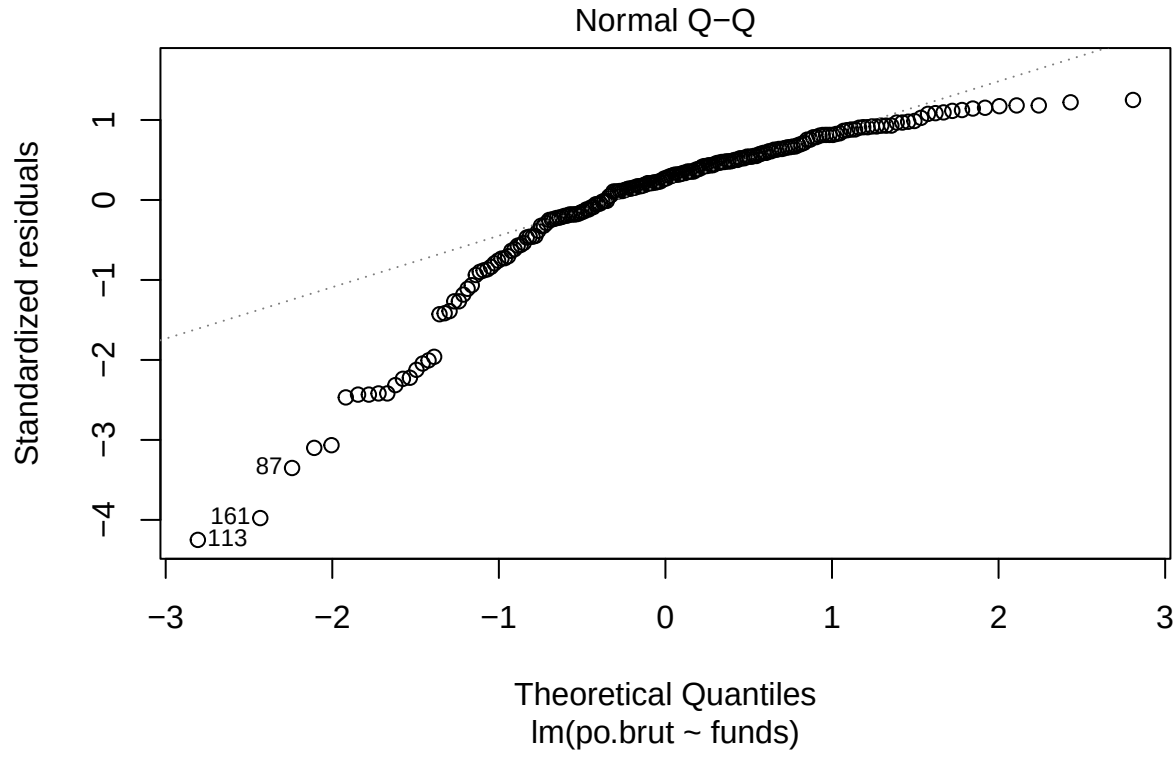
3) The equal variance assumption, or homoscedasticity, is not satisfied as the produced red line features significant trends, informing us that the residuals and hence errors have non-constant variance.

```
plot(reg.output, which = 3)
```



4) The normal population assumption does not appear to be satisfied as both ends of the QQ plot appear to be skewed, particularly the rightward portion which features significant outliers.

```
plot(reg.output, which = 2)
```



e. Answer the question of interest based on your analysis.

Question of interest: does having more funding in a police department lead to fewer incidents of police brutality?

Answer: The data has indicated that no significant relationship can be derived from the increased marginal funding and incidents of police brutality. While the initial linear regression estimated a significant relationship between the two variables, the four assumptions all failed to be satisfied. The incident highlights the importance of both running the linear regression model and assessing all four assumptions. This case may be the result of estimated correlatory figures depicting a significant relationship with erroneous sample data. We can thus justify the acceptance of the null hypothesis, despite a remarkably low t value. With none of the four assumptions being satisfied, my previous remarks on the estimated relationship must be rescinded. We accept the null hypothesis and remain skeptical of a bilateral relationship between police funding and incidents of police brutality.

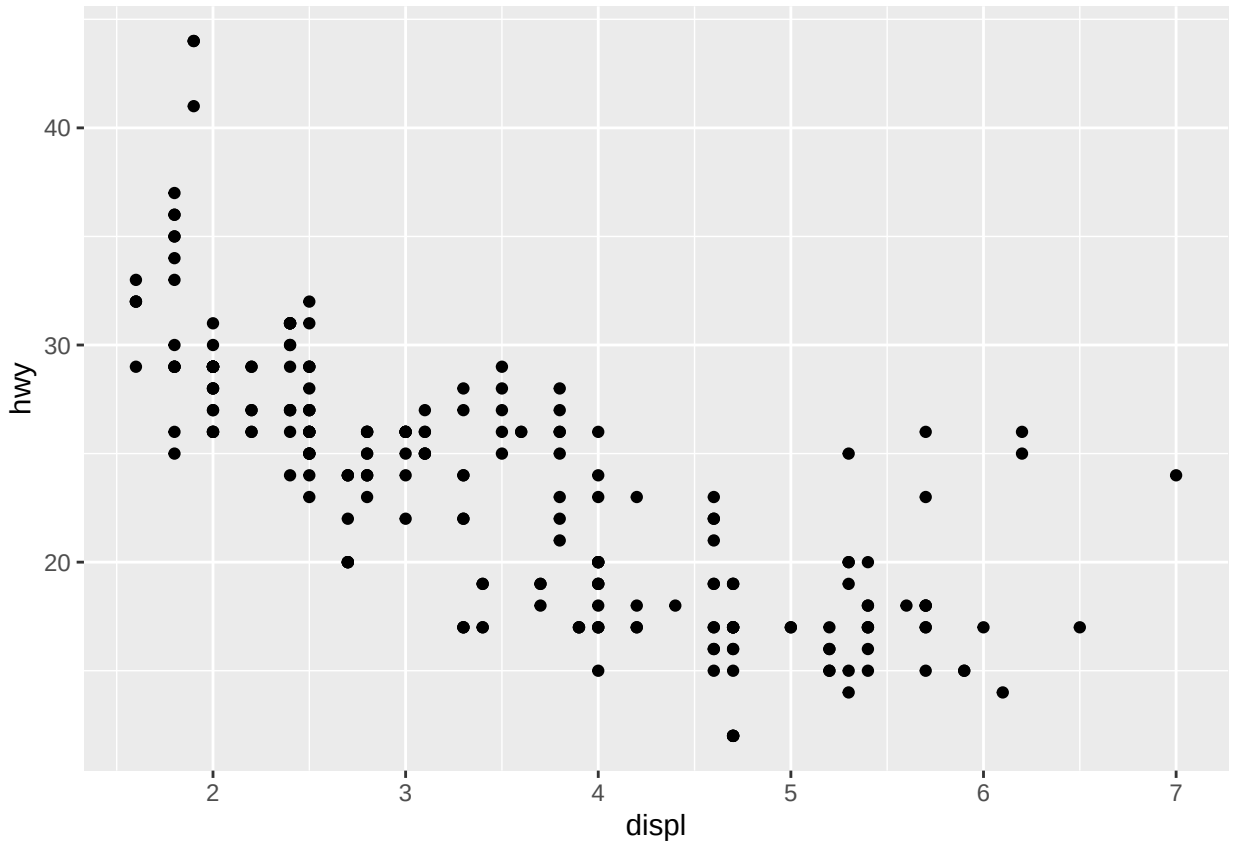
Problem 3:

Describe the dataset. Considering our lecture on data ethics, what concerns do you have about the dataset? Once you perform your analysis to answer the question of interest using this dataset, what concerns might you have about the results?

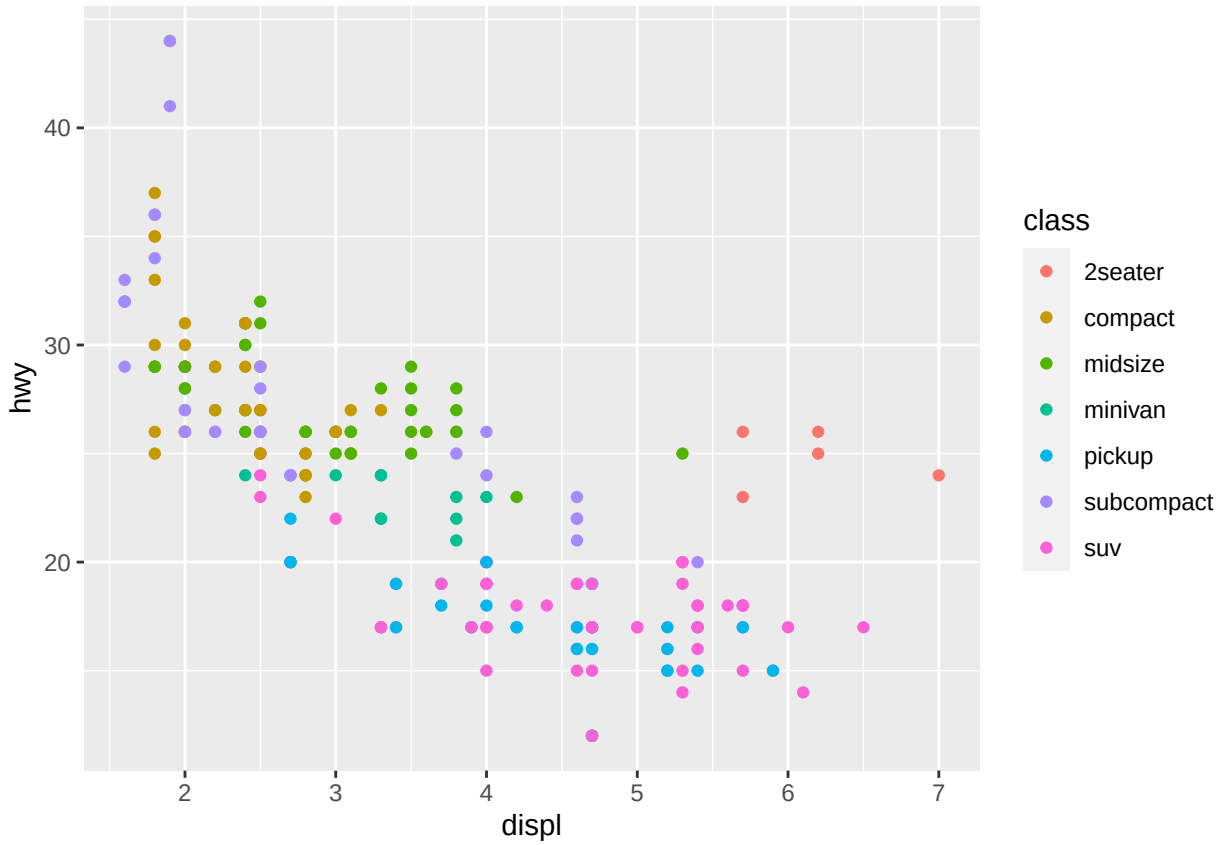
While assessing the data, I had various concerns regarding the dataset. For starters, there are continued issues in police reporting as law enforcement departments are often incentivized to underreport cases of brutality, skewing the figures. In fact, these very skewed figures may explain why the four tested assumptions failed despite a significant estimated relationship between the funding and incident variables. A reason behind the assumptions' failures may be that the data itself was errneous, which could occur if figures were intentionally misrepresented. The result is most concerning as an initial linear regression assessment could have led may, particularly those who failed to assess the assumptions, to believe there is indeed a relationship between the two variables. The finding could have translated into policy, which would have been founded on ill-assessed asssessments.

Assignment 4

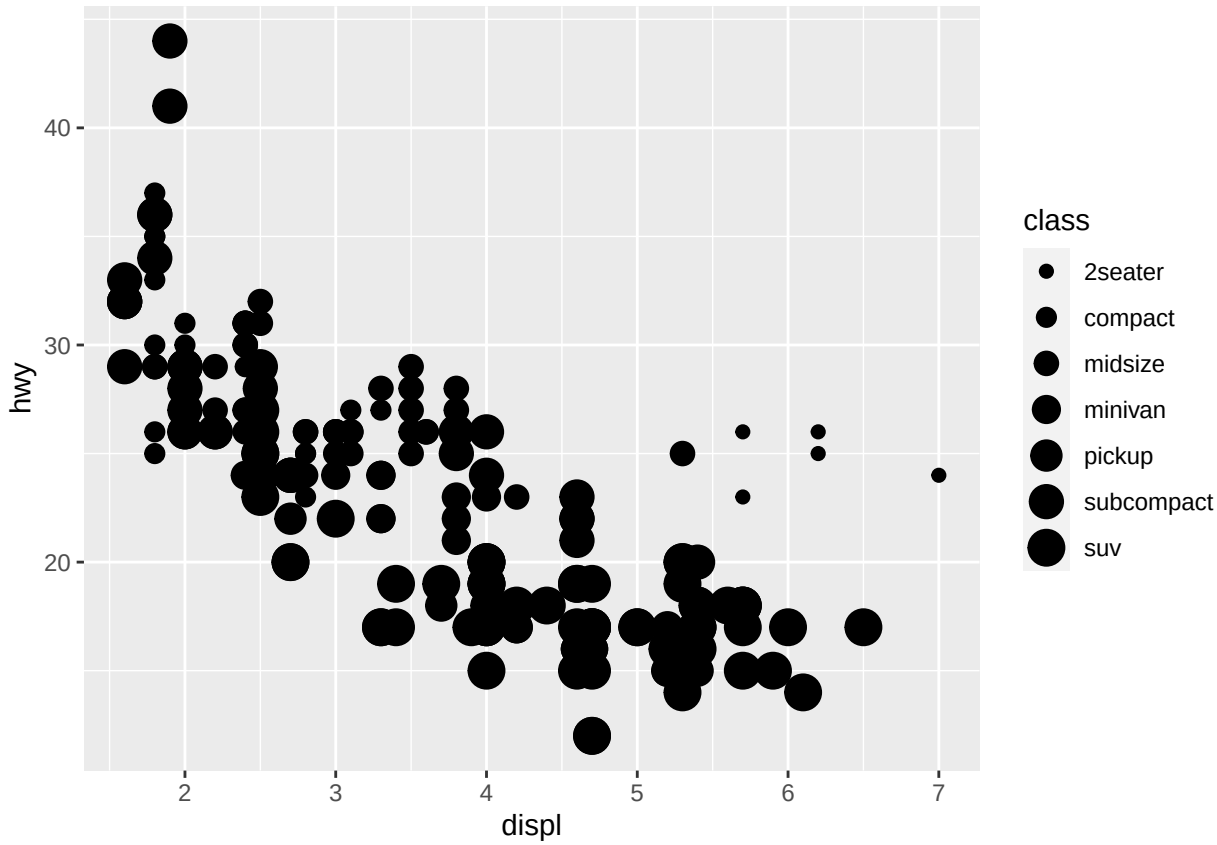
```
library(tidyverse)
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) ## makes a scatterplot
```



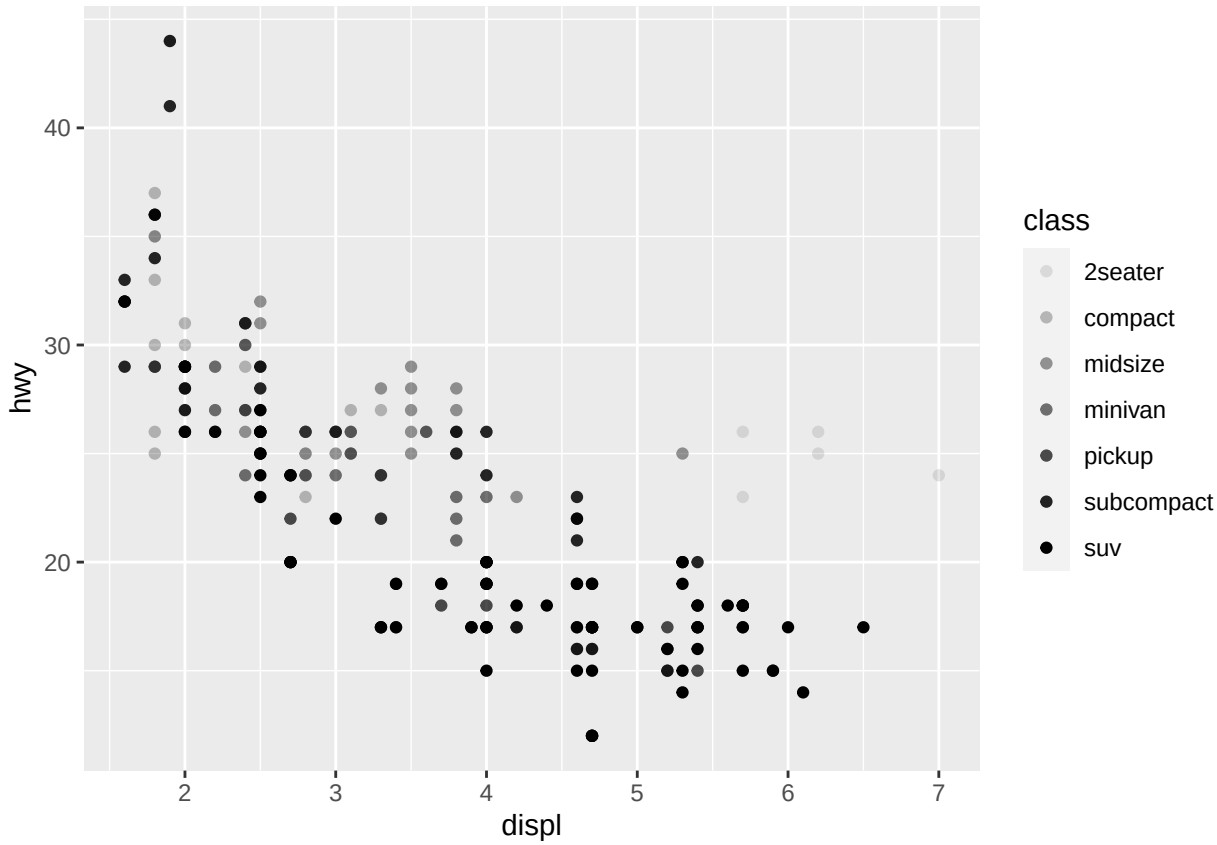

```
## template: ggplot(data = <DATA>) + <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>)); + goes on first row

ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = class)) ## aes variable allows for differentiate
```



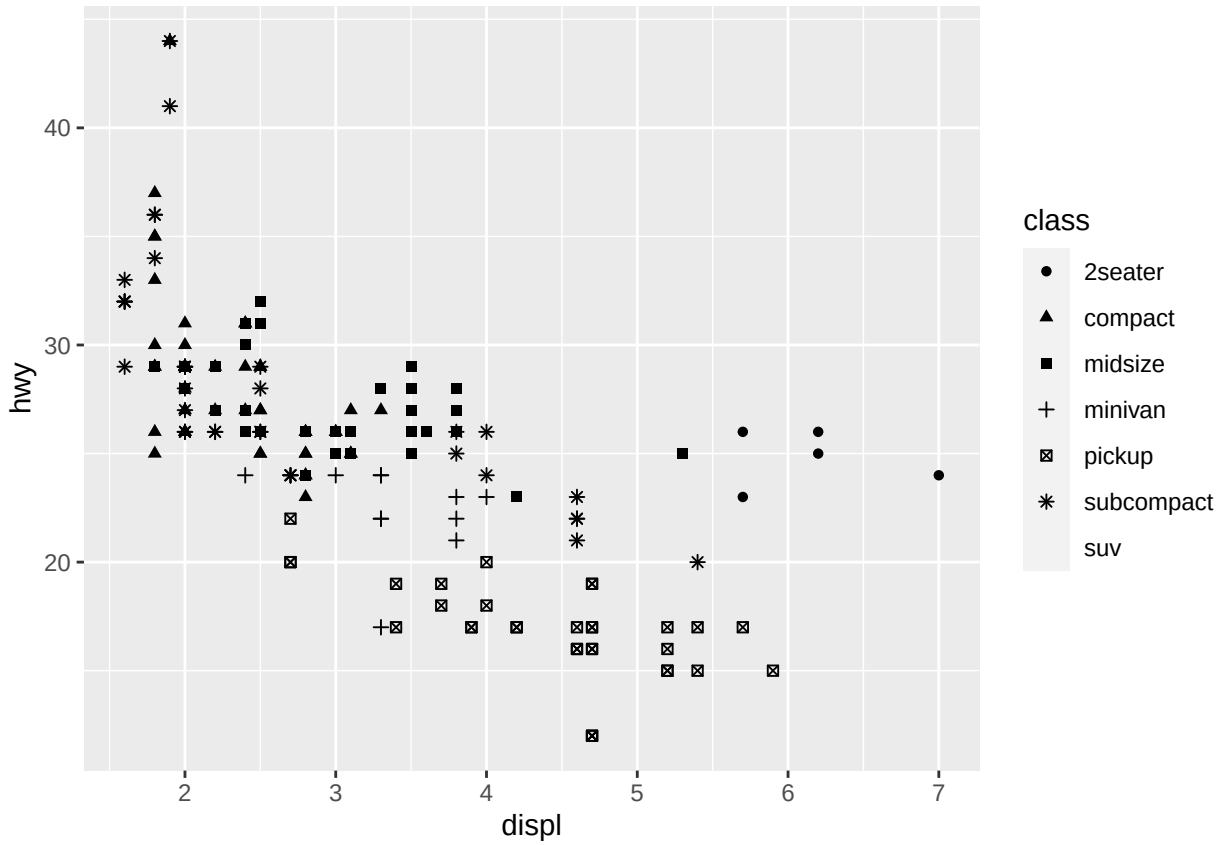

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = class)) ## use size as a class, rather than color
```



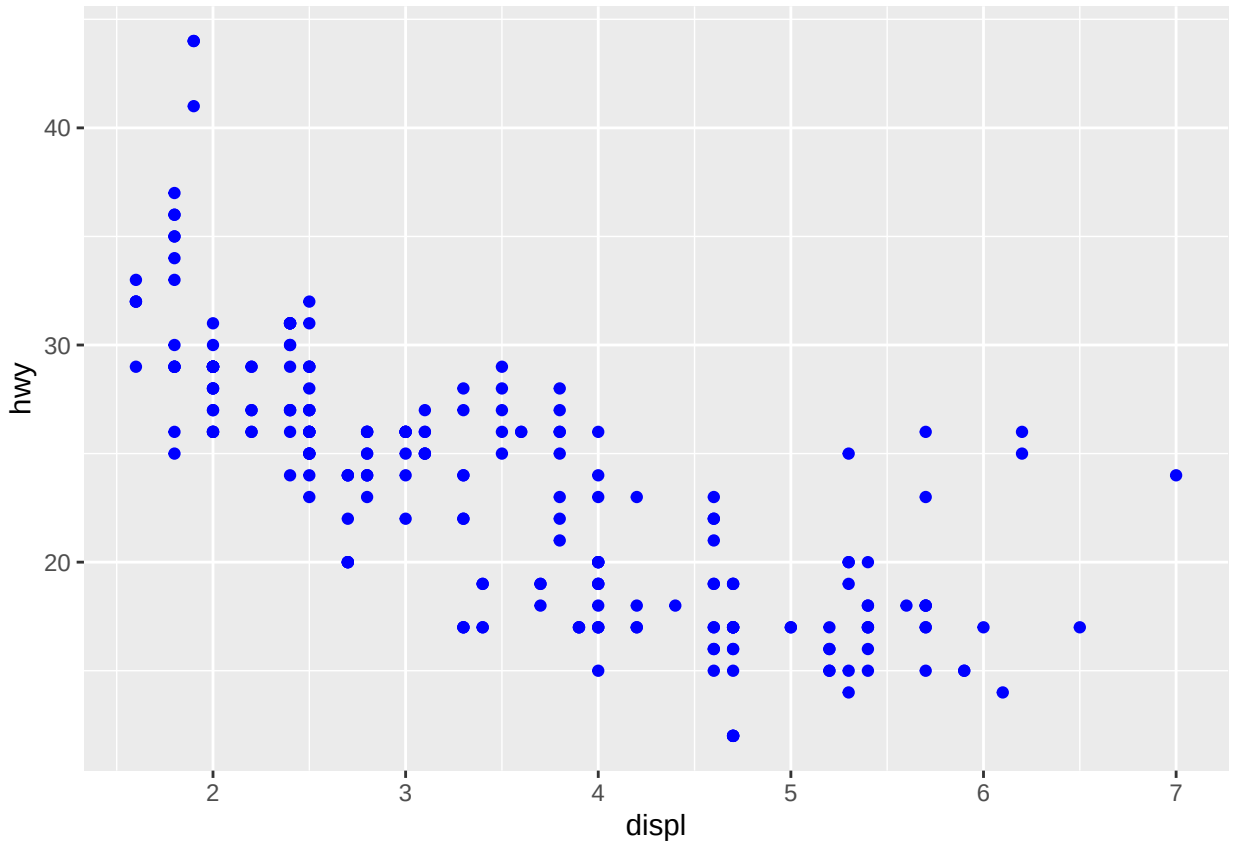

```
# Left: alpha
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, alpha = class)) ## alpha class plots with transparency
```



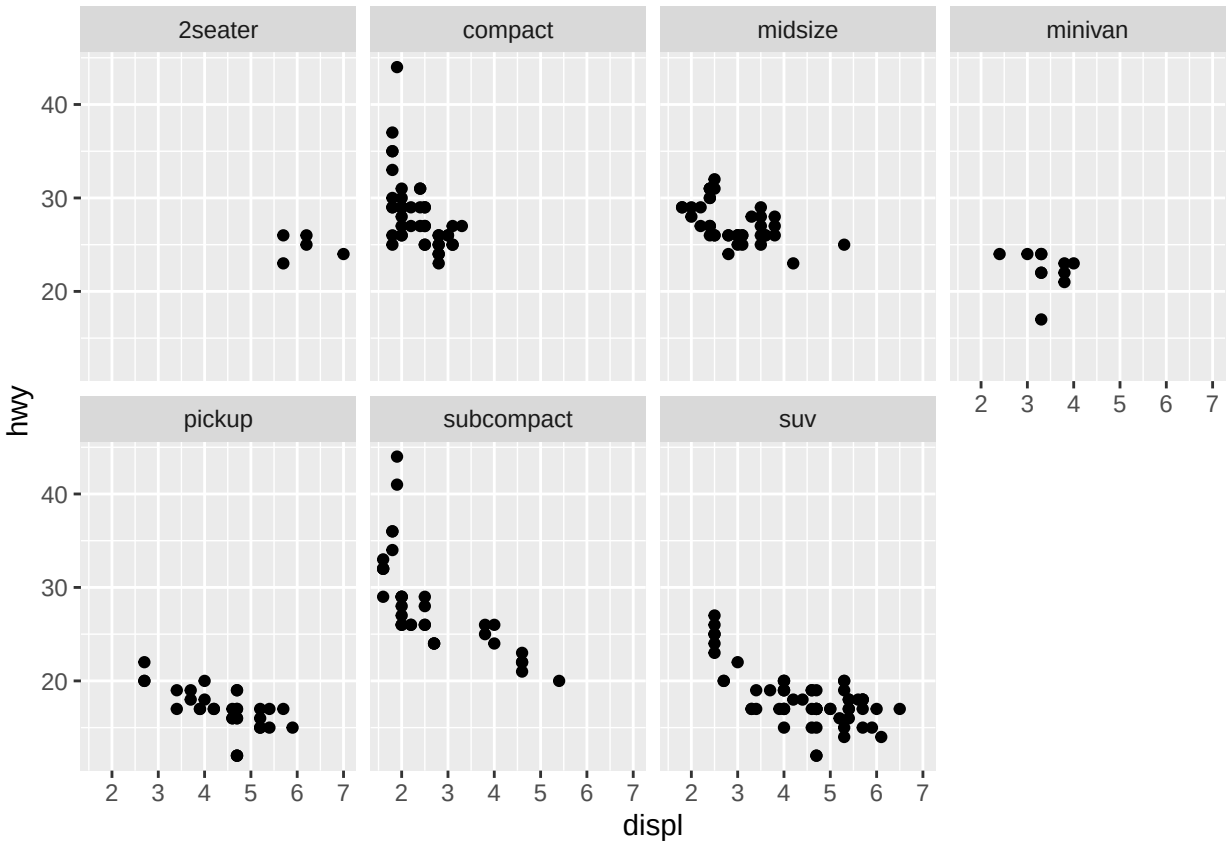

```
# Right
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, shape = class)) ## shape class plots with distinct shape
```



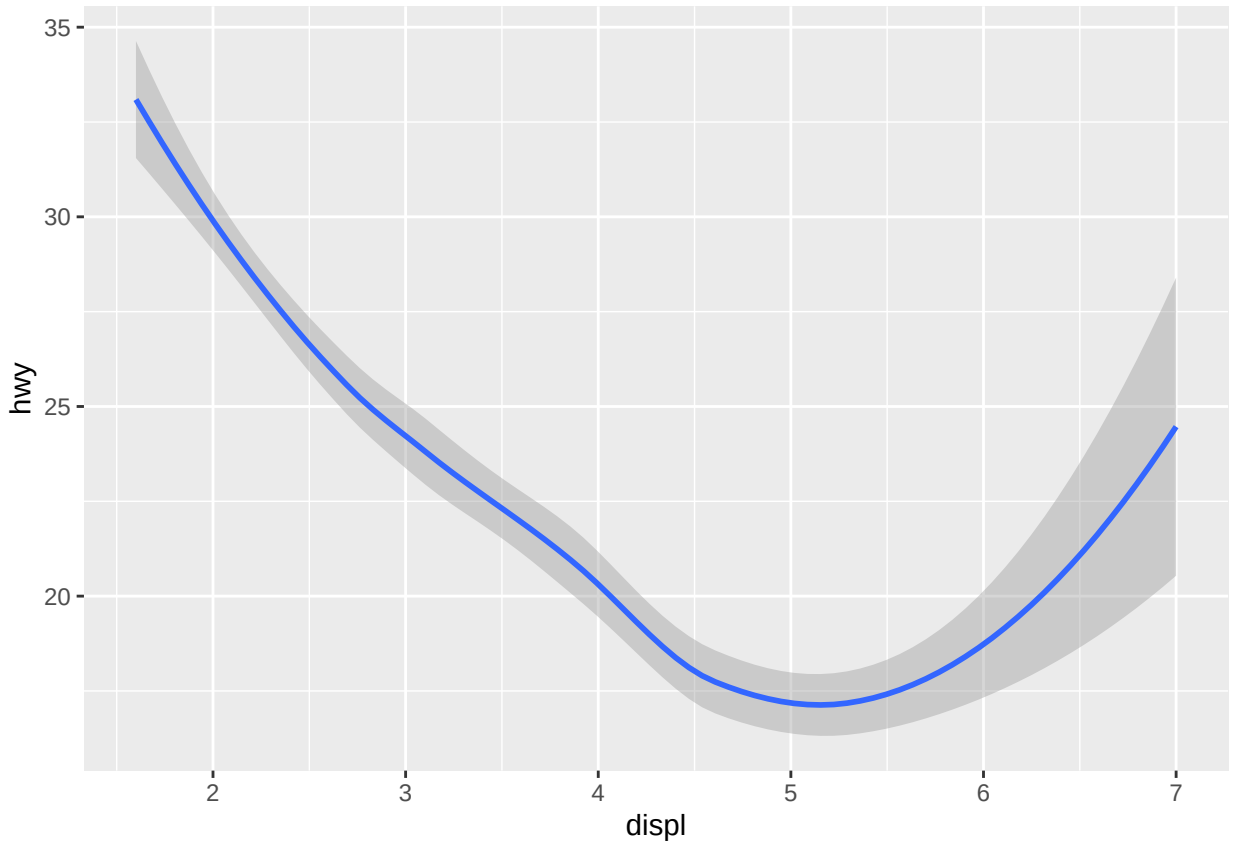

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy), color = "blue") ## color goes outside of the aes comman
```



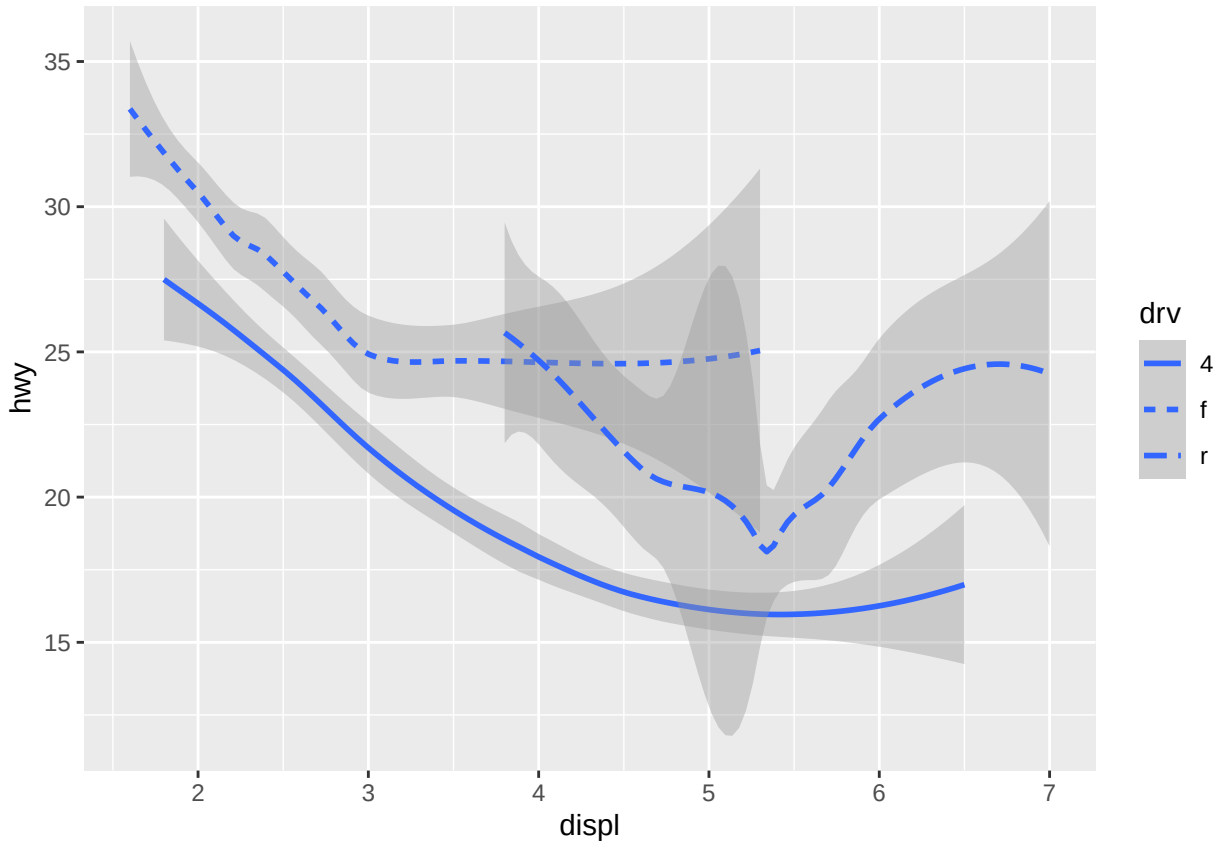

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +  ##create a facet formula; distinct than grid
  facet_wrap(~ class, nrow = 2)
```



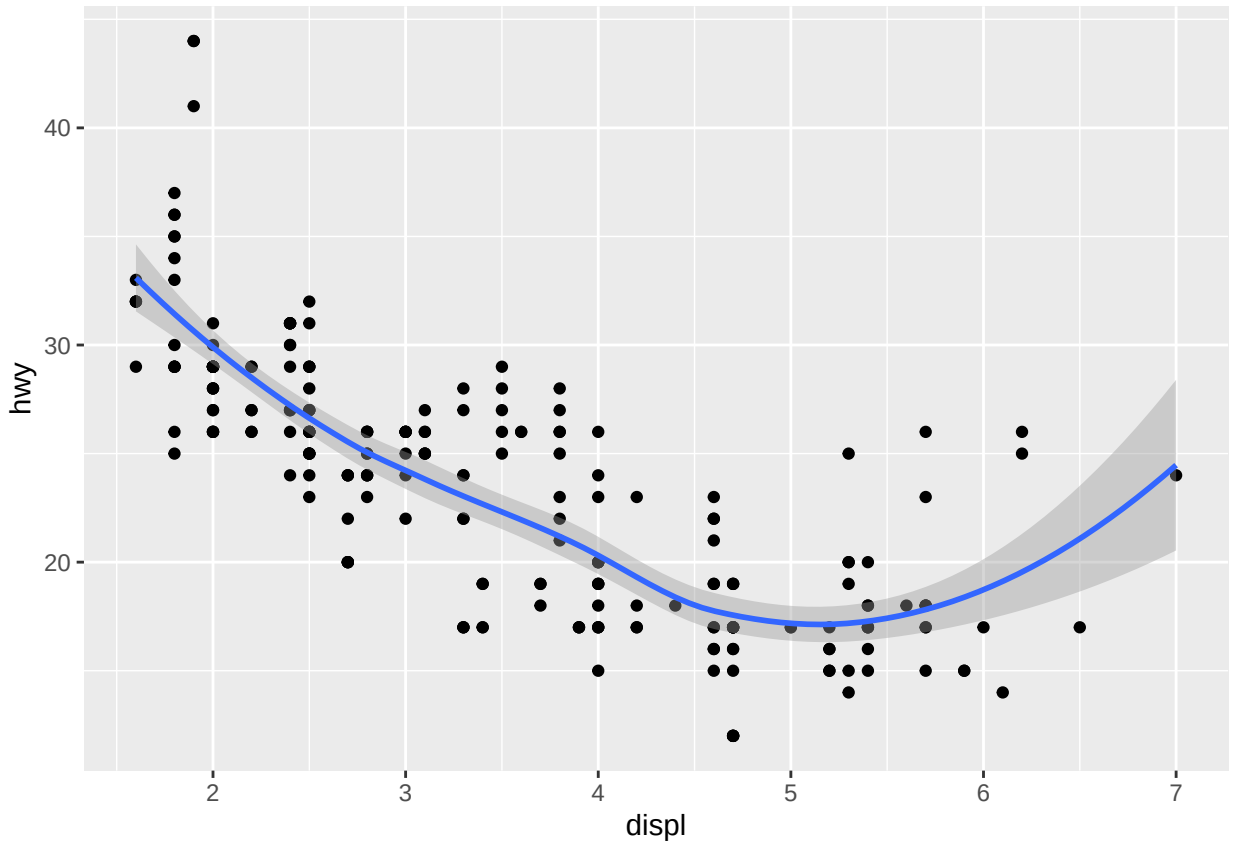

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy)) ## geom_smooth converts scatterplot to a range
```



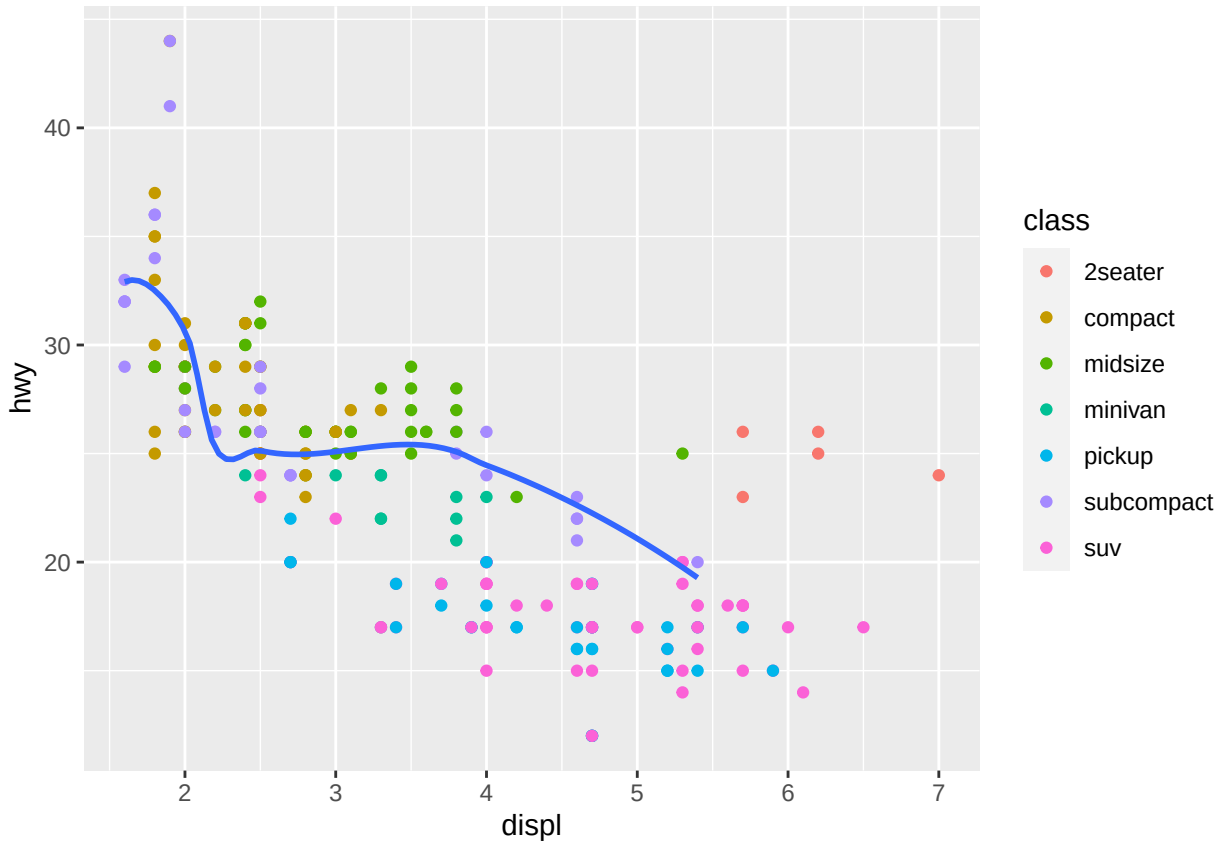

```
ggplot(data = mpg) +
  geom_smooth(mapping = aes(x = displ, y = hwy, linetype = drv)) ## "linetype = drv)" made overlapping
```



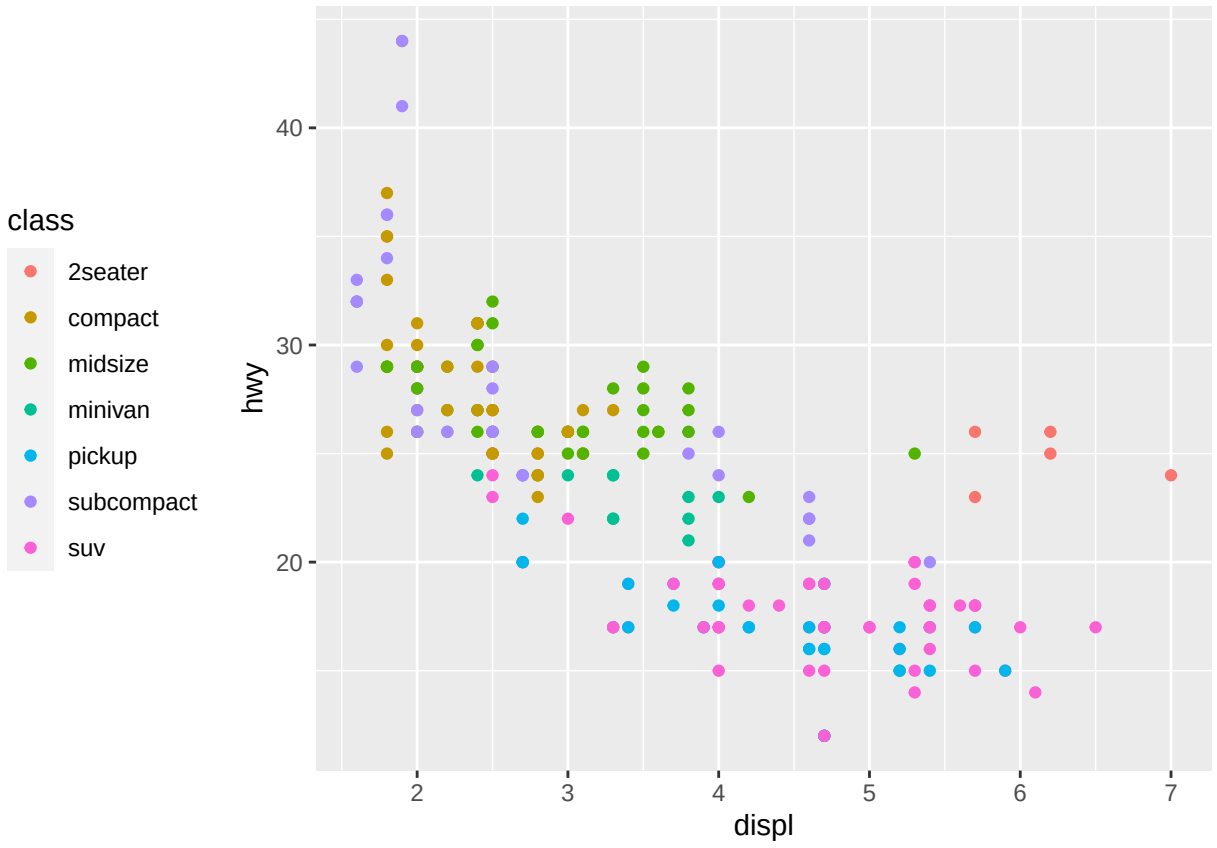

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) + ##include the plus on the end of the line for overlap
  geom_smooth(mapping = aes(x = displ, y = hwy))
```




```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping = aes(color = class)) +
  geom_smooth(data = filter(mpg, class == "subcompact"), se = FALSE) ## allows for local mapping classe
```

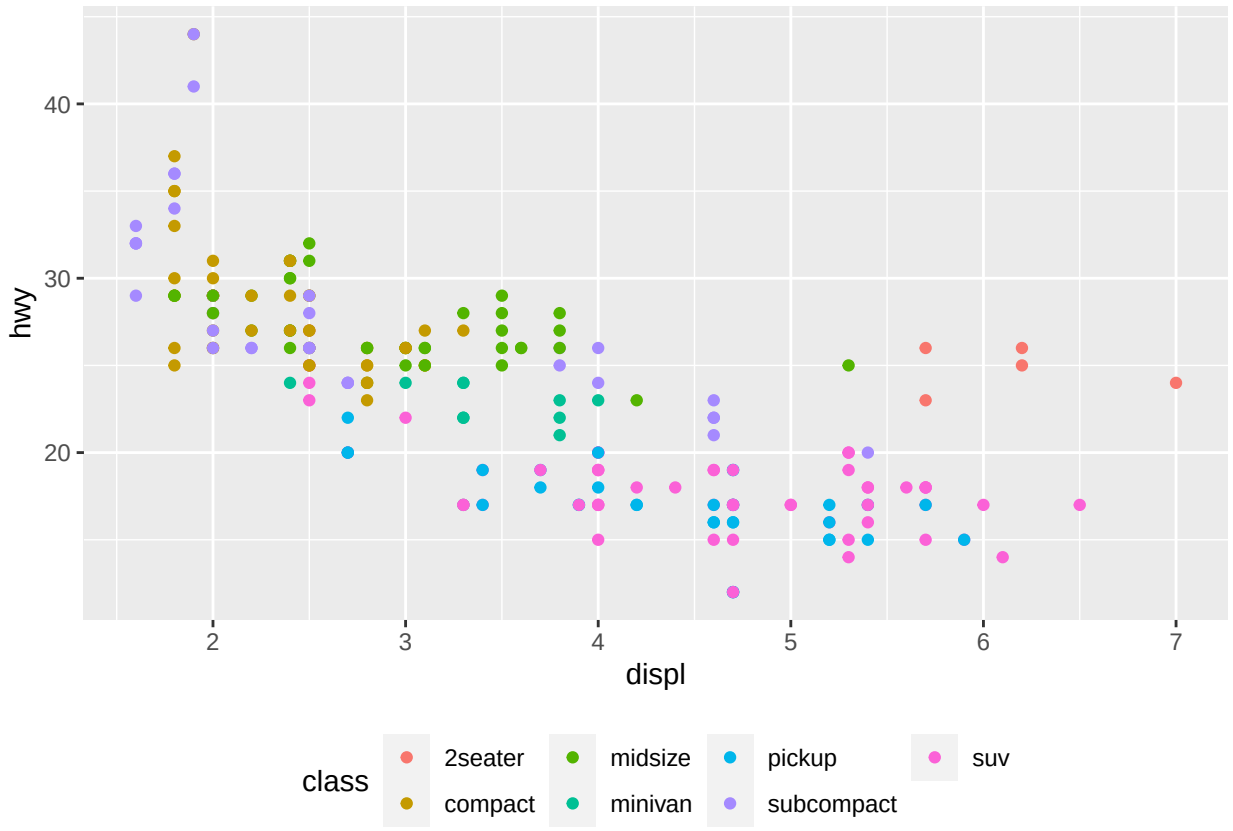



```
base <- ggplot(mpg, aes(displ, hwy)) +
  geom_point(aes(colour = class))

base + theme(legend.position = "left")
```

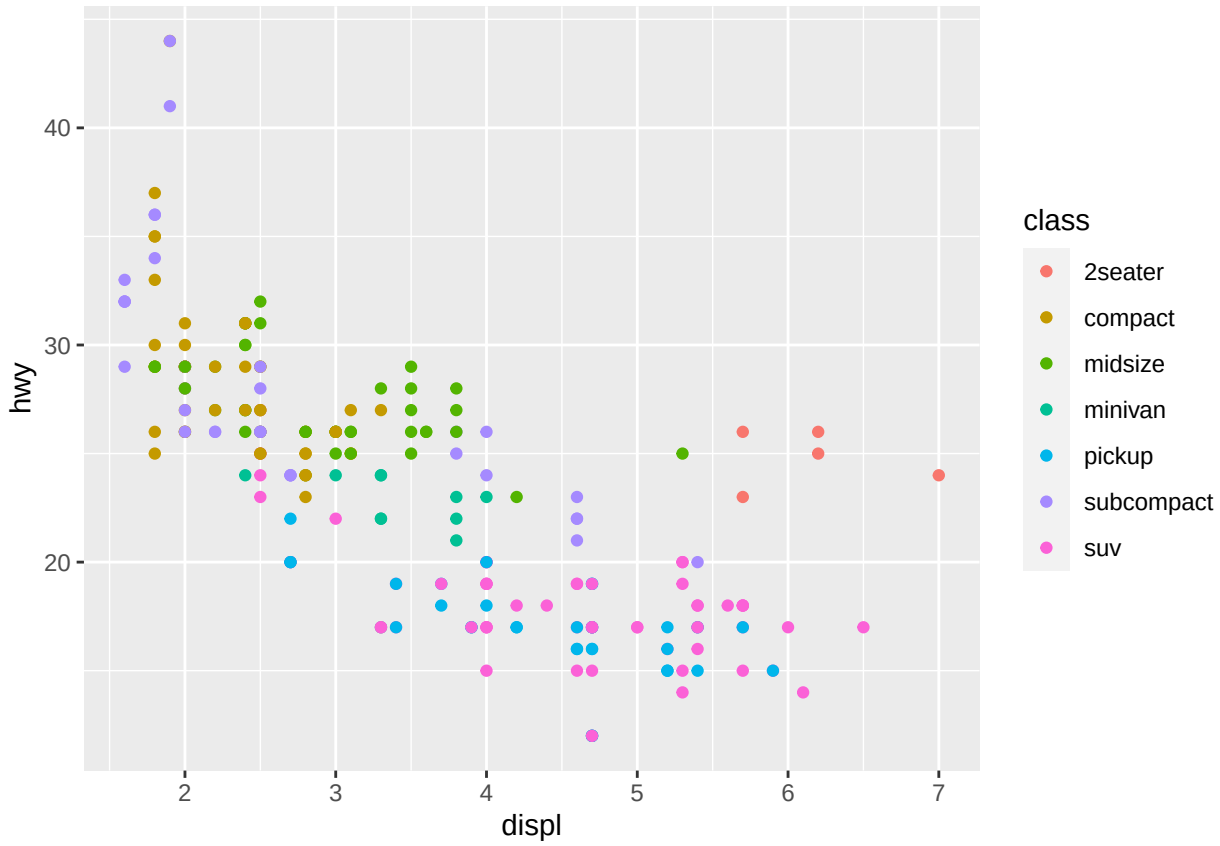



```
base + theme(legend.position = "top")
```

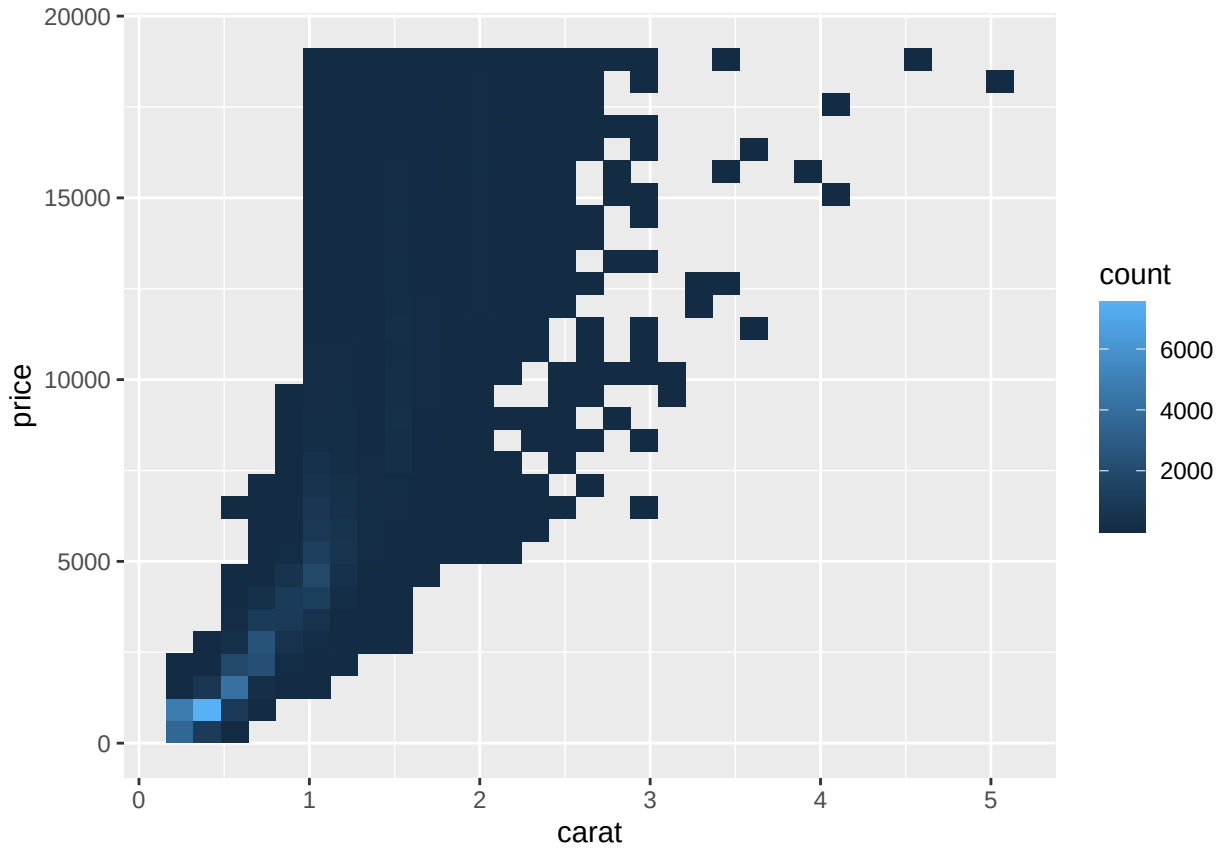

```
base + theme(legend.position = "bottom")
```



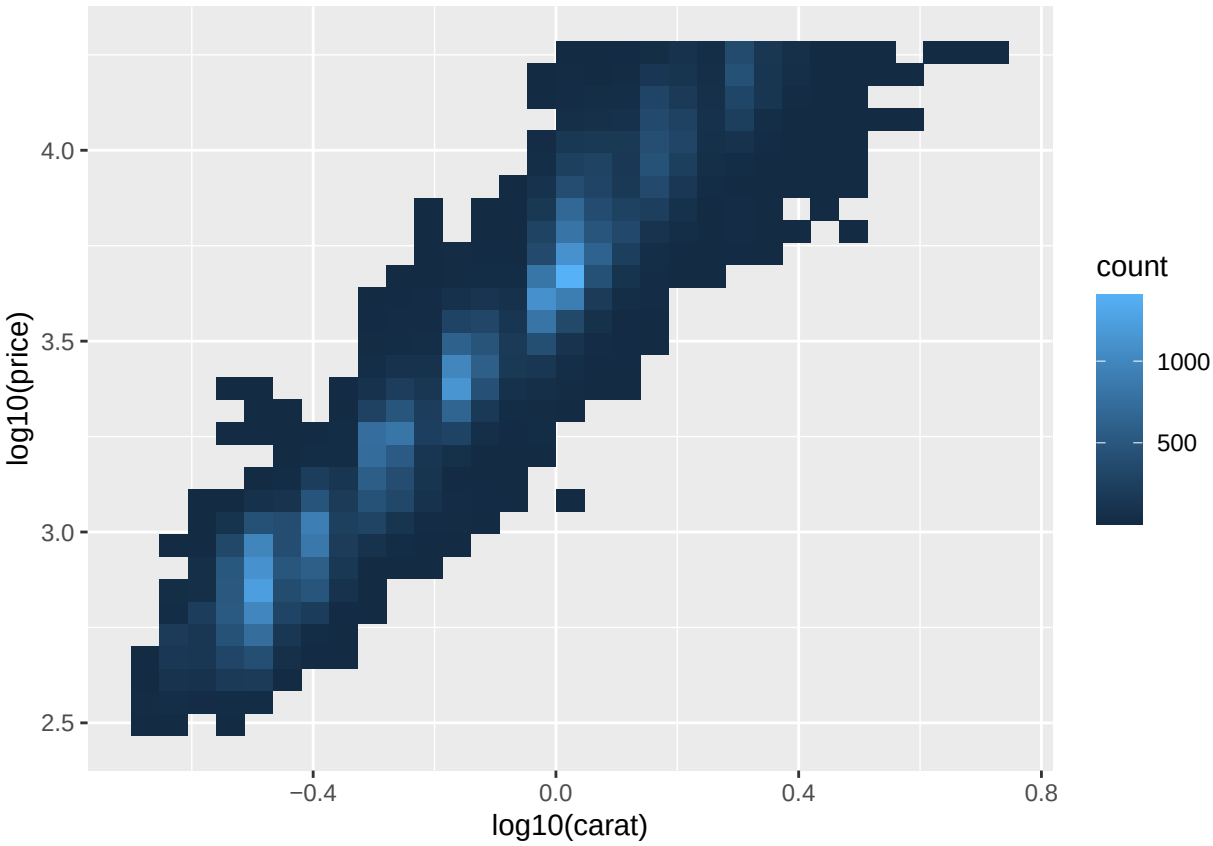

```
base + theme(legend.position = "right") # the default
```




```
ggplot(diamonds, aes(carat, price)) +
  geom_bin2d()
```




```
ggplot(diamonds, aes(log10(carat), log10(price))) +
  geom_bin2d()
```



Final Project Paper

Research Question

The primary objective of this project is to gauge FBI UCR reporting agency participation, assessing the potential for hate crime underreporting along state-lines. For this paper, we chose to narrow our goals in order to answer the question: what is the relationship between the number of UCR reporting agencies per state and the total count of reported hate crimes in the United States? We hypothesize that states with smaller populations are more likely to have fewer reporting agencies, which in turn leads to underreporting.

We chose this research question due to the growing concern for hate crimes in the country. The prejudice-motivated crime has become increasingly visible in recent years due to various factors including the COVID-19 pandemic. Thus, examining potential compounding causes that contribute to the underreporting of hate crimes is paramount, as well as identifying potential solutions to reduce the experienced effect of the dark figure of crime. We hope to contribute to broader studies by providing a specialized assessment of reporting agencies.

Description of Data

The data we examined for our assessment is derived from the Uniform Crime Reporting Program, also known as the UCR Program. The UCR Program includes data from more than 18,000 cities, universities/colleges, counties, states, tribal areas, and federal law enforcement agencies. Agencies participate voluntarily and submit their crime data either through a state UCR program or directly to the FBI's UCR Program. The UCR Program's foundation was created in 1929 by the Committee on Uniform Crime Records, which was in turn established by the International Association of Chiefs of Police in hopes of standardizing criminal reporting.

The UCR is regarded as the largest and arguably most reliable crime data source; however, researchers remain divided on its accuracy due to the nature of the program. We summarize popular arguments in order to provide context for our extrapolated findings.

The program's support lies in:

- It being the oldest and most consistent form of measuring crime, with the FBI providing accessible data as early as the 1960s;
- Its broad national 98.4% metropolitan and 92.7% non-metropolitan coverage;
- Its support for cross-year referencing and analysis;
- Its consistently reliable reporting of higher level crimes;
- And its agencies' broad coverage of 97% of the national population

Whereas, the program's criticism results from: - The continued existence of underreporting, which in turn stems from:

- Victims believing law enforcement cannot or will not act on their behalf;
- Victims fostering a standing fear of reprisal;
- Victims' exposure to deterring cultural or social reporting stigmas;
- And victims' normalization of certain crimes;
- Inconsistent law enforcement reporting practices;
- And the general absence of "Victimless" crime in reportings

We utilized the FBI UCR 2019 Report on Hate Crime Statistics for our specialized assessment. The dataset includes 14 tables detailing hate crimes on a national scale, with defined variables including bias motivation, crime types, and identifying offender/victim characteristics, among others. Our particular analysis required the use of tables 11 and 12. Jointly, both tables were used to contextualize the relationship between state's reporting agencies and reported hate crimes, accounting for differences in agency counts, crime frequencies, and populations.

The former, "Table 11: Offenses and Offense Classification by Participating States," provided a breakdown of state's reported hate crimes. Primary variables included an identifying state name, a state's total offenses, and specified offense counts for crimes against persons, property, and society, with respective subcategories by committed crimes. To elaborate, a total sum count for a state's reported hate crimes against persons or property were reported. Further data for particular crimes such as murder, assault, and arson were then additionally reported via specified crime counts within their respective crime classification. Reported incidents of murder, for example, could thus be studied either in isolation or within the broader parameters of a state's overall crime total or crimes against persons.

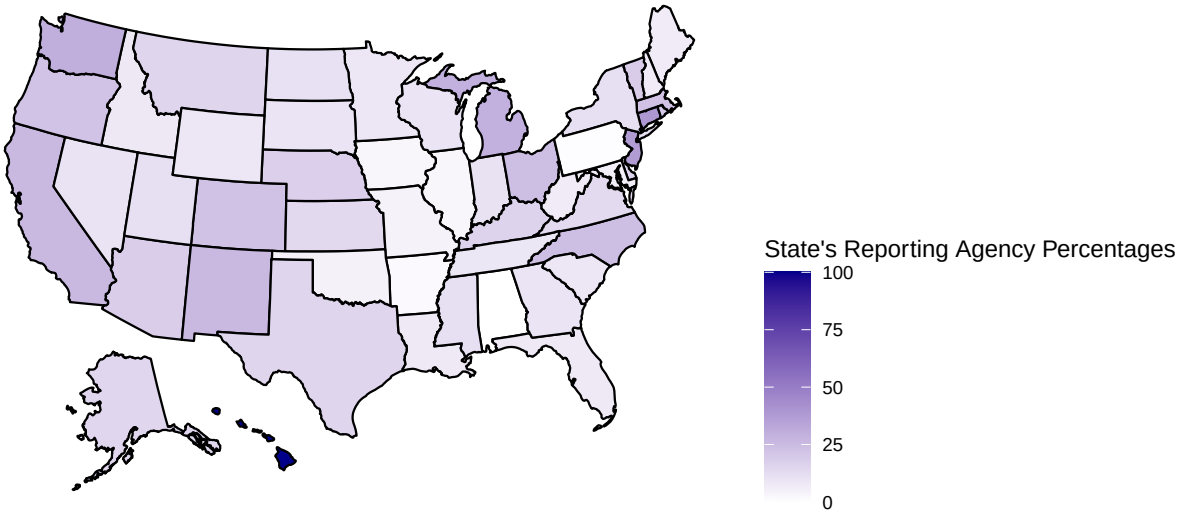
In contrast, the latter, "Table 12: Agency Hate Crime Reporting by State and Federal," provided a summary of each state's total participating agencies and reported incidents. Primarily variables included an identifying state name, state populations, a state's total number of incidents reported, and a state's overall number of UCR participating agencies, as well the number of participating agencies who submitted reports.

Table 12 was thus particularly valuable to the testing of our hypothesis and completion of our objective, while table 11 contextualized our findings.

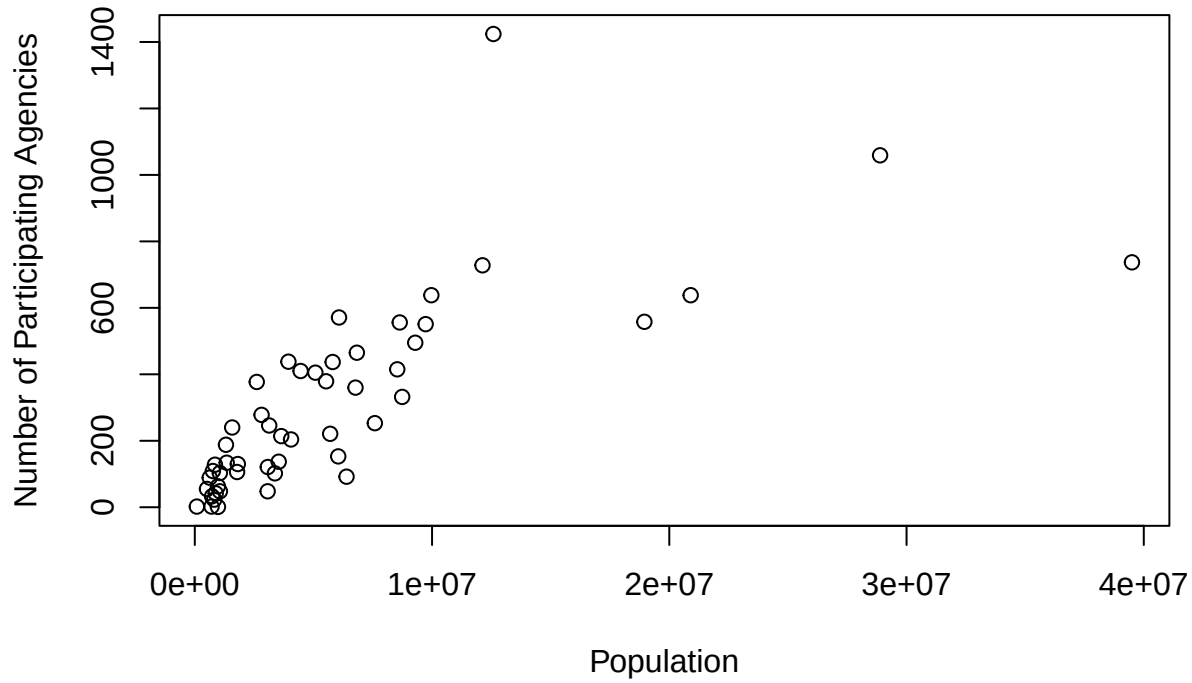
Exploratory Data Analysis

To reiterate, our primary objective was to examine the relationship between reporting agencies and reported hate crimes, as we anticipated states with smaller populations would be more likely to have fewer reporting agencies and thus under-report crimes more frequently. Before conducting a linear regression for total incidents reported and participating agencies, we completed two exploratory data analyses. Both expressed via scatterplots, we examined correlational relationships between a state's numbers of participating agencies and population, as well as a state's total incidents reported and population.

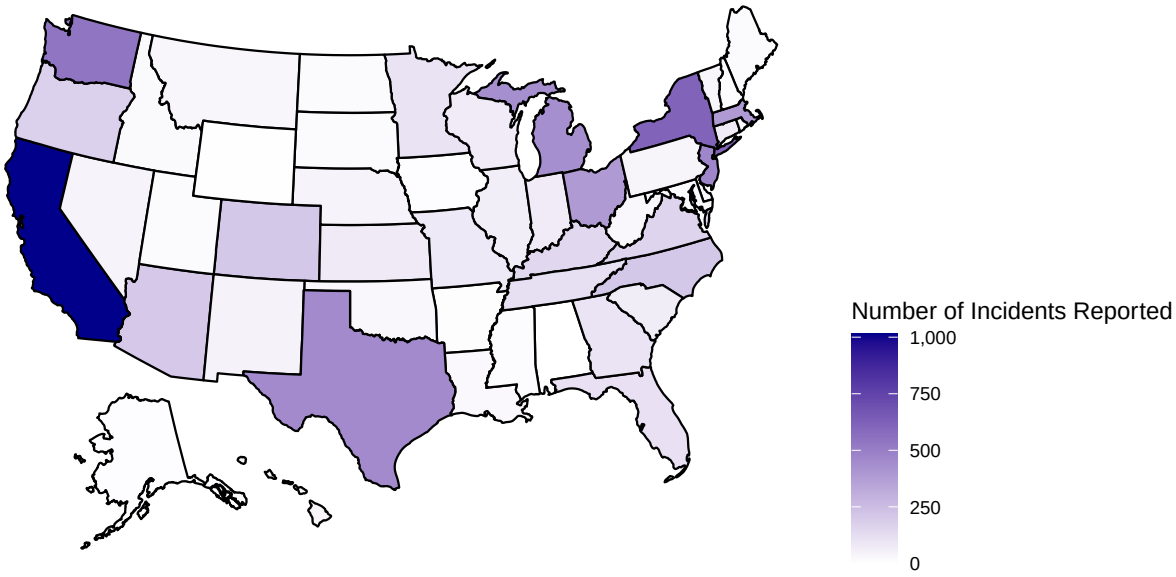
To contextualize our preliminary findings, we will first briefly summarize our variables. Our analyses included all fifty states, as well as the District of Columbia. However, while all states and DC have UCR participating agencies, some agencies and states fail to report hate crime incidents. Such was the case with Alabama, which reported no incidents despite having two participating agencies. Hawaii, in contrast, did indeed reporte with one sole agency. We will discuss the consequences of this phenomenon in the later discussion. A visual summary of state's percentages of agencies who reported cases out of their total participating agency count can be found below:

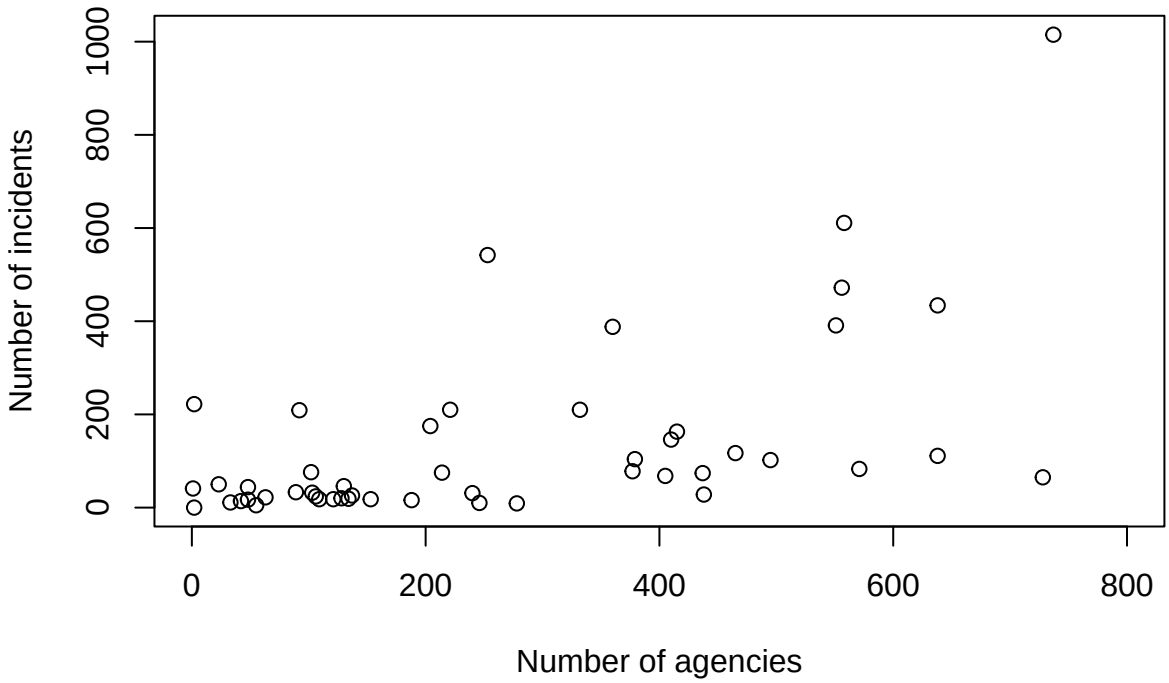


The prior graph provides context for the first exploratory data analysis, where we sought to identify the relationship between a state's population and its number of UCR participating agencies. Our quick assessment yielded a correlation coefficient of 0.7281031, reflected by a clear visual upward trend identifiable in both the raw and fitted graphs depicted below.



Observing a potential relationship, we conducted our second preliminary analysis, this time testing the relationship between a state's population and its total number of incidents reported. Depicted below, is the scatterplot depicting the assessment, which yielded a correlation coefficient of 0.7795704, as well as a map for states' overall incidents for context. Again, please note the outermost value for the x-axis on the scatterplot.





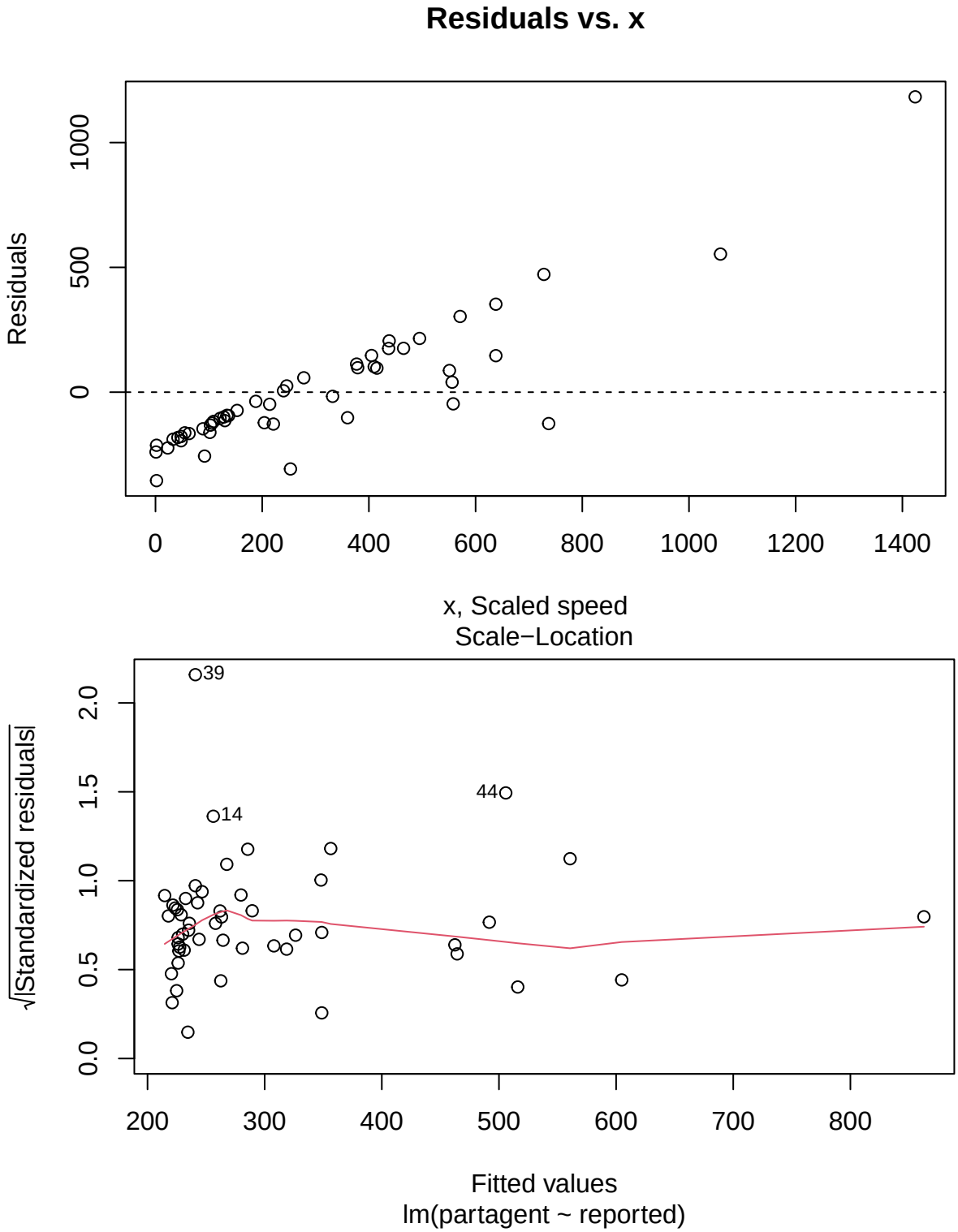
With promising exploratory data analysis results, we decided to test our hypothesis by performing a linear regression on reporting agencies and reported hate crimes.

Linear Regression

We conducted a linear regression to determine if there was an association between the number of agencies and the number of incidents that a state records. As we fit this linear model, it will help us see if an increase of the number of agencies increased the number of incidents a state reported. First, we got a summary of our linear model to see if we have obtained a significant p-value. We obtained a p-value of 0.001 which means we should be able to reject the null hypothesis, but when we look at the residual standard error, it is clear that we have obtained some obscure amount since this would mean that we deviated by 257.1 agents from our regression line. So, we must now test our assumptions in order to verify we have a valid model.

```
##
## Call:
## lm(formula = partagent ~ reported, data = table12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -354.40 -153.95  -92.78  100.04 1183.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 214.6460    44.2994   4.845 1.31e-05 ***
## reported      0.6385     0.1831   3.486  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 257.1 on 49 degrees of freedom
## Multiple R-squared:  0.1988, Adjusted R-squared:  0.1824
## F-statistic: 12.16 on 1 and 49 DF,  p-value: 0.001043
```

First, we test the linearity assumption. With this, we want to see if the relationship between the number of agencies and the mean of the number of reported incidents is linear. When we plot a graph of the residuals against the x value, we see that there is a clear, slightly linear pattern. Therefore, we can't say that this assumption holds.



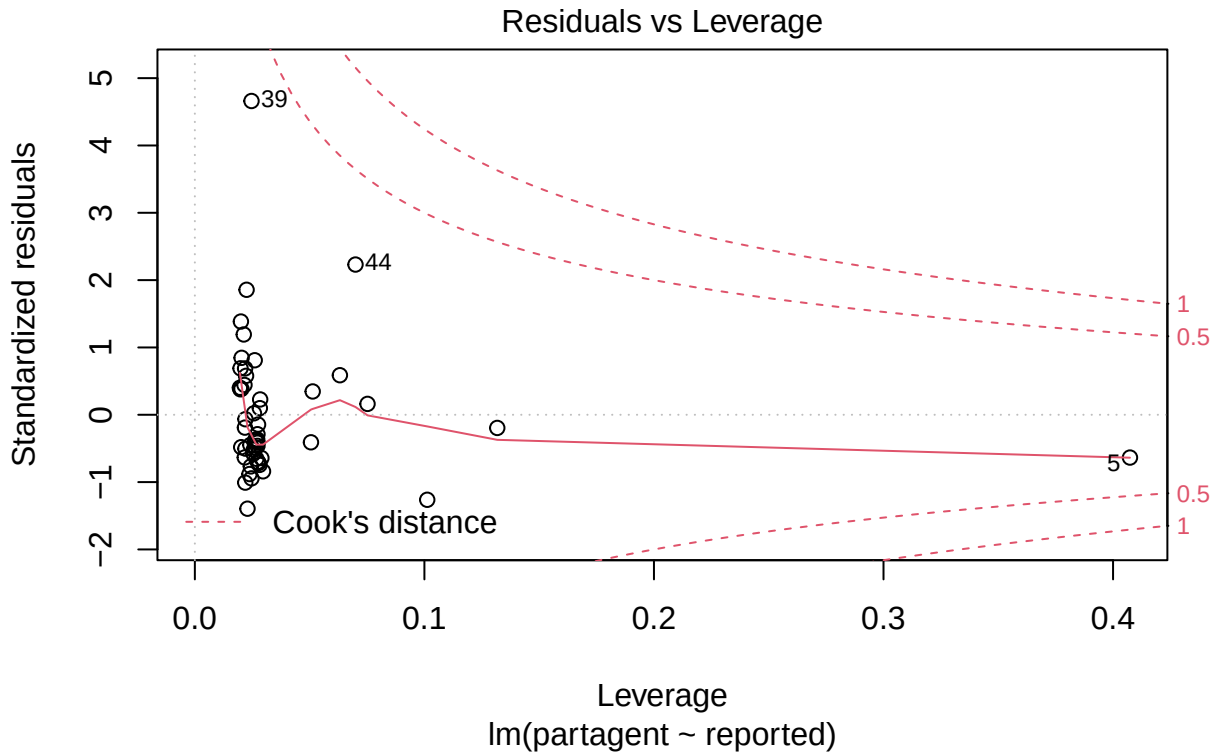
Second, we want to test the independence assumption. Although there is no real test for this, when we look

at the graph for the residuals against x, we can see that there is a large amount of clumping. This suggests a failure of independence. Thus, we cannot say that this assumption holds.

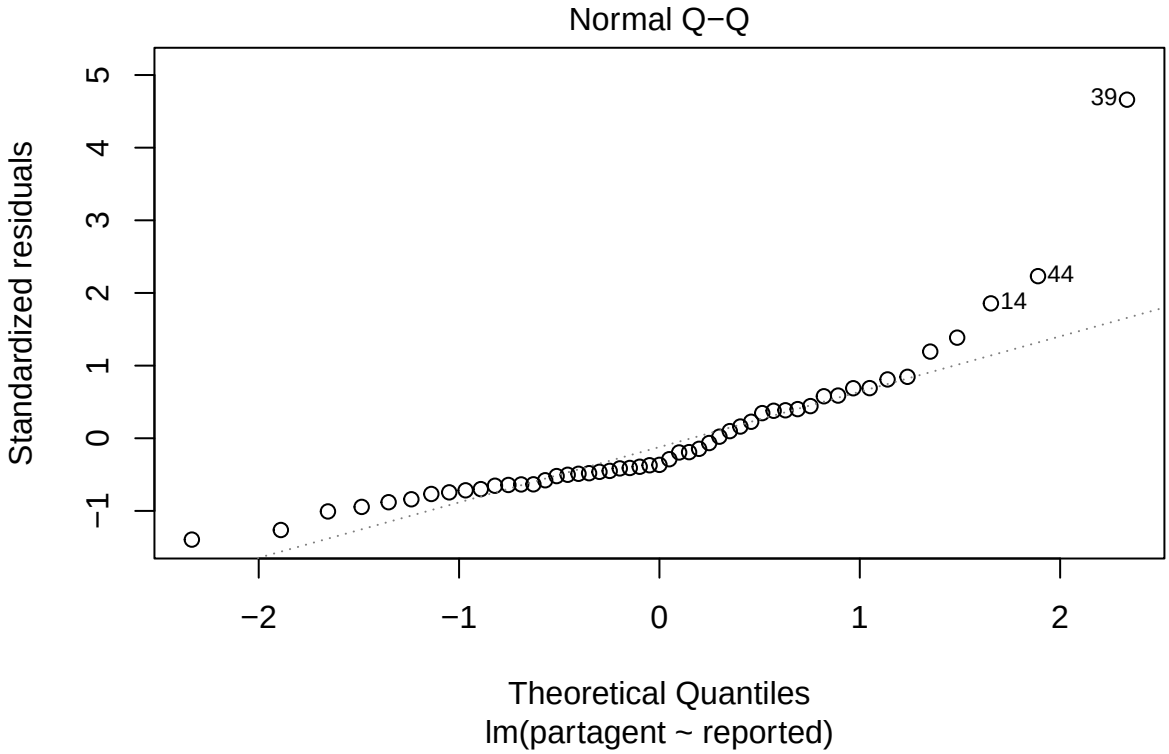
Next, we tested the equal variance assumption by looking at the scale-location plot and checking for deviations. Although we do not see significant trends in the red line, there is a significant amount of clumping in the beginning. Therefore the conclusion is inconclusive.

Finally, we test the normal population assumption by looking at the residuals against the leverage and a normal Q-Q plot. First, the normal q plot has a slight right skew. However, both ends of the tail become light. This implies that it is smaller than usual for a normal distribution. Additionally, as we analyze the residuals against the leverage. We can see the heavy clumping in the beginning alongside the outliers that are pulling the fit of this model. This supports the idea that this assumption does not hold.

```
plot(lm_agencies_report, which=5)
```




```
plot(lm_agencies_report, which=2)
```



As we conducted our linear regression we initially obtained significant p-values that would support our claim. However, after testing each of our assumptions, it is clear that this is not a good fit. Thus, we must reject the null hypothesis.

Causal Analysis and Causal DAG

Causally speaking, our hypothesis was made on the basis that smaller state populations would lead to fewer reporting agencies, which in turn would lead to underreporting of hate crimes in that state. The ideal data set would contain not only data on the reporting agencies themselves, but also data on hate crimes at an individual city level to see if those with reporting agencies have higher rates of reported hate crimes, which would support the argument that underreporting is a result of lack of agency. In a field such as hate crimes, there are many more factors that have an effect on the reporting. Bias by police officers means they may report hate crimes differently based on personal opinions, which directly leads to underreporting; there is also the indirect path in which this bias leads to victim mistrust of law enforcement as a whole, which means they may not report their incidents of hate crime. Police bias may also lead to fewer agencies reporting, as the voluntary nature of the UCR means that they can simply choose not to participate if they feel hate crimes are not worthy of reporting. From a policy standpoint, there are policies when an individual commits multiple crimes that can lead to underreporting, simply because another charge takes precedence. All of these confounding factors mean the causal connection between population, reporting agencies, and underreporting is far more complex and would require a more concise dataset.

Discussion

Our initial underlying interest lied in examining the relationship between reported variables in hopes of investigating potential incidents of hate crime underreporting. Given the widespread prevalence of hate crimes nationwide, we are aware that the crime category has plagued the justice system as members of marginalized communities are increasingly subjected to incidents. Along with the growing concern for hate crimes comes its association with underreporting, which is a result of issues such as misclassification, trivialization by those in power and victims, and negligence. As aforementioned, the UCR is entirely voluntary, and while

all states have participating agencies, some like Alabama fail to report any cases. In contrast, Hawaii had proportionally half their agencies report. This confound makes studying the underreporting of hate crimes very complex, as population, agencies, and policies all have to be taken into account. This, along with the confounding variables addressed in the DAG, explain why it is increasingly difficult to establish a causal relationship between any of these factors and the underreporting of hate crimes. Future research could investigate different ways to model this relationship, such as running multi variable regressions that can take into account the litany of factors that affect the reporting of hate crimes.

References

About the Uniform Crime Reporting (UCR) program. Federal Bureau of Investigation: Uniform Crime Reporting. (2011, July 25). Retrieved December 1, 2021, from https://ucr.fbi.gov/leoka/leoka-2010/aboutucrmain.

Criminal Justice Information Services Division, Federal Bureau of Investigation. "Table 11—Offenses, Offense Type, by Participating State and Federal, 2019." (n.d.). Retrieved December 1, 2021, from https://ucr.fbi.gov/hate-crime/2019/resource-pages/tables/table-11.xls.

Criminal Justice Information Services Division, Federal Bureau of Investigation. "Table 12—Agency Hate Crime Reporting by State and Federal, 2019." (n.d.). Retrieved December 1, 2021, from https://ucr.fbi.gov/hate-crime/2019/topic-pages/tables/table-12.xls.

FBI. (2018, September 10). Uniform Crime Reporting (UCR) Program. Federal Bureau of Investigation. Retrieved December 1, 2021, from https://www.fbi.gov/services/cjis/ucr/.