Lab 8

Snow Christensen

October 27, 2017

Using your own dataset (which may include more than one table) carry out the following data cleaning steps. Knit together the PDF document and commit both the Lab 8 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

Before you begin: as many of you have large datasets, you're going to want to select only the variables you're interested in utilizing for this project (ideally no more than twenty columns but perhaps much smaller) so you don't have R Studio's memory working on the entire dataset. The example code provided below can be modified to allow you to subset your data to only the variables you wish to use. First, read in your complete dataset and save it as `data`. Then, add the names of the variables you wish to use for your poster project to the `select` function, separated by commas. Run the two lines of code to save this new, smaller version of your data to `data_subset`. Use this smaller dataset to complete the rest of the lab

```
#load package for using select() function
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#load package for using "%%" function
install.packages("magrittr")
```

```
## Installing package into 'C:/Users/snowbc/Documents/Projects/sexual_violence_college/packrat/lib/x86_
## (as 'lib' is unspecified)
```

```
## package 'magrittr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Temp\20\RtmpG2Qfxf\downloaded_packages
```

```
library("magrittr")

#load package for reading .por file
# install.packages("haven")
library("haven")

# Read in your data with the appropriate function
data <- read_por("~/Projects/sexual_violence_college/ICPSR_03212/DS0001/03212-0001-Data.por")



 data_subset <- data %>%
  select(RACE, NOASSERT, PLEASE, NORISK, TRUSTFUL, CONSENT, ATTEMPT, PRESSSI, AUTHSI, SEXACTS)
```

1. To get a feel for its structure, look at the class, dimensions, column names, structure, and basic summary statistics of your data.

```
# to figure out the names of the variables
names(data_subset)
```

```
##  [1] "RACE"     "NOASSERT" "PLEASE"   "NORISK"   "TRUSTFUL" "CONSENT"
##  [7] "ATTEMPT"  "PRESSSI"  "AUTHSI"   "SEXACTS"
```

```
# to list the structure of my data
str(data_subset)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1580 obs. of  10 variables:
##  $ RACE    :Class 'labelled'  atomic [1:1580] 1 2 3 1 1 3 1 1 1 1 ...
##   .. ..- attr(*, "labels")= Named num [1:5] 0 1 2 3 9
##   .. .. ..- attr(*, "names")= chr [1:5] "No response" "White" "Black" "Other" ...
##  $ NOASSERT:Class 'labelled'  atomic [1:1580] 1 2 4 1 1 3 4 1 2 1 ...
##   .. ..- attr(*, "labels")= Named num [1:7] 0 1 2 3 4 5 9
##   .. .. ..- attr(*, "names")= chr [1:7] "No response" "Not at all like me" "A little like me" "Somewh
##  $ PLEASE  :Class 'labelled'  atomic [1:1580] 2 1 1 2 2 3 2 1 3 2 ...
##   .. ..- attr(*, "labels")= Named num [1:7] 0 1 2 3 4 5 9
##   .. .. ..- attr(*, "names")= chr [1:7] "No response" "Not at all like me" "A little like me" "Somewh
##  $ NORISK  :Class 'labelled'  atomic [1:1580] 1 4 1 3 4 2 3 4 4 4 ...
##   .. ..- attr(*, "labels")= Named num [1:7] 0 1 2 3 4 5 9
##   .. .. ..- attr(*, "names")= chr [1:7] "No response" "Not at all like me" "A little like me" "Somewh
##  $ TRUSTFUL:Class 'labelled'  atomic [1:1580] 2 2 1 2 2 4 2 2 2 1 ...
##   .. ..- attr(*, "labels")= Named num [1:7] 0 1 2 3 4 5 9
##   .. .. ..- attr(*, "names")= chr [1:7] "No response" "Not at all like me" "A little like me" "Somewh
##  $ CONSENT :Class 'labelled'  atomic [1:1580] 1 2 1 2 2 1 2 2 2 2 ...
##   .. ..- attr(*, "labels")= Named num [1:4] 0 1 2 9
##   .. .. ..- attr(*, "names")= chr [1:4] "No response" "Never" "At least once" "Missing"
##  $ ATTEMPT :Class 'labelled'  atomic [1:1580] 1 2 1 1 1 1 1 2 1 2 ...
##   .. ..- attr(*, "labels")= Named num [1:4] 0 1 2 9
##   .. .. ..- attr(*, "names")= chr [1:4] "No response" "Never" "At least once" "Missing"
##  $ PRESSSI :Class 'labelled'  atomic [1:1580] 1 2 1 1 1 1 1 2 1 1 ...
##   .. ..- attr(*, "labels")= Named num [1:4] 0 1 2 9
##   .. .. ..- attr(*, "names")= chr [1:4] "No response" "Never" "At least once" "Missing"
##  $ AUTHSI  :Class 'labelled'  atomic [1:1580] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..- attr(*, "labels")= Named num [1:4] 0 1 2 9
##   .. .. ..- attr(*, "names")= chr [1:4] "No response" "Never" "At least once" "Missing"
##  $ SEXACTS :Class 'labelled'  atomic [1:1580] 1 1 1 1 1 1 1 1 1 1 ...
##   .. ..- attr(*, "labels")= Named num [1:4] 0 1 2 9
##   .. .. ..- attr(*, "names")= chr [1:4] "No response" "Never" "At least once" "Missing"
```

```
# to look at the dimensions of my data
dim(data_subset)
```

```
## [1] 1580   10
```

```
# to look at the class of my data
class(data_subset)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
#to better understand the structure and columns
glimpse(data_subset)
```

```
## Observations: 1,580
```

```
## Variables: 10
## $ RACE     <dbl+lbl> 1, 2, 3, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3...
## $ NOASSERT <dbl+lbl> 1, 2, 4, 1, 1, 3, 4, 1, 2, 1, 1, 1, 1, 1, 2, 1, 3...
## $ PLEASE   <dbl+lbl> 2, 1, 1, 2, 2, 3, 2, 1, 3, 2, 3, 1, 1, 1, 1, 1, 3...
## $ NORISK   <dbl+lbl> 1, 4, 1, 3, 4, 2, 3, 4, 4, 4, 4, 3, 3, 4, 4, 3, 4...
## $ TRUSTFUL <dbl+lbl> 2, 2, 1, 2, 2, 4, 2, 2, 2, 1, 1, 2, 3, 3, 1, 1, 4...
## $ CONSENT  <dbl+lbl> 1, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ ATTEMPT  <dbl+lbl> 1, 2, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1...
## $ PRESSSI  <dbl+lbl> 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 1...
## $ AUTHSI   <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ SEXACTS  <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
```

2. Preview the first and last 15 rows of your data. Is you dataset tidy? If not, what principles of tidy data does it seem to be violating?

```
# to preview the first 15 rows of my data
head(data_subset, n=15)
```

```
## # A tibble: 15 x 10
##         RACE  NOASSERT    PLEASE    NORISK  TRUSTFUL   CONSENT   ATTEMPT
##    <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl>
##  1         1         1         2         1         2         1         1
##  2         2         2         1         4         2         2         2
##  3         3         4         1         1         1         1         1
##  4         1         1         2         3         2         2         1
##  5         1         1         2         4         2         2         1
##  6         3         3         3         2         4         1         1
##  7         1         4         2         3         2         2         1
##  8         1         1         1         4         2         2         2
##  9         1         2         3         4         2         2         1
## 10         1         1         2         4         1         2         2
## 11         1         1         3         4         1         2         1
## 12         1         1         1         3         2         2         1
## 13         1         1         1         3         3         2         1
## 14         2         1         1         4         3         2         1
## 15         2         2         1         4         1         2         1
## # ... with 3 more variables: PRESSSI <dbl+lbl>, AUTHSI <dbl+lbl>,
## #   SEXACTS <dbl+lbl>
```

```
# to preview the last 15 rows of my data
tail(data_subset, n=15)
```

```
## # A tibble: 15 x 10
##         RACE  NOASSERT    PLEASE    NORISK  TRUSTFUL   CONSENT   ATTEMPT
##    <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl>
##  1         1         2         2         4         3         2         1
##  2         1         2         2         2         3         2         1
##  3         1         3         2         3         2         1         1
##  4         1         1         1         4         4         2         1
##  5         1         1         1         3         4         2         1
##  6         1         3         3         3         4         2         1
##  7         1         1         1         4         1         1         1
##  8         2         2         2         4         3         2         2
##  9         1         2         1         2         3         2         1
## 10         2         1         1         4         1         2         2
## 11         2         1         1         3         4         2         1
```

```
## 12         1         2         2         4         2         1         1
## 13         1         3         1         2         1         2         1
## 14         1         2         4         4         4         2         1
## 15         1         1         1         5         2         2         1
## # ... with 3 more variables: PRESSSI <dbl+lbl>, AUTHSI <dbl+lbl>,
## #   SEXACTS <dbl+lbl>
```

3. Create a histogram for at least two variables you plan to focus on for your study. Describe what these plots show you about these variables.

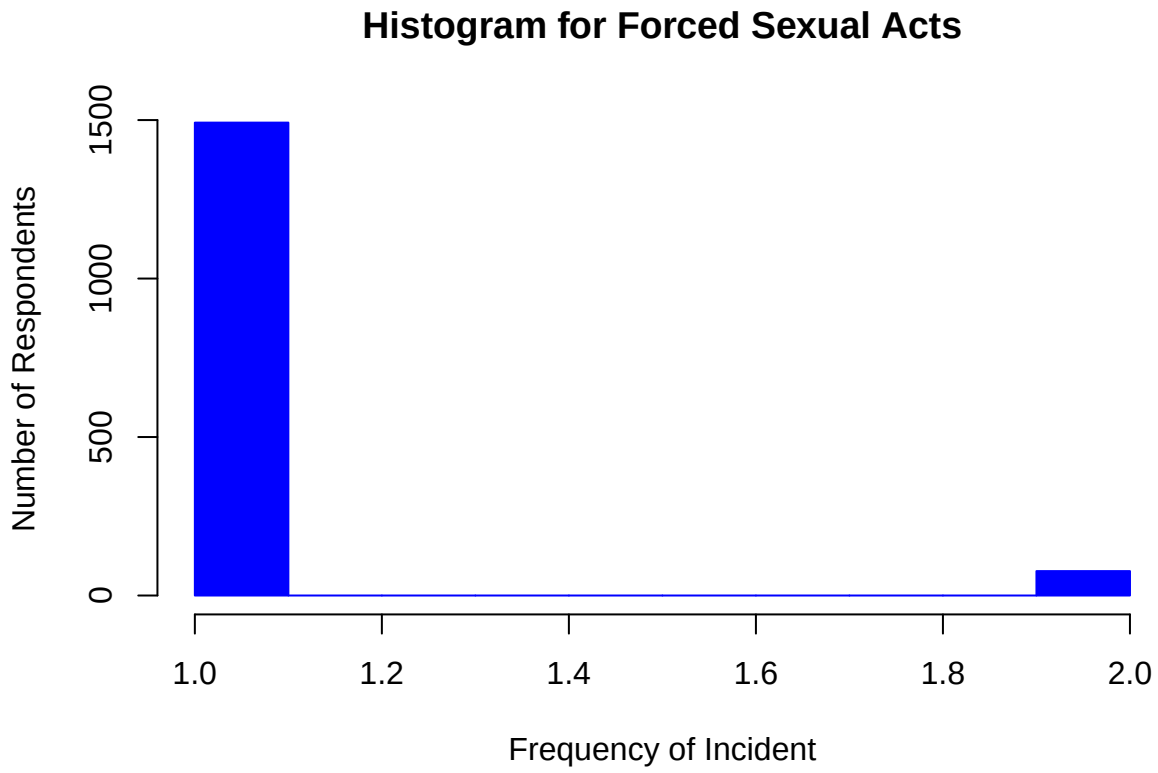
```
#load the package for the histogram
# install.packages("ggplot2")
library("ggplot2")

#convert dataframe to a matrix so it can be read by the hist function
data.matrix(data_subset[c(1:50), ])
```

```
##       RACE NOASSERT PLEASE NORISK TRUSTFUL CONSENT ATTEMPT PRESSSI AUTHSI
##  [1,]    1        1      2      1        2       1       1       1      1
##  [2,]    2        2      1      4        2       2       2       2      1
##  [3,]    3        4      1      1        1       1       1       1      1
##  [4,]    1        1      2      3        2       2       1       1      1
##  [5,]    1        1      2      4        2       2       1       1      1
##  [6,]    3        3      3      2        4       1       1       1      1
##  [7,]    1        4      2      3        2       2       1       1      1
##  [8,]    1        1      1      4        2       2       2       2      1
##  [9,]    1        2      3      4        2       2       1       1      1
## [10,]    1        1      2      4        1       2       2       1      1
## [11,]    1        1      3      4        1       2       1       1      1
## [12,]    1        1      1      3        2       2       1       1      1
## [13,]    1        1      1      3        3       2       1       1      1
## [14,]    2        1      1      4        3       2       1       2      1
## [15,]    2        2      1      4        1       2       1       2      1
## [16,]    2        1      1      3        1       2       1       1      1
## [17,]    3        3      3      4        4       2       1       1      1
## [18,]    2        2      2      2        2       2       1       1      1
## [19,]    2        1      1      2        1       2       1       1      1
## [20,]    1        2      2      4        5       1       1       1      1
## [21,]    1        2      4      4        3       2       1       2      1
## [22,]    2        1      1      5        2       2       1       1      1
## [23,]    2        1      1      4        1       2       2       2      1
## [24,]    1        1      1      1        1       2       1       1      1
## [25,]    2        1      2      3        3       2       1       1      1
## [26,]    2        2      1      5        2       2       2       2      1
## [27,]    1        1      1      4        2       1       1       1      1
## [28,]    1        1      1      5        2       2       2       2      1
## [29,]    2        1      1      1        1       2       1       2      1
## [30,]    1        1      2      2        3       2       1       1      1
## [31,]    1        1      1      4        1       2       2       1      1
## [32,]    1        1      1      5        3       2       1       2      1
## [33,]    2        1      1      4        2       1       1       1      1
## [34,]    1        3      2      1        4       1       1       1      1
## [35,]    1        1      1      3        1       1       1       1      1
## [36,]    1        2      1      5        1       2       1       2      1
## [37,]    1        1      1      4        3       2       1       1      1
```

```
## [38,]    2        1      1      1        2        2        1        1       1
## [39,]    1        1      1      5        1        2        1        1       1
## [40,]    1        1      1      4        1        2        1        1       1
## [41,]    1        1      2      4        2        1        1        1       1
## [42,]    2        2      2      3        2        2        1        1       1
## [43,]    2        1      3      2        1        1        1        1       1
## [44,]    2        2      3      2        2        2        1        1       1
## [45,]    1        1      1      5        2        2        1        1       1
## [46,]    2        1      1      3        1        2        1        1       1
## [47,]    2        1      1      3        1        2        1        2       1
## [48,]    1        1      2      5        1        2        1        1       1
## [49,]    1        3      4      3        4        2        2        2       1
## [50,]    1        2      2      4        1        2        1        1       1
##       SEXACTS
##  [1,]       1
##  [2,]       1
##  [3,]       1
##  [4,]       1
##  [5,]       1
##  [6,]       1
##  [7,]       1
##  [8,]       1
##  [9,]       1
## [10,]       1
## [11,]       1
## [12,]       1
## [13,]       1
## [14,]       1
## [15,]       1
## [16,]       1
## [17,]       1
## [18,]       1
## [19,]       1
## [20,]       1
## [21,]       1
## [22,]       1
## [23,]       1
## [24,]       1
## [25,]       1
## [26,]       1
## [27,]       1
## [28,]       1
## [29,]       1
## [30,]       1
## [31,]       1
## [32,]       1
## [33,]       1
## [34,]       1
## [35,]       1
## [36,]       1
## [37,]       1
## [38,]       1
## [39,]       1
## [40,]       1
```

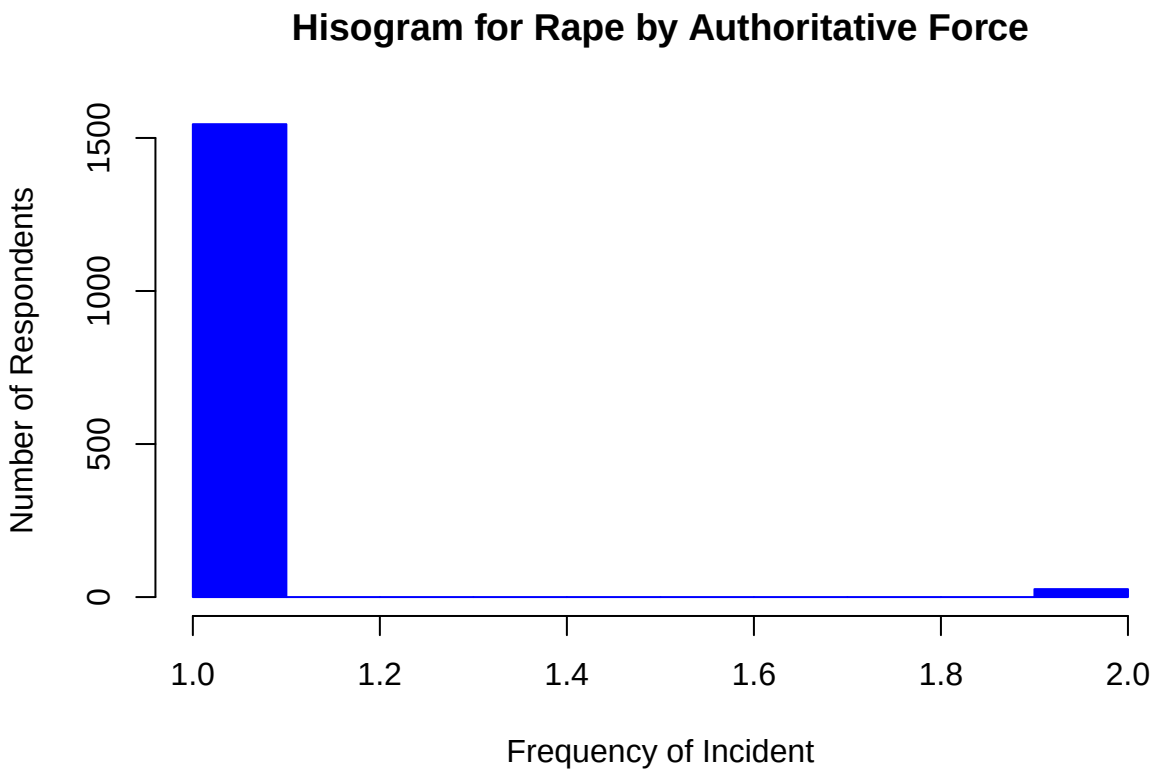
```
## [41,]       1
## [42,]       1
## [43,]       1
## [44,]       1
## [45,]       1
## [46,]       1
## [47,]       1
## [48,]       1
## [49,]       1
## [50,]       1
```

```r
#create histogram of frequency of forced sexual acts
hist(data_subset$SEXACTS,
     main = "Histogram for Forced Sexual Acts",
     xlab = "Frequency of Incident",
     ylab = "Number of Respondents",
     border = "blue",
     col = "blue"
     )
```

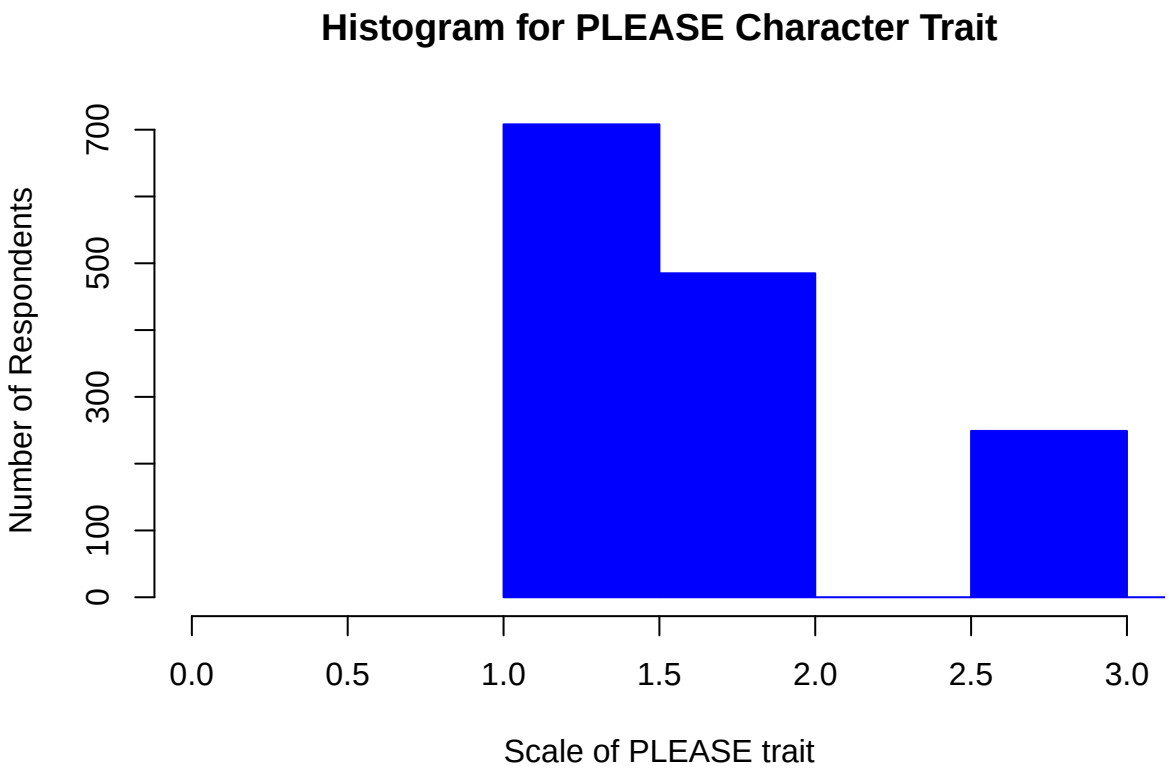



```r
#create histogram of frequency of forced sexual intercourse through authority
hist(data_subset$AUTHSI,
     main = "Hisogram for Rape by Authoritative Force",
     xlab = "Frequency of Incident",
     ylab = "Number of Respondents",
     border = "blue",
     col = "blue"
```

```
    )
```

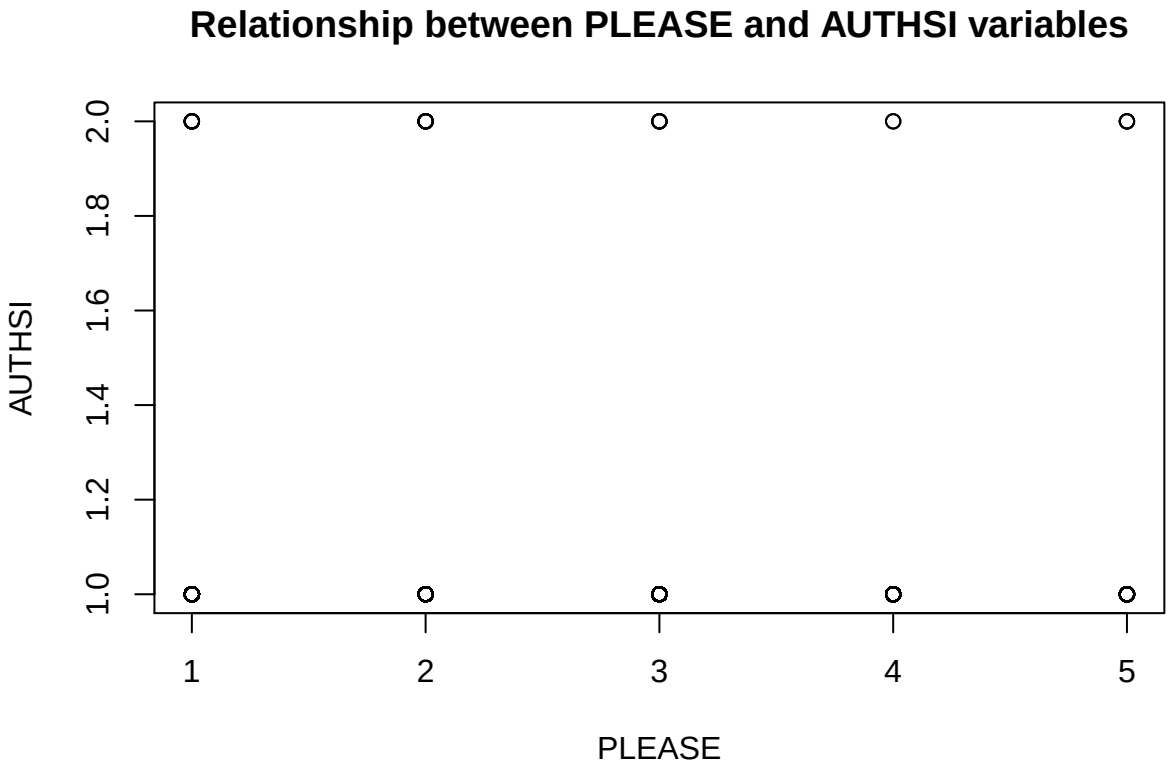



```
#create histogram of frequency of PLEASE variable
hist(data_subset$PLEASE,
     main = "Histogram for PLEASE Character Trait",
     xlab = "Scale of PLEASE trait",
     ylab = "Number of Respondents",
     border = "blue",
     col = "blue",
     xlim = c(0,3)
     )
```



4. Create at least one bivariate plot showing the relationship between two variables of interest. What does/do the(se) plot(s) tell you about the association between these two variables?

```
plot(data_subset$PLEASE, data_subset$AUTHSI,
     main = "Relationship between PLEASE and AUTHSI variables",
     xlab = "PLEASE",
     ylab = "AUTHSI")
```



This shows that there is no significant relationship between the two variables PLEASE and AUTHSI.

5. Load the `tidyr` package. Do all of your columns correspond to variables? Do any columns represent multiple variables? If your answer is yes to either question, carry out the appropriate `tidyr` function (`gather()` or `spread()` respectively) to tidy your data.

```
# I loaded the tidyr package in case I need to use it in the future
install.packages("tidyr")
```

```
## Installing package into 'C:/Users/snowbc/Documents/Projects/sexual_violence_college/packrat/lib/x86_
## (as 'lib' is unspecified)
```

```
## package 'tidyr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Temp\20\RtmpG2Qfxf\downloaded_packages
```

```
library("tidyr")
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

I do not think I need to use the gather() or spread() functions.

6. Do any columns need to be separated into two or more? Do any columns need to be combined into one? If so, carry out the appropriate the appropriate `tidyr` function (`separate()` or `unite()` respectively)

to tidy your data.

No, the columns are already separated.

At this stage each row in your data should represent one observation, each column should be a variable, and each table should be observational unit.

7. What is the class of each of the variables in your analysis? Are these classes appropriate for the type of measurement they purport to capture? Explain your reasoning.

```
# use glimpse function to determine the class of each variable
glimpse(data_subset)
```

```
## Observations: 1,580
## Variables: 10
## $ RACE     <dbl+lbl> 1, 2, 3, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3...
## $ NOASSERT <dbl+lbl> 1, 2, 4, 1, 1, 3, 4, 1, 2, 1, 1, 1, 1, 1, 2, 1, 3...
## $ PLEASE   <dbl+lbl> 2, 1, 1, 2, 2, 3, 2, 1, 3, 2, 3, 1, 1, 1, 1, 1, 3...
## $ NORISK   <dbl+lbl> 1, 4, 1, 3, 4, 2, 3, 4, 4, 4, 4, 3, 3, 4, 4, 3, 4...
## $ TRUSTFUL <dbl+lbl> 2, 2, 1, 2, 2, 4, 2, 2, 2, 1, 1, 2, 3, 3, 1, 1, 4...
## $ CONSENT  <dbl+lbl> 1, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2...
## $ ATTEMPT  <dbl+lbl> 1, 2, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1...
## $ PRESSSI  <dbl+lbl> 1, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 1...
## $ AUTHSI   <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ SEXACTS  <dbl+lbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
```

The labelled double data type is appropriate because it allows the coder to apply a text label to a numeric data type which makes it the data easier to understand and reproduce.

8. Do any of your variables need to be coerced into a different data type? If so, carry out the appropriate coercion methods below. (This includes transformation of any date objects using the `lubridate` package)

They needed to be coerced into a matrix so they could be read into a histogram in question 3 and converted into a numeric for the boxplot in a question below.

9. Are there any strings you need to manipulate for your analysis? If so, use the appropriate function from the `stringr` package.

As of right now I have not needed to manipulate any strings.

10. Do you have any missing values in your dataset? How many and how are they coded? **Be sure to look out for specific codebook values for missing values (i.e. `-1` for `NA`) as well as empty strings or other software-specific values for `NA`.** Don't worry about removing NAs yet - we'll tackle this question later once discern whether they're random or systematically distributed.

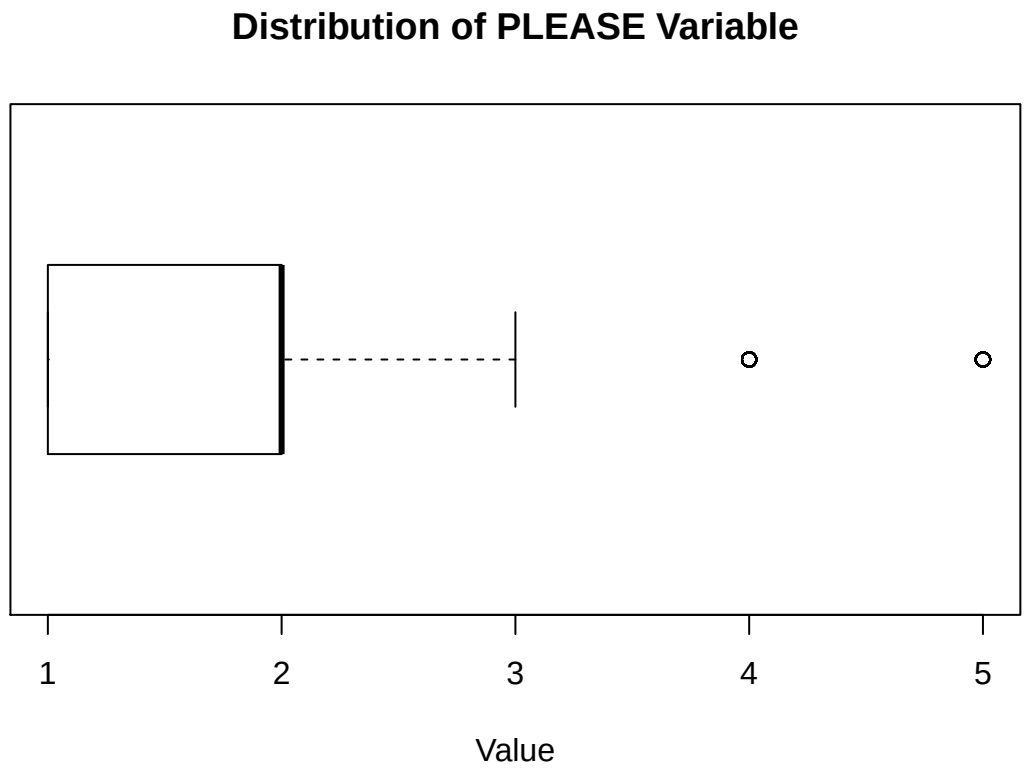
Yes, any missing values are coded as 0 for "No Response."

11. Are there any special values in your dataset? If so, what are they and how do you think they got there? *The presence of special values is less likely if you haven't performed any data manipulation yet so you should remember to return to this step each time you carry out a mathematical transformation of any values in your dataset.*

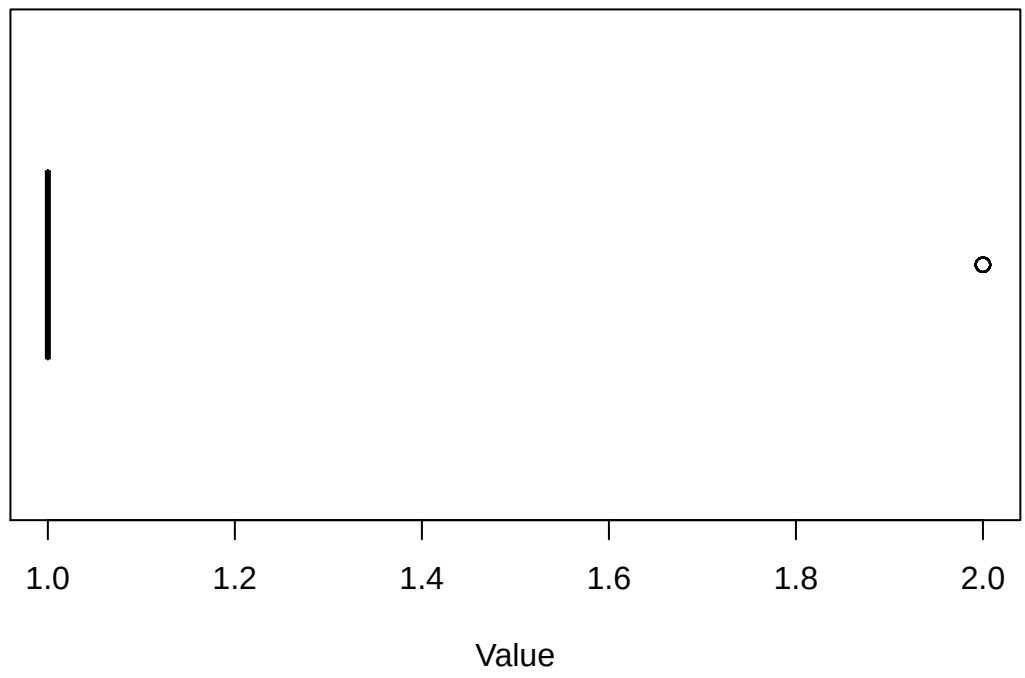
12. Create a boxplot of your data (you can create an individual boxplot for each variable if there are too many variables in your dataset to meaningfully visualize them all in one plot). Are there any outliers? If so, what are they and to which variable do they correspond? Do any of these outliers seem like obvious errors? If so, why?

```
# convert the PLEASE labelled variable to a numeric so it can be read into a boxplot
PLEASEnumeric <- as.numeric(data_subset$PLEASE)
```

```r
#create a boxplot for the PLEASE variable
boxplot(PLEASEnumeric, main = "Distribution of PLEASE Variable", horizontal = TRUE, xlab = "Value")
```




```r
# convert the AUTHSI variable to a numeric so it can be read into a boxplot
AUTHSInumeric <- as.numeric(data_subset$AUTHSI)

# create a boxplot for the AUTHSI variable
boxplot(AUTHSInumeric, main = "Distribution of AUTHSI Variable", horizontal = TRUE, xlab = "Value")
```



For both variables there are obvious outliers. These do not seem like errors because the responses could have just been different for very specific cases. However, I do need to be careful to make sure the outliers do not skew the analysis too much which is why I need to use the most appropriate measure of center, such as the mode.

13. For any outliers and/or obvious errors, what do you think is the best way to handle them (i.e. remove them entirely, run analyses including and excluding them and compare the results, manually change them to an appropriate measure of center, or something else?).

The best way to hand them is to include them but run analyses with the measure of center that will be least affected by the outliers.