Lab 8

Yana Chakalo

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

Installing packages that i might need to use.

```
install.packages("dplyr", repos = "https://CRAN.R-project.org/package=dplyr")
```

```
## Installing package into '/Users/soniachakalo/Downloads/Final.Porject.Labs/packrat/lib/x86_64-apple-da
## (as 'lib' is unspecified)
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=dplyr/bin/macosx/el
##   Line starting '<!DOCTYPE HTML PUBLI ...' is malformed!
```

```
## Warning: package 'dplyr' is not available (as a binary package for R
## version 3.4.1)
```

```
install.packages("readr", repos ="https://CRAN.R-project.org/package=readr")
```

```
## Installing package into '/Users/soniachakalo/Downloads/Final.Porject.Labs/packrat/lib/x86_64-apple-da
## (as 'lib' is unspecified)
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=readr/bin/macosx/el
##   Line starting '<!DOCTYPE HTML PUBLI ...' is malformed!
```

```
## Warning: package 'readr' is not available (as a binary package for R
## version 3.4.1)
```

```
install.packages("packrat", repos = "https://CRAN.R-project.org/package=packrat")
```

```
## Installing package into '/Users/soniachakalo/Downloads/Final.Porject.Labs/packrat/lib/x86_64-apple-da
## (as 'lib' is unspecified)
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=packrat/bin/macosx/
##   Line starting '<!DOCTYPE HTML PUBLI ...' is malformed!
```

```
## Warning: package 'packrat' is not available (as a binary package for R
## version 3.4.1)
```

```
install.packages("foreign", repos = "https://CRAN.R-project.org/package=foreign")
```

```
## Installing package into '/Users/soniachakalo/Downloads/Final.Porject.Labs/packrat/lib/x86_64-apple-da
## (as 'lib' is unspecified)
```

```
## Warning: unable to access index for repository https://CRAN.R-project.org/package=foreign/bin/macosx/
##   Line starting '<!DOCTYPE HTML PUBLI ...' is malformed!
```

```
## Warning: package 'foreign' is not available (as a binary package for R
## version 3.4.1)
```

```
install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Installing package into '/Users/soniachakalo/Downloads/Final.Porject.Labs/packrat/lib/x86_64-apple-da
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/c5/l5rjsd_92hn_5rrsxxbmt_0r0000gn/T//RtmpirMclA/downloaded_packages
```

Downloading my data set and which variables i want to this lab and to Rstudio.

Read in your data with the appropriate function

replace with variable's you wish to add

```
data2 <- read_tsv("/Users/soniachakalo/Downloads/Final.Porject.Labs/ICPSR_23625_2002/DS0002/23625-0002-
```

```
## Error in read_tsv("/Users/soniachakalo/Downloads/Final.Porject.Labs/ICPSR_23625_2002/DS0002/23625-000
```

```
data_subset2 <- data2 %>%
  select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, BIASMO1, VTYP_I1, LOCCOD1, VTYP_I1, LOCCOD1)
```

```
## Error in data2 %>% select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, : could not find functio
```

```
# Read in your data with the appropriate function
data13 <- read_tsv("/Users/soniachakalo/Downloads/Final.Porject.Labs/ICPSR_36118_2013/DS0002/36118-0002-
```

```
## Error in read_tsv("/Users/soniachakalo/Downloads/Final.Porject.Labs/ICPSR_36118_2013/DS0002/36118-000
```

```
data_subset13 <- data13 %>%
  select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, BIASMO1, VTYP_I1, LOCCOD1, VTYP_I1, LOCCOD1)
```

```
## Error in data13 %>% select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, : could not find functi
```

```
# replace with variable's you wish to add
```

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

My data is very tidy. I limited as many varaibles as possible so my data would be easy and get straight to the point in terms of what i want to study. One thing i should look at is where to place the column variables in terms of the best sycnchronization almost like reading a book. For example, I should but state code then city name first in order to see where these crimes were committed. And then the Number of offenders, number of victims per offense and then what the bias is and so on.

Making my two datasets for 2002 and 2013 into the first 15 and last 15 rows.

```
tail(data_subset2, 15)
```

```
## Error in tail(data_subset2, 15): object 'data_subset2' not found
```

```
tail(data_subset13, 15)
```

```
## Error in tail(data_subset13, 15): object 'data_subset13' not found
```

```
head(data_subset2, 15)
```

```
## Error in head(data_subset2, 15): object 'data_subset2' not found
```

```
head(data_subset13, 15)
```

```
## Error in head(data_subset13, 15): object 'data_subset13' not found
```

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.

Converting my 2002 and 2013 data to a matrix so i can make a histogram

```
data.matrix(data_subset2[c(1:50), ])
```

```
## Error in is.data.frame(frame): object 'data_subset2' not found
```

```
data.matrix(data_subset13[c(1:50), ])
```

```
## Error in is.data.frame(frame): object 'data_subset13' not found
```

Histogram of the 2002 data of bias motives

```
hist(data_subset2$BIASMO1,
     main = "Histogram for motive bias of the offense",
     xlab = "VTYP_I1",
     ylab = "BIASMO1",
     border = "blue",
     col = "blue")
```

```
## Error in hist(data_subset2$BIASMO1, main = "Histogram for motive bias of the offense", : object 'data
```

Histogram of 2013 data of Bias motives

```
hist(data_subset13$BIASMO1,
     main = "Histogram for motive bias of the offense",
     xlab = "VTYP_I1",
     ylab = "BIASMO1",
     border = "blue",
     col = "blue")
```

```
## Error in hist(data_subset13$BIASMO1, main = "Histogram for motive bias of the offense", : object 'dat
```

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?

Making a bivariate scatterplot for my two data sets.

```
plot(data_subset2$BIASMO1)
```

```
## Error in plot(data_subset2$BIASMO1): object 'data_subset2' not found
```

```
plot(data_subset13$BIASMO1)
```

```
## Error in plot(data_subset13$BIASMO1): object 'data_subset13' not found
```

```
plot(data_subset2$TNUMVTMS)
```

```
## Error in plot(data_subset2$TNUMVTMS): object 'data_subset2' not found
```

```
plot(data_subset13$TNUMVTMS)
```

```
## Error in plot(data_subset13$TNUMVTMS): object 'data_subset13' not found
```

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.

Installed tidyr package

```
install.packages("tidyr")
```

```
## Installing package into '/Users/soniachakalo/Downloads/Final.Porject.Labs/packrat/lib/x86_64-apple-da
## (as 'lib' is unspecified)
```

```
##
## The downloaded binary packages are in
##  /var/folders/c5/l5rjsd_92hn_5rrsxxbmt_0r0000gn/T//RtmpirMclA/downloaded_packages
```

All of my columns in my data do represent variables. So i am using the gather function on the 2002 and 2013 data sets.

```
gathered.data_subset2 <- gather(data_subset2, -CITY, -STATECOD, -POP1, -TNUMVTMS, -TNUMOFF, -GOFFRAC, -
```

```
## Error in gather(data_subset2, -CITY, -STATECOD, -POP1, -TNUMVTMS, -TNUMOFF, : could not find functio
```

```
head(gathered.data_subset2, 15)
```

```
## Error in head(gathered.data_subset2, 15): object 'gathered.data_subset2' not found
```

```
gathered.data_subset13 <- gather(data_subset13, -CITY, -STATECOD, -POP1, -TNUMVTMS, -TNUMOFF, -GOFFRAC,
```

```
## Error in gather(data_subset13, -CITY, -STATECOD, -POP1, -TNUMVTMS, -TNUMOFF, : could not find functi
```

```
head(gathered.data_subset13, 15)
```

```
## Error in head(gathered.data_subset13, 15): object 'gathered.data_subset13' not found
```

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.

None of my data needs this.

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.

My variables classs for my data set is categories and numericals ranges pertaining to how many victims and offenders there were in the incident. The categorical variables pertain to a numeric associated with a certain type of bias motive, city name, and state. These types of classes do appropriatly resemble the data becasue you cannot rate bias or city or state as better then the other. You do need to have a ordered numeric in order ot know how many victims and offenders there were in the incident.

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)

I believe my data type as it is works really well in understanding the information.

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

I believe all of my strings are fine.

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.

In the Numeric version of my data set in question 3 when i had to convert my data to a matrix. But other than that my data does not have any NA's or -1 or missing values.

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.*

I do have some special values for example the values indicating the race of the offender which is indicated by a first letter that is associated with the race word itself. For example white race would be indicated as W and black would use B and asian A.

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?

Boxplots for 2 different variables of my data. TNUMOFF is the total number of offenders and TNUMVTMS is the total number of victims in the incident

```
boxplot(data_subset2$TNUMOFF)
```

```
## Error in boxplot(data_subset2$TNUMOFF): object 'data_subset2' not found
```

```
boxplot(data_subset13$TNUMOFF)
```

```
## Error in boxplot(data_subset13$TNUMOFF): object 'data_subset13' not found
```

```
boxplot(data_subset2$TNUMVTMS)
```

```
## Error in boxplot(data_subset2$TNUMVTMS): object 'data_subset2' not found
```

```
boxplot(data_subset13$TNUMVTMS)
```

```
## Error in boxplot(data_subset13$TNUMVTMS): object 'data_subset13' not found
```

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 handle my data in regards to outliers would be to take them out and see how it affects my data as well as change them to the appropriate measure of center.