Lab 10 - Merging Data

Yana Chakalo

November 2, 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 10 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. For your poster project, do you have multiple tables you'd like to join together to create your complete dataset? If so, describe what each table represents.

I do have two data sets that i want to use. The first data set are incident reports of all hate crimes reported and committed in the US in the year 2002 listed by their city/state. The other data set is the same data but for the year 2013. I want to use these two tables to look at hate crime trends within these two years and observe if they have siginificant similarities or differences concerning bias motives, crime type, victim/offender type. I am still not sure if I want to combine the datasets or look at them separetly and then compare.

2. What is/are your primary key(s)? If you have more than one table in your data, what is/are your foreign key(s)? Do your primary key(s) and foreign key(s) have the same name? If not, what does this mean for the way you need to specify potential data merges?

The primary keys in both my datasets are the bias motivation for each incident because this is like the id for every incident report. It tells us what the bias motivation behind the crime was. For example, there are a list of bias motivations corresponding to a number like 12 = anti-black, 43 = anti-homosexual, 52 = anti-mental disability. The primary key identifies each report (row) what type of hate crime was committed. However, the primary key and foreign key are not the same because bias motive is independent key to each data set and each row. My dataset does not have any foreign keys that link the datasets together. Both datasets have the same primary keys and column variables but none are used in one table to uniqueley identify within another table. Both datasets have unique different incident reports.

3. If you do not need to merge tables to create your final dataset, create a new dataset from your original dataset with a `grouped_by()` summary of your choice. You will use this separate dataset to complete the following exercises.

If you are merging separate tables as part of your data manipulation process, are your keys of the same data type? If not, what are the differences? Figure out the appropriate coercion process(es) and carry out the steps below.

Grouping the city and state in dataset 2002 and wanting to get the mean number of offenders for each city. Also counting the amount of different Bias motives within each year to figure out more data munging.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## 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
```

```
test2 <- group_by(data_subset2, CITY, STATECOD)
```

```
## Error in group_by(data_subset2, CITY, STATECOD): object 'data_subset2' not found
```

```
summarise(data_subset2, avg = mean(TNUMOFF))
```

```
## Error in summarise(data_subset2, avg = mean(TNUMOFF)): object 'data_subset2' not found
```

```
summarise(test2, avg = mean(TNUMOFF))
```

```
## Error in summarise(test2, avg = mean(TNUMOFF)): object 'test2' not found
```

```
count(test2, BIASMO1)
```

```
## Error in group_vars(x): object 'test2' not found
```

```
count(data_subset2, BIASMO1)
```

```
## Error in group_vars(x): object 'data_subset2' not found
```

2013 version

```
test13 <- group_by(data_subset13, CITY, STATECOD)
```

```
## Error in group_by(data_subset13, CITY, STATECOD): object 'data_subset13' not found
```

```
summarise(data_subset13, avg = mean(TNUMOFF))
```

```
## Error in summarise(data_subset13, avg = mean(TNUMOFF)): object 'data_subset13' not found
```

```
summarise(test13, avg = mean(TNUMOFF))
```

```
## Error in summarise(test13, avg = mean(TNUMOFF)): object 'test13' not found
```

```
count(test13, BIASMO1)
```

```
## Error in group_vars(x): object 'test13' not found
```

```
count(data_subset13, BIASMO1)
```

```
## Error in group_vars(x): object 'data_subset13' not found
```

4. Perform each version of the mutating joins (don't forget to specify the `by` argument) and print the results to the console. Describe what each join did to your datasets and what the resulting data table looks like. For those joining two separate datasets, did any of these joins result in your desired final dataset? Why or why not?

My data does not need to be mutated and joined in way because my data sets are independent in their observations and just need to be analyzed individually and then compared/contrasted.

5. Do the same thing with the filtering joins. What was the result? Give an example of a case in which a `semi_join()` or an `anti_join()` might be used with your primary dataset

Commiting an anti-join function to see which Bias motives are in only one data set and not in another. In order to see which Bias motives are only in 2013 or in 2002 within certain city/state.

```
test13 <- group_by(data_subset13, CITY, STATECOD)
```

```
## Error in group_by(data_subset13, CITY, STATECOD): object 'data_subset13' not found
```

```
BIAS13 <- count(test13, BIASMO1)
```

```
## Error in group_vars(x): object 'test13' not found
```

```
BIAS2 <- count(test2, BIASMO1)
```

```
## Error in group_vars(x): object 'test2' not found
```

```
anti_join(BIAS2, BIAS13, by = c("CITY", "STATECOD"))
```

```
## Error in anti_join(BIAS2, BIAS13, by = c("CITY", "STATECOD")): object 'BIAS2' not found
```

```
anti_join(BIAS13, BIAS2, by = c("CITY", "STATECOD"))
```

```
## Error in anti_join(BIAS13, BIAS2, by = c("CITY", "STATECOD")): object 'BIAS13' not found
```

My result of the anti-join showed me that were the bias motives for only the specific year either 2002 or 2013 and what city/state they were in. So for example, bias motive 41 is only in the 2002 data in Abingdon, VA and no where else. So this allowed me to see what was unique to each data set and to each city/state.

6. What happens when you apply the set operations joins to your tables? Are these functions useful for you for this project? Explain why or why not. If not, give an example in which one of them might be usefully applied to your data.

Anti-join was helpful as it let me look at the unique bias motives only in certain years and in certain cities. However, the inner join or left join is not helpful because the way my dataset is set up and the way i want to annalyze my data has nothing to do with combing the evidence and incident reports i have. Each incident report in each city and year has unique identifiers in telling us what type of hate crime it is. So my questions deal with knowing all of these incidents and comparing and looking at patterns between these incident reports. But I have to first analyze each dataset and incident report by themselves so joining any data would not be effective and no even possible to do.

7. If you have any reason to compare tables, apply `setequal()` below. What were the results? ## Comparing tables for 2002 and 2013 hate crime data

```
setequal(data_subset2, specifc = c[, 5000], data_subset13, specific = c[, 5000])
```

```
## Error in setequal(data_subset2, specifc = c[, 5000], data_subset13, specific = c[, : object 'data_su
```

I tried comparing these two data sets that i have but it will not let me because the number of rows are not the same.

8. What is the purpose of binding data and why might you need to take extra precaution when carrying out this specific form of data merging? If your data requires any binding, carry out the steps below and describe what was accomplished by your merge.

Using the bind rows function on both subset data to and creating a new table from it. But the only difference i see is that it added the amount of rows together from both datasets but i am not sure how the columns changed to accommodate the new rows.

```
both <- bind_rows(data_subset2, data_subset13)
```

```
## Error in eval_bare(dot$expr, dot$env): object 'data_subset2' not found
```

I do not think my data needs binding but i tried it any way to see if it changed my data at all and what it did. The purpose of data binding is if you have the same rows or columns in the same order you can tack them on into one dataset. There are two functions, bind_rows which combines two datasets that contain the exact same columns in the same order into a single dataset and bind_cols which combines two or more datasets that contain the same rows in the same order into a single dataset. You have to be careful with this data merging because it may not show mutate the table correctly and possibly not account for some of the data you have.

9. Do you need to merge multiple tables together using the same type of merge? If so, utilize the `reduce()` function from the `purrr` package to carry out the appropriate merge below.

I do not need to merge any data together.

10. Are there any other steps you need to carry out to further clean, transform, or merge your data into one, final, tidy dataset? If so, describe what they are and carry them out below.

I want to choose which hates crimes i need to look at their frequency and then how to plot those frequencies in plots. I also want to look at which bias motives within the entire year of 2002 and 2013 are most prevelent.

counting the frequency of the bias motives for each year.

```
bias_count2 <- count(data_subset2, BIASMO1)
```

```
## Error in group_vars(x): object 'data_subset2' not found
```

```
bias_count13 <- count(data_subset13, BIASMO1)
```

```
## Error in group_vars(x): object 'data_subset13' not found
```

Trying to Make the bias frequency into a plot or histogram

```
ggplot(data = bias_count2,
    mapping = aes(x = BIASMO1,
            y = frequency(1:2000)))
```

```
## Error in ggplot(data = bias_count2, mapping = aes(x = BIASMO1, y = frequency(1:2000))): could not fi
```

```
geom_line(bias_count2, mapping = aes(x = BIASMO1, y = frequency(1:2000)))
```

```
## Error in geom_line(bias_count2, mapping = aes(x = BIASMO1, y = frequency(1:2000))): could not find fu
```

Filtering out th specific hate crimes i want to understand.(Creating a table with variables of the Offenders race and what was the Bias motive that was committed by the offenders. 2002 and 2013 - W = White offender, BIASMO1 = 42 (Anti-Female homosexual))

```
GB_2002_W42 <- data_subset2 %>% filter(GOFFRAC == "W" & BIASMO1 == 42)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset2' not found
```

```
GB_2013_W42 <- data_subset13 %>% filter(GOFFRAC == "W" & BIASMO1 == 42)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset13' not found
```

Same filter function but different Bias motive. W = white offender, BIASMO1 = 12 (anti black)

```
GB_2002_W12 <- data_subset2 %>% filter(GOFFRAC == "W" & BIASMO1 == 12)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset2' not found
```

```
GB_2013_W12 <- data_subset13 %>% filter(GOFFRAC == "W" & BIASMO1 == 12)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset13' not found
```

Same filter. GOFFRAC = B = black. biasmo1 = 11 (anti-white)

```
GB_2002_B11 <- data_subset2 %>% filter(GOFFRAC == "B" & BIASMO1 == 11)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset2' not found
```

```
GB_2013_B11 <- data_subset13 %>% filter(GOFFRAC == "B" & BIASMO1 == 11)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset13' not found
```

Adding an offense type column because i think it is relevant.

```
data_subset2 <- data2 %>%
  select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, BIASMO1, VTYP_I1, LOCCOD1, OFFCOD1)
```

```
## Error in eval(lhs, parent, parent): object 'data2' not found
```

```
data_subset13 <- data13 %>%
  select(CITY, STATECOD, POP1, TNUMVTMS, TNUMOFF, GOFFRAC, BIASMO1, VTYP_I1, LOCCOD1, OFFCOD1)
```

```
## Error in eval(lhs, parent, parent): object 'data13' not found
```

Counting the frequency of the type of crime committed for each year.

```
OFFCOD1_count2 <- count(data_subset2, OFFCOD1)
```

```
## Error in group_vars(x): object 'data_subset2' not found
```

```
OFFCOD1_count13 <- count(data_subset13, OFFCOD1)
```

```
## Error in group_vars(x): object 'data_subset13' not found
```

This is telling us what the offense was among the incidents where the offender was black and the bias was white

```
GB_countO_2002_B11 <- data_subset2 %>% filter(GOFFRAC == "B" & BIASMO1 == 11) %>% count(OFFCOD1)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset2' not found
```

```
GB_countO_2013_B11 <- data_subset13 %>% filter(GOFFRAC == "B" & BIASMO1 == 11) %>% count(OFFCOD1)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset13' not found
```

This is telling us what the most common type of offense was committed within the incidents where the offender was white and the bias motive was ant-black

```
GB_countO_2002_W12 <- data_subset2 %>% filter(GOFFRAC == "W" & BIASMO1 == 12) %>% count(OFFCOD1)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset2' not found
```

```
GB_countO_2013_W12 <- data_subset13 %>% filter(GOFFRAC == "W" & BIASMO1 == 12) %>% count(OFFCOD1)
```

```
## Error in eval(lhs, parent, parent): object 'data_subset13' not found
```