Tutorial 2: Merging Answers

Leah Brooks

February 1, 2021

F. PS 2: Try it yourself with bigger data

1. Use R programming commands to fix the one problematic observation in the `student2` dataframe and make it merge properly (by first and last name) with `student1`.

```r
# load the first student data
# extra option is because character variables were loading as factors
student1 <- read.csv("H:/pppa_data_viz/2019/tutorial_data/lecture02/2019-01-27_fake_student_data.csv",
                     stringsAsFactors = FALSE)

# how many students?
student1
```

```
##   First_name last_name GWID gpa degree remaining_sem
## 1      elpis    josefa  G37 2.9    mpa             1
## 2    longina   nedelya  G12 3.9    mpp             4
## 3   richelle    bjoern  G08 3.4    mpp             1
## 4    mozghan      mara  G62 3.5    mpa             2
## 5       runa   marcelo  G14 3.8    mpp             3
```

```r
dim(student1)
```

```
## [1] 5 6
```

```r
# load the second student data
# extra option is because character variables were loading as factors
student2 <- read.csv("H:/pppa_data_viz/2019/tutorial_data/lecture02/2019-01-27_fake_student_data_part2.c
                     stringsAsFactors = FALSE)

# how many students?
dim(student2)
```

```
## [1] 5 4
```

```r
# what variables?
names(student2)
```

```
## [1] "First_name" "last_name"  "GWID"       "age"
```

```r
# make student2 w/o duplicate variables
student2.nogw <- student2[,c("First_name","last_name","age")]

# look at whether names are factors
str(student1)
```

```
## 'data.frame':    5 obs. of  6 variables:
```

```
##  $ First_name   : chr  "elpis" "longina" "richelle" "mozghan" ...
##  $ last_name    : chr  "josefa" "nedelya" "bjoern" "mara" ...
##  $ GWID         : chr  "G37" "G12" "G08" "G62" ...
##  $ gpa          : num  2.9 3.9 3.4 3.5 3.8
##  $ degree       : chr  "mpa" "mpp" "mpp" "mpa" ...
##  $ remaining_sem: int  1 4 1 2 3
```

I had to clean up a bunch of variables that became factors for some reason.

```
# do the merge
students <- merge(x = student1,
                  y = student2.nogw,
                  by = c("First_name","last_name"),
                  all = TRUE)

print("this is bad!")
```

```
## [1] "this is bad!"
```

```
students
```

```
##   First_name last_name GWID gpa degree remaining_sem age
## 1      elpis    josefa  G37 2.9    mpa             1  22
## 2    longina   nedelya  G12 3.9    mpp             4  25
## 3    mozghan      mara  G62 3.5    mpa             2  25
## 4   richelle    bjoern  G08 3.4    mpp             1  40
## 5       runa   marcelo  G14 3.8    mpp             3  NA
## 6       runa  marcelo  <NA>  NA   <NA>            NA  24
```

This is the problem. Can I fix?

```
# fix student2
student2.nogw$last_name <- ifelse(test = student2.nogw$last_name == "marcelo ",
                                  yes = "marcelo",
                                  no = student2.nogw$last_name)
str(student1)
```

```
## 'data.frame':    5 obs. of  6 variables:
##  $ First_name   : chr  "elpis" "longina" "richelle" "mozghan" ...
##  $ last_name    : chr  "josefa" "nedelya" "bjoern" "mara" ...
##  $ GWID         : chr  "G37" "G12" "G08" "G62" ...
##  $ gpa          : num  2.9 3.9 3.4 3.5 3.8
##  $ degree       : chr  "mpa" "mpp" "mpp" "mpa" ...
##  $ remaining_sem: int  1 4 1 2 3
```

```
str(student2.nogw)
```

```
## 'data.frame':    5 obs. of  3 variables:
##  $ First_name: chr  "elpis" "longina" "richelle" "mozghan" ...
##  $ last_name : chr  "josefa" "nedelya" "bjoern" "mara" ...
##  $ age       : int  22 25 40 25 24
```

```
# does the merge now work?
students <- merge(x = student1,
                  y = student2.nogw,
                  by = c("First_name","last_name"),
                  all = TRUE)

students
```

```
##   First_name last_name GWID gpa degree remaining_sem age
## 1      elpis    josefa  G37 2.9    mpa             1  22
## 2    longina   nedelya  G12 3.9    mpp             4  25
## 3    mozghan      mara  G62 3.5    mpa             2  25
## 4   richelle    bjoern  G08 3.4    mpp             1  40
## 5       runa   marcelo  G14 3.8    mpp             3  24
```

Yes – we're back down to five observations.

2. Find a dataset with county information and merge it either with the data we used last class, or with a larger county-level dataset for all counties in the US, 1910 to 2010. This second dataset is here, and the variable definitions are here.

It is sufficient for your new dataset to have just one year of information, or information for just one state or one year.

The final dataset could be at the county level, or at some other unit of observation. (For example, if you found a dataset about power plants, which had one observation per power plant with a county ID, you would end up with a power plant-level dataset.)

Make sure you explain the following:

- what your new data describe
- the source of your additional data
- how many observations the new data have
- how many observations you expect from the merge and why; if this seems off, fix it and explain what you did

This one is on you! I'm looking for a clear explanation of all of the above.

3. Make three relevant summary statistics of your choice for these new merged data. What these output are depends on what you're merging in.

Same – I'm checking for logical summary stats, and no missing values.