DataMergeLab

Liam Niehus-Staab

10/8/2018

#### On My Own Questions ####  
#1. Read the three files into R, naming them \*\*books\*\*, \*\*authors\*\*, and   
#\*\*book\_authors\*\*.  
books <- read.csv("~/Shared/F18MAT295/r-tutorials-master/merging/data/books.csv", as.is = TRUE)  
authors <- read.csv("~/Shared/F18MAT295/r-tutorials-master/merging/data/authors.csv", as.is = TRUE)  
book\_authors <- read.csv("~/Shared/F18MAT295/r-tutorials-master/merging/data/book-authors.csv", as.is = TRUE)

#2. Use the appropriate `join` statement to add the ISBNs to the \*\*authors\*\*   
#data table. Why does the resulting data frame have 31 rows instead of 11?  
isbn = left\_join(authors, book\_authors, by=c("authorid" = "authorID"))  
#since there are 5 cases of George Martin rows in each dataset, each row in the left  
#matches with each each row on the right 5 times. This causes George Martin to have  
#25 rows, whereas the other authors only match once. So they only get 1 row.   
head(isbn)

## authorid first.name last.name born died nationality ISBN  
## 1 martinge01 George Martin 1948 NA American 9.780554e+12  
## 2 martinge01 George Martin 1948 NA American 9.780553e+12  
## 3 martinge01 George Martin 1948 NA American 9.780554e+12  
## 4 martinge01 George Martin 1948 NA American 9.780554e+12  
## 5 martinge01 George Martin 1948 NA American 9.780554e+12  
## 6 martinge01 George Martin 1948 NA American 9.780554e+12

#3. To eliminate the duplicate rows of your data frame from #6 (which we'll assume   
#you named \*\*df2\*\*) run the following code (change the object names to align with   
#your code as necessary):   
isbn = unique(isbn)  
head(isbn)

## authorid first.name last.name born died nationality ISBN  
## 1 martinge01 George Martin 1948 NA American 9.780554e+12  
## 2 martinge01 George Martin 1948 NA American 9.780553e+12  
## 3 martinge01 George Martin 1948 NA American 9.780554e+12  
## 4 martinge01 George Martin 1948 NA American 9.780554e+12  
## 5 martinge01 George Martin 1948 NA American 9.780554e+12  
## 26 prachete01 Terry Pratchett 1948 2015 British 9.780061e+12

#4. Use the appropriate `join` statement to add the author information table from #3   
#to the \*\*books\*\* data table.   
books = left\_join(books,isbn, by = "ISBN")  
head(books)

## ISBN  
## 1 9.780554e+12  
## 2 9.780061e+12  
## 3 9.780061e+12  
## 4 9.780341e+12  
## 5 9.780425e+12  
## 6 9.780553e+12  
## Title Year  
## 1 A Game of Thrones 1996  
## 2 Good Omens: The Nice and Accurate Prophecies of Agnes Nutter, Witch 2006  
## 3 Good Omens: The Nice and Accurate Prophecies of Agnes Nutter, Witch 2006  
## 4 Dune 1965  
## 5 The Moon is a Harsh Mistress 1966  
## 6 Darkness at Noon 1940  
## Genre authorid first.name last.name born died nationality  
## 1 Fantasy martinge01 George Martin 1948 NA American  
## 2 Fantasy prachete01 Terry Pratchett 1948 2015 British  
## 3 Fantasy gaimane01 Neil Gaiman 1960 NA British  
## 4 Sci Fi herbefr01 Frank Herbert 1920 1986 American  
## 5 Sci Fi heinlro01 Robert Heinlein 1907 1988 American  
## 6 Historical koestar01 Arthur Koestler 1905 1983 Hungarian/British

#5. Are there any authors in the \*\*authors\*\* data table that do not correspond to   
#books in the \*\*books\*\* data table? Use an appropriate join statement to determine  
#the answer.  
sum(row(anti\_join(books, authors)))

## Joining, by = c("authorid", "first.name", "last.name", "born", "died", "nationality")

## [1] 0

#0  
# So there are no authors that appear in "authors" that arent in "books"

#6. After reading \*A Game of Thrones\* the student decides to read the rest of the  
#series over the summer. `books2.csv` contains the updated books on the student's   
#reading list. Read this file into R, naming it \*\*books2\*\*.  
books2 <- read.csv("~/Shared/F18MAT295/r-tutorials-master/merging/data/books2.csv", as.is = TRUE)

#7. Use the same join statement that you did in #4, but using \*\*books2\*\* rather   
#than \*\*books\*\*.  
books2 = left\_join(books2,isbn, by = "ISBN")  
head(books2)

## ISBN  
## 1 9.780554e+12  
## 2 9.780553e+12  
## 3 9.780554e+12  
## 4 9.780554e+12  
## 5 9.780554e+12  
## 6 9.780061e+12  
## Title Year  
## 1 A Game of Thrones 1996  
## 2 A Clash of Kings 1998  
## 3 A Storm of Swords 2000  
## 4 A Feast for Crows 2005  
## 5 A Dance with Dragons 2011  
## 6 Good Omens: The Nice and Accurate Prophecies of Agnes Nutter, Witch 2006  
## Genre authorid first.name last.name born died nationality  
## 1 Fantasy martinge01 George Martin 1948 NA American  
## 2 Fantasy martinge01 George Martin 1948 NA American  
## 3 Fantasy martinge01 George Martin 1948 NA American  
## 4 Fantasy martinge01 George Martin 1948 NA American  
## 5 Fantasy martinge01 George Martin 1948 NA American  
## 6 Fantasy prachete01 Terry Pratchett 1948 2015 British