Homework 3 - Joining Data Frames

Diana Perez

19/10/2021

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

##First we loaded the library that will allow us to import the files  
##Used the link of Raw from Github  
  
library(vroom)  
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

library(tidyr)  
sort\_pop1 <- vroom("https://raw.githubusercontent.com/chrit88/Bioinformatics\_data/master/Workshop%203/to\_sort\_pop\_1.csv")

## Rows: 30 Columns: 29

## -- Column specification --------------------------------------------------------  
## Delimiter: "\t"  
## chr (4): species, primary\_threat, secondary\_threat, tertiary\_threat  
## dbl (24): pop\_1\_2003-01-01, pop\_1\_2004-01-01, pop\_1\_2005-01-01, pop\_1\_2006-0...  
## lgl (1): pop\_1\_1995-01-01

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

sort\_pop2 <- vroom("https://raw.githubusercontent.com/chrit88/Bioinformatics\_data/master/Workshop%203/to\_sort\_pop\_2.csv")

## Rows: 30 Columns: 28

## -- Column specification --------------------------------------------------------  
## Delimiter: "\t"  
## chr (4): species, primary\_threat, secondary\_threat, tertiary\_threat  
## dbl (21): pop\_2\_2000-01-01, pop\_2\_2001-01-01, pop\_2\_2002-01-01, pop\_2\_2003-0...  
## lgl (3): pop\_2\_1996-01-01, pop\_2\_1997-01-01, pop\_2\_1998-01-01

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#visualizing information about the datas  
dim(sort\_pop1)

## [1] 30 29

dim(sort\_pop2)

## [1] 30 28

##Joining two dataframes   
## Natural join or Inner Join: To keep only rows that match from the data frames, specify the argument all=FALSE.  
## Full outer join or Outer Join:To keep all rows from both data frames, specify all=TRUE.  
## Left outer join or Left Join:To include all the rows of your data frame x and only those from y that match, specify x=TRUE.  
## Right outer join or Right Join:To include all the rows of your data frame y and only those from x that match, specify y=TRUE.  
  
#full\_join to maintain all rows from both data frames  
sort\_pop1 %>%  
 full\_join(sort\_pop2, by="species")

## # A tibble: 52 x 56  
## species primary\_threat.x secondary\_threa~ tertiary\_threat~ `pop\_1\_2003-01-~  
## <chr> <chr> <chr> <chr> <dbl>  
## 1 Schistid~ Habitat destruc~ <NA> <NA> NA  
## 2 Paraleuc~ Exploitation Habitat loss <NA> NA  
## 3 Scapania~ Climate change <NA> <NA> NA  
## 4 Seligera~ Exploitation <NA> <NA> NA  
## 5 Tortula ~ Habitat loss Pollution Climate change 96  
## 6 Pohlia m~ <NA> <NA> <NA> 288  
## 7 Bryum we~ Exploitation <NA> <NA> 81  
## 8 Ceratodo~ <NA> <NA> <NA> NA  
## 9 Trichoco~ Climate change <NA> <NA> 245  
## 10 Bryum kl~ Habitat loss <NA> <NA> 22  
## # ... with 42 more rows, and 51 more variables: pop\_1\_2004-01-01 <dbl>,  
## # pop\_1\_2005-01-01 <dbl>, pop\_1\_2006-01-01 <dbl>, pop\_1\_2007-01-01 <dbl>,  
## # pop\_1\_2008-01-01 <dbl>, pop\_1\_2009-01-01 <dbl>, pop\_1\_2010-01-01 <dbl>,  
## # pop\_1\_2011-01-01 <dbl>, pop\_1\_2012-01-01 <dbl>, pop\_1\_2013-01-01 <dbl>,  
## # pop\_1\_2014-01-01 <dbl>, pop\_1\_2015-01-01 <dbl>, pop\_1\_2016-01-01 <dbl>,  
## # pop\_1\_2017-01-01 <dbl>, pop\_1\_2018-01-01 <dbl>, pop\_1\_2019-01-01 <dbl>,  
## # pop\_1\_2000-01-01 <dbl>, pop\_1\_2001-01-01 <dbl>, pop\_1\_2002-01-01 <dbl>, ...

#re-shaping wide to long format  
##our data frame  
long <- sort\_pop1 %>%  
 ##and then apply this function   
 pivot\_longer(cols = -c(species:tertiary\_threat),  
 names\_to = "population date",  
 values\_to = "abundance")  
long

## # A tibble: 750 x 6  
## species primary\_threat secondary\_threat tertiary\_threat `population dat~  
## <chr> <chr> <chr> <chr> <chr>   
## 1 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2003-01-01  
## 2 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2004-01-01  
## 3 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2005-01-01  
## 4 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2006-01-01  
## 5 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2007-01-01  
## 6 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2008-01-01  
## 7 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2009-01-01  
## 8 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2010-01-01  
## 9 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2011-01-01  
## 10 Schistidiu~ Habitat destru~ <NA> <NA> pop\_1\_2012-01-01  
## # ... with 740 more rows, and 1 more variable: abundance <dbl>