Joining Data Tables

James Waterford

2023-03-14

Joining Data Tables

Shortcut: Use `crtl + opt + i` to create a code block.

```
library(readr)
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
```

```
surveys <- read_csv("../197-raw_storage/surveys.csv")
```

```
## Rows: 35549 Columns: 9

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): species_id, sex
## dbl (7): record_id, month, day, year, plot_id, hindfoot_length, weight
##
## 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.
```

```
species <- read.csv( "../197-raw_storage/species.csv")
plots <- read.csv("../197-raw_storage/plots.csv")
```

How do we combine tables?

We can use the *shared columns* between different data frames to combine them.

Specifically, we will use `_join` functions to combine two or mate data tables.

The different functions of `_join` allow us to combine in different ways:

`inner_join` - connects data sets, and removes all outliers.

`left_join` - keeps all values in x, and adds matching values for y

`right_join` - keeps all values in y, and adds matching values in x

`full_join` - keeps all values of x AND y

```
surveys %>%
inner_join(species, by = "species_id") -> join_table
head(join_table)
```

```
## # A tibble: 6 x 12
##   record_id month   day  year plot_id speci~1 sex   hindf~2 weight genus species
##       <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>   <chr>   <dbl>  <dbl> <chr> <chr>
## 1         1     7    16  1977       2 NL      M          32     NA Neot~ albigu~
## 2         2     7    16  1977       3 NL      M          33     NA Neot~ albigu~
## 3         3     7    16  1977       2 DM      F          37     NA Dipo~ merria~
## 4         4     7    16  1977       7 DM      M          36     NA Dipo~ merria~
## 5         5     7    16  1977       3 DM      M          35     NA Dipo~ merria~
## 6         6     7    16  1977       1 PF      M          14     NA Pero~ flavus
## # ... with 1 more variable: taxa <chr>, and abbreviated variable names
## #   1: species_id, 2: hindfoot_length
```

```
str(join_table)
```

```
## spc_tbl_ [34,786 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ record_id      : num [1:34786] 1 2 3 4 5 6 7 8 9 10 ...
##  $ month          : num [1:34786] 7 7 7 7 7 7 7 7 7 7 ...
##  $ day            : num [1:34786] 16 16 16 16 16 16 16 16 16 16 ...
##  $ year           : num [1:34786] 1977 1977 1977 1977 1977 ...
##  $ plot_id        : num [1:34786] 2 3 2 7 3 1 2 1 1 6 ...
##  $ species_id     : chr [1:34786] "NL" "NL" "DM" "DM" ...
##  $ sex            : chr [1:34786] "M" "M" "F" "M" ...
##  $ hindfoot_length: num [1:34786] 32 33 37 36 35 14 NA 37 34 20 ...
##  $ weight         : num [1:34786] NA NA NA NA NA NA NA NA NA NA ...
##  $ genus          : chr [1:34786] "Neotoma" "Neotoma" "Dipodomys" "Dipodomys" ...
##  $ species        : chr [1:34786] "albigula" "albigula" "merriami" "merriami" ...
##  $ taxa           : chr [1:34786] "Rodent" "Rodent" "Rodent" "Rodent" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   record_id = col_double(),
##   ..   month = col_double(),
##   ..   day = col_double(),
##   ..   year = col_double(),
##   ..   plot_id = col_double(),
##   ..   species_id = col_character(),
##   ..   sex = col_character(),
##   ..   hindfoot_length = col_double(),
##   ..   weight = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

This is good! Now, let's filter out any plot types that aren't "Control".

```
surveys %>%
  inner_join(plots, by = "plot_id") %>%
  filter(plot_type == 'Control') %>%
  head() %>%
  str()
```

```
## tibble [6 x 10] (S3: tbl_df/tbl/data.frame)
##  $ record_id      : num [1:6] 1 3 7 14 16 18
##  $ month          : num [1:6] 7 7 7 7 7 7
##  $ day            : num [1:6] 16 16 16 16 16 16
##  $ year           : num [1:6] 1977 1977 1977 1977 1977 ...
##  $ plot_id        : num [1:6] 2 2 2 8 4 2
##  $ species_id     : chr [1:6] "NL" "DM" "PE" "DM" ...
##  $ sex            : chr [1:6] "M" "F" "F" NA ...
##  $ hindfoot_length: num [1:6] 32 37 NA NA 36 22
##  $ weight         : num [1:6] NA NA NA NA NA NA
##  $ plot_type      : chr [1:6] "Control" "Control" "Control" "Control" ...
```

Finding Relevant Data

What if we don't know what columns both data sets have?

What if sorting through columns is too much of a tedious task.

```
intersect(colnames(surveys), colnames(species))
```

```
## [1] "species_id"
```

Joining Multiple Tables

```
surveys %>%
  inner_join(species, by = "species_id") %>%
  inner_join(plots, by = "plot_id") -> combined
head(combined)
```

```
## # A tibble: 6 x 13
##   record_id month   day  year plot_id speci~1 sex   hindf~2 weight genus species
##       <dbl> <dbl> <dbl> <dbl>   <dbl> <chr>   <chr>   <dbl>  <dbl> <chr> <chr>
## 1         1     7    16  1977       2 NL      M          32     NA Neot~ albigu~
## 2         2     7    16  1977       3 NL      M          33     NA Neot~ albigu~
## 3         3     7    16  1977       2 DM      F          37     NA Dipo~ merria~
## 4         4     7    16  1977       7 DM      M          36     NA Dipo~ merria~
## 5         5     7    16  1977       3 DM      M          35     NA Dipo~ merria~
## 6         6     7    16  1977       1 PF      M          14     NA Pero~ flavus
## # ... with 2 more variables: taxa <chr>, plot_type <chr>, and abbreviated
## #   variable names 1: species_id, 2: hindfoot_length
```

```
str(combined)
```

```
## spc_tbl_ [34,786 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ record_id      : num [1:34786] 1 2 3 4 5 6 7 8 9 10 ...
##  $ month          : num [1:34786] 7 7 7 7 7 7 7 7 7 7 ...
##  $ day            : num [1:34786] 16 16 16 16 16 16 16 16 16 16 ...
##  $ year           : num [1:34786] 1977 1977 1977 1977 1977 ...
##  $ plot_id        : num [1:34786] 2 3 2 7 3 1 2 1 1 6 ...
##  $ species_id     : chr [1:34786] "NL" "NL" "DM" "DM" ...
##  $ sex            : chr [1:34786] "M" "M" "F" "M" ...
##  $ hindfoot_length: num [1:34786] 32 33 37 36 35 14 NA 37 34 20 ...
##  $ weight         : num [1:34786] NA NA NA NA NA NA NA NA NA NA ...
##  $ genus          : chr [1:34786] "Neotoma" "Neotoma" "Dipodomys" "Dipodomys" ...
##  $ species        : chr [1:34786] "albigula" "albigula" "merriami" "merriami" ...
##  $ taxa           : chr [1:34786] "Rodent" "Rodent" "Rodent" "Rodent" ...
##  $ plot_type      : chr [1:34786] "Control" "Long-term Krat Exclosure" "Control" "Rodent Exclosure"
##  - attr(*, "spec")=
##   .. cols(
##   ..   record_id = col_double(),
##   ..   month = col_double(),
##   ..   day = col_double(),
##   ..   year = col_double(),
##   ..   plot_id = col_double(),
##   ..   species_id = col_character(),
##   ..   sex = col_character(),
##   ..   hindfoot_length = col_double(),
##   ..   weight = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
surveys %>%
  inner_join(species, by = "species_id") %>%
  inner_join(plots, by = "plot_id") %>%
  filter(plot_type == 'Control' | plot_type == 'Long-term Krat Exclosure') %>%
  select(year, genus, species, weight, plot_type, taxa) %>%
  filter(taxa == 'Rodent', !is.na(weight))
```

```
## # A tibble: 19,344 x 6
##     year genus       species  weight plot_type                taxa
##    <dbl> <chr>       <chr>     <dbl> <chr>                    <chr>
##  1  1977 Dipodomys   merriami     40 Long-term Krat Exclosure Rodent
##  2  1977 Dipodomys   merriami     29 Control                  Rodent
##  3  1977 Dipodomys   merriami     46 Control                  Rodent
##  4  1977 Dipodomys   ordii        52 Control                  Rodent
##  5  1977 Perognathus flavus        8 Control                  Rodent
##  6  1977 Onychomys   sp.          22 Long-term Krat Exclosure Rodent
##  7  1977 Perognathus flavus        7 Control                  Rodent
##  8  1977 Dipodomys   merriami     22 Control                  Rodent
##  9  1977 Perognathus flavus        8 Control                  Rodent
## 10  1977 Dipodomys   merriami     41 Control                  Rodent
## # ... with 19,334 more rows
```