Outline

Code ▾

dplyr for beginners

Earl F Glynn

2021-02-13

Several `palmerpenguins` examples are based on Rebecca Barter's Across (dplyr 1.0.0): applying dplyr functions simultaneously across multiple columns (http://www.rebeccabarter.com/blog/2020-07-09-across/)

1 Setup

1.1 1.2 Packages 1.3 Helper function

2 Overview of Palmer Penguins

2.1 `str`

Hide

```
str(penguins)        # `str` from utils package
```

```
tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

2.2 `glimpse`

Hide

```
glimpse(penguins)    # `glimpse` from `tibble` package
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, A...
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...
$ sex               <fct> male, female, female, NA, female, male, female, m...
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
```

3 Slices

3.1 `slice_head`

Selected first and last rows

Hide

```
penguins   %>%    # `%>%` is `magrittr` pipe operator; read "then"
  slice_head(n = 3)
```

```
# A tibble: 3 x 8
  species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex  
  <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
1 Adelie  Torge~           39.1          18.7              181        3750 male 
2 Adelie  Torge~           39.5          17.4              186        3800 fema~
3 Adelie  Torge~           40.3          18                195        3250 fema~
# ... with 1 more variable: year <int>
```

Outline

Hide

```
penguins            %>%
  slice_head(n = 3)  %>%
  Show()          # see section 1.3 "Helper function"
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007

Hide

```
penguins      %>%
  slice(1:3)  %>%
  Show()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007

Hide

```
penguins                    %>%    # Use this with databases
  filter(row_number() <= 3)  %>%
  Show()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007

3.2 slice_tail

Hide

```
penguins            %>%
  slice_tail(n = 3)  %>%
  Show()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Chinstrap	Dream	49.6	18.2	193	3775	male	2009
Chinstrap	Dream	50.8	19.0	210	4100	male	2009
Chinstrap	Dream	50.2	18.7	198	3775	female	2009

Hide

```
penguins                          %>%    # Use this with databases
  filter(row_number() >= n() - 2)  %>%
  Show()
```

Outline

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Chinstrap	Dream	49.6	18.2	193	3775	male	2009
Chinstrap	Dream	50.8	19.0	210	4100	male	2009
Chinstrap	Dream	50.2	18.7	198	3775	female	2009

3.3 Largest mass

3.3.1 top_n

Hide

```
penguins                %>%
  top_n(3, body_mass_g) %>%    # name `top_n` is superseded by `slice_max`
  Show()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Gentoo	Biscoe	49.2	15.2	221	6300	male	2007
Gentoo	Biscoe	59.6	17.0	230	6050	male	2007
Gentoo	Biscoe	51.1	16.3	220	6000	male	2008
Gentoo	Biscoe	48.8	16.2	222	6000	male	2009

3.3.2 slice_max

Hide

```
penguins                        %>%
  slice_max(body_mass_g, n = 3) %>%
  Show()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Gentoo	Biscoe	49.2	15.2	221	6300	male	2007
Gentoo	Biscoe	59.6	17.0	230	6050	male	2007
Gentoo	Biscoe	51.1	16.3	220	6000	male	2008
Gentoo	Biscoe	48.8	16.2	222	6000	male	2009

4 Frequency Counts

4.1 “Old” way

4.1.1 1D

Hide

```
table(penguins$species, useNA = "ifany")
```

```

   Adelie Chinstrap    Gentoo 
      152        68       124 
```

4.1.2 2D

Hide

```
table(penguins$species, penguins$island, useNA = "ifany")
```

Outline

```
           Biscoe Dream Torgersen
  Adelie        44    56        52
  Chinstrap      0    68         0
  Gentoo       124     0         0
```

4.2 “New” Way

4.2.1 1D

Hide

```
penguins         %>%
  count(species) %>%
  Show()
```

species	n
Adelie	152
Chinstrap	68
Gentoo	124

More general approach

Hide

```
penguins             %>%
  group_by(species)  %>%
  summarise(n = n(),
            .groups = "drop") %>%
  ungroup()          %>%
  Show()
```

species	n
Adelie	152
Chinstrap	68
Gentoo	124

4.2.2 2D

“long” format

Hide

```
penguins                      %>%
  group_by(species, island)   %>%
  summarise(n = n(),                  # sometimes must spell as `summarise`
            .groups = "drop") %>%
  ungroup()                   %>%
  Show()
```

species	island	n
Adelie	Biscoe	44
Adelie	Dream	56
Adelie	Torgersen	52
Chinstrap	Dream	68
Gentoo	Biscoe	124

“wide” format

Hide

Outline

```
penguins                            %>%
  group_by(species, island)         %>%
  summarise(n = n(),
            .groups = "drop")       %>%
  ungroup()                         %>%
  pivot_wider(names_from  = island,             # `tidyr` function
              values_from = n,
              values_fill = list(n = 0))  %>%
  Show()
```

species	Biscoe	Dream	Torgersen
Adelie	44	56	52
Chinstrap	0	68	0
Gentoo	124	0	0

`spread` is a `retired` function but often works “better” IMHO than `pivot_wider`

5 Gentoo Subset

Hide

```
gentoo <-
  penguins                      %>%
  filter(species == "Gentoo")  %>%
  select(starts_with("bill_"), sex, year)   # options:  `ends_width`, `contains`

dim(gentoo)
```

```
[1] 124   4
```

Hide

```
head(gentoo, 3) %>% Show()
```

bill_length_mm	bill_depth_mm	sex	year
46.1	13.2	female	2007
50.0	16.3	male	2007
48.7	14.1	female	2007

5.1 Summaries

5.1.1 Longer

Hide

```
gentooBySexYear <-
  gentoo              %>%
  group_by(sex, year)  %>%
  summarise(n = n(),
            across(c(bill_length_mm, bill_depth_mm),
                   mean, .names = "mean_{col}"),
            .groups = "drop")         %>%
  ungroup()                           %>%
  rename_with(function(x){str_replace_all(x, "mean_bill_|_mm", "")},
         starts_with("mean_bill_"))

gentooBySexYear %>%
  Show(caption = "Mean Bill Length and Bill Depth [mm]")
```

Mean Bill Length and Bill Depth [mm]

sex	year	n	length	depth
female	2007	16	45.06250	13.99375

Outline

sex	year	n	length	depth
female	2008	22	45.29545	14.13182
female	2009	20	46.26000	14.55000
male	2007	17	49.00000	15.36471
male	2008	23	48.53913	15.70435
male	2009	21	50.88095	16.01905
NA	2007	1	44.50000	14.30000
NA	2008	1	46.20000	14.40000
NA	2009	3	NA	NA

5.2 Wider

Hide

```
gentooBySexYear        %>%
  filter(!is.na(sex))  %>%
  pivot_wider(
    names_from  = year,
    values_from = c(n, length, depth)
  )                    %>%
  Show(caption = "Mean Bill Length and Depth [mm]")
```

Mean Bill Length and Depth [mm]

sex	n_2007	n_2008	n_2009	length_2007	length_2008	length_2009	depth_2007	depth_2008	depth_2009
female	16	22	20	45.0625	45.29545	46.26000	13.99375	14.13182	14.55000
male	17	23	21	49.0000	48.53913	50.88095	15.36471	15.70435	16.01905

6 Palmer Penguin Summaries

6.1 Global

Hide

```
penguins                   %>%
  summarise(
             n             = n(),

             nSpecies      = n_distinct(species),
             nameSpecies   = str_flatten(species %>% unique() %>% sort(),
                                         collapse = "|"),

             nIsland       = n_distinct(island),
             nameIsland    = str_flatten(island %>% unique() %>% sort(),
                                         collapse = "|"),

             nSex          = n_distinct(sex),
             nameSex       = str_flatten(sex %>% unique() %>% sort(),
                                         collapse = "|"),

             nYear    = n_distinct(year),
             minYear  = min(year, na.rm = TRUE),
             maxYear  = max(year, na.rm = TRUE)
           )          %>%
  Show()
```

n	nSpecies	nameSpecies	nIsland	nameIsland	nSex	nameSex	nYear	minYear	maxYear
344	3	Adelie|Chinstrap|Gentoo	3	Biscoe|Dream|Torgersen	3	female|male	3	2007	2009

6.2 With Across

Hide

```
flatString <- function(variable)
{
  str_flatten({{variable}}  %>%  unique()  %>%  sort(),
              collapse ="|")
}
```

Outline

Factor with NA to string “-missing-”

Hide

```
penguins              %>%
  mutate(sex = as.character(sex))             %>%  # force factor to character string
  mutate(sex = replace_na(sex, "-missing-"))  %>%

  summarise(
             n        = n(),
             across(c(species, island, sex, year), n_distinct, .names = "n_{col}"),
             across(c(species, island, sex, year), flatString, .names = "names_{col}")
           )          %>%
  Show()
```

n	n_species	n_island	n_sex	n_year	names_species	names_island	names_sex	names_year
344	3	3	3	3	Adelie|Chinstrap|Gentoo	Biscoe|Dream|Torgersen	-missing-|female|male	2007|2008|2009

6.3 Counts of missing values

Hide

```
penguins                             %>%
  summarise(across(everything(),
                   ~sum(is.na(.))))  %>%
  Show()
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
0	0	2	2	2	2	11	0

7 Recoding

Hide

```
penguinRecoded <-
  penguins                        %>%
  mutate(species =
           recode(species,
                  "Adelie"    = "Penguin1",
                  "Chinstrap" = "Penquin2",
                  "Gentoo"    = "Penguin3"),

         island =
           recode(island,
                  "Biscoe"    = "Island1",
                  "Dream"     = "Island2",
                  "Torgersen" = "Island3"),

         sex = replace_na(as.character(sex), "-missing")
        )
```

Hide

```
glimpse(penguinRecoded)
```

```
Rows: 344
Columns: 8
$ species           <fct> Penguin1, Penguin1, Penguin1, Penguin1, Penguin1,...
$ island            <fct> Island3, Island3, Island3, Island3, Island3, Isla...
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...
$ sex               <chr> "male", "female", "female", "-missing", "female",...
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
```

Outline

8 Joins

8.1 Add “info”

Add `dimension` info to `fact` table

Hide

```
info <- read_xlsx("infoPenguins.xlsx")

info %>% Show()
```

species	information
Adelie	common along the entire coast of the Antarctic continent
Chinstrap	inhabits a variety of islands and shores in the Southern Pacific and the Antarctic Oceans
Gentoo	species in the genus Pygoscelis, most closely related to the Adélie penguin and the chinstrap penguin

Hide

```
infoPenguins <-
  penguins  %>%
  inner_join(info, by = "species")

glimpse(infoPenguins)
```

```
Rows: 344
Columns: 9
$ species           <chr> "Adelie", "Adelie", "Adelie", "Adelie", "Adelie",...
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...
$ sex               <fct> male, female, female, NA, female, male, female, m...
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
$ information       <chr> "common along the entire coast of the Antarctic c...
```

8.2 Control selections

Use “tagging” especially with large lists for selections without much typing.

Hide

```
target <-
  read_xlsx("targetPenguins.xlsx")  %>%
  filter(tag == "x")                %>%  # Can type "x" reliably
  select(-tag)

target %>% Show()
```

species
Adelie
Chinstrap

Hide Outline

```
targetPenguins <-
  penguins  %>%
  inner_join(target, by = "species")

glimpse(targetPenguins)
```

```
Rows: 220
Columns: 8
$ species           <chr> "Adelie", "Adelie", "Adelie", "Adelie", "Adelie",...
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...
$ sex               <fct> male, female, female, NA, female, male, female, m...
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
```

Hide

```
targetPenguins  %>%  count(species)  %>%  Show()
```

species	n
Adelie	152
Chinstrap	68

Use `filter` instead of `inner_join`

Hide

```
targetPenguins2 <-
  penguins  %>%
  filter(species %in% target$species)

glimpse(targetPenguins2)
```

```
Rows: 220
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, A...
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torge...
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34....
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18....
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, ...
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 347...
$ sex               <fct> male, female, female, NA, female, male, female, m...
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2...
```

9 Quantiles

See dplyr 1.0.0: new summarise() features (https://www.tidyverse.org/blog/2020/03/dplyr-1-0-0-summarise/#quantiles).

Hide

```
Qs <- c(0, 0.10, 0.25, 0.50, 0.75, 0.90, 1.00)
```

9.1 `bill_length all`

Hide

```
quantile(penguins$bill_length_mm, Qs, na.rm = TRUE)
```

```
    0%    10%    25%    50%    75%    90%   100% 
32.100 36.600 39.225 44.450 48.500 50.800 59.600 
```

Hide

Outline

```
penguins       %>%
  summarise(quantiles   = Qs,
            bill_length = quantile(bill_length_mm, Qs, na.rm = TRUE))  %>%
  Show()
```

quantiles	bill_length
0.00	32.100
0.10	36.600
0.25	39.225
0.50	44.450
0.75	48.500
0.90	50.800
1.00	59.600

9.2 `bill_length` by `sex`

Hide

```
penguinQuantiles <-
  penguins       %>%
  group_by(sex)  %>%
  summarise(quantiles   = paste0("Q", 100*Qs),
            bill_length = quantile(bill_length_mm, Qs, na.rm = TRUE),
            .groups = "drop")  %>%
  ungroup()                    %>%
  pivot_wider(
    names_from  = quantiles,
    values_from = bill_length
  )

penguinQuantiles  %>%
  Show()
```

sex	Q0	Q10	Q25	Q50	Q75	Q90	Q100
female	32.1	35.78	37.600	42.8	46.200	47.50	58.0
male	34.6	38.80	40.975	46.8	50.325	51.93	59.6
NA	34.1	36.82	37.800	42.0	44.500	46.42	47.3

10 Other Useful Notes

- dplyr 1.0.0: working across columns (https://www.tidyverse.org/blog/2020/04/dplyr-1-0-0-colwise/).
- dplyr 1.0.0: select, rename, relocate (https://www.tidyverse.org/blog/2020/03/dplyr-1-0-0-select-rename-relocate/).
- Dario Radečić's How to Analyze Data with R: A Complete Beginner Guide to dplyr (https://appsilon.com/r-dplyr-tutorial/).
- Rasmus Bååth's The Tidyverse in a Table (http://www.sumsar.net/blog/2020/12/tidyverse-in-a-table/).
- Emily Riederer's Generating SQL with {dbplyr} and sqlfluff (https://emilyriederer.netlify.app/post/sql-generation/).
- HighlandR's Solving small data problems with data.table (https://johnmackintosh.com/2020-08-11-short-problems/).
- Tidyverse Tips (https://www.r-bloggers.com/2020/11/tidyverse-tips/).
- Understanding Non-Standard Evaluation (https://thomasadventure.blog/posts/understanding-nse-part1/).

11 Fini

11.1   11.2 Session Info

Processing time: 3.7 secs

2021-02-13 11:33