~~Five~~ Three ways to do "tidy" Pairwise Operations

- corrr

- recipes

- dplyover

Bryan Shalloway
Data Scientist, NetApp
2022-06-23

@brshallo   Bryan Shalloway   bryanshalloway.com

`combn()`

I. nest...
II. expand combos...
III. filter...
IV. map fun(s)...

`expand()`

pwiser: https://github.com/brshallo/pwiser/
Tidy Pairwise Operations (blog-post and record of examples):
https://www.bryanshalloway.com/2020/06/03/tidy-2-way-column-combinations/

Palmer Penguins

```r
library(tidyverse)
library(palmerpenguins)
library(corrr)
library(recipes)
# devtools::install_github("TimTeaFan/dplyover")
library(dplyover)
options(digits = 2)

ppenguins <- na.omit(penguins) %>%
  rename(bi_len_mm = bill_length_mm,
         bi_dep_mm = bill_depth_mm,
         fl_len_mm = flipper_length_mm)

ppenguins
#> # A tibble: 333 x 8
#>    species island    bi_len_mm bi_dep_mm fl_len_mm body_mass_g sex      year
#>    <fct>   <fct>         <dbl>     <dbl>     <int>       <int> <fct>   <int>
#>  1 Adelie  Torgersen      39.1      18.7       181        3750 male     2007
#>  2 Adelie  Torgersen      39.5      17.4       186        3800 female   2007
#>  3 Adelie  Torgersen      40.3      18         195        3250 female   2007
#>  4 Adelie  Torgersen      36.7      19.3       193        3450 female   2007
#>  5 Adelie  Torgersen      39.3      20.6       190        3650 male     2007
#>  6 Adelie  Torgersen      38.9      17.8       181        3625 female   2007
#>  7 Adelie  Torgersen      39.2      19.6       195        4675 male     2007
#>  8 Adelie  Torgersen      41.1      17.6       182        3200 female   2007
#>  9 Adelie  Torgersen      38.6      21.2       191        3800 male     2007
#> 10 Adelie  Torgersen      34.6      21.1       198        4400 male     2007
#> # ... with 323 more rows
```

@brshallo   Bryan Shalloway   bryanshalloway.com

Pairwise operations

```r
ppenguins %>%
  select(contains("_mm")) %>%
  cor()
#>             bi_len_mm bi_dep_mm fl_len_mm
#> bi_len_mm      1.00      -0.23      0.65
#> bi_dep_mm     -0.23       1.00     -0.58
#> fl_len_mm      0.65      -0.58      1.00
```

```
ppenguins %>%
  select(contains("_mm")) %>%
  corrr::correlate()
#> # A tibble: 3 x 4
#> term        bi_len_mm bi_dep_mm fl_len_mm
#> <chr>           <dbl>     <dbl>     <dbl>
#> 1 bi_len_mm        NA     -0.23      0.65
#> 2 bi_dep_mm     -0.23        NA     -0.58
#> 3 fl_len_mm      0.65     -0.58        NA
```

```
ppenguins %>%
  select(contains("_mm")) %>%
  corrr::correlate() %>%
  corrr::shave()
#> # A tibble: 3 x 4
#> term        bi_len_mm bi_dep_mm fl_len_mm
#> <chr>           <dbl>     <dbl>     <dbl>
#> 1 bi_len_mm        NA        NA        NA
#> 2 bi_dep_mm     -0.23        NA        NA
#> 3 fl_len_mm      0.65     -0.58        NA
```

```
ppenguins %>%
  select(contains("_mm")) %>%
  corrr::correlate() %>%
  corrr::focus(bi_len_mm)
#> # A tibble: 2 x 2
#>   term      bi_len_mm
#>   <chr>         <dbl>
#> 1 bi_dep_mm    -0.229
#> 2 fl_len_mm     0.653
```

@brshallo   Bryan Shalloway   bryanshalloway.com

Mapping Arbitrary Pairwise Functions

`corrr::colpair_map()`

```r
cor_p_value <- function(x, y) stats::cor.test(x, y)$p.value

ppenguins %>%
  select(contains("_mm")) %>%
  corrr::colpair_map(cor_p_value)
#> # A tibble: 3 x 4
#>   term       bi_len_mm bi_dep_mm fl_len_mm
#>   <chr>          <dbl>     <dbl>     <dbl>
#> 1 bi_len_mm NA          2.53e- 5  7.21e-42
#> 2 bi_dep_mm  2.53e- 5 NA          4.78e-31
#> 3 fl_len_mm  7.21e-42   4.78e-31 NA
```
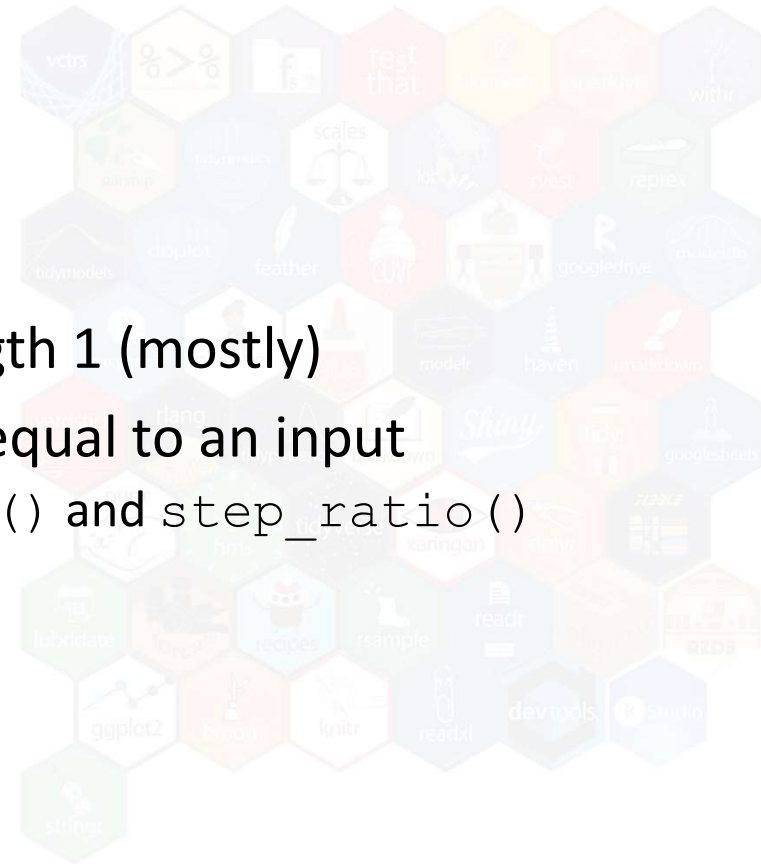
```r
joint_avg <- function(x, y) mean((x + y) / 2)

ppenguins %>%
  select(contains("_mm")) %>%
  corrr::colpair_map(joint_avg)
#> # A tibble: 3 x 4
#>   term       bi_len_mm bi_dep_mm fl_len_mm
#>   <chr>          <dbl>     <dbl>     <dbl>
#> 1 bi_len_mm      NA         30.6      122.
#> 2 bi_dep_mm      30.6       NA        109.
#> 3 fl_len_mm     122.       109.        NA
```

@brshallo   Bryan Shalloway   bryanshalloway.com

summarising vs mutating

- summarising operations return object of length 1 (mostly)
- mutating operations return object of length equal to an input
  - Is possible with `recipes::step_interact()` and `step_ratio()`

pairwise mutating operation ex

```r
rec <- recipe(body_mass_g ~ bi_len_mm + bi_dep_mm + fl_len_mm, data = ppenguins) %>%
  step_ratio(all_predictors(), denom = denom_vars(all_predictors()))

rec %>%
  prep(data = ppenguins) %>%
  juice() %>%
  glimpse()
#> Rows: 333
#> Columns: 10
#> $ bi_len_mm          <dbl> 39, 40, 40, 37, 39, 39, 39, 41, 39, 35, 37, 39, ~
#> $ bi_dep_mm          <dbl> 19, 17, 18, 19, 21, 18, 20, 18, 21, 21, 18, 19, ~
#> $ fl_len_mm          <int> 181, 186, 195, 193, 190, 181, 195, 182, 191, 198~
#> $ body_mass_g        <int> 3750, 3800, 3250, 3450, 3650, 3625, 4675, 3200, ~
#> $ bi_dep_mm_o_bi_len_mm <dbl> 0.48, 0.44, 0.45, 0.53, 0.52, 0.46, 0.50, 0.43, ~
#> $ fl_len_mm_o_bi_len_mm <dbl> 4.6, 4.7, 4.8, 5.3, 4.8, 4.7, 5.0, 4.4, 4.9, 5.7~
#> $ bi_len_mm_o_bi_dep_mm <dbl> 2.1, 2.3, 2.2, 1.9, 1.9, 2.2, 2.0, 2.3, 1.8, 1.6~
#> $ fl_len_mm_o_bi_dep_mm <dbl> 9.7, 10.7, 10.8, 10.0, 9.2, 10.2, 9.9, 10.3, 9.0~
#> $ bi_len_mm_o_fl_len_mm <dbl> 0.22, 0.21, 0.21, 0.19, 0.21, 0.21, 0.20, 0.23, ~
#> $ bi_dep_mm_o_fl_len_mm <dbl> 0.103, 0.094, 0.092, 0.100, 0.108, 0.098, 0.101,~
```

Some limitations with approaches mentioned

- recipes only supports a couple pairwise operations
- corrr doesn't support tidyselection or lambda functions
- neither supports grouping

... neither is really intended as a package for arbitrary pairwise operations

Would be nice if...

- Fit in a dplyr pipeline
- Consistent approach for both mutating and summarising pairwise operations
- https://github.com/TimTeaFan/dplyover

@brshallo   Bryan Shalloway   bryanshalloway.com

summarising example

```r
cor_p_value <- function(x, y) stats::cor.test(x, y)$p.value

ppenguins %>%
  summarise(across2x(.xcols = contains("_mm"),
                     .ycols = contains("_mm"),
                     .fns = cor_p_value,
                     .comb = "minimal"))
#> # A tibble: 1 x 3
#>   bi_len_mm_bi_dep_mm bi_len_mm_fl_len_mm bi_dep_mm_fl_len_mm
#>                 <dbl>               <dbl>               <dbl>
#> 1           0.0000253            7.21e-42            4.78e-31
```
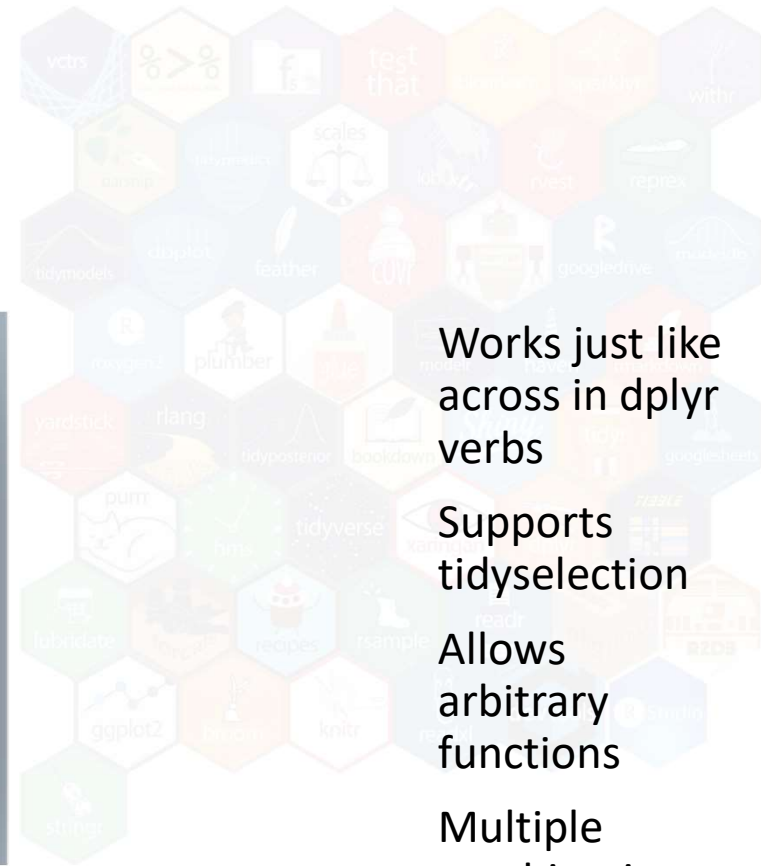
Works just like across in dplyr verbs

Supports tidyselection

Allows arbitrary functions

Multiple combination methods

summarising example 2

```r
cor_p_value <- function(x, y) stats::cor.test(x, y)$p.value

ppenguins %>%
  group_by(species) %>%
  summarise(n = n(),
            across2x(.xcols = contains("_mm"),
                     .ycols = contains("_mm"),
                     .fns = list(pval = ~cor_p_value(.x, .y)),
                     .names = "{fn}_{xcol}_{ycol}",
                     .comb = "minimal"))
#> # A tibble: 3 x 5
#>   species       n pval_bi_len_mm_bi_dep_mm pval_bi_len_mm_fl_l~ pval_bi_dep_mm_~
#>   <fct>     <int>                    <dbl>                <dbl>            <dbl>
#> 1 Adelie      146                  1.51e- 6             4.18e- 5          1.34e- 4
#> 2 Chinstrap    68                  1.53e- 9             4.92e- 5          2.16e- 7
#> 3 Gentoo      119                  7.34e-16             1.80e-16          1.40e-19
```

Respects groups

Other calculations

Named lists of funs

Lambda funs

Custom naming

mutating example

```
ppenguins %>%
  mutate(across2x(contains("_mm"),
                  contains("_mm"),
                  list(ratio = `/`, difference = `-`),
                  .names = "{fn}_{xcol}_{ycol}"))
#> # A tibble: 333 x 26
#>    species island    bi_len_mm bi_dep_mm fl_len_mm body_mass_g sex    year
#>    <fct>   <fct>         <dbl>     <dbl>     <int>       <int> <fct> <int>
#>  1 Adelie  Torgersen      39.1      18.7       181        3750 male   2007
#>  2 Adelie  Torgersen      39.5      17.4       186        3800 female 2007
#>  3 Adelie  Torgersen      40.3      18         195        3250 female 2007
#>  4 Adelie  Torgersen      36.7      19.3       193        3450 female 2007
#>  5 Adelie  Torgersen      39.3      20.6       190        3650 male   2007
#>  6 Adelie  Torgersen      38.9      17.8       181        3625 female 2007
#>  7 Adelie  Torgersen      39.2      19.6       195        4675 male   2007
#>  8 Adelie  Torgersen      41.1      17.6       182        3200 female 2007
#>  9 Adelie  Torgersen      38.6      21.2       191        3800 male   2007
#> 10 Adelie  Torgersen      34.6      21.1       198        4400 male   2007
#> # ... with 323 more rows, and 18 more variables:
#> #   ratio_bi_len_mm_bi_len_mm <dbl>, difference_bi_len_mm_bi_len_mm <dbl>,
#> #   ratio_bi_len_mm_bi_dep_mm <dbl>, difference_bi_len_mm_bi_dep_mm <dbl>,
#> #   ratio_bi_len_mm_fl_len_mm <dbl>, difference_bi_len_mm_fl_len_mm <dbl>,
#> #   ratio_bi_dep_mm_bi_len_mm <dbl>, difference_bi_dep_mm_bi_len_mm <dbl>,
#> #   ratio_bi_dep_mm_bi_dep_mm <dbl>, difference_bi_dep_mm_bi_dep_mm <dbl>,
#> #   ratio_bi_dep_mm_fl_len_mm <dbl>, difference_bi_dep_mm_fl_len_mm <dbl>, ...
```

References

- corrr: https://github.com/tidymodels/corrr
- recipes: https://github.com/tidymodels/recipes
- dplyover: https://github.com/TimTeaFan/dplyover

(Stuff that was cut)

- widyr: https://github.com/dgrtwo/widyr
- pwiser: https://github.com/brshallo/pwiser/
- Tidy Pairwise Operations (blog-post and record of examples): https://www.bryanshalloway.com/2020/06/03/tidy-2-way-column-combinations/

@brshallo   Bryan Shalloway   bryanshalloway.com