Data wrangling across columns

Warmup activity

Work on the activity (handout) with a neighbor, then we will discuss as a class

Warmup

Your friend writes the following code:

```r
starwars |>
  summarize(across(where(is.numeric),
                   list("median" = median),
                   .names = "{.fn}_{.col}"))
```

```
# A tibble: 1 × 3
  median_height median_mass median_birth_year
          <int>       <dbl>             <dbl>
1            NA          NA                NA
```

Why are they getting NAs?

Ignoring NAs

```r
starwars |>
  summarize(median_height = median(height))
```

```
# A tibble: 1 × 1
  median_height
          <int>
1            NA
```

What would I change to ignore missing values (NAs) when computing the median?

Ignoring NAs

```r
starwars |>
  summarize(median_height = median(height, na.rm=T))
```

```
# A tibble: 1 × 1
  median_height
          <int>
1           180
```

Now let's try with `across`...

Ignoring NAs

```r
starwars |>
  summarize(across(where(is.numeric),
                   list("median" = median(na.rm=T))))
```

```
Error in `summarize()`:
ℹ In argument: `across(where(is.numeric), list(median = median(na.rm =
  T)))`.
Caused by error in `median.default()`:
! argument "x" is missing, with no default
```

Why is this code failing?

Ignoring NAs

```
starwars |>
  summarize(across(where(is.numeric),
                   list("median" = median(na.rm=T))))
```

no longer a function
instead, it is function output
(which is failing)

median is a function:

```
median
```

```
function (x, na.rm = FALSE, ...)
UseMethod("median")
<bytecode: 0x13cf29ab8>
<environment: namespace:stats>
```

median() is *evaluating* (calling) the function:

```
median()
```

← problem: not calculating median of anything

```
Error in median.default(): argument "x" is missing, with no default
```

Ignoring NAs

```r
starwars |>
  summarize(across(where(is.numeric),
                   list("median" = median(na.rm=T))))
```

```
Error in `summarize()`:
ℹ In argument: `across(where(is.numeric), list(median = median(na.rm =
  T)))`.
Caused by error in `median.default()`:
! argument "x" is missing, with no default
```

What should we do instead?

Writing a new function

We want a function that calculates the median without the NAs, so we can do something like

```r
starwars |>
  summarize(across(where(is.numeric),
                   list("median" = median_no_na)))
```

However, this median_no_na function doesn't exist. We have to write it ourselves!

Writing a new function

```r
median_no_na <- function(x) {
  median(x, na.rm = T)
}
```

arguments to the function

x : input to the function (vector we want median of)

function name (should be informative)

body of function (what happens when we call the function)

Calculte median of input (x), removing missing values

output: median of x (ignoring NAs)

Writing a new function

```r
median_no_na <- function(x) {
  median(x, na.rm = T)
}
```

What will each of the following lines return?

```r
median(c(0, 1, 2, NA))
```

NA

```r
median(c(0, 1, 2, ~~NA~~), na.rm=T)
```

1

```r
median_no_na(c(0, 1, 2, NA))
```

1

→ x = c(0, 1,2, NA)

↳ median(c(0,1,2,NA), na.rm=T)

↳ 1

Writing a new function

```r
median_no_na <- function(x) {
  median(x, na.rm = T)
}

median(c(0, 1, 2, NA))
```

```
[1] NA
```

```r
median(c(0, 1, 2, NA), na.rm=T)
```

```
[1] 1
```

```r
median_no_na(c(0, 1, 2, NA))
```

```
[1] 1
```

Writing a new function

```r
median_no_na <- function(x) {
  median(x, na.rm = T)
}

starwars |>
  summarize(across(where(is.numeric),
                   list("median" = median_no_na)))
```

```
# A tibble: 1 × 3
  height_median mass_median birth_year_median
          <int>       <dbl>             <dbl>
1           180          79                52
```

What would I change if I want to calculate the mean instead of the median?

Writing a new function

```r
mean_no_na <- function(x) {
  mean(x, na.rm = T)
}

starwars |>
  summarize(across(where(is.numeric),
                   list("mean" = mean_no_na)))
```

```
# A tibble: 1 × 3
  height_mean mass_mean birth_year_mean
        <dbl>     <dbl>           <dbl>
1        175.      97.3            87.6
```

Will we need to use the `mean_no_na` or `median_no_na` functions many times?

Anonymous functions

If we don’t need a function repeatedly, we can make an *anonymous* function instead:

(defined when we need it; don't give it a name)

Anonymous function:

```
starwars |>
  summarize(across(where(is.numeric),
                   list("mean" = function(x) mean(x, na.rm=T))))
```

function definition, where we need it

```
# A tibble: 1 × 3
  height_mean mass_mean birth_year_mean
        <dbl>     <dbl>           <dbl>
1        175.      97.3            87.6
```

for short functions, can make anonymous if we won't use them repeatedly

Class activity

https://sta279-f25.github.io/class_activities/ca_08.html

- Work with a neighbor on the class activity
- At the end of class, submit your work as an HTML file on Canvas (one per group, list all your names)