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
1  starwars |>
2    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

```
1  starwars |>
2    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

```
1  starwars |>
2    summarize(across(where(is.numeric),
3                     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

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

`median` is a function:

```
1  median
```

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

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

```
1  median()
```

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

Ignoring NAs

```
1  starwars |>
2    summarize(across(where(is.numeric),
3                     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
1 starwars |>
2   summarize(across(where(is.numeric),
3                   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
1  median_no_na <- function(x) {
2    median(x, na.rm = T)
3  }
```

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))
```

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

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

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

```
1  median_no_na <- function(x) {
2    median(x, na.rm = T)
3  }
4
5  starwars |>
6    summarize(across(where(is.numeric),
7                     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:

Anonymous function:

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

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

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)