Chapter 2: Causality
Data Transformation with Tidyverse Functions

Claire Liow

University of Tokyo

5/31/2022

1 Aggregate data with `group_by()`

2 Summarize data with `summarize()`

3 Reshape the data with `pivot_wider()`

4 Summary

Section 1

Aggregate data with `group_by()`

What does it do?

- Group data for downstream analysis
- Commonly used with `summarize` and `mutate`

```
## count the observation in selected variables
race.call.summary <- resume %>%
  group_by(race, call) %>%
  count()

## calculate callback rates
callback_by_race <- resume %>%
  group_by(race, sex) %>%
  summarize(callback_rate = mean(call))
```

Section 2

Summarize data with `summarize()`

What does it do?

- Group ⤳ Summarize
- Collapse each group into a single row summary

```r
## calculate callback rate by race
resume %>%
group_by(race) %>%
summarize(callback = mean(call))

## calculate callback rate by race and sex
resume %>%
group_by(race, sex) %>%
summarize(callback = mean(call))
```

Section 3

Reshape the data with `pivot_wider()`

What does it do?

- Increasing the number of columns
- Decreasing the number of rows

```r
## tidyverse
## without pivoting the data
resume %>%
group_by(race, sex) %>%
summarize(callback = mean(call))

## after pivoting the data
resume %>%
group_by(race, sex) %>%
summarize(callback = mean(call)) %>%
pivot_wider(names_from = race, values_from = callback)
```

Section 4

Summary

Today's tidyverse functions:

- `group_by()` to aggregate data
- `summarize()` to summarize data
- `pivot_wider()` to reshape data