Chapter 2: Causality

Data Transformation with Tidyverse

Sho Miyazaki

Keio University

5/26/2022

- Introduction
- 2 dplyr and %>%
- Summary

Section 1

Introduction

Let's get started with Data

```
## load packages
library(tidyverse)
library(qss)
## load data
resume <- read csv("data/resume.csv")
# check data
resume
## # A tibble: 4,870 x 4
##
     firstname sex
                 race
                           call
## <chr> <chr> <chr> <chr> <dbl>
## 1 Allison female white
   2 Kristen female white
##
## 3 Lakisha female black
   4 Latonya female black
##
##
   5 Carrie female white
##
   6 Jay male white
##
   7 Jill female white
##
   8 Kenya female black
                              0
   9 Latonya female black
```

Today's Goal

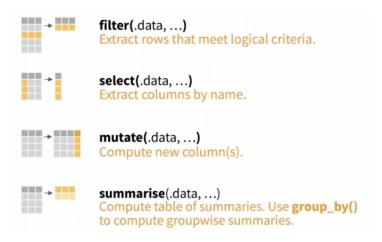
Combine functions to get informative output

sex	black	white	race_gap
female male		0.0989247 0.0886957	

Tools



dplyr from Tidyverse



Source: RStudio

Section 2

dplyr and %>%

What is "pipe %>%"?



• "a good way to pronounce %>% when reading code is "then"."

Source: R for Data Science

Extract Rows (filter)

• filter: Return rows by name/number/etc.

subset data with black names

```
resume %>%
 filter(race == "black")
## # A tibble: 2,435 x 4
##
     firstname sex race
                           call
##
     <chr> <chr> <chr> <chr> <chr> <dbl>
##
  1 Lakisha female black
   2 Latonya female black
##
   3 Kenya female black
##
##
   4 Latonya female black
   5 Tyrone male black
##
##
   6 Aisha female black
##
   7 Aisha female black
   8 Aisha female black
##
##
     Tamika
              female black
```

Extract Columns (select)

• select: Return columns by name/number/etc.

```
## Subset with sex and race columns
resume %>%
  select(sex, race)
  # A tibble: 4,870 x 2
##
      sex
             race
##
      <chr> <chr>
##
   1 female white
##
    2 female white
##
   3 female black
   4 female black
##
##
    5 female white
##
    6 male white
##
  7 female white
```

8 female black

female black

##

##

Compute New Columns (mutate)

mutate

```
## # A tibble: 4,870 x 5
##
  firstname sex race call type
## <chr> <chr> <chr> <chr> <dbl> <chr>
  1 Allison female white 0 WhiteFemale
##
   2 Kristen female white 0 WhiteFemale
##
##
   3 Lakisha female black 0 BlackFemale
## 4 Latonya female black 0 BlackFemale
##
   5 Carrie female white 0 WhiteFemale
##
   6 Jay male white 0 WhiteMale
## 7 Jill female white 0 WhiteFemale
## 8 Kenya female black 0 BlackFemale
   9 Latonya female black 0 BlackFemale
##
## 10 Tyrone male black
                            0 BlackMale
## # ... with 4.860 more rows
```

Compute Table Summaries (summarise)

```
## callback rate for black female names
resume %>%
  filter(race == "black" & sex =="female") %>%
  summarize(callback_rate = mean(call, na.rm = TRUE))
## # A tibble: 1 x 1
## callback rate
```

<dbl>

##

1

Section 3

Summary

Overwhelmed?

Don't worry!

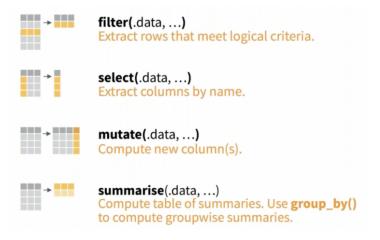
There are many resources you can use, and you don't have to memorize all the functions.

- QSS Textbook
 - Tidyverse Version is on Perusall
- Cheetsheets
 - Search "tidyverse cheetsheets"
 - https://www.rstudio.com/resources/cheatsheets/
- Online Resources
 - Google "tidyverse add column error"
 - official reference page, stackoverflow, RPubs, etc.

Teaching Team

We are here for you!

Let's practice!



Source: RStudio