

Intro to Data Manipulation

Joscelin Rocha Hidalgo

(she,her,hers)

@JoscelinRocha

**Slides adapted from David Keyes (@dgkeyes), inspired by Danielle Navarro
(@djnavarro) and Paul Campbell (@paulcampbell91)**

Child Health and Development Studies (CHDS)

"Birth weight, date, and gestational period collected as part of the Child Health and Development Studies in 1961 and 1962. Information about the baby's parents – age, education, height, weight, and whether the mother smoked is also recorded."



Tidyverse



Base R Syntax

```
table(data$age,  
      data$ed)
```

- In this example, table is our function and our data frame goes in the parentheses
- The dollar sign is used to specify the column or variable we want
(data_frame\$variable)

Con: you have to refer to the database every time (annoying)

Pro: there are thousands of add-on packages available for R at CRAN

Tidyverse Syntax

```
data %>%  
  tabyl(age, ed)
```

- This is the same as earlier but did you notice I only referred to my dataframe once?
- The pipes! (%>%). They chain together a series of functions

The Pipe



this is not a pipe

L
U

The Pipe

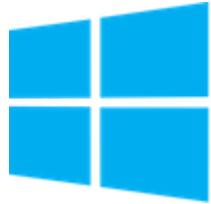
I would read each pipe as "then." For example:

```
data %>%
  filter(age < 25) %>%
  group_by(ed) %>%
  summarize(mean_gestation = mean(gestation, na.rm = TRUE))
```

Art by @allison_horst



Shortcuts



Windows

control-shift-M



Mac

command-shift-M

Your Turn

1. Open the 03-data-manipulation-exercises.Rmd file
2. Load the `tidyverse` package
3. Import the Child Health and Development Studies dataset:
`data/chds6162_data.csv` to a data frame called `data`
4. I will walk you through some examples and then it will be your turn
5. These are the functions we will go over:

1. select	5. summarize
2. mutate	6. group_by
3. case_when	7. across
4. filter	8. relocate