Wrangling Data in the Tidyverse

Joscelin Rocha-Hidalgo
(she, her, hers)
@JoscelinRocha

Tuesday, Oct 4th, 2022
R-Ladies St. Louis 2022

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

Agenda

Logistics

Quick intro to RMarkdown

Our dataset

Data Wrangling

Tips & Resources

Logistics

Download the materials: https://github.com/Joscelinrocha/Rladies-Wrangling-Data-in-the-Tidyverse

R markdown

Art by @allison_horst

RMarkdown Overview

Every RMarkdown document has the following:

Knitting (aka Export)

YAML

```yaml
---
title: "This workshop is awesome"
author: "Joscelin Rocha Hidalgo"
date: "07/18/2020"
output: word_document
---
```

Stands for "YAML Ain't Markup Language"

Where you add title, author, date, output options, etc.

Text


---
title: "My Super Fancy Report"
author: "David Keyes"
output: html_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
Text
```{r cars}
summary(cars)
```
Including Plots
You can also embed plots, for example:
Text
```{r pressure, echo=FALSE}
plot(pressure)
```

Text

Markdown

```
Text with **some words in bold**
and *some words in italics*
```

Output

Text with **some words in bold** and *some words in italics*

Headers

Markdown

```
# First-Level Header

## Second-Level Header

### Third-Level Subheader
```

Output

First-Level Header

Second-Level Header

Third-Level Subheader

Lists

Markdown

```
- Bulleted list item
- Bulleted list item


1. Numbered list item
1. Numbered list item
```

Output

- Bulleted list item #1
- Bulleted list item #2

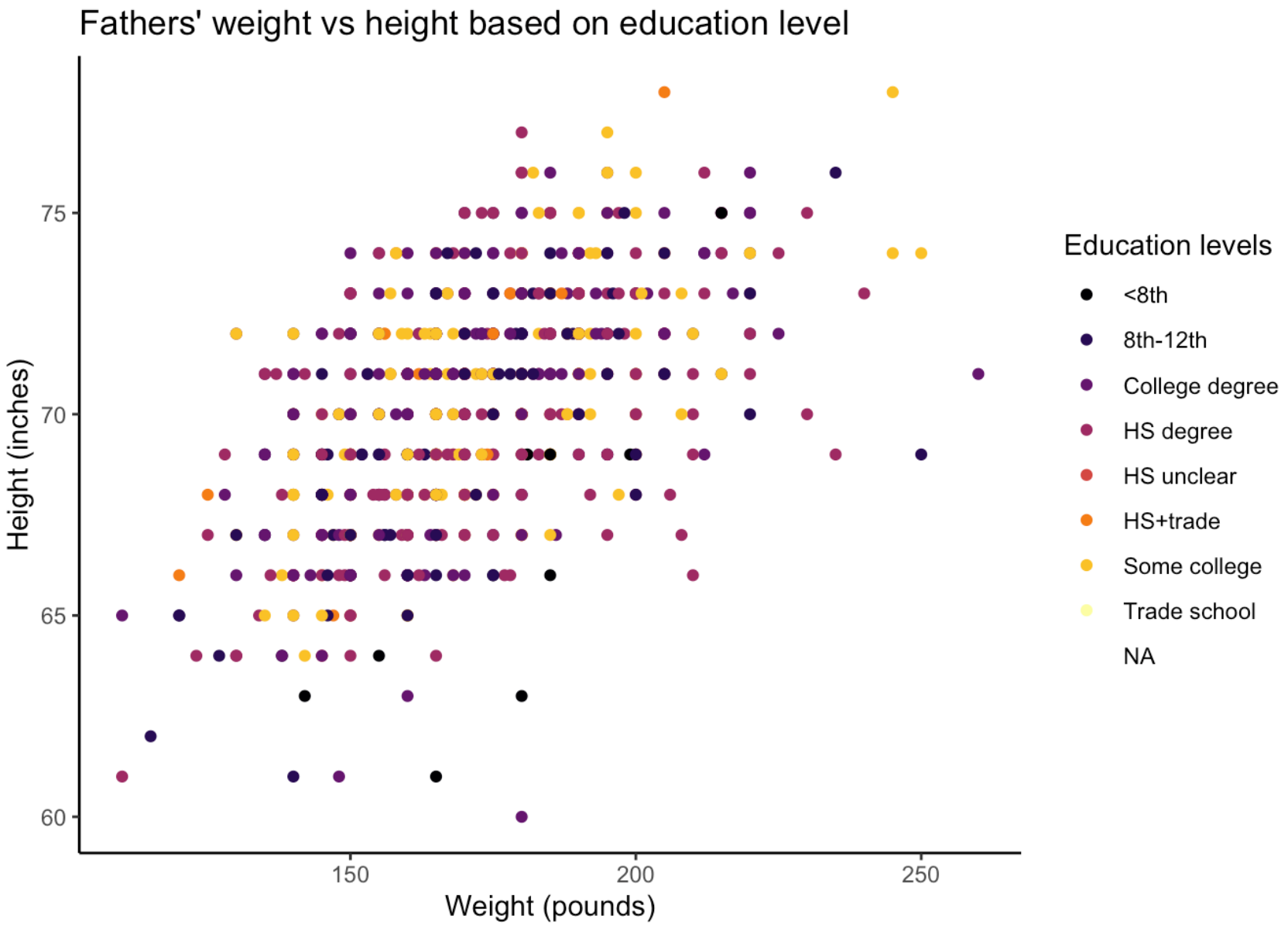
1. Numbered list item #1
2. Numbered list item #2

Code Chunk

They start with three backticks and {r} and end with three backticks.

````
```{r}
ggplot(data,aes(dwt,dht, color = ded_lbls)) +
  geom_point() + scale_color_viridis_d(option = "inferno") +
  labs (title = "Fathers' weight vs height based on education level",
        x = "Weight (pounds)",
        y = "Height (inches)",
        color = "Education levels") +
  theme_classic()

```
````


Fathers' weight vs height based on education level
Height (inches)
75
70
65
60
150
200
250
Weight (pounds)
Education levels
<8th
8th-12th
College degree
HS degree
HS unclear
HS+trade
Some college
Trade school
NA

Insert a Code Chunk: Button

Insert a Code Chunk: Keyboard Shortcut

Windows

control+alt+i

Mac

command+option+i

Chunk Options

Other options that we won't discuss today:

- **`warning`** (show any warnings that R throws)
- **`message`** (show any messages that R sends)
- **`fig.width`** (default figure width)
- **`fig.height`** (default figure height)
- **`echo`** (show the R code in the knitted report)
- and many more ...

Setup Code Chunk

A special code chunk with the text `setup` right after the `r`.

````
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
````

All chunk options can be set at the **global level** (in the setup code chunk) or at the **chunk level** (for individual chunks).

Options at the individual chunk level **override** global chunk options.

Our Dataset

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."

- Website: https://www.stat.berkeley.edu/users/statlabs/papers/sample.pdf
- R package: https://vincentarelbundock.github.io/Rdatasets/doc/mosaicData/Gestation.html

Data from the Child Health and Development Studies

Description

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.

Usage

```
data(Gestation)
```

Format

A data frame with 1236 observations on the following variables.

- `id` identification number
- `plurality` 5 = single fetus
- `outcome` 1 = live birth that survived at least 28 days
- `date` birth date where 1096=January 1, 1961
- `gestation` length of gestation (in days)
- `wt` birth weight (in ounces)
- `parity` total number of previous pregnancies (including fetal deaths and still births)
- `race` mother's race: 0-5=white 6=mex 7=black 8=asian 9=mixed

Tidyverse


stringr
forcats
tidyverse
www.rstudio.com
dplyr
tidyr
unknown

The Pipe

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

- These are the functions we will go over:

1. rename
2. clean_names
3. toupper/tolower
4. separate/unite
5. select
6. filter
7. mutate
8. case_when
9. summarize
10. group_by
11. relocate
12. pivot_longer/pivot_wider