PREREQUISITES

- We’ll be using **R** and **R Studio**
  - Download and Install **R**: https://cran.rstudio.com
  - Download and Install **R Studio**: https://www.rstudio.com/products/rstudio/download/
- Download the workshop source materials at:
  - github.com/jrogol/Intro2dplyr/
- Unzip and open **Intro2dplyr.Rproj**
  - In the console, type `source(“0-Prerequisites.R”)`
    - This will update the necessary packages.

Introduction to **dplyr** & the **tidyverse**

James Rogol

August 2017

AGENDA

- Background
  - What is “tidy” data?
  - What is the **tidyverse**?
- Manipulating data with **dplyr**
  - Single-table functions
  - Integrating **magrittr** and the **pipe**
  - Summary Functions
- Resources
- References

What is “tidy” data?

"All happy families are alike; every unhappy family is unhappy in its own way."

- Leo Tolstoy, *Anna Karenina*

"Tidy datasets are all alike, but every messy dataset is messy in its own way."

- Hadley Wickham

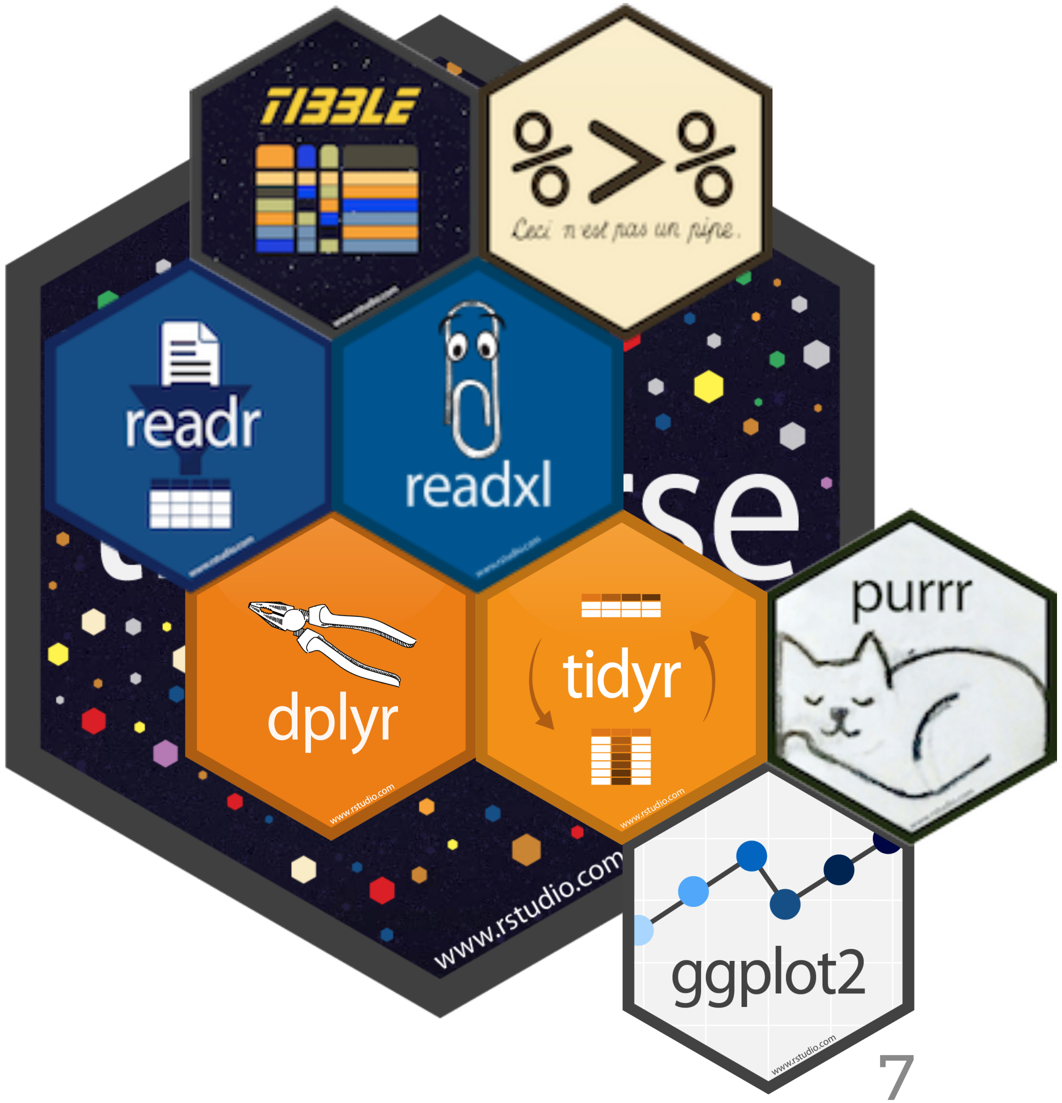
What is “tidy” data?

- Real-world data is messy
  - The 80% aphorism
- Standardizing data structures
  - Each column represents a unique variable
  - Each row corresponds to an observation
- “tidy” data facilitates analysis

What is the **tidyverse**?

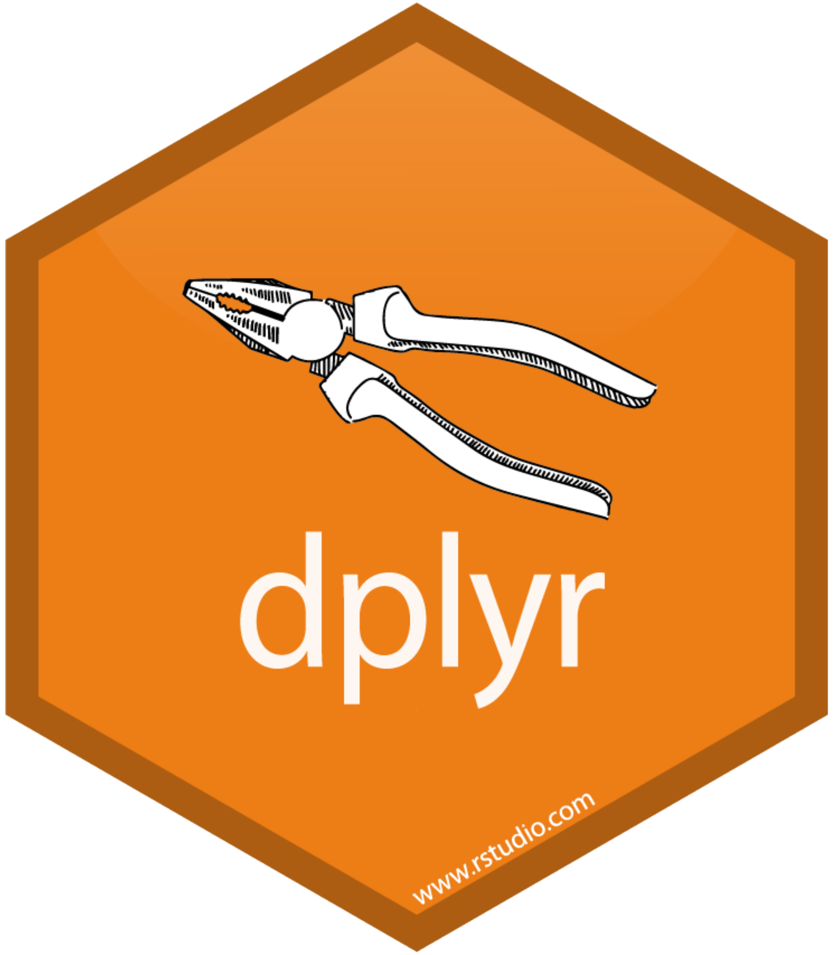
The **tidyverse**

- Suite of R **packages**
  - Import, Tidy, Transform
  - Visualize and Model
  - Program
- Uses the **tibble** to store data
  - Simplified **Data Frames**



Manipulating data with **dplyr**

- Brings database-like queries to R
  - `SELECT * FROM TABLE WHERE package = “dplyr”`
  - Single-table
  - Multi-table
- Faster than base R and its predecessor, **plyr**

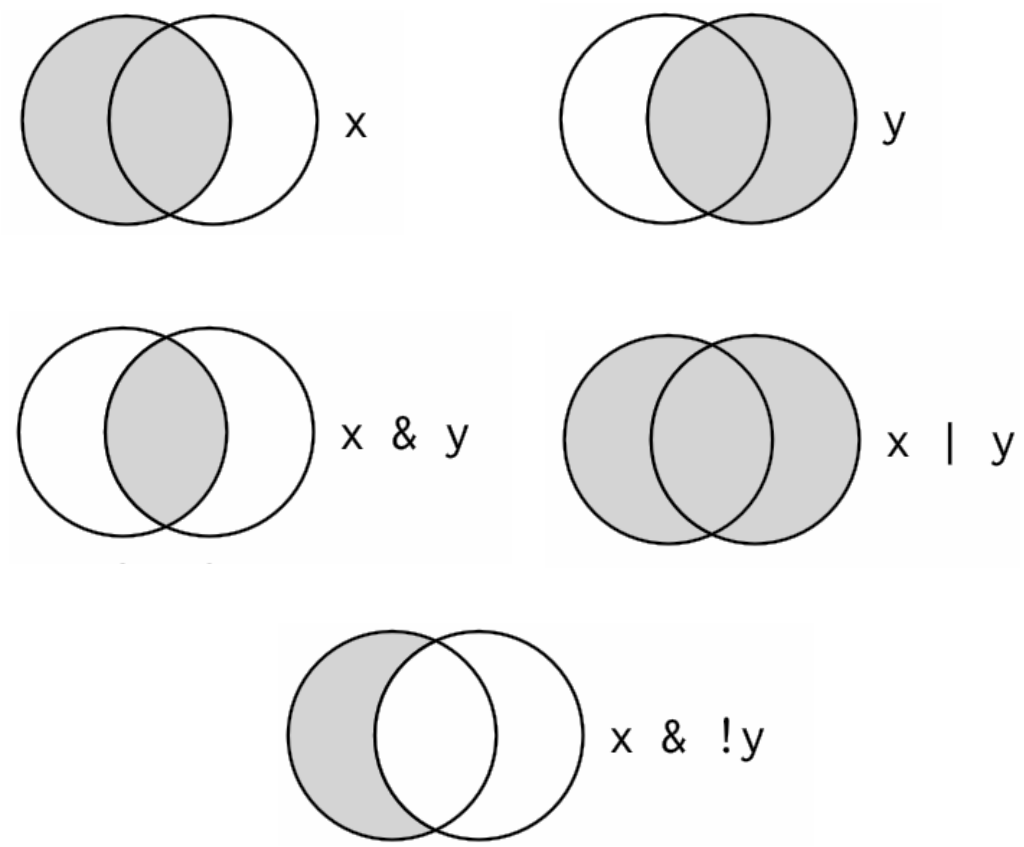


Manipulating data with **dplyr**

- Six main single-table “verbs”
  - `select()` columns/variables
  - `filter()` rows/observations
  - `arrange()` the order of rows
  - `mutate()` new columns/variables
  - `summarize()` the data
  - `group_by()` variable values
- Verbs take two arguments
  - `verb(data, what to do)`

Conditions and Operators

- Logical operators
  - `==` , `!=`
  - `>` , `!>` , `>=`
  - `<` , `!<` , `<=`
  - `%in%`



Adapted from *R for Data Science: r4ds.had.co.nz/transform.html*

The **pipe** and **magrittr**

- Why **magrittr**?
  - The 20th century painter, René Magritte
- %>% operator
  - Feeds output of one operation into another
  - Improves human-readable code

To **RStudio!**

What’s next?

Next Steps

- Review the workshop
- **dplyr**’s two-table verbs
  - Learn how to join two tables together
- Practice, practice, practice!
  - 2-Flights.R
  - Your own data

Resources

- R Cheatsheets: www.rstudio.com/resources/cheatsheets/
- R for Data Science: r4ds.had.co.nz (Print and Digital)
  - www.tidyverse.org (recently updated!)
  - www.rstudio.com/resources/videos/data-science-in-the-tidyverse/
- Questions? www.stackoverflow.com
- Need more data? archive.ics.uci.edu/ml/datasets.html

Thanks for coming!

- Questions? Comments? Coffee?
  - Email: rogol@virginia.edu
- Like this workshop? Want to learn more?
  - View upcoming sessions: cal.hsl.virginia.edu

References

- Pafka, Szlizard. “Dplyr and a very basic benchmark.” *DataScience.LA*, 2 Dec. 2014, datascience.la/dplyr-and-a-very-basic-benchmark/.
- Rickert, Joseph. “What is the tidyverse?” *R Views*, RStudio, 7 June 2017, rviews.rstudio.com/2017/06/08/what-is-the-tidyverse/.
- Wickham, Hadley. “dplyr 0.7.0.” *RStudio Blog*, RStudio, 13 June 2017, blog.rstudio.com/2017/06/13/dplyr-0-7-0/.
- Wickham, Hadley. “Introduction to dplyr.” useR! 2014. 30 June 2014, Los Angeles, CA., UCLA, www.dropbox.com/sh/i8qnluwmuieicxc/AAAgt9tIKoIm7WZKIyK25lh6a.
- Wickham, Hadley and Garrett Grolemund. *R for Data Science. O’Reilly, 2017.*
- Wickham, Hadley. “Tidy Data.” *Journal of Statistical Software*, vol. 59, no. 10, 12 Sept. 2014, www.jstatsoft.org/article/view/v059i10.
- Wickham, Hadley, et al. “Tidyverse.” *GitHub*, github.com/tidyverse/.