R Graphics Overview

Nick Ulle

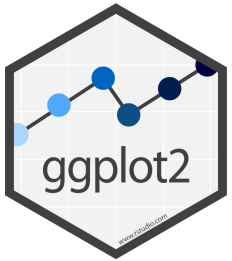
Graphics in R

Dominant systems:

1. **base**
    - R's built-in graphics functions, like `plot()` and `curve()`.
2. **lattice**
    - Similar interface to base, but more features, especially for making grouped plots.
3. **ggplot2**
    - "Grammar of graphics" interface, where graphics are assembled in layers.

Each is incompatible with the others.

The ggplot2 Package



We'll mainly use ggplot2.

Why?

- Excellent documentation, with examples: ggplot2.tidyverse.org
- More concise and featureful than base
- More popular than lattice
- Available for Python (plotnine) & Julia (Gadfly.jl)
- Part of the Tidyverse…

The Tidyverse

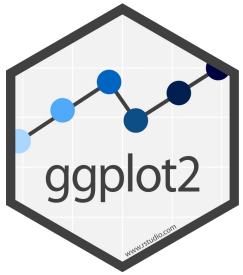
Nick Ulle

The Tidyverse

The **Tidyverse** is a collection of R packages for data analysis.

- Designed to work well together
- Made by many of the same people as RStudio
- Excellent documentation: www.tidyverse.org
  - Also many cheat sheets: rstudio.com/resources/cheatsheets
- Alternatives to R's built-in tools
  - A different dialect of R
  - Occasionally accused of **dividing** the R community

The Tidyverse Packages



Package	Tools For
readr, readxl	Reading files
tibble, dplyr, tidyr	Data frames
stringr	Strings
forcats	Factors
ggplot2	Graphics
purrr	Functional programming

These are not the only packages in the Tidyverse.

Images from tidyverse.org

Tidy Data

Most Tidyverse packages require tidy data sets.

In a **tidy** data set:

- Each observation has its own row.
- Each feature has its own column.
- Each value has its own cell.

The tidyr package has examples and tools to clean up untidy data.

Tibbles

Nick Ulle

The Grammar of Graphics

Nick Ulle

Saving Plots

Nick Ulle

Customizing Plots

Nick Ulle

Exploratory Data Analysis

Nick Ulle

Exploring Data

What does it mean to “explore” data?

- Get an overview of what’s included
- Look for errors (typos, extreme values, etc.)
- Look for patterns within features
- Look for relationships between features
- Check assumptions (for conclusions, models, etc.)

Use plots, summary statistics, and *with caution*, models.

Choosing a Plot

There isn’t a strict guideline for how to choose a plot.

This table has *suggestions:*

First Feature	**Second Feature**	**Suggested Plots**
categorical		bar, dot
categorical	categorical	bar, dot, mosaic
numerical		box, density, histogram
numerical	categorical	box, density
numerical	numerical	line, scatter, smooth scatter