Introduction to R Programming

Slide Set 3: Data Visualization

Maria Ptashkina

Barcelona GSE ITFD

September 2021

Table of Contents

Introduction

- There are functions in base `R` to visualize the data
- But we will learn the arguably best data visualization tool `ggplot2`
- `ggplot2` is part of `tidyverse` and contains functions that would build a graph *in layers*
- We will start with a simple graph and then add layers one at a time

ggplot

- In building a `ggplot` only the data and the mapping of the variables is required, the rest can appear in any order
- `ggplot` specifies
  - The data to be plotted
  - The mapping using the `aes` function (standing for aesthetics)
- In `ggplot2` functions are chained together using the + sign to build a final plot

Geoms

- Geoms are the geometric objects (points, lines, bars, etc.) that can be placed on a graph
- Parameters can be specified for `geom_` functions
- For `geom_point` the parameters are color, size, and alpha (transparency from 0 to 1)
- `geom_smooth` adds the best fit
- Parameters for `geom_smooth` are type of line (linear, quadratic, nonparametric), thickness, color, presence or absence of a confidence interval

Grouping, Scales, Facets, Labels and Themes

- Grouping
  - We map the variables in x and y axis
  - In addition then can be mapped in color, shape, size, transparency, and other visual characteristics of geometric objects
  - This would allow us to make plots for different groups
- Scales
  - Scales control how variables are mapped to the visual characteristics of the plot
  - Scale functions (which start with `scale_`) allow you to modify this mapping
- Facets
  - Facets reproduce a graph for each level a given variable (or combination of variables)
- Labels
  - `labs` function provides customized labels for the axes and legends
  - You can also add a custom title, subtitle, and caption
- Themes
  - Theme functions control background colors, fonts, grid-lines, legend placement, and other non-data related features of the graph

Placing data and mapping options

- If `mapping` is placed in the `ggplot` function, it will apply to the whole following functions (geoms)
- Alternatively, if you need to apply some mapping just to a particular geom, you can place it directly in that geom
- Usually you would place the mapping in the `ggplot` though
- Additional insight: graphs in `R` can be saved and manipulated as objects!

Table of Contents

Univariate Graphs

- Categorical
  - Bar plot (counts, percent)
  - Pie chart (not suggested)
  - Tree map (hardly ever)
- Quantitative
  - Histogram
  - Kernel density plot (estimated probability function, i.e. a smoothed histogram where the area under the curve is equal to one)

Bivariate Graphs

- Categorical vs. categorical
  - Stacked bar chart
  - Grouped bar chart
  - Segmented bar chart
- Quantitative vs. quantitative
  - Scatter plot
  - Line plot (usually if one of the variables is time)
- Categorical vs. quantitative
  - Bar chart on summary statistics (means, medians, etc.)
  - Grouped kernel density plots (pretty nice)
  - Box plots
  - Violin plots
  - Ridgeline plots
  - Mean / SE graphs
  - Cleveland dot charts (when you have large number of groups)

Time Series

- A time series is a set of quantitative values obtained at successive time points
- The intervals between time points (e.g., hours, days, weeks, months, or years) are usually equal
- Line plot is the most commonly used representation
- To plot several series you need to organize a dataset in a long format using `gather()`

Table of Contents

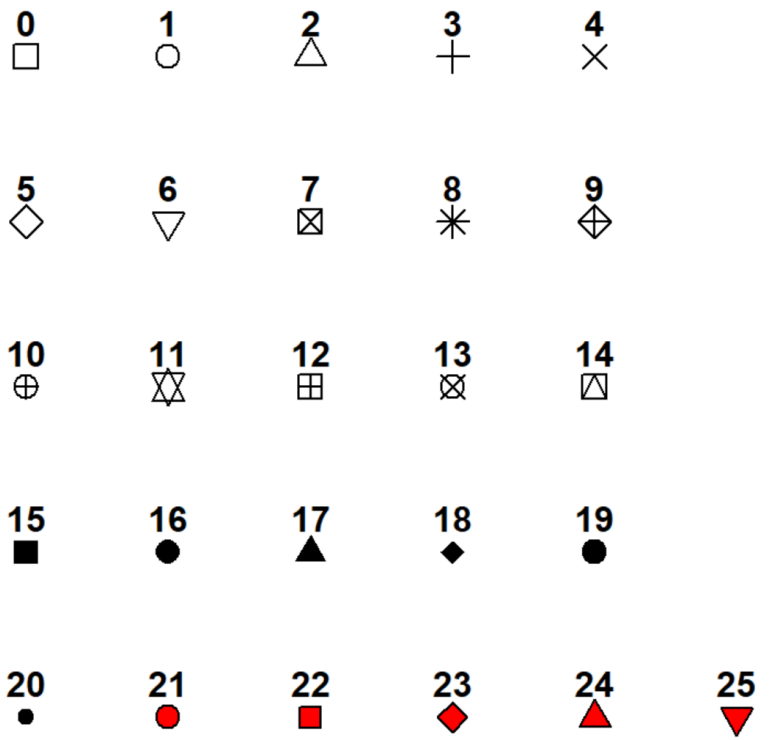
Customizing Graphs

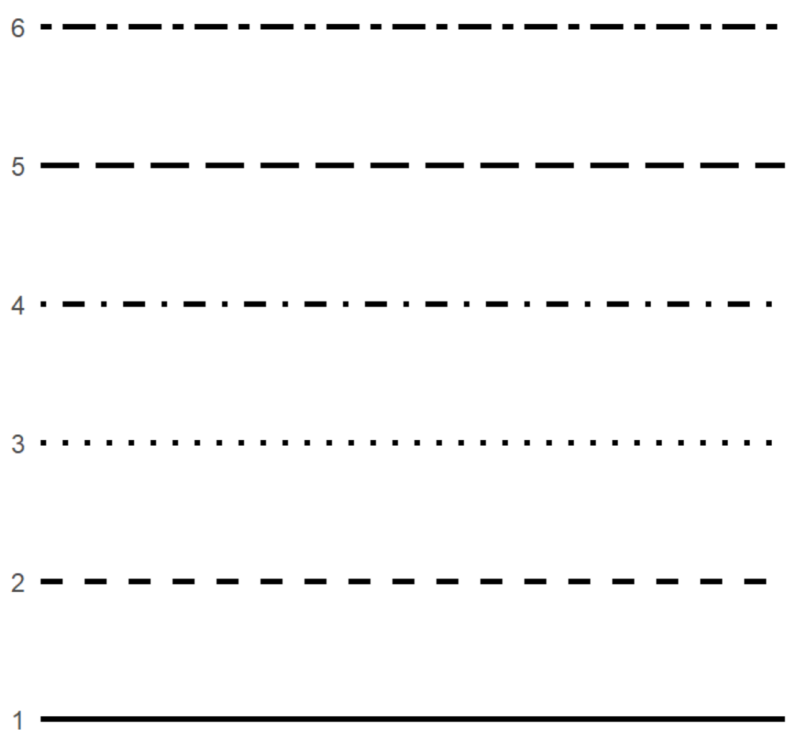
- Axes ▸ Help Sheet
  - Breaks and limits for continuous axes
  - Formatting numeric labels into percent, dollars, etc.
  - Limits and labels for categorical axes
  - Breaks and labels for dates (next slide)
- Colors
  - Specific colors for points, lines, bars, areas, and text
  - Specific colors mapped to the levels of a variable in the dataset
  - Color palettes ▸ ColorBrewer ▸ Vidris
- Points and lines
  - The default point is a filled circle, and the default line is a solid line
  - See other shapes in the next slide
- Legend
  - Position
  - Title
- Themes

Formatting Specifications for Dates

Symbol	Meaning	Example
%d	day as a number (0-31)	01-31
%a	abbreviated weekday	Mon
%A	unabbreviated weekday	Monday
%m	month (00-12)	00-12
%b	abbreviated month	Jan
%B	unabbreviated month	January
%y	2-digit year	07
%Y	4-digit year	2007

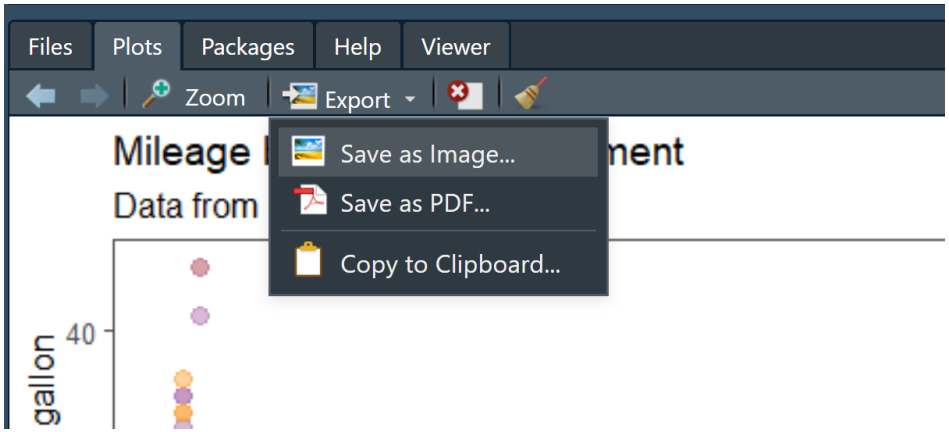
Points and Lines





Saving Graphs

- Via menu



- Via code (pdf, jpeg, tiff, png, svg)

Lots of Other Amazing Possibilities

- We cover only the basic graphs
- There are also graphs for statistical models ▸ Stats, 3D plots and other things ▸ Other graphs, maps ▸ Maps and even interactive graphs ▸ Maps

References and Resources

- Datavis with R ▸ Tutorial
- A Comprehensive Guide to Data Visualisation in R for Beginners ▸ Guide
- 7 Visualizations You Should Learn in R ▸ Tutorial
- R for Data Science ▸ Tutorial