Build your First Data Visualization with ggplot2

EXPLORING THE GGPLOT2 DATA VISUALIZATION PACKAGE

Martin Burger
STATS PROGRAMMING TUTOR

Basic Concepts

Course overview and recommended skill level

Importing the course dataset: lures.csv

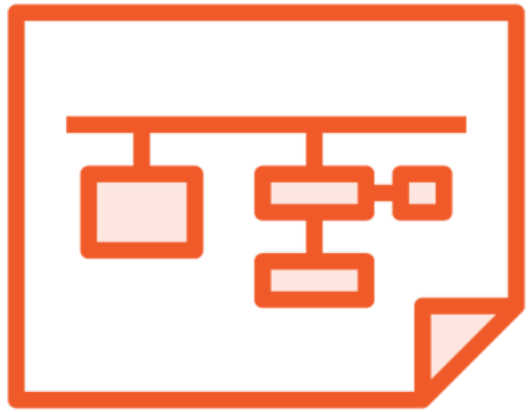
System setup for data visualizations with the library ggplot2

Characteristics of the ggplot2 syntax

Managing Expectations

Skill Level

Understanding of R syntax



Package management

RStudio orientation

Healthy System

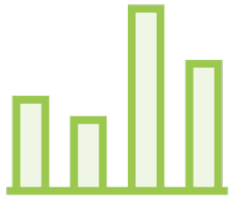
Make sure that your version of R and RStudio are up to date (not older than 8-10 months).

A very consistent plotting system

Customization is done for all chart types in a similar fashion

What You Will Learn

The ggplot2 environment



Standard plot types

Help and resources

Plot customization

General syntax and setup

Support elements

Recommended Courses on ggplot2

Annotating ggplot2 Visualizations in R

Formatting ggplot2 Visualization Elements in R

The Environment

Ggplot2 is an alternative to the R Base plotting system

- Enhanced results and a better user experience

Has developed to be the most popular charting library in R

- Translations are available in other languages

Get the library:

- install.packages("ggplot2")
- library(ggplot2)

What is the Tidyverse and how does it relate to ggplot2?

Tidyverse

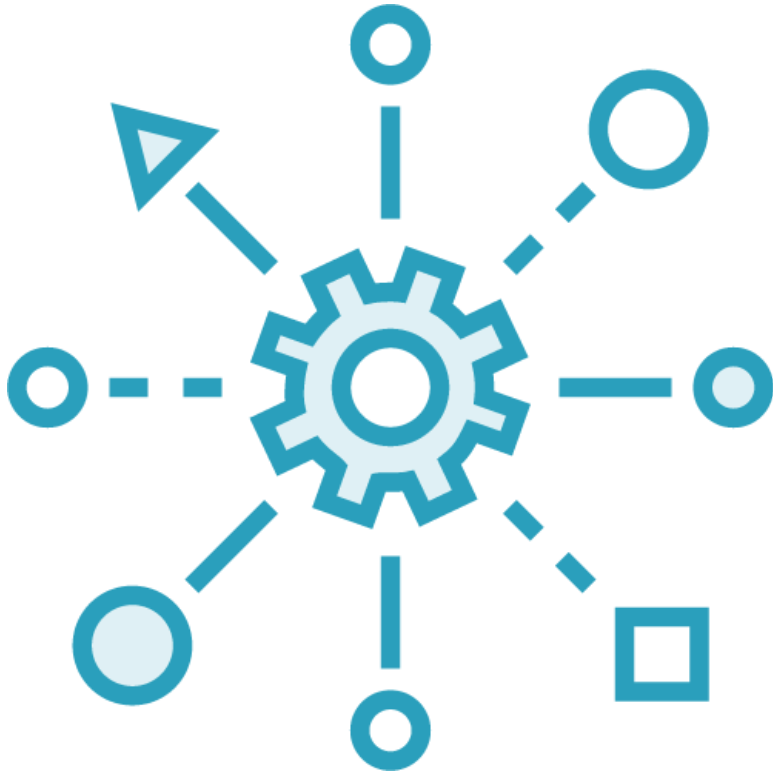
A curated collection of open source R libraries to clean, structure and visualize data. In this system ggplot2 takes the data visualization part.

Very popular data visualization system

Huge user community

Ongoing improvements

Extension libraries



Why Choose ggplot2

A well thought-out, consistent library

Quality data visualizations out of the box

Huge community of users and contributors

The Course Dataset

Course Dataset

Download the lures.CSV file to be able to follow along with the lectures.

Variable Classification

Factor: Shop ID, Sales Person ID, Item Nr

Double: Price, Revenue

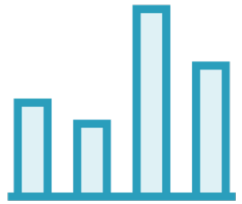
Integer: Quantity

Date: Date

Data Classes Factor and Character

Character: Text with semantic information

Factor: Grouping variable of pre-defined members



Available plot types depend on the data class

Variables with grouping functionality are better classified as factors

Many visualization types are built with factors, almost none with characters

- E.g. box plot, bar chart, pie chart

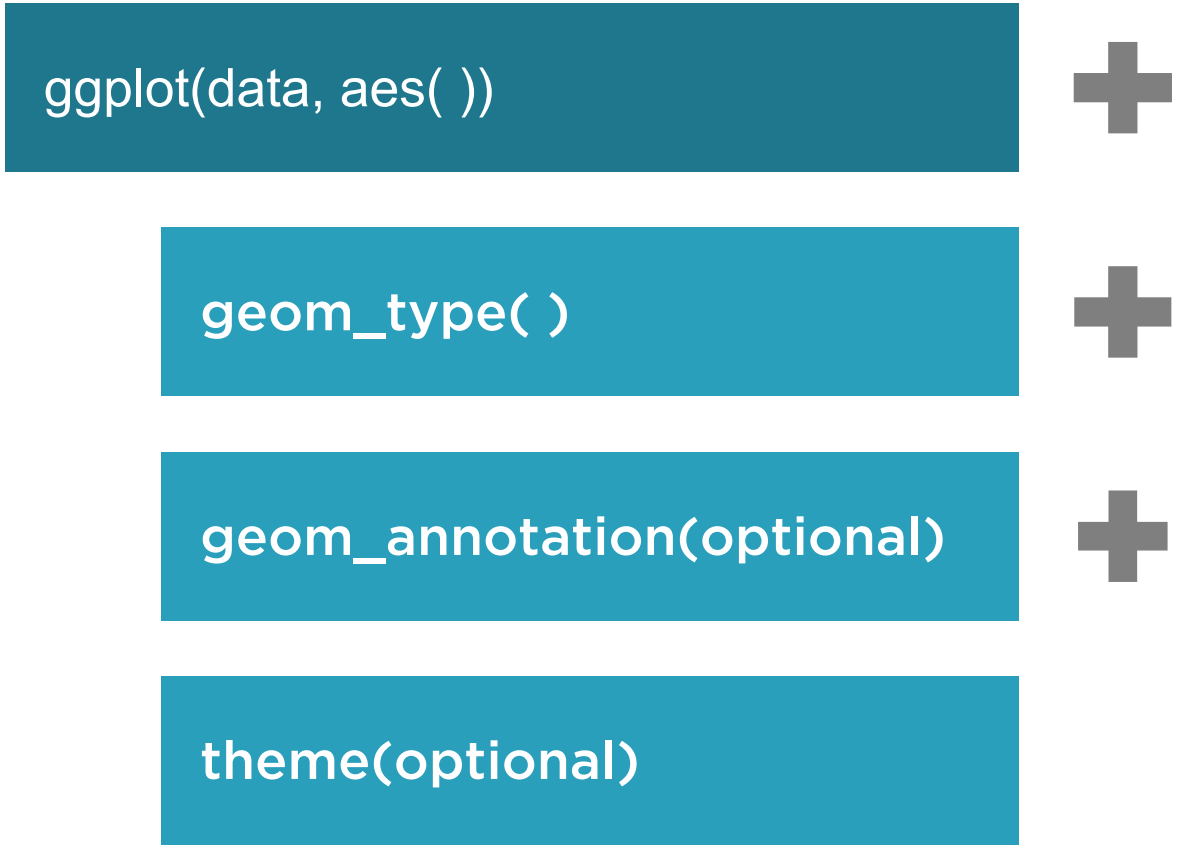
The ggplot2 Syntax

Code is organized in blocks of multiple lines

Indentation

Connecting lines with plus symbols (+)

The structure doesn't resemble R Base code



The Foundation of ggplot2 Visualizations

Function: ggplot()

- Accepts a data.frame object
- Aesthetics: X and Y variables

Alternative function for quick results: qplot()

- Coded similarly to R Base plots
- Less popular and rarely used

Defining the Plot Type

While ggplot() establishes the plot, the visualization type is set with geoms

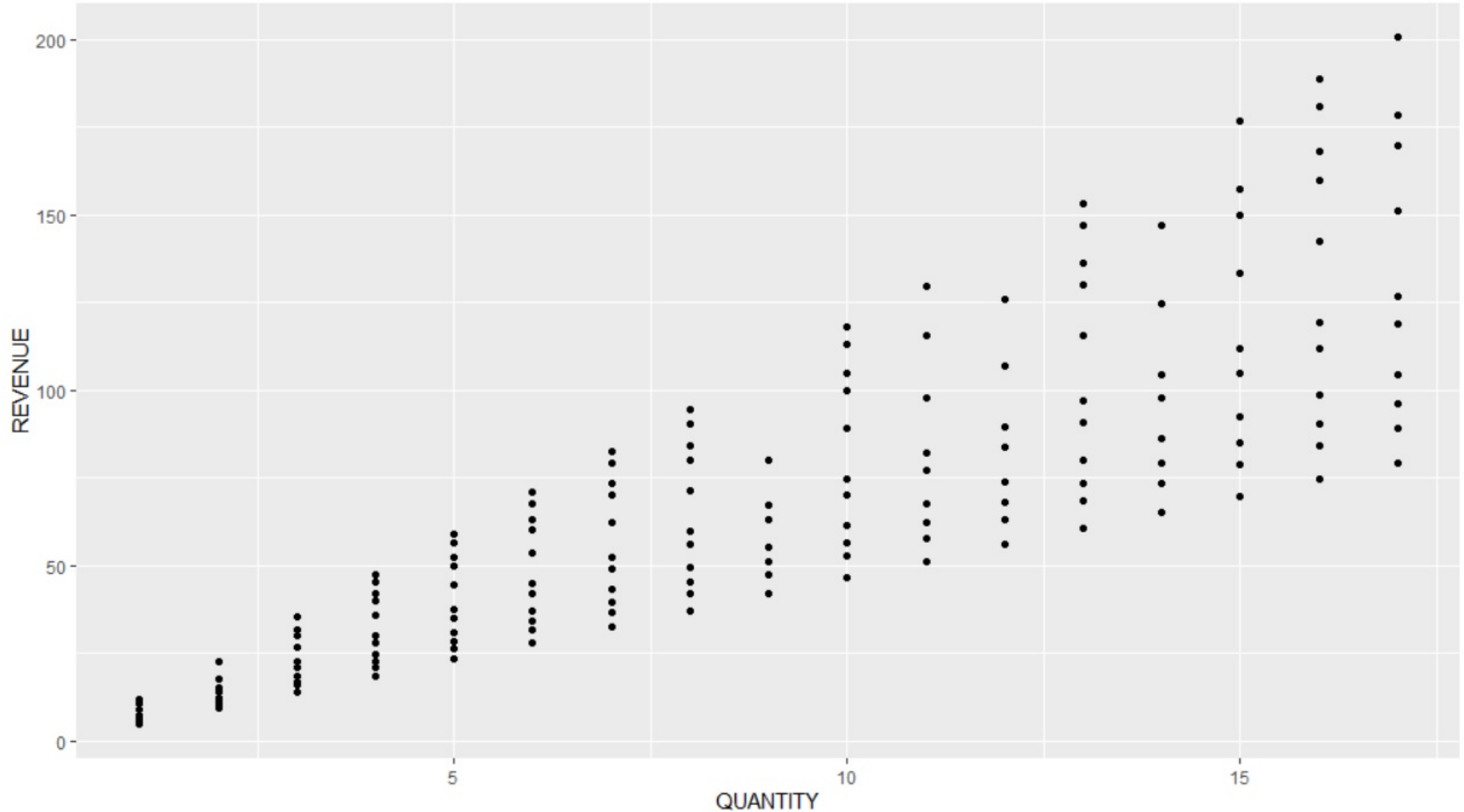
Dedicated geoms for standard plot types

- Scatterplot: geom_point()
- Bar chart: geom_bar()
- Line graph: geom_line()

Geoms can be combined, but avoid overloaded visualizations

```
ggplot(lures,
       aes(QUANTITY, REVENUE)) +

 geom_point()
```

- The data and the aesthetics are introduced via the ggplot() command
- The plus sign connects the two lines of code
- Indenting the lines below the ggplot() command is mandatory
- The plot type is set to scatterplot with geom_point()
- Additional aesthetics can be added at this point
- Formatting the appearance of the points is also done in the geom

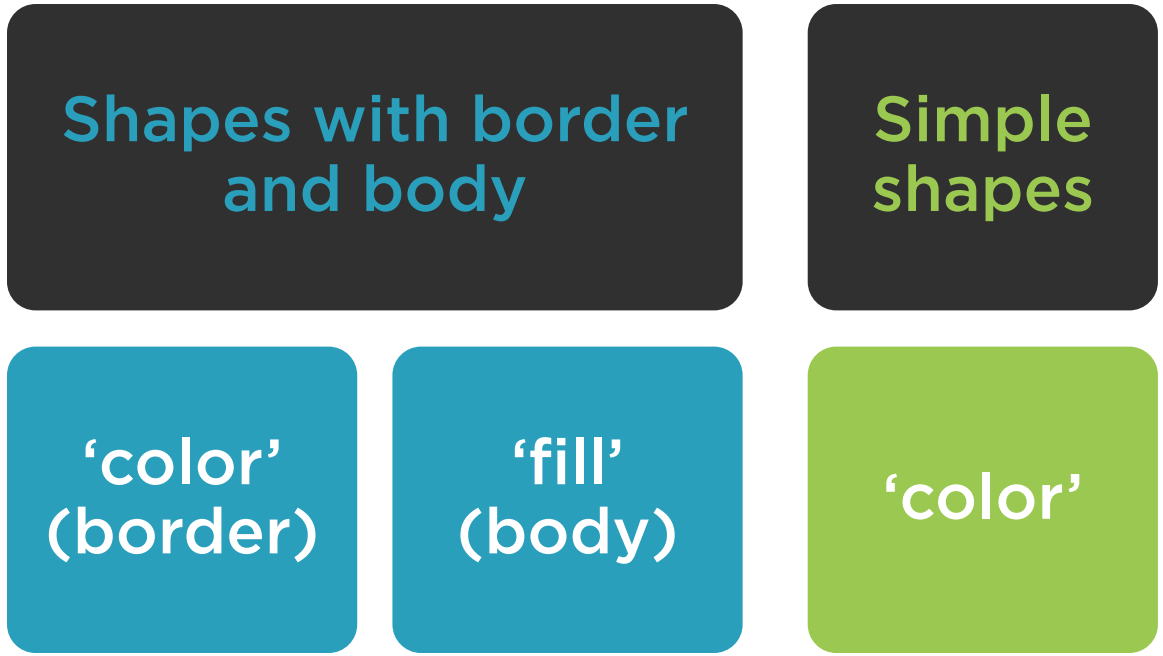

200
150
100
50
0
REVENUE
5
10
15
QUANTITY

Modifying the Plot Appearance within the Geom

Coloring Plot Elements

The appearance (size, color, shape) of aesthetics related markers is set in the geom.

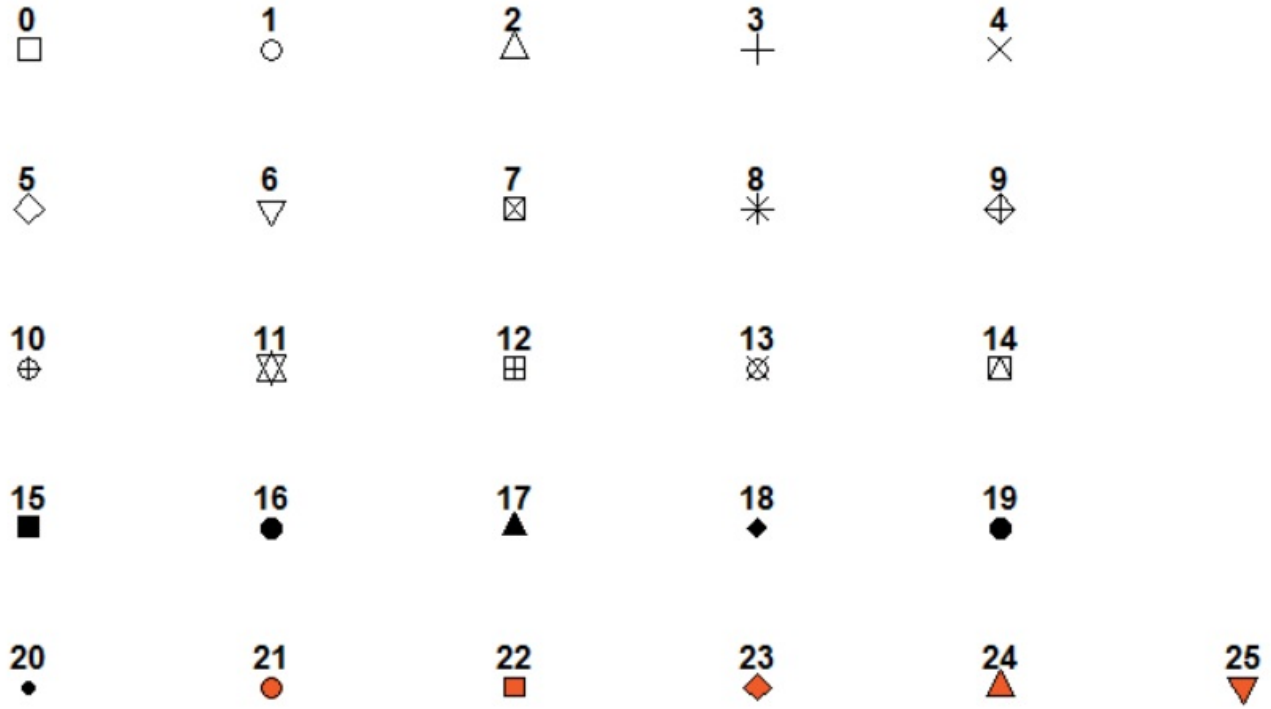
For some markers the body and border can be accessed separately



Coloring options depend on the shape

Shapes 21:25 are filled

Shapes 0:20 are solid



Summary: Exploring the Ggplot2 Data Visualization Package

Environment

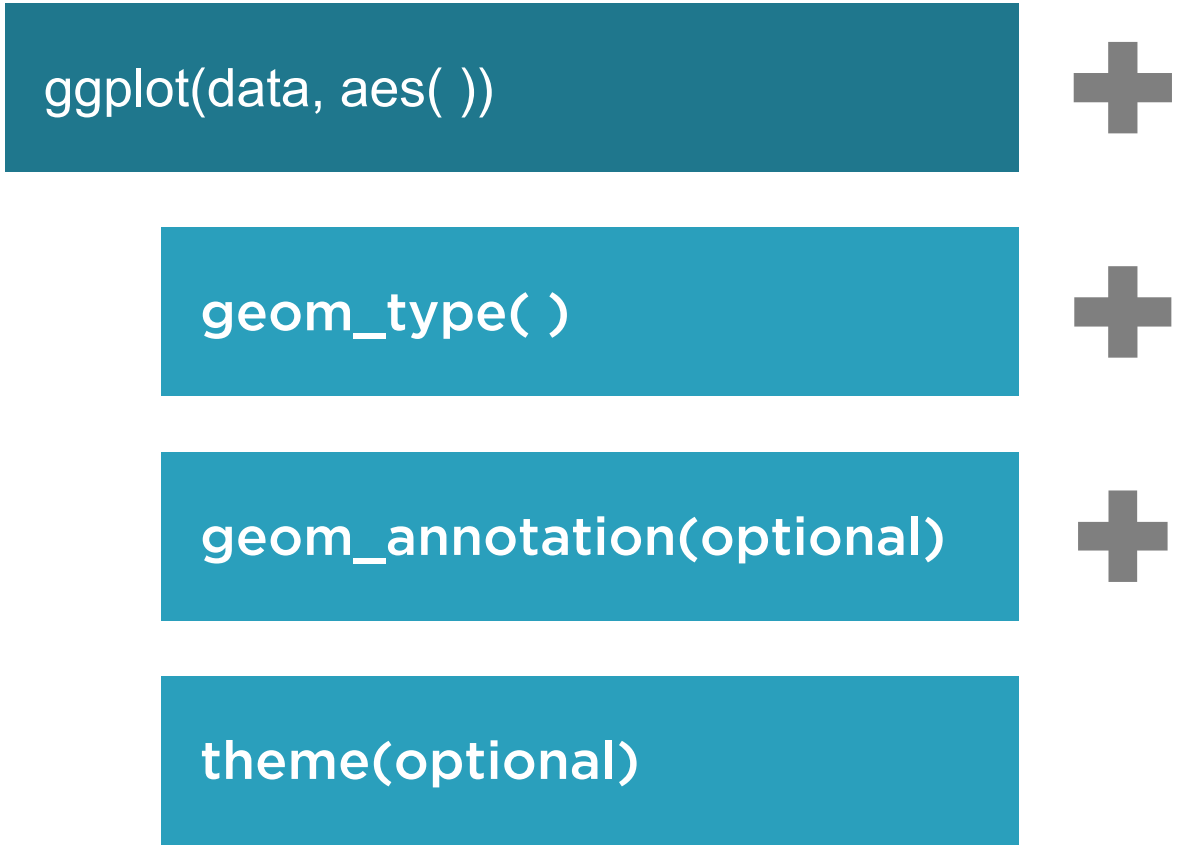
```
# Install gggplot2 once
intall.packages('ggplot2')

# Activate in each session
library(ggplot2)
```

Dataset and variables introduced in ggplot()

The geom() defines the plot type

Each layer builds (+) on one another

Multiple geoms are allowed



Up Next: Modifying a Ggplot

Further plot types with dedicated geoms

Modifying the appearance of the plot via the theme

Plot guides: Legend and the axes