Introduction to Data Science Using R

Lecture 5 - Introduction to Data visualization

Ben Rogers, Sept 20, 2022

Review

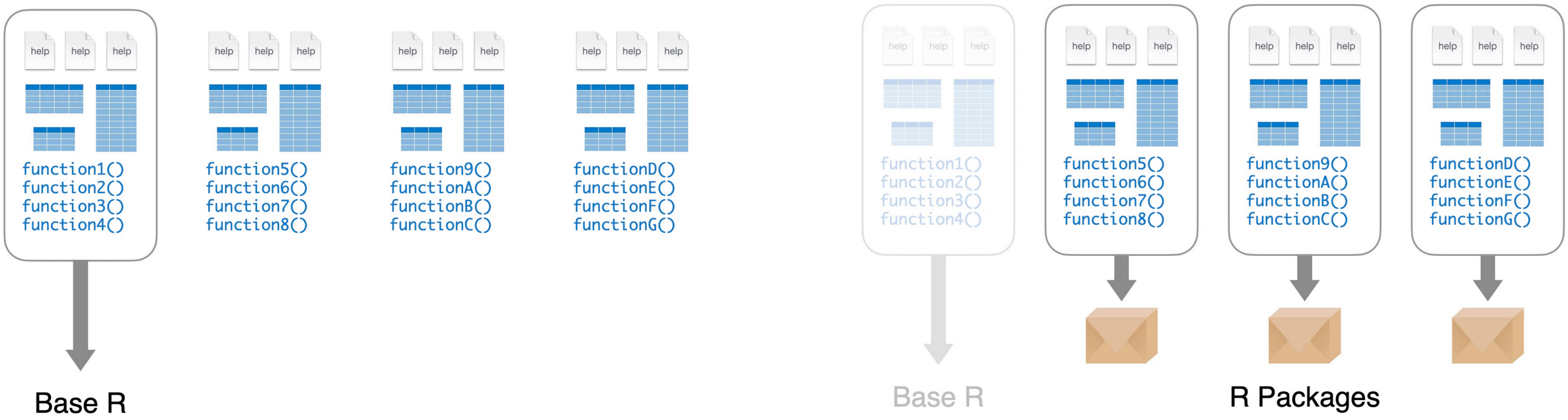
Last Class

- Introduction to R Markdown
  - Writing and formatting in markdown
  - Code chunks!
  - Knitting an R markdown document

Today

- Introduction to data visualization in R!

R Packages



- R packages extend the functionality of R by providing additional functions, data, and documentation.
- They are written by a worldwide community of R users and can be downloaded for free from the internet.

Tidyverse

Tidyverse Packages Blog Learn Help Contribute

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

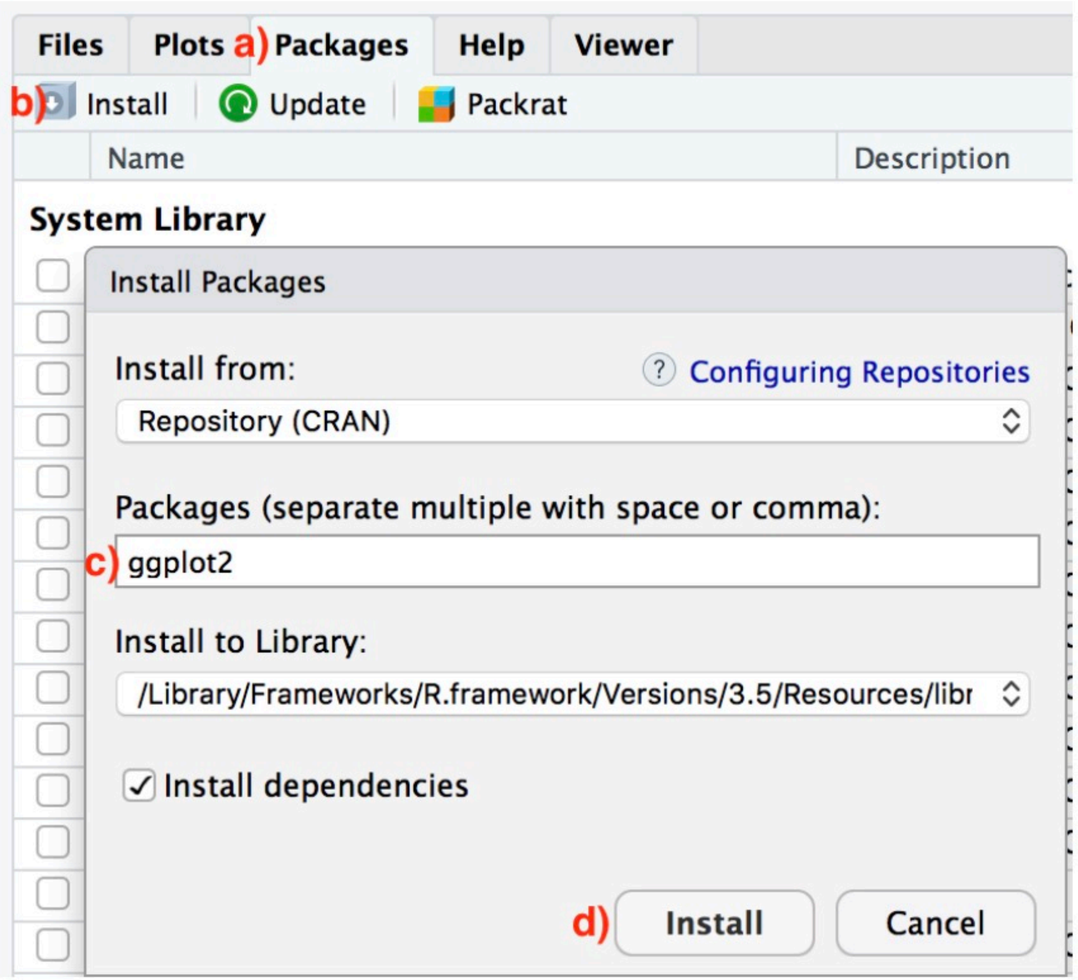
Install the complete tidyverse with:

```
install.packages("tidyverse")
```

ggplot2 package

- We will focus on ggplot2
- Allows for the user to concentrate on the visualizations instead of creating the underlying code.
- On top of this central philosophy, ggplot 2 has:
    - Increased flexibility over many plotting systems
    - An advanced theme system for professional/publication level graphics
    - Large developer base — many libraries extending its flexibility
    - Large user base — Great documentation and active mailing list

Package installation



- An alternative way is by typing

```
install.packages("ggplot2")
```

in the console pane of RStudio and pressing Return/Enter on your keyboard.

Load ggplot2

- To use an R package, you must:
    - Install the package so that the files for it are on your computer (this only needs to be done once per computer):

      ```
      install.packages("tidyverse")
      ```

    - Load the package so that R knows you will be using it in this session (this needs to be done every time you re-open R and want to use a package):

      ```
      library(tidyverse)
      ```

    - The ggplot2 package is contained within the tidyverse package.

Your turn

- Install the tidy verse and nycflights13 packages and load them.

mpg data

- You now have access to the data, help pages and functions in ggplot2.
- Let’s look at the mpg dataset, type mpg into the console and hit enter.
- To make the discussion easier we need to get familiar with some terms:
    - A **variable** is a quantity, quality, or property that you can measure
    - A **value** is at the state of the variable when you measure it. The value of a variable may change from measurement to measurement.
    - An **observation** is a set of measurements made under similar conditions. An observation may contain several values, each associated with a different variable. I will sometimes refer to an observation as a data point.

mpg data

```
> mpg
# A tibble: 234 × 11
   manufacturer model      displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4           2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4           2    2008     4 auto(av)   f        21    30 p     compact
 5 audi         a4           2.8  1999     6 auto(l5)   f        16    26 p     compact
 6 audi         a4           2.8  1999     6 manual(m5) f        18    26 p     compact
 7 audi         a4           3.1  2008     6 auto(av)   f        18    27 p     compact
 8 audi         a4 quattro   1.8  1999     4 manual(m5) 4        18    26 p     compact
 9 audi         a4 quattro   1.8  1999     4 auto(l5)   4        16    25 p     compact
10 audi         a4 quattro   2    2008     4 manual(m6) 4        20    28 p     compact
# … with 224 more rows
# ℹ Use `print(n = ...)` to see more rows
>
```

- A tibble is a specific kind of data frame in R. This specific data frame has:
    - 234 rows corresponding to different observations. Here, each observations is a car make, model and year.
    - 11 columns corresponding to different variables describing each observation.
    - To know the meaning of the variables, type ?mpg

Eploring the data

- view(mpg) — brings up RStudio's built-in data viewer
- glimpse(mpg) — gives the first few entries of each variable in a row after the variable name. In addition, the data type of the variable is given immediately after each variable's name.
    - int stands for integers
    - dbl stands for doubles, or real numbers.
    - chr stands for character vectors, or strings
    - lgl stands for logical, values can only be TRUE or FALSE
    - fctr stands for factors, or categorical variables
    - dttm stands for date-times
    - date stands for dates

Exploring the data

- kable() — you need to load knit package before using kable(). kable() helps you draw a table using Rmarkdown.
- $ operator — the $ operator allows us to extract and then explore a single variable within a data frame

Your turn

Explore flights data in nycflights13 package

1. How many observations and variables are in the dataset?
2. What do the variables represent, what is the meaning of their values?
3. What does each observation represent?

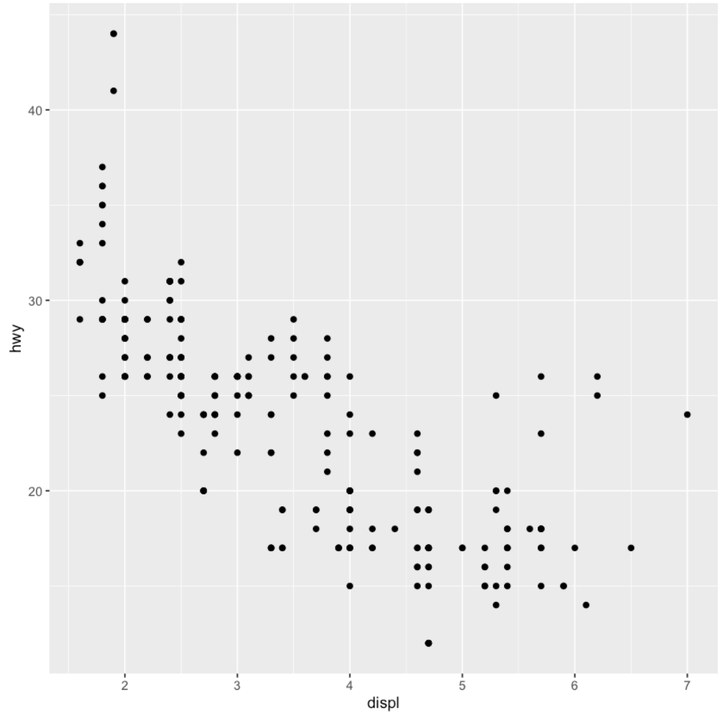
Data visualization using ggplot2

- Graphics/plots/charts provide a nice way to explore the patterns in data, such as the presence of outliers, distributions of individual variables, and relationships between groups of variables.
- Graphics are designed to emphasize the findings and insights you want your audience to understand.
    - This requires a balancing act between highlighting as many interesting findings as possible and including too much information as to overwhelm your audience.

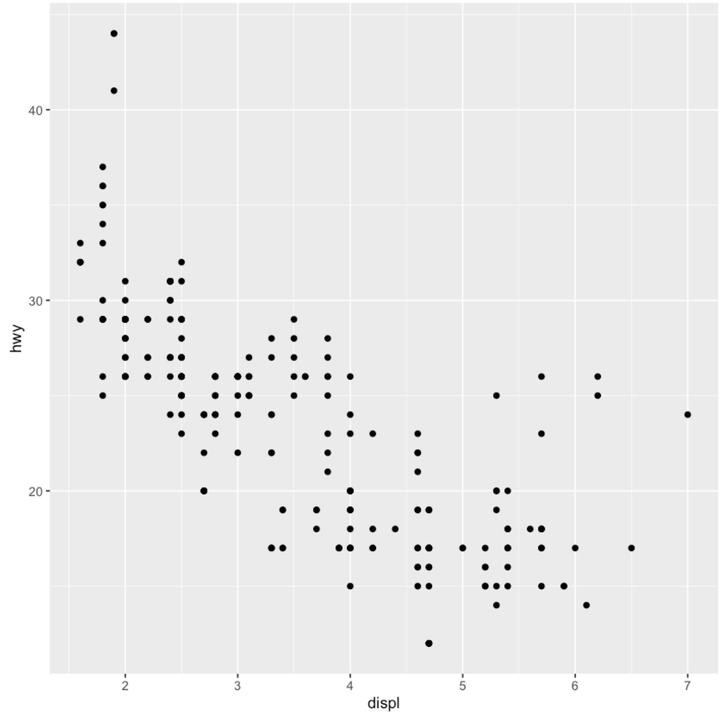
A simple plot

- Question: what relationship do you expect to see between engine size (displ) and gas mileage (hwy)?
- Let's plot two variables in the mpg dataset:

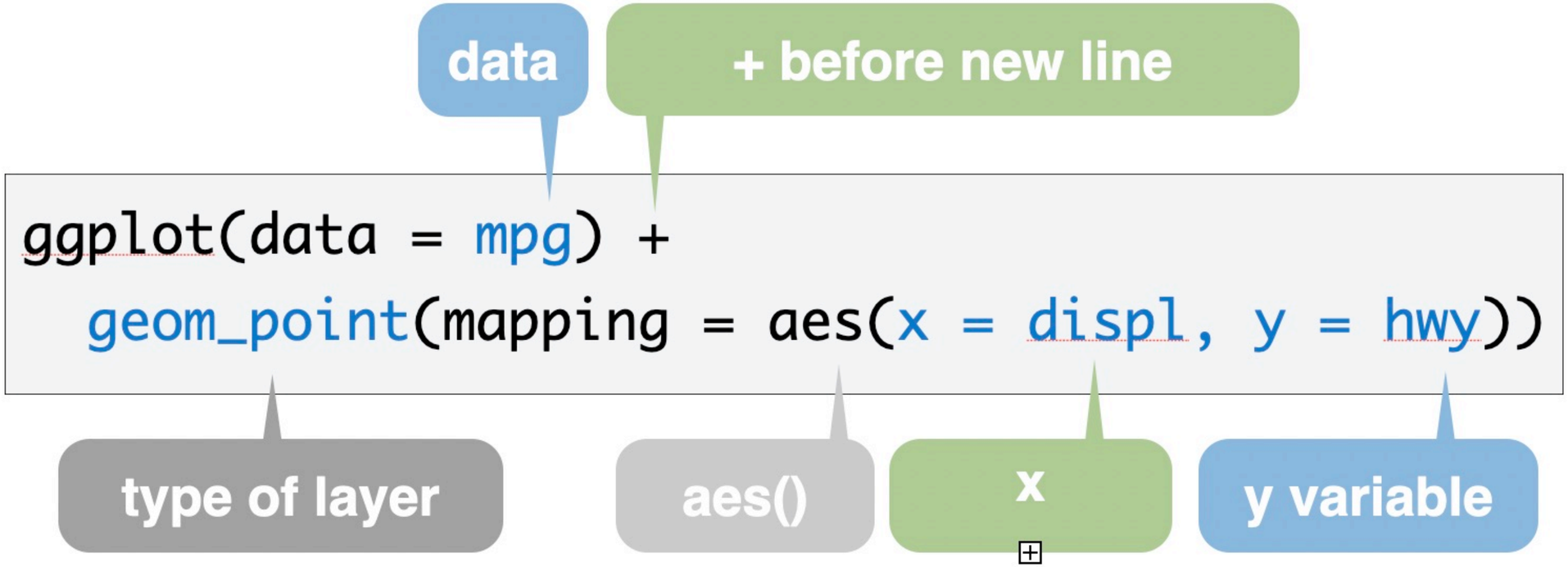
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```



```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```



- Three essential components:
  - data: the dataset containing the variables of interest.
  - geom: the geometric object in question. This refers to the type of object we can observe in a plot. For example: points, lines, and bars.
  - aes: aesthetic attributes of the geometric object. For example x/y position, color, shape, and size. Aesthetic attributes are mapped to variables in the dataset.


data
+ before new line
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy))
type of layer
aes()
x
y variable

Your turn

- Make a scatterplot of hwy vs cyl
- What happens if you use a constant for x or y (e.g. x=1)?