Getting started with data in R

*R, RStudio, and your first **data frame***

install.packages()
library()
glimpse()
View()

the nycflights23 tables

`flights`

Everything in this chapter is a step you will repeat all semester.

R vs RStudio

Two different pieces of software. You install both, but you only ever open one of them.

The Engine & The Dashboard

- **R** is the language. It does the computing.
- **RStudio** is the environment you write R in.
- You open **RStudio**; it talks to R for you.
- The **console** runs code now. An **R Markdown** file saves it for later.

install.packages()

Download a **package** from CRAN onto your computer. You do this **once** per package, in the console, never in an R Markdown file.

Install Once

A package is a bundle of functions and data someone else already wrote.

install.packages()

Example

download it (console only, once)

```
install.packages("ggplot2")

library(ggplot2)
```

load it (every session)

library()

Load an installed package so its functions are available. You do this **every time** you start a new R session.

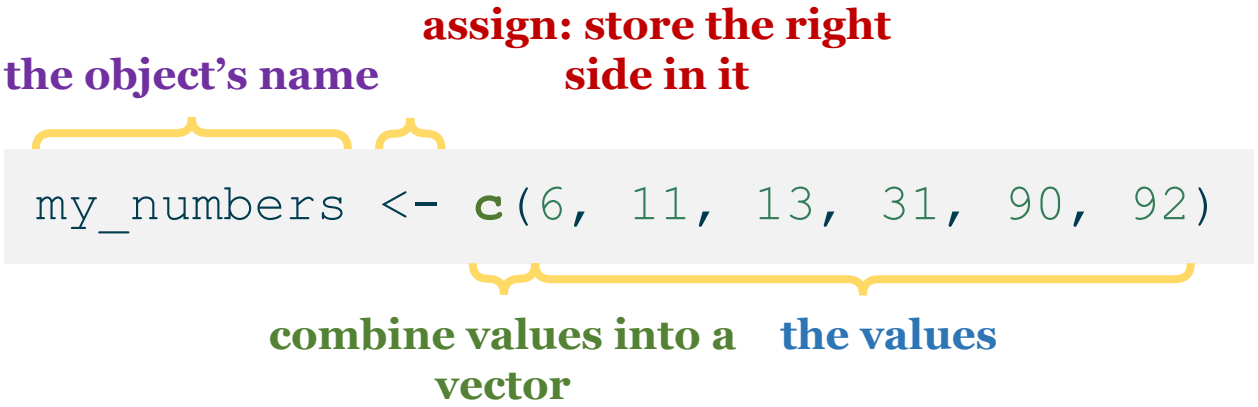
Load Every Session



<-

Save a value in an **object** so you can use it again. Type the object's name on its own to see what it holds.

Example



Data frames

A rectangle of data: every column is a **variable**, every row is an **observation**.

Rows & Columns

- A **variable** is something measured on each case: a column.
- An **observation** is one case: a row.
- **Categorical** variables record a group or label.
- **Quantitative** variables record a number you could average.

glimpse()

Print every variable in a data frame down the page, with its type and its first few values.

```
library(dplyr)
library(nycflights23)
glimpse(flights)
```

the verb (function) **the data frame**

Example

```
Rows: 435,352
Columns: 19
$ year           <int> 2023, 2023, 2023, 2023, 2023, 2023, 2023, 202…
$ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ dep_time       <int> 1, 18, 31, 33, 36, 503, 520, 524, 537, 547, 5…
$ sched_dep_time <int> 2038, 2300, 2344, 2140, 2048, 500, 510, 530, …
$ dep_delay      <dbl> 203, 78, 47, 173, 228, 3, 10, -6, 17, 2, -10,…
# ... 13 more lines, not shown
```

Three ways to look

Each one tells you something different about a data frame.

Explore Your Data

- **glimpse()** shows every variable, its type, and a few values.
- **View()** opens the spreadsheet viewer. Console only, never in a chunk.
- Typing the name prints the first ten rows.
- **?flights** opens the help page describing every variable.

The nycflights23 tables

The package holds five related tables. **flights** is the one this chapter works with.

Five Tables, One Package

- **flights** is one row per domestic flight leaving New York City in 2023.
- **airlines** matches a two-letter carrier code to an airline name.
- **planes** describes each physical aircraft that flew.
- **weather** holds hourly conditions at the three NYC airports.
- **airports** names and locates every destination.

Before lecture

Have these done and you will be ready to work along in class.

Checklist

- Sign in to RStudio and find the console, editor, and environment panes.
- Install **dplyr**, **nycflights23** and **knitr**, then load them without errors.
- Run **glimpse(flights)** and read what the variable types mean.
- Knit an R Markdown file to HTML at least once.