Ann & Robert H. Lurie Children's Hospital of Chicago®

INTRODUCTION TO R PROGRAMMING

Day 1: R Basics

July 26, 2023

Ann & Robert H. Lurie Children's Hospital of Chicago

Who is this workshop for?

- Those with little or no experience with R
- Those who would like an R refresher
- Those with a need or interest for using R in their work

Who are your instructors?

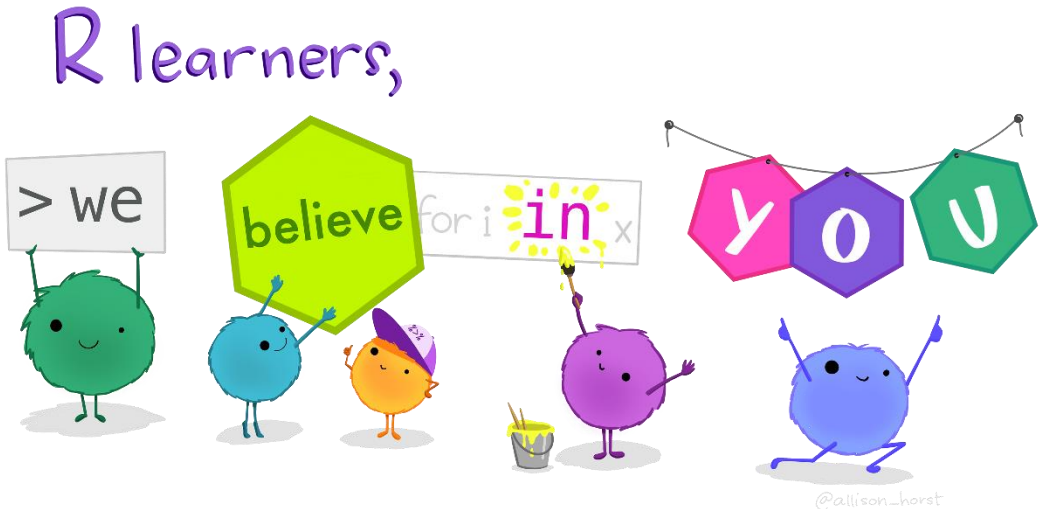
Joy Nyaanga, PhD
Senior Bioinformatician
Day 1 & 3

Stella Karuri, PhD
Statistician
Day 2

Ann & Robert H. Lurie Children's Hospital of Chicago®

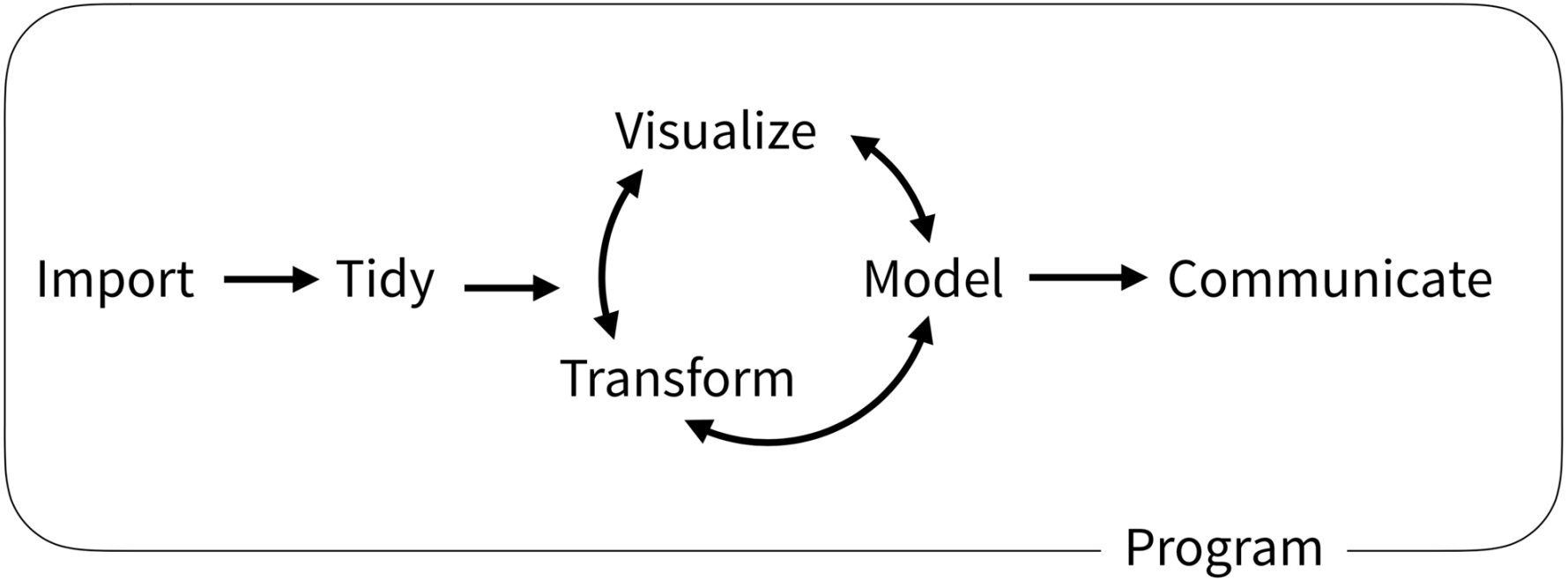
What will you learn?

- Anyone can learn R programming!
- R provides substantial flexibility and reproducibility over excel
- Data analysis can be streamlined in R



Orientation to the workshop

The data analysis pipeline



Workshop breakdown

Session 1: Lecture	11:00 a.m. – noon
Break	
Session 2: Interactive	2:00 p.m. – 4:00 p.m.

Course website: https://stanley-manne-childrens-research.github.io/introR/

Code repository: https://github.com/Stanley-Manne-Childrens-Research/introR

Day 1: The Basics of R

Learning objectives

- Understand basic coding principles
- Use R as a calculator
- Be able to load in data
- Basic operations on data
- Be able to make a plot
- Know how to get help

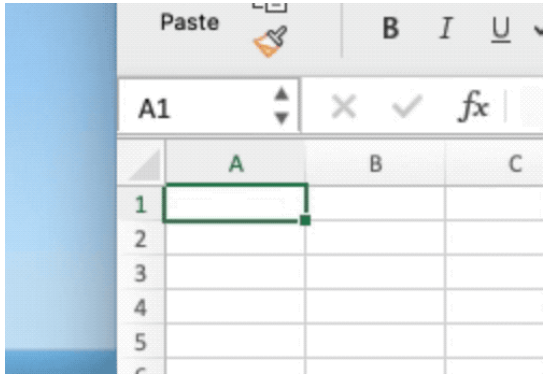
Why use something other than Excel?

- Excel can change your data and rename genes

Scientists rename human genes to stop Microsoft Excel from misreading them as dates / Sometimes it's easier to rewrite genetics than update Excel

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Aug 6, 2020, 7:44 AM CDT | 0 Comments / 0 New



Human Gene Nomenclature Committee

Symbols that affect data handling and retrieval. For example, all symbols that autoconverted to dates in Microsoft Excel have been changed (for example, SEPT1 is now SEPTIN1; MARCH1 is now MARCHF1); tRNA synthetase symbols that were also common words have been changed (for example, WARS is now WARS1; CARS is now CARS1).

Why use something other than Excel?

- Excel can change your data and rename genes
- Cut and paste errors are easy to do and can have bad consequences

Excel snafu costs firm $24m

Some cleric, some error

Drew Cullen

A simple spreadsheet error cost a firm a whopping US$24m.

The mistake led to TransAlta, a big Canadian power generator, buying more US power transmission hedging contracts in May at higher prices than it should have.

"… a cut-and-paste error in an Excel spreadsheet that we did not detect when we did our final sorting and ranking bids prior to submission" – chief executive Steve Snyder

https://www.theregister.com/2003/06/19/excel_snafu_costs_firm_24m/

Why use something other than Excel?

- Excel can change your data and r
- Cut and paste errors are easy to
- Making the wrong cell selection c
- Dates in Excel are an unmitigatec
- Excel has a relatively low (for some researchers) ceiling on number of available rows

Analysis

Covid: how Excel may have caused loss of 16,000 test results in England

Alex Hern

UK technology editor

Public Health England data error blamed on limitations of Microsoft spreadsheet

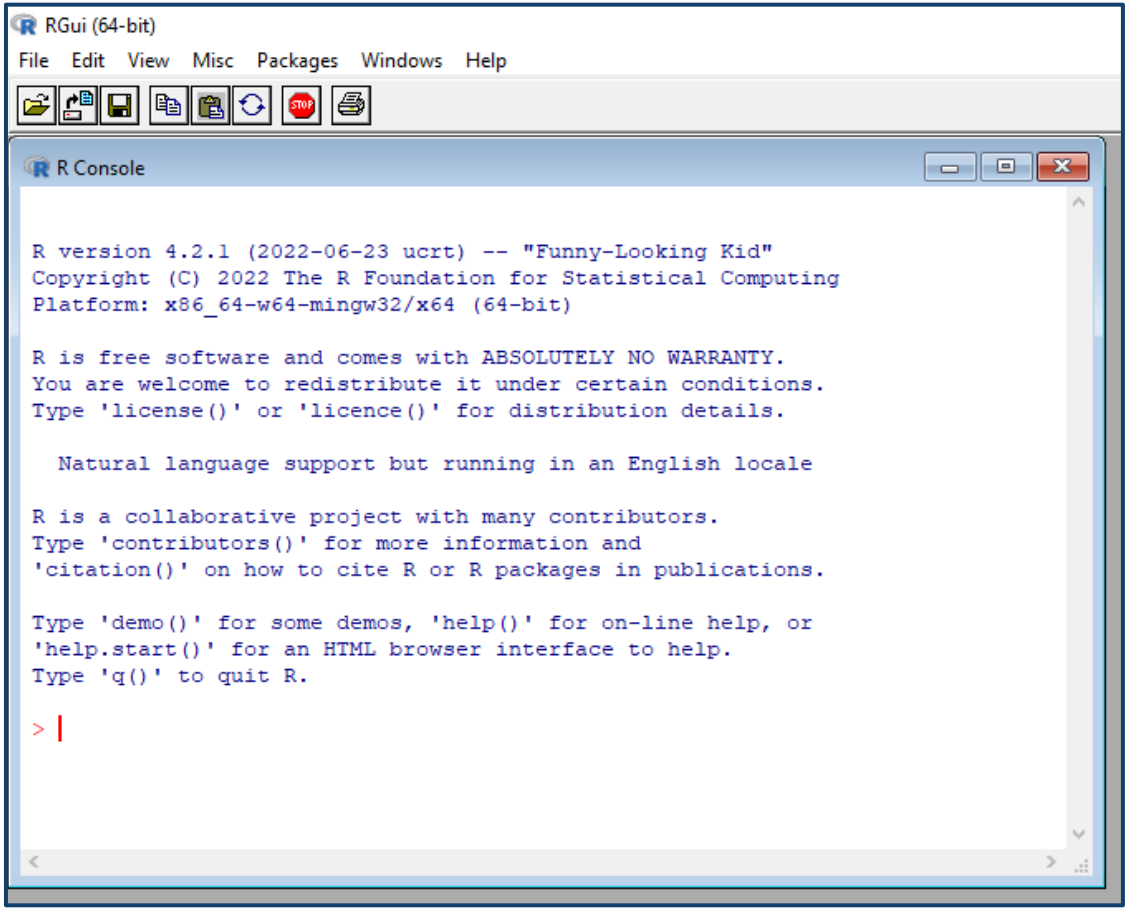
- Coronavirus - latest updates
- See all our coronavirus coverage

Point-and-click is not reproducible

What is R?

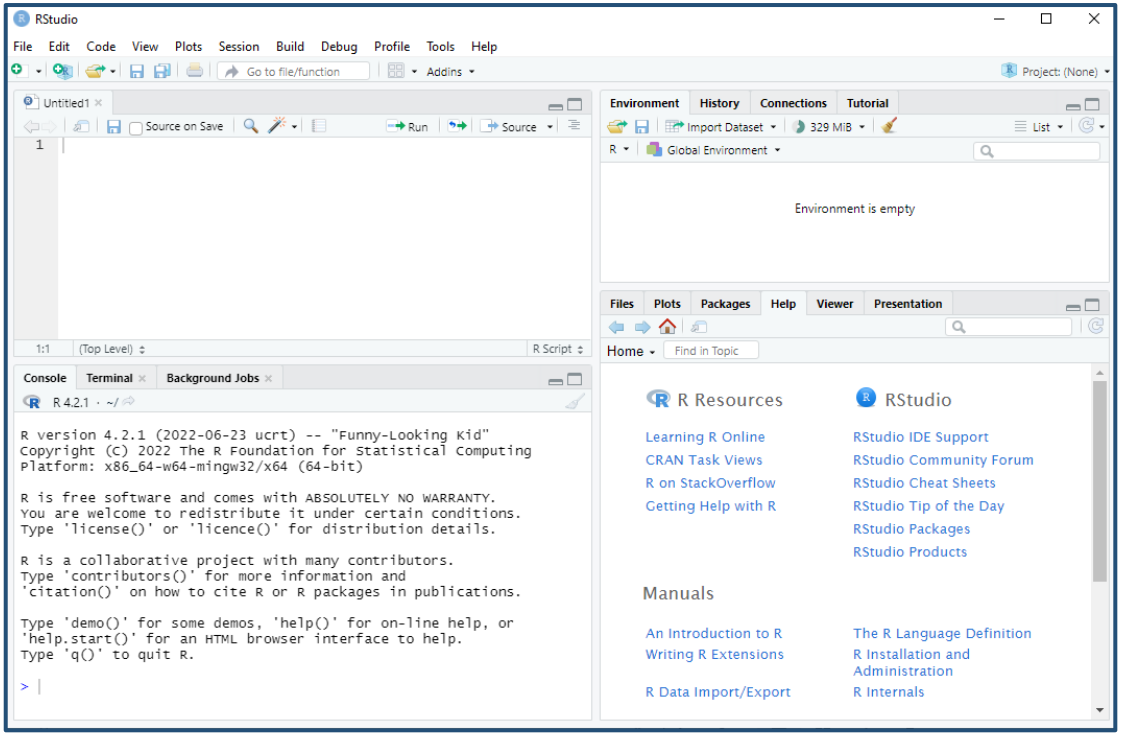
- A programming language
  - A successor of the S language
  - Created in 1991 by 2 statisticians to teach introductory statistics
  - In 1995 the source code for the R system was made accessible to the public
  - In 2000 R version 1.0.0 was officially released!
- Focus on statistical modeling and data analysis
- Interfaces with other languages (i.e. python, bash, etc.)

A programming language for data analysis



Early developers ran R via the command line console

What is RStudio?

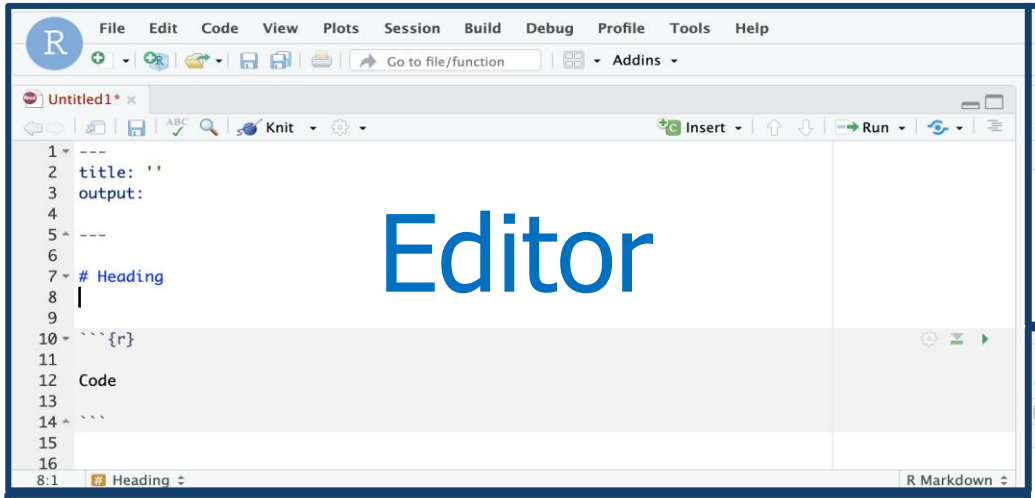


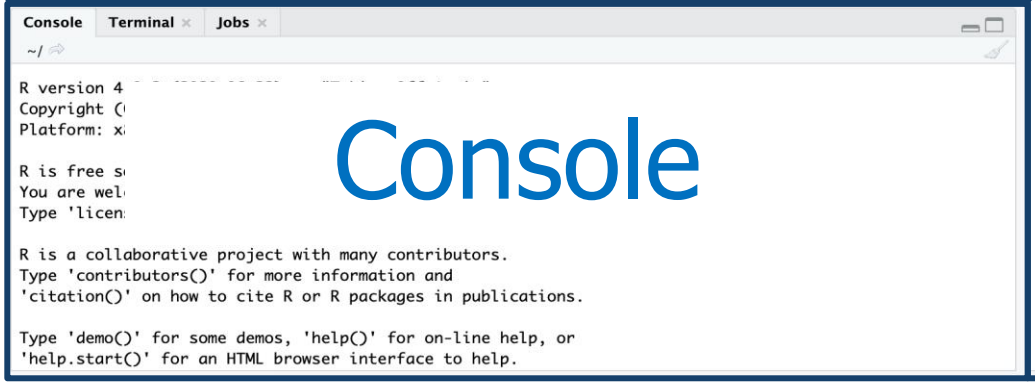
- An integrated development environment (IDE)
- Allows you to write, save, and open **R** code

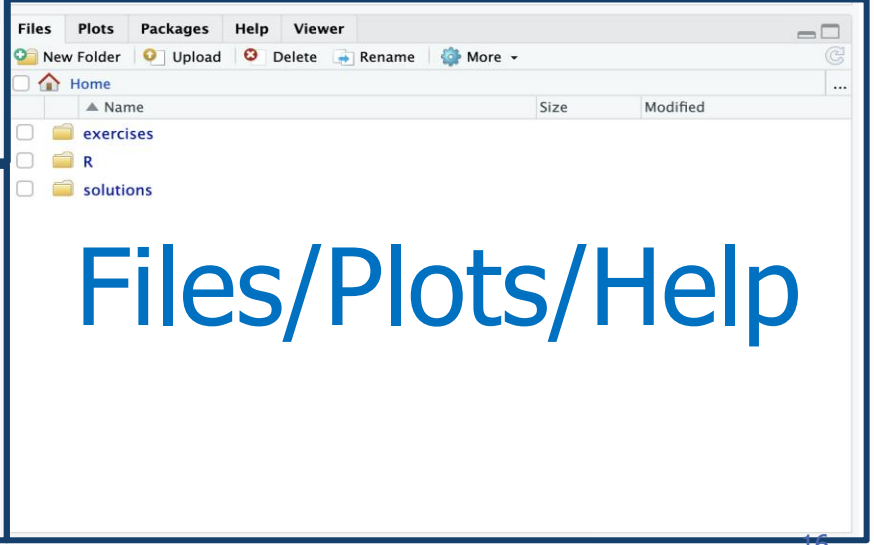
*"RStudio gives you a way to talk to your computer. **R** gives you a language to speak in"*

RStudio anatomy

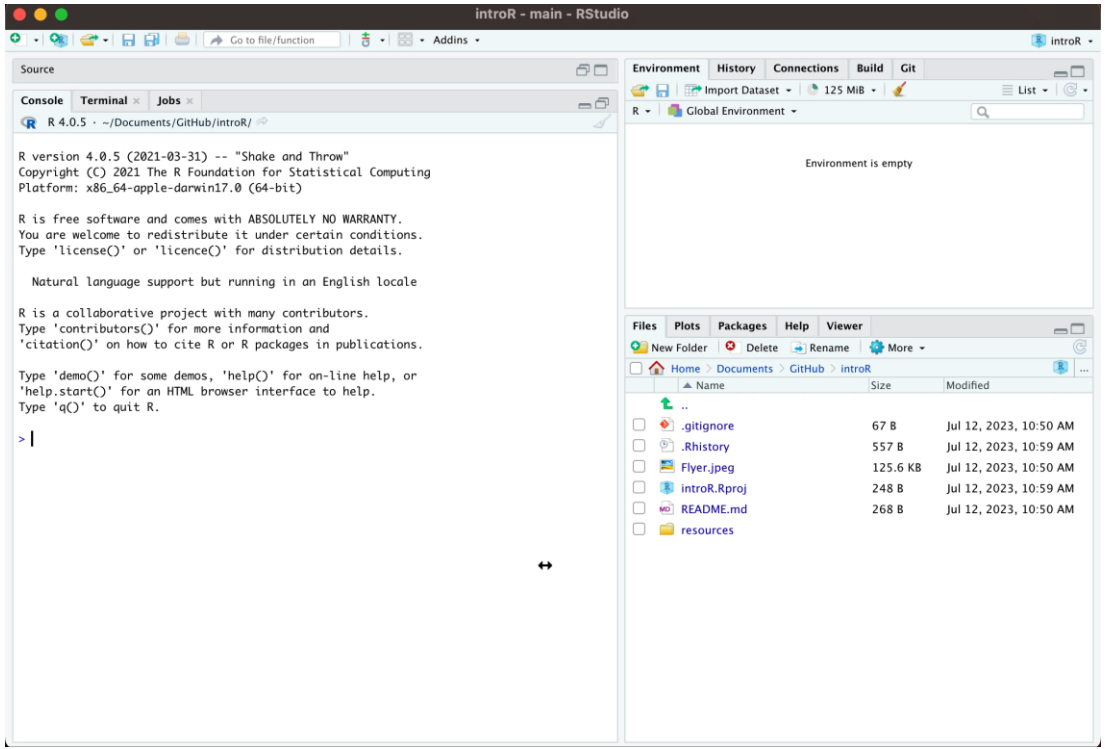
Ann & Robert H. Lurie Children's Hospital of Chicago®







Coding in the console

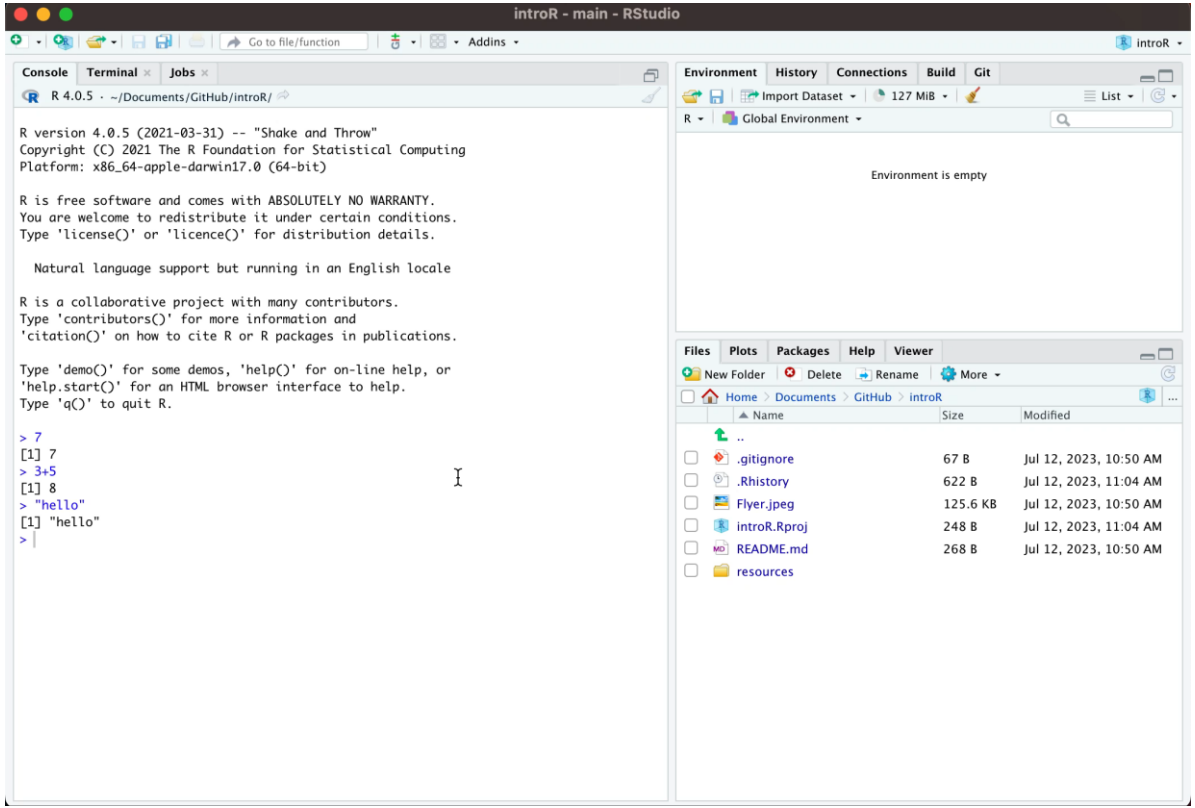


- Type code in the console
- Press **return** to execute code
- Output shown below

Coding in the console is not advisable for most situations!

- Only recommended for short pieces of code that you don't need to save

Coding in a script



Scripts allow you to…

- Execute code in blocks
- Save your work
- Easily share your work

Coding basics

Using R as a calculator

```
> 10^2

[1] 100
```

```
> 6/9

[1] 0.6666667
```

```
> 9-43

[1] -34
```

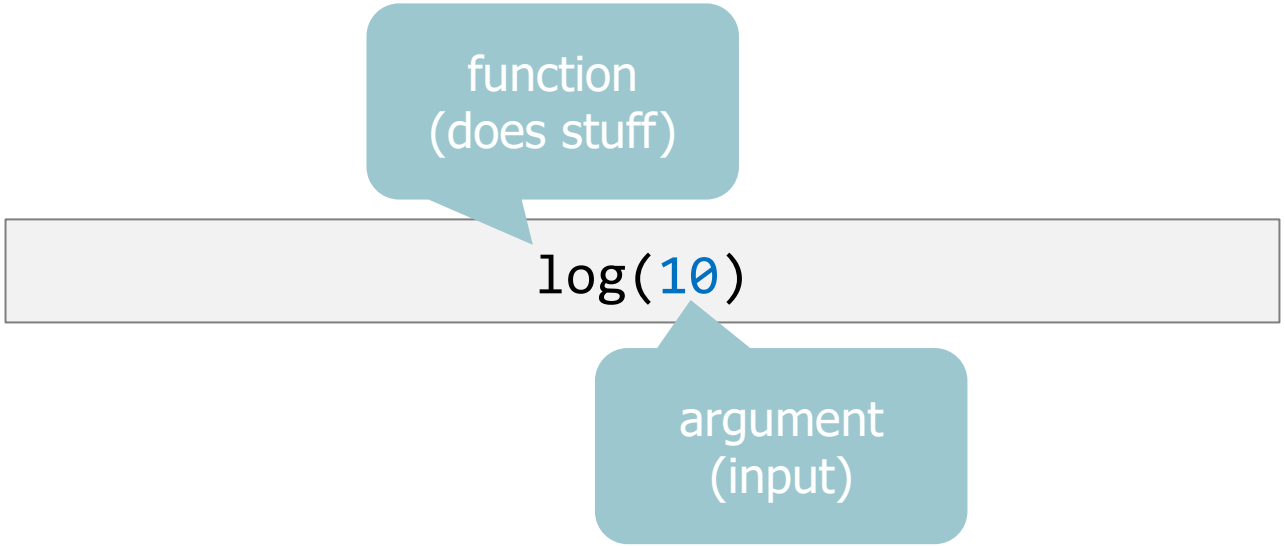
R follows the rules for order of operations and ignores spaces between numbers

```
> log(10)

[1] 2.302585
```

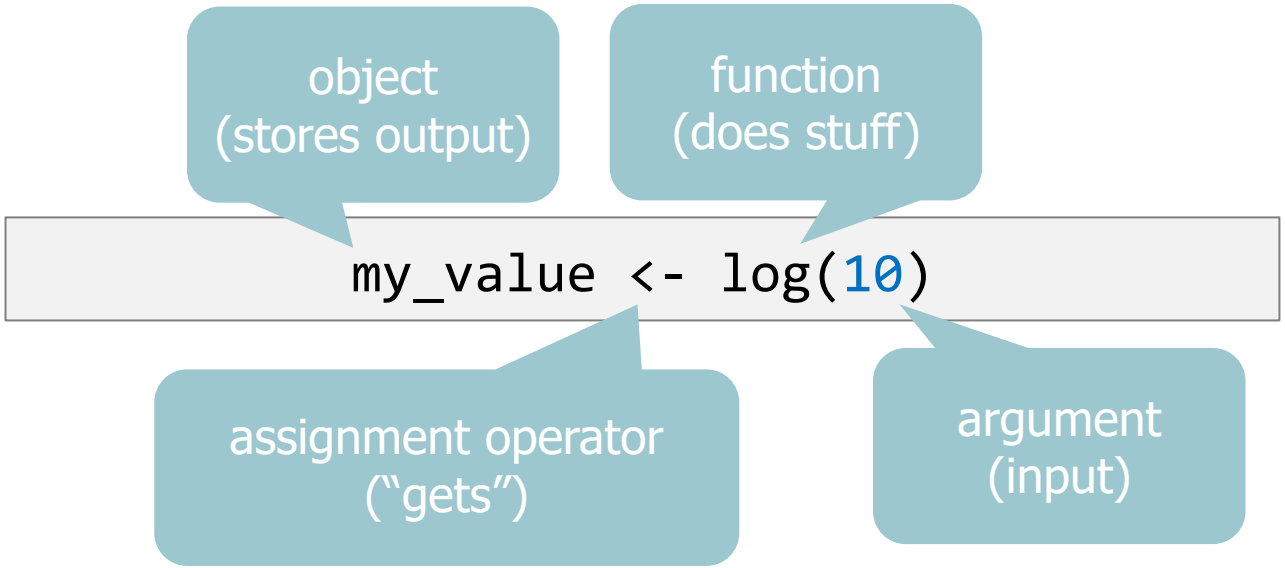
Functions

- `log()` is an example of a **function**
- Functions have "arguments"



Variables

- Everything is stored as a variable
- Variables are assigned using `<-`



Math with variables

Math using variables with just *one* value

```
> x <- 5
> x
```

```
[1] 5
```

```
> x + 3
```

```
[1] 8
```

Math using variables with *multiple* values

```
> a <- 3:6
> a
```

```
[1] 3 4 5 6
```

```
> a+2; a*3
```

```
[1] 5 6 7 8
```

```
[1]  9 12 15 18
```

Variable types

- `numeric` – `110, 55.2, 123`
- `integer` – `1L, 55L, 100L` (where "L" declares this as an integer)
- `character` – `"hello", "R is fun"` (sometimes called strings)
- `logical` – `TRUE` or `FALSE`
- `missing` - `NA`

We can use the `class()` function to check the data type of a variable

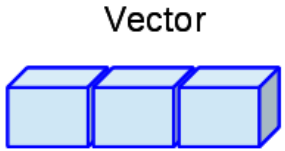
```
class(2023)
```

```
class("example")
```

Ann & Robert H. Lurie Children's Hospital of Chicago®

Object types

Vector: One-dimensional dataset of *one* data type

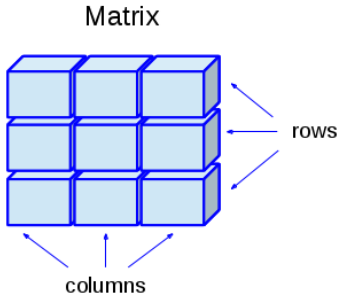



```
c(1, 2, 3)
```

```
c("one", "two", "three")
```

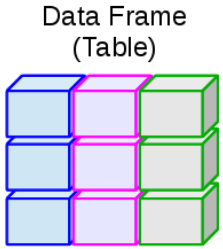
```
c(TRUE, FALSE, FALSE)
```

Matrix: Two-dimensional dataset of *one* data type




```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
```

Data frame: Two-dimensional dataset of *any* data type

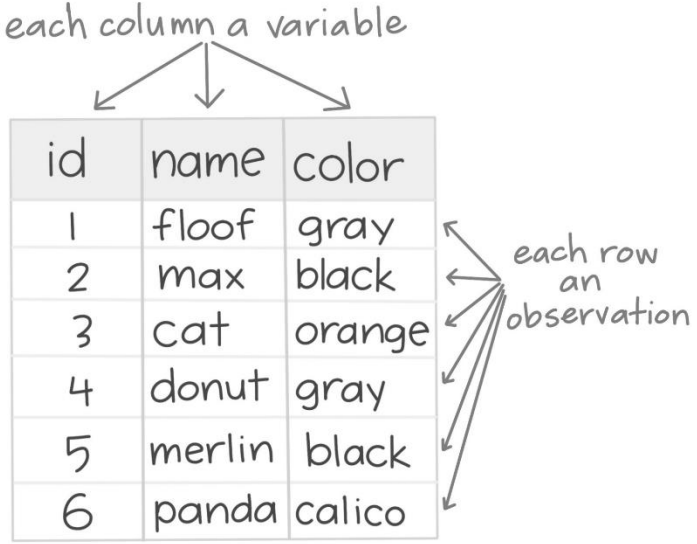



```
  month  day  year
1 May    1    2023
2 June   12   2023
3 July   10   2023
```

Tidy data frames

In tidy data:

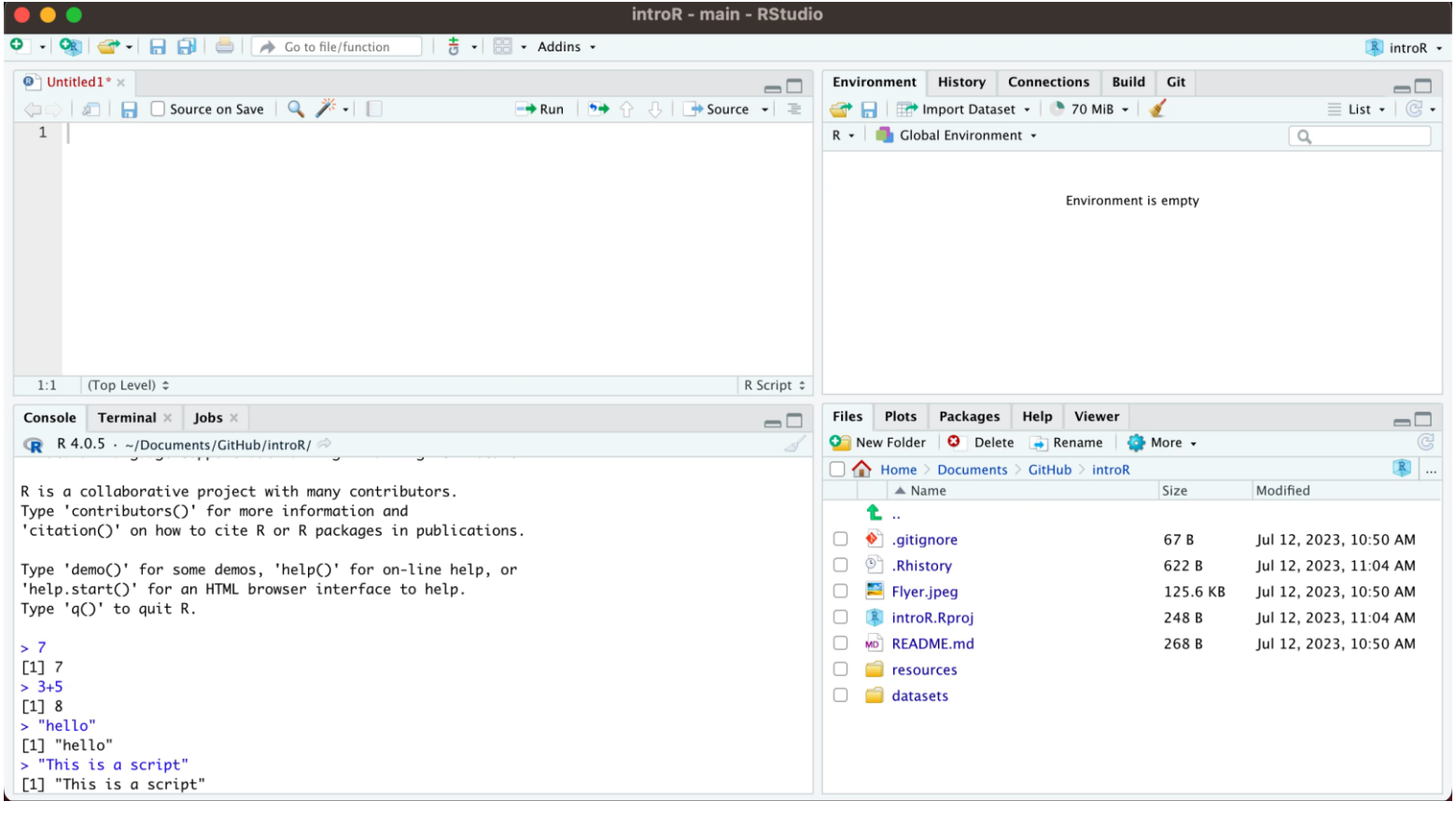
- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Getting data into RStudio

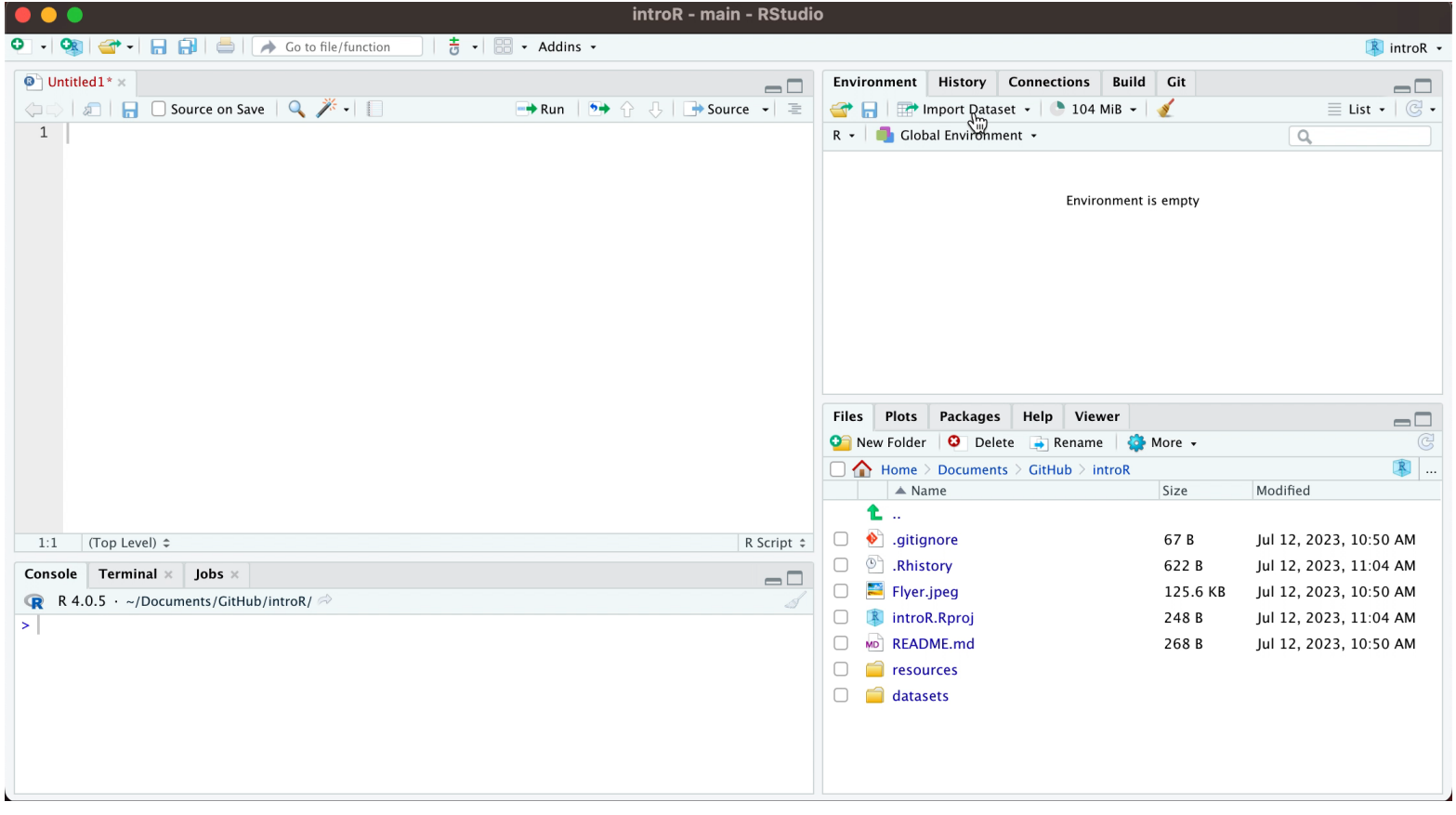
Importing data using the environment window

Ann & Robert H. Lurie Children's Hospital of Chicago



Importing **excel** data using the environment window

Ann & Robert H. Lurie Children's Hospital of Chicago



Importing data using code

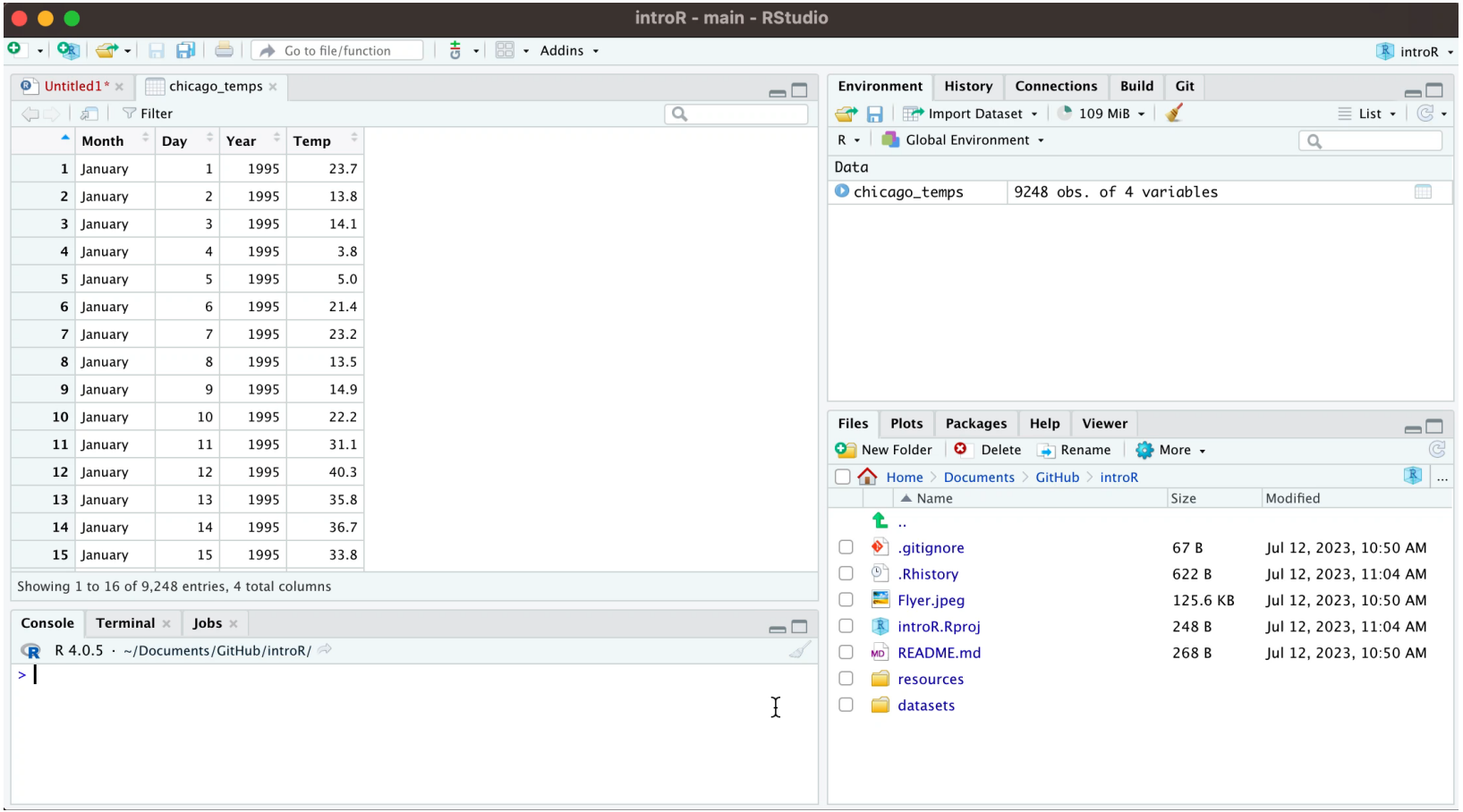
Code Preview:

```
library(readxl)
chicago_temps <- read_excel("datasets/chicago_temps
.xlsx")
View(chicago_temps)
```

- `readxl` is an R package
- R packages are a collection of data, code, and functions developed by the R user community
- Packages are stored under a directory called "library"

Library of packages

Ann & Robert H. Lurie Children's Hospital of Chicago



	Month	Day	Year	Temp
1	January	1	1995	23.7
2	January	2	1995	13.8
3	January	3	1995	14.1
4	January	4	1995	3.8
5	January	5	1995	5.0
6	January	6	1995	21.4
7	January	7	1995	23.2
8	January	8	1995	13.5
9	January	9	1995	14.9
10	January	10	1995	22.2
11	January	11	1995	31.1
12	January	12	1995	40.3
13	January	13	1995	35.8
14	January	14	1995	36.7
15	January	15	1995	33.8

Working with data

Basic operations on data: Selecting a column

Ann & Robert H. Lurie Children's Hospital of Chicago

A data frame named `2020_temp`:

	month	day	year	temp_F
1	January	1	2020	23
2	January	2	2020	14
3	January	3	2020	18
...				
363	December	29	2020	34
364	December	30	2020	36
365	December	31	2020	32

To single out the **month** column

```
2020_temp[, 1]
```

row

column

	month
1	January
2	January
3	January
...	
363	December
364	December
365	December

Basic operations on data: Selecting a column

Ann & Robert H. Lurie Children's Hospital of Chicago

A data frame named `2020_temp`:

	month	day	year	temp_F
1	January	1	2020	23
2	January	2	2020	14
3	January	3	2020	18
...				
363	December	29	2020	34
364	December	30	2020	36
365	December	31	2020	32

To single out the **month** column

```
2020_temp$month
```

name of column

	month
1	January
2	January
3	January
...	
363	December
364	December
365	December

Basic operations on data: Selecting multiple columns

Ann & Robert H. Lurie Children's Hospital of Chicago

A data frame named `2020_temp`:

	month	day	year	temp_F
1	January	1	2020	23
2	January	2	2020	14
3	January	3	2020	18
...				
363	December	29	2020	34
364	December	30	2020	36
365	December	31	2020	32

To single out **month**, **day**, and **temp_F**

```
2020_temp[, c(1, 2, 4)]
```

```
2020_temp[, c(month, day, temp_F)]
```

	month	day	temp_F
1	January	1	23
2	January	2	14
3	January	3	18
...			
363	December	29	34
364	December	30	36
365	December	31	32

Basic operations on data: Creating a new column

Ann & Robert H. Lurie Children's Hospital of Chicago

A data frame named `2020_temp`:

	month	day	year	temp_F
1	January	1	2020	23
2	January	2	2020	14
3	January	3	2020	18
...				
363	December	29	2020	34
364	December	30	2020	36
365	December	31	2020	32

To create a new column: **temp_C**

```
2020_temp$temp_C <-
(2020_temp$temp_F - 32)*(5/9)
```

	month	day	year	temp_F	temp_C
1	January	1	2020	23	-5
2	January	2	2020	14	-10
3	January	3	2020	18	-8
...					
363	December	29	2020	34	1
364	December	30	2020	36	2
365	December	31	2020	32	0

Basic operations on data: Subsetting data

Ann & Robert H. Lurie Children's Hospital of Chicago

A data frame named `2020_temp`:

	month	day	year	temp_F
1	January	1	2020	23
2	January	2	2020	14
3	January	3	2020	18
...				
363	December	29	2020	34
364	December	30	2020	36
365	December	31	2020	32

To select rows where **temp_F** < 20

```
2020_temp[2020_temp$temp_F < 20,]
```

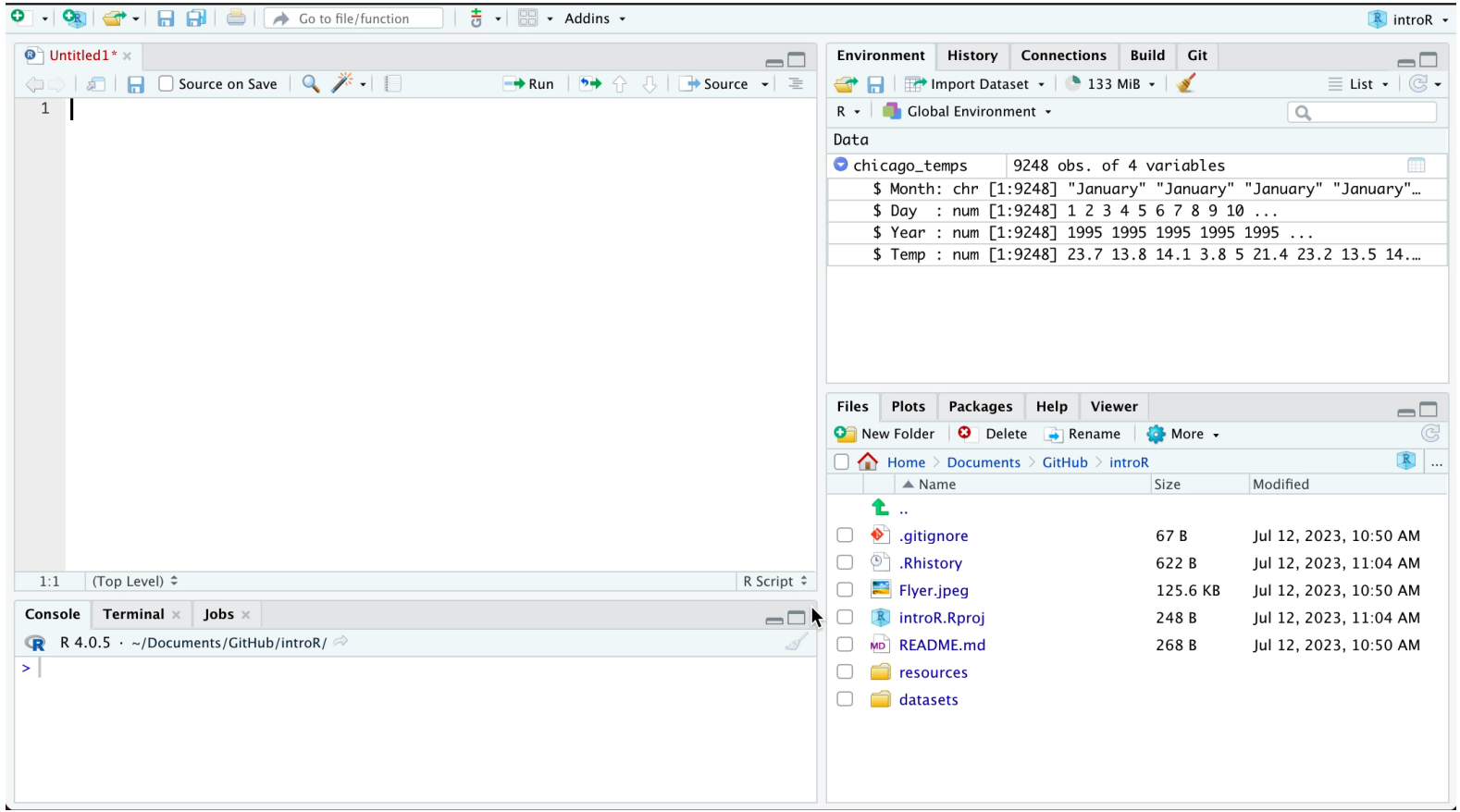
row

column

	month	day	year	temp_F
1	January	2	2020	14
2	January	3	2020	18
3	January	4	2020	19
...				
53	December	17	2020	10
54	December	18	2020	12
55	December	22	2020	15

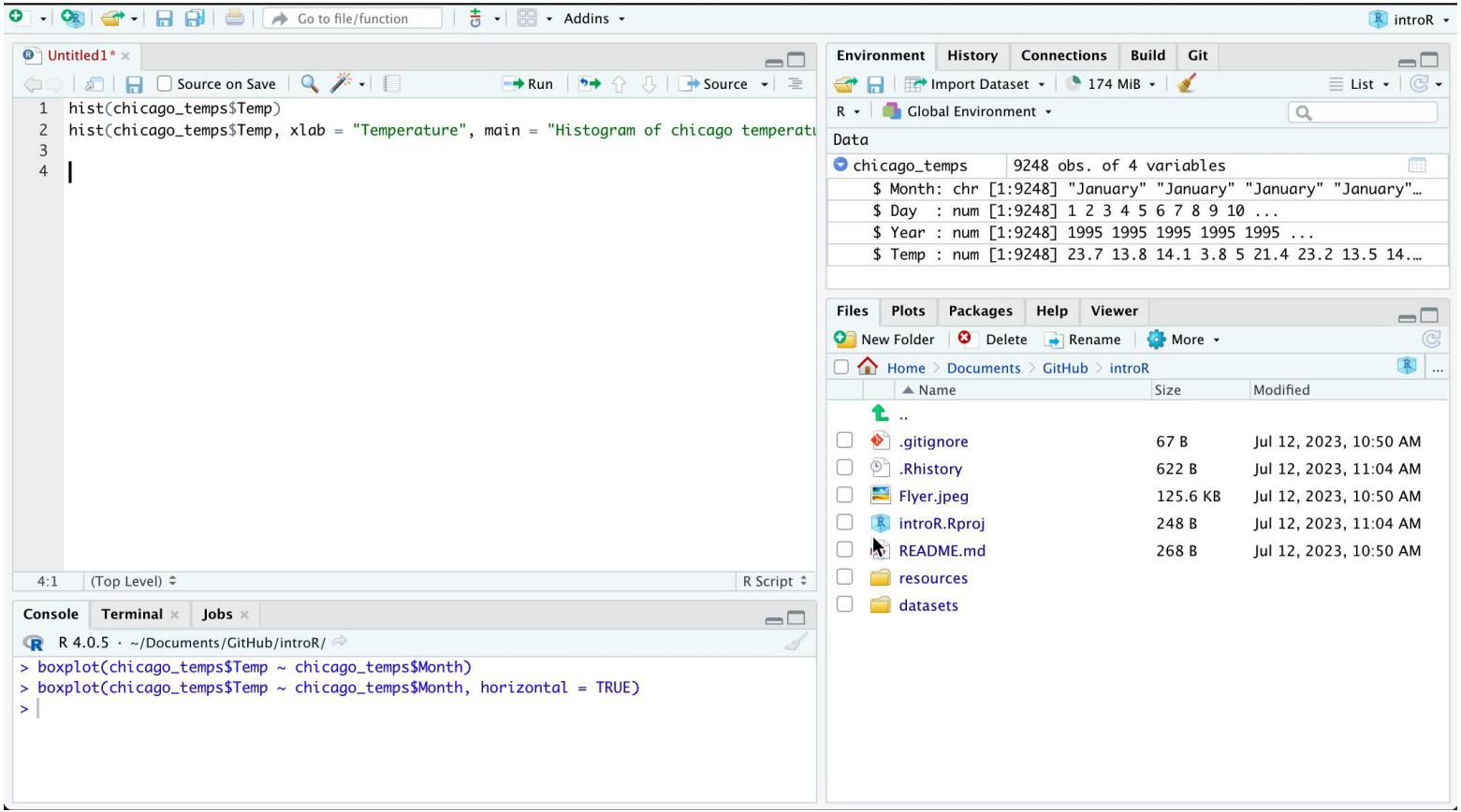
Visualizing data

Basic plots: Histogram

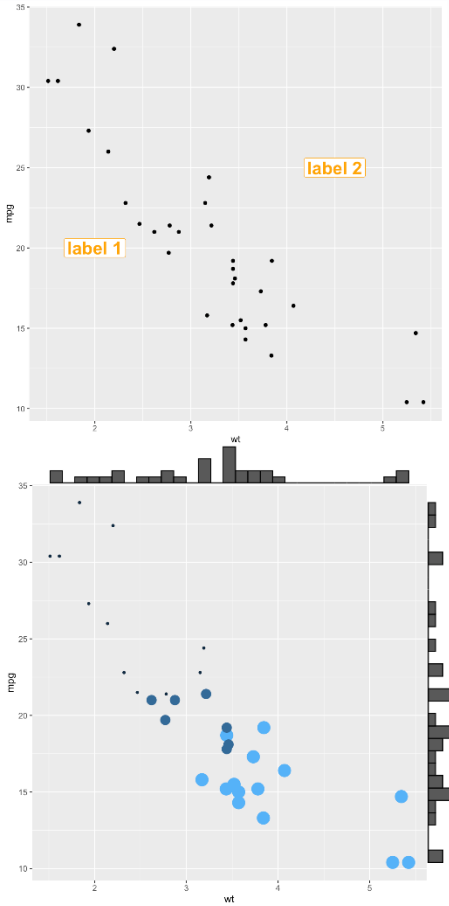


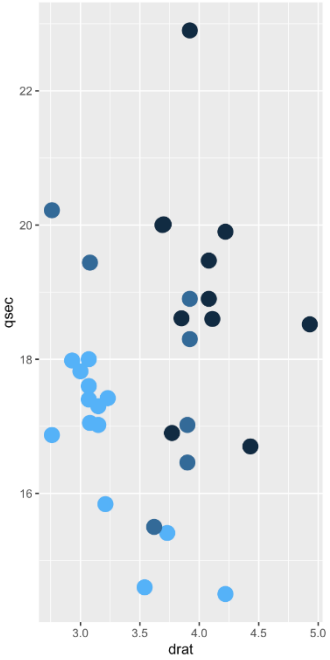
Basic plots: Boxplot

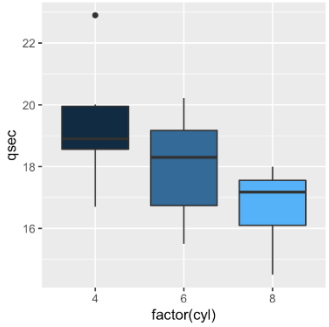
Ann & Robert H. Lurie Children's Hospital of Chicago

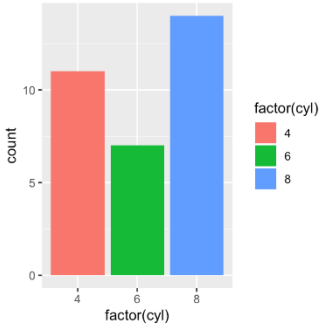


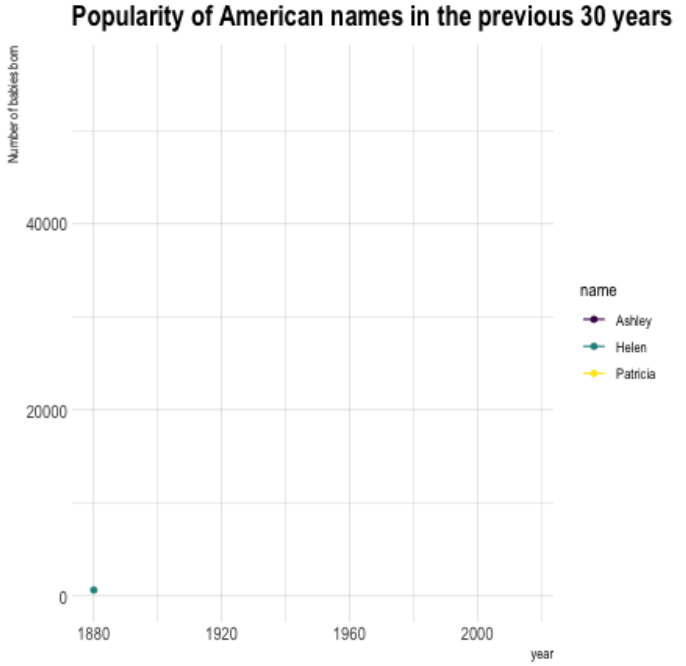
Endless plotting possibilities







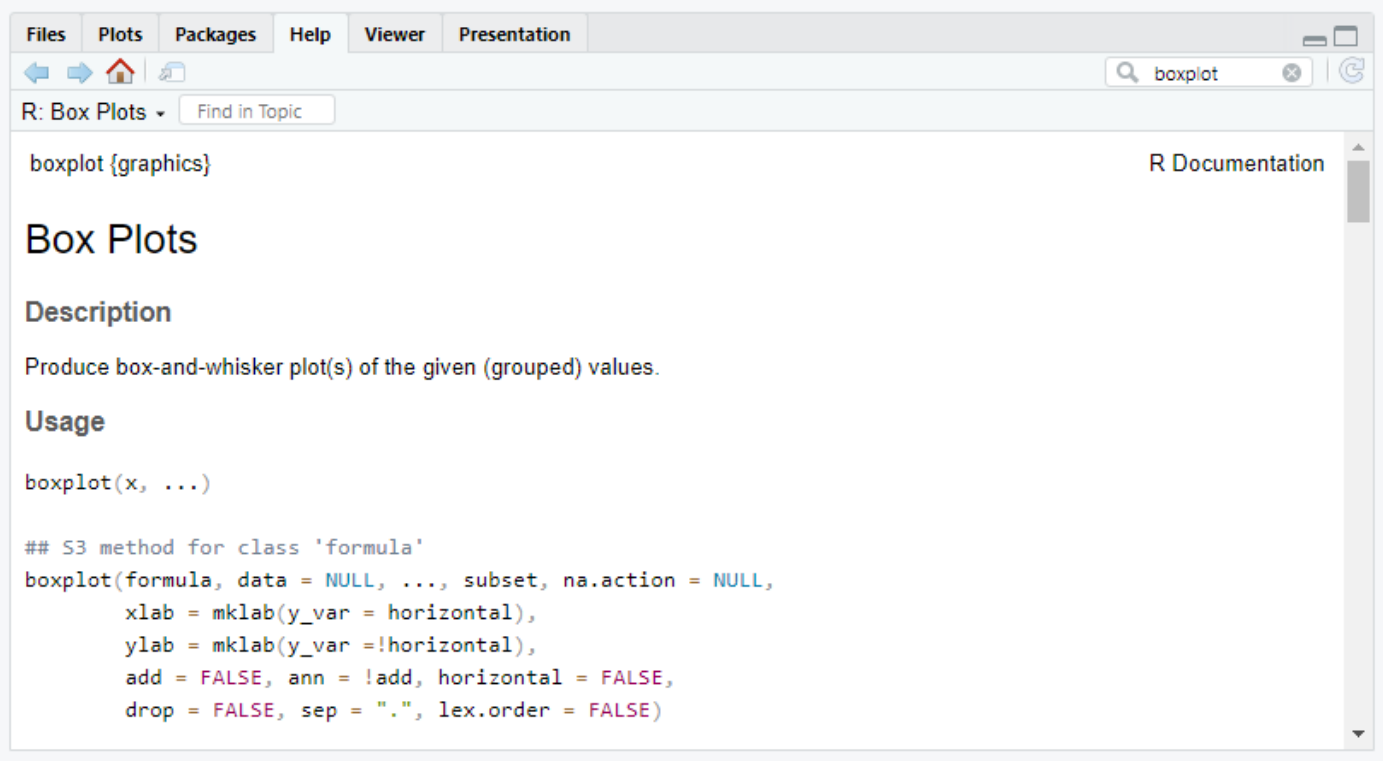




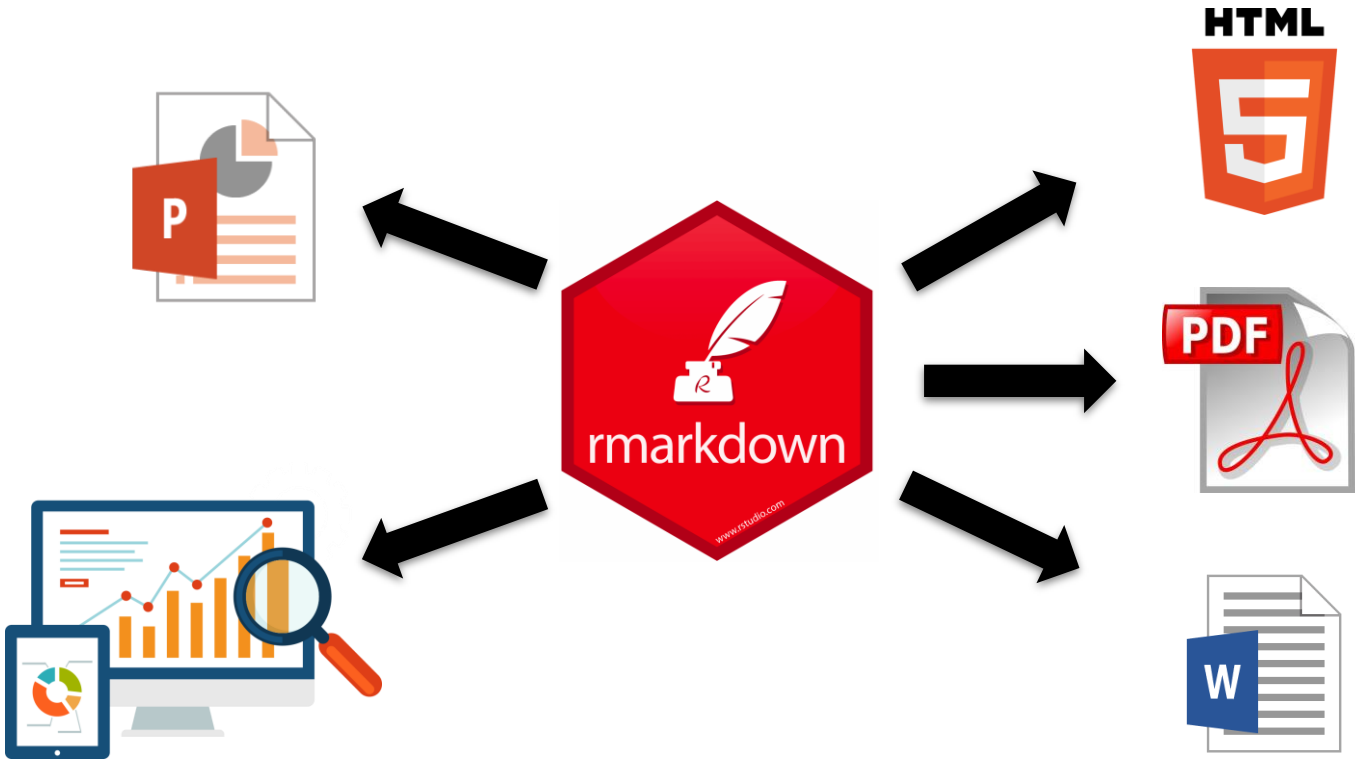
The more you know…

How to get help

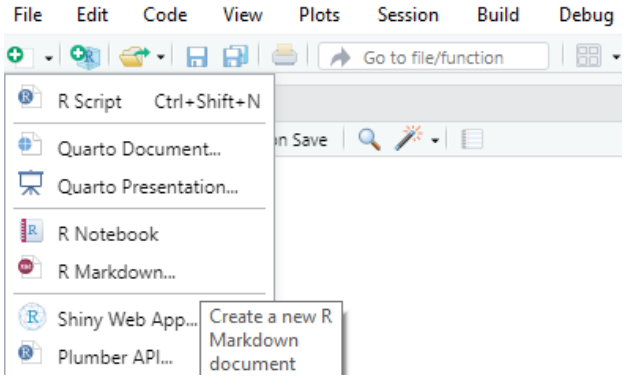
```
> ?boxplot
```

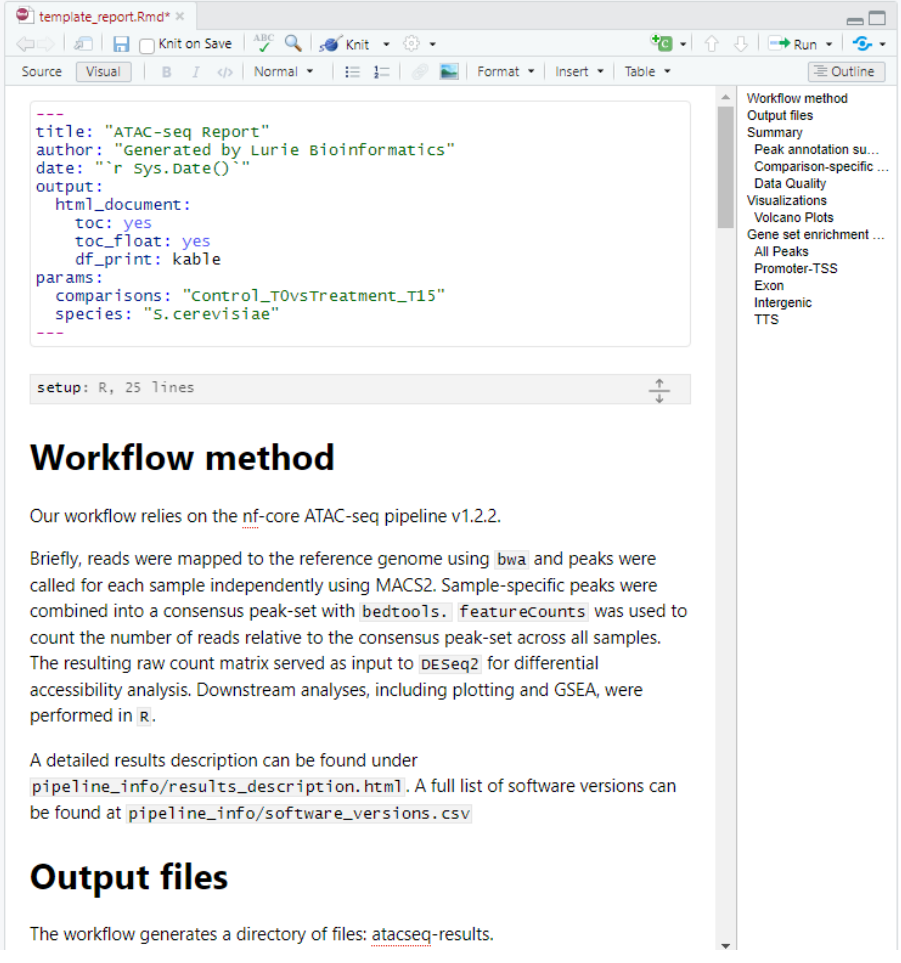


Reproducible reports with Rmarkdown



Rmarkdown reports





Interactive reports with Rmarkdown

Ann & Robert H. Lurie Children's Hospital of Chicago

ATAC-seq Report

Generated by Lurie Bioinformatics

2023-06-02

Workflow method

Our workflow relies on the nf-core ATAC-seq pipeline v1.2.2.

Briefly, reads were mapped to the reference genome using `bwa` and peaks were called for each sample independently using MACS2. Sample-specific peaks were combined into a consensus peak-set with `bedtools.` `featureCounts` was used to count the number of reads relative to the consensus peak-set across all samples. The resulting raw count matrix served as input to `DESeq2` for differential accessibility analysis. Downstream analyses, including plotting and GSEA, were performed in `R`.

A detailed results description can be found under `pipeline_info/results_description.html`. A full list of software versions can be found at `pipeline_info/software_versions.csv`

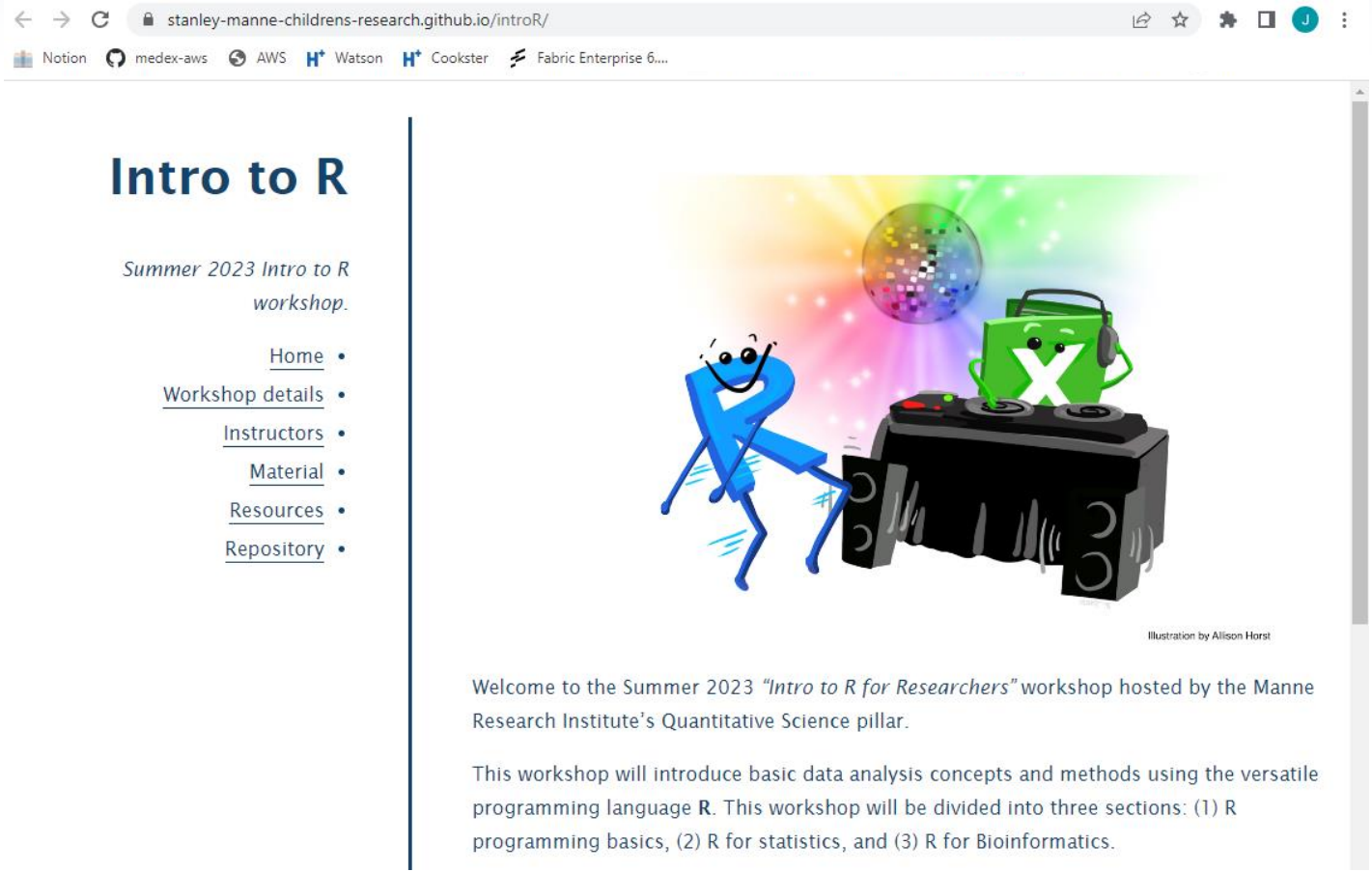
Output files

The workflow generates a directory of files: atacseq-results.

```
atacseq-results
├── peakannotatedresults.xlsx
├── bigwig
│   ├── ...
│   └── <sample>.mLb.clN.bigWig
├── DiffAnalysis
│   ├── consensus_peaks.mRp.clN.annotatePeaks.txt
```

Basic websites with Rmarkdown

Ann & Robert H. Lurie Children's Hospital of Chicago



Questions?