15th Bioinformatics in Health Sciences course

Introduction to R

Ana Gonçalves
ICVS/3B's, University of Minho

Nuno S. Osório
ICVS/3B's, University of Minho

ICVS

February 7th 2024

University of Minho
School of Medicine

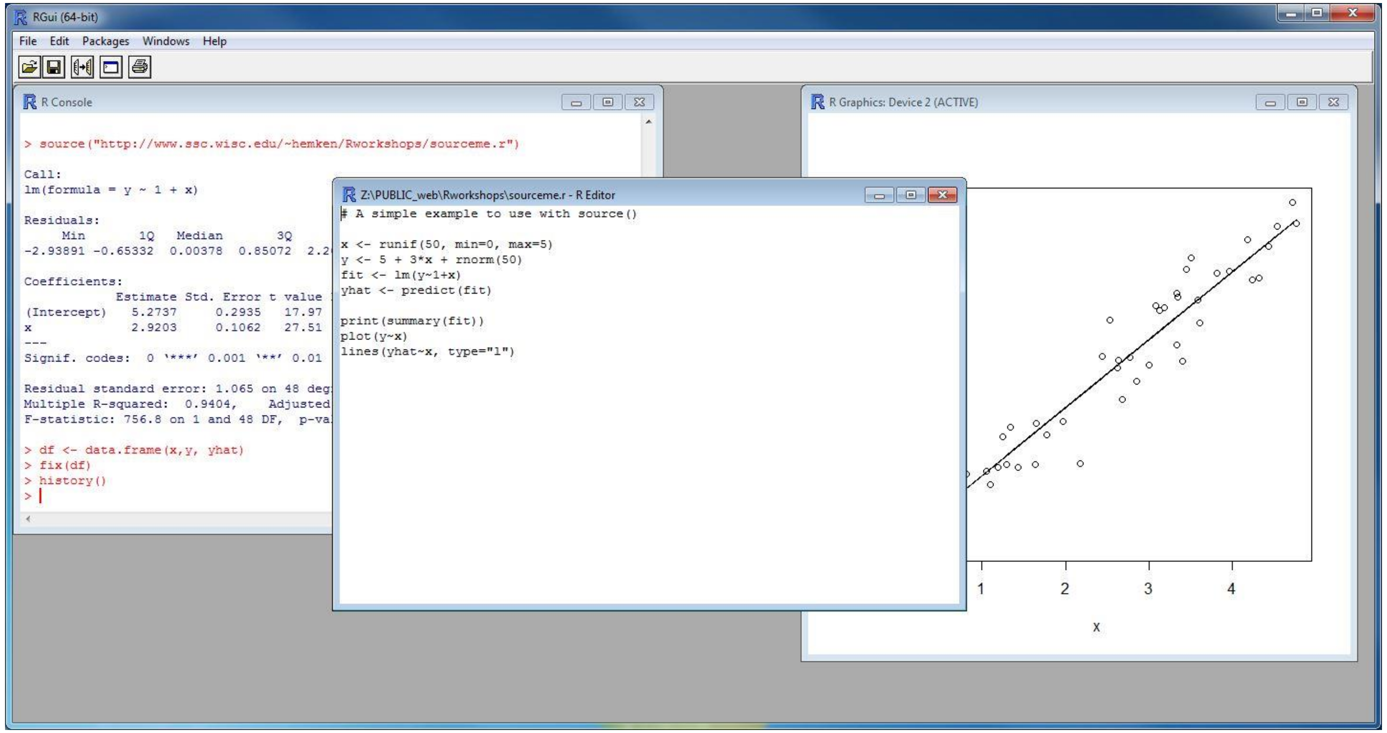
ICVS/3B's
Associate
Laboratory

WHY LEARN R?

- R is available for Windows, OS X, and Linux/Unix platforms.
- R is free and open-source software, with no license restrictions.
- R plays a key role in a wide variety of research and data analysis projects because it makes many modern statistical methods, both simple and advanced, readily available and easy to use.
- R provides a wonderfully flexible programming environment favored by the many researchers who do some form of data analysis as part of their work.
- Perhaps the most appealing feature of R is that any programmers can contribute code in the form of *packages* (or *libraries*), so the rest of the world has fast access to developments in statistics and data science.

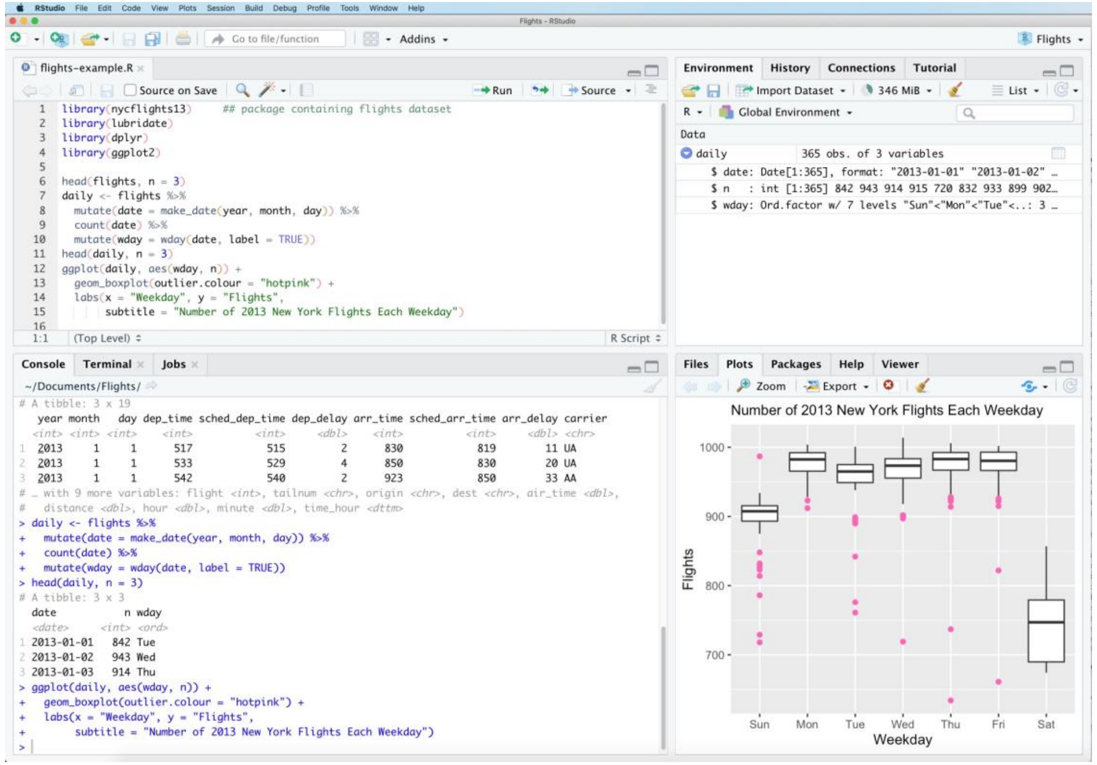
R vs RStudio®

R refers to a programming language as well as the software that runs R code.
RStudio is a software interface that can make it easier to write R scripts and interact with the R software.


RGui (64-bit)
File Edit Packages Windows Help
R Console
Z:\PUBLIC_web\Rworkshops\sourceme.r - R Editor
R Graphics: Device 2 (ACTIVE)
1
2
3
4
x

```
> source("http://www.ssc.wisc.edu/~hemken/Rworkshops/sourceme.r")

Call:
lm(formula = y ~ 1 + x)

Residuals:
     Min       1Q   Median       3Q
-2.93891 -0.65332  0.00378  0.85072  2.2

Coefficients:
            Estimate Std. Error t value
(Intercept)   5.2737     0.2935   17.97
x             2.9203     0.1062   27.51
---
Signif. codes:  0 '***' 0.001 '**' 0.01

Residual standard error: 1.065 on 48 deg
Multiple R-squared:  0.9404,    Adjusted
F-statistic: 756.8 on 1 and 48 DF,  p-va

> df <- data.frame(x,y, yhat)
> fix(df)
> history()
>
```

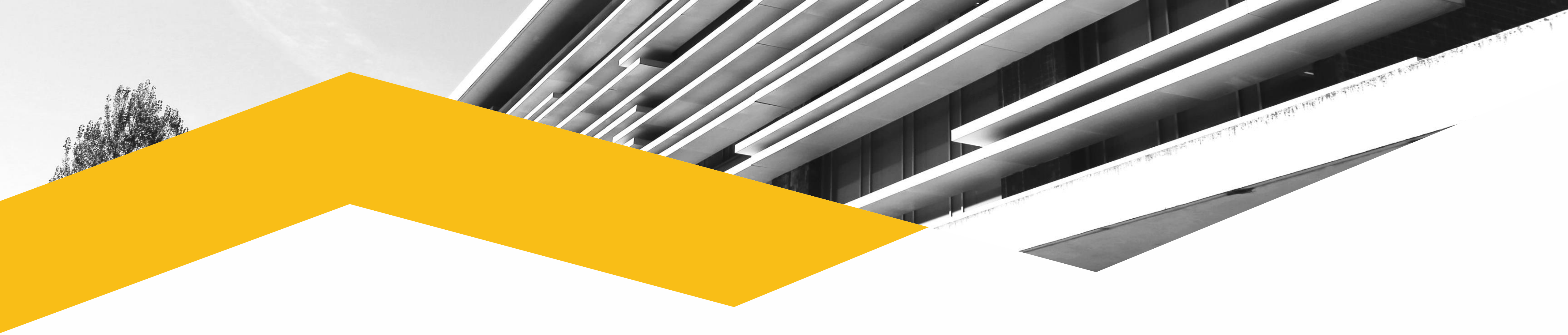
```
# A simple example to use with source()

x <- runif(50, min=0, max=5)
y <- 5 + 3*x + rnorm(50)
fit <- lm(y~1+x)
yhat <- predict(fit)

print(summary(fit))
plot(y~x)
lines(yhat~x, type="l")
```



```
library(nycflights13)     ## package containing flights dataset
library(lubridate)
library(dplyr)
library(ggplot2)

head(flights, n = 3)
daily <- flights %>%
  mutate(date = make_date(year, month, day)) %>%
  count(date) %>%
  mutate(wday = wday(date, label = TRUE))
head(daily, n = 3)
ggplot(daily, aes(wday, n)) +
  geom_boxplot(outlier.colour = "hotpink") +
  labs(x = "Weekday", y = "Flights",
       subtitle = "Number of 2013 New York Flights Each Weekday")
```

```
Data
daily                365 obs. of 3 variables
   $ date: Date[1:365], format: "2013-01-01" "2013-01-02" …
   $ n   : int [1:365] 842 943 914 915 720 832 933 899 902…
   $ wday: Ord.factor w/ 7 levels "Sun"<"Mon"<"Tue"<..: 3 …
```

```
~/Documents/Flights/
# A tibble: 3 x 19
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay carrier
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>     <dbl> <chr>
1  2013     1     1      517            515         2      830            819        11 UA
2  2013     1     1      533            529         4      850            830        20 UA
3  2013     1     1      542            540         2      923            850        33 AA
# … with 9 more variables: flight <int>, tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>,
#   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>
> daily <- flights %>%
+   mutate(date = make_date(year, month, day)) %>%
+   count(date) %>%
+   mutate(wday = wday(date, label = TRUE))
> head(daily, n = 3)
# A tibble: 3 x 3
  date           n wday
  <date>     <int> <ord>
1 2013-01-01   842 Tue
2 2013-01-02   943 Wed
3 2013-01-03   914 Thu
> ggplot(daily, aes(wday, n)) +
+   geom_boxplot(outlier.colour = "hotpink") +
+   labs(x = "Weekday", y = "Flights",
+        subtitle = "Number of 2013 New York Flights Each Weekday")
>
```



Time to do some coding!

Let’s go open both R and Rstudio to further understand the differences!

Sciences course15th Bioinformatics in Health
Introduction to R

February 7th 2024

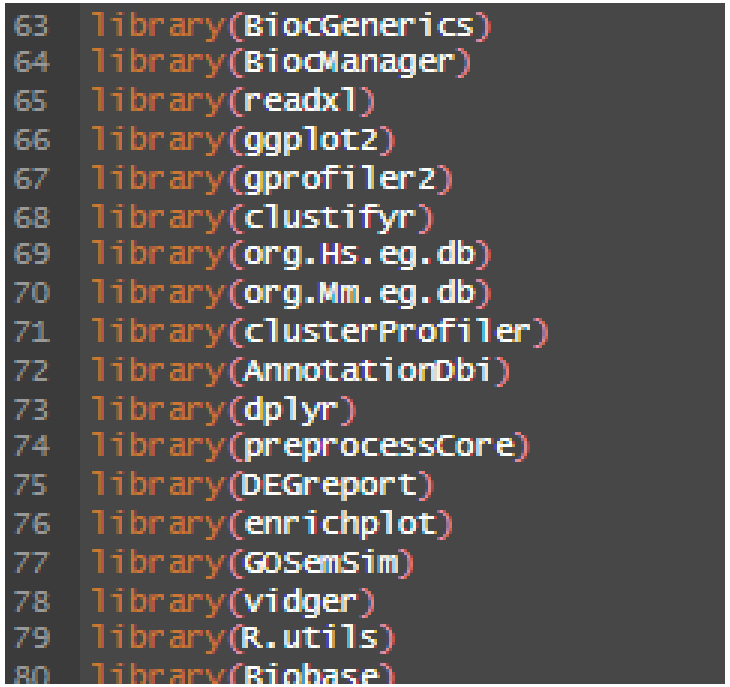
R: Packages

Packages are the fundamental units of reproducible R code. They include reusable R functions, the documentation that describes how to use them, and sample data.

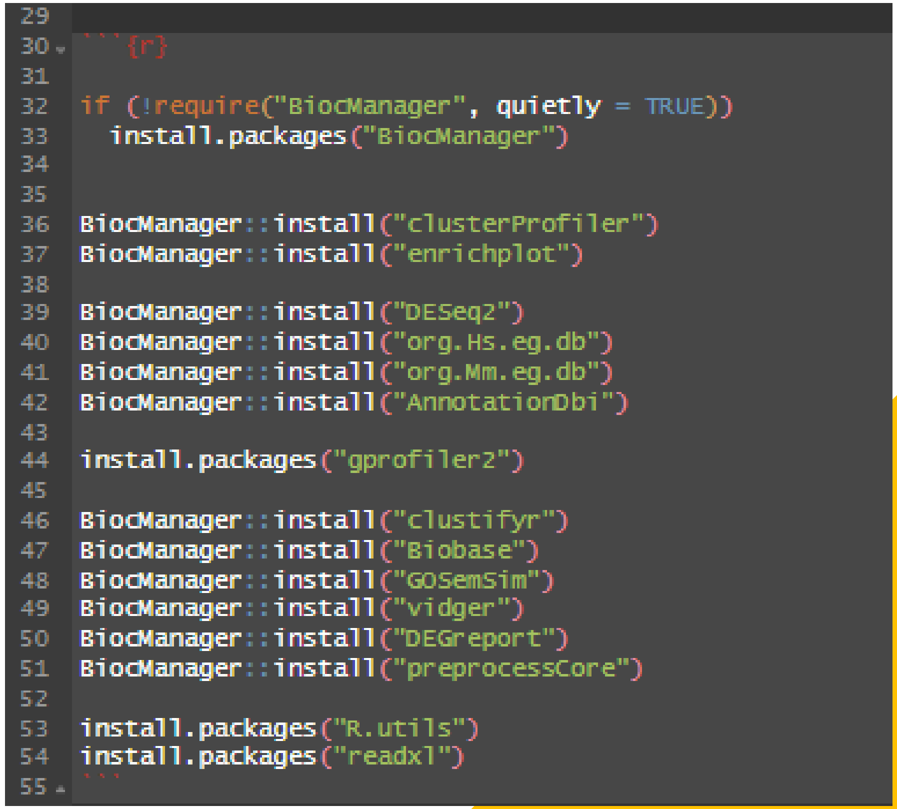
- Once a package is installed, we don't need to install it again in our R studio.

BUT…

- **We always need to call a library in order to run specific functions.**

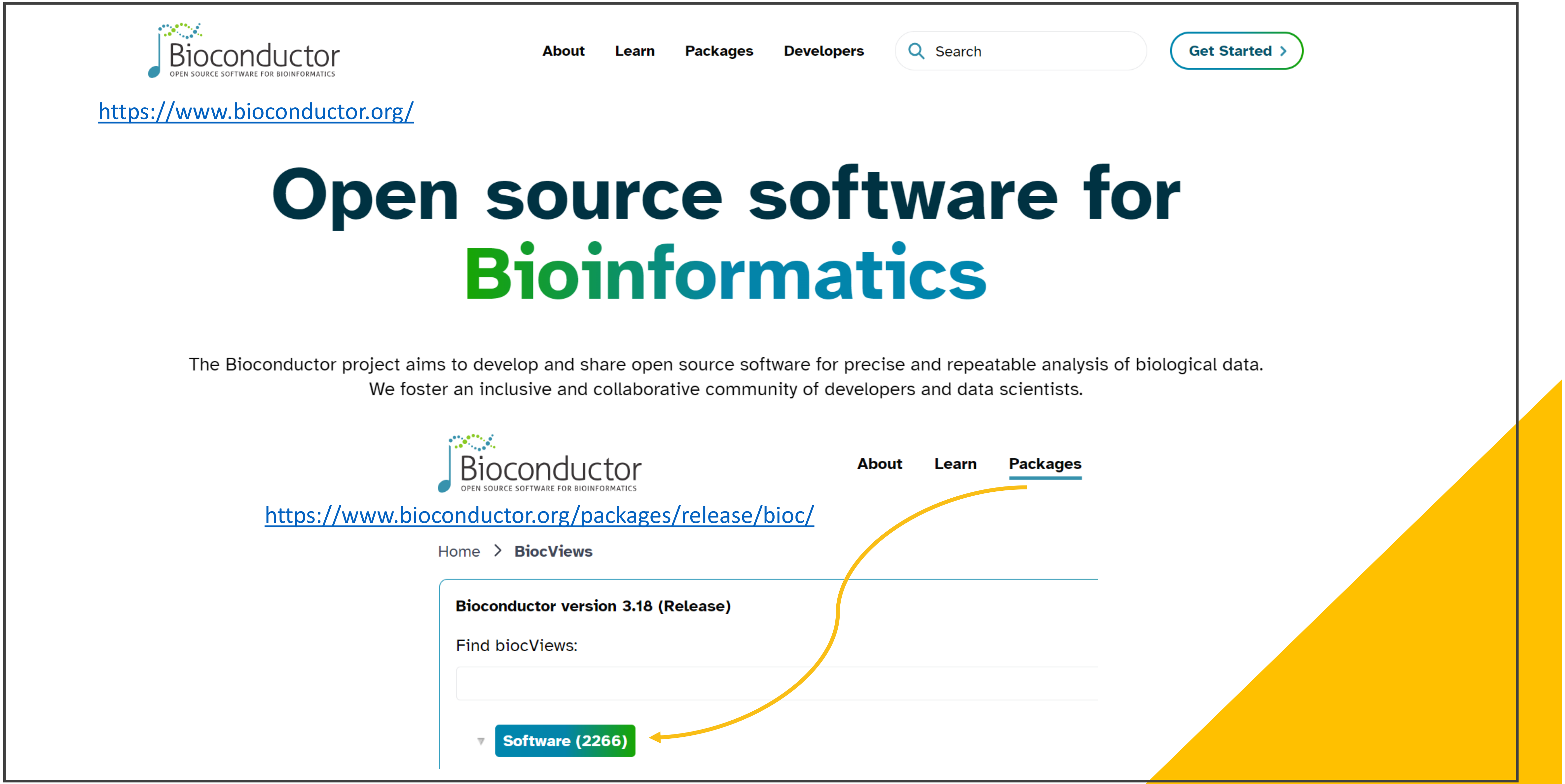


```r
library(BiocGenerics)
library(BiocManager)
library(readxl)
library(ggplot2)
library(gprofiler2)
library(clustifyr)
library(org.Hs.eg.db)
library(org.Mm.eg.db)
library(clusterProfiler)
library(AnnotationDbi)
library(dplyr)
library(preprocessCore)
library(DEGreport)
library(enrichplot)
library(GOSemSim)
library(vidger)
library(R.utils)
library(Biobase)
```



~~~r

```{r}

if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")


BiocManager::install("clusterProfiler")
BiocManager::install("enrichplot")

BiocManager::install("DESeq2")
BiocManager::install("org.Hs.eg.db")
BiocManager::install("org.Mm.eg.db")
BiocManager::install("AnnotationDbi")

install.packages("gprofiler2")

BiocManager::install("clustifyr")
BiocManager::install("Biobase")
BiocManager::install("GOSemSim")
BiocManager::install("vidger")
BiocManager::install("DEGreport")
BiocManager::install("preprocessCore")

install.packages("R.utils")
install.packages("readxl")
~~~
```
```

R: Packages - BiocManager



https://www.bioconductor.org/

Open source software for Bioinformatics

The Bioconductor project aims to develop and share open source software for precise and repeatable analysis of biological data. We foster an inclusive and collaborative community of developers and data scientists.

https://www.bioconductor.org/packages/release/bioc/

Home › BiocViews

Bioconductor version 3.18 (Release)

Find biocViews:

R: files

To open a .txt file:

```
# Install utils package
install.packages(" utils ")
```

```
# Access utils package
library(utils)
```

Now, depending if the txt file has decimals, a header, and other factors, three functions can be selected:

```
# Read data from .xlsx sheet called "Year1" as data frame and assign to object

1. Data <- read.delim(file = "my_file.txt", header = TRUE, sep = "t", dec = ".")

2. Data <- read.delim2(file = "my_file.txt", header = TRUE, sep = "t", dec = ",")

3. Data <- read.table(file = "my_file.txt", header = TRUE)
```

R: files

To open a CSV file:

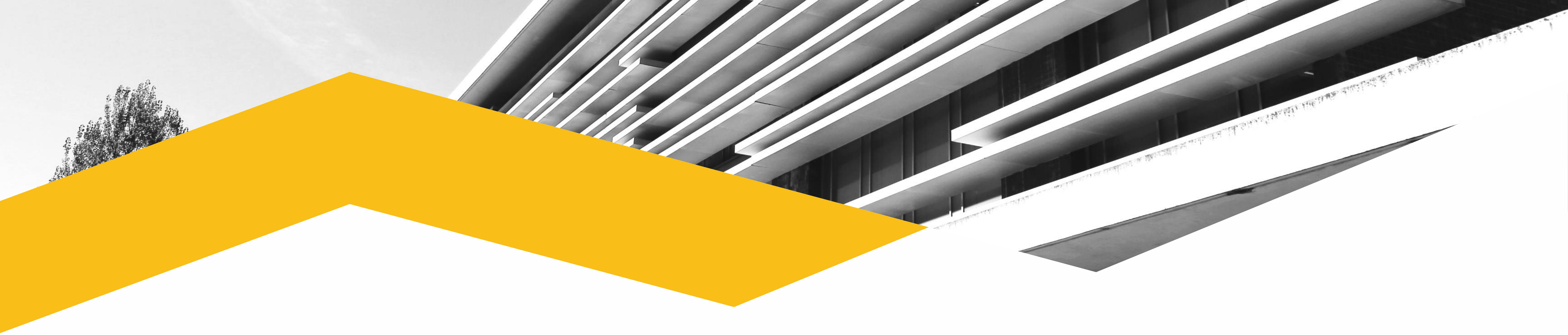
```
# Install readr package
install.packages("readr")

# Access readr package
library(readr)

# Read .csv file into R as data frame
Data <- read_csv(“transcriptomics.csv")

# Read .csv data file into R from a website
data <- read_csv("https://raw.githubusercontent.com/tutorial/transcriptomics.csv ")
```

We will use this method in our script

R: files

To open a .xlsx file:

```
# Install readxl package
install.packages("readxl")

# Access readxl package
library(readxl)

# Read data from .xlsx sheet called “transcriptomics" as data frame and assign to object
Data <- read_excel(“data.xlsx", sheet=“transcriptomics")
```



Time to do some coding!

RStudio®

Sciences course15th Bioinformatics in Health
Introduction to R

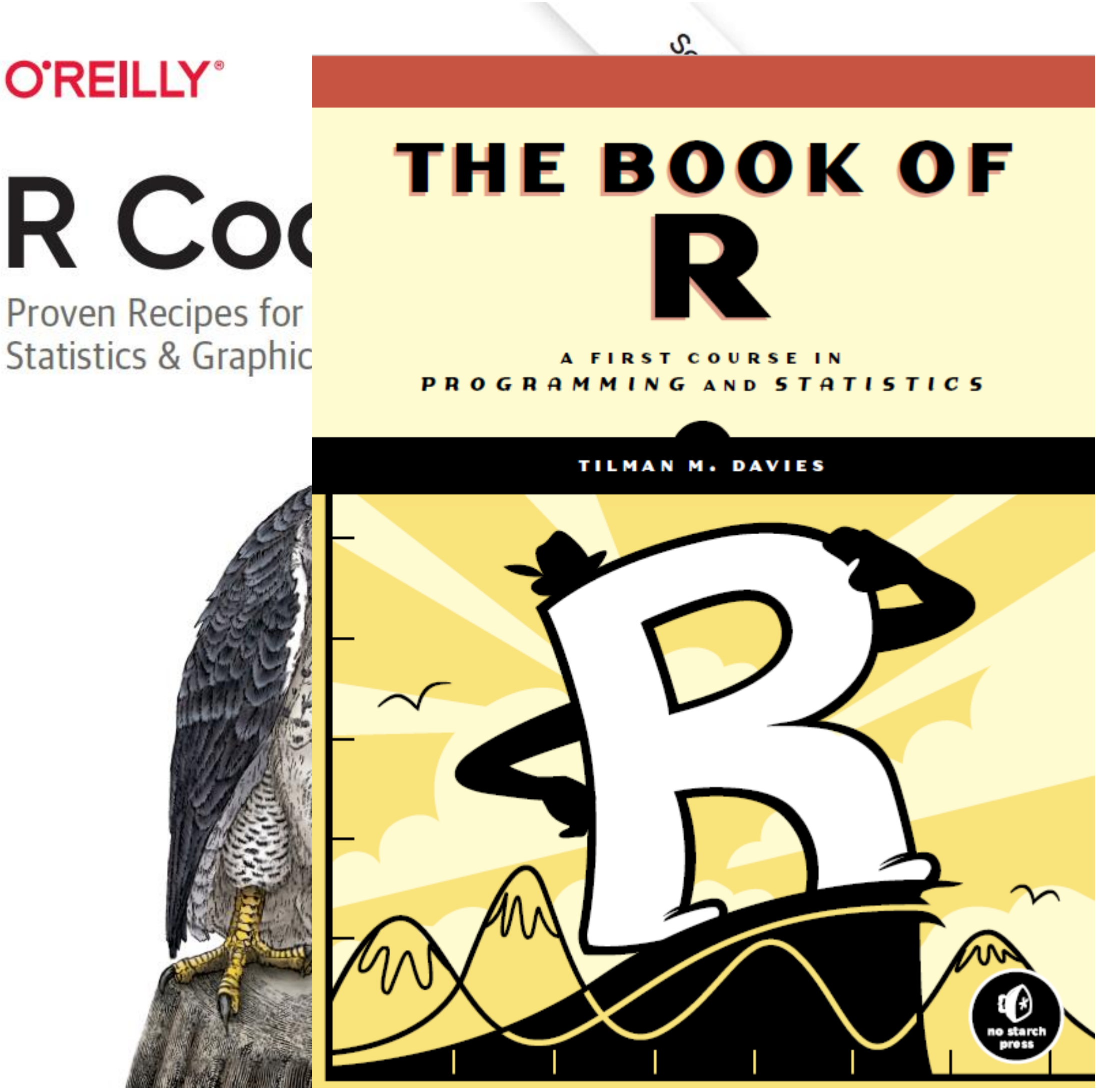
ICVS

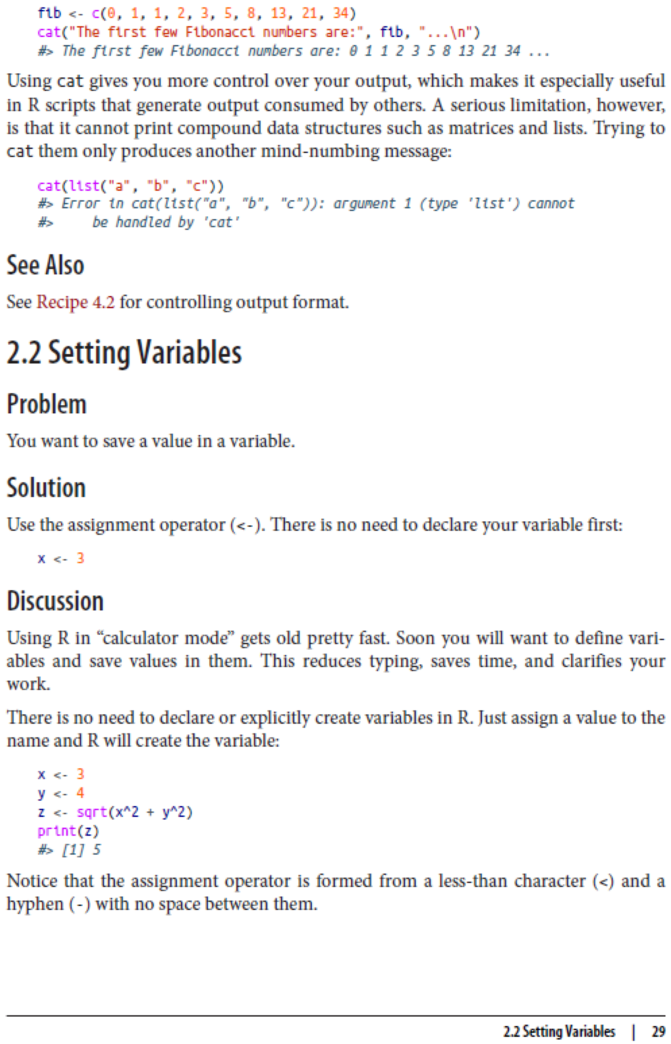
February 7th 2024

University of Minho
School of Medicine

ICVS/3B's
Associate
Laboratory

R: Books recommendations





```
fib <- c(0, 1, 1, 2, 3, 5, 8, 13, 21, 34)
cat("The first few Fibonacci numbers are:", fib, "...\n")
#> The first few Fibonacci numbers are: 0 1 1 2 3 5 8 13 21 34 ...
```

Using cat gives you more control over your output, which makes it especially useful in R scripts that generate output consumed by others. A serious limitation, however, is that it cannot print compound data structures such as matrices and lists. Trying to cat them only produces another mind-numbing message:

```
cat(list("a", "b", "c"))
#> Error in cat(list("a", "b", "c")): argument 1 (type 'list') cannot
#>     be handled by 'cat'
```

See Also

See Recipe 4.2 for controlling output format.

2.2 Setting Variables

Problem

You want to save a value in a variable.

Solution

Use the assignment operator (<-). There is no need to declare your variable first:

```
x <- 3
```

Discussion

Using R in "calculator mode" gets old pretty fast. Soon you will want to define variables and save values in them. This reduces typing, saves time, and clarifies your work.

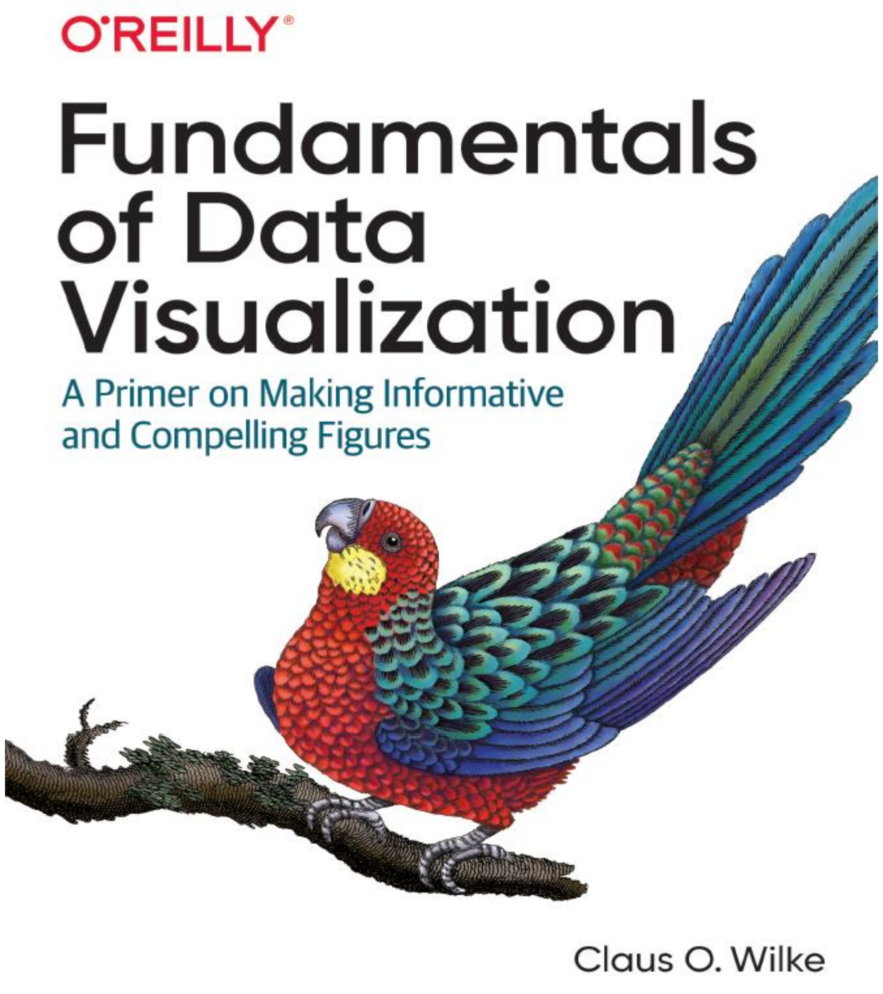
There is no need to declare or explicitly create variables in R. Just assign a value to the name and R will create the variable:

```
x <- 3
y <- 4
z <- sqrt(x^2 + y^2)
print(z)
#> [1] 5
```

Notice that the assignment operator is formed from a less-than character (<) and a hyphen (-) with no space between them.

2.2 Setting Variables | 29

R: Books recommendations

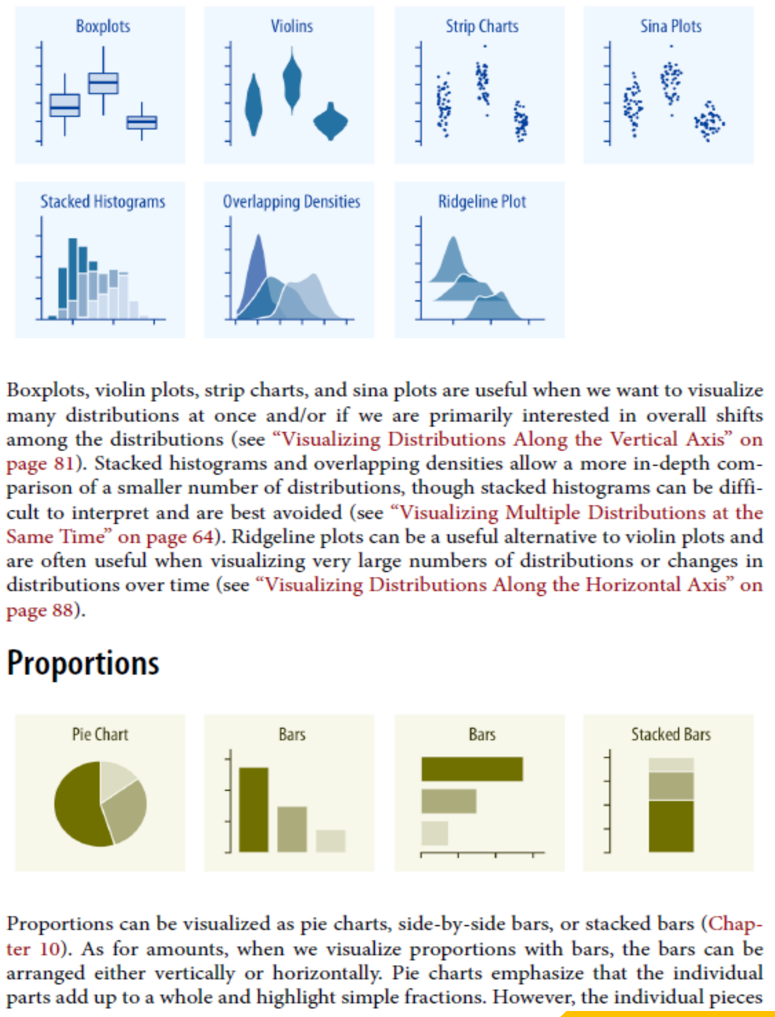


O'REILLY®

Fundamentals of Data Visualization

A Primer on Making Informative and Compelling Figures

Claus O. Wilke



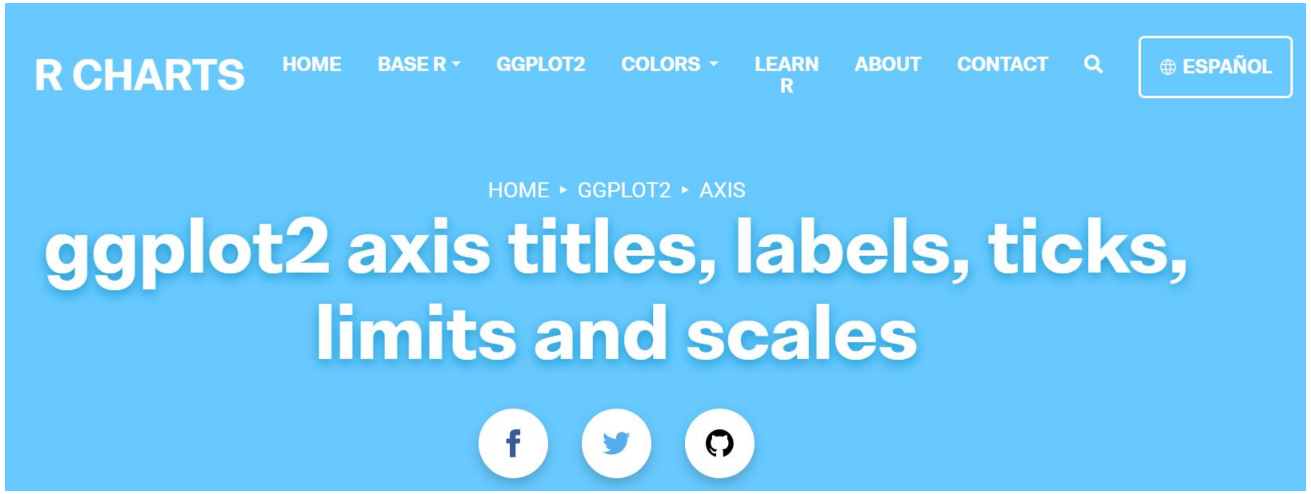
Boxplots, violin plots, strip charts, and sina plots are useful when we want to visualize many distributions at once and/or if we are primarily interested in overall shifts among the distributions (see "Visualizing Distributions Along the Vertical Axis" on page 81). Stacked histograms and overlapping densities allow a more in-depth comparison of a smaller number of distributions, though stacked histograms can be difficult to interpret and are best avoided (see "Visualizing Multiple Distributions at the Same Time" on page 64). Ridgeline plots can be a useful alternative to violin plots and are often useful when visualizing very large numbers of distributions or changes in distributions over time (see "Visualizing Distributions Along the Horizontal Axis" on page 88).

Proportions

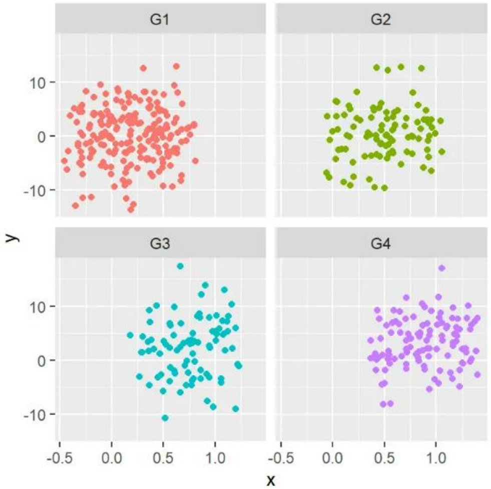
Proportions can be visualized as pie charts, side-by-side bars, or stacked bars (Chapter 10). As for amounts, when we visualize proportions with bars, the bars can be arranged either vertically or horizontally. Pie charts emphasize that the individual parts add up to a whole and highlight simple fractions. However, the individual pieces

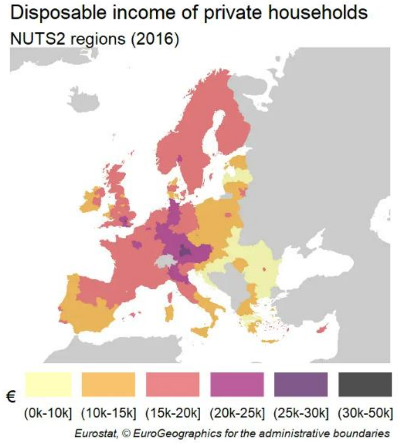
R: ggplot2 Package

https://cran.r-project.org/web/packages/ggplot2/index.html

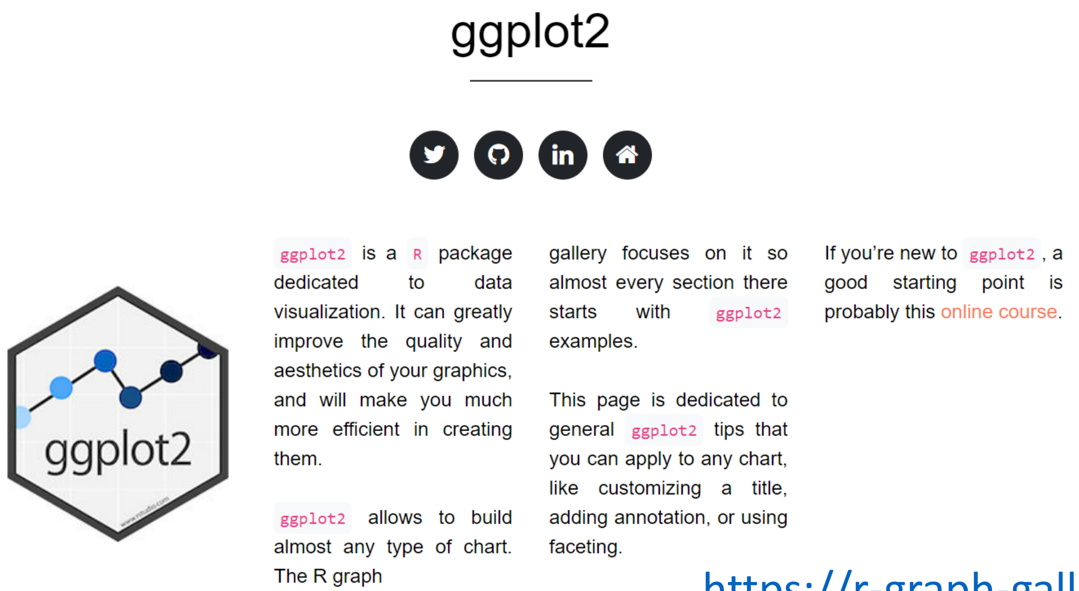


ggplot2 axis titles, labels, ticks, limits and scales




Disposable income of private households
NUTS2 regions (2016)
€
(0k-10k]
(10k-15k]
(15k-20k]
(20k-25k]
(25k-30k]
(30k-50k]
Eurostat, © EuroGeographics for the administrative boundaries

https://r-charts.com/ggplot2/



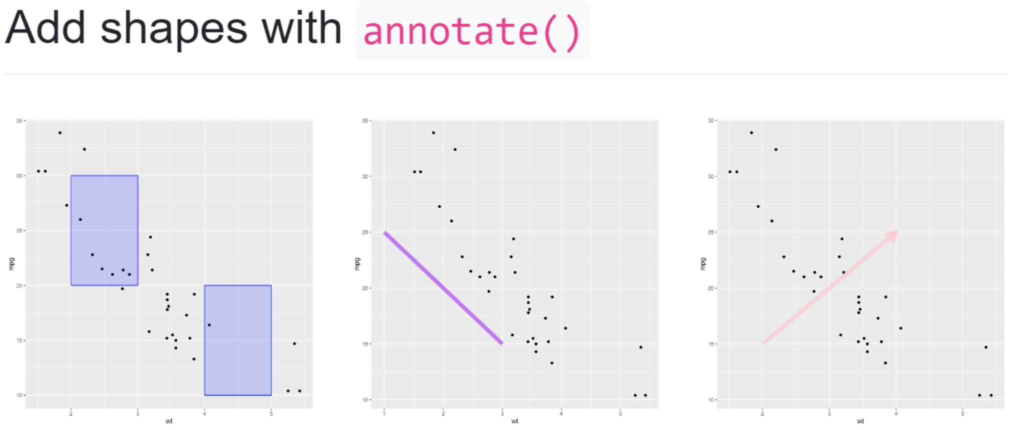
ggplot2

`ggplot2` is a `R` package dedicated to data visualization. It can greatly improve the quality and aesthetics of your graphics, and will make you much more efficient in creating them.

`ggplot2` allows to build almost any type of chart. The R graph gallery focuses on it so almost every section there starts with `ggplot2` examples.

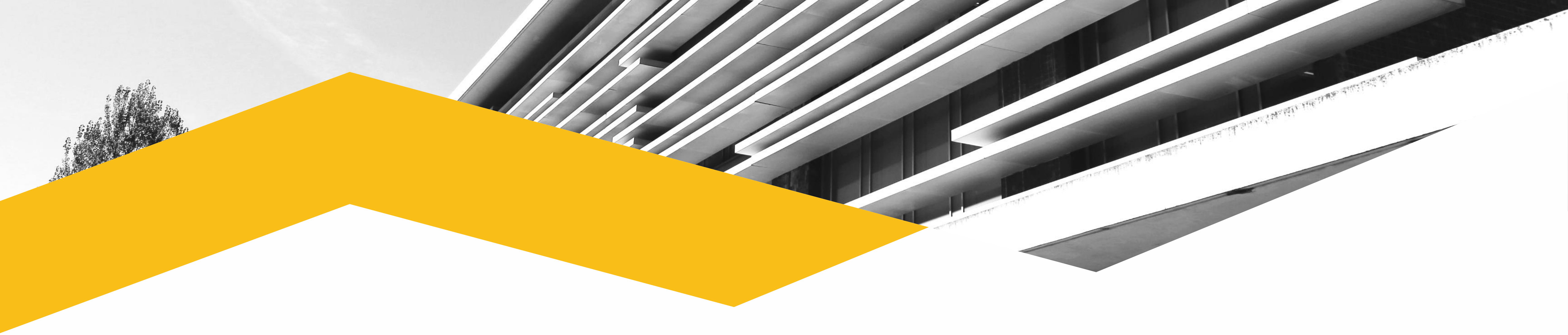
This page is dedicated to general `ggplot2` tips that you can apply to any chart, like customizing a title, adding annotation, or using faceting.

If you're new to `ggplot2`, a good starting point is probably this online course.



Add shapes with `annotate()`

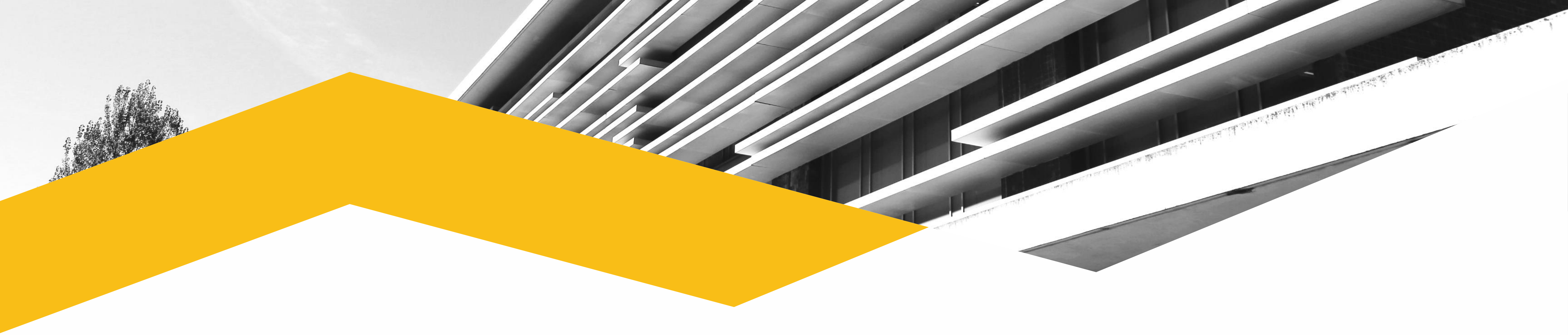
https://r-graph-gallery.com/ggplot2-package.html



Time to do some coding!

Sciences course15th Bioinformatics in Health
Introduction to R

February 7th 2024



15th Bioinformatics in Health Sciences course

Using R for Gene Expression Analysis

Ana Gonçalves
ICVS/3B's, University of Minho

Nuno S. Osório
ICVS/3B's, University of Minho

February 7th 2024

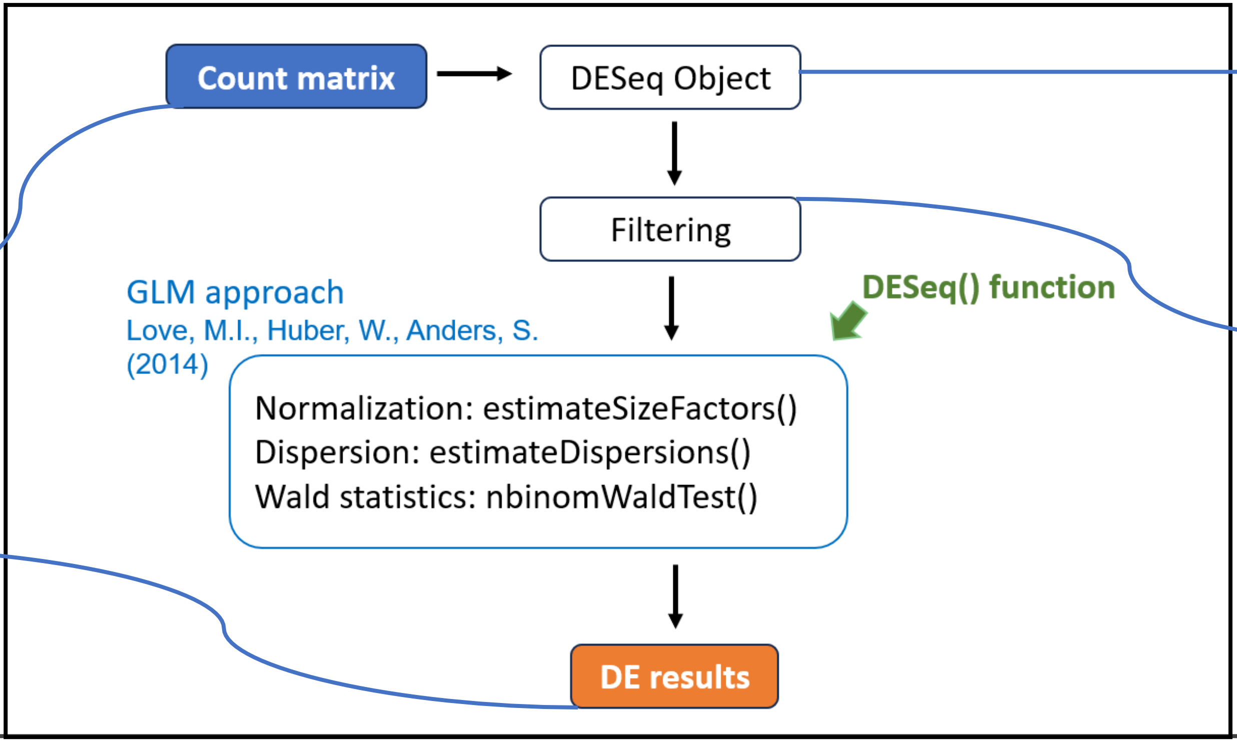
R: Package for DE genes analysis

DESeq2: A Popular RNA-Seq Analysis Package

Widely used and well-documented: DESeq2 is a widely used tool in the bioinformatics community, with a large user base and a well-documented user manual.

https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Workflow



Count matrix
DESeq Object
Filtering
GLM approach
Love, M.I., Huber, W., Anders, S. (2014)
DESeq() function
Normalization: estimateSizeFactors()
Dispersion: estimateDispersions()
Wald statistics: nbinomWaldTest()
DE results

An object used to store the input values, intermediate calculations and results of the analysis of differential expression.

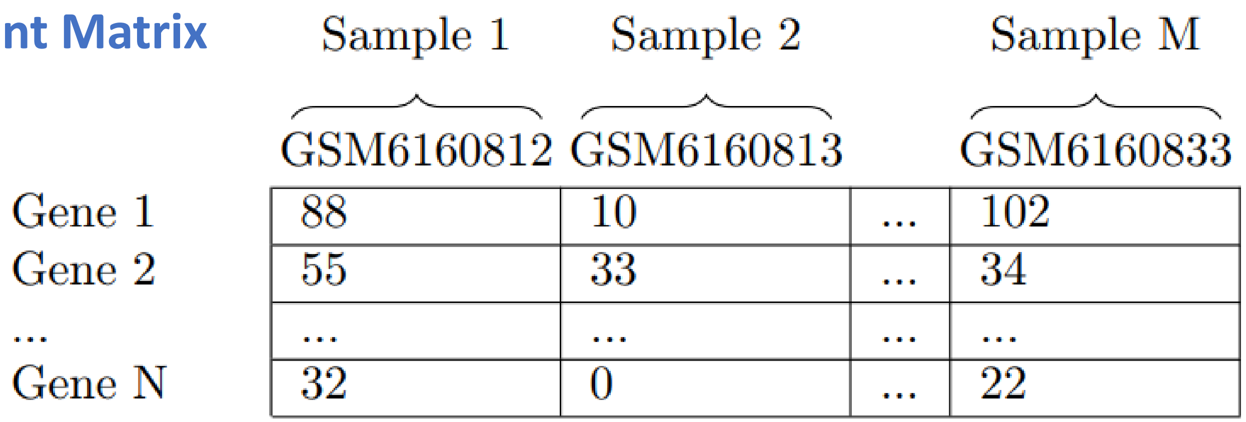
The reads from the RNA sequencing experiment

Removal of genes that have null reads across all samples

Acquire the differential expressed genes

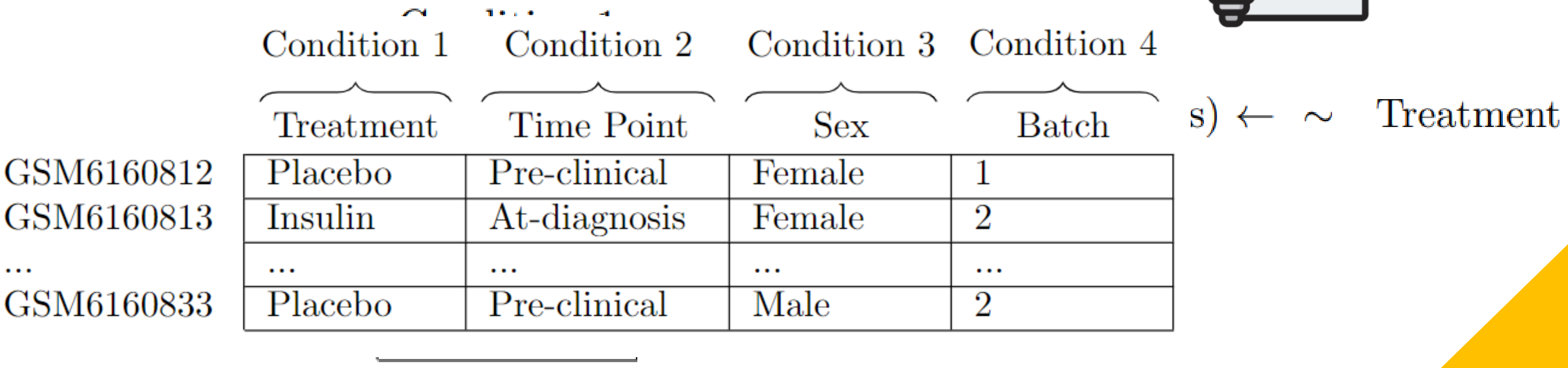
R: Information necessary for the DESeq Object

Count Matrix



	Sample 1	Sample 2		Sample M
	GSM6160812	GSM6160813		GSM6160833
Gene 1	88	10	...	102
Gene 2	55	33	...	34
...	...	...	...	...
Gene N	32	0	...	22

Study information → **Study design**

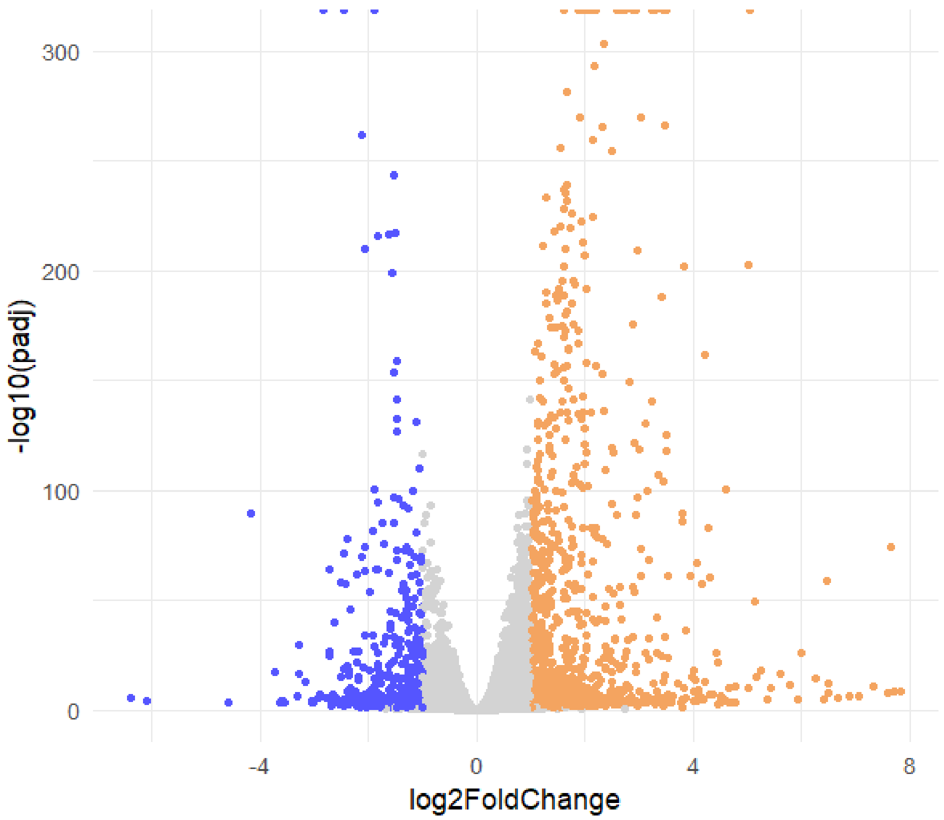


	Condition 1	Condition 2	Condition 3	Condition 4
	Treatment	Time Point	Sex	Batch
GSM6160812	Placebo	Pre-clinical	Female	1
GSM6160813	Insulin	At-diagnosis	Female	2
...	...	...	...	...
GSM6160833	Placebo	Pre-clinical	Male	2

s) ← ∼ Treatment

R: DE results

Volcano Plot



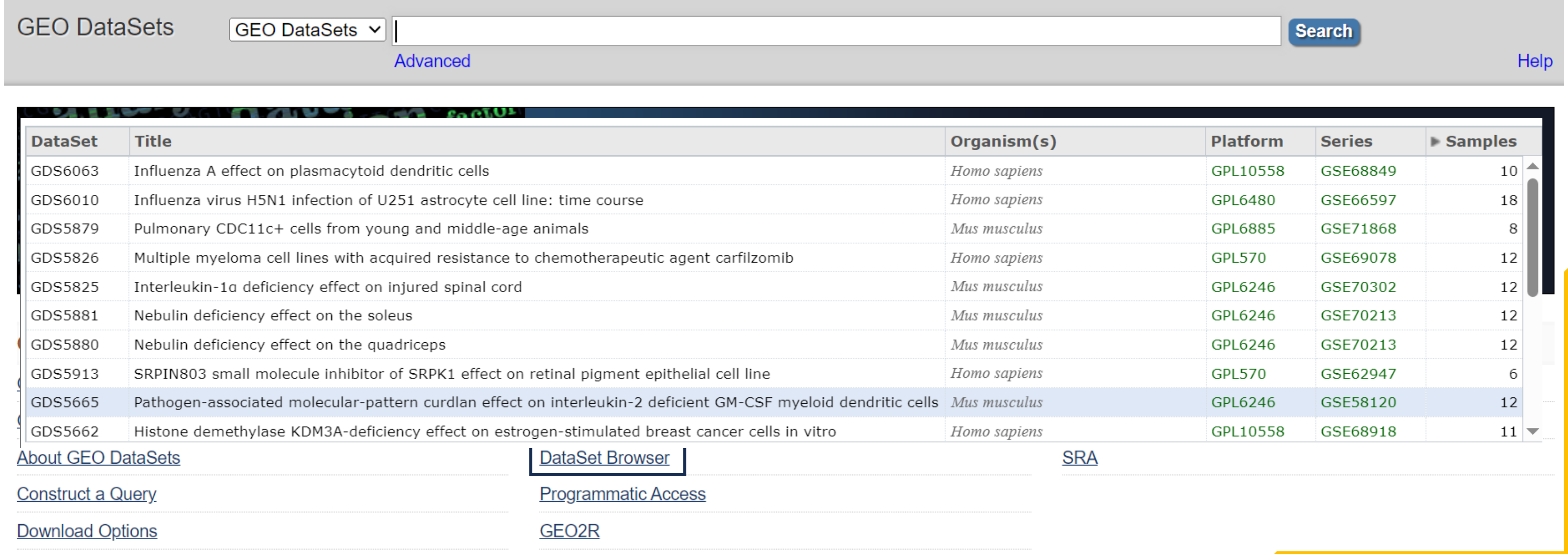
- Each dot represents a gene;
- Most statistically significant genes will be towards the top;
- Most upregulated genes will be towards the right;
- Most downregulated genes are towards the left;
- Fold change equal to 0 means that the expression does not change.

R: DESeq2

Let's get a count matrix from an RNA sequencing experiment from the National Center for Biotechnology Information repository!

Dataset: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE232750

[Home - GEO - NCBI (nih.gov)](https://www.ncbi.nlm.nih.gov/geo/)



DataSet	Title	Organism(s)	Platform	Series	▸ Samples
GDS6063	Influenza A effect on plasmacytoid dendritic cells	*Homo sapiens*	GPL10558	GSE68849	10
GDS6010	Influenza virus H5N1 infection of U251 astrocyte cell line: time course	*Homo sapiens*	GPL6480	GSE66597	18
GDS5879	Pulmonary CDC11c+ cells from young and middle-age animals	*Mus musculus*	GPL6885	GSE71868	8
GDS5826	Multiple myeloma cell lines with acquired resistance to chemotherapeutic agent carfilzomib	*Homo sapiens*	GPL570	GSE69078	12
GDS5825	Interleukin-1α deficiency effect on injured spinal cord	*Mus musculus*	GPL6246	GSE70302	12
GDS5881	Nebulin deficiency effect on the soleus	*Mus musculus*	GPL6246	GSE70213	12
GDS5880	Nebulin deficiency effect on the quadriceps	*Mus musculus*	GPL6246	GSE70213	12
GDS5913	SRPIN803 small molecule inhibitor of SRPK1 effect on retinal pigment epithelial cell line	*Homo sapiens*	GPL570	GSE62947	6
GDS5665	Pathogen-associated molecular-pattern curdlan effect on interleukin-2 deficient GM-CSF myeloid dendritic cells	*Mus musculus*	GPL6246	GSE58120	12
GDS5662	Histone demethylase KDM3A-deficiency effect on estrogen-stimulated breast cancer cells in vitro	*Homo sapiens*	GPL10558	GSE68918	11

About GEO DataSets

Construct a Query

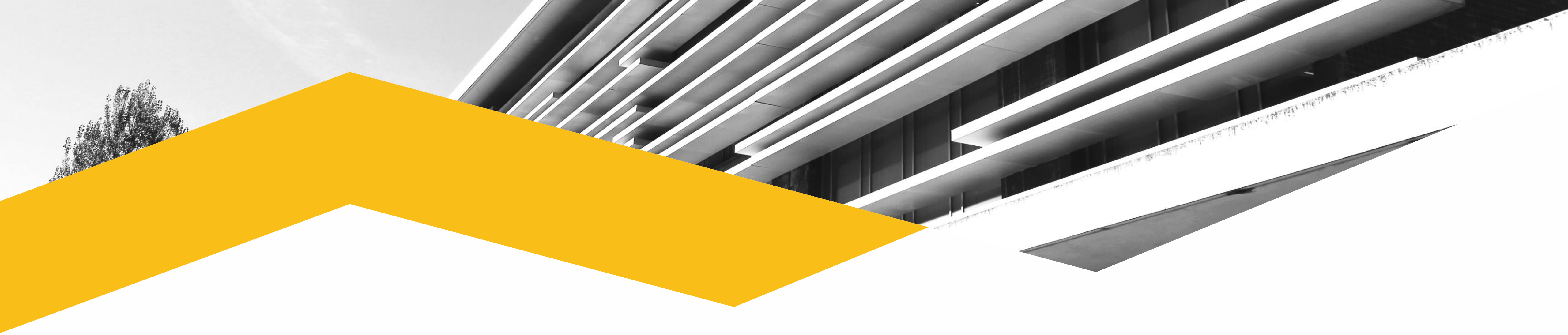
Download Options

DataSet Browser

Programmatic Access

GEO2R

SRA



Time to do some coding!

Let’s go back to the R visualization script

RStudio

Sciences course15th Bioinformatics in Health
Introduction to R

February 7th 2024