

# Curso básico de lenguaje R aplicado a las Ciencias Sociales



# Curso básico de lenguaje R aplicado a las Ciencias Sociales

Versión 2.0  
Abril 2019

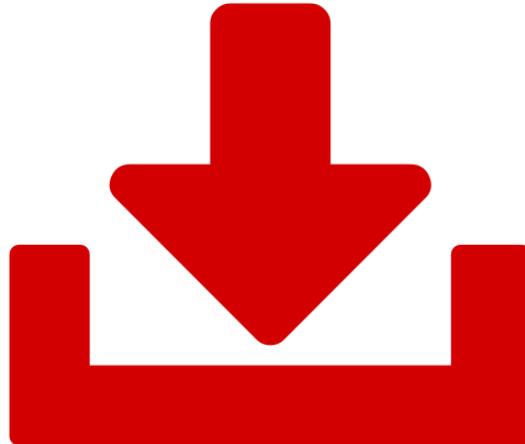
**Nicolás Robinson-Garcia**

<http://nrobinsongarcia.com>

Curso organizado por la  
**Oficina de Apoyo al  
Investigador** de la  
Universidad Internacional  
de la Rioja

10 de abril 2019  
Curso de 10 horas

**Todo el material del curso así como  
recursos adicionales están disponibles**





## ¿Quién soy yo y cuánto sé de este tema?

- ✓ Licenciado en Documentación
- ✓ Doctor en Ciencias Sociales
- ✓ Investigador Marie Skłodowska-Curie

Nicolas Robinson-Garcia

I am a social scientist specialized in bibliometrics. I work as a postdoctoral researcher at [Delft Institute of Applied Mathematics, TU Delft](#)

BIO CV

PUBLICATIONS

# Agenda

## PARTE 1

1. Presentación
2. ¿Qué es y por qué R?
3. Instalación y toma de contacto
4. Importación de datos

## PARTE 2

1. El uso de paquetes y librerías
2. Intro al paquete *ggplot2*
3. Análisis descriptivo y visualización
4. Puesta en común

# Parte 1

---

- Presentación
- ¿Qué es y por qué R?
- Instalación y toma de contacto
- Importación y exploración de datos



# Presentación

PARTE I-I



# ¿Qué aprenderéis en este curso?

## **Lenguaje estadístico R**

- Qué es y por qué merece la pena aprenderlo
- Los principios básicos para aprender a programar en R
- Paquetes de interés para la visualización de datos y para el análisis estadístico
- Recursos y consejos para buscar ayuda y seguir con la formación en R

**¿Qué NO aprenderéis en este curso?**

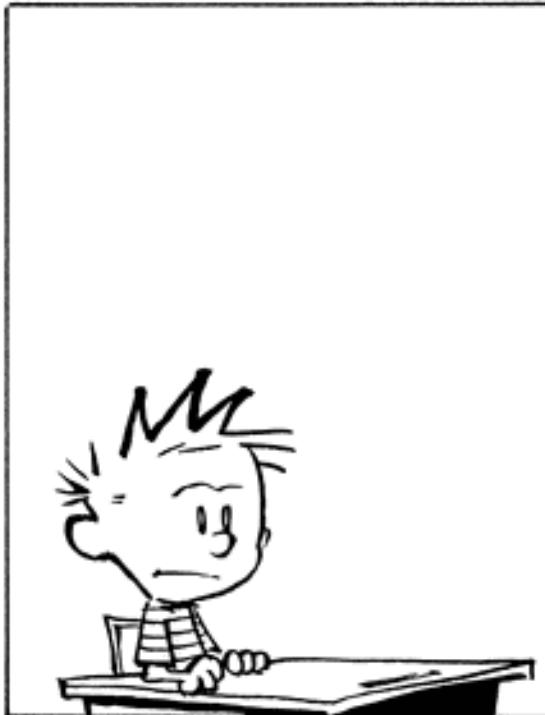
## **Análisis estadístico**

- Conocimiento pormenorizado de las técnicas de análisis
- La aplicación de técnicas de análisis avanzadas

# ¿Qué NO aprenderéis en este curso?

## **Lenguaje estadístico R**

- Poder realizar cualquier tipo de análisis estadístico con R
- Conocer todas las características del lenguaje
- Programar en R



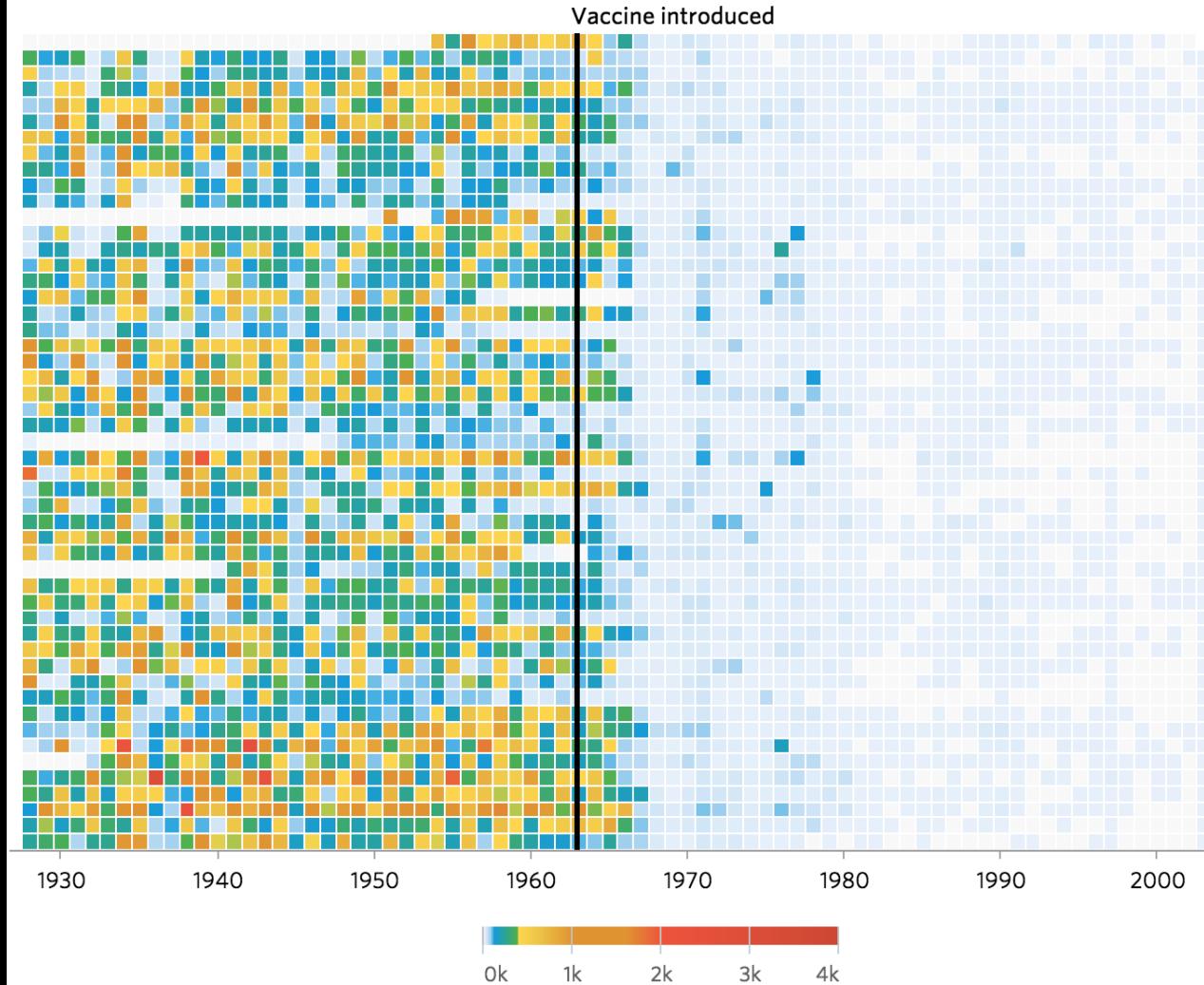
# Contenidos

## Programación en R

- Introducción al lenguaje de R
- Directrices básicas de comandos, software adicional y acceso a formación
- Paquetes básicos para la visualización y el análisis de datos
- Dinámicas de análisis descriptivo de datos: Importación, descriptivos y visualización de datos

# ¿Qué es R y por qué?

1. Por qué R
2. Un poco de historia
3. R y RStudio
4. Comandos básicos



# Carte Figurative des pertes successives en hommes de l'Armée Française dans la Campagne de Russie 1812-1813.

Dessiné par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; il somme de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie; le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chier, de Léger, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et de Maréchal Davout, qui avaient été détachés sur Minsk et Mohilow et se rejoignent vers Orsha en Witelsk, avaient toujours marché avec l'armée.

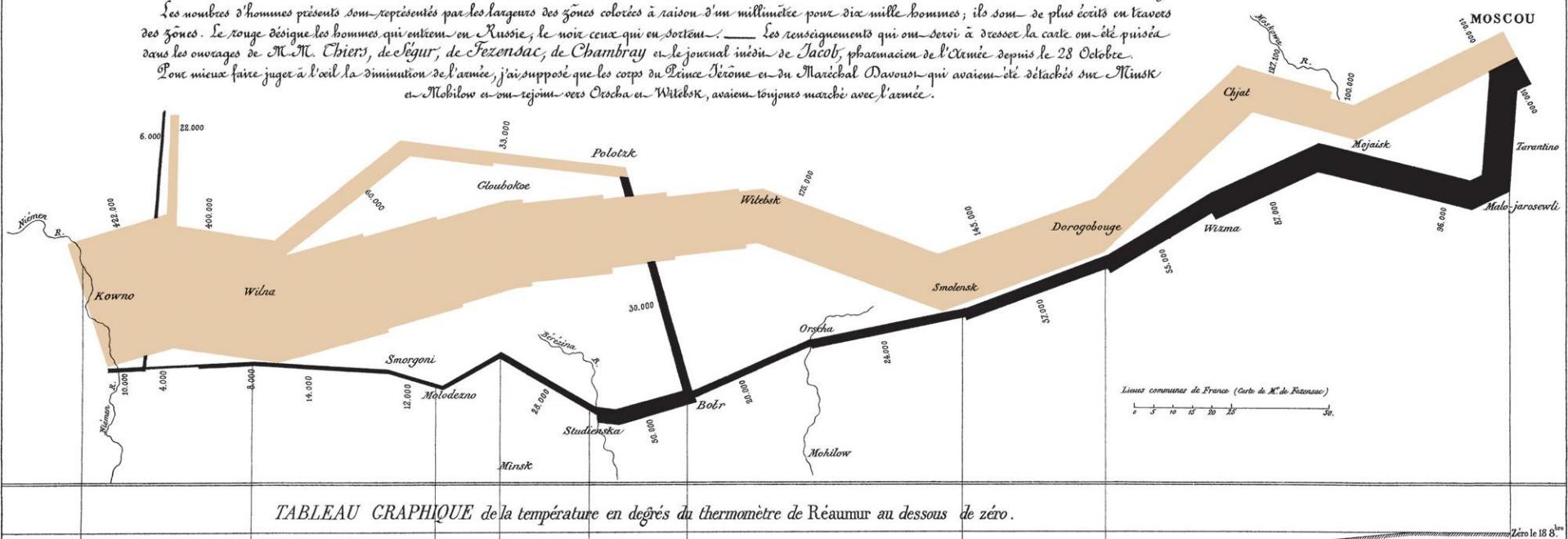


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Zéro le 18<sup>bre</sup>

# Razones por las que vale la pena aprender R

**1**

Gratis y abierto

**2**

Compatible con distintas plataformas

**3**

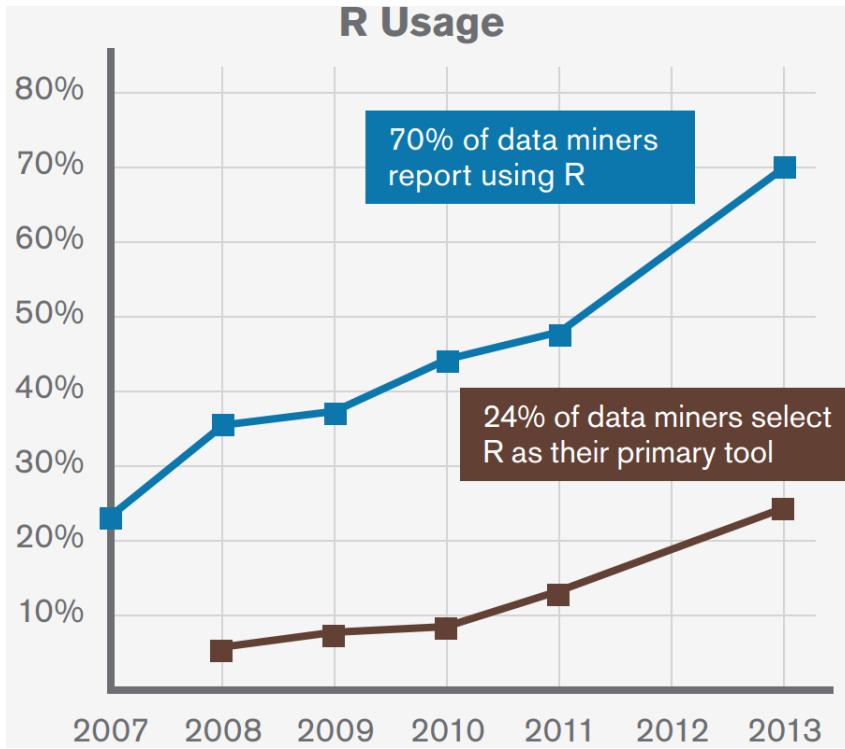
Reproducible

**4**

Universal y colaborativo

**5**

Flexible y en continuo crecimiento



## Alternativas:

- ✗ SPSS
- ✗ SAS
- ✗ Stata
- ✗ MatLab

1

Gratis y abierto

# Reading & Exporting Data in R



2

Compatible con distintas plataformas

## R es reproducible

Se dice que R es reproducible ya que al ser necesario escribir en código todos y cada uno de los pasos que se siguen en el tratamiento de los datos desde el momento en que se importan, cualquier persona con el mismo set de datos y el mismo código debería obtener los mismos resultados. Por ejemplo, todos tenemos el set de datos `mtcars` por defecto en R. Por tanto, si correlacionamos las cuatro primeras columnas del set de datos, todos deberíamos obtener los mismos resultados.

```
cor(mtcars[,1:4])
```

```
##          mpg         cyl        disp         hp
## mpg  1.0000000 -0.8521620 -0.8475514 -0.7761684
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475
## disp -0.8475514  0.9020329  1.0000000  0.7909486
## hp   -0.7761684  0.8324475  0.7909486  1.0000000
```

3

Reproducible

[Why GitHub?](#) ▾[Enterprise](#)[Explore](#) ▾[Marketplace](#)[Pricing](#) ▾[Sign in](#)[Sign up](#)[elrobin / introstatsconr](#) [Watch](#)

1

 [Star](#)

0

 [Fork](#)

1

 [Code](#) [Issues 0](#) [Pull requests 0](#) [Projects 0](#) [Insights](#)

Curso de introducción al análisis estadístico en Ciencias Sociales con R

 [32 commits](#) [1 branch](#) [0 releases](#) [1 contributor](#) [GPL-3.0](#)[Branch: master ▾](#)[New pull request](#)[Find File](#)[Clone or download ▾](#)

elrobin changed from Rmd to md

Latest commit b91200a 7 hours ago

 [introstat-v1](#)

changed from Rmd to md

7 hours ago

 [.gitignore](#)

First edit

a year ago

# 4

# Universal y colaborativo

# The tidyverse

## Components

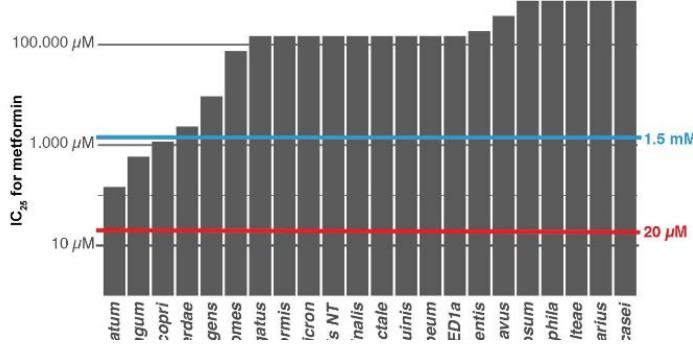


5

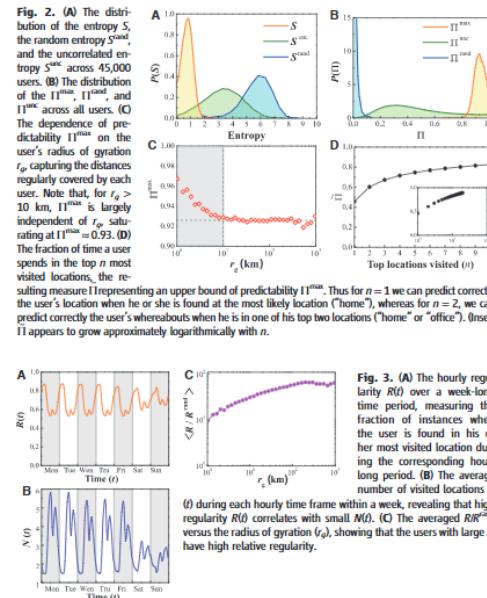
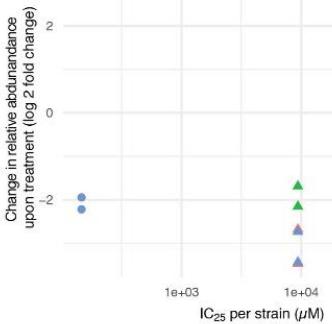
Flexible y en continuo crecimiento

# Algunos ejemplos

a



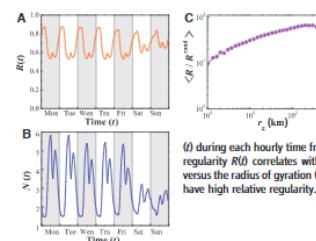
b



users' trajectories, a remarkable example of how their mobility pattern of the users hides an unexpectedly high degree of potential predictability. We have also determined the maximal predictability  $\Pi^{\text{max}}$  and the random predictability  $\Pi^{\text{rand}}$  extracted from  $S^{\text{unc}}$  and  $S^{\text{rand}}$ . As Fig. 2B shows, the result is strikingly different— $P(\Pi^{\text{max}})$  is extremely widely distributed and peaked at  $\Pi^{\text{max}} \approx 0.3$ , which indicates that, if we rely only on the heterogeneous spatial distribution, the predictability across the whole population is insignificant and varies widely from person to person. Similarly,  $P(\Pi^{\text{rand}})$  has a peak at  $\Pi^{\text{rand}} = 0$ , which suggests not only that  $\Pi^{\text{rand}}$  and  $\Pi^{\text{unc}}$  are ineffective as predictive tools, but also that a significant share of predictability is encoded in the temporal order of the visitation pattern.

How can we reconcile the wide variability in the observed travel distances, as captured by the fat-tailed  $P(r_g)$ , with the highly bounded predictability observed across the user population? To answer this, we measured the dependency  $\epsilon$  of  $\Pi^{\text{max}}$  on  $r_g$ , and found that, for  $r_g \geq 10$  km, predictability becomes largely independent of  $r_g$ , saturating at  $\Pi^{\text{max}} \approx 0.93$  (Fig. 2C). Therefore Fig. 2C explains the failure of our earlier hypothesis: Individuals with  $r_g \geq 100$  km, covering hundreds of kilometers on a regular basis, are just as predictable as those whose life is constrained to a  $\approx 10$ -km neighborhood, a saturation that lies behind the high predictability observed across the whole user base.

To determine how much of our predictability is really rooted in the visitation patterns of the top locations, we calculated the probability  $\Pi$  that, at a given moment, the user is in one of the top-most visited locations, where  $n = 2$  typically captures home and work. Thus,  $\Pi$  represents a upper bound for  $\Pi^{\text{max}}$ , as even if our predictive algorithm is 100% accurate, it can forecast the future location only when the user is found in one of the top  $n$  locations monitored by the algorithm. As Fig. 2D shows, the top two locations ( $n = 2$ )



# Algunos ejemplos

## Nature

### ► Software

---

Policy information about availability of computer code

#### 7. Software

Describe the software used to analyze the data in this study.

Data was processed using R (version 3.4.2) using scripts deposited at [https://git.embl.de/mkuhn/drug\\_impact\\_gut\\_bacteria](https://git.embl.de/mkuhn/drug_impact_gut_bacteria), Iris (doi:10.1038/nmicrobiol.2017.14)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

# Recursos

- Cursos Data Science - Johns Hopkins University
- R for Data Science - Hadley Wickham
- Introduction to Data Science - Rafael A. Irizarry
- R Journal - Q2 JCR
- Journal of Statistical Software - Q1 JCR

# Un poco de historia

1976

John Chambers desarrolla el lenguaje de programación S en Bell Labs.

1992

R. Ihaka y R. Gentleman comienzan a desarrollar R basado en S.

1995

M. Mächler se une y les anima a liberar el código en abierto

2000

Se publica la primera versión beta estable de R.

2009

New York Times se hace eco del éxito y crecimiento en el uso de R.

2011

Lanzamiento oficial de la primera versión de **RStudio**.

# Características de R

---

Lenguaje de programación sin interfaz

---

Curva de aprendizaje alta

---

Actualizaciones frecuentes

---

Estructura modular

---

<https://cloud.r-project.org/>

# RStudio



---

Interfaz para R libre y gratuita

---

Una amplia comunidad detrás

---

Actualizaciones constantes y recursos propios

---

Ofrece un entorno integrado

---

**<https://www.rstudio.com>**

## RStudio Connect



- ✓ Publicación de documentos en PDF
- ✓ Cuadernos de notas RMarkdown
- ✓ Creación de presentaciones y páginas web
- ✓ Creación de aplicaciones
- ✓ Control de versiones - GIT

No es solo una interfaz para programar

# ¿Cómo es RStudio?

The screenshot shows the RStudio interface with the following components:

- Top Bar:** RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Untitled1, Insert, Run, Addins.
- Code Editor (Untitled1):** Displays R Markdown code. Lines 1-17 show the template for a new notebook, and lines 18-20 show a plot of the 'cars' dataset.
- Console:** Shows the standard R license message.
- Environment:** Global Environment, showing an empty environment.
- File Browser:** Home directory listing files and folders.

```
1 ---  
2 title: "R Notebook"  
3 output: html_notebook  
4 ---  
5  
6 This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code  
within the notebook, the results appear beneath the code.  
7  
8 Try executing this chunk by clicking the *Run* button within the chunk or by placing  
your cursor inside it and pressing *Ctrl+Shift+Enter*.  
9  
10 ``{r}  
11 plot(cars)  
12 ...  
13  
14 Add a new chunk by clicking the *Insert chunk* button on the toolbar or by pressing  
*Ctrl+Alt+I*.  
15  
16 When you save the notebook, an HTML file containing the code and output will be saved  
alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML  
file).  
17  
4:1 R Notebook
```

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'licence()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

Name	Size	Modified
.Rhistory	10.6 KB	Apr 4, 2018, 7:22 AM
bajaautonomoSS.pdf	132.9 KB	Aug 15, 2015, 7:12 AM
drive-download-20161123T181038Z		
Facebook		
GitHub		
IBM		
Kutools for Excel		
Mi música		
Mis archivos de origen de datos		
Mis imágenes		
Mis videos		



# Zona de trabajo

- Scripts
- Notas
- Visualización de tablas

The screenshot shows the RStudio interface with a red box highlighting the left pane where code is written. A large red number '1' is overlaid on the code editor area.

RStudio interface components visible:

- File**, **Edit**, **Code**, **View**, **Plots**, **Session**, **Build**, **Debug**, **Profile**, **Tools**, **Help** menu bar.
- Toolbar with various icons.
- Untitled1** tab in the code editor.
- Code editor containing R Markdown code, including a plot chunk (line 11).
- Environment**, **History**, **Connections** tabs in the top right.
- Global Environment panel showing "Environment is empty".
- File, Plots, Packages, Help, Viewer tabs in the bottom right.
- File browser showing local files like .Rhistory, bajautonomoSS.pdf, etc.
- Console and Terminal tabs at the bottom.
- Console output showing R's license information.



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Untitled1

```
1 ---  
2 title: "R Notebook"  
3 output: html_notebook  
4 |---  
5  
6 This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code  
within the notebook, the results appear beneath the code.  
7  
8 Try executing this chunk by clicking the *Run* button within the chunk or by placing  
your cursor inside it and pressing *Ctrl+Shift+Enter*.  
9  
10 ``{r}  
11 plot(cars)  
12 ...  
13  
14 Add a new chunk by clicking the *Insert chunk* button on the toolbar or by pressing  
*Ctrl+Alt+I*.  
15  
16 When you save the notebook, an HTML file containing the code and output will be saved  
alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML  
file).  
17  
4:1 R Notebook
```

Console Terminal

```
~/  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'licence()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

# Zona de medio

- Objetos cargados
- Historial
- Conexiones externas

Project: (None)

Environment History Connections

Import Dataset

Global Environment

Environment is empty

2

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

Name	Size	Modified
.Rhistory	10.6 KB	Apr 4, 2018, 7:22 AM
bajaautonomoSS.pdf	132.9 KB	Aug 15, 2015, 7:12 AM
drive-download-20161123T181038Z		
Facebook		
Github		
IBM		
Kutools for Excel		
Mi música		
Mis archivos de origen de datos		
Mis imágenes		
Mis videos		



# Zona de ejecución

- Consola

The screenshot shows the RStudio interface with the following components:

- R Markdown Editor:** The main pane displays an R Markdown notebook titled "R Notebook". It contains code chunks and their corresponding output. A red box highlights the "Console" tab in the bottom-left corner.
- Global Environment:** The top-right pane shows the global environment, which is currently empty.
- File Explorer:** The bottom-right pane shows a file tree with various files and folders.
- Console Output:** The "Console" tab in the bottom-left pane shows the standard R startup message and help information. A large red number "3" is overlaid on this area.

```
1 ---  
2 title: "R Notebook"  
3 output: html_notebook  
4 |---  
5  
6 This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code  
within the notebook, the results appear beneath the code.  
7  
8 Try executing this chunk by clicking the *Run* button within the chunk or by placing  
your cursor inside it and pressing *Ctrl+Shift+Enter*.  
9  
10 ``{r}  
11 plot(cars)  
12 ...  
13  
14 Add a new chunk by clicking the *Insert chunk* button on the toolbar or by pressing  
*Ctrl+Alt+I*.  
15  
16 When you save the notebook, an HTML file containing the code and output will be saved  
alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML  
file).  
17  
41 R Notebook
```

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'licence()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

> |



RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Untitled1

```
1 ---  
2 title: "R Notebook"  
3 output: html_notebook  
4 |---  
5  
6 This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.  
7  
8 Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.  
9  
10 ``{r}  
11 plot(cars)  
12 ...  
13  
14 Add a new chunk by clicking the *Insert chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.  
15  
16 When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).  
17  
4:1 R Notebook
```

Console Terminal

```
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'licence()' or 'licence()' for distribution details.
```

```
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.
```

```
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```

Project: (None)

Environment History Connections

Global Environment

Environment is empty

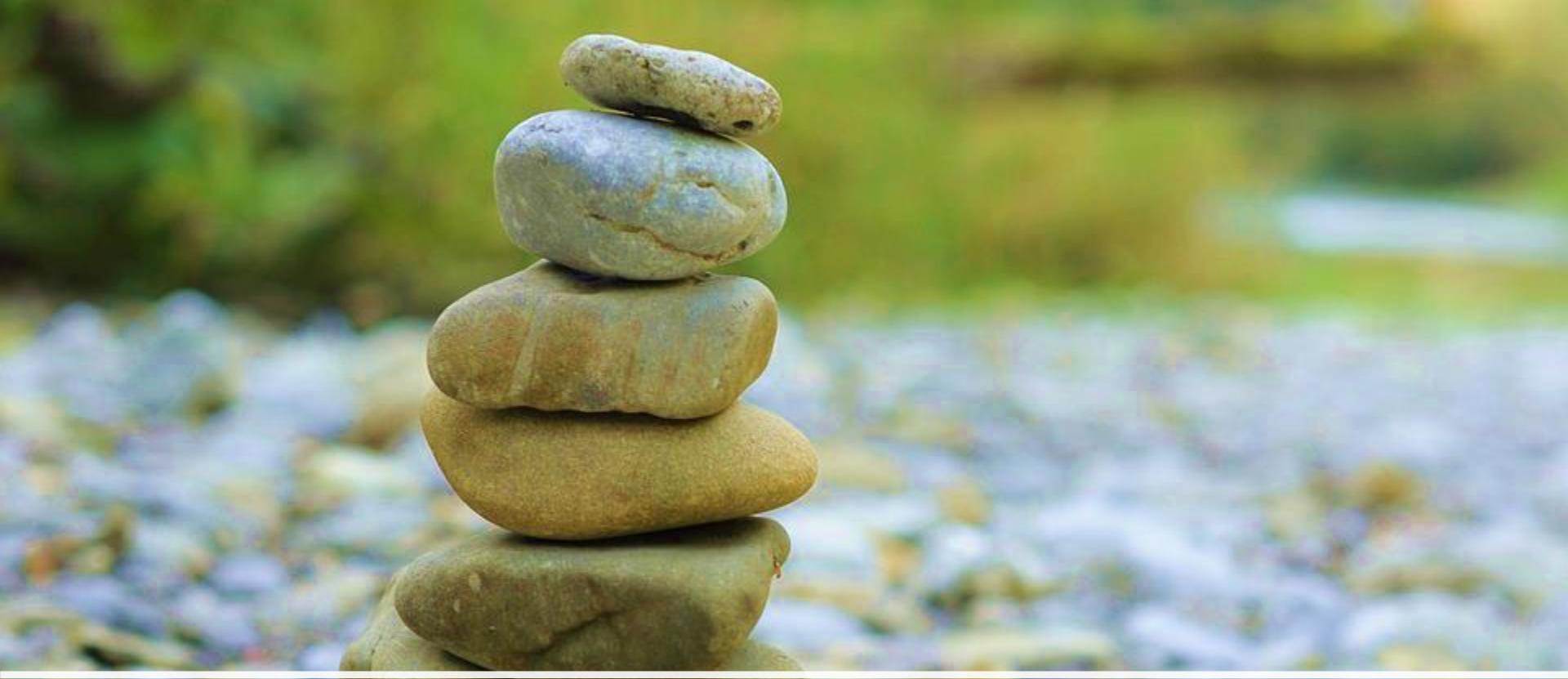
4

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

Name	Size	Modified
.Rhistory	10.6 KB	Apr 4, 2018, 7:22 AM
bajaautonomoSS.pdf	132.9 KB	Aug 15, 2015, 7:12 PM
drive-download-20161123T1148Z		
Facebook		
GitHub		
IBM		
Kutools for Excel		
Mi música		
Mis archivos de origen de datos		
Mis imágenes		
Mis videos		



Conoce bien las reglas, para poder romperlas de forma efectiva

# COMANDOS BÁSICOS

Todo es un objeto...

```
> presentacion <- "Buenos días mundo"  
> presentacion  
[1] "Buenos días mundo"  
> |
```

... una frase

# COMANDOS BÁSICOS

Todo es un objeto...

```
> suma <- 5+8  
> suma  
[1] 13  
|
```

... una operación matemática

# COMANDOS BÁSICOS

# ELEMENTOS BÁSICOS

Todo es un objeto...

```
[1] 13  
> lista_de_compra <- c("tomates", "lechuga", "pimientos", "patatas", "manzanas")  
> lista_de_compra  
[1] "tomates"    "lechuga"     "pimientos"  "patatas"    "manzanas"  
> |
```

... una lista

# COMANDOS BÁSICOS

## ELEMENTOS BÁSICOS

Todo es un objeto...

```
> vector <- c(3, 1, 2, 5, 10)
> vector
[1] 3 1 2 5 10
>
```

... un vector

# COMANDOS BÁSICOS

## ELEMENTOS BÁSICOS

Todo es un objeto...

```
> matriz <- matrix(1:6, 2, 2)
> matriz
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> |
```

... una matriz

# COMANDOS BÁSICOS

## ELEMENTOS BÁSICOS

Todo es un objeto...

```
> data <- as.data.frame(cbind(vector, lista_de_compra))
> data
  vector lista_de_compra
1      3      tomates
2      1      lechuga
3      2      pimientos
4      5      patatas
5     10      manzanas
.
```

... un data frame

# Diferencias entre data frame y matriz

En una matriz todas las columnas tienen el mismo tipo de datos, mientras que en un data frame se pueden incluir distintos tipos de datos (e.g., numérico, ordinal, etc.)

**Normalmente trabajaremos con data frames**

# COMANDOS BÁSICOS

```
[1] 13  
> lista_de_compra <- c("tomates", "lechuga", "pimientos", "patatas", "manzanas")  
> lista_de_compra  
[1] "tomates"   "lechuga"    "pimientos" "patatas"   "manzanas"  
> |
```

```
> presentacion <- "Buenos dias mundo"  
> presentacion  
[1] "Buenos días mundo"  
> |
```

# COMANDOS BÁSICOS

Los elementos entrecomillados se consideran valores de tipo textual.

Si no se trata de números, y no está entrecomillado, se considera un objeto o función

```
> data <- as.data.frame(cbind(vector, lista_de_compra))  
> data
```

```
> vector <- c(3, 1, 2, 5, 10)  
> vector
```

# Algunas funciones

```
> data <- as.data.frame(cbind(vector, lista_de_compra))  
> data  
  vector lista_de_compra  
1      3      tomates  
2      1      lechuga  
3      2      pimientos  
4      5      patatas  
5     10      manzanas
```

# Algunas funciones

```
> data <- as.data.frame(cbind(vector, lista_de_compra))  
> data  
  vector lista_de_compra  
1      3      tomates  
2      1      lechuga  
3      2      pimientos  
4      5      patatas  
5     10      manzanas
```

# Algunas funciones

```
> data <- as.data.frame(cbind(vector, lista_de_compra))  
> data  
vector lista_de_compra  
1      3          tomates  
2      1          lechuga  
3      2          pimientos  
4      5          patatas  
5     10          manzanas
```

1. **cbind()** - Une una serie de objetos en columnas  
**rbind()** - Une filas

2. **as.data.frame()** -  
Convierte el objeto en un data frame

# Kit de ayuda

1. Información sobre funciones y ayuda dentro de RStudio

>?cbind()

>??cbind

1. Ayuda en la web

r how to [función a realizar]

# Kit de ayuda

## Stackoverflow El lugar para preguntarle a la comunidad

The screenshot shows a Stack Overflow question page. At the top, there's a navigation bar with a logo, 'Questions' (which is underlined), 'Developer Jobs', 'Tags', 'Users', and a search bar. The main title of the question is 'Difference between data frame and matrix indexing'. Below the title, there's a voting section with upvote (triangle), downvote (triangle), and star icons, followed by a '0' indicating zero votes. The question text asks about the difference between reading integers from a file into a data frame versus a matrix. It includes two code snippets: one for a matrix showing output 'V1 0' and another for a data frame showing output '[1] 0'. The question ends with a call for explanation. Below the question are three tags: 'r', 'matrix', and 'dataframe'.

Difference between data frame and matrix indexing

▲ 0 ▾ ★

I have a text file of integers which I've been reading into R and storing as a data frame for the time being. However, coercing it to a matrix it (say `y` , using `as.matrix()` ) doesn't seem to be the same as the matrix I created ( `x` ). Namely, if I look at a single entry I get different output

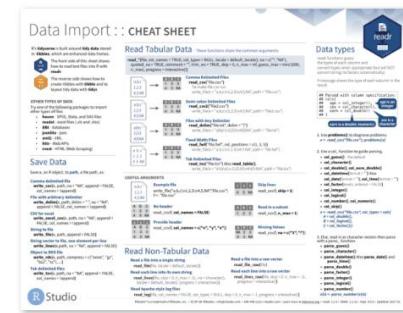
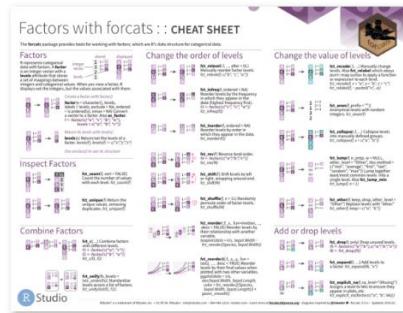
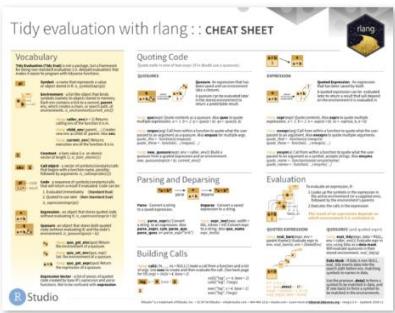
```
> y[1,1]
V1
0
```

as opposed to

```
> x[1,1]
[1] 0
```

Can anyone explain the difference?

r matrix dataframe



<https://www.rstudio.com/resources/cheatsheets/>



Comunidad Hispano R

La asociación de usuarios de R de España

¿QUIÉNES SOMOS? HAZTE SOCIO/A EMPLEO FORMACIÓN JORNADAS GRU

## La lista de correo R-help-es

7 marzo, 2016 / Pedro Concejero

R-hispano y en general la comunidad R en español mantiene una lista correo-e para ayuda sobre este lenguaje y sus aplicaciones.

Para suscribirte a la lista R-Help-es debes hacerlo desde  
<https://stat.ethz.ch/mailman/listinfo/r-help-es>.

## Ayuda en español

- Lista de correos muy activa
- Grupos locales de apoyo
- Anuncio de cursos formativos

<http://r-es.org>

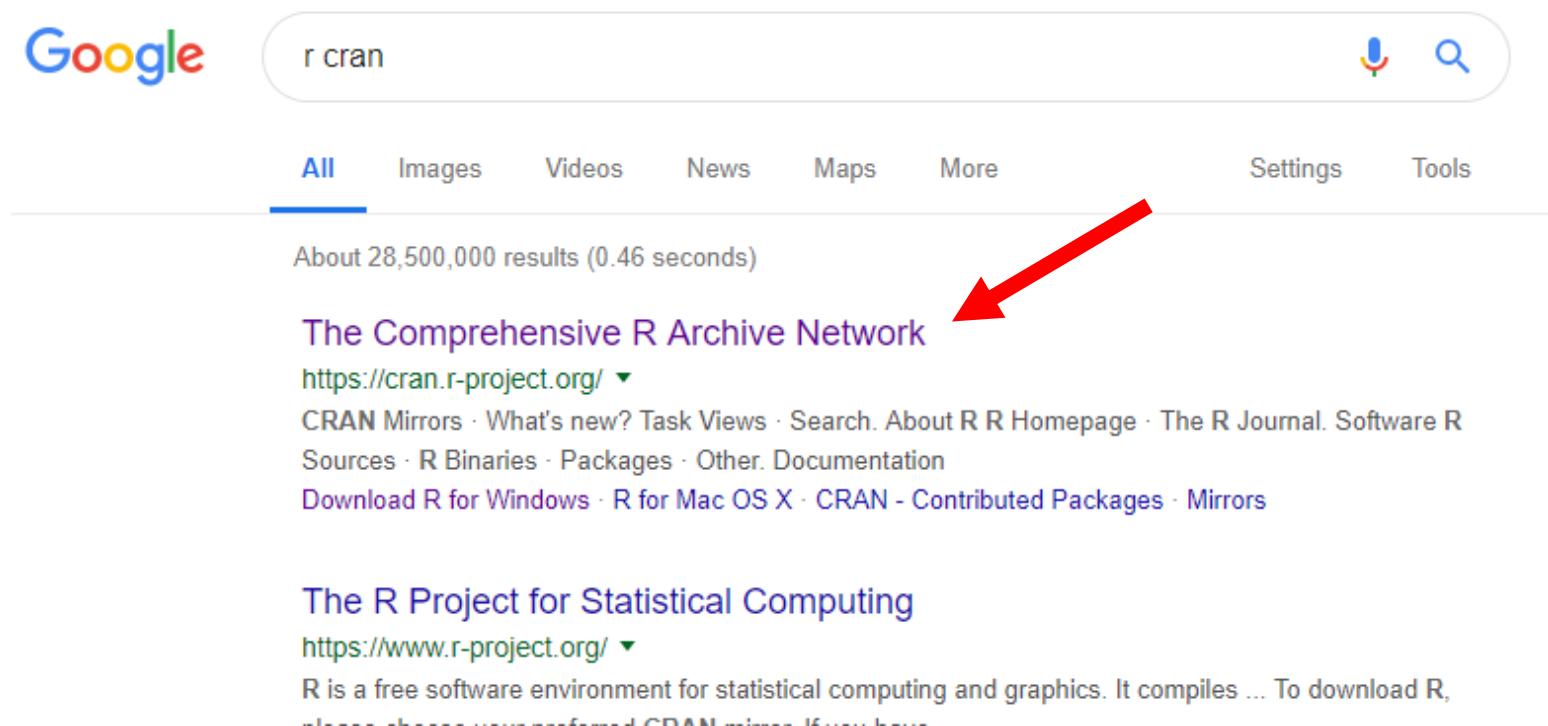
# Instalación y toma de contacto

1. R CRAN y versiones
2. RStudio: proyectos y tipos de ficheros
3. Primeros pasos en el análisis de datos

**PARTE I-3**



# Para instalar R debemos introducir en nuestro buscador *r cran*



A screenshot of a Google search results page. The search bar at the top contains the query "r cran". Below the search bar, there are navigation links for "All", "Images", "Videos", "News", "Maps", and "More", with "All" being underlined. To the right of these are "Settings" and "Tools" links. A red arrow points from the text "The Comprehensive R Archive Network" down towards the first search result. The search results section displays the following information:

About 28,500,000 results (0.46 seconds)

**The Comprehensive R Archive Network**  
[https://cran.r-project.org/ ▾](https://cran.r-project.org/)

CRAN Mirrors · What's new? · Task Views · Search · About R · R Homepage · The R Journal · Software R · Sources · R Binaries · Packages · Other · Documentation · Download R for Windows · R for Mac OS X · CRAN - Contributed Packages · Mirrors

**The R Project for Statistical Computing**  
[https://www.r-project.org/ ▾](https://www.r-project.org/)

R is a free software environment for statistical computing and graphics. It compiles ... To download R, please choose your preferred CRAN mirror. If you have ...

# Descarga la opción correspondiente con vuestro sistema operativo



[CRAN](#)  
[Mirrors](#)  
[What's new?](#)  
[Task Views](#)  
[Search](#)

[About R](#)  
[R Homepage](#)  
[The R Journal](#)

[Software](#)  
[R Sources](#)  
[R Binaries](#)  
[Packages](#)  
[Other](#)

[Documentation](#)  
[Manuals](#)  
[FAQs](#)  
[Contributed](#)

The Comprehensive R Archive Network

### Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows** and **Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

### Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2018-03-15, Someone to Lean On) [R-3.4.4.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension packages

# Incluso si se trata de una actualización, pincha en la opción de instalar R por primera vez

## R for Windows

Subdirectories:

[base](#)

Binaries for base distribution. This is what you want to [install R for the first time](#).

[contrib](#)

Binaries of contributed CRAN packages (for R >= 2.13.x; managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.

[old\\_contrib](#)

Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.x; managed by Uwe Ligges).

[Rtools](#)

Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.



R-3.4.4 for Windows (32/64 bit)

[Download R 3.4.4 for Windows](#) (62 megabytes, 32/64 bit)

[Installation and other instructions](#)

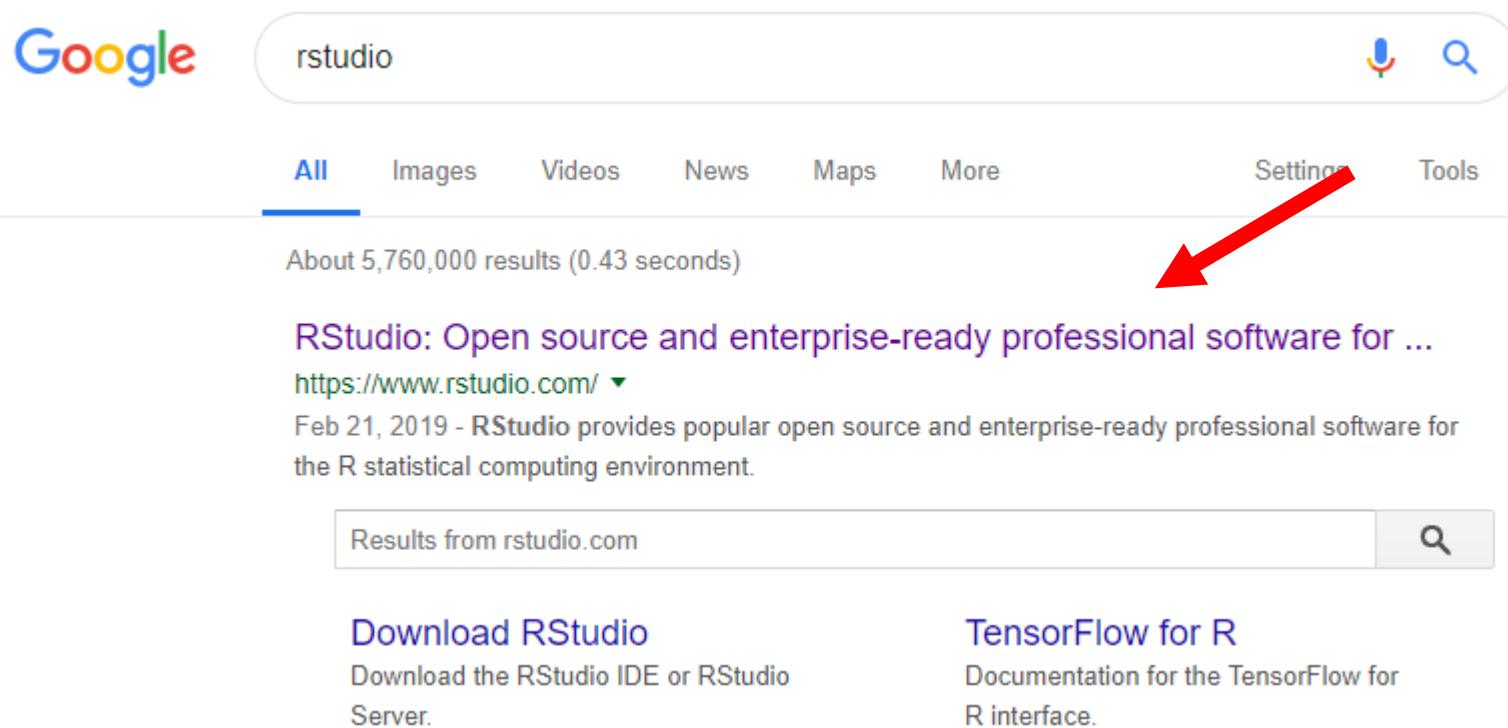
[New features in this version](#)

# Importante saber que....

- Al instalar R, es recomendable indicar que recuerde el número de versión
- Todos los años se lanza una nueva versión de R, normalmente en los meses de verano
- Para tener la última versión hay que volver a descargar e instalar R, al recordar el número de versión, el ordenador ignorará y eliminará las más antiguas

**Para instalar RStudio es  
esencial haber instalado  
previamente R**

# En primer lugar buscamos la web de RStudio desde nuestro buscador



A screenshot of a Google search results page. The search query "rstudio" is entered in the search bar. Below the search bar, there are filters for "All", "Images", "Videos", "News", "Maps", and "More". A red arrow points from the text "About 5,760,000 results (0.43 seconds)" down to the first search result. The result is for "RStudio: Open source and enterprise-ready professional software for ...". It includes a link to <https://www.rstudio.com/>, a snippet about RStudio providing software for the R statistical computing environment, and a "Results from rstudio.com" button with a magnifying glass icon.

Google

rstudio

All Images Videos News Maps More Settings Tools

About 5,760,000 results (0.43 seconds)

RStudio: Open source and enterprise-ready professional software for ...  
<https://www.rstudio.com/>  
Feb 21, 2019 - RStudio provides popular open source and enterprise-ready professional software for the R statistical computing environment.

Results from rstudio.com

Download RStudio  
Download the RStudio IDE or RStudio Server.

TensorFlow for R  
Documentation for the TensorFlow for R interface.

# RStudio

Open source and enterprise-ready  
professional software for R

[Download RStudio](#)

[Discover Shiny](#)

[Discover RStudio Package Manager](#)

[Discover RStudio Connect](#)



RStudio es un software libre, aunque cuenta con una versión de pago, la versión libre cuenta con todas las características necesarias, la única diferencia es que no cuenta con servicio de mantenimiento.

AT YOUR OWN RISK!

# Descarga la versión gratuita para escritorio de RStudio e instálala

<a href="#">RStudio Desktop Open Source License</a>	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server Pro + RStudio Connect Commercial License
FREE  <a href="#">DOWNLOAD</a>	\$995 per year <a href="#">BUY</a>	FREE <a href="#">DOWNLOAD</a>	\$9,995 per year <a href="#">DOWNLOAD</a>	\$29,995 per year <a href="#">TALK</a>

[Learn More](#)

[Learn More](#)

[Learn More](#)

[Learn More](#)

[Learn More](#)

Integrated  
Tools for R



Priority  
Support



# Ejercicio 1. Conociendo las herramientas

---

- **Objetivo:** Familiarizarse con la interfaz de RStudio y realizar operaciones básicas en la consola de R.
- **Tarea:** Inicia un proyecto en RStudio llamado curso-iniciacion. Utilizando la consola de RStudio, deberás crear tres objetos: un vector de números, un vector de nombres de personas y un data frame.
- **Planteamiento:** Utiliza la función `str()` para cada uno de los tres objetos creados. ¿Para qué sirve esta función? ¿Qué información ofrece sobre cada uno de los tres objetos creados?

# Ejercicio 2. Conociendo las herramientas

---

- **Objetivo:** Conocer los tipos de ficheros que crea RStudio
- **Tarea:** Basándote en el ejercicio anterior, crea un documento en R que pueda exportarse a formato html, Word o PDF en el que se explique cómo se ha creado cada objeto así como el análisis de la función str().
- **Planteamiento:** La utilidad de los documentos Rmarkdown radica en que nos permitirán guardar un *historial* de toda nuestra actividad, siendo capaces de integrar en un mismo software texto y comentarios, resultados y código.

# Importación de datos

1. Importación de datos
2. Comprobación de importación
3. Estructura de datos
4. Exploración

**PARTE I-4**

# Importación de set de datos

The screenshot shows the RStudio interface. On the left, there is a data grid with three columns: 'salario' (Salario actual), 'salini' (Salario inicial), and 'tiempemp' (Meses desde el contrato). The data consists of 12 rows of salary information. In the center, the 'Environment' tab is active, showing a list of datasets: 'Global', 'empl...', 'matr...', 'mtca...', 'values', 'Anto...', 'Isabe...', 'Lista...', 'Maria...', and 'RData'. A context menu is open over the 'values' dataset, with options: 'From Text (base)...', 'From Text (readr)...', 'From Excel...', 'From SPSS...', 'From SAS...', 'From Stata...', and 'From R...'. This menu is highlighted with a red box. To the right, the 'History' tab is active, displaying a list of recent R code and data frames. At the bottom, the 'Files' tab is active, showing a file tree with '.gitignore', '.RData', '.Rhistory', 'data', 'introstatscon.Rproj', and 'LICENSE' files.

# Importación de set de datos

Import Statistical Data

File/Url:

~/R/introstatsconr/data/EMPLEADOS.sav

Data Preview:

id Código de empleado	sexo Sexo	fechnac Fecha de nacimiento	educ Nivel educativo	catlab Categoría laboral	salario Salario actual	salini Salario inicial	tiempemp Meses desde el contrato	expprev Experiencia previa (meses)	minoría Clasificación de minoría
1	h	1952-02-03		15	3	57000	27000	98	144
2	h	1958-05-23		16	1	40200	18750	98	36
5	h	1955-02-09		15	1	45000	21000	98	138
6	h	1958-08-22		15	1	32100	13500	98	67
7	h	1956-04-26		15	1	36000	18750	98	114
15	h	1962-08-29		12	1	27300	13500	97	66
16	h	1964-11-17		12	1	40800	15000	97	24
17	h	1962-07-18		15	1	46000	14250	97	48

<  >

Previewing first 50 entries.

Import Options:

Name:

Model:

Format:   Open Data Viewer

Code Preview:

```
library(haven)
EMPLEADOS <- read_sav("data/EMPLEADOS.sav")
View(EMPLEADOS)
```

?

Reading data using haven

# Comandos básicos para explorar los datos

`str()`

- Muestra para cada variable el tipo de dato que contiene

`summary()`

- Indicadores descriptivos de cada variable (media, max., mín., etc.)

`complete.cases()`

- Indica para cada línea si hay datos incompletos o no

`head()`

- Muestra las seis primeras líneas del set de datos

`subset()`

- Filtrado condicional del set de datos

`nrow()`

- Cuenta el número de líneas del set de datos

`[,] $`

- Filtrado de datos de columnas o líneas específicas

# str()

```
> str(mtcars)
'data.frame': 32 obs. of 11 variables:
 $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num 160 160 108 258 360 ...
 $ hp : num 110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt : num 2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num 16.5 17 18.6 19.4 17 ...
 $ vs : num 0 0 1 1 0 1 0 1 1 1 ...
 $ am : num 1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num 4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num 4 4 1 1 2 1 4 2 2 4 ...
```

# summary()

```
> summary(mtcars)
```

	mpg	cyl	disp	hp	drat
Min.	10.40	Min. :4.000	Min. : 71.1	Min. : 52.0	Min. :2.760
1st Qu.	15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080
Median	19.20	Median :6.000	Median :196.3	Median :123.0	Median :3.695
Mean	20.09	Mean :6.188	Mean :230.7	Mean :146.7	Mean :3.597
3rd Qu.	22.80	3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:180.0	3rd Qu.:3.920
Max.	33.90	Max. :8.000	Max. :472.0	Max. :335.0	Max. :4.930
	wt	qsec	vs	am	
Min.	1.513	Min. :14.50	Min. : 0.0000	Min. :0.0000	
1st Qu.	2.581	1st Qu.:16.89	1st Qu.:0.0000	1st Qu.:0.0000	
Median	3.325	Median :17.71	Median :0.0000	Median :0.0000	
Mean	3.217	Mean :17.85	Mean : 0.4375	Mean :0.4062	
3rd Qu.	3.610	3rd Qu.:18.90	3rd Qu.:1.0000	3rd Qu.:1.0000	
Max.	5.424	Max. :22.90	Max. :1.0000	Max. :1.0000	
	gear	carb			
Min.	3.000	Min. :1.000			
1st Qu.	3.000	1st Qu.:2.000			
Median	4.000	Median :2.000			
Mean	3.688	Mean :2.812			
3rd Qu.	4.000	3rd Qu.:4.000			
Max.	5.000	Max. :8.000			

# complete.cases()

# head()

```
> head(mtcars)
```

		mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda	RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda	RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun	710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet	4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet	Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant		18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

# subset()

**subset(x, subset, select, ...)**

<b>x =</b>	Dataframe sobre el que se va a aplicar el filtro
<b>subset =</b>	Condición necesaria para filtrar
<b>select =</b>	Variables que queremos preservar

# subset()

```
> subset(x = mtcars, subset = cyl==6, select = c("mpg", "cyl"))
```

	mpg	cyl
Mazda RX4	21.0	6
Mazda RX4 Wag	21.0	6
Hornet 4 Drive	21.4	6
Valiant	18.1	6
Merc 280	19.2	6
Merc 280C	17.8	6
Ferrari Dino	19.7	6

# subset()

```
> subset(mtcars, mpg>20, c("mpg", "cyl"))
```

	mpg	cyl
Mazda RX4	21.0	6
Mazda RX4 Wag	21.0	6
Datsun 710	22.8	4
Hornet 4 Drive	21.4	6
Merc 240D	24.4	4
Merc 230	22.8	4
Fiat 128	32.4	4
Honda Civic	30.4	4
Toyota Corolla	33.9	4
Toyota Corona	21.5	4
Fiat X1-9	27.3	4
Porsche 914-2	26.0	4
Lotus Europa	30.4	4
Volvo 142E	21.4	4

# subset()

```
> subset(mtcars, mpg>10 & mpg<20)
```

		mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Hornet	Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant		18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster	360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc	280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc	280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc	450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc	450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc	450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac	Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln	Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler	Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Dodge	Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC	Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro	Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac	Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Ford	Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari	Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati	Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8

# nrow()

```
> nrow(mtcars)
[1] 32
> ncol(mtcars)
[1] 11
> length(mtcars)
[1] 11
```

[,] \$

## [filas, columnas]

```
> mtcars[1:4,]
      mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4     21.0   6 160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag 21.0   6 160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710    22.8   4 108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive 21.4   6 258 110 3.08 3.215 19.44  1  0    3    1
> mtcars[,1:4]
      mpg cyl disp  hp
Mazda RX4     21.0   6 160.0 110
Mazda RX4 Wag 21.0   6 160.0 110
Datsun 710    22.8   4 108.0  93
Hornet 4 Drive 21.4   6 258.0 110
Hornet Sportabout 18.7   8 360.0 175
Valiant       18.1   6 225.0 105
Duster 360    14.3   8 360.0 245
Merc 240D     24.4   4 146.7  62
Merc 230      22.8   4 140.8  95
Merc 280      19.2   6 167.6 123
Merc 280C     17.8   6 167.6 123
Merc 450SE    16.4   8 275.8 180
Merc 450SL    17.3   8 275.8 180
.. 1500 - 1500 1500 1500
```

[,] \$

## dataframe\$variable

```
> mtcars$cyl  
[1] 6 6 4 6 8 6 8 4 4 6 6 8 8 8 8 8 4 4 4 4 8 8 8 8 8 4 4 4 8 6 8 4
```

# Explorando los datos con visualizaciones

## plot()

- Gráfico de dispersion o lineal

## hist()

- Histograma

## barplot()

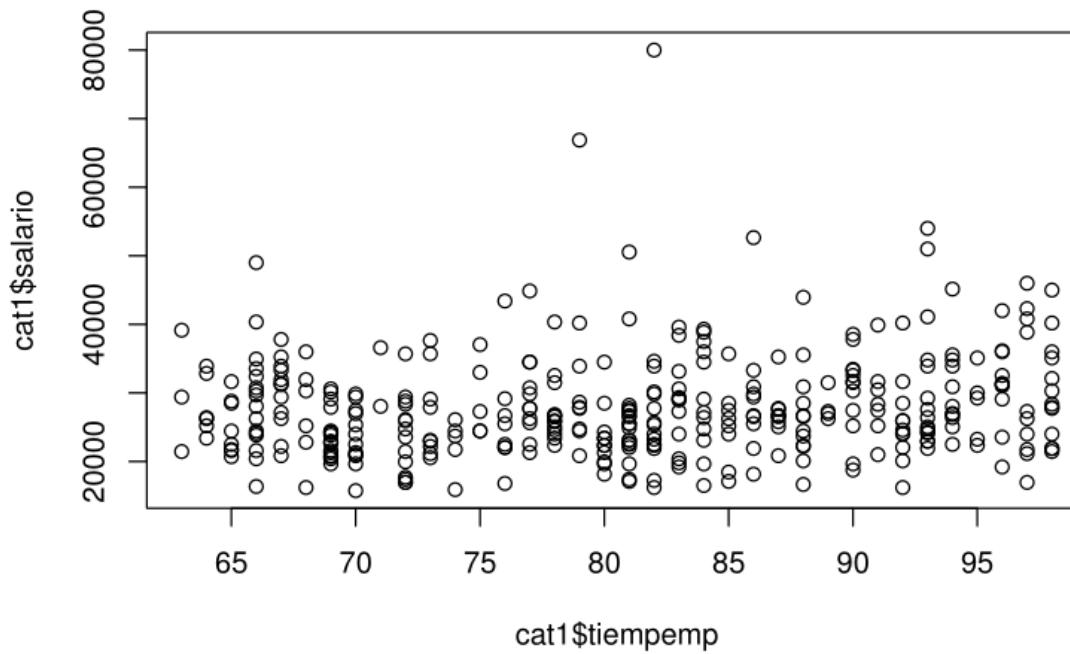
- Gráfico de barras

## boxplot()

- Gráfico de caja y bigotes

# plot()

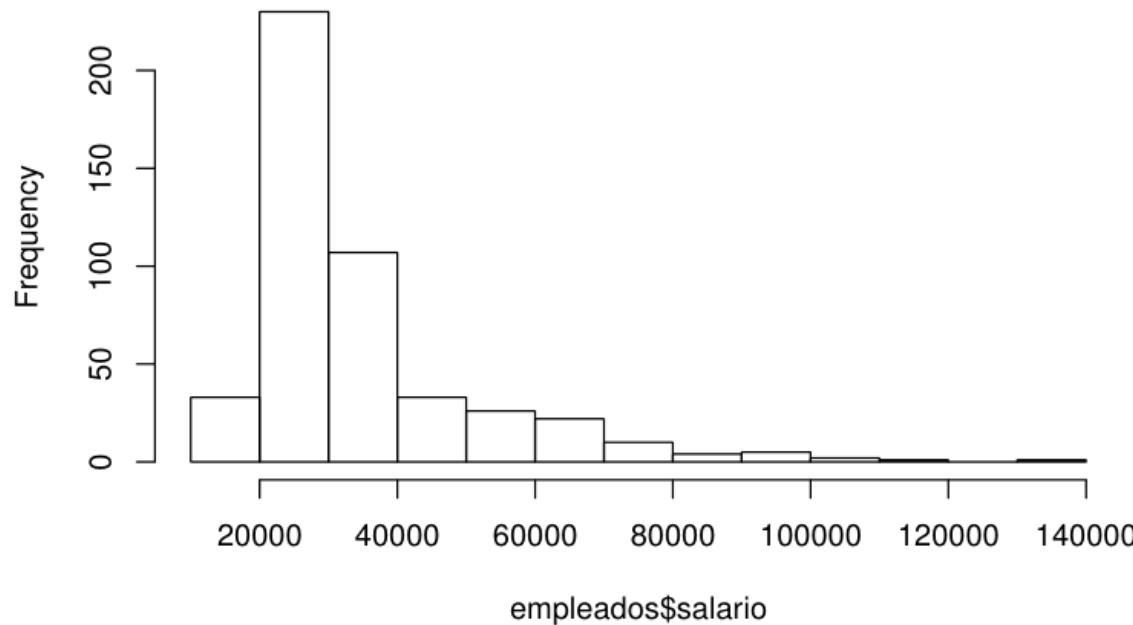
```
cat1 <- subset(empleados, catlab==1)
plot(cat1$tiempemp, cat1$salario)
```



# hist()

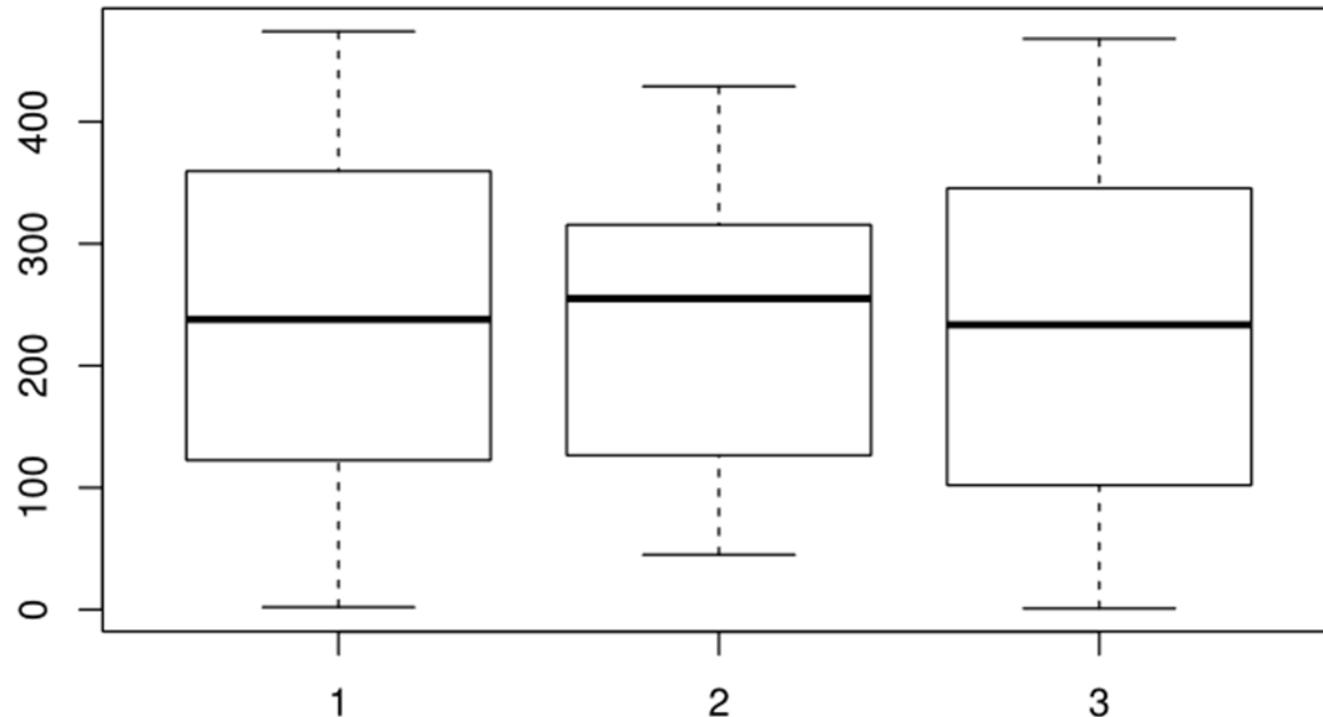
```
hist(empleados$salario)
```

Histogram of empleados\$salario



# boxplot()

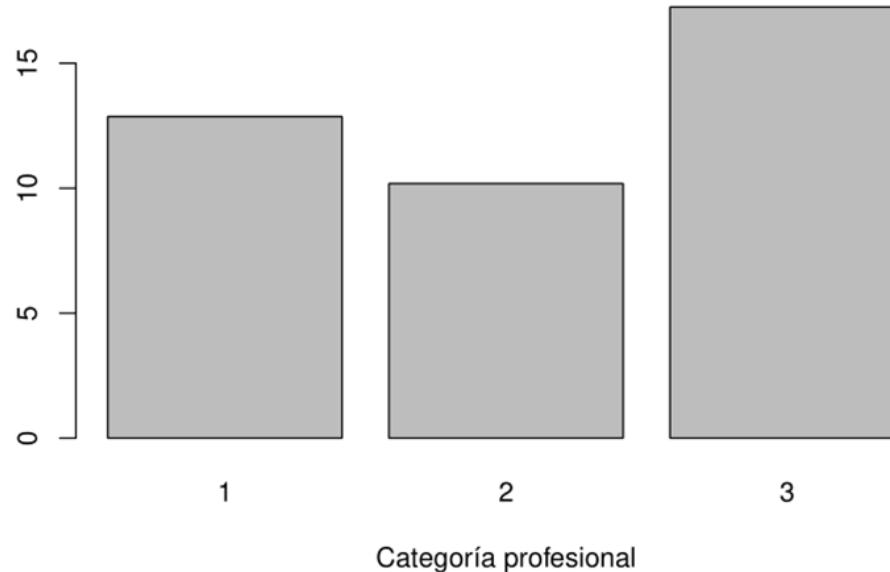
```
boxplot(id~catlab, data = empleados)
```



# barplot()

```
mean.educ <- aggregate(educ~catlab, data = empleados, mean)
mean.educ <- mean.educ[,2]
names(mean.educ) <- c("1", "2", "3")
barplot(mean.educ, main="Educación media",
        xlab = "Categoría profesional")
```

Educación media



# Ejercicio 3. Explorando datos

---

- **Objetivo:** Comienza a emplear las funciones básicas de R en un set de datos. Crea un documento donde toda la información esté disponible.
- **Tarea:** Utiliza el set de datos incluido por defecto en R llamado *USArrests*, aplica las funciones básicas tanto de exploración como de visualización para conocer el dataset.
- **Planteamiento:** Es evidente que no conseguiremos asimilar todas las funciones de R, pero combinando la información ofrecida aquí, la ayuda de R y ayuda en Internet, deberemos ser resolutivos y ser capaces de explorar de manera básica un set de datos.

# Parte 2

---

- El uso de paquetes y librerías
- Intro al paquete *ggplot2*
- Análisis descriptivo y visualización
- Puesta en común



# Paquetes y librerías

## 1. Instalación de paquetes

## 2. Dónde encontrarlos

## 3. Paquetes recomendados

**PARTE 2-I**

20 BEST LIBRARIES FOR DATA SCIENCE IN R			
	COMMITS	CONTRIBUTORS	FEATURES
DATA MANIPULATION	dplyr 4 354	136	<ul style="list-style-type: none"> <li>powerful library for data wrangling</li> <li>works with local data frames and remote database tables</li> <li>precise and simple command syntax</li> </ul>
	data.table 3 211	43	<ul style="list-style-type: none"> <li>quick aggregation of large data</li> <li>iconic flexible syntax and a wide suite of useful functions</li> <li>friendly file reader and parallel file writer</li> </ul>
	lubridate 1 427	45	<ul style="list-style-type: none"> <li>a set of functions to work with date and time format</li> <li>easy and fast parsing of date-time data</li> <li>expanded mathematical operations on time data</li> </ul>
	jsonlite 908	11	<ul style="list-style-type: none"> <li>robust and quick parsing JSON objects in R</li> <li>great tool for interacting with web APIs and building pipelines</li> <li>functions to stream, validate, and prettyify JSON data</li> </ul>
GRAPHIC DISPLAY	grid 3 903	133	<ul style="list-style-type: none"> <li>powerful implementation of the grammar of graphics visualization</li> <li>developed static graphics system</li> <li>takes care of plot specifications</li> </ul>
	gridExtra 299	8	<ul style="list-style-type: none"> <li>abilities to visualize correlation matrices and confidence intervals</li> <li>contains algorithms to do matrix reordering</li> <li>flexible appearance settings</li> </ul>
	lattice 132	0	<ul style="list-style-type: none"> <li>high-level visualization system</li> <li>emphasis on multivariate data</li> <li>efficiently copes with nonstandard requirements</li> </ul>
	grid 2 989	26	<ul style="list-style-type: none"> <li>rich features and plenty of available charts</li> <li>web-based toolbox for building visualizations</li> <li>abilities to make ggplot2 graphics interactive</li> </ul>
WEB INTERFACES	ggvis 2 159	21	<ul style="list-style-type: none"> <li>implementation of an interactive grammar of graphic</li> <li>incorporates shiny reactive programming model and dplyr grammar of data transformation</li> </ul>
	DT DataTables 1 919	21	<ul style="list-style-type: none"> <li>displays R matrices and data frames as interactive HTML tables</li> <li>creates sortable tables with a minimum of code</li> <li>many useful features and styling options for tables</li> </ul>
	rstudio	638	<ul style="list-style-type: none"> <li>interactive JS charts from R</li> <li>tools for creation, customization, and sharing</li> </ul>
	shiny 5 467	96	<ul style="list-style-type: none"> <li>transparent tool for easy dynamic report generation in R</li> <li>enables integration of R code into LaTeX, LyX, HTML, Markdown, AsciiDoc, and reStructuredText documents</li> </ul>
REPRODUCIBLE RESEARCH	knitr 2 297	56	<ul style="list-style-type: none"> <li>next generation implementation of R Markdown based on pandoc</li> <li>many static and dynamic output formats</li> <li>abilities to define new formats for custom publishing requirements</li> </ul>
	slidify 302	7	<ul style="list-style-type: none"> <li>generates reproducible HTML slides from r markdown</li> <li>allows embedded code chunks and mathematical formulas</li> <li>rich sharing and customizing opportunities</li> </ul>
	mlr 3 915	55	<ul style="list-style-type: none"> <li>extensible framework for classification, regression, survival analysis, and clustering</li> <li>easy extension mechanism through S3 inheritance</li> </ul>
	dmic XGBoost 3 188	259	<ul style="list-style-type: none"> <li>implementation of the Gradient Boosted Decision Trees algorithm</li> <li>reach tools for regression, classification, and ranking problems</li> <li>high speed and performance</li> </ul>
MACHINE LEARNING	caret 1 659	59	<ul style="list-style-type: none"> <li>many models for classification and regression</li> <li>powerful tools and algorithms for creating predictive models</li> </ul>
	gbm 731	26	<ul style="list-style-type: none"> <li>represents Generalized Boosted Regression Models</li> <li>includes plenty of regression methods</li> <li>tools variable selection and final stage precision modeling</li> </ul>
	Prophet 190	20	<ul style="list-style-type: none"> <li>high-quality forecasts for time series data</li> <li>manages data that has multiple seasonality with linear or non-linear growth</li> <li>robust to missing data, shifts in the trend, and large outliers</li> </ul>
	randomforest 56	0	<ul style="list-style-type: none"> <li>implements Breiman's random forest algorithm for classification and regression</li> <li>builds multiple decision trees and gives back the mean prediction of the individual trees</li> </ul>

Updated: December 2017

20 BEST LIBRARIES FOR DATA SCIENCE IN R			
	COMMITS	CONTRIBUTORS	FEATURES
DATA MANIPULATION	dplyr 4 354	136	<ul style="list-style-type: none"> <li>powerful library for data wrangling</li> <li>works with local data frames and remote database tables</li> <li>precise and simple command syntax</li> </ul>
	data.table 3 211	43	<ul style="list-style-type: none"> <li>quick aggregation of large data</li> <li>iconic flexible syntax and a wide suite of useful functions</li> <li>friendly file reader and parallel file writer</li> </ul>
	lubridate 1 427	45	<ul style="list-style-type: none"> <li>a set of functions to work with date and time format</li> <li>easy and fast parsing of date-time data</li> <li>expanded mathematical operations on time data</li> </ul>
	jsonlite 908	11	<ul style="list-style-type: none"> <li>robust and quick parsing JSON objects in R</li> <li>great tool for interacting with web APIs and building pipelines</li> <li>functions to stream, validate, and prettyify JSON data</li> </ul>
GRAPHIC DISPLAY	grid 3 903	133	<ul style="list-style-type: none"> <li>powerful implementation of the grammar of graphics visualization</li> <li>developed static graphics system</li> <li>takes care of plot specifications</li> </ul>
	gridExtra 299	8	<ul style="list-style-type: none"> <li>abilities to visualize correlation matrices and confidence intervals</li> <li>contains algorithms to do matrix reordering</li> <li>flexible appearance settings</li> </ul>
	lattice 132	0	<ul style="list-style-type: none"> <li>high-level visualization system</li> <li>emphasis on multivariate data</li> <li>efficiently copes with nonstandard requirements</li> </ul>
	grid 2 989	26	<ul style="list-style-type: none"> <li>rich features and plenty of available charts</li> <li>web-based toolbox for building visualizations</li> <li>abilities to make ggplot2 graphics interactive</li> </ul>
HTML WIDGETS	ggvis 2 159	21	<ul style="list-style-type: none"> <li>implementation of an interactive grammar of graphic</li> <li>incorporates shiny reactive programming model and dplyr grammar of data transformation</li> </ul>
	DT DataTables 1 919	21	<ul style="list-style-type: none"> <li>displays R matrices and data frames as interactive HTML tables</li> <li>creates sortable tables with a minimum of code</li> <li>many useful features and styling options for tables</li> </ul>
	rstudio 638	11	<ul style="list-style-type: none"> <li>interactive JS charts from R</li> <li>tools for creation, customization, and sharing</li> </ul>
	shiny 5 467	96	<ul style="list-style-type: none"> <li>transparent tool for easy dynamic report generation in R</li> <li>enables integration of R code into LaTeX, LyX, HTML, Markdown, AsciiDoc, and reStructuredText documents</li> </ul>
REPRODUCIBLE RESEARCH	knitr 2 297	56	<ul style="list-style-type: none"> <li>next generation implementation of R Markdown based on pandoc</li> <li>many static and dynamic output formats</li> <li>abilities to define new formats for custom publishing requirements</li> </ul>
	slidify 302	7	<ul style="list-style-type: none"> <li>generates reproducible HTML slides from r markdown</li> <li>allows embedded code chunks and mathematical formulas</li> <li>rich sharing and customizing opportunities</li> </ul>
	mlr 3 915	55	<ul style="list-style-type: none"> <li>extensible framework for classification, regression, survival analysis, and clustering</li> <li>easy extension mechanism through S3 inheritance</li> </ul>
	dmic XGBoost 3 188	259	<ul style="list-style-type: none"> <li>implementation of the Gradient Boosted Decision Trees algorithm</li> <li>reach tools for regression, classification, and ranking problems</li> <li>high speed and performance</li> </ul>
MACHINE LEARNING	caret 1 659	59	<ul style="list-style-type: none"> <li>many models for classification and regression</li> <li>powerful tools and algorithms for creating predictive models</li> </ul>
	gbm 731	26	<ul style="list-style-type: none"> <li>represents Generalized Boosted Regression Models</li> <li>includes plenty of regression methods</li> <li>tools variable selection and final stage precision modeling</li> </ul>
	Prophet 190	20	<ul style="list-style-type: none"> <li>high-quality forecasts for time series data</li> <li>manages data that has multiple seasonality with linear or non-linear growth</li> <li>robust to missing data, shifts in the trend, and large outliers</li> </ul>
	randomforest 56	0	<ul style="list-style-type: none"> <li>implements Breiman's random forest algorithm for classification and regression</li> <li>builds multiple decision trees and gives back the mean prediction of the individual trees</li> </ul>

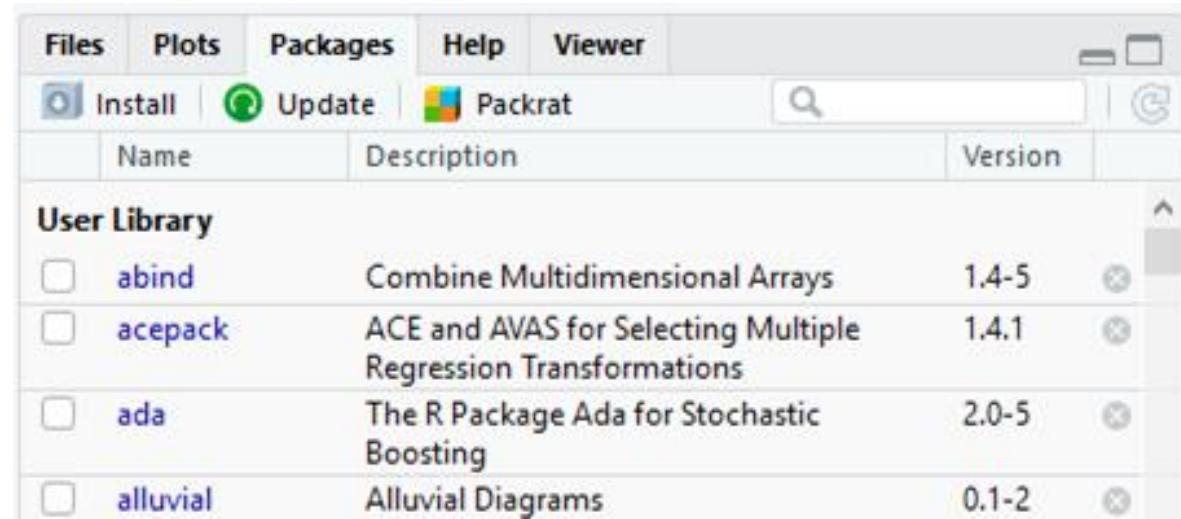
Created by ActiveWizards

Updated: December 2017

Created by ActiveWizards

# Acceso e instalación de paquetes

- Software descentralizado
- La comunidad crea sus propios paquetes que se van adhiriendo de manera modular
- Cada paquete tiene su propia versión y debe actualizarse



The screenshot shows the RStudio interface with the 'Packages' tab selected in the top navigation bar. Below the navigation bar, there are three buttons: 'Install' (blue icon), 'Update' (green icon), and 'Packrat' (yellow icon). A search bar is located to the right of these buttons. The main area displays a table titled 'User Library' with columns for Name, Description, and Version. The table lists several R packages:

Name	Description	Version	Actions
<b>User Library</b>			
<input type="checkbox"/> abind	Combine Multidimensional Arrays	1.4-5	
<input type="checkbox"/> acepack	ACE and AVAS for Selecting Multiple Regression Transformations	1.4.1	
<input type="checkbox"/> ada	The R Package Ada for Stochastic Boosting	2.0-5	
<input type="checkbox"/> alluvial	Alluvial Diagrams	0.1-2	

# R Package Discovery

## How do you currently discover and learn about R packages?

1027 responses

Email lists such as r-help, r-packages, or r-pkg-devel

15%

General search websites such as Google and Yahoo

57%

R-specific search websites such as METACRAN ([www.r-pkg.org](http://www.r-pkg.org)) or Rdocumentation (<https://www.rdocumentation.org/>)

11%

R packages built for search such as the sos package

2%

CRAN Task Views

22%

Your personal network, such as colleagues and professors

41%

Conferences, meet-ups, or seminars

24%

Books, textbooks, or journal articles (JSS, JOSS, R-Journal)

32%

Social media such as blogs, R-bloggers, Twitter, Slack, or GitHub contacts

79%

Other (send ideas to [@juliasilge](#) on Twitter!)

4%

# Acceso e instalación de paquetes

Normalmente, los paquetes finalizados se suben a **CRAN** -  
**<https://cran.r-project.org>**

Se pueden descargar directamente desde R con el siguiente comando:

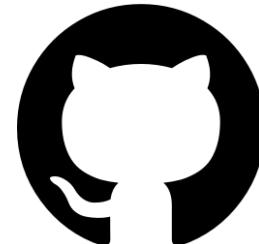
```
>install.packages("nombre_del_paquete")
```

# Acceso e instalación de paquetes

Sin embargo, muchas veces los paquetes están aún en desarrollo o estamos interesados en acceder a la última versión no testeada.

En muchas ocasiones podemos encontrar paquetes en la plataforma **GitHub**

```
>devtools::install_github("usuario/paquete")
```



# Algunos recursos de interés

Listado de paquetes creados por Hadley Wickham, el gurú detrás de RStudio. Incluye paquetes de procesamiento y limpieza de datos, visualización, etc.

## Tidyverse

[Packages](#)   [Articles](#)   [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.

# Algunos recursos de interés

Listado de paquetes mantenidos por la comunidad científica. Desde extracción de datos, procesamiento del lenguaje natural, visualización o de publicación de datos científicos

[ABOUT](#)[BLOG](#)[PACKAGES](#)[COMMUNITY](#)[DISCUSS](#)A blue-tinted photograph showing three people, two men and one woman, smiling and working on laptops. They appear to be in a collaborative environment, possibly a library or a computer lab.

Transforming science through  
open data and software



# Algunos recursos de interés

Familia de paquetes de R desarrollados por el Instituto de Estudios Cuantativos de las Ciencias Sociales de la Universidad de Harvard.



Everyone's Statistical Software

Zelig is an easy-to-use, free, open source, general purpose statistics program for estimating, interpreting, and presenting results from any statistical method. Zelig turns the power of R, with thousands of open source packages — but with free ranging syntax, diverse examples, and documentation written for different audiences — into the same three commands and consistent documentation for every method. Zelig uses R code from many researchers, making it "everyone's statistical software." We hope it becomes everyone's statistical software for applications too, as we designed it so anyone can use it or add their methods to it. We aim for Zelig to be the best way to do analysis, prepare replication files, learn new methods, or teach.

# Ejercicio 4. Instalación de librerías

---

- **Objetivo:** Comenzar a utilizar funciones de una librería externa a R
- **Tarea:** Instala la librería de análisis de datos [zeligverse](#). Una vez instalada, deberás activarla para comenzar a utilizar sus funciones con el comando `library()`
- **Planteamiento:** Crea un documento Rmarkdown en el que documentes el proceso de instalación. Entra en la web del proyecto Zelig y reproduce el primer ejemplo que muestran.

# GGPLOT2

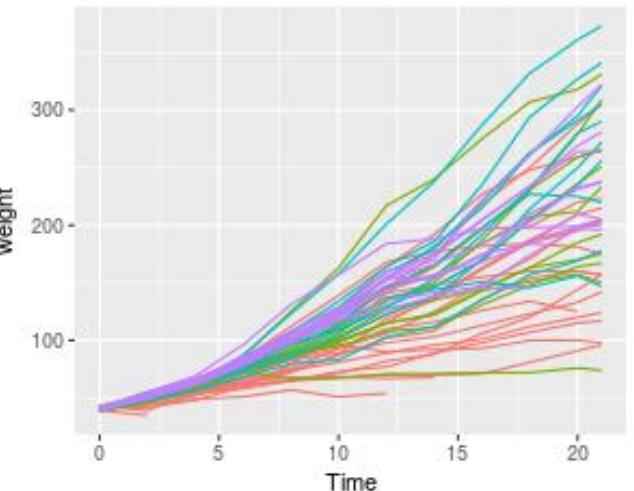
1. Gramática de ggplot2

2. Tipos de gráficos

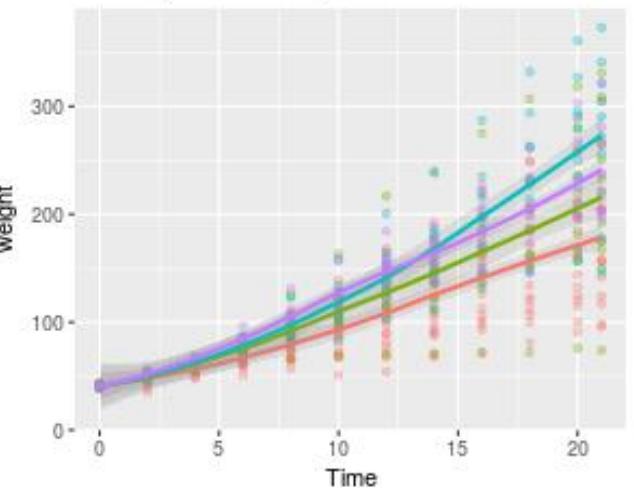
3. Personalización de etiquetas y temas

PARTE 2-2

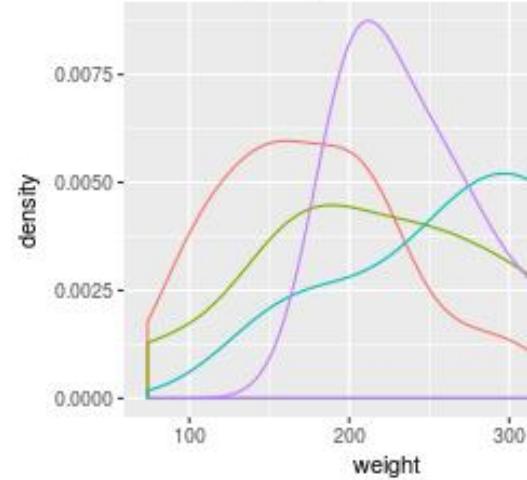
Growth curve for individual chicks



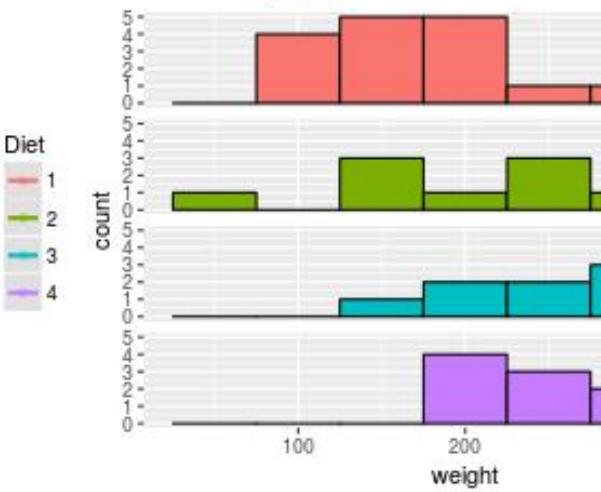
Fitted growth curve per diet



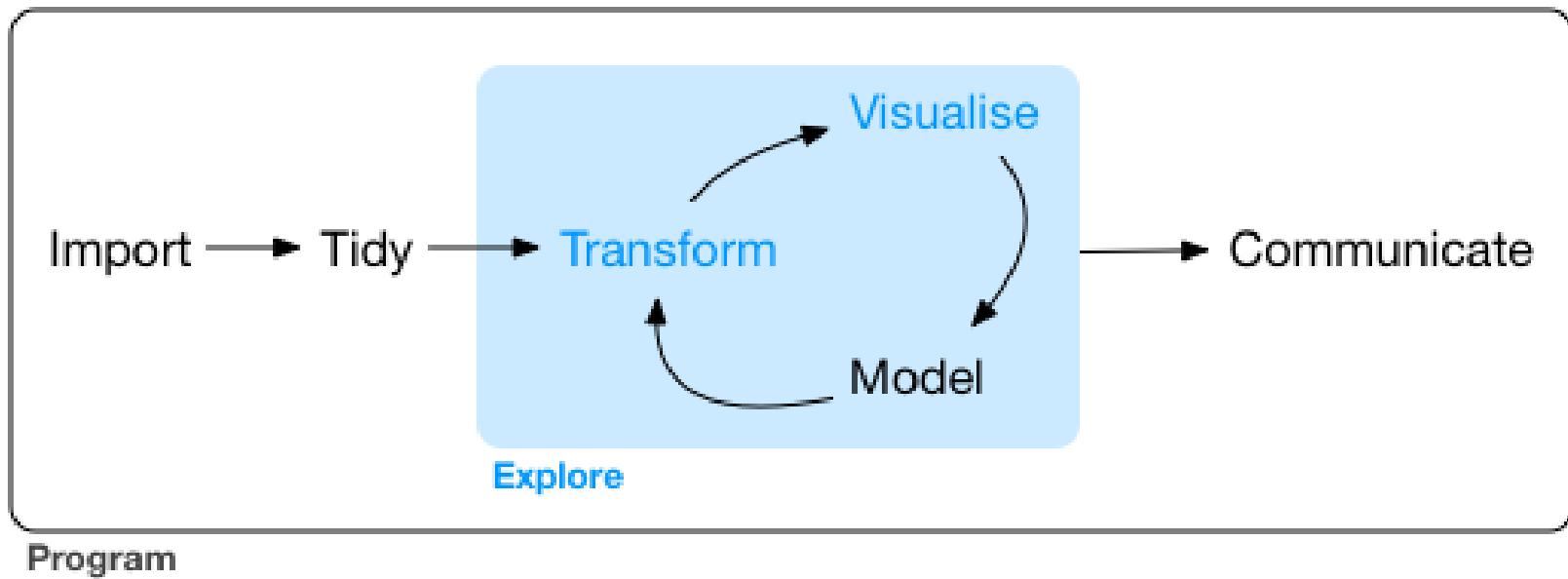
Final weight, by diet



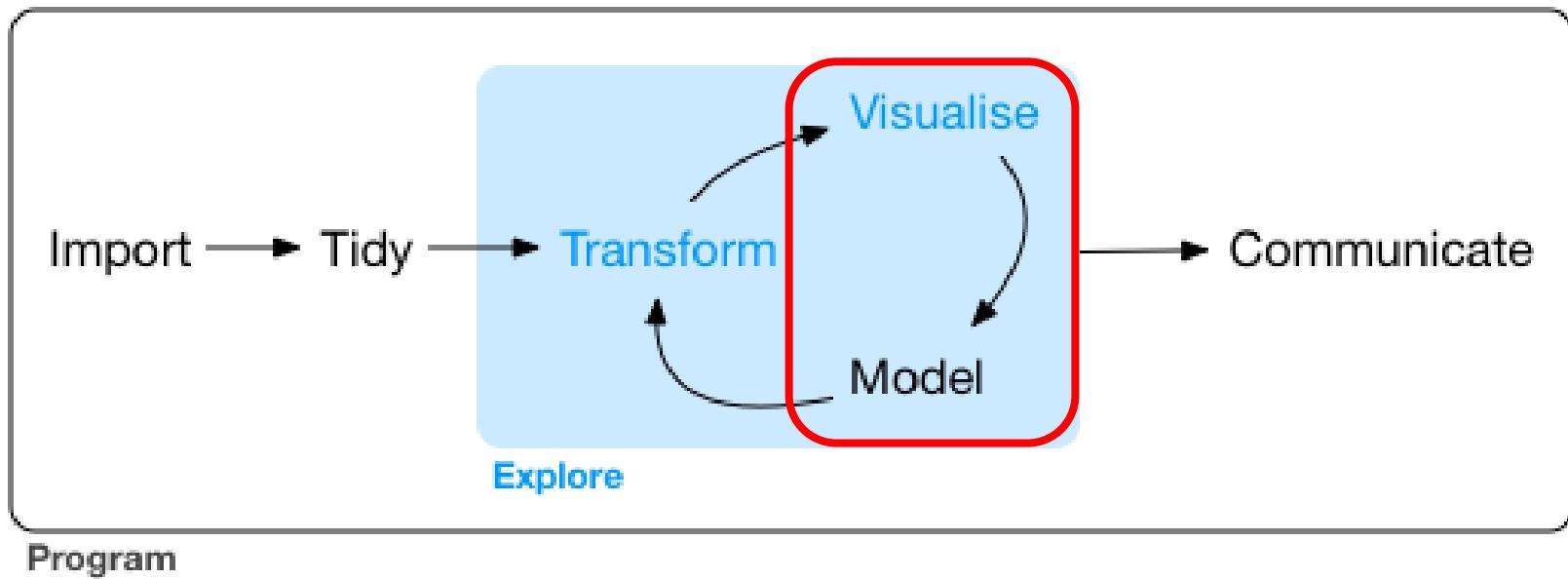
Final weight, by diet



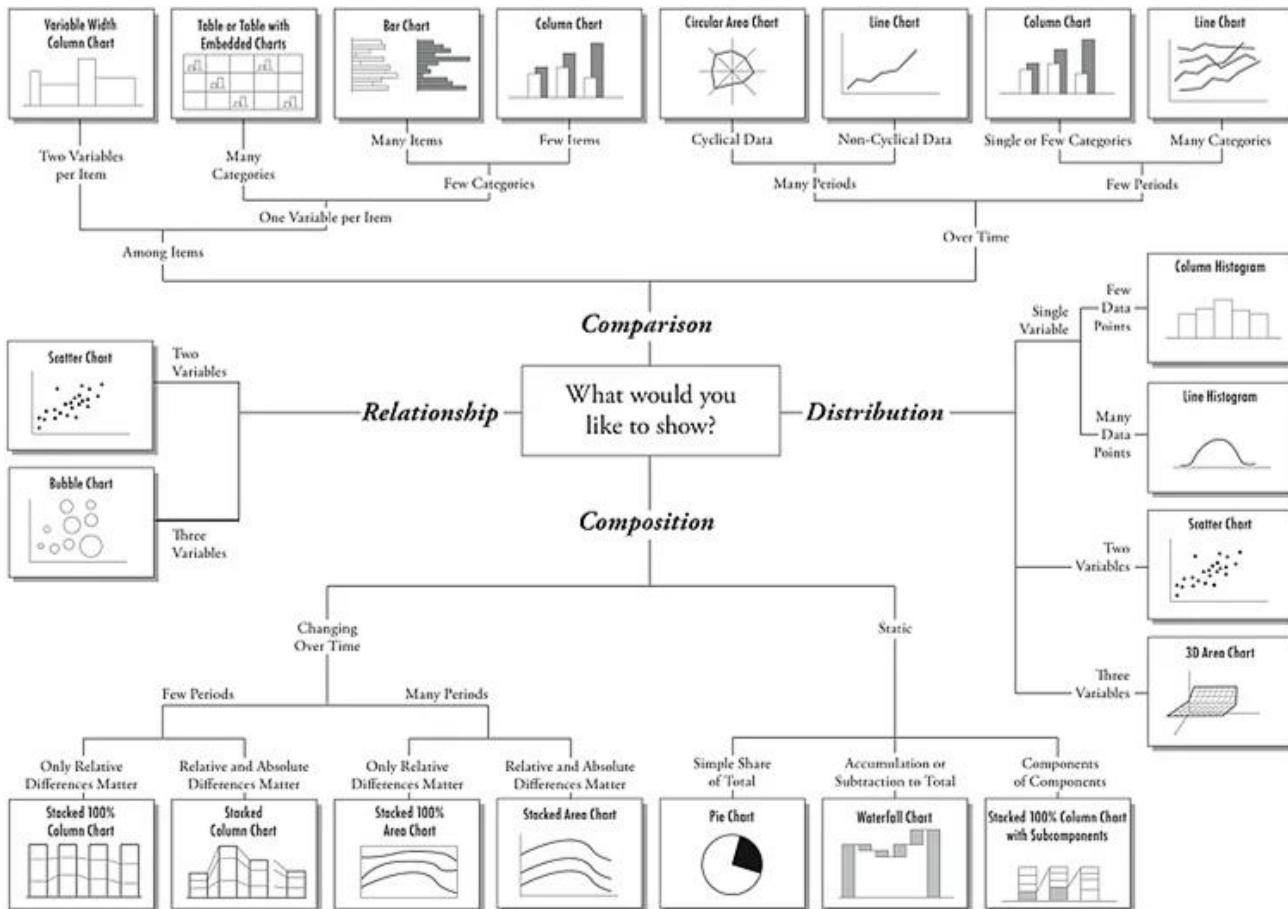
# Visualización de datos



# Visualización de datos



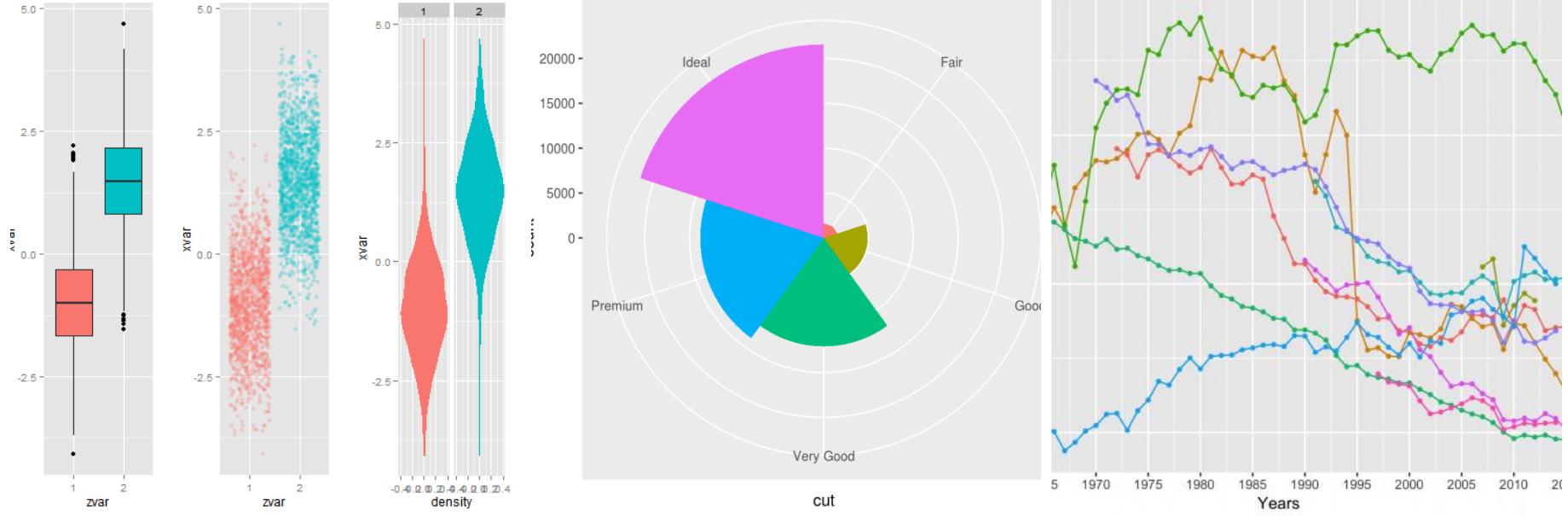
# Chart Suggestions—A Thought-Starter



# Visualización de datos - {ggplot2}

- El paquete de gráficos más extendido y empleado actualmente
- Su creador, Hadley Wickham, CEO de RStudio es autor de los paquetes más populares en la actualidad
- Numerosos paquetes de visualización están basados a su vez en **ggplot2**
- Se caracteriza por emplear la *gramática de gráficos*, una forma intuitiva de crear gráficas





# Visualización de datos - {ggplot2}

---

## Algunos ejemplos

# Visualización de datos - {ggplot2}

[Descárgala aquí](#)

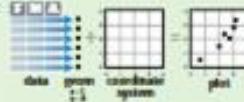
Cheat sheet

## Data Visualization with ggplot2 Cheat Sheet

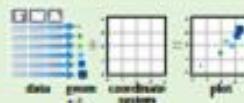


### Basics

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same few components: a **data set**, a set of **geoms**—visual marks that represent data points, and a **coordinate system**.



To display data values, map variables in the data set to aesthetic properties of the geom like **size**, **color**, and **x** and **y** locations.



Geoms - Use a geom to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.		
One Variable	Two Variables	Continuous Bivariate Distribution
<b>Continuous</b> <pre>a &lt;- ggplot(mpg, aes(hwy)) a + geom_area(stat = "bin") x, y, alpha, color, fill, linetype, size b + geom_area(aes(y = ..density..), stat = "bin") a + geom_density(kernel = "gaussian") x, y, alpha, color, fill, linetype, size, weight b + geom_density(aes(y = ..count..)) a + geom_dotplot() x, y, alpha, color, fill</pre>	<b>Continuous X, Continuous Y</b> <pre>f &lt;- ggplot(mpg, aes(cty, hwy)) f + geom_blank() f + geom_jitter() x, y, alpha, color, fill, shape, size f + geom_point() x, y, alpha, color, fill, shape, size f + geom_quantile() x, y, alpha, color, linetype, size, weight f + geom_rug(sides = "bl") alpha, color, linetype, size f + geom_smooth(model = lm) x, y, alpha, color, fill, linetype, size, weight f + geom_text(aes(label = cyl)) x, y, label, alpha, angle, color, family, fontface, hijust, lineheight, size, vjust</pre>	<b>i + geom_bin2d(binwidth = c(5, 0.5))</b> <pre>xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size, weight i + geom_density2d() x, y, alpha, colour, linetype, size i + geom_hex() x, y, alpha, colour, fill size</pre>
<b>Discrete</b> <pre>b &lt;- ggplot(mpg, aes(!)) b + geom_bar() x, alpha, color, fill, linetype, size, weight</pre>	<b>Continuous Function</b> <pre>j &lt;- ggplot(economics, aes(date, unemploy)) j + geom_area() x, y, alpha, color, fill, linetype, size j + geom_line() x, y, alpha, color, linetype, size j + geom_step(direction = "hv") x, y, alpha, color, linetype, size</pre>	<b>Visualizing error</b> <pre>df &lt;- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2) k &lt;- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))</pre>
<b>Graphical Primitives</b> <pre>c &lt;- ggplot(mtcars, aes(wt, mpg)) c + geom_abline(intercept = 0, slope = 1)</pre>		

# Visualización de datos - {ggplot2}

- Estructura del código por capas

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

# Visualización de datos - {ggplot2}

- Estructura del código por capas

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

→ **Primera capa:** Sistema de coordenadas y definición de set de datos

# Visualización de datos - {ggplot2}

- Estructura del código por capas

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

- **Primera capa:** Sistema de coordenadas y definición de set de datos
- **Segunda capa:** Tipo de gráfico y variables a mostrar

# Visualización de datos - {ggplot2}

- Estructura del código por capas

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



Para añadir nuevas capas

- **Primera capa:** Sistema de coordenadas y definición de set de datos
- **Segunda capa:** Tipo de gráfico y variables a mostrar

- geom\_dotplot()
- geom\_errorbarh()
- geom\_hex() stat\_bin\_hex()
- geom\_freqpoly() geom\_histogram()  
stat\_bin()
- geom\_jitter()
- geom\_crossbar() geom\_errorbar()  
geom\_linerange()  
geom\_pointrange()
- geom\_map()
- geom\_path() geom\_line()  
geom\_step()
- geom\_point()

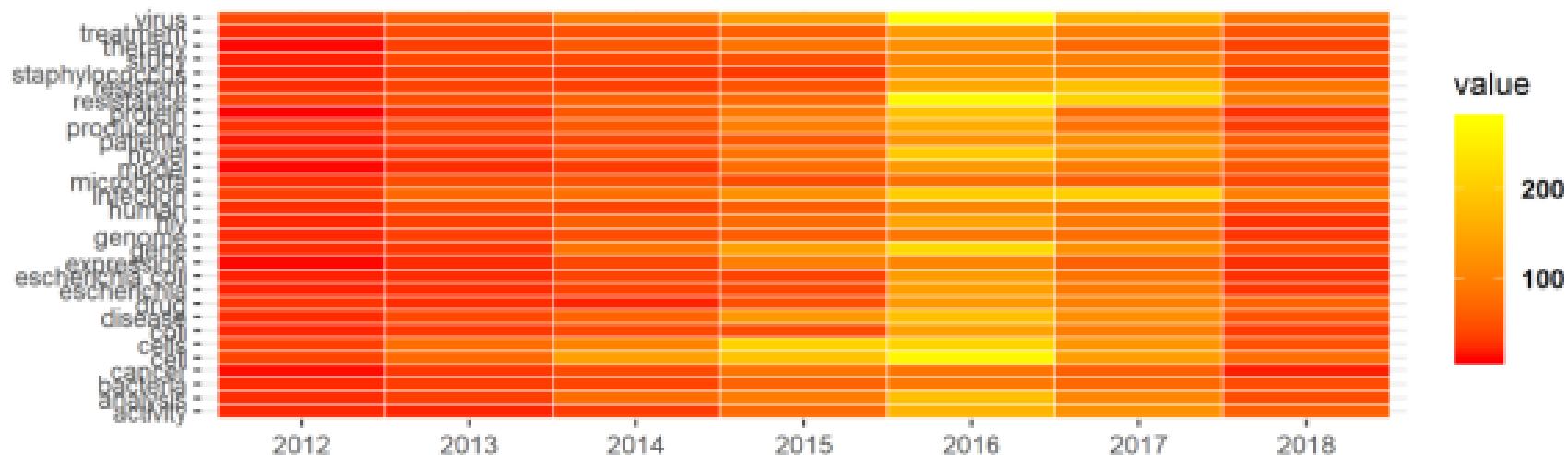
- geom\_abline() geom\_hline()
- geom\_vline()
- geom\_bar() geom\_col()  
stat\_count()
- geom\_bin2d() stat\_bin\_2d()
- geom\_blank()
- geom\_boxplot() stat\_boxplot()
- geom\_contour() stat\_contour()
- geom\_count() stat\_sum()
- geom\_density() stat\_density()  
geom\_density\_2d()  
stat\_density\_2d()

- geom\_polygon()
- geom\_qq\_line() stat\_qq\_line()  
geom\_qq() stat\_qq()
- geom\_quantile() stat\_quantile()
- geom\_ribbon() geom\_area()
- geom\_rug()
- geom\_segment() geom\_curve()
- geom\_smooth() stat\_smooth()
- geom\_spoke()
- geom\_label() geom\_text()
- geom\_raster() geom\_rect()  
geom\_tile()
- geom\_violin() stat\_ydensity()

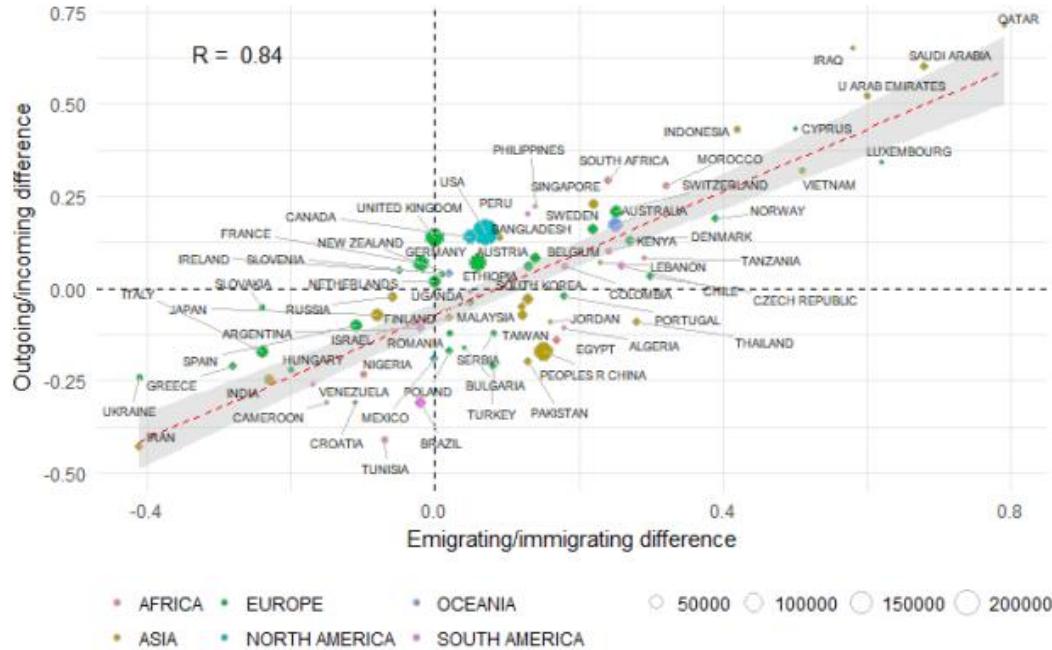
# Tipos de gráficos – {ggplot2}

<https://ggplot2.tidyverse.org/reference/>

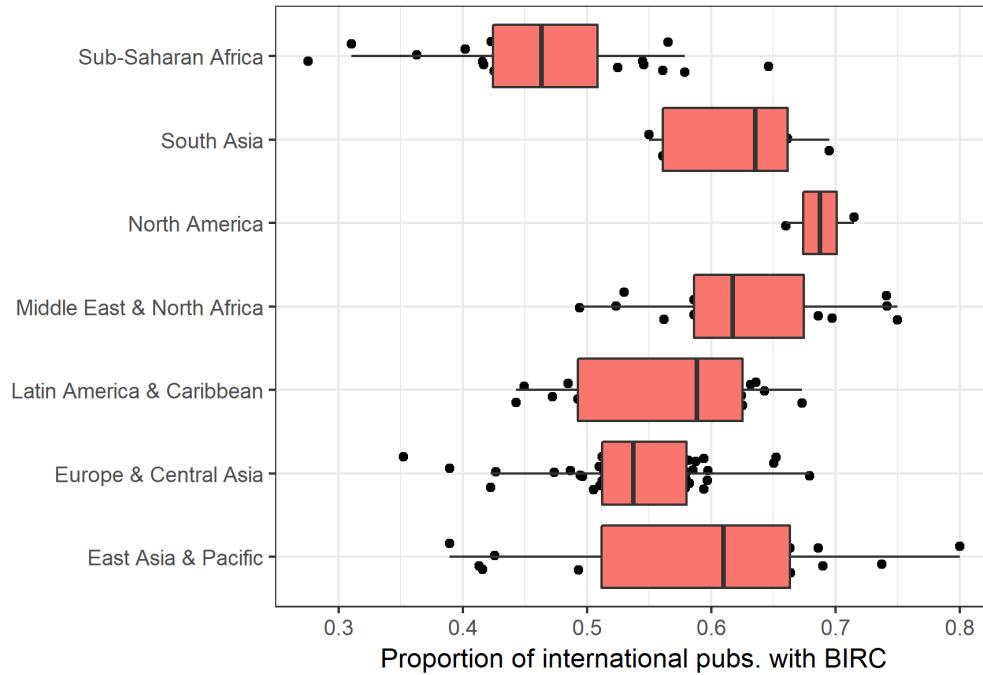
# Algunos ejemplos de elaboración propia



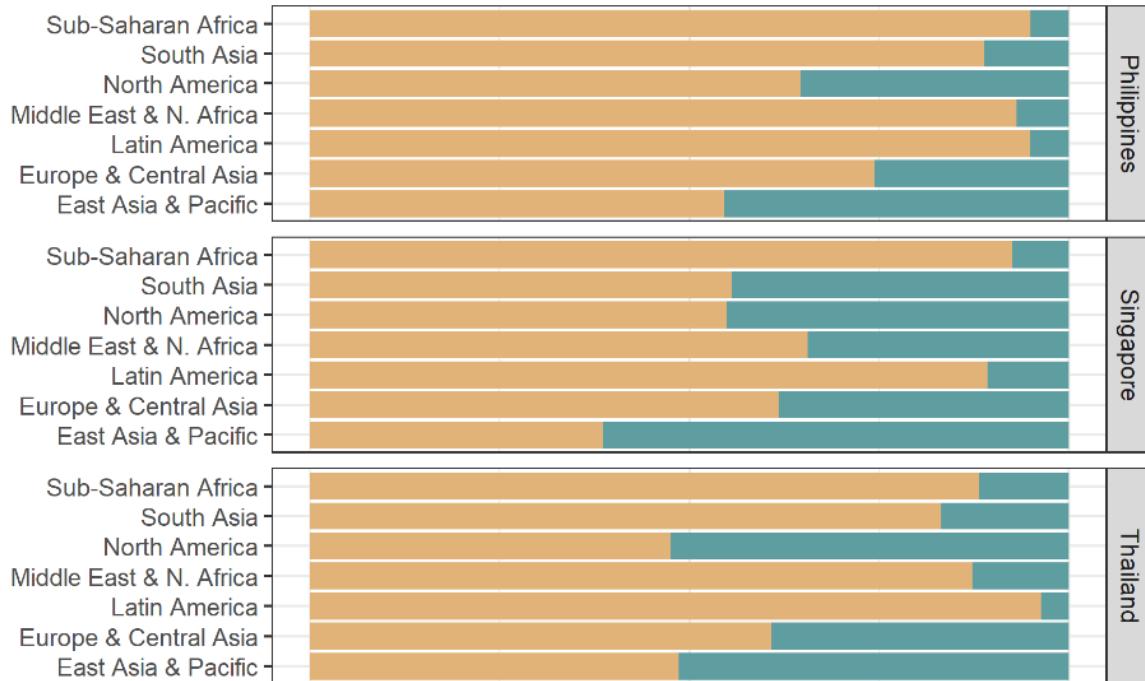
# Algunos ejemplos de elaboración propia



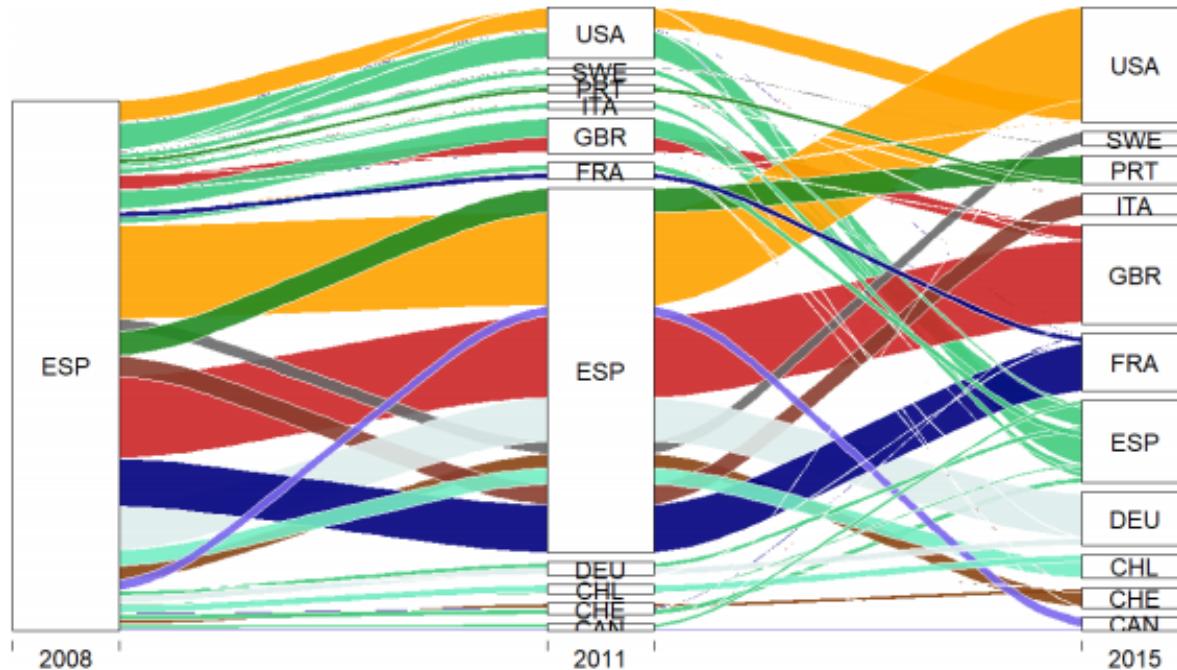
# Algunos ejemplos de elaboración propia



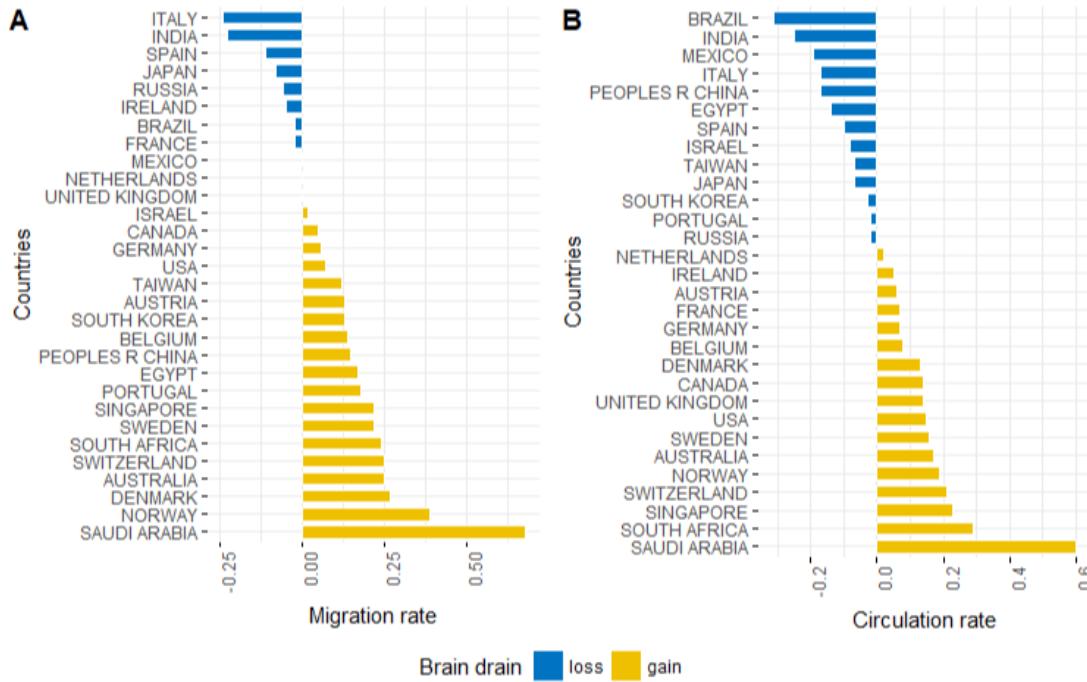
# Algunos ejemplos de elaboración propia



# En combinación con otros paquetes

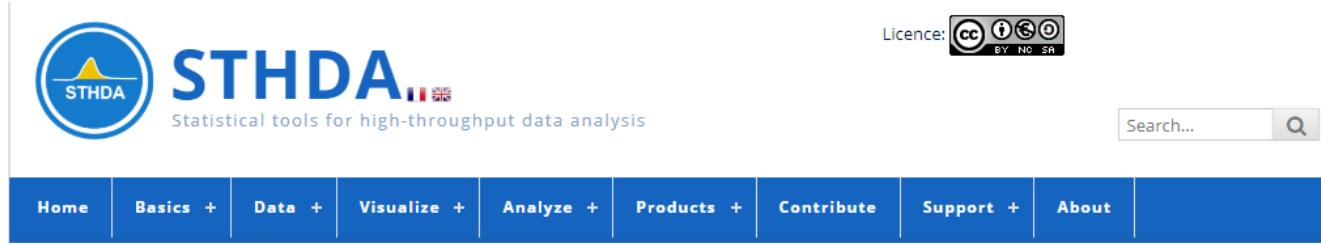


# En combinación con otros paquetes



# Recursos adicionales

- STHDA – {ggpubr} - <http://www.sthda.com/english/>



- ggplot2 extensions gallery - <http://www.ggplot2-exts.org/gallery/>



51 registered extensions available to explore

Sort: Github stars ▾

Text Filter: search name, author, ▾

Author Filter: ▾

Tag Filter: ▾

CRAN Only:

# Ejercicio 5. Gráficos con *ggplot2*

---

- **Objetivo:** Familiarizarse con la librería *ggplot2*
- **Tarea:** Instala la librería de visualización de datos *ggplot2*. Una vez instalada, deberás activarla para comenzar a utilizar sus funciones con el comando `library()`. Utilizando el set de datos USArrests, comienza a explorar gráficamente los datos
- **Planteamiento:** Crea un documento Rmarkdown en el que documentes las distintas gráficas que crees. Deberás crear un gráfico de dispersión, un gráfico de caja y bigotes y un gráfico de barras.



# Material extra

# Análisis estadísticos - Test de normalidad

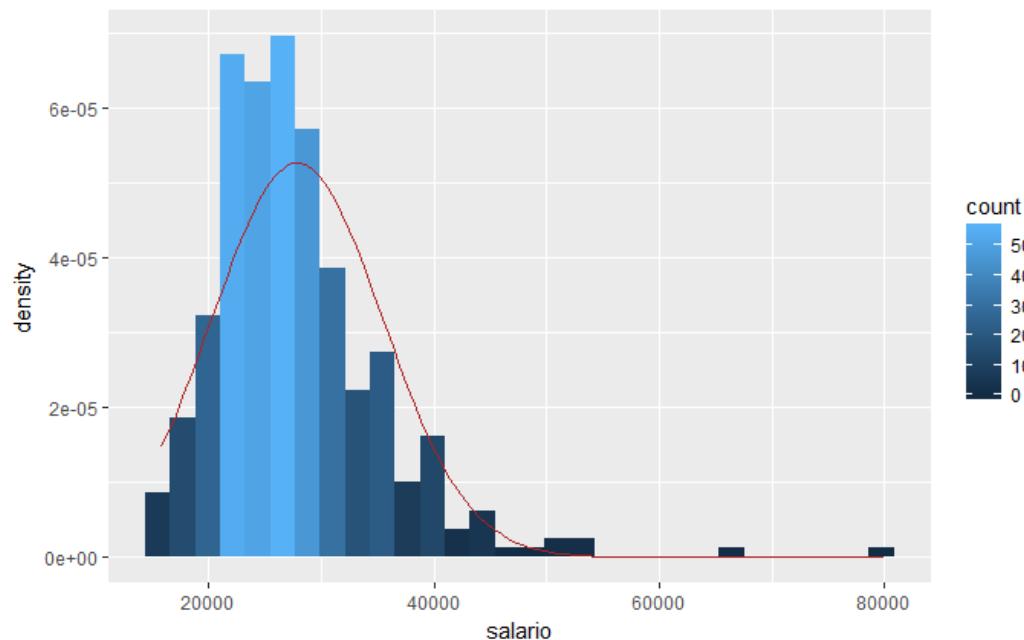
¿Cobran todos los técnicos de la empresa X un salario similar?

```
```{r}
aggregate(empleados2$salario, list(empleados2$catLab), FUN= mean)
```

Group.1 <fctr>	x <dbl>
1	27838.54
2	30938.89
3	63977.80

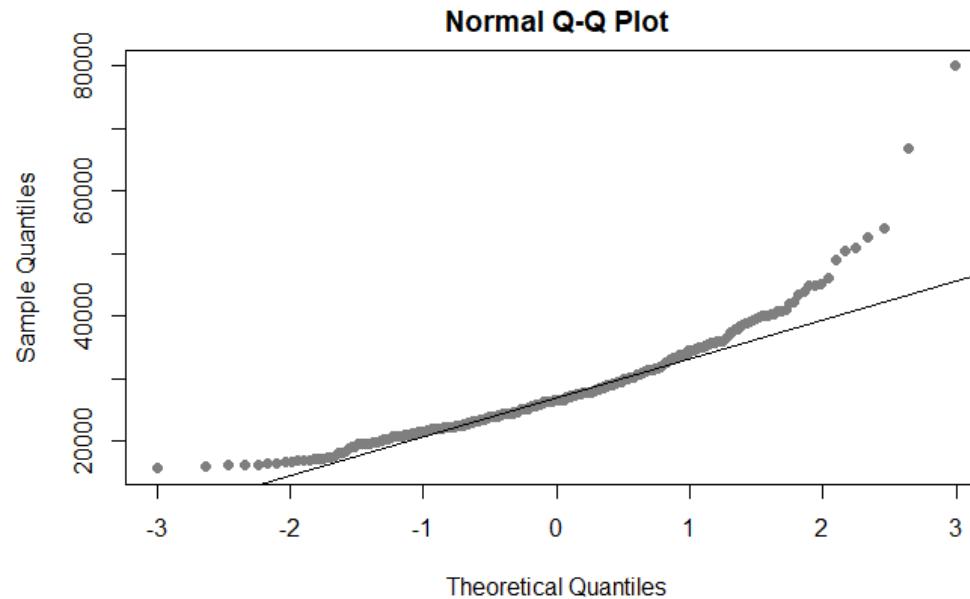
# Análisis estadísticos - Test de normalidad

El histograma nos permite ver rápidamente si la distribución es normal



# Análisis estadísticos - Test de normalidad

Otra opción es comparar la distribución por cuantiles reales con los teóricos



# Análisis estadísticos - Test de normalidad

Test de Shapiro-Wilk de normalidad

> shapiro.test()

```
shapiro-wilk normality test

data: tecnicos$salario
W = 0.88207, p-value = 4.568e-16
```

# Análisis estadísticos - Contraste de hipótesis

Comparación del salario de directivos y técnicos  
(catlab= 3 y 1)

> t.test()

```
welch Two Sample t-test

data: dire$salario and tec$salario
t = 17.803, df = 89.708, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 32106.31 40172.21
sample estimates:
mean of x mean of y
 63977.80 27838.54
```

# Análisis estadísticos - ANOVA

Comparación de salarios entre las tres categorías

```
> aov(var1 ~ var2)
```

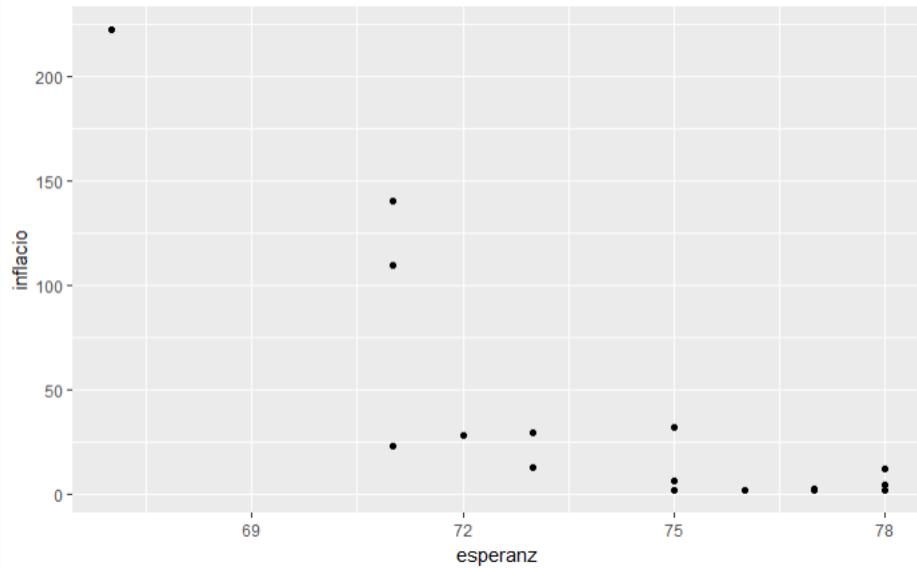
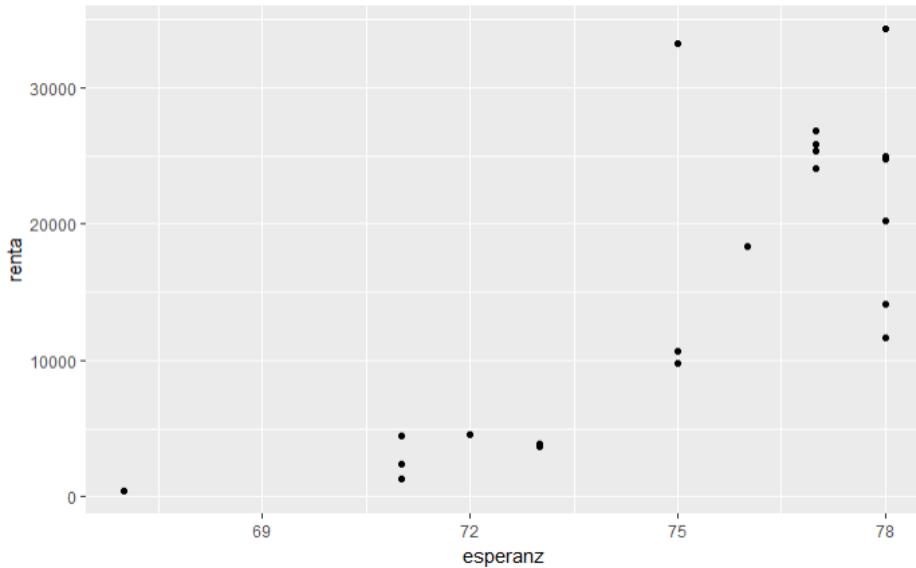
**Nota:** Esta permite visualizar distintos resultados. Debes asignarle un objeto. Después, prueba resumiendo los resultados o creando un gráfico:

```
> summary()
```

```
> plot()
```

# Regresiones y correlaciones

¿Influye la renta en la esperanza de vida de los países europeos? ¿Y la inflación?



# Regresiones y correlaciones

¿Influye la renta en la esperanza de vida de los países europeos? ¿Y la inflación?

```
```{r}
cor.test(europa$esperanz, europa$renta)
```
```

Pearson's product-moment correlation

```
data: europa$esperanz and europa$renta
t = 5.4165, df = 19, p-value = 3.163e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5234034 0.9060391
sample estimates:
cor
0.7790636
```

# Regresiones y correlaciones

¿Influye la renta en la esperanza de vida de los países europeos? ¿Y la inflación?

```
call:  
lm(formula = esperanz ~ renta + inflacio, data = europa)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-3.1205 -0.3580 -0.1185  0.9304  2.9355  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 7.395e+01 8.463e-01  87.378 < 2e-16 ***  
renta        1.272e-04 3.936e-05   3.232  0.00463 **  
inflacio    -2.985e-02 7.669e-03  -3.892  0.00107 **  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 1.54 on 18 degrees of freedom  
Multiple R-squared:  0.7866,    Adjusted R-squared:  0.7628  
F-statistic: 33.16 on 2 and 18 DF,  p-value: 9.196e-07
```