

powered by

R programming

Introducción general al uso de R - Clase 01 -

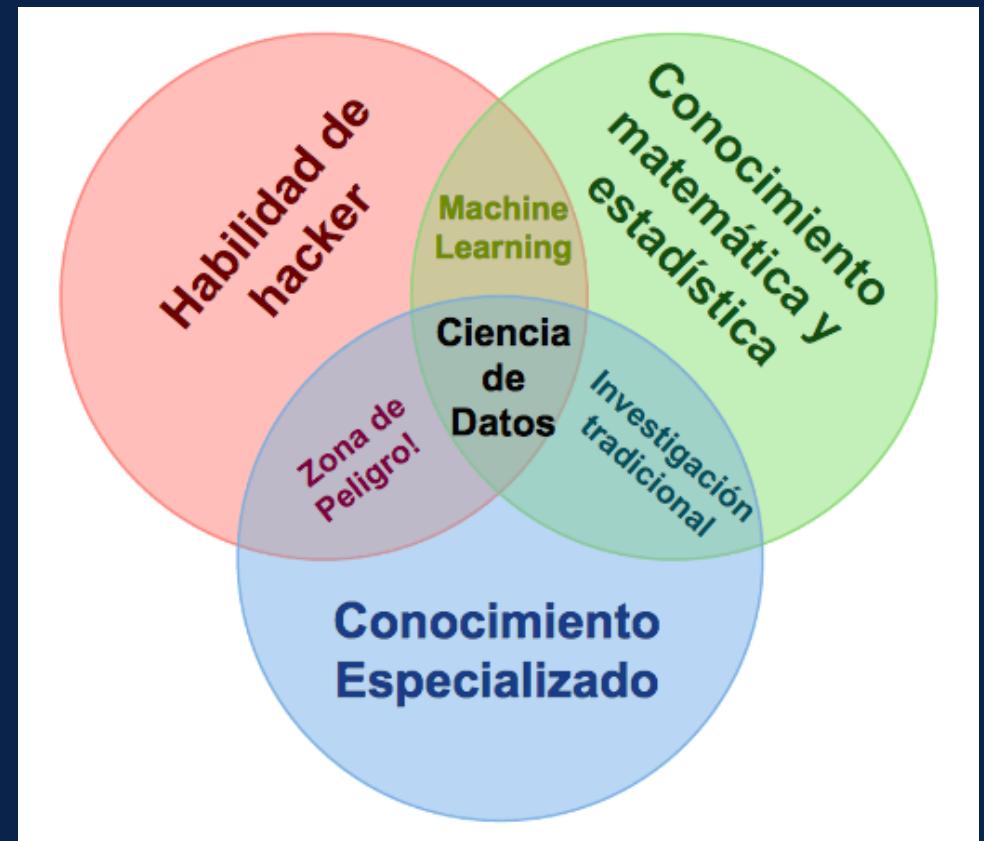
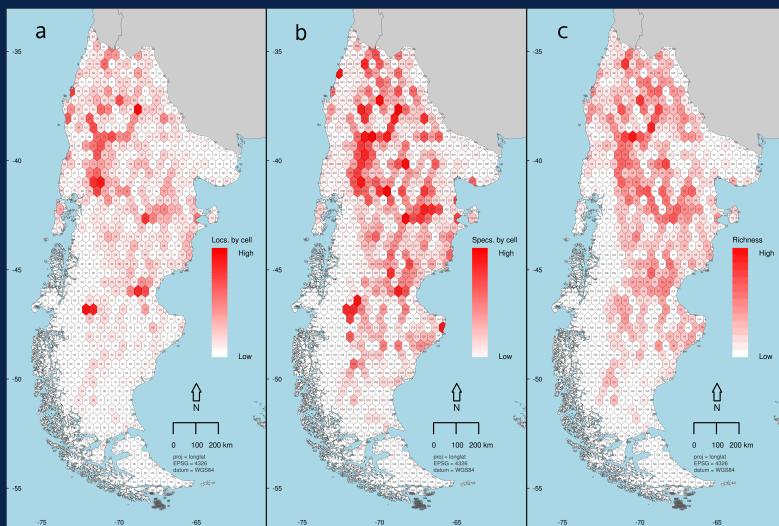
minolicnp@gmail.com

Ignacio Minoli *pwp*

Observatorio de Biodiversidad del Bosque Atlántico (OBBA)
Instituto de Biología Subtropical (IBS) - CONICET - UNaM.



Presentación



Formación y Experiencia

- Tecnologías: **Software Libre | Código Abierto | Linux | Programación en R | Programación en Bash-Shell | Programación en Python | Programación en HTML | SublimeText | Rstudio | Shiny Environment | Maxent | Marxan**
- Especialidades: **Biogeografía | Bases de Datos | Programación en R | Cartografía | Mapas | Cambio Climático | Modelado de Nicho Ecológico | GIS | Análisis Espaciales**
- Áreas de desarrollo e investigación: **Manipulación y curación de grandes bases de datos. Alcance internacional, nacional y regional. Biogeografía, Ecoinformática y Sistemas de Información Geográfica.**
- Áreas grales: **Programación, Estadística y Cartografía.**



Proyectos actuales



Historia personal con R

Recomendaciones...

- ¿Me conviene utilizar R u otro lenguaje de programación?
- Cursos de programación en R **versus** cursos de Estadística.
- Uso y/o lectura diaria - feeds.
- Autodidacta. Enfoque: preguntas - hipótesis - problemas puntuales.
- Biblioteca personal de scripts.

```
herramienta_principal <- "R is awesome and diverse"
#
scripts_propios <- c("Lecturas diarias", "Prueba error")
#
independencia_desarrollo <- (filosofia = "Software.Abierto_Gratis_Multiplataforma")
```

Posibles usos

- Análisis estadísticos

```
> LinearModel <- lm(dist ~ speed, data=cars)
> print(LinearModel)

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)      speed
-17.579         3.932

> |
```

```
> summary(LinearModel)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069 -9.525 -2.272  9.215 43.201 

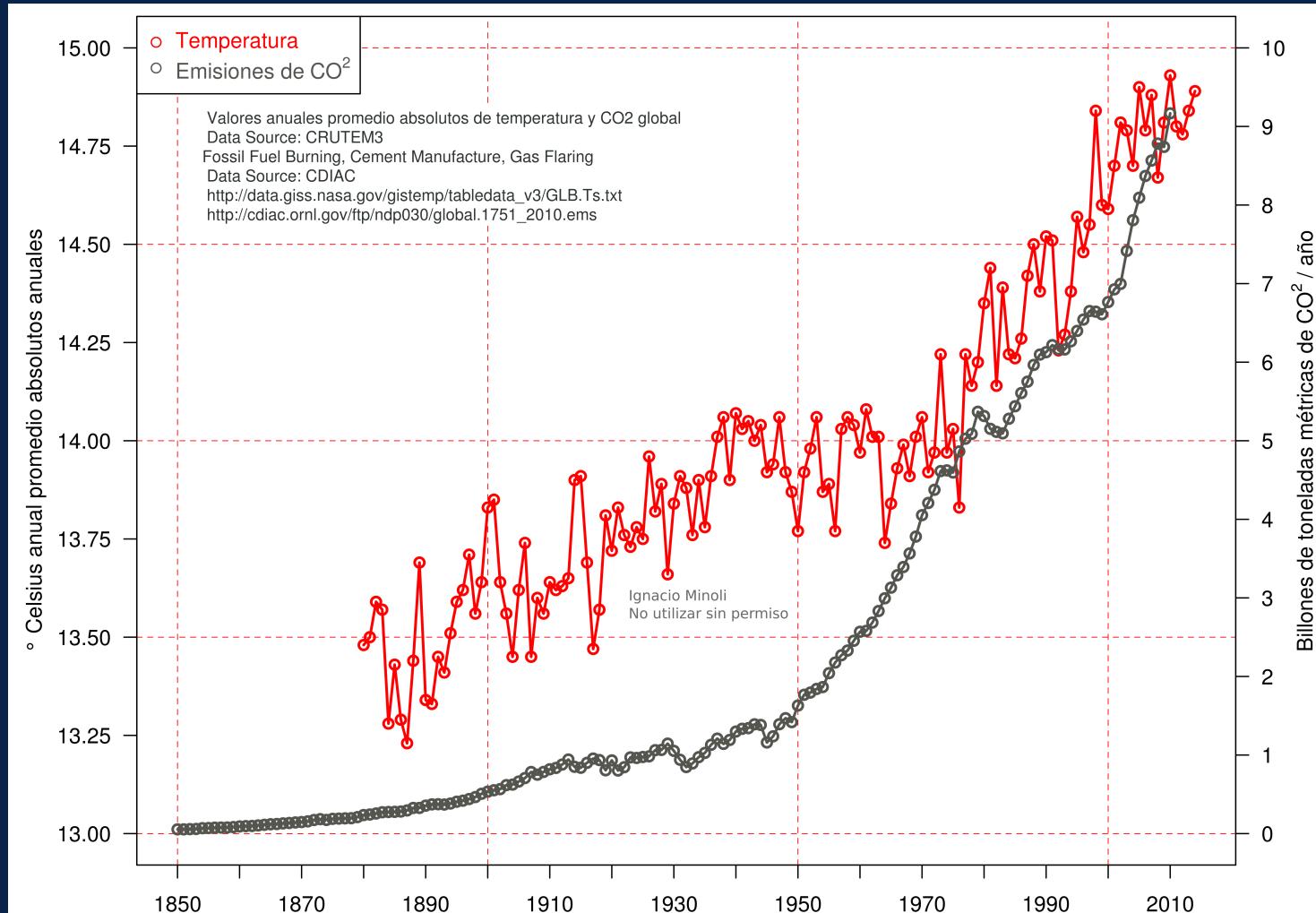
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791   6.7584 -2.601   0.0123 *  
speed        3.9324   0.4155  9.464 1.49e-12 *** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
> |
```

- Descriptiva / Inferencial (Paramétrica No Paramétrica). Uni, Multivariada, Bayesiana, etc.

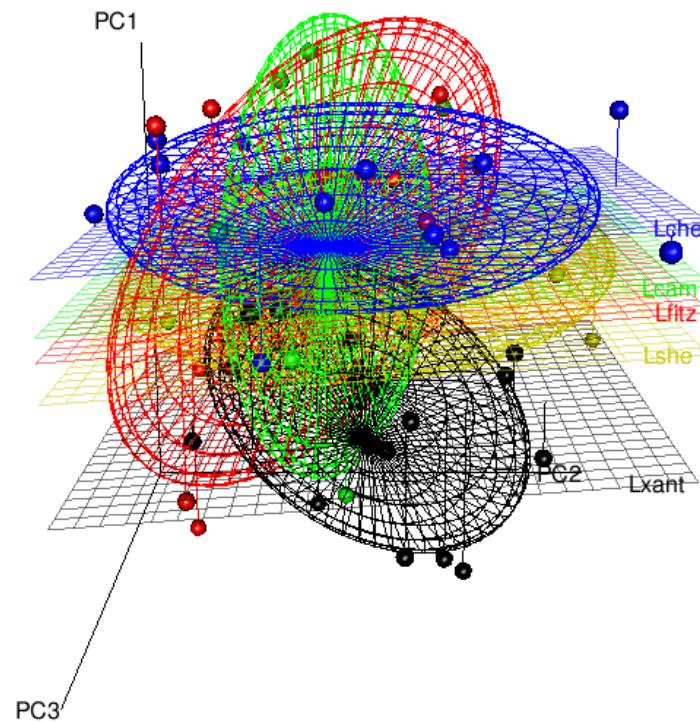
Posibles usos

- Plots



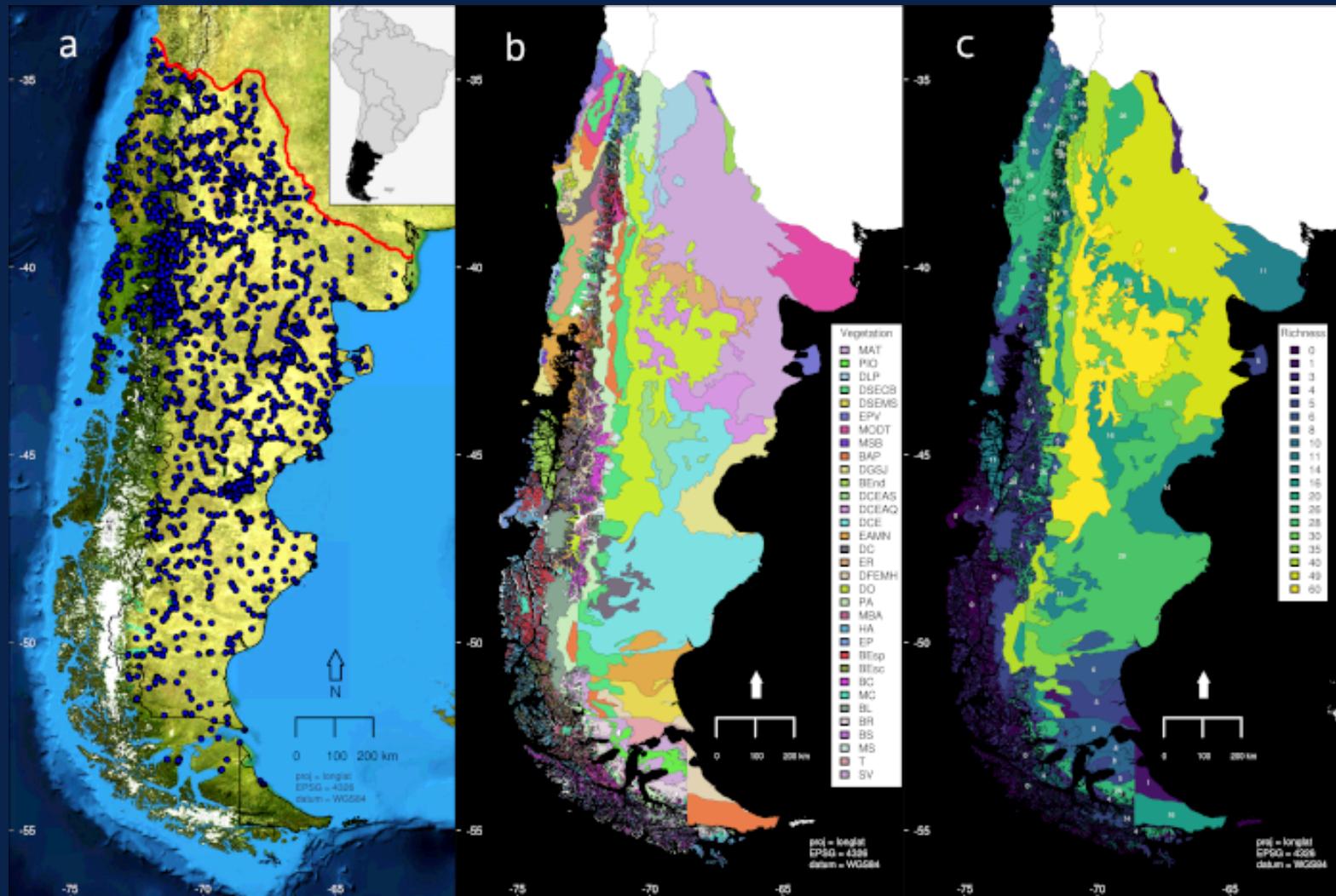
Posibles usos

- Plots



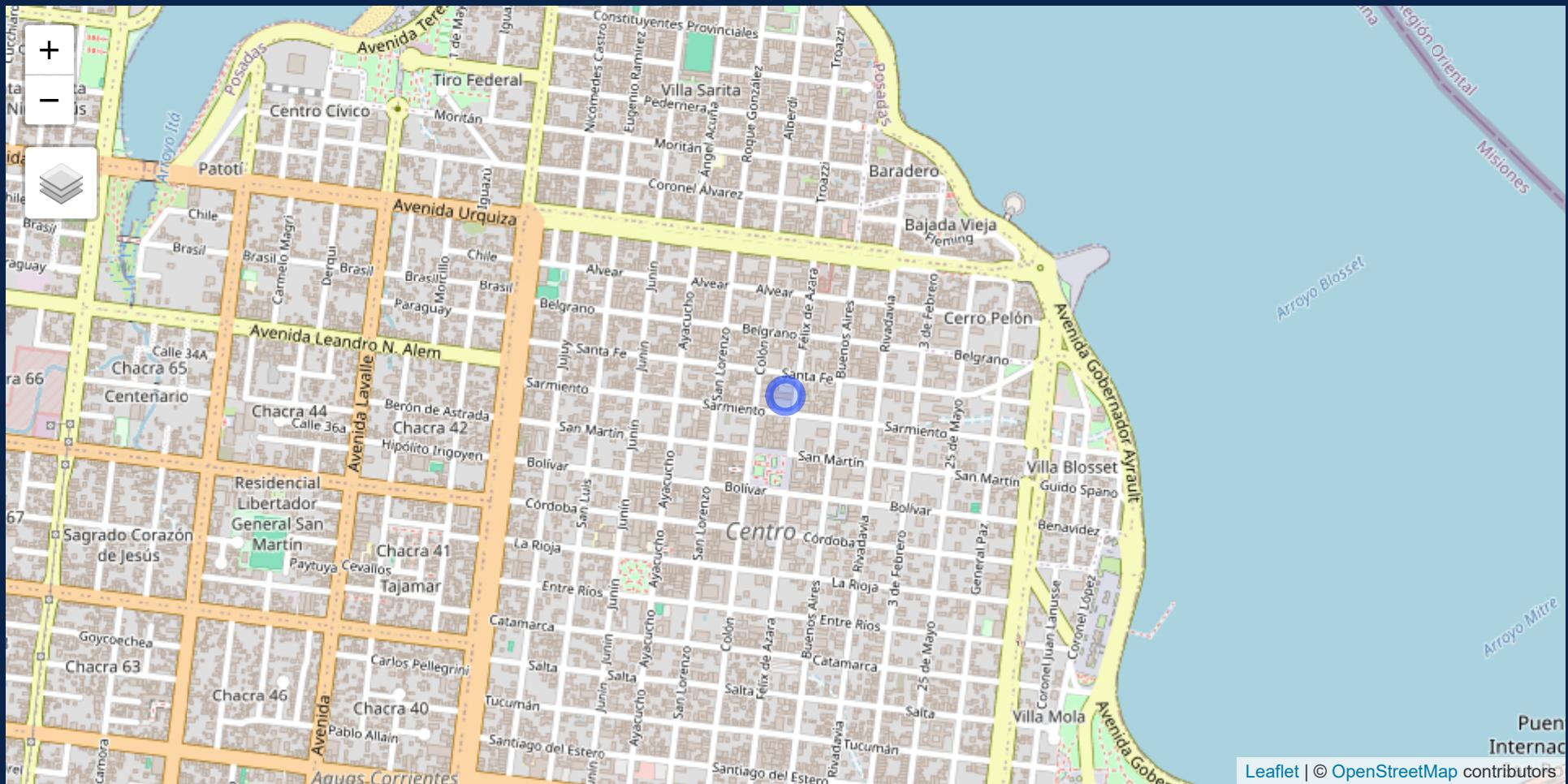
Posibles usos

- Mapas



Posibles usos

- Mapas interactivos



Posibles usos

- Tablas

2019 Women's World Cup Predictions												
TEAM	GROUP	Team Rating			Chance of Finishing Group Stage In ...			Knockout Stage Chances				
		SPI	OFF.	DEF.	1ST PLACE	2ND PLACE	3RD PLACE	MAKE ROUND OF 16	MAKE QTR-FINALS	MAKE SEMIFINALS	MAKE FINAL	WIN WORLD CUP
USA 6 pts.	F	98.3	5.5	0.6	83%	17%	–	✓	78%	47%	35%	24%
France 6 pts.	A	96.3	4.3	0.5	>99%	<1%	<1%	✓	78%	42%	30%	19%
Germany 6 pts.	B	93.8	4.0	0.7	98%	2%	–	✓	89%	48%	28%	12%
Canada 6 pts.	E	93.5	3.7	0.6	39%	61%	–	✓	59%	36%	20%	9%
England 6 pts.	D	91.9	3.5	0.6	71%	29%	–	✓	69%	43%	16%	8%
Netherlands 6 pts.	E	92.7	3.9	0.7	61%	39%	–	✓	59%	37%	19%	8%
Australia 3 pts.	C	92.8	4.2	0.9	13%	54%	34%	>99%	54%	26%	10%	5%

Posibles usos

- Tablas interactivas

Num	Cat	Lon	Lat	Clase	Orden	Genero
1		-54.4	-26.2		Entomobryomorpha	Lepidocyrtus
2		-54.4	-26.2		Sympyleona	Collophora
3		-54.4	-26.2		Sympyleona	Collophora
4		-54.4	-26.1		Sympyleona	Neosminthurus
5		-54.4	-26.1		Entomobryomorpha	Folsomina
6		-54.4	-26.1		Entomobryomorpha	Proisotoma
7		-54.4	-26.1		Poduromorpha	Neotropiella
8		-54.4	-26.1		Entomobryomorpha	Isotoma

[CSV](#)[Excel](#)[Previous](#)

1

2

3

4

5

...

327

[Next](#)

Posibles usos

- Apps web interactivas para webs de proyectos, disponibilizar resultados de papers, exposiciones en congresos, etc:

Ejemplos -> <https://fauna-atropellada.org.ar/filtro-por-especies/>

- Filtros reactivos - interactivos para una rápida vizualización de datos tabulados, búsqueda de información en big data, presentaciones de powerpoint, etc.

App

Estructura del Curso

- Carga teórica por las mañanas.
- Prácticas con scripts-data reproducibles por las tardes.
- Puesta en común teórica-práctica al final de las tardes.
- Ejercicio final del curso aplicando el contenido del mismo.
- Los pdf de las teorías y prácticas serán entregados.



¿Qué es el lenguaje R?

- R es un entorno y lenguaje de programación con un enfoque -> análisis estadístico + gráficos.
- Fue creado en 1993 por Ross Ihaka y Robert Gentleman. Fines 70' lenguaje S.
- El Comprehensive R Archive Network (CRAN) se creó en 1997. Alojamiento de: archivos ejecutables, código, documentación y librerías o paquetes.
- Actualmente desarrollado por el R Development Core Team.
- Sistema GNU y se distribuye bajo la licencia GNU GPL: Win, Mac, Linux.
- 2008 -> UTF-8. Idiomas; 2010 -> Soporte Win 64 bits
- Líneas de comandos (CLIs).
- Múltiples graphical user interface (GUIs).
- Múltiples Integrated development environment (IDEs).



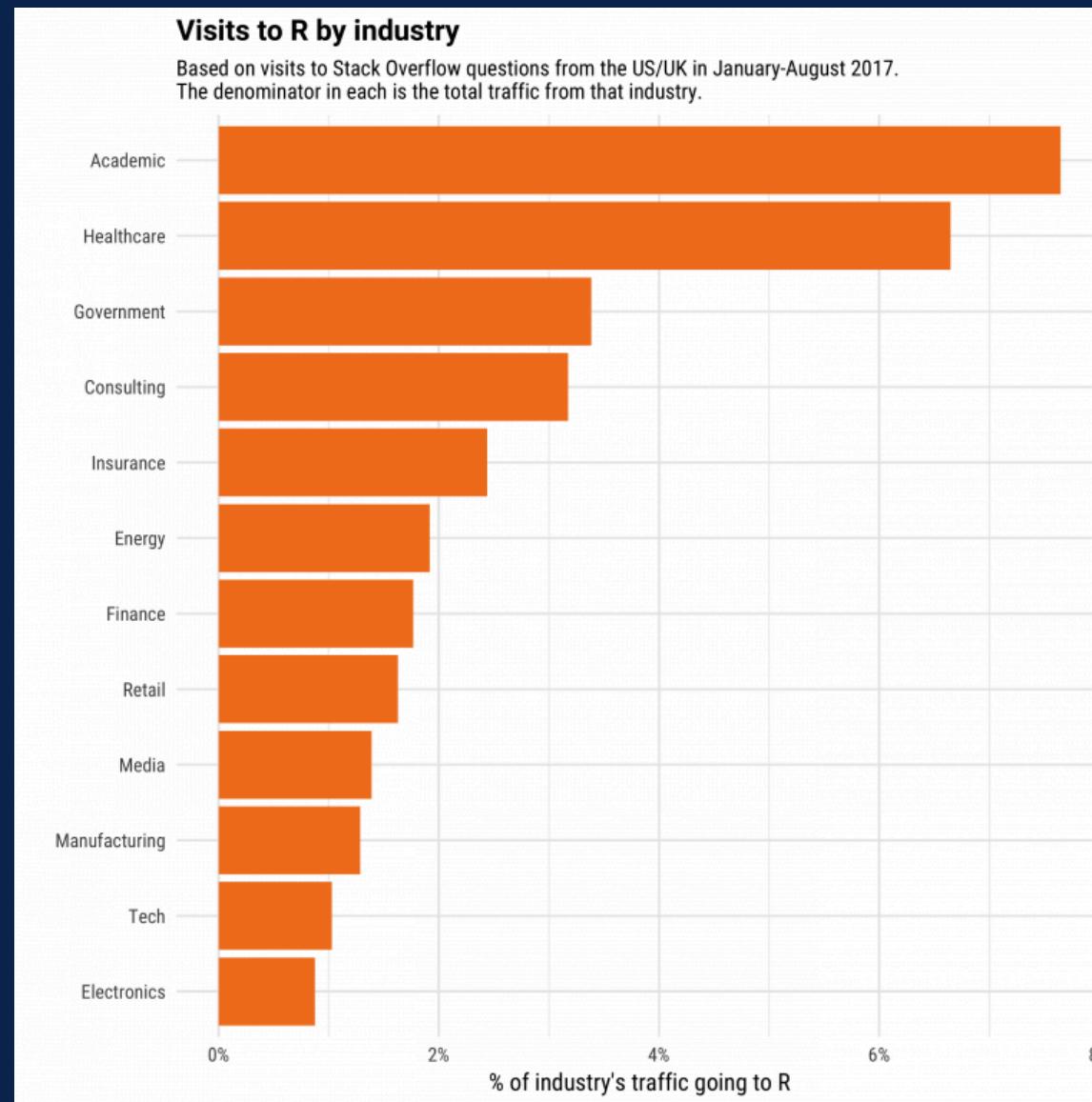
Características principales

- Open-source. Soporte multiplataforma. Compatible con otros lenguajes de programación.
- Programación orientada a objetos. Permite que los usuarios lo extiendan definiendo sus propias funciones.
- Enormes capacidades gráficas. Manejo y Almacenaje de datos.
- Comunidad de ayuda - soporte muy grande y diversa. Usuarios estadísticos: nuevas ideas y metodologías.
- Amplia oferta de paquetes - funciones. Realizar cálculos estadísticos complejos.

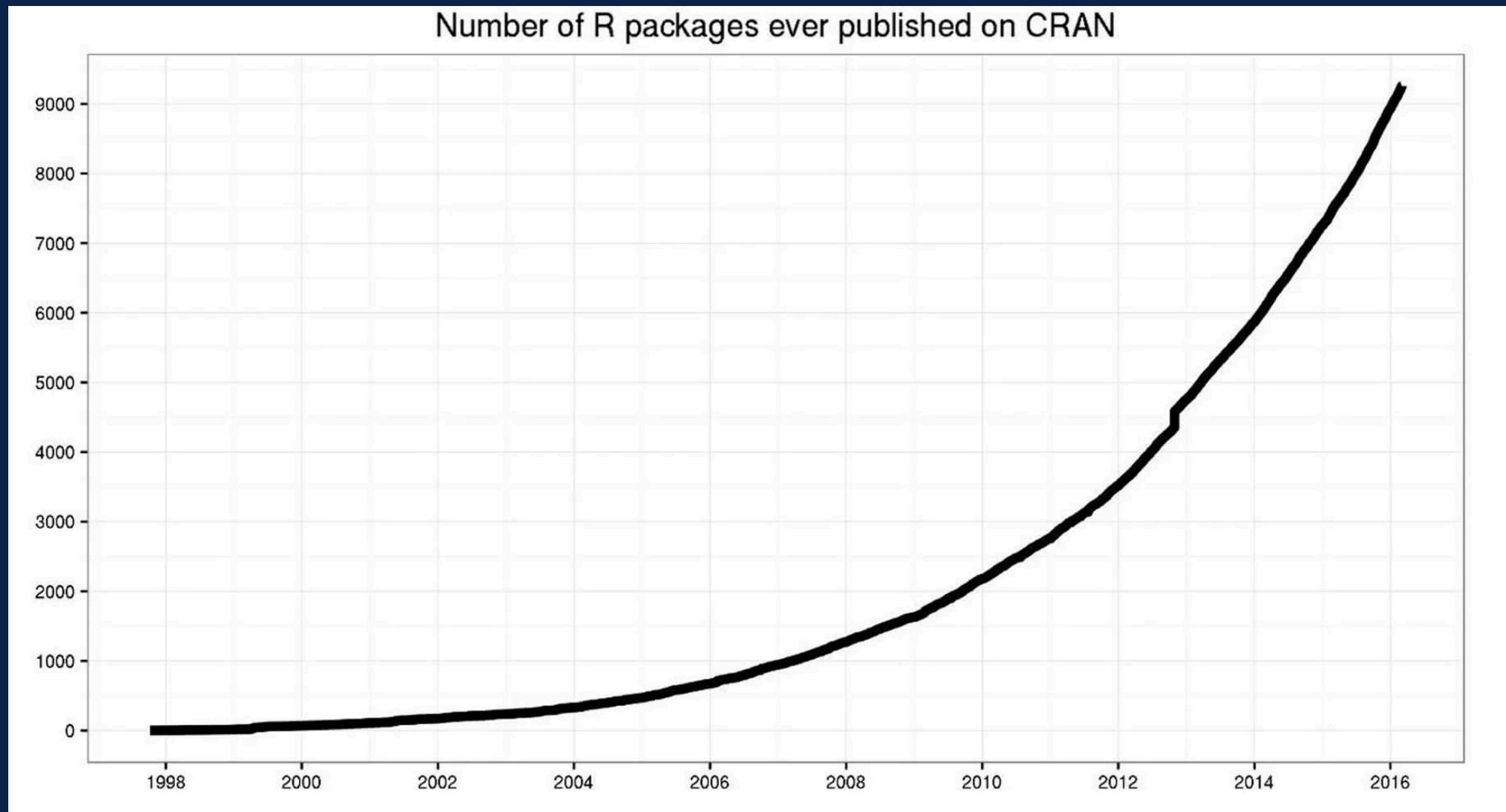
Características principales

- Paralelizar o trabajar en clusters. Operaciones aritméticas vectoriales.
- Interactúa con bases de datos: ROracle, Open Database Connectivity Protocol, RMySQL, etc.
- Amplia variedad de datos. Genera reportes en cualquier formato.
- Data Wrangling, Web Scraping, Machine Learning, Información de APIs, servers, etc.

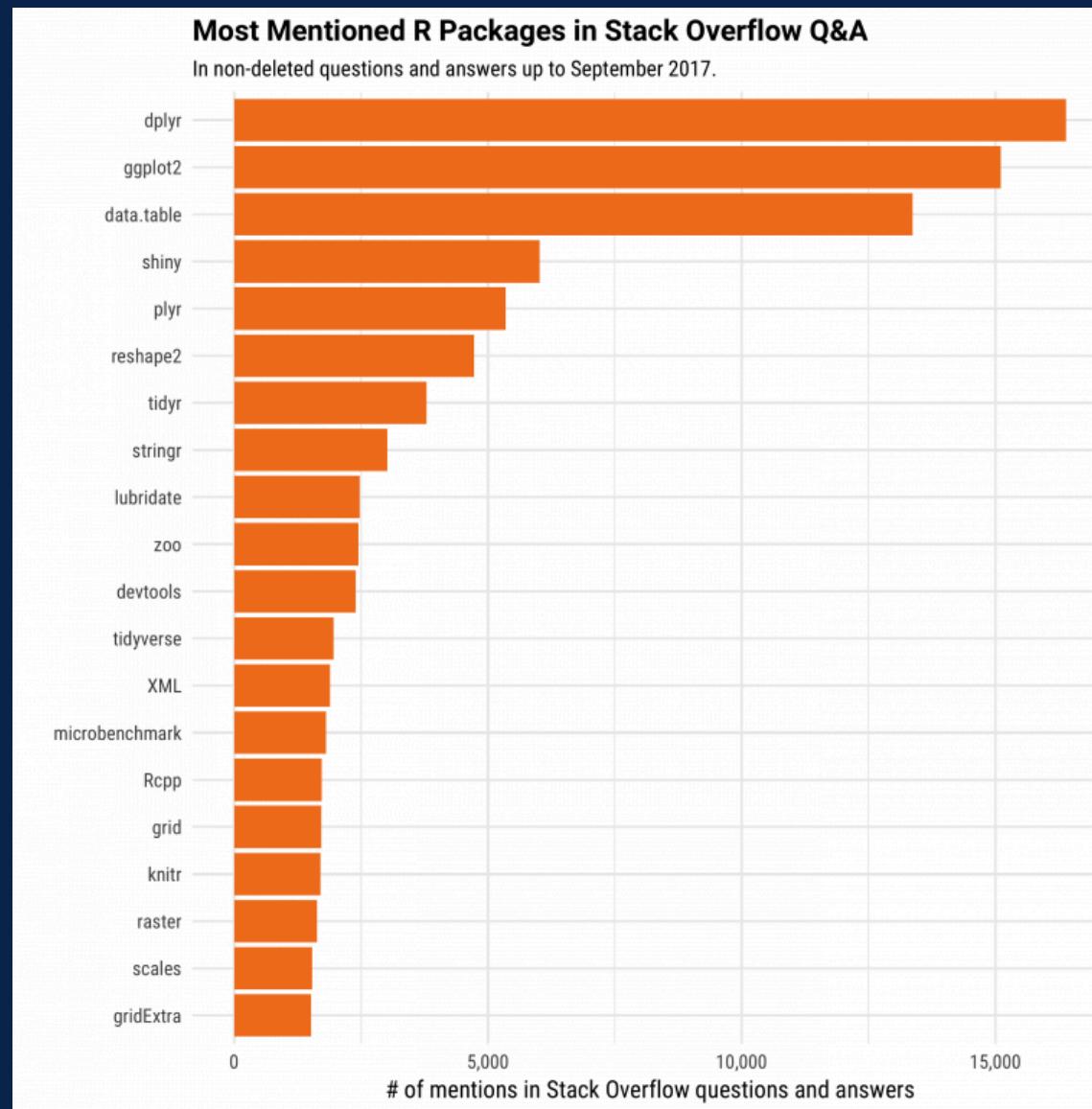
Características - usos



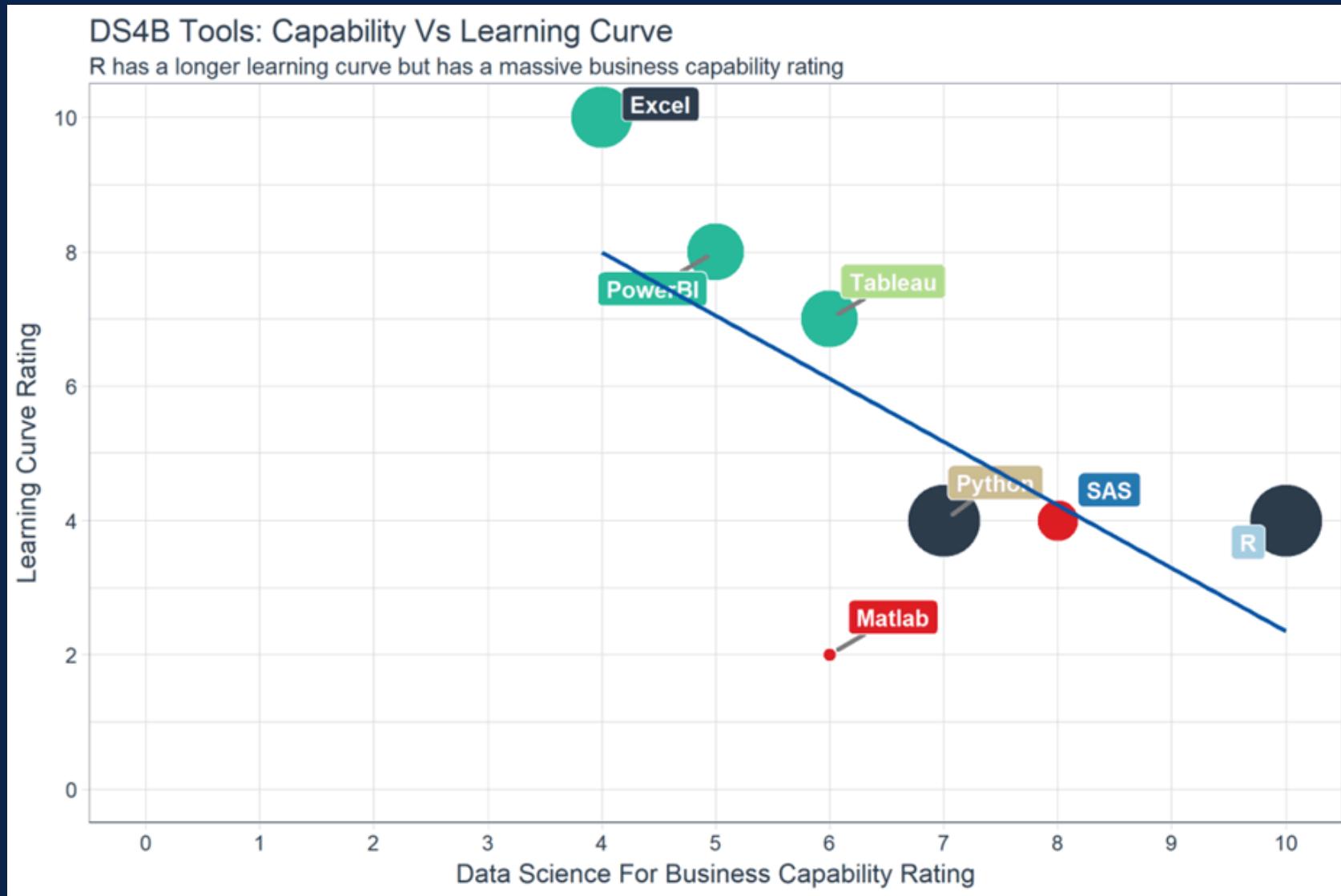
Características - paquetes



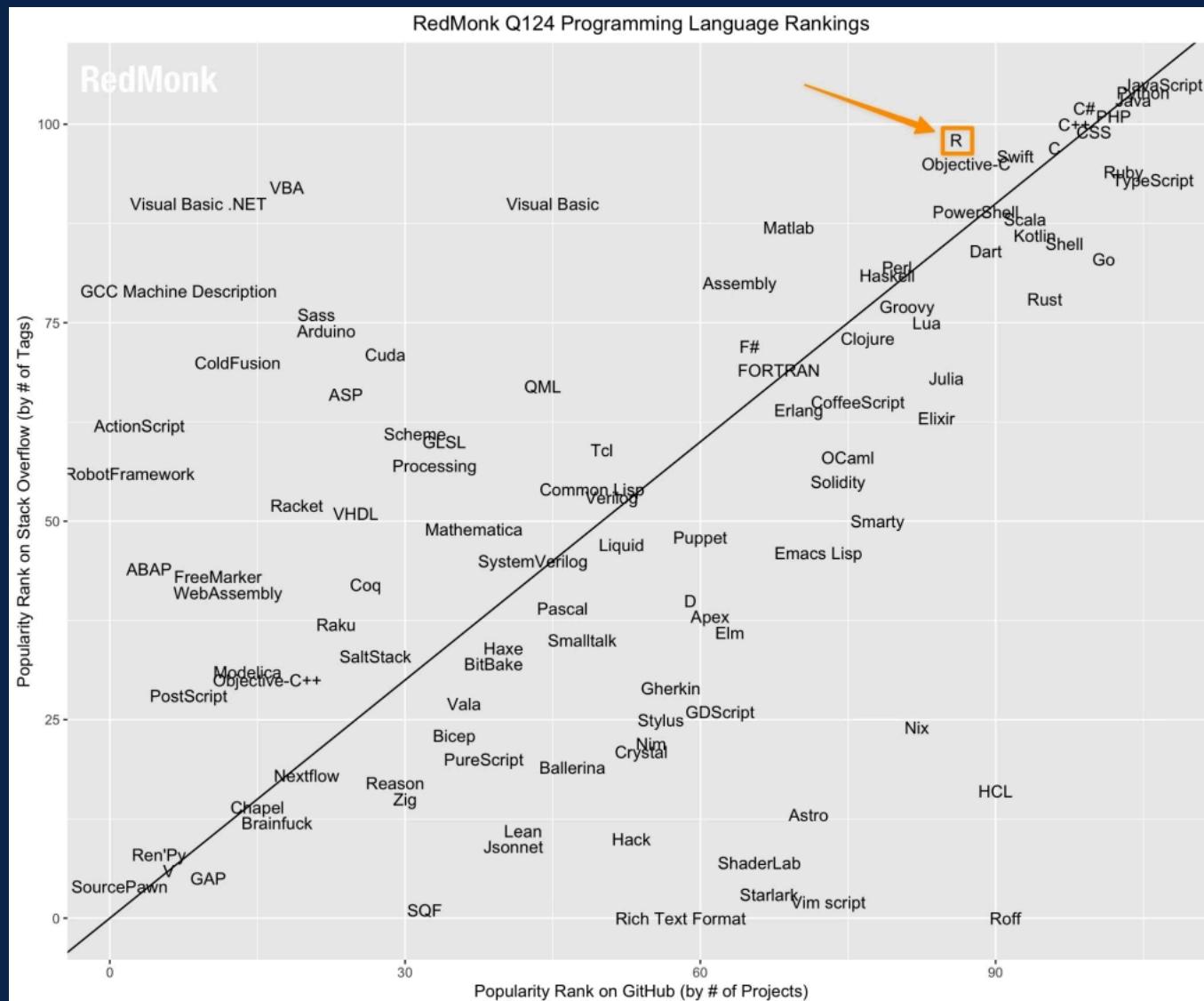
Características - paquetes + usados



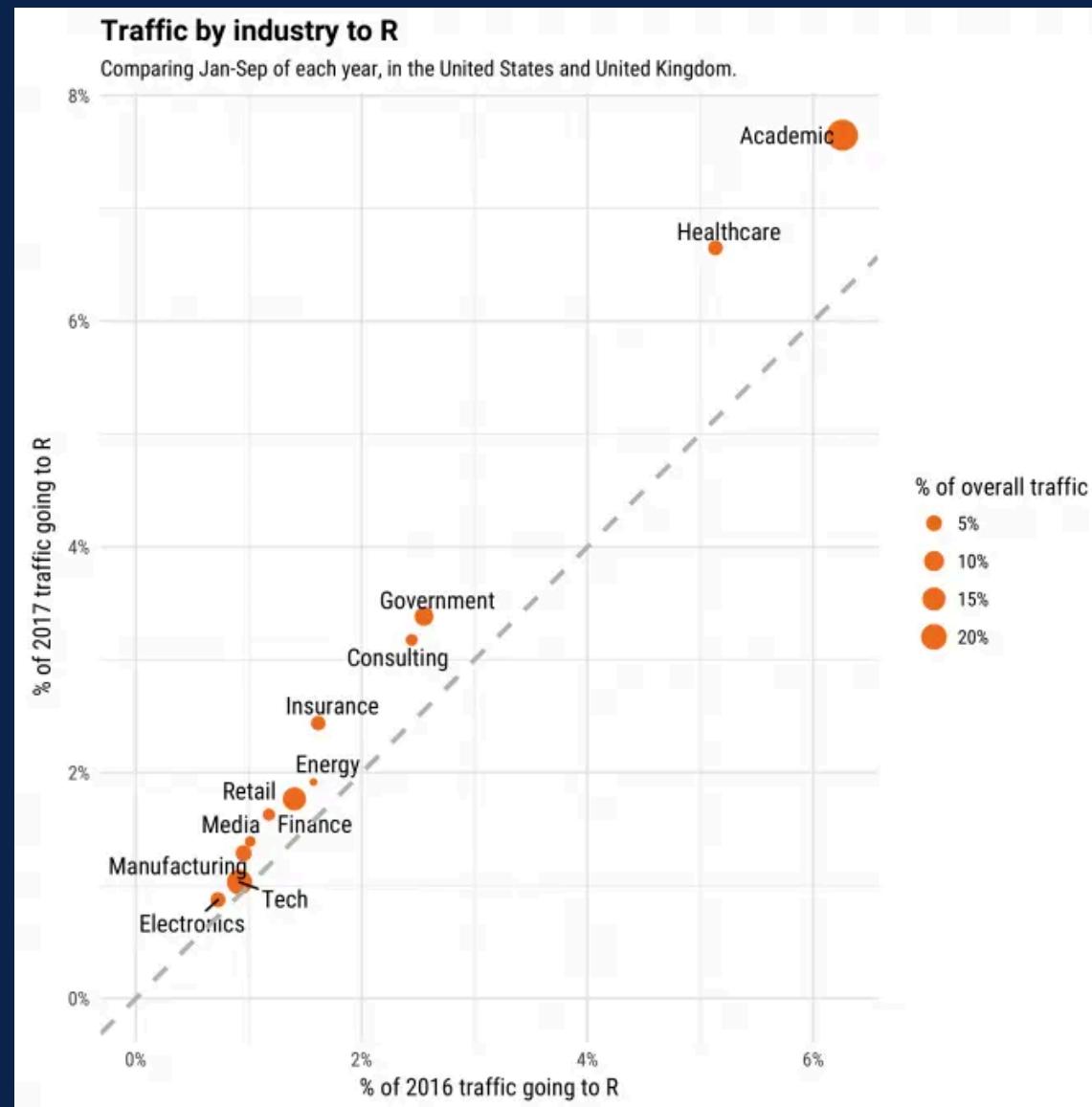
Características - aprendizaje



Características - popularidad



Características - académicas



Datos en R

- Posee 5 tipos básicos de datos:
character: "B", "cbz"
numeric: 9, 25.6 (números reales o decimales)
integer: 40L (números enteros)
logical: TRUE y FALSE (es para datos binarios)
- Adicionalmente, R puede utilizar estos otros tipos de datos:
double: doble precisión
complex: números complejos
raw: almacenar bytes
- Estos tipos de datos se pueden combinar en estructuras de datos: **vectores**.
- Un **vector** puede contener un solo tipo de datos. Es decir, habrá vectores de caracteres, numéricos, etc. Entonces, ¿Puedo reunir datos diferentes para ser analizados en conjunto?

Si, se puede !!

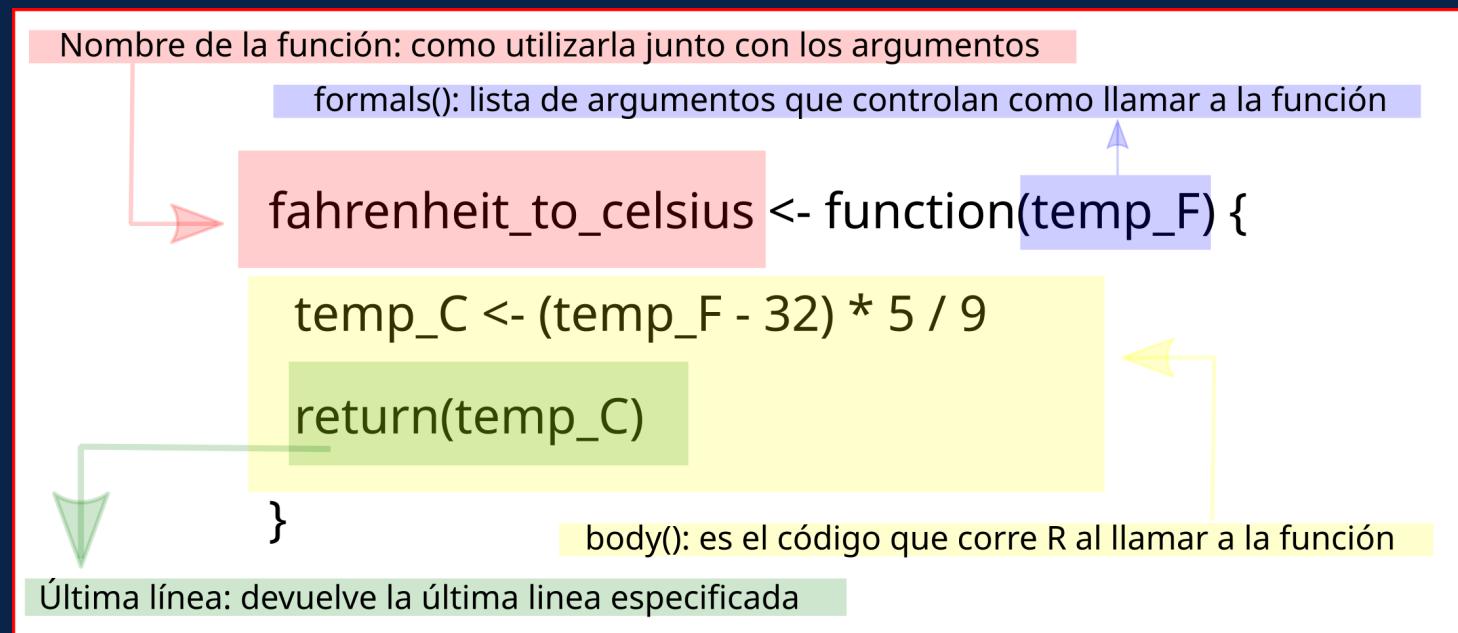


Estructuras de datos en R

- El conjunto de dos o más vectores y su organización forman estructuras de datos más complejas: matrices, listas y dataframes.
- La estructura de datos más común utilizada para análisis en R: **dataframe** o **tibble**.

▲	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.0	1.5	0.2	setosa

Función en R



```
fahrenheit_to_celsius <- function(temp_F) {
  temp_C <- (temp_F - 32) * 5 / 9
  return(paste0(temp_C, " grados celsius"))
}
# Test
fahrenheit_to_celsius(77)
```

[1] "25 grados celsius"

¿Qué es un paquete?

- ¿Qué es un **paquete** o **librería** en R? Un **paquete** es una colección de funciones de R, datos y código compilado. La ubicación en la cual los paquetes son almacenados es llamada **librería**.
- Un **paquete** es una manera adecuada de organizar objetivos y/o acciones a cumplir a través de **funciones** y puede ser compartida con otros.
- Además los **paquetes** incluyen ciertos requisitos como: tests para ver que todo funcione, warnings, datasets propios, ejemplos documentados, documentación sobre las funciones, etc.
- En relación a la especialidad del investigador/a hay áreas o líneas de trabajo que proponen sus propios “Recommended Packages” (e.g., los analistas espaciales).
- Los paquetes “base” al instalar R contiene funciones esenciales y primarias para el funcionamiento de R en cualquier entorno o IDE. Incluye funciones: aritmética, input/output, soporte básico para programación, etc... Ver `library(help = "base")`

R Base

- Por otro lado, en la instalación básica se instalan estos paquetes llamados “base”: base, compiler, datasets, graphics, grDevices, grid, methods, parallel, splines, stats, stats4, tcltk, tools, translations, utils. “Recommended Packages”.
- Al instalar R base ya estará disponible vía terminal en Linux, en el GUI para Windows e incorporado al IDE una vez instalado.
- Con el R base ya se pueden realizar los estadísticos más habituales y plots sin muchos argumentos.
- Datos más específicos en su formato y/o estructura y métodos puntuales usualmente requieren la instalación de paquetes especializados.
- Para un uso real y completo actualmente se necesita siempre la instalación de algún paquete.
- Los argumentos de las funciones fueron diversificándose así como las fuentes de instalación de los paquetes.

Repositorios - Oficial

- **CRAN package repository.** Es una red de servers ftp y web mantenidos por la comunidad de R alrededor del mundo.
- Políticas CRAN para revisión y publicación de los paquetes.

```
# Función base para instalar paquetes  
install.packages()
```

- Argentina mirror. Universidad Nacional de La Plata.
<http://mirror.fcaglp.unlp.edu.ar/CRAN/>
- CRAN con más de 21025 paquetes disponibles.
https://cran.r-project.org/web/packages/available_packages_by_name.html
https://cran.r-project.org/web/packages/available_packages_by_date.html
- Se utilizan los repositorios con funciones como: `install.packages()`, `update`, `remove`, `load()`, `unload()`.

Repositorios - Alternativos

- **Bioconductor.** Es un repositorio específico con orientación en paquetes dirigidos a la bioinformática.
- Políticas propias de sumisión y revisión de los paquetes.

```
# Función para instalar paquetes de Bioconductor  
BiocManager::install()
```

<https://www.bioconductor.org/>

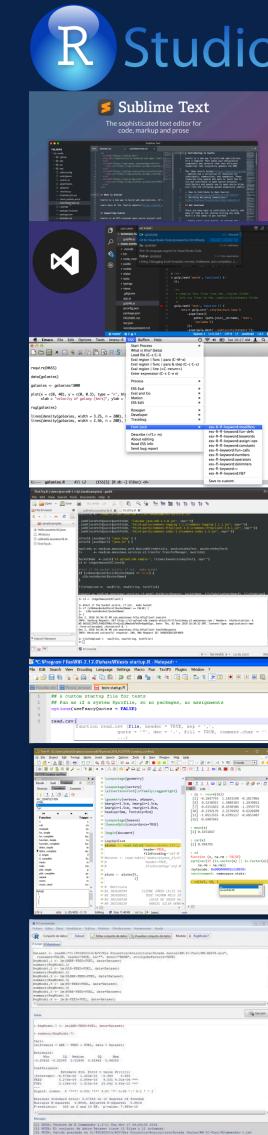
- **Github.** No es específico para R, pero es el la alternativa más utilizada después de CRAN y es un repositorio para proyectos open source completos.
- Muy popular. Espacio ilimitado para el código abierto. Integración con git, control de versiones, fácil de compartir y colaborar con otros. No hay proceso de revisión externo.

```
# Función para instalar paquetes de github  
devtools::install_github()  
remotes::install_github()
```

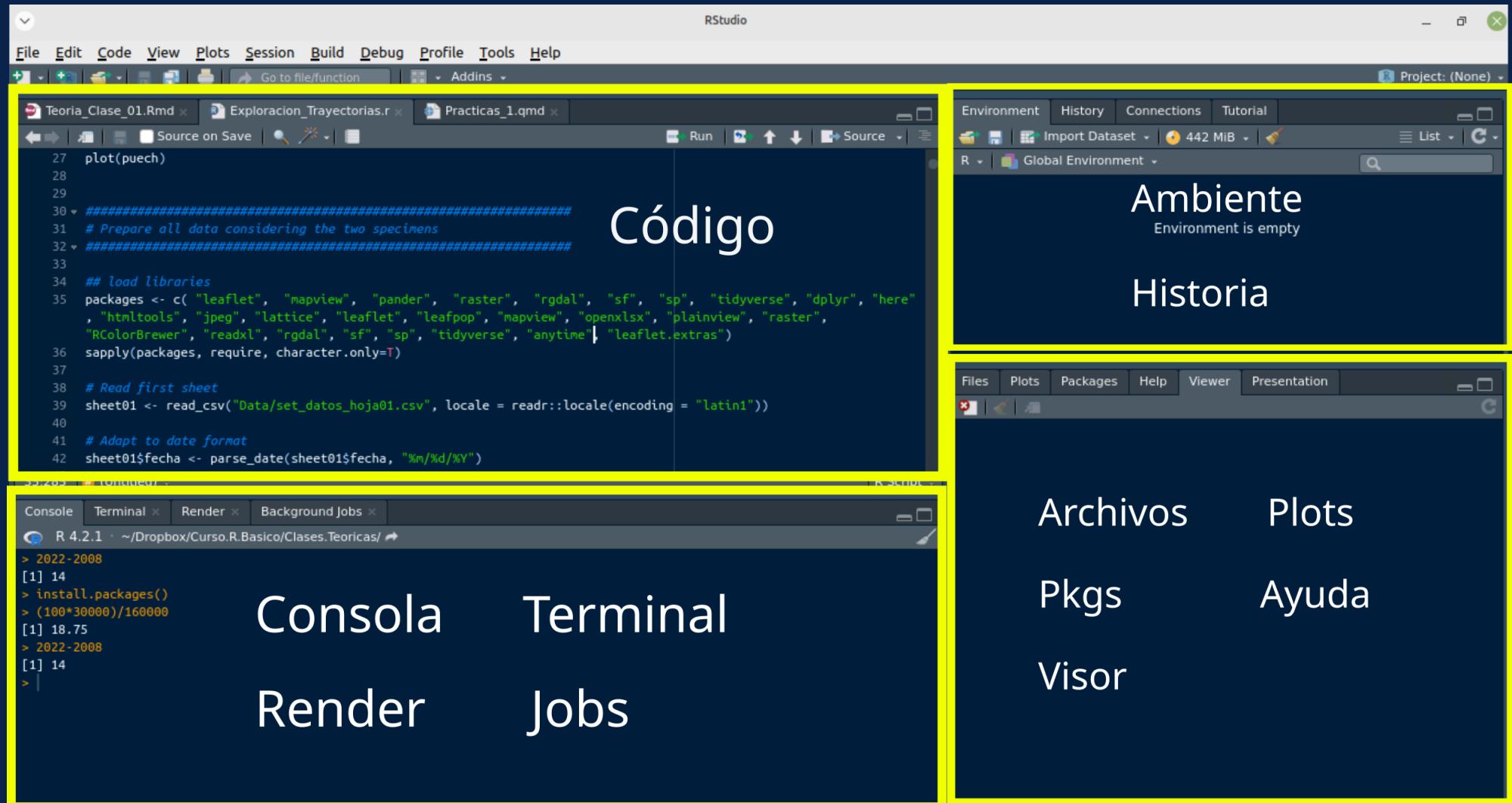
<https://www.github.com/>

Integrated Development Environment

- **RStudio.** Desktop (Windows, Linux, y macOS) y Server (solo en Linux server). Posit - Positron.
- **Visual Studio Code.** Microsoft, gratis y licencia open-source del MIT. Con extensiones.
- **Sublime Text.** Paquetes instalados. Registro.
- **ESS.** Emacs Speaks Statistics. Linux, Unix, Mac y Windows.
- **Gedit.** Editor de texto solo para Linux y Mac. Plugin.
- **Notepad ++.** Editor de texto con funciones avanzadas solo para Windows.
- **Tinn-R.** Editor de texto ASCII & UNICODE solo para Windows.
- **R-commander.** Editor de texto para Windows, Mac y Linux.



RStudio



Características

- Texto predecible. Code Snippets.
- Los objetos son guardados en la sesión y se ven en la ventana “Environment”.
- Permite visualizar los directorios, archivos, etc. que puedan estar relacionados con el script.
- Rstudio hace más accesibles los plots, herramientas interactivas, shortcuts, etc.
- La información relacionada a la sesión (como objetos) puede ser guardada como .RData. Esto permite tener una imagen que contiene los objetos contenidos en el “Environment”.
- Integración directa para proyectos colaborativos y con otros lenguajes de programación.
- sessionInfo().

Sesión de trabajo

```
# Información de la sesión

sessionInfo()

R version 4.3.3 (2024-02-29 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

time zone: America/Buenos_Aires
tzcode source: internal

attached base packages:
[1] stats      graphics   grDevices utils      datasets   methods    base

other attached packages:
[1] leafpop 0.1.0   lubridate 1.9.3 forcats 1.0.0   stringr 1.5.1
```

Documentación - Ayuda

- La ayuda y la documentación es imprescindible en R.

```
library(RDocumentation)  
help(package = "ggplot2")
```

- La ayuda puede buscarse dentro de R mismo: paquetes y funciones.

- **Paquetes**

```
help(ggplot2)  
?ggplot2  
??ggplot2
```

- **Funciones**

```
?barplot  
help("barplot")  
help("barplot", package = "graphics")  
help("barplot", package = "terra")
```

Documentación - Ayuda

- La documentación es fundamental para entender completamente el uso, funcionamiento, configuración e interpretación de resultados.
- Base R proporciona una forma estándar de documentar un paquete donde cada tema de documentación.
- Estos archivos usan una sintaxis personalizada, basada libremente en LaTeX, que se procesan en HTML, texto sin formato o pdf, según sea necesario, para su visualización.
- **Vignettes**: Es una guía larga y didáctica para el uso de funciones de un paquete.
- Un vignette es como un capítulo de un libro o un artículo académico. Puede describir el problema para el cual ese paquete está diseñado a resolver y luego mostrarle al lector cómo resolverlo e interpretarlo.
- **Vignettes**

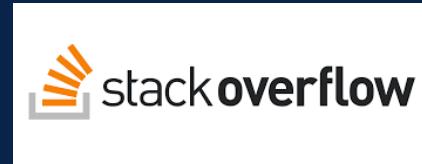
```
browseVignettes()  
browseVignettes("ggplot2")
```

Reproducibilidad en R

- Idioma -> Inglés. Encabezados del script.
- Sintaxis. R base versus Tidyverse.
- Ejemplos con datos reproducibles. Ayuda.
- Puesta a prueba -> pequeño set de datos.
- Optimización del código.
- Loops y for para tener menos líneas de código.
- Saltos de linea, pipes, etc.
- Nomencultura en nombres de objetos.
- Trabajar con rutas absolutas o relativas.
- Guardado de una imagen del workspace.

Reciclado en R

- Sitios de ayuda. Problemas similares. <https://stackoverflow.com/>



- Vignette, webs y blogs con tutoriales. <https://www.r-bloggers.com/>



- Biblioteca personal de scripts.
- Funciones propias.
- Partes dentro de un mismo script.
- Scripts en por partes de un proceso.

Citas

- Usualmente se cita el R con su versión (que ya trae paquetes del R Base asociado y sus versiones) y los paquetes más importantes o imprescindibles para el paper, libro, informe, etc. Vuelvo a recurrir a sessionInfo().

```
# Función para citar un paquete
citation("dplyr")
```

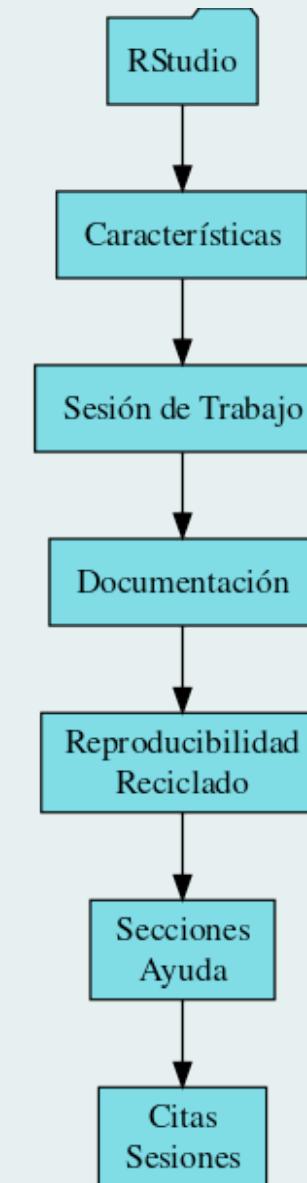
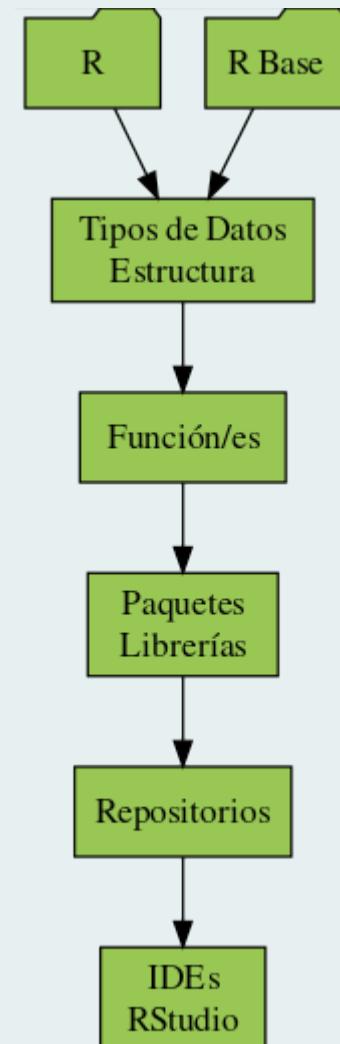
To cite package 'dplyr' in publications use:

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). `_dplyr: A Grammar of Data Manipulation_`. R package version 1.1.4, <<https://CRAN.R-project.org/package=dplyr>>.

A BibTeX entry for LaTeX users is

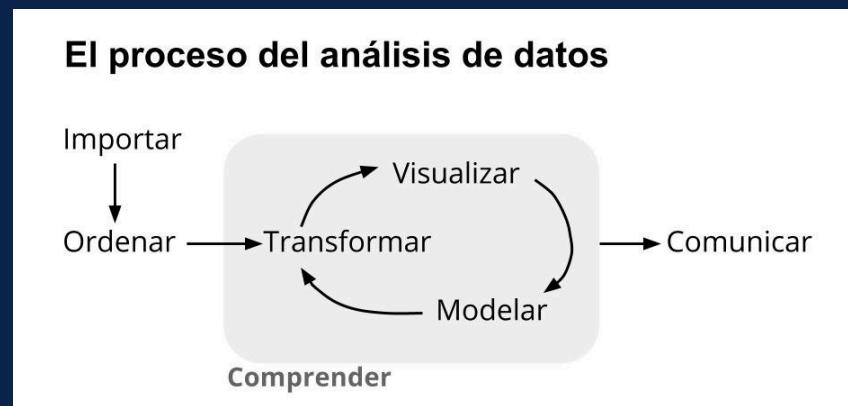
```
@Manual{,
  title = {dplyr: A Grammar of Data Manipulation},
  author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller and Davis Vaughan},
  year = {2023},
  note = {R package version 1.1.4},
  url = {https://CRAN.R-project.org/package=dplyr},
}
```

Resumen



Costos & Beneficios: programación

- Si, por supuesto que los costos-beneficios valen la pena y la curva de aprendizaje es diferente en lenguajes - personas.
- Poder de creación gratuito, innovador, colaborativo, escalable, reproducible, transparente, continuo e ilimitado...



Importancia de la programación

- Los resultados y productos expuestos anteriormente no serían posibles sin la programación...
- La idea de este curso es entender cada una de las acciones realizadas con la programación en R y aprovechar todo su poder potencial.



@DINOMITRON

QR - presentación

PDF Presentación: Clase 01 - Teoría



Fin. Clase 01 - Teoría

¡¡¡ Muchas gracias !!!

¿ Preguntas ? ... ¿ Consultas ?

