



Introducción a R

Un viaje a la Ciencia de Datos

William Hoyos

26 de mayo de 2020

¿Quien soy?



- William Hoyos
- Estudiante doctoral (EAFIT)
- Intereses:
 - Predicción de Enfermedades con Inteligencia Artificial
- Contacto:
 - wshoyosm@eafit.edu.co
 - Twitter: @whoyos21

Contenido

- Historia de R
- Descarga e Instalación
- Paquetes de R
- La Calculadora
- Estructuras de Datos
- Importar Datos

Historia de R

Historia de R ¹


- Desarrollado inicialmente por Robert Gentleman y Ross Ihaka
- Departamento de Estadística de la Universidad de Auckland en 1993
- IBM (Lenguaje S)
- *R Development Core Team*
- Proyecto colaborativo y abierto
- Mas de 15.000 paquetes a la fecha

¹: [https://es.wikipedia.org/wiki/R_\(lenguaje_de_programaci%C3%B3n\)](https://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n))

Descarga R y RStudio

Descarga de R

- Descarga de R <https://ftp.osuosl.org/pub/cran/>
- Click “Download R for (Mac) OS X” or “Download R for Windows”
- Descarga la última versión



[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 (2020-04-24, Arbor Day) [R-4.0.0.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](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear

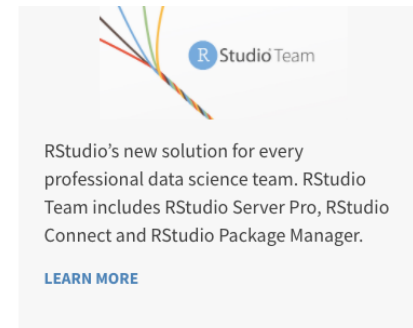
Descarga de RStudio

- Descarga de RStudio
- <https://rstudio.com/products/rstudio/download/>
- Click “Download”
- Descarga la última versión

Choose Your Version

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

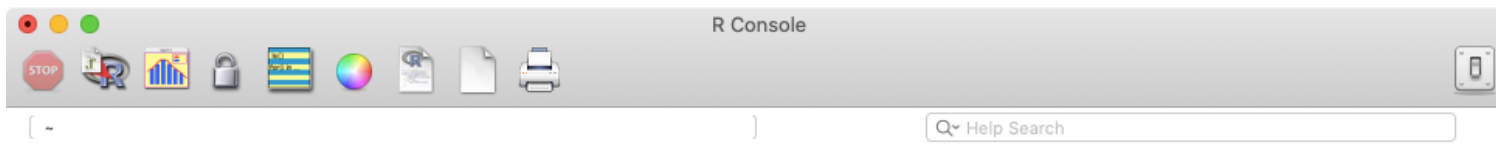
[LEARN MORE ABOUT RSTUDIO FEATURES](#)



RStudio Desktop	RStudio Desktop	RStudio Server	RStudio Server Pro
Open Source License	Commercial License	Open Source License	Commercial License
Free	\$995 /year	Free	\$4,975 /year (5 Named Users)
DOWNLOAD	BUY	DOWNLOAD	BUY
Learn more	Learn more	Learn more	Evaluation Learn more
Integrated Tools for R	✓	✓	✓

Estructura de R y RStudio

Estructura de R



```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (64-bit)

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

Natural language support but running in an English locale

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.

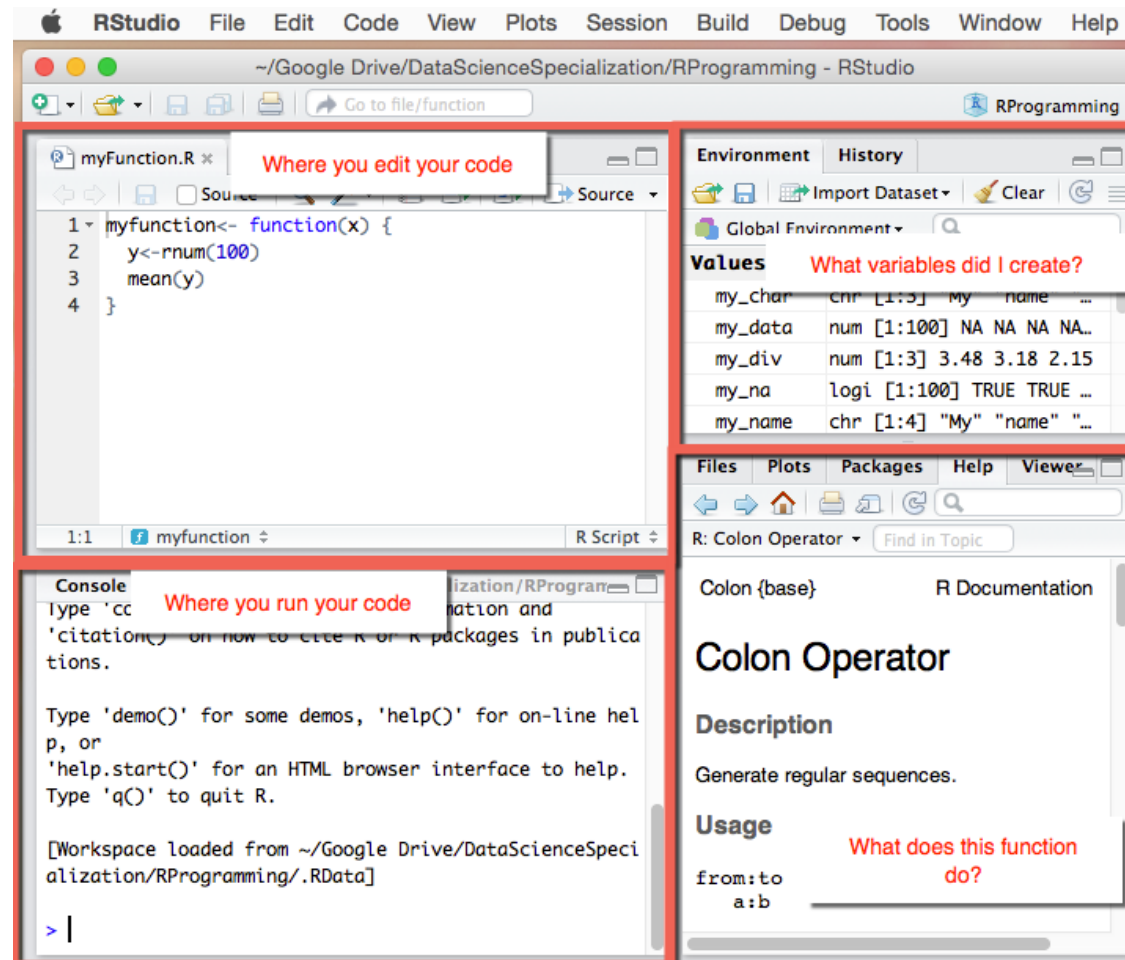
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.70 (7612) x86_64-apple-darwin15.6.0]

[History restored from /Users/whoyos21/.Rapp.history]

>
```

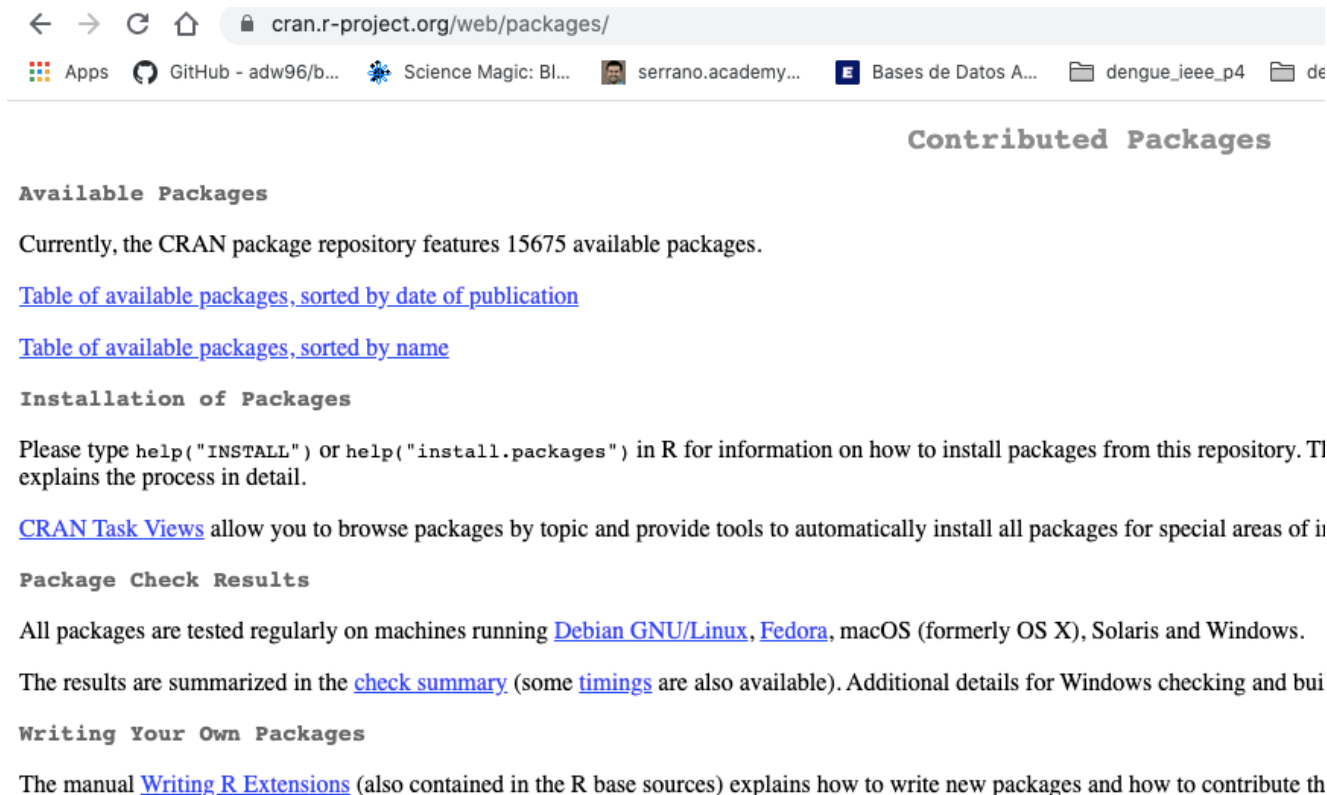
Estructura de RStudio



Los paquetes: la principal
herramienta de R

Paquetes

- Paquete base de R
- Conjunto de funciones para realizar una tarea específica
- También llamadas librerías



The screenshot shows the CRAN website at cran.r-project.org/web/packages/. The browser's address bar and tabs are visible at the top. The main heading is "Contributed Packages". Below it, the section "Available Packages" states that there are 15675 available packages and provides two links: "Table of available packages, sorted by date of publication" and "Table of available packages, sorted by name". The "Installation of Packages" section instructs users to use `help("INSTALL")` or `help("install.packages")` in R. The "CRAN Task Views" section mentions tools for installing packages for specific areas. The "Package Check Results" section notes that packages are tested on various operating systems and provides a link to the "check summary". The "Writing Your Own Packages" section refers to the manual "Writing R Extensions".

← → ↻ 🏠 cran.r-project.org/web/packages/

Apps GitHub - adw96/b... Science Magic: Bl... serrano.academy... Bases de Datos A... dengue_ieee_p4 de

Contributed Packages

Available Packages

Currently, the CRAN package repository features 15675 available packages.

[Table of available packages, sorted by date of publication](#)

[Table of available packages, sorted by name](#)

Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this repository. It explains the process in detail.

[CRAN Task Views](#) allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest.

Package Check Results

All packages are tested regularly on machines running [Debian GNU/Linux](#), [Fedora](#), macOS (formerly OS X), Solaris and Windows.

The results are summarized in the [check summary](#) (some [timings](#) are also available). Additional details for Windows checking and building are available in the [Windows FAQ](#).

Writing Your Own Packages

The manual [Writing R Extensions](#) (also contained in the R base sources) explains how to write new packages and how to contribute them.

Las funciones

Funciones en R

- Una función permite automatizar tareas de forma mas poderosa
- Vienen en los paquetes
- Pueden ser creadas por cualquier mortal
- Ejemplos:

Función que instala paquetes:

```
#install.packages("ggplot2") #Ojo las comillas
```

Función que carga el paquete a R

```
#library(ggplot2) #No es necesario comillas
```

R como una calculadora

R como una calculadora

```
5 + 7 #Pulsamos Enter y muestra el resultado
```

```
## [1] 12
```

```
pi
```

```
## [1] 3.141593
```

```
exp(2)
```

```
## [1] 7.389056
```

R como una calculadora

```
5 + 7.5 #Suma
```

```
## [1] 12.5
```

```
12 - 5.1 #Diferencia
```

```
## [1] 6.9
```

```
7 * 5 #Multiplicación
```

```
## [1] 35
```

```
18 / 6 #División
```

```
## [1] 3
```

R como una calculadora

Expresión matemática: $2(3+5)$

```
2 * (3 + 5)
```

```
## [1] 16
```

Funciones numéricas

Función	\sqrt{x}	e^x	$\ln(x)$	$\log_{10}(x)$	$\log_a(x)$	$n!$	$\binom{n}{m}$
Signo	<code>sqrt</code>	<code>exp</code>	<code>log</code>	<code>log10</code>	<code>log(, a)</code>	<code>factorial</code>	<code>choose</code>
Función	$\sin(x)$	$\cos(x)$	$\tan(x)$	$\arcsin(x)$	$\arccos(x)$	$\arctan(x)$	$ x $
Signo	<code>sin</code>	<code>cos</code>	<code>tan</code>	<code>asin</code>	<code>acos</code>	<code>atan</code>	<code>abs</code>

Objetos en R

Objetos en R

- R almacena los datos como objetos
- Se usa el símbolo <- para asignar
- Ejemplos:

```
x <- 5;x
```

```
## [1] 5
```

```
y <- 7;y
```

```
## [1] 7
```

```
z <- exp(x);z
```

```
## [1] 148.4132
```

Tipos de datos

Tipos de datos

- `logical` (lógicos: TRUE = verdadero; FALSE = falso)

```
25 > 5
```

```
## [1] TRUE
```

```
class(25 > 5) #¿De que clase es el dato resultante?
```

```
## [1] "logical"
```

- `integer` (números enteros)

```
class(7)
```

```
## [1] "numeric"
```

Tipos de datos

- `numeric` (números reales)

```
class(7.5)
```

```
## [1] "numeric"
```

- `character` (palabras)

```
class("Hola mundo")
```

```
## [1] "character"
```


Estructuras de datos

Vectores

- Conjunto de datos del mismo tipo
- Una dimensión
- Nombres de 5 personas
 - "Alvaro", "Alicia", "Pedro", "Jose", "Martha"

#Función c() que combina

```
nombres <- c("Alvaro", "Alicia", "Pedro", "Jose", "Martha")
```

```
nombres
```

```
## [1] "Alvaro" "Alicia" "Pedro"  "Jose"   "Martha"
```

```
class(nombres)
```

```
## [1] "character"
```

Vectores

- Edades de las 5 personas
 - 32, 12, 54, 76, 34

```
#Función c() que combina  
edades <- c(32, 12, 54, 76, 34)  
edades
```

```
## [1] 32 12 54 76 34
```

```
class(edades)
```

```
## [1] "numeric"
```

Vectores

- Creación de vectores de secuencias

#Secuencia desde el 0 hasta el 50

```
seq(from = 0, to = 50)
```

```
## [1] 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
## [26] 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
## [51] 50
```

#Secuencia del 0 hasta el 50, de 10 en 10

```
seq(from = 0, to = 50, by = 10)
```

```
## [1] 0 10 20 30 40 50
```

Operaciones con vectores

Suma de vectores

```
vector1 <- c(1, 2, 3, 4, 5)  
vector2 <- c(5, 10, 15, 20, 25)
```

```
vector3 <- vector1 + vector2  
vector3
```

```
## [1]  6 12 18 24 30
```

Operaciones con vectores

Subconjuntos de vectores

#Vector con seis números aleatorios

```
mi_vector <- c(12, 23, 34, 56, 75, 98)
```

#Seleccionar el número en la posición 5

```
mi_vector[5]
```

```
## [1] 75
```

#Seleccionar los primeros 3 números

```
mi_vector[1:3]
```

```
## [1] 12 23 34
```

Operaciones con vectores

Suma de vectores

```
vector1 <- c(1, 2, 3, 4, 5)
```

```
vector2 <- c(5, 10)
```

```
vector3 <- vector1 + vector2
```

```
## Warning in vector1 + vector2: longer object length is not a multiple of shorter
```

```
## object length
```

```
vector3
```

```
## [1]  6 12  8 14 10
```

Matrices

- Los datos deben ser del mismo tipo
- Tienen dos dimensiones (filas y columnas)
- Ejemplo: Matriz de 3 filas y 2 columnas

#Matriz de 3 x 3

```
mi_matriz <- matrix(data = c(1, 2, 3, 4, 5, 6), nrow = 3, ncol = 2, byrow = TRUE)
```

#Nombres de columnas y filas

```
colnames(mi_matriz) <- c("columna1", "columna2")
```

```
rownames(mi_matriz) <- c("fila1", "fila2", "fila3")
```

```
mi_matriz
```

```
##      columna1 columna2
## fila1         1         2
## fila2         3         4
## fila3         5         6
```


Operaciones con matrices

Subconjuntos de matrices

#Matriz con 6 números aleatorios

```
mi_matriz <- matrix(data = c(1, 2, 3, 4, 5, 6), nrow = 3, ncol = 2, byrow = TRUE)
```

#Nombres de columnas y filas

```
colnames(mi_matriz) <- c("columna1", "columna2")
```

```
rownames(mi_matriz) <- c("fila1", "fila2", "fila3")
```

```
mi_matriz
```

```
##      columna1 columna2
## fila1         1         2
## fila2         3         4
## fila3         5         6
```

#Seleccionar el número de la fila 2 y la columna 1

```
mi_matriz[2, 1]
```

```
## [1] 3
```

Factores

- Son usados para variables categóricas
- Las categorías son llamados niveles
- Ejemplo:
- Sexos de 5 personas
 - M, F, F, M, F

```
sexo <- factor(c("M", "M", "F", "M", "F"))  
sexo
```

```
## [1] M M F M F  
## Levels: F M
```

Dataframes (Conjuntos de datos)

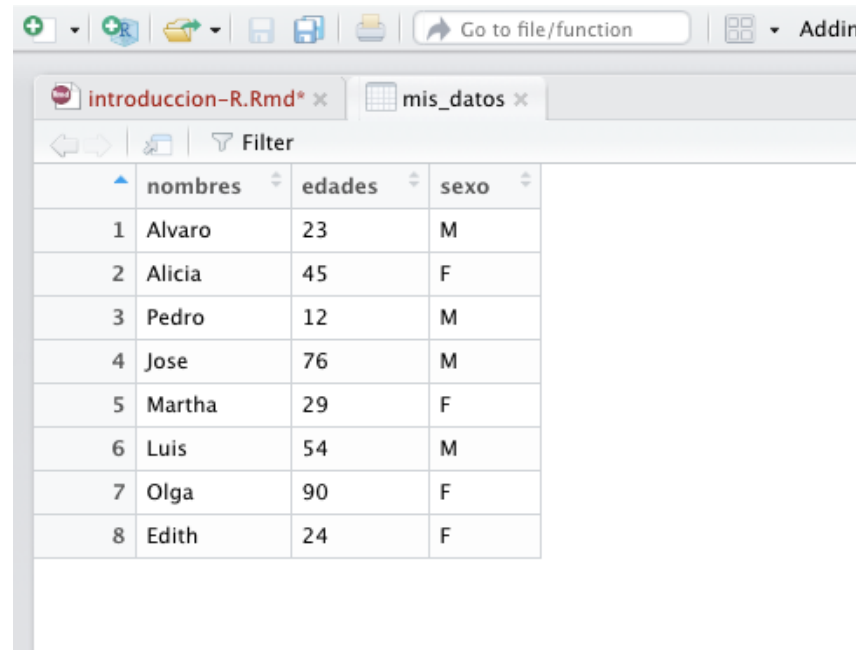
- Las columnas son variables
- Las filas son registros
- Ejemplo:

The diagram illustrates a dataframe structure. A table is shown with a header row and four data rows. Red dashed boxes highlight the header and the first column. Blue text labels with red arrows point to these elements: 'Nombres Filas' points to the first column, 'Nombres Columnas' points to the header row, and 'Columnas' points to the header cells. 'Filas' points to the first four rows.

	Nombre	Edad	Grado	Correo
1	María	18	Economía	maria@gmail.com
2	Luis	22	Medicina	luis@yahoo.es
3	Carmen	20	Arquitectura	carmen@gmail.com
4	Antonio	21	Economía	antonio@gmail.com

Función View()

```
nombres <- c("Alvaro", "Alicia", "Pedro", "Jose", "Martha", "Luis", "Olga", "Edith")
edades <- c(23, 45, 12, 76, 29, 54, 90, 24)
sexo <- c("M", "F", "M", "M", "F", "M", "F", "F")
mis_datos <- data.frame(nombres, edades, sexo)
#View(mis_datos)
```



The screenshot shows the RStudio interface with the 'mis_datos' data frame open in the View window. The table contains the following data:

	nombres	edades	sexo
1	Alvaro	23	M
2	Alicia	45	F
3	Pedro	12	M
4	Jose	76	M
5	Martha	29	F
6	Luis	54	M
7	Olga	90	F
8	Edith	24	F

Función head()

```
mis_datos <- data.frame(nombres, edades, sexo)  
head(mis_datos)
```

```
##  nombres edades sexo  
## 1  Alvaro    23    M  
## 2  Alicia    45    F  
## 3   Pedro    12    M  
## 4    Jose    76    M  
## 5  Martha    29    F  
## 6    Luis    54    M
```

Función tail()

```
mis_datos <- data.frame(nombres, edades, sexo)  
tail(mis_datos)
```

```
##   nombres edades sexo  
## 3   Pedro    12    M  
## 4    Jose    76    M  
## 5  Martha    29    F  
## 6    Luis    54    M  
## 7    Olga    90    F  
## 8   Edith    24    F
```

Función dim()

```
mis_datos <- data.frame(nombres, edades, sexo)  
dim(mis_datos)
```

```
## [1] 8 3
```

Acceder a columnas de un dataframe

```
(mis_datos <- data.frame(nombres, edades, sexo))
```

```
##   nombres edades sexo
## 1  Alvaro    23    M
## 2  Alicia    45    F
## 3   Pedro    12    M
## 4    Jose    76    M
## 5  Martha    29    F
## 6    Luis    54    M
## 7    Olga    90    F
## 8   Edith    24    F
```

```
mis_datos$edades
```

```
## [1] 23 45 12 76 29 54 90 24
```


Acceder a columnas de un dataframe

```
(mis_datos <- data.frame(nombres, edades, sexo))
```

```
##   nombres edades sexo
## 1  Alvaro     23    M
## 2  Alicia     45    F
## 3   Pedro     12    M
## 4    Jose     76    M
## 5  Martha     29    F
## 6   Luis     54    M
## 7   Olga     90    F
## 8  Edith     24    F
```

```
mis_datos[, 2]
```

```
## [1] 23 45 12 76 29 54 90 24
```

Importar datos a R

Espacio de trabajo en R

- No recomendable

#Obtener la carpeta de espacio de trabajo

```
getwd()
```

```
## [1] "/Users/whoyos21/Dropbox/William/Webinars_Ciencia_de_Datos/Introduccion_R"
```

#Configurar una carpeta como espacio de trabajo

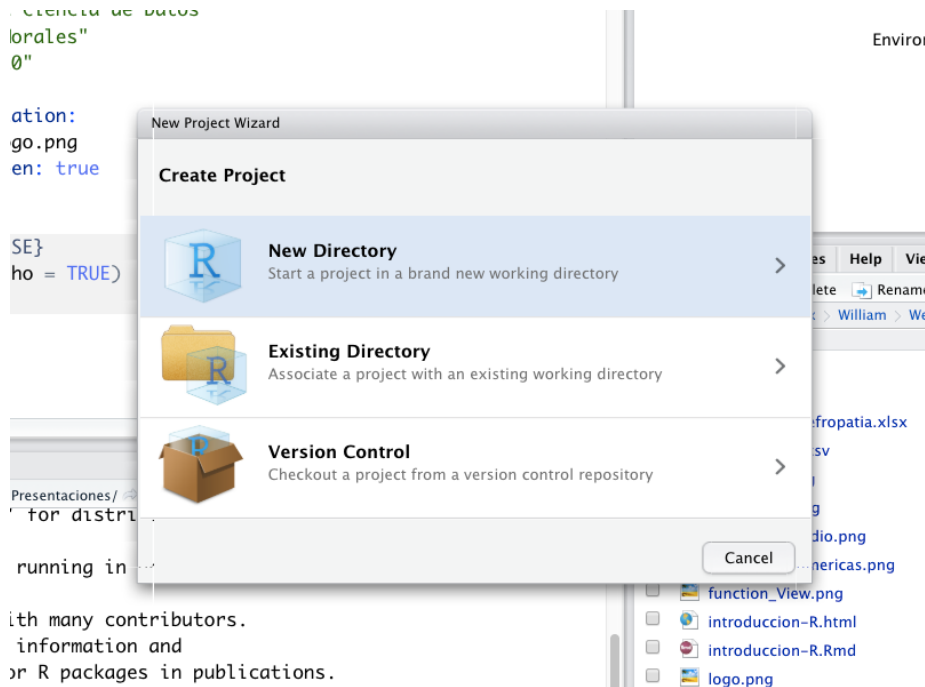
```
setwd("/Users/whoyos21/Desktop")
```

#Mostrar nueva carpeta de trabajo

```
getwd()
```

```
## [1] "/Users/whoyos21/Desktop"
```

Proyectos de R



- Altamente recomendable
- Acceso directo al proyecto
- Todos los archivos disponibles
- Sin rutas extensas

Importar csv

#Función read.csv() del paquete base

```
covid_colombia <- read.csv("covid_colombia.csv")
```

```
head(covid_colombia, 3)
```

```
## ID.de.caso Fecha.de.notificación Codigo.DIVIPOLA Ciudad.de.ubicación
## 1          1 2020-03-02T00:00:00.000          11001          Bogotá D.C.
## 2          2 2020-03-06T00:00:00.000          76111 Guadalajara de Buga
## 3          3 2020-03-07T00:00:00.000          5001          Medellín
## Departamento.o.Distrito atención Edad Sexo      Tipo Estado
## 1          Bogotá D.C. Recuperado   19    F Importado  Leve
## 2          Valle del Cauca Recuperado  34    M Importado  Leve
## 3          Antioquia Recuperado   50    F Importado  Leve
## País.de.procedencia          FIS Fecha.de.muerte
## 1          Italia 2020-02-27T00:00:00.000          -    -
## 2          España 2020-03-04T00:00:00.000          -    -
## 3          España 2020-02-29T00:00:00.000          -    -
##          Fecha.diagnostico          Fecha.recuperado          fecha.reporte.web
## 1 2020-03-06T00:00:00.000 2020-03-13T00:00:00.000 2020-03-06T00:00:00.000
## 2 2020-03-09T00:00:00.000 2020-03-19T00:00:00.000 2020-03-09T00:00:00.000
## 3 2020-03-09T00:00:00.000 2020-03-15T00:00:00.000 2020-03-09T00:00:00.000
```

Importar csv

El paquete readr también puede ser usado

```
library(readr)
covid_colombia <- read_csv("covid_colombia.csv")

## Parsed with column specification:
## cols(
##   `ID de caso` = col_double(),
##   `Fecha de notificación` = col_datetime(format = ""),
##   `Codigo DIVIPOLA` = col_double(),
##   `Ciudad de ubicación` = col_character(),
##   `Departamento o Distrito` = col_character(),
##   atención = col_character(),
##   Edad = col_double(),
##   Sexo = col_character(),
##   Tipo = col_character(),
##   Estado = col_character(),
##   `País de procedencia` = col_character(),
##   FIS = col_character(),
##   `Fecha de muerte` = col_character(),
##   `Fecha diagnostico` = col_datetime(format = ""),
##   `Fecha recuperado` = col_character(),
##   `fecha reporte web` = col_datetime(format = "")
## )
```

Importar xlsx

- Es necesario instalar una librería

```
library(readxl)
nefropatia <- read_excel("base_datos_nefropatia.xlsx")
head(nefropatia, 3)
```

```
## # A tibble: 3 x 8
```

```
##   anio  edad sexo municipio estancia hipertenso estado departamento
##   <dbl> <dbl> <chr> <chr>          <dbl>      <dbl> <dbl> <chr>
## 1  2019    10  1    Montería         6          0      1 Córdoba
## 2  2019    46  2    Tierralta         8          1      1 Córdoba
## 3  2019    52  2    Montería        20          1      1 Córdoba
```

Demo