

# Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS  
MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

## Tema 2. Estadística Computacional

# Instalación de R

Para instalar estos programas se puede usar lo siguiente:

## Instalar R para Windows:

<https://cran.r-project.org/bin/windows/base/>

## Instalar R para MacOS:



<https://cran.r-project.org/bin/macosx/>

## Instalar Rstudio para Windows o MacOS:

Descarga e instala la última versión que corresponda.

<https://www.rstudio.com/products/rstudio/download/>

Free-Open-Source

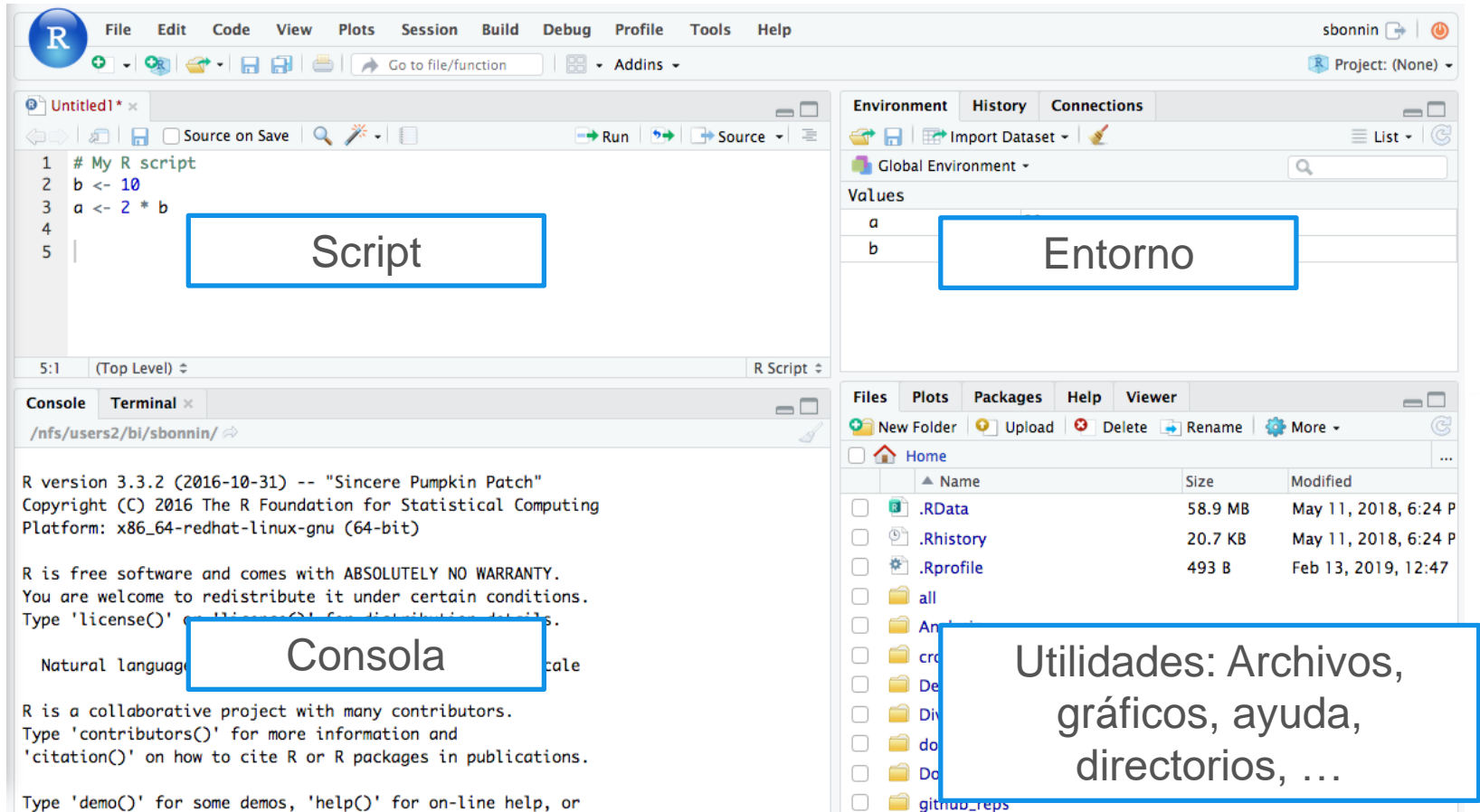
OS	Download	Size	SHA-256
Windows 10/11	 RStudio-2022.07.2-576.exe	190.49 MB	b38bf925
macOS 10.15+	 RStudio-2022.07.2-576.dmg	224.49 MB	35028d02

# Ideas básicas (R y Rstudio):



- R es un tipo de lenguaje de programación, pero Rstudio es un entorno de desarrollo integrado (IDE).
- R funciona de forma independiente, pero RStudio debe funcionar solo con el lenguaje R.
- La interfaz de usuario (GUI, Graphical User Interface) de R no es muy amigable ni versátil, así que interactuaremos con R a través de RStudio.
- RStudio es un programa que nos permitirá interactuar con R de forma más amigable

# Rstudio:

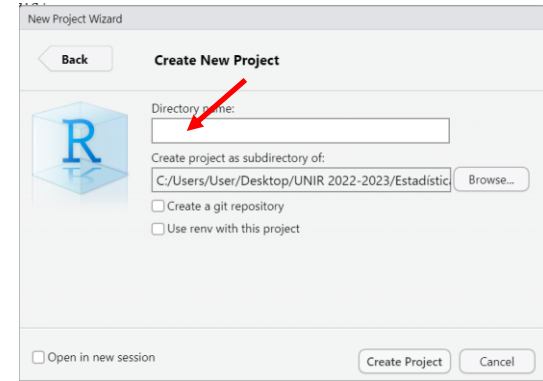
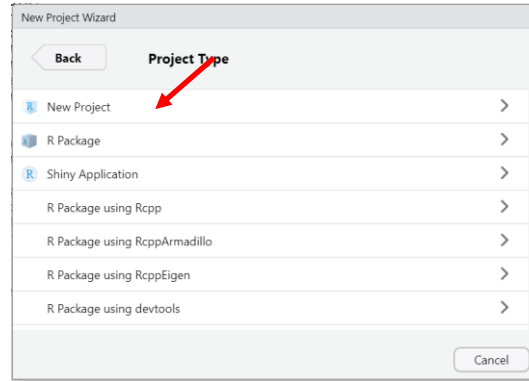
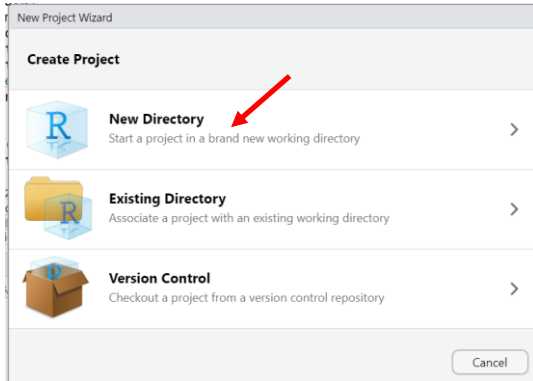


Una vez estamos en RStudio, podemos escribir y ejecutar las órdenes:

- directamente en la consola (código es interpretado)
- a través de un script (.R)
- con ficheros Rmarkdown (.Rmd)

# Proyectos en Rstudio:

Al crear un proyecto todos los ficheros quedan vinculados directamente al proyecto. Para crear un proyecto selección **File > New project...** Se abrirá la siguiente ventana:



- Asignamos un nombre al directorio (carpeta) que se va a crear y que al mismo tiempo será el nombre del proyecto de R. Para terminar, hacemos clic en el botón Create Project. Al seguir este proceso se habrá creado una carpeta en Documentos y un archivo con nombre\_carpeta.Rproj.
- Para crear un proyecto en una carpeta que ya existe, seleccionamos la carpeta ayudándonos del Browse si fuera necesario. Una vez elegida la carpeta, clicamos en Create Project.
- Para abrir un proyecto hacemos doble clic sobre el archivo con extensión .Rproj o lo abrimos desde el menú de RStudio: File > Open Project > ...

Cualquier archivo que creamos (script de R, documento de Rmarkdown, history, etc.) se guardará en la carpeta del proyecto.

# Instalar y utilizar paquetes:

```
install.packages("moments")
install.packages("tidyverse")

library(dplyr) #librería para la manipulación de datos
library(tidyr) #librería para la manipulación de datos
library(ggplot2) #librería gráficos especializados

library(moments) #calculo de sesgo y otros estadísticos

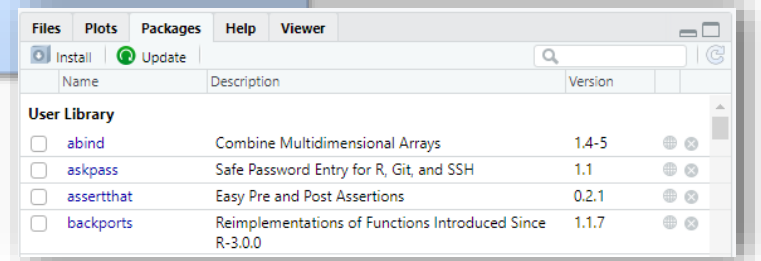
#presiona Ctrl + L #Borramos la consola
```



Hay que instalar antes  
el paquete



Usa el gestor de paquetes



Con **tydiverse** se instala un conjunto de paquetes que trabajan en armonía para la ciencia de datos:

- readr: para importar datos
- tidyr: para convertir los datos a tidy data
- dplyr: para manipular datos
- ggplot2: para hacer gráficos
- ...

# Operador pipe (%>%):

Pasa el elemento que está a su izquierda como un argumento de la función que tiene a la derecha;

f(object, argumentos de la función)	ES EQUIVALENTE a	object %>% f(argumentos de la función)
-------------------------------------	------------------	--

head(iris, n = 4)

iris %>% head(n = 4) #- %>% pasa lo que hay a la derecha como argumento de la función

Lo importante es que **las pipes se pueden encadenar.**

```
df %>% filter(X1 > 400) %>% group_by(X2) %>% summarise(media = mean(X3))
```

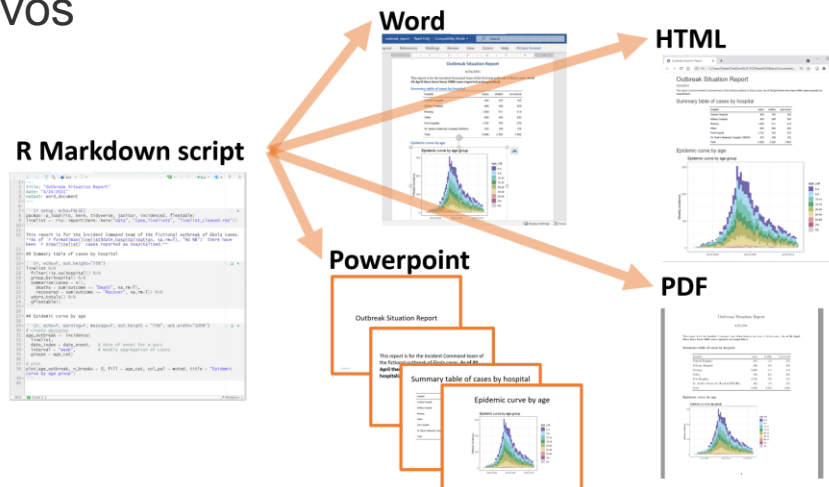
# R Markdown:

## Documentos RMarkdown (.Rmd).

Permite la elaboración de informes incluyendo:

1. Escritura de texto
2. Escritura de texto matemático
3. Realizar los cálculos estadísticos y gráficos. Código +evaluación de código
4. Pegar los gráficos y tablas en el documento de texto.
5. Crear documentos interactivos

...



<https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>

<https://bookdown.org/yihui/rmarkdown/>

<https://www.uv.es/vcoll/RMarkdown.html>



# Datos de ejemplo:

- Data Lending Club:  
<https://www.kaggle.com/wordsforthewise/lending-club>.  
Problema: Factores que determinan el Default en los créditos. Modelo de riesgo  
Obligaciones: 2015-2018
- Data desempeño de estudiantes en diferentes asignaturas:  
<https://www.kaggle.com/spscientist/students-performance-in-exams>  
Problema: Evaluar diferencias y encontrar determinantes de las notas de los estudiantes para las diferentes competencias evaluadas.
- Data **gapminder** contiene un fichero de datos de población, esperanza de vida y renta per cápita de los países del mundo entre 1952 y 2007.  
La fundación Gapminder es una organización sin fines de lucro con sede en Suecia que promueve el desarrollo global mediante el uso de estadísticas.  
`library(gapminder)`
- Data Covid-19:  
<https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide>

# Gráficas básicas:

Algunas funciones que usaremos:

**plot()** La función básica de R para hacer gráficos

**hist()** Histogramas

**boxplot()** Gráficos de caja y bigotes

**ggplot()** comando para generar el sistema de coordenadas dentro de las funcionalidades del paquete ggplot2

**ggpairs()** Crea una matriz de gráficos de dispersión

# ggplot():

**ggplot(datos, aes() ) + geom\_tipo()**

1

2

3

**aes()**: Características estéticas (aesthetic mappings). Cómo queremos que los datos se vean en el gráfico. Describe qué variables de los datos de la capa deben asignarse a qué estética

**geom\_tipo()**: representaciones gráficas de los datos en el gráfico (puntos, líneas, barras). ggplot2 ofrece muchos geoms diferentes; por ejemplo:

geom\_point (para puntos)

geom\_lines (para líneas)

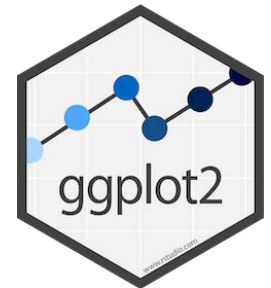
geom\_histogram (para histograma)

geom\_boxplot (para boxplot)

geom\_bar (para barras)

geom\_smooth (líneas suavizadas)

geom\_polygons (para polígonos en un mapa)



<https://link.springer.com/book/10.1007/978-3-319-24277-4>

<https://ggplot2-book.org/>

<https://github.com/rstudio/cheatsheets/blob/main/data-visualization.pdf>

# Gráficas básicas:

```
ggplot(data, mapping = aes(x,y)) + geom_***(aes(color,fill))  
                                     (+ geom_***(aes(color=varz)))  
+coord._***()  
+labs(title = "XXX", subtitle = "XX", caption = "X source: Y", x = "X", y = "Y", ...)
```

## Otros elementos:

- Stat (Stat), transformaciones estadísticas para resumir datos (por ejemplo: contar frecuencias, número de intervalos en los histogramas, etc.).
- Escala (Scale). Las escalas, por ejemplo, convierten datos en características estéticas (colores, etc.), crean leyendas... .
- Coordenadas (coordinates): sistema de coordenadas cartesianas, polares, proyecciones, etc.
- Faceting (Faceting), permite representar gráficos separados para subconjuntos de los datos originales.
- Theme.

```
ggplot(data = data, aes(y = y, x = x1 )) +  
geom_point()+  
geom_smooth() +  
facet_wrap(~factor)
```

```
ggplot(data = data, aes(y  
= y, x = factor )) +  
geom_boxplot()
```

# Gráficas básicas:

```
ggplot(datos, aes() ) + geom_tipo()
```

```
# Asigno gráfico a variables
```

```
surveys_plot <- ggplot(data = data, mapping = aes(x = x, y = y))
```

```
# Dibuja el gráfico
```

```
surveys_plot + geom_point()
```

# Gráficas básicas:

## **ggplot(datos, aes() ) + geom\_tipo()**

**ggplot(data = <DATA>, mapping = aes(<MAPPINGS>)) + <GEOM\_FUNCTION>()**

- Cualquier cosa que pongas en la función ggplot() puede ser vista por cualquier capa geom que añadas (es decir, se trata de ajustes de trazado universales). Esto incluye los ejes X e Y que configure en aes().
- También puede especificar la estética (aes) para un geom determinado independientemente de la estética definida globalmente en la función ggplot().
- El signo + utilizado para añadir capas debe colocarse al final de cada línea que contenga una capa.
- A veces se hace referencia a 'ggplot2' y a veces a 'ggplot'. Para aclarar, 'ggplot2' es el nombre de la versión más reciente del paquete. Sin embargo, cada vez que llamamos a la función en sí, se llama simplemente 'ggplot'.

### Ejemplos:

1. `ggplot(data = data, mapping = aes(x = v1, y = v2)) +  
 geom_point(alpha = 0.1, color = "blue")`
2. `ggplot(data = data, mapping = aes(x = v1, y = v2)) +  
 geom_point(alpha = 0.1, aes(color = factor))`
3. `ggplot(data = data, mapping = aes(x = v1_cat, y = v2)) +  
 geom_boxplot(alpha = 0.8) +  
 geom_jitter(alpha = 0.3, color = "tomato")`
4. `ggplot(data = data, aes(x = year, y = v1, color = factor1)) +  
 geom_line()+  
 facet_wrap(~ factor2)`

```
ggplot(Data, aes(x=x)) +  
  geom_histogram(bins=20, color="white",  
  fill="blue") +  
  ggtitle("Título") +  
  xlab("X_label") +  
  ylab("Y_label")
```

## □ Tema 3: Medidas que resumen la información.

- Medidas de tendencia central.
- Medidas de tendencia central robustas.
- Medidas de dispersión.
- Medidas de dispersión robustas.
- Medidas de posición y forma.
- Gráficos de caja.
- Datos atípicos y análisis exploratorio de datos.

Learn by **DOING**.





[www.unir.net](http://www.unir.net)