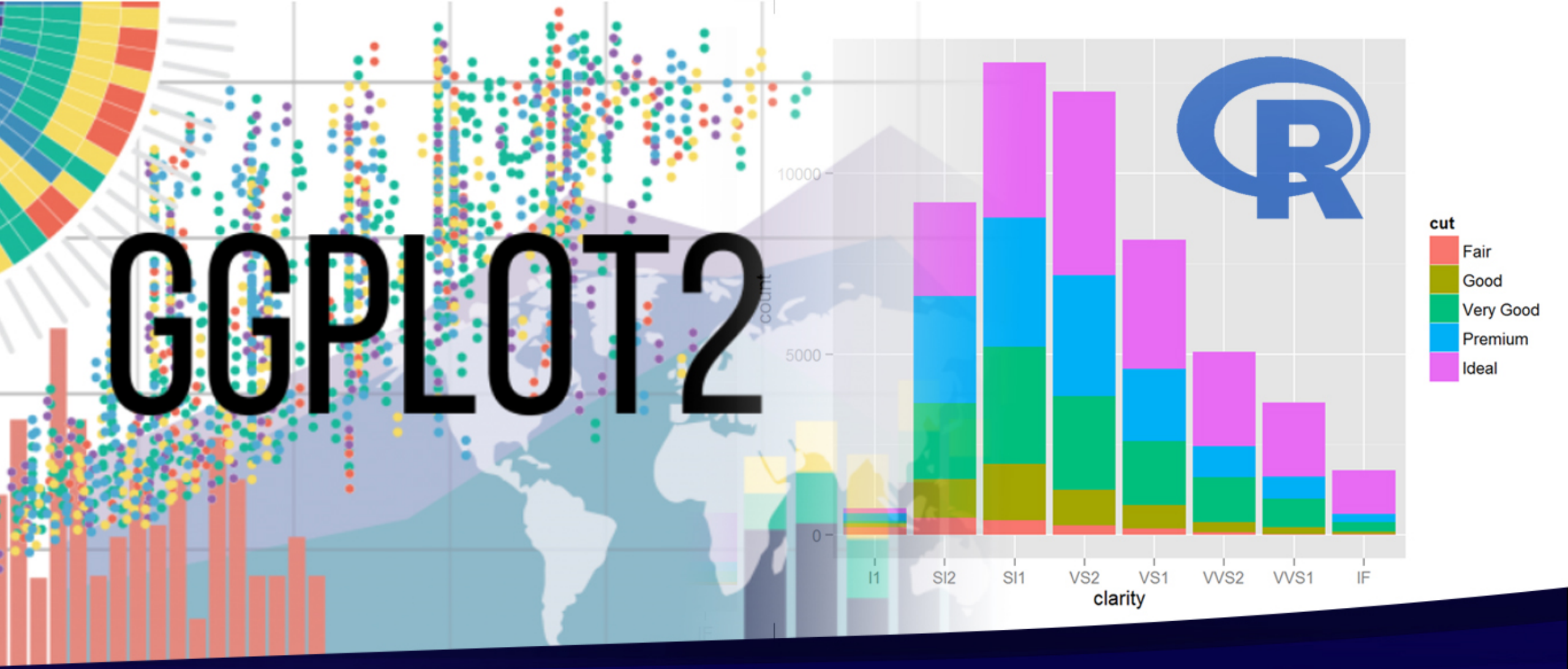


UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

Lenguaje R para la estandarización de visualizaciones de datos

Tema 1: Instalación de R, sintaxis y tipos de datos

Mgter. Danny Murillo

Objetivo

- Aprender a instalar R y el IDE Rstudio,
- Conocer los conceptos básicos del lenguaje y los tipos de datos que este utiliza.
- Conocer las estructuras básicas más comunes para guardar datos.

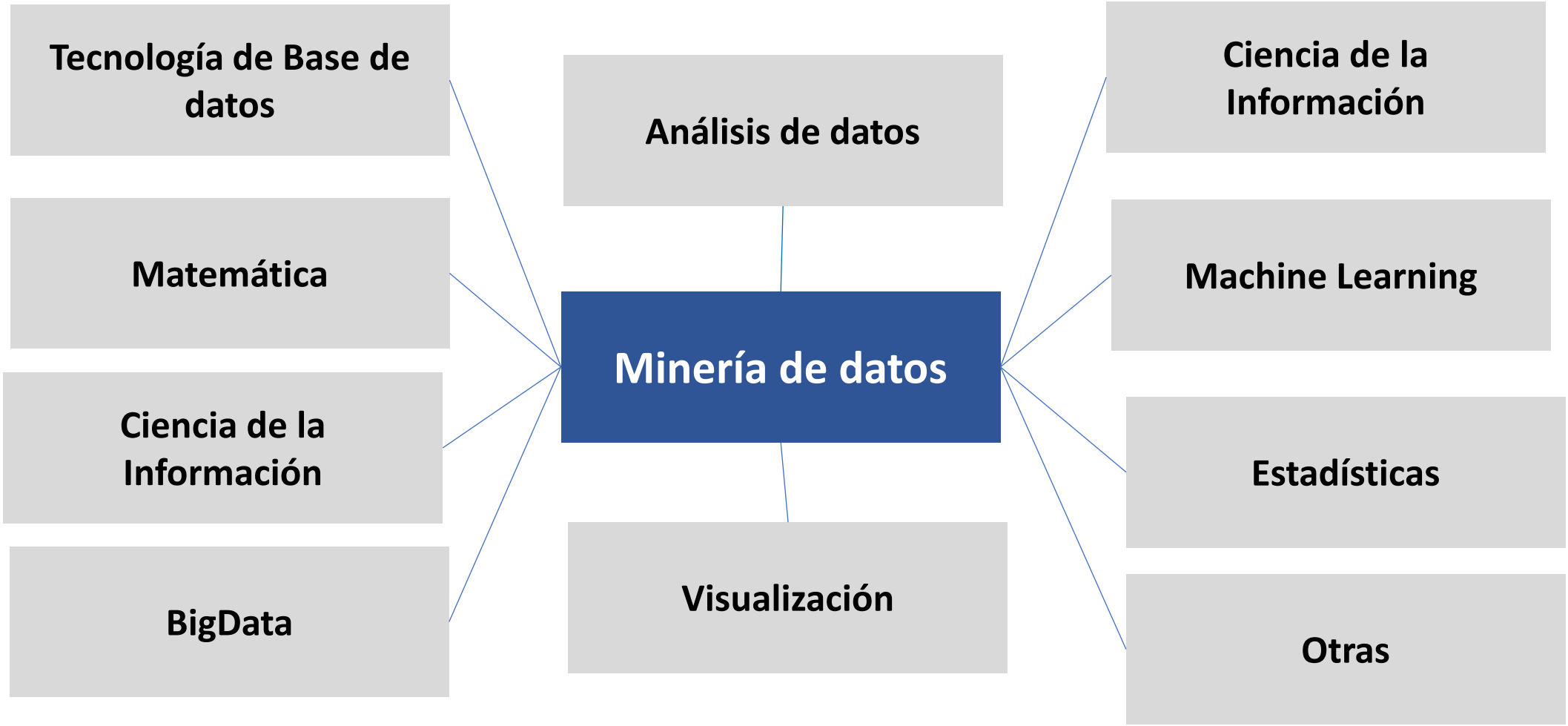
Minería de datos

Extracción de información (previamente desconocida y potencialmente útil) de grandes bases de datos para encontrar patrones ocultos usando medios automatizados.

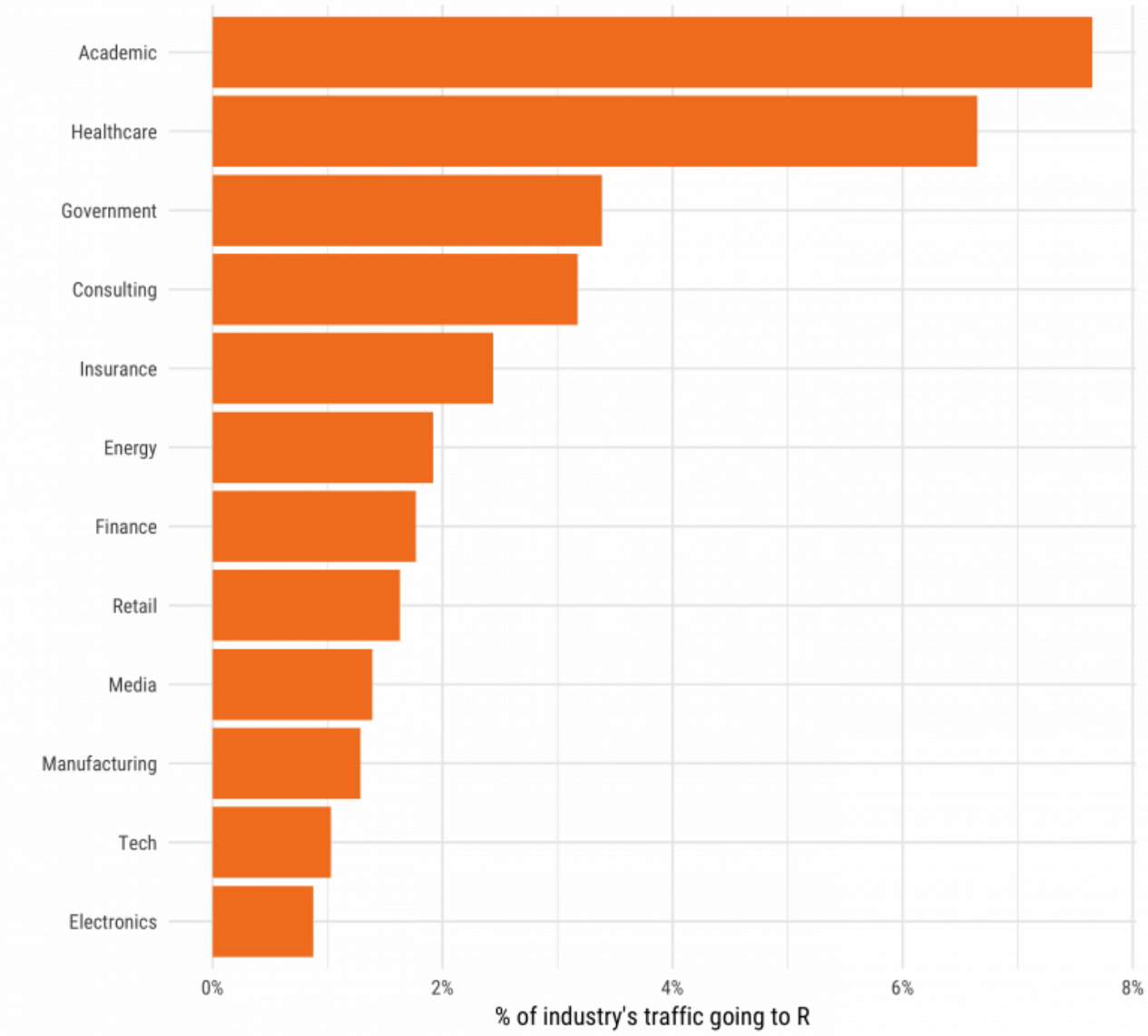
Esta búsqueda se lleva a cabo utilizando métodos matemáticos, estadísticos o algorítmicos.

El objetivo principal de la Minería de Datos es crear un proceso automatizado que toma como punto de partida los datos y cuya meta es la ayuda a la toma de decisiones.

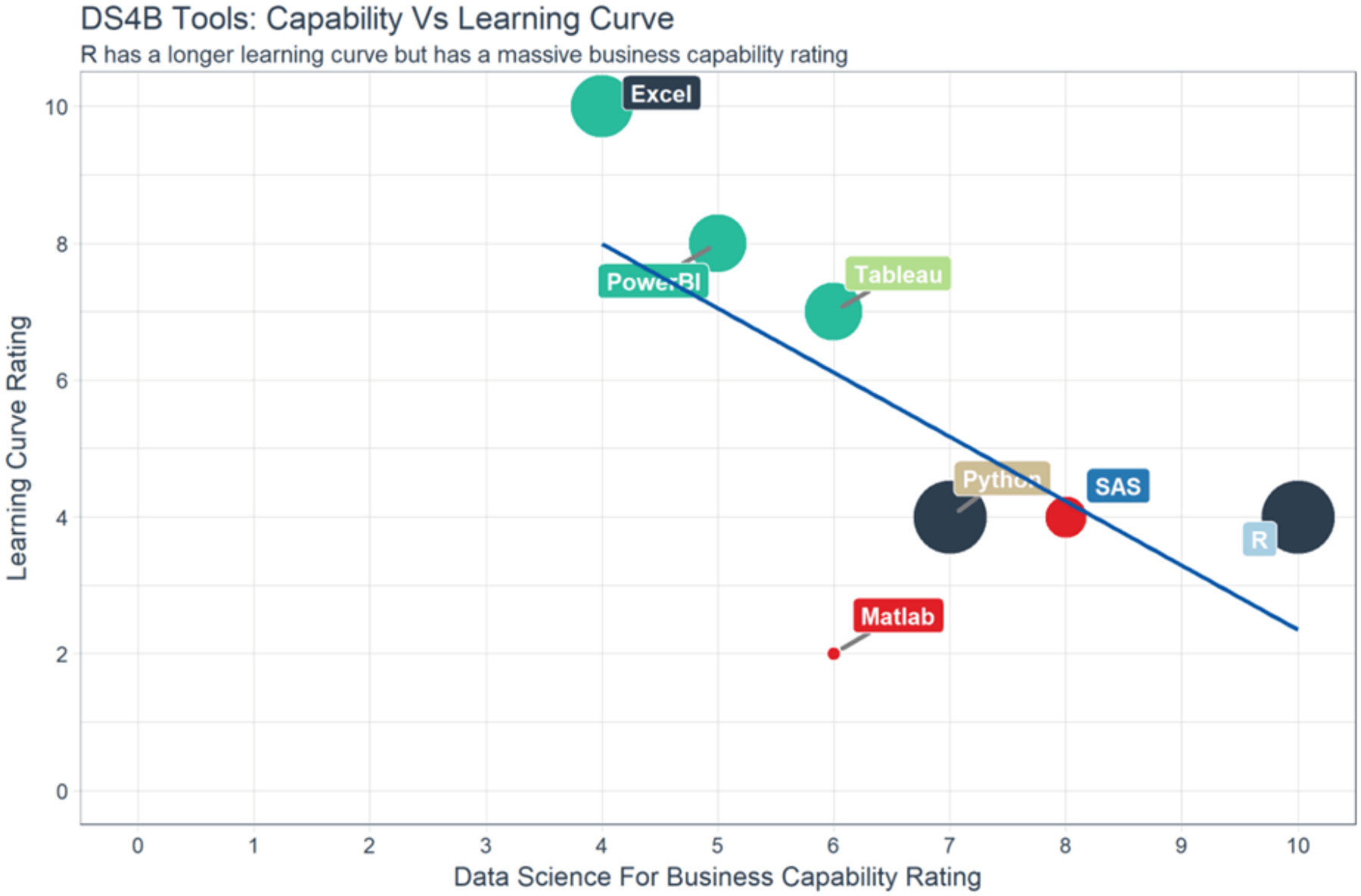
Minería de datos y otras disciplinas



Uso del lenguaje R



Curva de aprendizaje del lenguaje R



Lenguaje R

R es un lenguaje de programación usado para realizar procedimientos estadísticos y gráficos, fue creado en 1993 por los profesores e investigadores Robert Gentleman y Ross Ihaka, *miembros del Departamento de Estadística de la Universidad de Auckland, en Nueva Zelanda.*

Inicialmente el lenguaje se usó para apoyar los cursos de profesores. A partir de 1995 el código fuente de R está disponible bajo licencia GNU GPL para SO Windows, Macintosh y distribuciones Unix/Linux.

Características del Lenguaje R

1. Es un lenguaje de Script por lo cual no es necesario compilar el código.
2. Está dividido en paquetes modulares que responden a necesidades específicas.
3. Es libre, se distribuye bajo licencia GNU.
4. Es multiplataforma, hay versiones para Linux, Windows, Mac, iPhone… ¡web!
5. Se puede analizar en R cualquier tipo de datos.
6. Su capacidad gráfica difícilmente es superada por otro paquete estadístico.

INSTALACIÓN DE R

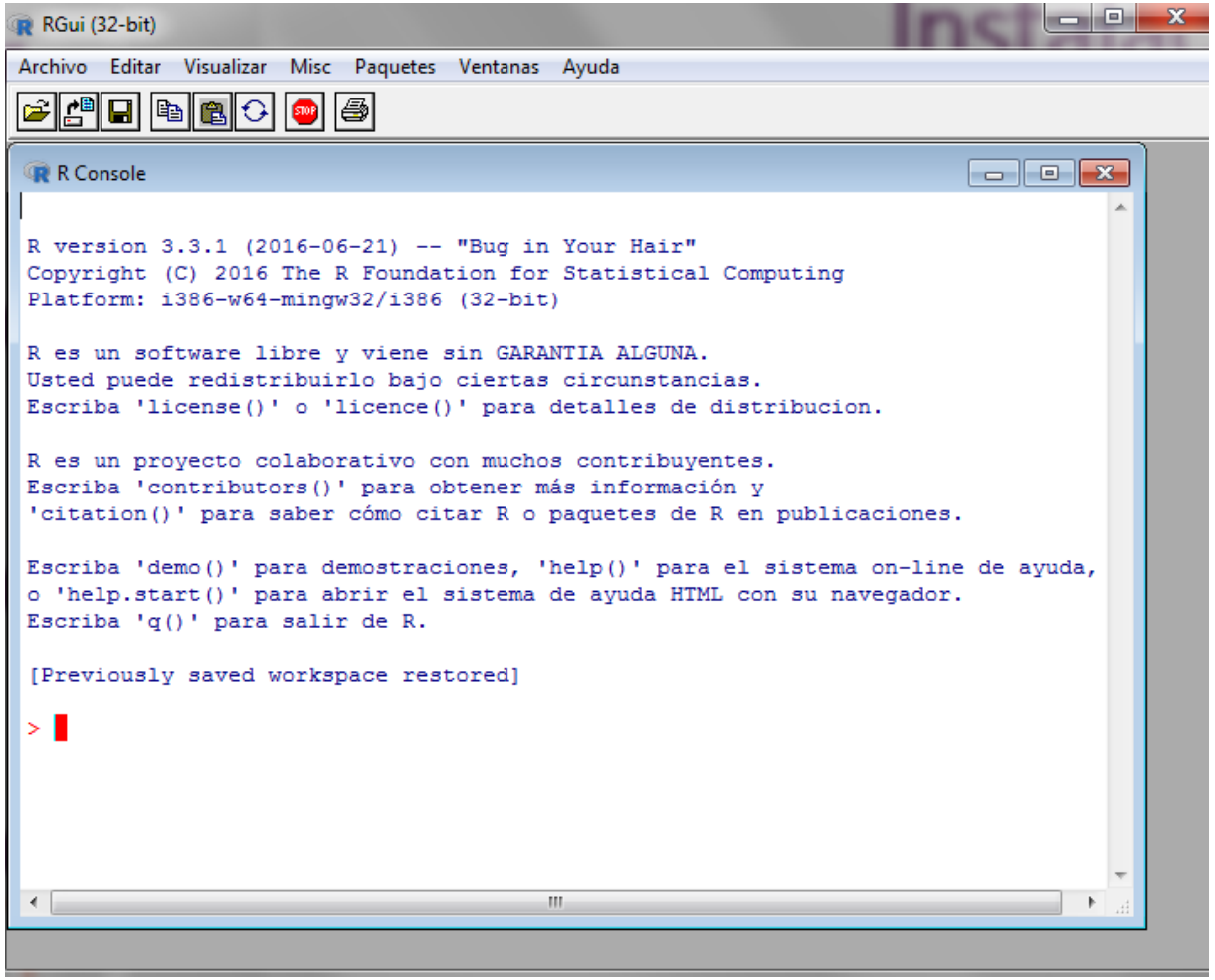
El sistema R está dividido en dos partes:

- El Sistema base con paquetes básicos que utiliza una consola para interactuar.
- Los paquetes en CRAN para cada necesidad de análisis de datos.

Para descargar R, lo haremos desde CRAN, un conjunto de servidores espejo distribuidos a lo largo del mundo y usado para distribuir R y paquetes R.
https://cran.r-project.org/

Apariencia del R en Rcommander

En R todo funciona en base a comandos y sintaxis de programación y en la consola de R commander se pueden ver los resultados de las variables y datos pero no es posible ver los gráficos que se generan.

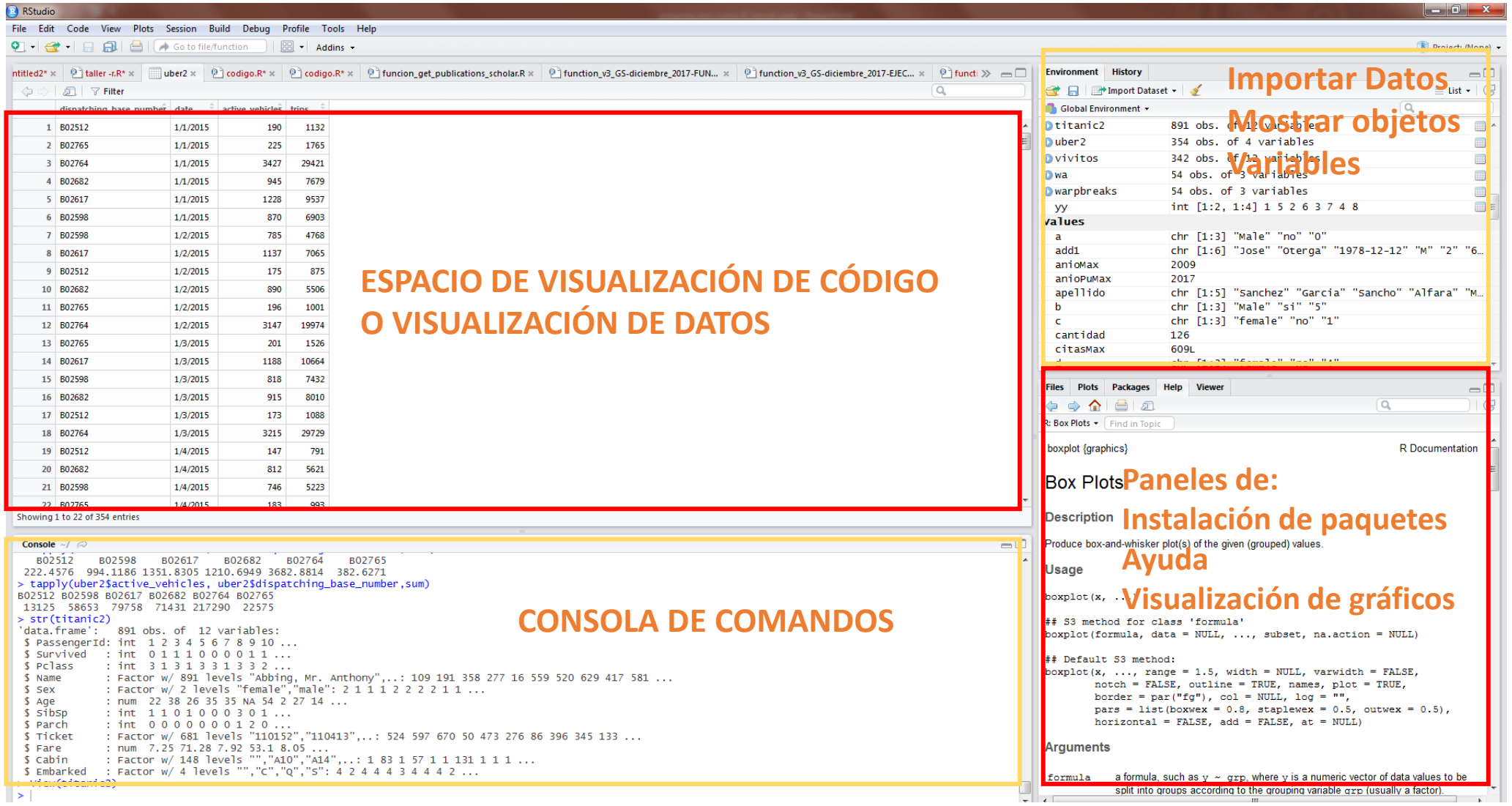


Instalación de Rstudio

Para mejorar la experiencia en el uso del lenguaje R se pude instalar el RStudio que es un entorno integrado de desarrollo, o IDE que facilita el entorno de trabajo y algunas tareas de programación.

Si ya hemos instalado R en nuestro equipo, RStudio lo detectará automáticamente y podremos utilizarlo desde este entorno.
http://www.rstudio.com/download.

Interface de Rstudio



Interface de Rstudio: Consola

Cuando hablamos de ejecutar nos referimos a darle una instrucción a R o una entrada para que realice algo.

Si escribimos en la consola lo siguiente : **25 + 10**
estamos pidiendo que se ejecute esta operación y al darle ENTER, nos será devuelto su resultado: **[1] 35**

Otros ejemplos de R como calculadora
1+2 #suma
20*3 #multiplicacion
50/9 #division
9-5 #resta
10%%3 #modulo
10%/%3 #coeficiente
9^3 #potencia

CTRL + L limpiar consola

¿Qué son los datos?

Un dato, es un valor que pertenece a un tipo de dato y que por lo regular debe estar contenida en una variable.

Una variable es una propiedad o característica de un individuo que puede variar su valor y que contiene un dato: color de ojos, estado civil, estura, edad.

Una colección de variables permiten describir un **individuo (***entidad, objeto, registro, caso, una observación*)

El conjunto de observaciones puede ser una **tabla** o una base de datos, que es necesario para hacer análisis de datos.

	Matematicas	Ciencias	Espanol	Historia	EdFisica
Carlos	6.3	6.4	8.2	9	7.2
Maria	6.8	7.2	8.7	9	7
Andres	6	6	7.8	8.9	7.3
Lucia	7	6.5	9.2	8.6	8
Ines	7.6	9.2	8	8	7.5
Ana	7.8	9.6	7.7	8	6.5
Jose	7.9	9.7	7.5	8	6
Pedro	7.5	9.4	7.3	7	7
Luis	5	6.5	6.5	7	9
Sonia	6	6	6.5	5.5	8.7
Carlos	6.3	6.4	8.2	9	7.2

Los datos y sus tipos

Tipos de datos en R:

Tipo	Ejemplo	Nombre en inglés
Entero	1	integer
Numérico	1.3	numeric
Cadena de texto	“uno”	character
Factor	uno	factor
Lógico	TRUE	logical
Perdido	`NA`	NA
Vacio	`NULL`	null

Variables en R

Una variable es una propiedad o característica de un individuo que puede variar su valor y que contiene un dato.

En R hay tres formas de asignar una variable.

Variable = valor

Variable -> valor

Variable <- valor

Ejemplos:

```
nombre<-"danny"
apellido <- "murillo"
estatura<- 1.80
Indice <-2.50
Nota_1<-80
Nota_2<-90
Bo1<- TRUE
Bo2 <- T
```

OPERADORES

Aritméticos		Comparativos		Lógicos	
+	Adición	==	Igual a	&	Y lógico
-	Substracción	!=	Diferente de	!	NO lógico
*	Multiplicación	<	Menor que	|	O lógico
/	División	>	Mayor que	is.na(x)	Ausente?
^	Potencia	<=	Menor o Igual que		
%/%	División Entera	>=	Mayor o Igual que		

EJEMPLO OPERADORES :

15 > 50

50 == 20

Bo1==Bo2

Bo2!=ind

Estructuras de datos en R

Las colecciones o conjunto de datos en R se organizan por su dimensión (1, 2, o varias dimensiones) y si son homogéneas (todos los objetos deben ser del mismo tipo) o heterogéneas (el contenido puede ser de diferentes tipos).

A continuación mostramos los cinco tipos de datos más usados en el análisis de datos:

	Homogénea	Heterogénea
1	Vector atómico	Lista
2	Matriz	Data frame
n	Array	

Estructuras de datos en R : VECTOR

El tipo más básico de estructura de dato en R es el *vector*.

Propiedades:

- **Tipo**. Un vector tiene el mismo tipo que los datos que contiene. Si tenemos un vector que contiene datos de tipo numérico, el vector será también de tipo numérico por lo que podemos decir que es **homogénea.**
- **Tamaño**. Es el número de elementos que contiene un vector. El largo es la única **dimensión** que tiene esta estructura de datos.

Estructuras de datos en R : VECTOR

Ejemplo para crear vector:

El uso de la función c() **para crear vector atómico**, que corresponde a la sigla de *combinar*:

#Crear vector numerico

c(4,2,-8)

 # Vector de cadena de texto

c("animal", "casa", "persona")

Crear vector y asignar a una variable

vector1 <- c(1, 2.5, 4.5)

Ejemplo: VECTOR

```
matematicas <- c(45,70,85,100,1000)  #crear vector
#seleccionar posición 5
matematicas[5]
#seleccionar posición del 3 al 5
matematicas[3:5]
#seleccionar todos los valores, menos el que esta en la posición 2
matematicas[-2]
#sumar dos valores de un vector
matematicas[1] + matematicas[2]
#restar dos valores de un vector
matematicas[3] - matematicas[4]
#multiplicar dos valores de un vector
matematicas[3] * matematicas[5]
#multiplicar todos los valores de un vector por 2
matematicas * 2
```

Estructuras de datos en R : MATRIZ

Una matriz es un vector numérico de dos dimensiones o de longitud 2 que define el número de filas y columnas.

Las matrices pueden ser descritas como **vectores multidimensionales**. Al igual que un vector, únicamente pueden contener datos de un sólo tipo, pero además de largo, tienen más dimensiones.

.

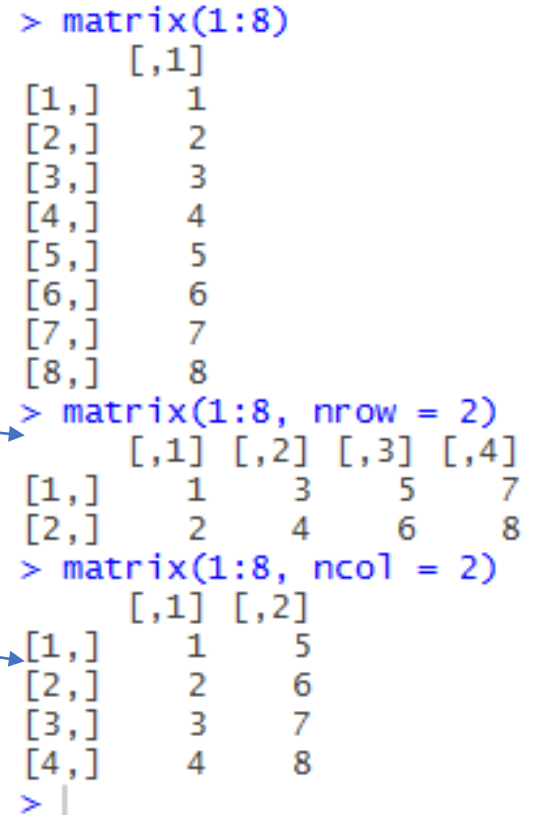
Estructuras de datos en R : MATRICES

Se crean con la función **matrix()**

#matriz de una columna
matrix(1:8)

#matriz de dos filas y 4 columnas
matrix(1:8, nrow = 2)

#matriz de dosc columnas y 4 filas
matrix(1:8, ncol = 2)



Estructuras de datos en R : MATRICES

#cargar matriz

matrix(1:8, nrow = 2)

Al crear una matrix con los datos del 1 al 8 esta guarda los valores ordenados por columnas.

```
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
```

#cambiar llenado de la matrix por fila utilizando byrow = TRUE

matrix(1:8, nrow = 2, byrow = TRUE)

```
> matrix(1:8, nrow = 2, byrow = TRUE)
     [,1] [,2] [,3] [,4]
[1,]    1    2    3    4
[2,]    5    6    7    8
>
```

Tipos de Variables

Cuantitativas (Numéricas)

Discretas

es aquella en la cual se puede contar el número posible de valores
(*son números enteros)*

continuas

puede tomar cualquier valor en un intervalo dado
(*son números reales)*

Cualitativas (Categorías)

Nominales

Las categorías son valores diferentes por una cualidad, no por una cantidad.
Sexo; estado civil.

Ordinales

Tienen un orden, pero no existe una distancia o intervalo definido entre los valores.
Bachiller, Licenciado, Máster, Doctor

Estructuras de datos en R : Data frames

Los *data frames* se utilizan en R para almacenar datos en forma de hoja de datos. Cada fila de la hoja de datos corresponde a una observación o valor de una instancia, mientras que cada columna corresponde a un vector que contiene los datos de una variable.

```
#creación de vectores
nombre <- c("Juan", "Margarita", "Ruben", "Daniel","Susana")
fecha_nacimiento <- c("1976-06-14", "1974-05-07", "1958-12-25", "1983-09-19","1975-07-18")
sexo <- c("M", "F", "M", "M","F")
nro_hijos <- c(1, 2, 3, 1,2)
edad<-c(45,67,34,90,85)
```

Data frames: Insertar filas y columnas

#crear dataframes con datos de vectores utilizando la función **cbind** añadir columnas)

```
estudiante<-data.frame()
estudiante<-cbind(estudiante, nombre)
estudiante<-cbind(estudiante,fecha_nacimiento)
estudiante<-cbind(estudiante,sexo)
estudiante<-cbind(estudiante,nro_hijos)
estudiante <-cbind(estudiante,edad)
```

Seleccionar datos en R : Data frames

Si queremos acceder a una variable

dataframe (fila, numero de columna)
estudiante[5, 1]

Aunque también podemos referirnos a la columna por su nombre:

estudiante[, “nombre"] o poniendo el nombre de la variable entre dobles corchetes y entre comillas:
estudiante[[“nombre"]]

Operador $ para uso de variables
estudiante$nombre

Funciones básicas en R

	Matemáticas		**Estadísticas**
sqrt(x)	Raíz de x	mean(x)	Media
exp(x)	Exponencial de x	sd(x)	Cuasidesviación
log(x)	Logaritmo natural de x	var(x)	Varianza
log10(x)	Logaritmo base 10	median(x)	Mediana
length(x)	Número de elementos	quantile(x,p)	Quantiles
sum(x)	Suma los elementos de x	cor(x,y)	Correlación
prod(x)	Producto de los elementos	max(x)	El máximo
sin(x)	Seno	min(x)	El mínimo
cos(x)	Coseno	range(x)	Retorna el máximo y mínimo
tan(x)	Tangente	sort(x)	Ordena las componentes de x
round(x,n)	redondea a n dígitos	which(condición)	los índices que cumplen la condición

Ejemplo de Funciones

#crear vector

data <- c(-58,46,28,69,22,18,18,42,62,78,18,210)

#conocer tamaño de un vector

length(data)

 # sumar datos de un vector

sum(data)

#calculo de la media

sum(data) / length(data)

#funciones estadísticas

#calculo de la media

mean(data)

#calculo de la mediana

median(data)

#Desviación Estandar

sd(data)