Unidad 3. Manipulación de datos en R

Del Curso introductorio al lenguaje de programación R orientado al análisis cuantitativo en Ciencias Sociales por Sarahí Aguilar González

Objetivo: Que el estudiante sea capaz de importar y exportar bases de datos de y en diferentes formatos, así como ejecutar transformaciones básicas sobre estas utilizando data frames y el paquete `data.table`.
Duración: 2 sesiones (4 horas)

Agenda

- Paquetes de R
- Data Frames
- data.table

Agenda

- Paquetes de R
- Data Frames
- data.table

corre en

Lenguaje de programación y entorno enfocado al cómputo estadístico y graficación.

R Base

La colección de funciones R que se carga cada vez que inicia R. Estas funciones proporcionan los conceptos básicos del lenguaje y no es necesario cargar un paquete antes de poder usarlas.

Entorno de desarrollo integrado (IDE) para R.

No eres la única persona que escribe sus propias funciones con R.

Muchos profesores, investigadores y programadores usan R para construir y actualizar sus propias funciones que se adecuan a sus necesidades.

Algunos, empaquetan sus funciones y las comparten con el mundo.
R promueve esto y lo permite hacer de forma fácil.

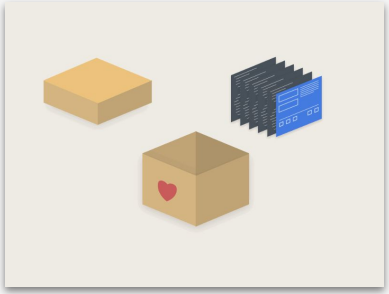
Un paquete R agrupa funciones, conjuntos de datos y archivos de ayuda.

Cualquiera puede usar estas funciones dentro de su propio código R una vez que cargue el paquete en el que residen.

Por lo general, el contenido de un paquete R **está relacionado con un solo tipo de tarea**, que el paquete ayuda a resolver.

Los paquetes necesitan ser **descargados** (una sola vez) y **cargados** al entorno de R (cada vez que se usarán).

Los paquetes son actualizados por sus creadores, y para correr correctamente en tu R, **deben ser compatibles con su versión**.



Paquetes de R

corre en

Lenguaje de programación y entorno enfocado al cómputo estadístico y graficación.

Entorno de desarrollo integrado (IDE) para R.

R Base

La colección de funciones R que se carga cada vez que inicia R. Estas funciones proporcionan los conceptos básicos del lenguaje y no es necesario cargar un paquete antes de poder usarlas.

corre sobre

Paquetes de R

Una colección de funciones escritas en R descargables. Estas funciones proporcionan conceptos más avanzada del lenguaje de forma simplificada, pero es necesario descargarlas y cargarlas antes de poder usarlas.

¿Qué paquetes de R conocen o han utilizado?

Instalando paquetes

(Descargar a tu computadora)

La forma más sencilla de instalar un paquete de R es con la función `install.packages("<package>")`

Esto buscará el paquete especificado en la colección de paquetes alojados en el sitio CRAN. Cuando R encuentre el paquete, lo descargara en una carpeta de bibliotecas en su computadora. R puede acceder al paquete aquí en futuras sesiones de R sin reinstalarlo.

Cargando paquetes

(Cargar sus funciones en tu sesión)

Para cargar un paquete de R, se utiliza la función R es con la función `library(<package>)`.

Esto pondrá a tu disposición todas las funciones, conjuntos de datos y archivos de ayuda del paquete hasta que cierre su sesión actual de R. La próxima vez que inicies una sesión de R, tendrás que volver a cargar el paquete si deseas usarlo, pero no tendrás que volver a instalarlo.

Buenas prácticas

- Utilizar Base R únicamente en medida de lo posible.
- Leer documentación.

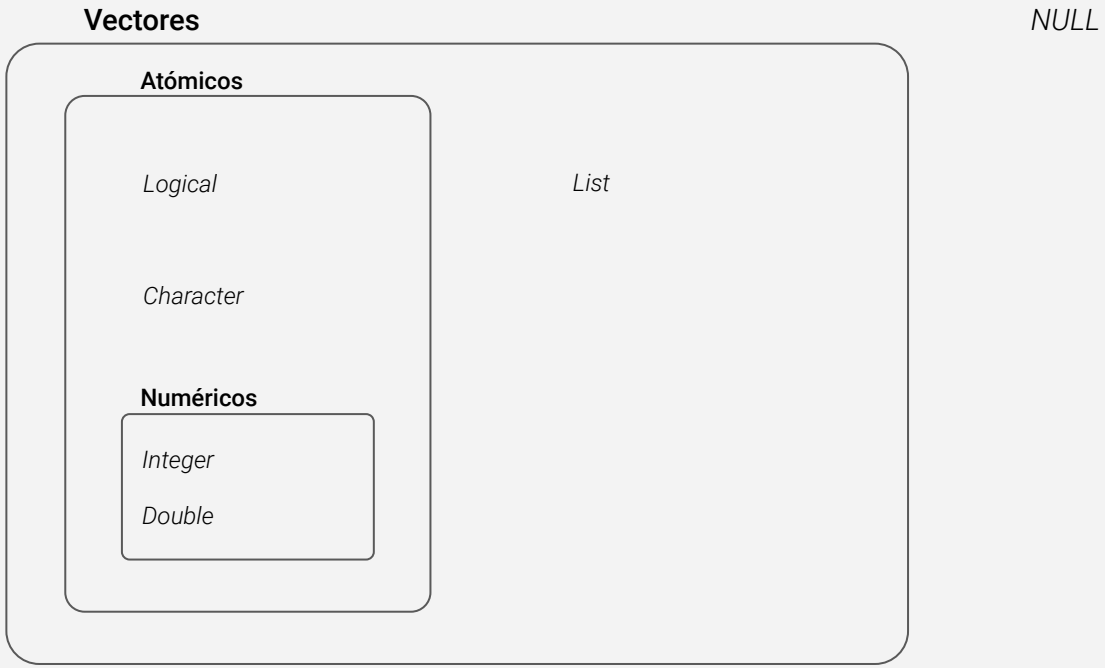
Agenda

Paquetes de R

Data Frames

data.table

Tipos de datos en R



Las 2 **propiedades** de un vector son:

Tipo

Tamaño

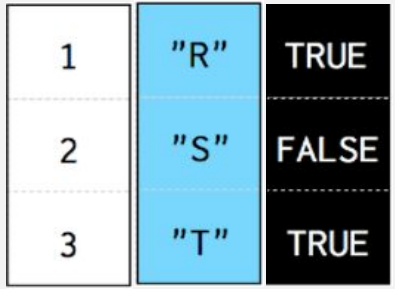
Además, pueden también contener **atributos adicionales** que los convierten en **vectores aumentados**.

Los **factores** están construidos *encima* de los vectores numéricos de tipo entero.
Las **fechas** están construidas *encima* de los vectores numéricos.
Los **data frames** están construidos encima de los vectores de tipo lista.

Los data frames son la versión bidimensional de una lista y son la estructura de almacenamiento más útil para el análisis de datos.

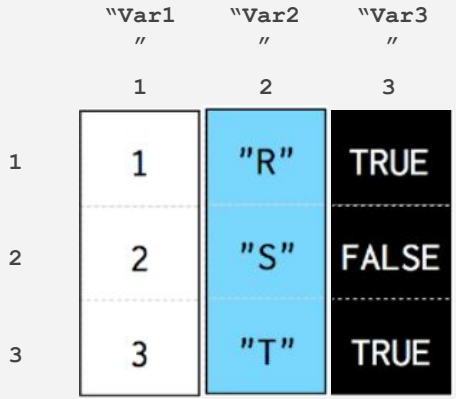
Los data frames agrupan los vectores en una tabla bidimensional. Cada vector se convierte en una columna de la tabla.

Como resultado, cada columna de un data frame puede contener un tipo de datos diferente; pero dentro de una columna, cada celda debe tener el mismo tipo de datos.



Data Frames

La extracción de datos, la creación de nuevas variables y la unión de bases de datos están basadas en **índices** y **nombres de columna**.



La forma general de sintaxis de un data frame es

```
df[i, j]
```

Textualmente, se leería de la siguiente forma:

1. Toma df
2. **Toma el subconjunto de filas usando i y columnas usando j**

`merge`; unión de bases de datos

- Solo es posible entre 2 bases de datos.
- Se requiere una o más variables compartida (llave).
- Se debe definir si se requieren conservar todos los elementos de la primer base de datos, todos los elementos de la segunda base de datos o todos los elementos de ambas bases de datos.

x

ID	X1
1	a1
2	a2

y

ID	X2
2	b1
3	b2

Todos los elementos de x

ID	X1	X2
1	a1	NA
2	a2	b1

Todos los elementos de y

ID	X1	X2
2	a2	b1
3	NA	b2

Todos los elementos de x y y

ID	X1	X2
1	a1	NA
2	a2	b1
3	NA	b2

En el ejemplo, la llave es la variable ID.

Agenda

- Paquetes de R
- Data Frames
- data.table

Agenda

Paquetes de R

Data Frames

¿Por qué no hemos utilizado el Tidyverse? Espera… ¿qué era el Tidyverse?

data.table

%>%

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.
tidyverse.org

Mientras más dependamos de paquetes de R, más y nuevas reglas tendremos que aprender, y nuestros programas serán menos homogéneos y difíciles de reutilizar.

Lo bonito es sencillo y lo sencillo es bonito.

Agenda

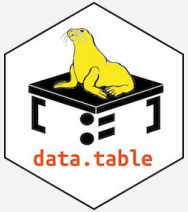
Paquetes de R

Data Frames

data.table

La manipulación de datos utilizando data frames suele ser *confusa* pues **hay muchas maneras** de extraer datos, crear nuevas variables y unir bases de datos.

data.table

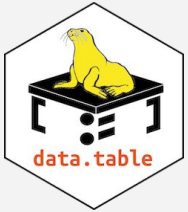


data.table es un paquete R que proporciona una versión *mejorada* de data frames.

data.table permite...

- Utilizar una **sintaxis concisa**, **coherente** e **independiente** del conjunto de operaciones.
- Realizar análisis de forma fluida **sin la carga cognitiva de tener que asignar cada operación a una función particular** de un conjunto potencialmente enorme de funciones disponibles antes de realizar el análisis.
- La **optimización automática** de las operaciones, produciendo un código rápido y eficiente en memoria.

En resumen, si uno está interesado en reducir enormemente el tiempo de programación y computación, este paquete es el indicado.



¿De qué un **data.table** proporciona una versión *mejorada* de un data frame?

A diferencia de un data frame, con un data.table se puede hacer mucho más que simplemente crear subconjuntos de filas y seleccionar columnas dentro del marco de [...].

La forma general de sintaxis de data.table es

```
DT[i, j, by]
```

Textualmente, se leería de la siguiente forma:

1. **Toma DT**
2. **Toma el subconjunto/ordena las filas usando i**
3. **Calcule j agrupado por**