**Notas de la clase 1**

1. [**Instalación**](#_4w86bihmmhgr)
2. [**Dataset**](#_y5ssvqju33l0)
3. [**Proyectos**](#_uv4pcedkvqcp)
4. [**Atajos dentro de Rstudio**](#_xzf5en535m9w)
5. [**Acceso a datos: valores de vector o data frame**](#_odjayytbqqw)
6. [**Funciones sobre dataframes**](#_y4chg2qiziqj)
7. [**Herramientas/paquetes utilizadas en la presentación**](#_y93dgxv2qn3o)
8. [**Algunos libros recomendados para leer**](#_zigt5k5pt5py)

## 1.Instalación

Los links para la instalación tanto R como Rstudio son los que se muestran a continuación:

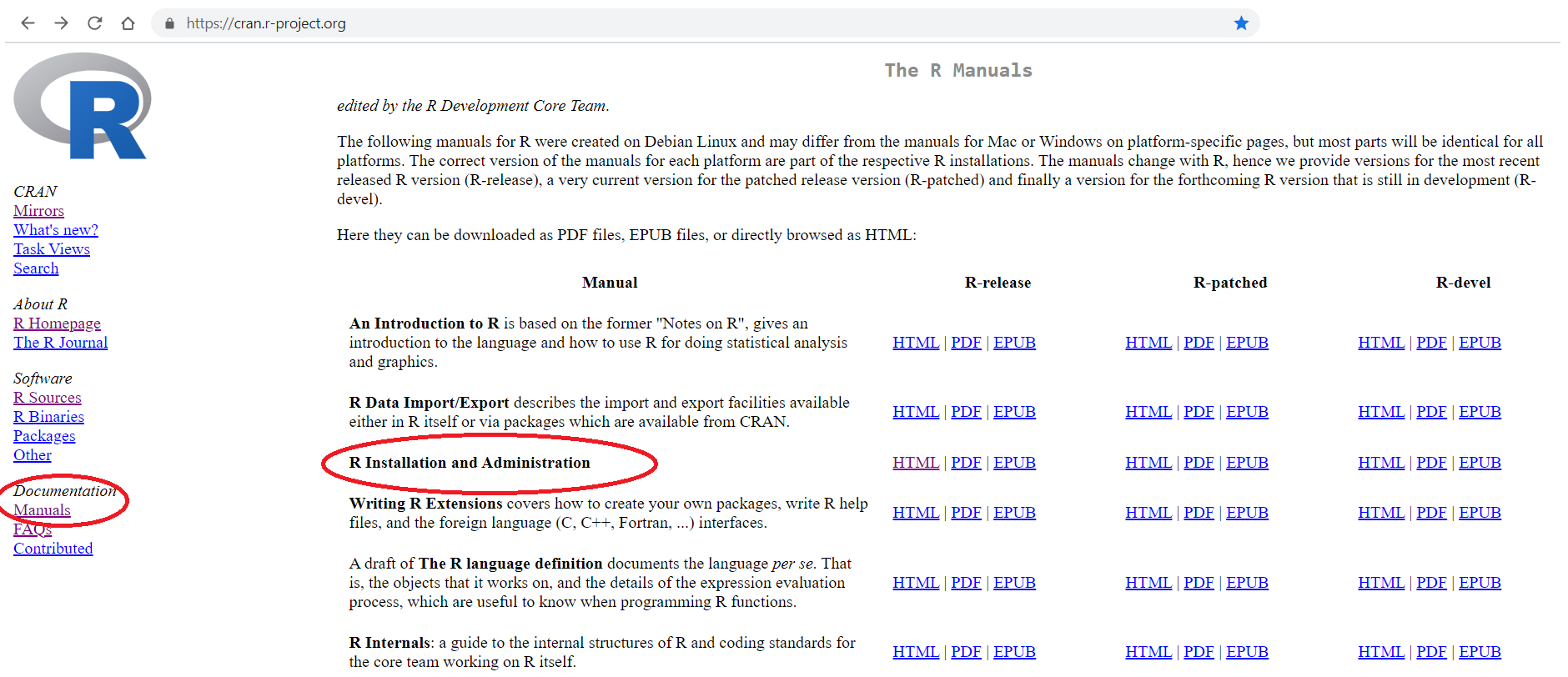
**R:** <https://cloud.r-project.org/>

**Rstudio:** <https://www.rstudio.com/products/rstudio/download/#download>

En el aula virtual encontrarán el tutorial con una guía paso a paso para instalar ambos.

* **Para instalar “R”, ¿Existe alguna restricción respecto al sistema operativo?**

La Instalación de R puede realizarse tanto en Windows, Linux o Mac OS. Para conocer más sobre la instalación pueden acceder al siguiente link <https://cran.r-project.org/> , luego ir a los manuales disponibles en la sección de Documentación o bien consultarnos por el aula virtual cualquier duda que tengan.



Para instalación en windows:

<https://cran.r-project.org/doc/manuals/r-release/R-admin.html#Installing-R-under-Windows>

Si su sistema operativo es de 32bit, este no es compatible con la versión actual de RStudio. Puede instalar versiones anteriores de Rstudio (las anteriores a la 1.4 van a funcionar bien).

<https://www.rstudio.com/products/rstudio/older-versions/>

* **¿Cuáles son las características base de Hardware para que trabajen bien las herramientas?**

Si bien oficialmente no existen requerimientos mínimos para utilizar R, es recomendable tener:

* sistema operativo de 64 bits (también es posible instalarlo en un s.o. de 32 bits)
* memoria ram como mínimo de 1 gb
* disco de 2 gb como mínimo

Obviamente que estas características van a variar dependiendo del uso que le den y del volumen de datos con el cual vayan a trabajar.

* **¿También hay que instalar Rtools? ¿O depende del sistema operativo que utilices?**

Si utilizan el sistema operativo Windows, en algunos casos será necesario instalar Rtools. Además, deberá indicar a R sobre Rtools. Pueden descargar Rtools y encontrar algunas indicaciones en el siguiente link: <https://cran.r-project.org/bin/windows/Rtools/>

## 2. Dataset

* **¿De qué trata la Encuesta Permanente de Hogares (EPH)?**

La Encuesta Permanente de Hogares (EPH) es una encuesta de manera continua por el instituto nacional de estadística de Argentina (INDEC). La misma presenta trimestral y semestralmente indicadores sobre el mercado laboral, los ingresos y condiciones de vida de los hogares. Los microdatos de la encuesta pueden encontrarse en la página oficial del INDEC:

https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos

* **¿Hay algún motivo para la elección de R sobre Python?**

Depende principalmente de la utilización de uno u otro lenguaje, según sea conveniente. Hay tareas para las que R tendrá más fortalezas sobre Python o a la inversa, pero este curso trabajará en su totalidad con lenguaje R.

* **¿Aprenderemos a trabajar con base de datos tipo censo, es decir, cuando se tiene más de 5 millones de observaciones?**

En principio no existen restricciones. En este curso trabajamos con bases chicas, pero las herramientas se podrán aplicar a cualquier volumen de datos con el que se quiera trabajar, ya que la funcionalidad es bastante independiente del tamaño de la base (va a depender de los recursos de tu computadora). Sin embargo, para aquellos casos en los que el tamaño de las bases sea más grande, es posible recurrir a algunas librerías pensadas para trabajar con grandes volúmenes de datos y optimizar los procesos (ver, por ejemplo, la librería [data.table](https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html))

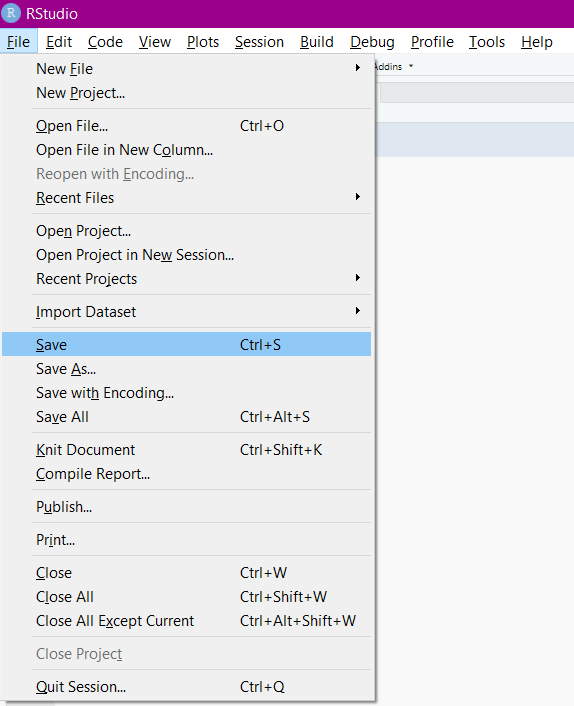
## 3. Proyectos

* **Para comenzar con un proyecto nuevo, ¿Cómo haces para que aparezca la opción Run?**

Es necesario abrir nuevo archivo o script, para ello los pasos son:

* Ingresar al menú **Files**, luego **New File** y por último seleccionar **New** **Script**
* **¿El archivo se va guardando mientras se trabaja?**

No, hay que guardar el archivo. Para esto existen dos maneras, la primera es ingresando al menú **Files** y luego seleccionando **Save** o mediante el atajo **Ctrl + s.** De todas formas, RStudio tiene un sistema de guardado paralelo que permite recuperar, en caso de que surja algún contingencia como un corte de luz,



## 4. Atajos dentro de Rstudio

**Atajos de líneas de código útiles para editar un script:**

* **ctrl + shift + c:** para comentar (#) o descomentar una o más líneas de código.
* **ctrl + D:**  para borrar una línea de código.
* **Alt + Arriba/Abajo**: para mover hacia arriba o hacia abajo una línea de código dentro de un script.
* **Ctrl + I:** para arreglar las indentaciones de las líneas de código.
* **Alt + -** (guión medio): para la asignación (<- )

**Atajos para la consola:**

* **Ctrl + (- o +):** Para aumentar o disminuir el tamaño de la fuente.
* **Ctrl + L:** Para limpiar la consola.

**Atajos para la gestión de archivos**

* **Ctrl + Shift + N:** paracrear un nuevo script.
* **Ctrl + O:** para abrir un archivo o script desde el explorador.
* **Ctrl + S:** para guardar el archivo o script en el que estemos trabajando.
* **Ctrl + W:** para cerrar el archivo o script en el que estemos trabajando.
* **Ctrl + Shift + W:** para cerrar los archivos o scripts que tengamos abiertos.
* **Ctrl + Alt + Shift + W:** para cerrar todos los archivos o scripts que tengamos abiertos, excepto el que estemos utilizando en ese momento.

## 5. Acceso a datos

* **Vectores**
* **Al ejecutar en la consola C(23:26) me muestra el siguiente mensaje "Error in C(23:26) : object not interpretable as a factor”**

La creación de vectores se realiza con c() con c minúscula. El lenguaje R es sensible a minúsculas y mayúsculas, por tanto, hay que tener cuidado al usar las sentencias propias de R como en el uso de variables que has nombrado en otras acciones. Sin embargo, el nombre de las variables puede ir en inglés o español, o cualquier otro idioma.

* Cuando combinamos dos valores que son de diferente “tipo”, R fuerza al tipo más general para resolver el problema. El orden de "forzado" para vectores atómicos es: logic -> numeric (integer -> double) -> character

Como ven, R fuerza desde lo más restringido a lo más general.

Ejemplo:

# Si nosotros tenemos un vector con valores lógicos (logic) y otro con una palabra (character):

log <- c(TRUE, FALSE)

num <- c(1:3)

car <- c("hola", "ok")

class (c(log, num))

# [1] "integer"

class (c(car, num))

# [1] "character"

class(c(log, car))

# [1] "character"

* **¿Qué significan los valores [1] y [4]?**

Es el número (la posición) de elemento dentro del vector.

* **Formas de acceso a una columna o variable**

Teniendo en cuenta el ejercicio de la práctica donde teníamos un data frame (llamado base\_personas) con 3 columnas nombre, localidad y tipo de alojamiento, si quisiéramos acceder a la variable tipo de alojamiento, lo podríamos resolver de diversas formas:

base\_personas$tipo\_alojamiento

base\_personas[[c("tipo\_alojamiento")]]

base\_personas[ , 3]

base\_personas[ , c("tipo\_alojamiento")]

* **Formas de acceso a un valor particular dentro de un data frame**

Si quisiéramos extraer del data.frame anterior el valor de la tercera fila y segunda columna, podríamos resolverlo así:

base\_personas [3, 2]

base\_personas$tipo\_alojamiento[3]

base\_personas[[“tipo alojamiento”]][3]

## 6. Funciones sobre dataframes

Algunas recomendaciones muy generales para la exploración de un dataset son:

- **str(iris)**: Devuelve la estructura general de las variables y estructura del dataframe

- **summary(iris):** Devuelve un resumen de cada una de las variables (dependiendo de su clase genera una salida específica).

**¿La función table representa la frecuencia de una variable?**

Si, table() arroja frecuencias.

**Manejo de NA en funciones base (argumento na.rm = T)**

Cuando ejecutamos algunas funciones de R Base tenemos un argumento opcional que nos permite manejar los NA (valores perdidos o del inglés ‘not available’), para que no interfieran en nuestros cálculos.

Ejemplo:

mean(1:10)

# 5.5

mean(c(1:10, NA))

# NA

mean(c(1:10, NA), na.rm = T)

# 5.5

En la clase se preguntó cómo manejaba el **argumento na.rm = T** el tratamiento de los NA en las operaciones, en el sentido de si el promedio se calcula colocando en el denominador la cuenta total o solo la cuenta de los válidos. La respuesta es que sólo considera los valores válidos. De hecho, na.rm es una abreviación de remove NA (remover NA), por lo cual podemos tener una intuición de qué hará con estos resultados en los cálculos.

**Lectura de datos que contienen NA**

Cuando leemos un data.frame desde un archivo externo (por ejemplo un archivo de texto separado por comas - csv) pueden existir valores que en el archivo original no tengan sus campos completos. Este es el tipo de campos que la mayoría de las funciones de lectura de datos externos de R considerará como NA. El archivo original no debe decir NA para que R sepa que es un valor NA. Sin embargo, muchas de las funciones de lectura de datos externos de R permiten definir cuáles valores se considerarán NA. Por otra parte, hay que recordar que para R “ ” es un character (un elemento que no tiene nada), por lo cual no es sinónimo de NA.

## 7. Herramientas/paquetes para la presentación

El paquete [rmarkdown](https://bookdown.org/yihui/rmarkdown/) y presentaciones del tipo [xaringan](https://github.com/yihui/xaringan).

# 8. Algunos libros recomendados para leer

## Hadley Wickham. R para Ciencia de Datos. <https://es.r4ds.hadley.nz/> (español) <https://r4ds.had.co.nz/> (inglés)

## Antonio Vazquez Brust. Ciencia de Datos para Gente Sociable <https://bitsandbricks.github.io/ciencia_de_datos_gente_sociable/>