



PONTIFICIA
UNIVERSIDAD
CATÓLICA
DEL PERÚ

Introducción al R

ESTADÍSTICA

Maria Teresa Villalobos Aguayo

mtvillalobosa@pucp.edu.pe

Sobre el R

Sobre el R, RStudio y RCommander

Qué es el R?

- R es un lenguaje computacional de alto nivel y un programa para realizar análisis estadístico y gráficos.
 - ✓ Permite aplicar una variedad de métodos estadísticos básicos y avanzados.
 - ✓ Produce gráficos de alta calidad.
 - ✓ R es un lenguaje de programación; es decir, podemos escribir nuevas funciones y extender el uso de R.
- R es un software open source que es mantenido por muchos contribuyentes. El R Core Team es el grupo de programadores responsables de modificar el código fuente de R.
- El sitio web oficial de R es: <http://www.R-project.org>
- R puede ser instalado libremente (no requiere pago ni registro alguno) en Windows, Mac o Linux.

Qué es el RStudio?

- RStudio es un IDE (Entorno de desarrollo integrado) para R.
 - ✓ Permite desarrollar código en R con más facilidad que usando el programa base.
 - ✓ R-Studio es desarrollado por RStudio, Inc y tiene versiones para una Desktop o para un servidor, también tienen la opción de ser libre o de pago. En el curso utilizaremos la opción para una Desktop libre.
- Para trabajar con el R es mejor trabajar con el Rstudio.
- El sitio web oficial del R-Studio es: <https://rstudio.com/>

Qué es el RCommander:

- El R-Commander es una interfaz que permite el manejo del programa R mediante una ventana de menús.
- Esta interfaz permite al usuario comenzar a manejar este programa sin conocer el lenguaje de instrucciones, y
- Permite el aprendizaje del lenguaje R de forma sencilla (si el usuario así lo prefiere).

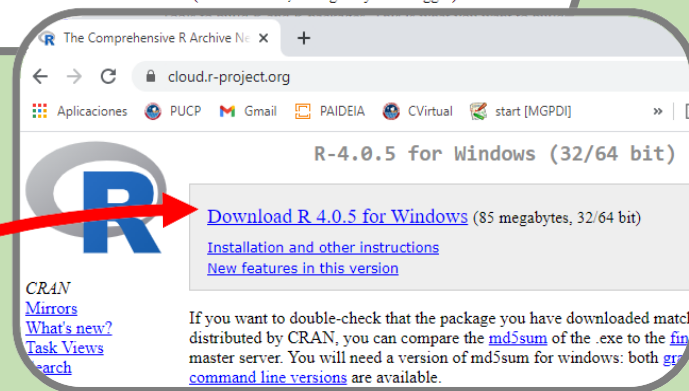
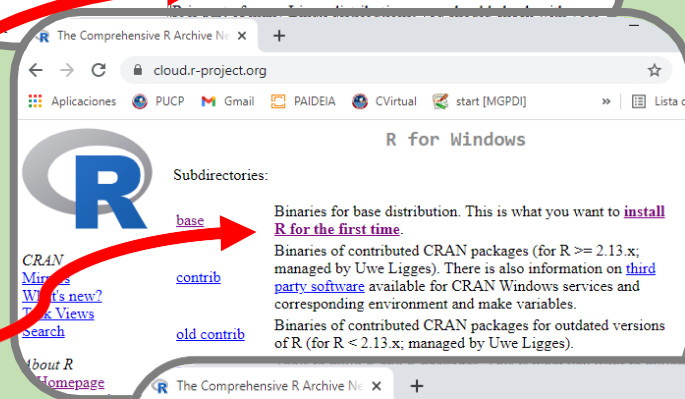
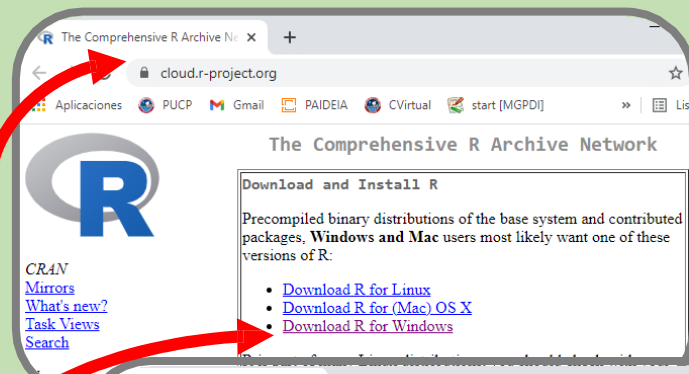
Instalación

Instalación

- Es necesario instalar 2 aplicativos:
 - ✓ El R, siguiendo las opciones default.
 - ✓ El RStudio, siguiendo las opciones default.
- Posteriormente, a través del RStudio, se puede instalar el RCommander.

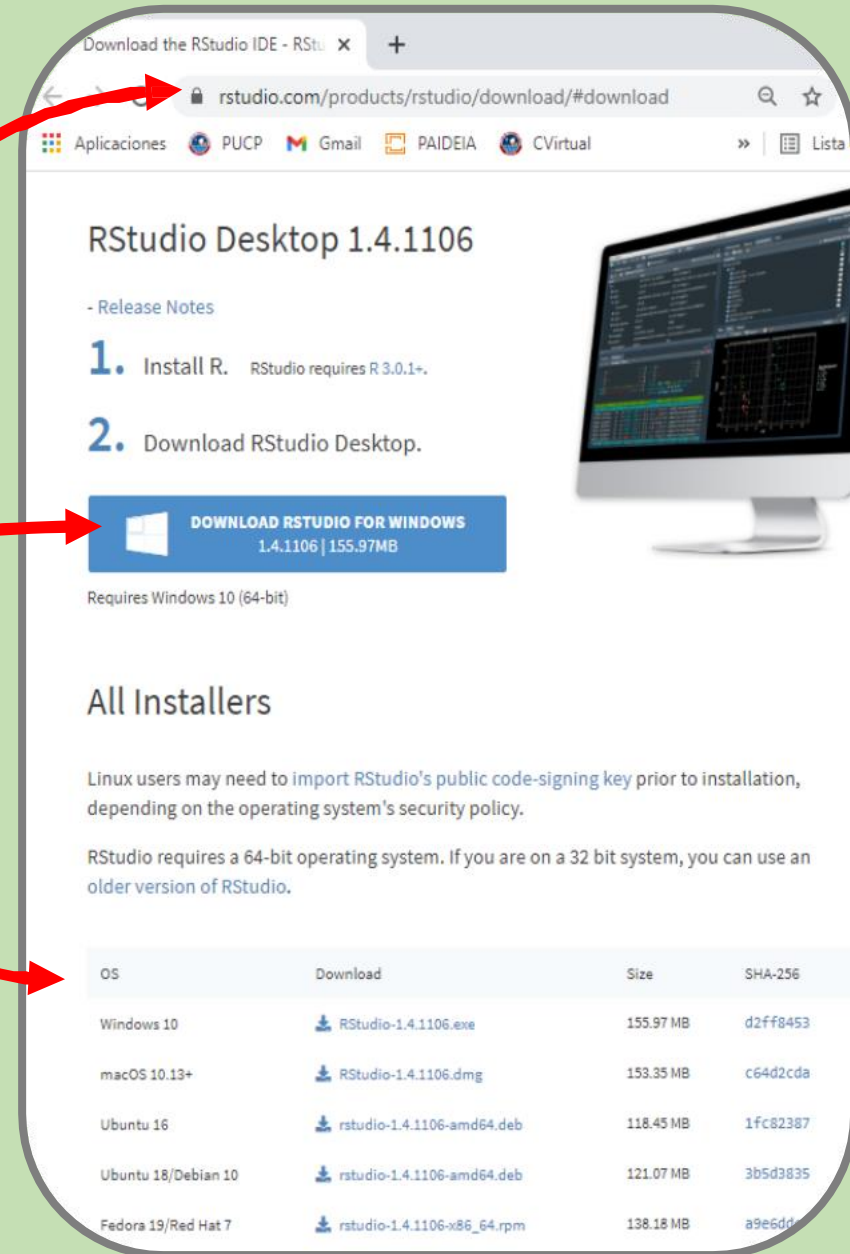
Instalación del R

1. Ir a la pagina <https://cloud.r-project.org/>.
2. Seleccione para descargar la versión de acuerdo a su sistema operativo.
3. En el caso de Windows, descargue la opción base.
4. Descargue el instalador y ejecute considerando todas las opciones que aparecen por default.



Instalación del RStudio

1. Ir a la página web
<https://rstudio.com/products/rstudio/download/#download>
2. Descargar el instalador de acuerdo a su sistema operativo.
3. Instalar después de haber instalado el R, y ejecutar considerando todas las opciones que aparecen por default.



Introducción al RStudio

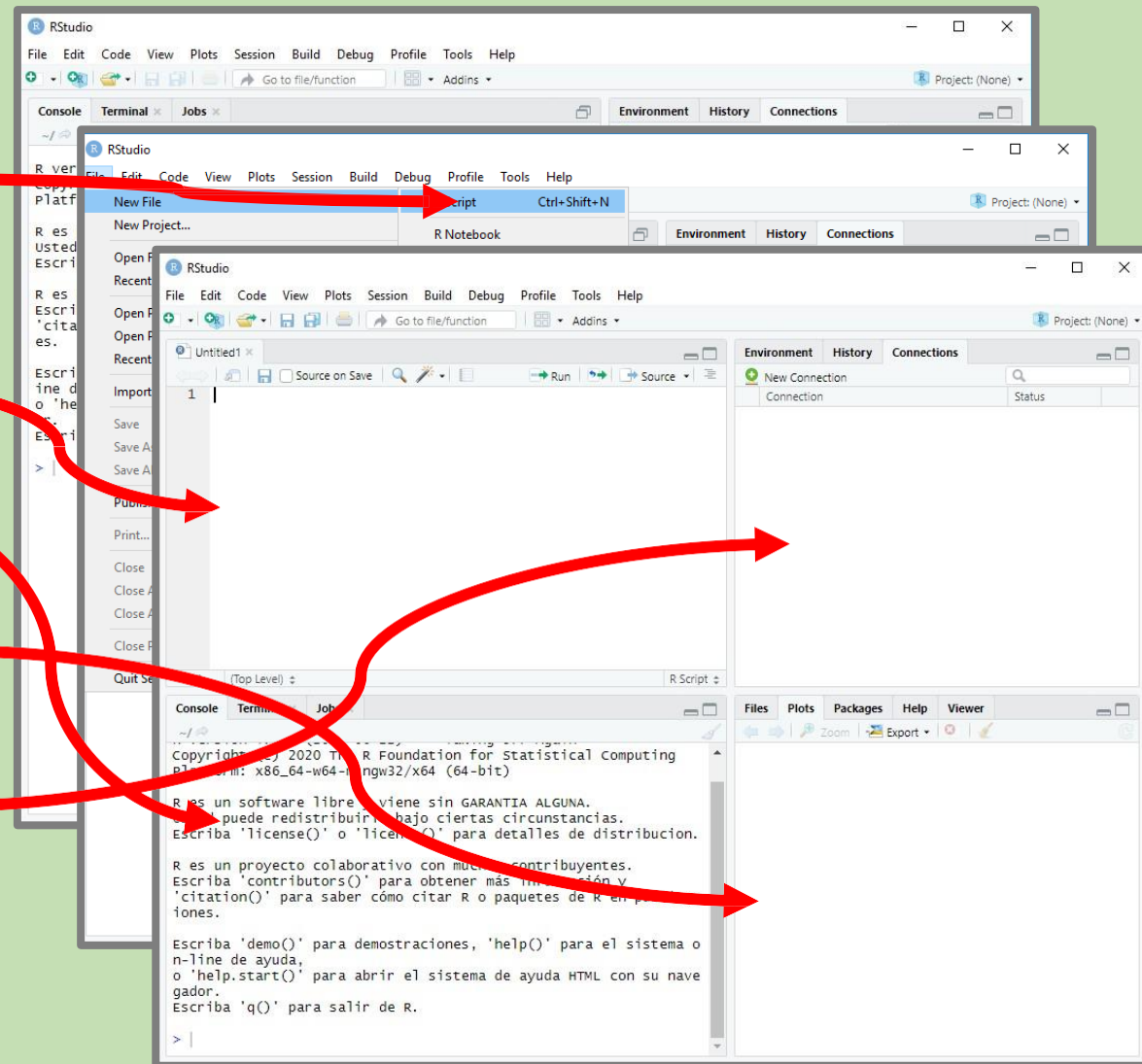
Explorando el RStudio

Después de abrir el RStudio, active la ventana para el Script:

➤ **> File > New File > R Script**

En seguida, verá que este sistema cuenta con 4 ventanas, que sirven para las siguientes funciones:

- ✓ **Script.** Programa en R que será guardado como texto.
- ✓ **Console.** Aquí se ejecutan las instrucciones en la línea de comando.
- ✓ **Files/Plots/Packages/Help/Viewer.** Aquí se presentarán: la ayuda, gráficos, los paquetes, etc.
- ✓ **Environment/History.** Aquí se presentan los objetos creados y el histórico de los comando ejecutados.



Lectura y manejo de datos en RStudio

Preparación de datos para el RStudio

Los conjuntos de datos deben ser organizados en una matriz, como una base de datos, donde las filas representan las observaciones (unidades estadísticas) y las columnas representan a las variables.

Ejemplo.

Datos Iris de Fisher. Es un conjunto de datos de flores de Iris de tres variedades diferentes. Presentan medidas, en cms., del largo y ancho del sépalo y del pétalo de las flores.

En este caso, tenemos 150 observaciones (filas) en las cuales se han medido 5 variables (columnas).

Estos se encuentran en el archivo iris.csv

	A	B	C	D	E
1	largo.sepalo	ancho.sepalo	largo.petaló	ancho.petaló	variedad
2	5.1	3.5	1.4	0.2	Setosa
3	4.9	3	1.4	0.2	Setosa
4	4.7	3.2	1.3	0.2	Setosa
5	4.6	3.1	1.5	0.2	Setosa
6	5	3.6	1.4	0.2	Setosa
7	5.4	3.9	1.7	0.4	Setosa
8	4.6	3.4	1.4	0.3	Setosa
9	5	3.4	1.5	0.2	Setosa
10	4.4	2.9	1.4	0.2	Setosa
11	4.9	3.1	1.5	0.1	Setosa

Los datos se pueden digitar en un archivo de texto separados por comas, con los nombres de las variables en la primera línea, o en un archivo del Excel, para posteriormente importarlos.

Además, se pueden importar de archivos de sistemas estadísticos como: SPSS, SAS o Stata.

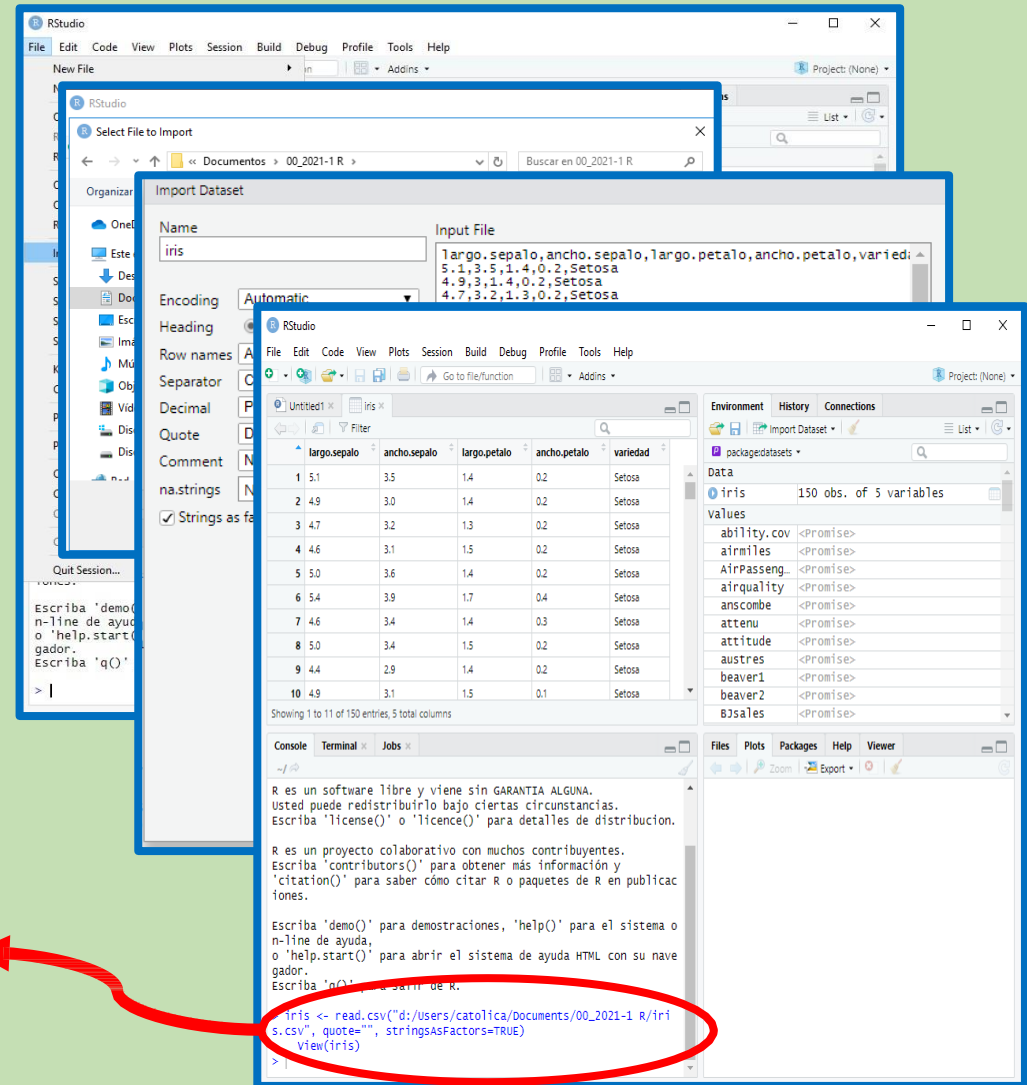
Importando los datos al RStudio

- Para poder analizar un conjunto de datos en R, debe estar disponible como un objeto en la memoria. Para ello, seleccione del menú, lo siguiente:

- > File > Import Dataset > From Text (base) ...
- > File > Import Dataset > From Excel ...

- Al ejecutar este comando para importación de datos, se genera el siguiente código, que posteriormente podrá ser reutilizado:

- `iris <- read.csv("d:/Users/catolica/Documents/00_2021-1 R/iris.csv", quote="", stringsAsFactors=TRUE)`
- `View(iris)`



Visualizando los datos en el RStudio

- Podemos **verificar los datos**:

- **View(iris)**

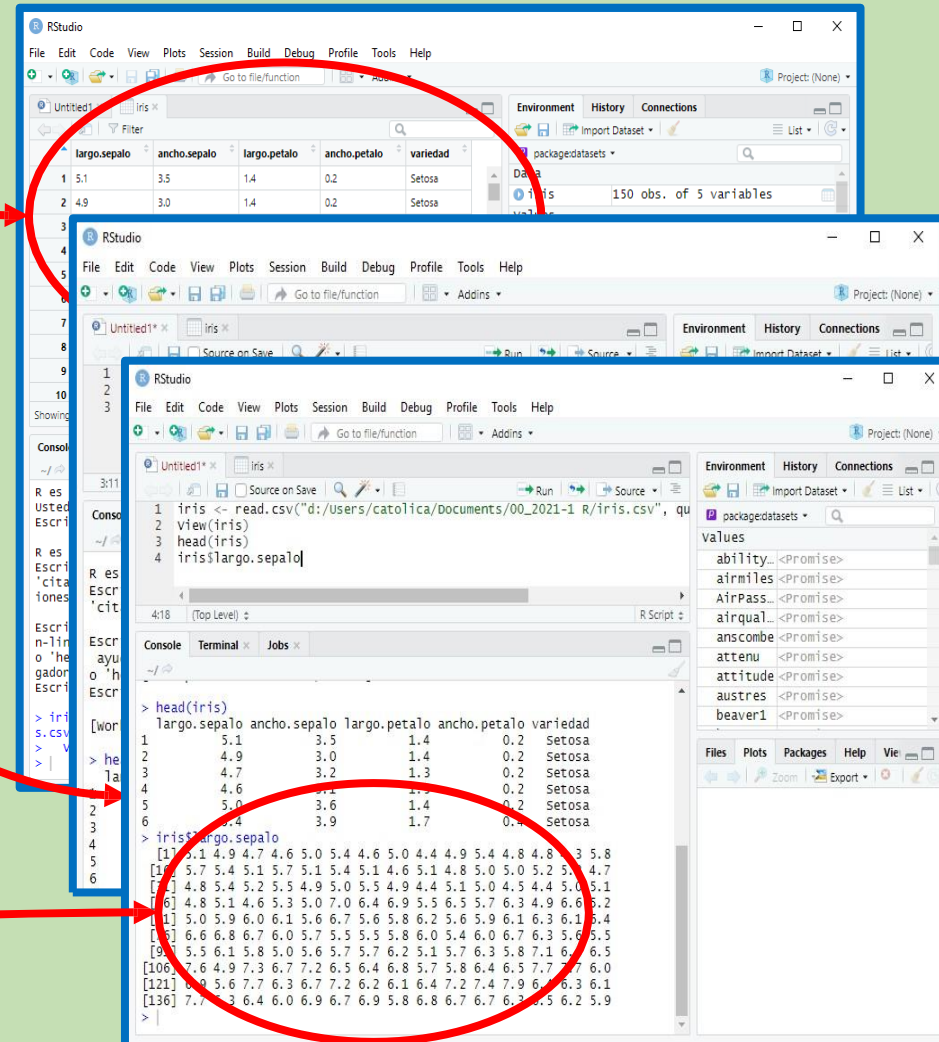
Se visualizan los datos subidos como un archivo.

- **head(iris)**

se presentarán las primeras filas del objeto en la consola.

- En R los conjuntos de datos suelen estar **almacenados en un objeto de tipo data.frame**. Esto es una matriz donde cada columna contiene la información de una variable. Para **visualizar los datos de una variable** de un data.frame, se utiliza el operador \$. Por ejemplo, si deseamos acceder a los datos de la variable largo.sepalo del conjunto de datos iris:

- **iris\$largo.sepalo**



Introducción al R-Commander

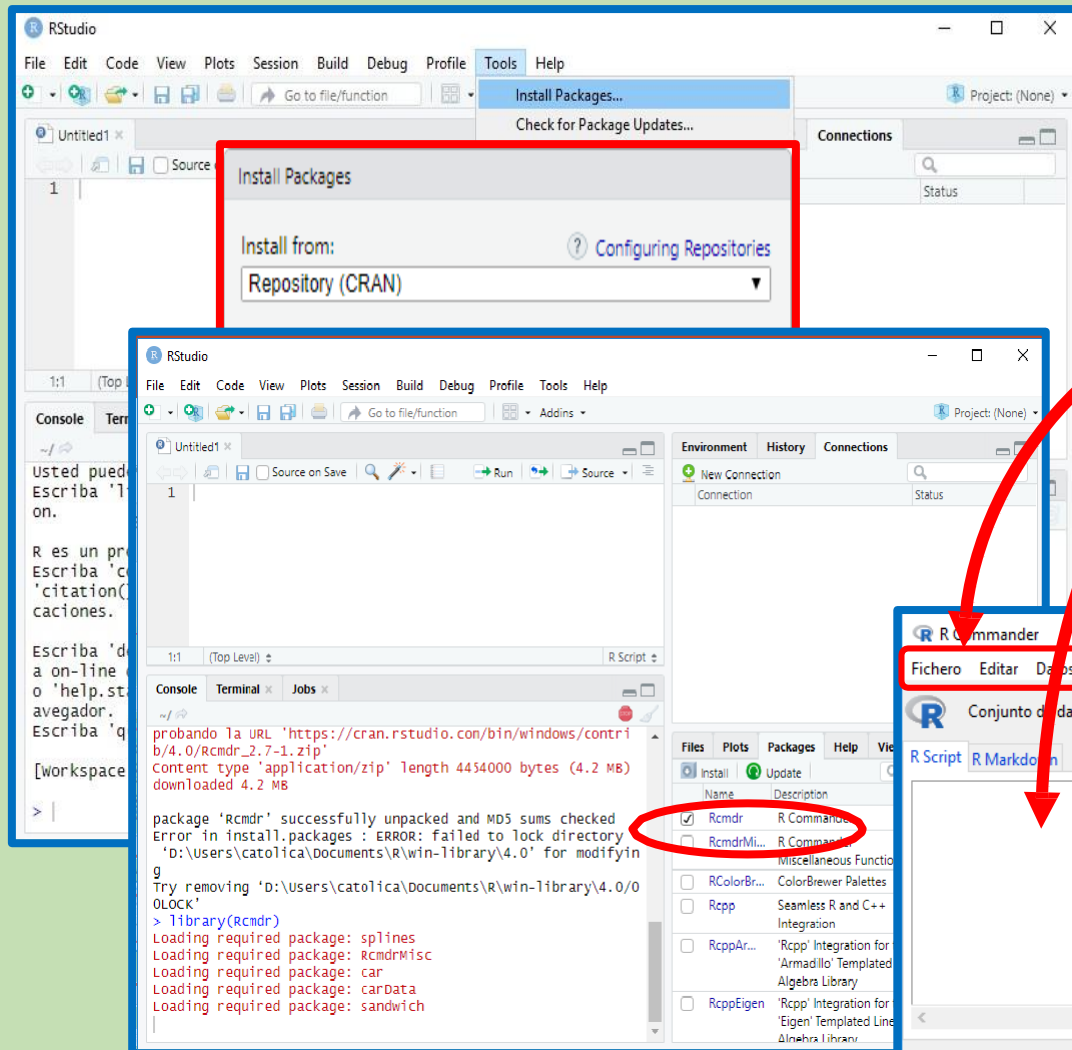
Este ejercicio será desarrollado con el Sistema Estadístico R y la interfaz del R Commander:

El objetivo de este laboratorio es que conozcan el sistema R-Studio y el R-Commander, para que puedan procesar un conjunto de datos, previamente digitados en Excel, de manera sencilla.

R es un software libre que permite realizar análisis estadísticos y el más usado en la comunidad científica. Este programa está disponible en la página web: <http://www.r-project.org> y consta de una aplicación central y de librerías de multitud de temas que se pueden instalar según la necesidad.

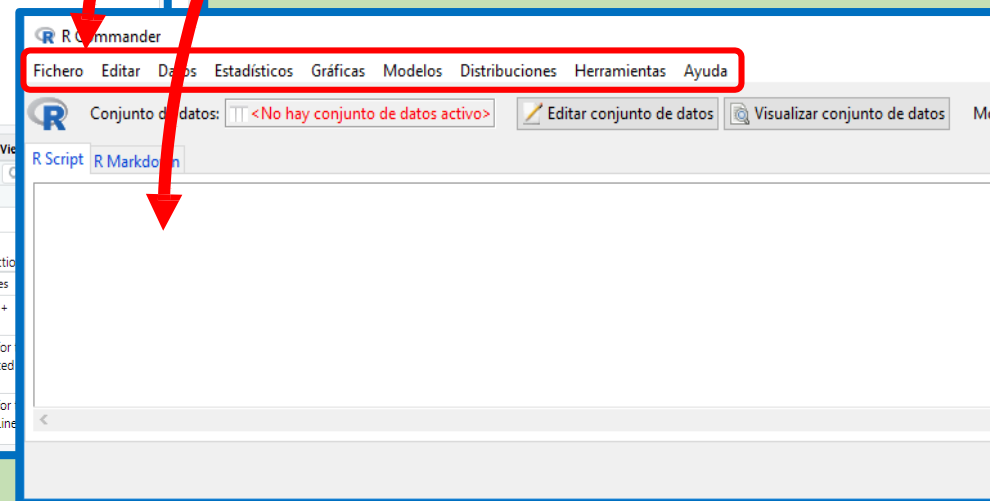
El R-Commander es una interfaz que permite el manejo del programa R mediante una ventana de menús. Este interfaz permite al usuario comenzar a manejar este programa sin conocer el lenguaje de instrucciones, y permite el aprendizaje de este lenguaje de forma sencilla (si el usuario así lo quiere).

Instalar del R-Commander por medio del RStudio



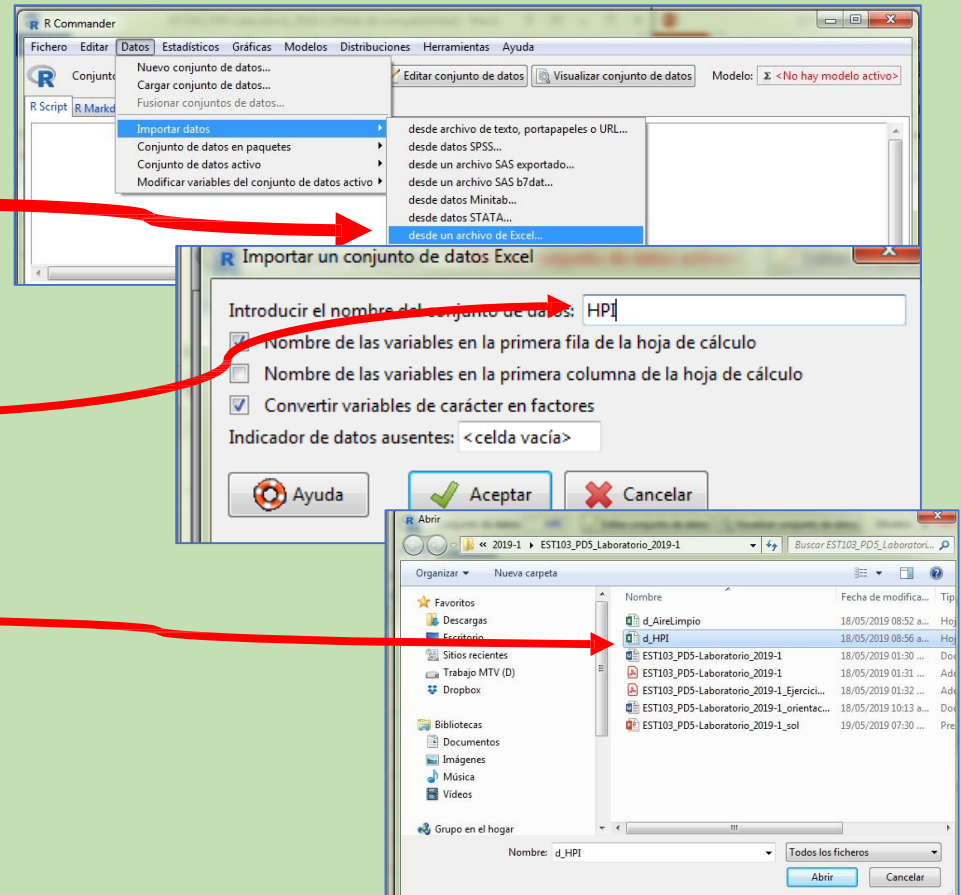
Un ambiente del RCommander, con:

- ✓ menús
- ✓ Área del Lenguaje R, el cual es generado al accionar el menú

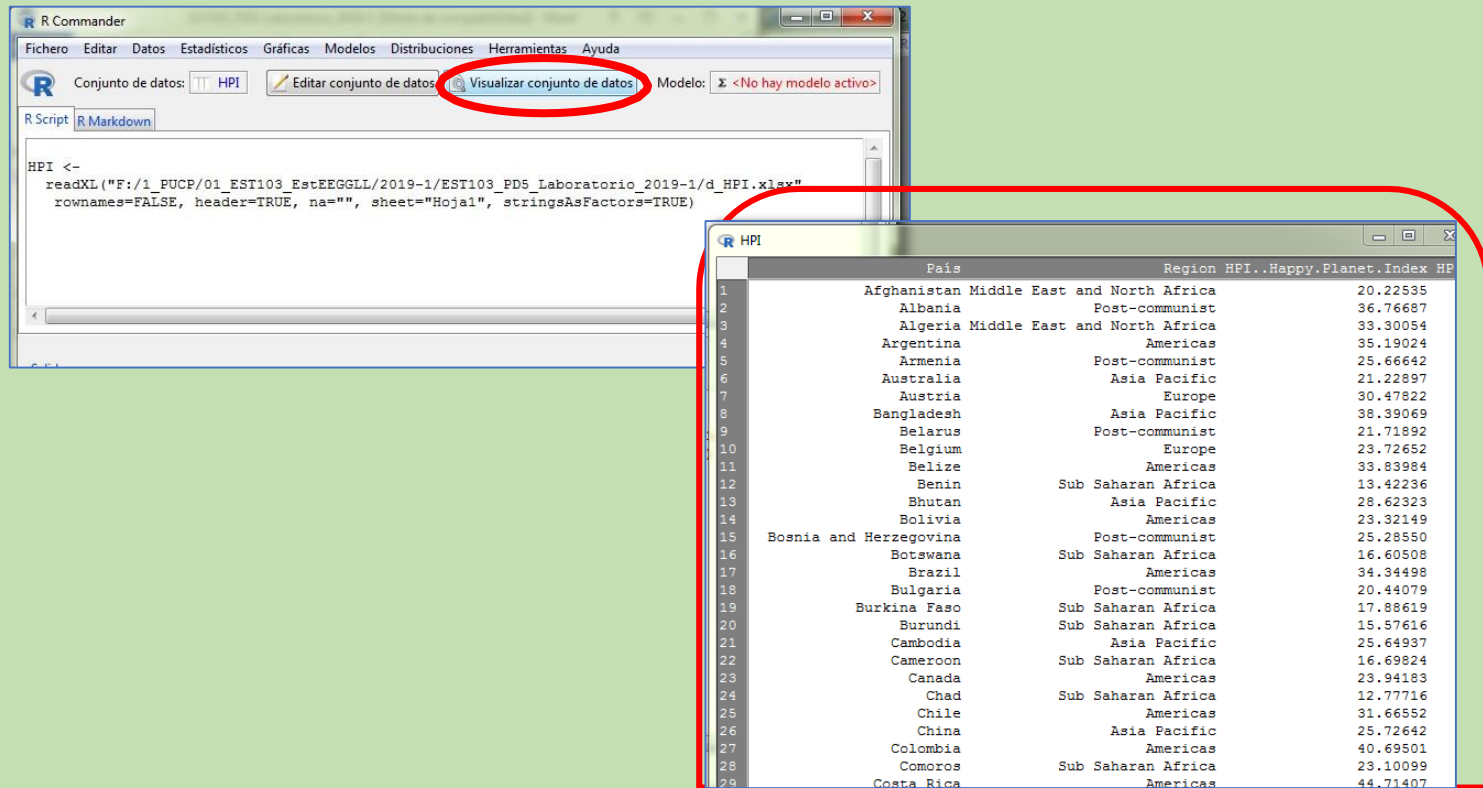


suba los datos del estudio del HPI:

- Menu Datos
- importar datos
- desde un archivo de Excel
- Introducir el nombre del conjunto de datos "HPI"
- Seleccionar el archivo Excel



Visualice los datos que acabó de subir:



The screenshot shows the R Commander interface. The 'Visualizar conjunto de datos' button is circled in red. Below it, the R script shows the command to read an Excel file. A preview window titled 'HPI' is open, displaying a table with the following data:

	Pais	Region	HPI..Happy.Planet.Index	HP
1	Afghanistan	Middle East and North Africa	20.22535	
2	Albania	Post-communist	36.76687	
3	Algeria	Middle East and North Africa	33.30054	
4	Argentina	Americas	35.19024	
5	Armenia	Post-communist	25.66642	
6	Australia	Asia Pacific	21.22897	
7	Austria	Europe	30.47822	
8	Bangladesh	Asia Pacific	38.39069	
9	Belarus	Post-communist	21.71892	
10	Belgium	Europe	23.72652	
11	Belize	Americas	33.83984	
12	Benin	Sub Saharan Africa	13.42236	
13	Bhutan	Asia Pacific	28.62323	
14	Bolivia	Americas	23.32149	
15	Bosnia and Herzegovina	Post-communist	25.28550	
16	Botswana	Sub Saharan Africa	16.60508	
17	Brazil	Americas	34.34498	
18	Bulgaria	Post-communist	20.44079	
19	Burkina Faso	Sub Saharan Africa	17.88619	
20	Burundi	Sub Saharan Africa	15.57616	
21	Cambodia	Asia Pacific	25.64937	
22	Cameroon	Sub Saharan Africa	16.69824	
23	Canada	Americas	23.94183	
24	Chad	Sub Saharan Africa	12.77716	
25	Chile	Americas	31.66552	
26	China	Asia Pacific	25.72642	
27	Colombia	Americas	40.69501	
28	Comoros	Sub Saharan Africa	23.10099	
29	Costa Rica	Americas	44.71407	

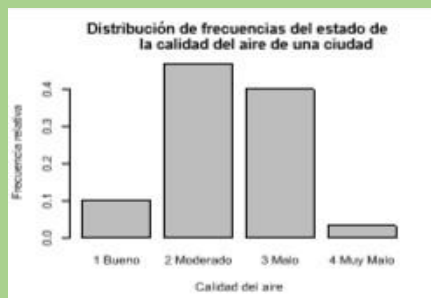
Ejemplo – Calidad del aire

Durante un mes se monitoreo el estado de la calidad del aire en una ciudad, estos fueron los resultados:

Bueno	Moderado	Bueno	Malo	Moderado	Malo
Malo	Moderado	Malo	Malo	Malo	Moderado
Moderado	Moderado	Moderado	Malo	Muy Malo	Malo
Moderado	Moderado	Malo	Moderado	Moderado	Malo
Malo	Moderado	Moderado	Bueno	Moderado	Malo

Obtenga su distribución de frecuencias y los gráficos de barras y de sectores circulares.

	n.j	f.j	p.j
1 Bueno	3	0.10000000	10.000000
2 Moderado	14	0.46666667	46.666667
3 Malo	12	0.40000000	40.000000
4 Muy Malo	1	0.03333333	3.333333



Menú del R-Commander

Datos >
Importar datos
> desde un
archivo de
Excel...

Estadísticos >
Resúmenes >
Distribución de
frecuencias...

Gráficas >
Gráfica de
barras...

Gráficas >
Gráfica de
sectores...

En RStudio:

```
library(readxl)
CalidadAire <-
readXL("d:/Users/catolica/Documents/00
2022-0_R/datos/CalidadAire.xlsx",
rownames=FALSE, header=TRUE, na="",
sheet="Hoja1", stringsAsFactors=TRUE)
View(CalidadAire)
```

```
install.packages("DescTools")
library(DescTools)
Freq(CalidadAire$CalidadDelAire)
```

```
n.j=table(CalidadAire$CalidadDelAire)
n=length(CalidadAire$CalidadDelAire)
f.j=n.j/n
```

```
barplot(f.j, xlab="Calidad del aire",
ylab="Frecuencia relativa",
main="Distribución de frecuencias de la
calidad del aire de una ciudad")
```

```
pie(f.j, main="Distribución de
frecuencias de la calidad del aire de
una ciudad")
```

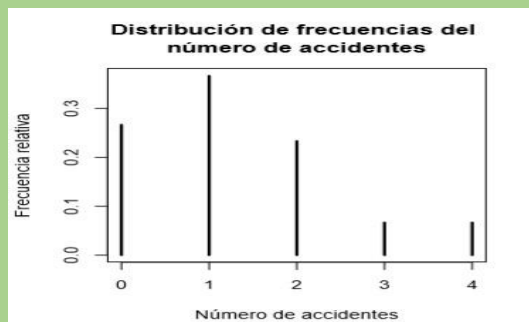
Ejemplo – Accidentes de tránsito por día

En un cierto distrito durante un mes se registró el número de accidentes de tránsito por día, estos fueron los resultados:

1	2	0	3	1	0	1	0	4	2
1	1	2	0	1	1	0	3	1	1
0	2	1	0	4	0	1	2	2	2

Note que la variable número de accidentes de tránsito por día en un distrito puede tomar los siguientes valores: 0, 1, 2, 3 y 4.

	level	freq	perc	cumfreq	cumperc
1	0	8	26.7%	8	26.7%
2	1	11	36.7%	19	63.3%
3	2	7	23.3%	26	86.7%
4	3	2	6.7%	28	93.3%
5	4	2	6.7%	30	100.0%



Menú del
R-Commander

Datos >
Importar datos
> desde un
archivo de
Excel...

Datos >
Modificar
variables del
conjunto de
datos activos >
Convertir
variable
numérica en
factor...

Estadísticos >
Resúmenes >
Distribución de
frecuencias...

Gráficas >
Dibujar una
variable
numérica
discreta...

En RStudio:

```
library(readxl)
accidentes <-
read_excel("d:/Users/catolica/Documents/
00_2021-1 R/accidentes.xlsx")
View(accidentes)
```

```
library(DescTools)
Freq(as.factor(accidentes$accidentes))
```

```
n.j=table(accidentes$accidentes)
n=length(accidentes$accidentes)
f.j=n.j/n
```

```
plot(f.j, type="h", xlab="Número de
accidentes", ylab="Frecuencia relativa",
main="Distribución de frecuencias del
número de accidentes", lwd=5)
```

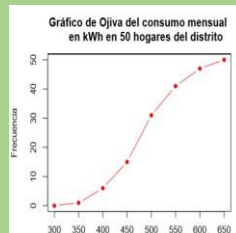
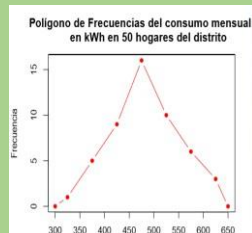
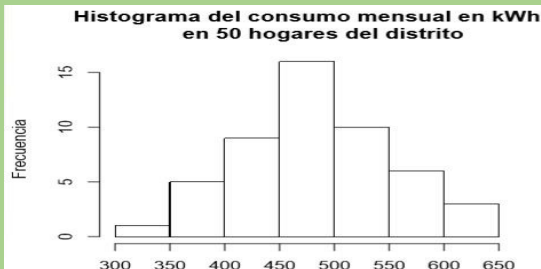
Ejemplo – Consumo de electricidad

Se registró el consumo de electricidad en kWh de 50 hogares obteniéndose:

589	493	531	355	469	432	415	468	617	426
300	439	464	430	403	525	478	392	432	459
398	372	488	481	620	484	509	522	488	502
596	567	466	477	580	555	520	525	425	650
384	497	438	501	521	452	508	462	457	577

Construya una distribución de frecuencias y muestre esta gráficamente.

	level	freq	perc	cumfreq	cumperc
1	[300,350]	1	2.0%	1	2.0%
2	(350,400]	5	10.0%	6	12.0%
3	(400,450]	9	18.0%	15	30.0%
4	(450,500]	16	32.0%	31	62.0%
5	(500,550]	10	20.0%	41	82.0%
6	(550,600]	6	12.0%	47	94.0%
7	(600,650]	3	6.0%	50	100.0%



Menú del
R-
Commander

Datos >
Importar
datos >
desde un
archivo de
Excel...

Estadísti
c os >
Resúmenes
>
Resúmenes
numéricos
...
Gráficas >
Histograma
...

En RStudio:

```
library(readxl)
consumo <-
  readXL("d:/Users/catolica/Documents/00 2022-
0_R/datos/consumo.xls",
         rownames=FALSE, header=TRUE, na="",
         sheet="Hoja1", stringsAsFactors=TRUE)
View(consumo)
```

```
library(DescTools)
Freq(consumo$electricidad)
```

```
hist(consumo$electricidad , xlab="Consumo en kwh",
      ylab="Frecuencia", main="Histograma del consumo
mensual en kwh en 50 hogares del distrito")
```

```
h=hist(consumo$electricidad, plot = FALSE)
```

```
x.pol=c(min(h$breaks),h$mids,max(h$breaks))
y.pol=c(0,h$counts, 0)
plot(x.pol,y.pol, type="b", col=2, lwd=2, pch=16,
      xlab="Consumo en kwh", ylab="Frecuencia",
      main="Polígono de Frecuencias del consumo mensual en
kwh en 50 hogares del distrito")
```

```
x.oj=c(h$breaks)
y.oj=c(0,cumsum(h$counts))
plot(x.oj,y.oj, type ="b", col=2, lwd=2,
      pch=16, xlab="Consumo en kwh",
      ylab="Frecuencia", main="Gráfico de la ojiva del
consumo mensual en kwh en 50 hogares del distrito")
```

Ejemplo – Consumo de electricidad, con 6 intervalos

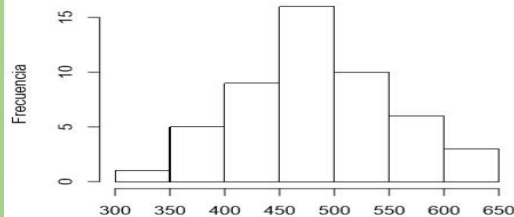
Se registró el consumo de electricidad en kWh de 50 hogares obteniéndose:

589	493	531	355	469	432	415	468	617	426
300	439	464	430	403	525	478	392	432	459
398	372	488	481	620	484	509	522	488	502
596	567	466	477	580	555	520	525	425	650
384	497	438	501	521	452	508	462	457	577

Construya una distribución de frecuencias y muestre esta gráficamente.

	level	freq	perc	cumfreq	cumperc
1	[300,350]	1	2.0%	1	2.0%
2	(350,400]	5	10.0%	6	12.0%
3	(400,450]	9	18.0%	15	30.0%
4	(450,500]	16	32.0%	31	62.0%
5	(500,550]	10	20.0%	41	82.0%
6	(550,600]	6	12.0%	47	94.0%
7	(600,650]	3	6.0%	50	100.0%

Histograma del consumo mensual en kWh en 50 hogares del distrito



Polígono de Frecuencias del consumo mensual en kWh en 50 hogares del distrito

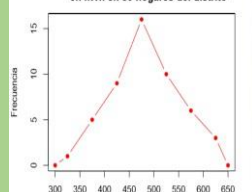
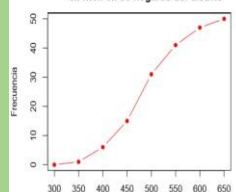


Gráfico de Ojiva del consumo mensual en kWh en 50 hogares del distrito



Menú del
R-Commander

Datos >
Importar
datos > desde
un archivo de
Excel...

Estadísticos
> Resúmenes >
Distribución
de
frecuencias...

Gráficas >
Dibujar una
variable
numérica
discreta...

En RStudio:

```
k=6 # Numero de intervalos
A=max(consumo$electricidad)-
min(consumo$electricidad) # Amplitud
c=ceiling(A/k) # Ancho de clase
b=seq(from=min(consumo$electricidad), by
=c, length.out=k+1) # Limites
```

```
hist(consumo$electricidad, breaks = b,
xlab="Consumo en kwh",
ylab="Frecuencia",
main="Histograma del consumo mensual en kwh
en 50 hogares del distrito" )
```