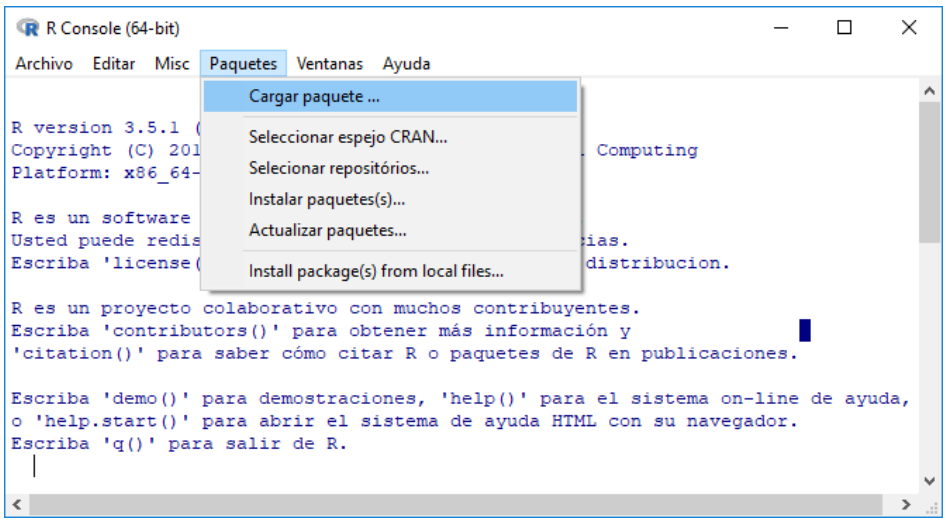
Práctica 1

Introducción a R Commander. Preparación de los datos

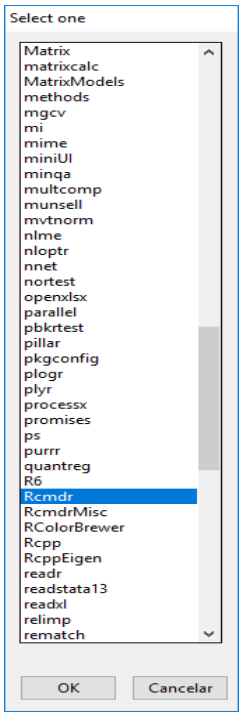
Contenido

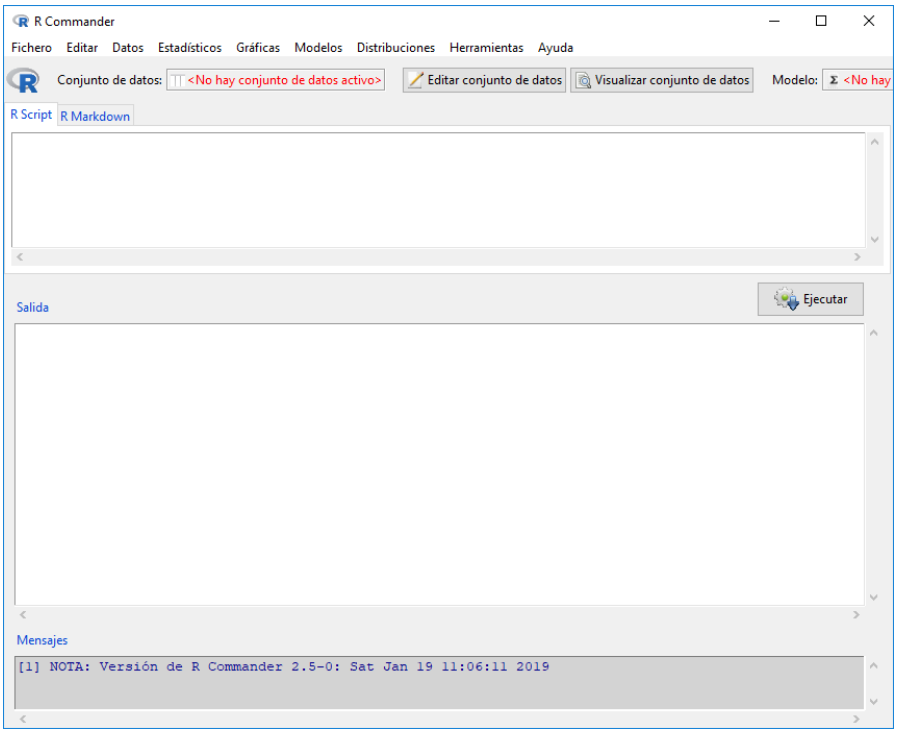
1. Estructura de R y R Commander

Una vez instalado el R con el R Commander debes cargarlo, este proceso se debe seguir cada vez que se inicia R. Tienes dos opciones para cargar el paquete, puedes escribir ***library(Rcmdr)*** en la R Console o bien usar el menú de la R Console eligiendo la opción "Cargar paquete…" en el menú Paquetes

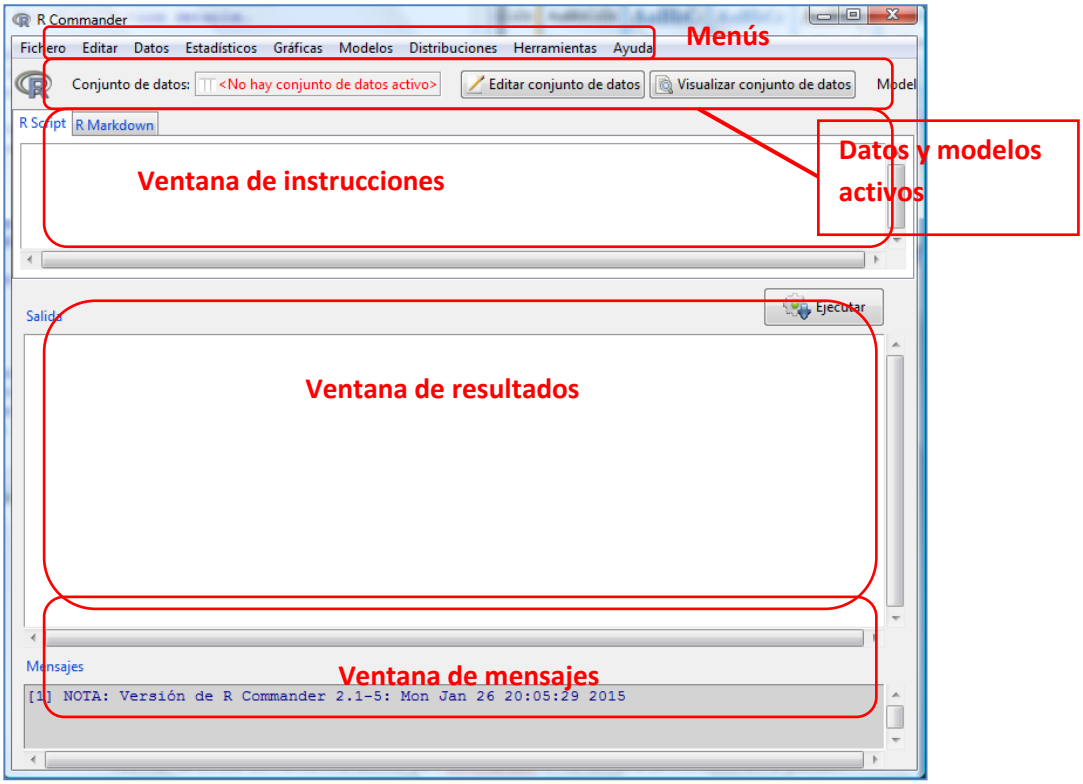


Eliges otra vez Rcmdr. La carga del paquete Rcmdr (R Commander), provoca que se abra la ventana interactiva R Commander que ofrece al usuario una interacción de menús más rica que la de R Console, y también permite trabajar en modo comando.





La ventana del R Commander consta de las siguientes partes: barra de menús, barra de elementos activos (conjuntos de datos y modelos), área de instrucciones, área de resultados y área de mensajes



Además de utilizar los menús interactivos de R Commander, en su ventana de instrucciones se pueden escribir una o varias instrucciones R, separadas por ";", o varias líneas de instrucciones, de modo similar a como se puede hacer en la ventana R Console. Para ejecutarlas con R Commander, se puede actuar, bien ubicando el cursor en una de las líneas de las instrucciones o bien seleccionando una o varias instrucciones en una misma línea o en varias, y pulsando el botón "*Ejecutar*" de la ventana R Commander, o también pulsando el botón derecho del ratón y seleccionando *"Ejecutar"*. Cuando se ejecutan instrucciones, en la ventana de resultados aparecen las instrucciones ejecutadas y los resultados que producen, y si hay mensajes/errores aparecen en la ventana de Mensajes. Se puede escribir texto en cualquiera de las ventanas de R Commander pero las instrucciones sólo se pueden ejecutar si están en su ventana asociada.

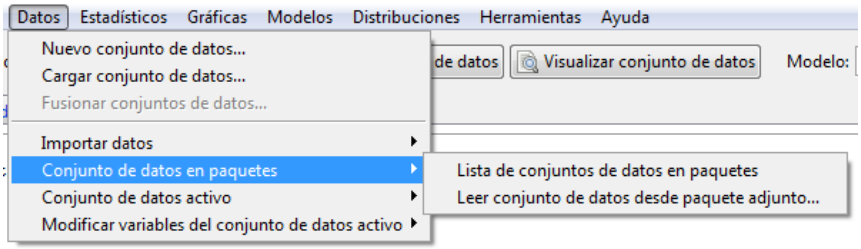
Menús Principales de *Rcmdr* :
Fichero: Menú desde el que podemos cambiar el directorio de trabajo, abrir y guardar archivos y ficheros y salir de R Commander y/o de R
Editar: Cortar, pegar, seleccionar, limpiar ventana,… Pulsando con el botón derecho encima de las ventanas de instrucciones y resultados aparece también un menú contextual.
Datos: Submenús para lectura y manipulación de datos.
Estadísticos: Submenús para realizar análisis estadísticos.
Gráficas: Submenús para realizar gráficos estadísticos.
Modelos: Submenús para obtener resúmenes numéricos, intervalos de confianza, tests de hipótesis y gráficos complejos.
Distribuciones: Submenús de distribuciones de probabilidad estándar.
Herramientas: Submenús para cargar otros paquetes o datos de *R*.
Ayuda: Menú con ayuda para *Rcmdr* y un manual del mismo.

2. Carga de un conjunto de datos

En *Rcmdr* siempre debe existir un conjunto de datos activo, cuando se abre el R Commander no está cargado ningún conjunto de datos, por tanto lo primero que debes hacer es introducir un conjunto de datos. Existen varias formas de introducir un conjunto de datos.

2.1. Conjunto de datos en paquetes

Puedes acceder a una colección de datos de ejemplos propios de R.

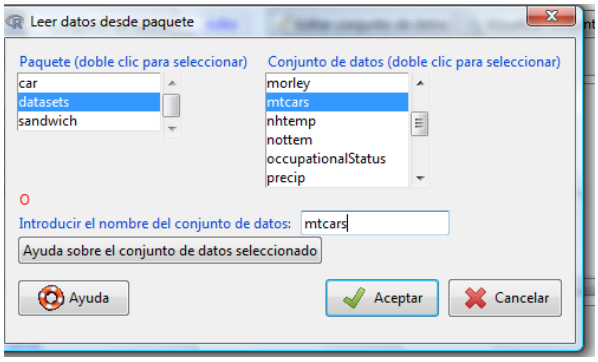


Si seleccionas *Lista de conjuntos de datos en paquetes*, te aparecerá una relación de conjuntos de datos ligados a paquetes, como los llamados *car* y *datasets*.

Si eliges *Leer conjunto de datos desde paquete adjunto*, podrás seleccionar unos datos determinados como conjunto de datos activo para el R Commander.

Ejemplo 1. Lee el conjunto de datos *mtcars* del paquete de datos *datasets*.

Solución: Sigue la ruta **Datos → Conjunto de datos en paquetes → Leer conjunto de datos desde paquete adjunto** y selecciona:



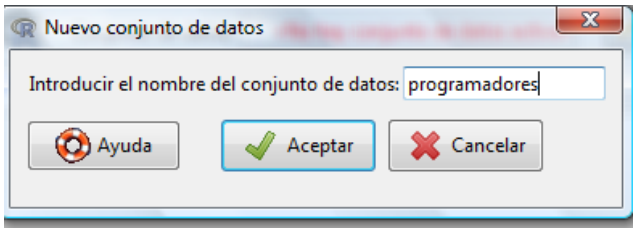
El conjunto *mtcars* será uno de los utilizados durante el curso. Los datos han sido extraídos de la revista *Motor Trend* US. Se trata de un estudio sobre el consumo de combustible y 10 aspectos del diseño de automóviles y rendimiento para 32 automóviles (modelos de 1973-1974). Tiene 32 observaciones (registros) y 11 variables:

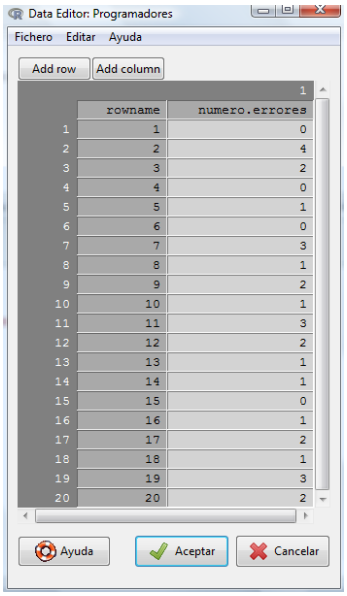
mpg: Consumo, en millas/galón
cyl: Número de cilíndros
disp: Cilindrada, en pulgadas cúbicas (cu.in.)
hp: Potencia, en cv
drat: Relación del eje trasero
wt: Peso, en lb/1000
qsec : Tiempo que tarda en recorrer ¼ milla, en segundos
vs: V/S (vs = 0 si tiene cilindros en V y vs = 1 si cilindros en serie)
am: Transmisión (0 = automática, 1 = manual)
gear: Número de velocidades hacia adelante
carb: Número de carburadores

2.2. Nuevo conjunto de datos

Si el conjunto de datos es pequeño puedes hacerlo directamente desde el menú eligiendo **Datos → Nuevo conjunto de datos** e introduciendo los nombres de las variables y las categorías (variable cualitativa o factor) o valores (variable cuantitativa o numérica) de cada una de ellas en una ventana emergente tipo hoja de cálculo. Cada columna corresponde a una variable y cada fila a un registro o elemento de la muestra). Cuando un dato no está disponible debes escribir en la celda correspondiente el código NA.

Ejemplo 2. Crea un nuevo conjunto de datos a partir de 20 valores de la variable número de errores en 100 líneas de código: 0, 4, 2, 0, 1, 0, 3, 1, 2, 1, 3, 2, 1, 1, 0, 1, 2, 1, 3, 2.
Solución: Sigue la ruta **Datos → Nuevo conjunto de datos**, dale un nombre al conjunto de datos (por ejemplo, *programadores*) y rellena la tabla escribiendo un dato en cada fila

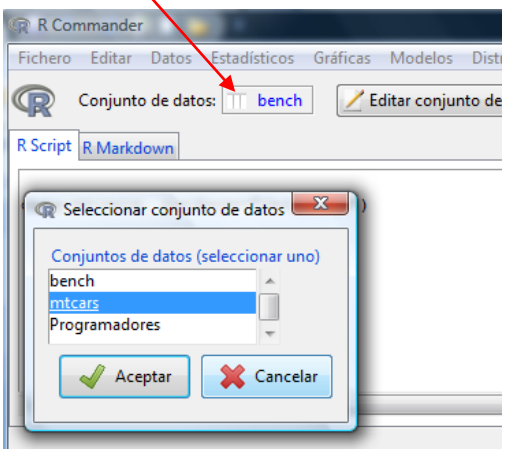




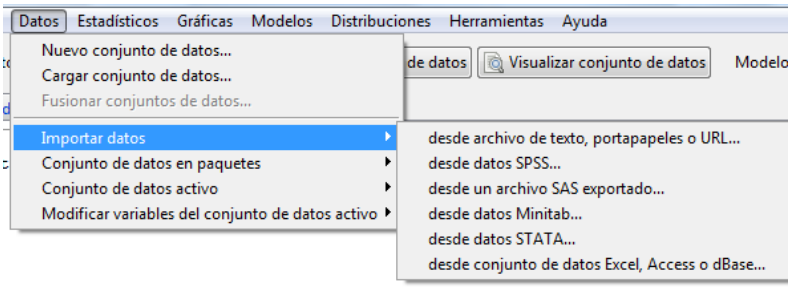
Alternativamente se puede crear el nuevo conjunto de datos desde la ventana de instrucciones, ejecutando las siguientes instrucciones:

```
error<-c(0,1,2,3,4)
freq<-c(4,7,5,3,1)
numero.errores<- rep(error,freq)
programadores<- as.data.frame(numero.errores)
```

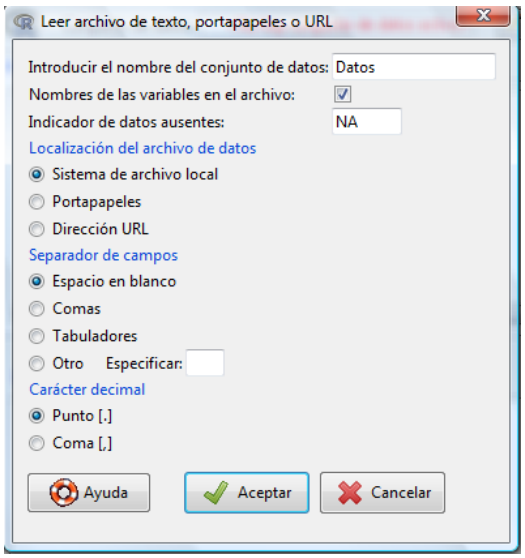
Ejemplo 3. Cambia el conjunto de datos activo y selecciona el conjunto *mtcars*.
Solución:



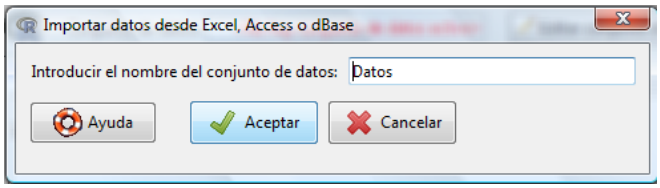
2.3. Importar datos

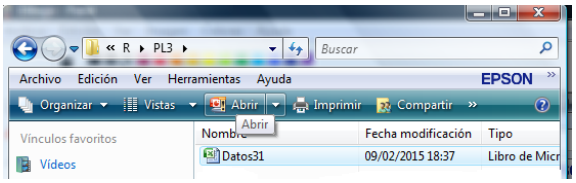


Desde un archivo de texto, portapapeles o URL, …



Desde conjuntos de datos Excel, Access o dBase, …



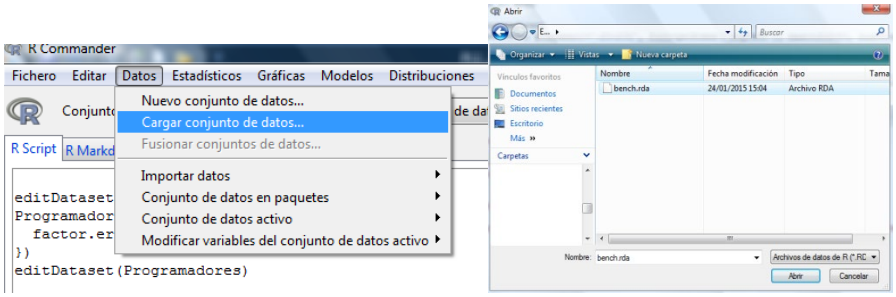


2.4. Cargar conjunto de datos

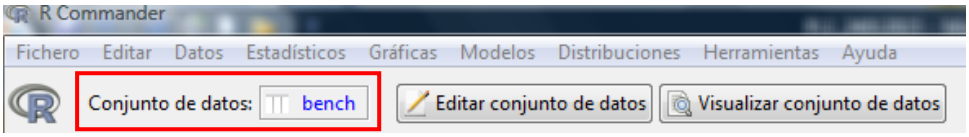
Lee los datos de un fichero que tienes almacenado en tu ordenador

Ejemplo 4. Abre el banco de datos *bench*.

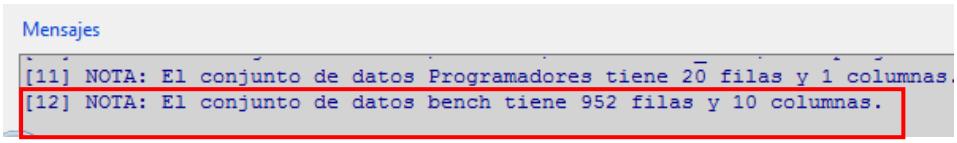
Solución:



Ejemplo 5. Identifica el número de variables y el número de registros (individuos en la muestra) del banco de datos *bench*.
Solución: En primer lugar debes comprobar que está activo el banco de datos *bench*.



y leer en la ventana de mensajes el que corresponde al conjunto de datos *bench*.



En total se dispone de 952 registros y 10 variables.
Los datos que acabas de cargar se han obtenido del Benchmarksgame de Debian y corresponden a la ejecución del mismo programa, en el mismo ordenador, con la misma carga de trabajo pero usando lenguajes de programación diferentes. Las variables observadas son:

name: Nombre del algoritmo ejecutado.
lang: Lenguaje en que se implementó el algoritmo.
id: ¿? Esta variable no nos interesa
n: Tamaño del problema resuelto.
size.B.: Tamaño en octetos (1 octeto = 1 Bytes = 8 bits) del fichero de código fuente comprimido sin comentarios ni espacios.
cpu.s.: Tiempo, en segundos, usado realmente por el procesador, tanto sistema como usuario.
mem.KB.: Cantidad de memoria usada, en KBytes.
status: Un cero significa que la ejecución concluyó con éxito. Como todos sus valores son iguales a 0, esta variable no nos interesa
load: Porcentaje de tiempo no ocioso en cada núcleo; es un procesador con cuatro núcleos.
elapsed.s.: Tiempo real ocupado, en segundos.

Puedes ver las variables en **Visualizar conjunto de datos** o siguiendo la secuencia **Datos → Conjunto de datos activo → Variables del conjunto de datos activo.**

Cambiar el nombre de un conjunto de datos activo. Para cambiar el nombre del conjunto de datos activo debes escribir en la ventana de instrucciones

Nuevo nombre del conjunto de datos <- Nombre actual del conjunto de datos

y ejecutar la instrucción.

Si quieres guardar el conjunto de datos con el nuevo nombre debes activarlo seleccionándolo en el botón de *conjunto de datos* y seguir las instrucciones siguientes.

Guardar el conjunto de datos activo. Para guardar el conjunto de datos activo debes seguir la secuencia **Datos → Conjunto de datos activo → Guardar el conjunto de datos activo**.

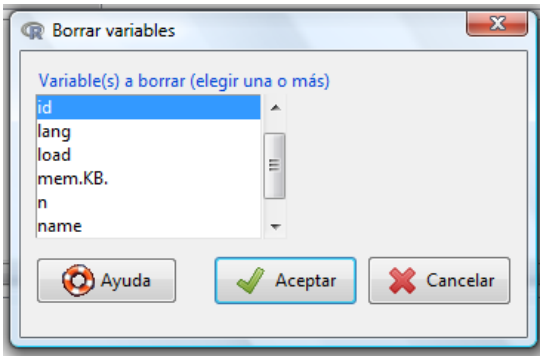
3. Preparación de los datos

En esta sección vamos a ver cómo puedes realizar algunas operaciones con los datos siguiendo la ruta **Datos → Conjunto de datos activo** o la ruta **Datos → Modificar variables del conjunto de datos activo**.

3.1. Borrar variables del conjunto de datos activo

Ejemplo 6. Ya que las variables *id* y *status* del conjunto de datos *bench* no nos interesan ¿cómo puedes eliminarlas?

Solución: Lo único que tienes que hacer es seguir la secuencia **Datos → Modificar variables del conjunto de datos activo → Eliminar variables del conjunto de datos**. En la ventana emergente, llamada **Borrar variable**s, debes seleccionar las variables a borrar, en este caso *id* y *status*, a continuación aceptar y después confirmar. El nuevo conjunto de datos *bench* tiene 8 variables.



3.2. Filtrar el conjunto de datos activo

En ocasiones nos conviene analizar solamente una parte de los datos, aquellos que cumplen determinada característica. Vamos a ver cómo puedes realizar un filtrado de datos.

Ejemplo 7. Sólo nos interesan los datos de las variables del conjunto *bench* para los que la variable size.B. >= 1000 ¿Cómo puedes obtener el conjunto de datos filtrado?

Solución: Sigue la ruta **Datos → Conjunto de datos activo → Filtrar el conjunto de datos activo.** Aparece una ventana titulada **Filtrar el conjunto de datos**, la parte superior de la misma permite escoger columnas (es decir, variables) concretas; lo habitual será que interese dejar seleccionada la opción por omisión: *Incluir todas las variables*.

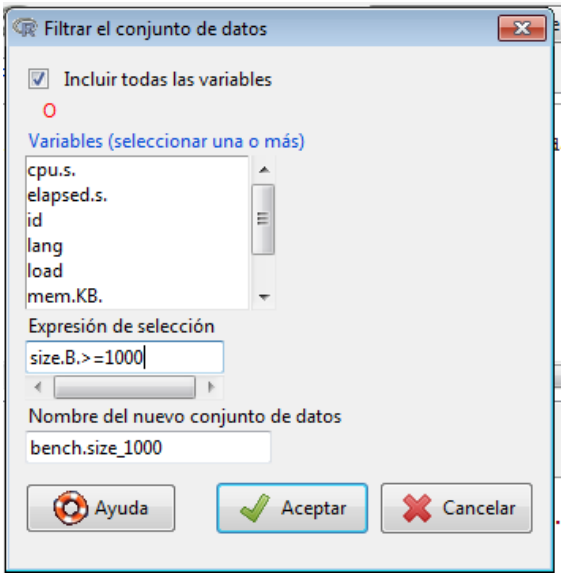
Expresión de selección. En este caso, debes escribir:

size.B. >= 1000

Nombre del nuevo conjunto de datos. Por omisión aparece *<igual que en el conjunto de datos activos>*. Conviene cambiarlo a otro nombre, porque si lo dejas así el resultado del filtro machaca (sustituye) al conjunto original, y los datos originales se pierden. Por ejemplo, le llamas *bench.size_1000*. El nombre puede usar letras, cifras, puntos y subrayados (guiones bajos), pero no espacios, ni guiones, ni muchos otros. Para finalizar, comprueba que en la barra de elementos activos aparece

Conjunto de datos: *bench.size_1000*

El conjunto de datos filtrado tiene 417 filas y 10 columnas.



Mensajes

```
[2] NOTA: El conjunto de datos bench tiene 952 filas y 10 columnas.
[3] NOTA: El conjunto de datos bench.size_1000 tiene 417 filas y 10 columnas.
```

El siguiente cuadro muestra las correspondencias entre expresiones habituales a la hora de establecer una condición y los símbolos que usa R-Commander:

igual (=)	==
distinto (≠)	!=
menor o igual (≤)	<=
mayor o igual (≥)	>=
conjunción (y)	&
disyunción (o)	|

También se pueden usar paréntesis para agrupar. Por otro lado, al usar un valor numérico, tienes que usar un punto decimal (no coma); un valor textual, debes entrecomillarlo con el símbolo " o el apóstrofo '.

Ejemplo 8. Selecciona solamente los datos correspondientes a ejecuciones del programa para las que el tiempo real ocupado haya sido de menos de un segundo y no hayan utilizado el lenguaje de programación Python 3.
Solución: Creamos un nuevo fichero de datos que llamaremos *filtro* y en este caso la expresión de selección será

(elapsed.s.<1) & (lang != "Python 3")

Mensajes

```
[8] NOTA: El conjunto de datos bench tiene 952 filas y 10 columnas.
[9] NOTA: El conjunto de datos filtro tiene 291 filas y 10 columnas.
```

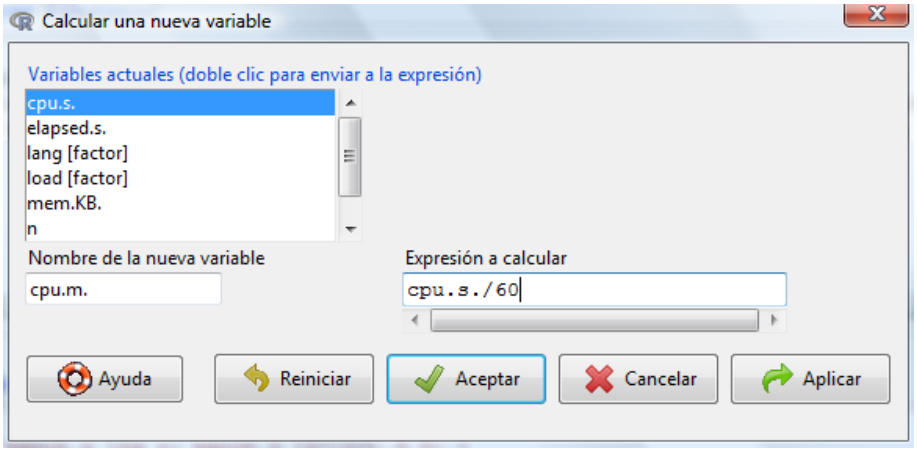
La selección ha creado el nuevo conjunto de datos con 291 de los 952 datos del fichero bench.

Advertencia: Es habitual olvidarse de que se ha establecido un filtro, y realizar subsiguientes cálculos con el conjunto filtrado sin pretenderlo. Recuerda volver al conjunto de datos original pinchando el botón *Conjunto de datos* en la barra de elementos activos de *Rcmdr*.

3.3. Calcular una nueva variable

Permite definir una nueva variable mediante una expresión matemática.

Ejemplo 9. Pasa de segundos a minutos la variable *cpu.s.* y llama *cpu.m.* a la nueva variable.
Solución: En primer lugar debes cerciorarte de que el conjunto de datos activo es el conjunto *bench*. A continuación sigue la ruta **Datos → Modificar variables del conjunto de datos activo → Calcular una nueva variable.** Tienes que rellenar la ventana emergente de la forma siguiente:



Ahora el conjunto de datos *bench* tiene 9 columnas (variables).

3.4. Convertir una variable numérica en factor

Los valores que toma una variable estadística se llaman modalidades. Si las modalidades son cantidades (números) la variable se dice cuantitativa o numérica (por ejemplo, las variables que representan alguna magnitud: velocidad, edad, tiempo, etc.); si las modalidades son nombres (etiquetas, atributos, niveles...) entonces la variable se dice cualitativa o categórica o factor.

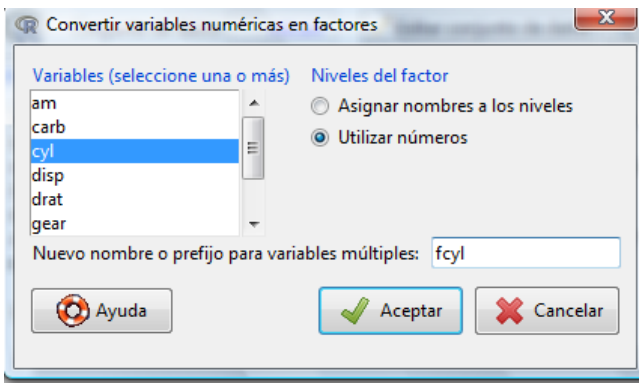
A la hora de trabajar con una variable en R Commander, es importante ser consciente de si se trata de una variable numérica o de un factor, pues hay procedimientos que sólo se pueden aplicar a uno de los tipos. Por ejemplo, sólo se puede hacer un gráfico de barras si la variable es factor; si es numérica, hay que convertirla antes a factor.

Al convertir a factor, tienes dos opciones: *asignar nombres a los niveles* o *utilizar números*. Para variables cuantitativas suele ser conveniente usar los mismos valores numéricos como etiquetas de las modalidades.

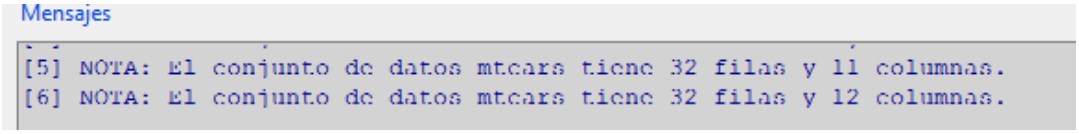
Si para *Nuevo nombre* del factor mantienes la opción por omisión *<igual que las variables>*, se pierde el carácter cuantitativo de la variable. Esto significa que, si posteriormente se desean obtener descriptivos cuantitativos hay que dar explícitamente un *Nuevo nombre* distinto del original.

Ejemplo 10. Convierte en factor la variable *cyl* del conjunto de datos *mtcars*.

Solución: Para realizar esta transformación debes seguir la secuencia **Datos → Modificar variables del conjunto de datos activo → Convertir variable numérica en factor**, en la ventana emergente selecciona *cyl*, marca *Utilizar números*, escribe el nuevo nombre (*fcyl*) y acepta.

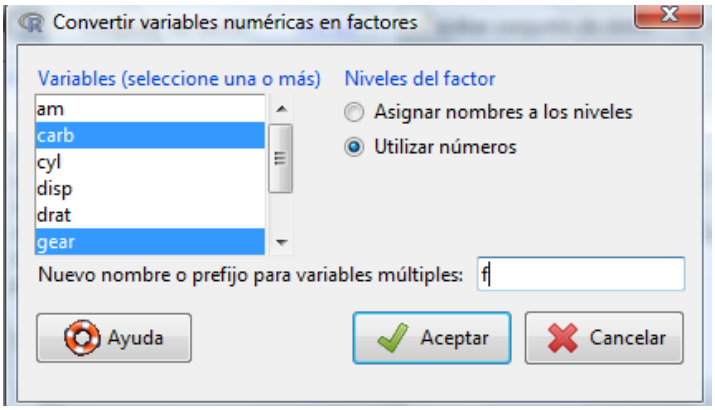


Observa, en la ventana de mensajes, que el conjunto de datos *mtcars* tiene ahora una variable más.

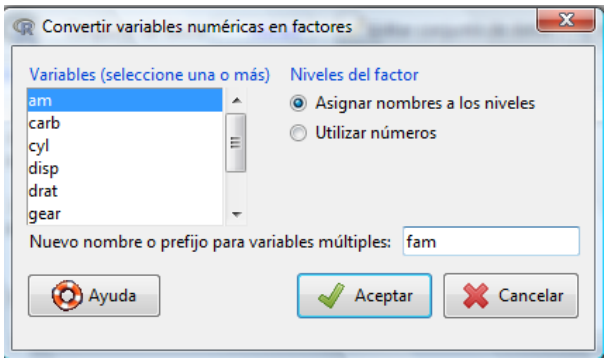


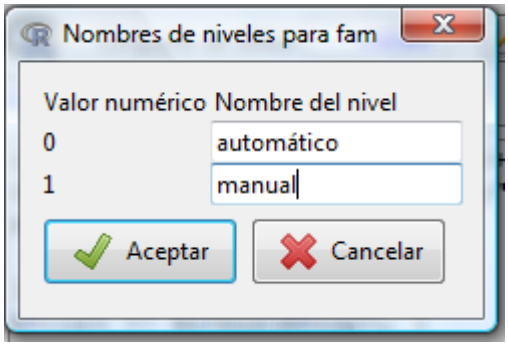
Si quieres puedes guardar el nuevo conjunto de datos con otro nombre, para ello debes seguir la secuencia **Datos → Conjunto de datos activo → Guardar el conjunto de datos activo.**

Ejemplo 11. Convierte simultáneamente en factor las variables *carb* y *gear* del conjunto de datos *mtcars*.
Solución: Para realizar esta transformación debes seguir la secuencia **Datos → Modificar variables del conjunto de datos activo → convertir variable numérica en factor**, en la ventana emergente selecciona *carb* y *gear*, marca *Utilizar números*, escribe el prefijo para variables múltiples (*f*) y acepta.



Ejemplo 12. Convierte en factor la variable *am* del conjunto de datos *mtcars* y cambia *0* por *automático* y *1* por *manual*.
Solución: Para realizar esta transformación debes seguir la secuencia **Datos → Modificar variables del conjunto de datos activo → convertir variable numérica en factor**, en la ventana emergente selecciona *am*, escribe el nuevo nombre (*fam*) y acepta.





3.5. Recodificar variables

Nos permite crear una nueva variable, generalmente es útil para crear una variable discreta a partir de una continua.
Las expresiones que irán en Introducir directrices de recodificación son:

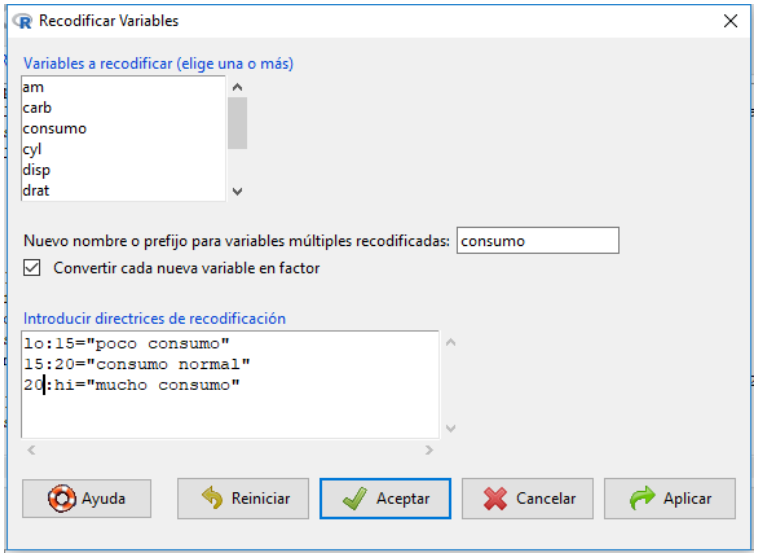
- un valor simple: "Alta"=1
- varios valores separados por comas: 7,8,9="alto"

- un rango de valores indicados por dos puntos: 7:9="alto". El caso especial de los valores no acotados lo (lowest) y hi (highest) son admitidos.
- el comando especial else, con el que hay que tener cuidado en su utilización, puesto que es aplicable en cualquier caso que no sean los anteriores, incluso si la celda está en blanco.

Ejemplo 13. Agrupa los valores de la variable mpg del conjunto de datos mtcars en una nueva variable que tenga 3 categorías:

- poco consumo si mpg es menor de 15 millas por galón
- consumo normal si mpg está entre 15 y 20 millas por galón
- mucho consumosi supera las 20 millas por galón en mpg

Solución: Para realizar esta transformación debes seguir la secuencia **Datos → Modificar variables del conjunto de datos activo → Recodificar variables**, en la ventana emergente selecciona *mpg*, escribe el nuevo nombre (*consumo*), activa *convertir cada nueva variable en factor*, *introduce directrices de codificación* y acepta.




```
             consumo
consumo normal:12
mucho consumo :14
poco consumo  : 6
```

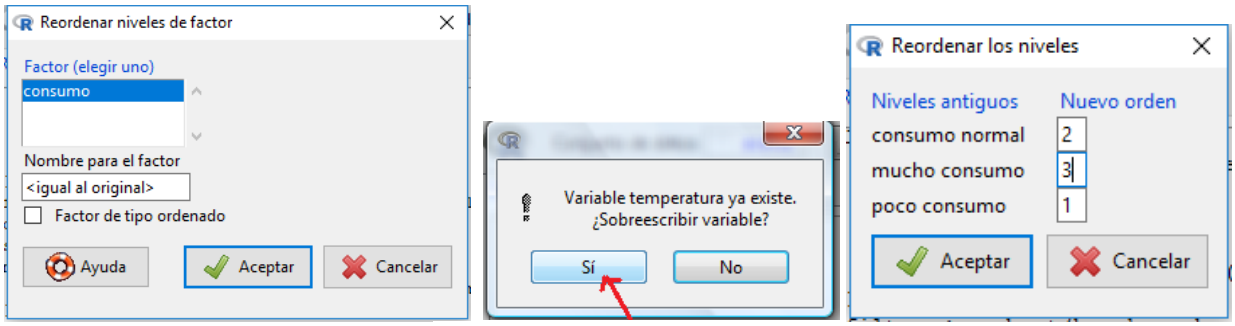
Nota: En casos especiales en los que se necesita agrupar los datos de una variable numérica en segmentos equidistantes (el rango total se divide en intervalos de la misma longitud) o segmentos de igual cantidad (todas las categorías tendrán igual porcentaje de individuos) se tiene la opción de **Segmentar una variable numérica** entre las opciones de **Modificar variables del conjunto de datos activo**.

3.6. Reordenar niveles de factor

A las modalidades de una variable estadística cualitativa (factor), *R* les llama *niveles*. Si la variable estadística cualitativa es ordinal debemos indicarle a *R* el orden de los niveles ya que, por defecto, *R* los ordena alfabéticamente.

Ejemplo 14. Reordena los niveles de la variable *consumo,* creada a partir de la recodificación de *mpg* en el *Ejemplo 13* de forma que el nuevo orden sea *Poco*, *Normal*, *Mucho*.

Solución: Para realizar esta transformación debes seguir la secuencia **Datos → Modificar variables del conjunto de datos activo → Reordenar niveles de factor**, en la ventana emergente selecciona *consumo,* activa *Factor de tipo ordenado* y acepta, sobrescribe (o no) la variable, proporciona el nuevo orden y acepta.




```
                consumo
poco consumo   : 6
consumo normal:12
mucho consumo  :14
```

Otra base de datos con la que trabajaremos a lo largo del curso es el *acero2*, cuyo origen es el siguiente.

Con el fin de analizar el consumo energético de una empresa productora de acero se inspeccionó cada una de las tres líneas de producción de dicha empresa. La inspección duró cinco días; en los cuatro primeros, se inspeccionó cada una de las ocho horas del turno; el último día, sólo se inspeccionaron siete horas (todas menos la última). La inspección suponía registrar los valores de las variables más relevantes, estas son:

consumo: Consumo energético de la empresa, en megavatios/hora.
pr.tbc : Producción del tren de bandas calientes, en toneladas de acero/hora.
pr.cc: Producción de colada continua, en toneladas de acero/hora.
pr.ca: Producción del convertidor de acero, en toneladas de acero/hora.
pr.galv1: Producción de galvanizado de tipo I, en toneladas de acero/hora.
pr.galv2: Producción de galvanizado de tipo II, en toneladas de acero/hora.
pr.pint : Producción de chapa pintada, en toneladas de acero/hora.
linea : Línea de producción empleada (A, B o C) esa hora.
hora: Hora en la que se recogieron los datos (1ª, 2ª,… , 8ª).
temperatura: Temperatura del sistema en esa hora laborable: Alta, Media y Baja.
averias: Presencia de averías (Si, No) en esa hora.
naverias: Número de averías detectadas por hora.
sistema : Activación de un sistema de detección de sobrecalentamiento en esa hora de trabajo: encendido (ON), apagado (OFF).
ProdTotal: Producción total de la empresa, en toneladas de acero/hora.
NOx: Emisiones de mezcla de óxidos de nitrógeno, en toneladas de NO2/hora.
CO: Emisión de óxidos de monóxido de carbón , en toneladas /hora.
COV : Emisión de compuestos orgánicos volátiles, en toneladas de Ch1.85/hora.
SO2: Emisión de dióxido de azufre, en toneladas /hora.
CO2 : Emisión de dióxido de carbono, en toneladas/hora.
N2O: Emisión de óxido nitroso, en toneladas /hora.

4. Ejercicios propuestos

Usa el conjunto de datos *acero2* para realizar los ejercicios siguientes.

Ejercicio 1. Identifica el número de registros y de variables del conjunto de datos *acero2* .

Ejercicio 2. ¿Cómo puedes obtener los datos que corresponden a la variable *consumo* cuando la temperatura es alta y no hay averías?

Ejercicio 3. Define una variable *pr.galv_total* que nos proporcione la producción total de galvanizado (tipos I y II) cuando hay producción de los dos tipos.

Ejercicio 4. Convierte en factor la variable numérica *naverias*, a la nueva variable llámale *fnaverias*.

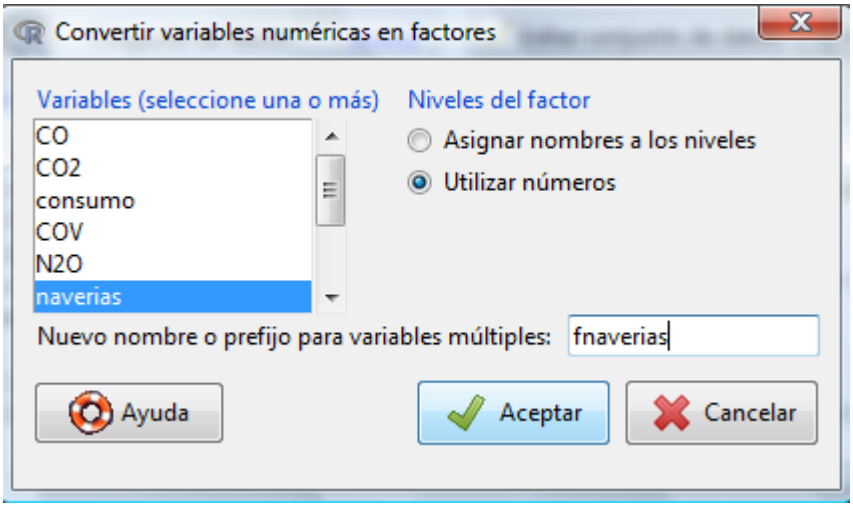
Soluciones

Ejercicio 1. Tiene 117 registros (filas) y 20 variables (columnas).

Ejercicio 2. Debes filtrar los datos (**Datos → Conjunto de datos activo → Filtrar el conjunto de datos activo**) de acuerdo al filtro: *temperatura == "Alta" & averias == "No"*. Recuerda renombrar el conjunto el nuevo conjunto de datos.

Ejercicio 3. Recupera el conjunto de datos *acero2*. Filtra los datos (**Datos → Conjunto de datos activo → Filtrar el conjunto de datos activo**) de acuerdo al filtro: *pr.galv1 > 0 & pr.galv2 > 0* y cambia el nombre del conjunto de datos. Ahora defines la nueva variable siguiendo la ruta **Datos → Modificar variables del conjunto de datos activo → calcular una nueva variable**, escribiendo *pr.galv_total* en *nombre de la nueva variable* y *pr.galv1 + pr.galv2* en *expresión a calcular* de la ventana emergente.

Ejercicio 4. Recupera el conjunto de datos *acero2*. Sigue la secuencia **Datos → Modificar variables del conjunto de datos activo → convertir variable numérica en factor**, en la ventana emergente selecciona *naverias*, marca *Utilizar números*, escribe el nuevo nombre y acepta



Observa que el conjunto de datos *acero2* tiene ahora una variable más: