

Práctica Gestión de Datos

Crear una nueva base de datos (CASAS)

La siguiente tabla contiene los datos de dos características de casas, concretamente los metros cuadrados útiles (área) y los precios de compra en miles de euros (precio).

Area	69	90	80	135	71	96	82	71	101	88	79	69	77	100	119
Precio	192	215	210	274	112	185	212	220	276	260	221	255	260	293	375

A. Mediante la opción Nuevo conjunto de datos crea un fichero de datos llamado CASAS y Cierra el editor.
CASAS <- edit(as.data.frame(NULL))

B. Comprueba que el conjunto de datos activo es el fichero Casas. Guarda el mismo en un directorio conveniente. La opción de guardar se encuentra dentro del menú conjunto de datos activo.
save("CASAS", file="D:/documentos/Dropbox/R OUC taller/Compartir/CASAS.rda")

C. Mediante la opción Editar conjunto de datos puedes continuar introduciendo datos en el fichero Casas. Si abandonas R-Commander no olvides volver a guardar el conjunto de datos activo.
fix(CASAS)

D. Crea una modalidad para indicar que las primeras 5 observaciones son de barrio A, las siguientes 5 del barrio B y las siguientes 5 del barrio C.
fix(CASAS)

E. Ejecuta el comando getwd() en la ventana de instrucciones y comprueba el directorio donde se guardan tanto ficheros de datos como de resultados. Ejecutando el comando setwd(directorio de trabajo) se cambia el directorio de trabajo al indicado dentro de las comillas.
getwd()
setwd("D:/documentos/Dropbox/R OUC taller/Compartir")

F. Mediante la opción Calcular una nueva variable del menú Modificar variables de los datos activos añadir una nueva variable al fichero que indique el precio por metro cuadrado.
CASAS\$Precio.m2 <- with(CASAS, Precio/ Área)

G. Crea una variable de modalidad (atributo) para indicar el tamaño de la casa, con la opción recodificar variables, de forma que las casas menores que 70 metros cuadrados sean pequeñas, las que tienen superficie entre 70 y 90 sean medianas y las que tienen más de 90 metros cuadrados sean grandes.
CASAS\$f.Área <- recode(CASAS\$Área, 'lo:69="Pequeña"; 70:90="Mediana"; 91:hi="Grandes"; ',
as.factor.result=TRUE)

También puede llevarse a cabo mediante la opción segmentar variable numérica, tiene diferentes opciones.
CASAS\$Área1 <- bin.var(CASAS\$Área, bins=3, method='intervals', labels=c('Pequeña','Mediana','Grande'))
Le decimos que nos
CASAS\$Área1 <- bin.var(CASAS\$Área, bins=3, method='intervals', labels=c('Pequeña','Mediana','Grande'))
CASAS\$Área2 <- bin.var(CASAS\$Área, bins=3, method='intervals', labels=NULL)
CASAS\$Área3 <- bin.var(CASAS\$Área, bins=3, method='proportions', labels=NULL)
CASAS\$Área4 <- bin.var(CASAS\$Área, bins=3, method='natural', labels=NULL)
La variable que crea con la segmentación sigue siendo numérica

```
mode(CASAS$Área4)
```

También podemos ver los intervalos que creo con la opción visualizar

```
library(relimp, pos=4)
```

```
showData(CASAS, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80, maxheight=30)
```

H. Guarda el fichero resultante mediante la opción Conjunto de datos activo

Guardar el conjunto de datos activo.

```
save("CASAS", file="D:/documentos/Dropbox/R OUC taller/Compartir/CASAS.rda")
```

El fichero **parqueeolico.dat**, contiene datos de la velocidad del viento, registrados durante 730 horas de forma simultánea, en dos localizaciones alternativas (Parque1 y Parque2). Se trataría de establecer la localización más aconsejable para la instalación de un parque de producción de energía eólica.

Abrir el archivo con el bloc de notas desde su ubicación.

A. La estructura de los datos es de dos columnas. Resulta más manejable si es transformada en dos variables, una continua que contenga las mediciones de viento y otra que sea un factor que indique la localización. Esto se realiza desde el menú: Datos -> Conjunto de datos activo-> Apilar variables del conjunto de datos activo ... En la ventana de diálogo se pide el nombre de la nueva base de datos que llamaremos eolico2, el nombre de la variable apilada, velocidad, y el nombre de la nueva variable factor, parque, cuyas clases se han denominado Parque1 y Parque2.

#Leemos los datos

```
DatosEolico <- read.table("D:/documentos/Dropbox/R OUC taller/Compartir/parque_eolico.dat",  
header=TRUE, sep=" ", na.strings="NA", dec=".", strip.white=TRUE)
```

```
fix(DatosEolico)
```

Función para apilarlos

```
Eolico.apilados <- stack(DatosEolico[, c("Parque1", "Parque2")])
```

```
names(Eolico.apilados) <- c("velocidad", "Parque")
```

```
library(relimp, pos=4)
```

```
showData(Eolico.apilados, placement='-20+200', font=getRcmdr('logFont'),  
maxwidth=80, maxheight=30)
```

Abrir una base de datos del programa, **Chile**, que esta dentro de la librería car.

```
data() # muestra las base de datos
```

```
data(Chile, package="car") # lee los datos y la selecciona como conjunto de datos activos
```

```
help("Chile") # información sobre los datos.
```

Observar que dice:

```
#education
```

```
#A factor with levels (note: out of order): P, Primary; PS, Post-secondary; S, #Secondary.
```

#Se ordenan los niveles del factor:

```
Chile$education <- factor(Chile$education, levels=c('P','S','PS'))
```

```
#Exportar datos a txt
```

```
write.table(Chile, "D:/documentos/Dropbox/R OUC taller/Compartir/Chile.txt", sep=" ", col.names=TRUE,
row.names=TRUE, quote=TRUE, na="NA")
```

```
# Importar los datos desde stata, SPSS
```

```
Autos <- read.dta("D:/documentos/Dropbox/R OUC taller/Compartir/auto.dta", convert.dates=TRUE,
convert.factors=TRUE, missing.type=TRUE, convert.underscore=TRUE,
warn.missing.labels=TRUE)
summary(Autos)
```

```
Osteoporosis <- read.spss("D:/documentos/Dropbox/R OUC taller/Compartir/osteoporosis.sav",
use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
```

```
# Importamos datos desde Excel la tabla Pregunta 2 del dades.xls
```

```
Democracia <- sqlQuery(channel = 1, select * from [Pregunta 2$])
names(Democracia) <- make.names(names(Democracia))
```

```
#Sale una nota verde que nos avisa que las filas cambiaron de nombre coloca #puntos en los espacios vacíos.
Podemos visualizar los datos.
```

```
showData(Democracia, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80, maxheight=30)
```

```
# se le puede asignar nombre a las filas, esto puede ser útil para los gráficos
```

```
row.names(Democracia) <- as.character(Democracia$Country)
```

```
Democracia$Country <- NULL
```

```
#Ejercicios con datos PULSE
```

```
#Cambiar el directorio de trabajo
```

```
setwd("D:/documentos/Dropbox/R OUC taller/Compartir")
```

```
# Abrir datos TXT
```

```
DatosPulso <- read.table("D:/documentos/Dropbox/R OUC taller/Compartir/EIPulso.txt", header=TRUE,
sep="", na.strings="NA", dec=".", strip.white=TRUE)
```

```
fix(DatosPulso)
```

```
# Guardar como rda
```

```
save("DatosPulso", file="D:/documentos/Dropbox/R OUC taller/Compartir/EIPulso.rda")
```

```
# a) Renombrar el nombre de las variables al castellano.
```

```
names(DatosPulso)[c(1)] <- c("Pulso1")
```

```
names(DatosPulso)[c(2)] <- c("Pulso2")
```

```
names(DatosPulso)[c(3)] <- c("Correr")
```

```
names(DatosPulso)[c(4)] <- c("Fumar")
```

```
names(DatosPulso)[c(5)] <- c("Sexo")
```

```
names(DatosPulso)[c(6)] <- c("Altura")
```

```
names(DatosPulso)[c(7)] <- c("Peso")
```

```
names(DatosPulso)[c(8)] <- c("Actividad")
```

```

#también se pueden hacer todos a la vez
names(DatosPulso)[c(1,2,3,4,5,6,7,8)] <-
c("Pulso1", "Pulso2", "Correr", "Fumar", "Sexo", "Altura", "Peso", "Actividad")

# b) Actualizar y visualizar los cambios realizados.
showData(DatosPulso, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80, maxheight=30)

# c) Cambiar el tipo de la variable Correr para que sea de tipo Caracter.
DatosPulso$Correr <- as.factor(DatosPulso$Correr)

# d) Recodifica la variable Correr en las categorías "Si" y "No".
DatosPulso$Correr <- recode(DatosPulso$Correr, '1==Si; 2==No; ', as.factor.result=TRUE)

# e) Recodifica la variable Sexo a "Hombre" y "Mujer".
DatosPulso$Sexo <- factor(DatosPulso$Sexo, labels=c('Hombre','Mujer'))

# f ) Recodifica la variable Fumar a "fuma" y "No fuma".
DatosPulso$Fumar <- factor(DatosPulso$Fumar, labels=c('Fuma','No Fuma'))

#g) Recodifica la variable Actividad a "Suave", "Moderada" y "Alta" esta hecha por la función de convertir
numérica
DatosPulso$Actividad <- factor(DatosPulso$Actividad, labels=c('Suave','Moderada','Alta'))

#h) La variable Altura viene dada en pulgadas (inches). Sabiendo que cada pulgada equivale a 2,54 cm,
cambiar esta variable a centímetros.
DatosPulso$Altura.cm <- with(DatosPulso, Altura*2.54)

#i ) La variable Peso viene dada en libras (pounds). Sabiendo que cada libra equivale a 454 gramos, cambiar
esta variable a kilos.
DatosPulso$Peso.kg <- with(DatosPulso, Peso*0.454)

#j ) Calcular una nueva variable difer.pulso que registre la diferencia entre las variables pulso2 y el pulso1
DatosPulso$difer.pulso <- with(DatosPulso, Pulso2-Pulso1)

#k) Guardar el conjunto de datos activo.. con el nombre EIPulso.rda
save("DatosPulso", file="D:/documentos/Dropbox/R OUC taller/Compartir/EIPulso.rda")

```