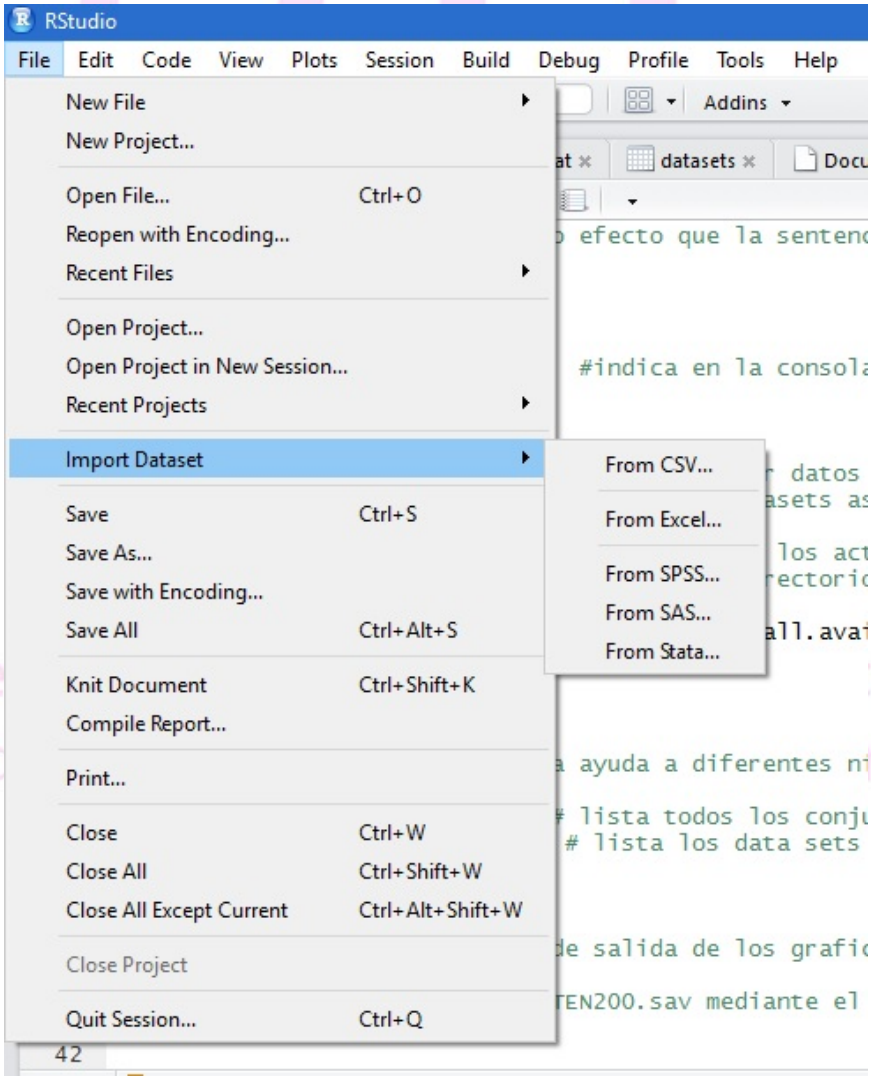
INTRODUCCIÓN

En esta práctica de laboratorio, que también puede considerarse de carácter introductorio al R, **pasaremos a un caso concreto de importación de datos.** Realizaremos su importación, veremos la estructura de los datos, y sencillas modificaciones de datos, tanto en sentido de las variables: creación, modificación, sustitución y eliminación, como en el caso de los datos: seleccionar casos de acuerdo a unas condiciones.

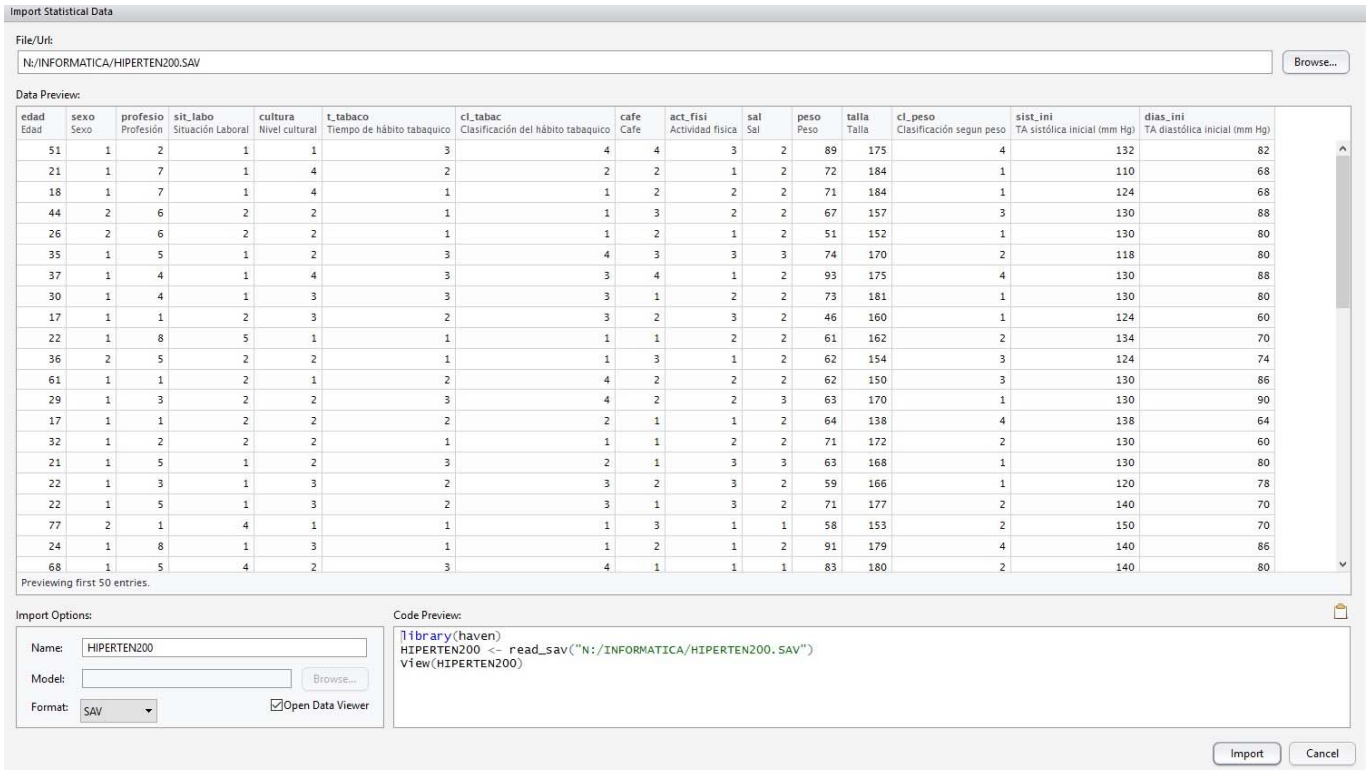
IMPORTACION – EXPORTACION DE CONJUNTOS DE DATOS A(de) R

Lo más usual es que tengamos nuestro propio conjunto de datos o lo descargamos desde la red, que usualmente no esta preparado para que se acceda directamente desde R. Nuestro conjunto de datos inicial es **HIPERTEN200.sav,** un fichero procedente del paquete IBMSPSS, con 200 casos y 19 variables.

Vamos a realizar esta tarea desde el interfaz de RStudio. Desde la primera opción del menú de la barra superior, podemos acceder a importar datos desde SPSS.



Posiblemente la primera vez nos indique que debemos actualizar unas librerías, accedemos a ello. Al final se nos abre una ventana para seleccionar el directorio y luego el fichero en cuestión. Entonces se nos aparece en el monitor la siguiente pantalla



edad Edad	sexo Sexo	profesio Profesión	sit_labo Situación Laboral	cultura Nivel cultural	t_tabaco Tiempo de hábito tabaquico	cl_tabac Clasificación del hábito tabaquico	cafe Cafe	act_fisi Actividad fisica	sal Sal	peso Peso	talla Talla	cl_peso Clasificación segun peso	sist_ini TA sistólica inicial (mm Hg)	dias_ini TA diastólica inicial (mm Hg)
51	1	2	1	1	3	4	4	3	2	89	175	4	132	82
21	1	7	1	4	2	2	2	1	2	72	184	1	110	68
18	1	7	1	4	1	1	2	2	2	71	184	1	124	68
44	2	6	2	2	1	1	3	2	2	67	157	3	130	88
26	2	6	2	2	1	1	2	1	2	51	152	1	130	80
35	1	5	1	2	3	4	3	3	3	74	170	2	118	80
37	1	4	1	4	3	3	4	1	2	93	175	4	130	88
30	1	4	1	3	3	3	1	2	2	73	181	1	130	80
17	1	1	2	3	2	3	2	3	2	46	160	1	124	60
22	1	8	5	1	1	1	1	2	2	61	162	2	134	70
36	2	5	2	2	1	1	3	1	2	62	154	3	124	74
61	1	1	2	1	2	4	2	2	2	62	150	3	130	86
29	1	3	2	2	3	4	2	2	3	63	170	1	130	90
17	1	1	2	2	2	2	1	1	2	64	138	4	138	64
32	1	2	2	2	1	1	1	2	2	71	172	2	130	60
21	1	5	1	2	3	2	1	3	3	63	168	1	130	80
22	1	3	1	3	2	3	2	3	2	59	166	1	120	78
22	1	5	1	3	2	3	1	3	2	71	177	2	140	70
77	2	1	4	1	1	1	3	1	1	58	153	2	150	70
24	1	8	1	3	1	1	2	1	2	91	179	4	140	86
68	1	5	4	2	3	4	1	1	1	83	180	2	140	80

Si pulsamos sobre ***import,*** podremos ver el contenido de dicho fichero en la ventana 1. Solo verlo, no podemos editarlo. Al mismo tiempo podemos observar que aparece como data frame en la ventana 2. Observamos que los datos han sido importados correctamente, veamos como los ha leído realmente, ejecutamos la sentencia estructura de un dataframe

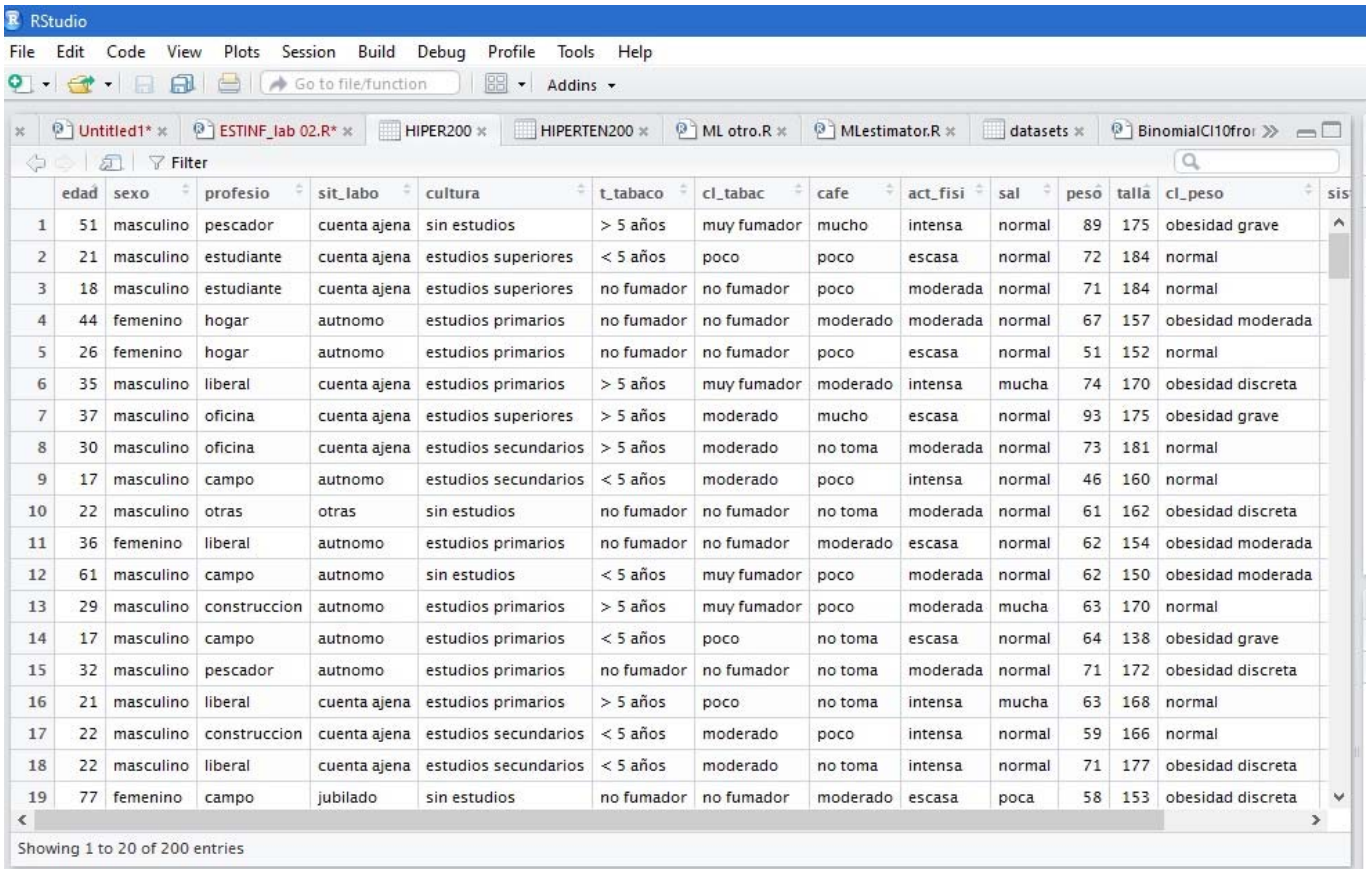
str(HIPERTENT200)

Si bien los valores recogidos son los adecuados, no lo es para el caso de variables cualitativas. Por ello procedemos a declarar como numéricas a todas las variables cuantitativas y a todas las cualitativas como factores indicando sus valores asignados y sus etiquetas. Además de indicar si son variables con orden entre las modalidades.

Son dos comandos aplicados de manera iterativa y se verán directamente en el script de la práctica.

Una vez ejecutados, si aplicamos de nuevo ***str(HIPER200)*** la información de las variables si es correcta.

Si miramos los datos en la primera pantalla. Solo veremos números en las variables cuantitativas y etiquetas de modalidades para las variables cualitativas.



	edad	sexo	profesio	sit_labo	cultura	t_tabaco	cl_tabac	cafe	act_fisi	sal	peso	talla	cl_peso	sis
1	51	masculino	pescador	cuenta ajena	sin estudios	> 5 años	muy fumador	mucho	intensa	normal	89	175	obesidad grave	
2	21	masculino	estudiante	cuenta ajena	estudios superiores	< 5 años	poco	poco	escasa	normal	72	184	normal	
3	18	masculino	estudiante	cuenta ajena	estudios superiores	no fumador	no fumador	poco	moderada	normal	71	184	normal	
4	44	femenino	hogar	autnomo	estudios primarios	no fumador	no fumador	moderado	moderada	normal	67	157	obesidad moderada	
5	26	femenino	hogar	autnomo	estudios primarios	no fumador	no fumador	poco	escasa	normal	51	152	normal	
6	35	masculino	liberal	cuenta ajena	estudios primarios	> 5 años	muy fumador	moderado	intensa	mucha	74	170	obesidad discreta	
7	37	masculino	oficina	cuenta ajena	estudios superiores	> 5 años	moderado	mucho	escasa	normal	93	175	obesidad grave	
8	30	masculino	oficina	cuenta ajena	estudios secundarios	> 5 años	moderado	no toma	moderada	normal	73	181	normal	
9	17	masculino	campo	autnomo	estudios secundarios	< 5 años	moderado	poco	intensa	normal	46	160	normal	
10	22	masculino	otras	otras	sin estudios	no fumador	no fumador	no toma	moderada	normal	61	162	obesidad discreta	
11	36	femenino	liberal	autnomo	estudios primarios	no fumador	no fumador	moderado	escasa	normal	62	154	obesidad moderada	
12	61	masculino	campo	autnomo	sin estudios	< 5 años	muy fumador	poco	moderada	normal	62	150	obesidad moderada	
13	29	masculino	construccion	autnomo	estudios primarios	> 5 años	muy fumador	poco	moderada	mucha	63	170	normal	
14	17	masculino	campo	autnomo	estudios primarios	< 5 años	poco	no toma	escasa	normal	64	138	obesidad grave	
15	32	masculino	pescador	autnomo	estudios primarios	no fumador	no fumador	no toma	moderada	normal	71	172	obesidad discreta	
16	21	masculino	liberal	cuenta ajena	estudios primarios	> 5 años	poco	no toma	intensa	mucha	63	168	normal	
17	22	masculino	construccion	cuenta ajena	estudios secundarios	< 5 años	moderado	poco	intensa	normal	59	166	normal	
18	22	masculino	liberal	cuenta ajena	estudios secundarios	< 5 años	moderado	no toma	intensa	normal	71	177	obesidad discreta	
19	77	femenino	campo	jubilado	sin estudios	no fumador	no fumador	moderado	escasa	poca	58	153	obesidad discreta	

Una vez comprobado que todo esta correcto, vamos a guardar en disco dicho dataframe:

```
save(HIPER200, file="HIPER200.RData")
```

El fichero se ha guardado, pero no lo encontramos. Estará en el directorio que le hubiésemos indicado o el que tomo por defecto.
Para verlo podemos escribir

```
getwd()                          # muestra el directorio de trabajo actual
setwd("N:/RTRABAJO/INF")         # establece el nuevo directorio de trabajo
```

Si ejecutamos de nuevo el comando *save*, nos aparece el fichero **HIPER200.RData**
Podemos vaciar todo el contenido del workspace con el comando

```
remove(list=ls())
```

No tenemos nada en memoria, ver la ventana 2.
Vamos a cargar en R el fichero que guardamos anteriormente mediante

```
load("HIPER200.RData")
```

Si no funciona el comando, es que no ha encontrado el fichero en el directorio de trabajo, podemos subsanarlo indicando totalmente la dirección del fichero con

load("N:/RTRABAJO/INF/HIPER200.RData")

Observamos que nuestro fichero es correcto, pero queremos realizar algunos cambios menores en la estructura de los datos.

Cambiar el nombre de una variable

names(HIPER200)[names(HIPER200)=="cultura"]<-c("nivel de estudios") #por su nombre

names(HIPER200)[2]<-"genero" #por su posición en el fichero

Recodificar las etiquetas de los niveles de un factor

```
library(DescTools)        #nueva librería

HIPER200$cl_tabac1<-Recode(HIPER200$cl_tabac,
 "no fumador" = c("no fumador"),
 "fumador" = c("poco","moderado","muy fumador"))
```

Crea una variable con el número secuencial de cada caso

HIPER200$id<-seq(dim(HIPER200)[1])

Reordenamos las columnas en el date frame

HIPER200<-HIPER200[c(23,1,12,13,15:18,2:10,14,11,19:22)] #tener cuidado

Seleccionar subconjuntos de datos y de variables

mydata02<-subset(HIPER200,select=c(-genero1,-cl_tabac1)) #eliminamos las variable duplicadas

mydata03<-subset(HIPER200, genero=="femenino") #seleccionamos casos

mydata04<-subset(HIPER200, edad>=30 & genero=="masculino" & t_tabaco=="no fumador") #seleccionamos casos mediante valores de tres variables

Recodificar una variable numérica a factor

HIPER200$edad1<-cut(HIPER200$edad, breaks=c(15,30,60,90), labels=c("joven", "maduro", "jubilado"))

Todas estas mismas operaciones realizadas en esta práctica, **en teoría** se puede hacer desde el Rcommander, abriendo el fichero script de la práctica y realizar secuencialmente las instrucciones, pero usualmente suelen haber errores de ejecución.

EJERCICIOS

1- También se suministran los ficheros HIPERTEN200.xlsx, HIPERTEN200.csv HIPERTEN200.dta, HIPERTEN200.sas7bdat y HIPERTEN200.dat, que se corresponden con el mismo fichero utilizado en la práctica pera en diferentes formatos, para que el alumno los importe igualmente y realice las operaciones oportunas.
[Pista: mydata<-read.table("HIPERTEN200.dat",header=TRUE, sep="\t", na.strings= "-9")]

2- Crear variables adicionales a partir del fichero de datos "HIPER200.RData", por ejemplo: log(edad), IMC=peso/talla^2, sist0=TAsist0/100, TA_med0=(TAsist0+2*TAdias0)/3, TA_med1=(TAsist1+2*TAdias1)/3, etc..

3.- La librería ISLR contiene un conjunto de datos relacionado con variables sobre ciertos coches denominado "auto". Se pide realizar un proceso similar al realizado con HIPERTEN200.sav. Podríamos resumirlo en: localizar el fichero de datos en el ordenador o via web, importarlo a través de RSudio, ver la estructura del fichero, definir como numéricas las variables cuantitativas y como factores las restantes. Dar etiquetas a estos últimos. Cambiar las unidades americanas de las variables a la del sistema métrico. Reordenar las variables dentro del fichero si se considera apropiado. Guardar dicho fichero. Eliminarlo del workspace. Cargarlo de nuevo, etc.