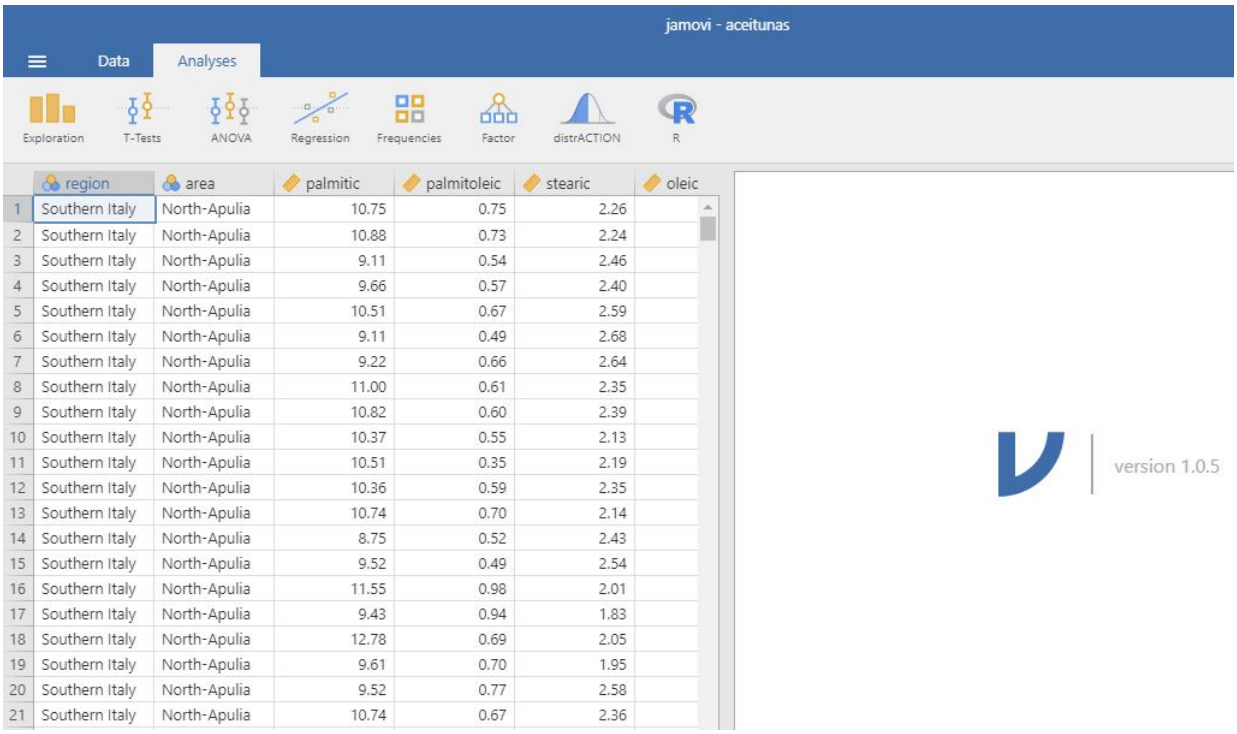
Práctica 1: Análisis exploratorio de datos

1. Ácidos grasos en aceitunas de distintas regiones italianas

El fichero `aceitunas.omv` (disponible en la página de Moodle de la asignatura) contiene datos sobre el porcentaje de ocho ácidos grasos en la fracción lipídica de aceitunas procedentes de nueve áreas de Italia correspondientes a tres grandes regiones: norte de Italia, Cerdeña y sur de Italia.

Al abrir el fichero de datos con `jamovi` veremos una ventana con la apariencia de una hoja de cálculo. A la derecha veremos un espacio en el que irán apareciendo posteriormente los resultados de los cálculos.



	region	area	palmitic	palmitoleic	stearic	oleic
1	Southern Italy	North-Apulia	10.75	0.75	2.26	
2	Southern Italy	North-Apulia	10.88	0.73	2.24	
3	Southern Italy	North-Apulia	9.11	0.54	2.46	
4	Southern Italy	North-Apulia	9.66	0.57	2.40	
5	Southern Italy	North-Apulia	10.51	0.67	2.59	
6	Southern Italy	North-Apulia	9.11	0.49	2.68	
7	Southern Italy	North-Apulia	9.22	0.66	2.64	
8	Southern Italy	North-Apulia	11.00	0.61	2.35	
9	Southern Italy	North-Apulia	10.82	0.60	2.39	
10	Southern Italy	North-Apulia	10.37	0.55	2.13	
11	Southern Italy	North-Apulia	10.51	0.35	2.19	
12	Southern Italy	North-Apulia	10.36	0.59	2.35	
13	Southern Italy	North-Apulia	10.74	0.70	2.14	
14	Southern Italy	North-Apulia	8.75	0.52	2.43	
15	Southern Italy	North-Apulia	9.52	0.49	2.54	
16	Southern Italy	North-Apulia	11.55	0.98	2.01	
17	Southern Italy	North-Apulia	9.43	0.94	1.83	
18	Southern Italy	North-Apulia	12.78	0.69	2.05	
19	Southern Italy	North-Apulia	9.61	0.70	1.95	
20	Southern Italy	North-Apulia	9.52	0.77	2.58	
21	Southern Italy	North-Apulia	10.74	0.67	2.36	

2. Descripción de una variable cualitativa

Consideramos la variable `region`, que corresponde a la región de procedencia de las aceitunas. Para describirla lo que procede es calcular una tabla de frecuencias y representar un

gráfico de barras.

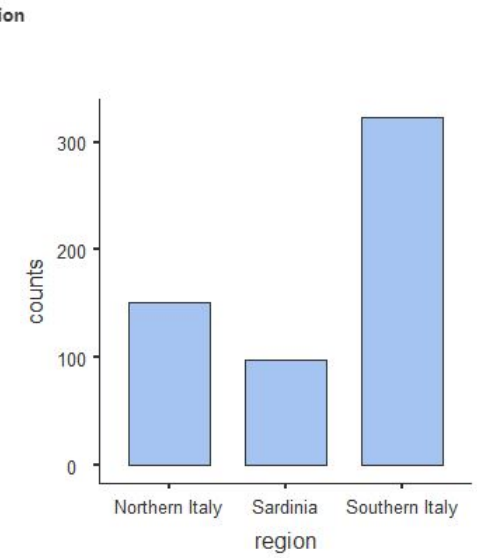
Para ello, hay que seleccionar la opción

`Analyses` ↪ `Exploration` ↪ `Descriptives`

A continuación seleccionamos la variable `region` y la pasamos (usando la flecha) al recuadro *variables*. En la parte inferior marcamos *Frequency tables* y, dentro del apartado *Plots*, *Bar plot*. Veremos a la derecha los resultados:

Frequencies of region

Levels	Counts	% of Total	Cumulative %
Northern Italy	151	26.4 %	26.4 %
Sardinia	98	17.1 %	43.5 %
Southern Italy	323	56.5 %	100.0 %



Si queremos saber el número de observaciones para cada región y cada área (lo que se llama habitualmente una tabla de contingencia), tenemos que ir a

`Analyses` ↪ `Frequencies` ↪ `Independent samples`

Podemos elegir la variable `region` para las columnas y `area` para las filas. ¿A qué caso corresponde el mayor número de observaciones? Debajo aparece un resultado relativo al contraste χ^2 que de momento podemos olvidar. En el apartado *Cells* podemos marcar las opciones que permiten calcular los porcentajes.

3. Descripción de una variable cuantitativa

De nuevo elegimos la opción:

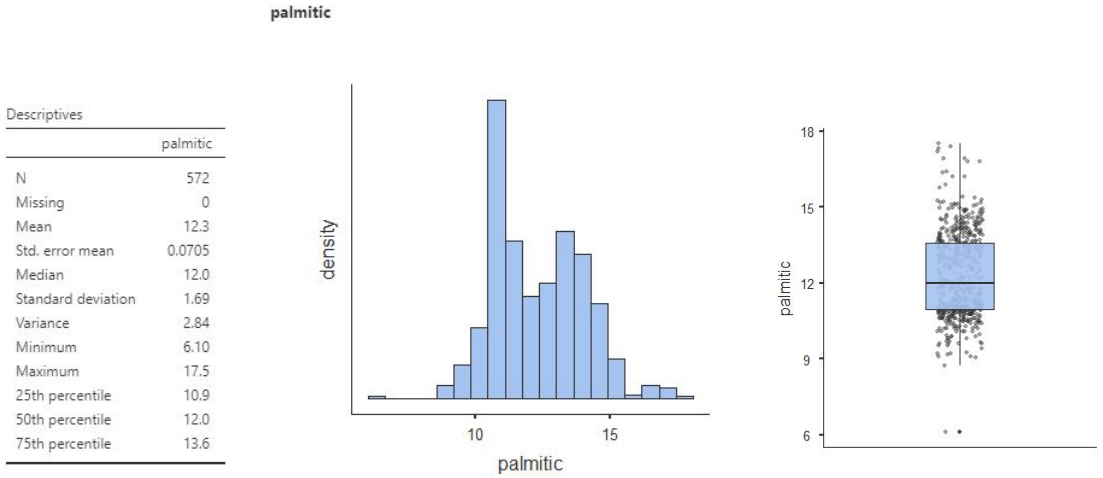
`Analyses` ↪ `Exploration` ↪ `Descriptives`

A continuación seleccionamos alguna de las variables cuantitativas, por ejemplo, `palmitic`. En la parte inferior marcamos:

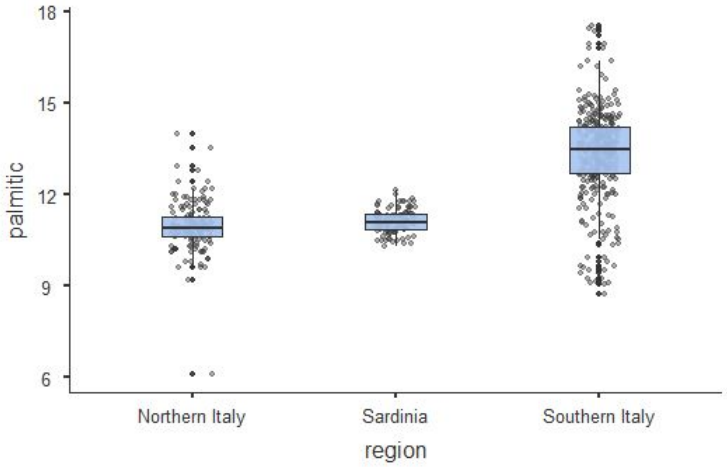
(a) En *Statistics*: las medidas numéricas que queramos calcular además de las que ya están marcadas por defecto. Por ejemplo: *quartiles, std. deviation, variance* y *s.e. mean.* (¿Entiendes bien todas ellas?)

(b) En *Plots*: los gráficos que queramos representar. Por ejemplo: *histogram* y *box plot*. Puedes añadir los puntos que dan lugar a los gráficos eligiendo *Data*.

¿Se observa algo relevante en los graficos o en las medidas numéricas de esta variable?



Claramente se observa que la distribucion es bimodal. Estamos considerando todas la observaciones a la vez, pero es conveniente separar el estudio por zonas. Para ello, tenemos que añadir al recuadro *Split by* la variable `region`. Vemos que ahora todos los resultados se dan por separado para cada una de las tres regiones. Por ejemplo, los diagramas de cajas resultan ser estos:



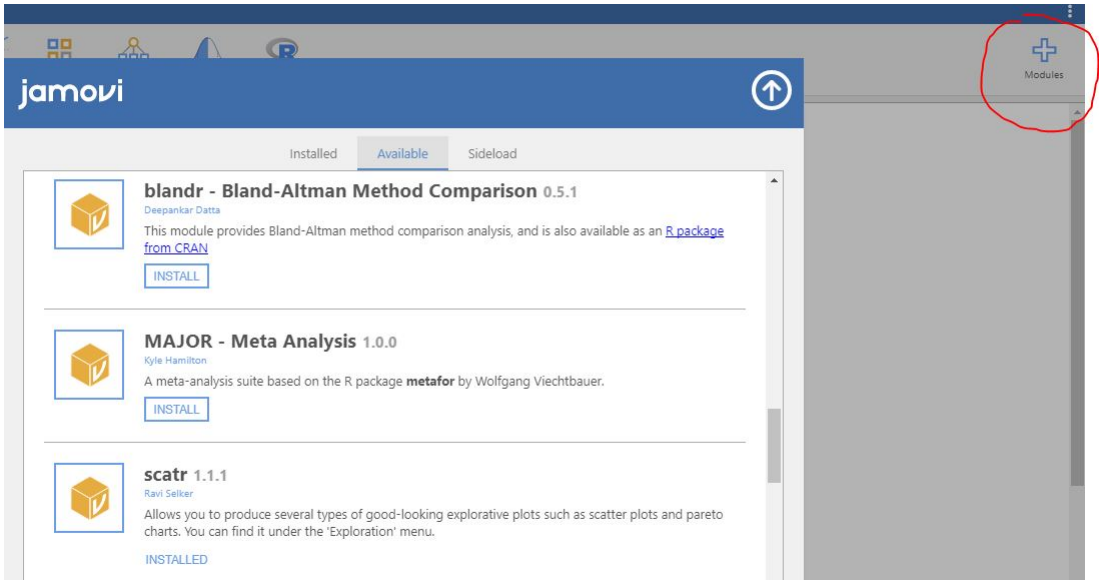
Una transformación habitual para una variable cualitativa consiste en recodificar los nombres de los valores de la variable. En el fichero que estamos manejando los nombres están en inglés pero tal vez preferiríamos que estuvieran en castellano. Para renombrar los niveles de

la variable `region` hacemos doble clic en el nombre de la variable y dentro del recuadro *levels* nos situamos en los niveles que queremos modificar y hacemos los cambios. Cuando hayamos terminado salimos.

Con un procedimiento similar se pueden cambiar también el nombre y el tipo de cada variable.

4. Relaciones entre variables: diagrama de dispersión y correlación

Existe una lista creciente de módulos que permite incrementar la funcionalidad del programa. Para representar un diagrama de dispersión necesitamos cargar previamente el módulo `scatr` pulsando la cruz de la parte superior derecha de la pantalla:



Para estudiar gráficamente el grado de asociación existente entre dos variables elegimos la opción:

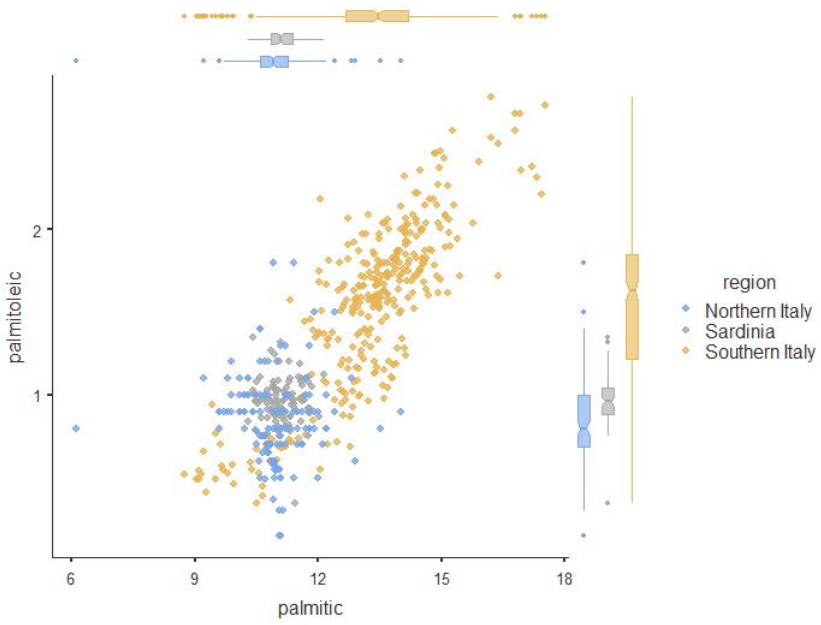
`Analyses` ↪ `Exploration` ↪ `Scatterplot`

Como variable Y debemos elegir la variable *dependiente*, la que queremos explicar. Por ejemplo, `palmitoleic`. Como variable X elegimos la variable *explicativa*. Por ejemplo, `palmitic`.

Si queremos que en el gráfico se utilice un color diferente según una variable cualitativa, la seleccionamos en el apartado *Group*. En nuestro ejemplo, podemos elegir `region`.

Finalmente es posible también marcar la opción *Boxplots* del apartado *Marginals*, con lo que se añade un diagrama de cajas de las dos variables involucradas.

El resultado es el siguiente:



Una vez obtenido el gráfico, ¿puede decirse que la relación entre las dos variables es lineal? ¿Es la asociación positiva o negativa? ¿Difiere mucho la situación para cada una de las regiones?

Para calcular la correlación entre las dos variables, elegimos

`Analyses` ↪ `Regression` ↪ `Correlation Matrix`

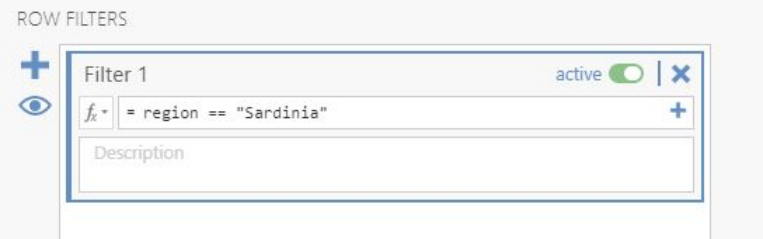
A continuación pasamos al cuadro todas las variables entre las que queremos calcular correlaciones. Puedes elegir las 8 variables numéricas del fichero original. ¿Entre qué par de variables se observa el mayor grado de relación lineal?

5. Seleccionar un subconjunto de los datos

Una operación que en la práctica suele ser necesaria frecuentemente es filtrar los datos, usar en el análisis solo un subconjunto de ellos que cumpla cierta condición. En nuestro caso, podríamos estar interesados en usar exclusivamente los datos de Cerdeña. Para filtrar los datos, elegimos la opción:

`Data` ↪ `Filters`

En el recuadro que se abre tenemos que escribir la condición lógica que queremos que cumplan los datos que queremos analizar. Para elegir los de Cerdeña, escribimos:



Obsérvese el uso del doble signo de igual y el de las comillas. En el conjunto de datos aparece una nueva columna que nos indica qué observaciones cumplen la condición y cuáles no.

Un segundo ejemplo: para seleccionar las observaciones que **no** sean de Cerdeña **y para las que además** la variable `palmitic` sea mayor que 10 escribimos:

Obsérvese el uso de `!=` para indicar que la variable *no es igual a...* y el de la palabra *and* para añadir una segunda condición.

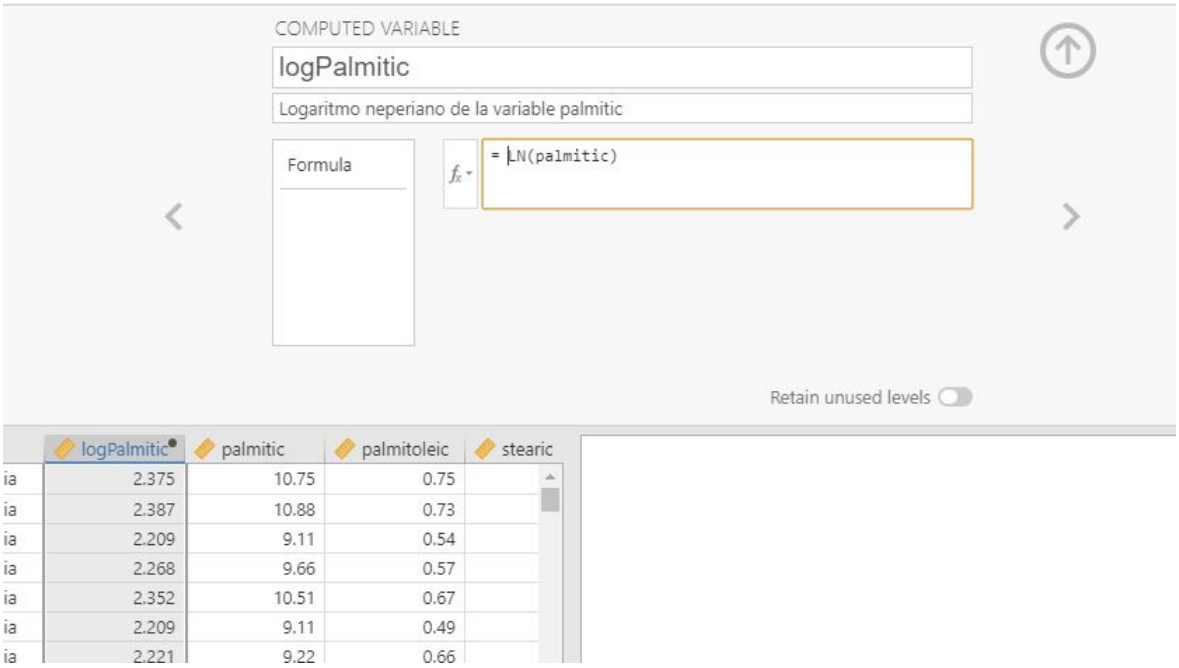
Cuando un filtro se activa, todos los cálculos que hayamos hecho se rehacen automáticamente considerando únicamente el subconjunto de variables seleccionadas.

6. Transformaciones de variables

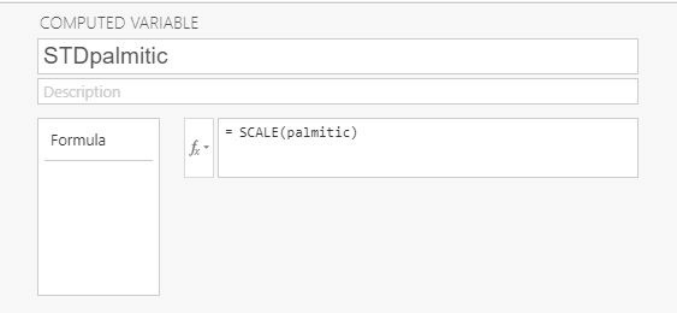
Para crear una nueva variable como resultado de aplicar una transformación a alguna de las ya existentes, se procede de la forma siguiente:

- Vamos a `Data` $\hookrightarrow$ `Compute`.
- En *Computed variable* se escribe el nombre de la nueva variable que vamos a crear. Debajo se puede redactar una breve descripción que nos recuerde en qué consiste esa variable.
- Pulsando el botón f_x se abren los listados de funciones matemáticas y de variables ya existentes. Elegimos la función y la variable que queremos transformar (por este orden) haciendo doble clic en la opción deseada.
- Inmediatamente vemos que en el fichero aparece la nueva variable que hemos creado. A partir de ahora la podremos usar en cualquier tipo de cálculo.

Por ejemplo, si queremos crear una variable que sea el logaritmo neperiano de la variable `palmitic`, las opciones a elegir son las siguientes:



Si lo que queremos es estandarizar la variable, en lugar de la función `LN` se usa `SCALE`:



7. Al salir de jamovi

Cada análisis individual (numérico o gráfico) se puede salvar usando el botón derecho del ratón. Se puede copiar al portapapeles para luego pegarlo en otro documento, o guardar en un fichero individual en diferentes formatos.

Es recomendable guardar los cambios y análisis que hemos llevado a cabo **en un nuevo fichero** de manera que siempre tengamos disponibles los datos originales. Para ello, en el menú de la parte superior izquierda (las tres rayas horizontales) elegimos *Save as...* y elegimos el nombre y la carpeta que queramos.

Ejercicio

Primera parte: Contestar a las preguntas siguientes (cada una de ellas se responde con un número, no hace falta añadir ninguna explicación):

(1) Calcula el rango intercuartílico de la variable `linolenic` considerando solo las observaciones procedentes del área de Calabria.
(2) Calcula el coeficiente de correlación entre `linolenic` y `arachidic` teniendo en cuenta todas las observaciones menos las procedentes del área de Calabria.
(3) Calcula el máximo valor de la variable `linolenic` estandarizada.
(4) ¿Cuántos datos mayores que la mediana aparecen marcados como atípicos en el diagrama de cajas de la variable `linolenic` estandarizada?
(5) Calcula el coeficiente de correlación entre los logaritmos (en base 10) de las variables `linolenic` y `arachidic` para la muestra de observaciones que procede del sur de Italia.

Segunda parte: Llevar a cabo un estudio descriptivo de los datos contenidos en el fichero `metabolismo.omv`. La tasa metabólica, es decir, la tasa a la que el cuerpo consume energía, es relevante en los estudios sobre ganancia de peso y dietética. El fichero `metabolismo.omv` contiene datos sobre masa corporal magra (en kg) y tasa metabólica (en cal por cada 24 horas) de 12 mujeres y 7 hombres que participaron en un estudio de dietética. El fichero está disponible en la página de Moodle de la asignatura.

Se trata de decidir y aplicar las medidas numéricas y los gráficos que sean más eficaces para describir los datos, de acuerdo con los objetivos. El informe debe contener los siguientes epígrafes:

(1) **Características de los datos:** número de observaciones y de variables, tipos de variables, etc.
(2) **Objetivos:** preguntas o cuestiones que puede ser interesante responder con estos datos.
(3) **Medidas numéricas y gráficos:** los que sean necesarios para responder a las preguntas del apartado anterior.
(4) **Conclusiones:** qué hemos aprendido en el apartado anterior de acuerdo a los objetivos que teníamos.

La extensión máxima para contestar a la segunda parte es de tres páginas.