

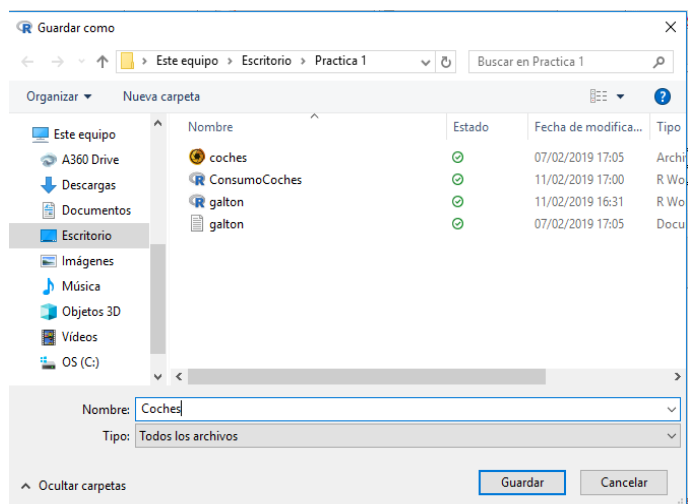
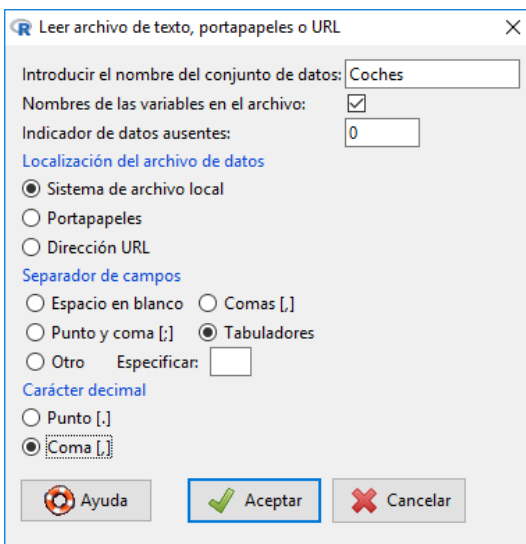
Práctica 3 R-Commander

NOMBRE: Jaime Osés Azcona

1. Lee el archivo desde R-Commander, sabiendo que la primera fila contiene los nombres de las variables, las columnas están separadas por tabuladores, los datos faltantes están codificados con "0" y el separador decimal es la coma. Guárdalo con nombre "coches.rda".

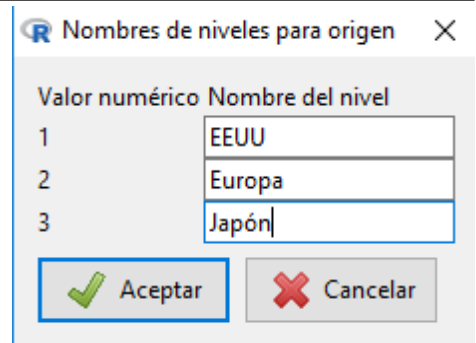
Primero importaremos nuestra base de datos desde un archivo (.dat). Para ello, iremos a "**Datos-Importar datos-desde archivo de texto...**" y especificaremos las características que tenga el archivo que queremos importar. Seleccionamos el archivo y se creará nuestro conjunto de datos.

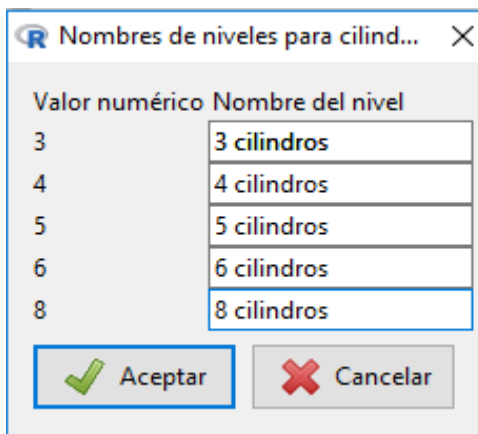
Para guardarlo en un archivo rda iremos a "**Datos-Conjunto de datos activo-Guardar el conjunto de datos activo...**" y lo guardaremos en el directorio de trabajo de esta práctica que hemos creado anteriormente.



2. Prepara los factores que hagan falta, teniendo en cuenta que la variable número de cilindros se tratará como una variable ordinal. Denomina las nuevas variables como las anteriores con el prefijo f.

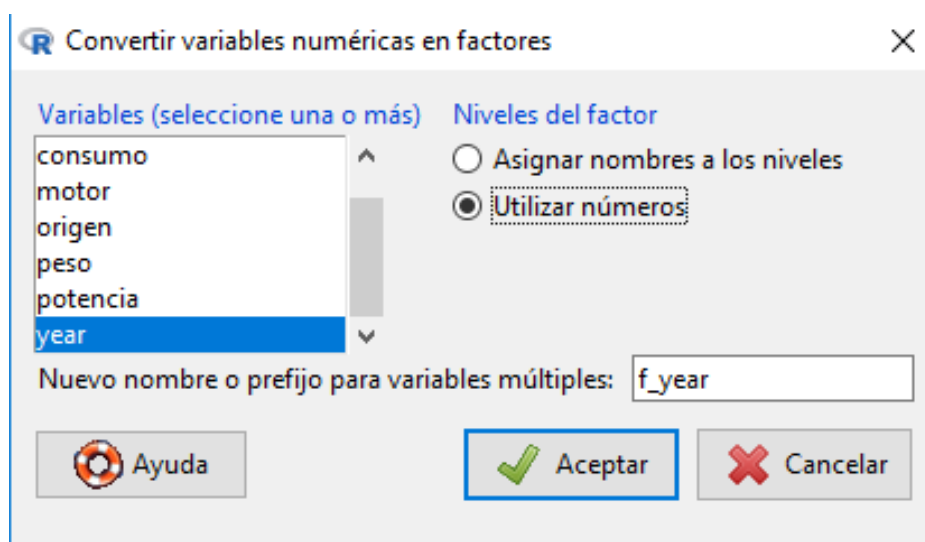
En el apartado origen, necesitaremos realizar un cambio de variable numérica a factor, ya que hay una serie de números asociados cada uno a diferentes países de origen. Para ello iremos a "**Datos-Modificar variables...-Convertir variable numérica en factor...**", seleccionaremos el campo al que queremos realizar la modificación y por último definiremos los distintos niveles.





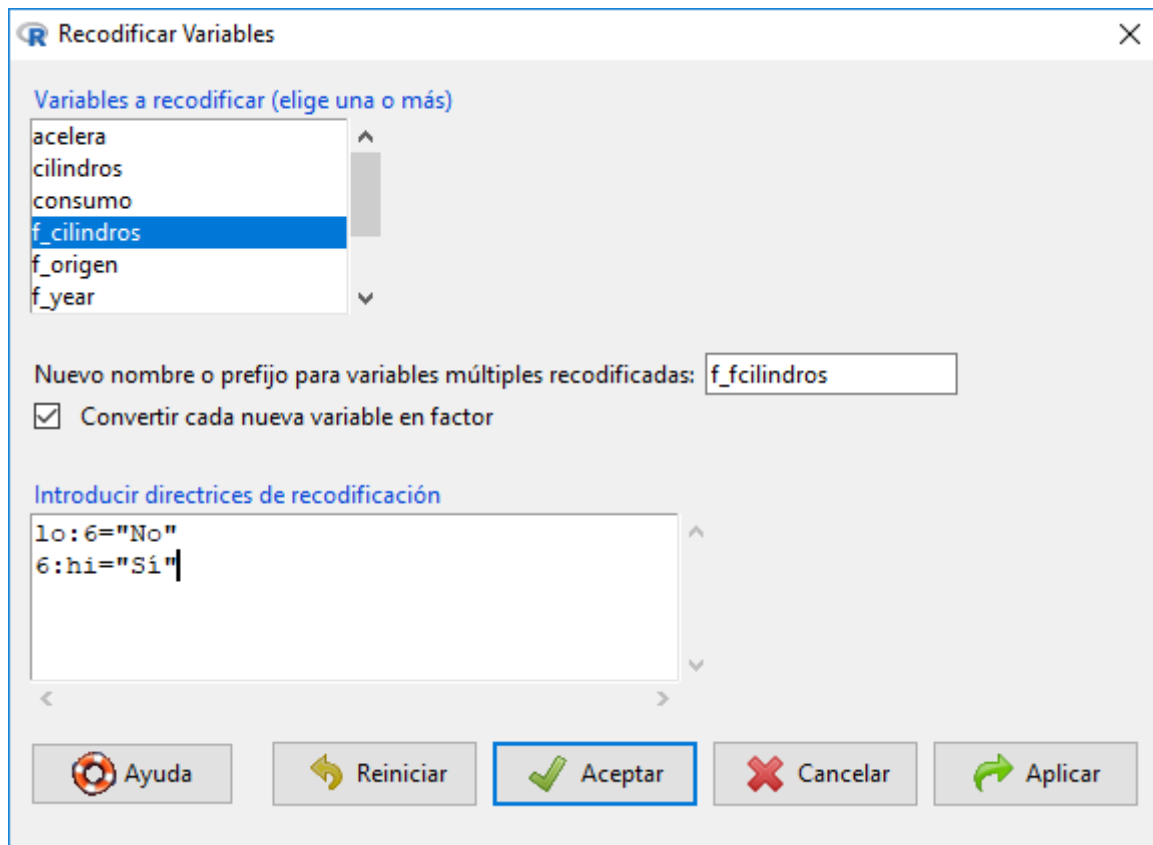
En el apartado cilindros, necesitaremos realizar un cambio de variable numérica a factor, ya que hay una serie de números asociados cada uno a diferentes números de cilindros. Para ello iremos a **“Datos-Modificar variables...-Convertir variable numérica en factor...”**, seleccionaremos el campo al que queremos realizar la modificación y por último definiremos los distintos niveles.

En el apartado year, necesitaremos realizar un cambio de variable numérica a factor, ya que, aunque sean todos los datos números, no queremos hacer ninguna operación estadística, sino diferencial los diferentes años. Para ello iremos a **“Datos-Modificar variables...-Convertir variable numérica en factor...”**, seleccionaremos el campo al que queremos realizar la modificación y por último definiremos los distintos niveles, esta vez con números en vez de nombres.



3. Crea un nuevo factor que recoja la información de si el coche tiene o no 6 o más cilindros, denomina a la nueva variable ffcilindros.

Para ello tendremos que ir a “**Datos-Modificar variables...-Recodificar variables...**” y tendremos que definir los limites de la recodificación y qué queremos que muestre cada uno de ellos. Tendremos que dar al factor un nombre diferente a las columnas anteriores para que no nos pise ninguno de ellos, ya que estamos realizando un calculo aparte.



4. Crea una variable que recoja el peso de los modelos en toneladas con dos decimales, denomínala PesoTm, y otra que recoja la potencia en Watts (1 C.V=735 Watts), denomínala PotenciaWatts.

Para calcular estas nuevas variables, operaremos con las diferentes columnas que ya tenemos. Para ello tendremos que ir a “**Datos-Modificar variables...-Calcular una nueva variable...**” y tendremos que definir las operaciones que queremos realizar. Tendremos que dar al factor un nombre diferente a las columnas anteriores para que no nos pise ninguno de ellos, ya que estamos realizando un calculo a partir de dichas columnas, las cuales necesitamos.

- **Peso a Toneladas**






Calcular una nueva variable

Variables actuales (doble clic para enviar a la expresión)

- f_year [factor]
- motor
- origen
- peso
- potencia
- year

Nombre de la nueva variable:

Expresión a calcular:

 Ayuda  Reiniciar  Aceptar  Cancelar  Aplicar

- **Potencia a Watios**






Calcular una nueva variable

Variables actuales (doble clic para enviar a la expresión)

- f_year [factor]
- motor
- origen
- peso
- PesoTm
- potencia

Nombre de la nueva variable:

Expresión a calcular:

 Ayuda  Reiniciar  Aceptar  Cancelar  Aplicar

5. Crea otra variable denomina fpeso que recoja 4 niveles de peso: ligeros hasta 0.8 toneladas; normales de 0.8 a 1 tonelada, pesados de 1 a 1.5 toneladas y muy pesados m´as de 1.5 toneladas.

Para ello tendremos que ir a “**Datos-Modificar variables...-Recodificar variables...**” y tendremos que definir los limites de la recodificacion que nos dividan los pesos en diferentes niveles. Tendremos que dar al factor un nombre diferente a las columnas anteriores para que no nos pise ninguno de ellos, ya que estamos realizando un calculo aparte.

Recodificar Variables

Variables a recodificar (elige una o más)

- origen
- peso
- PesoTm**
- potencia
- PotenciaWatts
- year

Nuevo nombre o prefijo para variables múltiples recodificadas:

☒ Convertir cada nueva variable en factor

Introducir directrices de recodificación

```
1:0.8="Ligeros"
0.8:1="Normales"
1:1.5="Pesados"
1.5:hi="Muy Pesdos"
```

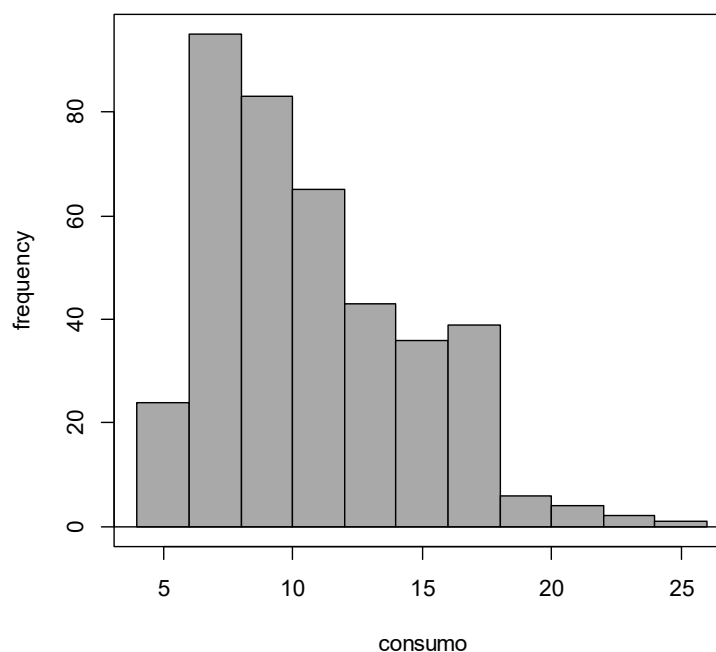
Ayuda Reiniciar **Aceptar** Cancelar Aplicar

6. Haz un estudio de la variable consumo, mediante un resumen de sus estadísticos y gráficos.

Para realizar un resumen de una variable tendremos que ir a “**Estadísticos–Resúmenes–Resúmenes numéricos...**” y deberemos seleccionar la variable de la cual se quiere realizar dicho resumen. También se deberá seleccionar desde la ventana estadísticos las operaciones que se quieren realizar: media, error típico de la media, desviación típica, cuartiles...

```
mean      sd  se(mean) IQR      cv  skewness  kurtosis  0%  25%  50%  75%  100%  n  NA
11.22864  3.946172  0.1978037   5  0.351438  0.7595016  0.07773354  5   8  10  13  26  398  8
```

Podemos observar que la media del consumo de los coches es de 11,22 L por cada 100 km. Esta media es así ya que hay una mayor cantidad de coches con un consumo pequeño, cosa que podemos deducir ya que los valores entre los primeros cuartiles cambian mas lento en intervalos de la misma amplitud, lo que evidencia que este tramo tiene una gran cantidad de valores. Esto hace que la media se desplace hacia la izquierda. Esto también se puede observar en el siguiente histograma:



7. Crear una nueva variable factor denominada fconsumo que recoja la información de la variable consumo en tres tramos, donde cada tramo tenga el mismo número de automóviles.

Para dividir una variable en partes iguales tenemos que ir a **“Datos-Modificar variables...-Segmentar variable numérica...”** y deberemos seleccionar las opciones de segmentación que queramos. En este caso dividiremos en 3 partes con el mismo número de automóviles, lo que haremos al seleccionar la opción segmentos de igual cantidad.

En nuestro caso los segmentos no tendrán el mismo número de automóviles ya que hay una gran cantidad de ellos con los mismos datos, lo que a la hora de agrupar hace que todos ellos vayan al mismo grupo, dejando los otros grupos con una menor cantidad.

Segmentar una variable numérica

Variable a segmentar (elegir una):
acelera
cilindros
consumo
motor
origen
peso

Nombre de la nueva variable:
f_consumo

Número de clases: 3

Nombres de niveles:
☒ Especificar nombres
☐ Números
☐ Rangos

Método de segmentación:
☐ Segmentos equidistantes
☒ Segmentos de igual cantidad
☐ Segmentos naturales (mediante agrupación por K-medias)

Ayuda Reiniciar Aceptar Cancelar

Nombres de los segmentos

Clase	Nombre
1	Bajo
2	Medio
3	Alto

Aceptar Cancelar

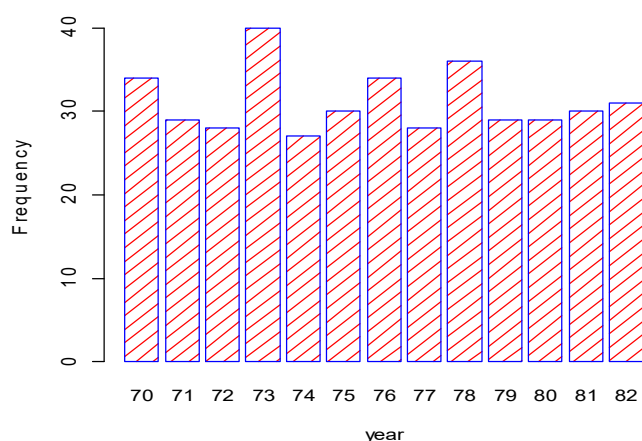
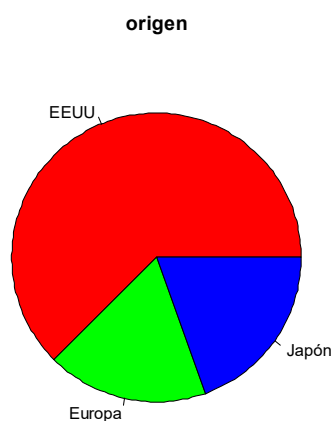
8. Elegir dos variables cualitativas y analízalas. Introduce algún cambio en los gráficos.

Como podemos observar en el gráfico de sectores, la variable `f_origen` tiene 3 grupos diferenciados: los coches que provienen de EEUU, los de Europa y los de Japón. La mayor parte de los coches (aproximadamente un 62%) provienen de EEUU y los restantes, los dos en cantidades parecidas (aproximadamente un 18% cada uno), provienen de Europa y Japón.

Como podemos observar en la grafica de barras, la variable `f_year` tiene 13 grupos diferenciados que se corresponden con los años de modelo de los coches. Casi todos los años de modelo tienen un número de coches muy parecidos que varía en un intervalo aproximadamente entre 25 y 30. Solo 4 años de modelo destacan entre los demás superando esa frecuencia de 30. Más en concreto, el año de modelo que mas coches tiene es el 73, con una cantidad de 40 coches.

Para realizar cambios en los gráficos, deberemos modificar las fórmulas desde la ventana R-Script.

- El cambio que hemos realizado en la gráfica de sectores es ponerle una paleta de colores ya existente. Para ello deberemos poner en la fórmula de R-Script **`col=rainbow(3)`**, indicando en el paréntesis el número de sectores que hay.
- El cambio que hemos realizado en la gráfica de barras es ponerle a las barras un borde del color que queramos y rellenar dichas barras en vez de con colores, con una especie de rayado con la densidad que queramos. Para ello deberemos poner en la fórmula de R-Script **`border="blue"`**, **`density=10`** y para indicar el color del rayado **`col="red"`**.

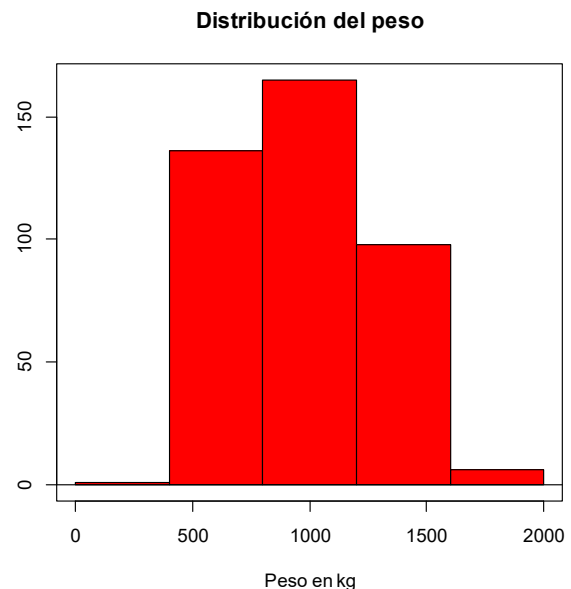
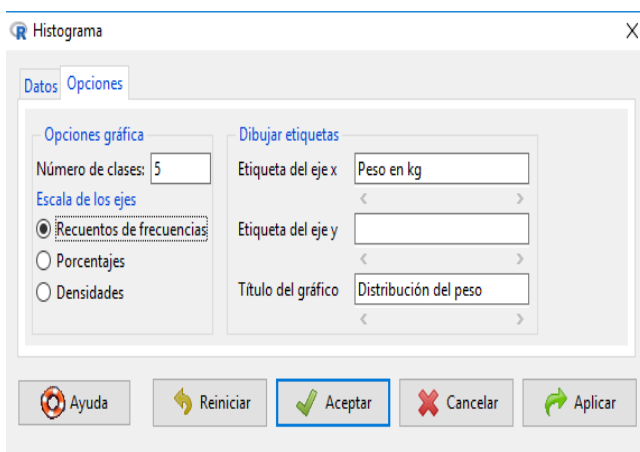


9. Construye el histograma de la variable peso en kg en color rojo, con 5 intervalos, un título principal (Distribución del peso), otro en el eje x (Peso en kg) y sin título en el eje y.

Para construir un histograma tendremos que ir a **Gráficas-Histograma...** y deberemos seleccionar la variable de la que se quiere realizar. Para poner las características que se especifican, deberemos ir al apartado opciones y hacer los cambios que queramos.

Hay algunos cambios que no se pueden hacer en el apartado opciones y deberemos realizarlos directamente en la fórmula de R-Script. Uno de ellos es el color, que deberemos modificarlo poniendo **col="(color que queramos)"**, en este caso **col="red"**.

Otra de ellas es los intervalos en que lo queramos dividir. En este caso desde el menú opciones lo que interpreta es una recomendación, pero no hace el número de intervalos que deseamos, lo que deberemos solucionar poniendo en la fórmula **breaks=c(0,400,800,1200,1600,2000)**.



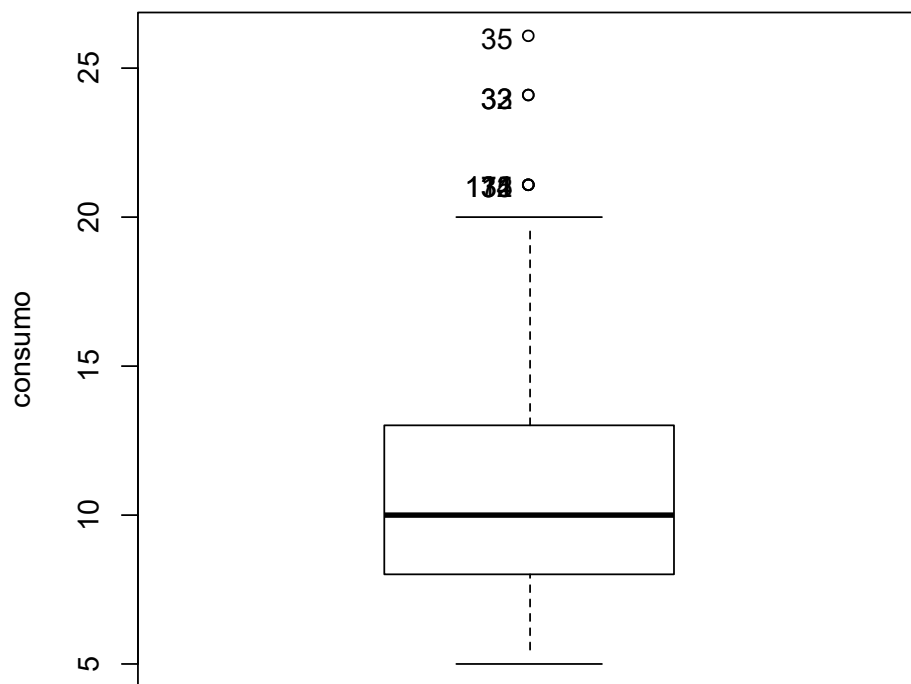
10. Calcula los cuartiles de la variable consumo e interprétalos. Construye el diagrama de cajas de esta variable.

Para calcular los cuartiles de una variable tendremos que ir a **Estadísticos-Resúmenes-Resúmenes numéricos...** y en el apartado Estadísticos seleccionar la opción cuartiles e indicar los que queramos realizar.

0%	25%	50%	75%	100%	n	NA
5	8	10	13	26	398	8

Esto nos indica los valores de la variable en ciertos intervalos, en este caso el 0, 20, 50, 75 y 100 % de la variable. Podemos deducir entonces que el menor consumo es 5 ya que este es el valor del P_0 . También podemos deducir que hay menos coches con un consumo mayor de 13, ya que de P_{75} a P_{100} hay un cambio mas grande que en los demás y en un intervalo de la misma amplitud.

Para realizar el diagrama de cajas tendremos que ir a **Gráficas-Diagrama de caja...** y seleccionar la variable de la cual queremos realizarlo.



11. Analiza las variables peso total y aceleración según la variable factor creada en el apartado 3. Obtén las medidas más relevantes y los diagramas de cajas asociados a cada variable según el grupo.

Como podemos ver en la gráfica de cajas correspondientes, la variable peso es dependiente del factor consumo creado en el apartado 3. Esto lo podemos notar ya que las cajas están en escalera. Esto significa que la variable peso tiene valores diferentes según en grupo de cilindros en el que se encuentren.

En el caso de la variable acelera, las cajas son muy parecidas y están casi alineadas. Esto significa que la variable acelera tiene valores muy parecidos en todos los grupos de cilindros, por lo que podemos deducir que es una variable independiente del consumo.

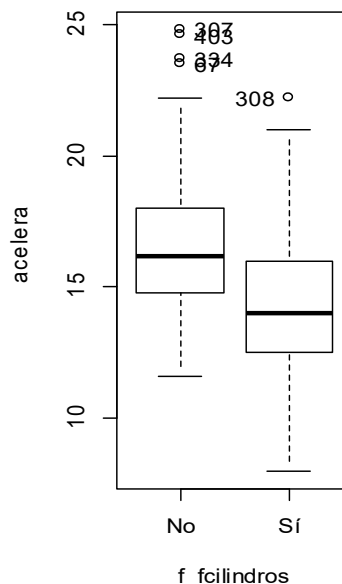
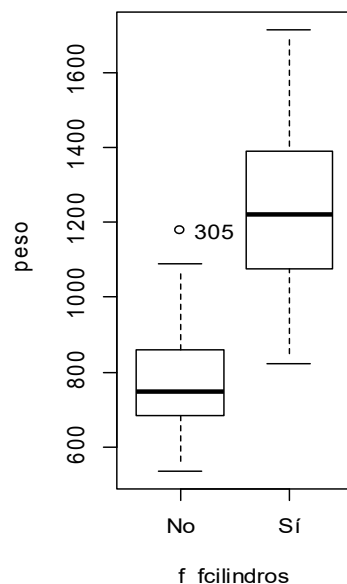
Haciendo esto con números y más exacto, podemos afirmar lo mismo:

- En la variable peso, la media de sus valores por cada grupo de cilindros es muy diferente en cada grupo. Por tanto, podemos afirmar que es una variable dependiente.

	mean	sd	se(mean)	IQR	cv	skewness	kurtosis	0%	25%	50%	75%	100%	peso:n
No	774.7757	120.3053	8.223899	177.75	0.1552775	0.6178855	-0.03934051	537	683	749.5	860.75	1176	214
Sí	1234.0105	199.6736	14.447873	313.00	0.1618087	0.1549236	-0.80759101	824	1075	1221.0	1388.00	1713	191

- En la variable acelera, la media de sus valores es muy parecida en todos los grupos de cilindros. Por tanto, podemos afirmar que es una variable independiente.

	mean	sd	se(mean)	IQR	cv	skewness	kurtosis	0%	25%	50%	75%	100%	acelera:n
No	16.58178	2.407947	0.1646039	3.2	0.1452165	0.8345478	0.7045842	11.6	14.8	16.2	18	24.8	214
Sí	14.31414	2.733738	0.1978064	3.5	0.1909817	0.2180893	-0.1597546	8.0	12.5	14.0	16	22.2	191



12. Haz un estudio de la variable peso factorizado por el país de origen, mediante un resumen de sus estadísticos y gráficos.

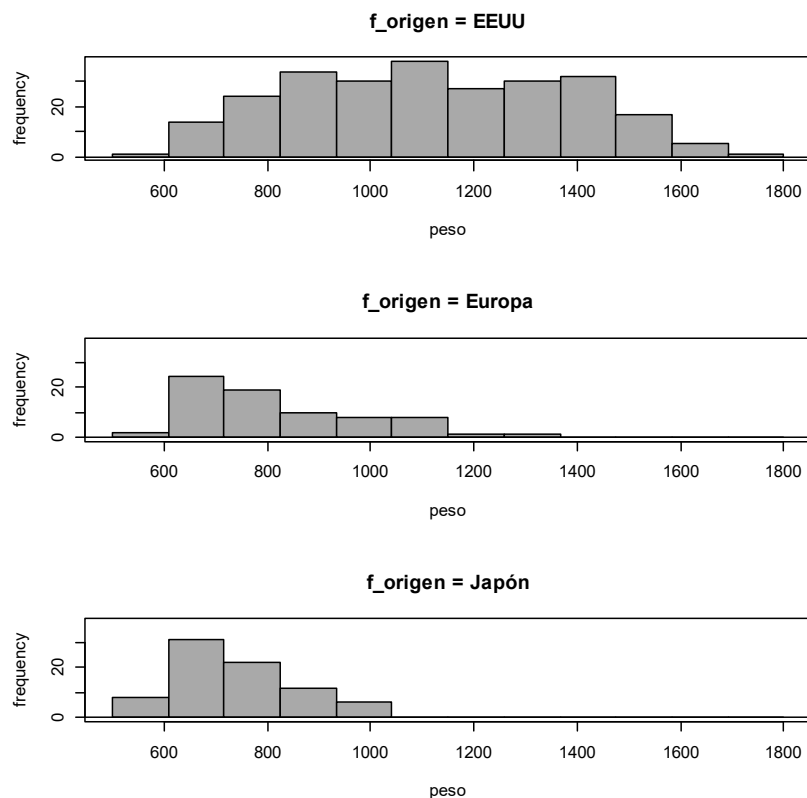
Para realizar un resumen de una variable tendremos que ir a “**Estadísticos–Resúmenes–Resúmenes numéricos...**” y deberemos seleccionar la variable de la cual se quiere realizar dicho resumen. También se deberá seleccionar desde la ventana estadísticos las operaciones que se quieren realizar: media, error típico de la media, desviación típica, cuantiles...

	mean	sd	se (mean)	IQR	cv	skewness	kurtosis
EEUU	1122.1146	262.8679	16.52636	445.0	0.2342612	0.05558777	-0.9528407
Europa	810.1233	163.6234	19.15067	245.0	0.2019734	0.77310789	-0.3265656
Japón	740.0506	106.7643	12.01192	142.5	0.1442662	0.49718432	-0.3790739

	0%	25%	50%	75%	100%	peso:n
EEUU	600	906	1126	1351.0	1713	253
Europa	608	688	748	933.0	1273	73
Japón	537	661	718	803.5	976	79

Podemos observar como influye el factor origen en el peso. La media en los diferentes países es diferente en cada uno de ellos. En este caso, los coches de EEUU tienen un peso mayor que los de Europa y Japón, siendo estos últimos los de menor peso. Por tanto, podemos afirmar que la variable peso es dependiente del origen.

En EEUU, hay mas coches con pesos medios. Esto se evidencia ya que los percentiles intermedios cambian en menor cantidad respecto a los exteriores, en la misma longitud de intervalos. Esto hace que la media sea más o menos la mitad de la muestra. En Europa y Japón, los coches tienen pesos bajos. Esto se evidencia ya que los primeros percentiles cambian en menor cantidad respecto a los exteriores, en la misma longitud de intervalos. Esto hace que la mediana se desplace hacia la izquierda de la muestra y sea menor que la de EEUU. También podemos observarlo en el siguiente histograma:



13. ¿Cuántos coches de EEUU hay en la muestra? Rellena la siguiente la tabla referida a estos coches.

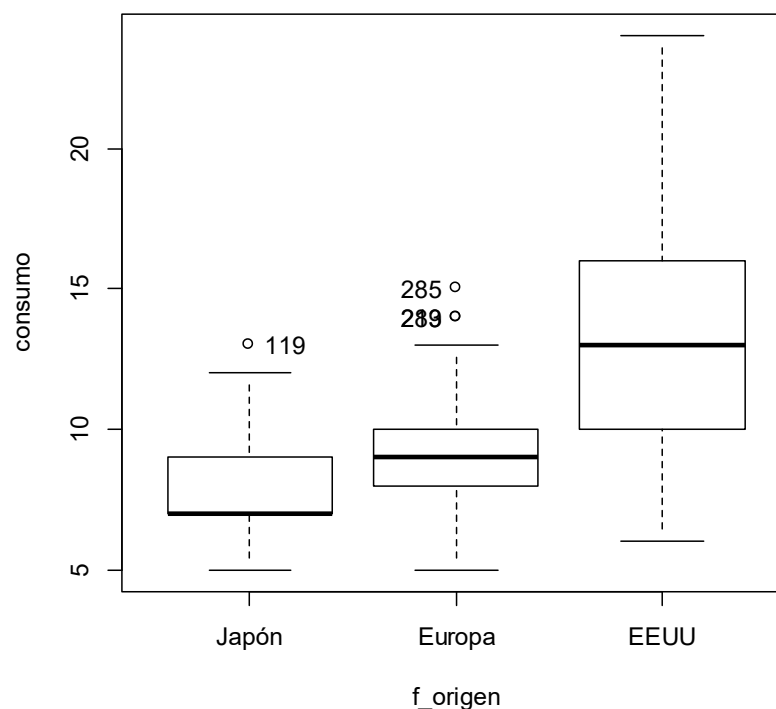
Para realizar los cálculos de esta tabla iremos a “**Estadísticos-Resúmenes-Resúmenes numéricos...**” y activaremos la opción por grupos, seleccionando el factor origen creado anteriormente, y elegiremos la variable de la cual queremos realizar el resumen. En este caso, haremos 3 resúmenes: el de consumo, el de potencia y el de peso.

De cualquiera de estos 3 resúmenes podemos obtener que hay 253 coches de EEUU.

	Media	Desviación Típica	Coefficiente Variación	Mínimo	Máximo	Mediana
Consumo	12.842742	3.793702	0.2953966	6	24	13
Potencia	119.60643	39.79916	0.3327511	52	230	105
Peso	1122.1146	262.8679	0.2342612	600	1713	1126

14. Realiza un diagrama de cajas de la variable consumo según el país de origen. ¿Existen valores anómalos en alguno de los países de origen? Analiza las similitudes y diferencias que puedas observar sobre variable consumo para cada país de origen.

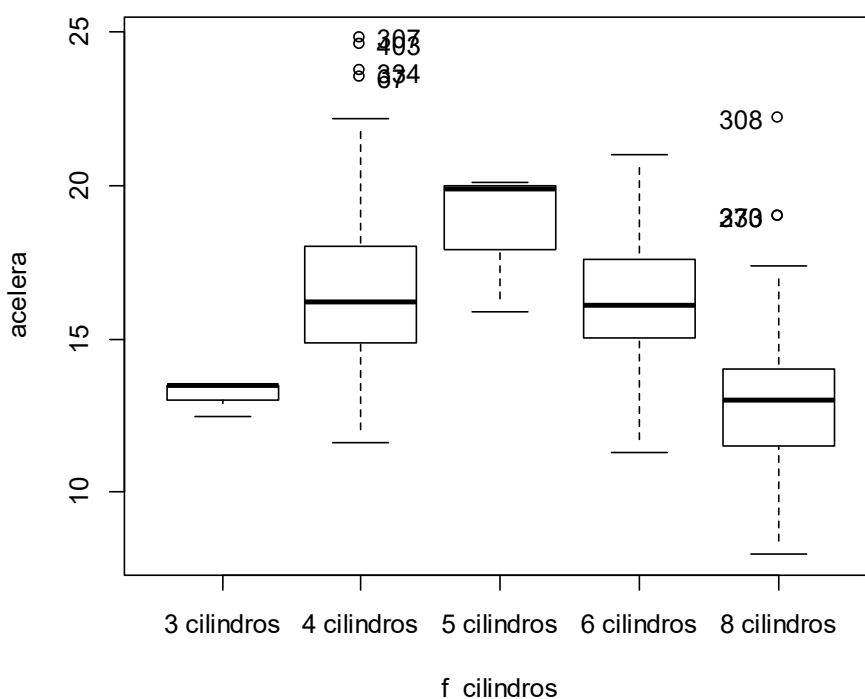
Para realizar el diagrama de cajas tendremos que ir a “**Gráficas-Diagrama de caja...**”, seleccionar la opción de gráfica por grupos con el factor origen y seleccionar la variable de la cual queremos realizarlo.



En la caja de Japón, el valor de la mediana está muy próximo o es el mismo que el de Q_1 en vez de estar entre Q_1 y Q_3 como es habitual. También podemos observar que la caja de Japón y Europa tienen valores parecidos, excepto la mediana mencionada anteriormente, mientras que la de EEUU es mucho más grande que las demás y con valores mayores.

15. Realiza un diagrama de cajas de la variable aceleración según el número de cilindros. ¿Existen valores anómalos en alguno de los grupos? Analiza las similitudes y diferencias que puedas observar sobre variable aceleración para cada número de cilindros.

Para realizar el diagrama de cajas tendremos que ir a “Gráficas-Diagrama de caja...”, seleccionar la opción de gráfica por grupos con el factor cilindros y seleccionar la variable de la cual queremos realizarlo.



En las cajas de 3 y 5 cilindros, el valor de la mediana está muy próximo o es el mismo que el de Q_3 en vez de estar entre Q_1 y Q_3 como es habitual. También podemos observar que la caja de 4 y 6 cilindros tienen valores parecidos y la de 8 cilindros tiene unas dimensiones parecidas a estas, pero con datos más bajos.

16. El 25 % de los coches que más consumen a los 100km, gastan más de, ¿cuántos litros?

Para calcular lo que nos pide el enunciado, en realidad lo que tenemos que hacer es hallar el percentil 75 (P_{75}). De esta manera obtenemos la cifra justa del último 25% de los coches que hay, por tanto los que más consumen. Esta cifra obtenida es la menor de este grupo de coches, por tanto todos esos coches gastan más de esa cifra.

75% n NA
13 398 8

Para ello tendremos que ir a **“Estadísticos-Resúmenes-Resúmenes numéricos...”** y obtendremos el siguiente resultado. En este caso, el 25 % de los coches que más consumen a los 100km, gastan más de 13 litros.

17. El 25 % de los coches de EEUU que menos consumen a los 100km, gastan menos de, ¿cuántos litros?

Para calcular lo que nos pide el enunciado, en realidad lo que tenemos que hacer es hallar el percentil 25 (P_{25}). De esta manera obtenemos la cifra justa del primer 25% de los coches que hay, por tanto los que menos consumen. Esta cifra obtenida es la mayor de este grupo de coches, por tanto todos esos coches gastan menos de esa cifra.

Para ello tendremos que ir a **“Estadísticos-Resúmenes-Resúmenes numéricos...”** y seleccionar la opción por grupos y hacer el resumen en función del f_origen.

	25%	n	NA
EEUU	10	248	5
Europa	8	70	3
Japón	7	79	0

Obtendremos el siguiente resultado, y de ahí deberemos elegir el dato perteneciente a los coches de EEUU. En este caso, el 25 % de los coches de EEUU que menos consumen a los 100km, gastan menos de 10 litros.

18. Estudia la relación entre: a) las variables niveles de consumo y de peso que hemos creado (fconsumo, fpeso). b) las variables niveles de consumo y país de origen.

Para estudiar la relación entre dos variables cualitativas deberemos realizar 2 tablas de frecuencias: la observada y la esperada. Deberemos estudiar la distancia que separa ambas tablas, y si están muy cerca, podremos decir que la variable depende de la otra. Para ello, utilizaremos el coeficiente chi-cuadrado que nos calcula R (p-value), y si este es menor que 0,05 podremos decir que la variable es dependiente de la otra.

Para ello tendremos que ir a **“Estadísticos-Tablas de contingencia-Tabla de doble entrada”**, y deberemos elegir las dos variables a estudiar. Seleccionaremos las opciones Test de independencia Chi-cuadrado e Imprimir las frecuencias esperadas.

a)

```
Frequency table:
      f_peso
f_consumo Ligeros Muy Pesdos Normales Pesados
Bajo      120      0      43      6
Medio     17      0      46     35
Alto       2     15      6    108

      Pearson's Chi-squared test

data:  .Table
X-squared = 309.32, df = 6, p-value < 2.2e-16

Expected counts:
      f_peso
f_consumo Ligeros Muy Pesdos Normales Pesados
Bajo  59.02261  6.369347 40.33920 63.26884
Medio 34.22613  3.693467 23.39196 36.68844
Alto  45.75126  4.937186 31.26884 49.04271
```

Como podemos observar, el valor de chi-cuadrado (p-value) es mucho menor que 0,05, por tanto podemos afirmar que la variable niveles de consumo depende de la variable niveles de peso.

b)

```
Frequency table:
      f_origen
f_consumo Japón Europa EEUU
Bajo      61      49     59
Medio     17      17     64
Alto       1       4    125

      Pearson's Chi-squared test

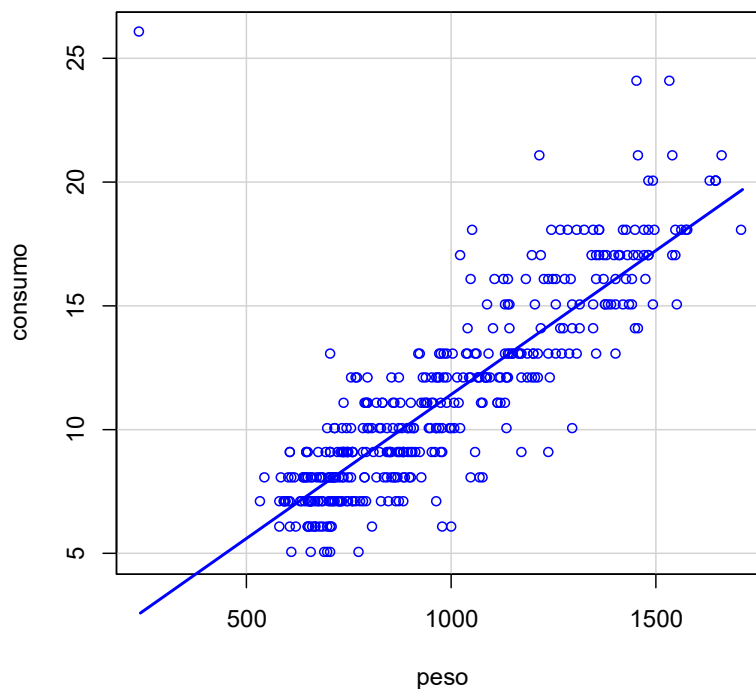
data:  .Table
X-squared = 118.79, df = 4, p-value < 2.2e-16

Expected counts:
      f_origen
f_consumo Japón Europa EEUU
Bajo  33.62972 29.79849 105.57179
Medio 19.50126 17.27960  61.21914
Alto  25.86902 22.92191  81.20907
```

Como podemos observar, el valor de chi-cuadrado (p-value) es mucho menor que 0,05, por tanto podemos afirmar que la variable niveles de consumo depende de la variable origen.

19. a) Haz un estudio de la relación entre el consumo y el peso, entendiendo que se quiere explicar el consumo en función del peso en kg. Analiza la bondad del ajuste, la presencia de relación lineal y valora la calidad del pronóstico de consumo que la recta de ajuste proporciona para un coche con peso 999kg. b) Estudia la relación lineal entre la aceleración y el peso en Tm.

a) Para realizar el estudio de dos variables cuantitativas debemos realizar el diagrama de dispersión y determinar la relación lineal entre dichas variables, además del ajuste de su recta. Para ello tendremos que ir a “**Gráficas- Diagrama de dispersión**” y seleccionar las dos variables a relacionar. La explicada deberá ir en el eje y y la explicativa en el eje x.



Ahora pasaremos a estudiar la relación lineal de las dos variables. Para ello tendremos que ir a “**Estadísticos- Ajuste de modelos-Regresión lineal**”, y seleccionar las variables. Obtendremos los siguientes datos:

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2791655  0.3929284  -0.71   0.478
peso         0.0116659  0.0003829  30.47  <2e-16 ***
-----
Residual standard error: 2.161 on 396 degrees of freedom
(8 observations deleted due to missingness)
Multiple R-squared:  0.701, Adjusted R-squared:  0.7002
F-statistic: 928.3 on 1 and 396 DF,  p-value: < 2.2e-16

```

Como podemos ver, el ajuste no es bueno, ya que engloba solo el 70% de los datos y esto está por debajo del 80% recomendado. En cuanto a la relación lineal de las variables, como el p-value es menor que 0,05 podemos decir que las 2 variables tienen una relación lineal, es decir, están relacionadas entre si.

De este ajuste podemos obtener la ecuación de la recta que es la siguiente:

$$\text{consumo} = 0,0116659 \cdot \text{peso} - 0,2791655$$

Ahora pasaremos a hacer un pronóstico de acuerdo a dicha ecuación, sabiendo el peso de un coche. Para un coche de 999 kg se espera un consumo de 11,3750686.

$$\text{consumo} = 0,0116659 \cdot 999 - 0,2791655 = 11,3750686 \text{ L por cada 100 km}$$

b) Ahora pasaremos a realizar un estudio de la relación lineal entre la aceleración y el peso en Tm. Para ello tendremos que ir a **“Estadísticos- Ajuste de modelos–Regresión lineal”**, y seleccionar las variables. Obtendremos los siguientes datos:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.5916     0.4639   42.229  <2e-16 ***
PesoTm        -4.1390     0.4507   -9.184  <2e-16 ***
-----
Residual standard error: 2.569 on 404 degrees of freedom
Multiple R-squared:  0.1727, Adjusted R-squared:  0.1707
F-statistic: 84.34 on 1 and 404 DF,  p-value: < 2.2e-16
```

Como el p-value es menor que 0,05 podemos decir que las 2 variables tienen una relación lineal, es decir, están relacionadas entre si.