Hoja 3 de problemas y prácticas con R

Estadística Computacional I. Grado en Estadística

Departamento de Estadística e Investigación Operativa. Universidad de Sevilla

Ejercicio 1

1. Acceder al fichero de datos USairpollution de la librería HSAUR2.

No vamos a cargar la librería completa para que no haya efectos sobre otros paquetes que vamos a utilizar.

```
library(HSAUR2)
```

```
## Loading required package: tools
```

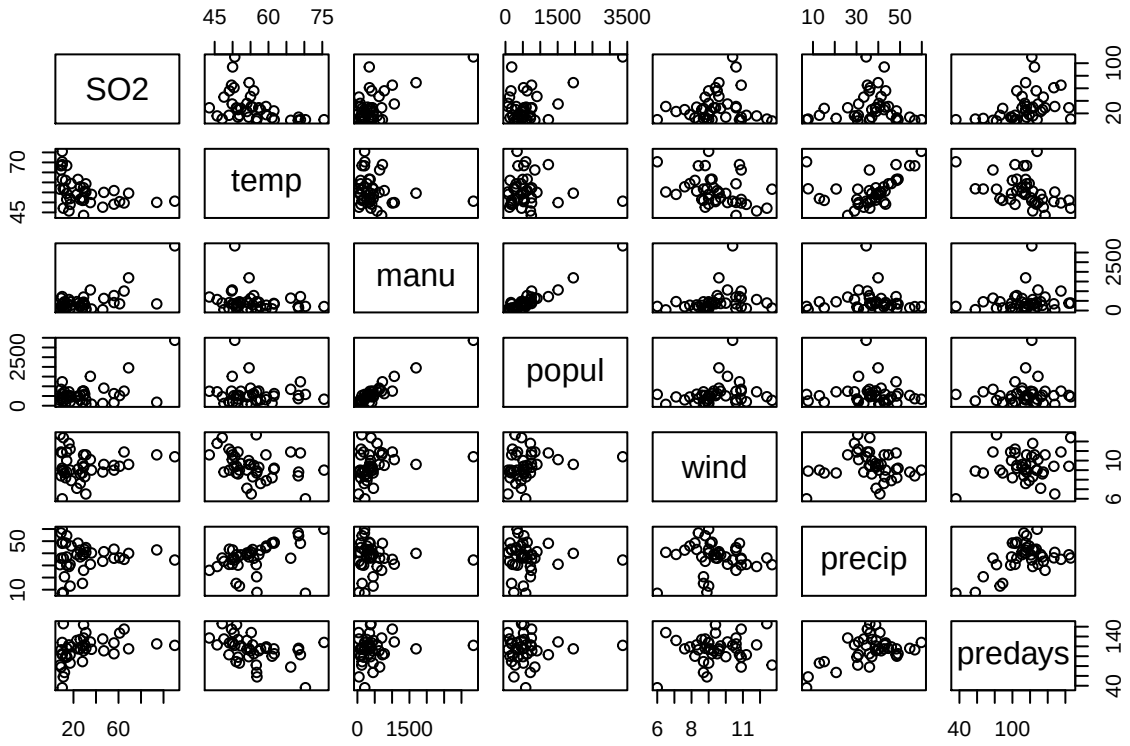
```
USairpollutioN=HSAUR2::USairpollution
glimpse(USairpollutioN)
```

```
## Observations: 41
## Variables: 7
## $ SO2     <int> 46, 11, 24, 47, 11, 31, 110, 23, 65, 26, 9, 17, 17, 35, 56,...
## $ temp    <dbl> 47.6, 56.8, 61.5, 55.0, 47.1, 55.2, 50.6, 54.0, 49.7, 51.5,...
## $ manu    <int> 44, 46, 368, 625, 391, 35, 3344, 462, 1007, 266, 641, 454, ...
## $ popul   <int> 116, 244, 497, 905, 463, 71, 3369, 453, 751, 540, 844, 515,...
## $ wind    <dbl> 8.8, 8.9, 9.1, 9.6, 12.4, 6.5, 10.4, 7.1, 10.9, 8.6, 10.9, ...
## $ precip  <dbl> 33.36, 7.77, 48.34, 41.31, 36.11, 40.75, 34.44, 39.04, 34.9...
## $ predays <int> 135, 58, 115, 111, 166, 148, 122, 132, 155, 134, 78, 86, 10...
```

Solución ejercicio 1

```
i)    Generar las nubes de puntos para cada par de variables.
```

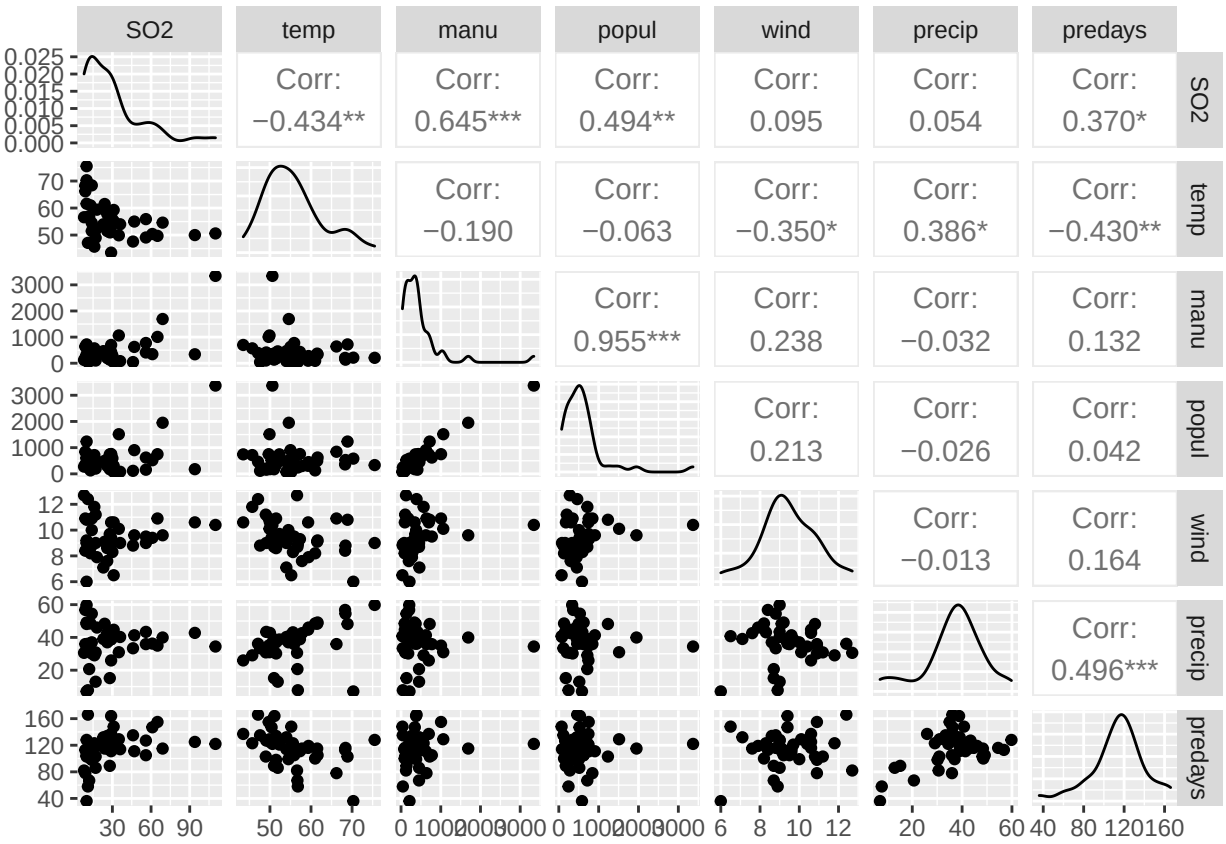
```
plot(USairpollution)
```




```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
USairpollutioN %>%
  ggpairs()
```



Calcular la matriz de correlaciones.

```
USairpollutioN %>% cor()
```

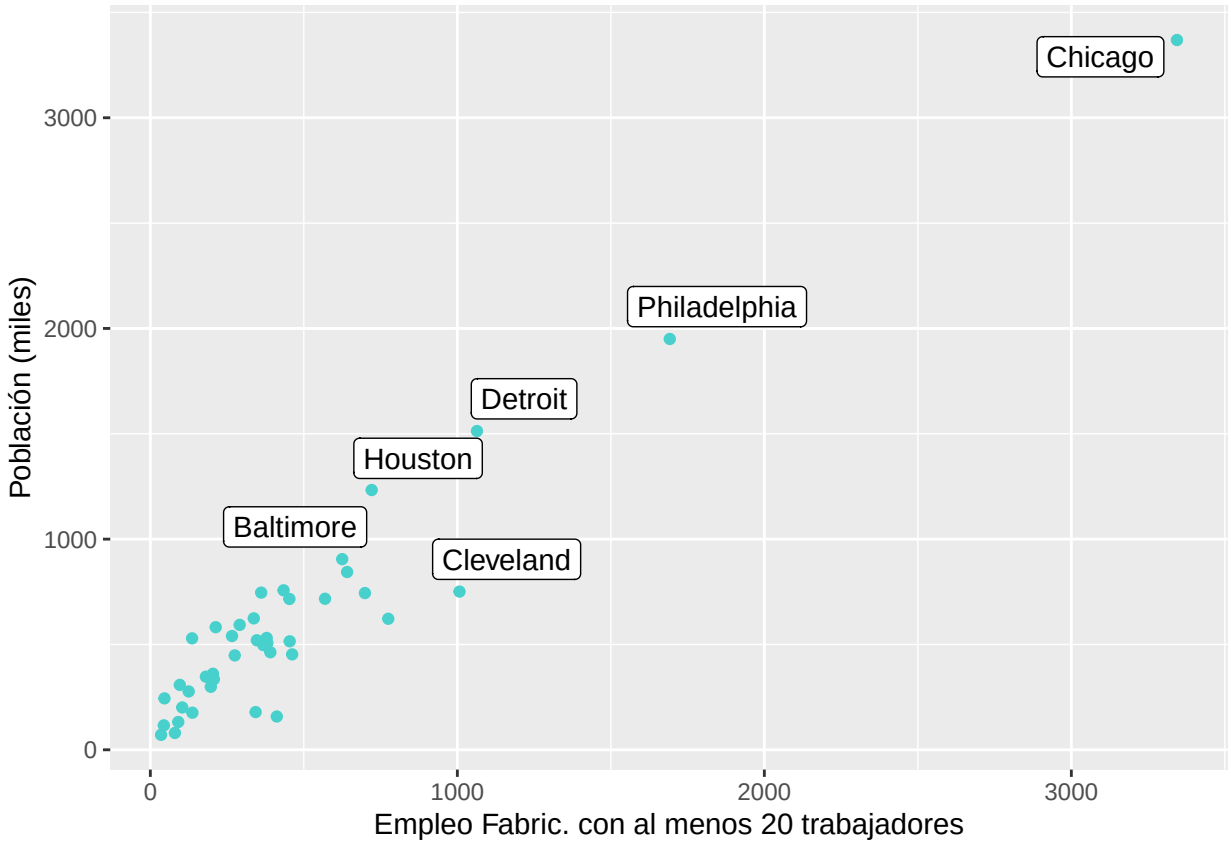
```
##                SO2        temp        manu       popul        wind      precip
## SO2      1.00000000 -0.43360020  0.64476873  0.49377958  0.09469045  0.05429434
## temp    -0.43360020  1.00000000 -0.19004216 -0.06267813 -0.34973963  0.38625342
## manu     0.64476873 -0.19004216  1.00000000  0.95526935  0.23794683 -0.03241688
## popul    0.49377958 -0.06267813  0.95526935  1.00000000  0.21264375 -0.02611873
## wind     0.09469045 -0.34973963  0.23794683  0.21264375  1.00000000 -0.01299438
## precip   0.05429434  0.38625342 -0.03241688 -0.02611873 -0.01299438  1.00000000
## predays  0.36956363 -0.43024212  0.13182930  0.04208319  0.16410559  0.49609671
##             predays
## SO2      0.36956363
## temp    -0.43024212
## manu     0.13182930
## popul    0.04208319
## wind     0.16410559
## precip   0.49609671
## predays  1.00000000
```

ii) Obtener la nube de puntos para las variables manu y popul.

```
library(ggrepel)
p1=USairpollutioN %>%
  ggplot(aes(x=manu , y=popul ,
             label=rownames(USairpollutioN))) +
  geom_point(color="mediumturquoise") +
  geom_label_repel(max.overlaps = 12)+
```

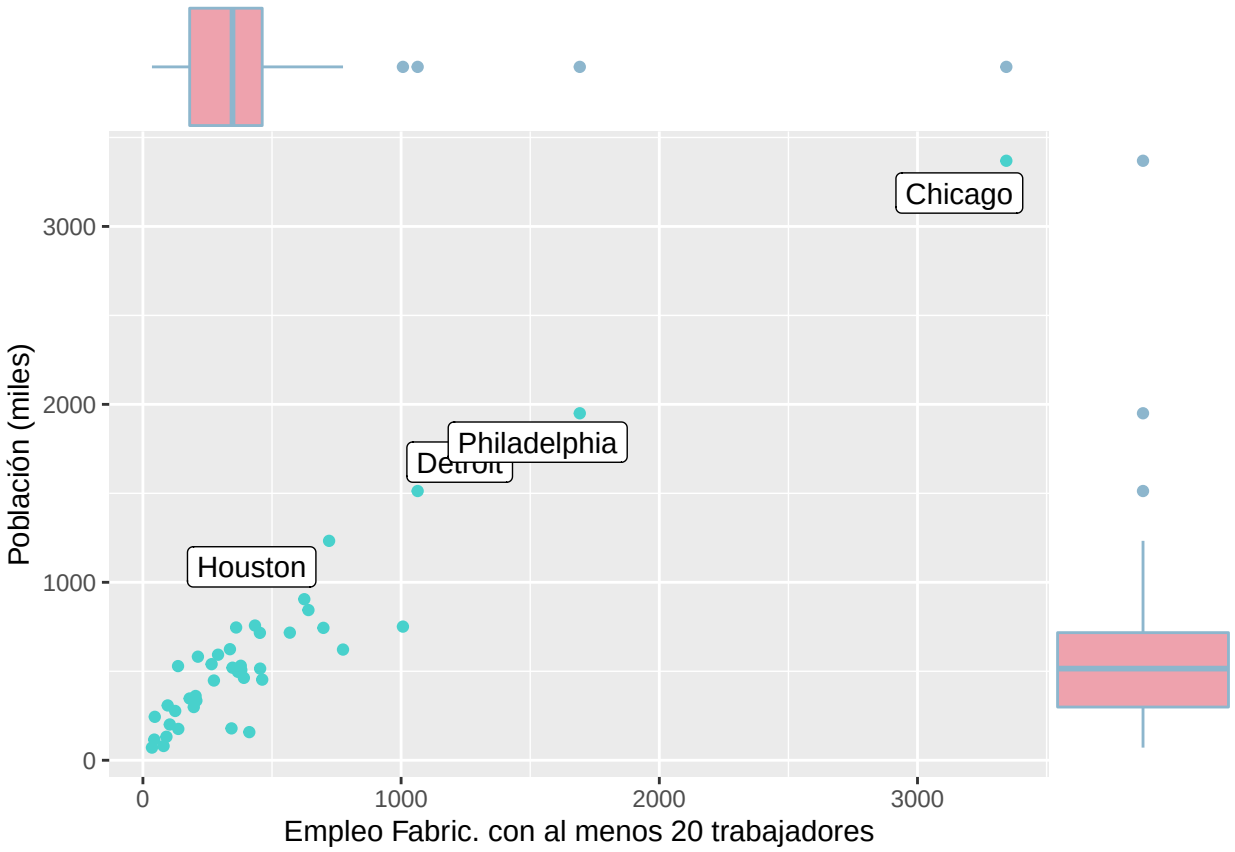
```
  labs(x="Empleo Fabric. con al menos 20 trabajadores",
       y="Población (miles)");p1
```

```
## Warning: ggrepel: 35 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



iii) Añadir a la nube de puntos anterior un gráfico caja y bigotes de popul y un histograma de manu.

```
library(ggExtra)
ggMarginal(
  p1,
  type="boxplot",
  colour="lightskyblue3",
  fill="lightpink2"
)
```



2. Leer el fichero de datos decathlon1989.sav (formato SPSS).

Solución ejercicio 2

```
i)  Dibujar todos los gráficos de caja y bigotes por separado.


ii) Generar las nubes de puntos y añadir histogramas en la diagonal, superponer rectas de minimos cuadra


iii)    Utilizar la librería corrplot para visualizar la matriz de correlaciones.


iv) Obtener las principales medidas descriptivas. Tipificar los datos.
```

Ejercicio 3

Solución ejercicio 3

3. Cargar el espacio de trabajo “Pisa2009.RData”. Dibujar representaciones multivariantes como las caras de Chernoff o los gráficos “star”.

Ejercicio 4

Solución ejercicio 4

4. Leer el fichero "familias.txt", que contiene información sobre 50 familias: identificación, ingresos, número de adultos y de 14 a 16 años, número de menores de 14 años, zona de residencia y máximo nivel de estudios alcanzado. La primera fila es una cabecera con los nombres de las variables.
   i) Calcular y almacenar en una variable del data frame las unidades de consumo de cada familia: 1 para el primer mayor de edad ó entre 14 y 16, 0'5 por cada uno de los restantes de este grupo y 0'3 por cada menor 14. (Definición oficial).
   ii) Calcular para cada familia los ingresos por cada unidad de consumo. Se fija el umbral de pobreza (relativa) como el 60% de la mediana de los ingresos por unidad de consumo. Una persona cuya familia tenga unos ingresos por unidad de consumo inferiores a dicho umbral estará en estado de pobreza relativa.
   iii) Detectar las familias en estado de pobreza relativa.
   iv) Hallar el porcentaje de familias en estado de pobreza.
   v) Hacer un diagrama de barra de dichos porcentajes.
   vi) Hallar los ingresos medios por unidad de consumo para cada grupo de pobreza.
   vii) Hallar la desviación típica para cada grupo de pobreza.
   viii) Hallar la mediana de ingresos para cada grupo definido por el nivel educativo máximo en la familia.
   ix) Calcular la brecha de pobreza definida como la diferencia entre la media de ingresos por unidad de consumo de las familias que representan el 20% de las familias con ingresos más altos, y la media de dichos ingresos para las familias que representan el 20% de las familias con ingresos más bajos.

Ejercicio 5

Solución ejercicio 5

5. El fichero "Preferencias_Marcas.txt" contiene en cada fila la edad, sexo y marca preferida de cierto producto para un conjunto de personas.
   i) Leer los datos y obtener un resumen inicial. El código "99" indica valor perdido.
   ii) Encontrar los casos incompletos. Seleccionar los casos completos para el resto del ejercicio.
   iii) Obtener y representar la función de distribución de la variable edad.
   iv) Construir una variable categórica para la edad, utilizando los siguientes puntos de corte: (18, 23, 29, 50, 58, 67).
   v) Obtener las tablas de frecuencias de cada variable categórica.
   vi) Cruzar las variables edad (intervalos) y marca.
   vii) Cruzar las tres variables categóricas.
   viii) Resumir la edad según la marca.

Ejercicio 6

Solución ejercicio 6

6. El fichero "matrimoniosA2009.txt" contiene un registro por cada matrimonio oficial realizado en España en 2009. Se desean leer las siguientes variables (entre paréntesis las columnas que ocupan, ya que se trata de un fichero con formato fijo y valores posibles):

   - Mes (6-7);
   - Tipo de celebración (12) (1= católico, 2=otra religión, 3= exclusivamente civil);
   - Código provincial (13-14) (1...52);
   - Año nacimiento cónyuge A (23-26); Sexo cónyuge A (41) (1: Hombre, 6: Mujer)
   - Año nacimiento cónyuge B (71-74); Sexo cónyuge B (89) (1: Hombre, 6: Mujer)

i) Leer los datos de acuerdo con este diseño de registro. Seleccionar los casos completos, eliminando los registros que presenten valores perdidos.
ii) Convertir en variables tipo factor las variables categóricas, definiendo apropiadamente los nombres de los códigos. Para el código provincial, leer los nombres de las provincias en el fichero SPSS "CódigosProvComINE.sav".
iii) Construir nuevas variables que contengan las edades de los cónyuges y la diferencia de edad entre los cónyuges A y B.
iv) Almacenar en un data frame permanente el fichero con todas las variables construidas. Borrar el espacio de trabajo y cargar el data frame.
v) Obtener los totales de matrimonios y los porcentajes correspondientes para el cruce de las variables Sexo Cónyuges.
vi) Tabular el código provincial.
vii) Obtener una tabla donde para cada provincia aparezca el total de matrimonios, el para cada par (Sexo A, SexoB) y según tipo de celebración.
viii) Construir una tabla similar a la anterior pero con los porcentajes dentro de cada provincia. Ordenar las provincias según el porcentaje de matrimonios católicos. Representar gráficamente las provincias según los porcentajes de matrimonios católicos y civiles.
ix) Analizar por separado las edades de los cónyuges, tanto en su escala original como categorizando con los puntos de corte (15,30,45,65,100).
x) Resumir la diferencia de edad, tanto globalmente como según otras características.