Fundamentos lenguajes
Proyecto

19 de Enero de 2018

Parte 1 (1 punto) Con el dataset `diamonds` (viene incluido en el paquete `ggplot2`), crear un nuevo data.frame que sea el resultado de aplicar las siguientes operaciones:

1. Filtrar los diamantes con corte "Ideal".
2. Seleccionar las columnas `carat`, `cut`, `color`, `price` y `clarity`.
3. Crear una nueva columna precio/quilate.
4. Agrupar los diamantes por color.
5. Calcular la media del precio/quilate para cada uno de los grupos anteriores.
6. Ordenar por precio/quilate de forma descendente.

Parte 2 (1 punto)

1. Cargar el paquete `tidyr`.
2. Leer el conjunto de datos `weather.txt`. Cuidado con los *missing values*, están codificados como "-" (ver parámetro `na.strings` de `read.table`).
3. Identificar cuales son las variables en los datos.
4. Agrupar las variables `d1`–`d31` en dos variables `día` y `temperatura` (función `gather`).
5. Convertir las columnas `element` y `temperatura` en dos variables `TMAX` y `TMIN` (función `spread`, es la operación contraria a `gather`).
6. Separar la columna `id` en dos variables, `país` e `id`.

Parte 3 (4 puntos) Con el conjunto de datos `diamonds` original (no el modificado en el ejercicio 1):

1. Ver el tipo de cada una de las variables.
2. Realizar un análisis estadístico de las variables numéricas: calcular la media, varianza, rangos, etc. ¿Tienen las distintas variables rangos muy diferentes?.
3. Hacer un gráfico de cajas de la variable `price` para cada uno de los distintos valores de `color`.
4. Hacer el mismo gráfico del punto anterior pero con un gráfico de cajas para cada uno de los valores de la variable `cut`.
5. Calcular la correlación de todas las variables numéricas con la variable `price`.
6. Crear un histograma de la variable `carat` para cada uno de los distintos valores de `color`. ¿Son muy diferentes las distribuciones?.
7. Realizar un gráfico de dispersión para las variables que tienen más y menos correlación con `price` y comentar los resultados. ¿Como sería el gráfico de dispersión entre dos vectores con correlación 1?.
8. Definimos los *outliers* como los elementos (filas) de los datos para los que cualquiera de las variables (numéricas) está por encima o por debajo de la mediana más/menos 3 veces el MAD (Median Absolute Deviation). Identificar estos outliers y quitarlos.
9. Separar el conjunto de datos en dos subconjuntos disjuntos de forma aleatoria, el primero conteniendo un 70% de los datos y el segundo un 30%.
10. Escalar los datos para que tengan media 0 y varianza 1, es decir, restar a cada variable numérica su media y dividir por la desviación típica. Calcular la media y desviación en el conjunto de train, y utilizar esa misma media y desviación para escalar el conjunto de test.

Parte 4 (4 puntos) Con el conjunto de datos `titanic.csv` de la práctica 3:

1. Representar, en un mismo gráfico, dos histogramas de la variable `age`, uno para los pasajeros con sexo masculino y otro para los pasajeros con sexo femenino. En caso de que se solapen los histogramas, usar colores con transparencias (ver función `rgb()`).

2. Examinar la variable `name`, ¿qué otra variable podemos extraer de la misma?. Extraer los distintos valores de esa variable.

3. Crear una nueva variable `title` con los valores `Master` (hombre soltero), `Miss` (mujer soltera), `Mr.` (hombre casado), `Mrs.` (mujer casada) y `Otro` a partir de la variable nombre. Es importante tener en cuenta que el título `Miss` está en ocasiones codificado con su abreviatura en frances `Mlle` (*mademoiselle*) y lo mismo ocurre con `Mrs.`, que en ocasiones aparece como `Ms.` ó `Mme` (*madame*).

4. Explorar la relación entre las variables `age` y la nueva variable `title` mediante un `boxplot` para cada uno de los valores de la misma. ¿Tienen alguna relación?.

5. Ver la relación entre la supervivencia la nueva variable `title` con un gráfico de barras. En el caso del valor `Otros` de la variable `title`, ¿nos proporciona este alguna información sobre la supervivencia?. ¿A qué se debe?.

6. Corregir el problema anterior con el grupo `Otros` dividiendo el mismo en dos nuevos títulos. Para ello se puede explorar los datos y hacer "trampas", es decir, ver qué títulos hasta ahora categorizados como `Otros` han sobrevivido y cuales no y si se puede encontrar un patrón común entre los mismos.

7. Explorar la relación entre `age`, `pclass` y `title` en varios gráficos de dispersión con colores, donde el color representa la supervivencia (Pista: usar facetas).

8. En la práctica 3 se han completado los *missing value* de la variable `age` con la mediana de sus valores. De acuerdo al gráfico del punto 7, ¿es esta la solución correcta?. Completar ahora los *missing values* pero con la mediana de los valores de acuerdo a las variables `pclass` y `title`.

Entrega La entrega se realizará se realizará a través de Moodle en un único fichero .zip que tenga 4 ficheros .R, uno con cada ejercicio. El nombre del fichero .zip debe de ser `P4_<apellidos>.zip`. Incluir comentarios en el código siempre que se considere necesario. Las respuestas planteadas a las preguntas deben responderse como comentarios en el fichero .R después de la línea de código correspondiente.

Criterios de evaluación Para resolver los ejercicios se pueden utilizar indistintamente funciones de R base o de paquetes adicionales. Es conveniente (y se valorará) utilizar un estilo de programación adecuado. Algunas directrices pueden encontrarse en la Guía de estilo de R de Google:

`https://google.github.io/styleguide/Rguide.xml`. Además del estilo, se valorará que el código R sea:

- Correcto
- Claro
- Conciso
- General

Ejemplo: para calcular la media de cada columna de una data.frame podemos hacerlo de, al menos, 3 formas:

1. ```
   mean(mtcars$gear)
   mean(mtcars$mpg)
   mean(mtcars$wt)
   ...
   ```
2. ```
   for(i in 1:ncol(mtcars)) {
      mean(mtcars[,i])
   }
   ```
3. ```
   lapply(mtcars, mean)
   ```

Aunque las tres obtienen el mismo resultado, en este ejemplo preferimos la tercera forma ya que el código es más claro, conciso y general.
   ```