Práctica 2: R

Enunciado

Vamos a trabajar con los ficheros "Estaciones.txt" y "Precios_2017_04_02.txt" (fuente), que contienen información sobre el precio de distintos carburantes el día 2 de Abril de 2017 en estaciones de toda España:

1. Lee fichero de precios. Asegúrate de que todas las variables de precios son numéricas.
2. Cambia los espacios en los nombres de las columnas por '_'.
3. Mira los tipos de datos del dataset de precios. Observa que la fecha no ha sido interpretada como tal. Conviértela a tipo fecha.
4. Calcula el porcentaje de NAs de cada columna para la del dataframe de precios.
5. Elimina las columnas que no tengan ningún valor distinto de NA.
6. Crea un dataframe con la media, máximo y mínimo de las columnas numéricas.
7. Habrás visto que el máximo del precio para gasolina 95 es un valor extredemadante alto. Para la variable precio de la gasolina 95, cambia todos los valores mayores a 3 (valores anómalos) por NA.
8. Selecciona las gasolineras que vendan gasolina 95 o 98 utilizano el dataframe resultado del ejercicio 5.
9. Selecciona las columnas que representan variables de precios usando el dataframe resultado del ejercicio 5.
10. Selecciona las 5 gasolineras con el precio de gasóleo B más barato.
11. Añade dos nuevas variables con los litros por euro de gasolina 95 y 98 usando el dataframe del ejercicio 5.
12. Crea un nuevo dataframe a partir del resultado del ejercicio 7 con las columnas `ID_estacion`, `fecha`, `tipo_gasolina` y `precio` y guarda el resultado en la variable `precios_long`. La variable `tipo_gasolina` tendrá los valores: "Precio gasolina 95", "Precio gasóleo A", "Precio gasóleo B", "Precio bioetanol", "Precio nuevo gasóleo A", "Precio biodiesel", "Precio gasolina 98", "Precio gas natural comprimido", "Precio gas natural licuado" y "Precio gases licuados del petróleo".
13. En la variable `tipo_gasolina` del ejercicio anterior, elimina el texto "Precio".
14. Haz un histograma de todas las variables de precio **usando facetas** de ggplot.
15. Lee el fichero de estaciones y júntalo con el dataframe resultado del ejercicio 5 utilizando como clave el `ID_ESTACION`. Considerar únicamente las gasolineras presentes en las dos tablas.
16. Calcula el precio medio de gasolina 95 por provincia.
17. Añade una variable de tipo lógico con nombre `ind_24H`, indicando si la gasolinera es 24H.
18. Realiza un gráfico de barras con el conteo de estaciones por cada tipo de rótulo.
19. Convierte la variable rótulo en un factor que tenga 6 niveles: los 5 rótulos más frecuentes y un nivel "Otros" con el resto.
20. Representa cada gasolinera como un punto en función de su longitud y latitud. Cada punto debe tener un tamaño proporcional al precio de la gasolina 95 y el color debe indicar si se trata de una gasolinera Cepsa, Repsol u otros.

Entrega

La fecha límite de entrega de la práctica es el día **26 de enero de 2020 a las 23.55h**.

Junto con el enunciado de la práctica se proporciona una plantilla de un archivo .Rmd (`practica_template.Rmd`). Este fichero esta dividido en tantos apartados como ejercicios y todos los análisis/comentarios que se deseen proporcionar para cada pregunta deberán incluirse como comentarios de código en el ejercicio correspondiente.

La entrega consiste en un fichero .zip con nombre `<apellido1>_<apellido2>_<nombre>.zip` que contiene el fichero .Rmd proporcionado y una versión compilada en HTML. El nombre del fichero .Rmd tendrá también el formato `<apellido1>_<apellido2>_<nombre>.Rmd`. Por ejemplo: `rodriguez_lujan_irene.Rmd`.

Criterios de evaluación

La práctica se califica sobre 10 puntos. Cada ejercicio tiene un valor de 0.5. Para resolver los ejercicios se pueden utilizar indistintamente funciones de R base o de paquetes adicionales, aunque se recomienda el uso de las funciones del tidyverse. Es conveniente (y se valorará) utilizar un estilo de programación adecuado. Algunas directrices pueden encontrarse en la Guía de estilo http://style.tidyverse.org/. Además del estilo, se valorará que el código R sea:

- Correcto
- Claro
- Conciso
- General

Ejemplo: para calcular la media de cada columna de una dataframe podemos hacerlo de, al menos, 4 formas:

1. Copy-paste

```
mean(mtcars$gear)
mean(mtcars$mpg)
mean(mtcars$wt)
# ...
```

2. Bucle

```
for(i in seq_along(mtcars)) {
  mean(mtcars[, i])
}
```

3. `purrr`

```
library(purrr)
map_dbl(mtcars, mean)
```

4. `dplyr`

```
library(dplyr)
summarize_all(mtcars, mean)
```

Aunque las cuatro obtienen resultados similares, en este ejemplo preferimos la tercera o la cuarta forma ya que el código es más claro, conciso y general.