Práctica 3: R

Enunciado

El conjunto de datos `comptagevelo2017.csv` contiene información sobre el número de personas que circulan en bicicleta por cada uno de los distritos de la ciudad de Montreal a lo largo del año 2017 http://donnees.ville.montreal.qc.ca/dataset/velos-comptage. En este conjunto de datos, las filas representan los dias del año y las columnas cada uno de los distritos. La columna 2 contiene un timestamp que vamos a ignorar. Sobre este conjunto de datos:

1. Leer el fichero `comptagevelo2017.csv` como un dataframe.
2. Eliminar la columna 2 del dataframe.
3. Identificar cuales son las variables que están contenidas en el dataframe. A continuación, transformar ese dataframe para que cada columna represente una única variable.
4. Crear tres nuevas variables en el dataframe que contengan la información del día, mes y año respectivamente (manteniendo la columna Date).
5. Convertir la columna Date a tipo date.
6. Modificar la columna de los distritos para eliminar los espacios alrededor de "/".
7. Calcular el porcentaje de días en los que faltan datos para cada uno de los distritos.
8. Calcular el total de ciclistas que pasa por cada uno de los distritos a lo largo de todo el año.
9. ¿Cuales son los cinco distritos con más número de ciclistas?
10. Realizar un gráfico de barras horizontales donde el eje x representa el total de ciclistas y el eje y los distritos.
11. Realizar un gráfico de lineas con la evolución mensual de ciclistas para cada distrito. En el gráfico tiene que aparecer una línea por distrito.
12. Ordenar las barras del gráfico del punto 10 de mayor (arriba) a menor (abajo) según el número de ciclistas.
13. Añadir sobre el gráfico del punto 11 una linea de color azul y más ancha que el resto con la media de ciclistas por mes.
14. Realizar un gráfico de barras del número de ciclistas para cada día de la semana en cada uno de los cinco distritos con más ciclistas (usando facetas).
15. Completar los missing values de la columna que representa el número de ciclistas con la media del resto de datos de esa variable pero agrupado por distrito y mes.
16. Leer el fichero `localisationcompteursvelo2015.csv`. Importante: la codificación del fichero no es UTF-8 sino ISO-8859-1.
17. Realizar un gráfico de puntos de las columnas coord X (eje x) y coord Y (eje y), con el color de los puntos representando la variable Type y la forma la variable Etat.
18. Hacer un join de los dos dataframes por las columnas con los nombres del districto en el primer dataframe y "nom comptage" en el segundo.
19. ¿Ver qué districtos del primer dataframe no se encuentran en el segundo?

20. Realizar un gráfico de puntos del dataframe resultante del ejercicio 18 de las columnas coord X (eje x) y coord Y (eje y), donde el tamaño de los puntos representa el número total de ciclistas que pasaron por ese districto a lo largo de todo el año.

Entrega

La fecha límite de entrega de la práctica es el día **8 de noviembre de 2020 a las 23.55h**.

La entrega consiste en un fichero .Rmd o .R de la práctica con nombre `<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.