Máster en Big Data. Fundamentos matemáticos del análisis de datos.

Examen.

Fernando San Segundo

Curso 2020-21.

Preliminares.

- Crea una carpeta para los ficheros del examen llamada:

  `examen_apellido_nombre`

  También puedes, si lo prefieres, crear esa carpeta como un proyecto de RStudio.

- Guarda el fichero `2020-Examen-MBDFME.Rmd` disponible en Moodle en esa carpeta. Es el fichero que usarás para introducir tus respuestas. Si tienes problemas para descargarlo desde Moodle abre un script vacío en el editor de RStudio y *eligiendo previamente la carpeta del examen como directorio de trabajo* copia este script y ejecútalo:

```
# Comprobación del directorio de trabajo
getwd()

nombreFichero =  "2020-Examen-MBDFME.Rmd"

urlFichero = paste0("https://musing-stonebraker-3eeaa5.netlify.app/chova/", nombreFichero)

if(!file.exists(nombreFichero)){
  download.file(urlFichero, nombreFichero)
} else {
  warning(paste0("Cuidado: el fichero ", nombreFichero, " ya existe en tu directorio de trabajo. ")
}
```

- La carpeta del proyecto debe contener una subcarpeta llamada `datos`.

- Vamos a trabajar con una tabla de datos llamada `aireMadrid.csv`, que puedes descargar desde este enlace:
  aireMadrid.csv
  Si ese enlace no funciona abre un script vacío en el editor de RStudio y *eligiendo previamente la carpeta del examen como directorio de trabajo* copia este script y ejecútalo:

```
# Comprobación del directorio de trabajo
getwd()

dataFile =  "aireMadrid.csv"

dataURL = paste0("https://musing-stonebraker-3eeaa5.netlify.app/chova/datos", dataFile)

localFile = paste0("./datos/", dataFile)
```

```r
if(!dir.exists("datos")) warning("CUIDADO: No existe la carpeta datos en tu directorio de trabajo."
if(!file.exists(localFile)){
  download.file(dataURL, destfile = localFile)
} else {
  warning(paste0("Cuidado: el fichero de datos", dataFile, " ya existe en tu carpeta de datos"))
}
```

- Los valores de esta tabla reflejan las mediciones de distintos contaminantes atmosféricos medidos a lo largo del año 2019 en las esraciones de la red del "Sistema Integral de la Calidad del Aire" del Ayunatmiento de Madrid.
  - Cada fila de la tabla corresponde a la medición de uno de esos contaminantes en una estación y fecha concreta.
  - El tipo de contaminante de esa medición se indica mediante un código numérico en la variable `magnitud`.
  - Todos los valores de la variable `medicion` están en $\mu g/m^3$.
  - Se incluye además para cada medición la informacion sobre los milímetros de lluvia caídos ese día en Madrid.

Instrucciones de entrega.

- Cuando termines de responder a las preguntas del examen, comprime la carpeta de tu proyecto en formato `zip`, de manera que el archivo comprimido tenga el nombre de esa carpeta y extensión `.zip`. Enviáme por correo ese fichero comprimido a la dirección `fsansegundo@comillas.edu`.
- Se valorará adicionalmente el uso de funciones del tidyverse (dplyr, ggplot) para las representaciones gráficas y las operaciones con tablas.

Apartado 1.

Lee el fichero de datos y crea un data.frame de R llamado `aireMadrid`, que usaremos para el resto del examen.

- C1: ¿Cuántas observaciones hay en la tabla? ¿Cuántas variables?
- C2: ¿Hay datos ausentes? ¿De qué tipo son las variables? Para empezar a entender la estructura de los datos haz alguna exploración preliminar de los datos. En particular, haz una tabla de `magnitud` frente a `estacion` (la tabla es grande). Verás que no disponemos de mediciones de todas las magnitudes para todas las estaciones.
- C3: Piensa en la estructura de la tabla y la forma en la que se codifica la información. Para guiarte en este apartado, analiza por ejemplo si las mediciones de `Dióxido de azufre` (magnitud 1) siguen una distribución normal (haz un análisis gráfico, incluyendo los datos de ozono de todas las estaciones). Los datos de la tabla ¿son *datos limpios* (tidy)? Atención: no se pide que los transformes en datos limpios, sino que digas si lo son.

Apartado 2.

- C4: Añade a la tabla un factor `lluvia` según si ese día llovía o no. Se considera que llovía si el valor de la variable `lluvia_mm` es distinto de cero. Haz una tabla de frecuencia relativa de ese factor. ¿Qué probabilidad están estimando esos valores? (Cuidado, no es la probabilidad de que un día elegido al azar haya llovido).
- C5: El `dióxido de azufre` (magnitud 1) solo se ha medido en algunas estaciones de la red. Calcula el valor medio de las mediciones de `dióxido de azufre` en cada una de esas estaciones según si llovía o no. ¿Qué conclusión extraes?

- C6: ¿Cuál es la probabilidad de que una medición de la tabla elegida al azar corresponda a `monóxido de nitrógeno` (magnitud 7)? Sabiendo que una medición correspomde a `monóxido de Nitrógeno`, ¿cuál es la probabilidad de que proceda de la estación 24?
  Usa estos valores para calcular la probabilidad de que una medición elegida al azar sea de `monóxido de nitrógeno` y se haya hecho en la estación 24. Comprueba el resultado calculando directamente esa probabilidad.
- C7: Si elegimos 10 mediciones al azar (y con remplazamiento) procedentes de la estación 49, ¿cuál es la probabiidad de que 3 o más sean de `monóxido de nitrógeno`?

Apartado 3.

- C8: Las estaciones 24 (Casa de Campo) y la 56 (Plaza Elíptica) representan ubicaciones muy distintas. La primera está situadas en una zonas verde mientras que la segunda tiene niveles altos de tráfico. Es natural preguntarse si los niveles de contaminantes de estas dos estaciones difieren. Para esta pregunta vamos a elegir como contaminante las partículas de menos de 10 μm (magnitud 10). Analiza gráficamente (con la herramienta que creas más adecuada) si la ubicación de esas estaciones afecta a ese contaminante.
- C9: Calcula un intervalo de confianza para el nivel medio de partículas de menos de 10 μm (magnitud 10) en las mediciones de la estación de Casa de Campo. Haz lo mismo para las de la Plaza Elíptica. ¿Se solapan los intervalos? ¿Qué conclusión extraes?
- C10: Haz un contraste de la hipótesis alternativa: *el valor medio de las mediciones de partículas de menos de 10 μm en la estación de Casa de Campo es menor que el de Plaza Elíptica.*
  Nota: si no sabes hacer el contraste de dos muestras mira la sesión 6 del curso. Si con eso no te aclaras, alternativamente puedes hacer un contraste de una muestra: calcula la media de las observaciones de Plaza Elíptica. Llama μ_0 a ese valor y contrasta la hipótesis alternativa *el valor medio de las mediciones de partículas de menos de 10 μm en la estación de Casa de Campo es menor que* μ_0.

Apartado 2.

- C11: Vamos a hacer un modelo de regresión lineal para las dos variables:
  x : nivel de partículas de menos de 10 μm (magnitud 10)
  y : nivel de partículas de menos de 2.5 μm (magnitud 9)
  En esta pregunta solo usaremos datos procedentes de la estación 38 (Cuatro Caminos) porque en esa estación se midieron los valores de ambas variables exactamente los mismo días. Para construir el modelo es necesario en primer lugar identificar las mediciones de esas dos variables correspondientes a un mismo día (aparecen las dos mediciones en la misma columna). Puedes hacer esto de dos maneras. La mejor es usar una función del `tidyR`. Si haces esto asegúrate de ajustar los nombres de las variables de la tabla para que sean sintácticamente correctos (los números no sirven como nombres). La función `names` permite examinar pero también modificar los nombres de una tabla.
  Alternativamente, puedes aprovecharte del hecho de que los días de medición *son los mismos y aparecen en el mismo orden.* Gracias a eso puedes: (i) dividir la tabla en dos (una conn los valores de x y otra con los de y), (ii) extraer los valores de y de una de esas dos tablas y (iii) añadirlos a la otra.
  Sea cual sea el método que uses, construye el modelo y dibuja el diagrama de dispersión de esas variables junto con la recta de regresión obtenida.
- C12: ¿Qué porcentaje de la variabilidad en el nivel de partículas de menos de 2.5 μm se explica con el modelo? Examina además los gráficos diagnósticos del modelo y úsalos para hacer un juicio sobre si ese modelo cumple las hipótesis del modelo lineal.