Master en Big Data. Fundamentos Matemáticos del Análisis de Datos (FMAD).

Tarea 1

Departamento de Matemática Aplicada

Curso 2021-22. Última actualización: 2021-09-09

Instrucciones preliminares

- Empieza abriendo el proyecto de RStudio correspondiente a tu repositorio personal de la asignatura.
- En todas las tareas tendrás que repetir un proceso como el descrito en la sección *Repite los pasos Creando un fichero Rmarkdown para esta práctica* de la *Práctica00*. Puedes releer la sección *Practicando la entrega de las Tareas* de esa misma práctica para recordar el procedimiento de entrega.

Ejercicio 0

- Si no has hecho los *Ejercicios* de la *Práctica00* (págs. 12 y 13) hazlos ahora y añádelos a esta tarea. Si ya los has hecho y entregado a través de GitHub no hace falta que hagas nada.

Ejercicio 1. Análisis exploratorio de un conjunto de datos y operaciones con dplyr.

- Vamos a utilizar el conjunto de datos contenido en el fichero (es un enlace):
  cholesterol.csv
  Los datos proceden de un estudio realizado en la *University of Virginia School of Medicine* que investiga la prevalencia de la obesidad, la diabetes y otros factores de riesgo cardiovascular. Se puede encontrar más información sobre el fichero en este enlace:
  https://biostat.app.vumc.org/wiki/pub/Main/DataSets/diabetes.html
- Carga el conjunto de datos en un data.frame de R llamado `chlstrl`.
- Empezaremos por información básica sobre el conjunto de datos. Cuántas observaciones contiene, cuáles son las variables y de qué tipos,...
- Asegúrate de comprobar si hay datos ausentes y localízalos en la tabla.
- El análisis exploratorio (numérico y gráfico) debe cubrir todos los tipos de variable de la tabla. Es decir, que al menos debes estudiar una variable por cada tipo de variable presente en la tabla. El análisis debe contener, al menos:

  - Para las variables cuantittativas (continuas o discretas).
    Resumen numérico básico.
    Gráficas (las adecuadas, a ser posible más de un tipo de gráfico).

  - Variables categóricas (factores).
    Tablas de frecuencia (absolutas y relativas).
    Gráficas (diagrama de barras).

- Los valores de `height` y `weight` están en pulgadas (inches) y libras (pounds) respectivamente. Una libra son ≈ 0.454kg y una pulgada son ≈ 0.0254m. Usa dplyr para convertir esas columnas a metros y kilogramos respectivamente. Las nuevas columnas deben llamarse igual que las originales.
- Ahora usa esos valores de `height` y `weight` para añadir una nueva columna llamada BMI, definida mediante:
$$BMI = \frac{weight}{height^2}$$
(se divide por el cuadrado de la altura).
- Crea una nueva columna llamada `ageGroup` dividiendo la edad en los siguientes tres niveles:

  `(10,40], (40,70], (70,100]`

- Usando `dplyr` calcula cuántas observaciones hay en cada nivel de `ageGroup` (indicación: usa `group_by`). Ahora, usando aquellas observaciones que corresponden a mujeres, ¿cuál es la media del nivel de colesterol y de BMI en cada uno de esos grupos de edad?

Ejercicio 2: Funciones de R.

- Crea una función de R llamada `cambiosSigno` que dado un vector `x` de números enteros no nulos, como

  `-12, -19, 9, -13, -14, -17, 8, -19, -14,`

  calcule cuántos cambios de signo ha habido. Es decir, cuántas veces el signo de un elemento es distinto del signo del elemento previo. Por ejemplo, en el vector anterior hay 4 cambios de signo (en las posiciones 3, 4, 7 y 8).
- Modifica la función para que devuelva como resultado las posiciones donde hay cambios de signo. Llama `cambiosSignoPos(x)` a esa otra función. Por ejemplo, para el vector anterior el resultado de esta función sería [1] 3 4 7 8

  También se valorará que incluyas en el código como usar `sample` para generar vectores aleatorios de 20 enteros *no nulos* (el vector debe poder tomar valores positivos y negativos).

Ejercicio 3. R4DS.

Es recomendable que esta semana del curso hagas al menos una lectura somera de los Capítulos 1 a 5 de R for Data Science (R4DS), de H. Wickham, con énfasis especial en los Capítulos 3 y 5 (los capítulos 1, 2 y 4 son muy breves). Los siguientes apartados pretenden motivar esa lectura y por eso mismo pueden resultar un poco más laboriosos.

- Haz el ejercicio 6 de la Sección 3.6.1 de R4DS.
- Haz el ejercicio 1 de la Sección 5.2.4 de R4DS.