Boletín Tema 2.

Tratamiento de Datos. Grado en Ciencia de Datos- UV

Alicia ESteban

2022-02-16

1. Un repositorio de datos muy utilizado es *UCI Machine Learning Repository* https://archive.ics.uci.edu/ml.

    a. Examina el repositorio y busca la información disponible de conjunto de datos **Wine Quality Dataset**. Indica el número de variables (*attributes*) y cuántos registros (*Instances*) contiene.
    b. Relaciona la información disponible (Metadata-codebook) con un Codebook ideal. ¿De que elementos, indicados en el *codebook ideal* dispones ?

2. El conjunto de datos **Wine Quality Dataset** está dividido en varios ficheros. Un fichero de texto *winequality.names* y dos ficheros con formato **CSV**. Uno para los vinos blancos y otro para los tintos. Usa la función **download.file**, descarga el fichero de texto y cárgalo realizando una lectura línea a línea. ¿Cuántas y cuáles son las variables medidas?. ¿Es posible localizar las variables automáticamente ?

3. Este ejercicio es una continuación de los anteriores, donde trabajamos con el mismo conjunto de datos. Realiza las siguientes tareas:

    a. Identifica la URL para el vino tinto, descárgalo y almacénalo en un un fichero llamado **UCIwineQualityRed.csv**.
    b. Usa **read.csv** e importa el fichero **UCIwineQualityRed.csv** en un data frame. Examina la estructura del data frame y comenta si los resultados son los esperados. ¿Cuántas variables tiene el conjunto ?
    c. Repite el apartado anterior utilizando **read.csv2** y observa la estructura del nuevo data frame. Consulta la ayuda del **read.csv** y averigua a qué se debe la diferencia.
    d. Importa el fichero con la herramienta de importación automática (**Import Dataset −> From Text (base)**), y observa el código generado. Copia dicho código en tu programa y realiza las modificiones adecuadas para que use una ruta relativa al proyecto para acceder a los datos.
    e. Importa el fichero con la herramienta de importación automática (**Import Dataset −> From Text (readr)**), y mira los valores de la variable *total sulfur dioxide* en las filas 1296 y 1297. ¿Coinciden con los valores del conjunto original ?
    f. El problema se ha producido ya que al importar, se ha elegido un tipo de dato *integer* para la columna *total sulfur dioxide* cuando realmente se trata de una variable no entera. La forma correcta sería utilizar `total sulfur dioxide` = col_double(). Otra alternativa es realizar la importación como cadenas de texto y posteriorme asignar el tipo de dato más adecuado.

(CON LA LIBRERÍA READR SE INTERPRETA CORRECTAMENTE SI SE DEJA EN MODO AUTOMÁTICO. SI LO ELEGIMOS MANUALMENTE A PARTIR DE LOS DATOS Y SELECCIONAMOS TIPO ENTERO FALLARIA. ES NECESARIO TENER INFORMACIÓN DE LOS VALORES DE LAS VARIABLES Y ESTE DATO NO ESTÁ DISPONIBLE EN EL CODEBOOK)

2. Realiza la importación de los ficheros **FileCodificado1.csv**, **FileCodificado2_Latin1.csv**. En primer lugar determina cuál es el tipo de codificación más probable. Analiza qué ocurre con los datos importados si no averiguas la codificación previamente. Observa los caracteres acentuados, ñ, etc. que aparecen en el fichero original.

3. Realiza los capítulos 1, 2 y 3 del curso **Importing Data in R (Part 1)**

4. En este ejercicio veremos el procedimiento para conectarnos a una base de datos remota. Esta base de datos se utiliza en el curso DataCamp **Importing Data in R (part2)**. Es necesario tener instalada la librería **RMySQL**. El siguiente código muestra como establecer la conexión con una base de datos llamada **tweater** que se encuentra en un host remoto **courses.csrrinzqubik.us-east-1.rds.amazonaws.com** y disponemos el puerto, usuario y contraseña para poder acceder.

a. Establece la conexión y determina el número de tablas que contiene (funcion ‘dbListTables‘).

a. Muestra la estructura de cada una de las tablas (función ‘dbListFields‘).

a. Puedes descargar una tabla con la instrucción **dbReadTable**. Descarga cada una de las tabla y alma

a. ¿Qué ocurriría si tuvieses una base de datos con múltiples tablas y **millones** de registros?.

Solución: Hacer una consulta **SQL** a la base de datos y descargar únicamente los registros necesarios

2. Importa el fichero de datos **ERCA.xls** y lee la información del codebook que se ha proporcionado (fichero **CODEBOOK ERCA.docx**. ¿Cuántas etapas crees que son necesarias para obtener un data frame adecuado ?

3. La información enviada por un gps se ha almacenado en el fichero `UNIFICADO.txt`. La información relativa al formato de datos de importación se ha extraído de http://aprs.gids.nl/nmea/#gga. Visualiza el fichero de datos con un editor de texto ¿Cómo crees que se podría importar este fichero ? (No se pide que lo importes)