Trabajo Tema 01: Programación en R

Fundamentos de Estadística y Programación en R (Diploma de Especialización en Data Science y Big Data)

Pedro L. Luque

19 de febrero de 2016

En este trabajo se va a trabajar con datos procedentes de la liga de béisbol de los Estados Unidos recopilados por Sean Lahman desde 1871 a nuestros días. Descargue el fichero que aparece como `2014 - comma-delimited version - Updated January 24, 2015` (`lahman-csv_2015-01-24.zip`) en la dirección http://seanlahman.com/baseball-archive/statistics/.

Cómo entregar el trabajo: Cree un fichero con código R en el que se obtenga respuesta a cada uno de los siguientes ejercicios que se plantean, y envíeme **antes del día 1 de abril de 2016** por email a `calvo@us.es` como adjunto (en un fichero comprimido) todo el material necesario para que se puedan reproducir todos sus cálculos y los ficheros resultantes. Nota: también podría ser un fichero Rmarkdown (Rmd).

Ejercicio 1

Cargue los ficheros: Master.csv y Batting.csv en los objetos R: `master` (datos de jugadores) y `bateos` (información sobre el juego de estos jugadores), con al menos dos métodos distintos (utilice funciones de paquetes distintos).

Apartado (a)

Extraiga los nombres de las variables que contienen los dos ficheros.

Apartado (b)

Muestre las primeras 6 filas de los dos objetos R creados.

Apartado (c)

Cree un data.frame que contenga solamente las siguientes variables del objeto `master` (llámelo: `master2`):

```
"lahmanID"     "playerID"     "birthYear"
"birthCountry" "birthState"
"deathYear"    "deathCountry" "deathState"
"nameFirst"    "nameLast"
"nameNick"     "weight"       "height"
"bats"         "throws"      "debut"
"finalGame"    "college"
```

Apartado (d)

¿De cuántos países distintos hay jugadores de béisbol? Muestre el peso (`weight`) y la altura (`height`) de los jugadores de "W.Germany". Represente esos puntos con la ayuda de `plot`.

Apartado (e)

¿Cuántos jugadores son de "England" y tienen un peso mayor que 180 (libras)?

Apartado (f)

Seleccione 200 jugadores al azar y calcule el siguiente índice:

$$Indice = \frac{peso}{altura^2} * 100$$

Guarde los datos en un fichero Excel `datos200.xlsx`.

Apartado (g)

¿De qué clase R es la variable `nameFirst`? Conviértela a clase `character`. Recodifique en la variable `nameFirst` (en `master2`) para que en lugar de `Charlie` aparezca `Carlos`. ¿Cuántos jugadores tienen como `nameFirst` el valor Carlos ahora? ¿Y antes?

Ejercicio 2

Combine las dos data.frame en un único data.frame (llámelo `todos`) uniéndolos por la variable que los relaciona `playerID`.

Apartado (a)

Guarde los 2000 primeros registros de `todos` en un fichero csv.

Ejercicio 3

Cree una función que calcule el momento de orden k de una variable, que por defecto calcule el momento de orden 2 (**nota**: elimine en la función los datos faltantes o `NA` que puediera tener la variable). Utilícela para calcular el momento de orden 2, 3 y 4, de las variables peso y altura de todos los jugadores, y de la variable `RBI` pero únicamente de los jugadores nacidos en `USA` con `yearID` igual a 2008 (usa la función `subset`).

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(x - \bar{x})^k$$

Ejercicio 4

NOTA: Añada este cuarto ejercicio libre con algún tipo de manipulación sobre estos datos que le resulte de interés.