A1: Preprocesamiento de los datos

7 de marzo de 2018

Introducción

Los datos a tratar corresponden a la informació del **World Happiness Report** del año 2016 que muestra una serie de variables asociadas a la felicidad en distintos países del mundo. El fichero se denomina **2016_raw.csv**, y contiene 157 registros y 13 variables. Las variables del fichero son: Country, Region, Happiness.Rank, Happiness.Score, Lower.Confidence.Interval, Upper.Confidence.Interval, GDP.per.Capita, Family, Life.Expectancy, Freedom, Government.Corruption, Generosity, Dystopia.Residual

Preprocesamiento de datos

El objectivo concreto de esta actividad es preparar el fichero para su posterior análisis. Para guiar la actividad, se listan a continuación qué tipos de preproceso se deben realizar para resolver esta actividad satisfactoriamente:

1. Cargar el fichero de datos en R. Antes de cargar el fichero, se debe inspeccionar qué tipo de formato csv es para realizar la lectura apropiada.
2. Cambiar los nombres de las variables que son muy largos por otros más cortos (al final del documento se especifica cómo).
3. Indicar el tipo de variable estadística de cada variable.
4. En el caso en que R no haya asignado el tipo apropiado a una variable, realizar la conversión necesaria para que el tipo final de cada variable sea el adecuado.
5. Corregir errores de variables cuantitativas con confusión de separador decimal.
6. Normalizar/Estandardizar variables cualitativas.
7. Revisar posibles inconsistencias entre variables.
    i. Lower.Confidence.Interval vs Upper.Confidence.Interval
    ii. Happiness.Rank vs Happiness.Score
8. Buscar valores atípicos en las variables cuantitativas
    i. Presentar un boxplot para cada variable cuantitativa.
    ii. Realizar un cuadro con las estimaciones robustas y no robustas de tendencia central y dispersión de cada variable cuantitativa.
9. Valores perdidos.
    i. Buscar qué variables y registros tienen valores perdidos.
    ii. Imputar los valores a partir de los k-vecinos más próximos usando la distancia de Gower con la información de les 6 últimas variables.
10. Finalmente, realizar un breve estudio descriptivo de los datos una vez depurados y crear el fichero de datos corregido.

Cuando se realiza un preprocesado, el analista decide como estandardizar/normalizar las variables. Para ello, establece unos criterios, que idealmente se escriben para homogeneizar versiones o para posteriores preprocesados que aparezcan sobre nuevos datos A continuación, se indican los criterios a seguir:

- El punto (.) es el separador decimal de cualquier variable numérica.

- Los nombres de las variables largos son los que están formados por varias palabras separadas por un punto, por ejemplo "Happiness.Score" o "Lower.Confidence.Interval". El criterio para cambiar el nombre de la variable es escoger la primera letra de cada palabra. Por ejemplo, la variable "Happiness.Score" queda como "HS", la variable "Lower.Confidence.Interval" queda com "LCI", ... El resto de variables cuyo nombre tenga una sola palabra, se pueden quedar igual (con el mismo nombre).
- Las variables cualitativas se estandarizan en inglés con la primera letra de cada palabra en mayúscula y el resto en minúsuculas. Por ejempplo: Albania, Angola, Latin America and Caribbean, ... En el caso que el texto contenga la palabra "and" se debe tener en cuenta que debe escribirse en minúsculas.
- En caso de inconsistencias entre las variables Lower.Confidence.Interval y Upper.Confidence.Interval, es decir, que el valor de Lower.Confidence.Interval sea mayor que Upper.Confidence.Interval, se intercambiarán los valores entre estas dos variables.
- La variable Happiness.Rank indica la posición del Happiness.Score para cada país. El caso de inconsistencia entre las variables Happiness.Rank vs Happiness.Score sucede cuando el valor de posición de Happiness.Rank no coincide con la ordenación que se realiza en Happiness.Score. La variable correcta es el Happiness.Score y por tanto, será necesario rectificar los valors erróneos de Happiness.Rank con la información sobre la ordenación de Happiness.Score.

Comentarios importantes

1. **No** se puede inspeccionar ni corregir de manera manual el archivo de datos. Por ejemplo, **no** se puede hacer una asignación del tipo:

   ```
   `mydata[1,5] <- 32.5`
   ```

   Este tipo de preprocesos se deben realizar con funcionalidades de búsqueda (buscar los registros donde hay errores o inconsistencias) y luego, hacer las correcciones necesarias aplicando las funciones adecuadas de R. De esta forma, el procedimiento de limpieza o de corrección es correcto, independientemente del fichero de datos y de la posición y valores concretos que éste contenga. O dicho de otro modo, si cambian los valores del fichero de datos, el código de preprocesado es igualmente válido.
2. Recordar que **no** se pueden realizar listados completos de los datos en la solución. El motivo es que se generan ficheros de salida con centenares de páginas que son muy difícil de trazar y corregir. Para comprobar las funcionalidades del código, podéis usar head y tail que solo muestran unas líneas del fichero.

Puntuaciones de los apartados

- Apartados 1 a 3 (10%)
- Apartados de 4 a 6 (20%)
- Apartado 7 (10%)
- Apartado 8 (20%)
- Apartado 9 (10%)
- Apartado 10 (20%)
- Calidad del informe dinámico (10%)