Autores:Jordi Real & Ramon Puig

Títulos

formatGe: Librería de R para preparar y depurar datos para su posterior anàlisis. **Format**ear y **Ge**nerar variables

formatGe: Librería de R para preparar y depurar datos durante el preproceso de datos (**Format**ear y **Ge**nerar variables)

formatGe: Librería de R asistente para preparar y depurar datos durante el preproceso de datos (**Format**ear y **Ge**nerar variables).

formatGe: Soporte adaptativo y funcional durante el preproceso de datos mediante R

formatGe: Mejora la calidad y eficiencia de tus análisis de datos con un asistente adaptativo y funcional mediante R

Resumen:

La mayoría de estadísticos o científicos de datos consume la mayor parte de su tiempo (sobre el 80%) limpiando, depurando y preparando la base datos antes de que ésta sea analizable. Así, la mayor parte del trabajo consiste en preparar los datos, en lugar de llevar a cabo los análisis.

La continua depuración de los datos, los cambios de criterios, y las actualizaciones de datos son motivos que obligan a reanalizar los mismos datos una y otra vez, y convierte algunos análisis en tareas interminables.

El software estadístico R de libre acceso se ha popularizado en entornos de investigación médica por parte de analistas de datos gracias a su gran flexibilidad, la gran cantidad de paquetes estadísticos existentes y la facilidad de funcionalizar tareas.

En este sentido hemos creado una librería de R con funciones específicas dirigidas hacia las primeras fases de un análisis, como son formatear y generar variables (formatGe).

Las funciones incluidas en formatGe pivotan sobre un sistema de análisis de datos basado en conductores. Un conductor es un fichero de datos estructurado que contiene información relativa a cada variable y análisis. Este fichero incluye variables y parámetros: tales como etiquetas de variables, niveles de valor, rango de valores válidos, puntos de corte de recodificaciones, criterios de exclusión etc. Las funciones de formatGe usan el conductor (análisis conductual) como vector de parámetros para ejecutar sus procesos. A su vez, el conductor es un fichero externo e independiente al código del análisis, editable por parte de cualquier miembro del equipo investigador y puede tener varios formatos (txt, csv, spss, Excel, rds, etc.)

FormarGe incluye funciones que permiten limpiar, recodificar, etiquetar, categorizar, y generar formulas reciclables para otros paquetes existentes de R. La característica de formatGe es que todas las funciones se aplican sobre el conjunto de datos según la información que contiene el conductor, adhoc para cada análisis.

La librería formatGe de R facilita los análisis, ya que agiliza la depuración de datos y minimiza los errores, y por lo tanto se convierte en una herramienta muy eficiente sobretodo en análisis de datos complejos y/o bases de datos con gran cantidad de información (ya sea en número de variables o de registros).

Además el uso de formatGe previene los errores habituales, permite tener un mayor control del análisis a medida que lo vamos ampliando y facilita los reanálisis que a menudo requiere cualquier investigación cuantitativa.

Ref:

(<https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#310ccbd26f63>

Most data scientists spend only 20 percent of their time on actual data analysis and 80 percent of their time finding, cleaning, and reorganizing huge amounts of data, which is an inefficient data strategy (https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html)

<https://www.reddit.com/r/datascience/comments/bupmyf/data_scientists_spend_up_to_80_of_time_on_data/>