Categorizar

**Introducción**

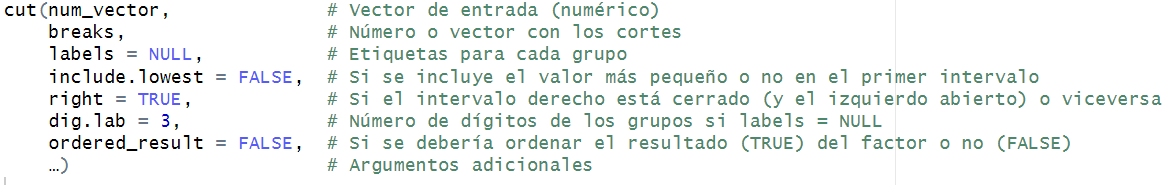
En R, categorizar se refiere al proceso de convertir datos en un formato categórico, es decir, agrupar valores en categorías o niveles definidos. Esto permite que el análisis y procesamiento de los datos trate estas categorías como **grupos discretos** en lugar de valores individuales, facilitando tareas como análisis estadístico, visualización o modelado. En términos técnicos**, implica convertir datos en un tipo de dato llamado factor,** que representa variables categóricas de manera eficiente.

**La función cut para categorizar variables**

La función cut permite cortar datos en grupos y especificar ‘etiquetas de corte’, por lo que es **muy útil para crear un factor a partir de una variable continua**.

En ocasiones resulta útil categorizar los valores de una variable continua en diferentes niveles de un factor. Para ese propósito puedes usar la función cut de R.

**Sintaxis**:



**El parámetro break:**

El argumento breaks permite cortar los datos en subgrupos y, por lo tanto, clasificarlos.

Ejemplo:

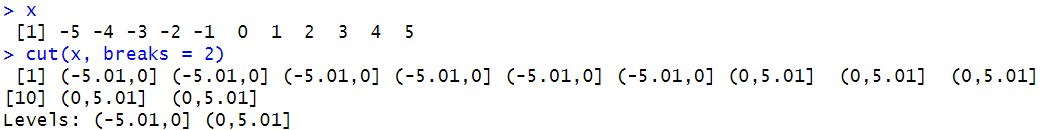
Vamos a dividir el rango de valores de **x en dos intervalos** (o niveles) de igual ancho, con base en el mínimo y máximo de los valores.

x <- -5:5

cut(x, breaks = 2)

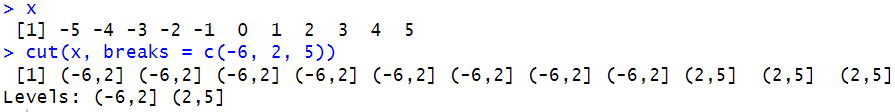
El resultado será algo:

* Un intervalo de [-5, 0) (de -5 hasta 0, excluyendo 0)
* Otro intervalo de [0, 5] (de 0 hasta 5, incluyendo 5)



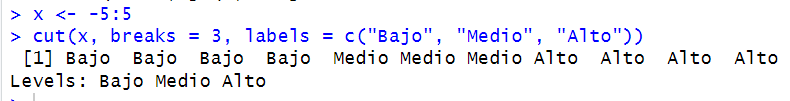
Los límites (como -5.01 y 5.01) ocurre porque R asegura que todos los valores del vector x caigan dentro de los intervalos, incluso si están en el borde. Esto es un comportamiento estándar de cut().

Vamos a dividir en intervalos específicos -6, 2, 5:



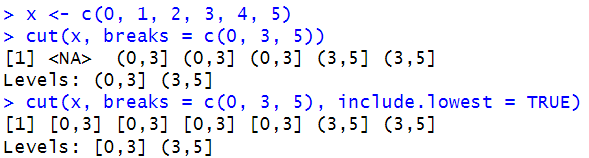
**El parámetro label:**

El parámetro label sirve para reemplazar las etiquetas por defecto (los rangos numéricos) con nombres personalizados, facilitando la interpretación de los datos categorizados.

****

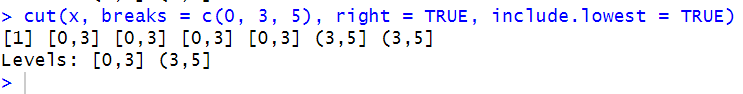
**El parámetro include.lowest:**

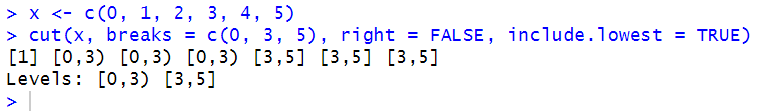
El parámetro include.lowest asegura que el valor más bajo del rango de datos se incluya en el primer intervalo.



**El parámetro right:**

El parámetro right controla si los intervalos son cerrados a la derecha (right = TRUE) o a la izquierda (right = FALSE). Cambiar este parámetro afecta qué valores caen en cada intervalo, especialmente los que están en los bordes de los límites.





**Ejemplo: ¿Cómo categorizar grupos de edad en R?**

Considera que deseas clasificar un vector numérico de edades en las siguientes categorías:

* 0-14: Niño.
* 15-24: Joven.
* 25-64: Adulto.
* 65 y más: Mayor.

Datos de ejemplo:

edad <- c(0, 12, 89, 14, 25, 2, 65, 1, 16, 24, 67, 61, 64)

**Ejemplo: ¿Cómo categorizar notas de exámenes en R?**

Las notas de exámenes son variables continuas que se pueden clasificar como suspensas, si la nota es inferior a 5 puntos sobre 10 o aprobadas en el otro caso. Generaremos un conjunto de datos muy sencillo para clasificar las calificaciones de los exámenes.

Datos de ejemplo:

notas <- c(0, 6.1, 5.3, 8.9, 5.0, 8.8, 1.9, 6.6, 7.2, 9.4, 4.9,

7.1, 3.9, 1.0, 9.3, 9.9, 5.9, 5.1, 8.4, 3.2, 10.0)

Categorías:

* Suspenso: 0 al 5 (sin incluir)
* Aprobado: 5 al 10 (incluido)