Material Imprimible

Curso de Análisis de datos con R

Módulo III

Contenidos:

- Librería dplyr y comando pipe
- Select
- Filter
- arrange
- mutate
- joins
- group by
- summarise

Librería dplyr y comando pipe

La librería dplyr en R ofrece una serie de funciones fundamentales para la manipulación de datos. Contiene funciones y métodos para limpiar, filtrar y agrupar datos en dataframes.

La combinación de estas funciones en secuencias (pipes) con el operador %>% facilita la escritura de código legible y expresivo, ofreciendo un enfoque eficiente para trabajar con datos en R.

Select

La función select se utiliza para elegir columnas específicas de un data frame.

Se escribe:

select()

Ejemplo: df <- select(data_frame, columna1, columna2)

Filter

Filtra filas de un data frame según condiciones específicas.

Se escribe:

filter()

Ejemplo: df_filtrado <- filter(data_frame, columna > 10)

arrange

Ordena filas de un data frame según una o más columnas.

Se escribe:

arrange()

df_ordenado <- arrange(data_frame, columna1, desc(columna2))

mutate

Agrega nuevas columnas o modifica las existentes mediante operaciones.

Se escribe:

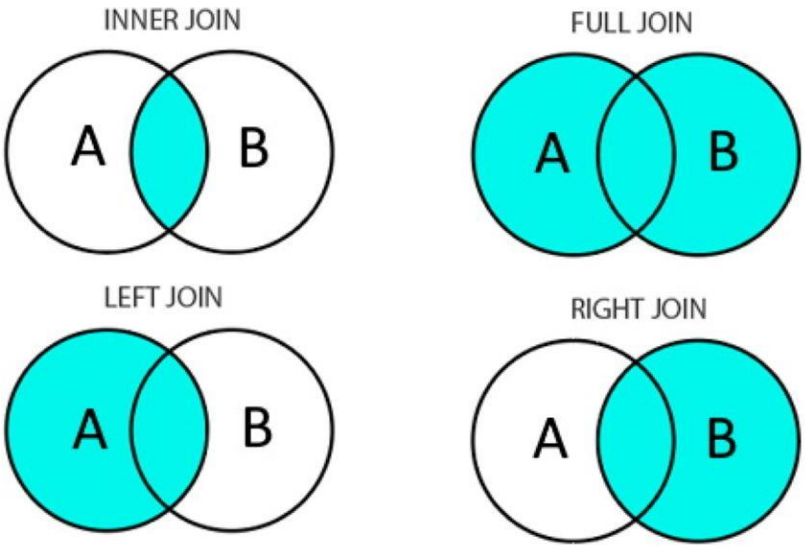
mutate()

Ejemplo: df_mutate <- mutate(data_frame, nueva_columna = columna1 + 10)

Joins

Diferentes funciones (inner_join(), left_join(), right_join(), full_join()) realizan combinaciones entre data frames.

Tipos de Join:



Los tipos de joins conceptualmente son similares a Python y SQL.

En teoría de los conjuntos, dado dos conjuntos, si se realiza un inner join se quedan todos los registros (observaciones en el caso de R) donde las claves crucen. Si se realiza un left o right join, se prioriza como "fuerte" uno de los conjuntos y se devuelven todos los registros del mismo, según el tipo de join elegido, crucen o no, con la entidad débil. Finalmente, un Full Join devuelve todos los registros existentes en ambas entidades crucen o no.

Es fundamental siempre definir la clave, que es una variable que se utiliza para unir ambos set de datos. Sin ella, no es posible realizar ningún tipo de join

Ejemplo: df_joineado <- left_join(data_frame1, data_frame2, by = "clave")

Group BY

Agrupa el dataframe por una o mas columnas, reduciendo su granularidad.
Sintaxis: group_by()

Summarise

Permite realizar agregaciones sobre el dataset agrupado, como el promedio o la suma.
Sintaxis: summarise()
La combinación de estas dos últimas funciones permite armar data frames agregados por variables con valores calculados resumidamente.

```
df_agrupado <- data_frame %>%
 group_by(columna_texto)  %>%
 summarise(promedio  = mean(columna_numerica))
```

En este caso, generamos un nuevo dataframe agrupando mediante una columna de texto y para cada valor categórico, calculamos el promedio a traves de la función “mean” aplicado sobre una columna numérica.

Las funciones que puede realizar summarise, siempre sobre columnas con números enteros o decimales, son las siguientes:

- sum(): Suma de los valores en una columna.
- mean(): Media aritmética de una columna.
- min(): Valor mínimo en una columna.
- max(): Valor máximo en una columna.
- median(): Mediana de una columna.
- n(): Número total de observaciones en un grupo.
- sd(): Desviación estándar de una columna.
- var(): Varianza de una columna.

Algunas las iremos usando a lo largo de los siguientes módulos.