Unidad 04:
Preparación de datos

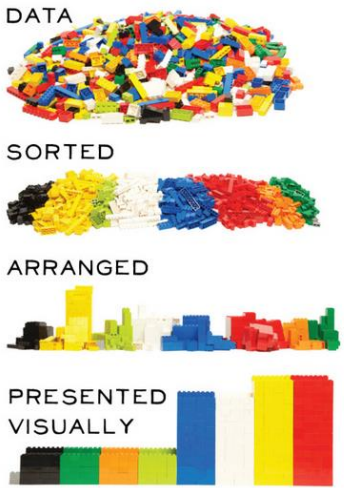
Mg. Jesús Eduardo Gamboa Unsihuay

Octubre del 2024

Introducción

Introducción

- Datos = diamante en bruto
- ***Garbage in, garbage out***
- Tareas de preparación o preprocesamiento:
  - Limpieza
  - Transformación
  - Integración
- Haremos uso de los paquetes **`dplyr`** y **`magrittr`**. Este último permite añadir el operador pipe **`%>%`** (Control + Shift + M en RStudio)
- No obstante, también es posible utilizar el pipe nativo **`|>`**



Funcionamiento del pipe

El operador pipe permite que dispongamos los elementos de una función de manera opuesta, permitiendo la concatenación. Por ejemplo:

```
> x = c(2,4,5,3)
> sum(x)
[1] 14
> x %>% sum()
[1] 14
> x |> sum()
[1] 14
```

Funcionamiento del pipe

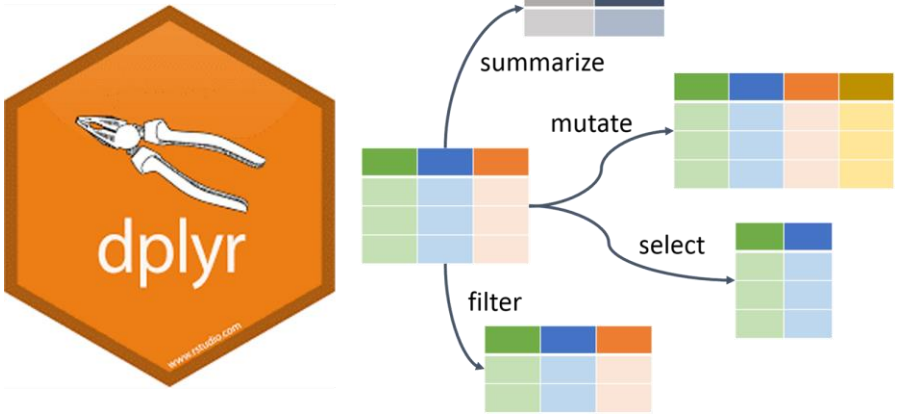
En general f(x) es expresado como x %>% f.

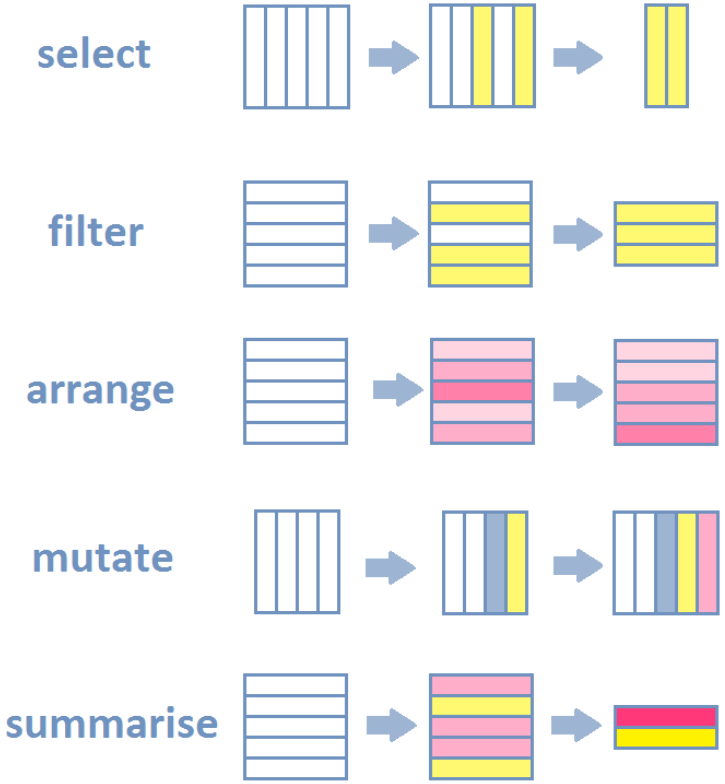
Tiene más utilidad cuando se usa una serie de funciones anidadas, por ejemplo: `h(g(f(x)))` es equivalente a `x %>% f %>% g %>% h` o `x %>% f() %>% g() %>% h()`

A partir de la versión 4.1.0, existe un nuevo pipe en R: |>

Es decir `h(g(f(x)))` es equivalente a `x |> f() |> g() |> h()`

Paquete dplyr





Selección de variables

Selección de variables

Función: select

Paquete: dplyr

¿Cómo se usa?

```
df |> select(posiciones o nombres de las variables a ser seleccionadas)

df |> select(-posiciones o nombres de las variables a ser eliminadas)
```

Selección de variables

Función: pull

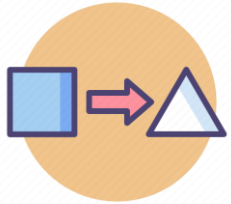
Paquete: dplyr

¿Cómo se usa?

```
df |> pull(posición o nombre de la variable a ser seleccionada)
```

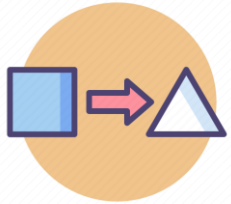
Transformación
de datos

Transformación de datos



En ocasiones, algunos datos pueden estar expresados en una escala que no corresponde a nuestros requerimientos o deseamos realizar una conversión conveniente, por ejemplo: extraer el mes de una fecha.

Transformación de datos



Función: mutate

Paquete: dplyr

¿Cómo se usa?

```
df |> mutate(nueva_variable = regla para crear la nueva variable)

df |> mutate(variable_existente = regla para modificar la variable)
```

Renombramiento
de variables

Renombramiento de variables

Función: rename

Paquete: dplyr

¿Cómo se usa?

```
df |> rename(nuevo_nombre = posición de la columna)
```

```
df |> rename(nuevo_nombre = antiguo_nombre)
```

Filtro de datos

Filtro de datos

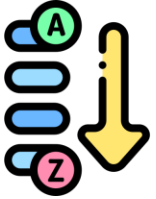
Función: filter

Paquete: dplyr

¿Cómo se usa?

```
df |> filter(reglas de filtro)
```

Ordenamiento
de datos



Ordenamiento de datos

Función: arrange

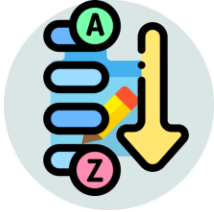
Paquete: dplyr

¿Cómo se usa?

```
df |> arrange(variable) # orden alfabético A - Z, o de menor a mayor

df |> arrange(desc(variable)) # orden alfabético Z - A, o de mayor a menor

df |> arrange(-variable numérica) # orden de mayor a menor
```



Ordenamiento de datos

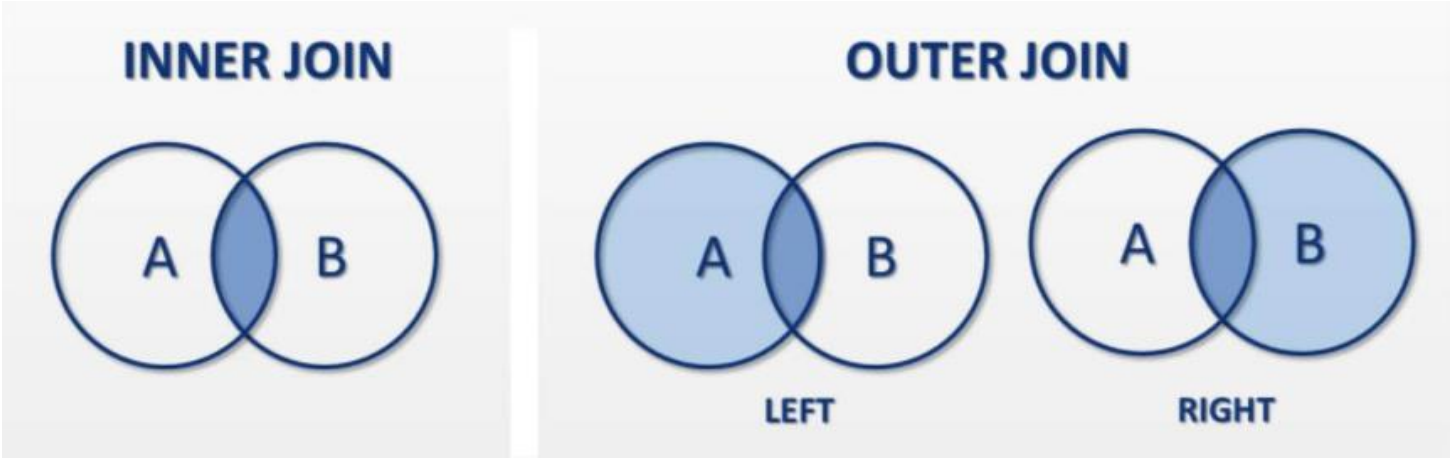
Usando datos21:

a) Ordenar por CÓDIGO, de manera alfabética, de la A a la Z
b) Ordenar por CÓDIGO, de manera alfabética, de la Z a la A
c) Ordenar por HABITACIONES, de menor a mayor
d) Ordenar por HABITACIONES, de mayor a menor

Integración
de datos

Integración de datos

En ocasiones, nuestros datos no están presentes en un único data frame y necesitamos juntar dos o más de estos.



Funciones para integración

- **`inner_join(x,y)`** retorna las filas de x que tienen valores coincidentes en y, considerando todas las columnas de x e y
- **`right_join(x,y)`** retorna las filas de y, además de todas las columnas de x e y. Completa con NA cuando es necesario.
- **`left_join(x,y)`** retorna las filas de x, además de todas las columnas de x e y. Completa con NA cuando es necesario.

Funciones para integración

```
inner_join(x,y)

x %>% inner_join(y)

x |> inner_join(y)
```

Aplicación

El archivo **04 - DATOS - 03.xlsx** contiene datos acerca de empleados de una empresa, su domicilio y el área a la que pertenecen (hoja Empleados), así como datos de cada área (hoja Areas).

1. Lea cada hoja del archivo como empleados y areas
2. Se desea saber qué tan distante es el distrito de domicilio con el distrito de trabajo de cada empleado, además del jefe de cada uno(a) de ellos(as). Enlace estos conjuntos de datos haciendo uso de inner_join, left_join, right_join, full_join.

En resumen

Funciones del paquete **dplyr** para manipulación de datos:

- rename
- select
- pull
- filter
- arrange
- mutate
- inner_join, full_join, right_join, left_join
- y varias más disponibles aquí