BeduTech

Programación orientada a objetos

BIG DATA

Sílvia Ariza Sentís

Agenda

BeduTech

15 min. Introducción
30 min. Tipos de datos
30 min. Funciones básicas
30 min. Importar ficheros
30 min. Dplyr
15 min. Introducción

Calentamiento

Tipos de datos

Numeric

Integer

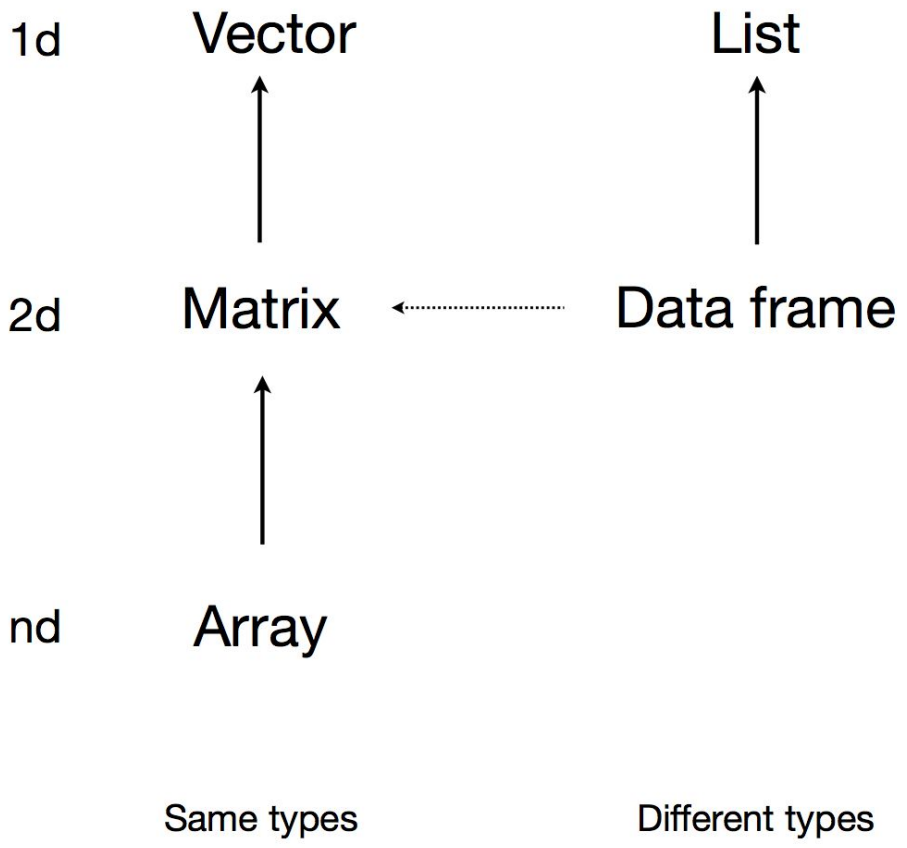
Complex

Logical

Character

Matrix

DF



Sintaxis

BeduTech

- () para usar funciones sobre una variable (i.e. colnames(mi_matriz))
- [] para hacer extracciones sobre una variable (i.e. mi_vector[2:5])
- <- Definir variables y guardarlas hasta que reiniciemos sesión
- class() función para conocer la clase de la variable
- colnames() función para conocer los nombres de las columnas (+ rownames)
- str función para conocer la estructura de la tabla
- ncol() función para conocer el número de columnas (+ nrow)
- $ sirve para observar una variable dentro de la tabla (i.e. tabla$columna1)

Importar ficheros

- setwd Definir el directorio (set working directory) buscando la ruta donde se encuentra el fichero dentro de nuestra laptop

- Definimos la tabla para que quede guardada en la memoria y hacer operaciones con ella posteriormente (<-)

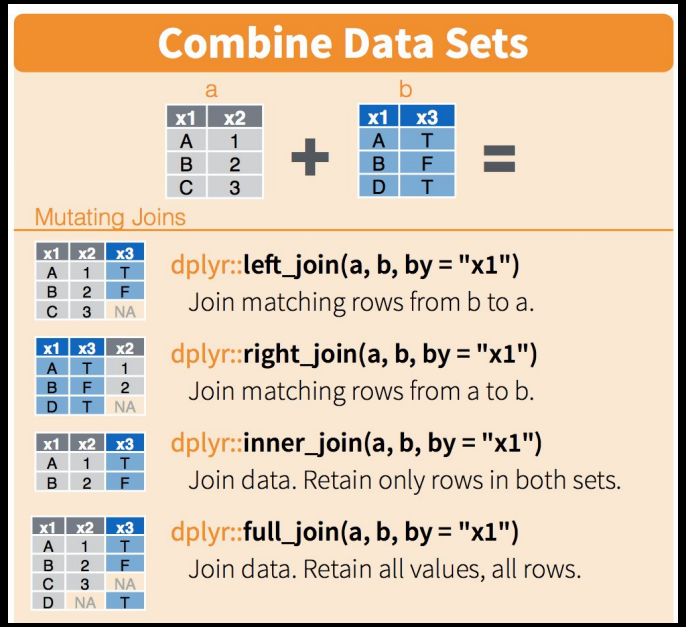
- Leemos el fichero con la función read.csv e indicamos si hay nombres en las columnas (HEADER = TRUE)

BeduTech

```
> dir.create('DataMaster')
> dir()
[1] "baltimore.csv" "DataMaster"
> setwd("./DataMaster")
> getwd()
[1] "C:/Users/Guest/Desktop/R/DataMaster"
> setwd("../")
> getwd()
[1] "C:/Users/Guest/Desktop/R"
> setwd("../")
> getwd()
[1] "C:/Users/Guest/Desktop"
> setwd('./R')
> getwd()
[1] "C:/Users/Guest/Desktop/R"
```

dplyr

El package dplyr sirve para agrupar y seleccionar datos de manera rápida (funciones similares a SQL)



Funciones:

- Select (i.e. Datos, -datos, starts_with,
- Rename
- Filter
- & | %in% and, or, in equivalentes en SQL
- Summarise resumen de la data según media, mediana, etc.
- Arrange ordenar los datos
- %>% separador de funciones

Actividad

Ejemplo 01

- Usar dplyr
- Seleccionar y filtrar información a través de dplyr.

Cognitive process:

1. Take the **ydat** dataset, ***then***
2. **filter()** for genes in the leucine biosynthesis pathway, ***then***
3. **group_by()** the limiting nutrient, ***then***
4. **summarize()** to correlate rate and expression, ***then***
5. **mutate()** to round *r* to two digits, ***then***
6. **arrange()** by rounded correlation coefficients

The old way:

```
arrange(
  mutate(
    summarize(
      group_by(
        filter(ydat, bp=="leucine biosynthesis"),
      nutrient),
    r=cor(rate, expression)),
  r=round(r, 2)),
r)
```

The dplyr way:

```
ydat %>%
  filter(bp=="leucine biosynthesis") %>%
  group_by(nutrient) %>%
  summarize(r=cor(rate, expression)) %>%
  mutate(r=round(r,2)) %>%
  arrange(r)
```

Revisa Repositorio

Actividad

Reto 01

BeduTech

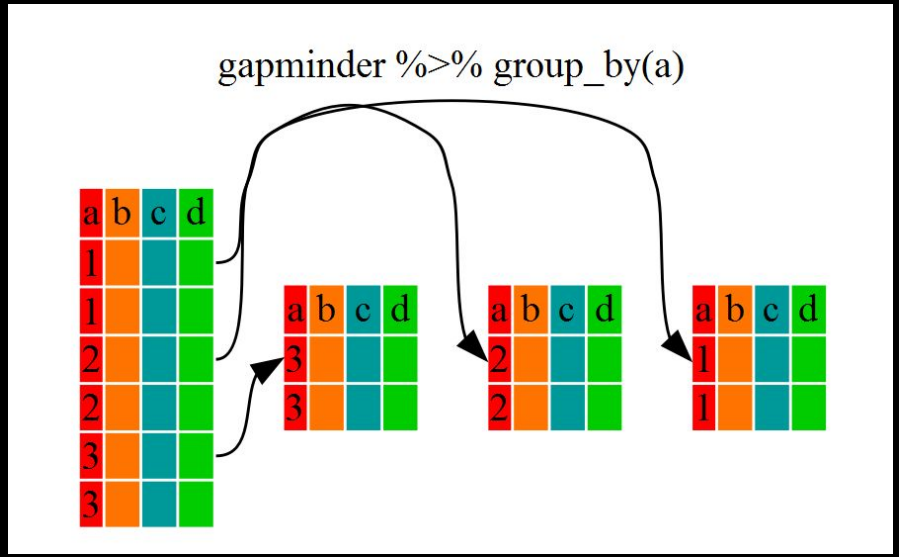
Importamos la tabla de Ecobici con la que hemos estado trabajando (hint: read_excel).

Revisa Repositorio

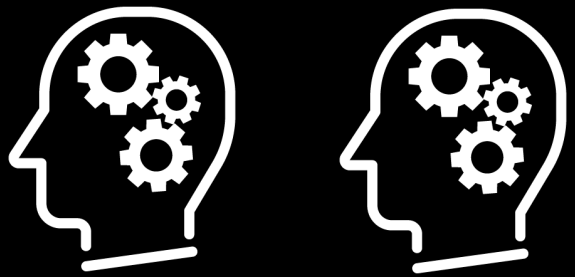
Actividad

Ejemplo 02

- Con dplyr crearemos una orden compleja en un solo comando de código.



Revisa Repositorio



Actividad

Reto 02

Sobre la tabla de ecobici que ya estamos trabajando, crearemos un comando optimizado de varias órdenes apoyándote de la función %>%.

Revisa Repositorio

Preguntas


BEDU