

Introducción

Entornos de Análisis de Datos: R

Alberto Torres Barrán

2020-02-06

Introducción

- R es un lenguaje de programación y un entorno para manipular datos, realizar cálculos y gráficos.
- Herramienta muy popular para tareas de Data Science (junto con Python)
- Comparado con herramientas clásicas (Excel, SaS, SPSS)
 - Más flexible
 - Curva de aprendizaje inclinada
 - **Librerías!**

Librerías

- R tiene una colección de más de 12000 librerías o paquetes de terceros
- La mayoría disponibles en un repositorio centralizado (CRAN)
- No forman parte del núcleo de R (R base)
- Se pueden instalar muy fácilmente

Entorno

- R está disponible para los principales sistemas operativos (Windows, Linux, MacOS):
 - <http://cran.r-project.org>
- Recomendado el uso del IDE RStudio
 - <http://www.rstudio.com>

Tidyverse

- Colección de paquetes diseñados para tareas de Data Science
- No son estrictamente necesarios, pero simplifican las tareas más comunes
- Los principales son: `dplyr`, `ggplot2`, `tidyr`, `readr`, `purrr`, `stringr`, `forcats` y `tibble`

Instalar y cargar librerías

- Podemos instalar nuevas librerías con la sentencia:

```
install.packages("tidyverse")
```

- Para usar las librerías tenemos que cargarlas en el entorno:

```
library(tidyverse)
```

- También se puede hacer de forma gráfica en RStudio

Operador de asignación

- El resultado de cualquier sentencia de R se pierde si no se asigna a una nueva variable

```
# este resultado se muestra y se pierde  
2 + 2  
## [1] 4
```

- El operador de asignación es `<-`

```
# el resultado de la operación se almacena en una nueva variable  
# `suma`  
suma <- 2 + 2
```

Data frames

- Tabla para almacenar datos en R
- Está compuesto por observaciones (filas) y variables (columnas)
- Cada variable puede ser de un tipo distinto (texto, categórica, numérica, etc.)
- Todas las observaciones de una misma variable tienen que ser del mismo tipo
- Cada variable tiene un nombre

Funciones

- Construcción de R que toma unos argumentos de entrada, realiza un cálculo y devuelve un resultado
- Elemento básico de cualquier lenguaje de programación
- Ejemplos de llamadas a funciones:

```
v <- c(4.6, 8.2, 9.2)
mean(v)
## [1] 7.333333
```

```
# siempre con parentesis, aunque no tengan ningun argumento
ls()
## [1] "best_vars"          "bs"                "bs_coef"
## [4] "bs_fit"             "bs_mse"            "bs_pred"
## [7] "bs_summ"           "center"            "coef_df"
## [10] "data"               "data_MB"           "data_num"
## [13] "data_red"          "df"                "df2"
## [16] "en_coef"           "en_fit"            "en_mse"
## [19] "en_pred"           "enet"              "factor_vars"
## [22] "fit"               "i"                 "idx"
## [25] "ind_train"         "input"             "input_dir"
## [28] "intercept"         "k"                 "km"
## [31] "km_def"            "kopt"              "la_coef"
## [34] "la_fit"            "la_mse"            "la_pred"
## [37] "lm_coef"           "lm_fit"            "lm_mse"
## [40] "lm_pred"           "load_image_file"   "load_label_file"
## [43] "logreg"            "model1"            "mse_list"
## [46] "ncomp_opt"         "numbers"           "nyc"
## [49] "output"            "output_dir"        "p"
## [52] "pfit"              "plot_decision_boundary" "pls_coef"
## [55] "pls_fit"           "pls_mse"           "pls_pred"
## [58] "preds"             "proj_pca"          "repeats"
```

Referencias y ayuda

- La referencia principal del curso es el libro "**R for Data Science**" de Hadley Wickham y Garret Grolemund (O'Reilly 2017)
- Tiene una versión online gratuita
- Hadley Wickham es además el creador de muchos de los paquetes que componen el **tidyverse**
- Acceder a la ayuda de R:

```
?mean  
help(mean)
```

Funciones de data frames

Número de filas

```
nrow(mpg)  
## [1] 234
```

Número de columnas

```
ncol(mpg)  
## [1] 11
```

Nombres de las columnas

```
colnames(mpg)
## [1] "manufacturer" "model"      "displ"      "year"      "cyl"
## [6] "trans"        "drv"        "cty"        "hwy"        "fl"
## [11] "class"
```

Primeras líneas

```
head(mpg)
## # A tibble: 6 x 11
##   manufacturer model displ year   cyl trans      drv   cty   hwy fl  class
##   <chr>         <chr> <dbl> <int> <int> <chr>    <chr> <int> <int> <chr> <chr>
## 1 audi         a4     1.8  1999     4 auto(l5) f      18    29 p   compa
## 2 audi         a4     1.8  1999     4 manual(m5) f      21    29 p   compa
## 3 audi         a4     2    2008     4 manual(m6) f      20    31 p   compa
## 4 audi         a4     2    2008     4 auto(av) f      21    30 p   compa
## 5 audi         a4     2.8  1999     6 auto(l5) f      16    26 p   compa
## 6 audi         a4     2.8  1999     6 manual(m5) f      18    26 p   compa
```

str

Estructura del data frame

```
str(mpg)
## Classes 'tbl_df', 'tbl' and 'data.frame':  234 obs. of  11 variables:
## $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
## $ model       : chr  "a4" "a4" "a4" "a4" ...
## $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
## $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
## $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
## $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
## $ drv         : chr  "f" "f" "f" "f" ...
## $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
## $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
## $ fl         : chr  "p" "p" "p" "p" ...
## $ class       : chr  "compact" "compact" "compact" "compact" ...
```

summary

Estadísticas de las variables

```
summary(mpg)
## manufacturer      model      displ      year      cyl
## Length:234      Length:234      Min.    :1.600      Min.    :1999      Min.    :4
## Class :character  Class :character      1st Qu.:2.400      1st Qu.:1999      1st Qu.:4
## Mode  :character  Mode  :character      Median :3.300      Median :2004      Median :6
##                                     Mean  :3.472      Mean  :2004      Mean  :5
##                                     3rd Qu.:4.600      3rd Qu.:2008      3rd Qu.:8
##                                     Max.   :7.000      Max.   :2008      Max.   :8
## trans      drv      cty      hwy
## Length:234  Length:234  Min.    : 9.00  Min.    :12.00
## Class :character  Class :character  1st Qu.:14.00  1st Qu.:18.00
## Mode  :character  Mode  :character  Median :17.00  Median :24.00
##                                     Mean  :16.86  Mean  :23.44
##                                     3rd Qu.:19.00  3rd Qu.:27.00
##                                     Max.   :35.00  Max.   :44.00
## f1      class
## Length:234  Length:234
## Class :character  Class :character
## Mode  :character  Mode  :character
##
##
##
```