

STA 326 2.0 R Programming and Data Analysis

Thiyanga S Talagala

2020-02-02

Contents

1	R environment	5
1.1	Getting help	5
1.2	Data permanency and removing objects	5
2	Introduction	7
3	Data structures in base R	9
3.1	Atomic vectors	9
3.2	Matrix	14
3.3	Array	16
3.4	List	16
3.5	Data frame	16
4	Functions in R programming	21
4.1	In-built functions	21
4.2	User-defined functions	21
4.3	Some useful built-in functions in R	22
5	Writing functions	23
5.1	When should we write functions?	23
5.2	Glogal variables vs local variables	23
5.3	Control structures	23
6	Data analysis with tidyverse	25
7	Data wrangling	27
8	Data visualisation	29

Learning outcomes functions:

In this tutorial we learned what functions in R programming are, the basic syntax of functions in R programming, in-built functions and how to use them to make our work easier, the syntax of a user-defined function, and different types of user-defined functions. In the next session, we are going to learn how

to read files in R programming.

Chapter 1

R environment

1.1 Getting help

1.2 Data permanency and removing objects

Chapter 2

Introduction

The entities R operates on are technically known as **objects**. There are two types of objects:

1. Data structures
2. Functions

Chapter 3

Data structures in base R

There are five data types in R

1. Atomic vector
2. Matrix
3. Array
4. List
5. Data frame

3.1 Atomic vectors

- This is a 1-dimensional
- All elements of an atomic vector must be the same type, Hence it is a **homogeneous** type of object. Vectors can hold numeric data, character data or logical data.

3.1.1 Creating vectors

Vectors can be created by using the function concatenation `c`

Syntax

```
vector_name <- c(element1, element2, element3)
```

Examples

```
first_vec <- c(10, 20, 50, 70)
second_vec <- c("Jan", "Feb", "March", "April")
third_vec <- c(TRUE, FALSE, TRUE, TRUE)
fourth_vec <- c(10L, 20L, 50L, 70L)
```

3.1.2 Types and tests with vectors

1. `typeof()` returns types of their elements

```
typeof(first_vec)
```

```
## [1] "double"
```

```
typeof(fourth_vec)
```

```
## [1] "integer"
```

2. To check if it is a

- vector: `is.vector()`

```
is.vector(first_vec)
```

```
## [1] TRUE
```

- character vector: `is.character()`

```
is.character(first_vec)
```

```
## [1] FALSE
```

- double: `is.double()`

```
is.double(first_vec)
```

```
## [1] TRUE
```

- integer: `is.integer()`

```
is.integer(first_vec)
```

```
## [1] FALSE
```

- logical: `is.logical()`

```
is.logical(first_vec)
```

```
## [1] FALSE
```

- atomic: `is.atomic()`

```
is.atomic(first_vec)
```

```
## [1] TRUE
```

3. `length()` returns number of elements in a vector

```
length(first_vec)
```

```
## [1] 4
length(fourth_vec)
```

```
## [1] 4
```

3.1.3 Coercion

Vectors must be homogeneous. When you attempt to combine different types they will be coerced to the most flexible type so that every element in the vector is of the same type.

Order from least to most flexible

logical -> integer -> double -> character

```
a <- c(3.1, 2L, 3, 4, "GPA")
typeof(a)
```

```
## [1] "character"
```

```
anew <- c(3.1, 2L, 3, 4)
typeof(anew)
```

```
## [1] "double"
```

3.1.4 Explicit coercion

Vectors can be explicitly coerced from one class to another using the functions `as.character`, `as.numeric`, `as.integer`, and `as.logical`.

```
vec1 <- c(TRUE, FALSE, TRUE, TRUE)
typeof(vec1)
```

```
## [1] "logical"
```

```
vec2 <- as.integer(vec1)
typeof(vec2)
```

```
## [1] "integer"
```

```
vec2
```

```
## [1] 1 0 1 1
```

Question

Why the below output produce NAs?

```
x <- c("a", "b", "c")
as.numeric(x)
```

```
## Warning: NAs introduced by coercion
```

```
## [1] NA NA NA
```

3.1.5 Simplifying vector creation

1. colon : produce regular spaced ascending or descending sequences.

```
a1 <- 10:16
a1
```

```
## [1] 10 11 12 13 14 15 16
```

```
b1 <- -0.5:8.5
b1
```

```
## [1] -0.5 0.5 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5
```

2. sequence `seq()`. There are three arguments we need to provide, i) initial value, ii) final value, and iii) increment

syntax

```
seq(initial_value, final_value, increment)
```

example

3. repeats `rep()`

```
rep(9, 5)
```

```
## [1] 9 9 9 9 9
```

```
rep(1:4, 2)
```

```
## [1] 1 2 3 4 1 2 3 4
```

```
rep(1:4, each=2) # each element is repeated twice
```

```
## [1] 1 1 2 2 3 3 4 4
```

```
rep(1:4, times=2) # whole sequence is repeated twice
```

```
## [1] 1 2 3 4 1 2 3 4
```

```
rep(1:4, each=2, times=3)
```

```
## [1] 1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4
```

```
rep(1:4, 1:4)
```

```
## [1] 1 2 2 3 3 3 4 4 4 4
```

```
rep(1:4, c(4, 1, 4, 2))
```

```
## [1] 1 1 1 1 2 3 3 3 3 4 4
```

3.1.6 Logical operations

3.1.7 Subsetting

There are situations where we want to select only some of the elements of a vector. Following codes show various ways to select part of a vector object.

```
data <- c(10, 20, 103, 124, 126)

data[1] # shows the first element

## [1] 10
data[-1] # shows all except the first item

## [1] 20 103 124 126
data[1:3] # shows first three elements

## [1] 10 20 103
data[c(1, 3, 4)]

## [1] 10 103 124
data[data > 3]

## [1] 10 20 103 124 126
data[data<20|data>120]

## [1] 10 124 126
```

Example: How do you replace the 3rd element in the data vector by 203?

```
data[3] <- 203
data

## [1] 10 20 203 124 126
```

3.1.8 Vector arithmetic

Vector operations are performed element by element.

```
c(10, 100, 100) + 2 # two is added to every element in the vector

## [1] 12 102 102
```

Vector operations between two vectors

```
v1 <- c(1, 2, 3)
v2 <- c(10, 100, 1000)
v1 + v2

## [1] 11 102 1003
```

Add two vectors of unequal length

```
longvec <- seq(10, 100, length=10)
shortvec <- c(1, 2, 3, 4, 5)

shortvec+longvec
```

```
## [1] 11 22 33 44 55 61 72 83 94 105
```

3.1.9 Missing values

Use NA to place a missing value in a vector.

```
z <- c(10, 101, 2, 3, NA)
is.na(z)
```

```
## [1] FALSE FALSE FALSE FALSE TRUE
```

3.2 Matrix

Matrix is a 2-dimensional and a homogeneous data structure

Syntax to create a matrix

```
matrix_name <- matrix(vector_of_elements,
                       nrow=number_of_rows,
                       ncol=number_of_columns,
                       byrow=logical_value, # If byrow=TRUE, then the matrix is filled by rows
                       dimnames=list(rnames, cnames)) # To assign row names and column names
```

Example

```
values <- c(10, 20, 30, 40)
matrix1 <- matrix(values, nrow=2) # Matrix filled by columns (default option)
matrix1
```

```
##      [,1] [,2]
## [1,]  10  30
## [2,]  20  40
```

```
matrix2 <- matrix(values, nrow=2, byrow=TRUE) # Matrix filled by rows
matrix2
```

```
##      [,1] [,2]
## [1,]  10  20
## [2,]  30  40
```

Naming matrix rows and columns

```
rnames <- c("R1", "R2")
cnames <- c("C1", "C2")
```

```
matrix_with_names <- matrix(values, nrow=2, dimnames=list(rnames, cnames))
matrix_with_names
```

```
##      C1 C2
## R1 10 30
## R2 20 40
```

3.2.1 Matrix subscript

`matraix_name[i,]` gives the *i*th row of a matrix

```
matrix1[1, ]
```

```
## [1] 10 30
```

`matraix_name[, j]` gives the *j*th column of a matrix

```
matrix1[, 2]
```

```
## [1] 30 40
```

`matraix_name[i, j]` gives the *i*th row and *j*th column element

```
matrix1[1, 2]
```

```
## [1] 30
```

```
matrix1[1, c(1, 2)]
```

```
## [1] 10 30
```

3.2.2 cbind and rbind

Matrices can be created by column-binding and row-binding with `cbind()` and `rbind()`

```
x <- 1:3
y <- c(10, 100, 1000)

cbind(x, y) # binds matrices horizontally
```

```
##      x      y
## [1,] 1     10
## [2,] 2    100
## [3,] 3   1000
```

```
rbind(x, y) # binds matrices vertically
```

```
##      [,1] [,2] [,3]
## x      1    2    3
## y     10   100 1000
```

3.2.3 Matrix operations

3.3 Array

- 3 dimensional data structure
-

3.4 List

3.5 Data frame

- A data frame is more general than a matrix in that different columns can contain different modes of data.
- It's similar to the datasets you'd typically see in SPSS and MINITAB.
- Data frames are the most common data structure you'll deal with in R.

3.5.1 Creating a dataframe

Syntax

```
name_of_the_dataframe <- data.frame(
  var1_name=vector of values of the first variable,
  var2_names=vector of values of the second variable)
```

Example

```
corona <- data.frame(ID=c("C001", "C002", "C003", "C004"),
  Location=c("Beijing", "Wuhan", "Shanghai", "Beijing"),
  Test_Results=c(FALSE, TRUE, FALSE, FALSE))
corona
```

```
##      ID Location Test_Results
## 1 C001  Beijing      FALSE
## 2 C002   Wuhan       TRUE
## 3 C003 Shanghai      FALSE
## 4 C004  Beijing      FALSE
```

To check if it is a dataframe

```
is.data.frame(corona)
```

```
## [1] TRUE
```

To convert a matrix to a dataframe

```
mat <- matrix(10:81, ncol=4)
mat
```



```
##      [,1] [,2] [,3] [,4]
## [1,]  10  28  46  64
## [2,]  11  29  47  65
## [3,]  12  30  48  66
## [4,]  13  31  49  67
## [5,]  14  32  50  68
## [6,]  15  33  51  69
## [7,]  16  34  52  70
## [8,]  17  35  53  71
## [9,]  18  36  54  72
## [10,] 19  37  55  73
## [11,] 20  38  56  74
## [12,] 21  39  57  75
## [13,] 22  40  58  76
## [14,] 23  41  59  77
## [15,] 24  42  60  78
## [16,] 25  43  61  79
## [17,] 26  44  62  80
## [18,] 27  45  63  81
```

```
mat_df <- as.data.frame(mat)
mat_df
```

```
##    V1 V2 V3 V4
## 1  10 28 46 64
## 2  11 29 47 65
## 3  12 30 48 66
## 4  13 31 49 67
## 5  14 32 50 68
## 6  15 33 51 69
## 7  16 34 52 70
## 8  17 35 53 71
## 9  18 36 54 72
## 10 19 37 55 73
## 11 20 38 56 74
## 12 21 39 57 75
## 13 22 40 58 76
## 14 23 41 59 77
## 15 24 42 60 78
## 16 25 43 61 79
## 17 26 44 62 80
## 18 27 45 63 81
```

3.5.2 Subsetting data frames

Select rows

```
head(mat_df) # default it shows 5 rows
```

```
##   V1 V2 V3 V4
##  1 10 28 46 64
##  2 11 29 47 65
##  3 12 30 48 66
##  4 13 31 49 67
##  5 14 32 50 68
##  6 15 33 51 69
```

```
head(mat_df, 3) # To extract only the first three rows
```

```
##   V1 V2 V3 V4
##  1 10 28 46 64
##  2 11 29 47 65
##  3 12 30 48 66
```

```
tail(mat_df)
```

```
##   V1 V2 V3 V4
## 13 22 40 58 76
## 14 23 41 59 77
## 15 24 42 60 78
## 16 25 43 61 79
## 17 26 44 62 80
## 18 27 45 63 81
```

To select some specific rows

```
index <- c(1, 3, 7, 8)
mat_df[index, ]
```

```
##   V1 V2 V3 V4
##  1 10 28 46 64
##  3 12 30 48 66
##  7 16 34 52 70
##  8 17 35 53 71
```

Select columns

1. Select column(s) by variable names

```
mat_df$V1 # Method 1
```

```
## [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
```

```
mat_df[, "V1"] # Method 2
```

```
## [1] 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
```

2. Select column(s) by index

```
mat_df[, 2]
```

```
## [1] 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
```

3.5.3 Built in dataframes

Note: All objects in R have a class.

Chapter 4

Functions in R programming

A function is a block of organized and reusable code that is used to perform a specific task in a program. There are two types of functions in R:

1. In-built functions
2. User-defined functions

4.1 In-built functions

These functions in R programming are provided by R environment for direct execution, to make our work easier. Some examples for the frequently used in-built functions are as follows.

```
mean(c(10, 20, 21, 78, 105))
```

```
## [1] 46.8
```

4.2 User-defined functions

These functions in R programming language are declared and defined by a user according to the requirements, to perform a specific task.

All R functions have three main components: (Check this with Hadley's book)

1. **function name:** name of the function that is stored as an R object
2. **arguments:** are used to provide specific inputs to a function while a function is invoked. A function can have zero, single, multiple or default arguments.

3. **function body:** contains the block of code that performs the specific task assigned to a function. **return value**

4.3 Some useful built-in functions in R

4.3.1 Maths functions

Operator	Description
abs(x)	absolute value of x
log(x, base=y)	logarithm of x with base y; if base is not specified, returns the natural logarithm
exp(x)	exponential of x
sqrt(x)	square root of x
factorial(x)	factorial of x

4.3.2 Basic statistic functions

Operator	Description
mean(x)	mean of x
median(x)	median of x
mode(x)	mode of x
var(x)	variance of x
scale(x)	z-score of x
quantile(x)	quantiles of x
summary(x)	summary of x: mean, minimum, maximum, etc.

4.3.3 Probability distribution functions

Chapter 5

Writing functions

5.1 When should we write functions?

- do many repetitive task

5.2 Global variables vs local variables

5.3 Control structures

Chapter 6

Data analysis with tidyverse

Some *significant* applications are demonstrated in this chapter.

Chapter 7

Data wrangling

Chapter 8

Data visualisation