R4DS

Hennie Bezuidenhout

03 October 2016

# Programming in R Studio - Pipes

## The point of the **Pipe** is to help us read our code in an easy to understand way!!

### magrittr package - but all packages in tidyverse automatiically make

When are we better off not using the *Pipe* ?

1. When **Pipes** are longer than ten steps (Rather create intermediate objects with meaningful names)
2. We have multiple *inputs* or *outputs* (the same solution as above applies)
3. When we are starting to think about a *directed graph* with a complex dependency structure

Remember that the%T>%can assist us with more complex *Pipes* -Effectively, the **T-Pipe** returns the left-hand side instead of the right-hand side incase we want to plot or print

# Introducing Functions

## Allows us to automate common tasks in a more powerful and general way than copy-and-pasting



We do not copy and paste

What are the advantages of using *functions* ?

* You can give a function an evocative name that makes your code easier to understand.
* As requirements change, you only need to update code in one place, instead of many.
* You eliminate the chance of making incidental mistakes when you copy and paste (i.e. updating a variable name in one place, but not in another)

##### We should seriously consider writing a function whenever we have copied and pasted a block of code more than **TWICE**

There are 3 steps to creating a new *function* ?

* You need to pick a name for the function
* You list the inputs, or arguments, to the function inside function
* You place the code you have developed in **body** of the function

rescale01 <- function(x) { rng <- range(x, na.rm = TRUE) (x - rng[1]) / (rng[2] - rng[1]) }

Most important attribute of functions: \*\* WE DO NOT WANT TO REPEAT OURSELVES\*\*

## Conditional Execution

An if statement allows us to conditionally execute code

if (condition) {  
# code executed when the condition is \*\*TRUE\*\*  
}  
else {  
# code executed when the condition is \*\*FALSE\*\*  
}

The condition must evaluate to TRUE or FALSE

Both if and function should (almost) always be followed up by squiggly brackets, except for very short easy to understand if statements

## Function Arguments

The arguments to a function typically fall into two broad sets: one set supplies the **DATA** to compute on, and the other supplies arguments that control the ***details*** of the computation

Generally, data arguments should come first, and detail arguments should go on the end *and usually should have default arguments*

##### Simple Example

mean\_ci <- function(x, conf = 0.95) { >se <- sd(x) / sqrt(length(x)) **(This is DATA)** alpha <- 1 - conf **(This is DATA)** mean(x) + se \* qnorm(c(alpha / 2, 1 - alpha / 2)) (This is detailed arguments to manupilate the DATA) }

# Vectors

There are 2 main types of Vectors

* **Atomic Vectors** (Logical, Integer, Double, Character, Complex and Raw)
* **Lists**

*I feel that in dlyr we will work primarily with Tibbles which automatically recognizes types of Vectors upon reading the Files*

# Iteration

Iteration addresses duplication.Why ?

* It is easier to see the intend of your code
* It is easier to respond to changes in requirements
* We are likely to have fewer bugs because each line of code is used in more places

**Iteration** helps us to do the same thing to multiple inputs: Repeating the same operation on different columns, or on different datasets.

*How will we be able to get the mean for each of the below columns ?*

library(knitr)  
df <- tibble::tibble(  
 a = rnorm(10),  
 b = rnorm(10),  
 c = rnorm(10),  
 d = rnorm(10)  
)  
kable(df)

|  |  |  |  |
| --- | --- | --- | --- |
| a | b | c | d |
| 0.6480840 | -1.0759851 | -0.0488765 | -0.9593299 |
| 0.3658974 | -0.7635152 | 0.6618515 | 0.7839848 |
| -0.1851962 | -1.9057835 | 0.9438019 | 1.4197782 |
| -0.5965279 | 0.5352681 | 0.3972675 | 1.5685785 |
| 2.0848867 | 1.4367584 | -2.6283927 | 0.6639166 |
| -0.0192297 | 1.3360714 | -0.8160732 | -0.8467633 |
| -0.1460706 | -0.3852539 | -0.2934559 | -0.4970061 |
| 0.6786517 | 1.2863697 | -0.3818876 | -1.8779094 |
| 0.5483603 | -0.7965556 | 0.5271912 | -0.6459500 |
| -1.7332035 | 0.6575525 | -0.4546370 | -0.9441542 |

output <- vector("double", ncol(df)) # 1. output  
for (i in seq\_along(df)) { # 2. sequence  
 output[[i]] <- median(df[[i]]) # 3. body  
}  
output  
#> [1] -0.2458 -0.2873 -0.0567 0.1443

Every loop has 3 components:

* The OUTPUT:output <- vector("double", ncol(df)) Before we start the loop, we must always allocate sufficient space for the output (This increases the speed at which the loop will be able to operate)
* The SEQUENCE: for (i in seq\_along(df)). This determines what to loop over. Each rum of the for loop will assign *i* to a different value from seq\_along(df)
* the BODY: output[[i]] <- median(df[[i] This is the code that does the work. It runs repeatedly, each time with a different value for *i*. The first iteration will run output[[1]] <- median(df[[1]]), the second will run output[[2]] <- median(df[[2]]), and so on.

##### Exercise Exercise Exercise Exercise