2. Base R - Subsetting

CT5102 - J. Duggan (University of Galway)

Subsetting

R's subsetting operators are fast and powerful. Mastering them allows you to succinctly perform complex operations in a way that few other languages can match.
— Hadley Wickham "Advanced R"

Overview

- Subsetting operations allow you to process data stored in atomic vectors and lists
- R provides a range of flexible approaches that can be used to subset data.
- Learning outcomes:
  - The four main ways to subset a vector, namely, by positive integer, the minus sign, logical vectors, and vector element names.
  - The role of the `[[` operator for processing lists, and how to distinguish this from the `[` operator.
  - The use of the `$` operator
  - How to use the `for` loop structure to iterate through a list
  - How to use the `if` statement when processing lists, and how that statement differe from the `ifelse()` function covered in Chapter Chapter @ref(ch2).
  - Additional R functions that allow you to process vectors
  - How to solve all three test exercises.

Atomic Vectors

- Let's model the daily number of customers arriving at a restaurant as a *Poisson distribution*, with a mean (λ) of one hundred customers per day (λ=100).
- The Poisson distribution describes a discrete random variable, and the mean and variance are both equal to λ
- In R, the `rpois()` function can be used to generate random numbers from a Poisson distribution, with mean (λ).

```
# set the seed to ensure replication
set.seed(111)
# Generate the count data, assume a Poisson distribution
customers <- rpois(n = 10, lambda = 100)
names(customers) <- paste0("D",1:10)
customers
```

```
##  D1  D2  D3  D4  D5  D6  D7  D8  D9 D10
## 102  96  97  98 101  85  98 118 102  94
```

Subsetting vectors - Positive Integers

- For vectors in R, the operator `[` is used to subset vectors
- Positive integers will subset atomic vector elements at given locations
- To extract the n^{th} item from a vector `x` the term `x[n]` is used
- This can also apply to a sequence, starting at `n` and finising at `m` can be extracted from the vector `x` as `x[n:m]`
- Indices can also be generated using the combine function `c()`, which is then passed in to subset a vector.

```
customers[1]
```

```
##  D1
## 102
```

```
customers[1:5]
```

```
##  D1  D2  D3  D4  D5
## 102  96  97  98 101
```

Subsetting vectors - Negative Integers

- Negative integers, expressed as a vector, can be used to exclude elements from a vector
- One or more elements can be excluded

```
customers[-1]
```

```
##  D2  D3  D4  D5  D6  D7  D8  D9 D10
##  96  97  98 101  85  98 118 102  94
```

```
customers[-c(1,length(customers))]
```

```
##  D2  D3  D4  D5  D6  D7  D8  D9
##  96  97  98 101  85  98 118 102
```

```
customers[-(2:(length(customers)-1))]
```

```
##  D1 D10
## 102  94
```

Subsetting vectors - Logical Vectors

- Logical vectors can be used to subset a vector
- This allows for the use of relational and logical operators.
- when a logical vector is used to subset a vector, only the corresponding cells of the logical vector element that contain TRUE will be retained in the operation.

```
##    D1    D2    D3    D4    D5    D6    D7    D8    D9   D1
##  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALS
```

```
##  D1  D5  D8  D9
## 102 101 118 102
```

- The two statements can be combined into the one expression.

```
(customers[customers > 100])
```

```
##  D1  D5  D8  D9
## 102 101 118 102
```

Recycling

- A nice feature of subsetting with logical vectors is that the logical vector size does not have to equal the size of the target vector.
- When the length of the logical vector is less than the target vector, R will *recycle* the logical vector by repeating the sequence of values until all the target values have been subsetted.

```
# Subset every third element from the vector
customers[c(TRUE,FALSE,FALSE)]
```

```
##  D1  D4  D7 D10
## 102  98  98  94
```

Subsetting vectors - By Element Names

- If a vector has named elements - usually set via the function `names()`, then elements can subsetted through their name.
- This is convenient if you want to retrieve an element but do not necessarily want to know its exact indexed location.

```
customers
```

```
##  D1  D2  D3  D4  D5  D6  D7  D8  D9 D10
## 102  96  97  98 101  85  98 118 102  94
```

```
# Show the value from day 10
customers["D10"]
```

```
## D10
##  94
```

```
customers[c("D1","D10")]
```

```
##  D1 D10
```

Subsetting Lists

In a similar manner to our exploration of atomic vectors, we first generate a simulated manufacturing data for two products, A and B.

```
# A small products database. Main list has two products
products <- list(
              A=list(product="A",
                     sales=12000,
                     quarterly=list(quarter=1:4,
                                    sales=c(6000,3000,2000,1000))),
              B=list(product="B",
                     sales=8000,
                     quarterly=list(quarter=1:4,
                                    sales=c(2500,1500,2800,1200))))
```

Exploring the list structure

```
str(products)
```

```
## List of 2
##  $ A:List of 3
##   ..$ product  : chr "A"
##   ..$ sales    : num 12000
##   ..$ quarterly:List of 2
##   .. ..$ quarter: int [1:4] 1 2 3 4
##   .. ..$ sales  : num [1:4] 6000 3000 2000 1000
##  $ B:List of 3
##   ..$ product  : chr "B"
##   ..$ sales    : num 8000
##   ..$ quarterly:List of 2
##   .. ..$ quarter: int [1:4] 1 2 3 4
##   .. ..$ sales  : num [1:4] 2500 1500 2800 1200
```

Visualising the List

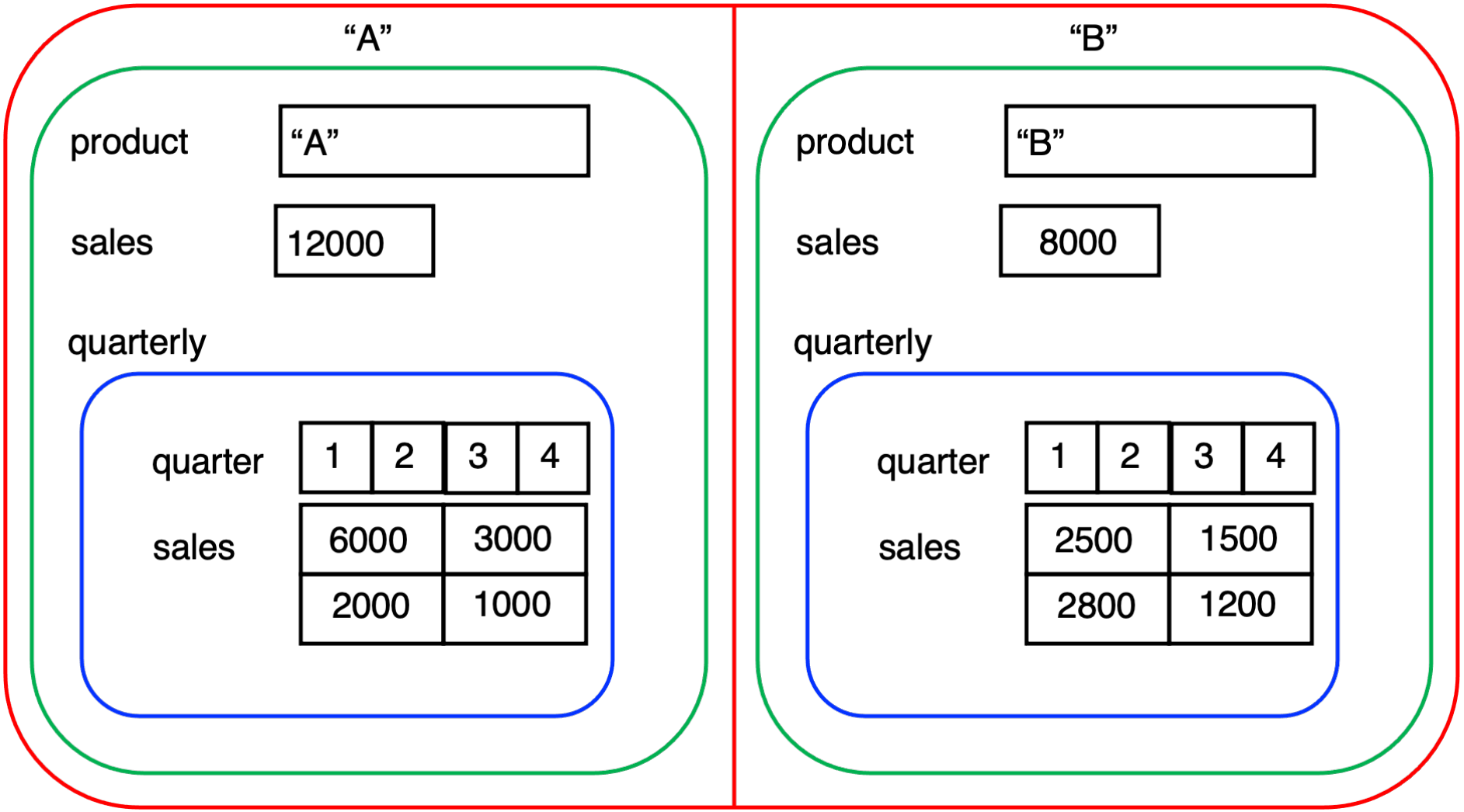


Figure 1: Visualisation of the products data structure

Some observations on the list

- At its core, the list is simply a vector of two named elements, and this is highlighted with the red lines.
- We can check this with R code to show (1) the length of the vector and (2) the name of each element.

```
# Show the vector length (2 elements)
length(products)
```

```
## [1] 2
```

```
# Show the names of each element
names(products)
```

```
## [1] "A" "B"
```

Some observations on the list

- However, even though there are just two elements in the list, each element has a significant internal structure.
- Each element contains a list, highlighted in green. This list contains three elements:
  - the product name, a character atomic vector,
  - the sales, a numeric atomic vector
  - an element named quarterly, which is another list (coloured blue).
- This third list contains two atomic vector elements:
  - the quarter number (1, 2, 3 and 4), and
  - the corresponding sales amount for each quarter, for the product.
- Note that the sum of the `sales` vector in this list equals the amount in the `sales` vector in the previous list.

Subsetting lists

- Subsetting lists is more challenging than subsetting atomic vectors.
- There are three methods that can be used, and to illustrate the core idea we define a list (l1) that contains three named elements.

```
# Create a simple list vector
l1 <- list(a="Hello",b=1:5,c=list(d=c(T,T,F),e="Hello World"))
# Show the structure
str(l1)
```

```
## List of 3
##  $ a: chr "Hello"
##  $ b: int [1:5] 1 2 3 4 5
##  $ c:List of 2
##   ..$ d: logi [1:3] TRUE TRUE FALSE
##   ..$ e: chr "Hello World"
```

Subsetting with [

- The single square bracket '[', when applied to a list, will *always return a list*.
- The same indexing method used for atomic vectors can also be used for filtering lists, namely: positive integers, negative integers, logical vectors, and the element name. Here are examples of filtering a list using each of these methods.

```
# extract the first element of the list
str(l1[1])
```

```
## List of 1
##  $ a: chr "Hello"
```

```
str(l1["a"])
```

```
## List of 1
##  $ a: chr "Hello"
```

Subsetting with [

Some additional examples:

```
# extract the first two list elements
str(l1[1:2])
```

```
## List of 2
##  $ a: chr "Hello"
##  $ b: int [1:5] 1 2 3 4 5
```

```
str(l1[c(T,F)])
```

```
## List of 2
##  $ a: chr "Hello"
##  $ c:List of 2
##   ..$ d: logi [1:3] TRUE TRUE FALSE
##   ..$ e: chr "Hello World"
```

Extracting list contents with [[

- The single bracket [return a list, but in many cases this is not sufficent for analysis
- We will need to access the data within the list (which can be an atomic vector, and also a list).
- For example, finding the value of element a or element b.
- To do this, we must use the '[[' operator, which extracts the *contents of a list* at a given location (i.e. element 1, 2, .., N), where N is the list length.

```
# extract the contents of the first list element
l1[[1]]
```

```
## [1] "Hello"
```

```
l1[["a"]]
```

```
## [1] "Hello"
```

Extracting list contents with [[

```
# extract the contents of the second  list element
l1[["b"]]
```

```
## [1] 1 2 3 4 5
```

```
# extract the contents of the third list element (a list!)
str(l1[["c"]])
```

```
## List of 2
##  $ d: logi [1:3] TRUE TRUE FALSE
##  $ e: chr "Hello World"
```

```
l1[["c"]][["d"]]
```

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

```
l1[[3]][[1]]
```

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

Extracting list contents with $

- There is a convenient alternative to the [[operator, and this is the tag operator $ which can be used once a list element is named.
- For example, for our list l1 the terms l1[[1]], l1[["a"] and l1$a are the same, and in the general case l[["y]] is equivalent to l$y

```
l1$b
```

```
## [1] 1 2 3 4 5
```

```
# extract the contents of the third list element (a list!)
str(l1$c)
```

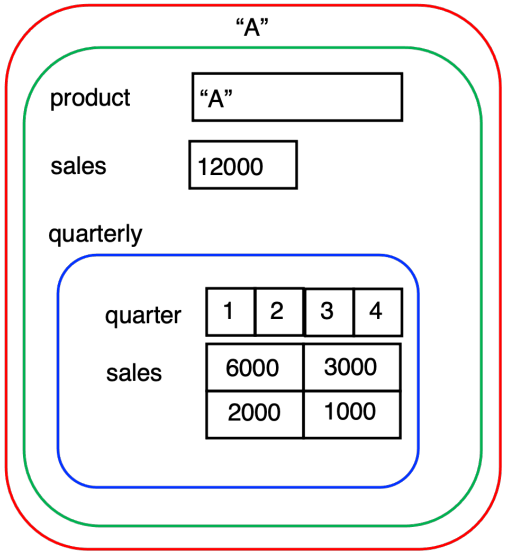
```
## List of 2
##  $ d: logi [1:3] TRUE TRUE FALSE
##  $ e: chr "Hello World"
```

```
l1$c$d
```

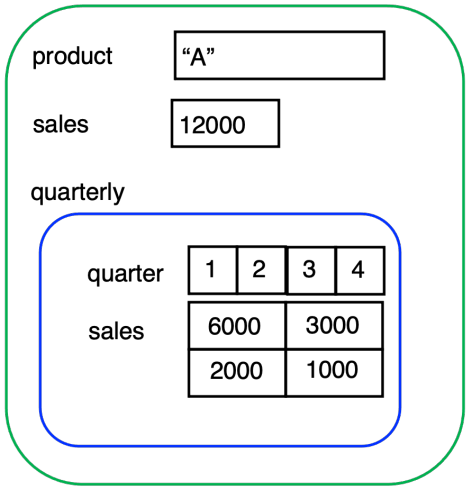
```
## [1]  TRUE  TRUE FALSE
```

Visualising subsets of `products`

(1) products[1] *or* products["A"]

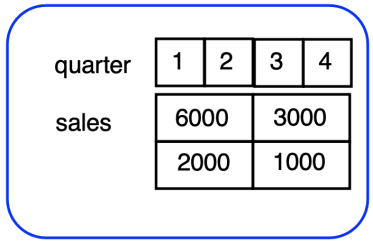


(2) products[[1]] *or* products[["A"]] *or* products$A

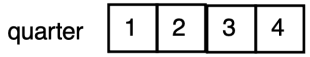


(4) products[[1]][[2]] *or* products[["A"]][["sales"]] *or* products$A$sales

(5) products[[1]][[3]] *or* products[["A"]][["quarterly"]] *or* products$A$quarterly



(6) products[[1]][[3]][[1]] *or* products[["A"]][["quarterly"]][["quarter"]] *or* products$A$quarterly$quarter



(7) products[[1]][[3]][[2]] or products[["A"]][["quarterly"]][["sales"]] *or* products$A$quarterly$sales

(8) products[[1]][[3]][[2]][1:2] *or*

Code for (1) - get the first list element

```
# Example (1) - get the first element of the list as a list
ex1.1 <- products[1]
ex1.2 <- products["A"]
str(ex1.1)
```

```
## List of 1
##  $ A:List of 3
##   ..$ product  : chr "A"
##   ..$ sales    : num 12000
##   ..$ quarterly:List of 2
##   .. ..$ quarter: int [1:4] 1 2 3 4
##   .. ..$ sales  : num [1:4] 6000 3000 2000 1000
```

Code for (2) - get the contents of the first list element

```
# Example (2) - get the contents of the first list element
ex2.1 <- products[[1]]
ex2.2 <- products[["A"]]
ex2.3 <- products$A
str(ex2.1)
```

```
## List of 3
##  $ product  : chr "A"
##  $ sales    : num 12000
##  $ quarterly:List of 2
##   ..$ quarter: int [1:4] 1 2 3 4
##   ..$ sales  : num [1:4] 6000 3000 2000 1000
```

Code for (3) - get the product name from the 1st list element

```
# Example (3) - get the product name for the first product
ex3.1 <- products[[1]][[1]]
ex3.2 <- products[["A"]][["product"]]
ex3.3 <- products$A$product
str(ex3.1)
```

```
##  chr "A"
```

Code for (4) - get annual sales of Product A

```
# Example (4) - get the annual sales for the first product
ex4.1 <- products[[1]][[2]]
ex4.2 <- products[["A"]][["sales"]]
ex4.3 <- products$A$sales
str(ex4.1)
```

```
##  num 12000
```

Code for (5) - get the list for quarterly sales data

```
# Example (5) - get as a list, the detailed quarterly sales
ex5.1 <- products[[1]][[3]]
ex5.2 <- products[["A"]][["quarterly"]]
ex5.3 <- products$A$quarterly
str(ex5.1)
```

```
## List of 2
##  $ quarter: int [1:4] 1 2 3 4
##  $ sales  : num [1:4] 6000 3000 2000 1000
```

Code for (6) - get the quarters vector

```
# Example (6) - get the quarters
ex6.1 <- products[[1]][[3]][[1]]
ex6.1 <- products[["A"]][["quarterly"]][["quarter"]]
ex6.1 <- products$A$quarterly$quarter
str(ex6.1)
```

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

Code for (7) - get the quarterly sales vector

```
# Example (7) - get the quarterly sales
ex7.1 <- products[[1]][[3]][[2]]
ex7.1 <- products[["A"]][["quarterly"]][["sales"]]
ex7.1 <- products$A$quarterly$sales
str(ex7.1)
```

```
##  num [1:4] 6000 3000 2000 1000
```

Code for (8) - subset the quarterly sales for product A

```
# Example (8) - get the quarterly sales for the first two quar
ex8.1 <- products[[1]][[3]][[2]][1:2]
ex8.2 <-products[["A"]][["quarterly"]][["sales"]][1:2]
ex8.3 <-products$A$quarterly$sales[1:2]
str(ex8.1)
```

```
##  num [1:2] 6000 3000
```

Updating and adding new elements

```
# Increase the sales of product A by 10,000
products$A$sales <- products$A$sales + 10000
# Add a new field to product A
products$A$type <- "Food"
str(products$A)
```

```
## List of 4
##  $ product  : chr "A"
##  $ sales    : num 22000
##  $ quarterly:List of 2
##   ..$ quarter: int [1:4] 1 2 3 4
##   ..$ sales  : num [1:4] 6000 3000 2000 1000
##  $ type     : chr "Food"
```

Some observations on list subsetting

- Clearly, for list manipulation, the tag operator is the most programmer-friendly, so it is recommended to use this, and also try and ensure that the list elements are named
- Indexing using `[[` by positive integer is very useful for looping structures, we will see an example of this shortly
- Functions such as those in the package `purrr` provide efficient and flexible ways to iterate through lists.

Iteration using Loops

- Iteration is fundamental to all programming languages, and R is no exception.
- There are a number of basic looping structures than can be used in R, and we will focus on one of these, the `for` loop. The general structure is 'for(var in seq)expr, where:
  - `var` is a name for a variable that will change its value for each loop iteration
  - `seq` is an expression that evaluates to a vector
  - `expr` which is an expression, which can be either a simple expression, or a compound expression of the form `{expr1; expr2}`, which is effectively a number of lines of code with two curly braces.
- A convenient method to iterate over a vector (a list or an atomic vector), is to use the function `seq_along()` which returns the indices of a vector.

Example loop structure (Atomic Vector)

```
set.seed(100)
(v <- sample(1:6,10,replace = T))
```

```
##  [1] 2 6 3 1 2 6 4 6 6 4
```

```
seq_along(v)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10
```

```
n_six <- 0
for(i in seq_along(v)){
  n_six <- n_six + as.integer(v[i] == 6)
}
n_six
```

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

Example loop structure using [[(List)

Our goal is to find the average sales for the two products, and for this we can use a list.

```
sum_sales <- 0

for(i in seq_along(products)){
  cat("Sales of product ",i," = ",products[[i]]$sales,"\n")
  sum_sales <- sum_sales+products[[i]]$sales
}
```

```
## Sales of product  1  =  22000
## Sales of product  2  =  8000
```

```
(avr_sales <- sum_sales / length(products))
```

```
## [1] 15000
```

The if statement

- `if(cond) expr` which evaluates `expr` if the condition `cond` is true
- `if(cond) true.expr else false.expr`, which evaluates `true.expr` if the condition is true, and otherwise evaluates `false.expr`

```
# create a test vector
v <- 1:10
lv <- vector(mode="logical",length(v))
for(i in seq_along(v)){
  if(v[i] > mean(v))
    lv[i] <- TRUE
  else
    lv[i] <- FALSE
}
v[lv]
```

```
## [1]  6  7  8  9 10
```

Mini-Case Star Wars Movies

- The CRAN package `repurrrsive` is used
- It has lists on films (`sw_films`), people (`sw_people`), planets (`sw_planets`), species (`sw_species`), starships (`sw_starships`) and species (`sw_species`).
- Here we focus on the `sw_films` list, which contains seven elements, and each element is a list that contains fourteen elements.

```
library(repurrrsive)
length(sw_films)
```

```
## [1] 7
```

The first movie

```
str(sw_films[1])
```

```
## List of 1
##  $ :List of 14
##   ..$ title         : chr "A New Hope"
##   ..$ episode_id    : int 4
##   ..$ opening_crawl: chr "It is a period of civil war.\r\nR
##   ..$ director      : chr "George Lucas"
##   ..$ producer      : chr "Gary Kurtz, Rick McCallum"
##   ..$ release_date  : chr "1977-05-25"
##   ..$ characters    : chr [1:18] "http://swapi.co/api/people
##   ..$ planets       : chr [1:3] "http://swapi.co/api/planets
##   ..$ starships     : chr [1:8] "http://swapi.co/api/starshi
##   ..$ vehicles      : chr [1:4] "http://swapi.co/api/vehicle
##   ..$ species       : chr [1:5] "http://swapi.co/api/species
##   ..$ created       : chr "2014-12-10T14:23:31.880000Z"
##   ..$ edited        : chr "2015-04-11T09:46:52.774897Z"
```

Movie Directors for 1st and last movie

```
# Get the first film name and movie director
sw_films[[1]][[1]]
```

```
## [1] "A New Hope"
```

```
sw_films[[1]][[4]]
```

```
## [1] "George Lucas"
```

```
# Get the last film name and movie director
sw_films[[length(sw_films)]][[1]]
```

```
## [1] "The Force Awakens"
```

```
sw_films[[length(sw_films)]][[4]]
```

```
## [1] "J. J. Abrams"
```

Movie Directors for 1st and last movie

```
# Get the first film name and movie director
sw_films[[1]]$title
```

```
## [1] "A New Hope"
```

```
sw_films[[1]]$director
```

```
## [1] "George Lucas"
```

```
# Get the last film name and movie director
sw_films[[length(sw_films)]]$title
```

```
## [1] "The Force Awakens"
```

```
sw_films[[length(sw_films)]]$director
```

```
## [1] "J. J. Abrams"
```

Find all movies directed by George Lucas

- A for-loop structure (along with `seq_along()`)is used to iterate over the entire loop and mark those elements as either a match (`TRUE`) or not a match (`FALSE`). This information is stored in an atomic vector.
- Before entering the loop, we create a logical vector variable (`is_target`) of size seven (the same size as the list), and this will store information on whether a list item should be marked for further processing.
- For each list element we extract the directors name and check if it matches the target ("George Lucas"), and store this value in the corresponding element of `is_target`.
- The vector `is_target` can then be used to filter the original `sw_films` list and retain all the movies directed by George Lucas.

Code Solution

```
# Search for movies by George Lucas and store these in a new
target <- "George Lucas"
# Create a logical vector that will hold information for posi
is_target <- vector(mode="logical",length = length(sw_films))
# Iterate through the entire sw_films list (of 7)
for(i in seq_along(sw_films)){
  is_target[i] <- sw_films[[i]]$director == target
}
is_target
```

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

```
target_list <- sw_films[is_target]
length(target_list)
```

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

An additional task

In this case, we look to extract the movie titles into a new data structure, in this case an atomic vector.

```
# Create a movies vector to store the movie names
movies <- vector(mode="character",
                 length = length(target_list))
# Iterate through the list to extract the movie title
for(i in seq_along(target_list)){
  movies[i]<-target_list[[i]]$title
}
movies
```

```
## [1] "A New Hope"              "Attack of the Clones" "The Phan
## [4] "Revenge of the Sith"
```

Creating a new list

- One feature of R you will discover is that there are often may ways to achieve the same outcome.
- For example, another way to access the movies of George Lucas would be to re-arrange the *list of lists* into a single list, where each list element is an atomic vector of values (each of size seven).
- The process for creating the new data structure is:
  - Create a new list (`sw_films1`) of elements you wish to store (for example, movie title, episode_id and director) from the original list. This new list initially contains empty vectors.
  - Loop through the `sw_films` list and append each movie title and director to the corresponding element of `sw_films1`

Sample solution

```
sw_films1 <- list(title=c(), episode_id=c(), director=c())
for(i in seq_along(sw_films)){
  sw_films1$title      <- c(sw_films1$title,
                            sw_films[[i]]$title)
  sw_films1$episode_id <- c(sw_films1$episode_id,
                            sw_films[[i]]$episode_id)
  sw_films1$director   <- c(sw_films1$director,
                            sw_films[[i]]$director)
}
str(sw_films1)
```

```
## List of 3
##  $ title     : chr [1:7] "A New Hope" "Attack of the Clones
##  $ episode_id: int [1:7] 4 2 1 3 6 5 7
##  $ director  : chr [1:7] "George Lucas" "George Lucas" "Geo
```

Parallel Vectors

- Notice that we now have one list, and this list has three elements, each an atomic vector of size seven.
- These can be viewed as *parallel vectors*, where each vector is the same size, and the i-th element of each vector are related.
- This feature can be exploited to filter related atomic vectors using logical vector subsetting.

```
cat(sw_films1$title[1],"-",sw_films1$episode_id[1],"-",
    sw_films1$director[1],"\n")
```

```
## A New Hope - 4 - George Lucas
```

```
sw_films1$title[sw_films1$director=="George Lucas"]
```

```
## [1] "A New Hope"              "Attack of the Clones" "The Phan
## [4] "Revenge of the Sith"
```

Useful R Functions

R Function	Description
`as.list()`	Coerces the input argument into a list.
`paste0()`	Converts arguments to character strings and then concatenates (with no spaces)
`rpois()`	Generates up to n random numbers from a Poisson distribution with mean lambda
`seq_along()`	Generates a regular sequence that can be used to iterate over vectors
`which()`	Give the TRUE indices of a logical object

Exercise 1

- Filter the list `sw_people` (87 elements), contained in `repurrrsive` to include only those whose height is *not unknown*, and name this list `sw_people1`.
- Use an atomic vector `has_height` to filter the list, and populate this vector using a loop structure.
- This new list (`sw_people1`) should have 81 elements.

```
sum(has_height)
```

```
## [1] 81
```

```
length(sw_people1)
```

```
## [1] 81
```

Exercise 2

- Using a `for` loop over the filtered list `sw_people1` from exercise 2, create a list of people whose height is greater than or equal to 225 inches.
- The resulting vector should grow as matches are found, as we do not know in advance how many people will be contained in the result.
- Use the command `characters <- c()` to create the initial empty result vector.
- The `if` expression may be useful here too, and make sure that the height value is converted to a numeric value before evaluating.

The following result should be obtained.

```
# These are the characters whose height is >= 225
characters
```

```
## [1] "Chewbacca"   "Yarael Poof" "Lama Su"     "Tarfful"
```

Exercise 3

- Based on the list `sw_species`, and given that each species has a `classification`, create the following tabular summary, again using a loop to iterate through the list.
- Make use of the `table()` function

```
# A tabular summary of the types of species
t_species
```

```
## c_species
##  amphibian artificial  gastropod  insectoid     mammal     r
##          6          1          1          1         16
##  reptilian   sentient    unknown
##          1          1          6
```

Lecture Summary

- The four main ways to subset a vector, namely, by positive integer, the minus sign, logical vectors, and vector element names.
- The role of the `[[` operator for processing lists, and how to distinguish this from the `[` operator.
- The use of the `$` operator
- How to use the `for` loop structure to iterate through a list
- How to use the `if` statement when processing lists, and how that statement differs from the `ifelse()` function
- Additional R functions that allow you to process vectors
- How to solve all three test exercises.