**Set-up: scraping some data from Wikipedia**

{lubridate} is yet another tidyverse package, that makes dealing with dates or duration data  
(and intervals) as painless as possible. I do not use every function contained in the package  
daily, and as such will only focus on some of the functions. However, if you have to deal with  
dates often, you might want to explore the package thoroughly.

Let’s get some data from a Wikipedia table:

library(tidyverse)

library(rvest)

page <- read\_html("https://en.wikipedia.org/wiki/Decolonisation\_of\_Africa")

independence <- page %>%

html\_node(".wikitable") %>%

html\_table(fill = TRUE)

independence <- independence %>%

select(-Rank) %>%

map\_df(~str\_remove\_all(., "\\[.\*\\]")) %>%

rename(country = `Country[a]`,

colonial\_name = `Colonial name`,

colonial\_power = `Colonial power[b]`,

independence\_date = `Independence date`,

first\_head\_of\_state = `First head of state[d]`,

independence\_won\_through = `Independence won through`)

This dataset was scraped from the following Wikipedia [table](https://en.wikipedia.org/wiki/Decolonisation_of_Africa#Timeline).  
It shows when African countries gained independence from which colonial powers. In Chapter 11, I  
will show you how to scrape Wikipedia pages using R. For now, let’s take a look at the contents  
of the dataset:

independence

## # A tibble: 54 x 6

## country colonial\_name colonial\_power independence\_da… first\_head\_of\_s…

##

## 1 Liberia Liberia United States 26 July 1847 Joseph Jenkins …

## 2 South … Cape Colony … United Kingdom 31 May 1910 Louis Botha

## 3 Egypt Sultanate of… United Kingdom 28 February 1922 Fuad I

## 4 Eritrea Italian Erit… Italy 10 February 1947 Haile Selassie

## 5 Libya British Mili… United Kingdo… 24 December 1951 Idris

## 6 Sudan Anglo-Egypti… United Kingdo… 1 January 1956 Ismail al-Azhari

## 7 Tunisia French Prote… France 20 March 1956 Muhammad VIII a…

## 8 Morocco French Prote… France Spain 2 March 19567 A… Mohammed V

## 9 Ghana Gold Coast United Kingdom 6 March 1957 Kwame Nkrumah

## 10 Guinea French West … France 2 October 1958 Ahmed Sékou Tou…

## # ... with 44 more rows, and 1 more variable:

## # independence\_won\_through

as you can see, the date of independence is in a format that might make it difficult to answer questions  
such as *Which African countries gained independence before 1960 ?* for two reasons. First of all,  
the date uses the name of the month instead of the number of the month (well, this is not such a  
big deal, but still), and second of all the type of  
the independence day column is *character* and not “date”. So our first task is to correctly define the column  
as being of type date, while making sure that R understands that *January* is supposed to be “01”, and so  
on.

**Using {lubridate}**

There are several helpful functions included in {lubridate} to convert columns to dates. For instance  
if the column you want to convert is of the form “2012-11-21”, then you would use the function ymd(),  
for “year-month-day”. If, however the column is “2012-21-11”, then you would use ydm(). There’s  
a few of these helper functions, and they can handle a lot of different formats for dates. In our case,  
having the name of the month instead of the number might seem quite problematic, but it turns out  
that this is a case that {lubridate} handles painfully:

library(lubridate)

##

## Attaching package: 'lubridate'

## The following object is masked from 'package:base':

##

## date

independence <- independence %>%

mutate(independence\_date = dmy(independence\_date))

## Warning: 5 failed to parse.

Some dates failed to parse, for instance for Morocco. This is because these countries have several  
independence dates; this means that the string to convert looks like:

"2 March 1956

7 April 1956

10 April 1958

4 January 1969"

which obviously cannot be converted by {lubridate} without further manipulation. I ignore these cases for  
simplicity’s sake.

Let’s take a look at the data now:

independence

## # A tibble: 54 x 6

## country colonial\_name colonial\_power independence\_da… first\_head\_of\_s…

##

## 1 Liberia Liberia United States 1847-07-26 Joseph Jenkins …

## 2 South … Cape Colony … United Kingdom 1910-05-31 Louis Botha

## 3 Egypt Sultanate of… United Kingdom 1922-02-28 Fuad I

## 4 Eritrea Italian Erit… Italy 1947-02-10 Haile Selassie

## 5 Libya British Mili… United Kingdo… 1951-12-24 Idris

## 6 Sudan Anglo-Egypti… United Kingdo… 1956-01-01 Ismail al-Azhari

## 7 Tunisia French Prote… France 1956-03-20 Muhammad VIII a…

## 8 Morocco French Prote… France Spain NA Mohammed V

## 9 Ghana Gold Coast United Kingdom 1957-03-06 Kwame Nkrumah

## 10 Guinea French West … France 1958-10-02 Ahmed Sékou Tou…

## # ... with 44 more rows, and 1 more variable:

## # independence\_won\_through

As you can see, we now have a date column in the right format. We can now answer questions such as  
*Which countries gained independence before 1960?* quite easily, by using the functions year(),  
month() and day(). Let’s see which countries gained independence before 1960:

independence %>%

filter(year(independence\_date) <= 1960) %>%

pull(country)

## [1] "Liberia" "South Africa"

## [3] "Egypt" "Eritrea"

## [5] "Libya" "Sudan"

## [7] "Tunisia" "Ghana"

## [9] "Guinea" "Cameroon"

## [11] "Togo" "Mali"

## [13] "Madagascar" "Democratic Republic of the Congo"

## [15] "Benin" "Niger"

## [17] "Burkina Faso" "Ivory Coast"

## [19] "Chad" "Central African Republic"

## [21] "Republic of the Congo" "Gabon"

## [23] "Mauritania"

You guessed it, year() extracts the year of the date column and converts it as a *numeric* so that we can work  
on it. This is the same for month() or day(). Let’s try to see if countries gained their independence on  
Christmas Eve:

independence %>%

filter(month(independence\_date) == 12,

day(independence\_date) == 24) %>%

pull(country)

## [1] "Libya"

Seems like Libya was the only one! You can also operate on dates. For instance, let’s compute the difference between  
two dates, using the interval() column:

independence %>%

mutate(today = lubridate::today()) %>%

mutate(independent\_since = interval(independence\_date, today)) %>%

select(country, independent\_since)

## # A tibble: 54 x 2

## country independent\_since

##

## 1 Liberia 1847-07-26 UTC--2018-12-15 UTC

## 2 South Africa 1910-05-31 UTC--2018-12-15 UTC

## 3 Egypt 1922-02-28 UTC--2018-12-15 UTC

## 4 Eritrea 1947-02-10 UTC--2018-12-15 UTC

## 5 Libya 1951-12-24 UTC--2018-12-15 UTC

## 6 Sudan 1956-01-01 UTC--2018-12-15 UTC

## 7 Tunisia 1956-03-20 UTC--2018-12-15 UTC

## 8 Morocco NA--NA

## 9 Ghana 1957-03-06 UTC--2018-12-15 UTC

## 10 Guinea 1958-10-02 UTC--2018-12-15 UTC

## # ... with 44 more rows

The independent\_since column now contains an *interval* object that we can convert to years:

independence %>%

mutate(today = lubridate::today()) %>%

mutate(independent\_since = interval(independence\_date, today)) %>%

select(country, independent\_since) %>%

mutate(years\_independent = as.numeric(independent\_since, "years"))

## # A tibble: 54 x 3

## country independent\_since years\_independent

##

## 1 Liberia 1847-07-26 UTC--2018-12-15 UTC 171.

## 2 South Africa 1910-05-31 UTC--2018-12-15 UTC 109.

## 3 Egypt 1922-02-28 UTC--2018-12-15 UTC 96.8

## 4 Eritrea 1947-02-10 UTC--2018-12-15 UTC 71.8

## 5 Libya 1951-12-24 UTC--2018-12-15 UTC 67.0

## 6 Sudan 1956-01-01 UTC--2018-12-15 UTC 63.0

## 7 Tunisia 1956-03-20 UTC--2018-12-15 UTC 62.7

## 8 Morocco NA--NA NA

## 9 Ghana 1957-03-06 UTC--2018-12-15 UTC 61.8

## 10 Guinea 1958-10-02 UTC--2018-12-15 UTC 60.2

## # ... with 44 more rows

We can now see for how long the last country to gain independence has been independent.  
Because the data is not tidy (in some cases, an African country was colonized by two powers,  
see Libya), I will only focus on 4 European colonial powers: Belgium, France, Portugal and the United Kingdom:

independence %>%

filter(colonial\_power %in% c("Belgium", "France", "Portugal", "United Kingdom")) %>%

mutate(today = lubridate::today()) %>%

mutate(independent\_since = interval(independence\_date, today)) %>%

mutate(years\_independent = as.numeric(independent\_since, "years")) %>%

group\_by(colonial\_power) %>%

summarise(last\_colony\_independent\_for = min(years\_independent, na.rm = TRUE))

## # A tibble: 4 x 2

## colonial\_power last\_colony\_independent\_for

##

## 1 Belgium 56.5

## 2 France 41.5

## 3 Portugal 43.1

## 4 United Kingdom 42.5

{lubridate} contains many more functions. If you often work with dates, duration or interval data, {lubridate}  
is a package that you have to master.