# Scraping leboncoin

```
library(httr)
library(rvest)
library(tidyverse)
library(knitr) # Just to have nice tables in the html document... -> function 'kable'
# And for geocoding:
# devtools::install_github(repo = 'rCarto/photon')
library(photon)
```

# Manage to read leboncoin pages

If you try to read a page in Leboncoin using the `read_html()` function directly you'll likely get a 403 error message, meaning you have been denied the access. Send a query as if it was a regular query sent **from your browser**.

Open your browser **Inspector** -> go to the Network tab and them go select the part of the answer that corresponds to the html part of the answer. Then have a look at the **headers** sent along with your query. We'll use 3 info items here and add them to our query :

- User-Agent
- Accept (accepted formats for the answer)
- Accept-Language (accepted languages for the answer)

```
go_GET <- function(url){
  result=GET(url,
             add_headers(
               "User-Agent" = "Mozilla/5.0 (Windows NT 6.1; Win64; x64; rv:62.0) Gecko/20100101 Firefox/62.0",
               "Accept"="text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8",
               "Accept-Language"="fr,fr-FR;q=0.8,en-US;q=0.5,en;q=0.3"))
  return(result)
}
go_GET("https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/")
```

```
## Response [https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/]
##   Date: 2018-09-28 14:44
##   Status: 200
##   Content-Type: text/html; charset=utf-8
##   Size: 503 kB
## <!DOCTYPE html>
## <html lang="fr">
## <head>
##  <meta charset="utf-8">
##  <meta http-equiv="x-ua-compatible" content="ie=edge">
##  <title data-react-helmet="true">Ventes immobilières, maisons à vendre R...
##
##  <meta data-react-helmet="true" name="google-site-verification" content=...
##
##
## ...
```

# Scrape all ads in the real estate category in leboncoin

… for now, just for 1 department: Rhône.

## Get links to all ads

Get the total number of ads and deduce the number of pages to scrape (35 ads are displayed per page).

```
url_base <- "https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/"
url_base_raw <- go_GET(url_base)
html_base <- read_html(url_base_raw)

nb_links <- html_base  %>%
  html_nodes("._2ilNG") %>%
  html_text() %>%
  first() %>%
  str_replace(" ","") %>%
  as.numeric()
nb_pages=ceiling(nb_links/35)


pages=c(url_base,
        str_c(url_base,"p-",2:nb_pages))
pages[1:5]
```

```
## [1] "https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/"
## [2] "https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/p-2"
## [3] "https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/p-3"
## [4] "https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/p-4"
## [5] "https://www.leboncoin.fr/ventes_immobilieres/offres/rhone_alpes/rhone/p-5"
```

Now we have the urls of all the pages we have to scrape to get links to all ads (vector `pages`).

# For each page, get link to individual ads

Definition of function `ads_by_page()` which takes a **page listing ads as an input and returns all ads' urls** as output.

I added some **random waiting time** to each call to `ads_by_page()` of 1 to 5 seconds.

```
ads_by_page <- function(page){
  Sys.sleep(runif(1,1,5))
  my_html <- read_html(go_GET(page))
  links <- my_html %>%
      html_nodes(".clearfix") %>%
      html_attr("href") %>%
      na.omit()
  tib <- tibble(urls=str_c("https://www.leboncoin.fr",links))
  return(tib)
}
ads_by_page(pages[1])
```

```
## # A tibble: 35 x 1
##    urls
##    <chr>
##  1 https://www.leboncoin.fr/ventes_immobilieres/1498833165.htm/
##  2 https://www.leboncoin.fr/ventes_immobilieres/1487886414.htm/
##  3 https://www.leboncoin.fr/ventes_immobilieres/1469672965.htm/
##  4 https://www.leboncoin.fr/ventes_immobilieres/1497446180.htm/
##  5 https://www.leboncoin.fr/ventes_immobilieres/1498454616.htm/
##  6 https://www.leboncoin.fr/ventes_immobilieres/1498828413.htm/
##  7 https://www.leboncoin.fr/ventes_immobilieres/1412219839.htm/
##  8 https://www.leboncoin.fr/ventes_immobilieres/1497620632.htm/
##  9 https://www.leboncoin.fr/ventes_immobilieres/1498826837.htm/
## 10 https://www.leboncoin.fr/ventes_immobilieres/1417925822.htm/
## # ... with 25 more rows
```

Now **apply iteratively** function `ads_by_page()` to all pages' urls listed in `pages`.

And I didn't actually do it on all 405 pages but only on 3 of them, to show you the principle!

```
tib_ads_urls <- map(pages[1:3],safely(ads_by_page)) %>%
  map("result") %>%
  bind_rows()
```

# For each ad, get info

Define function `ad_info()`, which takes **an ad's url** as an input and returns, as an output, a **tibble** with information regarding

- `url` : the ads' urls
- `title` : their titles,
- `type` : the type of property
- `surface` : the surface of the property
- `rooms` : the number of rooms
- `GHG` : Greenhouse gas emission category
- `energy_class` : Energy class category,
- `location` : Location of the property

I added some **random waiting time** to each call to `ad_info()` of 1 to 5 seconds.

```
ad_info <- function(ad){
    Sys.sleep(runif(1,1,5))
    html_ad <- read_html(go_GET(ad))
    title <- html_ad %>%
      html_nodes("._1KQme") %>%
      html_text()
    criteria <-
      tibble(name= html_ad %>% html_nodes("._3-hZF") %>% html_text(),
             value=html_ad %>% html_nodes("._3Jxf3") %>% html_text())
    f=function(x){if(length(x)==0){x=NA};return(x)}
    type    <- filter(criteria, name=="Type de bien")$value %>% f()
    surface <- filter(criteria, name=="Surface")$value %>%
      str_extract("^\\d*") %>% f()
    rooms   <- filter(criteria, str_detect(name,"Pi.ces"))$value %>%
      as.numeric() %>% f()

    price <- html_ad %>%
      html_nodes(".eVLNz") %>%
      html_text() %>%
      first() %>%
      str_replace_all("[^0-9]","") %>%
      as.numeric()
    GHG <- html_ad %>%
      html_nodes("._2BhIP") %>%
      html_text() %>%
      first()
    energy_class <-html_ad %>%
      html_nodes("._15MMC") %>%
      html_text() %>%
      first()
    location <- html_ad %>%
      html_nodes("._1aCZv") %>%
      html_text() %>%
      str_replace("Voir sur la carte","")
    ## Geocoding
    #
    #zipcode <- str_extract(location,"\\d+")
    #city <- str_extract(location,"[A-Za-z- ]+")
    # url <- str_c("https://geocode.xyz/",zipcode,"+",city,"?json=1&region=FR")
    # raw_json <- GET(url)
    # geocode <- content(raw_json,as="parsed")
    # latitude <- geocode$latt
    # longitude <- geocode$longt
    coord_table=photon::geocode(location)
    latitude=coord_table$lat[1]
    longitude=coord_table$lon[1]
    tib_ad=bind_cols(urls=ad,
                     title=title,
                     price=price,
                     type=type,
                     surface=surface,
                     rooms=rooms,
                     GHG=GHG,
                     energy_class=energy_class,
                     location=location,
                     latitude=latitude,
                     longitude=longitude)
    return(tib_ad)
}
ad_info(tib_ads_urls$urls[1]) %>% kable()
```

| urls | title | price | type | surface | rooms | GHG | energy_class | location | latitude |
|---|---|---|---|---|---|---|---|---|---|
| https://www.leboncoin.fr/ventes_immobilieres/1498833165.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1498833165.htm/) | Terrain viabilisé 919m² Cailloux-sur-Fontaines | 399000 | Terrain | 919 | NA | NA | NA | Cailloux-sur-Fontaines 69270 | 45.85238 |

Please note that while during the course we had geocoded using the geocode.xyz API, this might not be optimal for geocoding with R. This choice was due to the fact that we wanted to show you **how to use an API in a direct query**. You can also geocode using other APIs with API clients (see for instance function `geocode()` in package `photon`, which is the solution we finally used in this document).

Now **apply iteratively this function `ad_info()`** to all ads in `tib_ads_urls`, using `purrr` iteration.

I actually did not do it on all ads but just on 20 of them to show you the principle!

```
tmp=Sys.time()
tib_ads <- map(tib_ads_urls$urls[1:20],
                  safely(ad_info)) %>%
     map("result") %>% bind_rows()
time_for_20_ads <- Sys.time()-tmp
tib_ads %>% kable()
```

| urls | title | price | type | surface | rooms | GHG | energy_class | location |
|---|---|---|---|---|---|---|---|---|
| https://www.leboncoin.fr/ventes_immobilieres/1498833165.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1498833165.htm/) | Terrain viabilisé 919m² Cailloux-sur-Fontaines | 399000 | Terrain | 919 | NA | NA | NA | Cailloux-sur-Fontaines 69270 |
| https://www.leboncoin.fr/ventes_immobilieres/1487886414.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1487886414.htm/) | Appartement 4 pièces 70 m² | 235000 | Appartement | 70 | 4 | B | E | Caluire-et-Cuire 69300 |
| https://www.leboncoin.fr/ventes_immobilieres/1469672965.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1469672965.htm/) | Maison 9 pièces 260 m² | 590000 | Maison | 260 | 9 | B | E | Lentilly 69210 |
| https://www.leboncoin.fr/ventes_immobilieres/1497446180.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1497446180.htm/) | Maison de ville de 145m2 avec terrain | 349900 | Maison | 145 | 5 | B | E | Villefranche-sur-Saône 69400 |
| https://www.leboncoin.fr/ventes_immobilieres/1498454616.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1498454616.htm/) | Maison a LUCENAY | 200000 | Maison | 90 | 4 | B | E | Lucenay 69480 |
| https://www.leboncoin.fr/ventes_immobilieres/1498828413.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1498828413.htm/) | Appartement T3 | 159000 | Appartement | 70 | 3 | B | E | Villeurbanne 69100 |
| https://www.leboncoin.fr/ventes_immobilieres/1412219839.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1412219839.htm/) | T4 - Carré Ouest | 338000 | Appartement | 85 | 4 | NA | NA | Francheville 69340 |
| https://www.leboncoin.fr/ventes_immobilieres/1497620632.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1497620632.htm/) | Appartement 3 pièces 63 m² | 298000 | Appartement | 63 | 3 | B | E | Lyon 69001 |
| https://www.leboncoin.fr/ventes_immobilieres/1498826837.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1498826837.htm/) | T2 lyon 9e | 205000 | Appartement | 53 | 2 | B | E | Lyon 69009 |
| https://www.leboncoin.fr/ventes_immobilieres/1417925822.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1417925822.htm/) | Propriété 11 pièces 450 m² | 895000 | Maison | 450 | 11 | B | E | Saint-Pierre-la-Palud 69210 |
| https://www.leboncoin.fr/ventes_immobilieres/1454127092.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1454127092.htm/) | Maison de village 2 pièces 112 m² | 55000 | Maison | 112 | 2 | NA | NA | Saint-Clément-les-Places 69930 |
| https://www.leboncoin.fr/ventes_immobilieres/1495335959.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1495335959.htm/) | Maison de village 3 pièces 53 m² | 83000 | Maison | 53 | 3 | NA | NA | Courzieu 69690 |
| https://www.leboncoin.fr/ventes_immobilieres/1495336043.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1495336043.htm/) | Appartement 3 pièces 50 m² | 64000 | Appartement | 50 | 3 | B | E | Tarare 69170 |
| https://www.leboncoin.fr/ventes_immobilieres/1493308695.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1493308695.htm/) | T4 rez de jardin | 382000 | Appartement | 86 | 4 | B | E | Lyon 69009 |
| https://www.leboncoin.fr/ventes_immobilieres/1490785555.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1490785555.htm/) | Dernière Opportunité | 189000 | Appartement | 37 | 2 | B | E | Lyon 69005 |
| https://www.leboncoin.fr/ventes_immobilieres/1490737572.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1490737572.htm/) | T2 ideal investisseur ou premier aquisition | 245000 | Maison | 41 | 2 | B | E | Lyon 69005 |
| https://www.leboncoin.fr/ventes_immobilieres/1493272654.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1493272654.htm/) | Rare t3 rez de jardin BORD DE SAONE | 249700 | Appartement | 57 | 3 | B | E | Lyon 69009 |
| https://www.leboncoin.fr/ventes_immobilieres/1493366641.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1493366641.htm/) | T4 rez de jardin | 370000 | Appartement | 80 | 4 | B | E | Villeurbanne 69100 |
| https://www.leboncoin.fr/ventes_immobilieres/1495288729.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1495288729.htm/) | T3 derniere opportunite Dernier étage | 191585 | Appartement | 60 | 3 | B | E | Villeurbanne 69100 |
| https://www.leboncoin.fr/ventes_immobilieres/1495183177.htm/ (https://www.leboncoin.fr/ventes_immobilieres/1495183177.htm/) | Studio Ideal investisseur | 206000 | Appartement | 37 | 1 | B | E | Villeurbanne 69100 |

For 20 ads, it took us about 1.5 minutes to get the data so if we would like to do this on all ads (~14000 ads) then it would take along time (about 18 hours…)!