PROJECT

- Your goal

  L'objectiu principal és poder obtenir la informació principal de qualsevol ciutat (superfície, població, coordenades, ...) per tal d'explotar-la i representar-la de forma fàcil i ràpida.

- The data source and web technology it is based upon

  La font que utilitzarem serà Wikipedia. Buscarem l'enllaç que fa referencia a diverses ciutats d'Espanya i extraurem tota la informació que hi ha a la taula de la dreta de la pàgina (tots els enllaços referents a ciutats contenen aquesta taula informativa).

- The approach/technology that you will use for downloading

  Per obtenir les dades de Wikipedia, utilitzarem Rstudio i un package anomenat XLS.

- How you have scraped the data (be detailed, but not unnecessarily exhaustive).

  Per realitzar el web scraping, llegim cada cop una pàgina de la wikipedia seguint els següents passos:

1. Esborrar els primers registres que fan referencia a les imatges de la ciutat i de la seva bandera i que, per tant, apareixen buits o com NA's.
2. Agafar únicament aquelles files que fan referència a la informació que ens interessa: ubicació (coordenades), superfície, població i densitat.
3. Depurar cadascun dels camps llegits. Per exemple, la ubicació ve donada tant en graus com en coordenades. Ens quedem únicament amb les dues coordenades que vénen informades en un mateix camp i les separem en dues noves variables (coordX i coordY).

Variable	Abans de la depuració	Després de la depuració
Ubicació	"41°22'57"N 2°10'37"E / 41.3825, 2.1769444444444Coordenadas: 41°22'57"N 2°10'37"E / 41.3825, 2.1769444444444"	CoordX = 2.1769444444444 CoordY = 41.3825
Superfície	"102,15 km^2 "	102.15
Població	"1 608 746 hab. (2016)"	1608746
Densitat	"15 748,86 hab./km^2"	15748.86

4. Agrupar totes les variables en un sol data.frame en què cada fila fa referència a una ciutat. Aquest data.frame és el que retorna la funció creada.

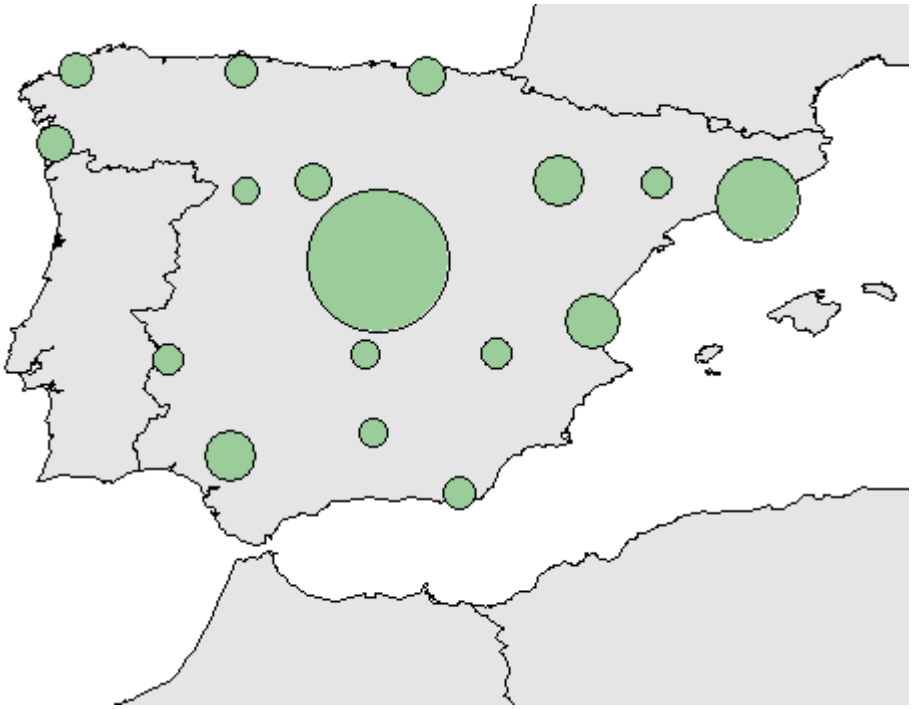
- If it makes sense, show (or provide) the results of the scraping process.

  Com a exemple, hem triat 17 ciutats d'Espanya. Un cop fet tot el procés comentat a l'apartat anterior, obtenim la següent taula:

	Superficie	Población	Densidad	CoordX	CoordY
Barcelona	102,15	1608746	15748,86	2,18	41,38
Madrid	605,77	3165541	5225,65	-3,69	40,42
Valencia	134,65	790201	5868,56	-0,38	39,47
Sevilla	140,8	690566	4904,59	-5,98	37,38
Bilbao	41,6	345122	8296,2	-2,95	43,26
Zaragoza	973,78	661108	678,91	-0,88	41,65
Vigo	109,06	292817	2684,92	-8,72	42,23
Oviedo	186,65	220567	1181,71	-5,85	43,36
La_Coruña	37,83	243978	6449,33	-8,38	43,37
Badajoz	1440,37	149946	104,1	-6,97	38,88
Valladolid	197,91	301876	1525,32	-4,73	41,65
Zamora	149,28	63217	423,48	-5,76	41,50
Almeria	296,21	194515	656,68	-2,45	36,83
Ciudad_Real	284,98	74054	259,86	-3,92	38,98
Albacete	1125,91	172426	153,14	-1,86	39,00
Jaen	424,3	114658	270,23	-3,79	37,77
Lérida	211,7	138144	652,55	0,63	41,62

- If it is reasonably short you can do an analysis of the data. If you think it adds too much extra work, explain how you would perform the analysis.

  L'anàlisi que s'ha fet ha consistit en la representació gràfica de les ciutats a partir de les coordenades proporcionades per la taula. A més, també s'ha tingut en compte la població de forma que els cercles representats són proporcionals a la quantitat d'habitants de la ciutat.



- Finish the report with a brief discussion of the limitations that you have found to do the scraping and what do you think that you should have done in a real situation

  El problema principal que hem trobat és que el nostre script està pensat per un tipus de pàgines de Wikipedia que segueixen una estructura determinada. Hi ha algunes ciutats en què aquesta estructura canvia i, per tant, no podríem obtenir les dades corresponents ni representar-les al mapa. A més, també existeixen ciutats amb dades mancants (per exemple, sense informació de les coordenades).
  Com a millora del projecte, es podria buscar una web alternativa que també contingui aquesta informació de manera que si a wikipedia no hi és, puguem anar a buscar-la a un altre link.

Codi R:

```
## PROJECTE WEB SCRAPING ##

library(XML)
library(dplyr)
library(rvest)
library(stringr)
library(rworldmap)
library(rworldxtra)
library(ggplot2)
library(maps)



ciutat <- function(vecnom){
  total <- matrix(ncol=5,nrow=length(vecnom))
  rownames(total)<-vecnom
  colnames(total) <- c("Superficie", "Población", "Densidad","CoordX",
"CoordY")
  for(i in 1:length(vecnom)){
    scotusURL <- paste0("https://es.wikipedia.org/wiki/", vecnom[i])
    temp <- scotusURL %>% html %>% html_nodes("table")


    data_total <- as.data.frame(html_table(temp[1], fill=T))[,1:2] #llegim
dades
    data_total[,1] <- unlist(lapply(data_total[,1], function(x)
str_extract(pattern = "[[:alpha:]]+",x))) #llegim bé els noms
    data <- data_total[data_total[,1] %in% c("Ubicación", "Superficie",
"Población", "Densidad"),] #info rellevant


    #Arreglem coordenades
    datacoord <-  data[1,2]
    coord <- strsplit(datacoord, "/ ")[[1]][3]
    coordX <- as.numeric(strsplit(coord,", ")[[1]][2])
    coordY <- strsplit(coord,", ")[[1]][1]
    coordY <- as.numeric(substr(coordY,2,nchar(coordY)))

    #Arreglem superfície
    sup <- data[data[,1]=="Superficie",2]
    sup <- gsub(" km²","", sup)
    sup <- as.numeric(gsub(",",".", sup))

    #Arreglem población
    pob <- data[data[,1]=="Población",2]
    pob <- gsub("[[:blank:]]hab.[[:blank:]][[:punct:]]\\d{4}[[:punct:]]",
"",pob)
    pob <- as.numeric(gsub("[[:blank:]]","", pob))

    #Arreglem densitat
    dens <- data[data[,1]=="Densidad",2]
    dens <- gsub(" hab./km²","", dens)
    dens <- gsub(",",".",dens)
    dens <- as.numeric(gsub("[[:blank:]]","",dens))
```

```
    total[i,1] <- sup
    total[i,2] <- pob
    total[i,3] <- dens
    total[i,4] <- coordX
    total[i,5] <- coordY
  }
  return(as.data.frame(total))

}
noms<-c('Barcelona','Madrid','Valencia','Sevilla','Bilbao','Zaragoza',
        'Vigo','Oviedo','La_Coruña','Badajoz','Valladolid','Zamora',
        'Almeria','Ciudad_Real','Albacete','Jaen','Lérida')
dd<-ciutat(noms)

newmap <- getMap(resolution = "high")
plot(newmap, xlim = c(-10, 4), ylim = c(39, 39), asp = 1, col='grey90')
max.symbol.size=10
min.symbol.size=2
bubble.size <- (dd$Población-min(dd$Población, na.rm=T))/((max(dd$Población,
na.rm=T)+.0001)-min(dd$Población, na.rm=T))*(max.symbol.size-
min.symbol.size)+min.symbol.size

points(dd[,'CoordX'], dd[,'CoordY'], pch=21, col='black',bg='darkseagreen3',
cex=bubble.size)
```