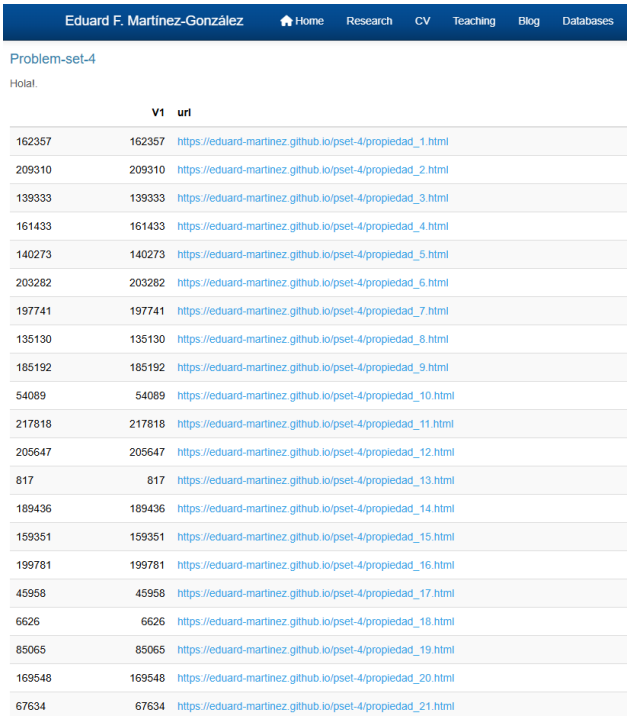
TALLER R - Estadistica y programación

Alejandro Correa Jimenez - 202022775

2024-05-31

1.Extracción de información

Inicialmente se identifica la fuente del conjunto de datos.



Eduard F. Martínez-González Home Research CV Teaching Blog Databases

Problem-set-4

Hola!.

	V1	url
162357	162357	https://eduard-martinez.github.io/pset-4/propiedad_1.html
209310	209310	https://eduard-martinez.github.io/pset-4/propiedad_2.html
139333	139333	https://eduard-martinez.github.io/pset-4/propiedad_3.html
161433	161433	https://eduard-martinez.github.io/pset-4/propiedad_4.html
140273	140273	https://eduard-martinez.github.io/pset-4/propiedad_5.html
203282	203282	https://eduard-martinez.github.io/pset-4/propiedad_6.html
197741	197741	https://eduard-martinez.github.io/pset-4/propiedad_7.html
135130	135130	https://eduard-martinez.github.io/pset-4/propiedad_8.html
185192	185192	https://eduard-martinez.github.io/pset-4/propiedad_9.html
54089	54089	https://eduard-martinez.github.io/pset-4/propiedad_10.html
217818	217818	https://eduard-martinez.github.io/pset-4/propiedad_11.html
205647	205647	https://eduard-martinez.github.io/pset-4/propiedad_12.html
817	817	https://eduard-martinez.github.io/pset-4/propiedad_13.html
189436	189436	https://eduard-martinez.github.io/pset-4/propiedad_14.html
159351	159351	https://eduard-martinez.github.io/pset-4/propiedad_15.html
199781	199781	https://eduard-martinez.github.io/pset-4/propiedad_16.html
45958	45958	https://eduard-martinez.github.io/pset-4/propiedad_17.html
6626	6626	https://eduard-martinez.github.io/pset-4/propiedad_18.html
85065	85065	https://eduard-martinez.github.io/pset-4/propiedad_19.html
169548	169548	https://eduard-martinez.github.io/pset-4/propiedad_20.html
67634	67634	https://eduard-martinez.github.io/pset-4/propiedad_21.html

1.1. Obtención de URLs

Mediante el identificador a y atributo href, recolectamos las URL presentes en la página web

```
dir <- "https://eduard-martinez.github.io/pset-4.html"
tmp <- read_html(dir)
url_full <- html_nodes(tmp, "a") %>% html_attr("href")
```

1.2. Filtro de URLs

Del conjunto completo de URLs, seleccionamos aquellas que en la posición 42 contienen la palabra propiedad

```
url_subset <- url_full[which(substring(url_full,42,50)=="propiedad")]
```

1.3. Extracción de tablas

En cada iteración del ciclo for:

1. Leemos el HTML.
2. Extraemos la informacion que necesitamos de la tabla.
3. Agregamos dicha información a lista_tablas.

```
lista_tablas <- vector("list",3)
i <- 1

for (dir in url_subset){
  #Leemos el HTML
  tmp <- read_html(dir)

  # De la tabla contenida en tbody, separamos la información identificada con td
  tabla <- html_nodes(tmp, "tbody") %>% html_elements("td")
  # guardamos y extraemos los datos que necesitamos
  lon_lat <- tabla[12] %>% html_text2()
  price <- tabla[8] %>% html_text2()

  #Y agregamos los elementos a la lista de tablas
  lista_tablas[[i]] <- list(coordenadas = lon_lat, precio = price)

  i <- i + 1
}
```

1.4. Preparando de la información

Utilizando la función `rbindlist()`, se agrupan las listas por nombre.

```
db_house <- rbindlist(lista_tablas, use.names=TRUE)

#Convertimos la columna precio a tipo num
db_house$precio <- as.numeric(db_house$precio)
```

2. Manipulando la información

2.1. Creación de objeto sf

```
sf_house <- st_as_sf(db_house, wkt = "coordenadas")
```

2.2. Pintando el mapa

Se utilizan las funciones `geom_sf()` y `scale_color_viridis()` para gráficas los puntos almacenados.

```
ggplot() + geom_sf(data = sf_house, aes(colour = precio)) + scale_color_viridis()
```