Taller de R: Estadística y Programación

Problem set 4

2024-04-20

En este taller se evalúan los temas vistos en las clases 11 y 13 del curso. Lea atentamente las instrucciones del taller.

Instrucciones

- Este taller representa el **25%** de la nota total del curso y puede ser realizado de manera individual o en grupos de hasta 3 personas. En las primeras líneas del script, escriba su nombre, código y la versión de R que está utilizando. Además, al inicio del código, debe incluir las librerías que utilizará en la sesión, por ejemplo: `pacman`, `rio`, `data.table`, `tidyverse`, `sf`, y `rvest`.
- Asegúrese de crear un nuevo repositorio en su cuenta de GitHub. Si trabaja en grupo, solo un integrante debe crear el repositorio y compartir el acceso con los demás. El repositorio debe ser público para permitir el acceso desde cualquier cuenta de GitHub. Incluya al menos tres carpetas en el repositorio: `input` (datos originales), `output` (datos procesados), y `code` (script con la respuesta del taller).
- Todos los integrantes del grupo deben publicar el enlace al repositorio de GitHub en la actividad **Problem-set-4** del Bloque Neón antes de las 23:59 horas del 06 de junio de 2024.
- Por favor, organice su trabajo cuidadosamente y comente paso a paso cada línea de código. Recuerde **NO** usar acentos ni caracteres especiales dentro del código para evitar problemas al abrir los scripts en diferentes sistemas operativos.
- No seguir estas instrucciones resultará en una penalización del **20%** en la nota final.

Solucionar:

Para resolver este problem set, deberá poner en práctica los conocimientos adquiridos en la clase de *web-scraping* y gestionar la información *GIS* que extraerá de internet. Finalmente, empleará *R-Markdown* para generar un HTML que posteriormente podrá publicar en su página de *GitHub*.

1. Extraer la información de internet (50%)

Primero, debe dirigirse a la página https://eduard-martinez.github.io/pset-4.html y examinar su contenido.

- **1.1 Obtener las URL**

Cree un objeto llamdo `url_full` que almacene el vector de URLs contenidas en la página ../pset-4.html.

- **1.2 Filtrar URL:**

Del objeto `url_full` mantenga únicamente las URLs que contengan la palabra `propiedad`, guarde el resultado en un objeto llamado `url_subset`.

- **1.3 Extraer las tablas de los HTML:**

Utilice un bucle o función que itere sobre todos los elementos del objeto `url_subset`. Para cada URL, extraiga de su correspondiente HTML la tabla que contiene la información de las coordenadas y el precio de la propiedad. Almacene el resultado de este loop en un objeto tipo lista llamado `lista_tablas`.

- **1.4 Preparar información:**

Utilice la función `rbindlist` del paquete `data.table` para convertir la lista `lista_tablas` en un dataframe. Almacene este resultado en un objeto llamado `db_house`, que contendrá toda la información de las tablas.

2. Manipular la información GIS (50%)

- **2.1 Cree un objeto `sf`**

Utilice la función `st_as_sf` del paquete `sf` para convertir el objeto `db_house` en un `SimpleFeature` de tipo punto. Nombre este nuevo objeto `sf_house`.

- **2.2 Pintar mapa**

Utilice la función `geom_sf` de la librería `ggplot2` para crear un mapa que visualice los puntos almacenados en el objeto `sf_house`. Utilice el valor de la vivienda como escala de colores, aplicando las paletas de colores de la función `scale_fill_viridis`. Una vez generado el mapa, exporte este objeto en formato `.pdf`.

3. Bonos:

Si resuelve al menos uno de los dos bonos, obtendrá un 5 en el problem-set. Si resuelve los dos bonos, recibirá un 5 como nota final del curso.

- **3.1 `R-Markdown`**

Resuelva este problem-set utilizando `R-Markdown`. Debe generar un archivo `HTML` que contenga las respuestas. Incluya tablas y/o gráficos con algunas estadísticas descriptivas de los datos y escriba un pequeño párrafo con el análisis de las tablas y/o mapas. Otros analísis son bienvenidos.

- **3.2 `GitHub`**

Suba el archivo `HTML` generado en el punto anterior a su repositorio de `GitHub` donde almacena su página web. Comparta la `URL` del `HTML` publicado en su página web.