Diplomatura en análisis de datos aplicados al desarrollo de políticas públicas

UNIVERSIDAD NACIONAL GUILLERMO BROWN

Trabajo práctico
Introducción al manejo de grandes volúmenes de datos

Andrea Ariza

Para la creación y normalización de un .db realizamos lo siguiente:

1) Seleccionamos la encuesta anual de hogares del portal de datos abiertos del gobierno de la Ciudad de Buenos Aires.
2) Luego, en Google collab empezamos a escribir el código necesario.
3) En primer lugar, importamos la librería pandas y recuperamos la base de datos a partir del url provisto por el gobierno.
4) Observamos que, para poder leer el url, necesitamos realizar un encoding.
5) Una vez obtenidos los datos, empezamos a observar la estructura de la tabla. Para eso realizamos lo siguiente: indagamos la estructura (cantidad de columnas y filas), luego pedimos que enliste el nombre de las 31 columnas que posee la tabla y, finalmente, le solicitamos que nos muestre las cinco primeras filas para poder tener un acercamiento a los datos.
6) Dada la información provista, decidimos quedarnos con 7 columnas: edad, comuna, sexo, ingreso per cápita familiar, nivel actual, lugar de nacimiento, sector educativo.
7) Luego de realizar el proceso de filtrado de datos, lo guardamos en un nuevo df. que se llama *df_encuesta_reducida*
8) Para realizar el proceso de guardado, importamos la librería sqlite3, generamos la conexión y utilizamos el método *tosql()* de nuestro df.reducido y le pasamos como parámetro el nombre que queremos que tenga la tabla, la conexión que creamos en el paso anterior y un tercer parámetro que indica que sí ya existe esa tabla que la reemplace. Actualizo y cierro la conexión.
9) Luego para realizar consultas, voy a abrir nuevamente la conexión y voy a indagar el tipo de dato de cada columna para chequear la consistencia con el modelo.
10) Finalmente, si quisiéramos iniciar con el proceso de análisis de los datos podríamos realizar un conjunto de *query* tales como ordenar los datos de acuerdo a los datos de comunas o filtrar los datos por edad.
11) Una vez realizadas las consultas, cerramos la conexión.