IIC2413 — Bases de Datos

Control 3: manejo de datos para un problema real

Durante esta actividad trabajaremos con los datos del sistema nacional de información municipal, del gobierno de Chile, y una planilla de datos preliminares del plebiscito 2020[1]
Tienes acceso a dos fuentes de datos.

- Una base de datos en SQLite llamada municipios.db, la que tiene información sobre el presupuesto municipal y la cantidad de personas de cada una de las comunas de Chile. Esos datos se almacenan en un esquema con tres tablas. Una tabla `Comuna(id,nombre`, que asigna a cada comuna en Chile un identificador; una tabla `Personal(id_comuna,personas)`, que asigna a cada comuna la cantidad de personas que trabajan ahí, y una tabla `Presupuesto(id_comuna, municipio, plata)` que contiene el presupuesto de dicha comuna.
- Recuerda que en SQLite puedes poner `.table` para ver las tablas que hay, y puedes poner `.schema <tabla>` para ver el esquema de una tabla llamada `<tabla>`.
- Una archivo csv llamad `Resultados_Pleb.csv`, que contiene la votación agregada de cada comuna para el plebiscito de Octubre.

La idea de este control es hacer un análisis de datos en torno a la siguiente problemática:

> ¿Cómo se relaciona el presupuesto de las comunas con su votación en el plebiscito?

1. Manejo de datos

En este control, asumimos que ya sabes trabajar tanto con SQLite como con la librería pandas de python. Para trabajar tus datos necesitas tener todo en pandas, pero puede ser que quieras importar datos a SQLite antes de eso. Puede ser también que prefieras hacer el join en python, y para eso no necesitas importar. Para importar a SQLite, tienes dos opciones.

Importar desde la consola de SQLite3 Esta opción requiere que instales SQLite3 en tu computador. Aquí tienes que abrir la base de datos en el cliente de SQLite3 (el comando debería ser `sqlite3 municipios.db`) y luego:

[1] compilada por Vergara Perucich, Francisco & Greene, Ricardo & Correa Parra, Juan & Aguirre-Nuñez, Carlos & Cancino Contreras, Francisca. (2020). Cartografías del apruebo: Análisis preliminar del plebiscito para cambio constitucional, Chile 2020. 10.13140/RG.2.2.24281.34408.

- Crea en SQL una tabla para los resultados del plebiscito, con atributos de acuerdo a lo que ves en el archivo .csv.
- Dentro de SQLite3, ejecuta el comando `.separator ,`, para indicar que el separador es el caracter `,`.
- Ejecuta `.import archivo.extension tabla` para importar el archivo a la tabla.
- Revisa si la primera línea del archivo se importó o no, y cómo quedaron los espacios vacíos. Puede que quieras arreglar eso.
- **Importante:** Nunca grabes un .csv desde excel, por que excel puede cambiar la codificación de esos archivos. En caso de problema, siempre puedes volver a bajar el .csv desde la página del curso.

Importar los datos utilizando Python Existen varias formas de importar los datos utilizando Python. Una forma es la propuesta en `https://stackoverflow.com/questions/2887878/importing-a-csv-file-into-a-sqlite3-database-table-using-python`.

2. Tareas a realizar

Ten cuidado: puede ser que los datos estén sucios, pues son llenados por humanos. Es tu responsabilidad limpiar esos datos: para las siguientes tareas no debes tomar en cuenta datos nulos o que no han sido reportados (aunque eso signifique dejar fuera algunas comunas).

2.1. Correlación entre habitantes y presupuesto

1. Crea un gráfico de puntos para visualizar la correlación entre el presupuesto de las comunas y la cantidad de habitantes de esa comuna.
2. Observa que la correlación es ligeramente positiva (como es de esperar). Sin embargo, hay un outlier cuyo presupuesto se escapa. Puedes ver qué comuna es?

2.2. Descartando ciertos outliers

Ahora refinemos un poco la tabla de comunas que vamos a utilizar.

1. Crea un vista llamada `ComunaLimpia(id,nombre)` que contenga todas las comunas, **excepto** i) El outlier que identificaste en el apartado anterior, y ii) todas las comunas que tengan 5000 habitantes o menos. Esta vista debería tener 286 comunas.

2.3. Correlación entre presupuesto y votación del apruebo

1. Crea un gráfico de puntos para visualizar la correlación entre el presupuesto de las comunas y el porcentaje de votos apruebo (sobre el total de votos) en esas comunas. Anota el coeficiente de correlación. ¿Puedes inferir algo?

2. Ahora visualiza el mismo gráfico, pero utilizando solo aquellas comunas que estén en la vista ComunaLimpia. Vuelve a anotar el coeficiente de correlación.

3. ¿Como cambia la correlación? ¿Por qué crees que pasó eso?

3. Entrega y detalles administrativos

Este control es individual. La entrega de este control debe ser un archivo comprimido donde se encuentre un jupyter notebook, junto a cualquier archivo .csv o .db que estés llamando desde tu código, y un readme con los comandos que usaste para importar, y para cualquier otro comando que hayas puesto directo a sqlite, sin pasar por el notebook.
El plazo para el control es el **Viernes 4 de Diciembre, a las 20:00 hrs.**.
La nota se calcula como un promedio ponderado entre las partes 2.1, 2.2 y 2.3.