IIC2413 — Bases de Datos

Bonus SQL

Durante esta actividad trabajaremos con los datos de la encuesta CASEN. En la carpeta puedes encontrar distintos archivos que terminan en .csv:

- `comunas.csv`
- `demografia.csv`
- `casen.csv`

Tu primera labor será **importar estas tablas a una base de datos** en SQLite3 que se encuentra subida junto con esta actividad (`casen.db`). La tabla de demografía ya está importada, en una tabla llamada `demografia` (puedes poner `.table` para ver las tablas que hay y `.schema <nombre_tabla>` para ver el esquema de una tabla llamada `<nombre tabla>`). Para las otras dos tablas 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 casen.db`) y luego:

- Crea en SQL tablas de para las comunas y para los datos de la encuesta CASEN, con atributos de acuerdo a lo que ves en los 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.

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`.

1. Descripción general

Pueden trabajar en grupos de hasta 3 personas. Deben elegir **una de las tres** tareas y trabajar en las tareas durante la clase del 27 de marzo. Los trabajos cuyo análisis sea excelente recibirán una bonificación en la próxima interrogación.

1.1. Actividad 1: Remuneración y delincuencia

Tu deber es soportar o refutar la siguiente hipótesis:

> Las comunas con un promedio de remuneración por debajo del promedio nacional tienen una mayor tasa de delincuencia

Además se te pide responder la siguiente pregunta

> ¿Existe alguna relación entre el promedio de la remuneracion de cada comuna y los rangos etáreos de sus habitantes? ¿qué permite sustentar tu afirmación?

1.2. Actividad 2: Diferencias entre PSU y SIMCE

Tu deber es soportar o refutar la siguiente hipótesis:

> Las comunas de Chile que tienen un promedio SIMCE por sobre el promedio nacional, y que tienen un promedio PSU por debajo del promedio nacional, se ven más afectadas por la delincuencia.

Para esto debes apoyarte en la base de datos de la encuesta CASEN, pero también debes lograr una cuota de análisis más allá de la parte técnica de SQL

1.3. Actividad 3: Diferencias de género y edad

Tu deber es realizar consultas que soporten o refuten la siguiente hipótesis:

> Las comunas en que mayoritariamente viven mujeres no necesariamente tienen una menor tasa de delincuencia.

Además se te pide responder la siguiente pregunta

> ¿Existe alguna relación entre la tasa de delincuencia y los rangos etáreos de cada comuna? ¿qué permite sustentar tu afirmación?

2. Entrega y evaluación

Si completaste la tarea de forma satisfactoria, debes enviar un **Jupyter Notebook** al buzón habilitado en el SIDING en la fecha discutida en clases. Este notebook debe contener la descripción de tu trabajo, código, consultas y análisis, además del **nombre de los integrantes**. El foco de esta tarea son las **consultas SQL realizadas**, por lo que no deberías hacer el análisis de datos usando Python o alguna de sus librerías. Además, si es necesario, debes incluir un readme para señalar cómo correr tu tarea.

Un trabajo excelente deberia contar al menos con ciertas visualizaciones. Dentro de python puedes generar visualizaciones con los datos que tomes de SQLite. Para esto recomendamos alguna de estas librerías: Pygal (`http://pygal.org/en/stable/`) o matplotlib (`http://matplotlib.org/`).

Importante: El informe puede ser evaluado con 0, 3 o 4 puntos. 4 puntos es un trabajo por sobre lo esperado y 3 puntos es un trabajo excelente. Cada punto **será una décima adicional a tu Interrogación 1**.

Aclaración

Al editar los datos para facilitar la importación se modificó el header original del archivo `casen.csv`, pero este contiene información relevante acerca del origen de los datos que pueden considerar dentro de su análisis, las columnas originales eran:

- Proyección de población a Junio 2013 (base Censo de Población 2002). Instituto Nacional de Estadísticas
- Remuneración imponible promedio de afiliados a seguro de cesantía (pesos), abril 2013. Administradora de Fondos de Cesantía de Chile
- Promedio de Puntaje Simce Lectura 4 básico, 2012. Agencia de Calidad de la Educación
- Promedio de Puntaje Simce Matemáticas 4 básico, 2012. Agencia de Calidad de la Educación
- Promedio de Puntaje Simce Historia, Geografía y Cs. Sociales 4 básico, 2012. Agencia de Calidad de la Educación
- Promedio de Puntaje Prueba de Selección Universitaria, 2012. Consejo de Rectores
- Promedio de Puntaje PSU Lenguaje, 2012. Consejo de Rectores
- Promedio de Puntaje PSU Matemáticas, 2012. Consejo de Rectores
- Tasa de casos policiales por delitos de mayor connotación social por cada 100.000 habitantes, 2012. Ministerio del Interior