Introducción a la Ciencia de Datos y sus Metodologías

Proyecto Final: Base de datos, procesamiento y análisis.

Alumno:

Fernando Luna Ponce

Profesor:

Dr. Juan Pablo Soto Barrera

Hermosillo, Sonora 25 de noviembre de 2022

Tabla de contenidos.

Introducción

Un sistema gestor de base de datos es un software que permite administrar una base de datos. Es decir, con el podemos utilizar, configurar y extraer la información almacenada.

En este proyecto estaremos utilizando un sistema de base de datos relacionales conocido como MySQL para almacenar y procesar la información sobre las muertes ocasionadas por diabetes en el Estado de Sonora.

Se describirán los pasos a seguir para lograr la reproducibilidad de nuestro proyecto, desde como generar una instancia de MySQL utilizando la nube de AWS para poder ejecutar nuestro código en una libreta de Google Colab. También veremos como realizar la conexión con nuestra instancia de MySQL para realizar algunas inserciones en nuestra base de datos, realizar el procesamiento de nuestros datos y posteriormente realizar consultas a la base de datos.

Objetivos

- Generar una instancia de MySQL utilizando la nube de AWS.
- Generación de nuestros objetos de base de datos (tablas, vistas, funciones, etc.)
- Realizar la conexión con la base de datos de MySQL mediante Python.
- Realizar inserción de información en las tablas creadas.
- Realizar el procesamiento de la información en la base de datos.
- Ejecutar consultas que nos retornen información hacia un dataframe de Pandas.

Descripción de nuestra fuente de datos

En este proyecto se realizará la descarga de tres fuentes de información diferentes que nos servirán para obtener la cantidad de muertes ocasionadas por diabetes a lo largo de los años en el Estado de Sonora.

Las fuentes son:

- Página de datos abierto del INEGI: La información está contenida en diferentes archivos con formato CSV, esta contiene la información de las muertes generadas a nivel nacional por año.
- Cuéntame de México: Esta página contiene la información sobre la cantidad de habitantes en el Estado de Sonora en el año 2020.
- Página de la CONAPO: Contiene la información sobre la estimación de la cantidad de habitantes por localidades de México.

Generación de instancia de MySQL en AWS

Se generó una instancia de MySQL utilizando la nube de AWS, aprovechando que nos ofrece este servicio gratuito durante 12 meses. Esto nos permite poder utilizar el servicio de bases de datos relacionales de manera online, por lo que podremos acceder a ella en todo momento mediante una conexión a internet. Con esto podremos ejecutar nuestros procesos ya sea desde nuestro equipo local o utilizando una libreta de Google Colab.

En el github del proyecto podremos encontrar un manual que nos permitirá realizar la configuración de nuestra instancia de MySQL.

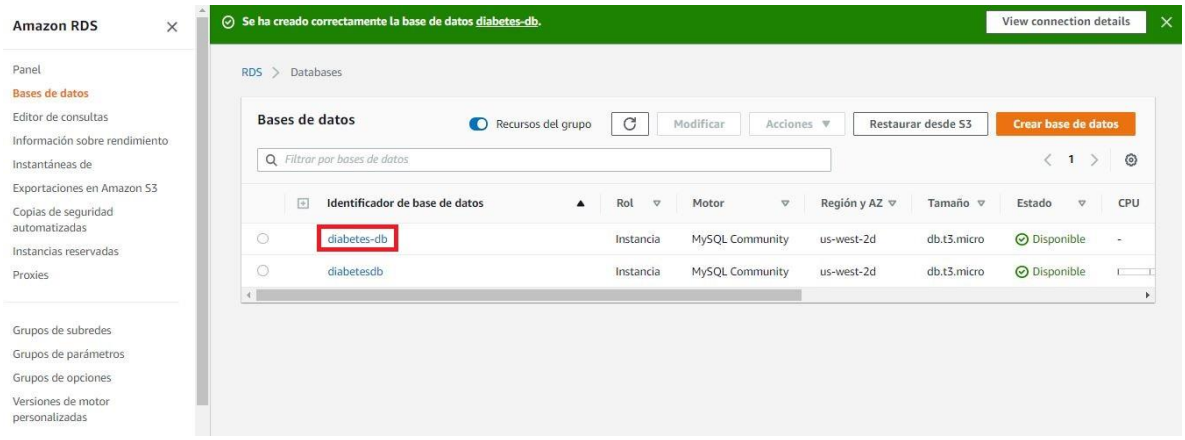

Fig. 1 Instancia de MySQL corriendo en la nube de AWS.

Generación de nuestros objetos de base de datos.

Si se quiere reproducir este proyecto, tendremos que utilizar los scripts que vienen en un archivo SQL, localizado en nuestro repositorio de GitHub.

Solo tenemos que copiar el contenido en una hoja de SQL de MySQL Workbench y ejecutar estos scripts. Con esto se generarán todos nuestros objetos de base de datos necesarios para poder ejecutar nuestra libreta de Colab.

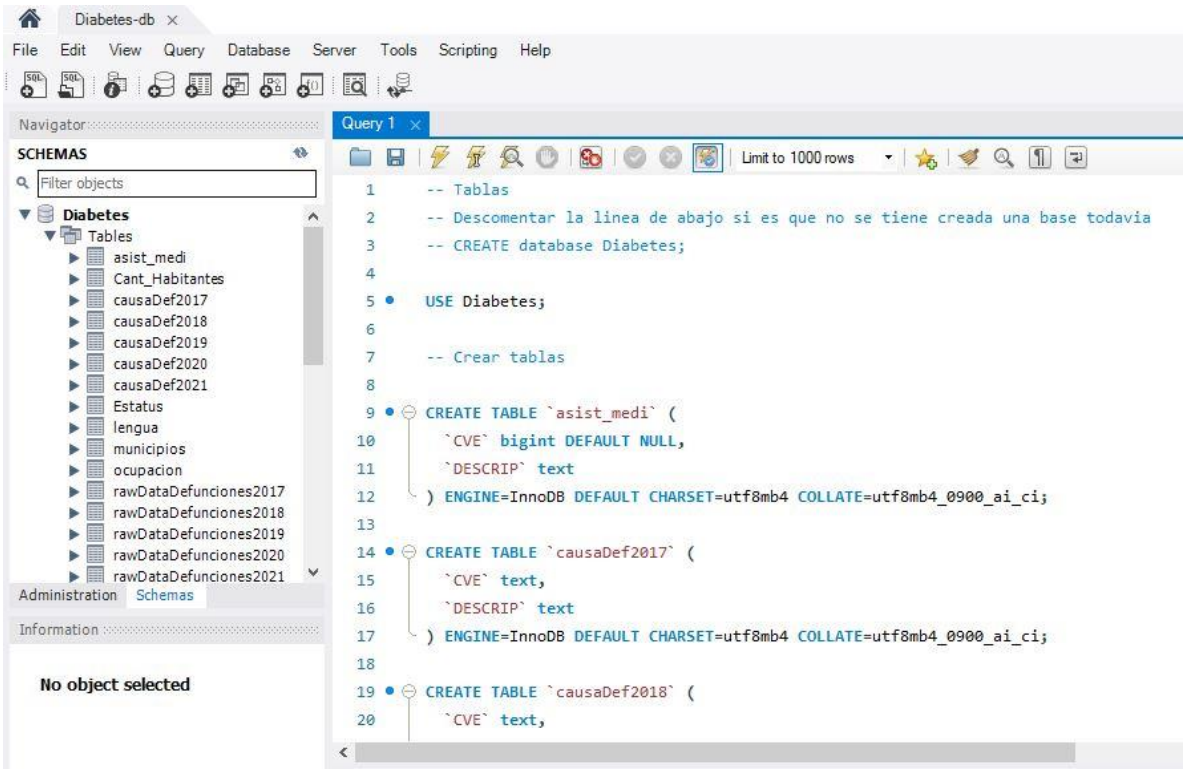

Fig. 2 Ejecución de scripts en MySQL Workbench.

Estos scripts generarán lo siguiente:

- Tablas para almacenar la información descargada.
- Tablas para almacenar la información ya procesada.
- Función para realizar la validación de nuestros procedimientos almacenados no se ejecuten nuevamente si la información ya se ha procesado.
- Registros en nuestra tabla de validación con el estatus de los procedimientos almacenados
- Procedimientos almacenados para realizar la limpieza de nuestros datos.
- Vistas para obtener la información que necesitamos de nuestras tablas.

Ejecución de libreta de Google Colab

Una vez que ya tenemos todos nuestros elementos anteriores, ya podremos ejecutar nuestra libreta de Colab.

Esta contiene lo siguiente:

- La descarga de nuestros datos de las diferentes fuentes utilizadas.
- La conexión a la base de datos mediante la librería SQLAlchemy para realizar la inserción de nuestros datos mediante los dataframes de Pandas.
- Ejecución de procedimientos almacenados y realización de consultas mediante la librería de pymysql.
- Generación de la gráfica con la mortalidad por diabetes por cada 100,000 habitantes por municipio del Estado de Sonora.

Gráfica de mortalidad por diabetes

Después de ejecutar nuestra libreta de Colab, se obtuvo la siguiente gráfica que nos muestra la mortalidad por diabetes en el Estado de Sonora.

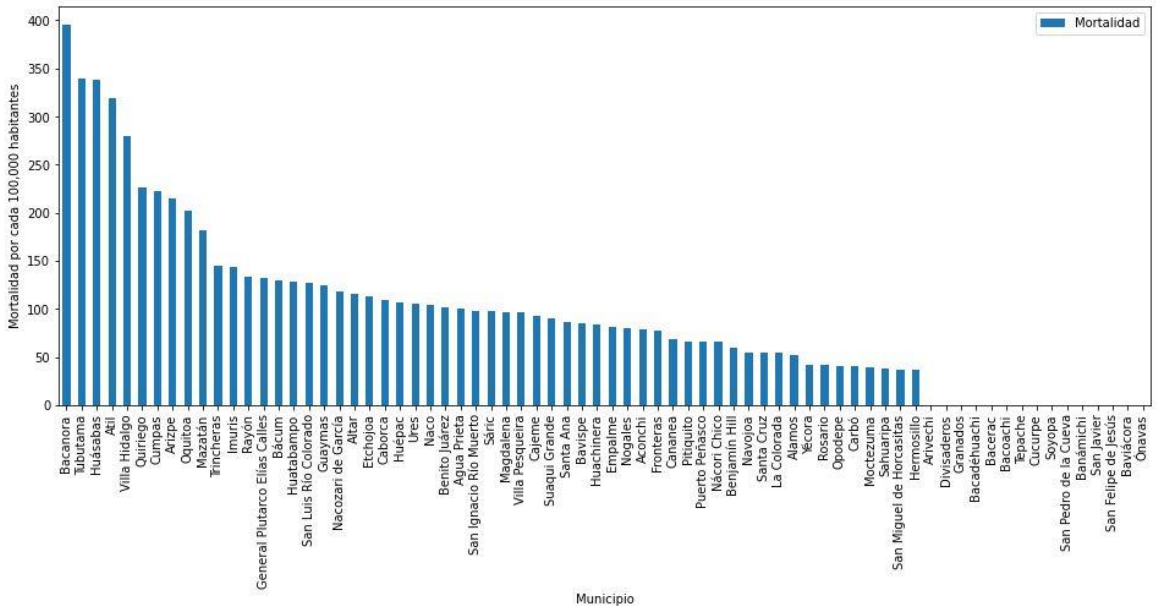

Fig. 3 Gráfica obtenida con información procesada en la base de datos.

Repositorio de GitHub

Toda la información necesaria para reproducir este proyecto, desde el manual para generar una instancia de MySQL hasta la libreta de Colab se encuentra disponible aquí.