Laboratorio de Datos - TP 1

Integrantes:
- Dorogov Cristina
- Pucciarelli Francisco Lautaro
- Salto Julian

Para poder correr los scripts de Python, es necesario:
- Python 3.6 o superior.

Que en el environment (donde ejecutan los scripts) tenga los módulos:
- Pandas
- Duckbd
- Seaborn
- MatPlotLib
- Pathlib (viene incluido en los paquetes básicos de Python)

Descripción General

La carpeta *TablasOriginales* contiene los DataSets originales.
- Padrón de Bibliotecas Populares.
- Padrón Oficial de Establecimientos Educativos 2022.
- Datos de Población por Departamento 2022 (DataSet del enunciado del TP).

El archivo llamado *Script_Main.py* contiene el código correspondiente a
- Limpieza y Procesamiento de Datos.
- Consultas SQL.
- Gráficos.

Está separado por secciones. Puede ejecutarse todo junto de corrido (esto puede tomar unos segundos… la carga de DataSets toma su tiempo, son archivos grandes), o ir ejecutándose por secciones. En caso de ir por la segunda opción, por favor ir en el orden en el que aparecen; ya que algunas secciones generan archivos que son utilizados en secciones posteriores.
Al ejecutar la totalidad del script, se generarán 3 carpetas:
- *TablasModelo* : aquí se almacenarán los DataSets (en formato CSV) producto de la limpieza y procesamiento.
- *ConsultasSQL* : aquí se almacenarán los resultados de las consultas SQL (en formato CSV).
- *Graficos_y_Visualizazciones* : aquí se almacenarán los gráficos.

El código está super comentado. No solo se encuentran los comentarios correspondiente a qué hace cada parte del código, sino también las distintas ideas y decisiones que se fueron tomando en el camino (los comentarios funcionaron como *cuaderno de laboratorio*, a lo largo del desarrollo de cada etapa).

El archivo *Verificaciones_Preliminares.py* no es necesario ejecutarlo. Simplemente tiene algunas verificaciones que fuimos realizando a medida que exploramos los DataSets. Además de algunas verificaciones posteriores, para ir viendo cómo eran los resultados obtenidos del proceso de limpieza de DataSets. No generan nada, simplemente imprimen en consola algunos mensajes y cosas que utilizamos para verificar. En caso que desee ejecutarlo, primero correr el archivo *Script_Main.py* (ya que trabaja sobre los DataSets originales y los DataSets limpios).

El archivo *Metricas_GQM.py* tampoco es necesario ejecutarlo. Al igual que el de *Verificaciones_Preliminares.py*, sólo tiene verificaciones que fuimos realizando para poder desarrollar el análisis de calidad de datos (sobre los DataSets originales). En caso que desee ejecutarlo, puede hacerlo en cualquier orden (trabaja sobre los DataSets originales). Mas solo va a imprimir algunos mensajes y variables en consola.

Finalmente, tenemos la carpeta llamada *Plan_B*. Esta carpeta contiene el mismo programa que *Script_Main.py*, pero separado en archivos individuales (como estaba originalmente). Dejamos esta carpeta en caso que haya algún problema con el archivo *Script_Main.py* (no manejamos muy bien la herramienta que separa el archivo en secciones… entonces podría fallar).
En caso de tener que recurrir a esta opción (suponiendo que falló el script original) los archivos deben ejecutarse en el siguiente orden:

1. Limpieza_DataSets_1.py
2. Limpieza_DataSets_2.py
3. Limpieza_DataSets_3.py
4. Consultas_SQL.py
5. Graficos.py

Aclaraciones y Comentarios Adicionales

Para el manejo automático de las rutas de los diferentes archivos, así como la generación automática de las carpetas (donde se almacenan los resultados de cada script), utilizamos herramientas del módulo *Pathlib* (incluido en los paquetes básicos de Python).

Los gráficos están configurados para generarse y guardarse en formato PNG (en la carpeta correspondiente). Sin embargo, hay 2 gráficos que, al guardarse como PNG, se generan vacíos. No sabemos exactamente por qué sucede, revisamos el código mil veces y no hallamos el bug que hace que se generen mal. De todos modos, si se ejecuta el script desde la IDE Spyder, en la sección de Gráficos (Plots) se generan todos los gráficos correctamente (por eso nos parece raro, todos los gráficos se generan de forma correcta; pero esos dos se exportan mal).