Laboratorio de Datos - TP 1.

Integrantes:

- Dorogov Cristina
- Pucciarelli Francisco Lautaro
- Salto Julian

Esta carpeta está pensada y preparada para que ingrese a cada uno de los scripts, los ejecute y listo. No es necesario completar `paths` ni añadir los DataSets originales.

Para poder correr los scripts de Python, solamente 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

Descripción general de lo que contiene la carpeta.

- 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).

- Los archivos titulados como `Limpieza_DataSets_X` contienen todo el tratamiento de limpieza que se hizo a cada DataSet original. Están separados ya que en cada archivo se trabaja un DataSets distinto. La idea era modularizar el trabajo sobre cada DataSet, para así aumentar la claridad y alivianar la carga sobre los archivos (son DataSets grandes en cuanto a cantidad de información). Estos archivos, al ejecutarlos, general una carpeta llamada `TablasModelo`. En ella se van a ir almacenando los DataSets limpios, que se corresponden con el Modelo Relacional planteado. **ESTOS ARCHIVOS DEBEN SER EJECUTADOS EN ORDEN DE NUMERACIÓN**: primero `Limpieza_DataSets_1`, luego `Limpieza_DataSets_2` y (por último) `Limpieza_DataSets_3`.

- El archivo `Consultas_SQL` tiene el trabajo de consultas SQL realizado sobre los DataSets limpios. Al ejecutar este archivo se genera una carpeta llamada `ConsultasSQL`, donde se almacenan (en formato *CSV*) los resultados de las consultas realizadas. Por favor, **ANTES DE EJECUTAR ESTE ARCHIVO, EJECUTAR TODOS LOS ARCHIVOS DE `Limpieza_Datasets_X`** (ya que las consultas trabajan sobre la carpeta que tiene los DataSets limpios).

- El archivo `Verificaciones_Preliminares` no es necesario ejecutarlo. Simplemente tiene algunas verificaciones que fuimos realizando a medida que explorábamos 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 cosas. **EN CASO QUE**

QUIERA EJECUTARLO, EJECUTAR TODOS LOS ARCHIVOS DE `Limpieza_DataSets` ANTES (ya que trabaja sobre los DataSets originales y los DataSets limpios).

- El archivo `Metricas_GQM` tampoco es necesario ejecutarlo. Al igual que el de `Verificaciones_Preliminares`, solo 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.

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).