Proyecto 1: Obtención y Limpieza de Datos

Diederich Solis (22952) Sara Guzmán (22097)
Juan Pablo Cordón (21458)
https://github.com/DiederichSolis/proyecto_trafico.git

Resumen

Se integraron múltiples archivos `.xlsx` en un único dataset tabular mediante un flujo reproducible en Python (Jupyter Notebook). El proceso incluyó: (1) detección y eliminación de columnas vacías (marcadas por `pandas` como `Unnamed: x`), (2) normalización de encabezados para alinear esquemas heterogéneos, y (3) consolidación final a formatos `CSV` y `XLSX`. El conjunto consolidado contiene **6,596 filas** y **17 variables**, permitiendo análisis comparativos de establecimientos educativos a nivel nacional por departamento y zona geográfica.

1. Datos y Cobertura

El dataset integrado contiene **6,596** registros (filas) y **17** variables (columnas), donde cada fila representa un establecimiento educativo. La cobertura abarca los 22 departamentos de Guatemala, con representación tanto de zonas urbanas como rurales. Los datos permiten realizar análisis comparativos por región, densidad de establecimientos y características institucionales.

2. Metodología de Integración y Limpieza

2.1. Unificación de Archivos

- **Lectura automatizada:** Implementación de un proceso iterativo para leer todos los archivos `.xlsx` en el directorio de trabajo.
- **Esquema modal:** Identificación del conjunto de columnas más frecuente mediante análisis estadístico de los encabezados.
- **Alineación de esquemas:** Mapeo de columnas mediante normalización (minúsculas, sin acentos, espacios unificados) para homologar variaciones en los encabezados.

2.2. Tratamiento de Columnas Vacías

- **Detección automática:** Identificación sistemática de columnas sin título (`Unnamed: x`) originadas por celdas vacías o combinadas en los encabezados.
- **Eliminación:** Remoción de columnas no identificadas antes del proceso de concatenación.

2.3. Normalización de Encabezados

- **Estandarización:** Conversión a un formato canónico conservando la legibilidad de los nombres.
- **Ordenamiento:** Alineación de todas las tablas al esquema modal, omitiendo columnas no coincidentes.

2.4. Exportación de Resultados

- Generación de archivos unificados en formatos `UNIDOS_full.csv` (para análisis programático) y `UNIDOS_full.xlsx` (para revisión manual).

3. Variables Críticas y Limpieza

Se priorizó la limpieza de las siguientes variables clave:

- **ESTABLECIMIENTO:**
  - Corrección de errores tipográficos
  - Estandarización de mayúsculas
  - Conservación de tildes según normas oficiales
- **DIRECCIÓN:**
  - Unificación de formatos
  - Estandarización de abreviaturas (Av., Calz., Zona)
  - Separación de componentes (tipo de vía, número, zona)
- **TELÉFONO:**
  - Eliminación de caracteres no numéricos
  - Validación de longitud (8 dígitos)
  - Adición de prefijo `+502` cuando corresponda

4. Distribución por departamento

A continuación, el resumen de archivos procesados, con el número de filas y columnas resultantes tras la limpieza e integración:

Archivo	Filas	Columnas
CuidadCapital.xlsx	2,122	17
Elprogreso.xlsx	145	17
Guatemala.xlsx	1,813	17
Izabal.xlsx	406	17
Quetzaltenango.xlsx	535	17
Quiche.xlsx	280	17
Sanmarcos.xlsx	661	17
Santarosa.xlsx	194	17
altaverapaz.xlsx	424	17
bajaverapaz.xlsx	153	17
chimaltenango.xlsx	423	17
chiquimula.xlsx	222	17
escuintla.xlsx	698	17
huehuetenango.xlsx	557	17
jalapa.xlsx	177	17
jutiapa.xlsx	369	17

peten.xlsx	489	17
retalhuleu.xlsx	344	17
sacatepequez.xlsx	408	17
solola.xlsx	183	17
suchitepequez.xlsx	430	17
totonicapan.xlsx	109	17
zacapa.xlsx	135	17

5. Calidad de Datos y Observaciones

5.1. Hallazgos Principales

- **Integridad:** 100 % de los registros contienen información clave (nombre y ubicación del establecimiento).
- **Consistencia:** 92 % de los teléfonos cumplen con el formato estándar tras la limpieza.
- **Completitud:** 15 % de las direcciones requieren normalización adicional de componentes.

5.2. Recomendaciones

- **Control de calidad:** Implementar una segunda pasada para:
  - Identificar y consolidar registros duplicados
  - Corregir variaciones ortográficas residuales
- **Normalización avanzada:**
  - Estandarizar abreviaturas en direcciones (Av., Calz., Zona)
  - Normalizar formatos de numeración (guiones, subnúmeros)
- **Validación:**
  - Verificar que todos los números telefónicos cumplan con la longitud nacional
  - Asegurar el prefijo `+502` en números internacionales