Limpieza de Datos de Establecimientos Educativos en Guatemala

Integrantes

- Flavio Galán
- Josue Say
- Isabella Miralles

Repositorio

- Enlace a GitHub
- No se trabajó google docs sino md en el repositorio en la carpeta docs/reporte

Proceso de Obtención de Datos

- Se desarrolló un script en Python utilizando Selenium para automatizar la descarga de datos desde el sitio del Ministerio de Educación.
- Se implementó un sistema de cacheo basado en el archivo links_cache.txt para evitar descargas repetidas.
- Los datos fueron almacenados en archivos .txt en formato Python.
- Se unificaron todos los datos en un único archivo .csv (data_unified.csv).

Estructura Original (previo a limpieza)

Estructura del Conjunto de Datos Crudo

En esta sección se presenta un análisis exploratorio del conjunto de datos unificado que contiene información de establecimientos educativos a nivel diversificado en Guatemala. El objetivo es comprender su composición inicial, identificar posibles irregularidades estructurales y sentar las bases para las futuras operaciones de limpieza.

Dimensiones Generales del Dataset

- **Total de filas (registros):** 6599
- **Total de variables (columnas):** 17
- **Filas completamente duplicadas:** 0
- **Valores nulos por columna:**
  - direccion: 2 valores nulos
  - telefono: 46 valores nulos

  - director: 26 valores nulos
  - Resto de columnas: 0 valores nulos
- **Tipo de datos por columna**: Todas las columnas son de tipo object (texto), incluso aquellas que podrían representar datos categóricos o numéricos.

Variables del Conjunto de Datos

A continuación se listan las variables que componen el dataset, junto con una breve descripción del tipo de información que contienen:

Variable	Descripción
codigo	Identificador único del establecimiento educativo. Formato compuesto por códigos departamentales, municipales y de nivel.
distrito	Código que representa el distrito educativo al que pertenece el establecimiento.
departamento	Nombre del departamento geográfico (ej. ALTA VERAPAZ).
municipio	Nombre del mu nicipio donde se encuentra ubicado el centro educativo.
establecimiento	Nombre oficial del establecimiento, que puede incluir variaciones en ortografía, uso de mayúsculas o símbolos.
direccion	Ubicación física del establecimiento, usualmente con calles, zonas o kilómetros.
telefono	Número telefónico de contacto. Presenta formato numérico pero se encuentra almacenado como texto.
supervisor	Nombre del supervisor a cargo del área o distrito educativo del establecimiento.
director	Nombre del director o responsable directo del establecimiento.
nivel	Nivel educativo ofrecido por el establecimiento (en todos los casos: DIVERSIFICADO).
sector	Define si el establecimiento es de carácter PRIVADO u OFICIAL.
area	Zona en la que opera el establecimiento: URBANA o RURAL.
status	Estado de funcionamiento del centro educativo (ej. ABIERTA).
modalidad	Modalidad lingüística o pedagógica (ej. MONOLINGÜE, BILINGÜE).
jornada	Jornada o turno en que se imparten las clases (ej. MATUTINA, VESPERTINA).
plan	Tipo de plan de estudios ofrecido (ej. DIARIO(REGULAR)).
departamental	Nombre del departamento, redundante con la columna departamento pero útil para verificar integridad.

Estrategia e Implementación del Proceso de Limpieza

Análisis Inicial

Antes de aplicar transformaciones específicas, se realizó un análisis exploratorio general del dataset para identificar:

- Presencia de valores vacíos explícitos ("") o simulados (espacios " ").
- Inconsistencias tipográficas o textuales.
- Variables candidatas a ser normalizadas o categorizadas.
- Posibles redundancias entre columnas como departamento y departamental.

Este análisis orientó las decisiones sobre estandarización, tratamiento de valores nulos, y generación de nuevas variables.

Establecimiento y Clasificación Tipológica

Se aplicó una doble estrategia a la columna establecimiento:

1. Limpieza y normalización textual:

- Conversión de todo el texto a **mayúsculas**.
- Eliminación de comillas, tildes y signos diacríticos.
- Reducción de múltiples espacios consecutivos.
- Corrección de errores ortográficos comunes.

2. Clasificación en tipos de establecimiento: Se creó la variable tipo_establecimiento mediante reglas basadas en coincidencias con palabras clave (COLEGIO, INSTITUTO, etc.) y análisis manual complementado por inteligencia artificial para resolver casos ambiguos. Esta clasificación permite filtrar y agrupar eficientemente los registros.

Supervisión y Dirección

Las columnas supervisor y director fueron limpiadas con el mismo enfoque:

- Normalización textual eliminando espacios redundantes y tildes.
- Conversión a mayúsculas.

Esto mejora la identificación y facilita futuros cruces institucionales.

Dirección

La columna direccion presentaba texto no estructurado con variaciones comunes como “km”, “zona”, entre otros. Se aplicaron las siguientes transformaciones:

- Limpieza básica y normalización.
- Reemplazo de abreviaciones comunes (KM → KILÓMETRO).
- Conversión a mayúsculas.

Esta estandarización facilita una futura georreferenciación o análisis espacial.

Teléfono

El campo telefono fue tratado como texto para evitar pérdidas de información. Se implementó:

- Eliminación de caracteres no numéricos.
- Validación para conservar solo teléfonos de 8 dígitos válidos.

Registros incompletos o inválidos fueron marcados como NaN.

Modalidad, Jornada, Plan, Status, Sector y Área

Estas variables presentaban inconsistencias tipográficas (espacios, mayúsculas/minúsculas). Se estandarizaron de la siguiente manera:

- Limpieza textual.
- Conversión a mayúsculas.
- Conversión a tipo category en los campos aplicables (status, sector, area).

Valores Nulos

Se reemplazaron todas las cadenas vacías ("") o compuestas solo por espacios por np.nan para asegurar una identificación correcta de valores faltantes.

En las columnas con datos faltantes (direccion, telefono, director), se imputaron valores explícitos para mejorar la legibilidad del dataset:

- direccion → "SIN DIRECCION"
- telefono → "SIN TELEFONO"
- director → "NO REGISTRADO"

Esto permite mantener registros completos sin eliminar información potencialmente útil.

Código Geográfico y Estructuración

Se diseñó un esquema para descomponer el campo codigo (cuando estaba presente) en sus componentes clave:

- codigo_departamento
- codigo_municipio
- codigo_establecimiento
- codigo_interno

La descomposición permitía construir un identificador único codigo_geografico, utilizado inicialmente como referencia en una tabla externa (codigos_geograficos.csv). Aunque se consideró mantener esta tabla como archivo auxiliar, se optó por **incorporar directamente los datos estructurados al dataset final**, dejando abierta la posibilidad de usar el archivo separado en futuras integraciones o validaciones externas.

Adicionalmente, se extrajo el componente codigo_distrito y se eliminaron columnas redundantes.

Departamento vs. Departamental

El análisis evidenció que las columnas departamento y departamental no son equivalentes:

- departamento representa la división político-administrativa oficial de Guatemala.
- departamental refleja subdivisiones operativas internas de gestión educativa.

Por tanto, se decidió **conservar ambas columnas**, ya que aportan información distinta y valiosa según el enfoque analítico.