Avance del Proyecto: Limpieza de Datos de Establecimientos Educativos de Nivel Diversificado en Guatemala

Entregable: Análisis del Estado Inicial del Conjunto de Datos

Este avance tiene como objetivo entregar una descripción detallada del conjunto de datos crudos descargado, identificar variables que requerirán mayor atención durante la limpieza, y definir una estrategia de limpieza sin implementarla aún.

Componentes del Avance

Descarga y Unificación de los 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 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 Propuesta para la Limpieza

Antes de aplicar las transformaciones específicas por variable, **se realizó un análisis exploratorio general de valores nulos**, identificando columnas con valores vacíos reales o simulados (como cadenas vacías " " o espacios " "). Como primer paso de limpieza, **se reemplazaron todos estos casos por valores nulos reales (np.nan)**, con el objetivo de facilitar su posterior tratamiento, análisis de calidad y filtrado.

A continuación, se detalla la estrategia específica aplicada por variable:

Variable(s)	Problema Detectado	Operación de Limpieza Propuesta	Justificación
establecimiento	Inconsistencias tipográficas: uso mixto de mayúsculas/minúsculas, dobles espacios, comillas, y abreviaciones	- Normalizar a MAYÚSCULAS.- Eliminar dobles espacios.- Estandarizar comillas.- Crear nueva columna tipo_establecimiento a partir de patrones comunes (ej. “Colegio”, “Instituto”)	Mejora consistencia textual y permite clasificar o agrupar fácilmente por tipo de centro educativo.
direccion	Texto no estructurado con errores ortográficos y abreviaciones irregulares (ej. “km”, “zona”)	- Corregir espacios.- Estandarizar abreviaciones comunes para guatemala (km → Km, zona → Zona, etc.).- Normalizar formato general	Facilita georreferenciación futura, análisis espacial y cruce con mapas.

Variable(s)	Problema Detectado	Operación de Limpieza Propuesta	Justificación
telefono	Verificación de formato inconsistente, con caracteres no numéricos y longitudes incorrectas	- Validar longitud de 8 dígitos.- Normalizar como texto limpio	Permite validar teléfonos, contactar a centros o analizar cobertura de comunicación.
departamento, departamen-tal	Campos redundantes. Posibles diferencias tipográficas	- Verificar coincidencia exacta entre ambos.- Si son equivalentes, eliminar departamental	Elimina duplicidad innecesaria y refuerza integridad de la variable geográfica.
codigo	Campo compuesto que incluye subcódigos concatenados	- Separar en columnas nuevas: codigo_departamento, codigo_municipio, codigo_establecimiento, codigo_interno (si aplica).- Eliminar campo original codigo	Facilita segmentación, análisis por región y comprensión jerárquica del código.
supervisor, director	Nombres con inconsistencias en tildes, uso de mayúsculas o espacios extra	- Normalizar: quitar espacios innecesarios, capitalizar correctamente, corregir tildes comunes	Mejora identificación de responsables y su cruce con otros registros institucionales.
modalidad, jornada, plan, status, sector, area	Variaciones tipográficas en valores iguales (ej. “MATUTINA”, “matutina”, ” MATUTINA ”)	- Estandarizar todos los valores en MAYÚSCULAS.- Eliminar espacios innecesarios	Asegura agrupamientos correctos al analizar por categoría o filtro.
General (todas las columnas)	Presencia de " " o cadenas vacías que aparentan ser datos válidos	- Reemplazar por valores NaN o nulos reales (np.nan)	Mejora la detección de valores faltantes y permite análisis precisos de calidad de datos.
Todo el dataset	Todos los campos son tipo object (texto), incluso numéricos o categóricos	- Mantener telefono como texto limpio.- Evaluar tipado adecuado para análisis posterior: convertir valores categóricos (status, sector, etc.) y códigos numéricos donde aplique	Permite eficiencia en análisis estadístico, creación de visualizaciones, y compatibilidad con otros sistemas.

Implementación de operaciones de limpieza

Establecimiento La limpieza de la columna establecimiento se realizó en dos etapas complementarias:

1. Limpieza básica y normalización textual:

Se aplicaron transformaciones orientadas a garantizar consistencia y eliminar caracteres no deseados:

- Eliminación de comillas simples y dobles (" y ').
- Sustitución de múltiples espacios por uno solo.
- Eliminación de comas al final de cadenas.
- Conversión de todo el texto a **mayúsculas**.
- Normalización de caracteres Unicode: se eliminaron tildes y signos diacríticos

(ej. É → E, Ñ → N).

Esto garantiza que cadenas visualmente iguales sean comparables computacionalmente, evitando errores por diferencias ortográficas o de codificación.

2. Clasificación tipológica (tipo_establecimiento):

A partir de la columna normalizada, se creó una nueva variable tipo_establecimiento mediante una función que identifica patrones y aplica reglas específicas:

- Corrección de errores comunes en palabras clave iniciales (ej. COLEGO → COLEGIO).
- Asignación condicional según coincidencias con substrings clave (COLEGIO, INSTITUTO, SCHOOL, COLLEGE, etc.).
- Inclusión de **excepciones** y **casos especiales** con cadenas exactas para evitar falsos positivos.
- Creación de una categoría especial "CORPORACION" si la cadena inicia con "CORPORACION".

Además, para una clasificación más precisa, se realizó un análisis manual complementado con la ayuda de una inteligencia artificial (ChatGPT) para explorar nombres de entidades, identificar patrones contextuales, y decidir correctamente el tipo real del establecimiento. Este proceso se documentó y desarrolló en este enlace.

Código Para trabajar con la columna codigo, primero se realizó una **normalización previa de los campos geográficos** (departamento, municipio y departamental) con el fin de garantizar consistencia textual. Esta limpieza eliminó tildes, diéresis, la letra ñ y otros signos diacríticos, transformando los valores a mayúsculas y eliminando espacios redundantes. Esta estandarización fue clave no solo para el análisis posterior, sino también para facilitar la comparación futura entre departamento y departamental, tarea destinada a otro integrante del equipo.

Una vez limpia la base, se descompuso el código estructurado (XX-YY-ZZZZ-WW) en partes clave:

- XX: código del departamento.
- YY: código del municipio.
- ZZZZ: código del establecimiento.
- WW: código interno adicional.

Los códigos XX y YY se utilizaron en combinación con los nombres limpios de departamento y municipio para construir un identificador único denominado codigo_geografico, almacenado en un archivo externo (codigos_geograficos.csv) y

luego reintegrado al dataset principal. Esto permite vinculación territorial precisa, trazabilidad geográfica y eliminación de ambigüedad entre municipios que comparten códigos repetidos bajo distintos departamentos.

Adicionalmente, se extrajo el segundo segmento del campo distrito (YY) y se renombró como codigo_distrito, ya que representa un identificador municipal útil y el campo original fue descartado. Finalmente, se eliminaron las columnas codigo, departamento, municipio y distrito, por ser reemplazadas o integradas en nuevas variables con mayor estructura y control de calidad.

Aquí tienes la sección **completa y redactada** para el título **"Implementación de operaciones de limpieza"**, basada fielmente en el código proporcionado:

Limpieza de nombres de personas (supervisor, director) Se aplicó una limpieza orientada a estandarizar nombres propios:

- Eliminación de espacios sobrantes.
- Capitalización con estilo título (.title()).
- Eliminación de tildes y signos diacríticos mediante normalización Unicode (NFKD).

Esto asegura la consistencia visual y semántica de nombres, clave para posteriores cruces con bases institucionales.

Estandarización de variables categóricas (modalidad, jornada, plan, status, sector, area) Se homogenizaron los valores de estas variables:

- Conversión a **mayúsculas**.
- Eliminación de espacios innecesarios.
- Conversión final del tipo de datos a category en las columnas status, sector y area, optimizando espacio y análisis.

Tratamiento de valores vacíos en todo el dataset Se reemplazaron todas las cadenas vacías ("") por valores nulos reales (np.nan), permitiendo un análisis confiable de valores faltantes.

Corrección del campo direccion Se normalizó el texto en esta columna para facilitar su análisis y posibles procesos de georreferenciación:

- Conversión a **minúsculas**.
- Eliminación de espacios sobrantes.
- Reemplazo de abreviaciones comunes: km → kilómetro, zona → zona (mantenido para uniformidad).

Validación del campo telefono Para asegurar que los números telefónicos sean válidos:

- Eliminación de cualquier carácter no numérico.
- Verificación de longitud: solo se conservaron números de **8 dígitos**, el formato oficial en Guatemala. Los valores inválidos se marcaron como NaN.

Verificación y eliminación de redundancia (departamento vs. departamental) Se compararon ambas columnas para detectar duplicidad:

- Si sus valores coincidían exactamente, se eliminó departamental por ser redundante.

Conversión de la columna codigo a formato numérico Se intentó convertir codigo a tipo numérico (int64). En caso de error, se asignó NaN. Esto prepara el campo para su posterior descomposición estructural.

Guardado del conjunto limpio Finalmente, el dataset transformado se guardó como archivo CSV con el nombre data/df_limpio.csv.