Proyecto 1. Obtención y Limpieza de datos

Code Book

Agosto

GRUPO 1

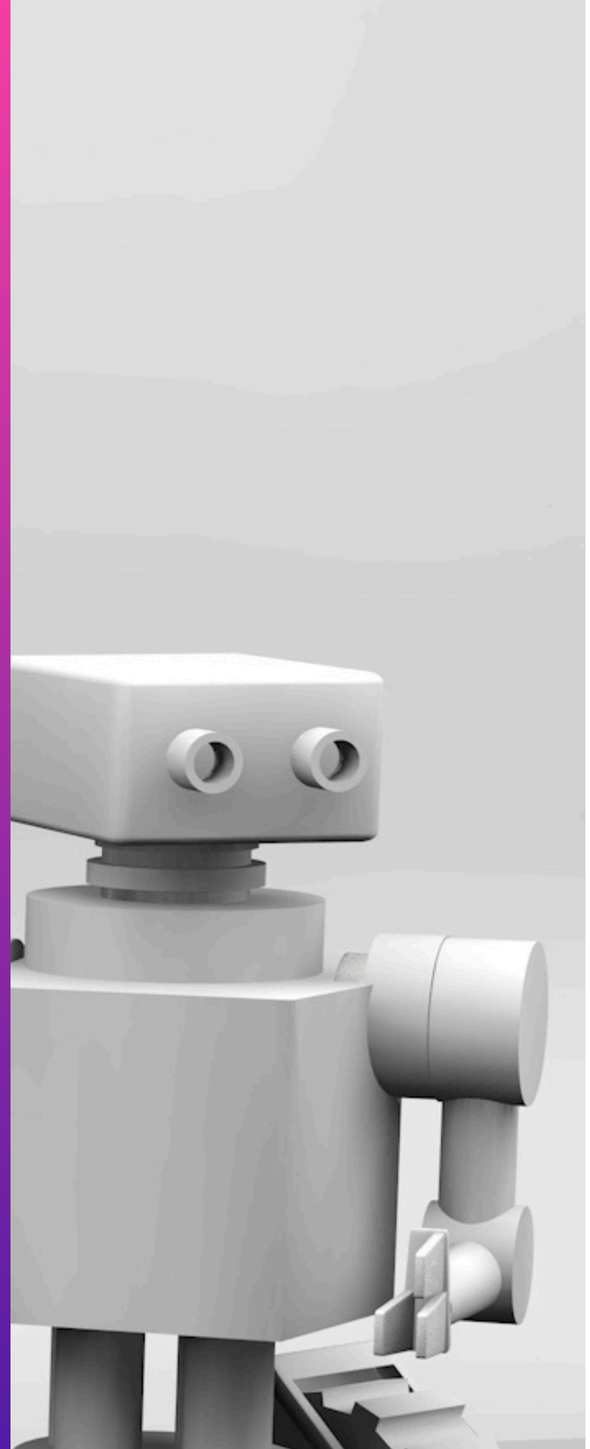
Descripción de los Datos y Problemas Iniciales

Descripción de los Datos

- **Origen:** Portal del Ministerio de Educación (MINEDUC) de Guatemala.
- **Contenido:** Información sobre instituciones educativas hasta el nivel diversificado.
- **Campos Principales:** Código único, departamento, municipio, nombre del establecimiento, dirección, teléfono.
- **Fecha de Extracción:** 5 de agosto de 2024.

Problemas Iniciales Encontrados

- Rendimiento del Sitio: Caídas frecuentes durante la descarga.
- Formato Inusual: Archivos en formato .html.xls.
- Tamaño de los Datos: Advertencia de datos pesados, dificultando la descarga completa.
- Solución: Algoritmo en Python para extraer y convertir datos a CSV.



Proceso de Limpieza

- **Corrección de Texto:**
  - Tildes corregidas en campos de texto (ej. "DOS DÍAS A LA SEMANA").
- **Manejo de Valores Faltantes:**
  - Reemplazo de NaN en columnas clave por "NO ESPECIFICADA".
- **Filtrado de Nivel Escolar:**
  - Se mantuvieron solo registros de "DIVERSIFICADO".
- **Estándarización de Nombres:**
  - Eliminación de caracteres no deseados, conversión a mayúsculas y limpieza de espacios en DIRECTOR y SUPERVISOR.
- **Limpieza de Teléfonos y Direcciones:**
  - Normalización de formato en TELEFONO.
  - Extracción de columnas adicionales en DIRECCION (ej. TIPO_VIA, ZONA).
- **Validación de Formato:**
  - Asegurar formatos correctos en DISTRITO y CODIGO.
- **Verificación de Duplicados:**
  - Confirmación de que no hay registros duplicados.

Resultados y Unificación

Variables Principales:

El conjunto de datos incluye 22 variables esenciales, como Código, Distrito, Departamento, Municipio, Establecimiento, y más. Cada una de estas variables ha sido cuidadosamente revisada y estandarizada para asegurar la integridad y consistencia de los datos.

Proceso de Unificación y Limpieza:

Todos los registros de los establecimientos educativos fueron unificados en un solo archivo CSV. Se realizaron procesos de limpieza detallados para corregir inconsistencias, como valores faltantes, errores de formato y duplicados, garantizando que el conjunto de datos esté listo para su análisis posterior.

Conclusión

El proceso de unificación y limpieza de datos fue crucial para asegurar la calidad y coherencia de la información. Esto no solo facilitó la generación de un conjunto de datos limpio y preparado para el análisis, sino que también permitió la elaboración de un detallado codebook, esencial para comprender y utilizar eficazmente las variables del conjunto de datos.

¡Muchas gracias!