Guatemala, julio 11 de 2025	Data Science.
Michelle Mejía	Carnet. 22596
Silvia Illescas	Carnet. 22376
Emilio Reyes	Carnet. 22674

Avances del Proyecto 1

Descripción general del dataset.

El conjunto de datos utilizado en este proyecto corresponde a los establecimientos educativos de Guatemala que imparten hasta el nivel diversificado, recopilado a partir del portal del Ministerio de Educación (MINEDUC). El dataset original fue descargado y consolidado en un único archivo .csv que contiene la información de todos los departamentos del país.

Hablando de las dimensiones del dataset, tenemos un total de 6,326 registros, ordenados en 17 columnas diferentes que representan las 17 variables de cada registro. A continuación, se detalla cada variable con su descripción.

Variable	**Descripción**	**Variable**	**Descripción**
Código	Código único del establecimiento	Distrito	Identificador del distrito educativo
Departamento	Departamento al que pertenece el establecimiento	Municipio	Municipio de ubicación del establecimiento
Establecimiento	Nombre del centro educativo	Dirección	Dirección física del establecimiento
Teléfono	Número de contacto telefónico	Supervisor	Nombre del supervisor asignado
Director	Nombre del director del centro educativo	Nivel	Nivel educativo (Diversificado)
Sector	Sector al que pertenece el establecimiento	Área	Área geográfica (urbana o rural)
Status	Estado de funcionamiento (abierta, cerrada, etc.)	Modalidad	Modalidad de enseñanza (monolingue, bilingue, etc.)
Jornada	Jornada en la que se imparten clases	Plan	Tipo de plan educativo

Departamental	Nombre del departamento (duplicado)		

Con esto en mente, a priori, las variables que necesitan una limpieza son las siguientes, por las razones correspondientes.

Variable	**Problema**
Establecimiento	Variaciones en mayúsculas o minúsculas, comillas, errores tipográficos y posibles duplicados.
Dirección	Uso inconsistente de abreviaturas como “Av.”, “calle”, caracteres especiales y formatos distintos.
Teléfono	Formatos no estandarizados o datos faltantes.
Municipio	Inconsistencia en nombres que se refieren al mismo lugar (“Guatemala” y “Ciudad Capital”)
Supervisor y Director	Diferencias por tildes, mayúsculas o minúsculas, y errores tipográficos.
Departamental	Variable redundante
Extra	Eliminar aparentes valores nulos que aparecen como “,,,,,”

Con eso en mente, se describe el plan de limpieza para las variables que presentan mayor nivel de inconsistencia en el dataset, para dejar el conjunto de datos lo más uniforme y coherente posible.

Establecimiento:

- Convertir todo a mayúsculas para facilitar comparaciones.
- Eliminar comillas dobles, signos de puntuación y espacios dobles.
- Revisar errores ortográficos o variaciones mínimas en los nombres.
- Identificar y marcar establecimientos duplicados con nombres ligeramente diferentes.

Dirección:

- Convertir todo a mayúsculas.
- Unificar abreviaturas (“AV” por “AVENIDA”, “Z.” por “ZONA”, etc.).
- Quitar espacios múltiples y tildes innecesarias.
- Establecer formato con base común (número-calle-zona).

Teléfono:

- Eliminar caracteres no numéricos.
- Validar que todos los números tienen 8 dígitos.

Municipio:

- Unificar nombres de municipios donde haya ambigüedad. Por ejemplo, “Ciudad Guatemala” se refiere a zonas, “Guatemala” se refiere a Mixco, Chinautla.. etc.
- Eliminar espacios extras o diferencias de formato.

Supervisor y Director:

- Convertir a mayúsculas.
- Eliminar tildes para evitar duplicidades.
- Arreglar errores ortográficos evidentes.

Departamental:

- Comparar su contenido con la variable “Departamento”
- Si es redundante, eliminar la variable.