Ignacio Méndez Alvarez
Pablo Herrera Barbales
Diego Soto Flores

14 de julio de 2025

Proyecto 1 – Data Science

Descripción de los datos:

El set de datos a utilizar está compuesto por 23 documentos de Excel, que posteriormente se combinaran en un mismo set de datos. Existe un documento por cada departamento de Guatemala (22) y un documento extra para la Ciudad Capital (1), lo que completa los 23 documentos. Cada documento incluye información sobre los centros educativos de Nivel Diversificado del departamento correspondiente. El set de datos fue extraído del Ministerio de Educación (MINEDUC).

En la parte superior de cada archivo, se incluyen las instrucciones para la obtención de datos, que se muestran en la página web del MINEDUC, de donde se extrajo la información. Estas instrucciones se eliminarán a la hora de realizar la carga de datos. Asimismo, en la parte inferior se encuentra información sobre los derechos de autor en el pie de página, la cual también se eliminará al cargar los datos.

El set de datos total tiene una cantidad de 6600 filas, distribuida por departamento de la siguiente forma:

- Alta Verapaz (294)
- Baja Verapaz (94)
- Chimaltenango (304)
- Chiquimula (136)
- Ciudad Capital (867)
- El Progreso (97)
- Escuintla (393)
- Guatemala (1038)
- Huehuetenango (295)
- Izabal (273)
- Jalapa (121)
- Jutiapa (296)
- Petén (270)
- Quezaltenango (365)
- Quiché (184)
- Retalhuleu (272)
- Sacatepéquez (208)
- San Marcos (432)
- Santa Rosa (133)
- Sololá (111)
- Suchitepéquez (296)
- Totonicapán (51)
- Zacapa (70)

Por parte de las columnas, el set de datos cuenta con las siguientes 17 variables:

- Código: código de 10 dígitos asignado al establecimiento.
- Distrito: código del distrito en el que se encuentra el establecimiento.
- Departamento: nombre del departamento en el que se encuentra el establecimiento. En el caso de la Ciudad Capital, se toma como otro departamento.
- Municipio: nombre del municipio en el que se encuentra el establecimiento. En el caso de la Ciudad Capital, se muestra la zona en la que se encuentra el establecimiento.
- Establecimiento: nombre del establecimiento.
- Dirección: dirección del establecimiento.
- Teléfono: teléfono de contacto del establecimiento.
- Supervisor: nombre del supervisor del establecimiento.
- Director: nombre del director del establecimiento.
- Nivel: nivel escolar del establecimiento. En este caso todos son Nivel Diversificado.
- Sector: tipo del establecimiento.
- Área: área en la que se encuentra el establecimiento (Urbana o Rural).
- Status: estado de funcionamiento del establecimiento.
- Modalidad: modalidad lingüística del establecimiento.
- Jornada: jornada de funcionamiento del establecimiento.
- Plan: plan de funcionamiento del establecimiento.
- Departamental: región del departamento en la que se encuentra el establecimiento.

Las variables que más necesitan limpieza son:

- Establecimiento
- Dirección
- Teléfono
- Supervisor
- Director
- Plan

Estrategias:

- Para “Código” / “Distrito”:
    - Verificar que se cumpla el formato
- Para “Establecimiento”:
    - Verificar uniformidad de mayúsculas
    - Eliminar distintos tipos de comillas al principio o final de nombres
    - Verificar duplicados
    - Verificar que no haya variaciones en los nombres
    - Revisar errores ortográficos
- Para “Dirección”:
    - Unificar variaciones de abreviaciones (como AVENIDA, AVE y AV)
    - Eliminar caracteres especiales innecesarios
    - Eliminar duplicaciones en número y letras
    - Verificar formato
- Para “Teléfono”:

    - Rellenar valores vacíos o dejarlos como nulo
    - Verificar la longitud
    - Verificar existencia de letras o símbolos
- Para “Supervisor” / “Director”:
    - Verificar uniformidad de mayúsculas
    - Revisar errores ortográficos
    - Verificar existencia de nombres duplicados con errores leves
- Para “Plan”:
    - Unificar categorías parecidas, si se puede
    - Eliminar paréntesis y espacios innecesarios