UNIVERSIDAD DEL VALLE DE GUATEMALA

Facultad de Ingeniería

Data Science – Mario Barrientos

Proyecto 1

Francis Aguilar – 22243

Angela García – 22869

Cesar Lopez - 22535

Repositorio: https://github.com/angelargd8/Proyecto1-ds.git

Descripción set de datos:

Este conjunto de datos contiene información sobre los centros educativos autorizados por el Ministerio de Educación (MINEDUC) de Guatemala, así como los niveles y planes de estudio que están autorizados a impartir. Puede utilizarse para consultar detalles administrativos, geográficos y académicos de cada establecimiento registrado.

Variable	Descripción
1. CODIGO	Código único del establecimiento autorizado por el MINEDUC.
2. DISTRITO	Código que identifica el distrito escolar al que pertenece el centro educativo.
3. DEPARTAMENTO	Nombre del departamento de Guatemala donde se ubica el establecimiento.
4. MUNICIPIO	Nombre del municipio correspondiente dentro del departamento.
5. ESTABLECIMIENTO	Nombre oficial del centro educativo
6. DIRECCION	Dirección física del establecimiento, que puede información más detallada
7. TELEFONO	Número telefónico registrado para el establecimiento.
8. SUPERVISOR	Nombre del supervisor designado por el MINEDUC
9. DIRECTOR	Nombre del director o responsable del centro educativo.
10. NIVEL	Nivel educativo que imparte el establecimiento (ej. Preprimaria, Primaria, Secundaria, Diversificado).
11. SECTOR	Tipo de gestión del establecimiento: privado u oficial
12. AREA	Área geográfica del establecimiento: urbana o rural.
13. STATUS	Estado de funcionamiento del establecimiento: abierta, cerrada, suspendida, etc.
14. MODALIDAD	Modalidad de enseñanza: monolingüe, bilingüe, intercultural, etc.
15. JORNADA	Jornada en la que opera el centro educativo: matutina, vespertina, nocturna, jornada extendida, etc.
16. PLAN	Plan de estudios autorizado que se imparte (por ejemplo: diario, fin de semana, entre otros).

17. DEPARTAMENTAL	Confirmación del departamento educativo al que pertenece (puede coincidir con "DEPARTAMENTO" o usarse como campo administrativo complementario).

Variables con más operaciones de limpieza:

Variable	Explicación
18. TELEFONO	Porcentaje alto de nulos, la columna es object, aunque la mayoria de sus datos parecen numeros, aparecen muchos formatos mixtos, como guiones o con extensiones.
19. DIRECTOR	Datos nulos, ademas hay muchas variaciones ortograficas, como acentos, mayusculas y pueden haber duplicados escritos de forma diferente
20. DIRECCION	Datos nulos, hay muchas abrebiaturas y mayusculas y minusculas que lo hacen inconsistentes, tambien tiene acentos.
21. ESTABLECIMIENTO	Datos nulos, demasiadas categorias, tambien se abrevian muchas palabras y nuevamente acentos. Al igual que hay nombres que tienen muchas comillas.
22. CODIGO	Datos nulos, muchos valores unicos, el largo de los numeros o el numero de digitos es inconsistente.
23. SUPERVISOR	Pocos nulos pero mezcla de abreviaturas de cargo y nombre.
24. STATUS	Verificar que no exista otro tipo de nombre para el status
25. SECTOR	Verificar que no exista otro tipo de nombre para el sector
26. PLAN	Verificar que no exista otro tipo de nombre para un mismo plan
27. NIVEL	Verificar que no exista otro tipo de nombre para diversificado

Estrategia de limpieza:

- Código
    - Revisar que todos los valores sean únicos

    - Revisar valores duplicados, validar que el resto de información sea duplicada también
- Distrito
    - Validar que el formato sea válido
    - Validar que el distrito haga sentido con el departamento
- Departamento
    - Convertir todo a mayúsculas
    - Validar que todo esté escrito sin tildes para que no haya datos diferentes
    - Quitar espacios al inicio y final
    - Validar que haya 23 datos únicos correspondientes a los 23 departamentos de Guatemala
- Municipio
    - Validar que el municipio sea del departamento
    - Convertir todo a mayúsculas
    - Validar ortografía
- Establecimiento
    - Validar caracteres especiales
    - Quitar comillas dobles de más
- Dirección
    - Convertir todo a mayúsculas o minúsculas para estandarizar.
    - Eliminar espacios al inicio o final.
    - Revisar errores ortográficos o cambios de letras
    - Unificar nombres que se refieren al mismo lugar.
    - Eliminar duplicados
- Teléfono
    - Eliminar guiones
    - Validar longitud
    - Revisar datos faltantes
- Supervisor

    - Convertir todo en mayúsculas
    - Unificar formato (nombres, apellidos)
    - Verificar caracteres especiales
    - Verificar diferentes que se refieren a la misma persona
    - Revisar ortografía
- Director
    - Convertir todo en mayúsculas
    - Unificar formato (nombres, apellidos)
    - Verificar caracteres especiales
    - Verificar diferentes que se refieren a la misma persona
    - Revisar ortografía
- Nivel
    - Todo en mayúsculas
    - Verificar que solo sea hasta diversificado
    - Validar ortografía
- Sector
    - Verificar que correspondan a los sectores válidos
    - Convertir todo en mayúsculas
    - Validar ortografía
- Área
    - Estandarizar nombres
    - Convertir todo a mayúsculas
    - Validar ortografía
- Status
    - Estandarizar nombres
    - Ver que corresponda a los valores válidos
- Modalidad

    - Ver que corresponda a valores válidos
    - Estandarizar nombres
- Jornada
    - Unificar términos que se refieren a lo mismo
    - Validar ortografía
    - Estandarizar nombres
- Plan
    - Estandarizar nombres
- Departamental
    - Corregir nombres de departamentos mal escritos.
    - Convertir en mayúsculas
    - Validar ortografía

Libro de codigos

Nombre	Descripción	Tipo de dato	Unidades	Valores posibles valores especiales /faltantes	Estadísticas básicas	Fuente original	Transformaciones	Notas
Codebook de importaciones								
CODIGO	Código único del establecimiento autorizado por el MINEDUC	Numérico / Texto	N/A	Valores únicos por establecimiento; sin duplicados 0	Cantidad total de establecimientos, % duplicados	MINEDUC	Normalizar a texto	Identificador principal
DISTRITO	Código que identifica el distrito escolar	Texto	N/A	Códigos alfanuméricos según MINEDUC	Conteo por distrito	MINEDUC	Validar consistencia de códigos	Puede usarse para agrupar por distrito
DEPARTAMENTO	Nombre del departamento	Texto	N/A	22 departamentos de Guatemala	Conteo por departamento	MINEDUC	Estándarizar nombres	Útil para análisis geográfico
MUNICIPIO	Nombre del municipio	Texto	N/A	Todos los municipios de Guatemala	Conteo por municipio	MINEDUC	Estándarizar nombres	Útil para análisis geográfico
ESTABLECIMIENTO	Nombre oficial del centro educativo	Texto	N/A	Cualquier nombre registrado	Conteo total, % duplicados	MINEDUC	Normalizar mayúsculas/minúsculas	Normalizar mayúsculas/minúsculas

DIRECCION	Dirección física del establecimiento	Texto	N/A	Cualquier dirección registrada Valores alfanuméricos	Validar longitud, % faltantes	MINEDUC	Limpiar caracteres especiales	Algunas direcciones pueden estar incompletas
TELEFONO	Número telefónico registrado	Integer	N/A	Formatos numéricos, no pueden incluir guiones o espacios	Valores únicos	MINEDUC	Uniformizar formato, validar los que no están disponibles	En las transformaciones se valida que los numeros que no son de longitud 8 se colocan como no disponibles
SUPERVISOR	Nombre del supervisor designado	Texto	N/A	Nombres únicos o pueden ser repetidos	Validar longitud, % faltantes	MINEDUC	Normalizar	
DIRECTOR	Nombre del director		N/A	Nombres únicos o pueden ser repetidos	Validar longitud, % faltantes	MINEDUC	Normalizar	
NIVEL	Nivel educativo que imparte el establecimiento	Texto	N/A	Diversificado	Conteo por nivel	MINEDUC	Normalizar nombres	
SECTOR	Si es privado o público el establecimiento	Texto	N/A	Privado Oficial	Conteo	MINEDUC	Normalizar nombres	
AREA	Area demografica	Texto	N/A	Urbana	Conteo por área	MINEDUC	Normalizar los datos y ver datos faltantes	

STATUS	Estado del establecimiento	Texto	N/A	Abierta o cerrada	Conteo por area	MINEDUC	Norallizar datos y ver faltantes	
MODALIDAD	Modalidad de enseñanza	Texto	N/A	MONOLINGUE BILINGÜE	Conteo por modalidad	MINEDUC	Normalizar datos y ver faltantes	
JORNADA	Jornada en la que opera el centro	Texto	N/A	matutina', 'vespertina', 'doble', 'nocturna', 'sin jornada', 'intermedia'	Conteo por modalidad	MINEDUC	Normalizar datos, validar datos faltantes	No se encuentran faltantes para estos datos
PLAN	Plan de estudios autorizado	Texto	NA	diario(regular), 'fin de semana', 'a distancia', 'semipresencial', 'semipresencial (fin de semana)','semipresencial (un dia a la semana)', 'virtual a distancia', 'semipresencial (dos dias a la semana)', 'sabatino', 'intercalado', 'dominical', 'mixto	Conteo por plan para estadística	MINEDUC	Normalizar datos ver faltantes y faltas ortograficas	
DEPARTAMENTAL	Confirmación de departamento al que pertenece el centro educativo	Texto	NA	Departamentos de Guatemala	Conteo por departamento	MINEDUC	Normalizar los datos y validar faltantes	