Universidad del Valle de Guatemala
Data Science
Javier Prado – 21486
Bryan España - 21550
Luis Monterroso – 21699
Angel Herrarte - 22873

Proyecto 1
Libro de código

Descripción general del conjunto de datos:

Análisis de la infraestructura educativa de Guatemala, sus niveles y modalidades. Las fuentes de información fueron extraídas del portal oficial del Ministerio de Educación de Guatemala (MINEDUC), en el cual eran tablas HTML exportadas desde su sitio web.

Diccionario de variables

Variable	Tipo	Descripción	Valores / Formato
CODIGO	String	Identificador único del establecimiento educativo	formato NN-NN-NNNN-NN
DISTRITO	String	Código del distrito educativo	Formato NN-NNN
DEPARTAMENTO	String	Nombre del departamento político-administrativo	EL PROGRESO
MUNICIPIO	String	Nombre del municipio dentro del departamento	GUASTATOYA
ESTABLECIMIENTO	String	Nombre oficial del centro educativo	Texto libre

Universidad del Valle de Guatemala
Data Science
Javier Prado – 21486
Bryan España - 21550
Luis Monterroso – 21699
Angel Herrarte - 22873

DIRECCION	String	Dirección postal completa	Texto libre
TELEFONO	int	Número de contacto	9450881 (solo dígitos)
SUPERVISOR	String	Nombre completo del supervisor académico	Texto libre
DIRECTOR	String	Nombre completo del director o rector	Texto libre
NIVEL	Categórica	Nivel educativo	DIVERSIFICADO, PREPIMARIO BILINGUE, PRIMARIA, BASICO
SECTOR	Categórica	Tipo de gestión	OFICIAL, PRIVADO, MUNICIPAL, COOPERATIVA
AREA	Categórica	Ubicación geográfica	URBANA, RURAL
STATUS	Categórica	Estado de registro	ABIERTA
MODALIDAD	Categórica	Modalidad de enseñanza	MONOLINGUE, BILINGUE
JORNADA	Categórica	Turno o jornada	MATUTINA, VESPERTINA, DOBLE, SIN JORNADA,

Universidad del Valle de Guatemala
Data Science
Javier Prado – 21486
Bryan España - 21550
Luis Monterroso – 21699
Angel Herrarte - 22873

			NOCTURNA, INTERMEDIA
PLAN	Categorica	Tipo de plan o programa	DIARIO(REGULAR) FIN DE SEMANA A DISTANCIA SEMIPRESENCIAL (FIN DE SEMANA) SEMIPRESENCIAL (UN DIA A LA SEMANA) SEMIPRESENCIAL (DOS DIAS A LA SEMANA) SEMIPRESENCIAL VIRTUAL A DISTANCIA MIXTO SABATINO DOMINICAL INTERCALADO
DEPARTAMELNTAL	String	Departamento al que reporta administrativamente	Departamentos de guatemala

Fecha de extraccion de datos

Descarga de datos:

Origen: http://www.mineduc.gob.gt/BUSCAESTABLECIMIENTO_GE/

Limpieza y transformación

1. Extraccion HTML --> CSV
    - Se utilizo `transform_html_to_csv()`para convertir la tabla de índice 9 de cada archivo HTML.

Universidad del Valle de
Guatemala
Data Science
Javier Prado – 21486
Bryan España - 21550
Luis Monterroso – 21699
Angel Herrarte - 22873

2. Eliminacion de archivos temporates
    - Remove_files() borró data que ya no se usará

3. Limpieza de encabezados y trailers
    - Clean_headers_and_trailers() eliminó filas de cabeceras duplicadas y filas vacías finales

4. **Detección y reporte de valores nulos**
    - `get_null_stats()` y `count_null_instances()` generaron `null_report.csv`.

5. **Verificación de tipos de datos**
    - `check_datatypes()` identificó inconsistencias de tipo.

6. **Normalización de tipos: texto único**
    - `change_dataframe_types_structure()` forzó todo a `string` y aplicó `clean_string()`.

7. **Relleno de valores faltantes**
    - `fill_nulls_with_values()` sustituyó `NaN` en:
    - DIRECCION → “DESCONOCIDO”
    - TELEFONO → “00000000”
    - DIRECTOR → “DESCONOCIDO”

8. **Estandarización de formatos y contenidos**
    - Quitar acentos (`remove_accents_and_special_chars()`).
    - Homogeneizar teléfonos (`standardize_phone_number()`).
    - Limpiar direcciones y abreviaturas (`standardize_address()`).
    - Texto en mayúsculas y sin comillas innecesarias (`comprehensive_text_cleaning()`).

9. **Detección y unificación de duplicados**
    - `detect_*()` generó `resumen_duplicados.csv`.

Universidad del Valle de Guatemala
Data Science
Javier Prado – 21486
Bryan España - 21550
Luis Monterroso – 21699
Angel Herrarte - 22873

- `clean_legitimate_duplicates()` eliminó/fusionó registros repetidos.

10. **Verificación estructural y concatenación**
    - `verify_dataframe_structure()` y `concatenate_dataframes()` produjeron el CSV final.