Proyecto 1: Obtención y Limpieza de Datos

Diederich Solis (22952) Sara Guzmán (22097)

Juan Pablo Cordón (21458)

https://github.com/DiederichSolis/PR_data.git

Semestre II – 2025

1. Introducción

El acceso a datos limpios y consistentes constituye un prerrequisito fundamental para el análisis en ciencia de datos. En el contexto guatemalteco, el Ministerio de Educación (MINEDUC) publica registros de establecimientos educativos a nivel nacional. Dichos registros, aunque valiosos, presentan limitaciones típicas de datos administrativos: inconsistencias tipográficas, duplicados, campos incompletos y formatos no estandarizados.

Este proyecto se centra en el procesamiento de establecimientos que ofrecen nivel diversificado, con la meta de transformar la base de datos cruda en un conjunto limpio, uniforme y apto para análisis posteriores. El trabajo se justifica porque un centro educativo puede estar registrado bajo múltiples códigos según sus servicios autorizados (nivel, plan o jornada). La ausencia de un proceso de limpieza adecuado conlleva a duplicaciones y distorsiones estadísticas, afectando la interpretación de la cobertura educativa.

2. Metodología

El desarrollo se realizó en Python 3.10 dentro de Jupyter Notebook. Se utilizaron las librerías:

- `pandas`: manipulación tabular y consolidación de archivos.
- `numpy`: operaciones numéricas auxiliares.
- `unicodedata`: normalización de acentos y caracteres especiales.
- `openpyxl`: lectura/escritura de Excel.

Las fases principales fueron:

1. **Consolidación de datos crudos**: 24 archivos departamentales en Excel fueron unificados en un único DataFrame con 33,831 registros iniciales.
2. **Análisis exploratorio**: diagnóstico de valores faltantes, duplicados y valores atípicos en campos clave.
3. **Procesos de limpieza**: normalización de nombres, direcciones, teléfonos y variables categóricas, apoyados en catálogos oficiales.
4. **Control de calidad**: generación de reportes auxiliares para validar las transformaciones aplicadas.
5. **Exportación**: creación de un CSV y un Excel limpio con 11,985 registros únicos, además de reportes de auditoría en CSV.

3. Análisis del estado crudo

3.1. Dimensiones iniciales

El csv contenía 33,831 registros y 17 variables. La Figura **??** muestra la comparación entre los datos crudos y los depurados: tras la limpieza se redujo a 11,985 registros, lo cual representa una reducción del 64.6 %.

3.2. Valores faltantes

Los campos con mayor proporción de valores ausentes fueron:

- **DIRECTOR** (13.29 %)
- **TELEFONO** (8.29 %)
- **SUPERVISOR** (4.68 %)
- **DISTRITO** (4.66 %)

3.3. Duplicados

Se identificaron:

- **Duplicados exactos**: 33,122 registros (97.9 %).
- **Duplicados aproximados**: mismos establecimiento–dirección–teléfono pero con variaciones tipográficas (ej. “INSTITUTO NAL. MIXTO” vs “INSTITUTO NACIONAL MIXTO”).

Esto evidencia deficiencias en los procesos de consolidación originales.

4. Procesos de limpieza

Cada variable recibió un tratamiento específico, documentado en reportes auxiliares.

4.1. CODIGO

Se preservó como texto para mantener ceros iniciales. Se verificó formato de 12 dígitos y unicidad relativa. Se mantuvo la regla de que un mismo establecimiento puede tener múltiples códigos.

4.2. DISTRITO, DEPARTAMENTO y MUNICIPIO

Se detectaron 247 combinaciones inusuales de pares departamento–municipio, registradas en el archivo `reporte_pares_departamento_municipio_raros.csv`. La limpieza consistió en estandarizar valores ortográficos y corregir inconsistencias frecuentes.

4.3. ESTABLECIMIENTO

Fue uno de los campos más problemáticos:

- Eliminación de caracteres especiales (comillas, tildes innecesarias).
- Expansión de abreviaturas: "INST." → "INSTITUTO".
- Normalización a mayúsculas sin acentos.

Se documentaron posibles duplicados semánticos en `reporte_sospechas_duplicados_establecimiento.c`

4.4. DIRECCION

Problemas más comunes:

- Abreviaturas heterogéneas ("Av.", "Ave.", "Avda." → "AVENIDA").
- Números de zona escritos con letras ("ZONA O" → "ZONA 0").

La estandarización permitió mejorar la consistencia espacial de los registros.

4.5. TELEFONO

Los problemas fueron:

- Presencia de paréntesis y guiones.
- Múltiples teléfonos en un solo campo.
- Números inválidos con menos de 8 dígitos.

El reporte `reporte_telefonos_invalidos.csv` documenta 1,888 casos corregidos o eliminados.

4.6. SUPERVISOR y DIRECTOR

Campos con alta proporción de faltantes. Se normalizó mayúsculas y se eliminaron caracteres no alfabéticos, preservando coherencia aunque sin imputación automática para evitar sesgos.

4.7. Variables categóricas (NIVEL, SECTOR, AREA, STATUS, MODALIDAD, JORNADA, PLAN, DEPARTAMENTAL)

Se estandarizaron mediante diccionarios de mapeo. Ejemplos:

- "Matutina" $\rightarrow$ "MATUTINA"
- "Ciudad Capital" $\rightarrow$ "GUATEMALA"
- "Priv." $\rightarrow$ "PRIVADO"

Los conteos antes y después están resumidos en `reporte_value_counts_categoricos.csv`.

5. Resultados

El dataset final cuenta con 11,985 registros y 17 variables limpias. Principales logros:

- Reducción de más del 60 % de los registros mediante deduplicación.
- Normalización completa de nombres, direcciones y teléfonos.
- Estándares homogéneos para variables categóricas.
- Exportación a CSV y XLSX para uso posterior.

6. Comparación entre datos crudos y datos limpios

Una parte fundamental del proceso de limpieza consistió en comparar sistemáticamente el estado inicial de la base de datos contra el estado final depurado. Esto permite evidenciar la magnitud de los problemas encontrados y la efectividad de las transformaciones realizadas.

6.1. Dimensiones generales

El dataset pasó de 33,831 registros a 11,985 registros únicos, manteniendo las mismas 17 variables. La reducción fue del 64.6 %, principalmente por la eliminación de duplicados y la normalización de códigos repetidos.

```
=== Dimensiones ===
Original: (33831, 17)
Limpia  : (11985, 17)
```

6.2. Comparación de valores nulos

La Tabla 1 muestra la reducción de valores nulos tras la limpieza. Destaca la disminución significativa en **TELEFONO** (de 2,805 a 1,888) y **DIRECTOR** (de 4,497 a 1,906), reflejo de la estandarización y correcciones aplicadas.

Cuadro 1: Comparación de valores nulos antes y después de la limpieza

Campo	Nulos original	Nulos limpio
CODIGO	0	0
DISTRITO	1575	527
DEPARTAMENTO	0	0
MUNICIPIO	0	0
ESTABLECIMIENTO	12	4
DIRECCION	231	78
TELEFONO	2805	1888
SUPERVISOR	1584	533
DIRECTOR	4497	1906

6.3. Ejemplos de cambios específicos

Se presentan transformaciones representativas aplicadas a campos clave:

6.3.1. ESTABLECIMIENTO

- Antes: `INSTITUTO DE EDUCACION DIVERSIFICADA ÇENTRO DE ESTUDIOS MERCADOLOGICOS Y PUBLICITARIOS"`
- Después: `INSTITUTO DE EDUCACION DIVERSIFICADA CENTRO DE ESTUDIOS MERCADOLOGICOS Y PUBLICITARIOS`

6.3.2. DIRECCION

- Antes: `10A. AVENIDA 9-42`
- Después: `10 AVENIDA 9-42`

6.3.3. TELEFONO

- Antes: `(502) 2251-2759`
- Después: `22512759`

6.4. Variables categóricas

En la Tabla 6.4 se resumen los principales cambios en categorías después de la limpieza. El efecto más notorio es la reducción proporcional en todas las categorías debido a la deduplicación.

Variable	**Original**	**Limpia**
NIVEL DIVERSIFICADO	33,831	11,985
SECTOR PRIVADO	29,085	10,250
SECTOR OFICIAL	3,393	1,235
SECTOR MUNICIPAL	510	186
SECTOR COOPERATIVA	843	314
AREA URBANA	27,654	9,784
AREA RURAL	6,168	2,198
STATUS ABIERTA	19,752	7,101
STATUS CERRADA DEFINITIVAMENTE	5,268	1,837

STATUS CERRADA TEMPORALMENTE	8,475	2,935
JORNADA DOBLE	11,112	3,976
JORNADA MATUTINA	8,811	3,087
JORNADA VESPERTINA	9,189	3,252

6.5. Interpretación

El análisis comparativo demuestra que:

- La limpieza permitió reducir de manera drástica la duplicación de registros sin perder información única.
- La estandarización disminuyó errores tipográficos y mejoró la consistencia.
- Los campos categóricos fueron mapeados a valores únicos coherentes, facilitando el análisis estadístico posterior.

7. Discusión

El proyecto muestra cómo datos administrativos requieren un esfuerzo de limpieza riguroso. En el caso de los establecimientos educativos:

- La duplicación refleja la existencia de múltiples códigos por centro, lo cual es real, pero debía documentarse y no confundirse con registros distintos.

- La falta de estandarización tipográfica en nombres y direcciones dificultaba la identificación de duplicados, justificando la normalización agresiva.
- La ausencia de números de teléfono válidos limita estudios de contacto, pero la depuración asegura que los números presentes cumplen con formato nacional.

8. Conclusiones

El proceso de limpieza permitió obtener un conjunto confiable y coherente. En particular:

1. Se documentaron las inconsistencias iniciales en campos clave.
2. Se aplicaron transformaciones reproducibles para cada variable.
3. Se entregaron reportes complementarios para transparencia.

Este dataset limpio constituye la base para análisis de cobertura educativa, planificación de recursos y estudios de accesibilidad escolar en Guatemala.

9. Referencias

Referencias

Ministerio de Educación de Guatemala. (2025). *Busca Establecimiento Educativo*. Recuperado de `http://www.mineduc.gob.gt/BUSCAESTABLECIMIENTO_GE/`

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference.

Van der Aalst, W. (2016). *Data Preparation and Cleaning*. Springer.