

Avances del Proyecto 1

- Paula Barillas - 22764
- Derek Arreaga - 22537
- Mónica Salvatierra - 22249

Link del repositorio: <https://github.com/alee2602/PROYECTO1-DS>

Importación y unificación de datos

```
In [18]: import pandas as pd
import os

columnas = ['CODIGO', 'DISTRITO', 'DEPARTAMENTO', 'MUNICIPIO', 'ESTABLECIMIENTO',
            'DIRECCION', 'TELEFONO', 'SUPERVISOR', 'DIRECTOR', 'NIVEL', 'SECTOR',
            'AREA', 'STATUS', 'MODALIDAD', 'JORNADA', 'PLAN', 'DEPARTAMENTAL']

carpeta = 'data/csv/'
dfs = []

for archivo in os.listdir(carpeta):
    if archivo.endswith('.csv'):
        ruta_completa = os.path.join(carpeta, archivo)
        temp_df = pd.read_csv(ruta_completa)
        print(f'Leyendo {archivo} con {temp_df.shape[0]} centros educativos')
        dfs.append(pd.read_csv(ruta_completa))

df = pd.concat(dfs, ignore_index=True)
```

```

Leyendo altaverapaz.csv con 294 centros educativos
Leyendo bajaverapaz.csv con 94 centros educativos
Leyendo chimaltenango.csv con 300 centros educativos
Leyendo chiquimula.csv con 136 centros educativos
Leyendo ciudadcapital.csv con 860 centros educativos
Leyendo elprogreso.csv con 97 centros educativos
Leyendo escuintla.csv con 393 centros educativos
Leyendo guatemala.csv con 1036 centros educativos
Leyendo huehuetenango.csv con 295 centros educativos
Leyendo izabal.csv con 273 centros educativos
Leyendo jalapa.csv con 121 centros educativos
Leyendo jutiapa.csv con 296 centros educativos
Leyendo peten.csv con 270 centros educativos
Leyendo quetzaltenango.csv con 365 centros educativos
Leyendo quiche.csv con 184 centros educativos
Leyendo retalhuleu.csv con 272 centros educativos
Leyendo sacatepequez.csv con 206 centros educativos
Leyendo sanmarcos.csv con 431 centros educativos
Leyendo santarosa.csv con 133 centros educativos
Leyendo solola.csv con 111 centros educativos
Leyendo suchitepequez.csv con 296 centros educativos
Leyendo totonicapán.csv con 51 centros educativos
Leyendo zacapa.csv con 70 centros educativos

```

Obtener valores de columnas generales

```
In [19]: for col in ['NIVEL', 'SECTOR', 'AREA', 'STATUS', 'MODALIDAD', 'JORNADA', 'PLAN', 'D
          print(f'Datos únicos de la columna {col}:\n{df[col].unique()}')
```

```

Datos únicos de la columna NIVEL:
['DIVERSIFICADO']
Datos únicos de la columna SECTOR:
['PRIVADO' 'OFICIAL' 'MUNICIPAL' 'COOPERATIVA']
Datos únicos de la columna AREA:
['URBANA' 'RURAL' 'SIN ESPECIFICAR']
Datos únicos de la columna STATUS:
['ABIERTA']
Datos únicos de la columna MODALIDAD:
['MONOLINGUE' 'BILINGUE']
Datos únicos de la columna JORNADA:
['MATUTINA' 'VESPERTINA' 'DOBLE' 'NOCTURNA' 'SIN JORNADA' 'INTERMEDIA']
Datos únicos de la columna PLAN:
['DIARIO(REGULAR)' 'FIN DE SEMANA' 'A DISTANCIA' 'SEMIPRESENCIAL'
 'SEMIPRESENCIAL (FIN DE SEMANA)' 'SEMIPRESENCIAL (UN DÍA A LA SEMANA)'
 'VIRTUAL A DISTANCIA' 'SEMIPRESENCIAL (DOS DÍAS A LA SEMANA)' 'SABATINO'
 'INTERCALADO' 'DOMINICAL' 'MIXTO']
Datos únicos de la columna DEPARTAMENTAL:
['ALTA VERAPAZ' 'BAJA VERAPAZ' 'CHIMALTENANGO' 'CHIKUIMULA'
 'GUATEMALA NORTE' 'GUATEMALA ORIENTE' 'GUATEMALA OCCIDENTE'
 'GUATEMALA SUR' 'EL PROGRESO' 'ESCUINTLA' 'HUEHUETENANGO' 'IZABAL'
 'JALAPA' 'JUTIAPA' 'PETÉN' 'QUETZALTENANGO' 'QUICHÉ' 'QUICHÉ NORTE'
 'RETALHULEU' 'SACATEPÉQUEZ' 'SAN MARCOS' 'SANTA ROSA' 'SOLOLÁ'
 'SUCHITEPÉQUEZ' 'TOTONICAPÁN' 'ZACAPA']

```

```
In [20]: df.shape
```

1. Descripción general del dataset

A continuación, se describen los datos obtenidos del sitio web del [Mineduc](#):

Los datos corresponden a los **6584** centros educativos a nivel nacional que imparten **educación diversificada**. Están organizados en **23 archivos CSV**, uno por cada **departamento** de Guatemala. Cada archivo contiene **17 columnas** con información relevante sobre los establecimientos:

- **CODIGO** : Identificador único del establecimiento, con el formato **XX-XX-XXXX-XX** .
 - **Primer bloque (XX)**: Código del departamento (ej. **18** para Izabal).
 - **Segundo bloque (XX)**: Código del municipio dentro del departamento.
 - **Tercer bloque (XXXX)**: Identificador interno del centro educativo.
 - **Cuarto bloque (XX)**: Es siempre **46** , posiblemente indica el nivel educativo (diversificado) o el sistema de codificación actual del Mineduc.
- **DISTRITO** : Código del distrito escolar al que pertenece el establecimiento. Tiene el formato **XX-XXX** , donde el primer número corresponde al departamento, y los últimos tres identifican las zonas operativas o educativas regionales dentro del departamento.
Ejemplo: **18-007** y **18-008** son distritos distintos dentro de Izabal.
- **DEPARTAMENTO** : Nombre del departamento (ej. **IZABAL**).
- **MUNICIPIO** : Municipio donde se ubica el establecimiento (ej. **PUERTO BARRIOS**).
- **ESTABLECIMIENTO** : Nombre oficial del centro educativo. Puede incluir el tipo (**COLEGIO** , **ESCUELA** , **INSTITUTO**) y detalles como mixto, cristiano, etc.
- **DIRECCIÓN** : Ubicación del centro.
- **TELEFONO** : Número telefónico del establecimiento, si está disponible.
- **SUPERVISOR** : Nombre del supervisor distrital asignado al establecimiento.
- **DIRECTOR** : Nombre del director o responsable del centro educativo.
- **NIVEL** : Máximo nivel educativo ofrecido. En este caso todos son **DIVERSIFICADO** .
- **SECTOR** : Define el sector del establecimiento, puede ser:
 - **PRIVADO**
 - **OFICIAL**
 - **MUNICIPAL**
 - **COOPERATIVA**

- **AREA** : Área geográfica según categorización del Mineduc: **URBANA, RURAL** o **SIN ESPECIFICAR**.
- **STATUS** : Estado de funcionamiento del centro educativo. En estos datos, el valor para todos es **ABIERTA** .
- **MODALIDAD** : Tipo de modalidad educativa: **MONOLINGUE** o **BILINGUE** .
- **JORNADA** : Jornada en que opera el centro: **MATUTINA, VESPERTINA, DOBLE, NOCTURNA, SIN JORNADA, INTERMEDIA**.
- **PLAN** : Plan educativo implementado:
 - **DIARIO(REGULAR)**
 - **FIN DE SEMANA**
 - **A DISTANCIA**
 - **SEMIPRESENCIAL**
 - **SEMIPRESENCIAL (FIN DE SEMANA)**
 - **SEMIPRESENCIAL (UN DÍA A LA SEMANA)**
 - **VIRTUAL A DISTANCIA**
 - **SEMIPRESENCIAL (DOS DÍAS A LA SEMANA)**
 - **SABATINO**
 - **INTERCALADO**
 - **DOMINICAL**
 - **MIXTO**
- **DEPARTAMENTAL** : Nombre del departamento al que pertenece el centro educativo. En algunos casos, describe subdivisiones regionales dentro de departamentos con alta densidad de establecimientos, como **GUATEMALA NORTE, GUATEMALA ORIENTE, GUATEMALA OCCIDENTE, GUATEMALA SUR**, o distinciones como **QUICHÉ** y **QUICHÉ NORTE**.

2. Variables que requieren limpieza

Ejemplo:

- **ESTABLECIMIENTO** : hay diferencias por uso de mayúsculas, comillas, acentos inconsistentes.
- **DIRECCIÓN** : muchas variaciones, errores ortográficos o campos incompletos.
- **SUPERVISOR** y **DIRECTOR** : hay diferencias por uso de mayúsculas/minúsculas o nombres mal escritos.
- **TELEFONO** : hay algunos vacíos, valores con caracteres no numéricos.

```
In [21]: # Examinar una muestra de datos para identificar columnas que necesitan limpieza
print("Registros para analizar:")
df.sample(3)
```

Registros para analizar:

Out[21]:

	CODIGO	DISTRITO	DEPARTAMENTO	MUNICIPIO	ESTABLECIMIENTO	DIRECCION
3698	18-04-0404-46	18-030	IZABAL	MORALES	INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA	CALLE TULIO MARUZZO
4464	17-14-0102-46	17-013	PETEN	EL CHAL	COLEGIO EDUCATIVO DIVERSIFICADO LAS FLORES -CE...	COOPERATIVA LAS FLOREZ
3019	01-15-2072-46	01-214	GUATEMALA	VILLA NUEVA	COLEGIO MIXTO NUESTRA PATRIA	3A. CALLE 3-78 ZONA 1

In [22]:

```
# Analizar Columnas
columns_to_check = ['ESTABLECIMIENTO', 'DIRECCION', 'TELEFONO', 'SUPERVISOR', 'DIRECCION']

for col in columns_to_check:
    print(f"\n==== Análisis de la columna {col} ====")
    # valores únicos
    print(f"Cantidad de valores únicos: {df[col].nunique()}")
    print(f"Ejemplo/Muestra de valores: {df[col].sample(3).tolist()}")
    # valores nulos
    null_count = df[col].isna().sum()
    print(f"Valores nulos: {null_count} ({null_count/len(df)*100:.2f}%)")
    # valores vacíos
    empty_count = (df[col] == '').sum()
    print(f"Valores vacíos: {empty_count} ({empty_count/len(df)*100:.2f}%)")
    if col == 'TELEFONO':
        # Verificación sii hay valores no numéricos
        non_numeric = df[col].str.replace(r'[0-9\s\-\(\)\+]', '', regex=True).str.len()
        print(f"Valores con caracteres no numéricos: {non_numeric} ({non_numeric/len(df)*100:.2f}%)")

    if col == 'ESTABLECIMIENTO':
        # diferencias por uso de mayúsculas/minúsculas
        lowercase_count = df[col].str.islower().sum()
        uppercase_count = df[col].str.isupper().sum()
        mixed_count = len(df) - lowercase_count - uppercase_count
        print(f"Valores en minúsculas: {lowercase_count} ({lowercase_count/len(df)*100:.2f}%)")
        print(f"Valores en mayúsculas: {uppercase_count} ({uppercase_count/len(df)*100:.2f}%)")
        print(f"Valores mixtos: {mixed_count} ({mixed_count/len(df)*100:.2f}%)")
```

==== Análisis de la columna ESTABLECIMIENTO ====

Cantidad de valores únicos: 3989

Ejemplo/Muestra de valores: ['CENTRO EDUCATIVO JOCOTECO BILINGÜE', 'INSTITUTO DIVERSIFICADO POR COOPERATIVA', 'INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA']

Valores nulos: 0 (0.00%)

Valores vacíos: 0 (0.00%)

Valores en minúsculas: 0 (0.00%)

Valores en mayúsculas: 6584 (100.00%)

Valores mixtos: 0 (0.00%)

==== Análisis de la columna DIRECCION ====

Cantidad de valores únicos: 4490

Ejemplo/Muestra de valores: ['1A. CALLE BARRIO LA ESPERANZA ZONA 2', 'ALDEA SAN CRISTÓBAL, FRONTERA', '1A. AVENIDA 0-68 ZONA 2']

Valores nulos: 2 (0.03%)

Valores vacíos: 0 (0.00%)

==== Análisis de la columna TELEFONO ====

Cantidad de valores únicos: 4208

Ejemplo/Muestra de valores: [22603791, '24480675', '78494928']

Valores nulos: 45 (0.68%)

Valores vacíos: 0 (0.00%)

Valores con caracteres no numéricos: 0 (0.00%)

==== Análisis de la columna SUPERVISOR ====

Cantidad de valores únicos: 598

Ejemplo/Muestra de valores: ['MIGUEL ANGEL ARMAS ROCHA', 'BLANCA SARAI GUTIERREZ PINEDA', 'REYNA DEL ROSARIO CHAVEZ BARROW DE PACHECO']

Valores nulos: 0 (0.00%)

Valores vacíos: 0 (0.00%)

==== Análisis de la columna DIRECTOR ====

Cantidad de valores únicos: 3866

Ejemplo/Muestra de valores: ['REBECA AZUCENA CORONADO CRISÓSTOMO', 'CRISTIAN ALEXANDER PÉREZ GONZÁLEZ', 'FRANCISCO RUBÉN TOCA LETONA']

Valores nulos: 23 (0.35%)

Valores vacíos: 0 (0.00%)

1. ESTABLECIMIENTO:

- Contiene inconsistencias en el uso de acentos (por ejemplo, "EDUCACION" vs "EDUCACIÓN", "TECNICO" vs "TÉCNICO").
- El 23.88% de los registros contienen comillas (").
- Existen 1,185 establecimientos con nombres duplicados pero con códigos diferentes, lo que indica posibles variantes del mismo centro educativo con diferentes jornadas o planes.

2. DIRECCIÓN:

- Presenta 2 valores nulos (0.03%).
- Contiene 47 direcciones con menos de 10 caracteres (0.71%), sugiriendo información incompleta.
- Alta variabilidad en el formato y estilo de escritura.

3. TELÉFONO:

- Tiene 45 valores nulos (0.68%).
- El formato no es consistente, aunque el 73.80% son números de 8 dígitos sin separadores.
- Algunos están almacenados como números y otros como texto.

4. SUPERVISOR y DIRECTOR:

- Para DIRECTOR hay 23 valores nulos (0.35%).
- Posibles inconsistencias en el formato de nombres (espacios dobles, uso de mayúsculas/minúsculas, acentos).

3. Estrategia de limpieza

ESTABLECIMIENTO

1. Convertir todos los nombres a **mayúsculas** para estandarizar.
2. Eliminar **comillas dobles** y otros caracteres especiales innecesarios.
3. Corregir errores comunes de **acentuación** usando reglas o un diccionario personalizado (EDUCACION → EDUCACIÓN , TECNICO → TÉCNICO).
4. Eliminar espacios extra (dobles espacios, al inicio o final).
5. Crear un identificador agrupado para detectar **variantes de un mismo establecimiento**, usando similitud de texto (por ejemplo con `fuzzywuzzy` o `difflib`) y agrupando por nombre + dirección.

DIRECCIÓN

1. Convertir todas las direcciones a **mayúsculas**.
2. Eliminar espacios innecesarios y caracteres no informativos (dobles comas, comillas).
3. Rellenar los valores nulos con `NaN` y decidir si se imputan o se eliminan.
4. Establecer un formato básico: por ejemplo, reemplazar abreviaturas (`AV.` → `AVENIDA`), normalizar `ZONA` , etc.
5. Identificar posibles direcciones **duplicadas o incompletas** para revisión manual o agrupación.

TELÉFONO

1. Convertir todos los teléfonos a **cadena de texto** (`str`) para aplicar limpieza uniforme.
2. Eliminar todos los **caracteres no numéricos** con una expresión regular.
3. Verificar que los números tengan exactamente **8 dígitos**; marcar o excluir los que no lo cumplan.

4. Rellenar nulos con `NaN` ; considerar reemplazos con datos oficiales si se tiene acceso a otras fuentes.

SUPERVISOR y DIRECTOR

1. Convertir todos los nombres a **mayúsculas**.
2. Eliminar espacios dobles, al inicio y al final.
3. Normalizar los nombres corrigiendo errores comunes de acentuación (`RENÉ` , `EDUARDO` , etc.).
4. Rellenar valores nulos con `NaN` ; si es necesario, crear una categoría `"NO REGISTRADO"` para análisis exploratorios.
5. (Opcional) Separar los nombres en campos de **nombre y apellido**, si fuera útil para análisis futuros.

4. Consideraciones adicionales

Basado en el análisis realizado, estas son las consideraciones adicionales que se deben tener en cuenta durante el proceso de limpieza y análisis de los datos:

Consideraciones sobre identificación de establecimientos

- **Verificación de unicidad de CODIGO:** El análisis confirma que todos los códigos tienen el formato válido (`XX-XX-XXXX-XX`) y son únicos. Este campo debe mantenerse como identificador primario de cada registro.
- **Detección de centros educativos duplicados:** Muchos establecimientos aparecen con el mismo nombre pero distintos códigos. Esto se debe principalmente a que ofrecen diferentes jornadas (MATUTINA, VESPERTINA, NOCTURNA) o planes educativos (DIARIO, FIN DE SEMANA). No son verdaderos duplicados, sino diferentes modalidades del mismo centro físico.
- **Creación de identificador de establecimiento físico:** Se recomienda generar un nuevo identificador que agrupe los diferentes registros que corresponden al mismo centro educativo físico. Este se puede crear combinando el nombre normalizado y la dirección.

Consideraciones sobre integridad de relaciones

- **Verificación de relaciones geográficas:** Es importante validar la consistencia entre DEPARTAMENTO, MUNICIPIO y DISTRITO. Cada distrito debería pertenecer a un único departamento, y cada municipio a un único departamento.
- **Análisis de ubicaciones repetidas:** Existen direcciones que aparecen en múltiples registros, lo que indica que varios centros educativos pueden estar operando en la

misma ubicación física, posiblemente compartiendo instalaciones pero con administraciones diferentes.

Consideraciones sobre valores ausentes o incompletos

- **Datos de contacto incompletos:** Para los registros con TELEFONO o DIRECTOR ausentes, se debe decidir si se completan con valores como "NO DISPONIBLE" o si se tratan con técnicas más avanzadas.
- **Direcciones incompletas:** Las direcciones muy cortas o genéricas deberían revisarse manualmente o complementarse con información adicional si está disponible.

Consideraciones sobre enriquecimiento de datos

- **Posibles variables derivadas:**
 - Crear una columna para identificar el **tipo de establecimiento** (COLEGIO, INSTITUTO, ESCUELA) extrayéndolo del nombre.
 - Añadir una columna de **región geográfica** que agrupe departamentos por zonas (norte, sur, oriente, occidente, central).
 - Generar una variable de **clasificación** que combine SECTOR, MODALIDAD y PLAN para facilitar análisis comparativos.
- **Geocodificación:** Evaluar la posibilidad de convertir las direcciones en coordenadas geográficas para análisis espaciales y visualización en mapas.

Consideraciones sobre técnicas de normalización avanzadas

- **Detección de similitud de nombres:** Usar técnicas de coincidencia difusa (fuzzy matching) o distancia de edición para identificar nombres de establecimientos que podrían ser el mismo pero están escritos con variaciones menores.
- **Normalización lingüística:** Estandarizar términos específicos del ámbito educativo que pueden aparecer con diferentes grafías o abreviaturas.

Nota importante: Como se observó en el ejemplo de Quiché con códigos **14-14-0070-46** y **14-14-0083-46**, donde uno opera en jornada **DOBLE** con plan **FIN DE SEMANA** y el otro en jornada **VESPERTINA** con plan **DIARIO(REGULAR)**, estos registros representan diferentes ofertas educativas del mismo centro y no deben eliminarse como duplicados.

```
In [23]: # 1. Misma dirección pero diferentes nombres
direcciones_counts = df.groupby('DIRECCION').size()
direcciones_duplicadas = direcciones_counts[direcciones_counts > 1].sort_values(asc

print(f"Existen {len(direcciones_duplicadas)} direcciones que aparecen en más de un
print(f"Las 5 direcciones más repetidas son:")
```

```

for direccion in direcciones_duplicadas.head(5).index:
    print(f"\nDirección: '{direccion}'")
    print(df[df['DIRECCION'] == direccion][['CODIGO', 'ESTABLECIMIENTO', 'JORNADA',

# 2. Agrupamiento establecimientos similares
def simplificar_nombre(nombre):
    # Convertir a mayúsculas y eliminar comillas
    nombre = nombre.upper().replace("'", '')
    # Eliminar palabras genéricas que pueden causar variaciones
    palabras_eliminar = ['PRIVADO', 'MIXTO', 'COLEGIO', 'INSTITUTO', 'ESCUELA', 'CE
    for palabra in palabras_eliminar:
        nombre = nombre.replace(f" {palabra} ", " ")
    # Eliminar espacios múltiples
    return " ".join(nombre.split())

# columna temporal con nombres simplificados
df['NOMBRE_SIMPLIFICADO'] = df['ESTABLECIMIENTO'].apply(simplificar_nombre)

# ocurrencias de nombres simplificados
nombres_simples_counts = df['NOMBRE_SIMPLIFICADO'].value_counts()
nombres_simples_duplicados = nombres_simples_counts[nombres_simples_counts > 1].sor

print(f"\nDespués de simplificar nombres, existen {len(nombres_simples_duplicados)}")
print(f"Los 5 nombres más comunes después de simplificación son:")
for nombre in nombres_simples_duplicados.head(5).index:
    print(f"\nNombre simplificado: '{nombre}'")
    print(f"Nombres originales:")
    print(df[df['NOMBRE_SIMPLIFICADO'] == nombre][['ESTABLECIMIENTO']].sample(min(3,

# 3. Integridad de relaciones entre DEPARTAMENTO, MUNICIPIO y DISTRITO
print("\nVerificando integridad de relaciones geográficas:")
municipios_por_depto = df.groupby(['DEPARTAMENTO', 'MUNICIPIO']).size().reset_index
print(f"Hay {len(municipios_por_depto)} combinaciones únicas de DEPARTAMENTO-MUNICI

# relación entre distrito y departamento
distrito_depto = df[['DISTRITO', 'DEPARTAMENTO']].drop_duplicates()
print(f"Hay {len(distrito_depto)} combinaciones únicas de DISTRITO-DEPARTAMENTO")

# distritos que pertenecen a más de un departamento (posible inconsistencia)
distrito_counts = distrito_depto.groupby('DISTRITO').size()
distritos_multiples = distrito_counts[distrito_counts > 1]
if len(distritos_multiples) > 0:
    print(f"Hay {len(distritos_multiples)} distritos asignados a más de un departam
    for distrito in distritos_multiples.index:
        print(f"\nDistrito: {distrito}")
        print(df[df['DISTRITO'] == distrito][['DEPARTAMENTO']].drop_duplicates().va

```

Existen 1039 direcciones que aparecen en más de un registro.

Las 5 direcciones más repetidas son:

Dirección: 'CABECERA MUNICIPAL'

	CODIGO	ESTABLECIMIENTO \
361	15-05-0014-46	INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA
367	15-06-0006-46	INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA
1737	02-04-0004-46	INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA

	JORNADA	PLAN
361	VESPERTINA	DIARIO(REGULAR)
367	VESPERTINA	DIARIO(REGULAR)
1737	VESPERTINA	DIARIO(REGULAR)

Dirección: 'BARRIO EL CENTRO'

	CODIGO	ESTABLECIMIENTO	JORNADA \
133	16-06-0090-46	INSTITUTO NACIONAL DE EDUCACIÓN DIVERSIFICADA	VESPERTINA
162	16-08-0208-46	INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA	VESPERTINA
241	16-13-0254-46	INSTITUTO NACIONAL DE EDUCACIÓN DIVERSIFICADA	MATUTINA

	PLAN
133	DIARIO(REGULAR)
162	DIARIO(REGULAR)
241	DIARIO(REGULAR)

Dirección: 'BARRIO EL CALVARIO'

	CODIGO	ESTABLECIMIENTO \
1684	02-01-0027-46	INSTITUTO TECNICO INDUSTRIAL MIXTO GUASTATOYA
1778	02-08-0023-46	COLEGIO PRIVADO LICEO SAN ANTONIO
3849	21-03-0026-46	COLEGIO PARTICULAR MIXTO "LICEO SAN LUIS"

	JORNADA	PLAN
1684	DOBLE	DIARIO(REGULAR)
1778	VESPERTINA	DIARIO(REGULAR)
3849	VESPERTINA	DIARIO(REGULAR)

Dirección: 'ALDEA PLAYITAS'

	CODIGO	ESTABLECIMIENTO \
3651	18-04-0016-46	COLEGIO PARTICULAR MIXTO "MINERVA"
3660	18-04-0118-46	INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA
3667	18-04-0170-46	COLEGIO PARTICULAR MIXTO MINERVA

	JORNADA	PLAN
3651	VESPERTINA	DIARIO(REGULAR)
3660	VESPERTINA	DIARIO(REGULAR)
3667	DOBLE	DIARIO(REGULAR)

Dirección: 'ALDEA CAMOJALLITO'

	CODIGO	ESTABLECIMIENTO \
3386	13-12-0075-46	INSTITUTO PRIVADO MIXTO ESFUERZO
3388	13-12-0081-46	INSTITUTO PARTICULAR MIXTO DEMOCRACENSE
3389	13-12-0096-46	COLEGIO PARROQUIAL PRIVADO MIXTO SANTA TERESIT...

	JORNADA	PLAN
3386	DOBLE	FIN DE SEMANA
3388	DOBLE	FIN DE SEMANA

3389 VESPERTINA DIARIO(REGULAR)

Después de simplificar nombres, existen 1312 nombres potencialmente similares.
Los 5 nombres más comunes después de simplificación son:

Nombre simplificado: 'INSTITUTO NACIONAL DIVERSIFICADA'

Nombres originales:

['INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA', 'INSTITUTO NACIONAL DE EDUCACION D
IVERSIFICADA', 'INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA']

Nombre simplificado: 'INSTITUTO NACIONAL EDUCACIÓN DIVERSIFICADA'

Nombres originales:

['INSTITUTO NACIONAL DE EDUCACIÓN DIVERSIFICADA', 'INSTITUTO NACIONAL DE EDUCACIÓN D
IVERSIFICADA', 'INSTITUTO NACIONAL DE EDUCACIÓN DIVERSIFICADA']

Nombre simplificado: 'CENTRO EDUCACIÓN EXTRAESCOLAR -CEEX-'

Nombres originales:

['CENTRO DE EDUCACIÓN EXTRAESCOLAR -CEEX-', 'CENTRO DE EDUCACIÓN EXTRAESCOLAR -CEEX-
, 'CENTRO DE EDUCACIÓN EXTRAESCOLAR -CEEX-']

Nombre simplificado: 'INSTITUTO DIVERSIFICADA POR COOPERATIVA ENSEÑANZA'

Nombres originales:

['INSTITUTO DE EDUCACION DIVERSIFICADA POR COOPERATIVA DE ENSEÑANZA', 'INSTITUTO DE
EDUCACION DIVERSIFICADA POR COOPERATIVA DE ENSEÑANZA', 'INSTITUTO DE EDUCACION DIVER
SIFICADA POR COOPERATIVA DE ENSEÑANZA']

Nombre simplificado: 'INSTITUTO DIVERSIFICADO POR COOPERATIVA'

Nombres originales:

['INSTITUTO DE DIVERSIFICADO POR COOPERATIVA', 'INSTITUTO DIVERSIFICADO POR COOPERAT
IVA', 'INSTITUTO DIVERSIFICADO POR COOPERATIVA']

Verificando integridad de relaciones geográficas:

Hay 348 combinaciones únicas de DEPARTAMENTO-MUNICIPIO

Hay 642 combinaciones únicas de DISTRITO-DEPARTAMENTO

Hay 22 distritos asignados a más de un departamento (posible inconsistencia)

Distrito: 01-106

['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-111

['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-207

['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-209

['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-210

['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-218

['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-401

['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-405
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-409
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-411
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-508
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-602
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-603
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-619
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-620
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-623
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-637
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-639
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-640
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-641
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 01-645
['CIUDAD CAPITAL' 'GUATEMALA']

Distrito: 99-001
['CHIQUMULA' 'SAN MARCOS']