Universidad del Valle de Guatemala
Facultad de Ingeniería
Data Science
Departamento de Ciencias de la Computación
Ciclo I, 2024

Proyecto I

Obtención y Limpieza de Datos

Adrián Ricardo Flores Trujillo	Carné 21500
Andrea Ximena Ramírez Recinos	Carné 21874
Daniel Armando Valdez Reyes	Carné 21240
Emilio José Solano Orozco	Carné 21212

Guatemala, 12 de agosto de 2024

1. Introducción

En esta actividad sé realizó el proceso de limpieza de datos obtenidos de establecimientos educativos de Guatemala hasta el nivel diversificado. El objetivo principal es preparar un conjunto de datos limpio y consistente que facilite el análisis posterior.

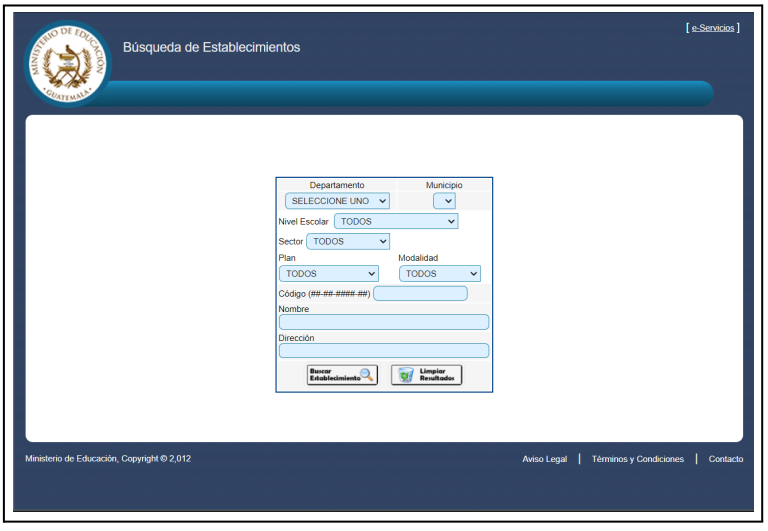
Durante la realización de este proyecto, se hizo uso de la técnica del web scraping para automatizar el acceso directo a la fuente de los datos brindados por el Mineduc, asegurando que la recolección de datos sea transparente y fácilmente reproducible. Se documentan todas las operaciones de limpieza llevadas a cabo, con el fin de permitir a otros verificar y comprender cada paso del proceso.

Las actividades incluyen la descripción del estado inicial de los datos y las operaciones de limpieza necesarias. Se presta especial atención a campos críticos, como nombre, dirección y teléfono, para garantizar la integridad de la información. Las inconsistencias, como representaciones múltiples de valores nulos y errores ortográficos, son abordadas de manera sistemática para asegurar que el conjunto de datos final esté listo para un análisis efectivo. Finalmente, se describe el codebook del conjunto de datos final, tras la limpieza, para facilitar futuras investigaciones y estudios.

2. Proceso de Obtención de los Datos

El proceso de obtención de los datos se llevó a cabo utilizando la base de datos de los establecimientos de Guatemala, la cual está disponible en la página proporcionada por el Mineduc para este propósito:

https://www.mineduc.gob.gt/BUSCAESTABLECIMIENTO_GE/



La página proporciona la posibilidad de obtener información de los establecimientos filtrando distintos valores para cada columna. En nuestro caso, se utilizó para filtrar por aquellos establecimientos que cuentan con nivel escolar que llegue hasta el Diversificado. Sin embargo, el problema principal de la página es que no permite seleccionar todos los departamentos al mismo tiempo y, para algunos departamentos, debido a la extensión de los datos que tienen no permiten la descarga de los mismos en formato csv. Por ende se optó por

hacer uso de una estrategia de automatización de la extracción de la información con la técnica denominada web scraping, la cual permite leer el apartado html de la página y obtener la información directamente de la fuente. Esto permitió también concatenar esta información en un único archivo llamado “establecimientos.csv”.

3. Breve Análisis y Procedimientos de Limpieza

3.1 Variables con Mayor Cantidad de Operaciones de Limpieza

1. adress
2. phone
3. school_name

3.2 Estrategias Principales de Limpieza

1. Se identificaron y eliminaron 22 entradas duplicadas en el conjunto de datos, lo que representa aproximadamente el 0.5% del total.

2. Para facilitar la comprensión y el manejo del conjunto de datos, se modificaron los nombres de las variables, permitiendo una organización más clara y una interpretación más precisa de la información.

3. En la limpieza de la variable departamento, se reemplazaron todas las entradas de "CIUDAD CAPITAL" por la cadena "GUATEMALA", que es más adecuada en el contexto de esta variable.

4. Para la limpieza de la variable municipalidad, se convirtieron todas las entradas con la subcadena "ZONA" a "CIUDAD DE GUATEMALA", ya que estas entradas representaban una porción significativa del conjunto de datos y habían sido erróneamente clasificadas como una municipalidad de la Ciudad de Guatemala.

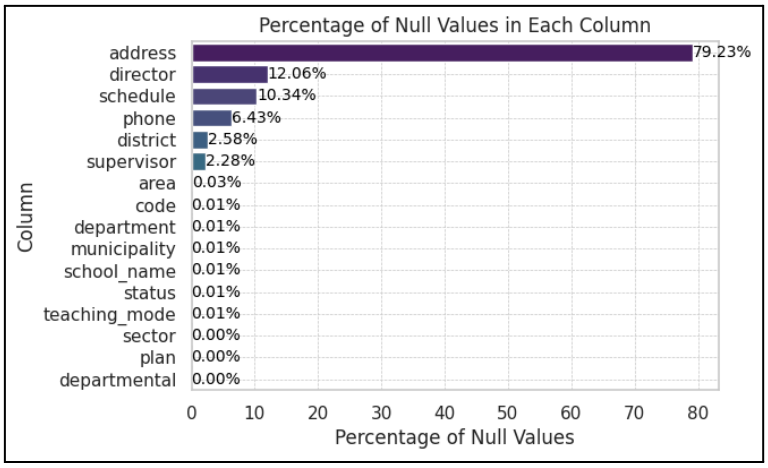
5. En las variables director y supervisor, se reemplazaron las entradas que no correspondían al nombre de una persona por la cadena "SIN DATO". Estas entradas se caracterizaban por consistir únicamente en caracteres especiales como "." o "-".

6. La variable nivel, que constaba exclusivamente de la entrada "DIVERSIFICADO", fue eliminada por no aportar información significativa para cualquier tipo de investigación que emplee este conjunto de datos.

7. En la variable plan, que contaba con 13 posibles entradas, se unificaron las categorías con significados similares, reduciendo la redundancia y clasificando todos los planes de estudio en una menor cantidad de categorías. Este problema se ejemplificó claramente en la categoría "SEMIPRESENCIAL", que tenía 4 clasificaciones adicionales.

8. Finalmente, en la limpieza de la variable departamental, se agruparon ciertas entradas de regiones en una sola para evitar inconsistencias en la nomenclatura de la variable, abordando específicamente las etiquetas "Guatemala" y "Quiché".

3.3 Manejo de Valores Faltantes



Aunque hay una cantidad significativa de valores nulos en una gran parte de las columnas del conjunto de datos, actualmente no se cuenta con información suficiente sobre el propósito de estos datos. Por lo tanto, no sería apropiado tomar decisiones drásticas en cuanto a la imputación de datos faltantes. En su lugar, se ha establecido una representación coherente para indicar la falta de información en todo el conjunto de datos, utilizando la etiqueta "Sin especificar" para todas las columnas.

Nota: Tomar en cuenta que todo este proceso con sus respectivas explicaciones y snippets de código se encuentran en el notebook que se adjunta en este repositorio/entrega.

4. Codebook

4.1 Descripción de los Datos

Este conjunto de datos se enfoca en la recopilación de información sobre establecimientos educativos en el país Guatemala. Los datos fueron obtenidos de la página oficial del Ministerio de Educación. Con alrededor de 9.354 registros y 16 columnas, cuenta con una dimensión relativamente grande. Cada uno de los registros representa una observación única, mientras que las 16 columnas corresponden a diferentes características o variables medidas para cada observación.

4.2 Clasificación de las Variables

Este conjunto de datos incluye aproximadamente cuatro variables cualitativas descriptivas que proporcionan información general sobre las observaciones, así como doce variables categóricas cualitativas que representan características específicas y atributos de los elementos analizados.

Nombre	Descripción	Clasificación
code	Identifica de manera única a cada establecimiento.	Cualitativa descriptiva.
school_name	Nombre del establecimiento educativo.	Cualitativa descriptiva.
address	Ubicación física del establecimiento.	Cualitativa descriptiva.
phone	Número(s) de teléfono del establecimiento.	Cualitativa descriptiva.

Nombre	Descripción	Clasificación
director	Nombre del director del establecimiento.	Cualitativa categórica.
district	Clasifica el establecimiento dentro de un distrito específico.	Cualitativa categórica.
department	Indica el departamento de Guatemala donde se encuentra cada establecimiento.	Cualitativa categórica.
municipality	Define el municipio de Guatemala donde se localiza cada establecimiento.	Cualitativa categórica.
sector	Tipo de sector educativo.	Cualitativa categórica.
area	Área geográfica del establecimiento.	Cualitativa categórica.
status	Estado actual del establecimiento.	Cualitativa categórica.
teaching_mode	Modalidad de enseñanza ofrecida.	Cualitativa categórica.
schedule	Tipo de jornada educativa.	Cualitativa categórica.
plan	Plan educativo implementado.	Cualitativa categórica.
departmental	Departamento administrativo al que pertenece el establecimiento.	Cualitativa categórica.
supervisor	Nombre del supervisor del establecimiento.	Cualitativa categórica.

Cuadro 1 - Descripción y Clasificación de Variables en el Conjunto de Datos

4.3 Variable *code*

- **Descripción**

  Contiene un identificador único para cada establecimiento.

- **Distribución**

  Esta variable contiene un valor único para cada establecimiento, por lo que no existen entradas duplicadas.

- **Código de Variable**

  No aplica.

- **Formato de la Variable**

  Esta variable consta de una secuencia de dígitos separados por guiones. La primera parte de la secuencia es el distrito del establecimiento, seguido por más dígitos para diferenciar cada establecimiento.

- **Datos Faltantes y Casos no Aplicables**

  Esta variable no contiene datos faltantes.

4.4 Variable *school_name*

- **Descripción**

  Contiene los nombres por el que se identifica cada establecimiento.

- **Distribución**

  Contiene 3829 entradas únicas, donde el valor con más repeticiones es “Instituto Nacional De Educación Diversificada Xekupilaj” con 430 veces.

- **Código de Variable**

  No aplica.

- **Formato**

  No aplica.

- **Datos Faltantes**

  Solo contiene un valor faltante el cual es expresado con el campo “Sin especificar”.

4.5 Variable *address*

- **Descripción**

  Contiene las direcciones estandarizadas de los registros de los establecimientos educativos de nivel diversificado de Guatemala.

- **Distribución**

  Consta de 559 entradas únicas, con el valor más repetido siendo "2 Avenida 1-04 Zona 1," que aparece un total de 24 repeticiones.

- **Código de Variable**

  No aplica.

- **Formato de la Variable**

  Estructura

  Las direcciones siguen el formato estándar utilizado en Guatemala:

  número, tipo de calle, número de propiedad, zona.

  Componentes

  →**Número de Calle o Avenida:** Un número entero que indica el número de la calle (Tomar en consideración que **NO** se incluyen sufijos como 'a', 'ra', etc.).

  →**Tipo de Calle:** Una de las siguientes palabras, capitalizadas: "Calle" o "Avenida".

  → **Número de Propiedad:** Un número entero seguido de un guión y otro número entero (e.g., 1-04), que indica un número de propiedad del establecimiento educativo.

  → **Zona:** La palabra "Zona" seguida de un número entero que indica la zona correspondiente a la dirección.

- **Datos Faltantes y Casos No Aplicables**

  1. La variable debe contener direcciones en el formato mencionado, y a cualquier dirección que no cumpla con este formato o carezca de suficiente información para cumplirlo se le deberá colocar **“Sin especificar”,** haciendo referencia a un dato nulo.

2. Por favor tome en consideración el **NO** ingresar tildes o caracteres especiales.

- **Ejemplos**
    1. 2 Avenida 1-04 Zona 1
    2. 11 Calle 0-97 Zona 3
    3. 7 Avenida 8-15 Zona 1

4.6 Variable *phone*

- **Descripción**

  Contiene una lista con los números telefónicos de 8 dígitos del establecimiento, separados en la mitad por un guión.

- **Distribución**

  Esta variable cuenta con 5403 valores únicos, con el valor más repetido siendo '2206-7425', que aparece un total de 21 veces en el conjunto de datos.

  Código de Variable

  No aplica

- **Formato de la Variable**

  Cadenas de texto compuestas por 8 dígitos sin extensión, separados por un guión al medio.

- **Datos Faltantes y Casos no Aplicables**

  Cualquier número telefónico de longitud menor a 8 dígitos se deberá colocar como **"Sin especificar"**, haciendo referencia a un dato nulo. Esto aplica para números telefónicos de 6 dígitos y fax.

4.7 Variable *director*

- **Descripción**

  Contiene el nombre completo de la persona encargada de la dirección del establecimiento.

- **Distribución**

  Cuenta con 4538 valores únicos. Donde el valor más repetido es el campo vacío "nan" con una frecuencia de 1125 veces y seguido de Jorge Granados Guzman con 12 repeticiones.

- **Código de Variable**

  No aplica.

- **Formato**

  El nombre sigue el formato usual de un nombre completo. Primero van los nombres o el nombre del individuo, seguido de sus apellidos.

- **Datos Faltantes**

Los registros llenados con un patrón de uno o varios guiones, incluidos aquellos que dicen “SIN DATO”, se deberán colocar como **“Sin especificar”** para referirnos a un valor nulo.

4.8 Variable *district*

- **Descripción**

  Contiene una secuencia de dígitos que especifican el distrito al que pertenece el establecimiento educativo.

- **Distribución**

  Consta de 691 valores únicos, con el valor más repetido siendo ”01-403”, que aparece un total de 268 veces en el conjunto de datos.

- **Código de Variable**

  No aplica.

- **Formato de la Variable**

  Los distritos están compuestos de dos partes, separadas por un guión. La primera parte es un número de 2 dígitos que especifica el departamento, y la segunda parte es un número de 3 dígitos que define el distrito específico al que pertenece el establecimiento.

- **Datos Faltantes y Casos no Aplicables**

  Cualquier distrito incompleto o desconocido se deberá colocar como **“Sin especificar”**, haciendo referencia a un dato nulo. Esto aplica para distritos incompletos, de los cuales solo se conoce el número del departamento.

4.9 Variable *department*

- **Descripción**

  Contiene los nombres de los departamentos de Guatemala a los cuales pertenece cada establecimiento educativo. Por favor note que cada nombre ha sido capitalizado para mantener la consistencia.

- **Distribución**

  Consta de 22 entradas únicas. El departamento más repetido es "Guatemala," que aparece un total de 3041 veces a lo largo del conjunto de datos. A continuación, se presentan los departamentos junto con su respectivo número de repeticiones:

Nombre	Conteo
Guatemala	3041
Escuintla	628
San Marcos	574
Huehuetenango	516
Quetzaltenango	491

Nombre	Conteo
Suchitepéquez	385
Alta Verapaz	374
Izabal	368
Petén	366
Chimaltenango	359
Sacatepéquez	319
Retalhuleu	316
Jutiapa	310
Quiche	244
Chiquimula	170
Santa Rosa	158
Jalapa	151
Sololá	138
El Progreso	125
Baja Verapaz	114
Zacapa	94
Totonicapán	90

Cuadro 2 - Distribución de la variable "department"

- **Código de Variable**
  No aplica.

- **Formato de la Variable**
  Estructura
  Los nombres de los departamentos están en formato de título (capitalizando la primera letra de cada palabra).

  Componentes
  → **Nombre de Departamento:** El nombre de uno de los departamentos oficiales del país de Guatemala, sin tildes o caracteres especiales y con la correspondiente capitalización.

- **Datos Faltantes y Casos No Aplicables**
  1. La variable debe contener nombres de departamentos válidos de Guatemala, todos capitalizados correctamente. Se le deberá colocar **"Sin especificar",**

haciendo referencia a un dato nulo, en dado caso no se cuente con información para la entrada.
2. Por favor tome en consideración el **NO** ingresar tildes o caracteres especiales.
3. Tampoco se pueden ingresar otros nombres de departamentos que no se encuentren ya en los listados anteriormente, ya que esto querrá decir que no forman parte de los 22 oficiales del país de Guatemala.

- **Ejemplos**
    1. Guatemala
    2. San Marcos
    3. Huehuetenango

4.10 Variable *municipality*

- **Descripción**

  Contiene los nombres de los municipios de Guatemala en los cuales se encuentra cada establecimiento del conjunto de datos.

- **Distribución**

  Consta de 22 entradas únicas. El departamento más repetido es "Guatemala," que aparece un total de 3041 veces a lo largo del conjunto de datos. A continuación, se presentan los departamentos junto con su respectivo número de repeticiones:

Nombre	Conteo
Guatemala	1567
Mixco	428
Villa Nueva	370
Quetzaltenango	248
Retalhuleu	185

Cuadro 3 - Distribución de la variable "municipality"

(Otras entradas varían en frecuencia, llegando hasta un total de 1 para varios supervisores menos comunes).

- **Código de Variable**

  No aplica.

- **Formato de la Variable**

  Estructura

  Los nombres de los municipios están en formato de título (capitalizando la primera letra de cada palabra).

  Componentes

  → **Nombre de Municipio:** El nombre de uno de los municipios oficiales del país de Guatemala, sin tildes o caracteres especiales y con la correspondiente capitalización.

- **Datos Faltantes y Casos No Aplicables**
    1. La variable debe contener nombres de municipios válidos de Guatemala, todos capitalizados correctamente. Se le deberá colocar **"Sin especificar",** haciendo referencia a un dato nulo, en dado caso no se cuente con información para la entrada.
    2. Por favor tome en consideración el **NO** ingresar tildes o caracteres especiales.

- **Ejemplos**
    1. Guatemala
    2. Villa Nueva
    3. Mixco

4.11 Variable *sector*

- **Descripción**

  Contiene la clasificación del sector en que se encuentra el establecimiento.

- **Distribución**

  Contiene 4 valores únicos de los cuales el más repetido es "PRIVADO" con 7956 veces.

- **Código de Variable**
    - **Privado:** Son aquellos que son administrados y financiados por entidades privadas, ya sean individuos, empresas o organizaciones sin fines de lucro. Estos centros educativos suelen ofrecer una variedad de niveles educativos y son conocidos por su independencia en cuanto a la currícula y los métodos de enseñanza.
    - **Oficial:** Bajo la administración y financiamiento del gobierno. Estos centros educativos públicos son gratuitos y están regulados por el Ministerio de Educación. Su objetivo principal es garantizar el acceso a la educación a todos los ciudadanos, cumpliendo con los estándares educativos nacionales.
    - **Cooperativas:** Son gestionados por cooperativas educativas, las cuales son organizaciones formadas por grupos de personas que se asocian voluntariamente para satisfacer sus necesidades educativas comunes. Estos centros combinan principios de la educación pública y privada, operando bajo un modelo de autogestión y colaboración comunitaria.
    - **Municipal:** Son aquellos gestionados y financiados por las municipalidades locales. Estos centros educativos suelen estar más enfocados en atender las necesidades específicas de la comunidad local, ofreciendo educación accesible y adaptada a las realidades y demandas de la población del municipio.

- **Formato**

  No aplica.

- **Datos Faltantes**

  Solo contiene un valor faltante el cual es expresado con el campo "Sin especificar".

4.12 Variable *area*

- **Descripción**

  Contiene el área al que pertenece el establecimiento en cuestión.

- **Distribución**

  Contiene 3 valores únicos de los cuales el sector Rural es el más repetido con 7606 veces.

- **Código de Variable**
  - **Urbano:** Se encuentran en zonas de alta densidad poblacional, como ciudades y grandes municipios.
  - **Rural:** Se encuentran en zonas con menor densidad poblacional, generalmente en el campo o en comunidades alejadas de los centros urbanos.

  Formato

  No aplica.

  Datos Faltantes

  Solo contiene un valor faltante el cual es expresado con el campo "Sin especificar".

4.13 Variable *status*

- **Descripción**

  Contiene el estatus actual del establecimiento en cuestión.

- **Distribución**

  Contiene 4 valores únicos de los cuales el área ABIERTA es el más repetido con 6545 veces.

- **Código de Variable**
  - **Abierta:** Están en pleno funcionamiento y ofrecen servicios educativos de manera regular.
  - **Cerrada temporalmente:** Han suspendido sus actividades educativas por un periodo determinado. Las razones para esta suspensión pueden variar, incluyendo renovaciones, falta de personal, o situaciones excepcionales.
  - **Temporal Títulos:** Están autorizados temporalmente para operar con el propósito específico de expedir títulos o certificados.
  - **Temporal Nombramiento:** Funcionan de manera provisional, generalmente debido a un nombramiento temporal de autoridades o personal.

  Formato

  No aplica.

  Datos Faltantes

  Solo contiene un valor faltante el cual es expresado con el campo "Sin especificar".

4.14 Variable *teaching_mode*

Descripción

Contiene el modalidad de enseñanza que imparten en el establecimiento

- **Distribución**

  Contiene 2 valores únicos de los cuales la modalidad MONOLINGUE es el más repetido con 9039 veces.

- **Código de Variable**
  - **Monolingüe:** Imparten la educación en un solo idioma, que generalmente es el idioma oficial del país, el español. Estos centros educativos se enfocan en enseñar todas las materias y contenidos utilizando exclusivamente este idioma, asegurando que los estudiantes desarrollen un alto nivel de competencia lingüística en él.
  - **Bilingüe:** Ofrecen educación en dos idiomas, combinando el idioma oficial con otro idioma adicional, que puede ser una lengua extranjera o una lengua indígena local.

  Formato

  No aplica.

  Datos Faltantes

  Solo contiene un valor faltante el cual es expresado con el campo "Sin especificar".

4.15 Variable *schedule*

- **Descripción**

  Contiene el horario en que imparte clases el establecimiento.

- **Distribución**

  Contiene 6 valores únicos de los cuales la modalidad DOBLE es el más repetido con 3037 veces.

- **Código de Variable**
  - **Doble:** Operan tanto en horario matutino como vespertino.
  - **Vespertina:** Imparten clases en horario de la tarde.
  - **Matutina:** Ofrecen clases durante las horas de la mañana.
  - **Nocturna:** Ofrecen clases en horario nocturno.
  - **Intermedia:** Imparten clases en un horario que se sitúa entre la jornada matutina y la vespertina, generalmente comenzando a media mañana y finalizando a primera hora de la tarde.

  Formato

  No aplica.

  Datos Faltantes

  Los registros llenados con un patrón de uno o varios guiones, incluidos aquellos que dicen "SIN JORNADA", se deberán colocar como **"Sin especificar"** para referirnos a un valor nulo.

4.16 Variable *plan*

- **Descripción**

  Especifica el plan educativo implementado en el establecimiento educativo.

- **Formato**
  Entrada textual.

- **Valores Posibles y Distribución**

Valor	Conteo
Diario	5723
Fin de semana	2869
Semipresencial	541
A distancia	193
Mixto/Intercalado	5
Otro	1

Cuadro 4 - Valores Posibles y Distribución de la variable "plan"

- **Código de Variable**
  No aplica.

- **Datos Faltantes y Casos No Aplicables**
  Esta variable no cuenta con valores faltantes.

4.17 Variable *departmental*

- **Descripción**
  Especifica la región administrativa creada por el decreto 70-86 del Congreso de la República en base al departamento del establecimiento.

- **Formato**
  Entrada textual.

- **Valores Posibles y Distribución**

Valor	Conteo
Metropolitana	3041
Suroccidente	1994
Central	1306
Noroccidente	760
Nororiente	757
Suroriente	619
Verapaz	488
Petén	367

Cuadro 5 - Valores Posibles y Distribución de la variable "departmental"

- **Código de Variable**
  No aplica.

- **Datos Faltantes y Casos No Aplicables**

  Esta variable no cuenta con valores faltantes.

4.18 Variable *supervisor*

- **Descripción**

  Contiene los nombres de los supervisores asociados a los establecimientos educativos de los registros en el conjunto de datos.

- **Distribución**

  Consta de 651 entradas, con las siguientes frecuencias de los supervisores **más** comunes:

Nombre	Conteo
Carlos Humberto Gonzalez De Leon	333
Miguel Angel Armas Rocha	228
Sin especificar	213
Remy Arturo Sinay Gudiel	167
Juan Enrique Martinez Solano	167

Cuadro 6 - Distribución de la variable "supervisor"

(Otras entradas varían en frecuencia, llegando hasta un total de 1 para varios supervisores menos comunes).

- **Código de Variable**

  No aplica.

- **Formato de la Variable**

  Estructura

  Los nombres de los supervisores están en formato de título (capitalizando la primera letra de cada palabra, sin tildes o caracteres especiales). Preferiblemente se deben incluir los dos nombres y dos apellidos del supervisor del establecimiento.

  Componentes

  →**Nombre Completo del Supervisor:** El nombre del supervisor asociado al establecimiento educativo, sin tildes o caracteres especiales y con la correspondiente capitalización.

- **Datos Faltantes y Casos No Aplicables**

  1. La variable debe contener nombres de supervisores válidos o el valor **"Sin especificar"** para los registros donde no se pueda determinar el supervisor.
  2. Por favor tome en consideración el **NO** ingresar tildes o caracteres especiales.
  3. Aunque no es obligatorio, se recomienda ingresar el nombre completo de los supervisores, de tal manera que se incluyan ambos nombres y apellidos.

- **Ejemplos**
    1. Carlos Humberto Gonzalez De Leon
    2. Miguel Angel Armas Rocha

5. Enlace a Repositorio

Enlace a repositorio → https://github.com/Andrea-gt/data-cleansing-mineduc