Proyecto 1

Data Science

Avances iniciales

1. **Descripción del set de datos**
   - El set de datos cuenta con 23 tablas correspondiente a cada uno de los departamentos de Guatemala, más uno que corresponde solamente a la ciudad de Guatemala.
   - Cada tabla es un archivo CSV separado.
   - Al inicio de los archivos (las primeras filas) hay una pequeña descripción sobre la tabla, además de información extra añadida a los archivos de Excel (esta información es innecesaria para los fines de análisis de datos, ya que es cualitativa, se debe quitar).
   - Al final de los archivos también hay filas extras (copyright, cuántos establecimientos se encontraron, etc.).

   - Columnas (17):
     1. **Código**: código asignado por el ministerio al establecimiento educativo.
     2. **Distrito**: código del distrito en el que se encuentra el establecimiento dentro del municipio.
     3. **Departamento**: nombre del departamento en el que se encuentra el colegio.
     4. **Municipio**: nombre del municipio en el que se encuentra el establecimiento.
     5. **Establecimiento:** nombre del colegio, escuela o instituto.
     6. **Dirección:** dirección específica del establecimiento.
     7. **Teléfono:** número de teléfono del colegio, escuela o instituto.
     8. **Supervisor:** nombre del supervisor encargado del establecimiento.
     9. **Director:** nombre del director encargado del establecimiento.
     10. **Nivel:** Nivel más alto al que llega el colegio, escuela o instituto (ej. Diversificado).
     11. **Sector:** si el establecimiento pertenece al sector oficial, privado, cooperativa, etc.
     12. **Área:** si el establecimiento se encuentra en área urbana o rural.
     13. **Status:** si el colegio está abierto o cerrado actualmente.
     14. **Modalidad:** bilingüe, multilingüe, etc.
     15. **Jornada:** vespertina, matutina, doble, etc.
     16. **Plan:** qué programa de enseñanza tienen, por ejemplo, un día a la semana, fin de semana, presencial, virtual, etc.
     17. **Departamental:** nombre de la región o departamento del país a la que pertenece el establecimiento.

   - **Filas:** Cada archivo tiene una cantidad variable de filas, que varía entre menos de 100 y 1000 establecimientos.

2. **Listar las variables que más operaciones de limpieza necesitarán + estrategia para limpiarlas.**

- **Teléfono:** hay algunas observaciones que no tienen teléfono y hay unas que tienen dos. Proceso de limpieza:
    1. Dependiendo del objetivo para el que se utilizarán los datos, se decidirá si se eliminan las observaciones con teléfono faltante o no (para el análisis del conjunto de datos no necesariamente es necesaria la columna teléfono, así que dependiendo el caso se puede omitir las faltas).
    2. Los colegios que tienen más de un número registrado, se eliminará el segundo, manteniendo los primeros 8 dígitos de la columna.

- **Departamento, Municipio, Departamental:** Revisar que no haya faltas de ortografía o caracteres extraños en algunas observaciones para que luego funcionen bien las agrupaciones.

- **Departamento:** El archivo de “Ciudad Capital” se tiene que cambiar a “Guatemala”, ya que pertenece a este departamento.

- **Director, Supervisor y Dirección:** revisar si la conversión al csv provocó el surgimiento de caracteres extraños en los nombres que llevaban tildes, ñ, etc. Si este es el caso, para realizar la limpieza, se tendrá que reemplazar los caracteres incorrectos por correctos. Se pueden eliminar las tildes y reemplazar por caracteres simples y reemplazar los caracteres extraños que aparecen en vez de la “ñ” por “n”.

- **Para todas las variables:**

    1. Revisar si hay datos faltantes y considerar la importancia de la columna para el análisis que se vaya a hacer con el set de datos. Si la columna es indispensable, eliminar las observaciones con información faltante (dado que no hay columnas numéricas para sacar el promedio y sustituir el valor, sino son todas cualitativas).
    2. El dataset está todo en mayúsculas. Si es necesario, se puede convertir a Sentence Case o a minúsculas.

3. **Estrategias adicionales para la limpieza de datos**

- Eliminar las primeras 27 y las últimas 5 filas de todos los archivos (información innecesaria y no tabular).
- Eliminar la segunda columna y las últimas dos de todos los archivos (están vacías porque fueron utilizadas solamente en el encabezado).
- Unir los datasets horizontalmente para obtener uno solo.