Amado Garcia 181469
Ricardo Valenzuela 18762
Sara Zavala 18893

Avances Proyecto 1 Data Science

1. Descripción de los datos:
   De la página del Ministerio de Educación se extrajeron todos los datos de los establecimientos educativos del país que lleguen hasta el nivel de diversificado. Para hacer esto se descargaron 23 datasets ya que la pagina solo brindaba la información por departamento, cada dataset cuenta con 17 variables, sin embargo, la cantidad de filas si cambia dependiendo del departamento:

	Filas
Alta Verapaz	280
Baja Verapaz	114
Chimaltenango	290
Chiquimula	290
Ciudad Capital	1347
El Progreso	111
Escuintla	496
Guatemala	1238
Huehuetenango	438
Izabal	304
Jalapa	116
Jutiapa	265
Peten	357
Quetzaltenango	431
Quiche	212
Retalhuleu	255
Sacatepequez	276
San Marcos	487
Santa Rosa	140
Solola	138
Suchitepequez	330
Totonicapan	81
Zacapa	93

   Para un total de 8099 filas a la hora de unificar todos los datasets. Todos los datasets estructuraron sus columnas y datos en el idioma español y utilizando mayúsculas, por lo que respetaremos este estándar. Todas las variables son de tipo categórico.

2. Lista de Variables que más operaciones de limpieza necesitan:
   a. Departamento
   b. Teléfono
   c. Establecimiento
   d. Director
   e. Departamental

Amado Garcia 181469
Ricardo Valenzuela 18762
Sara Zavala 18893

3. Operaciones de limpieza a realizar en los set de datos:
    a. Ubicar valores en vacíos y reemplazar con NaN.
        i. En todo el dataset hay datos faltantes ya sea que estén en blanco o tengan una línea para identificar que no existe (-).
        ii. Identificar valores faltantes o llenados incorrectamente con una línea.
        iii. Sustituir estos valores con NaN.
    b. Identificar filas repetidas y eliminarlas:
        i. Tras realizar la limpieza total del dataset se identificarán cuantas filas repetidas hay.
        ii. Se evaluará si realmente son repetidas o no, de serlo serán eliminadas.
    c. Variable Departamental mal identificada en el dataset de Guatemala:
        i. En el set de datos del municipio de Guatemala la variable DEPARTAMENTAL está mal identificada ya que tiene un ";" al final.
        ii. Esto impide que se unifiquen correctamente los set de datos, por lo que se debe corregir el nombre de esta variable en ese set de datos.
    d. Variable Establecimiento:
        i. El principal problema de esta variable es que muchos nombres de los establecimientos contienen comillas (“), esto causa conflicto al leer el .csv y hace que se pierdan muchos datos.
        ii. Se planea reemplazar estas comillas por otro signo de puntuación que no cause problemas con el lector de csv de pandas.
    e. Variable Teléfono:
        i. Hay un conflicto en esta variable relacionada a su tipo de dato, ya que podemos encontrar strings, floats o enteros.
        ii. Se planea convertir todos sus datos a tipo string, ya que no se pueden convertir a tipo int por qué algunos datos contienen caracteres.
    f. Variable Departamental:
        i. Esta variable es redundante ya que contamos con una que identifica el departamento, su único aporte es dividir algunos departamentos por regiones.
        ii. Adicionalmente en varios departamentos se agrega un ";" al final que hace que se tome como un nuevo dato y causa conflictos al querer hacer la tabla de frecuencia.
        iii. Por estos motivos se planea eliminar dicha variable.
    g. Variable Director:
        i. Existe mucha información faltante de los directores, además, esta información faltante se encuentra identificada por líneas, ejemplo: -, --, ---.
        ii. Esta inconsistencia hace que no se puedan analizar bien los datos.
        iii. Se planea reemplazar estos datos faltantes con NaN de numpy con el fin de tener una variable más organizada.