Guatemala 25 de julio de 2024
Arturo Argueta - 21527
Renatto Guzmán - 21646
Gabriel García - 21352
Dariel Villatoro - 20776

Data Science - S10

Avances Proyecto 1

Github: [Arthurs101/Proyecto1_Data_Science (github.com)](https://github.com/Arthurs101/Proyecto1_Data_Science)

Tipo de valor del set de datos

Al juntar todos los datos csv por departamento en uno solo, sin modificar las variables puesto que son las mismas para todos , se obtiene un total de 9331 registros con 17 variables o columnas, las cuales son:

Variable	Descripción
Código	ID único por establecimiento
Distrito	ID que representa el distrito donde se encuentra el establecimiento
Departamento	Nombre del departamento donde se ubica
Municipio	Nombre del municipio en donde se halla
Establecimiento	Nombre del establecimiento
Dirección	Dirección completa del establecimiento
Teléfono	Teléfono del lugar
Supervisor	Nombre del supervisor
Director	Nombre del Director
Nivel	Estudios que imparte el establecimiento (Bachiller, primario , etc)
Sector	Indica si es privado o público

Área	Indica si es Urbano o rural
Status	Indica si está Abierto o Cerrado a operaciones
Modalidad	Referente a la cantidad de lenguas impartidas, Monolingue, Bilingue etc
Jornada	Indica la jornada que tiene disponible (matutina, vespertina, nocturna)
Plan	Diario (regular) o fin de semana
Departamental	si es único o departamental

En DIRECCION, TELEFONO,SUPERVISOR,DIRECTOR hay menor cantidad de datos, indicando que hay faltantes

Variables que requieren de limpieza y sus respectivas operaciones

Antes de iniciar, se hará una limpieza de registros duplicados del dataset.

Código:

- Eliminación de valores no numéricos, para que sea un valor numérico.

Número de teléfono:

- Eliminación de valores no numéricos, es decir los guiones
- En caso de no haber entrada, reemplazar por string de 8 0's

Dirección

- Normalizar abreviaturas, es decir Avenida a Av.
- Convertir a mayúsculas para uniformidad

Supervisor y Director

- Convertir a mayúsculas para uniformidad
- Valores nulos, reemplazar por "no conocido"

Establecimiento

- Convertir a mayúsculas para uniformidad
- aplicar utf8 de encoding para evitar problemas con tildes.

Nivel

- Solo contiene un solo valor, no es necesaria y fue eliminada

Libro de códigos

CODIGO

Tipo de valor: Valores únicos numéricos de entre 7 y 10 dígitos

Ejemplos:
1601013846
1701407546
1703009546
1703009846

DISTRITO

Tipo de valor: Valor numérico de 5 dígitos separado por un guión entre el 2do. y 3er. dígito, o NO DISPONIBLE.

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	01-403	268
2	NO DISPONIBLE	208
3	11-017	176
4	05-033	167
5	01-411	167

DEPARTAMENTO
Tipo de valor: Nombre del departamento

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	CIUDAD CAPITAL	1564
2	GUATEMALA	1477
3	ESCUINTLA	628
4	SAN MARCOS	574
5	HUEHUETENANGO	516

DEPARTAMENTAL
Tipo de valor: Nombre de la región departamental, subsección de departamento

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	GUATEMALA NORTE	1050
2	GUATEMALA SUR	833
3	GUATEMALA OCCIDENTE	785
4	ESCUINTLA	628
5	SAN MARCOS	574

ESTABLECIMIENTO
Tipo de valor: Cadena de texto con el nombre de la institución

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	INSTITUTO NACIONAL DE EDUCACION DIVERSIFICADA	333
2	INSTITUTO NACIONAL DE EDUCACIÓN DIVERSIFICADA	94
3	CENTRO DE EDUCACION EXTRAESCOLAR -CEEX-	33
4	INSTITUTO DE EDUCACION DIVERSIFICADA POR COOPERATIVA DE ENSEÑANZA	27
5	INSTITUTO DIVERSIFICADO POR COOPERATIVA	23

DIRECCION

Tipo de valor: Cadena de texto con la dirección de la institución

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	CABECERA MUNICIPAL	271
2	BARRIO EL CENTRO	61
3	NO DISPONIBLE	51
4	BARRIO EL PORVENIR	20
5	BARRIO EL CALVARIO	19

TELEFONO

Tipo de valor: Número de 8 digitos de telefonos

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	22067425	21
2	79480009	14
3	77602663	13
4	22093200	12
5	45353648	12

SUPERVISOR

Tipo de valor: Nombre del supervisor

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	CARLOS HUMBERTO GONZALEZ DE LEON	333
2	MIGUEL ANGEL ARMAS ROCHA	228
3	NO CONOCIDO	209
4	REMY ARTURO SINAY GUDIEL	16
5	JUAN ENRIQUE MARTINEZ SOLANO	

SECTOR
Tipo de valor: Indica si es privado o público

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	PRIVADO	7956
2	OFICIAL	990
3	COOPERATIVA	245
4	MUNICIPAL	140

AREA
Tipo de valor: Indica si es Urbano o rural

Esta tabla solo cuenta con 3 valores, que son los siguientes:

Num.	Valor	Cantidad en los datos
1	URBANA	7606
2	RURAL	1723
3	SIN ESPECIFICAR	2

STATUS
Tipo de valor: Indica si está Abierto o Cerrado a operaciones

Los 4 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	ABIERTA	6545
2	CERRADA TEMPORALMENTE	2663
3	TEMPORAL TITULOS	120
4	TEMPORAL NOMBRAMIENTO	3

MODALIDAD
Tipo de valor: Indica si habla uno o dos idiomas

Los 2 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	MONOLINGUE	9039
2	BILINGUE	292

JORNADA
Tipo de valor: Indica la jornada que tiene disponible

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	DOBLE	3037
2	VESPERTINA	2567
3	MATUTINA	2382
4	SIN JORNADA	964
5	NOCTURNA	289

PLAN
Tipo de valor:

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	DIARIO(REGULAR)	5721
2	FIN DE SEMANA	2328
3	SEMIPRESENCIAL (FIN DE SEMANA)	481
4	SEMIPRESENCIAL	398

	(UN DIA A LA SEMANA)	
5	A DISTANCIA	143

MUNICIPIO

Tipo de valor: Nombre del municipio

Los 5 valores que más se repiten son los siguientes:

Num.	Valor	Cantidad en los datos
1	ZONA 1	633
2	MIXCO	428
3	VILLA NUEVA	370
4	QUETZALTENANGO	248
5	RETALHULEU	185