Guía para la validación de datos

Plataforma Ciudadana de Datos Abiertos

I. RECOMENDACIONES GENERALES

Para que tus datos sean más fáciles de **interpretar, procesar y utilizar** es importante seguir una serie de recomendaciones en cuanto a su estructura y el contenido de sus columnas.

Las herramientas de análisis y visualización de datos disponibles dentro de la Plataforma Ciudadana de Datos Abiertos están diseñadas para que conjuntos de datos con distintos tipos de información (geográfica, temporal, numérica o de texto) puedan ser aprovechados por la ciudadanía. En particular, la plataforma cuenta con tres herramientas principales:

1. Herramienta de visualización de datos (Sistema Ajolote)
2. Herramienta de validación de datos
3. Herramienta de cruce de datos

Cada una de estas herramientas requiere un nivel mínimo de estandarización de los datos que son cargados. Esta guía tiene como objetivo ayudar a la persona usuaria a revisar y en su caso ajustar sus conjuntos de datos antes de cargarlos a la plataforma.

La información que encontrarás aquí te ayudará a validar o mejorar la estructura de tus datos, lo cual tiene los siguientes beneficios:

1. Podrás **cruzar tu información** con otra información de la Plataforma Ciudadana de Datos Abiertos.
2. Podrás **generar gráficas y mapas** interactivos de tu conjunto de datos con la herramienta de visualización.
3. Tus datos podrán **aprovecharse de manera más efectiva** por otras personas que utilicen o colaboren en la plataforma.

Contenido de las columnas

1. Para **columnas de texto**:
    a. En la medida de lo posible, recomendamos que tus columnas de texto tengan un catálogo reducido de posibles valores. Revisa con detenimiento si existen faltas de ortografía o errores de escritura.
    b. La base de datos debe estar codificada en UTF-8 para evitar errores de interpretación de caracteres especiales (alternativamente puedes cargar tu info sin acentos).
2. Para **columnas numéricas**:
    a. Los valores nulos de las variables numéricas deberán estar marcados como "NaN".
3. Para **columnas de georeferenciación**:

   a. La georeferenciación debe estar en dos columnas separadas llamadas "latitud" y "longitud".
   b. Las columnas de georeferencia de latitud y longitud deben ser de formato numérico; aquellas que especifiquen nombres de alcaldías, colonias, localidades o AGEBs deben ser en formato de texto.

Estructura de la base

En general, es recomendable que tus tablas de datos cuenten con las siguientes características:

1. Cada columna es un campo, atributo o variable.
2. Cada fila es un registro u observación.
3. Los nombres de las columnas deberán ser entendibles y preferentemente estar en minúsculas, sin acentos y con guiones bajos en lugar de espacios.
4. No debe utilizarse más de un tipo de dato en una misma columna (ejemplo: mezclar texto con valores numéricos).
5. Cuando los valores representan magnitudes, es necesario que permanezcan como datos numéricos y que la unidad de medida se agregue a la descripción del título del campo, p. ej. “Distancia en KM”; o bien en el nombre de la columna, p. ej “distancia_km2”.
6. Los campos numéricos, incluyendo los monetarios, deben permanecer en un formato numérico. En este último caso se debe evitar el uso de símbolos monetarios y mejor indicarlo en el título del campo.
7. Con respecto al nombre de las columnas:
   a. En minúsculas, sin espacios ni acentos (sustituir los espacios por guiones bajos).
   b. No pueden repetirse los nombres de las columnas.
   c. Para las columnas de cruce, éstas deberán tener alguno de los siguientes nombres:
      i. ageb
      ii. colonia
      iii. alcaldia
      iv. localidad
      v. fecha

A continuación se muestra una ejemplificación de estos puntos:

Malas prácticas

id	edad	monto_anual	porcentaje
1	28	$550	67%
2	31 años	443.24	27.2
3	44 a	999 pesos	98.4
4	55	134	0.99
5	2 años y 1 mes	1567.65	25%
6	22	$134	69.2
7	43	98	4 porciento
8	78 anios	$135	2.40%
9	81	0.89	145%
10	53	134	21

Buenas prácticas

id	edad	monto_anual_mxn	porcentaje
1	28	550	67
2	31	443.24	27.2
3	44	999	98.4
4	55	134	0.99
5	2	1567.65	25
6	22	134	69.2
7	43	98	4
8	78	135	2.4
9	81	0.89	145
10	53	134	21

8. Es más recomendable que tus datos estén en formato largo (long) que ancho (wide).

En el formato long (largo o apilado), cada observación se representa en una única fila de la tabla, y las variables se encuentran en columnas separadas. Esto implica que puede haber múltiples filas para una misma colonia o fecha (por ejemplo) si se registran múltiples mediciones o categorías para esa colonia o fecha. En este formato, generalmente hay una columna que indica la variable y otra columna que contiene los valores correspondientes.

En el formato wide (ancho), cada observación se representa en una única fila de la tabla, y las categorías de las variables se expanden horizontalmente en columnas. Cada categoría tiene su propia columna en la tabla, y los valores correspondientes a cada observación se encuentran en las celdas respectivas.

A continuación encontrarás un ejemplo de formato wide y long para una tabla de datos:

Formato *(wide)*

id	alcaldia	pob_mujer	pob_hombre
1	BENITO JUAREZ	104789	103784
2	AZCAPOTZALCO	89402	88340
3	COYOACAN	122398	122200
4	MILPA ALTA	54897	51928
5	TLALPAN	68455	68245
6	MIGUEL HIDALGO	283779	282496
7	ALVARO OBREGON	72819	71536
8	IZTACALCO	182736	181453
9	IZTAPALAPA	4927112	4925829
10	TLAHUAC	39182	37899

Formato *(long)*

id	alcaldia	sexo	poblacion
1	BENITO JUAREZ	Mujer	104789
2	BENITO JUAREZ	Hombre	103784
3	AZCAPOTZALCO	Mujer	89402
4	AZCAPOTZALCO	Hombre	88340
5	COYOACAN	Mujer	122398
6	COYOACAN	Hombre	122200
7	MILPA ALTA	Mujer	54897
8	MILPA ALTA	Hombre	51928
9	TLALPAN	Mujer	68455
10	TLALPAN	Hombre	68245

Puedes consultar más ejemplos de bases de datos con estructuras correctas en el siguiente enlace:
https://archivo.datos.cdmx.gob.mx/plantillas_STI/plantilla_general.xlsx

II. CATÁLOGOS DE REFERENCIA

Los catálogos de referencia son documentos que definen los valores permitidos para ciertos datos. Estos catálogos son utilizados como datos maestros para establecer cuáles valores son válidos dentro de las columnas o variables de una base de datos.

Desagregaciones geográficas

Para esta plataforma, los catálogos de referencia son especialmente importantes para las columnas o variables geográficas. En este sentido, existen cuatro catálogos de referencia que es necesario tomar en cuenta para ordenar y limpiar tus bases de datos antes de cargarlas a la plataforma:

1. AGEB
2. Colonias
3. Localidades
4. Alcaldías

Es necesario que al cargar información, los nombres de las alcaldías, las localidades, las colonias y las claves de AGEB sean exactamente a los que se presentan en los catálogos de referencia.

En el siguiente enlace, podrás consultar estos documentos:
https://archivo.datos.cdmx.gob.mx/plantillas_STI/plantilla_geografica.xlsx

Coordenadas geográficas

Para la información que incluya coordenadas geográficas es importante lo siguiente:

1. Las coordenadas deben estar separadas en dos columnas (una de latitud y una de longitud)
2. Las columnas deben llamarse "latitud" y "longitud"
3. Las columnas deben estar en formato numérico.

III. FORMATOS DE FECHA

En el Repositorio Ciudadano de Datos Abiertos es necesario que toda la información que tenga un componente temporal esté acompañada o referida en la base de datos con al menos una columna de fecha.

Un componente temporal se refiere a toda aquella información que indica una fecha o un momento en el tiempo, por ejemplo un trimestre, un semestre o un año.

Periodos de tiempo

Los periodos de tiempo más comunes en los que se genera la información son: Anual, Semestral, Trimestral, Bimestral, Mensual, Semanal y Diario. Para poder generar gráficas con líneas de tendencia dentro del Repositorio Ciudadano de Datos Abiertos o para cruzar información de distintas fuentes a través de la fecha es indispensable que cada columna que indica un periodo de tiempo - en texto - vaya acompañada de otra columna con formato fecha. Esta fecha debe indicar el día de la fecha de cierre del periodo correspondiente.

En el siguiente enlace encontrarás los distintos periodos y ejemplos de cómo indicar las fechas de referencia para cada uno de ellos:
https://archivo.datos.cdmx.gob.mx/plantillas_STI/plantilla_fechas.xlsx

Formato de fecha

Existen diversos formatos para registrar una fecha; sin embargo, como buena práctica se adoptó el estándar de formato año-mes-día de la siguiente manera: AAAA-MM-DD

En la siguiente tabla se muestran ejemplos de otros formatos y de fecha y cómo deben adaptarse al formato estándar:

Otros formatos	**Formato estándar** (aaaa-mm-dd)
16-ene-21	2021-01-16
18-07-23	2023-07-18
08-21-2022	2022-08-21
05-junio-1996	1996-06-05

Para cargar tus datos es indispensable que las fechas estén en el estándar señalado.