GUÍA PRÁCTICA

CÓMO TENER DATOS DE CALIDAD EN HOJAS DE CÁLCULO

VERSIÓN 1.0

GOBIERNO DE LA CIUDAD DE MÉXICO | ADIP

Elaborado por la Agencia Digital de Innovación Pública de la Ciudad de México
Plaza de Las Vizcaínas 30, Centro Histórico de la Cdad. de México,
Centro, Cuauhtémoc, 06000 Cuauhtémoc, CDMX
Noviembre 2020

Guía práctica: Cómo tener datos de calidad en hojas de cálculo

---

Las interfaces de hojas de cálculo (como Microsoft Excel y Google Spreadsheets, entre otras) son muy útiles para que cualquier usuario pueda almacenar, procesar y analizar datos fácilmente, siendo uno de los principales formatos usados a nivel mundial. Sin embargo, las mismas funcionalidades que las hacen tan sencillas de utilizar nos permiten llevar a cabo prácticas que afectan la calidad de los datos y obstaculizan su uso a largo plazo.

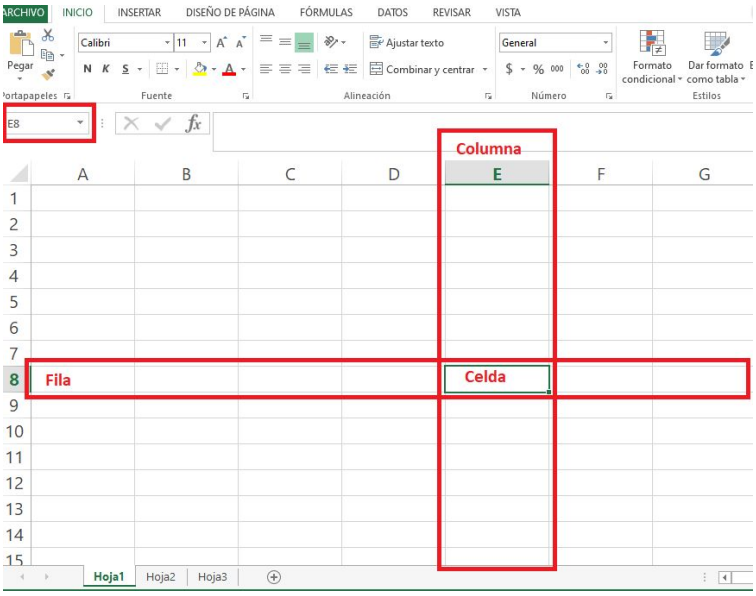
¿Para qué tener datos de calidad en hojas de cálculo?

- Para poder almacenarlos en otros formatos y disminuir el espacio de almacenamiento que requieren.
- Para poder procesarlos y analizarlos con herramientas como R, Stata, etc.
- Para poder hacer visualizaciones de datos.
- Para poder integrarlos con otros conjuntos de datos o bases de datos.
- Para reducir el tiempo de limpieza y transformación requerido para su uso.

Los tres principios de tidy data:

La mayoría de los conjuntos de datos en hojas de cálculo son tablas compuestas de filas y columnas, donde las filas representan un registro u observación y las columnas representan un atributo, variable o campo. Según los tres principios de tidy data:

1. Cada columna es un campo, atributo o variable.
2. Cada fila es un registro u observación.
3. Cada celda es un valor.



Recomendaciones básicas:

1. Sobre la estructura de los datos
    - La primera fila debe tener los nombres de los campos, atributos o variables (las etiquetas de cada columna).
    - Desde la segunda fila en adelante, sólo debe haber datos, pero nunca un encabezado.
    - No deben existir celdas vacías entre los encabezados y la primera fila de datos.
    - Los nombres de las columnas deben ser únicos.
    - Cada columna debe representar un atributo o campo.
    - La primera columna debe ser un identificador de registro único que permita identificar cada registro u observación, se recomienda usar el sufijo "id".
    - Cada fila debe representar un registro u observación.
    - No utilizar más de un tipo de dato en cada columna. Ejemplo:

Mala práctica	
numero_personas	edad
35	25 años
treinta	15

↓

Buena práctica	
numero_personas	edad
35	25
30	15

Ejemplo de una buena estructura:

ID	apellido_paterno	apellido_materno	nombres	fecha_nac	sexo	num_hijos
001	Mendoza	Fuentes	Alina	1992-10-29	M	0
002	Zavala	Araiza	Miguel	1975-09-23	H	1
003	Muñoz	Tapia	Alejandra	1989-01-09	M	2
004	Merino	Mora	Claudia	1983-07-14	M	0

2. Sobre los nombres de las columnas o encabezados
    - Los encabezados de columna deben ser claros y auto-descriptivos, en la medida de lo posible.
    - No utilizar espacios en blanco. Utilizar puntos (.) o guion bajo (_) para separar las palabras.
    - No dejar espacios en blanco al inicio o final de las palabras.

- No usar caracteres especiales como ?, $, *, +, #, (,), -, /, }, {, |, >, <, entre otros.
- No usar mayúsculas.
- No usar acentos, ñ o diéresis (¨).
- No utilizar números al inicio de los nombres de las columnas.
- No duplicar los nombres de las columnas.

3. Sobre el uso de las celdas:
   - No combinar/ fusionar celdas.
   - No ocultar filas o columnas.
   - No dejar filas vacías.
   - No dejar celdas vacías.
     - Cuando existan valores faltantes éstos se deben indicar de forma explícita (ya sea con NA, null, no disponible, etc.).
     - No utilizar el número cero (0) como equivalente a un valor faltante.
   - No utilizar comentarios o notas a las celdas.
   - No utilizar los distintos formatos disponibles para las celdas (fecha, porcentaje, moneda, etc.).
   - No hacer más de una tabla por pestaña u hoja de cálculo.
   - No guardar imágenes, gráficas u otros archivos sobre las celdas (ejemplo: evitar poner logos).

4. Recomendaciones para distintos tipos de datos

4.1) Fechas y horas[1]

- La fecha debe estar en formato AAAA-MM-DD. El año siempre debe escribirse a cuatro dígitos.
- Las horas deben estar en formato 24 horas HH:MM:SS

4.2) Números

- El separador decimal debe ser el punto (.)
- En números menores a 1 escribir el cero antes del punto.
- No utilizar separadores de miles (como comas o espacios).
- No agregar símbolos monetarios o de unidades de medición en la misma celda que los números. Utilizar una columna adicional para tal información o escribir en decimales en el caso de los porcentajes. Ejemplos:

Mala práctica
participacion_mercado
25%
45%

[1] *Se recomienda el uso según la Norma ISO-8601,* https://www.iso.org/iso-8601-date-and-time-format.html

30%

↓

Buena práctica
participacion_mercado
0.25
0.45
.30

Mala práctica	
peso_paciente	sexo
60 kg	F
80 kg.	M
96 kilos	M

↓

Buena práctica		
peso_paciente	unidad_medida	sexo
60	kg	F
80	kg	M
96	kg	M

- En los números negativos se debe incluir el símbolo menos "-" antes del número, sin dejar espacio en blanco entre ellos.

4.3) Texto

- No usar diferentes palabras o frases para referirse a la misma cosa. Por ejemplo, utilizar "CDMX", "Ciudad de México", "Cd. de México" y "Cd. de Mex." en la misma columna. Los catálogos te pueden ayudar a evitar este problema (link).

- No agregar columnas para resúmenes de datos (como promedios o totales) en el cuerpo de la tabla. Utilizar una tabla separada para esto.

5. Sobre cómo mejorar la captura de los datos

- Utiliza validadores para reducir el número de errores humanos en la captura. Puedes utilizar menús desplegables o incluso utilizar herramientas como google forms, monday, SurveyMonkey, entre otras, que alimenten automáticamente una hoja de cálculo.

- Cada pieza de información debe tener su propia celda. Es decir, es conveniente descomponer los campos en campos más pequeños para poder manejar la información más fácilmente.

Ejemplo de información que puede ser separada en elementos más pequeños:

nombre	**pago**
Pedro López R.	2500 pesos m.n. pagado a tiempo
Juan Pérez Pérez	3200 pesos m.n. pago atrasado
María Ramírez	5000 USD pago atrasado

↓

Ejemplo de la información descompuesta en pedazos pequeños:

apellido_pat	apellido_mat	nombre	monto	moneda	estatus
López	Ramírez	Pedro	2500	MXN	A tiempo
Pérez	Pérez	Juan	3200	MXN	Atrasado
Ramírez	Santillán	María	50	USD	Atrasado

- No utilizar colores, negritas u otros formatos como una forma de registrar información, ya que si se exporta el archivo a otro formato se pierde el formato y con eso la información.

Ejemplo:
No colorear celdas de un color para indicar que algo ya se atendió o algo es prioridad. En lugar de eso, poner una columna adicional para registrar esa información.

- Transformar las fórmulas en valores estáticos después de que cumplan con su cometido (después de haber hecho los cálculos que se requerían), para evitar errores humanos al manejar los datos y poder guardarlos en distintos formatos. Sin embargo, es importante mantener un registro de las fórmulas utilizadas, ya sea en un diccionario de datos u otra documentación o en una columna extra dentro del conjunto de datos.

6. Estandarización e interoperabilidad de algunos formatos de datos

Para mejorar el análisis, compartición y comprensión de los datos es importante homologar el formato de aquellos datos que se repiten y usan frecuentemente, como los siguientes:

- En todos los casos donde los datos puedan usar un catálogo, preferir el uso de catálogos consensados nacional o internacionalmente como, por ejemplo: Catálogo de unidades económicas del INEGI, Catálogo de Clasificación Internacional de Enfermedades de la OMS, etcétera.
- Para el registro de datos geográficos, utilizar el Catálogo Único de Claves de Áreas Geoestadísticas Estatales, Municipales y Localidades
- Se sugiere utilizar los códigos numéricos de estados, municipios y localidades, que asigna el catálogo.
- Para datos geográficos que referencian otros países utilizar el ISO 3166-1 que proporciona códigos para nombres de países y otras dependencias administrativas.
- Se recomienda usar el sistema de código de tres letras (alfa-3) para identificar los países, por ejemplo: BRA para Brasil, MEX para México, USA para Estados Unidos de América.
- Para códigos de monedas internacionales utilizar el estándar ISO 4217, por ejemplo: CLP para los pesos chilenos,USD para los dólares estadounidenses o EUR para Euros.
- Los códigos postales deben estar contenidos en un campo llamado "codigo_postal" y seguir el formato definido por el Servicio Postal Mexicano.
- La columna que contenga el campo “codigo_postal” debe tener un formato de texto, para evitar su confusión con valores numéricos o su transformación en fechas.
- Cuando se desagreguen datos por sexo, recordar que se refiere a la condición biológica que distingue a las personas entre “Hombre” y “Mujer” y no confundir con género (femenino, masculino, transgénero, no-binario, etcétera).
- Cuando se codifique el sexo, utilizar H o 1 para hombre y M o 2 para mujer, tal como lo hace el INEGI.
- Acompañar el archivo de hojas de cálculo con un Diccionario de Datos que permita entender a qué se refiere y qué valores puede tomar cada atributo o campo del conjunto de datos.
- Por último, utilizar la codificación de caracteres según el esquema de formato UTF-8, ya que representa todos los caracteres necesarios para la escritura de los idiomas hablados en la actualidad[2].

[2] Ejemplo: https://soporte.newslettersoft.com/hc/es/articles/216592343-C%C3%B3mo-generar-un-fichero-CSV-codificado-en-UTF-8-en-Excel