GUÍA

VERSIÓN 1.1

CALIDAD DE DATOS EN HOJAS DE CÁLCULO

GOBIERNO DE LA CIUDAD DE MÉXICO | ADIP

Elaborado por la Agencia Digital de Innovación Pública
Plaza de Las Vizcaínas 30, Centro Histórico de la Cdad. de México,
Centro, Cuauhtémoc, 06000 Cuauhtémoc, CDMX
Marzo 2021

Presentación

El presente documento fue elaborado por la Agencia Digital de Innovación Pública como parte de la implementación de la Política de Gestión de Datos de la Ciudad de México. Tiene como objetivo servir de guía a los Entes Públicos de la Administración Pública de la Ciudad de México para que los datos que generan, recolectan, analizan y publican en aplicaciones de hojas de cálculo cumplan con ciertos criterios.

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.

La presente guía está basada en los principios de *tidy data*. Si bien estos principios no son exclusivos para hojas de cálculo, éstas son una de los principales métodos de organización de conjuntos de datos. Por tal razón, este documento se enfoca en ese medio.

Contenido

¿Qué es tidy data?

Tidy data o "datos ordenados" se refiere a una serie de recomendaciones de cómo estructurar los conjuntos de datos para facilitar su análisis. Los principios de *tidy data* proporcionan una manera estándar de organización de los datos, lo cual facilita su limpieza y el procesamiento. El hecho de tener una estructura predeterminada significa que las personas usuarias no tienen que perder tiempo y esfuerzos en descifrarla.

El estándar de tidy data fue diseñado para facilitar la exploración y el análisis inicial de los datos, así como para simplificar el desarrollo de herramientas de análisis que sean compatibles entre ellas.[1]

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

- ➔ **Almacenarlos** en varios formatos y disminuir su volumen
- ➔ **Procesarlos y analizarlos** con distintas herramientas como R, Python y Stata.
- ➔ Hacer **visualizaciones y tableros** de datos
- ➔ **Integrarlos** con otros conjuntos de datos o bases de datos
- ➔ Reducir el tiempo de **limpieza y transformación** requerido para su uso.

[1] Para ver más sobre *Tidy Data*, consultar R-Project

Los tres principios de tidy data

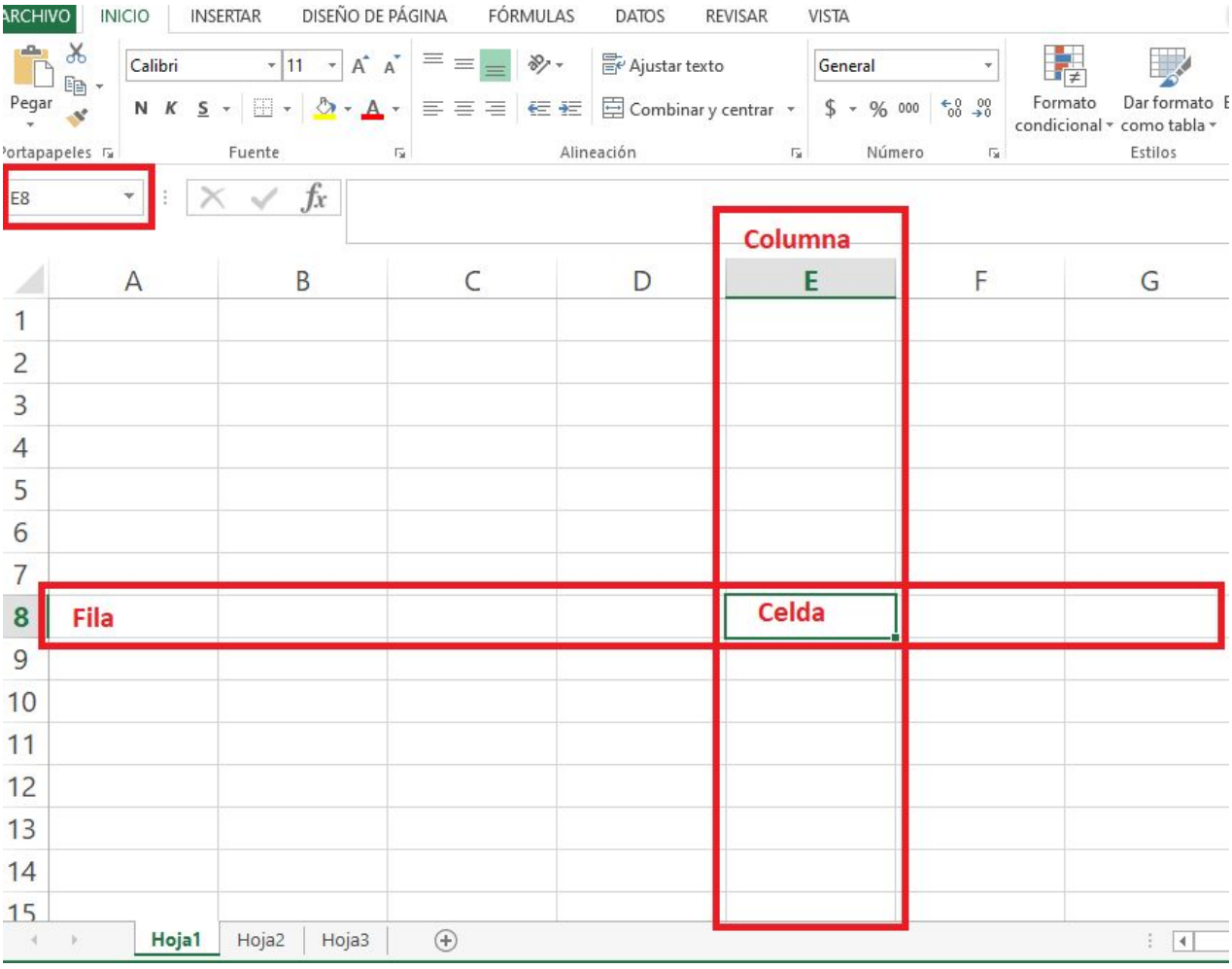
Los conjuntos de datos en hojas de cálculo son **tablas** compuestas de filas y columnas. Según los tres principios de tidy data:

1. Cada columna es un campo, atributo o variable[2]

2. Cada fila es un registro u observación[3]

3. Cada celda es un valor

Lo anterior se ilustra en la Imagen 1.



[2] Se entiende como atributo, campo o variable como cualidad o característica de una observación o entidad dentro de una base de datos o conjunto de datos.
[3] Se entiende como registro u observación como cada uno de los elementos (sujetos, observaciones, hechos) dentro de una base de datos o conjunto de datos; usualmente escritas en filas o tuplas.

Imagen 1. Los tres principios de Tidy Data.

Recomendaciones básicas: filas y columnas

➔ La primera fila debe tener la etiqueta[4] de cada una de las columnas. Es decir, el nombre de los campos, atributos o variables, tal como se puede ver en la imagen 2.

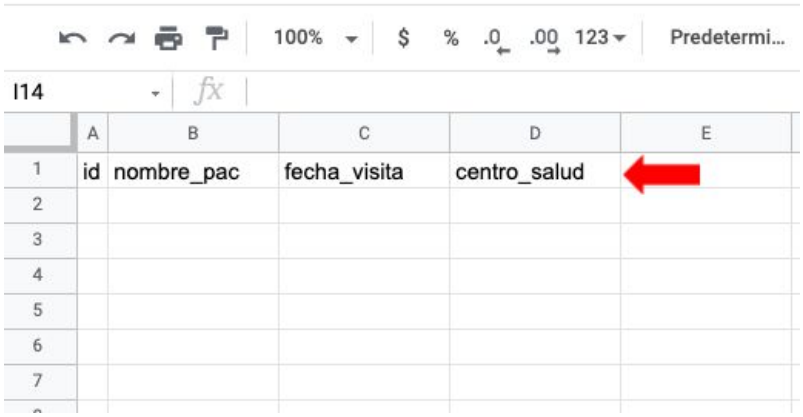


Imagen 2. Encabezados de columna

➔ Cada columna debe representar un campo, atributo o variable

➔ Cada fila debe representar un registro u observación

➔ La primera columna debe ser un identificador de registro único que permita identificar cada registro y observación. Se recomienda utilizar el sufijo “id”, como se muestra en la imagen 3.

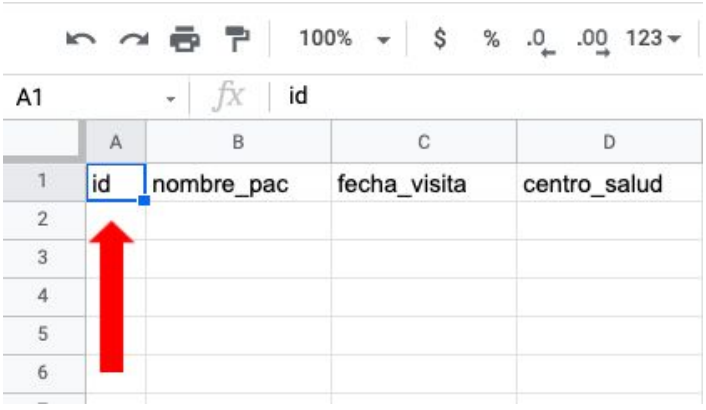


Imagen 3. Utilización de “id”

[4] Se entiende por etiqueta el nombre que se le da a un atributo, campo o variable

- ➔ Los encabezados de columna deben ser auto-descriptivos en la medida de lo posible.
- ➔ Los encabezados no deben contener espacios en blanco. De preferencia utilizar puntos (.) o guión bajo (_) para separar las palabras.
- ➔ Los encabezados no deben tener espacios en blanco al inicio o final.
- ➔ No se deben usar caracteres especiales como ?, !, %,=, “, ”, $, *, +, #, (,), -, /, }, {, |, >, <, entre otros.
- ➔ Los encabezados no deben contener mayúsculas
- ➔ No usar acentos (´) , ñ o diéresis (¨).
- ➔ No utilizar números al inicio de los encabezados
- ➔ No duplicar los nombres de las columnas. Es decir, cada columna debe tener un nombre único.
- ➔ No se debe utilizar más de un tipo de dato en cada columna. Es decir, cada variable, campo o atributo debe tener utilizar únicamente un tipo de dato. En la Imagen 4 se puede ver la comparación entre una mala y una buena práctica.

	A	B	C	D
1	numero_personas	edad		
2	35	25 años		
3	treinta	15	←	mala práctica
4	24 personas	18 años de edad		

	A	B	C	D
1	numero_personas	edad		
2	35	25		
3	30	15	←	buena práctica
4	24	18		
5				

Imagen 4. Buenas y malas práctica de usos de tipo de datos

Recomendaciones básicas: uso de celdas

➔ No deben existir celdas vacías en el documento. Particularmente, no deben existir entre el encabezado de columnas y la primera fila de observaciones, tal como se muestra en la imagen 5.

A	B	C	D	E	F	G
numero_personas	edad	dia	mes	año		
					←	mala práctica: filas vacías
35	25	2	1	2020		
30	15	27	2	2020		
24	18	8	3	2021		

Imagen 5. Ejemplo de mala práctica: filas vacías

➔ A partir de la primera fila sólo deben haber datos, nunca encabezados.

➔ No combinar/ fusionar celdas, como se puede observar en la imagen 6.

	A	B	C	D	E	F	G
1	**Eventos del año**					←	mala práctica: celdas combinadas
2	numero_personas	edad	dia	mes	año		
3	35	25	2	1	2020		
4	30	15	27	2	2020		
5	24	18	8	3	2021		

Imagen 6. Ejemplo de mala práctica: celdas combinadas

➔ No ocultar filas o columnas.

➔ 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), tal como se ve en la imagen 6.

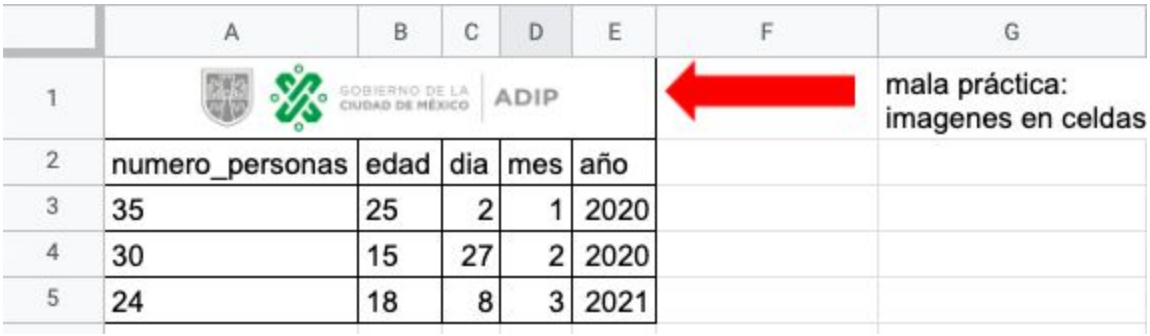


Imagen 6. Ejemplo de mala práctica: imágenes en hojas de cálculo

Recomendaciones básicas: uso de distintos tipos de datos

Fechas y horas[5]

- ➔ 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

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).
- ➔ En los números negativos se debe incluir el símbolo menos "-" antes del número, sin dejar espacio en blanco entre ellos.
- ➔ 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, como se muestra en las imágenes 7 y 8.

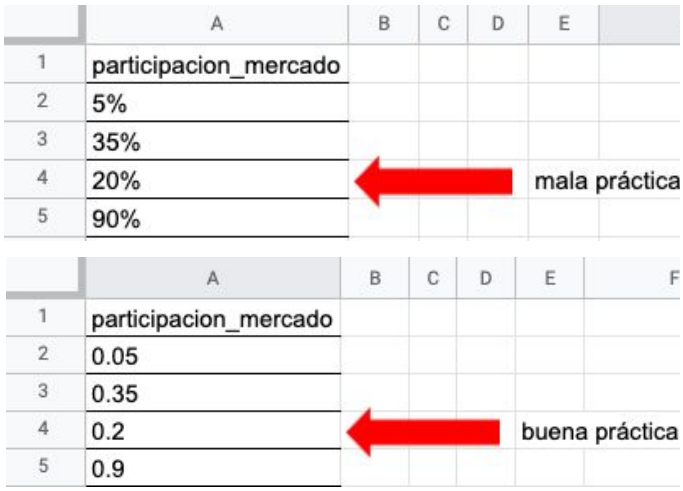


Imagen 7. Ejemplo de mala y buena práctica de uso de porcentajes

[5] Se recomienda el uso según la Norma ISO-8601

	A	B
1	peso_paciente	sexo
2	60 kg	F
3	80 kg.	M
4	96 kilos	M
5		
6		
7		
8	mala práctica	

vs.

	A	B	C
1	peso_paciente	unidad_ medida	sexo
2	60	kg	F
3	80	kg	M
4	96	kg	M
5			
6			
7			
8	buena práctica		

Imagen 8. Ejemplo de mala y buena práctica en el uso de unidades de medida

➔ No agregar columnas o filas con resúmenes de datos (como promedios o totales) en el cuerpo de la tabla, como se muestra en la imagen 9.

	A	B	C	D	E
1	id	peso_paciente	unidad_ medida	sexo	
2	1	60	kg	F	
3	2	80	kg	M	
4	3	96	kg	M	
5	promedio mes 1	78.66666667			
6	4	90	kg	F	
7	5	75	kg	F	
8	6	68	kg	M	
9	7	42	kg	M	
10	promedio mes 2	68.75	kg		

Imagen 9. Ejemplo de mala práctica: resúmenes de datos en el cuerpo de la tabla

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.

➔ Utilizar catálogos de datos de referencia siempre que sea posible para tener vocabularios controlados.

Recomendaciones básicas: cómo mejorar la captura de datos

➔ Utiliza validadores para reducir el número de errores humanos en la captura. Puedes utilizar menús desplegables, como se muestra en la imagen 10.

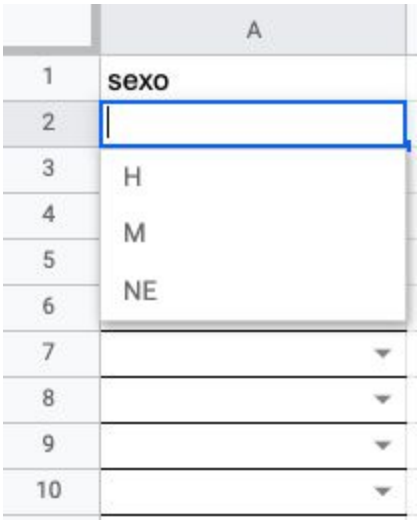


Imagen 10. Uso de validadores en un conjunto de datos para indicar sexo: Hombre (H), Mujer (M) y No especificado (NE)

➔ Cuando sea posible, utilizar formularios web, ya sean desarrollados desde cero o utilizando herramientas como formularios de Google, Monday, SurveyMonkey y Typeform que generen automáticamente un conjunto de datos.

➔ 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, como se muestra en las imágenes 11 y 12.

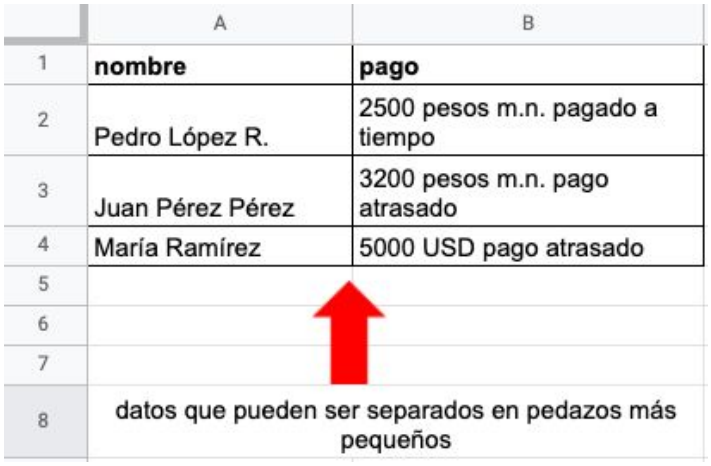


Imagen 11. Ejemplo de información que puede descomponerse en un mayor número de celdas

	A	B	C	D	E	F
1	**apellido_pat**	**apellido_mat**	**nombre**	**monto**	**moneda**	**estatus**
2	López	Ramírez	Pedro	2500	MXN	A tiempo
3	Pérez	Pérez	Juan	3200	MXN	Atrasado
4	Ramírez	Santillán	María	50	USD	Atrasado

Imagen 12. Ejemplo de la información de la imagen 11 descompuesta en partes para facilitar su procesamiento y análisis

➔ 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. Por ejemplo, no colorear celdas de un color para indicar que algo está atrasado, como se muestra en la imagen 13. En lugar de eso, poner una columna adicional para registrar esa información, como en la imagen 12.

	A	B	C	D	E
1	**apellido_pat**	**apellido_mat**	**nombre**	**monto**	**moneda**
2	López	Ramírez	Pedro	2500	MXN
3	Pérez	Pérez	Juan	3200	MXN
4	Ramírez	Santillán	María	50	USD

Imagen 13. Ejemplo de mal uso del formato de celdas para indicar 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 en un diccionario de datos.

Ejemplo de una estructura de datos *tidy*

	A	B	C	D	E	F	G
1	ID	apellido_paterno	apellido_materno	nombres	fecha_nac	sexo	num_hijos
2	1	Mendoza	Fuentes	Alina	1992-10-29	M	0
3	2	Zavala	Araiza	Miguel	1975-09-23	H	1
4	3	Muñoz	Tapia	Alejandra	1989-01-09	M	2
5	4	Merino	Mora	Claudia	1983-07-14	M	0
6							

GOBIERNO DE LA CIUDAD DE MÉXICO | ADIP

Elaborado por la Agencia Digital de Innovación Pública
Plaza de Las Vizcaínas 30, Centro Histórico de la Cdad. de México,
Centro, Cuauhtémoc, 06000 Cuauhtémoc, CDMX
Marzo 2021