TUIA

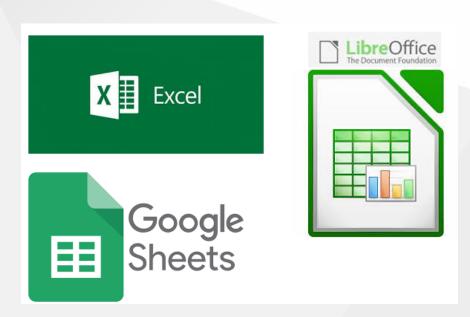
Fundamentos de la Ciencia de Datos

Unidad 1 - Planilla de cálculo



Planillas de cálculo

Los software de hojas de cálculo (tales como Microsoft Excel, Google Sheets y LibreOffice Calc, entre otros) son herramientas valiosas para ingresar, organizar y almacenar datos. También se pueden utilizar para cálculos, análisis y visualizaciones.



Las más conocidas, Excel 365, Google Sheets y LibreOffice Calc: ¿Cuál es la mejor opción?

Las aplicaciones más comunes son Excel 365, Google Sheets y LibreOffice:

- Excel 365 es una aplicación de Microsoft Office que se ejecuta en el escritorio de la computadora y también en la nube.
- Por otro lado, <u>Google Sheets</u> es una aplicación de hojas de cálculo en línea que se ejecuta en la nube de Google.
- Por último, **LibreOffice** es una suite de oficina gratuita y de código abierto que incluye una aplicación de hojas de cálculo llamada <u>Calc</u>.

Buenas prácticas en el uso de planillas de cálculo

A continuación, se describirán algunos principios considerados 'buenas prácticas' que serán de utilidad para proyectos futuros (Broman & Wo, 2017).

Los mismos permitirán crear hojas de cálculo menos propensas a errores, más fáciles de procesar con distintos lenguajes de programación y más fáciles de compartir con colaboradores y el público.

Consistencia/coherencia

Códigos de las variables categóricas

Seleccionar una opción, y usar siempre la misma. Por ejemplo, en la variable *sexo* podemos seleccionar entre:

Femenino	Masculino	No_binario
femenino	masculino	no_binario
f	m	nb
F	M	X
Fem	Masc	Nobin

Los nombres de las variables

Si es un archivo tenemos una variable llamada 'Provincias', usar el mismo nombre para la variable en otros archivos (no usar: 'Provincia', 'Prov', etc).

Por ejemplo, si en archivos que guardamos datos de experimentos de un laboratorio registro los datos de 'glucosa en sangre después de 3 horas' como 'Glucosa_3horas' en unos, 'Gluc_3h' y 'Glucosa 3 horas' en otros, el analista de datos tendrá que averiguar que todas estas variables corresponden a los mismos datos.

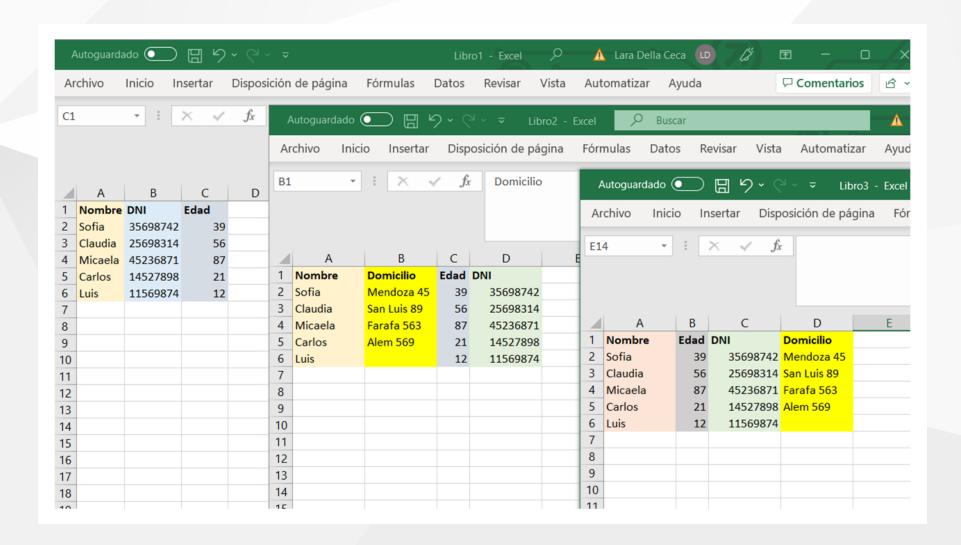
Celdas con datos faltantes

Valor nu lo	Problema	Compatibbilidad	Recomendación
0	Indistinguible de un valor real 0		No usar NUNCA
Blank	Es difícil distinguir los valores que faltan de los que se pasan por alto al ingresar. Es difícil distinguir los espacios en blanco de los espacios, que se comportan de manera diferente	R, Python, SQL	Mejor opción
-999, 999	No reconocido como nulo por muchos programas sin intervención del usuario. Se puede ingresar inadvertidamente en los cálculos		Evitar
NA, na	También puede ser una abreviatura (por ejemplo, North America), puede causar problemas con el tipo de datos (convertir una columna numérica en una columna de texto). NA es más comúnmente recognized than na.	R	Buena opción
N/A	Alternativa para NA, conviene evitar caracteres especiales		Evitar
NULL	Puede traer problemas con el tipo de datos		Evitar
None	Poco utilizado. Puede traer problemas con el tipo de datos	Python	Evitar
No data	Poco utilizado. Puede traer problemas con el tipo de datos, contiene un espacio		Evitar
Missing	Poco utilizado. Puede traer problemas con el tipo de datos		Evitar
-,+,.	Poco utilizado. Puede traer problemas con el tipo de datos		Evitar

Diseño de los datos en múltiples archivos

Si los datos están en distintos archivos y no tienen el mismo diseño (nombre de variables, posición de las columnas, etc.), combinar la información para poder realizar algún análisis tendrá mucho trabajo extra... y será más difícil automatizar el proceso.

Diseño de los datos en múltiples archivos



El nombre de los archivos

Establezca algún sistema para nombrar archivos. Si un archivo se llama "Incendios_por_provincia_2015.csv", entonces no llame al archivo para el siguiente año "Incendios_2016_por_provincia.csv" sino "Incendios_por_provincia_2016.csv".

- Alumnos_Tecnicatura_2021.xlsx
- Alumnos_TUIA_2022.xlsx
- Alumnos_2023_Tecnicatura.xlsx

El formato de las variables que contengan fechas

Preferiblemente se utiliza con el formato estándar YYYY-MM-DD (basado en la <u>ISO 8601</u>), por ejemplo 2023-01-20.

Si a veces escribe 20/1/2023 y a veces 20-1-23, será más difícil usar las fechas en los análisis o visualizaciones de datos.

Las notas en la columna de *Observaciones* (si existe)

NUNCA agregar comentarios en las celdas

Mejor agregar una columna con notas/observaciones. Si un comentario es 'por debajo del nivel de detección', siempre use el mismo (no lo modifique por, por ejemplo, 'por debajo del niv de det.').

Las notas en la columna de Observaciones

1	Α	В	С	D	Е	F		G	Н
1	persona_id	anio	sexo_id	sexo_descripcion	edad	maximo_grado_academico_id	disciplin	a_maximo_grado_academico_id	OBSERVACIONES
2	1	2020	2	MASCULINO	45	-1		-1	
3	5	2020	1	FEMENINO	57	1		255	
4	7	2020	2	MASCULINO	39	1		158	
5	10	2020	2	MASCULINO	60	5		248	
6	12	2020	2	MASCULINO	57	1		58	
7	13	2020	2	MASCULINO	50	3		281	
8	15	2020	1	FEMENINO	55	1		258	
9	17	2020	2	MASCULINO		Lara Della Ceca D9 Ø		281	Registro no seguro
10	19	2020	2	MASCULINO	(ID)	Edia Della Coca		255	
11	21	2020	1	FEMENINO		No confirmado		201	
12	22	2020	2	MASCULINO		09 de marzo de 2023, 09:21		65	
13	23	2020	2	MASCULINO		03 de maizo de 2023, 03.21		256	
14	24	2020	2	MASCULINO	Re	spuesta		34	
15	25	2020	2	MASCULINO				237	
16	26	2020	1	FEMENINO	69	1		251	
17	27	2020	1	FEMENINO	60	2		223	
18	28	2020	2	MASCULINO	58	1		11	

Cuidado con los espacios en blanco extra en las celdas

Una celda en blanco no es lo mismo que una celda que contiene un espacio, "femenino" es distinto de "femenino" (con un espacio antes y después).

Elegir buenos nombres (de archivos y variables)

Los nombres deben ser **cortos**, **pero significativos** (o sea que no tan corto...).

- No usar espacios, ni en nombres de variables ni en nombres de archivos. Hacen más difícil su lectura con los lenguajes de programación. Pueden usarse guiones en su remplazo, por ejemplo: glucosa_3_horas o glucosa-3-horas. No mezclar con _ !!! (p. ej: glucosa**_3-**horas)
- Evitar los caracteres especiales, excepto los guiones. Algunos otros símbolos (por ejemplo: \$,@,%,#,&,*,(,),!,/) suelen tener un

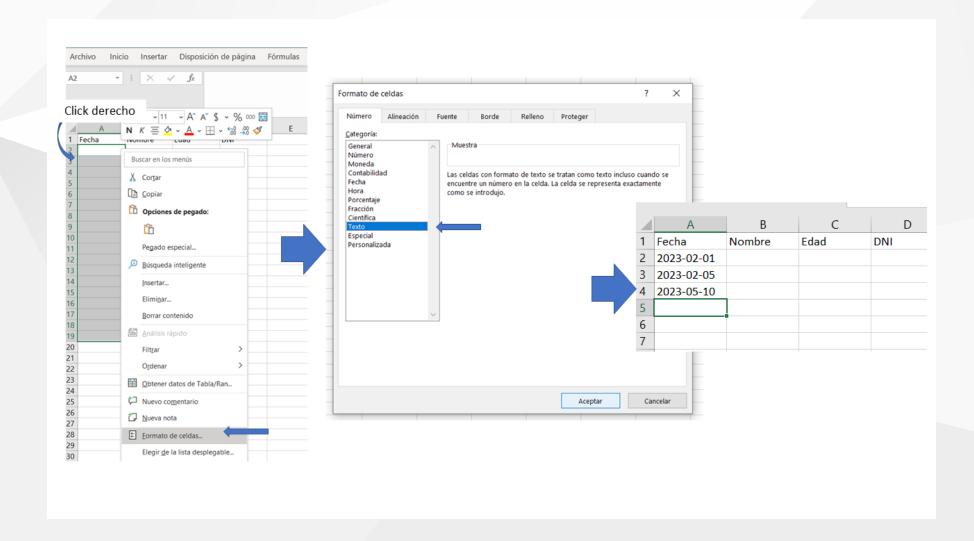
• Cuidado con los espacios extra...

Buen nombre	Buena alternativa	Evitar		
Max_temp_C	MaxTemp	Maximum Temp (°C)		
Precipitation_mm	Precipitation	precmm		
Mean_year_growth	MeanYearGrowth	Mean growth/year		
sex	sex	M/F		
weight	weight	W.		
cell_type	CellType	Cell Type		
Observation_01	first_observation	1st Obs		

Escribir fechas en el formato YYYY-MM-DD (formato estándar global <u>ISO 8601</u>)

Excel guarda las fechas internamente como un número, con distintas convenciones en la versión de Windows y Mac. Por lo tanto, es conveniente chequear la integridad de los datos de las fechas cuando provienen de Excel.

Escribir fechas en el formato YYYY-MM-DD



No dejar celdas vacías

Rellenar todas las celdas. Usar un valor común para los datos faltantes. Permite reconocer y dejar claro cuando el dato no está porque no se registró. Si no, no sabría si quien completó la tabla se olvidó de poner el dato o realmente no había dato que poner en dicha celda.

Colocar un único dato por celda

Por ejemplo, si precisamos una columna que contenga un período de tiempo, conviene dividir esta variable en dos, una columna para *Periodo_inicio* y otra para *Periodo_fin*.

O si deseamos incluir las unidades, por ejemplo '3505 g', podría colocarse la unidad en el nombre de la columna, p. ej. Peso_nacimiento_g.

Tampoco deben unirse celdas... no cumple con la regla de no dejar celdas vacías!

Diseño rectangular

El mejor diseño para contener datos dentro de una hoja de cálculo es un solo rectángulo cuyas filas contengan casos y las columnas variables. La primera fila debe contener nombres de variables, no utilizar más de una fila para los nombres de variables.

En general, en un mismo archivo de Excel encontramos tablas de datos en distintas hojas de cálculo. Es preferible tener múltiples archivos con una sola hoja de cálculo para poder guardar más fácilmente los datos como archivos csv.

	Α	В	С	D	Е	F	G	н	- 1
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	А	147	139	166	179	334	354	451	474
4	В	246	240	178	172	514	611	412	447

Celdas vacías Mas de una fila en el encabezado

	Α	В	С	D	E
1	strain	genotype	min	replicate	response
2	Α	normal	1	1	147
3	Α	normal	1	2	139
4	В	normal	1	1	246
5	В	normal	1	2	240
6	Α	mutant	1	1	166
7	Α	mutant	1	2	179
8	В	mutant	1	1	178
9	В	mutant	1	2	172
10	Α	normal	5	1	334
11	Α	normal	5	2	354
12	В	normal	5	1	514
13	В	normal	5	2	611
14	Α	mutant	5	1	451
15	Α	mutant	5	2	474
16	В	mutant	5	1	412
17	В	mutant	5	2	447

Diseño rectangular simple, facilita la lectura de la información con lenguajes de programación.

Crear un diccionario de los datos

Es útil tener un archivo separado que explique a qué corresponde cada variable, y en forma rectangular, para que el analista de datos pueda usarlo en los análisis. Dicho *diccionario de datos* podría contener:

- el nombre exacto de la variable como existe en el archivo de datos
- una versión del nombre de la variable que podría usarse en visualizaciones de datos
- una explicación más detallada de lo que significa la variable
- las unidades de medida
- los valores mínimos y máximos esperados

Diccionario de los datos

Esto es parte lo que suele denominarse metadatos: información sobre los datos.

4	Α	В	С
1	variable	nombre_plot	descripcion
2	raton	Raton	identificador del animal
3	sexo	Sexo	Macho (M) o Hembra (H)
4	fecha_sac	Fecha de sacrificio	Fecha en que el raton fue sacrificado
5	color	Color	Color del pelaje del raton
6	dias	Dias de tratamiento	Cantidad de dias en tratamiento
7			

No hacer cálculos en los archivos

A menudo, los archivos de Excel incluyen todo tipo de cálculos y gráficos. El archivo de datos principal debería contener solo los datos y nada más: sin cálculos, sin gráficos.

El archivo de datos primario debería almacenar los datos prístinos. Es conveniente protegerlo contra escritura y hacer una copia de seguridad.

Si se desea hacer algunos análisis en Excel, se puede crear una copia del archivo y hacer los cálculos y gráficos en la copia.

NO usar colores de fuente o resaltado como información

El formato de las celdas es agradable visualmente, pero es difícil extraer esa información para usarla en un análisis posterior. Los programas de análisis pueden manejar mucho más fácilmente los datos almacenados en una columna que los datos codificados como un formato diferente (resaltado de celda, fuente, negrita, etc.). Esa información, lo más probable es que se pierda completamente.

Es preferible incluir una columna de Observaciones.

Utilizar la validación de datos para evitar errores

La validación de datos es una función de Excel que permite controlar qué tipo de datos son ingresados en una celda o un grupo de celdas. Esta función se utiliza para evitar el registro de datos erróneos en las hojas de cálculo y permite establecer parámetros (rangos de fechas, valores límite, etc.) para que una misma columna o grupo de celdas conserven las mismas características o se encuentren dentro de un mismo intervalo.

<u>https://docs.google.com/presentation/d/1Xg5bBjX8sNoQtC6ATa6h9</u>
<u>o Ptmx7OPmbGN6kkuFheC0/edit#slide=id.g216f35db23d 0 0</u>

Funciones útiles de búsqueda en planillas de cálculo

Una función es una fórmula predefinida que realiza cálculos utilizando valores específicos en un orden particular. En <u>este link</u> pueden encontrar la lista de funciones disponibles a utilizar en Google Sheets.

A continuación se describen algunas funciones que consideramos útiles conocer.

VLOOKUP

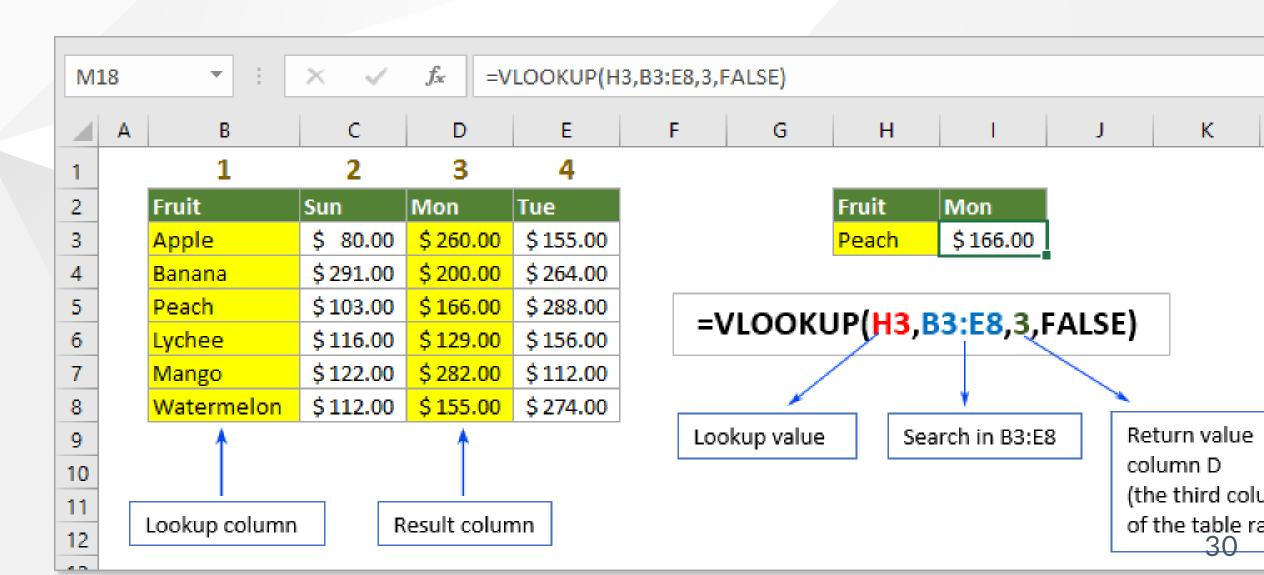
La función **VLOOKUP** (**BUSCARV**) o búsqueda vertical, es útil para buscar un valor específico en una columna de una tabla y recuperar un valor relacionado en otra columna.

Esta función recibe los datos de referencia como parámetros, incluyendo el valor que se busca en la primera columna (valor de búsqueda) y el rango de celdas que contiene la tabla de referencia (rango). También se especifica el número de columna donde se encuentra el valor deseado (índice) en la fila encontrada.

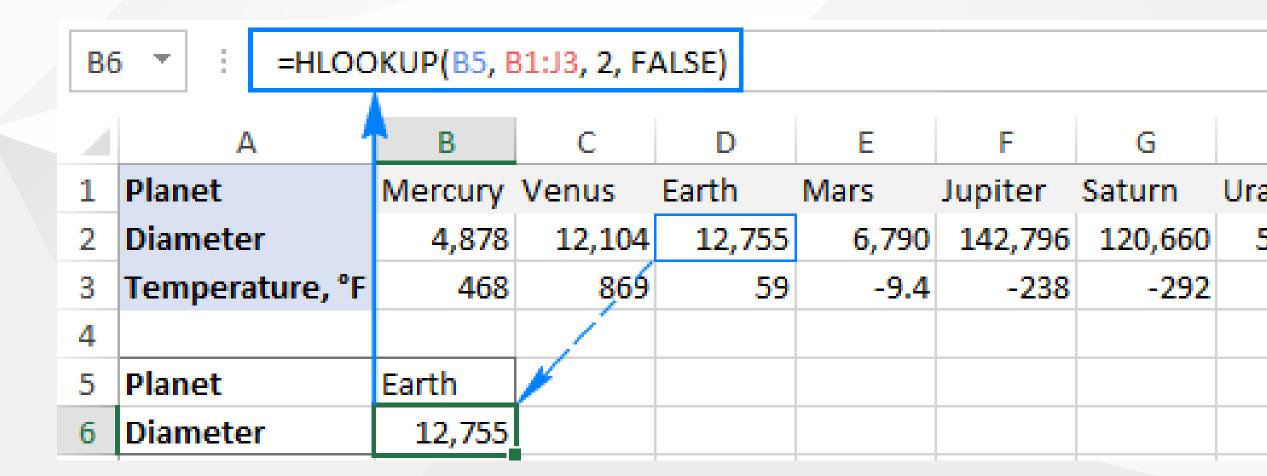
La forma general:

VLOOKUP(valor_búsqueda, rango, Índice, [está_ordenada])

VLOOKUP

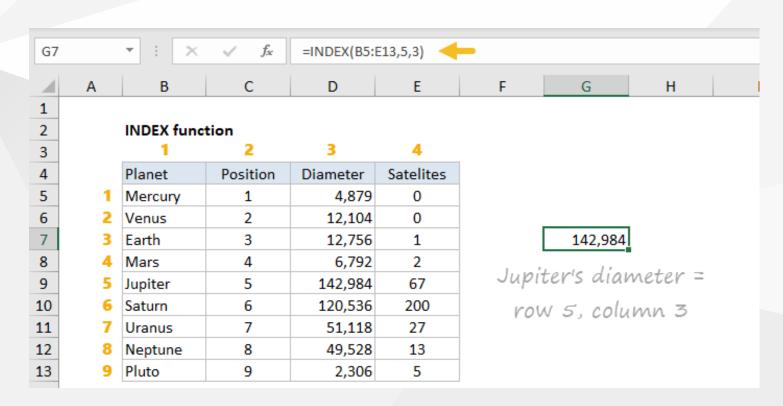


HLOOKUP



INDEX

INDEX (INDICE) devuelve el valor de una celda en un rango



https://docs.google.com/presentation/d/1iooid_p3ezDwxkr_VOSBLJKvymm2SziItrlXSr4OZw/edit#slide=id.p

MATCH

La función MATCH (COINCIDIR) devuelve la posición de un valor dado dentro de un rango

Forma general:

MATCH(valor_búsqueda, rango, [tipo_coincidencia])

https://support.microsoft.com/en-us/office/match-function-e8dffd45-c762-47d6-bf89-533f4a37673a

MATCH

MATCH (COINCIDIR) devuelve la posición de un valor dado en un rango

CC	COUNTIF \checkmark : \times \checkmark f_x =MATCH(H2,B3:B9,0)										
\square	Α	В	С	D	E	F	G	Н	1	J	
1											
2		ID	First Name	Last Name	Salary		ID	53			
3		72	Emily	Smith	\$64,901		Salary	5			
4		66	James	Anderson	\$70,855						
5		14	Mia	Clark	\$188,657						
6		30	John	Lewis	\$97,566						
7		53	Jessica	Walker	\$58,339						
8		56	Mark	Reed	\$125,180						
9		79	Richard	Lopez	\$91,632						
10											

https://docs.google.com/presentation/d/1tJrmzDhLD4V388v-4G5fBFZGOAz9vwqkRiurHYQL0ac/edit#slide=id.g215a933b4d9 0

INDEX + MATCH

Ejercicio comparando VLOOKUP vs INDEX+MATCH

https://docs.google.com/presentation/d/1FHrV3WKldcXO5-jVQodrLiWCSQUMf-BCXycB9llh7rs

XLOOKUP

Esta función es muy nueva en google sheet, se empezó a implementar en el año 2022. Y aumenta la complejidad y flexibilidad de VLOOKUP.

La Función **XLOOKUP** le permite buscar en una columna o fila específica y devolver un valor de la misma fila o columna. A diferencia de la función VLOOKUP anterior se puede especificar la columna y no tomar todo el rango. Esto permite buscar a la izquierda cosa que no era posible en VLOOKUP.

XLOOKUP

Además tiene algunas otras ventajas, por ejemplo si cambiamos el rango de columnas no cambia el resultado y limita la posibilidad de error.

```
XLOOKUP(
valor_búsqueda, rango_búsqueda, ^
rango_resultado, [falta_valor],
[modo_de_coincidencia], [modo_de_búsqueda]
)
```

XLOOKUP

НЗ	H3 ▼ : × ✓ <i>f</i> = XLOOKUP(H2,B3:B9,E3:E9)										
4	Α	В	С	D	Е	F	G	Н	1	J	
1											
2			First Name	Last Name	Salary		ID	53			
3		72	Emily	Smith	\$64,901		Salary	\$58,339			
4		66	James	Anderson	\$70,855						
5		14	Mia	Clark	\$188,657						
6		3 <mark>0</mark>	John	Lewis	\$97,566						
7		53	Jessica	Walker	→ \$58,339						
8		56	Mark	Reed	\$125,180						
9		79	Richard	Lopez	\$91,632						
10											

Manejo de errores

Otra característica interesante es la posibilidad de inidicar un valor de retorno cuando la función encuentra error. Esto mismo puede realizarse con la función IFERROR anidando a VLOOKUP:

=IFERROR(VLOOKUP(...,...), "No se encuentra el valor buscado")

O sea, utilizando una función condicional que mediante un input error (en este caso resultado de VLOOKUP) permite indicar un mensaje o valor de reemplazo como retorno alternativo al error. Una buena práctica para utilizar con estas funciones.

Busquedas parciales

Hablamos del parámetro donde introducimos **0** para especificar que lo que queremos es una búsqueda exacta. También en XLOOKUP podemos aplicar el parametro **2** (comodín) para traer todo lo que puede contener el valor a buscar pasado en el primer parámetro **que debe ser pasado entre asteriscos.**

Tablas dinámicas

Una tabla dinámica (o pivot table en inglés) en Excel es una herramienta avanzada para calcular, resumir y analizar grandes cantidades de datos de manera interactiva y visual en una hoja de cálculo.

Es creada a partir de un rango de datos en la hoja de cálculo que contiene varias columnas con información.

Las tablas dinámicas funcionan de forma un poco distinta dependiendo de la plataforma que use para ejecutar Excel (software de escritorio o Google Sheets).

Tablas dinámicas

Las tablas dinámicas o tablas pivote permiten:

- Consultar y realizar cálculos de grandes cantidades de datos de manera sencilla y rápida
- Visualizar los datos por categorías o subcategorías
- Expandir y contraer datos
- Filtrar, ordenar y agrupar datos
- Obtener distintos resúmenes de los datos: trasladar filas a columnas y viceversa