PLANILLA DE CÁLCULO

UNIDAD 1

Fundamentos de Ciencia de Datos 2025

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 de datos y elaboración de visualizaciones.

Excel 365, Google Sheets y LibreOffice Calc

- **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, **Google Sheets** 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 **Calc**.

Buenas prácticas en el uso de planillas de cálculo

A continuación, se describirán algunos principios considerados como ***buenas prácticas*** que serán de utilidad para proyectos futuros.

Los mismos permiten la creación de hojas de cálculo menos propensas a errores y más fáciles de procesar con distintos lenguajes de programación y de compartir con colaboradores.

1. Consistencia/coherencia en códigos de variables categóricas

Seleccionar una única opción y mantener siempre la misma.

Por ejemplo, en la variable ***estado civil*** podemos elegir entre:

Soltero	Casado	Divorciado	Viudo
soltero	casado	divorciado	viudo
s	c	d	v
S	C	D	V
Sol	Cas	Div	Viu

2. Sobre los nombres de las variables

Usar el mismo nombre para variables que aparezcan en múltiples archivos.

Ejemplo

Si en archivos en los que guardamos datos de experimentos de un laboratorio registramos los datos de *glucosa en sangre después de 3 horas* utilizando nombres variables como: **Glucosa_3horas**, **Gluc_3h** y **Glucosa 3 horas**, el analista de datos tendrá que averiguar que todas estas variables contienen el mismo tipo de datos.

3. Datos en múltiples archivos

Si los datos están en distintos archivos que no tienen el mismo diseño, combinar la información para poder realizar algún análisis tendrá mucho trabajo extra y será más difícil automatizar el proceso.

4. *Missing data*

Valor nulo	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

5. Sobre los nombres de los archivos

Establezca algún sistema para nombrar archivos.

Ejemplo

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”**.

6. Formato de las variables que contengan fechas

Preferiblemente se utiliza con el formato estándar **YYYY-MM-DD** (basado en la ISO 8601), por ejemplo: 2025-03-07.

Sea cual sea el formato que se utilice, es esencial mantener una coherencia a lo largo del documento. Si a veces se escribe 07/03/2025 y a veces 07-03-25, será más difícil usar las fechas en los análisis o visualizaciones de datos.

7. Las notas en la columna *Observaciones*

NUNCA agregar comentarios en las celdas.

Es mejor agregar una columna con notas/observaciones.

Si un comentario es *“por debajo del nivel de detección”*, siempre use esa misma estructura. No lo modifique por *“por debajo del niv de det.”*, por ejemplo.

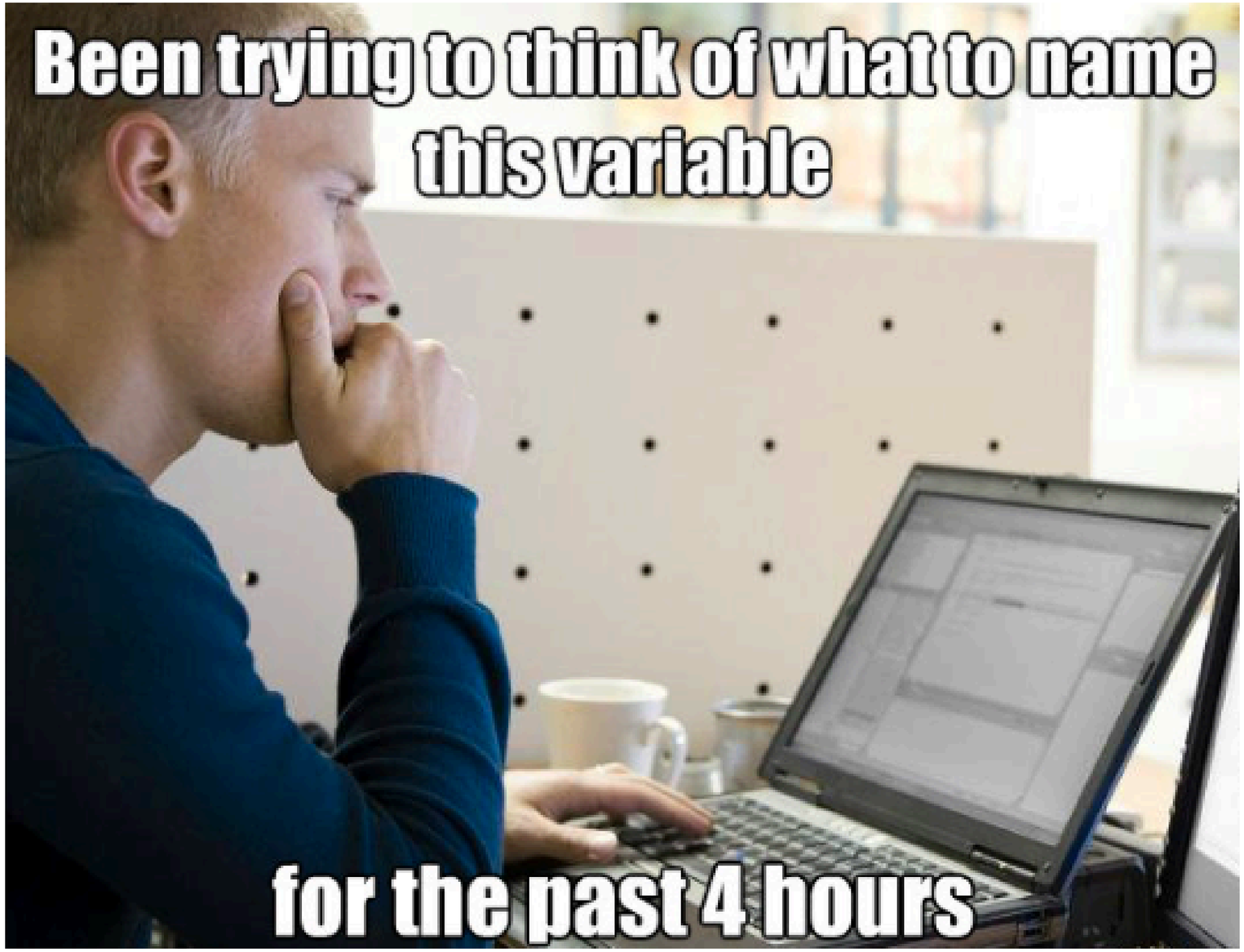
7. Las notas en la columna *Observaciones*

	A	B	C	D	E	F	G	H	I
1	**persona_id**	**anio**	**sexo_id**	**sexo_descripcion**	**edad**	**maximo_grado_academico_id**	**disciplina_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			281	Registro no seguro	
10	19	2020	2	MASCULINO			255		
11	21	2020	1	FEMENINO			201		
12	22	2020	2	MASCULINO			65		
13	23	2020	2	MASCULINO			256		
14	24	2020	2	MASCULINO			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		

8. ¡Cuidado con los espacios en blanco en las celdas!

- Una celda en blanco o vacía no es equivalente a una celda que contiene un espacio.
- “Femenino” es distinto de ” Femenino ” (con un espacio antes y otro después).

9. Elegir buenos nombres



9. Elegir buenos nombres

Aplica tanto a archivos como a variables. Los nombres deben ser cortos, pero significativos (o sea que no tan cortos...).

- **No usar espacios**, ni en nombres de variables ni en nombres de archivos. Hacen más difícil su lectura por parte de los lenguajes de programación. En su lugar, usar guiones bajos o medios (elegir un tipo y mantener siempre el mismo). Ejemplo: *glucosa_3_horas* o *glucosa-3-horas*.
- **Evitar los caracteres especiales**, excepto los guiones. Algunos otros símbolos (por ejemplo: $,@,%,#,&,*,(,),!,/) suelen tener un significado especial en los lenguajes de programación.

9. Elegir buenos nombres

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

10. Colocar un único dato por celda

Ejemplos:

- Si precisamos una columna que contenga un periodo de tiempo, conviene dividir esta variable en dos: una columna para **Periodo_inicio** y otra para **Periodo_fin**.
- Si deseamos incluir las unidades en las que está medida una variable, por ejemplo ‘3505 g’, podría colocarse la unidad en el nombre de la columna (por ejemplo, **Peso_nacimiento_g**).

11. Utilizar un diseño rectangular o formato largo

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.

11. Utilizar un diseño rectangular o formato largo

12. Hacer un diccionario de los datos

Es útil tener un archivo separado con una tabla en forma rectangular que explique **a qué corresponde cada variable**, para que el analista de datos pueda usarlo en el análisis posterior.

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

12. Hacer un diccionario de los datos

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

	A	B	C
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			

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

14. 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.). Lo más probable es que esa información se pierda completamente.

Es preferible incluir una columna de *Observaciones*.

15. Utilizar la validación de datos para evitar errores

La **validación de datos** es una función de Excel/Sheets 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.

SLIDES

Funciones de búsqueda en planillas de cálculo

Las **funciones** son fórmulas predefinidas que efectúan cálculos utilizando valores específicos llamados **argumentos**, en un orden o estructura particular.

En este link puede encontrarse la lista de funciones disponibles a utilizar en Google Sheets.

En particular veremos las siguientes funciones de búsqueda **VLOOKUP, HLOOKUP, INDEX y MATCH.**

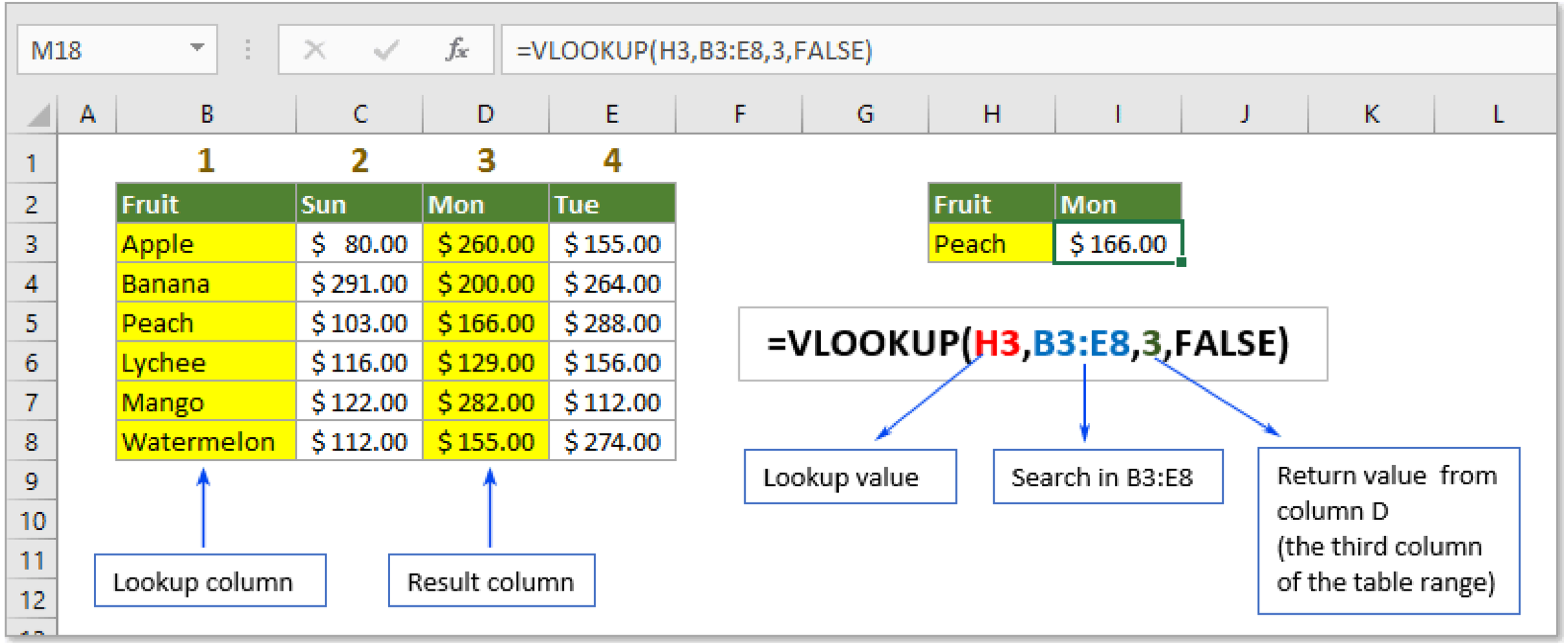
VLOOKUP (BUSCARV)

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 de la misma.

La forma general es la siguiente:

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

VLOOKUP (*BUSCARV*)



VLOOKUP (*BUSCARV*)

Sobre el argumento [está_ordenada].

Si es igual a **FALSE**, la función hace una **búsqueda exacta**. Sólo devuelve un resultado si encuentra exactamente el valor de búsqueda en la primera columna.

Si es **TRUE** (o si se deja vacío el argumento), la función asume que la columna de búsqueda está ordenada (de menor a mayor si son números o alfabéticamente en el caso de texto), y hará una **búsqueda aproximada.**

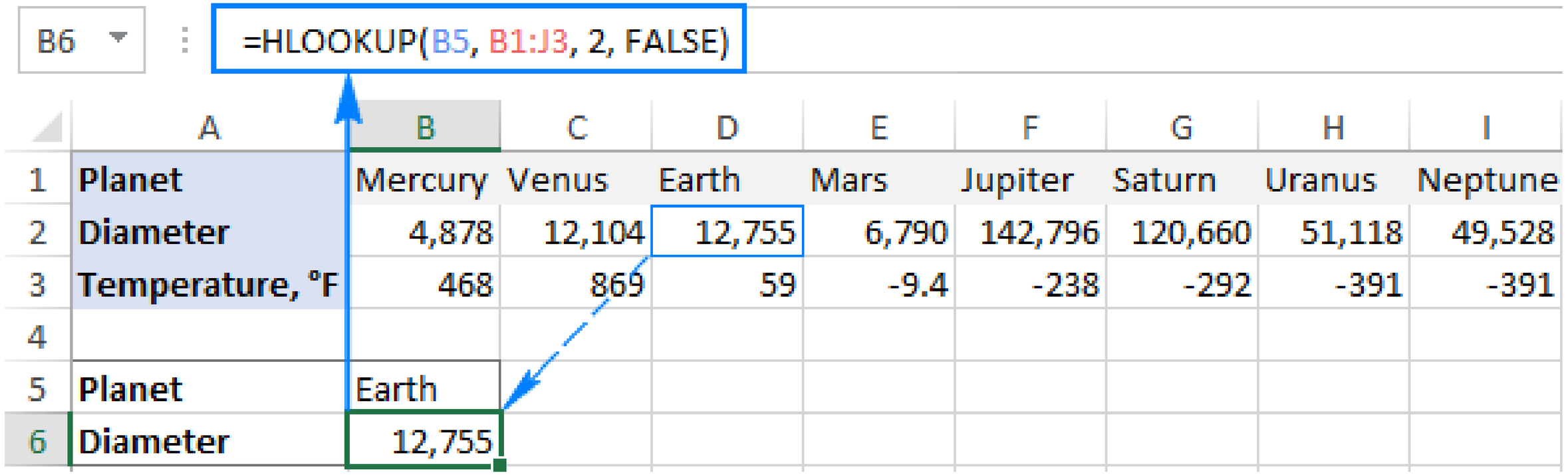
HLOOKUP (BUSCARH)

La función **HLOOKUP(BUSCARH)**, o búsqueda horizontal, busca un valor específico en la primera fila de una tabla y devuelve el valor de una celda específica en la columna encontrada.

La forma general es la siguiente:

HLOOKUP(valor_búsqueda, rango, índice, [está_ordenada])

HLOOKUP (*BUSCARH*)



	A	B	C	D	E	F	G	H	I
1	**Planet**	Mercury	Venus	Earth	Mars	Jupiter	Saturn	Uranus	Neptune
2	**Diameter**	4,878	12,104	12,755	6,790	142,796	120,660	51,118	49,528
3	**Temperature, °F**	468	869	59	-9.4	-238	-292	-391	-391
4									
5	**Planet**	Earth							
6	**Diameter**	12,755							

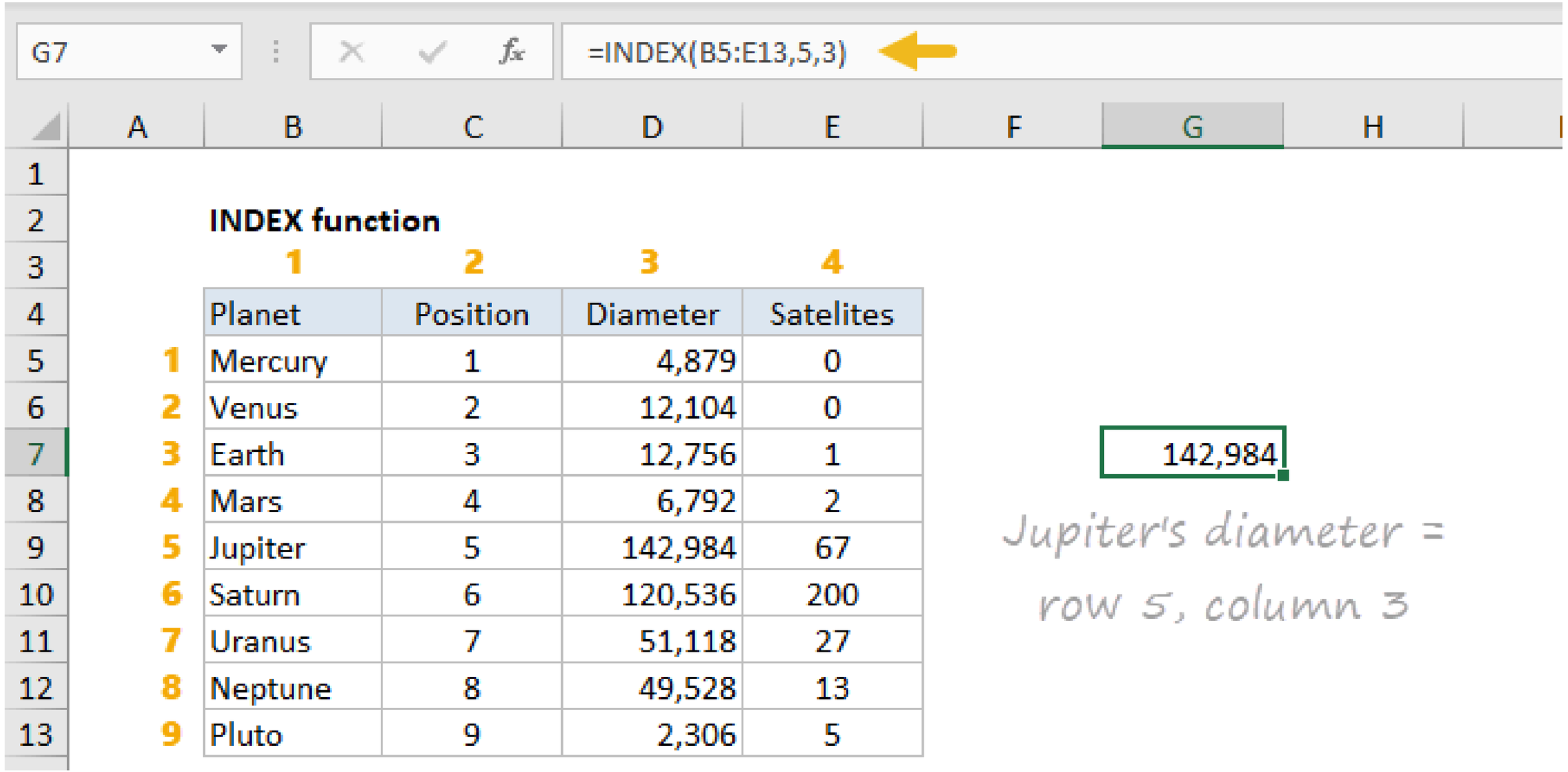
INDEX (INDICE)

Ofrece el **contenido** de una celda, especificado por los índices de número de fila y de columna.

La forma general es la siguiente:
INDEX(referencia; fila; columna)

EJEMPLO

INDEX (*INDICE*)



INDEX function

	1	2	3	4
	Planet	Position	Diameter	Satelites
1	Mercury	1	4,879	0
2	Venus	2	12,104	0
3	Earth	3	12,756	1
4	Mars	4	6,792	2
5	Jupiter	5	142,984	67
6	Saturn	6	120,536	200
7	Uranus	7	51,118	27
8	Neptune	8	49,528	13
9	Pluto	9	2,306	5

142,984

Jupiter's diameter = row 5, column 3

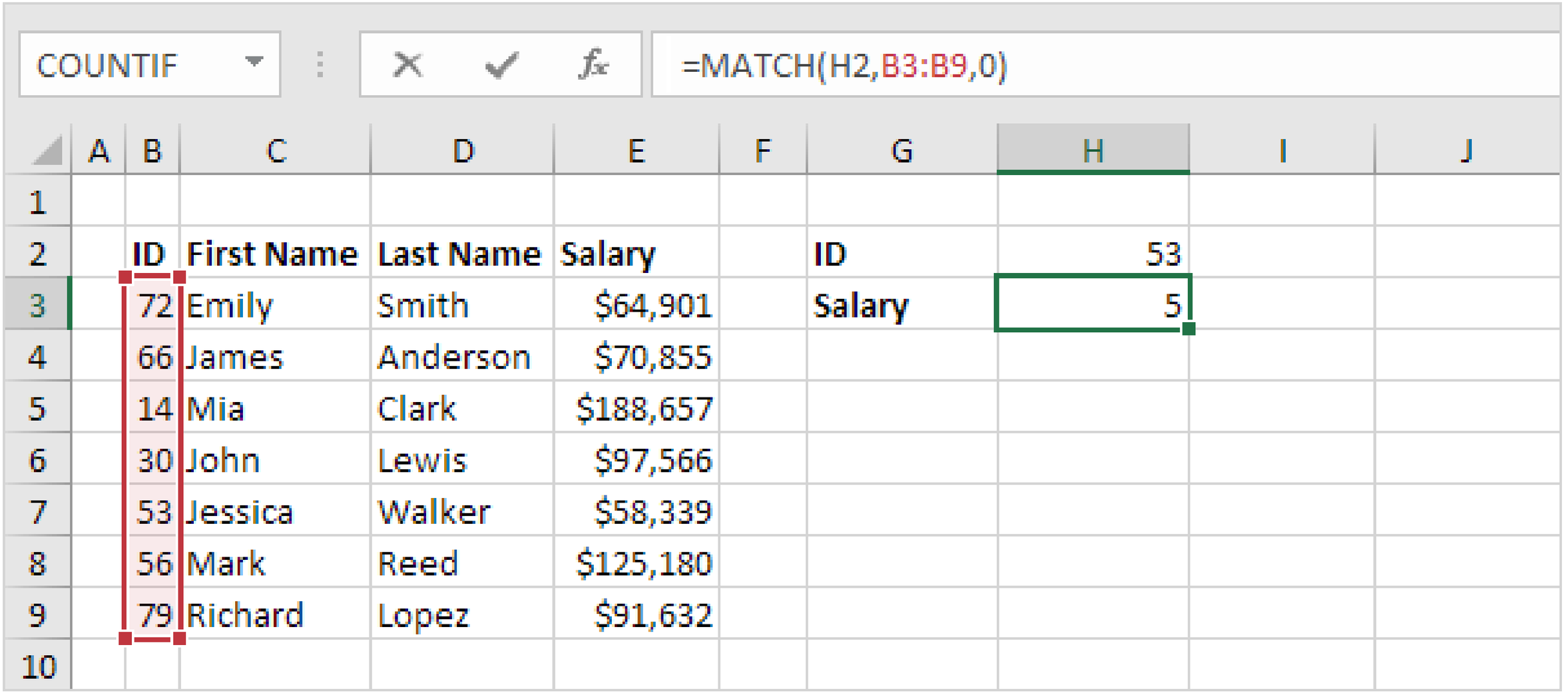
MATCH (COINCIDIR)

Busca un elemento determinado en un intervalo de celdas y devuelve la **posición relativa** de dicho elemento en el rango.

La forma general es la siguiente:
MATCH(valor_búsqueda; intervalo; [tipo_búsqueda])

EJEMPLO

MATCH (*COINCIDIR*)



	A	B	C	D	E	F	G	H	I	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										

INDEX + MATCH

Es posible combinar las funcionalidades de búsqueda de **INDEX** y **MATCH** para lograr una alternativa superadora a **VLOOKUP**.

EJEMPLO

INDEX + MATCH vs. VLOOKUP

Búsqueda de derecha a izquierda.

VLOOKUP no puede mirar a su izquierda, lo que significa que su valor de búsqueda siempre debe residir en la columna más a la izquierda de la tabla. **INDEX + MATCH** puede realizar búsquedas a la izquierda con facilidad.

INDEX + MATCH vs. VLOOKUP

Pueden insertarse o eliminarse columnas de forma segura.

Las fórmulas de **VLOOKUP** se rompen o entregan resultados incorrectos cuando se elimina o agrega una nueva columna a una tabla de búsqueda porque la sintaxis de **VLOOKUP** requiere especificar el número de índice de la columna de la que desea extraer los datos. Esto no ocurre con **INDEX + MATCH**.

INDEX + MATCH vs. VLOOKUP

No hay límite para el tamaño del valor de búsqueda.

Al usar la función **VLOOKUP**, la longitud total de sus criterios de búsqueda no puede exceder los 255 caracteres.

INDEX + MATCH vs. VLOOKUP

Mayor velocidad de procesamiento.

Si las tablas son relativamente pequeños, apenas habrá una diferencia significativa en el rendimiento de Excel/Sheets. Pero si sus hojas de trabajo contienen cientos o miles de filas y, en consecuencia, cientos o miles de fórmulas, **INDEX+MATCH** funcionará mucho más rápido que **VLOOKUP** porque Excel/Sheets tendrá que procesar sólo las columnas de búsqueda y devolución en lugar de toda la matriz de la tabla.

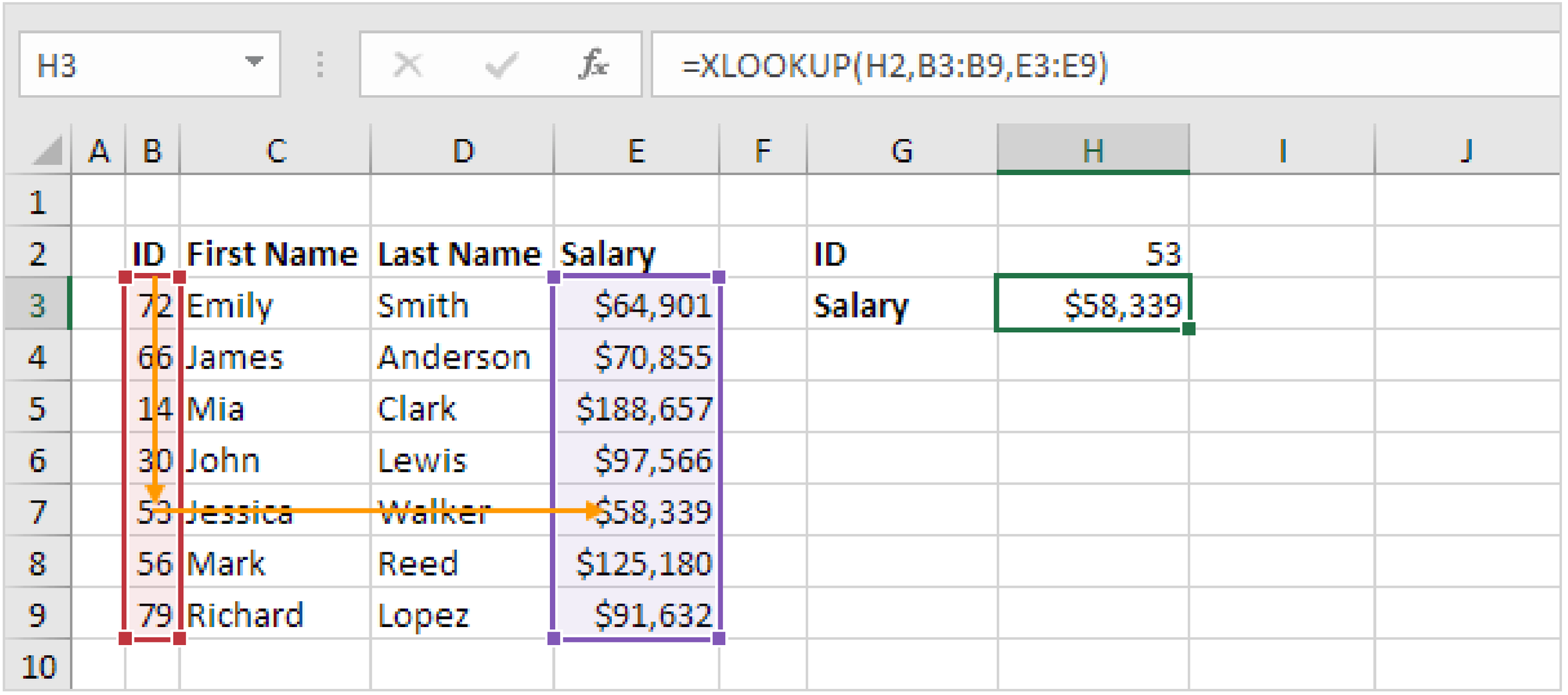
XLOOKUP

Esta función es muy nueva en Google Sheets (se empezó a implementar en el año 2022) y aumenta la complejidad y flexibilidad de VLOOKUP.

La función **XLOOKUP** permite buscar un elemento en una columna o fila específica y devolver un valor de la misma fila o columna.

A diferencia de VLOOKUP se puede especificar la columna y no tomar toda la matriz de datos, lo que permite realizar búsquedas hacia la izquierda.

XLOOKUP



	A	B	C	D	E	F	G	H	I	J
1										
2		ID	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		30	John	Lewis	$97,566					
7		53	Jessica	Walker	$58,339					
8		56	Mark	Reed	$125,180					
9		79	Richard	Lopez	$91,632					
10										