VLOOKUP

What is a VLOOKUP?

- VLOOKUP (short for “vertical lookup”) is a function that is used to extract a particular value from a spreadsheet, given a unique identifier
- Let’s say we wanted to know the weight of a particular breed of dog
- To find this out, we would use VLOOKUP!

	A	B
1	breed	weight
2	chihuahua	5.5
3	corgi	31
4	labrador retr	67.5
5	pitbull	62
6	pug	16
7	standard sch	40

Source

`VLOOKUP` Syntax

- The VLOOKUP function take three arguments:
- `=VLOOKUP(lookup_value, table_array, col_index_num)`

`VLOOKUP` Syntax

- The VLOOKUP function take three arguments:
- `=VLOOKUP(lookup_value, table_array, col_index_num)`
- **`lookup_value`:** the index of the cell that contains what we want to search for, in this case a breed of dog

VLOOKUP Syntax

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	corgi	31		Breed:	
4	labrador retr	67.5		Weight:	
5	pitbull	62			
6	pug	16			
7	standard sch	40			

`VLOOKUP` Syntax

- The VLOOKUP function take three arguments:
- `=VLOOKUP(lookup_value, table_array, col_index_num)`
- **`lookup_value`:** the index of the cell that contains what we want to search for, in this case a breed of dog
  - Let's set the `lookup_value` to be `E3`, and ask the user to enter a breed of dog there

`VLOOKUP` Syntax

- The VLOOKUP function take three arguments:
- `=VLOOKUP(lookup_value, table_array, col_index_num)`
- **`lookup_value`:** the index of the cell that contains what we want to search for, in this case a breed of dog
  - Let's set the `lookup_value` to be `E3`, and ask the user to enter a breed of dog there
- **`table_array`:** a range `start:end` where start is the upper-left cell of the data table and end is the bottom-right cell

VLOOKUP Syntax

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	corgi	31		Breed:	
4	labrador retr	67.5		Weight:	
5	pitbull	62			
6	pug	16			
7	standard sch	40			

`VLOOKUP` Syntax

- The VLOOKUP function take three arguments:
- `=VLOOKUP(lookup_value, table_array, col_index_num)`
- **`lookup_value`:** the index of the cell that contains what we want to search for, in this case a breed of dog
  - Let's set the `lookup_value` to be `E3`, and ask the user to enter a breed of dog there
- **`table_array`:** a range `start:end` where start is the upper-left cell of the data table and end is the bottom-right cell
  - The data table spans A2 to B7, so `table_array` is `A2:B7`

`VLOOKUP` Syntax

- The VLOOKUP function take three arguments:
- `=VLOOKUP(lookup_value, table_array, col_index_num)`
- **`lookup_value`:** the index of the cell that contains what we want to search for, in this case a breed of dog
  - Let's set the `lookup_value` to be `E3`, and ask the user to enter a breed of dog there
- **`table_array`:** a range `start:end` where start is the upper-left cell of the data table and end is the bottom-right cell
  - The data table spans A2 to B7, so `table_array` is `A2:B7`
- **`col_index_num`:** the column number where the lookup value resides

VLOOKUP Syntax

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	corgi	31		Breed:	
4	labrador retr	67.5		Weight:	
5	pitbull	62			
6	pug	16			
7	standard sch	40			

`VLOOKUP` Syntax

- The VLOOKUP function take three arguments:
- `=VLOOKUP(lookup_value, table_array, col_index_num)`
- **`lookup_value`:** the index of the cell that contains what we want to search for, in this case a breed of dog
  - Let's set the `lookup_value` to be `E3`, and ask the user to enter a breed of dog there
- **`table_array`:** a range `start:end` where start is the upper-left cell of the data table and end is the bottom-right cell
  - The data table spans A2 to B7, so `table_array` is `A2:B7`
- **`col_index_num`:** the column number where the lookup value resides
  - The weight of the dog resides in the second column of the data table, so `col_index_num` is 2

VLOOKUP Syntax

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	corgi	31		Breed:	corgi
4	labrador retr	67.5		Weight:	31
5	pitbull	62			
6	pug	16			
7	standard sch	40			

=VLOOKUP(E3,A2:B7,2)

`VLOOKUP` defaults to Approximate Match

- `VLOOKUP` looks for an approximate match if it doesn't find an exact match
- It assumes ordered data, and takes as its approximate match the value that is one cell *above* (i.e., before) the input

`VLOOKUP` defaults to Approximate Match

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	corgi	31		Breed:	corgo
4	labrador retr	67.5		Weight:	31
5	pitbull	62			
6	pug	16			
7	standard sch	40			

`corgi` comes before `corgo` alphabetically, so its weight is returned

range_lookup

- VLOOKUP can take in an optional 4th parameter called range_lookup
- range_lookup can be either 1, for TRUE, or 0, for FALSE
- The default value (when nothing is specified) is TRUE
- FALSE means only return a value when an exact match is found

Invalid Input

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	corgi	31		Breed:	corgo
4	labrador retr	67.5		Weight:	#N/A
5	pitbull	62			
6	pug	16			
7	standard sch	40			

=VLOOKUP(E3,A2:B7,2,FALSE)

VLOOKUP with Approximate Matching

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	labrador retr	67.5		Breed:	labrador retriever
4	corgi	31		Weight:	31
5	pitbull	62			
6	pug	16			
7	standard sch	40			

Be careful! When using a range_lookup of TRUE, the column that you're searching from has to be sorted in **ascending** order, or you might get some weird results—even with an exact match!

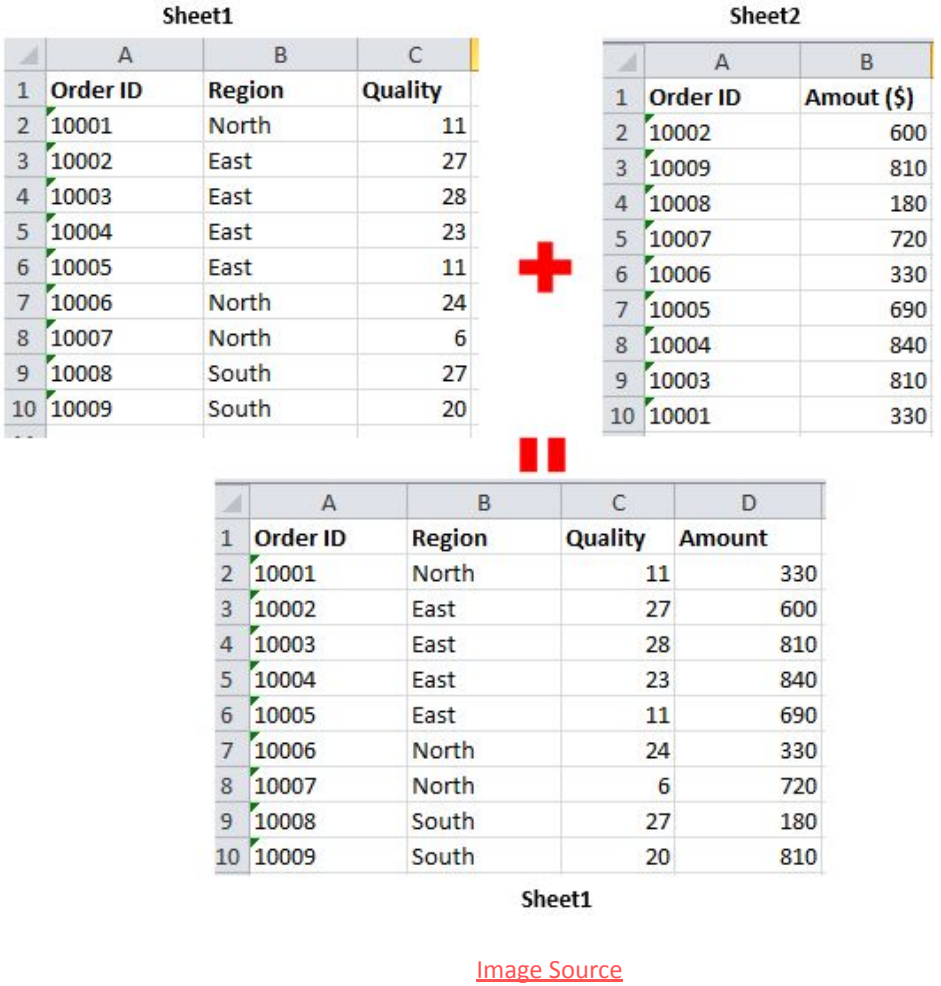
Join

The Join Operation

- Join is a fundamental database operation.
- As its name suggests, it combines two databases into one.
- In order to perform a join, both databases need to share a variable, and the values that the variable takes on must be unique to each observation.
- Join combines the databases by matching on these values.

An Example

In this example, the `Region` and `Quality` columns from `Sheet1`, and the `Amount` [sic] column from `Sheet2`, are joined by matching on the `Order ID` column.



	A	B	C
1	Order ID	Region	Quality
2	10001	North	11
3	10002	East	27
4	10003	East	28
5	10004	East	23
6	10005	East	11
7	10006	North	24
8	10007	North	6
9	10008	South	27
10	10009	South	20

	A	B
1	Order ID	Amout ($)
2	10002	600
3	10009	810
4	10008	180
5	10007	720
6	10006	330
7	10005	690
8	10004	840
9	10003	810
10	10001	330

	A	B	C	D
1	Order ID	Region	Quality	Amount
2	10001	North	11	330
3	10002	East	27	600
4	10003	East	28	810
5	10004	East	23	840
6	10005	East	11	690
7	10006	North	24	330
8	10007	North	6	720
9	10008	South	27	180
10	10009	South	20	810

Image Source

Back to our running example...

- Next, let's consider two tables, one that lists the average weight of male dogs, and a second that lists the average weight of female dogs.

	A	B	C	D	E	F
1	breed	male_weight			breed	female_weight
2	chihuahua	5.5			labrador retr	62.5
3	corgi	32.5			pug	16
4	labrador retr	72.5			standard sch	37.5
5	pitbull	62			chihuahua	5.5
6	pug	16			pitbull	62
7	standard sch	42.5			corgi	29.5

Joining tables in a spreadsheet

- Let’s say that we wanted to combine these two tables into one that lists the average male and female weights for each breed of dog.
- We can use `VLOOKUP` to join the data!

	A	B	C
1	breed	male_weight	female_weig
2	chihuahua	5.5	5.5
3	corgi	32.5	29.5
4	labrador retr	72.5	62.5
5	pitbull	62	62
6	pug	16	16
7	standard sch	42.5	37.5

Joining tables in a spreadsheet

- To start, we type `=VLOOKUP(A2, E2:F7, 2, FALSE)` in cell `C2`, because we want to find an exact match (`range_lookup = FALSE`) for `chihuahua` (`lookup_value = A2`) from the female weights table (`table_array = E2:F7`), and we want the value from its second column (`col_index_num = 2`)

Joining the first value

	A	B	C	D	E	F
1	breed	male_weight	JOIN		breed	female_weight
2	chihuahua	5.5	5.5		labrador retr	62.5
3	corgi	32.5			pug	16
4	labrador retr	72.5			standard sch	37.5
5	pitbull	62			chihuahua	5.5
6	pug	16			pitbull	62
7	standard sch	42.5			corgi	29.5

Joining tables in a spreadsheet

- Next, we type `=VLOOKUP(A3, E2:F7, 2, FALSE)` in cell `C3`, because we want to find an exact match (`range_lookup = FALSE`) for `corgi` (`lookup_value = A3`) from the female weights table (`table_array = E2:F7`), and we want the value from its second column (`col_index_num = 2`).
- Wait a minute! That sounds an awful lot like what we did before.
- The only difference is that `A2` became `A3`, and `C2`, `C3`.
- So is there a shortcut, then, for entering the rest of these formulas?
- Let’s try copying the formula we typed to the rest of the cells in the column by selecting the cell and dragging it downwards.

Joining the rest of the data

- What went wrong?

	A	B	C	D	E	F
1	breed	male_weight	JOIN		breed	female_weight
2	chihuahua	5.5	5.5		labrador retr	62.5
3	corgi	32.5	29.5		pug	16
4	labrador retr	72.5	#N/A		standard sch	37.5
5	pitbull	62	62		chihuahua	5.5
6	pug	16	#N/A		pitbull	62
7	standard sch	42.5	#N/A		corgi	29.5

Joining tables in a spreadsheet

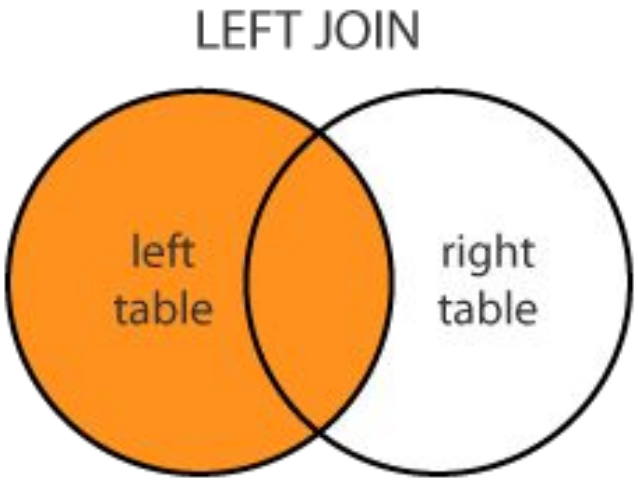
- Our original formula was `=VLOOKUP(A2, E2:F7, 2, FALSE)`.
- But if you take a look at cell `C3`, you can see that its formula is `=VLOOKUP(A3, E3:F8, 2, FALSE)`.
- Likewise, the formula for `C4` is `=VLOOKUP(A4, E4:F9, 2, FALSE)`.
- The problem is that the row numbers in `table_array` are changing, but the table location is constant.
- We need absolute references to refer to the rows in the `table_array` in our original formula: `=VLOOKUP(A2, E$2:F$7, 2, FALSE)`.
- Let's try copying this new formula to the rest of the cells in the column.

Voila!

	A	B	C	D	E	F
1	breed	male_weight			breed	female_weight
2	chihuahua	5.5	5.5		labrador retr	62.5
3	corgi	32.5	29.5		pug	16
4	labrador retr	72.5	62.5		standard sch	37.5
5	pitbull	62	62		chihuahua	5.5
6	pug	16	16		pitbull	62
7	standard sch	42.5	37.5		corgi	29.5

More About Joins

- VLOOKUP joins are **left outer joins** (aka “left joins”), meaning that it will try to fill in all values of the original table (the left table) with values from the table that you’re copying from (the right table), even if the values don’t exist



Source

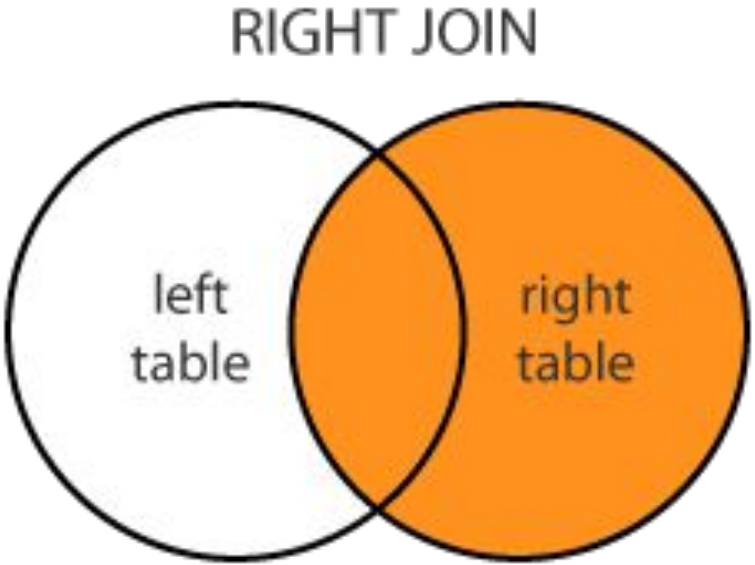
Left Outer Join with Missing Data

- In this example, `corgi` is not in the female weights table.

	A	B	C	D	E	F
1	breed	male_weight	JOIN		breed	female_weight
2	chihuahua	5.5	5.5		labrador retr	62.5
3	corgi	32.5	#N/A		pug	16
4	labrador retr	72.5	62.5		standard sch	37.5
5	pitbull	62	62		chihuahua	5.5
6	pug	16	16		pitbull	62
7	standard sch	42.5	37.5			

More About Joins

- **Right outer joins** (aka “right joins”) are the opposite of left outer joins, keeping all values from the second table, regardless of whether or not they are in the first table



Source

Right Outer Join with Missing Data

- In this example, `corgi` is not in the male weights table.

	A	B	C	D	E	F
1	breed	male_weight		breed	female_weig	JOIN
2	chihuahua	5.5		chihuahua	5.5	5.5
3	labrador retr	72.5		corgi	29.5	#N/A
4	pitbull	62		labrador retr	62.5	72.5
5	pug	16		pitbull	62	62
6	standard schı	42.5		pug	16	16
7				standard schı	37.5	42.5

VLOOKUP with Approximate Match

	A	B	C	D	E
1	breed	weight			
2	chihuahua	5.5			
3	corgi	31		Breed:	labrador retrieve
4	labrador retr	67.5		Weight:	31
5	pitbull	62			
6	pug	16			
7	standard sch	40			

Even though “labrador retrieve” is closer to “labrador retriever,” the VLOOKUP returns the value of “corgi” since it the closest identifier that comes before the input