

DATA REFERENCING IN EXCEL

DATA REFERENCING IN EXCEL

LEARNING OBJECTIVES

- Use data functions, VLOOKUP and HLOOKUP, to manipulate datasets.
- Use data functions, INDEX and MATCH, to lookup values in other tables.

DATA REFERENCING IN EXCEL

OPENING

OPENING

- VLOOKUP (and its siblings HLOOKUP, INDEX, and MATCH) are often considered advanced topics, and knowing how to use them effectively can make you stand out in the world of Excel users.
- Interview Tip: When interviewing for a new job, there is a big difference between someone who claims "experience in Excel" and someone who can name-drop lookup functions or INDEX/MATCH.

DATA REFERENCING IN EXCEL

INTRODUCTION: BUSINESS SCENARIO

INTRODUCTION: BUSINESS SCENARIO

- In the work of an analyst, it is rare when you have all the data you need right in your dataset. Oftentimes, you will be given data in different formats, different spreadsheets, etc. One big step of data cleansing is simply getting the data into one dataset.
- In the next unit, SQL, we will see how databases can be extremely helpful in this regard, but if you don't know how to use SQL, or if none of the data you are working with is in SQL, it can be easier to do this step in Excel.

DATA REFERENCING IN EXCEL

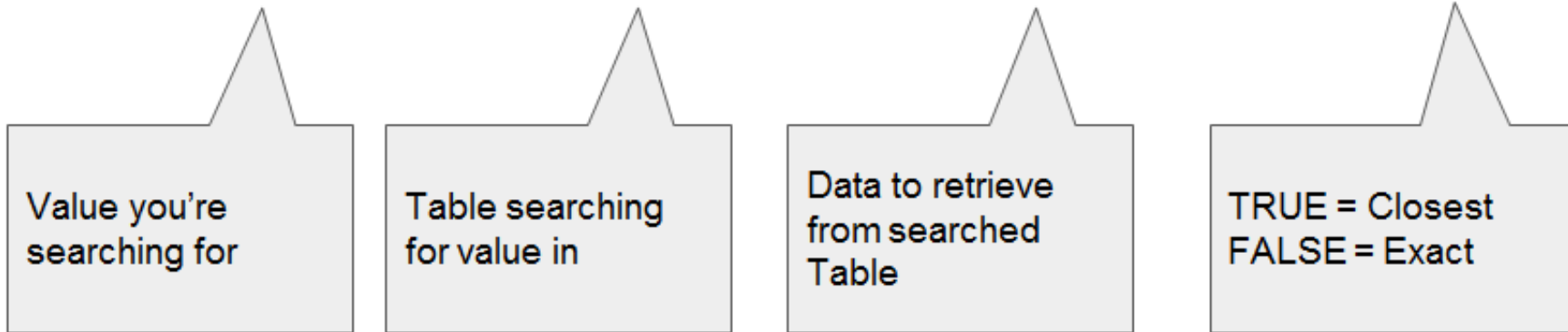
DEMO: A SIMPLE VLOOKUP EXAMPLE

DATA REFERENCING IN EXCEL



DATA REFERENCING IN EXCEL

VLOOKUP (lookup_value, table_array, col_index_num, [range_lookup])



Value you're
searching for

Table searching
for value in

Data to retrieve
from searched
Table

TRUE = Closest
FALSE = Exact

DATA REFERENCING IN EXCEL

**PRACTICE: USING
VLOOKUP TO COMBINE
DATA ACROSS
WORKSHEETS**

INTRODUCTION: BUSINESS SCENARIO

- We will see how to combine data from another worksheet into our ACS dataset.
- We will also learn the concept of a “lookup table” and how to use one to categorize our quantitative data.

SAMPLE DATASET

- Open demo1_accounts worksheet from the workbook L2_demo_worksheets.xlsx
 - This list contains account names and numbers from a small company.
- Open demo1_emails worksheets
 - This list contains email accounts with account numbers, but no name.
- In order to combine these two lists, meaning to make a fourth column in demo1_accounts with email addresses, we can use VLOOKUP.
 - VLOOKUP will “look up” one value in another table and return another column of that row.

VLOOKUP SYNTAX

- The syntax is:

=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

- **Lookup_value** is the value that will be used to match. This is usually an identifier (an ID of some kind). It must exist in both worksheets.
- **Table_array** is the table that you will want to retrieve data from.
- **Col_index_num** is the number of the column from the left side of the table_array that you want to retrieve data from.

VLOOKUP SYNTAX

- The syntax is:

=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

- **Range_lookup** defines whether or not lookup_value is an approximate match or an exact match of the value you are comparing it to in the left-most column of table_array.
 - **TRUE** is the default and means “approximate.” BUT you should ALWAYS put **FALSE** here, unless you know what you are doing/have a specific reason to look for “approximate” value.
 - We will use the **FALSE** value here when we look at lookup tables later today.

VLOOKUP SYNTAX

- In cell D2 of demo1_accounts, enter:

=VLOOKUP(C2, demo1_emails!A:B, 2, FALSE)

- The e-mail should populate cell D2.
- Double-click the bottom-right corner to expand this down the entire table.

This is how VLOOKUP works in general. Now we will try another example with our ACS dataset.

GUIDED PRACTICE

- Earlier, we looked at the scatterplot comparing two columns in our dataset:
 - Population estimate of our census tract, and
 - % of people who commute with public transit.
- There was no obvious relationship. This makes sense as population and population *density* are not the same thing.
- While we expect people in urban areas to use public transit more: the number of people in a census tract is not necessarily a function of density.

GUIDED PRACTICE

Let's look at the scatterplot between census tract density and public transit usage. Density is not currently data that we have in our dataset. Follow these steps to create this scatterplot:

- Open census_tract_areas_WA.xlsx
 - Here we have a few variables; the one of interest is area_sqkm
 - It would be error-prone and time-consuming to copy and paste this data into our spreadsheet.

PREPARE TO VLOOKUP

- Before we can combine these data sources, we need to ensure they have a column/identifier in common. Take a look at column A in our original dataset and column D in the new dataset.
 - Column A in the original is equivalent to column D in the new dataset, but each value has “14000000US” on the front.
 - In census_tract_areas_WA.xlsx, create a new column E and enter this formula into cell E2: =“14000000US”&D2. Expand it to all rows.
 - Now we have a common identifier in our two worksheets.

PERFORM THE VLOOKUP

- In our original worksheet, 2014_acs_select, go to column AJ.
- Name it “area_sqkm” by typing this in cell AJ1.
- In cell AJ2, enter:

```
=VLOOKUP(A2,census_tract_areas_WA!E:N,10,FALSE)
```
- Expand this formula to all rows by double-clicking the bottom-right corner of the cell.
- Now we have the tract area in our dataset.

CALCULATE THE DENSITY

- In cell AK1, type “density”.
- In cell AK2, enter formula =E2/AJ2 and expand.
- Create a scatterplot between % public transit vs. your new density variable.
- Do we see any interesting patterns now?
- Now let's complete one more VLOOKUP.
 - At the moment, our only identifier for each tract is the ID number.
 - What county are these tracts in?

USE VLOOKUP FOR COUNTY NAMES

- Open census_tract_county_names_WA.xlsx
- This file also includes the IDs, so we can use that as our common identifier to match the county names with the census tracts in our dataset.
- In our dataset, label a new column “County”.

USE VLOOKUP FOR COUNTY NAMES

- In the second row, enter our formula for VLOOKUP:

```
=VLOOKUP(A2,[census_tract_county_names_WA.xlsx]ACS_14_5YR_B01002_with_ann!$A:$B,2,FALSE)
```

- When selecting the table array, it is easy to use the cell selector (the mouse icon that is a white '+') to select the columns of our table_array. This is especially useful when crossing over to another file, as we are doing here.
- Copy the contents down by double-clicking the square in the bottom-right.
- We've now successfully used VLOOKUP with two examples! What else can VLOOKUP do?

DATA REFERENCING IN EXCEL

**PRACTICE: USING
VLOOKUP TO CREATE
CATEGORICAL
VARIABLES**

PERFORM THE VLOOKUP

- Often it is very helpful to create categorical values from numeric values. For example, a test that is scored 0-100 could be classified as A, B, C, D, or F, depending on the score.
- Let's classify our census tracts as either Low, Medium, or High.
- What are some ways to create three classification groups?

PERFORM THE VLOOKUP

- Using 33rd, 66th, and 99th percentile is one option. While there are various ways to calculate a percentile, it generally means which value has X% less than the value in question. For example, if 13 is the 50th percentile, half of the numbers in the range are less than 13. Let's use this to break our data up into three equal-sized groups.
- Create a new worksheet called Density Lookup.
- In cells A1, A2, A3, enter values 0, 350, 1420
 - The lookup values ALWAYS have to go on the left side of the table_array when using VLOOKUP.
 - The lookup will go from the value on the row *up to* the next entry (or infinity). In other words, they are *left-bound*.

PERFORM THE VLOOKUP

- In cells B1, B2, B3, enter values Low, Medium, High.
- Back on worksheet 2014_acs_select, in cell AL1 type “density group”.
- In AL2, complete the lookup:
`=VLOOKUP(AK2,'Density Lookup'!A:B,2,TRUE)`
- Expand to all rows. The densities have now been classified using our lookup table.
- If we ever want to change our definition of the three classifications, all we have to do is modify the lookup table on the Density Lookup table tab and the values will change accordingly.

DATA REFERENCING IN EXCEL

DEMO: VLOOKUP VS HLOOKUP

HLOOKUP

- HLOOKUP is very closely related to VLOOKUP, but instead of providing the *column number*, you provide the *row number*.
- Looking at worksheet demo2_hlookup in L2_demo_worksheets.xlsx, consider the top table of grades for different students.
- In this case, the students are columns instead of rows.
- Suppose we needed to fill out the second table using only formulas. We would use an HLOOKUP, since we need to do the *lookup* across row A instead of down a column, as we have seen before.

HLOOKUP SYNTAX

- Complete the following steps:

- In cell B8, enter:

- `=HLOOKUP($A8,$B:$L,2,FALSE)`

- We are looking up the name, so \$A8 is our lookup_value
 - The table_array is the table of grades.
 - The row_index_number is how far *down* we need to go.
 - The range_lookup is the same as VLOOKUP: does our match need to be exact or not?
 - Copy the formula down two cells by dragging the bottom-right corner.

CREATE OTHER HLOOKUPS

- In cell C8, enter:

`=HLOOKUP($A8,$B:$L,3,FALSE)`

- Copy the formula down two cells.

- In cell D8, enter:

`=HLOOKUP($A8,$B:$L,4,FALSE)`

- Copy the formula down two cells.

- Admittedly, HLOOKUP can be quite rare. It can be used whenever your variables are each a row and your observations are across a column.

DATA REFERENCING IN EXCEL

INTRODUCTION: INDEX AND MATCH

INTRODUCTION TO INDEX AND MATCH

- VLOOKUP (and HLOOKUP) are great for “looking up” data, but they are not perfect. To see why not, do the following:
 - Open our dataset to see the county names we have added via VLOOKUP previously.
 - Verify that they look as expected.
 - Go to the file from which they are imported:
(census_tract_county_names_WA.xlsx)
 - Insert a new column B, moving the county names over to column C.
 - Return to our dataset and look at the county names – what

INTRODUCTION TO INDEX AND MATCH

- Because the VLOOKUP references a col_index, and that col_index is a simple number and not an actual column or cell reference, it is unable to automatically update it when you insert a column or columns to the table_array.
- Don't worry, INDEX/MATCH to the rescue!

DATA REFERENCING IN EXCEL

DEMO: A SIMPLE INDEX/MATCH EXAMPLE

INDEX/MATCH EXAMPLE

- Open L2_demo_worksheets.xlsx and go to the demo3_index worksheet.
- In cell D2, type:
`=INDEX(A:A, 4)`
- In cell E2, type:
`=MATCH("Coats", A:A, 0)`
- For now, we will always use a 0 as the third argument.
 - In the next section, we will see how to use an inexact match.
 - A 0 means an exact match.

INDEX/MATCH EXAMPLE

- In cell H2, we will now combine INDEX and MATCH into a *nested formula*:
`=INDEX(A:A,MATCH(G2,B:B,0))`
- Copy it down to H3.
 - The inner MATCH looks up the Product ID of interest in the B:B column, returning the row number that matches.
 - The row number from the MATCH above is then used in the INDEX to lookup a value in column A:A and return it.
 - And we get the result!
- Not only can we insert a column between A and B and not break our lookup table (try it), but we also looked up a value on the *left* of the matched value. That's another thing we cannot do with VLOOKUP!

DATA REFERENCING IN EXCEL

GUIDED PRACTICE: REDOING OUR VLOOKUPS

REDO OUR VLOOKUPS WITH INDEX/MATCH

- To practice, let's recreate our VLOOKUPs from earlier using INDEX/MATCH.
- In a new column, let's use INDEX/MATCH to look up the area of the census tracts again:
 - In row 2 of the new column, enter:
`=INDEX(census_tract_areas_WA!N:N,
MATCH('2014_acs_select'!A2,census_tract_areas_WA!E:E,0))`
 - Copy it down to all rows.
 - To check equality with our VLOOKUP, in the next column enter
`=AN2=AJ2` (adapt as necessary if columns differ).
 - Copy down and ensure all values are TRUE.

REDO OUR VLOOKUPS WITH INDEX/MATCH

- In a new column, let's redo the density classification:
 - In row 2 of the new column, enter:
`=INDEX('Density Lookup'!B:B,MATCH('2014_acs_select'!AK2,'Density Lookup'!A:A,1))`
- Notice we are using a '1' as the third argument of MATCH. This is the inexact match that is equivalent to the TRUE fourth argument of VLOOKUP we used for the density classification earlier.
- Copy it down to all rows.
- To check equality with our VLOOKUP, in the next column enter `=AP2=AL2` (adapt as necessary if columns differ). Copy down and ensure all values are TRUE.

DATA REFERENCING IN EXCEL

**INDEPENDENT
PRACTICE: PRACTICING
WHAT WE HAVE LEARNED**

ACTIVITY: PRACTICING VLOOKUP AND INDEX/MATCH



EXERCISE

DIRECTIONS

1. Open L2_independent_activity.xlsx
2. Based on your experience, choose either the BASE or STRETCH tab to complete (25 min).

You may work with a partner, checking in with each other after answering each question.

DELIVERABLE

Complete BASE or STRETCH tab in L2_independent_activity.xlsx

DATA REFERENCING IN EXCEL

Q & A

DATA REFERENCING IN EXCEL

CONCLUSION

CONCLUSION

- Today we learned two advanced and very valuable techniques in Excel: VLOOKUP and INDEX/MATCH.
- VLOOKUP's cousin, HLOOKUP, was also introduced.
- Next, we will continue with the two more advanced Excel topics: aggregate functions and pivot tables!

DATA REFERENCING IN EXCEL

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET

DATA REFERENCING IN EXCEL

CREDITS

DATA REFERENCING IN EXCEL

CITATIONS

- Census tract densities: <https://www.census.gov/geo/maps-data/data/tiger-line.html>
- HLOOKUP example adapted from:
http://www.exceltrick.com/formulas_macros/hlookup-in-excel-with-examples/
- INDEX/MATCH example adapted from:
<http://www.randomwok.com/excel/how-to-use-index-match/>

DATA REFERENCING IN EXCEL

RESOURCES

- <http://www.randomwok.com/excel/how-to-use-index-match/>
- <https://www.deskbright.com/excel/using-index-match-match/>