

DATA REFERENCING IN EXCEL

DATA REFERENCING IN EXCEL

LEARNING OBJECTIVES

In today's lesson, we will:

1. Build relationships between cells in Excel.
2. Use named ranges to easily reference data subsets.
3. Manipulate data sets using **VLOOKUP** and **HLOOKUP**.
4. Look up values in other tables using **INDEX** and **MATCH**.

DATA REFERENCING IN EXCEL

RECAP: DATA CLEANING



RECAP: DATA CLEANING - BEST PRACTICES

1. Keep a copy of all untouched, raw data.
2. Document your steps through cleaning and preparation:
 - a. Create new columns for converted data.
 - b. Use color to highlight prepared columns.
 - c. Use conditional formatting to identify outliers.
3. Create a summary sheet with “metadata” including:
 - a. A directory of other sheets.
 - b. An explanation of analysis.
 - c. A sample summary of results.

RECAP: DATA CLEANING - FUNCTIONS AND METHODS

- Duplicates
- Auto filter
- Concatenate()
- Trim blanks/spaces
- Boolean operators
- Define a strategy for NULLs.
- Clean non-printing charts
- Uniform appearance (Upper, Lower, Proper)
- Date field selections
- Numeric column settings
- Text to columns
- Aggregation (% , avg, sum)
- Substitute

RECAP: DATA CLEANING - MISSING DATA

STRATEGIES FOR NULLS

What are the **four** primary strategies?

1. **Delete them** (with caution).
2. **Ignore them** (some have meaning).
3. **Impute values** (median or zeros).
4. **Find the value** (reference resources).

Guideline: If your data has more than 15% null values, you should investigate the data quality further!

EXCEL: POWER SHORTCUTS



DATA REFERENCING IN EXCEL

DATA REFERENCING

DATA REFERENCING

Referencing, in its basic form, means pulling the value of one cell into another cell.

- A simple example of this is setting cell A2 to the value of A1 with “=A1.”
- Now A2 ***references*** A1, and the value of A2 will always be equal to the value of A1.

DATA REFERENCING


We'll take this concept of referencing to the next level.

- **VLOOKUP**, **HLOOKUP**, **INDEX**, and **MATCH** are often considered advanced Excel tools, and using them effectively can greatly increase your efficiency while reducing data integrity issues.
- **Named Ranges** are another useful way to keep your formulas clear and flexible.

DATA REFERENCING

As you build these new Excel functions into your work, keep in mind that there are at least two ways to quickly see the fields that Excel requires to complete the action.

- For example, when building a **VLOOKUP** statement:
 - In cell, type “=**VL**.”
 - Double click on the function name that’s presented for syntax.
 - Alternatively, click **fx** in the menu ribbon, then select “function.”

Font	Alignment	Number
	=VLOOKUP(
	VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])	
	Matching Codes to ACS Census	NAME
	1400000US53077940003	9400.03 Census Tract G5020

Function Arguments

VLOOKUP

Lookup_value = any

Table_array = number

Col_index_num = number

Range_lookup = logical

Looks for a value in the leftmost column of a table, and then returns a value in the same row from a column you specify. By default, the table must be sorted in an ascending order.

Lookup_value is the value to be found in the first column of the table, and can be a value, a reference, or a text string.

Formula result =

[Help on this function](#)

DATA REFERENCING: NAMED RANGE

Naming ranges in Excel is a simple way to identify, define, or refer to a single cell, or group of cells, to be identified in a formula as a table array.

Here's how to define a name for a cell or cell range in a worksheet:

- Select the cell, range of cells, or non-adjacent cells you want to name.
- Click the Name box at the left end of the formula bar.
- Type the name you want to use to refer to your selection.*
- Press return/enter.
- This “Named Range” can now be referenced in formulas, functions, etc.

DATA REFERENCING IN EXCEL

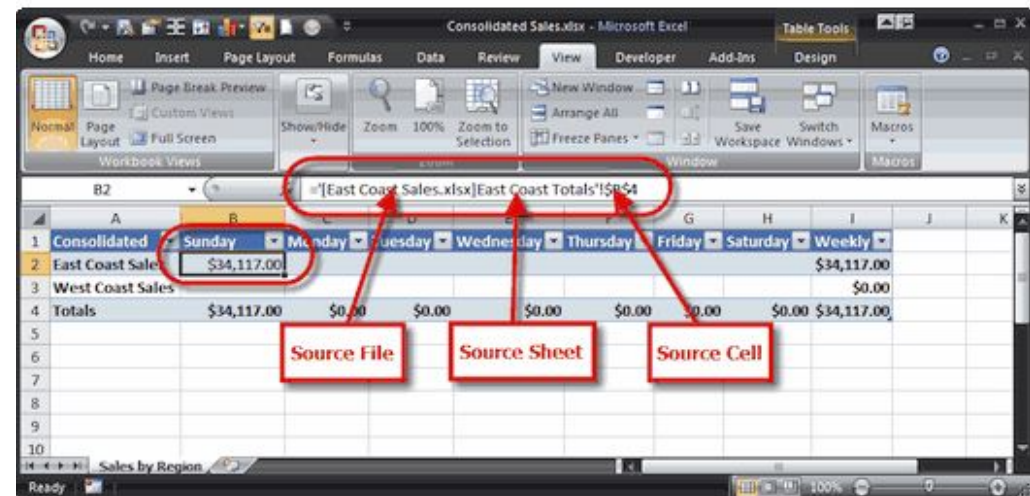
GUIDED PRACTICE: BUSINESS SCENARIO

GUIDED PRACTICE: BUSINESS SCENARIO

We'll combine data from another worksheet into our ACS data set.

We will also explore the concept of a “lookup table” and how to use one to categorize our quantitative data.

How to Link Cells in Different Excel Spreadsheets



DATA REFERENCING IN EXCEL

DEMO: VLOOKUP

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 data. This is usually an identifier (an ID of some kind). It must exist in both worksheets.

Table_array is the table from which you want to retrieve data.

Col_index_num is the number of the column from the left side of the table_array from which you want to retrieve data.

VLOOKUP SYNTAX

The syntax is:

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

Range_lookup defines whether or not the lookup_value is an approximate match or an exact match of the value you are comparing it to in the left-most column of the table_array.

- **TRUE**: Approximate match is needed.*
- **FALSE**: An exact match is required.

SAMPLE DATA SET

1. Open the “**demo1_accounts**” worksheet from the “**L3_demo_worksheets.xlsx**” workbook.
2. Then open the “**demo1_emails**” worksheet.

In order to combine these two lists — where we’ll create a new column in “**demo1_accounts**” with email addresses — we can use **VLOOKUP**.

VLOOKUP will “look up” one value in another table and return another column in that row.

VLOOKUP SYNTAX

1. In cell D2 of “**demo1_accounts**”, enter:
=VLOOKUP(C2,demo1_emails!A:B,2,FALSE)
 - a. The email address should populate cell **D2**.
2. Double click the bottom-right corner of the cell — or click and drag — to replicate this function down the entire column.
3. Implement a “**Named Range**” to simplify what’s required.

DATA REFERENCING IN EXCEL

**GUIDED PRACTICE:
USE VLOOKUP TO COMBINE
DATA ACROSS WORKSHEETS**

GUIDED PRACTICE

Earlier, we looked at the scatterplot comparing two columns in our data set:

- Population estimate of our census tract.
- Percent of people who commute with public transit.

There was no obvious relationship. This makes sense, as population and *population density* are not the same.

While we expect people in urban areas to use public transit more often, 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. Currently, we don't have density data in our data set.

Follow these steps to create this scatterplot:

1. Open “**Part_2_census_tract_area**”
 - a. Here we have a few variables; the one of interest is **area_sqkm (M)**.
 - b. 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 that they have a column/identifier in common.

Take a look at **column A** in our original data set and **column D** in the new data set. **Column A** in the cleaned ACS data from last class is equivalent to **column D** in the new data set — but each value has “**1400000US**” on its front.

1. In “**Part_2_census_tract_area**,” create a new column: **E**, and enter this formula into cell **E2**: **=“1400000US”&D2** or **CONCATENATE(1400000,“US”,D2)**
2. Expand it to all rows.
3. Check column data type for appropriate designation (general/text).
4. Expand it to all rows.

VLOOKUP

1. In our original worksheet, “**PART_1_ACS_Dataset**”, go to column **AJ**.
2. Name this “**area_sqkm**” by typing this in cell **AJ1**.
3. In cell **AJ2**, enter:
=VLOOKUP(A2,PART_2_census_tract_area!E:N,10,FALSE)
4. Expand this formula to all rows by double clicking the bottom-right corner of the cell.

Now we have the tract area in our data set!

CALCULATE THE DENSITY

1. In cell **AK1**, type “density.”
2. In cell **AK2**, enter the formula “**=E2/AJ2**” and expand.
3. **Create** a scatterplot between % public transit vs. your new density variable.
 - a. Do we see any interesting patterns 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

1. Open “**census_tract_county_names**”.

This file also includes the IDs, so we can use them as our common identifier to match the county names with the census tracts in our data set.

2. In our data set, label a new column “**County**”.

USE VLOOKUP FOR COUNTY NAMES

1. In the second row, enter our formula for **VLOOKUP**:

=VLOOKUP(A2, census_tract_county_names!\$A:\$B, 2, FALSE)

When selecting the table_array, it's easy to use the cell selector (the mouse icon that's a white "+") to select the columns of our table_array. This is especially useful when crossing over to another file, as we're doing here.

2. Copy the contents down by double clicking the square in the bottom-right corner. We've now successfully used **VLOOKUP** with two examples.

DATA REFERENCING IN EXCEL

**GUIDED PRACTICE:
USE VLOOKUP TO CREATE
CATEGORICAL VARIABLES**

VLOOKUP

Often, it's 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.

1. Let's classify our census tracts as either *low*, *medium*, or *high*.

VLOOKUP

What are some ways to create classification groups? One option is to break them down into **percentiles**, using the 33rd, 66th, and 99th percentiles.

A **percentile** is generally used to define which value has x percent 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.

VLOOKUP

1. Create a new worksheet called ***Density Lookup***.
2. In cells **A1**, **A2**, and **A3**, enter values 0, 350, and 1420.
3. In cells **B1**, **B2**, and **B3**, enter values low, medium, and high.

Execution notes:

- The lookup values *must always* be located on the left side of the data used for a table array when using **VLOOKUP**.
- In this case, we select the TRUE parameter for the **range_lookup** value, causing the lookup function to compare the ascending range boundaries and return the category designation in the second column.

VLOOKUP

1. On worksheet “**Part_1_ACS_Dataset**”, type “**Density Group**” in cell **AL1**.
2. In **AL2**, complete the lookup:
=VLOOKUP(AK2, 'Density Lookup' !A:B, 2, TRUE)
3. 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/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 “**L3_demo_worksheets.xlsx**”, consider the top table of grades for different students. In this case, the students are columns instead of rows.

HLOOKUP

Suppose we needed to fill out the second table using only formulas?

We would use an **HLOOKUP**, as we need to do the *lookup* across row A, instead of down a column.

Pro Tip: In the real world, **HLOOKUP** is used much less frequently. However, it is a method that can be used whenever your variables are each contained in a row and your observations are organized across a column.

HLOOKUP SYNTAX

Complete the following steps:

1. In cell B8, enter: **=HLOOKUP(\$A8,\$B:\$L,2,FALSE)**
 - a. We're looking up the name, so **\$A8** is our `lookup_value`.
 - b. The **table_array** is the table of grades.
 - c. The **row_index_number** is how far *down* we need to go.
 - d. The **range_lookup** is the same here as in as **VLOOKUP**
 - e. Does our match need to be **FALSE** (exact match) or **TRUE** (range)?
2. Copy the formula down two cells by dragging the bottom-right corner.

CREATE OTHER HLOOKUPS

1. In cell C8, enter: **=HLOOKUP(\$A8 , \$B:\$L , 3 , FALSE)**
2. Copy the formula down two cells.
3. In cell D8, enter: **=HLOOKUP(\$A8 , \$B:\$L , 4 , FALSE)**
4. Copy the formula down two cells.

VLOOKUP & HLOOKUP: SYNTAX SUMMARY

Function arguments:

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

=HLOOKUP(lookup_value, table_array, row_index_num, [range_lookup])

- **Value:** The value to look for in the first column of a table.
- **Table:** The table from which to retrieve a value.
- **col_index:** The column (row) in a table from which to retrieve a value.
- **range_lookup:** [optional]
 - **TRUE** = Approximate match. (Default)
 - **FALSE** = Exact match.

INTRODUCTION: INDEX AND MATCH

INTRODUCTION TO INDEX AND MATCH

VLOOKUP (and **HLOOKUP**) are great for “looking up” data, but they do have two significant limitations:

1. They are **unidirectional**, meaning they must work with indexes fixed to the left side (**VLOOKUP**) or top (**HLOOKUP**) of the work area.
2. Because the **VLOOKUP** references a **col_index**, it's unable to **dynamically** update whenever you insert a column or columns in the table_array.

For additional flexibility and features, we'll need to use **INDEX/MATCH**.

INTRODUCTION TO INDEX AND MATCH

INDEX/MATCH

- Faster than LOOKUPS.
- More flexible.

INDEX returns the **value** at the **intersection** of a row and column in a given range.

- **Syntax:** =**INDEX**(Array, Row_num, Column_num).
- **Example:** I want to know name of the customer in the fifth row of the “customer_name” column of my Sales table.

INTRODUCTION TO INDEX AND MATCH

MATCH returns the **position** of an **item** in an **array** that **matches** a **value**.

- **Syntax:** **MATCH**(Lookup_value, Lookup_array, Match_type)
 - **Lookup_value:** Value or cell reference.
 - **Lookup_arrange:** Named range or specified cells.
 - **Match_type:** “0”: Exactly equal to lookup_value.
 - “-1”: Smallest value $>$ or $=$ lookup_value.*
 - “1”: Largest value $<$ or $=$ lookup_value.*
- **Example:** I want to find the first occurrence of the words “Order Error” in the column containing order completion status.

INTRODUCTION TO INDEX AND MATCH

These functions are powerful when combined. An **INDEX/MATCH** lookup allows you to specify lookup column and return the value column **independently**.

Syntax (Combined):

=**INDEX**(Return_value_range, **MATCH**(Lookup_value,Lookup_value_range, Match_type))

- The search is not case sensitive.
- If looking up text, you may use wildcards to search for a partial string.
- An unsuccessful match will return #N/A.

Example: I want to return the name of the customer whose order completion status is “Order Error.”

DATA REFERENCING IN EXCEL

DEMO: A SIMPLE INDEX/MATCH EXAMPLE

INDEX/MATCH EXAMPLE

1. Open “**L3_demo_worksheets.xlsx**” and go to the “**demo3_index**” worksheet.
2. In cell **D2**, type: =**INDEX**(**A:A**, 4)
3. In cell **E2**, type: =**MATCH**(“Coats”, **A:A**, 0)

For now, we’ll always use a “0” as the third argument. The “0” parameter requires an **exact** match. In the next section, we’ll see how to use an **inexact** match.

GUIDED PRACTICE: REVISING OUR VLOOKUPS

INDEX/MATCH EXAMPLE

1. In cell **H2**, we will now combine **INDEX** and **MATCH** into a **nested formula**:
=INDEX(A:A, MATCH(G2, B:B, 0))
2. Copy it down to **H3**.
3. The inner **MATCH** looks up the “Product ID” of interest in the **B:B** column, returning the matching row number.
4. The row number from the **MATCH** above is then used in the **INDEX** to look up and return a value in column **A:A**.

REVISE VLOOKUPS WITH INDEX/MATCH

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:

1. In row 2 of the new column, type:
=INDEX(PART_2_census_tract_areas!N:N, MATCH(A2, PART_2_census_tract_areas!E:E, 0))
2. Copy this down to all rows.
3. To check equality with our **VLOOKUP** by entering “**=AN2=AJ2**” in the next column (adapt as necessary if columns differ).
4. Copy down and ensure all values are **TRUE**.

REVISE VLOOKUPS WITH INDEX/MATCH

In a new column, let's redo the density classification.

1. In **row 2** of the new column, enter: **=INDEX('Density Lookup'!B:B, MATCH(AK2, 'Density Lookup'!A:A, 1))**
2. Copy it down to all rows.
3. To check equality with our **VLOOKUP**, enter “**=AP2=AL2**” in the next column (adapt as necessary if columns differ).
4. Copy down and ensure all values are **TRUE**.

DATA REFERENCING IN EXCEL

INDEPENDENT PRACTICE

ACTIVITY: PRACTICE VLOOKUP AND INDEX/MATCH



EXERCISE

DIRECTIONS

1. Open “**PART_2_Independent_Activity_Dat**”
2. Based on your experience, choose to complete either the BASE or STRETCH tab (25 min).

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

DELIVERABLE

Complete BASE or STRETCH tab in “*Day 1- Data Analytics with Excel.*”

DATA REFERENCING IN EXCEL

CONCLUSION

CONCLUSION

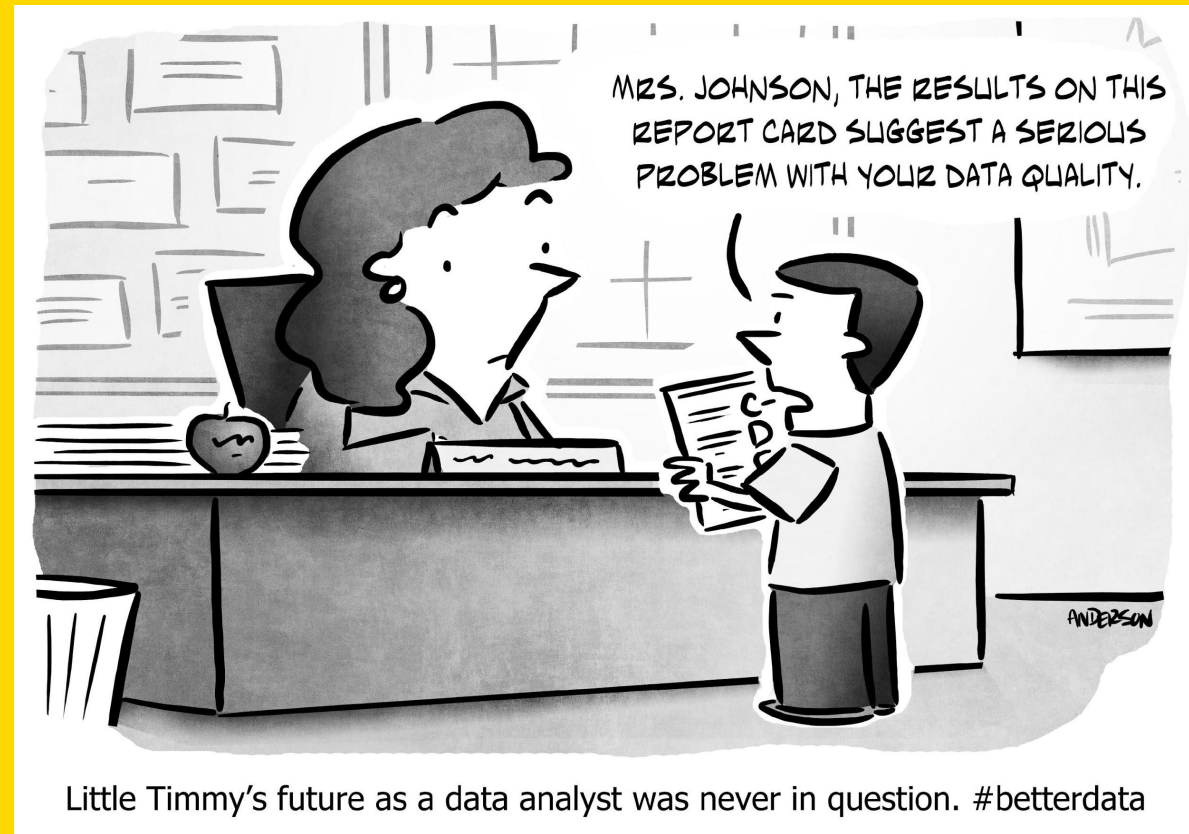
In this lesson, we:

1. Reviewed data cleaning techniques and strategies.
2. Introduced **NULLs** with four options for preparation.
3. Learned how to use **named ranges**.
4. Demonstrated and applied two valuable data referencing techniques in Excel:
 - a. **VLOOKUP** and **HLOOKUP**
 - b. **INDEX** and **MATCH**

In our next lessons, we'll continue on with more advanced Excel topics:
aggregate functions, PivotTables, and conditional formatting!

DATA REFERENCING IN EXCEL

Q&A



DATA REFERENCING IN EXCEL

RESOURCES

DATA REFERENCING IN EXCEL

RESOURCES

- Excel Keyboard Shortcuts: <https://goo.gl/EhClMw>
- <http://www.randomwok.com/excel/how-to-use-index-match/>
- <https://www.deskbright.com/excel/using-index-match-match/>

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/>