APPEND AND MERGE

Climate change affects the environment is many ways. This past winter produced little snow in California which is likely to compound their drought. The lack of snow also negatively impacts the $12 billion winter sports economy in the United States.

|  |  |
| --- | --- |
|  | http://www.theodora.com/wfb/photos/canada/big_white_ski_resort_village_kelowna_british_columbia_canada_photo_tourism_bc-don_weixl.jpg |

The National Oceanic and Atmospheric Administration (NOAA) provides a lot of data that is freely available. Several regional weather offices also publish data. Data from the Western Regional Climate Center (Website: <http://www.wrcc.dri.edu/>) is used in this handout.

The following graphs of Colorado snowfall were computed from a dataset that required a substantial amount of data management before summaries could be constructed.

|  |  |
| --- | --- |
| Colorado average snowfall amounts - Aggregated by County and Month | The effect of elevation on January snowfall for weather stations in Colorado |

The data from this handout is from the Western Regional Climate Center. This data was provided in two files. One file contained typical snowfall amounts for each month of the year. A second file contained relevant auxilary information about the weather station for which snow measurements were obtained.

Procedural Steps

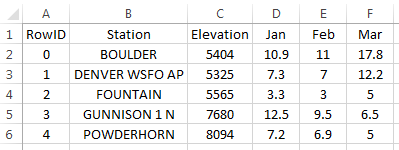
1. Create a stacked version of the snowfall amount dataset by appending columns
2. Merge various demographic features of the stations into the dataset created above
3. Construct appropriate summaries and visualizations

Data Technologies

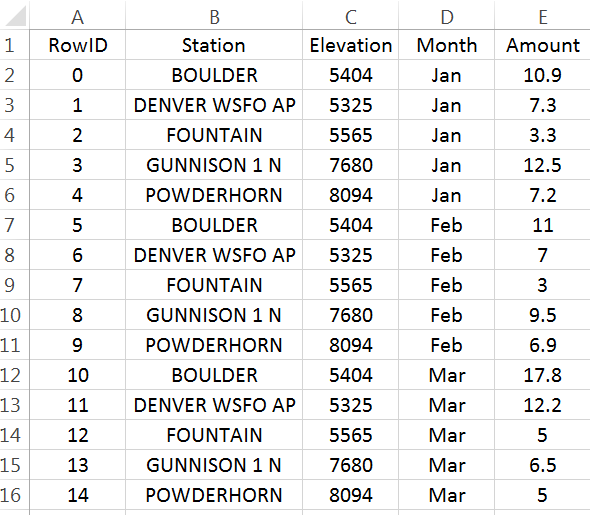
1. OFFSET function to append columns of data
2. VLOOKUP, MATCH, and INDEX functions to merge elements of one dataset into another
3. Summaries and visualizations through PivotTables

Enter the following data into Excel. This is a very small subset of the snowfall dataset and will be used to understand how to append columns in Excel.

Enter the following data into Excel



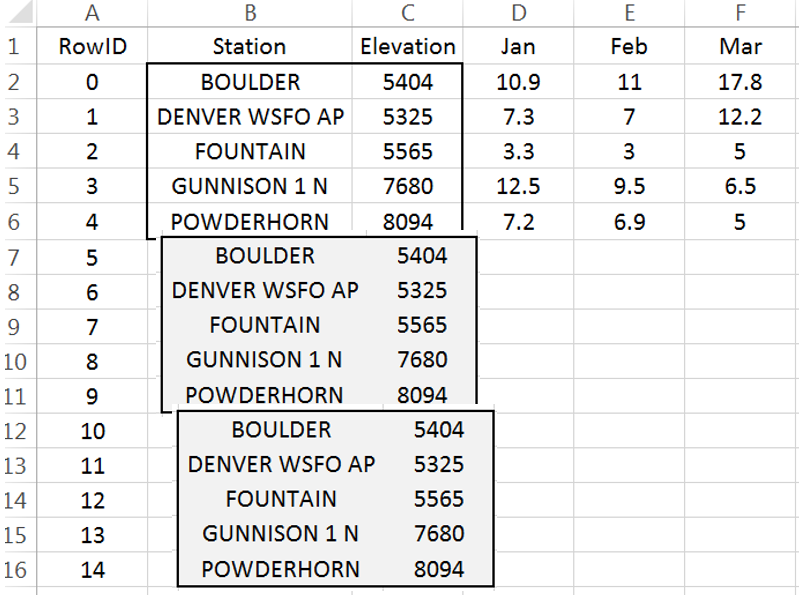
The task here is to create a stacked version of the data by appending the Jan, Feb, and Mar columns



Replicating Row Content

The first step to appending columns is to create replicates of the information not being stacked. For our example, station and elevation need to be replicated twice in our example.

# Replicates required = # Columns to be stacked – 1



Insert a new column to the left of Station. Label this column RowID. Starting with 0, create a sequence from 0 to 14. This data has 5 rows and 3 columns are to be stacked. The number of rows needed for a stacked version of the dataset is 15.

Modular arithmetic will be used and starting the sequence with 0 will prove to be easier than starting the sequence at 1.

Next, insert a column between RowID and Station. Label this new column Row Reference. This column will identify which row of the original dataset is being referenced. Enter the following into cell B2 and copy down for all cells.

Cell B2: =MOD(A2,5)

|  |  |
| --- | --- |
| Enter MOD formula into cell B2 as shown here | The Row Reference provides the needed reference for the replicates. |

The =OFFSET() function in Excel will be used to replicate the necessary contents for each row. This function returns the contents of another cell. The contents being returned is determined by the number of cells down and to the right from a reference cell.

|  |  |
| --- | --- |
| =OFFSET() requires the specification of a reference cell, i.e. cell C2 here. The row reference is specified by B7 and the column reference should be set to 0.    Cell B7 contains 0, thus in this instance =OFFSET() will shift 0 rows down and 0 rows to the right to obtain the contents. Here =OFFSET() places BOULDER in cell C7 as desired. | Cell C7: =OFFSET($C$2,B7,0)  Copy this down for remaining cell. The Station names are replicated as needed. |

A similar process can be used to replicate the Elevation values.

|  |  |
| --- | --- |
| Put the following formula in cell D7 and copy down for the remaining cells.  Cell D7: =OFFSET($C$2,B7,1) | Elevation should now be replicated as shown here |

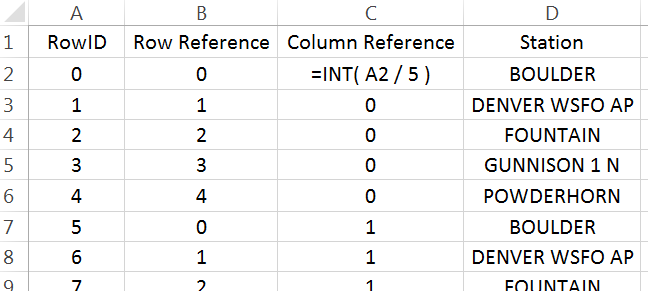
Appending Columns

The process needed for the columns to be appended is slightly different than above. The =OFFSET() function must automatically shift to the right for each replicate.

|  |  |
| --- | --- |
| Cell F2 will be used as the reference cell | For RowIDs 5 through 9, the column index should be set to 1; however, the column index should be 2 for RowIDs 10 through 14. |

Insert another column for the Column Reference. Enter the following formula into Cell C2.

Cell C2: =INT(A2 / 5)

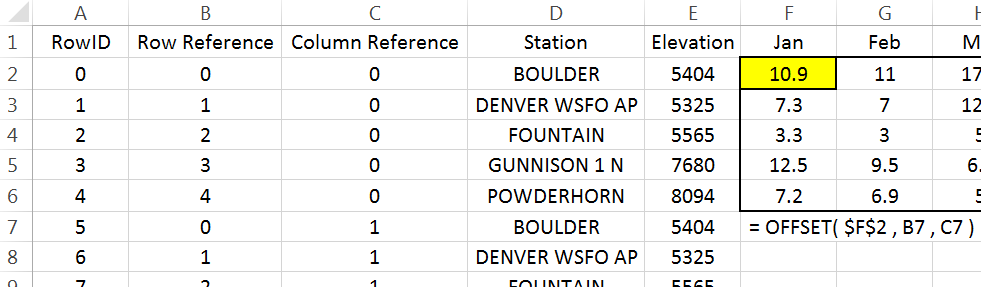


The =INT() function is equivalent to the floor function and simply returns the integer part of a number.

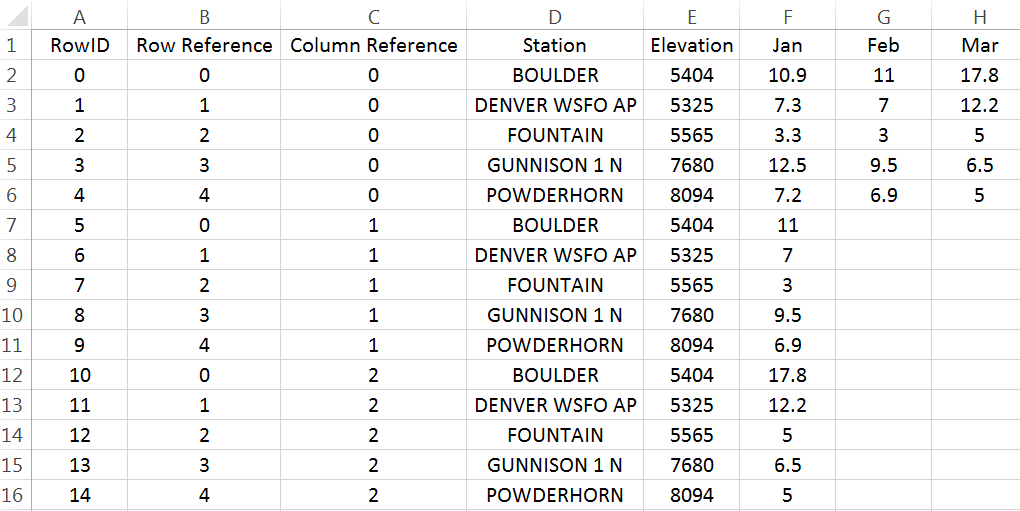
|  |  |
| --- | --- |
|  |  |
|  |  |

After creating the Column Reference column, type the following into cell F7. $F$2 will be used as the reference cell. This function also makes use of the row and column references. Copy this formula down for all remaining cells.

Cell F7: =OFFSET($F$2, B7, C7)

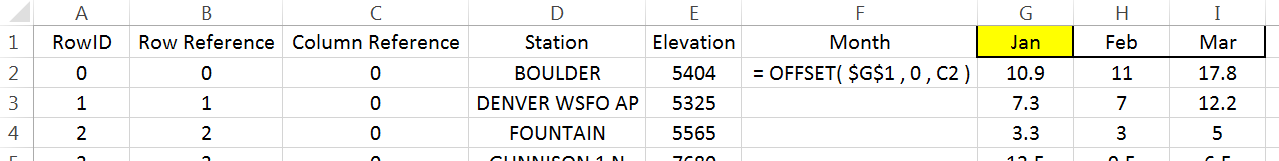


The snowfall amounts should now be stacked.

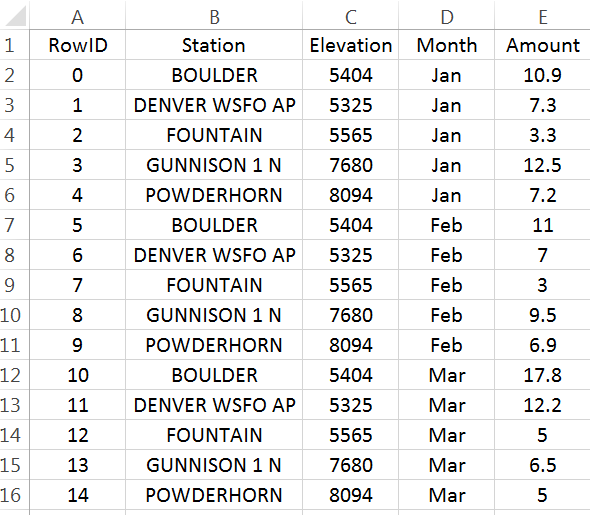


The last step is to identify the month for each row. Click on column F, right click and select Insert. Name this new column Month. Enter the following into cell F2.

Cell F2: =OFFSET($G$1, 0, C2)



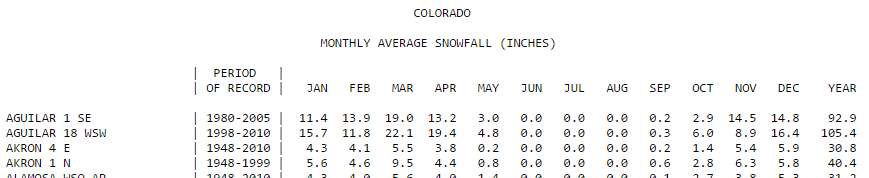
A final version of the stacked dataset is shown here. Unfortunately, the contents in many of these cells rely on the contents of other cells. You may find it beneficial to create a second version of the data that does not contain such dependencies. When making the second copy, select Paste Values to remove the dependencies.



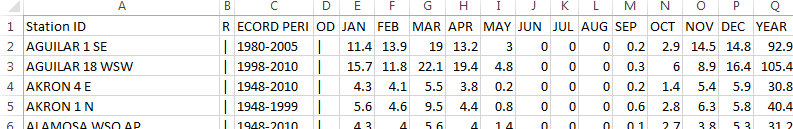
Working with Complete Dataset

|  |  |
| --- | --- |
| Data Source | |
| Address | <http://course1.winona.edu/cmalone/workshops/uscots2015/> |
| Description | CO Snowfall Datasets   |  |  | | --- | --- | | The Western Regional Climate Center provides a text file of the historic monthly average snowfall (inches) amounts for weather stations across Colorado. Information about each weather station is provided in a second text file.  Link to data:  <http://www.wrcc.dri.edu/htmlfiles/co/co.sno.html>  Link for station information: <http://www.wrcc.dri.edu/inventory/sodco.html> |  | |

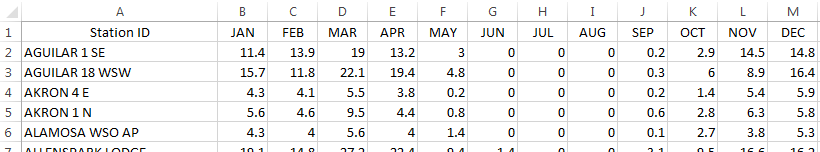
Often data downloaded from the internet must be cleaned before importing into Excel or other software packages. For example, the header content on this file should be removed before importing. The files provided on the workshop website have the unwanted header information removed.



Import the monthly snowfall data into Excel. Select Data > From Text, specify Fixed width in Step 1 of the import wizard. Continue through the remaining steps of the import wizard. You should delete Columns B, C, D, and Q as these columns will not be used here.

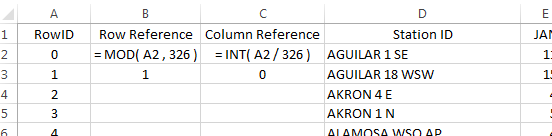


The following snippet show the data that needs to be stacked.



Similar to the example discussed above, the goal here is to stack the Month columns. There are a total of 326 rows in this dataset and 12 columns are to be stacked.

Insert a RowID column. Again, starting with 0, create a sequence from 0 to 3911 with increments of size 1.

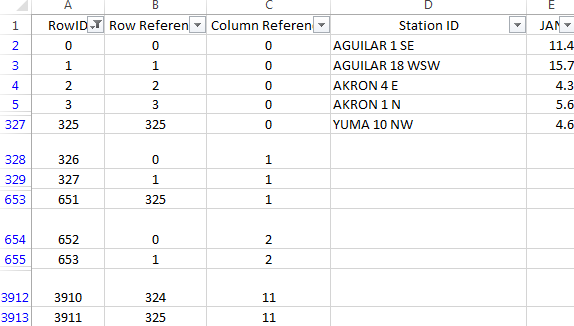


Next, insert two columns which will be used for the row and column reference. Type the following equations into cells B2 and C2. Copy these formulas down for all rows.

Cell B2: =MOD(A2, 326)

Cell C2: =INT(A2 / 326)

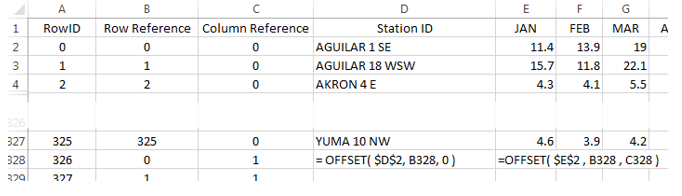
Before starting with the =OFFSET() function, verify these formulas have produced the desired outcomes.



If the row and column references have been correctly specified, then we can proceed with the =OFFSET() function. Akin to the example above, the Station ID column must be replicated several times. A 0 is used for the column reference when stacking the Station IDs. However, a column reference is needed for the column to be stacked. This reference is contained in Column C.

Cell D328: =OFFSET( $D$2,B238 , 0 )

Cell E328: =OFFSET( $E$2 , B328 , C328 )

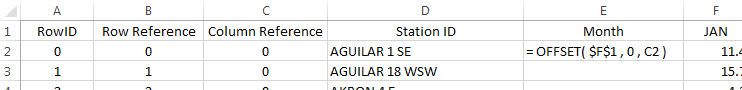


Question

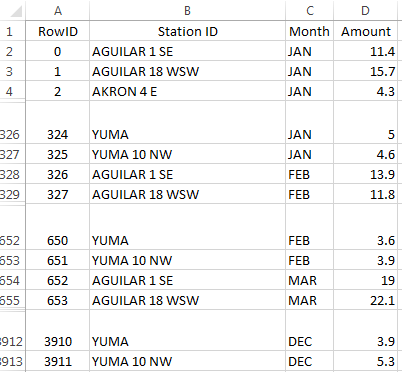
1. How would the formula for the =OFFSET() function in column E be written if $D$2 is used as the reference cell?

Finally, insert a column to the left of Jan and name this column Month. Copy this formula down for all cell.

Cell E2: =OFFSET( $F$1 , 0 , C2 )



Verify that all columns have been properly stacked and the content of all rows has been correctly specified. Obtain a copy of the this data using Paste Values to remove all cell dependencies. A snippet of the final dataset is provided here for reference.

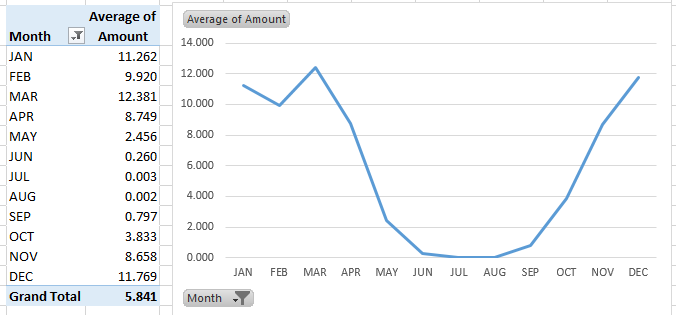


Getting Summaries Stacked vs. Unstacked

The efficient use of PivotTables requires the data be in a stacked structure.

|  |  |
| --- | --- |
| Stacked data: getting  averages by Month | Unstacked data: getting averages by Month |

The PivotTable output and its associated visualization using the stacked version of the data.



Questions

1. Which month has the highest average snowfall?

1. It is true that it has snowed in every month at some point in Colorado’s history. Explain how this data supports this statement.

Merging Content from Tables

A data scientist must often merge data from one table into another table. The discussion here will again use the subset of the CO snowfall data from above. Table A consists of the stacked data from the example discussed above. Table B contains information auxiliary information for various stations in CO.

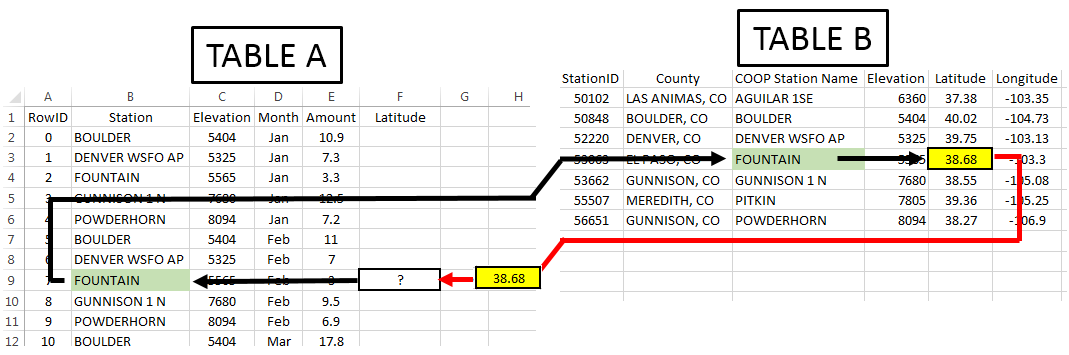
|  |  |
| --- | --- |
| Stacked data from above example | Auxiliary information for weather stations in CO |

The =VLOOKUP() function in Excel can be used to merge the content from one table into another.



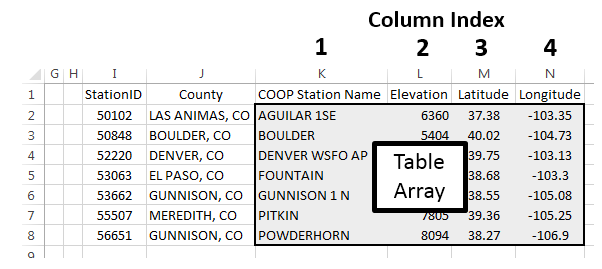
* First argument: specifies the value to be looked up
* Second argument: specifies the cell range of Table B or a subset of Table B
* Third arugment: an index that specifies which column is used to retrieve content from Table B
* Fourth argument: This should be set to FALSE. FALSE forces =VLOOKUP() to find an exact match

The following provides a visualization of the =VLOOKUP() function for cell F9.



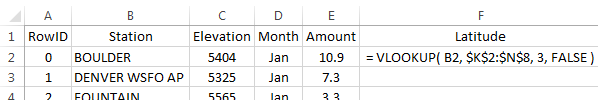
Note: Table B may or may not be contained in the same worksheet as Table A.

In this example, the second argument (i.e. the table array) is specified as K2:N8. If the Latitude value is to be returned, then the column index should be set to 3.

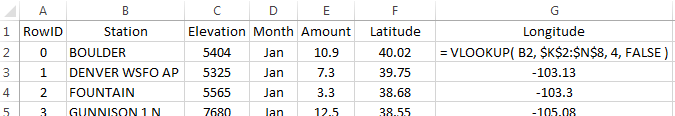


Type the following formula into cell F2.

Cell F2: =VLOOKUP( B2 , $K$2:$N$8 , 3 , FALSE )

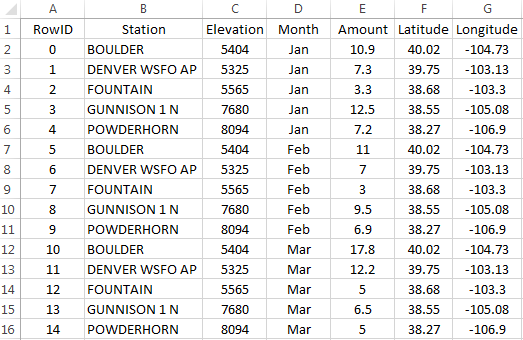


Copy this formula down for the remaining cells.



Repeat this process for Longitude by typing the following into cell G2 and copying down for all cells.

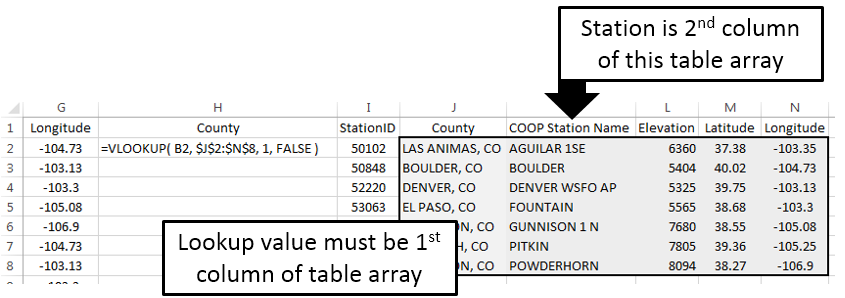
Cell G2: =VLOOKUP( B2 , $K$2:$N$8 , 4 , FALSE )



Next, suppose County is to be merged into Table A. The =VLOOKUP() function will not work for County as this function always uses the left-most column of the table array to search for a match. The =VLOOKUP() function fails here because Station is not the left-most column of the table array.

|  |  |
| --- | --- |
| The following will not work.  Cell H2: =VLOOKUP(B2 , $J$2:$N$8 , 1 , FALSE ) |  |

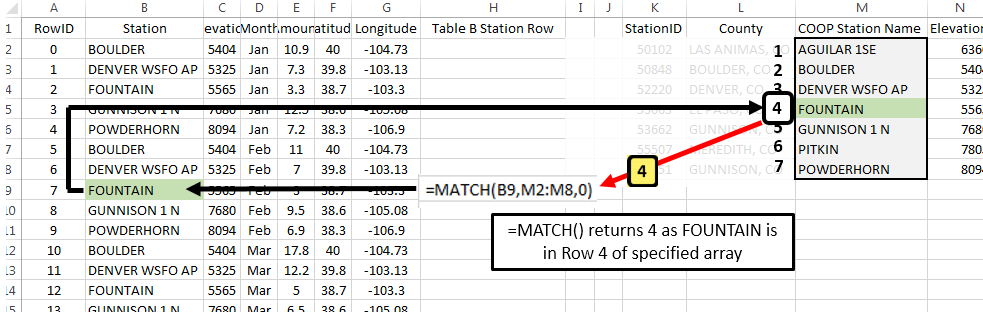
The column containing the lookup value must be the left-most column of the table array.



Using =MATCH() and =INDEX() to Merge Tables

The =MATCH() / =INDEX() approach to merging tables in Excel is considered to be better than =VLOOKUP(). This method requires two steps.

Suppose the County for RowID 7 is to be obtained. The =MATCH() function does not return the requested content from Table B, but instead returns the row number of Table B that matches the lookup value.



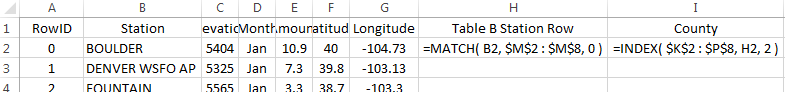
The second step of this process is to retrieve the actual contents from Table B. This is done using the =INDEX() function. The =INDEX() function is similar to the =OFFSET() function used earlier. =OFFSET() required the specification of a single cell to be used as a reference. INDEX() requires specification of the entire range of cells along with information regarding which row and column to return.

|  |  |
| --- | --- |
| Specify $K$2:$P$8 as the cell range for =INDEX() function | =INDEX() will return EL PASO, CO for RowID 7 |

Type the following formulas into Excel and copy these formulas down for all remaining cells.

Cell H2: =MATCH( B2 , $M$2,$M$8 , 0 )

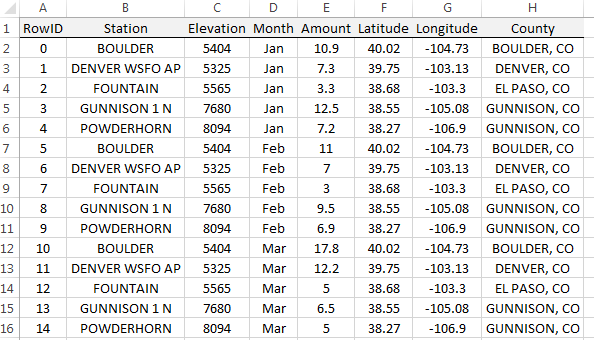
Cell I2: =INDEX( $K$2:$P$8,H2 , 2 )



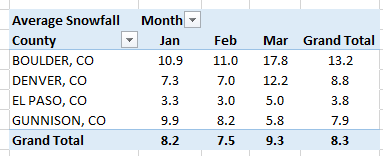
Questions

1. What is the purpose of the third argument in the =MATCH() function?
2. Why is 2 specified as the last argument in the =INDEX() function?

The following table is a successful merge of County from Table B into Table A.

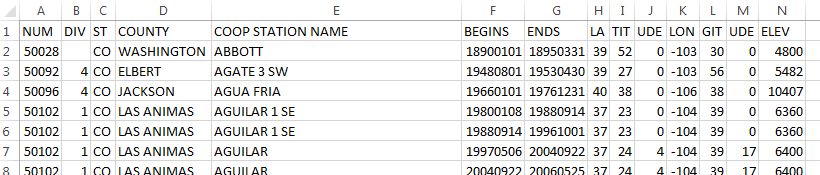


After the successful merging these two tables, averages can now be computed over county as is shown here. The latitude and longitude values are necessary for mapping snowfall.



Return to Complete Dataset

Import the station data into Excel. Select Data > From Text, specify Fixed width in Step 1 of the import wizard. The following snippet shows the first few rows of the station dataset.



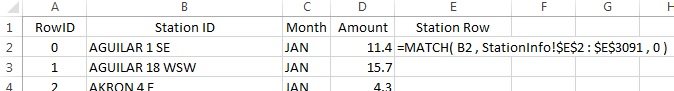
In Excel, name this worksheet StationInfo on the tab near the lower-left corner.



The station dataset contains information on many stations that are not present in our dataset. Also, several stations are replicated because new weather stations are added and others are removed from time to time. The =MATCH() and =VLOOKUP() functions use the first instance of a match. These functions ignore rows after an exact match is found.

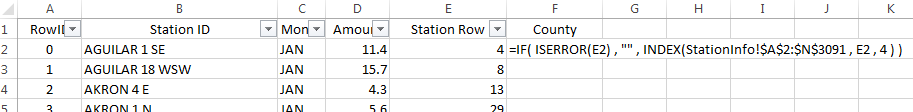
Assuming you have named the worksheet containing the station data StationInfo, type the following into cell E2 in the stacked version of the snowfall dataset. Column E is being used in this formula as this column contains the Station IDs in Table B.

Cell E2: =MATCH(B2, StationInfo!$E$2:$E$3091, 0 )

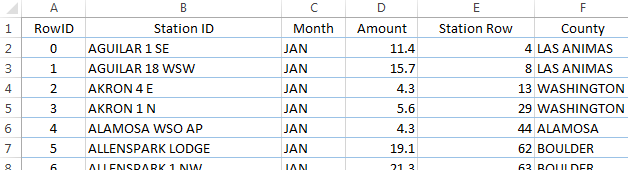


Copy this formula down for all cells. Some Station IDs from Table A cannot be found in Table B. In this case, a #N/A values is appropriately retuned by the =MATCH() function. This formula will not provide a County name when an #N/A is returned by the =MATHC() function. Next, the =INDEX() function can be used to retrieve County from Table B.

Cell F2: =IF( ISERROR(E2) , ”” , INDEX( StationInfo!$A$2:$N$3091 , E2 , 4 ))



The following shows a successful merge of the County information from the StationInfo worksheet into the Snowfall dataset.



Next, in column G, the following formula can be used to merge Elevation from the StationInfo worksheet into the dataset.

=IF( ISERROR(E2), "", INDEX(StationInfo!$A$2:$N$3091, E2, 14 ) )

Questions

1. What is the purpose of the empty string, i.e. “”, in the formula above?
2. What happens if the following is used in cell F2 instead of the formula provided above for merging County?

Cell F2: =INDEX(StationInfo!$A$2:$N$3091,E2,4)

1. Some software packages will create maps based on county names. However, abbreviations for state must be included with the county name. Use the following formula to concatenate County with the state abbreviation for CO.

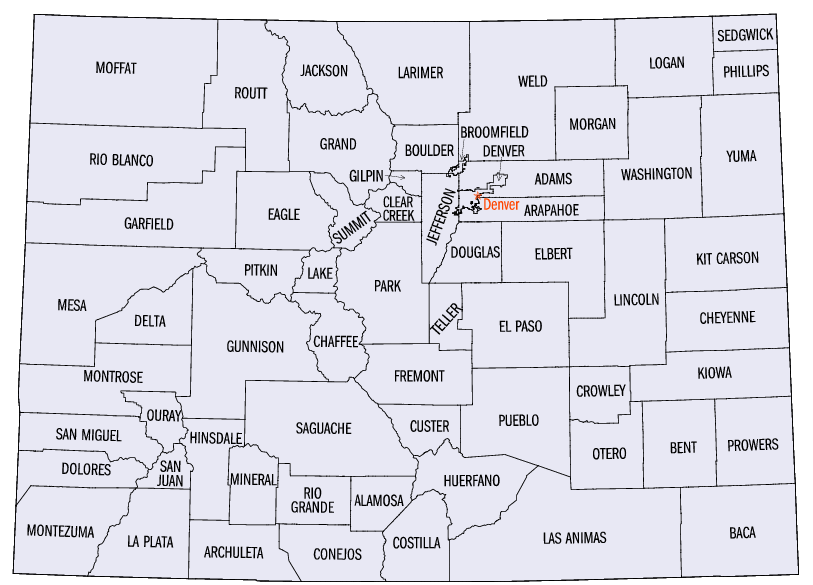
Cell G2: =IF(ISERROR(E2), “”, CONCATENATE(F2,”, CO“) )

Summaries using Merged Content

A summary of total snowfall by county is being requested by your boss. You have successfully merged these dataset and create the following PivotTable.

|  |  |
| --- | --- |
| Snowfall by County | The PivotTable structure used  to create this table |

A map of the counties in Colorado is given here for reference.



Questions

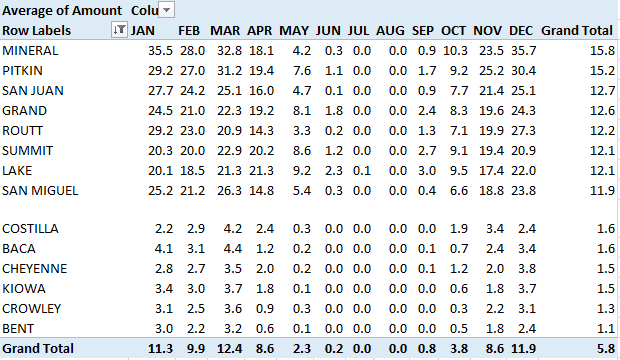
1. Your boss makes the following comment, “There is no way El Paso County has 120 weather stations.” Your boss is correct. How many weather stations does El Paso County have in this dataset?
2. The SUM is being used here as the total snowfall over the entire year is of interest. I’d argue that a SUM should not be used as the number of stations per county is not the same. Do you agree or disagree? Explain.

A PivotTable based on averages, instead of totals, is shown below.

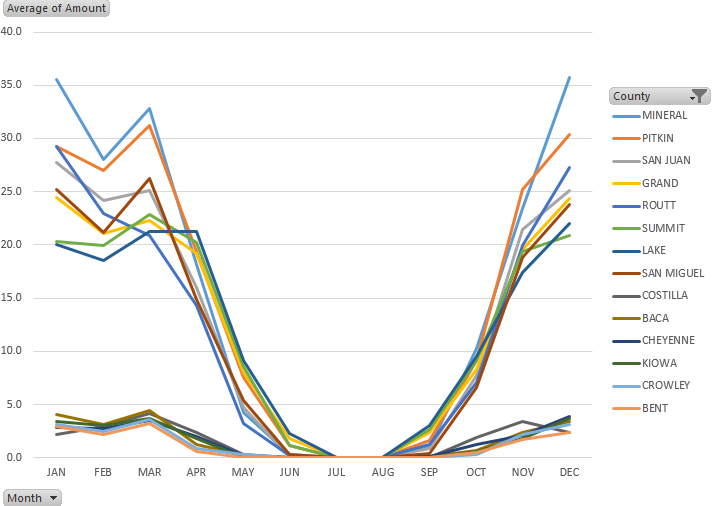
|  |  |
| --- | --- |
| Average Snowfall by County | Pareato-type chart of average snowfall amounts |

Questions

1. The average snowfall for Mineral County is 15.8 inches. Provide an interpretation for this value.
2. I’d argue that the averages provided here collapse the data too much. For example, do you believe the average for Mineral County provided above is a good estimate for snowfall in July? How about January? Discuss.
3. The following table shows the snowfall by month across counties in CO. Recreate this table in Excel.



1. The following visualization is from the PivotTable provided above. Is it true that for most of these counties the snowfall amount increase from Oct through Dec? Is it true that snowfall tends to steadily decrease from Jan through Apr?



1. Consider the following graphs that show the relationship between elevation and snowfall. For January, does elevation have much impact below 6,000 feet? What can be said about Point A in this plot? Consider the plot that includes data from March? Is there much of a difference in the relationship between elevation and snowfall amounts between January and March? Discuss.

|  |  |
| --- | --- |
|  |  |