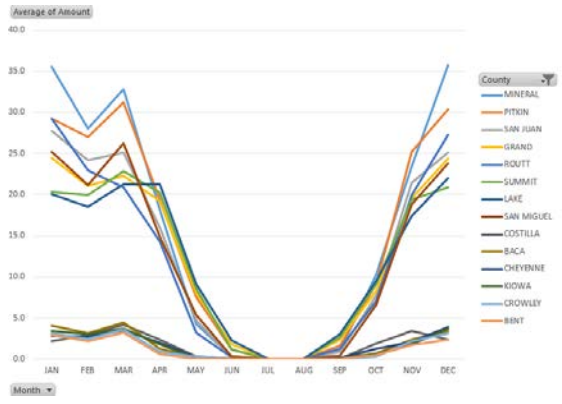
APPEND AND MERGE

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.

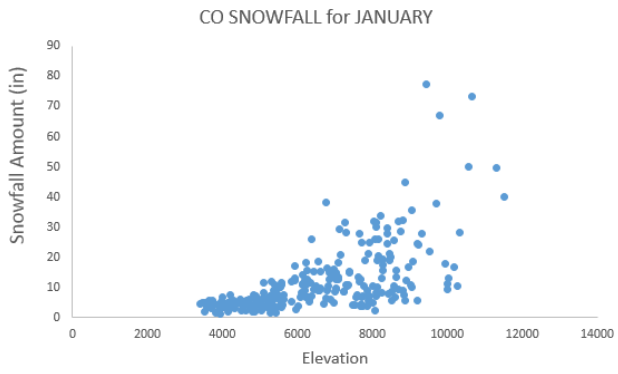
The National Oceanic and Atmospheric Administration (NOAA) provides a lot of data that is freely available. Several regional weather offices also publish data. Data from the Western Regional Climate Center (Website: http://www.wrcc.dri.edu/) is used in this handout.

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

	A	B	C	D	E	F
1	RowID	Station	Elevation	Jan	Feb	Mar
2	0	BOULDER	5404	10.9	11	17.8
3	1	DENVER WSFO AP	5325	7.3	7	12.2
4	2	FOUNTAIN	5565	3.3	3	5
5	3	GUNNISON 1 N	7680	12.5	9.5	6.5
6	4	POWDERHORN	8094	7.2	6.9	5

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

	A	B	C	D	E
1	RowID	Station	Elevation	Month	Amount
2	0	BOULDER	5404	Jan	10.9
3	1	DENVER WSFO AP	5325	Jan	7.3
4	2	FOUNTAIN	5565	Jan	3.3
5	3	GUNNISON 1 N	7680	Jan	12.5
6	4	POWDERHORN	8094	Jan	7.2
7	5	BOULDER	5404	Feb	11
8	6	DENVER WSFO AP	5325	Feb	7
9	7	FOUNTAIN	5565	Feb	3
10	8	GUNNISON 1 N	7680	Feb	9.5
11	9	POWDERHORN	8094	Feb	6.9
12	10	BOULDER	5404	Mar	17.8
13	11	DENVER WSFO AP	5325	Mar	12.2
14	12	FOUNTAIN	5565	Mar	5
15	13	GUNNISON 1 N	7680	Mar	6.5
16	14	POWDERHORN	8094	Mar	5

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

	A	B	C	D	E	F
1	RowID	Station	Elevation	Jan	Feb	Mar
2	0	BOULDER	5404	10.9	11	17.8
3	1	DENVER WSFO AP	5325	7.3	7	12.2
4	2	FOUNTAIN	5565	3.3	3	5
5	3	GUNNISON 1 N	7680	12.5	9.5	6.5
6	4	POWDERHORN	8094	7.2	6.9	5
7	5	BOULDER	5404			
8	6	DENVER WSFO AP	5325			
9	7	FOUNTAIN	5565			
10	8	GUNNISON 1 N	7680			
11	9	POWDERHORN	8094			
12	10	BOULDER	5404			
13	11	DENVER WSFO AP	5325			
14	12	FOUNTAIN	5565			
15	13	GUNNISON 1 N	7680			
16	14	POWDERHORN	8094			

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.

$$(5\ rows * \ 3\ columns\ to\ stacked) = 15$$

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

	A	B	C	D	
1	RowID	Row Reference	Station	Elevation	Ja
2	0	= MOD(A2 , 5)	BOULDER	5404	1
3	1	1	DENVER WSFO AP	5325	7
4	2	2	FOUNTAIN	5565	3
5	3	3	GUNNISON 1 N	7680	1
6	4	4	POWDERHORN	8094	7
7	5	0			
8	6	1			
9	7	2			
10	8	3			
11	9	4			
12	10	0			
13	11	1			
14	12	2			
15	13	3			
16	14	4			

The Row Reference provides the needed reference for the replicates.

	A	B	C	D	
1	RowID	Row Reference	Station	Elevation	Ja
2	0	0	BOULDER	5404	1
3	1	1	DENVER WSFO AP	5325	7.
4	2	2	FOUNTAIN	5565	3.
5	3	3	GUNNISON 1 N	7680	12
6	4	4	POWDERHORN	8094	7.
7	5	0			
8	6	1			
9	7	2			
10	8	3			
11	9	4			
12	10	0			
13	11	1			
14	12	2			
15	13	3			
16	14	4			

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.

	A	B	C	D
1	RowID	Row Reference	Station	Elevation
2	0	0	BOULDER	5404
3	1	1	DENVER WSFO AP	5325
4	2	2	FOUNTAIN	5565
5	3	3	GUNNISON 1 N	7680
6	4	4	POWDERHORN	8094
7	5	0	= OFFSET($C$2, B7, 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.

	A		C	D
1	RowID	Row Reference		
2	0	0	BOULDER	5404
3	1	1	DENVER WSFO AP	5325
4	2	2	FOUNTAIN	5565
5	3	3	GUNNISON 1 N	7680
6	4	4	POWDERHORN	8094
7	5	0		

0 **1**

0 **1** **2** **3** **4**

Cell C7: =OFFSET($C$2,B7,0)

Copy this down for remaining cell. The Station names are replicated as needed.

	A	B	C	D
1	RowID	Row Reference	Station	Elevation
2	0	0	BOULDER	5404
3	1	1	DENVER WSFO AP	5325
4	2	2	FOUNTAIN	5565
5	3	3	GUNNISON 1 N	7680
6	4	4	POWDERHORN	8094
7	5	0	BOULDER	
8	6	1	DENVER WSFO AP	
9	7	2	FOUNTAIN	
10	8	3	GUNNISON 1 N	
11	9	4	POWDERHORN	
12	10	0	BOULDER	
13	11	1	DENVER WSFO AP	
14	12	2	FOUNTAIN	
15	13	3	GUNNISON 1 N	
16	14	4	POWDERHORN	

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)

	A	B	C	D
1	RowID	Row Reference	Station	Elevation
2	0	0	BOULDER	5404
3	1	1	DENVER WSFO AP	5325
4	2	2	FOUNTAIN	5565
5	3	3	GUNNISON 1 N	7680
6	4	4	POWDERHORN	8094
7	5	0	BOULDER	= OFFSET($C$2, B7, 1)

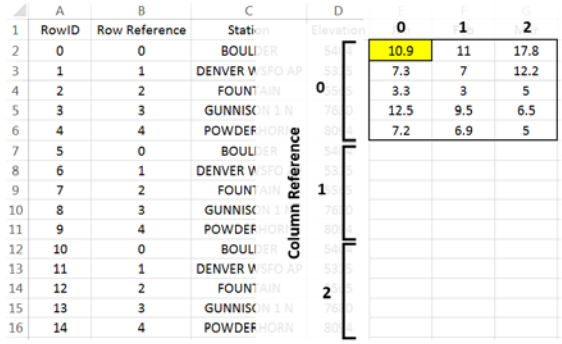
Elevation should now be replicated as shown here

	A	B	C	D
1	RowID	Row Reference	Station	Elevation
2	0	0	BOULDER	5404
3	1	1	DENVER WSFO AP	5325
4	2	2	FOUNTAIN	5565
5	3	3	GUNNISON 1 N	7680
6	4	4	POWDERHORN	8094
7	5	0	BOULDER	5404
8	6	1	DENVER WSFO AP	5325
9	7	2	FOUNTAIN	5565
10	8	3	GUNNISON 1 N	7680
11	9	4	POWDERHORN	8094
12	10	0	BOULDER	5404
13	11	1	DENVER WSFO AP	5325
14	12	2	FOUNTAIN	5565
15	13	3	GUNNISON 1 N	7680
16	14	4	POWDERHORN	8094

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.

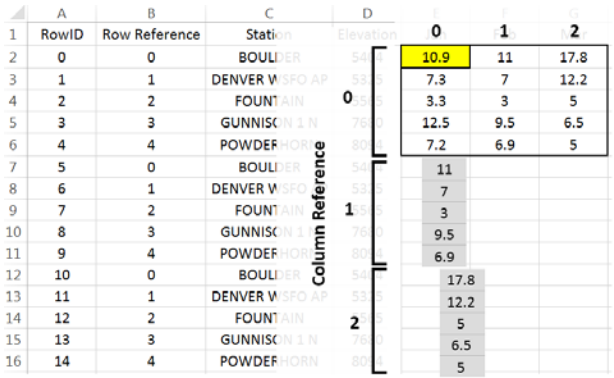
Cell F2 will be used as the reference cell



	A	B	C	D	F	G	H
1	RowID	Row Reference	Station	Elevation	0	1	2
2	0	0	BOULDER	5427	10.9	11	17.8
3	1	1	DENVER WSFO AP	5325	7.3	7	12.2
4	2	2	FOUNTAIN	5445	3.3	3	5
5	3	3	GUNNISON 1 N	7680	12.5	9.5	6.5
6	4	4	POWDERHORN	8014	7.2	6.9	5
7	5	0	BOULDER	5427			
8	6	1	DENVER WSFO AP	5325			
9	7	2	FOUNTAIN	5445			
10	8	3	GUNNISON 1 N	7680			
11	9	4	POWDERHORN	8014			
12	10	0	BOULDER	5427			
13	11	1	DENVER WSFO AP	5325			
14	12	2	FOUNTAIN	5445			
15	13	3	GUNNISON 1 N	7680			
16	14	4	POWDERHORN	8014			

Column Reference: 0, 1, 2

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.



	A	B	C	D	F	G	H
1	RowID	Row Reference	Station	Elevation	0	1	2
2	0	0	BOULDER	5427	10.9	11	17.8
3	1	1	DENVER WSFO AP	5325	7.3	7	12.2
4	2	2	FOUNTAIN	5445	3.3	3	5
5	3	3	GUNNISON 1 N	7680	12.5	9.5	6.5
6	4	4	POWDERHORN	8014	7.2	6.9	5
7	5	0	BOULDER	5427	11		
8	6	1	DENVER WSFO AP	5325	7		
9	7	2	FOUNTAIN	5445	3		
10	8	3	GUNNISON 1 N	7680	9.5		
11	9	4	POWDERHORN	8014	6.9		
12	10	0	BOULDER	5427	17.8		
13	11	1	DENVER WSFO AP	5325	12.2		
14	12	2	FOUNTAIN	5445	5		
15	13	3	GUNNISON 1 N	7680	6.5		
16	14	4	POWDERHORN	8014	5		

Column Reference: 0, 1, 2

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

Cell C2: =INT(A2 / 5)

	A	B	C	D
1	RowID	Row Reference	Column Reference	Station
2	0	0	=INT(A2 / 5)	BOULDER
3	1	1	0	DENVER WSFO AP
4	2	2	0	FOUNTAIN
5	3	3	0	GUNNISON 1 N
6	4	4	0	POWDERHORN
7	5	0	1	BOULDER
8	6	1	1	DENVER WSFO AP
9	7	2	1	FOUNTAIN

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

$$\left\lfloor \frac{1}{5} \right\rfloor = \lfloor 0.2 \rfloor = 0 \qquad \left\lfloor \frac{6}{5} \right\rfloor = \lfloor 1.2 \rfloor = 1$$

$$\left\lfloor \frac{4}{5} \right\rfloor = \lfloor 0.8 \rfloor = 0 \qquad \left\lfloor \frac{14}{5} \right\rfloor = \lfloor 2.8 \rfloor = 2$$

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)

	A	B	C	D	E	F	G	H
1	RowID	Row Reference	Column Reference	Station	Elevation	Jan	Feb	M
2	0	0	0	BOULDER	5404	10.9	11	17
3	1	1	0	DENVER WSFO AP	5325	7.3	7	12
4	2	2	0	FOUNTAIN	5565	3.3	3	5
5	3	3	0	GUNNISON 1 N	7680	12.5	9.5	6.
6	4	4	0	POWDERHORN	8094	7.2	6.9	5
7	5	0	1	BOULDER	5404	= OFFSET($F$2 , B7 , C7)		
8	6	1	1	DENVER WSFO AP	5325			

The snowfall amounts should now be stacked.

	A	B	C	D	E	F	G	H
1	RowID	Row Reference	Column Reference	Station	Elevation	Jan	Feb	Mar
2	0	0	0	BOULDER	5404	10.9	11	17.8
3	1	1	0	DENVER WSFO AP	5325	7.3	7	12.2
4	2	2	0	FOUNTAIN	5565	3.3	3	5
5	3	3	0	GUNNISON 1 N	7680	12.5	9.5	6.5
6	4	4	0	POWDERHORN	8094	7.2	6.9	5
7	5	0	1	BOULDER	5404	11		
8	6	1	1	DENVER WSFO AP	5325	7		
9	7	2	1	FOUNTAIN	5565	3		
10	8	3	1	GUNNISON 1 N	7680	9.5		
11	9	4	1	POWDERHORN	8094	6.9		
12	10	0	2	BOULDER	5404	17.8		
13	11	1	2	DENVER WSFO AP	5325	12.2		
14	12	2	2	FOUNTAIN	5565	5		
15	13	3	2	GUNNISON 1 N	7680	6.5		
16	14	4	2	POWDERHORN	8094	5		

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	B	C	D	E	F	G	H	I
1	RowID	Row Reference	Column Reference	Station	Elevation	Month	Jan	Feb	Mar
2	0	0	0	BOULDER	5404	= OFFSET($G$1 , 0 , C2)	10.9	11	17.8
3	1	1	0	DENVER WSFO AP	5325		7.3	7	12.2
4	2	2	0	FOUNTAIN	5565		3.3	3	5

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.

	A	B	C	D	E
1	RowID	Station	Elevation	Month	Amount
2	0	BOULDER	5404	Jan	10.9
3	1	DENVER WSFO AP	5325	Jan	7.3
4	2	FOUNTAIN	5565	Jan	3.3
5	3	GUNNISON 1 N	7680	Jan	12.5
6	4	POWDERHORN	8094	Jan	7.2
7	5	BOULDER	5404	Feb	11
8	6	DENVER WSFO AP	5325	Feb	7
9	7	FOUNTAIN	5565	Feb	3
10	8	GUNNISON 1 N	7680	Feb	9.5
11	9	POWDERHORN	8094	Feb	6.9
12	10	BOULDER	5404	Mar	17.8
13	11	DENVER WSFO AP	5325	Mar	12.2
14	12	FOUNTAIN	5565	Mar	5
15	13	GUNNISON 1 N	7680	Mar	6.5
16	14	POWDERHORN	8094	Mar	5

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.

COLORADO

MONTHLY AVERAGE SNOWFALL (INCHES)

	PERIOD OF RECORD	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	YEAR
AGUILAR 1 SE	1980-2005	11.4	13.9	19.0	13.2	3.0	0.0	0.0	0.0	0.2	2.9	14.5	14.8	92.9
AGUILAR 18 WSW	1998-2010	15.7	11.8	22.1	19.4	4.8	0.0	0.0	0.0	0.3	6.0	8.9	16.4	105.4
AKRON 4 E	1948-2010	4.3	4.1	5.5	3.8	0.2	0.0	0.0	0.0	0.2	1.4	5.4	5.9	30.8
AKRON 1 N	1948-1999	5.6	4.6	9.5	4.4	0.8	0.0	0.0	0.0	0.6	2.8	6.3	5.8	40.4
ALAMOSA WSO AP	1948-2010	4.3	4.0	5.6	4.0	1.4	0.0	0.0	0.0	0.1	2.7	3.8	5.3	31.2

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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Station ID	R	ECORD PERI	OD	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	YEAR
2	AGUILAR 1 SE	|	1980-2005	|	11.4	13.9	19	13.2	3	0	0	0	0.2	2.9	14.5	14.8	92.9
3	AGUILAR 18 WSW	|	1998-2010	|	15.7	11.8	22.1	19.4	4.8	0	0	0	0.3	6	8.9	16.4	105.4
4	AKRON 4 E	|	1948-2010	|	4.3	4.1	5.5	3.8	0.2	0	0	0	0.2	1.4	5.4	5.9	30.8
5	AKRON 1 N	|	1948-1999	|	5.6	4.6	9.5	4.4	0.8	0	0	0	0.6	2.8	6.3	5.8	40.4
6	ALAMOSA WSO AP	|	1948-2010	|	4.3	4	5.6	4	1.4	0	0	0	0.1	2.7	3.8	5.3	31.2

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

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Station ID	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
2	AGUILAR 1 SE	11.4	13.9	19	13.2	3	0	0	0	0.2	2.9	14.5	14.8
3	AGUILAR 18 WSW	15.7	11.8	22.1	19.4	4.8	0	0	0	0.3	6	8.9	16.4
4	AKRON 4 E	4.3	4.1	5.5	3.8	0.2	0	0	0	0.2	1.4	5.4	5.9
5	AKRON 1 N	5.6	4.6	9.5	4.4	0.8	0	0	0	0.6	2.8	6.3	5.8
6	ALAMOSA WSO AP	4.3	4	5.6	4	1.4	0	0	0	0.1	2.7	3.8	5.3

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.

$$(326\ rows * 12\ columns) = 3912$$

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

	A	B	C	D	E
1	RowID	Row Reference	Column Reference	Station ID	JAI
2	0	= MOD(A2 , 326)	= INT(A2 / 326)	AGUILAR 1 SE	1
3	1	1	0	AGUILAR 18 WSW	1
4	2			AKRON 4 E	
5	3			AKRON 1 N	
6	4			ALAMOSA WSO AP	

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.

	A	B	C	D	E
1	RowID	Row Referen	Column Referen	Station ID	JAN
2	0	0	0	AGUILAR 1 SE	11.4
3	1	1	0	AGUILAR 18 WSW	15.7
4	2	2	0	AKRON 4 E	4.3
5	3	3	0	AKRON 1 N	5.6
327	325	325	0	YUMA 10 NW	4.6
328	326	0	1		
329	327	1	1		
653	651	325	1		
654	652	0	2		
655	653	1	2		
3912	3910	324	11		
3913	3911	325	11		

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)

	A	B	C	D	E	F	G	
1	RowID	Row Reference	Column Reference	Station ID	JAN	FEB	MAR	A
2	0	0	0	AGUILAR 1 SE	11.4	13.9	19	
3	1	1	0	AGUILAR 18 WSW	15.7	11.8	22.1	
4	2	2	0	AKRON 4 E	4.3	4.1	5.5	
326								
327	325	325	0	YUMA 10 NW	4.6	3.9	4.2	
328	326	0	1	= OFFSET($D$2, B328, 0)	=OFFSET($E$2 , B328 , C328)			
329	327	1	1					

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)

	A	B	C	D	E	F
1	RowID	Row Reference	Column Reference	Station ID	Month	JAN
2	0	0	0	AGUILAR 1 SE	= OFFSET($F$1 , 0 , C2)	11.4
3	1	1	0	AGUILAR 18 WSW		15.7

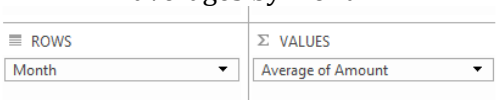
Verify that all columns have been properly stacked and the content of all rows has been correctly specified. Obtain a copy of the this data using Paste Values to remove all cell dependencies. A snippet of the final dataset is provided here for reference.

	A	B	C	D
1	RowID	Station ID	Month	Amount
2	0	AGUILAR 1 SE	JAN	11.4
3	1	AGUILAR 18 WSW	JAN	15.7
4	2	AKRON 4 E	JAN	4.3
326	324	YUMA	JAN	5
327	325	YUMA 10 NW	JAN	4.6
328	326	AGUILAR 1 SE	FEB	13.9
329	327	AGUILAR 18 WSW	FEB	11.8
652	650	YUMA	FEB	3.6
653	651	YUMA 10 NW	FEB	3.9
654	652	AGUILAR 1 SE	MAR	19
655	653	AGUILAR 18 WSW	MAR	22.1
3912	3910	YUMA	DEC	3.9
3913	3911	YUMA 10 NW	DEC	5.3

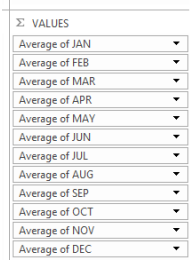
Getting Summaries Stacked vs. Unstacked

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

Stacked data: getting averages by Month

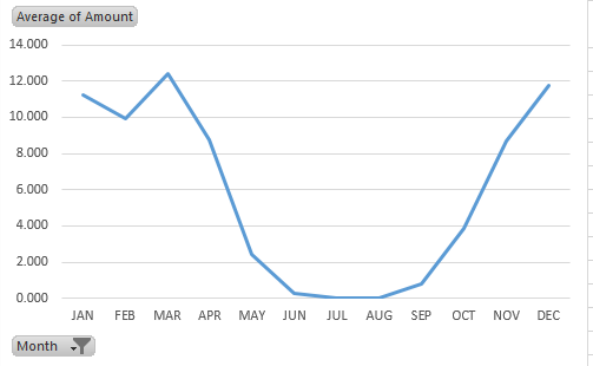


Unstacked data: getting averages by Month



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

Month	Average of Amount
JAN	11.262
FEB	9.920
MAR	12.381
APR	8.749
MAY	2.456
JUN	0.260
JUL	0.003
AUG	0.002
SEP	0.797
OCT	3.833
NOV	8.658
DEC	11.769
Grand Total	5.841



Questions

2. Which month has the highest average snowfall?

3. 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

RowID	Station	Elevation	Month	Amount
0	BOULDER	5404	Jan	10.9
1	DENVER WSFO AP	5325	Jan	7.3
2	FOUNTAIN	5565	Jan	3.3
3	GUNNISON 1 N	7680	Jan	12.5
4	POWDERHORN	8094	Jan	7.2
5	BOULDER	5404	Feb	11
6	DENVER WSFO AP	5325	Feb	7
7	FOUNTAIN	5565	Feb	3
8	GUNNISON 1 N	7680	Feb	9.5
9	POWDERHORN	8094	Feb	6.9
10	BOULDER	5404	Mar	17.8
11	DENVER WSFO AP	5325	Mar	12.2
14	FOUNTAIN	5565	Mar	5
15	GUNNISON 1 N	7680	Mar	6.5
16	POWDERHORN	8094	Mar	5

TABLE A

Auxiliary information for weather stations in CO

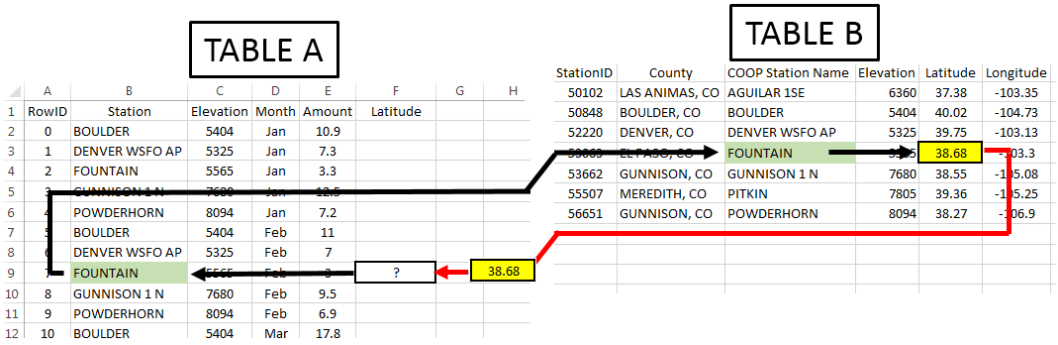
StationID	County	COOP Station Name	Elevation	Latitude	Longitude
50102	LAS ANIMAS, CO	AGUILAR 1SE	6360	37.38	-103.35
50848	BOULDER, CO	BOULDER	5404	40.02	-104.73
52220	DENVER, CO	DENVER WSFO AP	5325	39.75	-103.13
53063	EL PASO, CO	FOUNTAIN	5565	38.68	-103.3
53662	GUNNISON, CO	GUNNISON 1 N	7680	38.55	-105.08
55507	MEREDITH, CO	PITKIN	7805	39.36	-105.25
56651	GUNNISON, CO	POWDERHORN	8094	38.27	-106.9

TABLE B

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

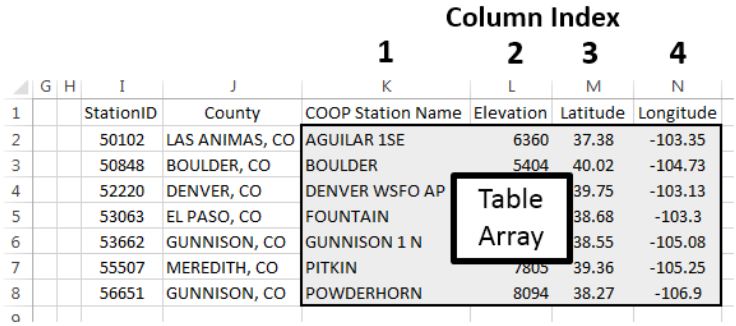
- First argument: specifies the value to be looked up
- Second argument: specifies the cell range of Table B or a subset of Table B
- Third arugment: an index that specifies which column is used to retrieve content from Table B
- Fourth argument: This should be set to FALSE. FALSE forces =VLOOKUP() to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



	G	H	I	J	K	L	M	N
1			StationID	County	COOP Station Name	Elevation	Latitude	Longitude
2			50102	LAS ANIMAS, CO	AGUILAR 1SE	6360	37.38	-103.35
3			50848	BOULDER, CO	BOULDER	5404	40.02	-104.73
4			52220	DENVER, CO	DENVER WSFO AP		39.75	-103.13
5			53063	EL PASO, CO	FOUNTAIN		38.68	-103.3
6			53662	GUNNISON, CO	GUNNISON 1 N		38.55	-105.08
7			55507	MEREDITH, CO	PITKIN	7805	39.36	-105.25
8			56651	GUNNISON, CO	POWDERHORN	8094	38.27	-106.9

Type the following formula into cell F2.

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

	A	B	C	D	E	F
1	RowID	Station	Elevation	Month	Amount	Latitude
2	0	BOULDER	5404	Jan	10.9	= VLOOKUP(B2, $K$2:$N$8, 3, FALSE)
3	1	DENVER WSFO AP	5325	Jan	7.3	
4	2	FOUNTAIN	5565	Jan	3.3	

Copy this formula down for the remaining cells.

	A	B	C	D	E	F	G
1	RowID	Station	Elevation	Month	Amount	Latitude	Longitude
2	0	BOULDER	5404	Jan	10.9	40.02	= VLOOKUP(B2, $K$2:$N$8, 4, FALSE)
3	1	DENVER WSFO AP	5325	Jan	7.3	39.75	-103.13
4	2	FOUNTAIN	5565	Jan	3.3	38.68	-103.3
5	3	GUNNISON 1 N	7680	Jan	12.5	38.55	-105.08

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)

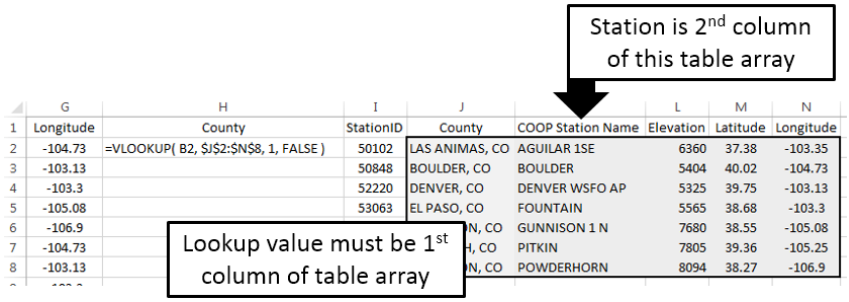
	A	B	C	D	E	F	G
1	RowID	Station	Elevation	Month	Amount	Latitude	Longitude
2	0	BOULDER	5404	Jan	10.9	40.02	-104.73
3	1	DENVER WSFO AP	5325	Jan	7.3	39.75	-103.13
4	2	FOUNTAIN	5565	Jan	3.3	38.68	-103.3
5	3	GUNNISON 1 N	7680	Jan	12.5	38.55	-105.08
6	4	POWDERHORN	8094	Jan	7.2	38.27	-106.9
7	5	BOULDER	5404	Feb	11	40.02	-104.73
8	6	DENVER WSFO AP	5325	Feb	7	39.75	-103.13
9	7	FOUNTAIN	5565	Feb	3	38.68	-103.3
10	8	GUNNISON 1 N	7680	Feb	9.5	38.55	-105.08
11	9	POWDERHORN	8094	Feb	6.9	38.27	-106.9
12	10	BOULDER	5404	Mar	17.8	40.02	-104.73
13	11	DENVER WSFO AP	5325	Mar	12.2	39.75	-103.13
14	12	FOUNTAIN	5565	Mar	5	38.68	-103.3
15	13	GUNNISON 1 N	7680	Mar	6.5	38.55	-105.08
16	14	POWDERHORN	8094	Mar	5	38.27	-106.9

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)

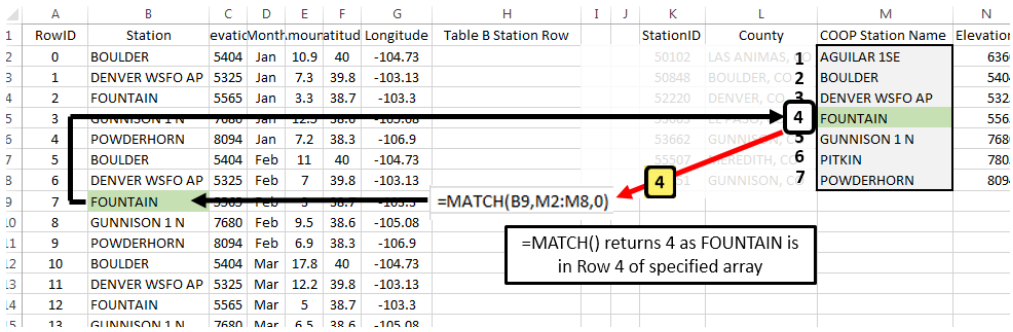
The column containing the lookup value must be the left-most column of the table array.



<u>Using =MATCH() and =INDEX() to Merge Tables</u>

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.

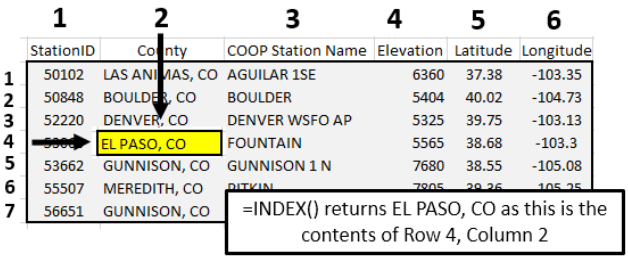


The second step of this process is to retrieve the actual contents from Table B. This is done using the =INDEX() function. The =INDEX() function is similar to the =OFFSET() function used earlier. =OFFSET() required the specification of a single cell to be used as a reference. INDEX() requires specification 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

	J	K	L	M	N	O	P
1		StationID	County	COOP Station Name	Elevation	Latitude	Longitude
2		50102	LAS ANIMAS, CO	AGUILAR 1SE	6360	37.38	-103.35
3		50848	BOULDER, CO	BOULDER	5404	40.02	-104.73
4		52220	DENVER, CO	DENVER WSFO AP	5325	39.75	-103.13
5		53063	EL PASO, CO	FOUNTAIN	5565	38.68	-103.3
6		53662	GUNNISON, CO	GUNNISON 1 N	7680	38.55	-105.08
7		55507	MEREDITH, CO	PITKIN	7805	39.36	-105.25
8		56651	GUNNISON, CO	POWDERHORN	8094	38.27	-106.9

=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)

	A	B	C	D	E	F	G	H	I
1	RowID	Station	evatic	Month	mour	atitud	Longitude	Table B Station Row	County
2	0	BOULDER	5404	Jan	10.9	40	-104.73	=MATCH(B2, $M$2 : $M$8, 0)	=INDEX($K$2 : $P$8, H2, 2)
3	1	DENVER WSFO AP	5325	Jan	7.3	39.8	-103.13		
4	2	FOUNTAIN	5565	Jan	3.3	38.7	-103.3		

Questions

4. What is the purpose of the third argument in the =MATCH() function?

5. 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.

	A	B	C	D	E	F	G	H
1	RowID	Station	Elevation	Month	Amount	Latitude	Longitude	County
2	0	BOULDER	5404	Jan	10.9	40.02	-104.73	BOULDER, CO
3	1	DENVER WSFO AP	5325	Jan	7.3	39.75	-103.13	DENVER, CO
4	2	FOUNTAIN	5565	Jan	3.3	38.68	-103.3	EL PASO, CO
5	3	GUNNISON 1 N	7680	Jan	12.5	38.55	-105.08	GUNNISON, CO
6	4	POWDERHORN	8094	Jan	7.2	38.27	-106.9	GUNNISON, CO
7	5	BOULDER	5404	Feb	11	40.02	-104.73	BOULDER, CO
8	6	DENVER WSFO AP	5325	Feb	7	39.75	-103.13	DENVER, CO
9	7	FOUNTAIN	5565	Feb	3	38.68	-103.3	EL PASO, CO
10	8	GUNNISON 1 N	7680	Feb	9.5	38.55	-105.08	GUNNISON, CO
11	9	POWDERHORN	8094	Feb	6.9	38.27	-106.9	GUNNISON, CO
12	10	BOULDER	5404	Mar	17.8	40.02	-104.73	BOULDER, CO
13	11	DENVER WSFO AP	5325	Mar	12.2	39.75	-103.13	DENVER, CO
14	12	FOUNTAIN	5565	Mar	5	38.68	-103.3	EL PASO, CO
15	13	GUNNISON 1 N	7680	Mar	6.5	38.55	-105.08	GUNNISON, CO
16	14	POWDERHORN	8094	Mar	5	38.27	-106.9	GUNNISON, CO

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.

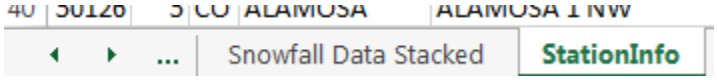
Average Snowfall	Month			
County	Jan	Feb	Mar	Grand Total
BOULDER, CO	10.9	11.0	17.8	13.2
DENVER, CO	7.3	7.0	12.2	8.8
EL PASO, CO	3.3	3.0	5.0	3.8
GUNNISON, CO	9.9	8.2	5.8	7.9
Grand Total	**8.2**	**7.5**	**9.3**	**8.3**

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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	NUM	DIV	ST	COUNTY	COOP STATION NAME	BEGINS	ENDS	LA	TIT	UDE	LON	GIT	UDE	ELEV
2	50028		CO	WASHINGTON	ABBOTT	18900101	18950331	39	52	0	-103	30	0	4800
3	50092	4	CO	ELBERT	AGATE 3 SW	19480801	19530430	39	27	0	-103	56	0	5482
4	50096	4	CO	JACKSON	AGUA FRIA	19660101	19761231	40	38	0	-106	38	0	10407
5	50102	1	CO	LAS ANIMAS	AGUILAR 1 SE	19800108	19880914	37	23	0	-104	39	0	6360
6	50102	1	CO	LAS ANIMAS	AGUILAR 1 SE	19880914	19961001	37	23	0	-104	39	0	6360
7	50102	1	CO	LAS ANIMAS	AGUILAR	19970506	20040922	37	24	4	-104	39	17	6400
8	50102	1	CO	LAS ANIMAS	AGUILAR	20040922	20060525	37	24	4	-104	39	17	6400

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)

	A	B	C	D	E	F	G	H
1	RowID	Station ID	Month	Amount	Station Row			
2	0	AGUILAR 1 SE	JAN	11.4	=MATCH(B2 , StationInfo!$E$2 : $E$3091 , 0)			
3	1	AGUILAR 18 WSW	JAN	15.7				
4	2	AKRON 4 E	JAN	4.3				

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))

	A	B	C	D	E	F	G	H	I	J	K
1	RowI	Station ID	Mon	Amou	Station Row	County					
2	0	AGUILAR 1 SE	JAN	11.4	4	=IF(ISERROR(E2) , "" , INDEX(StationInfo!$A$2:$N$3091 , E2 , 4))					
3	1	AGUILAR 18 WSW	JAN	15.7	8						
4	2	AKRON 4 E	JAN	4.3	13						
5	3	AKRON 1 N	JAN	5.6	29						

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

	A	B	C	D	E	F
1	RowID	Station ID	Month	Amount	Station Row	County
2	0	AGUILAR 1 SE	JAN	11.4	4	LAS ANIMAS
3	1	AGUILAR 18 WSW	JAN	15.7	8	LAS ANIMAS
4	2	AKRON 4 E	JAN	4.3	13	WASHINGTON
5	3	AKRON 1 N	JAN	5.6	29	WASHINGTON
6	4	ALAMOSA WSO AP	JAN	4.3	44	ALAMOSA
7	5	ALLENSPARK LODGE	JAN	19.1	62	BOULDER
8	6	ALLENSPARK 1 NW	JAN	21.3	63	BOULDER

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

6. What is the purpose of the empty string, i.e. "", in the formula above?

7. 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)

8. 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"))

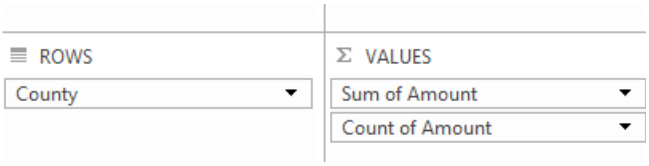
Summaries using Merged Content

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

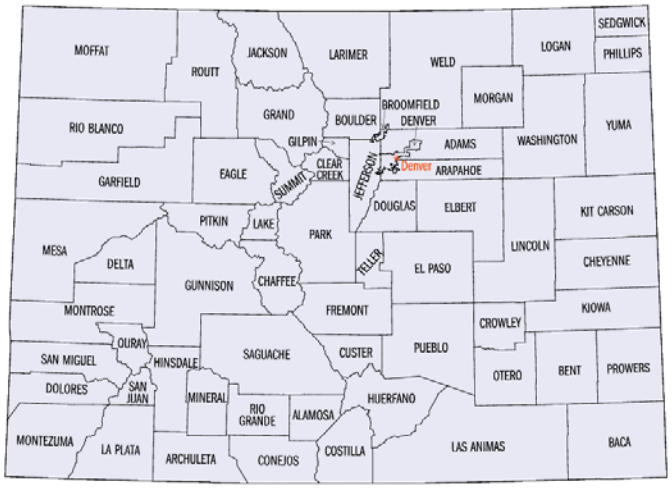
Snowfall by County

Row Labels	Sum of Amount	Count of Amount
GRAND	1212.4	96
MINERAL	947.5	60
PITKIN	909.4	60
EL PASO	828.9	120
JEFFERSON	792.7	108
LARIMER	789.3	120
LAS ANIMAS	717.5	120
SAN MIGUEL	713.7	60
PROWERS	47.3	24
PHILLIPS	32.8	12
BENT	27	24
COSTILLA	19.7	12
CROWLEY	16	12
Grand Total	**19568.9**	**3360**

The PivotTable structure used to create this table



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



Questions

9. 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?

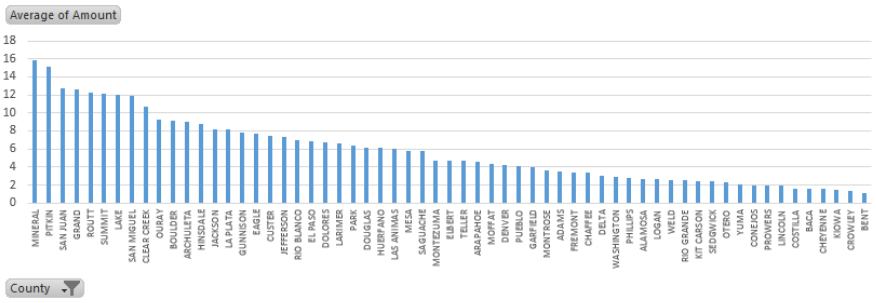
10. The SUM is being used here as the total snowfall over the entire year is of interest. I'd argue that a SUM should not be used as the number of stations per county is not the same. Do you agree or disagree? Explain.

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

Average Snowfall by County

County	Average Snowfall
MINERAL	15.8
PITKIN	15.2
SAN JUAN	12.7
GRAND	12.6
ROUTT	12.2
SUMMIT	12.1
LAKE	12.1
SAN MIGUEL	11.9
BACA	1.6
CHEYENNE	1.5
KIOWA	1.5
CROWLEY	1.3
BENT	1.1
Grand Total	**5.8**

Pareato-type chart of average snowfall amounts



Questions

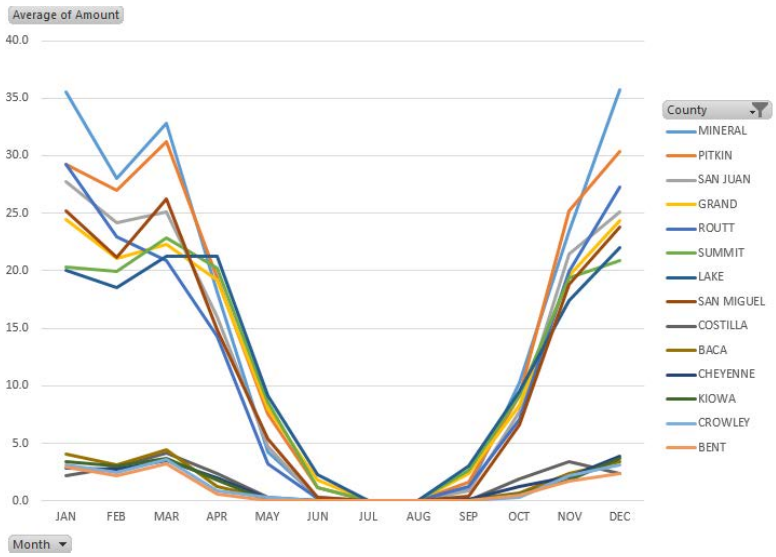
11. The average snowfall for Mineral County is 15.8 inches. Provide an interpretation for this value.

12. 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.

13. The following table shows the snowfall by month across counties in CO. Recreate this table in Excel.

Average of Amount	Colu												
Row Labels	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	Grand Total
MINERAL	35.5	28.0	32.8	18.1	4.2	0.3	0.0	0.0	0.9	10.3	23.5	35.7	15.8
PITKIN	29.2	27.0	31.2	19.4	7.6	1.1	0.0	0.0	1.7	9.2	25.2	30.4	15.2
SAN JUAN	27.7	24.2	25.1	16.0	4.7	0.1	0.0	0.0	0.9	7.7	21.4	25.1	12.7
GRAND	24.5	21.0	22.3	19.2	8.1	1.8	0.0	0.0	2.4	8.3	19.6	24.3	12.6
ROUTT	29.2	23.0	20.9	14.3	3.3	0.2	0.0	0.0	1.3	7.1	19.9	27.3	12.2
SUMMIT	20.3	20.0	22.9	20.2	8.6	1.2	0.0	0.0	2.7	9.1	19.4	20.9	12.1
LAKE	20.1	18.5	21.3	21.3	9.2	2.3	0.1	0.0	3.0	9.5	17.4	22.0	12.1
SAN MIGUEL	25.2	21.2	26.3	14.8	5.4	0.3	0.0	0.0	0.4	6.6	18.8	23.8	11.9
COSTILLA	2.2	2.9	4.2	2.4	0.3	0.0	0.0	0.0	0.0	1.9	3.4	2.4	1.6
BACA	4.1	3.1	4.4	1.2	0.2	0.0	0.0	0.0	0.1	0.7	2.4	3.4	1.6
CHEYENNE	2.8	2.7	3.5	2.0	0.2	0.0	0.0	0.0	0.1	1.2	2.0	3.8	1.5
KIOWA	3.4	3.0	3.7	1.8	0.1	0.0	0.0	0.0	0.0	0.6	1.8	3.7	1.5
CROWLEY	3.1	2.5	3.6	0.9	0.3	0.0	0.0	0.0	0.0	0.3	2.2	3.1	1.3
BENT	3.0	2.2	3.2	0.6	0.1	0.0	0.0	0.0	0.0	0.5	1.8	2.4	1.1
Grand Total	**11.3**	**9.9**	**12.4**	**8.6**	**2.3**	**0.2**	**0.0**	**0.0**	**0.8**	**3.8**	**8.6**	**11.9**	**5.8**

14. 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?



15. 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.