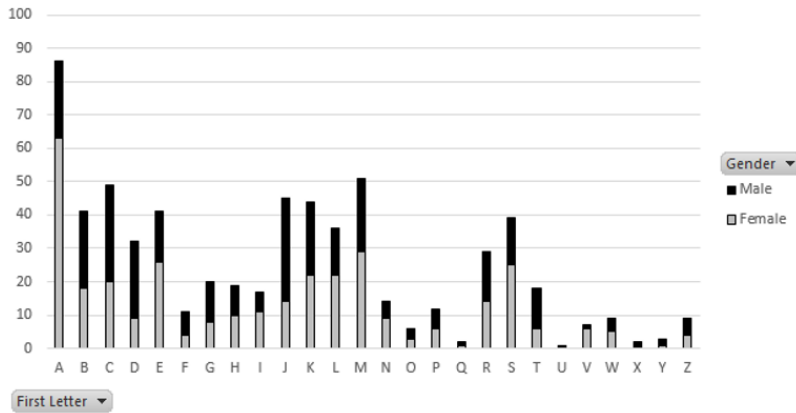


## WORKING WITH STRINGS & PIVOTTABLES

Charlotte is the name given to the most recent addition to the family of Prince William and Kate Middleton. As a result, other parents are likely to name their daughter Charlotte as well. Baby names follow certain patterns over time.



Consider the following graph that compares the first letter of a baby's name across gender. We can see that A is the most popular first letter for girls, but this is not the case for boys. Consider the fact that for the data being investigated here, a first letter of Z is almost as popular as F.



My name is Chris and my dad's name is Greg. My sister's name is Ann. Simple names with uncomplicated spellings. Are names in 2014 longer in length? Do baby names today contain a higher proportion of vowels than other generations?

Summary Measure	Year	
	2014	1900
Average	0.417	0.414
Median	0.40	0.40
Std Dev	0.127	0.133
Minimum	0	0
Maximum	0.75	0.80

This handout will cover two tasks that will involve the manipulation of strings in Excel. These tasks include obtaining the first letter for each baby name. The second task will be to compute the proportion of vowels for each baby name.

### Procedural Steps

1. Obtain the first letter for each baby name by sub-setting a string
2. Use the PivotTable feature in Excel to obtain summaries and visualizations
3. Develop a process to count the number of vowels in a string

### Data Technologies

4. String functions in Excel
5. Summaries and Visualizations through PivotTables

Data Source	
Address	<a href="http://course1.winona.edu/cmalone/workshops/uscots2015/">http://course1.winona.edu/cmalone/workshops/uscots2015/</a>
Description	<p>BabyNames Dataset</p> <p>This dataset contains the unique names of all babies born at the Olmstead Medical Center in 2014. These names are published in the Rochester Post Bulletin's Mother's Day Weekend Edition</p>

Open the BabyNames dataset in Excel. Convert this dataset to an Excel Table. This data contains a total of 643 unique names – 336 unique names for girls and 307 for boys.

Data in Excel

	A	B	C	D
1	RowID	Year	Gender	Name
2	1	2014	Female	Abel
3	2	2014	Female	Abigail
4	3	2014	Female	Adalyn
5	4	2014	Female	Addison
6	5	2014	Female	Adeline
7	6	2014	Female	Adelynn
8	7	2014	Female	Adelynnne
9	8	2014	Female	Adison
10	9	2014	Female	Adyson
11	10	2014	Female	Ahlam
12	11	2014	Female	Alaina

Dataset as a Table in Excel

	A	B	C	D
1	RowID	Year	Gender	Name
2	1	2014	Female	Abel
3	2	2014	Female	Abigail
4	3	2014	Female	Adalyn
5	4	2014	Female	Addison
6	5	2014	Female	Adeline
7	6	2014	Female	Adelynn
8	7	2014	Female	Adelynnne
9	8	2014	Female	Adison
10	9	2014	Female	Adyson
11	10	2014	Female	Ahlam
12	11	2014	Female	Alaina

The first task will be to obtain a subset of the name, i.e. the first letter. This can be accomplished using the =MID() function in Excel.

=MID( [Name], 1, 1)

- First argument: Original string from which the subset will be obtained
- Second argument: Starting position from which to begin the subset
- Third argument: Number of characters to include in subset

In cell E1, specify a name for this new variable, e.g. First Letter. Next, enter the function specified above into cell E2. This function will autofill all cells in the table.

	A	B	C	D	E
1	RowID	Year	Gender	Name	First Letter
2	1	2014	Female	Abel	=MID( [Name],1,1)
3	2	2014	Female	Abigail	

The =COUNTIF() function could be used to obtain the number of baby names that start with an A. From the table below, we see that 86 of the 643 names, about 13%, start with the letter A.

Counts by letter

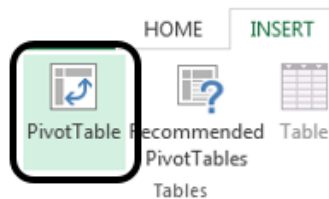
First Letter	Count
A	86
B	41
C	49
D	32
E	41
F	44

Using =COUNTIF() to obtain counts

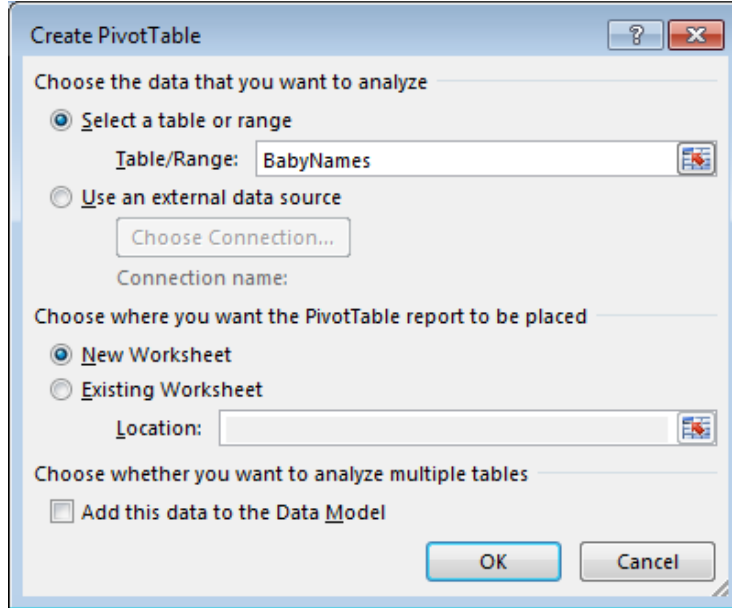
First Letter	Count
A	=COUNTIF(BabyNames[First Letter],"A")
B	=COUNTIF(BabyNames[First Letter],"B")
C	=COUNTIF(BabyNames[First Letter],"C")
D	=COUNTIF(BabyNames[First Letter],"D")
E	=COUNTIF(BabyNames[First Letter],"E")
F	=COUNTIF(BabyNames[First Letter],"F")

### PivotTables in Excel

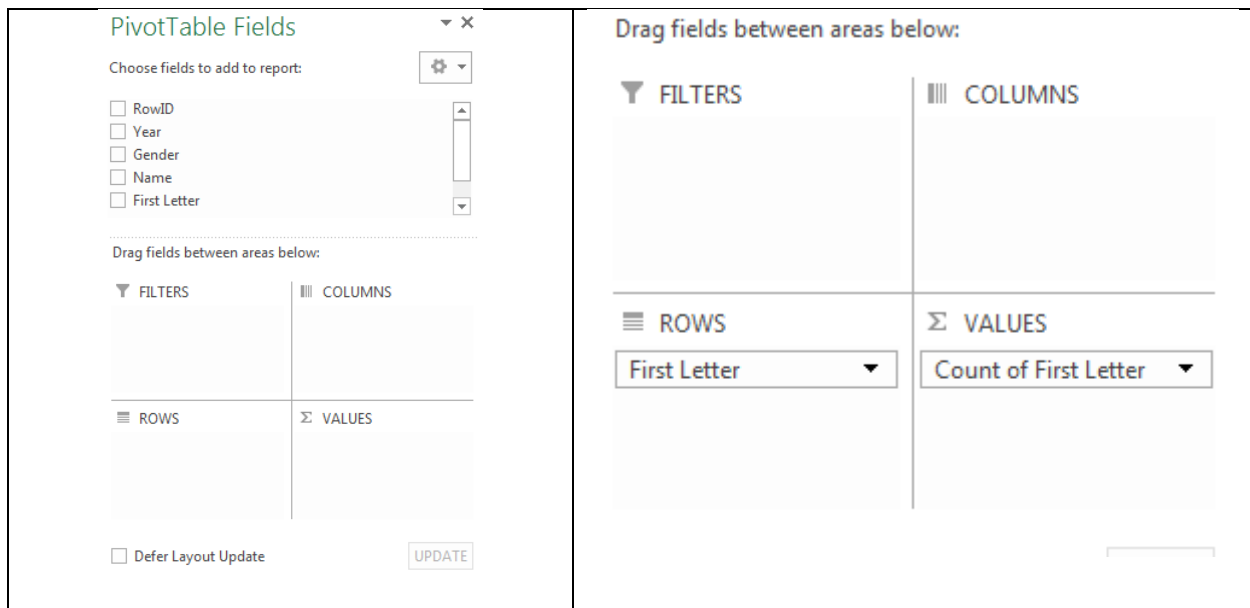
PivotTables are a commonly used feature in Excel. This is Excel's equivalent to the apply() function or Hadley Wickham's notion of group\_by(), i.e. aggregation. To construct a PivotTable, select Insert > PivotTable. On a MAC, select Data > PivotTable.



The initial window provided by PivotTables includes specification of the data to be summarized and the location of the output. I named my Table BabyNames, so this is specified under Select a table or range. A New Worksheet is best for output as output will not be placed over existing content.



Click OK. After the data and location for output has been specified. The PivotTable Field list is provided and is used to specify the structure for the resulting summary table. For example, if a frequency count of each letter is required, the First Letter can be dragged into the ROWS box and Frist Letter should also be dragged into the VALUES box. The VALUES box specifies what is to be calculated, e.g. a count or an average.



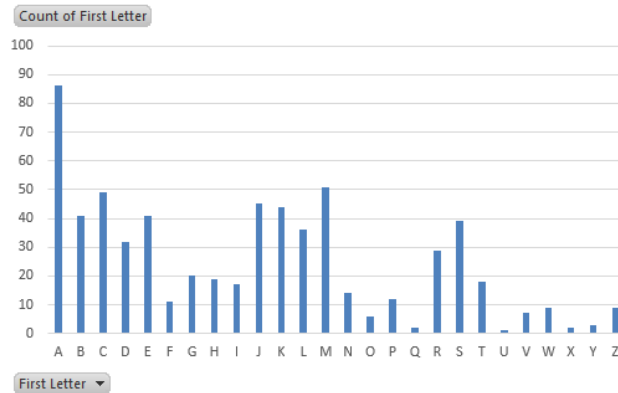
The following table is produced. A visualization of this table can easily be obtained.

Table of Counts

Row Labels	Count of First Letter
A	86
B	41
C	49
D	32
E	41
F	11
G	20
H	19
I	17
J	45
K	44
L	36
M	51
N	14
O	6
P	12
Q	2
R	29
S	39
T	18
U	1
V	7
W	9
X	2
Y	3
Z	9
<b>Grand Total</b>	<b>643</b>

Making a simple bar chart in Excel

- Step 1: Place cursor within PivotTable.
- Step 2: Select the graph of your choice under the Insert ribbon.

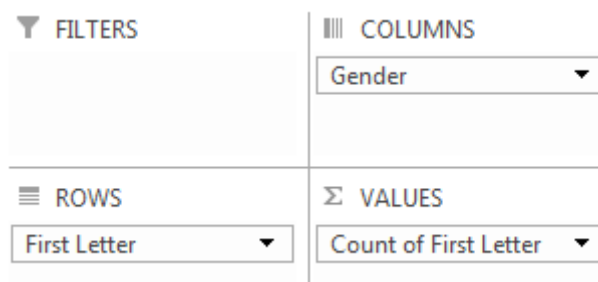


### Questions

1. Which first letter is most frequent?
2. Which letter is least frequent?
3. C is the first letter of my name. This is the 3<sup>rd</sup> most common first letter for a baby's name. How common is your first letter?

Note: Select a cell in the Count column of the PivotTable. Right click and select Sort to sort the table from the most frequent letter to the least.

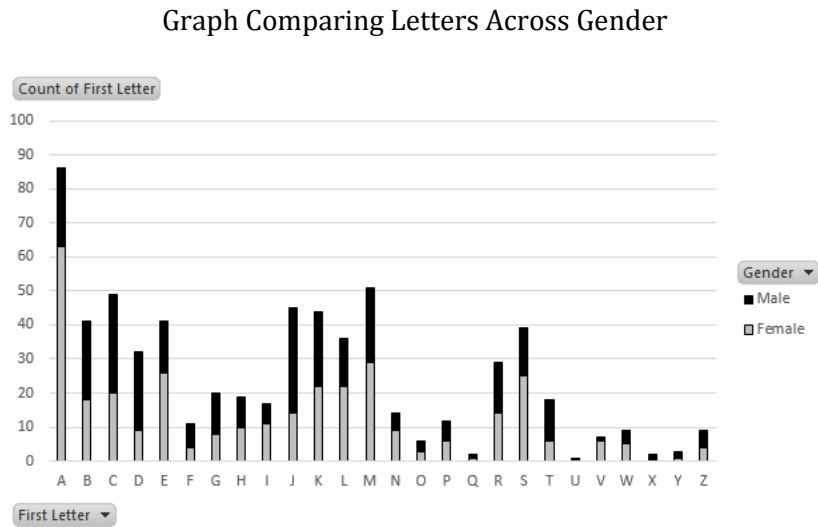
Next, we will have Excel separate the counts across Gender. This can be accomplished by simply dragging Gender in to the Columns box in Excel.



The following table of counts is produced. Once again, a graphical representation of this table may be beneficial for comparing genders.

First Letter Counts by Gender

Count of First Letter	Gender		Grand Total
Row Labels	Female	Male	
A	63	23	86
B	18	23	41
C	20	29	49
D	9	23	32
E	26	15	41
F	4	7	11
G	8	12	20
H	10	9	19
I	11	6	17
J	14	31	45
K	22	22	44
L	22	14	36
M	29	22	51
N	9	5	14
O	3	3	6
P	6	6	12
Q	1	1	2
R	14	15	29
S	25	14	39
T	6	12	18
U		1	1
V	6	1	7
W	5	4	9
X		2	2
Y	1	2	3
Z	4	5	9
Grand Total	336	307	643



### Questions

- Which first letter is most frequent for Females? How about Males?
- Consider only First Letter = S. Provide a measure of discrepancy between Females and Males for S. Briefly explain how you developed this measure.
- Use your measure of discrepancy to measure the discrepancy for other letters.  
Note: You should use Excel to automate the calculations here.
- Your friend decides to use the following measure of discrepancy. Do you believe this is a good measure for discrepancy? Discuss any advantages and disadvantages of this measure.

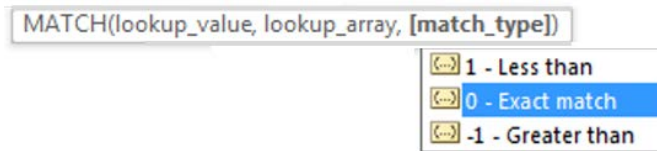
$$|\#Females\ with\ this\ Letter - \#Males\ with\ this\ Letter|$$

An Investigation of Vowels in Baby's Names

This section will involve an investigation of vowels. For the sake of our discussion here, y will be excluded from the vowel list.

Reference List of Vowels: a, e, i, o, u

The =MATCH() function in Excel can be used to identify whether or not the first letter is a vowel.



A reference list of vowels is needed and has been specified in cells H2: H6.

Cell F2: =MID( [First Letter] , H2:H6 , 0 )

	D	E	F	G	H
1	Name	First Letter	Start with Vowel		List of Vowles
2	Abel	A	=MATCH( [ First Letter ] , H2:H6 , 0 )		A
3	Abigail	A			E
4	Adalyn	A			I
5	Addison	A			O
6	Adeline	A			U
7	Adelvnn	A			

Provided your data is an Excel Table, this formula will autofill for all rows. Realize, this formula does not appear to be working for the remaining cells as the reference list for the vowels is incorrect for all rows except the first.

	D	E	F	G	H
1	Name	First Letter	Start with Vowel		List of Vowles
2	Abel	A	1		A
3	Abigail	A	=MATCH([First Letter], H3:H7, 0)		E
4	Adalyn	A	#N/A		I
5	Addison	A	#N/A		O
6	Adeline	A	#N/A		U
7	Adelvnn	A	#N/A		

Range incorrect when copied down

Absolute cell referencing should be used in this instance. An absolute cell reference will force the formula to retain the specified range. Absolute cell referencing is invoked by using a \$ around the letter and number reference for the cells.

Cell F2: =MID( [First Letter] , \$H\$2:\$H\$6 , 0 )

	D	E	F	G	H
1	Name	First Letter	Start with Vowel		List of Vowles
2	Abel	A	=MATCH([First Letter], \$H\$2:\$H\$6, 0)		A
3	Abigail	A	=MATCH([First Letter], \$H\$2:\$H\$6, 0)		E
4	Adalyn	A	=MATCH([First Letter], \$H\$2:\$H\$6, 0)		I
5	Addison	A	1		O
6	Adeline	A	1		U
7	Adelvnn	A	1		

The output from this function should be verified. A subset of rows is provided here and it appears the function is correct.

Verify =MID() function is working correctly

	D	E	F
1	Name	First Letter	Start with Vowel
2	Abel	A	1
3	Abigail	A	1
4	Adalyn	A	1
68	Bela	B	#N/A
69	Belladonna	B	#N/A
111	Dreena	D	#N/A
112	Eden	E	2
113	Elaine	E	2

Counts of Start with Vowel

Row Label	Count of Start with Vowel
1	86
2	41
3	17
4	6
5	1
#N/A	492
<b>Grand Total</b>	<b>643</b>

Questions

- Consider the value returned by the =MATCH() function. What does this value represent? Explain.
- Use the table of counts provided above to determine how often the first letter is a vowel?
- Modify the PivotTable provide above to determine how often the first letter is a vowel for a Female.

Comment

The following function can be used to relabel the Start with Vowel column as either a Yes or No.

=IF(ISNUMBER([Start with Vowel]),"Yes","No")

Verify the above formula for several rows

	D	E	F	G
1	Name	First Letter	Start with Vowel	Start with Vowel 2
2	Abel	A	1	Yes
3	Abigail	A	1	Yes
4	Adalyn	A	1	Yes
69	Belladonna	B	#N/A	No
70	Betty	B	#N/A	No
112	Eden	E	2	Yes
113	Elaine	E	2	Yes
114	Eleanor	E	2	Yes

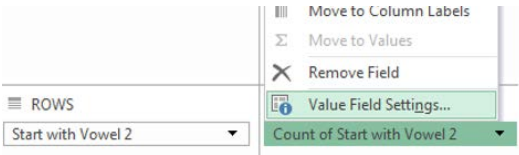
Summary of Count and Percentage Start with Vowel

Row Label	Count of Start with Vowel 2	% Start with Vowel 2
No	492	76.52%
Yes	151	23.48%
<b>Grand Total</b>	<b>643</b>	<b>100.00%</b>

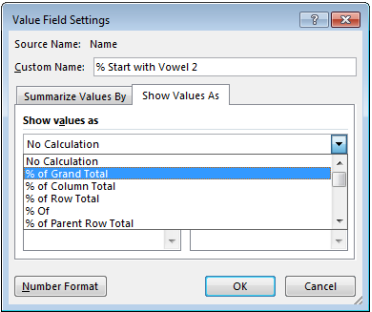


### Computing Percentages with PivotTables

In the VALUES box, right click on the variable for which a percentage is to be computed



Under the Show Values As tab, select % of Grand Total.



Note: On a MAC, the Show Values As menu can be found under the Options tab.

### Finding Particular Text within a String

I have a daughter whose name is Abbylyn. This name is somewhat uncommon; however, the use of “lyn” happens more often.

Cell E2: =FIND( “lyn”, [Name], 1)

- First argument: String to find
- Second argument: String to be searched
- Third argument: Starting position from which to begin search

Create a new column in your table and type the above into the first row of the data table in Excel.

	A	B	C	D	E
1	RowID	Year	Gender	Name	Contain lyn
2	1	2014	Female	Abel	= FIND( "lyn" , [Name] , 1 )
3	2	2014	Female	Abigail	
4	3	2014	Female	Adalyn	
5	4	2014	Female	Addison	

The =FIND() function returns the location within the string of the “lyn” instance. If “lyn” does not exist, the function returns a #VALUE error.

	A	B	C	D	E
1	RowID	Year	Gender	Name	Contain lyn
2	1	2014	Female	Abel	#VALUE!
3	2	2014	Female	Abigail	#VALUE!
4	3	2014	Female	Adalyn	4
5	4	2014	Female	Addison	#VALUE!
7	6	2014	Female	Adelynn	4
8	7	2014	Female	Adelynne	4
71	70	2014	Female	Bradlee	#VALUE!
72	71	2014	Female	Braelynn	5
74	73	2014	Female	Bricelyn	6
46	145	2014	Female	Grace	#VALUE!
47	146	2014	Female	Gracelynn	6

The =ISERROR() function is akin to the =ISNUMBER() function and can be used to relabel the Contain lyn column as “Yes” or “No” for whether or not it contains the text “lyn”.

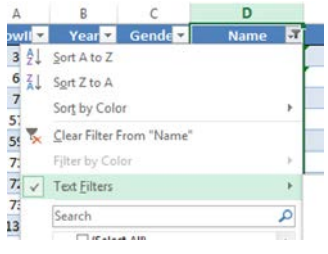
=IF(ISERROR( [Contain lyn] ) , "No" , "Yes" )

A summary of the Contain lyn variable suggests that about 3% of the names contain “lyn”.

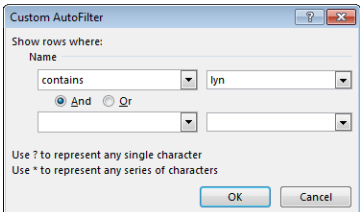
Row Label	Count of Contain lyn	% for Contain lyn
No	624	97.05%
Yes	19	2.95%

A text filter can be applied to the Name column as well to identify Names that contain the text “lyn”. This is shown here.

**Apply a Filter on Name**



**Specify contains lyn in the Custom AutoFilter box**



**The rows that contain lyn**

	A	B	C	D	E	F	G	H
1	RowID	Year	Gender	Name	Contain lyn	Contain lyn 2		
4	3	2014	Female	Adalyn	4	Yes		1
7	6	2014	Female	Adelynn	4	Yes		2
8	7	2014	Female	Adelynnne	4	Yes		3
58	57	2014	Female	Avalynn	4	Yes		4
60	59	2014	Female	Avelynn	4	Yes		5
72	71	2014	Female	Braelynn	5	Yes		6
73	72	2014	Female	Braylyn	5	Yes		7
74	73	2014	Female	Bricelyn	6	Yes		8
135	134	2014	Female	Evelyn	4	Yes		9
147	146	2014	Female	Gracelynn	6	Yes		10
185	184	2014	Female	Kaelyn	4	Yes		11
186	185	2014	Female	Kaelynn	4	Yes		12
220	219	2014	Female	Lillylynn	6	Yes		13
223	222	2014	Female	Locklynn	5	Yes		14
227	226	2014	Female	Luxlyn	4	Yes		15
231	230	2014	Female	Madalynn	5	Yes		16
233	232	2014	Female	Madelyn	5	Yes		17
236	235	2014	Female	Magdalynn	6	Yes		18
279	278	2014	Female	Raelyn	4	Yes		19
550	549	2014	Male	Lyncoln	#VALUE!	No		20
551	550	2014	Male	Lynkin	#VALUE!	No		21

Consider the table above. The two names at the bottom contain “Lyn”, but were not identified by the =FIND() function. The reason this discrepancy exists is because the =FIND() function is *case-sensitive*. That is, “Lyn” is different from “lyn” for this function.

The following table compares the behavior =FIND(), =SEARCH(), and =MATCH().

<b>=FIND()</b> <i>case-sensitive</i>		A	B	C	
	1		=FIND("A",A1,1)	=FIND("a",A1,1)	
	2	Abel	1	#VALUE!	
	3	Anna	1	4	
	4	Braelynn	#VALUE!	3	
	5	Christina	#VALUE!	9	
<b>=SEARCH()</b> <i>case-insensitive</i>		A	B		
	1		=SEARCH("A",A1,1)		
	2	Abel	1		
	3	Anna	1		
	4	Braelynn	3		
	5	Christina	9		
<b>=MATCH()</b> <i>case-insensitive exact matches only</i>		A	B	C	D
	1		=MATCH("A",A1,1)	=MATCH("Abel",A1,1)	=MATCH("abel",A1,1)
	2	Abel	#N/A	1	1
	3	Anna	#N/A	#N/A	#N/A
	4	Braelynn	#N/A	#N/A	#N/A
	5	Christina	#N/A	#N/A	#N/A

The =LOWER() and =UPPER() functions can be used to convert all text within a string to lowercase and uppercase, respectively.

### Replacing Text within a String

The following procedure will be used to count the number of vowels in a baby's name.

- Obtain the length of the baby name
- Remove the vowels using the =SUBSTITUTE() function
  - This will be done in successive steps
    - First, remove the a's from the original string
    - Next, remove the e's from the string that contains no a's
    - Continue to remove i's, o's, and u's in a successive manner
- Obtain the length of the name after removing all vowels
- Compute the percentage of vowels for each name

Consider the following applications of the =SUBSTITUTE() function. The LOWER() function is being used here because the =SUBSTITUTE() is case-sensitive.

	A	B	C	D
1		=SUBSTITUTE(A1,"a","#")	=SUBSTITUTE(LOWER(A1),"a","#")	=SUBSTITUTE(LOWER(A1),"a","")
2	Abel	Abel	#bel	bel
3	Anna	Ann#	#nn#	nn
4	Braelynn	Br#elynn	br#elynn	brelynn
5	Christina	Christin#	christin#	christin

Task #1

Compute the proportion of vowels for each name in this dataset. The process for doing this is described above. My output is provided here for the first few names. The following should help get you started.

- Cell E2 contains the function =LEN( [Name] )
- Cell F2 contains the function =SUBSTITUTE( LOWER( [Name] ),"a","")
- Cell G2 contains the function = SUBSTITUTE( LOWER( 'Remove A's' ), "e", "" )
- The following is used in cell L2

$$1 - \frac{\text{Lenth of name without vowels}}{\text{Length of name}}$$

	D	E	F	G	H	I	J	K	L
1	Name	Length of Name	Remove A's	Remove E's	Remove I's	Remove O's	Remove U's	Length No Vowel	Proportion Vowel
2	Abel	4	bel	bl	bl	bl	bl	2	0.5
3	Abigail	7	bigil	bigil	bgl	bgl	bgl	3	0.571428571
4	Adalyn	6	dlyn	dlyn	dlyn	dlyn	dlyn	4	0.333333333
5	Addison	7	ddison	ddison	ddsn	ddsn	ddsn	4	0.428571429
6	Adeline	7	deline	dlin	dln	dln	dln	3	0.571428571
7	Adelynn	7	delynn	dlynn	dlynn	dlynn	dlynn	5	0.285714286

The following summaries were obtained for the Proportion of Vowels column.

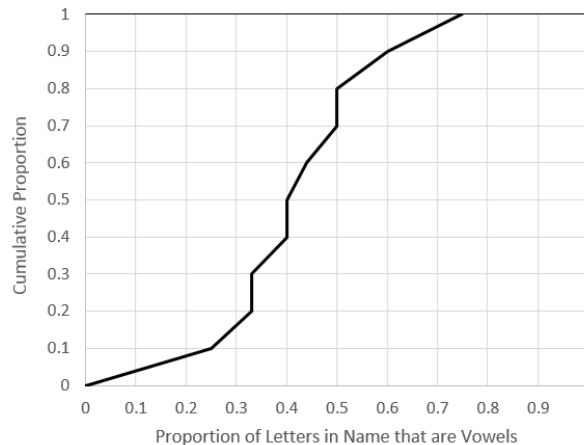
Standard Statistical Summaries

Summary Measure	Value
Average	0.417
Median	0.40
Standard Deviation	0.127

Distribution of Counts

Percentage Vowels in Name	Counts
0.000	3
0.111	1
0.125	1
0.143	3
0.167	13
0.200	22
0.222	2
0.250	31
0.273	1
0.286	45
0.333	95
0.364	1
0.375	23
0.400	90
0.429	44
0.444	14
0.462	2
0.500	158
0.556	4
0.571	17
0.600	44
0.625	1
0.667	24
0.750	4
<b>Total</b>	<b>643</b>

Cumulative Density Plot for Percentage Vowels in Name



Questions

11. Consider the following statement, “A majority of names have more than half their letters as vowels.” Is this statement true? Discuss.
12. From the Distribution of Counts table, three names do not contain any vowels. What are these names?
13. The Social Security Administration of the United States Government maintains a website that contains information on baby names dating back to the late 1800’s. Website: <http://www.ssa.gov/oact/babynames/limits.html>

I have computed the summary measures for names from 1900. Has the distribution of vowels changed much? Discuss.

Summary Measure	Year: 2014	Year: 1900
Average	0.417	0.414
Median	0.40	0.40
Standard Deviation	0.127	0.133
Minimum	0	0
Maximum	0.75	0.80

Task #2

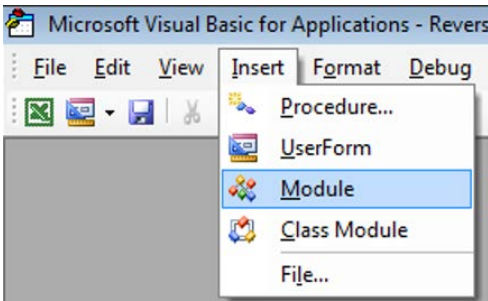
A palindrome is a word that is spelled exactly the same forward and backward. The name of our second oldest daughter is a palindrome.

ANNA  
→
ANNA  
←

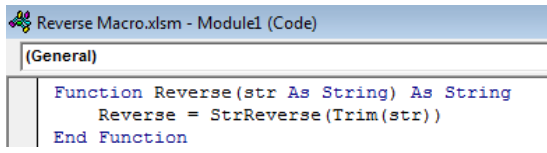
The evaluation of whether or not a name is a palindrome requires that a string be searched backwards. Excel does not have a built-in function for this. However, Visual Basic does contain a StrReverse() function. The following can be used to create a custom function in Excel using Visual Basic.

The Excel Visual Basic Editor can be obtained using Alt + F11. This editor can also be found on the Developer ribbon – which may have to be added to the list of visible ribbons.

Insert > Module will provide a new module window



Creating a custom formula named =Reverse()



```
Function Reverse(str As String) As String
    Reverse = StrReverse(Trim(str))
End Function
```

Save your Excel file as a macro-enabled file. This is required for your new function to work. You should be able to use your new function. A simple application is shown here.

	A	B	C
1			
2		Anna	=Reverse(B2)
3			annA

The following columns were used to identify whether or not each name was a palindrome.

	A	B	C	D	E	F	G
1	Name	Length	int(Length/2)	Left Half	Reverse	Right Half	Palindrome
2	Abel	4	2	Ab	lebA	le	No
3	Abigail	7	3	Abi	liagibA	lia	No
4	Adalyn	6	3	Ada	nyladA	nyl	No
5	Addison	7	3	Add	nosiddA	nos	No
6	Adeline	7	3	Ade	eniledA	eni	No
7	Adelynn	7	3	Ad	nnyledA	nnv	No

Questions

14. What is the purpose of Column C? What function might one use in Column G?
15. My process identified Anna, Ava, Aziza, Hannah, and V as palindromes. Verify that no palindromes were missed by my procedure.