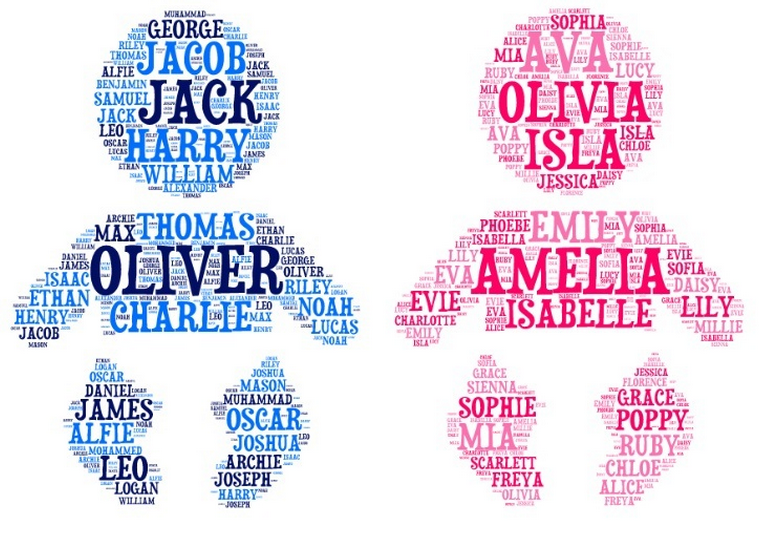
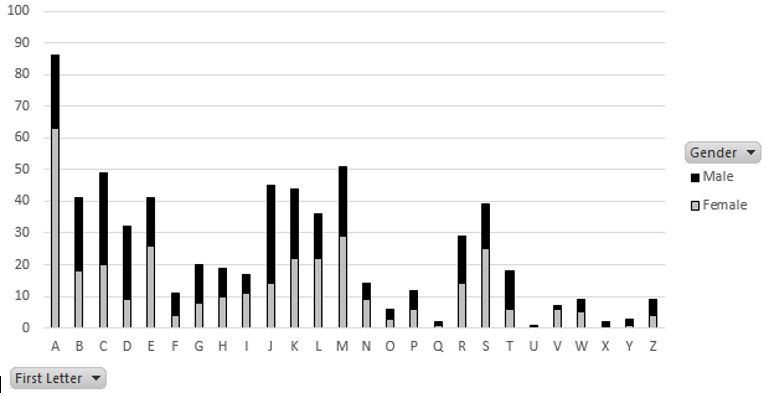
Excel: Working with Strings, i.e. MID(), MATCH(), INDEX()

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 then 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

1. String functions in Excel
2. Summaries and Visualizations through PivotTables

|  |  |
| --- | --- |
| Data Source | |
| Address | <https://github.com/WINSTATS-REU/Materials> |
| Description | BabyNames Dataset   |  |  | | --- | --- | | 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 | http://cdn.smosh.com/sites/default/files/ftpuploads/bloguploads/0613/tech-baby-names-google.jpg | |

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.

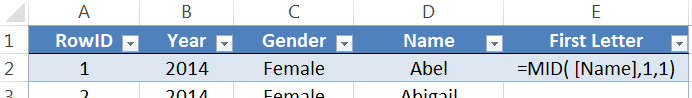
|  |  |
| --- | --- |
| Data in Excel | Dataset as a Table in Excel |

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.

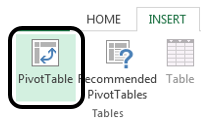


The =COUNTIF() function could be used to obtain the number of baby names that start with an A. From the table below, we see that 86 of the 643 names, about 13%, start with the letter A.

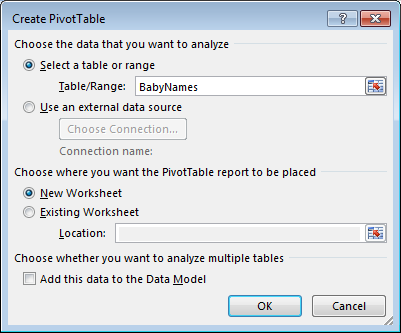
|  |  |
| --- | --- |
| Counts by letter | Using =COUNTIF() to obtain counts |

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.

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

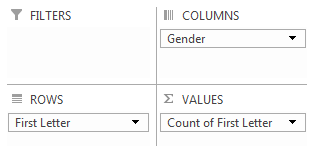
The following table is produced. A visualization of this table can easily be obtained.

|  |  |
| --- | --- |
| Table of Counts | 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 3rd 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 accmomplished 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 | Graph Comparing Letters Across Gender |

Questions

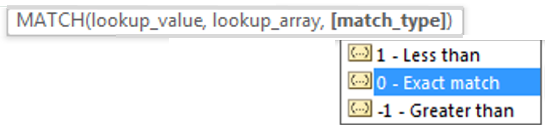
* 1. Which first letter is most frequent for Females? How about Males?
  2. Consider only First Letter = S. Provide a measure of discrepancy between Females and Males for S. Briefly explain how you developed this measure.
  3. Use your measure of discrepancy to measure the discrepancy for other letters.   
     Note: You should use Excel to automate the calculations here.
  4. 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.

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.

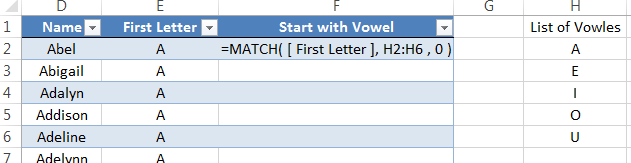
Reference List of Vowels: a, e, i, o, u

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

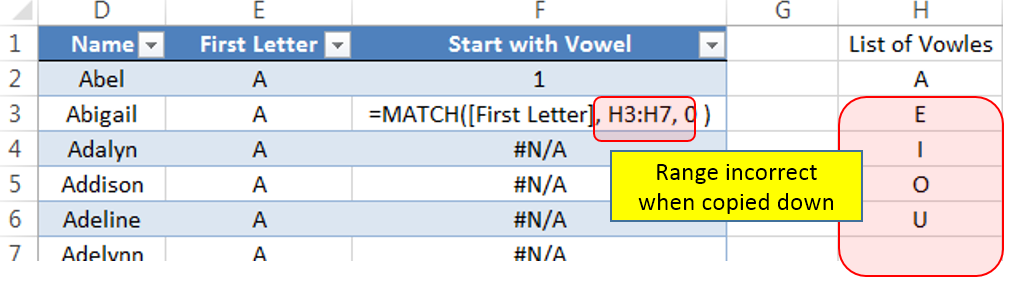


A reference list of vowels is needed and has been specified in cells H2: H6.

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

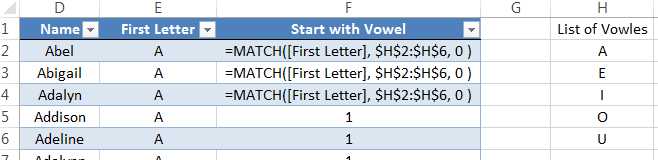


Provided your data is an Excel Table, this formula will autofill for all rows. Realize, this formula does not appear to be working for the remaining cells as the reference list for the vowels is incorrect for all rows except the first.



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 )



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

|  |  |
| --- | --- |
| Verify =MID() function is working correctly | Counts of Start with Vowel |

Questions

* 1. Consider the value returned by the =MATCH() function. What does this value represent? Explain.
  2. Use the table of counts provided above to determine how often the first letter is a vowel?
  3. 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 | Summary of Count and Percentage  Start with Vowel |

|  |  |
| --- | --- |
| 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. |

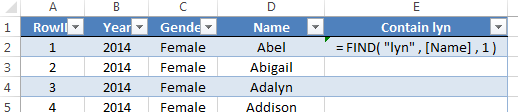
Finding Particular Text within a String

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

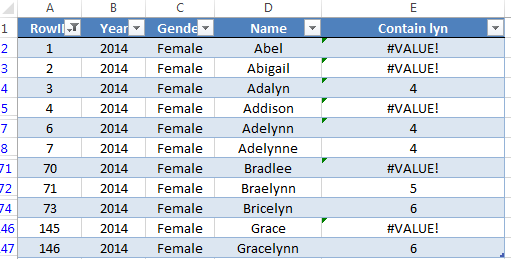
Cell E2: =FIND( “lyn”, [Name], 1)

* 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.



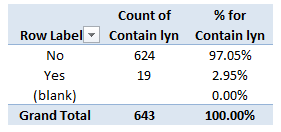
The =FIND() function returns the location within the string of the “lyn” instance. If “lyn” does not exist, the function returns a #VALUE error.



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”.



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 |

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().

|  |  |
| --- | --- |
| =FIND()  *case-sensitive* |  |
| =SEARCH()  *case-insensitive* |  |
| =MATCH()  *case-insensitive*  *exact matches only* |  |

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

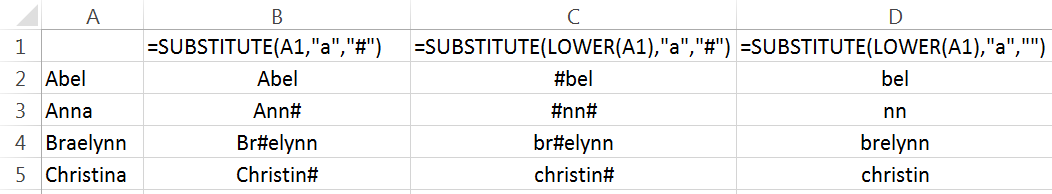
The following procedure will be used to count the number of vowels in a baby’s name.

* 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

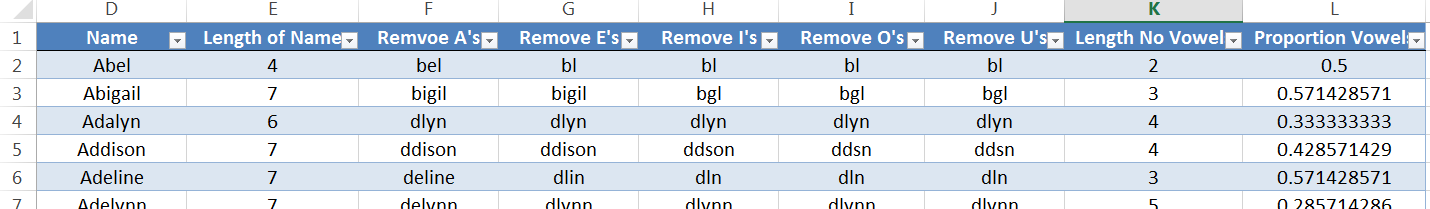
Consider the following applications of the =SUBSTITUTE() function. The LOWER() function is being used here because the =SUBSTITUTE() is case-sensitive.



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



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 | Cumulative Density Plot for Percentage Vowels in Name |

Questions

* 1. Consider the following statement, “A majority of names have more than half their letters as vowels.” Is this statement true? Discuss.
  2. From the Distribution of Counts table, three names do not contain any vowels. What are these names?
  3. 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

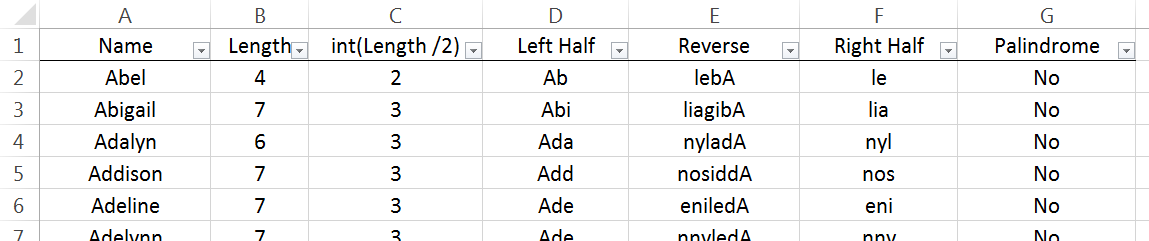
A palindrome is a word that is spelled exactly the same forward and backward. The name of our second oldest daughter is a palindrome.



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. |

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



Questions

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