LVAIC Excel Workshop 1/11/2023

Zane Kratzer, Strategic Analytics Manager
Lehigh University, Office of Institutional Research and Strategic Analytics

Use with "Vendor Spend_Sample.xlsx"
(Reminder: This is a fake dataset created explicitly for use in the workshop.)

A. Questions to Ask Before Beginning Your Analysis
    a. Do you need to create a copy of the raw data before making edits/revisions?
    b. Does the file contain extra headers, row totals, additional notes and comments?
    c. Is the data already in a simplified Rows/Columns format?
    d. Do you know what all the columns in the dataset represent?
    e. Are there additional columns not needed for your analysis?
    f. Is there documentation? Notes on data definitions, calculated fields?
    g. Should you keep record of all data transformations for future use?

B. Checking for Data Types, Column Formatting and Missing Values
    a. First, do a quick audit of the primary fields you will be working with in the dataset
        i. Click anywhere on the spreadsheet and then go to upper right menu, select dropdown for "Sort & Filter" and select "Filter"
        ii. Use the Filter dropdown arrows to examine the data in each field
        iii. Consider what data should be expected in each field. Do you notice anything that seems questionable?
            1. Are negative values in the "Amount" column accurate?
            2. Should the "Vend_ID" column be numeric or text?
            3. Why are there UNKNOWN and blank values for "Vendor Name"?
        iv. Highlight a specific field, then go to upper right menu, select dropdown for "Sort & Filter" and select either "Sort A to Z" or "Sort Z to A" to further examine the content.
    b. Highlight fields, right-click > Format Cells to check on data type formatting
    c. Trans_ID: Text
    d. Date: change data formatting (Custom vs Date in Format Cells window)
    e. Amount: change to Currency and adjust decimal points
    f. Vend_ID: General, Excel automatically treats as a Numeric
    g. Vendor Name: General, Excel automatically treats as a Text
    h. Purchasing Department: General, Excel automatically treats as a Text

C. Working with Pivot Tables
    a. Ctrl-A to select all relevant fields and rows in a spreadsheet
        i. Be sure to remove empty columns and empty header rows
    b. Insert > Pivot Table
    c. Choose between "New Worksheet" and "Existing Worksheet"
    d. Use the "PivotTable Analyze" menu in the top toolbar to make changes to the PivotTable:
        i. Refresh
        ii. Change Data Source
    e. Values: The field to be summarized (i.e. Trans_ID provides a Count of all Trans_ID values)

i. To change how the sum works, click on dropdown arrow and select “Value Field Settings”

ii. Change the “Summarize value field by” to Sum, Count, Average, Max, Min, etc. (Note that if your field is a text value then numeric options such as Sum and Average will not work)

iii. To show the summed values as a percentage or other form, click on the “Show Value As” tab and select an option such as “% of Column Total”

f. Rows and Columns: Fields to be inserted into either the Row or Column of the pivot table

g. Filters: Fields to be used to filter out specific values from the data

h. Sort: Click on a value in one of the columns that you want to sort the table by. Right-click on select “Sort” and then “Sort Smallest to Largest” or “Sort Largest to Smallest”

D. Identifying Duplicate Records

a. Counting Duplicate Records by Formula:

i. Formula to use: =IF(COUNTIF($A$2:A2,A2)=1,COUNTIF(A:A,A2),"")

ii. Replace ‘A’ with whatever column your target field is in

iii. If your data does not begin on row 2, replace ‘2’ with whatever row your data begins on

b. Counting the Number of Unique Records by Formula:

i. Formula to use: =SUM(IF(A2:A7579<>"",1/COUNTIF(A2:A7579,A2:A7579), 0))

ii. Change the cell range to match the column and rows you are working with

iii. ‘A2’ gets replaced with the first cell where data begins

iv. ‘A7579’ gets replaced with the last cell where data ends

c. Finding Duplicate Records with Conditional Formatting:

i. Ctrl-A to select all relevant fields and rows in a spreadsheet

ii. Conditional Formatting > Highlight Cells Rules > Duplicate Values

iii. Check default settings on Duplicate Values dialog box and select OK

iv. To remove formatting, go to Conditional Formatting > Clear Rules > Clear Rules from Entire Sheet

E. Fixing the Loss of Leading Zeros in an ID field

a. Vend_ID is supposed to have leading zeros but it was converted to a numeric variable and no longer has the leading zeros.

b. Recheck formatting: Highlight the Vend_ID column, right-click > Format Cells > General

c. Insert a new column next to it and name it “Vend_ID2”

d. =CONCATENATE(“000”,D2)

e. Change the ‘D2’ to reflect wherever your target field is located

f. Double-click on dropdown button (in bottom right corner) to populate all rows with the new formula

g. The new field should show as a left-justified Text field with leading zeros, compared to the original right-justified Numeric field with leading zeros stripped.

F. Merging Multiple Tables/Pulling in Fields from other Sheets

a. VLOOKUP

i. Create a new field called “Vendor Name Edited”

ii. Use VLOOKUP to pull in the correct Vendor Name from the “Vendor Info-SAMPLE” tab

1. =VLOOKUP(‘Vendor Spend-SAMPLE’!E2,‘Vendor Info-SAMPLE’!$A$2:$C$1481, 2, FALSE)

2. The first part assigns the value that is being searched for, in this case, you are looking for the Vend_ID in the E column of the original sheet ('Vendor Spend-SAMPLE'!E2)
3. The second part assigns the table range where you are going to look for a match, in this case, all columns and rows of the Vendor Info sheet ('Vendor Info-SAMPLE'!$A$2:$C$1481)
4. The third part assigns which column from the new sheet you want to pull in new data from, in this case, column 2 since we want the Vendor Name column (2)
5. The last part tells Excel whether you want to use an approximate match (TRUE) or if you need an exact match (FALSE). In this case, we use FALSE because we have a specific ID value we need to match.

iii. Some other Tips and Tricks for working with VLOOKUP:

1. Be sure to add $ to the table range formula (2nd part of VLOOKUP input) or else when you go to drag down your new VLOOKUP formula to the remaining rows, the VLOOKUP will not properly work because your table array range will change as you drag down the formula.
2. VLOOKUP does not respond well to column sorting. Once you have successfully finished the VLOOKUP, you have 2 options to make it easier to work with your new data:
    a. Copy the new VLOOKUP data to a new field with Paste Options > Values (V), this will create a new column that is not dependent on the original VLOOKUP formula and is now populated with the new values
    b. Remove the beginning part of the VLOOKUP formula that points to the original sheet name ('Vendor Spend-SAMPLE'!). This makes column sorting with new VLOOKUP fields much easier to work with.

G. Combining Years

a. What is the unit of analysis?
b. If each row must contain a unique ID, then the data should be in a wide format, with each row representing a unique ID and all subsequent fields or measurements for that ID added as new columns.
c. If the unit of analysis is the point of measurement for each ID (a specific term, date, event, etc.), then the data should be in a long format, with each row representing a unique data entry for each ID. Some IDs will have duplicate rows if data was recorded for that particular ID more than once.

H. Other Data Analysis/Management Tools to Consider

a. SQL/SQL Developer/SQL Server Management Studio
b. Tableau Desktop/Tableau Prep Builder/Tableau Public
c. Microsoft Power BI
d. Argos
e. Microsoft Access
f. R or Python