Assignment #4: ETL in Excel

(Due Sunday, March 12, *9:00am*)

Guidelines

- You must **submit** the final version of your **EXCEL file** "Assignment #4 - ETL Workbook.xlsx" worksheet to Blackboard.

If you do not follow these instructions, your assignment will be counted late.

This assignment requires you to perform the Extract, Transform, Load (ETL) process in Excel workbook. *You will be taking two sets of data and combining them into a single source* that can be analyzed. Right now, the two sets of data are similar, but have too many differences which prevent records from being directly compared across sets. You will need to create and implement rules that resolve the differences in the data. So you will be:

- **Extracting** the data from source worksheets.
- **Transforming** the data using Excel formulas.
- **Loading** the data into a new worksheet that contains a single set of combined data.

You'll be working with data in the Excel workbook "ETL Workbook.xlsx." This spreadsheet is a collection of orders, organized by line item. *Multiple rows can be associated with a single order because you can order multiple items in a single order.*

There are 5 worksheets in this workbook:

- Source 1: The first data source (29 records)
- Source 2: The second data source (30 records)
- Full Set: An empty data source that will contain a consolidated set of all 59 records
- Lookups: Where we'll store the tables used to look up values. You'll use this throughout the assignment.
- Description: Documentation regarding where this data set originated.

To understand the inconsistencies between the data, open the workbook and look at the Source 1 and Source 2 worksheets. You'll notice that the data doesn't quite match up. For example, order is represented in Source 1 as a five-digit number (i.e., 10001) but in source 2 as an "A" followed by a five-digit number (i.e., A10001). Left as is, an analysis (such as a Pivot Table) would see this as two different orders. The data must be reconciled so that the format is the same.

You should not make any changes to the Source 1 or Source 2 worksheets to complete this assignment!

Part 1: ETL with the OrderID field (FOLLOW-ALONG)

Let's decide that the rule is to leave OrderID in Source 1 alone and remove the "A" from Source 2. Try this:

1) Click on the "Full Set" tab.
2) Click on cell B2.
3) Press "=" to start a formula, switch to the Source 1 tab, and click on A2 there.
4) Press Enter and you'll see the OrderID from Source 1.
5) Copy that formula down to cell B30 on the Full Set tab (you'll see it's labeled "Data From Source 1").
6) Now click on cell B31.
7) Press "=" to start a formula, and type RIGHT('Source 2'!A2,LEN('Source 2'!A2)-1)
   (don't forget the space between "Source" and "2"!!)
8) Press Enter and you'll see the OrderID from Source 2 without the leading "A"
9) Copy that formula down to cell B60 on the Full Set tab (the part labeled "Data From Source 2").

Dissecting the formula (READ THIS – IT'S IMPORTANT!):

- RIGHT(value, n) is an Excel function that takes the right n characters of a string value. So RIGHT("HELLO", 2) will return "LO".
- LEN(value) returns the number of characters contained in a string value. So LEN(123) and LEN("DOG") both return "3".
- So LEN('Source 2'!A2)-1 looks at the length of the cell A2 and returns everything except the first character. Here's an example: Let's say the cell contains "A12345". The LENgth is 6, so length -1 is 5. Now if you take the right 5 characters of A12345 you get only 12345.
- So you've transformed your data into a new format!

Part 2: ETL with the Customer State/Province field (FOLLOW-ALONG)

Now let's look at the "Customer State/Province" field. Our rule will be that state and provinces (for Canada) names will be displayed using their abbreviation (i.e., PA instead of Pennsylvania, ON instead of Ontario). To do this, we will use the "State/Province Lookup" table that has been created in the "Lookups" worksheet.

Before you begin: Take a look at the "State/Province Lookup" table in the Lookups tab. Then look at how State/Province is represented in the Source 1 and Source 2 tabs (they are different). Now follow the instructions below:

1) Click on the "Full Set" tab.
2) Click on cell E2.
3) Press "=" to start a formula, switch to the Source 1 tab, and click on D2 there.
4) Press Enter and you'll see the state abbreviations from Source 1.
5) Copy that formula down to cell E30 on the Full Set tab (you'll see it's labeled "Data From Source 1").
6) Now click on cell E31.
7) Press "=" to start a formula, and type VLOOKUP('Source 2'!E2,Lookups!$A$3:$B$62,2,FALSE)
8) Press Enter and you'll see the state abbreviation from Source 2 ("KS") instead of the full name ("Kansas")
9) Copy that formula down to cell E60 on the Full Set tab (the part labeled "Data From Source 2").

Dissecting the formula (READ THIS TOO – IT'S ALSO IMPORTANT!):

- VLOOKUP(lookup_value, table_array, column_index, range_lookup) is an Excel function that will match a value with another value in a separate table.
- So "lookup_value" is value that you're looking for. So in this case Excel will look for the value contained in cell E2 in the Source 2 worksheet. In this case, that value is "Kansas".
- And "table_array" is the table where you're going to do your search. The table is from A3 to B62 on the "Lookups" worksheet. Notice that the first column of that table is in alphabetical order. **That is what it uses to find a match; if the first column of your lookup table isn't in alphabetical order (or ascending numerical order) the function won't work.**
- Also, you need to use the dollar signs to keep the cell references from changing when you copy the formula to the other cells on the Full Set worksheet. In other words, you're lookup value keeps changing, but your lookup table is always the same.
- The parameter "column_index" indicates column number with the value that is returned. Notice that column 2 has all of the state abbreviations.
- Finally, "range_lookup" is TRUE if we are looking for approximate matches and FALSE if we are looking for exact matches. Unless you have a good reason to do so, always use FALSE.

Part 3: NOW…Finish the worksheet (ON YOUR OWN)

Perform the ETL process on the rest of these fields in the Full Set worksheet:

- **Customer Full Name***
- Customer City
- **Customer Status***
- Order Date
- Product ID
- Product
- Unit Price
- Quantity
- Discount
- Full Price
- Extended Price
- **Total Discount***

* These are fields with inconsistent data between the fields.

In most cases you'll just be copying the data from each worksheet without transformation (like you did in the first five steps in Parts 1 and 2). For example, Order Date is represented in the same way in Source 1 and Source 2.

For Customer Full Name, Customer Status, and Total Discount, you'll need to transform the data.

Here are a summary of the remaining inconsistencies:

Source 1 Field	Source 2 Field	In the Full Set tab
Customer Full Name as one field	Customer First Name and Customer Last Name as separate fields	Customer Full Name should appear as FirstName LastName with a single space in-between for all customers (use CONCATENATE – see below)
Customer Status as "Silver," "Gold," and "Platinum." Platinum is the best.	Customer Status as 1, 2, and 3. (Silver = 1, Gold = 2 and Platinum = 3)	Customer Status should appear as Silver, Gold, or Platinum for all customers (use VLOOKUP and the Lookups worksheet. You need to fill in the values in the Lookups worksheet for the "Customer Status Lookup" section.)
Total Discount included	Total Discount not computed	Total Discount should be computed for all customers (use a formula to subtract the extended price from the full price)

You can use whatever transformation you'd like, but when you are done the data has to be consistently formatted across the entire set of data. Make the changes to the "Full Set" tab.

One more formula that might be useful to you…

CONCATENATE(value1, value2…): Combines two or more string values or data in cells

Example:
CONCATENATE(A2, ", HELLO") will append the string ", HELLO" to the end of whatever is in cell A2. Like this:

	A	B
1	Name	NewCell
2	Bob	Bob, HELLO
3	Jack	Jack, HELLO
4	Sue	Sue, HELLO
5	Janet	Janet, HELLO

Part 4: Credit Line field

Add the data for credit line to the "Full Set" worksheet. A minimum credit line of $2,000 has been established, so that even if the customer has a credit line of $0 it is changed to $2,000. Use the VLOOKUP() function to put this data into the "Full Set" worksheet, using the Credit Line Lookup table in the Lookups worksheet.

If you do it correctly, at first there will be three errors ("N/A" values in three cells). That's because there is a problem with the data in the Credit Line Lookup table.

Make the necessary change to the data in the Credit Line lookup table to correct the issue so that Credit Line data appears for all the customers.