Assignment # 1

Submission Deadline: Sunday, 9th February 2020, 11:55 PM

Only slate submissions will be accepted. NO EMAIL Submissions.

NO Deadline Extensions. Please.

Assignment: Extract, Transform, Load (ETL)

This assignment requires you to perform the Extract, Transform, Load process on an Excel workbook. You will be taking two sets of data and combining it 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 the original worksheet.
- **Transforming** the data using an Excel formula.
- **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-Data for Assignment 1.xlsx." It 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.
- Source 3(for Assignment Part 4): The Lookup data for Part4 of the assignment

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.

Part 1: ETL with the OrderID field

Perform the necessary Transformation operation using Excel’s formula’s to Extract the OrderID in Source 1 as it is and Load it into Full Set worksheet. Whereas remove the “A” in OrderID from Source 2 and then load it into the Full Set. At the end, you have all the orderID’s in the same format in Full Set worksheet.

Part 2: ETL with the Customer State/Province field

Now let’s look at the “Customer State/Province” field. Your 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, you will use the “State/Province Lookup” table that has been created in the “Lookups” worksheet.

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). Source 1 is storing ‘PA’, where as Source 2 is storing Pennsylvania.

Perform the necessary transformation steps to Extract the Source 1 State/Province field and load it in the Full Set sheet as it is, where as extract the Source 2 State/Province field and perform the transformation by looking up the extracted value in the “Lookups” worksheet to find its two letter abbreviation and then load that abbreviation in the Full Set worksheet.

At the end, you have all the Customer State/Province in the same format in Full Set worksheet.

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. 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 is 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
Customer Status as "Silver," "Gold," and "Platinum." Platinum is the best.	Customer Status as 1, 2, and 3. 3 is the best.	Customer Status should appear as Silver, Gold, or Platinum for all customers (use Status Lookup in the 'Lookups' worksheet)
Total Discount included	Total Discount not computed	Total Discount should be computed for all customers Inspect the data very carefully of Source 1, to find the pattern of calculating 'Total Discount' and then perform the same pattern for generating values for missing Source 2 worksheet Total Discount column

You can use whatever transformation (excel formulas) you'd like, but when you are done the data has to be consistently formatted across the entire set of data.

Part 4: Credit Line field

Add the data for credit line to the "Full Set" worksheet from "Source 3(for Assignment Part 4)" worksheet. A minimum credit line of $2,000 has been established, so that even if the customer previously had a credit line of $0 it is changed to $2,000. Use the Excel functions to put this data, i.e., New Credit Line column into the "Full Set" worksheet, by using "Source 3(for Assignment Part 4)" as a lookup information for customer names to match with Customer names in "Full Set" worksheet.

If you do it correctly, at first there will be three errors ("N/A" values in three cells) of Credit Line column in "Full Set" worksheet. That's because there is a problem with the data in the "Source 3(for Assignment Part 4)".

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

Upload the assignment on Slate, as an Excel workbook with your name/number appended with "Full Set" worksheet, before the deadline.