Task
You have done such a good job with cleaning data from Vandelay Industries that they have asked you to do some further cleaning of their data. The sales group is suspicious that there might be errors in the data for January.

You will be working with a new set of 3,296 orders with 5,182 line items from January 2014. The data is in a file called “VandelayJan2014.xlsx.” A “line item” is just an order for a specified number of a particular product – there can be multiple line items per order.

You’ll be looking for errors in the data in several places:
1) Errors in the product names.
2) Errors in the promotional codes.
3) Errors in the total product price.

You will find, document, and correct the errors in the Excel workbook.

Deliverables
Submit both your answer sheet and Excel file to Blackboard.

Part 1: Errors in Product Names
Verify if the product names in Column J are correct by using the master product list in the Lookups tab. You should assume that the information in the Lookups tab is always right. So if there is a mismatch, the error is in your data set.

Use the MATCH function (see below) and place your function in Column N. Make the title of the column “ProdMatch” (in Cell N1) and start your MATCH formulas in Cell N2.

List the products names with errors, the corrected name, and how many rows of data had the same error on your answer sheet.

HINT: Try sorting by product_name. You only need to list each incorrect product name.
Dissecting the MATCH function (READ THIS – IT'S IMPORTANT!)

MATCH(value, lookup_array, match_type) is an Excel function that searches a list of values for a single value (i.e., looking for the number “105” in a list of house numbers).


If the value is in the table, it returns the row number where that value is found. If the value isn’t in the table, it returns “#N/A” (an error!). This give us an easy way of checking to see if a value is in a list.

Part 2: Errors in Promotional Codes

Verify if the promotional codes in Column E are correct by using the master list in the Lookups tab. As in Part 1, use the MATCH function and place your function in Column O. Make the title of the column “PromMatch” (in Cell O1) and start your MATCH formulas in Cell O2.

List the promotional codes with errors, the corrected codes, and how many rows of data had the error.

HINT: Try sorting by promo_code. You only need to list each incorrect promotional code once.
HINT: Using the Sort and Filter features in Excel can also help you!

Suppose that you want to know how many orders are for “Baby Blue T-Shirt.” Then, sort the entire table by product_name (Column J).

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>total_product_price</td>
</tr>
<tr>
<td>95</td>
<td>19.41</td>
</tr>
<tr>
<td>96</td>
<td>16.73</td>
</tr>
<tr>
<td>97</td>
<td>21.87</td>
</tr>
<tr>
<td>98</td>
<td>340.38</td>
</tr>
<tr>
<td>99</td>
<td>312.62</td>
</tr>
<tr>
<td>100</td>
<td>218.68</td>
</tr>
</tbody>
</table>

As you can see, Baby Blue T-Shirt begins to appear at Row 98 and ends at Row 115. This indicates that there are 18 (=115-97) orders for this product.

Part 3: Errors in the Total Product Price

Verify if the total product prices in Column I are correct for each line item. We know that the product prices in Column H were recorded correctly, but we’re just not sure the total product price in Column I was calculated correctly, which is the price of the entire order and the amount we bill to customers.

The total product price is the item product price multiplied by the product quantity. For the first line item in the data set, we see this is true (3 x 16.73 = 50.19).

<table>
<thead>
<tr>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>product_quantity</td>
<td>product_price</td>
<td>total_product_price</td>
</tr>
<tr>
<td>3</td>
<td>16.73</td>
<td>50.19</td>
</tr>
</tbody>
</table>

First, see if there are any outliers by creating a scatter plot of total_product_price and find out how many outliers there are in Column I.

Now sort the records by total_product_price in a descending order to identify those outliers.
1) List the lineitem_ids and the total product price for the outliers as listed.

**Second**, check for 0 values for total_product_price.

2) How many 0 values are there in Column I?

**Third**, check to see if there are any other errors in the data set. You can do this by comparing the product_quantity (Column G) X item_product_price (Column H) to total_product_price (Column I). If the item price or the total price is incorrect, then these two values won’t match, indicating a problem.

**HINT**: Use an IF function in Excel. Place your IF function in Column P. Make the title of the column “TotalCheck” (in Cell P1) and start your IF formulas in Cell P2.

**BIGGER HINT**: As an example, if we wanted to compare whether the sum of the values in Cells A2 and B2 were equal to the value in Cell C2, we could do this:

```excel
=IF((A2+B2)=C2,"RIGHT","WRONG")
```

Which says that if the equation is true ((A2+B2)=C2), then display the word RIGHT in the cell. Otherwise, display the word WRONG.

This will allow you to find out which rows have a problem.

3) List the lineitem_ids for each row with an error (with “WRONG”) and the incorrect total_product_price (EXCEPT the outliers and the zeros in Q1 and Q2).

**Submission Instruction**

- Submit both your completed answer sheet and Excel file into Blackboard by Monday, Nov. 16th, 11:59PM EST. This deadline is firm, and the instructor will not take any extraneous circumstance into consideration that occurs to you such as a PC malfunction or network outages.

- Late submission is allowed, but there will be 10% penalty per each 12 hours. For example, if you submit in the morning of Nov. 18th, a 30% penalty is imposed on your submission. Therefore, your submission will be graded zero after the noon of Sat, Nov 21th.