IS0855: Data Science
Assignment 3: Cleaning a Data Set

Task:

You have done such a good job with forecasting and 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.

Make sure you complete the in-class exercise "Finding Bad Data in Excel" before going any further! It will help you!

Deliverables:

This assignment is to be completed individually. Complete the worksheet at the end of this document. Email the completed worksheet with the "cleaned" Excel file to your instructor by the start of class the day the assignment is due.

Evaluation:

You will be graded based on the number of correct answers. There are 10 questions overall.

Part 1: Errors in Product Names

Verify that the product names (Column J) are correct using the master list in the Lookups tab and correct any errors. You can assume the information in the Lookups tab is always right. So if there is a mismatch, the error is in your data set.

To do this, you will use the MATCH formula (use in-class exercise 7.2 as a guide). Place your MATCH calculation in column N of the "Vandelay Orders (Jan)" worksheet. Make the title of the column "ProdMatch" (in cell N1) and start your MATCH formulas in cell N2.

HINT: Using the Sort and Filter features in Excel can also help you! You've used both of those features in the exercises we've done so far in this course.

ANOTHER HINT: Remember, there is a list of correct product names in the Lookups worksheet.

Answer the following questions:

1) How many line items (rows) had incorrect product names?
2) List the products names with errors, listing the incorrect name, the corrected name, and how many rows of data had the error.
   (Try sorting by product_name. You only need to list each incorrect product name once.)

Now fix the incorrect product names in the "Vandelay Orders (Jan)" worksheet.

HINT: Use "Find and Replace" to speed up fixing the errors. You can find this feature under the "Find & Select" button under the HOME tab.

Part 2: Errors in Promotional Codes

Verify that the promotional codes (Column E) are correct using the master list in the Lookups tab and correct any errors. Use the MATCH function and place your function in Column O of the "Vandelay Orders (Jan)" worksheet. Make the title of the column "PromMatch" (in cell O1) and start your MATCH formulas in cell O2.

Answer the following questions:

1) How many line items (rows) had incorrect promotional codes?
2) List the promotional codes with errors, listing the incorrect codes, the corrected codes, and how many rows of data had the error.
   (Try sorting by promo_code. You only need to list each incorrect promotional code once.)

Now fix the incorrect promotional code values in the "Vandelay Orders (Jan)" worksheet.
Remember, there is a list of correct promotional codes in the Lookups worksheet.

Part 3: Errors in the Total Product Price

Verify that the total product price is correct for each line item. We know that the product prices were recorded correctly, we're just not sure the total product price was calculated correctly, which is the price of the entire order and the amount we bill our customers.

To do this, keep in mind a few things:

- 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:

G	H	I
product_quant	**item_product_price**	**total_product_price**
3	16.73	50.19

First, see if there are any outliers by creating a scatter plot of total_product_price.

1) How many outliers are there?
2) Copy/paste/screenshot the plot into the worksheet at the end of this document. Be sure to do this step before proceeding since the chart will change as you make changes. Double check to make sure your chart does not change after you make corrections at the end.

Now sort by total product price to identify those outliers.

3) List the lineitem_ids and the total product price for the outliers as listed.

By looking at the quantity purchased and the total price, it seems unlikely that the item product price is incorrect (this would make the products very expensive!). So correct the total product price for these rows in column I of the spreadsheet. Remember, `total_product_price` is `product_quantity` times `item_product_price`. *Don't delete the rows, fix them.*

Second, check for 0 values for total product price.

4) How many 0 values are there?

Now correct the total product price for these rows in column I of the spreadsheet. *Don't delete the rows, fix them.*

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 in the spreadsheet) **times** the item_product_price (column H in the spreadsheet) to the total_product_price (column I in the spreadsheet). 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:

```
=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.

5) How many line items still have errors (rows with "WRONG")?
6) List the lineitem_id for each row with an error and the incorrect total_product_price.

Now correct the total product price for these rows in column I of the spreadsheet. *Don't delete the rows, fix them.*

Assignment 4: Cleaning a Data Set - Worksheet
Name __Reed Ceniviva_____

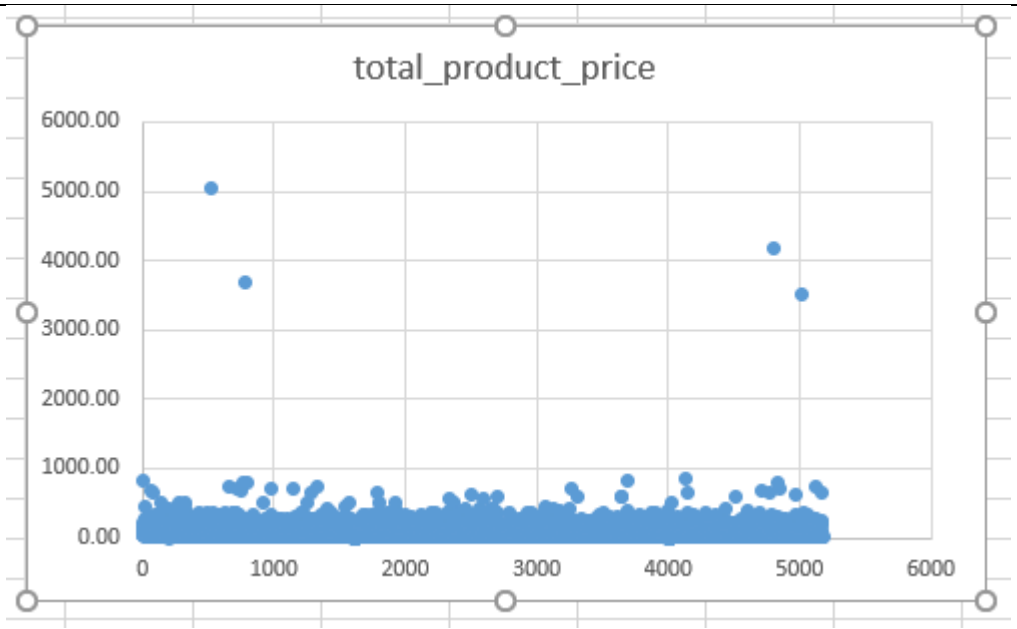
Part 1: Errors in Product Names **Add rows to question 2 as needed (there may be more than two incorrectly named products!).**			
Question	**Answer**		
1	111		
2	**Wrong Name**	**Right Name**	**Number of Rows**
	MOPS Sweatshirt	MOOPS Sweatshirt	10
	Proper Comic T-Shirt	Prop Comic T-Shirt	7
	Goretec Sweatshirt	Gore-tex Sweatshirt	80
	Cotton Deckers Jeans	Cotton Dockers Jeans	14

Part 2: Errors in Promotional Codes **Add rows to question 2 as needed.**			
Question	**Answer**		
1	2438		
2	**Wrong Code**	**Right Code**	**Number of Rows**
	GROUPN	GROUPON	32
	No Promotional Code	No Promo Code	2406

(continued on next page)

Part 3: Errors in Product Price and Total Product Price

Add rows to questions 3 and 6 as needed.

Question	Answer	
1	4	
2		
3	**Line Item ID**	**The listed total product price**
	332	5035.80
	4366	4189.00
	947	3700.00
	4847	3500.00
4	20	
5	7	
6	**Line Item ID**	**The listed total product price**
	550	339.80
	551	150.00
	1095	96.96
	3500	82.35
	2089	35.00
	1847	33.98
	322	25.00