Fox School of Business
TEMPLE UNIVERSITY®

MIS 0855 Data Science (Section 002) – Spring 2015
Assignment #3 – Cleaning a Data Set
Due by Monday, March 16th, 11:59 PM EST

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 5,192 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" (Day 19) before going any further! It will help you.

Deliverables:

Submit both your answer sheet and the "cleaned" Excel file on Blackboard.

Part 1: Errors in Product Names

Verify that the product names (Column J) are correct by 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.

Answer the following questions:

1) How many line items (rows) had incorrect product names?
2) List the products names with errors, the corrected name, and how many rows of data had the same 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.

ANOTHER HINT: The sum of the number of rows in Q2 must be equal to Q1.

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).

So MATCH (E2, ZipCodeStateLookup!$A$2:$A$42524,0) will search for the value in Cell E2 in Cell A2-A42524 in the ZipCodeStateLookup worksheet. Make sure you put $ for absolute referencing.

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 that the promotional codes (Column E) are correct by using the master list in the Lookups tab and correct any errors. As in Part 1, 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.

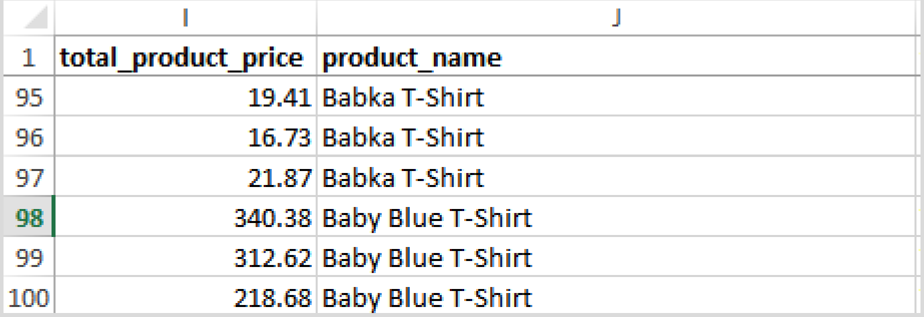
Answer the following questions:

1) How many line items (rows) had incorrect promotional codes?
2) List the promotional codes with errors, 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.

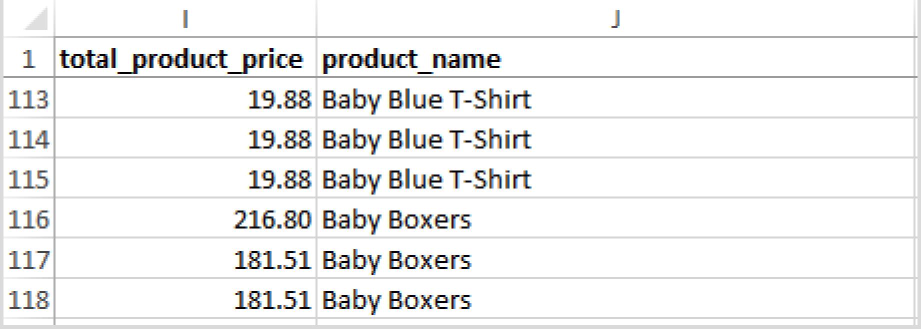
HINT: Using the Sort and Filter features in Excel can also help you!

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).



	I	J
1	**total_product_price**	**product_name**
95	19.41	Babka T-Shirt
96	16.73	Babka T-Shirt
97	21.87	Babka T-Shirt
98	340.38	Baby Blue T-Shirt
99	312.62	Baby Blue T-Shirt
100	218.68	Baby Blue T-Shirt

As you can see, Baby Blue T-Shirt begins to appear at Row 98 and



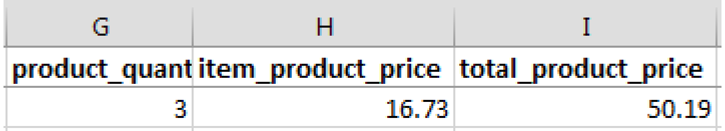
	I	J
1	**total_product_price**	**product_name**
113	19.88	Baby Blue T-Shirt
114	19.88	Baby Blue T-Shirt
115	19.88	Baby Blue T-Shirt
116	216.80	Baby Boxers
117	181.51	Baby Boxers
118	181.51	Baby Boxers

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 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.

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).



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 in Column I?
2) Copy and paste the plot into your answer sheet.

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 in Column I?

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** `item_product_price` (Column H in the spreadsheet) to `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.*

Submission Instruction

- Submit both your completed answer sheet and your cleaned data file into Blackboard by Monday, Mar. 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 Mar. 18th, a 30% penalty is imposed on your submission. Therefore, your submission will be graded zero after the noon of Fri, Mar 20th.