Fox School of Business
TEMPLE UNIVERSITY®

MIS 0855 Data Science (Section 006) – Fall 2017
Assignment #3 – Finding Bad Data in Excel (10% of the Final Grade)
Due by Monday, Oct 16th, 11:59 PM EST

Please read all the instructions carefully.

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 and document the errors in the Excel workbook.

Make sure you complete the in-class exercise "Finding Bad Data in Excel" (Day 18) before going any further! It will help you.

Deliverables

Submit both your answer sheet and Excel file.

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.

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

HINT 1 – First, sort the records by `product_name`.

HINT 2 – Using the MATCH function (see below) will save you time, but it is not required. Make the title of the column "`ProdMatch`" (in Cell N1) and start your MATCH formulas in Cell N2.

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 if the promotional codes in Column E are correct by using the master list in the Lookups tab.

List the promotional codes with errors, the correct codes, and how many records of data had the error.

HINT 1 – Sort the records by `promo_code`.

HINT 2 – 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.

HINT: Using the Sort in Excel can 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 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).

G	H	I
product_quant	item_product_price	total_product_price
3	16.73	50.19

First, see how many zeros there are in `total_product_price`, which by definition cannot be zero.

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

Second, 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.

Make the title of the column "`TotalCheck`" (in Cell P1) and start your IF formulas in Cell P2.

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.

2) List the `lineitem_id`s for each row with an error and the incorrect `total_product_price` (EXCEPT the zeros in Q1).

Alternative HINT – Instead of using an IF function, you can calculate (`total_product_price` - `product_quantity` **X** `item_product_price`). If this is zero, it means that `total_product_price` is correct.

Submission Instruction

- Submit both your completed answer sheet and Excel file by Monday, Oct. 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 Oct. 18th, a 30% penalty is imposed on your submission. Therefore, your submission will be graded zero after the noon of Sat, Oct 21st.