Dear Clients,

I am pleased to present an assessment of the 2017 sales data that you have provided in Excel format. The data is divided into four sheets, including transactions, new customer list, customer demographics, and customer address.

Firstly, I would like to commend you for capturing essential metrics of your client base. However, to ensure the accuracy of the analysis, there are some limitations in the data that need to be addressed.

Here is a summary of my observations:

- Correct Values: The formatting of the columns was not ideal, and data about money was not linked to currency. In addition, the number of significant digits for certain datasets varied widely, and geographic data was not coded in a way to optimize visualization and analysis. Each column needed to be ascribed to text values, standardized dates formatting, and other labels that captured the attributes of the data more effectively.
- Data Fields with Values: Approximately 5% of data is missing from essential fields, such as the brand product data and the product ID column. To capture the extent of the missing data issues, I have created a read-only pivot table, which can be filtered by each column in the transaction sheet to identify non-standardized values. I would recommend a closer inspection of each of these cells to ensure that they do not interfere with calculations on the sums, averages, and other metrics that would be used to analyze the data sets. A screen shot of one of the workbooks demonstrates how Conditional Formatting was used to highlight missing data. The filters that were applied are flexible but include highlighting missing, null, NA, and zero values.
- Up to Date Values: The values seem up to date with regard to the 2017 sales year. However, other information about Product First Sold Date in the Transactions workbook appears to be in a unique format that is not DD/MM/YY and would require further consultations to identify how this data might be useful in analytics.
- Data/Data Interactions: There are additional opportunities to link customer IDs with sales data that have not been optimized as individual data sets are in independent sheets. Having customer IDs in each sheet is a helpful way to combine critical data on past sales and customer demographics that could better target clients.
- Data with Allowable Values: The allowable values metric is something we need to work with the client to refine, given that certain measures, particularly in the new customer list workbook, such as property valuation, rank, and value, are not standardized and make it difficult to optimize how to use this data.
- Duplicate Values: Duplicate values were found in several spreadsheets, and the Transaction sheet contains the most errors as well as the most data. Unfortunately, the transaction IDs are not a good tool to identify unique sales, but the data in Column M “Product First Sold Date” provides a more accurate way to conditionally format data with highlighting to capture duplicates.

Overall, I recommend moving the data to Rstudio for further data analysis, given that the data set pushes the limitations of Excel. I suggest some initial processing in Excel, followed by designing a dashboard in Tableau to capture metrics that align with the clients' goals.

Please feel free to reach out to me if you have any questions or need further clarification on any of the points raised in this email.

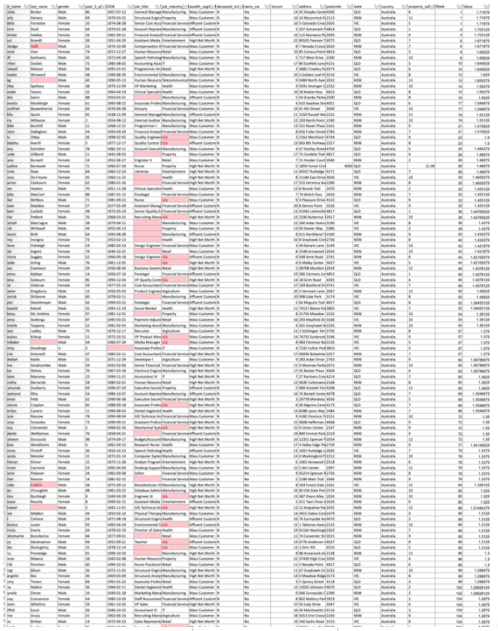
Thank you for providing me with this opportunity to work on your customer data.

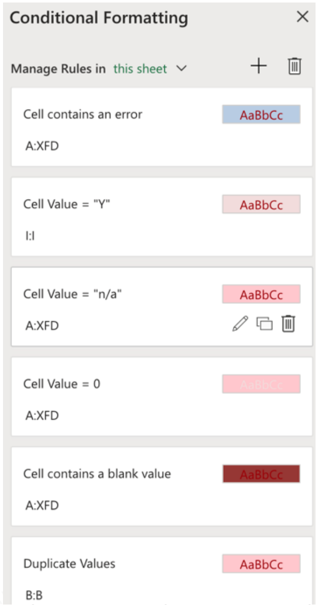
Best regards,

Beth Mara, PhD

*See samples of data cleaning and processing on next page.*

The top table and adjacent window demonstrate the power of conditional formatting for analysis of the customer database. This screen capture shows highlighted cells with null, zero, blank, n/a or errors. The bottom window is a Pivot table (link embedded) that allows screening for outliers in each column. I recommend data cleaning to ensure that these values do not skew averaging or other metrics. Data visualization will also aid in identifying outliers in like we could best be captured using Tableau or PowerBI.

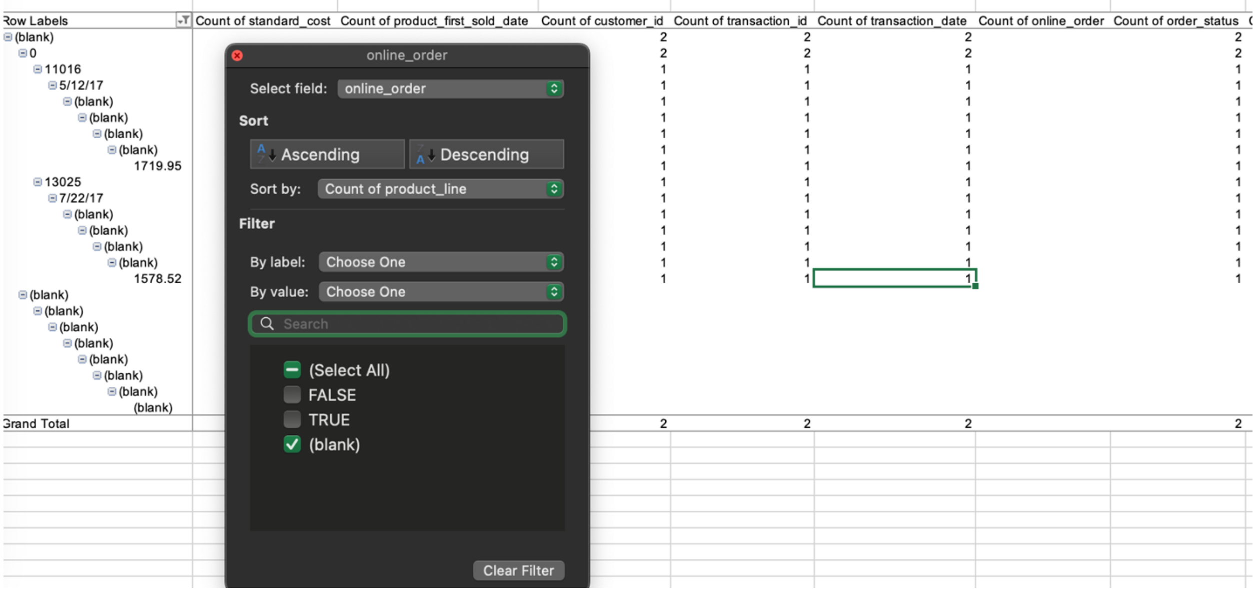




Conditional Formatting

Manage Rules in this sheet

| Rule | Applies to | Format |
|---|---|---|
| Cell contains an error | A:XFD | AaBbCc |
| Cell Value = "Y" | I:I | AaBbCc |
| Cell Value = "n/a" | A:XFD | AaBbCc |
| Cell Value = 0 | A:XFD | AaBbCc |
| Cell contains a blank value | A:XFD | AaBbCc |
| Duplicate Values | B:B | AaBbCc |



| Row Labels | Count of standard_cost | Count of product_first_sold_date | Count of customer_id | Count of transaction_id | Count of transaction_date | Count of online_order | Count of order_status |
|---|---|---|---|---|---|---|---|
| (blank) | | | 2 | 2 | 2 | | 2 |
| 0 | | | 2 | 2 | 2 | | 2 |
| 11016 | | | 1 | 1 | 1 | | 1 |
| 5/12/17 | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| 1719.95 | | | 1 | 1 | 1 | | 1 |
| 13025 | | | 1 | 1 | 1 | | 1 |
| 7/22/17 | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| (blank) | | | 1 | 1 | 1 | | 1 |
| 1578.52 | | | 1 | 1 | 1 | | 1 |
| (blank) | | | | | | | |
| (blank) | | | | | | | |
| (blank) | | | | | | | |
| (blank) | | | | | | | |
| (blank) | | | | | | | |
| (blank) | | | | | | | |
| (blank) | | | | | | | |
| (blank) | | | | | | | |
| Grand Total | | | 2 | 2 | 2 | | 2 |

online_order

Select field: online_order

Sort

Ascending | Descending

Sort by: Count of product_line

Filter

By label: Choose One

By value: Choose One

Search

- ▬ (Select All)
- □ FALSE
- □ TRUE
- ☑ (blank)

Clear Filter