Hi,

We have been going through the datasets shared and have found some data quality issues and problems it may cause for the analysis if the issues are not dealt with properly.

The following table summarises the issues we are facing with datasets and how we are planning to mitigate the issues.

Dataset	Issues
Transactions	1. Missing values are present as follows: a. online_order (360) b. brand (197) c. product_line (197) d. product_line (197) e. product_size (197) f. standard_cost (197) g. product_first_sold_date (197) If we are unable to get values in these columns, then we are planning to drop these rows. After dropping the rows, we can see that we have dropped only 2.77% of the data. Hence, effect on the analysis after dropping these rows will be negligible. 2. The values in product_first_sold_date column is unable to be interpreted. It should be a date value whereas we have some numbers. We tried converting the numbers to dates and it was giving ambiguous dates. We would request to get the proper dates in this column if possible so that we can use the column data for the analysis. If we are unable to get that information, we might need to drop the column for the analysis.
CustomerDemographic	1. The column "default" is having some ambiguous values. It would be better if we get some information on what column and the values are representing. For the analysis we won't be able consider this data. 2. Missing values are present as follows: a. last_name (125) b. DOB (87) c. job_title (506) d. job_industry_category (656) e. tenure (87) We are more worried about the missing values in the columns DOB, job_title and job_industry_category as we feel this will

	help to yield meaningful insights during the analysis. And here as in Transactions table we cannot drop the records with missing values as there will be a lot of data lost if we do that. So, we request for this data and the other option we are having is a vague imputation of the values.
CustomerAddress	1. The table is fine with no missing and ambiguous values. 2. The details of customers with customer id 3, 10, 22 and 23 are missing when compared with customer demographic and transactions table. The records belonging to these customers sums up to 31 and hence even if we drop the records it is a negligible data loss in case if we are not able to get the required data.

When we have combined the datasets to identify if there is anymore information missing, we found out that the customer id, 5034 exists only in the transactions table and no information on this customer is available. There are 3 transactions under the customer id and if we are not able to get the information on this customer, we are planning to drop these records as well from the analysis.

Overall, the data shared with is pretty much fine with respect to our data quality framework apart from what we have found above.

We are hopeful that you will be able to help us to get the data and resolve some of the data quality issues above.

Thank you,

Joyal Joy Madeckal