Dear Sprocket Central Team Member,

Thank you for providing the Customer data to our team. We have had a chance to review the data. Below is a summary of what was received – please let us know if it does not match your understanding:

Data set	Number of Records	Number of Product IDs	Number of Customer IDs
Transactions (from cal. Year 2017)	20,000	101	3,494
Customer Demographics	4,000	n/a	4,000
Customer Addresses	3,999	n/a	3,999
New Customers	1,000	n/a	

Some data quality problems were uncovered. The information provided here should offer insight and allow you to take steps towards collecting cleaner data.

Data completeness and validity were the areas of worst quality across the information we received. Please see below for notes on each of the datasets provided, along with notes on our suggested fixes:

Missing information:

About 2% of orders are **missing the 'online order' status**. An additional 1% of orders **are missing information about the products being sold** – brand, product line, product class, cost, and product id. While that is a relatively small percentage of the transactions, knowing the source of an order is important. It would be worth investigating these transactions with missing statuses and determining where the leak is. If these are incomplete orders, they should be removed from the data set.

Employment information (job title/industry category) was not captured for over 10% of new customers. If Sprocket plans to use that information for targeting purposes, this is a significant of missing data. Requiring that field or providing additional options for customers to select during checkout could fill that gap.

As you can see in the above table, **we have more customers in the Demographic table than we have in the Transactions table.** Please ensure all data is from the same time period – **we will assume customers in the Demographic table, not appearing in the Transaction table had no sales in the period** – this will be included in our analysis.

Data Validation:

There are a few instances where **there are multiple values being entered to denote the same thing.** Gender and State in the Demographics and Address tables are the prime examples (e.g. Female being represented as "Female", "Femal", and "F"). Enforce the use of standard values during customer data entry using dropdown fields in checkout forms.

There are **fields where the data type is off**. Ensure data fields have constraints on what types they will accept (e.g. date type for dates).

Understanding the process that customers use to input information would allow KPMG to better assist your team in finding solutions to these data integrity problems. Moving forward, we will continue our analysis. Questions and assumptions that arise during our process will be documented. After we complete our work, it will be great to meet with your team and ensure we are aligned.

Best regards,

Zach Zazueta