

Dear Sprocket Central,

Thank you for providing the Customer data to our team. We have had a chance to review the data and have found some data quality problems. I hope the information provided here will offer insight and allow you to take steps towards collecting cleaner data.

Data completeness, accuracy, 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:

Transactions:

- About 2% of orders are missing the 'online order' status. 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.
- 1% of orders are missing information about the products being sold – brand, product line, product class, cost, and product id. If these are incomplete orders, they should be removed from the data set.
- Product_first_sold_date is being processed as a float – it would be best to process that as a date data type.

New Customers:

- 3% of new customers did not provide a last name. Personal information should be required fields during order placement; not having last names could lead to difficulty with shipping or other customer service needs.
- 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.
- There are customers included in this dataset which show 0 purchases in the past 3 years. The purpose of this data is to capture new customers, so prospective customers with no sale should not be included in this list.

Customer Demographic:

- Once again, there are customers without last names provided – this can lead to difficulties with customer support.
- DOB dates back to the year 1843 – we would suggest a data validation field be used as customers provide their information.
- There were two deceased customers in the list – they should be removed from any future outreach.
- There were multiple ways to enter gender (Male/M, Female/Female/F). This should be standardized, again with a validation feature.

Customer Address

- There were multiple ways that states labels were input (NSW/New South Wales, VIC/Victoria). Like gender, there should be a standardized data entry format.

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.

We look forward to working with you to collect the cleanest data possible.

Best regards,

Zach Zazueta