

Hello [Contact],

Thank you for providing us with three datasets from Sprocket Central Pty Ltd. My name is Kashish Saxena, and I am a team member of the Forage KPMG Virtual Participation Data Analytics, Information & Modelling team and I am writing to provide recommendations for improving the data after assessing the quality of your data. Below were some of the issues identified in the data and our recommendations for improvement are as follows:

- 1. Inconsistent values for the same attribute:** The gender column in the Customer Demographics has inconsistent values. The F's can be replaced by Female and the M's can be replaced by Male and the U's be replaced by null values as it also does not impact the analysis results significantly. The Femal can be manually replaced by Female as it is only one value. The default column has random values and is irrelevant to the analysis. So, it can be dropped altogether.

In the Customer Address dataset, there are inconsistent values in the state column. New South Wales is represented by both NSW and New South Wales. Thus, NSW can be replaced by New South Wales, and similarly, VIC can be replaced by Victoria.

Mitigation recommendation: Ensure that forms and tables have constraints on input values. Having inconsistent values makes analysis difficult and generates inaccurate results. Alternatively, have a drop-down selection to avoid such errors.

- 2. Missing values:** In the Customer Demographics dataset, there are missing last names which can be replaced by a null value as it would not have any significant impact on the analysis. There are some missing job titles and can be replaced by null. There are also missing tenures. They are an important part of the analysis and can be replaced by the average of the filled values.

In the Transactions dataset, there is missing data in the online order column, which can be replaced by NULL as it does not affect the sales. Line 1004 can be omitted as it contains important missing data that is crucial for analysis and can alter results.

Mitigation recommendation: Ensure that forms and tables have a mandatory requirement message for all skipped input fields.

- 3. Inconsistent data type:** In the datasets, numeric values have general data type, and customer IDs are numeric, which should be general (as ID variables are categorical)

Mitigation recommendation: Ensure that forms and tables have constraints on input values. Having inconsistent data types makes analysis difficult and generates inaccurate results.

Moving forward, the team will continue with data cleaning and pre-processing for the purpose of analysis. Any assumptions will be documented, and we will reach out if we have any questions along the way. Once this process is complete, we will share the updates in a meeting with your team members to ensure that the cleaned data and any assumptions are aligned with Sprocket Central's understanding.

Kind regards,

Kashish Saxena (Junior consultant, Information and modeling team)