Hello Kathleen,

Thank you for providing us with the dataset from Sprocket Central Pty Ltd. The below table tries to shed some light on the data quality of the four datasets forwarded to me. Please do let me know if you have any queries.

A Quick Rundown:

	Accuracy	**Completeness**	**Consistency**	**Validity**
Transactions		● Missing values.		● Improper format in 'product first sold' column.
New Customer List		● Missing values.		
Customer Demographic	● Inaccurate DOB.	● Missing values.	● Inconsistency in gender entries.	
Customer Address		● Missing values.	● Inconsistency in State Column.	

A Deeper Dive:

Below are the more in-depth description of the data quality issues found and the methods used to mitigate the same. Further explanation is also provided so that the data quality of any future dataset from Sprocket Central Pty. Ltd would not be imperilled.

Accuracy:

Some of the values of DOB were inaccurate for 'Customer Demographic'

Mitigation: Removed the inaccurate data point, Calculated age using the DOB entries column.

Recommendation: Try adding a column for age, hence the Customer would not be made to enter the DOB, which is long and arduous work. Hence giving an option for age would mean less prone to mistakes from the customer.

Completeness :

A lot of missing entries present in all four datasets. (NaNs entries)

Mitigation: Filtered out all the missing blanks from the dataset.

Recommendation: Ensure the data is up-to-data, as filtering out the data would take away the importance of some crucial data points

Consistency :

Inconsistency in the gender entries in 'Customer Demographic' dataset and State column in 'Customer Address' Dataset.

Mitigation: Filtered all the M and male entries all into Male Catagory and filtered all F and Femal entries into Female Catagory. Also Catagoriesed all U and Unknown into Unknown. For state column in Customer address dataset, I have filtered out 'NSW', 'QLD' and 'VIC' to 'New South Wales', 'Queensland' and 'Victoria'

Recommendations: Since the entries are limited in the options, having a dropdown option could help alleviate the issue.

Validity :

Improper format in Product first sold column in 'Transaction' Dataset.

Mitigation: Converted the column from "int64" to "date" format

Recommendation: Set up the column in the "date" format in the intial stage, so that any future entries would be recorded in "date" format.

That captures all the data quality issues that came across during the initial stage of data quality analysis. The recommendation would provide a simple and effective way to improve the data quality of the dataset. These suggestions will not only improve the quality of the data but also improve the final accuracy of the analysis and also improve the level of analysis of the data.

Thank you.

Regards
Sanjaya J Shetty