

Hi Sir,

Thanks for providing the datasets from Sprocket Central Pty Ltd. I have reviewed the quality of the datasets and table below is the summary of the quality test.

	Transactions	Customer Address	Customer Demographic	New Customer List
Accuracy			Date of Birth: <i>Not Accurate</i>	
Completeness	1.Online Order Column: <i>Blanks</i> 2.Brand Column: <i>Blanks</i>		1.Date of Birth: <i>Blanks</i> 2.Job Title: <i>Blanks</i>	1.Date of Birth: <i>Blanks</i> 2.Job Title: <i>Blanks</i>
Consistency		States: <i>Not Consistent</i>	Gender: <i>Not Consistent</i>	
Currency				
Relevancy				
Validity	Product first Sold Date: <i>Format</i>			1.Past 3 year purchase column: <i>Data in text format and not in Numeric format.</i> 2.Postcode: <i>Data in text format and not in Numeric format</i> 3.Property valuation: <i>Data in text format and not in Numeric format</i>
Uniqueness				Rank Column: <i>Duplicated</i>

The in-depth details regarding the Quality issues mentioned above along with the strategies to mitigate these issues.

1. Accuracy

Issue: Date of Birth is Not Accurate in the Customer Demographic Dataset.

Mitigation: Apply filters to the Date of Birth Column and remove the data's that are wrong

2. Completeness

Issue: The Transactions, Customer Demographic, New Customer List Dataset's has columns with blank data.

Mitigation: Filter out these Blank Data's from these columns

3. Consistency

Issues: The States column from the Customer Address dataset were not consistent as New South Wales State in Australia was used in the dataset as NSW and New South Wales. Similarly, Victoria State and VIC both were used.

The same issue was also found the Gender column from the Customer Demographic Dataset where the gender was represented as Male, Female, M, F, Femal.

Mitigation: Apply Filter to both these columns and then replace M to Male, F & Femal to Female. Similarly, New South Wales to NSW and Victoria to VIC

4. Currency

Issues: No issue found

5. Relevancy

Issues: No issue found

6. Validity

Issues: In Transaction dataset the format of the date in Product first sold date column. Similarly, in New Customer List Dataset columns- Past 3 years purchase, Postcode, Property Valuation had the data wrongly in the Text format

Mitigation: Changing the given format of the date in Product first sold column to short date format.

Format the data in the Past 3 years purchase, Postcode, Property Valuation columns of New Customer List Dataset from Text to Numeric Format

7. Uniqueness

Issues: Duplicate Rank columns in the New Customer List Dataset

Mitigation: Delete the entire column that is not affecting the previous or the next columns.

So, from the Data Quality Analysis done based on the Data Quality Framework the above issues were found and we believe the strategies suggested to mitigate these issues are easy enough to solve the issues found and to be utilized for future data preparations.

Feel free to contact us if there are any questions regarding the Data Quality issues found or mitigation strategies suggested.

Regards,

Melvin Alex Mathew Andoor