Data Quality Assessment Report Email Draft to Client:

Sai Vamsi Krishna Kamatham

Dear [Client point-of-contact],

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if you have any queries surrounding the issues presented.

Table name	**Accuracy**	**Completeness**	**Consistency**	**Currency**	**Relevancy**	**Validity**
Customer Demographic	DOB inaccurate, Age: missing.	Job title: blanks Customer id: incomplete	Gender: inconsistency	Diseased Customers: Filter out	Default Column: Delete	
Customer Address		Customer id: Incomplete	States: Inconsistency			
Transaction Data	Profit: missing	Customer id: incomplete Online Order: blanks Brand: blanks			Cancelled Status order: Filter out	List Price: Format Product sold date: Format

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

ACCURACY ISSUES

- **DOB was inaccurate for "Customer Demographic" and missing an age column. Missing a profit column for "Transactions"**

  Mitigation: Filter out outlier in DOB.

  Recommendation: Create an age column, allowing for more comprehensible data and easier to check for errors. Create a profit column in "Transactions" to check accuracy of sales.

  Creating additional columns for age and profit will allow for easier identification of errors. The profit column will assist in future monetary analysis.

COMPLETENESS ISSUES

- **Additional customer _ids were inconsistent among "Customer Demographic," "Customer Address," and "Transactions."**

  Mitigation: Filter all customer_ids from 1 to 3500

  Recommendation: Ensure tables are up to date (from the same time period). For our model, only customer_ids from 1 to 3500 will be used as they have complete data.

  The data received may not be in sync across all spreadsheets, with incomplete data the analysis results may be skewed. This is a 'completeness' issue, to prevent future occurrences it is encouraged to cross check spreadsheets and sync data.

- **Blanks in job_title for "Customer Demographic," in online_order and brand_column for "Transactions"**

  Mitigation: Filter out 'blanks' for job_title, online_order, and brand_column.

Recommendation: Simplify job_title to another category such as industry_industry or provide dropdown options for job_title. Provide dropdown options for online_order and brand_column.

Blanks are treated as incomplete data and can skew further analysis results. The addition of dropdown options will allow us to have more complete data and will result in more accurate analysis.

CONSISTENCY ISSUES

- **Inconsistency in gender for "Customer Demographic" and states for "Customer Address" respectively**

  Mitigation: Filter all 'M' under category of 'Male,' filter all 'Femal' and 'F' under 'Female' for gender. Filter all 'New South Wales' to 'NSW' and 'Victoria' to 'VIC' for states.

  Recommendation: Create dropdown options for 'Male," 'Femal," and 'U' in gender. Create dropdown options for all state abbreviations.

  Dropdown options minimize manual entry and human error. Allows for increase of consistency of terminology. Gender identity can be a sensitive topic, proceed with caution when creating options.

CURRENCY ISSUES

- **People that are 'Y' in deceased_indicator are not current customers for "Customer Demographic."**

  Mitigation: Filter out customers checked 'Y' in deceased _indicator.

  Recommendation: Can be difficult to check for deceased customers, but once this information is received one should update data accordingly.

  Deceased customers are not current customers, removing them from data will increase the currency of data and will result in more accurate estimates in future analysis.

RELEVANCY ISSUES

- **Lack of relevancy or comprehensibility in default_column for "Customer Demographic" and order_status for "Transactions"**

  Mitigation: Deleted Metadata in default_column. Filter out 'Cancelled' order_status.

  Recommendation: Check for incomprehensible Metadata and delete or format to make comprehensible.

  'Cancelled' order_status is irrelevant information for future analysis, as it can skew data — for example total number of customers per annum will be an overestimate. Victoria' to 'VIC' for states.

VALIDITY ISSUES

- **Format of list_price, product_sale_date for "Transactions"**

  Mitigation: Format product_sale_date to short date format, format list_price to currency.

  Recommendation: Set up columns so that formats such as price and decimals are already in place when entering new data.

  Allowable values will make data to be interpreted more easily. Formatting into price and allowing for I either 2 or 3 decimals placed consistently will increase readability. This will reflect positively on speed and accuracy of analysis for business decisions.

This summarizes all data quality issues discovered through the first stage of the data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis.

Please let us know if you have questions regarding mitigation or any data quality issues.

Moving forward, the team will continue with the data cleaning, standardization and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central's understanding. identified.

Kind regards,

Sai Vamsi Krishna Kamatham