Re: Data analysis – Data cleaning update

Dear [Client point-of-contact]

Thank you for providing us with the four datasets from Sprocket Central Ltd.

The summary table below highlights key data quality issues we have discovered in the data cleaning process. Please let us know if you have any queries concerning the issues presented.

Dataset	Accuracy	Completeness	Consistency	Currency	Relevancy	Uniqueness	Validity
(Dataset 1) **Transactions**		Data has 1542 missing values	Format of list price and std cost	Data is up to date		Data has no duplicate values	Format of Product first sold date is not correct
(Dataset 2) **New Customer List**		Data has 317 missing values	Unspecified gender 'U'		Irrelevant: unnamed columns	Data has no duplicate values	
(Dataset 3) **Customer Demographic Data**	DOB column	Data has 1461 missing values	Format: gender		Corrupted: default column deleted	Data has no duplicate values	
(Dataset 4) **Customer Address**		Data has no missing value	Format of state NSW and new south Wales			Data has no duplicate value	

DATA QUALITY ASSESSMENT

1. Accuracy Issues:

- **DOB column in "Customer Demographics" had few inappropriate values.**

Mitigation: Filter out outliers in DOB.
Recommendations: Create an age column, allowing for more comprehensible data. Set a range for values of age column using Tableau Prep Builder.

2. Completeness:

- **Null Values in many columns.**

Mitigation: Please ensure that all tables are free from Null values.
Recommendations: Fill null values using appropriate measure like mean, mode etc

3. Consistency:

- **Consistency should be maintained in all the values of each attribute.**

Mitigation: Change all 'F','Femal' to Female. All 'M' to 'Male' for 'gender' column. Change all 'New South Wales' to 'NSW' and 'Victoria' to 'VIC', in state column of Customer Demographics.
Recommendation: Edit values using replace function in python

- **Inconsistent data type for the attributes**

Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from string.

Recommendation: To avoid different representations of the same value, the data type should be categorical rather than a variable text field. Dropdown options minimise inconsistencies and human error in manual entries by different personnel and improves the data interpretability and readability. As gender is a protected characteristic, those identified as others may fall under the category of 'U'.

4. Currency:

- **Records with 'Y' in deceased_indicator in "Customer Demographics" are irrelavent.**

Mitigation: Retain only those customers who have a 'F' in deceased_indicator.

Recommendations: Keep only records with 'F' in deceased indicator

5. Relevancy:

- **Records with 'Cancelled' order_status in "Transactions" data are irrelevant.**

Mitigation: Retain only those customers who have a 'Approved' in order_status.

Recommendations: Keep only records with Transactions" order_status

6. Validity:

- **Ensure all the columns are assigned proper data types.**

Mitigation: Format columns to relevant data types.

Recommendations: use astype function in python

Other data quality issues

- There were many missing datapoints across various features/columns.
- Some of the data was also out of sync, i.e., there was some mismatch between the datasets.
- Inconsistent data types were used for the same attributes, e.g., integer for some fields and float for others which can introduce unintended bugs due to discrepancy in precision.
- Mitigation: If the number of null-value is small, I have filled the records using appropriate statistical methods. Otherwise, if the number of null-value is significant, the records have been dropped from the master datasets. The only exception I made was if the sample size is small and the datapoints are critical. This achieved a standardisation of all fields to achieve constraints on the permitted data types.

Please let us know if you have comments or questions on the above as I would be happy to discuss to ensure that all assumptions applied align with Sprocket Central Ltd.'s understanding.

Kind regards

Sagar Sarolia

Data Analytics Team
KPMG