Dear Manager,

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 cleaning process. Please let us know if you have any queries concerning the issues presented.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Accuracy** | **Completeness** | **Consistency** | **Timeliness** | **Relevancy** | **Uniqueness** | **Validity** |
| 1. Customer Demographics | DOB  Age missing | Job title blank  Customer id | gender | Deceased customer | Default column deleted |  |  |
| 2. Customer Address |  | Customer id | States |  |  |  |  |
| 3. Transactions | Profit missing | Customer id, online order, Brand blanks |  |  | Cancelled orders filtered |  | List price and product sold date format |
| 4. New Customer |  |  |  |  |  |  |  |

In depth descriptions of data quality issues discovered and methods of mitigations used.

I will describe the methods used to mitigate any data inconsistencies and recommendation with explanation the improve the accuracy of the data source to avoid further data quality issues in the future.

Data issues, strategies to mitigate these issues. Recommendations going forward to mitigate current data quality concerns. This will improve the accuracy inform any future business decisions.

I have joined the keys between the tables.

1. Customer demographics
2. Customer addresses
3. Transaction data in the past three months

According to the data quality dimensions framework

* Accuracy - Correct values
* Completeness - Data fields with values
* Consistency - Values free from contradiction
* Timeliness - Values up to date
* Relevancy - Data item with value meta-data
* Validity - Data containing allowable values
* Uniqueness - Records that are duplicated

Accuracy issues

DOB inaccurate for dataset 1, missing an age column, missing a profit column for dataset 3

Mitigation: filtered out outliers in DOB

Recommendation: Create an age column, for more comprehensible data to check for errors. Create a profit column in dataset 3 to check for accuracy of sales.

Completeness issues

Additional customers ids inconsistent among all datasets

Mitigation: filter all customer\_ids from 1-3500

Recommendation: ensure tables are up to date. Only complete data will be used.

If the data is not in sync across all spreadsheets, data analysis with incomplete data may skew the results. To prevent future occurrences, it is advised to cross check spreadsheets for completeness.

Blanks in job title for dataset 1, online order and brand column for dataset 3.

Mitigation: filtered out blanks for job title, online-order and brand column

Recommendation: provide dropdown options for job title, online order and brand column.

Blanks were treated as incomplete data and may skew the result of data analysis. The use of pre-defined options will allow for more complete data.

Consistency

Inconsistency in gender for dataset 1 and dataset 2.

Mitigation: filtered all variations of Men under the category of ‘Male’, and all variations of Women under ‘Female’. Filtered all names of states to abbreviations, e.g. ‘Victoria’ to ‘VIC’.

Dropdown options minimises inconsistencies and human error in manual entries by different personnel. As gender is a protected characteristic, those identified as others may fall under the category of ‘U’.

Timeliness

Those reported deceased are not current customers for dataset 1.

Mitigation: filtered out the customers marked as deceased.

Recommendation: it may be difficult to verify this information, but this should be updated promptly as soon as possible.

Relevancy

Lack of relevancy or comprehensibility in default column for dataset 1 and order status for dataset 3.

Mitigation: Dropped the metadata column. Filtered out cancelled order status.

Recommendation: remove or reformat any incomprehensible metadata to make it comprehensible.

Cancelled order status was ignored as it was deemed irrelevant.

Validity

Format of list price, product sale date for dataset 3

Mitigation: standardised the product sale date, and convert list price to currency.

Recommendation: Set the correct number of precision (2-3 d.p.) and value format for entries. Enforce restrictions on allowable values will improve the data interpretibility and readability.

‘Customer ID’ datapoints in dataset 2 and dataset 3 but missing in dataset 1.

Mitigation: Ensure that all datasets are from the same time period, otherwise any duplicate or missing data records may skew the data analysis. The datasets have been merged to a single master dataset in a consistent format.

Many missing datapoints across various columns.

Mitigation: If the number of null-value is small, these have been filled using appropriate statistical methods. Otherwise, if the number of null-value is significant, these records have been dropped from the master datasets. The only exception is if the sample size is small and the datapoints are critical.

Inconsistent data value and or format for the same attribute. ‘VIC’, ‘Victoria’, ‘V’.

Mitigation: Find all alternative values and replaced all alternative spellings into abbreviations to ensure consistency across the datasets. To avoid different representations of the same value, the data type should be categorical rather than a variable text field. Gender records have been filled according to the names.

Inconsistent data type for the same attribute e.g. integer for some fields, float for others which can introduce unintended bugs due to discrepancy in precision.

Mitigation: standardise all fields to have constraints on the permitted data types.

This summarises all data quality issues discovered through the first stage of the data quality analysis.

Moving forward the team will continue with the Extract, Transformation and Load process for the purpose of model analysis in stage 2 and stage 3.

Please let us know if you still have outstanding questions, I would be happy to ensure that all assumptions aligns with Sprocket Central Ltd.’s understanding.

Kind regards,

Man Kit Yip

KPMG consultant\*

\*Disclaimer: this is a hypothetical scenario in the role of a data and insights consultant.