Dear Team

Sprocket Central Pty Ltd,

I hope this email finds you well.

Thank you for providing us the datasets from Sprocket Central Pty Ltd. This is Jiarui Xu, from the KPMG Data & Analytics team. This email is in response to the preliminary data assessment of the dataset provided to us. The table below summarises the data quality issues that our team identified and mitigation strategies based on the Data Quality Dimensions.

**Summary Table**

|  | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| --- | --- | --- | --- | --- | --- | --- |
| Transactions | * Profit:   missing | * Brand: blanks * Customer id: incomplete * Date: blanks |  |  | * Cancelled order status: filter out | * List price: format issue |
| NewCustomerList |  | * Customer id: incomplete * DOB: blanks * Job title: blanks |  |  | * Unnamed columns: filter out | * Postcode: format issue * Value: format issue |
| Customer  Demographic | * DOB: inaccurate * Age: missing | * Job title: blanks | * Gender: inconsistent | * Deceased customer: inappropriate | * Default columns: irrelevant |  |
| CustomerAddress |  | * Customer id: incomplete | * State: inconsistent |  |  | * Postcode: format issue |

|  | No. of Records | Distinct Customer IDs | Receiving Date |
| --- | --- | --- | --- |
| Transactions | 20000 | 3501 | 2022-07-11 |
| Customer Demographic | 4000 | 4000 | 2022-07-11 |
| Customer Address | 3999 | 3999 | 2022-07-11 |

**Accuracy**

* DOB of “CustomerDemographic” was inappropriate, profit and age\_ column were missing for “Transactions” and “CustomerDemographic” datasets

Mitigation: Filter out the outsider for DOB.

Recommendation: Insert an age\_column for increasing the data accuracy. Create a profit\_column to provide a more clear view of business conditions.

**Completeness**

* Customer\_ids and job\_title were incomplete

Mitigation: Filter and keep customer\_ids from 1 to 3501.

Recommendation: By comparing these four datasets, the number of distinct customer\_id was 3501 and other data were overlapped in the datasets. First try to get all the customer information based on the “Transactions” dataset, then only use included customers with complete data.

**Consistency**

* Gender and state are inconsistent for “CustomerDemographic” and “CustomerAddress” respectively

Mitigation: Filter “M” to “Male”, “Femal” and “F” to “Female” and filter out “U” category.

Recommendation: When gender\_column is filtered, only rows that meet the filter criteria will display and other rows will be hidden to maintain the uniform format.

**Currency**

* Deceased customers should not be chosen as current customers in “CustomerDemographic”

Mitigation: Filter out deceased customers.

Recommendation: Always check customer status in further data analysis to make sure that the datasets contain the latest information.

**Relevancy**

* Unnamed and default columns were irrelevant for “NewCustomerList” and “CustomerDemographic”

Mitigation: Filter out garbled data in “CustomerDemographic” dataset and delete unnamed columns.

Recommendation: Notes can be added for better understanding of data valuation.

**Validity**

* Postcode didn’t conform a correct data format

Mitigation: Convert postcode from general format to number format and decrease decimal places of value\_column for “NewCustomerList”.

Recommendation: Conduct unit and integration testing in datasets when updates are deployed. Validating that data is correct should be an integral, ongoing process and will ensure confidence in the data.

Please do not hesitate to contact us if you have any questions regarding identified issues or mitigation.

Kind regards,

Jiarui Xu