To whom it may concern,

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 the figures are not aligned with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | No. of records | Distinct Customer IDs | Date Data Received |
| Customer Demographic | *4000* | *4000* | 31/07/2023 |
| Customer Address | *3999* | *3999* | 31/07/2023 |
| Transaction Data | *20000* | *3495* | 31/07/2023 |

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 re- occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Customer Demographic | Customer Address | Transaction Data |
| Accuracy | DOB: Inaccuracy  Job Industry Category: Misspelling |  |  |
| Completeness | DOB: Blanks  Job Title: Blanks  Job Industry Category: Blanks  Tenure: Blanks  Customer IDs: Not in synchronization | Customer IDs: Not in synchronization | Standard cost: Blanks  Brand: Blanks  Product Line: Blanks  Product Class: Blanks  Product Size: Blanks  Product First Sold Date: Blanks  Customer IDs: Not in synchronization |
| Consistency | Gender: Inconsistency | States: Inconsistency |  |
| Currency | Deceased Customers: Filter Out |  |  |
| Relevancy | Default: Exclude Field |  | Order Status: Exclude Cancelled |
| Validity |  |  | Product First Sold Date: Format |

Detailed descriptions of the discovered data quality issues and the steps taken to mitigate these issues in later phases of our analysis is given in the following sections of this email. Explanations and strategies to avoid these issues are also included. The following recommendations will improve the quality of data used to take important future business decisions for Sprocket Central Pty Ltd.

Accuracy Issues

*Issues:* The “DOB” field in Customer Demographic dataset contain inaccurate values. The “Job Industry Category” field contains misspelled category values

Mitigations: Filtered out outliers in the dataset. Fixed Misspelled categories

Recommendations: Include additional fields e.g “Age” in Customer Demographic Dataset and “Profit” in Transactions Dataset for easier detection of outliers.

Completeness Issues

*Issues:* “DOB”, Job Title, Job Industry Category and Tenure fields in the Customer Demographic Dataset contains missing values. ‘Standard Cost’, ‘Brand’, ‘ Product Line’, ‘Product Class’, ‘Product Size’ and ‘Product First Sold Date’ fields in the Transactions Dataset contain missing values for a number of transactions

*Mitigation:* Datasets were not in synchronization with join keys, ‘Customer IDs’ found to be inconsistent among datasets. Filtered out datasets containing incomplete data

*Recommendation:* Ensure datasets are complete and update regularly.

Consistency Issues

*Issues:* ‘Gender’ field in the Customer Demographics dataset contains acronyms and misspellings of the major categorical values, the ‘States’ field in the Customer Addresses contains multiple versions to identify the same categories. Also, ‘Customer IDs’ in the Transactions and Customer Addresses datasets indicate missing records in the Customer Demographics dataset.

*Mitigation:* Replaced extended values and acronyms in the necessary cases using regular expressions. Customers with complete data shall be considered.

*Recommendation:* Ensure consistency of terminologies across datasets for categorical fields. Regularly run sanity checks on datasets to reduce trails of human error.

Currency

*Issues:* The customer demographic datasets contain records pertaining to deceased

*Mitigation:* Filtered out demographic data for deceased customers by checking the ‘Deceased indicator field’

*Recommendation:* Can be difficult to track. Update datasets regularly and filter out records for more accurate analysis.

Relevancy

*Issues:* Incomprehensible ‘default’ field in the customer demographics dataset, cancelled orders present in the Transactions dataset which may not be relevant and may lead to inaccuracies and incorrect forecasts.

*Mitigation:* Dropped the ‘Default’ field and dropped records pertaining to cancelled orders in the respective datasets.

*Recommendation*: Delete or format customer metadata in comprehensible formats.

Validity

*Issues:* ‘Products First Sold Date’ incorrectly formatted as numeric data, may lead to errors in analysis.

*Mitigation:* Field Converted to date type

*Recommendation:* Format Data fields to correct date types

Acting upon the aforementioned recommendations in the first phase of analysis shall allow Sprocket Central Pty Ltd to make better business decisions, increase revenue and enhance operational capabilities. KPMG shall continue providing its expertise to provide a holistic view of your data in future phases of the analysis to enable your company to grow further.

Kind Regards,

Lucy Adhiambo

Moving forward, the team will continue with the data cleaning, standardisation 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.

Kind regards,

[Junior Consultant Name]