**Respected Manager,**

**Sprocket Central Pty Ltd.**

Thank you for coming to KPMG for help in analyzing their data to optimize marketing strategy. The below table highlights the summary statistics of the data has been received.

|  |  |  |  |
| --- | --- | --- | --- |
| Datasheet | Number of records | Distinct Customer ID | Date Data Received |
| Customer Demographic | 4000 | 4000 | 09/07/2020 |
| Customer Address | 4003 | 4003 | 09/07/2020 |
| Transactions | 20000 | 3500 | 09/07/2020 |

Notable data quality issues that were encountered are as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Datasheet | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity | Uniqueness |
| Customer  Demographic | DOB:  Inaccurate | Last names:  Blanks  Job title:  Blanks  Customer ID:  incomplete | Gender: inconsistent | Deceased Customers: filtered out | Default:  Irrelevant values |  |  |
| Customer  Address |  | Customer ID: incomplete | State: inconsistent |  |  |  |  |
| Transactions |  | Customer ID: incomplete  Online Order: blank  Brand: blanks |  |  |  | Product first sold date: format  List price: format |  |

Here is an in-depth analysis of the issues discovered in the datasets. Explanations of the problems and how to mitigate those issues have been given below. Please go through it to ensure prevention of further data quality issues. Following the recommendations given below will improve the data quality of datasets which will also help in better business decisions for Sprocket Central Pty Ltd.

**Accuracy**

**DOB is inaccurate in “Customer Demographic” Dataset.**

*Mitigation: Filtered out the outliers in DOB column.*

*Recommendations: Create “Age” column in Customer Demographic to get clearer view of the customers and also add “Profit” column in Transactions for actual details of sales per transactions.*

*Explanation: Creating additional age column will provide outliers in granular way and it will be easy to filter out them. Adding a profit column will help in future analysis of sales.*

**Completeness**

**Customer ID is incomplete in all three datasets.**

**Last name & Job title has blanks in “Customer Demographic” dataset**

**Online order & Brand has blanks in “Transactions” dataset**

*Mitigation: Filtered Customer ID on 1-3500 since this is the common details on three datasets.*

*Filtered out blanks in last name, job title, online order and brands.*

*Recommendation: Make sure the databases are up to date*

*Do not make the job title granular*

*Give more dropdown options to brands*

*Explanation: If the customer details of all customer is not up to date it will be difficult to do further analysis. Job title should be less redundant. Customers job designation will be sufficing rather than their rank. All the brands the company sells have to be listed otherwise there will be difficulty in analyzing brand wise sales.*

**Consistency**

**Gender column is inconsistent in “Customer Demographic”**

**State column is inconsistent in “Customer Address”**

*Mitigation: Change ‘F’ and ‘femal’ to female and ‘M’ to male in gender column.*

*Change ‘NSW’ to New South Wales and ‘VIC’ to Victoria in state column.*

*Recommendation: Create dropdowns on Gender and State to ensure that the inputs are consistant*

*Explanation: Creating drop down for gender and abbreviation of state category will ensure there is no error in data entry. It will improve consistency of the datasets.*

**Currency**

**Customers who have ‘Y’ in deceased\_indicator are not current customers of Company.**

*Mitigation: Filtered out ‘Y’ in deceased\_indicator*

*Recommendation: Change the dropdown option in deceased\_indicator to only ‘N’*

*Explanation: Deceased customers are not current customers of the company. Removing them will improve currency of dataset and improve the analysis of future sales.*

**Relevancy**

**Default column in “Customer demographic” is incomprehensible**

*Mitigation: Deleted metadata in Default column.*

*Recommendation: Check for incomprehensible data and edit them to make comprehensible*

*Explanation: Irrelevant information can skew the analysis, removing them will increase the analyzing power.*

**Validity**

**Product\_first\_sold\_date and price column has inaccurate format in “Transactions”**

*Mitigation: Change the format of product\_first\_sold\_date to short date format and price to appropriate currency*

*Recommendation: Set up the column such that they take product sold date as date format and price as currency when entering the new data*

*Explanation: Accurate format of data make it easier to interpret them and it increases speed of analysis.*

These summaries all probable quality issues in datasets. Following the mitigations methods and recommendations will not only improve data quality for future analysis, it will improve the business strategies after analysis.

Furthermore, our team will continue the data cleaning and preparation process for model analysis. Any issues encountered will be documented. After completion it will be great to have a meeting with your SME to ensure all the assumptions are aligned with Sprocket Central Pty Ltd.

Please let us know if you have any queries regarding the mitigation and recommendations mentioned in the report.

Regards,

Gargi Podder.