Subject: Analysis of Raw Datasets from SP Rocket Central Pty Limited

Dear Client,

We have evaluated the four raw datasets received from Sprocket Central Pty Ltd. Based on our preliminary assessment, we have identified several quality issues within the data that require attention. Additionally, we have provided recommendations to address these issues and enhance the overall data efficacy.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Dataset 1**  **Transactions** | **Dataset 2**  **New Customer List** | **Dataset 3**  **Customer Demographic** | **Dataset 4**  **Customer Address** | **Recommendations** |
| Accuracy |  |  | ‘default’ does not have correct values |  | Clean the values. |
| Completeness | Blank values found in ‘Online Order’, ‘Brand Name’, ‘Product Line’, ‘Product Class’, ‘Product Size’, ‘Standard Cost’, and ‘product\_first\_sold\_date’ | Missing value in ‘last\_name’, ‘DOB’, ‘job\_title’, ‘job\_industry’ | Missing value found in ‘Job Title’, ‘Job Category’, ‘DOB’ and ‘Tenure’ |  | Remove missing value as the percentage of missing value is low |
| Consistency | Format for list price, standard cost is not consistent | ‘DOB’ value is not consistent, some with years and without month and date |  | In ‘State’ column of Customer Address Table, multiple duplicate values were found such as ‘VIC’ & ‘Victoria’, ‘NSW’ & ‘New South Wales. Same goes to ‘Gender’ | Consistent use of the data structure type  Use abbreviations of the states for all records to ensure consistency. Value ‘U’ under Gender can be imputed with mode value |
| Currency |  |  | Customer ID ‘34’ named Jephthah Bachmann was born in 1843, currently 175 years old is not reasonable |  | Remove redundant data |
| Relevancy | Filter cancelled order status |  | ‘past\_3\_years” should be redefined | ‘Country’ can be removed as they are all same | Perform analysis based on the synced data of all three customer tables based on customer\_id |
| Validity | Attribute ‘Standard Cost’ in Transaction table has some records with special string character. ‘Product\_first\_sold\_date’ need to be formatted. |  |  |  | Remove the special characters and convert them to numerical data. Format the |
| Uniqueness – records that are duplicated |  | Duplication of customers |  |  |  |

Please investigate the above-mentioned quality issues along with the recommended changes to ensure the consistent quality of the dataset across all the tables. If all the suggestions are matched, we can proceed with further analysis of the data to find some suitable insights for the company.

Kind regards,

Hui En Ngu

Data Consultant, KPMG