Dear (Name),

I hope this message finds you well. I would like to notify you that we have reviewed the 3 datasets that have been provided to us for data quality. Please take note that there are various data quality issues that needed to be addressed. Kindly refer to the table below for your reference:

|  |  |  |  |
| --- | --- | --- | --- |
| **Standard Data Quality Dimensions** | **Customer Demographic** | **Customer Addresses** | **Transactions** |
| Accuracy | DOB: outlier (ref. Customer ID 34) | 1. Customer ID: inaccurate numbering (missing 3, 22 & 23) 2. Postcode: same address with different postcode | Customer ID: outlier 5034 |
| Completeness | Job title: blank |  | 1. Online order: blank 2. Brand: blank |
| Consistency | Gender: Inconsistent classification | States: inconsistent abbreviations | Product ID: inconsistent classification (ref. 0 and 19) |
| Currency | Deceased customers: delete |  | Customer ID: only 3500 customer ID provided as compared to 4000 in Customer Demographic & Addresses |
| Relevancy | Age: missing |  | Profit: missing |
| Validity | Default: invalid text/value |  | 1. List price: format 2. Standard cost: format (ref. transaction ID 16701, 8708 & 17469) 3. Product sold date: format |

As a recommendation to mitigate these issues above, kindly refer to the table listed below:

|  |  |  |
| --- | --- | --- |
| **Standard Data Quality Dimensions** | **Issues** | **Recommendations** |
| Accuracy | 1. DOB outlier 2. Customer ID inaccurate numbering/missing ID 2,22 & 23. 3. Customer ID outlier. | 1. Filter out the outlier (1843-12-21) 2. Cross check the customer’s name with their relevant ID with data source regarding their address to check for accuracy 3. Filter out the outlier (5034) |
| Completeness | Blank in job title, online order, and brand. | Cross check the data with source for completeness |
| Consistency | 1. Inconsistent gender classification 2. Inconsistent states abbreviations 3. Product ID are inconsistent (i.e., product ID 0 and 19 resulted the same brand and product details.) | 1. Ensure that the data use standardized classification of gender: Male and Female for ease of data analysis 2. Ensure that the data use proper abbreviations. For example, VIC for Victoria so the outcome of data analysis will be proper 3. Cross check the product ID data with source to ensure that correct value is reflected in the excel. |
| Currency | 1. Deceased customer information still in the data sets 2. Transactions data only have 3500 customer ID as compared to 4000 in Customer Demographic and Customer Addresses | 1. The data sets should be up-to-date and contains only living customers 2. The transactions data need to be up-to-date and include the 500 remaining customers transaction data for proper full data analysis. |
| Relevancy | Missing customers age and profits column | Both data are relevant in performing data analysis as we can determine the customer preference by age bracket and performing profit analysis by product. |

|  |  |  |
| --- | --- | --- |
| Validity | 1. Default column consist invalid value/text 2. Unstandardized format for list price, standard cost, and product sold date. | 1. Check whether it provides metadata, if no, then can proceed to delete the column. 2. Make sure the format is consistent with the relevant data (For example, list price using currency format and product sold date using data format) |

Should you have any questions or concerns, please feel free to respond to this email.

Best regards,

Amir Syamil

Data & Analytics (KPMG Lighthouse)