Dear Client,

Thank you for providing us with the datasets for the Sprocket Central Pvt Limited. The summary table below mention the quality issues we have discovered during the analysis.

**Summary table**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Customer Demographics** | **Customer Address** | **Transactions** |
| **Accuracy** | DOB: inaccurate  Age: missing |  | Profit made on each transaction is not mentioned |
| **Completeness** | Job title: Blanks  Customer id: few ids were missing | Customer id: few ids were missing | Customer id: few ids were missing Brand: Blanks  Online order:  Blanks |
| **Consistency** | Gender: inconsistent | States: Inconsistent values |  |
| **Currency** | Deceased customer column: not updated |  |  |
| **Relevancy** | Default Column with irrelevant values. |  |  |
| **Validity** |  | Property Valuation format is not clear. | List price, product sale format is not accurate. |
| **Uniqueness** |  |  |  |

Below are more in-depth descriptions of data quality issues discovered and methods of meditation used. Recommendations and explanations have also been included to avoid further data quality issues in the future. Following Recommendations will improve accuracy of data used to influence business decisions of Sprocket Central Pvt limited.

**Accuracy**

**Issues:**

* DOB was inaccurate for “Customer Demographic” And missing an age column
* Missing a Profit column for “Transactions”.

***Mitigation*** : Filter out outlier in DOB

***Recommendations***: Create an ***Age column*** ,allowing for more comprehensive data and easier to check for errors. Create a ***Profit column*** in “Transactions” check accuracy of sales.

Creating additional columns for age and profit will allow for easy identification of errors. The Profit column will assist in future monetary analysis.

**Completeness**

**Issue**:

* Also, few of the ***Customer ids*** were missing amongst “Customer Demographic”, “Customer Address” and in “Transaction” data sheet.
* Blank values were found in ***job title*** for “Customer Demographic” and in ***brand, online order, job title, job industry*** columns of “Transactions” sheet.

***Mitigation***: filter all customer IDs from 1 to 3500 and filter out the blank values from the column where ever required.

***Recommendations***: Ensure tables are up to date. For our model, only customer IDs from 1 to 3500 will be used as they have complete data.

The data received may not be in Sync across all spreadsheets, with incomplete data the analysis result may be skewed. This is a completeness issue, to prevent future occurrences it is encouraged to crosscheck spreadsheet and sync data.

Blanks are treated as an incomplete data and can affect before the analysis result. The addition of drop-down options will allow to have more complete data and will result in more accurate analysis.

**Consistency**

**Issue**:

* Inconsistent values in ***Gender*** and ***States*** for “Customer Demographic” and “Customer Address” respectively.

***Mitigation***: Filter all ‘M’ under category of “Male”, filter all ‘Femal’ and ‘F’ under “Female” for ***Gender*** column.

Filter all ‘New South Wales’ to ‘NSW’ and ‘Victoria’ to ‘VIC’ for ***States*** column.

***Recommendations***: Create drop down options for ‘Male’ ,‘Female’ and ‘U’ in ***Gender***. Create drop down options for all state abbreviations.

Drop down options, minimizes manual entry and human error. And allows for increase of consistency of terminology. Gender identity can be sensitive topic, proceed with caution when creating options.

**Currency**

**Issues:**

* ***Deceased indicator*** column is not updated, people that are deceased are not current customers for “Customer Demographic”.

***Mitigation***: Filter out customers checked ‘Y’ in ***Deceased indicator***.

***Recommendations***: Can be difficult to check for deceased customers, but once this information is received one should update the data accordingly, deceased customers are not current customers & removing them from the data will increase the currency of the data and will result in more accurate estimates in future analysis.

**Relevancy**

**Issues**:

* Lack of relevancy or comprehensibility in default column for “Customer Demographic” and order status for “Transactions”.

***Mitigation***: Deleted meta data in ***default*** column, filter out ‘**Cancelled’** ***order status***.

***Recommendations***: Check for incomprehensible meta data and delete or format to make comprehensible

‘**Cancelled’** ***order status*** is irrelevant information for future analysis it can skew data.

**Validity**

**Issues**:

* Format of ‘***list price***’, ‘***product sale date***’ for “Transactions” and the format of ‘***property valuation***’ for “Customer Address”.

***Mitigation***: Format ‘***product sale date***’ to short date format and ‘***list price***’ to currency. ***Recommendations***: Set up columns so that formats such as price and decimals are already in place when entering new data.

Allowable values will make the data to be interpreted more easily formatting into price and allowing for either two or three decimals placed consistency consistently will increase readability this will reflect positively on speed.

This summarizes all the data quality issues discovered through the first stage of data quality analysis. The mitigation strategies suggested are simple and effective ways of improving data quality for future analysis. They will not only improve the analysis output that one can perform within the company what will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

Please let us know if you have any questions regarding mitigation or any data quality issues identified.

Kind regards,

Sejal Pol

(Data Analytics Consulting Team)