Dear **Sprocket Central** Data Team,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the key data quality issues from the three datasets received. Please let us know if you have any queries surrounding the issues presented.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer**  **Demographic** | * DOB: inaccurate * Age: missing | * Job title:   Blanks   * Customer id: incomplete | * Deceased indicator: filter out | Default column: delete |  |  |
| **Customer Address** |  | * Customer id: incomplete | * States: inconsistency |  |  |  |
| **Transactions** | * Profit: missing | * Customer id: incomplete * Online orders: blanks * Brand: blanks |  |  | * Cancelled order: filter out | * List price: format * Product sold date: format |

Below are more in depth descriptions of data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Recommendations have been included to avoid the further data quality issues and to improve the accuracy of the underlying data used to drive business decisions of Sprocket Central Pvt Ltd in the future.

**Accuracy Issues:**

* **DOB was inaccurate for “Customer Demographic” and missing an age column**
* **Missing an profit column for “Transactions”**

*Mitigation: Filter outliers in DOB.*

*Recommendation: Create an* ***age*** *column, allowing for more comprehensible data and easier to check for errors. Create a* ***profit*** *column in “Transactions” to check the accuracy of sales.*

*Creating an additional columns for age and profit will allow for easier identification of erros, The profit column will assist in future monetary analysis.*

**Completeness:**

* **Additional customer id’s were inconsistent among “Customer Demographics”, “Customer Address”, and “Transactions”.**

*Mitigation: Filter all customer\_id’s from* ***1 to 3500.***

*Recommendation: Ensure tables are up to date(from the same time period). For our model, only customer\_id’s from 1 to 3500 will be used as they have complete data.*

* **Blanks in job\_title “Customer Demographic”, in online\_order and brand\_column for “Transactions”**

*Mitigation: Filter out ‘blanks’ for* ***job\_title, online\_order and brand columns.***

*Recommendation: Simplify job\_title to another category such as* ***Industry*** *or provide dropdown options for* ***job\_title*** *and for* ***online\_order*** *and* ***brand*** *columns.*

*Blanks are treated incomplete data and can skew further analysis results. The addition of dropdown options will allow to have more complete data and will result in more accurate analysis.*

**Consistency**

* **Inconsistency in gender for “Customer Demographic” and “Customer Address” respectively**

*Mitigation: Filter all ‘M’ under category of ‘Male’, filter all ‘Femal’ and ‘F’ under ‘Female’ for* ***gender****. Filter all ‘New South Wales’ to ‘NSW’ and ‘Victoria’ to ‘VIC’ for* ***states.***

*Recommendation: Create dropdown options for ‘Male’, ‘Female’ and ‘U’ in* ***gender****. Create a dropdown options for all* ***state*** *abbreviations.*

*Dropdown options, minimizes manual entry and human error. Allows for increase of consistency of terminology. Gender identify can be sensitive topic, proceed with caution when creating options.*

**Currency**

* **People that are ‘Y’ in deceased\_indicator are not current customers for “Customer Demographic”.**

*Mitigation: Filter out customers checked ‘****Y****’ in* ***deceased\_indicator.***

*Recommendation: Can be difficult to check for deceased customers, but once this information is received one should update data accordingly.*

*Deceased customers are not current customers, removing them from data will increase currency of data and will result in more accurate estimates in future analysis.*

**Relevancy**

* **Lack of relevancy or comprehensibility in default column for “Customer Demogrpahic” and order\_status in “Transactions”**

*Mitigation: Deleted Metadata in default\_column. Filter out “****Cancelled****”* ***order\_status****.*

*Recommendation: Check for incomprehensible metadata and delete or format to make comprehensible.*

*‘****Cancelled****’* ***order\_status*** *is irrelevant information for future analysis, as it can skew data – for example total number of customers per annum will be an overestimate.*

**Validity**

* **Format of list\_price, product\_sale\_data for “Transactions”**

*Mitigation: Format* ***product\_sale\_date*** *to short date format, format* ***list\_price*** *to currency.*

*Recommendation: Set up columns so that fomats such as price and decimals are already in place when entering new data.*

*Allowable values will make data to be interpreted more easily. Formatting into price and allowing for either 2 or 3 decimal places consistently will increase readability. This will reflect positively on speed and accuracy of analysis for business decisions.*

That summarises all key data quality issues discovered through the first stage of the 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 but will increase the level of analysis that can be performed by KMPG team.

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

Kind regards,

**Anudeep Basavaraju**