Dear Kathleen,

Thank you for providing us with the three datasets from sprocket central Pty Ltd. The Summary table below highlights key quality issue that we discovered within the three data sets. Please let us know if you have any queries surrounding the issues presented.

**Summary Table**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Completeness | Consistency | Currency | Relevancy | Validity |
| Customer  Demographic | * DOB:   Inaccurate   * Age:   Missing | * Job title:   Blanks   * Customer id:   Incomplete | * Gender:   Inconsistency | * Decreased   Customers:  Filter out | * Default   Columns:  delete |  |
| Customer  Address |  | * Customer id:   Incomplete | * States:   Inconsistency |  |  |  |
| Transactions | * Profit:   Missing | * Customer id:   Incomplete   * Online orders:   Blanks   * Brands: blanks |  |  | * Cancelled status order:   Filter out | * List price:   Formula   * Product sold date format |

Below are more in-depth descriptions of data quality issues discovered and methods of mitigation used. Recommendation and explanations have also been included to avoid further data quality issues in the future. Following recommendation will improve accuracy of data used to influence business decisions of Sprocket Central Pty Ltd in the future.

**Accuracy Issues**

* **DOB was inaccurate for “Customer Demographic” and missing an age\_columns:**

**Missing a profit column for “Transactions”.**

Mitigation: Filter out outlier in DOB.

Recommendation: Create an age column, allowing for more comprehensible data and easier to check for error. Create a profit column in “Transaction” to check accuracy of scale.

Creating additional columns for age and profit will allow for easier identification of errors. The

Profit column will assist in future monetary analysis.

**Completeness**

* **Additional customer ids were inconsistent among “Customer Demographic,”**

**“Customer Address,” and “Transactions”.**

Mitigation: Filter all customer from 1 to 3500

Recommendation: Ensure tables are up to date (from the same time period). For 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 results may be skewed. This is a ‘Completeness’ issue, to prevent occurrence it is encouraged to cross check spreadsheets and sync data.

* **Blanks in job title for “Customer Demographic,” in or in online order and brand column for “Transaction”.**

Mitigation: Filter out ‘Blanks’ for job title, online order, and brand column.

Recommendation: Simplify job title to another category such as industry or provide

dropdown options for job title. Provide dropdown options for online order and brand column.

Blanks are treated as 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 ‘Female’ and ‘F’ under ‘Female’ for gender. Filter all ‘New South Wales’s to ‘NSW’ and ‘Victoria’ to ‘VIC’ for state.

Recommendation: Create dropdown option for ‘Male’, ‘Female’, and ‘U’ in gender. Create dropdown options for all state abbreviation.

**Currency**

* **People that are ‘Y’ in decreased indicator are not current customer for ‘Customer Demographic’.**

Mitigation: Filter out customer checked ‘y’ in decreased indicator.

Recommendation: Can be difficult to check for decreased customers. But once this information is received one should update data accordingly.

Decreased customers are not current. Removing them form data will increase currency of data and will result in more accurate estimate future analysis.

**Relevancy**

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

Mitigation: Format product sale date to short date format. Format list price to currency.

Recommendation: set up columns so that formats 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 decimals placed consistently will increase readability. This will reflect positively on speed and accuracy of analysis for business decision.

That summarises all 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 preform within the company but will increase the level of analysis that can be performed by KPMG and other hired analysis teams.

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

Kind regards,

Kishore Devaragudi