Accuracy: Correct values

Completeness: Data fields with values

Consistency: Values free from contradiction

Currency: Values up to date

Relevancy: Data items with value meta-data

Validity: Data containing allowable values

Uniqueness: Records that are duplicated

Dear John Doe,

Thank you for providing us with 3 datasets from Sproket Central Pty Ltd. The quality issues that were discovered in the 3 datasets are put into a summary below. Please reach out to us if you have in kind of inquiry.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| **Customer Demographic** | * Inaccurate DOB * Missing Age: | * Customer id: incomplete * Job title blanks | * Gender Inconsistency | * Filtered out deceased customers | * Deleted Default Column |  |
| **Customer Address** |  | * Incomplete customer ID | * Inconsistency in states |  |  |  |
| **Transactions** | * Missing Profit | * Customer id: incomplete * Removed Blanks from online order * Removed Blanks from Brand |  |  | * Filtered out cancelled status order | * List Price: Format * Product Sold date: Format |

Below are the more in-depth descriptions of data quality issues that we discovered and the methods of mitigation used. We have also added recommendations and explanations so that data quality issues in the future can be avoided. The recommendations 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\_column**

**Missing a profit column from “Transactions”**

*Mitigation: Filter out outlier 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 accuracy of sales.*

Creating these additional columns will help identify errors faster. The **profit\_column** will also help with monetary analysis in the future

**Completeness**

* **Additional customer\_ids were inconsistent among “Customer Demographic,” “Customer Adress,” and “Transactions”**

*Mitigation: Filter out all* ***customer\_ids*** *from 1-3500*

*Recommendations: Ensure tables are up to date (from the same time-line). We will be using* ***customer\_ids from 1-3500*** *will be used as they have complete data.*

The Data we received may not be synced across all the spreadsheets, with incomplete data analysis results may be skewed. This is a ‘completeness, issue, for the future it is encouraged to sync data in all the spreadsheets

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

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

*Recommendations: Provide dropdown option for* ***job\_title****,* ***online\_order****, and* ***brand\_column****.*

Blanks are treated as incomplete data and can skew further analysis of data. The drop down boxes will ensure completeness of data

**Consistency**

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

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

*Recommendation: Create Drop down boxes for ‘Male’, ‘Female’ and ‘U’ in* ***gender****. Create drop down boxes for* ***states*** *with abbreviation.*

Drop down boxes decrease manual entry and reduces Human error.

**Currency**

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

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

*Recommendations: Should update data regularly to keep customer tacking up to date*

Deceased Customers are not current customers, removing data will increase currency of data and will result in more accurate analysis and estimate in the future.

**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 and so that formats like price and decimals are already in place.*

Allowable values will make interpreting data more easy. Formatting it to 2 to 3 decimal place will increase the data’s readability.

**Relevancy**

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

*Mitigation: Deleted Metadata in* ***default\_column****. Filter out* ***‘Cancelled’ order\_status****.*

*Recommendations: Check for incomprehensible Metadata and delete or format to make them comprehensible.*

**‘Cancelled’ order\_status** is not relevant data for future analysis, as it can skew data. One of the problems the company could face is, the total number of customers annually will be an over estimate.

Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central’s understanding.

Kind regards,

Tanjinul