Dear Client,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd.

The table below shows some data quality issues that I discovered during my assessment of the three dataset. I have also make some recommendation to improve the quality of the dataset for better analysis and good business decision-making.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** | **uniqueness** |
| **Customer Demographic** | * DOB – Inaccurate * Age missing | * Job title- Blanks * Customer id- incomplete | * Gender- Inconsistent | * Deceased Customer- Filter out | * Default column- Deleted |  |  |
| **Customer Address** |  | * Customer Id- Incomplete * Tenure - Blanks | * State - Inconsistent |  |  |  |  |
| **Transactions** | * Profit missing | * Customer ID- Incomplete * Online Order –Blanks * Product Line –Blanks * Product Class- Blanks * Product Size –Blanks * Standard Cost- Blanks * Product First Date Sold- blanks |  |  |  | * Product First Sold Date- Format * List Price - Format |  |

The descriptions of data quality issues discovered and the methods of mitigation used are briefly discussed below. In addition, recommendation and explanations are included to avoid further data quality issues in the future.

**Accuracy Issues**

* DOB was inaccurate for Customer Demographic, age column missing and profit column missing for Transactions.

**Mitigation**: filter out outlier in DOB

**Recommendation**: Create an age column to be able to check for errors. Create profit column in Transaction for accuracy of sales and it will assist in future monetary analysis.

**Incompleteness**

* Customer ID were inconsistent among Customer Demographic, Transaction and Customer Address.

**Mitigation**: Filter all customer ids from 1 – 3500. Filter out blanks cells

**Recommendation**: Customer ids from 1 to 3500 will be used for our data model as they contains complete data. Ensure complete data and double check data across all spreadsheets.

Blanks are treated as incomplete data. Hence, they are removed to avoid skewing our analysis result.

**Consistency**

* Inconsistency in gender for Customer Demographic and Customer Address
* Inconsistency in State for Customer Address

**Mitigation**: Filter all ‘M’ to Male and all ‘F and Female’ to Female for Gender.

Filter New South Wales to NSW and Victoria to VIC for State

**Recommendation**: Create a dropdown option for state abbreviation. Create a dropdown option for gender.

**Currency**

* ‘Y’ customer in Deceased Indicator under Customer Demographic are not current customer.

**Mitigation:** Filter out ‘Y’ in Deceased Indicator

**Recommendation**: Deceased customers are not current customer, removing them will increase the efficiency and currency of the data and will improve the accuracy of our analysis.

**Relevancy**

* Lack of relevancy for Default column in Customer Demographic

**Mitigation**: Default column deleted

**Recommendation**: Check and double check for irrelevant and incomprehensible data to be deleted or removed.

**Validity**

* Format List price and Product first sold date in Transaction

**Mitigation**: Format (Data type) List\_Price to currency and Product\_first\_sold\_date to short date

**Recommendation**: Set up columns so that format or data types such as price or date are already in place when entering data.

The above summary are the data quality issues discovered, mitigations and recommendations. Taking all these into account will improve the quality our data and also increase the efficiency of our analysis for good business decision-making.

Kindly reach out to me for further questioning and clarification.

Kind regards,

Paul Aderounmu