**Data Quality Assessment**

**Overview:** *Sprocket Central Pty Ltd , a medium size bikes & cycling accessories organisation. Primarily, Sprocket Central Pty Ltd needs help with its customer and transactions data. The organisation has a large dataset relating to its customers, but their team is unsure how to effectively analyse it to help optimise its marketing strategy.*

**Task:** *Review the data quality to ensure that it is ready for our analysis in phase two. Remember to take note of any assumptions or issues we need to go back to the client on. As well as recommendations going forward to mitigate current data quality concerns.*

**Solutions**

This email is written to express our evaluation for the datasets after carefully review:

We received 3 datasets ***Customer Demographic , Customer Addresses, Transaction*** data in the past three months. Here is the table expressing data quality under ***7 criterias*** in ***Standard Quality Dimensions.***

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Customer Demographic** | **Customer Addresses** | **Transaction data** |
| ***Accurracy*** | *Likely-good* | Good | Good |
| ***Completeness*** | *Average* | *Likely-good* | *Likely-good* |
| ***Consistency*** | *Average* | *Likely-good* | Good |
| ***Currency*** | Good | Good | Good |
| ***Relevancy*** | Good | Good | Good |
| ***Validity*** | Good | Good | Good |
| ***Uniqueness*** | Good | Good | Good |

We will explain clearly about criterias that don’t meet expectation “Good”

1. **Completeness**

*Customer Demographic:* Columns that are not filled completely are DOB **(0.02%)** and Job title at high level **(12.65%)**

*Transaction data:* Columns that are not filled completely are Online Oder **(1.8%),** Brand along with Product Line, Product Class, Products Standard Cost, Product first sold date **(4.5%).**

*Customer Addresses:* Column that is not filled completely is CustomerId with id **22, 23.**

1. **Consistency**

*Customer Demographic:* In Gender Column, Female is written as “Female” and “F”. Similarity, Male is written as “Male” and “M”.

*Customer Addresses:* In State column, there are inconsistent is “New South Wales”-“NSW” and “Victoria”-“VIC”. We think that they express the same values.

There are some customer\_id that appear in Transaction and Customer Adress but not in Customer Demographic.

Inconsistent data type for the same attribute (number value and string data type in the same column)

After assessing, we conclude that the overall quality of current datasets is good. There are some criterias need to improve:

1. Should fill the blank values in column ***Job title*** in Customer Demographic, it’s currently has high missing value rate and Job title is idea criteria to segment the customer.
2. For other missing data, we can filter because it’s not significant.
3. Should transform the ***inconsistency values*** we mentioned above into one dimension by replacing. Recommend using dropdown list instead free text field when inserting data.
4. Make sure all tables are from the same period to avoid missing values and inconsistent data
5. About inconsistent data type, should have data type constraints for each column.

After solving the issues above, the datasets will be ready for analyze.

Thank you so much!