**Task 1 – Data Quality Assessment**

Dear Sprocket Central Pty Ltd,

We have assessed the quality of the datasets you have provided to us and identified the issues and have proposed strategies to mitigate them.

Chart, diagram, funnel chart

Description automatically generated

The dimensions of the framework are detailed as such and will give you a better understanding of the analysis completed on the datasets you have provided. If a more in depth framework is required, please do not hessite to request one as we will be happy to provide it to you.

Below are summary tables of the data quality issues identified within each dataset, in line with the data quality framework.

|  |  |
| --- | --- |
| **Dataset** | **Transactions** |
| **Accuracy** | - |
| **Completeness** | 7 columns have *missing data*: online\_order, brand, product\_line, product\_class, product-size, standard\_cost and product\_first\_sold\_date [low % of missing data] |
| **Consistency** | *Inconsistent data types* for some columns– some fields are numeric and others strings |
| **Currency** | - |
| **Relevancy** | - |
| **Validity** | *Incorrect data type* for columns product\_first\_sold\_date, list\_price and standard cost |
| **Uniqueness** | - |

|  |  |
| --- | --- |
| **Dataset** | **New Customer List** |
| **Accuracy** | - |
| **Completeness** | 4 columns have *missing data*: last\_name, DOB, job\_title, job\_industry\_category [2 columns have over 10% of missing data] |
| **Consistency** | - Gender categories in gender column have a different *format*: U (for Unspecified) does not fit the same format as male and female (words not abbreviations)  - Columns deceaded\_indicator and owns\_car have a different *format* for Yes/No, one column has it formatted as abbreviations and the others not |
| **Currency** | - |
| **Relevancy** | There are 5 *un-named columns* |
| **Validity** | - |
| **Uniqueness** | - |
| **Dataset** | **Customer Demographic** |
| **Accuracy** | - Additional Customer IDs  - There’s a DOB with an *incorrect birth year* of 1843 and there are 87 DOBs with birth year NaT  - Categories in gender column are *titled incorrectly* (and thus repeated categories present) |
| **Completeness** | 6 columns have *missing data*: last\_name, DOB, job\_title, job\_industry\_category , default, tenure [2 columns have over 10% of missing data] |
| **Consistency** | - Gender categories in gender column have a different *format*, some genders are formatted as abbreviations while others are not  - Columns deceaded\_indicator and owns\_car have a different *format* for Yes/No, one column has it formatted as abbreviations and the others not |
| **Currency** | - |
| **Relevancy** | The default column contains *no valuable data* |
| **Validity** | - |
| **Uniqueness** | - |

|  |  |
| --- | --- |
| **Dataset** | **Customer Address** |
| **Accuracy** | Additional Customer IDs |
| **Completeness** | - |
| **Consistency** | The states in the state column have different *formats* as some states are formatted as abbreviations while others are not (and thus repeated states present). |
| **Currency** | - |
| **Relevancy** | - |
| **Validity** | - |
| **Uniqueness** | - |

We have proposed some strategies that can be taken to mitigate these issues in the future:

***Accuracy Issues:***

From the ‘Customer Demographic’ dataset there were two accuracy errors: incorrect birth years in the DOB column and incorrectly titled gender categories in the gender column.

*Mitigation*: data filtering on the fields before any analysis of the data can be taken. This data filtering allows you to check the accuracy of your data and either drop it or alter it depending on the nature of the analysis.

Additional Customer IDs in the ‘Customer Demographic’ and ‘Customer Address’ dataset but not in the

‘Transactions’ dataset. This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records

*Mitigation*: ensure that all tables are from the same period. Only customers in the ‘Transactions’ will be used during the analysis of the data.

***Completeness Issues:***

Datasets ‘Transactions’, ‘New Customer List’ and ‘ Customer Demographic’ had missing data in some of their columns. The ‘Transactions’ dataset had a very low % of missing data and with cases such as this, the best approach would be to drop the values before the analysis. However for the other two datasets, few of their (categorical) columns had over 10% of missing data.

*Mitigation:* for those columns we propose transforming the data to replace the null categorical null values with an appropriate replacement depending on the nature of the analysis.

***Consistency Issues:***

Datasets ‘New Customer List’, ‘ Customer Demographic’ and ‘Customer Address’ have format inconsistencies for some of their categorical columns, some of the categories within columns are formatted as abbreviation while other are not which results in repeats of categories.

*Mitigation:* transform data so that all the data follow the same formatting, in addition to eliminating repeats of categories. This is done by replacing one format for the other (abbreviations are simpler so should be kept, granted the abbreviations are understood by all involved)

*Recommendation:* Enforce a drop-down list for the user entering the data rather than a free text field.

In order to construct meaningful variables for the model, the data has been cleaned to avoid

multiple representations of the same value. Additionally, gender records where ‘U’ have been

replaced based on the distribution from the training dataset.

***Relevancy Issues:***

There are five un-named columns in the ‘New Customer List’ dataset that seem to be holding unusable data and the approach we propose is to drop this data.

The default column in ‘Customer Demographics’ dataset contains no valuable data instead it is mainly made up of symbols and should also be dropped. In the off chance that this column has been corrupted it would be advisable in the future to always have access to the original (uncorrupted data).

***Validity Issues:***

The ‘Transactions’ dataset has an invalid data type for a few of its columns :

1. The product\_first\_sold\_date column should be a date data type but instead has no specific format.
2. The list\_price should be a currency data type but instead has no specific format.
3. The standard\_cost column has a custom currency format, however it has not been applied to every field

*Mitigation*: transforming the data types of the column via the formatting options in Excel.

Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations need to be made to ensure consistent data types for a given

field.

Moving forward, the team will continue with the data cleaning, standardisation and transformation process

for the purpose of model analysis. If you have any further questions or request regarding this assessment of the data quality and the assumption we have made, do let us know and we’d be happy to meet your needs before we move on to the analysis of the data.

Kind Regards,

Nusayba Ahmed