**Subject: Data Quality Result**

Dear Manager,

Thank you for providing us with the datasets from Sprocket Central Ltd. As, requested by you we have reviewed the data to ensure that it is ready for analysis in phase two. But while analysing the we have found the issues in some dataset.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Datasets** | **Accuracy** | **Completeness** | **Consistency** | **Timeliness** | **Relevancy** | **Uniqueness** | **Validity** |
| (Dataset 1)  Transactions |  | Missing: profits (list price - standard cost),  online order, brand, product line/class/size, standard cost, first sold date | Unrealistic: product id |  |  |  | Format:  product sold date, |
| (Dataset 2)  New Customer List | Unreliable:  Customer Id | Missing: last name, DOB, job title, job industry | Format:  gender |  | Irrelevant:  unnamed columns |  |  |
| (Dataset 3)  Customer Demographics | Erroneous: DOB | Missing: last name, job title, job industry, tenure | Format: gender  Unrealistic: job title | Out of date: deceased customer | Corrupted:  default column deleted |  | Unreliable:  age |
| (Dataset 4)  Customer Address |  |  | Format: state |  |  |  |  |

The summary table below highlights key data quality issues we have discovered in the data cleaning process using Standard Data Quality Dimensions.

I provide a more thorough explanation of the data quality problems we identified in the sections that follow, along with suggestions and justifications on how to increase the reliability of the data sources in order to prevent such problems in the future. As a result, the data that can be used to guide future business decisions will be more accurate.

As mentioned earlier I have evaluated the datasets according to the data quality dimensions framework as follows:

* Accuracy – correct values
* Completeness – data fields with values
* Consistency – values free from contradiction
* Timeliness – values up to date
* Relevancy – data item with value meta-data
* Uniqueness – records that are not duplicated
* Validity – data containing allowable values

***Accuracy issues:***

In Dataset 2(NewCustomerList), the Customer Id.

* Mitigation: NA
* Recommendation: Customer Id, as it may ensure the unique Id of Customer and to identify the Duplicated value.

In Dataset 3(CustomerDemographic), the DOB is incorrect, E.g., one entry indicate that the customer was born in year 1843.

* Mitigation: In DOB, I filtered the outliers.
* Recommendation: Create an age column for more comprehensible data and to allow checking for errors.

***Completeness issues:***

In Dataset 1(Transactions), some records are null/missing, such as online order, brand, product line/class/size, standard cost, first sold date.

In Dataset 2(NewCustomerList), some records are null/missing, such as last name, DOB, job title, job industry.

In Dataset 3(CustomerDemographic), some records are null/missing, such as last name, job title, job industry, tenure.

Mitigation: I used forward/backward filling techniques or, when appropriate, an average of the cluster group to fill in the missing null values.

Recommendation: Provide drop-down menus for the brand column, online order, and job title. The first sales date should be transformed into a common format. The introduction of filled data could bias the analysis's findings. Utilizing pre-defined options will enable the usage of more thorough data.

***Consistency issues:***

In Dataset 1(Transactions), product\_id was unrealistics

In Dataset 2(NewCustomerList) and Dataset 3(CustomerDemographic), the gender was in inconsistent formats.

In Dataset 4(CustomerAddress), the state was in inconsistent formats.

Mitigation: I located all variations of Men under the category of "Male" and all variations of Women under the category of "Female" and all variations U under the category of "Unspecified" and replaced them. To establish consistency among the datasets, I also abbreviated state names, such as "Victoria" to "VIC." Each product\_id should have its own brand, its own line, its own class, and its own size and with it the list price and standard cost corresponding to each product\_id

Recommendation:The data type must be categorical, not a variable text field, to avoid different representations of the same value. Drop-down options minimize manual input inconsistencies and human error by various collaborators, improving the interpretability and readability of the data. Gender is a protected property, so anything marked as otherwise may fall into the "U"category.

***Timeliness issues:***

In Dataset 3(CustomerDemographic), Some customers are not current customers as they are reported deceased.

Mitigation: In Deceased, I filtered the customers marked as Deceased.

Recommendation: This information can be difficult to verify, but should be updated as soon as possible if it is available.

***Relevancy issues:***

For Dataset 2(NewCustomerList), there were several hidden columns.

For Dataset 3(CustomerDemographic), there was a default column with incomprehensible or corrupted data.

Mitigation:I dropped the default column and also dropped the unnamed/hidden column showing the data that wasn’t making sense or corrupted. Also filtered the order status for cancelled order.

Recommendation: Remove any incomprehensible meta-data to make it comprehensible.

***Validity Issues:***

In Dataset 1(Transactions), the product sold date is a float and which may cause confusion.

In Dataset 3(CustomerDemographic), missing an age column.

Mitigation: I standardised the product sold date using datetime

Recommendation: Make sure all datasets are from the same period. Otherwise, duplicate or missing datasets can distort your data analysis. The dataset has been merged into a single master dataset in a unified format.

***Other Data Quality Issues:***

Across numerous features and columns, there were a large number of missing datapoints.

Additionally, there were some data inconsistencies, or dataset mismatches, in some of the data.

The same attributes were assigned to different data types, such as integer for certain fields and float for others, which can result in unexpected problems due to the difference in precision. Across numerous features and columns, there were a large number of missing datapoints.

Additionally, there were some data inconsistencies, or dataset mismatches, in some of the data.

The same attributes were assigned to different data types, such as integer for certain fields and float for others, which can result in unexpected problems due to the difference in precision.

Mitigation: I have filled in the data using the proper statistical techniques. Otherwise, the entries have been removed from the master datasets if the number of null values is considerable. I only made an exception when the sample size was tiny and the data points were extremely important. This led to the standardisation of all fields and the enactment of restrictions on the types of data that were allowed.

The main data quality issues identified during the initial data quality analysis step are summarised here. The team will carry on with the extraction, transformation, and loading procedure going ahead in order to analyse the model in stages 2 and 3.

If you have any feedback or inquiries on the aforementioned, kindly let me know.

With Regards

Duy Duong

Data Consultant, KPMG\*

\*Disclaimer: This is a hypothetical scenario in the role of a data analysis consultant.