**Re: Data Analysis: Data Quality Assessment**

Dear Manager

Thank you for providing us with 3 datasets from Sprocket Central Pty Ltd.

The summary table below highlights the data quality issues discovered in the data quality assessment process. Please let us know if you have any queries regarding the findings presented.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** | **Uniqueness** |
| **Customer Demographic** | Error:  DOB  Spelling Errors | Missing Values:  customer\_id  first\_name  last\_name  gender  past\_3\_years  DOB  job\_title  job\_industry  wealth\_segment  deceased\_indicator  default  owns\_car  tenure | Format:  gender (F, Female, Female, M, U, Male) |  | Corrupted:  default |  |  |
| **Transactions** | Error:  brand (Solex, Sole, Selex, Silex, Soled, Soles, Scolex)  Missing standard\_price | Missing Values:  product\_id  customer\_id  transaction\_date  online\_order  order\_status  brand  product\_line  product\_class  product\_size  list\_price  standard\_cost  product\_first\_sold\_date |  |  |  |  |  |
| **Customer Address** | Error:  Spelling Error  (Like Mcbride, McBride, MacBride) | Missing Values:  customer\_id  address  postcode  state  country  property\_valuation | Format:  state (New South Wales, QLD, VIC, NSW, Victoria) |  |  |  |  |

**Evaluation of Dataset is done based on:**

*Accuracy*: Correct values

*Completeness*: Data without missing values

*Consistency*: Values with correct format, free from contradiction and confusion

*Timeliness*: Values up to date

*Relevancy*: Data item with meta data

*Uniqueness*: No duplicate values

*Validity*: Data with allowable values

**Data Quality Issues**:

1. Accuracy:

DOB data is not accurate there are some outliers.

*Mitigation:* Removing of Outliers

*Suggestions:* Make new column for DOB with less and easily manageable errors.

1. Completeness:

There are missing values present in the datasets.

*Mitigation:* Applying missing value treatment on all columns

*Suggestions:* Using all the available data for analysis

1. Consistency:

There is inconsistency present in the data, the data needs to be in one format to be easily readable

*Mitigation:* Replace the wrong format data with the format standard format Ex: in gender column replace F with Female, M with Male etc.

1. Validity:

There are some entries with id value zero Ex: product\_id, customer\_id having values zero.

The above data quality issues are discovered in the first data cleaning stage, the team will further continue to work with data for model analysis stage.

Please let us know if you have any questions or comments on the above summary as I would be happy to discuss to ensure that all assumptions are align with Sprocket Central Pyt Ltd’s understanding.

Kind Regards

Priyanka

Data Analyst