Dear Client,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd, to begin with, I have summarised the key facts of each table below, this is prior to any cleaning or transformation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table Name | Total Rows | Total Distinct Customers | Total Columns | Columns Included |
| Customer Demographic | 4,000 | 4,000 | 13 | customer\_id  first\_name  last\_name  gender  past\_3\_years\_bike\_related\_purchases  DOB  job\_title  job\_industry\_category  wealth\_segment  deceased\_indicator  default  owns\_car  tenure |
| Customer Address | 3,999 | 3,999 | 6 | customer\_id  address  postcode  state  country  property\_valuation |
| Transactions (3 Months) | 20,000 | 3,494 | 13 | transaction\_id  product\_id  customer\_id  transaction\_date  online\_order  order\_status  brand  product\_line  product\_class  product\_size  list\_price  standard\_cost |

**Data Quality Issues and Mitigation**

The data was then measured against the below attributes for quality and cleanliness, any issues that were encountered are detailed below as well as the mitigation carried out to address these issues prior to analysis. The table below summarises the findings of this exercise, but further detail is given below.

|  |  |
| --- | --- |
| **Data Quality Attribute** | **Tables With Issues Present** |
| Accuracy | N/A |
| Completeness | Customer Demographic  Transactions |
| Consistency | N/A |
| Timeliness | Customer Demographic  Customer Address  Transactions |
| Relevancy | Customer Demographic |
| Validity | Customer Demographic  Customer Address  Transactions |
| Uniqueness | N/A |

**Accuracy**

The degree to which information accurately reflects an event or object described.

Issues

* No identified issues.

Mitigations

* No identified issues.

**Completeness**

The degree to which required data is present, gaps in the data is nearly always a given but this may only be impactful if key data (required for answering the given question) is missing.

Issues

Missing values in the transactions (online\_order, brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date) and customer demographic (job\_industry\_category, job\_title, default, last\_name, DOB, tenure) datasets.

Mitigations

* Any records with nulls were dropped due to being small percentages of the entire dataset. With the exception of job\_industry\_category and job\_title, these could have been input using the distribution of existing values however as these are key fields this could influence the model.

**Consistency**

If data is replicated in multiple places, it needs to be consistent across all instances.

Issues

* No identified issues.

Mitigations

* No identified issues.

**Timeliness**

When business required data is available in the time frame that it is needed and that it is current/up to date.

Issue

* All customers in each of the three datasets are not necessarily present in all of them, indicating the data for each dataset could be from different time periods.

Mitigations

* Only data for customers present in all three datasets has been used for modelling.

**Relevancy**

Data contains genuinely valuable information to the business, and we are collecting it for a valid given purpose.

Issues

* ‘default’ column in customer demographic, non-sensical and non-valuable.

Mitigations

* Dropped non-required/non-valuable columns.

**Validity**

This refers to information that doesn’t conform to a specific format or doesn’t follow business rules.

Issues

* ‘state’ in the customer address table had variations (‘VIC’ and ‘Victoria’). Values were standardised to abbreviations only.
* ‘gender’ in the customer demographic table had variations (‘F’, ‘Femal’ and ‘Female’.
* ‘first sold date’ in transactions not formatted as date.

Mitigations

* Ensure data provided are aligned to the same time periods to contain all relevant customers in each, only customers present in all three tables will be used for modelling.
* Implement drop downs, or other data validation methods, to ensure users can only input pre-defined values to ensure consistency.
* Converted to consistent date format.

**Uniqueness**

Ensuring that there’s only one instance of it appearing in a database.

Issues

* No identified issues.

Mitigations

* No identified issues.

Below is a summary of the key facts of each table following data cleaning. With the newly cleaned data, I shall begin model development to support customer segmentation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table Name | Total Rows | Total Distinct Customers | Total Columns | Columns Included |
| Customer Demographic | 2,327 | 2,327 | 13 | customer\_id  first\_name  last\_name  gender  past\_3\_years\_bike\_related\_purchases  DOB  job\_title  job\_industry\_category  wealth\_segment  deceased\_indicator  owns\_car  tenure  age |
| Customer Address | 2,327 | 2,327 | 6 | customer\_id  address  postcode  state  country  property\_valuation |
| Transactions (3 Months) | 12,970 | 2,327 | 13 | customer\_id  product\_id  transaction\_date  online\_order  order\_status  brand  product\_line  product\_class  product\_size  list\_price  standard\_cost  new\_product\_first\_sold\_date |

If any of the actions/assumptions made are not aligned with your business’ approach, please let me know so we may discuss further.

Many thanks,

Sam Dejean