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