

**From:** Quim Bolós Fernández bolosfernandez@hotmail.es  
**Subject:** Data Quality Assessment and Strategies for Mitigation  
**Date:** 12 July 2023 at 17:09  
**To:** Client



Dear [Client],

I hope this email finds you well.

I have thoroughly assessed the datasets provided, namely the Customer Demographic, Customer Address, New Customer List, and Transactions. The table below highlights the summary statistics for the 4 datasets received. Please let us know if the figures are not aligned with your understanding or could provide more information to complete it.

DataSets Statistics	Nº of Records	Distinct Customer ID	Date Data Received
Customer Demographic	4000	4000 [1 - 4000]	Not Registered
Customer Address	3999	3999 [1 ... 4003] - missing values	Not Registered
New Customer List	1000	Not Registered	Not Registered
Transactions	20000	3494 [1 ... 3500] - missing values	2017

As it can be seen, customer\_ids can vary from datasets. Additionally, using customer\_id as the join key for the demographic and address datasets, only 5 entries from the New Customer List exist in the joined dataset, which may indicate that the data received may not be in sync with each other.

On the other hand, the individual notable data quality issues identified for each dataset are detailed below. Ultimately, I propose strategies to mitigate these issues and improve the overall data quality.

## 1. Customer Demographic

<b>Accuracy</b> Correct Values	One value in the Date of Birth (DOB) column is incorrect, showing the year as 1843.
<b>Completeness</b> Data Fields with Values	Several data fields have missing values, including last_name, DOB, job_title, job_industry_category, default, and tenure.
<b>Consistency</b> Values Free From Contradiction	<p>The Gender column contains various inputs such as 'M', 'Femal', 'F', and 'U,' which should be translated into Male or Female for semantic consistency.</p> <p>There are inconsistencies between the job_title and job_industry_category columns, where certain job titles are incorrectly assigned to different job industry categories.</p> <p>Overall, n/a should be replaced by blanks or the other way around for semantic consistency.</p>
<b>Currency</b> Values Up to Date	Columns such as deceased_indicator and owns_car may not be up to date.
<b>Relevancy</b> Data Items with Value Meta-Data	<p>The deceased_indicator column includes some 'Yes' values, which may not be useful since they are very few in number.</p> <p>The job_title column includes various jobs with different levels, potentially segmenting the dataset unnecessarily.</p> <p>The past_3_years_bike_related_purchases and wealth_segment columns may disagree if they are reliant on each other since all wealth_segments have the same range of values for past_3_years_bike_related_purchases.</p>
<b>Validity</b> Data Containing Allowable Values	The default column does not contain any allowable values.
<b>Uniqueness</b> Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different customer_id.

## 2. Customer Address

<b>Completeness</b> Data Fields with Values	Customers with IDs [3, 10, 22, 23] do not have address data.
<b>Consistency</b>	The state column contains various inputs such as 'New South Wales' and 'Victoria', which are

Values Free From Contradiction	also represented as 'VIC' or 'NSW'. These values should be should be all in the format 'VIC' or 'NSW' for semantic consistency.
<b>Currency</b> Values Up to Date	The address data may not be up to date.
<b>Relevancy</b> Data Items with Value Meta-Data	All customers in this dataset live in Australia, making the Country column irrelevant.
<b>Uniqueness</b> Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different customer_id.

### 3. New Customer List

<b>Completeness</b> Data Fields with Values	Several data fields have missing values, including last_name, DOB, job_title, and job_industry_category.
<b>Consistency</b> Values Free From Contradiction	<p>The Gender column contains 'U' values, which could be interpreted as 'Unknown' for semantic consistency.</p> <p>There are inconsistencies between the job_title and job_industry_category columns, where certain job titles are incorrectly assigned to different job industry categories.</p> <p>Overall, n/a should be replaced by blanks or the other way around for semantic consistency.</p>
<b>Currency</b> Values Up to Date	Columns such as deceased_indicator, owns_car and address-related data may not be up to date.
<b>Relevancy</b> Data Items with Value Meta-Data	<p>The deceased_indicator column does not contain any positive values, rendering it irrelevant.</p> <p>The job_title column includes various jobs with different levels, potentially segmenting the dataset unnecessarily.</p> <p>The past_3_years_bike_related_purchases and wealth_segment columns may disagree if they are reliant on each other since all wealth_segments have the same range of values for past_3_years_bike_related_purchases.</p> <p>The Country column contains only 'Australia' values, which may not provide valuable insights.</p> <p>Assess the relevancy of the default column and consider its removal.</p> <p>Evaluate the calculation and relevance of scoring criteria in columns 16 to 19.</p>
<b>Validity</b> Data Containing Allowable Values	The past_3_years_bike_related_purchases and the property_valuation are in text format, and should be in numeric format.
<b>Uniqueness</b> Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different customer_id.

### 4. Transactions

<b>Accuracy</b> Correct Values	The customer_id column has an outlier value of 5034
<b>Completeness</b> Data Fields with Values	<p>The online_order column has missing data for 19640 out of 20000 entries.</p> <p>The product_id 0 has missing data in multiple columns, including brand, product_line, product_class, product_size, standard_cost, and product_first_sold_date.</p>
<b>Consistency</b> Values Free From Contradiction	The product_first_sold_date values are different for each product_id, even when differentiated by online_order. However, list_price and standard_cost are consistent for each product_id.
<b>Validity</b> Data Containing Allowable Values	The product_first_sold_date values are in the wrong format.
<b>Uniqueness</b> Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different transaction_id.

### Mitigation Strategies

To address these data quality issues, I recommend the following strategies:

<b>Overall</b>	<p>Continue maintaining uniqueness within the dataset.</p> <p>Standardise gender inputs to either 'Male', 'Female' or 'Unknown'.</p> <p>Review and rectify inconsistencies between the job_title and job_industry_category columns.</p> <p>Regularly update columns like deceased_indicator, owns_car, and address-related data to ensure currency.</p>
<b>Customer Demographic</b>	<p>Correct the incorrect value in the DOB column.</p> <p>Impute missing values in fields such as last_name, DOB, job_title, job_industry_category, default, and tenure.</p> <p>Update columns like deceased_indicator and owns_car to ensure currency.</p> <p>Assess the relevancy of columns such as deceased_indicator, default, job_title, and wealth_segment, and consider removing unnecessary columns.</p> <p>Validate and clean the data to ensure the default column contains allowable values.</p>
<b>Customer Address</b>	<p>Investigate and retrieve missing address data for customers with IDs [3, 10, 22, 23].</p> <p>Evaluate the relevancy of the Country column and consider removing it if it does not provide meaningful insights.</p> <p>Standardise state inputs to 'VIC' and 'NSW'</p>
<b>New Customer List</b>	<p>Address missing values in fields such as last_name, DOB, job_title, and job_industry_category.</p> <p>Evaluate the relevancy of columns such as deceased_indicator, Country, and default, and remove unnecessary columns.</p> <p>Validate the data in the default column to ensure it contains allowable values.</p> <p>Convert into numeric format all the numbers in string format</p>
<b>Transactions</b>	<p>Investigate and rectify outlier in the customer_id column.</p> <p>Investigate and rectify missing data in the online_order column.</p> <p>Investigate and rectify missing data in columns related to product_id 0.</p> <p>Assess the consistency of product_first_sold_date values and consider formatting them correctly.</p>

Please let me know if you have any questions or require further clarification regarding these data quality issues or the proposed mitigation strategies.

Moving forward, the team will continue with the data cleaning, standardisation and transformation process for model analysis. After we have completed this, it would be great to spend some time with your data SME to ensure that all mitigation plans are aligned with Sprocket Central's understanding.

Best regards,

Joaquim Bolós Fernández

Junior Consultant

KPMG