

Dear Client,

This is Rasmo Wanyama from KPMG AU Data Analytics team. I am writing to you in response to the following datasets your company (Sprocket Central PTY Ltd) that you shared with our team:

- Customer Demographic
- Customer Addresses
- Transaction data in the past three months

We have analyzed the quality of the raw data and found several data quality issues these datasets. They are presented in the table below.

Dataset	Total no. of Records	Unique ID Records
Customer Demographic	4000	4000
Customer Addresses	4003	4003
Transaction Data	20000	3427

We strongly believe that the following issues need to be addressed before we can continue with processing this data.

Customer Demographic Dataset	
Accuracy	- DOB Inaccurate for some records (One record had 1843 as the DOB)
Completeness	- Age column missing - Last Name, DOB and Job Title Column contains blanks
Relevancy	- Default column is irrelevant as it contains a mix of special characters, nulls, and blanks - Job Industry column contains records with 'N/A' which might not provide relevant information.
Consistency	- Inconsistent gender values (M, Male, F, Female, Femal, U).
Customer Addresses	
Accuracy	- Incorrect data types as numbers in some columns are captured as text.
Consistency	- States column is inconsistent. Contains both VIC and Victoria & NSW and New South Wales as different states yet they're same
Transactions	
Completeness	- Profit column missing - Online orders and Brands columns contains blanks
Relevancy	- Cancelled Status in the Order Status is irrelevant.
Validity	list_price and product_first_sold_date are of the wrong data format

To mitigate these data quality issues in the future, these are some of our recommendations:

a) ACCURACY ISSUES

i. Customer Demographics Dataset

One DOB record with an improbable year 1843.

Recommendation:

- Validate and correct the DOB entry with the improbable year or filter out outliers in the DOB Column.

ii. Customer Addresses

Incorrect data types for numeric entries in some columns are captured as text.

Recommendation:

- Review and adjust data types for accurate representation.

b) COMPLETENESS ISSUES

i. Customer Demographics Dataset

1. Age Column Missing
2. Last Name, DOB, and Job Title columns contain blanks.

Recommendation:

- For the missing Age column, consider calculating age based on the Date of Birth (DOB) and the current year using this excel formula:

=TRUNC(((NOW()-F2)/365),0)

- Populate missing Last Name, DOB, and Job Title entries with accurate information to ensure a complete dataset and double-check columns to ensure the columns do not have blanks.

ii. Transactions Data

1. Profit column missing
2. Online orders and Brands columns contains blanks

Recommendation:

- Calculate and include the Profit column to provide comprehensive financial insights.
- Populate missing entries in the Online Orders and Brands columns for a complete transaction record and double-check columns to ensure the columns do not have blanks.

c) RELEVANCY ISSUES

i. Customer Demographics Dataset

1. Default column is irrelevant as it contains a mix of special characters, nulls, and blanks
2. Job Industry column contains records with 'N/A' which might not provide relevant information.

Recommendation:

- Remove the column or correct entries with special characters, nulls, or blanks in the Default column.
- Review and potentially enhance the Job Industry column to provide meaningful and relevant data.

d) CONSISTENCY ISSUES**i. Customer Demographics Dataset**

Inconsistent gender values (M, Male, F, Female, Femal, U).

Recommendation:

- Standardize gender values to ensure consistency across the dataset (e.g., Male, Female, Unknown).

ii. Customer Addresses

States column is inconsistent. Contains both VIC and Victoria & NSW and New South Wales as different states yet they're the same.

Recommendation:

- Standardize state abbreviations and names for consistency (e.g., use VIC for Victoria, NSW for New South Wales)

e) VALIDITY ISSUES**i. Transactions Data**

list_price and product_first_sold_date are of the wrong data format

Recommendation:

Validate the two columns for accuracy to ensure they are of the correct format.

By addressing these grouped issues, we aim to improve the overall quality, accuracy, and relevance of our datasets and effectiveness of our analysis. If you require further clarification or assistance, please don't hesitate to contact us.

Regards,

Rasmo Wanyama
KPMG AU Data Analytics Team