

Dear Sprocket Central Pty Ltd employee,

thank you for providing us with the datasets from Sprocket Central Pty Ltd. The summary table below, highlights the main quality issues our team discovered within the datasets. Please let us know if you have any questions for the presented issues.

	<i>Accuracy</i>	<i>Completeness</i>	<i>Consistency</i>	<i>Currency</i>	<i>Orderliness</i>	<i>Relevancy</i>
Transactions	<ul style="list-style-type: none">• Missing column 'profit'	<ul style="list-style-type: none">• 'customer_id' incomplete• 'brand' & 'online_orders' have blank values			<ul style="list-style-type: none">• Incorrect formats for 'list_price', 'product_sold_date'	
New Customer List		<ul style="list-style-type: none">• 'Q', 'R', 'S', 'T' 'U' are missing column names		<ul style="list-style-type: none">• Deceased customers filtered out	<ul style="list-style-type: none">• 'postcode' has incorrect format	<ul style="list-style-type: none">• 'Q', 'R', 'S', 'T' 'U' filtered out
Customer Demographic	<ul style="list-style-type: none">• 'DOB' is inaccurate• Missing column 'age'	<ul style="list-style-type: none">• 'job_title' and 'customer_id' have blank values	<ul style="list-style-type: none">• Gender specification inconsistent	<ul style="list-style-type: none">• Deceased customers filtered out		<ul style="list-style-type: none">• Column 'default' deleted
Customer Address		<ul style="list-style-type: none">• 'customer_id' is incomplete	<ul style="list-style-type: none">• 'state' is inconsistent			

Below is a detailed description of the data quality issues we discovered, and methods of the mitigation we have provided. We have also included explanations and recommendations so future data quality issues can be avoided.

Accuracy Issues

- The table “CustomerDemographic” contains inaccurate data for the column ‘DOB’ (Date of birth) and was also missing columns for customer age and transaction profit.

Mitigation: Filter out extreme outliers in ‘DOB’.

Recommendation: Create an ‘age’ column for better readability and comprehension, and a ‘profit’ column for a better overview over the transactions.

Creating additional columns for age and profit, will help with errors, and the profit column will also assist with future monetary analysis.

Completeness

- The columns ‘customer_id’ were inconsistent within the tables “Transactions”, “CustomerDemographic” and “CustomerAddress”.

Mitigation: Filter all ‘customer_id’s from 1 to 3900.

Recommendation: Ensure all the data is up to date .

It is possible that the received data is not in sync across all spreadsheets, so the analysis results might be skewed because of the incomplete data. It is recommended to cross check spreadsheets and sync data, so the completeness issues do not occur.

- Blank spaces for the column ‘job_title’ in the table ”CustomerDemographic”, and ‘online_order’ in the table “Transactions”.

Mitigation: Blank spaces were filtered out for ‘job_title’ and ‘online_order’.

Recommendation: Create dropdown options for job titles and online order for simplicity.

Consistency

- Inconsistency in the columns ‘gender’ from the table “CustomerDemographic” and ‘state’ from the table “CustomerAddress”

Mitigation: Use filter to switch ‘M’ to show ‘Male’, ‘Femal’ for ‘Female’, ‘New South Wales’ for ‘NSW’ and ‘Victoria’ to show ‘VIC’.

Recommendation: Create dropdown lists for gender and states.

Dropdown lists minimize manual entries as well as human errors. Naming genders can be a sensitive topic, so caution is recommended when doing so.

Currency

- All the customers who are marked as ‘Y’ for yes in the column ‘deceased’, don’t belong to current customers anymore.

Mitigation: Filter out customers who are marked as deceased.

Recommendation: It is sometimes difficult to find out if a customer has deceased or not, but once the information is available, the data should be updated accordingly.

Deceased customers are not current customers, so removing them will result in an accurate dataset which will provide better analysis results.

Orderliness

- The formats for the columns 'list_price' and 'product_sold_date' in the "Transactions" table as well as the 'postcode' column in the "NewCustomerList" table were out of order.

Mitigation: *Formatting 'list_price' from text to currency, 'product_sold_date' from text to short date, and 'postcode' from text to number.*

Recommendation: *Set up the formats beforehand, before entering the data.*

Correct formats are easier to read, so it is important to have dates in the date format or data about finance in the currency format.

Relevancy

- No need to further analyze the nameless columns 'Q', 'R', 'S', 'T' and 'U' in the "NewCustomerList", and the column 'default' in the table "CustomerDemographic", is deleted.

Mitigation: *The columns 'Q', 'R', 'S', 'T' and 'U' contain data which results are listed in the column 'rank', so we can hide them. The column 'default' contains unrelated symbols, so we've deleted them.*

Recommendation: *Incomprehensible Metadata should be formatted to be comprehensible.*

This would be a summary of the first stage of the data analysis. The mitigation methods are simple but effective in a way that they not only simplify the further analysis for the KPMG team, but it will help other analysis companies you work with as well, in finding valuable insights for your company.

Please let me know if there are any uncertainties and questions regarding the mitigations or any other data quality issue identified.

Best regards,

Selmir Kalender