To,

The client

Sub: Data Quality Framework Table – quality issues

Dear client,

Thank you for providing datasets but there are few quality issues with the dataset, that we discovered. I am explaining them in the table below followed by the descriptive summary.

File name	Accuracy	Completeness	Consistency	Relevancy	Validity
Customer Demographic	DOB: inaccurate and age was also missing Default: had error values	Last name, DOB, Job title, job industry category and tenure had many empty columns, data was not complete	Gender: inconsistency	Default: deleted Empty value rows deleted	
Customer Address			State: inconsistency		
Transactions	Profit column was missing and has been calculated and added	Online order, brans, product line, product size, standard cost and product first sold out had empty (missing) values		All missing value rows has been deleted	Product first Sold out: dates were in wrong formats

More description in depth about data quality issues, that's been discovered. There are also recommendations and explanations that can help to avoid further data quality issues. These recommendations will help to improve the quality of data that will influence business decisions of Sprocket Central Pty Ltd in the future.

Accuracy issue: DOB was inaccurate for 'Customer demographic' and the age column was missing. The same way in 'Transaction' table profit column was missing

Recommendation- Create an age column in Customer Demographic and profit column in Transaction. Make data more comprehensive to understand the data clearly and easy to find the error.

Completeness: There were data missing in many columns of the table (Customer Demographic and Transaction). In Customer Demographic table, customers id was total 4000 but last name, DOB, job title, job industry category and tenure were respectively 3875, 3913, 3494, 3344 and 3913. So, the data was not complete. In Transaction table also, transaction ids were 20000 but online order were 19640 and brand, product details and even product first sold date and standard cost were 19803.

May be while syncing across the spreadsheet, the data has been missed. Because of missing data or incomplete data, the analysis cannot be correct or can mislead the result.

Recommendation- To avoid this in future, it is recommended to cross check or validate the missing values of the data.

Consistency: There was inconsistency in the data. While defining gender, there were used Male and also M, Female, Femal and F. Because of that sorting was not easy and data were coming in different categories. Same in state column, states were defined in their full name like, New South Wales and also in abbreviation NSW, Victoria and VIC.

Recommendation-To avoid that in future, I would recommend to use dropdown option. It will save you from human error in data and also bring consistency of terminology in the data.

Relevancy: Default column in Customer Demography was irrelevant and had crypted data, which had no meaning. It can bring noise in the analysis and can skew data.

Recommendation- To avoid irrelevant information, check for incomprehensible metadata and delete it or if possible format it to make comprehensible.

Validity: The format of the Product First Sold Date was wrong in Transactions table. It should be in date format to read and calculate for analysis.

Recommendation- Define columns format such as date, decimal, so that while filling data in those columns, it will appear in right format and interpretable. This will also bring accuracy and speed in analysis to make right decisions.

So that's the summary of data quality check and all the recommendations will help to improve data and also will increase the quality and reduce the error and time in analysis. The data that's been missing or irrelevant data, has been deleted to avoid any noise and any mistake in analysis. This will help to bring the result more understandable format.

If there is anything that is not clear or any question about the data quality check or recommendations, please write me back and I'll try to explain everything. Or if any data, that's been missed and deleted, you wanted to include in analysis, please let me know.

Thanks and regards,

Sugandh Bansal

KPMG analysis team