Forage KPMG Data Assessment

Dear **Kathleen**,
Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

I used Google Colab and Python to accomplish the task. I uploaded the dataset with the extension of .xlsx. It has four different sheets: Transaction, NewCustomerList, Customer Address and Customer Demographic.

Introduction:

Sprocket Central Pty Ltd , a medium size bikes & cycling accessories organisation, has approached Tony Smith (Partner) in KPMG's Lighthouse & Innovation Team. Sprocket Central Pty Ltd is keen to learn more about KPMG's expertise in its Analytics, Information & Modelling team.

Smith discusses KPMG's expertise in this space (you can read more here). In particular, he speaks about how the team can effectively analyse the datasets to help Sprocket Central Pty Ltd grow its business.

Primarily, Sprocket Central Pty Ltd needs help with its customer and transaction data. The organisation has a large dataset relating to its customers, but their team is unsure how to effectively analyse it to help optimise its marketing strategy.

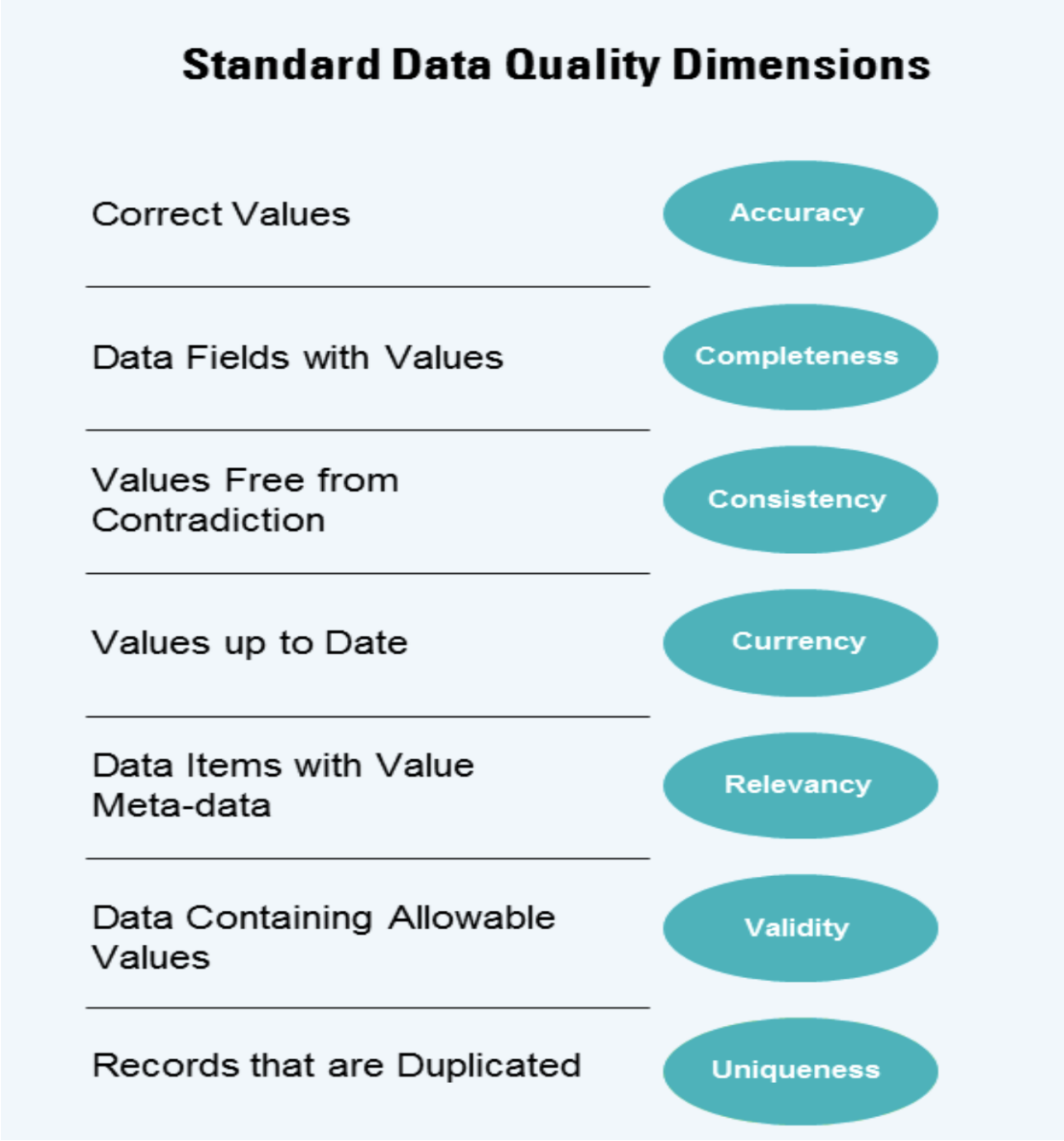
However, in order to support the analysis, you spoke to the Associate Director for some ideas and she advised that "the importance of optimising the quality of customer datasets cannot be underestimated. The better the quality of the dataset, the better chance you will be able to use it to drive company growth."

Objective:

I have been asked to do the preliminary data exploration task and identify ways to improve the quality of Sprocket Central Pty Ltd's data. The main task is to

assess the quality of the data and make some assumptions and provide recommendations to mitigate the issues.
Draft an email to the client identifying the data quality issues and strategies to mitigate these issues. Refer to 'Data Quality Framework Table' and resources below for criteria and dimensions which you should consider.

Data Quality Framework Table



Solutions:

Questions-Assumptions-Hypothesis-Recommendations

1. We can make a null hypothesis that there is no relationship between Property Evaluation and Value. I would check whether Tenurity has an impact on the value or not. Based on the z_score and p-value we can conclude the assumption that we made
2. If we look into the NewCustomerList dataset, we could regress the Value of the customer based on the explanatory variables. So we can assume that there is a linear relationship exists between the variables
3. In the transaction tab Product Class and Product Size have the same categories. It's redundant.
4. A few features from the Transactions tab could be significant to assess the customer value. So those variables can be added in NewCustomerList dataset considering the model building part
5. What is meant by the 'default' column in the demographic sheet?

I attached a .ipynb file considering the practical task in python. Here, I have written the workflow--

First, I will work on NewCustomerList Data as this will be used in the Data Modelling and Data Visualization tasks.
Check with the shape of the dataset
Check the description and information
Identify missing values and treat them accordingly

Then, I worked with the Demographic dataset.
Identified the inconvenient date value and dropped it.
Imputed the missing values with mode values for job_title and job_industry_category columns
For tenure column applied the mean imputation
Dropped the missing values in DOB and last_name columns as this won't be logical to impute those with mode values
Dropped default column as the records are not interpretable and didn't make any sense
Replaced the inconsistent values for the same category names in gender attribute

```
{ 'M' : 'Male', 'F' : 'Female', 'Femal' : 'Female'}
```

Coming to the Transaction dataset

```
txn_data = txn_data.replace({'online_order' : { 0.0 : False, 1.0 : True}})
```

Made online_order column binary by replace the category values
Mode Imputation for categorical features and Median imputation for continuous feature

Quality issue is not found in Customer Address dataset

Key Findings:

Additional customer_ids in the 'Transactions table' and 'Customer Address table' but not in 'Customer Master (Customer Demographic)'

Various columns, such as the brand of a purchase, or job title, have empty values in certain records

Inconsistent values for the same attribute (e.g. Victoria being represented as "V", "Vic" and "Victoria")

Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others)

Moving forward, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Questions will be raised along the way and assumptions documented. After we have completed this, it would be great to spend some time with your data SME to ensure that all assumptions are aligned with Sprocket Central's understanding.
Kind Regards,
Somnath Banerjee