**Task 1: Data Quality Assessment**

Here is the background information on your task

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 transactions 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 speak 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 drive company growth.”

The client provided KPMG with 3 datasets:

* Customer Demographic
* Customer Addresses
* Transactions data in the past 3 months

You decide to start the preliminary data exploration and identify ways to improve the quality of Sprocket Central Pty Ltd’s data.

**Here is your task**

You arrive at your desk after the initial client meeting. You have a voicemail on your phone which contains the following instructions.

[Voicemail transcript below]

“Hi there – Welcome again to the team! The client has asked our team to assess the quality of their data; as well as make recommendations on ways to clean the underlying data and mitigate these issues. Can you please take a look at the datasets we’ve received and draft an email to them identifying the data quality issues and how this may impact our analysis going forward?

I will send through an example of a typical data quality framework that can be used as a guide. Remember to consider the join keys between the tables too. Thanks again for your help.”

[Read email below]

Hi there,

As per voicemail, please find the 3 datasets attached from Sprocket Central Pty Ltd:

Customer Demographic

Customer Addresses

Transaction data in the past three months

Can you please review the data quality to ensure that it is ready for our analysis in phase two. Remember to take note of any assumptions or issues we need to go back to the client on. As well as recommendations going forward to mitigate current data quality concerns.

I’ve also attached a data quality framework as a guideline. Let me know if you have any questions.

Thanks for your help.

Kind Regards

Your Manager

**Here is your task:**

**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.

Dear [Client point-of-contact],

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.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | No. of records | Distinct Customer IDs | Date Data Received |
| Customer Demographic | 4,000 | 4,000 | 19/02/2023 |
| Customer Address | 3,999 | 3,999 | 19/02/2023 |
| Transaction Data | 20,000 | 3,494 | 19/02/2023 |

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

**Additional customer\_ids in the ‘Transactions table’ and ‘Customer Address table’ but not in ‘Customer Master (Customer Demographic)’**

**Mitigation:** Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model.

This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records. Please refer to excel file ‘data\_outliers.xlsx’ for the list of outliers between tables.

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

**Mitigation:** If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.

For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.

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

**Mitigation:** Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.

**Recommendation:** Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Additionally, gender records where ‘U’ have been replaced based on the distribution from the training dataset.

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

**Mitigation:** Convert selected records in characters to numeric. Remove non-numeric characters from string. Recommendation: Ensure that fact tables in the given database have constraints on data types.

Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | No. of records | Distinct Customer IDs | Date Data Sent |
| Customer Demographic | 3,911 | 3,911 | 16/03/2023 |
| Customer Address | 3,997 | 3,997 | 16/03/2023 |
| Transaction Data | 13,910 | 3,493 | 16/03/2023 |

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,

Idris