KPMG Virtual Internship Task 1

Dear Client,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The following table highlights the summary statistics from the three datasets. Please let us know if the figures align with your understanding.

Table Name	No. of Records	Distinct Customer IDs	Date of Data Reception
Customer Transactions	20000	3494	02/11/2020
Customer Demographic	4000	4000	02/11/2020
Customer Address	3999	3999	02/11/2020

Notable data quality issues that were encountered and the methods used to mitigate them are given below. Also, recommendations have been provided to avoid the re-occurrence of data quality issues and improve the accuracy of underlying data used to drive business decisions.

1. **Additional Customer_IDs in the Customer Transactions and Customer Address Sheets. Explanation:** Customer_ID of 5034 found in the Customer Transactions Sheet which is not present in the Customer Master sheet/ Customer Demographics Sheet. Additional Customer_IDs of 4001, 4002, and 4003 were found in the Customer Address Sheet and these are not present in the Customer Master Sheet. **Mitigation:** Please ensure that all tables are from the same period. Customers in the Customer Master Sheet will only be used for training our model. **Recommendation:** The data received may not be in sync with each other which may skew the analysis if there are missing records. Please refer to the excel file 'data outliers.xlsx' for the list of outliers between tables.
2. **Various columns have empty values. Explanation:** In the Transactions Table, brand, product_class, product_line and product_size have 197 missing values. In the Demographics Table, there are 87 missing values for DOB, 506 missing values for job_title, 656 missing values for job_industry_category, and 87 missing values for tenure. **Mitigation:** If a small number of rows are empty, then it can be removed from the dataset. If it is a core field, data has to be imputed accordingly. **Recommendation:** For the Transactions dataset, if less than 1% (totalling less than 0.1% of revenue) of the transactions have something missing it can be removed.

3. **Inconsistent values for the same attribute. Explanation:** We have the fields F, Femal, Female, M, Male and U for Gender. For the Customer Address Dataset, we have New South Wales, NSW and Vic and Victoria for the same attribute. **Recommendation:** Use regular expression to maintain consistency throughout the dataset. Put a drop-down list for the user instead of asking them to enter text. **Mitigation:** In order to construct meaningful variables for the model, data has been cleaned to make one variable and gender records with U have been replaced based on distribution from the training dataset.
4. **Inconsistent data types for the same attribute. Explanation:** Some of them are numeric and some are strings. Having different types of data types makes it difficult to interpret results at the later stage. Therefore, appropriate data transformations have to be made. **Mitigation:** Convert string characters to numeric. **Recommendation:** Put a constrain on data types.

Moving forward, the team will continue with 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,

Pradipti Thakur.