KPMG internship

Task 1

Dear Sir/Mam,

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	4000	4000	
Customer Address	3999	3999	
Transaction Data	20000	3494	

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.

- **The number of Customer_id is different in 'Customer Demographic' and 'Customer Address' table and one customer is missing.**
  Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Address 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.

- **Various columns such as brand, job title, family name and online order have empty values.**
  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.
  Missing values in brand, job title and online order has been filtered.

- **Inconsistent values for the same attribute (for example in gender and state columns)**
  Mitigation: use consistent value for values with more than two representation.
  Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field.

- **Invalid data type (for example text format instead of numeric)**
  Mitigation: change format of some columns.
  Recommendation: Ensure that fact tables in the given database have constraints on data types.

- **Irrelevant values (for example default column and cancelled order in order status column)**
  Mitigation: delete default column and filter cancelled orders.
  Cancelled order should not be consider because they may skew the analysis results.

- **Currency issue (in deceased column)**
  Mitigation: Filter deceased customer for training.

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,

Atefeh