KPMG Virtual Internship (Sprocket Central Property Ltd)

Module 1 Task (View and Identify Data Quality Issues)

Dear Sprocket Central Property Ltd,

Thank you for providing us with the four 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 (Rows)**	**No. of columns**
Customer Demographic	4,000	13
Customer Address	3,999	6
Transaction Data	20,000	13
NewCustomerList	1,000	18

Preliminary exploratory data analysis was done to view the quality of the provided datasets. Notable data quality issues encountered are listed below. Recommendations have also been provided to avoid the re- occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

Firstly, I had to rename the headers in each of the 4 datasheets for easy readability.

The Customer Demographic sheet had 4000 rows and 13 columns. 125 customers had their last name omitted. In the gender column, 1 record was filled in as F, 1 record as Femal, 2037 records as female, 1 record as M, 1872 records as male and 88 records as U. 87 customers omitted their DOB, one customer's DOB returned as 1843-12-21 (an outlier). 506 customers didn't indicate their job title while 656 customers filled n/a as their Job industry category. 2 customers were deceased, while others recorded were as alive. 1976 customers didn't own a car, 2024 owned. 87 customers had their tenure omitted. The default column seems to have errors, 240 records were omitted. 979 affluent customers in the wealth segment, 1021 for high net worth and 2000 for mass customer. Tenure ranged from 1- 22.

The CustomerAddress sheet had 3999 rows and 6 columns namely: Customer_id, Address, Postcode, State, Country and Property Valuation. Of the 6 columns, the Country column happened to be unique, Australia was the only country inputted. There were 3 different states (NSW, QLD & VIC), although some states were recorded as New south Wales in full (86), while 2054 records were abbreviated as NSW, 82 values were recorded in full as Victoria state, while 939 as VIC for some customers, QLD had 838 recorded values. Property valuation ranged from 1-12.

The Transaction sheets had 20,000 rows and 13 columns. Product_id ranged from 0 to 100. Transaction date was from January to December, 2017. 360 online orders were omitted, 9811 recorded with 0 code while 9829 were recorded with 1 code. 19821 orders were approved, 179 orders cancelled. There was a total 6 Brands, 3312 for Giant bicycles, 2910 for Norco bicycles, 3043 for Ohm cycles, 4253 for Solex, 2990 for Trek bicycles, 3295 WeareA2B, 197 missing values in the brand column. For product line column, 4 in all, 423 for mountain, 3970 for road, 14176 for standard & 1234 for touring, 197 missing values in the product line column. There were 3 classes in the product class column, 3013 for high, 2964 for low, 13826 for medium & 197 missing values. 3

product sizes, 3976 for large, 12990 for medium, 2837 for small, & 197 missing values. There were 197 missing values in the product first sold date column & standard cost column.

The NewCustomerList sheet had 1000 rows, 18 useful columns. Columns 17 to 21 werent properly named, thus understanding of these 5 columns posed a threat. 29 customers had their last name omitted. There were 470 male customers, 513 female customers while 17 customers were inputted with Unknown (U) gender. The DOB column had dates in different format, 17 customers didn't indicate their DOB. 106 customers didn't indicate their Job title, while 165 customers filled n/a as their Job Industry category. None of the customers were recorded as deceased. 507 customers didn't own a car, 493 did. Tenure ranged from 0 to 22. Records for state showed 506 for NSW, 228 for QLD and 266 for VIC. Property value ranged from 1 to 12. The wealth segment had 3 categories, 241 were recorded as affluent customers in the wealth segment, 251 for high net worth and 508 for mass customer.

Some of the encountered issues as well as identified data inconsistencies and recommended mitigation methods are as follows:

- 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.

- 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 (totaling 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 same attribute (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.

In conclusion, the team will continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. Hypothesis will be raised and observations documented. After we have completed this, it would be great to spend some time the data SME to ensure that all assumptions are aligned with Sprocket Central's understanding.

Kind Regards,
Tolulope Ogunsami.