Hello,

This is Eiti Mittal from KPMG Data Analytics-Virtual Internship team. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset\_Name | No. of Records | | Analysis |
| Before | After |
| Transaction | 2000 rows  13 columns | 19803 rows  14 columns | -Total profit: $10,930,284  -‘Solex’ is the most  purchased brand name  -The most and least sold product line is ‘Standard’ and ‘Mountain’ respectively |
| NerCustomerList | 1000 rows  23 columns | 978 rows  18 columns | -Most new customers are from the New South Wales, Australia  -Most customers own cars |
| Customer\_demographic | 4000 rows  13 columns | 3895 rows  13 columns | -Most customers are ‘mass customers’ in wealth segment  -Most customers are working in manufacturing and financial services industry |
| Customer Address | 3999 rows  6 columns | 3999 rows  6 columns | -Most customers are from New Sales Wales (NSW) |

# Some of the data quality issues have been solved by applying the following modifications in datasets-

# 1)Transactions-

# Less than 1% of transactions records (totaling less than 0.7% of revenue) have missing fields. Therefore, we have deleted the records containing missing values in several columns simultaneously. For eg there were 197 rows out of 20,000 where fields like 'brand', 'product\_line', 'product\_class', 'product\_size', 'standard\_cost', 'product\_first\_sold\_date' were missing altogether.

# Added a new column containing the profit earned of each product. The net profit generated is $10930283.9649719

# 2)NewCustomerList-

# Deleted unnamed fields.

# Replaced ‘U’ with ‘Unspecified’ in gender column.

# Removed records where both ‘job\_industry\_Category’ and ‘job\_title’ and missing together.

# The remaining empty values ‘job\_industry\_category’ can be populated using other records of similar ‘job\_title’.

# 3)Customer\_Demographic

# Additional column named ‘default’ has been removed as is consisted of trash values.

# Removed records where both ‘job\_industry\_Category’ and ‘job\_title’ and missing together.

# Gender containing values ‘F’, ‘Femal’, ‘Female’ were all replaced by ‘Female’ and values ‘M’, ‘Male’ by ‘Male’ to reduce data redundancy.

# Added a new column containing ages of all customers.

# Data type transformations such as DOB in series to datetime object are made to ensure consistent data types for a given field.

# Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies 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, online order 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.

# ● 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.

# ● 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.

# 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.

Regards,

Eiti Mittal