**Note:** The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

Dear Sprocket Central Pty Ltd.,

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* | *2023-07-20* |
| Customer Address | *3999* | *3999* | *2023-07-20* |
| New Customer List | *1000* | *1000* | *2023-07-20* |
| Transaction Data | *20000* | *3494* | *2023-07-20* |

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 re- occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

**OBSERVATIONS IN THE DATASET FOR:**

1. COMPLETENESS:

(Completeness: How much information all entities have. Number of missing values.)

# In Transactions table, order\_online column has about 1.83% of null values i.e., there are 360 records in which order\_online was not captured

# In Transactions table, the columns: brand, product\_line, product\_class, product\_size, standard\_cost, product\_first\_sold\_date also has a percentage of 0.995% missing values that is 197 null values, which should not be missing if product\_id is inherited and the details of the product cannot be missing.

* In NewCustomerList table, the columns: last\_name has 29 null values, DOB has 17 null values, job\_title has 106 null values and job\_industry\_category has 165 null values.
* In NewCustomerList 19.76% of job\_industry\_category values are missing almost equal to CustomerDemographic which is 19.61%.

# 11.85% of job\_title values are missing in NewCustomerList a little less as compared to CustomerDemographic that has 14.48% of missing values.

# 3.22% of last\_name values were missing in CustomerDemographic yet 2.98% of last\_name values are missing in NewCustomerList.

* CustomerDemographic has 2.22% of missing DOB values which is slighlty decreased to 1.72% NewCustomerList.
* There is a 2.22% of missing tenure values in CustomerDemographic but there are no missing values of tenure in NewCustomerList.
* CustomerDemographic has 2.22% of missing DOB values which is slighlty decreased to 1.72% in NewCustomerList.

1. CONSISTENCY AND VALIDITY:

(Consistency: How consistent is your Data. And, Validity: Validated data with allowable values.)

* **In Transactions table, the column product\_first\_sold\_date should be in the standard \*\*datetime64\*\* format but is of datatype ‘float64’**

*Mitigation: The datatype of product\_first\_sold\_date is changed to datetime64*

* In DOB column of transactions table, the standard date format (datetime64) used to capture DOB of customers is ‘YYYY-MM-DD’. But, the date values in the dataset were typed in format ‘MM/DD/YYYY’
* **Data Captured in Gender column in the dataset CustomerDemographic is not consistent. It should be "Male", "Female" and "U" as per the NewCustomerList. But, it has the following categories: Female – 2037, Male – 1872, U – 88, F – 1, Femal – 1 and M – 1.**

*Mitigation: All the categories were renamed and grouped to 3 categories :– Male - 1873, Female - 2039 and U (Unspecified) – 88*

* **Number of columns in NewCustomerList are inconsistent because there are 4 columns which are Unnamed and they contain some values as well, however are not labeled so cannot be identified.**

*Mitigation: These 4 columns were dropped.*

1. RELEVANCY:

(Relevancy/Auditability: Relevant data in your entities. Also used to measure number of irrelevant values.)

* **In CustomerDemographic table, the default column is of ‘object’ datatype but this column was not captured after and is not present in ‘NewCustomerList table’. Some of its values are dates but most of them are irrelevant making the whole column inconsistent.**

*Mitigation: Default column is dropped from the dataset.*

* Online Order is captured in a column of ‘float64’ datatype, however the values are ‘boolean’, that is true and false.
* The column ‘Value’ in NewCustomerList is captured in float64 datatype but this was not captured before and is not present in CustomerDemographic table or CustomerAddress table.
* Deceased Indicator column in both NewCustomerList table and CustomerDemographic table is of object datatype but It only has Boolean values i.e., yes or no.

1. ACCURACY:

(Accuarcy: How accurate is your Data i.e., number of errors in the data.)

* One date value is wrong. 1843 year is not possible.

1. UNIQUENESS:

(Uniqueness: How much uniques is your data. Number of duplicated values.)

* All the records of all the tables in the dataset are unique.

1. TIMELINESS:

(Timeliness: Updated data/Current data.)

* The transactions table data were collected in the year 2017.

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,

Vidhyambika