Dear Mr Smith,

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. Below I shall be highlighting the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding

Customer demographic data set-

1. DOB column has 87 blank entries or missing entries,Tenure field has 87 blank entries and 506 customers have blank job titles.

2. The Customer with the customer id 34 has a birth year of 1843 which is practically impossible as that would mean he/she is 174 years old.

3. 88 customers have gender defined as U.

4. Job industry has been mentioned as n/a for 656 customers.

5. Since Gender column is a categorical data column it would be better and easier for analysis if all the values were in the form of F/M or Female/Male.

6. Customers with customer id 753 and 3790 have yes as their deceased indicator, it would be better for only customers who're still alive to be present in the dataset.

7. The default column in the dataset seems unnecessary or a column out of an error. So it's irrelevant.

8. Age column is missing.

New Customer list data set-

1. 29 customers have missing last names.

2. DOB column is blank for 17 customers.

3. 106 customers haven’t specified their job titles and job industry category is n/a for 165 customers.

4. Age column is missing from the dataset.

Customer Address data set-

1. the state column has state codes for some data rows and entire state names for the other. Since it is a categorical data it is better to have all the state entries as state codes.

Transaction data set

1. Online order column has blank values for 360 customers,197 blank values for product line, product size and product class columns

2. Product_first_sold_date column should have values in the date format but it has entries in the form of 6 digit numbers which can’t be distinguished as dates.

3. Standard cost column has some values like 312.7350159,667.4000244 while other values are in dollars and there are 197 blank values.

4. Profit Column is missing from the dataset.

5. Present in the transactions dataset are also cancelled transactions. Analysis on this dataset is primarily focused on only approved transactions with the company. So the inclusion of cancelled transactions seems irrelevant.

Mitigation Measures

1. Various columns, such as the brand of a purchase, or job title, have empty values in certain records. 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 data set.

2. Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”). 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 removed while modelling.

3. Inconsistent data type for the same attribute (e.g. numeric values for some fields and strings for others.Having different data types for a given field makes it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field.

4. All blank, missing and n/a values have been removed since they constituted a very small portion of the entire data set.

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.

Kind regards,

Inalegwu John Ocho

[Junior Consultant]