**Sub:** Data quality issues and observations

Hi,

In this mail, I have focused fully on the data quality, observations, identifications which had been made regarding the data and the strategies which we would be taking to overcome the issues.

I would request you to go through all the 12 points mentioned below and get back with you views and opinions about them.

**“Sprocket Central Pty Ltd”** Data Quality *issues*, *observations* and *strategies*:

1. The **“NewCustomerList”** data needs **“customer\_id”** for each new customers. This is because **“customer\_id”** is an unique reference to each customers.We might not be able to fetch their transaction details from the DB if their ID’s are unknown.
2. Some column labels are missing from the **“NewCustomerList”** data. The column addresses are **Q, R, S, T, U**. As they all seem to be numeric index values, they might contain high information value and we can’t ignore them at all. These columns along with **“Rank”** and **“Value”** columns in the **“NewCustomerList”** data are not present in the old customer data as well.

As these columns are newly added in the **“NewCustomerList”** data, getting the old customer data related to these columns are important for customer profiling.

1. The columns **“job\_title”** and **“job\_industry\_category”** in **“CustomerDemographic”** and **“NewCustomerList”** have some missing data-points. If left empty without information, those data-points will be removed straight away if we perform a Market segmentation to identify potentiality.
2. There are some incorrect data in **“CustomerDemographic”** data. The **DOB** for **“customer\_id”: 34** is **1843-12-21.** A non-biblical human being is highly unlikely to be around 174 years and having a **“deceased\_indicator”: N** at the same time. Some **DOB** data-points for the customersare missing as well**.** We might be able to create a feature **Age** for modeling or visualization purpose if we get all of them for every customers.We will remove the **DOB** data-points which are **NULL** if unable to be provided.
3. The **“NewCustomerList”** data contain both thecustomer demographic andcustomer address data.We will look forward to merge **“CustomerDemographic**” and **“CustomerAddress”** data-sheet on **“customer\_id”.** After successfully receiving all the customer ids for the **“NewCustomerList”** data-sheet, we will merge it with the old prepared customer list.
4. The basis of classifying customers according to their wealth is unclear. Taking an example of **“customer\_id”: 108,** the customer is marked as **“High Net Worth”** under the **“wealth\_segment”** column. The customer is seen to fall under a valuation bin of **8** whereas the highest class value for **“property\_valuation”** in the **“CustomerAddress”** data-sheet is **12.** That customer doesn’t own a car as well as per **“owns\_car”** data in **“CustomerDemographic”** sheet. A detailed understanding on this feature is required.
5. The **“default”** column in **“CustomerDemographic”** contain weird encoded characters. They are irrelevant and are not useful at all to solve the desired problem.
6. The **“gender”** column in **“CustomerDemographic**” data has to be modified. We will replace the **M** and **F** class with **Male** and **Female** respectively in order to follow the **“gender”** format of **“NewCustomerList”** data-sheet.
7. The column **“deceased\_indicator”** in **“CustomerDemographic”** and **“NewCustomerList”** datahas only one class **N**. As there is no variation in the column in terms of values, we will not consider this column.
8. Some of the **tenure** data-points in **“CustomerDemographic”** data-sheetare missing. Some data-points under **“online\_order”** column in **“Transactions”** data-sheet are missing as well. As the missing data-points are handful, we might replace the **NULL** values with **Median**.
9. The **“product\_id”: 0** in **“Transactions”** data-sheet have some missing data-points. There are NULL values under the columns **“brand”, “product\_line”, “product\_class”, “product\_size”, “standard\_cost”** and **“product\_first\_sold\_date”** all at the same time for some data-points. If those missing data can’t be provided, we might be able to perform a **Linear Regression** and estimatethe **NULL** values for the **“standard\_cost”** against **“list\_price”** as the corresponding **“list\_price”** areavailable in the data.
10. The **“property\_valuation”** column in **“NewCustomerList”** are provided in **String** format and contain some **Float** values as well.The desired format for the column has to be of type **Integer** as provided in the old customer demographic data. We will change the datatype of the column.

Thank You,

Regards,

Adipta Biswas

Virtual Intern @KPMG