*Hi,*

*Hope you are doing great. This is with reference to the improvement in the data sets sent to us for preliminary data quality analysis. We received the following three data sets:*

* *Customer Demographic*
* *Customer Addresses*
* *Transactions data in the past 3 months*

*After reviewing we found the following issues and how to mitigate these issues;*

***CUSTOMER DEMOGRAPHIC***

* *In the* ***gender*** *column, the classes are supposed to be Male/Female/U. Since some of the rows have value F/M, this will create inconsistency while further analysing the model. Replace all the M’s with Male and F’s with Female.*
* *An extra column* ***Age*** *should be added so that in further model building we can use it as a feature. It will be feasible to use age as a factor than date of birth. Age can be calculated as (NOW()-column date of birth)/365.*
* *In the* ***job\_title*** *field, there is incompleteness in form of rows with blank values. Adjust the same*
* *In order to maintain that the data is up to date, alter the* ***deceased*** *column and eradicate the rows where the customer is deceased.*
* ***Default*** *column cannot be comprehended. Deletion of the same would be preferable.*

***CUSTOMER ADDRESSES***

* *There were inconsistent values in the* ***state*** *column. Replace New South Wales with NSW and Victoria with VIC. This will ensure smoother data analysis.*

***TRANSACTIONS***

* *Column* ***Online\_order*** *contains blanks- incompleteness. Also, to predict further growth I am assuming that all the orders are online. Remove null and false values for this column.*
* ***Order\_status*** *column contains invalid data as the orders a=which are cancelled can no more contribute to predict revenue. Delete the rows for which order status is cancelled.*
* ***Brand*** *column has blank values.*
* ***Product\_first\_sold\_date*** *should be formatted to short date syntax for better understanding.*
* *A new* ***profit*** *column should be added which will help to predict increase in sales.*

*Kindly, give a view on the above-mentioned strategies and update the 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. 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,*

*Tony Smith*

*(Partner) KPMG’s Lighthouse & Innovation Team.*