Dear Customer,

I am writing this mail with the recommend solution and identified the issues in the 3 Dataset which is provided by the Sprocket Central Pty Ltd. Please find the Changes and the files attached.

- KPMG_VI_output.xlsx
- KPMGDataQualityASS1.ipynb

The following Recommended Changes will give a serious of Good Impacts on the Data. Those Data are now data quality error free. Please follow the instructions to change the data.

3 Set of Portions are created for the 3 Datasets.

In **Customer Demographic** Data, we found some data quality issues that directly affect the accuracy and the result of the business outcome. So the following list of points are mentioned below illustrate the issues and the metrics taken to solve the data quality issue.

- ✓ Initially we have 13 data attributes in the data set with values, Specifically" ***Default***" is one of the attribute those values are standard less, neither numeric nor Strings. It contains in symbols which cannot be understand or transformed to any format. So we drop that particular attribute.
- ✓ In "***Gender***" we have six data attributes *{Female, male, U, Femal, M, F}* from this data we can understand that Gender has '3' attributes and nominated in different ways so data transformation has injected we got the outcome of *{Female, Male, Unknown}* for better Understanding and Results.
- ✓ Handling of Missing Values is a tedious process, based on an attribute we can handle in a specific way, here "***Last Name***", "***Job Title***", "***Job Industry Category***" having missing values. In the whole dataset our main focus is on Customer Id, First Name which are absolute values. So According to this data if we tried to change or modify the 'Last Name' ,"job title" ,"job industry category" that will make an Error in details which leads to Inaccurate Statements and Results so the remedy for this attributes are Leave as it is.
- ✓ "***Age***" is the new column that has been created using "***DOB***" by calculating age from the date of birth which can be used for easy operation missed values represent as 0.
- ✓ Missing Values data can be drop or we can replace the fields using functions.

That's All for this Data.

In **Customer Address** datasheet, we found some data quality issues that directly affect the accuracy and the result of the business outcome. So the following list of points are mentioned below illustrate the issues and the metrics taken to solve the data quality issue.

- ✓ Around 6 Data Attributes in the Sheet, we found 5 are fine values to carry the operations here one data need to change.
- ✓ "***State***" is the one needed to change. It has totally 3 different variables but categorized as 5 we need to rectify using replace function and placed in proper category for accuracy.
- ✓ {NSW, VIC} changed to {New South Wales, Victoria} for a described meaning so State has now {New South Wales, Victoria, QLD}.

That's All for this Data.

In **Transactions** datasheet, we found some data quality issues that directly affect the accuracy and the result of the business outcome. So the following list of points are mentioned below illustrate the issues and the metrics taken to solve the data quality issue.

- ✓ 13 Data Attributes found in the Dataset. Here ***transaction_id, product_id, customer_id, transaction_date, order_status, list_price*** are finite values without changing we can proceed rest need some alteration.
- ✓ Here ***product_first_sold_date*** is in integer, that format is changed into date format while using in excel sheet that is easy for calculation
- ✓ ***Online_Order*** is having missing value, those values are filled with the mean.
- ✓ ***List_price and standard_cost*** are not enough to make a desirable solution, so we injected a new attribute called ***profit*** from ***List_price - standard_cost.***
- ✓ ***Brand, product_line, product_class, product_size, standard_cost, product_first_sold_date*** are the data attributes contains categorical values with **0.98** which is less than 1 percent of the total value, so modification of values may cause a lack of accuracy. To resolve that missing values are eliminated which makes a data standard rate.
- ✓ Total of 20000 is reduced to 19803 with all the filled values for the better accuracy.

From this we can conclude that Data Quality Issues are addressed and the Solutions are presented in the above section for the 3 Datasets Provided. By applying these Changes those datasets become a Standard Data for any Operations.

Kindly let us know the Feedback. Glad to accept the Comments.

With Thanks
Sandeep Malayalan
KPMG Data Analytics Virtual Intern.