**Re: Data Quality Assessment Review**

Greetings Sprocket team,

The KPMG Data Analytics team have successfully completed the data quality assessment task. Some datasets were joined in order to ensure consistency of information.

Here are our findings:

**I.) Old Customers Dataset**

- Customer Demographics Dataset joined with customer addresses on customer\_id column.

*a.) Relevance*

- Default column holds no sensible information and therefore can be dropped.

- The country column is redundant since it only has one value; Australia hence can also be dropped.

*b.) Accuracy*

- A customer is recorded to be 179 years old which is inaccurate. This can be rectified by changing the DOB year from 1843 to 1943 as it is most likely a typing error.

*c.) Uniformity*

- Gender values have been represented using various methods. ‘Female’, ‘F’, and ‘Femal’ can be assumed to represent Female and ‘Male’ and ‘M’ can be assumed to represent Male. The gender U can be assumed to represent Unknown.

- Similarly, state values ‘Victoria’ can be represented as ‘VIC’ and ‘New South Wales’ can be represented as ‘NSW’.

- Both state and gender values can have a drop down in order to avoid errors while entering values.

- DOB column can be converted to date data type as it represents date information.

*d.) Completeness*

- Rows with only location information(address, postcode, state, property\_valuation) : These 3 rows can be dropped.

- Rows missing last\_name information: Can be imputed with NA meaning Not Applicable as it has no great significance to our analysis.

- Rows missing job category values but have job titles: Can be imputed with most common job category for the title provided.

- Rows missing job titles but have job categories: Can be imputed with top 5 job titles depending on the category provided.

- Rows missing both job titles and job categories: Can be imputed with NA since this information cannot be determined.

- Rows missing DOB: First, we can generate age column and then impute missing age values with mean age values with respect to job industry category, job title and state. The age value can then be used to calculate the DOB.

- Rows missing tenure values: Can be imputed with average tenure value by age since there is a strong positive relationship(covariance) between age and tenure.

- Rows missing location information: Can be imputed with NA since they hold other important information and therefore cannot be dropped.

*e.) Consistency*

- Data has no duplicate rows.

**ii.) New Customers**

*a.) Relevance*

- Unnamed: 16 column has similar values to rank hence can be dropped.

*b.) Accuracy*

- The data provided seems to be accurate.

- Customer id data is missing but is represented in both transactions and customer demographics tables.

*c.) Uniformity*

- DOB column can be converted to date data type.

- Postcode column can be converted to string since it holds no numerical significance.

*d.) Completeness*

- Rows missing job category and have job titles with no job category representation in the data can be imputed with most common job category in the dataset.

- Other missing row values can be imputed similarly to the above dataset.

*e.) Consistency*

- Data has no duplicate rows.

**iii.) Transaction Data in the past 3 Months.**

- Joined with state column from old customers data on customer\_id column.

*a.) Relevance*

- Drop transaction\_id column which is replica of index.

- We need to investigate what the product\_first\_sold\_date column represents as it may be wrongly labeled.

- No transactions information is provided for the new customers.

*b.) Uniformity*

- We can convert transaction date to date data type.

- We can also convert standard cost to float data type by removing $ and commas.

- We can restructure online order values such that ‘1’ represents True and ‘0’ represents False.

*c.) Accuracy*

- The data contains records for 11 months yet stated to have only 3 months.

- Products with similar product attributes have different product Ids.

*d.) Completeness*

- Rows missing state values: Can be imputed with the state with the highest count.

- Rows missing product information(product\_id, brand, product\_line, product\_class, product\_size, standard\_cost): Can be imputed using the most common product information by state.

- Rows missing online\_order information: Can be imputed by the value of how frequent the product is bought.

*e.) Consistency*

- The data contains no duplicates.

**Summary**

The following cases need further clarification:

1. Investigate why missing DOB also have missing tenure values, have gender represented as U and are mostly in IT job category.

2. Investigate why products with similar product attributes have different product Ids.

Kind regards,

Elizabeth Kihungi,

KPMG Junior Data Analyst.