**Subject: Preliminary Data Quality Analysis for Sprocket Central Pty Ltd**

**Hello there,**

This is K.Poornima from KPMG Data Analytics team. We have reviewed the datasets and performed the necessary quality checks based on KPMG's standard data quality framework key points. The datasets were checked for accuracy, completeness, consistency, currency, relevancy, validity and uniqueness. Below, we have listed the data qualities issues we faced together with our recommended strategies in mitigating the aforementioned issues.

## Transactions dataset

***General Overview:***
There are 13 columns, 20000 rows. Data types are int64, datetime64, float64 and object with no duplicate records. The data is relevant and consistent.

***Issues Faced:***
- **'product_first_sold_date'** column isn't in the appropriate date time format
- Missing values in columns = **'online order','product_line', 'product_class', 'product_size', 'standard_cost', 'product_first_sold_date'.**

***Strategies to overcome the issues***:
- The required column will be converted into appropriate datetime format
- As the percentage of missing values are not alarmingly high, they will not be dropped but can be treated with mode/mean values.

## Customer Demographic dataset

**General Overview:**
There are 4000 rows (unique values of 'customer-id') with 13 columns. Data types are int64, datetime64, float64 and object. No duplicates.

***Issued Faced:***
- Missing values in columns = **'last_name', 'DOB', 'job_title', 'job_industry_category', 'default' and 'tenure'.**
- Inconsistencies in the **'gender'** column
- Irrelevant and incomprehensible data in **'default'** column
- There is an error in the year of the **'DOB'** column.
- More clarification needed in regards of how **'wealth-segment', 'owns_car' and 'tenure' columns** are related.

***Strategies to overcome the issues***:
- For columns with missing values, we would employ replacement methods such as KNN imputer or through statistical techniques like mode or mean. Below are the columns which have more than 5 percent of missing values:

| Column | % of missing values |
|---|---|
| Job_title | 12.65 |
| Job_industry_category | 16.40 |
| Wealth_segment | 7.55 |

- To synchronize the values in the **'gender'** column, we will change 'female' to 'F', 'femal' to 'F', 'Male' to 'M' and 'U' to 'NA'.
- **'default'** column will be dropped as it is irrelevant to our analysis.
- Ages of the customers can be calculated from the **'DOB'** column. As there is an error of mistype error, that would be rectified too. From 1844 to 1944.

**Customer Address Dataset**

**General Overview:**
There are 3999 rows with 6 columns. Data types are int64 and object. No duplicates.

***Issued Faced:***
- The 'state' column has names of states in full form and abbreviations mixed up.

***Strategy to overcome***:
- The following will be changed: NSW to New South Wales, VIC to Victoria, QLD to Queensland

**New Customer Dataset**

**General Overview:**
There are 1000 rows with 22 columns. Data types are int64, datetime64, float64, int64 and object. No duplicates.

***Issued Faced:***
- Missing values in columns = **'last_name', 'DOB', 'job_title', and 'job_industry_category'**
- The values in **'gender'** column to be replaced to match with the old customer details.
- There are some unnamed columns with data which are vague to other columns.
- There is no unique customer id assigned to any of the customers.

***Strategies to overcome***:
- For columns with missing values, we would employ replacement methods such as KNN imputer or through statistical techniques like mode or mean.
- To synchronize the values in the **'gender'** column, we will change 'female' to 'F', 'Male' to 'M' and 'U' to 'NA'.
- The unnamed columns will be dropped.
- We can assign a new column 'customer_id' where the unique ids will be a continuation of the old customers'.

If the missing data can be filled up from your side, kindly proceed so that it will aid our analysis at later stage.

Since **'customer_id'** column is present in ***Customer Demographic, Customer Address and Transactions datasets,*** we can perform an inner join on 'customer_id' to join the datasets moving forward.

In case of any queries or feedback, please do not hesitate to contact us.

Thank you.

**Best Regards,**

K.Poornima
Virtual Intern
KPMG