Dear Sprocket Central,

My name is Mario Paul and I am a Data Science intern for KPMG.

I was tasked with assessing the three datasets on customer demographics, addresses, and transaction data in the past three months that we have received from you. This is a process that is done before we begin to work with any data we receive to ensure we can get the most accurate results and answers we are seeking.

The below table highlights the summary statistics from the three datasets received. Please let us know if the figures are not aligned with your understanding.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | No. of Records | Distinct Customer ID’s | Date Data Recieved |
| Customer Demographic | 4000 | 4000 | 04/25/2020 |
| Customer Address | 3999 | 3999 | 04/25/2020 |
| Transaction Data | 20000 | 20000 | 04/25/2020 |

Notable data quality issues that were encountered and the methods used to mitigate the identified data inconsistencies are as follows. Furthermore, recommendations have been provided to avoid the re-occurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions

**Customer Demographic:**

**Not Accurate:**

Jephthah Bachmann DOB claims he was born in 1843

**Not Consistent:**

Gender:

1 entry is ‘Femal’

1 entry is ‘M’

1 entry is F

**Not Valid:**

Gender:

88 entries are ‘U’

**Completeness:** The following Columns have missing / null values:

**Last\_name 125 (0.03125%)**

**DOB 87 (0.02175%)**

**Job\_title 506 (0.12650%)**

**Job\_industry\_category 656 (0.16400%)**

**default 302 (0.07550%)**

**tenure 87 (0.02175%)**

**Not Current:**

Two of the customers are deceased, Josey St. Quentin and Kurtis Morson.

**Relevancy:**

“Default column is not relevant.

**Customer Addresses**

**Not Consistent:**

Customer\_id’s not consistent

Address house / building numbers are not consistent. Some single digit numbers have a 0 in the front of them and some do not.

Some addresses are not correct. Some addresses start with the number 0

**Relevancy:**

Property\_valuation column is not revellant. Has nothing to do with address information.

Country column is not needed since all entries are from the same country.

**Transaction data**

**Completeness:** The following Columns have missing / null values:

**online\_order 360 (0.01800%)**

**brand 197 (0.00985%)**

**product\_line 197 (0.00985%)**

**product\_class 197 (0.00985%)**

**product\_size 197 (0.00985%)**

**standard\_cost 197 (0.00985%)**

**product\_first\_sold\_date 197 (0.00985%)**

**Not Current:**

Values are supposed to be transactions from the last 3 months. They are dated from 2017.

**Relevancy:**

There are 179 cancelled transactions where are not needed

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.

Kind Regards,

Mario Paul