Dear Manager,

RE: REVIEW OF DATA QUALITY OF DATASETS.

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. The table below 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 IDs	Date Data Received
Customer Demographic	4000	4000	Unknown
Customer Address	3999	3999	Unknown
Transaction Data	20000	20000	2017

After a thorough assessment of the three datasets here are some data quality issues identified:

- **Data Completeness issues:**

Data in Transactions and Customer Demographics sheets appear incomplete, these incompleteness can be found in columns: 'last name', 'DOB', 'job title', 'job industry category', 'default' and 'tenure' in the Customer Demographics dataset all have null values, same with columns: 'online order', `brand', 'product line', 'product class', 'product size', 'standard cost', 'product first sold date' in the Transactions Dataset.
Mitigation: depending on the importance of these columns, solving missing values issues could vary. If the columns aren't important for the subsequent analysis, consider dropping the columns entirely. This preserves the information in the other more important columns. Else, you could consider filling in the missing rows based on the distribution of the column. This is done where the number of missing value is minute.

- **Data Validity issues:**

In the Customer demographics dataset, validity of entry in `product first sold date` column are questionable as the data therein aren't in date format. Default column in Customer Demographic dataset also have data in special character format. This validity issues will need further clarification.

- **Data Consistency issue**:

Inconsistency issues arise in DOB column in Customer Demographic sheet where a customer has a value of '1843' as Date of Birth. Also, in the Customer Address dataset, the address column has inconsistent values for the same attribute (e.g Victoria being represented as "V", "Vic" and "Victoria"). This could be solved by enforcing a drop-down list for the user entering any categorical data.

Subsequently, the team will continue with data cleaning, standardization and transformation process for the purpose of model analysis with clear documentation along the way. Afterwards, 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,
Nwangene, Andrew