Dear Manager,

Many thanks for attaching the data quality framework. Based on that, I have reviewed the data from Sprocket Central Pty Ltd, summarised the issues and recommendations of mitigating current data quality concerns.

I will demonstrate the issues through seven data quality dimensions depending on the framework table and the contents are shown below:

Table name	No of records	Distinct Customer IDs
CustomerDemographic	4000	4000
CustomerAddress	3999	3999
Transactions	20000	3494

Notation: CustomerDemographic simplified by CD, CustomerAddress simplified by CA, Transactions simplified by T.

Correct values (Accuracy):

1. **DOB(CD)**: There is a person born in 1843, which is clearly wrong.
   Mitigation: Ask clients for the correct values, if not possible, try to drop the value or replace the value by the average date of born.

2. **Gender (CD)**: A few errors occur in the gender like 'F', 'Femal','M'.
   Mitigation: Replaced by 'Male' and 'Female'.

3. **State (CA)**: the 'VIC' and 'Victoria', 'New South Wales' and 'NSW' belong to same values.
   Mitigation: 'Victoria' and 'New South Wales' should be replaced by 'VIC' and 'NSW', respectively.

Data Fields with Values (Completeness):

1. **CD**: There are totally six columns which have missing values. Job_title and job_industry_category have highest missing values percentage with 12.65% and 16.4%.
   Mitigation: Ask the clients for the missing values, if not accessible, the missing values can be replaced by median values or most frequent values.

2. **T**: There are seven columns having missing values and only online_order is over 1%.
   Mitigation: Use strategy similar to CD.

Values Free from Contradiction (Consistency):

1. The values of Customer_id from all three datasets are not consistent. Customer_id (T) has only 3494 unique values, from 1 to 5034. The id number is not consistent. The other two datasets have 4000 unique customer_id.

Mitigation: Ask for the clients getting consistent Customer_id(T). If the consistent datasets are not available, we need to consider selecting the Customer_id occurring in all three datasets when joining the tables.

2. The datatypes of the sameattribute are not consistent.
   Mitigation: Convert the records to same datatype.

Values up to Date (Currency):

1. The transaction_date (T) contain values of whole year from '2017-01-01' to '2017-12-30'. What we need is transaction data in the past three months.
   Mitigation: Communicate with the clients to make sure the datasets are the latest and if it is true, we can select the data from '2017-09-30' to '2017-12-30'.

Data items with Value Meta-data (Relevancy)

1. First_name (CD), last_name (DC) are just names, they cannot show any correlations with other factors.
   Mitigation: Drop them when modelling.

2. Job_title (CD) has too many unique values and it is difficult to categorize.
   Mitigation: Feature engineering by simplifying the unique values or just dropping.

3. Address (CA) has too many unique values and almost every customer has one distinct address, which is hard to reflect any correlations with other features.
   Mitigation: Drop the features when modelling.

Data Containing Allowable Values (Validity)

1. **Default (CD)**: the default values are messy code.
   Mitigation: Ask the clients to check the reasonable values of default.

2. **Product_first_sold_date (T) :** the values of the feature are confused and cannot reflect any reasonable meanings.
   Mitigation: Go back to clients for retrieving the reasonable data.

Thanks again for offering the Framework table and hopefully the summary of data quality assessment is helpful. If you have any questions, please feel free to contact me.

Best wishes,

Zhen