Dear Client of Greetings Sprocket Central Pty Ltd,

Thank you for providing us with your datasets. My name is Manuel Martinez and I'm in charged of analyzing your data.

We had reviewed the quality of the 3 datasets:

- Customer Demographic
- Customer Addresses
- Transaction data in the past three months

General Summary:

Table Name	No. of records	Unique Customer Ids
Customer demographic	4000	4000
Customer address	3999	3999
Transactions	20000	3494

Customer Demographic Summary:

- The overview of the data is:

Dataset statistics		Variable types	
Number of variables	13	**Categorical**	4
Number of observations	4000	**Numerical**	3
Missing cells	1763	**Boolean**	2
Missing cells (%)	3.4%	**Date**	1
Duplicate rows	0	**Not Categorical – String**	2
Duplicate rows (%)	0.0%	**Unsuppoted**	1
Total size in memory	406.4 KiB		

- Data analysis:

Standard Data Quality Dimensions	Customer Demographic
Correct Values **(Accuracy)**	• DOB: There is a active customer with 177 years old, his customer_id is 34. It could be a typing error.
Data Field with Values **(Completeness)**	• There are some features with missing values: o last_name -> 125 o DOB -> 87 o job_title -> 506 o job_industry_category -> 656 o default -> 302 o tenure -> 87
Values Free from Contradiction **(Consistency)**	• There is inconsistency in the data: o Gender: [Female, Male, U, Femal, M]
Values up to Date **(Currency)**	• They are updated

Data Items with Values Meta-data **(Relevancy)**	• There is a column not relevant: ○ Default: doesn't have useful information
Data Containing Allowable Values **(Validity)**	• They are validated
Records that are Duplicated **(Uniqueness)**	• There are not duplicated rows

Customer Addresses:

- The overview of the data is:

Dataset statistics		Variable types	
Number of variables	6	**Categorical**	2
Number of observations	3999	**Numerical**	3
Missing cells	0	**Not Categorical – String**	1
Missing cells (%)	0.0%		
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	187.6 KiB		

- Data analysis:

Standard Data Quality Dimensions	Customer Addresses
Correct Values **(Accuracy)**	• There is not accuracy: ○ Customer_id: Not in sync
Data Field with Values **(Completeness)**	• There are not missing values
Values Free from Contradiction **(Consistency)**	• There is not consistency: ○ State: ['New South Wales' 'QLD' 'VIC' 'NSW' 'Victoria']
Values up to Date **(Currency)**	• There are updated
Data Items with Values Meta-data **(Relevancy)**	• There are Relevant
Data Containing Allowable Values **(Validity)**	• There are validated
Records that are Duplicated **(Uniqueness)**	• There are not duplicated rows

Transaction data in the past three months:

- The overview of the data is:

Dataset statistics		Variable types	
Number of variables	13	**Categorical**	5
Number of observations	20000	**Numerical**	5
Missing cells	1542	**Date**	2
Missing cells (%)	0.6%	**Boolean**	1
Duplicate rows	0		
Duplicate rows (%)	0.0%		
Total size in memory	2.0 MiB		

- Data analysis:

Standard Data Quality Dimensions	Transaction data in the past three months
Correct Values **(Accuracy)**	• There is not accuracy: ○ Customer_id: Not in sync

Data Field with Values **(Completeness)**	• There are some features with missing values:
	  ○ online_order -> 360
	  ○ brand -> 197
	  ○ product_line -> 197
	  ○ product_class -> 197
	  ○ product_size -> 197
	  ○ standard_cost -> 197
	  ○ product_first_sold_date -> 197

Values Free from Contradiction **(Consistency)**	• There is not consistency: ○ product_first_sold_date: Format
Values up to Date **(Currency)**	• They are updated
Data Items with Values Meta-data **(Relevancy)**	• There is not relevant values: ○ Order_status: Exclude Cancelled
Data Containing Allowable Values **(Validity)**	• There are values not validated: ○ product_id: It has 1378 zeros values (6.9%).
Records that are Duplicated **(Uniqueness)**	• There are not duplicated rows

Conclusions:

- **There are a difference between the customer_id in the three data.**
  - Customer Demographic has 4000 customers
  - Customer Addresses has 3999 customers
  - Transaction data in the past three months has 3494 customers

  So there are two datasets with missing customers_id.

  Mitigation: ensure that all tables are from the same period.

Solution: The datasets are not in sync with each others so only customers in the Customer Demographic (Master) will be used as training data for our model. If we don't do this it could skew the analysis results.

- **In many features there are many missing values as last_name, DOB, job_title, tenure, job_industry_category, etc.**
  Mitigation: if the features have more than 30% of missing values, that feature is excluded. If they only have a small number of rows empty, filter out the records from the dataset but if the features are important, we imputed based on the distribution of the dataset.
  Solution: in two datasets, we are going to impute the missing values, except on the features id.

- **Some features are inconsistent values for the same attribute. As the feature gender and state.**
  Mitigation: use regular expression to replaced extended values into abbreviations to ensure consistency across addresses and try to enforce a drop-down list for the user entering the data rather than a free text field.
  Solution: the datasets have been cleaned to avoid multiple representations of the same value. In State New South Wales 'NSW' and Victoria for 'VIC', in Gender Femal for 'Female', M for 'Male' and additionaly, U was replaced based on the distribution

- **Some features are not validated as product_id.**
  Mitigation: every dataset has to be a file with a explication of each feature.
  Solution: ask the client if the product id 0 exists. If exists we don't do anything, if not exists, we have to impute or eliminated these records.

- **There is a record that is no accuracy. DOB has a active customer with 177 years old, his customer_id is 34.**
  Mitigation: It could be a typing error. So as soon as the user enters a date that is older than a set age, a warning appears.
  Solution: This record has to be eliminated.

- **There is a feature and some records are no relevant. As the feature default and the order status cancelled.**
  Mitigation: Create just features with sense
  Solution: The feature default has not sense and records with a status of cancelled are eliminated for the analysis.

The team is going to continue with the data cleaning, standardisation and transformation process for the purpose of model analysis. All the assumptions are going to be documented. As soon as we finish this. We would like to have a metting with your team. That way, we can ensure that all assumptions are aligned with Sprocket Central's understanding.

Manuel Martínez.