Dear [the client],

I’ve got your raw datasets which including 4 sheets CustomerDemographic, CustomerAddress, NewCustomerList, and Transactions. Below is the data quality issues we preliminary identified following by some comments on how these issues impact our analysis going forward, how should we fix them and also advise on mitigating data quality issues in the future.

1. **Accuracy**

CustomerDemographic:  
- Customer\_id: 3, 10, 22, 23 don’t have any address

Transactions:  
- customer\_id: id number ‘5034’ doesn't exist in CustomerDemographic

* Impact: If data is not accurate, any derived insights or decisions based on the data can be misleading
* How to fix: Fill up all the missing information (addresses of customers have id 3, 10, 22, 23; demographic info of customer 5034)
* Advise: Only fill up information of other sheets if we have information in CustomerDemographic. Otherwise, back to sign up for a new customer\_id with their details first, then use customer\_id to create a new row in other sheets.

1. **Completeness (missing)**

**CustomerDemographic**:

- last name: 125 null values

- DOB: 87 null values

- Job\_title: 506 Null values

- job\_industry\_category: 656 nulls

**NewCustomerList**:

- Last\_name: 29 null values

- DOB: 19 null values

- job\_title: 106 null values

- job\_industry\_category: n/a

**Transactions**:

- online\_order: 360 null values.

- Brand, product\_line, product\_class, product\_size, product\_first\_sold\_date: 197 null values

* Impact: Incomplete data can lead to partial or biased analysis. It can cause underestimation or overestimation of certain metrics, leading to skewed results.
* How to fix: Double check any other data sources and see if we still have any these missing info somewhere else to re-fill the datasets.
  + DOB: use mode values to replace the null values.
  + Last\_name: use ‘-‘ to replace the null values
  + Job title/category: if there’s one valid values in these two columns, use that to find the other one as well based on the knowledge about the job industry. If there’s all null, use “Other”.
  + Online\_order: Check if we have any other kind of order rather than Online. Otherwise, use ‘Other’ to replace null values.
  + Brand, product\_line, product\_class, product\_size, product\_first\_sold\_date: random check the transaction\_id in other system (delivery, CRM, POS) to find out the reason. It might happen cause of deleted old products in another dataset.
* Advise: Let the entry data admins know the problems we got. Request strict requirements on the columns and values we expect in the data system.

**3. Consistensy (format)**

**CustomerDemographic**:

- Gender: 6 values represent for Female, Male and Unisex

**CustomerAddress**:

- State: should have formatted values as 3 character, for example: VIC/NSW

**NewCustomerList:**

- gender: has a value named "U” when the other two are Female and Male

- Data types in postcode, property\_valuation are mixed of text, decimal and whole number

**Transactions**:

- product\_first\_sold\_date: nonsense number because of wrong format

* Impact Inconsistent data can lead to confusion and difficulty to aggregate or compare them accurately.
* How to fix: Quickly replace above values by correct values following the formatted values of the company. For example: F/M/U, VIC/NSW/QLD.
  + product\_first\_sold\_date: change to Date type.
  + postcode, property\_valuation: change to whole number type
* Advise: Change values in any order system we have to formatted values we rule. Let entry data staff know about the issues to avoid using wrong formats.

**4. Relevancy**

Customer\_Demographic:

- defaut columns has irrelevant values

* Impact: It can make the data processing slower and might introduce noise into the results, causing confusion or misinterpretation.
* How to fix: Drop the column.
* Advise: Avoid creating irrelevant columns to reducing the efficiency of the data system if the amount of irrelevant data is too large

**5. Validity**

**CustomerDemographic:**

- DOB: 21/12/1843

* Impact: It can corrupt the analysis, make the values distribution skewed.
* How to fix: Find the correct year of birth of this customer and change the information. Otherwise, use the mode value to replace.
* Advise: On the order system, make sure we change code to only accept the age not over 100 or 120 years old depends on the knowledge of the target audience.

Please have a look into the issues and comments above. If all the suggestions are matched we can proceed with further analysis of the data to find some suitable insights for the company.

Regards,

Vi Nguyen

Dear [Client point-of-contact],

Thank you for providing us with the three datasets from Sprocket Central Pty Ltd. 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 IDs | Date Data Received |
| Customer Demographic | *-insert value-* | *-insert value-* | *-insert value-* |
| Customer Address | *-insert value-* | *-insert value-* | *-insert value-* |
| Transaction Data | *-insert value-* | *-insert value-* | *-insert value-* |

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 reoccurrence of data quality issues and improve the accuracy of the underlying data used to drive business decisions.

* **Additional customer\_ids in the ‘Transactions table’ and ‘Customer Address table’ but not in ‘Customer Master (Customer Demographic)’**  
  *Mitigation: Please ensure that all tables are from the same period. Only customers in the Customer Master list will be used as a training set for our model.*  
  This indicates that the data received may not be in sync with each other which may skew the analysis results if there are missing data records. Please refer to excel file ‘data\_outliers.xlsx’ for the list of outliers between tables.
* **Various columns, such as the brand of a purchase, or job title, have empty values in certain records**  
  *Mitigation: If only a small number of rows are empty, filter out the record entirely from the training set for prediction. Else, if it is a core field, impute based on distribution in the training dataset.*  
  For key datasets, such as transactions, less than 1% of transactions (totalling less than 0.1% of revenue) have missing fields. These records have been removed from the training dataset.
* **Inconsistent values for the same attribute (e.g. Victoria being represented as “V”, “Vic” and “Victoria”)**  
  *Mitigation: Use regular expression to replaced extended values into abbreviations to ensure consistency across addresses.*  
  Recommendation: Enforce a drop-down list for the user entering the data rather than a free text field. In order to construct meaningful variables for the model, the data has been cleaned to avoid multiple representations of the same value. Additionally, gender records where ‘U’ have been replaced based on the distribution from the training dataset.
* **Inconsistent data type for the sameattribute (e.g. numeric values for some fields and strings for others)**   
  *Mitigation: Convert selected records in characters to numeric. Remove non-numeric characters from string. Recommendation: Ensure that fact tables in the given database have constraints on data types.*   
  Having different data types for a given field make it difficult to interpret results at the later stage. Therefore, appropriate data transformations are made to ensure consistent data types for a given field. Table name No. of records Distinct Customer IDs Date Data Received Customer Demographic -insert value- -insert value- -insert valueCustomer Address -insert value- -insert value- -insert valueTransaction Data -insert value- -insert value- -insert valueNote: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

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,   
[Junior Consultant Name]