**Task 1:**

**Dear John Doe**,

We appreciate your contribution of the three datasets from Sprocket Central Pty Ltd. The summary table provided below outlines the main quality concerns identified within these datasets. If you have any questions or require further clarification regarding the highlighted issues, please let us know.

**Summary Table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Accuracy** | **Completeness** | **Consistency** | **Currency** | **Relevancy** | **Validity** |
| Customer Demographics | **1.DOB**:  Inaccurate | **1.Job title**:  Blanks  **2.Customer ID**: Incomplete | **1. Gender**:  Inconsistency | **1.Deceased Customers**:  Filter out | 1.**Default Column**:  Deleted |  |
| Customer Address |  | **1.Customer ID**: Incomplete | **1.States**:  Inconsistency |  |  |  |
| Transactions | **1.Profit**:  Missing | **1.Customer ID**: Incomplete  **2.Online** Order: Blank  **3.Brand**: Blank |  |  | **1.Cancelled status order**: Filter out | **1.List Price**: Format  **2.Product sold Date**: Format |

Below are more details about all quality issues that is identified within these three datasets and method of mitigation used. Recommendation and description have given to avoid further data quality concerns in future. Following recommendation will improve accuracy of data` used to influence business decisions.

# **Accuracy Issues:**

## **DOB was inaccurate for “Customer Demographic” and missing a profit column in “Transactions”.**

**Mitigation:** Filter out outliers in “DOB”.

**Recommendation:** Create an age column allowing for more comprehensive data and easier to check for errors. Create a profit column to check accuracy of column.

# **Completeness:**

## **Customer ID was inconsistence in “Customer Demographic”, “Customers Address” and “Transactions”.**

**Mitigation:** Filter out all customer id

**Recommendation:** Ensure table are up to date. For our model customer id from 1 to 3500 will be used as they have complete data.

**Solution:**

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]

**Note:** 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.