

Dear [Name of the Client],

We have received the three raw datasets from SP rocket central Pty Limited. As per the preliminary task, in the below-mentioned list, we have analysed the quality of the raw data and we found multiple quality issues that need to be addressed. Also, we have suggested recommendations to mitigate the quality issues and improve the effectiveness of the data.

High-Level information four dataset provided by your side to us.

Worksheets Names	Records (Rows * Columns)	No. Distinct Customer IDs	Date of Data Received
Customer Demographic	4000 * 13	4000	2017-12-30
Customer Address	3999 * 6	3999	2017-12-30
Transaction Data	20000 * 13	3494	2017-12-30
New Customer List	1000 * 23	Not Applicable	2017-12-30

After Data Quality Assessment, key-findings are as below:

**1. Redundant Outliers.**

**Issue:** Some of the data values are outliers and can disrupt the whole dataset. For example, The customer ID “34” with the name of Jephthah Bachmann was born in 1843, meaning that he is 175 years old which is an error in the data in the Customer Demographic Table.

**Recommendation:** Remove the redundant data as it may skew the distribution of the dataset.

**2. Missing Values.**

**Issue:** Multiple attributes like “Online Order”, “Brand Name”, “Product Line”, “Product Class”, “ Product Size”, “Standard Cost”, and “product\_first\_sold\_date” in the Transactions table had blank values. Also, In the Customer Demographic “Job Title”, “Job Category” and “Tenure” some of the records are missing.

**Recommendation:** As the percentage of missing values in the datasets is low as compared to the whole dataset we can proceed by removing them.

**3. Inconsistent Entries across the datasets.**

**Issue:** There are an additional number of entries in customer\_ids in the Transactions table than Customer Demographic and Customer Address Table. Hence, the skewed data cannot be used if there are any missing records.

**Recommendation:** We will only perform the analysis on the synced data of all the three customer tables across the customer\_ID.

#### **4. Multiple DataTypes for a Single Column.**

**Issue:** For the attribute “Standard Cost” in the Transaction table there are some records with special string characters which causes inconsistency in the dataset.

**Recommendation:** Remove the special characters from the records and convert all the characters into numeric data to ensure consistent data types.

#### **5. Duplicate values for the same column.**

**Issue:** In the “State” Column of the Customer Address Table multiple duplicate values were found such as “VIC” & Victoria, “NSW” & “New South Wales”. Also, the issue is in the “Gender” column of the Customer Demographic Dataset.

**Recommendation:** To use abbreviations of the states instead of full names for all the records to ensure consistency across addresses. For Gender Column, the records “U” can be imputed with the distribution of the dataset.

Please look into the above-mentioned quality issues along with the recommended changes to ensure the consistent quality of the dataset across all the tables. If all the suggestions are matched we can proceed with further analysis of the data to find some suitable insights for the company.

Regards,  
Parvej alam Ansari.