DATA QUALITY ASSESSMENT

Dear [client - name],

I have looked through each of the data sets provided and run some basic checks on them. I have summarized my results with regards to the problems present in the data, along with possible solutions (in green). The summary statistics for each table is given below. Please let us know if any discrepancies arise with respect to the same.

TABLE NAME	NO. OF RECORDS	DISTINCT CUSTOMER IDs	DATE DATA RECIEVED
Customer demographic	4000	4000	-
Customer Address	3999	3999	-
Transactions	20000	3494	-

ISSUES IN DATA QUALITTY

1. **No associated meta data or data description**
   For instance, there is no information about the units of prices or any description of what each column means leading to issues in Relevancy.

   The solution for this needs to come from the client end as only data collectors are aware of correct data descriptions.

2. **Missing Values**

   - Around 3% of the transactions data and 14% of the customer demographic data contains at least one missing value/column entry. This is quite significant given the sample sizes. It is reasonable to assume this does impact analysis and predictions quite significantly.

   - Majority of missing values in transaction data are regarding the brand and details like size , class, std cost etc. and also on whether the order was online or not. This could be significant while trying to analyse online vs offline purchase trends.

   - Last names, job titles and categories are predominantly missing in demographic data. This will hinder customer segmentation and the understanding purchase preferences of each demographic.

   Missing values can be imputed if and when possible, else can be ignored if insignificant enough.

3. **Inconsistency across data sets**

Some data IDs present in transactions and customer address data sets are not present in the customer demographic database. This leads to inaccuracies while analysing data.

It is imperative to ensure that all the data sets are in sync and have a uniform customer base. Otherwise any analysis will be inapplicable. Only the customer IDs in the master demographic data set will be used for drawing conclusions.

4. **Inconsistent encoding of similar attributes**

For instance – "NSW" and "New South Wales" are used to represent the same state. Similarly gender is "F" and "Female". This is problematic while analysing categorical data. A standard convention should be adopted for representing categories.

The categories can be re-encoded to enforce consistency. It is better to ensure that, in the future, constraints are imposed during data collection itself – for instance provide radio button selection of gender instead of asking customers to enter text.

5. **Values not up to date**

From the earliest and latest date values in the transaction data we can conclude that the values are 3 years old - not up to date – currency issue.

To obtain accurate and reliable findings it is essential to have recent data. If possible, the client is requested to provide us with the same.

6. **Other notable discrepancies**

- One DOB is 1843 - which is impossible and inaccurate.

- Further, there appear to be around 82 entries having DOBs in the years 2002,2001 and 2000 having rather descriptive and senior job titles – these people only 15-16 years of age during the time of transactions(which are all in 2017), which seems implausible since they cannot logically hold senior/managerial titles at such ages. This brings into question the accuracy of data provided.

- Many people of this age group also appear to own homes and cars - again dubious. This could lead to severe inaccuracies in understanding the customer demographic, especially given the large number of missing values.

There is no solution to this besides advocating for better data collection practices. The onus is not only on the client – sometimes customers provide incorrect data regardless of the practices followed.