Dear Manager,

Thank you for the provided datasets in transactions, customer demographic and customer address. After analysing the data quality in these datasets, I have found serval data quality issues.

|  |  |  |
| --- | --- | --- |
| Datasets | Customer ID records | Records |
| Transactions Dataset | 3496 | 20000 |
| Customer Demographic Dataset | 4000 | 4000 |
| Customer Address | 3999 | 3999 |

**Accuracy problem**

* Incorrect value in Customer Demographic dataset where customer was born in the year of 1843
* Missing profit column in Transaction dataset

**Data Mitigation:**

Filter out the unwanted year of 1843 in DOB and create an additional column called profits, allowing for monetary analysis.

**Completeness problem**

* Blank job title and job industry category based on the existing job title in Customer Demographic dataset
* 198 records have the product number in 0 and blank in product details and cost
* The additional customer\_ids were found in Transactions dataset and Customer Address dataset are not in Customer Demographic dataset

**Data Mitigation:**

1. Filter all blanks for standard\_cost, product\_size, product\_class, product\_line , job\_title, job\_industry, online\_order and brand\_column.
2. Please ensure the data are in sync and from the same period. Incomplete data may distort analysis results.

**Consistency problem**

* Data inconsistency in gender in Customer Demographic and Customer Address dataset
* Data inconsistency in state in Customer Address dataset

**Data Mitigation:**

Please use a standardise name to replace the mislabelled categories or classes (e.g. miss-spelling female in femal and f). Create a drop-down function for the customer to input information that ensures data consistency.

**Currency problem**

* The ‘Y’ in Deceased indicator indicate that customer isn’t a current customer in Customer Demographic dataset.

**Data Mitigation:**

Filter all the yes in Deceased indicator that ensures the data currency.

**Relevancy problem**

* Default column contains incorrect and irrelevant values in Customer Demographic dataset
* The cancelled order status in Transaction dataset is irrelevant to the analysis

**Data Mitigation:**

Remove the Default column and filter the cancelled order status to ensure all relevant data are available for analysis.

**Validity problem**

* Product first sold date in Transaction dataset didn’t specific the date in date data type
* The standard cost uses currency data type, whereas list price doesn’t have specific data type in Transaction dataset.
* Some numeric and string values uses Inconsistent data type

**Data Mitigation:**

Covert the Product first sold date into date data type, list price into currency data type, non-string into numeric and non-numeric to character.

The given data quality issue has been discovered at the first stage of data quality analysis. The mitigation has adopted an efficient and effective method to improve the data quality for the following analysis stage. Please have a look at the following identified data quality issues and let us know if you have any problems.

Yours sincerely,

Yu Hing Tang