Subject: Data Quality Assessment and Recommendations

Hi there,

I hope this message finds you well. We appreciate the opportunity to work with your valuable data and want to ensure its quality for meaningful analysis. After a thorough assessment of the datasets you have provided, we have identified several data quality issues in the Customer Demographics, Customer Address, and Transactions datasets, which may impact our analysis going forward. We have also offered suggestions for rectifying these issues and cleaning the data.

Here's an overview of the issues:

Customer demographics dataset:

- **Accuracy**: A single customer has an inaccurate date of birth.
- **Completeness**: Approximately 88 customer records are missing date of birth, 125 customer records are missing last names, 506 customer records are missing job titles, and 88 customer records are missing tenure information.
- **Consistency**: Inconsistent gender values (e.g., "Female", "F", "Femal", "M", "Male"). There are mismatches between Customer_id in the Demographics dataset and the Address dataset. For example, Customer_id "3," "10," "22," and "23" do not exist in the Address dataset, while Customer_id "4001," "4002," and "4003" are in the Address dataset but absent from the Demographic dataset.
- **Validity**: The "default" field is invalid as it fails to convey the data type contained within. It contains meaningless symbols that are not readable. In addition, there were 88 instances of Invalid gender value (“U”).

Customer Address dataset:

- **Consistency**: Inconsistent State values ("NSW," "New South Wales," "VIC," "Victoria").

Transactions dataset:

- **Completeness**: Approximately 360 customer records lack Online_Order values, and 197 customer records are missing the Brand, product_line, product_class, product_size, standard_cost and product_first_sold_date information.

Recommendations To Effectively Address And Mitigate These Data Quality Issues

To ensure the integrity of the data and the success of our analysis, we recommend addressing these issues in the following ways.

Customer Demographics Dataset:

- **Accuracy:** Review the customer with an inaccurate date of birth and correct the entry.
- **Completeness:** For missing date of birth, last names, job titles, and tenure information, contact customers to complete their profiles or use data imputation techniques like mean/median Imputation for numeric values (i.e. replacing missing data with the mean or median of the available data in the same column), mode imputation for text or categorical data (i.e. filling the missing values with the most frequently occurring value) or K-Nearest Neighbours

imputation which involves finding the K nearest neighbours with complete data and using their values to fill in missing data.

- **Consistency:** Standardise gender values by mapping inconsistent entries to a common format (e.g., "Male" and "Female").
- **Inconsistency (Join Keys):** Investigate the mismatches between Customer_id in the Demographics and Address datasets. Correct data inconsistencies by updating or removing records as needed.
- **Validity:** Review and change the "default" field to accurately convey the data contained in the field. But given that there are lots of meaningless symbols in that field which will lead to a noisy data doing analysis its best the column is removed. Address invalid gender values ("U") by correcting or contacting customers for accurate information.

Customer Address Dataset:

- **Consistency:** Standardize “State” values to a consistent format, such as using abbreviations (e.g., "NSW" and "VIC"). Correct any inconsistent entries.

Transactions Dataset:

- **Completeness:** For missing Online_Order values, consider filling the missing values with the most frequently occurring value of the same customer (i.e. considering fields with the same customer_id).
    - For missing Brand, product_line, product_class, product_size, standard_cost, and product_first_sold_date information, fill in gaps using product databases or perform data imputation.

General Recommendations:

- **Data Validation:** Implement data validation rules and checks such as mandatory fields, data type validation, range checks, format check, consistency checks, use list or dropdown validation, and pattern matching at the point of data entry to ensure data accuracy and completeness.
- **Data Enrichment:** Consider utilising data enrichment services (i.e. third-party solutions that enhance existing datasets by adding additional information or attributes) or external data sources to fill in missing data and maintain data accuracy.
- **Data Standardization:** Standardize data formats and representations (e.g., gender values and state names) to ensure consistency.
- **Data Deduplication:** Identify and merge duplicate customer records within each dataset to maintain data uniqueness.
- **Data Retention Policy:** Implement a data retention policy to archive or remove outdated or irrelevant records to reduce clutter.

We are committed to partnering with you to resolve these data quality concerns and are eager to proceed with your guidance on the next steps. Should you have any inquiries or wish to discuss this further, please feel free to reach out.

Kind regards,
Mary Ogbuka Kenneth
Data Analyst