Klynveld Peat Marwick Goerdelerhorizontal line

Level 38 Tower 3/300

Barangaroo Ave,

Sydney NSW 2000, Australia

February 23, 2023

Dear Client,

I hope this email finds you well. As per our recent discussion, I have reviewed the datasets you provided to us: Customer Demographic, Customer Addresses, and Transaction data in the past three months. I have also conducted an analysis to assess the quality of the data, identify data quality issues, and recommend strategies to mitigate these issues.

Before I delve into the findings of my analysis, I want to briefly explain the data quality framework our team uses to evaluate a dataset. We use the following dimensions to assess data quality: Accuracy, Completeness, Consistency, Currency, Relevancy, Validity, Uniqueness. I have attached a Data Quality Framework Table for your reference.

Now, let me share with you my findings for each dataset:

**Customer Demographics.**

Findings

1. Completeness

There are missing values in 'last\_name', 'DOB', 'job\_title', 'job\_industry\_category', 'default' and 'tenure' columns.

The missing value counts are as follows: 125 in 'last\_name', 87 in 'DOB', 506 in 'job\_title', 656 in 'job\_industry\_category', 302 in 'default', and 87 in 'tenure'.

1. Accuracy

The 'gender' column which is supposed to be categorical with choices between male and female has 6 unique values namely: F, Male, Female, U, Femal , M and blanks.

In the 'DOB' column, there is one customer with a year of birth listed as '1843', which is likely an error.

1. Uselessness

The 'default' column seems to be a useless chunk of data and has no relation to the goal set out to be met.

Based on these findings, I recommend the following

1. Completeness

The missing values in the 'last\_name', 'DOB', 'job\_title', 'job\_industry\_category', 'default', and 'tenure' columns should be investigated and filled in wherever possible, especially in cases where the missing values might affect the analysis or conclusions that can be drawn from the data.

1. Accuracy

The 'gender' column should be cleaned and standardized to have only two unique values, male and female, and any incorrect or invalid values should be removed or corrected.The year of birth listed as '1843' in the 'DOB' column should be investigated and corrected if it is indeed an error.

1. Uselessness

The 'default' column seems to have no relevance to the goals of the analysis and can be safely removed from the dataset to improve data quality and simplify the dataset.

**Customer Address**

Findings

1. Completeness:

There are 3999 non-null values in the address dataset, while there are 4000 unique customers. Customer with customer ID '3' doesn't have data in the address dataset.

1. Consistency:

The 'state' column has 5 unique values: 'New South Wales', 'QLD', 'VIC', 'NSW', and 'Victoria'. However, 'New South Wales' and 'NSW' are the same place, as are 'Victoria' and 'VIC'.

1. Uniqueness

There are 3 sets of duplicates in the CustomerAddress table, with each set having 2 rows. The duplicates are identified based on the values in the address, postcode, state, country, and property\_valuation columns. The sets are as follows:

Rows with customer\_id 732 and 2470, both with address '3 Talisman Place', postcode 4811/4017, state QLD, country Australia, and property\_valuation of 2/5 respectively.

Rows with customer\_id 2315 and 3535, both with address '64 Macpherson Junction', postcode 2208/4061, state NSW/QLD, country Australia, and property\_valuation of 11/8 respectively.

Rows with 'customer\_id' 2333 and 2985, both with address '3 Mariners Cove Terrace', postcode 3108/2216, State 'VIC' and 'NSW' and property\_valuation of 10/10.

Based on these findings, I recommend the following

1. Completeness

For the missing data in the address dataset for customer ID '3', try to gather the missing information or confirm if the customer does not have any address data.

1. Consistency

Standardize the values in the 'state' column to 'New South Wales', 'Queensland', and 'Victoria' to avoid inconsistencies.

1. Uniqueness

Based on the identified duplicates, it is recommended to further investigate and verify whether they represent actual duplicate customer records or unique customers with similar addresses. The investigation could involve reviewing additional customer information, such as names and contact details, and verifying them against the identified duplicate records. Once the investigation is complete, appropriate actions can be taken, such as merging the duplicate records or keeping them as separate records if they represent unique customers. Additionally, it is recommended to implement measures to prevent or minimize the occurrence of duplicate records in the future, such as using unique customer identifiers and performing regular data quality checks.

**Transaction Data**

Findings

1. Completeness:

There are missing values in the 'online\_order', 'brand', 'product\_line', 'product\_class', 'product\_size', 'standard\_cost', 'product\_first\_sold\_date' columns.

1. Completeness:

Explicitly there are 197 rows that have no data on 'brand', 'product\_line', 'product\_class', 'product\_size', 'standard\_cost', 'product\_first\_sold\_date'. These are the missing values in those columns.There are 360 missing values in the 'online\_order' column.

1. Accuracy:

The 'product\_first\_sold\_date' column doesn't have data that represents what its column name describes.

1. Consistency:

The 'Transaction\_date' column is in the format mm/dd/yyyy and needs to be standardized to sync with that of 'DOB' column in both the CustomerDemographics and NewCustomerList tables which is in the format yyyy-mm-dd.

Based on these findings, I recommend the following

1. Completeness

We recommend that the missing values in the 'online\_order', 'brand', 'product\_line', 'product\_class', 'product\_size', 'standard\_cost', and 'product\_first\_sold\_date' columns be investigated and filled in where possible. If it is not possible to retrieve the missing data, we recommend that these rows be removed from the dataset.

1. Completeness

We recommend that the root cause of the missing values in the 197 rows that have no data on 'brand', 'product\_line', 'product\_class', 'product\_size', 'standard\_cost', and 'product\_first\_sold\_date' columns be identified and addressed.

1. Completeness

We recommend that the missing values in the 'online\_order' column be investigated and filled in where possible. If it is not possible to retrieve the missing data, we recommend that these rows be removed from the dataset.

1. Accuracy

We recommend that the values in the 'product\_first\_sold\_date' column be reviewed to ensure that they accurately represent the first sale date for each product. If it is determined that the values are inaccurate, we recommend that they be updated accordingly.

1. Consistency

We recommend that the 'Transaction\_date' column be standardized to match the format of the 'DOB' column in both the CustomerDemographics and NewCustomerList tables, which is in the format yyyy-mm-dd. This will help ensure consistency across all tables and make it easier to perform analyses and comparisons.

**New Customer List**

Findings

1. Completeness

There are missing values in 'last\_name', 'DOB', 'job\_title', and 'job\_industry\_category' columns. The missing value counts are as follows: 29 in 'last\_name', 17 in 'DOB', 106 in 'job\_title', and 165 in 'job\_industry\_category'.

1. Accuracy

The 'property\_valuation', 'past\_3\_years\_bike\_related\_purchases', and 'postcode' columns have some data stored as text instead of numbers (int or float).

1. Consistency

The 'State' column needs to be standardized to sync with the data from the CustomerDemographics table. Hence values 'QLD, 'NSW' and 'VIC' must be changed to full names of the states.

Based on these findings, I recommend the following

1. Completeness

Missing values in 'last\_name', 'DOB', 'job\_title', and 'job\_industry\_category' columns should be handled appropriately. Depending on the percentage of missing data, it may be necessary to drop rows or impute missing values using techniques such as mean, median, or mode imputation.

1. Accuracy

The 'property\_valuation', 'past\_3\_years\_bike\_related\_purchases', and 'postcode' columns should be checked for inconsistencies and corrected. Data stored as text instead of numbers should be converted to their appropriate data types.

1. Consistency

The 'State' column should be standardized to match the values in the CustomerDemographics table. This can be achieved by replacing 'QLD' with 'Queensland', 'NSW' with 'New South Wales', and 'VIC' with 'Victoria'.

In addition, we recommend performing a join between the datasets to ensure that the data is complete and accurate. Specifically, we recommend joining the Customer Demographic and Customer Address datasets on the 'customer\_id' column to obtain the complete customer information.

Please let us know if you have any questions or concerns about these recommendations. We look forward to continuing to work with you on this project.

Best regards,

Osikanyi Essandoh

Data Analytics Intern

KPMG. Australia