From: **Quim Bolós Fernández** bolosfernandez@hotmail.es
Subject: Data Quality Assessment and Strategies for Mitigation
Date: 12 July 2023 at 14:27
To:

Dear Manager,

I hope this email finds you well.

I have conducted a thorough assessment of the datasets provided, namely the Customer Demographic, Customer Address, New Customer List, and Transactions, and I would like to highlight the data quality issues identified for each dataset. Additionally, I will propose strategies to mitigate these issues and improve the overall data quality.

1. Customer Demographic

Accuracy Correct Values	One value in the Date of Birth (DOB) column is incorrect, showing the year as 1843.
Completeness Data Fields with Values	Several data fields have missing values, including last_name, DOB, job_title, job_industry_category, default, and tenure.
Consistency Values Free From Contradiction	The Gender column contains various inputs such as 'M', 'Femal', 'F', and 'U,' which should be translated into Male or Female for semantic consistency. There are inconsistencies between the job_title and job_industry_category columns, where certain job titles are incorrectly assigned to different job industry categories. Overall, n/a should be replaced by blanks or the other way around for semantic consistency.
Currency Values Up to Date	Columns such as deceased_indicator and owns_car may not be up to date.
Relevancy Data Items with Value Meta-Data	The deceased_indicator column includes some 'Yes' values, which may not be useful since they are very few in number. The job_title column includes various jobs with different levels, potentially segmenting the dataset unnecessarily. The past_3_years_bike_related_purchases and wealth_segment columns may disagree if they are reliant on each other since all wealth_segments have the same range of values for past_3_years_bike_related_purchases.
Validity Data Containing Allowable Values	The default column does not contain any allowable values.
Uniqueness Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different customer_id.

2. Customer Address

Completeness Data Fields with Values	Customers with IDs [3, 10, 22, 23] do not have address data.
Currency Values Up to Date	The address data may not be up to date.
Relevancy Data Items with Value Meta-Data	All customers in this dataset live in Australia, making the Country column irrelevant.
Uniqueness Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different customer_id.

3. New Customer List

Completeness Data Fields with Values	Several data fields have missing values, including last_name, DOB, job_title, and job_industry_category.
Consistency Values Free From Contradiction	The Gender column contains ‘U' values, which could be interpreted as 'Unknown' for semantic consistency. There are inconsistencies between the job_title and job_industry_category columns, where certain job titles are incorrectly assigned to different job industry categories.

	certain job titles are incorrectly assigned to different job industry categories. Overall, n/a should be replaced by blanks or the other way around for semantic consistency.
Currency Values Up to Date	Columns such as deceased_indicator, owns_car and address-related data may not be up to date.
Relevancy Data Items with Value Meta-Data	The deceased_indicator column does not contain any positive values, rendering it irrelevant. The job_title column includes various jobs with different levels, potentially segmenting the dataset unnecessarily. The past_3_years_bike_related_purchases and wealth_segment columns may disagree if they are reliant on each other since all wealth_segments have the same range of values for past_3_years_bike_related_purchases. The Country column contains only 'Australia' values, which may not provide valuable insights. Assess the relevancy of the default column and consider its removal. Evaluate the calculation and relevance of scoring criteria in columns 16 to 19.
Validity Data Containing Allowable Values	The default column does not contain any allowable values.
Uniqueness Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different customer_id.

4. Transactions

Completeness Data Fields with Values	The online_order column has missing data for 19640 out of 20000 entries. The product_id 0 has missing data in multiple columns, including brand, product_line, product_class, product_size, standard_cost, and product_first_sold_date.
Consistency Values Free From Contradiction	The product_first_sold_date values are different for each product_id, even when differentiated by online_order. However, list_price and standard_cost are consistent for each product_id.
Validity Data Containing Allowable Values	The product_first_sold_date values are in the wrong format.
Uniqueness Values that are Duplicated	No duplicated rows were found, and each row corresponds to a different transaction_id.

Mitigation Strategies

To address these data quality issues, I recommend the following strategies:

Overall	Continue maintaining uniqueness within the dataset. Standardise gender inputs to either ‘Male', ‘Female' or ‘Unknown’. Review and rectify inconsistencies between the job_title and job_industry_category columns. Regularly update columns like deceased_indicator, owns_car, and address-related data to ensure currency.
Customer Demographic	Correct the incorrect value in the DOB column. Impute missing values in fields such as last_name, DOB, job_title, job_industry_category, default, and tenure. Update columns like deceased_indicator and owns_car to ensure currency. Assess the relevancy of columns such as deceased_indicator, default, job_title, and wealth_segment, and consider removing unnecessary columns. Validate and clean the data to ensure the default column contains allowable values.
Customer Address	Investigate and retrieve missing address data for customers with IDs [3, 10, 22, 23]. Evaluate the relevancy of the Country column and consider removing it if it does not provide meaningful insights.

New Customer List	Address missing values in fields such as last_name, DOB, job_title, and job_industry_category. Evaluate the relevancy of columns such as deceased_indicator, Country, and default, and remove unnecessary columns. Validate the data in the default column to ensure it contains allowable values.
Transactions	Investigate and rectify missing data in the online_order column. Investigate and rectify missing data in columns related to product_id 0. Assess the consistency of product_first_sold_date values and consider formatting them correctly.

Please let me know if you have any questions or require further clarification regarding these data quality issues or the proposed mitigation strategies. I look forward to your response.

Thank you for your attention to this matter.

Best regards,

Joaquim Bolós Fernández
Intern Data Analyst
KPMG