Subject: Data Quality Issues, Observations, and Recommendations

Hi,

We have recently received three datasets from SP Rocket Central Pty Limited, and I have conducted a thorough data quality check based on our predefined parameters. The purpose of this email is to provide you with an overview of the data quality issues we identified, along with our observations and recommended strategies to address these issues. I kindly request you to review the following points and share your views and opinions on them.

* The "NewCustomerList" dataset lacks a "customer\_id" for each new customer. This is a crucial requirement as the "customer\_id" serves as a unique reference for each customer. Without this information, we won't be able to retrieve their transaction details from the database.
* We noticed incorrect data in the "CustomerDemographic" dataset. Specifically, the date of birth (DOB) for "customer\_id" 34 is recorded as 1843-12-21. It is highly improbable for someone to be around 174 years old and simultaneously classified as "deceased\_indicator: N." For further analysis, we will assume the DOB values are correct.
* The criteria for classifying customers based on their wealth is unclear. We require a more detailed understanding of this feature. For instance, considering "customer\_id" 108, the customer is labeled as "High Net Worth" in the "wealth\_segment" column. However, the customer falls into a valuation bin of 8, whereas the highest class value for "property\_valuation" in the "CustomerAddress" dataset is 12. Additionally, the customer does not own a car according to the "owns\_car" data in the "CustomerDemographic" sheet.
* The "default" column in the "CustomerDemographic" dataset contains irrelevant and encoded characters that do not contribute to solving the desired problem. These characters should be removed as they are not useful for our analysis.
* We need to modify the "gender" column in the "CustomerDemographic" dataset. To align with the format in the "NewCustomerList" dataset, we will replace the classes "M" and "F" with "Male" and "Female," respectively.
* The "deceased\_indicator" column in both the "CustomerDemographic" and "NewCustomerList" datasets contains only one class, "N." Since there is no variation in values within this column, we will exclude it from our analysis.
* Although there are some missing data points, they constitute a negligible percentage compared to the total number of data points. For example, the "DOB" column has 2.18% missing values, "job\_title" has 16.65% missing values, "job\_industry\_category" has 16.40% missing values, and "tenure" has 2.18% missing values. We can handle these missing data points through replacement methods.

In addition to the above observations, we received another report highlighting several quality issues in the raw data. I have summarized them below along with the corresponding recommendations:

* Redundant outliers: Some data values are outliers, such as the customer ID "34" with the name Jephthah Bachmann, whose birth date is listed as 1843. This indicates an error in the Customer Demographic Table. We recommend removing redundant data to prevent skewing the dataset's distribution.
* Missing values: Various attributes in the Transactions table, including "Online Order," "Brand Name," "Product Line," "Product Class," "Product Size," "Standard Cost," and "product\_first\_sold\_date," have blank values. Similarly, the Customer Demographic dataset has missing records for "Job Title," "Job Category," and "Tenure." Given that the percentage of missing values is low compared to the total dataset, we recommend removing these records.
* Inconsistent entries across datasets: The Transactions table contains more customer IDs than the Customer Demographic and Customer Address tables. To ensure accurate analysis, we should only consider data from customer IDs present in all three tables.
* Multiple data types for a single column: The "Standard Cost" attribute in the Transactions table includes records with special characters, causing inconsistencies in the dataset. We suggest removing the special characters and converting all values to numeric data for consistency.
* Duplicate values for the same column: The "State" column in the Customer Address table has duplicate entries, such as "VIC" and "Victoria," as well as "NSW" and "New South Wales." Similarly, the "Gender" column in the Customer Demographic dataset has issues. To address this, we recommend using consistent state abbreviations and imputing the "U" records in the Gender column based on the dataset's distribution.

Please review the mentioned quality issues and recommended changes to ensure consistent data quality across all tables. Once these suggestions are implemented, we can proceed with further analysis to derive valuable insights for the company.

If you have any questions or require additional clarification, please don't hesitate to reach out. Thank you for your attention to this matter.

Best regards,

Mirza Hanan

Virtual Intern @KPMG