Dear Mr. John,

Thank you providing the datasets from your organization. The datasets we received from you were:

1. Transaction data in the past three months
2. Customer Demographic
3. Customer Addresses

Firstly, here is a summary of the number of records in the datasets we received from you (The column titles have not been included in this count). Please note that we made use of Microsoft Excel for this evaluation.

**Dataset Count**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Date Received** | **Total number of records** | **Total number of unique customer IDs** | **Comments** |
| **Transactions Table** | 17/11/2022 | 20,000 | 3,494 | Duplicate records found but this is allowed because it is the transactions table. One customer may have made several purchases. |
| **Customer Demographics Table** | 17/11/2022 | 4,000 | 4,000 | No duplicate records found |
| **Customer Address Table** | 17/11/2022 | 3,999 | 3,999 | No duplicate records found |

Having determined the number of records in the datasets, we examined the 3 datasets to assess the quality of the data. While doing that, we discovered some data quality issues which might impact our analysis. However, we have given recommendations to mitigate those concerns. The quality of the data was measured based on Accuracy, Completeness, Consistency, Currency, Relevancy, Validity, and Uniqueness. They are given in the summary table below:

**Data Quality Summary Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Data Quality Dimensions** | **Transactions Table** | **Customer Demographic Table** | **Customer Address Table** |
| **Accuracy** | **Issue:**   * The list\_price, and standard\_cost columns were in a text format instead of currency while the transaction\_date, and product\_first\_sold\_date columns were in a text form instead of date format.   **Recommendation:**   * Ensure all columns are in their correct formats by clicking on the “Number Format” function of the “Home” tab and selecting the right formats. | **Issue:**   * Incorrect date of birth in the DOB column.   **Recommendation:**   * Valid date of birth should be inputted to avoid encountering outliers during analysis. |  |
| **Completeness** | **Issues**:   * Blank row. * A lot of missing records in the online\_order, brand, product\_line, product\_class, product\_size, standard\_cost, and product\_first\_sold\_date columns. * Profit not calculated in a new column.   **Recommendations:**   * Remove blank rows by applying the filter function, then delete them. This will prevent calculation or applied formula errors. * Apply the filter function on the columns -as stated above, to check for missing values. Then fill up the blank spaces or remove them completely if the fields will not be useful for the analysis. * Create a new column to calculate the profit by subtracting the values in the standard\_cost columns from those of the list\_price. | **Issues:**   * Missing records in the last\_name, DOB, job\_title, owns\_car, and tenure fields. * Blank rows in the entire dataset. * Age not calculated.   **Recommendations:**   * Apply the filter function to check for blank rows and/or records to ensure thoroughness during analysis. * Date of Birth does not help in the analysis phase, so make sure that date of birth is converted to Age in a new column. |  |
| **Consistency** |  | **Issue:**   * Inconsistent fields in the gender, job\_industry\_category columns.   **Recommendation**:   * Utilize complete words to classify gender types instead of the first/single letters. This will be particularly useful in the analysis phase. | **Issue:**   * Disparity in the state names. Some state names are in full, while others are abbreviated.   **Recommendation:**   * To avoid clutter during the analysis phase, state names should be abbreviated. |
| **Currency** |  | **Issue:**   * Contains records in the deceased\_indicator column still includes customers that have passed on.   **Recommendation:**   * Filter out records for customers that are deceased so that analysis can be done on the company’s present customers. |  |
| **Relevancy** |  | **Issue:**   * Irrelevant records in the column named “default”.   **Recommendation:**   * The entire column should be deleted as the records are not useful. |  |
| **Validity** | **Issue:**   * Using the number “0” as product ID.   **Recommendation:**   * Avoid using “0” as a numbering for columns. It can be mistaken for a null or missing value during data clean-up or analysis. It is best to begin numberings with “1”. |  |  |

The recommendations given in the data quality summary table are effective ways in which you can improve the quality of data to avoid wrong analysis or insights which could lead to business setback.

Kindly let us know if you have any enquiries bordering the issues raised.

Best Regards,

Victoria J. Abdulkadir