Name - *Vishnav Tuli*

Task-1

Subject: Data Quality Analysis and Recommendations for Spreadsheets

Hi there,

I wanted to bring your attention to some data quality concerns I have identified while analyzing the spreadsheets. Below, I have outlined the issues we found in each dataset and provided recommendations on how to mitigate them.

1. Transactions (first dataset):

- The "list_price" field has not been assigned any unit of currency. To ensure clarity, it is recommended to specify the currency used.

- Several fields such as "online_order," "product_line," "product_class," "product_size," "standard_cost," and "product_first_sold_date" contain empty values. It is crucial to fill in these missing data points for accurate analysis.

- Multiple cells/sections have been assigned a "product_id" of 0, which seems incorrect. The correct product IDs should be used to ensure data integrity.

- The "product_first_sold_date" is mentioned in the wrong format. I suggest converting it to a short date format for consistency.

- There are empty spaces/cells in the "product_size" column, which should be removed to maintain a clean dataset.

- If "online_order" is marked as false, it implies that the order was not placed online. In such cases, the "order_status" should be updated to "cancelled" or "not approved" to reflect the correct status.

2. CustomerDemographic (second dataset):

- Duplicate values have been found within the dataset. These duplicates should be eliminated to avoid data redundancy.

- Empty spaces are present in the "last_name," "DOB," "job_title," and "tenure" fields. It is essential to remove these empty values for consistency.

- The "job_industry_category" contains null values (N/A). It is recommended to address these null values by either filling them or assigning a specific category, as appropriate.

- In the "gender" column, there are inconsistencies such as using "F" instead of "female" and abbreviations like "U," "Femal," and "M." We suggest replacing these entries with the appropriate genders to ensure uniformity.

- The "DOB" value for entry F-36 is mentioned as 1843-12-21, which appears to be invalid. It should be reviewed and corrected accordingly.

- The "default" column appears to be irrelevant and lacks relevance to the provided data. I recommend removing this column from the dataset.

3. CustomerAddress (third dataset):

- The "customer_id" values are not in sequential order, and some IDs are missing from the dataset. It is advisable to review and ensure that the customer IDs are consistent and sequential.

- The "property valuation" should be in a specific currency. Additionally, if the values are in large amounts, it would be helpful to include the currency symbol or appropriate formatting to enhance clarity.

- The states "New South Wales" and "Victoria" should be replaced with their respective abbreviations: "NSW" and "VIC" for consistency.

By addressing these data quality concerns, we can improve the accuracy and reliability of our analysis.

Best regards,

Vishnav Tuli