PHASE-I: DATA PRE-PROCESSING AND CLEANING

Team Member's Details:

• Group Name: “I am the TEAM”

• Name: Krishnakant Sonji

• Email: cronjobs alo.edu, krishnasonji@gmail.com

• Country: United States

• College/Company: University of New York at Bu alo

• Specialization: Data Science

Problem Description: XYZ Bank wants to send personalized Christmas offers to different customer groups instead of sending the same o er to all. The goal is to create no more than 5 customer segments using machine learning to improve the efficiency of the campaign.

GitHub Repo link: **CLICK HERE OR (https://github.com/Thir13een/DG-week-9)**

Data cleansing and transformation done on the data:

1) **Column Renaming Using a Dictionary Mapping:**
   First, I created a dictionary called column_mapping, which mapped the original names to more meaningful and descriptive names for better readability and interpretability of the dataset. Here is what changed:

   - 'fecha_dato' was renamed to 'data_partition_date'.
   - 'ncodpers' was renamed to 'customer_id'.
   - 'ind_empleado' was renamed to 'employee_status'.
   - 'pais_residencia' was renamed to 'residence_country'.
   - 'sexo' was renamed to 'gender'.

2) **Removing Columns with High Null Percentages:**
   Therefore, any columns that contained over 50-60% null values, such as last_primary_customer_date and spouse_employee_index, were dropped from the dataset as they contribute little to no value in such an analysis. By doing so, this step ensured that columns with too much missing data, which did not provide any value to the segmentation process, were excluded, hence improving the quality of the data by reducing the noise.

3) **Imputation and Dropping of Missing Values:**

- The missing values were imputed with the median of the column so that the distribution would be preserved while logically filling in the gaps.
- The column 'customer_joining_channel' contained a few unique entries other than a few missing values. Since each value depicted a certain channel, the rows containing NaNs were dropped in this column.
- We then clean the dataset further by removing the remaining rows containing NaN values in other columns after handling 'customer_joining_channel'.
- These steps ensure that logical imputation was done and removed rows that have missing critical data.

4) **Data Type Handling**:
   The columns' data type was changed, if required, to handle appropriate data types of values that each column possessed. Numeric columns were changed to the appropriate format while maintaining consistency in categorical columns.

5) **Handling Outliers**:
   - Unique test: Tracked down unexpected entries in columns by checking their unique values.
   - Manual Changes: Removed/corrected anomalies identified in unique value check.
   - Ensured Consistency: Related columns have consistent data types and ranges of values.