PHASE-I: DATA PRE-PROCESSING AND CLEANING

Team Member's Details:

- Group Name: “I am the TEAM”
- Name: Krishnakant Sonji
- Email: krsonji@buffalo.edu, krishnasonji@gmail.com
- Country: United States
- College/Company: University of New York at Buffalo
- Specialization: Data Science

Problem Description:

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

Data Understanding:

- The dataset has 1,000,000 rows and 48 columns, but the column names were not clearly defined and contained inconsistent naming.
- Some column names did not match the actual data they represented, causing confusion in understanding the dataset.
- There were mismatches in data types: some columns were marked as object but contained numeric or date values, requiring type conversion.
- The dataset had too many dimensions, making it harder to analyze without narrowing the focus to key features.
- A significant number of columns had considerable null values, especially in columns like renta (household income) and conyuemp (spouse employee index).

What type of data you have got for analysis?

Key data types include:

- Demographics: Age, gender, country of residence.
- Account/Product Usage: Savings, credit cards, mortgages.
- Income/Financial: Household income, customer activity.
- Geographical: Province and residence details.
- Employment: Employee status, spouse information.
- Transaction/Time-based: Payroll, direct debit, contract start date.

The data has the following issues:

- Missing Values: Several columns have a significant amount of missing data.
- Outliers: There are extreme values that deviate from typical ranges.
- Data Type Mismatch: Some columns have incorrect data types, requiring conversion.
- Inconsistent Column Naming: Column names are unclear and need standardization.
- Too Many Dimensions: The dataset has a high number of columns, making it complex to analyze.

Approach towards problems:

- Renaming Columns: All columns were renamed for clarity and consistency.
- Handling Outliers: Outliers were identified using the unique() function and removed by dropping rows with extreme values.
- Correcting Data Types: Data types were converted to their appropriate formats to ensure accurate analysis.
- Dealing with Missing Values:
    - Columns with a high percentage of missing values (e.g., 90% or more) were removed.
    - Rows with significant missing values were assessed for importance and removed if necessary.
    - Techniques like mean imputation were considered, but unique missing values that couldn't be meaningfully replaced were removed.
    - Removing Columns with Single Unique Values: Columns containing only one unique value were removed as they provided no useful information for analysis.