Problem Description

XYZ Credit Union, a financial institution based in Latin America, has successfully marketed and sold individual banking products such as credit cards, deposit accounts, retirement accounts, and safe deposit boxes. However, the credit union faces a significant challenge in cross-selling; existing customers are typically purchasing only one product. This limitation in cross-selling indicates untapped potential for increasing customer engagement and revenue through the sale of additional products to the current customer base. XYZ Credit Union has enlisted the help of ABC Analytics to address this issue and improve their cross-selling capabilities.

EDA Performed on the data

Missing Values

1. **Handling Missing Values:**
   - **Before Treatment:**
     - Several columns had significant missing values, such as ult_fec_cli_1t (927,932 missing), conyuemp (929,511 missing), and antiguedad (3 missing).
   - **After Treatment:**
     - Missing values were successfully imputed using mean for numeric columns and mode for non-numeric columns. The KNN imputer was also applied, resulting in zero missing values.

Skewness

2. **Skewness:**
   - **Before Treatment:**
     - High skewness was observed in columns like ind_nuevo (5.739031), antiguedad (-555.491690), indrel (23.438419), indrel_1mes (185.543646).
   - **After Treatment:**
     - Skewness was significantly reduced, particularly for ind_nuevo, indrel, and indrel_1mes, which became zero after treatment. Some columns like ncodpers and cod_prov still had moderate skewness.

Outliers

3. **Outliers:**
   - **IQR Method:**
     - **Before Treatment:** High number of outliers in ind_nuevo (25,889), age (7,035), indrel (1,683).
     - **After Treatment:** Outliers significantly reduced for ind_nuevo, indrel, indrel_1mes to zero, but increased for ncodpers (49,069), age (1,236), and cod_prov (32,518).
   - **Z-score Method:**

        - **Before Treatment:** Similar high outliers in ind_nuevo (25,861), age (5,990), indrel (1,680).
        - **After Treatment:** Outliers significantly reduced for ind_nuevo, indrel, indrel_1mes to zero, but increased for ncodpers (18,095), age (2,003), antiguedad (4,308), and cod_prov (10,371).

Final Recommendations

1. **Data Cleansing:**
    - Apply mean/mode imputation followed by KNN imputation for handling missing values. This ensures no missing data and leverages patterns within the data for imputation.
2. **Transformation:**
    - Use log transformation as the primary method to address skewness and outliers. This method effectively normalized data distributions and reduced the number of outliers.
    - Monitor residual skewness and apply additional transformations if necessary.
3. **Outliers:**
    - Use a combination of clipping and log transformation to handle outliers. For columns with persistent outliers, consider further methods like Winsorization.
    - Regularly review the data and apply business rules to handle outliers contextually.

By applying these methods, the dataset is better prepared for analysis, with minimized impact from missing values, skewness, and outliers. These steps enhance data quality, leading to more reliable and valid analytical outcomes.