Week 8: Deliverables

Group Name: Solo Analyst

Name: Munirah Alfehaid

Email: munirah9hamad@gmail.com

Country: Saudi Arabia

Specialization: Data Analyst

Problem Description

XYZ Credit Union in Latin America has been successful in selling individual banking products like credit cards, deposit accounts, and retirement accounts. However, they are facing a challenge in cross-selling, as existing customers are typically not purchasing more than one product. The goal of this project is to analyze the current situation and suggest strategies to increase cross-selling opportunities among existing customers without relying on machine learning solutions.

What type of data do you have for analysis?

The dataset consists of a mix of data types:

- **Numerical Data**: These include columns like ncodpers, age, ind_nuevo, antiguedad, indrel, tipodom, cod_prov, and ind_actividad_cliente. These are mostly integers, representing counts, IDs, or numerical values.
- **Categorical Data**: These include columns like sexo, ind_empleado, pais_residencia, segmento, and others, which are represented as text strings and indicate categories or groups.
- **Date/Time Data**: Columns like fecha_dato and fecha_alta appear to be dates, which can be useful for time-based analysis.

What are the problems in the data?

The dataset has several issues:

- **Missing Values**:
    - **High Missingness**: Columns like ult_fec_cli_1t and conyuemp have an extremely high percentage of missing values, with over 99% of their data missing.
    - **Moderate Missingness**: Columns like canal_entrada, cod_prov, nomprov, and segmento have missing values but to a lesser extent.

    - **Low Missingness**: Columns like sexo, indrel_1mes, and tiprel_1mes have very few missing values.
- **Mixed Data Types Warning**: There's a warning about mixed data types in the conyuemp column, which could lead to issues during analysis.
- **Potential Outliers**: Without seeing the actual distribution, certain columns (like age and renta) might have outliers, which can skew analysis results.

3. **What approaches are you trying to apply to overcome problems like NA values, outliers, etc., and why?**

To address these problems:

- **Handling Missing Values**:
    - **Drop Columns**: Consider dropping columns like ult_fec_cli_1t and conyuemp due to the overwhelming amount of missing data.
    - **Imputation**: For columns with moderate or low missing values, you can impute the missing values using the median or mode, depending on the data type. For example:
        - **Categorical Data**: Impute with the most frequent category.
        - **Numerical Data**: Impute with the median, especially if the data is skewed.
- **Mixed Data Types**: Convert the conyuemp column to a consistent data type if it's necessary for analysis, or consider dropping it if it's not valuable.
- **Handling Outliers**: Once identified, outliers can be managed by either capping them to a maximum value or transforming the data (e.g., log transformation) to reduce their impact.

4. Why are you applying these approaches?

- **Dropping Columns with High Missingness**: Columns with over 99% missing data are likely to be unreliable and contribute little to the analysis. Removing them helps streamline the dataset and avoids introducing noise.
- **Imputation**: Imputing missing values ensures that you don't lose valuable data from rows where only a few values are missing, which helps maintain the dataset's integrity and allows for more robust analysis.
- **Handling Outliers**: Outliers can distort statistical analyses and machine learning models. By managing them appropriately, you ensure that your analysis is more representative of the true data distribution.