Data Glacier Virtual Internship – Final Project (Week 8)

Cross-Selling Opportunity Analysis for XYZ Credit Union

1. Team Member Details

Group Name	Name	Email	Country	College	Specialization
DataVision	Raj Pawar	rajpawar32646@gmail.com	UK	University of Liverpool	Data Analysis
DataVision	Naga Pavithra Jajala	pavithrajajala8naga@gmail.com	UK	Birmingham City University	Data Analysis

2. Problem Description

XYZ Credit Union is experiencing challenges with **cross-selling** banking products. While individual product uptake (e.g., credit cards, savings accounts) is strong, most customers are using only one product. Our goal is to understand customer behaviour using structured banking data and generate **data-driven insights and recommendations** for cross-selling — **without using machine learning**. We aim to support business teams by identifying customer segments likely to adopt additional products.

3. Data Understanding

- Train.csv : 13M rows, 48 columns, Size: 2.13 GB, Contains - Customer info + product indicators (targets)
- Test.csv : 929,615 rows, 24 columns, Size: 105 MB, Contains - Customer info only (no product columns)
- Includes demographics, income, customer tenure, and binary product ownership columns
- Most fields are categorical or numeric

4. Types of Data in the Dataset

Category	Fields
Demographic	sexo, pais_residencia, age, segmento
Behavioral	ind_actividad_cliente, antiguedad, indfall
Financial	renta, ind_nomina_ult1, product flags (ind_cco_fin_ult1...)
Temporal	canal_entrada, tiprel_1mes, ind_empleado
Target indicators	fecha_dato, fecha_alta, ult_fec_cli_1t

5. Problems found in the Data

1. **Missing Values:**
   - Train: Up to ~13M missing in **ult_fec_cli_1t, conyuemp**
   - **renta**: ~2.7M missing (needs attention)
   - **segment:** ~189k
   - **canal_entrada:** ~186k
   - **indrel_1mes:** ~149k
   - **ult_fec_cli_1t:** ~13.6 million
   - **conyuemp:** ~13.6 million

2. **Dirty Data:**
   - **age** and **antiguedad** contain **" NA"** and **" NA"** - need to be cleaned
   - Some categorical fields contain mixed types (e.g., **indrel_1mes** = 1, 2, "P")

3. **Outliers:**
   - **renta** is highly skewed, contains extreme values above 1,000,000

6. Approaches to Handle Data Issues

1. **Handling Missing Values:**
   - **renta:** Impute using **median** or segment-wise averages (based on **segmento, nomprov**)
   - **age, antiguedad:** Convert strings to numeric and impute invalid values with median
   - **ult_fec_cli_1t, conyuemp:** Drop or fill with "Unknown" depending on their usage in EDA

2. **Handling Outliers:**
   - Use **IQR** or **Z-score** to detect income outliers in **renta**
   - Cap outliers if necessary for visualization clarity

3. **Handling Skewed Variables:**
   - Product flags are binary and expected to be skewed
   - Will visualize distribution to inform cross-sell strategies by segment