

## 1. Team Member's Details:

- **Group Name:** Data Innovators
  - **Name 1:** Stephanie Dawsonn-Andoh
    - **Email:** stephandoh@gmail.com
    - **Country:** Ghana
    - **Specialization:** Data Analytics
  - **Name 2:** Sreedhar Rongala
    - **Email:** rongalasreedhar@gmail.com
    - **Country:** India
    - **Specialization:** Data Analytics
  - **Name 3:** Brittney Smith
    - **Email:** brittneysintership@gmail.com
    - **Country:** USA
    - **Specialization:** Data Analytics
- 

## 2. Problem Description:

XYZ Bank is aiming to enhance its marketing campaign by delivering personalized Christmas offers to its customers. Instead of rolling out a generic offer for everyone, they want to target specific customer segments with relevant offers. To solve this problem efficiently, XYZ Bank approached ABC Analytics Company to help with customer segmentation. The bank's requirement is to group customers into no more than 5 segments, as more segments would be inefficient for their marketing efforts.

---

## 3. Data Understanding:

### Dataset Context

#### 1. Customer Demographics and Attributes

- The dataset includes a unique identifier for each customer (ncodpers), their country of residence (pais\_residencia), gender (sexo), and age. This information helps understand the geographic distribution and demographic characteristics of the customer base.

#### 2. Customer Relationship and Status

- It captures the employment status of customers (ind\_empleado), which provides insight into their financial stability. The dataset also includes an indicator for new customers (ind\_nuevo), the length of their relationship with the bank (antiguedad), and the type of relationship they have with the bank (indrel). Additionally, it tracks the customer's relationship status at the beginning of the month (indrel\_1mes) and the nature of that relationship (tiprel\_1mes).

### 3. Financial Products and Services

- This section outlines various financial indicators, such as whether the customer is active (`ind_actividad_cliente`), their gross income (`renta`), and whether they hold specific products like savings accounts, current accounts, mortgages, and credit cards. Each product has a corresponding indicator that shows if the customer has that type of account or service.

### 4. Temporal Aspects

- The dataset records key dates, including the date the customer became the primary holder of a contract (`fecha_alta`) and the last date they were considered a primary customer (`ult_fec_cli_1t`). These temporal aspects are essential for understanding customer engagement over time.
- 

## 4. Data Types in the dataset:

### Categorical Data:

- The dataset contains various categorical variables that help classify customers and their attributes. Examples include:
  - Customer Demographics: `pais_residencia` (Country of residence), `sexo` (Gender), `ind_empleado` (Employment status), and `canal_entrada` (Entry channel).
  - Relationship Status: `indrel` (Customer relationship type), `indrel_1mes` (Relationship type at the beginning of the month), and `tiprel_1mes` (Type of relationship).

### Numerical Data:

- Numerical variables are essential for quantitative analysis and can be further categorized into discrete and continuous data:
  - Discrete Data:
    - `age` (Customer's age) and `ind_nuevo` (New customer index) are examples of discrete numerical data.
  - Continuous Data:
    - `renta` (Gross income) and `antiguedad` (Seniority in months) are examples of continuous data that can provide insights into customer financial capacities and durations of relationships with the bank.

### Date/Time Data:

- The dataset includes date-related variables that track customer interactions with the bank:
    - `fecha_alta` (Date of becoming the primary holder of a contract) and `ult_fec_cli_1t` (Last date as a primary customer) are crucial for analyzing customer engagement over time.
-

## 5. Challenges and Solutions:

### Language Barrier:

- **Problem:** The dataset contained column names in Spanish, creating a barrier to understanding.
- **Solution:** Translated all column names from Spanish to English to facilitate better comprehension of the data.

### High Percentage of Missing Data:

- **Problem:** The columns `ult_fec_cli_1t` (latest date as primary customer) and `conyuemp` (spouse employee indicator) had more than 90% missing data, rendering them ineffective for analysis.
- **Solution:** Removed these columns from the dataset.

### Irrelevant Column:

- **Problem:** The column `Unnamed: 0` did not provide useful information.
- **Solution:** Dropped this column to streamline the dataset.

### Redundant Geographic Information:

- **Problem:** The columns `cod_prov` (province code) and `nomprov` (province name) provided redundant information since country data was more pertinent.
- **Solution:** Removed these columns to focus on more relevant geographic data.

### Recency Overlap:

- **Problem:** The `fecha_alta` (signup date) column was not necessary for analyzing customer recency, as `antiguedad` (seniority) could serve that purpose.
- **Solution:** Dropped the `fecha_alta` column.

### Missing Values in Categorical Columns:

- **Problem:** Categorical columns such as `sexo` (gender) and `pais_residencia` (country of residence) contained missing values.
- **Solution:** Replaced missing values with the most frequent values (mode) in those columns.

### Missing Values in Numerical Columns:

- **Problem:** The columns `renta` (income), `ind_nomina_ult1` (payroll product indicator), and `ind_nom_pens_ult1` (pension nomination product indicator) had missing values.
- **Solution:** Replaced missing values with median values for these numerical columns to minimize the impact of outliers.

### Data Type Adjustments:

- **Problem:** Several columns had inappropriate data types, such as float where integer was more suitable.

- **Solution:** Changed the data types from float to int for several columns, including ind\_empleado (employee indicator) and indresi (resident indicator), among others.

#### **Date Format Correction:**

- **Problem:** The fecha\_dato (data date) column was in an object format instead of a date format.
- **Solution:** Converted fecha\_dato to a date data type for accurate temporal analysis.

#### **Outliers in Age Column:**

- **Problem:** The age column had a maximum age of 116 and several outliers above the upper bound of 92.
- **Solution:** Capped all outliers in the age column at the upper bound of 92 to ensure realistic values.

#### **Negative Values in Seniority:**

- **Problem:** The antiguedad (seniority) column had a minimum value of -999999, which was nonsensical.
- **Solution:** Capped outliers in this column to a minimum sensible value of 0 after investigation.

Data storage location:

[https://github.com/stephandoh/Data-Glacier-Internship/tree/main/Week\\_8](https://github.com/stephandoh/Data-Glacier-Internship/tree/main/Week_8)