Team Members:

- Ray Ng
  - rayng2018@gmail.com
  - Canada
  - University of British Columbia
  - Specialization: Data Science
- Rita Uzoka
  - rita.uzoka@yahoo.com
  - United Kingdom
  - Sheffield Hallam University
  - Specialization: Data Science
- Fatemeh Bagheri
  - f.bagheri13@gmail.com
  - France
  - Université Jean Monnet St Etienne - Université de Lyon
  - Your Specialization: Data Science

Final Project Report & Code

Problem Statement

XYZ bank wants to roll out Christmas offers to their customers. But the bank does not want to roll out the same offer to all customers, instead they want to roll out personalized offers to particular sets of customers. If they manually start understanding the category of customer then this will not be efficient and also they will not be able to uncover the hidden pattern in the data (pattern which groups certain kinds of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want more than 5 groups as this will be inefficient for their campaign.

Business Understanding

ABC analytics should try to understand certain patterns in customer behaviour, which include relations between province, sex, seniority, etc. and household income. ABC must find at most five groups in which customers share common behaviors.

Data Issues, Transformation and Cleaning

We found that the issue with the original data set were:

- **Data Incompleteness** - various columns, such as the ind_empleado (Employee index), pais_residencia (Customer's Country residence), sexo (Customer's sex) etc., have empty

values (NA) in certain records.
    - Mitigation: Impute the missing values based on distribution in the training dataset. Depending on the context of the attribute, fill in the missing values with an appropriate value. If the variable is categorical, use the mode. If the variable is quantitative, use the median if discrete, and the mean if continuous.
    - Recommendation: Make input for such fields required, especially easily obtainable ones such as customer sex. If information cannot be obtained, then do not include. Alternatively, may consider removing some attributes which are not so important.
- **Inconsistent data type for the same attribute** (e.g. numeric values for some fields and strings for others).
    - Mitigation: Make data transformations to ensure consistent data types in each column. Examples: Convert selected records in strings to numeric, and remove non-numeric characters from the string.
    - Recommendation: Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field makes it difficult to interpret results at the later stage.
- **Inconsistent values for the same attribute** (e.g. -999999 as a value for seniority in months).
    - Mitigation: The seniority in months cant have a negative number. Use regular positive numbers to replace extended values into positive numbers or zero to ensure consistency across seniority in months. Or replace -999999 with NA and perform the missing data transformations as above using the median
    - Recommendation: Enforce a mandatory field where only positive integers are valid for this attribute.

Data Models

The model we have decided to use is based on the clustering principle. The model is called the **k-means clustering**, which divides the data into k clusters.

Results

One way to cluster the customers is to do it based on the age and household income. The customers could be segmented by:

- High-income (above $250,000) customers
- Old age (above 60) customers with household income not exceeding $250,000
- Low income (under $70,000) customers with age not exceeding 60
- Other middle aged (35-60) customers
- Other young (under 35) customers

Another clustering scheme is to extract the seniority and household income into five groups:

- Customers with lower income (under $100,000) and less seniority (under 96 months)
- Customers with lower income (under $100,000) and more seniority (over 96 months)
- Customers with medium income (between $100,000 and $250,000) and less seniority (under 96 months)
- Customers with medium income (between $100,000 and $250,000) and more seniority (over 96 months)
- Customers with higher income (over $250,000)