Customer Segmentation

Team Name: **DGI**

Team Member Details

No	Name	Email	Country	College/ Company	Specialization
1	Rahma Mahjoub Abker Habeeb	Rahma.mahgoub@gmail.com	Kuwait	Computer Science	Data Science
2	Nonhlanhla L Luphade	nonnynyathi4611@gmail.com	Zimbabwe	University of Cape Town	Data Science
3	Ajaegbu Ebuka Emmanuel	ajaegbu35@gmail.com	Nigeria	Grand Treasury Ltd	Data Science
4	Robin Masawi	Masawirobin69@gmail.com	Zimbabwe	Econet	Data Science

Problem Statement:

The Bank XYZ wants to roll out personalized Christmas offers for certain customers instead of rolling out the same offers for all customers. As an alternative of trying to manually decide which customer is which category. The bank seeks an efficient approach that enables them to uncover hidden patterns in their customer data and categorize customers into a 5 unique groups.

Business Understanding:

- Customer segmentation is the process of categorizing the customers into various groups according to their characteristics or behaviors.
- This will help the companies effectively match their products to the exact customers groups.

Data Understanding:

The data contains customer data for Latin American bank. Each customer has a customer id which uniquely identifies them. The data contains customers local customers and foreign customers, and the demographic data of the customers. Additionally, the data contains information about how active each customer is, customer seniority and the number of bank products each customer use. The table below gives a description of the data:

Variable name	**Description**	**Type of data**	**Number of missing values**
ncodeper	Customer code/id	numeric	0
Ind_empleado	Employee index: A-active, B – ex employed, F-filial, N – not employed, P pasive	categorical	10782
Pais_residencia	Customer's country residence.	object	10782
sexo	Customer's sex	categorical	10786
age	Customer's age	numeric	10782
Fecha_alta	Date which customer became first contract holder at the bank.	date	10782
Ind_nuevo	New customer index: 1- registered in the last 6 months	categorical	10782
Antiguedad	Customer seniority in months	Numeric	0
Indrel	1(primary/first), 99 (primary beginning of month and not at the end of month)	categorical	10782
Ult_fec_cli_1t	Last date as primary customer	date	998899
Indrel_1mes	Customer type beginning of month. 1-primary, 2-co-owner,P-potential, 3 former primary, 4- former co-owner	Categorical	10782
Tiprel_1mes	Customer relation type at the beginning of month. A – active, I-inactive, P- former customer, R-potential	Categorical	10782
Indresi	Residence index (S-yes,N -no)	Categorical	10782

Indext	foreign index (S-yes,N -no)	Categorical	10782
Conyuemp	Spouse index. 1 if customer spouse is employee	Categorical	999822
Canal_entrada	Channel customer used to join bank	categorical	10861
Indfall	Deceased index (S-yes,N -no)	Categorical	10782
Tipdom	Address type 1 primary	numeric	10782
Cod_prov	Province code	object	17734
Nomprov	Province name	object	17734
Ind_actividad_cliente	Activity index (1 active , 0 inactive)	Categorical	10782
Renta	Gross income of household	Numeric (float)	175183
Ind_ahor_fin_ult1	Savings account	categorical	0
Ind_aval_fin_ult1	Guarantees	categorical	0
Ind_cco_fin_ult1	Current account	categorical	0
ind_cder_fin_ult1	Derivada account	categorical	0
ind_cno_fin_ult1	Payroll account	categorical	0
ind_ctju_fin_ult1	Junior account	categorical	0
ind_ctma_fin_ult1	Mas particular account	categorical	0
ind_ctop_fin_ult1	Particular account	categorical	0
ind_ctpp_fin_ult1	Particular plus account	categorical	0
ind_deco_fin_ult1	Short_term deposits	categorical	0
ind_deme_fin_ult1	Medium_term deposits	categorical	0
ind_dela_fin_ult1	long_term deposits	categorical	0
ind_ecue_fin_ult1	e-account	categorical	0
ind_fond_fin_ult1	Funds	categorical	0
ind_hip_fin_ult1	Mortgage	categorical	0
ind_plan_fin_ult1	Pensions	categorical	0
ind_pres_fin_ult1	Loans	categorical	0
ind_reca_fin_ult1	Taxes	categorical	0
ind_tjcr_fin_ult1	Credit cards	categorical	0
ind_valo_fin_ult1	Securities	categorical	0
ind_viv_fin_ult1	Home account	categorical	0
ind_nomina_ult1	Payroll	categorical	0
ind_nom_pens_ult1	Pensions	categorical	0
ind_recibo_ult1	Direct debit	categorical	0

Suggested methods for handling NAs and outliers.

For handling NAs, the suggested methods are:

- Drop variables with more than 80% missing values.
- Replace numeric values like gross income with the mean.
- Replace categorical variables with the mode (most common category in that variable).

For handling outliers, the suggested methods are:

- Use the interquartile ranges to determine outliers in numeric variables. (lower quartile – 1.5*interquartile range < x<upper quartile + 1.5 * interquartile range).

Suggested methods for handling features.

1. The fecha_alta variable which is the date which the customer became the first holder of a contact at the bank can be changed by subtracting the customer's date with the max date 2015-01-28 to get the number of years since first contract.
2. The 24 variables that start with the letters ind_ and end with the letter _ult1 which describe the different accounts or products the customers use can be summed up to one variable which counts the number of products each customer makes use of in the bank.

Suggested methods for categorical variables (Label encoding).

The categorical variables can be separated to ordinal variables and nominal variables.

Ordinal variables:

1. Indrel – there is an order 1 is more important than 99.
2. Indrel_1mes – order of importance is 1(primary), 2 (co-owner), 3 (former primary)
3. Tiprel_1mes – order of importance is A (active), P (potential), I (inactive)

Nominal Variables:

1. Multicategory variables: Cod_prov, nomprov, canal_entrada and pais_residencia, these variables have more than 30 categories thus, one hot encoding would not be best. We can make us of frequency encoding.
2. Binominal variables: The rest of the nominal variables have 2 categories; thus, we can use one hot encoding.

Github Link: https://github.com/shelovescode000/CustomerSegmentationProject