GROUP NAME: DATA ALCHEMIST

TEAM MEMBERS:

Name: Yuhong Chen

Email: ryechenn@gmail.com

Country: Canada

College/Company: Mcgill University

Name: Haoyue Chang

Email: kerrychy1215@gmail.com

Country:USA

College/Company: Northeastern University

Name: Gifty Osei

Email: gifty18osei@gmail.com

Country: USA

College/Company: Montana State University

SPECIALIZATION: DATA ANALYST

Problem description

We are determined to help XYZ Bank improve its cross-selling strategies and enhance customer engagement. The bank offers a wide array of financial products and services, including savings accounts, credit cards, mortgages, loans, and investment options. However, we've observed that many of our customers have limited product adoption and aren't fully utilizing the range of services available to them.

To tackle this challenge head-on, we plan to implement customer segmentation techniques to gain deeper insights into our customer base. By dividing our customers into distinct groups based on their demographics, financial behavior, and product usage

patterns, we hope to identify specific customer segments that are more likely to use products and services. Armed with this valuable information, we aim to create personalized marketing strategies and tailored cross-selling initiatives to boost customer satisfaction and encourage higher product adoption.

As part of our data analysis team, the objective is to thoroughly analyze the extensive customer dataset provided by XYZ Bank and conduct a comprehensive customer segmentation analysis. The dataset includes detailed information about each customer, such as age, gender, income, transaction history, product holdings, and tenure with our bank.

Data understanding

- Customer demographics: Age, gender, location, and purchase history.
- Website interactions: Clickstream data, session duration, and product views.
- Purchase behavior: Cart abandonment, order history, and customer feedback.
- Customer support interactions: Queries, response times, and issue resolution.
- Different products for sale: Credit Card, particular Account, loans and deposits.

What type of data you have got for analysis

Floats, integers, and objects

What are the problems in the data (number of NA values, outliers , skewed etc)
Missing values in training dataset:

#missing values checking df1.isnull().sum() ind ahor fin ult1 0 fecha_dato 0 ind_aval_fin_ult1 0 ncodpers 0 ind_cco_fin_ult1 0 ind_empleado 27734 ind cder fin ult1 0 pais_residencia 27734 ind cno fin ult1 0 27804 ind ctju fin ult1 0 age ind_ctma_fin_ult1 0 fecha_alta 27734 ind_ctop_fin_ult1 0 27734 ind ctpp fin ult1 0 ind nuevo antiguedad ind deco fin ult1 0 indrel 27734 ind deme fin ult1 ind_dela_fin_ult1 0 ult_fec_cli_1t 13622516 ind_ecue_fin_ult1 0 indrel_1mes 149781 ind_fond_fin_ult1 0 tiprel_1mes 149781 ind hip fin ult1 0 indresi 27734 ind plan fin ult1 0 indext 27734 ind pres fin ult1 0 conyuemp 13645501 ind_reca_fin_ult1 0 canal_entrada 186126 ind_tjcr_fin_ult1 0 indfall 27734 ind valo fin ult1 0 tipodom 27735 ind viv fin ult1 0 cod prov 93591 ind nomina ult1 16063 nomprov 93591 ind_nom_pens_ult1 16063 ind actividad cliente 27734 ind recibo ult1 renta 2794375 dtype: int64 189368 segmento

missing values in testing dataset:

fecha_dato	0
ncodpers	0
ind_empleado	0
pais_residencia	0
sexo	5
age	0
fecha_alta	0
ind_nuevo	0
antiguedad	0
indrel	0
ılt_fec_cli_1t	927932
indrel_1mes	23
tiprel_1mes	23
indresi	0
indext	0
conyuemp	929511
canal_entrada	2081
indfall	0
tipodom	0
cod_prov	3996
nomprov	3996
ind_actividad_cliente	0
centa	0
segmento	2248

Outliers in the training dataset:

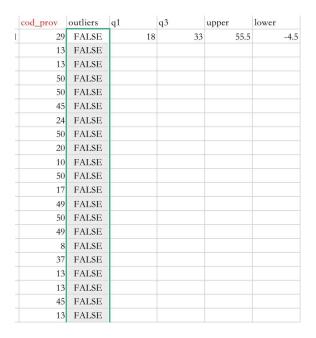
For age: There are 15891-11370 outliers, which is 4521.



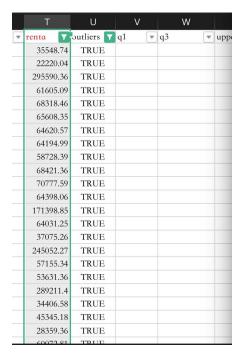
For antigucdad: There are 11374-11370 outliers, which is 4.

antiguedad	outliers	q1	q3	upper	lower
6	FALSE	24	154	349	-171
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
35	FALSE				
25	DATCE				

For cod_prov: There are 18332 outliers outcome but they are all from NA values so no outliers.



For renta: There are 431706 outliers.



What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

For missing object values in the dataset, we will delete them by removing rows directly. This approach is straightforward but maybe lead to a loss of valuable data. For those columns that contain floats and integers, we will fill in the missing values with estimated or substituted values. Common methods include using mean, median, or mode for

numerical variables, or using the most frequent category for categorical variables. And for outliers in the dataset, first we would handle with missing values and then we deal with outliers. There are some columns have outliers because of NA values, and there are less outliers that are real outliers in the dataset, so we would deal the real outliers in same way as missing values, whatever substitute them with median, mean or else.