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:

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

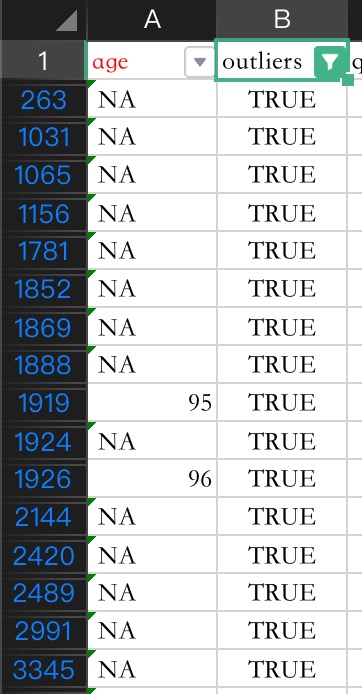
missing values in testing dataset:

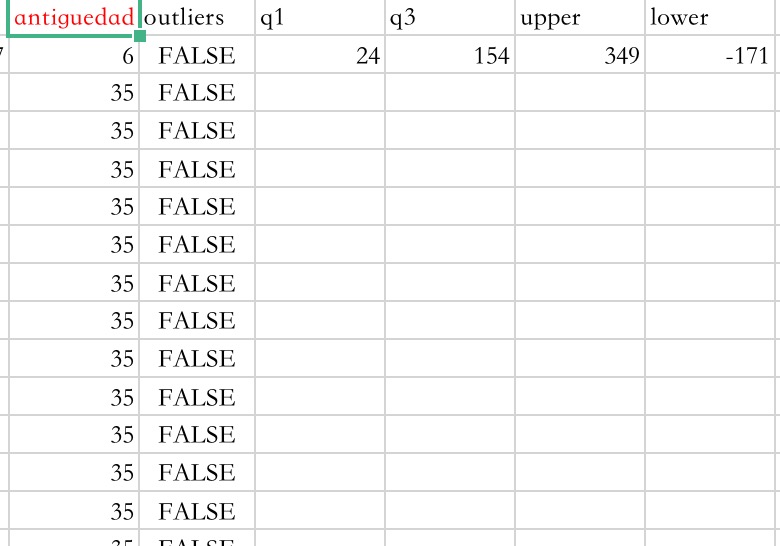
A screenshot of a computer code

Description automatically generated

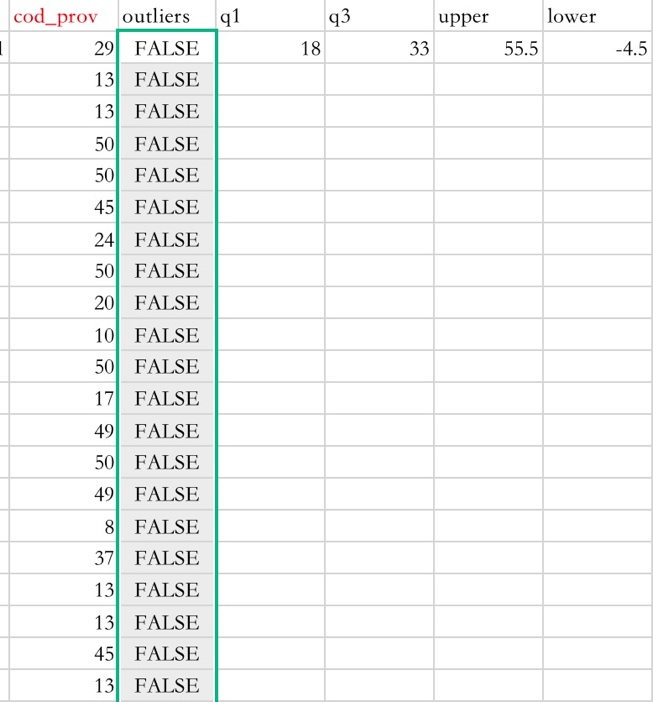
Outliers in the training dataset:

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

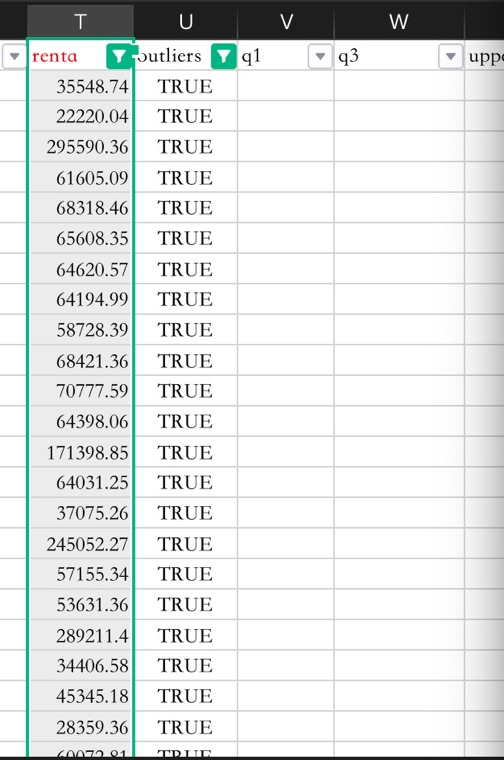


For antigucdad: There are 11374-11370 outliers, which is 4.

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 missing values and then we deal with outliers. There are some columns that have outliers because of NA values, and there are fewer outliers that are real outliers in the dataset, so we would deal with the real outliers in the same way as missing values, whatever substitute them with median, mean, or else.