Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis

Customer Segmentation for XYZ Bank

December 1, 2022

Team Member Details

Group Name: Elizabeth's Analytics

Name: Elizabeth Banning

Email: estall@hotmail.com

Country: USA

College: Western Governors University

Specialization: Data Science

Data Glacier
Your Deep Learning Partner

Agenda

Executive Summary

Problem Statement

Approach

Missing Values and Data Cleaning

Distributions of Features

Important Customer Categories

EDA Summary

Recommendations

Executive Summary

- **Purpose:** Segment customers into 2-5 groups for marketing campaign
- **Methods:** Clean data, then use k-means clustering analysis
- **Timeline:** Final results by December 30, 2022
- **Results of EDA:** Dataset cleaned, correlations and distributions explored

Problem Description

In order to develop its promotional campaign, XYZ Bank needs to know the answers to the following questions:

- What is the best number of groups to divide customers into?
- What are the primary characteristics of each group?

To answer these questions, the k-means clustering algorithm will be used to segment the customers, and the inertia metric will be used to determine the optimal number of groups (k). Finally, the characteristics of each group will be summarized so that XYZ Bank can determine which offers to develop and target to each group.

Problem Statement

- XYZ is a bank that wants to do a promotion
- 1,000,000 customers: need to tailor different promotions to different types of customers
- Maximum 5 groups
- **How can customers be grouped?**
- **What are the characteristics of each group?**

Approach

1 file used

- 1,000,000 customers (rows)
- 48 features (columns)

Clean the data:

- Check for duplicates and remove
- Check for missing values (treatment depends on type of data)
- Check for impossible/nonsense data and correct if necessary
- Drop irrelevant features (ID number, etc.)

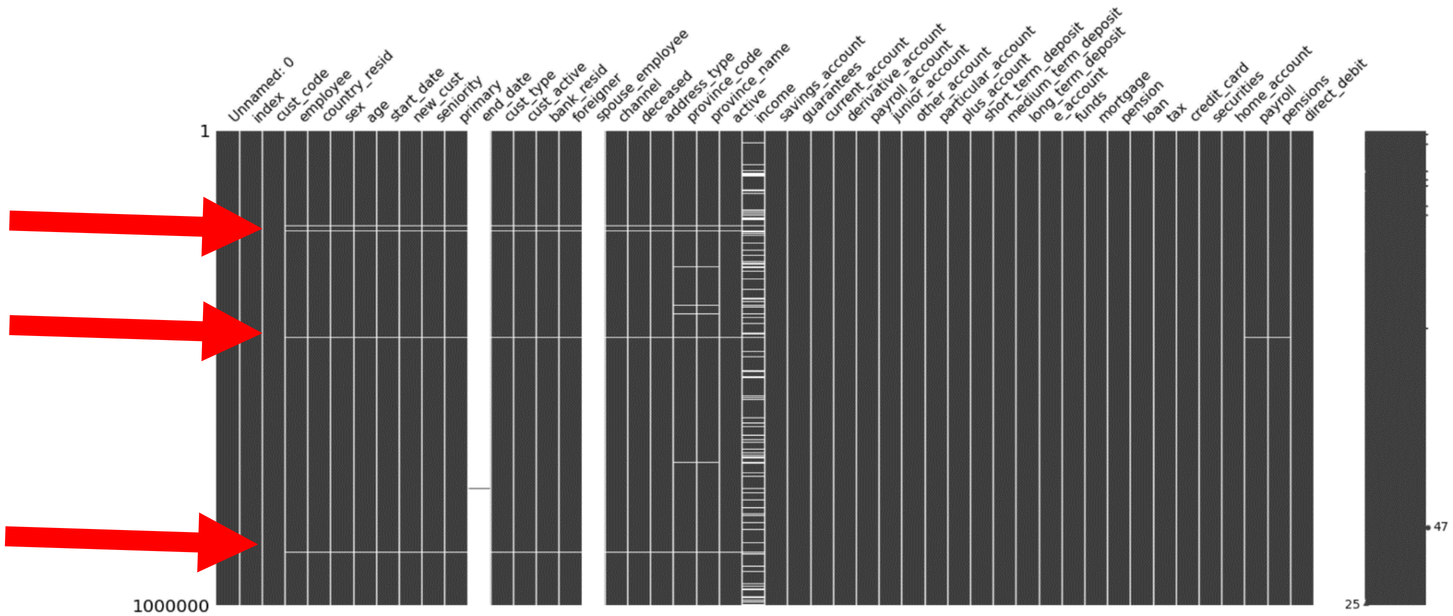
Approach

Explore the data:

- Distributions of numeric features
- Categories with high numbers of customers in each category
- Correlations of features (likely to be related to grouping customers)

Missing values

- Missing values are consistent in certain groups of rows
- Customer demographic information is missing, but services/accounts are present
- 10,782 rows (1.08% of total data)
- Rows were deleted (do not provide enough customer information)

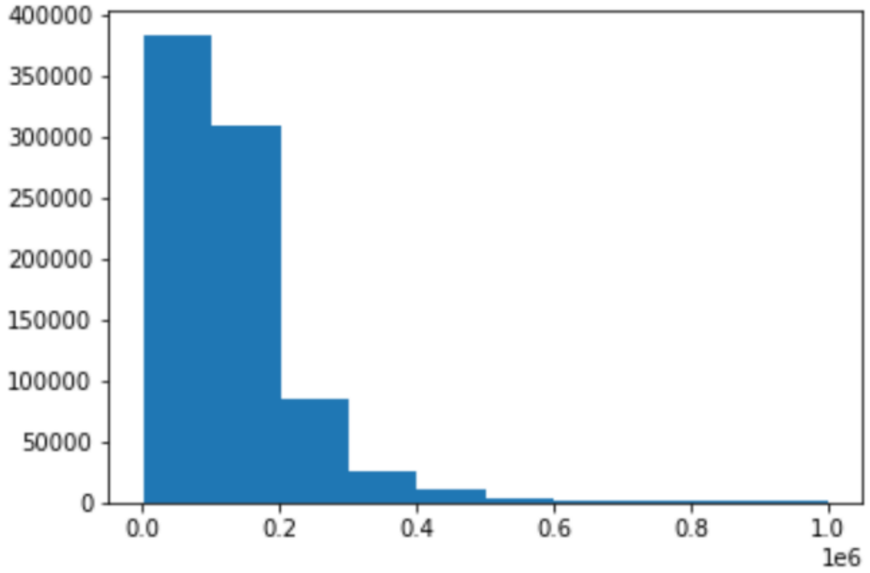


Missing values

Remaining missing values were imputed with:

- Mode – most common value (gender, spouse of employee, payroll, pensions)
- Median (income) due to positively-skewed distribution

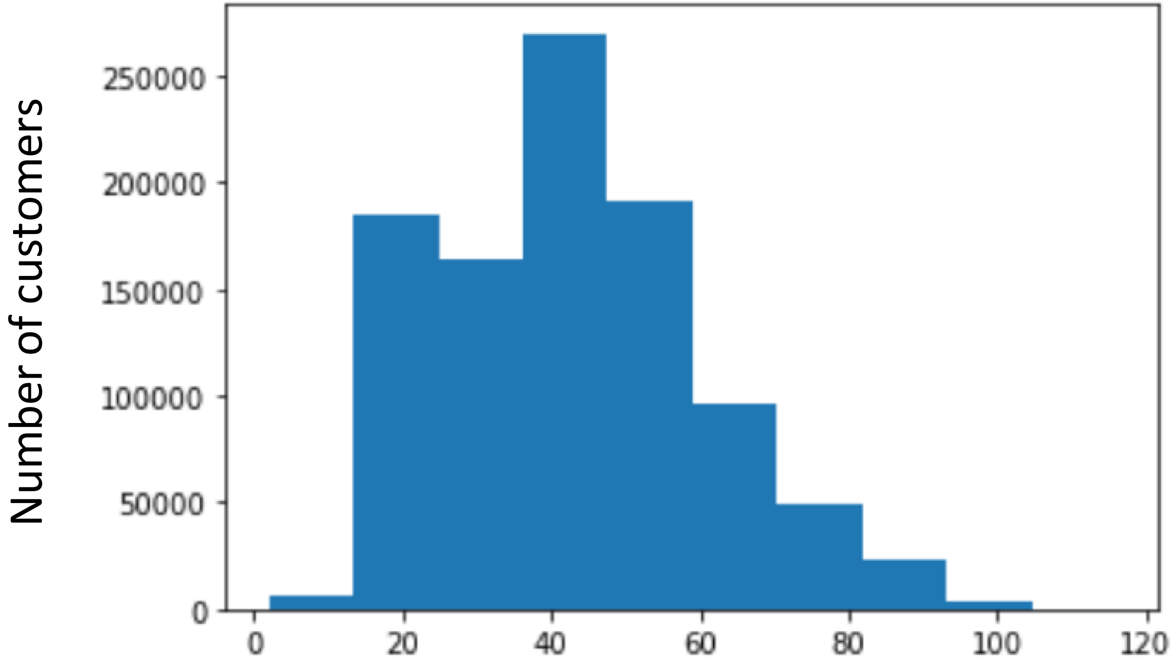
Distribution of income less than 1,000,000:



Other distributions of data

Note: Negative seniority values did not make sense and were changed to equal zero.

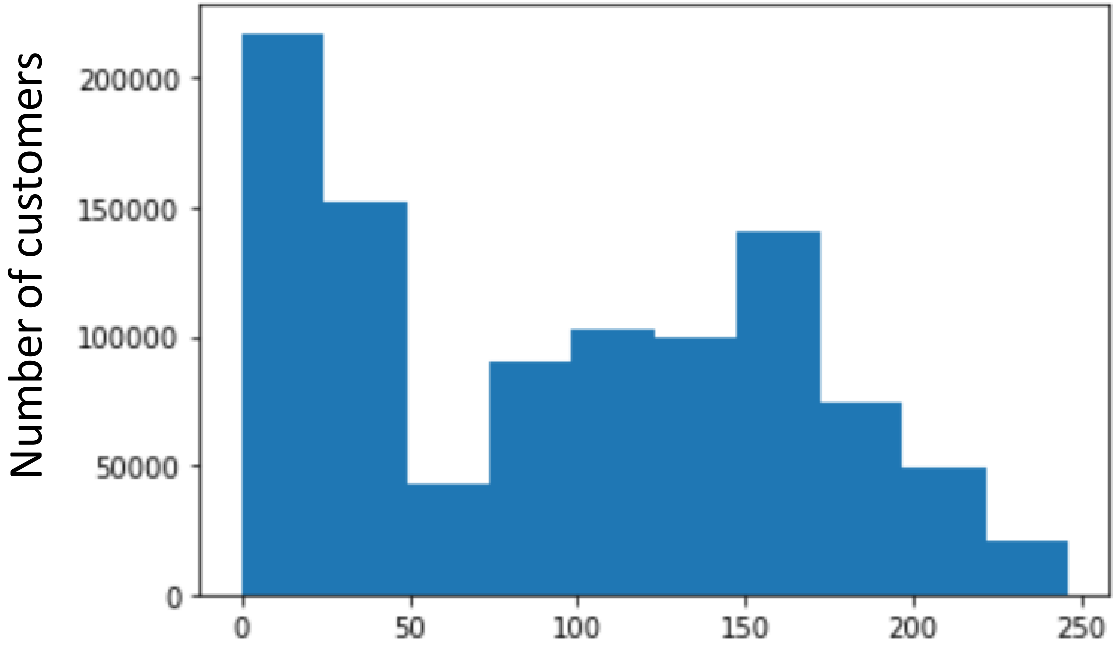
Age distribution



Number of customers

Customer age (years)

Seniority distribution



Number of customers

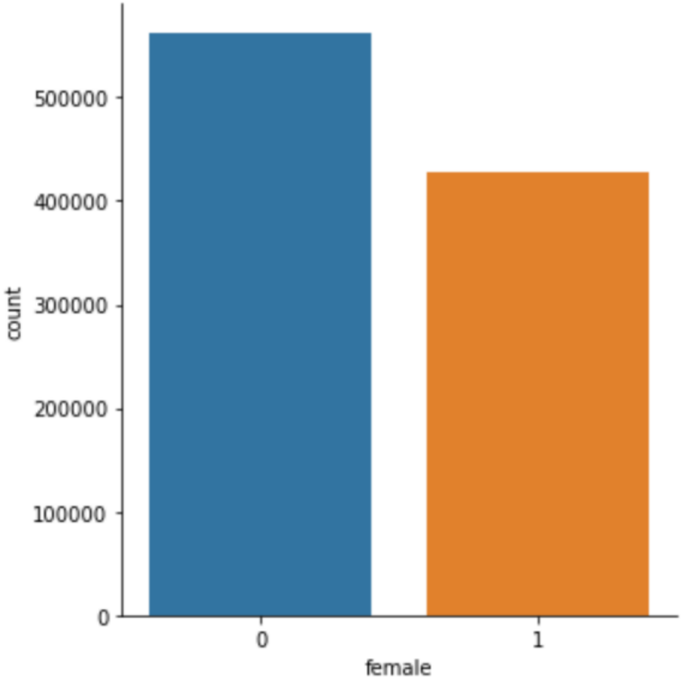
Customer seniority (months)

Data Glacier

Factors most likely to be involved in segmenting customers into groups:

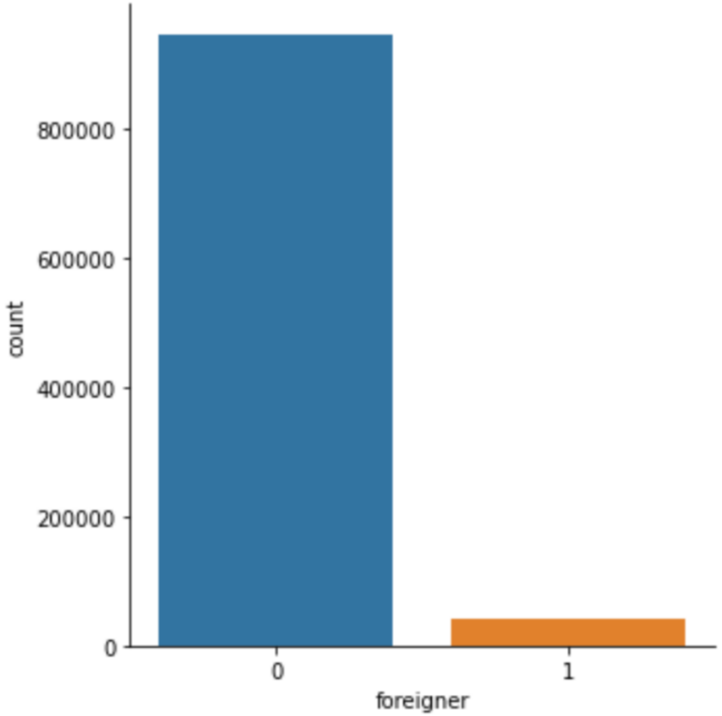
Sex (0: male, 1: female)

Number of customers



Foreigner (0: no, 1: yes)

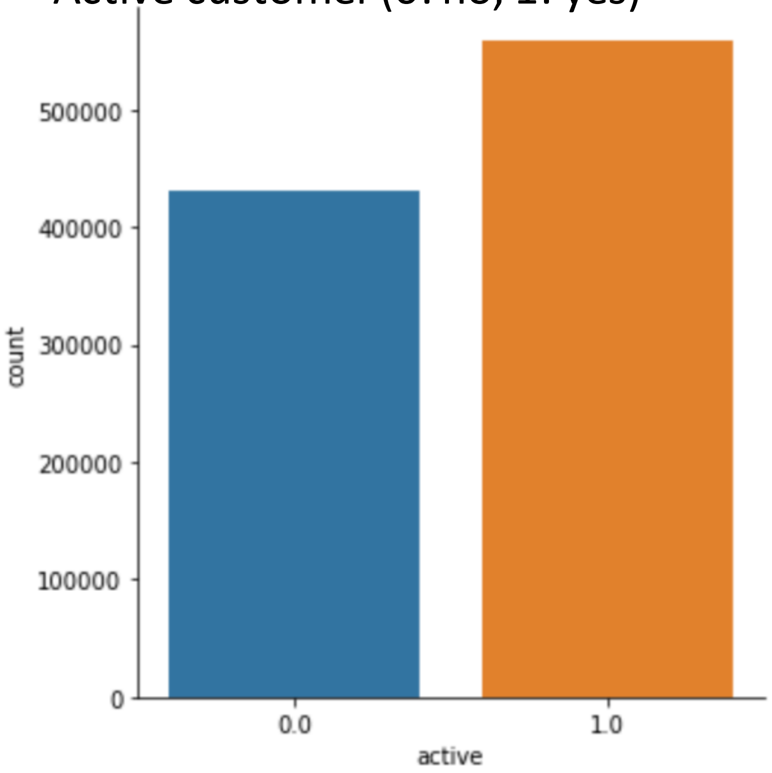
Number of customers



Factors most likely to be involved in segmenting customers into groups (continued):

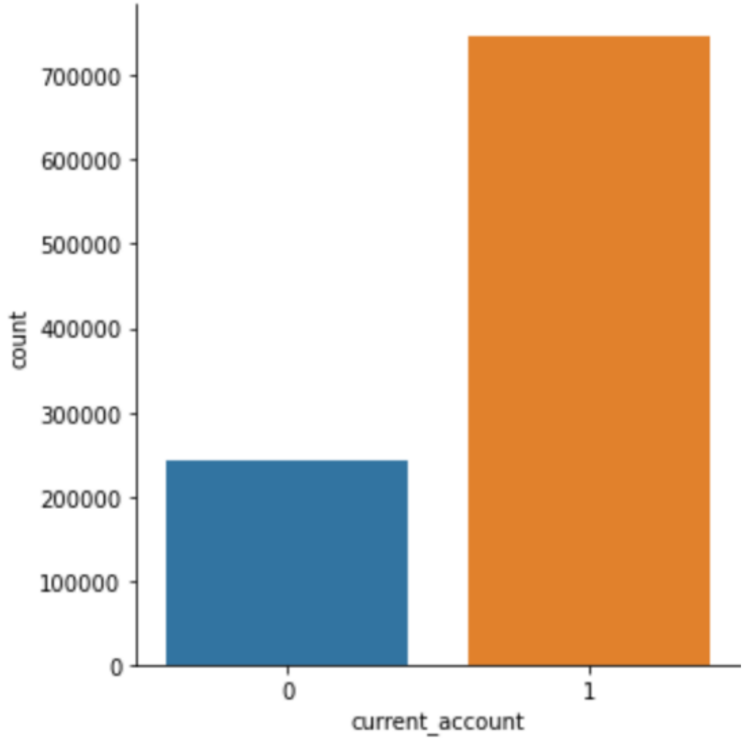
Active customer (0: no, 1: yes)

Number of customers



Customer has a current account (0: no, 1: yes)

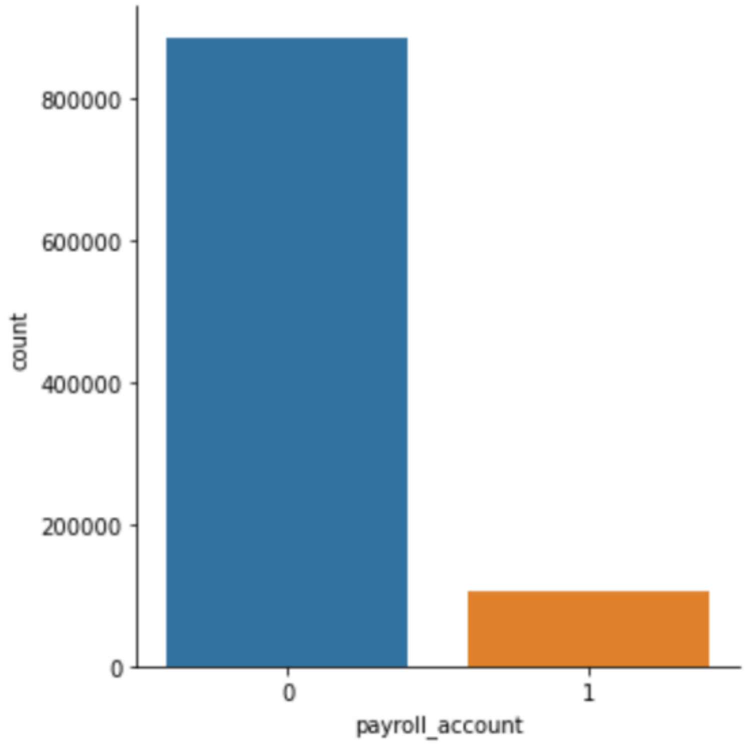
Number of customers



Factors most likely to be involved in segmenting customers into groups (continued):

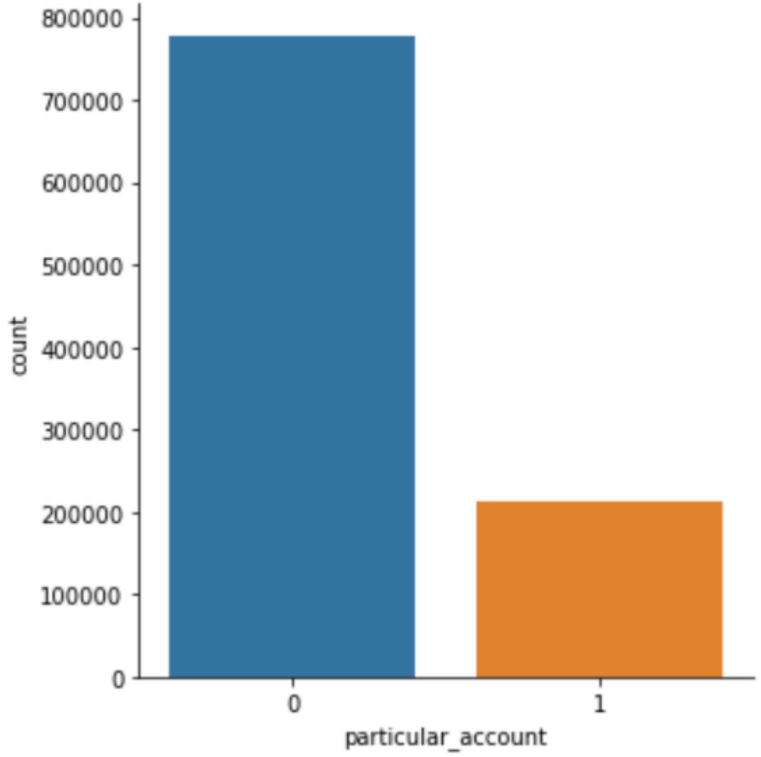
Customer has a payroll account (0: no, 1: yes)

Number of customers



Customer has a particular account (0: no, 1: yes)

Number of customers

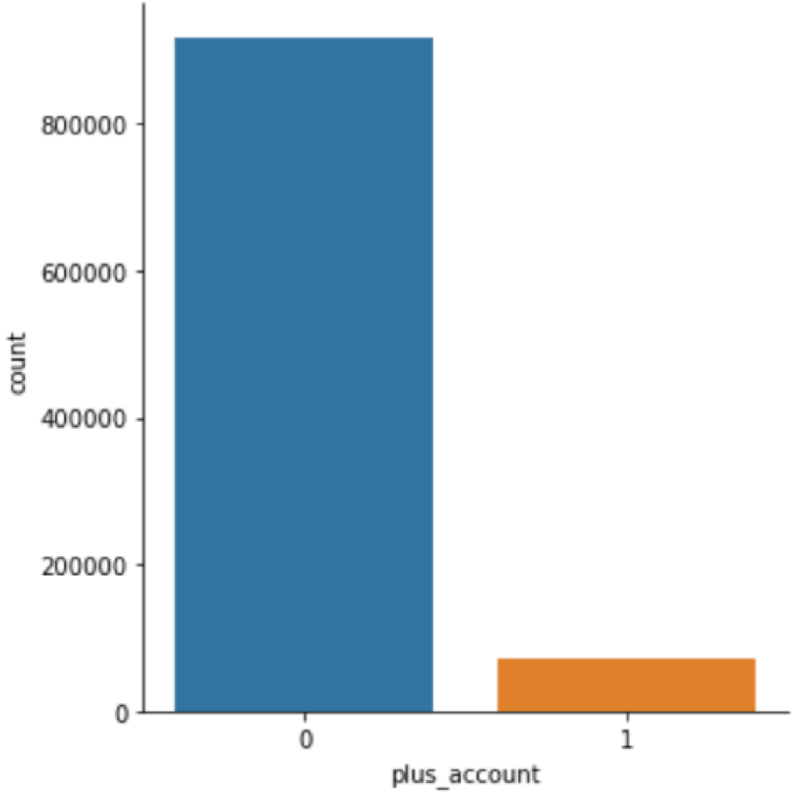


Data Glacier

Factors most likely to be involved in segmenting customers into groups (continued):

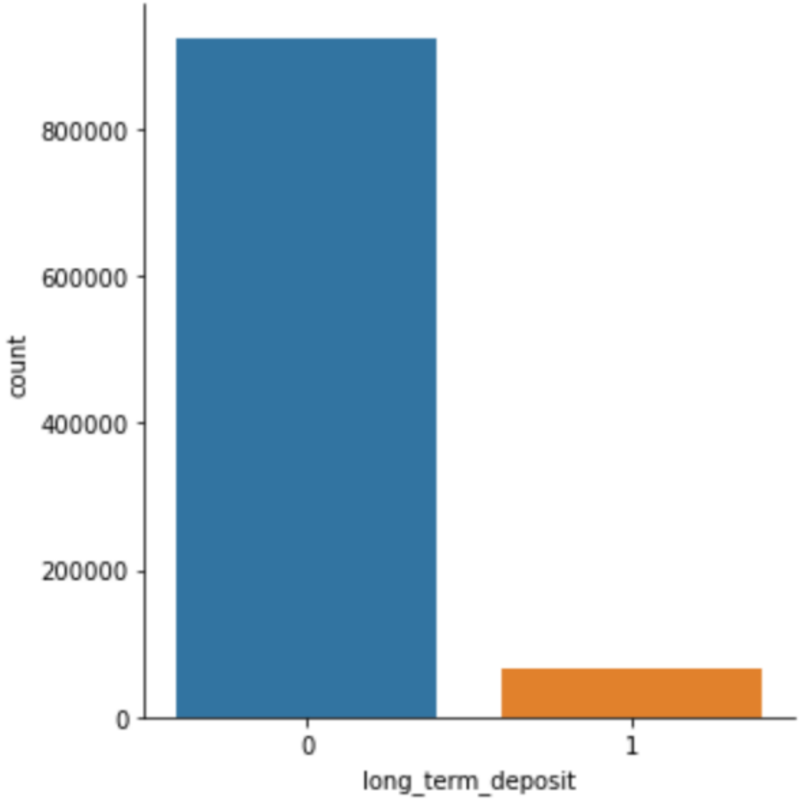
Customer has a plus account (0: no, 1: yes)

Number of customers



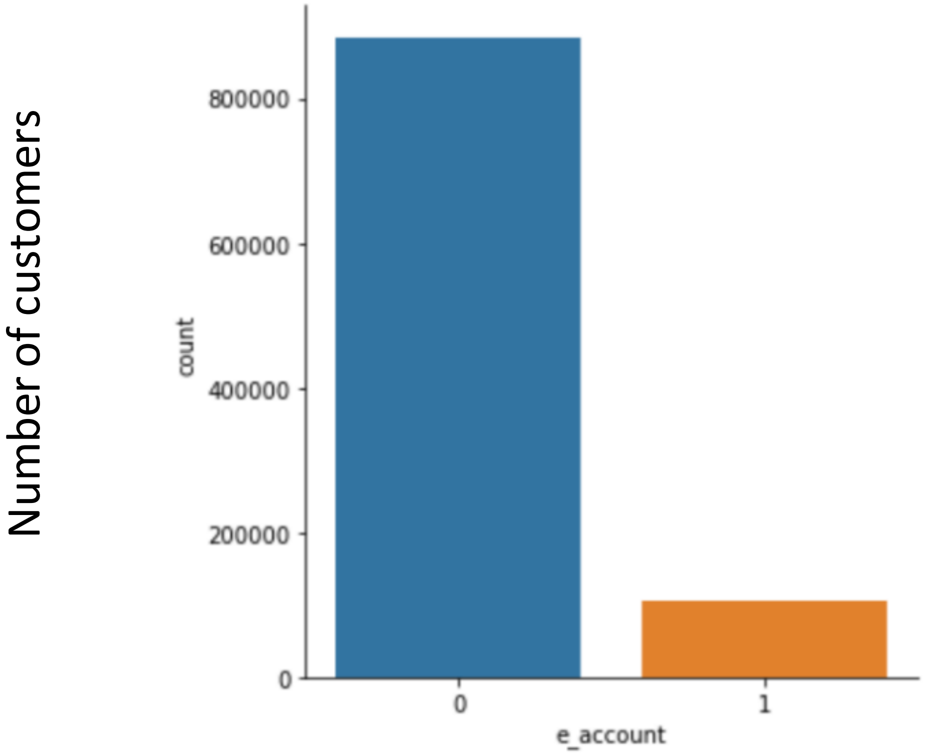
Customer has a long-term deposit account (0: no, 1: yes)

Number of customers

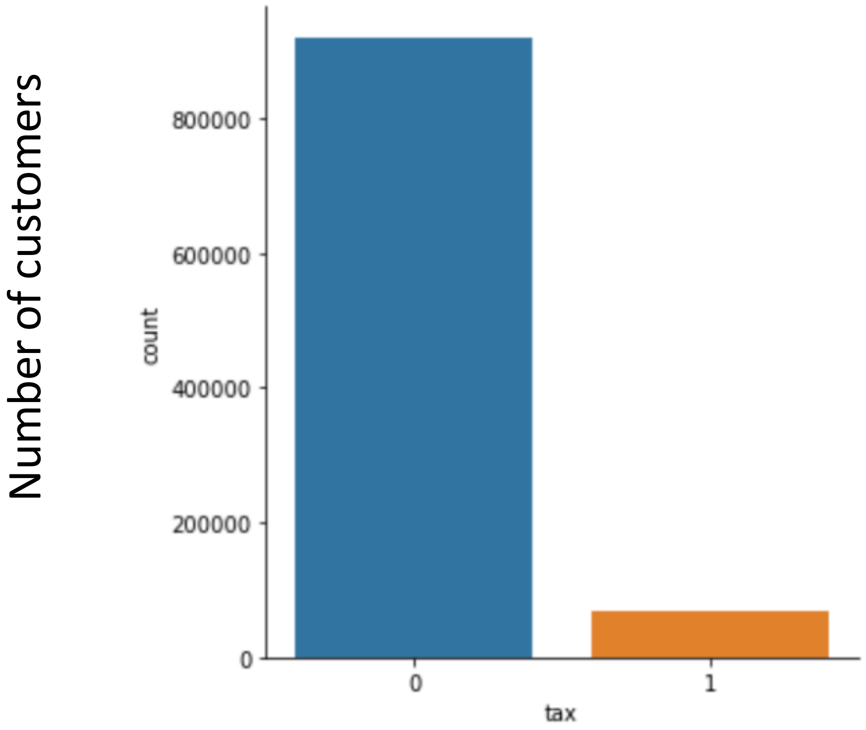


Factors most likely to be involved in segmenting customers into groups (continued):

Customer has an e-account (0: no, 1: yes)



Customer has a tax account (0: no, 1: yes)

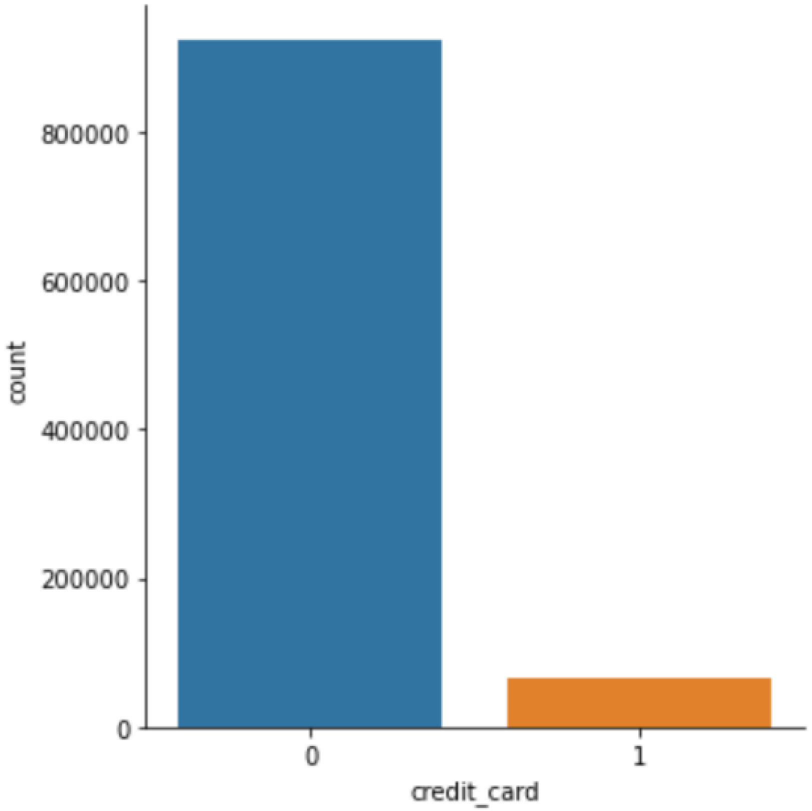


Data Glacier

Factors most likely to be involved in segmenting customers into groups (continued):

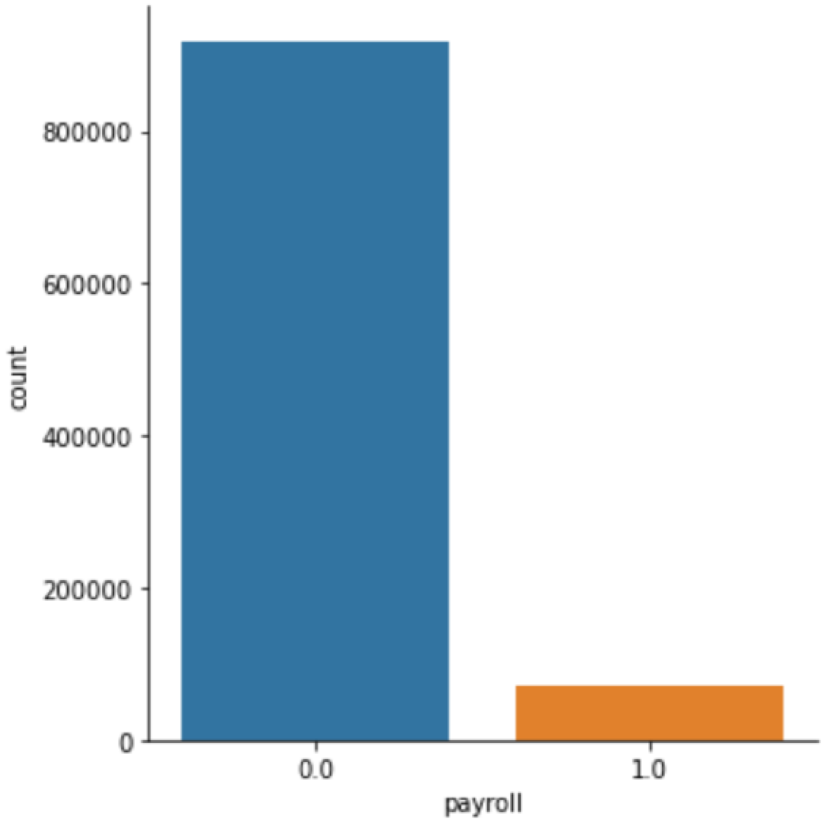
Customer has a credit card account (0: no, 1: yes)

Number of customers



Customer has a payroll account (0: no, 1: yes)

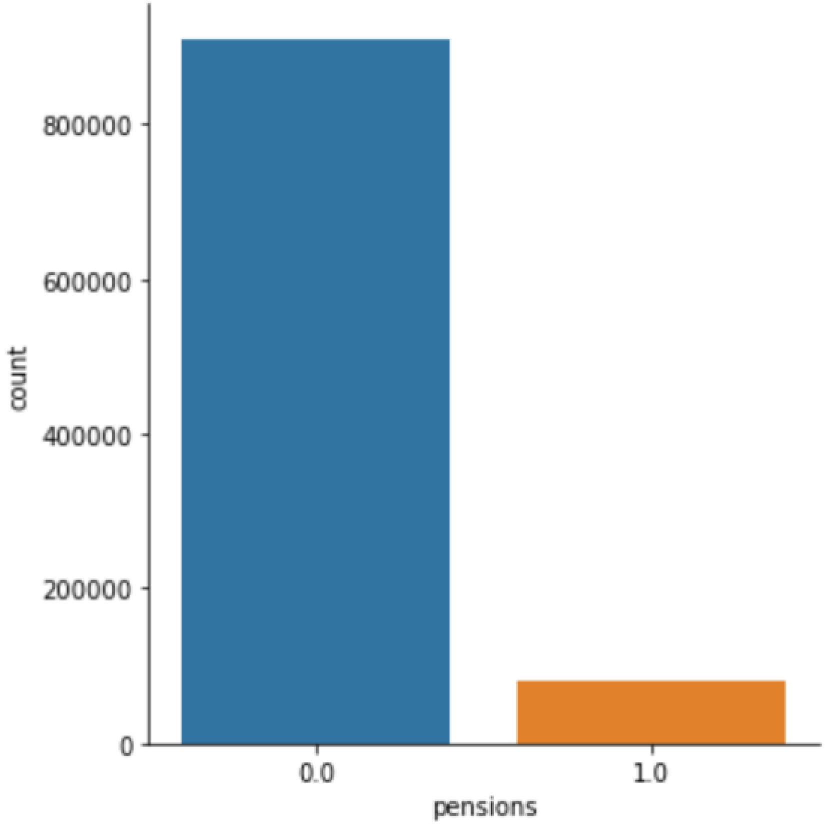
Number of customers



Factors most likely to be involved in segmenting customers into groups (continued):

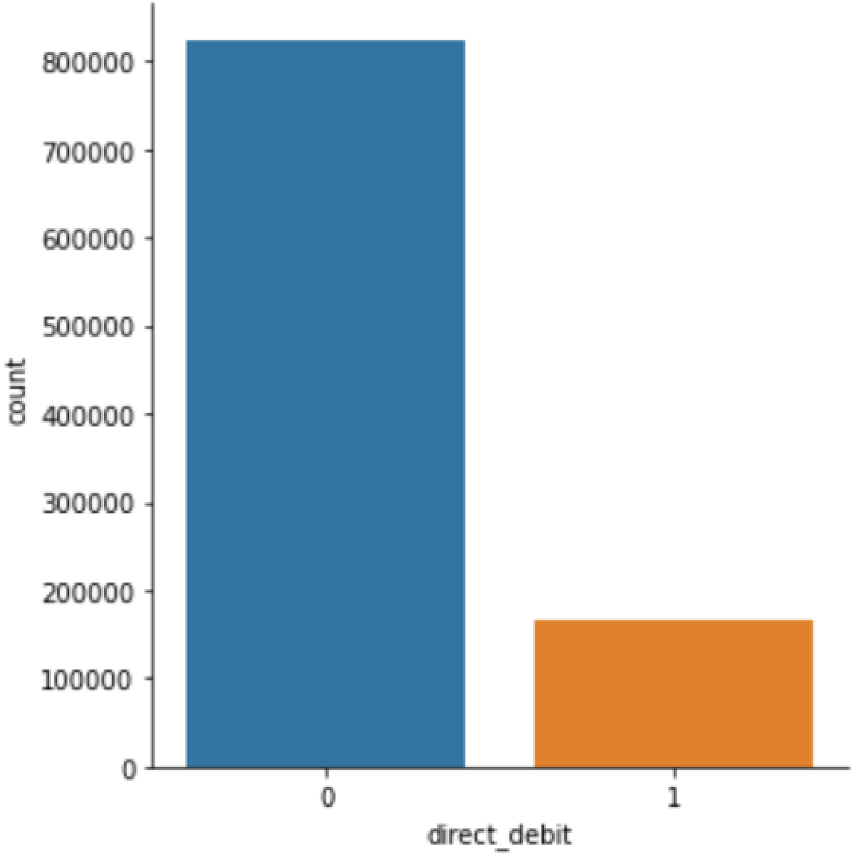
Customer has a pensions account (0: no, 1: yes)

Number of customers

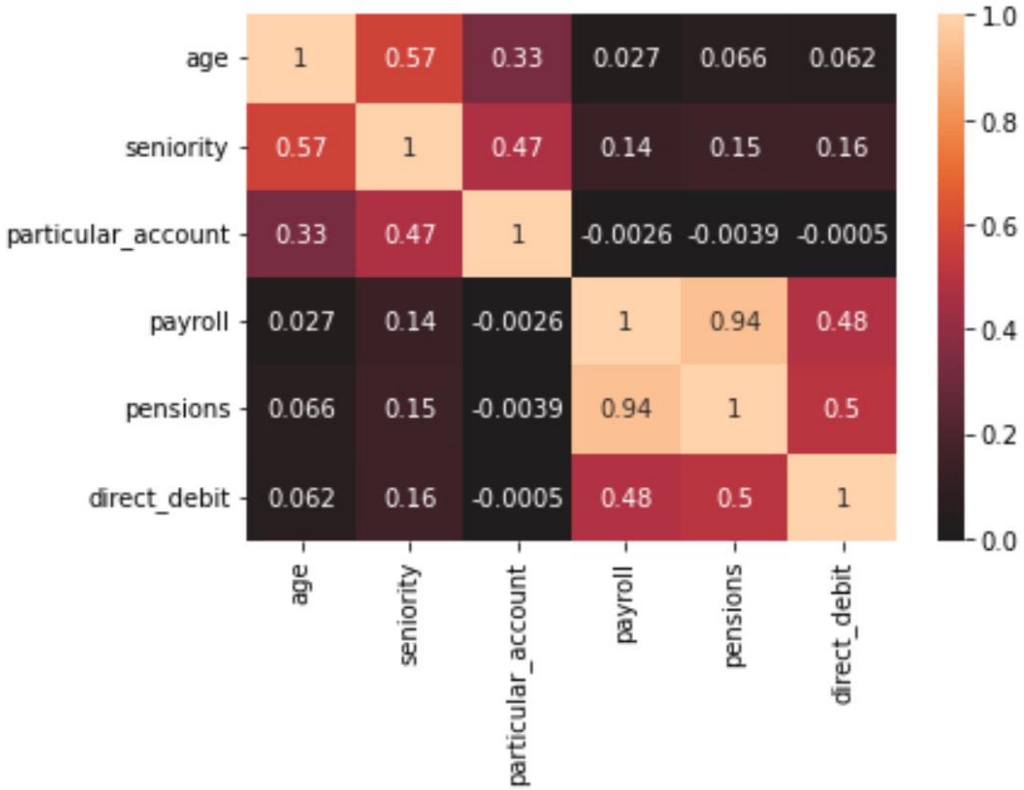


Customer has a direct-debit account (0: no, 1: yes)

Number of customers



Correlations



	age	seniority	particular_account	payroll	pensions	direct_debit
age	1	0.57	0.33	0.027	0.066	0.062
seniority	0.57	1	0.47	0.14	0.15	0.16
particular_account	0.33	0.47	1	-0.0026	-0.0039	-0.0005
payroll	0.027	0.14	-0.0026	1	0.94	0.48
pensions	0.066	0.15	-0.0039	0.94	1	0.5
direct_debit	0.062	0.16	-0.0005	0.48	0.5	1

- Age, seniority, and particular account were correlated. These may be important together in grouping customers.
- Payroll, pensions, and direct debit accounts were also correlated.
- These groups may help distinguish different customer groups. For example, older customers are more likely to have a particular account.

Data Glacier

Summary

- Dataset was cleaned: no more missing or nonsensical values
- Customer age, seniority, and income were all positively skewed and could help distinguish customer groups
- Some categories could also distinguish customer groups, such as gender, active level, and various accounts held
- Age, seniority, and particular account are correlated with each other
- Payroll, pensions, and direct debit are correlated with each other

Recommendations

- Perform k-means clustering with 2, 3, 4, and 5 groups
- Use inertia metric to determine optimal number of groups
- Investigate characteristics of each group to create customer profiles
- Provide groups with characteristics to marketing team to develop individual promotions for each customer segment

Thank You

GitHub repository link:

https://github.com/ebanning/DataGlacierProject

Data Glacier

Your Deep Learning Partner