Team Member Details

Group Name: Elizabeth's Analytics

Name: Elizabeth Banning

Email: estall@hotmail.com

Country: USA

College: Western Governors University

Specialization: Data Science

Problem Description

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

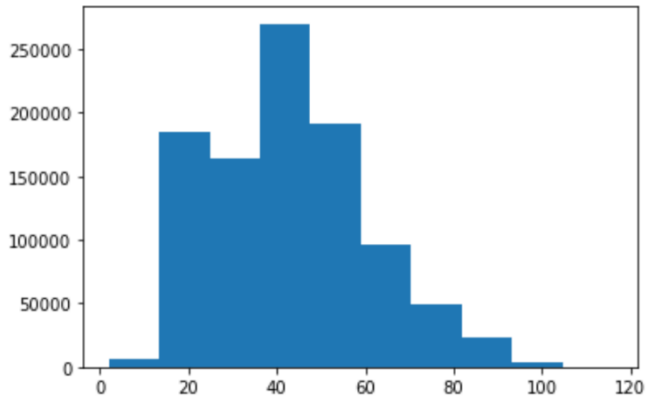
- What is the best number of groups to divide customers into?
- What are the primary characteristics of each group?

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.

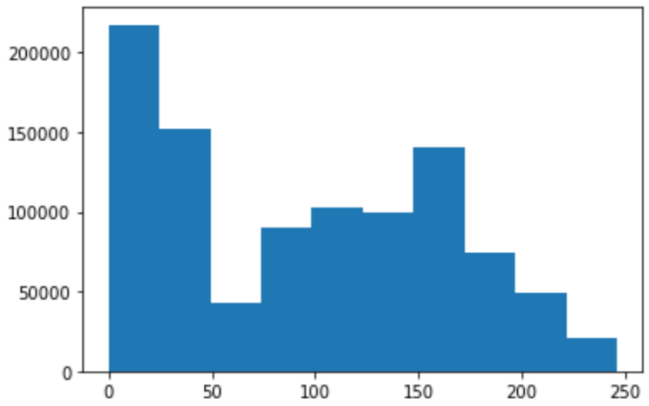
Exploratory Data Analysis

After cleaning the data and dropping the irrelevant columns, three continuous numeric columns remained: age, income, and seniority. The distribution of each of these columns is shown in the graphs below, along with the descriptive statistics listed. Note that the units of income are unknown. Age is assumed to be in years, and seniority represents the number of months the customer has been with the company.

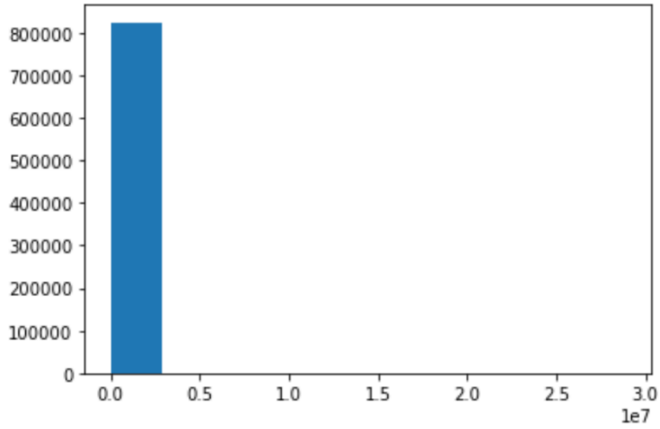
Distribution of age



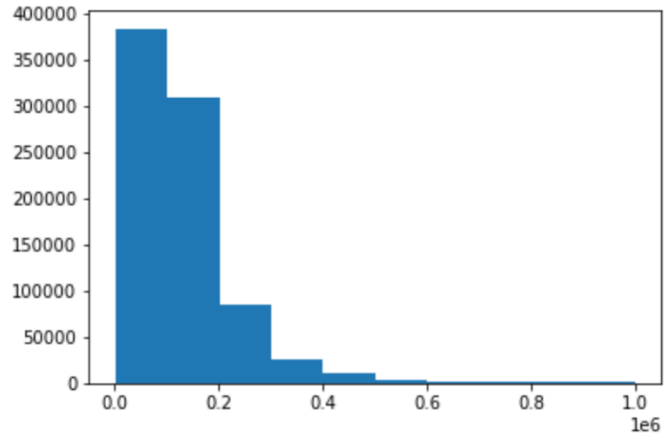
Distribution of seniority



Distribution of income (all values)



Distribution of incomes less than 1,000,000



All three of these variables are positively skewed, with few values at the high end, as expected for each of these. It is anticipated that all three could be involved in clustering customers, for example, a group of customers at the low end of income and another group at the high end. Modeling will determine if this is the case or not.

Descriptive statistics for the 3 continuous columns

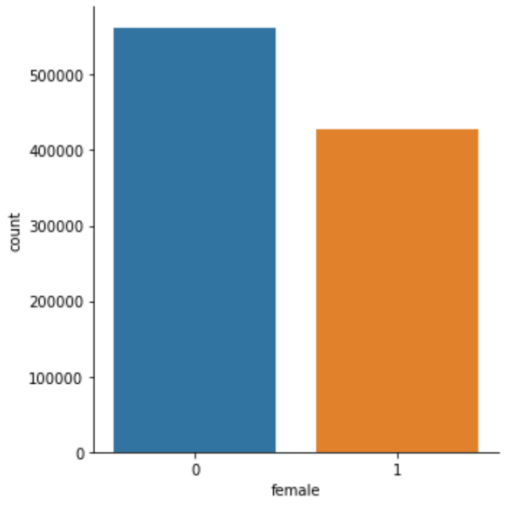
	income	age	seniority
count	9.892180e+05	989218.000000	989218.000000
mean	1.341627e+05	43.269624	97.137569
std	2.185706e+05	17.158355	65.828743
min	1.202730e+03	2.000000	0.000000
25%	7.795836e+04	27.000000	33.000000
50%	1.066519e+05	43.000000	97.000000
75%	1.482880e+05	53.000000	157.000000
max	2.889440e+07	116.000000	246.000000

Examination of categorical variables

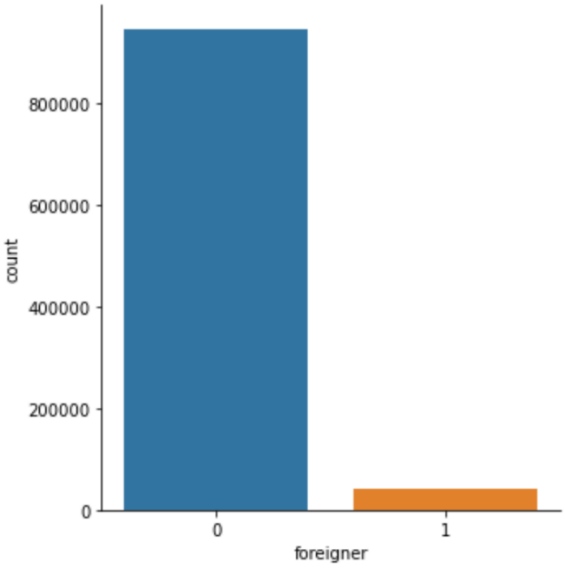
Some categories included relatively large numbers of customers in each group, while for other categories, the vast majority of customers belonged to only one of the groups. The former are more likely to be of use in clustering customers, since customers are more likely to be differentiated across the categories; the latter are less likely to be a factor in the clustering. Bar plots show the distribution of customers in each category.

The following categories have relatively larger numbers of customers in each group and are more likely to be a factor in clustering. Three of these categories (sex, foreigner, and active) are related to the customer's features, and the rest are related to the types of accounts and services held by the customer.

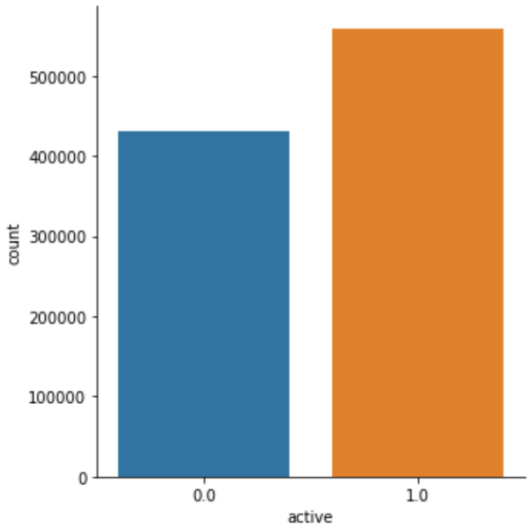
Sex



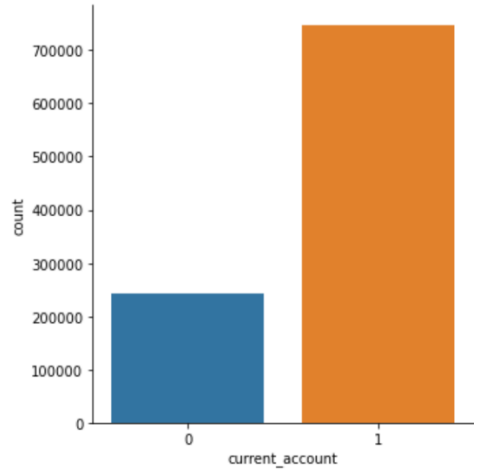
Foreigner



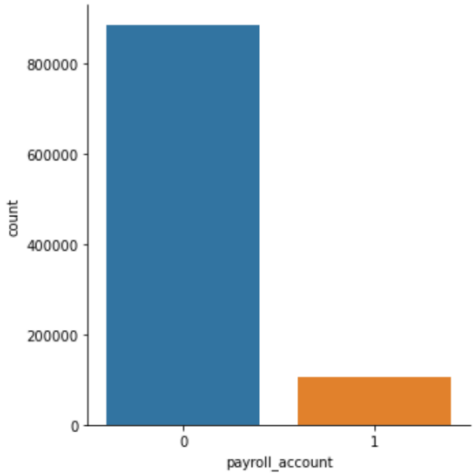
Active



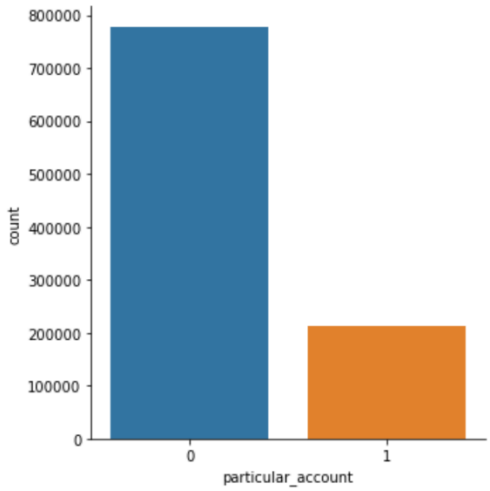
Current account



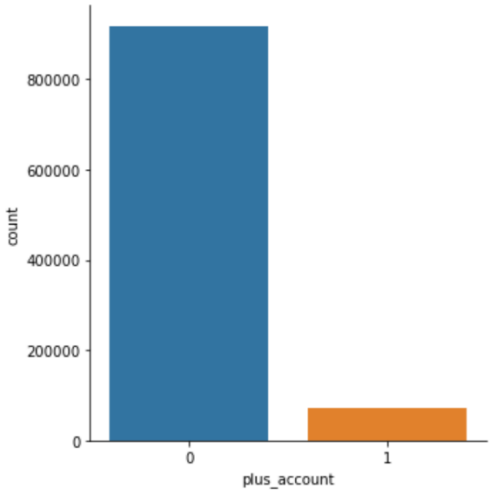
Payroll account



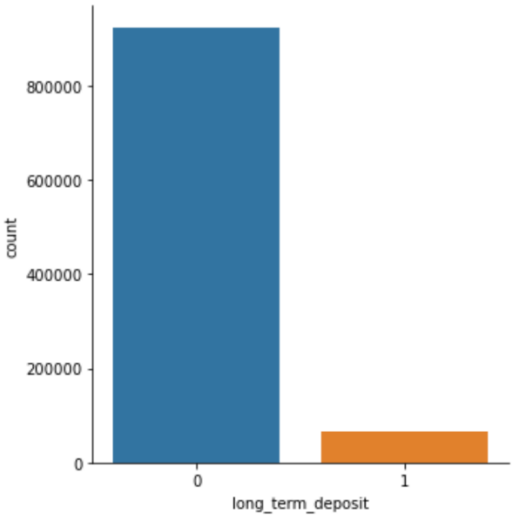
Particular account



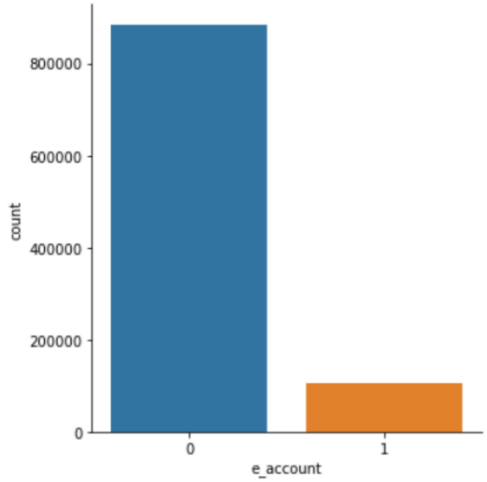
Plus account



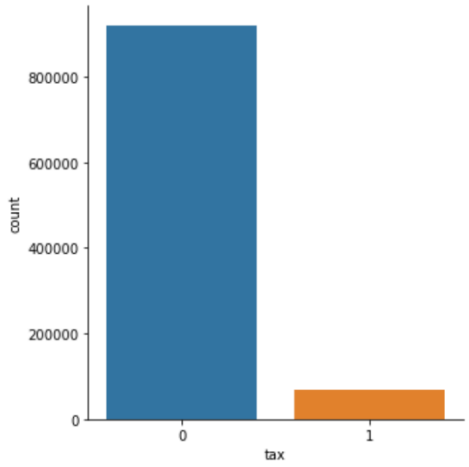
Long-term deposit



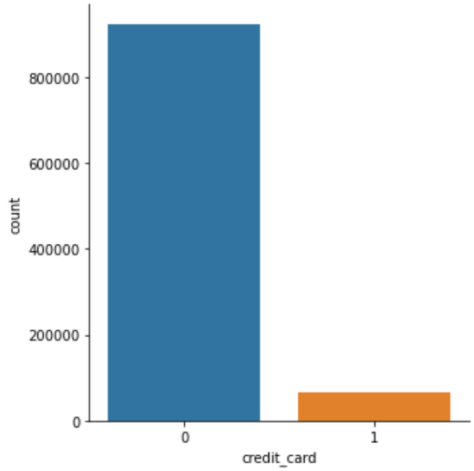
E-account



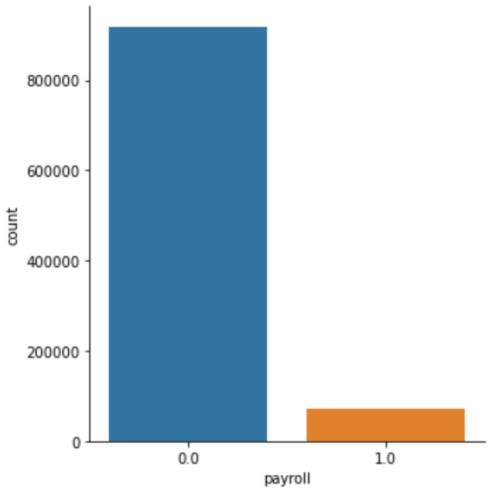
Tax



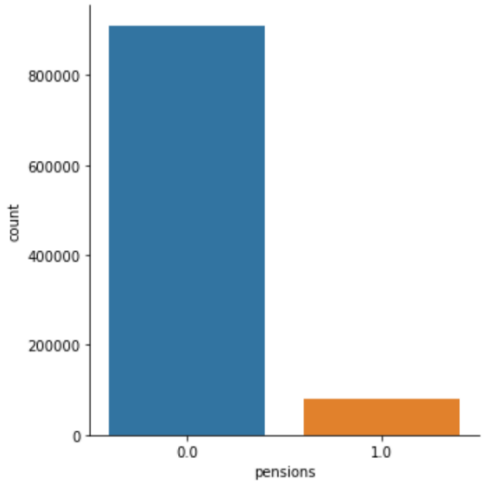
Credit card



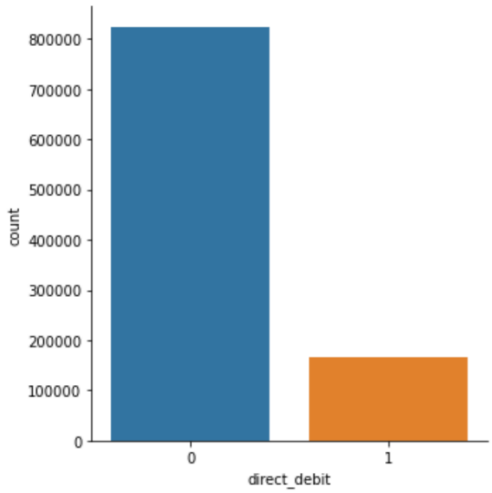
Payroll



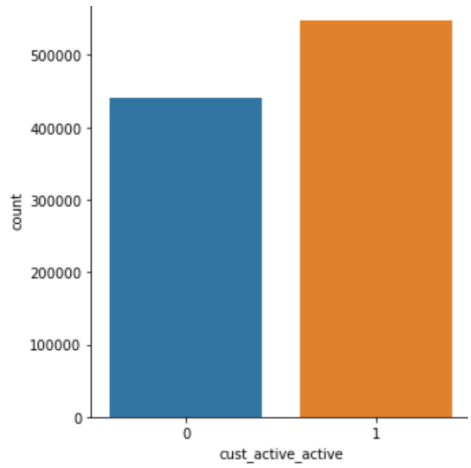
Pensions



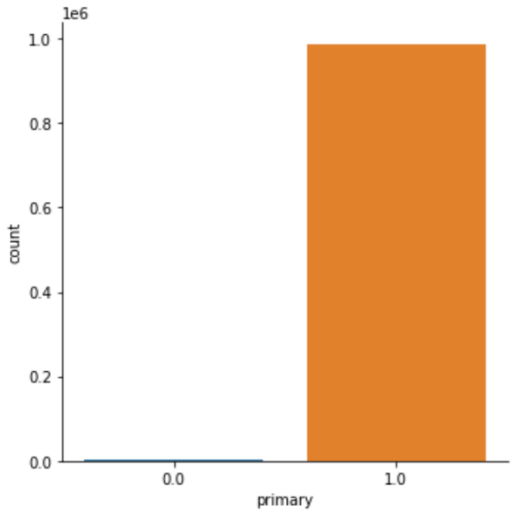
Direct debit



Customer active



The remaining variables are highly unbalanced, with nearly all customers belonging to just one of the categories. An example is the "Primary" variable, with nearly all customers being the primary account holder:



These variables are less likely to be important factors in clustering, since almost all customers belong to just one of the categories.

Heatmap with correlations

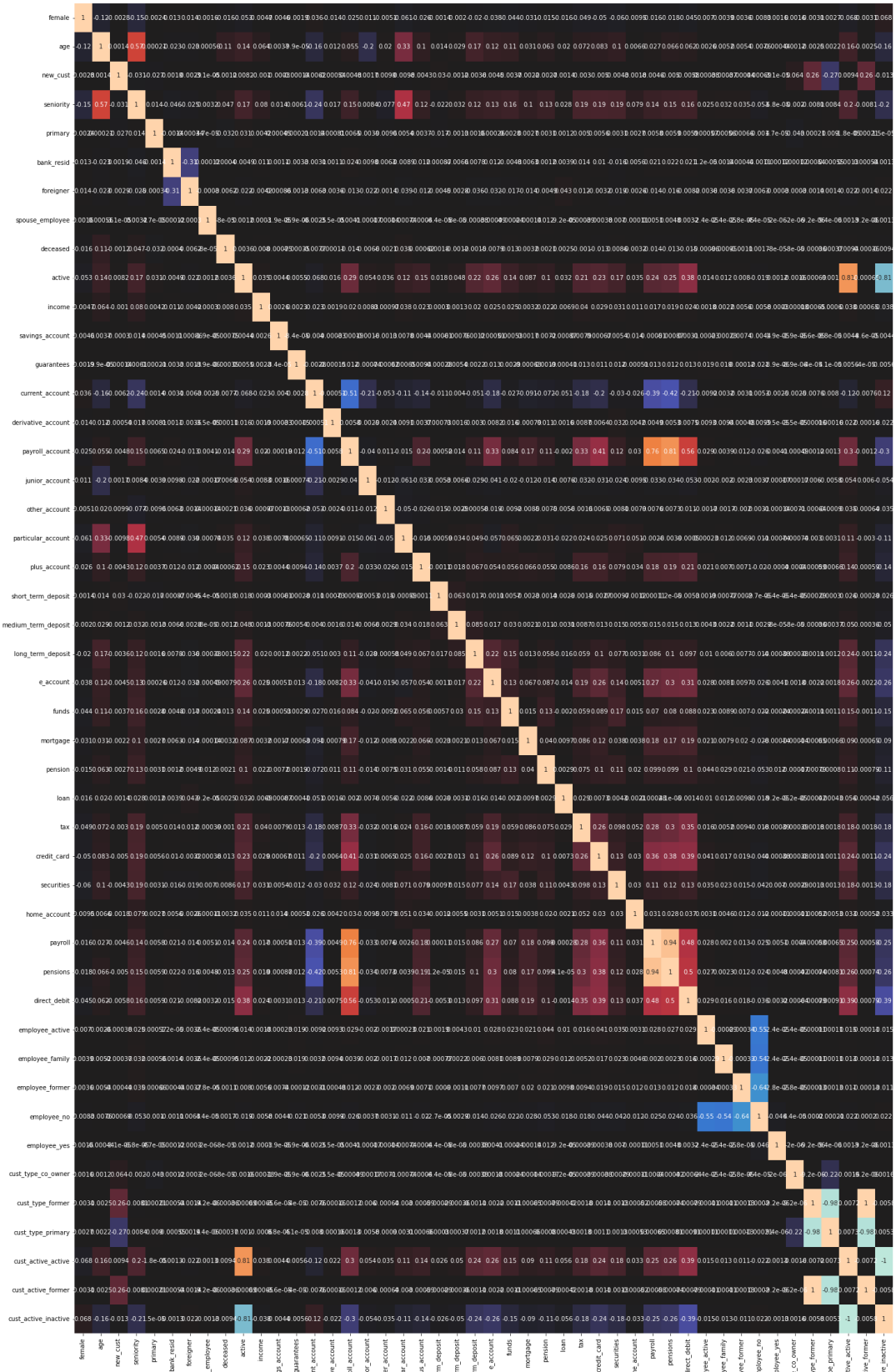
A heatmap was generated that included all variables in the cleaned, scaled dataset to investigate any potential correlations and is shown on the next page. Several interesting correlations were observed. Age is positively correlated with seniority (r = 0.57), suggesting that customers tend to stay with the company as they age. Both are also positively correlated with particular account, suggesting that customers who are older and have been with the company longer are more likely to have a particular account. In addition, payroll, pensions, and direct debit were positively correlated with each other; customers who have one of these services are more likely to have either or both of the others as well. Finally, current account was negatively correlated with payroll, pensions, and direct debit. This suggests the presence of two potential customer groups: those who use business services such as payroll, pensions, and direct debit, and those who are private customers with a current account.

Final Recommendation

Perform k-means clustering with 2-5 groups. Use inertia to determine which number of groups best clusters the customers into similar groups. Investigate the characteristics of each group to create a profile of the typical customer in that group. Provide these groups to the marketing team to develop individualized promotions for each customer segment.

Github Repository Link

ebanning/DataGlacierProject: This is a customer segmentation project for the DataGlacier Data Science virtual internship. (github.com)


female
age
new_cust
seniority
primary
bank_resid
foreigner
spouse_employee
deceased
active
income
savings_account
guarantees
current_account
derivative_account
payroll_account
junior_account
other_account
particular_account
plus_account
short_term_deposit
medium_term_deposit
long_term_deposit
e_account
funds
mortgage
pension
loan
tax
credit_card
securities
home_account
payroll
pensions
direct_debit
employee_active
employee_family
employee_former
employee_no
employee_yes
cust_type_co_owner
cust_type_former
cust_type_primary
cust_active_active
cust_active_former
cust_active_inactive
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75