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.

Data Understanding and Types

The dataset consists of 1,000,000 observations across 48 features. The features include customer demographic information as well as details on the services the customer uses at the bank. The features, their datatypes, sample values, and the number of missing values are listed in the following table. Feature names were renamed to meaningful English names. The original Spanish names are listed in parentheses.

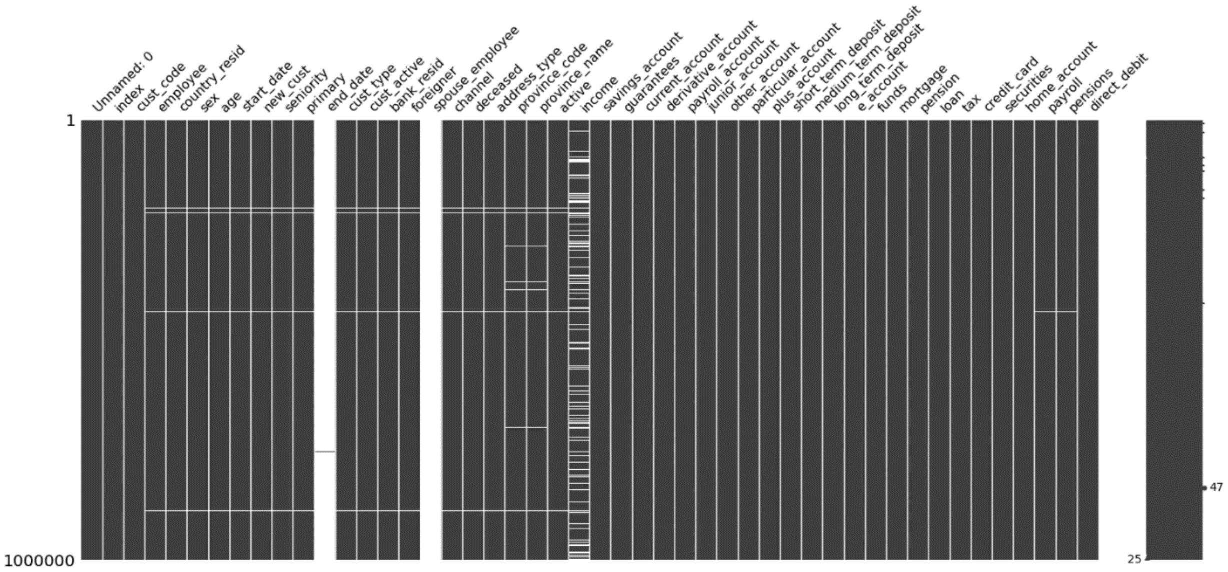
Feature Name	Feature Type	Datatype	Sample Values	Number of Missing Values
Unnamed: 0	index	integer	1, 999999	0
index (fecha_dato)	date	string	2015-01-28, 2015-02-28	0
cust_code (ncodpers)	numeric	integer	15889, 1379131	0
employee (indempleado)	categorical	string	N, S	10,782
country_resid (pais_residencia)	categorical	string	ES, AL	10,782
sex (sexo)	dichotomous categorical	string	V, H	10,786
age (age)	numeric	string	2, 116	10,782

start_date (fecha_alta)	date	string	1995-01-16, 2015-02-27	10,782
new_cust (ind_nuevo)	dichotomous categorical	float	0.0, 1.0	10,782
seniority (antiguedad)	numeric	string	-999999, 246	10,782
primary (indrel)	dichotomous categorical	float	1.0, 99.0	10,782
end_date (ult_fec_cli_1t)	date	string	2015-07-01, 2015-07-30	998,899
cust_type (indrel_1mes)	categorical	integer	1.0, 2.0	10,782
cust_active (tiprel_1mes)	categorical	string	A, I	10,782
bank_resid (indresi)	dichotomous categorical	string	S, N	10,782
foreigner (indext)	dichotomous categorical	string	S, N	10,782
spouse_employee (conyuemp)	dichotomous categorical	string	S, N	999,822
channel (canal_entrada)	categorical	string	KAT, KGC	10,861
deceased (indfall)	dichotomous categorical	string	S, N	10,782
address_type (tipodom)	categorical	float	1.0	10,782
province_code (cod_prov)	categorical	integer	1.0, 52.0	17,734
province_name (nom_prov)	categorical	string	AVILA, MADRID	17,734
active (ind_actividad_cliente)	dichotomous categorical	integer	0.0, 1.0	10,782
income (renta)	continuous numeric	float	1202.73, 28894395.51	175,183
savings_account (ind_ahor_fin_ult1)	dichotomous categorical	integer	0, 1	0
guarantees (ind_aval_fin_ult1)	dichotomous categorical	integer	0, 1	0
current_account (ind_cco_fin_ult1)	dichotomous categorical	integer	0, 1	0
derivative_account (ind_cder_fin_ult1)	dichotomous categorical	integer	0, 1	0
payroll_account (ind_cno_fin_ult1)	dichotomous categorical	integer	0, 1	0
junior_account (ind_ctju_fin_ult1)	dichotomous categorical	integer	0, 1	0

other_account (ind_ctma_fin_ult1)	dichotomous categorical	integer	0, 1	0
particular_account (ind_ctop_fin_ult1)	dichotomous categorical	integer	0, 1	0
plus_account (ind_ctpp_fin_ult1)	dichotomous categorical	integer	0, 1	0
short_term_deposit (ind_deco_fin_ult1)	dichotomous categorical	integer	0, 1	0
medium_term_deposit (ind_deme_fin_ult1)	dichotomous categorical	integer	0, 1	0
long_term_deposit (ind_dela_fin_ult1)	dichotomous categorical	integer	0, 1	0
e_account (ind_ecue_fin_ult1)	dichotomous categorical	integer	0, 1	0
funds (ind_fond_fin_ult1)	dichotomous categorical	integer	0, 1	0
mortgage (ind_hip_fin_ult1)	dichotomous categorical	integer	0, 1	0
pension (ind_pan_fin_ult1)	dichotomous categorical	integer	0, 1	0
loan (ind_pres_fin_ult1)	dichotomous categorical	integer	0, 1	0
tax (ind_reca_fin_ult1)	dichotomous categorical	integer	0, 1	0
credit_card (ind_tjcr_fin_ult1)	dichotomous categorical	integer	0, 1	0
securities (ind_valo_fin_ult1)	dichotomous categorical	integer	0, 1	0
home_account (ind_viv_fin_ult1)	dichotomous categorical	integer	0, 1	0
payroll (ind_nomina_ult1)	dichotomous categorical	integer	0, 1	5,402
pensions (ind_nom_pens_ult1)	dichotomous categorical	integer	0, 1	5,402
direct_debit (ind_recibo_ult1)	dichotomous categorical	integer	0, 1	0

Missing Data

It is notable that exactly 10,782 observations are missing from 14 features. The missing matrix suggests that these observations are all missing from the same rows:



These rows contain account information but not customer demographic information. Examining these rows further, there does not appear to be a pattern between types of accounts and the missingness of the data. The 10,782 rows account for only about 1.08% of the total data and do not provide useful customer demographic information for the purpose of clustering. Therefore, these rows were dropped from the dataset.

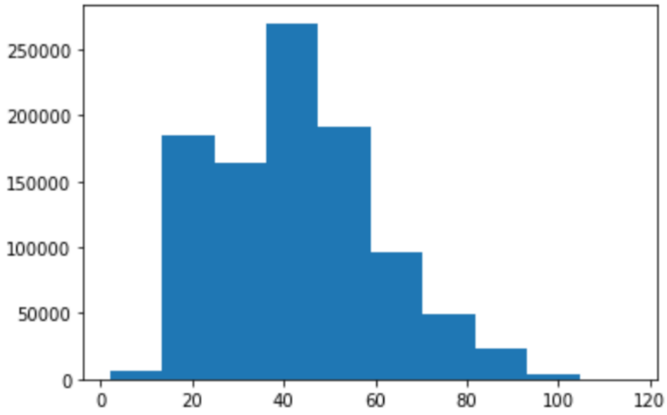
After these rows were dropped, some missing values remained. These were treated differently depending on the column as follows:

- sex (4 missing): Impute with the mode (male). Since this is a dichotomous category, the column name was changed to Female and values were changed: all male values (V) were changed to 0 and all female values (H) were changed to 1.
- end_date (988,117 missing): The column was dropped. It does not provide useful information for grouping the customers.
- spouse_employee (989,040 missing): The missing values were imputed as N (not spouse). Assume that the number of spouses of employees is low and that the missing values should be interpreted as non-spouses. Since this is a dichotomous category, the values were changed from N to 0 and from S to 1.
- channel (79 missing): The column was dropped. It does not provide useful information for grouping the customers due to the high number of channel values (156 different channels).
- province_code and province_name (6,952 missing values in each): The columns were dropped. They do not provide useful information for grouping the customers due to the high number of provinces (52 different provinces).
- income (164,401 missing): The income column is positively skewed (as is typical for incomes, with few people at the very high end). Therefore, missing values were replaced with the median (106,651.86).
- payroll (100 missing): Imputed with the mode (0).
- pensions (100 missing): Imputed with the mode (0).

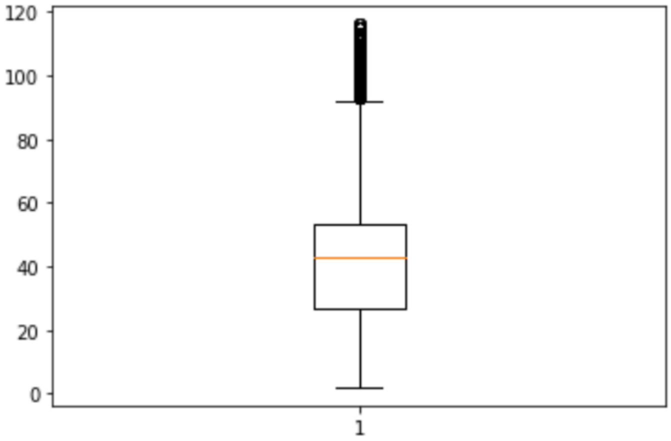
Distribution, Outliers, and Skew

The age column

The age column appears to be positively skewed:



There are outliers on the high end:

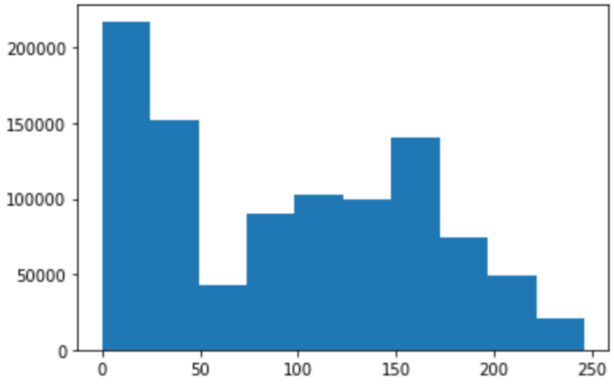


There is no reason to suspect that the outliers are unreasonable (for example, outside the expected age range for people) or otherwise incorrect. They were retained.

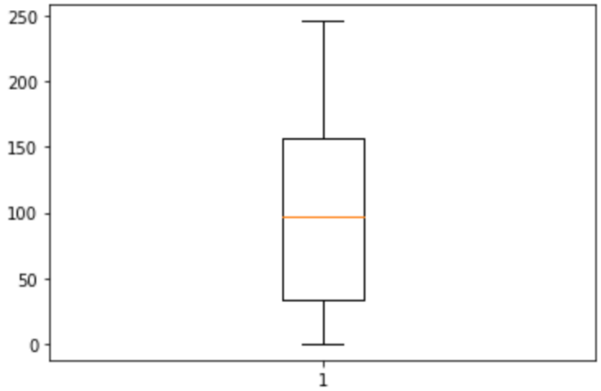
The seniority column

The seniority column contains many negative values, which do not make sense. The column is intended to represent the number of months the customer has been with the company. Because negative numbers do not make sense, any negative values in this column were changed to equal zero.

After this, the data appear to be positively skewed:

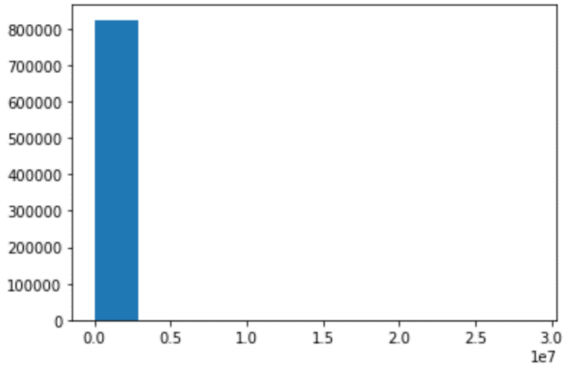


There are no outliers:

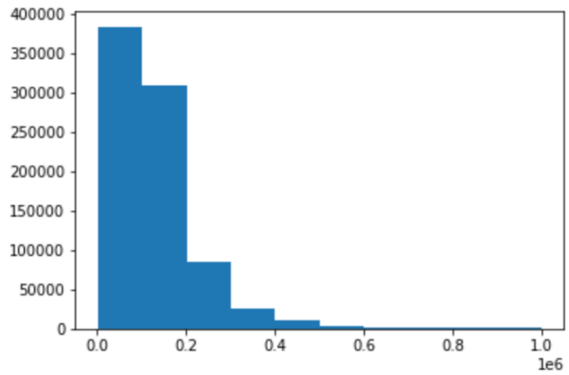


The income column

The distribution of the income column is positively skewed, which is typical of incomes, with very few values at the very high end:

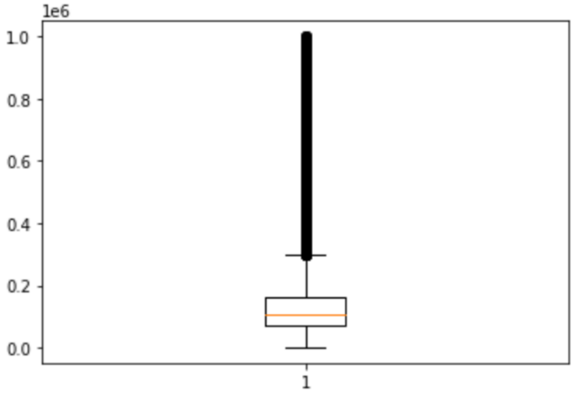


Excluding the 166,499 values above 10,000,000 (note that units of income are unknown) gives a clearer sense of the distribution:



Because of the skew, the missing values were imputed with the median, not the mean.

There are many significant outliers at the high end. The boxplot shows those for incomes less than 10,000,000:



There is no indication that the outliers are mistakes in the data, so they were retained.

Deletion of Some Columns

Certain columns were redundant or not of use in this analysis for reasons given below. Therefore, these columns were deleted:

- country_resid: This column indicates the country of residence of the customer. There are too many values (113 countries) to be of use in the analysis.
- start_date: This column is redundant. The seniority of the customer is already indicated by the seniority column.
- address_type: All values in this column are 1 (primary residence). It does not provide any useful information to distinguish different customers.

Re-expression of Categorical Values

- There were 5 values in the employee column. These were replaced with meaningful words to describe the values: A: active, B: former, N: no, S: yes, F: family. This column was then one-hot encoded.

- In the primary column, the values were 1 (primary) and 99 (not primary at the end of the month). To maintain consistency with the rest of the data, values equal to 99 were replaced with zero.
- The values in the customer type column were meaningless numbers and were replaced with meaningful words: 1: primary, 2: co-owner, 3: former. The column was then one-hot encoded.
- The "customer active" column had 3 values and these were replaced with meaningful words to describe the values: A: active, I: inactive, P: former. The column was then one-hot encoded.
- The "bank residence" column values were changed from S (yes) to 1 and from N (no) to zero.
- The foreigner column values were changed from S (yes) to 1 and from N (no) to zero.
- The deceased column values were changed from S (yes) to 1 and from N (no) to zero.

Scaling Data

Because clustering will be used to group the customers, and because clustering uses distance measurements, the variables were scaled using scikit-learn's StandardScaler function.

After these preprocessing and cleaning steps, the dataset is now ready for modeling.

Github Repository Link

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