

Data Glacier Data Scientist Internship

Batch: LISUM23: LISUM39

Week8: Deliverables

Project: Bank Customer Segmentation

Group name: Apple Analytics

Name: Madoka Fujii

Email address: mdkfji@gmail.com

Country: United States

Company: Omdena

Specialization: Data Analytics

Problem Description:

XYZ Bank plans to enhance its marketing campaign as Christmas offers for its customers. However, instead of offering the same deal to all customers as generic, the bank wants to provide personalized offers to specific customer groups to fit their preferences. Identifying customer categories manually would be inefficient and fail to uncover hidden patterns in the data that could inform better segmentation. To address this, the bank has sought the assistance of ABC Analytics. Additionally, the bank has specified that customer segmentation should result in no more than 5 groups to ensure the campaign's efficiency.

Data Understanding:

The dataset includes various information for the bank customers. For segmentation analysis, we need to identify the specific groups with specific characteristics. To find the uniquenesses, unsupervised learning algorithms such as clustering are the best way to analyze.

What type of data have you got for analysis?

-The shape: 1000000 rows × 48 columns

Column Name	Description
fecha_datos	The table is partitioned for this column
ncodpers	Customer code
ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive
pais_residencia	Customer's Country residence
sexo	Customer's sex
age	Age
fecha_alta	The date in which the customer became as the first holder of a contract in the bank
ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.
antiguedad	Customer seniority (in months)
indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)
ult_fec_cli_1t	Last date as primary customer (if he isn't at the end of the month)
indrel_1mes	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)
tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)
conyuemp	Spouse index. 1 if the customer is spouse of an employee
canal_entrada	channel used by the customer to join
indfall	Deceased index. N/S

tipodom	Addres type. 1, primary address
cod_prov	Province code (customer's address)
nomprov	Province name
ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
renta	Gross income of the household
ind_ahor_fin_ult1	Saving Account
ind_aval_fin_ult1	Guarantees
ind_cco_fin_ult1	Current Accounts
ind_cder_fin_ult1	Derivada Account
ind_cno_fin_ult1	Payroll Account
ind_ctju_fin_ult1	Junior Account
ind_ctma_fin_ult1	Más particular Account
ind_ctop_fin_ult1	particular Account
ind_ctpp_fin_ult1	particular Plus Account
ind_deco_fin_ult1	Short-term deposits
ind_deme_fin_ult1	Medium-term deposits
ind_dela_fin_ult1	Long-term deposits
ind_ecue_fin_ult1	e-account
ind_fond_fin_ult1	Funds
ind_hip_fin_ult1	Mortgage
ind_plan_fin_ult1	Pensions
ind_pres_fin_ult1	Loans

ind_reca_fin_ult1	Taxes
ind_tjcr_fin_ult1	Credit Card
ind_valo_fin_ult1	Securities
ind_viv_fin_ult1	Home Account
ind_nomina_ult1	Payroll
ind_nom_pens_ult1	Pensions
ind_recibo_ult1	Direct Debit

-Data type in original

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	1000000 non-null	int64
1	fecha_dato	1000000 non-null	object
2	ncodpers	1000000 non-null	int64
3	ind_empleado	989218 non-null	object
4	pais_residencia	989218 non-null	object
5	sexo	989214 non-null	object
6	age	1000000 non-null	object
7	fecha_alta	989218 non-null	object
8	ind_nuevo	989218 non-null	float64
9	antiguedad	1000000 non-null	object
10	indrel	989218 non-null	float64
11	ult_fec_cli_1t	1101 non-null	object
12	indrel_1mes	989218 non-null	float64

13	tiprel_1mes	989218	non-null	object
14	indresi	989218	non-null	object
15	indext	989218	non-null	object
16	conyuemp	178	non-null	object
17	canal_entrada	989139	non-null	object
18	indfall	989218	non-null	object
19	tipodom	989218	non-null	float64
20	cod_prov	982266	non-null	float64
21	nomprov	982266	non-null	object
22	ind_actividad_cliente	989218	non-null	float64
23	renta	824817	non-null	float64
24	ind_ahor_fin_ult1	1000000	non-null	int64
25	ind_aval_fin_ult1	1000000	non-null	int64
26	ind_cco_fin_ult1	1000000	non-null	int64
27	ind_cder_fin_ult1	1000000	non-null	int64
28	ind_cno_fin_ult1	1000000	non-null	int64
29	ind_ctju_fin_ult1	1000000	non-null	int64
30	ind_ctma_fin_ult1	1000000	non-null	int64
31	ind_ctop_fin_ult1	1000000	non-null	int64
32	ind_ctpp_fin_ult1	1000000	non-null	int64
33	ind_deco_fin_ult1	1000000	non-null	int64
34	ind_deme_fin_ult1	1000000	non-null	int64
35	ind_dela_fin_ult1	1000000	non-null	int64
36	ind_ecue_fin_ult1	1000000	non-null	int64

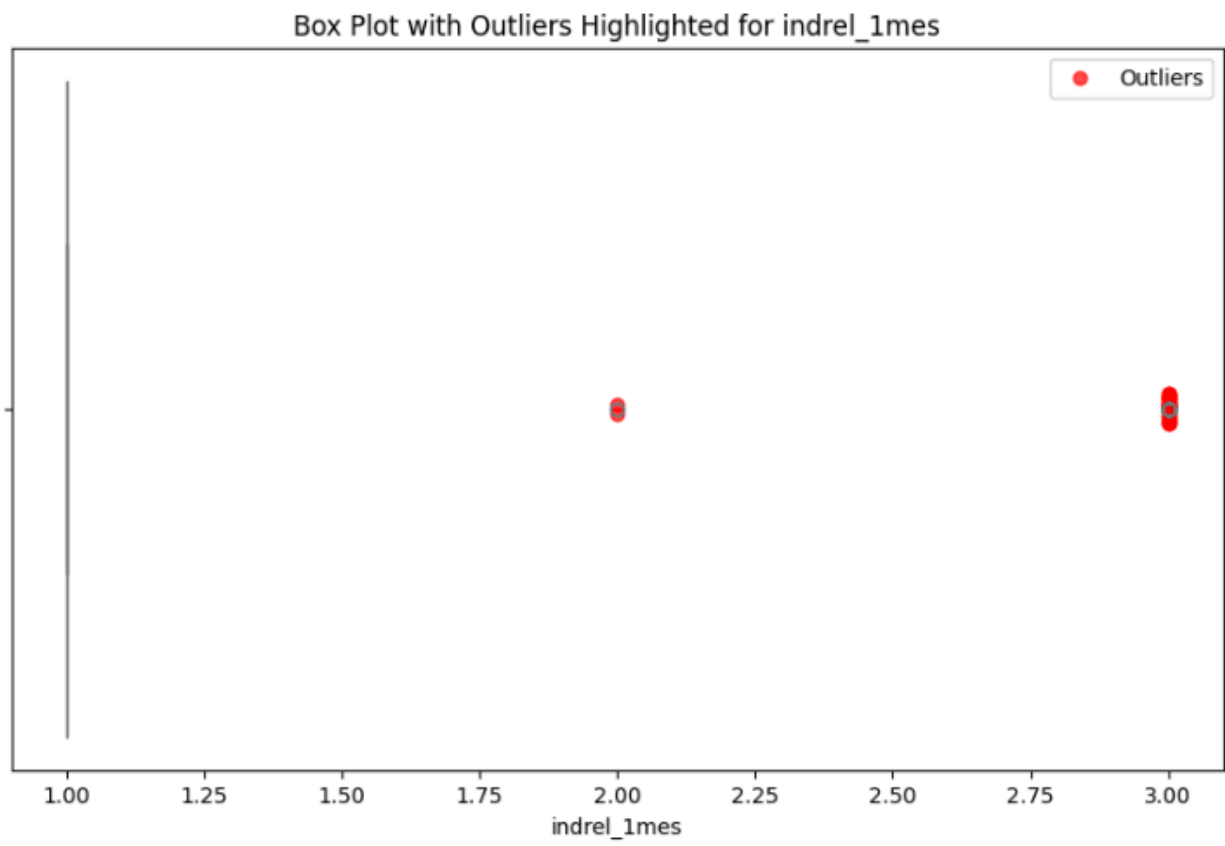
```
37 ind_fond_fin_ult1    1000000 non-null int64
38 ind_hip_fin_ult1     1000000 non-null int64
39 ind_plan_fin_ult1    1000000 non-null int64
40 ind_pres_fin_ult1    1000000 non-null int64
41 ind_reca_fin_ult1    1000000 non-null int64
42 ind_tjcr_fin_ult1    1000000 non-null int64
43 ind_valo_fin_ult1    1000000 non-null int64
44 ind_viv_fin_ult1     1000000 non-null int64
45 ind_nomina_ult1      994598 non-null float64
46 ind_nom_pens_ult1    994598 non-null float64
47 ind_recibo_ult1      1000000 non-null int64
dtypes: float64(9), int64(24), object(15)
```

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

-The dataset has a lot of missing data (2,371,207 missing datas) as below.

	Column_Name	aggregate	percent
0	conyuemp	999822	0.999822
1	ult_fec_cli_1t	998899	0.998899
2	renta	175183	0.175183
3	nomprov	17734	0.017734
4	cod_prov	17734	0.017734
5	canal_entrada	10861	0.010861
6	sexo	10786	0.010786
7	indresi	10782	0.010782
8	ind_actividad_cliente	10782	0.010782
9	tipodom	10782	0.010782
10	indfall	10782	0.010782
11	indext	10782	0.010782
12	tiprel_1mes	10782	0.010782
13	indrel_1mes	10782	0.010782
14	indrel	10782	0.010782
15	ind_nuevo	10782	0.010782
16	fecha_alta	10782	0.010782
17	pais_residencia	10782	0.010782
18	ind_empleado	10782	0.010782
19	ind_nomina_ult1	5402	0.005402
20	ind_nom_pens_ult1	5402	0.005402
21	ind_pres_fin_ult1	0	0.000000
22	ind_fond_fin_ult1	0	0.000000
23	ind_ecue_fin_ult1	0	0.000000
24	ind_hip_fin_ult1	0	0.000000

-Outliers

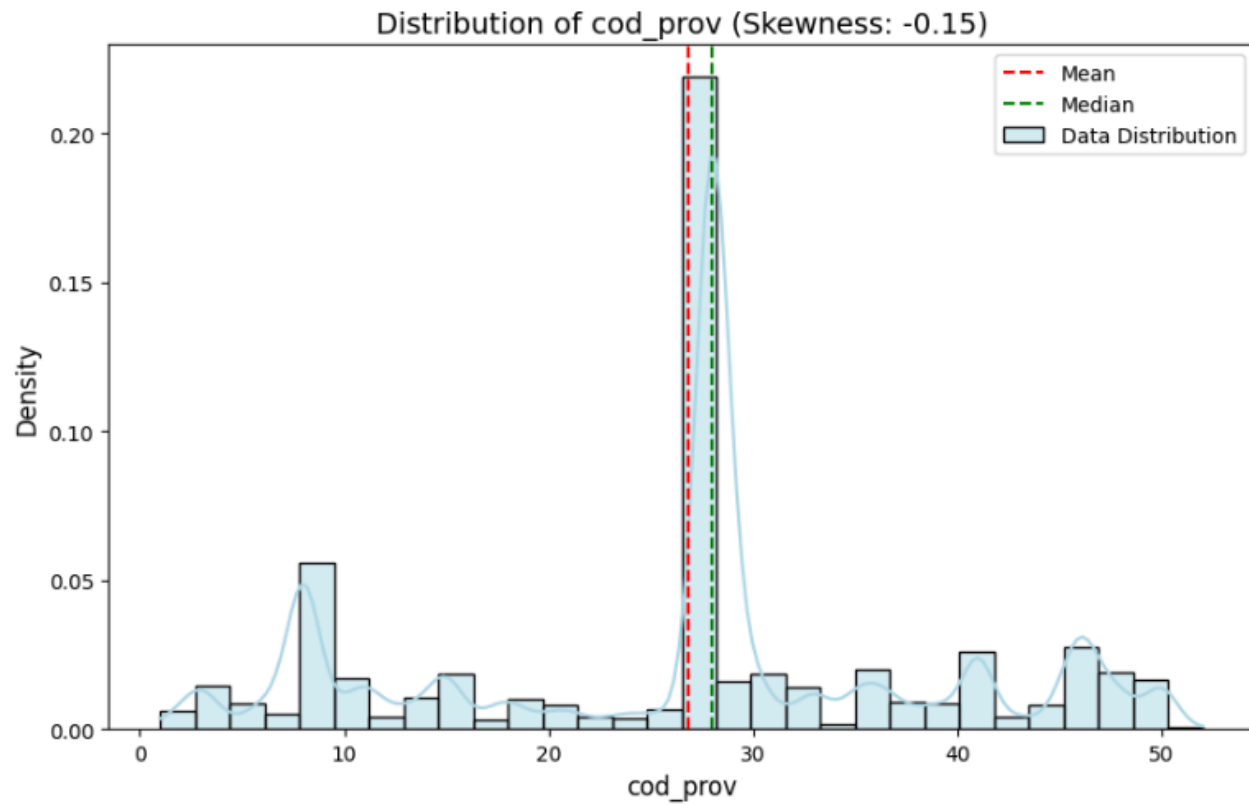


Regarding renta(Gross income of the household), there are so many outliers.

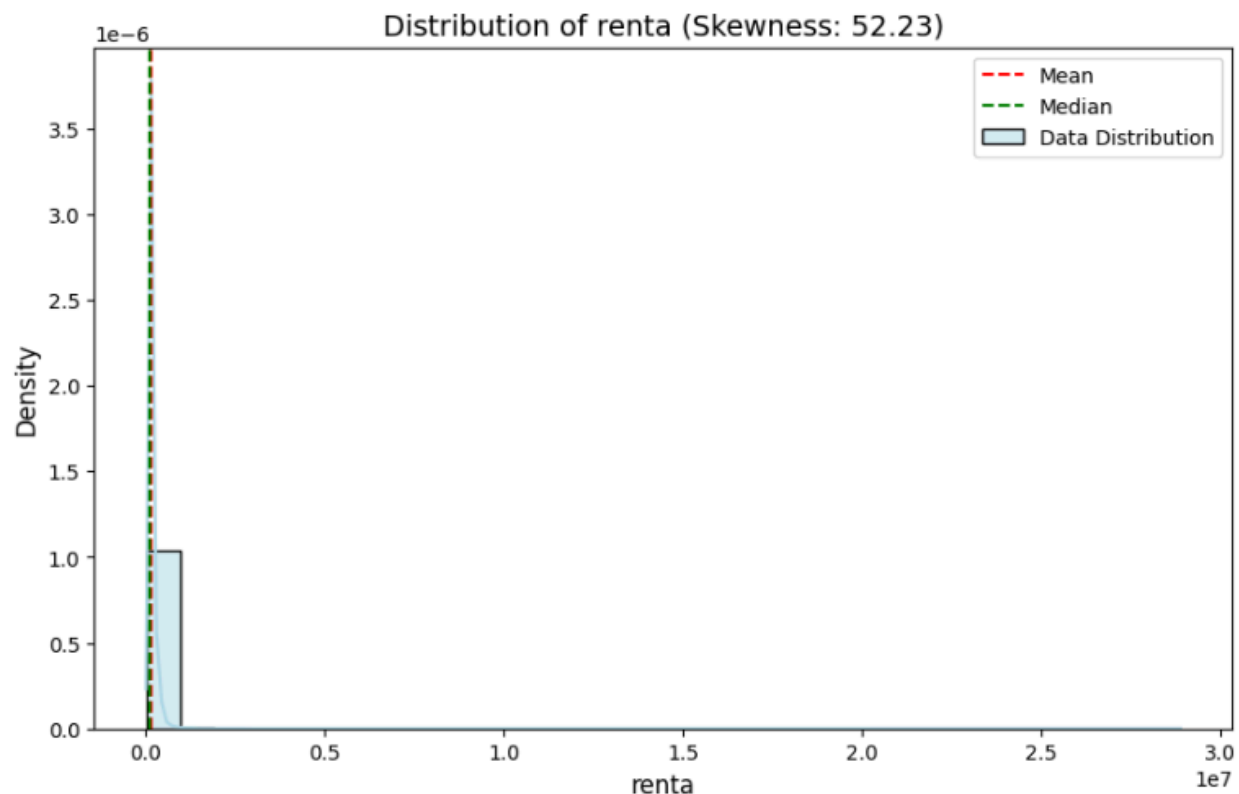


-Skewness

Regarding cod_prov(Province code (customer's address)), the median and the mean are around 27-28 and most of them were located to this bin.



Regarding renta(Gross income of the household), the median and the mean are skewed to right.



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

-NA value: Future Engineering, Imputing method using mean, median, and mode, etc.. If there are too many NA values in a variable, then dropping the column itself may be the best method because assuming missing data cannot be predictable.

Additionally, using the algorithm is good for missing data treatment such as NMM imputation and MICE(Multiple Imputation by Chained Equations)

-Outliers: May apply to omit them. Depending on the case.

Identify based IQR, Zscore, and anomaly detection models such as Isolation Forest and OBSCAN.

If the outlier indicates significant characteristics, we apply robust scaling (log) and omit specific upper/ lower limits.

-Skewness: In this case, there are many like one hot encoding and categoricals. And we apply the cluster method. Adjusting skewness like normalization may not good idea.

Project life cycle along with deadline:

Project weeks	Deadline	Lifecycle
Week7	Dec 19, 2024	Problem statement, Pre-process
Week8	Dec 26, 2024	Data process, understanding
Week9	Jan 02, 2025	Data Cleaning, Merge, Review
Week10	Jan 09, 2025	EDA, Final recommendation
Week11	Jan 16, 2025	EDA presentation for business users
Week12	Jan 23, 2025	Model Selection and Model Building/Dashboard
Week13	Jan 30, 2025	Final Project Report and Code

Tabular data details: forecasting_case_study.xlsx:

Total number of observations	1218
Total number of files	1
Total number of features	12
Base format of the file	.xlsx
Size of the data	80KB