Data Science Project

**Week 8 Deliverable**

**Group name: Data Explorers**

**Specialization:** Data Science

**Batch:** LISUM12

**Team members Details:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Email** | **Country** | **College\Company** |
| **Jalpa Deepak Patel** | pjalpa015@gmail.com | Canada | University of Windsor |
| **Dhvanilkumar Kiritkumar Prajapati** | dkprajapati46@gmail.com | Canada | University of Ottawa |
| **Sri Ram Prasad Commuri** | [csriramprasad@gmail.com](mailto:csriramprasad@gmail.com) | India | Data Glacier |

**Project Title:** Bank Customer Segmentation

**Problem Description**: A bank wants to create customer segments/categories to send personalized Christmas offers to its customers. We need to identify feature(s) to group customers into different categories.

Customer segmentation is an approach to creating smaller customer groups relevant to the bank’s product marketing and services. Based on customer accounts and usage details, sending different offers to customers increases the chances of increasing the business for a bank and reaching maximum customers as possible.

We need to create at most five customer segments. We plan to employ Data Science techniques to understand, pre-process and create customer segments.

**Data Source**: <https://drive.google.com/drive/folders/1bfCpJIKmp6IHxiLPWvOS2nU1dc24pViB>

**Data Understanding:**

* The dataset consists of 47 columns.
* The dataset consists of bank customer information. It consists of different account details, customer information (age, address, gender), and information on banking services (pensions, loans, credit cards, etc.)
* The data contains information about 1,000,000 customers of XYZ bank.
* The employee information (Present, ex-employee, active, passive) is present for each customer.
* Customer Details available are –
  + Age
  + Gender
  + date of joining the bank
  + customer type (owner, co-owner)
  + customer relation (active, inactive, potential, former)
  + Residence
  + Foreigner (different birth country)
  + Deceased
  + Address
  + Gross income
* Customer account and services information present in the dataset is as follows –
  + Saving account
  + Guarantees
  + Current Account
  + Derived Account
  + Payroll account
  + Junior account
  + Particular account, particular plus account
  + Short, medium, long-term deposits
  + E-account
  + Funds
  + Mortgage
  + Pensions
  + Loans
  + Taxes
  + Credit card
  + Securities
  + Home account
  + Direct debit
* As the account types and services are present in different columns, there is a possibility that not all customers would opt for all types of accounts and services.
* The data information column wise is as follows –

Data columns (total 48 columns):

# 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)

**NULL Values** –

* The null values columns are Employee Index, Customer country of residence, gender, date of customer joining, new customer index, Primary customer, late date for primary customer, customer type, address details (province, type), gross income, nominee payroll and pensions. The details are as follows –

ind\_empleado 10782

pais\_residencia 10782

sexo 10786

fecha\_alta 10782

ind\_nuevo 10782

indrel 10782

ult\_fec\_cli\_1t 998899

indrel\_1mes 10782

tiprel\_1mes 10782

indresi 10782

indext 10782

conyuemp 999822

canal\_entrada 10861

indfall 10782

tipodom 10782

cod\_prov 17734

nomprov 17734

ind\_actividad\_cliente 10782

renta 175183

ind\_nomina\_ult1 5402

ind\_nom\_pens\_ult1 5402

**Data Distribution** –

Data distribution helps us determine if any outliers are present in the data and how data is arranged.

The skewness of data helps understand the probability distribution of the data. It can be performed on financial data. The distribution for the gross income of customers is not a bell shaped distribution. There are many customers whose income lies between $1202 – $300,000. The maximum income recorded is 28894395.

The distribution of numeric data is identified as follows –

**Distribution of numeric data**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| ncodpers | 1000000.0 | 690596.670395 | 404408.432011 | 15889.00 | 336411.00 | 664476.00 | 1074511.25 | 1379131.00 |
| ind\_nuevo | 989218.0 | 0.000489 | 0.022114 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| indrel | 989218.0 | 1.109074 | 3.267624 | 1.00 | 1.00 | 1.00 | 1.00 | 99.00 |
| indrel\_1mes | 989218.0 | 1.000085 | 0.012954 | 1.00 | 1.00 | 1.00 | 1.00 | 3.00 |
| tipodom | 989218.0 | 1.000000 | 0.000000 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| cod\_prov | 982266.0 | 26.852131 | 12.422924 | 1.00 | 18.00 | 28.00 | 33.00 | 52.00 |
| ind\_actividad\_cliente | 989218.0 | 0.564971 | 0.495761 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| renta | 824817.0 | 139646.150940 | 238985.824907 | 1202.73 | 71571.84 | 106651.86 | 163432.47 | 28894395.51 |
| ind\_ahor\_fin\_ult1 | 1000000.0 | 0.000177 | 0.013303 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_aval\_fin\_ult1 | 1000000.0 | 0.000039 | 0.006245 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_cco\_fin\_ult1 | 1000000.0 | 0.749626 | 0.433229 | 0.00 | 0.00 | 1.00 | 1.00 | 1.00 |
| ind\_cder\_fin\_ult1 | 1000000.0 | 0.000591 | 0.024303 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_cno\_fin\_ult1 | 1000000.0 | 0.105296 | 0.306935 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_ctju\_fin\_ult1 | 1000000.0 | 0.013623 | 0.115920 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_ctma\_fin\_ult1 | 1000000.0 | 0.009894 | 0.098975 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_ctop\_fin\_ult1 | 1000000.0 | 0.212486 | 0.409067 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_ctpp\_fin\_ult1 | 1000000.0 | 0.072079 | 0.258619 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_deco\_fin\_ult1 | 1000000.0 | 0.002158 | 0.046404 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_deme\_fin\_ult1 | 1000000.0 | 0.003150 | 0.056036 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_dela\_fin\_ult1 | 1000000.0 | 0.066881 | 0.249816 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_ecue\_fin\_ult1 | 1000000.0 | 0.106267 | 0.308179 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_fond\_fin\_ult1 | 1000000.0 | 0.027182 | 0.162614 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_hip\_fin\_ult1 | 1000000.0 | 0.009982 | 0.099410 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_plan\_fin\_ult1 | 1000000.0 | 0.014553 | 0.119755 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_pres\_fin\_ult1 | 1000000.0 | 0.004661 | 0.068112 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_reca\_fin\_ult1 | 1000000.0 | 0.072581 | 0.259448 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_tjcr\_fin\_ult1 | 1000000.0 | 0.066084 | 0.248429 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_valo\_fin\_ult1 | 1000000.0 | 0.039378 | 0.194493 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_viv\_fin\_ult1 | 1000000.0 | 0.006442 | 0.080003 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_nomina\_ult1 | 994598.0 | 0.071629 | 0.257873 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_nom\_pens\_ult1 | 994598.0 | 0.079543 | 0.270584 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |
| ind\_recibo\_ult1 | 1000000.0 | 0.166275 | 0.372327 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

The time series classification of data is computed for the number of customers for the date when customer joined bank and the date the records were fetched. The results as follows –

To check the time series of numeric data by daily, monthly and yearly frequency

Plotting daily data

Chart, line chart

Description automatically generated

Plotting monthly data

Chart, line chart

Description automatically generated

Plotting daily data

Chart

Description automatically generated

Plotting monthly data

Graphical user interface, application

Description automatically generated

Plotting yearly data

Chart, line chart

Description automatically generated

As observed in the graphs, the account creation of customers is not equally distributed. There was a significant increase in customers between 2010 and 2015.

Approach for Data cleaning for the upcoming week –

1. Columns with account details have Null values or 0; however, not all customers have all types of accounts and services. Hence, we can still consider information from those columns.
2. In columns such as account creation date and income, the null values can maybe be imputed by considering the mean values. We will conduct experiments to see what options fit the best.
3. We also intend to conduct experiments on treating outliers. In the initial study of data, outliers have been observed. However, by conducting a study about the bank sector and analyzing the outlier data, we decide how to treat the outlier (trimming or imputation with values).

The code for the task can be found at - <https://github.com/jalpa015/DataScience_Project/blob/master/EDA.ipynb>