Total number of rows in raw data: **13647309**

Null value count for each column:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name (English)** | **Column Name (Original)** | **Null Values** | **Description** |
| Date | fecha\_dato | 0 | The table is partitioned for this column |
| customer\_code | ncodpers | 0 | Customer code |
| employee\_index | ind\_empleado | 27734 | Employee index: A active, B ex employed, F filial, N not employee, P pasive |
| customer\_residence | pais\_residencia | 27734 | Customer's Country residence |
| gender | sexo | 27804 | Customer's sex |
| age | age | 0 | Age |
| customer\_join \_date | fecha\_alta | 27734 | The date in which the customer became as the first holder of a contract in the bank |
| index\_customer\_reg\_span | ind\_nuevo | 0 | New customer Index. 1 if the customer registered in the last 6 months. |
| customer\_seniority | antiguedad | 0 | Customer seniority (in months) |
| index\_customer | indrel | 0 | 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month) |
| customer\_primary\_last\_date | ult\_fec\_cli\_1t | 13622516 | Last date as primary customer (if he isn't at the end of the month) |
| customer\_type\_month | indrel\_1mes | 149781 | Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner ),P (Potential),3 (former primary), 4(former co-owner) |
| customer\_relation\_type\_month | tiprel\_1mes | 149781 | Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential) |
| customer\_residence\_local\_index | indresi | 27734 | Residence index (S (Yes) or N (No) if the residence country is the same than the bank country) |
| customer\_residence\_foreign\_index | indext | 27734 | Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country) |
| index\_spouse | conyuemp | 13645501 | Spouse index. 1 if the customer is spouse of an employee |
| customer\_channel | canal\_entrada | 186126 | channel used by the customer to join |
| index\_deceased | indfall | 27734 | Deceased index. N/S |
| Index\_primary\_address | tipodom | 0 | Address type. 1, primary address |
| Province\_code | cod\_prov | 0 | Province code (customer's address) |
| Province\_name | nomprov | 93591 | Province name |
| index\_customer\_activity | ind\_actividad\_cliente | 0 | Activity index (1, active customer; 0, inactive customer) |
| customer\_gross\_income | renta | 2794375 | Gross income of the household |
| customer\_type | segmento | 189368 | segmentation: 01 - VIP, 02 - Individuals 03 - college graduated |
| Product\_saving\_account | ind\_ahor\_fin\_ult1 | 0 | Saving Account |
| Product\_guarantees | ind\_aval\_fin\_ult1 | 0 | Guarantees |
| Product\_current\_account | ind\_cco\_fin\_ult1 | 0 | Current Accounts |
| Product\_derivative\_coount | ind\_cder\_fin\_ult1 | 0 | Derivada Account |
| Product\_pryroll\_account | ind\_cno\_fin\_ult1 | 0 | Payroll Account |
| Product\_junior\_account | ind\_ctju\_fin\_ult1 | 0 | Junior Account |
| Product\_more\_particular\_account | ind\_ctma\_fin\_ult1 | 0 | Más particular Account |
| product\_particular\_account | ind\_ctop\_fin\_ult1 | 0 | particular Account |
| Product\_particular\_plus\_account | ind\_ctpp\_fin\_ult1 | 0 | particular Plus Account |
| Product\_short\_term\_deposits | ind\_deco\_fin\_ult1 | 0 | Short-term deposits |
| Product\_medium\_term\_deposits | ind\_deme\_fin\_ult1 | 0 | Medium-term deposits |
| Product\_long\_Term\_deposits | ind\_dela\_fin\_ult1 | 0 | Long-term deposits |
| Product\_e\_account | ind\_ecue\_fin\_ult1 | 0 | e-account |
| Product\_funds | ind\_fond\_fin\_ult1 | 0 | Funds |
| Product\_mortgage | ind\_hip\_fin\_ult1 | 0 | Mortgage |
| Product\_pensions | ind\_plan\_fin\_ult1 | 0 | Pensions |
| Product\_loans | ind\_pres\_fin\_ult1 | 0 | Loans |
| Product\_taxes | ind\_reca\_fin\_ult1 | 0 | Taxes |
| Product\_credit\_card | ind\_tjcr\_fin\_ult1 | 0 | Credit Card |
| Product\_securities | ind\_valo\_fin\_ult1 | 0 | Securities |
| Product\_home\_account | ind\_viv\_fin\_ult1 | 0 | Home Account |
| Product\_payroll | ind\_nomina\_ult1 | 0 | Payroll |
| Product\_pensions | ind\_nom\_pens\_ult1 | 0 | Pensions |
| Product\_direct\_Debit | ind\_recibo\_ult1 | 0 | Direct Debit |

**Datatype for each column variable:**

|  |  |
| --- | --- |
| **Column Name** | **DataType** |
| fecha\_dato | timestamp (nullable = true) |
| ncodpers | double (nullable = true) |
| ind\_empleado | string (nullable = true) |
| pais\_residencia | string (nullable = true) |
| Sexo | string (nullable = true) |
| Age | string (nullable = true) |
| fecha\_alta | timestamp (nullable = true) |
| ind\_nuevo | string (nullable = true) |
| antiguedad | string (nullable = true) |
| Indrel | string (nullable = true) |
| ult\_fec\_cli\_1t | timestamp (nullable = true) |
| indrel\_1mes | string (nullable = true) |
| tiprel\_1mes | string (nullable = true) |
| Indresi | string (nullable = true) |
| Indext | string (nullable = true) |
| conyuemp | string (nullable = true) |
| canal\_entrada | string (nullable = true) |
| Indfall | string (nullable = true) |
| tipodom | string (nullable = true) |
| cod\_prov | string (nullable = true) |
| nomprov | string (nullable = true) |
| ind\_actividad\_cliente | string (nullable = true) |
| Renta | double (nullable = true) |
| segmento | string (nullable = true) |
| ind\_ahor\_fin\_ult1 | integer (nullable = true) |
| ind\_aval\_fin\_ult1 | integer (nullable = true) |
| ind\_cco\_fin\_ult1 | integer (nullable = true) |
| ind\_cder\_fin\_ult1 | integer (nullable = true) |
| ind\_cno\_fin\_ult1 | integer (nullable = true) |
| ind\_ctju\_fin\_ult1 | integer (nullable = true) |
| ind\_ctma\_fin\_ult1 | integer (nullable = true) |
| ind\_ctop\_fin\_ult1 | integer (nullable = true) |
| ind\_ctpp\_fin\_ult1 | integer (nullable = true) |
| ind\_deco\_fin\_ult1 | integer (nullable = true) |
| ind\_deme\_fin\_ult1 | integer (nullable = true) |
| ind\_dela\_fin\_ult1 | integer (nullable = true) |
| ind\_ecue\_fin\_ult1 | integer (nullable = true) |
| ind\_fond\_fin\_ult1 | integer (nullable = true) |
| ind\_hip\_fin\_ult1 | integer (nullable = true) |
| ind\_plan\_fin\_ult1 | integer (nullable = true) |
| ind\_pres\_fin\_ult1 | integer (nullable = true) |
| ind\_reca\_fin\_ult1 | integer (nullable = true) |
| ind\_tjcr\_fin\_ult1 | integer (nullable = true) |
| ind\_valo\_fin\_ult1 | integer (nullable = true) |
| ind\_viv\_fin\_ult1 | integer (nullable = true) |
| ind\_nomina\_ult1 | string (nullable = true) |
| ind\_nom\_pens\_ult1 | string (nullable = true) |
| ind\_recibo\_ult1 | integer (nullable = true) |

**Exploratory Analysis:**

1. Number of Female Customers (46%) are less than Number of Male customers (54%)
2. Last Date being Primary Customer the max value is 2016-05-30
3. Individual Customers (58%) are more than VIP (6%) and Students (36%) of total number of customers
4. Customer type has values of data type String, Double and Int
5. Average gross income of VIP customers is more than Individual and Student customers
6. Average gross income of female is lesser than male
7. Male with Individual customer types are maximum
8. Count of Guarantees is maximum for male individual type customers with average gross income approx. 139593
9. Number of savings account maximum for male individual type of customers with average gross income 139593
10. Number of derivada account maximum for male individual type of customers with average gross income 139593
11. Further analysis shows that all the products are consumed maximum by male individual type of customer
12. There are more university students as customer than other types (Individual and VIP)

Based on Data Exploration process we will be handling missing values in following way:

**Data Cleaning Process**

**15 columns** have missing values.

**Age:**

NA - 27734

Step 1 - Remove outliers firsts.

- Divide age in to sections replace with mean values. This will hep us fix the age distribution.

mean(age >=18 and <=30) - Replace rows with age < 18

mean(age >= 30 and age <= 100) - Replace rows with age > 100

Step 2 - Fix NA values

- Impute NA and replace all the NA values with median of age columns.

- round(age) to convert it into integer from float

**fecha\_alta (date of join)**

NA - 27734

Step - replace NA values with median (middle value) of fetcha\_alta column

**nomprov (providence name)**

NA - 93591

Step - replace NA values with “Unknown”

**sexo (gender)**

NA - 27804

Step - we will replace the NA values with ratio

**indfall**

NA - 27734

Step - we will replace NA values with ratio

**indresi**

NA - 27734

Step - we will replace NA values with ratio

**ind\_empleado**

NA - 27734

Step - we will replace NA values with ratio

**indrel\_1mes**

NA - 149781

Step - we will replace NA values with ratio

**tiprel\_1mes**

NA - 149781

Step - we will replace NA values with ratio

**indext**

NA - 27734

Step - we will replace NA values with ratio

**canal\_entrada**

NA - 186126

Step - we will replace NA values with ratio

**renta**

NA - 186126

Step - we will replace NA values with average income of the customers in same province

**segmento (customer type)**

NA - 189368

Step - we will replace the NA values with “Unknown”

Columns to remove -

**tipodom** : not useful

**cod\_prov**: we already have this information in nomprov (province name)

**convuemp** (index\_spouse): so many null values

**ult\_fec\_cli\_lt** (last primary date) - so many null values