Insights from the Zeppelin

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)

Insights from the Zeppelin

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)

Insights from the Zeppelin

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

Insights from the Zeppelin

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)

Insights from the Zeppelin

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