**Zeppelin: Exploratory aNnaysis & Data Cleaning**

**CSYE 7200 Big Data Engineering using Scala**

Northeastern University | Spring 2018

**Team 7 Santander Product Recommendation System**

# **Exploratory Analysis**

1. Total number of rows in raw data: **13647309**
2. Renamed all Spanish columns names to English names and then took null count for each column. Below table describes information about the same:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column Name [English]** | **Column Name [Original]** | **Null Values** | **Description** |
| Partitioned\_Date | fecha\_dato | 0 | The table is partitioned for this column |
| Cust\_Code | Ncodpers | 0 | Customer code |
| Emp\_Index | ind\_empleado | 27734 | Employee index: A active B ex employed F Filial N not employee P passive |
| Cust\_Residence | pais\_residencia | 27734 | Customer's Country residence |
| Cust\_Gender | Sexo | 27804 | Customer's sex |
| Cust\_Age | Age | 0 | Age |
| First\_Holder\_Date | fecha\_alta | 27734 | The date in which the customer became as the first holder of a contract in the bank |
| New\_Cust\_Index | ind\_nuevo | 0 | New customer Index. 1 if the customer registered in the last 6 months. |
| Cust\_Seniority | Antiguedad | 0 | Customer seniority [in months] |
| Cust\_Index\_Primary | Indrel | 0 | 1 [First/Primary] 99 [Primary customer during the month but not at the end of the month] |
| Last\_Date\_Primary\_Cust | ult\_fec\_cli\_1t | 13622516 | Last date as primary customer [if he isn't at the end of the month] |
| Cust\_Type | 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] |
| Cust\_Relation\_Type | tiprel\_1mes | 149781 | Customer relation type at the beginning of the month A [active] I [inactive] P [former customer] R [Potential] |
| Residence\_Index | Indresi | 27734 | Residence index [S [Yes] or N [No] if the residence country is the same than the bank country] |
| Foreigner\_Index | Indext | 27734 | Foreigner index [S [Yes] or N [No] if the customer's birth country is different than the bank country] |
| Spouse\_Index | Conyuemp | 13645501 | Spouse index. 1 if the customer is spouse of an employee |
| Channel | canal\_entrada | 186126 | channel used by the customer to join |
| Deceased\_Index | Indfall | 27734 | Deceased index. N/S |
| Addres\_Type | Tipodom | 0 | Address type. 1 primary address |
| Province\_Code | cod\_prov | 0 | Province code [customer's address] |
| Province\_Name | Nomprov | 93591 | Province name |
| Activity\_Index | ind\_actividad\_cliente | 0 | Activity index [1 active customer; 0 inactive customer] |
| Gross\_Income | Renta | 2794375 | Gross income of the household |
| Cust\_Identification | Segmento | 189368 | segmentation: 01 - VIP 02 - Individuals 03 - college graduated |
| Saving\_Acc | ind\_ahor\_fin\_ult1 | 0 | Saving Account |
| Guarantees | ind\_aval\_fin\_ult1 | 0 | Guarantees |
| Current\_Acc | ind\_cco\_fin\_ult1 | 0 | Current Accounts |
| Derivada\_Acc | ind\_cder\_fin\_ult1 | 0 | Derivada Account |
| Payroll\_Acc | ind\_cno\_fin\_ult1 | 0 | Payroll Account |
| Junior\_Acc | ind\_ctju\_fin\_ult1 | 0 | Junior Account |
| Mas\_Acc | ind\_ctma\_fin\_ult1 | 0 | Más particular Account |
| Particular\_Acc | ind\_ctop\_fin\_ult1 | 0 | particular Account |
| Particular\_Plus\_Acc | ind\_ctpp\_fin\_ult1 | 0 | particular Plus Account |
| Short\_Term\_Deposit | ind\_deco\_fin\_ult1 | 0 | Short-term deposits |
| Medium\_Term\_Deposits | ind\_deme\_fin\_ult1 | 0 | Medium-term deposits |
| Long\_Term\_Deposits | ind\_dela\_fin\_ult1 | 0 | Long-term deposits |
| e\_Acc | ind\_ecue\_fin\_ult1 | 0 | e-account |
| Funds | ind\_fond\_fin\_ult1 | 0 | Funds |
| Mortgage | ind\_hip\_fin\_ult1 | 0 | Mortgage |
| Pensions | ind\_plan\_fin\_ult1 | 0 | Pensions |
| Loans | ind\_pres\_fin\_ult1 | 0 | Loans |
| Taxes | ind\_reca\_fin\_ult1 | 0 | Taxes |
| Credit\_Card | ind\_tjcr\_fin\_ult1 | 0 | Credit Card |
| Securities | ind\_valo\_fin\_ult1 | 0 | Securities |
| Home\_Acc | ind\_viv\_fin\_ult1 | 0 | Home Account |
| Payroll | ind\_nomina\_ult1 | 0 | Payroll |
| Nom\_Pensions | ind\_nom\_pens\_ult1 | 0 | Pensions |
| Direct\_Debit | ind\_recibo\_ult1 | 0 | Direct Debit |

1. **Datatype for each column variable:**

|  |  |  |
| --- | --- | --- |
| **Column name [English]** | **Column name [Original]** | **DataType** |
| Partitioned\_Date | fecha\_dato | timestamp [nullable = true] |
| Cust\_Code | Ncodpers | double [nullable = true] |
| Emp\_Index | ind\_empleado | string [nullable = true] |
| Cust\_Residence | pais\_residencia | string [nullable = true] |
| Cust\_Gender | Sexo | string [nullable = true] |
| Cust\_Age | Age | string [nullable = true] |
| First\_Holder\_Date | fecha\_alta | timestamp [nullable = true] |
| New\_Cust\_Index | ind\_nuevo | string [nullable = true] |
| Cust\_Seniority | Antiguedad | string [nullable = true] |
| Cust\_Index\_Primary | Indrel | string [nullable = true] |
| Last\_Date\_Primary\_Cust | ult\_fec\_cli\_1t | timestamp [nullable = true] |
| Cust\_Type | indrel\_1mes | string [nullable = true] |
| Cust\_Relation\_Type | tiprel\_1mes | string [nullable = true] |
| Residence\_Index | Indresi | string [nullable = true] |
| Foreigner\_Index | Indext | string [nullable = true] |
| Spouse\_Index | Conyuemp | string [nullable = true] |
| Channel | canal\_entrada | string [nullable = true] |
| Deceased\_Index | Indfall | string [nullable = true] |
| Addres\_Type | Tipodom | string [nullable = true] |
| Province\_Code | cod\_prov | string [nullable = true] |
| Province\_Name | Nomprov | string [nullable = true] |
| Activity\_Index | ind\_actividad\_cliente | string [nullable = true] |
| Gross\_Income | Renta | double [nullable = true] |
| Cust\_Identification | Segment | string [nullable = true] |
| Saving\_Acc | ind\_ahor\_fin\_ult1 | integer [nullable = true] |
| Guarantees | ind\_aval\_fin\_ult1 | integer [nullable = true] |
| Current\_Acc | ind\_cco\_fin\_ult1 | integer [nullable = true] |
| Derivada\_Acc | ind\_cder\_fin\_ult1 | integer [nullable = true] |
| Payroll\_Acc | ind\_cno\_fin\_ult1 | integer [nullable = true] |
| Junior\_Acc | ind\_ctju\_fin\_ult1 | integer [nullable = true] |
| Mas\_Acc | ind\_ctma\_fin\_ult1 | integer [nullable = true] |
| Particular\_Acc | ind\_ctop\_fin\_ult1 | integer [nullable = true] |
| Particular\_Plus\_Acc | ind\_ctpp\_fin\_ult1 | integer [nullable = true] |
| Short\_Term\_Deposit | ind\_deco\_fin\_ult1 | integer [nullable = true] |
| Medium\_Term\_Deposits | ind\_deme\_fin\_ult1 | integer [nullable = true] |
| Long\_Term\_Deposits | ind\_dela\_fin\_ult1 | integer [nullable = true] |
| e\_Acc | ind\_ecue\_fin\_ult1 | integer [nullable = true] |
| Funds | ind\_fond\_fin\_ult1 | integer [nullable = true] |
| Mortgage | ind\_hip\_fin\_ult1 | integer [nullable = true] |
| Pensions | ind\_plan\_fin\_ult1 | integer [nullable = true] |
| Loans | ind\_pres\_fin\_ult1 | integer [nullable = true] |
| Taxes | ind\_reca\_fin\_ult1 | integer [nullable = true] |
| Credit\_Card | ind\_tjcr\_fin\_ult1 | integer [nullable = true] |
| Securities | ind\_valo\_fin\_ult1 | integer [nullable = true] |
| Home\_Acc | ind\_viv\_fin\_ult1 | integer [nullable = true] |
| Payroll | ind\_nomina\_ult1 | string [nullable = true] |
| Nom\_Pensions | ind\_nom\_pens\_ult1 | string [nullable = true] |
| Direct\_Debit | ind\_recibo\_ult1 | integer [nullable = true] |

1. **Detail Exploratory Analysis:**

* Number of Female Customers [46%] are less than Number of Male customers [54%]
* Last Date being Primary Customer the max value is 2016-05-30
* Individual Customers [58%] are more than VIP [6%] and Students [36%] of total number of customers
* Customer type has values of data type String Double and Integer
* Average gross income of VIP customers is more than Individual and Student customers
* Average gross income of female is lesser than male
* Male with Individual customer types are maximum
* Count of Guarantees is maximum for male individual type customers with average gross income approx. 139593
* Number of savings account maximum for male individual type of customers with average gross income 139593
* Number of derivada account maximum for male individual type of customers with average gross income 139593
* Further analysis shows that all the products are consumed maximum by male individual type of customer
* There are more university students as customer than other types [Individual and VIP]
* The Madrid region has maximum number of customers
* Number of customers as per their last date being primary customers increases toward year end
* Number of customers as per their Join date [ First\_Date\_Acc\_Holder] is steady for initial 6 months span of a year

# **Approach to follow towards data cleaning activities:**

**Total 15 columns** have missing [null] values, also for few columns need to check for datatypes. The approach we will follow is as below:

1. **Age [Cust\_Age]:**

*Null value count* = 0

*Step 1* - Remove outliers firsts.

- Divide age in to sections replace with mean values. This will help 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 data type

- round[age] to convert it into integer from float

1. **fecha\_alta [First\_Holder\_Date]**

*Null value count* - 27734

*Step* – replace null values with median [middle value] of fetcha\_alta column

1. **nomprov [Province\_Name]**

*Null value count* - 93591

*Step* - replace NULL values with Unknown

1. **sexo [Cust\_Gender]**

*Null value count* - 27804

*Step* - we will replace the NULL values with ratio

1. **indfall [Deceased\_Index]**

*Null value count* - 27734

*Step* - we will replace NULL values with ratio

1. **indresi [Residence\_Index]**

*Null value count* - 27734

*Step* - we will replace NULL values with ratio

1. **ind\_empleado [Emp\_Index]**

*Null value count* - 27734

*Step* - we will replace NULL values with ratio

1. **indrel\_1mes [Cust\_Type]**

*Null value count* - 149781

*Step* - we will replace NULL values with Unknown

1. **tiprel\_1mes [Cust\_Relation\_type]**

*Null value count* - 149781

*Step* - we will replace NULL values with ratio

1. **indext [Foreigner\_Index]**

*Null value count* - 27734

*Step* - we will replace NULL values with ratio

1. **canal\_entrada [Channel]**

*Null value count* - 186126

*Step* - we will replace NULL values with Unknown

1. **renta [Gross\_Income]**

*Null value count* - 2794375

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

1. **segmento [Cust\_Identification]**

*Null value count* - 189368

*Step* - we will replace the NULL values with Unknown

1. **Columns to remove**

* **tipodom** [Address\_Type]: not useful
* **cod\_prov** [Province\_Code]: we already have this information in nomprov [Province\_Name], so can be removed
* **conyuemp** [Spouse\_Index]: approx. 99% values are null, so can be removed
* **ult\_fec\_cli\_lt** [Last\_Date\_Primary\_Cust]: approx. 99% values are null, so can be removed