Customer Segmentation Project

Week 8

**Name**: Customer Segmentation Project

**Report date**: 02-Sep-2021

**Internship Batch**: LISUM02

**Specialization**: Data Science

**Group Name**: Data Explorers

**GitHub link**: https://github.com/joeanton719/Customer-Segmentation-Project

Team member’s details:

* **Joseph Antony**
  + Email: [joeanton719@gmail.com](mailto:joeanton719@gmail.com)
  + Country: Turkey
  + Company: UrbanStat
* **Melisa Gozet**
  + Email: [mgozet@gmail.com](mailto:mgozet@gmail.com)
  + Country: Turkey
  + College: Ankara University Artificial Intelligence Technology PhD student
* **Dilem Unal**
  + Email: diilemunal[@gmail.com](mailto:cemreaka@gmail.com)
  + Country: Turkey
  + College/Company: Istanbul Aydın University Software Engineer Student
* **Aynur Cemre Aka**
  + Email: [cemreaka@gmail.com](mailto:cemreaka@gmail.com)
  + Country: Turkey
  + College: Yaşar University Software Engineering Student

Problem Description

Bank XYZ wants to offer Christmas offers to its customers. However, the bank does not want to offer the same offer to all its customers. Instead, they want to deploy the personalized offer to a particular group of customers. It will not be efficient to manually start understanding the category of the customer and they’ll not be able to uncover the hidden pattern in the data. ABC analytics assigned this task to their analytics team and instructed their team to come up with the approach and feature which groups similar behavior customers in one category and others in different categories. There shouldn’t be more than 5 groups as this will be inefficient.

Data Understanding

The dataset provided contains a list of the Bank’s customers from 1995 to 2015. Each observation is supposed to represent the different attributes belonging to a unique customer. The attributes are related to the customer activities with the bank account and other personal information such as the customer’s gender, joining date, where the customer is active, the customer’s residence, the bank products utilized by the customer, etc.

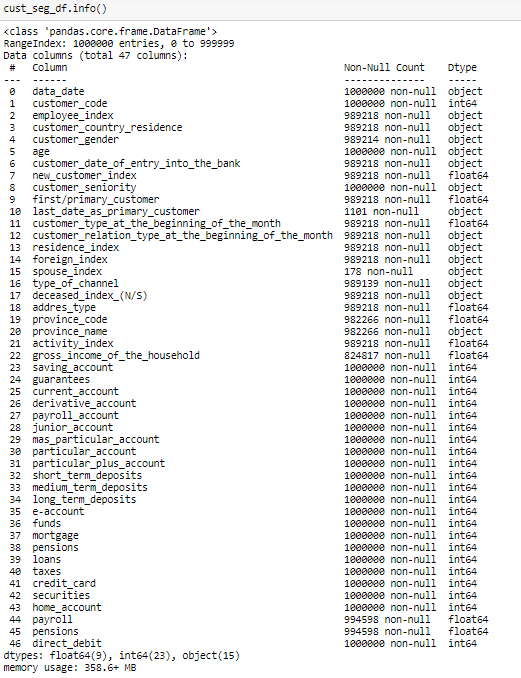
| **Column Name (Spanish)** | **Column Name (English)** | **Description** |
| --- | --- | --- |
| **fecha\_dato** | **data\_date** | **The table is partitioned for this column** |
| **ncodpers** | **customer\_code** | **Customer code** |
| **ind\_empleado** | **employee\_index** | **Employee index: A active, B ex employed, F filial, N not employee, P pasive** |
| **pais\_residencia** | **customer\_country\_residence** | **Customer's Country residence** |
| **sexo** | **customer\_gender** | **Customer's sex** |
| **age** | **age** | **Age** |
| **fecha\_alta** | **customer\_date\_of\_entry\_into\_the\_bank** | **The date in which the customer became as the first holder of a contract in the bank** |
| **ind\_nuevo** | **new\_customer\_index** | **New customer Index. 1 if the customer registered in the last 6 months.** |
| **antiguedad** | **customer\_seniority** | **Customer seniority (in months)** |
| **indrel** | **first/primary\_customer** | **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** | **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** | **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** | **Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)** |
| **indresi** | **residence\_index** | **Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)** |
| **indext** | **foreign\_index** | **Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)** |
| **conyuemp** | **spouse\_index** | **Spouse index. 1 if the customer is spouse of an employee** |
| **canal\_entrada** | **type\_of\_channel** | **channel used by the customer to join** |
| **indfall** | **deceased\_index\_(N/S)** | **Deceased index. N/S** |
| **tipodom** | **addres\_type** | **Addres type. 1, primary address** |
| **cod\_prov** | **province\_code** | **Province code (customer's address)** |
| **nomprov** | **province\_name** | **Province name** |
| **ind\_actividad\_cliente** | **activity\_index** | **Activity index (1, active customer; 0, inactive customer)** |
| **renta** | **gross\_income\_of\_the\_ household** | **Gross income of the household** |
| **ind\_ahor\_fin\_ult1** | **saving\_account** | **Saving Account** |
| **ind\_aval\_fin\_ult1** | **guarantees** | **Guarantees** |
| **ind\_cco\_fin\_ult1** | **current\_account** | **Current Accounts** |
| **ind\_cder\_fin\_ult1** | **derivative\_account** | **Derivada Account** |
| **ind\_cno\_fin\_ult1** | **payroll\_account** | **Payroll Account** |
| **ind\_ctju\_fin\_ult1** | **junior\_account** | **Junior Account** |
| **ind\_ctma\_fin\_ult1** | **más\_particular\_account** | **Más particular Account** |
| **ind\_ctop\_fin\_ult1** | **particular\_account** | **particular Account** |
| **ind\_ctpp\_fin\_ult1** | **particular\_plus\_account** | **particular Plus Account** |
| **ind\_deco\_fin\_ult1** | **short\_term\_deposits** | **Short-term deposits** |
| **ind\_deme\_fin\_ult1** | **medium\_term\_deposits** | **Medium-term deposits** |
| **ind\_dela\_fin\_ult1** | **long\_term\_deposits** | **Long-term deposits** |
| **ind\_ecue\_fin\_ult1** | **e-account** | **e-account** |
| **ind\_fond\_fin\_ult1** | **funds** | **Funds** |
| **ind\_hip\_fin\_ult1** | **mortgage** | **Mortgage** |
| **ind\_plan\_fin\_ult1** | **pensions** | **Pensions** |
| **ind\_pres\_fin\_ult1** | **loans** | **Loans** |
| **ind\_reca\_fin\_ult1** | **taxes** | **Taxes** |
| **ind\_tjcr\_fin\_ult1** | **credit\_card** | **Credit Card** |
| **ind\_valo\_fin\_ult1** | **securities** | **Securities** |
| **ind\_viv\_fin\_ult1** | **home\_account** | **Home Account** |
| **ind\_nomina\_ult1** | **payroll** | **Payroll** |
| **ind\_nom\_pens\_ult1** | **pensions** | **Pensions** |
| **ind\_recibo\_ult1** | **direct\_debit** | **Direct Debit** |

What type of data have you got for analysis?

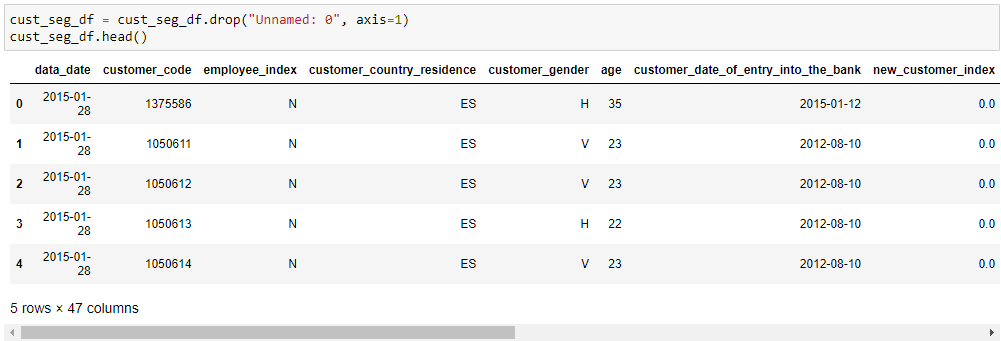
The dataset provided to us is of CSV format. It is quite large in size and dimension. The dataset contains 1 million rows and 48 columns, having a size of around approximately 366 MB. The datasets mostly contain numerical and categorical data types. A couple of columns are of date-time format.

Fewer categorical columns have higher cardinality, i.e, they have more than 10 categories. Most of the categorical columns are binary. Among the numerical features, only the `renta` variable is continuous. The rest are integers. It is important to note that some of the binary categorical columns are of float data type.

Below, we have attached snapshots of the datasets and its data types.



*Fig 1: Datasets Features and its Data Types (Before Dedup)*



*Fig 2: First five observations and 7 columns of the dataset*

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

The dataset is far from perfect. During initial dedup analysis, a significant number of missing values and duplicated observation has been observed. Among the categorical variables, there is higher cardinality for some of the categories, with some categories having only negligible counts compared to other categories. This can potentially make our model more biased towards those categories with higher value counts.

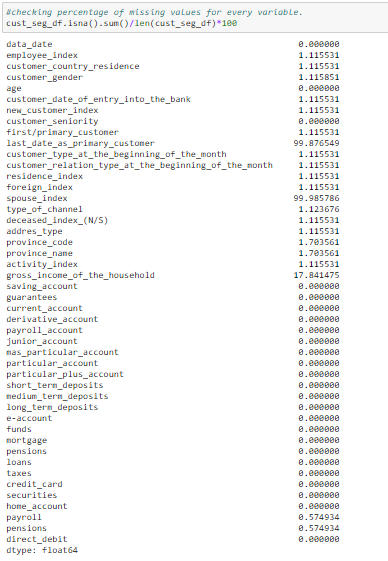
Below, snapshots of the data illustrating the problem are attached.

### 

### 

### 

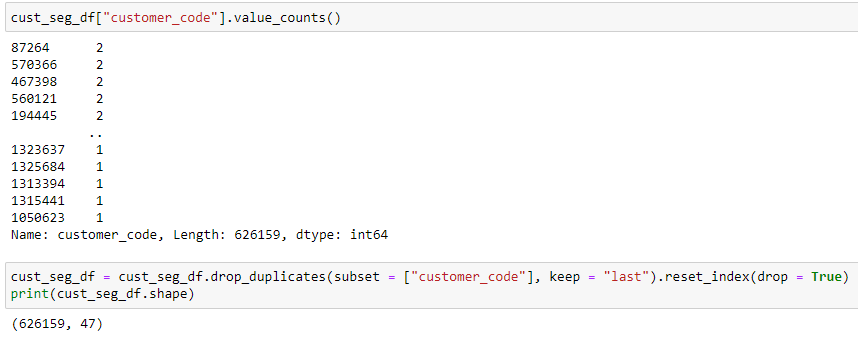
### Missing Values



*Fig 3: Percentage of Missing values across all features.*

A couple of columns have approximately 99.9 % missing data, followed by a third feature with around 17% missing data. There are other columns with much fewer missing data (less than 1%).

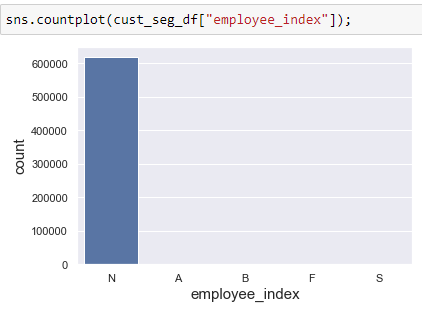
### Duplicated Observations



*Fig 4: Count of unique Customer codes*

The above picture represents a count of unique customer codes. Out of a million observations, there are only approximately 626K unique customer codes. This means that there are around 370k duplicate observations, where there are 2 customer codes.

### High Cardinality / Disproportionate ratio of categories



*Fig 5: Countplot illustrating value counts of each category for `employee\_index` feature*

The above figure illustrates the value counts of each category for a chosen feature. The counts of some categories are negligible. Moreover, some features have a very high number of categories.

### Outliers/Skewed Features



*Fig 6: Distribution of Customer household income*

Some of the features are extremely skewed, as shown above. There are significant outliers. This can affect the clustering model. Hence, appropriate outlier engineering techniques must be utilized.

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

All of these approaches below ensure to preserve most of the useful observations within the dataset.

### Duplicated Observations

Most of the customer codes are repeated one time. We only need unique customer codes. Therefore, we will drop duplicated customer code observations and keep only the last observation unique for each customer code. We keep the last observation unique as the last observation for each customer code seems to be the latest entry.

### Missing Values

We will drop those features column-wise having more than 20% missing values. For the rest of the observations having missing values, we will consider them as Missing at Random, and try to impute those missing values using the information from other features.

For example, for the customer income feature (having around 17% missing values), we will impute this feature by aggregating the province name and taking the median household income for each province. Those observations having several missing values across most of the other columns are dropped row-wise as there is no useful information that can be used from other features.

For most categorical features with missing values, we will impute them with the mode category of that particular column. But for the province name column, which seems to have no name for a particular column, we will fill in the missing value for these observations as “Foreign”. We do this, because we noticed the missing values are for those observations that do not have a customer country residence category as ‘ES’ (Spain).

### Outlier

For the Age variable, we will replace ages below 20 with the mean of ages between 20 and 35. For ages above 85, we will replace ages with mean of ages between 35 and 85.

For the customer house income variable, we will apply BoxCox transformation, in order to transform its distribution to nearly normal distribution. The resulting outliers will then be winsorized at both ends. Winsorizing will help distribute the outliers at both ends closer to other values within the normal distribution.