Project: Customer Segmentation

Team member:

- MAMADI FOFANA
- mamadi.fofana@edu.dsti.institute
- Republic of Guinea
- Data Science Tech Institute (DSTI)
- Data Science Specialization

I. Problem description

Problem Statement: XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also, they will not be able to uncover the hidden pattern in the data (pattern which group certain kind of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want **more than 5 group** as this will be inefficient for their campaign.

Business understanding (Customer Segmentation)

We will propose customer segmentation approach to the Bank.

Our goal will be to create models which we use in turn to find a solution

We will adopt three predictive models:

- Clustering models which group similar behavior customers in one category and others in different category.
- Classification or Recommender models to predict group of a customer and which will be used to evaluate quality of the clustering model.

Project lifecycle along with deadline

1. Business Understanding (Week 7)

2. Data Understanding (week 8)

3. EDA (Week 8)

4. Feature Engineering (Week 9)

4. Model Building (Week 10)

5. Model Evaluation (Week 11)

6. Presentation (week 12)

7. Document the challenges (Week 13)

II. Data understanding

Data source

Data source used is: cust_seg.csv

Data source Link:
https://drive.google.com/drive/folders/1bfCpJIKmp6IHxiLPWvOS2nU1dc24pViB

Column Name	Description
fecha_dato	The table is partitioned for this column
ncodpers	Customer code
ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive
pais_residencia	Customer's Country residence
sexo	Customer's sex
age	Age
fecha_alta	The date in which the customer became as the first holder of a contract in the bank
ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.
antiguedad	Customer seniority (in months)
indrel	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 (if he isn't at the end of the month)
indrel_1mes	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, A (active), I (inactive), P (former customer),R (Potential)
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)

conyuemp	Spouse index. 1 if the customer is spouse of an employee
canal_entrada	channel used by the customer to join
indfall	Deceased index. N/S
tipodom	Addres type. 1, primary address
cod_prov	Province code (customer's address)
nomprov	Province name
ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)
renta	Gross income of the household
ind_ahor_fin_ult1	Saving Account
ind_aval_fin_ult1	Guarantees
ind_cco_fin_ult1	Current Accounts
ind_cder_fin_ult1	Derivada Account
ind_cno_fin_ult1	Payroll Account
ind_ctju_fin_ult1	Junior Account
ind_ctma_fin_ult1	Más particular Account
ind_ctop_fin_ult1	particular Account
ind_ctpp_fin_ult1	particular Plus Account
ind_deco_fin_ult1	Short-term deposits
ind_deme_fin_ult1	Medium-term deposits
ind_dela_fin_ult1	Long-term deposits
ind_ecue_fin_ult1	e-account
ind_fond_fin_ult1	Funds
ind_hip_fin_ult1	Mortgage
ind_plan_fin_ult1	Pensions
ind_pres_fin_ult1	Loans
ind_reca_fin_ult1	Taxes
ind_tjcr_fin_ult1	Credit Card
ind_valo_fin_ult1	Securities
ind_viv_fin_ult1	Home Account
ind_nomina_ult1	Payroll
ind_nom_pens_ult1	Pensions
ind_recibo_ult1	Direct Debit

Type of data

Data source contains several sources of data:

- Numerical: integer and float
- Categorical:

Sweetviz (EDA Tool) overview

1000000	ROWS
0	DUPLICATES
1.1 GB	RAM
47	FEATURES
39	CATEGORICAL
3	NUMERICAL
5	TEXT

Pandas overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 47 columns):
 #    Column                  Non-Null Count    Dtype
---   ------                  --------------    -----
 0    fecha_dato              1000000 non-null  object
 1    ncodpers                1000000 non-null  int64
 2    ind_empleado            989218 non-null   object
 3    pais_residencia         989218 non-null   object
 4    sexo                    989214 non-null   object
 5    age                     1000000 non-null  object
 6    fecha_alta              989218 non-null   object
 7    ind_nuevo               989218 non-null   float64
 8    antiguedad              1000000 non-null  object
 9    indrel                  989218 non-null   float64
 10   ult_fec_cli_1t          1101 non-null     object
 11   indrel_1mes             989218 non-null   float64
 12   tiprel_1mes             989218 non-null   object
 13   indresi                 989218 non-null   object
 14   indext                  989218 non-null   object
 15   conyuemp                178 non-null      object
 16   canal_entrada           989139 non-null   object
 17   indfall                 989218 non-null   object
 18   tipodom                 989218 non-null   float64
 19   cod_prov                982266 non-null   float64
 20   nomprov                 982266 non-null   object
 21   ind_actividad_cliente   989218 non-null   float64
 22   renta                   824817 non-null   float64
 23   ind_ahor_fin_ult1       1000000 non-null  int64
```

```
 24  ind_aval_fin_ult1      1000000 non-null  int64
 25  ind_cco_fin_ult1       1000000 non-null  int64
 26  ind_cder_fin_ult1      1000000 non-null  int64
 27  ind_cno_fin_ult1       1000000 non-null  int64
 28  ind_ctju_fin_ult1      1000000 non-null  int64
 29  ind_ctma_fin_ult1      1000000 non-null  int64
 30  ind_ctop_fin_ult1      1000000 non-null  int64
 31  ind_ctpp_fin_ult1      1000000 non-null  int64
 32  ind_deco_fin_ult1      1000000 non-null  int64
 33  ind_deme_fin_ult1      1000000 non-null  int64
 34  ind_dela_fin_ult1      1000000 non-null  int64
 35  ind_ecue_fin_ult1      1000000 non-null  int64
 36  ind_fond_fin_ult1      1000000 non-null  int64
 37  ind_hip_fin_ult1       1000000 non-null  int64
 38  ind_plan_fin_ult1      1000000 non-null  int64
 39  ind_pres_fin_ult1      1000000 non-null  int64
 40  ind_reca_fin_ult1      1000000 non-null  int64
 41  ind_tjcr_fin_ult1      1000000 non-null  int64
 42  ind_valo_fin_ult1      1000000 non-null  int64
 43  ind_viv_fin_ult1       1000000 non-null  int64
 44  ind_nomina_ult1        994598 non-null   float64
 45  ind_nom_pens_ult1      994598 non-null   float64
 46  ind_recibo_ult1        1000000 non-null  int64
dtypes: float64(9), int64(23), object(15)
memory usage: 358.6+ MB
```

Problems in the data

After descriptive analysis (univariate analysis), correlation analysis (bivariate analysis), we notice following:

- **Missing value in some fields**

    We have a total of **2 371 207** missing values as per column

    - 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

For missing value in categorical columns, we replace missing value with a new category

For missing value in numerical values, we replace missing value with mean or median value of the columns

- **Duplicate values**

No duplicated values were found before data preprocessing, but found out duplicated values after removing some columns

Those duplicated value has been removed

- **High percentage of missing values in some columns in features:**

We dropped columns with high percentage of missing value which cannot decrease performance of the model

  - ult_fec_cli_1t
  - conyuemp

- **Column with unique categorical value**

We remove column with unique categorical value since it brings no insight to the model

  - tipodom

- **Low variance in some numeric attributes**

So far, no variance issue has been observed in the numerical column

- **Low entropy in some categorical attributes.**

The identified problem is a very small entropy of the feature, which means that most of the records have the same categorical values. In most cases, it is safe to remove categorical attributes with low entropy. This will not harm the performance of the model, and it can reduce the complexity of the model.

  - ind_nuevo
  - indrel
  - indrel_1mes
  - indresi

- indfall
- ind_ahor_fin_ult1
- ind_aval_fin_ult1
- ind_cder_fin_ult1
- ind_deco_fin_ult1
- ind_deme_fin_ult1
- ind_pres_fin_ult1
- ind_viv_fin_ult1
- ind_hip_fin_ult1
- ind_plan_fin_ult1
- ind_valo_fin_ult1
- ind_fond_fin_ult1
- ind_ctma_fin_ult1
- ind_ctju_fin_ult1
- indext
- ind_empleado
- pais_residencia

- **high correlated columns (*cor* with ind_nom_pens_ult1 >0.8)**

We remove columns highly correlated with others (we choose abs(correlation) > 0.8)

- ind_cno_fin_ult1
- ind_nomina_ult1
- fecha_alta

- **Columns with high cardinality**

The identified problem occurs when the number of unique values is too large for categorical attributes. This high cardinality creates problems for the typical one-hot-encoding process, creating a representation in an extremely high-dimensional space

Removing the feature if the number of occurrences per unique value of the attributes is too low

- ncodpers
- fecha_alta

- **Data in wrong format**

Some data are in wrong format need to be converted in the right format

- Age
- Antiguedad

- **Outliers and errors management**

Field *antiguedad* contains an errors value (-999 999) which need to be replaced by the median value of the field

- **Data not scaled for some numerical variable**

For the clustering algorithm to consider all numerical attributes as equal, they must all have the same scale

Moreover, in order to comply to some assumptions of analysis and to reduce skewness, we applied gaussian transformation to those variables

- Age
- Antiguedad
- Renta

We have to scale those columns using Z-score and log transformation

III. Exploratory Data Analysis (EDA) performed on the data

After EDA performed of data, we finally got this new this data structure:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 961479 entries, 0 to 961478
Data columns (total 18 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   fecha_dato             961479 non-null  object
 1   sexo                   961479 non-null  object
 2   age                    961479 non-null  float64
 3   antiguedad             961479 non-null  float64
 4   tiprel_1mes            961479 non-null  object
 5   canal_entrada          961479 non-null  object
 6   nomprov                961479 non-null  object
 7   ind_actividad_cliente  961479 non-null  int32
 8   renta                  961479 non-null  float64
 9   ind_cco_fin_ult1       961479 non-null  int64
 10  ind_ctop_fin_ult1      961479 non-null  int64
 11  ind_ctpp_fin_ult1      961479 non-null  int64
 12  ind_dela_fin_ult1      961479 non-null  int64
 13  ind_ecue_fin_ult1      961479 non-null  int64
 14  ind_reca_fin_ult1      961479 non-null  int64
 15  ind_tjcr_fin_ult1      961479 non-null  int64
 16  ind_nom_pens_ult1      961479 non-null  int32
 17  ind_recibo_ult1        961479 non-null  int64
dtypes: float64(3), int32(2), int64(8), object(5)
memory usage: 124.7+ MB
```

Data set transformed is located in the file: ***cust_seg_Updated***

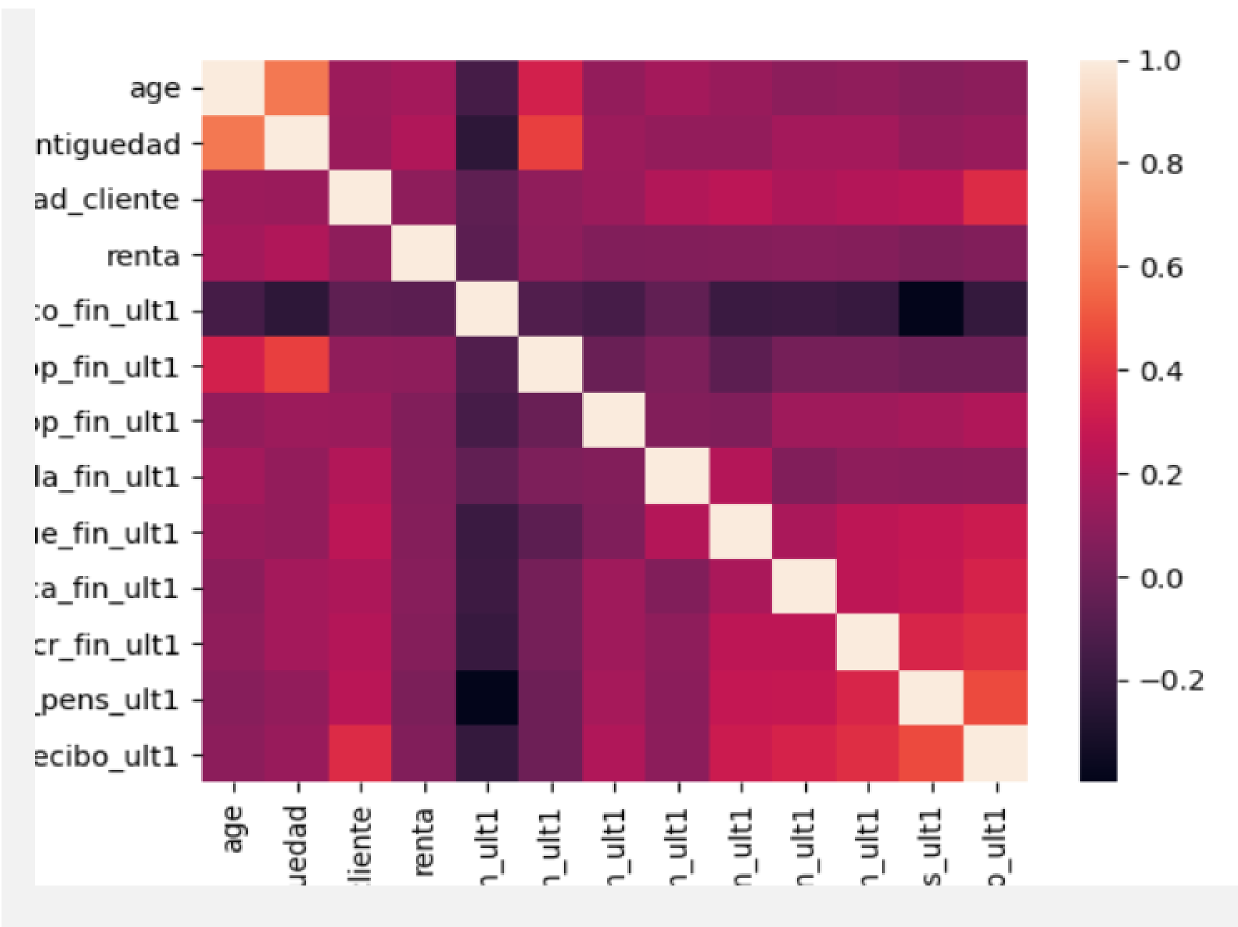
Automated exploratory data analysis for this new data set could be found in the file:
EDA_Cust_Fin.html

Correlation analysis

Numerical correlation

	age	antigueda	ind_activic	renta	ind_cco_fi	ind_ctop_f	ind_ctpp_f	ind_dela_f	ind_ecue_	ind_reca_f	ind_tjcr_fi	ind_nom_p	ind_recibo_
age	1	0.602371	0.146407	0.173596	-0.14495	0.32942	0.11987	0.173536	0.134264	0.095878	0.1062	0.082601	0.096204
antigueda	0.602371	1	0.139859	0.208611	-0.23161	0.441233	0.143707	0.120671	0.116528	0.171683	0.171289	0.114477	0.133276
ind_activic	0.146407	0.139859	1	0.10007	-0.05822	0.108476	0.141541	0.218361	0.251626	0.201941	0.220466	0.241334	0.372467
renta	0.173596	0.208611	0.10007	1	-0.06705	0.100847	0.060196	0.064223	0.069446	0.077423	0.067651	0.044455	0.056074
ind_cco_fi	-0.14495	-0.23161	-0.05822	-0.06705	1	-0.10264	-0.13634	-0.04728	-0.17945	-0.17292	-0.19257	-0.39773	-0.20332
ind_ctop_f	0.32942	0.441233	0.108476	0.100847	-0.10264	1	-0.01964	0.045601	-0.06203	0.020359	0.021212	-0.00975	-0.00702
ind_ctpp_f	0.11987	0.143707	0.141541	0.060196	-0.13634	-0.01964	1	0.065078	0.051241	0.158959	0.158365	0.176897	0.21122
ind_dela_f	0.173536	0.120671	0.218361	0.064223	-0.04728	0.045601	0.065078	1	0.222399	0.056459	0.100476	0.092534	0.094016
ind_ecue_	0.134264	0.116528	0.251626	0.069446	-0.17945	-0.06203	0.051241	0.222399	1	0.19003	0.253194	0.277997	0.302286
ind_reca_f	0.095878	0.171683	0.201941	0.077423	-0.17292	0.020359	0.158959	0.056459	0.19003	1	0.256445	0.280134	0.343754
ind_tjcr_fi	0.1062	0.171289	0.220466	0.067651	-0.19257	0.021212	0.158365	0.100476	0.253194	0.256445	1	0.355023	0.384043
ind_nom_p	0.082601	0.114477	0.241334	0.044455	-0.39773	-0.00975	0.176897	0.092534	0.277997	0.280134	0.355023	1	0.472749
ind_recibo	0.096204	0.133276	0.372467	0.056074	-0.20332	-0.00702	0.21122	0.094016	0.302286	0.343754	0.384043	0.472749	1

Numerical correlation could be found in the file ***corr_Fin.csv***
We notice a weak correlation between features



Relationship between two categorical features

Chi-Square Test for independence between 'tiprel_1mes' and ‘ind_actividad_cliente'

Null hypothesis: ‘features tiprel_1mes’ and ‘ind_actividad_cliente’ are independent

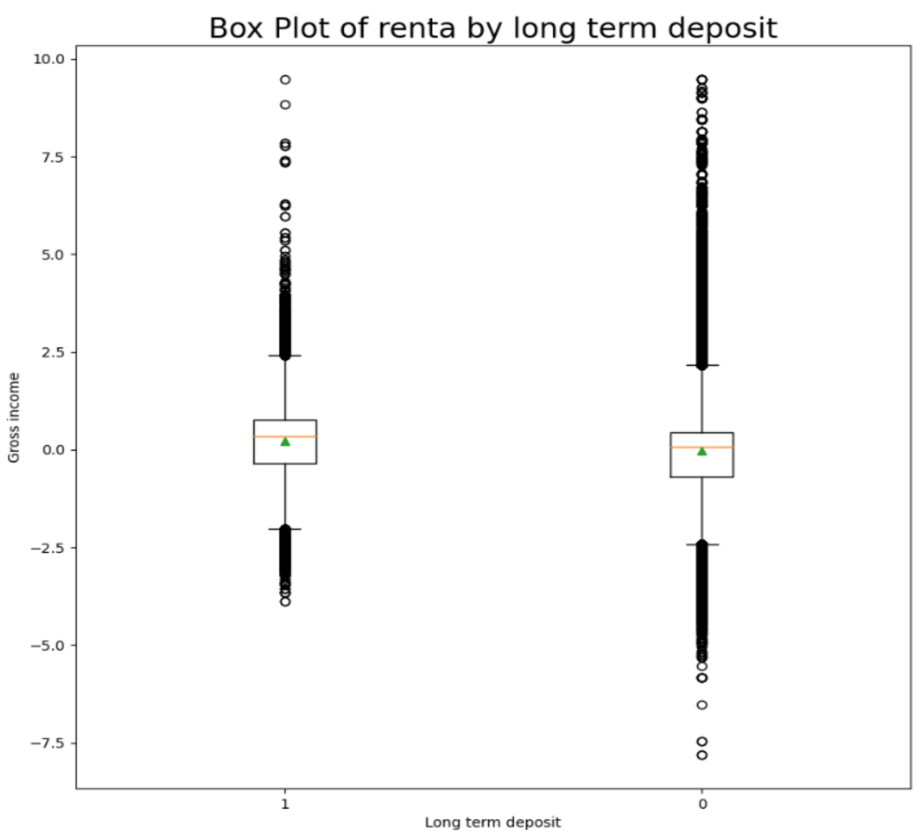
pValue equal 0 which means that we can reject null hypothesis:
So ‘tiprel_1mes’ and ‘ind_actividad_cliente’ are strongly dependent

Categorical and numerical relationship

Anova test between feature 'renta' and 'ind_dela_fin_ult1'

We need to test if it exists a relationship between long term deposit and Gross income of the household
The null hypothesis is the mean of gross income between category of long-term deposit is equal

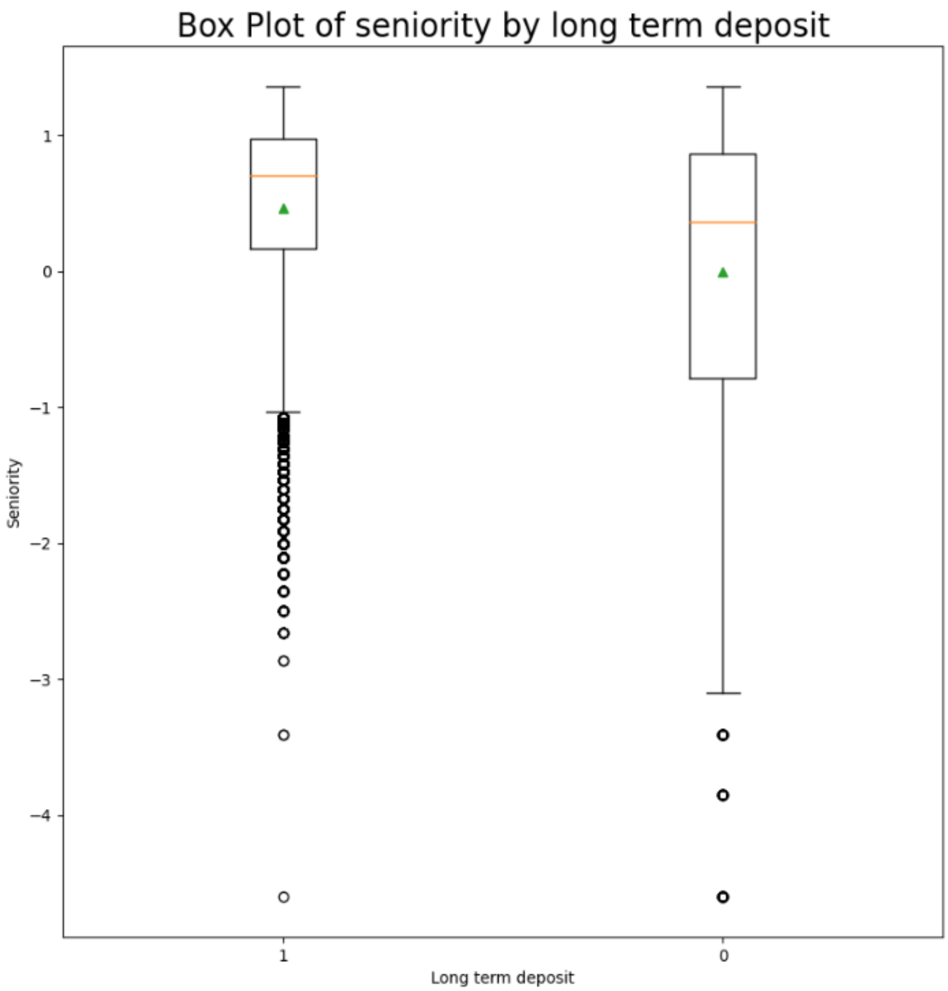
pValue equal 0 which means that we can reject null hypothesis:
So, long term deposit could be related to gross income
We can observe this fact also in the Boxplot



Anova test between feature 'antiguedad' and 'ind_dela_fin_ult1'

We need to test if it exists a relationship between long term deposit and customer seniority The null hypothesis is the mean of customer seniority between category of long-term deposit is equal

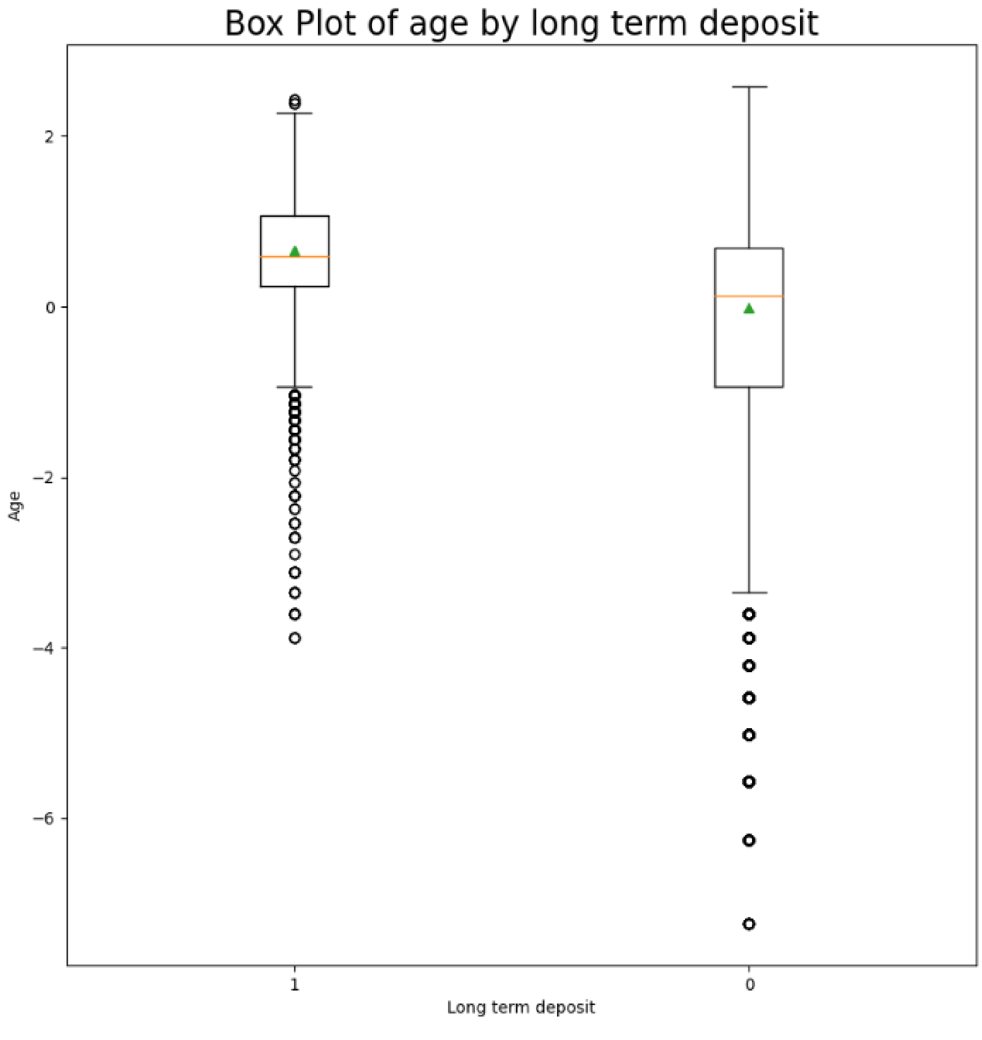
pValue equal 0 which means that we can reject null hypothesis:
So, long term deposit could be related to customer seniority
We can observe this fact also in the Boxplot



Anova test between feature 'age' and 'ind_dela_fin_ult1'

We need to test if it exists a relationship between long term deposit and age
The null hypothesis is the mean of age between category of long-term deposit is equal

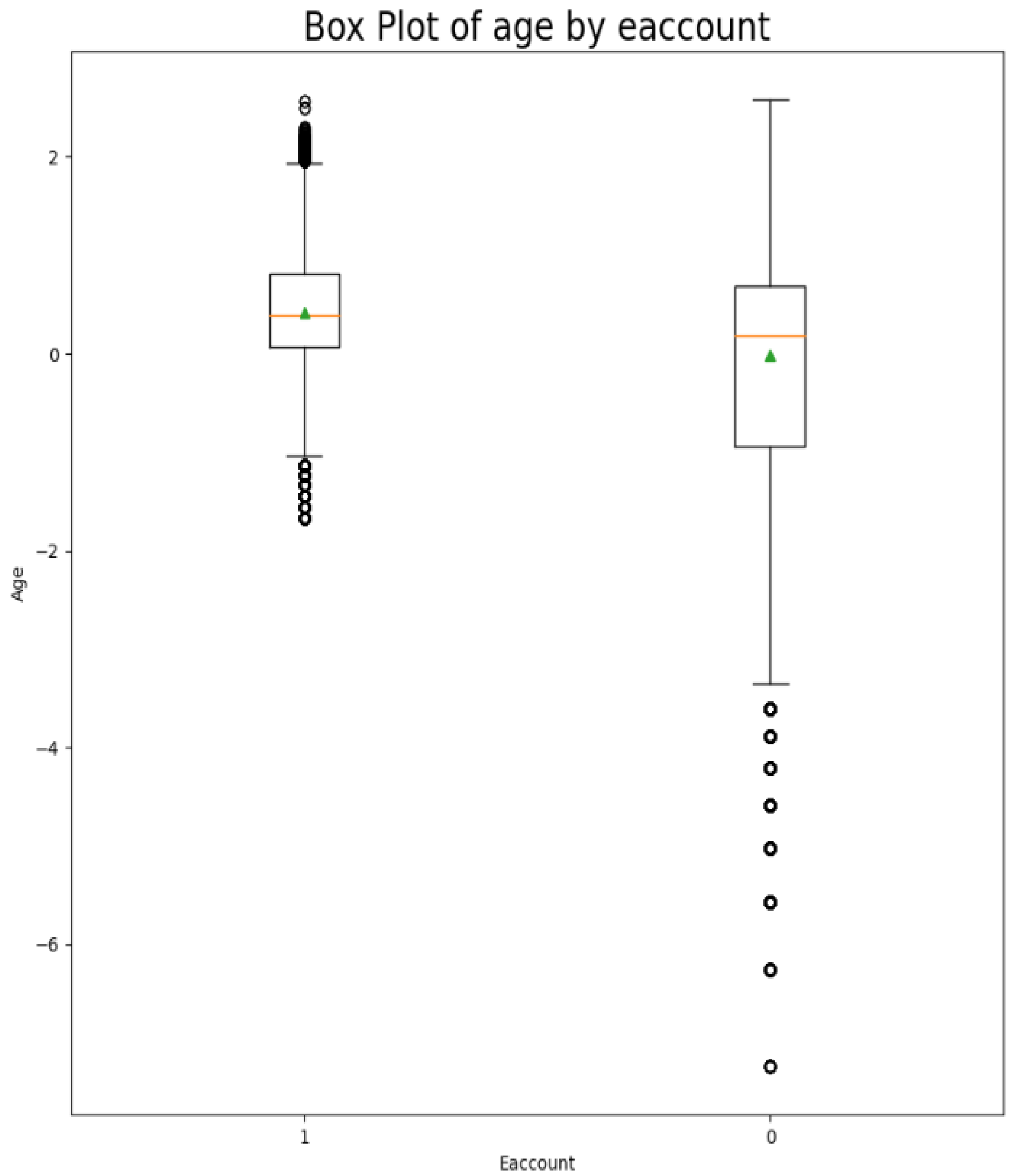
pValue equal 0 which means that we can reject null hypothesis:
So, long term deposit could be related to customer age
We can observe this fact also in the Boxplot



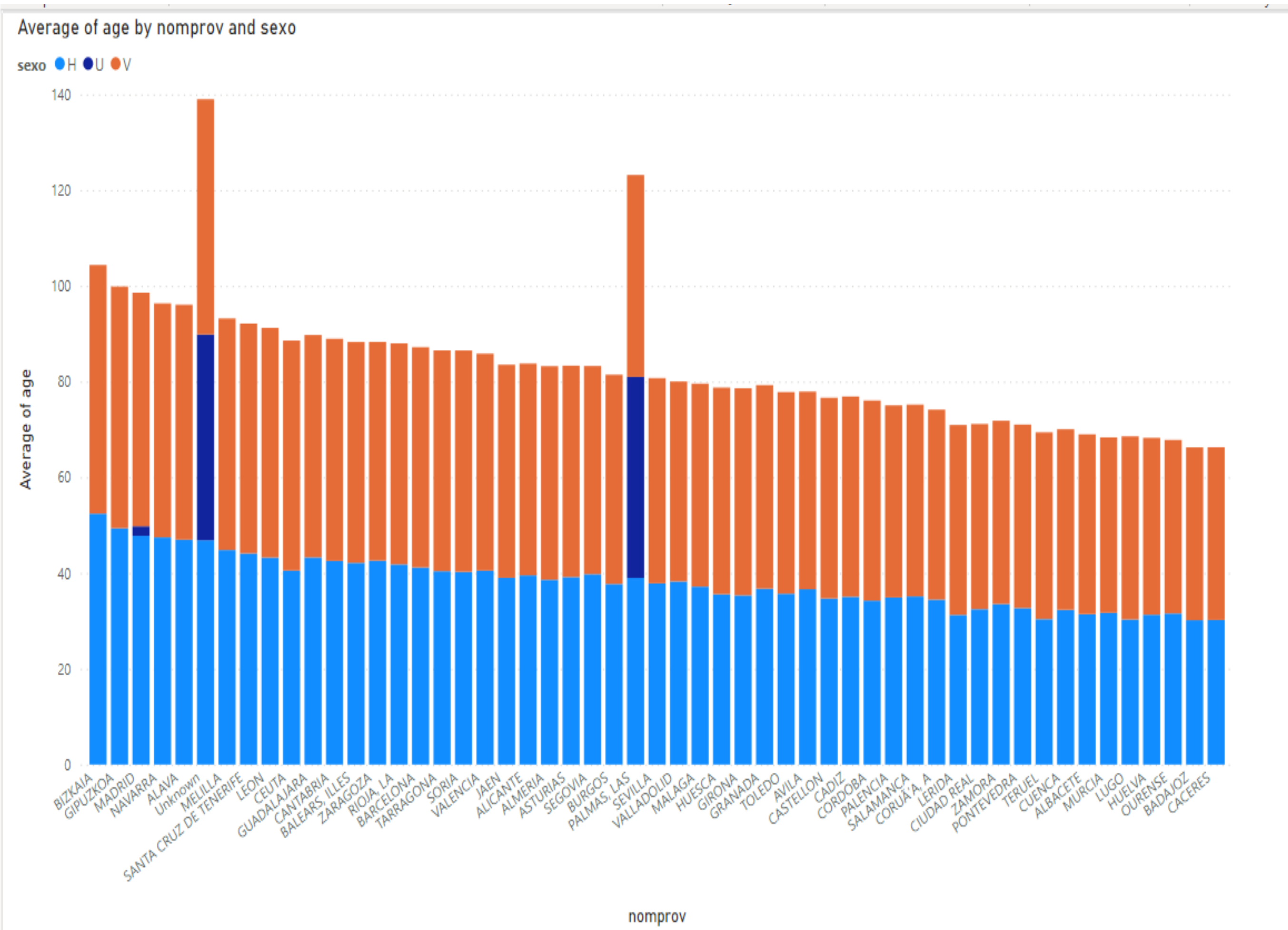
Anova test between feature 'age' and 'ind_ecue_fin_ult1'

We need to test if it exists a relationship between eaccount and age
The null hypothesis is the mean of age between category of eaccount is equal

pValue equal 0 which means that we can reject null hypothesis:
So having eaccount could be related to age
We can observe this fact also in the Boxplot



Average age by province and gender

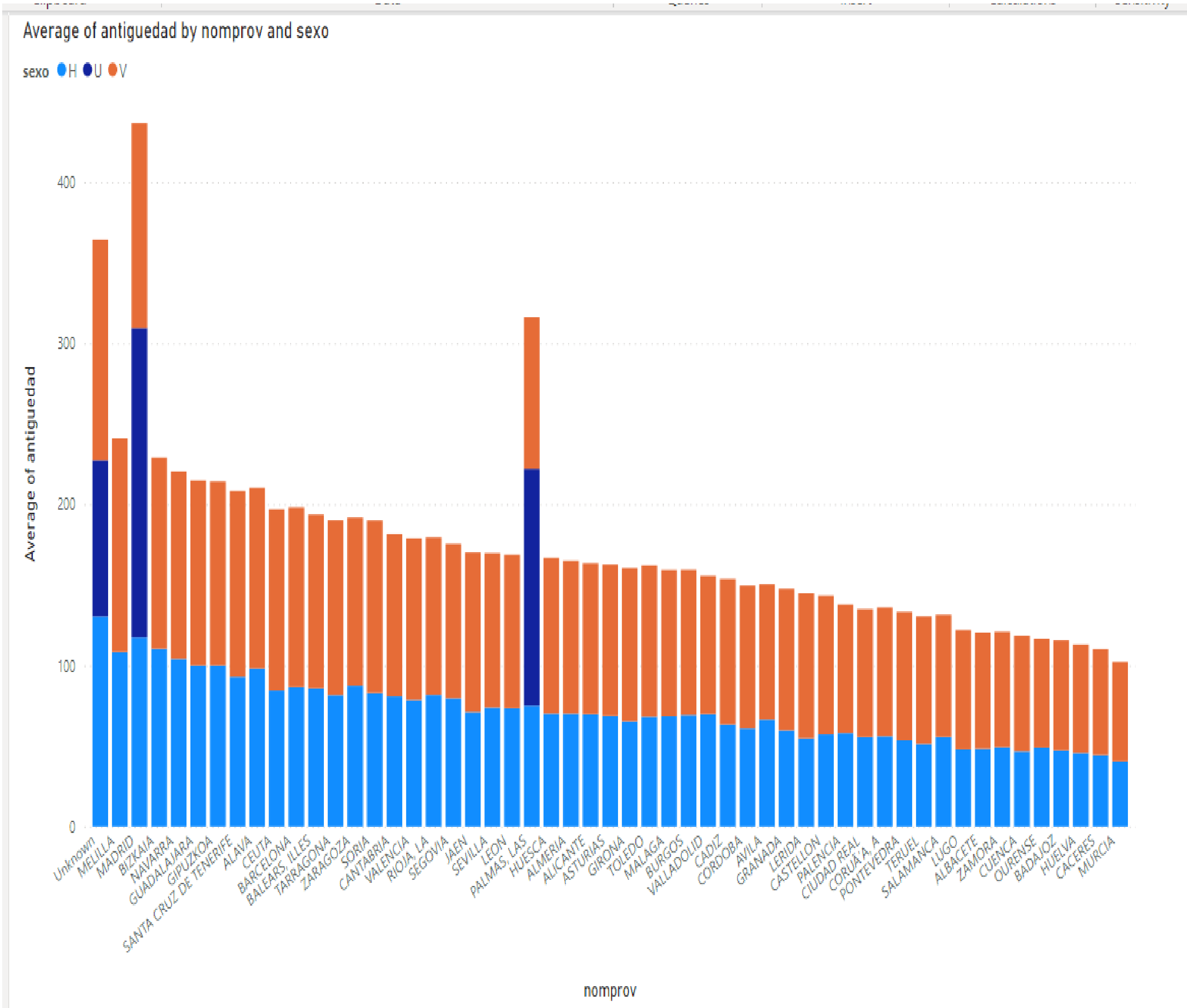


We noticed that average age is quite similar for both gender for most of provinces (nomprov) except for few provinces.

Only province ‘Las palmas ‘has a significant missing value for age variable

Also, a new category (Unknown) has been created to impute missing value for variable ‘nomprov’

Average seniority by province and gender

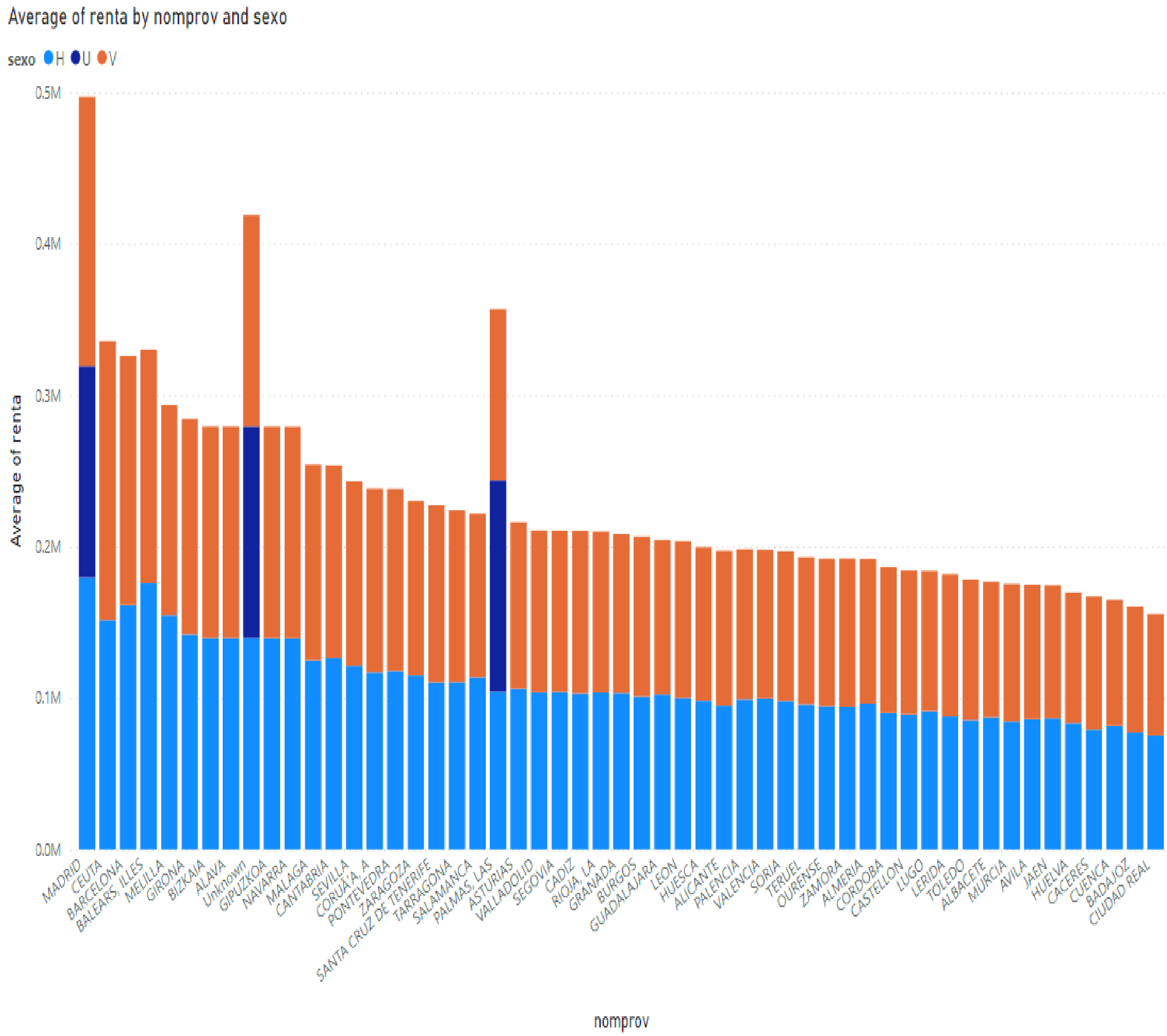


We noticed that average seniority is quite similar for both gender for most of provinces (nomprov) except for few provinces

Only province 'Las palmas ' and 'madrid' have a significant missing value for seniority variable

Also, a new category (Unknown) has been created to impute missing value for variable 'nomprov'

Average gross income by province and gender



We noticed that average seniority is quite similar for both gender for most of provinces (nomprov) except for few provinces (ie 'madrid' , 'balnear iles' etc.)

Only province 'Las palmas ' and 'madrid' have a significant missing value for gross income variable

Also, a new category (Unknown) has been created to impute missing value for variable 'nomprov'

Recommendations

We can reduce number of study variables since most of them are not necessary to the analysis

We need to improve data collect in some provinces, especially for ‘madrid’ and ‘Las palmas’ since those variables contains a lot of missing data

We have tested 5 hypotheses with following conclusions:

- Customer relationship at the beginning of the month ‘tiprel_1mes’ and customer activation ‘ind_actividad_cliente’ are strongly dependent

- Long term deposit could be related to gross income

- Long term deposit could be related to customer seniority

- Long term deposit could be related to customer age

- Having eaccount could be related to age