Project: Customer Segmentation

Week 7: Deliverables

Team Member Details:

- Brennan Clinch, bclincher98@gmail.com, USA, North Carolina State University, Data Science
- Rohit Sunku, rgs8890@gmail.com, UK, Le Wagon, Data Science
- Kutay Selçuk, kutay.selcuk@ozu.edu.tr, Turkey, Ozyegin University, Data Science
- Zhan Shi, zhanshi@g.ucla.edu, USA, University of California Los Angeles, Data Science

Problem Description:

XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out the same offer to all customers. Instead, they want to roll out personalized offers to a particular set of customers. If they manually start understanding the category of the 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 groups** as this will be inefficient for their campaign.

ABC analytics assigned this talk to their analytics team and instructed their team to come up with the approach and feature which group similar behavior in one category and others in a different category.

Data Understanding

Numeric variables: *age, antiguedad, renta*

- One thing to note here is that *age* and *antiguedad* were originally read in as object variables, but are really numeric variables. So when data cleaning we will have to change variables to numeric.

Categorical Variables: *Everything else*

- It was also found when looking at the data that the following variables were read in as float64: *ind_nuevo*, *indrel*, *indrel_1mes*, *tipodom*, *cod_prov*, *ind_actividad_cliente*, *ind_nomina_ult1*, *ind_nom_pens_ult1*.

- We will need to convert to categorical when cleaning the data.

Problems with the data

Missing Values

· There are null values in the file as they are mentioned below.

Column Name	Number of NaN
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

- From the table above, it appears that we could safely remove the missing values from the categorical variables since most columns don't have more than 2% of missing data.

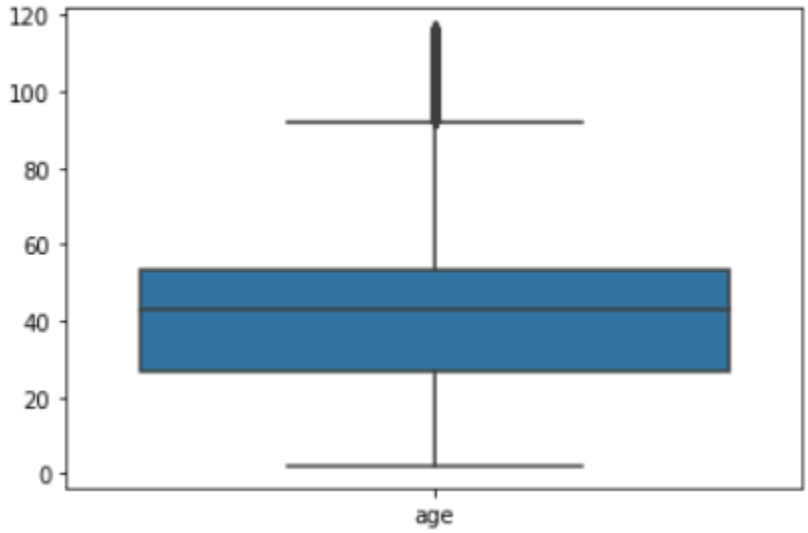
- For continuous variables like renta, it is an important variable for us to have in the analysis so for missing values removal of them would be too much so maybe we could try replacing them with a median/mean value.

- For variables with missing values greater than 50% of the data, we could safely remove them especially since there are a couple and they are not continuous but categorical.

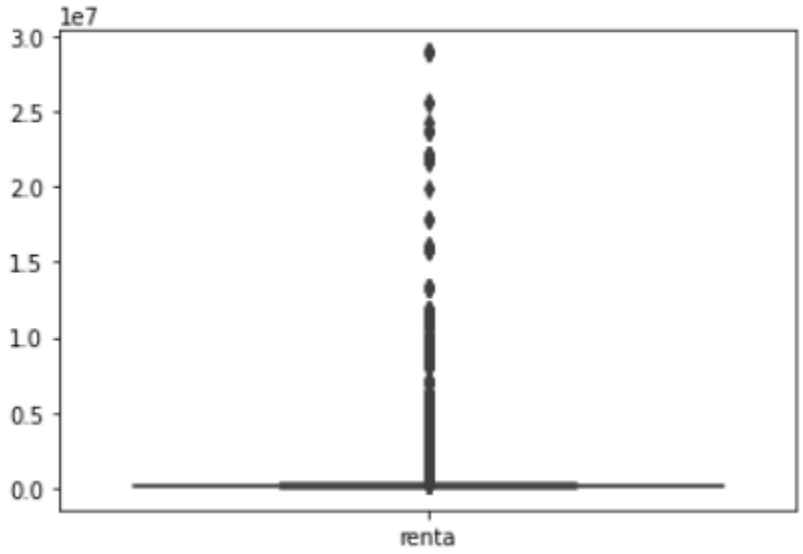
ind_empleado	1.08
pais_residencia	1.08
sexo	1.08
age	0.00
fecha_alta	1.08
ind_nuevo	1.08
antiguedad	0.00
indrel	1.08
ult_fec_cli_1t	99.89
indrel_1mes	1.08
tiprel_1mes	1.08
indresi	1.08
indext	1.08
conyuemp	99.98
canal_entrada	1.09
indfall	1.08
tipodom	1.08
cod_prov	1.77
nomprov	1.77
ind_actividad_cliente	1.08
renta	17.52
ind_nomina_ult1	0.54
ind_nom_pens_ult1	0.54

Outliers

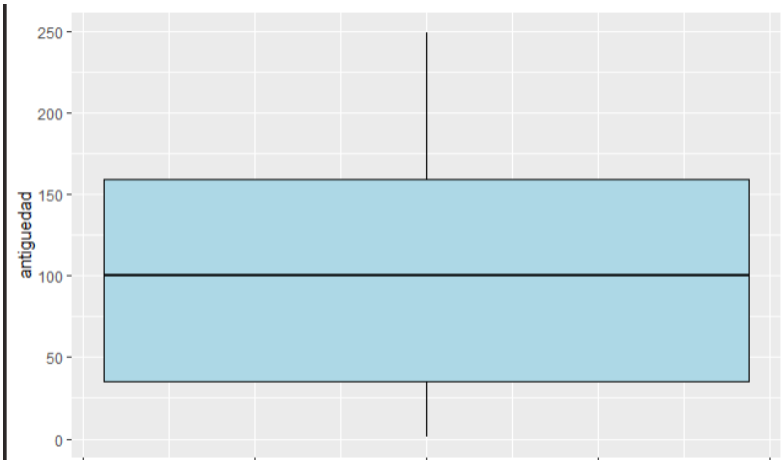
Age: We can see that there are outliers in the data around the age of 90, which can be expected since that is an old age. There is strong evidence of skewness that could be caused by the outliers.



renta: There are obviously many outliers in this variable, which in turn really makes the data very skewed and hard to follow the distribution outside of the outliers.



antiguedad: We can see that there aren't any outliers in the variable. So no cleaning other than changing the type of object when using



Approach for problems with data

Missing value removal/replacing

For missing/NA values, we see that some columns are mostly filled with NA values while some others have less NA values. It would be a bad choice to remove the NA values from the columns that contain more than half the data as that would really mess up our interpretability of the analysis. Fortunately, there are only a couple columns where this is the case. From the analysis on the null values that was studied above, we can see that we could easily remove null values

for a lot of our categorical variables since they only contain around 17% of the total data overall. But since one of the important numeric variables, renta, contains about 20% null values, it would be better if we could first remove all the null values from most categorical variables that contain around 1% of null values. After removing those values, with the remaining null values that were used in variables such as renta, we could try replacing some of the values with the mean or the median of the variable after removal of any outliers. This could not mess up the data too much depending on how many outliers there are. One other solution would maybe be to remove the variables who have more than 50% null values such as conyump or ult_fec_cli_lt.

Removal of outliers

2 of our continuous variables have a lot of outliers (renta and age), which add to those variable's skewness as well. One possible way we could handle outliers for age would be if we remove all data of members who are older than 90 along with starting with members who are 18 or older. For outliers for the renta variable, we might have to remove highly significant ones (5 *IQR) but work on replacing the null values with mean/median/etc.