Project: Customer Segmentation

Week 9: 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.

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.

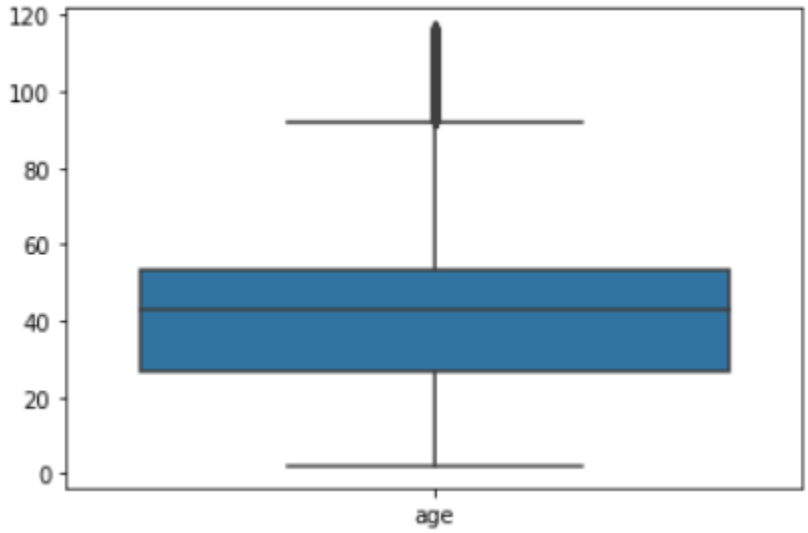
- For continuous variables like renta, it is an important variable for us to have in the analysis so for missing values removal of them would be too much so maybe we could try replacing them with a median/mean value.

- 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.

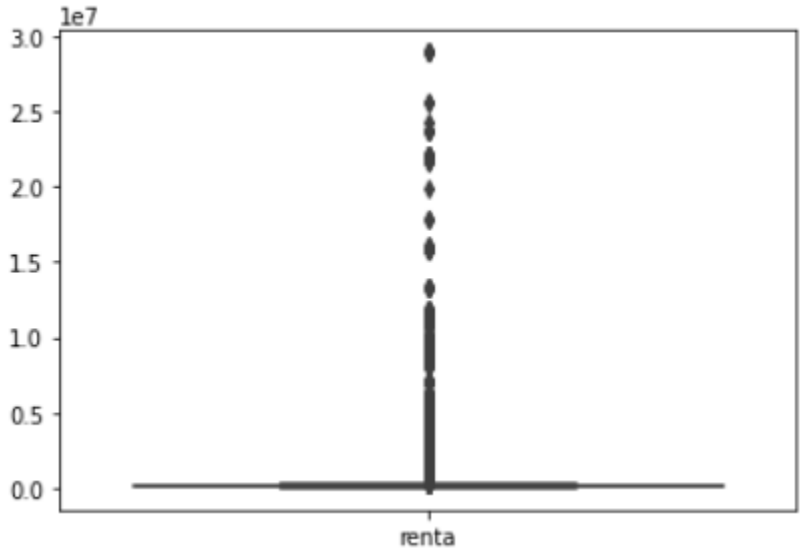
ind_empleado	1.08
pais_residencia	1.08
sexo	1.08
age	0.00
fecha_alta	1.08
ind_nuevo	1.08
antiguedad	0.00
indrel	1.08
ult_fec_cli_1t	99.89
indrel_1mes	1.08
tiprel_1mes	1.08
indresi	1.08
indext	1.08
conyuemp	99.98
canal_entrada	1.09
indfall	1.08
tipodom	1.08
cod_prov	1.77
nomprov	1.77
ind_actividad_cliente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

Outliers

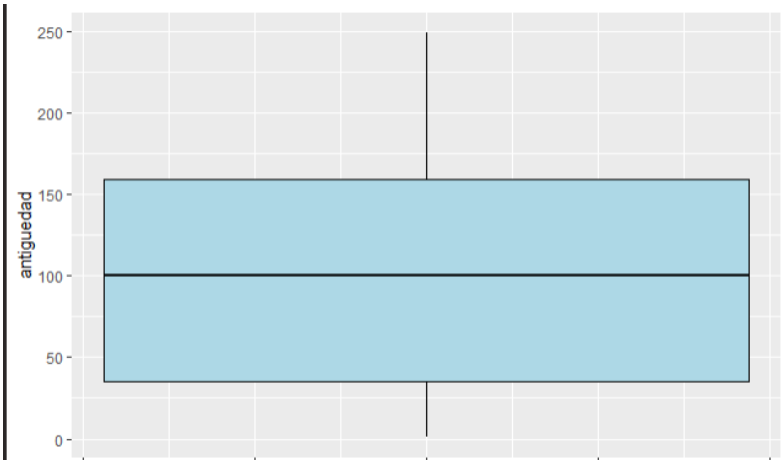
Age: We can see that there are outliers in the data around the age of 90, which can be expected since that is an old age. There is strong evidence of skewness that could be caused by the outliers.



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.

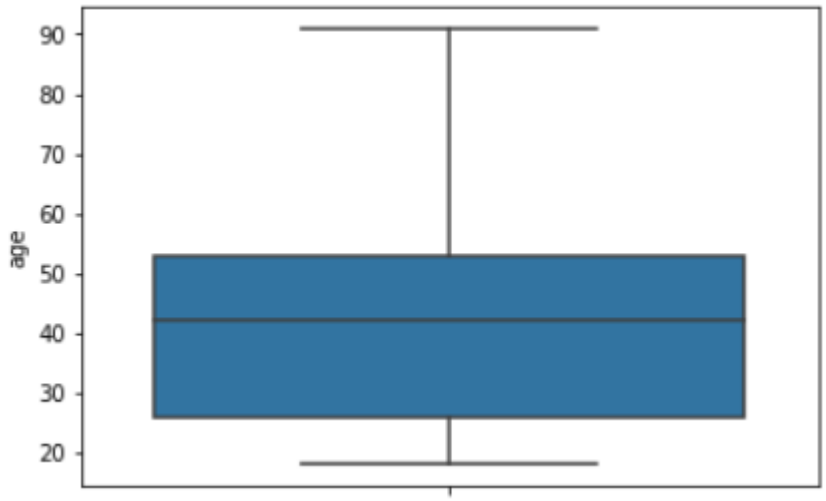
Removal of outliers

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.

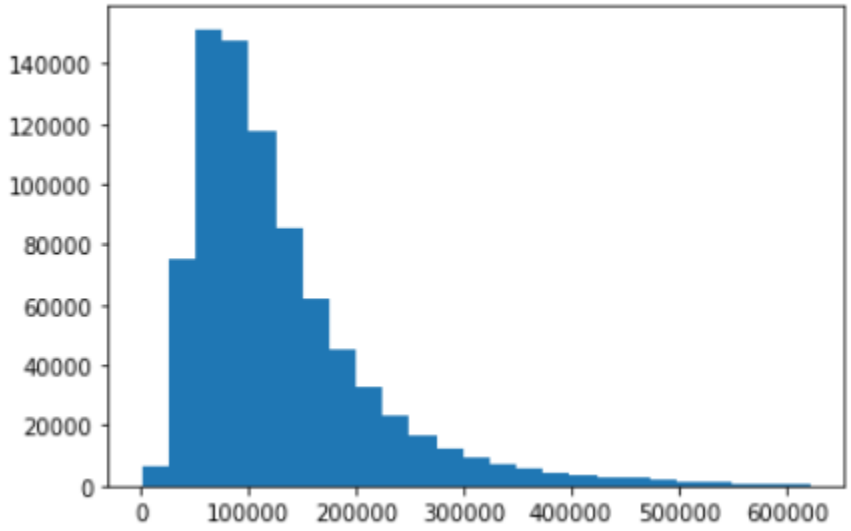
Data Cleaning (week 9)

Outlier removal for age and renta

As noted earlier, for the removal of outliers for the "age" variable, we will subset our data to only include values for the age ranging from 18-90 since young adults are the most popular demographic. After filtering, we can see our distribution looks a lot nicer with only a slight skew.



For renta, we removed all outliers that were greater than 5 * IQR. We still have some skewness in the data, but that is ok since removal of all outliers could impact our results, so we only got rid of the extreme ones.



Missing data removal/replacement

Removal of variables

One other thing that was mentioned before was the removal and replacement of NA values. To start, we removed the variables “conyuemp” and “ult_fec_cli_1t” because they have 99% missing values, along with that they aren’t very important for the analysis since they just contain info on whether the customer was a spouse of the employee.

Approach 1 for renta

For “renta”, after outlier removal, we have came up with two approaches for handling the missing data, the first one is to replace all NA values with the median value for “renta”. After replacement of the values with the median (105990), we saw that it didn’t change the overall shape of distribution at all since the mean stayed the same along with the other important values that make up the variable’s five number summary.

Approach 2 for renta

The second approach was to find another variable in the dataset that was categorical and replace the NA values with the mean/median of the most commonly occurring category of the variable when filtering for the gross household income (“renta”). The variable that would be most suitable for this that we chose was “nomprov” which listed the province of the customer. We saw from looking at the distribution of the province that Madrid was the most commonly occurring province so we filtered the renta data for that province and found the mean value to be 164013 and the median to be 141383. By replacing the NA values with 141383, we get the

same mean and median values as we did before replacing the NA values. So either way would work fine as it won't really affect the distribution too much, but we decided to replace the NA values with the median from the values of renta pertaining to customers that reside in Madrid.

Other variables

nomprov

As far as other NA values go, we replaced the values for nomprov with "Madrid" since it is the most occurring province along with it only having 1% of missing data.

canal_entrada

For this variable, which is the channel used by the customer to join, we once again replaced all null values with the mode which was "KAT", which replaced 1% of the data that was missing.

ind_nomina_ult1 & ind_nom_pens_ult1

The variable "ind_nomina_ult1" shows whether or not a customer has Payroll. The variable "ind_nom_pens_ult1" shows us whether or not the customer has Pensions. We replaced the null values with 0 for ind_nomina_ult1 and ind_nom_pens_ult1, which indicates that the customer does not have Payroll or Pensions.