Exploratory Data Analysis

Customer Segmentations

10th Aug 2021

Background – Customer Segmentation

- **Objective:** 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.
- **Contents of this Presentation**
  - Data Exploration and Cleaning
  - Analysis
  - Findings
  - Recommendations

Data Exploration

- Final Dataset
  - 85023 observations
  - 9 columns
- Actions Performed
  - Handling outlies and null values
    - Z Score Method
    - Mean method
  - Changing Appropriate dtypes
  - Changing column names and observations from Spanish to readable English
  - Selected appropriate features for modeling

NA Values and Outliers

- Few observations which have null values are overlapping, i.e same records have missing values in all columns have been removed.
- The columns ult_fec_cli_1t and conyuemp have 99% null values so the best action would be to drop these two columns. The renta column represents the gross income of the family, and has about 17% of null values, these null values can be replaced by the average of column. There are null values in other columns but they amount to around 1% of the dataset so I have decided to remove them from the dataset.
- Replaced nulls in renta column with average value
- Outliers in gross income and age column have been identified using a box plot and have been removed based on their z-score.
- Features are selected based upon the relevance of the column and the type of data.Most of the binary data is removed only important columns are retained as binary data doesnt work well with Clustering algorithms.

NA Values

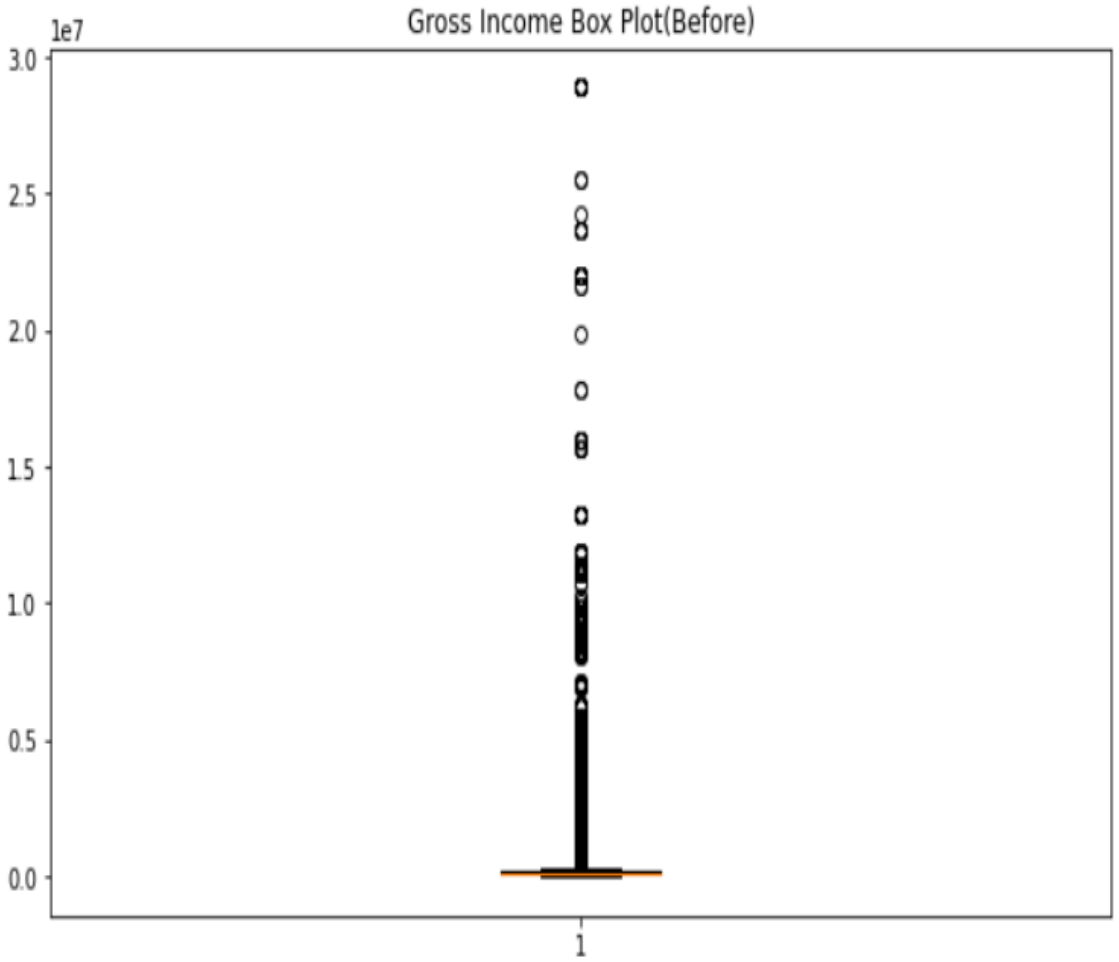
fecha_dato	0
ncodpers	0
ind_empleado	10782
pais_residencia	10782
sexo	10786
age	0
fecha_alta	10782
ind_nuevo	10782
antiguedad	0
indrel	10782
ult_fec_cli_1t	998899
indrel_1mes	10782
tiprel_1mes	10782
indresi	10782
indext	10782
conyuemp	999822
canal_entrada	10861
indfall	10782
tipodom	10782
cod_prov	17734
nomprov	17734
ind_actividad_cliente	10782
renta	175183

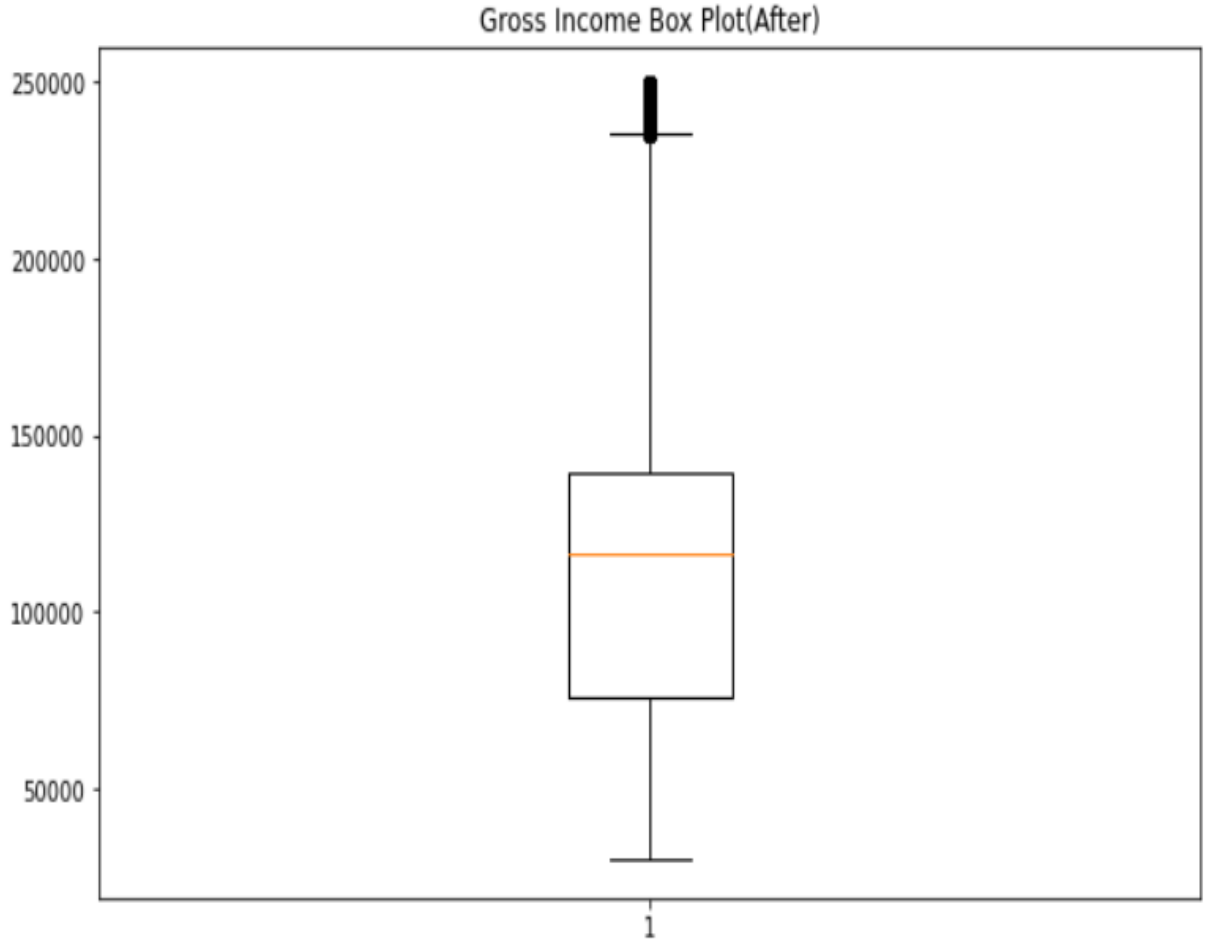
Before

fecha_dato	0
ncodpers	0
ind_empleado	0
pais_residencia	0
sexo	0
age	0
fecha_alta	0
ind_nuevo	0
antiguedad	0
indrel	0
indrel_1mes	0
tiprel_1mes	0
indresi	0
indext	0
canal_entrada	0
indfall	0
tipodom	0
cod_prov	0
nomprov	0
ind_actividad_cliente	0
renta	0

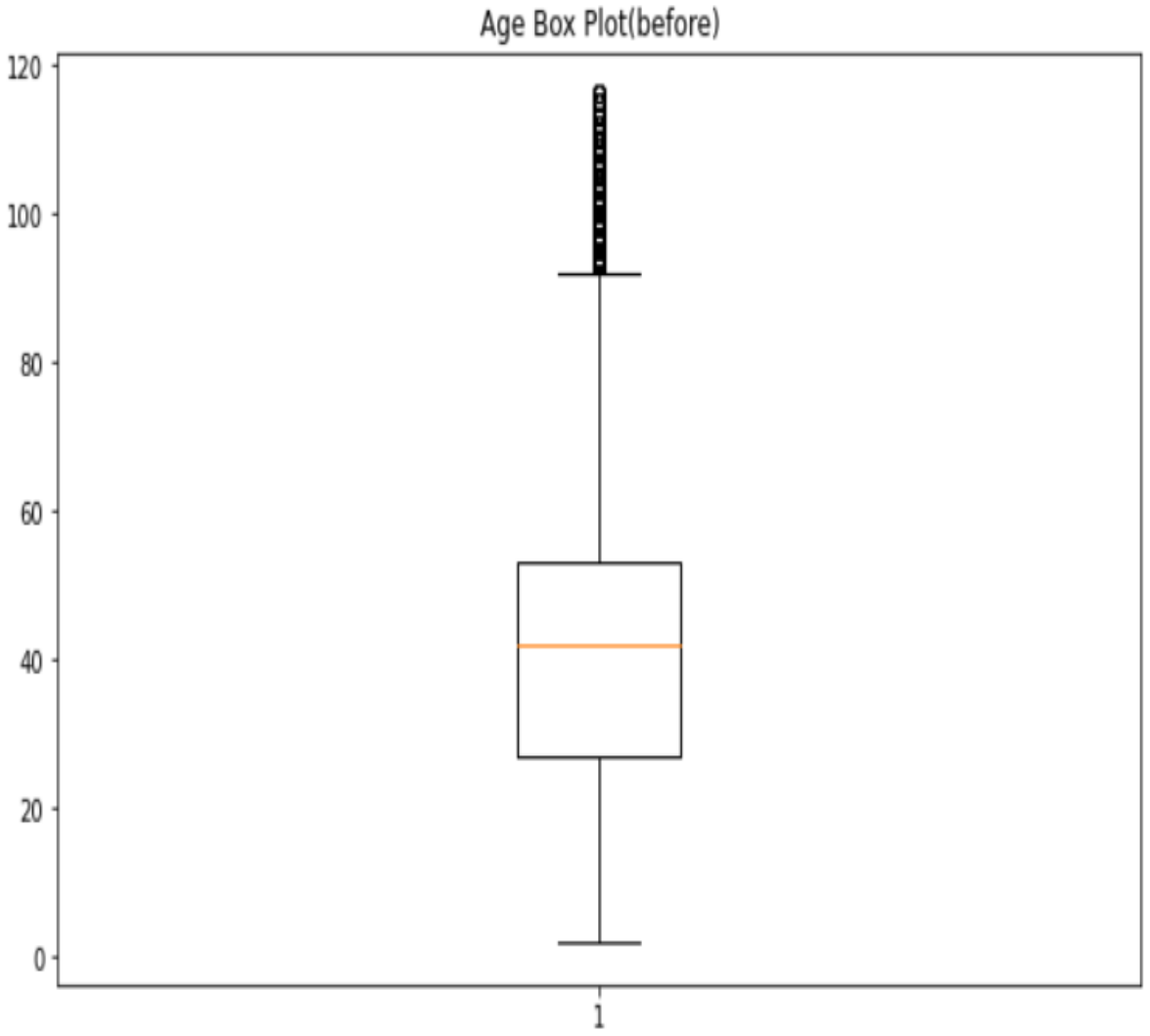
After

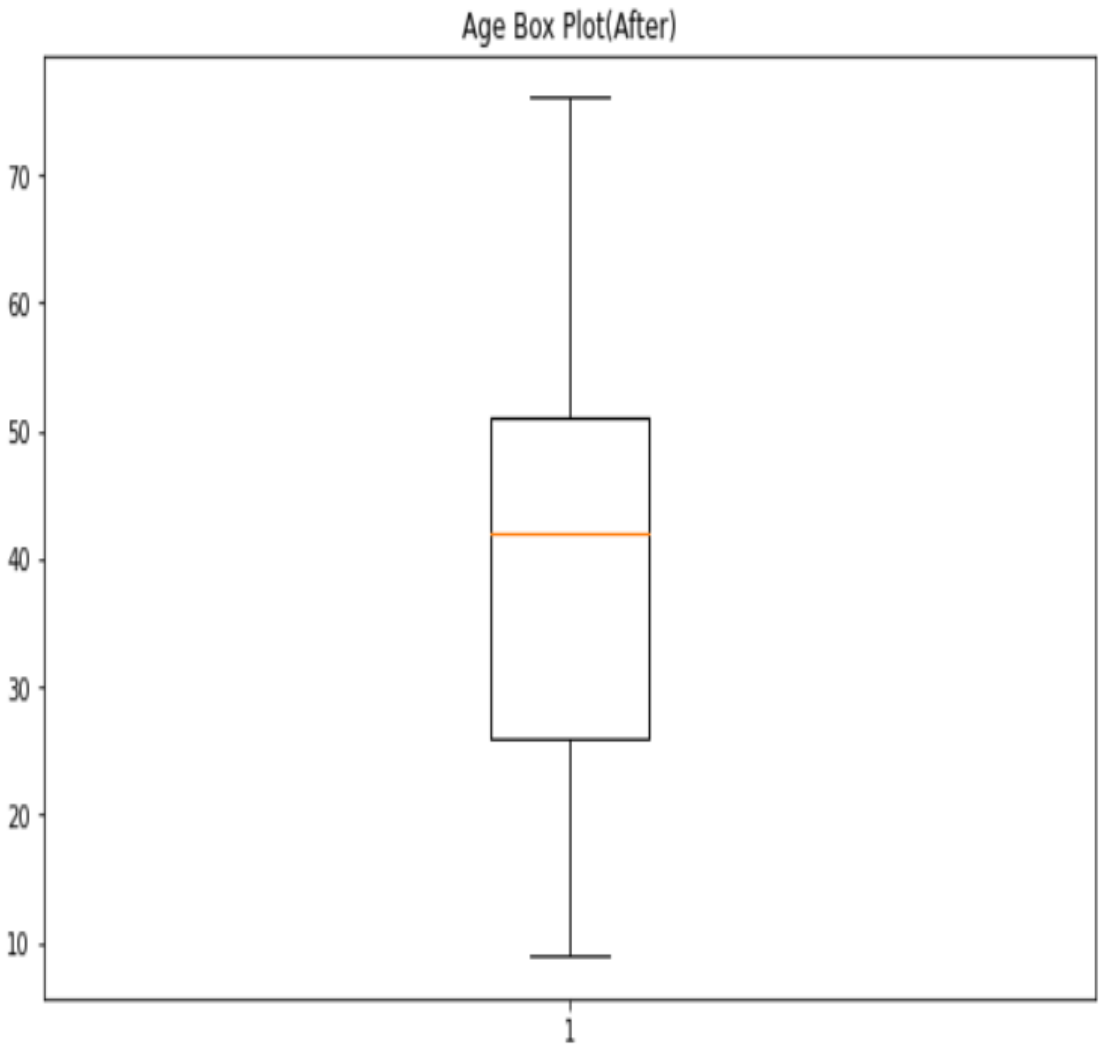
Outliers(Using z-score)





Outliers(Using z-score)

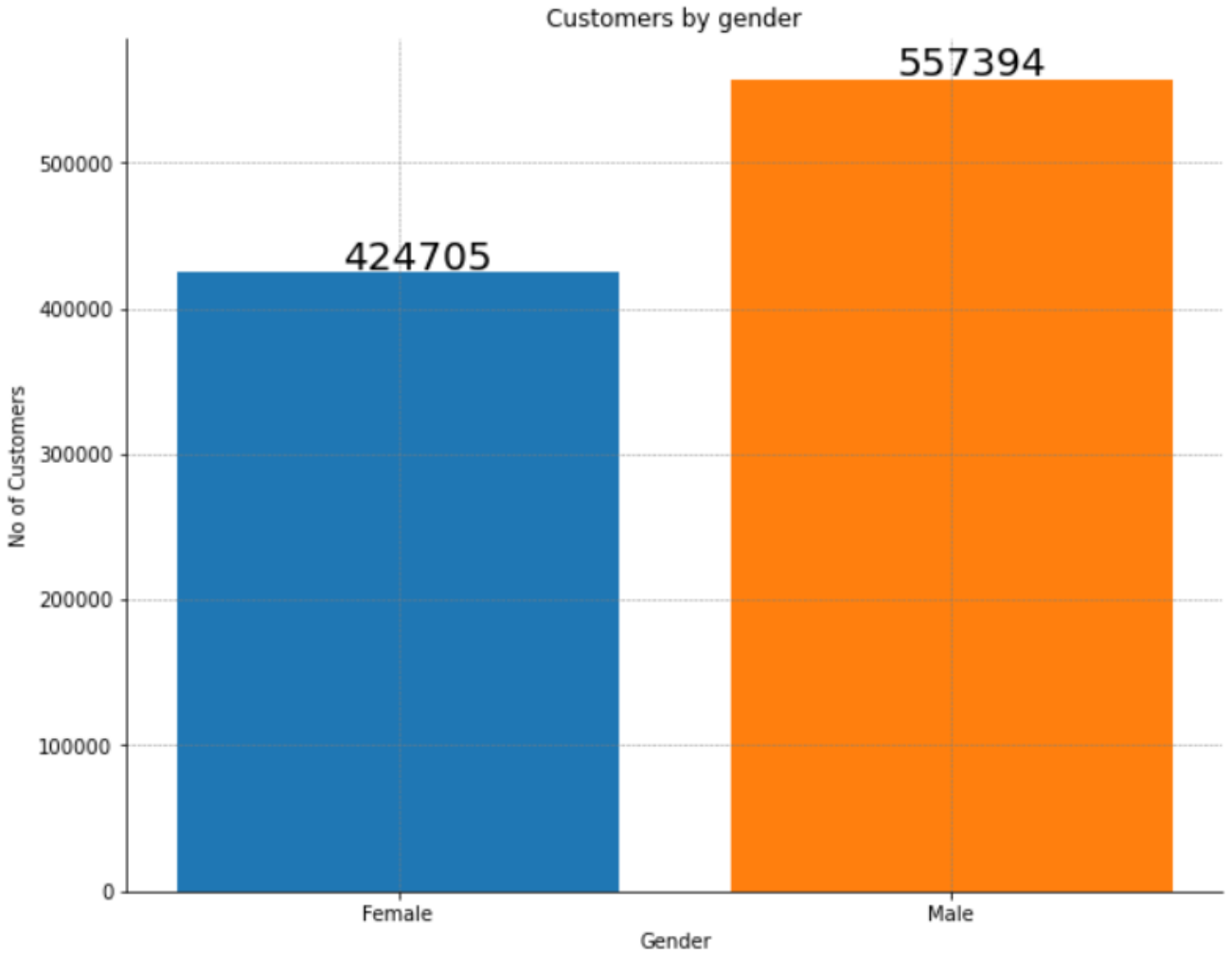




General Observations

- Employee Index values
  - N : 981145
  - A : 285
  - F : 282
  - B : 385
  - S : 2
- Country values
  - ES : 982089
  - IT : 4
  - DE : 2
  - BO : 2
  - PY : 2
- Customer Type values
  - 1 : 982095
  - 2 : 2
  - 3 : 2
- These column will not be very useful in clustering, as they will introduce bias, and can thus be removed.

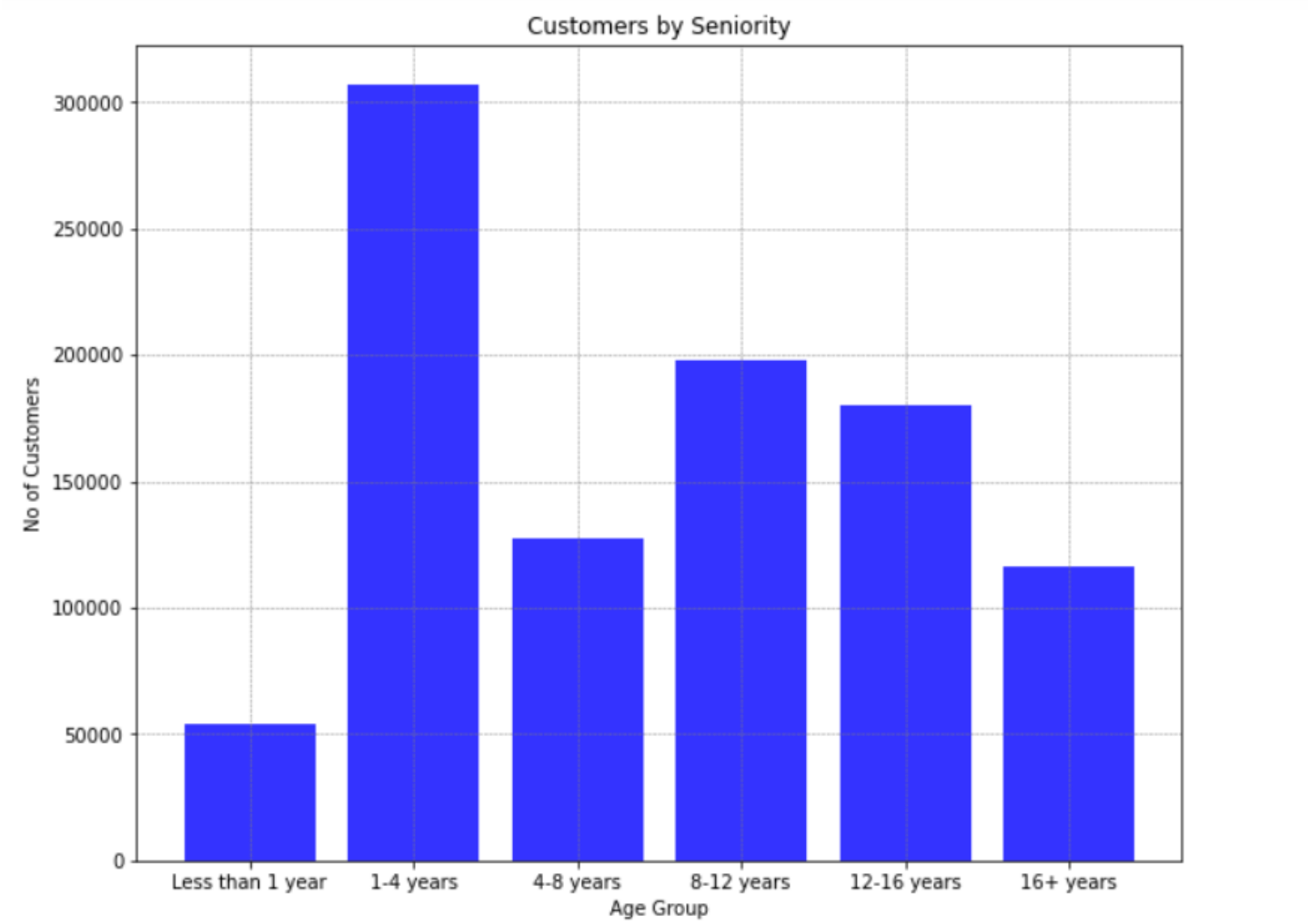
General Observations



General Observations

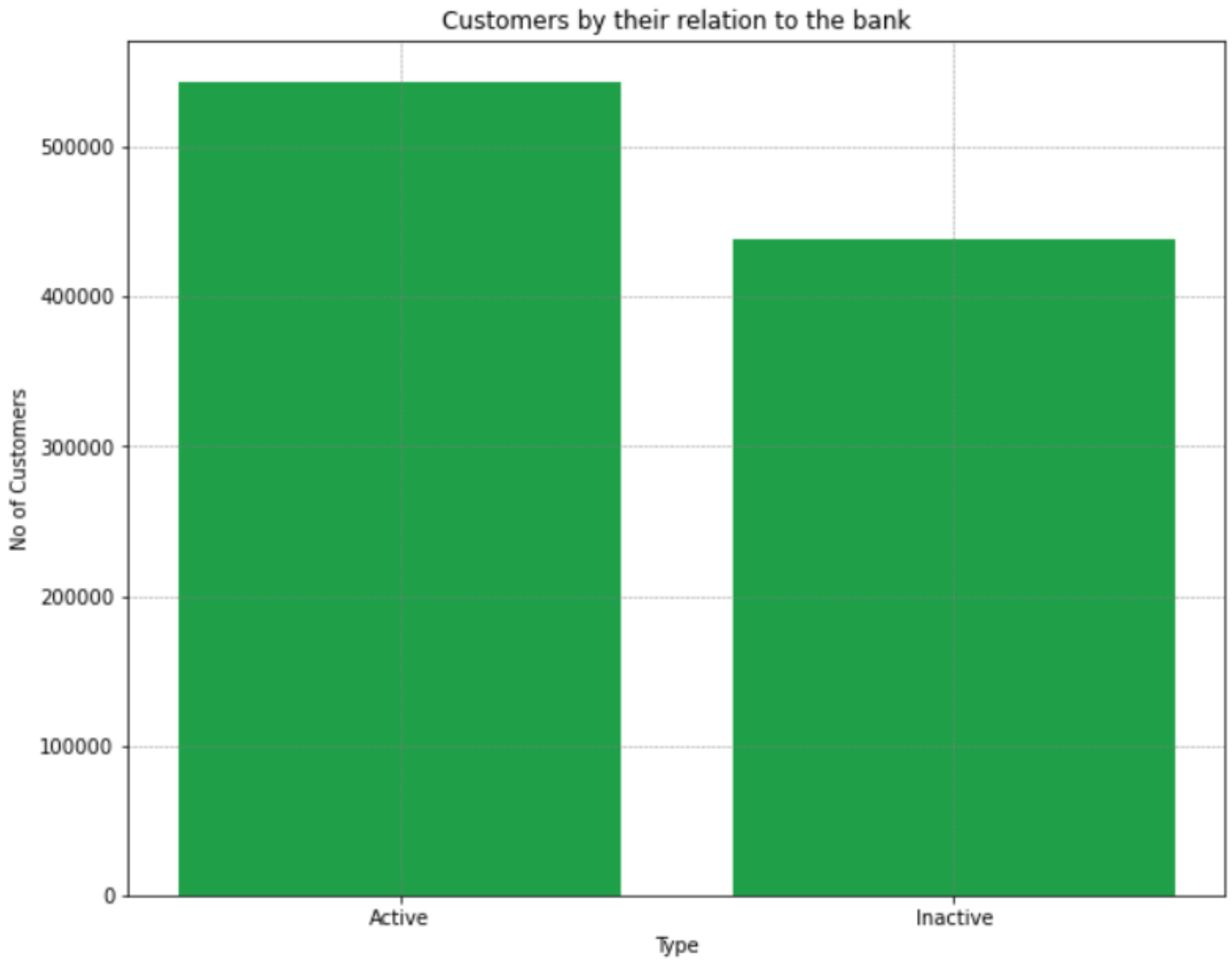
The maximum amount of customers are in the age group of 45-60.

General Observations

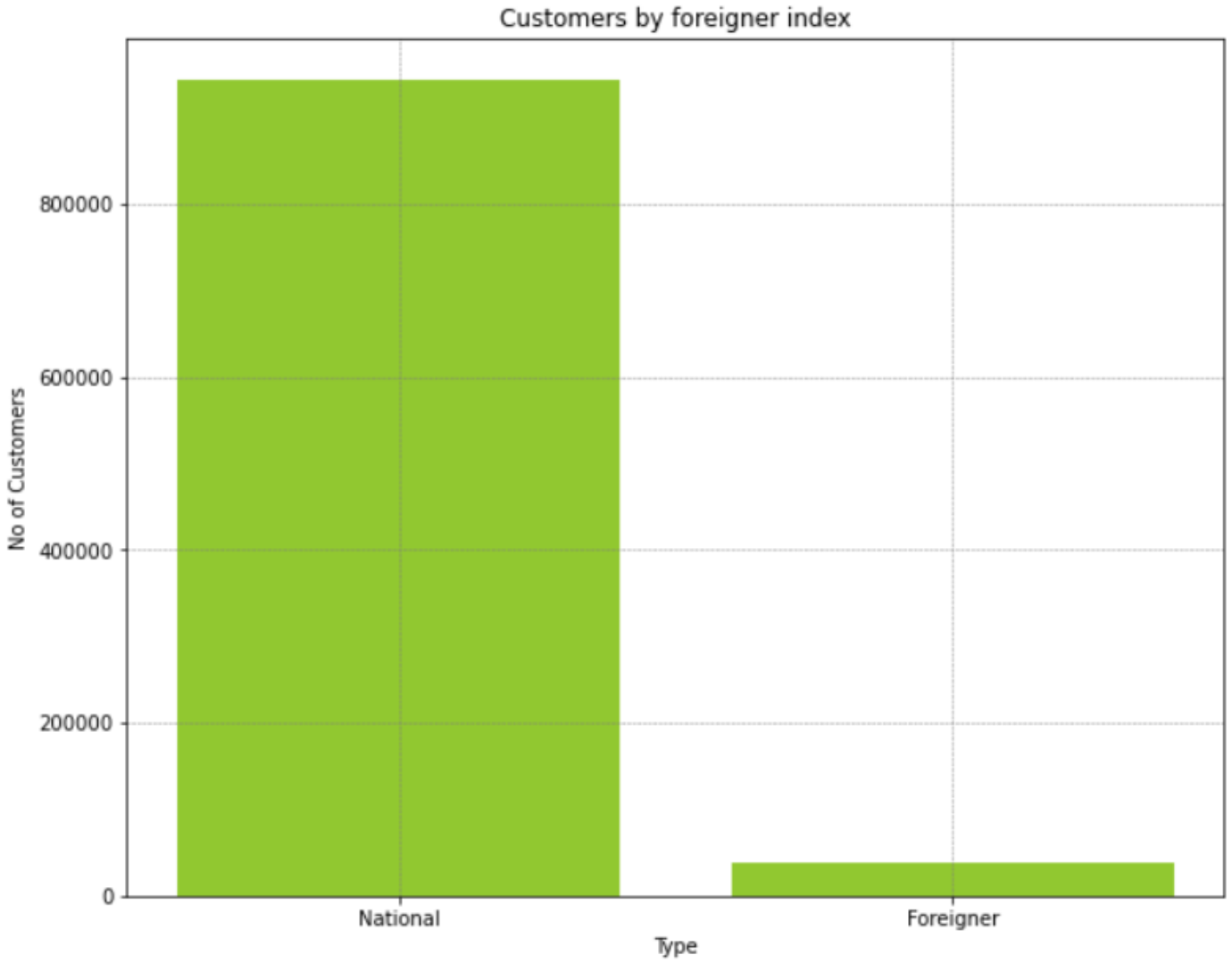


Customer Seniority Distribution, as we can see most of the customers joined 1-4 years ago.

General Observations



General Observations



Recommendation

- Finally our data has been cleaned and all outliers and NA values have been dealt with.
- The columns we have chosen are relatively balanced and will be good for modeling
- A clustering model can be created on the given dataset to divide customers into segments for targeted promotions.
- The best algorithm to use would be the K-means algorithm

Thank You