Final Presentaion

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
  - Modeling
  - Evaluation
  - 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

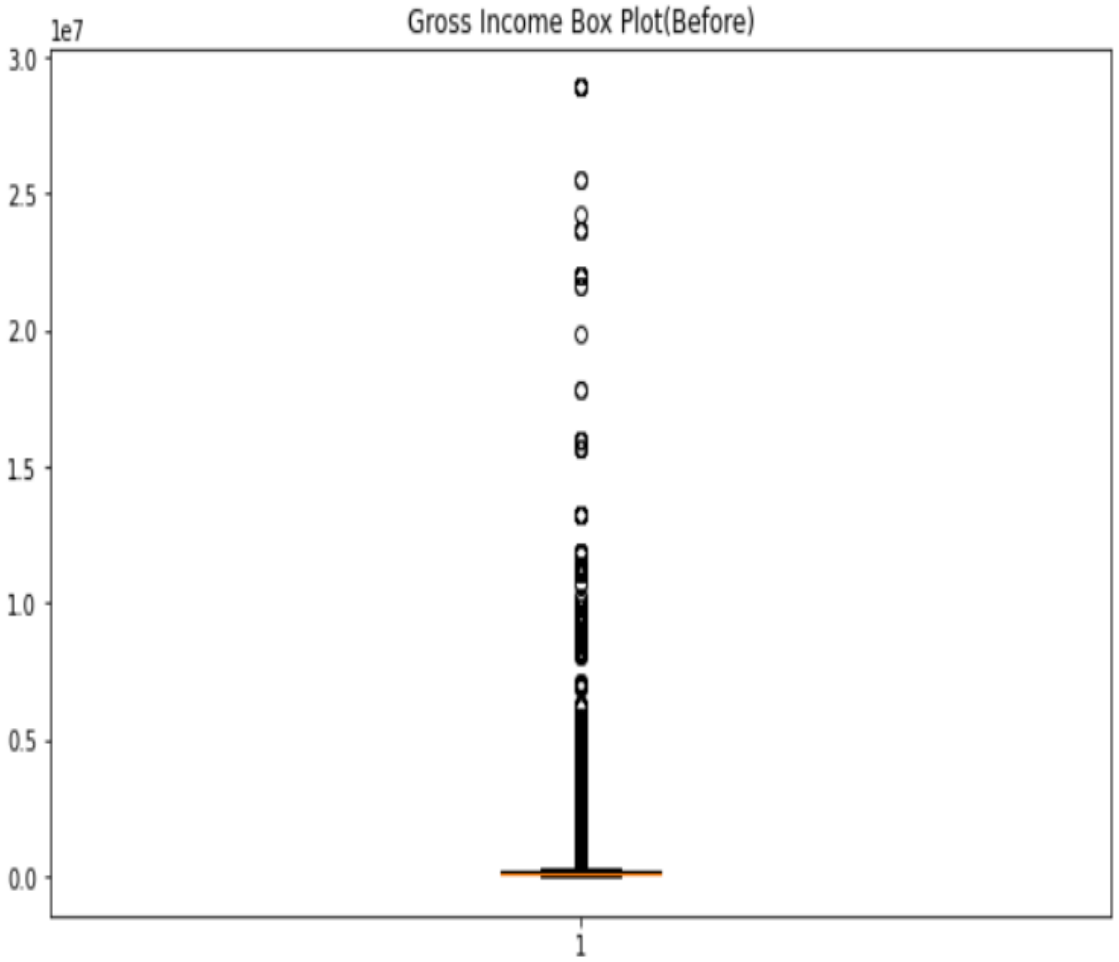
fecha_dato	0
ncodpers	0
ind_empleado	10782
pais_residencia	10782
sexo	10786
age	0
fecha_alta	10782
ind_nuevo	10782
antiguedad	0
indrel	10782
ult_fec_cli_1t	998899
indrel_1mes	10782
tiprel_1mes	10782
indresi	10782
indext	10782
conyuemp	999822
canal_entrada	10861
indfall	10782
tipodom	10782
cod_prov	17734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

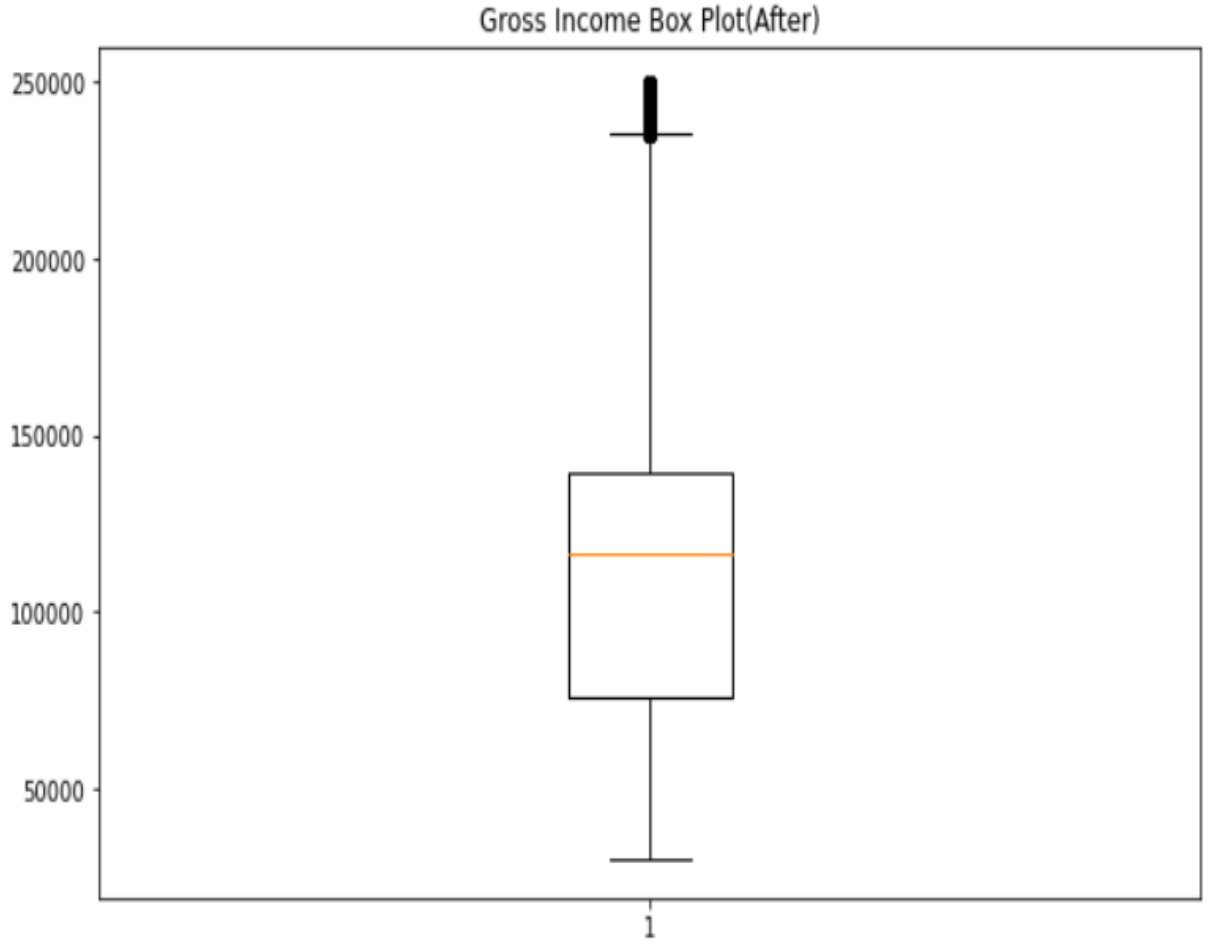
Before

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

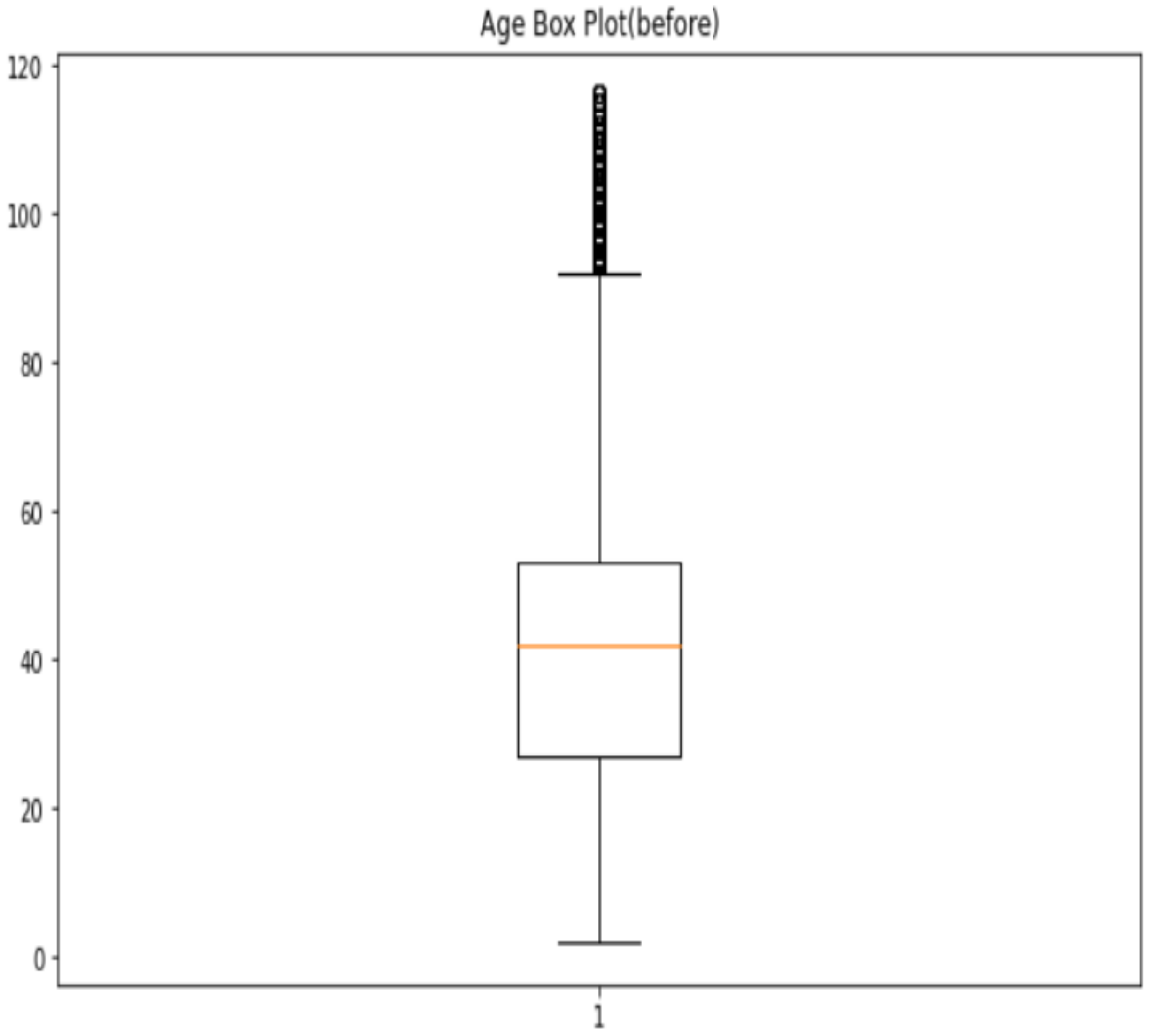
After

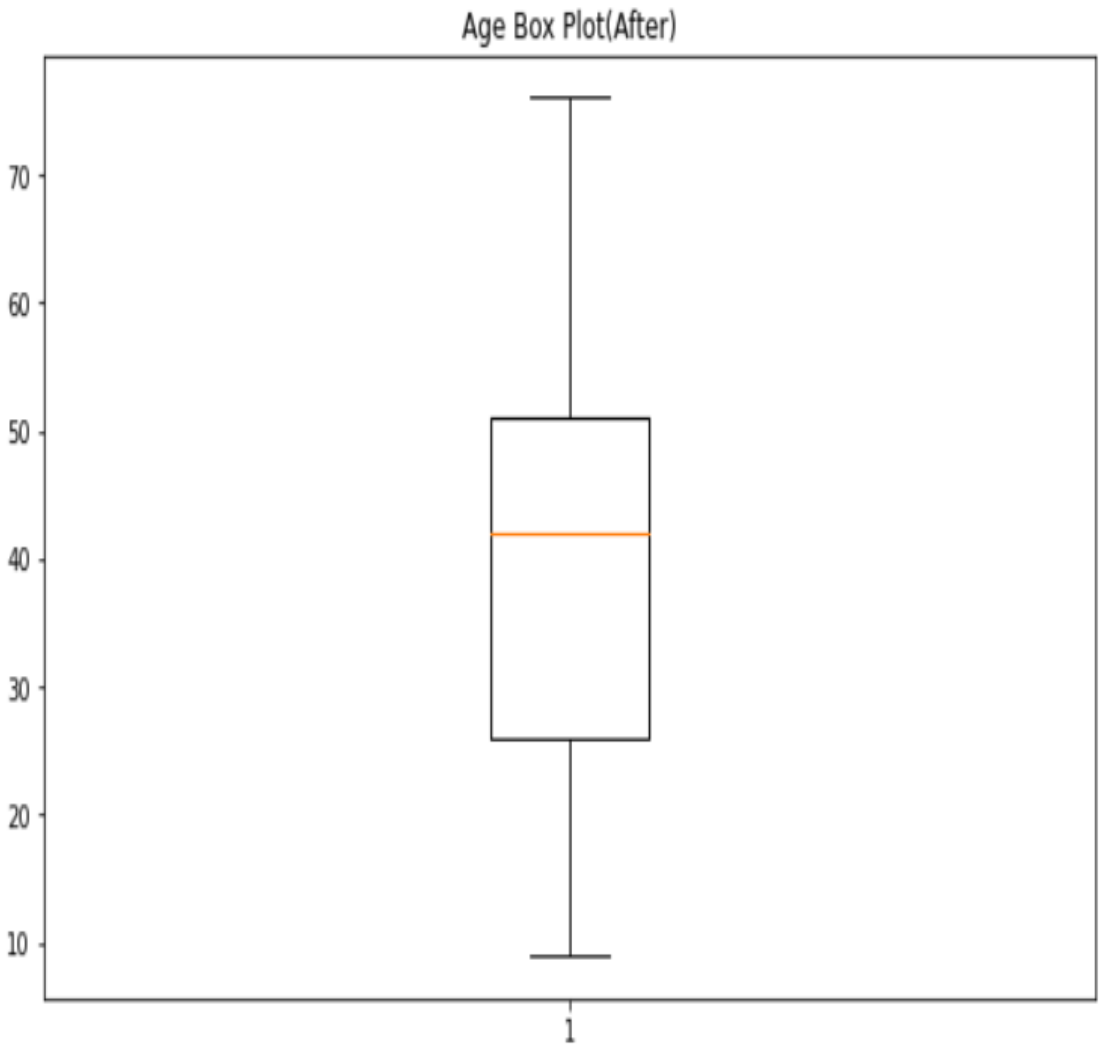
Outliers(Using z-score)





Outliers(Using z-score)

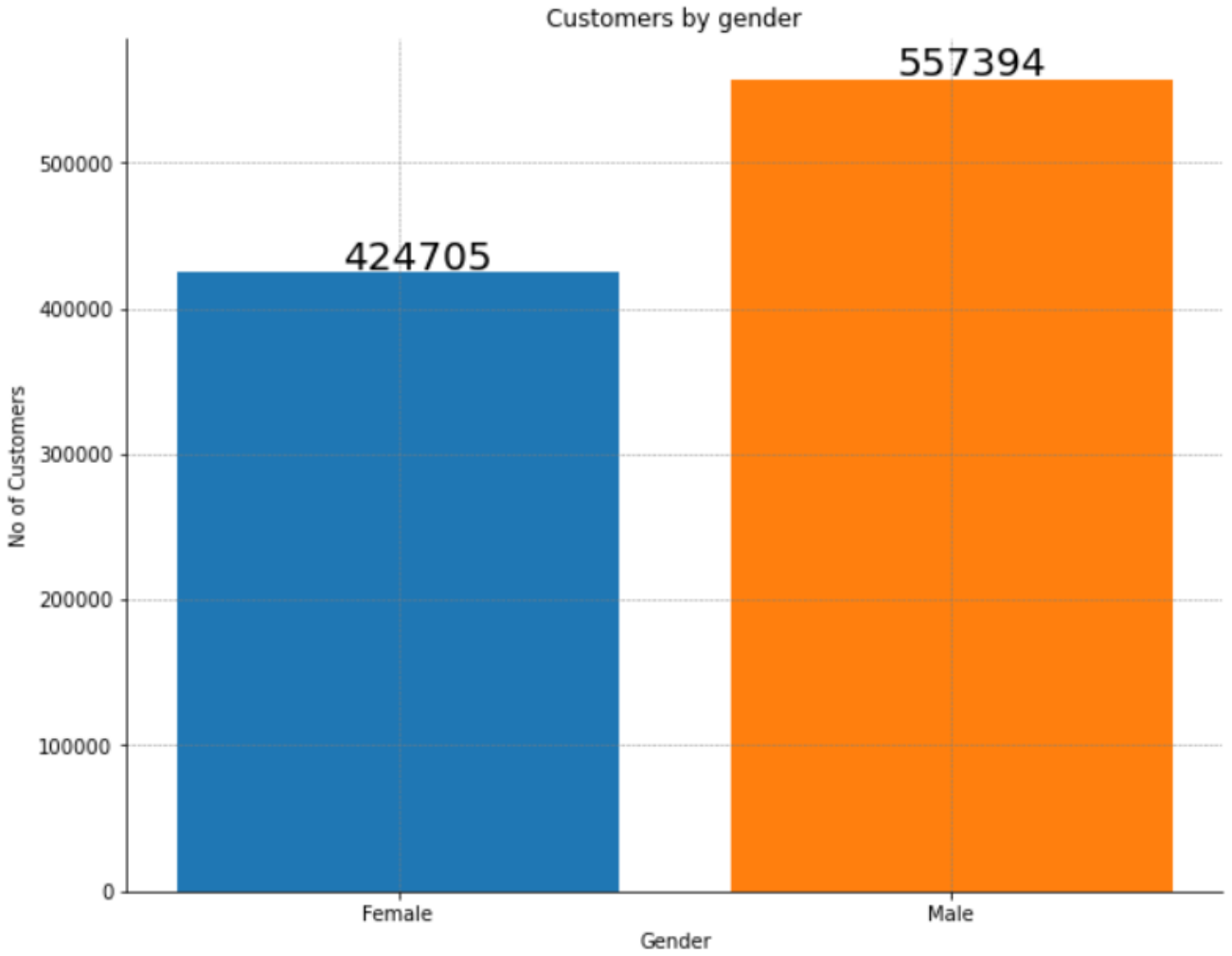




General Observations

- 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.

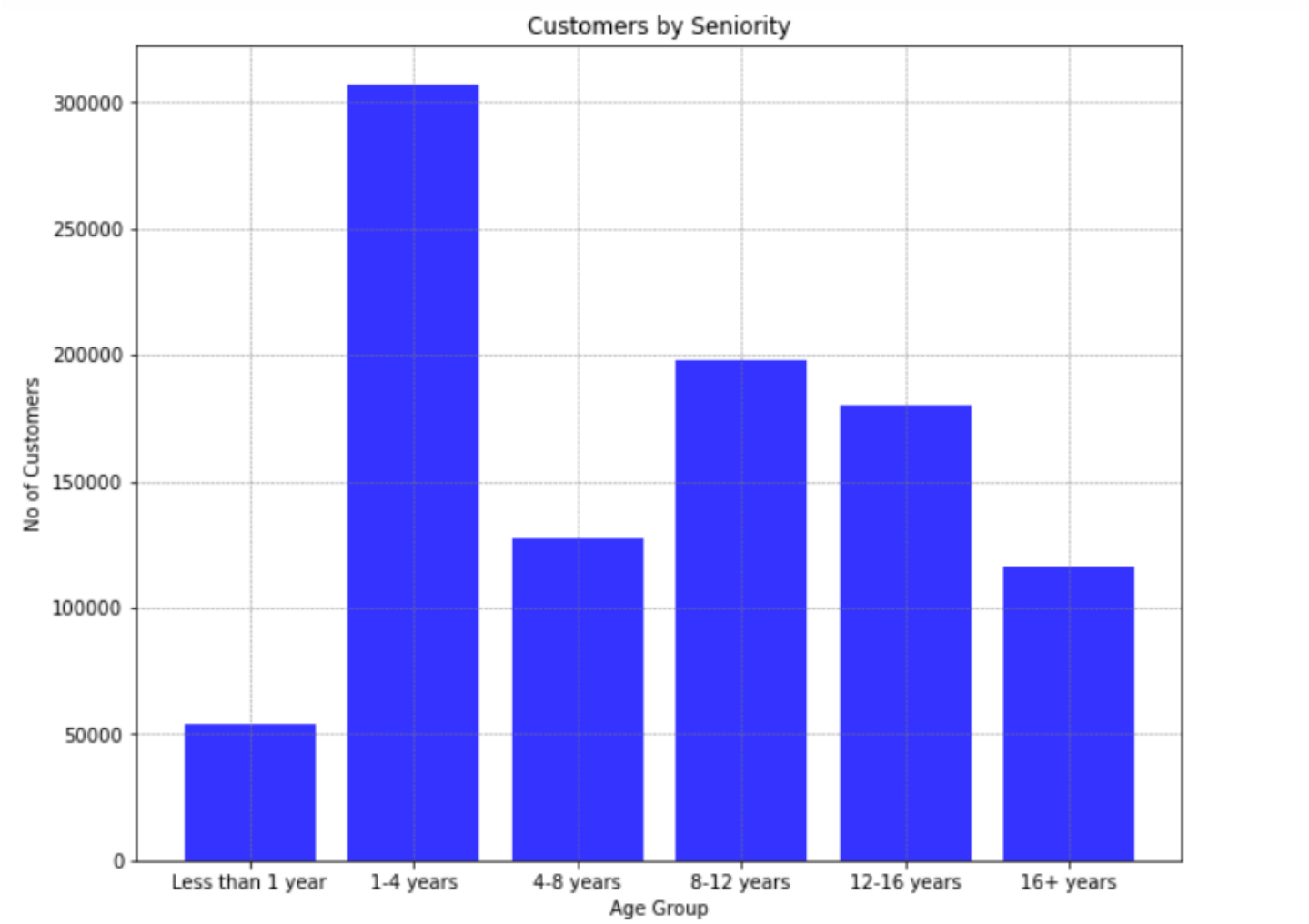
General Observations



General Observations

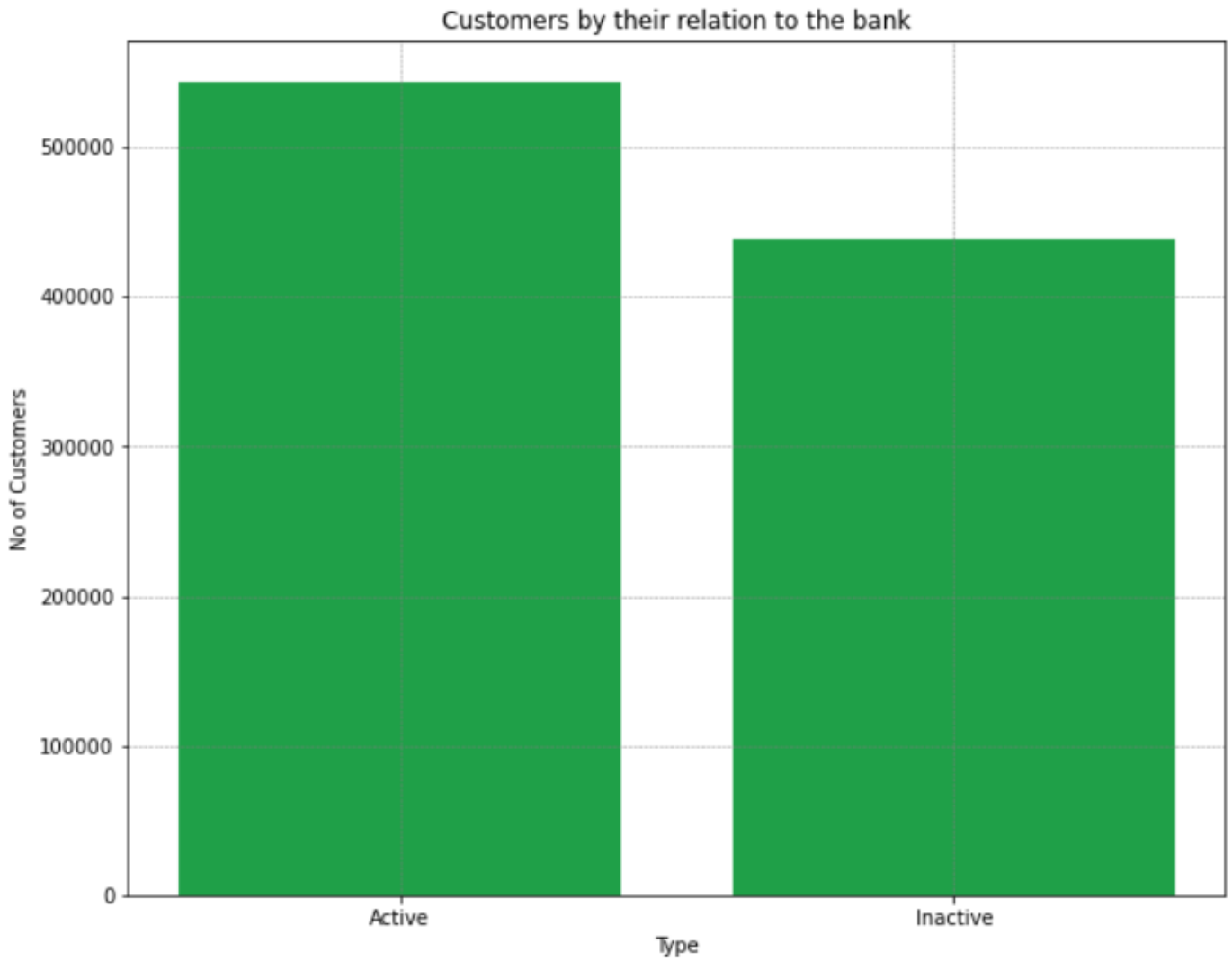
The maximum amount of customers are in the age group of 45-60.

General Observations

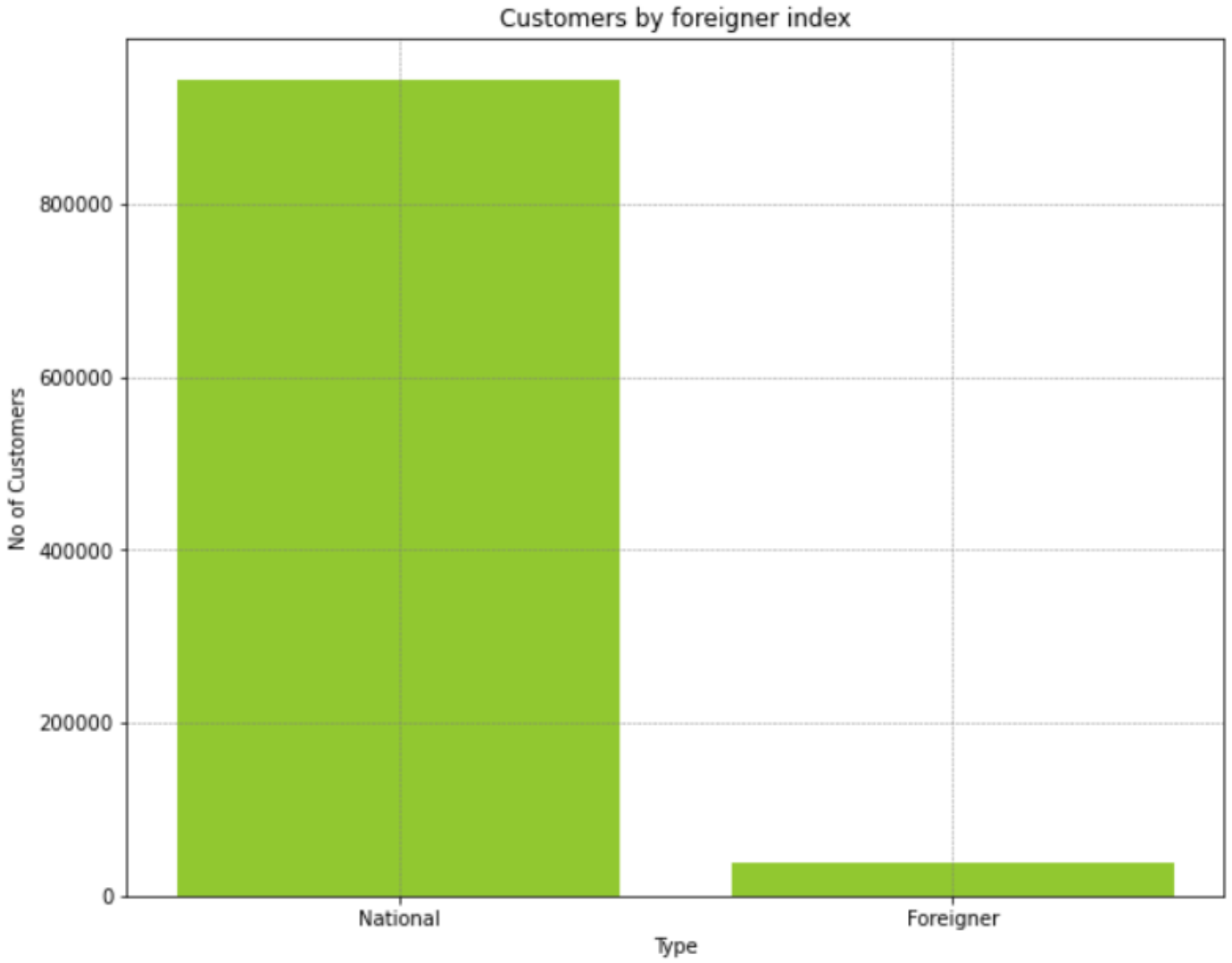


Customer Seniority Distribution, as we can see most of the customers joined 1-4 years ago.

General Observations



General Observations



Modeling

- Final Dataset looks something like

	Gender	Age	Customer Seniority(months)	Customer relation type	Channel joined	Province	Gross Income
0	F	35	6	A	KHL	MALAGA	87218.10000
1	M	23	35	I	KHE	CIUDAD REAL	35548.74000
2	M	23	35	I	KHE	CIUDAD REAL	122179.11000
3	F	22	35	I	KHD	ZARAGOZA	119775.54000
4	M	23	35	A	KHE	ZARAGOZA	139646.15094
...	...	...	...	...	...	...	...
853018	F	27	22	A	KFC	MADRID	55516.98000
853019	F	56	22	A	KFC	CORUÑA, A	75654.84000
853020	M	39	22	A	KFC	CORUÑA, A	80634.87000
853021	M	36	22	A	KFC	MADRID	57818.46000
853022	F	38	22	A	KFC	CADIZ	85903.44000

853023 rows × 7 columns

Modeling

- For this particular problem we have to use the K-means clustering algorithm for determining the group of each of the customer in the dataset.
- K-means is a clustering technique which calculates the distance between observations and groups them into clusters by iterative technique.
- K-means works very well with data of such sort.
- For K-means we have to convert all the variables into numeric values
- This can be done by using the Label encoder this is shown next slide.

Modeling(Encoding into numeric data)

```
Gender_lb = preprocessing.LabelEncoder()
relation_lb = preprocessing.LabelEncoder()
joined_lb = preprocessing.LabelEncoder()
prov_lb = preprocessing.LabelEncoder()

Gender_lb.fit(data['Gender'])
relation_lb.fit(data['Customer relation type'])
joined_lb.fit(data['Channel joined'])
prov_lb.fit(data['Province'])

data['Gender'] = Gender_lb.transform(data['Gender'])
data['Customer relation type'] = relation_lb.transform(data['Customer relation type'])
data['Channel joined'] = joined_lb.transform(data['Channel joined'])
data['Province'] = prov_lb.transform(data['Province'])

data.head()
```

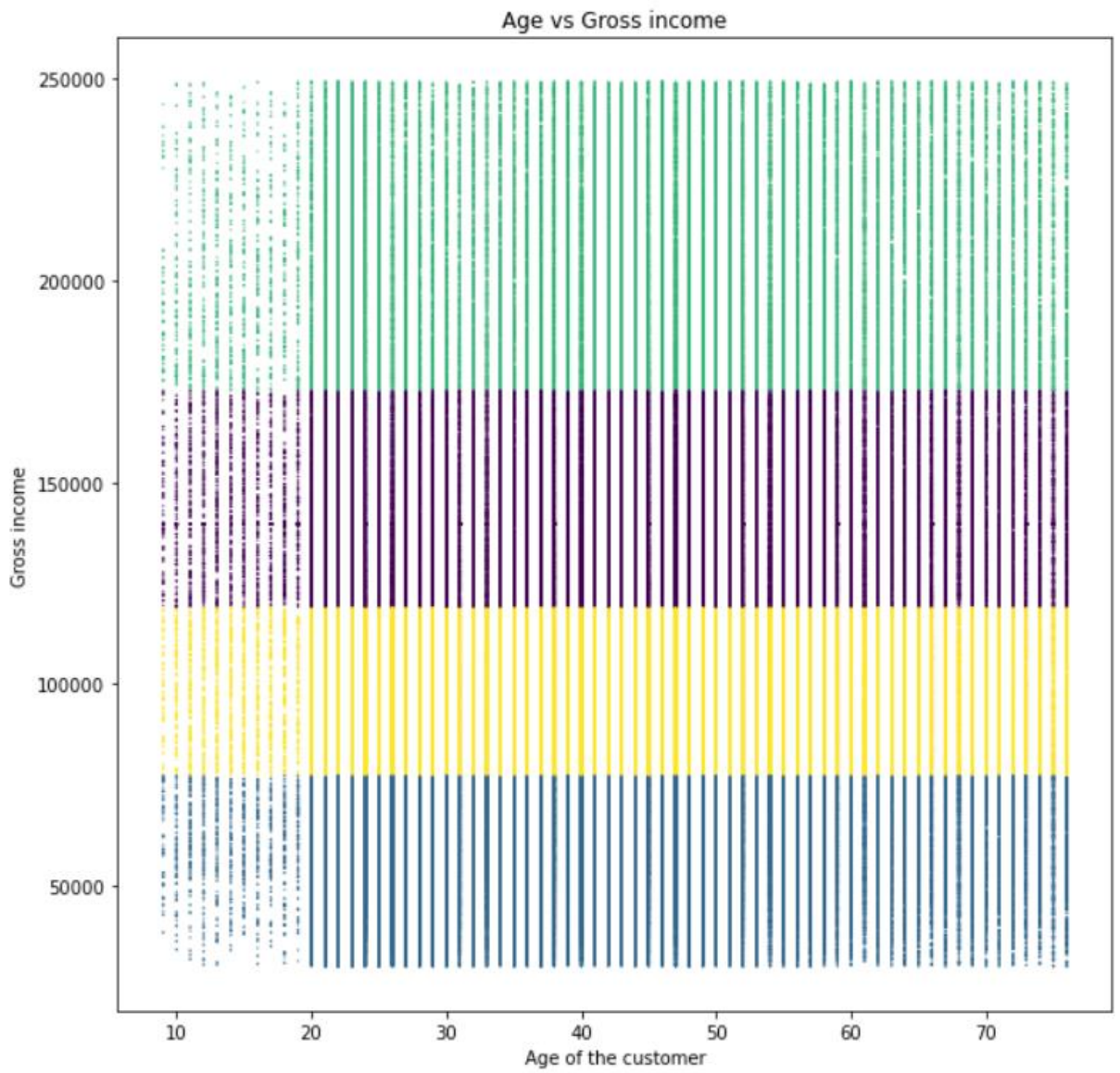
	Customer Code	Gender	Age	New Customer	Customer Seniority(months)	Customer relation type	Channel joined	Province	Gross Income
0	1375586	0	35	0	6	0	149	31	87218.10000
1	1050611	1	23	0	35	1	146	16	35548.74000
2	1050612	1	23	0	35	1	146	16	122179.11000
3	1050613	0	22	0	35	1	145	51	119775.54000
4	1050614	1	23	0	35	0	146	51	139646.15094

Modeling(Creating the model)

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4)
kmeans.fit(data[['Gender','Age','Customer Seniority(months)','Customer relation type','Channel joined',
                 'Province', 'Gross Income']])
y_kmeans = kmeans.predict(data[['Gender','Age','Customer Seniority(months)','Customer relation type','Channel joined',
                 'Province', 'Gross Income']])
```

Model has been created and predictions have been made

Modeling(Visualization)



Modeling

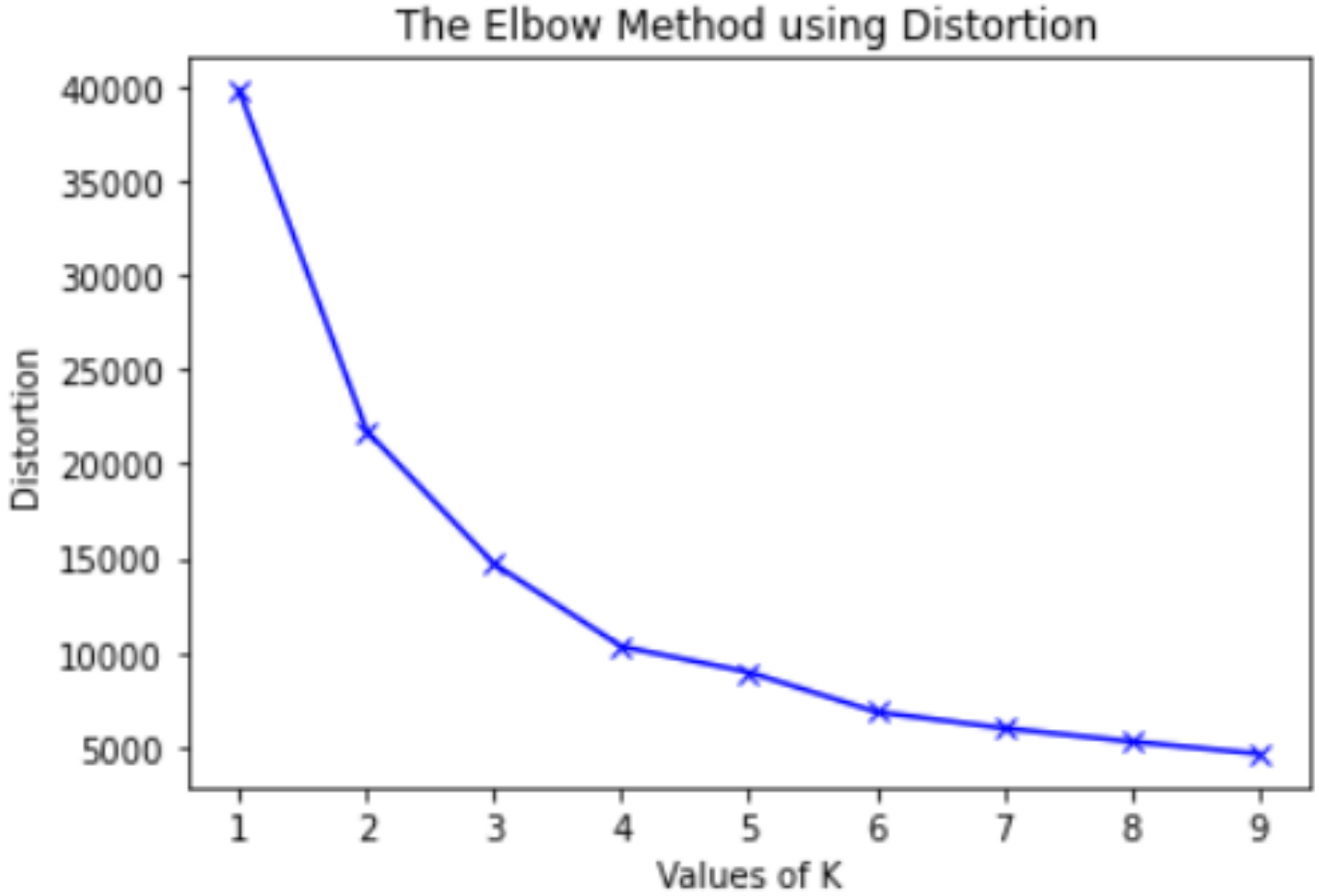
- The model has been created successfully and fit on the particular dataset and is able to distinguish customers into 4 segments
- The graph shows the clusters in relation with Gross Pay of the customer and Age of the customer.
- The model has created clusters based on the Gross pay of the customers and as our client is a bank this makes sense to provide targeted advertisements for the customers.
- Next step is to evaluate the model and find the optimal solution this has been shown in the next slide.

Modeling(Evaluation)

- To evaluate the model we use distortion method to calculate the distortion for different values of K and then find the Elbow method to find the optimal value of K

- We get the following distortion for each value of K.
  - 1 : 39751.68823745483
  - 2 : 21745.029006819208
  - 3 : 14706.331814293326
  - 4 : 10326.245693782306
  - 5 : 8968.418520063888
  - 6 : 6863.19407177712
  - 7 : 6005.101307798649
  - 8 : 5296.478246176699
  - 9 : 4646.272938309745

Modeling(Evaluation)



The optimal value of K is 4 as seen from the graph above.

Recommendation

- Finally our model is ready and has been evaluated.
- By evaluating different models we have found that 4 clusters perform the best in such dataset.
- The model determined that the clusters based on mainly the gross income of the customers and the bank should create better targeted advertisements using this cluster information

Thank You