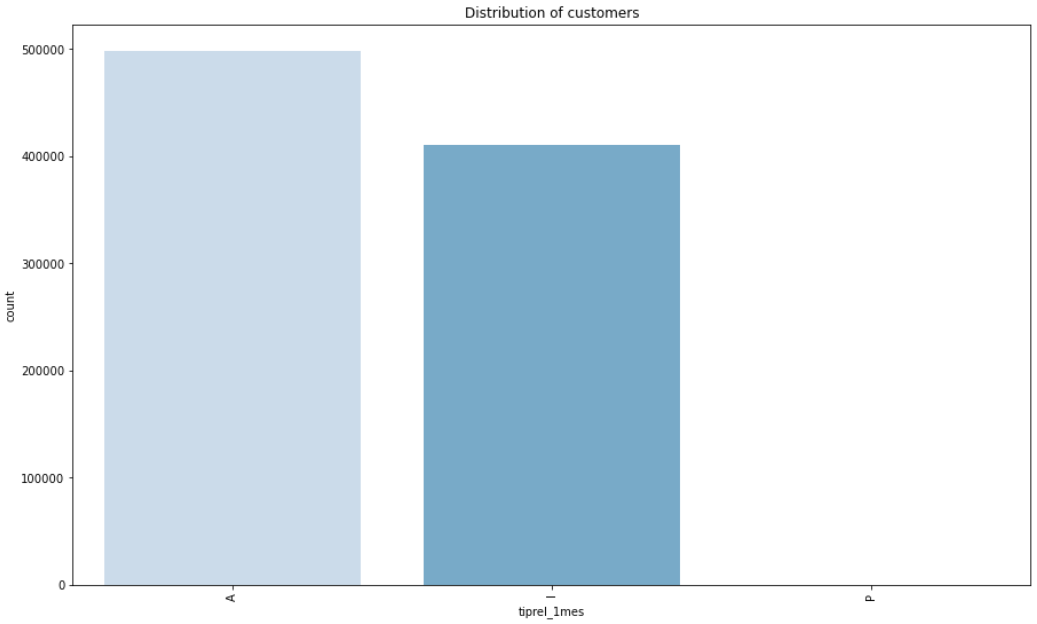
Customer Segmentation

- After preliminary analysis the following are the results of the data
  - As of the beginning of the month we had an almost even distribution between active and passive customers



Country of Origin

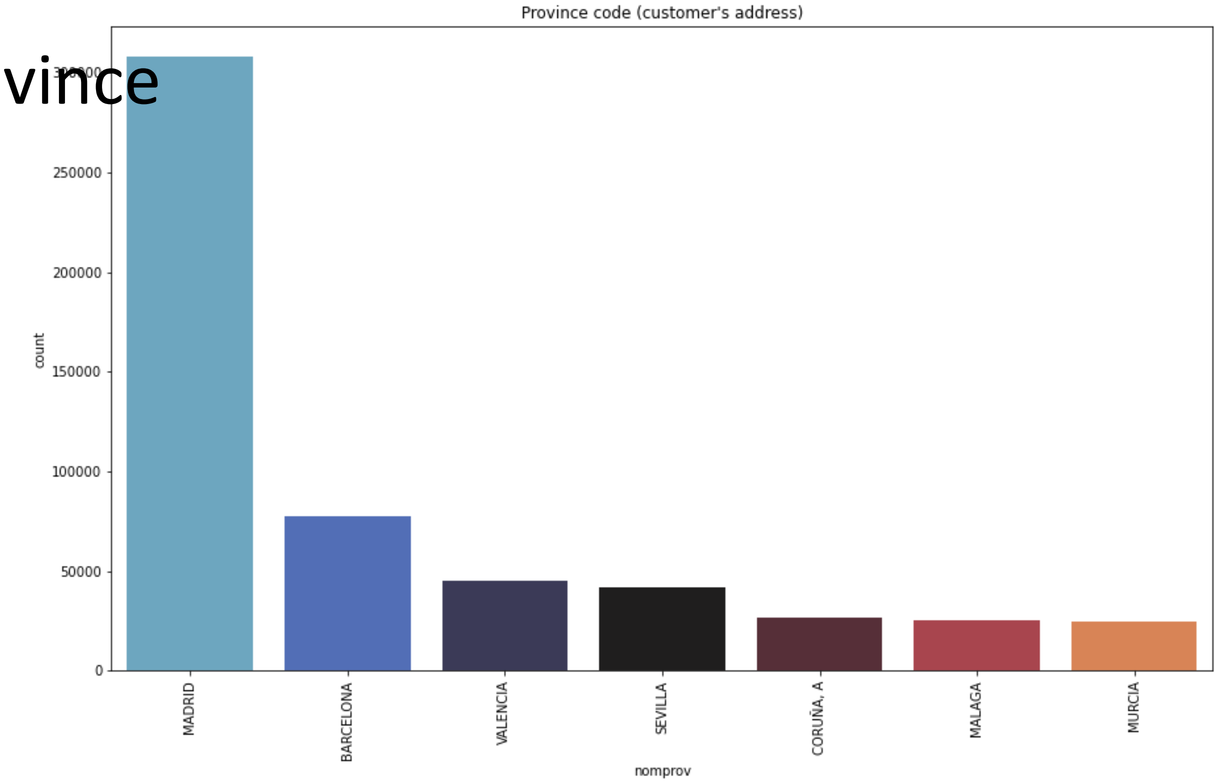
- Most of the customers reside in Spain which is the actual bank residency.
- This was dropped as a feature du to its heavy bias to the nationality and hence would create some bias in the model

Gender of customers

- After analyzing the sex column in the dataset we found that the gender distribution favors males as compared to females

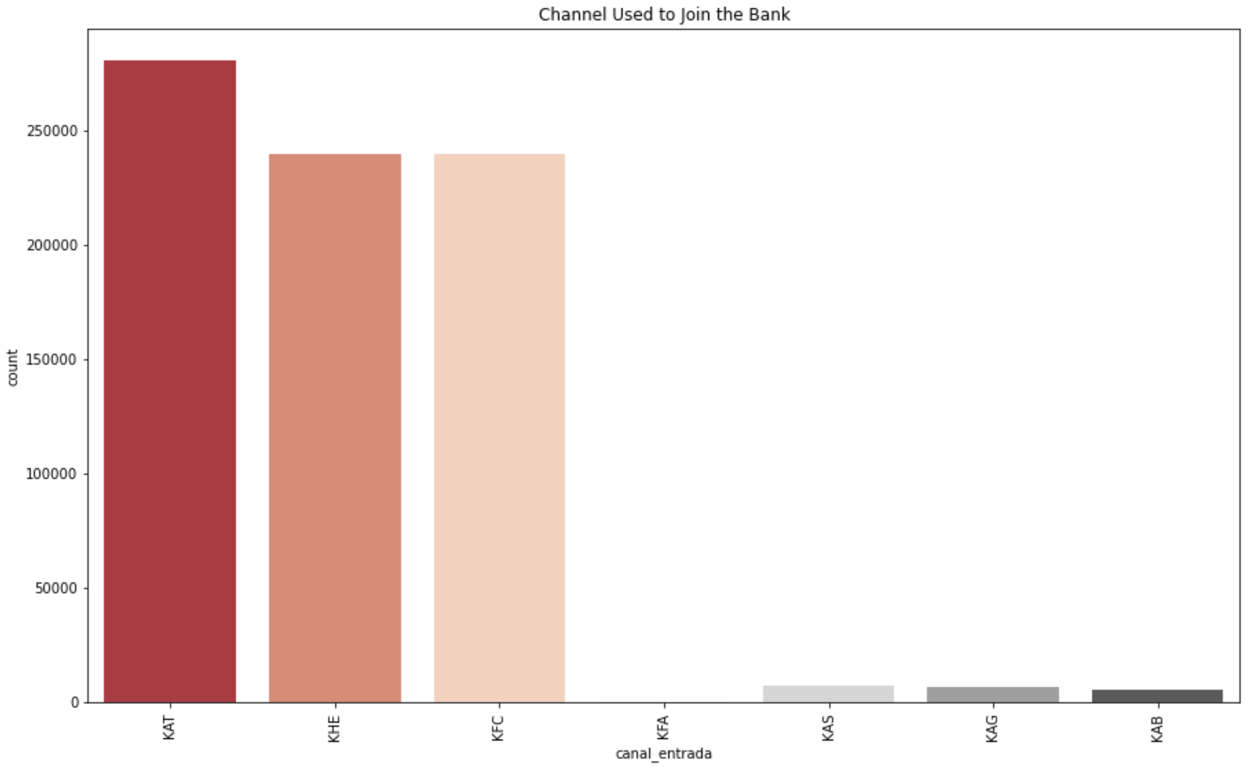
Area of Residence

- From the residential feature column we were able to obtain the distribution of customer’s areas
- Most come from the Madrid province



Channel used to join

- A high percentage of the customers joined the bank through KAT ,KHE and KFC



Technical EDA

- Most of the features were dropped due to their low variance of information
- Some columns had one record dominating over 90% of the records. To ensure easy generalization we dropped this features.
- Columns that showed a high correlation index needed to be addressed and hence dropped on of them
- Date column were converted into datetime format
- Some of the columns had misinterpreted data types and had to be corrected
- Many numeric column displayed high skewness levels and a large number of outlier. This was addressed for all of the discrete features. Such as age and rent

Model Selection

- For this segmentation analysis we shall employ the use of Kmeans clustering algorithm.
- With a cluster of 5
- We however decompose the data using Principle Component Analysis specifying 2 components.