Exploratory Data Analysis

Customer Segmentation Project

Data Explorers

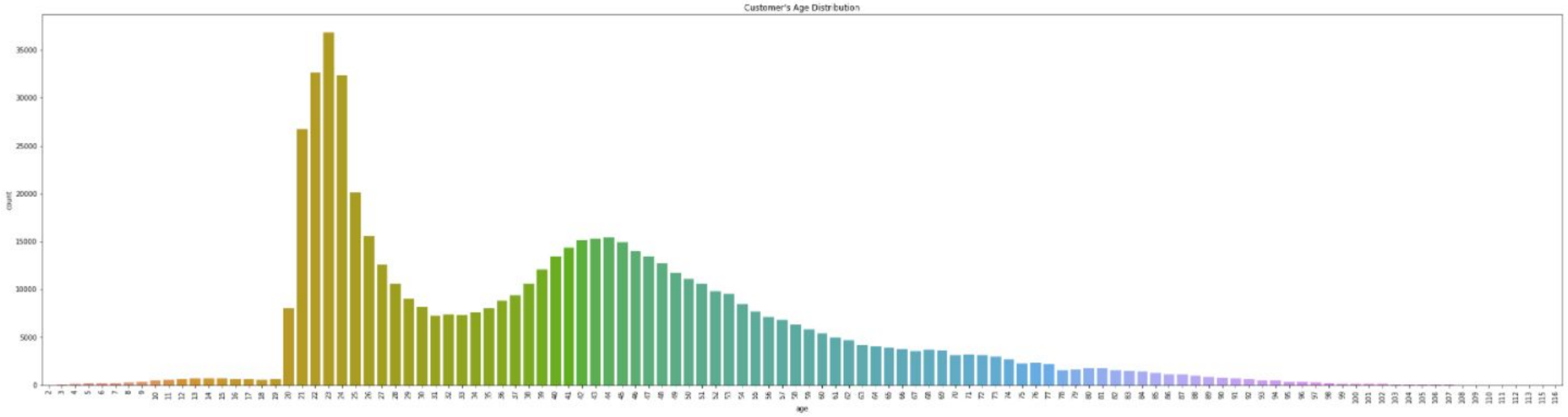
- Joseph Antony
- Dilem Ünal
- Melisa Gözet
- Aynur Cemre Aka

03-Oct-2021

Link Of The EDA File

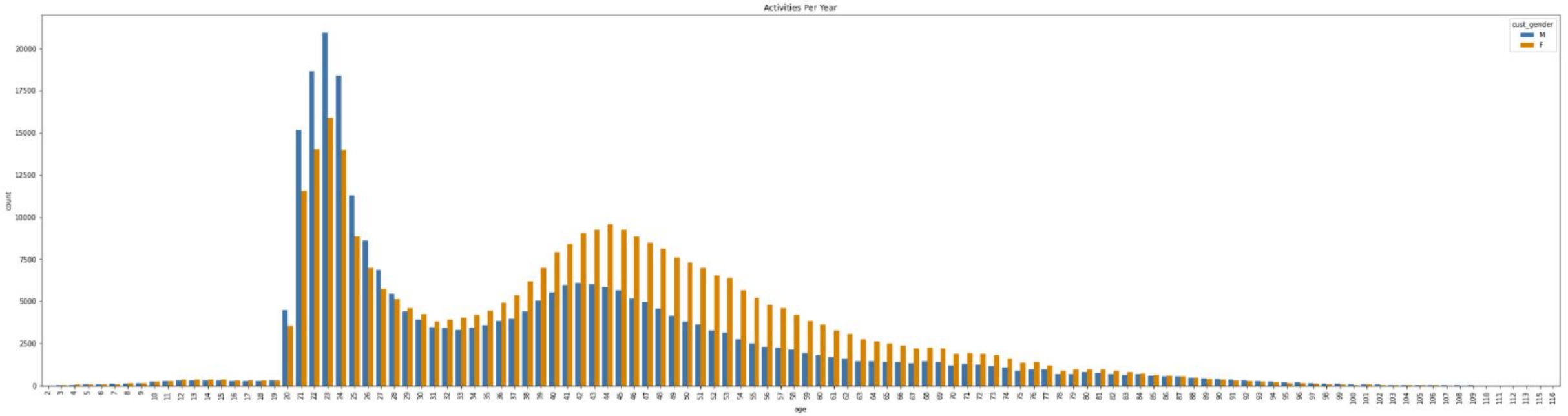
- https://github.com/joeanton719/Customer-Segmentation-Project/blob/main/Week%2010%2611/melisa_eda.ipynb

Customer’s Age Distribution



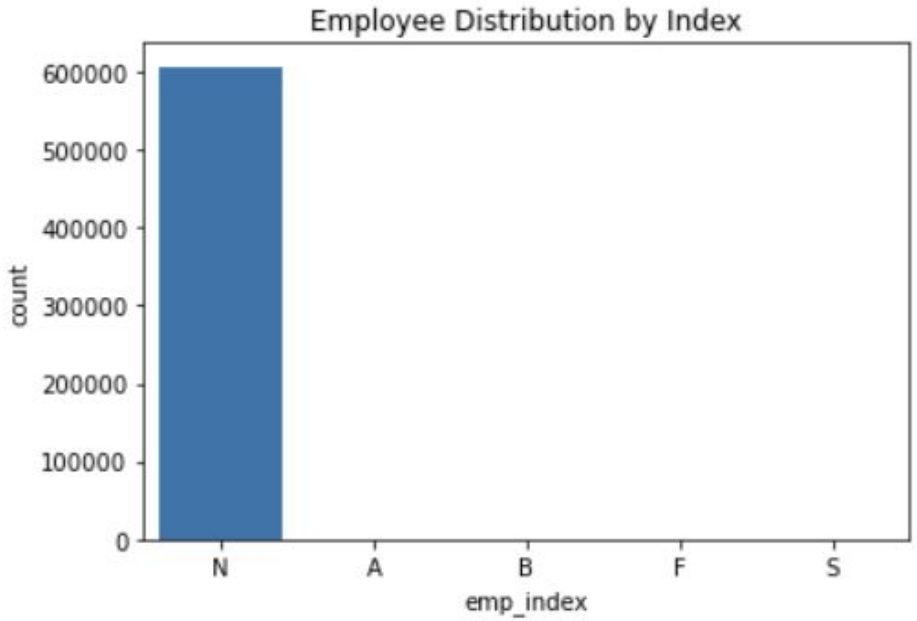
- The distribution is quite high between the ages of 20-30, followed by ages between 35 and 85.

Customer Gender Age distribution



- "F" category is higher in proportion compared to "M"

Customer Activity Index

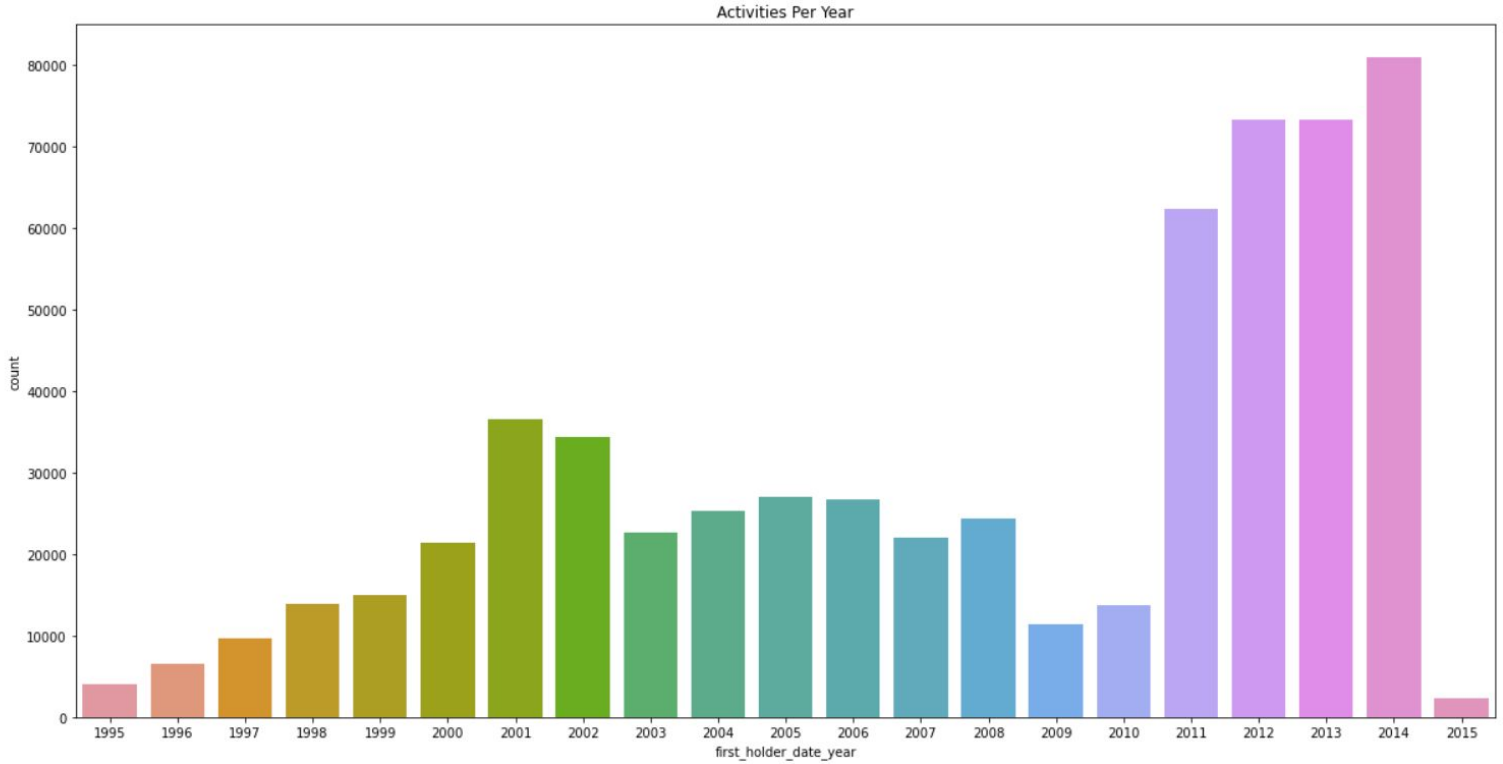


Employee index:

- A: active,
- B: ex employed,
- F: filial,
- N: not employee,
- P: pasive

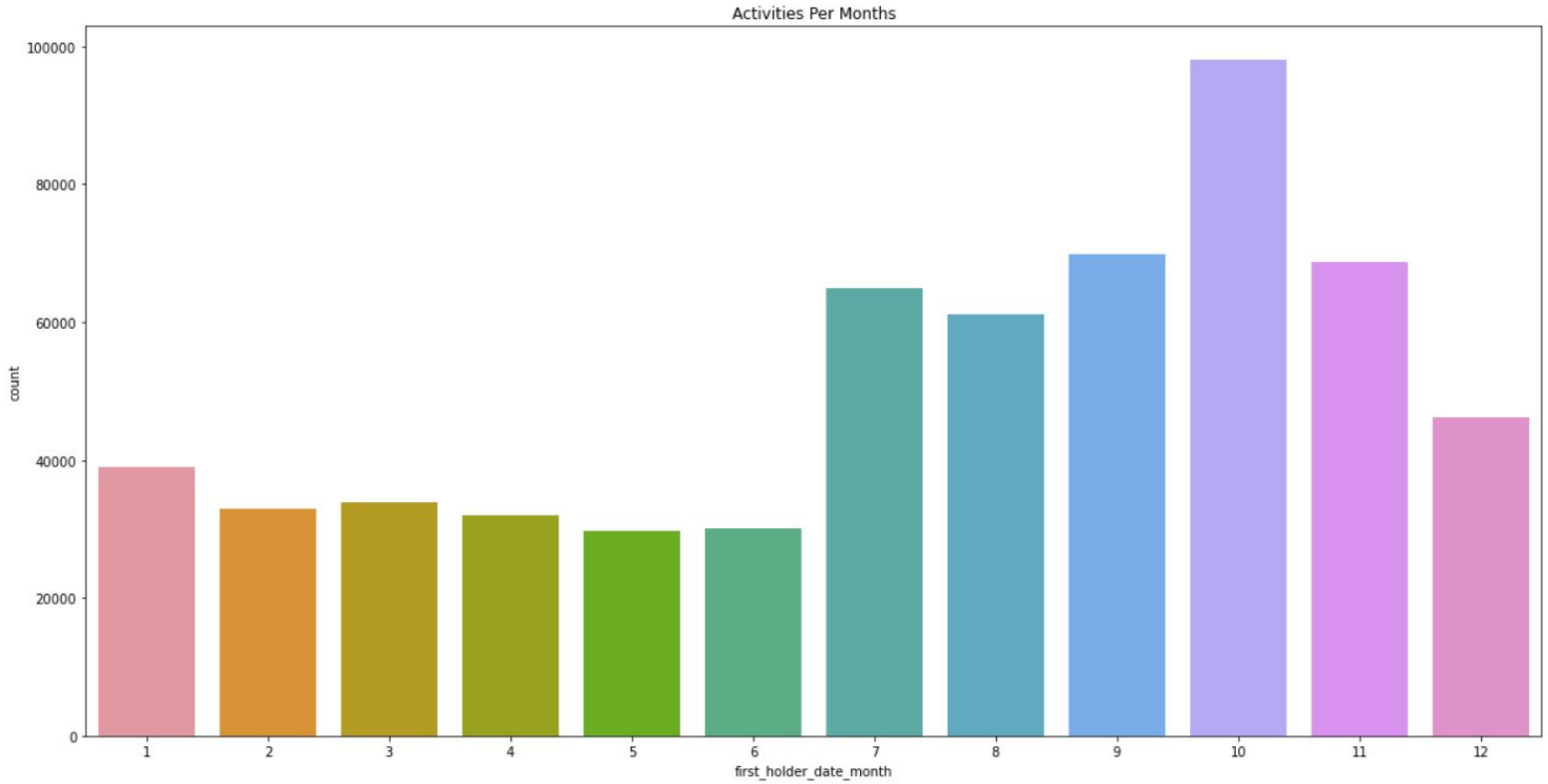
- Almost all customers are not employee. Only a small negligible portion of customers belong to other categories.
- Vast majority of customers are Active.

Activities Per Year



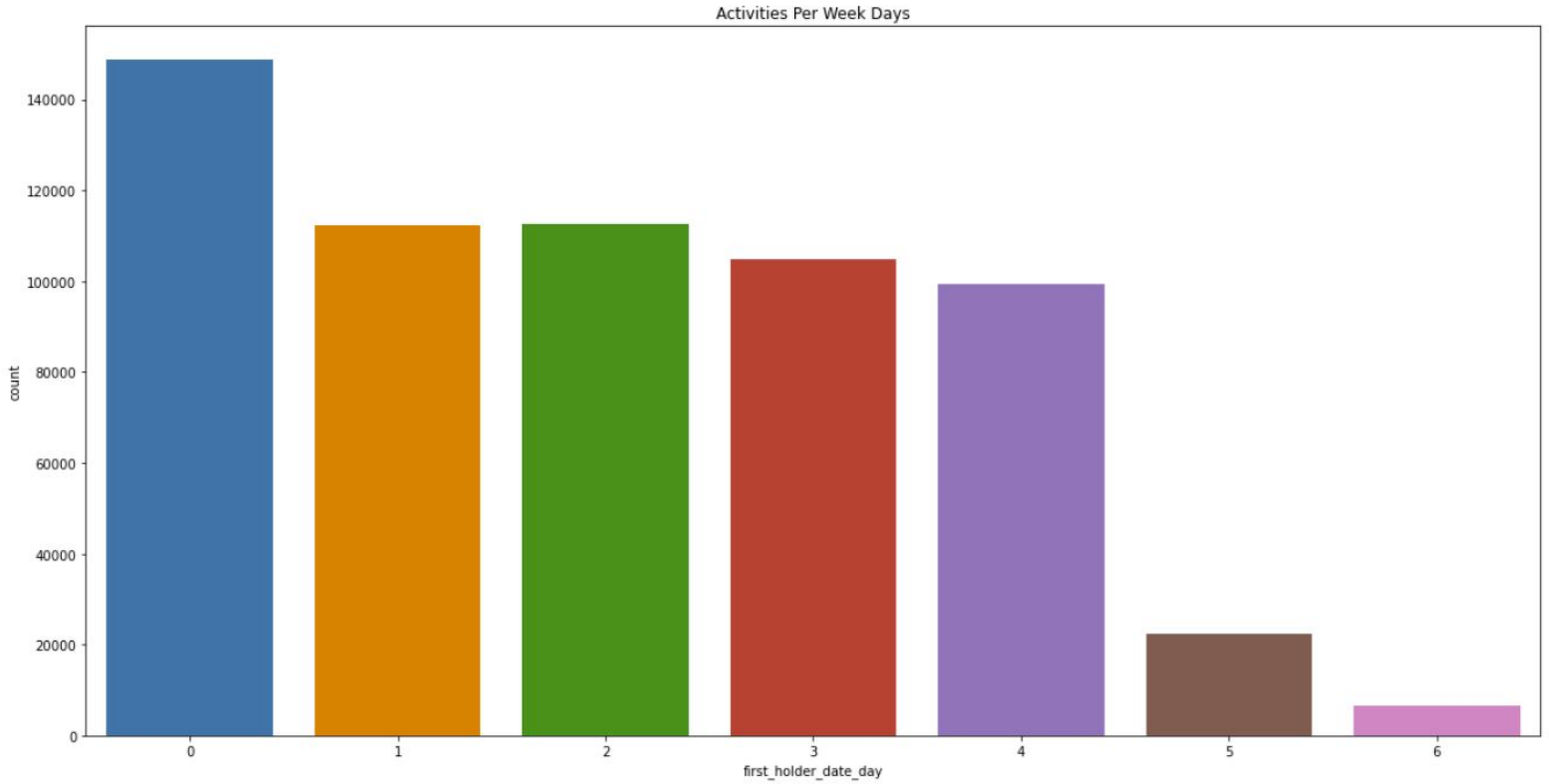
- Years between 2011 and 2014 saw the highest number of applicants compared to other years.

Activities Per Month



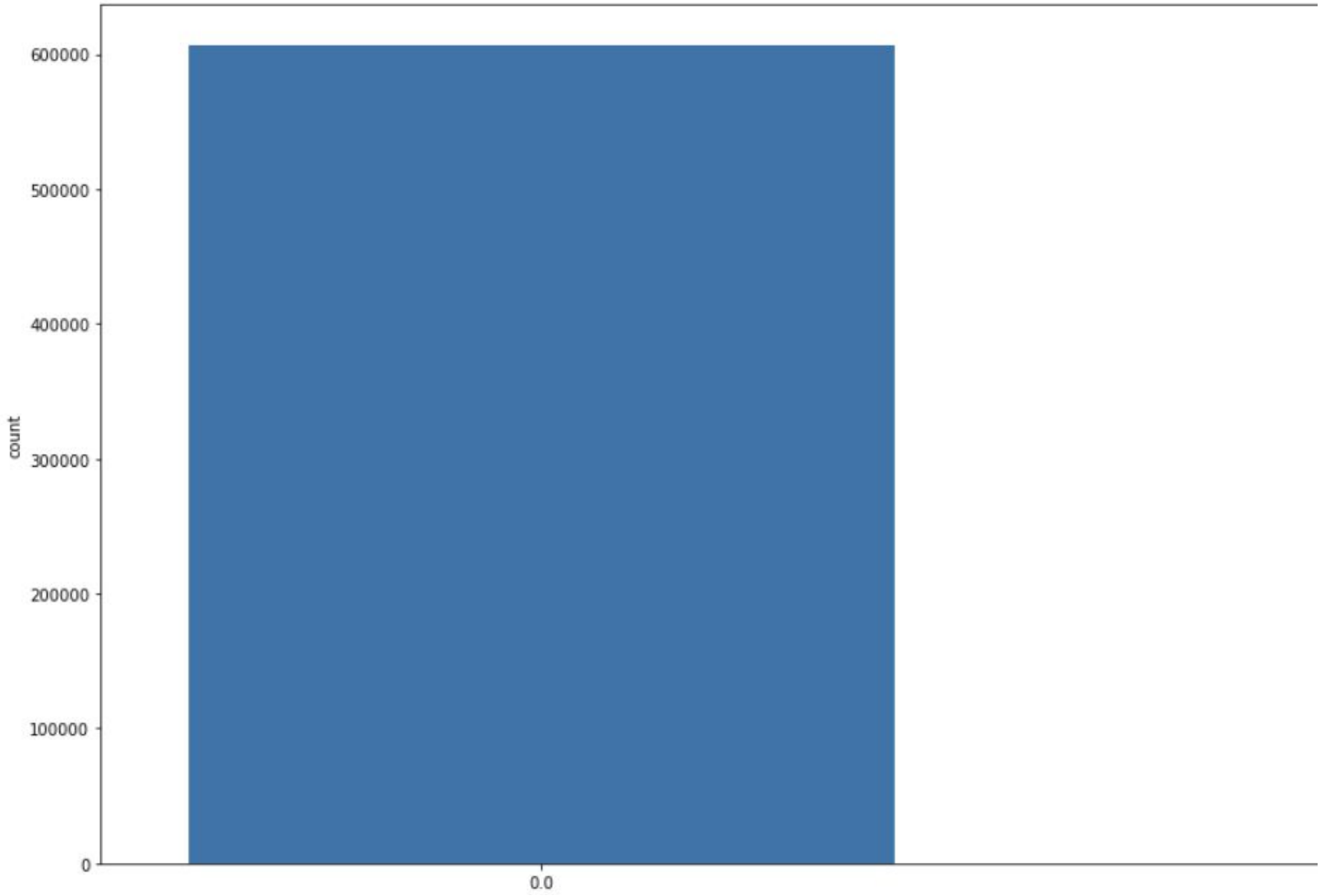
- For all years, most customers join during the month of October.

Activities Per Week Days



- For all years, the most weekly activity has been on Monday.
- Activities are minimal during the weekend.

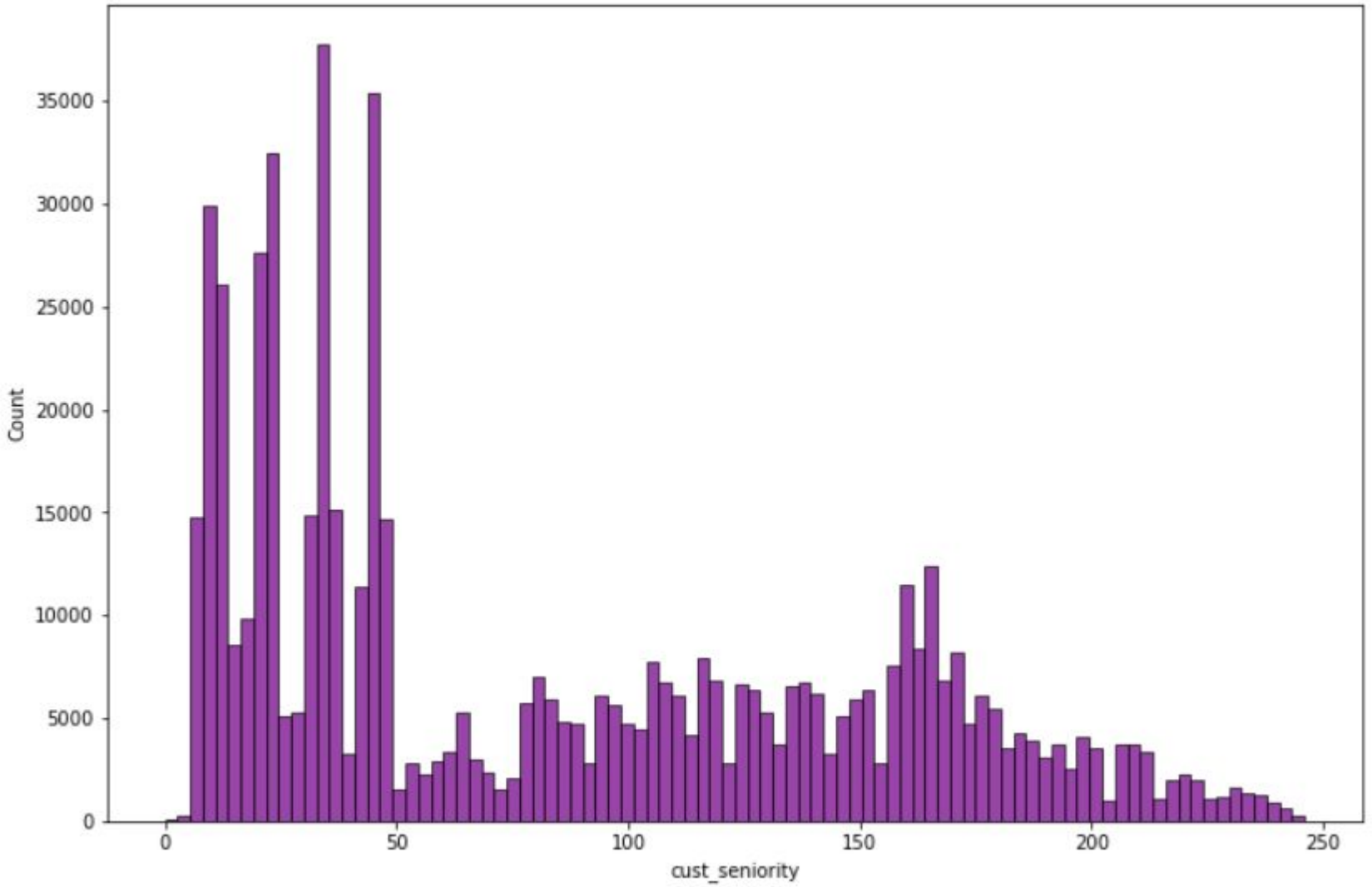
Distribution of New Customer



Only a negligible proportion of customers were registered 6 months ago.

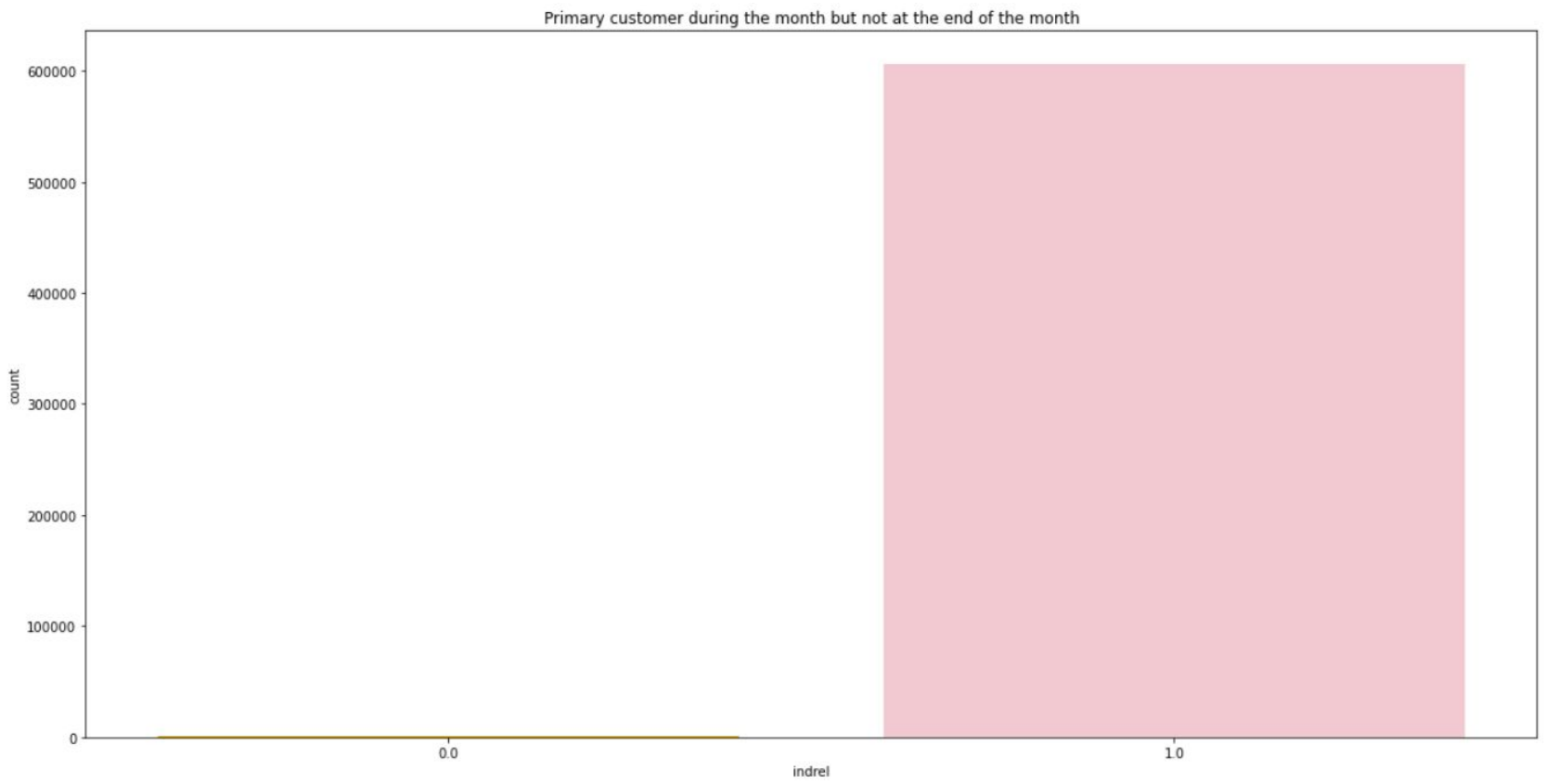
Majority of the customers are older than 6 months.

Distribution of Customer Seniority



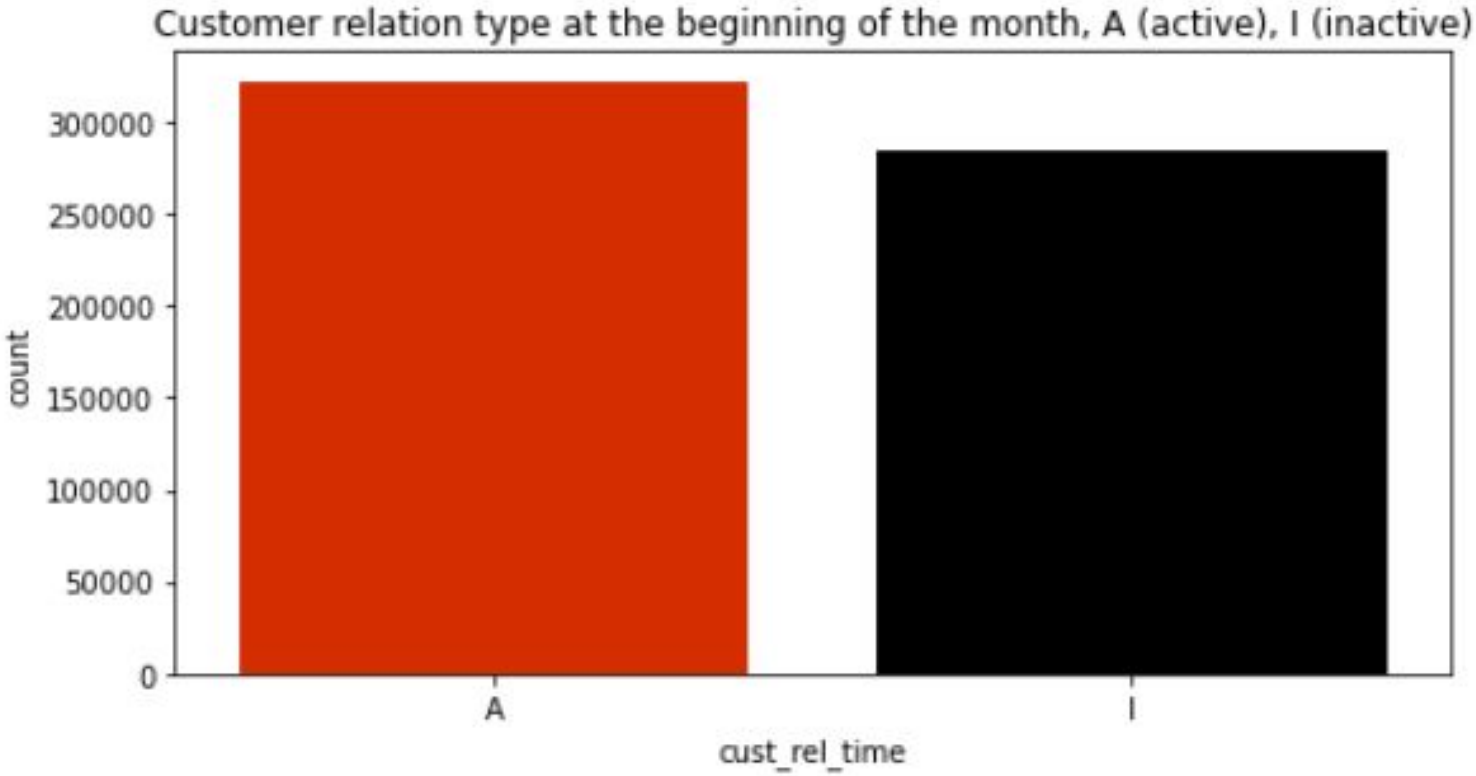
There was a huge spike in registrations recent few months compared to older months.

Distribution of Primary Customer But Not At the End Of Month



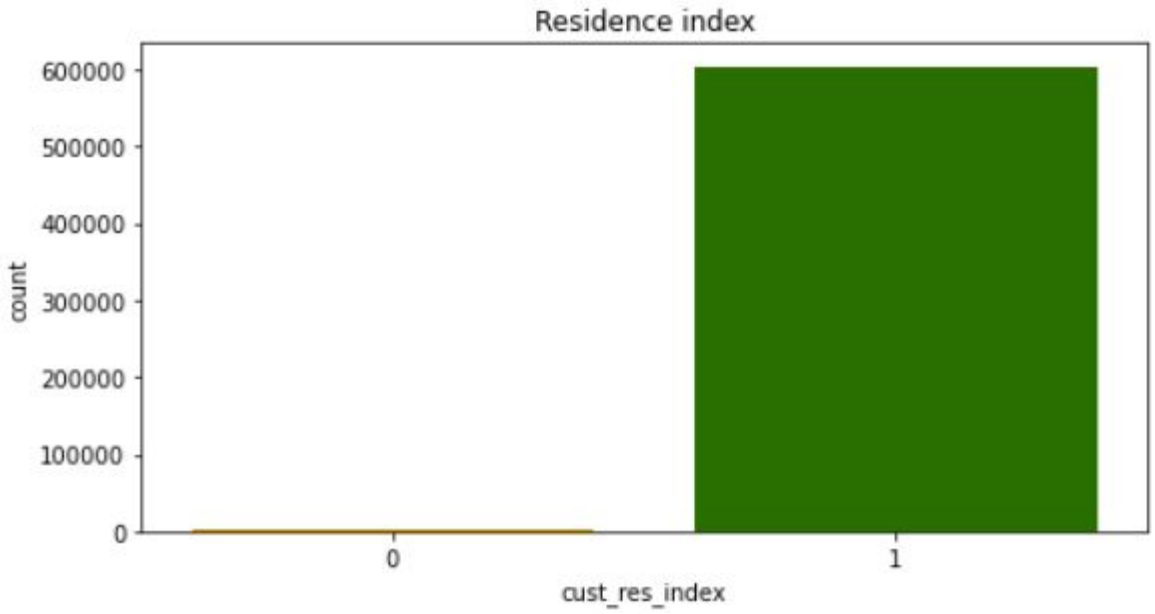
- indrel: 1 (First/Primary), 99 (Primary customer during the month but not at the end of the month).

Customer Relation Type At The Beginning Of The Month



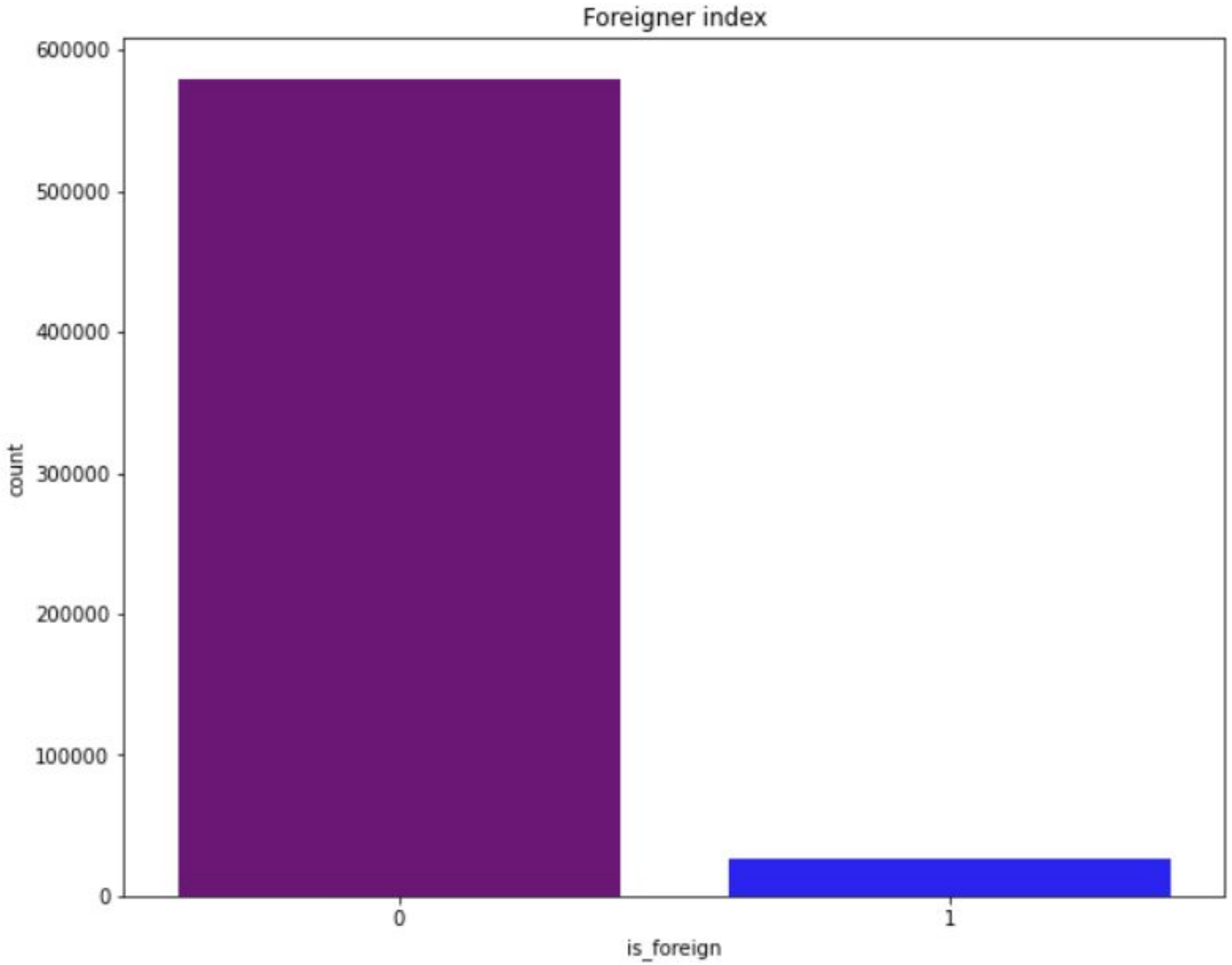
Most of the customers are almost active during the start of the month.

Customer Relation Type At The Beginning Of The Month



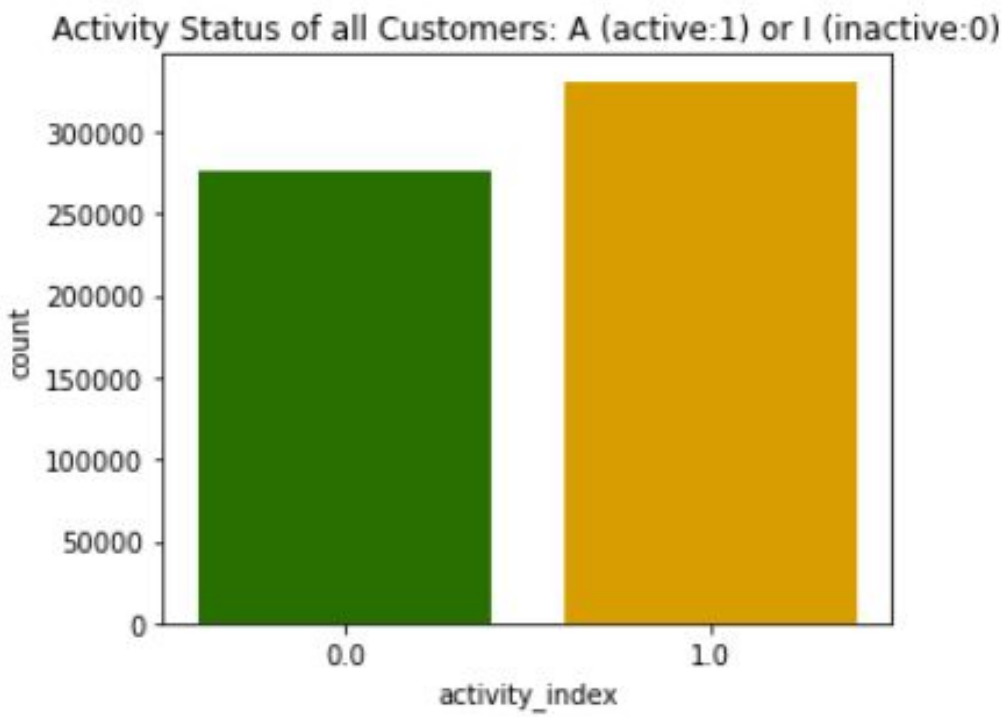
- Only a small proportion of people have bank accounts outside of their country of residence.

Total Customers by Foreigner Index



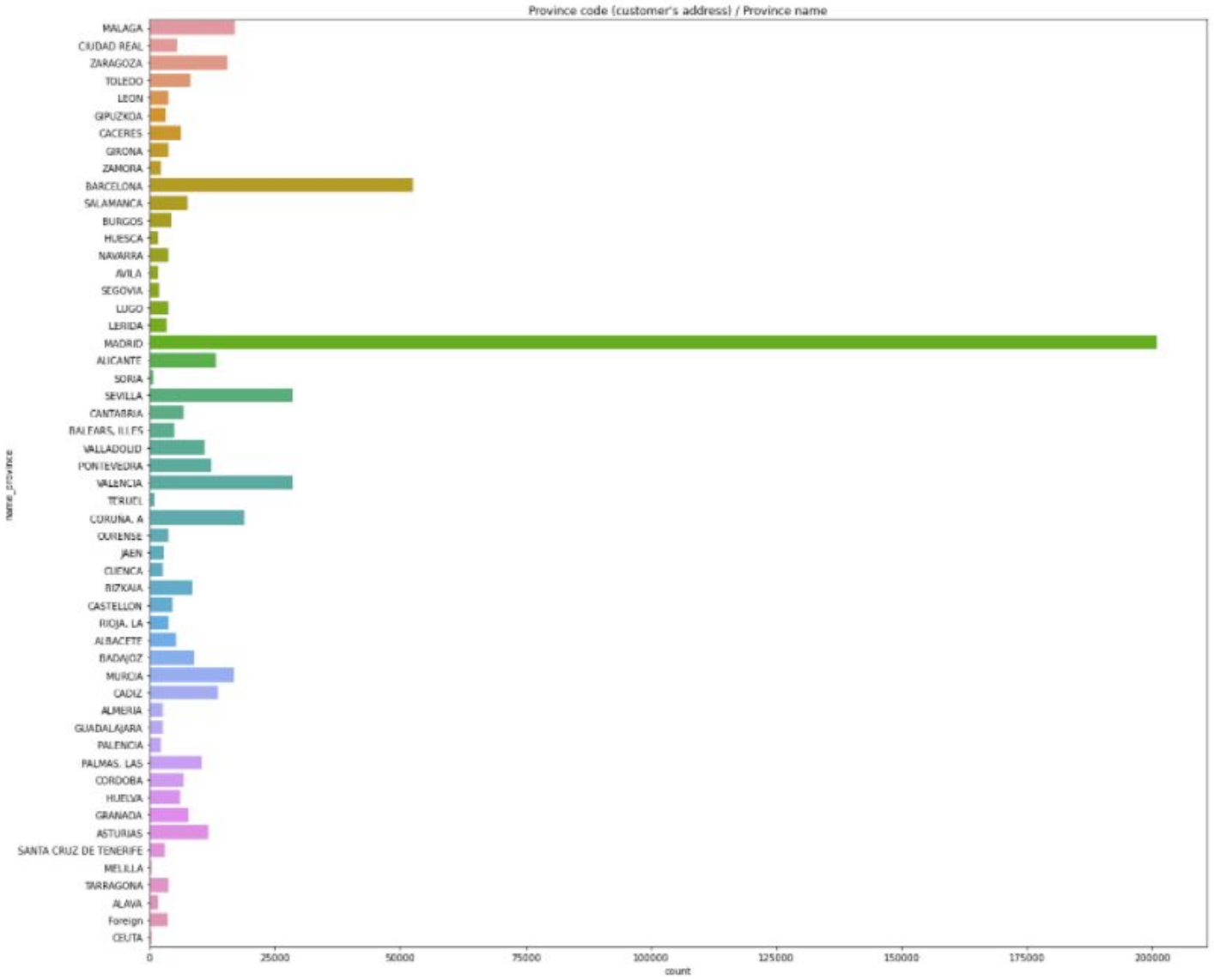
Majority of their customers are local (from Spain).

Activity Status Of All Customers



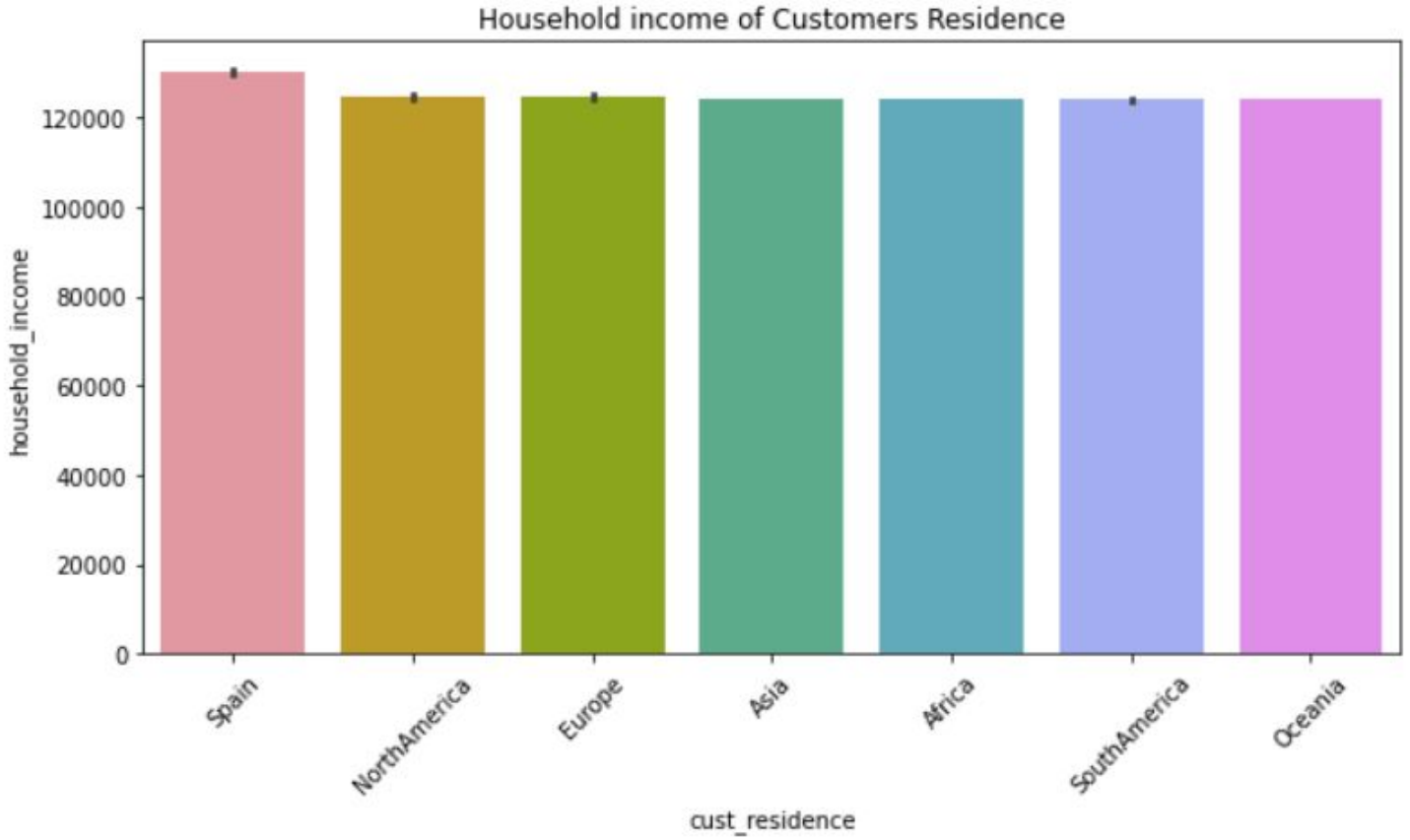
Majority of their customers are active

Distribution Of Customers By Province Code



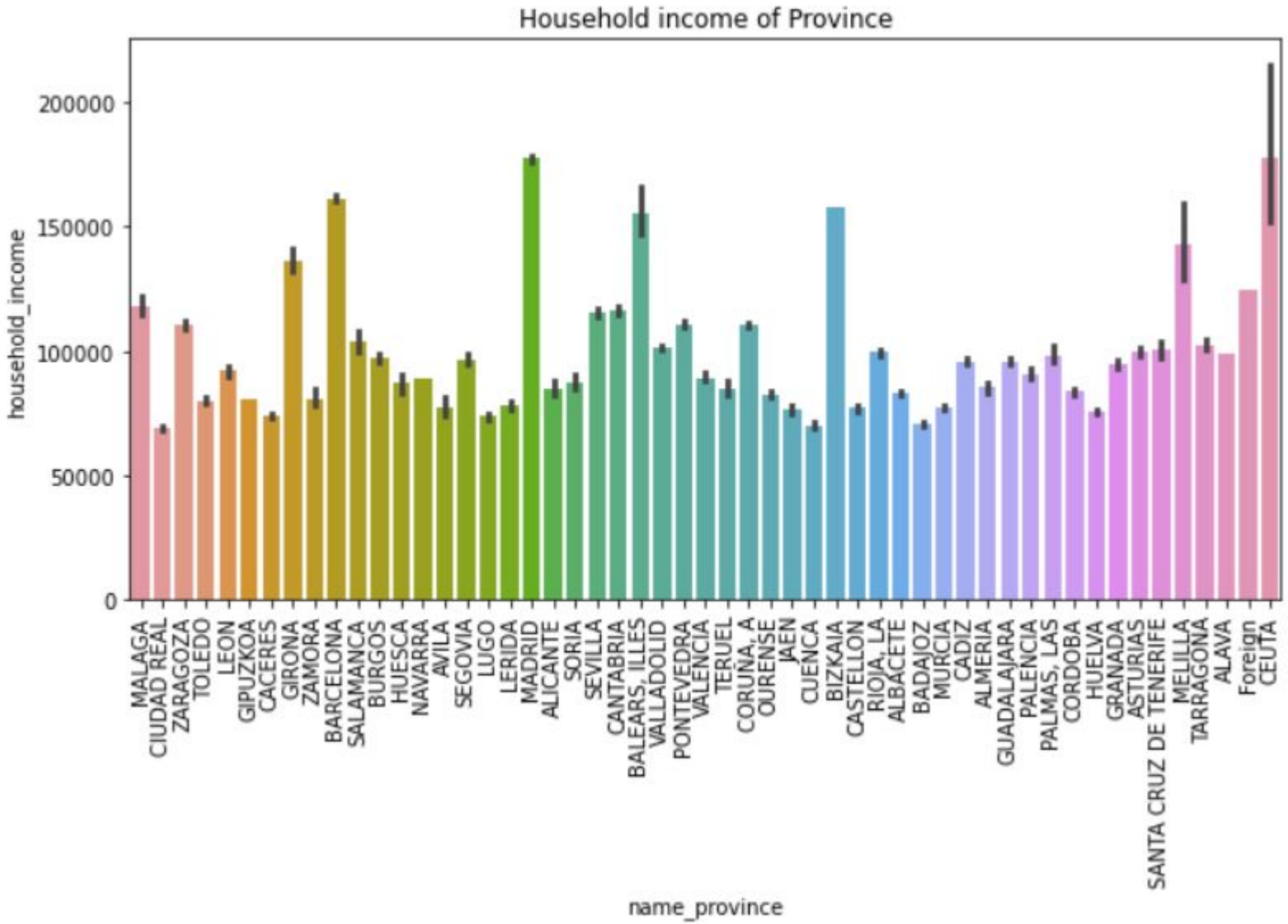
- The vast majority are in Madrid and Barcelona.

Household Income of Customer Residence



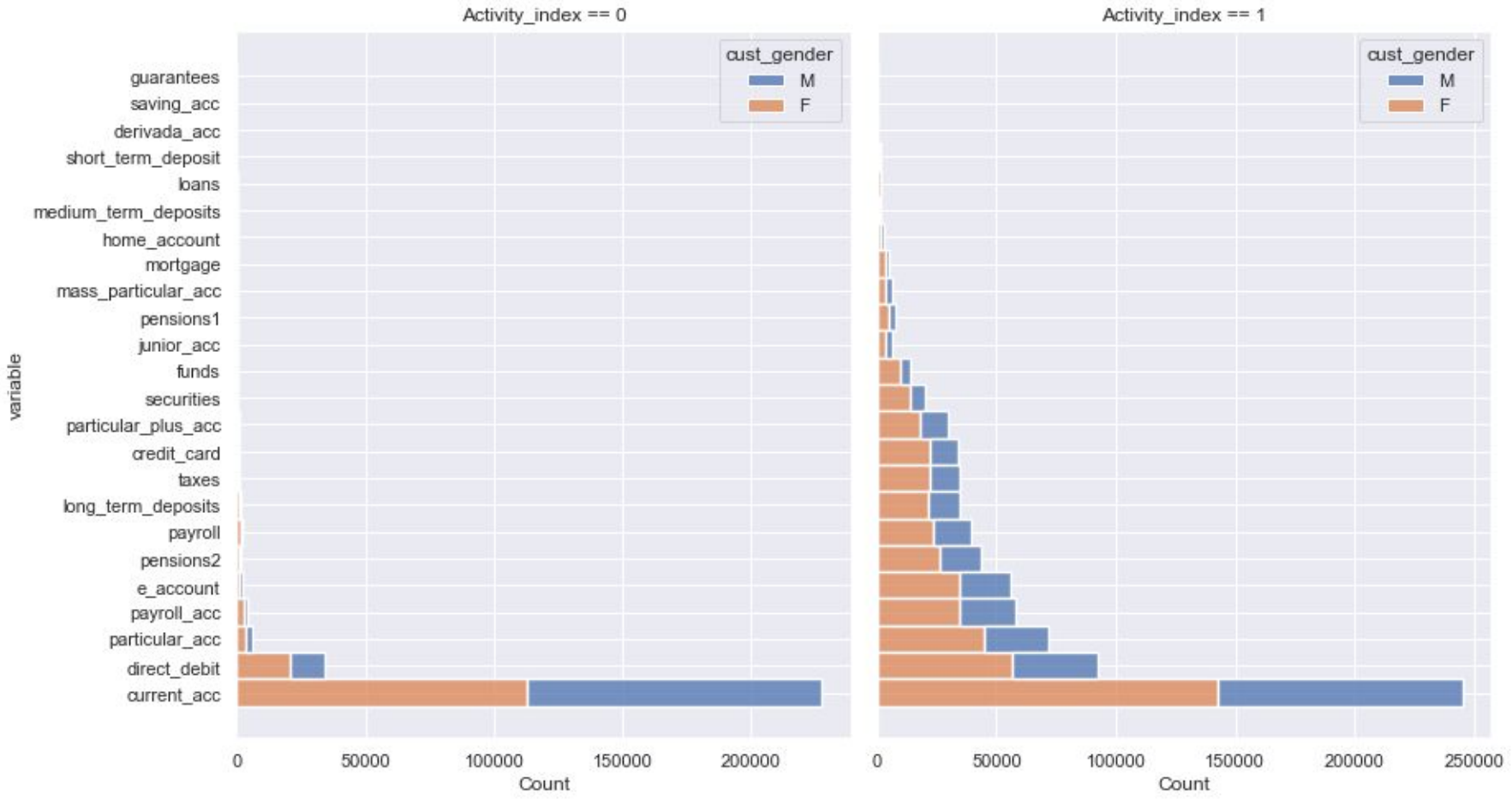
- Although the number of customers by country is mostly in Spain, there are-high-income customers in other countries.

Household Income of Province



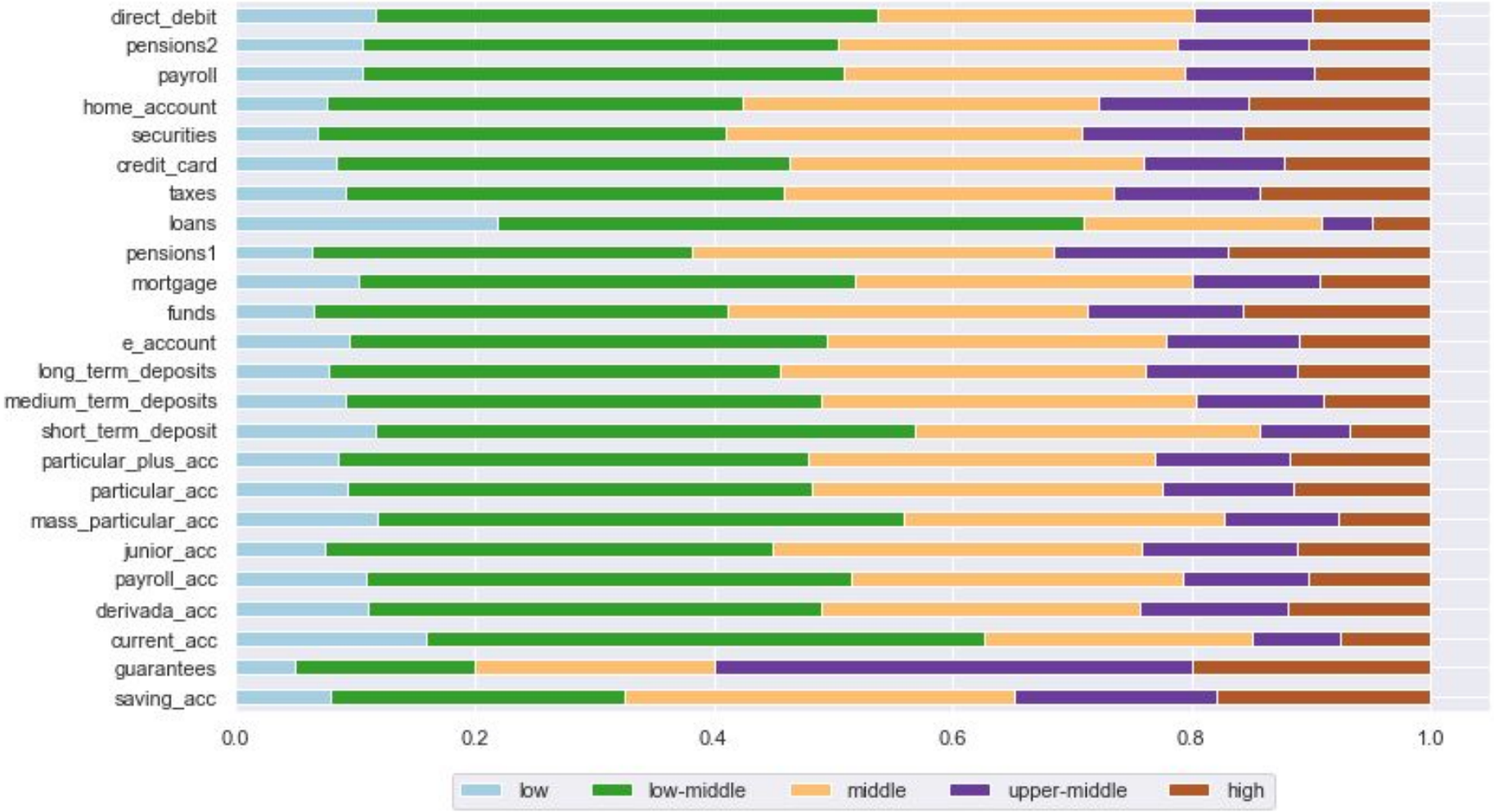
Customers from Cueta and Madrid has the highest median income compared to other provinces in Spain.

Popularity of Products by Customers Gender and Activity Index



Most of the customers holds a current account. Very few customer's holds deposits or loans.

Distribution of Products Among Customers by Income Group

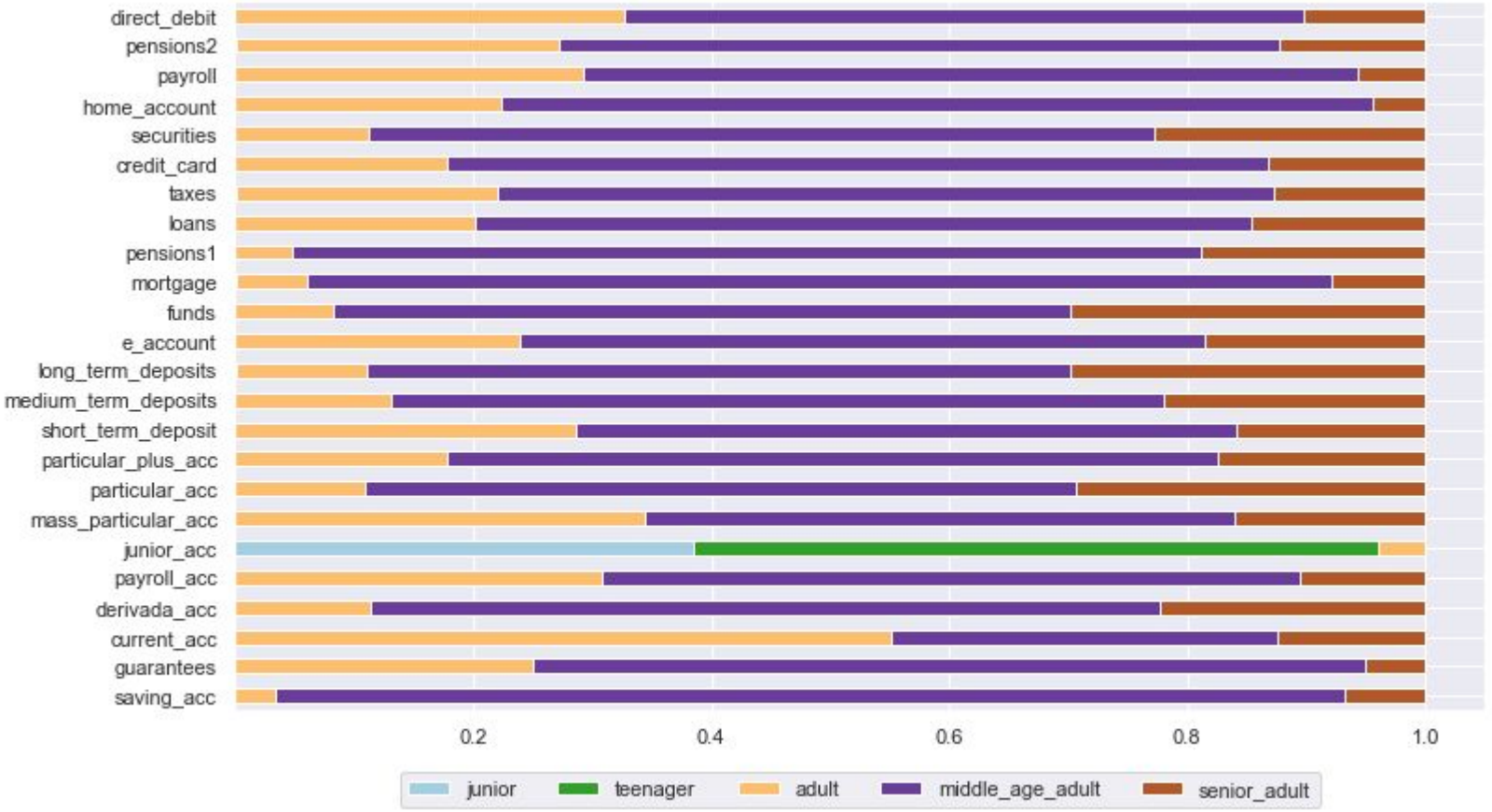


The plots shows the customers categorized by income level and the proportion of the types of accounts they hold.

Low income customers holds higher proportion of loan accounts.

upper income and high income customers holds a higher proportion of guarantees account.

Customers Age Distribution of Different Products



The plot shows proportion of accounts owned by customers categorized by age groups.

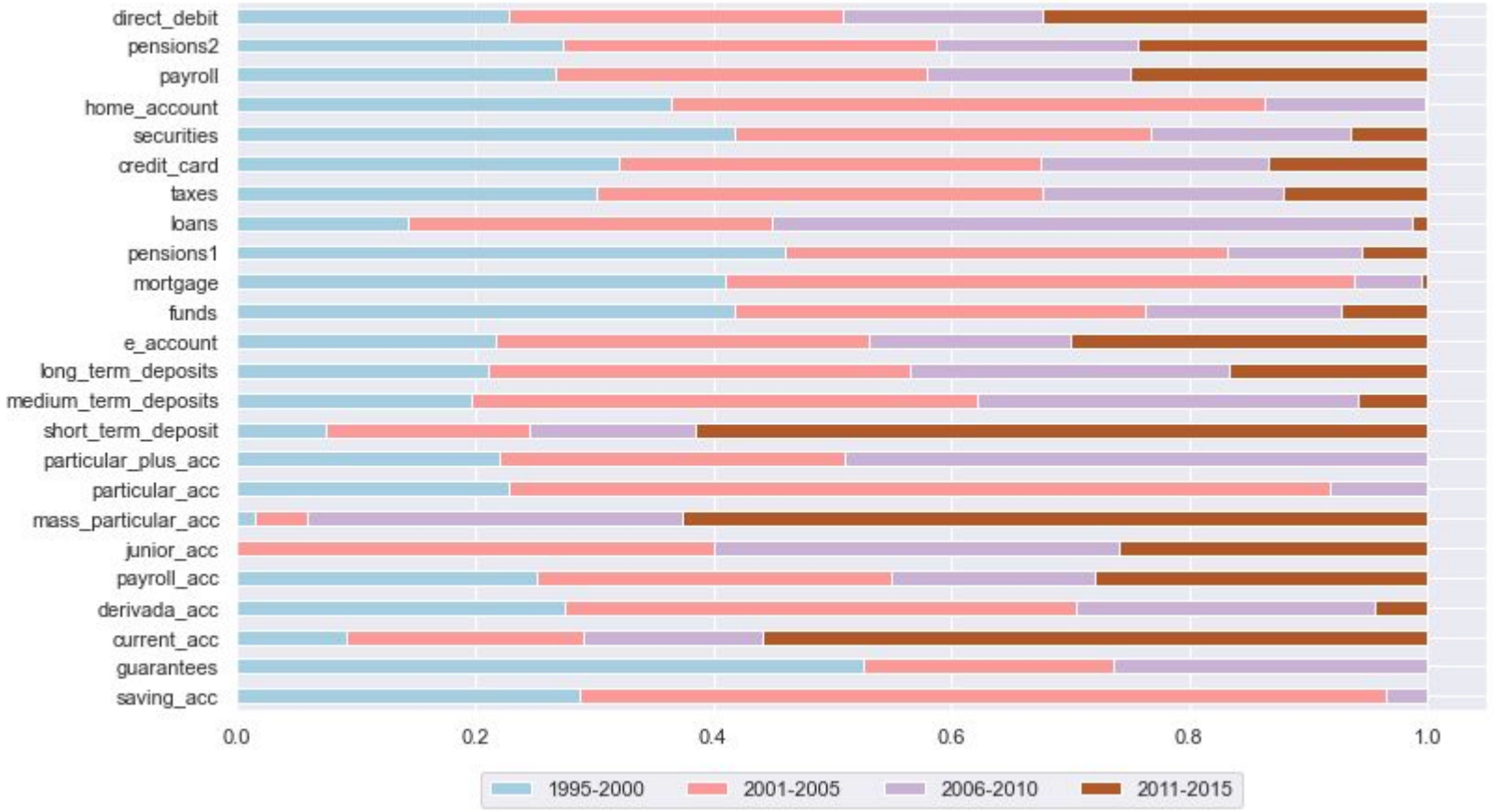
Junior and Teen customers holds the highest proportion of junior accounts.

Middle aged customers hold the highest proportion savings accounts.

Senior customers makes up the highest proportion of funds, securities and long term deposits.

Adult customers are the highest proportion using current accounts.

Customers Age Distribution of Different Products



This plots shows the types of accounts owned by customers categorized y their tenure using the banks services.

Latest customer’s mostly hold current accounts followed by short term and mas particular accounts.

Proposed Modelling Technique

1. **Feature Scaling:**
   First we will apply QuantileTransformer feature scaling on the dataset to normalize the range of independent variables. It also transforms the features to follow a uniform or a normal distribution.

2. **Dimensionality Reduction:**
   As the dataset contains lots of features, the clustering algorithms will suffer in performance due to the curse of dimensionality. Therefore, we will apply PCA on the dataset to reduce the features so that clustering algorithms can achieve better performance.

3. **Clustering:**
   We will use various clustering algorithms such as kmeans, meanshift etc that will help find 5 clusters of data with similar attributes. We will evaluate the model performances using metrics such as Silhouette coefficient and Davies-Bouldin score. Additionally we will visualize the clusters to see where the intra distance within clusters are shorter and the inter-distance among the clusters are higher.

Thank You