Capstone Project 2
Buidling Recommender System for Satander Bank Customer.

Table of Contents

Introduction

Shopping is necessity of human being as we would like to shop either product we like or our friends like.We tend to buy products recommended by people because we trust them.Taking the advantages of the same phenomenon,and because of growing amount of information on the internet and modern data analysis tools and techniques , the modern age technical revolution is taking shape in forms of systems like Machine Learning and to be specific Recommendation Engines.

Recommendation engines are data filtering tools to make use of algorithms and data to recommend most relevant items to a particular user.Not only it suggests products but also relates one which you could buy leveraging different techniques such as similar items,most frequently brought together.

With growing amount of information and significant rise in number of users ,it is becoimg important for companies to search,map and provide them with the relevant chunk of information according to their preferences and tastes.
The same phenomenon can be applied to various industries but not limited to such as healthcare,banking,manufacturing etc.In this project we are trying to apply these techniques in banking industry.

Problem Statement:

1.5 years of customers behavior data from Santander bank to predict what new products customers will purchase. The data starts at 2015-01-28 and has monthly records of products a customer has, such as "credit card", "savings account", etc.

Need to predict what additional products a customer will get in the last month, 2016-06-28, in addition to what they already have at 2016-05-28.

These products are the columns named: ind_(xyz)_ult1, which are the columns #25 - #48 in the training data. We need to predict what a customer will buy in addition to what they already had at 2016-05-28.

- Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?

According to Kaggle ,To support needs for a range of financial decisions, Santander Bank offers a lending hand to their customers through personalized product recommendations.

Under their current system, a small number of Santander's customers receive many recommendations while many others rarely see any resulting in an uneven customer experience. In their competition, Santander is challenging to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers.

OverView of Dataset:

The dataset contains 930k rows and 24 columns.
However for this project only 4% of data is selected considering the limited computing power.

The dataset used for analysis contains 54589 rows and 48 columns.

Data Wrangling:

Data Wrangling involves the checking of Data consistency and data types for the variables.

The variable names are changed in order to make them understand in English language using pandas.

Below is the list of variables with data types after column name changes.

```
Date                                datetime64[ns]
Customer_Code                       int64
Employee_Index                      object
Country                             object
Gender                              object
Age                                 float64
First_holder_Date                   datetime64[ns]
New_Customer_Index                  float64
Customer_Seniority                  float64
Primary_Customer                    float64
Customer_type_beginning_of_month    object
Customer_relation_month_begining    object
Residence_index                     object
Foreigner_index                     object
```

```
Joining_channel                    object
Deceased_index                       object
primary_address                    float64
Province_code                     float64
Province_name                        object
Active_OrNot                     float64
HouseHold_Gross_Income                  float64
Segmentation                      object
Saving_Account                       int64
Guarantees                        int64
Current_Accounts                     int64
Derivada_Account                      int64
Payroll_Account                     int64
Junior_Account                      int64
Más_particular_Account                   int64
particular_Account                   int64
particular_Plus_Account                  int64
Short_term_deposits                    int64
Medium_term_deposits                     int64
Long_term_deposits                     int64
e_account                        int64
Funds                          int64
Mortgage                         int64
Pensions                         int64
Loans                          int64
Taxes                          int64
Credit_Card                       int64
Securities                       int64
Home_Account                         int64
Payroll                       float64
Pensions                        float64
Direct_Debit                       int64
tot_products
```

MISSING VALUES:

```
Date                        0.000000
Customer_Code                   0.000000
Employee_Index                  0.001906
Country                      0.001906
Gender                       0.001906
Age                         0.000000
First_holder_Date                0.001906
```

```
New_Customer_Index               0.001906
Customer_Seniority               0.000000
Primary_Customer                 0.001906
Last_date_as_primary_customer    0.998245
Customer_type_beginning_of_month 0.011152
Customer_relation_month_begining 0.011152
Residence_index                  0.001906
Foreigner_index                  0.001906
Spouse_index                     0.999811
Joining_channel                  0.014190
Deceased_index                   0.001906
primary_address                  0.001906
Province_code                    0.006736
Province_name                    0.006736
Active_OrNot                     0.001906
HouseHold_Gross_Income           0.206072
Segmentation                     0.014341
Saving_Account                   0.000000
Guarantees                       0.000000
Current_Accounts                 0.000000
Derivada_Account                 0.000000
Payroll_Account                  0.000000
Junior_Account                   0.000000
Más_particular_Account           0.000000
particular_Account               0.000000
particular_Plus_Account          0.000000
Short_term_deposits              0.000000
Medium_term_deposits             0.000000
Long_term_deposits               0.000000
e_account                        0.000000
Funds                            0.000000
Mortgage                         0.000000
Pensions                         0.000000
Loans                            0.000000
Taxes                            0.000000
Credit_Card                      0.000000
Securities                       0.000000
Home_Account                     0.000000
Payroll                          0.001132
Pensions                         0.001132
Direct_Debit                     0.000000
dtype: float64
```

There are missing values in various columns as observed above.

Here I tried to delete columns which have more than 80% of missing data such as "Last_date_as_primary_customer" and spouse Index

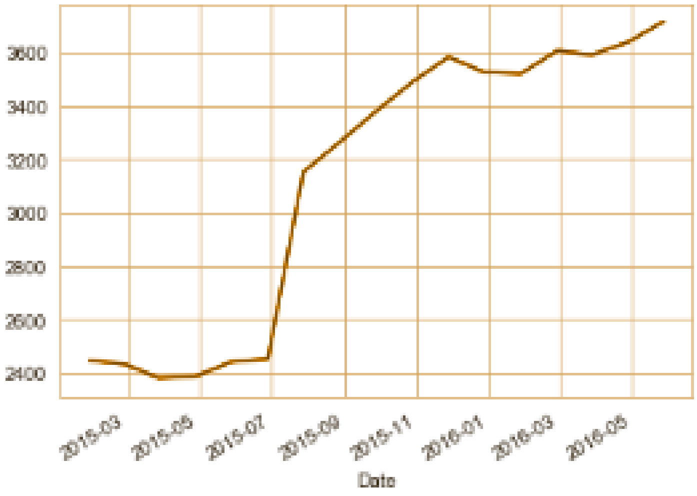
Exploratory Data Analysis:

The first few columns of Dataframe are metadata about the customers such as Age,Gender,Demographic information.

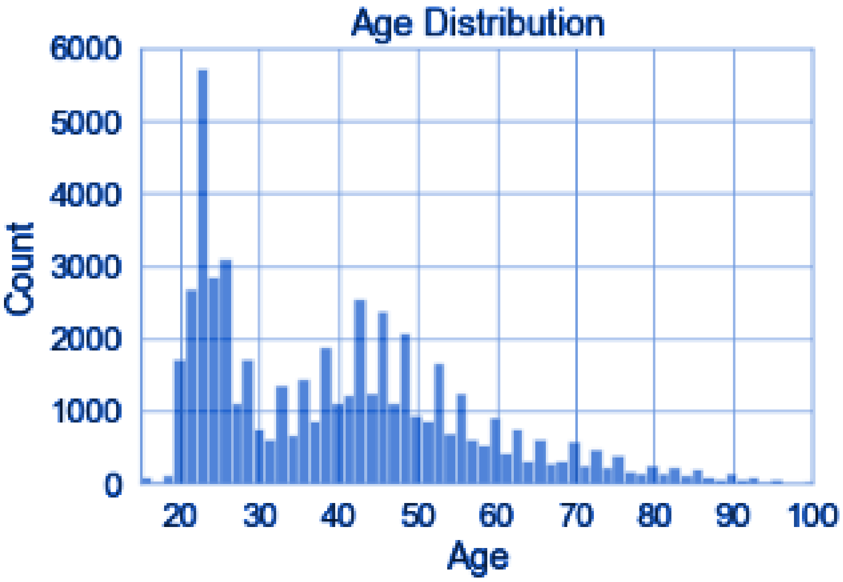
The column no's 24 onwards in dataframe are about the various accounts customer holds.Such information is core to build recommendation engine in order to find patterns and compare user behaviour as to purchase pattern using different algorithms.

Before that lets to Exploratory data analysis in order to understand the data.

It can be seen from below graph there is significant rise in number of customers bank owns.

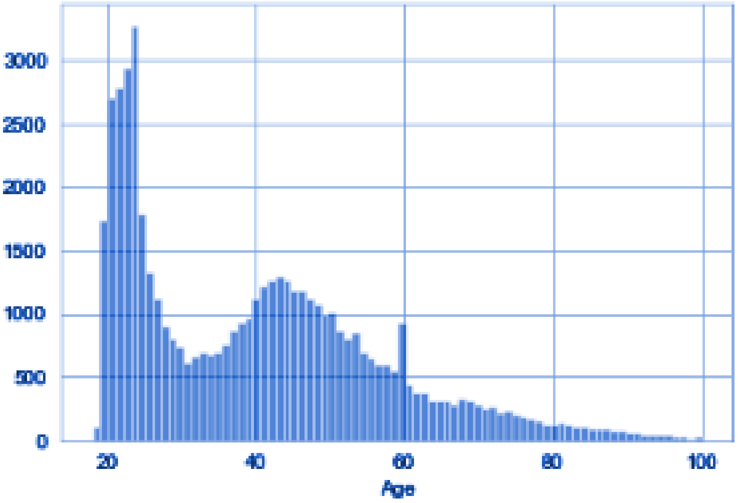


When we try to age distribution of different customers we can see the following pattern



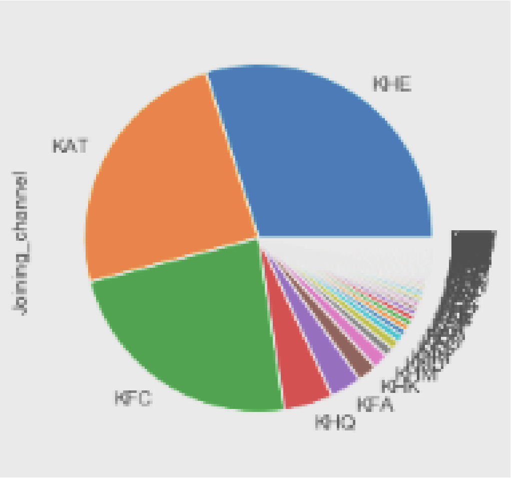
Based on above graph we can see the patterns in 0-30,30-60 and 60-90 range.

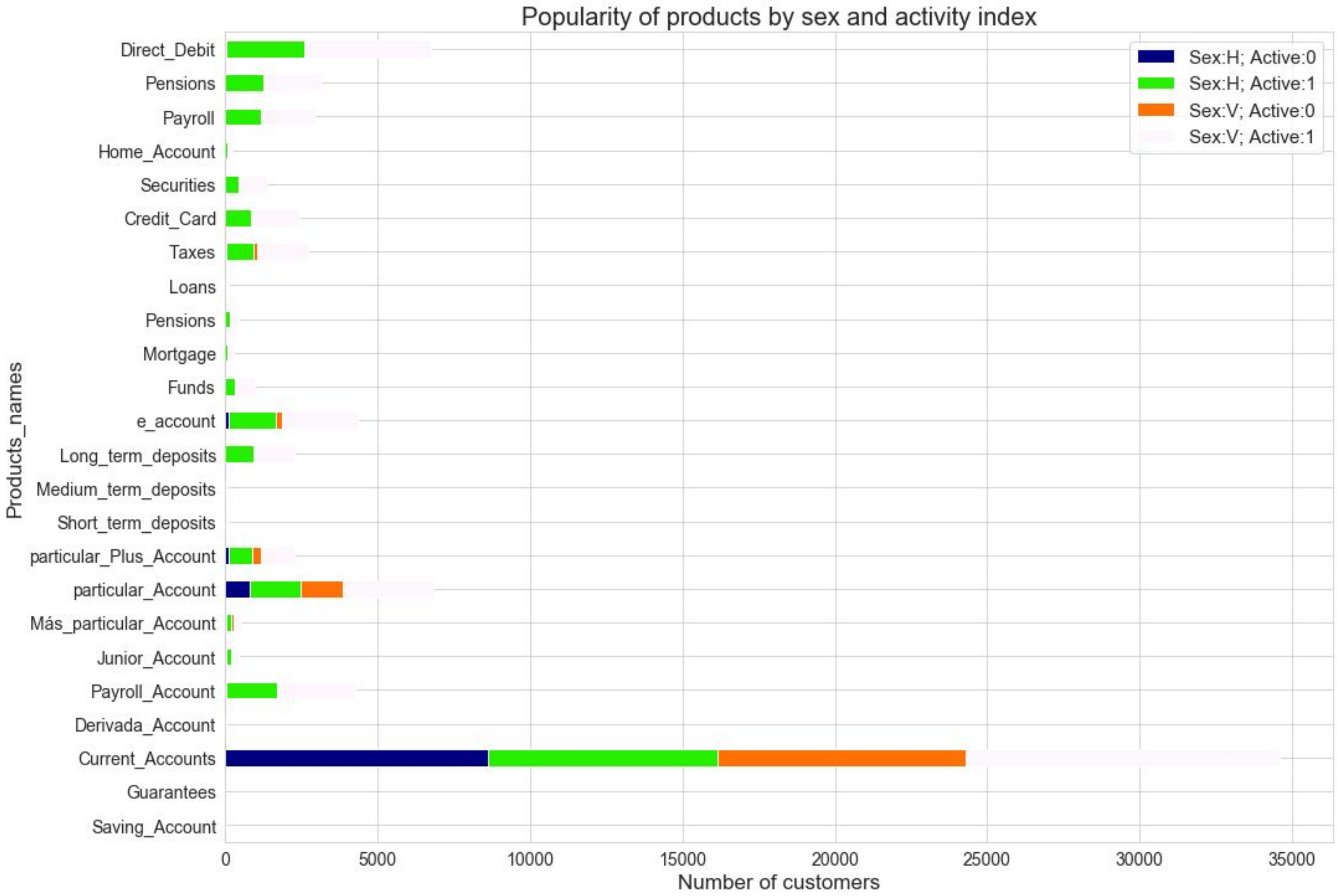
We can normlize them using Pandas library so the final graph appears to be as below:



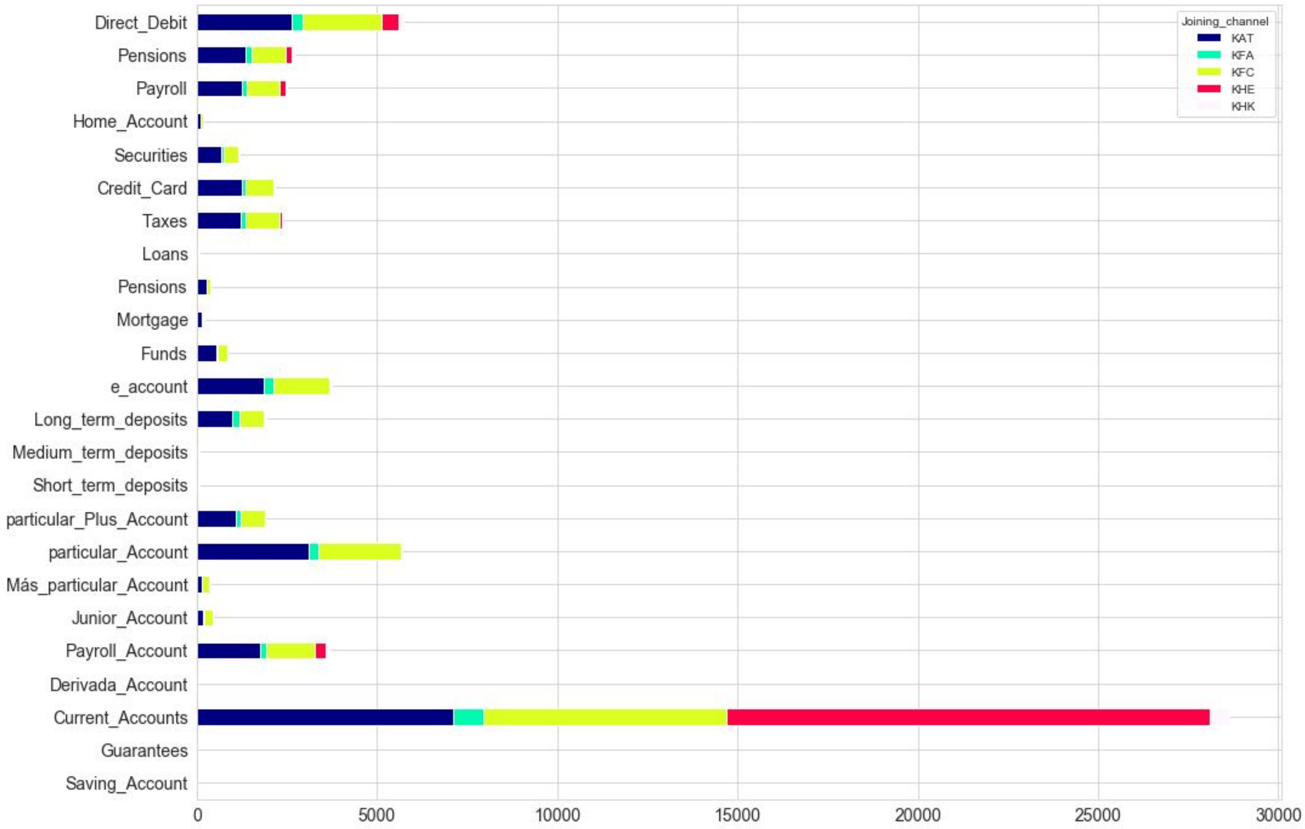
The piechart shows the various joining channels customers used as below:

We can see that KHE,KFC AND KAT are dominant.

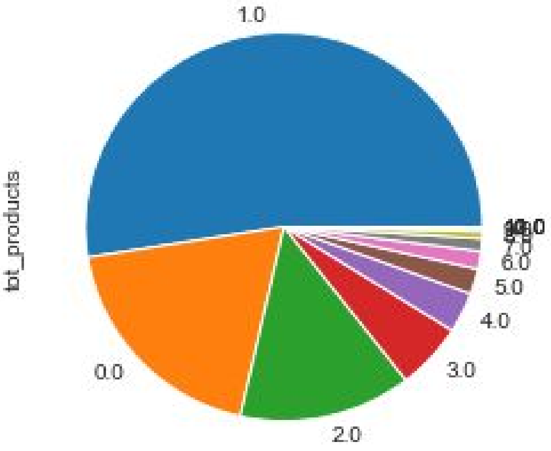




Following plot gives the idea of various channel used by customers against different accounts.

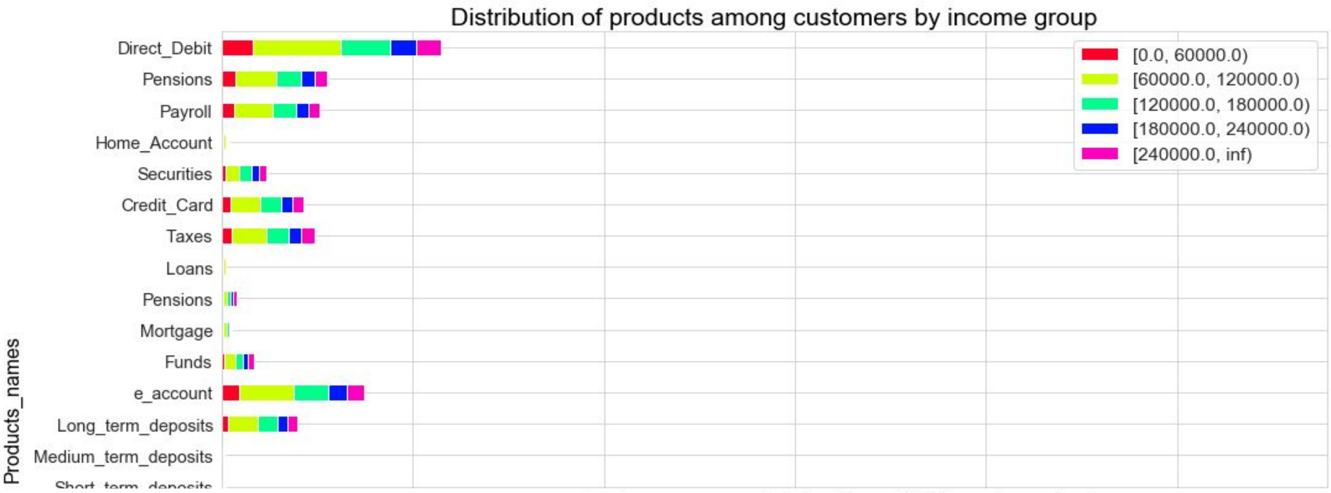


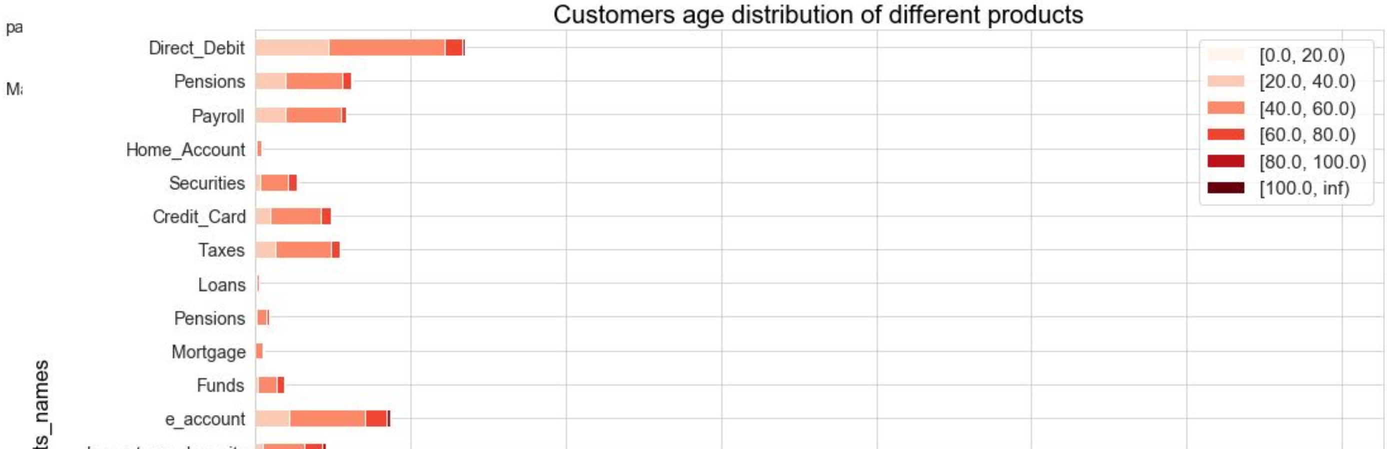
Following pie chart gives the information about total product owned by Customers.

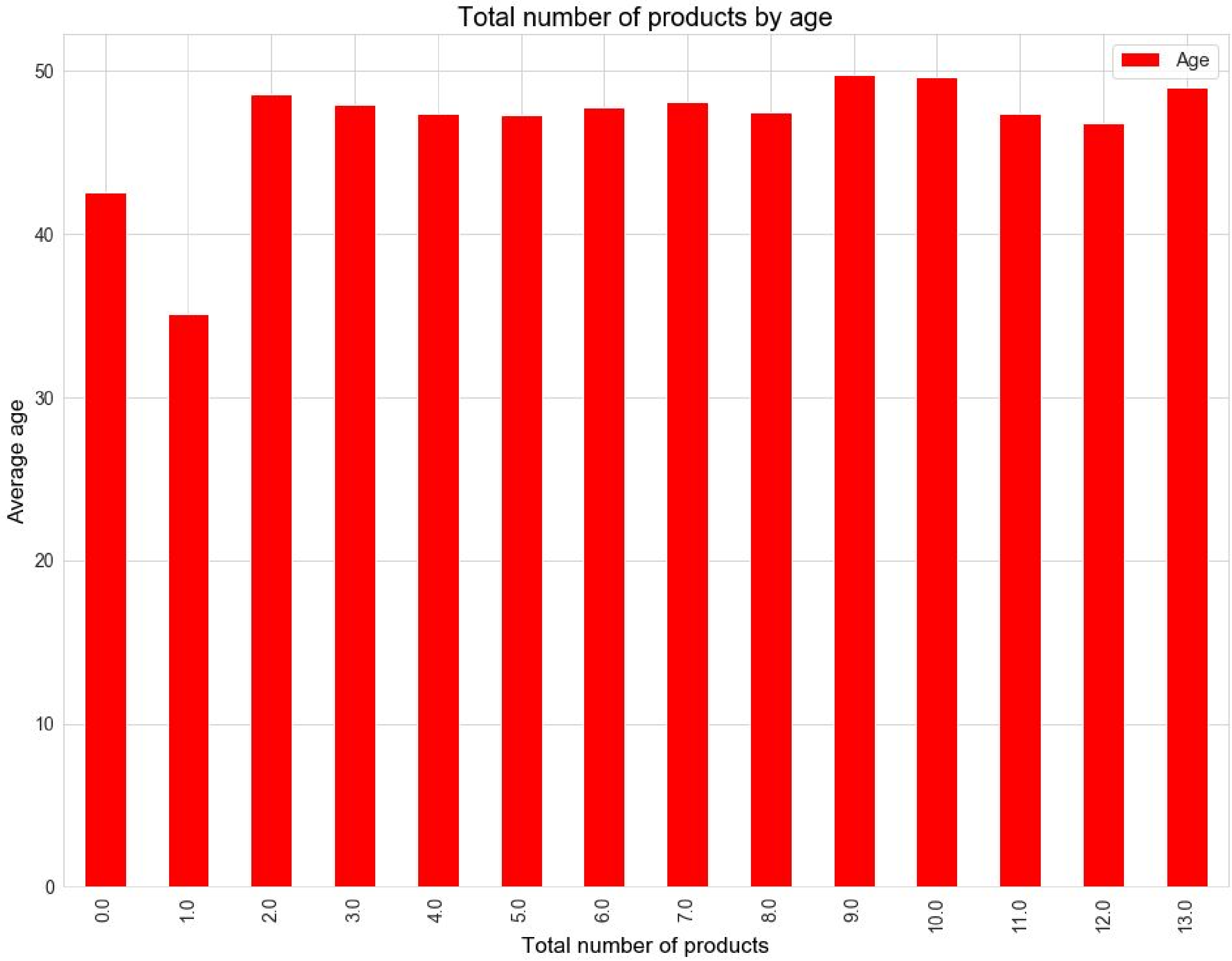


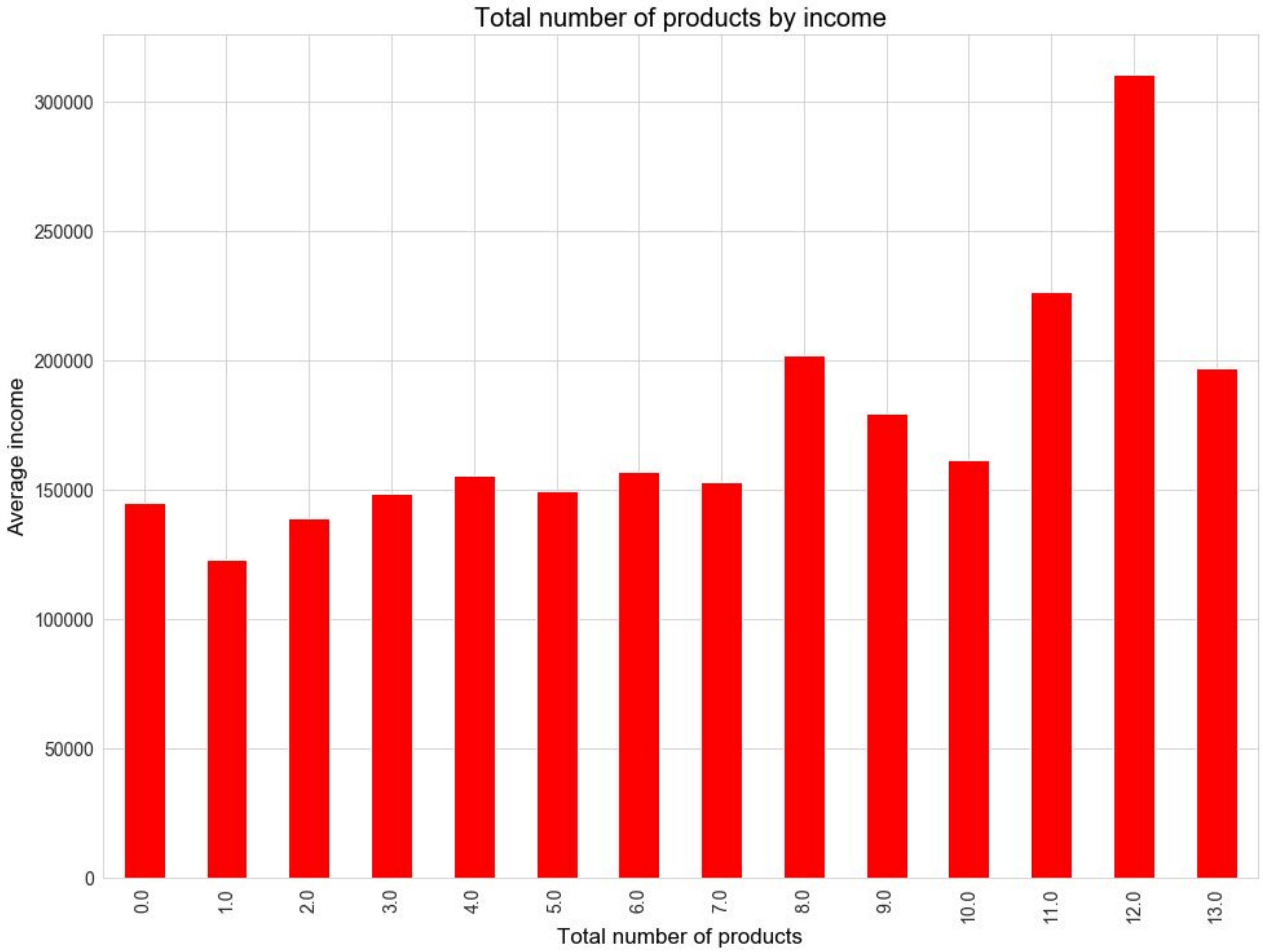
We can observe that most customers have atleast 1 Account.Also observe the 0 accounts ,that is the possibility of inactive customers which we will explore in coming sections.

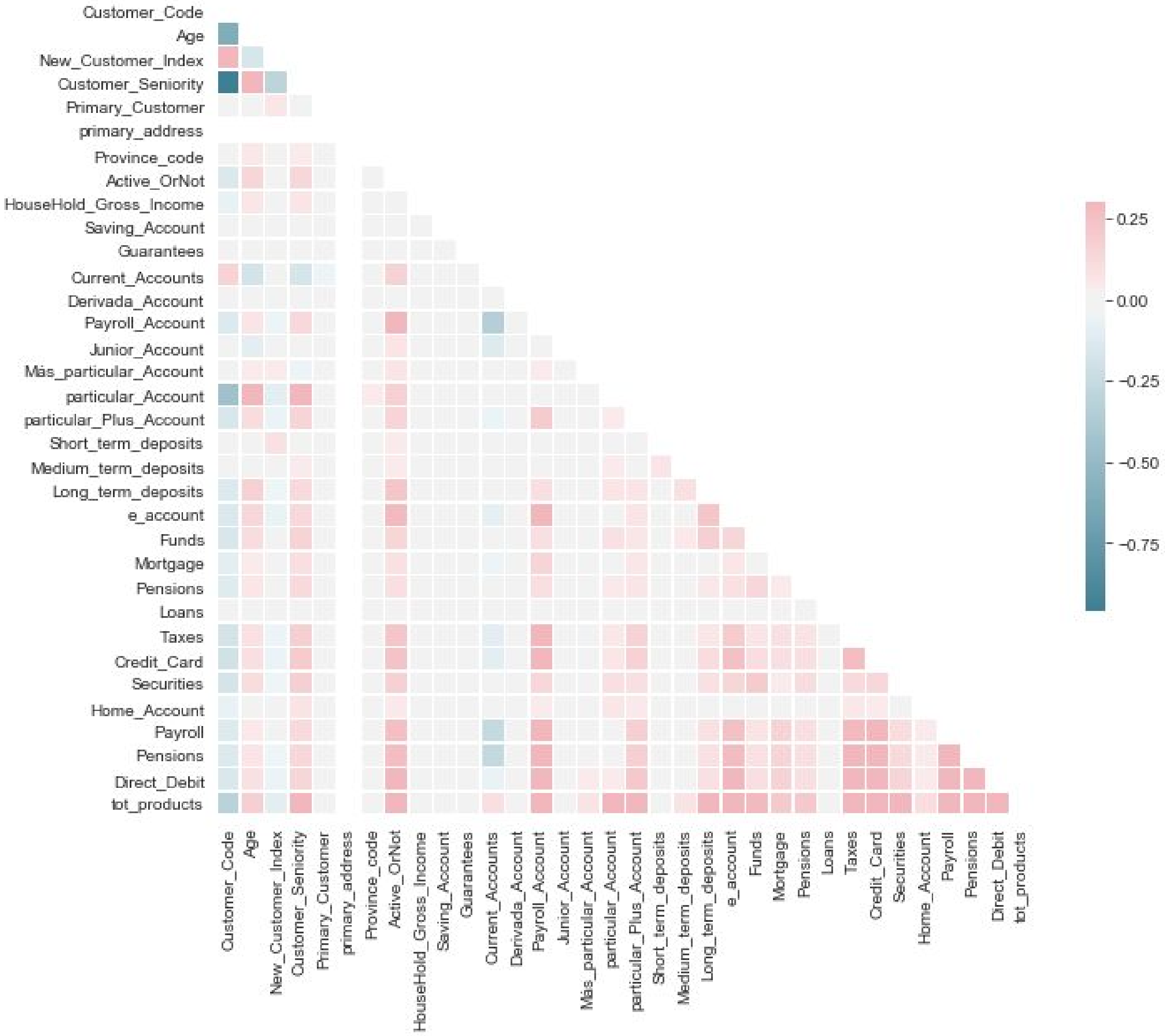
Lets try to visualise the customer purchase behaviour based on there income.






Total number of products by age
Age
Average age
0
10
20
30
40
50
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
Total number of products


Total number of products by income
Average income
300000
250000
200000
150000
100000
50000
0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
11.0
12.0
13.0
Total number of products


Customer_Code
Age
New_Customer_Index
Customer_Seniority
Primary_Customer
primary_address
Province_code
Active_OrNot
HouseHold_Gross_Income
Saving_Account
Guarantees
Current_Accounts
Derivada_Account
Payroll_Account
Junior_Account
Más_particular_Account
particular_Account
particular_Plus_Account
Short_term_deposits
Medium_term_deposits
Long_term_deposits
e_account
Funds
Mortgage
Pensions
Loans
Taxes
Credit_Card
Securities
Home_Account
Payroll
Pensions
Direct_Debit
tot_products
0.25
0.00
−0.25
−0.50
−0.75

Inferential Statistics

Following hypoyheis were tested (see the notebook for detailed analysis.

Total number of products owned by men are more than women

Null Hypothesis=Total number of products owned by men and women are same.

Alternate Hypotheis = Total numnber of product owned by men and women are different.

Conclusion- Based on P value it is observed that there is difference between the total number of products owned by men and women.

People with more Age have more accounts

Null Hypothesis**:** There is a relationship between Age and Total Products

Alternate Hypotheis **:** There is not a relationship between Age and Total Products
Conclusion- Based on P value it was observed that there is no relation between Age and number of Prodcuts

People with more income have more accounts

Based on p-value and correlation we can conclude that there is no relation between

income and number of products owned.

Summary :

1.There is overall increase in number of customers from 2015 to 2016

2.52% Customers own 1 Account, 14% owns 2 Accounts,5% owns 3 Accounts

3.Current account is most Dominant product owned by customers

4.There are inactive customers.

5.There is no relation between Age and Number of Products owned

6.There is no relation between salary and number of Products owned.