DATA GLACIER VIRTUAL INTERNSHIP

CROSS SELLING RECOMMENDATION-GROUP PROJECT

WEEK 10: DELIVERABLES -EDA

GROUP NAME: HEGY

Team members:

Name: B. Harika
Email: harikabreddy444@gmail.com
Country: India
College/ Company: Data Glacier
Specialization: Data Analyst

Name: Yusuf Yuhan
Email: yusufyuhan98.yy@gmail.com
Country: Srilanka
College/ Company: The Open University of Srilanka
Specialization: Data Analyst

Name: Ebaghae Imhanlahimi
Email: imhanlahimiw@gmail.com
Country: America
College/ Company: Data Glacier
Specialization: Data Analyst

Name: Gladys Kalas
Email: gladys@kalas.me
Country: USA
College/ Company: Data Glacier
Specialization: Data Analyst

Contents

Problem description:

XYZ Credit Union is a financial institution based in Latin America that offers a variety of banking products to its customers, including credit cards, deposit accounts, retirement accounts, and safe deposit boxes. While the credit union has been successful in selling these products individually, it has not been as successful in cross-selling its products to existing customers. The lack of success in cross-selling suggests that there may be several barriers preventing XYZ Credit Union from selling additional products to its existing customers. To address this problem, XYZ Credit Union has decided to work with ABC Analytics, a data analytics consulting firm, to identify the barriers to cross-selling and develop strategies to overcome them. ABC Analytics will work with the credit union to analyze Customer data and information to identify patterns and trends, and develop targeted marketing strategies that are designed to increase their possibilities and revenues in the credit union's quest to cross sell banking products to the customers.

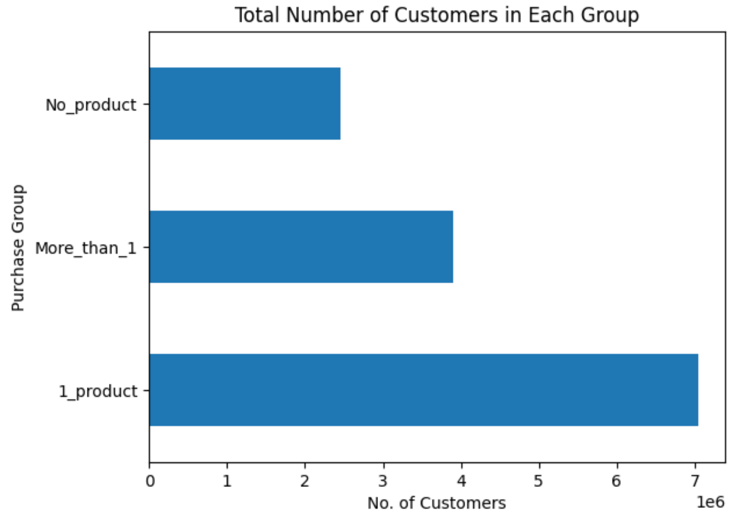
Data:

The data for the EDA is a cleaned data set which is

- Free from duplicates
- Checked for missing values and handled the missing values
- Detected outliers and treated the outliers with appropriate methods
- Checked the data types and appropriate conversions in places of incorrect types
- Organized feature headings (from Spanish language to English)

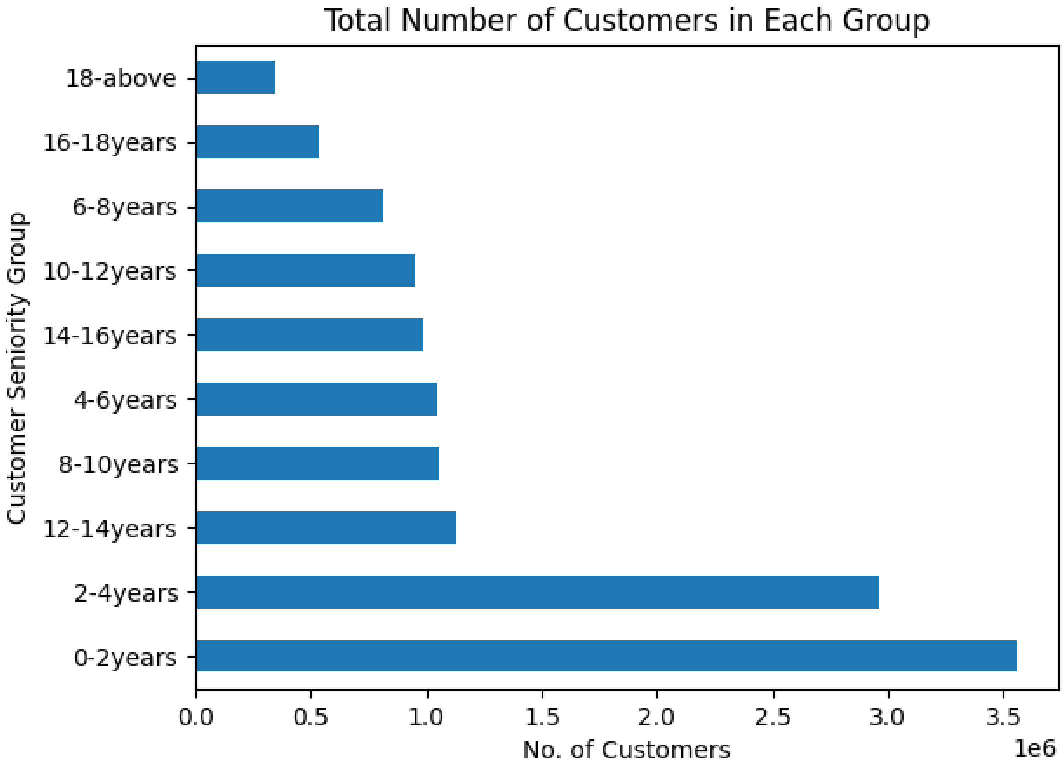
EDA performed on the data:

- Visualizing categorical data (Purchase group):



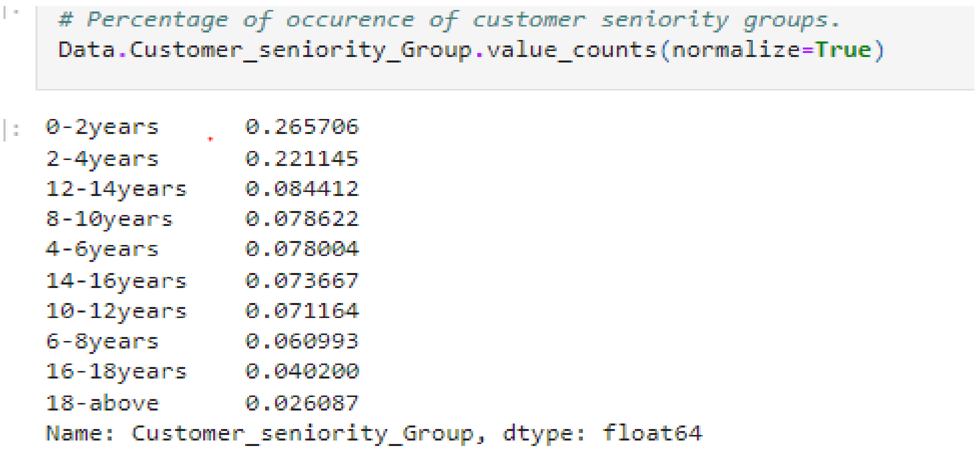
1. From the bar chart above, we can see that majority of customers purchased only 1 product.
2. Some customers did not purchase any product.
3. Some customers purchased more than 1 product

- Visualize the categorical data (Customer_seniority_Group):



1. From the bar chart above, we can see that majority of the customers have been customers for at most 2years.
2. This is closely followed by customers who have been in the bank for at most 4years.
3. Customers who have been in the bank for at least 18 years are the smallest group of customers.

- Percentage of occurrence of customer seniority groups:

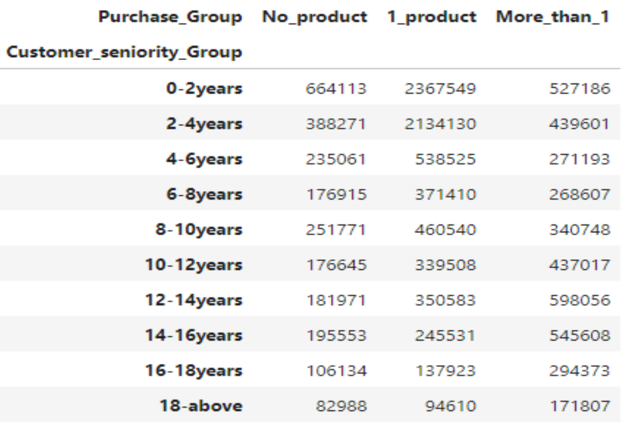


```
# Percentage of occurence of customer seniority groups.
Data.Customer_seniority_Group.value_counts(normalize=True)
```

```
0-2years       0.265706
2-4years       0.221145
12-14years     0.084412
8-10years      0.078622
4-6years       0.078004
14-16years     0.073667
10-12years     0.071164
6-8years       0.060993
16-18years     0.040200
18-above       0.026087
Name: Customer_seniority_Group, dtype: float64
```

1. Approximately 27% of customers have been in the bank for at most 2 years.
2. Approximately 22% of customers have been in the bank for at least 2 years and at most 4 years.

- Contingency Table: Frequency of a particular customer seniority group's product purchase

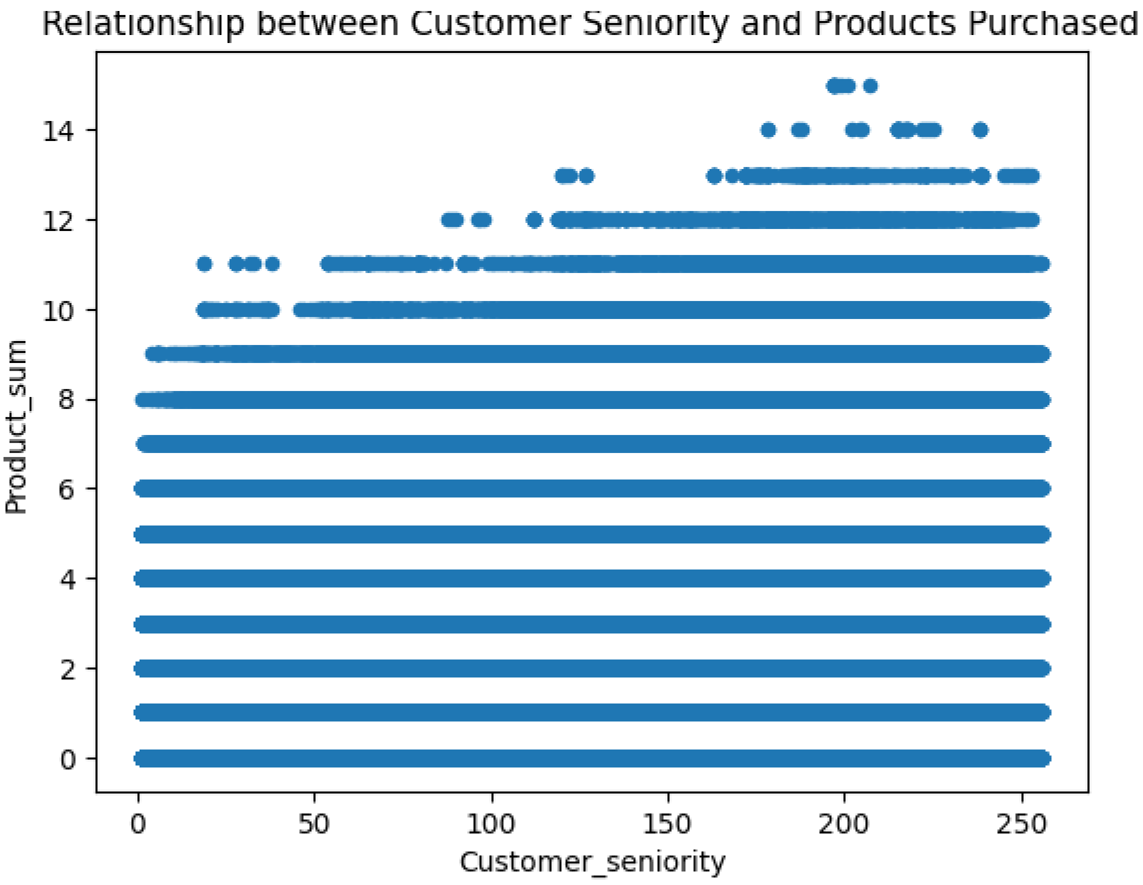


Purchase_Group	No_product	1_product	More_than_1
Customer_seniority_Group			
0-2years	664113	2367549	527186
2-4years	388271	2134130	439601
4-6years	235061	538525	271193
6-8years	176915	371410	268607
8-10years	251771	460540	340748
10-12years	176645	339508	437017
12-14years	181971	350583	598056
14-16years	195553	245531	545608
16-18years	106134	137923	294373
18-above	82988	94610	171807

1. Majority of customers who have been with the bank within 0 - 10 years purchase only one product.

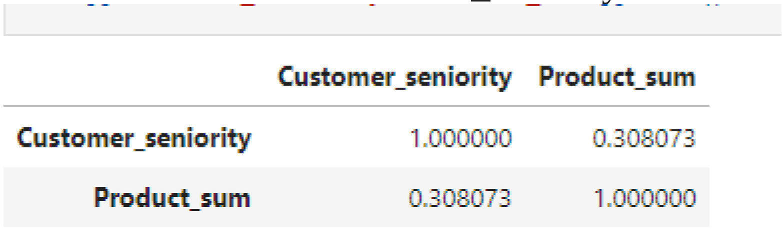
2. Majority of customers who have been in the bank for at least 10 years and above purchase more than 1 product.

- Relationship between customer seniority and product purchase:



The scatter plot above shows that there is no noticeable relationship between customer seniority and product sum.

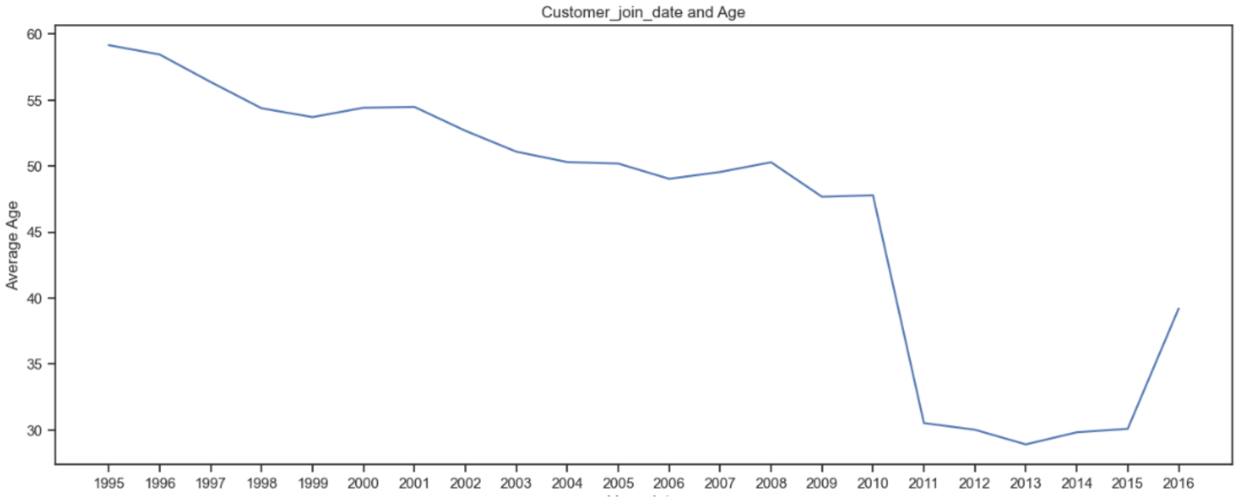
- Correlation between Customer_seniority and Product_sum:



	Customer_seniority	Product_sum
Customer_seniority	1.000000	0.308073
Product_sum	0.308073	1.000000

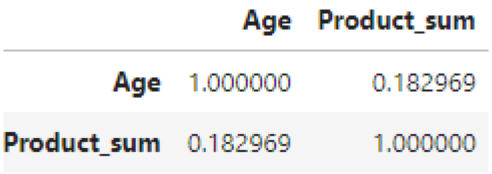
1. The correlation coefficient of customer seniority and product sum is 0.303. The correlation coefficient indicates that there is a very weak positive relationship customer seniority and product sum.

- Customer_join_date and Age:



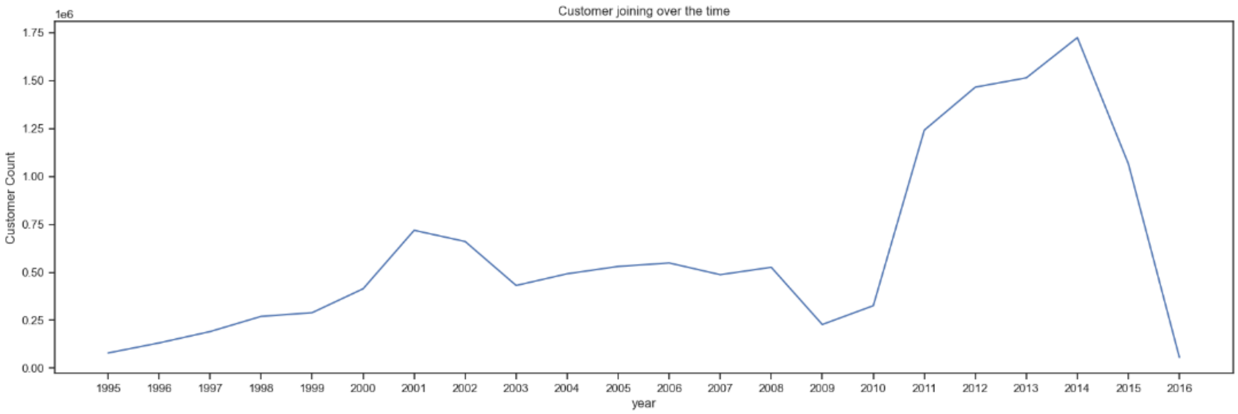
1.The average age of customers who joined the bank was high in early years, and dropped overtime. It dropped strongly in 2011, and started to rise again in 2015.

- Correlation between Customer Age and Product Sum:

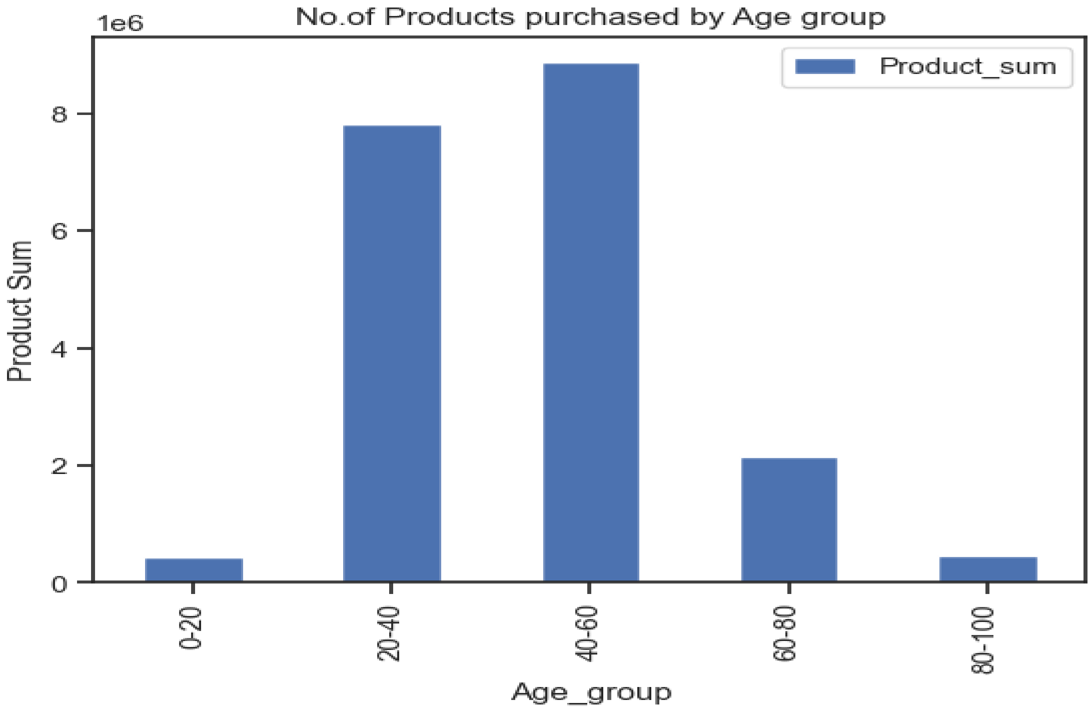


	Age	Product_sum
Age	1.000000	0.182969
Product_sum	0.182969	1.000000

1. The correlation coefficient between Age and Product sum is 0.182969, there is a negligible positive relationship between them.

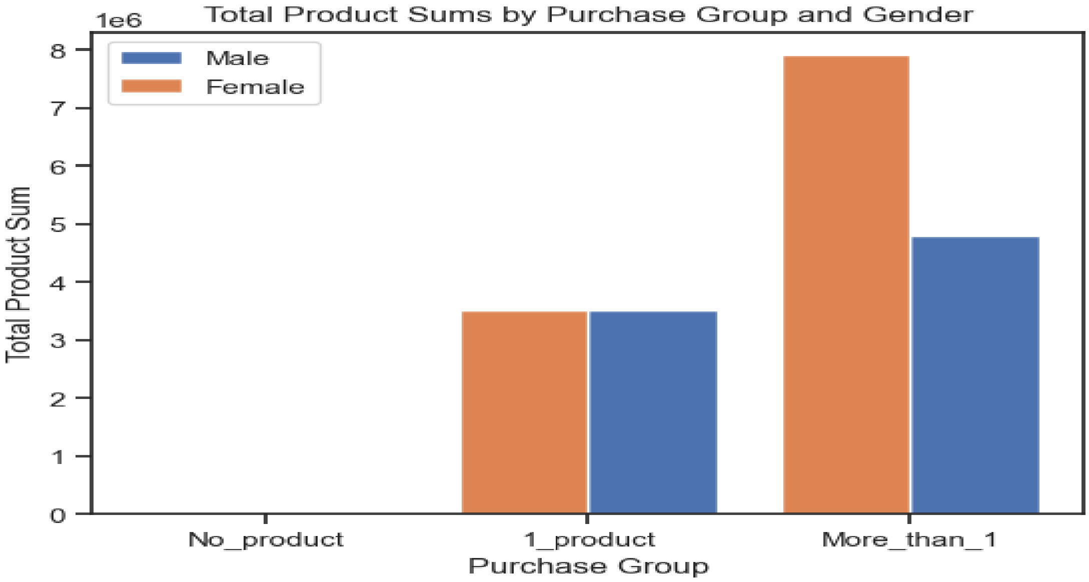


- Products purchased by Age group:



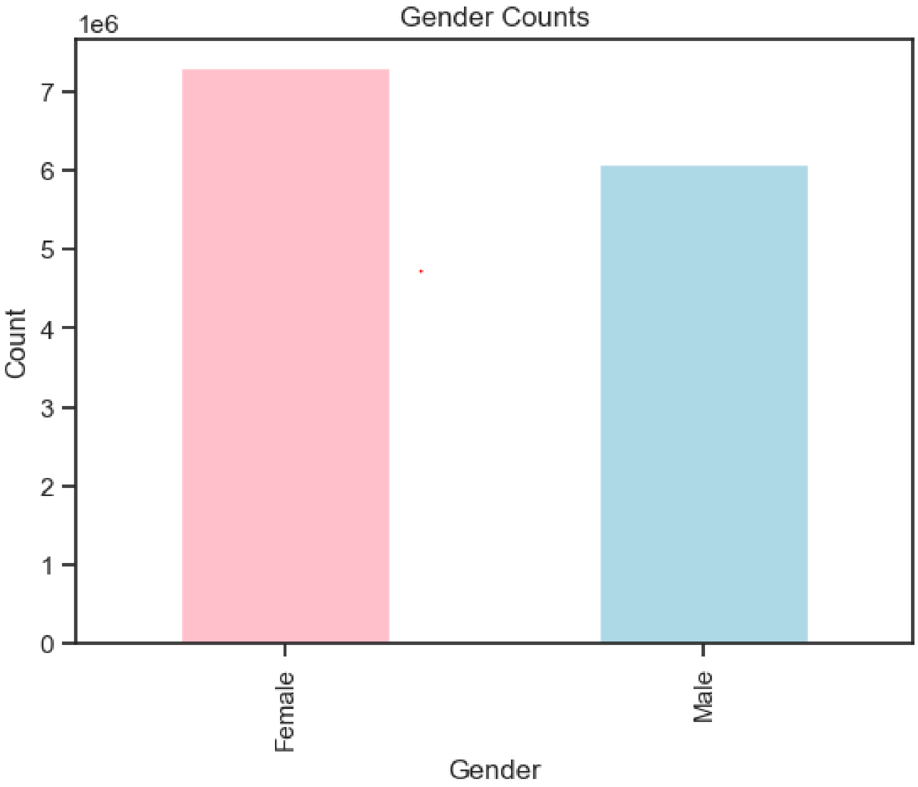
1. Age group of 40-60 has purchased more products than the other groups

- Total product sums by Purchase Group and Gender:



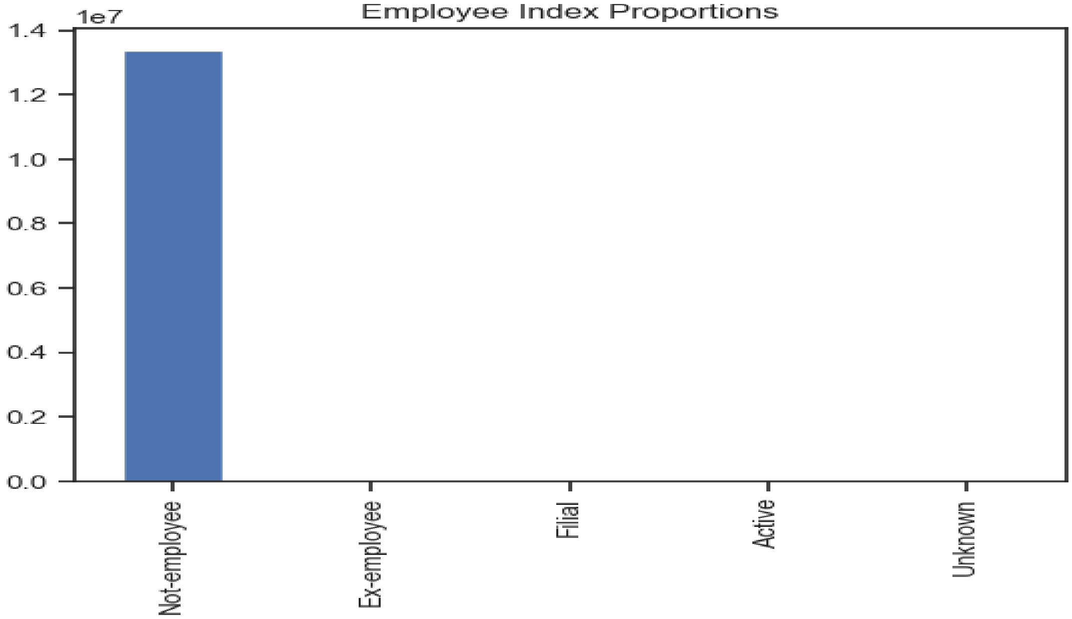
1. The bar chart above shows that both the same number male and female customers purchased only 1 product.
2. It also shows that female customers purchased more than 1 product more than male customers.

- Gender counts:



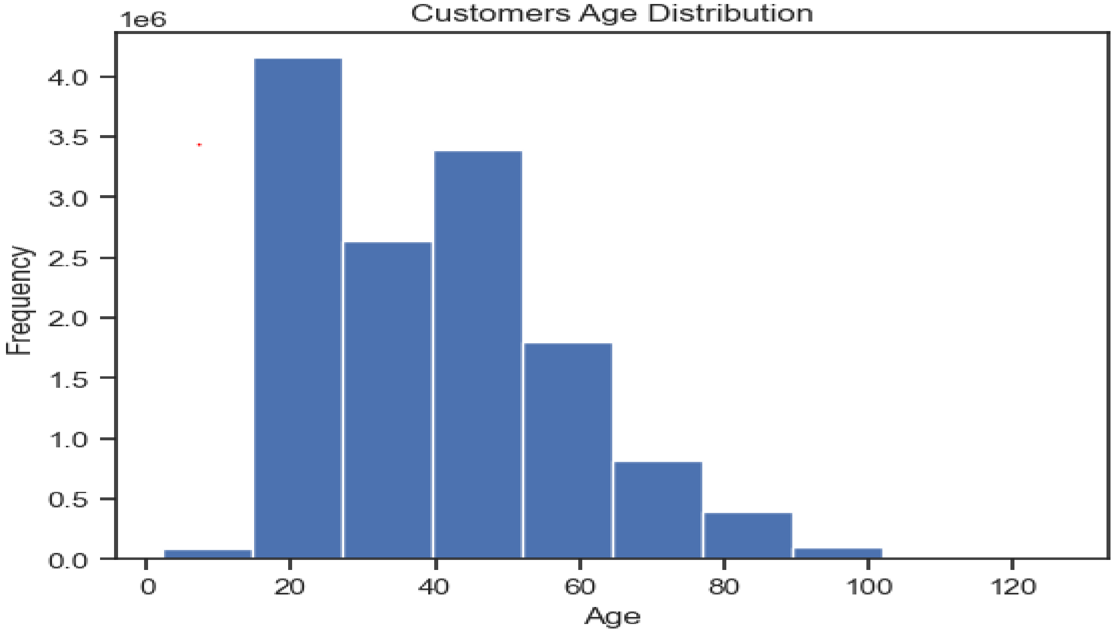
- The bar chart above shows that there are more female customers in the company than male customer

- Employee_index proportions:



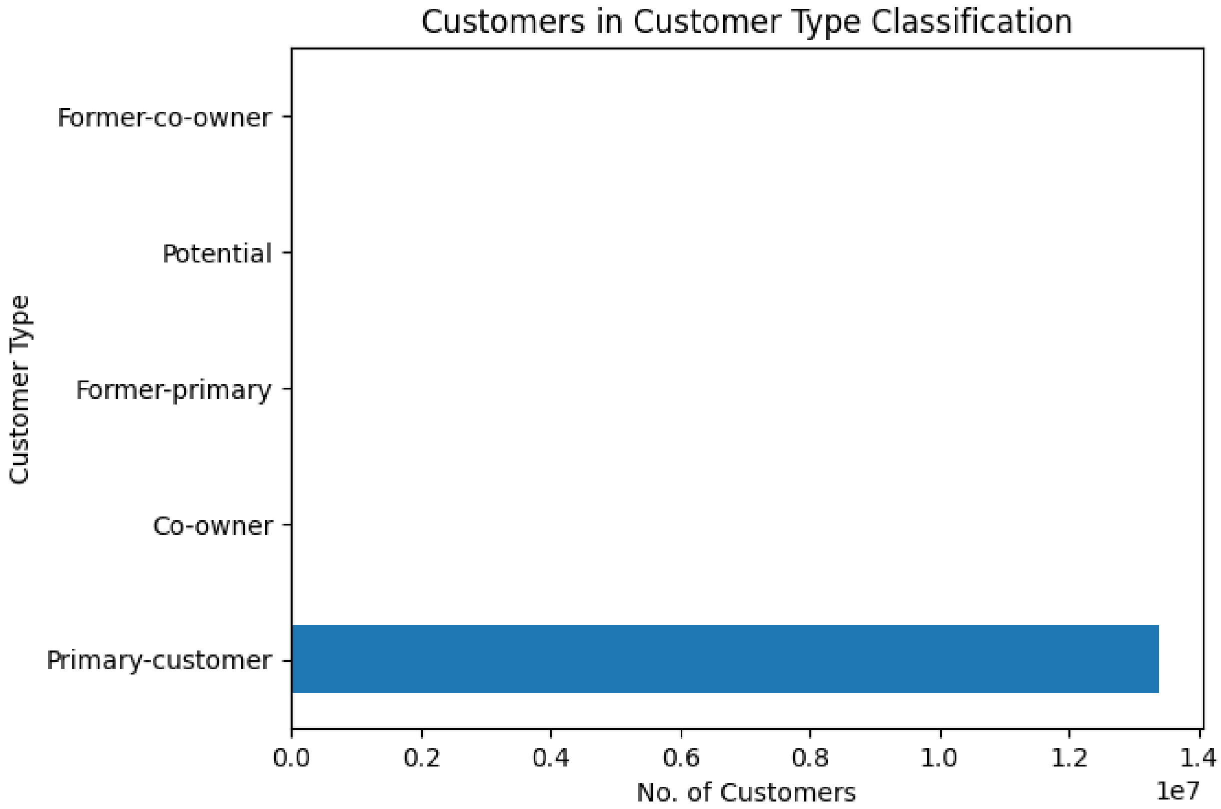
1. The bar chart above shows that majority of customers are not employees

- Customers Age distribution:

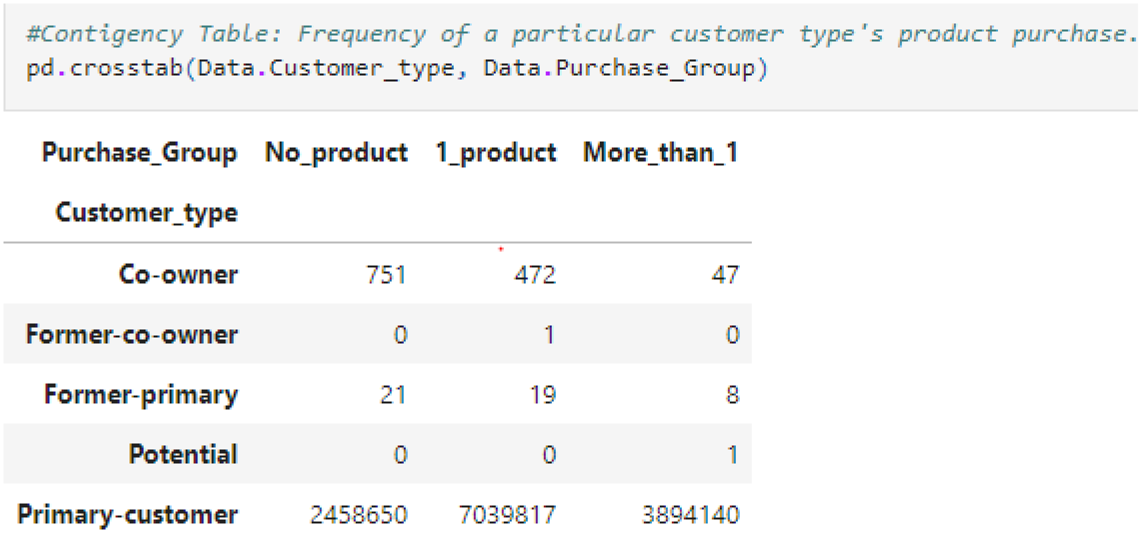


1. The bar chart above shows that majority of customers are in their 20's, this is followed customers in their 50's

- Customer_type Classification:



1. From the graph above, we can see that majority of customers are classified as primary customers.

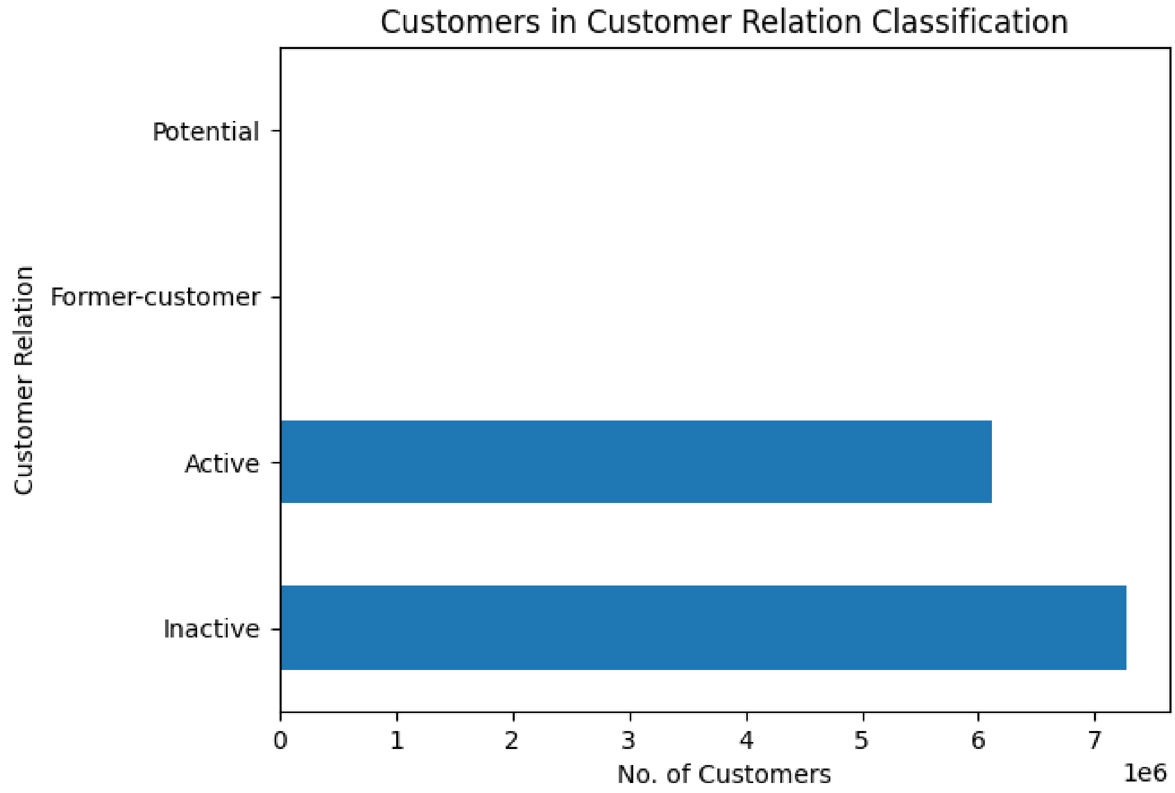


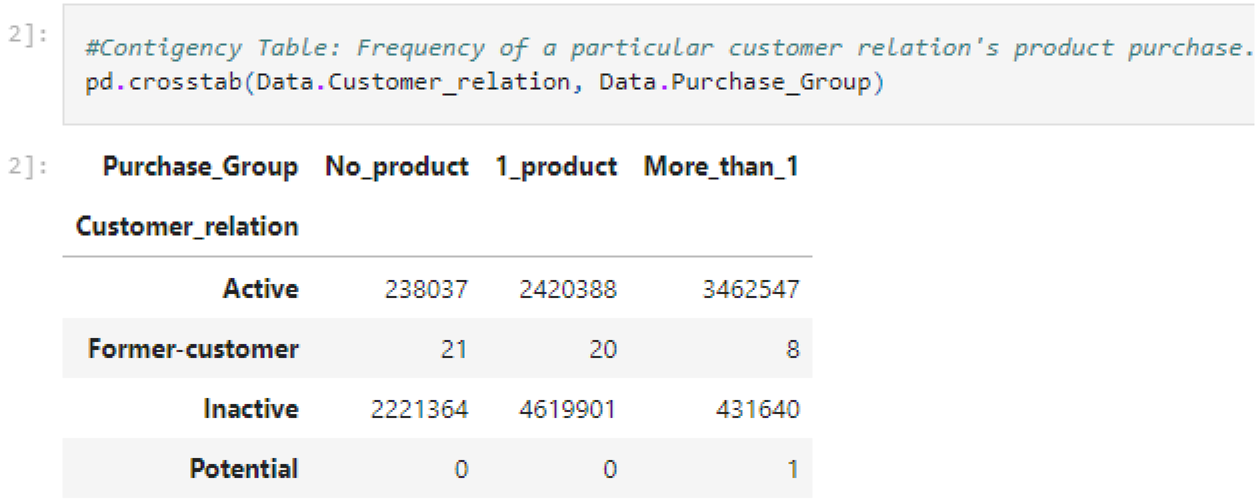
```
#Contigency Table: Frequency of a particular customer type's product purchase.
pd.crosstab(Data.Customer_type, Data.Purchase_Group)
```

Purchase_Group	No_product	1_product	More_than_1
Customer_type			
Co-owner	751	472	47
Former-co-owner	0	1	0
Former-primary	21	19	8
Potential	0	0	1
Primary-customer	2458650	7039817	3894140

1. The contingency table above shows that majority of customers in all the customer type classification bought only product. A good number of Customers in co-owner ond former-primary customer type group did not buy any product. We cannot say that there is any customer type that is more likely to buy multiple products.

- Customer relation classification:





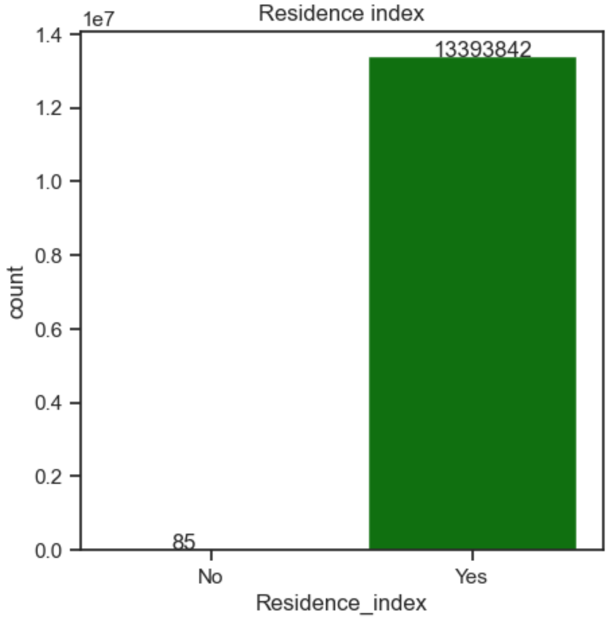
```
2]:  #Contigency Table: Frequency of a particular customer relation's product purchase.
     pd.crosstab(Data.Customer_relation, Data.Purchase_Group)
```

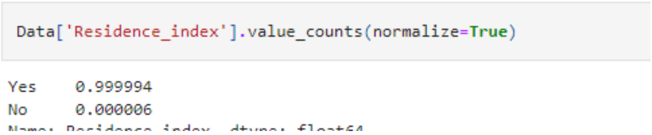
2]:

Purchase_Group Customer_relation	No_product	1_product	More_than_1
Active	238037	2420388	3462547
Former-customer	21	20	8
Inactive	2221364	4619901	431640
Potential	0	0	1

1. The contingency table above shows that majority of customers who have an active relation with the bank purchased more than 1 product. Whereas majority of customers who have an inactive relation with the bank purchased only 1 product, or no product.
2. This may imply that customers with active relation with the bank are more likely to purchase multiple products.

- Residence_index:



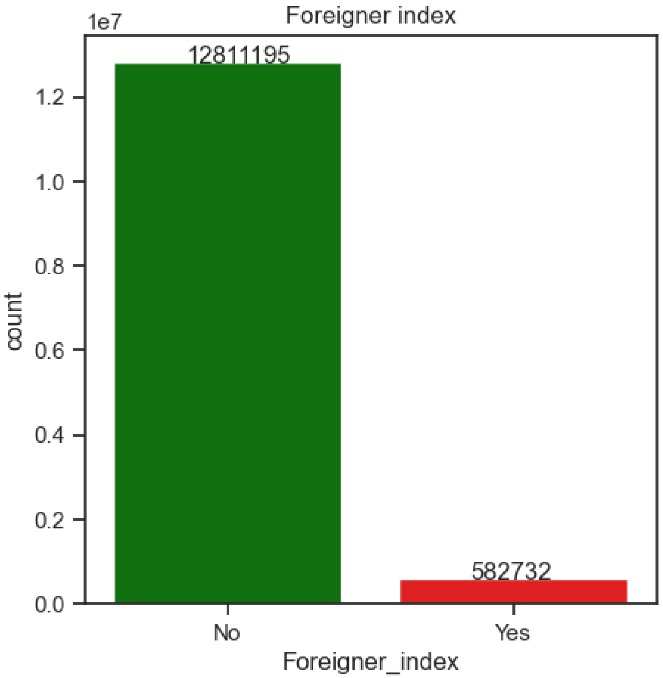


```
Data['Residence_index'].value_counts(normalize=True)
```

```
Yes    0.999994
No     0.000006
Name: Residence_index, dtype: float64
```

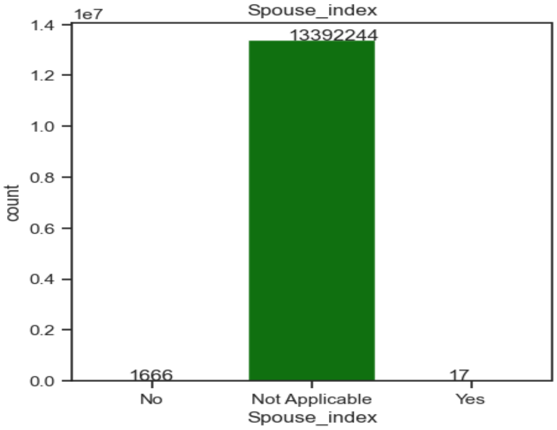
1. 99% of the customers reside in the country same as the XYZ credit union bank.

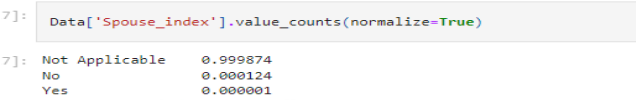
- Foreigner_index:



1. 95% of the customers are non-foreigners while the 4% population are foreign national

- Spouse_index:

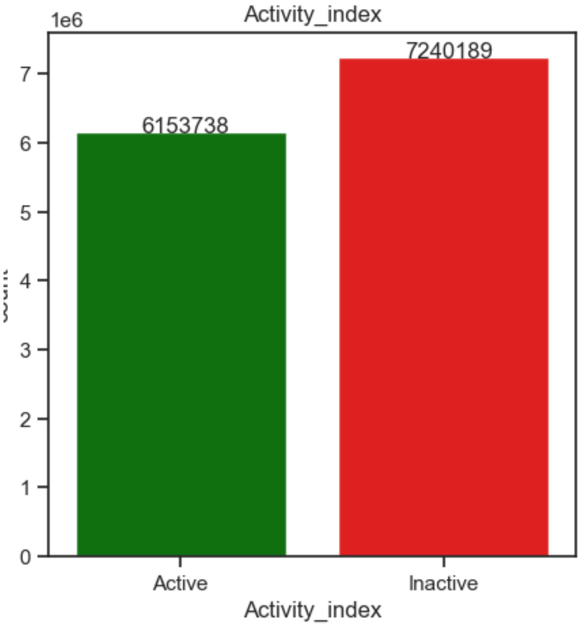




```
7]:  Data['Spouse_index'].value_counts(normalize=True)

7]:  Not Applicable    0.999874
     No                0.000124
     Yes               0.000001
```

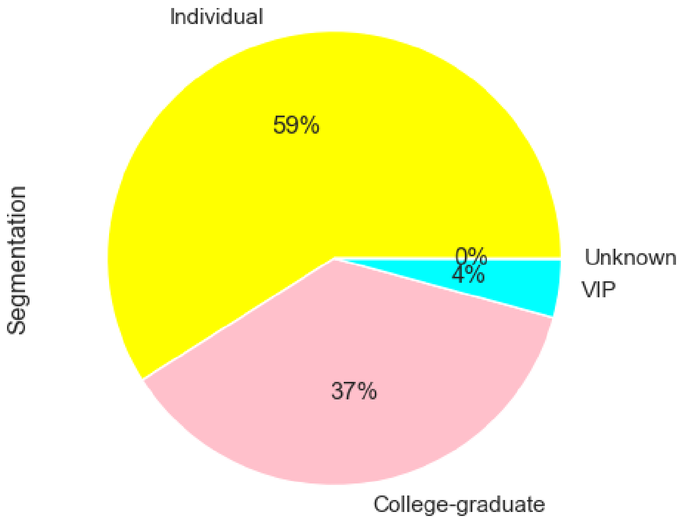
1. 99.9% of the data contains null or no data hence this is inconclusive and unclear to interpret.

- Activity_index:

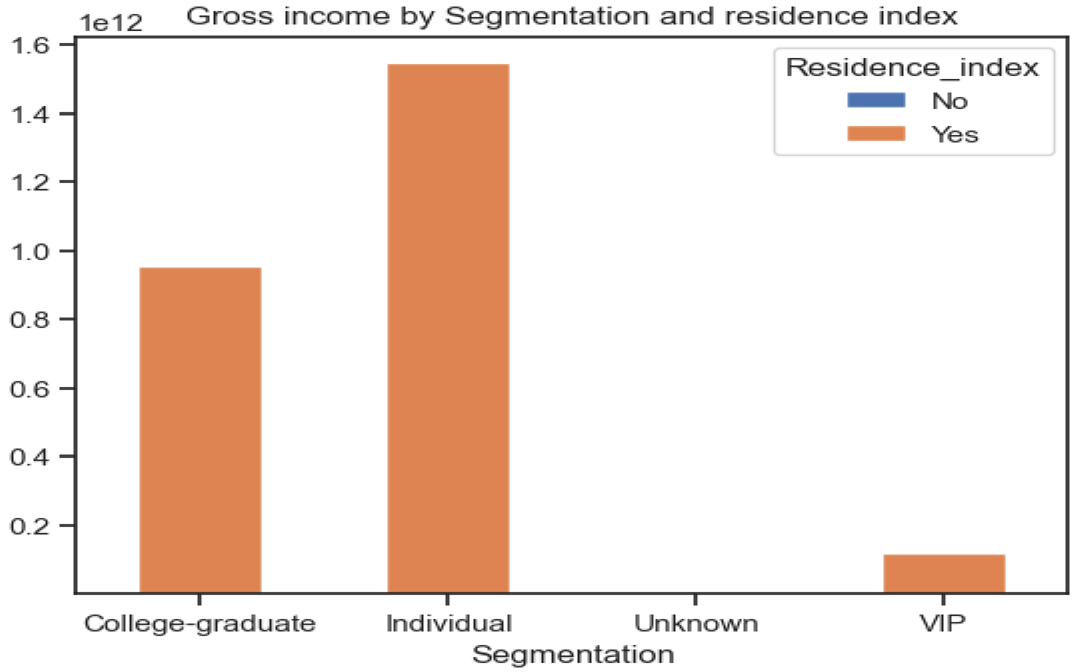


1. 45.9% are active account holders in the bank

- Segmentation:

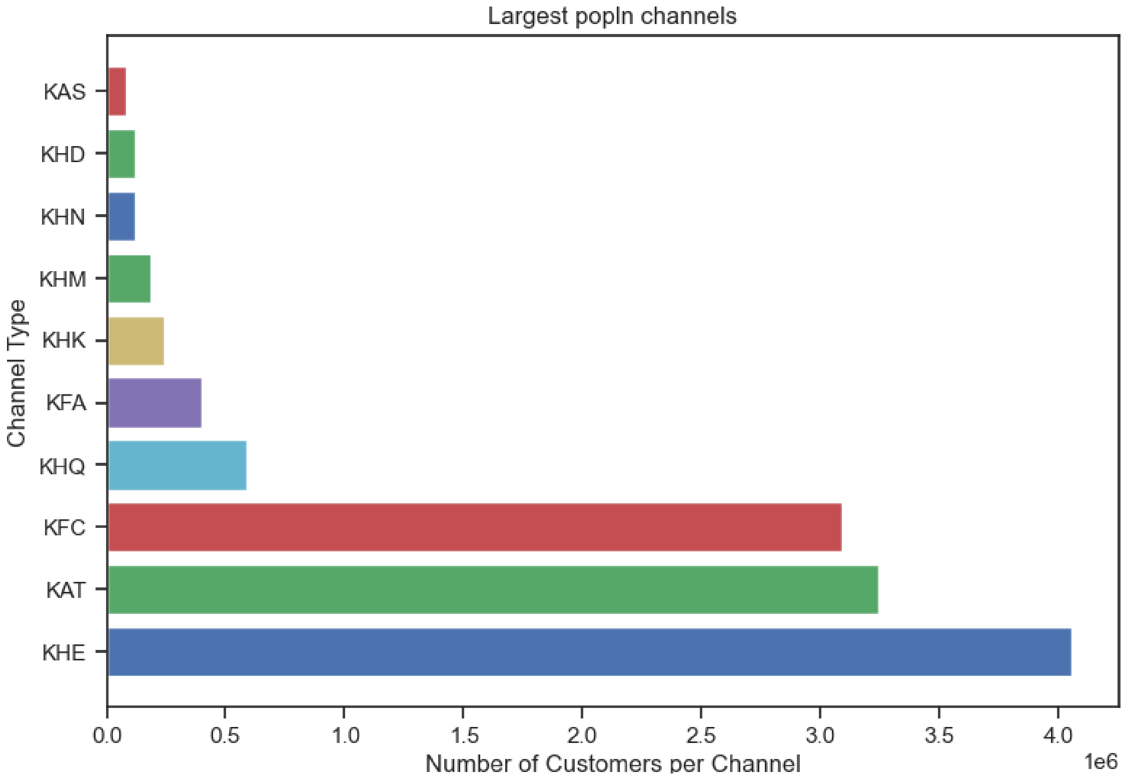


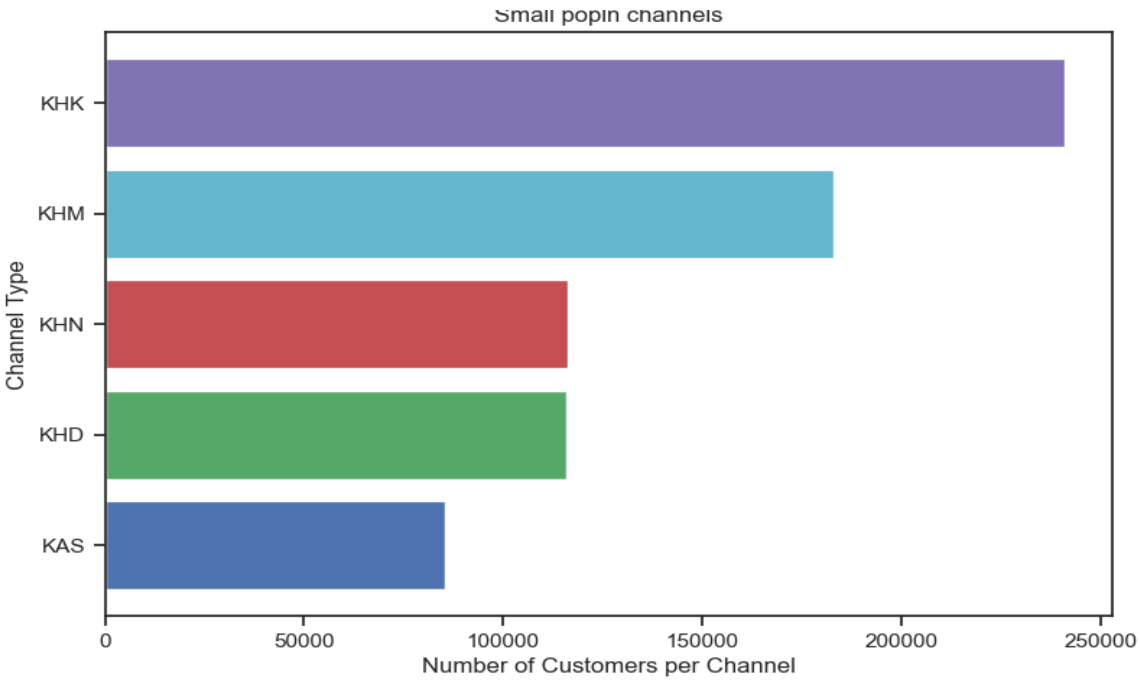
- Gross income by Segmentation and residence index:



1. Individuals form the highest income earners and belong to the same country as the bank

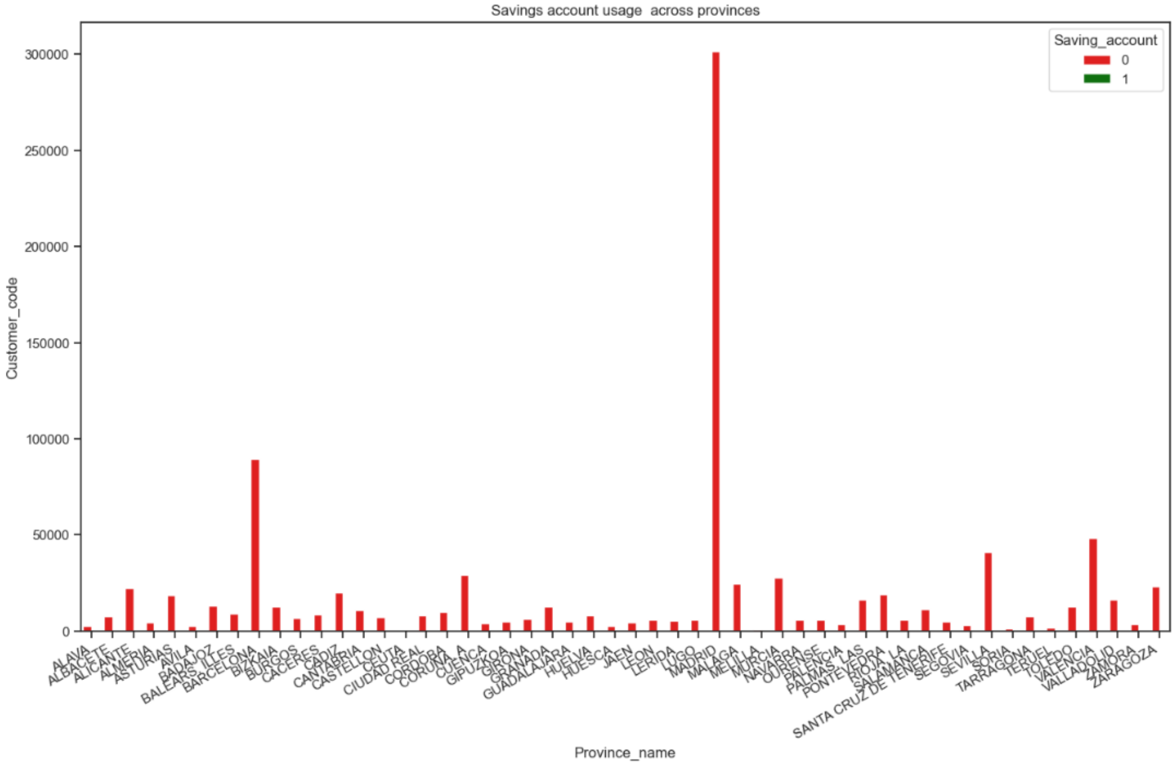
- Customers across Channels:

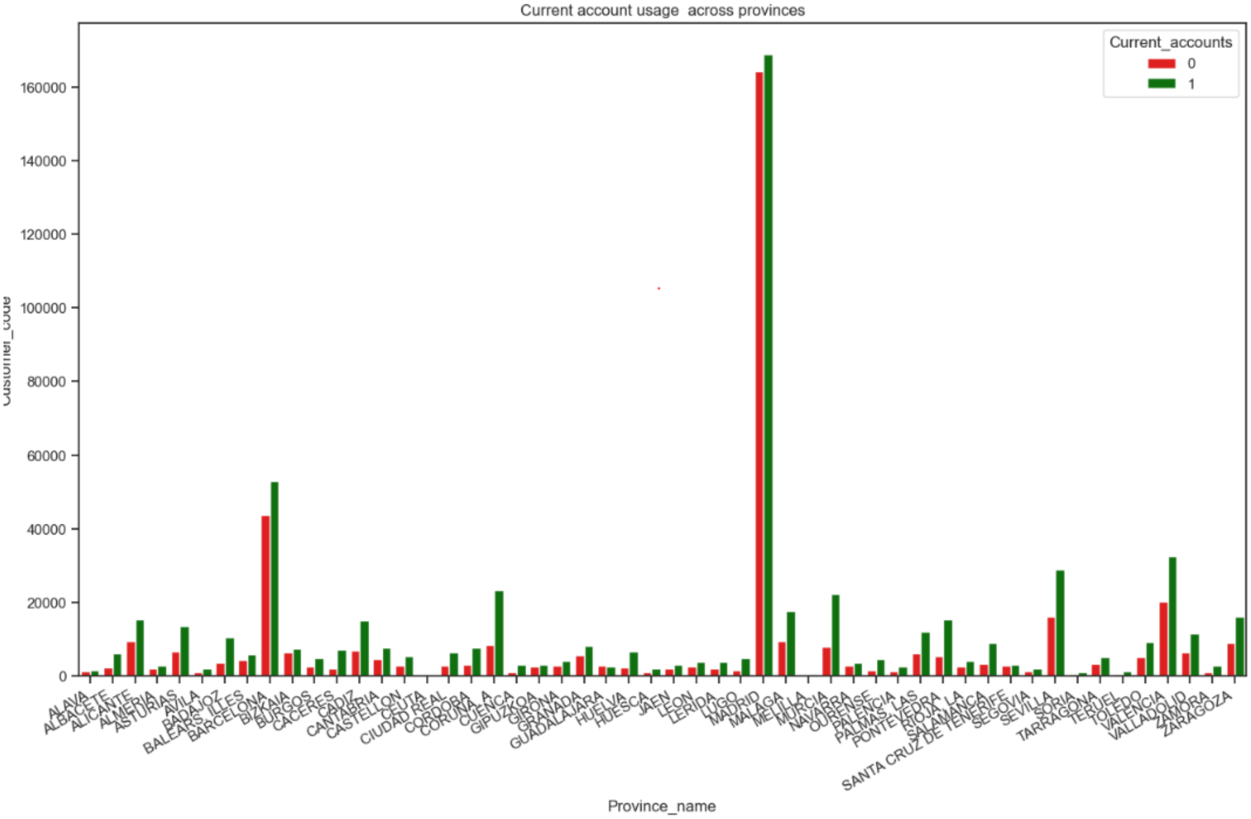




1. Th higher customer outreach are from Channel KHE while the lowest is from Channel KAS

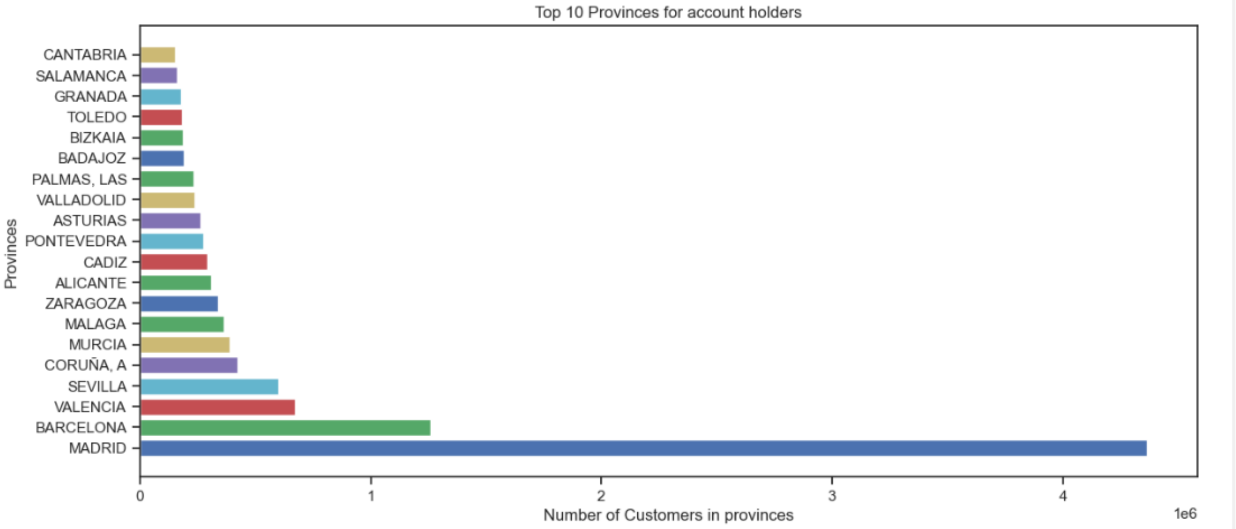
- Products Savings account and Current account across provinces:

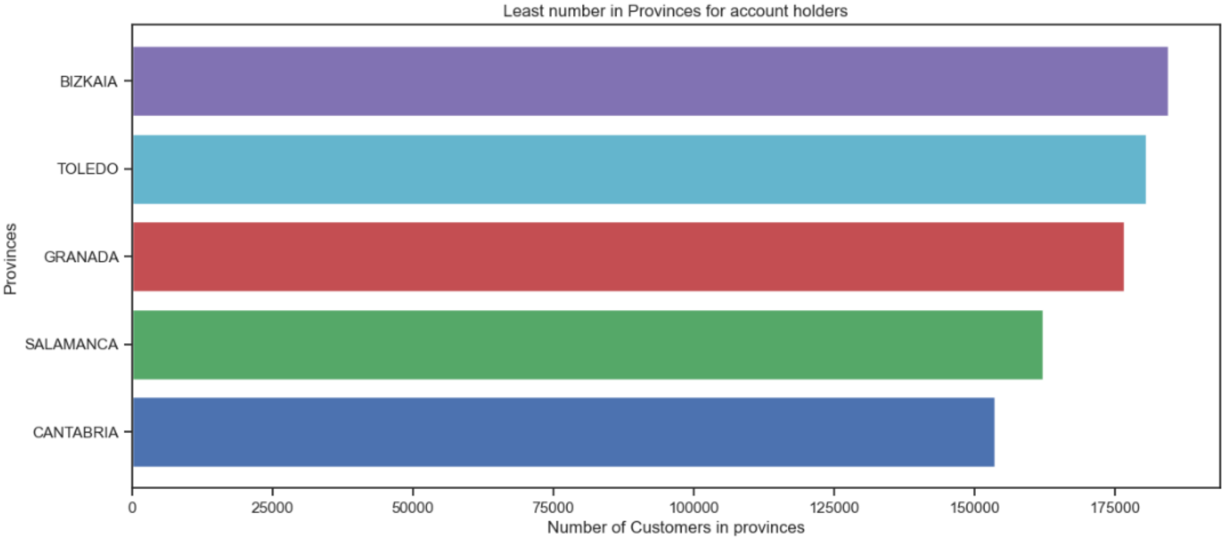




1. Provice Madrid tops in non saving account holders.
2. Madrid has highest current account holders followed by Barcelona and Valladolid

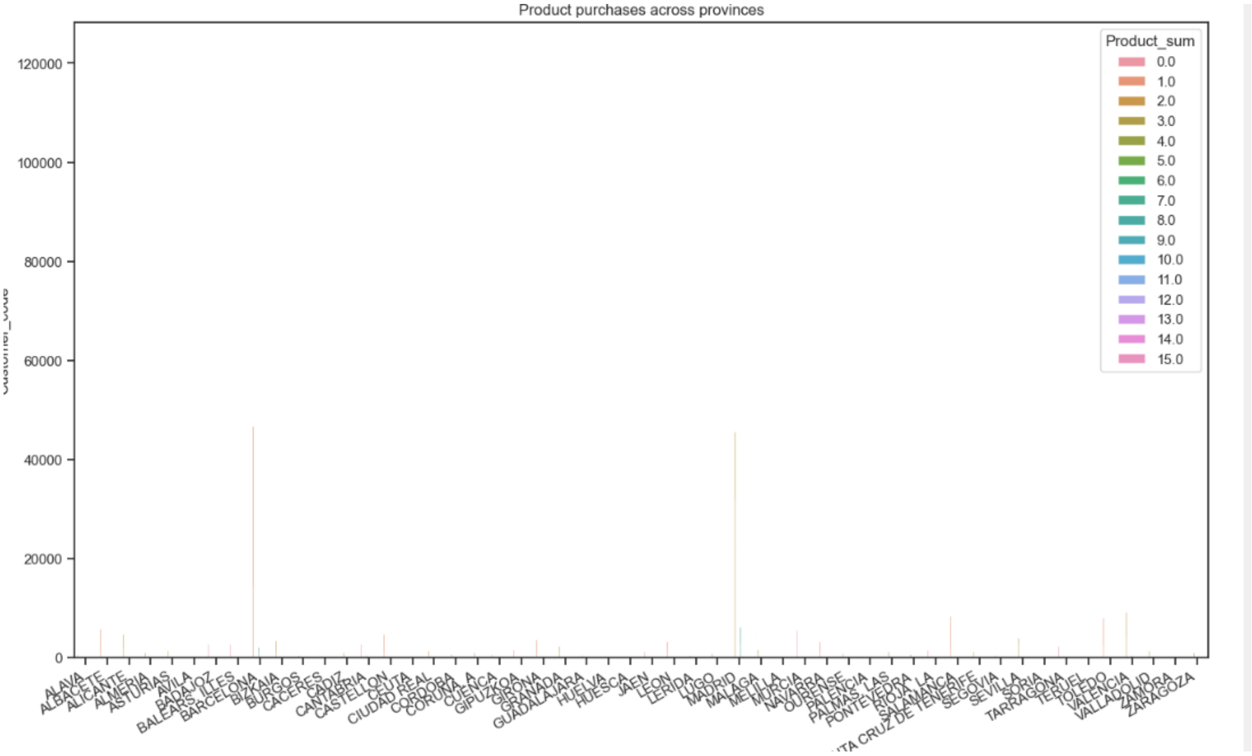
- Customers across Provinces:



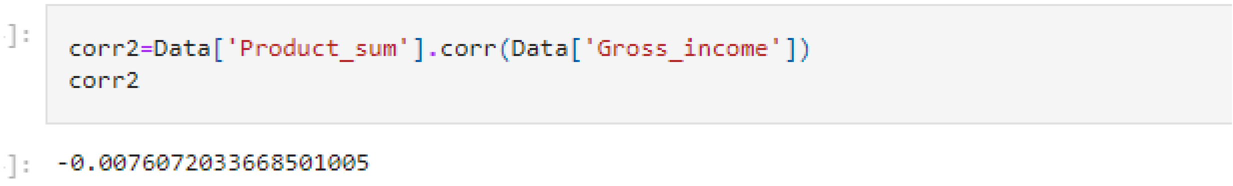


1. The above depictions show the population of customers across different provinces in Latin America

- Product purchase across provinces:



13. Correlation between Gross_income and number of product purchases:

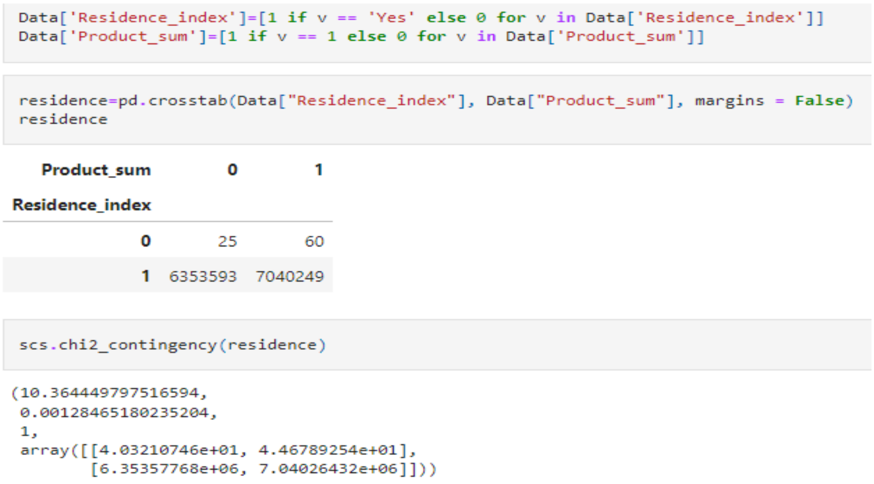


```
corr2=Data['Product_sum'].corr(Data['Gross_income'])
corr2
```

```
-0.0076072033668501005
```

Gross income depicts a negative correlation with the number of products customers hold

- Contingency table:



```
Data['Residence_index']=[1 if v == 'Yes' else 0 for v in Data['Residence_index']]
Data['Product_sum']=[1 if v == 1 else 0 for v in Data['Product_sum']]
```

```
residence=pd.crosstab(Data["Residence_index"], Data["Product_sum"], margins = False)
residence
```

Product_sum	**0**	**1**
Residence_index		
0	25	60
1	6353593	7040249

```
scs.chi2_contingency(residence)
```

```
(10.364449797516594,
 0.00128465180235204,
 1,
 array([[4.03210746e+01, 4.46789254e+01],
        [6.35357768e+06, 7.04026432e+06]]))
```

Hypothesis Testing:

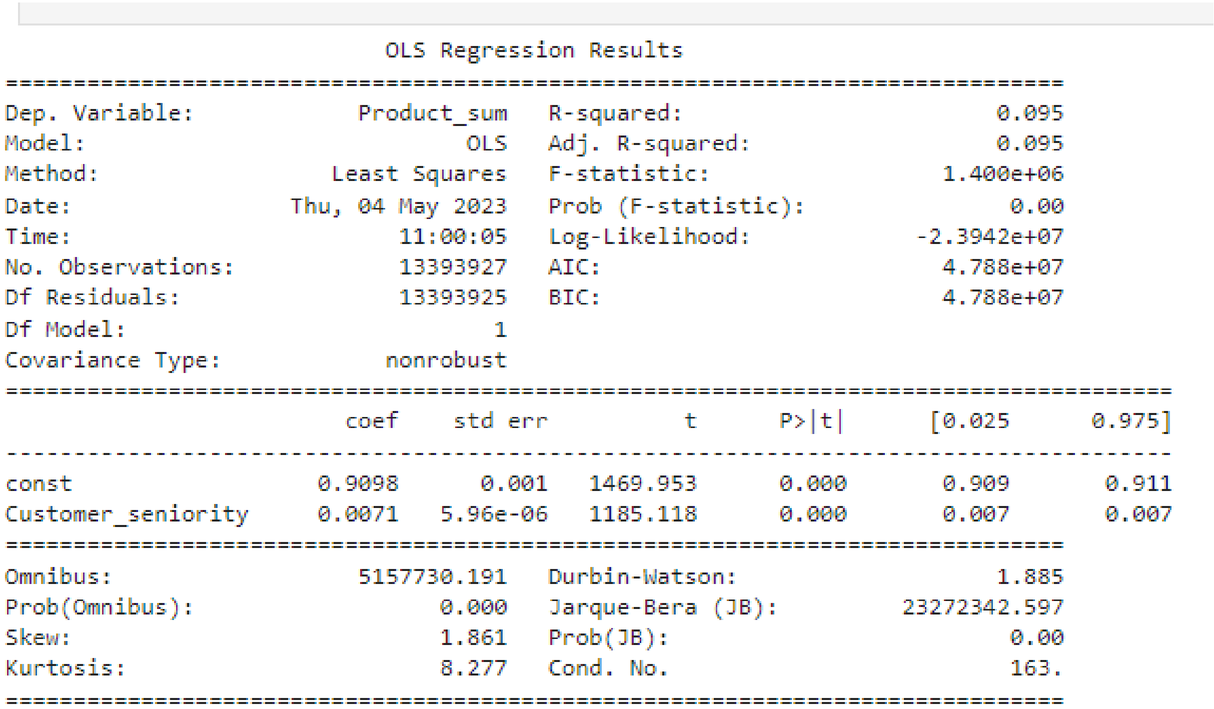
The output of the hypothesis test is generally a statistical value called the p-value. This value is the probability of occurrence of the null hypothesis. There is a threshold preset (generally 5% chance of the null hypothesis been true and the alternative hypothesis is false, or 95% chance of the alternative hypothesis been true and the null hypothesis false.) where if the pvalue is less than 0.05 or 5% we reject the null hypothesis and if more than 0.05 we fail to reject the null hypothesis.

The following hypotheses will be tested using various statistical methods.

Hypothesis 1: Does customers' purchases depend on how long they have been with the Bank?

H0: Product purchase does not depend on Customer seniority.

H1: Product purchase depends on customer seniority.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Product_sum   R-squared:                       0.095
Model:                            OLS   Adj. R-squared:                  0.095
Method:                 Least Squares   F-statistic:                 1.400e+06
Date:                Thu, 04 May 2023   Prob (F-statistic):               0.00
Time:                        11:00:05   Log-Likelihood:            -2.3942e+07
No. Observations:            13393927   AIC:                         4.788e+07
Df Residuals:                13393925   BIC:                         4.788e+07
Df Model:                           1
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                  0.9098      0.001   1469.953      0.000       0.909       0.911
Customer_seniority     0.0071   5.96e-06   1185.118      0.000       0.007       0.007
==============================================================================
Omnibus:                  5157730.191   Durbin-Watson:                   1.885
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         23272342.597
Skew:                           1.861   Prob(JB):                         0.00
Kurtosis:                       8.277   Cond. No.                         163.
==============================================================================
```

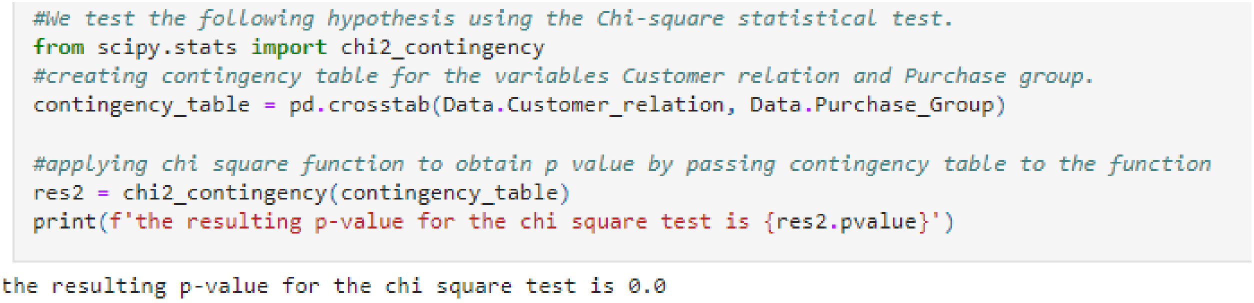
INFERENCE

- From the summary result above, the fitted regression model is: Product sum = 0.9098 + 0.0071(Customer seniority).
- This means that an increase in a customer's seniority is associated with an average increase in product purchase by 0.0071 points.
- The intercept value of 0.9098 tells us the average expected product purchase for customers.
- The pvalue of 0.000 is lesser than our significance treshhold of 0.05, thus we cannot say that there is no statistically significant association between product sum and customer seniority.
- The R-squared value tells us the percentage of variation in product sum that can be explained by customer seniority.

- The R-Squared value of 0.095 tells us that only approximately 9.5% of variations in product sum can be explained by changes in customer seniority.
- The number of residuals from the result above is more than zero implying that the model may be over/under predicting.
- Thus we fail to accept the null hypothesis which states that product sum depends on customer seniority.

Hypothesis 2: Is customer relations associated with amount of product purchased?

- H0: There is no significant association between the 2 variables - Customer relation and Purchase group.
- H1: There is a significant association between the 2 variables - Customer relation and Purchase group.



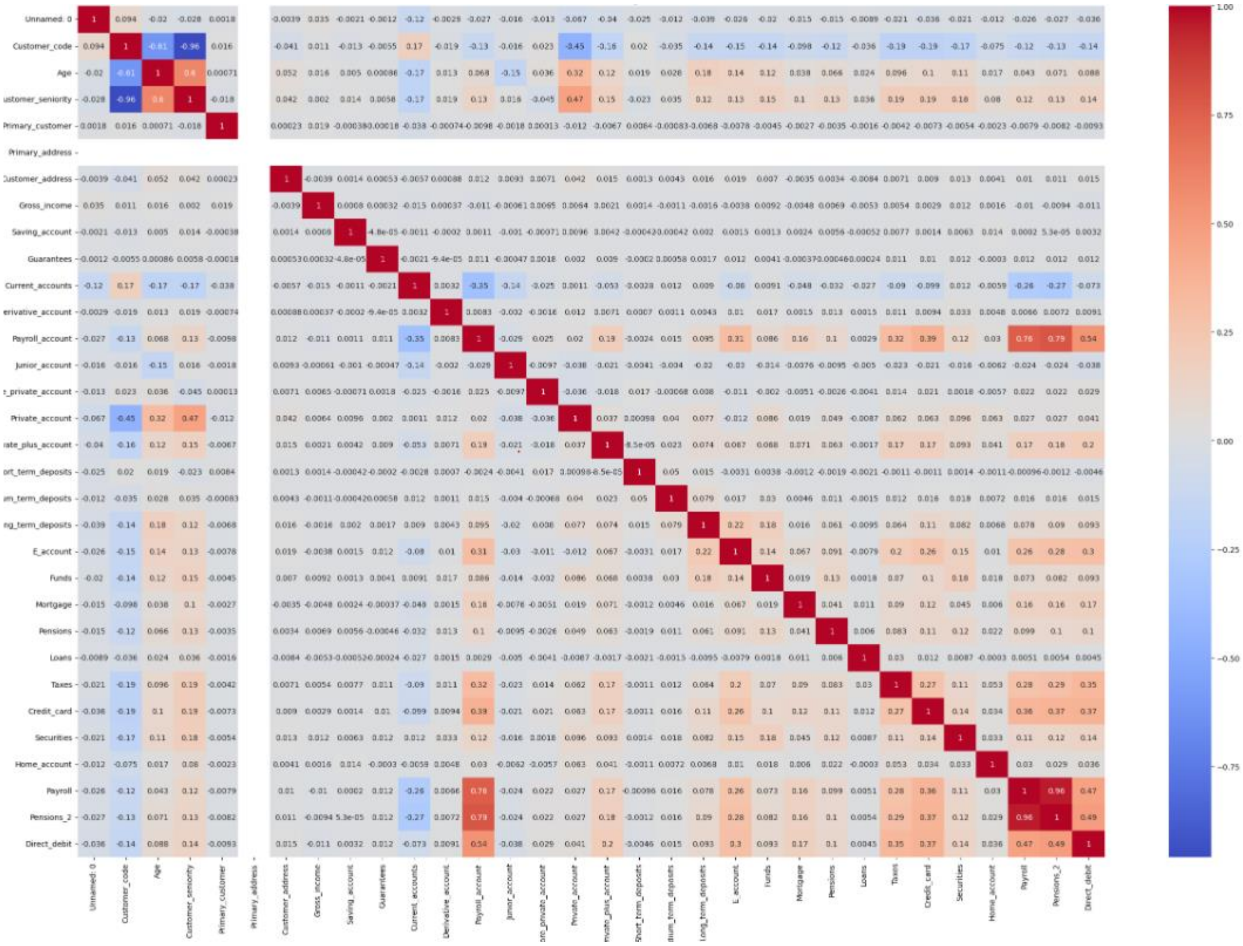
```
#We test the following hypothesis using the Chi-square statistical test.
from scipy.stats import chi2_contingency
#creating contingency table for the variables Customer relation and Purchase group.
contingency_table = pd.crosstab(Data.Customer_relation, Data.Purchase_Group)

#applying chi square function to obtain p value by passing contingency table to the function
res2 = chi2_contingency(contingency_table)
print(f'the resulting p-value for the chi square test is {res2.pvalue}')
```

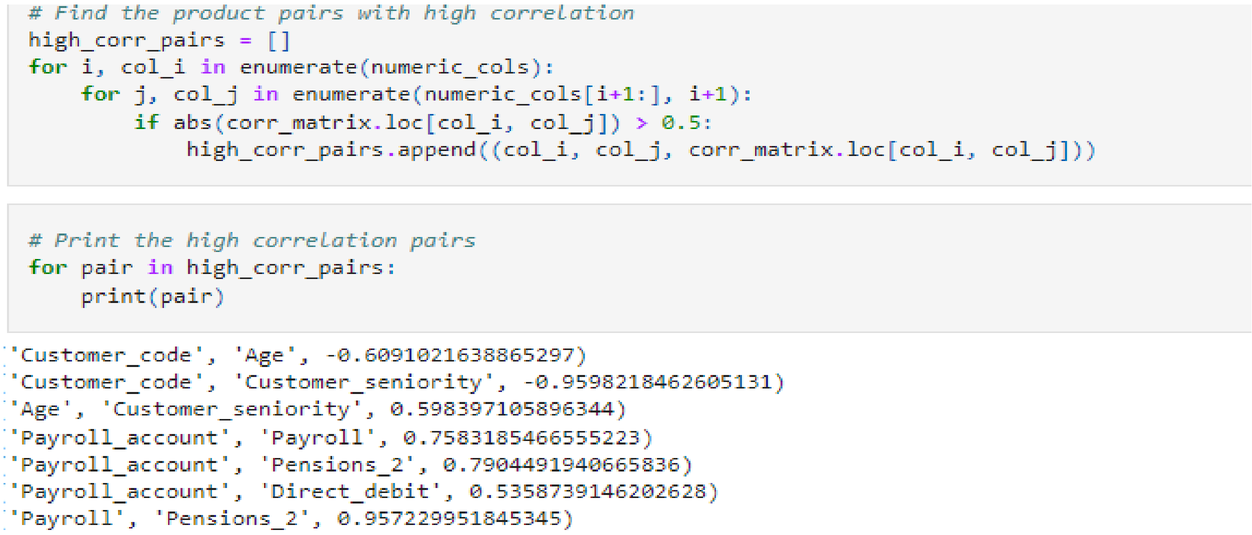
```
the resulting p-value for the chi square test is 0.0
```

INFERENCE

- The result provides a p value of 0.0 which is lesser than the significance threshold of 0.05.
- Using this we can conclude for our hypothesis test that we fail to accept the null hypothesis.
- Hence we can assume our alternative hypothesis to be true which says — There is a significant association between the 2 variables — Customer_relation and Purchase_Group.
- Thus we can say that purchase of products by customers is associated with the customer's relation with the bank.

Hypothesis 3: Are there any products that are commonly purchased together?





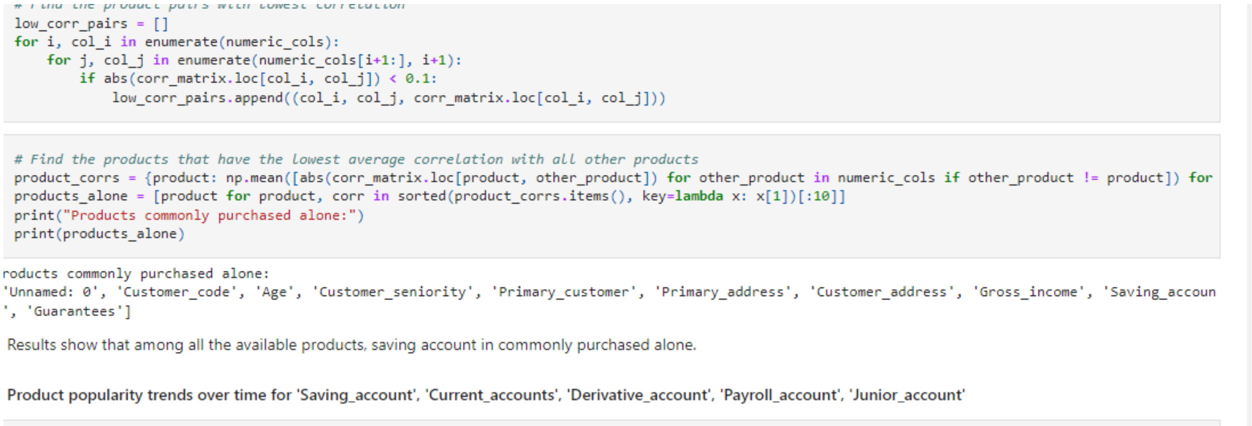
```
# Find the product pairs with high correlation
high_corr_pairs = []
for i, col_i in enumerate(numeric_cols):
    for j, col_j in enumerate(numeric_cols[i+1:], i+1):
        if abs(corr_matrix.loc[col_i, col_j]) > 0.5:
            high_corr_pairs.append((col_i, col_j, corr_matrix.loc[col_i, col_j]))
```

```
# Print the high correlation pairs
for pair in high_corr_pairs:
    print(pair)
```

```
'Customer_code', 'Age', -0.6091021638865297)
'Customer_code', 'Customer_seniority', -0.9598218462605131)
'Age', 'Customer_seniority', 0.598397105896344)
'Payroll_account', 'Payroll', 0.7583185466555223)
'Payroll_account', 'Pensions_2', 0.7904491940665836)
'Payroll_account', 'Direct_debit', 0.5358739146202628)
'Payroll', 'Pensions_2', 0.957229951845345)
```

The results show that Payroll account and Direct_debit are commonly purchased together by customers.

Hypothesis 4: Are there any products that are commonly purchased alone?



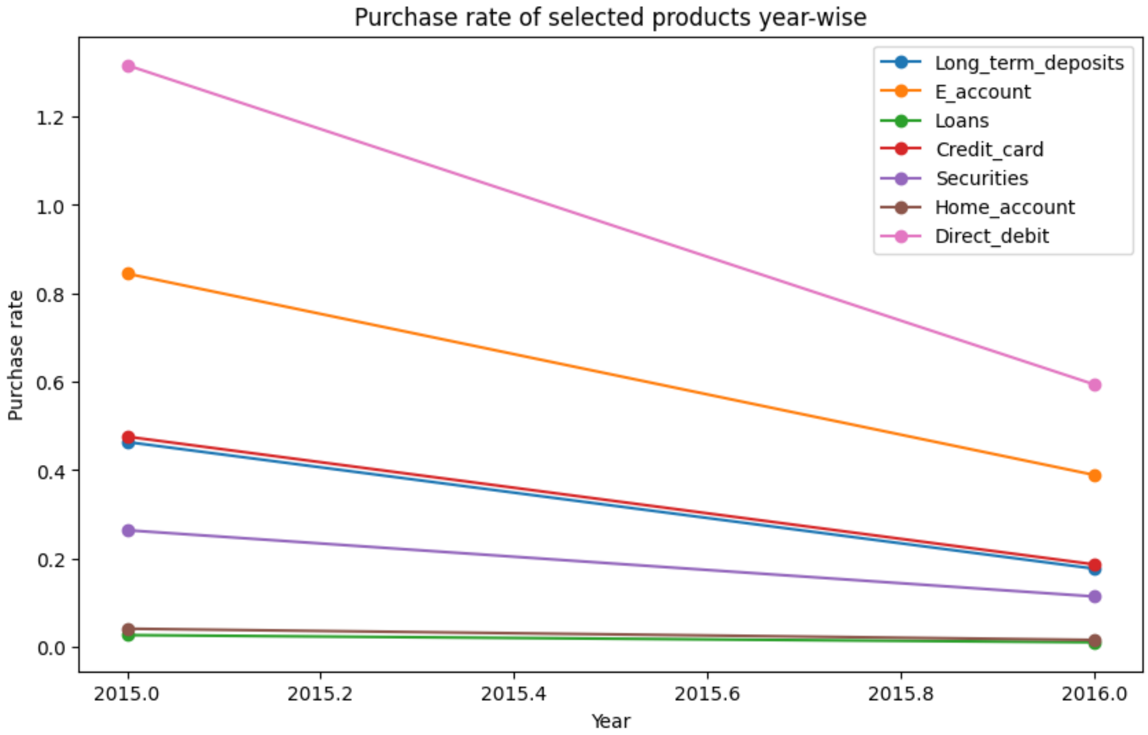
```
# Find the product pairs with lowest correlation
low_corr_pairs = []
for i, col_i in enumerate(numeric_cols):
    for j, col_j in enumerate(numeric_cols[i+1:], i+1):
        if abs(corr_matrix.loc[col_i, col_j]) < 0.1:
            low_corr_pairs.append((col_i, col_j, corr_matrix.loc[col_i, col_j]))
```

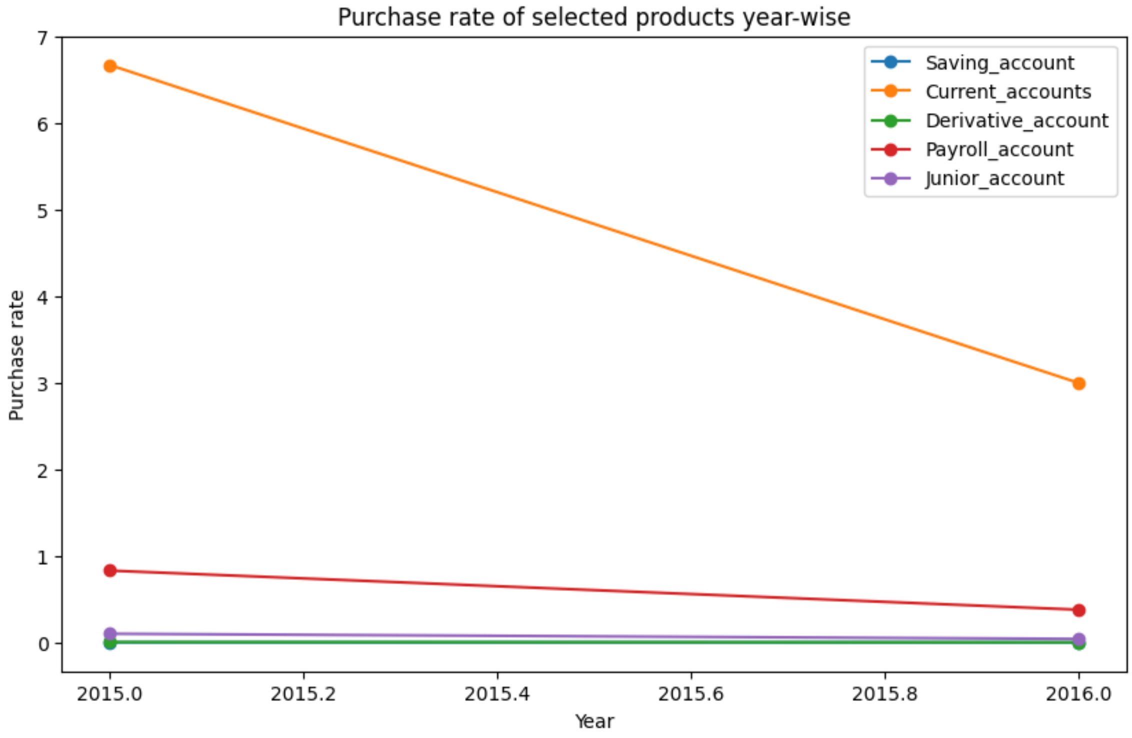
```
# Find the products that have the lowest average correlation with all other products
product_corrs = {product: np.mean([abs(corr_matrix.loc[product, other_product]) for other_product in numeric_cols if other_product != product]) for
products_alone = [product for product, corr in sorted(product_corrs.items(), key=lambda x: x[1])[:10]]
print("Products commonly purchased alone:")
print(products_alone)
```

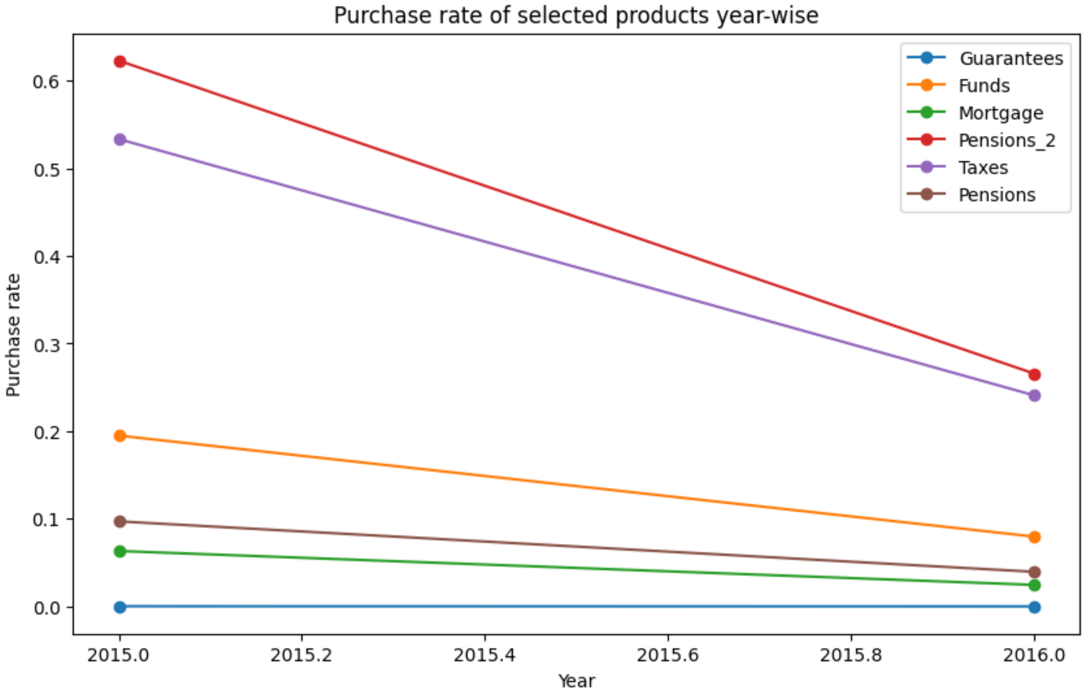
```
roducts commonly purchased alone:
'Unnamed: 0', 'Customer_code', 'Age', 'Customer_seniority', 'Primary_customer', 'Primary_address', 'Customer_address', 'Gross_income', 'Saving_accoun
', 'Guarantees']
```

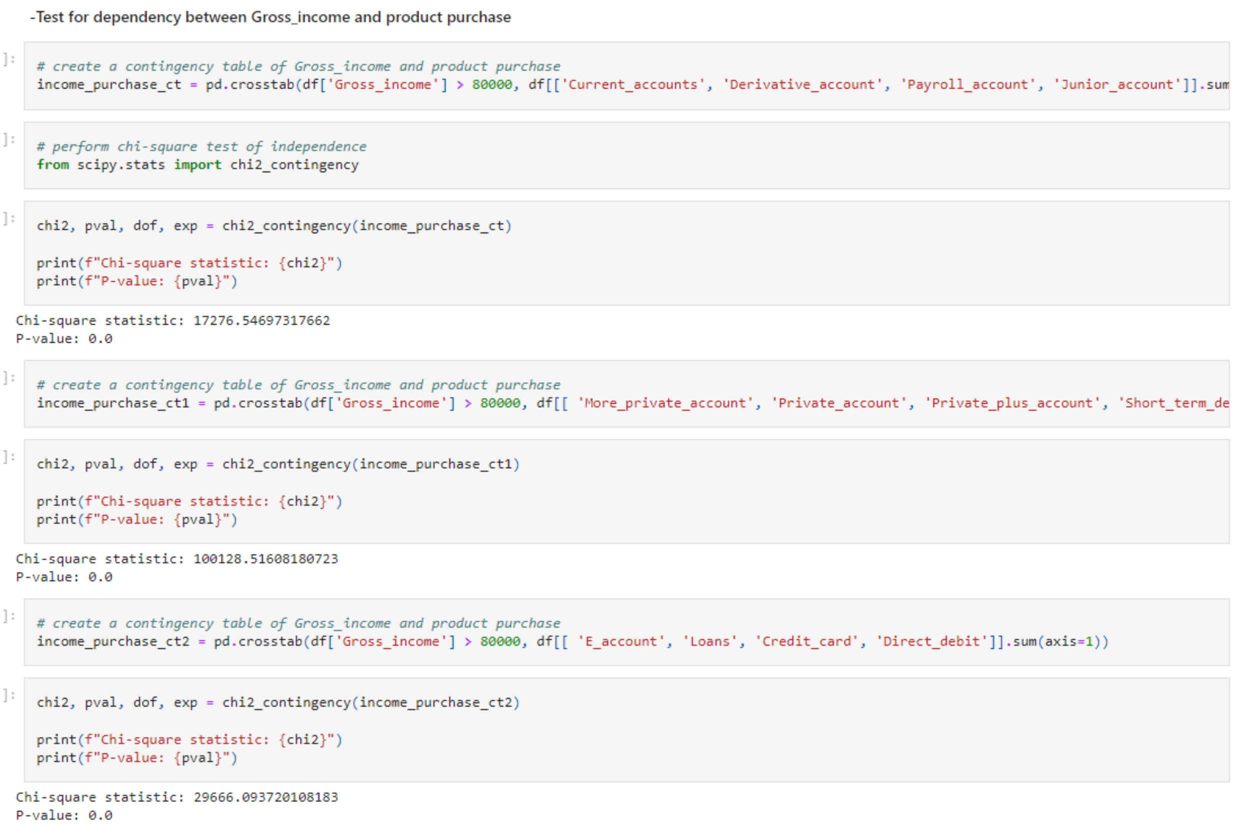
Results show that among all the available products, saving account in commonly purchased alone.

Product popularity trends over time for 'Saving_account', 'Current_accounts', 'Derivative_account', 'Payroll_account', 'Junior_account'




Purchase rate of selected products year-wise
Saving_account
Current_accounts
Derivative_account
Payroll_account
Junior_account
Purchase rate
0
1
2
3
4
5
6
7
2015.0
2015.2
2015.4
2015.6
2015.8
2016.0
Year

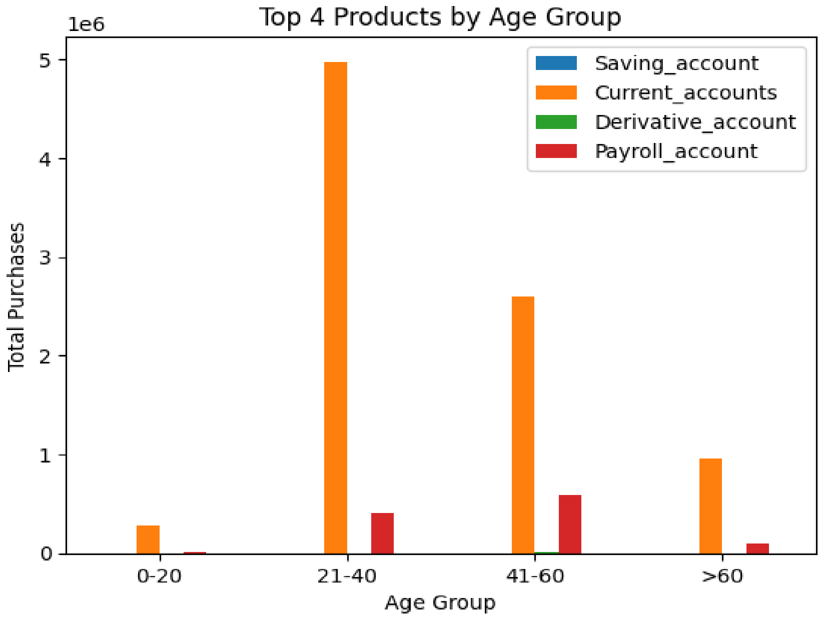

Purchase rate of selected products year-wise
Guarantees
Funds
Mortgage
Pensions_2
Taxes
Pensions
Purchase rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
2015.0
2015.2
2015.4
2015.6
2015.8
2016.0
Year



-Test for dependency between Gross_income and product purchase

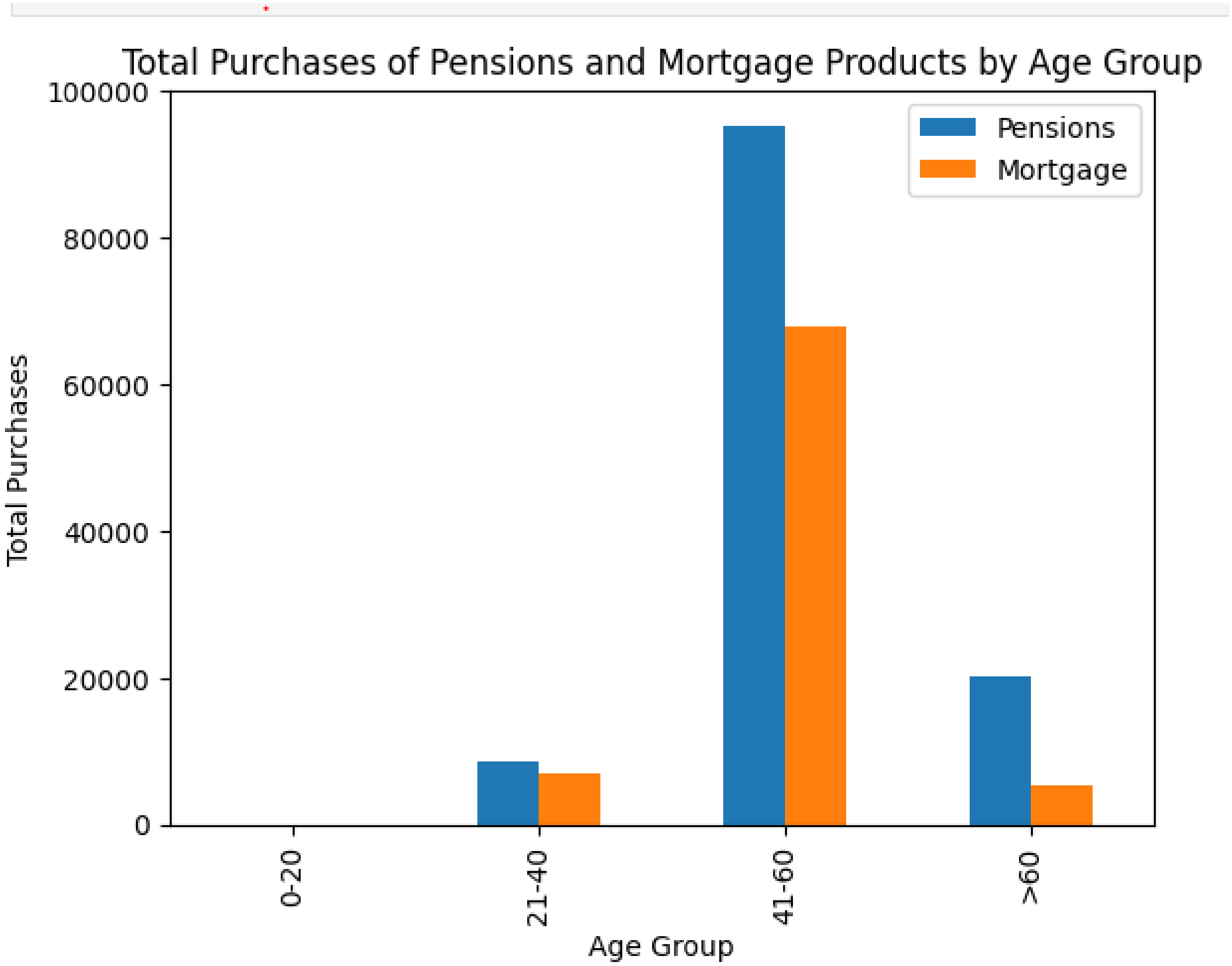
```
]: # create a contingency table of Gross_income and product purchase
   income_purchase_ct = pd.crosstab(df['Gross_income'] > 80000, df[['Current_accounts', 'Derivative_account', 'Payroll_account', 'Junior_account']].sum

]: # perform chi-square test of independence
   from scipy.stats import chi2_contingency

]: chi2, pval, dof, exp = chi2_contingency(income_purchase_ct)

   print(f"Chi-square statistic: {chi2}")
   print(f"P-value: {pval}")

  Chi-square statistic: 17276.54697317662
  P-value: 0.0

]: # create a contingency table of Gross_income and product purchase
   income_purchase_ct1 = pd.crosstab(df['Gross_income'] > 80000, df[[ 'More_private_account', 'Private_account', 'Private_plus_account', 'Short_term_de

]: chi2, pval, dof, exp = chi2_contingency(income_purchase_ct1)

   print(f"Chi-square statistic: {chi2}")
   print(f"P-value: {pval}")

  Chi-square statistic: 100128.51608180723
  P-value: 0.0

]: # create a contingency table of Gross_income and product purchase
   income_purchase_ct2 = pd.crosstab(df['Gross_income'] > 80000, df[[ 'E_account', 'Loans', 'Credit_card', 'Direct_debit']].sum(axis=1))

]: chi2, pval, dof, exp = chi2_contingency(income_purchase_ct2)

   print(f"Chi-square statistic: {chi2}")
   print(f"P-value: {pval}")

  Chi-square statistic: 29666.093720108183
  P-value: 0.0
```

Hypothesis 5: younger customers may be more likely to purchase products like savings accounts, payroll accounts, Derivative account and current accounts



Younger people have more purchase interests in Current_account and Payroll_accounts

Hypothesis 6: older customers may be more interested in pension and mortgage products



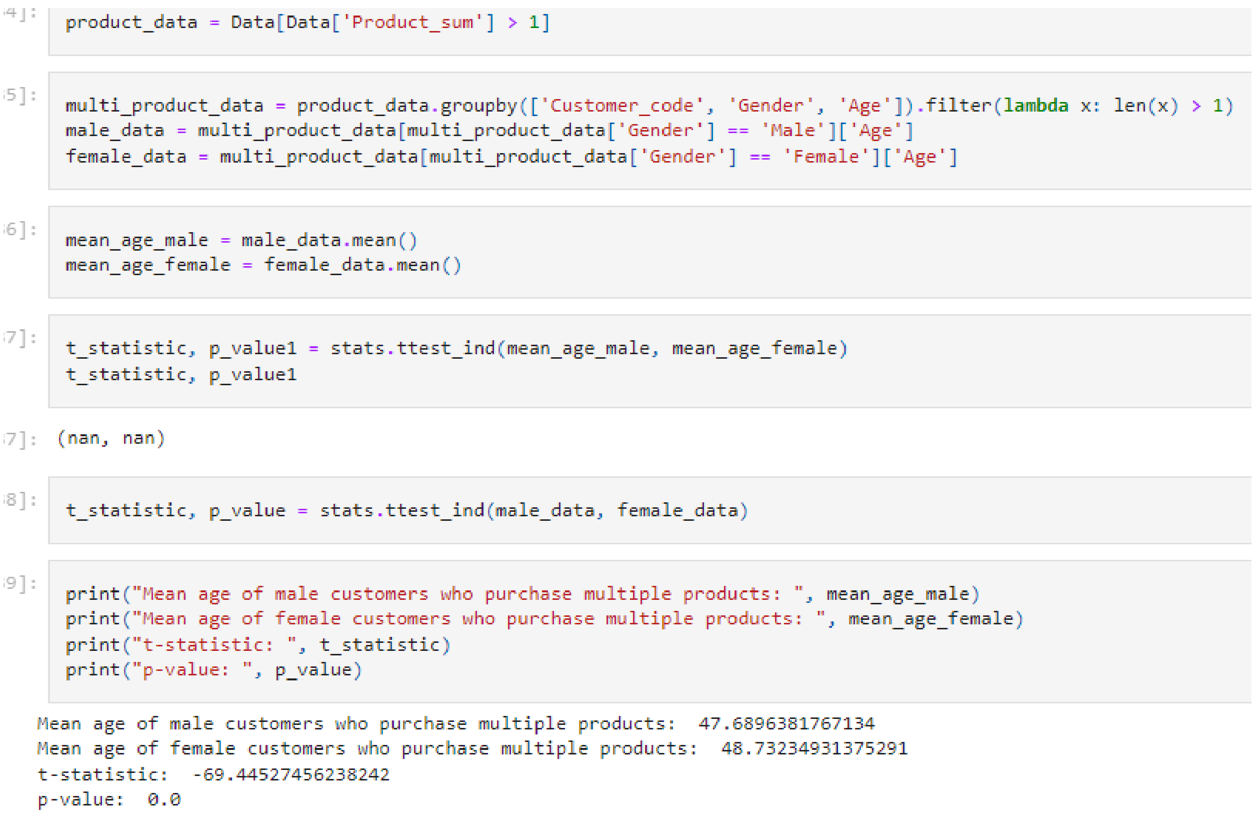
The bank has less amount of age >60 people so when its comes to age, older people buy more Pensions and Mortgages compared to Younger group.

Hypothesis 7: Are there any patterns in the gender and age of customers who purchase multiple products?

H0: There is no significant relationship between Gender and Age of Customers who purchase Multiple products.

H1: There is relationship between Gender and Age of Customers who purchase Multiple products.

- We will conduct a two-sample t-test to test the hypothesis. The significance level is set to 0.05.

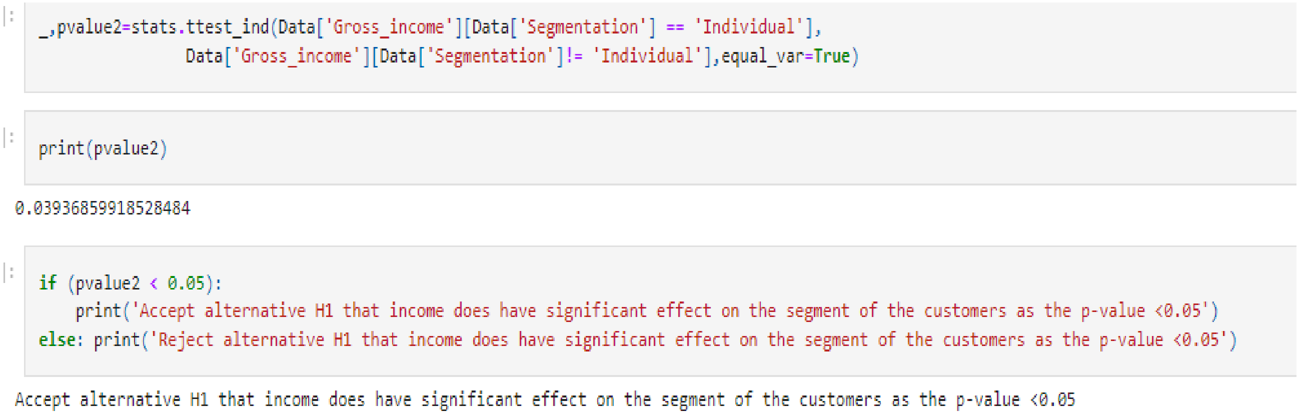


```
4]:  product_data = Data[Data['Product_sum'] > 1]

5]:  multi_product_data = product_data.groupby(['Customer_code', 'Gender', 'Age']).filter(lambda x: len(x) > 1)
     male_data = multi_product_data[multi_product_data['Gender'] == 'Male']['Age']
     female_data = multi_product_data[multi_product_data['Gender'] == 'Female']['Age']

6]:  mean_age_male = male_data.mean()
     mean_age_female = female_data.mean()

7]:  t_statistic, p_value1 = stats.ttest_ind(mean_age_male, mean_age_female)
     t_statistic, p_value1

7]: (nan, nan)

8]:  t_statistic, p_value = stats.ttest_ind(male_data, female_data)

9]:  print("Mean age of male customers who purchase multiple products: ", mean_age_male)
     print("Mean age of female customers who purchase multiple products: ", mean_age_female)
     print("t-statistic: ", t_statistic)
     print("p-value: ", p_value)

   Mean age of male customers who purchase multiple products:  47.6896381767134
   Mean age of female customers who purchase multiple products:  48.73234931375291
   t-statistic:  -69.44527456238242
   p-value:  0.0
```

Hypothesis 8:

H0: There is no statistically significant effect of the segment of customers (Individuals/College-graduates/VIPs) and the income factor.

H1: There is a statistically significant effect of the segment of customers (Individuals/College-graduates/VIPs) and the income factor



```
]:  _,pvalue2=stats.ttest_ind(Data['Gross_income'][Data['Segmentation'] == 'Individual'],
                  Data['Gross_income'][Data['Segmentation']!= 'Individual'],equal_var=True)

]:  print(pvalue2)

 0.03936859918528484

]:  if (pvalue2 < 0.05):
        print('Accept alternative H1 that income does have significant effect on the segment of the customers as the p-value <0.05')
    else: print('Reject alternative H1 that income does have significant effect on the segment of the customers as the p-value <0.05')

 Accept alternative H1 that income does have significant effect on the segment of the customers as the p-value <0.05
```

Recommendations:

- From the data, we saw that customers who have been with the bank for at least 10 years are few, but majority of these customers purchase more than 1 product. It is recommended that promotional products should be offered to customers who have been in the bank for at least 10 years as a way of rewarding their loyalty, this may encourage purchase of more than one product by these customers.

- The data also showed that majority of existing customers have been in the bank for less than 10 years, and a good number of these customers purchase only one product. It is recommended that customer loyalty be built by implementing policies that will build customer trust in the bank and its products. This may help to build customers confidence in the bank and encourage purchase of multiple products.

- Analysis of the data provided results which showed that majority of customers have an inactive relation with the bank. Majority of these customers purchased only 1 product, or no product. The results from the chi square test shows that purchase of products is associated with customer relation categories; Strategies should be taken to improve the activity level of customers in the hope of increasing customers interest in multiple products.

- From analysis of the data, we find that majority of existing customers who have an active relation with the bank purchased more than 1 product. The results from the chi square test shows that purchase of products is associated with customer relation categories; this implies that customers who have active relation with the bank are more likely to purchase multiple products. XYZ credit union should target active customers when advertising products, this may increase purchase of multiple products by customers.

- Since Payroll Account and Direct Debit are commonly purchased together, it would be beneficial for the bank to create bundle offers for these two products, this may help to encourage purchase of multiple products by customers.

- Saving Account is commonly purchased alone, which could indicate that customers are not aware of the benefits of bundling their products. The bank should consider creating awareness campaigns targeted towards customers with only savings account, to create awareness of the benefits of bundling.

- As there has been a decline in the purchase of several products year on year, the bank should review and analyze the causes of the decline. This will enable them to identify the issues and take corrective actions to prevent future occurrence.

- The four most popular products are Current Accounts, Private Account, Direct Debit, and E-Account. The bank should focus on improving these products further and continue to promote them to attract more customers.

- The data suggests that older age customers are buying more Pensions and Mortgage products than younger age customers. The bank should consider promoting these products more to the older age group, while also introducing more age-specific products.

- Younger age customers are buying more Current Accounts and Payroll Accounts than older age customers. The bank could create more personalized products to attract more younger age customers. These products should cater to their needs and offer additional benefits to meet their specific requirements.

- From analyses, the trend in customers joining the bank overtime showed that in the year 2014, the highest number of customers joined the bank. After this year, there was a great decline in the number 0f customers who joined the bank. It is recommended that the bank implements strategies to attract more customers towards the bank.

- From the analysis, we also see that female customers purchase multiple products more than male customers. Based on this, it is recommended that promotions should be made to encourage increased purchase of products by male customers. Also, since female customers purchase multiple products more, they can serve as a target market when advertising or marketing products for cross selling.

- Introducing loyalty programs and benefits for existing inactive customers to involve and explore relevant banking products to increase revenue through them

- Establishments of more banking facilities and XYZ branches across provinces with lower customer strength for provinces with higher customer presence like Madrid, Barcelona engaging customers with new or unique programs in mortgages (for businesses), travel/foreigner accounts (leisure activity, travel industry), insurances, car loans etc. Marketing through a broad spectrum of Channels and introducing engagement programs will increase the customer presence throughout demographics Engaging more with college-graduates with their upcoming transitions may be from short term account to long term, in terms of Education loans, travel insurances, rental loans

GitHub Repo Link:

https://github.com/HarikaReddyB/Cross_selling_recommendation---Group_Project/tree/main/Week%2010