Customer Segmentaton Project

Virtual Internship

10/27/2022

Group Information

Group Name: cust_seg

Specialization: Data Science

Submitted to: Data Glacier

Internship Batch: LISUM 12

Github Repo: https://github.com/Brennan-Clinch/Customer-Segmentation-Project

Team Member Details:

- Brennan Clinch, bclincher98@gmail.com, USA, North Carolina State University, Data Science
- Rohit Sunku, rgs8890@gmail.com, UK, Le Wagon, Data Science (Not contributing)
- Kutay Selçuk, kutay.selcuk@ozu.edu.tr, Turkey, Ozyegin University, Data Science (Not contributing)
- Zhan Shi, zhanshi@g.ucla.edu, USA, University of California Los Angeles, Data Science

Problem Description

XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out the same offer to all customers. Instead, they want to roll out personalized offers to a particular set of customers. If they manually start understanding the category of the 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 groups** as this will be inefficient for their campaign.

ABC analytics assigned this talk to their analytics team and instructed their team to come up with the approach and feature which group similar behavior in one category and others in a different category.

Overall Goals

Total Customers: 1000000

Goals:

- Since we are interested in segmenting customers into no more than 5 groups, we will first perform EDA on the distributions of products used by the customers along with other demographics to get an idea on recommendations for grouping the customers along with our modeling.
- After doing EDA, we will use modeling to identify and group customers using unsupervised learning techniques since we are not dealing with any dependent variables. We will choose at least two techniques and compare results with one another to come up with final recommendations.

Goals for EDA

Total Customers: 1000000

Goal: We want to segment customers based on their behavior. The main things we are interested in for our Exploratory Data Analysis are:

1. What products are most used by the customers?
2. What channel has the most amount of customers?
3. What age range has the most customers? Does the age ranges have an effect on the types of products being bought?
4. What products are most used based on the seniority of the customer from when they started?
5. Does the city/province of the customer have an effect on the products that are bought or used?
6. What effect does gender have on the products being bought?
7. Income by city, age, seniority, and gender.

EDA: Channel Distribution

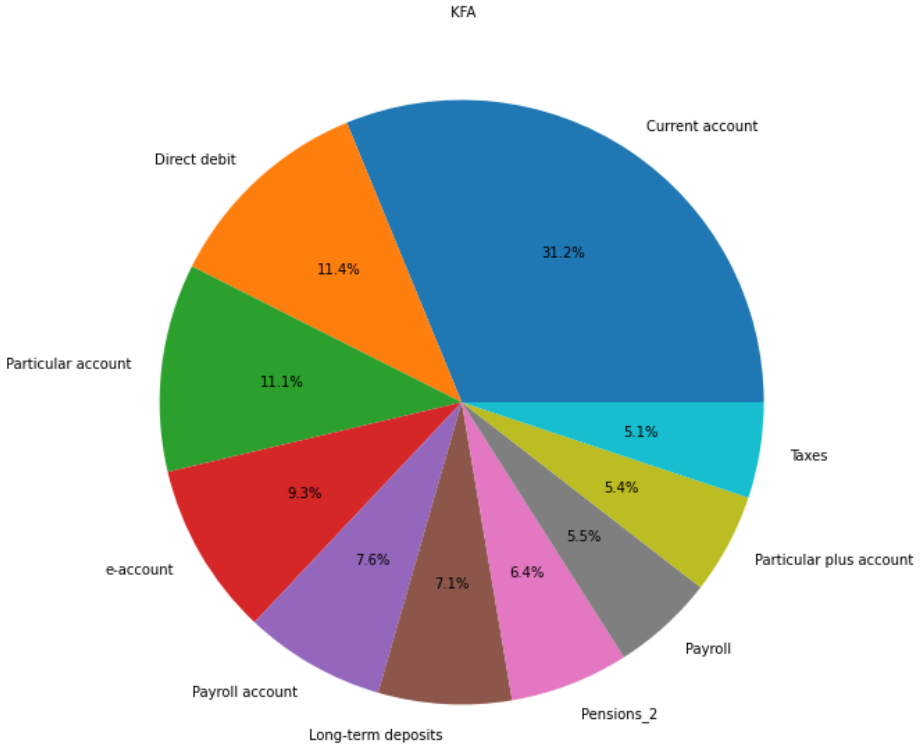
- From the table below, we see that channel “KAT” is the most popular channel being used by the customers.

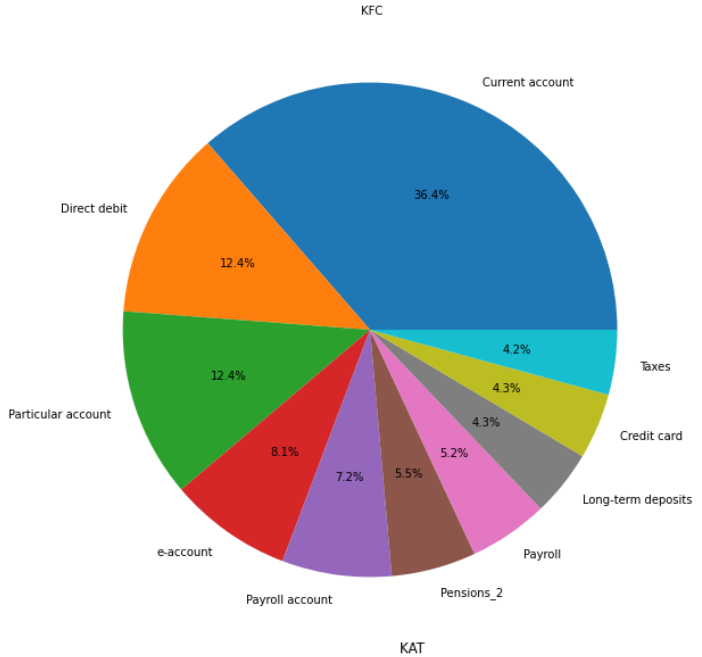
Table of the 10 most popular channels being used

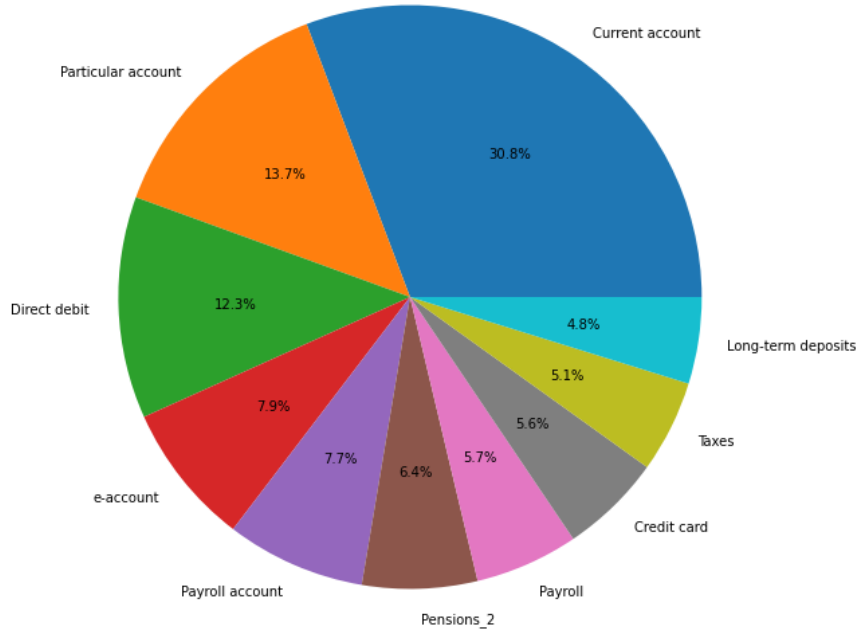
KAT	262959
KFC	215771
KHE	201024
KFA	32524
KAS	7087
KAG	6847
KAA	5342
KAB	5245
KAY	4992
KHK	4823

EDA: Products by Channel Distribution

- For the distribution of products based on channel, we can see that for the top 3 channels that take up the most amount of data, there is no significant difference of products between customers.

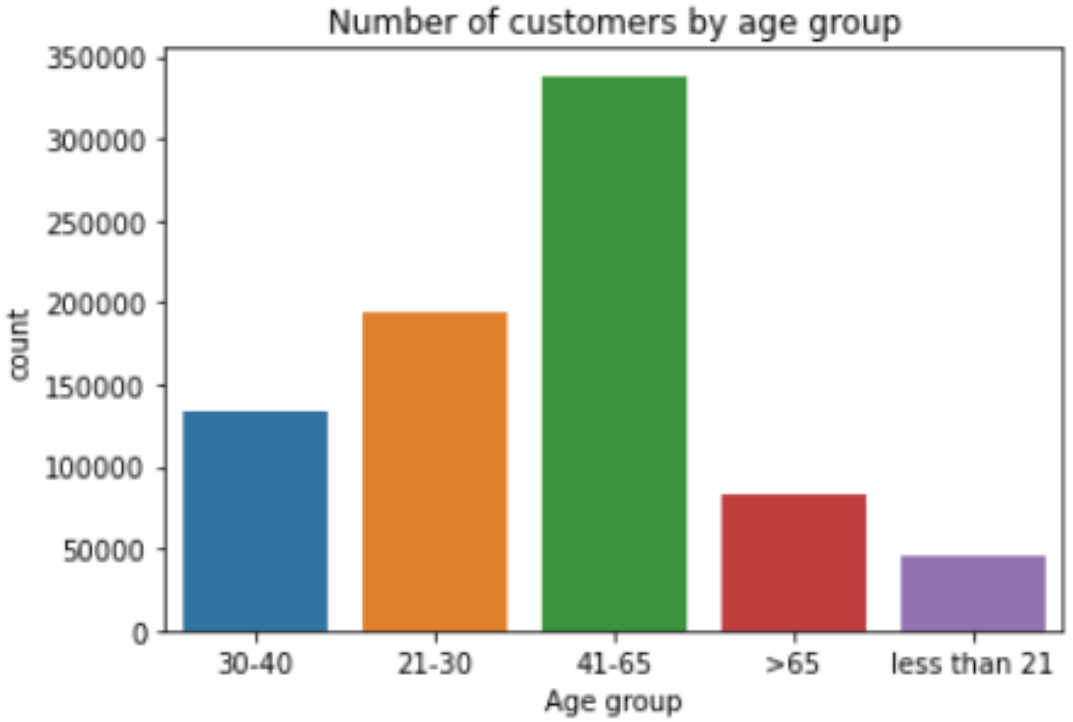






EDA: Age Distribution

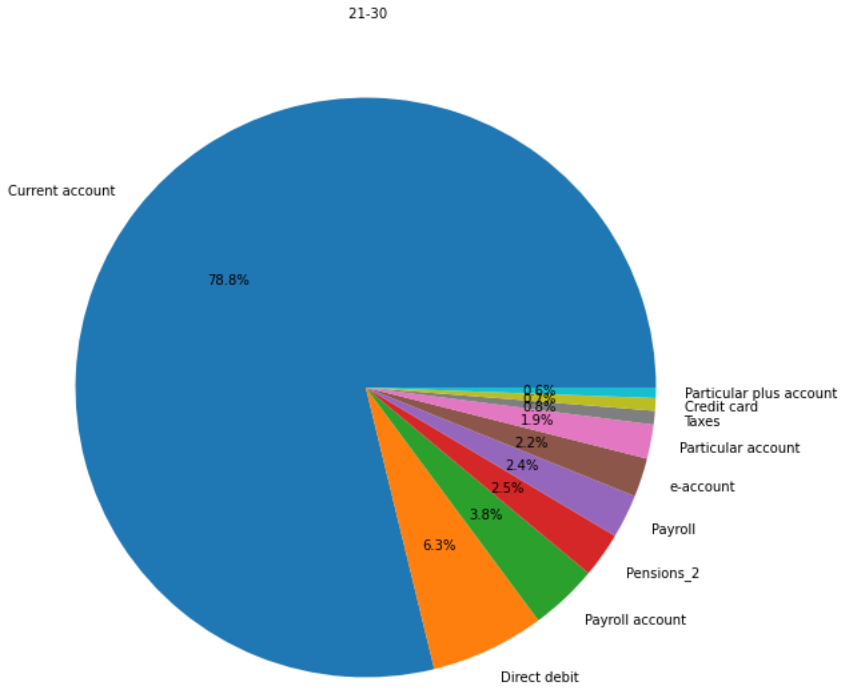
- For the general customer distribution, people aged 41-65 make up the biggest portion of total customers.



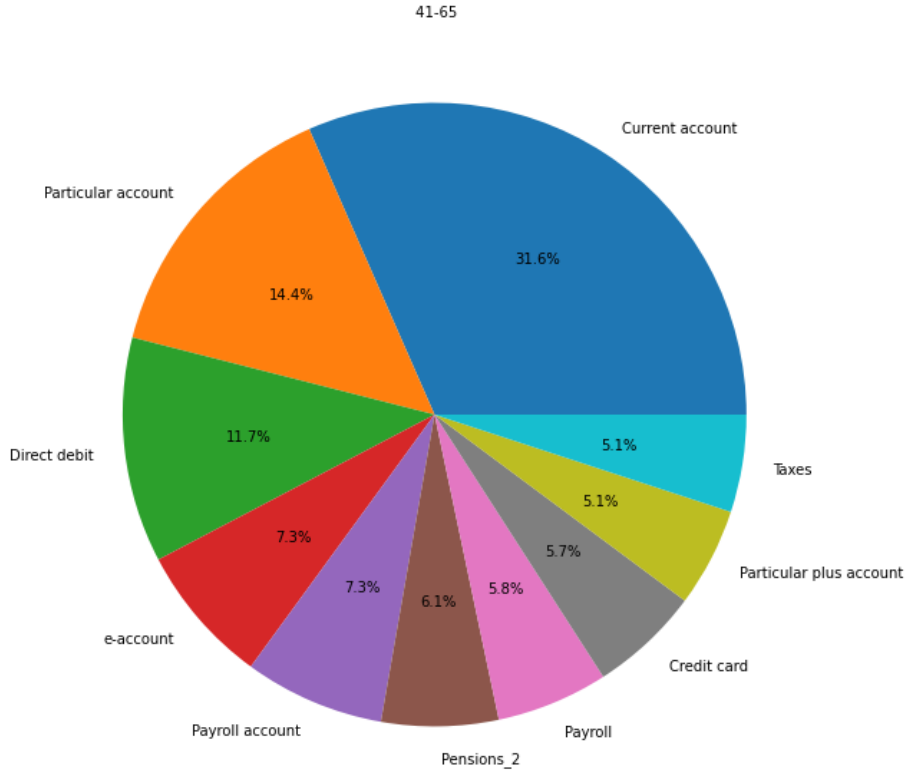
EDA: Product Distribution by Age

- However, customers who are young adults (under 21 and 21-30) make up the biggest proportion of purchases with a Current Account compared to those aged 41-65.
- Also, not shown here, but a Junior Account is a popular product used by those who are under 21.

21-30



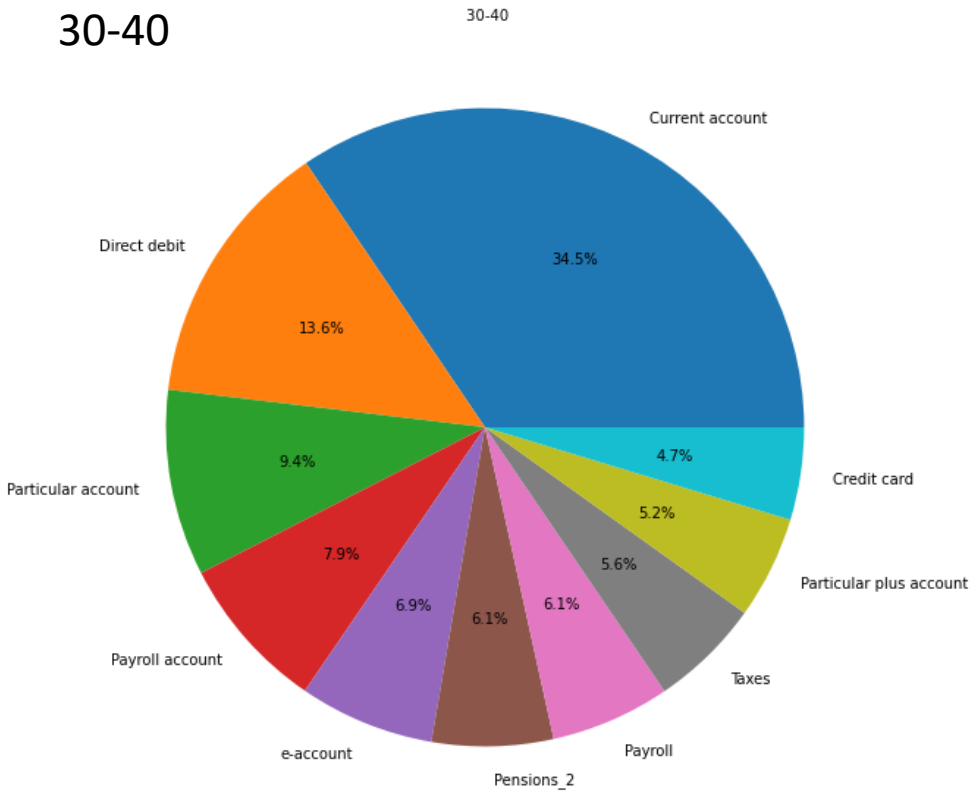
41-65



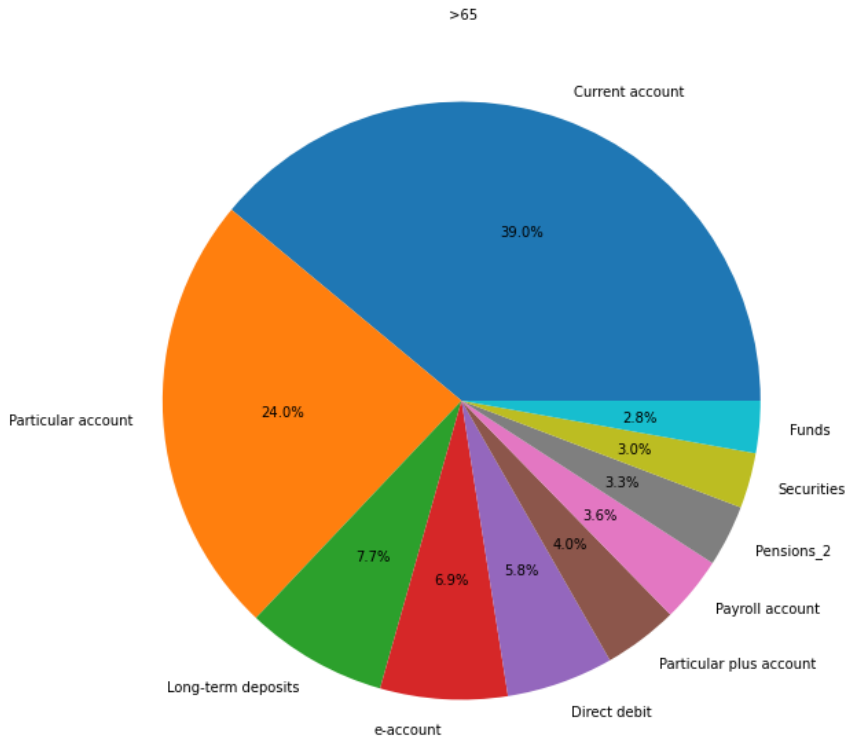
EDA: Product Distribution by Age

- For customers who are over 65, the other most popular account for them is Particular account.
- Direct debit is popular among individuals aged 30-40.

30-40

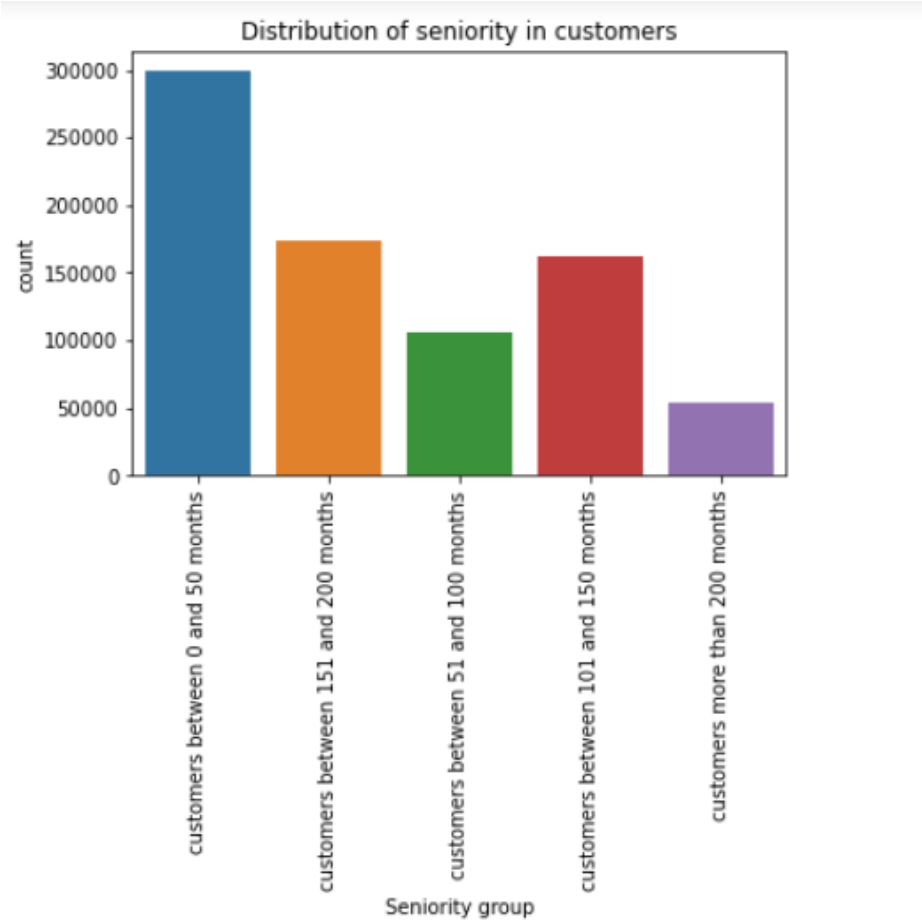


> 65



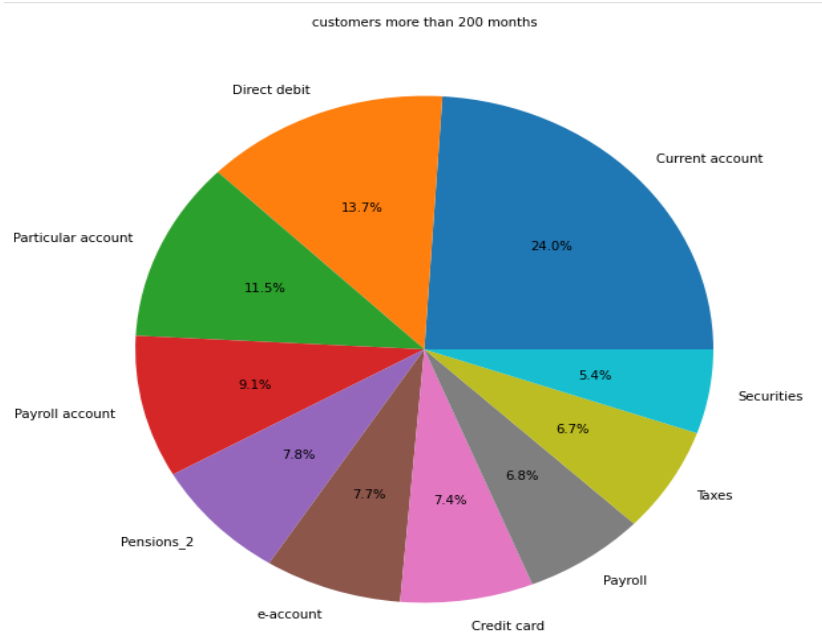
EDA: Seniority of customers

- Most customers are brand new, meaning they have been there less than 50 months (or 4 years).
- The next popular groups are those who have been at company between 100 and 200 months.



EDA: Product Distribution by Seniority

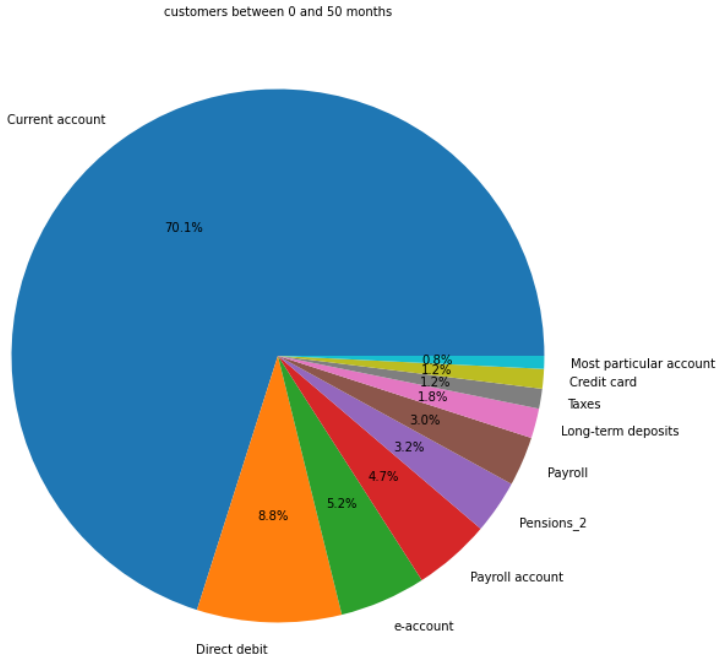
- The amount of customers who have a current account decreases as the time they have been a member increases.
- Customers who have been there 4 years or less are most likely to have a Current Account.

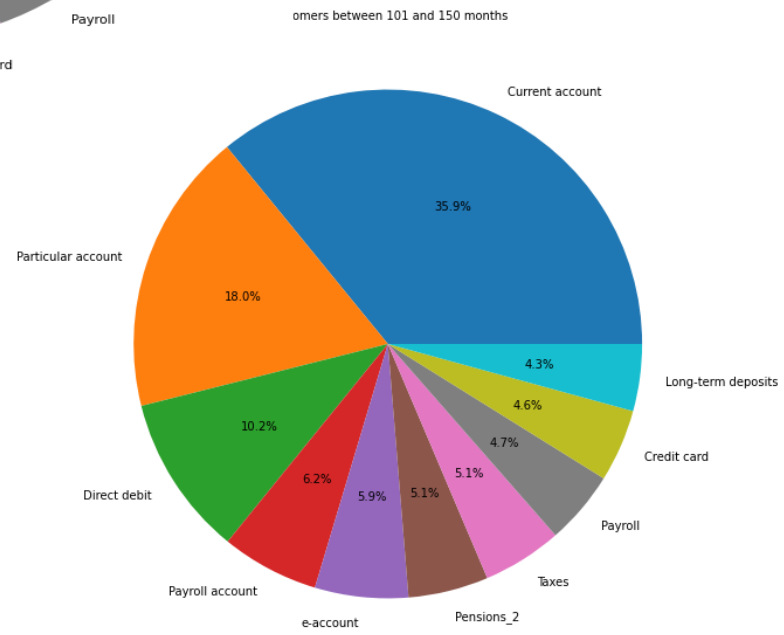


> 200 months

101 and 150 months

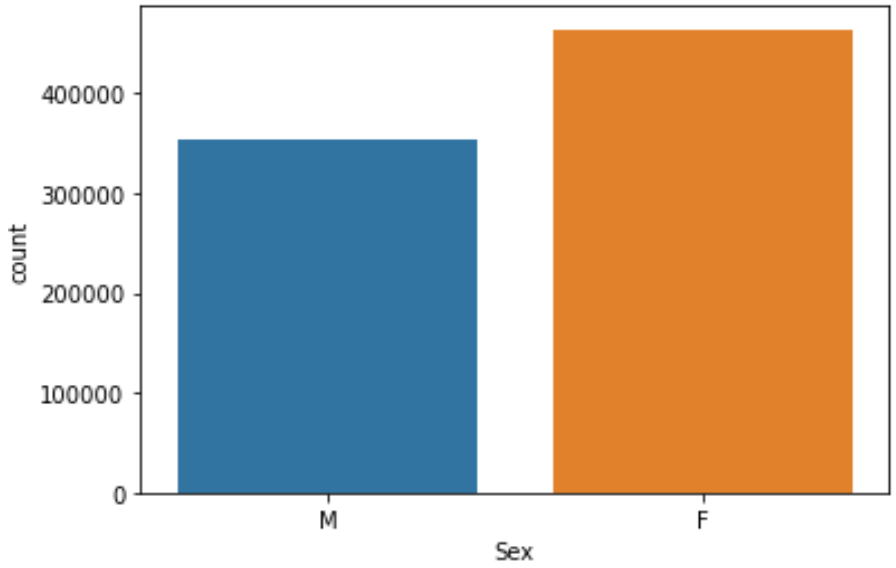
0 and 50 months (~4 years)





EDA: Customers by Gender

- Overall, females are the most popular customer gender.

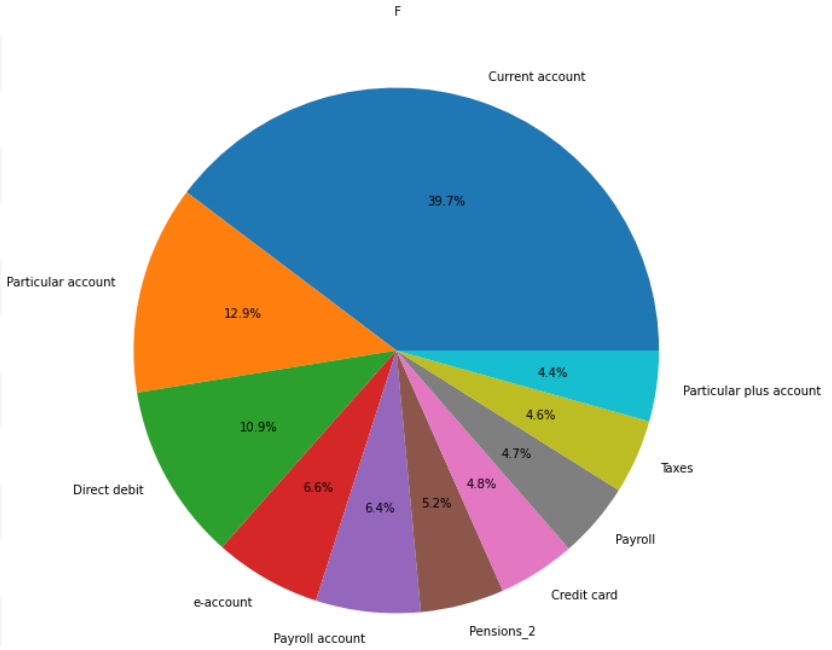


EDA: Product Distribution by Gender

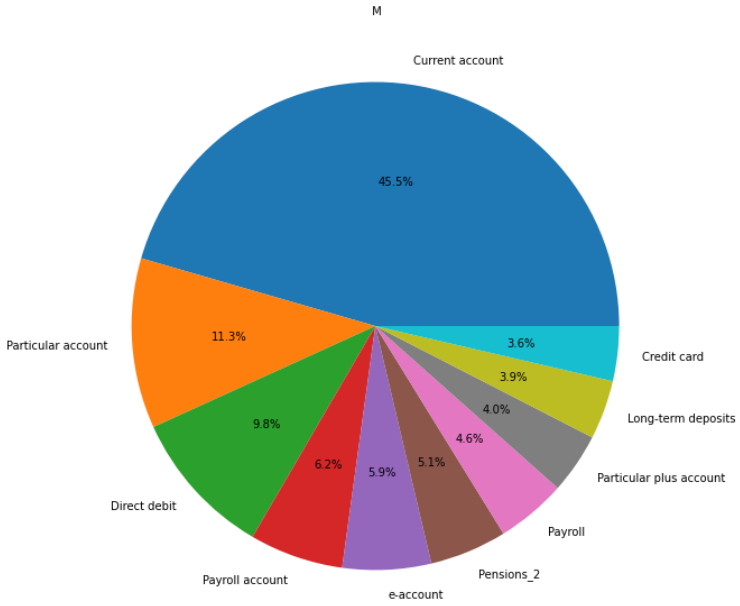
Sex	F	M
Saving account	113.0	40.0
Guarantees	26.0	11.0
Current account	342323.0	270689.0
Derivada account	416.0	71.0
Payroll account	54990.0	36653.0
Junior account	6085.0	5663.0
Most particular account	4830.0	3309.0
Particular account	110931.0	67061.0
Particular plus account	37421.0	23802.0
Short-term deposits	974.0	680.0
Medium-term deposits	1543.0	1107.0
Long-term deposits	34742.0	22836.0
e-account	56494.0	34764.0
Funds	15834.0	7125.0
Mortage	6188.0	2494.0
Pensions_1	7886.0	4740.0
Loans	2295.0	1065.0
Taxes	39577.0	21086.0
Credit card	37272.0	19537.0
Securities	23401.0	9493.0
Home account	3498.0	2114.0
Payroll	37079.0	25392.0

- Overall, females have more products than males, but although the difference isn't significant, there are a little bit more males that have a Current account than females.

Females



Males



EDA: Customers by Country and City

- Based on the tables below, we can see that almost all customers are from Spain, along with Madrid being the most popular city/province.

Top countries of customers

ES	818243
MX	4
PA	2
BE	2
IT	2
DE	2
BO	2
PY	2

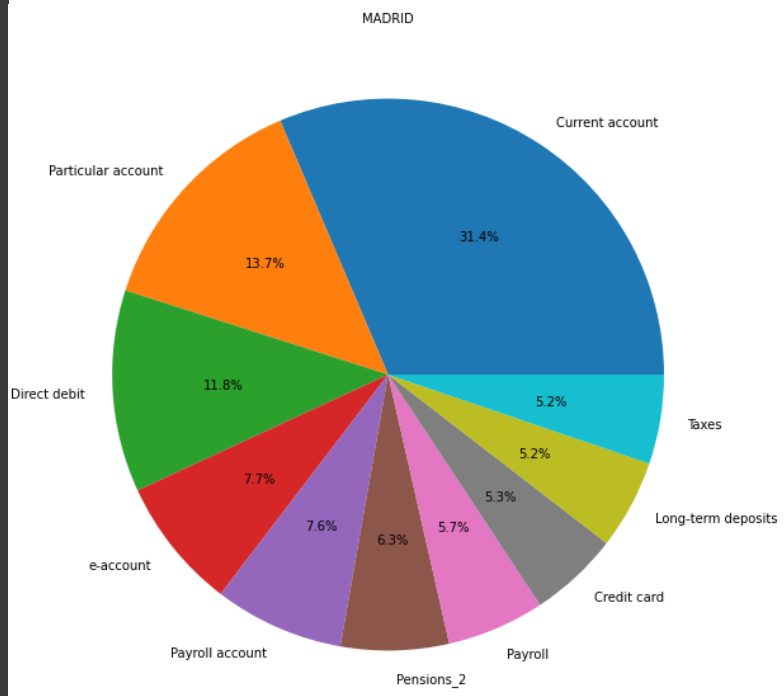
Top 10 cities of customers

MADRID	329081
BARCELONA	74838
VALENCIA	39516
SEVILLA	38786
ZARAGOZA	21882
MALAGA	21501
CORUÑA, A	20801
MURCIA	19378
CADIZ	16692
ALICANTE	16198

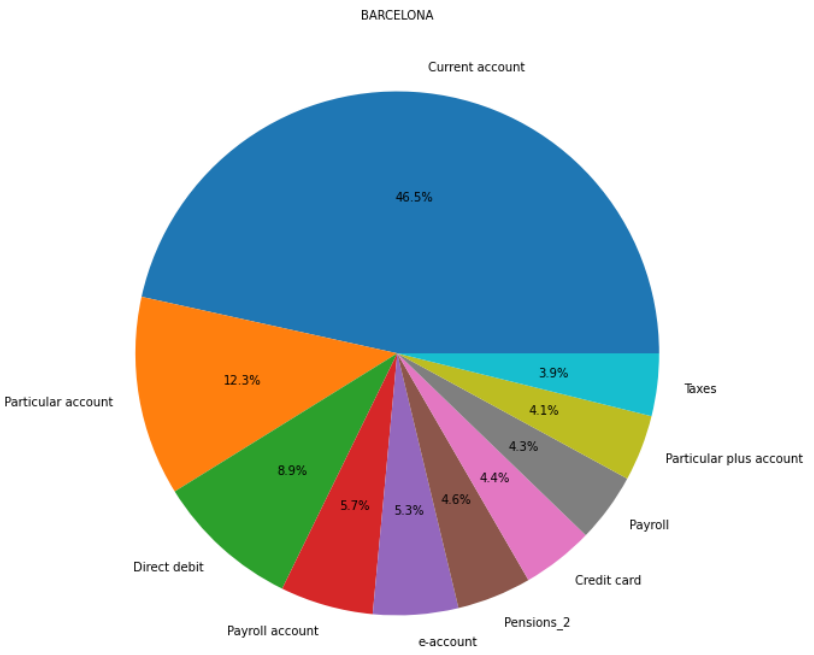
EDA: Customer Distribution by City

- Based on the pie charts, we can see that the city/province can have a significant effect on the product distributions from customers.

Madrid

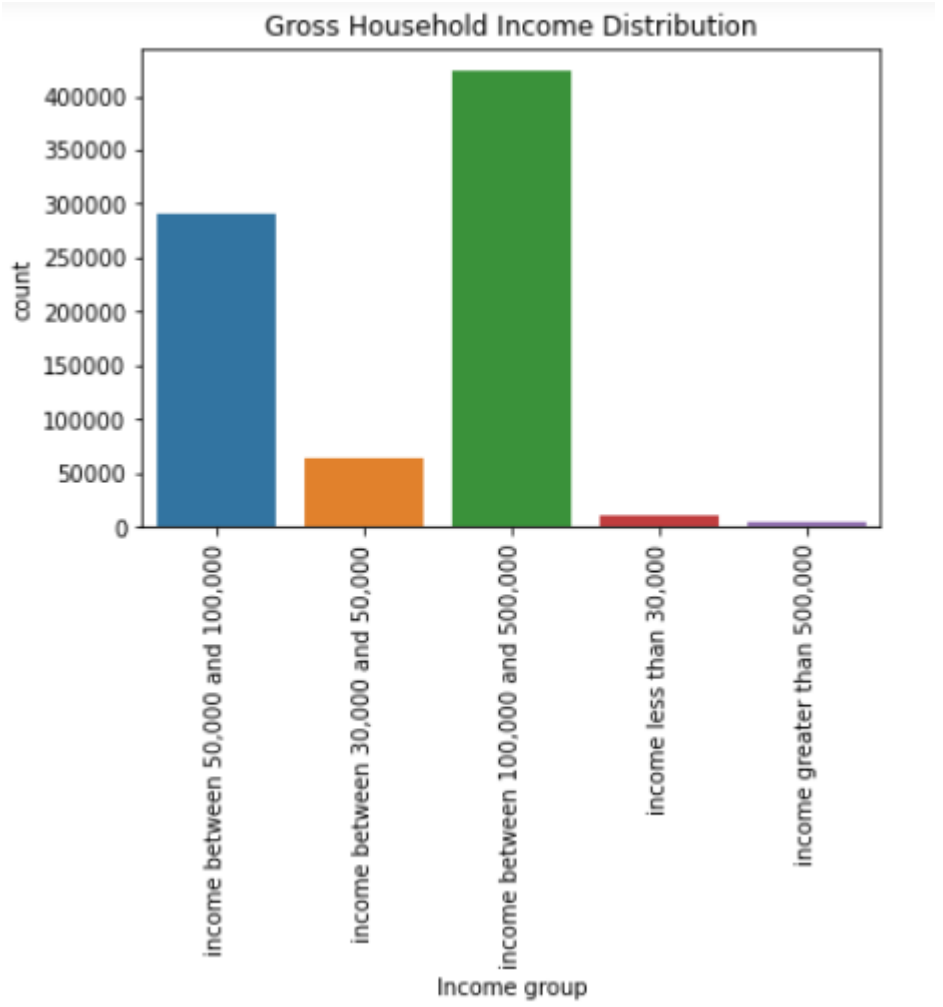


Barcelona



EDA: Customers by income

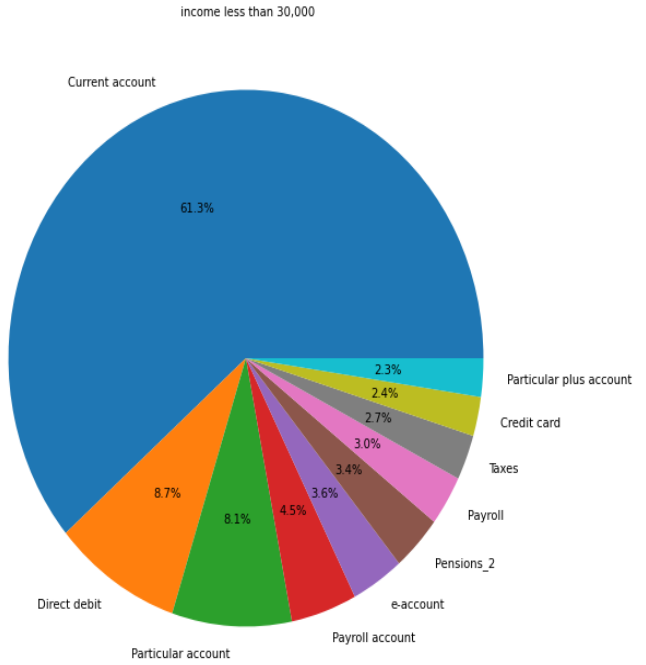
- From the bar graph, we can see that customers usually make between $100,000 and $500,000.



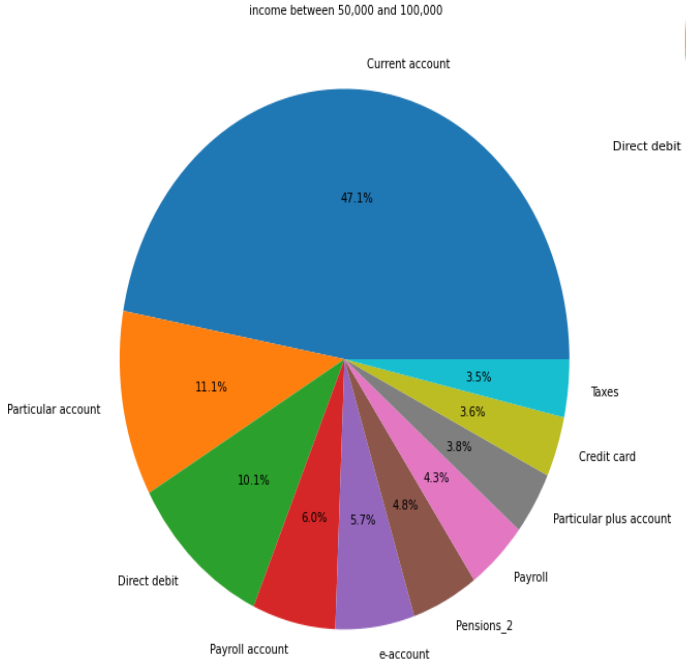
EDA: Product Distribution by Income

- Based on the pie charts, we can conclude that there is a significant difference between the product distributions based on the gross income of customers.

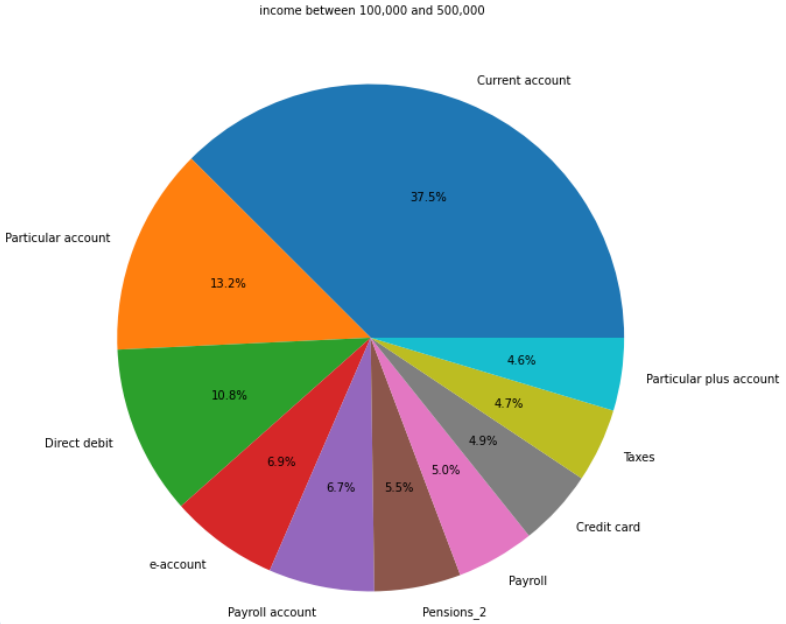
Income less than $30k



Income between $50k and $100k

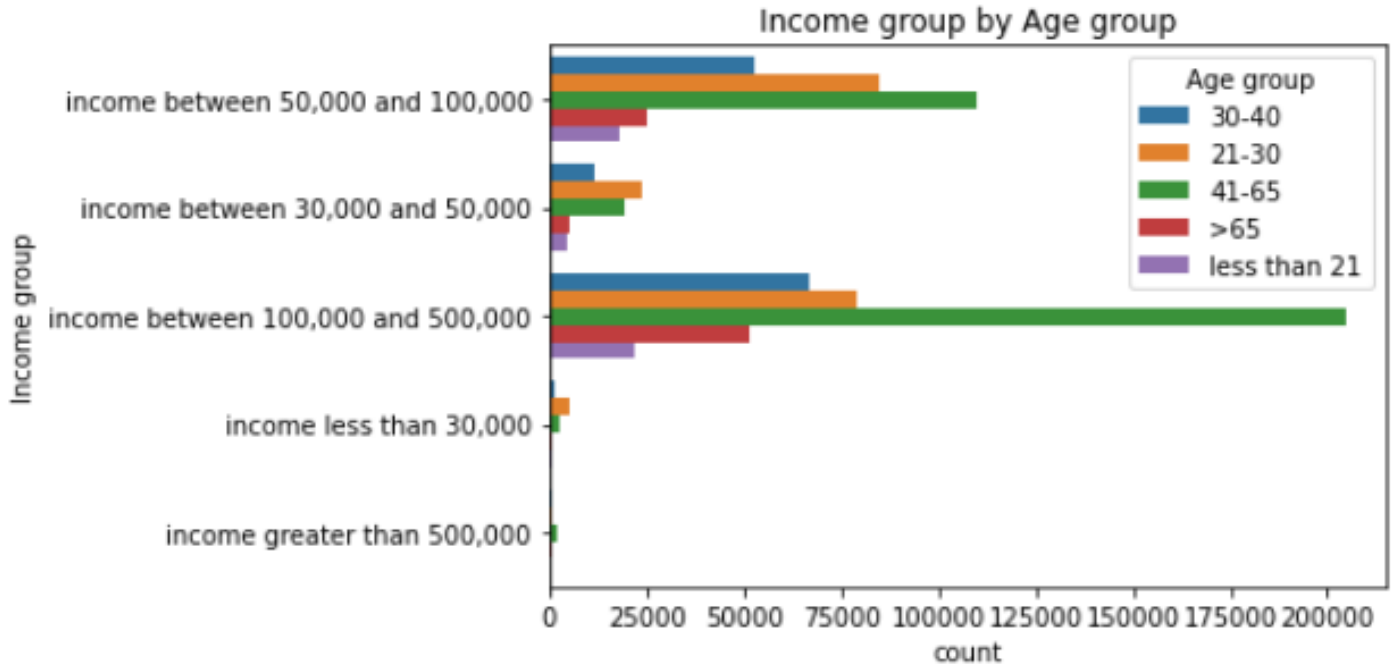


Income between $100k and $500k



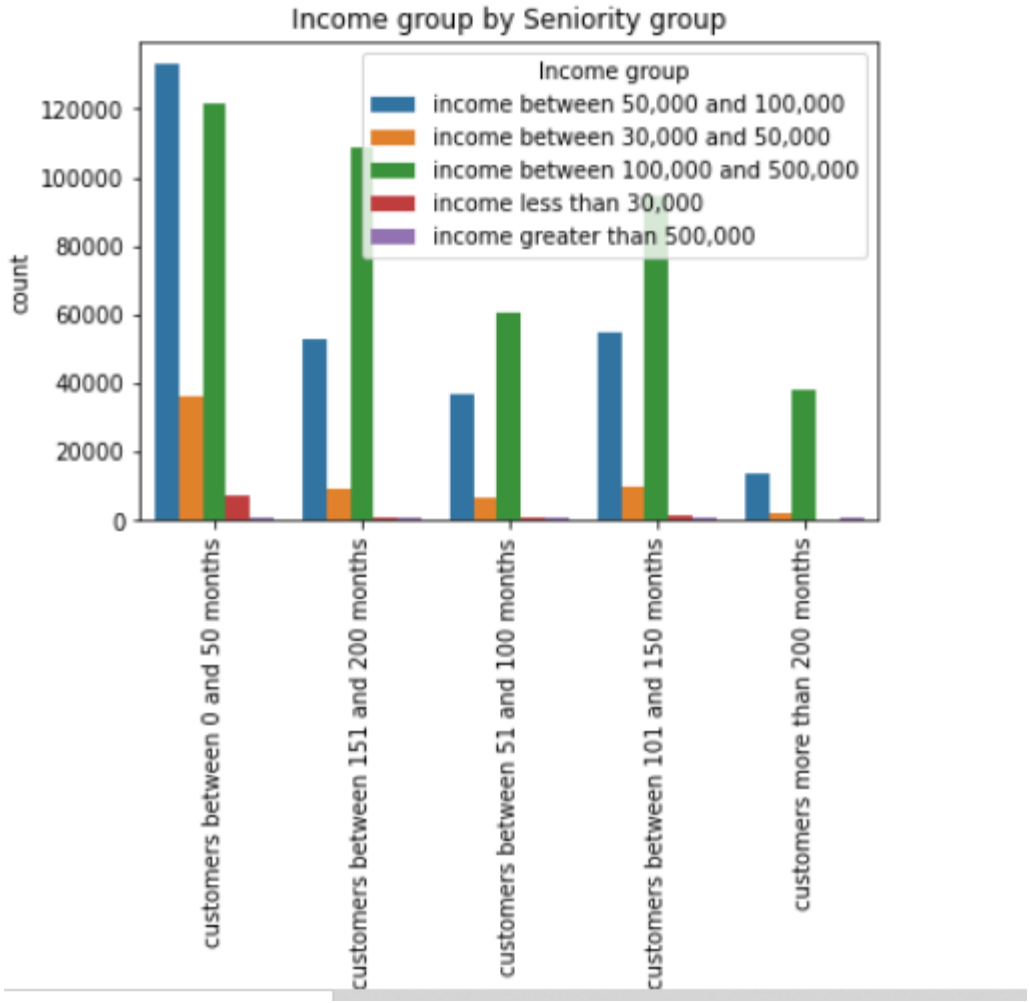
EDA: Income by Age Group

- Customers who are young adults (21-30) are most likely to have a gross household income of between $50,000 and $100,000.
- Customers who are middle aged (41-65) are more likely to have a gross household income greater than $100,000 but less than $500,000. This would most likely make sense as this age group has been working for a while.



EDA: Income by Seniority

- Customers who have gross household income between $50,000 and $100,000 most likely have been a customer for less than 4 years, along with those who have income between $100,000 and $500,000.



EDA: Income by City

- Overall, Madrid was the most popular city by customers for most income groups. However, it was shown that for customers whose gross household income was less than $30,000, other cities had a higher number of customers.

Income group	income greater than 500,000	income less than 30,000
Province name		
ALBACETE	0	233
ALICANTE	33	972
ALMERIA	2	87
ASTURIAS	6	194
AVILA	2	88
BADAJOZ	2	790
BALEARS, ILLES	50	6
BARCELONA	792	70
BIZKAIA	0	0
BURGOS	0	40
CACERES	8	396
CADIZ	63	555
CANTABRIA	28	55
CASTELLON	1	147
CEUTA	0	0
CIUDAD REAL	2	415
CORDOBA	6	401
CORUÑA, A	56	131
CUENCA	0	295
GIPUZKOA	0	0
GIRONA	32	25
GRANADA	19	175
GUADALAJARA	0	30
HUELVA	2	280
HUESCA	1	72
JAEN	0	153
LEON	4	137
LERIDA	3	133
LUGO	0	97
MADRID	3092	382
MALAGA	71	206
MELILLA	4	0
MURCIA	2	562
NAVARRA	0	0

Modeling Choices

- Since we will be working with no dependent variables, we will use **unsupervised** learning methods for our modeling.
- Since we are wanting to divide customers into different groups that share similar product and other demographic behavior in order to break up the groups for special Christmas offers, we will most likely be using clustering, which is when we divide data points up into a certain number of groups so that the data points in same group are similar to the other data points in same group than other groups.
- We will use two different types of clustering methods, shown below.

K-Means Clustering: For K-Means Clustering, the goal of this method is to partition n observations into k clusters in which every observation belongs to the cluster with the nearest mean.

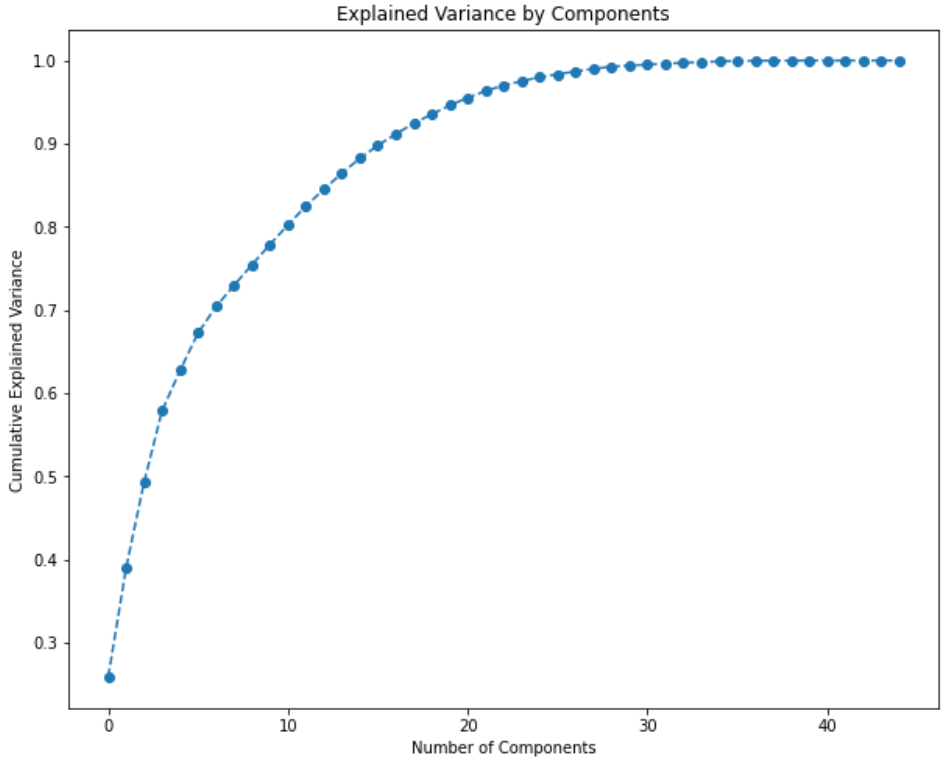
Hierarchical Clustering: For Hierarchical Clustering, the goal of this method is build a hierarchy of clusters.

- **Agglomerative Hierarchical Clustering**: For this type of hierarchical clustering, we start with individual clusters for each data point and then merge the clusters through ranking on their closeness. The final cluster contains all data.
- **Divisive Hierarchical Clustering**: For this type of hierarchical clustering, we start with the cluster containing all the data. For each step, we split the most distant data in the cluster, and is repeated until we have individual data points.
- We will only use Agglomerative Hierarchical Clustering for our analysis

- We will use both methods and compare for validation to finalize our analysis.

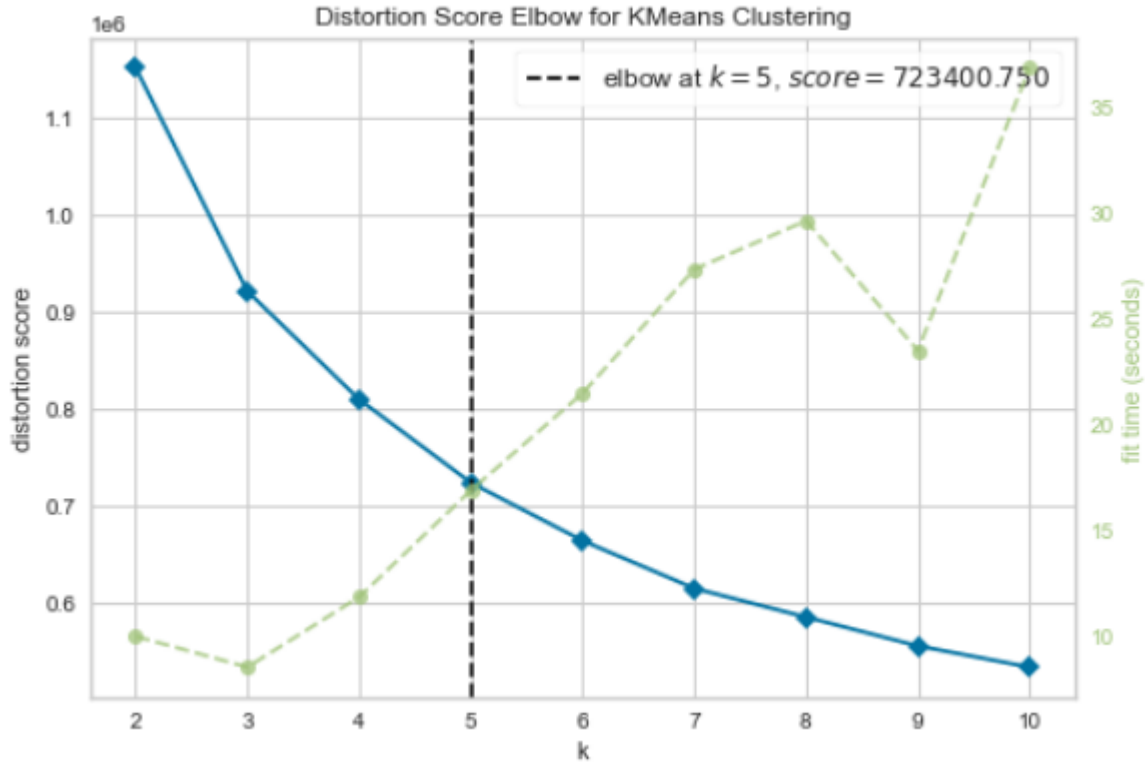
PCA: Principal Component Analysis

- When doing K-Means, it is important that our model is as interpretable as possible. One way we can improve interpretability is through dimensional reduction techniques such as **PCA (Principal Component Analysis).** The goal of PCA is to identify patterns in a data set, and then reduce the dimensionality of the data while retaining as much as possible variation in the original dataset.
- The plot below shows the explained variance by the number of components. A good rule of thumb is to still have roughly 80% of the total variance retained, so the best number of components we should keep is 10.



K-Means: Determining Number of Clusters

- For our analysis, we do not want more than 5 groups as it would be inefficient for the campaign. Let us use the Elbow Method to determine how many clusters we should use.
- Based on the plot below, the optimal number of clusters should not be more than 5, and it is obvious as the plot shows the optimal number is exactly 5 clusters. 5 clusters will allow us enough variety between the groups while also allowing us to not be inefficient to the marketing campaign with having the maximum number of groups.



from sklearn.metrics import silhouette_score

K-Means: Results

Group 1:
- No Saving Account
- No Guarantees
- No Derivada Account
- No Pensions_1
- No Mortgage
- No more than 4 total products
- Employee index filial or not employee
- Seniority less than 200 months

Group 2:
- 2 or more products
- Not including 1 city

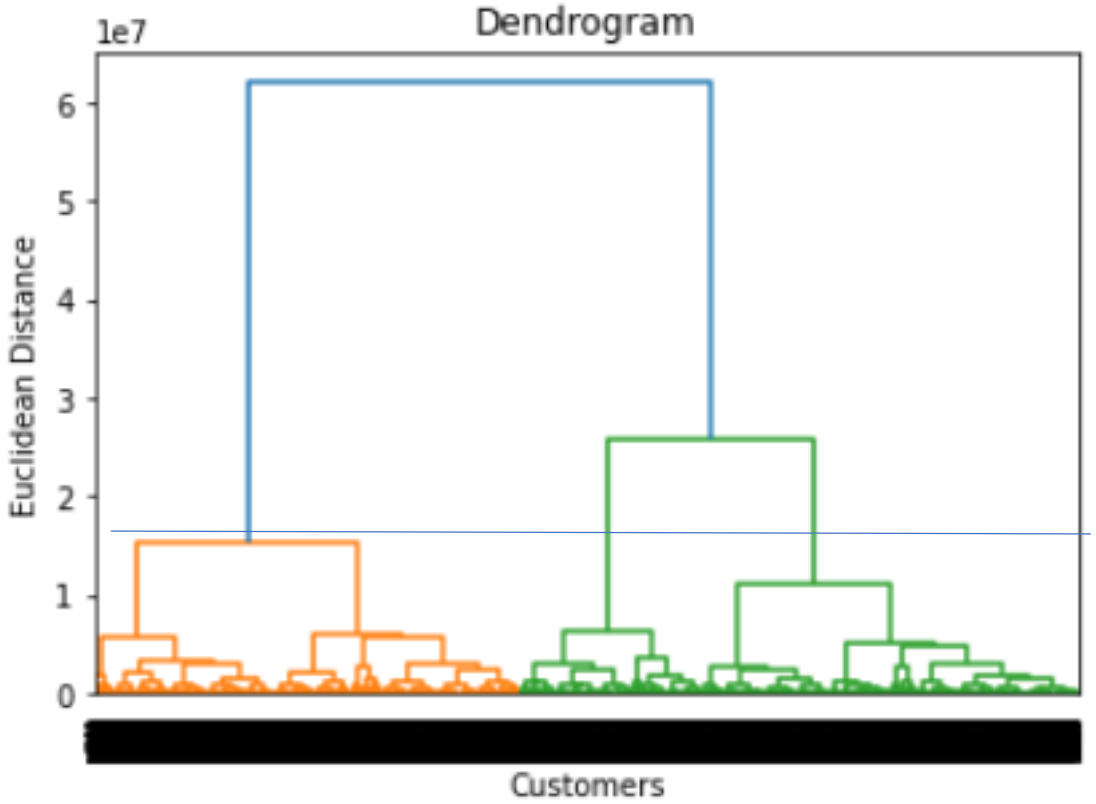
Group 3:
- No Junior Account
- No more than 11 total products
- Not including 2 cities
- Ages 21, 20, 19 not included
- Channels KHE and KHK not included

Group 4:
- No Guarantees
- No more than 8 total products
- All but 1 city

Group 5:
- No Guarantees
- No more than 6 total products
- Not including 2 cities

Hierarchical Clustering: Dendrogram for Determining Optimal Number of Clusters

- When determining the optimal number of clusters in Hierarchical Clustering, one way of doing so is through a cluster tree, or Dendrogram.
- To find the optimal number of clusters, we first pick an area where the vertical distance between each linkage is the greatest. From the horizontal line below, the optimal number is 3.



Hierarchical Clustering: Results

Group 1:

- Not including those with 12 total products
- Employee Index not ex employed
- Excluding 2 cities
- Income less than 500k

Group 3:

- No Guarantees
- No more than 12 total products
- Registered after 6 months
- Excluding 6 cities
- Income around 150k to 300k

Group 2:

- No Savings Account
- No Guarantees
- No Loans
- Excluding those with 11 products
- Employee Index not an employee
- Registered after 6 months
- Excluding many cities
- Income greater than 250k

Comparison

K-Means:

Pros:

- Good for large data sets
- Simple to implement

Cons:

- Scaling with multiple dimensions
- Lacks consistency
- Sensitive to outliers

Agglomerative Hierarchical:

Pros:

- Hierarchy is easier to interpret the number of clusters than in K-Means

Cons:

- Unable to cover very large amount of data, so you can only use a small percentage
- Sensitive to outliers

- So, the best model choice for our analysis is K-Means since we are dealing with a very large amount of data (1 Million records, 40+ variables)

Final Recommendation

Based on the results from K-Means clustering and Hierarchical Clustering, we will be going with most of the results from the K-Means model. However, there are still a few recommendations based on our EDA and the Hierarchical model that could also be used in the splitting of customers into groups.

- From the Hierarchical model, we did see that Income was a significant factor when dividing up customers, so maybe, we could have a couple groups that didn't have Guarantees have income between 150,000 and 300,000.
- We could also consider the time of registration as a couple groups from the Hierarchical model contained customers who registered after the first 6 months.
- The bank could also focus maybe on other cities/provinces in Spain.

Thank You