Week 3 — Perform Exploratory Data Analysis

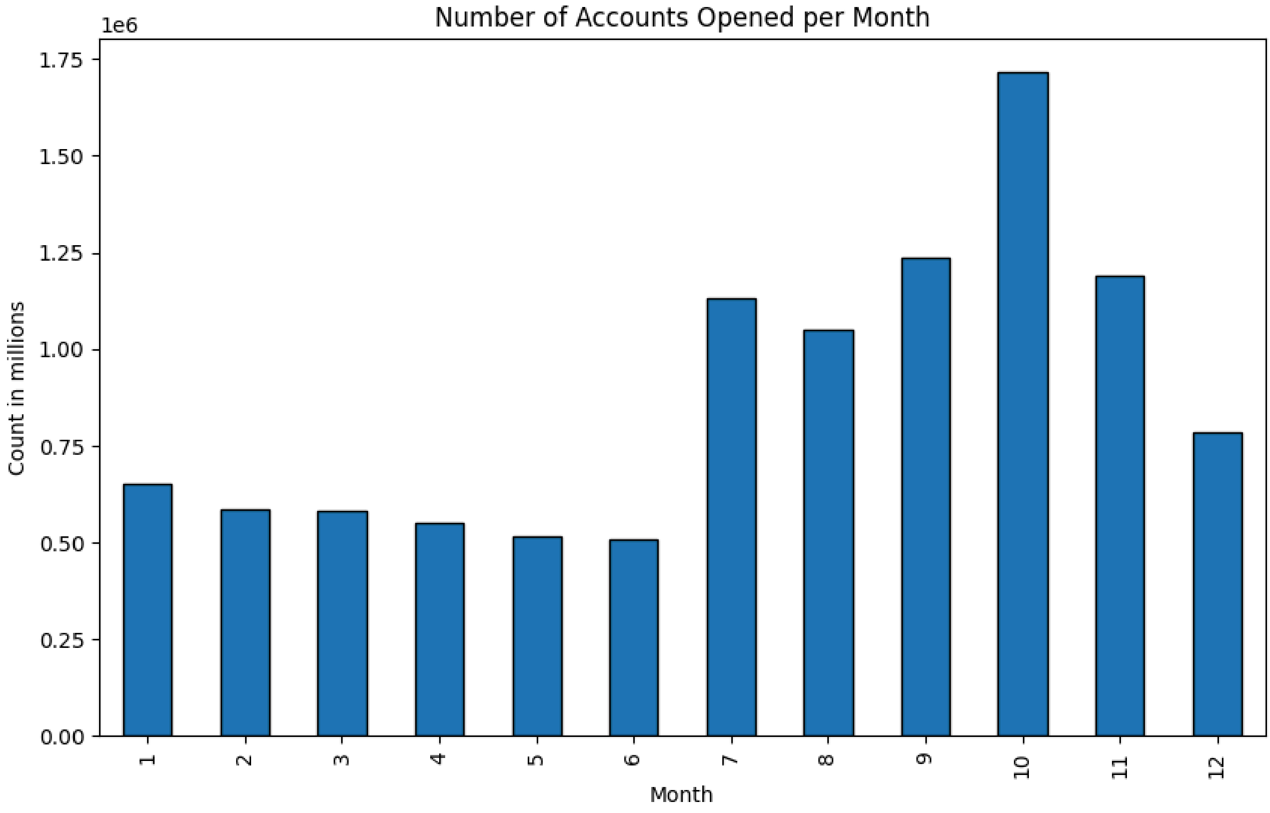
Maria Alice Fagundes Vieira
Aritra Ray

Review the dataset partition strategy and explain why you selected the specific strategy

For this project, we have selected to divide the dataset in an 80:20 ratio, with 80% set aside for training and 20% for validation. We have a separate test dataset for testing. The objective for this divide is to guarantee that the model has enough data to learn the underlying patterns while still retaining some data for evaluating its performance on previously unseen examples. This fraction is extensively employed in machine learning because it provides an appropriate balance between model training and evaluation.

EDA analysis, insights: what types of EDA did you run? Show the results.

Before we dropped the first contract date column, we wanted to explore it a little bit to make sure that we were extracting important information from it.We found out that when we plot contract dates by month to see if there is any seasonality, we can conclude that the last half of the year has significantly more new contracts than the first half. This can maybe be explained by raises, bonus, new people getting hired after summer, and other factors

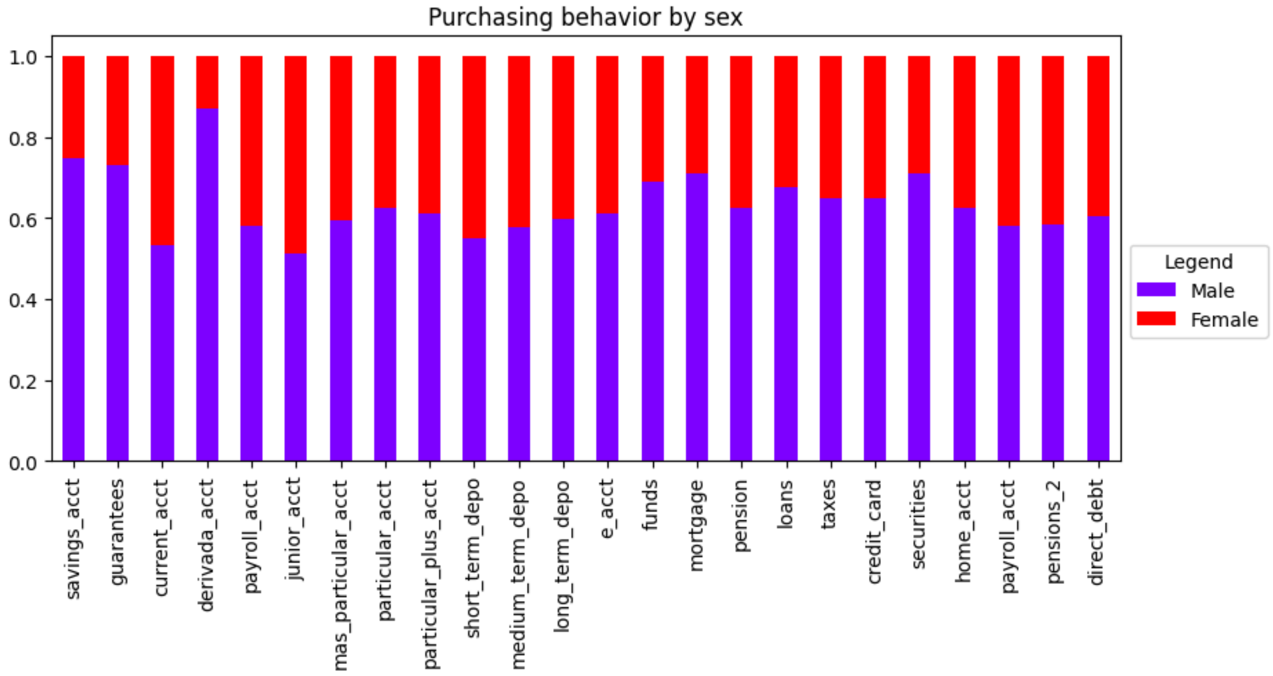


For the EDA analysis we started with a table of descriptive statistics that showed us an important summary of the data and helped us identify any extreme outlier present in the data. It also helps us to identify some initial insights from the data, such as:

- There are more man than women in the dataset
- The average and median age in the dataset is 40 years old
- There is a good range of seniority in the dataset, ranging from 0 to 256 months, or 21.3 years
- Most of the clients in the dataset have their primary residency and birth place in Spain

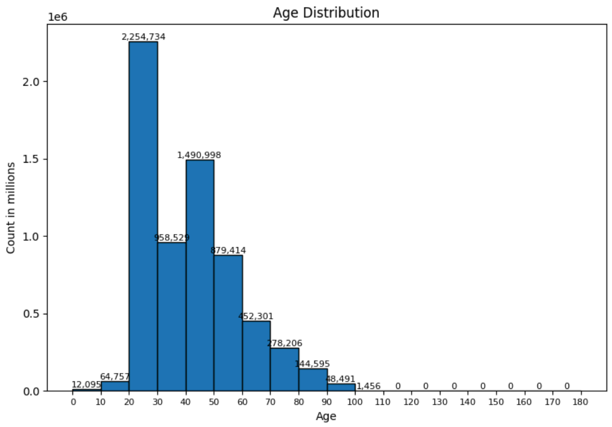
After the descriptive statistics, we proceeded to analyse each column more individually and what are the purchasing behaviour from the different clients within each column. For that, we used a mix of value_counts(), bar and pie plots and a final bar plot showing the distribution of each value in the columns for every product bought.

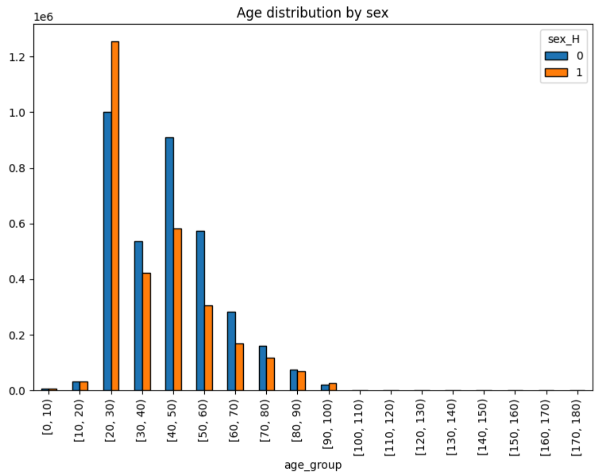
Sex column analysis:

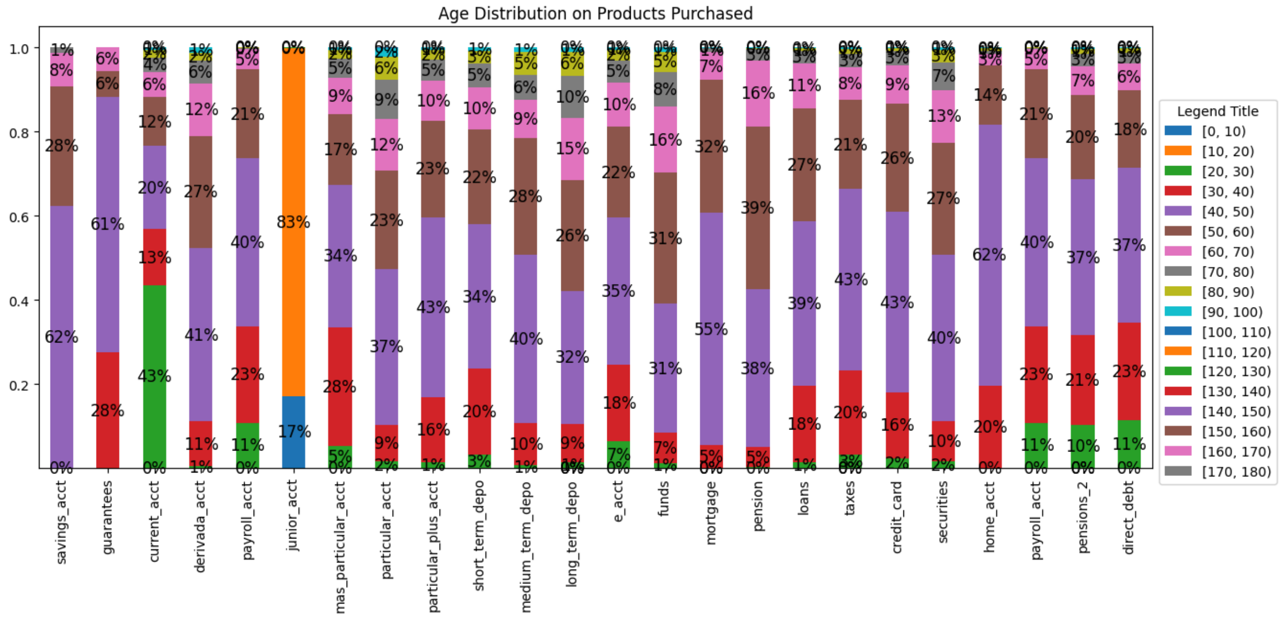


We can see that the bank has more male customers and they have bought more products than female customers, showing a skewed distribution in the total products bought. All the products, except the derivada account have a relatively similar ratio to the amount of male and female customers, showing that there may not be a product preference when it comes to gender, and the difference comes from the number of customers.

Age column analysis:







There are more male than female customers in all age groups, except for those aged 90-100. The majority of customers are young, with the 20-30 age range being the largest group, followed by the 40-50 age group. Most clients purchase their first product in their 20s, and the junior account is primarily held by customers aged 10-20. Overall, there is a good variety in products bought across age groups, with product dominance shifting mainly between the 40-50 and 50-60 age ranges, followed by the 30-40 range, even though there are more clients in the 20-30 age group.

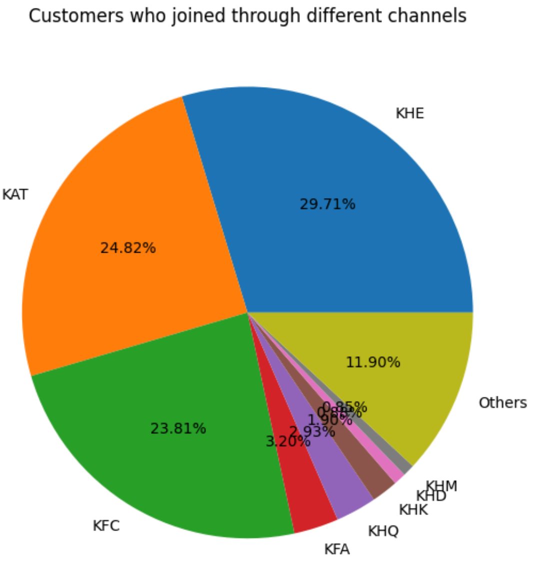
New Customer column analysis:

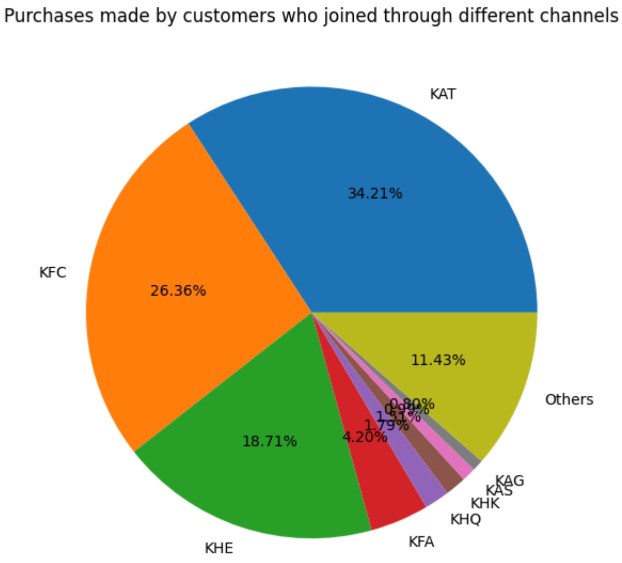
New customers primarily purchase short-term deposits and the "mas particular" account, with minimal participation in other products. This could suggest that only a few of the bank's offerings are effectively attracting new customers.

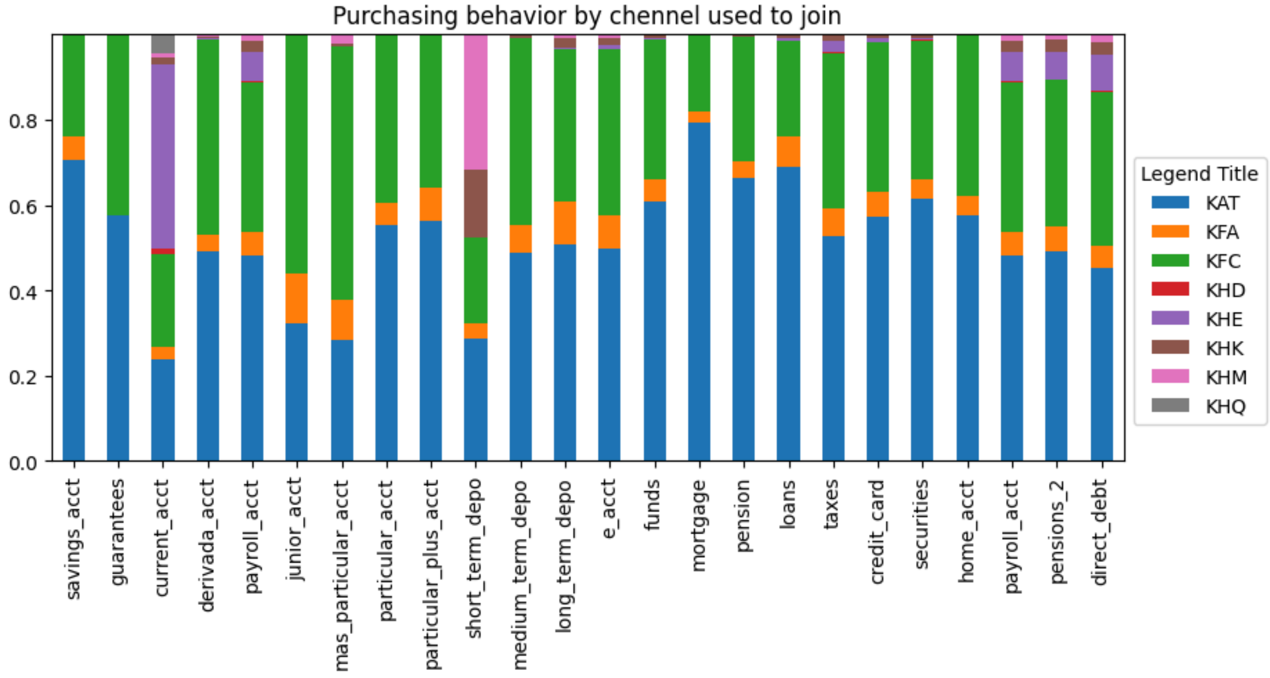
Seniority column analysis:

The bank's newest clients tend to purchase products like the current account, junior account, mas account, and short-term account. In contrast, the oldest customers have acquired savings accounts, particular_plus accounts, mortgages, loans, and home accounts.

Join Channel column analysis:







The top 8 channels used to join the bank account for ~88% of the total customers. When comparing customer purchases with the channels they joined through, 7 out of these 8 channels show consistency in both the number of clients and purchases, except for "others," which had missing values. Customers who joined via the KAT channel made the most overall purchases, while those who joined through KFC surpassed KAT in purchases for the mas particular account and junior account.

Province column analysis:
Most of the bank's clients are from Madrid, followed by cities like Barcelona, Valencia, and Sevilla. The distribution of products purchased closely mirrors the number of clients from each city.

Notably, only clients from Madrid hold guarantees accounts, while those from Barcelona predominantly buy loans and savings accounts.

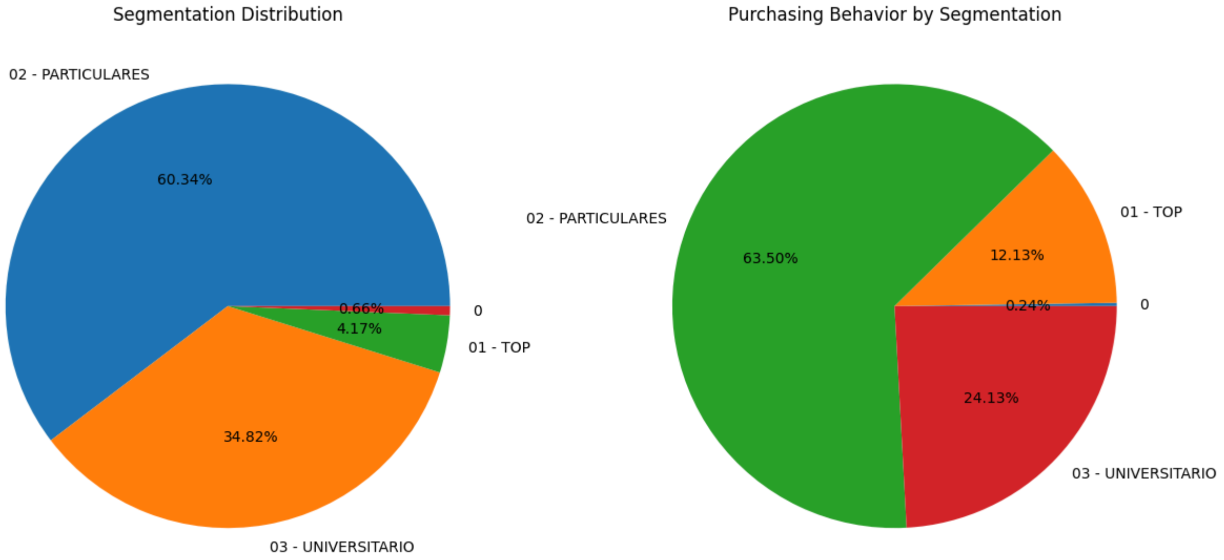
Active Customer column analysis:

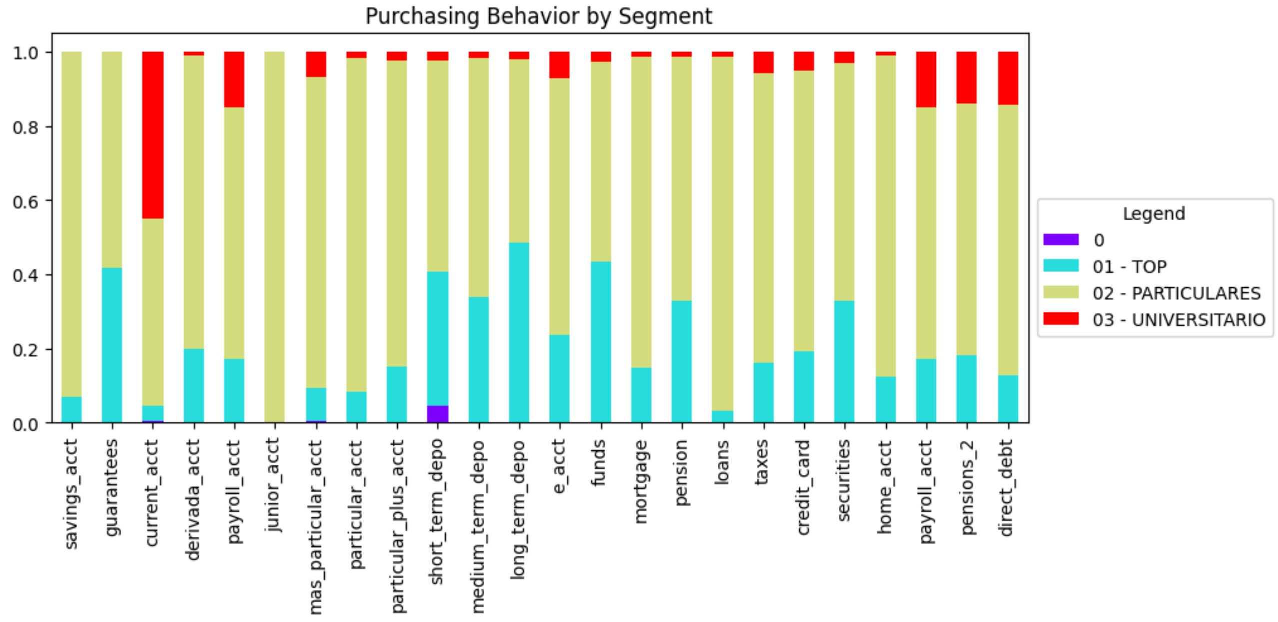
The dataset shows many inactive customers, though active customers outnumber inactive ones for all products. Savings accounts, current accounts, and particular accounts have the highest numbers of inactive customers, highlighting potential areas for improved customer retention strategies.

Income column analysis:

The median income of customers across all products is very similar, indicating a consistent economic profile among the user base. The gross household income for the region of Gipuzkoa stands out as the highest.

Segmentation column analysis:





The majority of customers fall within segment #2, known as "Particulares." There is a clear correlation between the count of clients in each segment and the specific products they have purchased. Notably, all customers who have acquired a junior account belong to segment #2.

Provide any insights you may have gained as a result of the EDA. What does the EDA tell you about the dataset and your problem statement? What insights can you share using your EDA about the opportunities or challenges in solving the problem statement as defined in Week 1

The EDA helped us identify some products that are more aligned with a specific kind of customer, such as: junior accounts and people under the age of 20; new customers, who are more likely to get short term deposit and checking accounts while older customers are more likely to get other products such as mortgage savings, and home account.
This helps us to understand patterns of what kind of clients we can expect to buy certain products, it is definitely not enough to predict what products they will buy, since it is hard to account for all the variables at once, but we can now have expectations based on what we have seen happening. For instance:

- The bank has more customers in the age group 20-30, however they are more concentrated in one product, such as current account, while the age groups 30-40 and 40-50 are dominant in the majority of other products offered by the bank, so we know that the younger clients are not going to purchase a lot of products, which can show an opportunity for the bank.
- The most used products for new customers are short term deposit accounts and the "mas particular" account, which could indicate that people are joining the bank with their eyes on this specific account.

Overall, we have a relatively younger population in the dataset, which could imply that the data is skewed toward younger adults. Additionally, we can now also focus on the impact of other variables that we were previously not sure how it related to both the products bought and the amount of purchases, such as the join channel column. The challenge now is to determine how much impact these variables have on the final objective of predicting which products the client will buy.

What data problem(s) did you identify? What recommendation(s) do you have, based on your EDA, in terms of data pre-processing that you will need to do?

We removed some rows and columns in this phase of the project:

- We dropped the column primary customer and kept the last date as primary customer column since we can have all the information in one column -customers that do not have a date are still primary
- We dropped rows where deceased was true. Deceased clients have few products, and they make up 0.2% of the database. However, deceased clients are not going to buy any more products, which is what we are trying to predict, so we will drop these rows and drop the column
- We removed the "address type" column since all the values are the same, meaning it doesn't provide any useful variation
- We dropped "province code" because we already have the "province name," which makes the code redundant.
- We removed "employee spouse" column due to the high number of null values, with over 10 million missing entries, making it less informative for the analysis.
- The income value was missing from over 2 million clients of the bank - these values will either be dropped or filled in with the median of the client's province

We dropped a few rows on some columns where we believe that the value was a filling value or an error, such as:

- Seniority_in_months: dropped rows where seniority was -999999
- Province_name: dropped the rows where province name = others
- Age: dropped rows where age was 0 or higher than 100

Seniority in months vs first contracted date: We had the impression these two columns were giving us the same information and we decided to check the correlation. Turns out it is highly correlated, so we'll keep the column seniority in months. Before we deleted the first contract date column, we wanted to extract any information we might need from it as discussed earlier.

Next week, for the pre-processing phase, we will deal with our categorical values and use a label encoder so we can fit those in the model better. Once we have those encoded, we will run a correlation heatmap to make sure the variables we are keeping are not highly correlated, which could affect the performance of our model.