**Week 4 — Make Data Model Ready**

**Aritra Ray**

**Maria Alice Fagundes Vieira**

To avoid data leakage, we split our data into separate files: one for the training dataset, one for validation, and one for testing. This approach ensures that we preprocess data only on the training set first. We then replicate the preprocessing code on the validation and test datasets to maintain consistency with the training set. This method prevents the model from accessing validation and test data, hence, eliminating the risk of data leakage.

For the preprocessing, we started by making sure our dataset had the correct types, since we exported the train set to a new file, we ended up losing some of the types we had previously set.

For the data preprocessing, the first problems we identified were:

- **Non-primary and deceased clients**: These represent only 0.2% of the dataset, creating imbalance and offering little value for predicting future product purchases. Since deceased clients cannot purchase additional products, we dropped these rows and removed the related column. As we are focusing only on primary and living customers, both columns were unnecessary and were thus removed.
- **last_date_primary column**: Given that we are working only with primary customers, this column is no longer useful and was dropped.
- We removed rows where **seniority_in_months** was -999999 and null values under **province_name** (previously filled with 'other') as these variables introduced noise.
- We will drop columns where age is 0 and over 100. We believe these clients are not going to be valuable on our model
- We had the impression the columns **seniority_in_months** and **first_contract_date** 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.

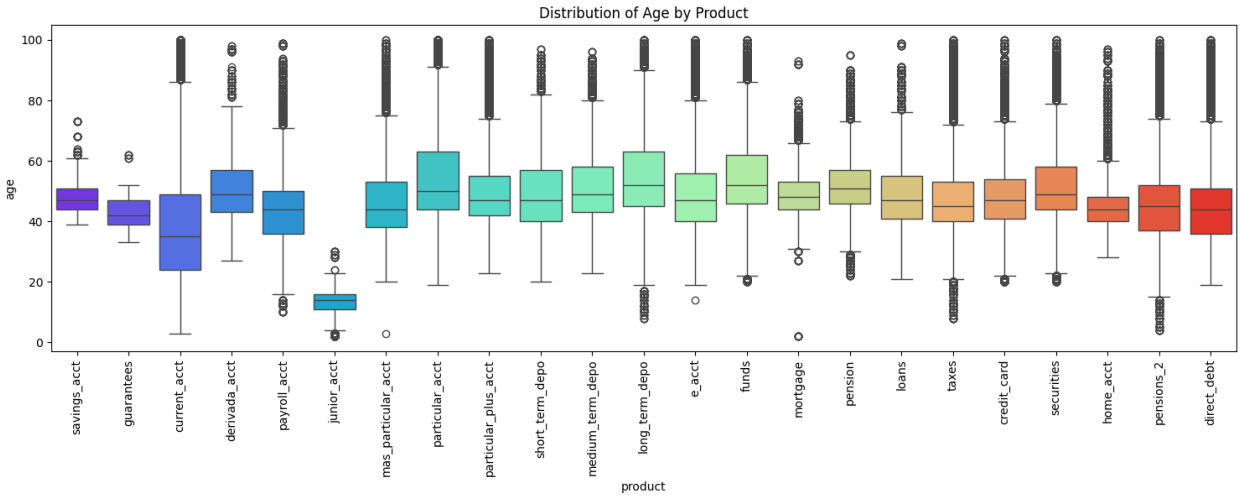
| | seniority_in_months | first_contract_date |
|---|---|---|
| seniority_in_months | 1.000000 | -0.966175 |
| first_contract_date | -0.966175 | 1.000000 |

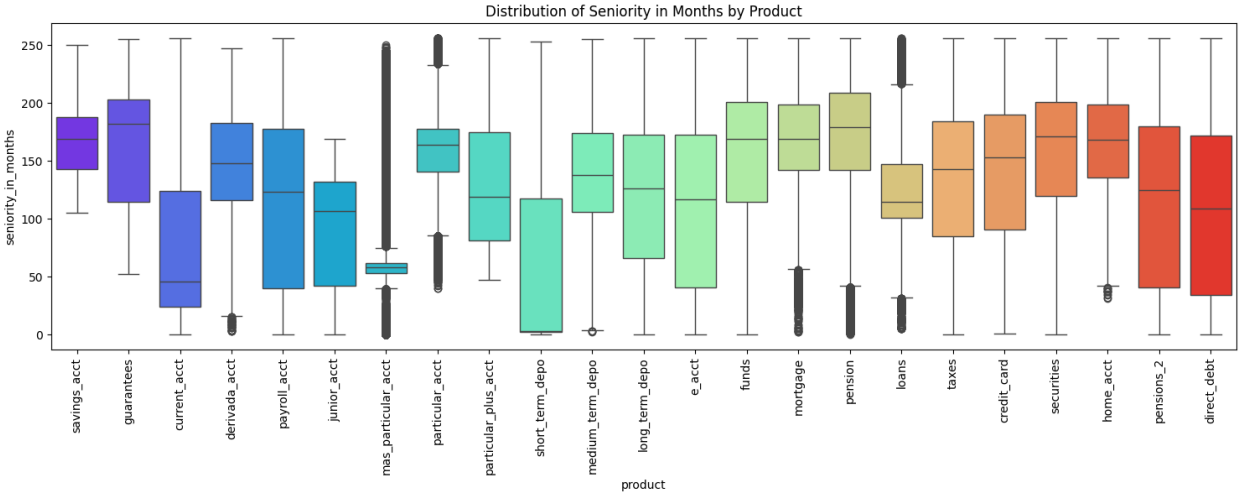
Handling Missing Income Data:

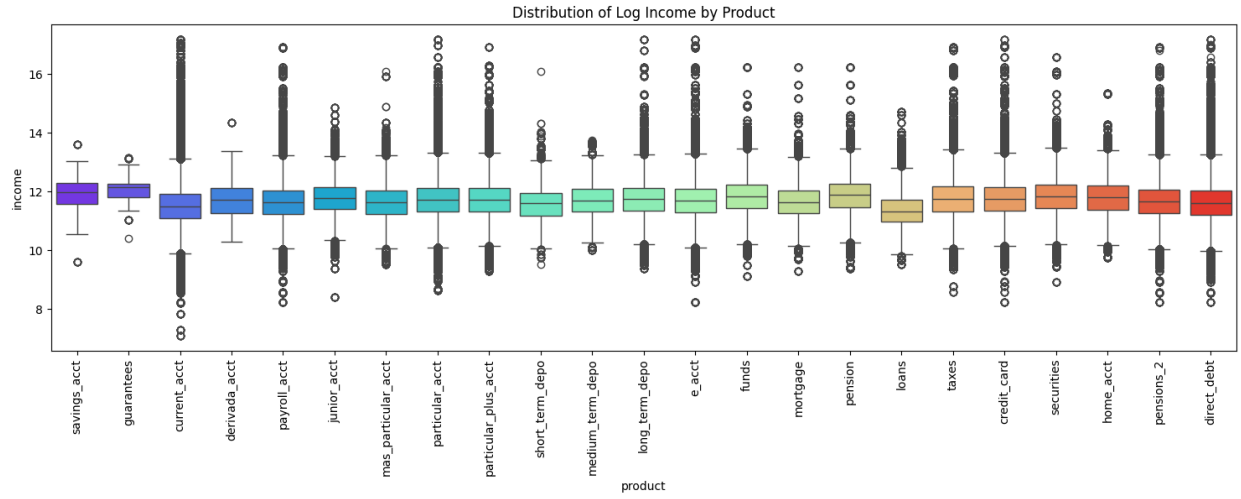
- A significant portion of the income values was missing. We considered imputing these values using the median income per province but found no strong correlation between income and province.
- After further analysis, we decided to drop rows where income was 0 and the customer had no products. Despite this, 16% of income data remained missing, and we ultimately decided to drop these values due to a lack of reliable imputation options.

We analyzed the correlation between province and income and there was no strong correlation. So, imputing values w.r.t province might not be a good idea. Since we lack any good sources to impute the data, we decided to drop the values.

After treating missing data, we proceeded to analyze outliers in **age, seniority_in_months** and **log_income**. Box plots for each product category offered by the bank helped us understand the distribution of these variables:



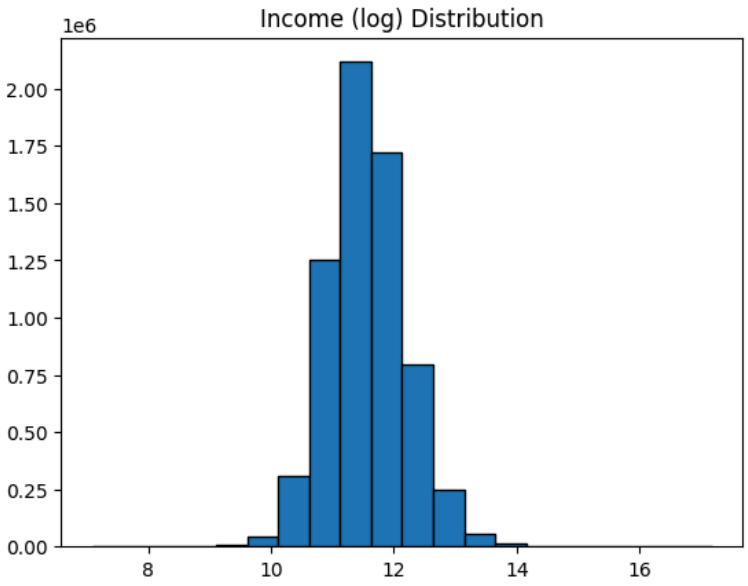




**Age**: Clear differences were observed between age groups subscribing to products like the junior account, where the median age was significantly lower than for other products.

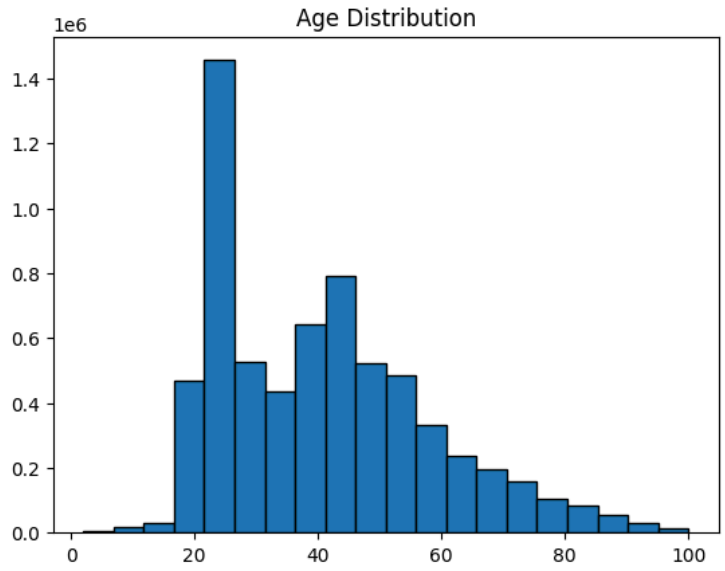
**Seniority in months**: Customers purchasing products like current, junior, mas, and short-term accounts had more recent purchases, while older customers tended to have products like mortgages, pensions, and derivada accounts.

**Income**: The median income of customers was similar across all product categories.

Next we plotted the income log distribution and age distribution to see if they were normally distributed.

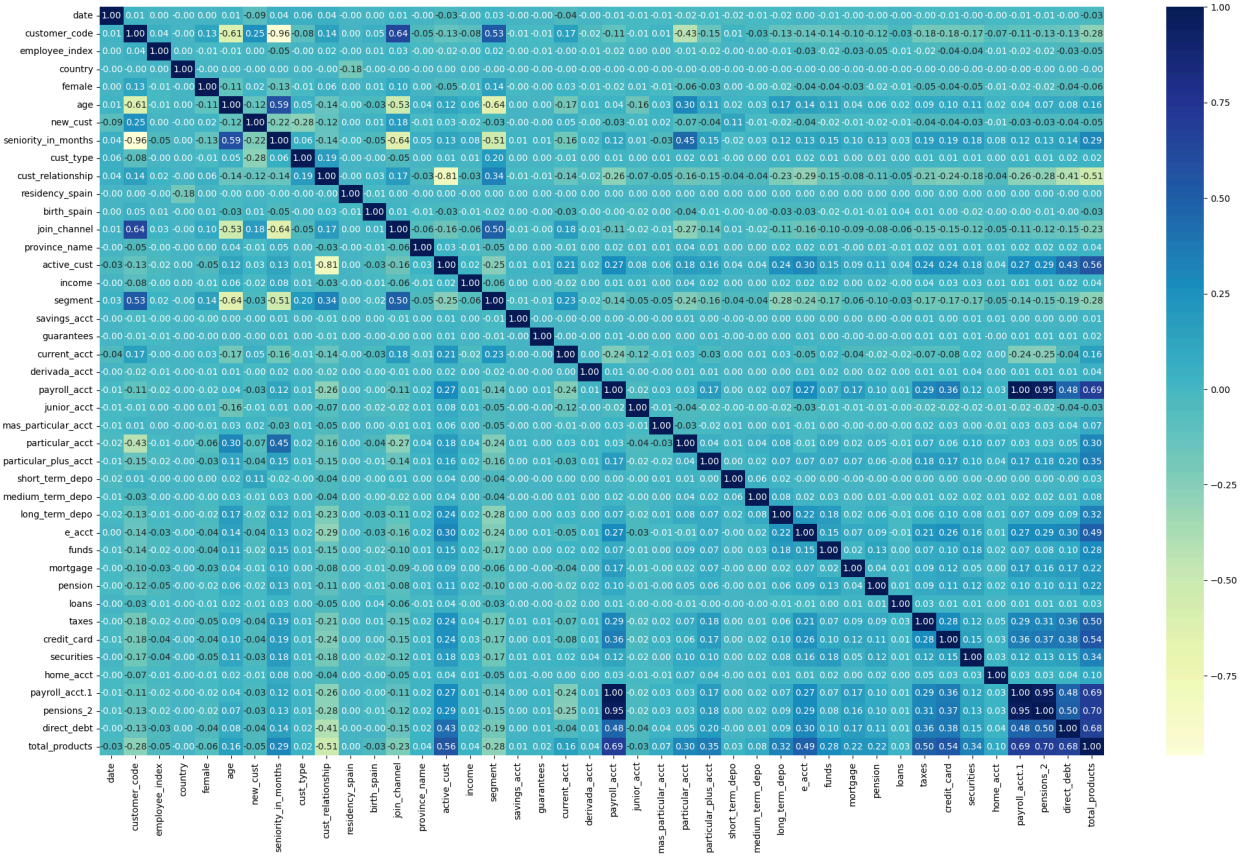



The histogram for log income appears to show a distribution that is approximately normal. It is symmetric and bell-shaped, with most of the data concentrated around the center. Since the distribution is of log-transformed income data, this might imply that the original data was right-skewed, and the log transformation helped to make it more normal-like. In this case, this distribution may be considered approximately normal in log-space.




For the age distribution histogram, the graph shows a right-skewed distribution with a peak around age 20. The tail extends more towards the older ages, which is not typical of a normal distribution. A normal distribution would have a single peak near the center with symmetrical tails on both sides, but this plot shows a clear imbalance. The large spike at age 20 suggests a sharp peak, which deviates from the smooth, bell-shaped curve of a normal distribution. Overall, this is not normally distributed and shows a skewed pattern typical of age distributions, where younger people tend to outnumber older individuals.

Next, we conducted a correlation analysis. Categorical variables were label-encoded temporarily to detect correlations. This transformation was not final, as it was used solely for correlation analysis.



There are two correlations that draw attention to us. The first one is active_cust and cust_relationship, which is 0.81. The second one is payroll_acct and payroll_acct.1 which is 1. They will both be dropped, the payroll_acct.1 gives us the same information as the payroll_acct and we believe that having such a high correlation between active_cust and cust_relationship would hinder our model.

Categorical Data Transformation:

We transformed the categorical data into numerical formats:

- **segment column**: We used one-hot encoding to create binary columns for the variables.
- **join_channel, province_name, and employee_index**: Target encoding was applied to avoid dimensionality issues, as using one-hot encoding would result in too many columns. We chose total_products as the target variable for the encoding since it represents the number of products purchased by each customer.
- **cust_type**: Label encoding was applied since most variables were already in ordinal order. "P" was replaced with 5.

Normalizing and Standardizing Data:

- We normalized the data for **age** and **seniority_in_months** since they were not normally distributed.
- We standardized **income** because its log-transformed distribution was approximately normal.