DESCRIPTION OF VARIABLES

account_id : This variable is the ID key of the final Table, simply the ID of each account owned by a client present in the final Table. This is an integer Variable. Each line of the Table has an unique value in this variable.

bank_district_id : This variable is the ID of the bank branch location. This is an integer Variable. There are 77 different IDs of this variable present in the final Table.

date : This variable is the date in which each client has created his account. This is an Integer variable.

year : This variable is the year corresponding to the year of account creation for each client present in the final table. This is an integer variable. The variable has values from 1993 to 1995.

LOR : This Variable indicates the number of years the account has existed. This is an integer variable.

disp_id : This variable shows the disposition to an account. This is an integer variable

client_id : This is a variable identifying each client present in the final table. There are 692 different clients present in the table. This is an integer variable.

client_district_id : This variable is an ID for each client address. There are 74 different districts in the final table. This is an Integer variable.

gender : This variable indicates the gender corresponding to each client. 'F' indicates Female, 'M' indicates Male. The type of this variable is object.

age : This variable indicates the age of each client present in the final table. This is an integer variable

age_group : This variable indicates the group age corresponding to the client. This is an integer variable

total_credit96 : This variable indicates the total credit amount calculated from all the transactions made in 1996 per each client. This is a float variable.

total_withdrawal96 : This variable indicates the total withdrawal amount calculated from all the transactions made in 1996 per each client. This is a float variable.

RFM_score : This variable indicates the RFM score for each client present in the final table, This variable indicates the loyalty level of each client on the computed Recency, Frequency and Monetary criteria.This is an integer variable.

total_credit97 : This variable indicates the total credit amount calculated from all the transactions made in 1997 per each client. This is a float variable.

total_withdrawal97 : This variable indicates the total withdrawal amount calculated from all the transactions made in 1996 per each client. This is a float variable.

had_loan_1997 : This variable is one of the two target variables present in the final table. It indicates if the corresponding client had a loan. This is an integer variable

had_card_issued_1997 : This variable is one of the two target variables present in the final table. It indicates if the corresponding client had a card issue. This is an integer variable

Data correction and transformation

Handling Missing Data:

- **For Loans:**
  - Loans issued in 1997 were filtered using date conversion
  - Missing values for loans were handled by assigning a default value of 0 for accounts without loans in 1997.
- **For Credit Cards:**
  - The credit card issuance dates were converted into datetime format.
  - Missing credit card issuance statuses were imputed with 0 for accounts with no card issued in 1997.

Categorical Variables:

- Binary flags were added for:
  - "had_loan_1997" (1 for loans issued in 1997, 0 otherwise).
  - "had_card_issued_1997" (1 for credit cards issued in 1997, 0 otherwise).

Feature Engineering:

- Derived features like "total_credit96" and "total_withdrawal96" were provided, representing financial metrics for the year 1996.
- Age grouping was implemented by binning the age variable into decades (e.g. 20, 30, 40).

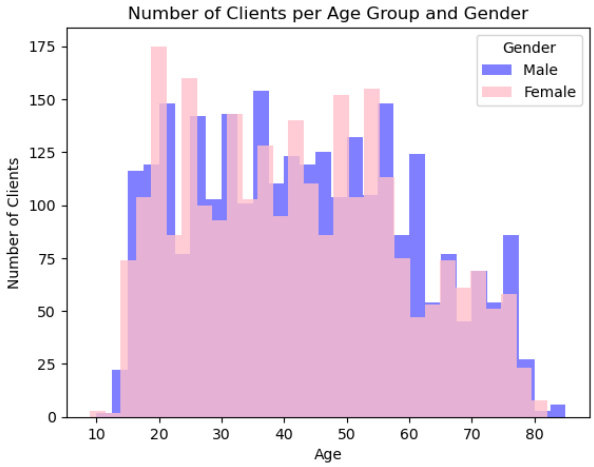
Outlier Treatment:

- Potential outliers in financial data such as "total_credit96", "total_withdrawal96", were identified through box plots or scatter plots.

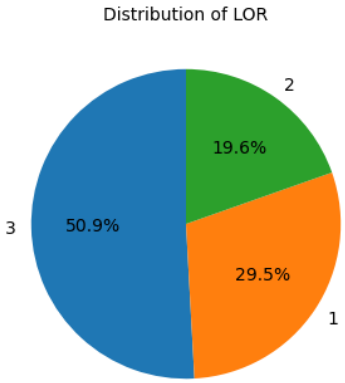
Merging Data:

- Several tables (e.g. loan and card data) were merged in the process to make the base table to consolidate all relevant features into one dataset.

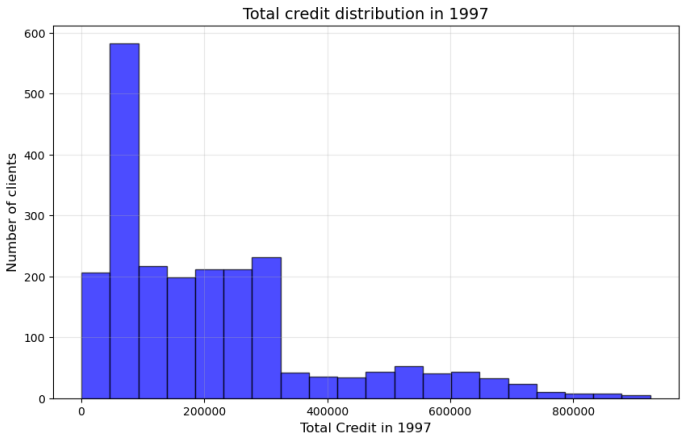
Analyze and visualize the independent variables and dependent variables



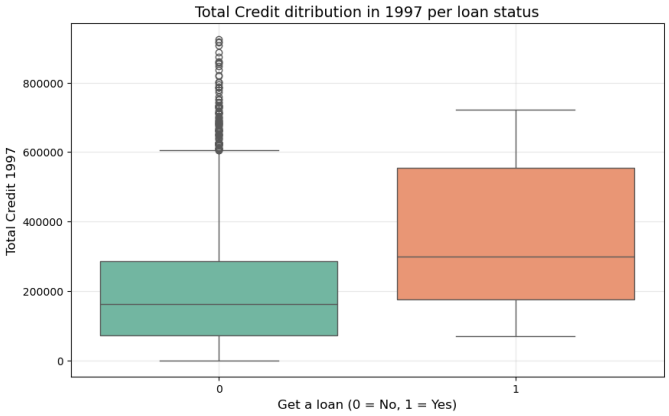
This visualization shows the number of clients from each gender distribution per. age group. Most women are aged 20, while most men are in the late 30.



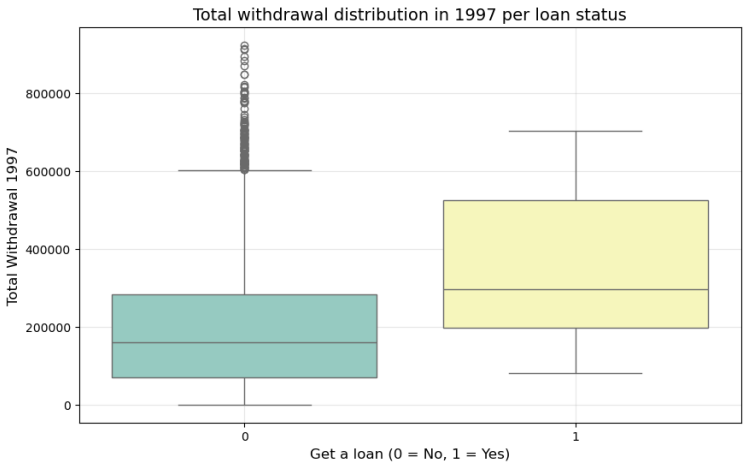
This Pie Chart shows the distribution of the clients relationship length (1, 2, or 3 years).



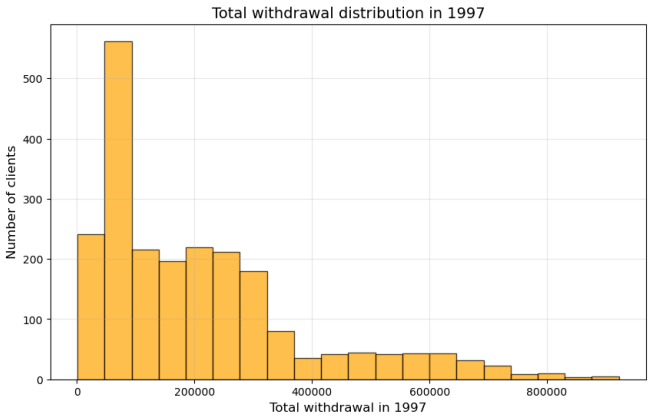
This Histogram shows the distribution of clients per total credit in 1997. Highest amount of clients have a credit of 40.000 to 80.000.



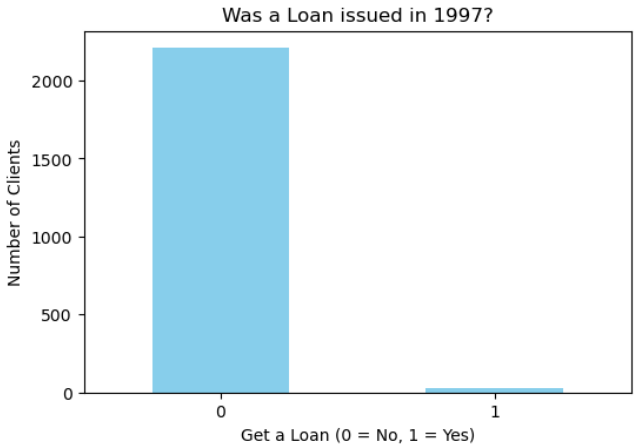
This BoxPlot shows the total Credit distribution in 1997 transactions per loan status. People with a loan has a higher average, with the lowest and largest values also being higher.



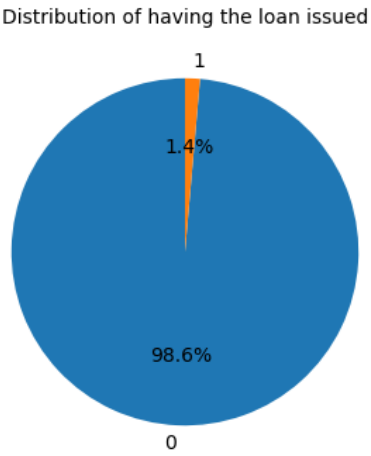
This boxplot shows the distribution of withdrawal amount based on the clients loan status in 1997. People with a loan has a higher average, with the lowest and largest values also being higher.



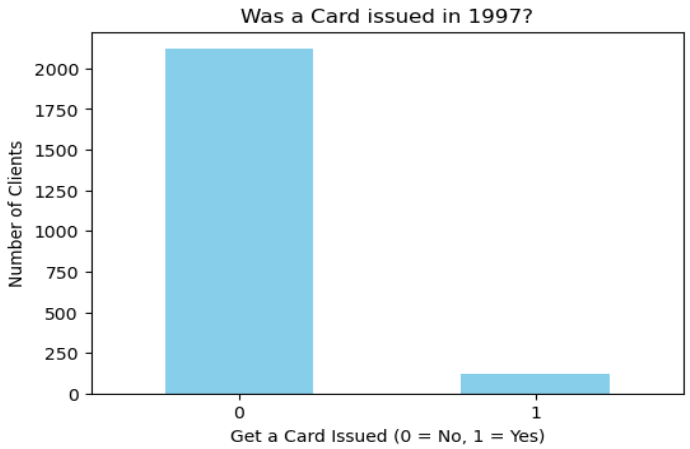
The bar chart visualize the total withdrawal amount and customer. The largest amount of clients have a withdrawal amount between 40.000 and 80.000 in 1997.



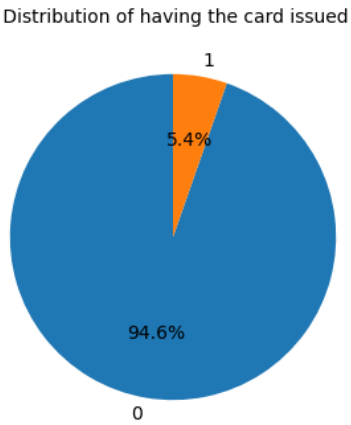
This Bar Chart shows the number of clients who got a loan issued in 1997.



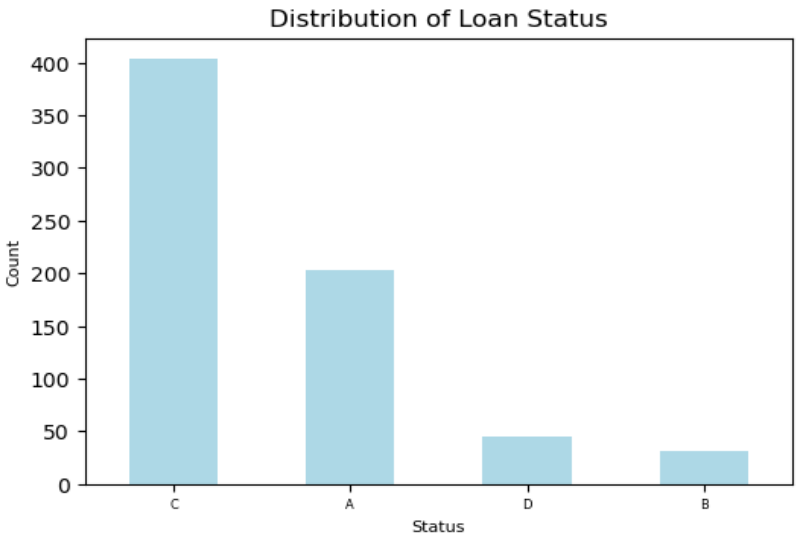
This Pie Chart shows the distribution of clients who get a loan issued in 1997.



This Bar Chart shows the number of clients who got a card issued in 1997.



This Pie Chart shows the distribution of clients who get a card issued in 1997.



This Bar Chart shows the number of clients per Loan Status.
'A' stands for contract finished, no problems,
'B' stands for contract finished, loan not payed,
'C' stands for running contract, OK so far,
'D' stands for running contract, client in debt.

Analysis summary

The analysis of the dataset provides insights into client demographic, financial behavior and the issuance card and loan trends in 1997.

Most clients have held their account for three years, followed by accounts held in one year, where 2 year holders are fewest. Despite 1 year holders being second in amount, the distribution still indicates a stable and loyal customer base, with over 70% being 2 or 3 year holders.

Gender is distributed relatively balanced in the dataset, with grouping age into groups of decades, showing most of the bank's clients being from 20 to 60 years old. The demographic analysis helps the bank tailor financial products in line with customer ages.

The RFM scores contribute to identifying the most loyal clients, who often and recently made transactions, which are valuable insight for engagement targets, because they represent possibilities for cross-selling or loyalty programs.

Loan and card issuance in 1997 were selective, with a minority of clients receiving loans or cards. Financial metrics show that clients with loans had higher credit and withdrawal activity, suggesting greater liquidity needs. The positively skewed distribution of total credit highlights a small group with significantly higher financial activity.

Overall the findings reveal a diverse and active client base, with notable trends in financial behavior and product engagement. Higher financial activity, including credits and withdrawals, possibly might show a significant correlation with loan issuance, which makes a key insight for future predictions.

To better understand and create services for the bank’s clientele, demographic and loyalty metrics can help inform strategic decisions, enhance customer engagement and increase the quality of the financial services the bank delivers.