IESEG
SCHOOL OF MANAGEMENT

DECEMBER 2021

BERKA BANK CHALLENGE PROJECT

AS A REQUIREMENT FOR FINANCIAL PROGRAMMING COURSE

PRESENTED BY

ENITA OMUVWIE
PEDRO ROMANINFANTE

Table of Contents

INTRODUCTION 2
PROJECT OVERVIEW 2
PROBLEM DESCRIPTION 2
DATA EXPLORATION 2
DATA MART PREPARATION 3
THE VISUALIZATION AND INTERPRETATION OF DATA 4
CLIENT BY AGE GROUP AND GENDER 4
ACCOUNTS GROUPED BY YEAR 5
ACCOUNTS GROUPED BY MONTH 5
ACCOUNTS GROUPED BY THE FREQUENCY OF STATEMENTS 5
NUMBER OF ACCOUNTS GROUPED BY DISTRICT NAME AND AVERAGE SALARY 6
NUMBER OF ACCOUNTS GROUPED BY REGION AND AVERAGE SALARY 7
GROUPING TRANSACTION AMOUNT BY PAYMENT CHARACTERIZATION 7
ACCOUNTS GROUPED BY GENDER AND AGE GROUP 8
NUMBER OF LOANS GROUPED BY STATUS 9
LOAN GROUPING BY LOAN DURATION 9
ORGANIZING THE NUMBER OF LOAN STATUSES BY AGE GROUP 10
CLASSIFYING LOAN STATUS BY GENDER 10
GROUPING LOANS BY LENGTH OF RELATIONSHIP 11
GROUPING TRANSACTIONS BY MONTH 12
GROUPING OF TRANSACTION TYPES 12
GROUPING TRANSACTION CHARACTERIZATION BY 1996 13
RECENCY, FREQUENCY AND MONETARY VALUE 14
BASETABLE VARIABLES DESCRIPTION 15

INTRODUCTION

PROJECT OVERVIEW

The financial institution strives to enhance its services. Bank managers have only a hazy concept about who is a good customer (to whom they should offer significant solutions) or who is a terrible customer (whom to watch carefully to minimize the bank loses).

Certainly, the financial institution file contains information about their customers, their account holders (money transfers over a specific length of time), the loan repayments those that have already granted, and the bank cards they have approved. The bank's executives hope to gain a better understanding of customer and to take concrete steps to improve services. They will not be persuaded by a simple application of a verification tool.

PROBLEM DESCRIPTION

Create the base table using the following steps below:

1. Extract information from the financial data set and create a data science base table with the following criteria:

- Rows: the granularity (lowest level of observation) is client who is account owner.
- Columns: independent and dependent variables following the timeline. The time window of the independent variables: 1996 (1 year).
- The time window of the dependent variables: 1997 (1 year).

2. Create the independent variables (e.g. gender, age, age group, RFM, LOR, etc.):

- Should be calculated only for the independent variables time window (i.e. 1996).
- Should be calculated only for clients having sufficient information in the independent variables time window (i.e. 1996).

3. Create the dependent variables (or target variables): o Target variable #1: Client had granted loan in the dependent variables time window (i.e. 1997), binary value (0 = did not have granted loan, 1 = had granted loan).

- Target variable #2: Client had credit card issued (for both account owner and disponent) in the dependent variables time window (i.e. 1997), binary value (0 = did not have credit card issued, 1 = had credit card issued)

DATA EXPLORATION

This is a financial dataset containing eight relational tables of data. The following tables are to be used for loan classification. Below are the listed tables:

- **Account table:** It contains 4500 objects with no null or missing values and provides the characteristics that are constant
- **Client table:** A customer's characteristics are described in the record. This record consists of 5369 objects.
- **Disposition table:** This table contains 5369 objects in the file. It connects a client to an account demonstrating a client's ability to operate an account
- **Permanent Order table:** It has 6471 objects and describes the features of a payment order.
- **Transaction table:** This table has 1056320 objects and is a description of a single transaction on an account.
- **Loan table:** It describes a loan made for a specific account and has 682 objects.

- **Credit Card table:** It has 892 objects and refers to a credit card issued to a specific account.
- **Demographic table:** It contains 77 objects that describe a district's demographic characteristics.

For us to create a data mart table we combined data from the tables above. Certain unique variables were selected from the tables to construct the data mart table.

DATA MART PREPARATION

1. We used the Pandas, NumPy, and Matplotlib libraries for data analysis and visualization. All tables were derived from the presented datasets, and new variables were generated from all these existing features.
2. There were new client table features such as birth year, birth month, birth day, gender, age, and age group. Following that, the client number was removed from the table after new features have been derived from it.
3. Using a dictionary of new variable names, the account table was cleaned up and the frequency column was renamed. Account year and account month are two new features that have been added.
4. Some columns in the credit card table were renamed, and the cards were further classified by issue year and type before being merged to the disposition table.
5. Using a dictionary, all district columns were renamed to match old column names with new data frame, and the total number of accounts were counted by district, total account by region, average salary per region and merged to the disposition table. Because they are the least important features, three columns were removed from the table.
6. We proceeded by renaming the k symbol column in the order table, and afterwards we set the minimum and maximum limits for the amounts by the k symbol. Three columns were removed from the district table and merged. A pivot table with k symbol and the cumulative sum of account ids, as well as other graphs, had been generated.
7. We considered several loans that were completed prior to 1996. Graphs were created by combining loan id, duration, and status. We calculated the total amount of loans based on age and status to create some detailed graph.
8. The k symbol, type, and operation columns in the transaction table were transformed into new variable names, and the cumulative sum of each of these types was calculated. The transaction table was used to derive aggregated functions such as mean, min, and max. We also created some graphs to show the changes in the dependent variable based on k symbols and types.
9. 9. The RFM was computed and derived from the transaction and account tables, and additional analysis was performed.
10. Throughout the data preparation process, all tables were carefully inspected for null values, and measures were implemented to make sure there were no other null values in the current data set.
11. Following the completion of the preparation, the dependent variables were chosen, and the base table was created with a containing 77 new variables and 2239 observations.

	card_issue_year	credit_card	lor	amount_Household	amount_Insurance	amount_Leasing	amount_Loan	amount_Unknown	order_count	total_amount	...
0	0	0	3	3662.0	0.0	0.0	0.0	0.0	1.0	3662.0	...
1	0	0	3	3596.0	4065.0	0.0	0.0	1474.0	3.0	9135.0	...
2	0	0	3	2141.0	0.0	0.0	0.0	1197.0	2.0	3338.0	...
3	0	0	3	9612.0	0.0	0.0	0.0	0.0	1.0	9612.0	...
4	0	0	3	2042.0	2300.0	0.0	0.0	41.0	3.0	4383.0	...
...	...	...	...	...	...	...	...	...	...	...	...
2234	0	0	1	0.0	0.0	0.0	2770.3	0.0	1.0	2770.3	...
2235	0	0	1	0.0	0.0	0.0	0.0	0.0	NaN	NaN	...
2236	0	0	1	0.0	0.0	0.0	0.0	0.0	NaN	NaN	...
2237	0	0	1	0.0	0.0	0.0	7876.0	0.0	1.0	7876.0	...
2238	0	0	1	4418.0	0.0	0.0	0.0	0.0	1.0	4418.0	...

2239 rows × 77 columns

Table 1: Showing Base table with new features

THE VISUALIZATION AND INTERPRETATION OF DATA

CLIENT BY AGE GROUP AND GENDER

This graph depicts clients who have been classified based on their age and gender. According to the graph, the age ranges 20 to 50 have the highest number of bank clients, whereas the age group 80 does have the lowest number of bank clients. Other age groups have also made a significant contribution to the bank's customer base. Other age groups also have contributed to the bank's customer base. In the client base, the age group 20 had most females, while the age group 30 had far more males with over 500 clients respectively.

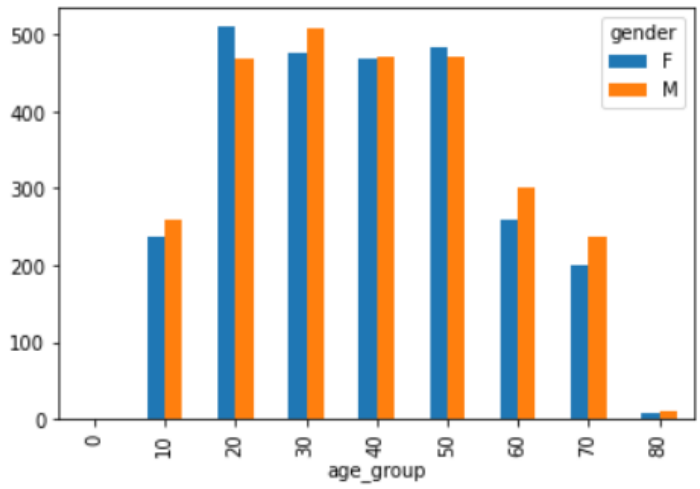

Figure 1: Grouping Clients by Age Group and Gender

ACCOUNTS GROUPED BY YEAR

According to the graph, the highest number of accounts was recorded in 1993, with over 1000 accounts, and the lowest number of accounts was recorded in 1994, with approximately 400 accounts. There is a decent amount of account holders in 1995, with approximately 600 accounts during that year.

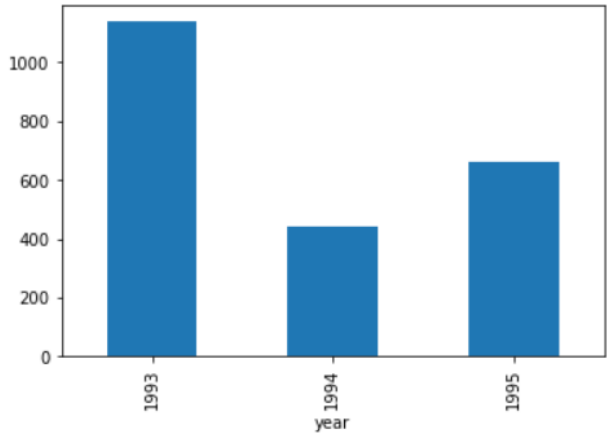

Figure 2: Grouping Accounts by Year

ACCOUNTS GROUPED BY MONTH

This graph depicts the cumulative variation of accounts in the different months of the year. The month of November saw the most account openings, with over 200 new accounts opened at the bank. Other accounts contributed at varying degrees of account openings, with no month having fewer than 100 accounts on average.

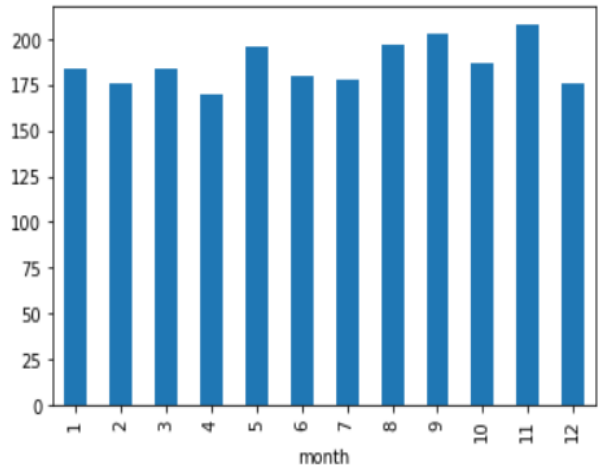

Figure 3: Grouping Accounts by Months

ACCOUNTS GROUPED BY THE FREQUENCY OF STATEMENTS

This pie chart depicts the frequency of accounts, with monthly transactions accounting for 92.4 percent, weekly transactions accounting for 5.4 percent, and transactional transactions accounting for the smallest

percentage of 2.1 percent. As a result, the vast bulk of account transactions were carried out every month rather than on a weekly or transactional basis by account holders.

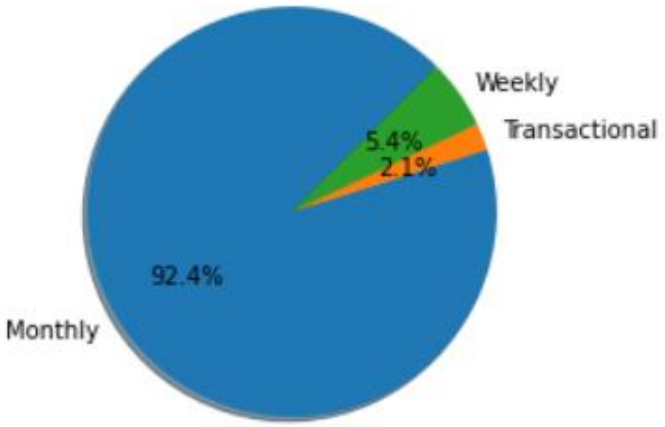

Figure 4: Grouping Accounts by Frequency

NUMBER OF ACCOUNTS GROUPED BY DISTRICT NAME AND AVERAGE SALARY

The chart illustrates a breakdown of various districts' bank branches, and the average salary earned by customer accounts as well as the number of accounts for each district. The redline represents the number of accounts with variations over each branch, and the bar plot represents the average salary earned in the districts. Hl.m. Praha had over 12000 average salaries and 250 accounts in their district, and the rapid decrease in the number of accounts distributed through other districts, while the average salary had shown a slow indication of decline across various districts.

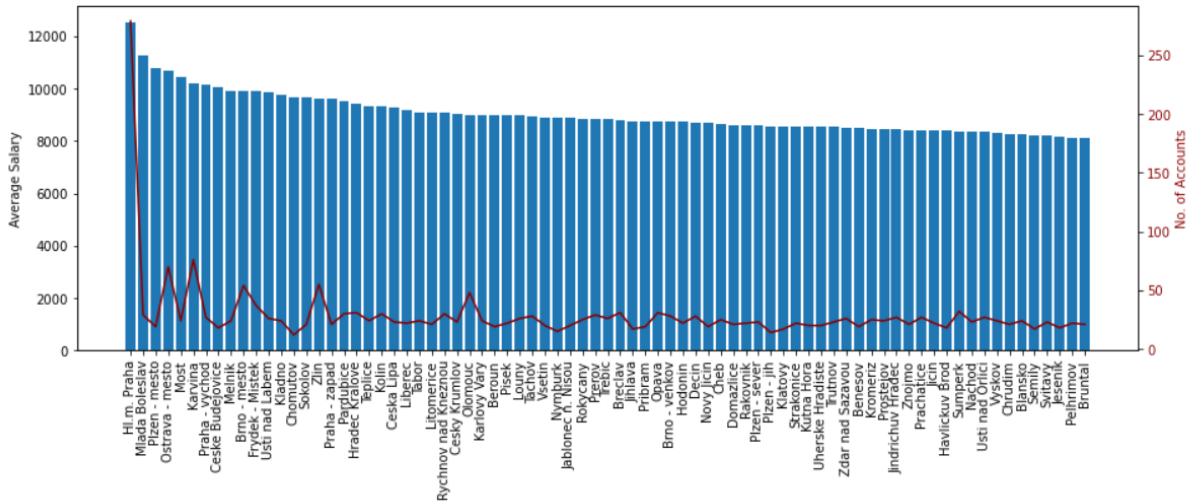

Figure 5: Grouping Accounts by District and Average Salary

NUMBER OF ACCOUNTS GROUPED BY REGION AND AVERAGE SALARY

This graph depicts the number of accounts and the average salary in each region. The red line counts the set of accounts, and the bar chart represents the average salary in each region. South Bohemia Region has the fewest accounts, with the point indicating less than 200 accounts and an average salary of more over 8000. North and South Monrovia had approximately 400 accounts in their respective regions, with an average salary of over 8000. The Prague region had the highest average salary of 12000, but only 300 accounts.

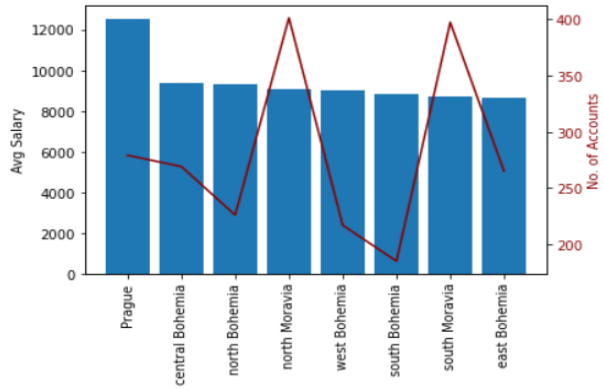

Figure 6: Grouping Accounts by Region and Average Salary

GROUPING TRANSACTION AMOUNT BY PAYMENT CHARACTERIZATION

The chart shows the various payment characteristics and the amounts associated with these qualities. Insurance and leasing had the lowest financial figures, whereas household had the highest financial move. Other payments that did not match into these categories and were categorized as unknown are also represented in the graph.

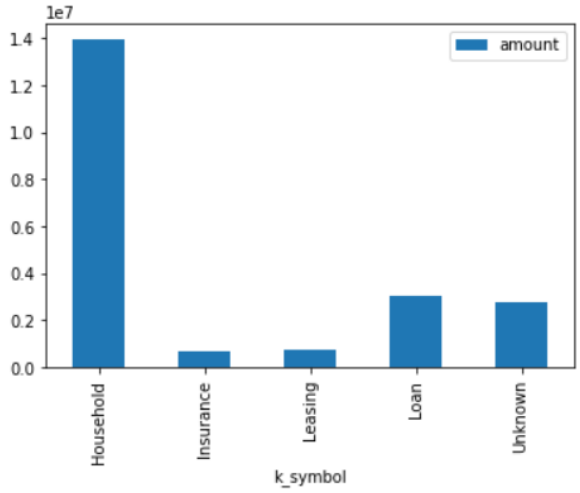

Figure 7:Grouping Payments by Characteristics

LOR GROUPED BY GENDER AND NUMBER OF ACCOUNTS

This depicts the number of accounts grouped by gender and the length of client relationship, which essentially indicates the lifecycle of account holders in the bank, or the number of years account holders have been in the bank for the period indicated in the timeline. We can say that clients of the same gender who have been in a relationship for three years have the most account holders, and males had about 600 account holders. However, in the one-year LOR, females had approximately 400 customer accounts than males.

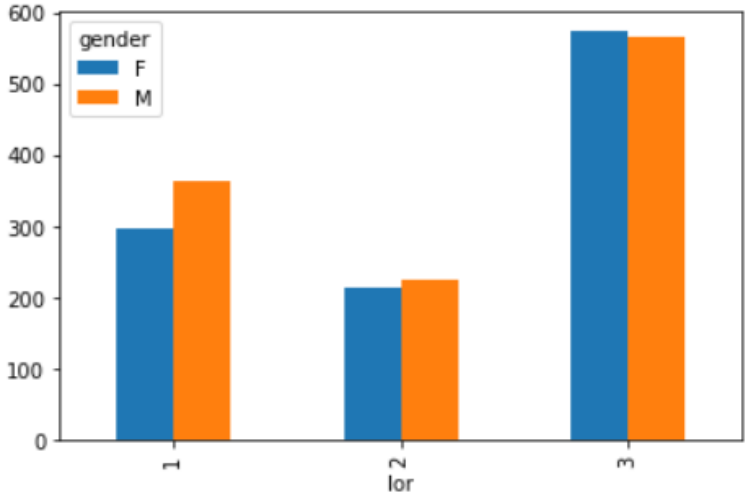

Figure 8:Grouping Accounts by Gender and LOR

ACCOUNTS GROUPED BY GENDER AND AGE GROUP

The number of accounts by gender and age group is depicted in this graph. The bars show that males in age groups 20 and 40 had the highest number of accounts, while females in age groups 10, 30, 50, 60, and 70 had the highest number of accounts with a variation in account count.

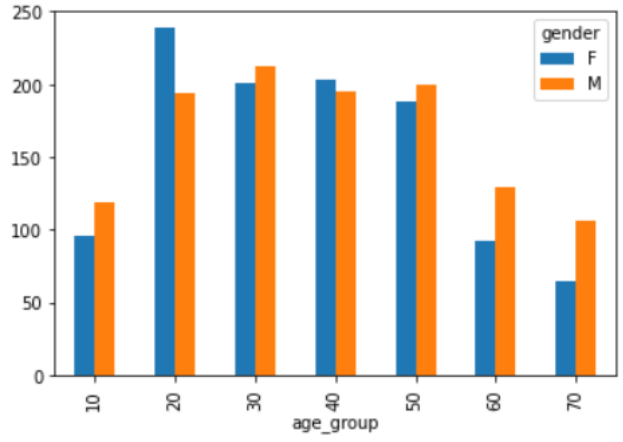

Figure 9:Grouping Number of Accounts by Gender and Age Group

NUMBER OF LOANS GROUPED BY STATUS

This chart depicts the number of loans by loan status, where A, B, C, D stand for Finished Contract, Loan Not Paid, Running Contract, and Client in Debt, respectively. C, which stands for Running contract, has the highest value, indicating that the bank has many running loan contracts with multiple account holders who seem to be able to meet their scheduled payments and are therefore not bad loans for the bank. A denotes loans that have been fully repaid to the bank, implying that many clients are creditworthy and possibly take out loan from a bank. There is a low percentage of loans which were not paid but completed the contract, and a percentage of bad loans, indicating that the client was in debt and has been unable to pay up on time.

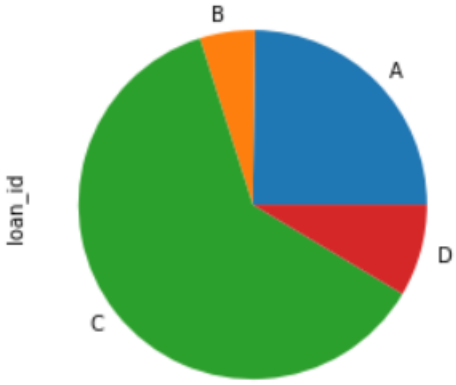

Figure 10:Grouping number of Loans by status

LOAN GROUPING BY LOAN DURATION

This graph displays the loan time duration grouped by the number of loans for each duration. We can see that long term loans of more than two years were highly approved by the bank, which could be due to the yielding interest, and the shortest period for any loan was one year, with the lowest number loans of about fifteen. 4 years had most loans with a total of about 30 which is still a long-term loan which could also be largely attributable to the customers' interest at the time of request. All these points are indicated in months in the graph below.

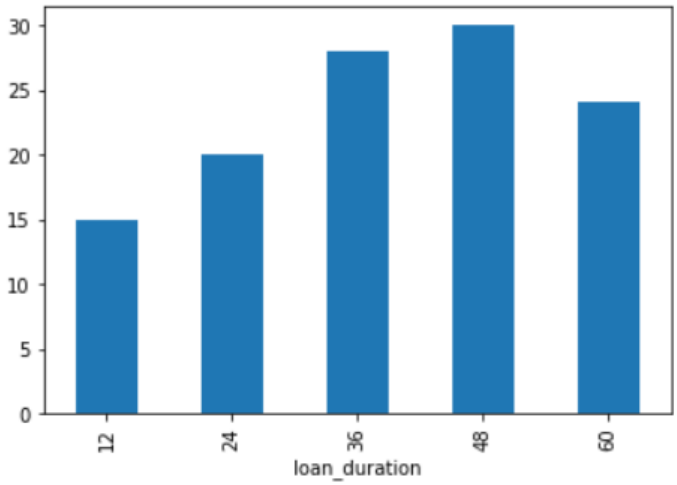

Figure 11: Grouping Loan Duration by Number of Loans

ORGANIZING THE NUMBER OF LOAN STATUSES BY AGE GROUP

We can see that C, which represents running contract, has the most age groups, with most of the lending being obtained by people in their twenties. The graph also shows that B, which indicates a completed contract but not yet paid, had the least among all age groups. The other two statuses are equitably distributed both across classifications.

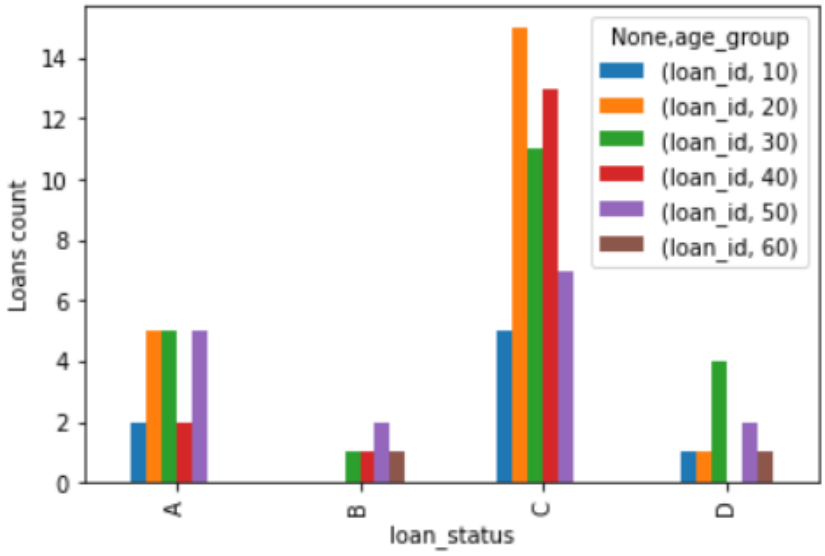

Figure 12:Grouping Loan Status by Age Group

CLASSIFYING LOAN STATUS BY GENDER

The graph below demonstrates the peak level on C (running contract) for both females and males, with males having taken out more loans in that category than females. However, in finished contact, which is

the first status, there have been more females than males who already had reimbursed their debts in overall percentage. Loan not paid in the second status, on the other hand, had the lowest number for both genders when categorized.

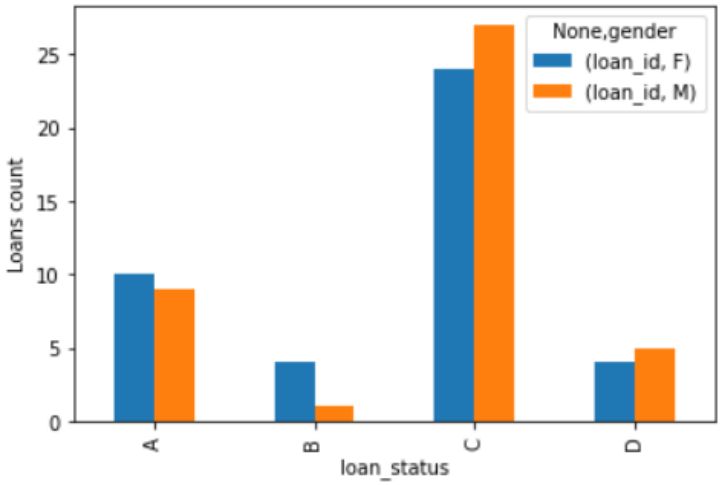

Figure 13:Loan Classification by Gender

GROUPING LOANS BY LENGTH OF RELATIONSHIP

As shown in the bar chart, this encapsulates the possibility that we do have more clients with a longer length of relationship in running contract status and the least number of clients in loan not paid. When compared to status B, which stands for loan not paid but contract completed. The other categories have relatively low points on the graph, but they are not the least once categorized on the graph points.

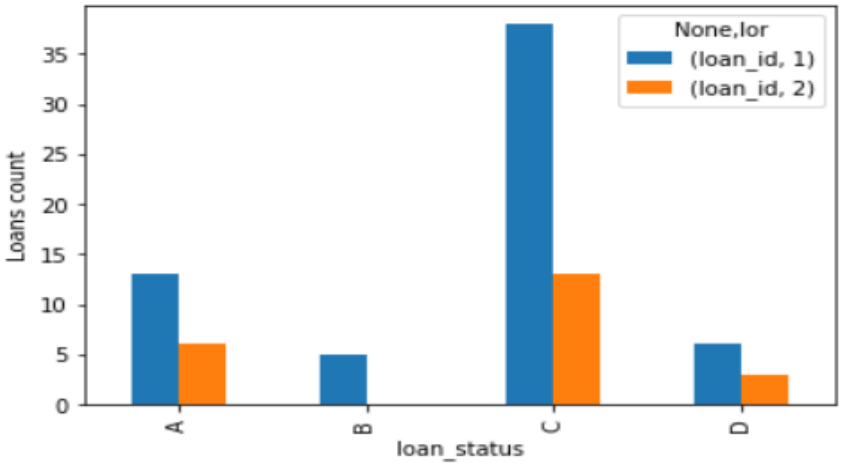

Figure 14:Grouping Loan by LOR

GROUPING TRANSACTIONS BY MONTH

We can see that in the graph below that credit already has its peak inflow in the sixth month, as did withdrawal, which is also recorded in the total. The red line represents the total number of transactions in the lump sum. All transactions apart from vyber which is card transaction and remained on the same point all around the period on the graph. All transactions increased in December, which is also the end of the banking year.

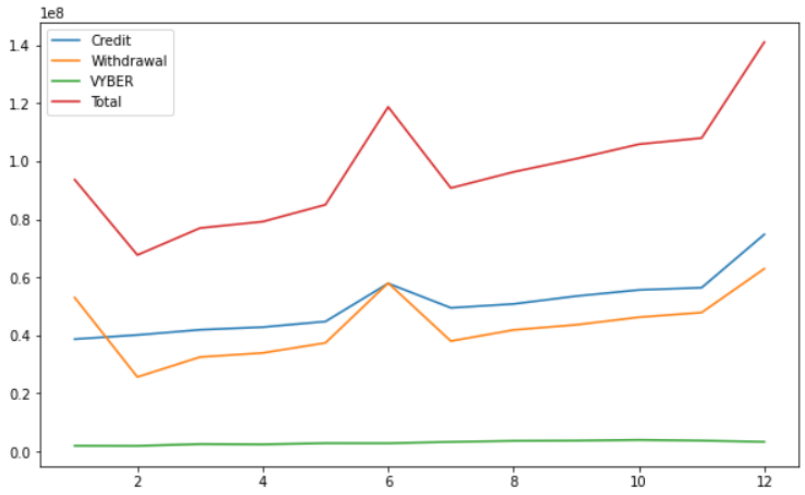

Figure 15:Grouping Transactions by Month

GROUPING OF TRANSACTION TYPES

The graph illustrates the different transaction types, which include credit, card, and withdrawal. As shown in the chart, withdrawal had the highest transaction rates of around 120000, while vyber, which stands for card, had the lowest number of transactions.

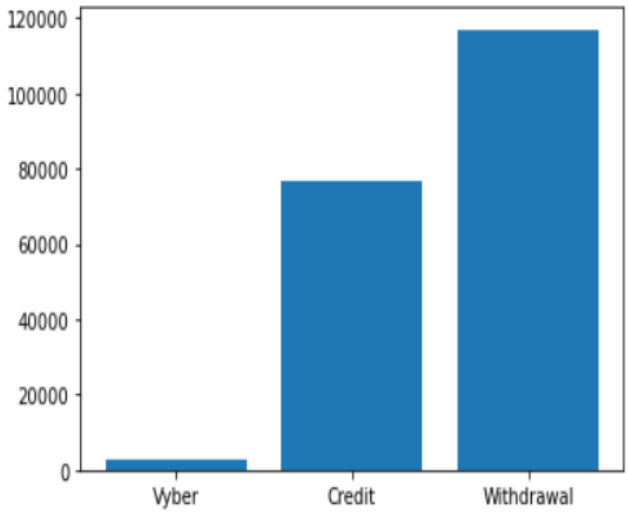

Figure 16:Grouping by Transaction Type

BANKING OPERATIONS GROUPED BY AMOUNT

This graph depicts the various operations as well as the amount generated by each operation. As shown, Cash Credit had the highest amount, followed by Cash Withdrawal. We could see that collections from other banks seemed to be slightly higher, that also simply refers to transactions done through the bank by other bank customers, when compared to remittances to other banks, which describes transfers done by bank customers to other customers. Credit card withdrawal was the operation with the smallest amount.

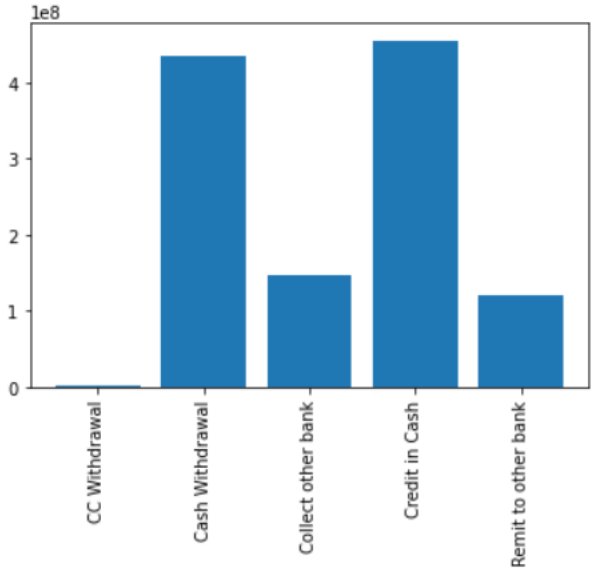

Figure 17:Grouping Amount by Operations

GROUPING TRANSACTION CHARACTERIZATION BY 1996

This graph depicts transaction characteristics. There were numerous unidentified transactions. Transactions with negative interest were the fewest. Interest, on the other hand, was the highest transaction in 1996.

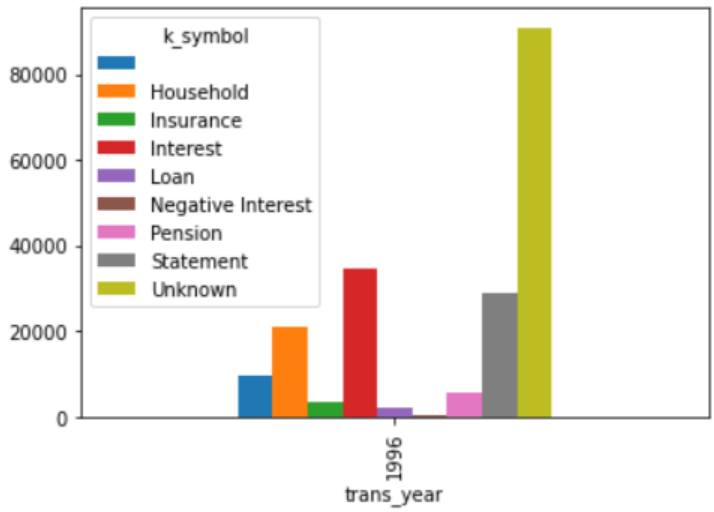

Figure 18:Grouping Transactions in 1996

TRANSACTIONS GROUPED BY AGE GROUPS

According to this graph, the age groups of 20 to 50 had the highest number between both transactions. We could see the age categories preceding and following these age groups, which could also be made up of younger developed people that have yet to understand the system and older generations who may not have been able to carry out these transactions on a regular basis.

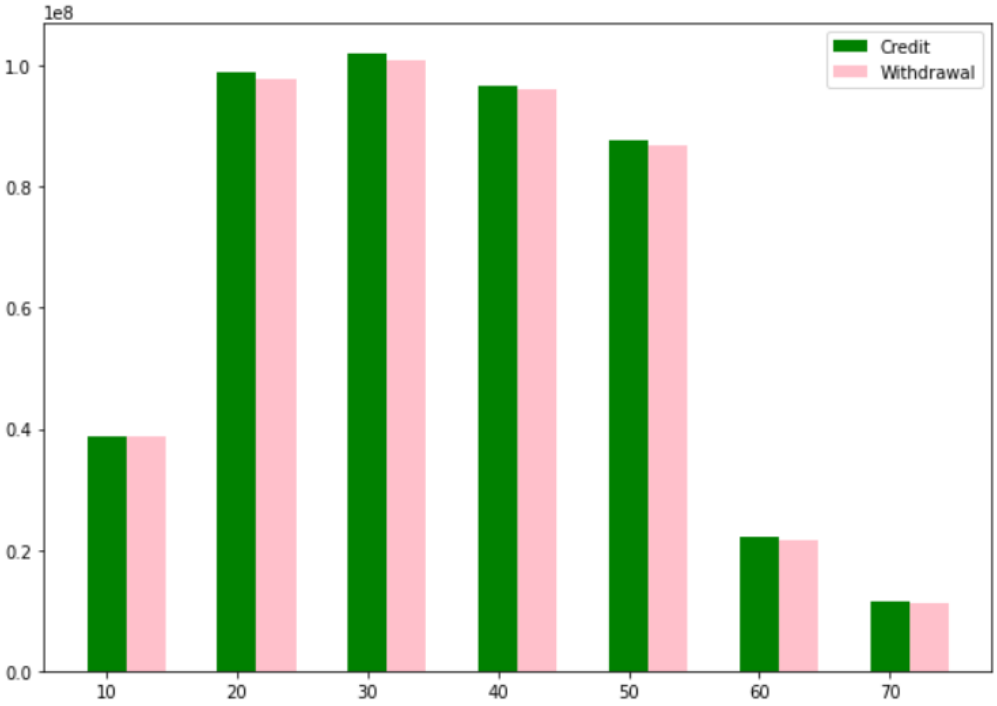

Figure 19:Grouping transactions by Age group

RECENCY, FREQUENCY AND MONETARY VALUE

RFM is an acronym for recency, frequency, and monetary value. This concept is frequently used in business analytics to separate customers into different groups which include high-value, medium-value, or low-value clients.

Recency: When was the last time the client did business with us?

Frequency: How frequently does the client demand a product from us?

Monetary: How much does the client spend on our products?

We computed the recency of customer transactions, the frequency of transactions processed, and the monetary values of clients. Then we ranked each client and calculated their RFM score from 1 to 5 and grouped it by average in scoring accounts to determine our top and low clients.

The graph below depicts RFM by age groups, indicating a variation across age demographics of clients as well as the recency, frequency, and monetary value of clients in the groups.

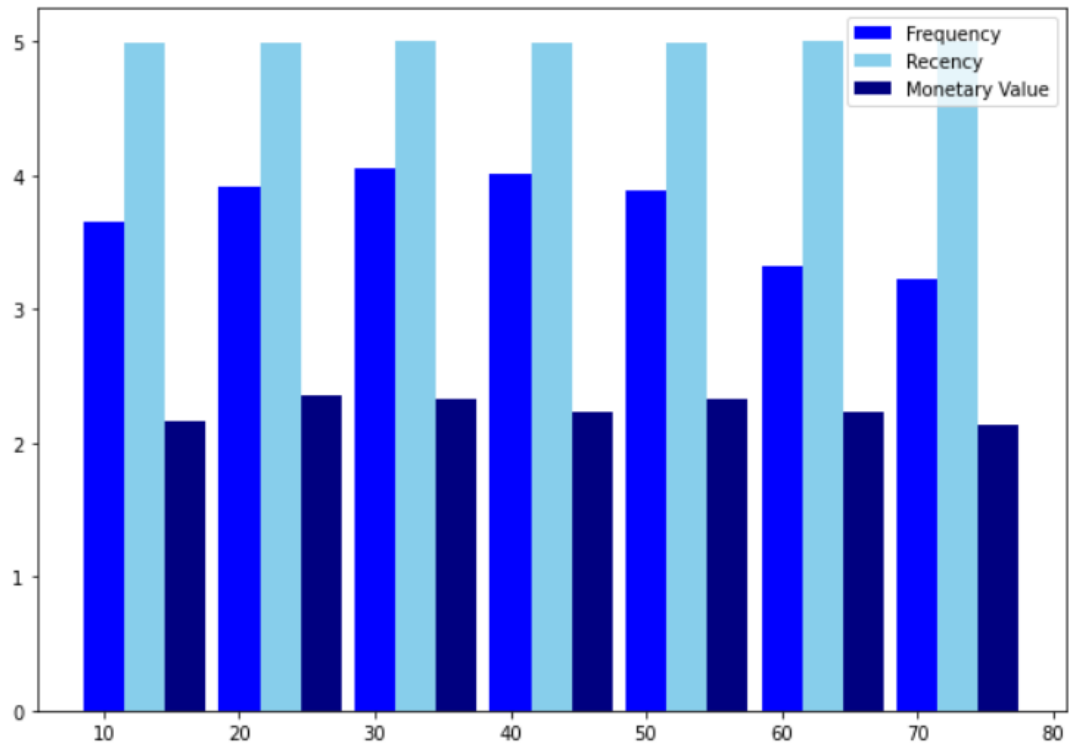

Figure 20: Grouping RFM by Age Groups

BASETABLE VARIABLES DESCRIPTION

VARIABLE NAME	MEANING	DATA TYPE	REMARK
card_issue_year	Year of card issuance	object	
credit_card	Number of Credit card	object	
lor	Relationship duration	int32	
amount_Household	Amount spent on household expenses	float64	
amount_Insurance	Insurance expenditures	float64	
amount_Leasing	Lease payment	float64	
amount_Loan	Amount of the loan	float64	
amount_Unknown	Uncategorized transaction amount	float64	
order_count	Payment order total	float64	
total_amount	Total Amount of order	float64	
gender	Gender of Client	object	Male or Female
age	Age of clients	int32	
age_group	Age demographics of clients	int32	Grouping ages in 10 - 70
Frequency	Number of times a client carries out transactions	int64	
recency	Most recent transactions of clients	int32	

Credit	Inflow of clients	float64	
Withdrawal	Outgoing funds of clients	float64	
Total_MV	Total monetary value of customers	float64	
trans_amount_avg	Mean of transaction amount	float64	
trans_amount_max	Maximum transaction amount	float64	
trans_amount_min	Minimum transaction amount	float64	
trans_No Description_amt	Number of Transaction without Description	int32	
trans_Household_amt	Amount for transaction relating to household	int32	
trans_Insurance_amt	Amount for Insurance transactions	int32	
trans_Interest_amt	Interest transaction amount	int32	
trans_Loan_amt	Loan transaction value	int32	
trans_Negative Interest_amt	Amount of Negative Interest	int32	
trans_Pension_amt	Amount of pension transaction	int32	
trans_Statement_amt	Amount for a transaction on a statement of account	int32	
trans_Unknown_amt	Amount for unidentified transactions	int32	
oper_CC_Withdrawal_amt	The total amount of credit card withdrawal operations	int32	
oper_Cash_Withdrawal_amt	Amount of cash withdrawal operations	int32	
oper_Collect_other_bank_amt	Amount derived from other banking operations	int32	
oper_Credit_in_Cash_amt	Amount of credit in cash operations	int32	
oper_Remit_to_other_bank_amt	Amount remitted to other bank operations	int32	
Monetary_Categ	Client account monetary category ranking	int32	
Frequency_Categ	Client account frequency category ranking	int32	
Recency_Categ	Client account ranking in the category of recency	int32	
acct_categ	Account classification	float64	
avg_credit	Credit inflows on average	float64	
min_credit	Credit inflows at a minimum	float64	
max_credit	Credit inflows at a maximum	float64	
count_credit	Total amount of credit per account	int64	
avg_withdrawal	The average amount withdrawn	float64	
min_withdrawal	Minimum withdrawal amount	float64	
max_withdrawal	Maximum withdrawal amount	float64	
count_withdrawal	Number of withdrawals in total	int64	

loan_amount	Amount of loan per client account	float64	
loan_duration	Each client's loan duration	float64	
loan_payments	Each account's loan payment	float64	
loan_status	Each account's loan status check	object	
loan_recency	Most recent loans	float64	
inhabitants	Number of residents	int64	
munic_499	municipalities with a population of at least 500 people	int64	
munic_500_to_1999	Number of municipalities with populations greater than 500 less than 2000	int64	
munic_2000_to_9999	Number of municipalities with populations greater than 2000 less than 10000	int64	
munic_10000	Municipalities with a population of more than ten thousand people	int64	
cities	Count of cities	int64	
ratio_inhab	Ratio of city dwellers	float64	
avg_salary	Average wage	int64	
unemployment_rate96	Unemployment rate in 1996	float64	
entrepreneurs_per1000	Number of entrepreneurs per 1000 people	int64	
crimes96	The crime rate in 1996	int64	
has_loan	Loaned accounts	object	1 for has loan 0 for no loan
has_cc	Credit card transactions	object	1 for credit card 0 for no credit card
frequency_Transactional	Accounts with a transaction frequency	uint8	
frequency_Weekly	Accounts with a weekly frequency	uint8	
card_type_classic	Accounts using a classic card	uint8	
card_type_gold	Accounts using a gold card	uint8	
card_type_junior	Account using a junior card	uint8	
region_central Bohemia	Account in the region of Central Bohemia	uint8	
region_east Bohemia	Account in the region of East Bohemia	uint8	
region_north Bohemia	Accounts in the region of North Bohemia	uint8	
region_north Moravia	Account in the region of North Moravia	uint8	
region_south Bohemia	Account in the region of South Bohemia	uint8	

region_south Moravia	Account in the region of South Moravia	uint8	
region_west Bohemia	Account in the region of West Bohemia	uint8	