Marijose Marcos Iglesias
Nguyet Han Nguyen
Mario Serrano

Master's in Big Data Analytics

Financial Programming Group Project

30/11/2021

Introduction

In this project, the authors were asked to construct a data science base table from the financial data set “Berka”. The base table is created through merging the 8 provided tables in the data set (account, card, client, disp, district, loan, order and trans), after each of the individual tables had been organized and cleaned through the removal of missing values, columns, etc. The goal of merging these tables is to make data manipulation easier in order to get insights on clients and loans, such as loans per client, clients, and cards for both types of accounts such as number of credit cards per client and account type, etc. To have more complete information, new variables were created to analyze the previously described insights. These variables are individually described in the Variable guide attached to this report.

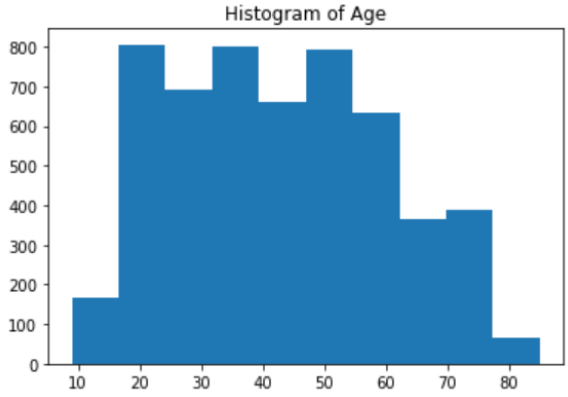
The final merged table displays the 2 required targeted variables (granted loans per client and credit card per client and account type). To have enough data to study the variables, the authors chose solely clients that had a bank history of at least a year /a length of relationship in years (LOR) >0.

In this report, the cleaning process of each table before merging as well as the merging process and target variable creation will be described to detail to reach deep understanding of the goal of the project and to show the importance of the acquisition of the targeted variables and the data analysis.

Data Processing of Tables

1. Client

For the client table, we started off by checking if there were any missing values, and none were found. The variables for year, month, and day of birth for each client were generated, along with the gender variable per client. The age for each client in the year 1996 was calculated; for the data to be more detailed, age groups were also created to classify clients into age groups, which had 15-year ranges. The following graphs show the histogram of age, and the distribution of the age groups, respectively:

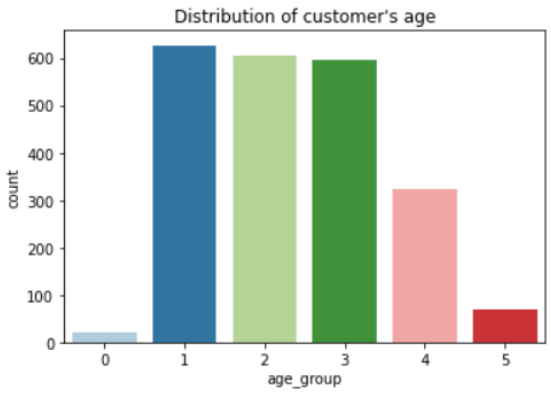


We used this table as the main table to merge the subsequent treated tables and new variables.

To get deeper insights on clients and their age, clients were grouped and counted by age, with which we found out only 72 of the clients are 18 or younger, representing 0.0134% of total clients.

Analyze and visualize independent variables:

- **age_group**



age_group	min	max	count
0	14	14	21
1	15	29	627
2	30	44	604
3	45	59	595
4	60	74	322
5	75	78	70

The bar graph above shows the number of client in each age group. There is only 21 clients below 14, accounting for 0.94% total clients in the basetable. The highest number of clients are from age group 15 to 29 years old. The percentage of each age grop is 0.94%, 28%, 27%, 26.6%, 14.4% and 3.1% respectively.

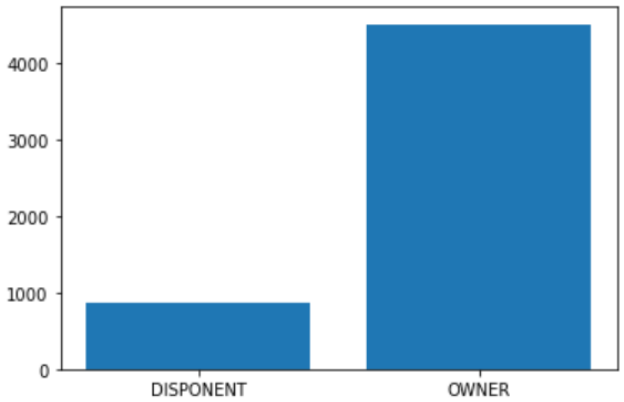
2. Card

In the cards table we started by creating independent variables for the year, month, and day of when the card was issued, which we extracted from the issue date of the card. Then, we made sure no missing values needed to be treated. With the issue year, we were able to estimate the amount of years that the client had the credit card for, up until the year 1996. Then, we checked how many disponents there were by card after grouping the cards per type and we found out there was only 1 disponent per card.

For this table we filtered with a subset only credit cards that were issued until 1996, excluding all information from the year 1997 and beyond. This table was joined with the client table, and we generated a dummy variable called “has_card_96” that shows if a client has a client (1) or not (0).

3. Disposition

For this table, we checked for missing values, which were also not found. Afterwards, the distribution of disponents and owners of the accounts were extracted:



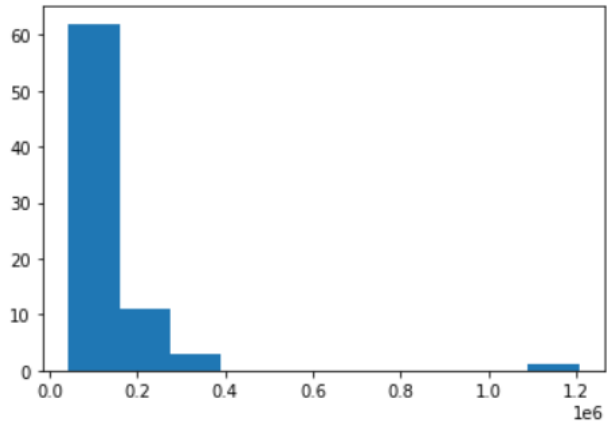
We generated two subsets from this table, one for the owners and a second one for disponents. With the owner subset we merged and filtered out only clients who are owners by doing an inner join with the client

table, to identify which client IDs correspond to owners. After that, we calculated the number of disponents by owner, which is 1 per owner.

After retrieving this information with the previous process, we joined the disponent's subset for each client id to identify which owners have disponents and which do not, then replaced the missing values for the "num_dispo" variable with zeros. After that, a dummy variable "has_card_96" was created to identify which clients hold a card up to year 1996, with which we found out only 95% of clients hold a card. Finally, we grouped the count of disponents by client id and joined that new variable to the main client table.

4. District

For this table, we started off by renaming the variables and assigning descriptive names for easier data manipulation. After no missing values were identified, a histogram for inhabitants in millions was performed:



Then, the variable "district_size" was created to classify the size of the district according to the number of inhabitants per district. This variable would then calculate and categorize each district as small, medium, or large according to their percentile on population.

After grouping by district size, we extracted information on the number of districts that are small medium in large, which in this case, most districts showed to have a medium size.

For unemployment rate and crime, some question mark ("?") signs were identified in small sized districts, which we then proceeded to replace for the mean of the small district's unemployment rate and crime. After, crime values for both years were changed to ratios for it to be easier to read in the data and we dropped the crime values for both years as well, since they were not needed anymore. Finally, we merged this table with the main client table.

5. Order

For this table, we checked for missing values as for the other tables before treating the information and none were identified. We checked for outliers in the data that needed treatment, but none were identified. To get deeper insights on the table we grouped per bank and amount and calculated the mean, min, max and median. After, the observations for the variable "k_symbol" were translated into English and we filled in the missing values with zeros.

Due to this table not having a date variable, we cannot know when the orders happened, and since we were asked to filter for the year 1996, we decided not to use this table to be certain that we were not adding any non-required data.

6. Trans

For this column we started off by separating the date variable into individual variables for transaction year month and day, then we checked for missing values in the table. Since they were categorical variables only that had missing values, we filled the missing values with the word "others" and translated to English variables “k_symbol” and “operation". Then, we subset the data for the year 1996, and extracted mean, min, and max of transactions to get more insights on the data.

Transactions were then grouped by account id and displayed by type of transaction to get the mean of every transaction type by account id as well as the sum of transaction types by account id. To retrieve more insights on this table, the group also calculated mean and sum of k_symbol by account id. The same procedure was applied for operation and account id for mean and sum. The variable “n_trans” was created to calculate number of transactions per account and after, the mean of the transactions for each month in the year were calculated by account id, to identify the months where most transactions were performed.

The variable “to_bank” was created to identify if transactions were made to an external bank (1) or not (0). After treating all the table by creating new variables and retrieving important insights through data manipulation.

From the transactions table we created the recency, frequency and monetary values:

- **Recency**: We created four recency variables, two for debit (in days and months) and two for credit (in days and months). These variables were calculated in months and in days using the 31 of December 1996 as the maximum date.
- **Frequency:** This variable was created as the number of transactions during 1996 by each client.
- **Monetary Value:** This is represented by several variables in the base table that show the sum and mean amount by k-symbol.

Analyze and visualize independent variables:

- op_collect_other_bank_mean, op_credit_cash_mean, op_debit_card_mean, op_debit_cash_mean, op_remit_other_bank_mean

The histogram illustrates the distribution of average transactions amount by character, namely collecting from other banks, credit cash, debit card, debit cash, remitting other banks. All the histogram from each character is right skewed, except for average transactions amount from credit cash.

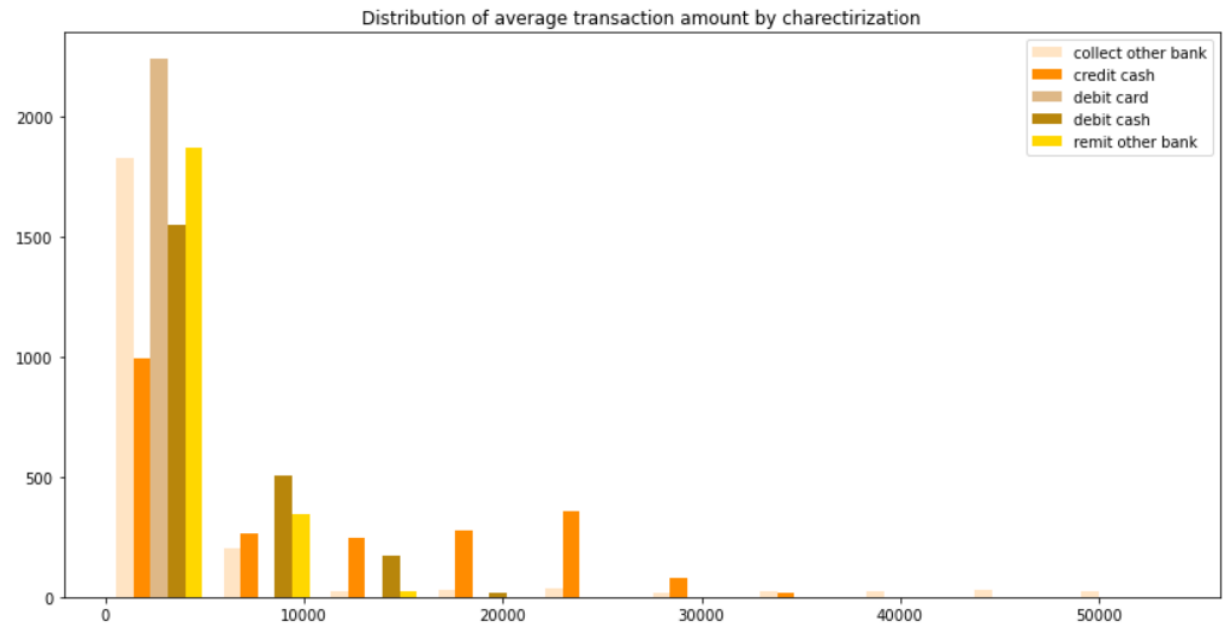
For collecting from other banks, the range of the transactions amount is 54,000 while the median of the transactions is 0, indicating many large outliers.

50% of transactions amount from credit cash is under 20,000. The most considerable frequency of transactions is under 10,000. In addition, there is a smaller hill whose peak at around 25,000.

For debit cards, almost all the transactions amount to 0.

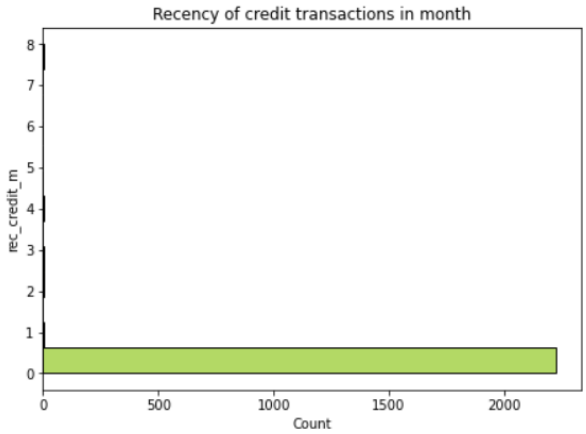
Regarding debit cash, most of the transaction amount is under 10,000 with mean equals 4335.

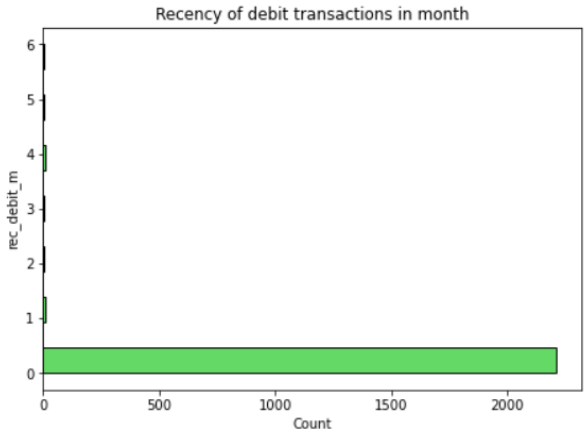
The average transaction amount from remitting other banks is from 0 to nearly 15,000 with 50% of the transactions is between 1000 and 4000 approximately.



	index	op_collect_other_bank_mean	op_credit_cash_mean	op_debit_card_mean	op_debit_cash_mean	op_remit_other_bank_mean
0	count	2238.0	2238.0	2238.0	2238.0	2238.0
1	mean	4201.7	10454.9	173.8	4335.2	2950.8
2	std	9960.7	9880.6	634.7	3776.8	2667.1
3	min	0.0	0.0	0.0	14.6	0.0
4	25%	0.0	0.0	0.0	1411.1	1005.1
5	50%	0.0	7837.8	0.0	3048.0	2418.5
6	75%	4202.2	19553.7	0.0	6285.7	4089.4
7	max	54000.8	35331.3	3900.0	19639.4	14658.0

- **rec_credit_m and rec_debit_m:**





The first bar chart reports the recency of credit transactions in months. Most of the recent credit transactions are under one month, specifically 2225 credit transactions, accounting for 99.4%. There are only 10 credit transactions from one to two months ago, and 3 transactions from three to eight months ago.

The second graph shows the recency of debit transactions in month. Again, more than 98.8% of recent debit transactions is under one month, 2211 transaction to be more precise. The number of recent debit transactions falls between one month to six months is 27, represents for 1.2% total debit transactions.

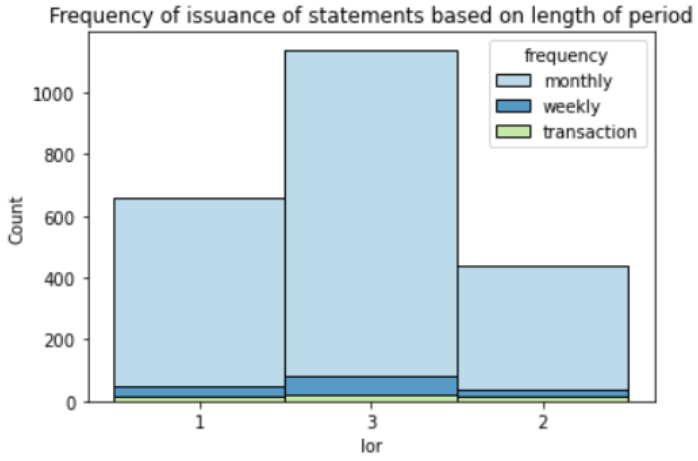
7. Account

The variable “frequency” of this table was first treated by translating it to English, after checking for missing values which were not identified. An independent variable for year, month and day was created from the date variable as well as a variable (“lor”) to define the length of relationship of the client and the bank, since we want information only before the year 1996, we filter the length of the relationship only before that year for the data to be accurate. With this information, we then plotted a histogram to get the frequency of issuance of statements based on length of period of accounts:

Then the table was merged to the main client table.

Analyze and visualize independent variables:

- frequency, lor



The bar chart represents the frequency of statement issuance in each card issued cycle length. Most of the statement issuance frequencies in all three cards issued cycle length groups is monthly, followed by week and finally by transaction.

8. Loan

Firstly, the table was checked for missing values, which were not identified; separate variables were created for year, month, and day from the date variable. The “status” variable and was then renamed with clearer names to easily identify whether the loan was paid or not. With this information we found out most of the clients are categorized as “ok” and only 45 clients are in debt. The loans were then filtered up to the year 1996 and the remaining data was dropped since it was not useful.

The table was then merged to the main client table.

9. Base Table

Acquisition of Target Variables:

There are two target variables (dependent variables) in the base table:

- Target variable 1 (has_loan_97) represents whether a customer had granted loan in 1997. If the value equals 0 then the client has not been granted a loan, otherwise the client has been granted a loan.
- Target variable 2 (card_issued_97) shows the customer's credit card issuance for both owners and disponents. The clients having a credit card issued are assigned a value of 1, while clients who did not have a credit card issued receive a value of 0.

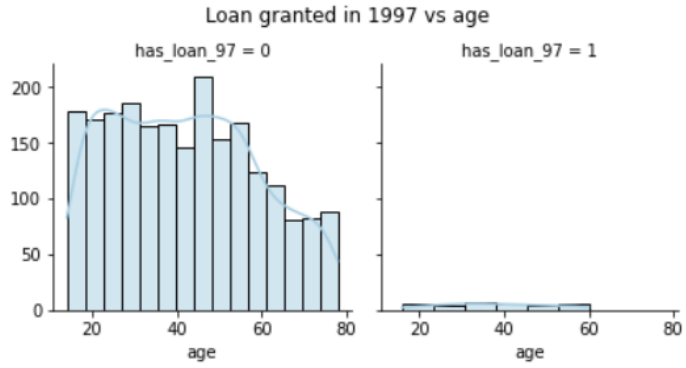
For the first target variable, we choose the year of loan table equals 1997 to fit in the timeline and merge with the owners_complete table.

For the second target variable, we filter the year from card table equals 1997 and merge with the previous table to create the base table.

The final base table contains 2239 observations and 98 variables. To be specific, 91 variables are numerical (including 2 target variables), and 7 variables are categorical. Out of 2239 clients, 2208 clients accounting for 98.6% total clients in the base table did not have loan granted in 1997 while 31 clients accounting for 1.38% total clients had granted loan in 1997. Considering credit cards issued in 1997, the number of clients who did not have a credit card issued and had a credit card issued are 2119 and 120, respectively. The percentage of each group is approximately 94.6% and 5.4%.

Analyze and visualize target variables.

Age vs has_loan_97 :



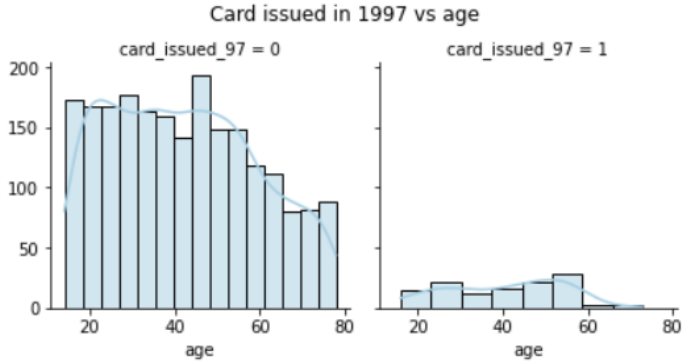
	age_group	card_issued_97	number_of_clients
0	0	0	21
1	1	0	593
2	1	1	34
3	2	0	574
4	2	1	30
5	3	0	543
6	3	1	52
7	4	0	318
8	4	1	4
9	5	0	70

Regarding clients having a loan granted in 1997, four age groups out of six have clients with granted loans in 1997, in which age ranges from 15 to 74. Age group between 30 and 44 has the highest number of clients having loan granted in 1997 (13 clients), while age group from 9 to 14 and age group from 75 to 85 do not include any clients granted a loan.

Regarding clients who did not have a loan granted in 1997, there is a decrease in the number of clients across age. Age group from 15 to 29 has the highest number of clients not having loan granted in 1997 (619 clients).

The percentage of customers having loan granted in each age group are 0%, 1.28%, 2.2%, 1.5%, 0.3% and 0% respectively.

Age vs card_issued_97:



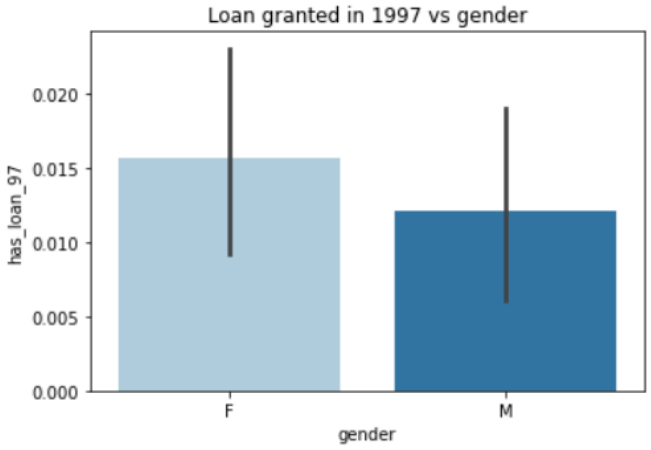
	age_group	card_issued_97	number_of_clients
0	0	0	21
1	1	0	593
2	1	1	34
3	2	0	574
4	2	1	30
5	3	0	543
6	3	1	52
7	4	0	318
8	4	1	4
9	5	0	70

Like the first target variable, four age groups out of six have customers having a card issued in 1997. The highest number of clients having a card issued in 1997 is 52, which belongs to an age group from 33 to 44 years old.

A decline of people who do not have a card issued in 1997 appears during the increase of age. Again, the highest number of customers not having a card issued in 1997 is between 15 and 29 years old.

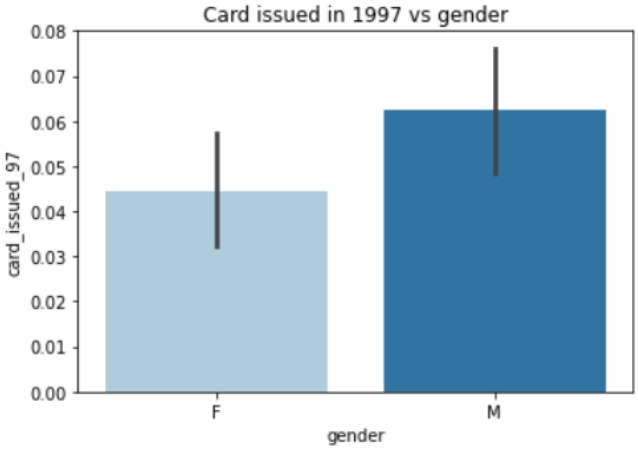
The percentage of customers having card issued in each age group are 0%, 5.4%, 5%, 8.7%, 1.2% and 0% respectively.

Gender vs has_loan_97:



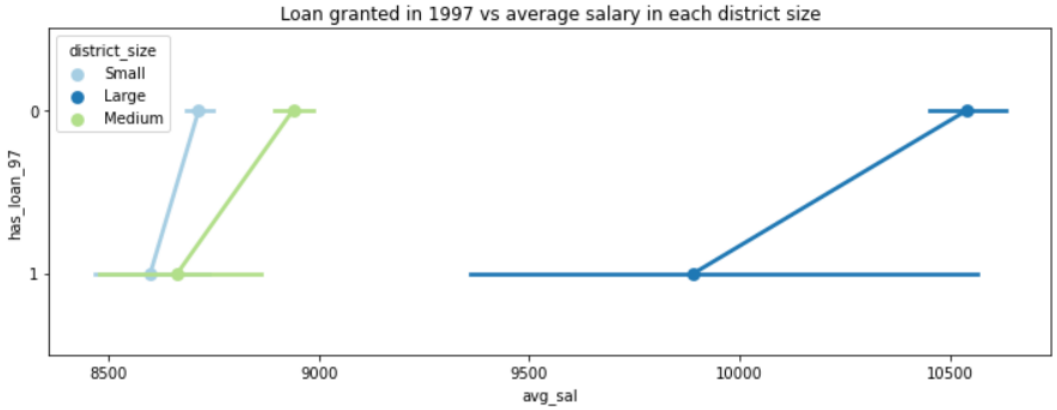
Among female clients, the average of having loan granted in 1997 was more than 1.5%, while for males, the average of having loan granted in 1997 was around 1.25%. Specifically, 17 female and 14 male customers had a loan granted in 1997.

Gender vs card_issued_97:



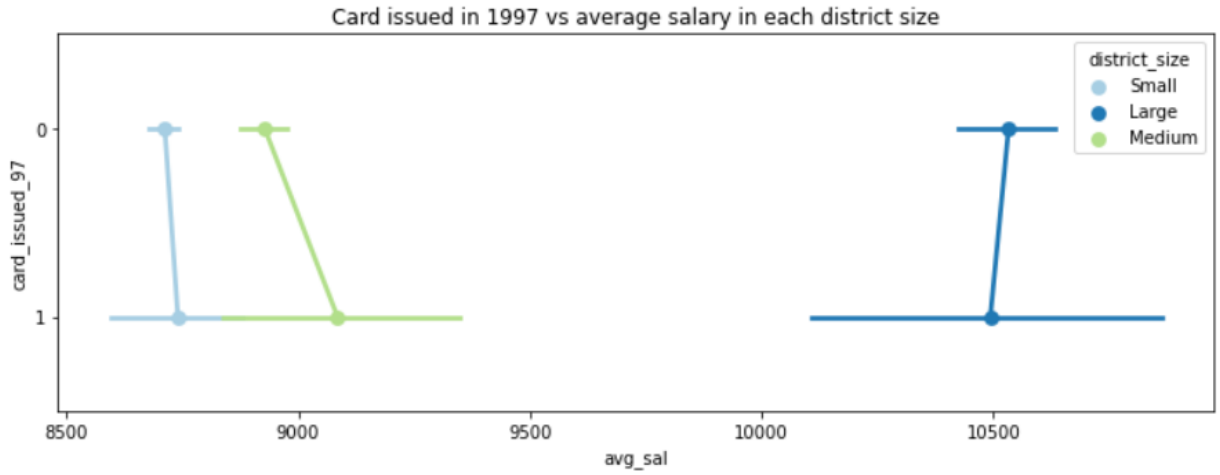
The average of having card issued in 1997 for females was less than 5% and for male customers was more than 6%. There are 48 females and 72 males had card issued in 1997.

Avg_sal vs has_loan_97 in each district size:

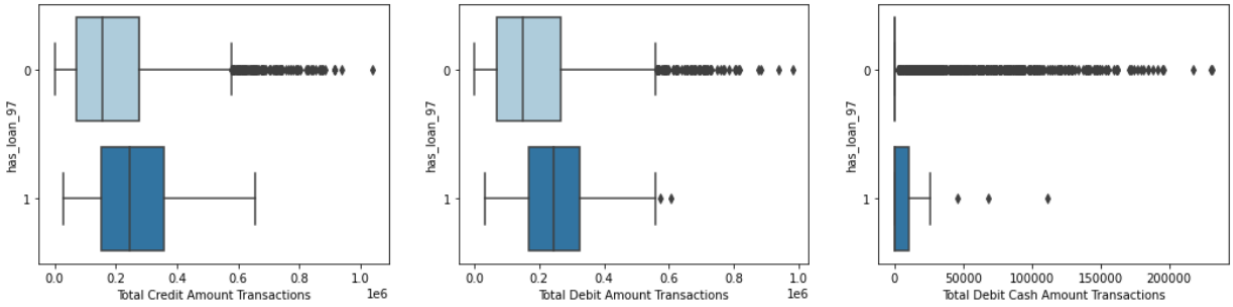


The mean average salary of the group having granted a loan in 1997 is lower than that of the other group in every district size. To be more specific, in large district size, the mean average salary of the clients in the first group, which was not granted a loan in 1997, is around 10,500 while it is only 9,800 in the other group. The bands that go through the point represent the confidence interval. In addition, the graph shows that district size is proportional to the mean average salary. In another way, the larger the district size, the higher the mean average salary.

For small and medium district size, the point plot shows that the mean average salary of group that had a card issued in 1997 is higher than that of the other group. However, for large district size, the opposite case applies. The mean average salary of cardholders opening in 1997 is lower than that of the other group. To be more precise, clients who were granted a credit card in 1997 had a salary of around 10,500 in 1996, while it is more than 10,500 for the other group.



credit_sum, debit_sum, debit_cash_sum vs has_loan_97:

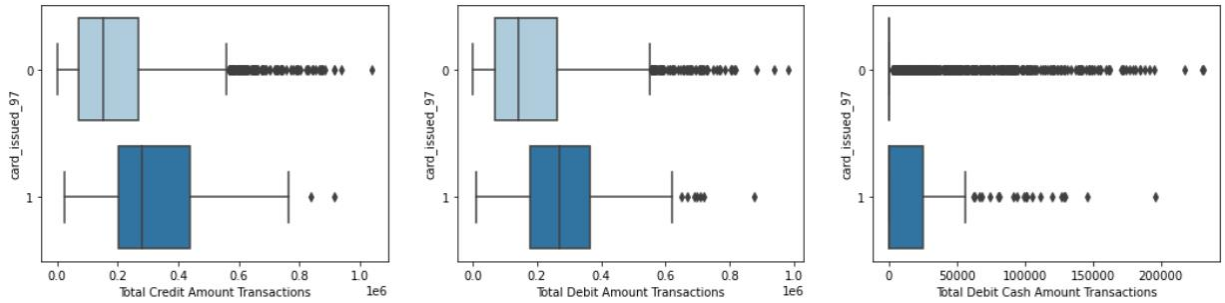


The first box plot illustrates the distribution of total credit amount transactions. The group had not granted a loan in 1997 has a median total credit amount transaction around 150,000. 50% of the transaction amount falls between 70,000 to under 300,000 approximately. There are many outliers, and the maximum amount is proximately 1,000,000. Regarding the group having granted a loan in 1997, the median of the total credit amount transaction is remarkably close to 250,000, which is higher than that of the other group. 50% of the transaction amount is between 150,000 to under 400,000. There is no outlier, and the maximum amount is about 650,000.

The second box plot describes the distribution of total debit amount transactions. There is a greater variability for debit amount transactions as well as larger outliers in the group who had not granted a loan in 1997. The median of the debit amount transaction in this group is close to 150,000. 50% of the transaction amount falls between 70,000 to 250,000 approximately. The median of the total debit amount transaction in the second group, which had granted a loan in 1997, is roughly 250,000. There are only two outliers in this group.

The third box plot represents the distribution of total cash debit amount transactions. There are too many outliers that run from 0 to more than 200,000 in the first group. The other group, which had granted loan in 1997 has only three outliers and 50% of transaction is from 0 to 10,000.

credit_sum, debit_sum, debit_cash_sum vs card_issued_97:



The first box plot shows the distribution of total credit amount transactions. The larger frequency of the first group which did not have a card issued in 1997 is between 70,000 to more or less than 250,000. There is a long tail to the right with many outliers, and the maximum amount is larger than 1,000,000. For the other group that had a card issued in 1997, the median of the total credit amount transaction is remarkably close to 250,000, which is higher than that of the other group. 50% of the transaction amount is between 200,000 to nearly 450,000 relatively. There are only two outliers, and the maximum amount is more than 900,000.

The second box plot defines the distribution of total debit amount transactions. There is a greater variability for debit amount transactions as well as larger outliers in the group had card issued in 1997. The median of the debit amount transaction in this group is close to 140,000. Regarding the second group, most of the transactions fall between 200,000 to 450,000. The median and the mean of the transactions are approximately 290,000 and 270,000 respectively.

The third box plot expresses the distribution of total cash debit amount transactions. For the first group, various outliers which run from 0 to more than 200,000 appeared. The other group, which had a card issued in 1997 had several outliers and 50% of transaction is from 0 to nearly 25,000.

Conclusion

After following a detailed process of data manipulation and having created several new variables we found relevant for the analysis of the data, the results displayed are our base table with the following dimensions:

Rows	2239
Columns	98

The variables were created based on the insights that we got on each table and thinking about the relevance to qualify a person for a loan next year.

The created variables highlight the recency, frequency, and monetary value for each client, as well as specific characteristics. This could be useful to create clusters and classify clients by risk of default or on knowing if they are going to pay the loan on time, etc. A strict timeline was followed, taking all the independent variables from 1996 and the target variables for 1997, meaning that historical data was used to predict the year to come. After this process is concluded, the complete base table is ready to go for further analysis and ready to start with a feature selection procedure.

Variable Guide

Variable Name	Description	Data Type	Value
client_id	client identifier	integer	unique identification
birth_number	birthday and gender of client	integer	the number is in the form YYMMDD for men, the number is in the form YYMM+50DD for women, where YYMMDD is the date of birth
district_id	district identifier	integer	unique identification
mean_amount	average of amout of money per account	float	amount
sum_amount	total of amout of money per account	float	amount
min_amount	minimum amout of money per account	float	amount
max_amount	maximum amout of money per account	float	amount
n_trans	number of transactions per account	float	amount
credit_mean	average amount of money by credit transaction per acount	float	amount
debit_mean	average amount of money by debit transaction per acount	float	amount
debit_cash_mean	average amount of money by debit cash transaction per acount	float	amount
credit_sum	total credit amount of money	float	amount
debit_sum	total debit amount of money	float	amount
debit_cash_sum	total debit cash	float	amount
others_sum	sum of other transaction types	float	amount
trans_household_sum	total household payment per account	float	amount
trans_insurance_sum	total insurance payment per account	float	amount
trans_interest_sum	total sanction interest if negative balance per account	float	amount
trans_interest_cred_sum	total interest credited per account	float	amount
trans_loan_sum	total loan payment per account	float	amount
trans_others_sum	total other amount of money per account	float	amount
trans_pension_sum	total old-age pension per account	float	amount
trans_statement_sum	total payment for statement per account	float	amount

birth_year	year of birth of client	integer	year
birth_month	month of birth of client	integer	from 1 to 12
birth_day	day of birth of client	integer	from 1 to 31
gender	gender of client	object	Male or Female
age	age of client	integer	from 1 to 100
age_group	age group of client	integer	6 age groups by every 15 years, starting from 0 and finishing in 85
disp_id	record identifier	integer	unique identification
account_id	account identifier	integer	unique identification
num_dispo	number of disponents per account owner	float	amount of disponents
card_type	type of card the owner holds	object	owner or disponent
issued	card issue date	object	n the form YYMMDD
issue_year	card issue year	float	in form YYYY
issue_month	card issue month	float	in form MM
issue_day	card issue day	float	in form DD
lor_card	lenght of relationship with card (for how long clients have owned a card)	float	measured in years
has_card_96	clients who own a card in 1996	integer	dummy variable: 1 if they own a card or 0 if they do not
district_name	name of district	object	district name
region	region of district	object	region name
inhabitants	number of inhabitants by district	integer	amount
nmuni_5	no. of municipalities with inhabitants < 499	integer	amount of municipalities
nmuni_2k	no. of municipalities with inhabitants 500-1999	integer	amount of municipalities
nmuni_10k	no. of municipalities with inhabitants 2000-9999	integer	amount of municipalities
nmuni_more_10k	no. of municipalities with inhabitants >10000	integer	amount of municipalities
ncities	number of cities	integer	cities amount
ratio_urb_inhab	share of urban inhabitants	float	number of urban inhabitants divided by inhabitants and multiplied by 1000
avg_sal	average salary in district	integer	average amount
unemp_95	unemployement rate 1995	float	number of unemployed divided by inhabitants and multiplied by 1000
unemp_96	unemployement rate 1996	float	number of unemployed divided by inhabitants and multiplied by 1000
entrep	no. of enterpreneurs per 1000 inhabitants	integer	entrepreneur amount
district_size	size of district	object	in categories "small", "medium" or "large" "small" if < than the 25th percentile, "medium" if > than the 25th percentile and "large" if > than the 75th percentile
crime_rate_95	number of crimes in 1995 pero 1000 inhabitants	float	number of crimes divided by inhabitants and multiplied by 1000

crime_rate_96	number of crimes in 1996 pero 1000 inhabitants	float	number of crimes divided by inhabitants and multiplied by 1000
others_mean	mean of other transaction types	float	transaction amount
trans_household_mean	average household payment per account	float	transaction amount
trans_insurance_mean	average insurance payment per account	float	transaction amount
trans_interest_mean	average sanction interest if negative balance per account	float	transaction amount
trans_interest_cred_mean	average interest credited per account	float	transaction amount
trans_loan_mean	average loan payment per account	float	transaction amount
trans_others_mean	average l other amount of money per account	float	transaction amount
trans_pension_mean	average old-age pension per account	float	transaction amount
trans_statement_mean	average payment for statement per account	float	transaction amount
balance	mean balance during the year 1996	float	transaction amount
op_collect_other_bank_mean	amount mean of "collect other bank" operation	float	transaction amount
op_credit_cash_mean	amount mean of "credit cash" operation	float	transaction amount
op_debit_card_mean	amount mean of "debit card" operation	float	transaction amount
op_debit_cash_mean	amount mean of "debit cash" operation	float	transaction amount
op_remit_other_bank_mean	amount mean of "remit other bank" operation	float	transaction amount
ntrans_banks	number of transactions to other banks	float	transaction amount
mean_amount_1	mean amount for the month of January	float	transaction amount
mean_amount_2	mean amount for the month of February	float	transaction amount
mean_amount_3	mean amount for the month of March	float	transaction amount
mean_amount_4	mean amount for the month of April	float	transaction amount
mean_amount_5	mean amount for the month of May	float	transaction amount
mean_amount_6	mean amount for the month of June	float	transaction amount
mean_amount_7	mean amount for the month of July	float	transaction amount

mean_amount_8	mean amount for the month of August	float	transaction amount
mean_amount_9	mean amount for the month of September	float	transaction amount
mean_amount_10	mean amount for the month of October	float	transaction amount
mean_amount_11	mean amount for the month of November	float	transaction amount
mean_amount_12	mean amount for the month of December	float	transaction amount
to_bank	transactions to external banks	float	amount of transactions
frequency	frequency of issuance of statements	object	monthly issuance weekly issuance issuance after transaction
open_year	year account was opened	integer	year
open_month	month acount was opened	integer	from 1 to 12
open_day	day account was opened	integer	from 1 to 31
lor	lenght of relationship with the bank (since account was opened)	integer	measured in years
time_since_loan	time in years since loan was granted	float	years
has_loan_97	clients with loans in 1997	integer	number of clients
card_issued_97	cards issued in 1997	integer	amount of cards
rec_debit_m	recency of debit transactons in months	integer	number of months
rec_credit_m	recency of credit transactons in months	integer	number of months
rec_debit_d	recency of debit transactons in days	integer	number of days
rec_credit_d	recency of credit transactons in days	integer	number of days