Financial Programming

Group Project - Financial Data Set

By Omar Mantilla, Juliana Sanchez Ramirez & George Simonsen

Table of Contents

Introduction

This report aims to construct a data science basetable using the PKDD'99 financial data set. The financial data set has tables which provide information about a bank, its clients, and client accounts with regards to recorded transactions, characteristics, and demographic information. This report will build on the currently vague idea of who is a good and bad client and improve the bank's understanding of its customers through the improvement of services.

Through this report our group aims to develop, analyse, and visualise independent and dependent variables during a specific time window. The report is split into 3 parts:

1. ***Development*** - Describe the independent and dependent variables created for the purpose of this report.
2. ***Analysis*** - Describe the data transformation implementation and methods applied throughout the basetable.
3. ***Visualisation*** - Visualize the independent variables and dependent variables aforementioned.

Basetable Variables

We have developed a basetable from the different datasets within the PKDD'99 financial data set. In this section we will describe each variable which has been incorporated into the basetable along with the variables name, description, data type and output (values). We will also provide the decision making process behind the inclusion of each variable into our basetable. Please note that all variables which are highlighted have multiple variables within them as mentioned.

Variable Name	***Variable Description***	***Data Type***	***Output (Values)***	***Reason for inclusion***
Client_Id	Provides a value for each client in the dataset	Integer	From 1 to ... (in our case there are 2239 OWNER clients)	Useful to merge across different data sets. Primary key
disp_id	Links account, credit and client data sets	Integer	From 1 to … (in our case there are 892 card accounts linked)	Useful to merge across different data sets.
account_id	Identification value for each account	Integer	From 1 to ... (in our case there are 2239 OWNER accounts)	Useful to merge across different data sets.
Age	Provides the age of each client in terms of years	Integer	Youngest OWNER client is 14 and the oldest is 78	Useful to understand the age distribution of the clients
bank_district_id	Provides a value for each district in the dataset	Integer	From 1 to ... (in our case there are 77 districts)	Useful to merge across different data sets. Districts information

lor	Length of relationship since opening an account before 1996	Integer	From 0 to 3 years as lowest date in datasets is for accounts opened in 1993	Indicates the number of years since the opening of the client's account
lor_card_m	Length of relationship since starting a credit card (in months)	Float	From 0 to 11 months (Analysis only for the cards issued in 1996)	Number of months since the issue of the credit card (In case that the client has a card)
Age_group_	4 variables which group OWNERS based on age 1. Teenagers 2. Young Adults 3. Middle Age 4. Senior	Integer	Four group types which are classified using dummy variables 0 and 1	Useful in grouping datasets to be able to focus on a specific group when necessary
gender_M	Gender type, male or female	Integer	Using dummy variables, if 1 is a Male, if 0 is a Female	Demographic information can be useful to understand the client's profile
Frequency_	Frequency of issuance of statements 1. Monthly 2. Weekly	Integer	Dummy variables for monthly and weekly, in case of zero in both variables the frequency is after transaction	Gives information about the periodicity of the account
Card_type_	3 card type variables which outline the cards used by OWNERS: 1. Junior 2. Classic 3. Gold	Integer	Dummy variables for the type of card in case of having. In case of zero for the 3 variables the client does not have a credit card	Information about having or not a credit card and the type of card which can give information about how premium or not the client is
Nb_inhabitants	Provides values for the total number of inhabitants in the district of the client	Float	Values are normalised and go from 0.035537 to 1. Normalization: Divided by the max. value	Information about the size of the district where the client lives
******_ ***municipality***	4 variables which consider municipality sizes with regards to cities and districts 1. Small 2. Medium 3. Large 4. VLarge	Integer	Count of municipalities per size in the district where the client lives	Demographic information to profile the customers
nb_cities	Gives the number of cities per district according to the district of the client	Integer	There is anywhere from 1 to 11 cities depending on the district	Demographic information to profile the customers
urbinhabitants_ratio	Provides a ratio for the number of inhabitants which live in urban areas per district according to the district of the client	Float	The ratio varies from 33.9% to 100%	Demographic information to profile the customers
avg_salary	Provides values for the avg_salary in the district of the client	Float	Values are normalised and range from 0.64667 to 1. Normalization: Divided by the max. value	Demographic information to profile the customers
96unemp_rate	Provides the rate for unemployment in 96 in the district of the client	Float	The rate ranges from 0.43% to 9.4% depending on the district	Demographic information to profile the customers
Entrepreneurs_per 1000_inhad	Provides a figure for the number of entrepreneurs in each district as per 1000 inhabitants. According to the the district of the client	Integer	Ranges from 81 to 167 as per the district	Demographic information to profile the customers
crimepct96	Provides the rate for crime in 96 as per district	Float	The rate ranges from 1.59% to 8.22% depending on the district	Provides figures for crime different district in1996
Crime_pct change	Shows the change in crime from 1995 to 1996	Float	The rate ranges from -0.82% to 7.01% depending on the district	Provides a visual difference between crime in different district between 1995 and 1996

Unemp_pct change	Shows the change in unemployment from 1995 to 1996	Float	The rate ranges from -0.36% to 3.17% depending on the district	Provides a visual difference between unemployment levels in different district between 1995 and 1996
loan_duration	Provides the total length of the loan at the moment of granting	Float	The loan duration can go from 0 to 60 months	For the loans granted in 1996 information about the duration
loan_recency_m	The recency analyses the number of months since the granting	Float	As per each client the loan recency left is anywhere from 0 to 11 months in case that the client has a loan from 1996	Information about the loans recency
running_loan	Variable to identify clients with loan granted in 1996 and still running	Float	Dummy variable	Identifying clients with loan running
num_trans_	Provides the number of transactions per account for the month in the variable name, includes: ● Jan96 ● Feb96 ● Mar96 ● Apr96 ● May96 ● Jun96 ● Jul96 ● Ago96 ● Sep96 ● Oct96 ● Nov96 ● Dec96	Float	The output is the sum of the value transactions done from January to December1996	Insight about monthly transactions
Avg_trans_	Provides the average number of transactions per account for the month in the variable name. Includes: ● Jan96 ● Feb96 ● Mar96 ● Apr96 ● May96 ● Jun96 ● Jul96 ● Ago96 ● Sep96 ● Oct96 ● Nov96 ● Dec96	Float	The output is the average of transactions of 1996. From 1st to 31st for each month	Insight about monthly transactions
Credit_transac	This variable was renamed from PRIJEM to Credit_transac.The variable provides information about the credit transactions done by account IDin 1996	Float	The output value is the sum of credit transactions done per account ID.	Insight about credit transactions is good to see account credit consumption.
Withdrawal_transac	This variable was renamed from VYDAJ to Withdrawal_transac. The variable provides information about the withdrawal transactions done by account ID in 1996	Float	The output value is the sum of withdrawal transactions done per account ID.	Insight about credit transactions is good to see account withdrawal behavior.
Other_ operations	This variable provides information about operations that are not identified in the Top 5.	Float	The output value is the sum of other operations done per account.	Insight about other operations per account.
Another_bank_ remittance	This variable has been renamed from PREVOD NA UCET to Another_bank_remittance. This provides information about this operation per account in the year of 1996.	Float	The output value is the sum of bank remittance done per account.	Insight about bank remittance per account.
Ext_bank_ collection	This variable has been renamed from PREVOD Z UCTU to Ext_bank_collection. This provides information about this operation per account in the year of 1996.	Float	The output value is the sum of collections done per account from another bank.	Insight about bank remittance per account.

Credit_in_cash	This variable has been renamed from VKLAD to Credit_in_cash. This provides information about this operation per account in the year of 1996.	Float	The output value is the sum of credit in cash done per account.	Insight about credit in cash per account.
Cash_ withdrawal	This variable has been renamed from VYBER to Cash_withdrawal. This provides information about this operation per account in the year of 1996.	Float	The output values are the sum of the amounts of the cash withdrawal operations.	This variable helps us to identify the accounts that most use this type of operation.
Credit_card_ withdrawal	This variable has been renamed from VYBER KARTOU to Credit_card_withdrawal. This provides information about this operation per account in the year of 1996.	Float	The output values are the sum of the amounts of the credit card withdrawal operations. The majority of the accounts has 0 in the this operation	This variable provides information about this operation per account in 1996.
Other_trans_ charact	This variable provides information other kind of transactions types	Float	The output value is the sum of other kind of transactions types per account	Insight about other kind of transactions types per account.
Old-age_ pension	This variable has been renamed from DUCHOD to Old-age_pension and provides information about this operation per account in the year of 1996	Float	The output value is the sum of old age pensions payments per account	Insight about old age pension per accounts
Insurrance_ payment	This variable has been renamed from POJISTNE to Insurrance_payment and provides information about this operation per account in the year of 1996.	Float	The output value is the sum of insurance payments per account.	Insight about account insurance payments.
Sanction_ interes	This variable has been renamed from SANKC. UROK to Sanction_interes and provides information about this operation per account in the year of 1996.	Float	The output value is the sum of negative sanction interests per account.	Insight about account negative sanction interests
Household	This variable has been renamed from SIPO to Household and provides information about this operation per account in the year of 1996.	Float	The output value is the sum of household amounts per account.	Insight about account household amounts.
Statement_ payment	This variable has been renamed from SLUZBY to Statement_payment and provides information about this operation per account in the year of 1996.	Float	The output value is the sum of statement payment per account.	Insight about account statement payment amount.
Interest_ credited	This variable has been renamed from UROK to Interest_credited and provides information about this operation per account in the year of 1996.	Float	The output value is the sum of interest credited per account.	Insight about account interest credited.
Loan_payment	This variable has been renamed from UVER to Loan_payment and provides information about this operation per account in the year of 1996.	Float	The output value is the sum of loan payment amounts done per account.	Insight about account loan payment
Other_bank	This variable provides information about account transactions to banks that are not identified in the top 10 banks. This shows information per account in the year of 1996	Float	The output values are the sum transactions done per account in the year of 1996 to other banks	Insight about account transactions to other banks
bank__ trans**	This variable provides information about transactions done per account to one of the banks in the year of 1996. This code bank is generated internally for easy bank identification processes. This includes: • AB • • CD • EF • GH • IJ • KL	Float	The output values are the sum transactions done per account in the year of 1996 to this bank	Insight about account transactions to this bank

	• MN • OP • QR • ST • UV • WX • YZ			
R_quartile	Indicator of recency according to the quartiles distribution. In case that the recency is in the first quartile the variable is 4, which represents that the client has recent transactions. The same principle for quartiles 2, 3 and 4 in case that the recency of transactions is higher	Integer	Number [1,2,3,4]	Create a ranking for clients according to the recency of transactions. Transactions of withdrawal in cash or remittance to another bak were excluded because those are not generating profits to our bank.
F_quartile	Indicator of frequency according to the quartiles distribution. In case that the frequency is in the fourth quartile the variable is 4, which represents that the client frequently makes transactions. The same principle for quartiles 1, 2 and 3 in case that the frequency is smaller.	Integer	Number [1,2,3,4]	Create a ranking for clients according to the frequency of transactions. Transactions of withdrawal in cash or remittance to another bak were excluded because those are not generating profits to our bank.
M_quartile	Indicator of monetary value according to the quartiles distribution. In case that the amount of transactions is in the fourth quartile the variable is 4, which represents that the client makes transactions of high amounts. The same principle for quartiles 1, 2 and 3 in case that the amount is smaller.	Integer	Number [1,2,3,4]	Create a ranking for clients according to the value of transactions. Transactions of withdrawal in cash or remittance to another bak were excluded because those are not generating profits to our bank.
RFM_score	Total indicator for the client calculated as the average of the 3 previous variables	float	Float between 1 and 4	Total score for the client.
target_1	Client had granted loan in 1997	Float	Binary variable	Target to predict
target_2	Client had credit card issued in 1997	Float	Binary variable	Target to predict

The variables above have all been included due to their relationship with the data sets which we initially analysed. It is also important to explain 2 aspects about the timeline and variables definition:

1. The variable status from the loans table was recalculated due to the fact that the status updating date is unknown. The assumption was that the status is updated at the end date of the loan. In this case any of the loans granted in 1996 have ended.

2. The information from the table permanent order does not have a date. For this reason the information of debits was not taken into account.

Data Correction & Transformation

Throughout the analysis, development and integration of the different datasets into a single basetable, we conducted a number of different data correction and transformation techniques.

Regarding data correction, we implemented three main methods which include removing missing values, changing variable names and removing redundant variables. We manipulated missing values by removing or changing them with relation to their value type. For example, we replaced *NAN* (missing) values in the data set we built for transactions in 1996 *(trans96 data set).* We replaced the values with zeros so that the observations could still be included throughout the rest of our data sets and basetable. The same method was incorporated for the following data sets within the process:

- ➔ Trans_per_month
- ➔ Avg_trans_month
- ➔ Trans_per_typ
- ➔ Trans_per_operation
- ➔ Trans_per_k_symbol
- ➔ trans_per_bank

We also changed multiple variable names to better align with our basetable. This was to ensure that future users of the basetable could understand the differences with regards to similarly named variables. Additionally, we aimed to minimise the length of variable names, all the while ensuring that they could be understood. For example, we renamed payment types from words including *POJISTNE* and *SIPO* to English words such as *Insurance* and *Household*. Additionally, we renamed the frequencies to outline whether payments were monthly or yearly and ensured that the user could see the different types of cards whether they be Junior, Gold or Classic.

Finally, we removed redundant or repetitive variables which output the same values. We did this to ensure a clean basetable which was still informative without the need for a lot of variables. We also deleted accounts opened before 1996 to have a client base of client ids which was indicative of the purpose of this report, to analyse the different data sets for that year.

Regarding data transformation, our group incorporated the use of dummy variables across a number of different variables within our basetable. We also normalised certain data to decrease the likelihood of data redundancy and improve the integrity of our final basetable. We normalised values within the order data set which had very high values for variables such as the number of inhabitants *(nb_inhabitants)* and average salary *(avg_salary)*. This normalisation consisted in dividing all the values by the maximum value of the variable. Using this technique ensured that we improved the integrity of the final basetable to maintain the distribution of our values along with the different ratio values.

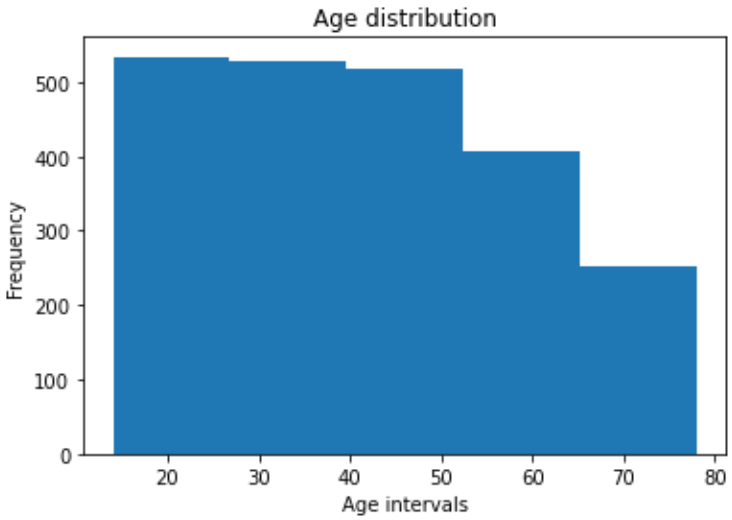
Dummy variables were used for variables including gender, credit grants, payment types, bank payment recipients and more. Regarding each variable type, the dummy variables were incorporated for different functions. For example, regarding the bank recipients and payment types (from order data set), we were able to incorporate the dummy variables in order to count the number of transactions/transaction types per client *(that is considered 'OWNER')*. This enabled us to produce a cleaner basetable all the while providing clarity in the client payment behaviours. Additionally, when building our target variables, we developed dummy variables which represent whether a client did *(1)* or did not *(0)* have a loan grant. Such dummy variables highlight the necessary clients we wanted within our dependent variables.

In turn, we incorporated a number of different data correction and transformation techniques in order to build and provide future users with an informative and concise basetable, all the while building on the 8 starting data sets with additional variables built within our dashboard.

Variable Traits

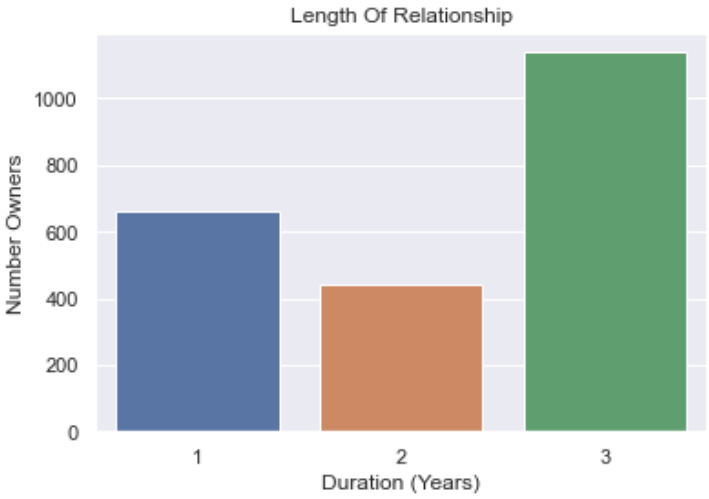
Age Distribution Table

The age distribution chart provides an overview of the varying ages of OWNER clients within the bank. It can be observed that ages vary from early teens to late 70s, with the predominant amount of account owners aging between 20 to 40 years of age. This chart can be presented to the Bank to show the types of age groups which can be focused for bank-client relationships. We highly recommend strengthening relationships with younger clients to avoid the dip seen at 50 years of age and be able to retain clients when they have increased wages, pensions etc.



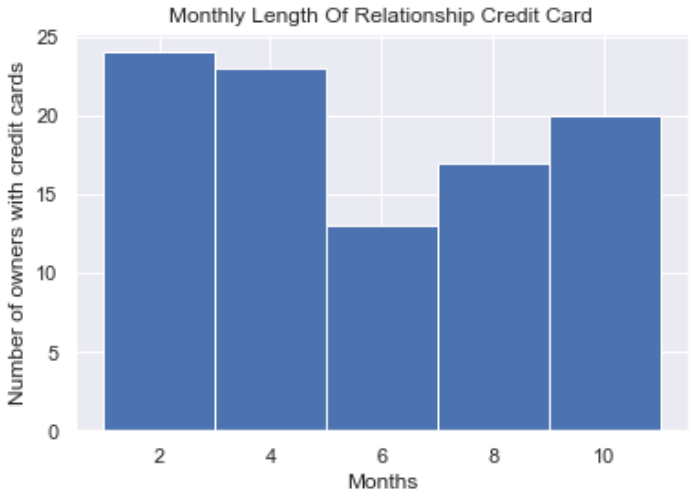
Length of Relationship in Years

The LOR distribution enables us to observe the number of new OWNERs and those that have been with the bank for multiple years. It would be interesting to further analyse why there is a drop in the 2nd year compared to 3 and 1 years ago. Furthermore, we can observe that a lot of clients have been with the Bank a longer amount of time.



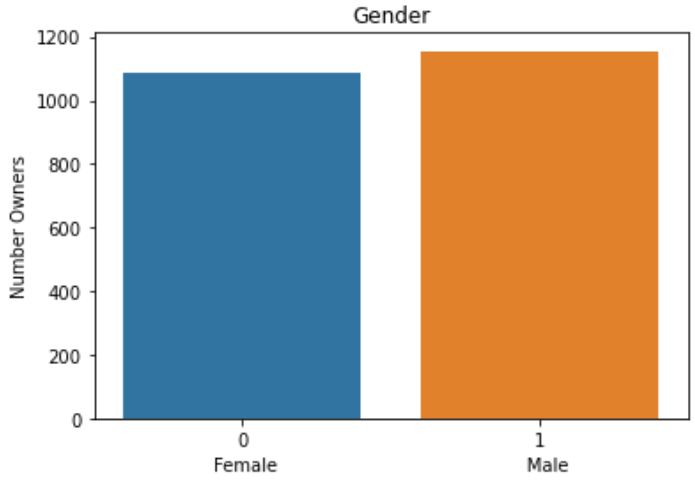
Credit Cards Length of Relationship in Months

The credit card LOR distribution enables us to observe the number of OWNERs that have a current credit card along with the number of months since they first started the credit card program. We can observe a strong need in the last few months with a dip 6 months back. THe bank should further analyse factors which could have influenced the dip including poor product offerings or high interest rates in order to increase the number of credit card users.



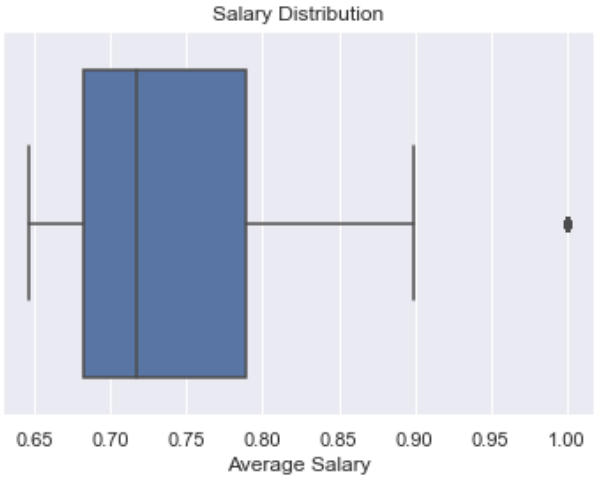
Gender Distribution

The gender distribution graph, as named, shows the number of Male and Female OWNERs within the bank. We can observe that there is an even distribution between both genders with a difference of approximately 100 OWNERs more that are male. In turn, the bank can outline the need to focus on both genders when developing product offerings and other such incentives.



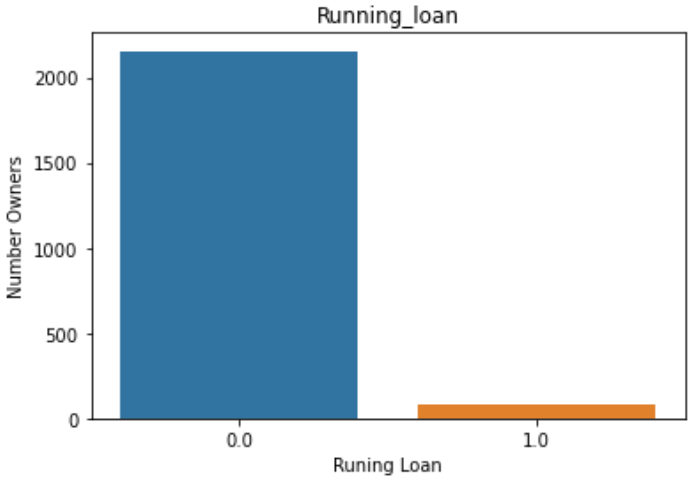
Average Salary Distribution

The distribution of average salary among districts shows that most areas have a similar average salary ranging from late 60s to the late 80s. Furthermore, we can observe that values are closer to the lower end of the distribution with higher end salaries much lower. This can be attributed to Pragues wealth in comparison to the rest of the districts throughout the Czech Republic.



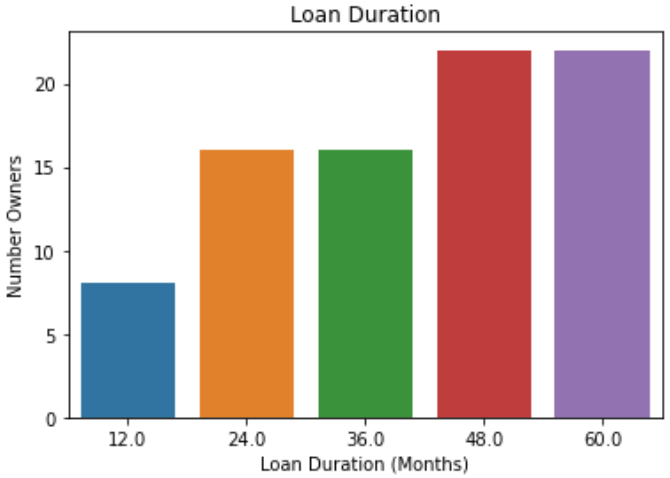
Number of running loans

This graph allows us to understand owner behaviours within the bank. We can observe that a small number of OWNERS have taken a loan within the bank. This graph can be interpreted by the fact that a very small number of clients take loans at the bank. The bank should find ways in which the clients without loans could be incentivized to take loans with the bank.



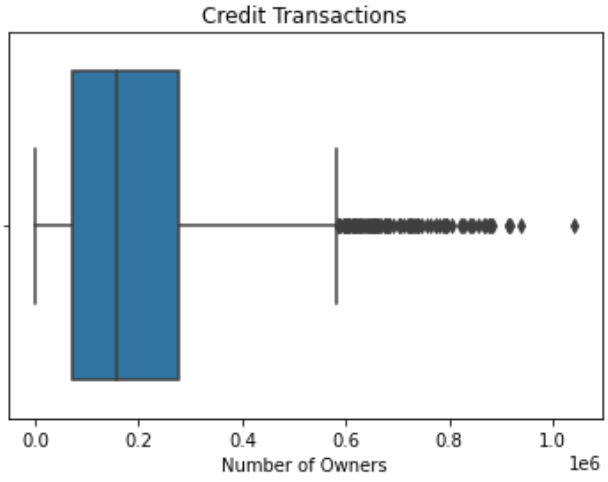
Duration of running loans

This table compliments the previous table regarding OWNERs that have taken out loans. This shows the duration left of their loan payments. We can observe that a high number of loanees still have a long time to pay off their loans.

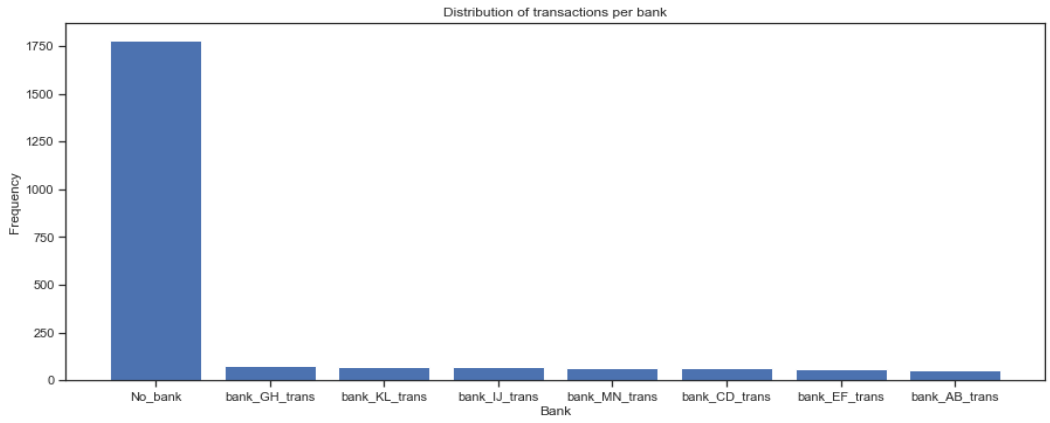


Number of credit transactions

The transaction boxplot provides a visual for the number of owners that perform credit transactions within the bank. Furthermore, we can see that the majority of OWNERs have small amounts of credit transactions, nevertheless, the large amount of outliers is not considered within the boxplot, but can be attributed to an increased amount of credit transactions for a minority group of OWNERs.

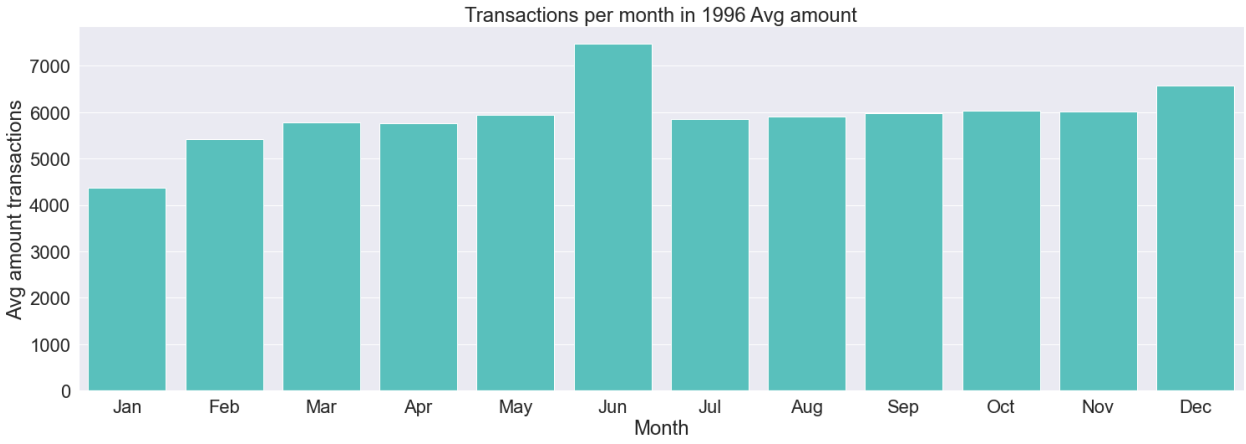


Distribution of transactions per bank



When analysing the amount of transactions per bank it is possible to observe a similar distribution for all the banks but it is also important to mention that most of the transactions do not correspond to banks.

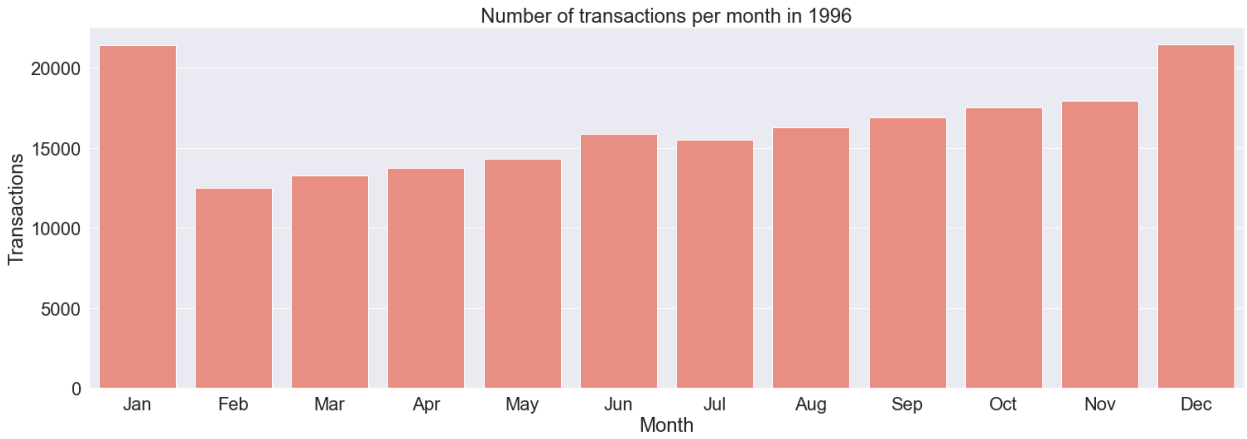
Distribution of avg. amount of transactions per month



The distribution of the average amount in transactions provides an illustration of monthly client behaviours with regards to increased months of purchases. We can see the obvious illustrators with regards to an increased spending in December surrounding the month of gift purchases for the holiday festivities. However, we can see that June has the highest amount of transactions, which could correlate to incoming and outgoing funds due to it being the end of the financial year.

Furthermore, it could also be due to summer holiday spending, however, to determine the exact reasons for increased amounts spent the bank would need to analyse further.

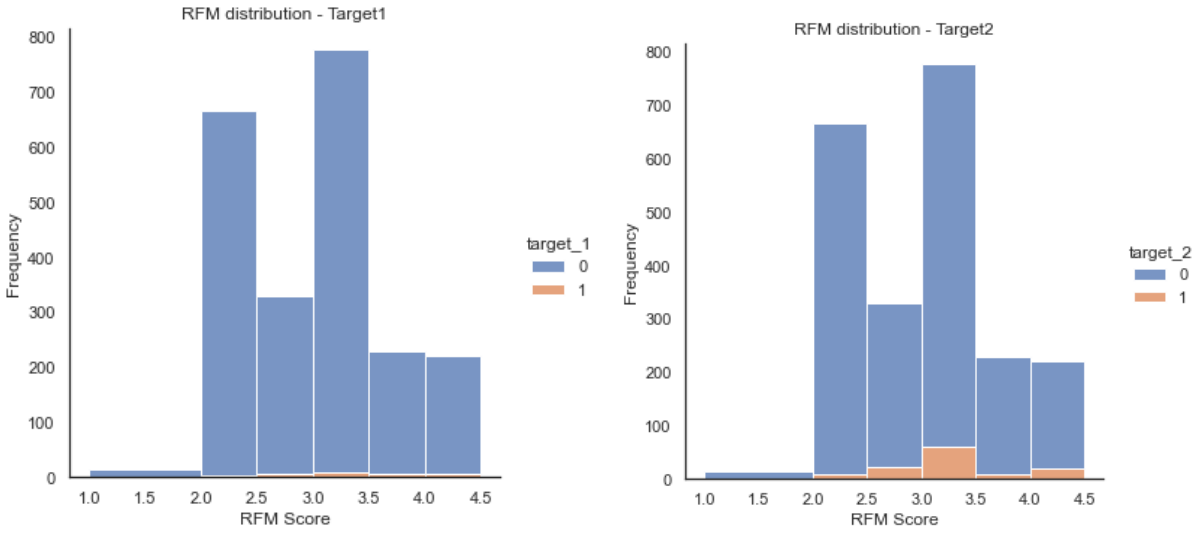
Distribution of transactions per month



Similarly to the previous graph regarding the average value amounts of transactions, this graph focuses on the number of transactions per month for the year of 1996. Once again, we can observe a peak in December which can be attributed to increased transactions surrounding purchasing behaviours of the gift giving season. Nevertheless, contrary to the previous graph, we can see that the highest amounts of transactions occurred in January.

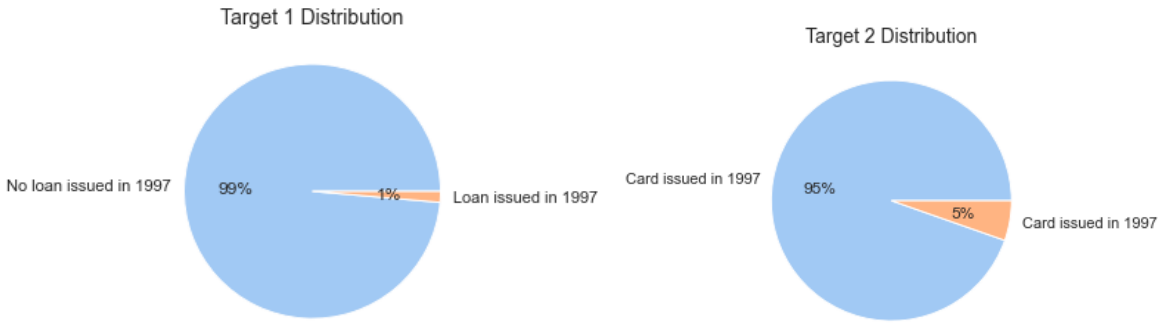
We can also state that although it recorded the highest amount of transactions, these transactions must have been for small amounts due to the lowest figure amount on average of transactions for the same month in the previous graph. The contradictory bars should be analysed further by the bank to see whether there are transaction issues or inefficiencies they could improve in order to increase customer satisfaction or provide customer incentives.

RFM Score distribution vs targets



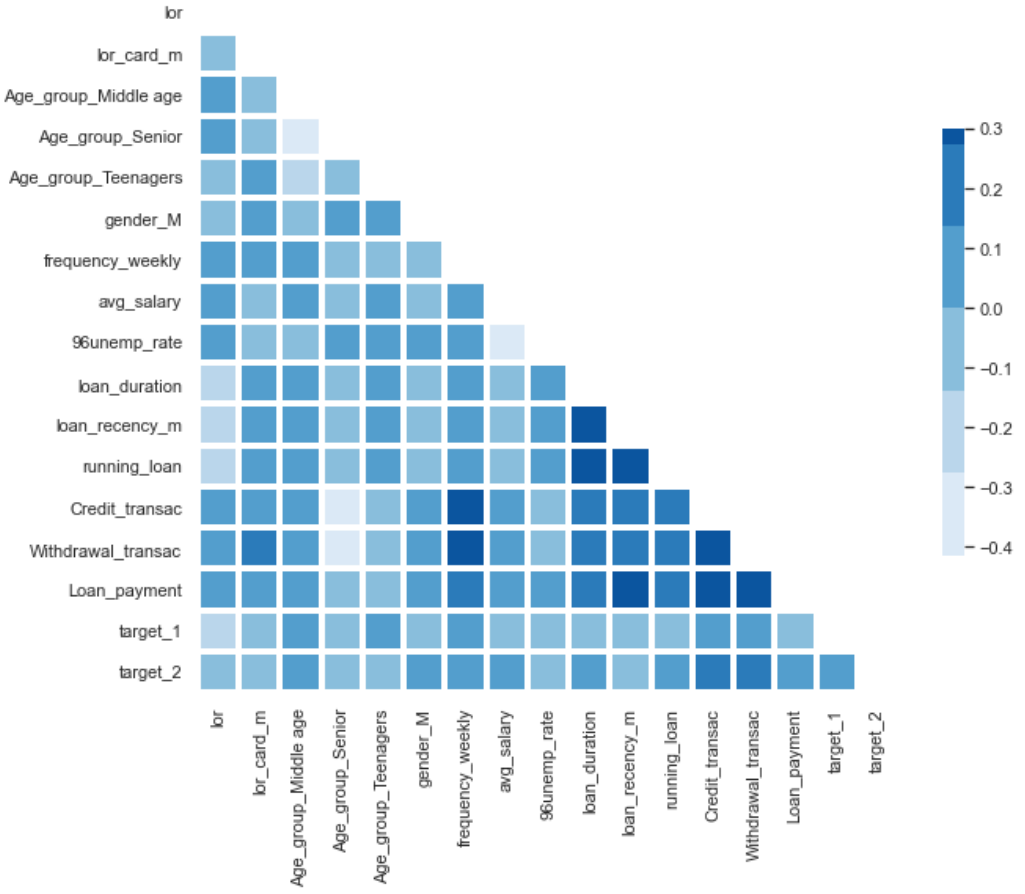
According to the charts it is possible to notice that most of the clients have an rfm score between 2 and 2.5 or 3 and 3.5. Also, the clients with score between 3 and 3.5 are more propense to issue a card than to grant a loan in 1997.

Targets distribution



For the target variables it is possible to observe that the clients who granted a loan are 1% and on the other hand 5% of the clients issued a card in 1997. It is also important to mention that a client can have just 1 loan at the same time and this can cause the number of loans to be smaller.

Correlation matrix for relevant variables



In order to understand better the relation between the variables, the correlation matrix presented shows that variables related with type of transaction have a negative correlation close to 0.4 with age groups and a positive correlation of 0.3 with weekly frequency of issuance. However, the correlation between the variables presented is in general small.