Data Mart Creation

Group 5: Abdon Ahuile, Jorge Polanco and Rishabh Sharma

Introduction

The purpose of the project is to create a financial Datamart based on the information available with the bank about their clients. The bank is based in Czech Republic. The datamart is to be a summary of the data taken from each of these tables in order to obtain insights about the clients and allow the bank to make further business decisions about the clients.

Data processing

1. Treatment of original tables

a) Client:

For the clients table, the age and age group was calculated and added as an additional variable so that it could be used to make demographic segmentations with these values. All the clients (Client_id) is kept for the final data mart.

b) Account:

The frequencies names in this table where replaced by their equivalent in the english language. Also the year month and date of when the account was created is extracted.

c) Card:

In card the year, month and day of when the card was issued is created and added as an additional variable.

d) Transaction:

Missing values in this table where replaced with a string “Missing” and the observations in the variables type, k_symbol and operation is translated into english.

e) Loan:

In loan the observations in the “Status” variable where changed to something more explicative, done with values from a dictionary.

f) Order

The observations in K_symbol where changed to english and blank spaces with a string called “Missing”.

g) Disposition:

From this table the amount of disponents per owners is extracted.

2. Final Data Mart

The final data mart is created by merging the already cleaned original tables and by creating new variables from this merging, this is a data mart that contains one observation per client.

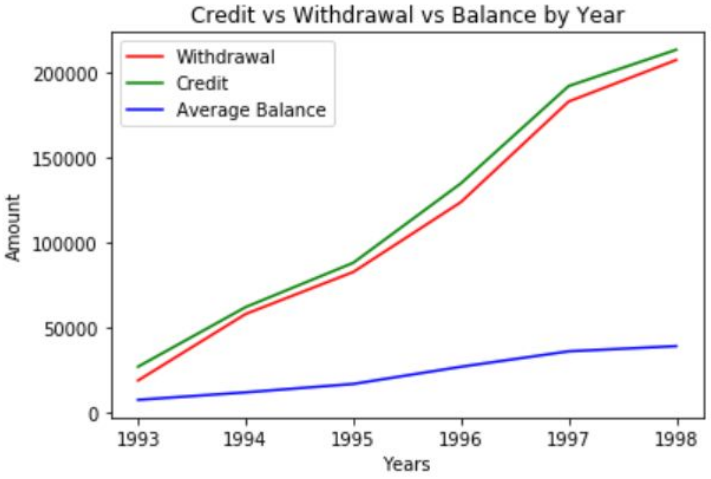
Amount of variables: **300**
Amount of observations: **5369**

As an additional element a grand average vector (grand_avg) is created so that it is useful for analysis with variables that are counts and sums, keep in mind that this vector is extracted directly from the final data mart.

List of Variables					
'client_id'	'sum_Withdrawal_8'	'count_1994_11'	'sum_1993_2'	'sum_1997_5'	'UV'
'birth_year'	'sum_Withdrawal_9'	'count_1994_12'	'sum_1993_3'	'sum_1997_6'	'WX'
'birth_day'	'sum_Withdrawal_10'	'count_1995_1'	'sum_1993_4'	'sum_1997_7'	'YZ'
'birth_month'	'sum_Withdrawal_11'	'count_1995_2'	'sum_1993_5'	'sum_1997_8'	'count_HouseholdPayment'
'age'	'sum_Withdrawal_12'	'count_1995_3'	'sum_1993_6'	'sum_1997_9'	'count_InsurancePayment'
'age_group'	'count_Credit_1993'	'count_1995_4'	'sum_1993_7'	'sum_1997_10'	'count_Leasing'
'gender'	'count_Credit_1994'	'count_1995_5'	'sum_1993_8'	'sum_1997_11'	'count_LoanPayment'
'account_id'	'count_Credit_1995'	'count_1995_6'	'sum_1993_9'	'sum_1997_12'	'count_Missing'
'count_Credit_1'	'count_Credit_1996'	'count_1995_7'	'sum_1993_10'	'sum_1998_1'	'sum_HouseholdPayment'
'count_Credit_2'	'count_Credit_1997'	'count_1995_8'	'sum_1993_11'	'sum_1998_2'	'sum_InsurancePayment'
'count_Credit_3'	'count_Credit_1998'	'count_1995_9'	'sum_1993_12'	'sum_1998_3'	'sum_Leasing'
'count_Credit_4'	'count_Withdrawal_1993'	'count_1995_10'	'sum_1994_1'	'sum_1998_4'	'sum_LoanPayment'
'count_Credit_5'	'count_Withdrawal_1994'	'count_1995_11'	'sum_1994_2'	'sum_1998_5'	'sum_Missing'
'count_Credit_6'	'count_Withdrawal_1995'	'count_1995_12'	'sum_1994_3'	'sum_1998_6'	'loan_amount'
'count_Credit_7'	'count_Withdrawal_1996'	'count_1996_1'	'sum_1994_4'	'sum_1998_7'	'loan_duration'
'count_Credit_8'	'count_Withdrawal_1997'	'count_1996_2'	'sum_1994_5'	'sum_1998_8'	'loan_emi'
'count_Credit_9'	'count_Withdrawal_1998'	'count_1996_3'	'sum_1994_6'	'sum_1998_9'	'loan_status'
'count_Credit_10'	'sum_Credit_1993'	'count_1996_4'	'sum_1994_7'	'sum_1998_10'	'loan_year'
'count_Credit_11'	'sum_Credit_1994'	'count_1996_5'	'sum_1994_8'	'sum_1998_11'	'loan_month'
'count_Credit_12'	'sum_Credit_1995'	'count_1996_6'	'sum_1994_9'	'sum_1998_12'	'loan_day'
'count_Withdrawal_1'	'sum_Credit_1996'	'count_1996_7'	'sum_1994_10'	'count_Credit'	'count_DISPONENT'
'count_Withdrawal_2'	'sum_Credit_1997'	'count_1996_8'	'sum_1994_11'	'count_Withdrawal'	'disp_type'
'count_Withdrawal_3'	'sum_Credit_1998'	'count_1996_9'	'sum_1994_12'	'sum_Credit'	'card_type'
'count_Withdrawal_4'	'sum_Withdrawal_1993'	'count_1996_10'	'sum_1995_1'	'sum_Withdrawal'	'loan_year'
'count_Withdrawal_5'	'sum_Withdrawal_1994'	'count_1996_11'	'sum_1995_2'	'mean_balance_1993'	'loan_month'
'count_Withdrawal_6'	'sum_Withdrawal_1995'	'count_1996_12'	'sum_1995_3'	'mean_balance_1994'	'loan_day'
'count_Withdrawal_7'	'sum_Withdrawal_1996'	'count_1997_1'	'sum_1995_4'	'mean_balance_1995'	'frequency'
'count_Withdrawal_8'	'sum_Withdrawal_1997'	'count_1997_2'	'sum_1995_5'	'mean_balance_1996'	'card_year'
'count_Withdrawal_9'	'sum_Withdrawal_1998'	'count_1997_3'	'sum_1995_6'	'mean_balance_1997'	'card_month'
'count_Withdrawal_10'	'count_1993_1'	'count_1997_4'	'sum_1995_7'	'mean_balance_1998'	'card_day'
'count_Withdrawal_11'	'count_1993_2'	'count_1997_5'	'sum_1995_8'	'mean_Credit'	'A1'
'count_Withdrawal_12'	'count_1993_3'	'count_1997_6'	'sum_1995_9'	'mean_Withdrawal'	'A2'
'sum_Credit_1'	'count_1993_4'	'count_1997_7'	'sum_1995_10'	'Household'	'A3'
'sum_Credit_2'	'count_1993_5'	'count_1997_8'	'sum_1995_11'	'Insurancepayment'	'A4'
'sum_Credit_3'	'count_1993_6'	'count_1997_9'	'sum_1995_12'	'Interestcredited'	'A9'
'sum_Credit_4'	'count_1993_7'	'count_1997_10'	'sum_1996_1'	'Loanpayment'	'A10'
'sum_Credit_5'	'count_1993_8'	'count_1997_11'	'sum_1996_2'	'trans_k_symbol_missing'	'A11'
'sum_Credit_6'	'count_1993_9'	'count_1997_12'	'sum_1996_3'	'Oldagepension'	'A12'
'sum_Credit_7'	'count_1993_10'	'count_1998_1'	'sum_1996_4'	'Paymentforstatement'	'A13'
'sum_Credit_8'	'count_1993_11'	'count_1998_2'	'sum_1996_5'	'Sanctioninterest'	'A14'
'sum_Credit_9'	'count_1993_12'	'count_1998_3'	'sum_1996_6'	'AB'	'A15'
'sum_Credit_10'	'count_1994_1'	'count_1998_4'	'sum_1996_7'	'CD'	'A16'
'sum_Credit_11'	'count_1994_2'	'count_1998_5'	'sum_1996_8'	'EF'	'mean_Balance'
'sum_Credit_12'	'count_1994_3'	'count_1998_6'	'sum_1996_9'	'GH'	
'sum_Withdrawal_1'	'count_1994_4'	'count_1998_7'	'sum_1996_10'	'IJ'	
'sum_Withdrawal_2'	'count_1994_5'	'count_1998_8'	'sum_1996_11'	'KL'	
'sum_Withdrawal_3'	'count_1994_6'	'count_1998_9'	'sum_1996_12'	'MN'	
'sum_Withdrawal_4'	'count_1994_7'	'count_1998_10'	'sum_1997_1'	'OP'	
'sum_Withdrawal_5'	'count_1994_8'	'count_1998_11'	'sum_1997_2'	'Otherbank'	
'sum_Withdrawal_6'	'count_1994_9'	'count_1998_12'	'sum_1997_3'	'QR'	
'sum_Withdrawal_7'	'count_1994_10'	'sum_1993_1'	'sum_1997_4'	'ST'	

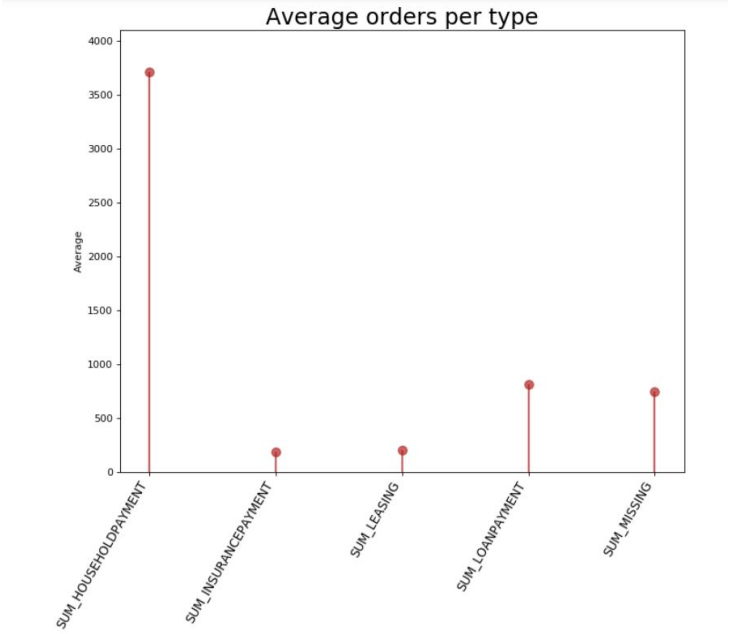
Insights

Growth over the years



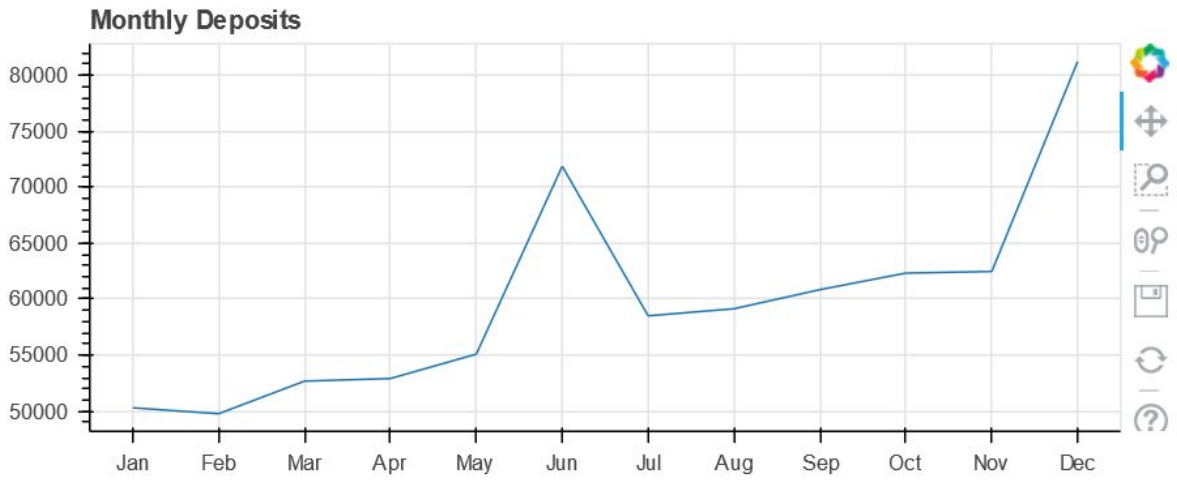
There is a clear rise in the average balance maintained by the clients indicating that the bank is performing well in attracting more and more deposits from its customers.
Moreover, the trajectory for both withdrawal and credit amount is similar during the last 5 years. This indicates that the bank is not coming under duress due to any major deviations between withdrawals and deposits

Average orders per type:



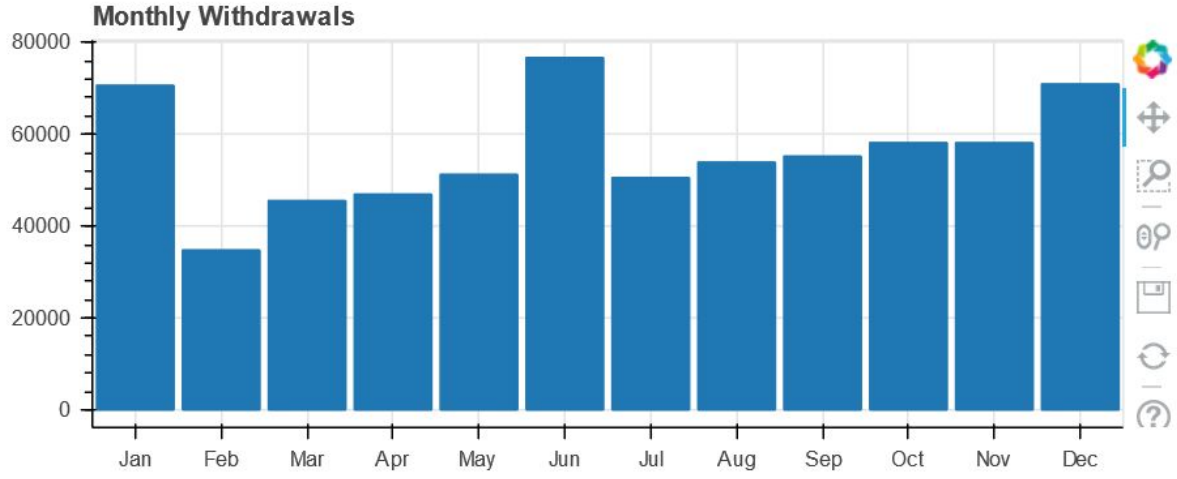
As it can be observed in the lollipop chart above, the household payment represents the most significant type of order for the bank, overcoming the rest of the orders by a relevant amount. Therefore clients in this group could be the focus for further analysis in order to watch their behaviour throughout the year so that they are classified as either good or bad clients.

Seasonality in Deposits and Withdrawals

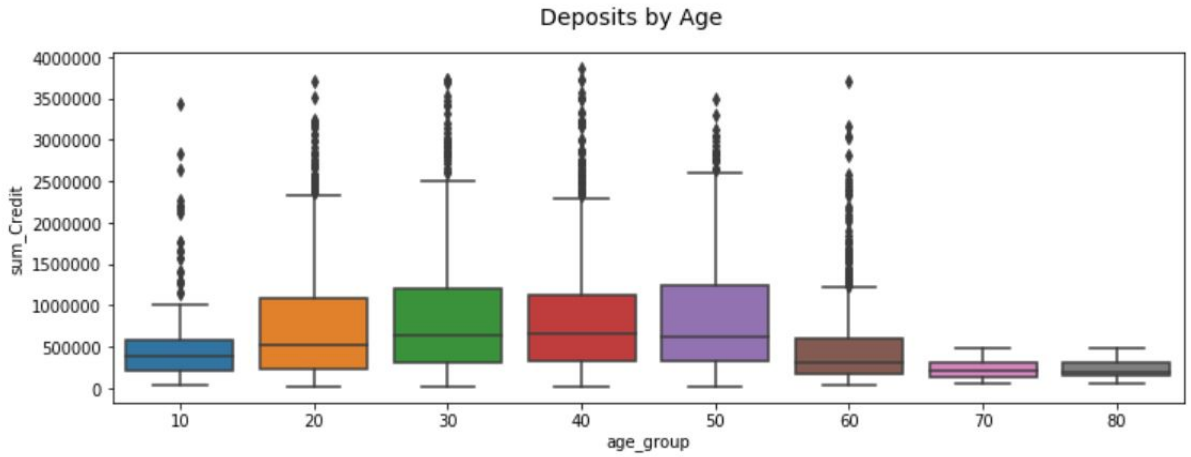


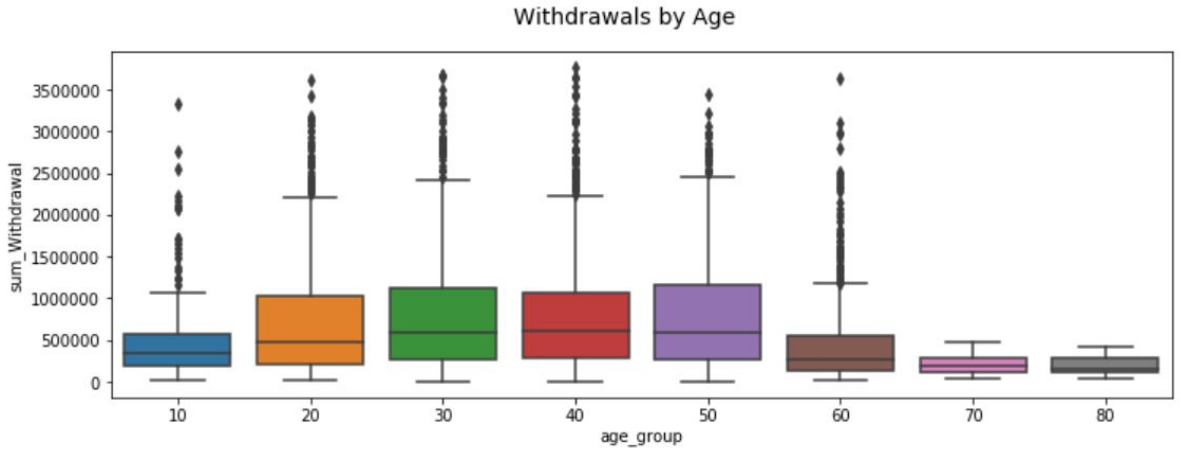
The chart above tracking average deposits made in each month over the given period 1993 to 1998 shows that there is a clear upward trend from January to December indicating that people save more as they go through the full year. There is also a surge in the deposit amounts during the months of June and December. We suspect the June surge could be because of salary revisions or bonus payments. The december surge could be for the same reason and it being a holiday season.

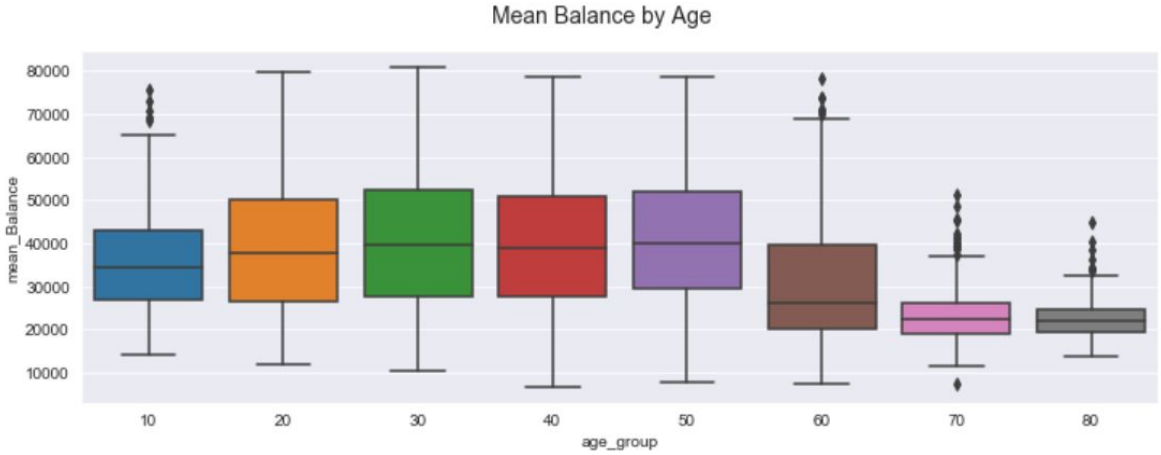
For withdrawals, we see that there are a lot more withdrawals in January compared to the deposits. But, we do see a similar surge in the month of June. Similar to the deposits, the monthly withdrawals also steadily rise as customers progress through the year.



Account Activities by Age

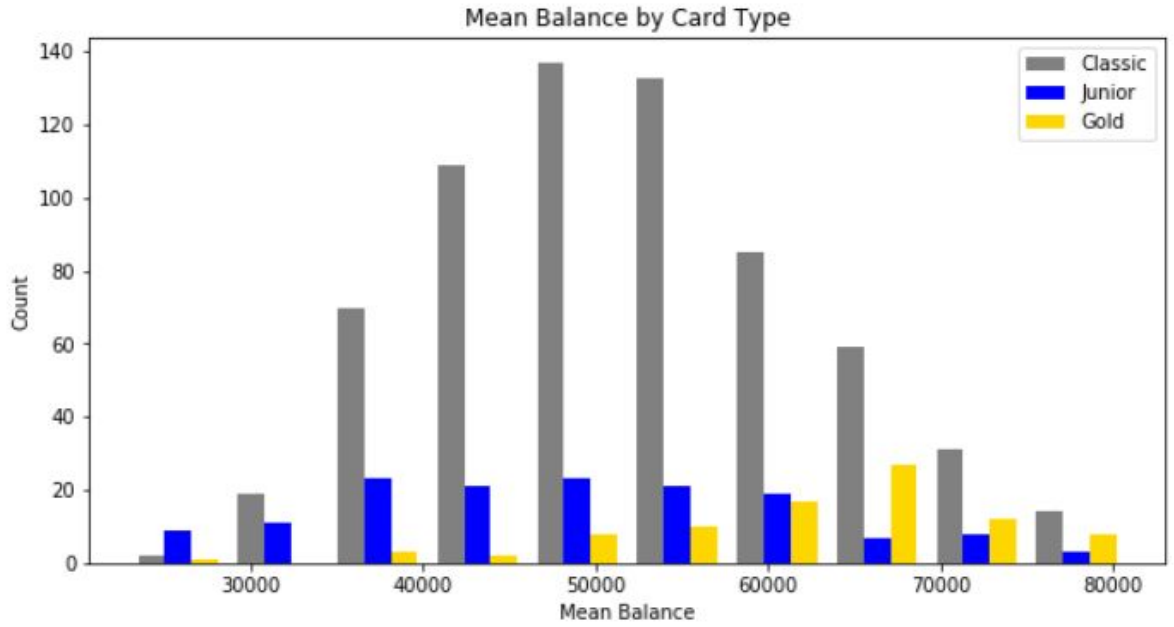






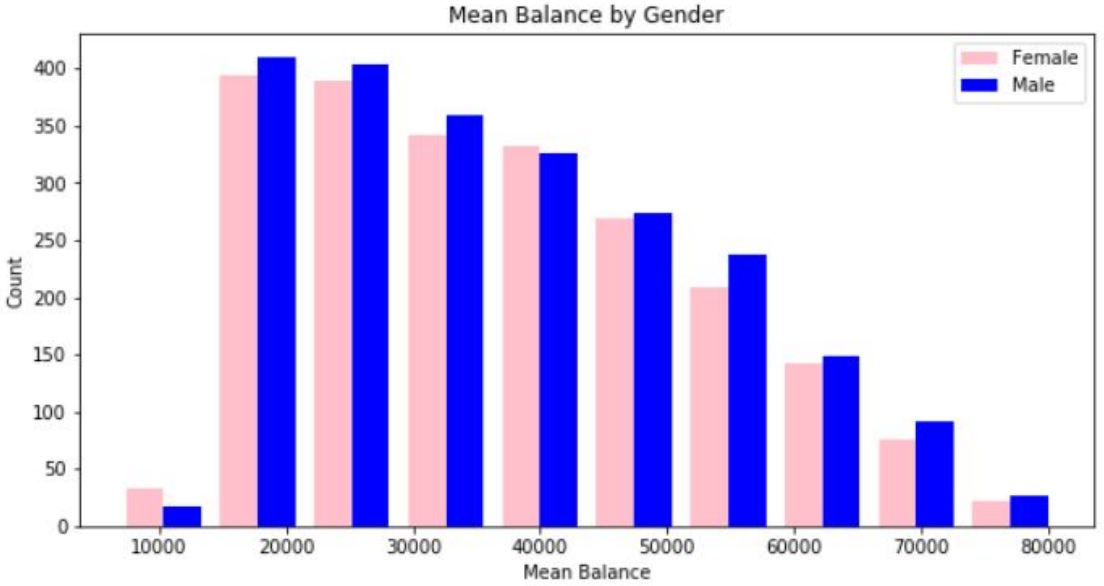
The above graphs show that among the customers older than 60 years of age, the account activity is really low. Most of them do not maintain very high balances and do smaller amounts of withdrawals and deposits. Clients in their 40s have the highest median deposit at CZK 656,902 and highest median withdrawals at CZK 611,761. In terms of mean annual balance, clients in the 50s have the highest median at CZK 40,009.

Balance by card type



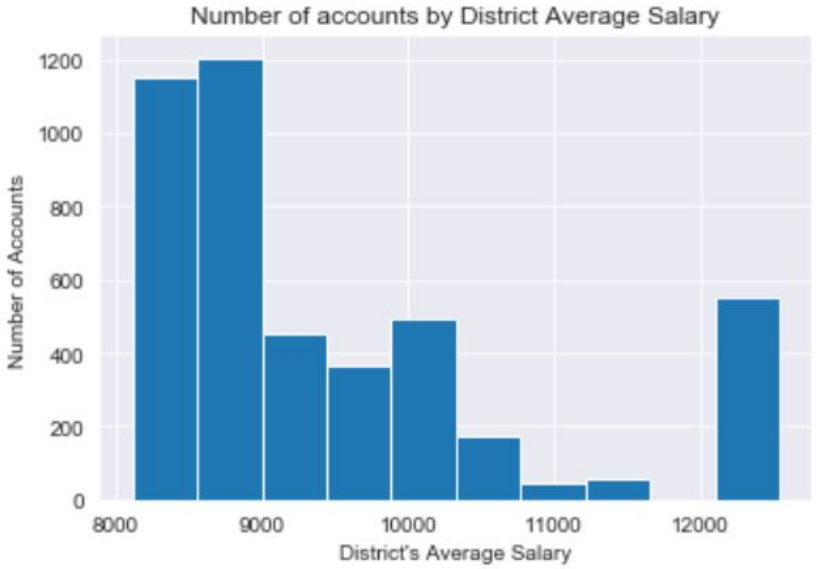
As can be observed above, the golden card holders maintain the highest mean balance over the year. Between the Classic and Junior card holders, there is no big difference in the mean balance among the card holders.

Balance by gender

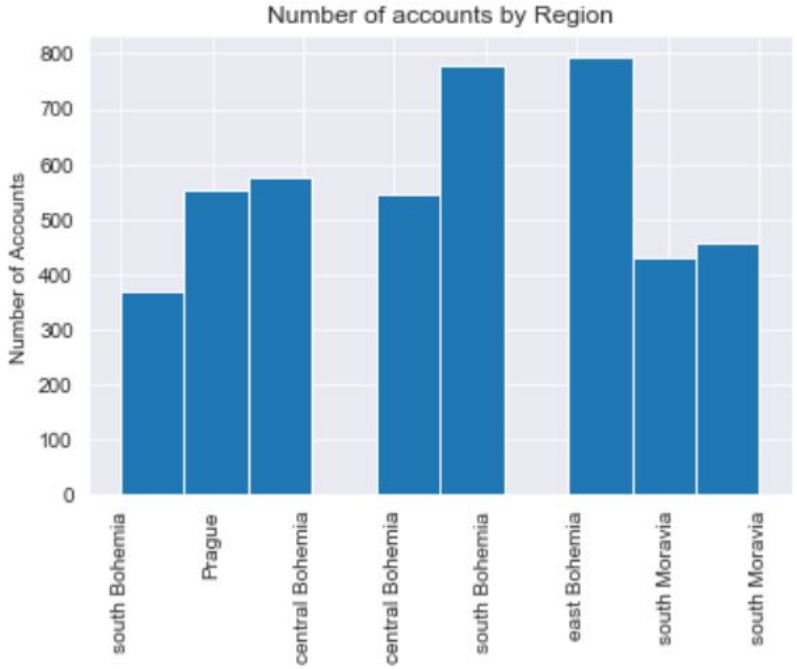


From this graph we can observe that when it comes to accounts with low mean annual balance, more of such accounts belong to females than males. Except the lowest balance category, in all other accounts, the males account holders outnumber female account holders.

Demographic

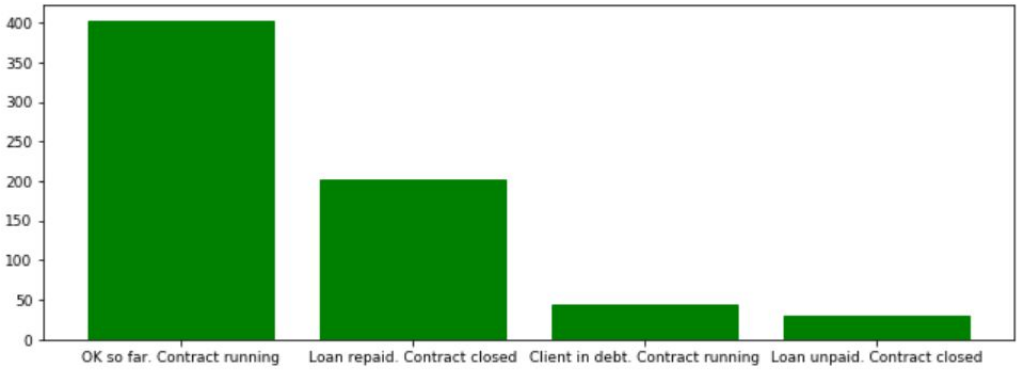


Most of the account holders of the bank come from low income districts of the country. There are a number of clients from the high income districts as well. Overall, the distribution is skewed towards the bottom and top ranges.



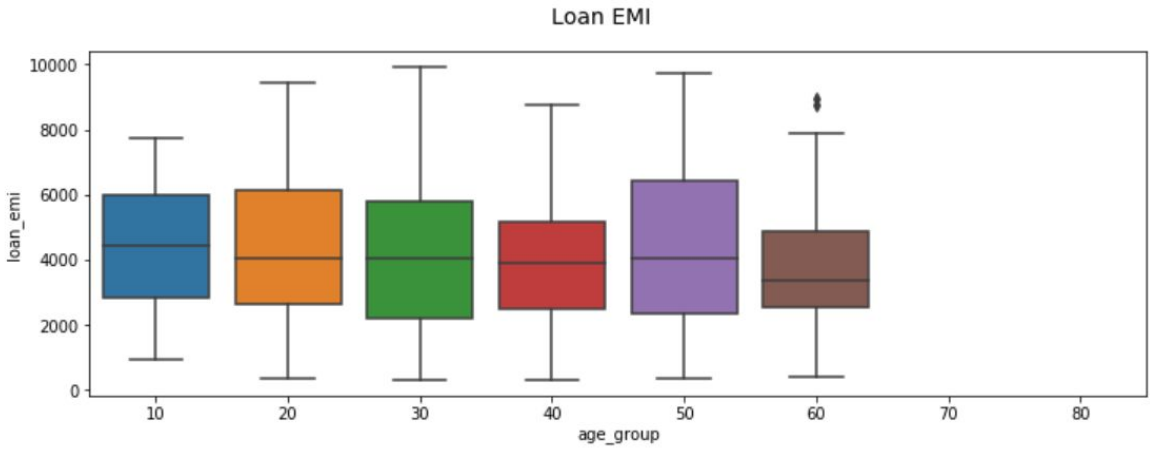
Most of the clients come from the South Bohemia and East Bohemia region, followed by Central Bohemia and Prague.

Loans Analysis:



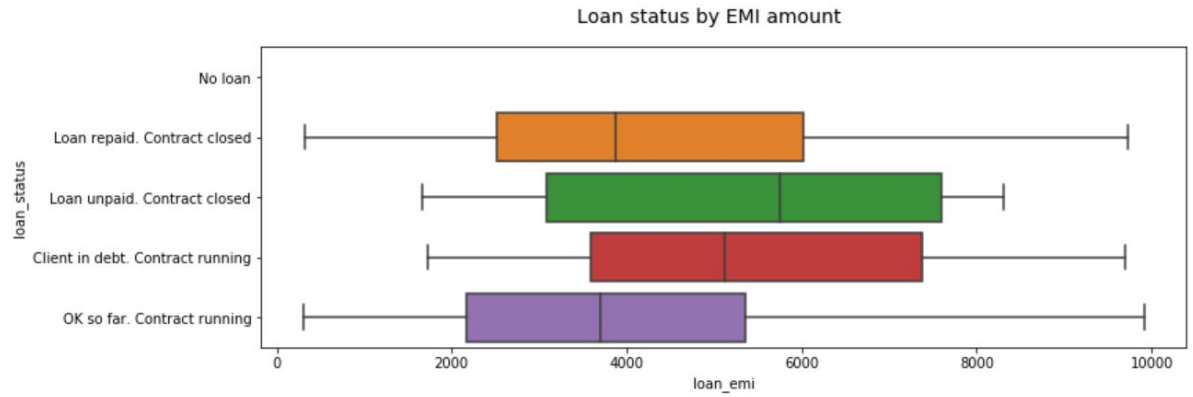
Only 15% of the clients have taken a loan from the bank. There is a business opportunity to target customers to whom the bank can grant loans. Over the last few years, the balance maintained by clients has gone up. It indicates that there are customers with repayment capacity who can be targeted for cross selling of loan products.

Loan burden (EMI) by age group:



The range of equated monthly installment (EMI) for loans taken by clients who are in their teens is the narrowest and by those in their 30s is the widest. At the same time, the median EMI amount for loans to teen clients is the highest. This could be because the loan purpose for younger clients is limited and high denomination (e.g. Education loans).

Loan Status by EMI

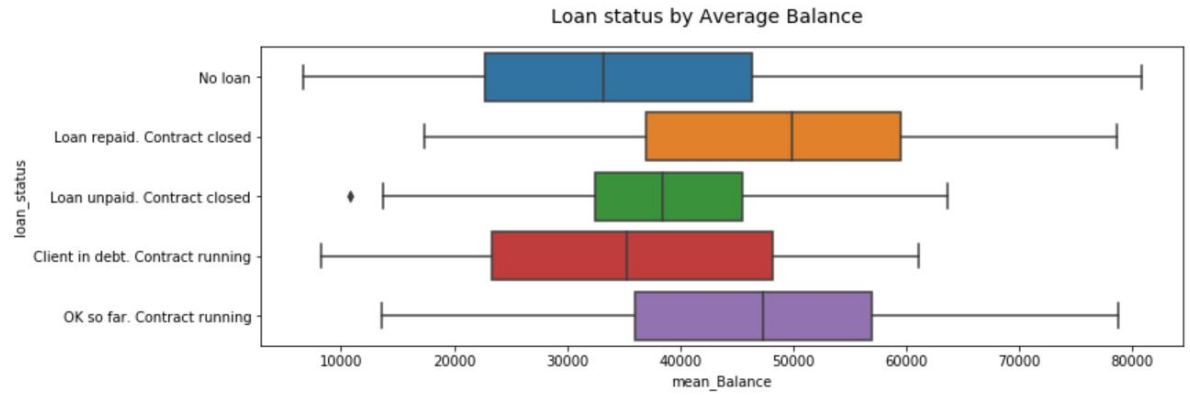


The loans which were unpaid and the contract was closed, i.e. they were classified as defaults can be seen to have higher the highest median EMI (CZK 5,746). The range of such loans is narrow and high. On the other hand, the loans which were repaid in full by the clients had comparatively lower median EMI (CZK 3,874). The same pattern can be seen in loans which are open at the moment with the bank: The loans with no issues yet have a median EMI of CZK 3,698 versus those where the client is in debt have a median EMI of 5,120.

Another interesting finding is that the loans which were declared as defaults as well as the ones which are having issues at the moment have a high minimum EMI (CZK 1,671 and CZK 1,728 respectively). The loans which perform well have very low minimum EMI (CZK 319 for repaid loans and CZK 304 for open loans with no issues).

Recommendation: The bank should be more careful giving out loans which have higher EMI as it could become difficult to repay them back.

Loan Status by Average balance



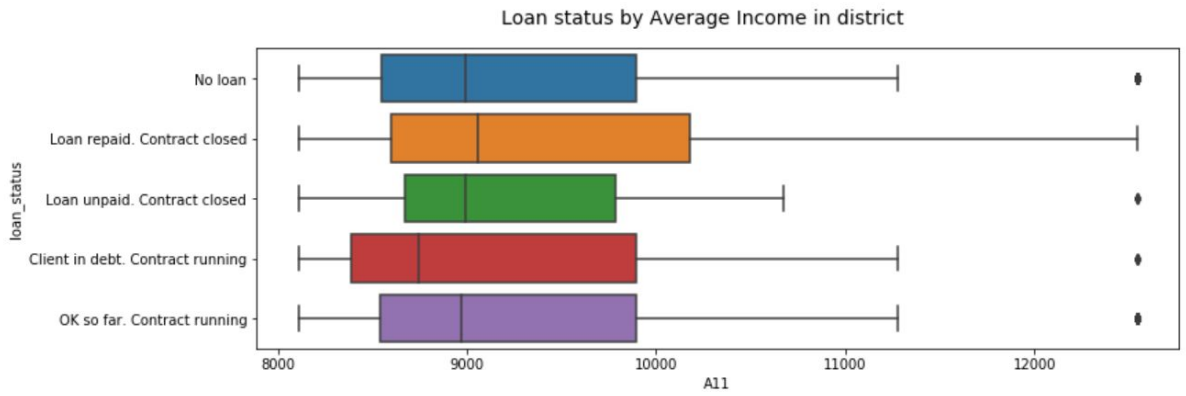
Analyzing further the possible indicators of loan performance, we look at the average balance maintained by the clients who have taken loans from the bank. As we can see below, the loans repaid on time or are being serviced by their borrowers without any issues are to clients who maintain high annual balance in their accounts. Loans repaid on time were given to clients with a median of CZK 49,907 as balance in their account. The loans running without any issues belong to clients with a median of CZK 47,343 in their account.

Loans which perform badly have distinctly low balance, with the minimum balance being maintained by a client who defaulted at CZK 10,812 versus a client that repaid back fully had a minimum balance of CZK 17,351.

The loans closed as defaults by the bank have a median balance of CZK 38,379 and those ongoing with issues have a median balance of CZK 35,295.

Recommendation: The bank should be more careful while giving loans the clients who maintain low balance on their account.

Loan status by Average Income in District

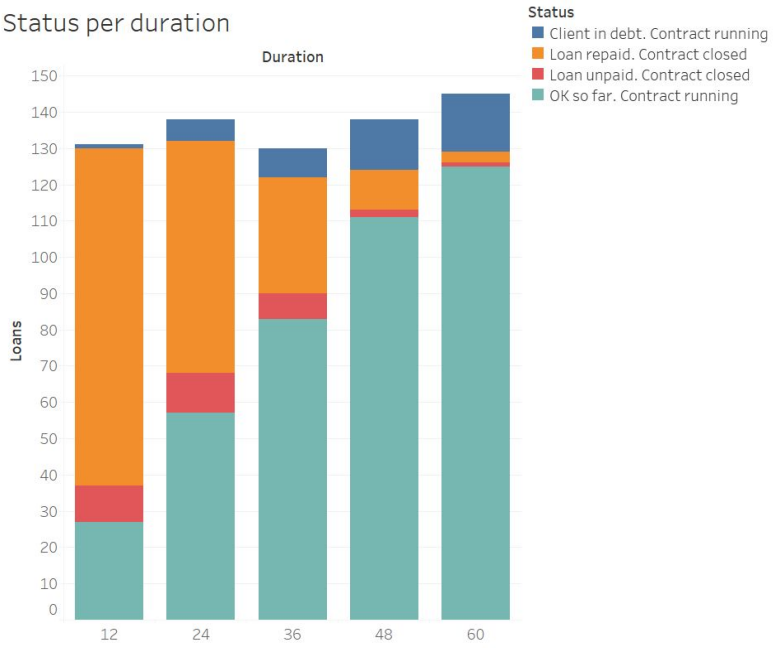


The chart below shows that the bank gives loans to clients living in districts with a range of average income. However, the loans which are ongoing but have issues, are given to clients in districts with a median average income of CZK 8,746.

On the other hand, the loans which were repaid successfully were given to clients living in districts where the median of the average income is CZK 9,060. The chart also shows that the loans that were repaid in full have no outlier average income districts, which implies that the distribution is tighter.

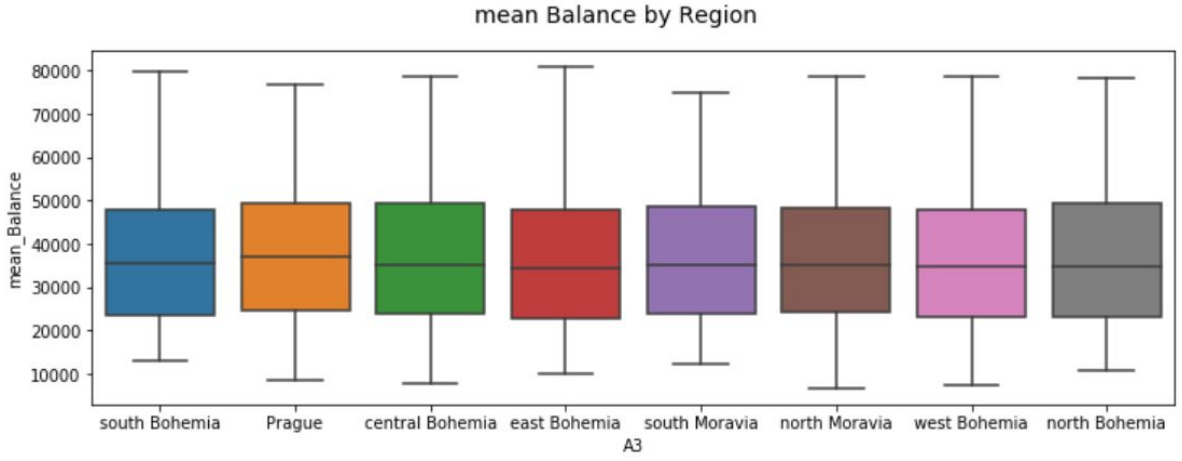
Recommendation: the bank should keep its loan portfolio diverse enough so as to avoid having exposure to any one particular district.

Loan Status by Region



For this case it can be observed that for loans with high durations the behaviour of the clients is good ("Ok so far, contract, running") as it can be seen that there is a positive correlation between the increase in duration and good behaviour. On the other hand it can also be observed a slight increase of clients in debts ("Client in debt.Contract running") for higher durations but also there are more unpaid loans in loans with lower durations ("Loan unpaid. Contract closed").

Mean Balance by Region



Clients from the Prague region maintain the highest median balance among all the regions (CZK 36,981). On the other hand, clients from the East Bohemia region maintain the lowest median balance among all the regions (CZK 34,429).

There are no major differences between different regions when it comes to the balance maintained by the clients residing there.

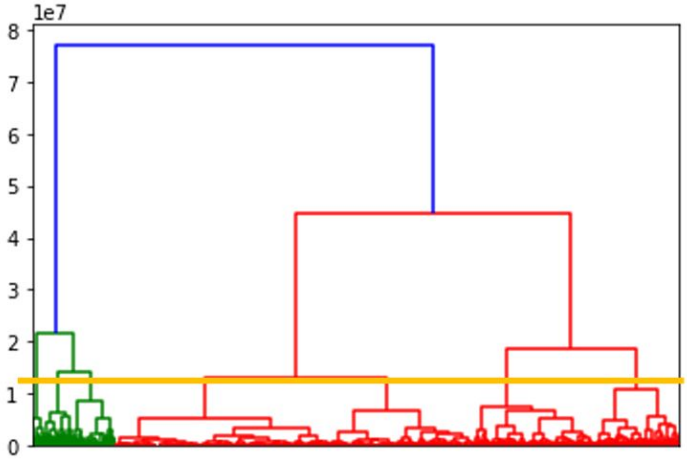
Segmentation of clients

In order to identify potential segments, which could help us to create more targeted marketing strategies, we ran a Cluster Analysis model. We ended up with three different models: a) all features, b) behaviour features and c) demographic features.

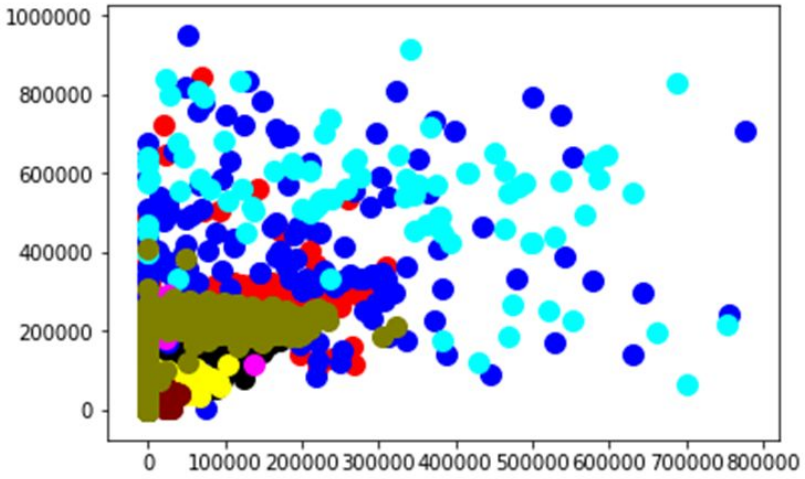
Finally, we decided to stick to the behaviour model because it made more sense from a commercial perspective. Besides, in the cluster plotting process of each one of the clients, it was easier to separate visually.

The features selected for this model were: 'sum_Credit_1993', 'sum_Credit_1994', 'sum_Credit_1995', 'sum_Credit_1996', 'sum_Credit_1997', 'sum_Credit_1998', 'sum_Withdrawal_1993', 'sum_Withdrawal_1994', 'sum_Withdrawal_1995', 'sum_Withdrawal_1996', 'sum_Withdrawal_1997', 'sum_Withdrawal_1998', 'mean_Credit_1993', 'mean_Credit_1994', 'mean_Credit_1995', 'mean_Credit_1996', 'mean_Credit_1997', 'mean_Credit_1998', 'mean_Withdrawal_1993', 'mean_Withdrawal_1994', 'mean_Withdrawal_1995', 'mean_Withdrawal_1996', 'mean_Withdrawal_1997', 'mean_Withdrawal_1998','sum_Credit', 'sum_Withdrawal', 'mean_Credit', 'mean_Withdrawal','mean_Balance'.

For each particular case, the number of clusters was determined by a Dendrogram Analysis. In the following chart, the orange line shows the cutoff that was chosen to establish the number of clusters for the 'behaviour model'.



Once we had established the number of clusters, we proceeded with running the model. As previously mentioned, the following charts show us how the clients are distributed by cluster number.



Crossing the number cluster with some statistics metrics, we can show some differences between the behaviour of these groups. The following chart is only one example of the analysis that can be made with this tool.

num Cluster	age	sum_Credit	sum_Withdrawal
0	42	1668749.0	1612534.0
1	41	742642.0	691999.0
2	40	2267225.0	2197684.0
3	40	3052860.0	2983608.0
4	48	410547.0	368721.0
5	41	1219491.0	1151070.0
6	49	166592.0	136575.0
7	40	1225191.0	1163292.0

Further analysis can be developed after meeting with the commercial department of the bank and getting more clarity of the different hypotheses that could be tested on this model.