# Analysis of the Financial Banking data

## Group 1

**Report prepared by:**

Liu Yi

Anne Kirika

Priya Varadarajan

# Table of Contents

## Overview

The main goal of the project is to prepare a DataMart from financial banking data sets that can be used for predictive analysis to improve the bank's customer service or describe the different kind of clients the bank has. In addition to the data mart creation, some exploratory and descriptive visualizations will be done to provide some useful insights on the clients.

## Background

The data availed for the financial analysis is from a bank that offers services to private clients. Services include managing accounts, offering loans, etc. With the development of the bank, there is a need for them to better understand their clients to improve their services. The data mart and the visualisation analysis done in this project can help the bank to improve decision making also proposing new services to clients.

Key focus on the analysis is to draw a user picture or 360-degree view of the customer based on all available data i.e. who is a good client and what additional services he can be provided with and who are bad clients to watch carefully to minimize the bank loses.

For the analysis the following data sets were provided each bearing descriptions of the data it contained.

- Account
- Client
- Disposition
- Permanent Order
- Transaction
- Loan
- Credit Card
- Demographic data

The data mart creation was broken down into chunks: -

1. Exploring the different datasets and identifying the relationship between them to check which tables could be merged.
2. Formulation of questions that needed to be answered while creating the variables to be used in the data mart.
3. Variable extraction and data mart creation.
4. Exploratory and descriptive graphical presentation of the data mart.

## Data Exploration:

A brief overview of datasets is as follows:

- **Loan data** that consists of accounts that have been granted loans, the duration, amount, date loan was granted, monthly payments and loan status.
- **Account data** that consists of account id, date of account opening, location where the account was opened and the frequency of account statements.
- **Client data** that consists of client id, birth date and district id.
- **Disposition data** that consists of account id, disposition id, client id and type of account.
- **Permanent Order data** that consists of order id, account id, bank to, account to debit amount and transaction type**.**
- **Credit card** data that consists of card id, disp id, type and date of card issue
- **Demographics data** that consists of figures of its client's districts
- **Transactions data that consists** of transaction dates, amounts, type of transaction, mode of transaction, balance after transaction, transaction amount, bank of partner and account of the partner.

## Variable creation and extraction

In order to know your customer its best practice to have KYC (Know your customer data) for each customer. Given the current setting most of the client information was in silos (different datasets) hence we started the exercise by formulating questions that we thought were essential in capturing key information about a client. We also decided to have our data mart to have one row per client that would give an overview of every client details.

In order to achieve this, we came up with the below questions that drove us to create the variables for our data mart.

- What kind of client demographics data could we derive? (Age, Gender, Region)
- When did the client open the first account with the bank?
- Which kind of statement(s) does the client receive from the bank? (Weekly, Monthly, Transactions)
- What is the customer value segmentation based on transactions? (Total credits, Total withdrawals)
- What kind of client disposition do we have based on account details? (Owner/Disponent)
- What kind of transactions are the clients making and their frequencies? (insurance payment, payment for statement, interest credited, sanction interest if negative balance, household, old-age pension, loan payment)

- How many automatic debits/standing orders does a client have?
- What is the clients current balance? (Balance as of max date in transaction table)
- When does the client do most of his/her transactions: count of transactions done during weekdays and weekends? (Weekday/Weekend)
- What is the mode of transaction does the client use to make transactions; two were found to be interesting hence took the frequency (collection from another bank, remittance to another bank), all other modes of transactions were grouped under others.
- What is the average transaction amount per client in a monthly basis?
- After how many years did the client obtain a credit card after opening the first account with the bank?
- What kind of credit card does the client have? (Junior, Gold, Classic)
- What is the clients Loan Amount?
- What is the clients loan payment duration?
- What is the loan payment status for the client? (finished, contract finished but loan not payed, running contract loan payment ok so far, running contract, client in debt)
- What are the kind of information that can help us know the client better in terms of the region they come from? (Average Salary per client region, average crime rate in the clients' region year 1996 were used as it was the most recent and what is the clients' region unemployment rate, 1996 was also considered for the reasons aforementioned)

## Preparation of Datamart

After coming up with all the questions in the previous section, we started cleaning and creating new variables from the different tables available to create the data mart as follows

### Deriving demographics data

From the client table the birth number variable was used to calculate the clients' age and gender. This column had encoded information that is the birth number value meant that if the 4th and 5th values are over 50 then client was female if less than 50 then the client was male. The age was calculated from subtracting the birth year from 1999.Birth year was represented by the first two numbers in the birth number variable. The region data was derived from the demographic data by merging with the client data table.

*This kind of data would be beneficial to the bank in knowing how to best position their services based on the clients age, gender and region that they come from.*

### Deriving when the client first opened an account with the bank

These variables were derived from the accounts table date variable by considering the first two number values as year of account opening and third and fourth as account opening month.

*This kind of information would be beneficial in establishing a customer life time cycle.*

### Deriving dummy variables for the account statement type the client receives from the bank

Three dummy variables were created from the accounts table to check the frequency of the accounts sent to the clients (Monthly, Weekly or Transactional Statements)

**Deriving dummy variables for the kind of credit card that the client owns**
Three dummy variables were created from the credit cards table (Junior, Classic or Gold)
*This kind of information can be used to upsell credit cards to holders whose have active or consistent credit transactions*

**Deriving dummy variables for client account type**
Two dummy variables were created from the disponent, Client is either owner or disponent.
*For example, while running a loan upsell promotion all clients under disponent are to be excluded.*

**Deriving dummy variables for client account type**
Four dummy variables were created from the loan table status variable to check the loan status per client based on the four statuses that the bank has.
*High loan risk clients are easily tracked when each client has a loan status tag to their profile.*

**Deriving the number of years, the client took before getting a credit card**
This was done by getting the difference in years from when the client opened an account and the year the client was issued with a credit card. The tables merged were the account, disp and cards tables.

**Deriving the value of the customer based on total credits and total withdrawals**
Summation of all the withdrawals and all the credits done by the client over the whole period covered in the transactions dataset. Two variables were created from this Credit_amount and Withdrawal_amount.

**Deriving frequency of transactions based on weekdays and weekends.**
The total number of transactions done by a client during weekdays and during weekends. Two variables were derived Weekend to depict the number of transactions done on weekends only, and Weekday to depict the number of transactions done on weekdays.
*These variables tell us when the transactions were made the most (weekday or weekend)*

**Deriving frequency of transactions based on mode of transaction, two key ones stood out**
(Collection from another bank and Remittance to another bank)
Total count of transactions based on these two modes as competitor insight can be drawn when remittance transactions are more that the collections from other banks. Frequency of all other mode of transactions except these two were categorized under others. Three variables were derived in total.

**Deriving frequency by transaction type**
Different transaction types were considered for variable creation. Available transactions type were: insurance, negative balance interest, household, payment for statement, interest credited, old age pension and loan payments.
*Having these frequencies could better help the bank upsell or cross sell products to the clients based on the transaction types they are more inclined to embrace.*

**Deriving the average transactions amount per month**

This was derived by first grouping the total transactions amounts based on monthly basis and then doing an average of the same to find the average monthly transaction amount per client.

**Deriving the balance amount**

This was derived by taking the balance per client from the transactions table only for the max transaction date.

**Deriving the number of automatic debits made by the client**

This was derived by summing up the number of permanent orders per client from the permanent order table.

**Deriving loan amount and duration**

This was derived as is from the loans table.

**Deriving Avg Salary, Avg crime rate '96 and Avg Unemployment rate '96 per region**

Year 1996 was considered as it was the most recent in the data provided. The region average was calculated based on the summation of all district per that region figures and then doing the average. Three variables were created.

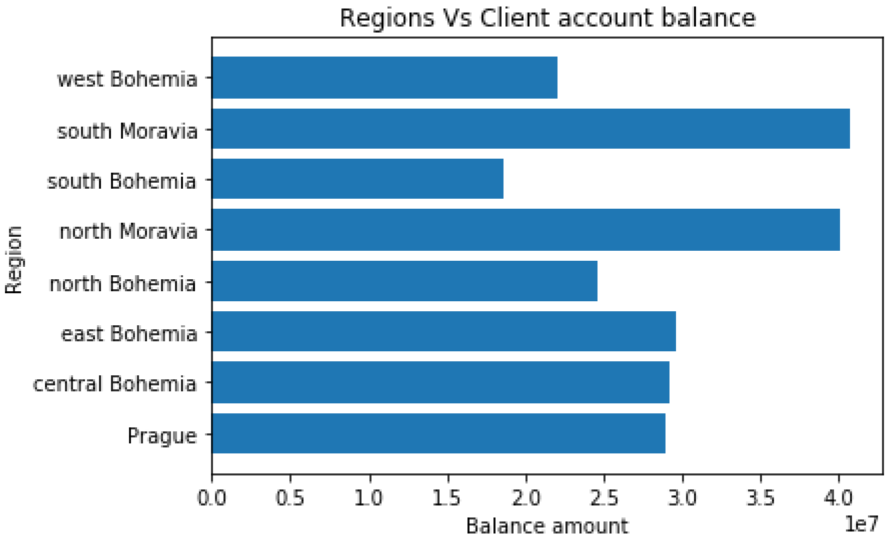
## Final data mart columns

| Variable | Tables Used for Variable Creation |
|---|---|
| client_id | Used as is from client table |
| gender | Calculated from client table |
| age | Calculated from client table |
| Weekday | Frequency calculated from transaction table |
| Weekend | Frequency calculated from transaction table |
| Credit_amount | Amount calculated from transaction table |
| Withdrawal_amount | Amount calculated from transaction table |
| collect_otherbank | Frequency calculated from transaction table |
| remit_otherbank | Frequency calculated from transaction table |
| others | Frequency calculated from transaction table |
| oldage_pension | Frequency calculated from transaction table |
| insurance_payment | Frequency calculated from transaction table |
| negative_payment | Frequency calculated from transaction table |
| household | Frequency calculated from transaction table |
| payment_statement | Frequency calculated from transaction table |
| credit_interest | Frequency calculated from transaction table |
| loan_payment | Frequency calculated from transaction table |
| Avg_trans_permonth | Amount calculated from transaction table |
| Balance_amount | Amount calculated from transaction table |
| years_since_creditcard | Years calculated from account,disp and credit cards tables |

| | |
|---|---|
| classic | Dummy variable retrieved from credit cards table |
| gold | Dummy variable retrieved from credit cards table |
| junior | Dummy variable retrieved from credit cards table |
| statement_monthly | Dummy variable retrieved from the accounts table |
| statement_transaction | Dummy variable retrieved from the accounts table |
| statement_weekly | Dummy variable retrieved from the accounts table |
| account_year | Calculated from the accounts table |
| Disponent_account | Dummy variable retrieved from the disp table |
| Owner_account | Dummy variable retrieved from the disp table |
| loan_amount | Retrieved as is from the loans table |
| loan_duration | Retrieved as is from the loans table |
| Loan_complete_paid | Dummy variable retrieved from the loans table |
| Loan_complete_unpaid | Dummy variable retrieved from the loans table |
| Loan_Contract_running | Dummy variable retrieved from the loans table |
| Loan_client_debt | Dummy variable retrieved from the loans table |
| Debit_Freq | Frequency calculated from permanent orders table |
| Region | Retrieved as is from the demographics table |
| Region_AvgUnemploymentRate_96 | Calculated from the demographics table |
| Region_AvgSalary | Calculated from the demographics table |
| Region_AvgCrimeRate_96 | Calculated from the demographics table |

# Data Visualisation

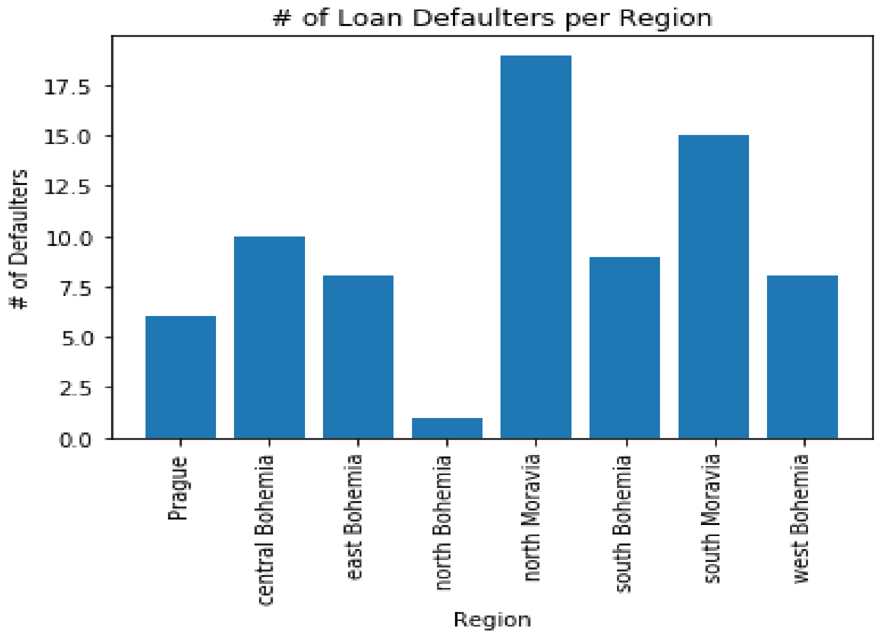
## Regions having most account balance



This plot indicates that people who live in Moravia have the highest account balance, among, south Moravia ranks the 1st and north Moravia ranks the 2nd.

The 3 regions followed are respectively East Bohemia, central Bohemia and Prague, however there is not a huge difference between these 3 places. And where people keep least money in their bank is the south Bohemia. With this kind of information which can be drilled to the client level it’s easy to segment the clients and propose services based on their respective regions.

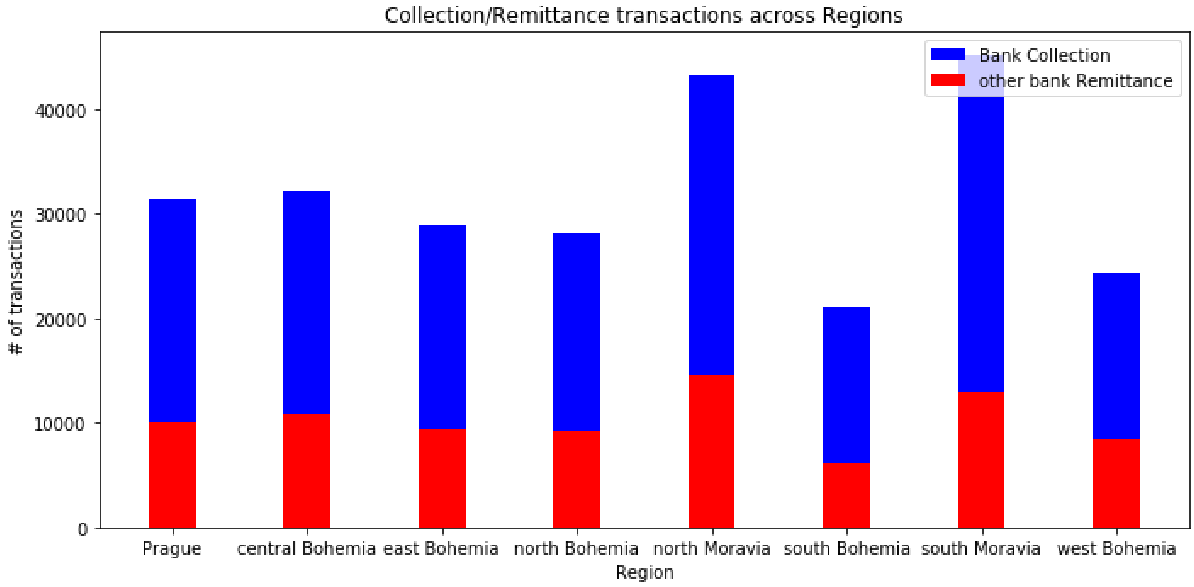
**Regions having the most loan defaulters**



The above bar chart shows that the north Moravia and the south Moravia are also respectively the top 2 region that have most people default the loan (calculated by summing the number of clients that had a loan status as loan duration finished and loan not paid and loan ongoing but client in debt). Followed by Central Bohemia, south Bohemia and east Bohemia. While north Bohemia has least loan defaulter, although it is the last 3rd region in balance amount.

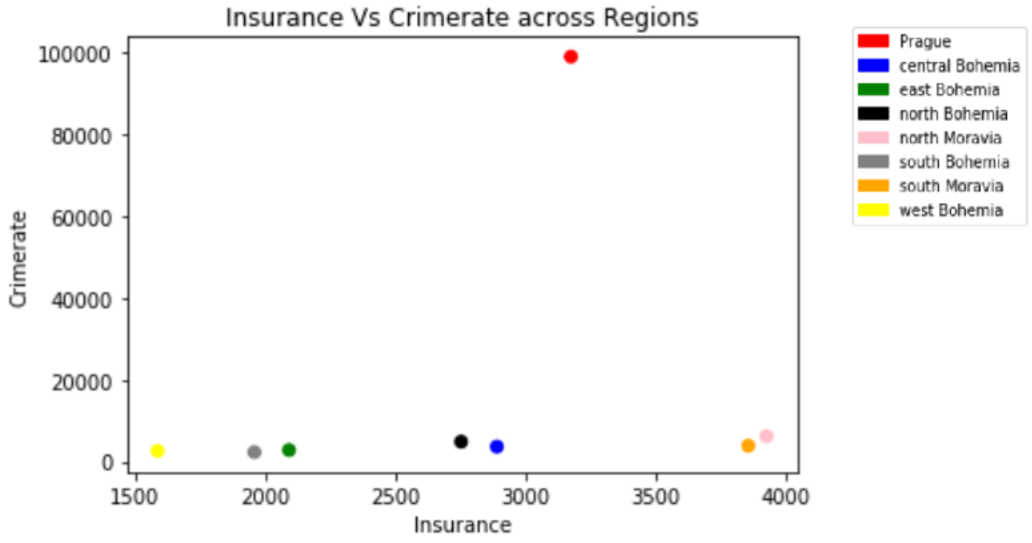
This may give information that people living in north and south Moravia are richer and at the same time they often borrow a lot of money from bank. For people in these 2 region, the bank could make a loosen money return policy. While for north Bohemia apply a tighter one.

### How does Collection/Remittance to other banks differ based on regions



Based on region, we can see that south Moravia has the most bank transfer amount, north Bohemia follows, and in all the regions people benefit much more from bank transfer rather than remitting to other banks. So in this aspect, collection from other banks to this bank was more however in some regions this count is almost equal to remittance to other banks which could mean that people were actually operating other banks other than this. Overall this shows that there more collection compared to remittance which is a good sign for the bank.

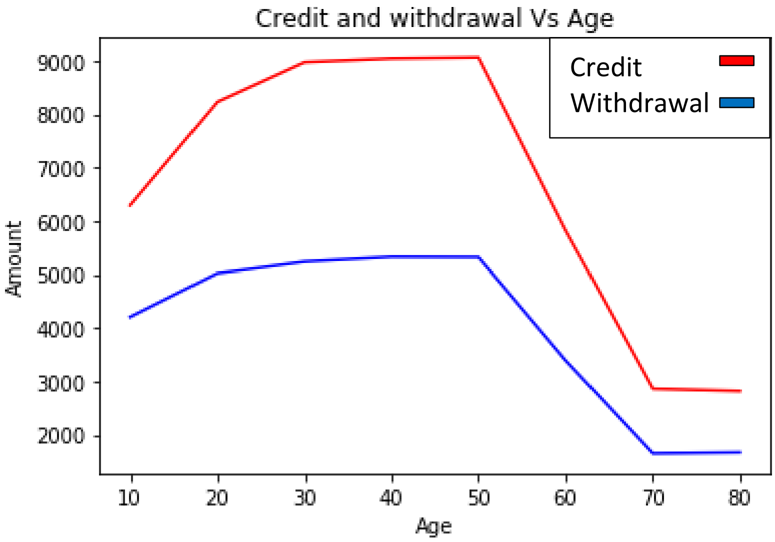
### Relationship between payment of insurance and the average crime rate in each region



Here we could find that Prague has the highest crime rate over all regions, and it ranks 3 in the amount of insurance people buy. This information could be a tip of the bank's insurance advertising that to Prague would be a good target for insurance selling.

Among civilizations in north Moravia pays most for insurance, followed by south Moravia. While the former one has an average crime amount ranks the 2nd, and the latter one ranks the 3rd. This means north Moravia and south Moravia are also good targets for insurance advertising.

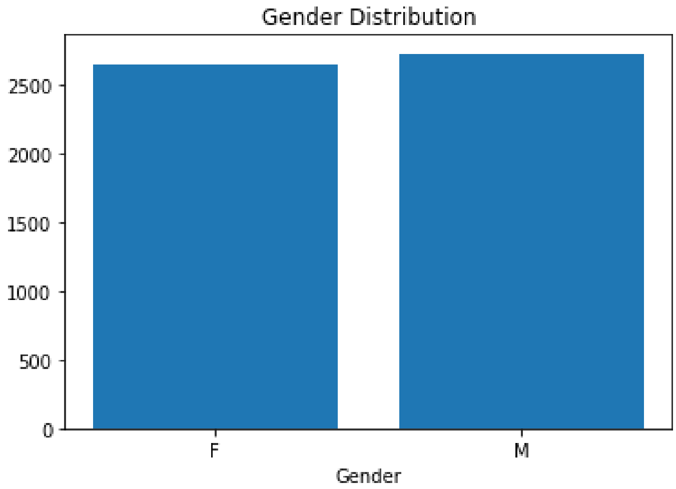
**Age groups with most frequent credit and withdrawal amounts**



From the chart it is obvious that people from 30-50 years old have both most credit and withdrawal frequency, especially people at their 50s. It means bank is most frequently used by this age group. At the same time we found people between 60-80 years old merely use bank services, especially people aged 70. As a result, age group of 30-50 would be the bank's most important target.

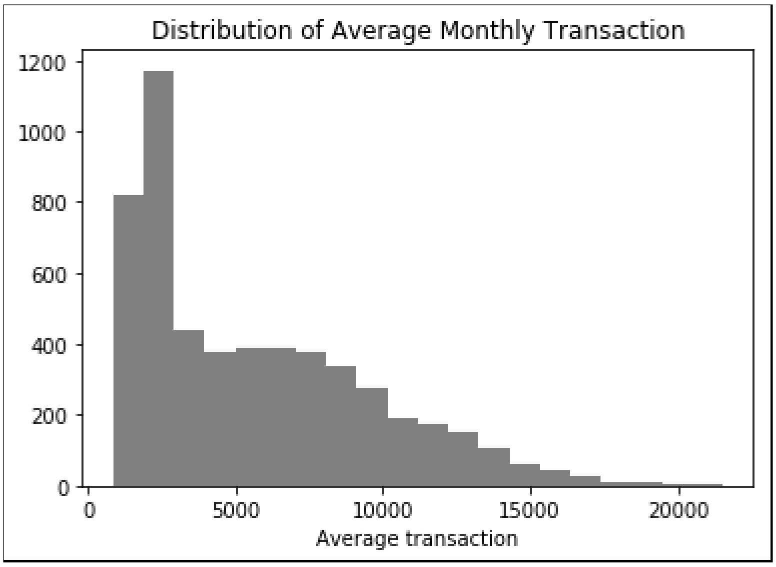
Also the plot indicates in general people tend to have more credit rather than withdrawal, that is people prefer saving rather than consuming.

### Gender Distribution across clients



The bar chart above shows the bank has more male costumers than the females, so a male might be more possible to use a bankcard and bank services, however the difference is not very huge.

### Distribution of Average monthly transaction



The above graph shows that for most people the average transaction amount on a monthly basis is less than 2500. Also there are very few people who are have an average monthly transaction of more than 15000. By this information bank can provide some special service to these clients whose Average transactions are more than 15000 and retain them

## Conclusion:

Based on our data mart and analysis we can say that our high value customers can be from the region of Moravia (north and south) with the age group between 30 and 50 with good average transactions on a monthly basis. We can say these people are good clients and more promotional offers can be send to them in order to retain them and also do promotional activities or services to increase their usage. But there are also more loan defaulters in this regions which has to be looked upon. Customers from Prague should be kept watch on as there is more crime rate in that region and they have more insurance transactions as well. People from North bohemia have lesser loan defaulters who can be targeted for more loans.

Apart from these conclusions that can be drawn based on the visualisation, the data mart variables can further be leveraged by the bank by applying some predictive analytics on them that can help find behaviour patterns that could help them increase usage and retain customers whilst keeping cost in mind.