Carlos Montenegro
Sébastien Pavot
Edward Vrijghem

Group Assignment Financial Programming 2019

Prof. Dr. Minh Phan

Contents

Report

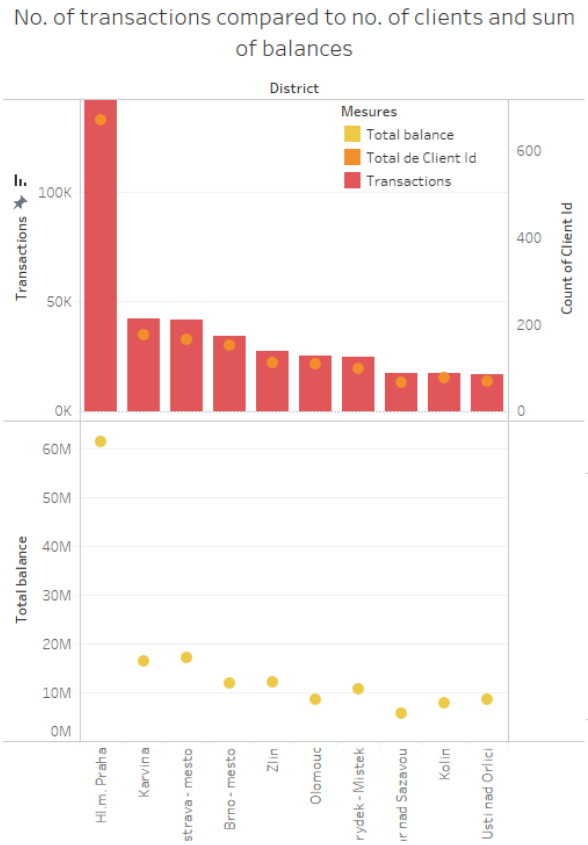
General data overview

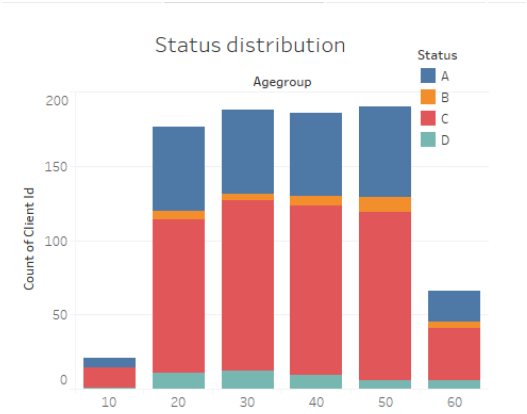
We found out that in general the total number of transactions and the total balance per district were highly correlated, this came as no surprise. The possession of a high number of clients results in a high number of transactions, as well as a high total balance. More clients result into more accounts, which results in more transactions as well as a higher total balance.

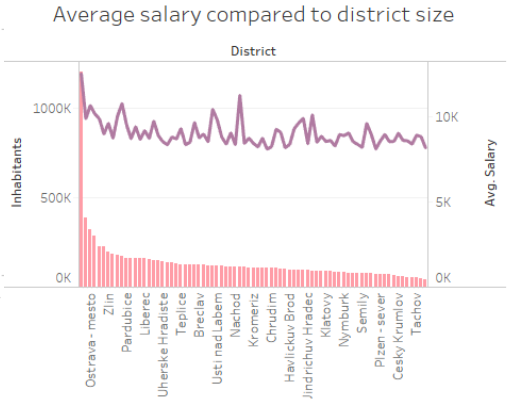
When comparing the average salary per district size we concluded that Prague, the capital of Czechoslovakia, scored best in both categories. Furthermore we observed a slight correlation between district size and average salary. Our clients were earning more in the capital district, compared to the in the smaller, lesser populated districts.

When digging deeper into loan statuses we start to notice some age-group related differences. When observing contracts where the loan was not payed, we see and increase for the age group of 50 year olds. This could indicate an increased risk when borrowing to this group, the risk assessment division should take this into consideration. Next to this, the high level of indebted clients (indicated in green) who take a loan in their 20's to 50's should also be considered.

Furthermore, every trend goes as expected. Younger and older age-groups tend to borrow less in general. This is firstly because the group of 60-69 tends to spend less money on so called investments (houses, cars and others) than the younger generation (10-19yo.) which doesn't have a lot of loans. This is most likely because the legal minimum age to borrow money is 18 in Czechoslovakia.







Insights and opportunities

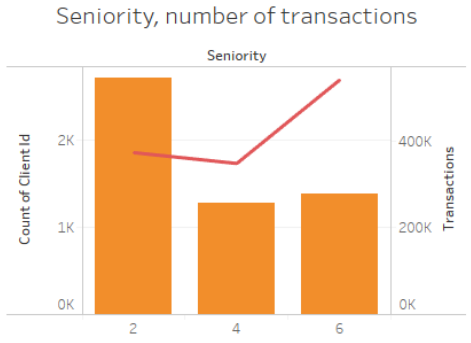
As of the relationship between seniority and number of transactions our findings were just as expected. Senior clients tended to have a higher total number of transactions compared to the newer clients. This means the newer clients possess a high potential and should thoroughly be targeted. More transactions mean more transaction fees which translates into more revenue for the bank.

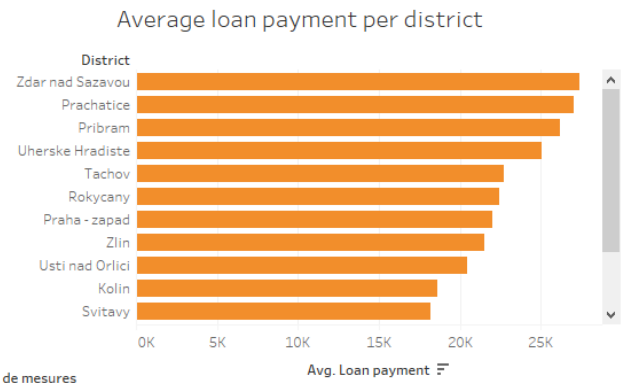
As for the cards we can see that there are huge opportunities, the grey part of the bar charts represent clients which are not in possession of a card yet. The clients situated in "the grey zone" could form a target group for later advertising actions.

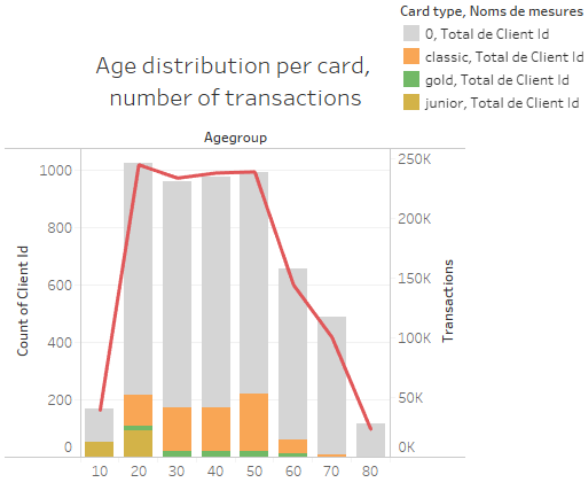
As per average loan payments per district (a direct derivative of average loan), we included the top regions. Surprisingly the capital district Praha didn't came out on top. This information could be useful to analyze in terms of marketing opportunities. Maybe this could be a result of the high population which results in a better indication of the average loan. Another reason could be that there is more competition in the capital to get a decent loan, resulting in banks being more reluctant in giving out big loans. In both cases the bank could take this into account when doing credit scores/assigning loans.

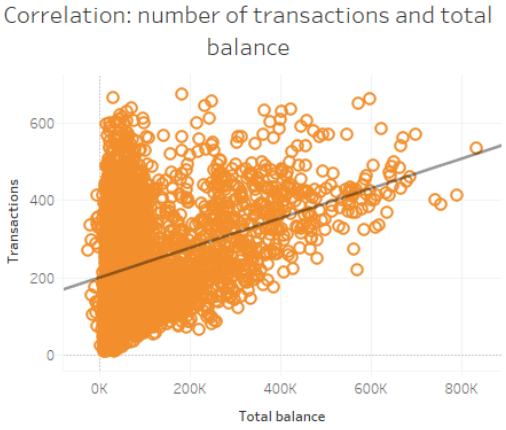
Carlos Montenegro
Sébastien Pavot
Edward Vrijghem

As per the relation between the number of transactions and the balance of your clients you can see a strong positive correlation. Wealthier clients tend to do more transactions, as expected.





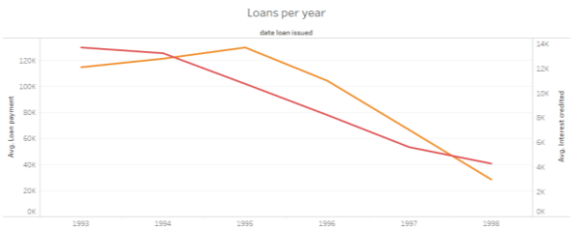


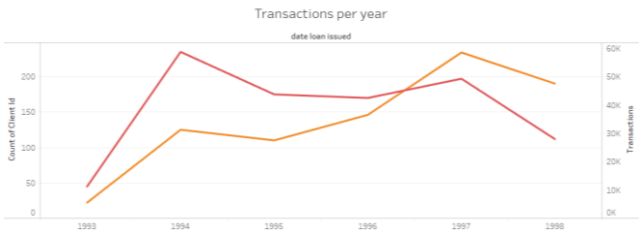


Trends

As for trends we see a steady decrease in the average yearly loan payment as well as the interests credited. These obviously go together. This could be an industry trend or could a company specific trend, in both cases this is a very worrying indicator and should thoroughly be examined.

Yearly transactions and total clients increased from 1993 to 1994, remained steady until 1997. After 1997 we notice a decrease, this is equally worrying as the decreasing average loan payments and average interests credited. All of these factors need a solid analysis and should be reviewed.





Carlos Montenegro
Sébastien Pavot
Edward Vrijghem

Technical Aspects

General Structure of the code

Explore the data

We explore the data using the following main functions:

1. nunique(): which gave us the number of unique values for specific variables
2. head(): which in most cases was used to return the first 5 rows of each dataset
3. info(): which gave us a summary for each variable, containing the number of non-missing values and its type.
4. Describe(): which gave the maximum and minimum numbers, which was specially important to check in which ranges our variables fluctuated.

Clean the data

1. We changed the name of the categories of some variables to more meaningful ones.
2. We left the missing values untouched and created new dummy variables for them.
3. We read the dates and calculated some new variables from them (ex. birthday or recency)
4. We decided which variables where not important to maintain and we dropped them.

Merge the tables

1. We found out that there was a problem with the information in the data set of transactions ("trans"). While most of the other tables were account based, the "trans" data set was transaction based, which mean that each row of data was uniquely identifying each transaction made in each account. We modified this data set and made it account based.
2. Once the previous step was done, we created to temporal big data sets:
    a. One containing the 'loan', 'order', 'trans', 'district' and 'account', in an account-based dataset.
    b. And the other containing the 'card', 'disp', and 'client', in an account-based dataset.
3. Finally, we were able to combine this temporal datasets, and group them by client Id to be able to get a client-based dataset and have a DataMart that uniquely identifies each client.

Create new variables

1. We created more variables which were mostly client oriented with the exception of the demographic variables which were aggregated figures.
2. We explored the categories and values of our variables and created new variables based on certain condition that we found meaningful.

Libraries

- **pandas**: this library is used to convert the variables to date time format, to merge tables, to create dummies for categorical variables, to get dummies, among others.
- **numpy**: this library is used to calculate a time difference, to find NaN values, to load a data file, to fix a random state, to obtain absolute values, to arrange bins, among others.

Functions created

- **explore:**
  - Logic: this function makes the data exploration simpler by generating the main information about the data set
  - Input: data set to be analyzed
  - Output: it returns a print of the following functions ouput: describe(), info(), nunique()
- **to_month_gender**
  - Logic: this function returns the gender of a text that contains the birthday and gender of a person (50 + MonthNumber if the person is a woman)
  - Input: data set value
  - Output: the person's gender

Variable information

Variable name	Explanation	From table
disp_id	Record identifier of disposition	Disp
client_id	Record identifier for each client	Client
account_id	Record identifier for each account	Account
Owner / Disp	Is the client owner or disponent of the account	Disp
Is_shared?	Is the account shared? 1 for yes and 0 for no.	Disp
card_id	Record identifier of credit card	Card
Card type	Type of card ('Classic', 'gold', 'junior')	Card
Date card issued	Date the card was issued to the client	Card
Time since card issued	Time in days since the card was issued based on 1999/01/01 as the today date	Card
Has card	Flag column to identify client with card. Yes, means that the client has a card and no if he has not.	Card
district_id	Record identifier of district	District
age	Age of the client in years	Client
gender	Gender of the client, F for female and M for male.	Client
Issuance type	Type of issuance of statements (Monthly, weekly, immediately)	Account
Date account opened	Date the account was opened	Account
Seniority	Time in years since the account was opened	Account
Loan amount	Amount of money the loan is valued	Loan
Loan duration	Time in months of the loan	Loan
Loan payment by month	Amount of money due by month	Loan
Loan status	Status of paying of the loan. A equal loan finished and paid, B equal loan finished but not payed, C for contract still running and OK so far, D for contract running but client in debt.	Loan

contracted a loan	Flag column to identify people who contracted a loan. Yes for people who contracted a loan and No else.	Loan
date loan issued	Date the loan was issued	Loan
Loan_finished	Yes if the loan is finished regarding time, no else. Based on difference between the 1999/01/01 and the date the loan was issued.	Loan
Loan months remaining	If the loan isn't finished, difference in month from 1999/01/01 since the loan was issued. Means the number of months remaining in the loan contract.	Loan
Amount loan remaining	Amount of money remaining to pay if the client is paying each month based on number of months remaining by the money due by month.	Loan
Total_order	Total of amount of order.	Order
Leasing	Total of amount of order characterize as leasing.	Order
Credit	Total of credit that the account has by doing the sum of all transactions characterize as Credit.	Trans
Debit	Total of debit that the account has by doing the sum of all transactions characterize as debit.	Trans
Cash deposit	Total of amount characterize as operation type cash deposit.	Trans
Cash withdraw	Total of amount characterize as operation type cash withdraw.	Trans
Money transfer to other bank	Total of amount characterize as operation type money transfer to other bank.	Trans
Recovering other bank	Total of amount characterize as operation type recovering other bank.	Trans
Debit card	Total of amount characterize as operation type debit card.	Trans
Other operation	Total of amount characterize as operation type other.	Trans
Insurance payment	Total of amount characterize as transaction type insurance payment.	Trans
Statement payment	Total of amount characterize as transaction type statement payment.	Trans
Interest credited	Total of amount characterize as transaction type interest credited.	Trans
Sanction interest negative	Total of amount characterize as transaction type sanction interest negative.	Trans
Household	Total of amount characterize as transaction type household.	Trans
Age pension	Total of amount characterize as transaction type age pension.	Trans
Loan payment	Total of amount characterize as transaction type loan payment.	Trans

Other transaction	Total of amount characterize as transaction type other.	Trans
Number transactions	Count the number of transactions	Trans
Total balance	Difference between credit and debit columns to know the actual solde.	Trans
district_name	Name of the district	District
Region	Region of the district	District
inhabitants	Number of inhabitants per district	District
ratio_urban	Ratio of inhabitants	District
avg_salary	Average salary per district	District
unempl_95	Unemployment rate in 1995	District
unempl_96	Unemployment rate in 1996	District
entrepren	Percentage of entrepreneurs per district	District
crime_95	Crime rate in 1995	District
crime_96	Crime rate in 1996	District
number_urban	Number of urban inhabitants per district	District
number_country	Number of inhabitants minus number of number of urban inhabitants	District
unemployment_trend	Increase or decrease of unemployment rate between 1995 and 1996	District
crime_per	Number of crime per inhabitants	District
crime_rate	Increase or decrease of crime rate between 1995 and 1996	District
has crime rate	Flag column to identify district that have a crime rate available.	District
has unempl rate	Flag column to identify district that have a unemployment rate available.	District