Exploratory Analysis Project

Introduction

This project is part of the Exploratory Data Analysis discipline of the Getulio Vargas Foundation in Business Analytics and Big Data MBA.

Members of the group: Matheus Amaral Mões, Marcelo Semerene Farah, Luísa Belus Henriques and Daniela de Góes N. Georg.

Context and objective: The manager of a Czech bank wants to get a better understanding on his clients. To do so, it has an extensive database that records customers and their transactions as follows:

To assist the manager, our team works with the database in the following steps: Understand, clean and organize data; Search relationships between data; Find product and service opportunities for the bank.

To assist the manager, our team worked with the database in the following steps:

1. Understand, clean and organize data;
2. Search relationships between data;
3. Find product and service opportunities for the bank.

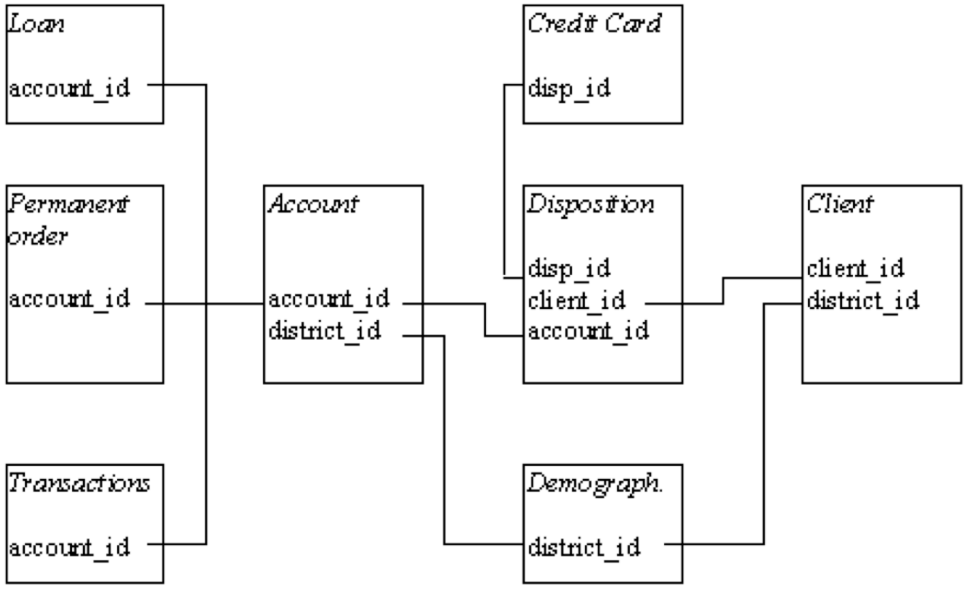
This book is organized according to the steps taken to reach a conclusion about the clients.

For the HTML version:

https://github.com/moesmatheus/exploratory_analysis

Read Files

The first step was to read the files and organize them. In each file, some columns had to be translated and some data reordered to facilitate the exploratory analysis work.



data

Accounts

Here we find the account data. There are 4500 records, each containing information about the account district, frequency of statement issuance, and date of account creation.

To make the data easier to understand, we translate Czech frequency values into English and separate dates with hyphens.

Table 1: accounts frame

account_id	district_id	frequency	date
576	55	monthly	1993-01-01
3818	74	monthly	1993-01-01
704	55	monthly	1993-01-01
2378	16	monthly	1993-01-01
2632	24	monthly	1993-01-02

Clients

The Clients list contains 5369 bank customers, tabulated with the district code and their birthday code - which informs the birthday and gender. For the analysis, we separated the date of birth and gender in separate columns.

Table 2: clients frame

client_id	birth_number	district_id	gender_code	gender	birth_date
1	706213	18	1	W	1970-12-13

2	450204	1	0	M	1945-02-04
3	406009	1	1	W	1940-10-09
4	561201	5	0	M	1956-12-01
5	605703	5	1	W	1960-07-03

Disposition

The Disposition relationship contains interactions between customers and accounts, classifying as owner / dependent. Some explanations of possible account-owner-dependency interactions: Every account has an owner and may or may not have dependents. Dependents may own another account. A customer may own more than one account.

Table 3: disposition frame

disp_id	client_id	account_id	type
1	1	1	OWNER
2	2	2	OWNER
3	3	2	DISPONENT
4	4	3	OWNER
5	5	3	DISPONENT

Order

List of payment orders issued to accounts. In addition to the money order code and the account code, the records contain information about the bank and account you received, the amount and type of payment (household, insurance, leasing and loan). To make it easier to understand, it was necessary to translate Czech payment types into English.

Table 4: order frame

order_id	account_id	bank_to	account_to	amount	k_symbol
29401	1	YZ	87144583	2452.0	household
29402	2	ST	89597016	3372.7	loan
29403	2	QR	13943797	7266.0	household
29404	3	WX	83084338	1135.0	household
29405	3	CD	24485939	327.0	

Transactions

List of all transactions made by customers, informing: Account that performed the transaction Transaction Date Transaction Type (Inbound and Outbound) Transaction (credit card withdrawal, credit in cash, collection from another bank, withdrawal in cash, remittance to another bank) Amount (transaction amount) Balance after transaction K-symbol (characterization of transaction) Bank (receiving bank) Account (receiving account)

Table 5: transaction frame

trans_id	account_id	date	type	operation	amount	balance	k_symbol	bank	account
695247	2378	1993-01-01	credit	credit cash in	700	700			NA
171812	576	1993-01-01	credit	credit cash in	900	900			NA
207264	704	1993-01-01	credit	credit cash in	1000	1000			NA
1117247	3818	1993-01-01	credit	credit cash in	600	600			NA
579373	1972	1993-01-02	credit	credit cash in	400	400			NA

Loans

Loan list, with 682 occurrences. Each customer can only receive one loan. In the table we have Loan key Account Key Date Value Duration (in months) Installment Payment Amount Status (finished - ok; Finished - not ok; Running - ok; Running - in debt) To make understanding easier during the analysis, we classify the status according to its conditions rather than leaving the original groups A through D from the database.

Table 6: loan frame

loan_id	account_id	date	amount	duration	payments	status
5314	1787	1993-07-05	96396	12	8033	Finished - not payed
5316	1801	1993-07-11	165960	36	4610	Finished - OK
6863	9188	1993-07-28	127080	60	2118	Finished - OK
5325	1843	1993-08-03	105804	36	2939	Finished - OK
7240	11013	1993-09-06	274740	60	4579	Finished - OK

Credit Card

The credit card list contains 892 occurrences and stores the card code, disposition id, card type (junior / classic / gold) and date of issue. Treatment was performed so that the issue date was in the Brazilian model.

Table 7: card frame

card_id	disp_id	type	issued
1005	9285	classic	1993-11-07
104	588	classic	1994-01-19
747	4915	classic	1994-02-05
70	439	classic	1994-02-08
577	3687	classic	1994-02-15

Demographic data

Finally, demographic data records 77 municipalities distributed in the 7 regions of the Czech Republic. The table contains the following elements:

District code;

District name;

Region;

Nº of inhabitants

Nº of municipalities with <499 inhabitants

Nº of municipalities with 500-1999 inhabitants

Nº of municipalities with 2000-9999 inhabitants

Nº of municipalities with> 1000 inhabitants

Nº of cities

Proportion of urban inhabitants

Average Salary

unemployment rate '95

Unemployment Rate '96

Nº of entrepreneurs per 1000 inhabitants

Nº of crimes committed in '95

Nº of crimes committed in '96

Table 8: demographic frame

region	district_name	inhabitants
Prague	Hl.m. Praha	1204953
central Bohemia	Benesov	88884
central Bohemia	Beroun	75232
central Bohemia	Kladno	149893
central Bohemia	Kolin	95616

Join Frames

In order to explore our database, we decided to combine the different tables. This was a two-step process:

1. Aggregate every table but the transactions in which the granularity was the account, the clients were agregated to show only the account owner and the number of the dependants in the account.

2. Next step was to join this new table with Transactions, creating a big dataset with all the data. The granularity for this is the single transaction.

With this two new structures we were able to study the data and draw some ideas and conclusions to help the manager.

Join frames without transactions:

```
#Base frame accounts
frame_no_transaction <- account %>%
  #Join accounts
  left_join(loan, by = 'account_id') %>%
  #Join Demographic
  left_join(demographic, by = 'district_id') %>%
  #Join disposition
  left_join(
    dplyr::filter(disposition, type == 'OWNER' ) %>%
      left_join(disposition %>% group_by(account_id) %>%
                  dplyr::summarise(dependents = n()-1), by =
'account_id')
    , by = 'account_id') %>%
  #Join client
  left_join(client,by = 'client_id') %>%
  #Join card
  left_join(card, by = 'disp_id') %>%
  #Join order
  left_join(order, by = 'account_id') %>%
  #Join transactions
  left_join(
    transaction %>% group_by(account_id) %>%
      dplyr::summarise(amount_transactions = sum(amount)),
    by = 'account_id')
```

Join frames with transactions:

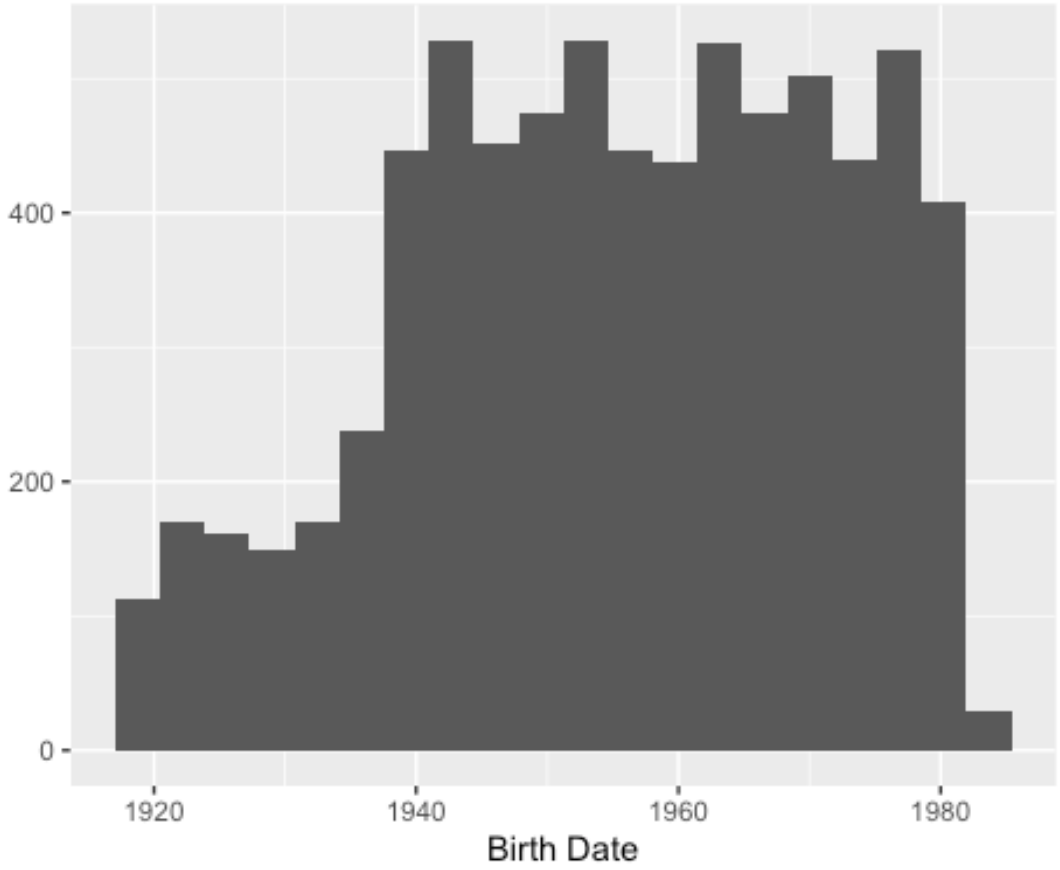
```
# Join previous frame with transactions
frame <- transaction %>%
  left_join(frame_no_transaction, by = 'account_id')
```

Client Profile

Understanding who the customer is is a critical step in generating hypotheses and moving on with the analysis. To do this, we explored the data and found some interesting information shared below:

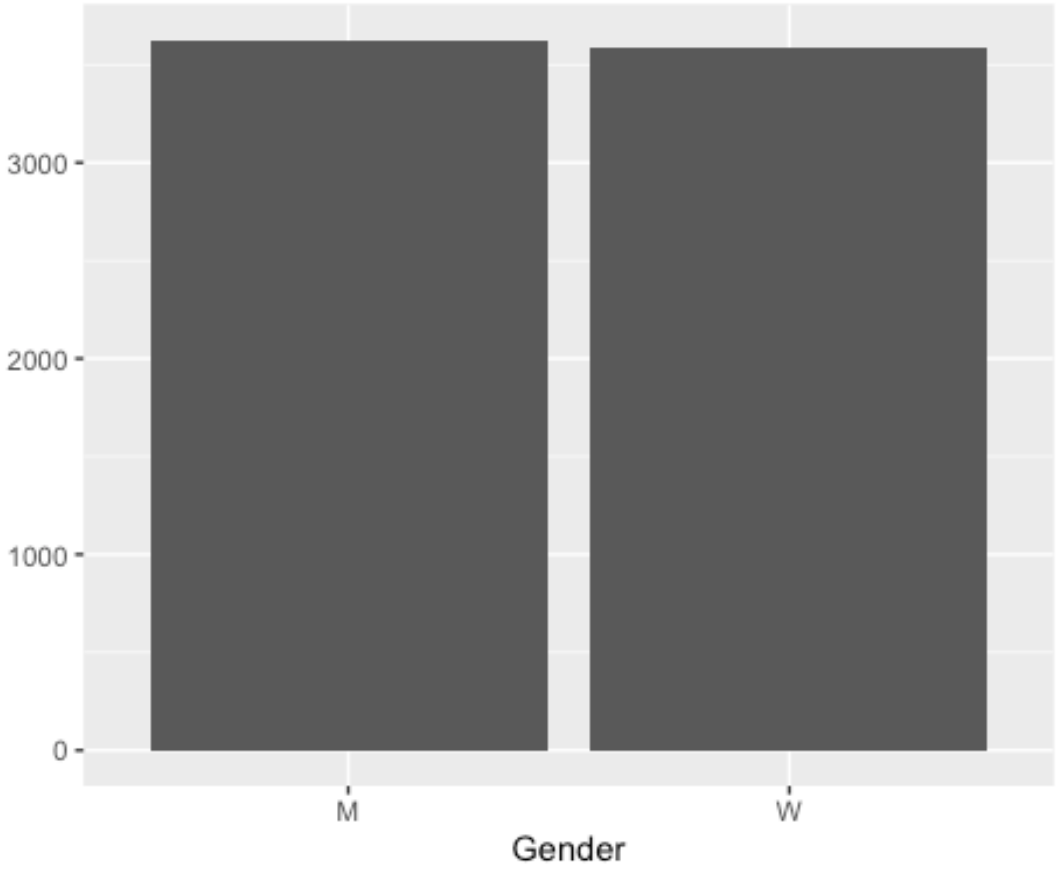
Birth Date

The majority of the bank's clients were born after the 40s.



Gender

There is a close distribution between male and female clients in the bank.



Number of Dependents

Most of the bank's clients don't have any dependents on their bank and account and the ones that do, only have one dependant. If having dependents is a revenue stream of the bank, this should be taken into consideration as an opportunity to grow.

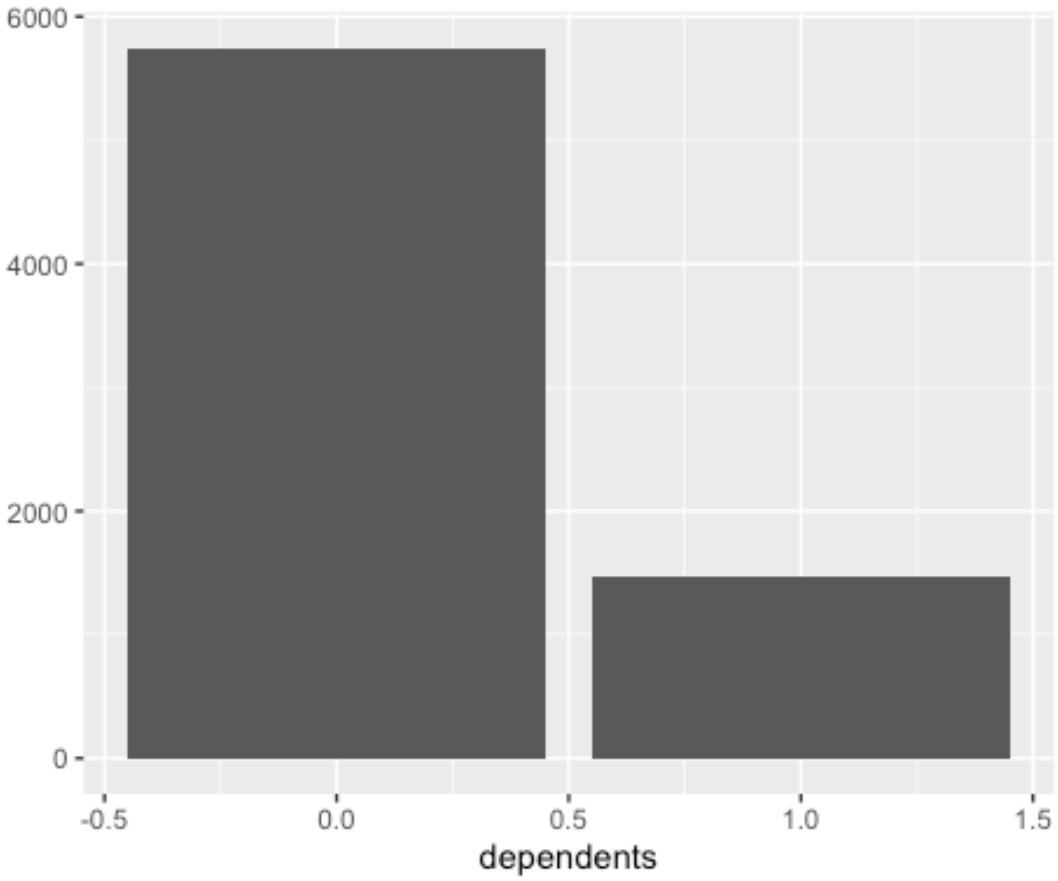


Figure 3: Number of Dependents

Account creation by time

There was a sharp decrease in accounts created after 1994 and since then the bank has been recovering from it.

Recommendation: investigate the root causes on this decrease, this way the bank can become aware of what impacts it's business.

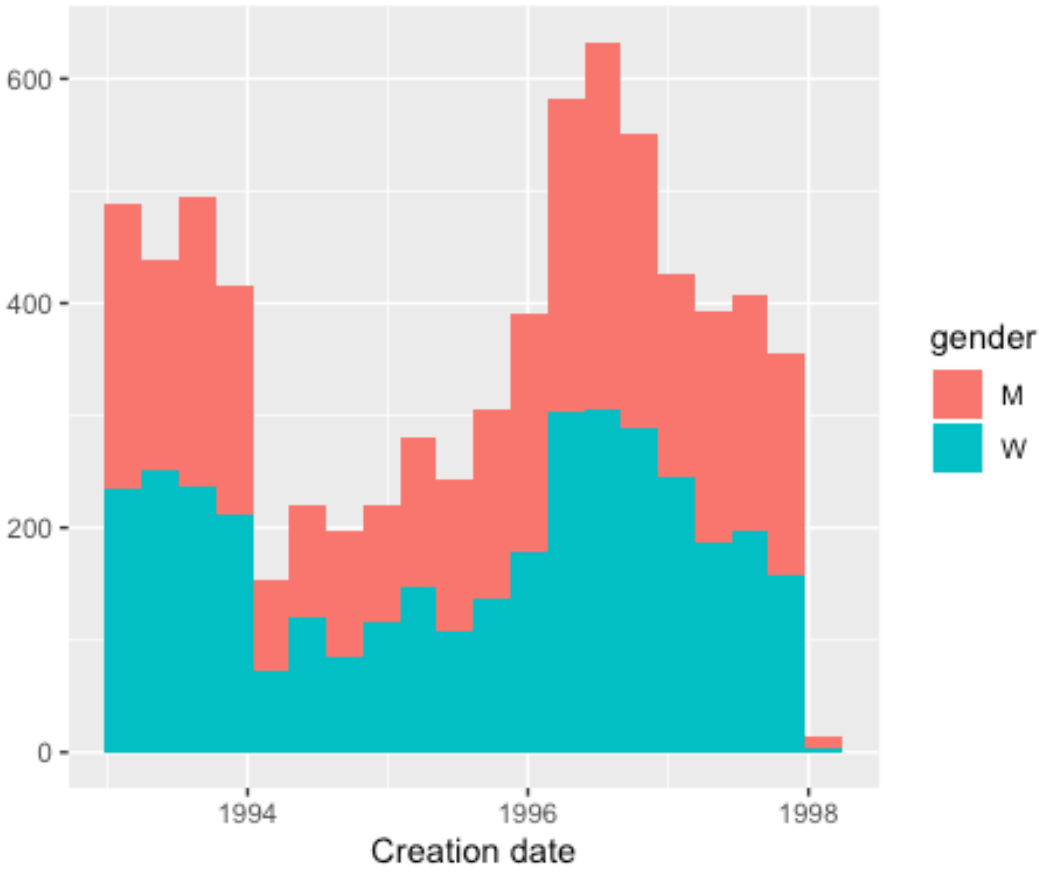


Figure 4: Distribution of account creation by time

Age at account creation

The youngest clients that the bank manage to capitate are on their late teens when they create their bank account and there is a sharp decrease of account creation after the 60s.

Having young clients is great for the business in the long term, but they are usually not as profitable as older clients (who already have higher incomes and do more bank transactions). It is suggested to seek greater penetration in the elderly market.

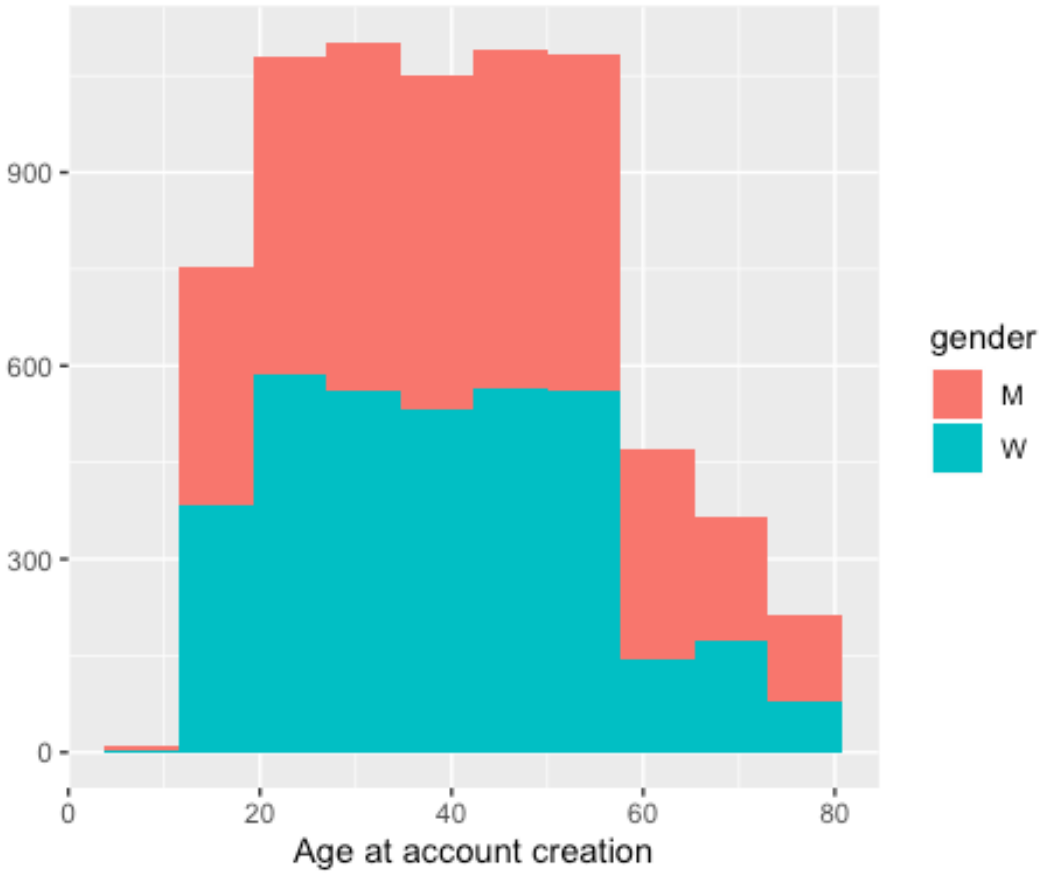


Figure 5: Distribution of age at account creation

Regions

There is no concentration of clients in any particular region in the country.

Given this scenario, a possible strategy would be to focus on some more profitable regions and gain space in them, then expand and gain more customers in all regions.

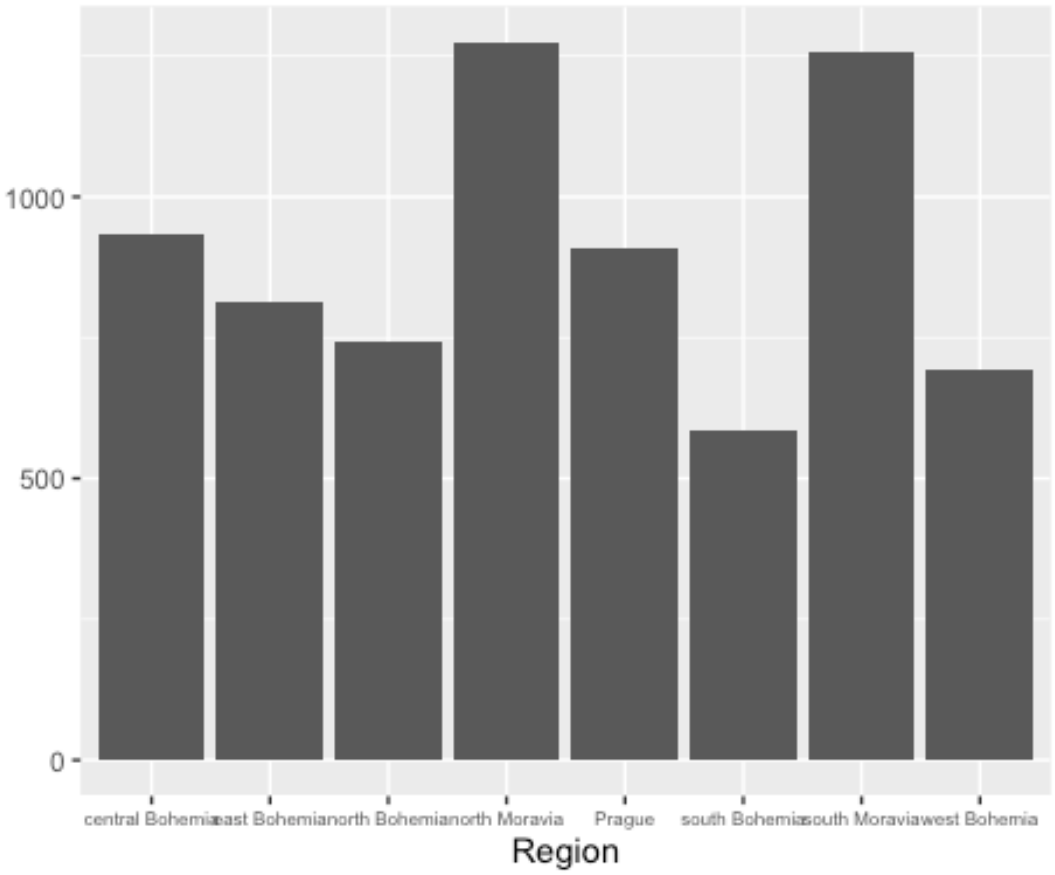


Figure 6: Clients by Region

Looking closely to number of inhabitants in each region and the number of clients we see that there is a major opportunity for improvement in South Moravia.

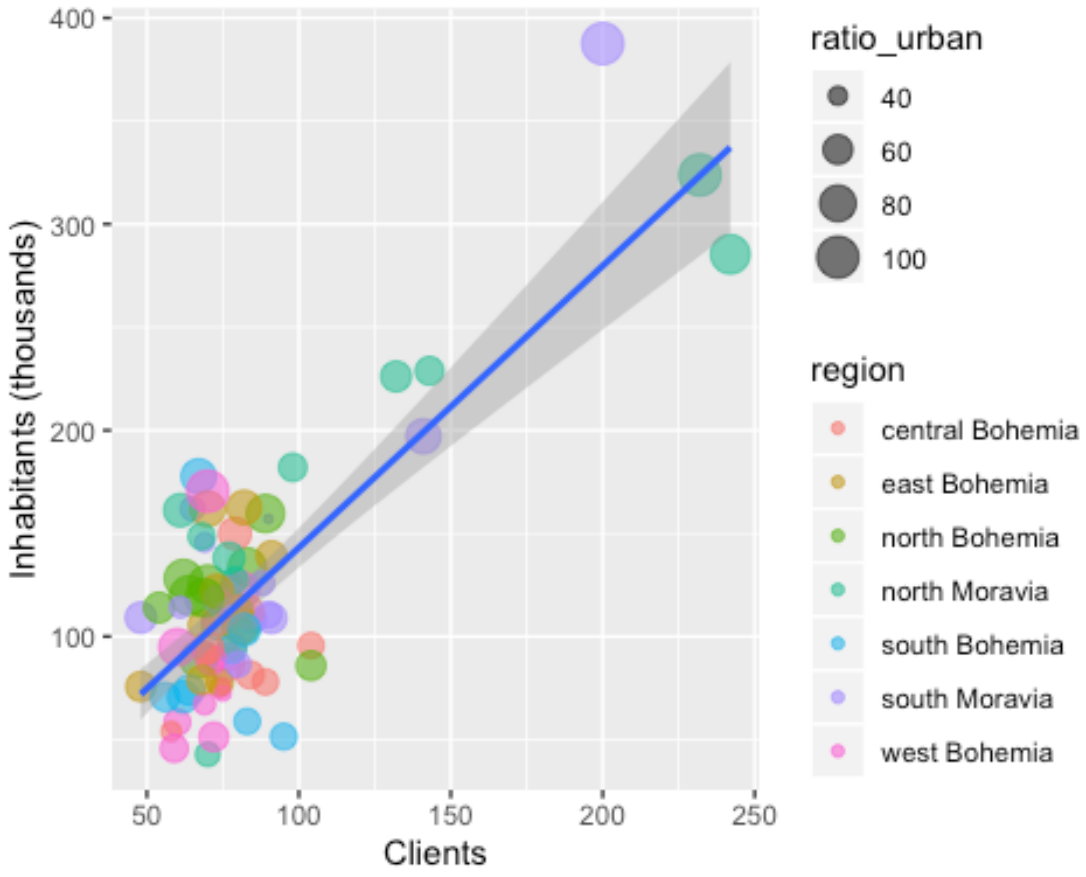


Figure 7: Clients by district

Frequency of statements

Almost all clients prefer to have their bank statements sent to them monthly.

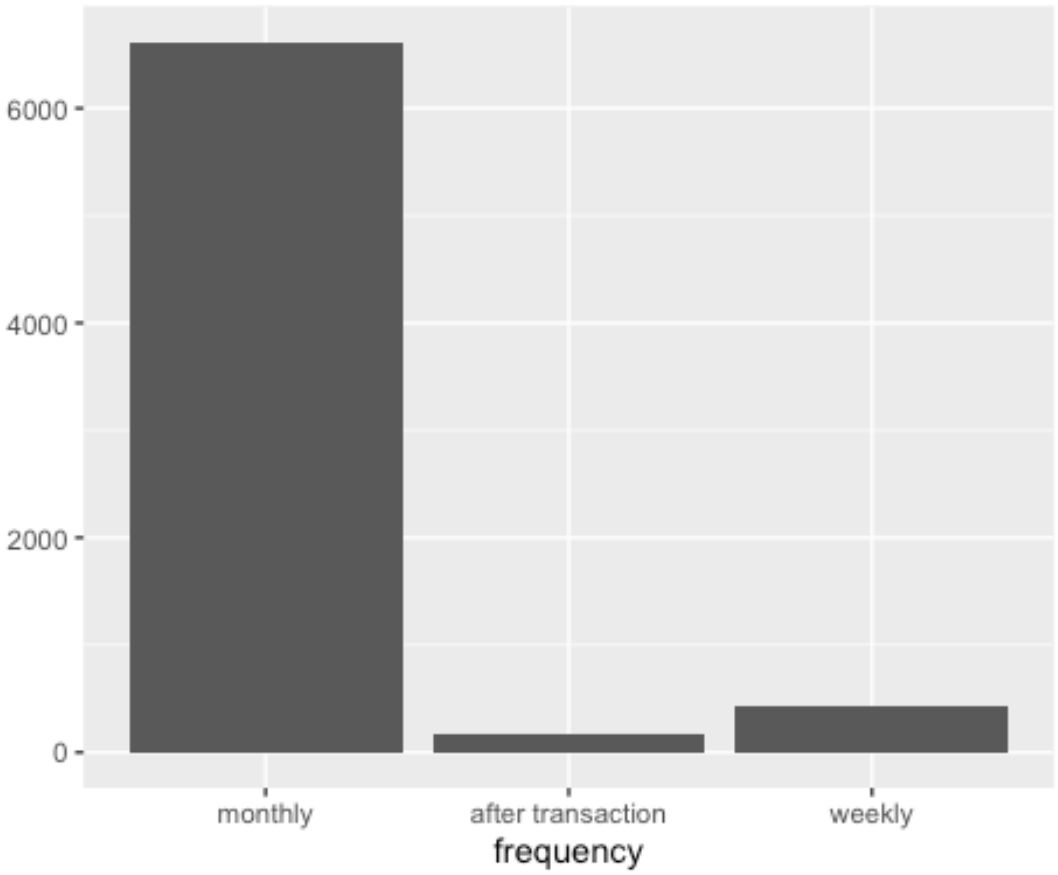


Figure 8: Frequency of statements

No bank client born before the 30s choose to have its bank statement sent after transaction or weekly.

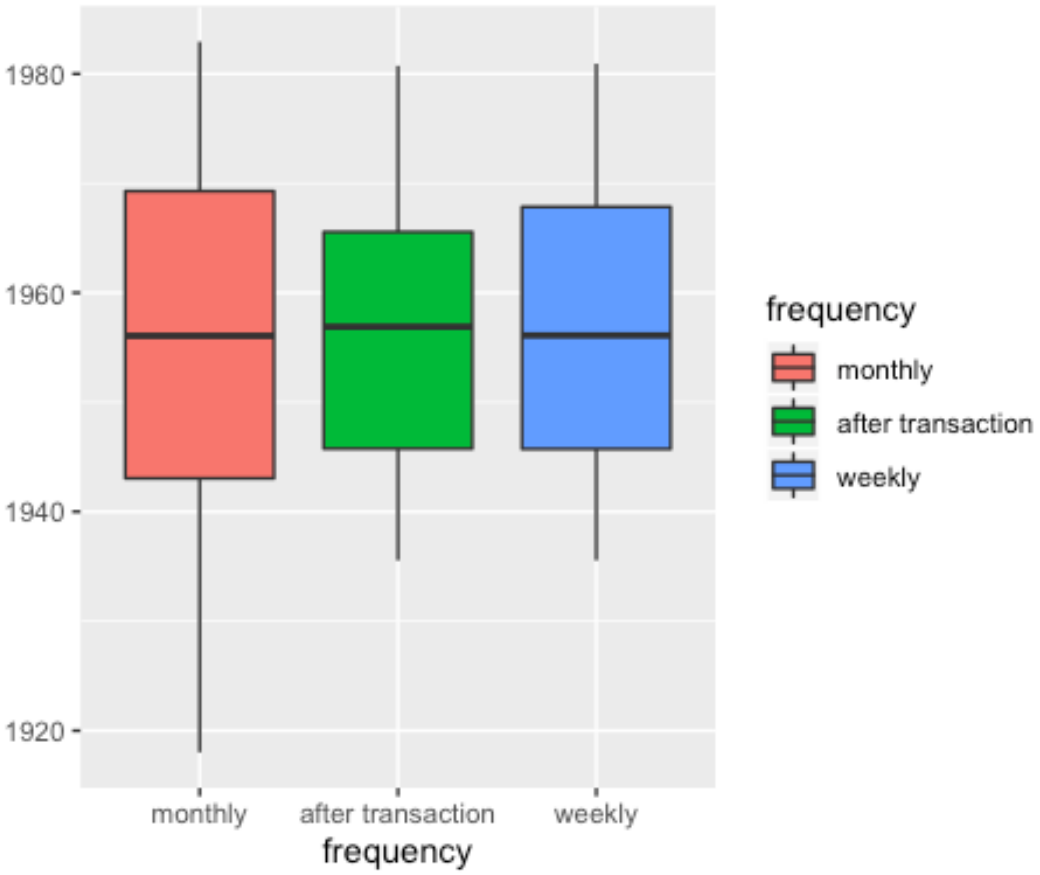


Figure 9: Frequency of statements by age

Loans

Loans are a great source of revenue for the bank, so we look at data for opportunities and outlier patterns.

Loans Status

The bank currently has a healthy record of loans with the majority of loans finished being paid correctly and less then 15% not being paid back. The current loans are also in good state with the majority of those running without problems and aproximatly 10% with the client in debt.

As the numbers indicate that the credit division is doing well in selecting their clients, it is suggested to extend the possibility of extending this lending to clients with similar profiles to borrowers. Accordingly, we may increase our profits from the volume of such interest.

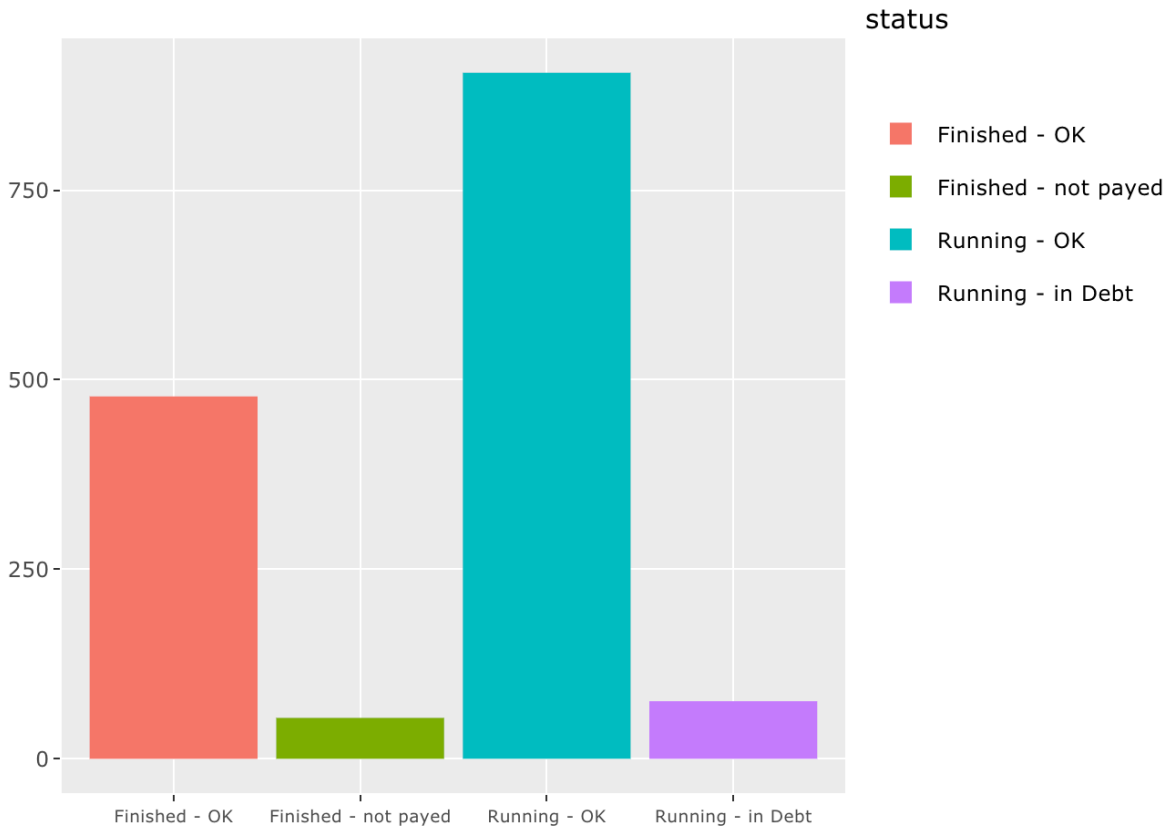


Figure 10: Loan Status

We can also see that that the number of loans with some problem does not appear to have a growing tendency, although we can see that there is a decline in loans being issued by the bank in 1998 showing some room for improvement. Also in 1997 there was a great number of bad loans, problaby being the caus for the reduction of loans in 1998.

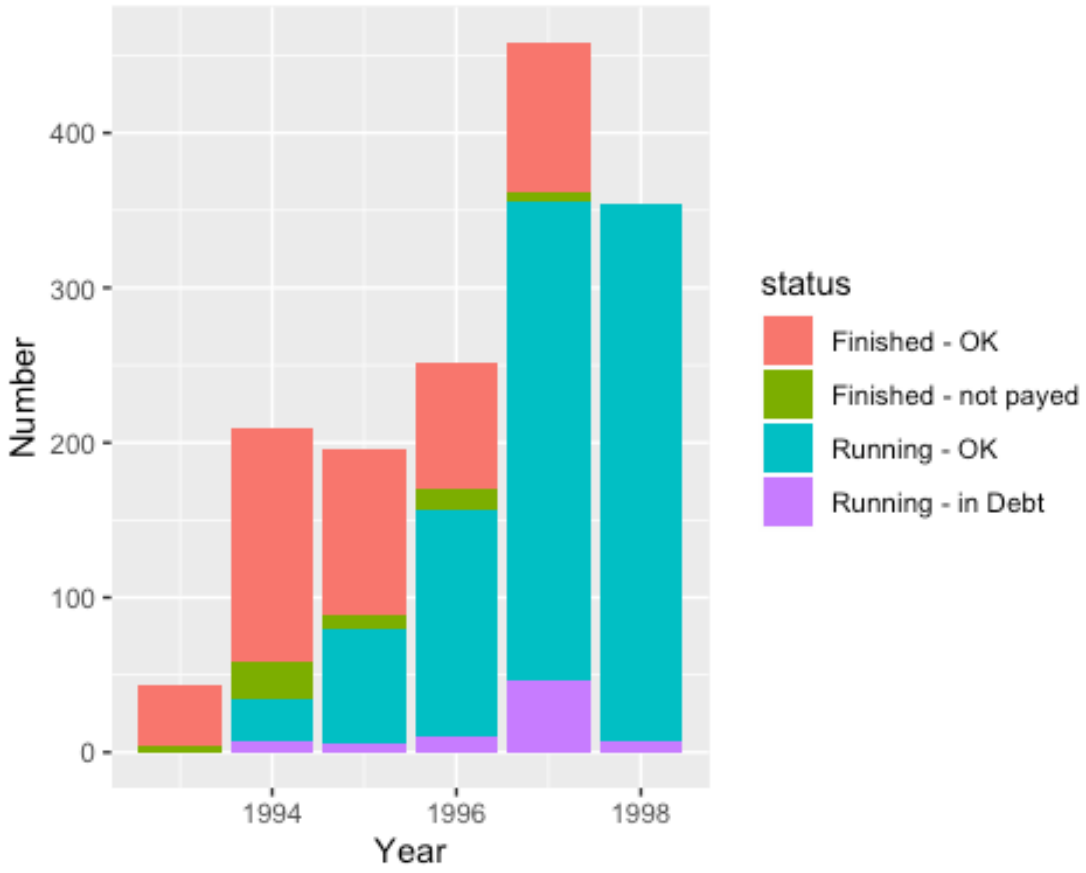


Figure 11: Loan status by creation date

Age distributions of loan

We can see that the bank spreads its loan through different age groups with the bulk of loans occurring from the 20s to the late 50s of its costumers and there is no major concentration of bad loans in any age group.

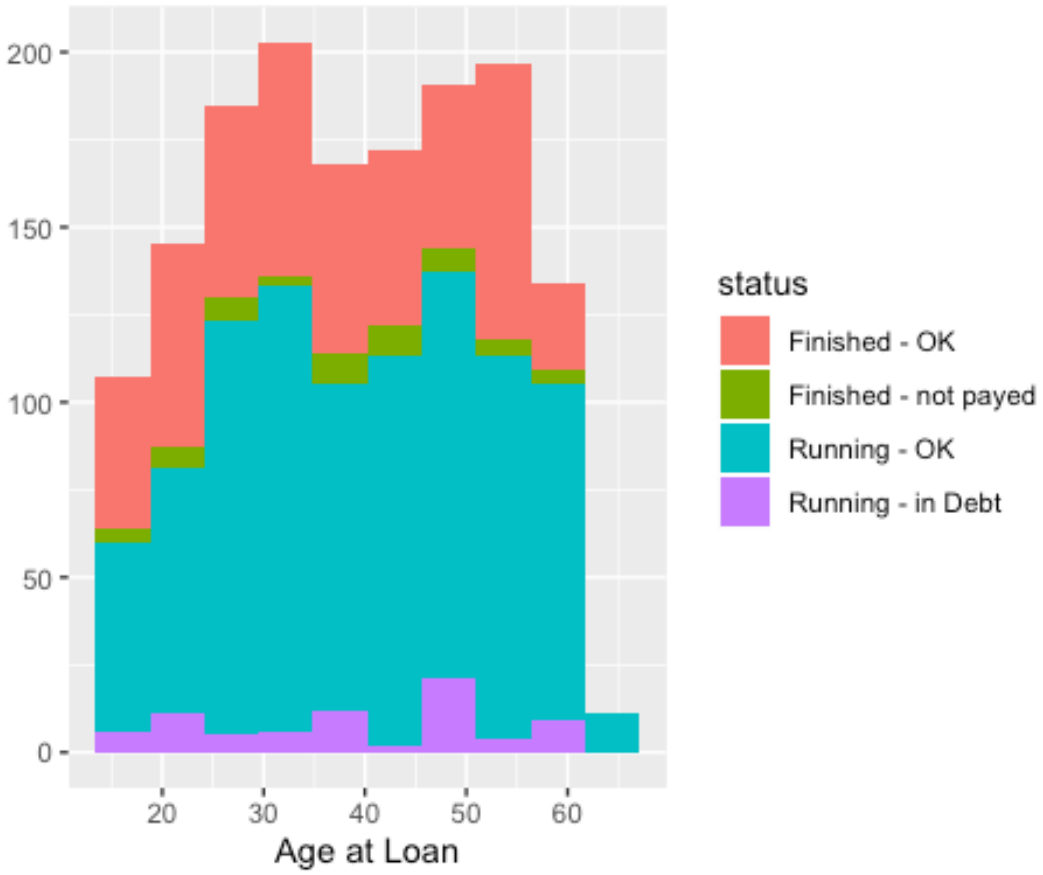


Figure 12: Loans by age at loan

By the Boxplot we can confirm the same conclusions with the addition that loans finished and not payed happened more with slighlty older people than the others.

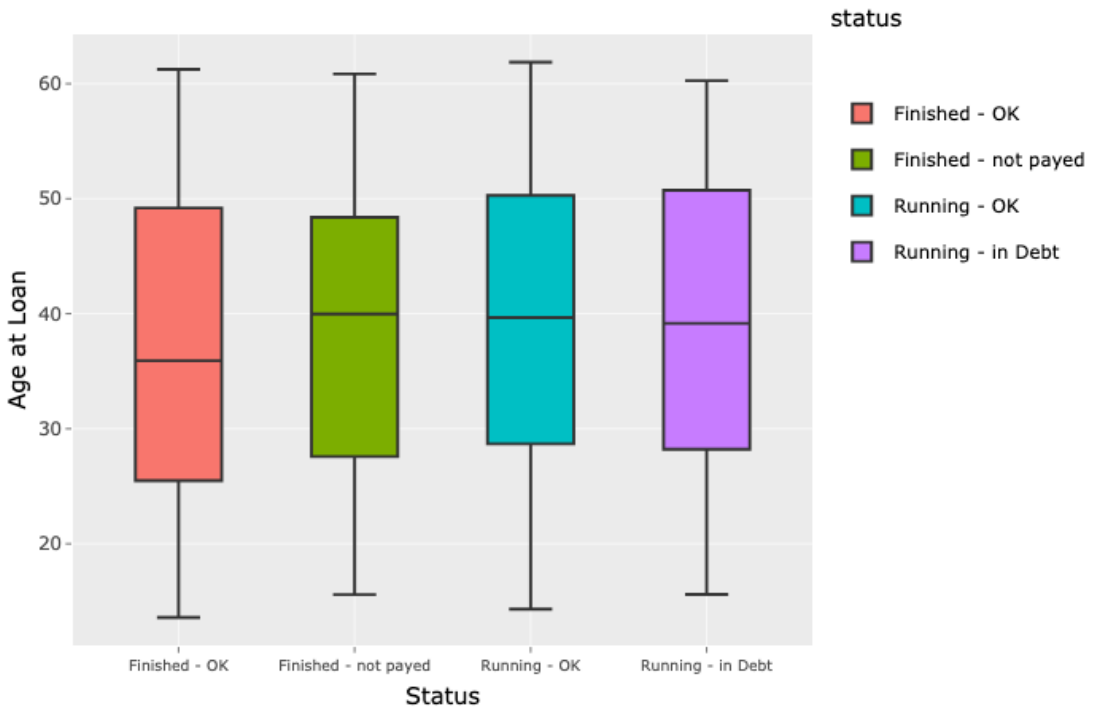


Figure 13: Boxplot loans by birth date

Loans status by size of loan

Analyzing the number size of loans taken we can see that bigger loans have higher risks. Comparing the loans finished and the ones still running we can see that in both groups there is ones with problems have a higher median value. Also with we can see that ther is a great number of outliers in the analysis.

By this information, is suggested to focus on bigger volumes of small loans, so the bank can make profit from it with less risk.

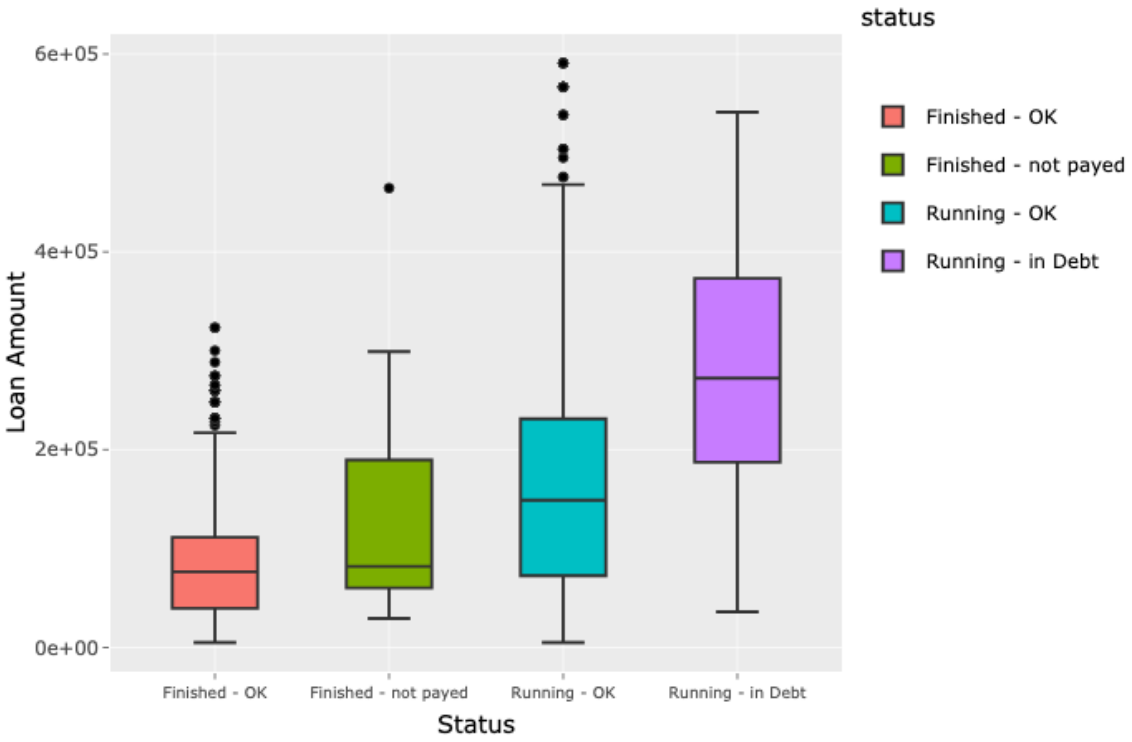


Figure 14: Distribution of birth date

Loans status by monthly payment

The number of outliers previously seen can be explained by the loans with a larger duration, in an analysis of loan size divided by duration we can see that there are no outliers in the boxplot. That analysis also shows that there are more problems in the loans with the higher monthly payment, so focusing on smaller loans could be a better strategy for the company. An interesting insight is that currently there is no default in the loans with a monthly payment smaller then 1.670 so the bank could be more agressive on those loans.

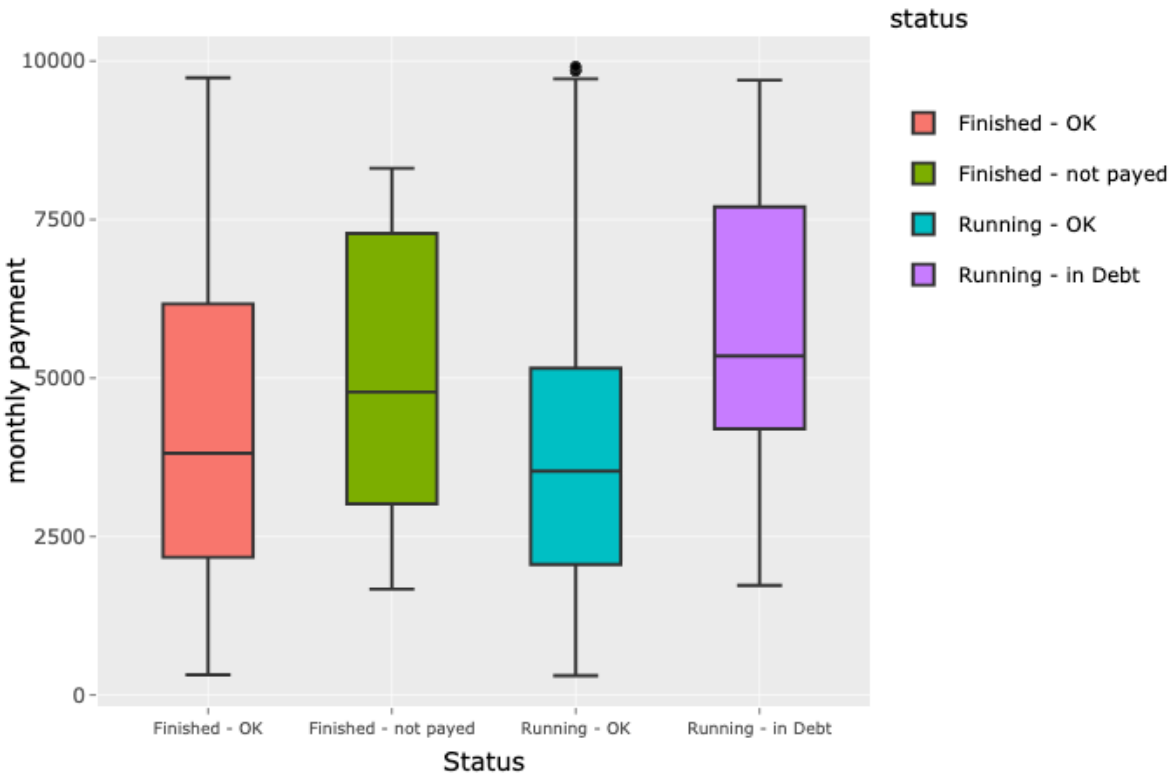


Figure 15: Loans status by monthly payment

Transactions

Amount of transactions annually

The amount of transactions done by the bank is growing continuously, showing the bank's growth. The majority of those transactions involves physical cash, either withdrawal or credit, and those are the most important drivers in the transaction's growth. The share of credit card in the banks transactions is very small.

This indicates an opportunity of growth from offering more credit cards and making incentives for the clients to use them.

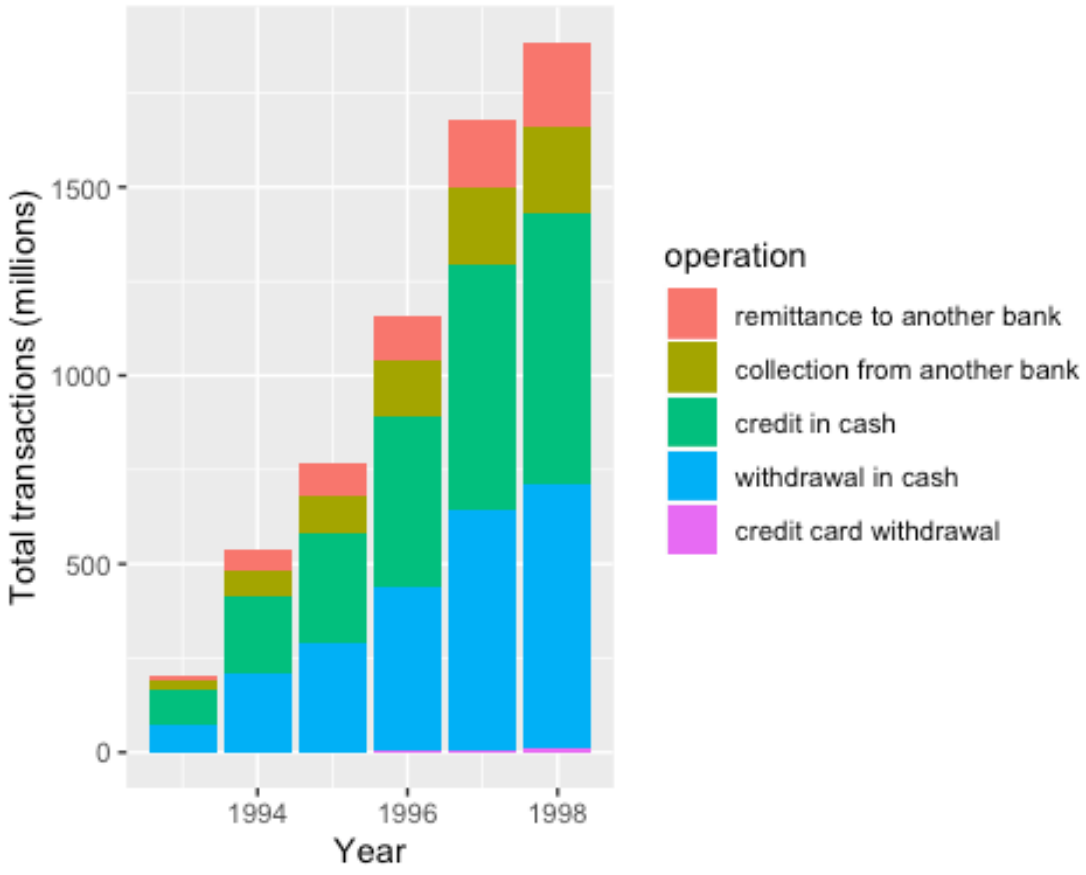


Figure 16: Amount of transactions annually

Net transactions amount annually

We can see that in every year on the time series the net flow of transactions is positive, which indicates that the bank shouldn't have a liquidity problem.

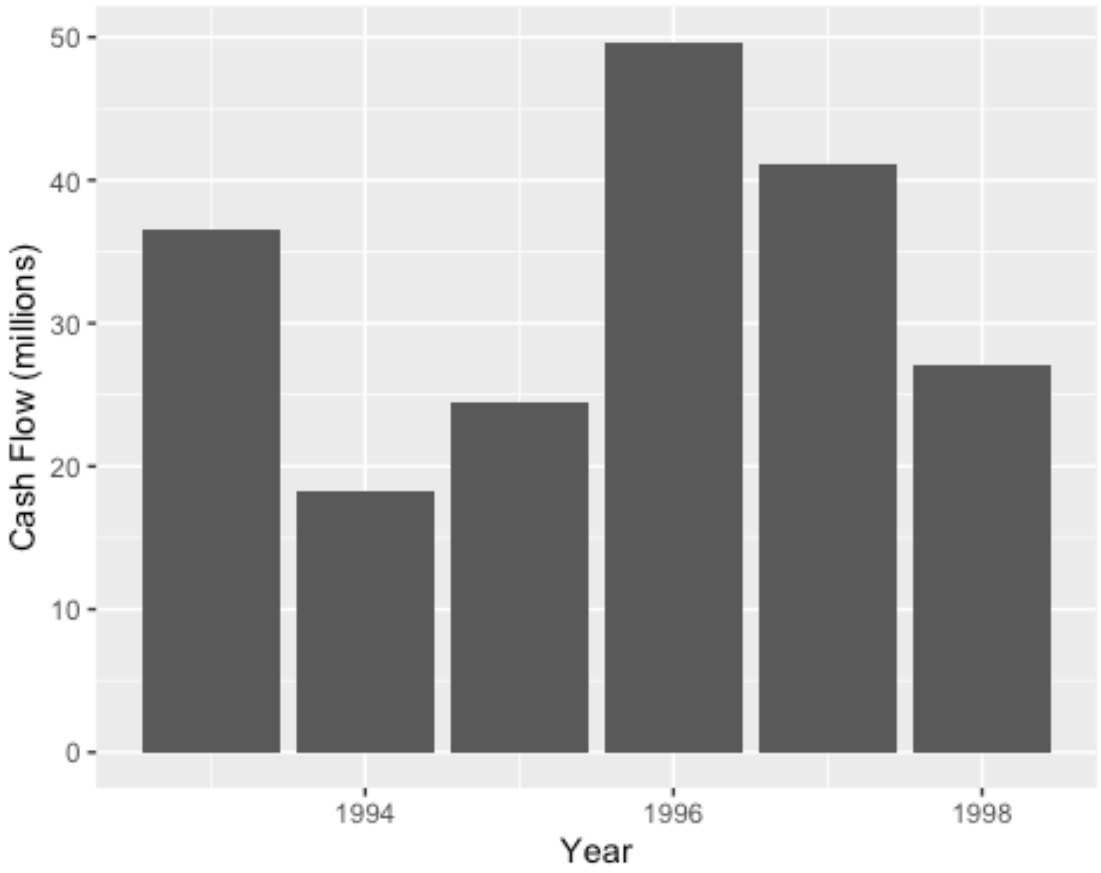


Figure 17: Net transactions amount annually

Net transactions amount monthly

If we take a closer look to the transactions monthly, we can that on every January and June of every year there is a negative flow of transactions, specially on January. Those negative flows should be a cause for concern to the bank and it should be more cautious on those months.

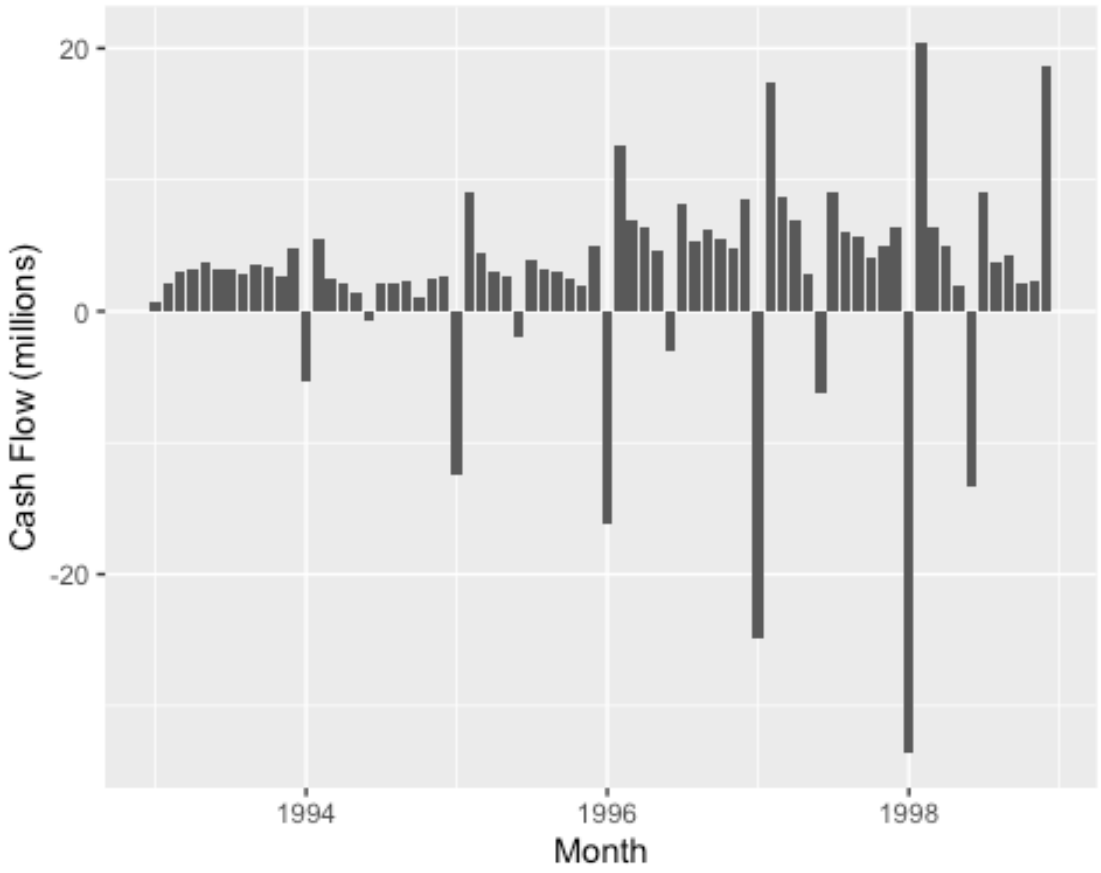


Figure 18: Net transactions amount monthly

Looking in the types of transactions we can see that the negative cash flow is caused by a surge in the number of withdrawals.

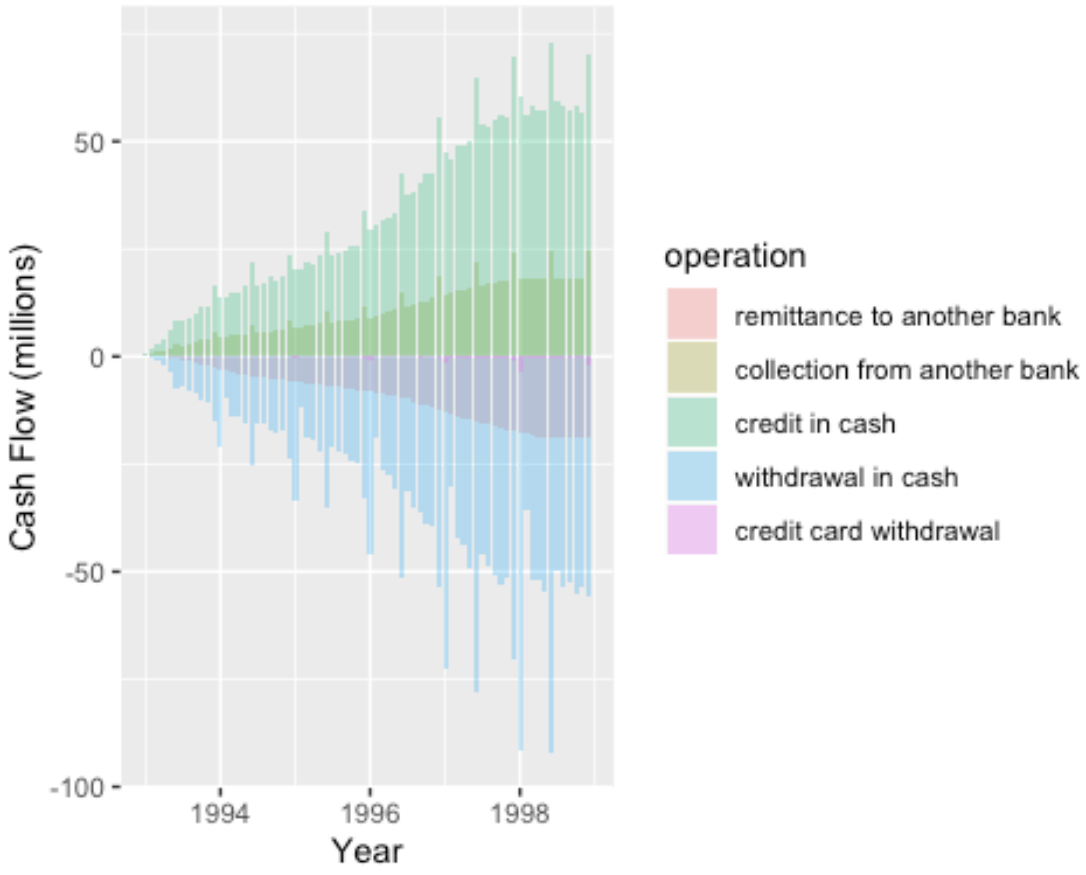


Figure 19: Cash Flow Year by Year

From the products categorized by the bank we can see that household has the fastest growth, even though it and the other products are going through a period of stagnation recently.

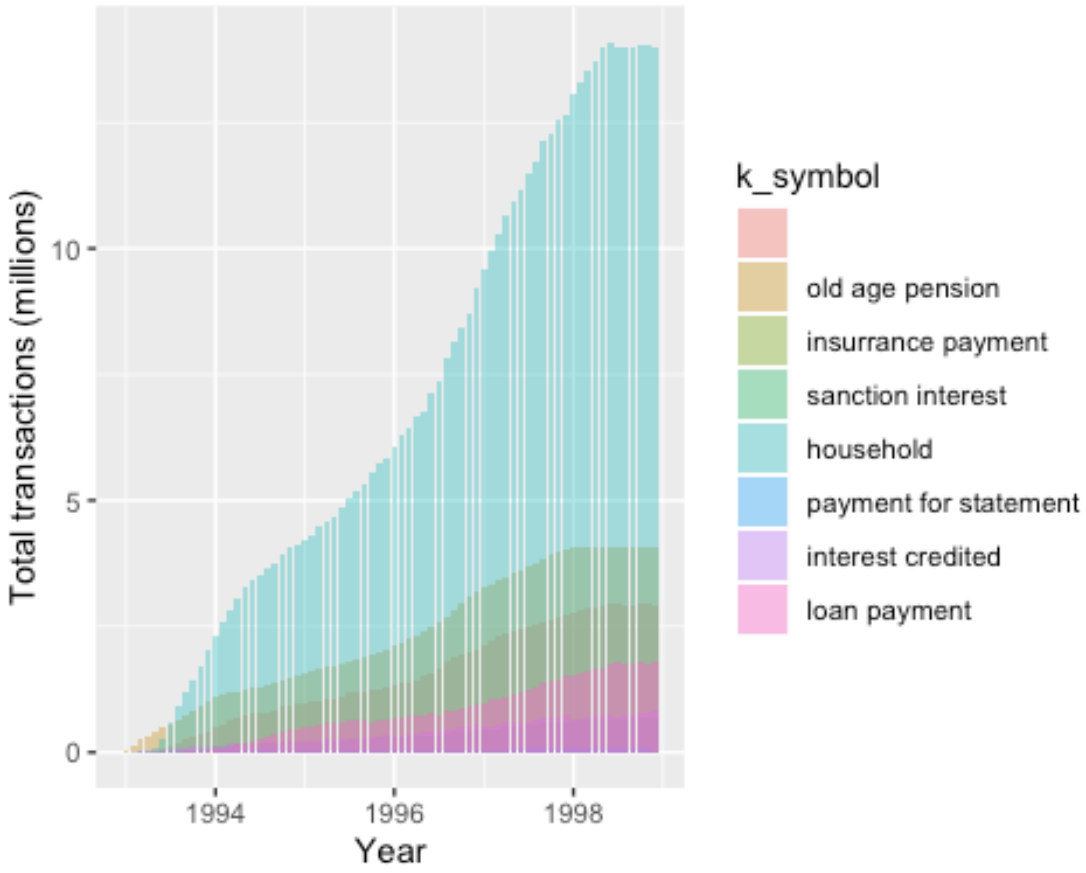


Figure 20: Cash Flow Year by Year

Orders

Order types

From the orders segmented by the bank household stands out as the most relevant one.

Creating special deals for this type of order may be a good opportunity for growing the bank but it's important to analyse if there's still a tendency in this direction. The risk is that this order type was very significant in the last years due to the economic situation of the country (from socialism to capitalism) and maybe from now on other order types will become more relevant for the population.

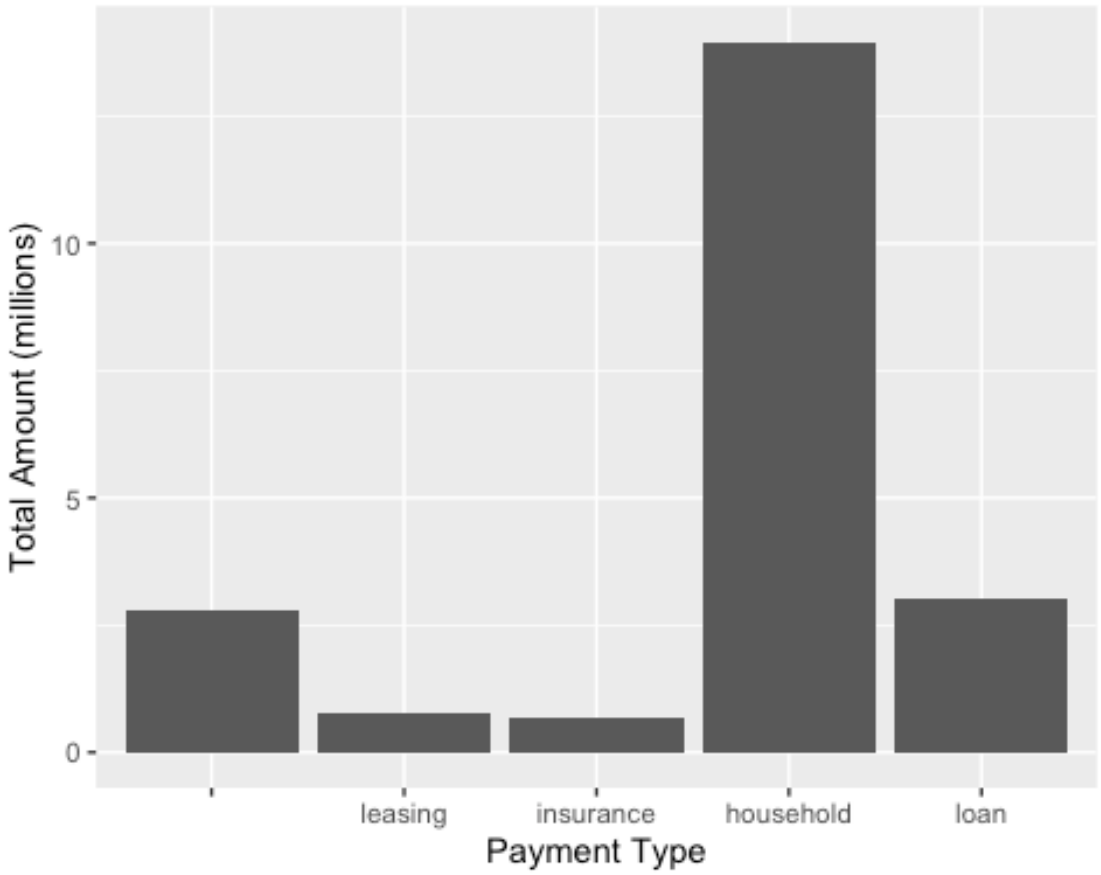


Figure 21: Payment Type

Cards

Card type

The majority of the cards issued by the bank are of the classic type. Since credit cards are not yet widely used by customers, offering premium cards can be a good opportunity to encourage their use.

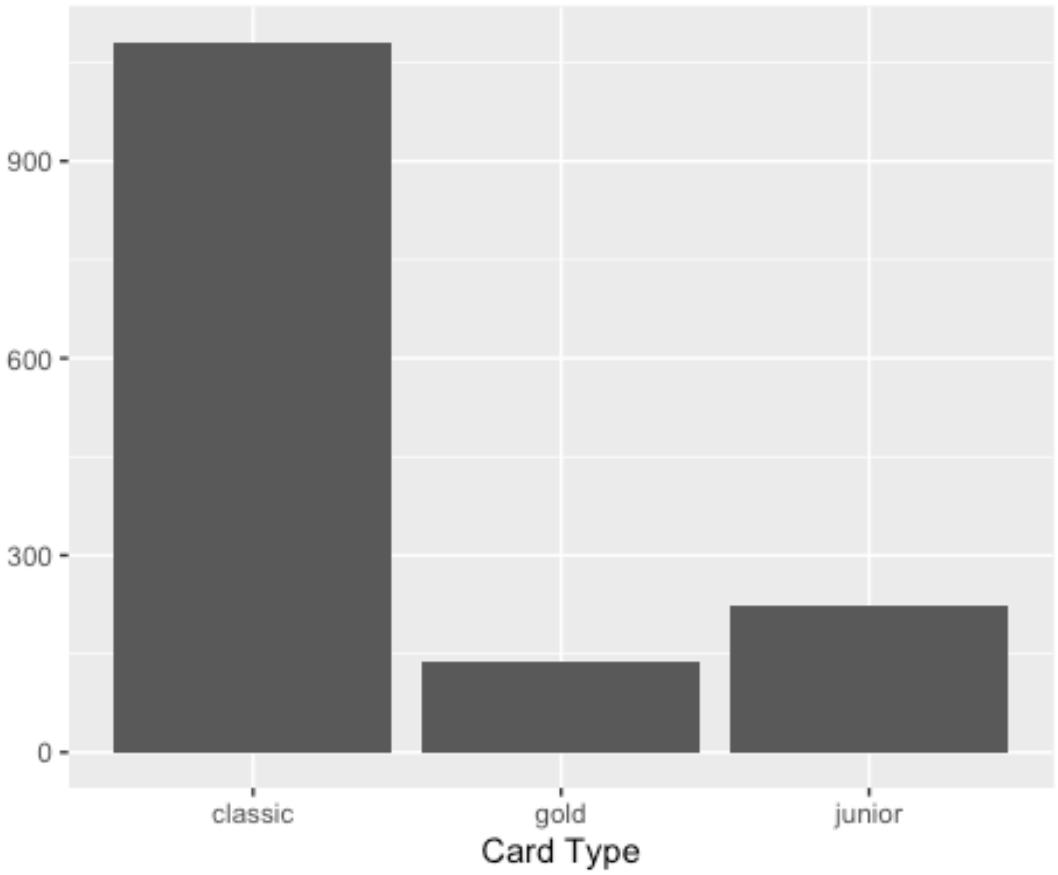


Figure 22: Card Type

Card type issued by year

There has been an increase in the number of cards issued by the bank pushed mostly by the classic card.

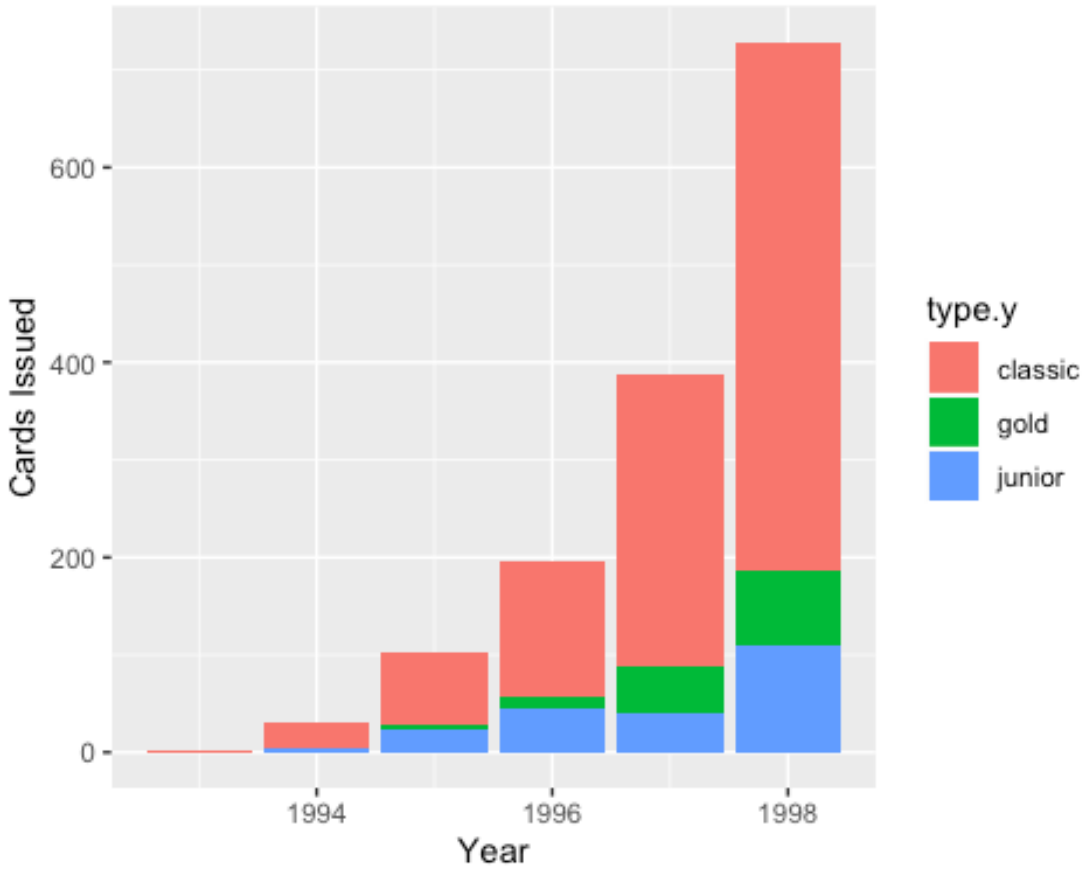


Figure 23: Cards issued by year

Credit card transactions by type

The growth in the number of credit cards has also been accompanied by a grow in the number of transactions done by credit card, mostly in the classic card.

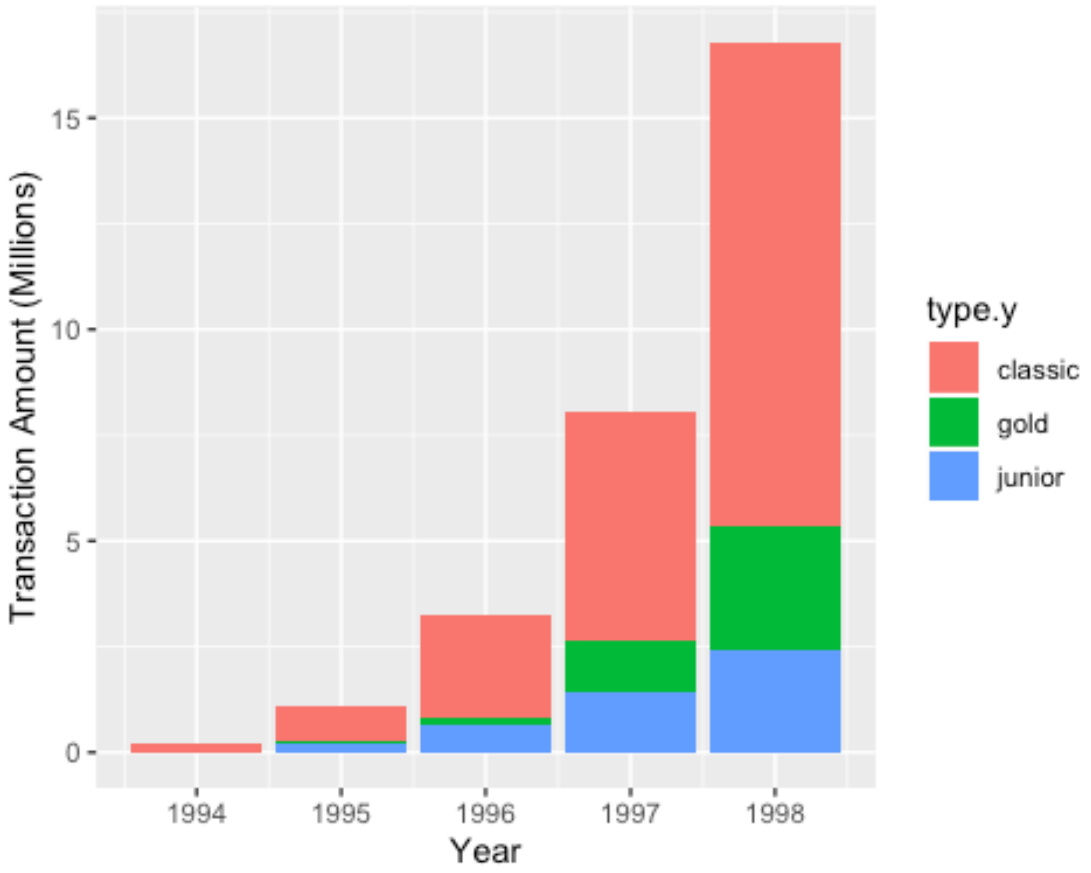


Figure 24: Amount of transactions by card type

Credit card transactions by type

The growth in the number of credit cards has also been accompanied by a grow in the number of transactions done by credit card, mostly in the classic card.

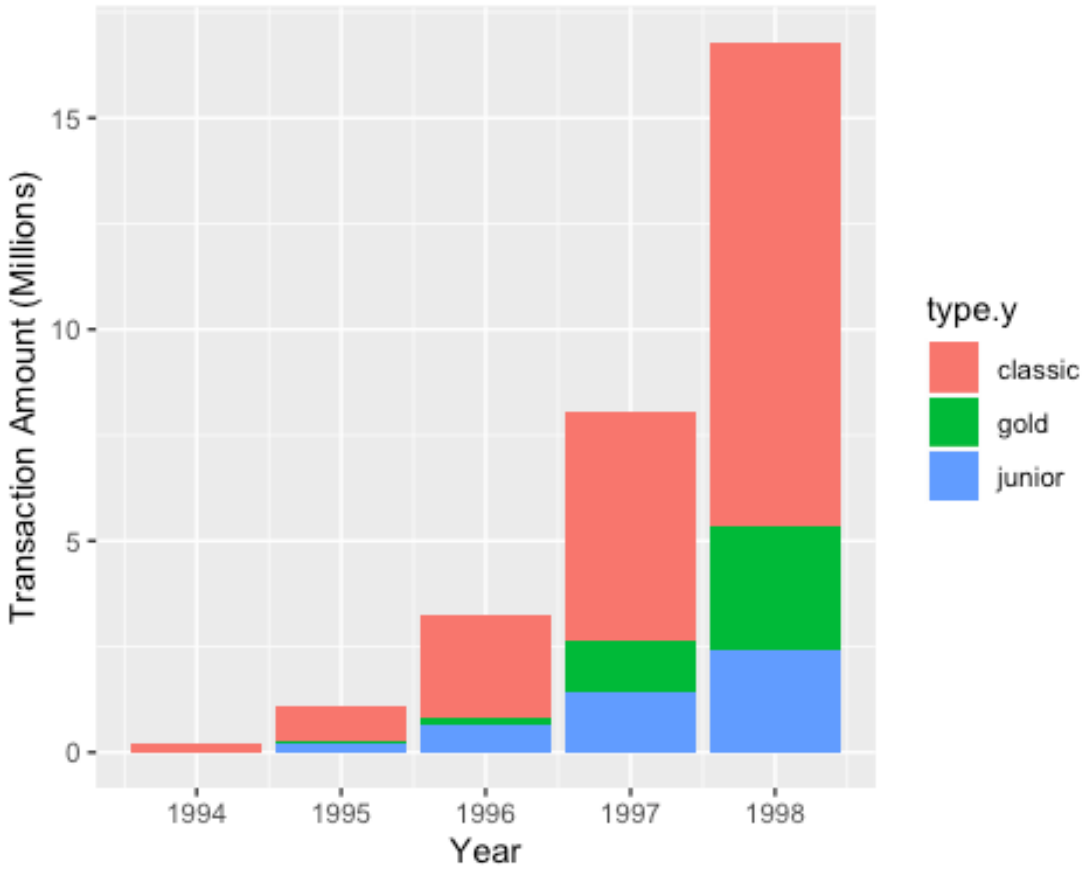


Figure 25: Amount of transactions by card type

Credit card transactions by month

As it happens with the amount of transactions there is also a spike in the use of credit cards every year in January, especially amongst the users of the classic card.

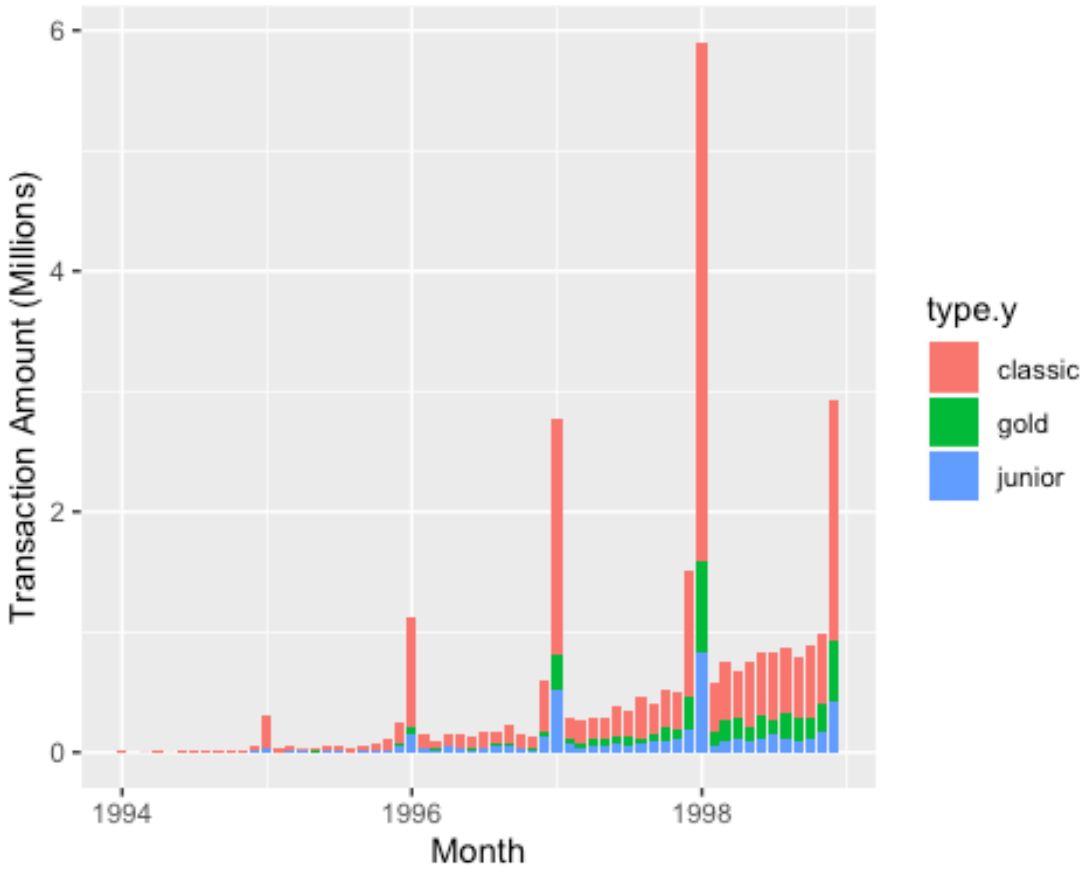


Figure 26: Amount of transactions by month

Amount of card transactions by account

The gold card has a slightly higher volume of transactions per client then the other cards. Looking at the clients with the classic card we see that there is a great number of outliers regarding the total of transactions with the card, those clients could be updated to the gold card.

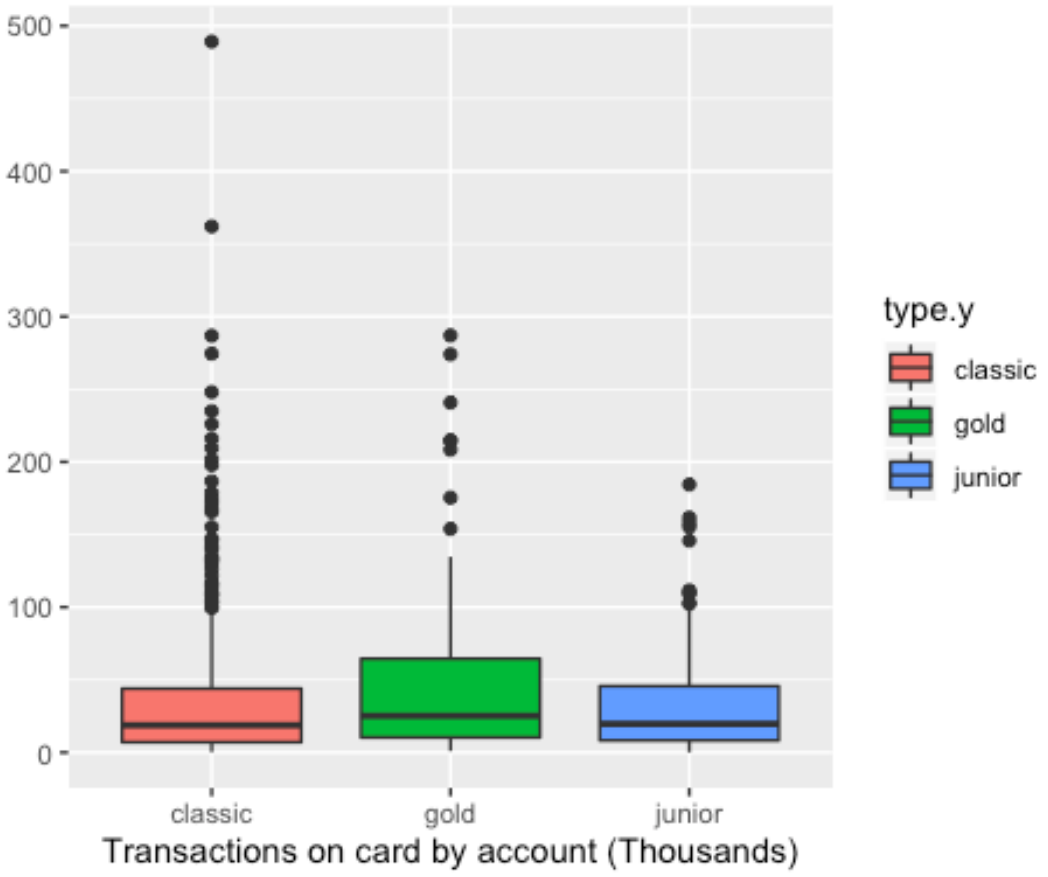


Figure 27: Amount of card transactions by account

Conclusion

The bank has seen growth in customer numbers but there is plenty of room to grow.

There is still little penetration in the different regions of the Czech Republic and the economic climate is very conducive to winning more customers. Recommendation: Focus on growth first in South Moravia, where population density is high and the bank does not have many clients yet. Focusing on growth in these regions, we can then expand to the rest of the country. Risks: It is important to keep in mind the possible risks of losing room for competitors in other regions.

The bank has a good repayment rate for loans made, especially those with lower values. Recommendation: Increase the volume of micro credits, which are performing well and may attract more customers to the bank. Risks: As we increase the volume, we increase our default risk, so we need to work directly with the credit bureau to ensure good customer reviews before lending.

A critical point noticed in the analysis is the large amount of withdrawals in July and January, leaving the bank balance negative. It is essential that the bank understands

the motivations of customers in these months, to anticipate and not suffer in these months. Recommendation: Conduct qualitative customer research to understand this movement. Based on this, structure action paths to minimize impacts. Risks: Customers may feel overrun when asked about their movements, so it is important to drive carefully to avoid dissatisfaction with the bank.

Finally, we see a growth in the number of cards issued in recent years, but still with little movement from them. Recommendation: Create more incentives for credit card use. This can be done by offering more premium cards and ensuring more credit per user. Risks: Because credit cards are not yet part of the Czech Republic's culture, we may have a period of learning how to use the card, which may cause delinquent invoices.

As next steps, we suggest further analysis and exploration of the above points.