PKDD'99 DISCOVERY CHALLENGE – "BERKA DATASET"

Submitted By:

Carl CHAHINE, Alessio ROSATO, Prineet Kaur BHURJI

(GROUP 5)

Problem Statement

Creating a Datamart to analyse the financial status of customers and find correlations between the various features useful for understanding the business trends to improve bank services.

Reference**:** https://sorry.vse.cz/~berka/challenge/pkdd1999/berka.htm

Data Exploration

We used Customer data that was been provided to us in the form of various tables like credit card, daily transactions, account, loan, demographics, disposition, orders, and client information to create the data mart. There are 5,369 unique clients and observations with 25 columns in our Datamart.

Data Preparation

1. The required libraries numpy, math, pandas, matplotlib were imported.

2. The datasets were read, and each record describes static characteristic of an account.

3. For Client Table new columns “birth_year”, “birth_month”, and “birth_day” were extracted from “birth_number” and a new column “Gender” was extracted from “birth_month”.

4. The “Age” and “Age Group” variables were calculated assuming current year as 1999.

5. The Disp and Card tables where merged using disp_id and Pivot Table created to fetch desired columns.

6. For Trans table –

- Converted “trans_id” to integer so that we can index it
- Checked that the column doesn't have values that should be in Operations
- Replaced the values in the type column with the value they should have and changing names for better understanding
- Found where the missing values of operations appear in relation with the type (Missing values in operations are either "Collection from another bank", "Credit in cash" or "Remittance from another bank")
- Filled missing values with “cc withdrawal” and “cash withdrawal” as per their % contribution in operation type
- Renamed the k_symbols to more understandable names while making sure there are no wrong values
- Saw that 99% of k_symbols that pay less than 30$ fall under “payment for statement”. Thus, we replaced all missing values <30 with “payment for statement”
- Used group by to analyse the missing values and thereby assigning them the suitable k_symbols
- Checked for the outliers in Amount and replaced them as per their quartile

7. For Order table fixed the names for all k_symbols and created a pivot table to get one row per client.

8. For Loan table checked outliers and replaced them as per their quartile in terms of amounts. Further grouped the loan table as per customer.

9. For District table renamed the columns named as “A1 – A16” to more meaningful names.

10. Explored joining all datasets using different approaches (In terms of Functions & Join Types) –

- Joined all Tables One by One Using CONCAT (INNER JOIN)
- Joined all Tables One by One Using MERGE (FULL JOIN)
- Joined all Tables Together Using MERGE (Every Client should have one row)

11. Fixed types for 'client_id', 'account_id', 'A', 'B', 'C', 'D', 'disp_id' and 'card_id' from “Float” to “Integer”.

12. Final Datamart –

```
In [37]: Join_total.head()
```

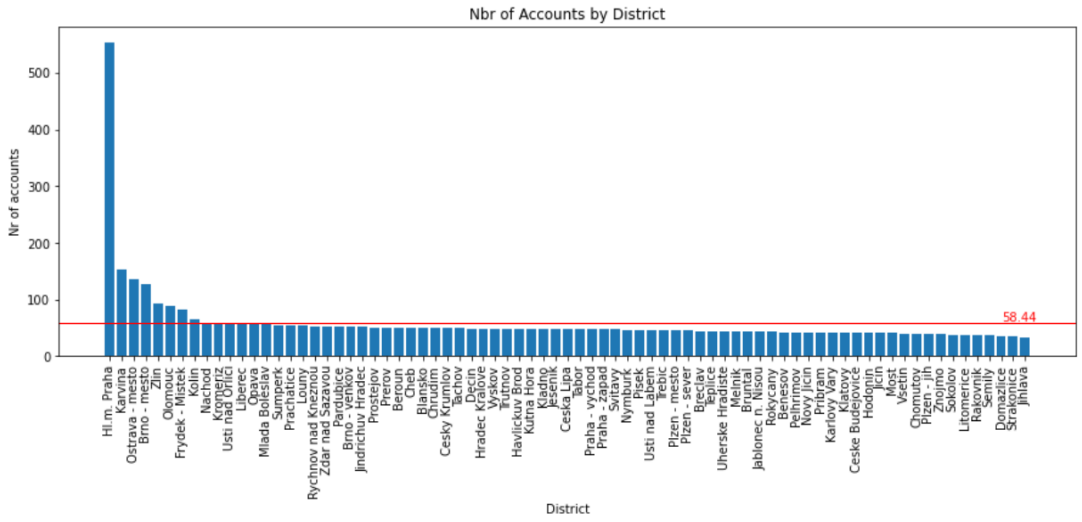
Out[37]:

	client_id	account_id	A	B	C	D	disp_id	card_id	LEASING	household_order	...	loan payment_trans	old age pension_trans	payment for statement_trans	sanctions_trans	birth_
0	1	120901	1	0	2	0	0	0	0.0	0.0	...	0.0	0.0	0.0	0.0	
1	2	1641922	34	3	43	4	0	0	0.0	0.0	...	0.0	0.0	0.0	0.0	
2	3	1641922	34	3	43	4	0	0	0.0	0.0	...	0.0	0.0	0.0	0.0	
3	4	182281	2	0	8	0	0	0	0.0	0.0	...	0.0	0.0	0.0	0.0	
4	5	182281	2	0	8	0	0	0	0.0	0.0	...	0.0	0.0	0.0	0.0	

5 rows × 25 columns

Data Visualization

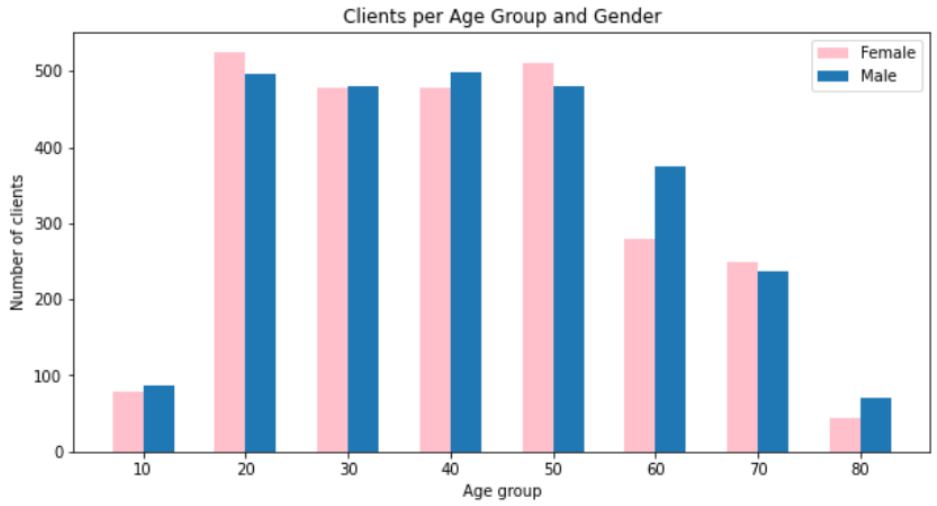
- *Number of Accounts by District*



From the above graph we can see that ‘Hl.m.Praha’ district has exceptionally high no. of accounts (500+) followed by 8 other districts (like ‘Karvina’, ‘Ostrava-mesto’, ‘Brno-mesto’ etc) which have total no. of accounts more than the mean (i.e. 58.44).

This may be because of many reasons like – ease of account opening, attractive offerings, seamless experience, and service provided to the customers or just mainly because of the higher population.

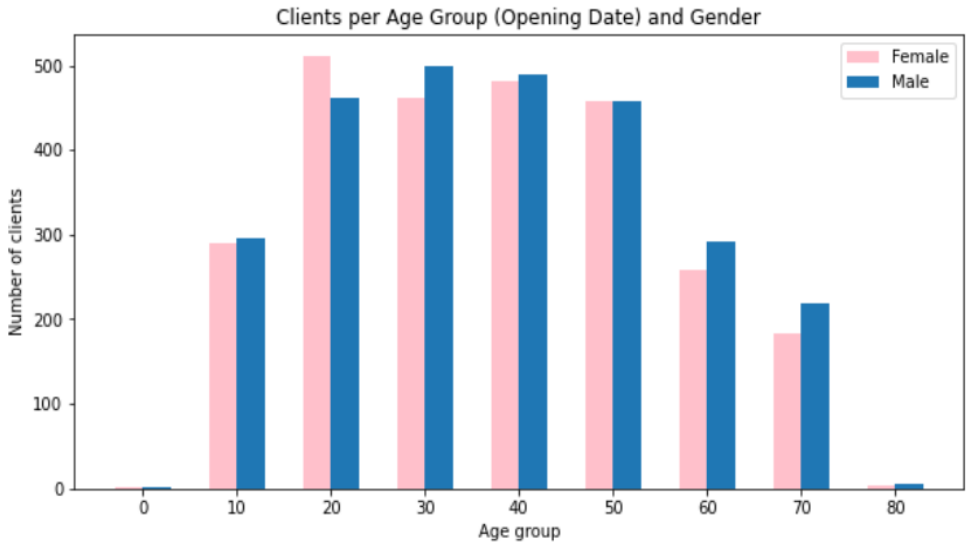
- *Number of Clients per Age Group and Gender*



After a low number of clients in age group 10, we can observe that the age groups from 20 to 50 contribute maximum to the bank's number of clients. As the age further increases, a downward trend is observed for age groups 60 and 70 and the minimum amount is reached with the age group of 80's, as expected.

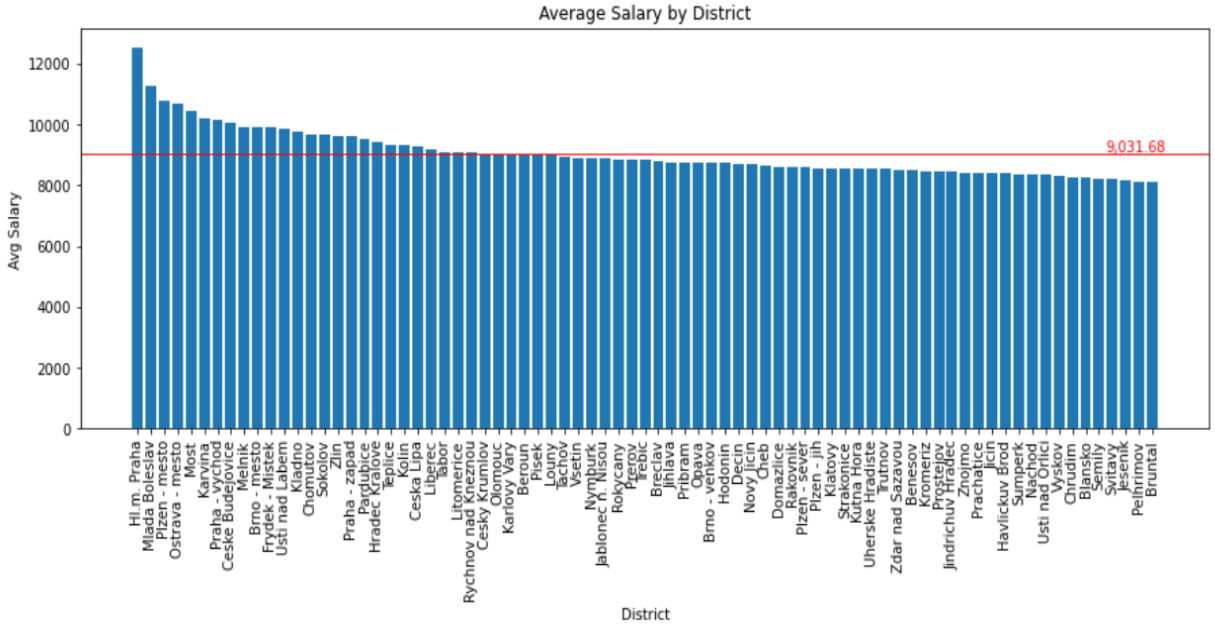
On the other hand, there is no obvious trend seen among the gender groups to see an inclination towards a specific gender. Across the age groups, the no. of female to male clients are almost comparable with the only exception for the 60's age group where the total no. of male clients is almost 133% of female clients.

- *Number of Clients per Age Group and Gender at the Time of Account Opening*



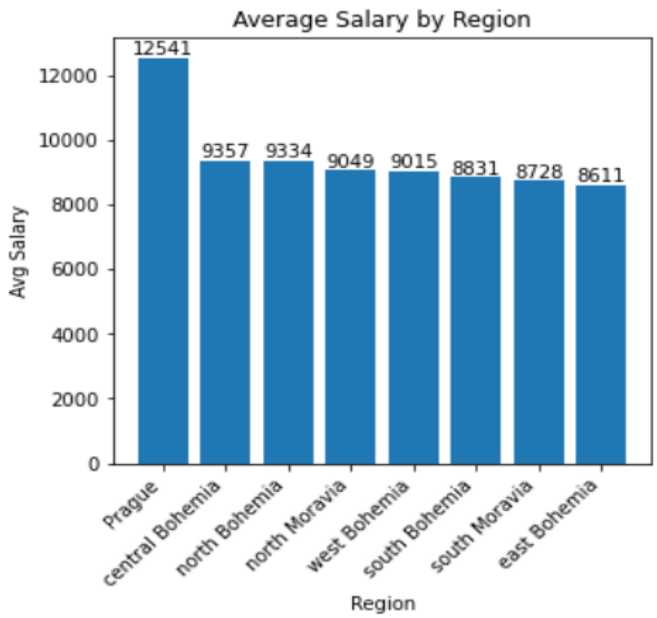
We see from the above graph that there are a very few observations in the two extreme bins. This supports the fact that to open a bank account as account owner you need to be of legal age (usually 18+). On the other hand, very few people open accounts in their 80's, which could be because they are already having an account by that age.

- *Average Salary by District*



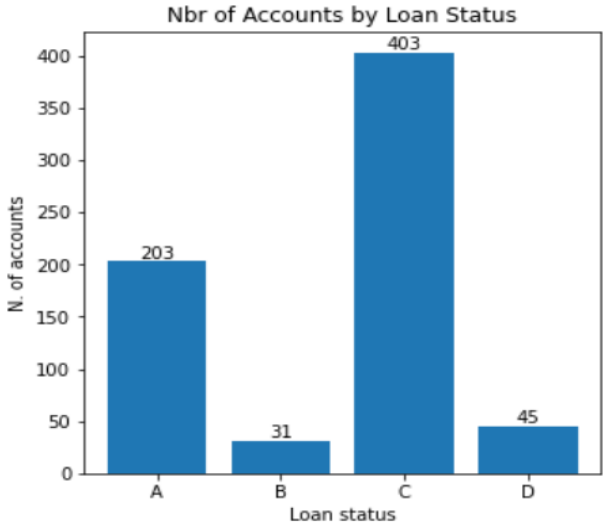
The above graph shows that 'Hl.m.Praha' is the district with highest avreage salary followed by 'Mlada Boleslav' and 'Plzen – mesto'. Further ~ 30% of the districts have their average salaries more than the mean salary (which happens to be 9,031).

- *Average Salary by Region*



It can be easily distinguished that 'Prague' region is offering much lucrative salaries as compared to any of the other regions, whose values are all very similar to each other. One may even say that the region has more well-established businesses as compared to others and that the companies are having a vision of rewarding people with good pay scales.

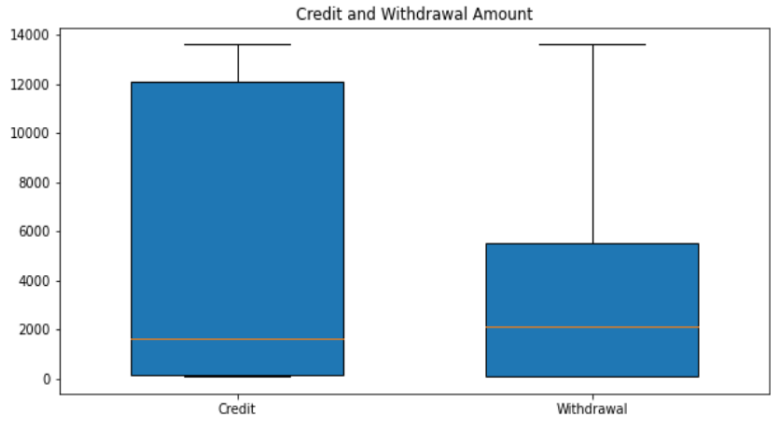
- *Number of Accounts by Loan Status*



The maximum number of accounts (i.e. 403) fall under loan category C (contract is running and ok so far), followed by 203 accounts for loan category A (contract finished with no problems). There are 45 and 31 accounts that fall under loan category D & B respectively where loan payment is stuck (the contract may be running or finished).

Thus, almost 15% of running contracts and 11% of finished contract loan payments are pending and these are the accounts the bank should work to ensure the loan payments happen.

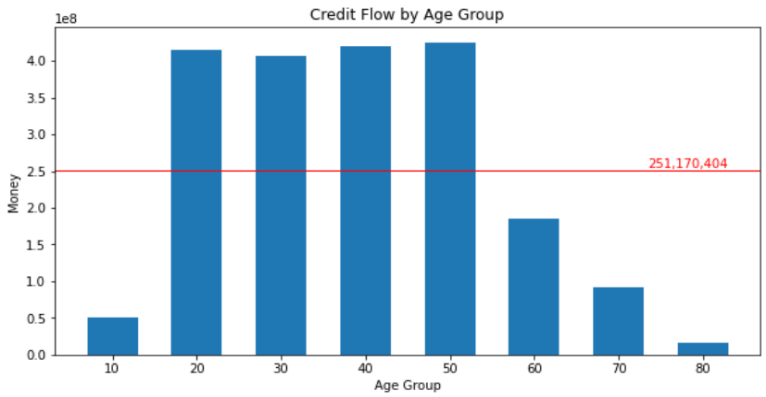
- *Credit and Withdrawal by Age Groups –*
    - *Credit and Withdrawal Amounts*



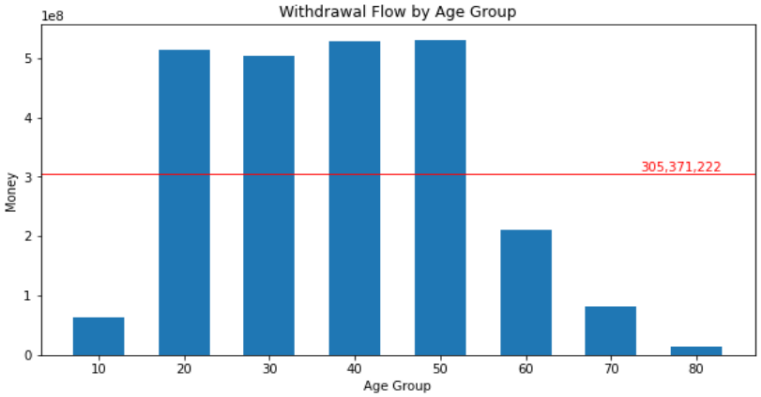
The above boxplot shows that while the interquartile range for Credit ranges between 0-12,000, the one for Withdrawal ranges between 0-6,000. Therefore, the interquartile range for Credit is almost twice as much.

However, the median values are quite similar (~2,000). This suggests that while the withdrawal is more concentrated towards lower amounts, the credit amounts follow a more uniform distribution across min and max values.

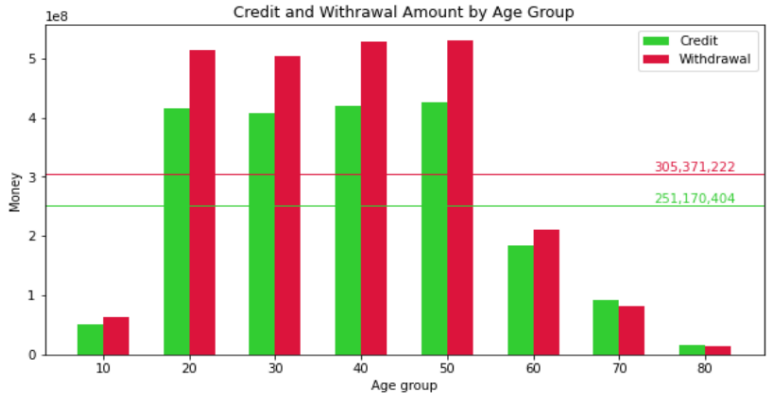
- *Credit Flow by Age Group*



- *Withdrawal Flow by Age Group*



- *Credit and Withdrawal Amount by Age Group*

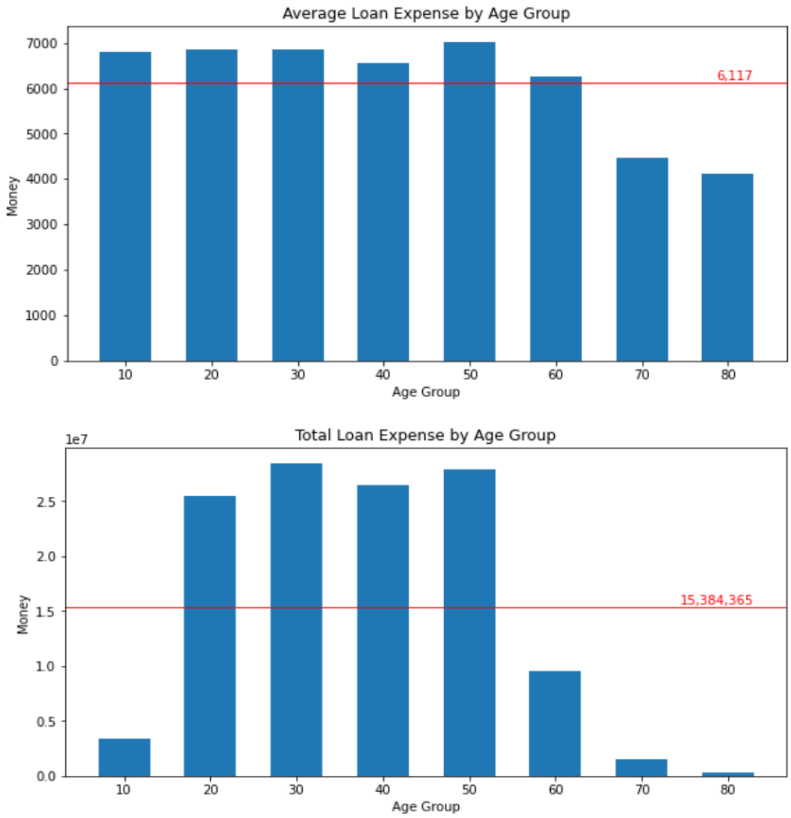


While the first two graphs (Credit flow and Withdrawal flow grouped by age) show that the age groups from 20 to 50 are the main contributors to both the transaction types, the third graph helps us compare the two types to get some further insights.

It is interesting to observe that all age groups (irrespective of how much total amount of transactions they do) have more withdrawals than credit inflow. Only the last two age groups (i.e. 70-80 years) as an exception follow an opposite trend with their credit inflows being more than their withdrawals.

One may conclude that people of the age groups 10-60 can be classified as "*Spenders*", while on the other hand people having age between 70-80 can be considered to be the "*Savers*".

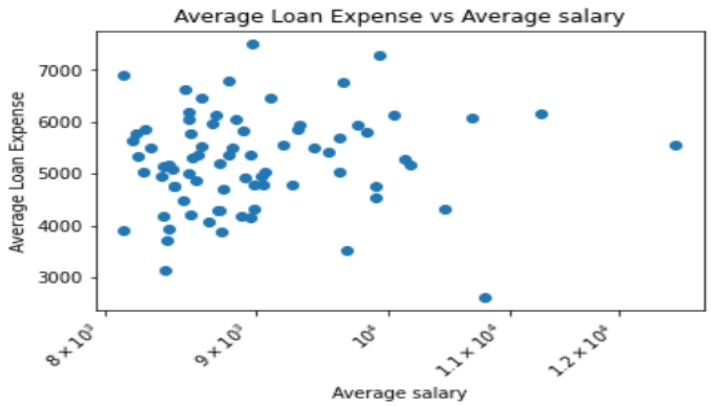
- *Average vs Total Monthly Loan Expenses by Age Group*



From the first graph above we can say that the average monthly loan expenses are very similar (all above the mean of 6,117) for age groups 10-50, while being very close to the mean for clients in the 60's and below mean for the older groups.

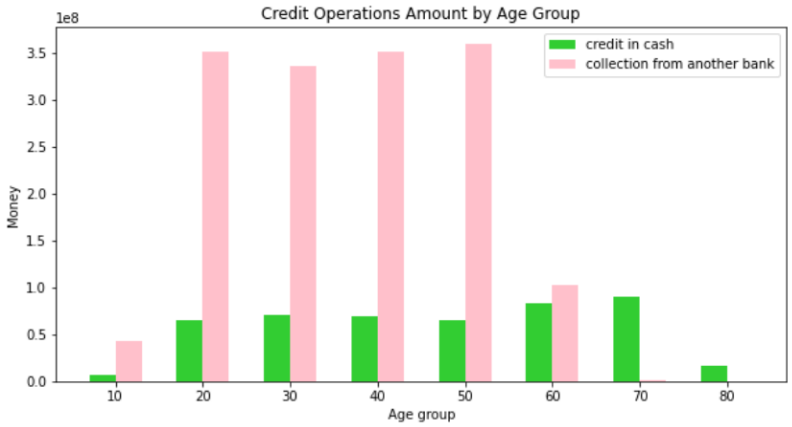
However, by looking at the total loan expense, we can see that the age groups spending the most are the ones between 20-50, while people below 20 and above 60 years old (especially above 70) spend much less. This is supported by the fact the few people below 20 sign loans and few people are still repaying their loans in their 70's or 80's (here we are assuming the bank is not willing to issue new loans to older people because of a higher mortality risk).

- *Average Monthly Loan Expense against Average salary*

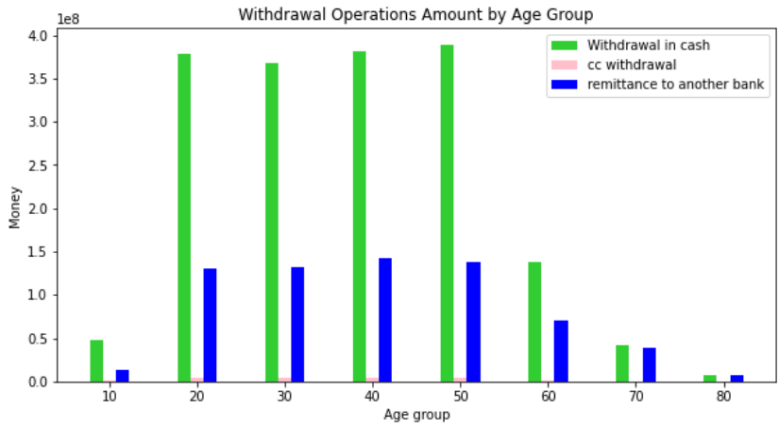


Above scatter plot shows that there is no correlation between average loan expense and average salary.

- *Operations Types by Age Groups*



The above graph for 'credit operations amounts by age group' depicts that the age groups 10-50 highly prefer direct collection from other banks over credit in cash. While, people in their 60's can be said to be liking both operation types almost equally, people of the age group 70-80 rely completely on cash credits with almost nil direct collections from other banks.



The above graph for 'withdrawal operations amounts by age group' depicts that across all age groups from 10-60 withdrawal in cash is highly preferred mode over remittance to another banks. But specifically for 70-80 age groups the acceptance towards both types of withdrawal operations can be considered almost the same.

On the other hand, withdrawal through credit card is not at all a likeable mode across the age groups, with a very few exceptional cases observed under 20-50 age groups.

Appendix:

(Final Table Variable explanation)

The final merged dataset has 25 variables. Following is the table which describes each column name –

Column Names	Description
client_id	Client number
Account_id	Client account number
A	Status of loan – contract finished; no problems
B	Status of loan – contract finished; loan not payed
C	Status of loan – running contract; ok so far
D	Status of loan – running contract; client in debt
District_id	Client district identification
Disp_id	Disposition to the account
Card_id	Client card identification
LEASING	Leasing amount in order table
Household_order	Household amount in order table
Insurance_order	Insurance amount in order table
loan_order	loan amount in order table
Household_trans	Household amount in trans table
Insurance_trans	Insurance amount in insurance table
interest credited_trans	Interest amount in trans table
loan payment_trans	Loan amount in trans table
old age pension_trans	Old age pension amount in trans table
payment for statement_trans	Payment for statement in trans table
sanctions_trans	Sanctions amount in trans table
Birth_year	Year of birth for client
Birth_month	Month of birth for client
Birth_day	Day of birth for client
Gender	Gender of the client (male or female)
Region	Client region