FINANCIAL PROGRAMMING

GROUP 3 PROJECT REPORT-FINANCIAL DATASET

Fajar Tri **ANGGORO**

Harikrishnan **GOPALAKRISHNAN**

Tristan **HELLE**

Table of Contents

Identifying Business Problem

A Base Table refers to one large table where it usually contains data from multiple sources. In business, a base table could contain different kinds of information and usually have what is defined as granularity. A granularity refers to the lowest level of instances within our data and the value should be unique for each instance.

A Base Table provides ease for a data analyst/data scientist to do some exploratory analysis as it contains a lot of information. In addition, a base table is also needed when one wants to develop a machine learning model as it usually has all of the variables needed.

In this case, we will develop a Base table from our financial dataset and do some exploratory data analysis from it.

Data Cleaning & Preparation

We will be using a financial dataset from a bank in czech republic. The dataset was created in 1999 and it contains information from 1993 - 1998. There are 8 tables within our dataset:

- Account table, contains information about accounts
- Client table, contains information about clients
- Disposition table, contains information about clients and it's disposition in a certain account
- Order table, contains information about payment orders
- Transaction table, contains information about transactions
- Loans table, contains information about loans given to an account
- Credit Cards table, contains information about credit cards issued to an account
- District table, contains demographic information

The granularity of the base table has been defined as a client who is an account owner. In addition, the time window of the independent variables is 1996 (1 year) while the time window of the dependent variables is 1997 (1 year).

Account Table

We start building our base table from the account table, this table contains information about account id, date of the opening of the account, and the district id in which the account was opened. This table contains 4500 accounts without any missing values. As the time window of the independent variables is 1996, we will only consider accounts that were created before 1996 (1993, 1994, 1995). The year of the opening of the account was extracted from the date information, and then filtered for years before 1996. In addition, the Length of Relationship (LOR) is also calculated.

Disposition Table

This table contains information about client id, account id, and the disposition type (OWNER or DISPONENT). It contains 5369 disponents without any missing values. A filter is applied to remove the disponents data, as we will only consider account owners in our base table.

Client Table

:count_id	bank_district_id	frequency	date	year	lor	disp_id	client_id	type	birth_number	client_district_id	birth_year	birth_day	birth_month	gender
576	55	POPLATEK MESICNE	930101	1993	3	692	692	OWNER	365111	74	1936	11	1	F
3818	74	POPLATEK MESICNE	930101	1993	3	4601	4601	OWNER	350402	1	1935	2	4	M
704	55	POPLATEK MESICNE	930101	1993	3	844	844	OWNER	450114	22	1945	14	1	M
2378	16	POPLATEK MESICNE	930101	1993	3	2873	2873	OWNER	755324	16	1975	24	3	F
2632	24	POPLATEK MESICNE	930102	1993	3	3177	3177	OWNER	380812	24	1938	12	8	M

Client table contains information about the client's birth information and it's district. It contains 5369 disponents without any missing values. The client table is merged with the disposition table and then merged with the account table. After, more information about the client (Client birthday, gender, age, and age group) are added as well.

Transaction Table

This table contains information about transactions (credit and withdrawal), account id, date of transaction, the type of transaction, etc. It has more than a million rows and a lot of missing values. First, we filtered the data so that it only contains transactions from 1996. After, we divided the table into two parts, one which contains information about withdrawals, and one about credits. From both of these tables, RFM variables are created by aggregating the data on the account level. Recency is determined by the most recent date of the transaction, Frequency is determined by the number of transactions, while Monetary is determined by the total and average value of transactions. Unused columns are then dropped and both tables are merged with our base table.

total_credit	average_credit	frequency_credit	most_recent_credit_date	total_withdrawal	average_withdrawal	frequency_withdrawal	most_recent_withdrawal_date
76097.3	3170.72	24.0	961231.0	70419.2	1853.14	38	961231
234806.4	9392.26	25.0	961231.0	223535.2	3062.13	73	961231
228514.9	9521.45	24.0	961231.0	218531.2	3642.19	60	961231
664545.4	19545.45	34.0	961231.0	633310.2	12924.70	49	961231
186658.9	7777.45	24.0	961231.0	191071.2	2582.04	74	961231

After merging, one missing value is detected and after investigation, it turns out that there is one account without any credit transaction in 1996. The missing value is replaced by 0.

```
# Check which data has NA values
base[base.isna().any(axis = 1)]
```

	account_id	bank_district_id	frequency	date	year	lor	disp_id	client_id	type	birth_number	...	age	age_group	total_credit	average_credit	fr
1815	1720	35	POPLATEK MESICNE	950516	1995	1	2086	2086	OWNER	546029	...	42	40	NaN	NaN	

District Table

This table contains regional information about unemployment rate, number of inhabitants, number of entrepreneurs, etc. It contains 77 districts without any missing values. We filtered the table so it remains the district name and the region name columns. This filtered column is then merged with our base table.

Order Table

This table contains information about payment orders, account ids, recipient accounts, etc. It contains 6471 orders without any missing values. Information from this table will not be included in our base table because this table does not contain any information about the date of the orders.

Loan Table

This table contains information about loans, the date of the loan was granted, the account id who received the loan, the amount, the status, etc. It has 682 different loans without any missing values. The information from the loan table will be extracted as one of our dependent variables in the base table. The table is filtered so that it only contains loans granted in 1997. A new column is also added to identify accounts that have been granted loan in 1997. After, unused columns are dropped and the table is merged with our base table.

After merging, missing values are detected and this is because not every account has been granted a loan. Missing values are replaced by 0 and we now have an identifier of whether an account has been granted a loan in 1997 (the value corresponds to 1 if a client received a loan in 1997 and 0 otherwise).

total_withdrawal	average_withdrawal	frequency_withdrawal	most_recent_withdrawal_date	A1	A2	A3	loan_id	status	had_granted_loan_in1997
70419.2	1853.14	38	961231	74	Ostrava - mesto	north Moravia	NaN	NaN	0
223535.2	3062.13	73	961231	1	Hl.m. Praha	Prague	NaN	NaN	0
218531.2	3642.19	60	961231	22	Domazlice	west Bohemia	NaN	NaN	0
633310.2	12924.70	49	961231	16	Jindrichuv Hradec	south Bohemia	NaN	NaN	0
191071.2	2582.04	74	961231	24	Karlovy Vary	west Bohemia	NaN	NaN	0

Credit Cards Table

This table contains information about credit cards issued to a certain disposition, the type of the card, and the date of the card issuance. It has 892 cards without any missing values.The information from the credit cards table will be extracted as one of our dependent variables in the base table. The table is filtered so that it only contains cards issued in 1997. This table is also merged with the disposition table because it has no information on neither account id nor client id. A new column is also added to identify records that have been issued a credit card in 1997. After, unused columns are dropped and the table is merged with our base table.

frequency_withdrawal	most_recent_withdrawal_date	A1	A2	A3	loan_id	status	had_granted_loan_in1997	type_x	had_creditcard_issued_in1997
38	961231	74	Ostrava - mesto	north Moravia	NaN	NaN	0	NaN	0
73	961231	1	Hl.m. Praha	Prague	NaN	NaN	0	NaN	0
60	961231	22	Domazlice	west Bohemia	NaN	NaN	0	NaN	0
49	961231	16	Jindrichuv Hradec	south Bohemia	NaN	NaN	0	NaN	0
74	961231	24	Karlovy Vary	west Bohemia	NaN	NaN	0	NaN	0

After merging, missing values are detected and this is because not every account has been issued a credit card. Missing values are replaced by 0 and we now have an identifier of whether an account has been issued a credit card in 1997 (the value corresponds to 1 if a client received a credit card in 1997 and 0 otherwise).

Final Output

Our final base table consist of a total 31 variables and information about 2239 distinct clients who is an account owner. Information ranges from their account info (year opened, length of relationship), client info (Gender, Age, Age group), transaction behavior (Recency, Frequency, Monetary) during 1996, whether the client has been granted a loan in 1997, and whether the client has been issued a credit card in 1997.

Dataset Description

No	Variable Name	Description	Type	Remarks

1	account_id	Account identifier	Categorical	
2	bank_district_id	location identifier of the branch	Categorical	
3	frequency	frequency of issuance of statements	Categorical	"POPLATEK MESICNE" stands for monthly issuance "POPLATEK TYDNE" stands for weekly issuance "POPLATEK PO OBRATU" stands for issuance after transaction
4	date_opened	date of creating of the account	Numeric	in the form YYMMDD
5	year	year of creating of the account	Numeric	
6	lor	Length of Relationship in years	Numeric	Calculated from 1996
7	disp_id	disposition identifier	Categorical	
8	client_id	client identifier	categorical	
9	type	type of disposition (owner/user)	categorical	Only account owners are available within the dataset
10	birth_number	Date of birth	numeric	the number is in the form YYMMDD for men the number is in the form YYMM+50DD for women where YYMMDD is the date of birth
11	client_district_id	address identifier of the client	categorical	

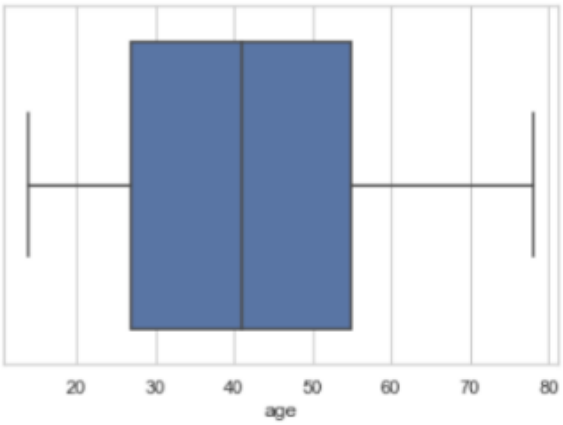
12	birth_year	Year of birth	numeric	
13	birth_day	day of birth	numeric	
14	birth_month	Month of birth	numeric	
15	gender	Gender of the client	categorical	M = Male F = Female
16	age	Age of the client	numeric	Calculated from 1996
17	age_group	Group age of the client	numeric	
18	total_credit	Total credit during 1996	numeric	Calculated only in 1996 (independent variable timeline)
19	average_credit	average credit during 1996	numeric	Calculated only in 1996 (independent variable timeline)
20	frequency_credit	credit transaction frequency during 1996	numeric	Calculated only in 1996 (independent variable timeline)
21	most_recent_credit_date	Last date of credit transaction	numeric	Calculated only in 1996 (independent variable timeline)
22	total_withdrawal	Total credit during 1996	numeric	Calculated only in 1996 (independent variable timeline)
23	average_withdrawal	average credit during 1996	numeric	Calculated only in 1996 (independent variable timeline)
24	frequency_withdrawal	credit transaction frequency during 1996	numeric	Calculated only in 1996 (independent variable timeline)
25	most_recent_withdrawal_date	Last date of credit transaction	numeric	Calculated only in 1996 (independent variable timeline)
26	client_district	client district name	string	
27	client_region	client region name	string	
28	loan_id	loan identifier	categorical	a value of 0 means the client has not

				been granted a loan in 1997
29	loan_status	loan status	categorical	a value of 0 means the client has not been granted a loan in 1997
30	had_granted_loan_in1997	status whether a client has been granted a loan in 1997	categorical	1 = granted a loan in 1997 0 = not granted a loan in 1997
31	creditcard_type	credit card type	categorical	a value of 0 means the client has not been issued a credit card in 1997
32	had_creditcard_issued_in1997	status whether a client has been issued a credit card in 1997	categorical	1 = issued a credit card in 1997 0 = not issued a credit card in 1997

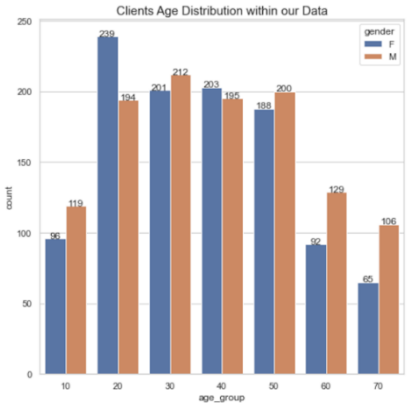
Exploratory Data Analysis

Age:

We can see thanks to this boxplot that the age of the youngest clients is 14 and the age of the oldest clients is 78. We can see that there are no real outliers as we have around 20 values for both 14 and 78 years old. The mean age of the clients is 42.

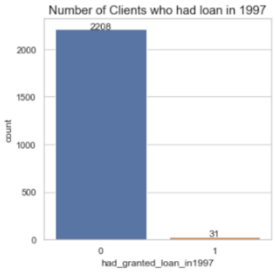


The bar chart also allows us to complete the boxplot analysis. We can see that the division of the different quartiles reflects well the distribution within the different age groups. The young (10-20 years old) and oldest (60-80 years old) are not the most abundant.



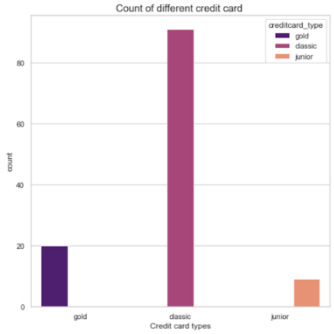
Here we can see that the distribution between male and female is rather equal for all group categories except for some special age groups such as the 20-30 and 60-70 & 70-80 years old.

Loans:



Thanks to this graph we can see that out of all the clients only 31 were able to have their loan granted in 1997

Credit cards:



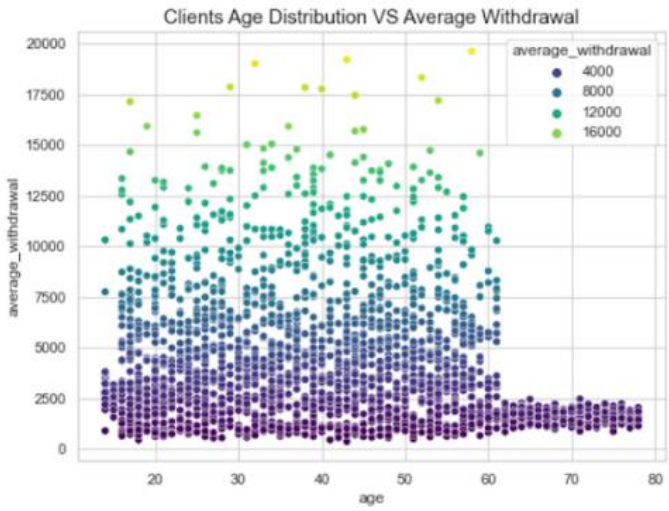
120 of the clients were able to get a credit card in 1997

From the graph below we can observe that first of all the classic credit card is by far the most common one.

Relationship between client age and average withdrawal:

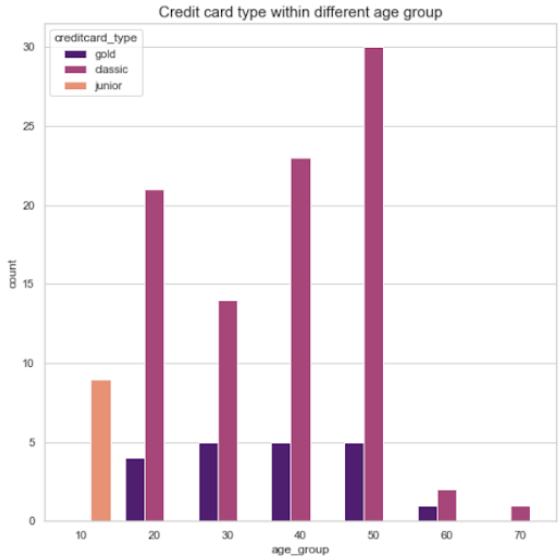
We can see on the scatter plot that there is a direct correlation between age and the average withdrawal of a person. This is only clearer when we look at the data of clients above 60 years old.

This is surely due to a lifestyle that is less prompt to consume as age increases.

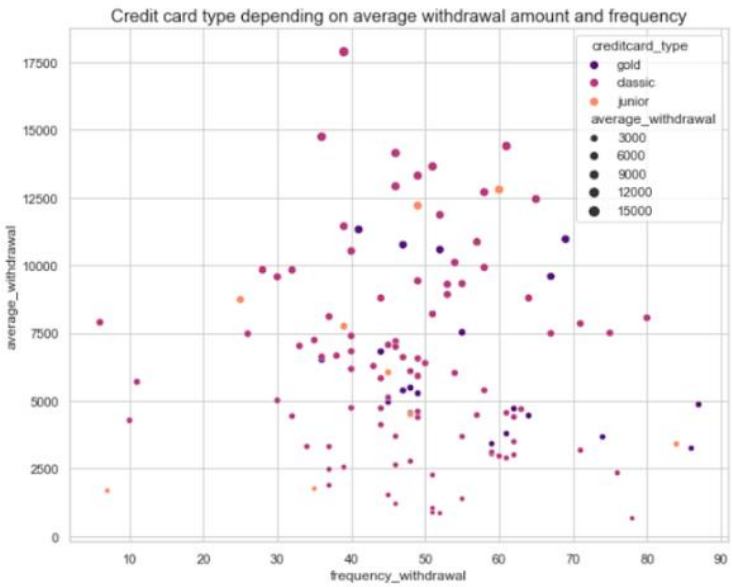


Depending on the different goals and perceptions of the bank this can be a very interesting insight. This can also be very important to take into account within the credit scoring algorithm.

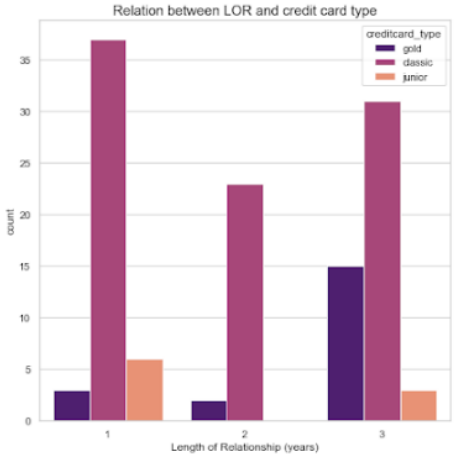
Relationship between credit card and age, withdrawal, and frequency.



From the first graph, we can see that the majority of the population has classic cards. We can also see that the gold credit cards are only assigned to people who are young adults & adults as the older people get the lower their average spending is. This means that owning a gold credit card has fewer advantages as they are less prompt to consume and are therefore logical.



The graph above provides information on the amount that is spent and the frequency of withdrawal. We can see that there is a correlation between the increase of withdrawal and the increase of frequency regarding the type of card that the clients own.



We can see that there is a relationship between the type of credit card that a customer has and the Length of the Relationship. Customers that have been with the bank for 3 are more likely to have a gold card. This can be explained by loyalty benefits/advantages as well as an increase in the trust of the bank's capacities to provide added value for the customer.

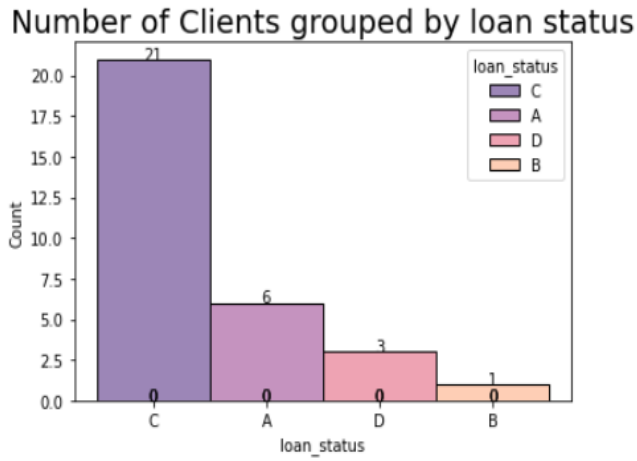
Loans:

Loan status:
The loan status is categorized into 4 types. They are:

1. **'A'** stands for contract finished, no problems.
2. **'B'** stands for contract finished, loan not paid.
3. **'C'** stands for running contract, OK so far.
4. **'D'** stands for running contract, client in debt.

Let us find the whether the Loan status feature has a relation with other features

Clients grouped by loan status

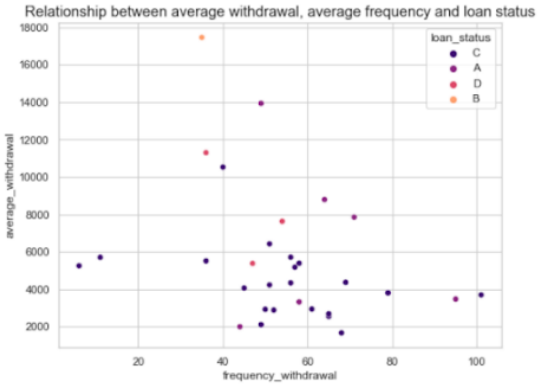


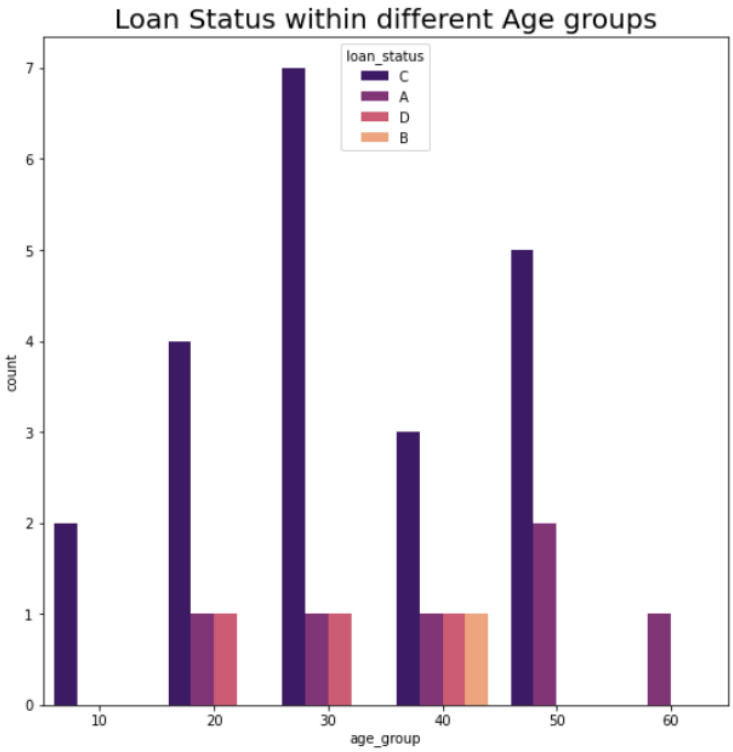
From this histogram plot, we can observe the number of clients grouped by loan status. It is evident that almost 68% of the total client's loan status is in running contract i.e in the loan repayment period and they paid the amount on time without causing any hassle to the bank.

Loan status in each gender group

Here From this Graph, we can see the relationship between Gender and loan status. Both Male and Female clients' loan status is active and made their payment on time so far(**"C"**).On the other hand, Category **"B"** and **"D"** represent the loan status where the loan is not paid and the client is in debt respectively. There are 2 Female and 5 male clients under these categories. According to the bank, these clients are considered as 'Bad clients'.

We can see on the graph that is on the right-hand side that people who remain below 6000 average withdrawal are almost all in good loan conditions (loans A ‘contract finished, no problems’ & C ‘running contract, OK so far’) while other customers seem to have more problems remaining on track with their loans (B & D status loans)

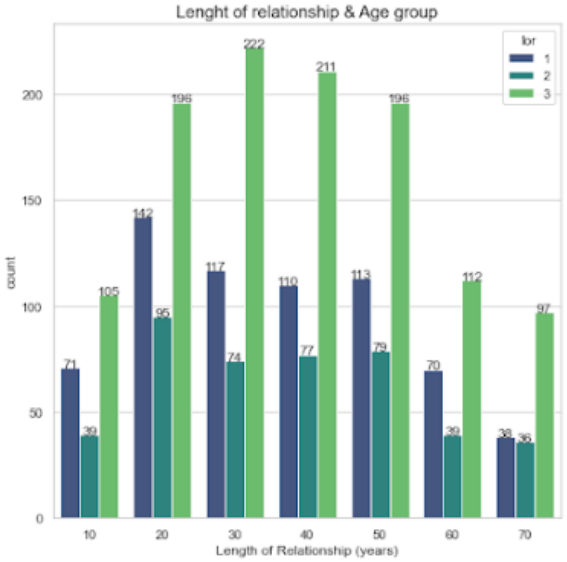




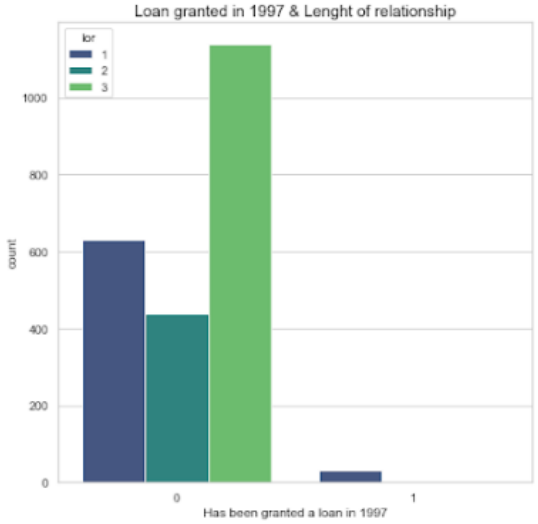
The relationship between the age and type C loan is clear as from the 10 - 30 year old categories we can see that there is an increase. This is very common as people in this age group tend to do important purchases (cars, houses, entrepreneurship, …)

We can see that after their 40s these kinds of loans tend to decrease until there are none left in the 60 - 80 age group.

Length of Relationship

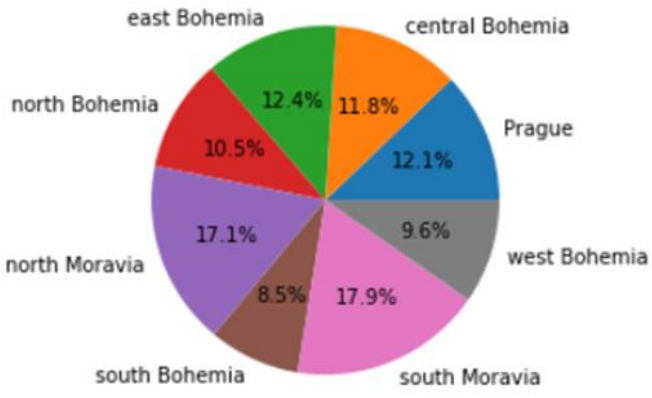


Thanks to the visualizations we can see that there is an even distribution between the length of relationship and the different age groups within this customer segment.



The graph on the left-hand side shows that the totality of the graphs that were accepted have been granted to new customers only. This can be seen as a weakness from the bank as it shows that they might be using loans as a feature to attract new customers. It is surprising to see that loans are not granted to customers that have a longer relationship than 1 year.

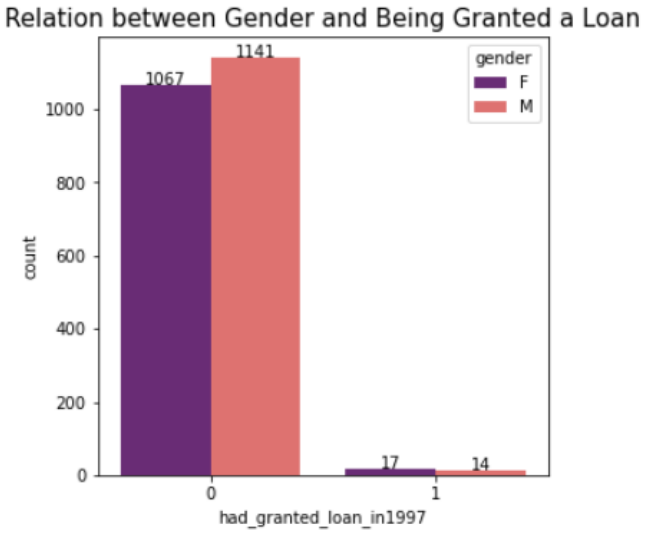
Proportion of clients per region



client_region	No_of_clients
Prague	272
central Bohemia	265
east Bohemia	277
north Bohemia	235
north Moravia	383
south Bohemia	190
south Moravia	401
west Bohemia	216

Here from this Pie-chart, We can observe the Proportion of clients with respect to the Regions. South Moravia is the region with the highest proportion of clients(17.9%) i.e 401 clients and followed by Region North Moravia with 17.1% proportion i.e 383 clients.

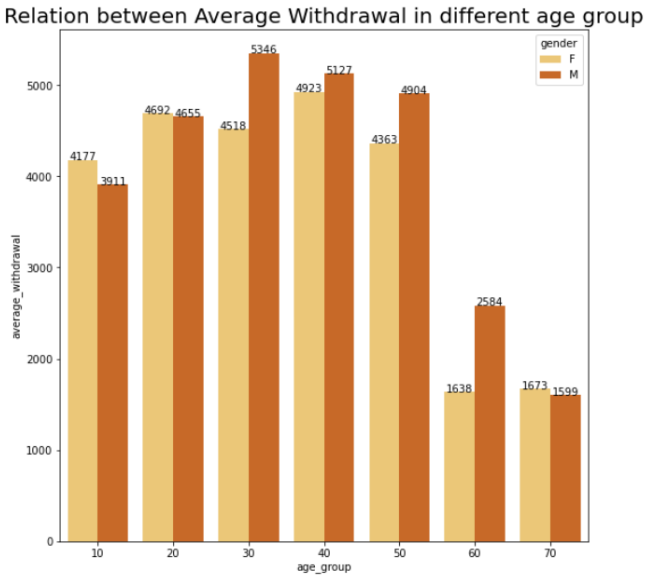
Relation between gender and being granted a loan



In order to show the number of males and females who had been granted a loan, We have used a counterplot from the seaborn library to represent the observation values using bars.

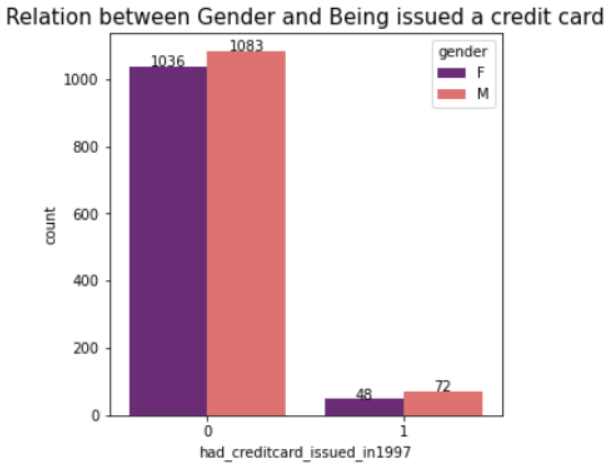
From the graph, we can notice that the bank had granted loans in 1997 only to 17 Female and 14 Male clients. From this, we can't assume that the Feature - Gender has a correlation with the Loan granted by the bank. Let us dive deep into other insights for the ' Gender' variable to find whether it actually has an impact on the target variables.

Relation between average withdrawal in different age group



From this graph, It is clear that age is one of the key features which has a direct correlation with the average withdrawal. Though there is a small difference in average withdrawal between male and female clients, we can notice that from age 20-50, both the clients have a high average withdrawal and drastically reduced after Age 50. This information will, in turn, help the bank institutions to grant loans and issue credit cards based on the client's behavior.

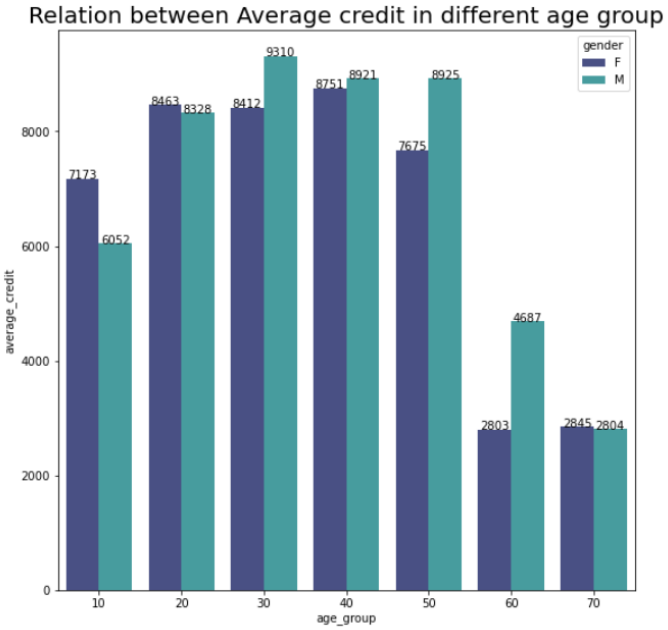
Relation between gender and being issued a credit card



In order to show the number of males and females who had been issued a credit card, We have used a counterplot from the seaborn library to represent the observation values using bars.

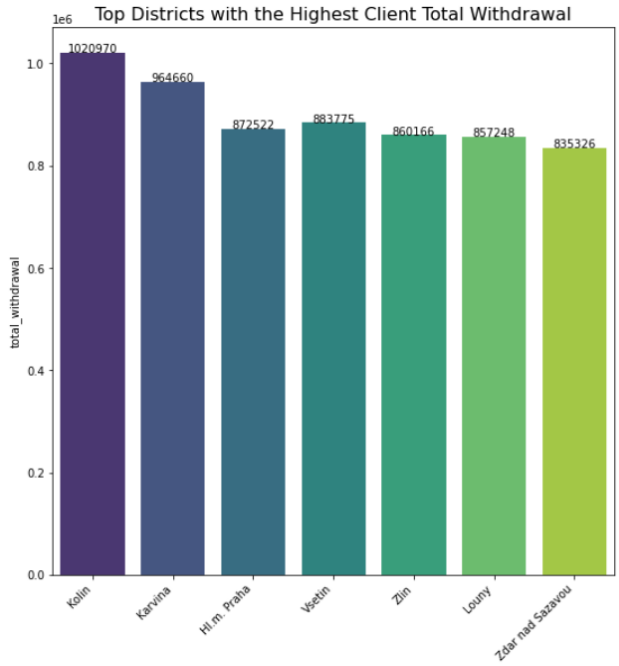
From the graph, we can notice that the bank had issued credit cards in 1997 only to 48 Female and 72 Male clients. From this, we can't assume that the Feature - Gender has a correlation with the credit card by the bank. Let us dive deep into other insights for the ' Gender' variable to find whether it actually has an impact on the target variables.

Relation between average credit in different age groups



This graph shows the relation between the average credit and Age group. This Graph's pattern is very similar to the graph of the relation between average withdrawal and Age group. It is pretty common that clients who earn more will spend more. So there is no doubt about age groups from 20-50 has a high average credit.
This client's credit information is one essential feature to determine the credit score, thereby helping the bank institutions to grant loans and issue credit cards.

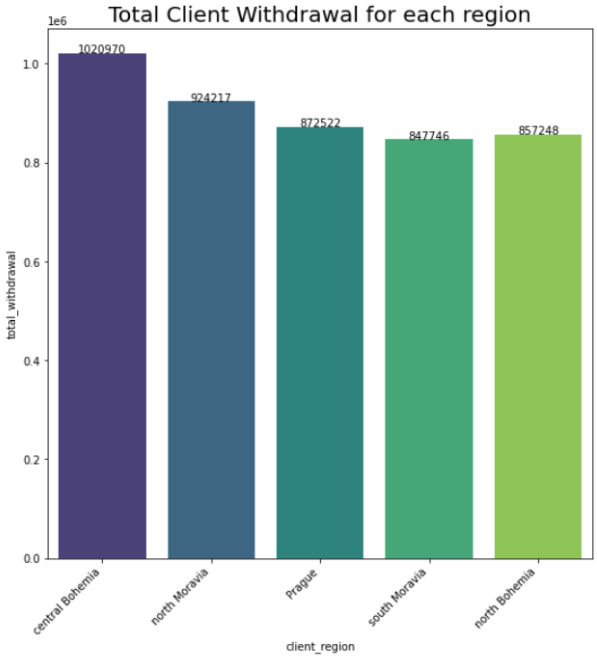
Top districts with the highest Total withdrawal



From this graph, we can observe the Top districts with the highest client total withdrawal. This district-wise total withdrawal information will allow the bank institutions to segment the client's withdrawal behavior in their respective districts. Kolin is the district with the highest total withdrawal and followed by the district Karvina.

The banks can use these details to prioritize the clients from the districts with the highest total withdrawal and provide them additional self-service opportunities to enhance their banking experience.

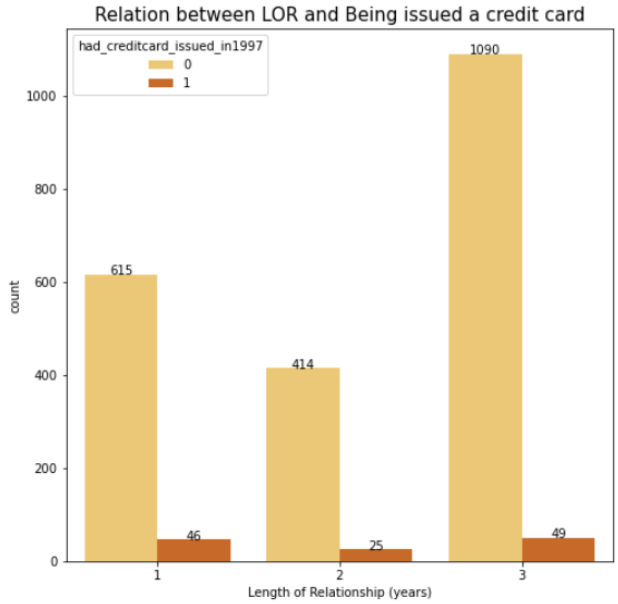
Total withdrawal for each region



From this graph, we can observe the Total client withdrawal for each region. This information will allow the bank institutions to segment the client's withdrawal behavior in their respective districts. Central Bohemia is the Region with the highest total withdrawal and followed by the Region North Moravia. Comparing this graph with the Region-Client Proportion pie chart, we can state that the region with more clients does not necessarily have the highest total withdrawal.

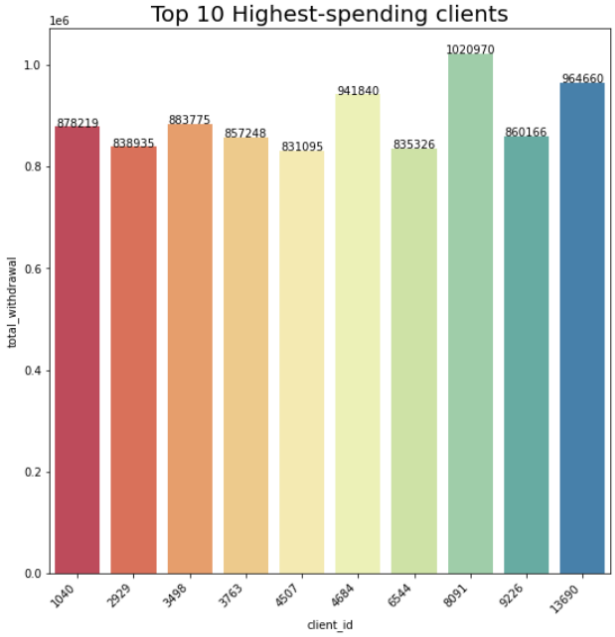
Banks can provide self-service opportunities through digital platforms and promote financial literacy through customer education to the prioritized clients to build a deeper relationship with customers.

Relation between length of relationship(lor) and being issued credit card



This graph illustrates the relationship between the LOR and credit card issued by the bank. We can observe that there is no direct correlation between LOR and credit cards issued by the bank. So, it is apparent that the client's credit and withdrawal behavior influences their credit score which in turn helps the bank to decide the issuance of a credit card.

Top 10 highest spending clients



This Bar graph shows the sample of top-spending clients. The bank institutions can segment them as value-added clients, observe their spending behavior and offer some additional services based on their interests. These instances will definitely enhance the customer-bank relationship beyond the transactional services and will help retain the customers.