Data mining project

PKDD'99 Financial dataset

DECEMBER, 2021

IESEG School of Management
MEDINA MARTINEZ Juan Jose
THAYANIDHI Kamalakannan
ZHAKSYBEK Dilda

Contents

Dataset overview

PKDD'99 is a financial dataset for a Czech bank. Dataset timeframe is from 1993 to 1998. Dataset includes several tables with information on accounts, clients, loans, orders, transactions, cards, districts, as well as the table to connect all the data together: dispositions.
This project is aimed at creating a basetable for the year 1996 and dependent variables for 1997, that can later be used for various data science applications, such descriptive and predictive analytics.

Data Correction and Transformation

Account table

- Created year, month, day integer variables by subsetting original date string.
- Translated frequency by mapping column's values to a function 'translate_frequency'.
- Created a 'dateInt' variable, which is a concatenated string of year, month and date values.
- Created a Date variable which is an account creation date in datetime format.
- Created a 'LOR' variable, which is a difference between Dec 31, 1996 and the creation date of the account.
- Renamed columns Date and Frequency.
- Dropped columns that will not be used in the final merged base table.
- Downloaded final cleaned table as a csv file.

Client table

- Created 'birth_year', 'birth_month' and 'birth_day' variables.
- Created 'gender' variable.
- Created 'age' variable.
- Created 'client_birth_date' variable.
- Dropped variables that are not used in the final merged base table.
- Downloaded final cleaned table as a csv file.

Loan table

- Created 'year', 'month', 'day' variables.
- Created 'loan_date' variable .
- Renamed columns .

- Dropped variables that are not used in the final merged base table.
- Downloaded final cleaned table as a csv file.

District table

- Renamed columns using '.rename' method.
- Replaced '?' values in columns 'UNEMP_95' and 'Crime_Comm_95' by 0.
- Replaced 0 in the above-mentioned columns by means of the respective columns.
- Downloaded final cleaned table as a csv file.

RFM table

- Subsetting transactions table for only those transactions with type = "Prijem", which stands for credit transactions.
- Creating a 'date' variable in a datetime format.
- Further subsetting transactions to only those that took place before and including 1996.
- Creating Recency variable as a difference between Dec 31, 1996 and latest (max) transaction date.
- Dropping duplicate values.
- Re-defining Recency as a minimum value per each account.
- Defining Monetary as a count of transactions per each account.
- Defining Frequency as a sum of transactions amount per each account.
- Mapping R, F and M values to each account ID in the RFM table.
- Defining scores for R, F and M values at 0.2, 0.4, 0,6, 0.8 quantiles.
- Calculating total RFM value as a total of R, F and M scores.

Orders table

- Replacing empty values with 'NONE' in 'k_symbol' column.
- Translating 'k_symbol' values by mapping to a pre-defined function.
- Downloaded final cleaned table as a csv file.
- Calculated the number of orders per account id, transformed into a dataframe.
- Dropping unnecessary columns.
- Calculating average amount sent to other banks per account.
- Pivoting the table to display banks as columns. Taking amount as values.
- Calculating total amount of orders sent to other banks per account.

Card table

- Created ‘card_issue_date’ variable.
- Dropped unnecessary columns.
- Downloaded final cleaned table as a csv file.

Transactions table

- Extracting ‘year’, ‘month’, ‘day’ variables from ‘date’ variables.
- Apply the translation function created previously to rename observations of ‘type’, ‘operations’ and ‘k_symbol’ variables.
- Replace miswritten entry in ‘type’ variable.
- Subsetting transactions table for 1996 as trans_1996.
- Subsetting trans_1996 into trans_1996_1 to only include credit type transactions and count the number of credit type transactions using groupby method.
- Subsetting transactions table for 1996 as trans_1996.
- Subsetting trans_1996 into trans_1996_2 to only include withdrawal type transactions and count the number of withdrawal type transactions using groupby method.
- Merge trans_1996_1 and trans_1996_2 with the original transactions table.
- Create a variables that create average credit and withdrawals per account id, merged to original table.
- Create variables with the maximum and minimum amount per credit and withdrawal transactions per account id and merge to the original table.
- Create the maximum and minimum balance per account id and merge to the original table.
- Group final merged table by account id, taking the mean of all variables created above.
- Merge RFM table with the final transactions table to create transactions_cleaned.
- Downloaded final cleaned table as a csv file.

Creating Dependent Variables

Loan_granted:

- Subsetting cleaned account table for those accounts that were created before and including 1996.
- Subsetting loans table for those loans that were taken in 1997.
- Creating a column in a subsetted loans table with a value 1 – loan granted.
- Dropping unnecessary columns in a subsetted loans table.
- Merging Account and Loan table. Resulting table named ‘account_dv’. Now for each account id we have a value or 0 or 1 depending on if the loan was granted for that account id in 1997.

Card_granted:

- Subsetting card table for those cards that were issued in 1997.
- Creating a column with a value of 1 – card granted.
- Dropping unnecessary columns.
- Merging the table with ‘account_dv’. Now for every account id we have a value of 1 or 0 depending on if the card was issued for the account or not.

Basetable

Overview

The final basetable comprises of selected variables, merged from several tables in the following order: dispositions, accounts, loans, orders, transactions, cards, district, client. Since only one client can be an account owner, all the tables were left merged by 'account_id', yielding a basetable with information for clients that are account owners.

Variables description

Below is the table containing information per each variable of the basetable.

VARIABLE	DATA TYPE	SOURCE TABLE	DESCRIPTION
ACCOUNT_ID	int64	account_cleaned	Unique account identifier
FREQUENCY	object	account_cleaned	Frequency of statements issue
ACCOUNT_YEAR	int32	account_cleaned	Year of account creation
LOR	int64	account_cleaned	Length of relationship with the client as a difference between current date and account creation date
LOAN_AMOUNT	float64	loan_cleaned	Total loan amount
LOAN_DURATION	float64	loan_cleaned	Duration of the loan (months)
LOAN_PAYMENTS	float64	loan_cleaned	Monthly payments specified by loan amount divided by loan duration
LOAN_STATUS	object	loan_cleaned	Loan status: where 'A' stands for contract finished, no problems; 'B' stands for contract finished, loan not paid; 'C' stands for running contract, OK so far; 'D' stands for running contract, client in debt.
AMOUNT	float64	order_cleaned	Number of orders per account
AB	float64	order_cleaned	Amount sent to AB bank
CD	float64	order_cleaned	Amount sent to CD bank
EF	float64	order_cleaned	Amount sent to EF bank
GH	float64	order_cleaned	Amount sent to GH bank
IJ	float64	order_cleaned	Amount sent to IJ bank
KL	float64	order_cleaned	Amount sent to KL bank
MN	float64	order_cleaned	Amount sent to MN bank
OP	float64	order_cleaned	Amount sent to OP bank
QR	float64	order_cleaned	Amount sent to QR bank

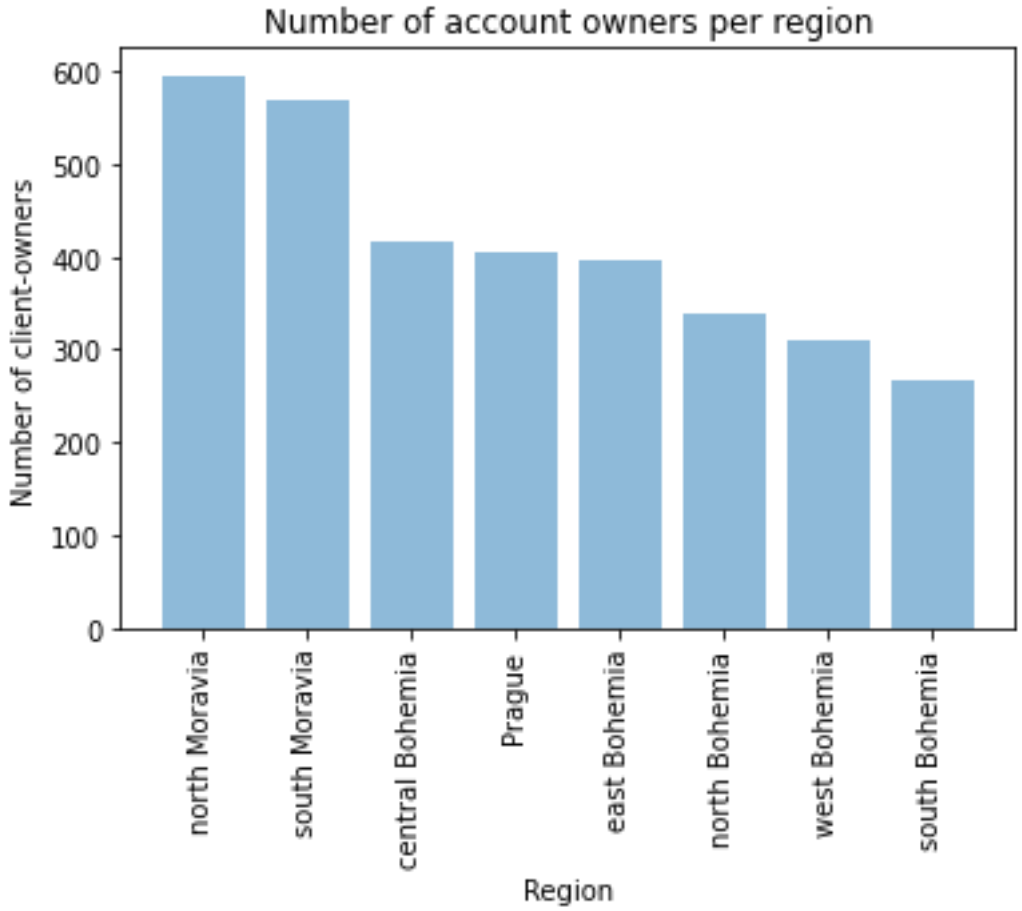
ST	float64	order_cleaned	Amount sent to ST bank
UV	float64	order_cleaned	Amount sent to UV bank
WX	float64	order_cleaned	Amount sent to WX bank
YZ	float64	order_cleaned	Amount sent to YZ bank
SUM_OF_ORDERS	float64	order_cleaned	Sum of orders sent to other banks per account
CREDIT_CNT	float64	trans_cleaned	Number of credit transactions
CREDIT_AVG	float64	trans_cleaned	Average amount of credit transactions per account
WITHDRAWAL_CNT	float64	trans_cleaned	Number of withdrawal transactions
WITHDRAWAL_AVG	float64	trans_cleaned	Average among of withdrawal transactions per account
AMOUNTMAX_CRED	float64	trans_cleaned	Max amount of credit transactions per account
AMOUNTMIN_CRED	float64	trans_cleaned	Min amount of credit transactions per account
AMOUNTMAX_WITH	float64	trans_cleaned	Max amount of withdrawal transactions per account
AMOUNTMIN_WITH	float64	trans_cleaned	Min amount of withdrawal transactions per account
BALANCEMAX	float64	trans_cleaned	Max balance per account
BALANCEMIN	float64	trans_cleaned	Min balance per account
R	float64	RFM table	Days since the most recent to Dec 31, 1996 transaction
F	float64	RFM table	Number of transactions before (including) 1996
M	float64	RFM table	Total amount of transactions before (including) 1996
RS	float64	RFM table	Recency score based on quantiles
FS	float64	RFM table	Frequency score based on quantiles
MS	float64	RFM table	Monetary score based on quantiles
RFM	float64	RFM table	Total RFM score as a sum of R,F,M scores. The higher the score the better.
DISP_TYPE	object	disp_owner	Subset from disposition table. Only those clients that are account owners
CARD_TYPE	object	card_cleaned	Type of card issued
CARD_YEAR	float64	card_cleaned	Year card was issued
REGION	object	district_cleaned	Region related to account
POPULATION	int64	district_cleaned	Population of the account's region
CITIES	int64	district_cleaned	Number of cities in the account's region
RATIO_URBAN	float64	district_cleaned	Ratio of urban inhabitants in the account's region
AVG_SLRY	int64	district_cleaned	Average salary in the account's region

UNEMP_96	float64	district_cleaned	Unemployment rate in the region for year 1996
ENP_PER_1000	int64	district_cleaned	Number of entrepreneurs per 1000 inhabitants
CRIME_COMM_96	int64	district_cleaned	Number of committed crimes in 1996
GENDER	object	client_cleaned	Account owner's gender
AGE	int32	client_cleaned	Account owner's age
AGE_GROUP	int32	client_cleaned	Account owner's age group
LOAN_GRANTED	float64	client_cleaned	Loan was granted to the client in 1997 (1 – yes, 0 – no)
CARD_GRANTED	float64	client_cleaned	Card was granted to the client in 1997 (1 – yes, 0 – no)

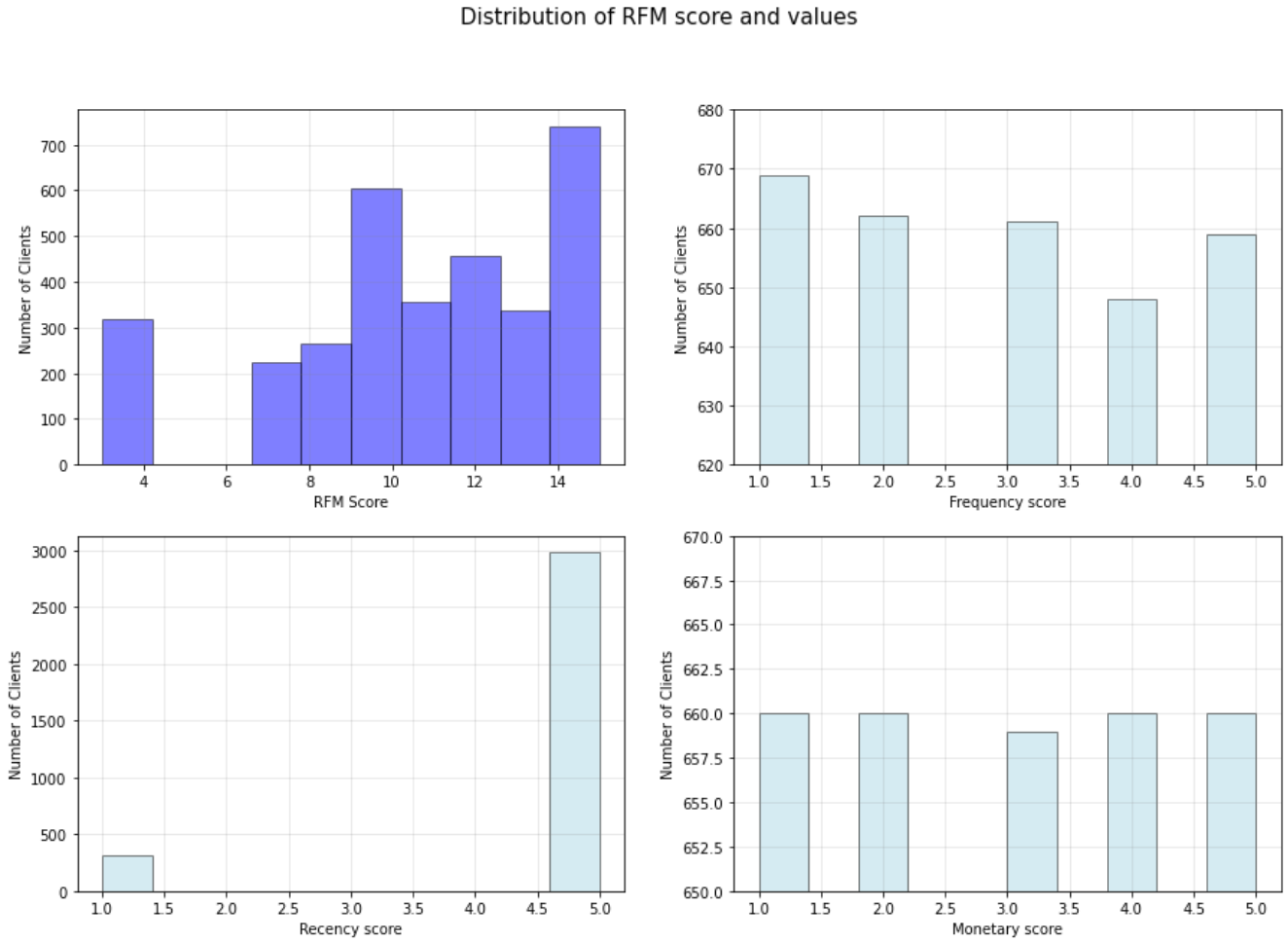
Data Analysis and Visualization

In this section we analyze the basetable using visualization tools. It should be noted that we assume all client's applied for loan or card, and those clients that were granted a loan or card - were approved, all the rest that were not granted a loan or a card are considered as rejected applications.

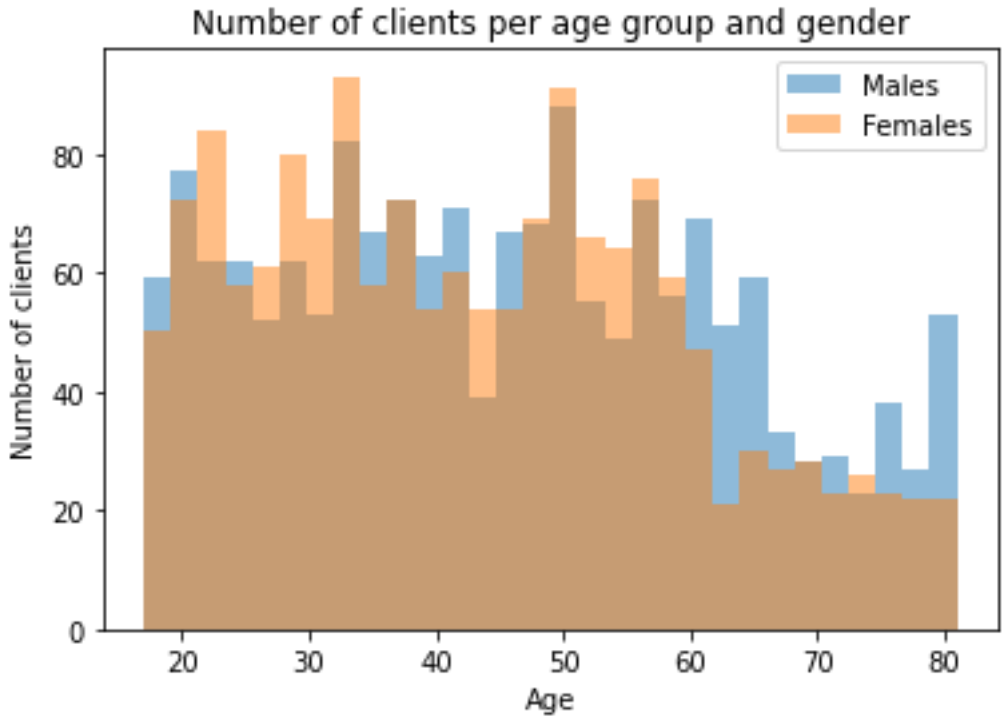
For the distribution of clients who are account owners, the leading region is north Moravia, with under 600 client-owners under this region, followed by south Moravia and Central Bohemia.



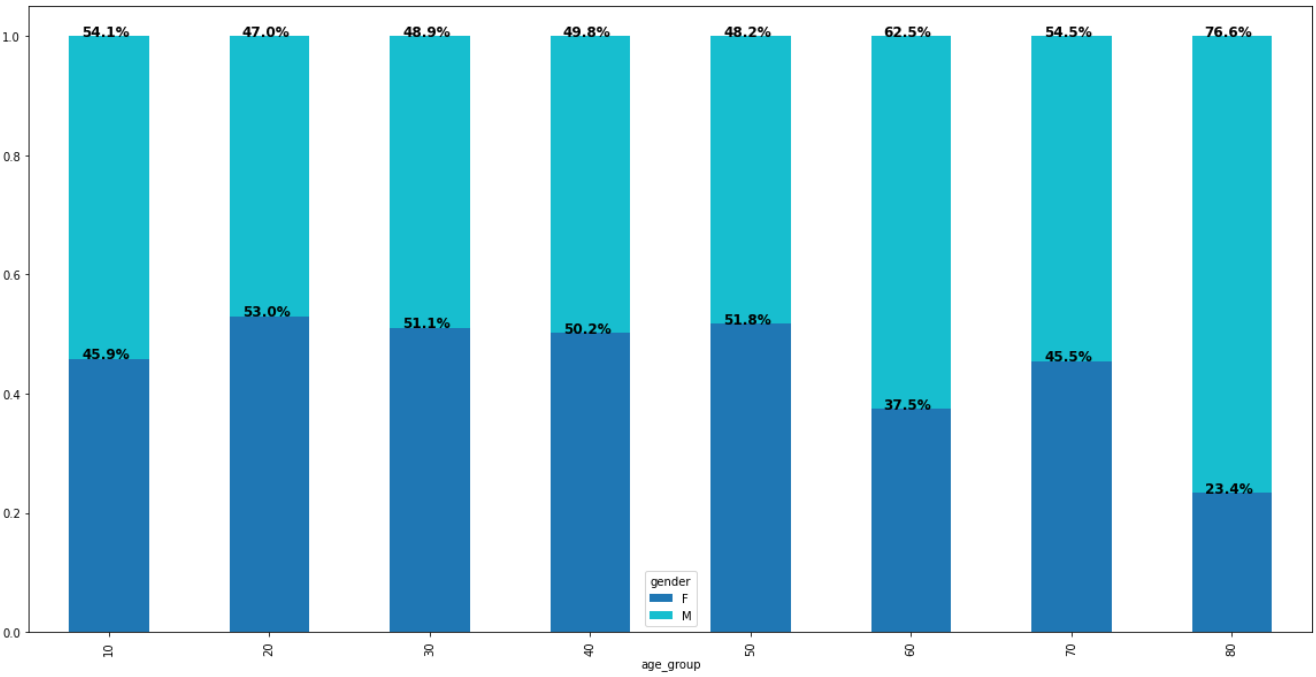
RFM score is a sum of Recency, Frequency and Monetary scores combined. Analyzing distribution of RFM scores, we can observe than most of our clients have an RFM score of 9 and above.
We can imply that the majority of the weight towards a higher RFM score is coming from a recency score, where only under 500 clients have a score of 1, while the rest of the clients have a score of 4 and above. For Frequency and Monetary scores, the distribution is even.



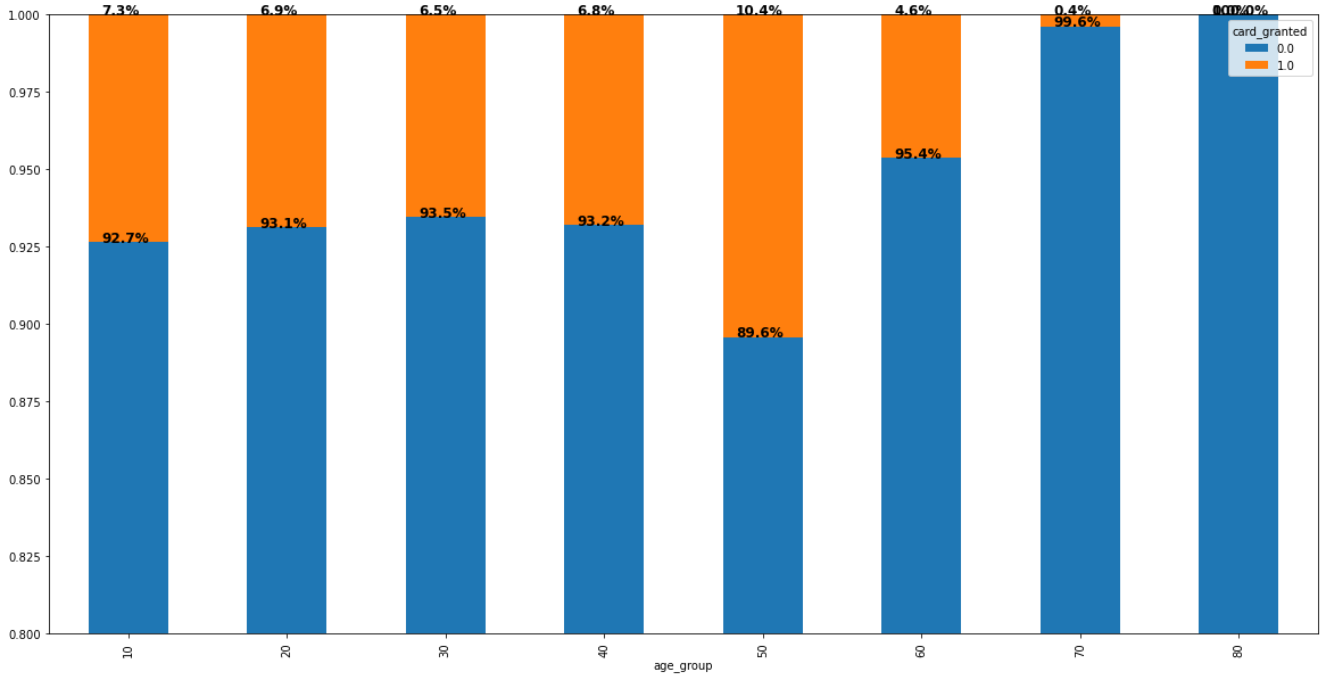
Below, we plot a histogram with a distribution of clients by age group and gender. We can observe than majority of our clients are under 60 years of age. Under 60, the number of male and female clients are approximately the same.



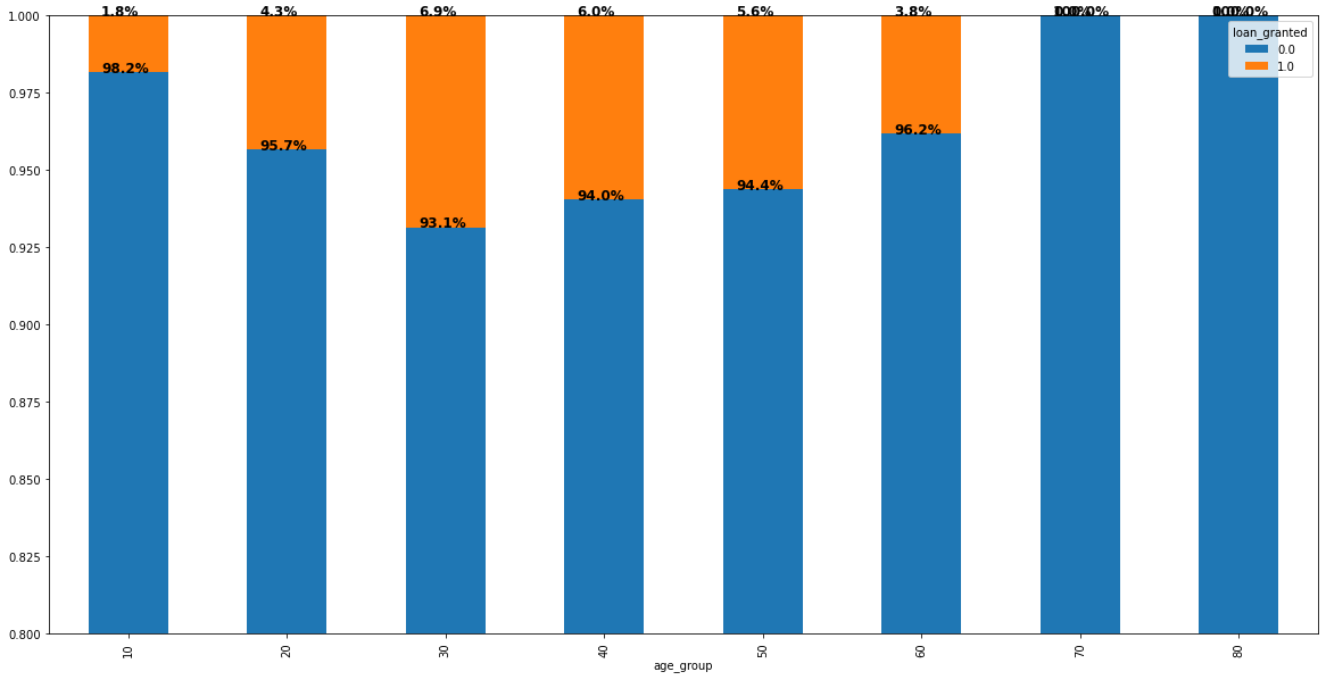
In a graph below we can see that for age groups from 20 to 50 division between male and female is equal, for the age group 60 and above the proportion of male clients increases greatly.



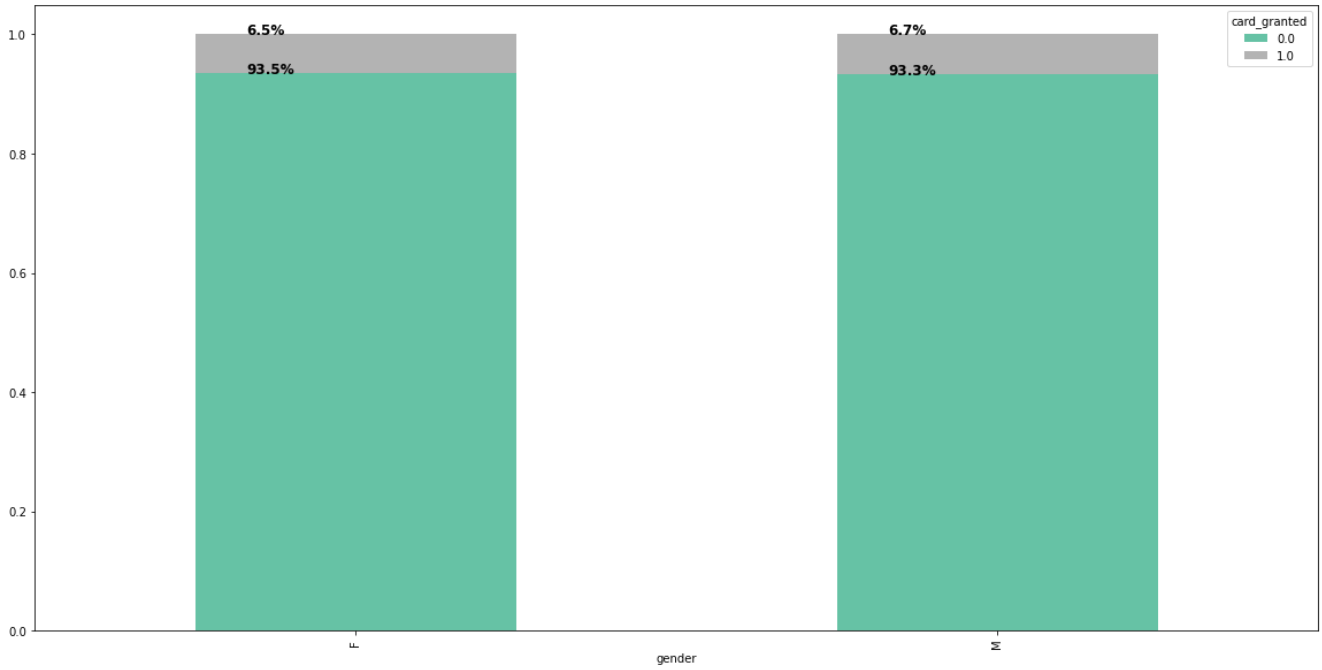
Comparing age groups, 50 year-olds as a group are most likely to be granted a card in 1997. Meanwhile, likelihood of age groups 70 and 80 being granted a card is almost none. It is unclear, however, whether 70 and 80 year-olds are interested in owning one. So we make an assumption that all people applied for a card.

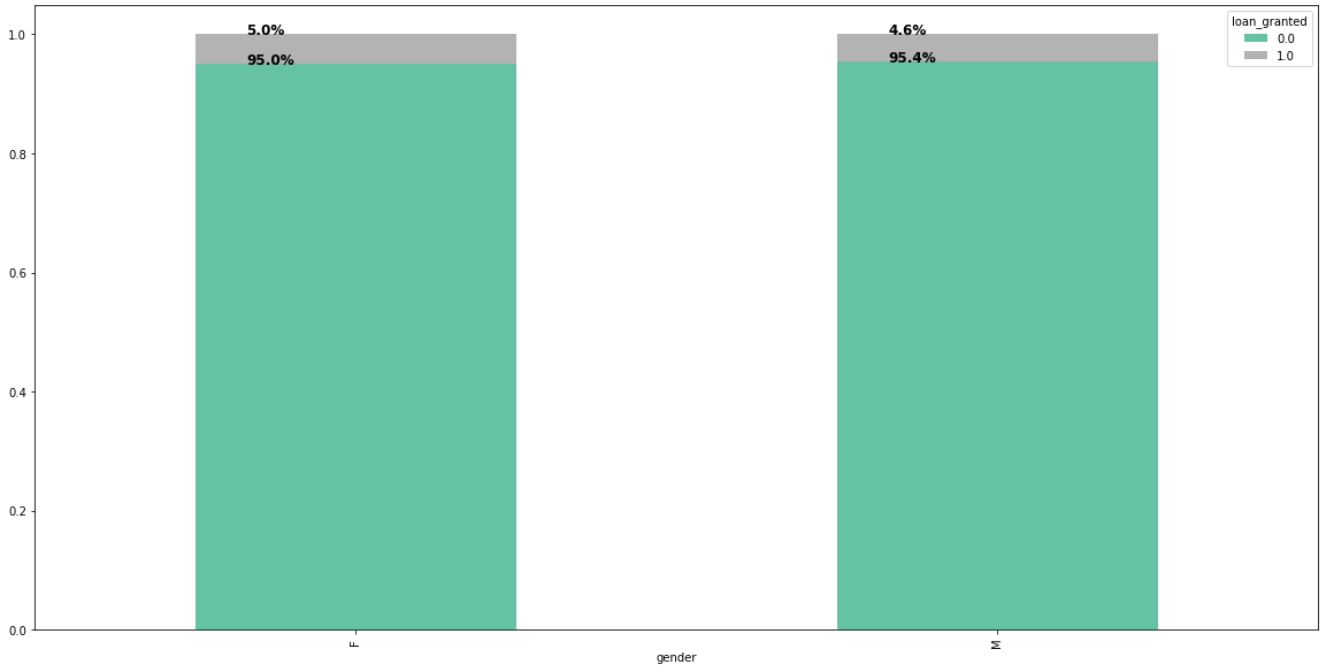


For the loan granted, age groups of 20 to 60 are more likely to be granted a loan in 1997, while age groups of 70 and 80 have almost no chance of getting a loan. Once again, however, we do not whether a client applied for one. Here we assume that all clients apply for a loan.



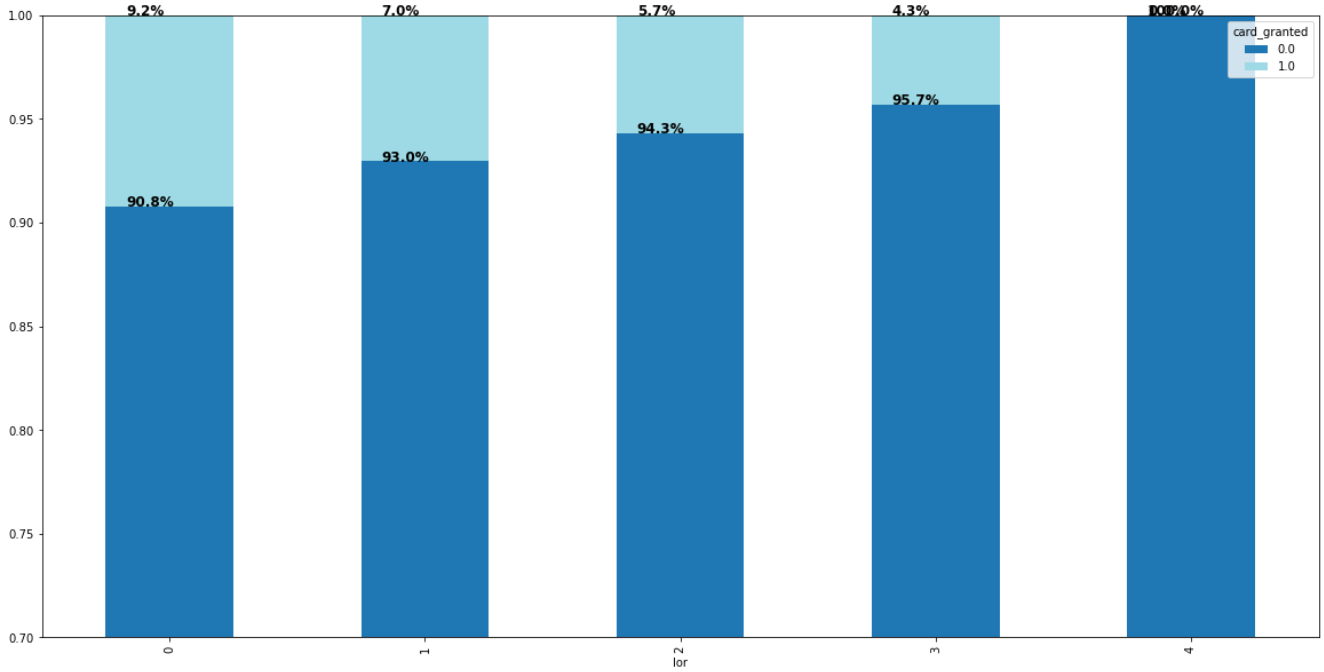
We can see that there is no gender bias when it comes to granting a loan or a card, as both genders are equally likely to be granted either.

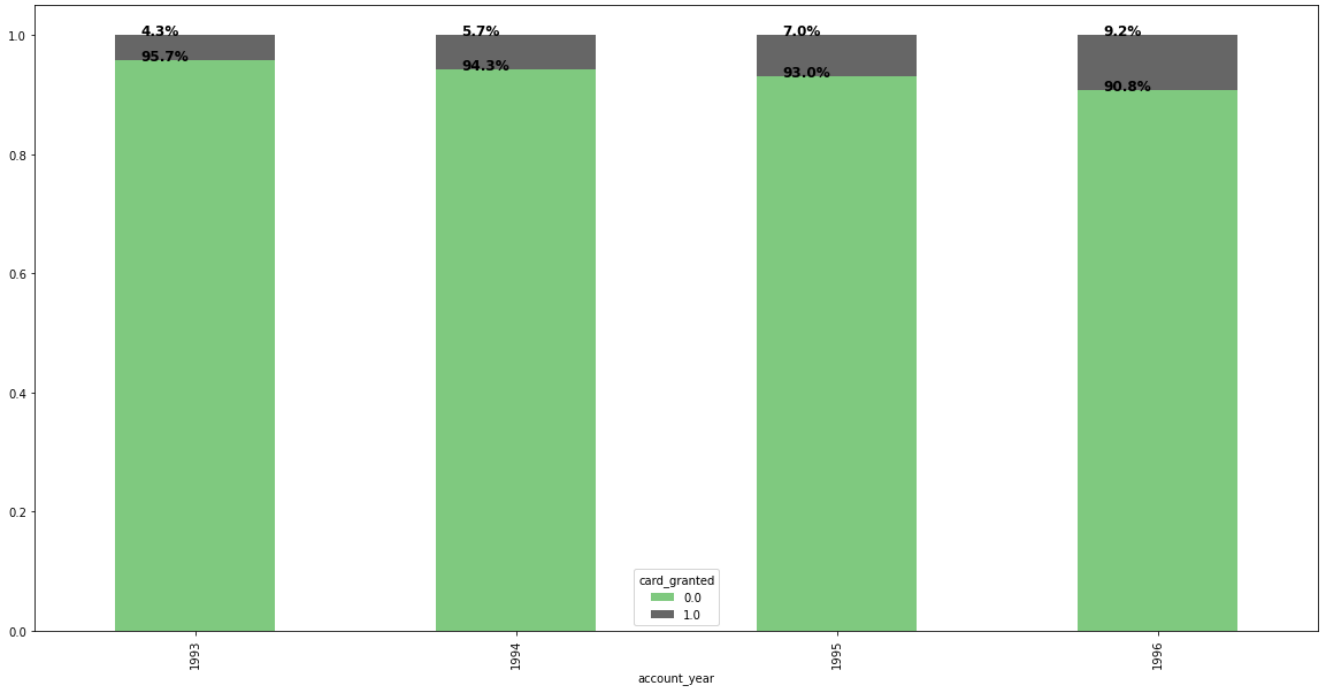




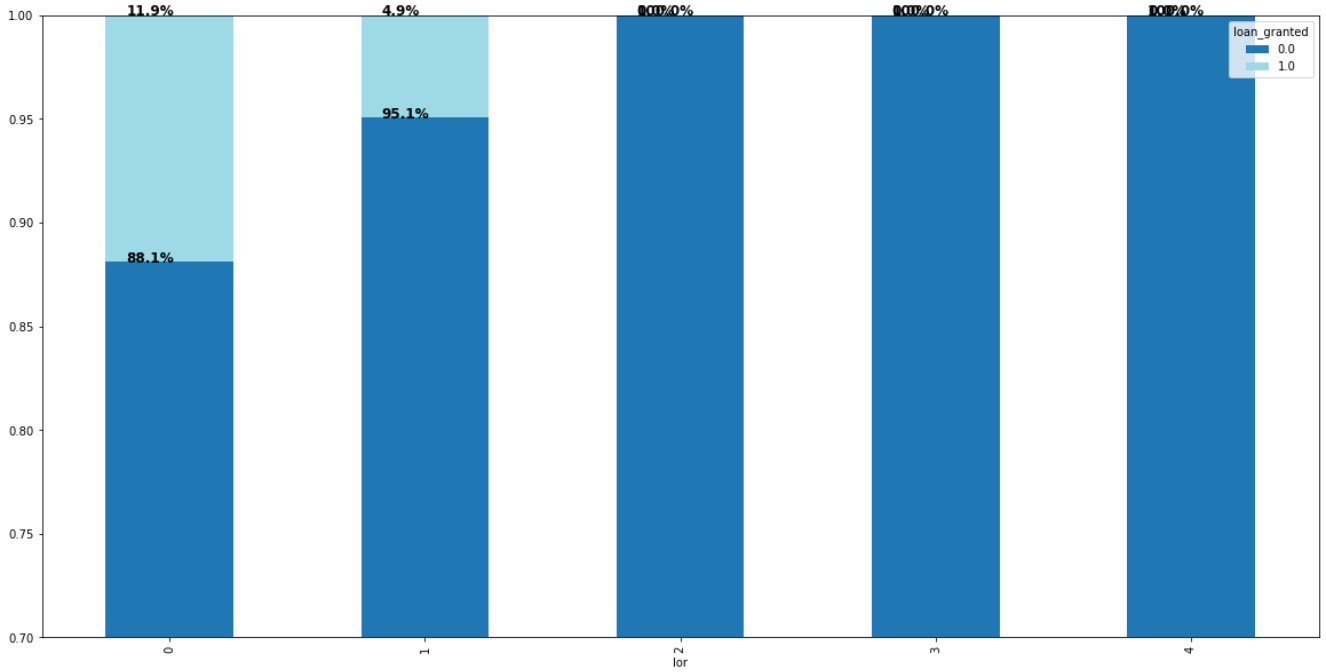
Analyzing LOR relationship with cards granted, the longer the client is with the bank, the less cards are granted in 1997. From this we can infer that almost all clients are granted their cards during the first 3 years of relationships with the bank. And the longer the client stays with the bank, the less are the number of cards granted in 1997. This finding is supported in analyzing account creation date to a dependable variable. For clients who created their accounts earlier, the percentage of cards granted in 1997 was lower, as those client's should have already got their cards. While for those clients who

created their accounts later, the percentage of the cards granted in 1997 is higher, as their LOR with the bank is shorter.

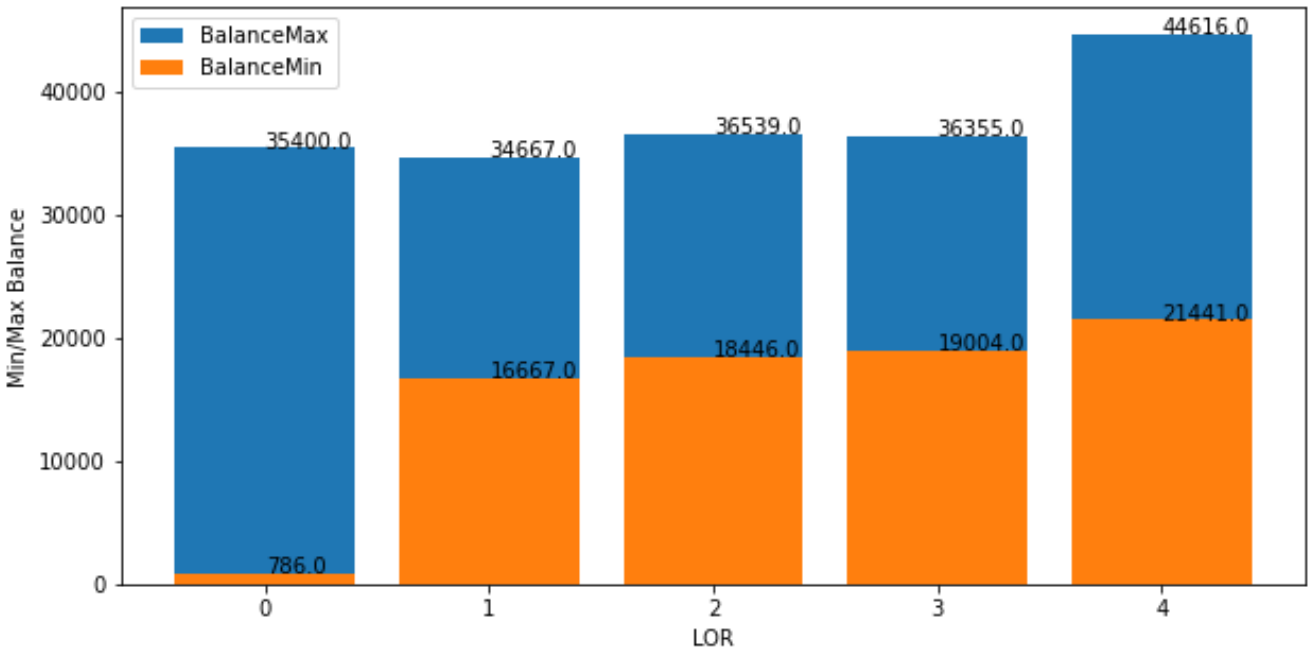




Analyzing LOR to loans granted, we can infer the similar interpretation as with the cards granted. Most of the clients are granted loans in 1997 in the first year of relationship with the client. Or some clients become bank's clients exclusively to apply for a loan, and they do so immediately after they become a client of the bank.



In the graph below we can see that clients with longer LOR with the bank have a larger minimum and maximum balance than those who opened the account recently. we are able to se that those clients that have a LOR of less than one year the gap from the maximum balance and the minimum balance is greater than those who have been with the bank for more than 1 year.


BalanceMax
BalanceMin
Min/Max Balance
LOR
0
10000
20000
30000
40000
0
1
2
3
4
35400.0
34667.0
36539.0
36355.0
44616.0
786.0
16667.0
18446.0
19004.0
21441.0