Datamart Report - PKDD'99 Discovery Challenge

December 18, 2019

Deborah Kewon, Ekapope Viriyakovithya, Sabir Akhtar

Intro

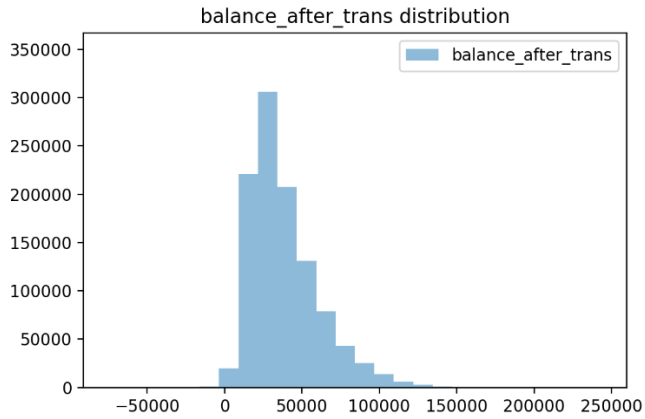
The purpose of this report is to analyze the financial status of clients in Czech Republic between January 1st ,1993 and December 31st, 1998. In order to create data mart, we used the following datasets 1. Card 2. Client 3. Account 4. Trans 5. Orders 6. Loans 7. Demographics 8. Disp. The total observations and variables of this data mart are 5,369 and 81 respectively.

I. Transaction table

Transaction table contains 4,500 account ids and 1,056,320 transactions.

1) Distribution of account balance after transaction

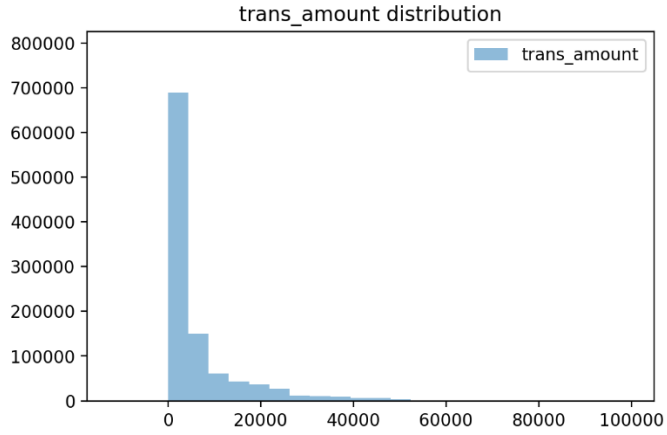
The distribution is right-skewed. Average balance after transaction is 38k CZK. Median balance is 33k CZK and the mode is 25k CZK.




```
count   1056320.0
mean      38518.3
std       22117.9
min      -41125.7
25%       22402.5
50%       33143.4
75%       49603.6
max      209637.0
Name: balance_after_trans, dtype: float64
```

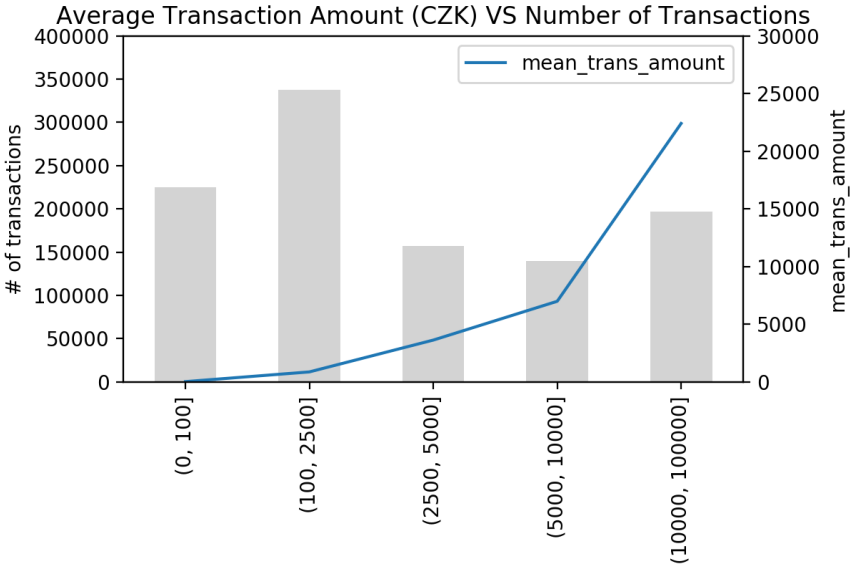
2) Distribution of transaction

The histogram of all transactions and summary statistics are shown below. The distribution is right-skewed, with average (whole timeframe, year 1993-1998) of 5.9k CZK. About 50% of total transactions were less than 2.1k CZK.




```
count   1056320.0
mean       5924.1
std        9522.7
min           0.0
25%         135.9
50%        2100.0
75%        6800.0
max       87400.0
Name: trans_amount, dtype: float64
```

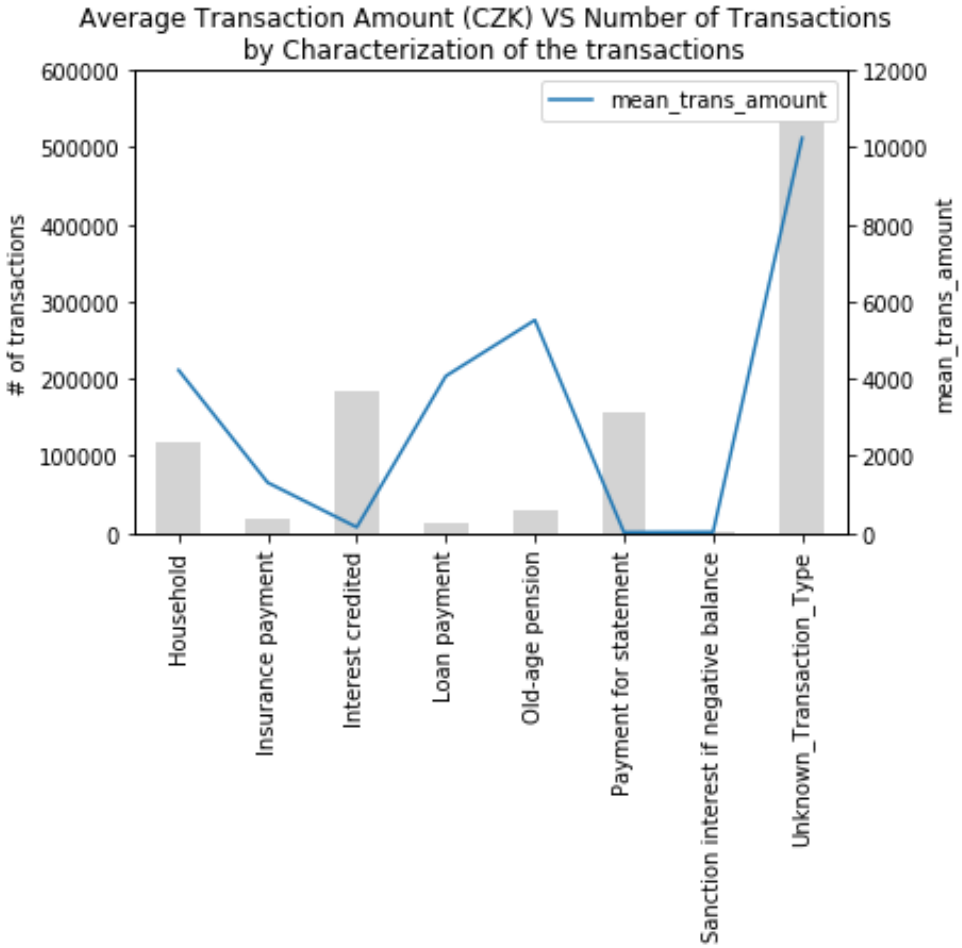
Based on the rough summary above, we performed discretizing into 5 groups. It shows that transaction amount between 100-2500 CZK occurred most frequently in the dataset.



Transaction amount (range)	Average amount	# of transactions
(0, 100]	32	225,242
(100, 2500]	873	337,789
(2500, 5000]	3,631	156,765
(5000, 10000]	6,992	139,705
(10000, 100000]	22,406	196,805

3) Distribution of average transaction by characterization of transaction

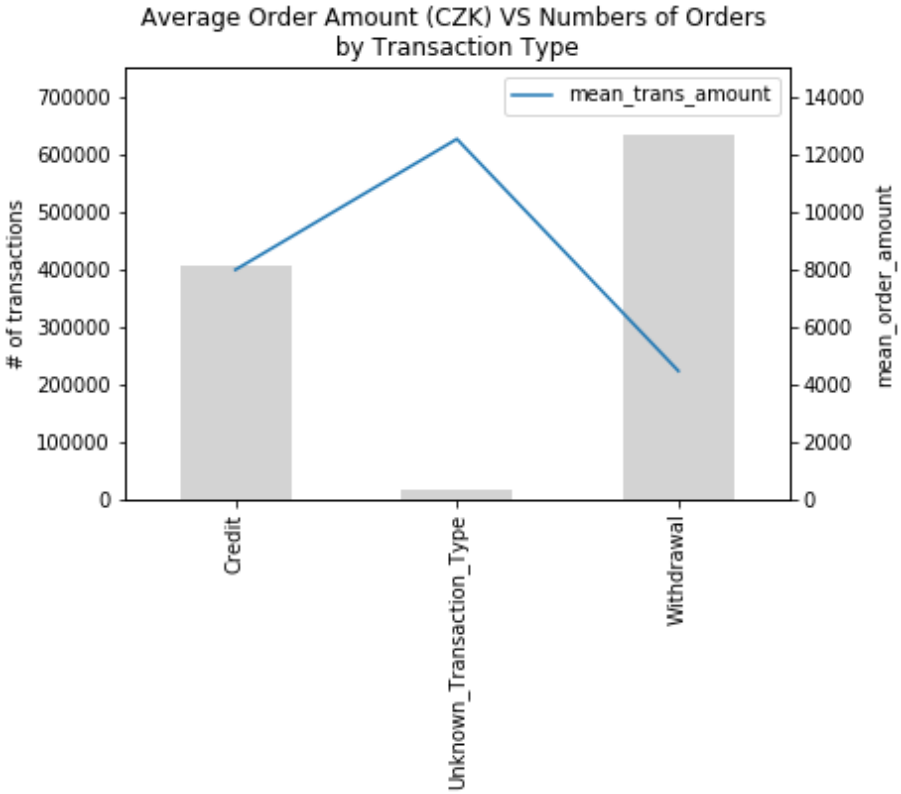
Most transactions are falling under 'Unknown type' group. This group has the highest average transaction amount. It would be more useful for the bank to identify this to leverage their data usage.



Characterize of Transaction	Average amount	# of transactions
Household	4,222	118,065
Insurance payment	1,307	18,500
Interest credited	150	183,114
Loan payment	4,069	13,580
Old-age pension	5,520	30,338
Payment for statement	17	155,832
Sanction interest if negative balance	24	1,577
Unknown Transaction Type	10,241	535,314

4) Distribution of average transaction by transaction type

More than 60% of transactions are 'Withdrawal', with the average amount of 4.5k CZK. 'Unknown Transaction Type' does not have a high number of transactions. However, this group has the highest average transaction amount once more. We would suggest the bank to have a closer look and identify this transaction type.



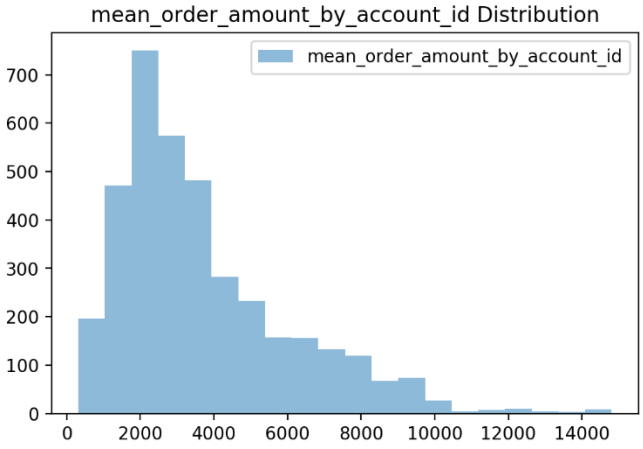
Transaction type	Average amount	# of transactions
Credit	7,967	405,083
Unknown Transaction Type	12,517	16,666
Withdrawal	4,447	634,571

II. Permanent Order table

There are 3,758 unique account ids in the order table with total of 6,471 records.

1) Distribution of order amount by account id

The distribution is right-skewed. Average amount per order for each account id is about 3.7k with the median of 3.1k CZK and mode of 2.5k CZK.



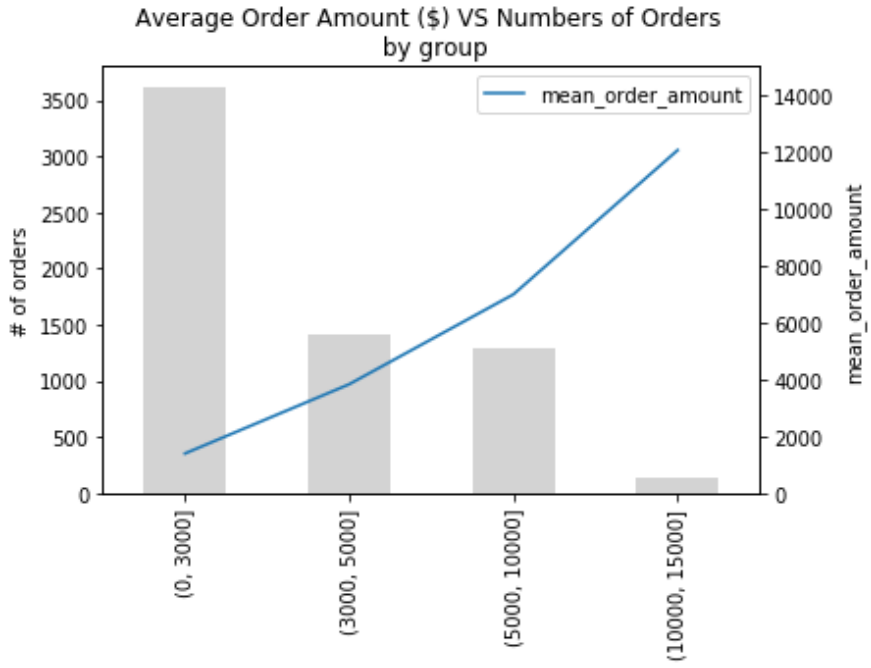

```
count     3758.0
mean      3710.6
std       2343.2
min        312.0
25%       2037.8
50%       3085.5
75%       4839.9
max      14811.0
Name: mean_order_amount_by_account_id, dtype: float64
```

2) Distribution of overall order amount

The distribution is right-skewed. Average amount per order is about 3.2k CZK with the median of 2.6k CZK. It is also interesting to note that there are some orders, which have the order amount as low as 1 CZK in the database.


```
count     6471.0
mean      3280.6
std       2714.5
min          1.0
25%       1241.5
50%       2596.0
75%       4613.5
max      14882.0
Name: order_amount, dtype: float64
```

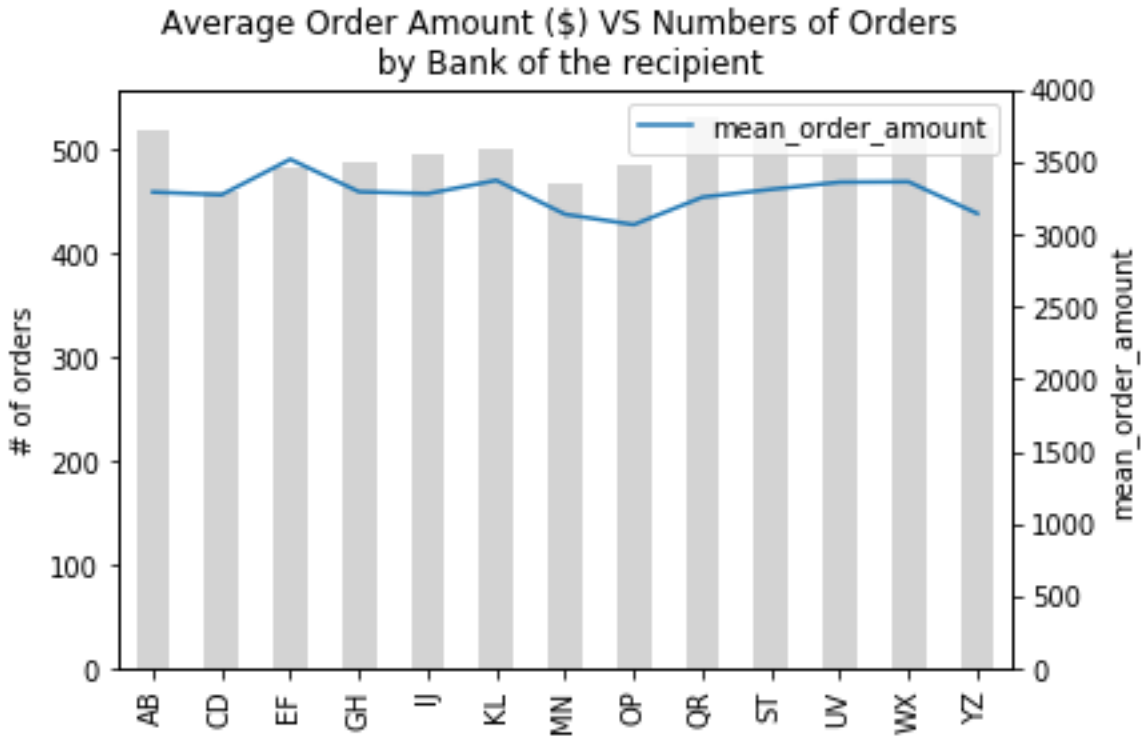
Based on the rough summary above, we performed discretizing into 4 groups. It shows that most transaction amount below 3k CZK has the most occurrence in the dataset.



Order amount (range)	Average amount	# of Orders
(0, 3000]	1,398	3,618
(3000, 5000]	3,833	1,416
(5000, 10000]	6,992	1,300
(10000, 15000]	12,062	137

3) # Order amount and # Orders for each Recipient Bank

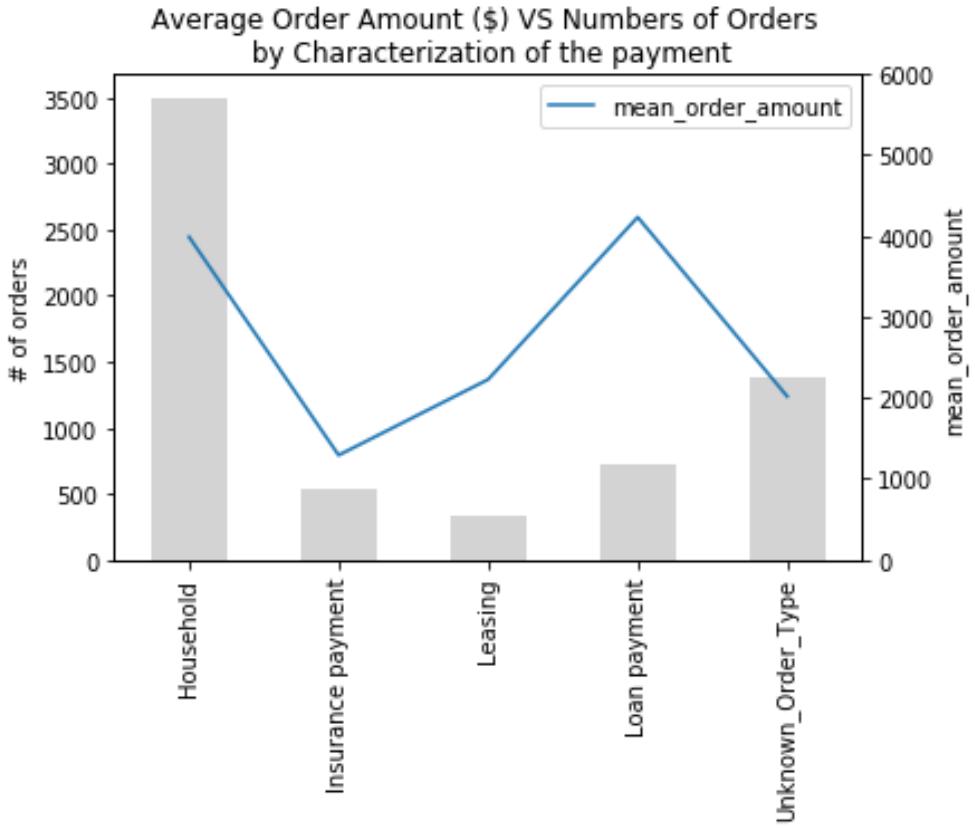
Average amount per order is about 3k CZK. All banks are equally distributed for both number of orders and the average amount.



Order Recipient Bank	Average amount	# of Orders
AB	3,290	519
CD	3,271	458
EF	3,516	483
GH	3,292	487
IJ	3,279	496
KL	3,371	500
MN	3,136	466
OP	3,065	485
QR	3,255	531
ST	3,309	511
UV	3,358	499
WX	3,361	515
YZ	3,142	521

4) Distribution of average order amount by characterization of payment

Most transactions are falling under 'Unknown type' group. This group has the highest average transaction amount. It would be more useful for the bank to identify this to leverage their data usage.



Characterize of Order	Average amount	# of Orders
Household	3,988	3,502
Insurance payment	1,291	532
Leasing	2,227	341
Loan payment	4,233	717
Unknown Order Type	2,017	1,379

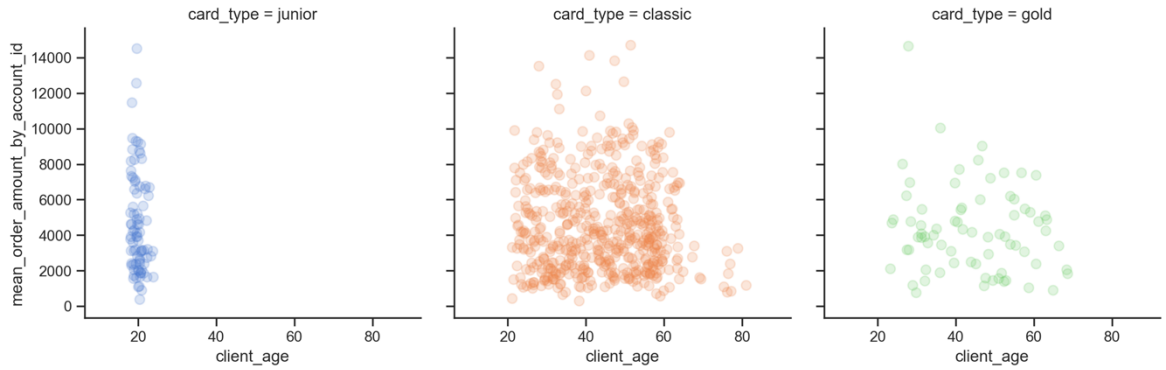
III. Datamart (basetable) Analysis

'trans_ratio_98_97' column derives from the ratio of transaction growth in the past year.

The client data was exported into two csv files; bottom10percent_customer.csv and top10percent_customer.csv.

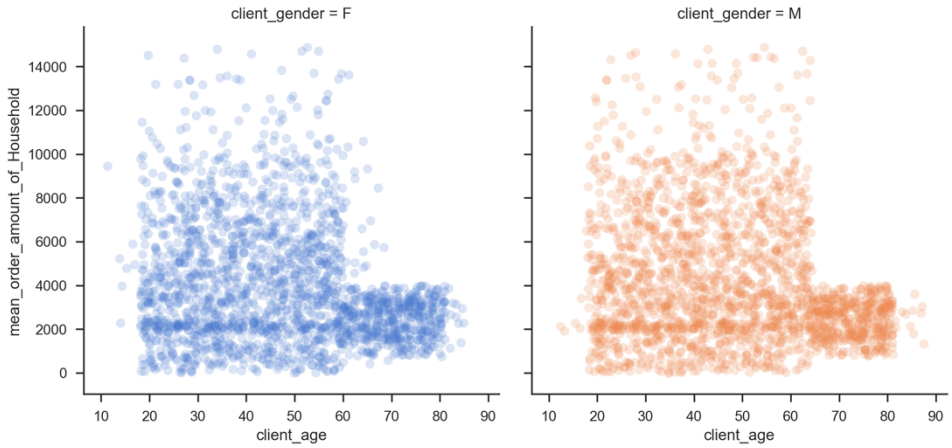
A) Orders

1) Order amount VS age VS card type



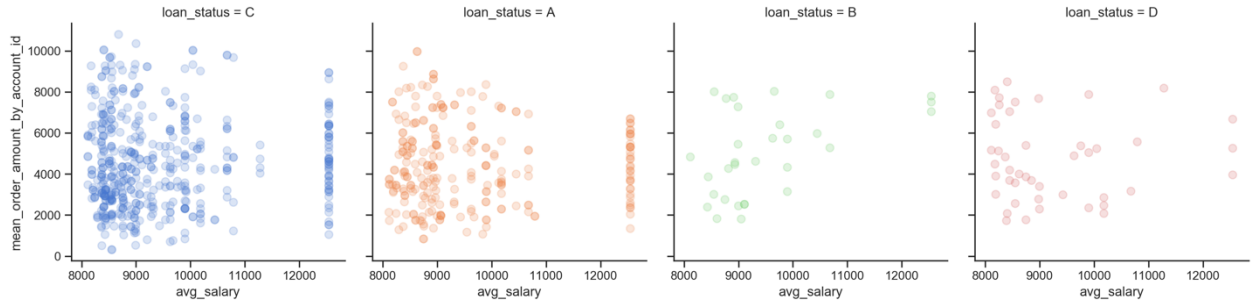
Most clients in their 20's own the card type 'junior'. The average order amount is between 2000-8000 CZK. The elders (age between 60-80) own 'classic' card type only, no other types presented in the data set.

2) Order amount VS age VS Gender



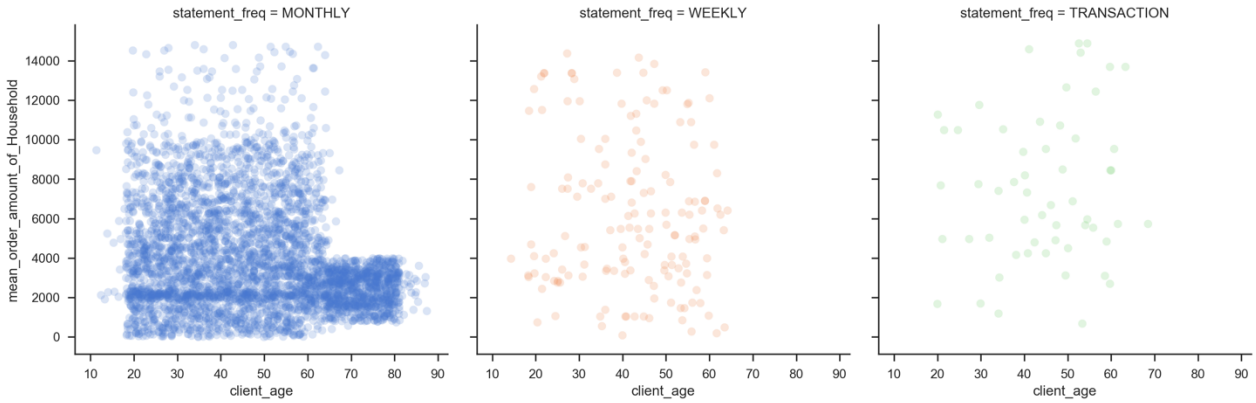
The distribution between two genders is quite similar when plotting the average order amount and age.

3) Order amount VS age VS loan status



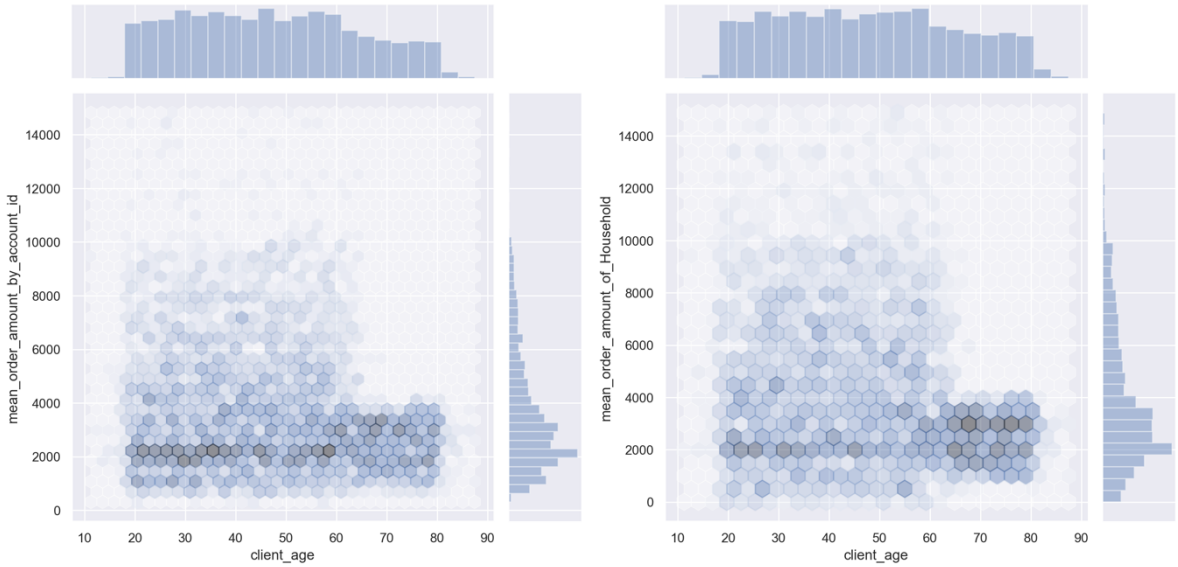
The average salary does not really determine the order amount and loan status. The maximum average order amount for 'B' and 'D' groups is 8k CZK when that of 'C' and 'A' groups is 10k CZK.

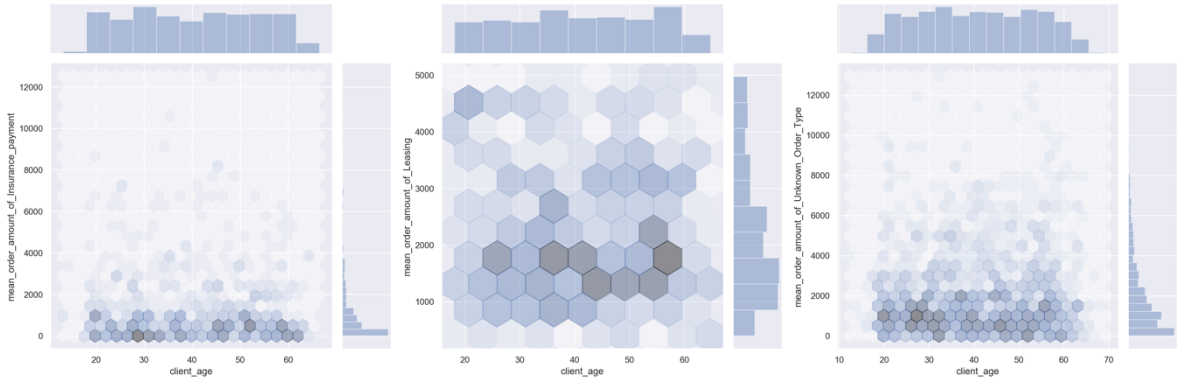
4) Order amount VS age VS statement frequency



In terms of issuance frequency of statements, elder customers (age above 60) prefer monthly statements. Elder customers do not appear in the 'weekly' and 'transaction' groups.

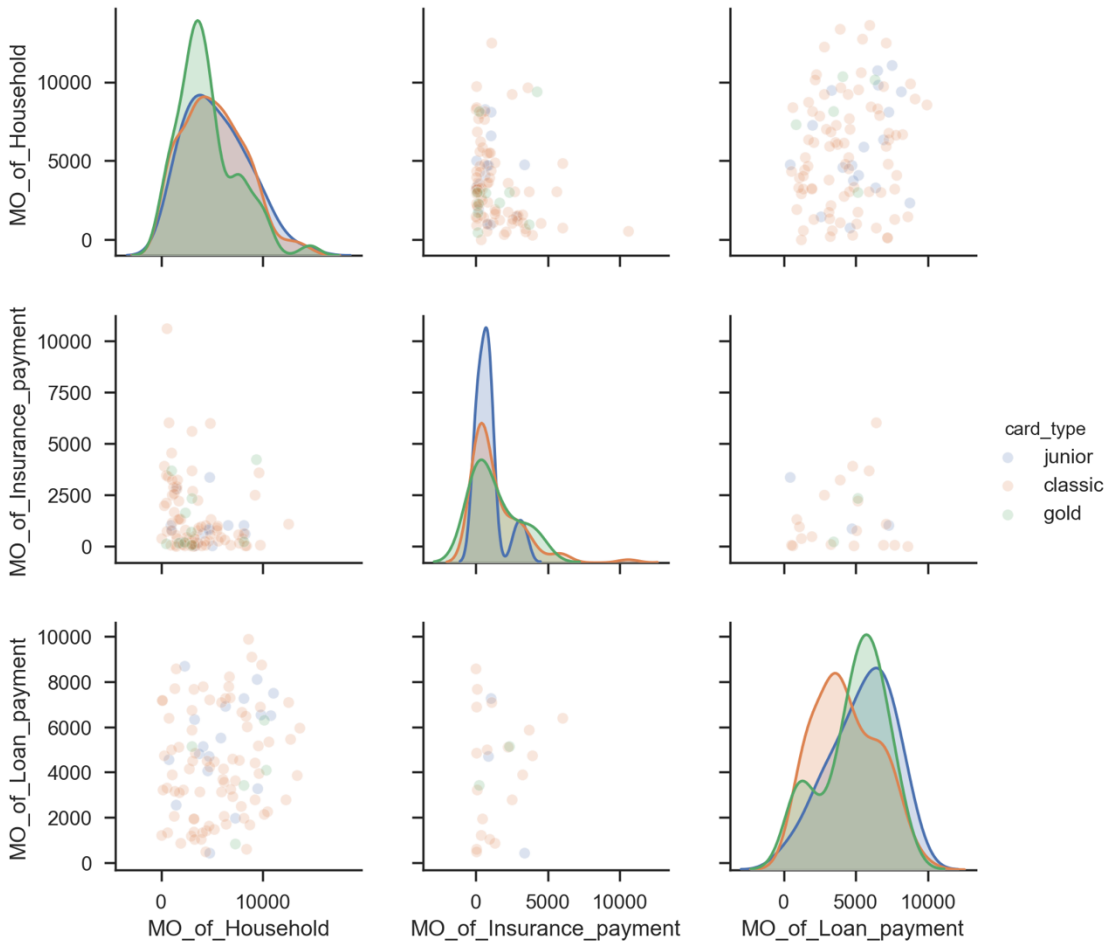
5) Average order amount distribution VS age distribution per account id - in each order category





Most average order amount is between 2000-4000 CZK. Elder customers (age above 60) do not show up in the average order above 4000 CZK. The elder group presents only in 'Household' category. No elder presents in 'Insurance Payment', 'Leasing', 'Loan Payment' or 'Unknown' categories.

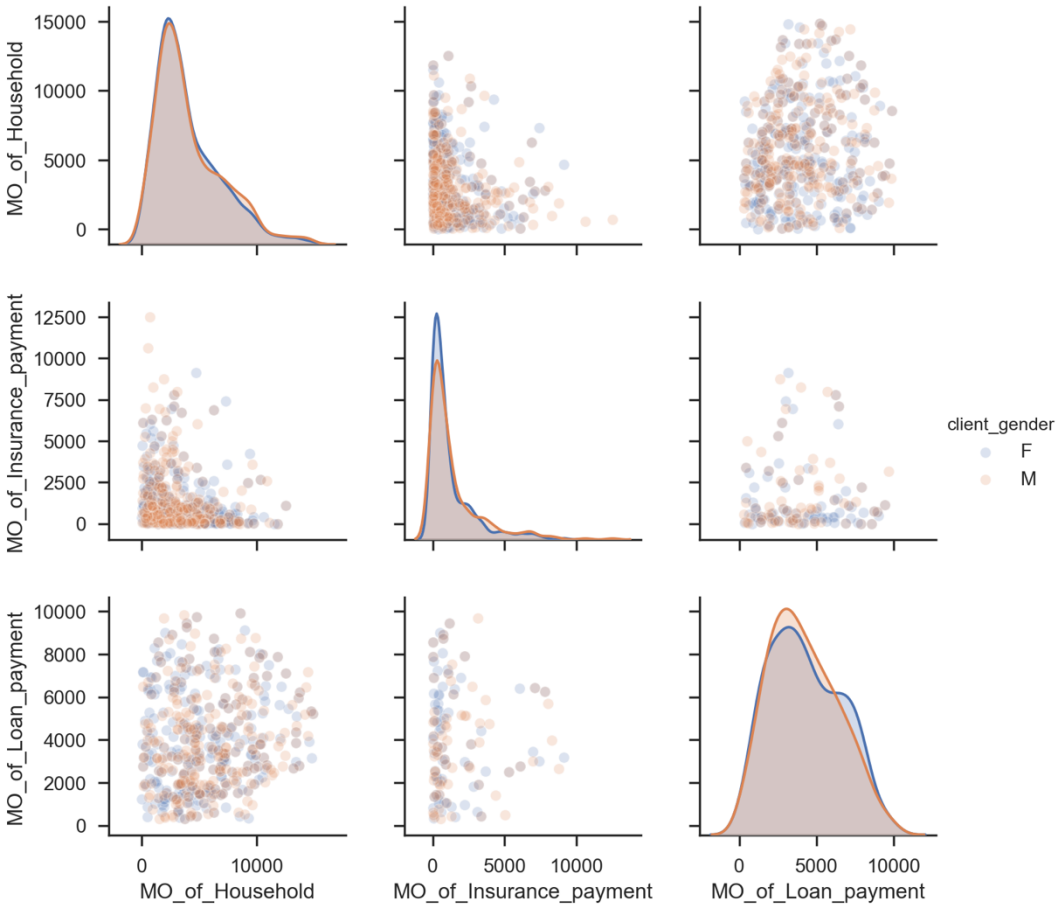
6) Card type distribution in each order category - plot with average amount of order



Since the sample size for leasing is very small, we only compare three categories; 'Household', 'Insurance Payment' and 'Loan Payment'. It is worth to note that most 'gold card' customers spend their

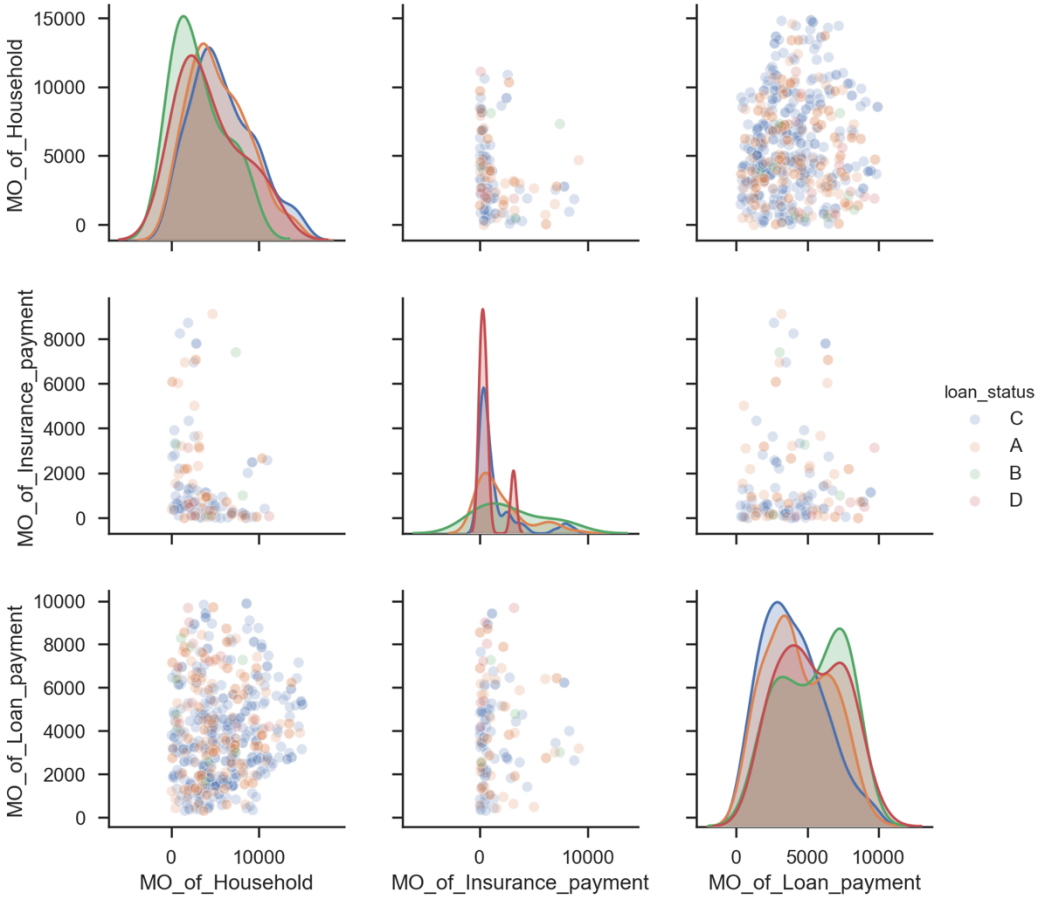
money on ‘Household’ and ‘Loan Payment’. Most ‘junior card’ customers spend their money on the ‘Insurance Payment’ category.

7) Gender distribution in each order category - plot with average amount of order



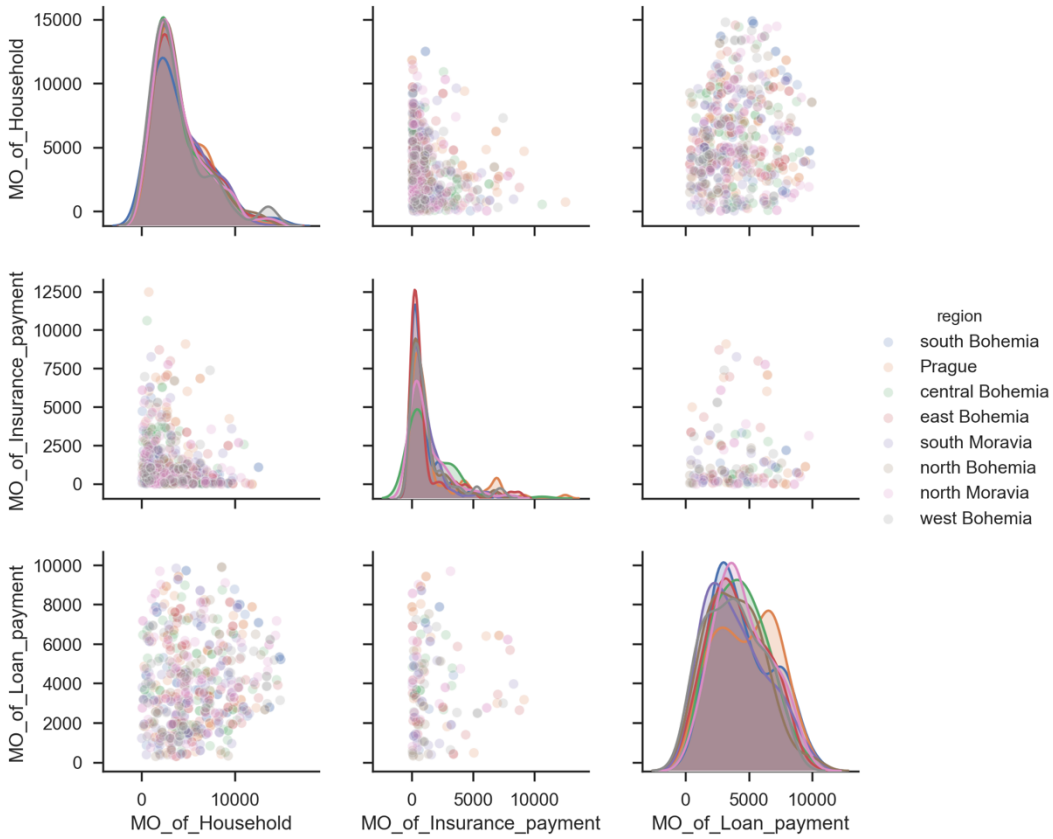
For each category, there is no significant difference between genders. This is in-line with the earlier plot (Overall order amount VS gender).

8) Loan status distribution in each order category - plot with average amount of order



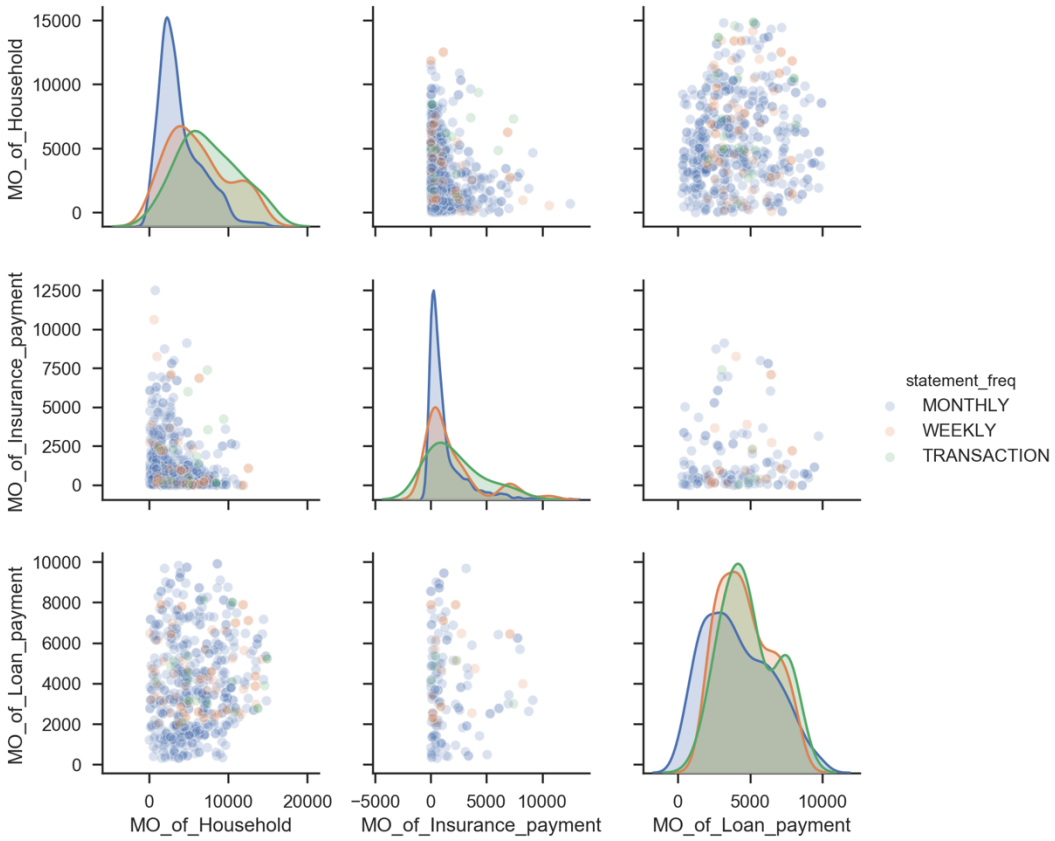
For the loan status in each category, 'B' group has the highest amount of loans in the 'Household' category, and 'C' and 'D' have the highest amount of loans in the 'Insurance Payment' group. The distributions of 'Loan Payment' are similar regardless of the loan status.

9) Regions in each order category - plot with average amount of order



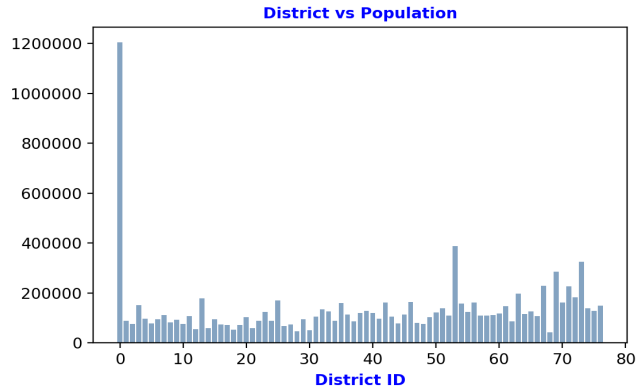
For each category, the differences among regions are insignificant.

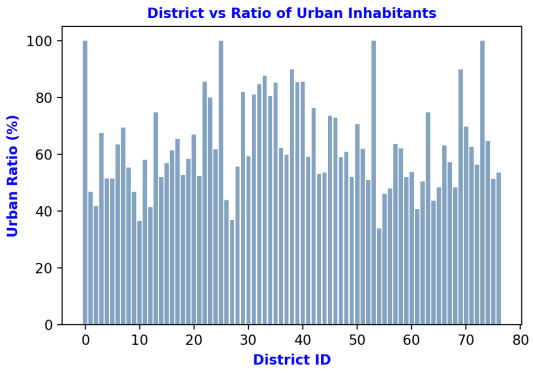
10) Statement Frequency in each order category - plot with average amount of order

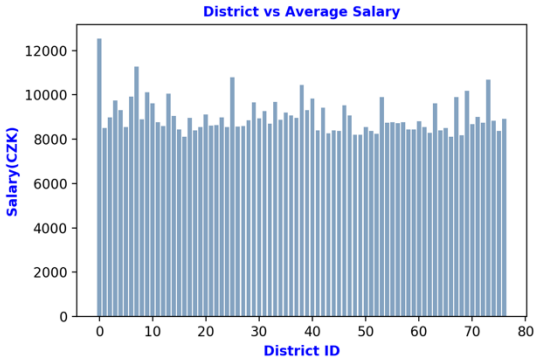


Most ‘Household’ and ‘Insurance Payment’ orders are ‘Monthly’. However, for ‘Loan Payment’, it is by ‘Weekly’ or ‘Transaction’.

B) Demographics (Population and Salary)



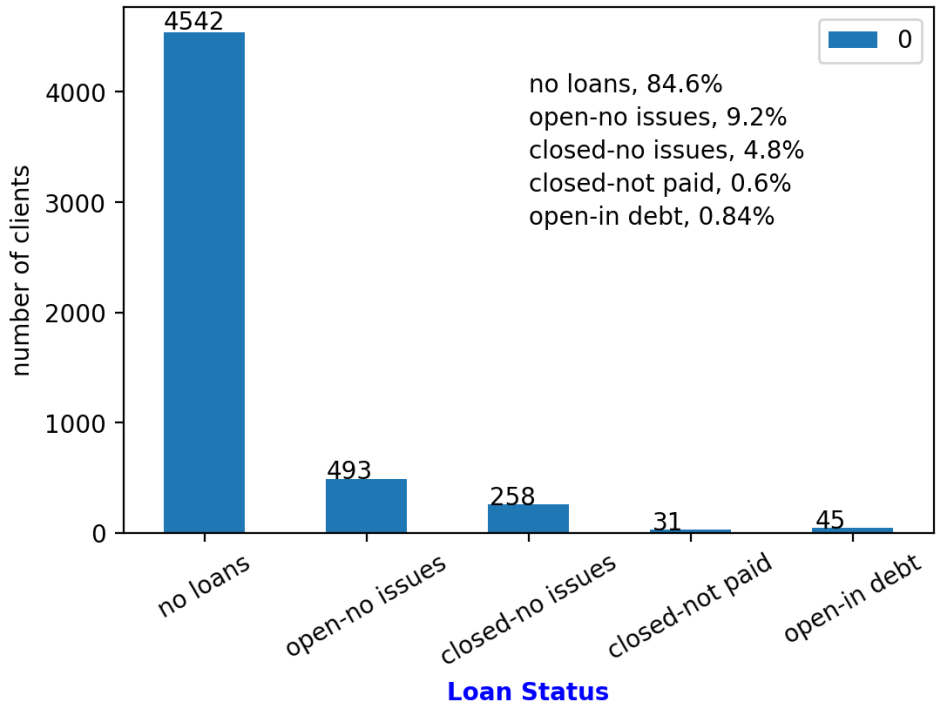




Considering the fact that the cities (urban ratio 100%) and other suburban/rural regions have a small difference in salary range, it is hard to say that urban residents always get paid more.

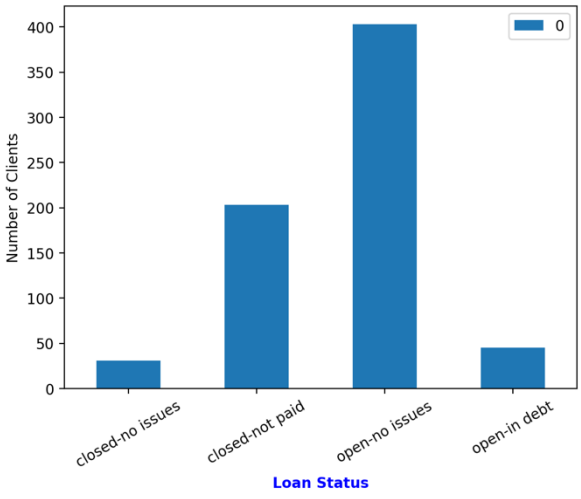
District 1 (Praha), 26(Plzen),54(Brno) and 74(Ostrava) have the urban ratio of 100%, but all the districts except district 1 have 400K or fewer inhabitants. Only District1 has the highest number of inhabitants and is categorized as city.

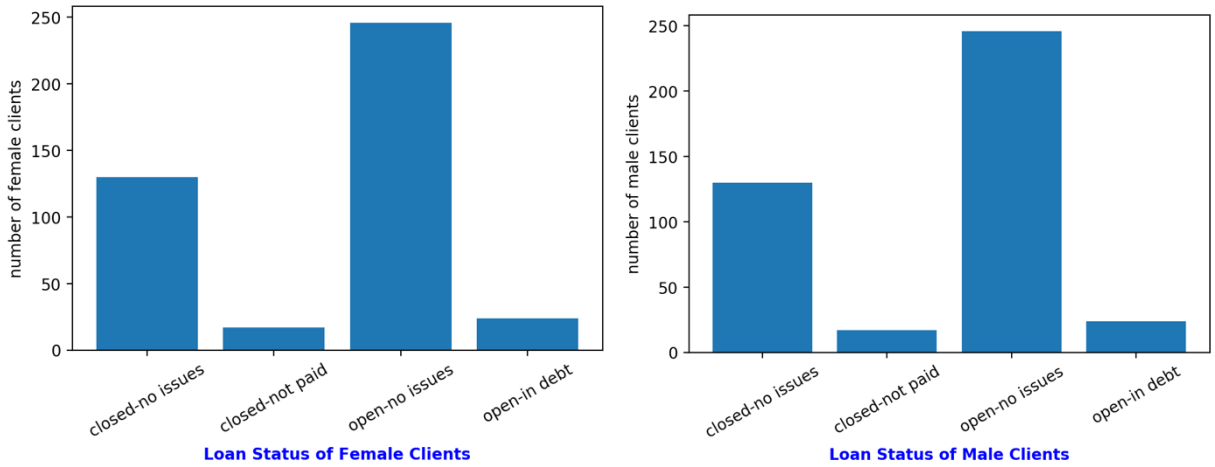
Clients with Loans



84.6% of clients do not have loans. Among clients who have open accounts and loans, only 0.84% of clients are in debt and have a hard time paying them on time.

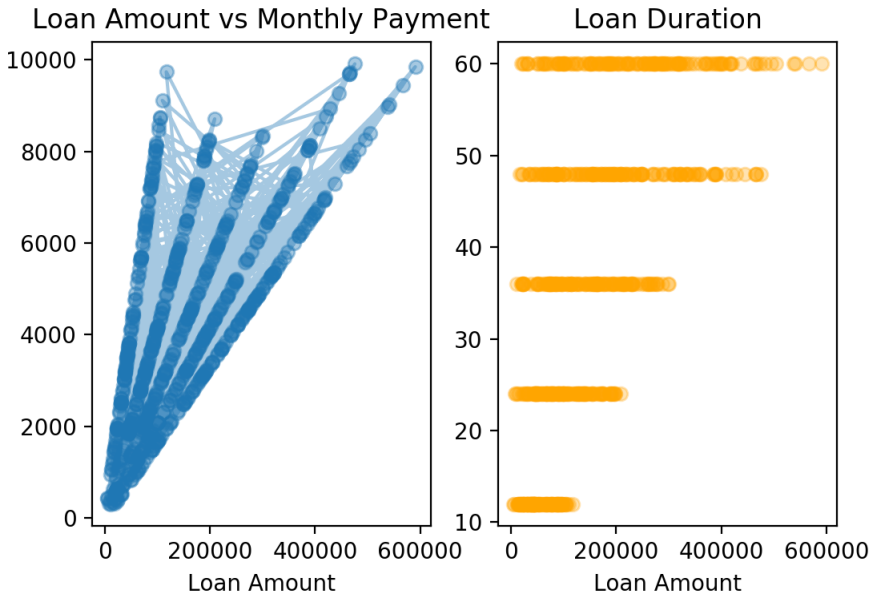
Loan Status vs. Gender





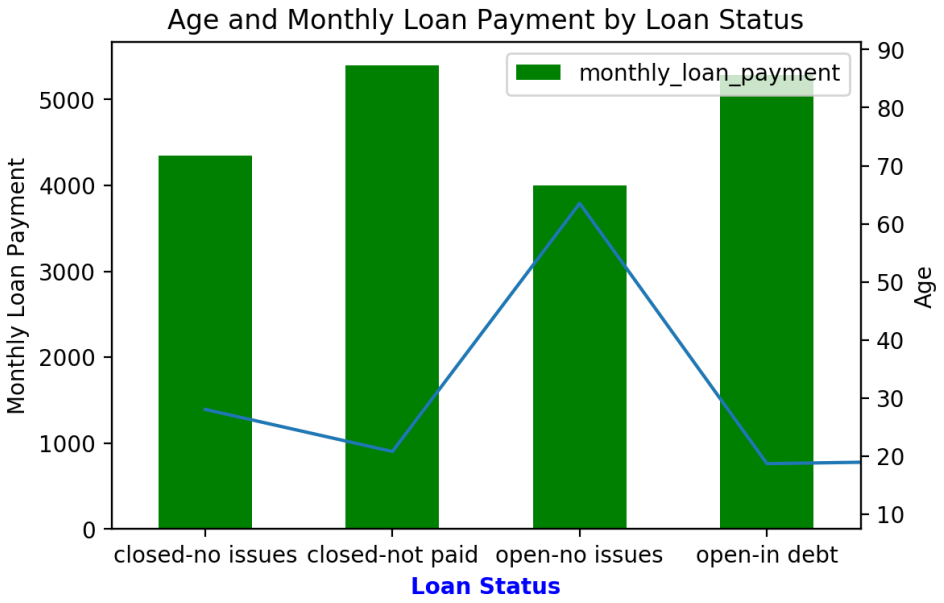
Most of clients with loans have open accounts and do not have any issues.

Loan Amount vs. Monthly Payment and Loan Duration



Clients with low amount of loans do not necessarily make low amount of monthly payments.
Clients with higher amount of loans tends to have longer loan duration.

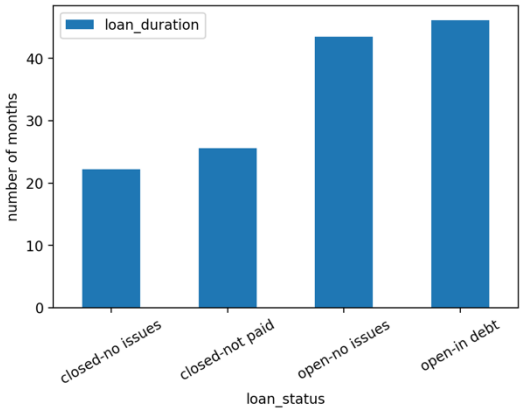
Age and Monthly Loan Payment by Loan Status

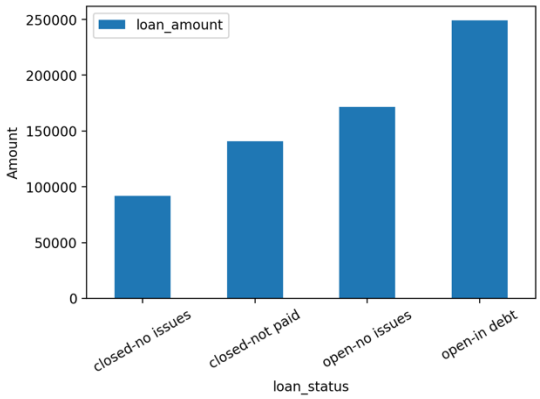


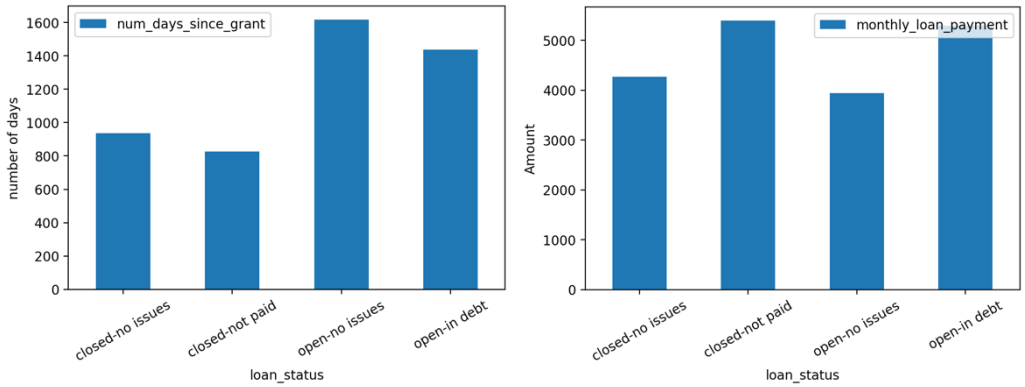
Clients with issues (not paid, in debt) make higher monthly loan payments.

Clients who are in the age range of 60 to 70 tend to have no issues paying loans. On the other hand, clients who are in the age range of 20 to 30 tend to have issues paying loans on time.

Loan Status vs. Loan Duration, Loan Amount, Number of Days Since Grant and Monthly Loan Payment

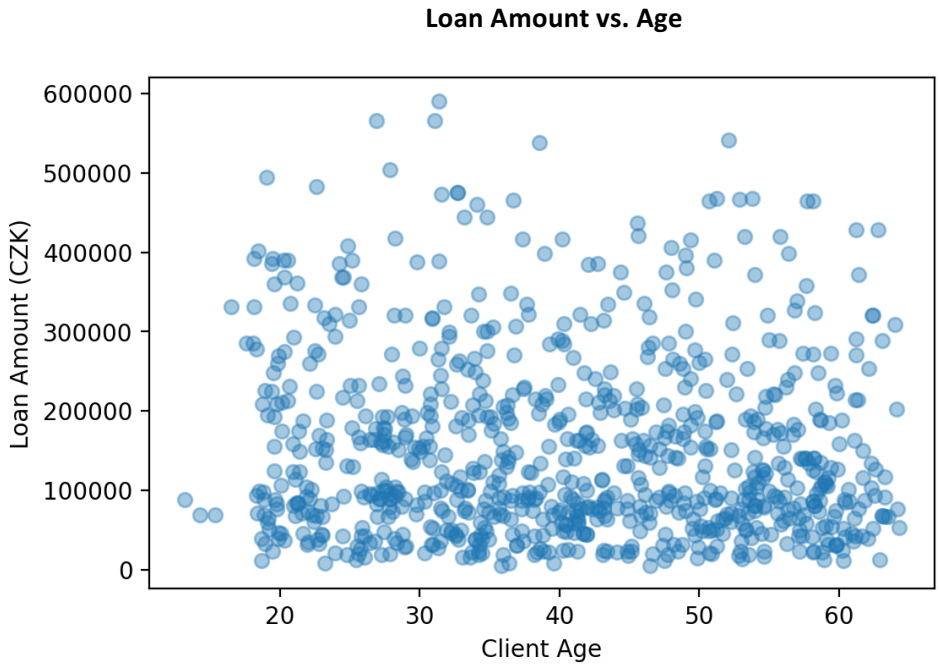






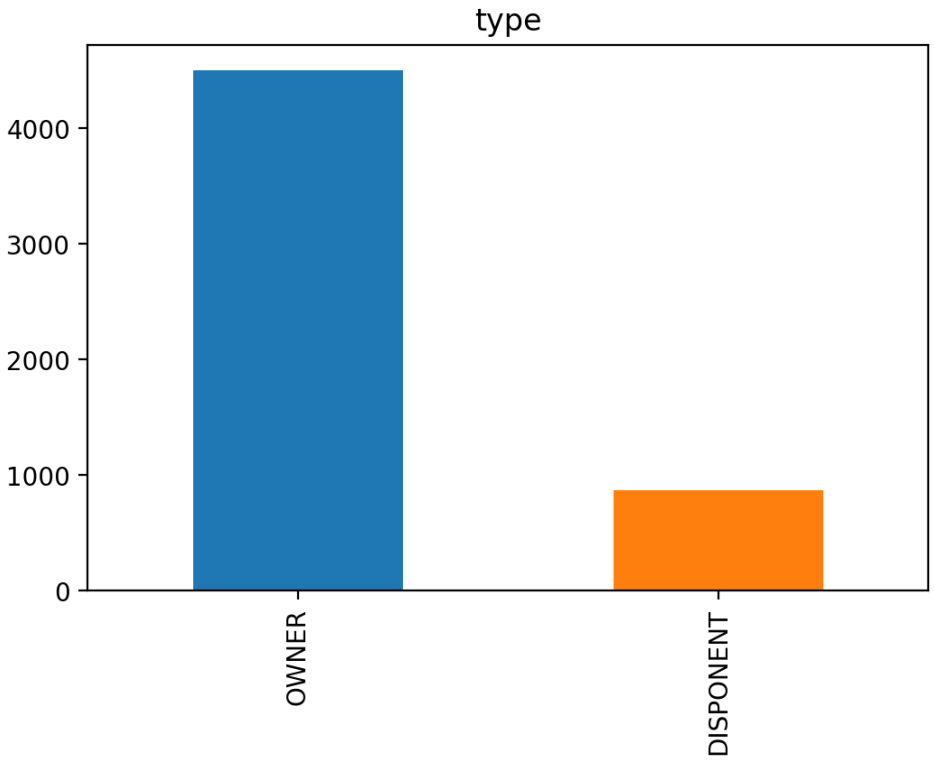
According to these four bar charts above, clients with open accounts in debt tend to make higher amount of monthly loan payments than the clients without any debt.

Clients with closed accounts had the shorter loan duration and the smaller amount of loans than the clients with open accounts.



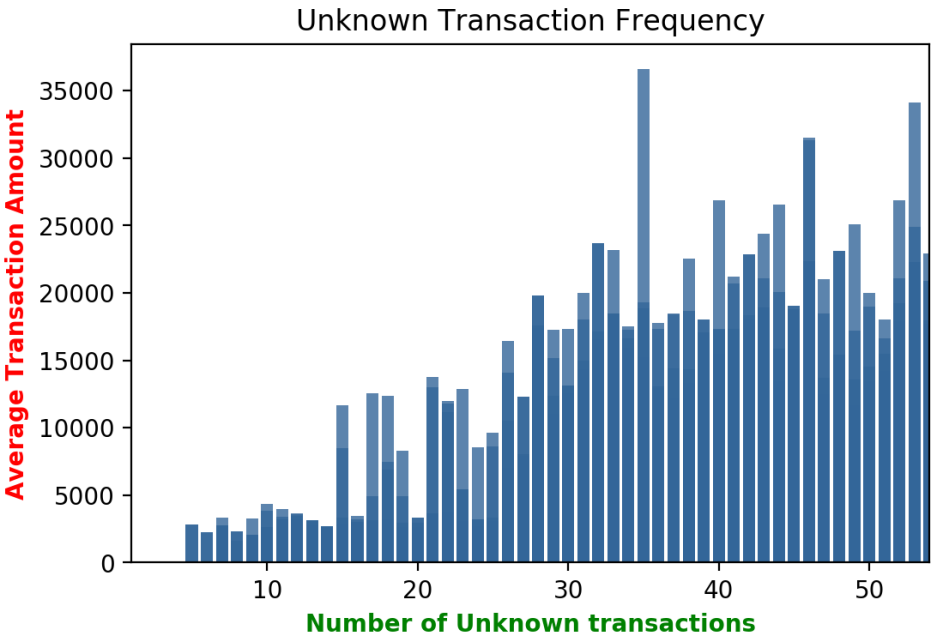
Clients regardless of the age seem to borrow loans between 50000 and 150000

Type of Clients



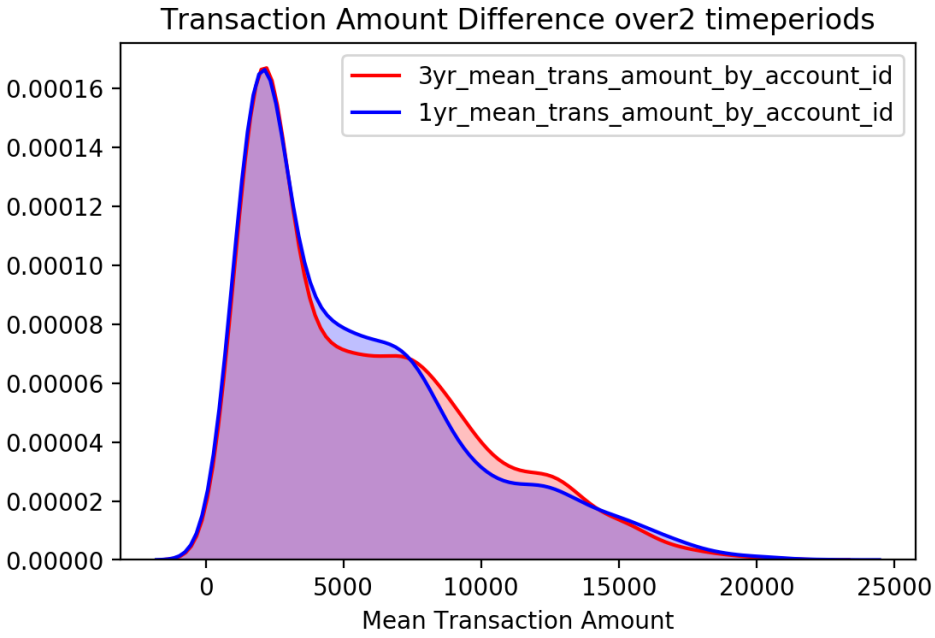
The Numbers of owners are 4 time higher than the number of disponents.

Unknown Transaction amounts and their frequencies



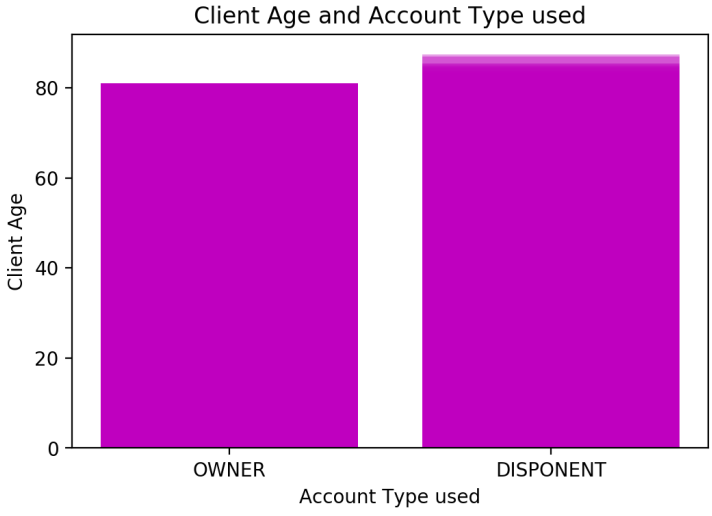
There have been a lot of occurrences of unknown transactions. It was analyzed and found that the highest number of such occurrences are 35 and has an average transaction amount of around 36000 CZK.

Average transactions over last 3 years vs last 1 year



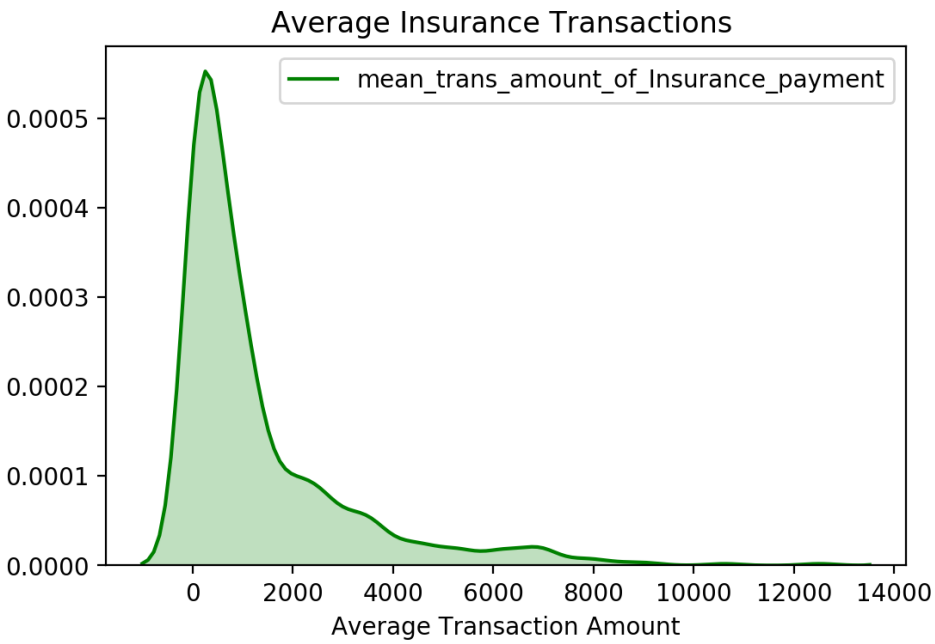
The average amount of transactions over the last 3 years and the past 1 year are almost the same showing a negligible difference. The highest average transaction amount is around 1500-2000 CZK.

Types of Accounts used by Clients of Different ages



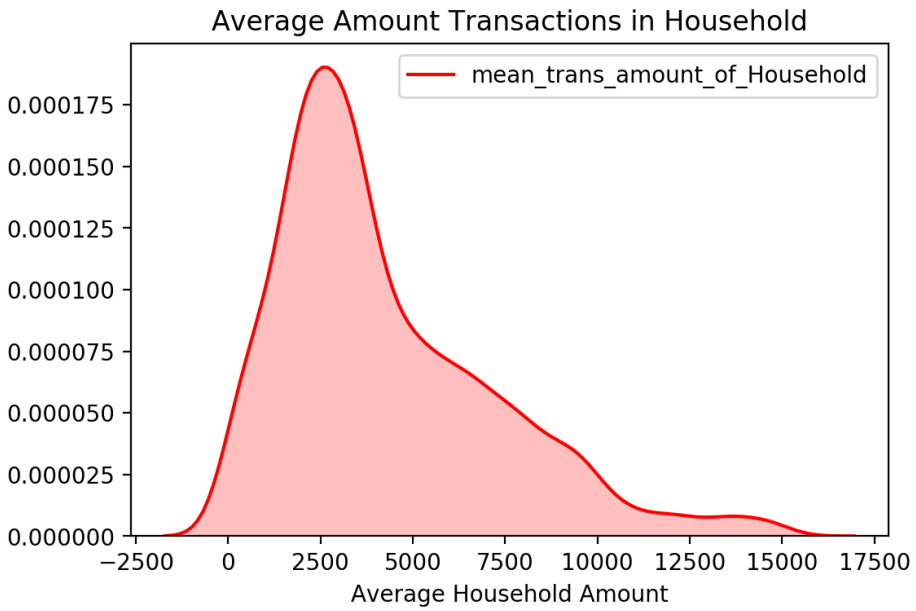
There are 2 distinct type of accounts - **Owner** and **Disponent.** The **Disponent** account holders are slightly more in the age range of 80 and above than the **Owner** account holders.

Average transaction amount for Insurance



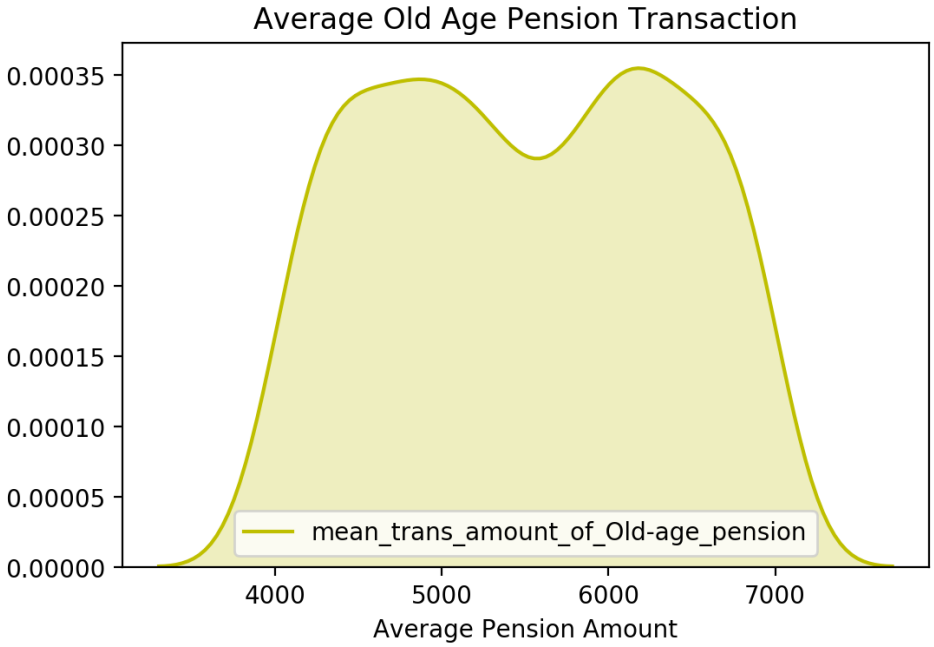
The average transaction amount for insurance is found to be around 1000 CZK.

Average transaction amount for household



The average transaction amount for household is found to be around 2500 CZK.

Average transaction amount for household



The average transaction amount for old age pension revolves around 5500 CZK.

Appendix – Datamart table explanation

Column_name	Remarks
client_id	client identifier
client_district_id	address of the client
client_age	
client_gender	
disp_id	Relation disposition record identifier
account_id	identification of the account
type	type of disposition (owner/user) only owner can issue permanent orders and ask for a loan
alltime_sum_trans_amount_by_account_id	sum of all transaction amount by account id
alltime_mean_trans_amount_by_account_id	mean of all transaction amount by account id
alltime_total_nbr_trans	total number of transactions by account id
mean_trans_amount_of_Household	mean of Household transaction amount by account id
mean_trans_amount_of_Insurance_payment	mean of Insurance_payment transaction amount by account id
mean_trans_amount_of_Interest_credited	mean of household transaction amount by account id
mean_trans_amount_of_Loan_payment	mean of Loan_payment transaction amount by account id
mean_trans_amount_of_Old-age_pension	mean of Old-age_pension transaction amount by account id
mean_trans_amount_of_Payment_for_statement	mean of Payment_for_statement transaction amount by account id
mean_trans_amount_of_Sanction_interest_if_negative_balance	mean of Sanction_interest_if_negative_balance transaction amount by account id
mean_trans_amount_of_Unknown_Transaction_Type	mean of Unknown_Transaction_Type transaction amount by account id
nbr_trans_Household	number of Household transaction by account id
nbr_trans_Insurance_payment	mean of Insurance_payment transaction by account id
nbr_trans_Interest_credited	number of household transaction by account id
nbr_trans_Loan_payment	number of Loan_payment transaction by account id
nbr_trans_Old-age_pension	number of Old-age_pension transaction by account id
nbr_trans_Payment_for_statement	number of Payment_for_statement transaction by account id
nbr_trans_Sanction_interest_if_negative_balance	number of Sanction_interest_if_negative_balance transaction by account id
nbr_trans_Unknown_Transaction_Type	number of Unknown_Transaction_Type transaction by account id
mean_trans_amount_of_Credit	mean transaction amount of credit by account id

mean_trans_amount_of_Unknown_Transaction_Type	mean transaction amount of Unknown_Transaction_Type by account id
mean_trans_amount_of_Withdrawal	mean transaction amount of Withdrawal by account id
nbr_trans_Credit	number transaction of credit by account id
nbr_trans_Unknown_Transaction_Type	number transaction of Unknown_Transaction_Type by account id
nbr_trans_Withdrawal	number transaction of Withdrawal by account id
1yr_sum_trans_amount_by_account_id	sum of transaction amount by account id in the past 1 year
1yr_mean_trans_amount_by_account_id	mean of transaction amount by account id in the past 1 year
1yr_total_nbr_trans	total number of transactions by account id in the past 1 year
3yr_sum_trans_amount_by_account_id	sum of transaction amount by account id in the past 3 year
3yr_mean_trans_amount_by_account_id	mean of transaction amount by account id in the past 3 year
3yr_total_nbr_trans	total number of transactions by account id in the past 3 year
trans_ratio_98_97	Ratio of sum of transaction (growth rate between year 1997-1998)
trans_ratio_97_96	Ratio of sum of transaction (growth rate between year 1996-1997)
sum_order_amount_by_account_id	sum of order amount by account id
mean_order_amount_by_account_id	mean of order amount by account id
total_nbr_orders	total number of order by account id
mean_order_amount_of_Household	mean order amount of Household by account id
mean_order_amount_of_Insurance_payment	mean order amount of Insurance payment by account id
mean_order_amount_of_Leasing	mean order amount of Leasing by account id
mean_order_amount_of_Loan_payment	mean order amount of Loan payment by account id
mean_order_amount_of_Unknown_Order_Type	mean order amount of Unknown Order Type by account id
nbr_orders_Household	Number of orders Household by account id
nbr_orders_Insurance_payment	Number of orders Insurance payment by account id
nbr_orders_Leasing	Number of orders Leasing by account id
nbr_orders_Loan_payment	Number of orders Loan payment by account id
nbr_orders_Unknown_Order_Type	Number of orders Unknown Order Type by account id
account_district_id	location of the branch
statement_freq	frequency of issuance of statements
account_date_opened	Date when account was opened
loan_id	Loan ID

loan_amount	Amount of loan sanctioned
loan_duration	Duration of the loan
monthly_loan_payment	Monthly payment of loan
loan_status	Status of paying off the loan
num_days_since_grant	Number of days since loan was granted
card_id	ID od card issued
card_type	Type of Cards issued
card_issued_date	Date when card was issued in YYMMDD
district_id	District ID
district_name	Name of district
region	Name of region
total_inhabitants	Total inhabitants
num_munipalities_less_499	Number of municipalities with inhabitants less than 499
500_to999	Number of municipalities with inhabitants between 500-999
2000_to_9999	Number of municipalities with inhabitants between 2000-9999
greater_than0000	Number of municipalities with inhabitants between 10000
num_cities	Number of cities
ratio_urban	Percentage of urbanization
avg_salary	Average Salary
unemp_rate_95	Unemployment rate in 1995
unemp_rate_96	Unemployment rate in 1996
enterpreneurs_per1000	Number of entrepreneurs per 1000 inhabitants
num_crimes95	Number of Crimes in 1995
num_crimes96	Number of Crimes in 1996