DATA CORRECTION AND TRANSFORMATION

- Modules used:
    - Pandas
    - numpy
    - Datetime
    - Matplotlib
    - Scipy
    - Seaborn

- **Account**
    - This table has four columns with one account id per row/observation, containing the district id, frequency of statement issuance, and date that the account was created.
    - The date column was renamed to 'date_account_created' and then converted from integer to date format using pd.to_datetime().
        - From the converted date column, new variables for the year, month, day, and weekday were created.
    - An LOR in days variable was created to calculate for the length of relationship in days of each account by the start of the dependent variable, which is Jan 1, 1997.
    - Dummy variables were created for each frequency type of statement issuance using pd.get_dummies().
        - The original frequency column was then dropped to avoid duplication.
    - The dataset was then subset to accounts created before 1996 and with owner type to serve as the independent variable.

- **Transaction**
    - The 'Vyber' values in the type column were replaced with the mode, 'VYDAJ'.
    - A year column was derived from the date column by transforming the date column to an integer and getting the first two characters and then converting it back to an integer before adding 1900.
    - The date format was then converted from integer to date format using pd.to_datetime().
    - Checking the k_symbol column, there are observations with only spaces as values. The spaces were replaced with NaN values.
    - Transactions were then subset to transactions that occurred in the year 1996 for the independent variables.
    - An additional dataframe was created to get the last balance to serve as the year-end balance in 1996.
    - Other additional variables were created for each transaction type, operation type, and k_symbol type, grouped by account ID:
        - Recency (last transaction in days)
        - Frequency (total count of transactions)

        - Monetary (total amount of transactions)
        - Average Transaction Size
    - The following data manipulations were done after creating above variables:
        - reset_index() was used to add the grouped account id as a column
        - The recency columns were divided by np.timedelta64(1, 'D') to remove the 'days' word in the values and then converted to integer type
        - The NaN values were replaced with 0 using replace attribute
    - All dataframes created for the variables were then merged into one table.

- **Disp**
    - Creation a new table as we only need the type 'OWNER'
        - We filtered and kept only owner type
        - With 'isin()'
- **Order**
    - This table as 6 columns
    - Replacement of missing values (we are defined by ' ') in the k_symbol column by NaN value
        - With 'replace()'
    - Creation of dummies for k_symbol, we now have with :get_dummies()':
        - k_symbol_LEASING
        - k_symbol_POJISTNE
        - k_symbol_SIPO
        - k_symbol_UVER
    - We have created a new table for the k_symbol dummies called "order2"
    - Creation of a new table called "order_client" to have an order table grouped by account_id
        - With 'pd.DataFrame()'
    - Additional variables to summarize the total amount of recurring payments by client ID were also added in the order_amount df by pivoting the order dataset.
- Demo
    - Replace the '?' value in A15 variable by NaN
        - With 'replace()'
    - Replace the '?' value in A12 variable by NaN
        - With 'replace()'
    - Creation of a new variable 'rate_crime_96' to have the growth rate for the crime in 1996, rounded to 2 decimals
    - Creation of a new variable 'num_mun' to the the number of municipalities per district
        - With 'sum()'
    - We rename all the columns by what they represent so it is more understandable
        - With 'rename()'

- **Loan**

- Cleaning data, we checked missing values and there was no need to replace anything because there were no missing values.
- Changed date format in Loan table to year-month-day using pd.to_datetime with format=%y%m%d, and renamed the column as 'loan-date' using rename function to make it easier when we merge everything together.
- Created two data-frames from table Loan:
  - The first data frame was filtered between 1-1-1997 and 31-12-1997 and it was called Loan1997.
  - The second data frame was filtered to before 1-1-1997 and it was called Loan_before_1997.
- Created dummy variables for each loan status using pd.get_dummies, and then combined the created dummy data frame with the main loan_before_1997 table using concat function.
- Dropped the status column and loan id column using drop function.
- Merged Loan1997 table with Loan_before_1997 table using merge funtion by outer join.
- The target from this merge was to be able to create a final table Loan1997 having accounts ids specified by who had a loan before 1997 and who didn't, and we achieved that by replacing NaN values in the table by 1 as account ids before 1997 and 0 as accounts ids created in 1997.

- **Client Table**
  - Based on previous work in class
  - Created multiple variables from birth number, taking first two numbers and adding 19 we were able to get birth_year, taking the third and four number we got the birth_month, and using the last two numbers we got birth_day.
  - Created gender using a condition if birth > 50 then female, if lower than Male, and fixed the month number by using birth_month – 50.
  - Created age using the year we want 1996 minus the client birth_year, and used Client['age'] // 10 * 10 to get age_group.
  - Created dummy variables for each gender and combined it with the original table.

- **Credit Card**

  - Cleaning data, we checked missing values and there was no need to replace anything because there were no missing values.
  - Changed date format in Loan table to year-month-day using pd.to_datetime with format=%y%m%d, and renamed the column as 'credit_card_issue_date' using rename function to make it easier when we merge everything together.
  - Created two data-frames from table Credit Card:
    - The first data frame was filtered between 1-1-1997 and 31-12-1997 and it was called Credit_Card_1997.

        - The second data frame was filtered to before 1-1-1997 and it was called Credit_Card_before_1997.
    - Created dummy variables for each credit card type using pd.get_dummies, and then combined the created dummy data frame with the main Credit_Card_before_1997 table using concat funtion.
    - Dropped the type of column and card id column using drop funtion.
    - Merged Credit_Card_1997 table with Credit_Card_before_1997 table using merge funtion by outer join.
    - The target from this merge was to be able to create a final table Credit_Card_1997 having disp ids specified by who had a credit card issued before 1997 and who didn't, and we achieved that by replacing NaN values in the table by 1 as disp ids before 1997 and 0 as disp ids created in 1997.

- Merging
    - After data cleaning, setting each table to account owner level, and the creation of variables per table, all datasets were merged into one consolidated basetable. The dependent variables were also added with a value of 1 for YES and 0 for NO.

BASETABLE DESCRIPTION AND ANALYSIS

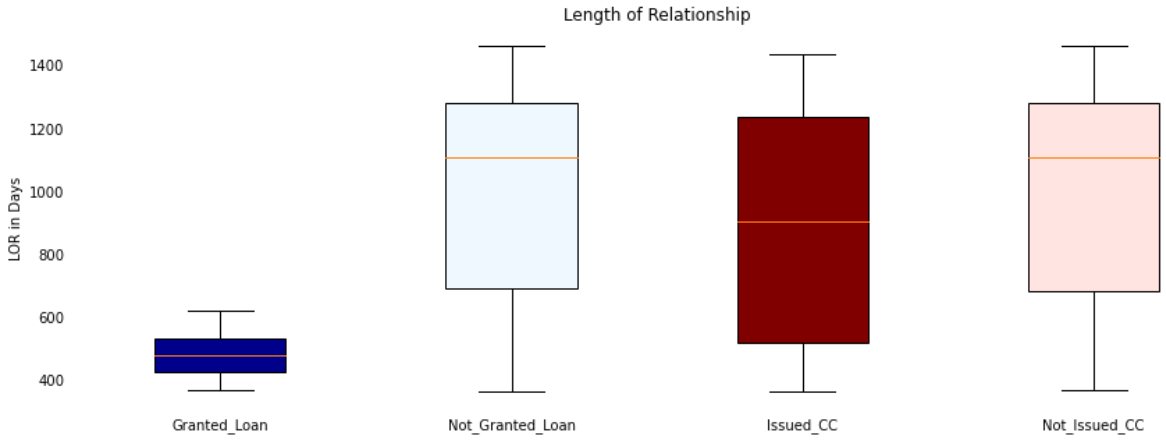
After data preparation and merging, the consolidated basetable now has 124 columns and 2,239 observations. There are 31 clients who are account owners and whose accounts were created before 1996 that were granted a loan in 1997 and 120 clients with the same criteria who were issued a credit card in 1997. Therefore one must take caution in analyzing the group with granted loan given the loan base.

Below charts illustrate the independent variables and dependent variables.

LENGTH OF RELATIONSHIP

The boxplot below illustrates the distribution of the length of relationship in days across: (1) clients who were granted a loan (Granted_Loan), clients who were not granted a loan (Not_Granted_Loan), clients who were issued a credit card (Issued_CC), and clients who were not issued a credit card (Not_Issued_CC). (Caveat: For the succeeding graphs, Only Granted_Loan vs Not_Granted_Loan and Issued_CC vs Not_Issued_CC are mutually exclusive.)

- Granted_Loan has the smallest distribution and the lowest LOR, indicating that these are with the bank for the less than 2 years.
- Issued_CC has a wider range of relationship with the bank.
- Clients without a loan granted or a credit card issued in 1997 display a higher median than the other groups.



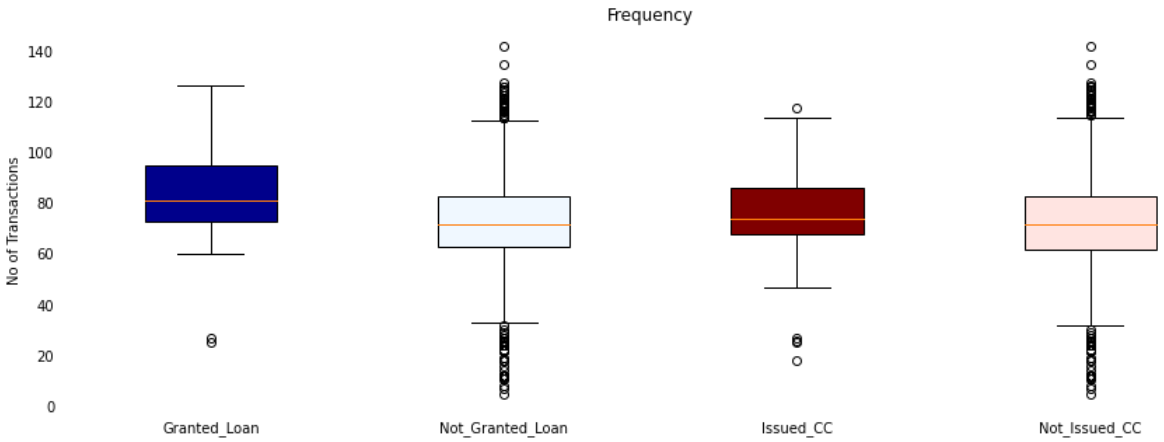
- A t-test done between the two groups displays a significant difference in the length of relationship between Granted_Loan vs. Not_Granted_Loan and Issued_CC and Not_Issued_CC at 99% CI.

	Granted_Loan	Not_Granted_Loan	Issued_CC	Not_Issued_CC
Mean	485.645	998.024	900.842	996.032
Var	4794.570	111,051.611	126,181.008	111,999.980
T-Test		-35.790		-2.864
p-Value		1.313e-38		0.005

FREQUENCY

The boxplot below illustrates the distribution of frequency (number of transactions) across the four groups.

- The median across 4 groups range between 70 to 80 transactions for the year however the Granted_Loan and Issued_CC groups show a smaller scale.



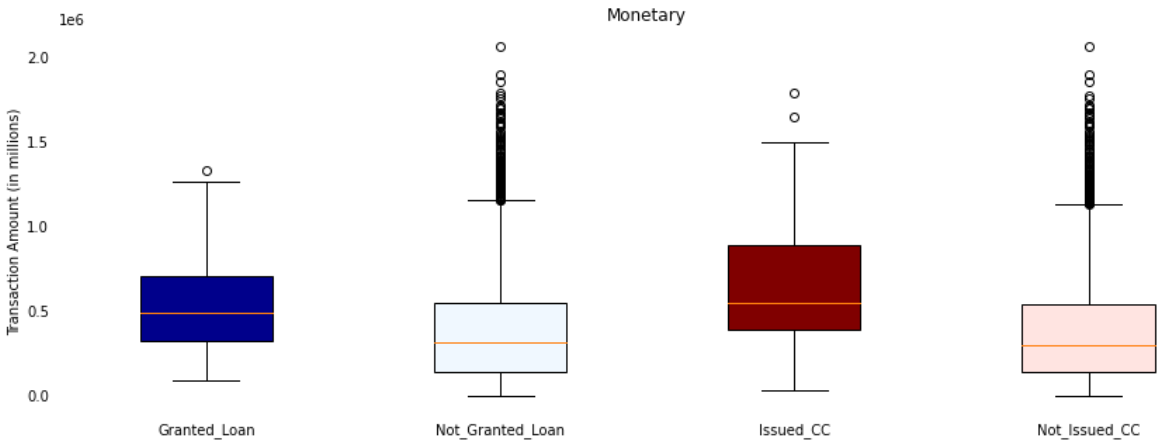
- A t-test done between the two groups displays that there is no significant difference in the frequency between Granted_Loan vs. Not_Granted_Loan and Issued_CC and Not_Issued_CC at 99% CI.

	Granted_Loan	Not_Granted_Loan	Issued_CC	Not_Issued_CC
Mean	81.645	73.053	76.283	72.995
Var	453.837	310.479	310.793	312.978
T-Test		2.235		1.9872
p-Value		0.039		0.049

Monetary

The boxplot below illustrates the distribution of monetary (amount of transactions) across the four groups.

- The middle boxes of Granted_Loan and Issued_CC sits higher than Not_Granted_Loan and Not_Issued_CC, respectively, indicating more spending from the former groups.

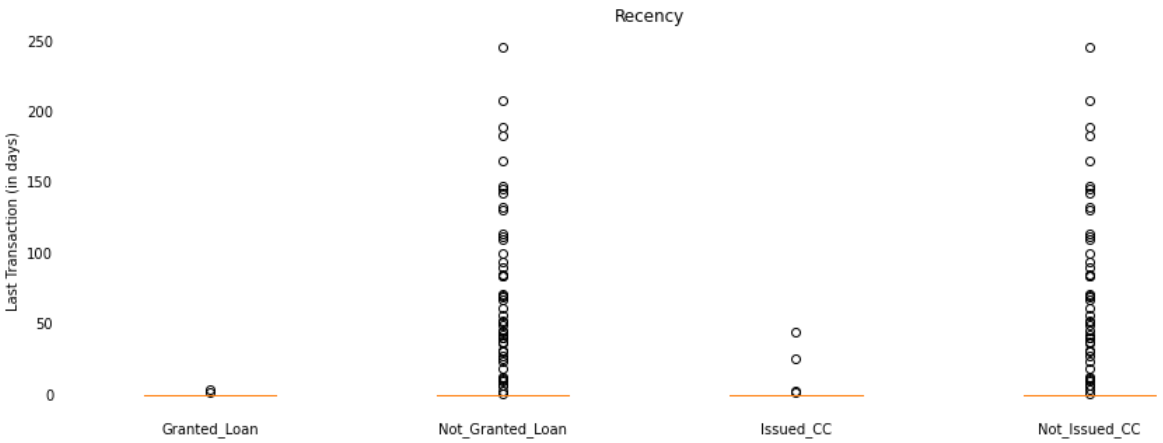


- A t-test done between the two groups displays that there is a significant difference in the monetary between Issued_CC and Not_Issued_CC at 99% CI.

	Granted_Loan	Not_Granted_Loan	Issued_CC	Not_Issued_CC
Mean	559,968.739	404,444.403	642,134.765	393,259.132
Var	108,693,640,346.355	118,334,539,447.979	151,505,850,243.243	113,362,157,856.824
T-Test		2.607		6.860
p-Value		0.014		2.559e-10

Recency

The boxplot below illustrates the distribution of recency (last transaction(in days)) across the four groups. Majority of the clients remain active as of the last day of 1996 but the Not_Granted_Loan and Not_Isuued_CC tend to show more outliers who have not been less active than their groups.



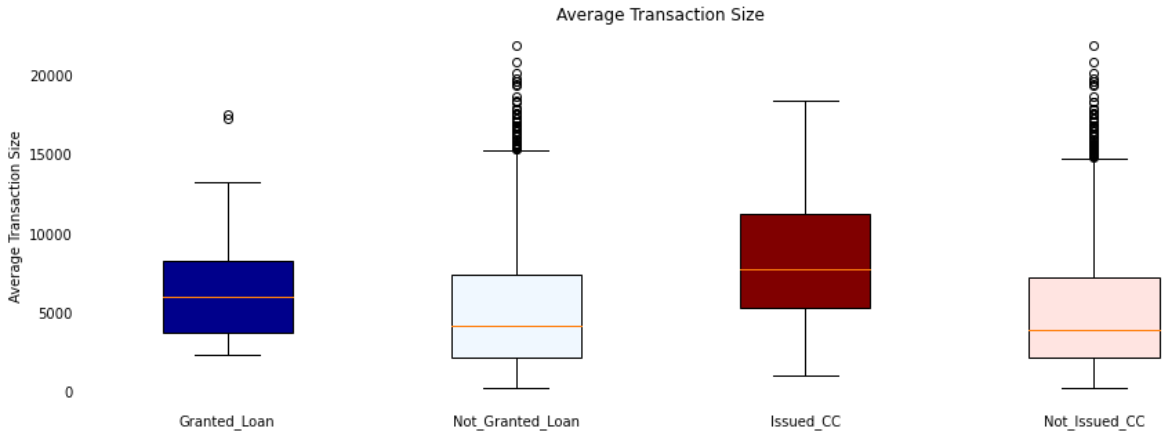
	Granted_Loan	Not_Granted_Loan	Issued_CC	Not_Issued_CC
Mean	0.194	1.632	0.617	1.668
Var	0.628	193.483	21.247	200.402

T-Test	-4.377	-2.017
p-Value	1.390e-05	0.045

Average Transaction Size

The boxplot below illustrates the distribution of average transaction size (amount per transaction) across the four groups.

- Issued_CC displays a higher transaction size vs. the Not_Issued_CC group.



- Similar with monetary, a t-test done between the two groups displays that there is a significant difference in the average transaction size between Issued_CC and Not_Issued_CC at 99% CI.

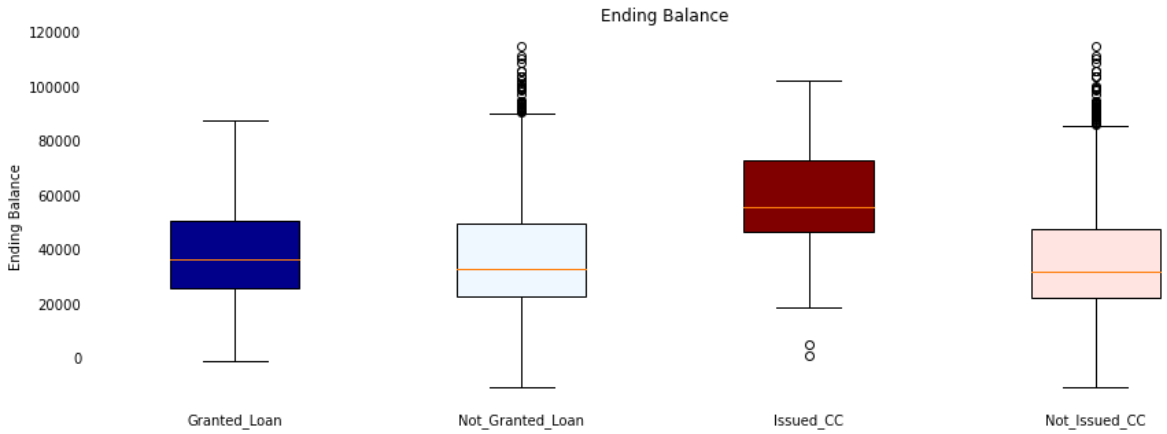
	Granted_Loan	Not_Granted_Loan	Issued_CC	Not_Issued_CC
Mean	6,860.406	5,280.422	642,134.765	393,259.132
Var	16,198,050.117	15,585,117.982	151,505,850,243.243	113,362,157,856.824
T-Test		2.171		7.923
p-Value		0.038		8.739e-13

Ending Balance

The boxplot below illustrates the distribution of monetary (amount of transactions) across the four groups.

- The Granted_Loan and Not_Granted_Loan boxes are almost of the same level and size, showing parity ending balances across the two graphs.

- On the other hand, the Issue_CC box display a higher account balance than those clients who were not issued a CC in 1997.

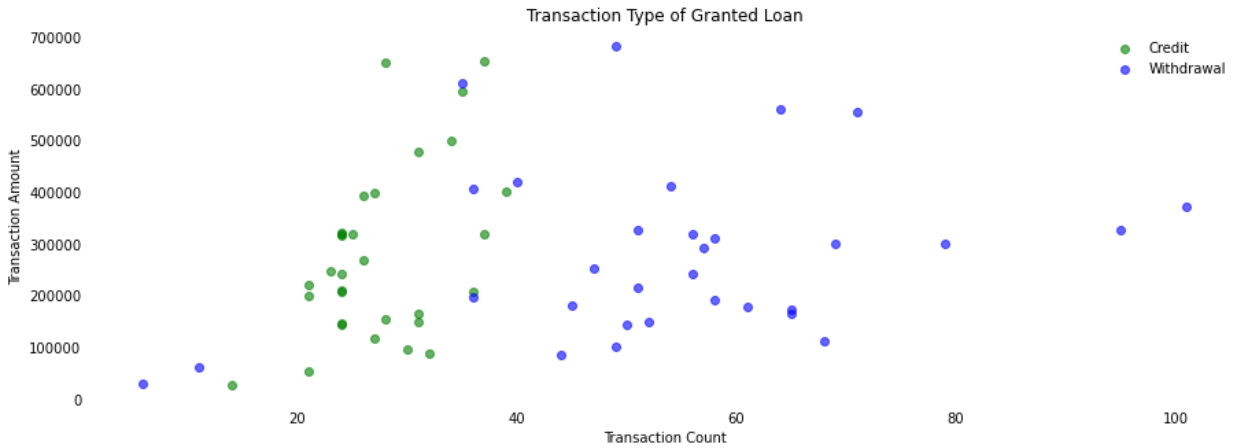


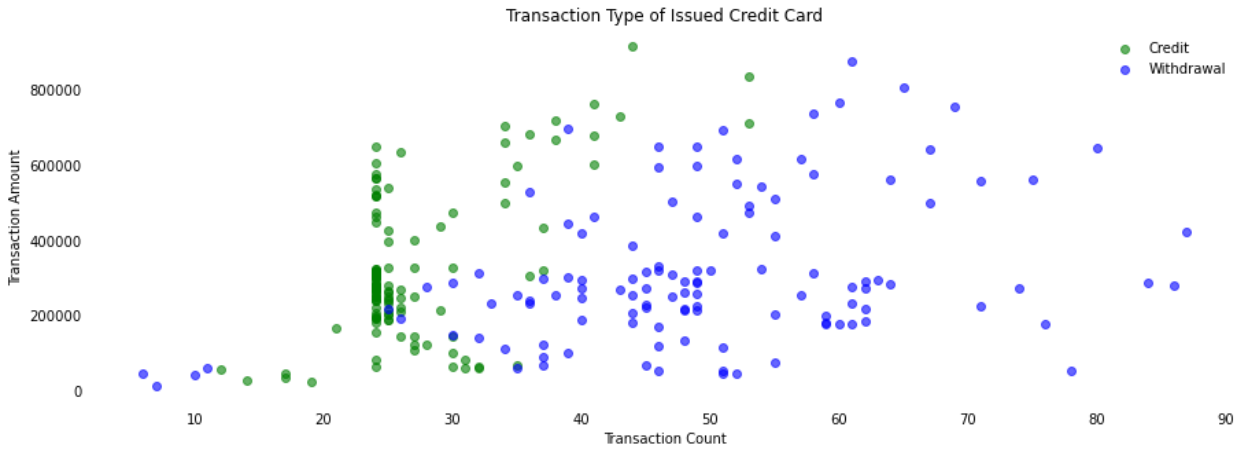
- A t-test done between the two groups displays that there is a significant difference in the ending account balance between Issued_CC and Not_Issued_CC at 99% CI.

	Granted_Loan	Not_Granted_Loan	Issued_CC	Not_Issued_CC
Mean	39,552.310	37,708.315	58,042.941	36,583.732
Var	395,739,987.952	394,913,859.797	367,114,435.446	371,844,366.424
T-Test		0.513		11.931
p-Value		0.612		9.151e-23

Transaction Type

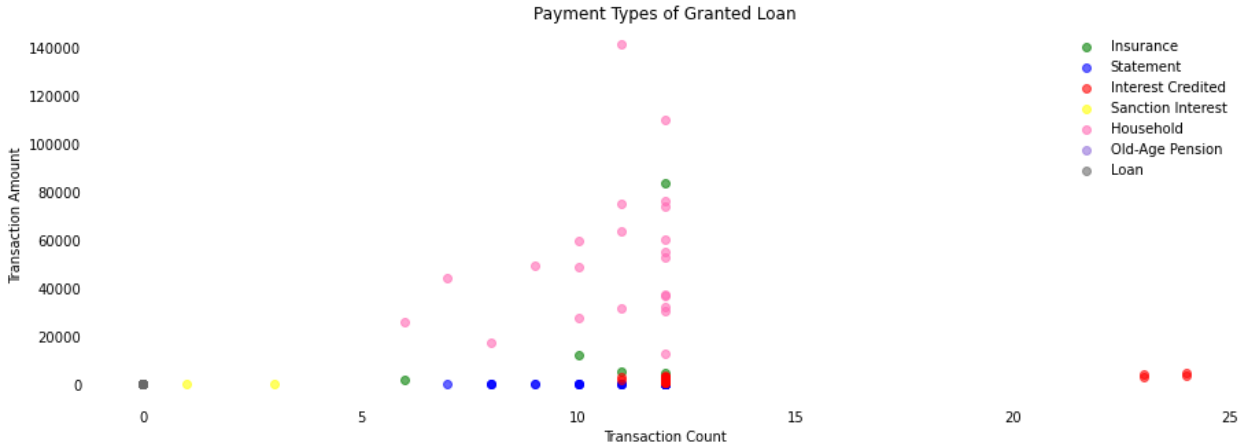
Both the Granted_Loan and Issued_CC groups perform more withdrawals vs credits.

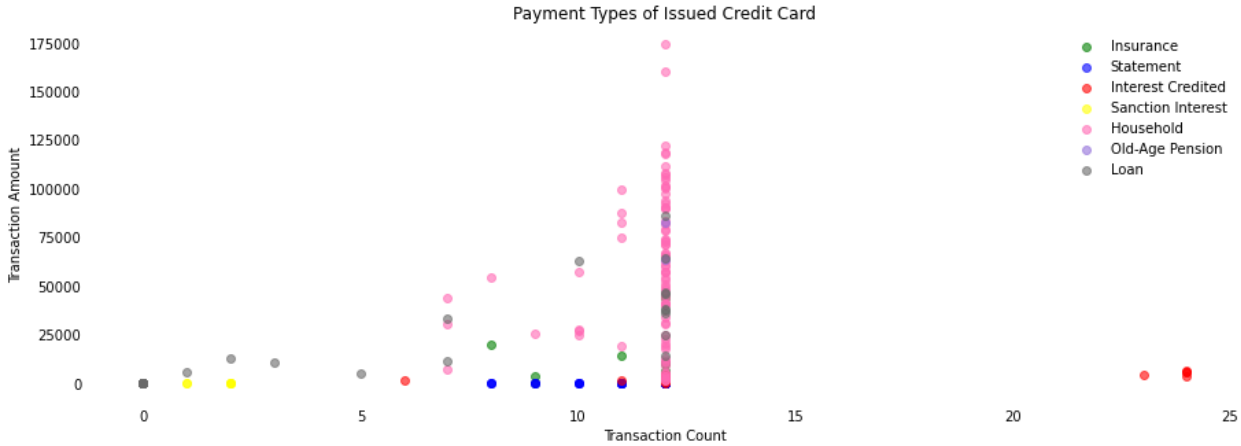




Payment Types

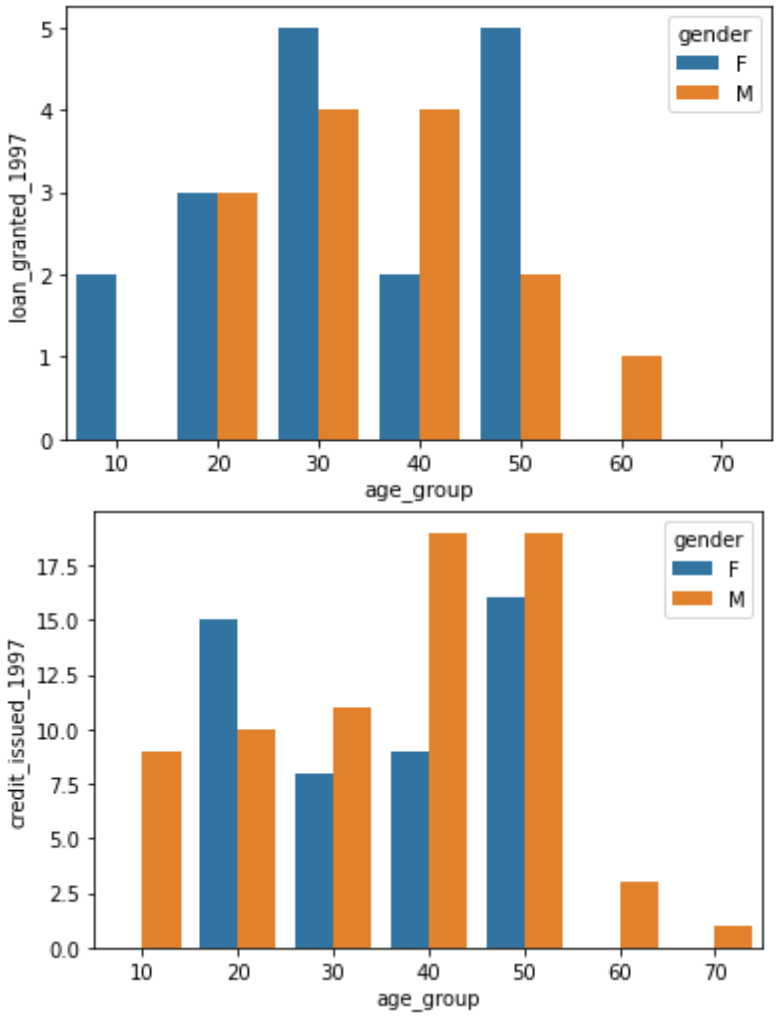
The majority of the transactions of the Granted_Loan and Issued_CC groups are for household payments.





Demographics

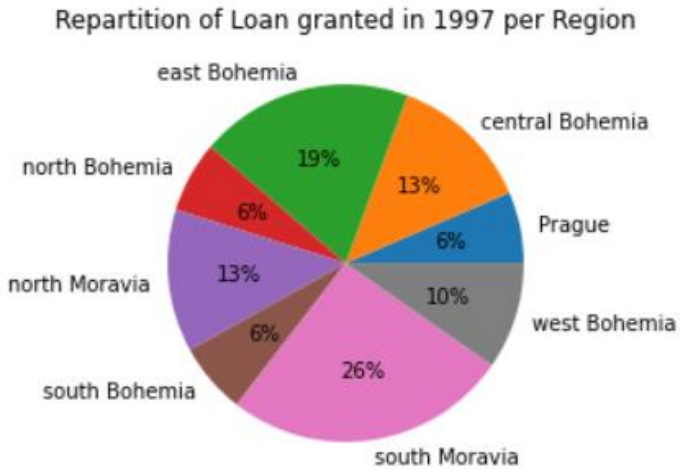
Bar Charts showing distribution of Loan Granted in 1997 and Credit Card issued in 1997 by gender and age group
(Seaborn library was used for creating these two bar graphs)

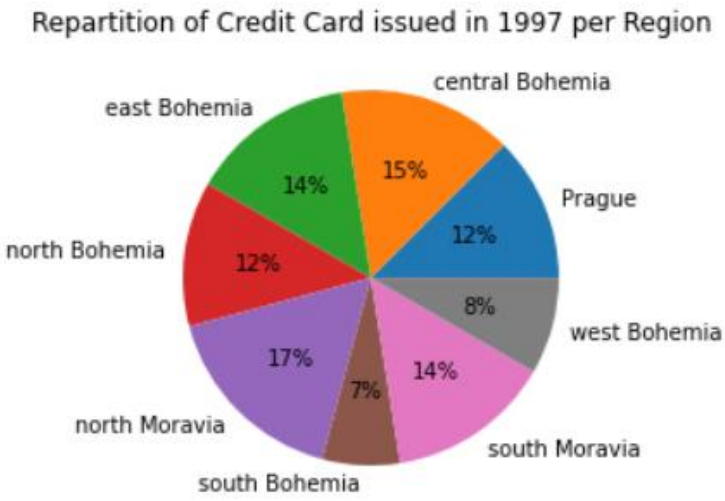


The first bar graph shows that clients who were granted a loan are spread among male and female and t the age groups. On the other hand, age groups 20, 40 and 50 and the male group have been issued the most amount of credit cards in 1997. These ages are likely clients who are buying houses or cars, and students who are finding a way to pay their learning and living expenses.

Pie Charts showing distribution of Loan Granted in 1997 and Credit Card issued in 1997 by Region

Those pie charts are here to show the distribution of loans granted and credit cards issued by region. We can see if the distribution of credit cards or loans is the same in regions. Here the top 3 in both charts is the same: South Moravia, East Bohemia and Central Bohemia.





(Seaborn library was used for creting these two bar graphs)

LIST OF VARIABLES

Variable Name	Table Source	Description	Data Type	Value
disp_id	disposition	record identifier	integer	
client_id	disposition	client identifier	integer	
account_id	disposition	identification of the account	integer	
type	account	type of disposition (owner/user)	string	
date_account_created	account	date of creating of the account	date	date
year_account_created	account	year of creating of the account	integer	Year
month_account_created	account	month of creating of the account	integer	1 to 12
day_account_created	account	day of creating of the account	integer	1 to 31
weekday_account_created	account	weekday of creating of the account	integer	1 to 7
LOR_in_days	account	length of relationship as of 01-01-1997 in days	integer	
Statement_Issuance__After_Trans	account	dummy variable for issuance of statement after every transaction	integer	1 or 0
Statement_Issuance__Monthly	account	dummy variable for monthly issuance of statement	integer	1 or 0
Statement_Issuance__Weekly	account	dummy variable for weekly issuance of statement	integer	1 or 0
district_id	client	location of the branch	integer	
birth_year	client	year of birth of client	integer	
birth_day	client	day of birth of client	integer	
birth_month	client	month of birth of client	integer	
age	client	age of client as of 1997	integer	
age_group	client	age group of client	integer	
Client_Female	client	dummy variable for female clients	integer	1 or 0
Client_Male	client	dummy variable for male clients	integer	1 or 0

credit_amount	transaction	total amount for credit transactions in 1996 per client	float	
wd_amount	transaction	total amount for withdrawal transactions in 1996 per client	float	
credit_count	transaction	total count of credit transactions in 1996 per client	float	
wd_count	transaction	total amount of withdrawal transactions in 1996 per client	float	
last_credit_in_days	transaction	number of days since client made a credit transaction from 12-31-1996	integer	
last_wd_in_days	transaction	number of days since client made a withdrawal transaction from 12-31-1996	integer	
credit_tran_size	transaction	average amount per credit transaction per client	float	
wd_tran_size	transaction	average amount per withdrawal transaction per client	float	
ending_balance_1996	transaction	ending account balance as of 12-31-1996 per client	float	
total_amount	transaction	total amount for transactions in 1996 per client	float	
total_tran_count	transaction	total count of transactions in 1996 per client	float	
ave_tran_size	transaction	average amount per transaction per client	float	
last_tran_in_days	transaction	number of days since client made a transaction from 12-31-1996	integer	
old_age_pension_amount	transaction	total amount for old age pension transactions in 1996 per client	float	

insurance_pmt_amount	transaction	total amount for insurance payment transactions in 1996 per client	float	
sanction_interest_amount	transaction	total amount for sanction interest transactions in 1996 per client	float	
household_pmt_amount	transaction	total amount for household transactions in 1996 per client	float	
statement_pmt_amount	transaction	total amount for statement pmt transactions in 1996 per client	float	
interest_credited_amount	transaction	total amount for interest credited transactions in 1996 per client	float	
loan_pmt_amount	transaction	total amount for loan payment transactions in 1996 per client	float	
old_age_pension_count	transaction	total count of old age pension transactions in 1996 per client	float	
insurance_pmt_count	transaction	total count of insurance payment transactions in 1996 per client	float	
sanction_interest_count	transaction	total count of sanction interest transactions in 1996 per client	float	
household_pmt_count	transaction	total count of household transactions in 1996 per client	float	
statement_pmt_count	transaction	total count of statement pmt transactions in 1996 per client	float	
interest_credited_count	transaction	total count of interest credited transactions in 1996 per client	float	
loan_pmt_count	transaction	total count of loan payment transactions in 1996 per client	float	
last_insurance_payments_in_days	transaction	number of days since client made a insurance payment transaction from 12-31-1996	float	

last_statement_payments_in_days	transaction	number of days since client made a statement pmt transaction from 12-31-1996	float	
last_interest_credited_in_days	transaction	number of days since client made a interest credited transaction from 12-31-1996	float	
last_sanction_interest_in_days	transaction	number of days since client made a sanction interest transaction from 12-31-1996	float	
last_household_payments_in_days	transaction	number of days since client made a household transaction from 12-31-1996	float	
last_oldage_pension_credited_in_days	transaction	number of days since client made a old age pension transaction from 12-31-1996	float	
last_loan_payments_in_days	transaction	number of days since client made a loan payment transaction from 12-31-1996	float	
old_age_pension_tran_size	transaction	average amount per old age pension transaction per client	float	
insurance_pmt_tran_size	transaction	average amount per insurance payment transaction per client	float	
sanction_interest_tran_size	transaction	average amount per sanction interest transaction per client	float	
household_pmt_tran_size	transaction	average amount per household transaction per client	float	
statement_pmt_tran_size	transaction	average amount per statement pmt transaction per client	float	
interest_credited_tran_size	transaction	average amount per interest credited transaction per client	float	

loan_pmt_tran_size	transaction	average amount per loan payment transaction per client	float	
remittance_other_bank_amount	transaction	total amount for remittance (to another bank) transactions in 1996 per client	float	
collection_other_bank_amount	transaction	total amount for collection (from another bank) transactions in 1996 per client	float	
credit_cash_amount	transaction	total amount for credit transaction in cash in 1996 per client	float	
cash_wd_amount	transaction	total amount for cash withdrawal transactions in 1996 per client	float	
credit_card_wd_amount	transaction	total amount for credit card withdrawal transactions in 1996 per client	float	
remittance_other_bank_count	transaction	total count of remittance (to another bank) transactions in 1996 per client	float	
collection_other_bank_count	transaction	total count of collection (from another bank) transactions in 1996 per client	float	
credit_cash_count	transaction	total count of credit transaction in cash in 1996 per client	float	
cash_wd_count	transaction	total count of cash withdrawal transactions in 1996 per client	float	
credit_card_wd_count	transaction	total count of credit card withdrawal transactions in 1996 per client	float	
last_credit_card_wd_in_days	transaction	number of days since client made a credit card withdrawal transaction from 12-31-1996	integer	

last_credit_cash_in_days	transaction	number of days since client made a credit transaction in cash from 12-31-1996	integer	
last_collection_other_bank_in_days	transaction	number of days since client made a collection (from another bank) transaction from 12-31-1996	integer	
last_cash_wd_in_days	transaction	number of days since client made a cash withdrawal transaction from 12-31-1996	integer	
last_remittance_other_bank_in_days	transaction	number of days since client made a remittance (to another bank) transaction from 12-31-1996	integer	
remittance_other_bank_tran_size	transaction	average amount per remittance (to another bank) transaction per client	float	
collection_other_bank_tran_size	transaction	average amount per collection (from another bank) transaction per client	float	
credit_cash_tran_size	transaction	average amount per credit transaction in cash per client	float	
cash_wd_tran_size	transaction	average amount per cash withdrawal transaction per client	float	
credit_card_wd_tran_size	transaction	average amount per credit card withdrawal transaction per client	float	
credit_card_issue_date	card	issue date	date	
credit_type_classic	card	dummy variable if credit card is classic	float	
credit_type_gold	card	dummy variable if credit card is gold	float	1 or 0
credit_type_junior	card	dummy variable if credit card is junior	float	1 or 0
loan_date	loan	date when the loan was granted	date	
amount	loan	amount of money loaned	float	
duration	loan	duration of the loan	float	

payments	loan	monthly payments	float	
loan_status_A	loan	dummy variable if loan status is A	float	1 or 0
loan_status_B	loan	dummy variable if loan status is B	float	1 or 0
loan_status_C	loan	dummy variable if loan status is C	float	1 or 0
loan_status_D	loan	dummy variable if loan status is D	float	1 or 0
total_order	order	total orders	float	
total_recurring_amount	order	total amount of recurring payments	float	
total_recurring_leasing	order	total amount of recurring leasing payments	float	
total_recurring_insurance	order	total amount of recurring insurance payments	float	
total_recurring_household	order	total amount of recurring household payments	float	
total_recurring_loan_payment	order	total amount of recurring loan payments	float	
total_recurring_leasing_amount	order	total count of recurring leasing payments	float	
total_recurring_insurance_amount	order	total count of recurring insurance payments	float	
total_recurring_household_amount	order	total count of recurring household payments	float	
total_recurring_loan_payment_amount	order	total count of recurring loan payments	float	
district_name	demographics	district name	string	
region	demographics	region	string	
population	demographics	population	integer	
mun_499	demographics	no. of municipalities with inhabitants < 499	integer	
mun_500_1999	demographics	no. of municipalities with inhabitants 500-1999	integer	

mun_2000_9999	demographics	no. of municipalities with inhabitants 2000-9999	integer	
mun_10000	demographics	no. of municipalities with inhabitants >10000	integer	
nb_cities	demographics	no. of cities	integer	
ratio_urb_pop	demographics	ratio of urban inhabitants	float	
avg_salary	demographics	average salary	integer	
unemployement_rate_95	demographics	unemploymant rate '95	float	
unemployement_rate_96	demographics	unemploymant rate '96	float	
A14	demographics	no. of enterpreneurs per 1000 inhabitants	integer	
nb_crimes_95	demographics	no. of commited crimes '95	float	
nb_crimes_96	demographics	no. of commited crimes '96	integer	
rate_crime_96	demographics	growth rate of crime from '95 to '96	float	
num_mun	demographics	total number of municipalities	integer	
loan_granted_1997	dependent variable	loan_granted 1997	integer	1 or 0
credit_issued_1997	dependent variable	credit_issued 1997	integer	1 or 0

REFERENCES:

- https://www.statology.org/pandas-rename-columns-with-dictionary/
- https://stackoverflow.com/questions/23668427/pandas-three-way-joining-multiple-dataframes-on-columns
- https://stackoverflow.com/questions/30222533/create-a-day-of-week-column-in-a-pandas-dataframe-using-python
- https://www.pythonprogramming.in/change-box-color-in-boxplot.html
- https://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.ttest_ind.html
- https://benalexkeen.com/comparative-statistics-in-python-using-scipy/
- https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots
- Class materials (jupyter notebooks and pdfs)