Financial Programming

Project : Financial Data Set

Group Number 06

Group Members :

Anais Mellas
Liam Scholl
Praveen Subramani

Dataset Overview

The dataset used in this project is from a real Czech bank in the period of 1993 to 1998. Included in the dataset are the tables district, account, client, card, disposition, loan, order, and trans.
The aim of the project is to use the common IDs located throughout the dataset to create a basetable combining all of the tables with created independent variables from the year of 1996 and dependent variables from 1997. This basetable is then used to find insights and useful information in the form of reporting and visualizations.

Data Preparation

For each table duplicates and missing values were checked

Account table:

- Convert 'date' column to string

Disp table:

- Renamed ‘type’ column for clarity
    - There is another ‘type’ column in another table so it is renamed to ‘account_type’

Order table:

- Replace values with ‘NOT_AVAILABLE’ if k_symbol has whitespace

Transaction table:

- Replace missing values
    - Add the string ‘Not available’ in the columns 'k_symbol’, ‘bank’, and ‘account’ for missing values
- Rename columns for clarity
    - Rename columns ‘amount’, ‘balance’, ‘bank’, ‘account’, ‘type’, and ‘operation’
- Impute missing values in 'trans_operation'
    - The missing values in the 'trans_operation' column are filled with the mode (most frequent value) of the same column
- Convert 'date' column to string
- Extract year, month, and day as separate columns
    - New columns 'trans_year', 'trans_month', and 'trans_day' are created based on substrings of the 'date' column.
- Combine day, month, and year to create 'DD-MM-YYYY' format
    - A new column 'trans_date' is created by combining 'trans_day', 'trans_month', and 'trans_year' with hyphen separators.
- Convert 'trans_date' column to date type with format '%d-%m-%Y'.

Loan table:

- Create new column ‘loan_granted_date’ using the ‘date’ column with format '%y%m%d'
- Rename columns for clarity
    - The columns ‘payments’, ‘amount’, ‘duration’, ‘payments’, and ‘status’
- Drop the columns ‘loan_id’ and ‘date’
- Filter the loan table to inlcude only loans granted in 1996

Credit card table:

- Change date column format to date type
- Extract year, month, and day as separate columns
    - New columns 'card_issued_year ', 'card_issued_month', and 'card_issued_day' are created based on substrings of the 'date' column.
- Filter the loan table to inlcude only cards issued in 1996

Demographic (district) table:

- Columns datatypes are converted and NaN values are filled
    - Columns ‘A12’ and ‘A15’ are changed to float and non-numeric values are replaced with 0.0
- Columns A1 to A16 are mapped to new column names and renamed
- Columns 'unemp_rate_95' and 'nbr_crimes_95' are dropped

Independent Variables

Account table:

- Extract year, month, and day as separate columns
    - New columns 'account_year', 'account_month', and 'account_day' are created based on substrings of the 'date' column.
    - 'account_year' is adjusted to the correct year format (assuming a two-digit year representation) and converted to an integer.
- Combine day, month, and year to create 'DD-MM-YYYY' format
    - A new column 'account_creation_date' is created by combining 'account_day', 'account_month', and 'account_year' with hyphen separators.
- Change 'account_creation_date' into date format:
    - 'account_creation_date' is converted to a datetime format using pd.to_datetime with the specified date format '%d-%m-%Y'.
- Add length of relationship ("LOR") column
    - Represents the length of the relationship by subtracting the 'account_year' from 1996.
- Rename 'district_id' to 'account_district_id'
- Convert 'account_district_id' from float64 to int64
- Change 'frequency' column name to 'account_freq_statement'
- Filter accounts created in 1996 or before

Client table:

- Create client birth year, month, and day variables using “birth_number”
    - 'client_birth_year' is created by concatenating '19' and converting the result to an integer.
    - 'client_birth_month' is created by extracting the characters at positions 2 to 4 in the 'birth_number' and converting them to an integer.
    - 'client_birth_day' is created by extracting the last two digits of the 'birth_number' and converting them to an integer.
- Extract gender from “birth_number”
    - 'client_gender' is initially set to 'M'.
    - For rows where 'client_birth_month' is greater than 50, indicating female, 'client_gender' is set to 'F'.
- Correct birth month for females
    - For rows where 'client_birth_month' is greater than 50, the month is corrected by subtracting 50.

- Create age and age group variable
    - 'client_age' is created by subtracting 'client_birth_year' from 1996.
    - 'client_age_group' is created by dividing 'client_age' by 10, flooring the result, and multiplying by 10.

Order table:

- Create pivot_k_symbol_order variable
    - Pivot the order table based on ‘k_symbol’ and ‘account_id’
    - Add ‘account_id’ column by resetting index
    - Add new column for sum of amounts per account called ‘k_symbol_SUM’
    - Add prefix 'order_' to every column
    - Rename 'order_account_id' to 'account_id'
- Create pivot_bank_to_order variable
    - Pivot the order table based on ‘bank_to and ‘account_id’
    - Add ‘account_id’ column by resetting index
    - Add prefix 'order_' to every column
    - Rename ‘bank_to’ to ‘order_bank_to’
    - Rename 'order_account_id' to 'account_id'

Transaction table:

- Create trans_96 variable
    - Selects transactions in 1996
- Create trans96_agg_credit variable
    - Aggregates credit transactions for each account ID
- Create trans96_avg_credit variable
    - Aggregates average credit transactions for each account ID
- Create trans96_agg_withdrawal variable
    - Aggregates withdrawals for each account ID
- Create trans96_avg_withdrawal variable
    - Aggregates average withdrawals for each account ID
- Create trans96_agg_total_spend variable
    - Aggregates the total spent for each account ID
- Create trans96_total_spend_flag variable
    - Flag variable for total spend greater than or equal to 0
- Create earliest_balance_trans96 variable
    - Displays the earliest balance for each account ID in the year 1996
- Create latest_balance_trans96 variable
    - Displays the latest balance for each account ID in the year 1996
- Create trans96_frequency variable
    - Calculates the frequency of transactions in the year 1996
- Convert all series to DataFrames for easy merging
- Merge the aggregated data onto the unique account ids
    - Create trans_iv table

Dependent variables:

- Create a copy of loan called loan_DV to extract ‘account_id’ and ‘loan_granted_date’
    - Create new column ‘loan_granted_97’ and set it to 1 if the loan was granted in 1997, else set to 0
- Create a copy of card table called card_DV with where year issued in 1997
    - Merge card table with disp table (merged_card_DV_Disp) to check if accounts have been issued a card in 1997
    - Create new column ‘credit_card_issued97’ and set it to 1 if the card was issued in 1997, else set to 0

Basetable

Basetable part 1:

Once the tables were cleaned and new variables were added, the first part of the basetable was created by first merging the two loan tables of loan_iv and loan_DV containing the new variables. Next, the two pivoted order tables are merged. Then, the tables loan, order, and trans were merged in that order on account_id, and district was merged on district_id to create BaseTable_part1.

Basetable part 2:

For the second part of the basetable, the two card tables of card_iv and card_dv with updated variables are combined to create a final card table. Then, the tables disp, client, and card are merged on client_id and disp_id in that order. The basetable is then filtered to only have account owners.

Final basetable:

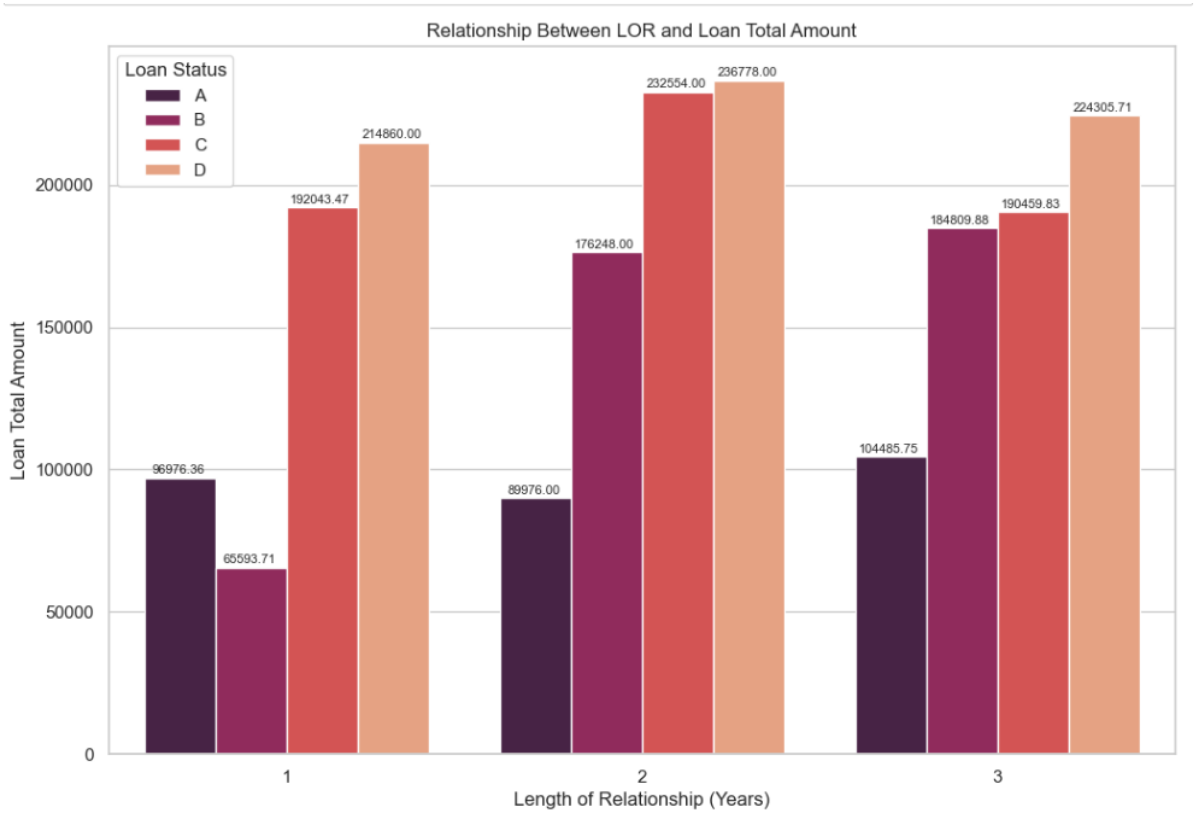
The final basetable is created by merging the two basetables BaseTable_part1 and BaseTable_part2_Filtered on account_id using an inner join. The columns district_id_x, date_x, date_y, loan_granted_date_y, district_id_y, and birth_number are dropped. The final basetable is then sorted by account_id, NaN values are filled with 0 and 9999, and outliers are removed.

Variable Name	*Data Type*	*Source Table*	*Description*
account_id	int64	account	Unique identifier for account
account_freq_statement	object	account	Frequency of statement issuance
account_year	int32	account	Year of account creation
account_month	object	account	Month of account creation
account_day	object	account	Day of account creation
account_creation_date	datetime64[ns]	account	Full date of account creation
account_LOR	int32	account	Length of relationship with client
loan_id	float64	loan	Unique identifier for loan
loan_total_amount	float64	loan	Total loan amount
loan_duration	float64	loan	Duration of loan in months
loan_monthly_payments	float64	loan	Monthly payments on the loan
loan_status	object	loan	Status of paying off the loan ('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)
loan_granted_date	datetime64[ns]	loan	Date that loan is granted
loan_granted_97	float64	loan	Client had granted loan in the 1997, binary value (0 = did not have granted loan, 1 = had granted loan)
order_LEASING	float64	order	Leasing payment
order_NOT_AVAILABLE	float64	order	No type of payment
order_POJISTNE	float64	order	Insurance payment
order_SIPO	float64	order	Household payment
order_UVER	float64	order	Loan payment
order_k_symbol_SUM	float64	order	Sum of payments
order_AB	float64	order	Bank partner
order_CD	float64	order	Bank partner
order_EF	float64	order	Bank partner
order_GH	float64	order	Bank partner
order_IJ	float64	order	Bank partner
order_KL	float64	order	Bank partner
order_MN	float64	order	Bank partner
order_OP	float64	order	Bank partner

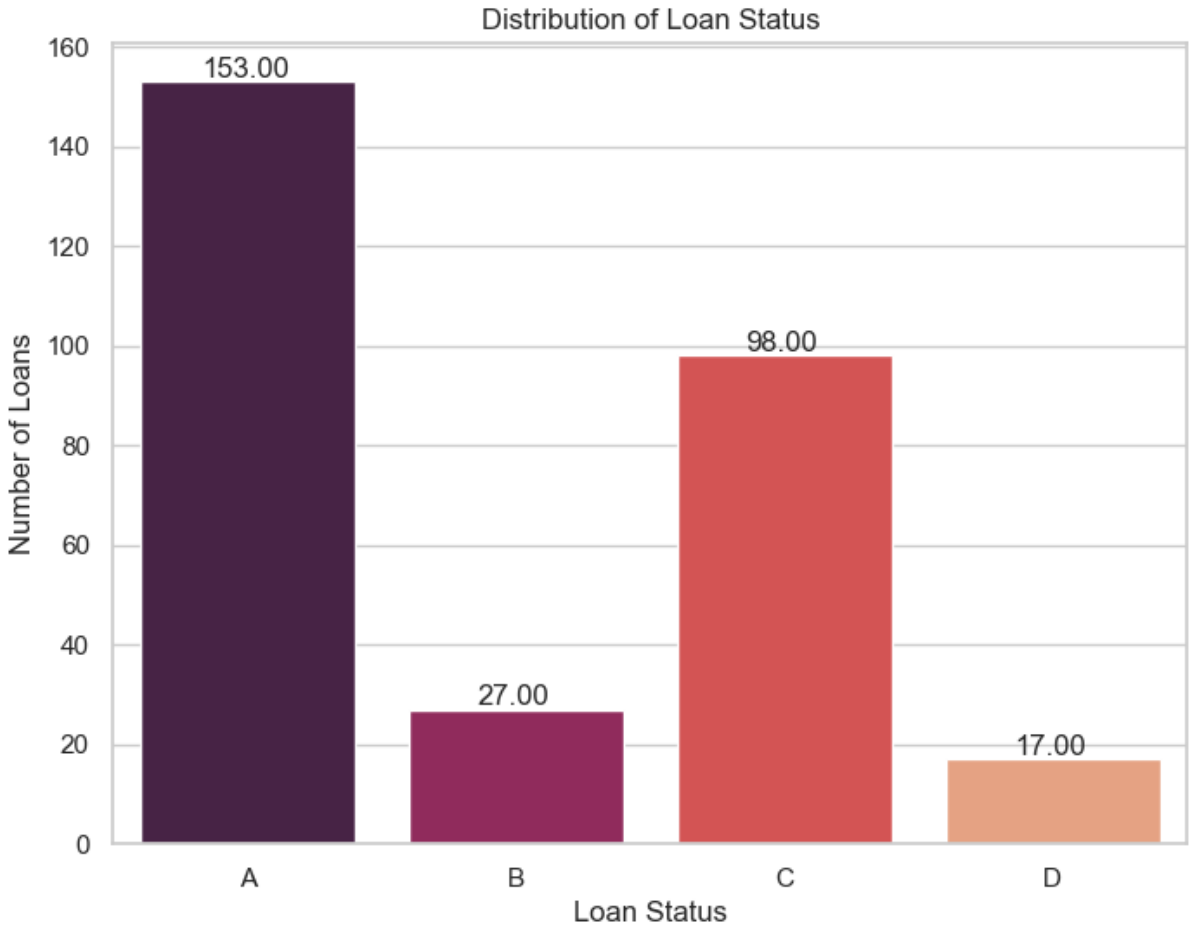
order_QR	float64	order	Bank partner
order_ST	float64	order	Bank partner
order_UV	float64	order	Bank partner
order_WX	float64	order	Bank partner
order_YZ	float64	order	Bank partner
trans_total_credit	float64	trans	Total credit transactions per account
trans_avg_credit	float64	trans	Average credit transactions per account
trans_total_withdrawal	float64	trans	Total withdrawn from account
trans_avg_withdrawal	float64	trans	Average withdrawn from account
trans_total_spend	float64	trans	Total spent per account
flag_spend	object	trans	Flag variable for total spend (1 = Positive spend, 0 = Negative spend)
trans_first_balance96	float64	trans	First account balance in 1996
trans_latest_balance96	float64	trans	Last account balance in 1996
trans_frequency_96	int64	trans	Frequency of transactions in 1996
district_name	object	district	Name of district
region_name	object	district	Name of region
nbr_inhab	int64	district	Number of inhabitants
nbr_muni_inhab_499	int64	district	Number of municipalities with inhabitants less than 499
nbr_muni_inhab_500_1999	int64	district	Number of municipalities with inhabitants between 500 and 1999
nbr_muni_inhab_2000_9999	int64	district	Number of municipalities with inhabitants between 2000 and 9999
nbr_muni_inhab_10000	int64	district	Number of municipalities with inhabitants greater than 10000
nbr_cities	int64	district	Number of cities
ratio_urban_inhab	float64	district	Ratio of urban inhabitants
avg_salary	int64	district	Average salary
unemp_rate_95	float64	district	Unemployment rate in 1995
unemp_rate_96	float64	district	Unemployment rate in 1996
nbr_entrepreneur_inhab	int64	district	Number of entrepreneurs per 1000 inhabitants
nbr_crimes_95	int64	district	Number of crimes in 1995
nbr_crimes_96	int64	district	Number of crimes in 1996

disp_id	int64	card	Unique identifier for disposition
client_id	int64	client	Unique identifier for client
account_type	object	disp	Type of disposition (owner or user)
client_birth_year	int32	client	Client's birth year
client_birth_day	int32	client	Client's birth day
client_birth_month	int32	client	Client's birth month
client_gender	object	client	Client's gender
client_age	int32	client	Client's age
client_age_group	int32	client	Client's age group
card_id	float64	card	Unique identifier for card
type	object	card	Card type
issued	object	card	Card issuance date
card_issued_year	float64	card	Card issuance year
card_issued_month	float64	card	Card issuance month
card_issued_day	float64	card	Card issuance day
credit_card_issued97	float64	card	Flag variable for card issuance in 1997 (1 = issued, 0 = not issued)

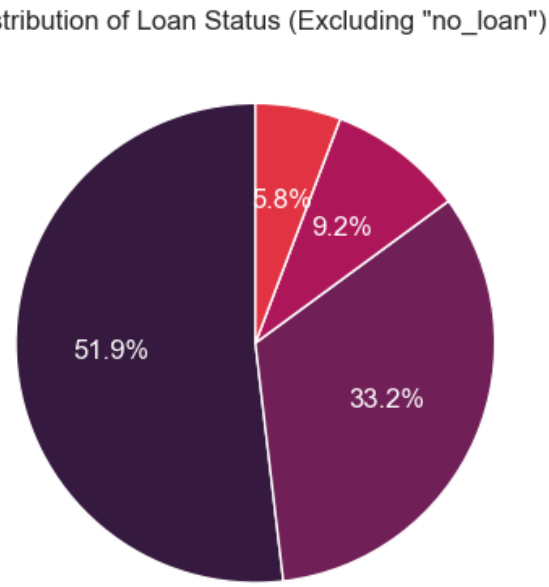
Visualizations for Independent Variables



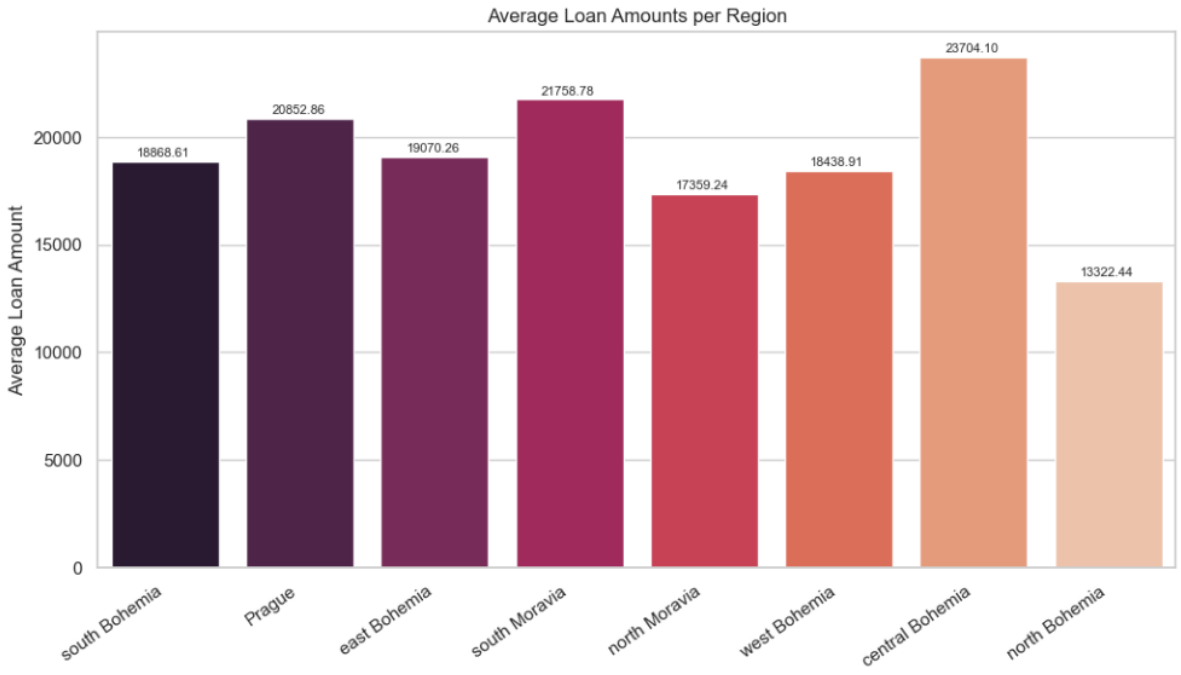
The chart is a clustered bar graph showing the relationship between the length of relationship (LOR) with a lending institution and the total amount of loans. The trends indicate that the total loan amount tends to increase with the length of the relationship for loan statuses A (contract finished and paid) and D (running contract, client in debt). For B (contract finished and not paid) and C (running contract, OK so far), the amounts fluctuate over the years. This shows that clients in running contracts who have debt are usually the ones with the highest loan amounts and have a high LOR as well.



The bar chart displays the distribution of loan statuses, with four categories labeled A, B, C, and D. Status A has the highest number of loans at 153, indicating it is the most common status. Status C follows with 98 loans, while B and D have significantly fewer loans, with 27 and 17 loans respectively. This distribution suggests that statuses A and C are the most prevalent loan conditions, whereas B and D are less common.

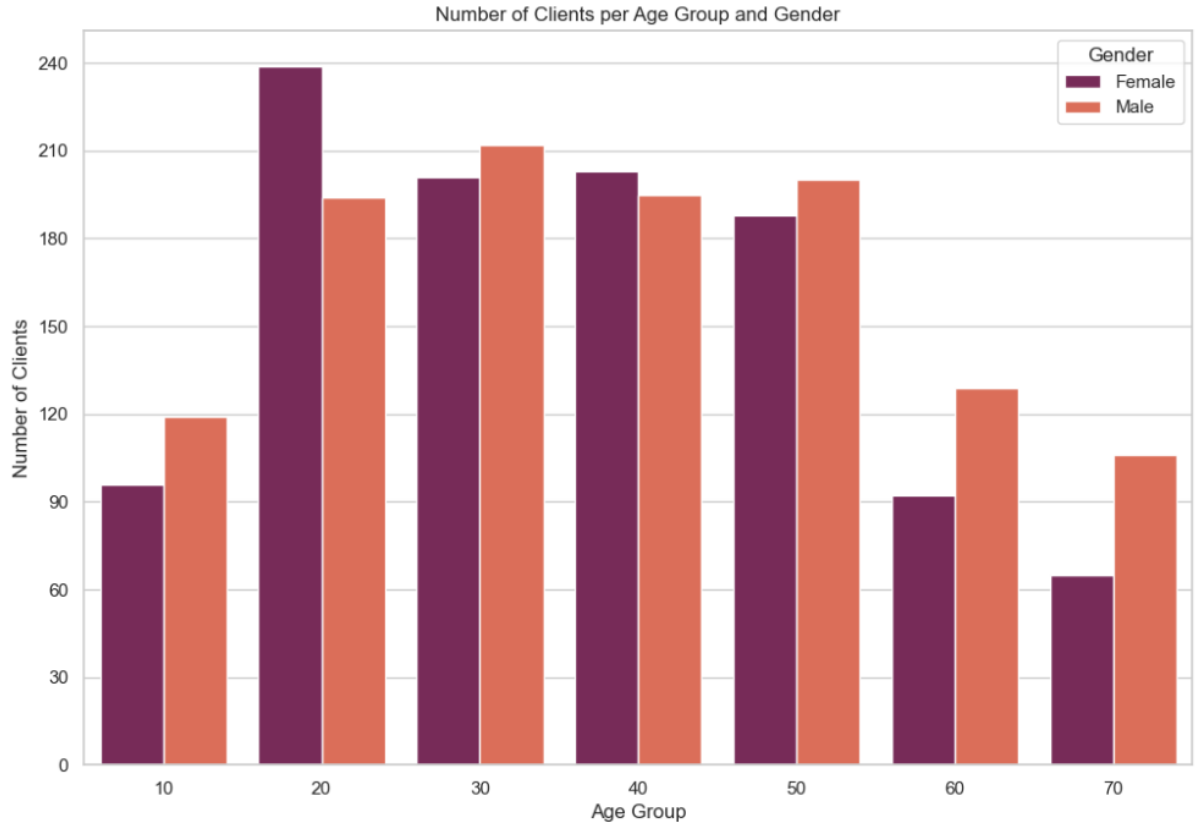


The pie chart shows the proportion of various loan statuses, with the category "no_loan" excluded. The largest segment, at 51.9%, represents the most common status among the loans. The next sizeable segment is 33.2%, followed by a smaller segment at 9.2%. The smallest segment is at 5.8%. This visualization indicates a majority status for more than half of the loans, with the remaining statuses making up less than half of the loans combined.

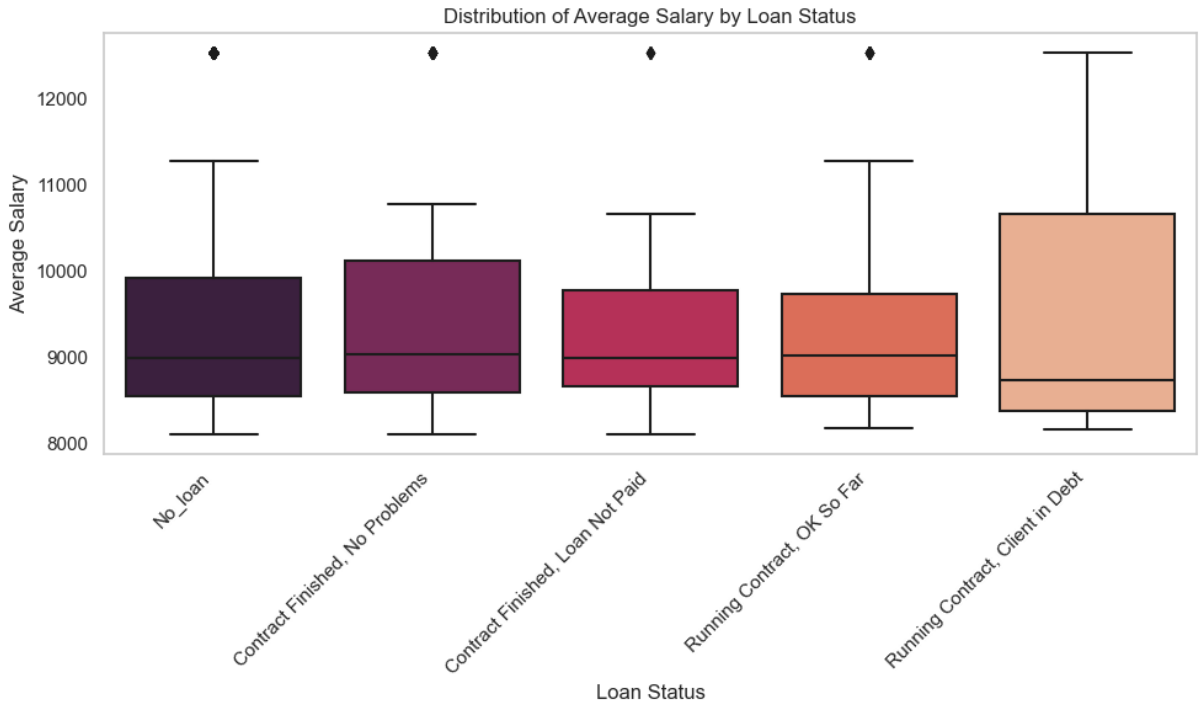


The bar chart presents average loan amounts per region. The highest average loan amount is in Prague at approximately 23,704 units, followed by Central Bohemia with around 20,853 units. South Moravia has the third highest average at roughly 21,759 units. The lowest averages are in North Bohemia at about 13,322 units and North Moravia at approximately 17,359 units. This indicates a

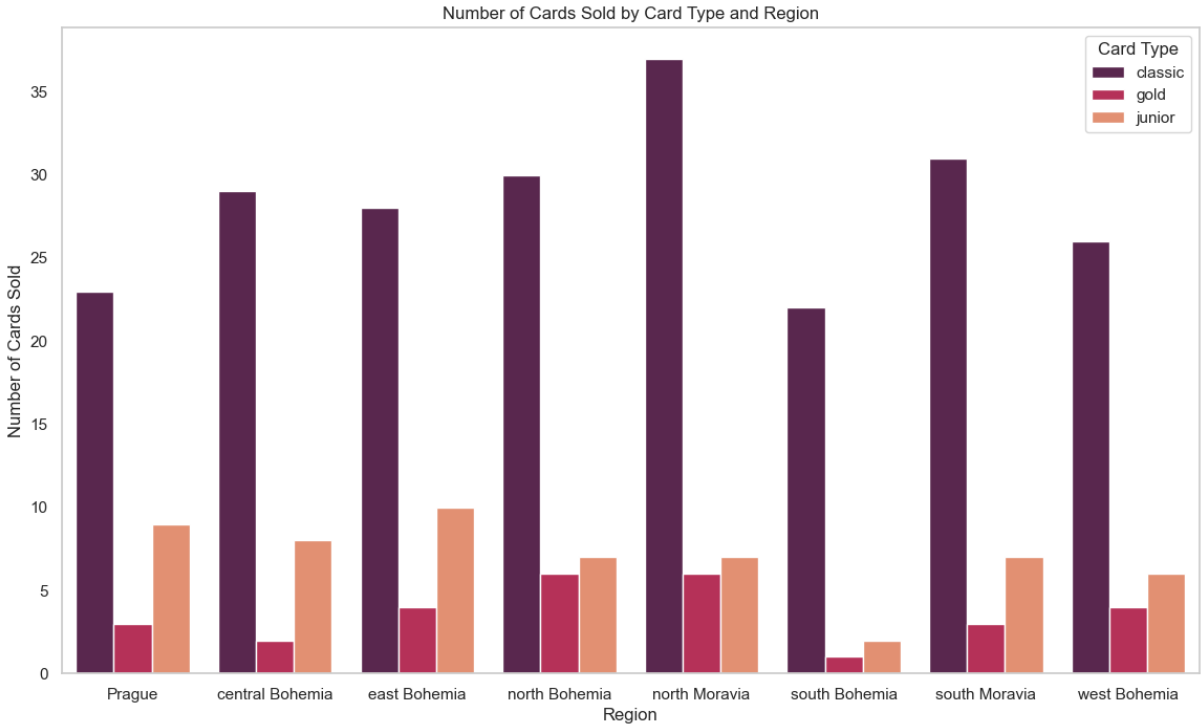
regional disparity in average loan amounts, with Prague and Central Bohemia seeing higher averages than other regions.



This is a clustered bar chart showing the number of clients per age group and gender. It shows that many of the clients are between the age groups 20s and 50s. The distribution of clients for each gender is relatively equal other than 20s, 60s, and 70s. They have more females in the 20s age group, while 60s and 70s have more males in the group.

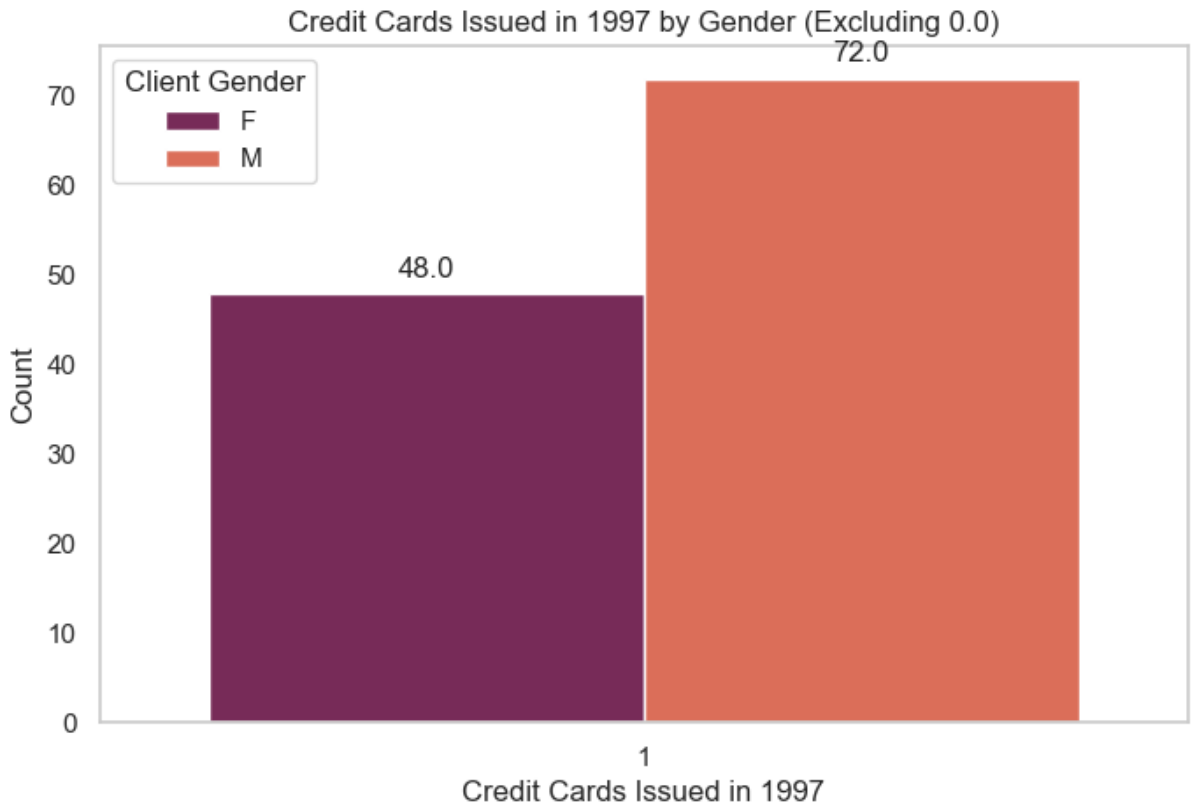


The boxplot shows that individuals with no loans or finished loans without problems have similar median salaries, while those currently in debt have higher salaries with greater variability. Unpaid finished loans correspond to the lowest median salaries, and successfully managed ongoing loans tend to have higher median salaries, with outliers present in all but the unpaid loan category.

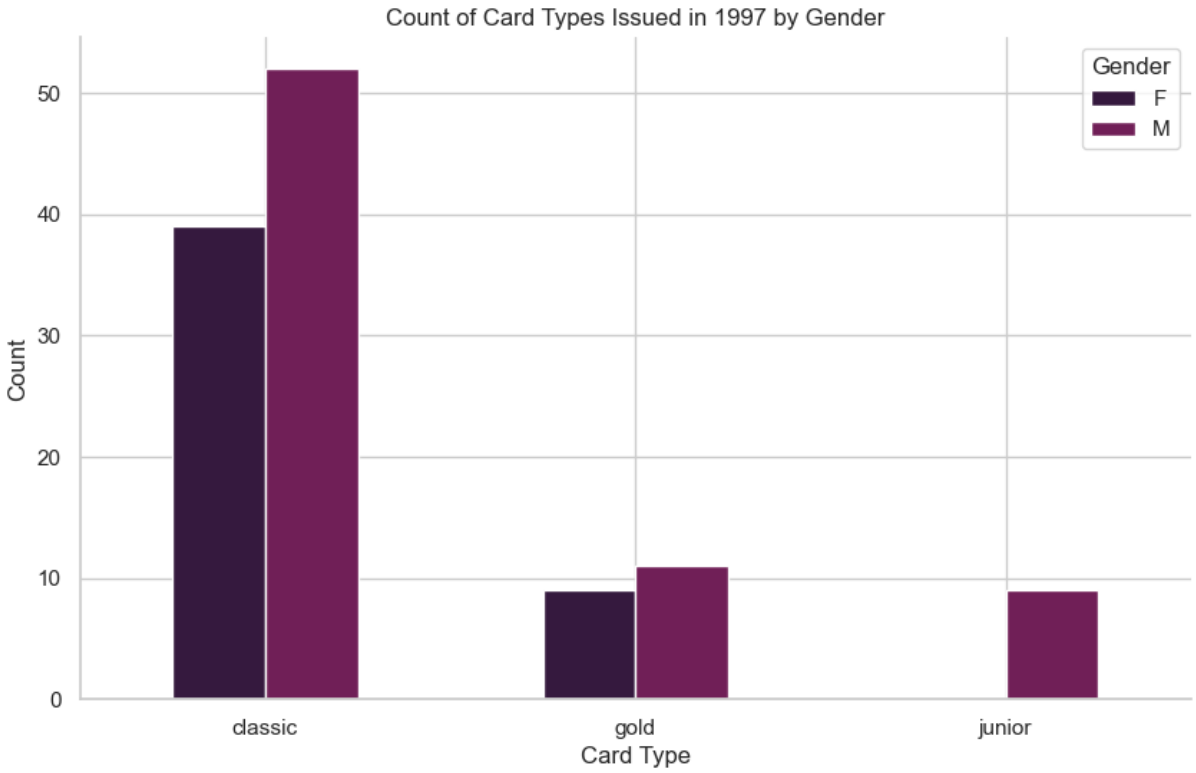


The bar chart depicts the number of credit cards sold by card type and region. Classic cards are the most sold across all regions, with South Moravia having the highest sales. Gold cards are less common, with the highest sales in Prague. Junior cards have the fewest sales, with relatively low numbers across all regions. The data indicates regional differences in the popularity of card types.

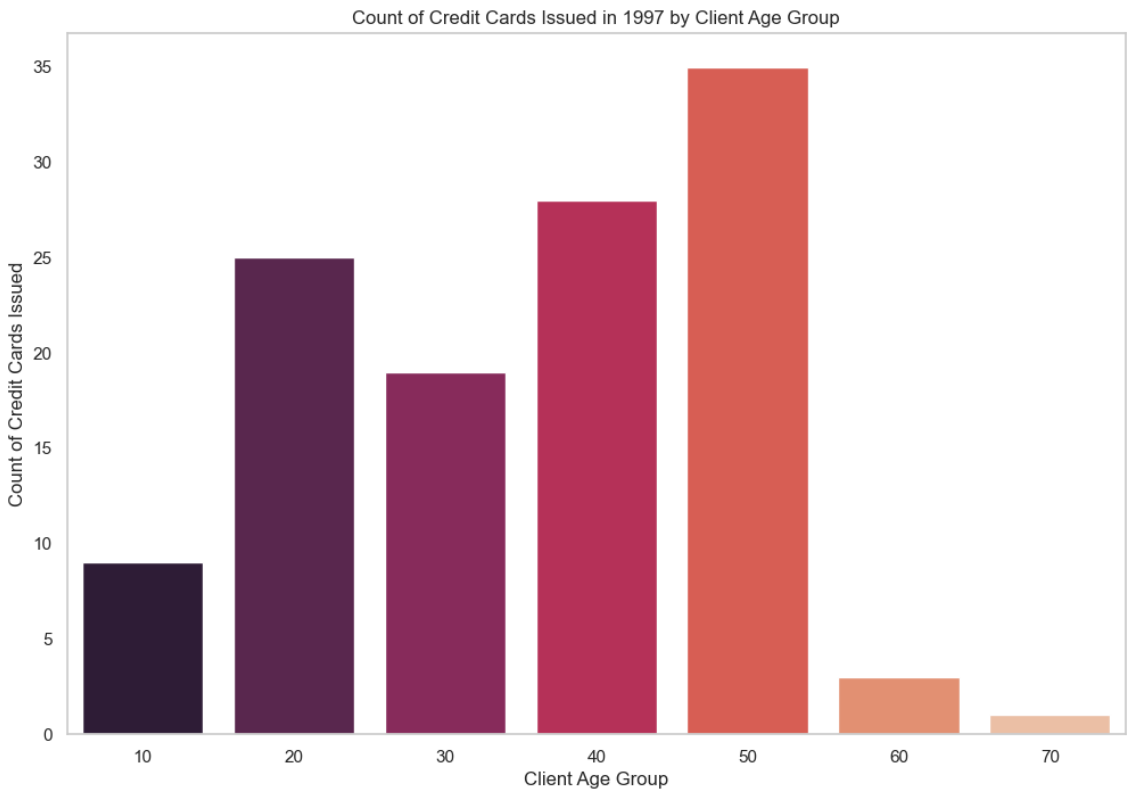
Visualizations for Dependent Variables



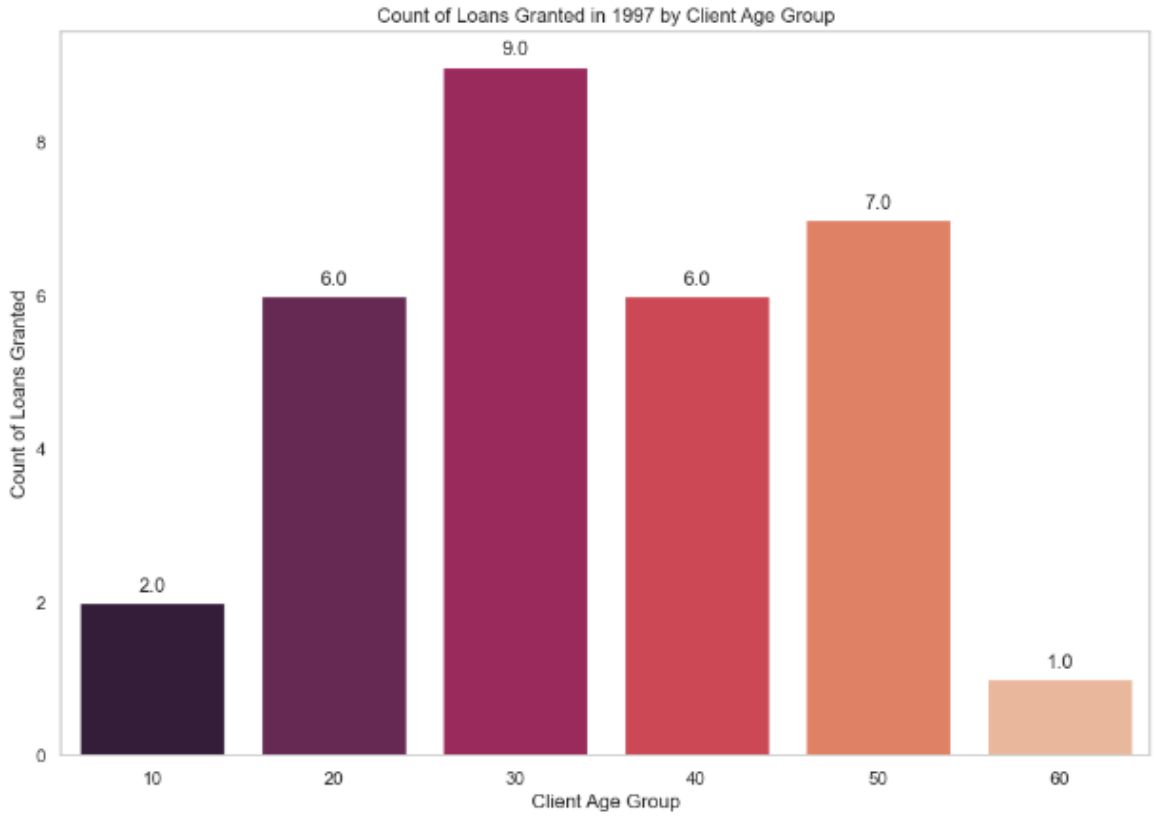
The bar chart displays the number of credit cards issued in 1997 by gender, with the data for those who did not receive a card ('0') excluded. It shows that 72 credit cards were issued to males (M) and 48 to females (F), indicating that more males received credit cards than females in that year. This could reflect the credit card issuance policy or the demand for credit cards among different genders at that time.



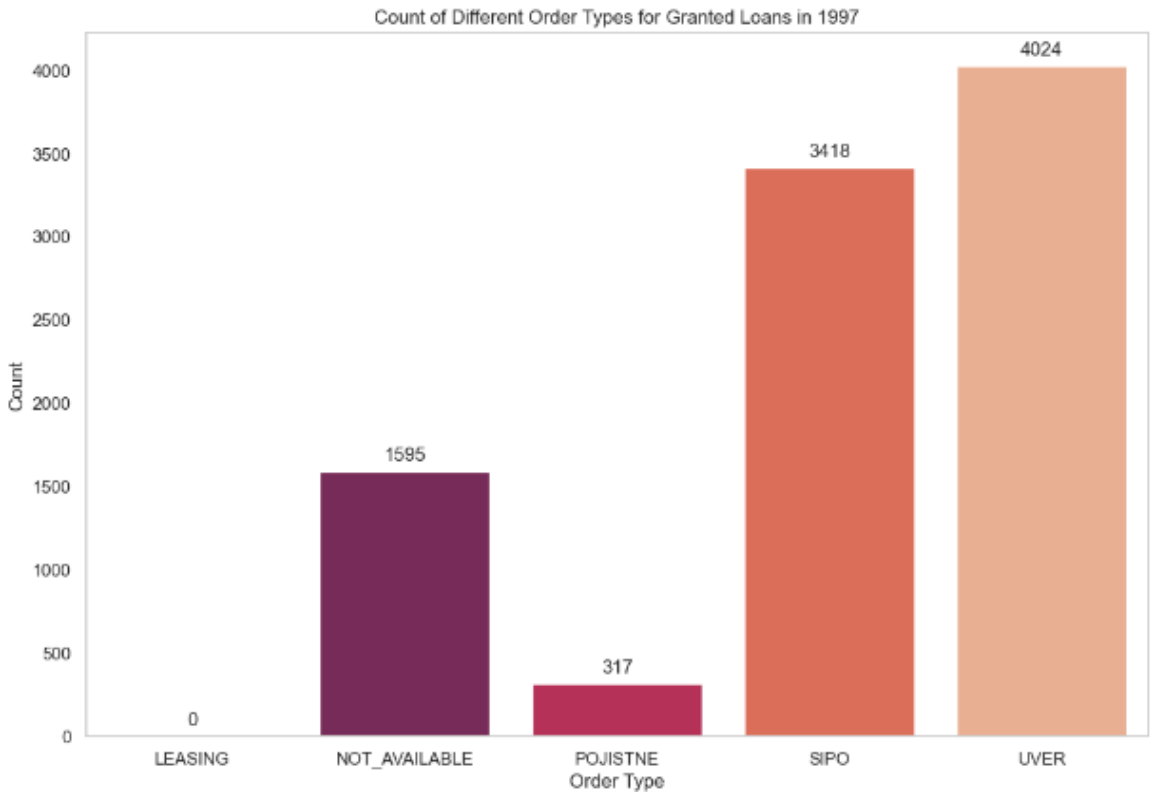
In 1997, classic credit cards were the most issued, with women receiving more than men. Gold cards, likely a premium option, were issued more to men. Junior cards saw the least issuance, distributed almost evenly between both genders.



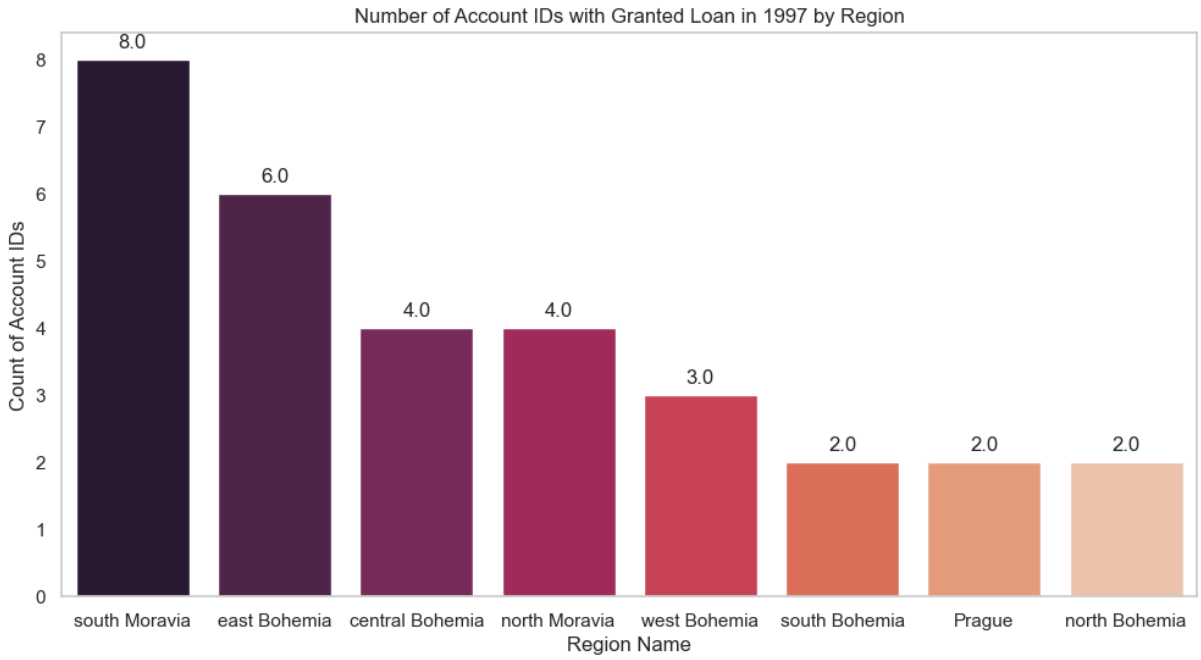
The bar chart shows credit card issuance peaking with clients in their 50s in 1997, tapering off for those in their 60s and significantly lower for the 70s age group. Clients in their 20s and 30s received a moderate number of cards, suggesting that credit card companies might have targeted middle-aged customers preferentially.



In 1997, clients in their 30s were granted the most loans, while those in their 20s and 40s received an equal, slightly lower number. Loan issuance dropped for clients in their 50s and was least for those in their 60s.



The bar chart shows the count of different order types for loans granted in 1997. 'UVER' was the most common order type with 4024 counts, followed by 'SIPO' with 3418. 'LEASING' had 1595 occurrences, and 'POJISTNE' had the fewest with only 317. There's also a category labeled 'NOT_AVAILABLE' indicating some data might be missing or not classified.



The bar chart depicts the number of account IDs with loans granted in 1997 by region. South Moravia had the highest count with 8 loans, followed by East Bohemia with 6. Central and North Moravia both had 4 loans each. West Bohemia, South Bohemia, North Bohemia, and Prague each had a count of 2 loans granted. This suggests a regional variation in loan distribution, with South Moravia leading.