FINANCIAL PROGRAMMING

Financial Dataset Analysis

About

Financial Dataset Analysis and Creation of Base-table with Dependent and Independent Variable Using Analysis and Visualization

Members:

RAMANNA Nithesh
NGUYEN Thi Quynh Anh
NITHARWAL Ashwani

SUMMARY

The purpose of this report was the analysis of one financial institution. With multiple datasets provided by the organization from 1993 to 1998, we aimed to perform a critical evaluation of the business in several aspects including but not limited to the influence of demographics and socioeconomic status on the issuance of credit card or loan acceptance. To achieve this goal, we utilized python to perform data cleaning and transformation to generate useful insights that could further be used for predictive analytics. Correlations between various parameters were explored to determine influence factors on the issuance of credit or loan grant. This analysis would lead us to get an overview of how a financial organization worked and how client profiles were evaluated based on their spending, balance and usage pattern and some other factors.

METHODOLOGY

Eight separate datasets tables were provided with customer information including age, gender, geographical locations, household income, loans, dispositions, credit cards, and transactions.

These datasets included a lot of varied details about spending rationale, spending patterns, loan status, issuance frequency, card types, mode of transactions, bank's tendency to transact, the region where customers opened an account, etc.

In this report, several data manipulation techniques were employed using python to transform and correct the data. Below was the list of data and manipulation techniques we have utilized.

- birth_number was used to extract the birthday, month, year and hence birth date.
- birth_number was used to extract the gender of the client. If the birth month was greater than 50 then gender was noted as Female.
- Age_group was calculated by dividing the age by 10, followed by the multiplication of resulting integer value by 10.
- Irrelevant and redundant dataset columns were deleted.
- Data was subset based on the required timeline of 1996 - 1997.
- Values in code or non-understandable language were replaced with custom words for readability and understanding.
- A few of the columns were renamed to know what the column values were related to.
- Data tables were pivoted as per the customization required.
- Columns and datasets were grouped to get the better insights or the broader sense.
- Data was filtered to get only the data required.
- Data columns or values were aggregated to get the pattern or sometimes also to summarize the data.
- Datasets were merged to get the base-table variables.
- Missing values were filled with appropriate values such as 0 or any aggregated value.
- Data columns were sorted by their values to perform further operations.
- Data was visualized using the charts and plots to get quick insights and gain pattern knowledge.

BASE-TABLE AND VARIABLES

Variable	**Data Type**	**Value**	**Description**
client_id	int	1,2,…	Identifier for client
district_id	int	1,2,…	Identifier for district
gender	string	M,F	Gender values for Male, Female
age	int	1,2,…	Age of clients (Years)
age_group	int	10,20,..	Age rounded to lowest 10th number
birth_date	date	2/4/1945	Birth date of the client
disp_id	int	1,2,…	Identifier to find the relation between client and its disposition for the same account
account_id	int	1,2,…	Identifier for account
district name	string	Pribram	Name of the district
region	string	Prague	Name of the region
inhabitants	int	70699	Number of inhabitants living in a district
municipalities with inhabitants < 499	Int	1,2,…	number of municipalities with inhabitants < 499
municipalities with inhabitants 500-1999	Int	1,2,…	number of municipalities with inhabitants between 500-1999

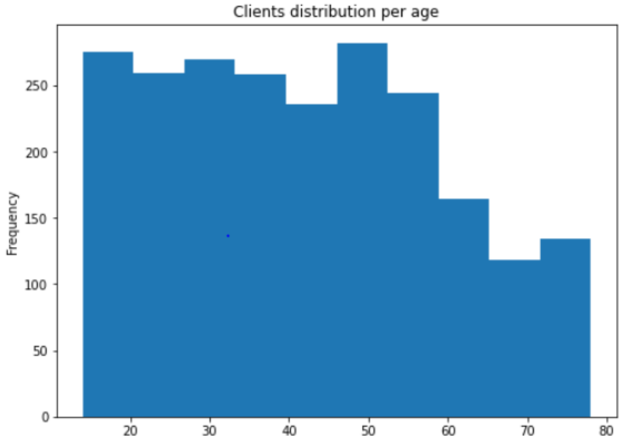
municipalities with inhabitants 2000-9999	Int	1,2,…	number of municipalities with inhabitants between 2000-9999
municipalities with inhabitants >10000	Int	1,2,…	number of municipalities with inhabitants > 10000
Cities	Int	1,2,…	Number of cities
ratio of urban inhabitants	Float	1.25	ratio of urban inhabitants
average salary	Int	1,2,…	Average salary
unemployment rate 1996	Float	1.25	unemployment rate 1996
entrepreneurs per 1000 inhabitants	Int	1,2,…	Number of entrepreneurs per 1000 inhabitants
commited crimes 1996	Int	1,2,…	Number of committed crimes in a district
frequency	String	Monthly issuance	Frequency of issuance of statements
LOR	Int	1,2,…	Length of relationship
Average Collection from another bank	Float	1.25	Average credit coming from another bank to the client bank
Average Credit in Cash	Float	1.25	Average cash credit in the client’s account

Average Interest Credited	Float	1.25	Average Interest credited in the client’s account
Average remittance to another bank	Float	1.25	Average fund remittance debit from the client’s account
Average withdrawal in cash	Float	1.25	Average withdrawal in cash from client’s account
Average Total Credit	Float	1.25	Average total credit by month in client’s account
Average Total Debit	Float	1.25	Average total debit by month from client’s account
Average Monthly Savings	Float	1.25	Monthly (Average total credit – average total debit)
Average Monthly Balance	Float	1.25	Monthly average balance in year 1996 (mean last day balance of every month)
Median Monthly Balance	Float	1.25	Median monthly balance in year 1996 (median last day balance of every month)
total_credit	Float	1.25	Total amount credited in year 1996 (all credits with all modes of transactions)
total_withdrawal	Float	1.25	Total amount withdrawals in year 1996 (all debits with all modes of transactions)
household payment	Float	1.25	sum of transaction payments amounts for household purpose

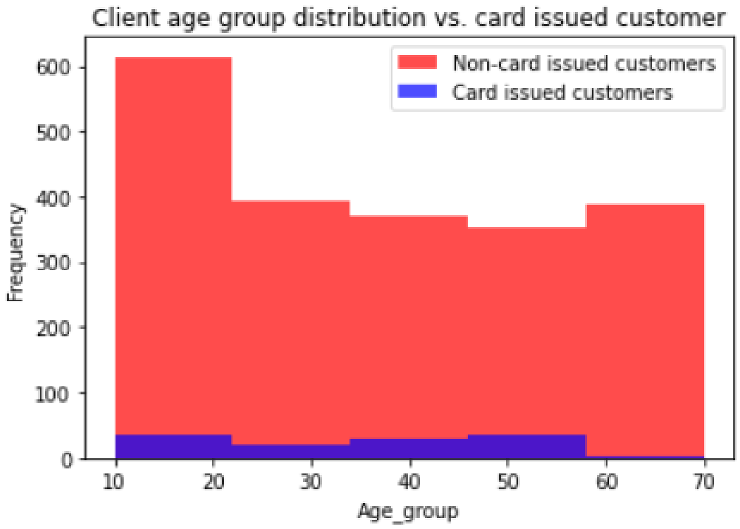
insurance payment	Int	1,2,…	sum of transaction payments amounts for insurance purpose
Lease	Float	1.25	sum of transaction payments amounts for lease purpose
loan payment	Float	1.25	sum of transaction payments amount for loan monthly payments purpose
Other	Int	1,2,…	sum of transaction payments amount for other purpose leaving the above column reasons
total order payment	Float	1.25	Sum of all order payments by each client
Card Type	string	Gold	Type of card a client posses
Card Issued before 1997	string	Yes	To show that the client had the card already issued before 1997
Loan Amount	Int	1,2,…	Loan amount for the clients having loans before year 1997
Loan Duration	Int	1,2,…	Duration(Months) of loans of clients having loans before year 1997
Loan Payments	Int	1,2,…	Monthly repayment (EMI) amount assigned to the clients having loans before year 1997
Status	string	Ok so far	Loan status of client loan before year 1997
Loan granted before 1997	string	Yes	To show that the client had the loan granted before 1997

Card Issued	Int	1 or 0	Target Variable #1: To show that the client can get the card issued in year 1997
loan granted	Int	1 or 0	Target Variable #2: To show that the client can get the loan granted in year 1997

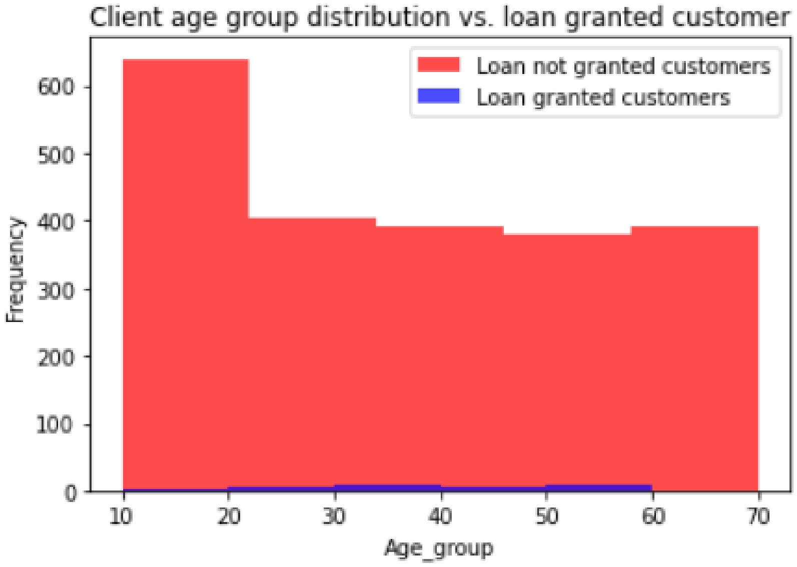
INSIGHTS REPORTING AND VISUALIZATION FOR BASETABLE VARIABLES



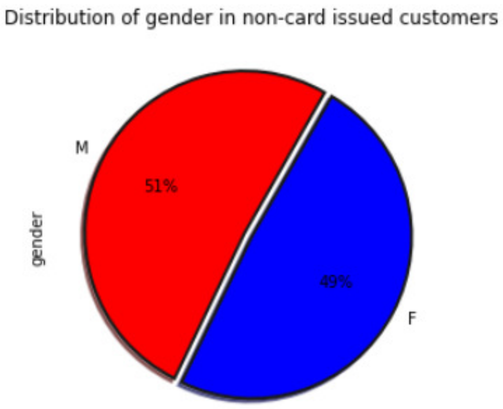
The histogram illustrated the distribution of clients according to their age. The chart adopted multimodal distribution with many peaks close together and a skewed distribution on the left. The X-axis represented the age of customers while the Y-axis represented the frequency. From this histogram, it showed that the majority of the clients fell into the age range of 20 to 60 and the number of clients declined with old age. To be specific, the frequency of clients below age 60 was about 250 while those with age 60 and above are less in number and nearly half if we looked closely at the frequency of clients.



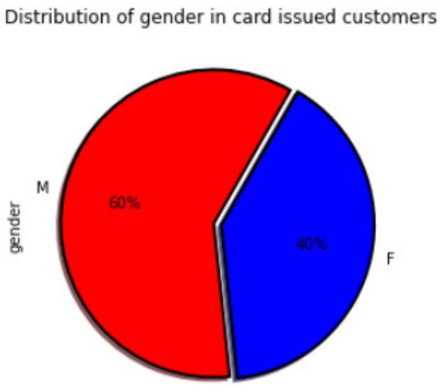
The histogram illustrated how age groups contributed to the credit card issue status. The blue bar represented the distribution of age groups among those who were approved for credit card issuance, while the red bar showed age group distribution for those who were not. The majority of clients were not approved for credit card issuance. Specifically, we could infer that clients from all age groups had the eligibility to file for card issuance but those aged 60 and 70 had the least probability of getting approved. Therefore, age group did influence the card issue status as presented by the declined probability associated with old age.



The histogram illustrated how age groups contributed to the loan grant status. The blue bar represented the distribution of age group among those who were approved for loan grant vs red bar for those who were not. Similar to the card issuance, only a small proportion of people had their loan granted. In fact, loan grant status was highest among clients in the age group of 30 - 40 or 50 - 60. Clients belonging to the age group of 60 - 70 had the least probability of getting the loan granted. In fact, no loan was approved for clients above the age of 70. This finding indicated that age group played a factor in loan grant status, especially for clients above age 60.

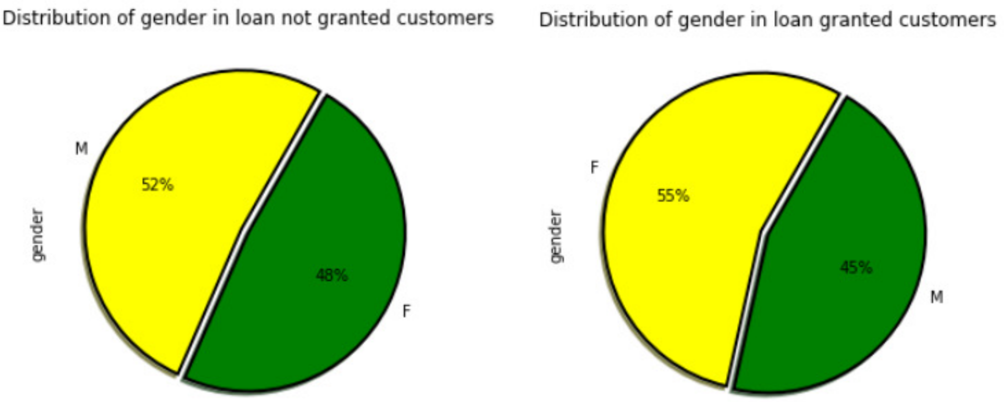


Distribution of gender in non-card issued customers

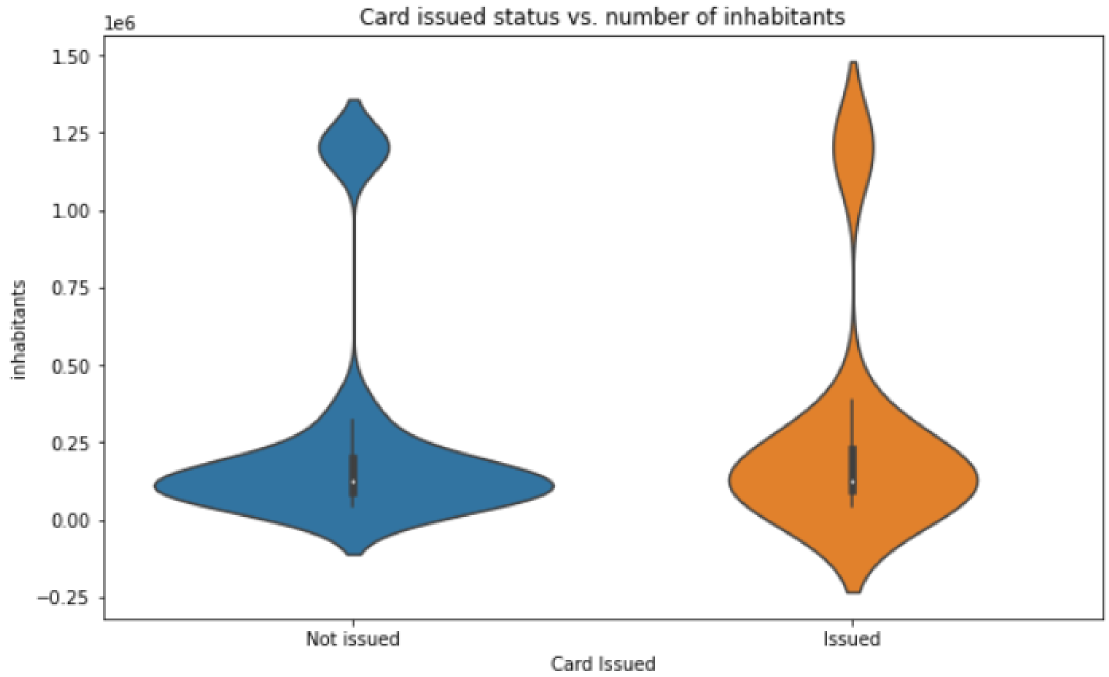


Distribution of gender in card issued customers

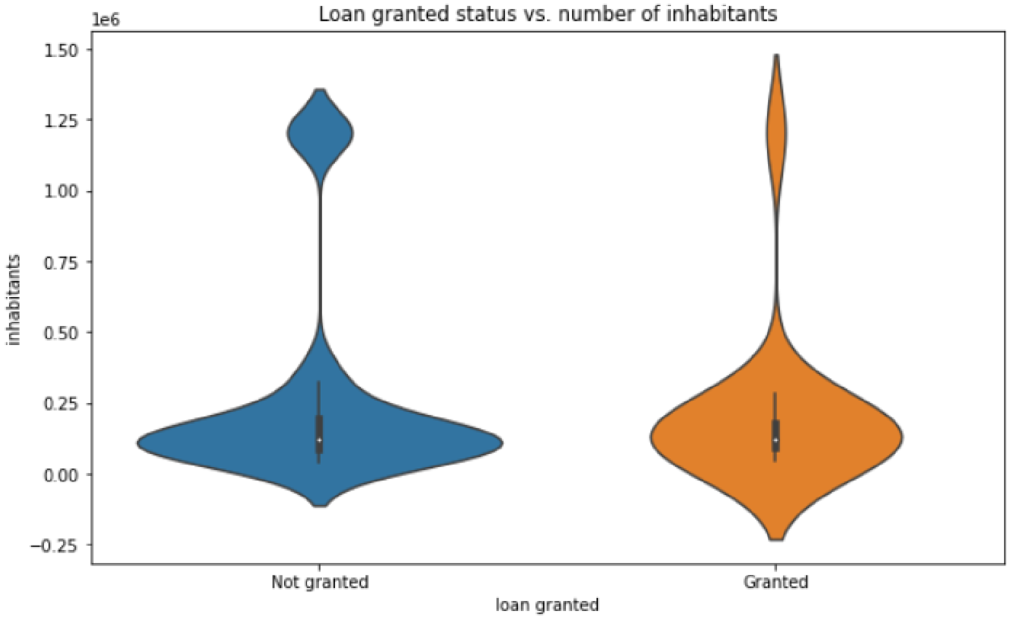
The two pie charts illustrated how gender contributed to the credit card issue status. The slices represented the distribution of gender among those who were approved for credit cards vs those who were not approved. In general, clients were distributed nearly independent of gender as they accounted for equal proportion regardless of credit card issue status. To be specific, for non-card issued customers, male accounted for 51% while females accounted for 49%. For card issued customers, females made up for 40% and male made up for 60%. From the above pie chart, it was clear that there was no relationship between the investigated variable (gender) and the target variable (credit card issue status) though it seems like card issuance is biased towards males.



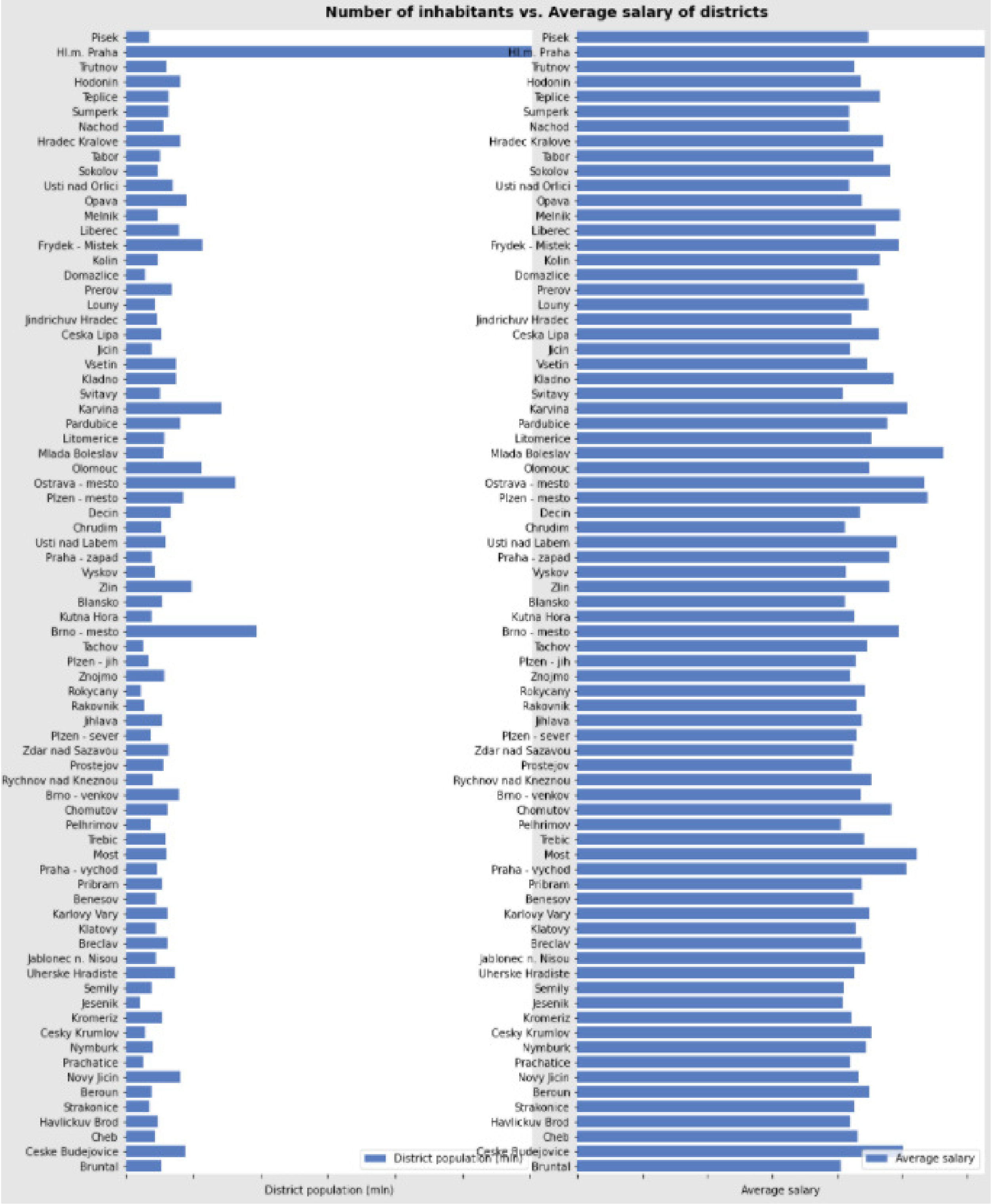
The two pie charts illustrated how gender contributed to the loan grant status. The slices represented the distribution of gender among those who were approved for grant of loan vs those who were not. In general, clients were distributed nearly independent of gender as they accounted for equal proportion regardless of loan grant status. To be specific, for no loan granted customers, male accounted for 52% while females accounted for 48%. For loan granted customers, females made up for 55% and male made up for 45%. From the above pie chart, it was clear that there was no relationship between the investigated variable (gender) and the target variable (loan grant status).



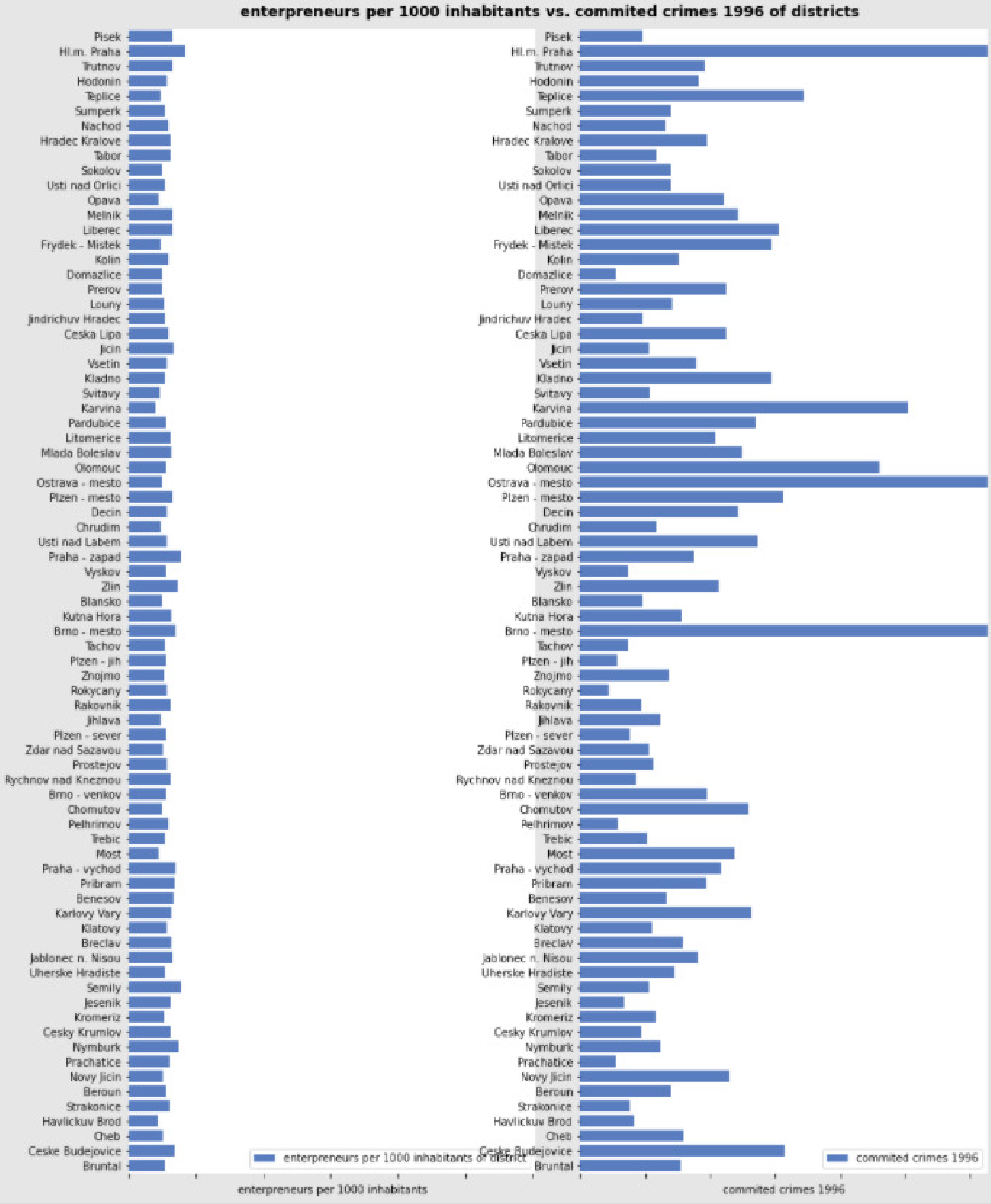
The violin plots illustrated how the number of inhabitants influenced the credit card issue status. The plots showed the distribution of the number of inhabitants among those who were approved for credit cards vs those who were not approved. In the comparison between the two plots in terms of the center, spread, and distribution of the data sets, we could see that both datasets had similar overall shape and distribution of the tips with the median and quartiles very close to each other. Graphically, both the plots had two modes with the wider sections representing higher probability and thinner sections corresponding to lower probability. Therefore, we could safely assume that the number of inhabitants did not influence the credit card issuance.



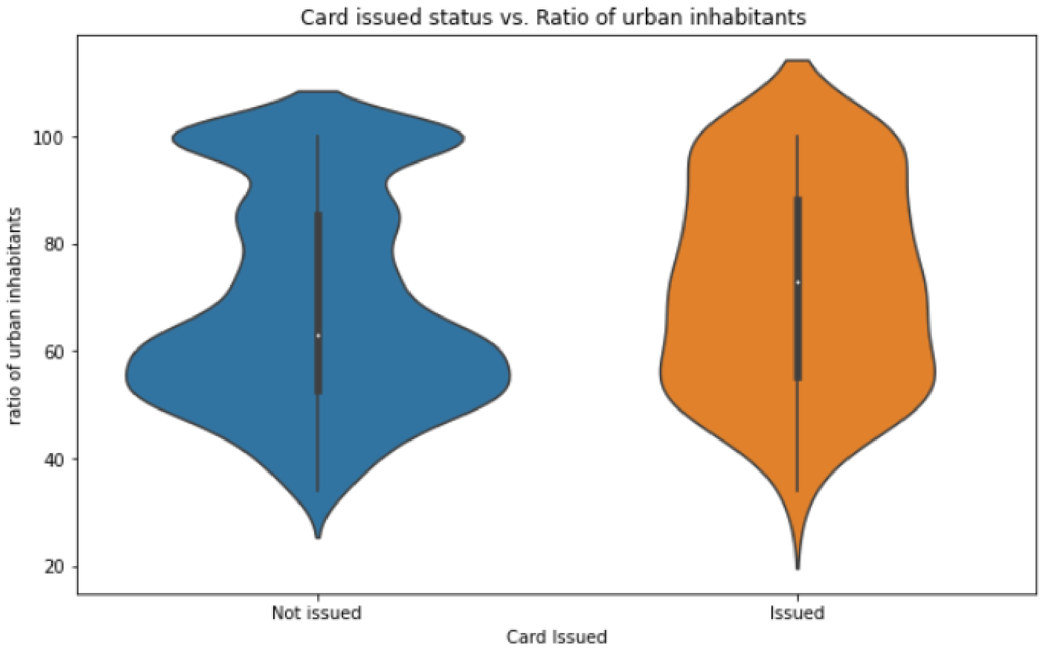
The violin plots showed how the amount of residents affected the loan grant status. The plots showed the population distribution between those who were authorized for a loan grant and those who were not. Both plots had two modes which indicated two distinct subpopulations with the wider sections corresponding to higher probability. The median and quartile range of both plots were relatively similar, suggesting that the investigated parameter had little effect on the loan approval.


Number of inhabitants vs. Average salary of districts
Pisek
Hl.m. Praha
Trutnov
Hodonin
Teplice
Sumperk
Nachod
Hradec Kralove
Tabor
Sokolov
Usti nad Orlici
Opava
Melnik
Liberec
Frydek - Mistek
Kolin
Domazlice
Prerov
Louny
Jindrichuv Hradec
Ceska Lipa
Jicin
Vsetin
Kladno
Svitavy
Karvina
Pardubice
Litomerice
Mlada Boleslav
Olomouc
Ostrava - mesto
Plzen - mesto
Decin
Chrudim
Usti nad Labem
Praha - zapad
Vyskov
Zlin
Blansko
Kutna Hora
Brno - mesto
Tachov
Plzen - jih
Znojmo
Rokycany
Rakovnik
Jihlava
Plzen - sever
Zdar nad Sazavou
Prostejov
Rychnov nad Kneznou
Brno - venkov
Chomutov
Pelhrimov
Trebic
Most
Praha - vychod
Pribram
Benesov
Karlovy Vary
Klatovy
Breclav
Jablonec n. Nisou
Uherske Hradiste
Semily
Jesenik
Kromeriz
Cesky Krumlov
Nymburk
Prachatice
Novy Jicin
Beroun
Strakonice
Havlickuv Brod
Cheb
Ceske Budejovice
Bruntal
District population (mln)
District population (mln)
Average salary
Average salary

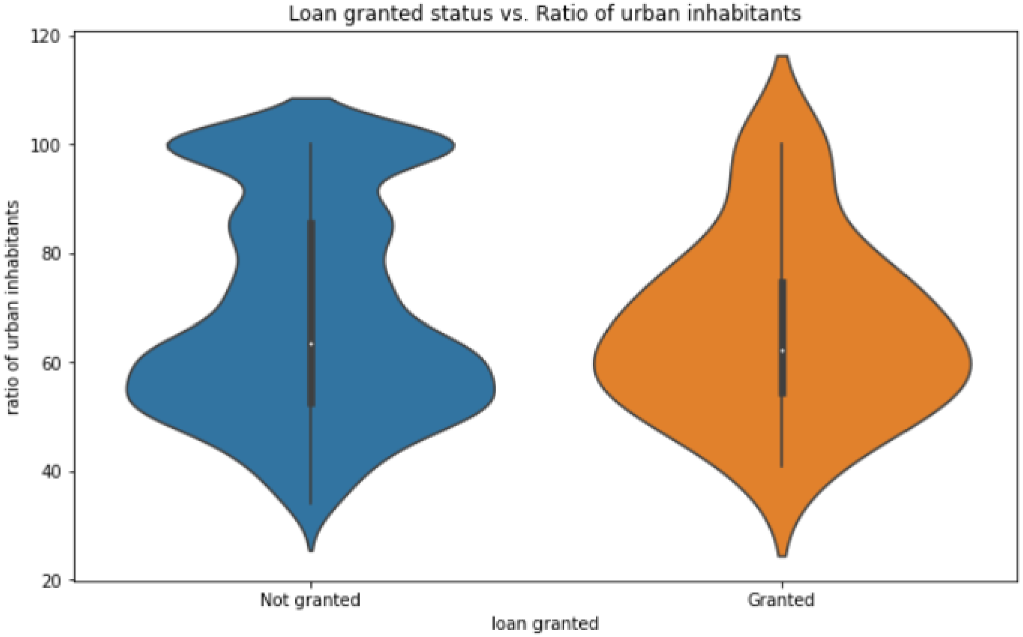
The bar charts showed the comparison between the number of inhabitants in a district and average salaries of clients in that district. Graphically, most of the districts had nearly comparable salaries but “Hl.M.Praha” had the highest average salary. Coincidentally, the number of inhabitants in the same district is highest among all the districts. However, compared to the 3-fold difference in the number of inhabitants with the second highest “Brno-mesto", the increase in the average salary was less than 1.5 fold. On the other hand, other districts had less number of inhabitants but the amount of salary was not that differentiated. The findings indicated that there was no relation between the number of inhabitants and the average salary.


enterpreneurs per 1000 inhabitants vs. commited crimes 1996 of districts
Pisek
Hl.m. Praha
Trutnov
Hodonin
Teplice
Sumperk
Nachod
Hradec Kralove
Tabor
Sokolov
Usti nad Orlici
Opava
Melnik
Liberec
Frydek - Mistek
Kolin
Domazlice
Prerov
Louny
Jindrichuv Hradec
Ceska Lipa
Jicin
Vsetin
Kladno
Svitavy
Karvina
Pardubice
Litomerice
Mlada Boleslav
Olomouc
Ostrava - mesto
Plzen - mesto
Decin
Chrudim
Usti nad Labem
Praha - zapad
Vyskov
Zlin
Blansko
Kutna Hora
Brno - mesto
Tachov
Plzen - jih
Znojmo
Rokycany
Rakovnik
Jihlava
Plzen - sever
Zdar nad Sazavou
Prostejov
Rychnov nad Kneznou
Brno - venkov
Chomutov
Pelhrimov
Trebic
Most
Praha - vychod
Pribram
Benesov
Karlovy Vary
Klatovy
Breclav
Jablonec n. Nisou
Uherske Hradiste
Semily
Jesenik
Kromeriz
Cesky Krumlov
Nymburk
Prachatice
Novy Jicin
Beroun
Strakonice
Havlickuv Brod
Cheb
Ceske Budejovice
Bruntal
enterpreneurs per 1000 inhabitants of district
enterpreneurs per 1000 inhabitants
Pisek
Hl.m. Praha
Trutnov
Hodonin
Teplice
Sumperk
Nachod
Hradec Kralove
Tabor
Sokolov
Usti nad Orlici
Opava
Melnik
Liberec
Frydek - Mistek
Kolin
Domazlice
Prerov
Louny
Jindrichuv Hradec
Ceska Lipa
Jicin
Vsetin
Kladno
Svitavy
Karvina
Pardubice
Litomerice
Mlada Boleslav
Olomouc
Ostrava - mesto
Plzen - mesto
Decin
Chrudim
Usti nad Labem
Praha - zapad
Vyskov
Zlin
Blansko
Kutna Hora
Brno - mesto
Tachov
Plzen - jih
Znojmo
Rokycany
Rakovnik
Jihlava
Plzen - sever
Zdar nad Sazavou
Prostejov
Rychnov nad Kneznou
Brno - venkov
Chomutov
Pelhrimov
Trebic
Most
Praha - vychod
Pribram
Benesov
Karlovy Vary
Klatovy
Breclav
Jablonec n. Nisou
Uherske Hradiste
Semily
Jesenik
Kromeriz
Cesky Krumlov
Nymburk
Prachatice
Novy Jicin
Beroun
Strakonice
Havlickuv Brod
Cheb
Ceske Budejovice
Bruntal
commited crimes 1996
commited crimes 1996

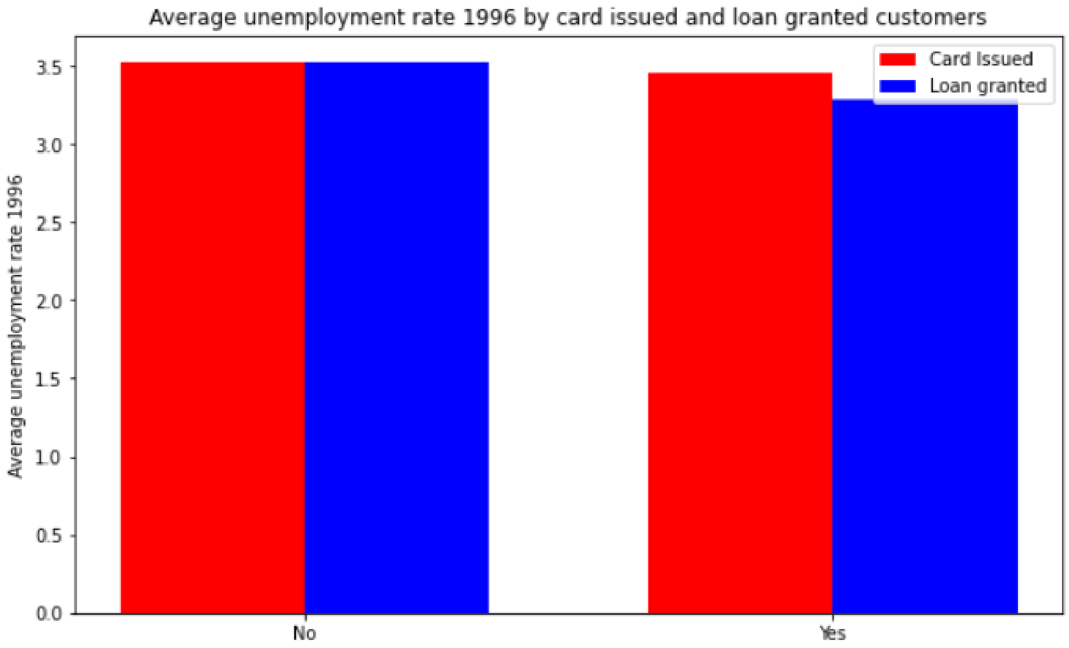
The graph showed the number of entrepreneurs per 1000 inhabitants and the number of crimes committed in 1996 in various districts. Graphically, most of the districts had a comparatively similar number of entrepreneurs in a district but the numbers of crimes committed showed random and different values. In fact, “Hl.M.Praha”, “Brno-mesto", and “Ostrava-mesto" experienced the highest committed crimes in 1996 as compared to “Rokycany" with the lowest number of committed crimes. As the number of entrepreneurs did not affect the probability of crimes committed, it could be said that the two factors were independent of each other.



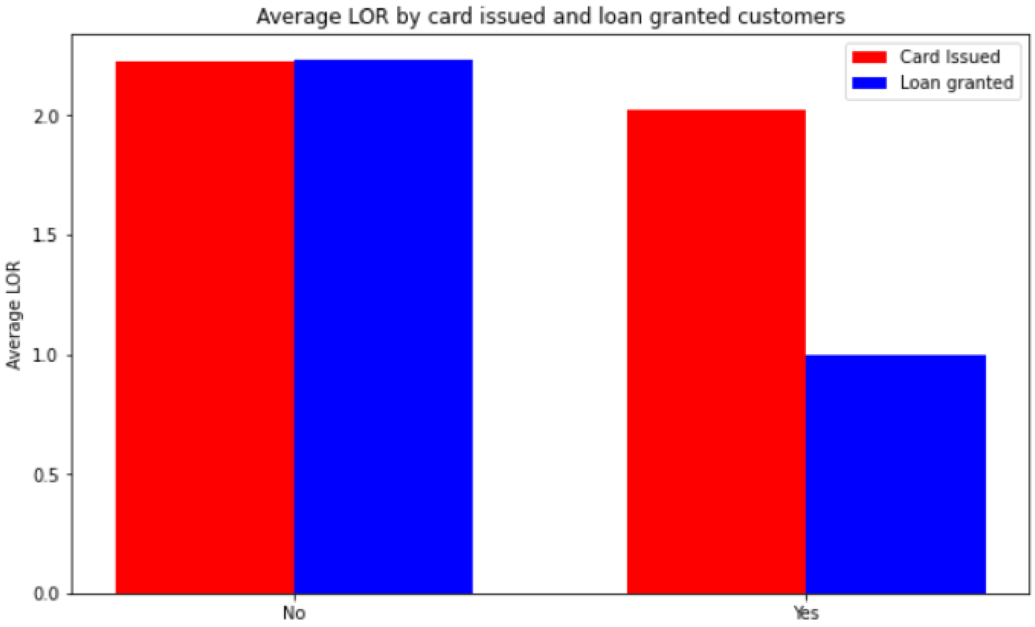
The violin plots illustrated how the ratio of urban inhabitants affected card issuance. The plots depicted the ratio of urban inhabitants’ distribution between those who were authorized for card issuance and those who were not. The plot for clients approved for card issuance adopted a more uniform shape with well distributed range as compared to the plot for clients with cards not approved. For the not-issued dataset, slightly lower probability was observed at the ratio of urban inhabitants ranging from 70 to 90 while the dataset for issued cards was distributed evenly in all ratios. Median value of the non-issued plot was lower than that of the issued plot with the ratio of urban inhabitants of about 60 as compared to 70. Therefore, there might be a relation between the ratio of urban inhabitants and card issuance status.



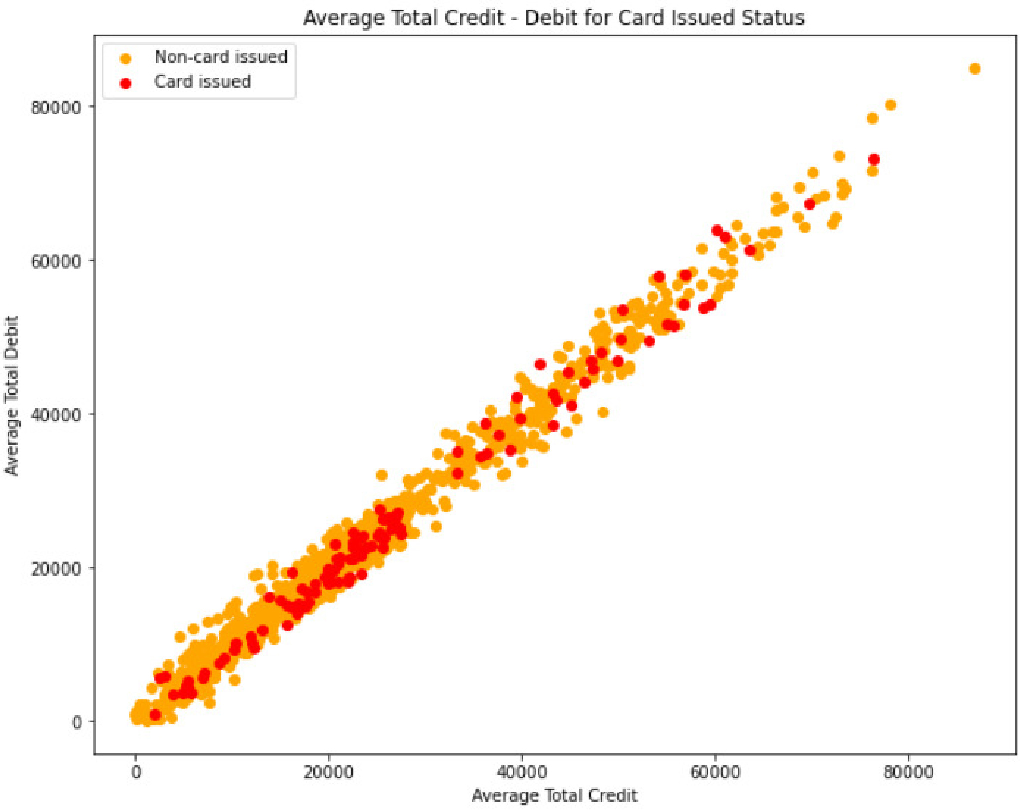
The violin plots illustrated how the ratio of urban inhabitants affected the loan granted status. The plots showed the ratio of urban inhabitants' distribution between those who were authorized for loan grant and those who were not. Graphically, both plots adopted the different shapes, but both datasets showed high probability in the range of 40 - 60. The median value of the two plots were also comparable in the 60s. Hence, there was no relation between loan grant status and number of urban inhabitants.



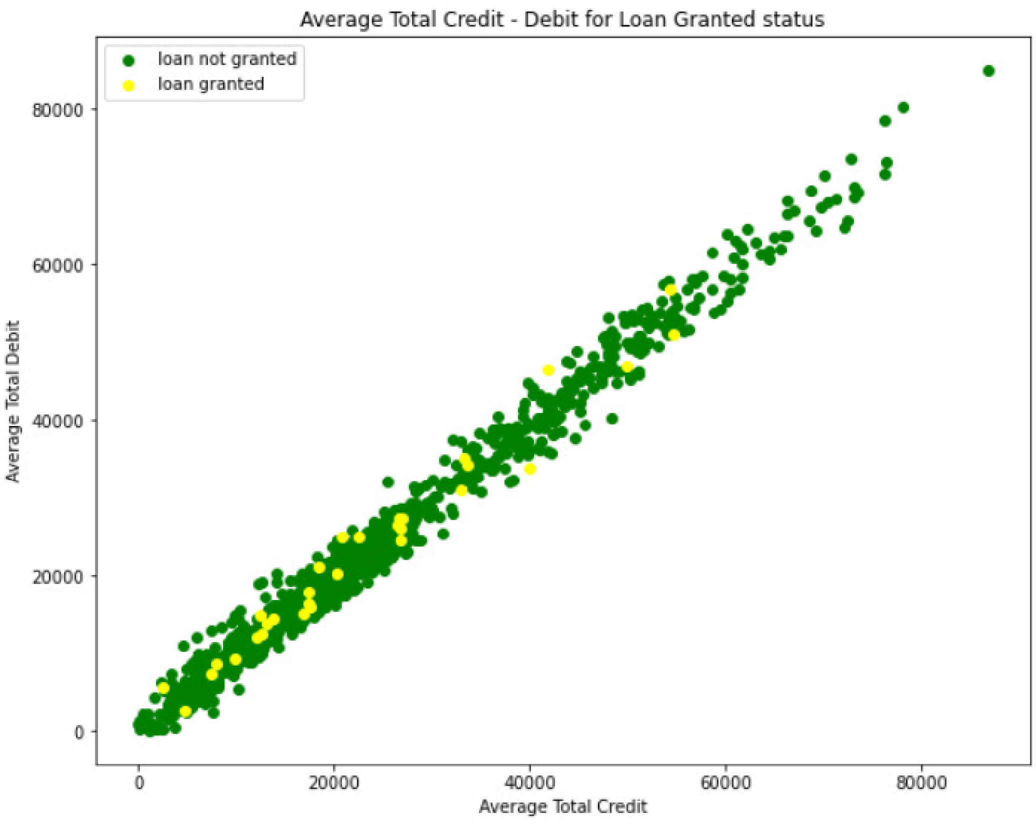
The bar plots presented the comparison between card issued and the loan granted clients based on the average unemployment rate in the year 1996. The red bars showed the card issue status and the blue bars showed the loan grant status. For the X-axis, Yes indicated the approval while No indicated rejection. No difference in the average unemployment rate in 1996 was observed between clients who had or did not have cards issued. Similar findings were observed for grant loan status as the average unemployment rate in 1996 remained comparatively similar despite acceptance or rejection status of loan. Therefore, we could say that there was no specific relation between the unemployment rate of 1996 and the issuance of cards and loan grants.



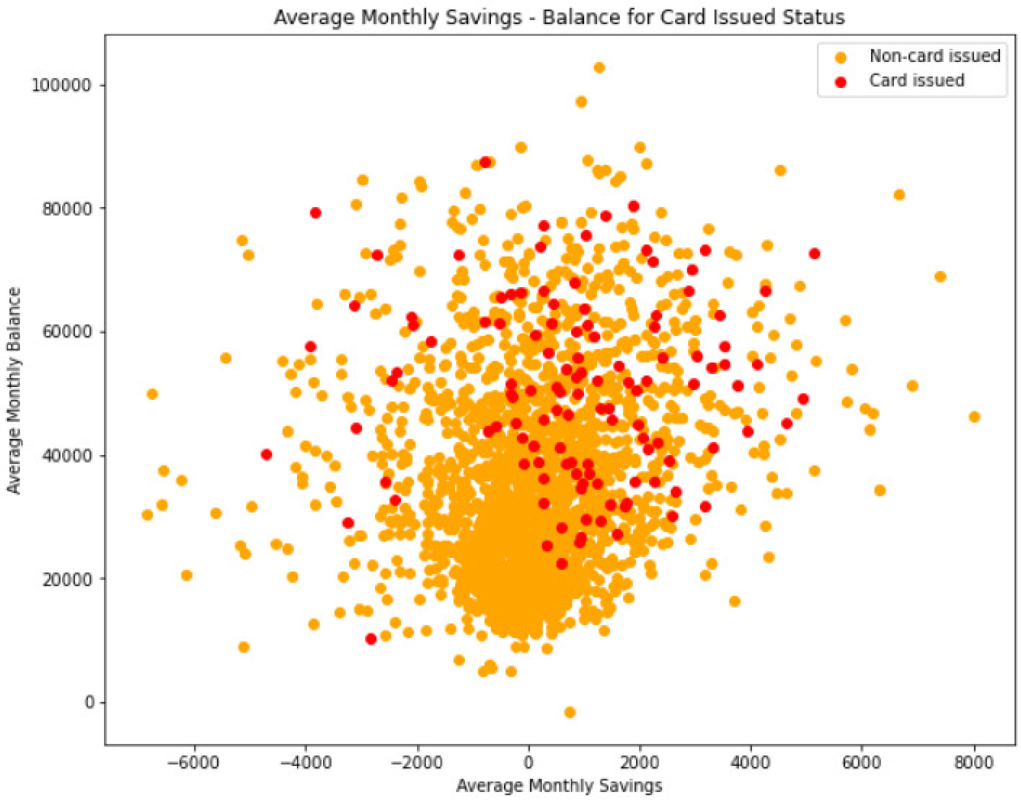
The bar plots showed the comparison between card issued and the loan granted clients based on the average length of relationship (LOR). The red bars showed the card issue status and the blue bars showed loan grant status. Graphically, for card issue status, clients with approved cards had a shorter LOR as compared to those who got rejected. Similar phenomenon was observed for loan grant status when clients who got loan granted exhibited a significantly lower LOR than those who were not granted. Also there were no clients with more than 1 LOR that were approved for loan. There was clearly a correlation between the average LOR and the card issuance/loan grant status.



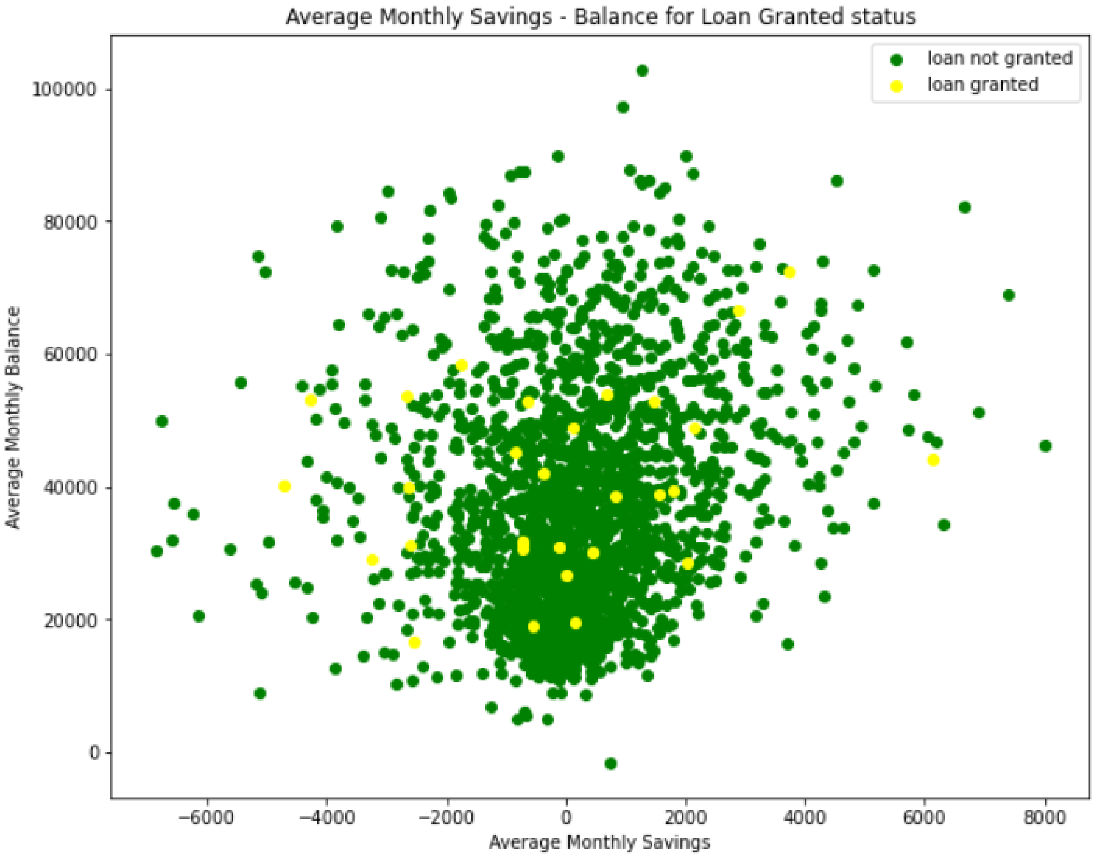
The scatter plots showed the relationship between average total credit/debit and card issuance for the clients. The red dots presented clients who did not have cards issued and the yellow ones were those with issued cards. By looking at the plots, we could see an uphill pattern which indicated a positive relationship between average total credit and average total debit. Although the data points seemed to spread widely across the graph area, more data points were concentrated on the lower region around 30000 and below for both average total credit and debit. From this finding, we could infer that most of the clients did small transactions of under 30000. On the other hand, no difference in the average total credit and average total debit was detected when comparing clients approved or rejected for card issuance. Therefore, there was no relation between average total credit / debit and card issue status.



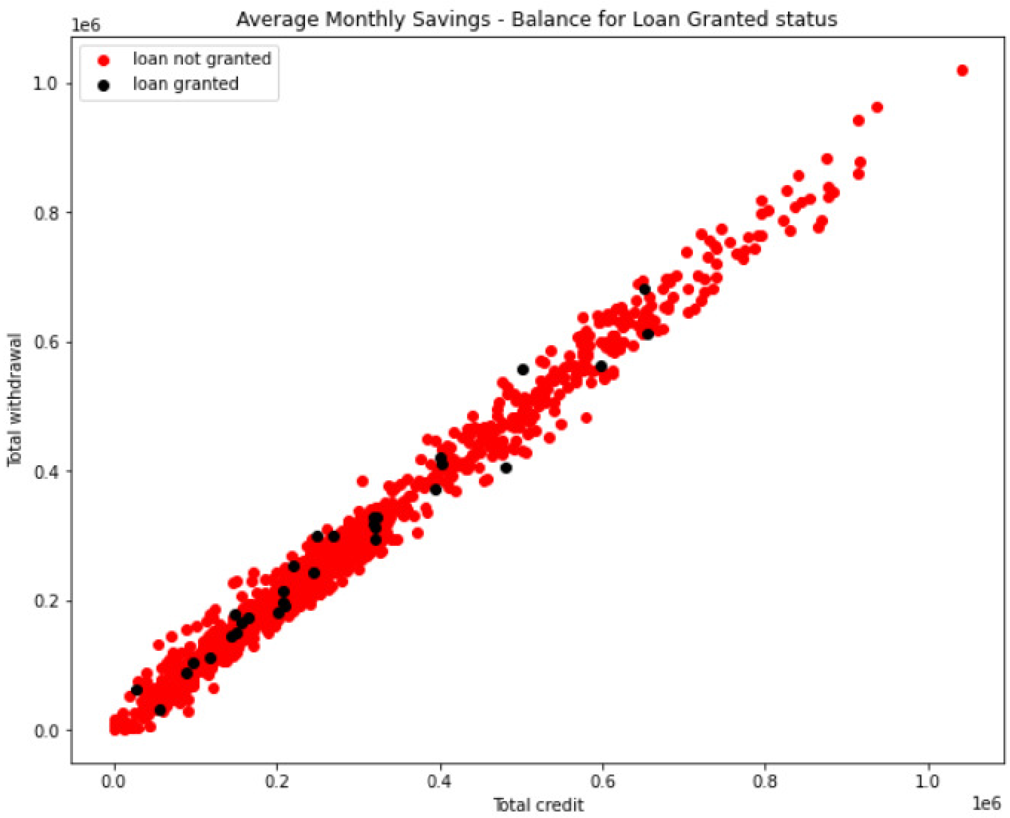
The scatter plot showed the relationship between average total credit/debit and loan grant for the clients. The green dots presented the clients with loans not granted and the yellow ones are those with loans granted. As the data points clustered around a line that runs from the lower left to upper right of the graph area, the relationship between the average total credit and average total debit was said to be positive or direct. In addition, when comparing the average total credit/debit and loan grant for the clients, no difference was observed. Hence, there was no relation between average total credit / debit and loan grant status.



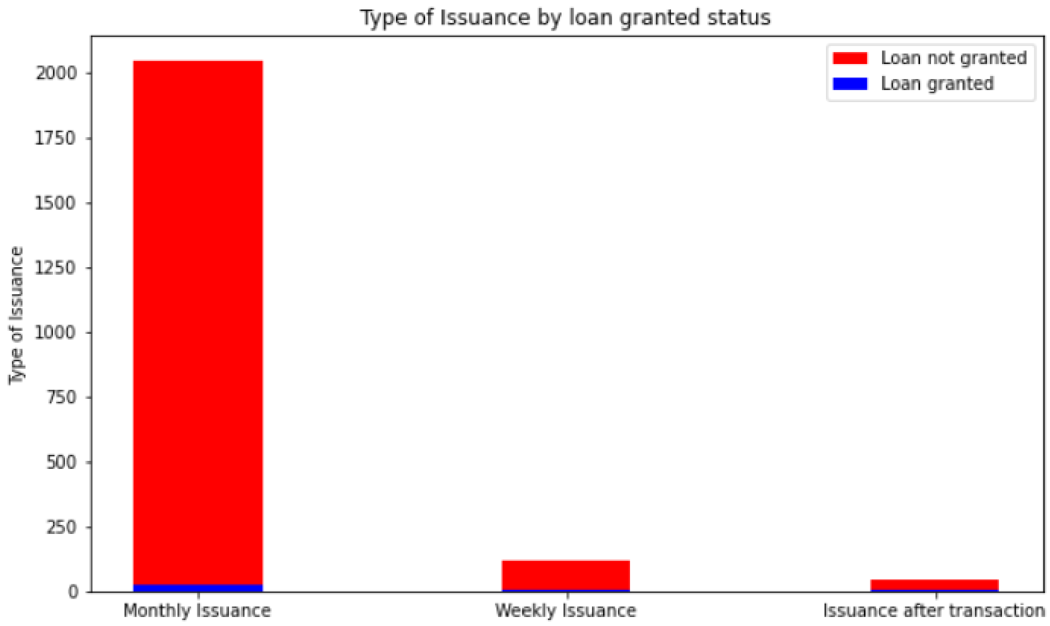
The scatter plot showed the distribution profile of average monthly balance and average monthly savings for both card issued clients and non card issued clients. The majority of clients had average monthly savings of less than 4000 despite the differences in average monthly balance as presented by the clustering of data points in the vertical shape. This suggested that average monthly balance was irrelevant to the average monthly savings. On the other hand, a comparison between non-card issued and card-issued groups showed that most clients were expected to have an average of more than 4000 to get the cards approved. However, there was no clear trend between the card issuance and average total savings or average total balance.



The scatter plot showed the distribution profile of average monthly balance and average monthly savings for clients who had or did not have loans granted. As data points clustered in the vertical shape, we could infer that most of the clients preferred to keep the average monthly savings below 8000 and average monthly balance did not have any relationship with the average monthly savings. When comparing loan grant status, it was observed that clients with low average monthly balance were at a higher risk of being rejected for loan. Nonetheless, there was no clear relation between the loan grant with average total savings and average total balance as the data points were distributed quite evenly across the graph area.

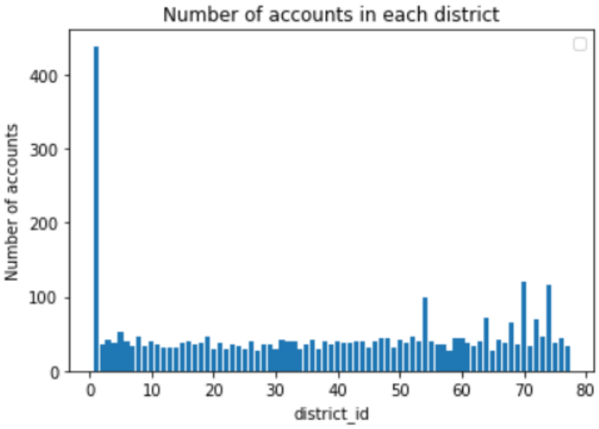


The above graph illustrated the relationship between total credit/total withdrawal and loan grant status for clients. The red dots indicated clients with rejected loans and the black dots indicated those with granted loans. Graphically, a linear pattern was observed where the data has a general look of a line going uphill, thus indicating a positive linear relationship between total credit and total withdrawal. Additionally, data points were widely distributed across the graph area and there was no difference between clients who had or did not have loan grants. This suggested that there was no relation between the total withdrawal and total credit with loan issuance status.

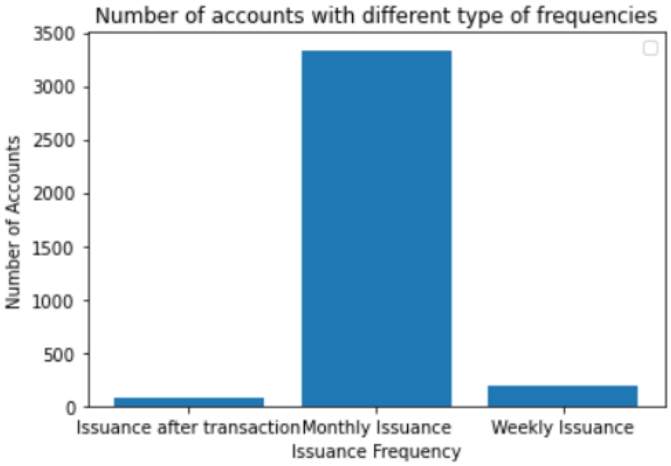


The graph illustrated the number of clients in relation to the type of issuance for their accounts. The red bars showed clients whose loans were not granted, and the overlapping blue bars showed the comparative number of clients for the same type of issuance whose loans were granted. As observed in the graph, the majority of clients opted for monthly issuance despite the loan grant status. However, it seemed that there was no relation between the issuance type and the loan grant status.

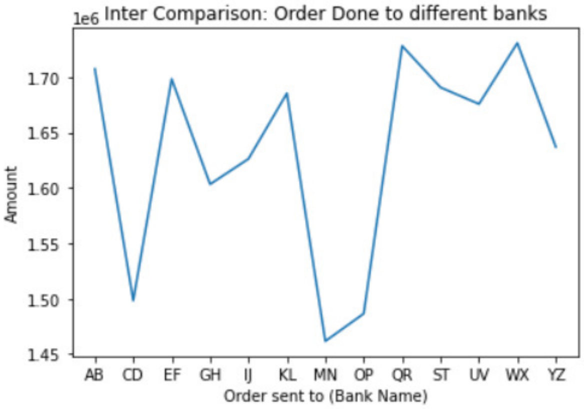
Some other visualizations (based on given individual datasets)



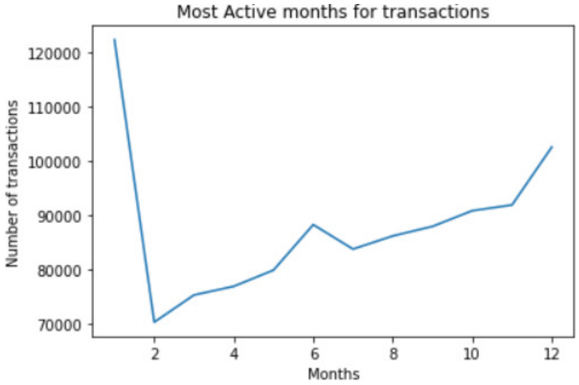
The bar plot showed the number of accounts in every district. The height of the bar represented the number of accounts in that district. The district with district_id 1 had the most number of clients and a huge difference could be seen.



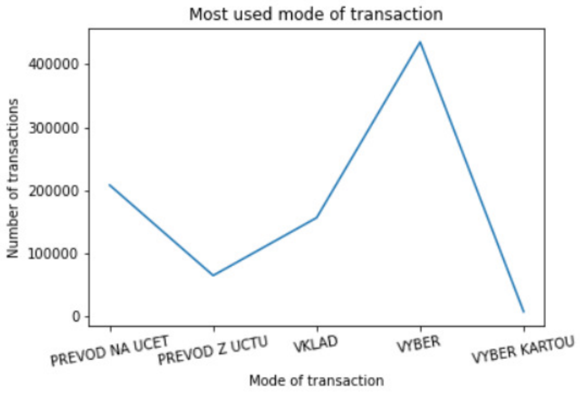
The graph showed the number of accounts with different types of issuance frequency. We could see from the above bars that most of the customers preferred monthly issuance frequency.



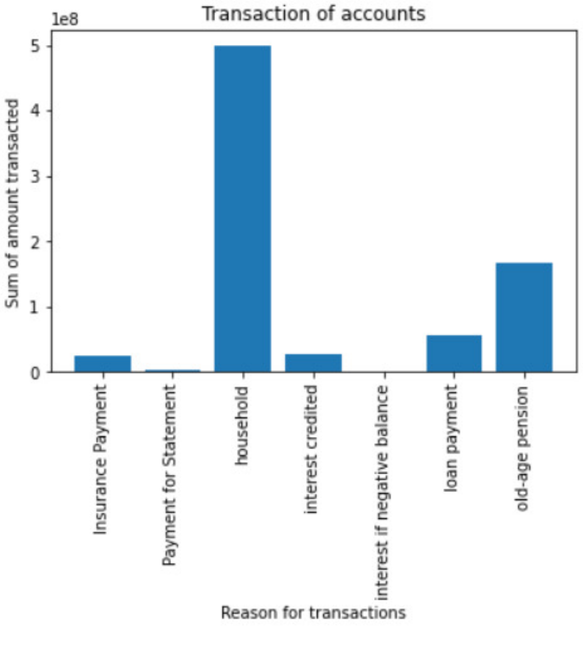
The graph showed the line plot with all the orders amount and the bank orders were sent to. This showed varied behavior of clients and their preferences to send orders to different banks. The most used banks were AB, WX, QR and the least sent banks were CD, MN, OP.



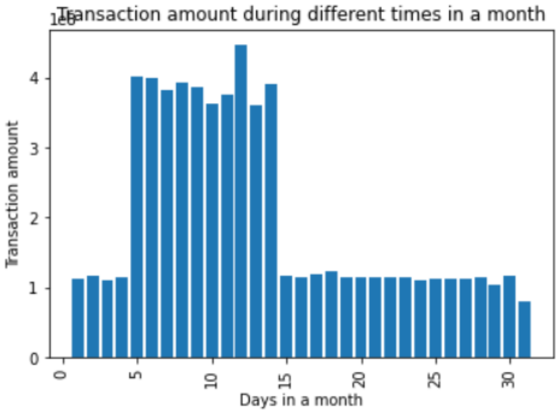
The graph showed the number of transactions done by the clients month-wise. This showed the usage load, as well as financial activity data of the most active month in the year. The above graph showed that clients did most transactions during the starting and the ending of the year. This could be inferred as the clients performed the highest transactions during Christmas and new year and it was the least in February.



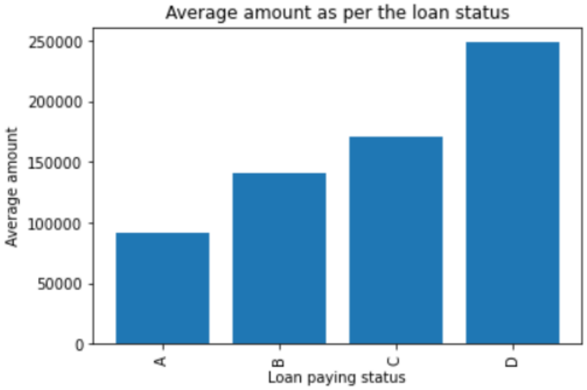
The above graph showed the mode of transactions versus the number of transactions that meant to tell us the most used transaction method. It could be seen that VYBER was the most used mode of transaction which meant withdrawal in cash and the least used mode of transaction was the VYBER KARTOU which meant credit card withdrawal.



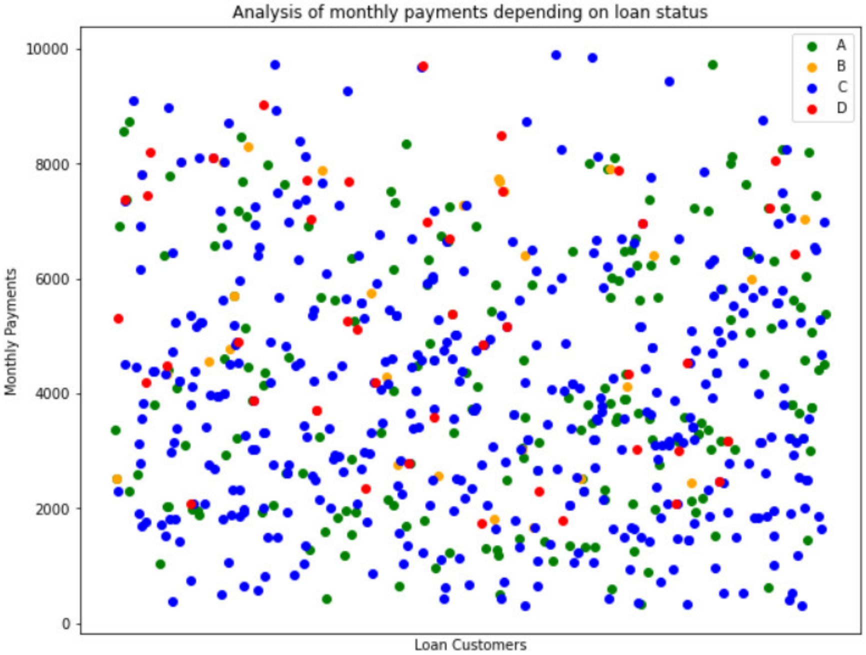
The graph showed the reasons for which the clients did the transactions versus the amount of transactions for that reason. We could see that most of the transactions were performed due to the household reason and the least used reason was 'interest if negative balance'.



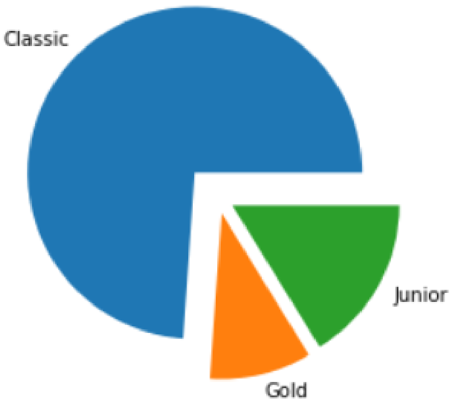
The graph showed the days of the month and the transaction done on them. It told us about the client transactions that when in the month did the transactions happen more often and when the least. It could be seen that transactions during the 5th of every month to 14th were the most transacted days of the month. The least transacted day in the month was the last day of the month.



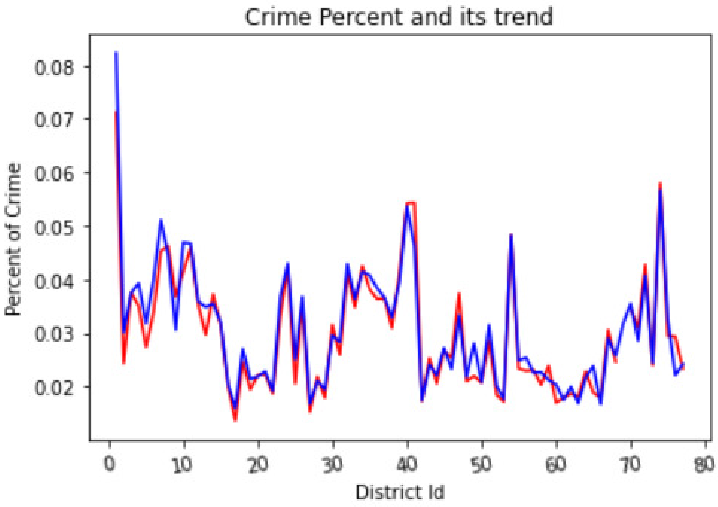
The graph showed the average number of clients with different loan status. The clients with the loan paying status D had the most loaned amount. Additionally, the status D showed that the customer was in debt which here meant that most of the loan amount was granted to the client which was in debt and was not paying it back.



The scatter plot showed the monthly payments of loans in relation with the loan customers with different loan status. The above graph showed that monthly payments were randomly distributed and there was no pattern based on monthly payments to analyze why loan status D clients were in debt or how it was different from other clients. Similarly by plotting other plots, we found that there was no relation between status D and duration or amount for their reason of not being able to pay on time.



The above pie chart showed the divisions of card users based on card type. There were 3 types of card issued to clients which were Classic, Junior and Gold. We could see that Classic was the most used credit card type and the Gold was the card with very few clients.



The above chart showed the crime percent of all the districts in 1995 and 1996. The red lines showed the crime percent of 1995 and the blue line showed the crime percent in 1996. The plot shows that there was no change in the crime percent in the 1 year gap. It was the same.

CONCLUSION

By utilizing different techniques to analyze and visualize data using python programming language, we have successfully extracted the relationships between multiple factors such as average monthly balance, length of relationship, urban inhabitant with the client usage and have tried to gain insights by visualizing them to a better understanding.

To investigate the influence of demographics and socioeconomic status on the issuance of credit card or loan acceptance, we have examined multiple factors. Some of which did not have any relation to the card issuance and loan grant status, some showed slight impact and few had a significant influence on the card issuance as well as loan grant status.

- Factors that had no impact on the card issuance and loan grant status: gender, monthly issuance frequency, total credit – total debit
- Factors that had slight impact on the card issuance and loan grant status: age group, number of inhabitants, average monthly balance, and unemployment rate
- Factors that had a significant impact on the card issuance and loan grant status: LOR (length of relation).

Based on the analysis of past data, the financial institution could decide on the criteria for granting the loan or the issuance of loan based on different investigated factors. The factors which were keeping the customers from paying back their loan on time should be monitored and evaluated in priority. The root cause for them to not pay the loan or credit card bill should be considered the primary support or the guarantee for the issuances of new cards or loans to protect the funds from getting wasted or unreturned. The analysis was crucial and should be performed regularly to ensure a good banking model and save the funds from getting into unsafe hands.

LEARNING

While working in this project we have gained knowledge about the different datasets, working on the raw data, the impurities they might contain, the data transformation they required. We learned different techniques to analyze and visualize data using python programming language, organizing and structuring the steps to carry out a successful project. Other skills gained include collaborative skills as well as working in teams and supporting each other and learning among us.

REFERENCES

- Class jupyter notebooks
- Google
- Matplotlib.org
- Several python libraries