FINANCIAL PROGRAMMING

GROUP REPORT

SWETHA KALLAM, ROHIT BHALERAO, NOÉMIE GAUTIER

Introduction:

We were asked to look at the data collected from a bank and to gather insights from this data. Throughout this report we will look at how we prepared the data and the steps we followed. Indeed, before being able to gather the insight, we had to do multiple steps. We started by looking at our data, cleaning the eight different tables that made up the data. Then we added some variables that we believed would be useful for us to gather insights. Once each table had been prepared, we were able to merge them in a data mart.

Then we will look at the insights gathered from this datamart. We created multiple visuals that highlights the key information in the data and allowed us to gather these insights. We split these into categories to make it clearer: Clients information, Accounts, Credit Cards, Transaction, Loans, and Comparing Loans to other criteria. We looked at insights for the bank on all these aspects, which allow us to conclude the factors that affect tendency of loan repayment.

Finally, we will conclude on our key findings based on this data and what we believe are some of the main issues that the bank is facing.

I. The Datamart:

a. Cleaning and preparing the data:

One of the main goals of this project was to create a DataMart that combined the eight different tables. To do so we started by importing all the data and looking at each of the table individually. Our first step was to clean each table one by one and add variables that could be useful later for our insights.

We modified the column names for each of the tables to make them clearer, we also checked for for any duplicates in the data. Then we did specific modification for each table depending on the data:

ACCOUNTS TABLE:

For the Accounts table we fixed the dates as well as the values for the Frequency column and added an "Account age" column that gives us the number of years since the account has been opened. Additionally, we checked for any duplicates in the data.

CREDIT CARDS TABLE:

For the Credit Cards table, we once again changed the column names, then we fixed the values in the Card Type column and put the dates in proper format. Similarly, to the Accounts table we also added a "Card Age" column to see the how long each card has been active.

CLIENT DATA TABLE:

For the Client Data table, we changed column names and formatted the data of birth column from YYMMDD to YYYYMMDD which converted it to string. Additionally, we created a function to add Gender and we extracted the Year, Day, and Month from the Date of Birth column. We added an Age column as well as an Age group column. Each client is either a disponent or an account owner and cannot be both. Each account will have an account owner but might or might not have a disponent.

ORDERS TABLE:

For the Orders table we replaced the values for the Payment Type column, we also created a new table as the table order1 is a more detailed version of table order we needed to separate the two. We therefore added new columns for each payment type. Moreover, for each of the OrderID we added the Amount value to new columns corresponding to the Payment type. We also did a Group by Account ID (one row per Account ID), this way for each account ID there is the count of orders and recipients, sum of total payments and payments per Payment Type, i.e., one row for each account ID describing all the types of orders for it.

LOAN TABLE:

For the Loan Table we fixed the date column and added a column to describe Loan Status.

DISTRICT TABLE:

For the District Table we had to deal with missing values by replacing '?' values with NaN. Then we changed columns to numeric replacing NaN values with the mean of the Region Group.

TRANSACTION TABLE:

For the Transaction Table we fixed the Date Column into proper format. Then we changed the column values for three columns: Transaction Type, Operation Type, Payment Type. We also made the decision to dropping two columns with many NA values as they were not too much use: Recipient Bank, Recipient Account. Additionally, we created a new table 'trans2' as a duplicate of 'trans' table and removed a few columns. We duplicated the column 'Date of Transaction' as 'First' & 'Last Transaction' which was needed in the final group by of this table. Moreover, we created a dummy variable for Credit and Withdrawal Transaction types as this was also needed for the final group by. We also created columns for each of the Operation Types, for each transaction ID, we added the amount to the corresponding operation (dummy variables) needed for the final group by. Then we removed unwanted columns (they were not needed as we have dummy variables). Finally, we did the group by Account ID, and aggregated the other columns. The dummy variables are also aggregated here. We changed the column names one last time and reordered the columns in the data frame. To finish on this table, we added a new column to count days since the last transaction. The transaction table at the end had one row for each account with a summary of all the transactions for that account.

b. Merging the Tables together:

Now that all our tables have been cleaned and useful variables have been added we can finally create the Datamart by combining all the tables together. To do so we merged all the tables together using multiple outer joins, we combined the tables one by one until we got the final table. Then we had to reorganise that table properly to get all the clients as rows.

c. Final output:

The final output of our DataMart can be found in the file datamart.csv. The DataMart has 5369 rows, one for each of the bank's client, and 73 columns, for each of the key variables initially in the tables, and the ones we added.

II. Insights:

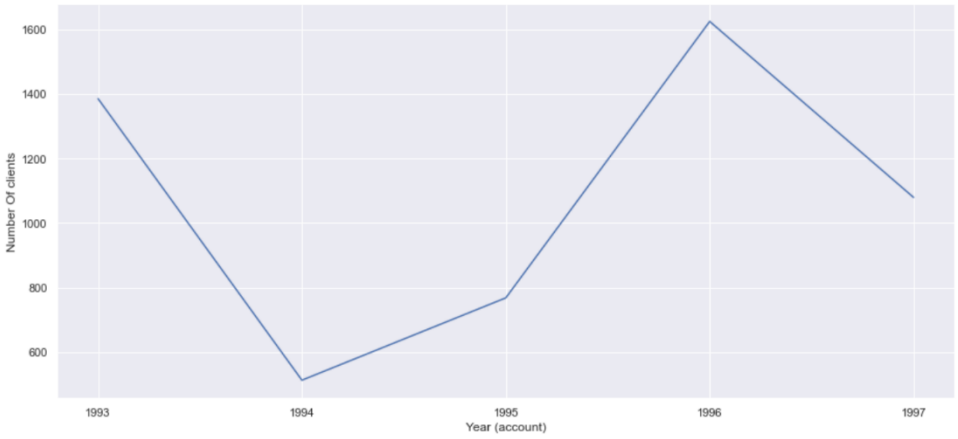
From the Datamart we created multiple visuals on different aspects of the data. Here we will discuss all the insights available with the data and visuals.

a. Clients Information:

A key aspect of gathering insights on any company is looking into their client base and the demographics of these clients. We therefore visualized the demographics of the bank clients by years, age groups, gender and date of birth. We also looked at all the client demographics and compared their transactions and balance amounts.

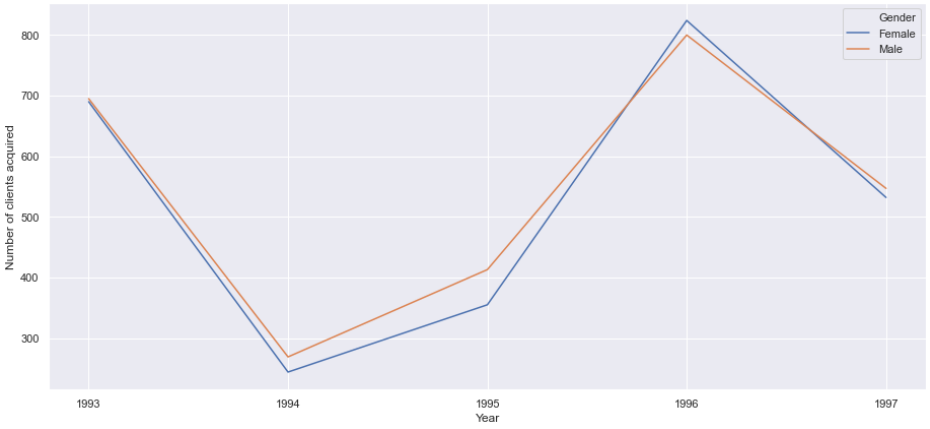
CLIENTS DISTRIBUTION BY YEAR:

The number of clients decreased in the year 1994, then steadily increased, with a peak of more than 1600 clients in 1996 and then we see a sudden drop again in the year 1997.



CLIENT DISTRIBUTION BY YEAR AND GENDER:

We see that the distribution for Male and Female customers is very similar to each other. But it is noticeable that the Female customers were least in number in 1994 and 1995 but also highest in number in 1996, although by very small amounts. It could be interesting to investigate this and whether there is a reason for and increase in female clients. Although as the difference is so small there might not be any specific reason.



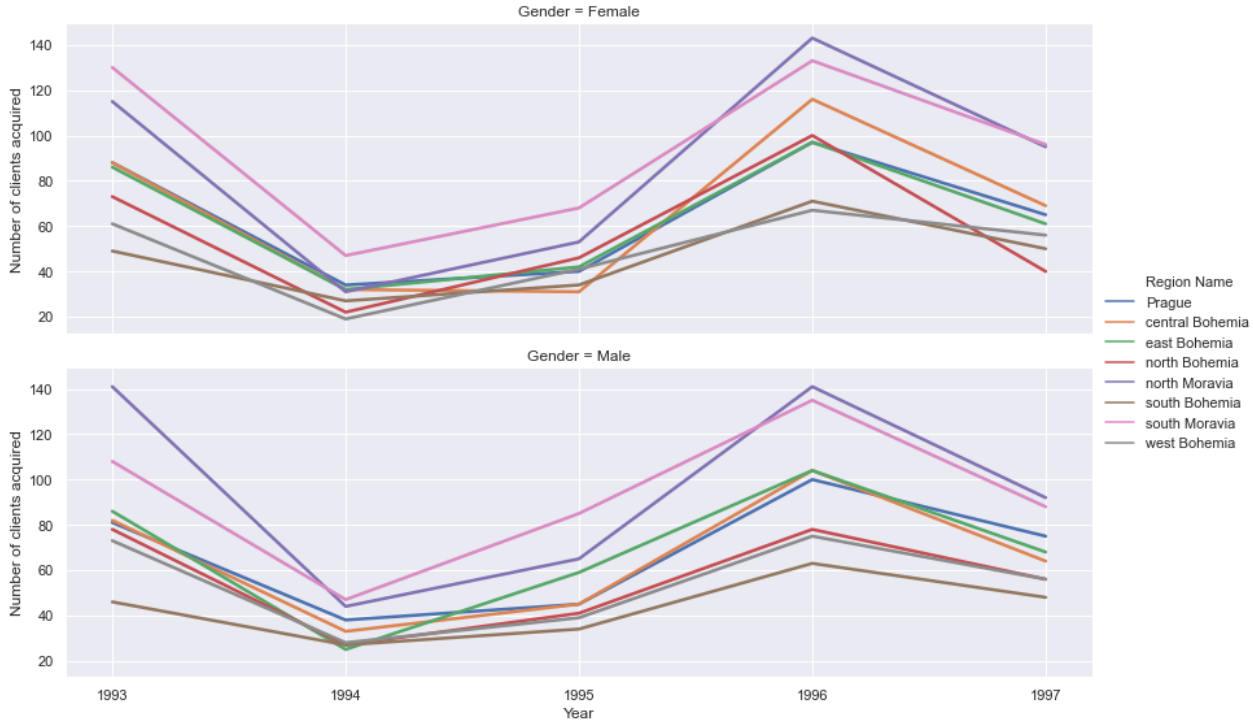
CLIENT DISTRIBUTION BY REGION, YEAR AND GENDER:

It should be noted that the distributions of Male and Female clients are a little different with respect to region.

Overall, we can say that the regions North Moravia and South Moravia are performing well, while the regions South Bohemia and West Bohemia are performing poorly with respect to number of customers.

Here are some key features:

- The highest number of clients for both genders is in the region North Moravia, closely followed by South Moravia.
- For the year 1997, there is a steep decrease of Female clients in North Bohemia region, the reasons for which should be further investigated.
- Overall, the least number of Female clients is for the West Bohemia region, in the year 1994.
- Overall, the least number of Male clients is for the East Bohemia region, in the year 1994.

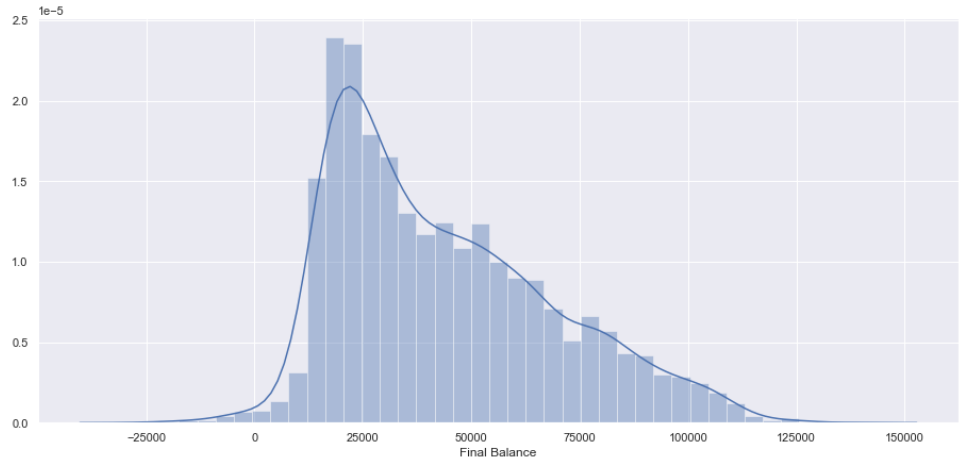


b. Accounts:

Another key feature we investigated was the data collected regarding the accounts. This is a key criterion to analyse for a bank as it can differentiate customers.

DISTRIBUTION OF FINAL ACCOUNT BALANCE:

We see that the distribution of the balance amount is left skewed / right tailed. The highest probability is in the range of around 25000.



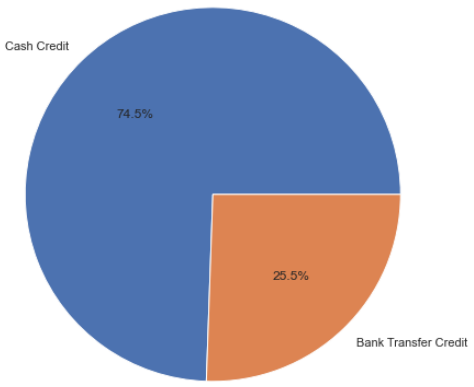
This tells us that there is a range of account balance from a little bellow 0 to 125,000. But most accounts are located around the 25,000-total balance. Finally, it should be noted that in 1999, when this data was collected, 39 accounts have a negative balance. These accounts and clients need to have a eye kept on them as they will need to pay back this money.

c. Transactions:

Transactions are also a key aspect of a bank's data. Knowing if there is a type of transaction that is more common than another could be helpful for the bank to make marketing decisions, pricing decisions and to guide their clients into certain types of services the bank can provide.

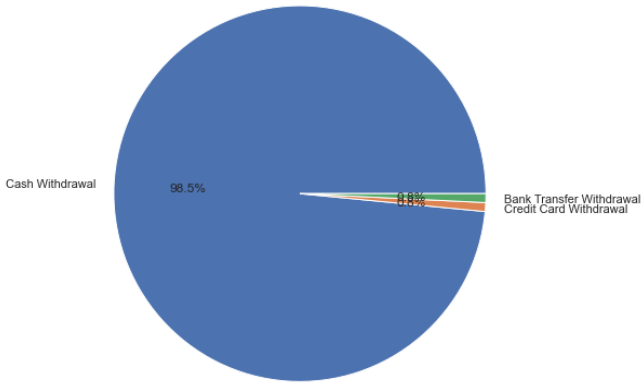
COMPARING DIFFERENT TYPES OF CREDIT TRANSACTIONS:

Cash credit transactions are more common with 74.5% of the total while Bank Transfer credits are only 25.5% of the total.



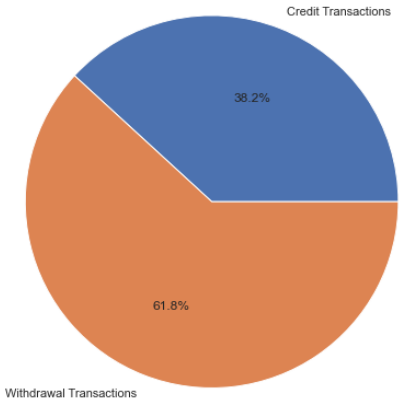
COMPARING DIFFERENT TYPES OF WITHDRAWALS TRANSACTIONS:

In withdrawals, we see that cash withdrawals dominate with 98.5% of the total, while Bank transfer and credit card withdrawals are only ~0.8% each.



COMPARING TOTAL CREDIT AND WITHDRAWAL TRANSACTIONS FOR ALL ACCOUNTS:

Here we can see that most transactions are withdrawals, and there is less credit transaction. This makes perfect sense as the data is from 1999. Customers still prefer to pay in cash.

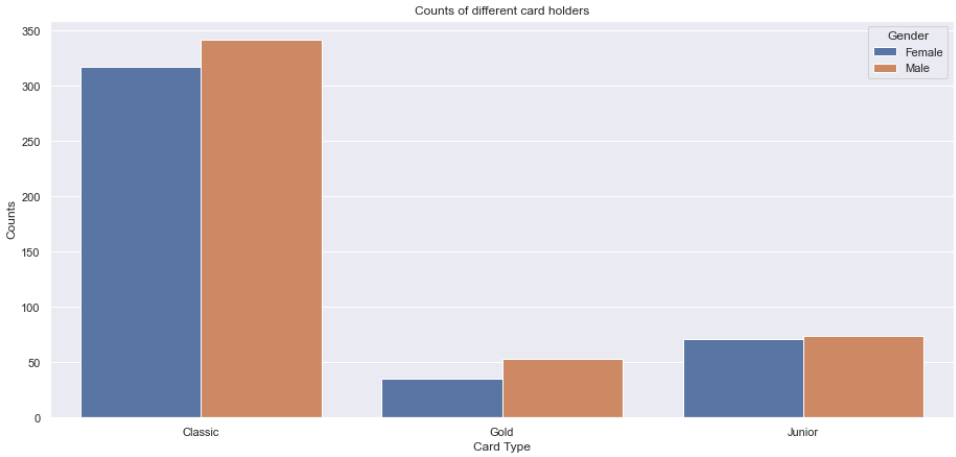


d. Credit Cards:

Credit cards are also a key aspect of bank products and services. They can be a differentiating factor as well as a way to target different groups of customers.

CREDIT CARD HOLDER BY TYPE AND GENDER:

We see that Classic is the most common credit card type. And we also see that males are slightly more in number for all types. This can be linked to the fact that there are more male clients overall in the bank. For each card there is a slight difference between females and males, but this gap is bigger for the Classic and Gold cart types. Junior Cards are almost equal for males and females.

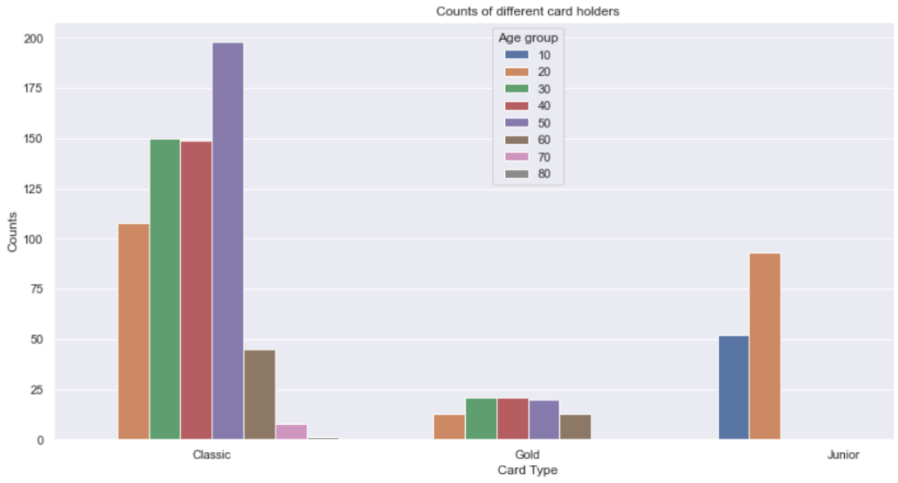


CREDIT CARD HOLDER BY TYPE AND AGE GROUP

The age group of 50 is the most dominating in the Classic type. Age groups 30 and 40 have approximately same numbers in each category which suggest they have similar preferences regarding their Credit Card. People with age groups 70 and 80 are very few and only exist in the Classic type. This age group most likely hasn't transitioned fully to credit card usage which seems normal.

People having Gold type of credit card are comparatively less in number. This is probably because Gold cards are generally more elite, and only for a selected number of customers that can afford them.

Junior group, as we can expect, is for children, so the most common age group in Junior type is 20 followed by 10.



e. Loans:

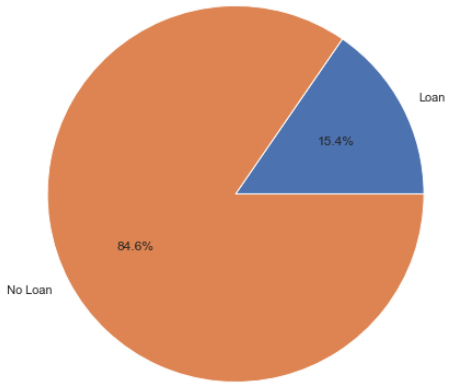
Then we analysed the loans to know different patterns in loan repayment and defaulter. Indeed, loans are a key aspect for banks, and defaulters can cause a real problem so knowing which factors most influence repayment or defaulters can help the bank predict which clients should be investigated.

Some of the factors that can help identify loan defaulters are the total loan amount, region, type of card and disposition, amount of credit or withdrawal transactions, frequency of transactions can help determine clients that will or will not repay loans in the future.

We also established some factors that do not affect loan defaulters such as age, age group, gender and Total Transaction amounts do not pay a major role in knowing Loan Defaulter patterns.

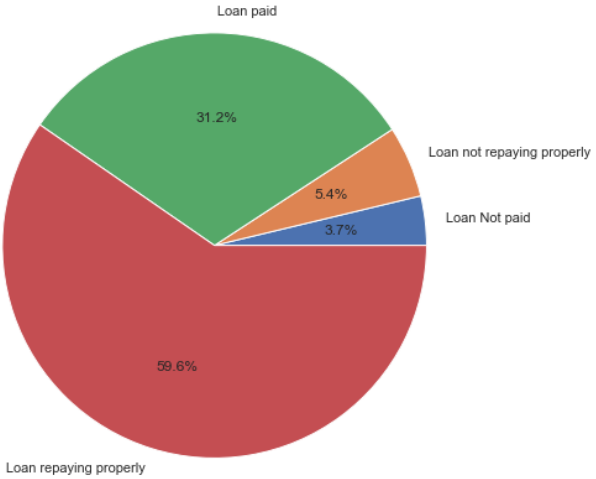
ACCOUNT & LOAN:

Here we can see the percentage of clients that have loans, which corresponds to 15.4%. Most clients don't have current loans in the bank. The bank could decide to advertise loans more if they want more of their clients to take loans, but this should probably not be advertised to all customers or not the same types of loans.



DISTRIBUTION OF LOANS BY STATUS:

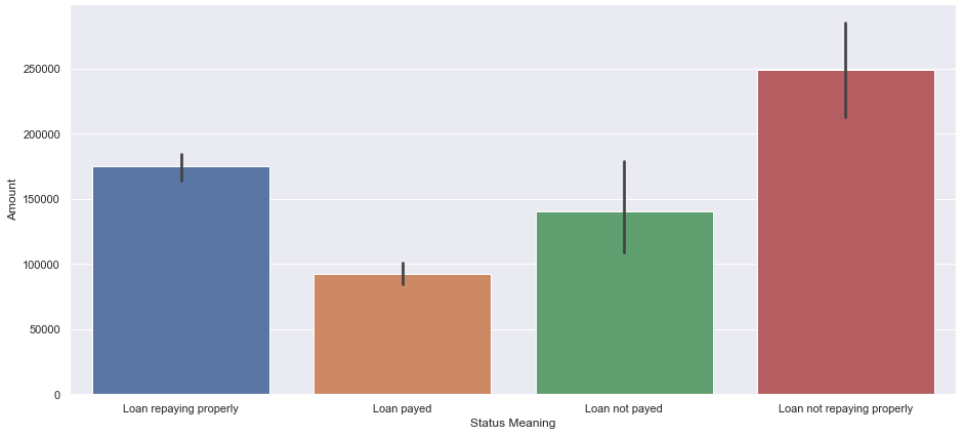
We see that majority of the clients are repaying or repaid their loan amounts on time, only a small percentage of clients are defaulters. The bank should focus on these clients to avoid any other clients with the same profile making loans and defaulting as well.



On the next graph we can see the amounts for these defaulters, which is quite shocking. Even if only a small percentage of clients with loans default, if the amount not payed is very high than this creates a big problem of the bank.

LOAN STATUS VS AMOUNT:

We see that we have a total loan amount of 250,000 where the clients are not repaying the monthly payments on time and around 140,000 where the loan term is over, but the clients have not yet repaid. Overall the total amount with defaulters is more than the total amount with those who are repaying correctly.



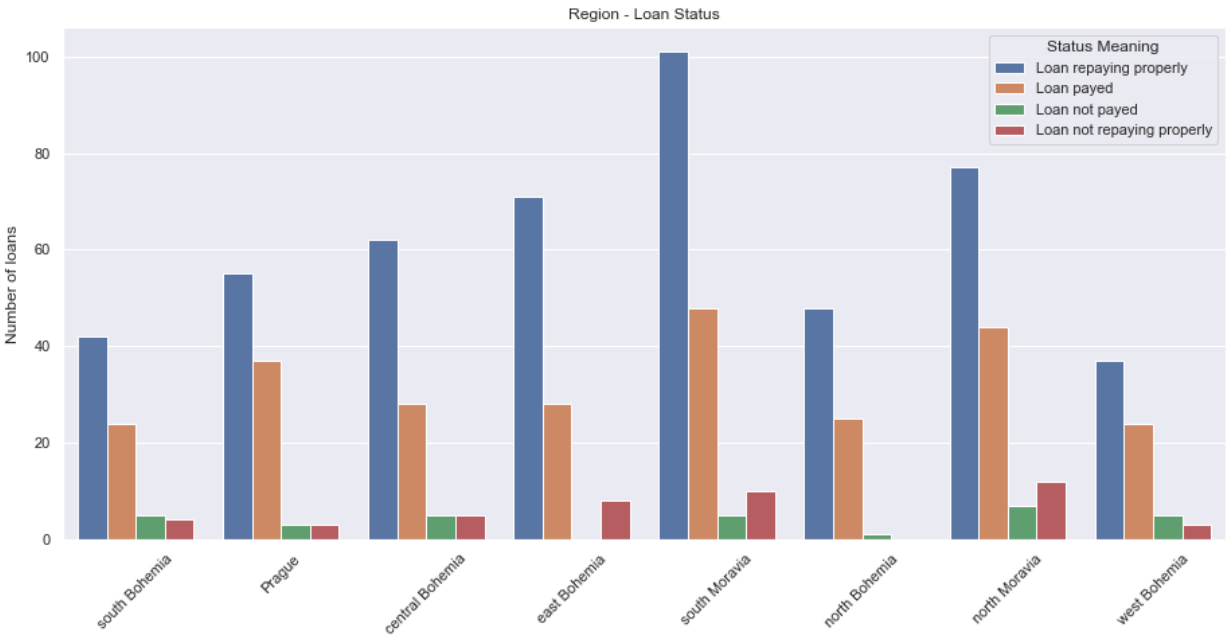
f. Comparing Loans to other criteria:

As loans are such an important aspect of a bank, and that from the last graph we saw there was a clear issue with defaulters we decided to investigate comparing loans to other criteria as well.

LOAN STATUS BY REGION:

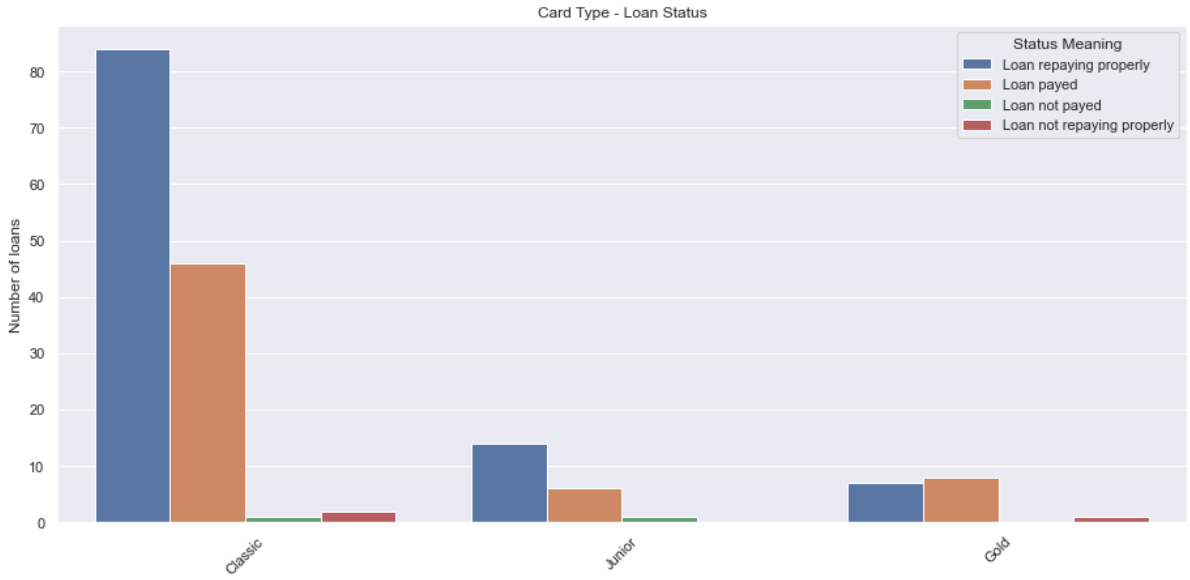
Here we can see that depending on the region loans status are different. Overall, as seen before there are more repaying loans then not but some regions have a higher amount of loans not payed back. Indeed, we can see that in North Morava there a quite a lot of no repaid loans or nor properly repaid in comparison to the number of repaid loans. On the other hand, in North Bohemia and East Bohemia

there are almost no unpaid loans. The regions with high loans not being repaid should review their clients and see what the issue is.



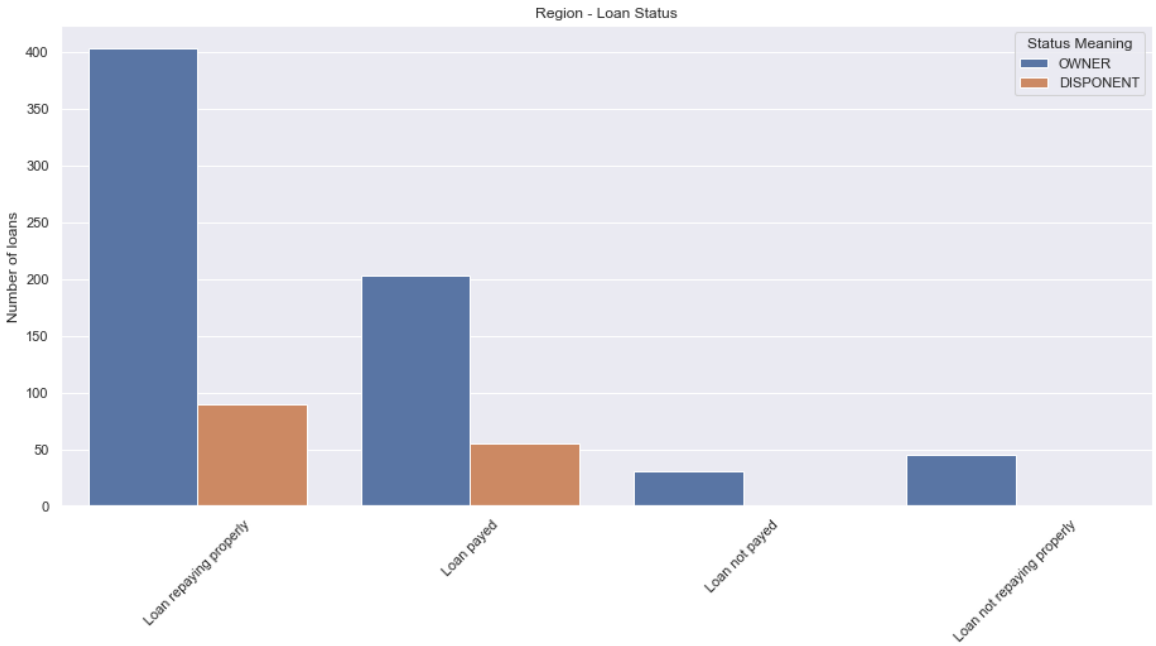
LOAN STATUS VS CARD TYPE:

When comparing the loan status to the type of card owned, we can see that clients with Gold card have always paid their loans in the past. While the same is not true about clients with Classic or Junior cards.

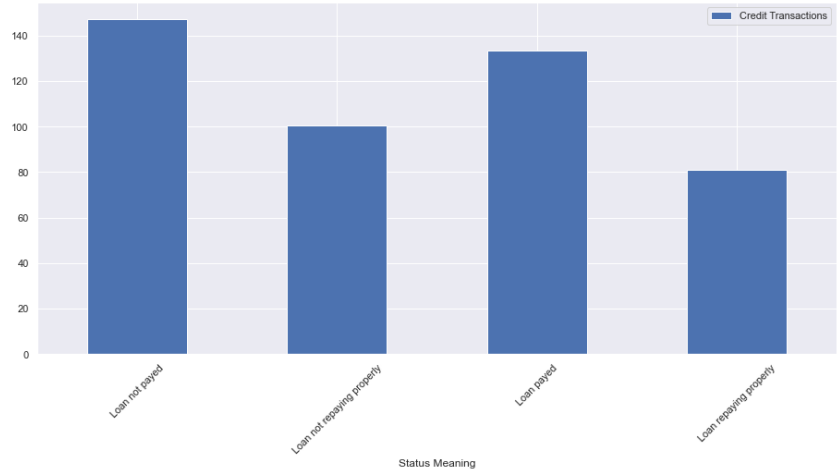


TYPE OF ACCOUNT VS LOAN:

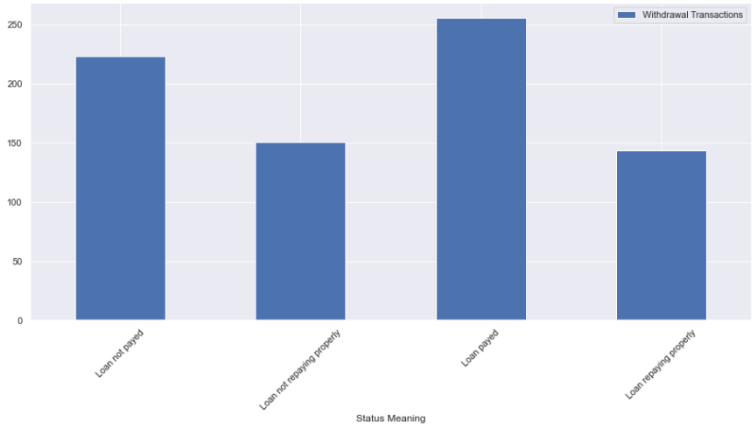
While account owners sometimes fail to repay loan, the Disponents (non-owning account clients) have always paid loans on time.



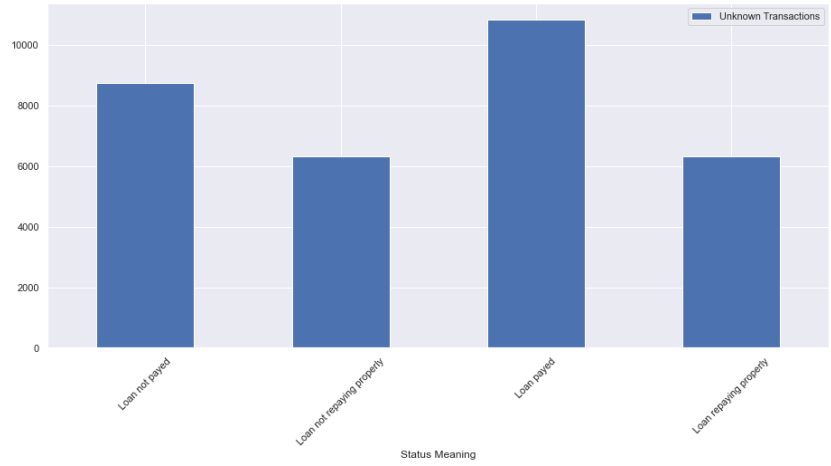
LOANS STATUS VS TRANSACTIONS:



Accounts which do not pay loans on time on an average have higher credit transaction amounts. This is interesting as it means they still spend a lot of money but do not pay back their loans.



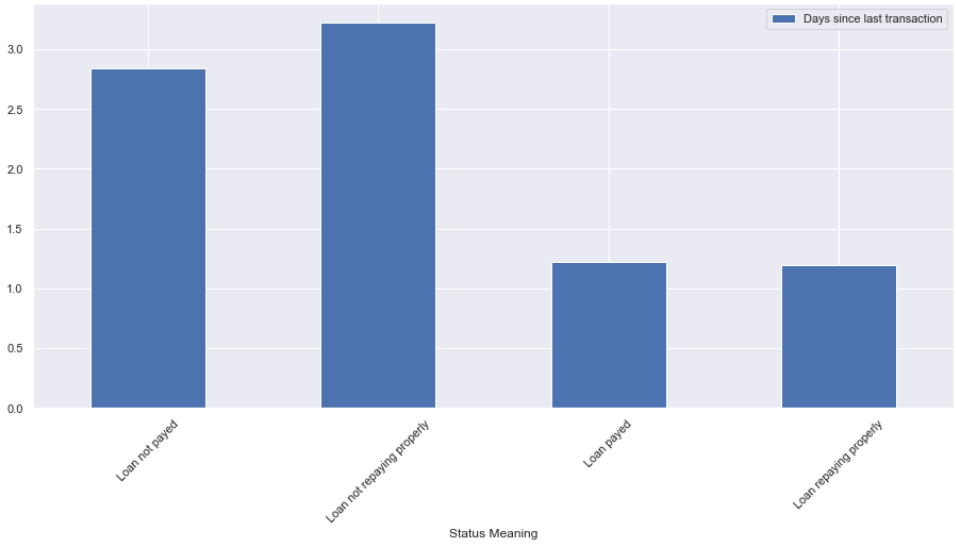
On the other hand accounts which pay loans on time on an average have higher withdrawal transaction amounts



Additionally accounts which pay loans on time on an average have higher unknown type transaction amounts.

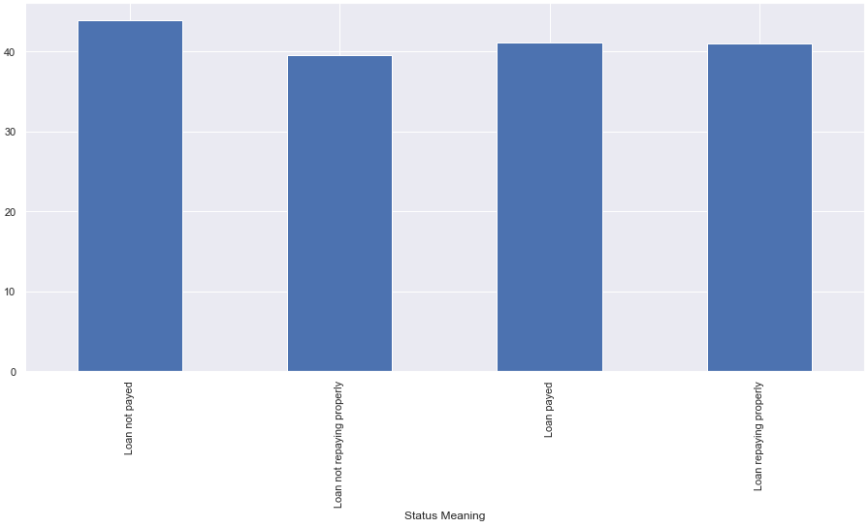
LOANS STATUS VS ACCOUNT USAGE:

Clients which repay their loans also use accounts more frequently, meaning that they are frequent clients of the bank.

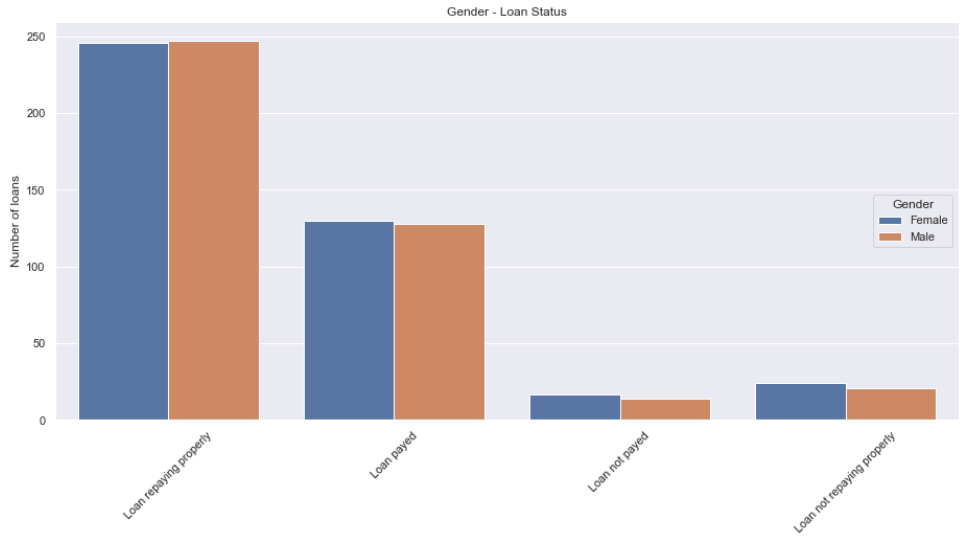


Finally, as mentioned above we found that age, age Group, gender and total transaction amounts do not pay a major role in knowing Loan Defaulter patterns. This can be seen on the following visuals.

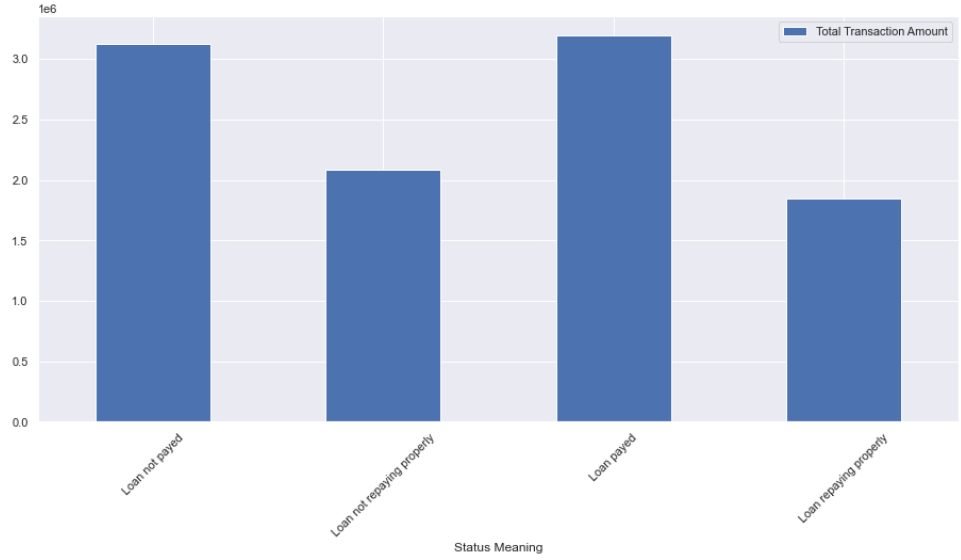
LOANS STATUS VS AGE:



LOAN STATUS VS GENDER



LOAN STATUS VS TOTAL TRANSACTIONS



Conclusion:

To conclude, this data allowed us to create a data frame and then analyse it with various visuals. Regarding the demographics, accounts, credit cards, and transactions we can see that the bank targets different types of customers. There are Junior customers who are less than 20, but there are also older customers that are over 70 and hardly use credit cards. In between there are various clients, with different incomes and the different credit card types, and transaction habits. This highlights the various types of clients a bank can have and it's important for the bank to be able to know all these customers habits and preferences. This is what our insights allows the bank to do.

Some of the key insights we found and that the bank should consider is that there is a real problem with loan defaulters. When loan is high there is more chances that it won't be paid back properly, and the amount of money not paid back is therefore very high. We investigated the criteria that can impact loan payments to help the bank predict future defaulters. We found that total loan amount, region, type of card and disposition, amount of credit or withdrawal transactions, frequency of transactions can help determine clients that will or will not repay loans in the future. On the other hand, Age, Age Group, Gender and Total Transaction amounts do not play a major role in knowing Loan Defaulter patterns. Additionally, the loan status varies from one region to another, indeed clients of East Bohemia and North Bohemia tend to pay the loan, while clients from North Moravia are more likely to not repay the loan than other regions.