**Sai Jayanth Kanukolan**

**Jay.srivatsa07@gmail.com**

**understanding the Loan Business of a Bank**

**Data Science Evaluation -Merilytics**

Contents

[1 Introduction 3](#_Toc21626272)

[2 Objective 3](#_Toc21626273)

[3 Questions 3](#_Toc21626274)

[4 Data Provided 3](#_Toc21626275)

[5 Software/Libraries Used 3](#_Toc21626276)

[6 Overview 4](#_Toc21626277)

[7 Lending Patterns 5](#_Toc21626278)

[7.1 Lending Patterns Based on Income 5](#_Toc21626279)

[7.1.1 Income Group vs Number of Loans Given and Rejected 6](#_Toc21626280)

[7.1.2 Income Group vs Loans Repaid 7](#_Toc21626281)

[7.1.3 Types of Loan vs Income Groups 7](#_Toc21626282)

[7.1.4 Income Group vs Loan Grant for First Time 8](#_Toc21626283)

[7.2 Lending Pattern Based on Loan Purpose 9](#_Toc21626284)

[7.2.1 Loan Purpose vs Number of Loans Given and Rejected 9](#_Toc21626285)

[7.2.2 Loan Purpose vs Loans Repaid 10](#_Toc21626286)

[7.2.3 First Time Loans Analysis based on Purpose of Loan 11](#_Toc21626287)

[7.2.4 Purpose of Loan vs Average Savings 12](#_Toc21626288)

[7.3 Lending Pattern based on Age 13](#_Toc21626289)

[7.3.1 Age vs Loan Purpose 13](#_Toc21626290)

[7.3.2 Age vs Loans Granted and Rejected 13](#_Toc21626291)

[7.3.3 Age vs Loans Granted for Various Purposes 14](#_Toc21626292)

[7.4 Lending Patterns based on Credit Card Limit Used 14](#_Toc21626293)

[7.5 Lending Patterns based on Number of Dependents 15](#_Toc21626294)

[7.5.1 Number of Dependents vs Loans Granted and Rejected 15](#_Toc21626295)

[7.5.2 First Time Loan Application, Loans Granted and Repaid 16](#_Toc21626296)

[7.5.3 Number of Dependents vs Income 16](#_Toc21626297)

[7.5.4 Number of Dependents vs Yearly Salary 17](#_Toc21626298)

[7.5.5 Number of Dependents vs Loan Purpose 17](#_Toc21626299)

[7.6 Summary 18](#_Toc21626300)

[7.6.1 Understand the loan business of the bank. 18](#_Toc21626301)

[7.6.2 Identifying customers who haven’t been granted a loan which are more likely to repay the loan and provide loans to them to maximize the profits of the bank. 19](#_Toc21626302)

[7.7 Recommendations to Maximize Bank’s Profits 19](#_Toc21626303)

[7.7.1 Bank Should Continue to do 19](#_Toc21626304)

[7.7.2 Bank Should Discontinue 19](#_Toc21626305)

[7.7.3 New Strategies that the Bank could Implement 19](#_Toc21626306)

[8 Process of Forecasting 20](#_Toc21626307)

[8.1 Step 1: Importing the Required Libraries 20](#_Toc21626308)

[8.2 Step 2: Importing Data and Preliminary Analysis 21](#_Toc21626309)

[8.3 Step 3: Exploratory Data Analysis 21](#_Toc21626310)

[8.4 Step 4: Data Manipulation and Preparation 21](#_Toc21626311)

[8.5 Step 5: Modelling 22](#_Toc21626312)

[8.6 Testing 22](#_Toc21626313)

[8.7 Forecasting 22](#_Toc21626314)

[9 Result 23](#_Toc21626315)

# Introduction

A Bank want to understand its loan business. It wants to identify who would be the right target for the loans. The data contains both customers who have been given loans and those who haven’t yet received loans.

# Objective

1. Understand the loan business of the bank.
2. Identify the customers who haven’t been granted a loan which are more likely to repay the loan and provide loans to them to maximize the profits of the bank.

# Questions

1. Perform an EDA of the features in the data. Describe any interesting insight you might get when you go through the data.
2. Implement a model to make maximum profit for the bank when providing loans. Please implement at least 2 algorithms of your choice. Explain the reason for choosing the algorithms.

# Data Provided

Two data sets have been provided:

1. Borrower\_table – Contains Information about borrowers
2. Loan\_table – Contains information about loans

# Software/Libraries Used

1. Software –Jupyter Notebook, Tableau-Public
2. Libraries used – Numpy, Pandas, Matplotlib, sklearn

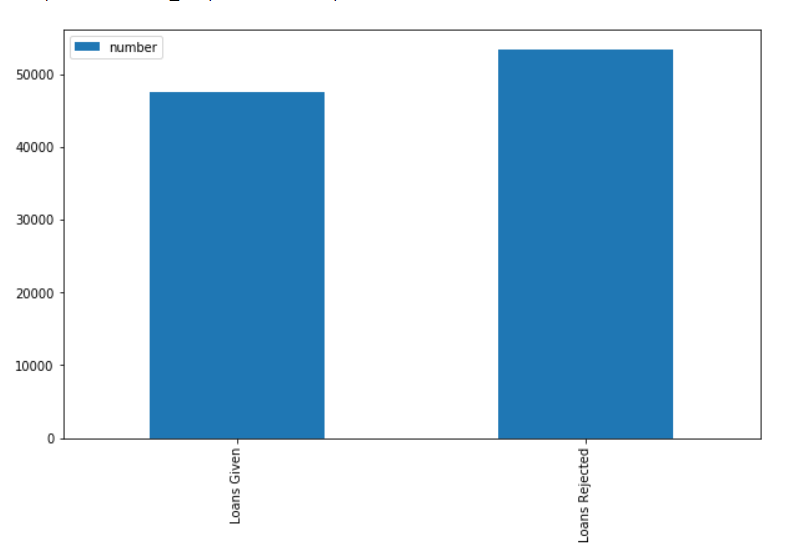
# Overview

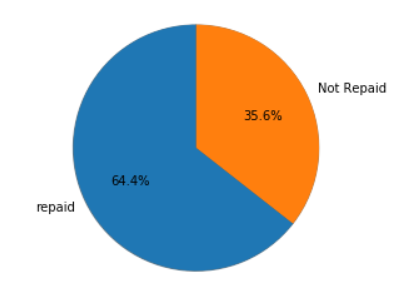
An EDA has been performed on the given data sets of bank customers. The given data has been classified based on

1. Income of the customers
2. Purpose of Loan
3. Number of Dependents
4. Age
5. Credit Card Limit

Considering the above classifications lending patterns of the bank have been studied.

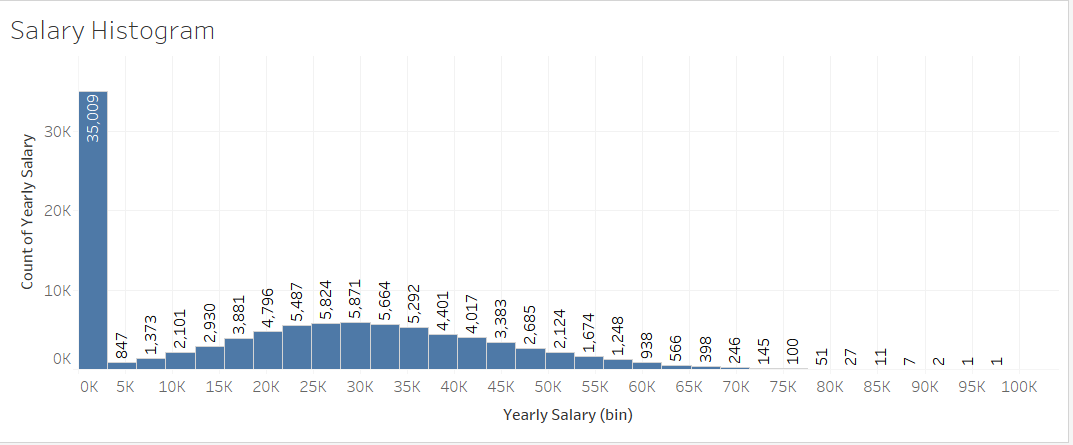
The number of loans given and rejected by the bank are shown below. Out of the loans given about 64.4% of loans have been repaid and 35.6% of the loans have not been repaid.



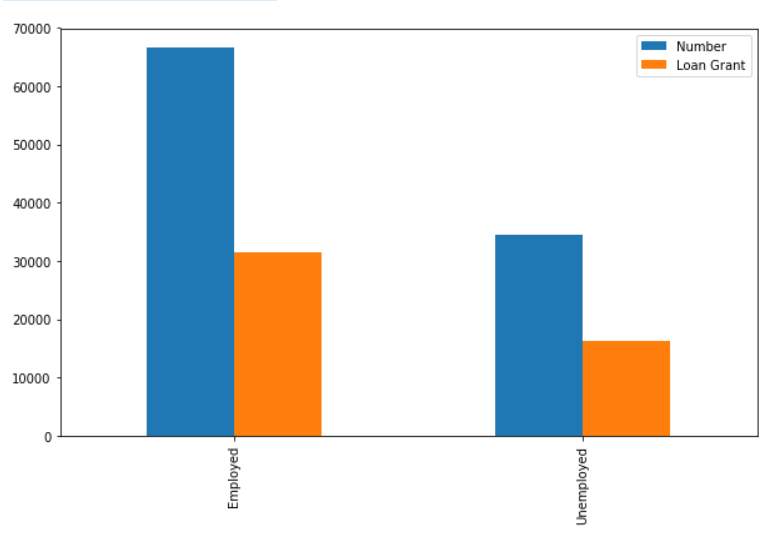


# Lending Patterns

## Lending Patterns Based on Income



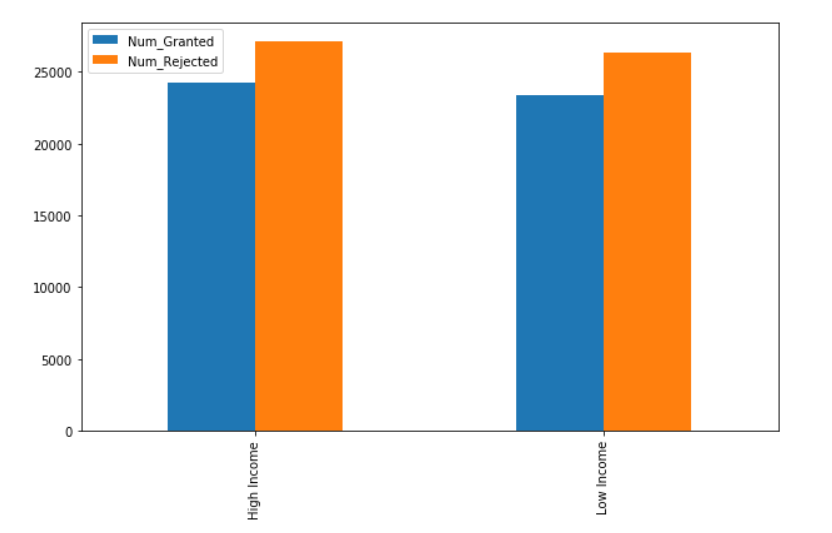
The above graph represents the salary distribution of the data set. Maximum number of people in the data set, about 35000, lie in the first bin, which implies they are either unemployed or have very low income.



About 50% of employed and 50% of unemployed customers get loans from the bank.

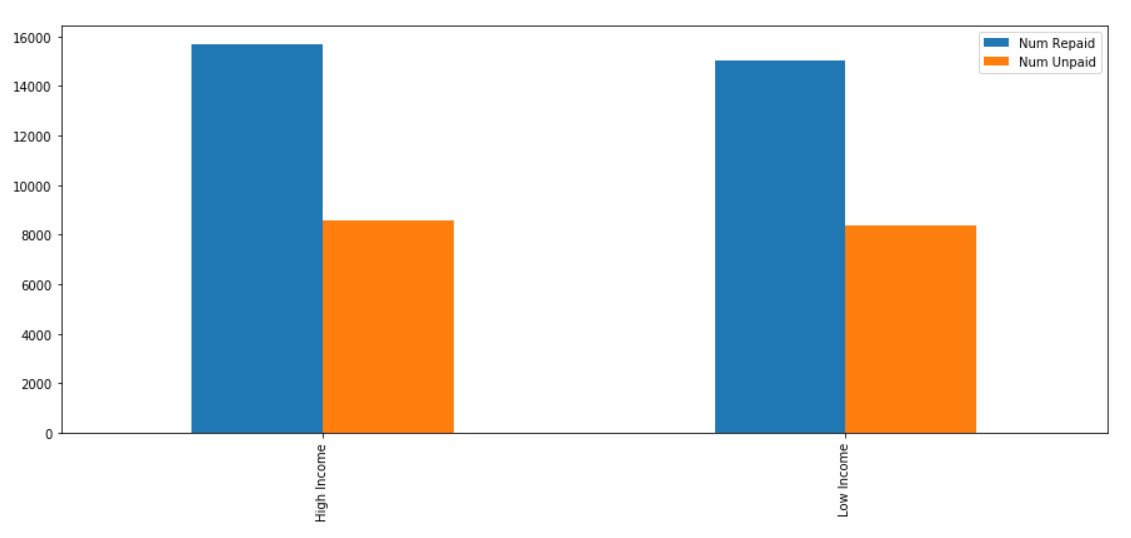
### Income Group vs Number of Loans Given and Rejected

For the purpose of classifying borrowers based on their yearly income, average income of all the borrowers was considered. People with income above the average are classified as high-income group where as people with income less than the average are classified as low-income.



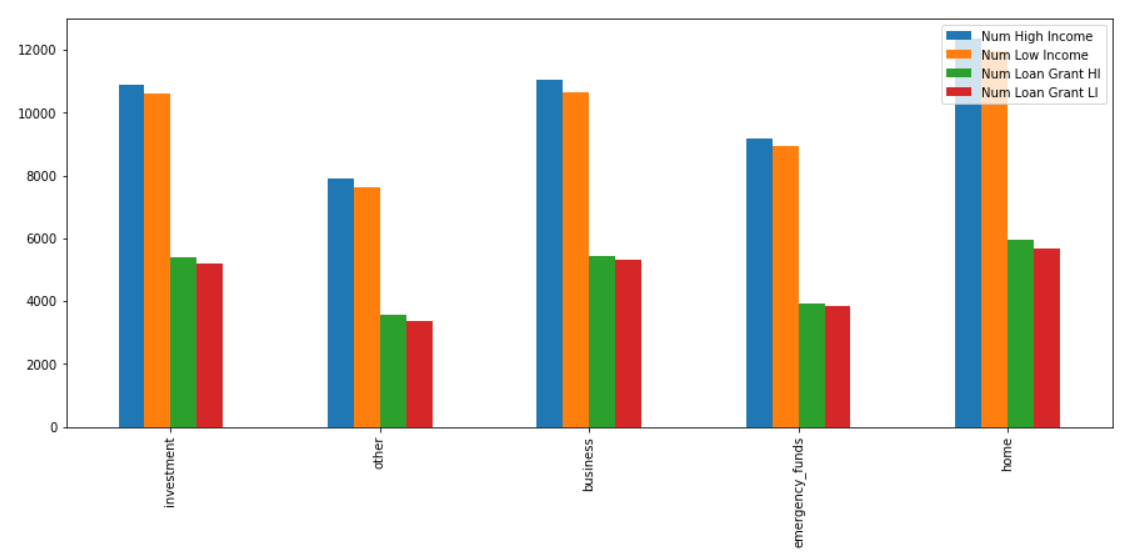
As observed from the graph the bank provides loan almost equally to high income as well as low income groups.

### Income Group vs Loans Repaid



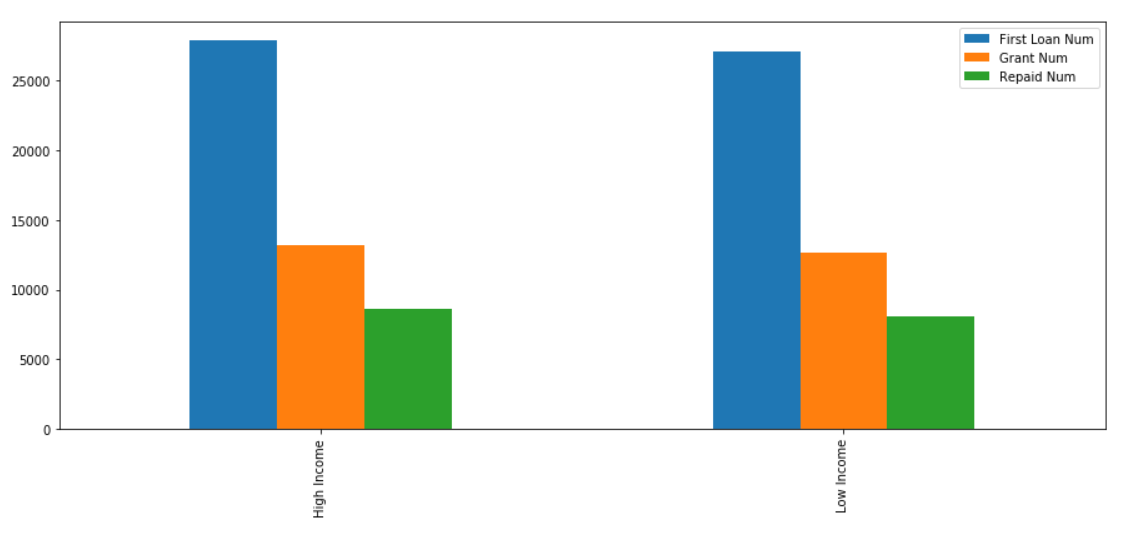
As observed the loans repaid by high income and low-income groups are approximately the same**.** Which means the loan amount granted by the bank is well within the ability of the customer to repay.

### Types of Loan vs Income Groups



Number of High-income customers for the investment, business and home are almost the same. And the loan grants for low-income and high-income groups for these purposes are also very similar. We can infer that the bank specializes in home loans.

### Income Group vs Loan Grant for First Time



As observed from the graph the number of people applying for first time loans from high income groups and low-income groups is almost the same. The number of people who repaid these loans are also almost equal.

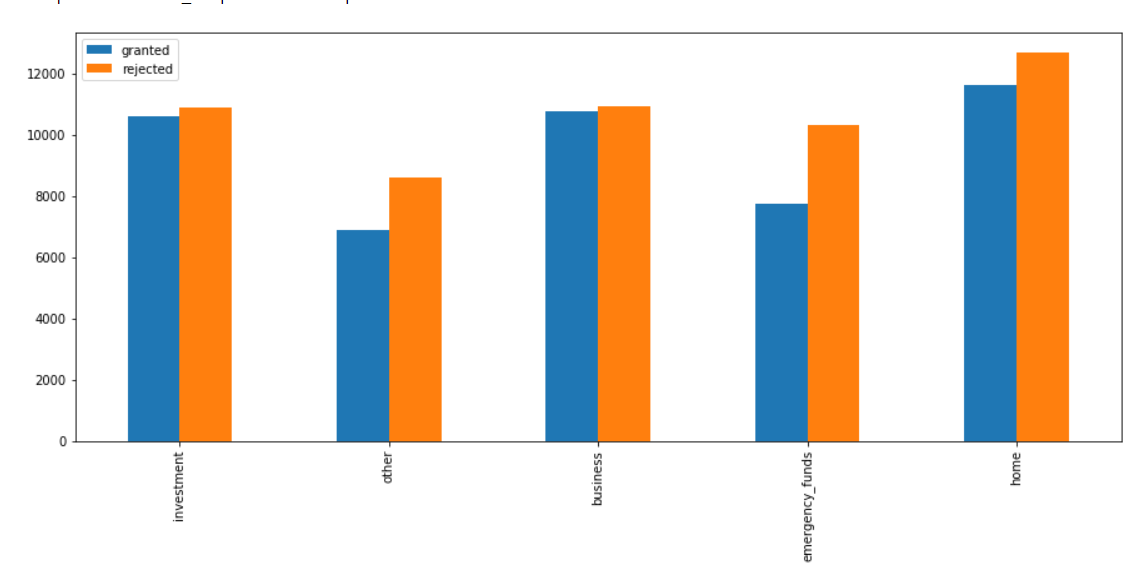
From high income group, 47% of the first-time loan applicants have been granted loans. From low income group, 43% of the first-time loan applicants have been granted loans.

From high income group, 65% of the first-time loan applicants have repaid their loans. From low income group, 64% of the first-time loan applicants have repaid their loans.

Loans which have been taken for emergency purpose have highest default rate for high income and low-income groups.

## Lending Pattern Based on Loan Purpose

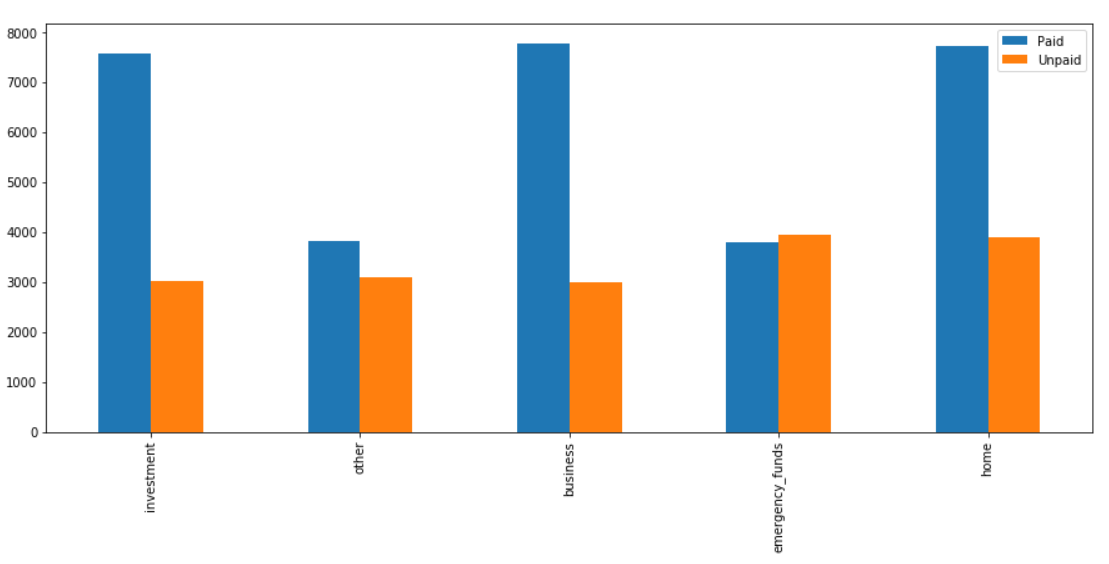
### Loan Purpose vs Number of Loans Given and Rejected



Graph shows the number of loans granted and rejected for different loan purposes. The following can be inferred from the graph.

* + As observed from the graph the maximum number of rejections are for loans asked for emergency purpose.
  + Maximum number of loans granted and rejected are home loans.
  + When compared to the number of loans granted, the number of rejects seem to be relatively high for ‘Other’ and ‘Emergency’ purpose.

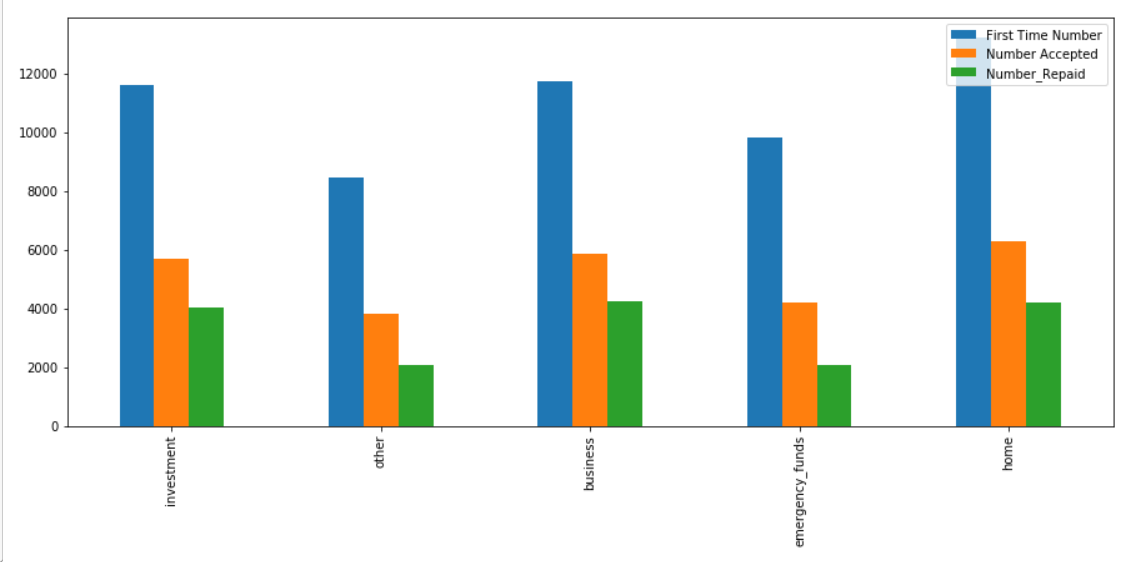
### Loan Purpose vs Loans Repaid



As observed from the graph, majority of people who take loans for investment, business and home repay their loans. But people who take loans for emergency tend to default on loan payments.

Out of all the customer that were granted loans for various purposes the maximum repayment rate is for investment and business which are at 72% followed by home loans.

### First Time Loans Analysis based on Purpose of Loan

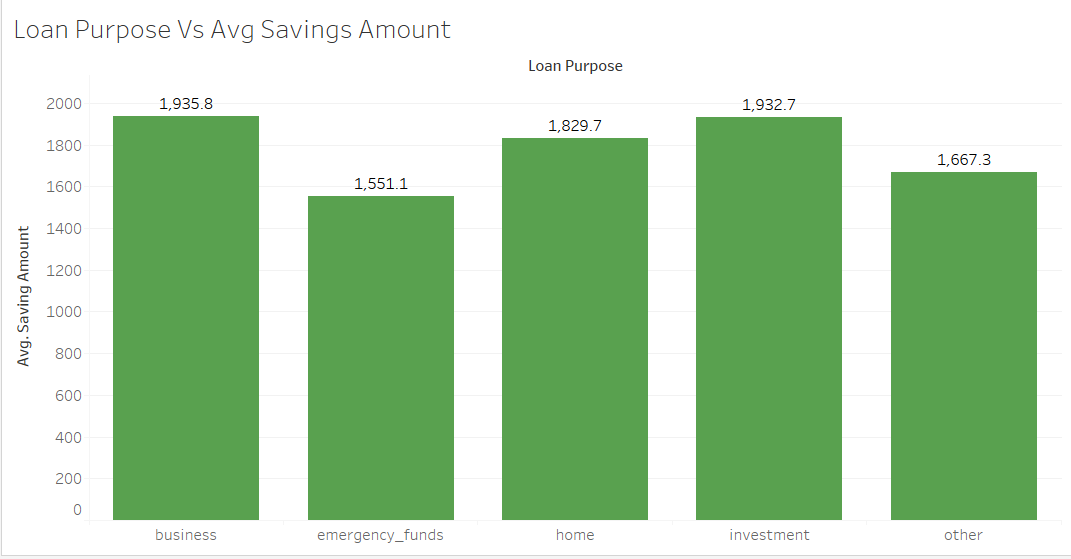


In the above graph, the blue bar represents first loan applicants. The green bar indicates the number of first loans approved. The Orange bar represents number of first loans repaid.

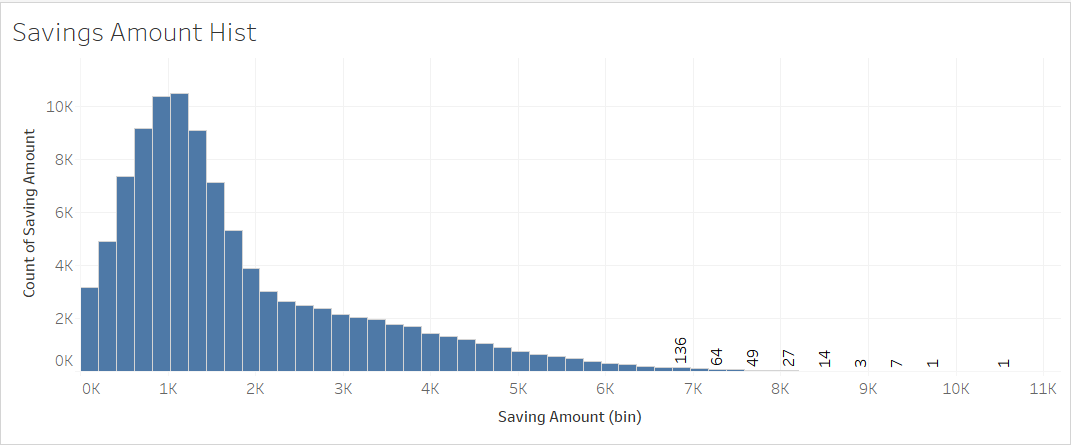
About 53% of the people who were given loans for investment purpose were first timers and out of those about 70% repaid their loans.

The bank, on an average grants’ loans to about 54% of people who apply for first time loans for various purposes. Out of those, the maximum repayment rate is for Business, followed by investment, followed by home. For other and emergency purposes the repayment rate is about 50%.

### Purpose of Loan vs Average Savings



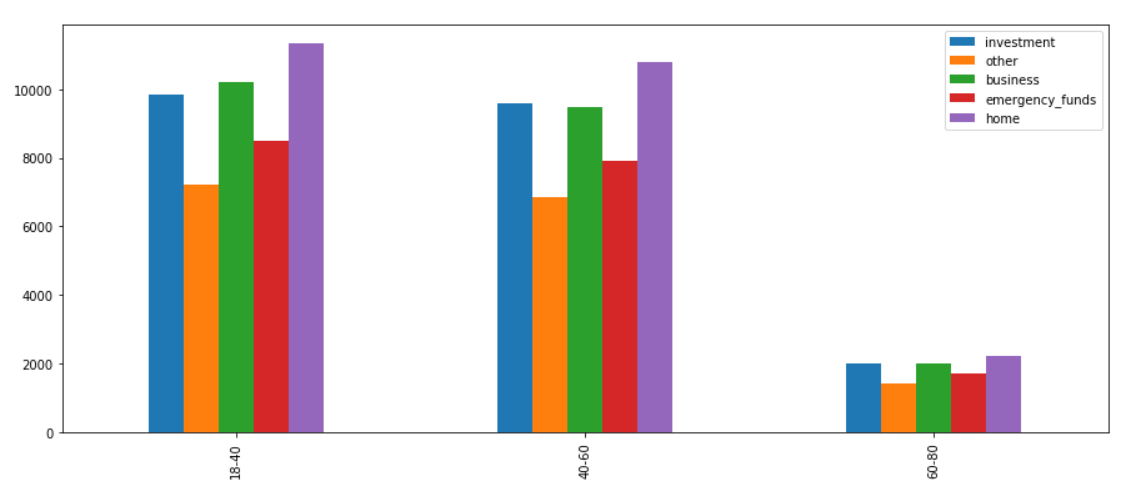
Highest repayment rate is for investment, business and home. As observed from the chart, more the savings more the probability of honouring the loan. This graph supports the fact **that more savings a person has the easier it is to repay the loan.**



As observed from the above histogram maximum number of people save around 1000.

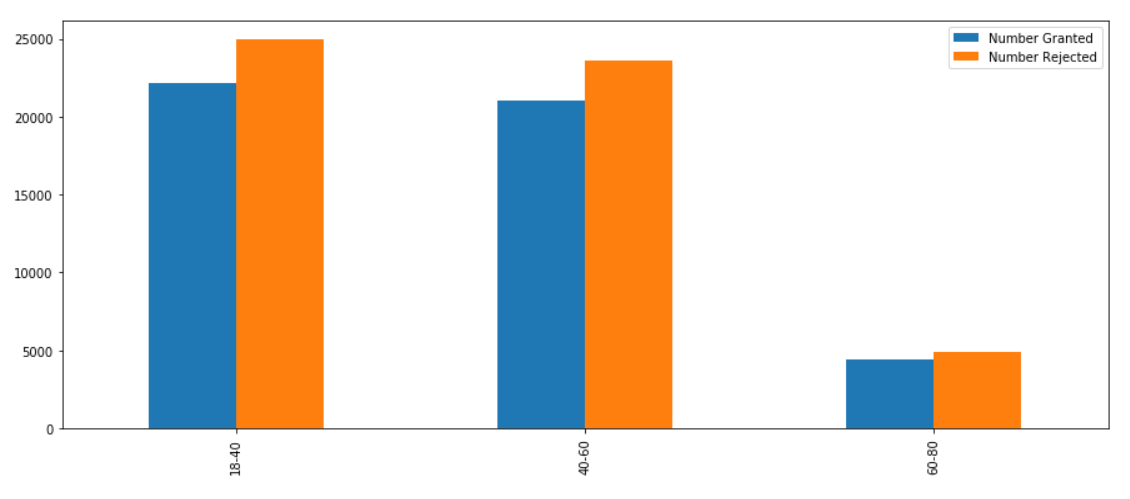
## Lending Pattern based on Age

### Age vs Loan Purpose



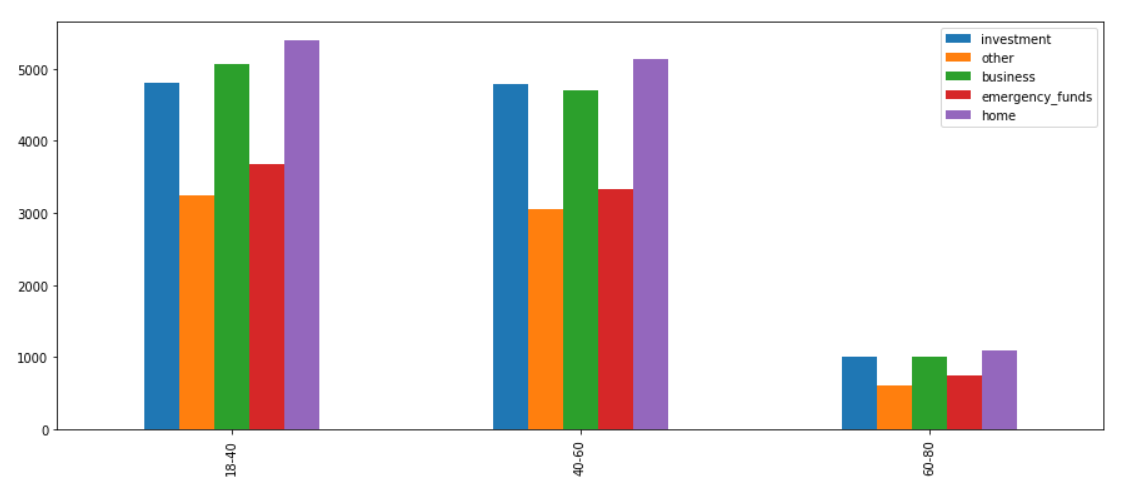
Maximum number of loans are applied by people in the age group of 18-40 out of which highest number is for home loans.

### Age vs Loans Granted and Rejected



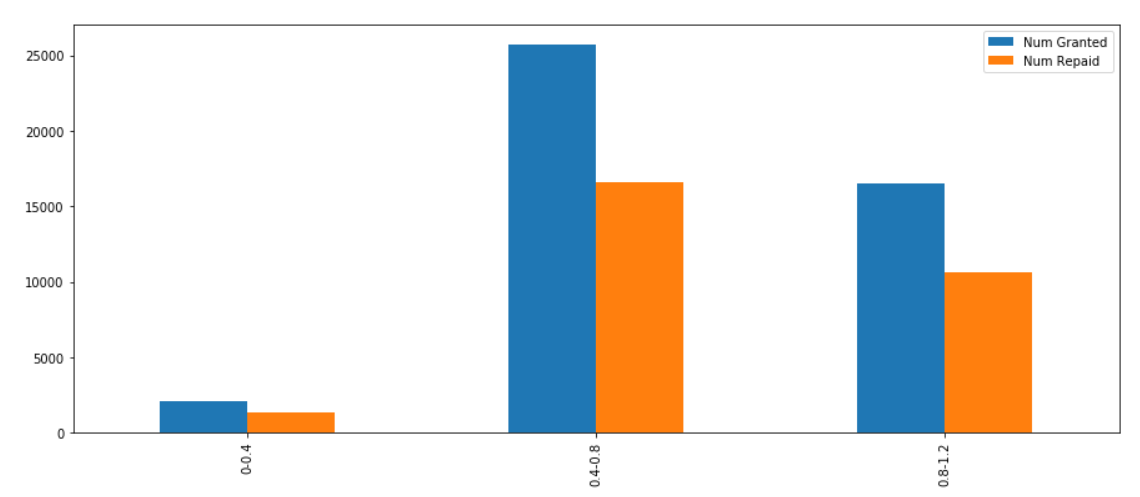
The average approval rate is about 47% across all age groups.

### Age vs Loans Granted for Various Purposes



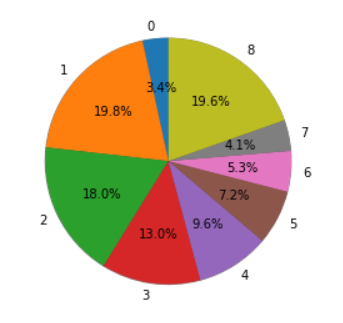
Home loan is the top category across all age groups.

## Lending Patterns based on Credit Card Limit Used



Maximum loans are being granted to customers who use credit card usage is in between 0.4-0.8 %. The overall credit utilization ratio is high for this group.

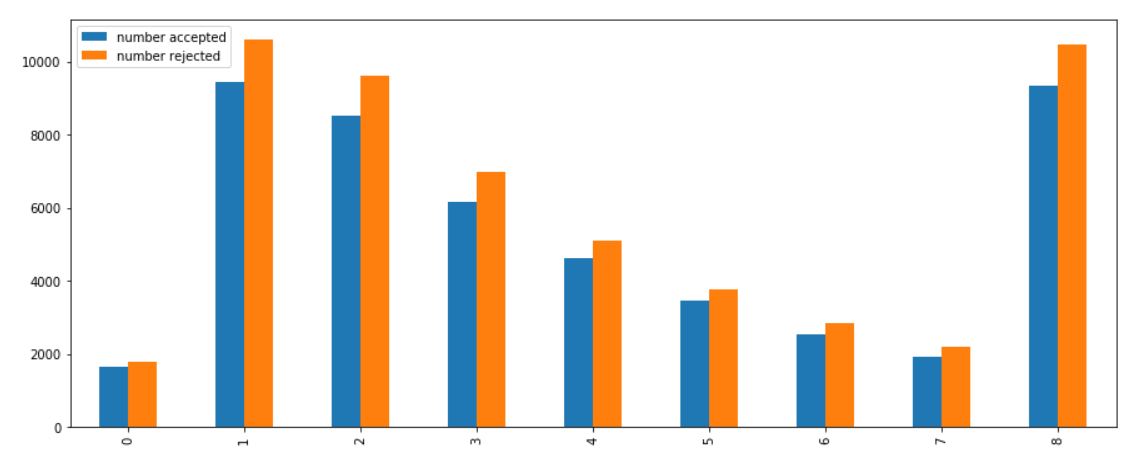
## Lending Patterns based on Number of Dependents



The pie chart shows the number of dependents of each customer. Maximum number of customers have only 1 person dependent on them.

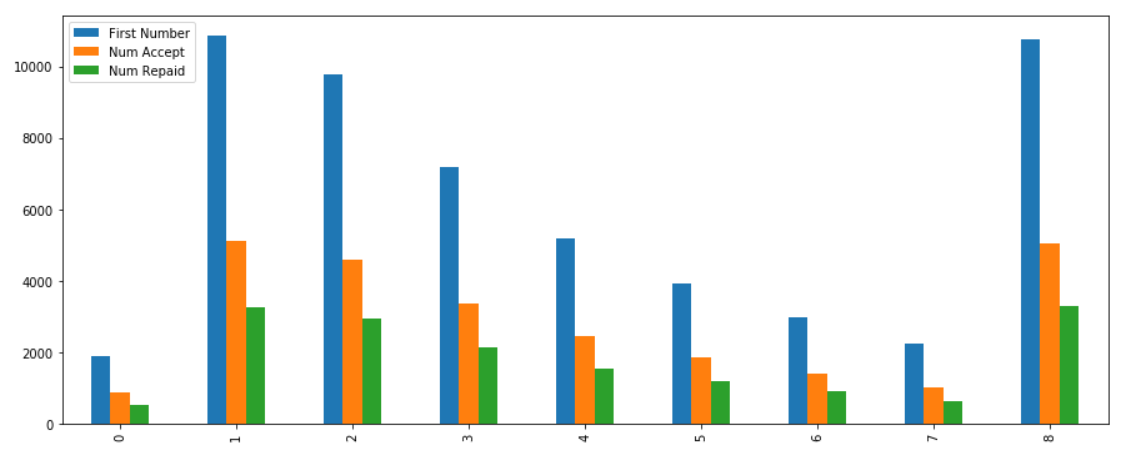
Second highest number of customers have 8 dependents.

### Number of Dependents vs Loans Granted and Rejected



As observed from the graph maximum number of loans are applied by people who have 1 dependent followed closely by people who have 8 dependents. The third place is occupied by people with 2 dependents.

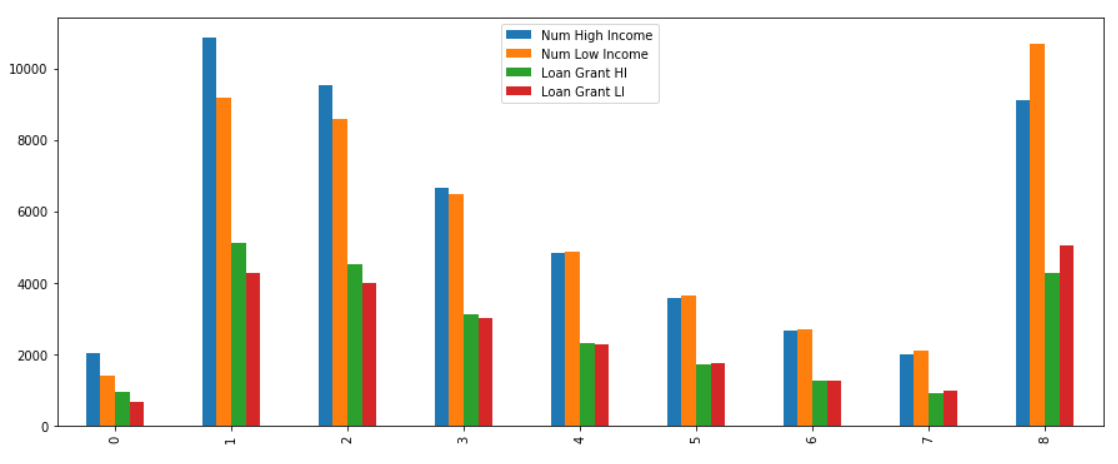
### First Time Loan Application, Loans Granted and Repaid



Maximum number of first time loan applications are received from customers with 1,2 and 8 dependents.

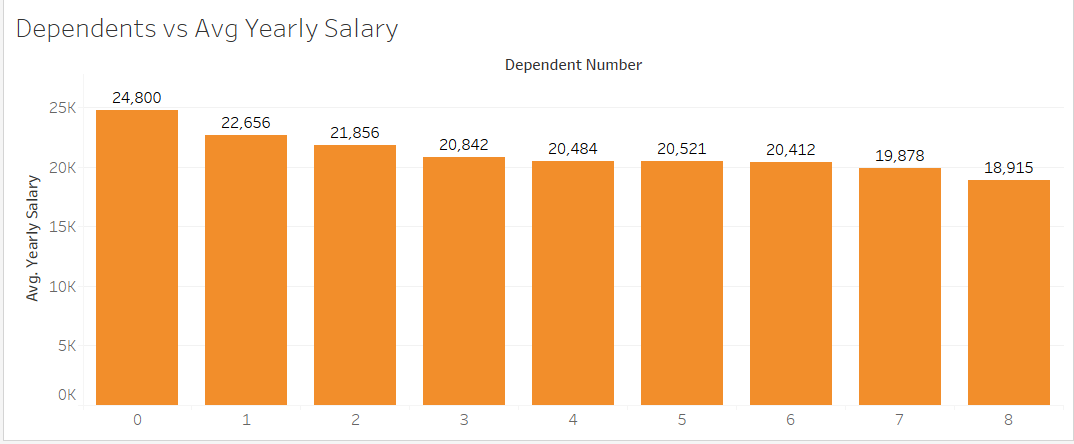
The average acceptance rate for first time loans is about 54%.It might suggest that the bank is conservative in granting first time loans.

### Number of Dependents vs Income



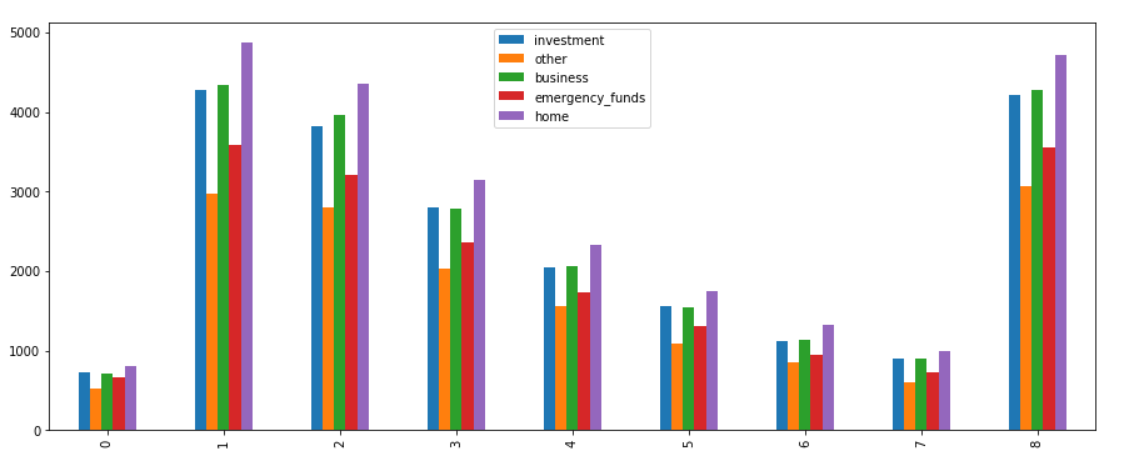
As observed from the graph, smaller family belong to the high-income group. As the number of family members increases the income decreases.

### Number of Dependents vs Yearly Salary



As observed from the graph, more the number of dependents less the savings. However, the average yearly salary for people with more dependents is comparable with that of families with smaller number of dependents, suggesting that people with higher number of dependents belong to the high-income group.

### Number of Dependents vs Loan Purpose



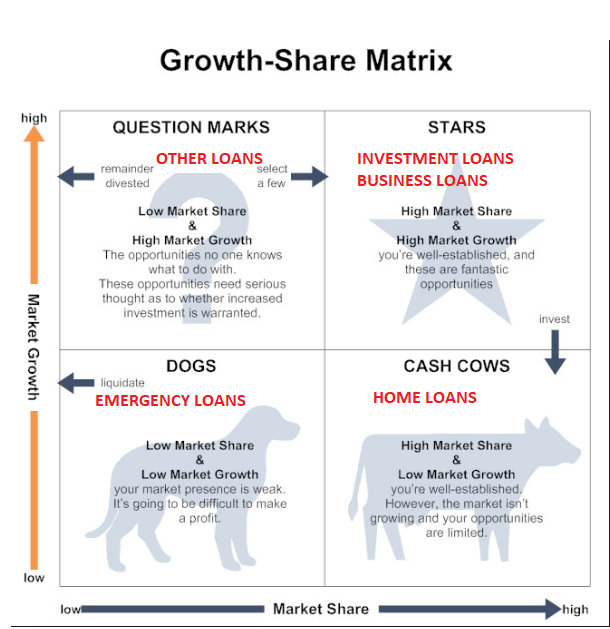
Maximum number of loans are applied for housing followed by business.

## Summary

Based on the lending patterns the following conclusions can be drawn.

### Understand the loan business of the bank.

1. Since only 50% of the first-time loan applicants are granted loans, this might imply that the bank is conservative towards first time loan applicants.
2. Number of loans granted to high income and low-income groups is almost the same suggesting that the bank caters to all economic segments of society.
3. Number of loans repaid by the high-income group and low-income groups are approximately the same which means that the loan amount approved by the bank commensurate with the customer’s repaying capacity.
4. The maximum loans are being granted for the purpose of housing followed by business and investment. The least is for the ‘Other’ category. Here ‘Other’ category may mean small consumer loans or personal loans.
5. Below is the depiction of the BCG matrix for the various loan categories



### Identifying customers who haven’t been granted a loan which are more likely to repay the loan and provide loans to them to maximize the profits of the bank.

Please find attached the list of customers (loan\_id) to whom no loans have been granted and are likely to repay the loan. There are total of 18316 customers to whom loans can be granted.



## Recommendations to Maximize Bank’s Profits

### Bank Should Continue to do

1. They should continue to give more investment and business loans. The repayment rate for these categories and quite high.
2. Continue to reject loans to people who have exceeded their credit card limit.
3. Continue to give less home loans to people above the age of 70.

### Bank Should Discontinue

1. The bank should either avoid giving emergency loans as the emergency loan category has maximum number of defaulters or develop strategies to ensure that emergency loans are repaid timely. One of the measures could be to have collateral security.
2. Discontinue giving loans to unemployed segment as their repayment rate is low.
3. Discontinue giving loans to customers with 0 credit limit. 903 customers with 0 credit limit have been granted loans and only 156 have repaid them.
4. Discontinue giving loans to customers who have exceeded their credit card limit. 994 customers with exceeded credit card limits have been granted a loan and out of which only 220 repaid.

### New Strategies that the Bank could Implement

1. Granting more loans to people with high income and low income for the purpose of investment, home and business.
2. Accepting more first-time loan applications.
3. Loans in the ‘Other’ category can be increased but measures should be taken that the defaulters are less. For loans like gold loans the interest rate could be lowered so as to attract more customers.
4. Increase number of loans given to employed people with significant saving and checking amount.
5. Grant more loans to people who have already repaid their previous loans and currently repaying any other loans.

# Process of Forecasting

Models were evaluated to maximize profits of the bank when providing loans.

## Step 1: Importing the Required Libraries

The following Libraries are imported:

1. Pandas - Pandas is a software library written for the Python programming language for data manipulation and analysis.
2. NumPy - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
3. Matplotlib – Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.

## Step 2: Importing Data and Preliminary Analysis

Pandas library has a function which has the function to read csv file. The function used id pandas.read\_csv(‘file\_name.csv’)

For preliminary analysis the function ‘.info()’ or ‘.describe()’ can be use.

1. ‘.info()’ – It gives information about the number of values in each column of the data set.
2. ‘.describe()’ – It gives detailed information about the data set like the number of values, mean, standard deviation etc.

The above-mentioned functions only work for numerical data. If the data set has any strings or alpha-numeric values they are ignored.

## Step 3: Exploratory Data Analysis

In this step the library Matplotlib, Pandas Data Frame were extensively used. The library NumPy was used to get data. The extracted data was then put into data frame and interesting insights from the data analysis were noted.

## Step 4: Data Manipulation and Preparation

In this step the data set is prepared for modelling. The following is done for data preparation.

1. All the text data is replaced with numerical data. In the given data set the following values were given to various loan purposes
   1. Investment = 1
   2. Other = 2
   3. Business = 3
   4. Emergency = 4
   5. Home = 5
2. Another process that is carried out in data manipulation and preparation is filling any unavailable data. In the given data set, places where data was not available was filled with 404.
3. The rows that are not required for analysis are dropped.

## Step 5: Modelling

In the case study a library called Scikit-Learn (sklearn) is used for modelling. Scikit-learn is a free software machine learning library for the Python programming language. It is extensively used for predictive analysis and predictive modelling.

The following is done to fit the data into a suitable model.

1. X, y are defined. X has all the values that are required to predict the target variable y.
2. They are then divided in training set and testing set. In the given problem 20% of the data is considered for testing set, whereas the remaining 80% is the training set.
3. The data is fit and then error is calculated.
4. The model with least error is selected.

In the problem at hand the data was fit into Decision Tree and classification report was generated for both training and testing data.

## Testing

For testing the correctness of the model **in-sample forecasting** was performed. In-sample forecasting is forecasting an observation that was part of the data sample.

The following cases were checked:

1. Repayment of loan after the loan was granted.
2. Defaulting on loan when the loan was granted.
3. Repayment was not possible because the loan was not granted

## Forecasting

Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends.

In the given problem the bank wanted to identify the customers who haven’t been granted a loan which are more likely to repay the loan and provide loans to them.

From the given data set the data of people who have been granted a loan was removed. Assuming the that the customers who haven’t been granted a loan were granted a loan forecasting to predict repayment of the loan.

If the probability of repayment of loan was greater than 70% those customers were selected.

# Result

1. The loan business of the bank was understood from their lending patterns.
2. About 18316 customers were identified.