**MIS 587**

**Group 5**

****

**Analyzing Loan Repayment Behavior**

Megan McKenna

Apurva Atalkar

Ahhaan Badhwar

Priyal Devanshu Dani

Rashmi Rao

Table of Contents

[1. Overview of Client and Problem Statement 3](#_Toc165643008)

[2. Dataset Description 4](#_Toc165643009)

[3. Data Warehouse Design and Implementation 6](#_Toc165643010)

[4. Data Preparation 9](#_Toc165643016)

[5. Data Exploration 12](#_Toc165643020)

[6. Data Analysis and Results 15](#_Toc165643021)

[7. Business Implications 18](#_Toc165643022)

[8. Appendix 19](#_Toc165643023)

# Overview of Client and Problem Statement

Home Credit is a global consumer finance provider that focuses on providing loans primarily to people with little or no credit history. They operate in various countries, particularly in emerging markets such as Asia, Eastern Europe, and Central Asia. Home Credit typically offers installment loans, cash loans, and revolving loans for purchasing consumer durables, such as electronics, appliances, and other goods.

Their business model often involves providing loans directly at the point of sale, allowing customers to make purchases on credit with minimal hassle. They utilize alternative data sources and advanced risk assessment techniques to evaluate the creditworthiness of individuals who may not have traditional credit histories. The company has grown significantly in recent years and has become one of the leading consumer finance providers in many of the markets where it operates.

In emerging markets, access to credit is vital for individuals and businesses to seize economic opportunities and foster growth. However, a significant obstacle hinders this access: the lack of detailed credit histories for most people. Traditional credit scoring mechanisms, which rely heavily on credit history, often fail to adequately assess the creditworthiness of individuals in these markets. This deficiency in credit assessment systems manifests in two critical problems:

1. **Limited Access to Credit**: Many individuals are unable to obtain credit at all due to the absence of a credit history. This denies them the opportunity to invest in education, entrepreneurship, or other ventures that could improve their lives and contribute to economic development.
2. **Unsustainable Loan Terms**: For those who do manage to secure loans, they often receive terms that are unsustainable, leading to default and financial instability. Without accurate assessments of creditworthiness, lenders may either deny credit altogether or offer loans with excessively high interest rates or stringent repayment terms, exacerbating financial stress for borrowers.

To address these challenges, there is an urgent need to develop alternative methods for evaluating creditworthiness that do not rely solely on traditional credit history. By leveraging innovative data analysis techniques and machine learning algorithms, we aim to predict loan repayment ability based on a broader set of factors beyond credit history.

Our approach involves utilizing the Home Credit Default Risk dataset, which provides comprehensive information on loan applicants, including demographic data, financial transactions, and other relevant variables. We will build a robust data warehouse and apply advanced analytical techniques to identify patterns and indicators of creditworthiness within this dataset. By considering factors like income, employment status, and spending behavior, we aim to provide financial institutions with a more accurate and reliable tool for assessing the credit risk of applicants.

The successful implementation of this approach has the potential to significantly enhance financial inclusion in emerging markets. By enabling financial institutions to make more informed lending decisions, we can expand access to credit for underserved populations, empowering individuals, and businesses to pursue economic opportunities and improve their livelihoods.

# Dataset Description

**Loan Application Fact Table:** The table describes details about every loan application submitted.Each row represents one loan in the data sample.

|  |  |  |
| --- | --- | --- |
| **Row Name** | **Data Type/Structure** | **Description** |
| SK\_ID\_CURR | Primary Key | ID of the loan |
| SK\_ID\_BUREAU | Foreign Key | ID of previous credit bureau credit related to the loan |
| SK\_ID\_BALANCE\_PREV | Foreign Key | ID of previous credit in home credit related to the loan |
| SK\_ID\_PREV | Foreign Key | ID of previous credit in home credit related to the loan |
| CUSTOMER\_ID | Foreign Key | ID of the customer |
| TARGET | - 1 is a client with payment difficulties, 0 is all other cases. | Target Variable |
| NAME\_CONTRACT\_TYPE |  | Identification if loan is cash or revolving |
| AMT\_INCOME\_TOTAL |  | Income of the applicant |
| AMT\_CREDIT |  | Credit amount of the loan |
| AMT\_ANNUITY |  | Loan annuity |

**Bureau Dimension:** All client’s previous credits provided by other financial institutions that were reported to the Credit Bureau.

|  |  |  |
| --- | --- | --- |
| **Row Name** | **Data Type/Structure** | **Description** |
| SK\_ID\_CURR | Foreign Key | ID of the loan |
| SK\_ID\_BUREAU | Primary Key | ID of previous credit bureau credit related to the loan |
| CREDIT\_ACTIVE |  | Status of the credit bureau reported credits |
| AMT\_CREDIT\_SUM\_LIMIT |  | Current credit amount for the credit bureau credit |
| AMT\_CREDIT\_SUM\_DEBT |  | Current debt on credit bureau credit |
| CREDIT\_TYPE |  | Type of credit bureau credit (Car, Cash, etc.) |
| CREDIT\_DURATION |  | Length of time for the credit bureau credit |
| CURRENT\_DEBT\_TO\_  CREDIT\_RATIO |  | Debt to credit ratio that is contained within credit bureau |

**Credit Card Balance Dimension:** Monthly balance point in times of previous credit cards that the applicant has with home credit.

|  |  |  |
| --- | --- | --- |
| **Row Name** | **Data Type** | **Description** |
| SK\_ID\_PREV | Primary Key | ID of previous credit in home credit related to the loan |
| MONTHS\_BALANCE | -1 is the freshest balance date | Month of balance relative to the application date |
| AMT\_RECIVABLE |  | Amount receivable on the previous credit |
| NAME\_CONTRACT\_STATUS |  | Contract status (active, signed, etc.) on the previous credit |
| AMT\_INTEREST\_RECIEVABLE |  | Amount of the interest receivable |

**Previous Application Dimension:** All previous applications for home credit loans of clients who have loans in the sample.

|  |  |  |
| --- | --- | --- |
| **Row Name** | **Data Type** | **Description** |
| SK\_ID\_PREV | Primary Key | ID of previous credit in home credit related to the loan |
| NAME\_CONTRACT\_TYPE |  | Contract product type (Cash loan, consumer loan, home loan, etc.) of the previous application |
| AMT\_ANNUITY |  | Annuity of previous application |
| AMT\_CREDIT |  | Final credit amount on the previous application |
| AMT\_DOWN\_PAYMENT |  | Down payment on the previous application |
| RATE\_INTEREST\_PRIMARY |  | Primary interest rate |
| NAME\_CLIENT\_TYPE |  | Type of client (New, Repeater, etc.) |

**Customer Dimension**: The customers that are applicants for home credit or loans in the system.

|  |  |  |
| --- | --- | --- |
| **Row Name** | **Data Type** | **Description** |
| CUSTOMER\_ID | Primary Key | ID of the customer |
| GENDER |  | Gender of the customer |
| NAME\_EDUCATION\_TYPE |  | Type of education that the customer received |
| AGE |  | Age of the customer |
| YEARS\_EMPLOYED |  | Years the customer has been working |
| INCOME\_BAND |  | Income group that the customer fits into |
| CNT\_FAM\_MEMBERS |  | Number of family members for the customer |
| NAME\_FAMILY\_STATUS |  | Married, Single etc. |
| OCCUPATION\_TYPE |  | Occupations of the client |
| ORGANIZATION\_TYPE |  | Organization which a client belongs to |

# Data Warehouse Design and Implementation

## Dimensional Modeling Process

|  |  |
| --- | --- |
| **Identify Business Process** | Determining the credit worthiness of applicants.  The business process would be assessing whether or not an applicant is able to repay their loan based on their limited credit. |
| **Declare the Grain** | Each row represents a single loan application. |
| **Identify Facts** | AMT\_CREDIT, AMT\_INCOME\_TOTAL, AMT\_ANNUITY, PERCENT\_CREDIT\_INCOME |
| **Identify Dimension** | AGE, GENDER, OCCUPATION\_TYPE, NAME\_EDUCATION\_TYPE |

## A diagram of a credit card Description automatically generated Star Schema

## Details of Warehouse Design and Implementation

For the Warehouse design, we did some transformations in Python as well as in SSIS.

A screenshot of a computer code

Description automatically generatedIn Python, we created 2 new columns namely Percent Annuity Income and Percent Credit Income. Percent Annuity Income indicates the loan annuity amount relative to the person's total income. Percent Credit Income indicates the income relative to the credit.

A screenshot of a computer

Description automatically generated

## 

## A screenshot of a credit report Description automatically generated

We imported flat files in SSMS in database named HomeCredit\_Group5. In SSIS, we created control flow to first truncate the table data that already exists in HomeCredit\_DW\_Group5. Next the control flow is designed for data to be loaded from source (HomeCredit\_Group5) to destination (HomeCredit\_DW\_Group5) for 4 dimensions namely, Previous application, Credit card Balance, Customer and Bureau and then into fact table which is Loan Application Fact table.

A screenshot of a computer

Description automatically generated

In the Bureau data flow, we created 2 columns CURRENT\_CREDIT\_DEBT\_RATIO (which is difference of AMT\_CREDIT\_SUM - AMT\_CREDIT\_SUM\_DEBT) and CURRENT\_DEBT\_TO\_CREDIT\_RATIO (AMT\_CREDIT\_SUM\_DEBT/(AMT\_CREDIT\_SUM+0.01)).

* 0.01 is added in the denominator because for cases where AMT\_CREDIT\_SUM is 0 the value will be undefined. Next, we did a conditional split on column CURRENT\_DEBT\_TO\_CREDIT\_RATIO to filter only those values where ratio is less than equal to 1.

A screenshot of a computer

Description automatically generated

Finally, in the data flow of Loan Application we connected lookups from all dimensions and finally loaded data into Loan Application Fact table.

A screenshot of a computer

Description automatically generated

Screenshots of successful loading of data into HomeCredit\_DW\_Group5 database.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

# Data Preparation

## Outlier Detection and Removal

**Detection** – By looking at the descriptive statistics for both the features, we identified that they have outliers.

A screenshot of a number of numbers

Description automatically generated

* The attributes – DAYS\_EMPLOYED & AMT\_INCOME\_TOTAL
* DAYS\_EMPOLOYED - How many days before the application the person started current employment.
* AMT\_INCOME\_TOTAL - Income of the client, taken from the Loan Application (Fact) Table
* YEARS\_EMPLOYED is derived from DAYS\_EMPLOYED and is later dropped.
* DAYS\_EMPLOYED contains inconsistent data. It contains negative values (since it is in the form of 'days' relative to the application), then suddenly there is one very big positive number (365243 days - which equates to 1000 years!) with a significant number of individuals having it.
* Similarly, AMT\_INCOME\_TOTAL has an outlier. As per the data, the 75th percentile is equal to 202,500, while the maximum value is equal to a one hundred seventeen million. It is quite a big difference as it means that 75% of our customers are already making a total income of 202,500 or below - the remaining 25% have a total income higher than 202,500. AMT\_INCOME\_TOTAL has a wide range of values, from min value 25,650 to max value117,000,000.
* Assumption for currency is INR.

**Removal**

AMT\_INCOME\_TOTAL

* Outliers are usually greater than 3 standard deviations away from the mean.
* We calculated a threshold and decided to retain only those rows which are less than the calculated threshold and consider only those rows where DAYS\_EMPLOYED is less than or equal to 0.

DAYS\_EMPLOYED

* Convert anomalous data in DAYS\_EMPLOYED which is 365243 days (1000 years) to -14600 days (equal to 40 years)
* It’s safe to assume that he might retire after working for 50 years.
* The assumption is taken on the basis that the dataset has values where an individual has been employed for 49 years. And after that there is a jump to 1000 years.

*Note: About 18% of the outliers were removed from the dataset.*

A screenshot of a computer

Description automatically generated

A white rectangular sign with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

## Missing values

Detection

* The attributes – EXT\_SOURCE1, EXT\_SOURCE2, EXT\_SOURCE3 & OCCUPATION\_TYPE
* The external sources are in the fact table, they were highly correlated to Target, which is why it’s an important feature.
* Occupation Type is the occupation of the clients. Examples are Working, Commercial associate, Businessman etc. Occupation Type has 31% of its values missing. Most of the null values are from pensioners.

Before Missing Values Removal

A screenshot of a computer code

Description automatically generated



Removal

* For the continuous variables, we used the mean to fill in missing values, as mean imputation is well-suited for numerical data.
* Created a new category ‘Data Not available’ to fill in the missing values.

A screenshot of a computer program

Description automatically generatedAfter Missing Values Removal

A screenshot of a computer code

Description automatically generated

## Label Encoding

* The attributes CODE\_GENDER, NAME\_EDUCATION\_TYPE, ORGANIZATION\_TYPE.
* Transformed the categorical columns using Label Encoding. Considering the space, our dataset is large enough which is why we decided to not do One-hot encoding.

Before Encoding

A close-up of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

# Data Exploration

1. **Analysis on Gender**

A graph with red squares

Description automatically generated

* More women have applied for loans (about twice the number of men).
* But looking at the target variables, men default more than women. Around 10% of the total men default compared to 7% for women applicants.

1. **Analysis on Family Status**

A screenshot of a graph

Description automatically generated

* Married people have applied for a large proportion of loans. But the single, civilly married, and separated applicants have a higher chance of being defaulters.

1. **A group of blue and orange dotted graphs

   Description automatically generatedA screenshot of a computer screen

   Description automatically generatedCorrelation Analysis**

* More women have applied for loans
* But men default more than women
* AMT\_CREDIT - Credit amount of the loan.
* AMT\_GOODS\_PRICE - For consumer loans it is the price of the goods for which the loan is given.
* AMT\_ANNUITY - annuity is the monthly due amount.
* AMT\_CREDIT and AMT\_GOODS\_PRICE are highly correlated (scoring 0.99) and has a positive linear slope - because as the price of goods for which the loan is given gets higher, the credit amount of the loan gets higher too.
* AMT\_CREDIT and TARGET are correlated; it is a good indicator to predict loan.

1. **Distribution of income for different Occupations**

**A blue and orange lines

Description automatically generated**

* The plot shows a wide range of incomes across different occupations. Some occupations like Managers, IT staff, and High skill tech staff appear to have higher income ranges compared to occupations like Laborers, Cleaning staff, and Low-skill laborers.
* Defaulters are spread across all income brackets, there seems to be a noticeable presence of defaulters in the lower to middle income range across various occupations.
* This might suggest that individuals in these income brackets could be more vulnerable to defaulting, possibly due to less financial stability.
* Right off the bat, it seems as if the laborers have the highest difficulty in repaying. This is not a better way to conclude, because this contains biased number of applicants. A better way is to find a metric that incorporates relative relationship between applicants count and repayors count.

A graph of different colored bars

Description automatically generated with medium confidence A screenshot of a document

Description automatically generated

* According to the ratio of Number of repayors to Number of applicants in every occupation type, we see that it is most safe to lend money to Accountants with an R/A ratio of 0.9516 and it is least safe to lend money to low skilled laborers with an R/A ratio of 0.8284.

1. **Analysis on Income Band**

A screenshot of a computer screen

Description automatically generated

A graph of a number of individuals with blue bars

Description automatically generated

* Created income bands to see the distribution of incomes and which income band defaults the most.
* The income band between 130k and 160k default the most.

# Data Analysis and Results

1. **Risk Assessment for the Lender**

* The line chart shows debt-to-credit ratio by age.
* Target audience: lenders (real estate, banking, home loans).
* Visual helps lenders predict loan repayment likelihood.

A graph of a graph

Description automatically generated with medium confidence

* The debt-to-credit ratio compares total liabilities to total credit limits, with lower ratios indicating better financial health.
* The chart shows that younger age groups tend to have higher debt-to-credit ratios than older age groups.
* One of the reasons why younger people might default more is because they might be carrying student loan debt. Another reason is that younger people may have less established credit histories, which can make it difficult to qualify for lower interest rates on loans.

1. **Duration Assessment by Lenders**

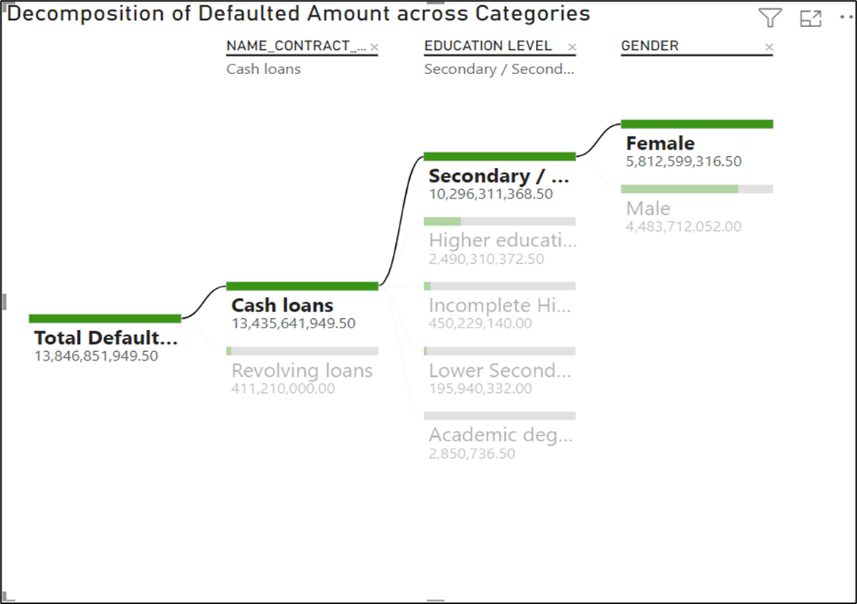
The funnel chart highlights the following insights:

* Mortgages and credit cards have the longest average durations (15.07 years and 13.91 years, respectively).
* A screenshot of a computer screen

  Description automatically generatedLoan durations vary based on the loan amount, with shorter durations observed for smaller loans. For instance, car loans between $30k and $65k have an average duration of 3.97 years, while loans for business development of less than $30k have an average duration of 3.62 years.

1. **Segmentation by Education Level & Gender**

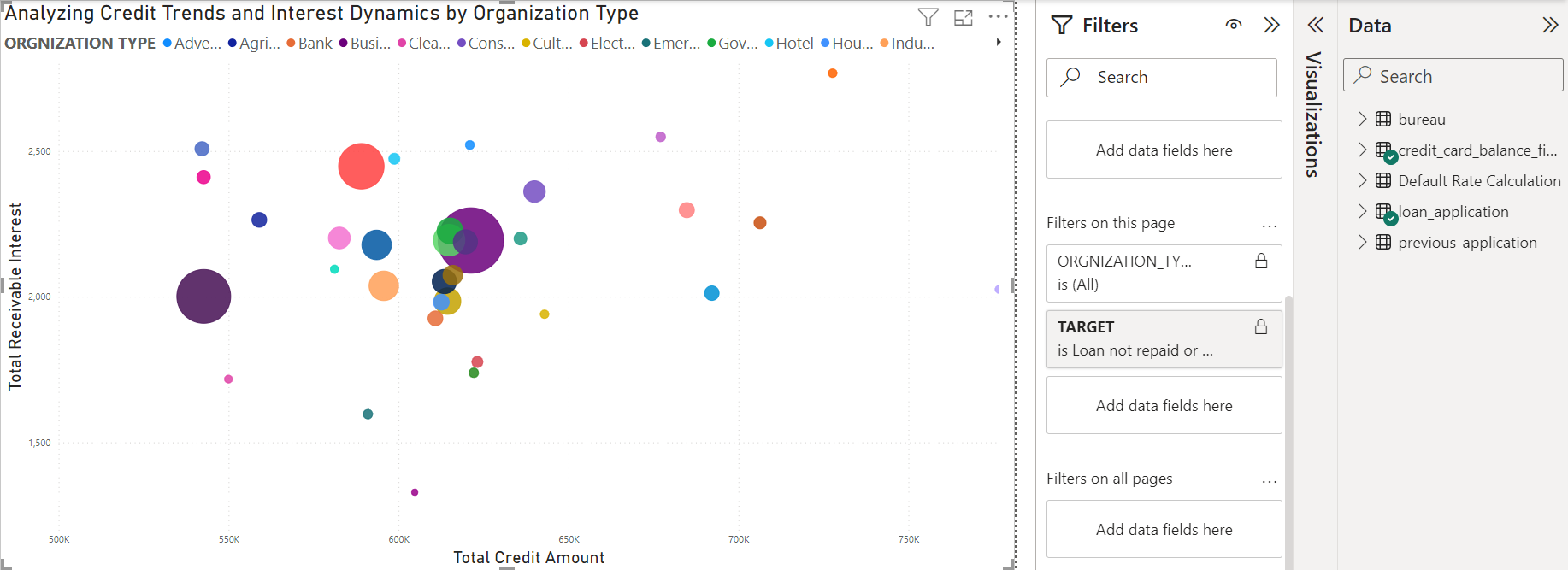
* We used decomposition tree to understand the different levels affecting the default amount. It shows the total defaulted amount broken down by education level and gender for two loan types: cash loans and revolving loans.
* Cash loans had higher default amounts compared to revolving loans, suggesting higher default risk due to larger loan amounts or longer repayment terms.
* Customers with secondary education had higher default amounts, indicating challenges in financial management or limited access to financial education resources.
* Females showed higher default rates compared to males within the analyzed segmentation.



1. **Analyzing Credit Trends & Interest Dynamics By Org Type**

A scatterplot was created showcasing that the data points are the different organization types. The size of the circle represents the count of the number of organizations, and the color represents the average of AMT Credit and AMT Receivable.

The filters on the chart are - Average of AMT CREDIT, Average of AMT RECEIVABLE, Count of ORGANIZATION TYPE, ORGANIZATION TYPE- (the organization in which the individuals work in)



The presence of realtors, insurance companies, and universities with high total receivable interest indicates that the individuals in these organizations may represent high-interest borrowers. Despite their small representation in the borrower base, there are opportunities for growth and expansion in lending relationships with these individuals. Home credit may explore strategies to deepen relationships and capture additional market share.

# Business Implications

1. **Enhanced Risk Evaluation:** Implementing alternative credit assessment methods lowers default rates by improving risk evaluation. Loan terms need to consider a wider range of factors other than traditional credit history. The analysis shows differences in default rate in gender, across single and civilly married people, less educated people, and lower skill professions which embarks upon the fact that there is a bigger picture to consider than credit history. This shows the scope for a more holistic analysis of applicants when processing loans with the goal of reducing loan default while also increasing access to capital for the masses.
2. **Loan Duration Strategies:** Our analysis highlights that for smaller loan amounts, shorter durations are preferred, indicating a need for flexible repayment options and quicker turnaround for smaller loans to cater to customer preferences and market demands. On the other hand, for mortgages and credit cards, which have longer durations, lenders can focus on offering competitive interest rates and value-added services to attract and retain customers over the extended loan periods.
3. **Tailored Loan Solutions:** The analysis shows the window of opportunity for tailoring loan products to address loan repayment and borrowing disparities. As seen in the analysis, there is a high default rate in younger demographics, low skilled occupations, and less educated groups. This reveals an opportunity for tailored products for such demographics with flexible repayment options such as graduated repayment plans, income-driven repayment plans, or extended repayment terms, integrated debt consolidation services which will consolidate high-interest debts into a single cumulating to lower-interest loan which can simplify repayment and reduce overall borrowing costs, and loan products with built-in credit building features, such as reporting loan payments to credit bureaus or offering secured credit cards as part of the loan package.
4. **Customer Education:** Our analysis correlates loan repayment success with age, education levels and professions. As seen in our analysis, consumers belonging to certain low skilled professions, less educated demographics, and younger age groups face challenges in repayment.
5. **Competitive Advantage:** Our analysis reveals disparities in access to credit and loan repayment across demographics. Younger age groups carry higher debt-to-credit ratios, while men are less represented among borrowers and have lower success rates. Additionally, though married applicants dominate loan applications, singles, civilly married, and separated individuals are more prone to default. These disparities underscore the need for remedial actions outlined in points 2, 3, and 4, offering a competitive edge and fostering brand trust and loyalty.
6. **Targeted Marketing**: Analyzing customer demographics reveals key insights into loan repayment success, with women, specific professions, education levels, family statuses, and organizations being overrepresented among borrowers and underrepresented among defaulters. This reveals the potential for targeted digital marketing and outreach efforts.

# Appendix

Screenshots of all the tables from data warehouse database:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated