Loan Default Analysis

Lending Club Case Study

Submitted by,
Narayana Manikarnike
Geevanandam Murugesan

Table of Contents

Objective

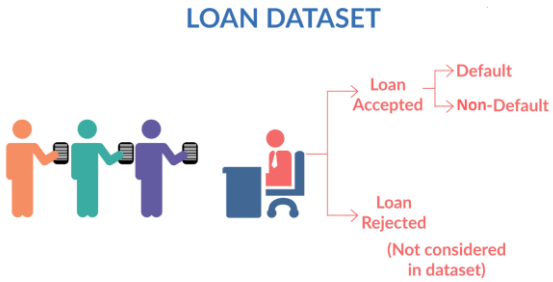
Problem Statement

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:

- **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
- **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
- **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)



Expected Results:

Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

Data Overview

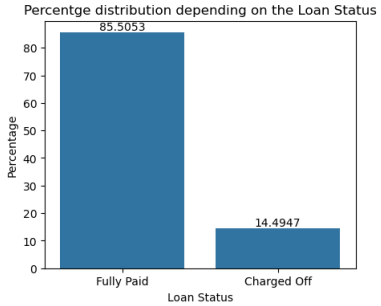
- The dataset includes various attributes related to loan applications, such as loan_amnt, int_rate, annual_inc, and loan_status. This information is crucial for performing an in-depth analysis to identify patterns and relationships that might indicate a risk of default.
- **Code Snippet**:

```
# Reading Loan.csv file.
df_loan = pd.read_csv("loan.csv")
df_loan.head()
```

- **Data Sample:**

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length	home_ownership	annual_inc	verification_status	issue_d	loan_status	pymnt_plan	url	desc	purpose	
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	NaN	10+ years	RENT	24000.0	Verified	Dec-11	Fully Paid	n	https://lendingclub.com/browse/loanDetail.acti...	Borrower added on 12/22/11 > I need to upgra...	credit_card	C
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	Ryder	< 1 year	RENT	30000.0	Source Verified	Dec-11	Charged Off	n	https://lendingclub.com/browse/loanDetail.acti...	Borrower added on 12/22/11 > I plan	car	

- **Analysis based on Loan Status**



Observation:

About 14.5% of the total applicants end up becoming defaulters.

Data Cleaning

- **Handling Missing Data**: Columns with significant null values were removed to ensure the analysis is based on complete and reliable data.
  - Removing columns which contain only null values.
  - Removing the records for the customers with the Loan Status as 'Current'.
  - Removing columns with high amount of missing values.
  - Identifying columns which will be used in the analysis.
  - We are ignoring the behavioural columns and the columns which have only 1 unique value.
  - Identify percentage of missing values.
  - Remove rows with null values for columns 'emp_length' and 'pub_rec_bankruptcies‘
  - Converting the 'int_rate' column from string to float
  - Converting the 'pub_rec_bankruptcies' column from float to integer
  - Creating Month and Year columns from the issue date
- **Categorical Column Creation:** Continuous variables like loan_amnt, int_rate, and annual_inc were segmented into categorical columns to facilitate easier analysis.
  - Identifying quartiles for the Loan Amount
  - Creating a categorical column based on the Loan Amount
  - Identifying quartiles for the Interest Rate
  - Creating a categorical column based on the Interest Rate
  - Identifying quartiles for the Annual Income
    - Creating a box plot to check the outliers for Annual Income
    - Since there are very high outliers for the Annual Income, we will be removing those records.
  - Creating a categorical column based on the Annual Income
  - Identifying the numerical and categorical columns

Numerical Columns

loan_amnt
funded_amnt
funded_amnt_inv
int_rate
installment
annual_inc
dti

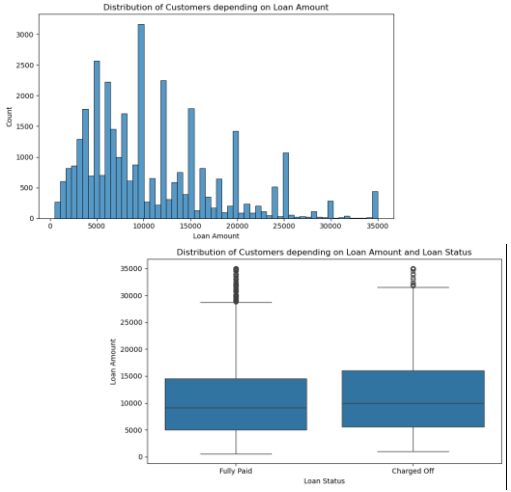
Categorical Columns

term
grade
emp_length
home_ownership
verification_status
loan_status
purpose
addr_state
issue_year
loan_amnt_cat
int_rate_cat
annual_inc_cat
pub_rec_bankruptcies

Univariate Analysis

- **Loan Amount**: The distribution of loan amounts reveals common loan sizes and their frequencies.
- **Annual Income**: Variations in borrowers' annual income may influence their ability to repay loans.
- **Interest Rate**: Different interest rate ranges have been analyzed to understand their distribution and potential impact on defaults.

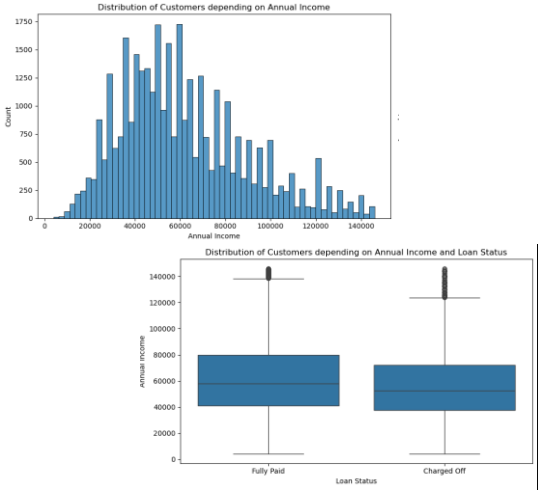
Univariate analysis on the Loan Amount



Observations

1. Most of the loan amount applied was in the range of 5K - 15K.
2. For Defaulters, most of the loan amount applied was in the range of 6K - 16K

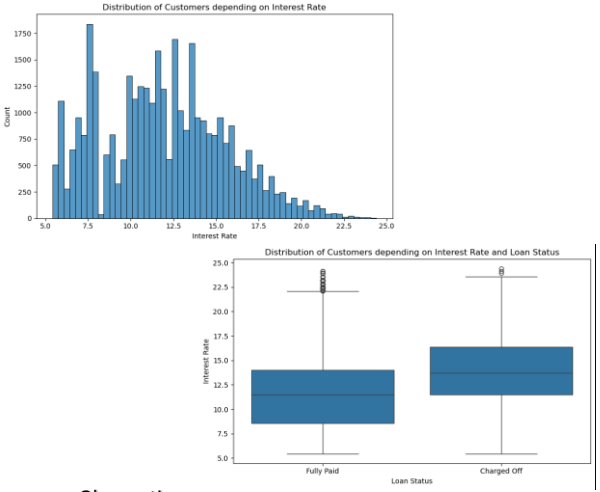
Univariate analysis on the Annual Income



Observations

1. The annual income of most applicants lies between 40K - 80K.
2. The annual income of most defaulters lies between 37K - 72K.

Univariate analysis on the Interest Rate

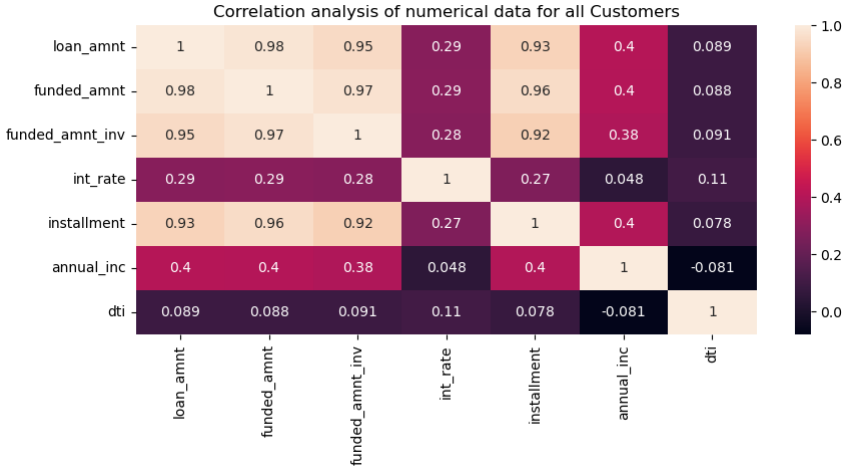


Observations

The rate of interest for most applicants lies between 9% - 14.5%.
The rate of interest for most defaulters lies between 11.5% - 16.5%.

Correlation Analysis

- **Correlation Analysis**: A correlation matrix was created to examine relationships between numerical variables, revealing key associations that may impact loan defaults.
- **Loan Status Analysis**: The relationship between loan_status and other key variables, such as int_rate and loan_amnt, was explored to identify significant factors influencing defaults

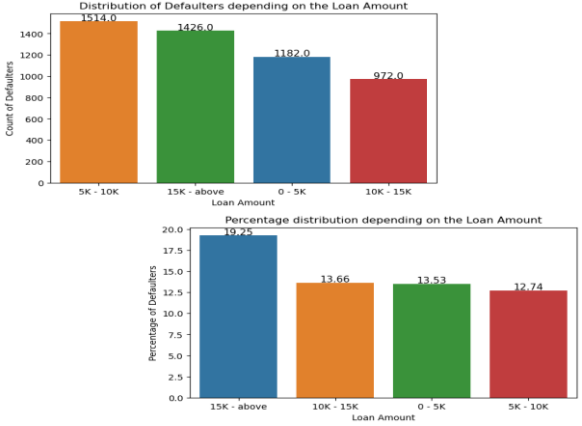


Observations:

- There is an excellent correlation between the loan amount, the funded amount and the funded amount by investors.
- There is a moderate correlation between the loan amount and annual income. There is a weak correlation between the loan amount and rate of interest.
- There is hardly any correlation of the DTI with the other attributes.

Categorical Data Analysis

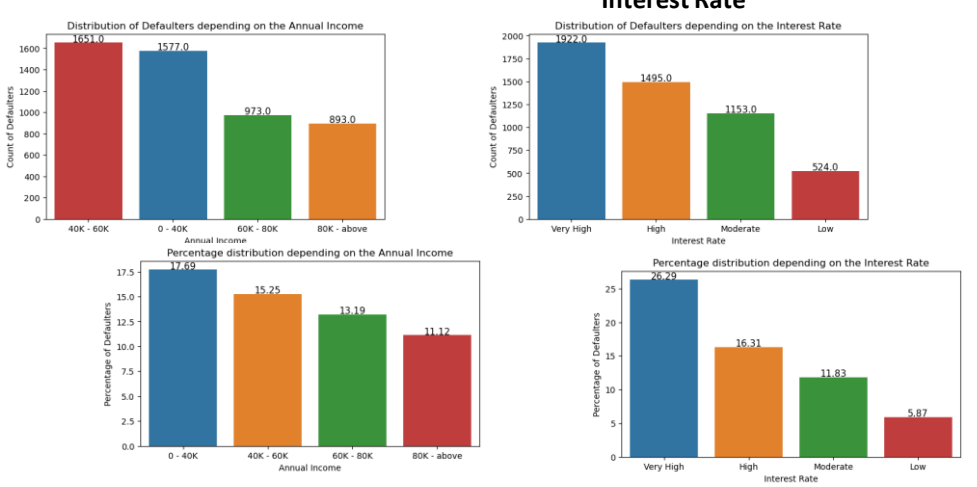
Loan Amount



Observations:

1. For loan amount, the most number of defaulters fall under the range of 5K-15K and the least under 10K-15K.
2. Looking at the proportion, the highest is for the range 15K-Above and the lowest is for 5K-10K.

Annual Income



Observations:

1. For annual income, the most number of defaulters fall under the range of 40K-60K and the least under 80K-Above.
2. Looking at the proportion, the highest is for the range 0-40K and the lowest is for 80K-Above.

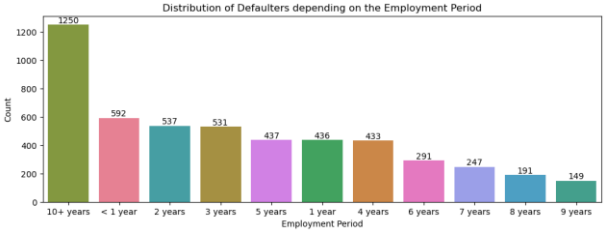
Interest Rate

Observations:

1. For rate of interest, the most number of defaulters fall under the category 'Very High' and the least under 'Low'.
2. Looking at the proportion, the highest is for the category 'Very High' and the lowest is for 'Low'.

Categorical Data Analysis(Ctnd)

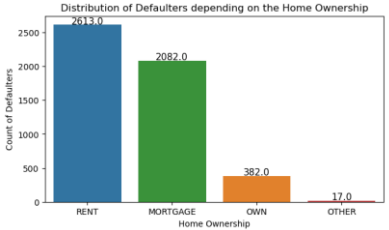
Employment Period

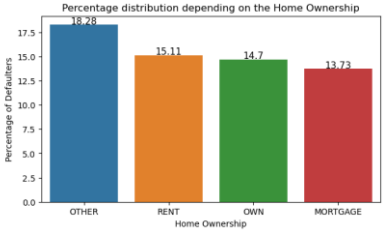


Observations:

1. For the employment period, the most number of defaulters fall under the category '10+ Years' and the least under '9 Years'.
2. There is not a high variation in the proportion for the different categories under employment period.

Home Ownership

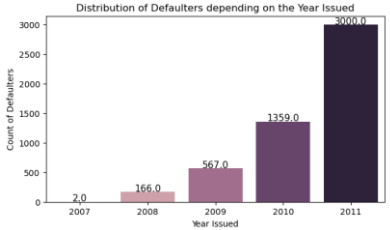


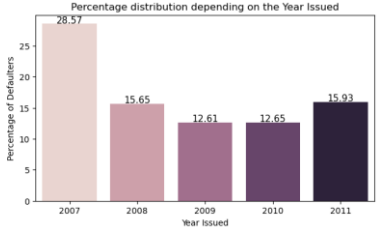


Observations:

1. For home ownership, the most number of defaulters fall under the category 'Rent' and the least under 'Other'.
2. Looking at the proportion, the highest is for the category 'Other' and the lowest is for 'Mortgage'.
3. There is a high proportion for the category 'Other', but the number of defaulters in it is very low compared to the other categories.

Year Issued



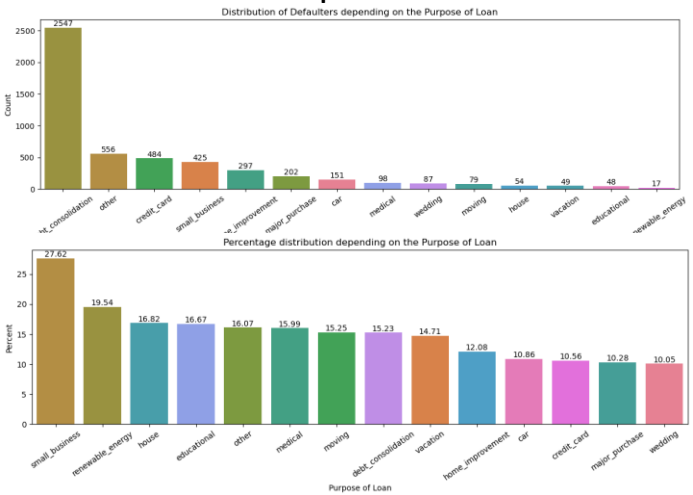


Observations:

1. For year of issue, the most number of defaulters fall under the year 2011 and the least under 2007.
2. Looking at the proportion, the highest is for the year 2007 and the lowest is for 2009.
3. There is a high proportion for the year 2007, but the number of defaulters in it is very low.
4. There was gradual decrease in the proportion over the years from 2007-2010, but then there wan an increase in the year 2011.

Categorical Data Analysis(Ctnd)

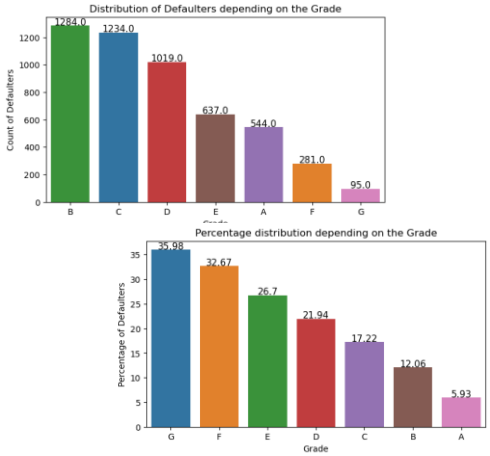
Purpose of Loan



Observations:

1. Looking at the proportion, the highest is for the category 'Small Business' and the lowest is for 'Wedding'.
2. 'Renewable Energy' has the second highest percentage, but it has the lowest number of defaulters.
3. 'Debt Consolidation', which has the highest number of defaulters, has an average proportion when compared to the others.

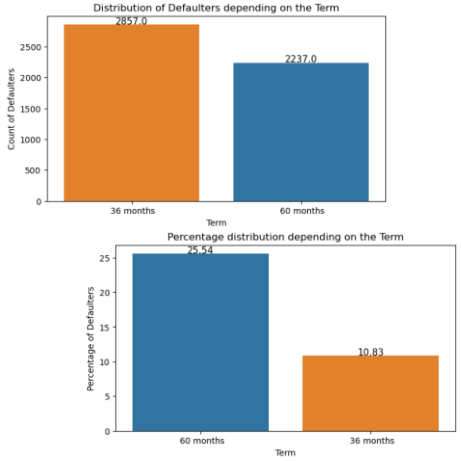
Grade



Observations:

1. For the grade, the most number of defaulters fall under the category 'B' and the least under 'G'.
2. Looking at the proportion, the highest is for the category 'G' and the lowest is for 'A'.
3. If we ignore the category A, then there is reverse correlation between the proportion and the actual number of defaulters

Term

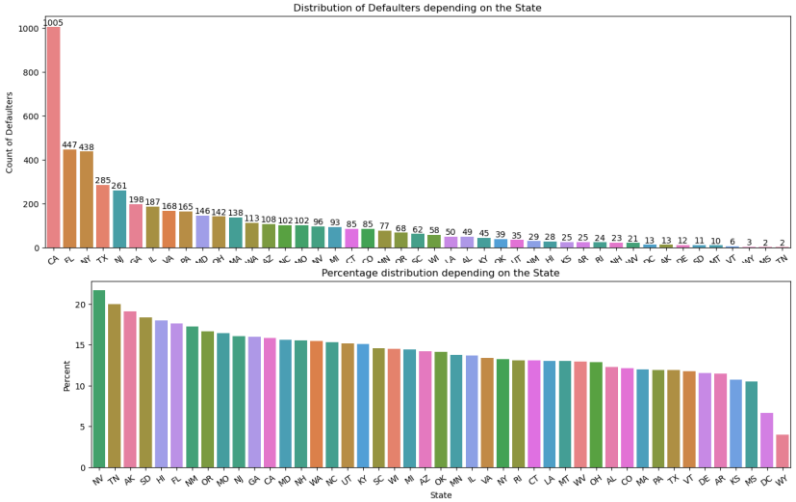


Observations:

1. For the term, the most number of defaulters fall under the category of '36 Months' and the least under '60 Months'.
2. Looking at the proportion, the highest is for the category '60 Months' and the lowest is for '36 Months'.
3. There is reverse correlation between the proportion and the actual number of defaulters
4. Even though there is not a huge difference in the number of defaulters between the two categories, the proportion difference is high.

Categorical Data Analysis(Ctnd)

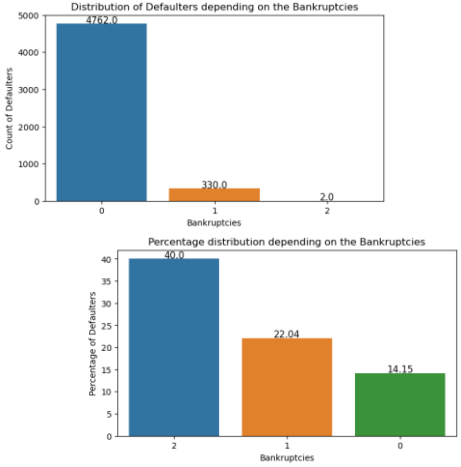
Address State



Observations:

1. Looking at the proportion, the highest is for the state of 'NV' and the lowest is for 'WY'.
2. Even though proportion is high for the state 'TN', the number of defaulters is very low.
3. The number of defaulters is high for the state of 'CA', but the percentage is almost average when compared to the others

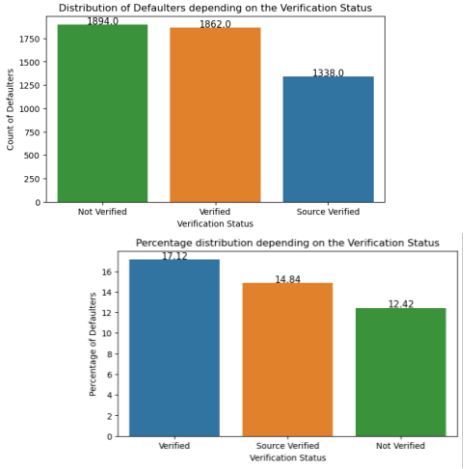
Public Record Bankruptcies



Observations:

1. For the bankruptcies, the most number of defaulters fall under the category of '0' and the least under '2'.
2. Looking at the proportion, the highest is for the category '2' and the lowest is for '0'.
3. There is reverse correlation between the proportion and the actual number of defaulters.
4. Even though proportion is high for the category '0', the number of defaulters is very low.

Verification Status



Observations:

1. For the verification status, the most number of defaulters fall under the category of 'Not Verified' and the least under 'Source Verified'.
2. Looking at the proportion, the highest is for the category 'Verified' and the lowest is for 'Not Verified.
3. Even though there is hardly any difference in the number of defaulters between the two categories 'Verified' and 'Not Verified', the proportion difference is high.

Key Insights

Interest Rates and Default Risk:
Higher interest rates are strongly correlated with a higher likelihood of loan default. Borrowers with rates in the "Very High" category had the highest default rates.
Income and Loan Default:
Borrowers with lower annual incomes (particularly those below $40K) are more likely to default on their loans. This emphasizes the need for stringent income verification and assessment.
Employment Tenure:
A shorter employment period, particularly under 5 years, is associated with higher default risks. Borrowers with over 10 years of employment have lower default rates.
Loan Amount and Default:
Loans in the mid-range ($5K-$15K) see the highest number of defaults. However, the proportion of defaulters is highest in the $15K and above category.
Home Ownership:
Renters have a higher default rate compared to homeowners, particularly those with a mortgage. This indicates that home ownership status is a significant factor in assessing loan risk.
Purpose of Loan:
Loans taken for small businesses have the highest default proportions, while debt consolidation loans have the highest number of defaulters. This insight can guide the development of targeted loan products.
Geographic Influence:
Certain states, such as Nevada (NV), show higher proportions of defaults, suggesting geographic location plays a role in loan repayment behavior.
Verification Status:
Loans that were not verified show a higher rate of default, highlighting the importance of thorough verification processes in loan approval.

Conclusion

The EDA provided valuable insights into factors influencing loan defaults.

These findings can guide strategies to minimize default risks.

Thank You!

- Thank you for joining this journey into loan default analysis. Your engagement is invaluable.
- We hope this analysis has provided you with useful insights into the complex world of lending.
- Should you have any questions, thoughts, or need further discussion, don’t hesitate to reach out.
- Together, let’s pave the way for a more insightful and responsible lending future.
- Thank you for your attention and interest in our findings!