Loan Default Analysis

Exploratory Data Analysis for Risk Assessment

By: Sahil Awasthi and Deepak Kor

Introduction

Brief description of the project:

- This project focuses on analyzing loan default data to understand patterns and factors that influence the likelihood of default.

Importance of understanding loan defaults for financial institutions:

- Loan defaults pose a significant risk to financial institutions, leading to financial losses.
- Understanding patterns can help mitigate these risks.

Objective:

- Identify patterns indicating the likelihood of loan default to inform risk assessment and decision-making.

Problem Statement

The business problem:

- Financial institutions aim to minimize financial loss by identifying risky loan applicants.

Two types of risks:

- Loss of business if a loan is not approved.
- Financial loss if a loan defaults.

Analysis Approach

Load and clean the dataset:

- Import the dataset and perform initial data cleaning.

Perform Exploratory Data Analysis (EDA) to understand data characteristics:

- Analyze the dataset to uncover initial insights and patterns.

Identify key factors influencing loan defaults:

- Determine which factors are most predictive of loan defaults.

Data Understanding

Brief description of the dataset:

- The dataset 'loan.csv' contains the complete loan data for all loans issued through the period 2007 to 2011. It includes information about past loan applicants and whether they defaulted or not.

Key variables:

- Loan amount, interest rate, term, loan status, annual income, debt-to-income ratio (DTI), employment length, etc.

Handling missing values and data types:

- Addressing missing values through imputation or removal and ensuring correct data types for analysis.

Data Cleaning

Process of handling missing values:

- Techniques used include dropping columns with more than 50% missing values and filling remaining missing values with the mean for numerical columns and the mode for categorical columns.

Steps to convert categorical variables to numerical:

- Encoding categorical variables such as 'term' by extracting numerical values.

Replacement of infinite values with NaN and subsequent handling:

- Addressing infinite values and cleaning the dataset by replacing them with NaN and dropping rows with any remaining NaN values.

Univariate Analysis

Analysis of individual features:

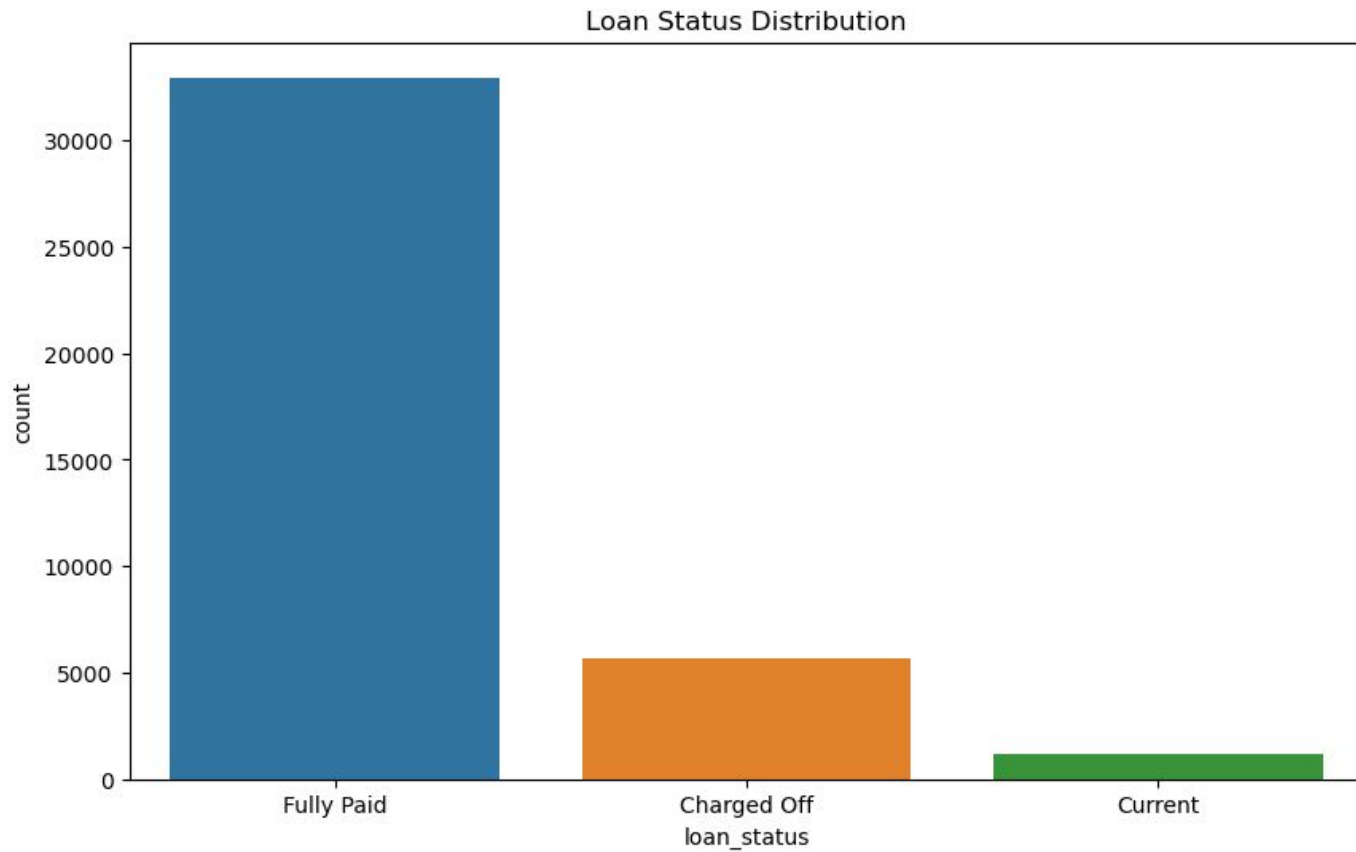- Examining the distribution of key variables to understand their individual characteristics.

Example: Distribution of loan statuses:

- Analyzing how loan statuses are distributed across the dataset.

Visualization:

- **Bar chart:** Showing the distribution of 'loan_status'.

Bar chart: Showing the distribution of 'loan_status'.

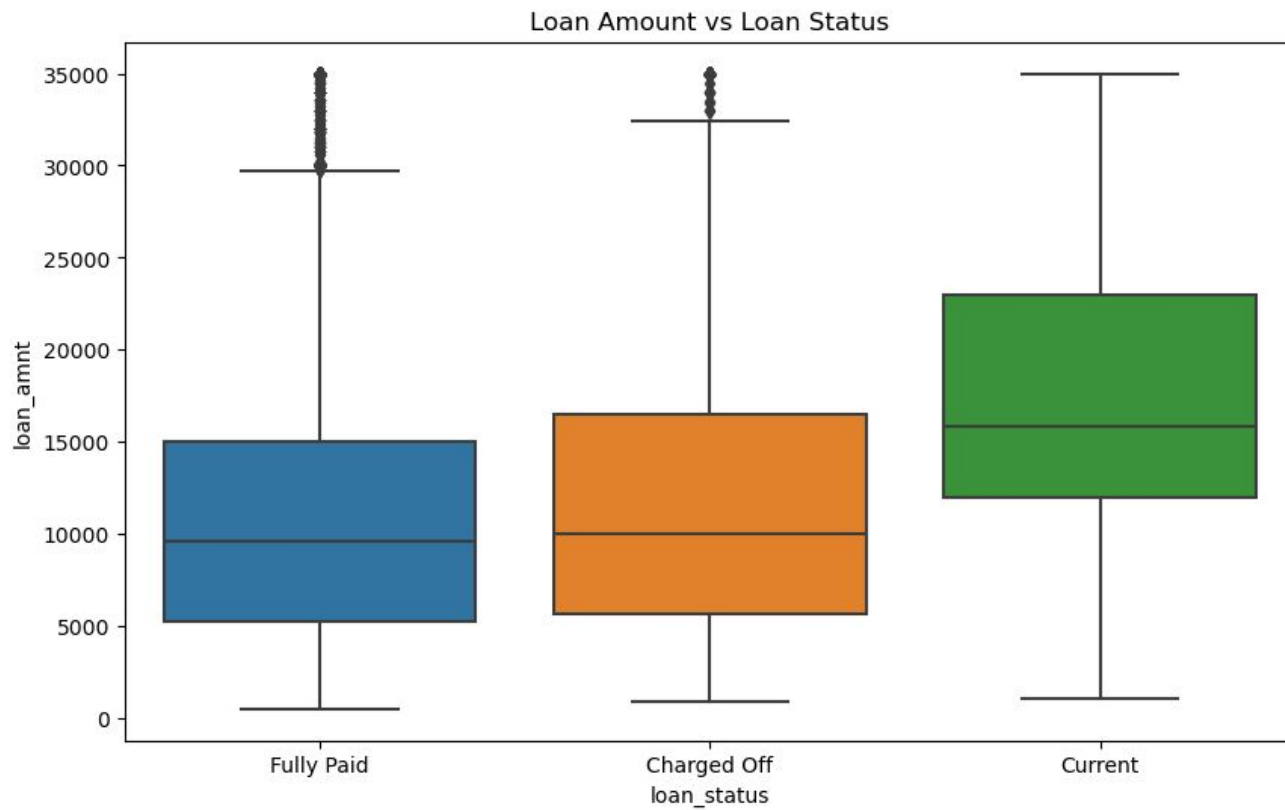# Bivariate Analysis - Loan Amount vs Loan Status

Analysis of the relationship between loan amount and loan status:

- Investigating how loan amounts vary with different loan statuses.

Visualization:

- **Boxplot:** Showing 'loan_amnt' across different 'loan_status'.

Boxplot: Showing 'loan_amnt' across different 'loan_status'.

Loan Amount vs Loan Status

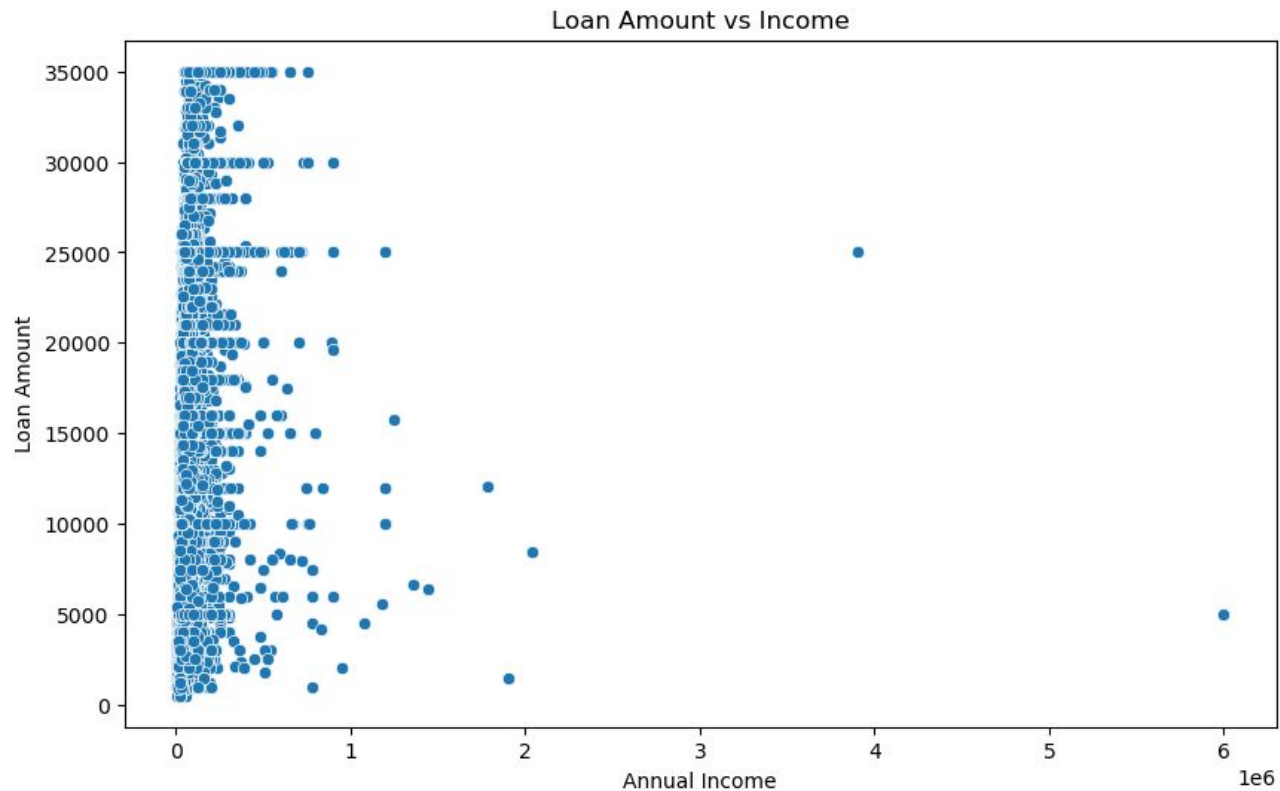# Bivariate Analysis - Loan Amount vs Income

Analysis of the relationship between loan amount and income:

- Exploring how loan amounts correlate with applicant incomes.

Visualization:

- **Scatter plot:** Showing 'loan_amnt' vs 'annual_inc'.

Scatter plot: Showing 'loan_amnt' vs 'annual_inc'.

Loan Amount vs Income
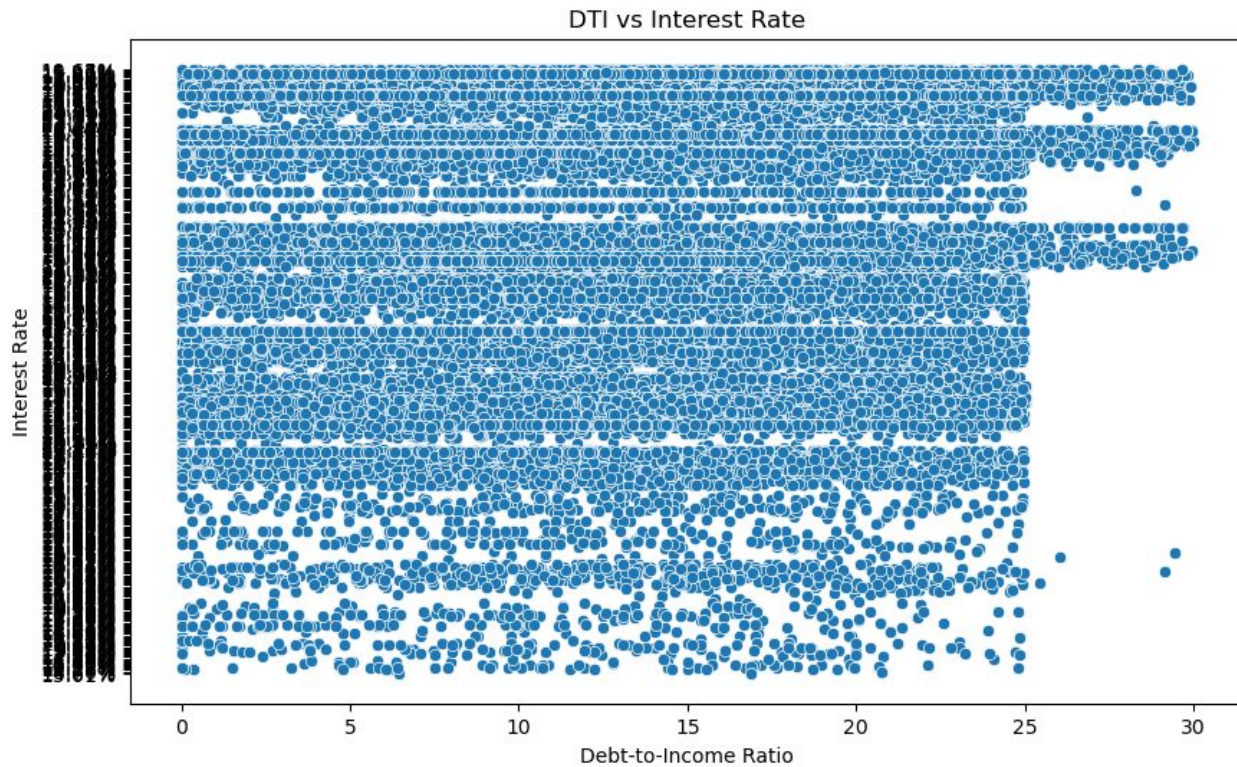
Bivariate Analysis - DTI vs Interest Rate

Analysis of the relationship between debt-to-income ratio (DTI) and interest rate:

- Understanding how DTI influences the interest rates of loans.

Visualization:

- **Scatter plot:** Showing 'dti' vs 'int_rate'.

Scatter plot: Showing 'dti' vs 'int_rate'.

DTI vs Interest Rate
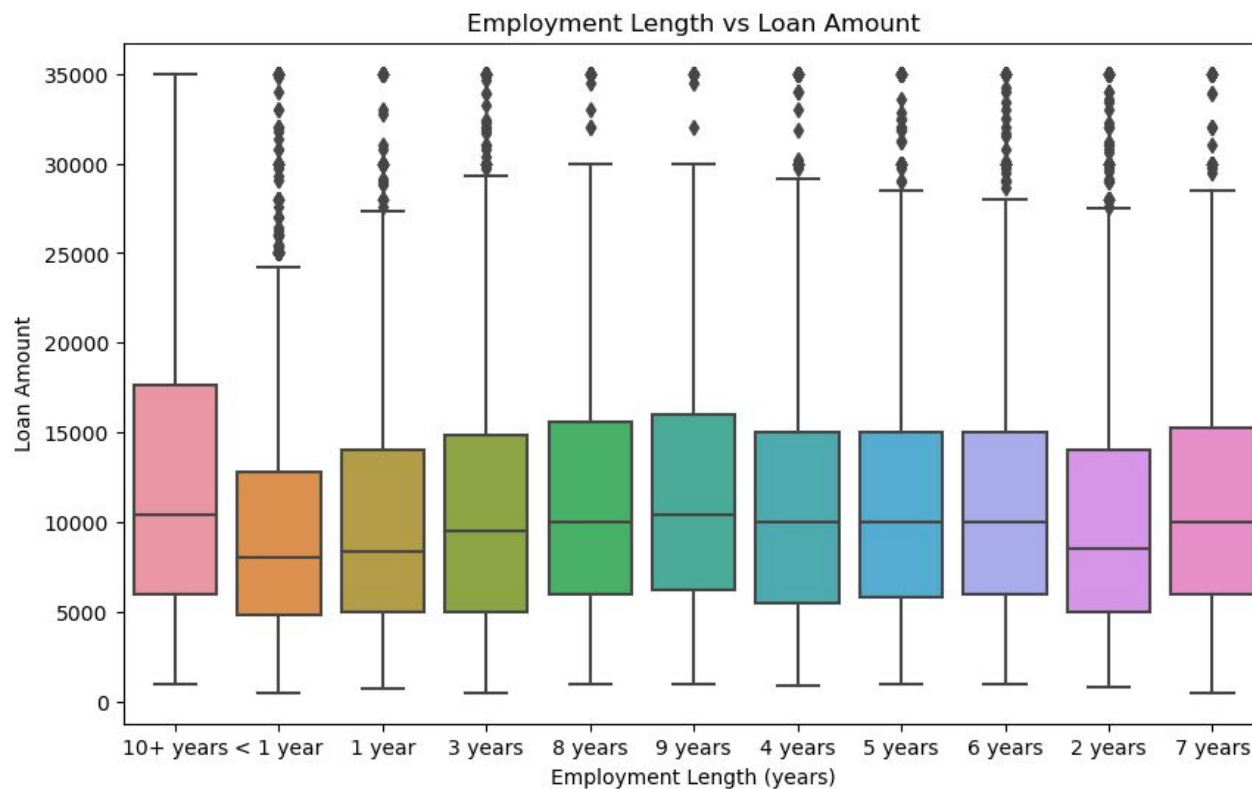
Employment Length vs Loan Amount

Analysis of the relationship between employment length and loan amount:

- Investigating the impact of employment length on loan amounts.

Visualization:

- **Boxplot:** Showing 'emp_length' vs 'loan_amnt'.

Boxplot: Showing 'emp_length' vs 'loan_amnt'.

Employment Length vs Loan Amount

Multivariate Analysis

Analysis of relationships among multiple variables:

- Examining the interplay between several key variables.
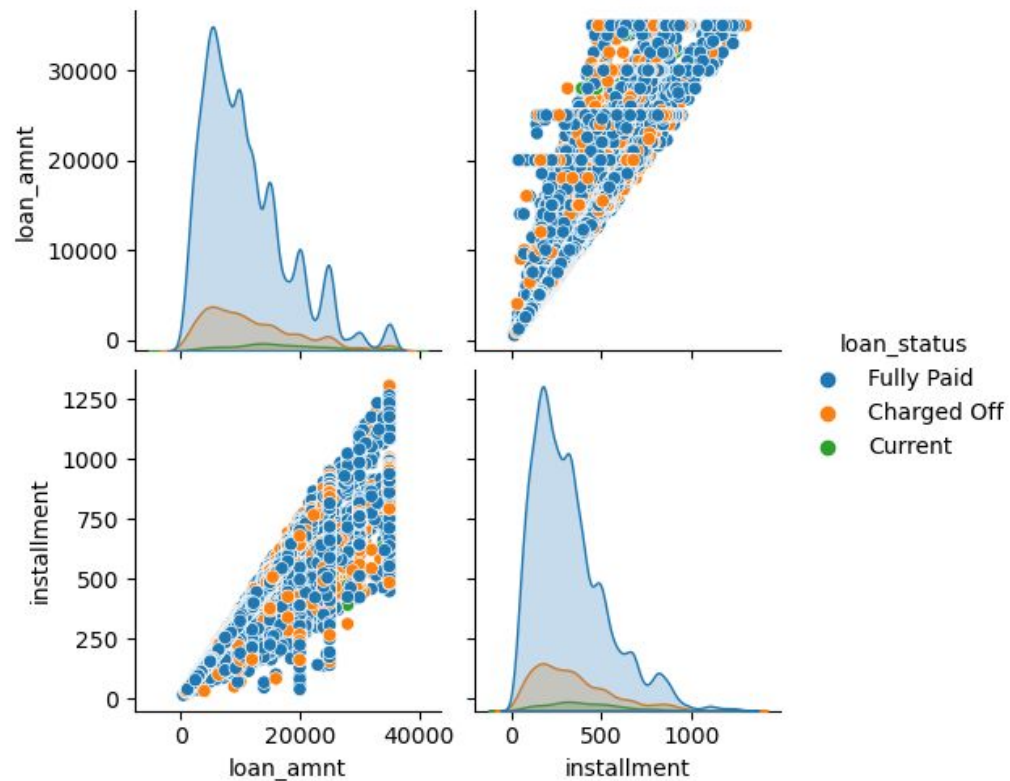
Example: Pairplot of loan amount, interest rate, installment, and loan status:

- Visualizing how these variables interact.

Visualization:

- **Pairplot:** Visualizing relationships among key variables.

Boxplot: Showing 'emp_length' vs 'loan_amnt'.

Key Insights

Summary of Findings from the Analysis:

- Through the exploratory data analysis, we uncovered several key insights regarding the patterns and factors influencing loan defaults.

Default Rate Among Loan Applicants:

- The default rate within the dataset was significant, highlighting the importance of robust risk assessment mechanisms. Approximately 14.17% of loan applicants defaulted on their loans.

Importance of Loan Amount and Interest Rates:

- Loan amount and interest rates were found to be crucial factors influencing the likelihood of default. Higher loan amounts and higher interest rates were correlated with increased default rates.

Correlations with Consumer Attributes:

- Various consumer attributes, such as annual income, employment length, and debt-to-income ratio (DTI), showed significant correlations with loan outcomes. Higher incomes and longer employment lengths were associated with lower default rates, while higher DTI ratios were linked to higher default risks.

Business Implications

Recommendations for Reducing Financial Loss:

- Implement stricter loan approval criteria for high-risk applicants identified through the analysis.
- Offer lower loan amounts or higher interest rates to applicants with higher default risks to mitigate potential losses.

Strategies for Handling High-Risk Applicants:

- Develop targeted risk assessment models to better evaluate the creditworthiness of applicants.
- Provide financial counseling and support to high-risk applicants to improve their financial stability and reduce default rates.

Potential Use of Risk Assessment Models:

- Leverage predictive models and machine learning algorithms to assess the risk of default for each loan applicant.
- Continuously update and refine these models based on new data to ensure their accuracy and effectiveness in predicting defaults.

Conclusion

Recap of the Analysis and Key Findings:

- The analysis provided valuable insights into the factors influencing loan defaults, with significant findings related to loan amounts, interest rates, and consumer attributes.

Importance of EDA in Understanding Loan Defaults:

- Exploratory Data Analysis (EDA) played a crucial role in identifying patterns and correlations within the dataset, helping to inform better risk assessment and decision-making processes.

Next Steps for Implementing Insights into Business Processes:

- Implement the recommendations for reducing financial loss and handling high-risk applicants.
- Develop and deploy risk assessment models to improve loan approval decisions.
- Continuously monitor and update the models and strategies based on new data and insights to maintain their effectiveness.

Thank You

Thank you all for your attention. I appreciate the time you have taken to view our presentation and the interest you have shown in this analysis.

Contact information:

If you have any further questions or would like to discuss this analysis in more detail, please feel free to reach out.

- **Email:** connectwithsahilawasthi@gmail.com
- **Phone:** +91 6265851105, +91 98672 65364
- **LinkedIn:** Sahil Awasthi, Deepak Kor
- **GitHub:** Sahil Awasthi, Deepak Kor