

Lending Club Case Study

Submitted By

Pradeep Harry Michael

Aman Gupta

Contents



Problem Statement



Analysis and Outcomes

- 1> Data Understanding and Cleaning
- 2> Univariate/Bivariate for numerical
- 3> Bivariate for Categorical Data



Business Summary



Problem Statement

- The consumer finance company faces a critical challenge in making loan approval decisions that minimize financial risks. When an applicant applies for a loan, the company needs to assess whether the applicant is likely to repay the loan. Approving loans to applicants who are likely to default results in significant financial losses, while rejecting loans from creditworthy applicants means missed business opportunities.
- The primary concern for the company is identifying "risky" applicants—those who are most likely to default. These defaulters cause the largest amount of credit loss. To address this, the company aims to analyze past loan data to identify patterns that predict whether an applicant is likely to default. By uncovering these patterns, the company can take proactive measures, such as denying loans to high-risk applicants, adjusting loan amounts, or charging higher interest rates. This data-driven approach will help the company reduce credit loss and better manage its loan portfolio, ensuring financial stability.



Analysis : Data Understanding And Cleaning

Data Understanding and Cleaning activities performed

- Drop Columns which do not have any values across all Rows.
- Drop Columns which have more than 50% of the Rows with empty values.
- Change values of columns like loan status to business standard like "Default" and "Non-Default"
- Drop Rows which have all column values as NA or Null
- Data type Changes
- Drop Rows which are duplicate

Outcome

- The initial Data Frame of shape(rows , columns) (39717, 111) was reduced to (39717, 54)
- The Key Columns for analysis were identified

Numerical Fields

- int_rate
- annual_inc
- loan_amnt
- delinq_2yrs
- inq_last_6mths

Categorical Field

- term,grade,sub_garde
- emp_length , home_ownership,issue_d
- verification_status , purpose , addr_state
- pub_rec, pub_rec_bankruptcies

Target Field

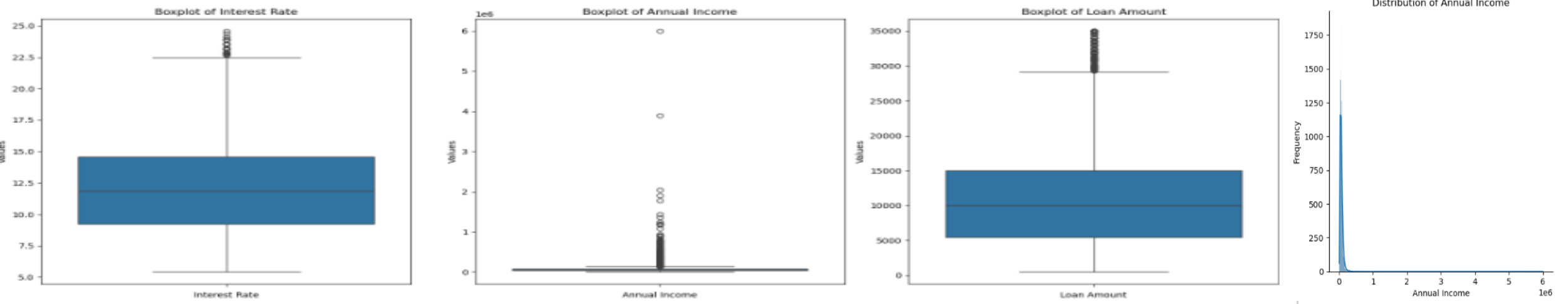
- loan_status



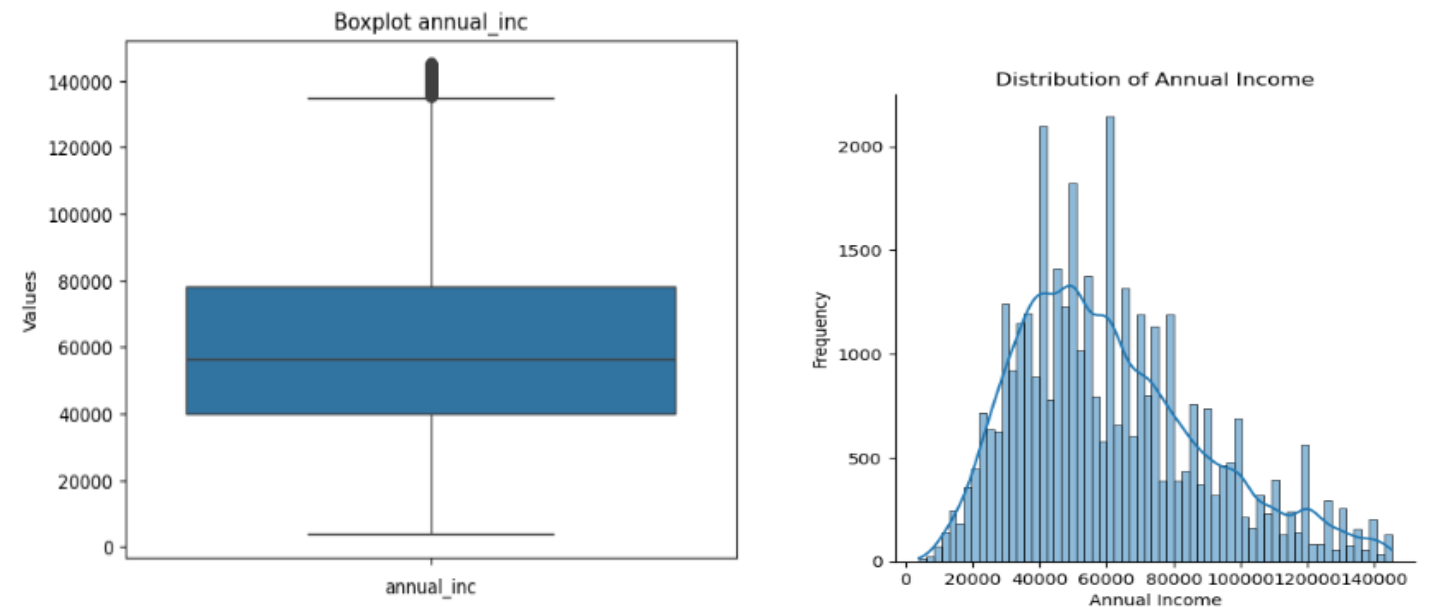
Analysis :Univariate and Bivariate for Numeric

Identified Outliers and removed

- Box plots indicated Outliers for 'annual_inc' field . The IQR logic was used to remove Outliers since it did not have a Normal Distribution



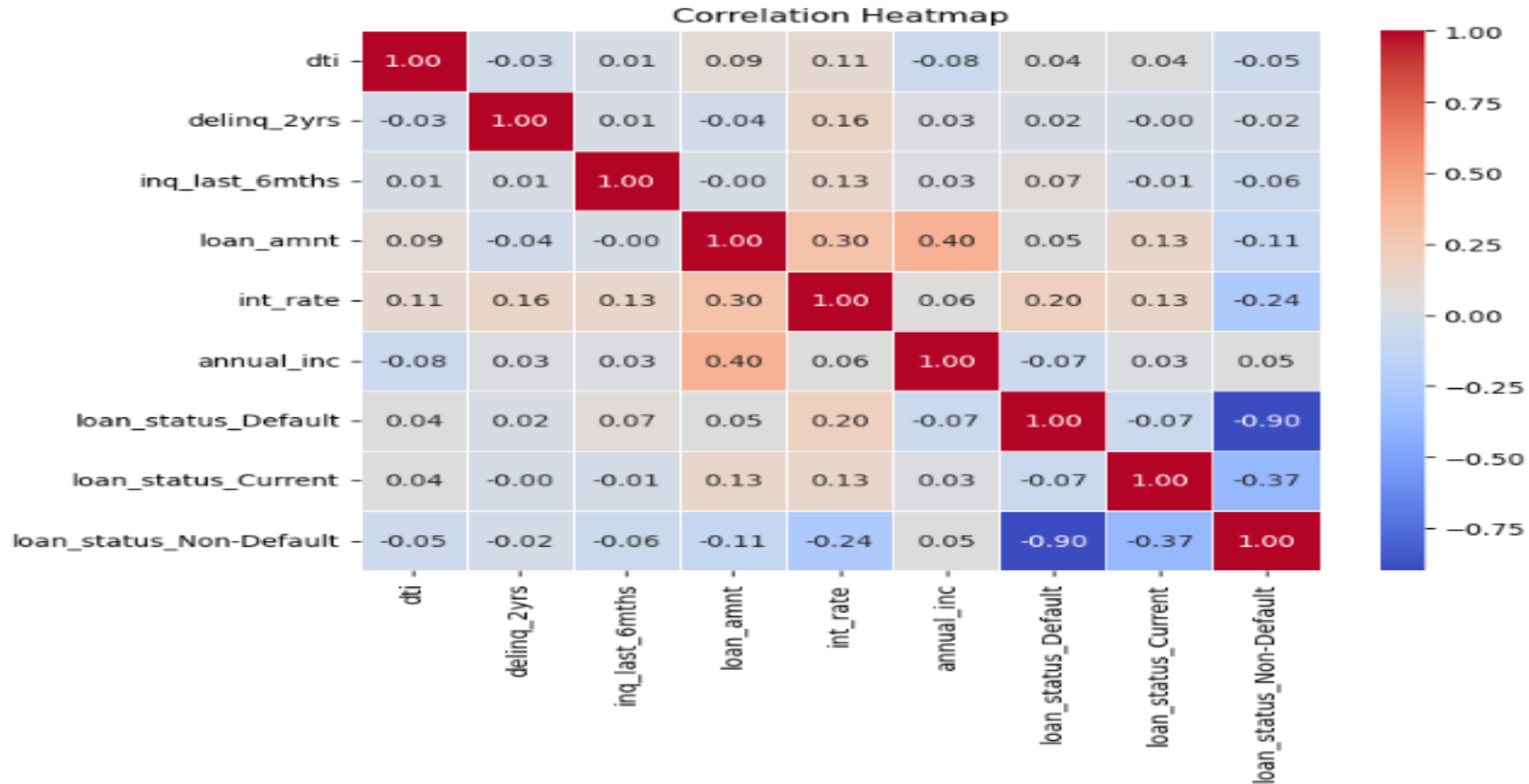
- Outcome** After removing Outlier for 'annual_inc'





Analysis :Univariate and Bivariate for Numeric

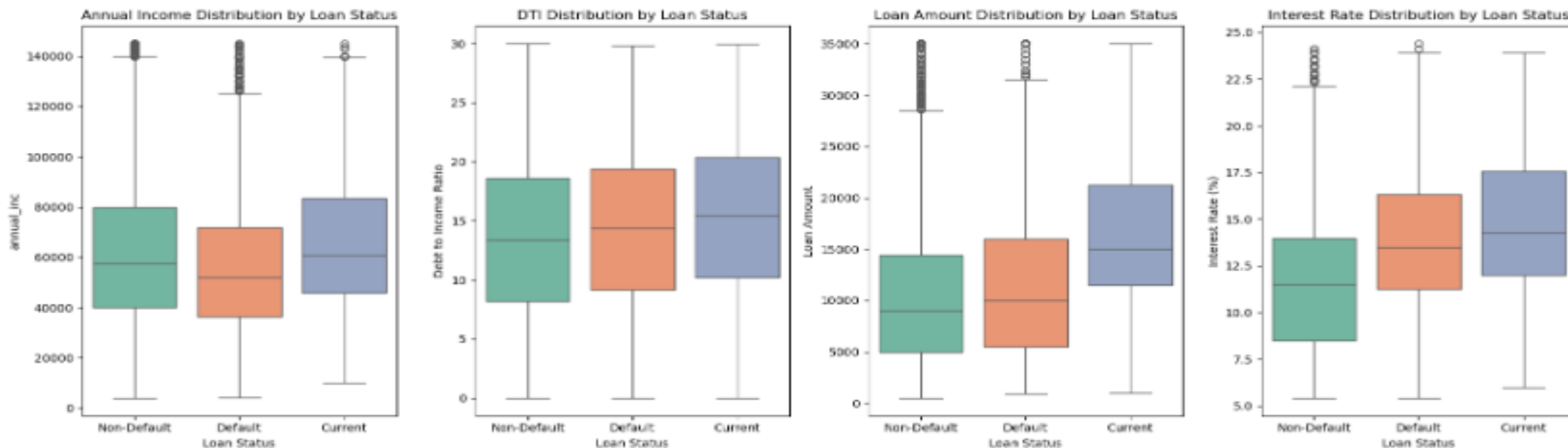
Heat Map between the Numeric Fields along with Loan Status





Analysis :Univariate and Bivariate for Numeric

Distribution Analysis was performed with fields 'annual_inc', 'dti', 'loan_amt', 'int_rate' against 'loan_status'



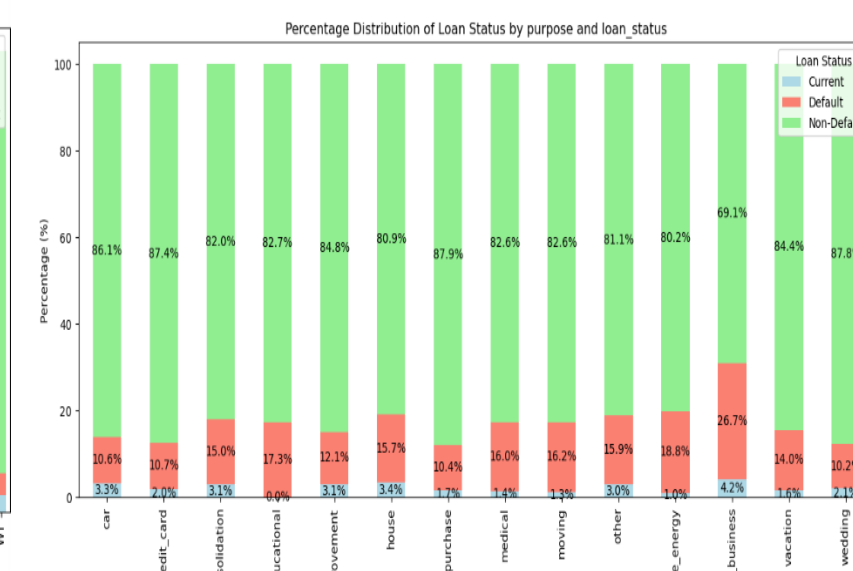
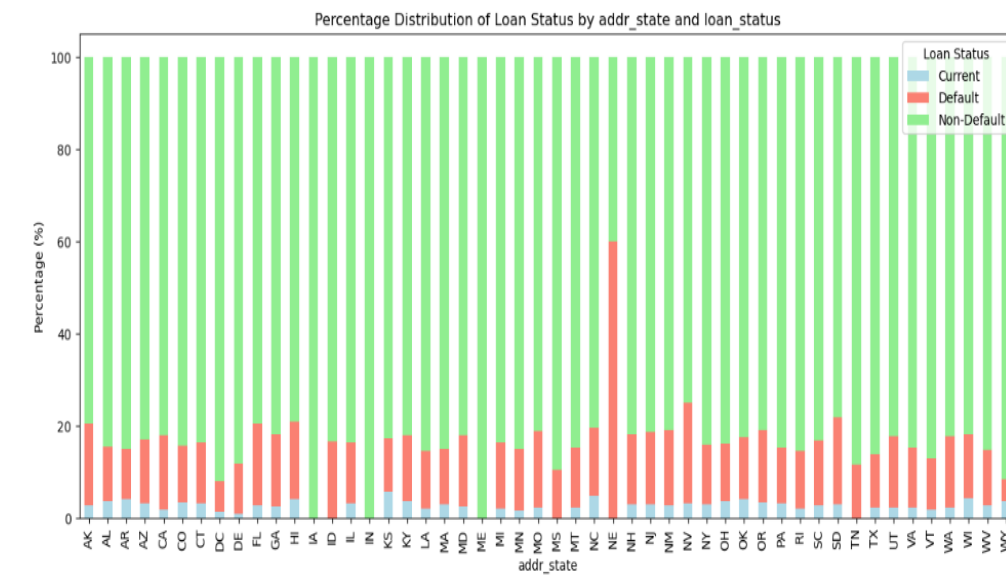
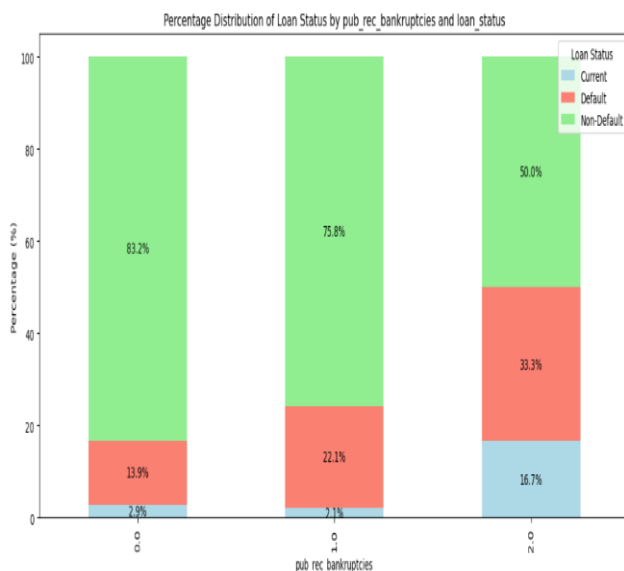
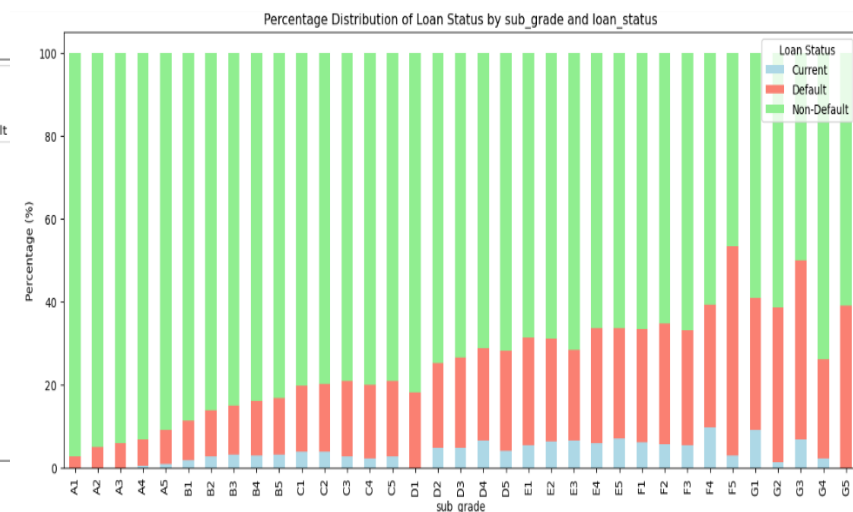
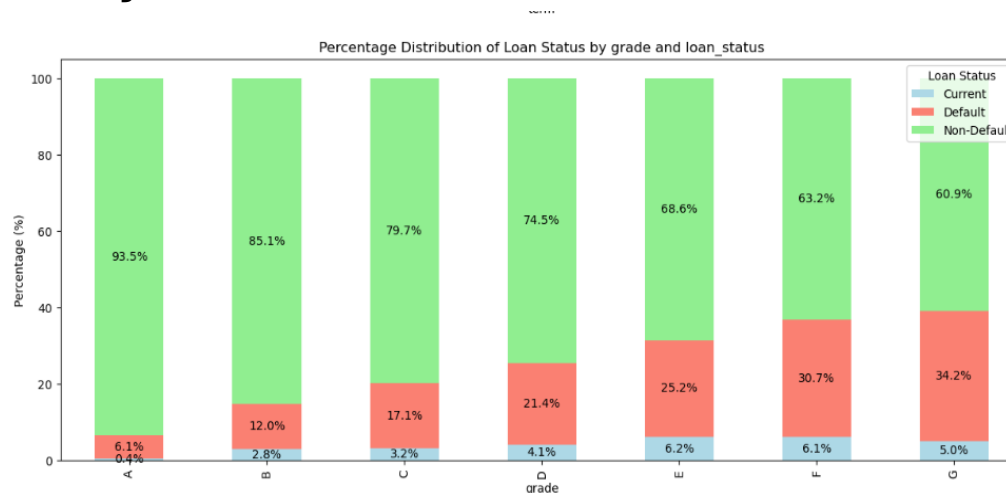
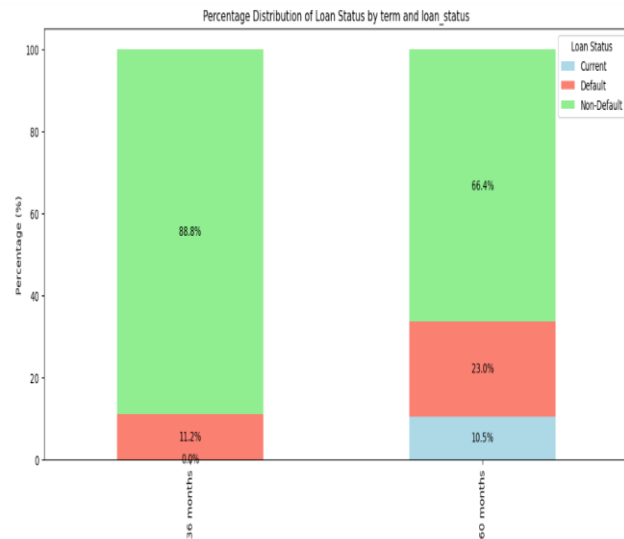
Outcome

- **int_rate** (Interest Rate) indicates that Loans with higher int_rate have more risk of getting Defaulted
- **dti** (Debt to Interest Rate Ratio) indicates that higher dti ratio are at higher risk of getting Defaulted
- **annual_inc** (Annual Income) even though the correlation is not very prominent but still lower annual_inc have higher risk of getting Defaulted
- **loan_amt** (Loan Amount) even though the correlation is not very prominent but still higher loan_amt have higher risk of getting Defaulted



Analysis :Bivariate Analysis of Categorical Fields

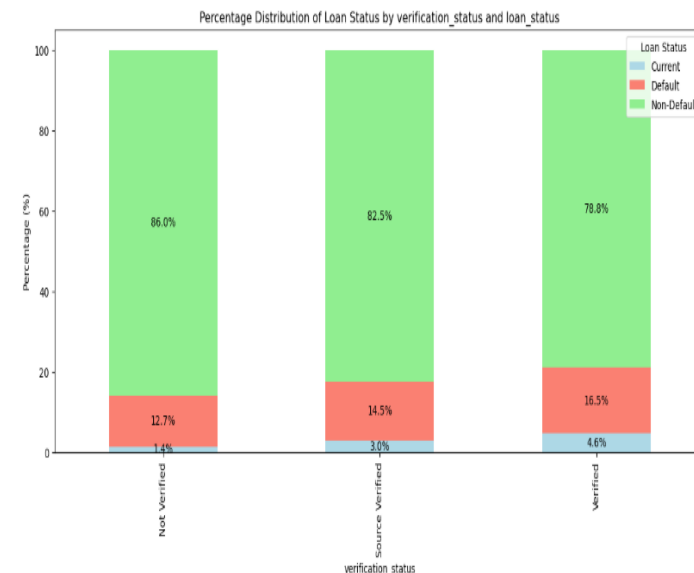
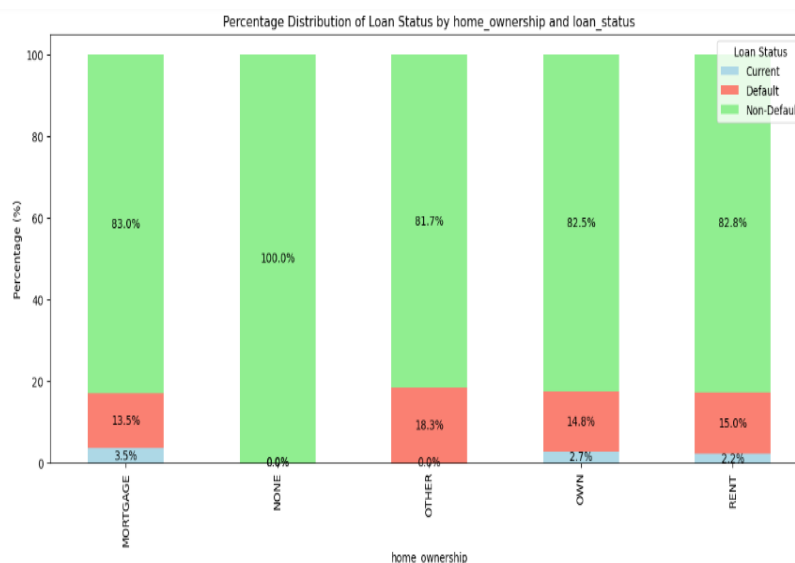
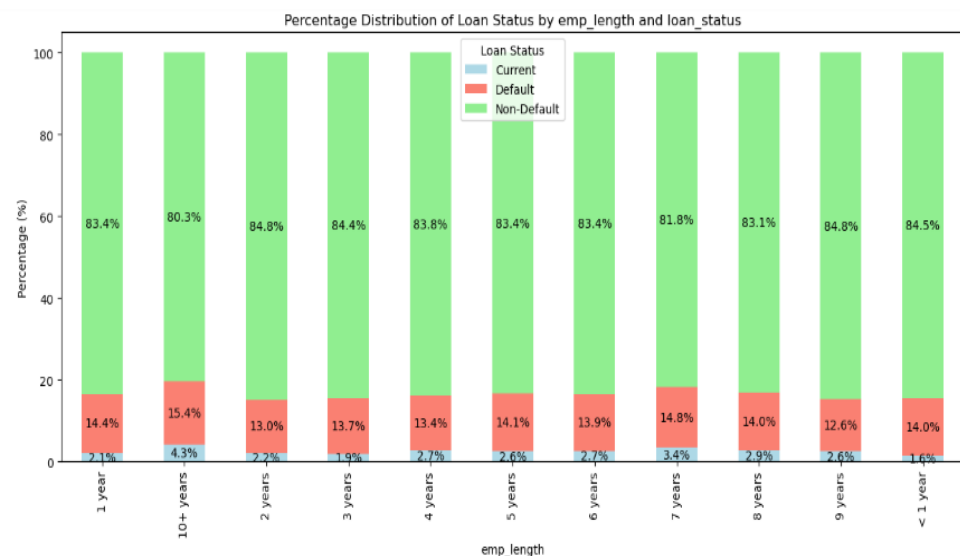
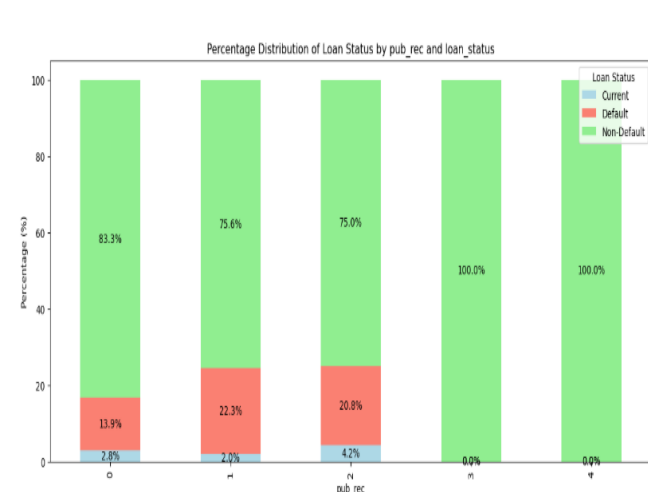
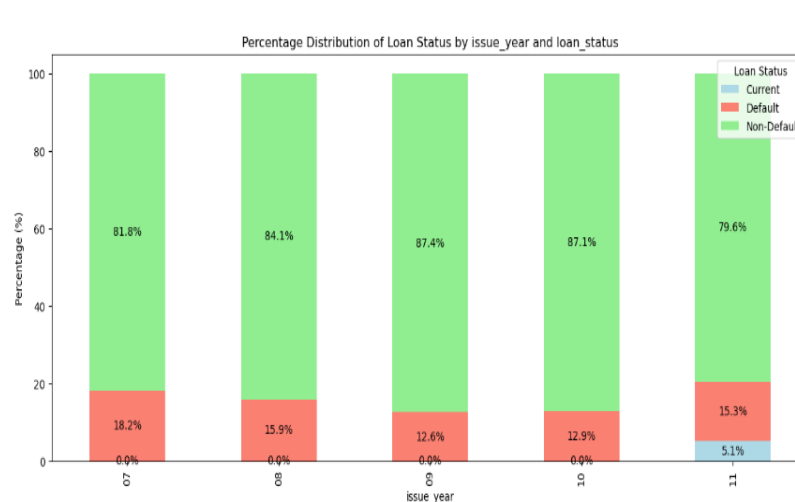
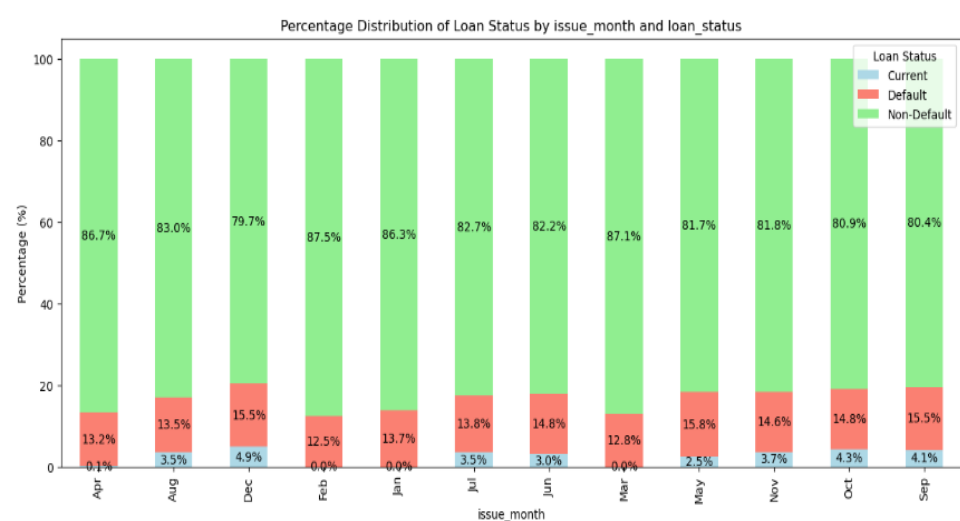
Bivariate analysis indicating following fields have a considerable impact on the status of loan and contribute to whether it is likely to Default





Analysis :Bivariate Analysis of Categorical Fields

Bivariate analysis indicating following fields do not have a considerable impact on loan status





Analysis :Bivariate Analysis of Categorical Fields

OutCome of Bivariate Analysis for Categorical Fileds

- The following Categorical Variables have a considerable impact on Loan getting Defaulted
 - **term** , Higher Term like 60 months are at a higher risk of getting defaulted
 - **grade and sub_grade** , lower the grade then higher is the risk of getting defaulted
 - **purpose** , the purpose named “**small business**” has a higher probability of getting defaulted in comparison to other purposes
 - **addr_state** , the state 'NE' has a higher probability of getting defaulted though the count of samples considered are very less
 - **pub_rec_bankruptcies**, higher the number of publicly recorded bankruptcies relates to higher risk of getting Defaulted
- The following Categorical Variables do not have a significant or prominent impact on Loan getting Defaulted
 - **emp_length**
 - **home_ownership**
 - **verification_status**
 - **pub_rec**
 - **issue_d**



Business Summary

- The consumer finance company is facing a critical challenge in balancing the risk of financial losses with the opportunity for growth through loan approvals. The company needs to identify high-risk applicants—those likely to default—while ensuring it does not miss potential business opportunities by rejecting creditworthy customers. The objective is to minimize credit loss while maintaining a healthy portfolio. The company intends to achieve this by leveraging past loan data to identify patterns and factors that predict loan default.
- Through analysis, several key factors have been identified that influence the likelihood of default. **Numerical analysis**, loans with **higher interest rates (int_rate)** or **higher Debt to Income Ratio** show a higher risk of default, as well as loans taken by employees with lower annual income (**annual_inc**) or higher loan amounts (**loan_amt**). While the correlations for **annual income** and **loan amounts with loan status** are not overwhelmingly strong, these factors still play a role in assessing risk.
- From a **categorical analysis** perspective, certain variables significantly impact the likelihood of default. **Loan term** plays a crucial role, with longer terms (e.g., 60 months) being associated with higher default risk. Additionally, **loan grade and sub-grade** indicate that lower grades correlate with a higher risk of default. **Purpose** of the loan also affects risk, particularly loans for "small business" purposes, which show a higher probability of default. The state **'NE'** also demonstrates a higher default rate, though this is based on a smaller sample size. **Publicly recorded bankruptcies** (pub_rec_bankruptcies) have a strong correlation with default risk—the higher the number of bankruptcies, the greater the chance of default.
- On the other hand, some factors do not show a significant impact on the likelihood of default, such as **employment length** (emp_length), **homeownership status** (home_ownership), **verification status**, **public record** (pub_rec), and the **issue date** of the loan.
- By understanding these patterns and utilizing this data-driven approach, the company can proactively manage its loan portfolio. It can deny loans to high-risk applicants, adjust loan amounts, or increase interest rates for those who pose a higher risk, thus reducing credit loss and ensuring financial stability. This strategic approach not only protects the company's financial health but also allows it to make more informed, risk-aware lending decisions.