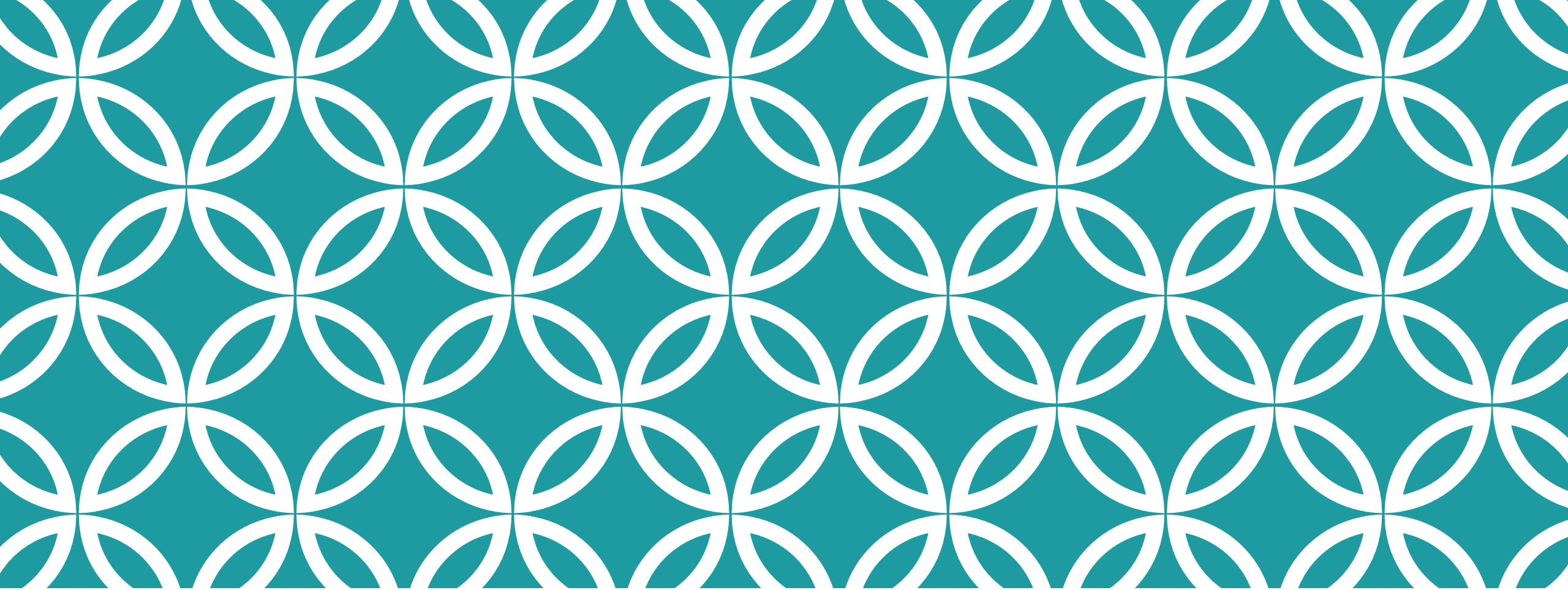


# LENDING CLUB CASE STUDY

AI ML C65 2024

Raghavendra Siddappa
Rajesh Sinha

# INTRODUCTION

## Overview:

This is a data science project focused on risk analytics in the banking and financial services sector, utilizing the Lending Club dataset to predict loan defaults.

## Problem Statement:

This **Company** is the largest **online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures**. Borrowers can easily access lower interest rate loans through a fast online interface. Like most other lending companies, lending loans to **'risky'** applicants is the largest source of **financial loss (called credit loss)**. **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 **credit loss**.

## Goal

Utilize **Exploratory Data Analysis(EDA)** to determine which types of **customers** are likely to **default on a loan**. By identifying the driving factors behind loan defaults to **reduce credit loss effectively**.

# DATA OVERVIEW

**Dataset Description:**

Our analysis is based on a dataset comprising historical loan application data from Lending Club.
The dataset contains various features such as loan amount, interest rate, loan grade, annual income, and loan status.

**Data Source:**

The data given contains information about past loan applicants and whether they 'defaulted' or not. Data has details regarding approved loan not the rejected ones. It has 3 status of loan which is Fully Paid, Current and Charged-Off.

**Key Features:**

- **Loan Amount:** The principal amount of the loan.
- **Interest Rate:** The interest rate charged on the loan.
- **Loan Grade:** The credit grade assigned to the loan.
- **Annual Income:** The borrower's annual income.
- **Loan Status:** Indicates whether the loan was fully paid or defaulted

# APPROACH

**Step-by-Step Analysis:**

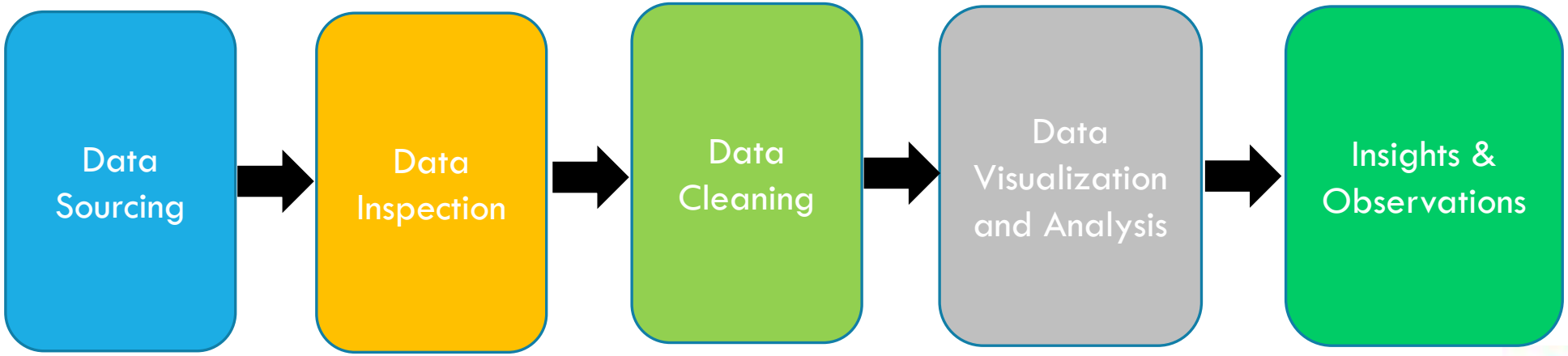
**Data Sourcing**: Load the provided csv file data.

**Data Inspection**: Understanding the given data.

**Data Cleaning**: Address missing values, outliers, and data inconsistencies.

**Data Visualization and Analysis**: Perform univariate, segmented univariate and bivariate analysis.

**Insights and observations** : Derive insights and suggest actionable recommendations.



# DATA SOURCING AND DATA INSPECTION

**Data Loading:**

- Open the provided loan.csv and Data_Dictionary excel file to get a feel and understanding of data and columns.
- Open Anaconda software and create a new Jupyter Notebook.
- Import the required libraries.
- Load the given loan.csv file in DataFrame using pandas.
- Use different methods like info, describe, dtypes, columns etc. to do data inspection.

# DATA CLEANING

**Initial Data Cleaning:**

- Handling missing values: Dropped columns with more than 50% missing values and imputed remaining missing values.
- Removed Duplicate Entries: Ensured data integrity by removing duplicate records.
- Corrected Data Types: Ensured consistency by converting data types (e.g., numeric strings to integers).

**Outliers:**

- Identified outliers using the Interquartile Range (IQR) method.
- Treated outliers by capping or removing extreme values to prevent skewing the analysis.

# DATA VISUALIZATION AND ANALYSIS

- **Distribution of Loan Amounts:**

  •Most loans are below 20,000 indicating a higher frequency of smaller loans.

- **Interest Rate Trends:**

  •Interest rates show a concentration around 10-15%, with some loans having higher rates.

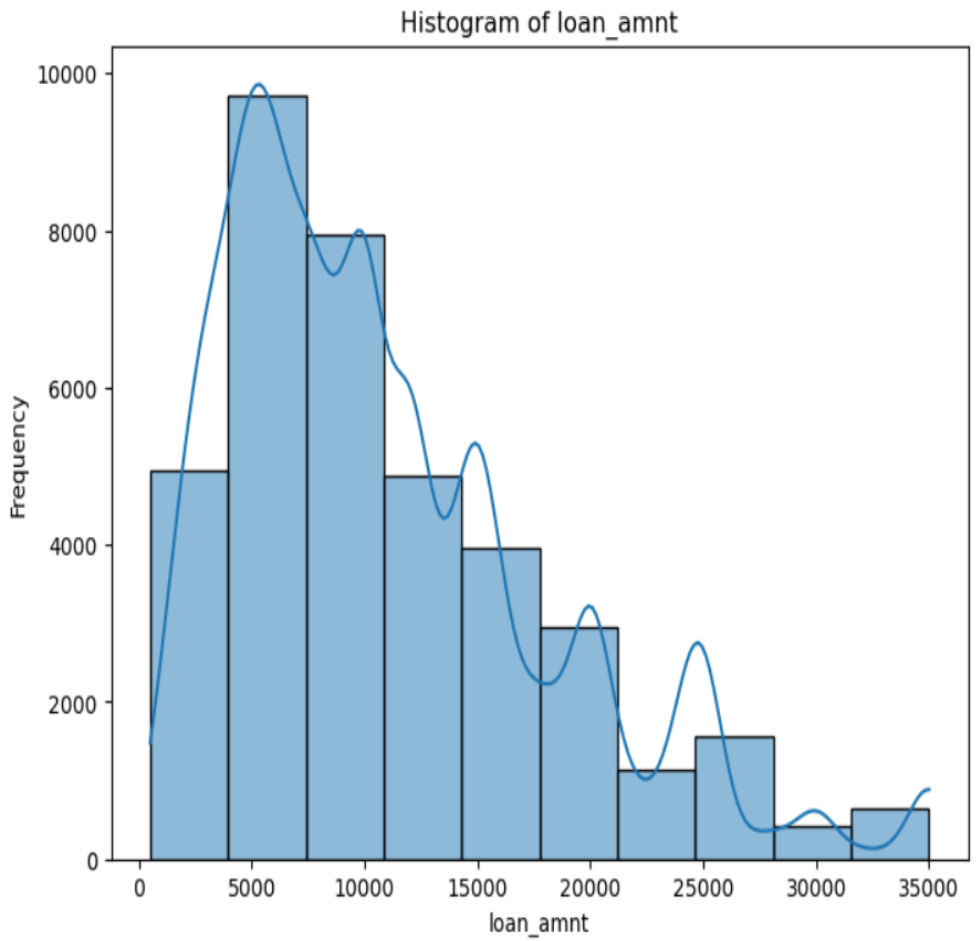
- **Loan Grades:**

  •Higher quality loans (grades A, B) are more common, while lower quality loans (grades D, F, G) are less frequent.

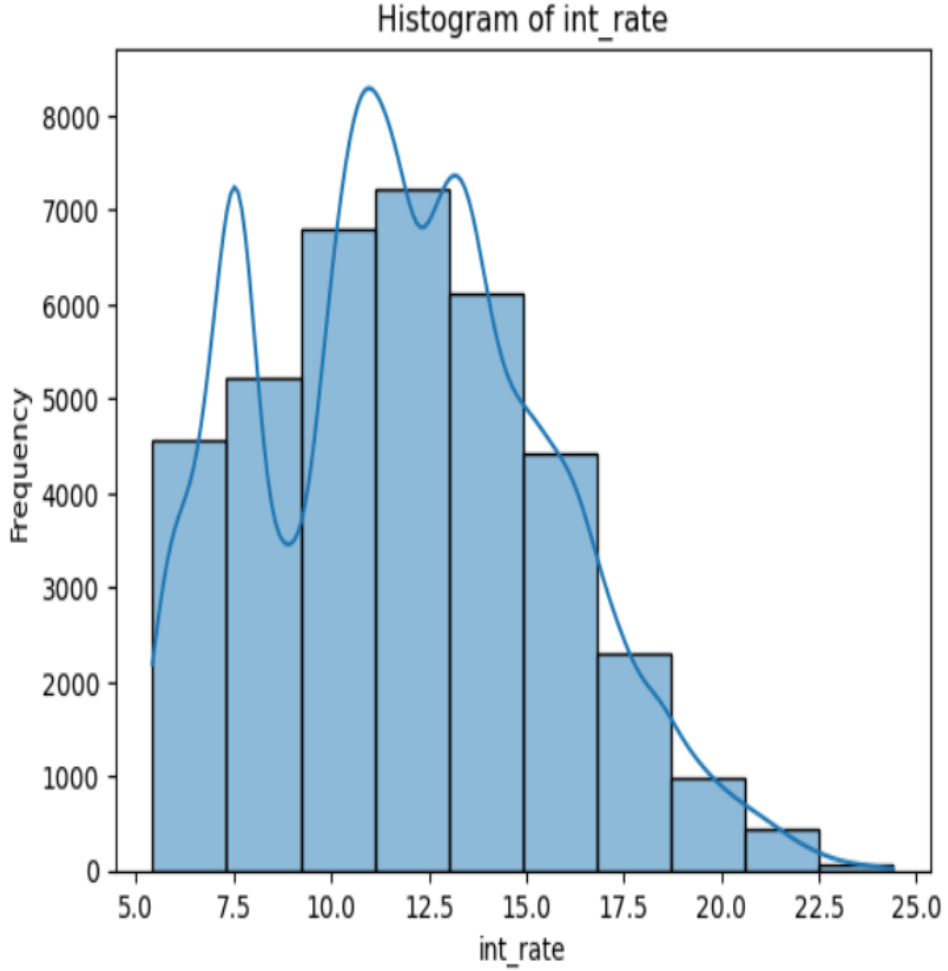
- **Correlation Analysis:**

  •Scatter plot showing relationships between features. Notable correlations include interest rate with loan amount and loan amount with annual income.

# UNIVARIATE ANALYSIS: LOAN AMOUNT AND INTEREST RATE DISTRIBUTION




**Inference:** The loan amount varies from 500 to 35000 with a mean of 9600.

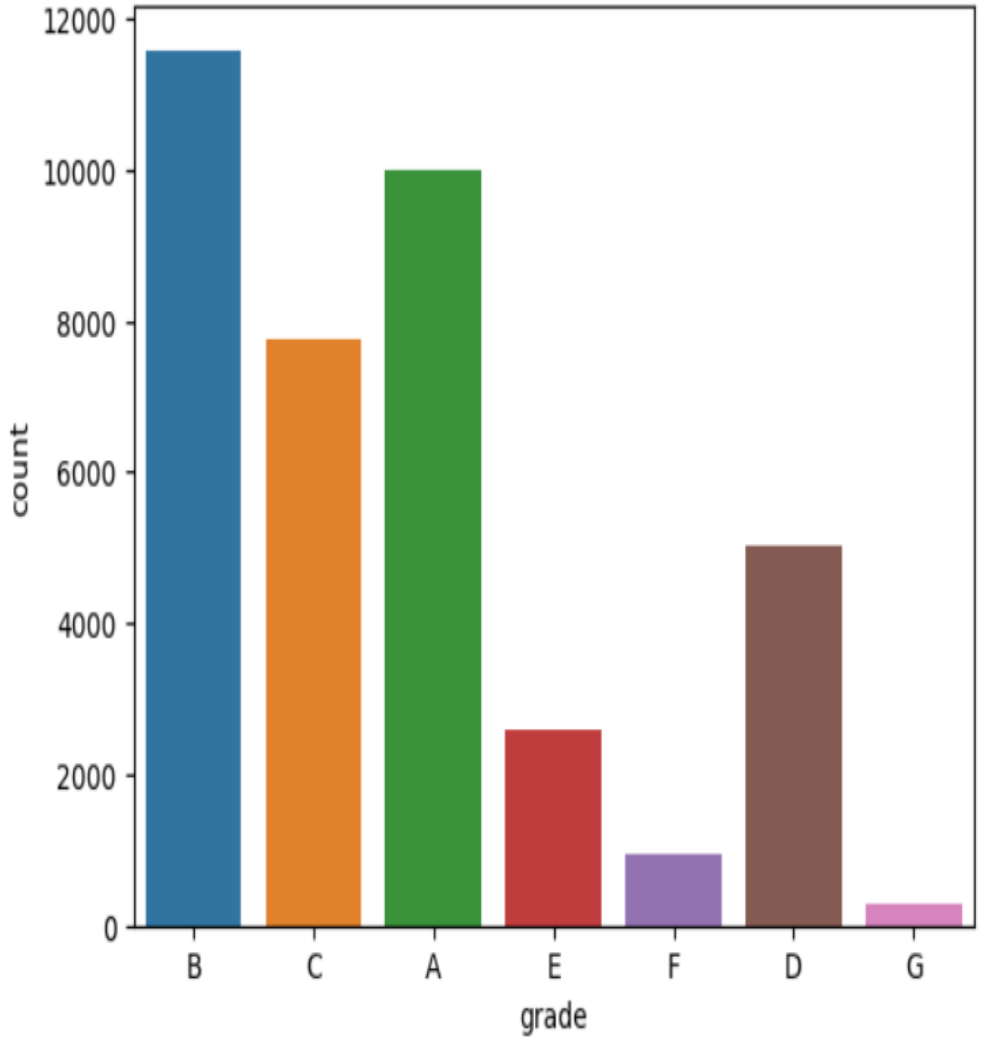



**Inference:** The highest interest rate is more than 24% and minimum is just more than 5%. Most Borrowers have taken loan at 12.5% and least at just less than 25%.

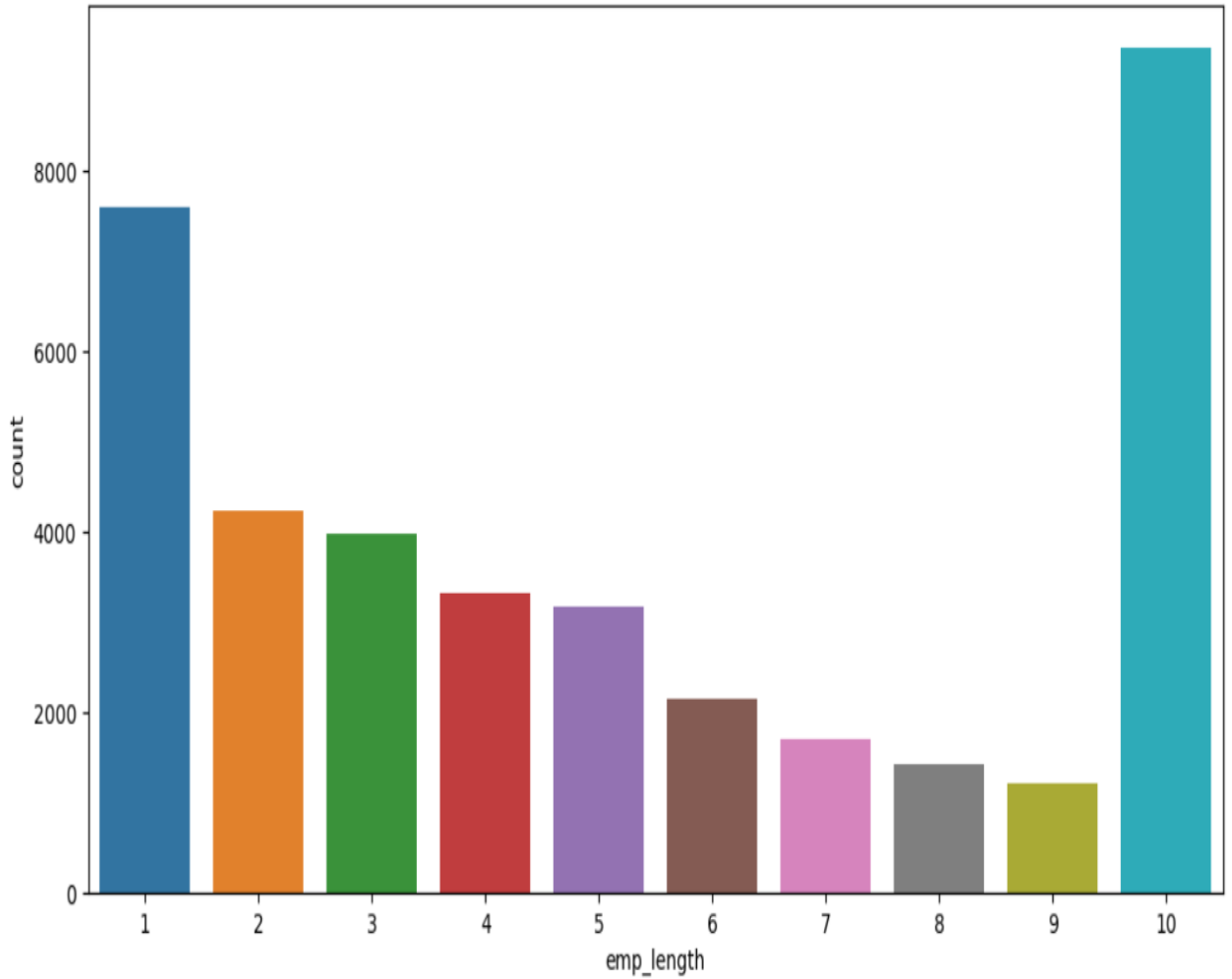
iit-b ज्ञानमुत्तमम

upGrad

# UNIVARIATE ANALYSIS: GRADE AND EMPLOYEMENT LENGTH DISTRIBUTION




**Inference:** The "B" grade loans are highest with more than 30% and "G" lowest.

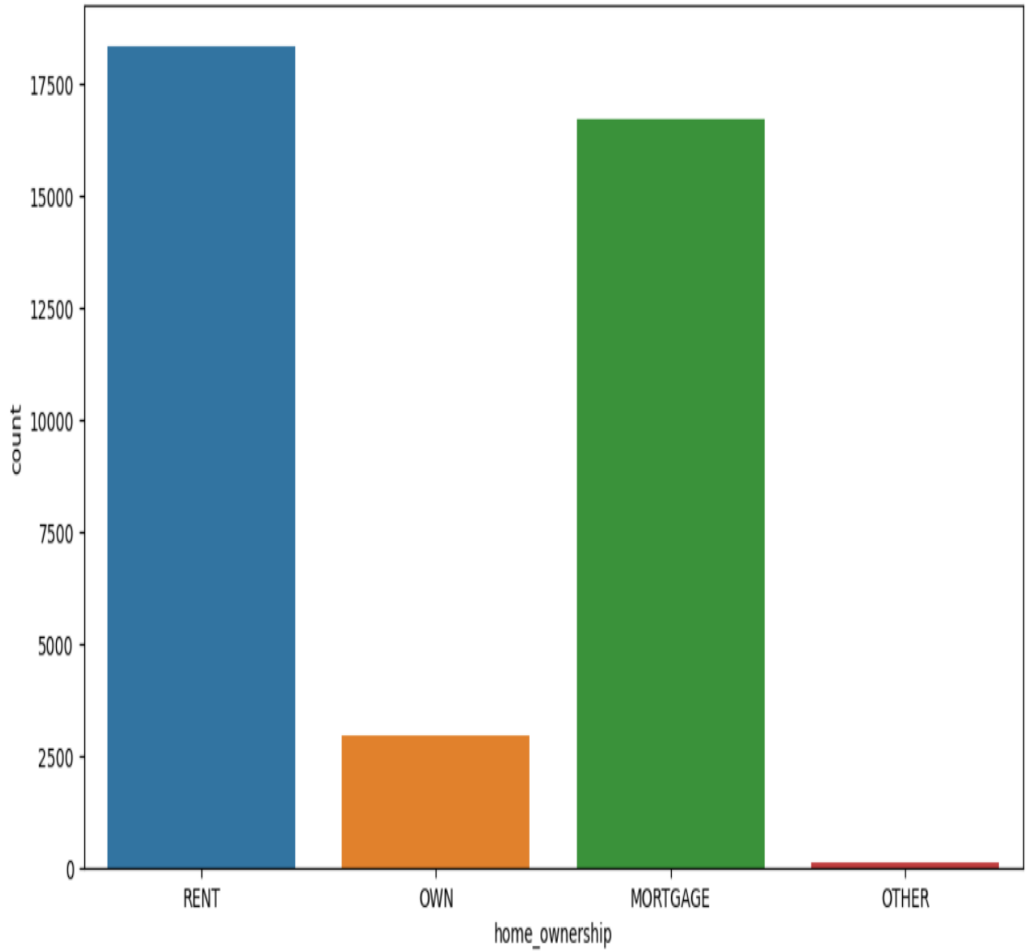



**Inference:** Most of the borrowers have work experience greater than 10 years.

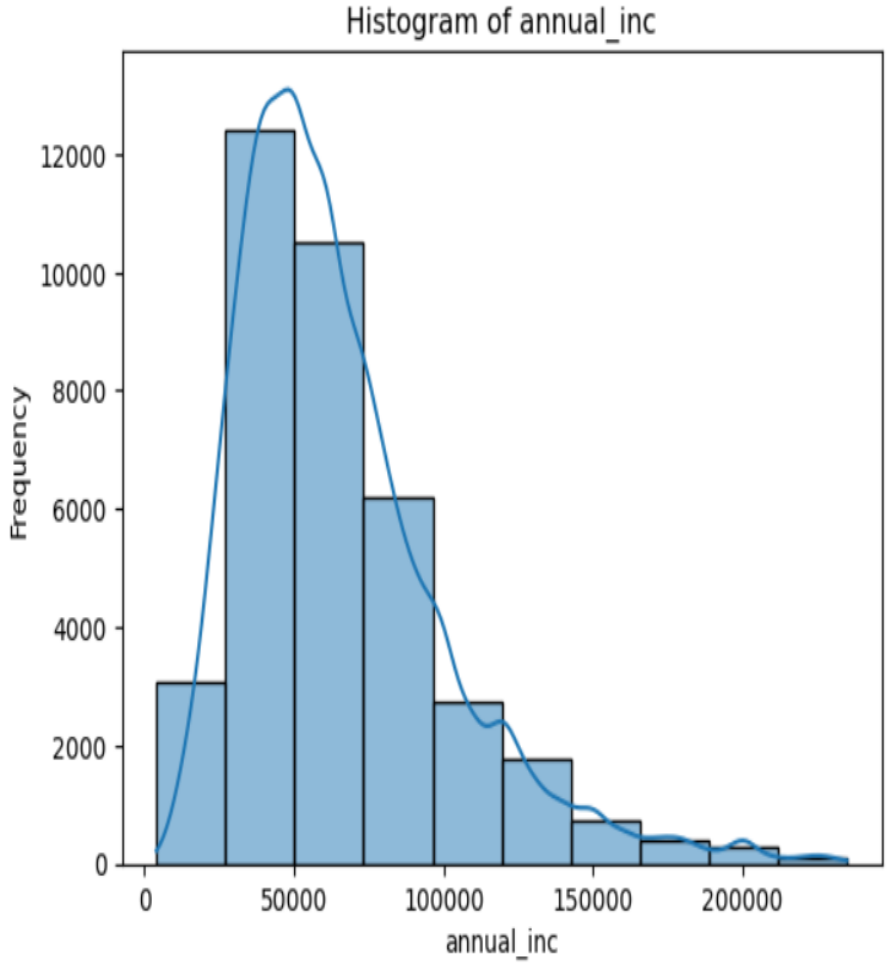
iiit-b ज्ञानमुत्तमम

upGrad

# UNIVARIATE ANALYSIS: ANNUAL INCOME AND HOME OWNERSHIP DISTRIBUTION




**Inference:** Most of the borrowers are not owing house but on either mortgage or rent.

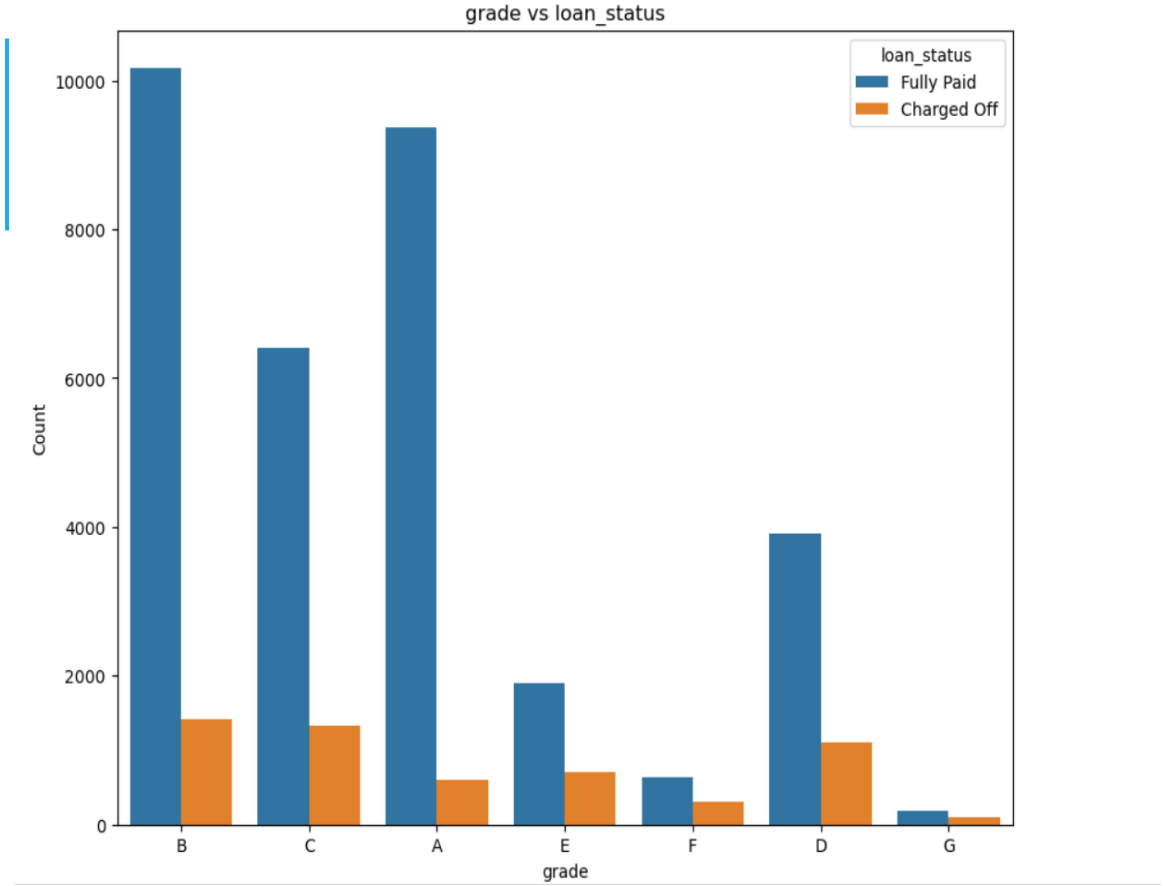



**Inference:** Most of the borrowers are in low income range of around 60k.

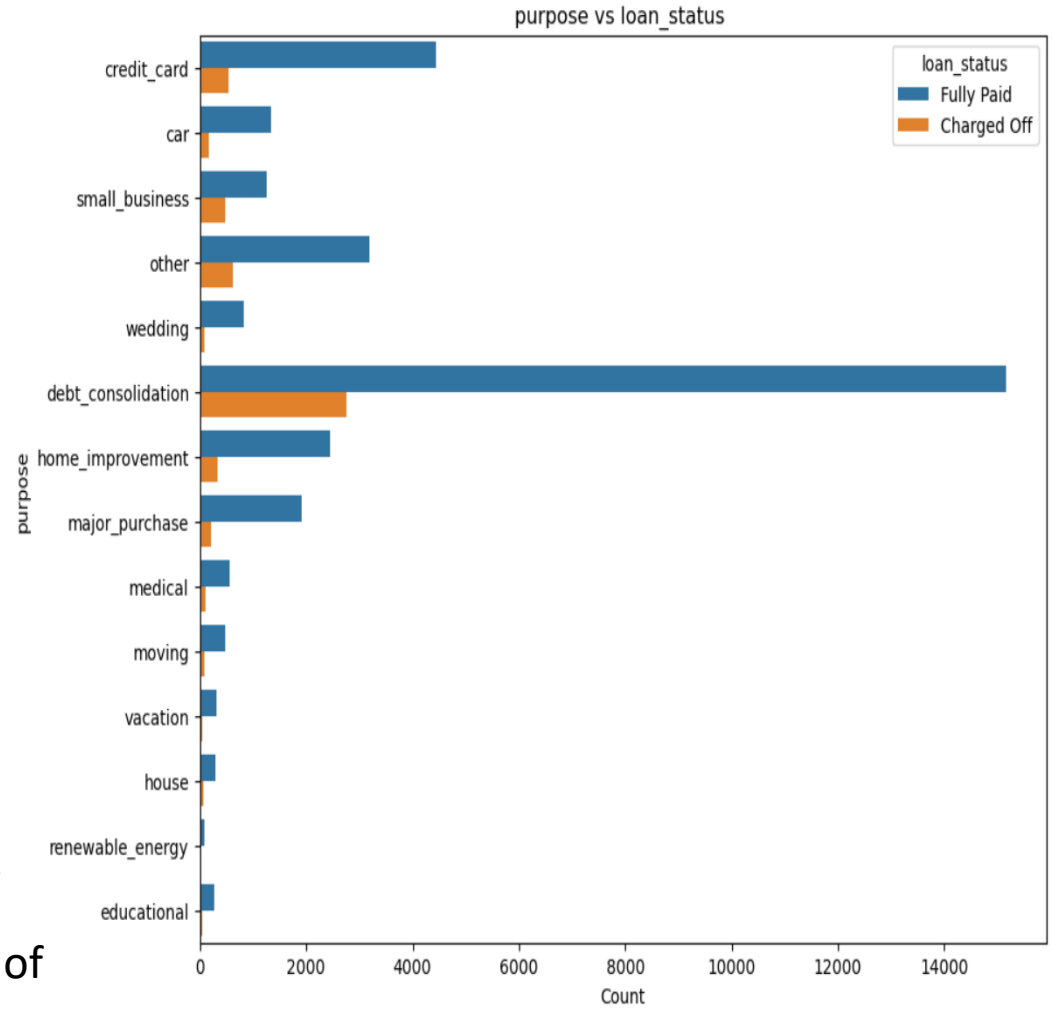
iit-b ज्ञानमुत्तमम

upGrad

# SEGMENTED UNIVARIATE ANALYSIS: GRADE VS LOAN STATUS AND PURPOSE VS LOAN STATUS

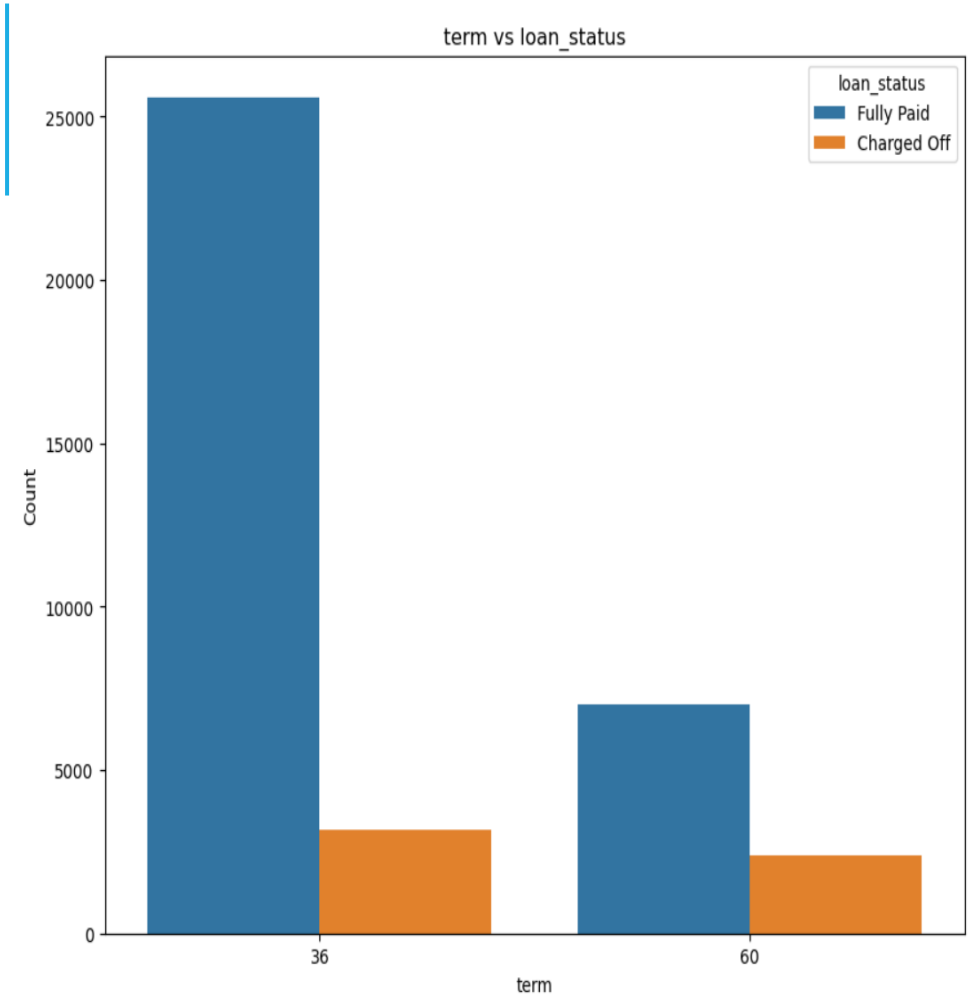



**Inference:** we can say that Grade F,D,G have higher chance of defaulting with G the highest.

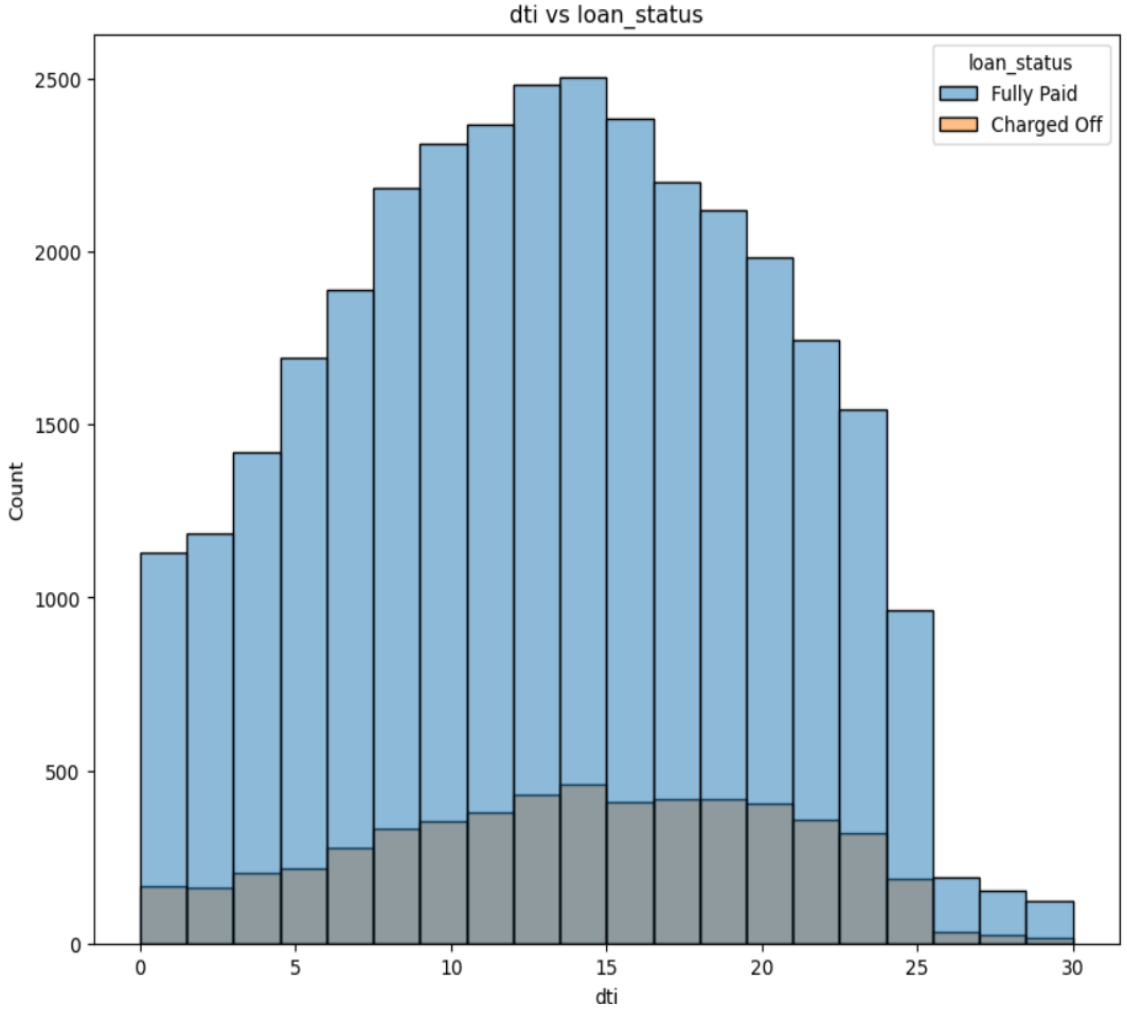



**Inference:** we can say that Borrowers who take small_business loans have higher chance of defaulting.

# SEGMENTED UNIVARIATE ANALYSIS: TERM VS LOAN STATUS AND DTI VS LOAN STATUS




**Inference:** we can say that 60 month loan term has higher chance of defaulting.




**Inference:** we can say that loans in dti of 15 has higher chance of defaulting.

# DERIVED METRICS

**New Metrics Created:**

- Created metrics such as issue_year, issue_month etc. to add depth to the analysis.
- Create groups such as int_rate_groups, annual_inc_groups, open_acc_groups etc. to help understand the data better.

**Utilization of Metrics:**

- These metrics were used to enhance the analysis of borrower risk profiles and identify patterns related to defaults.

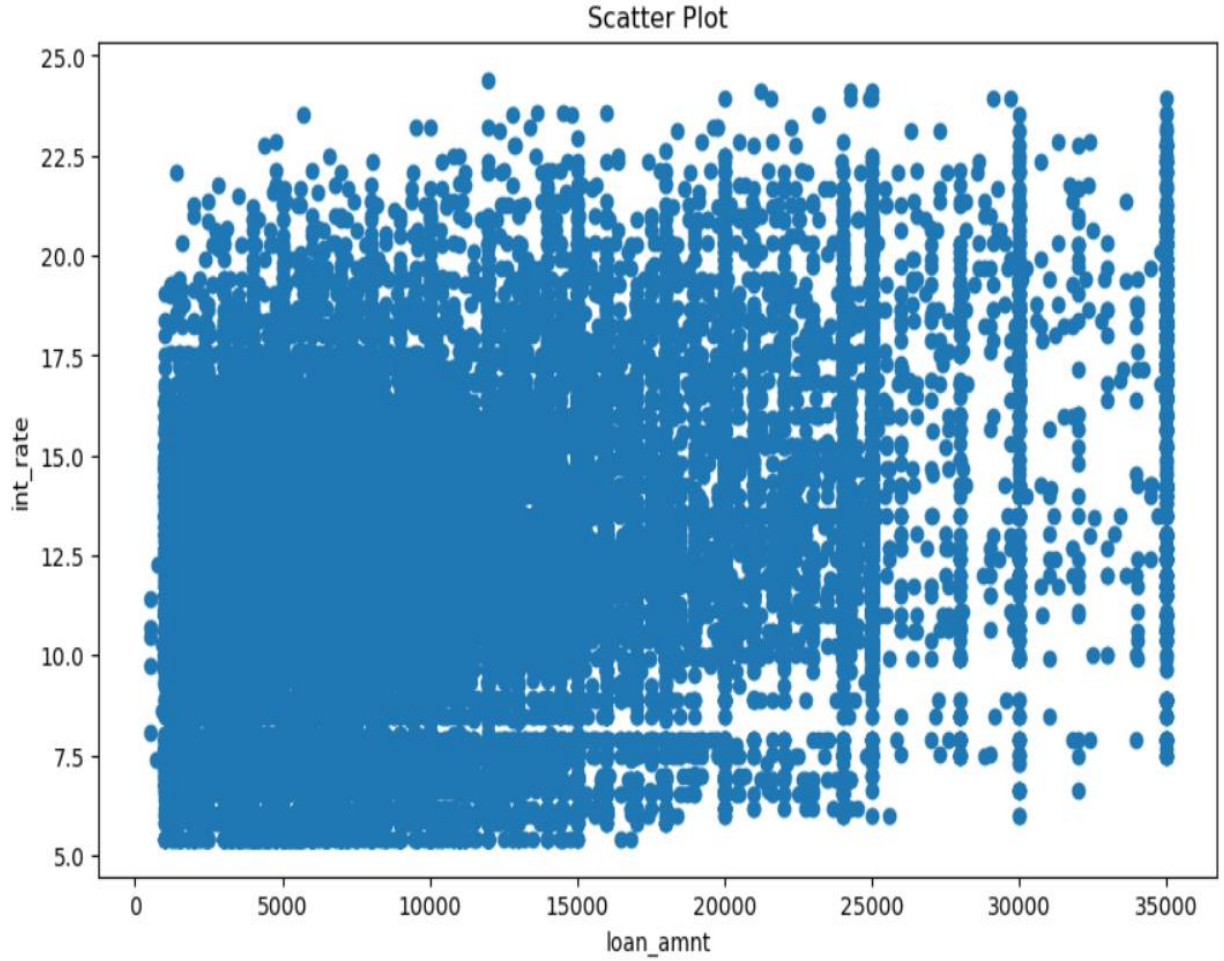
# BIVARIATE METRICS

**Important Combinations:**

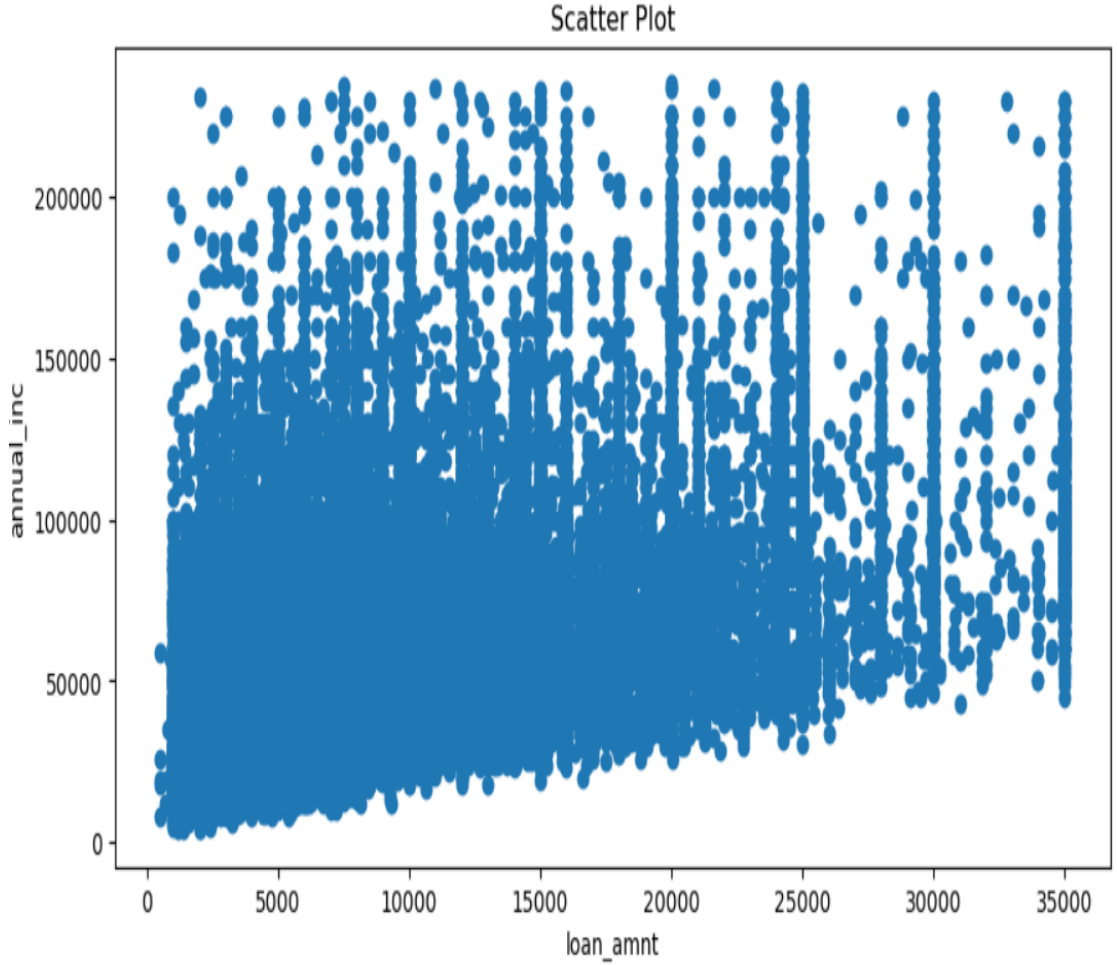
• Loan Amount vs. Interest Rate: Higher loan amounts and interest rates are correlated with higher default rates.
• Grade vs. Default Rate: Lower grade loans (D,F,G etc.) show higher default rates compared to higher grades (A, B).

**Insights:**
• Bivariate analysis helps identify significant patterns and interactions between variables that contribute to loan defaults.

# BIVARIATE ANALYSIS: LOAN AMOUNT VS INTEREST RATE AND LOAN AMOUNT VS ANNUAL INCOME

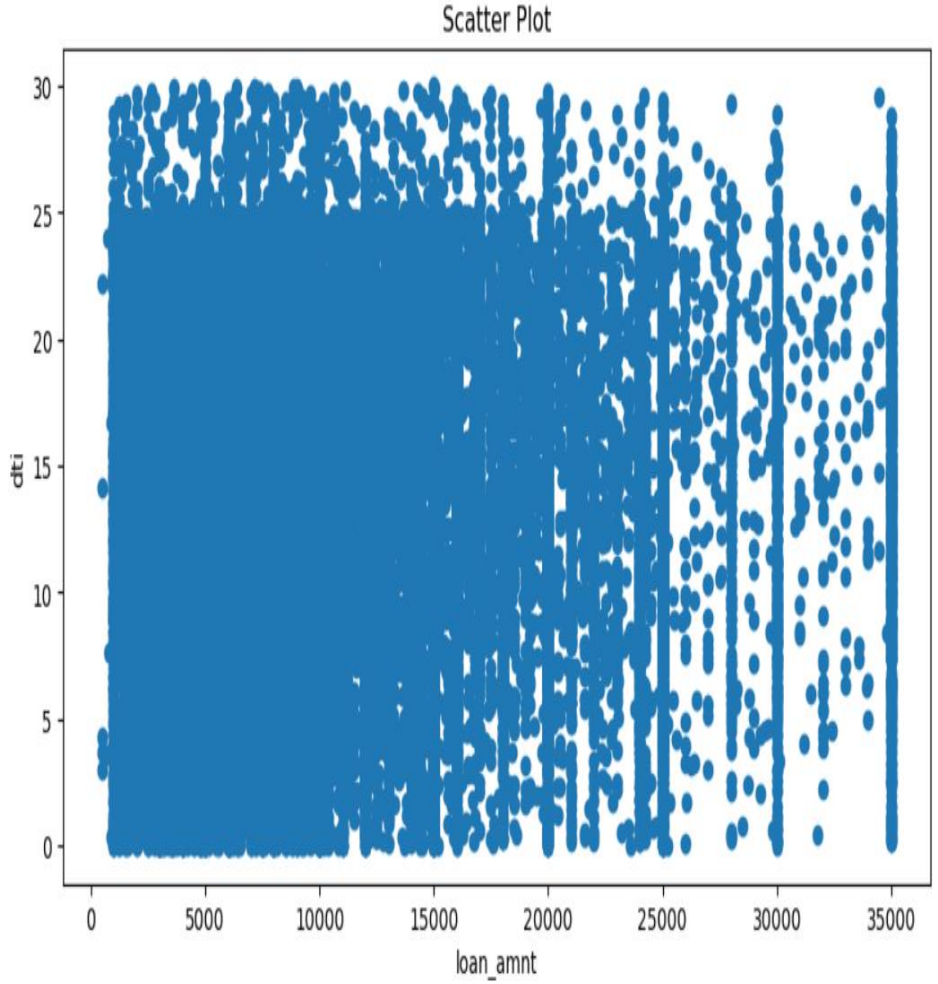



**Inference:** we can say that there is a moderate positive correlation between the loan amount and interest rate. As the loan amount increases, the interest rate tends to increase as well, but not very strongly.

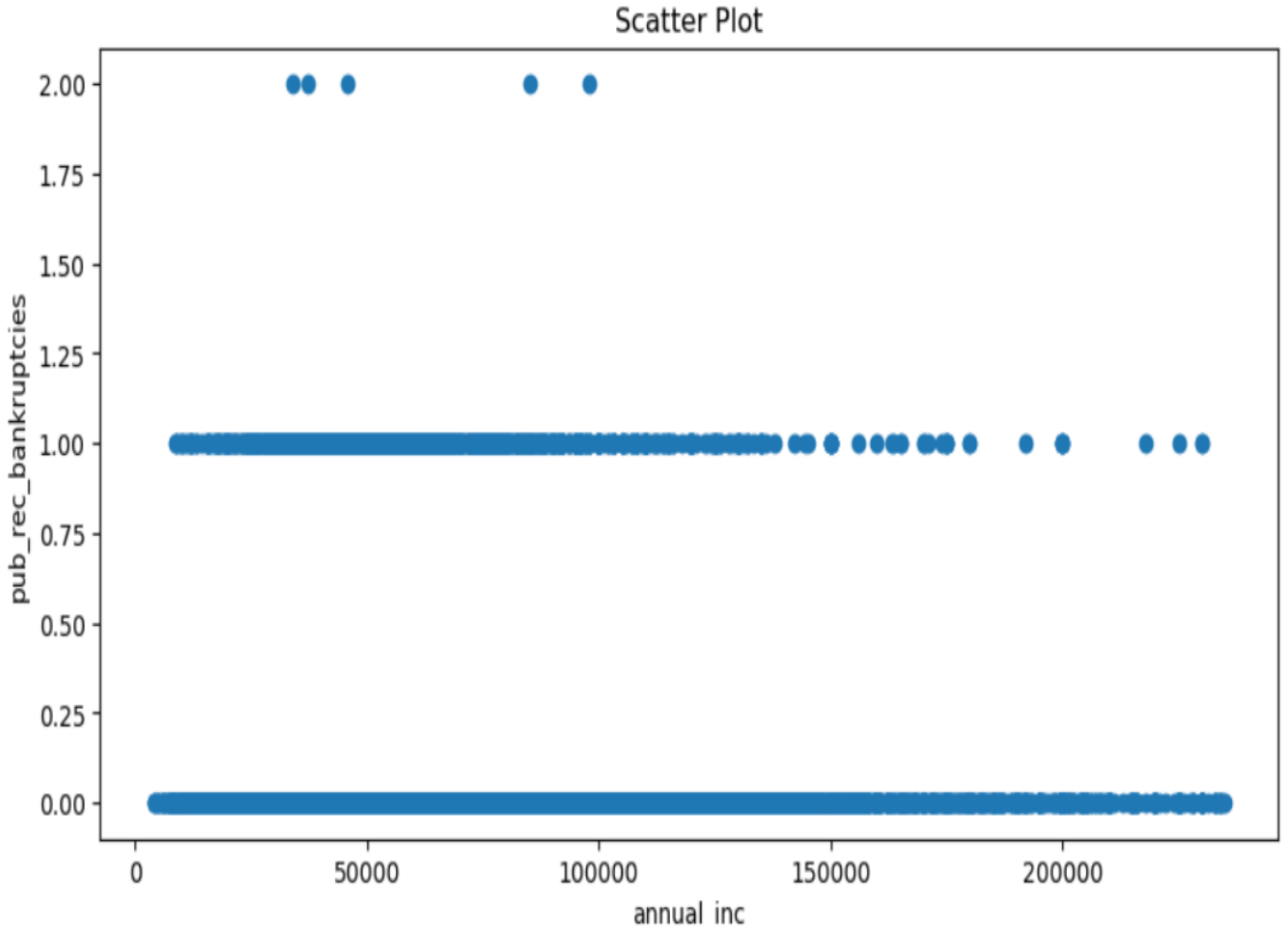



**Inference:** we can say that there is a moderate positive correlation between the loan amount and annual income. As the annual income increases, the loan amount tends to increase as well, but not very strongly.

# BIVARIATE ANALYSIS: LOAN AMOUNT VS DTI AND ANNUAL INCOME VS PUBLIC RECORD OF BANKRUPTCIES




**Inference**: we can say that this indicates a very weak positive correlation between the loan amount and the debt-to-income ratio. As the debt-to-income ratio increases, the loan amount does not show a strong tendency to increase or decrease.




**Inference:** we can say that this indicates an extremely weak negative correlation between annual income and the number of public record bankruptcies. As the number of public record bankruptcies increases, the annual income tends to decrease slightly, but the relationship is almost negligible.

# KEY INSIGHTS

**Factors Influencing Defaults:** Summary of key factors identified through EDA.

- Loan amount: Larger loans have a higher likelihood of default.
- Interest rate: Higher interest rates correlate with increased default rates.
- Loan grade: Lower grades (D,F,G) are more prone to default.
- Annual income: Lower annual incomes are associated with higher default rates.

**Patterns Observed:**

- Seasonal trends and borrower profiles that are more likely to default were identified.
- Specific borrower characteristics, such as employment length and debt-to-income ratio, were significant predictors of default.

# RECOMMENDATIONS

**Risk Mitigation:**

- Tighten Credit Policies: Implement stricter credit checks and criteria for high-risk applicants.
- Offering better terms to low-risk applicants.

**Further Analysis:**

- Segmentation: Further segment borrower profiles for targeted analysis and risk assessment.
- Additional Data: Collect additional data on employment history, credit scores, and other relevant factors to improve risk assessment and predictive models.

# CONCLUSION

**Summary:**

- The analysis identified key factors influencing loan defaults, including interest rate, loan grade, dti, loan purpose, public record bankruptcies, employment length and annual income.

- Recommendations for mitigating risk and improving lending decisions were provided based on the findings.

**Future Work:**

- Potential development of machine learning models to predict loan defaults.
- Exploration of additional data sources to enhance the analysis and improve predictive accuracy.