# Lending Club Case Study

*Harshit Jain*
*RaviChandra*

# Summary

- Problem Statement
- Input Overview
- Data Cleaning
- Exploratory Data Analysis
- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis
- Correlation Analysis
- Conclusions
- Recommendations

# Problem Statement

In a certain consumer finance company which specializes in lending various types of loans to urban customers receives loan application and it has to make a decision for loan approval based on applicant's profile.

The bank has to deal with two type of risks associated to these applications

A. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

B. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Based on the above risks, the bank would take two below decisions

A. **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

B. **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)

***<u>Objective : Use EDA to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.</u>***

# Data Summary & Insights

[2]:
```python
# Loading the data from CSV file into a DataFrame
loan_data = pd.read_csv('loan.csv')

# Checking the dimensions of the DataFrame
loan_data.shape
```

[2]: (39717, 111)

[3]:
```python
# Displaying summary statistics of the data
loan_data.describe()  # Summary of numeric columns
```

[3]:

| | id | member_id | loan_amnt | funded_amnt | funded_amnt_inv | installment | annual_inc | dti | delinq_2yrs | inq_last_6mths | ... | num_tl_9 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 3.971700e+04 | 3.971700e+04 | 39717.000000 | 39717.000000 | 39717.000000 | 39717.000000 | 3.971700e+04 | 39717.000000 | 39717.000000 | 39717.000000 | ... | |
| **mean** | 6.831319e+05 | 8.504636e+05 | 11219.443815 | 10947.713196 | 10397.448868 | 324.561922 | 6.896893e+04 | 13.315130 | 0.146512 | 0.869200 | ... | |
| **std** | 2.106941e+05 | 2.656783e+05 | 7456.670694 | 7187.238670 | 7128.450439 | 208.874874 | 6.379377e+04 | 6.678594 | 0.491812 | 1.070219 | ... | |
| **min** | 5.473400e+04 | 7.069900e+04 | 500.000000 | 500.000000 | 0.000000 | 15.690000 | 4.000000e+03 | 0.000000 | 0.000000 | 0.000000 | ... | |
| **25%** | 5.162210e+05 | 6.667800e+05 | 5500.000000 | 5400.000000 | 5000.000000 | 167.020000 | 4.040400e+04 | 8.170000 | 0.000000 | 0.000000 | ... | |
| **50%** | 6.656650e+05 | 8.508120e+05 | 10000.000000 | 9600.000000 | 8975.000000 | 280.220000 | 5.900000e+04 | 13.400000 | 0.000000 | 1.000000 | ... | |
| **75%** | 8.377550e+05 | 1.047339e+06 | 15000.000000 | 15000.000000 | 14400.000000 | 430.780000 | 8.230000e+04 | 18.600000 | 0.000000 | 1.000000 | ... | |
| **max** | 1.077501e+06 | 1.314167e+06 | 35000.000000 | 35000.000000 | 35000.000000 | 1305.190000 | 6.000000e+06 | 29.990000 | 11.000000 | 8.000000 | ... | |

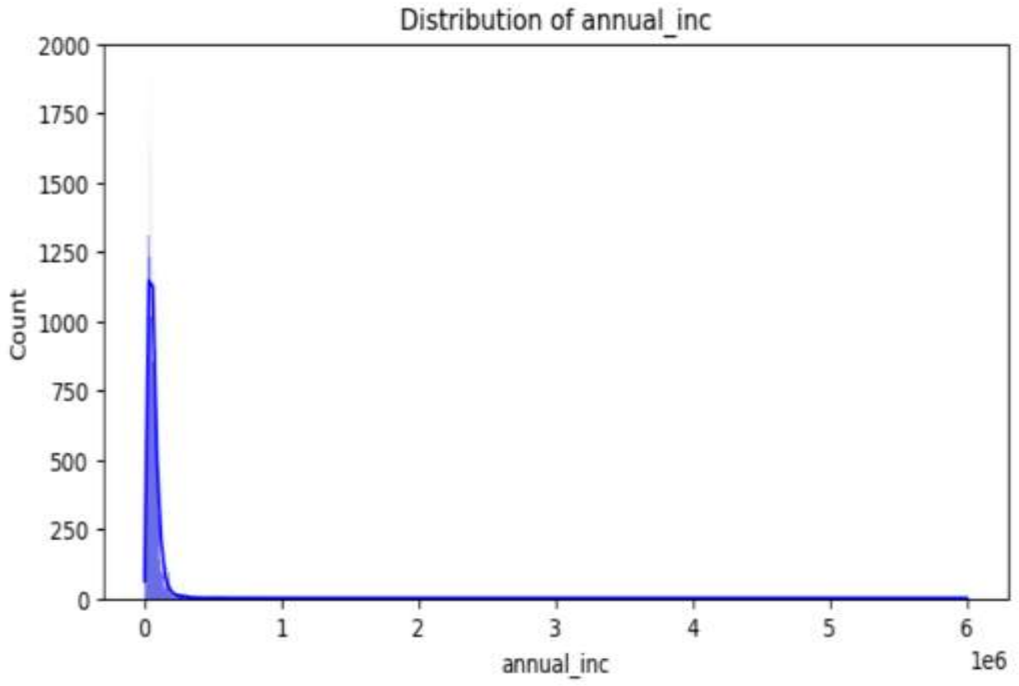
# Data Cleaning Steps

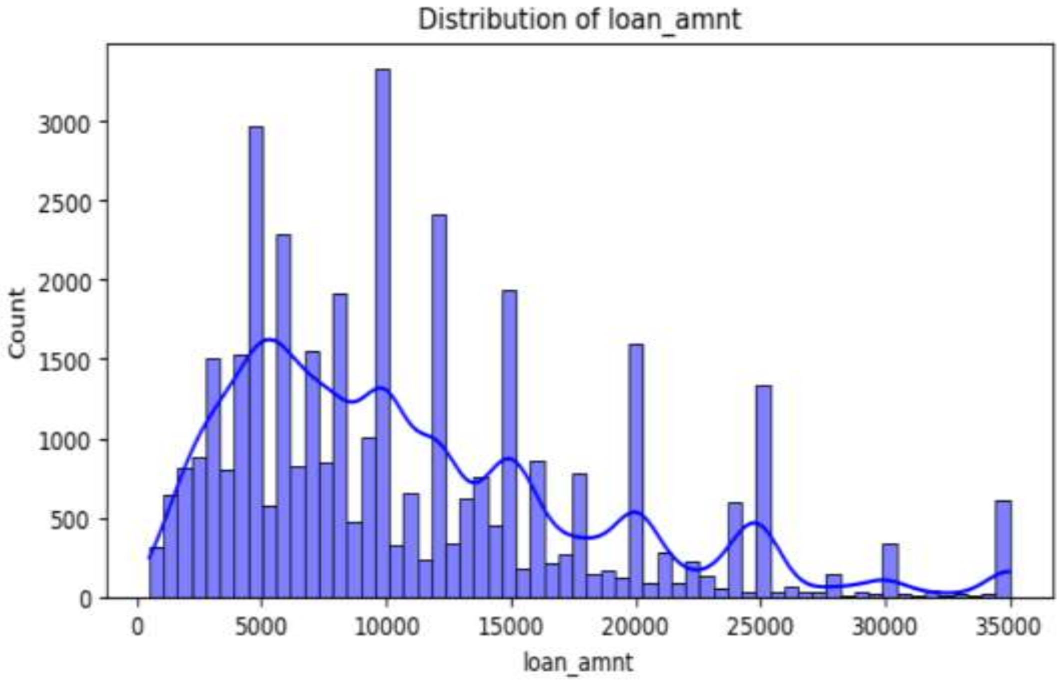
1. Removed records with 'Current' loan status as the tenure is not completed.
2. Removed columns with 100% null values
3. Dropped columns with only one unique value as they don't contribute to analysis.
4. Removed columns irrelevant to loan approval process (post-approval behavioral columns).
5. Converted data types of int_rate, term, loan_amnt, funded_amnt, and issue_d.
6. Handled missing values in emp_length and pub_rec_bankruptcies columns by dropping rows.

# Exploratory Data Analysis

## UNIVARIATE ANALYSIS

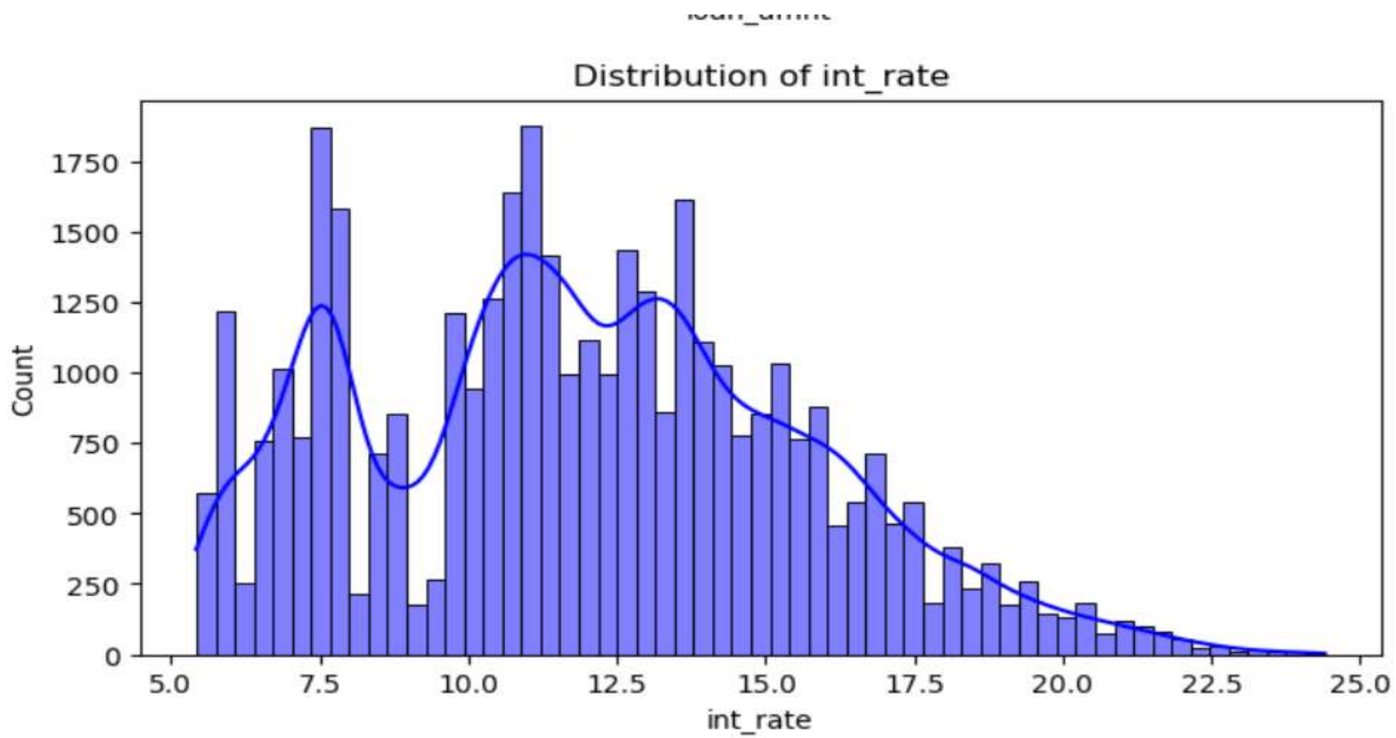
1. Distribution plot for Annual income and Loan amount



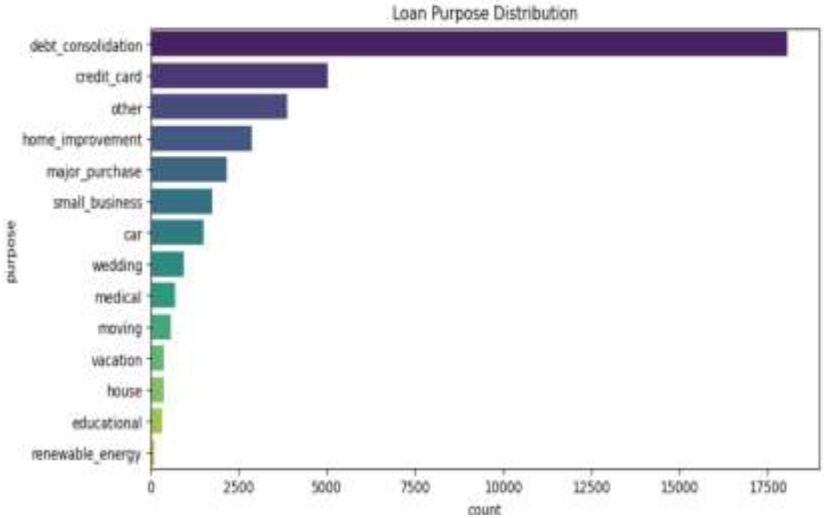



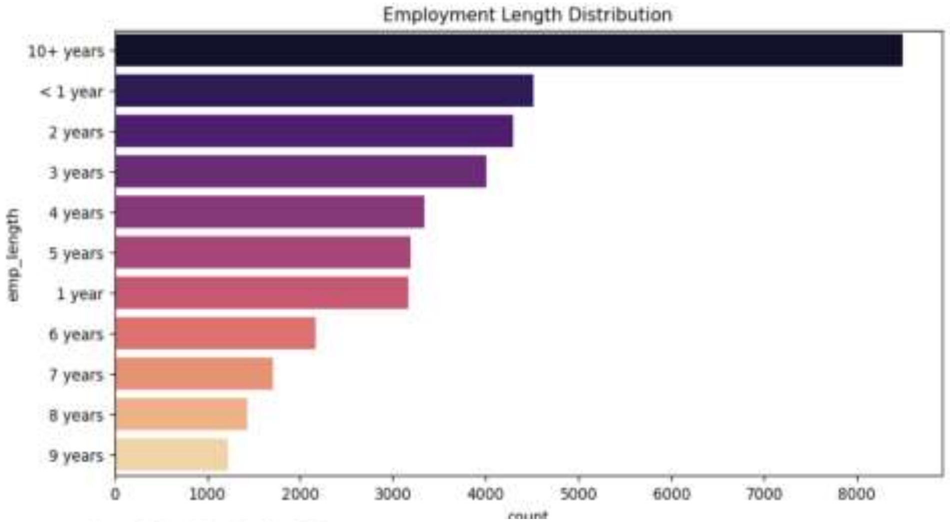
# UNIVARIATE ANALYSIS

2. Distribution plot for Count and Interest Rate

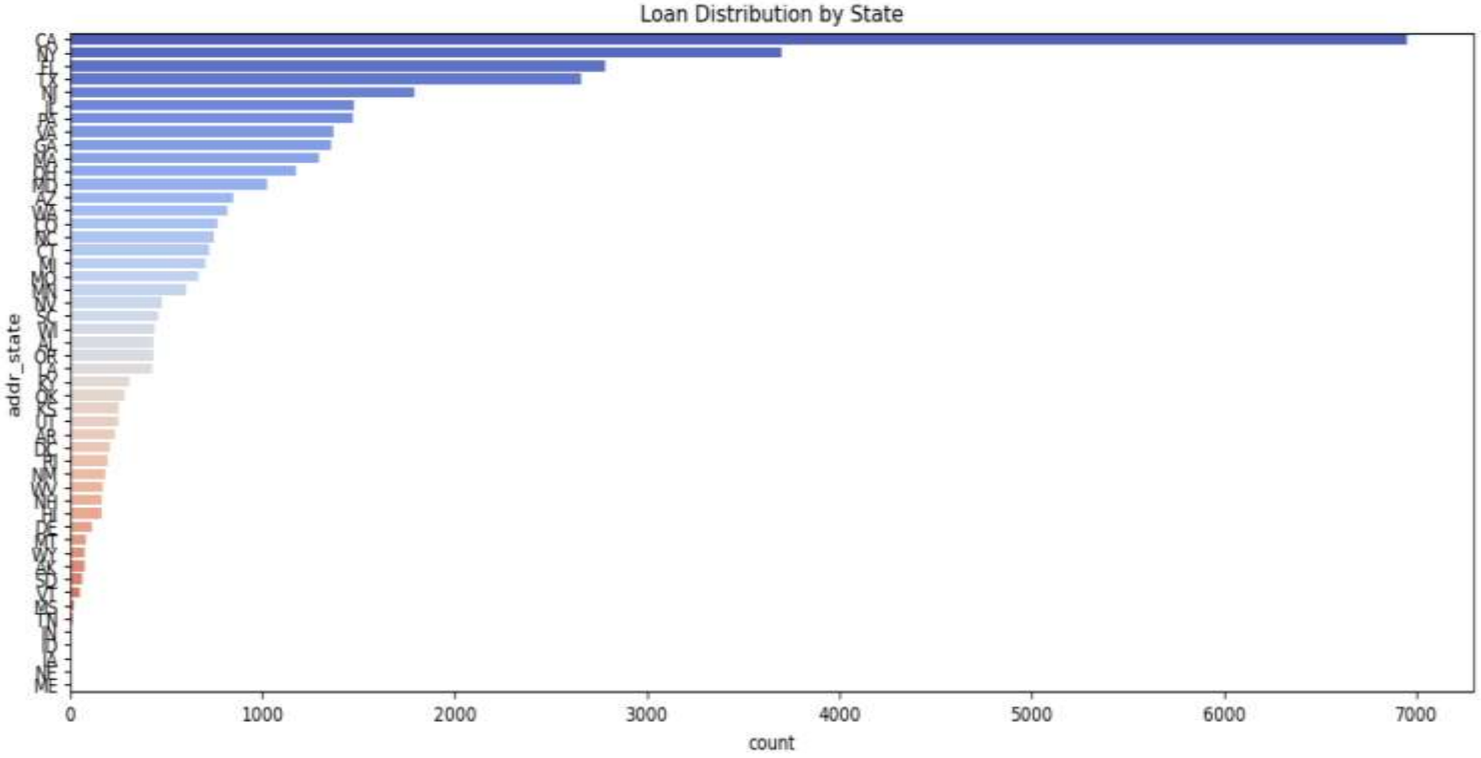


# Univariate Analysis catagorical variable



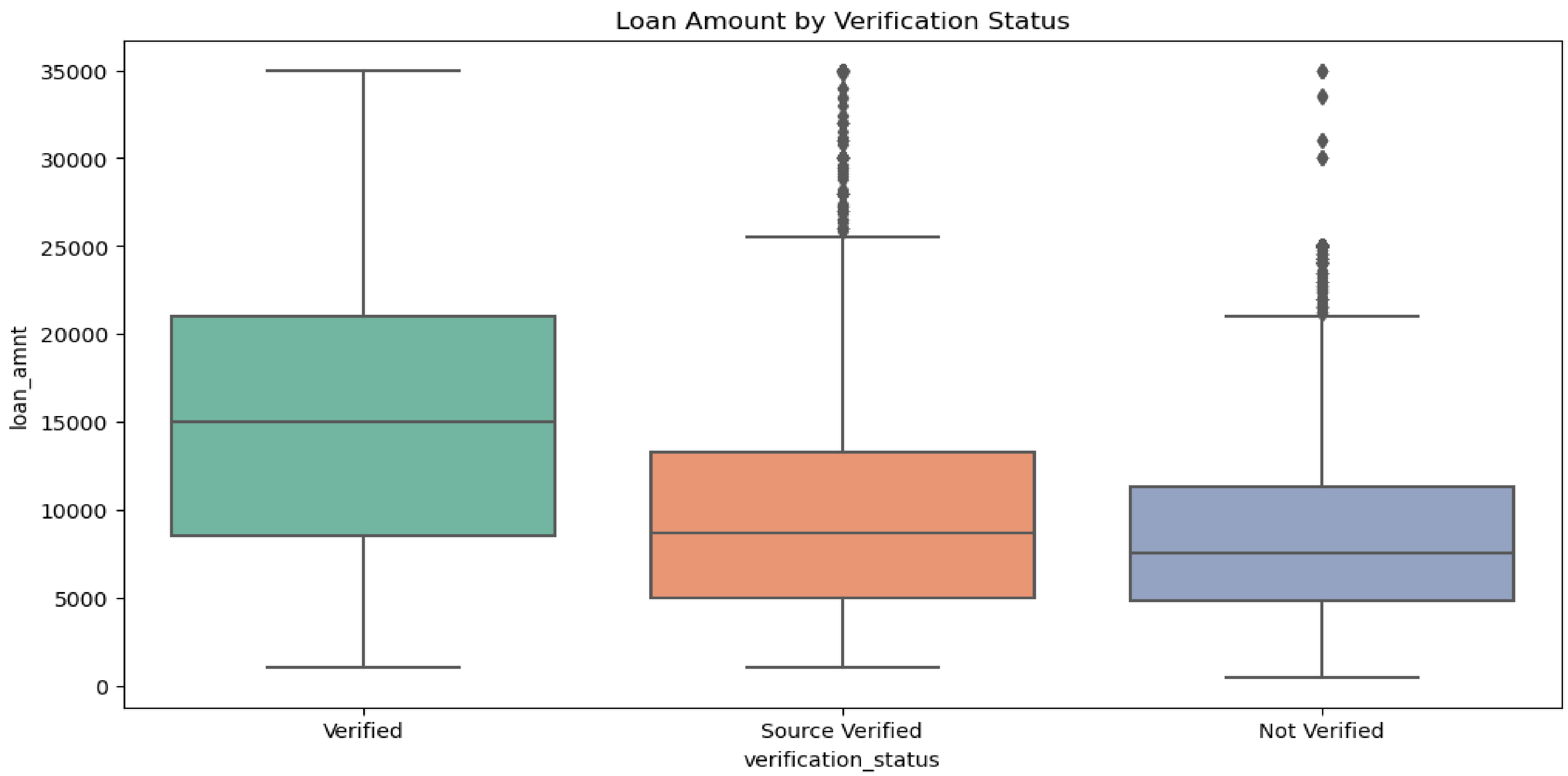




1. This analysis shows that highest number of people took loan for the purpose of debt consolidation
2. The Bar Plot shows that maximum people who takes loan is 10+ years Experienced
3. The Bar Plot shows that Maximum people who is taking Loan is from CA



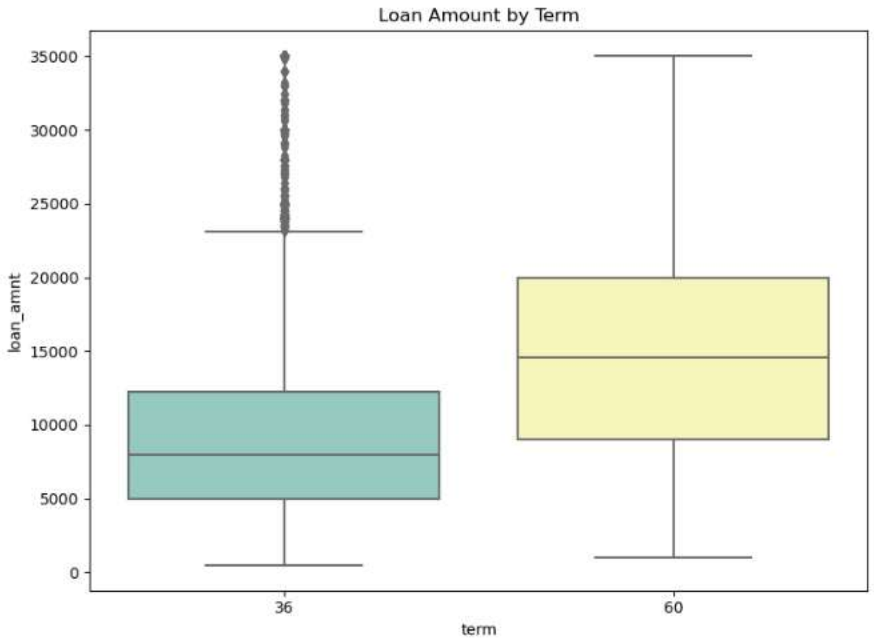
# Segmented Univariate Analysis

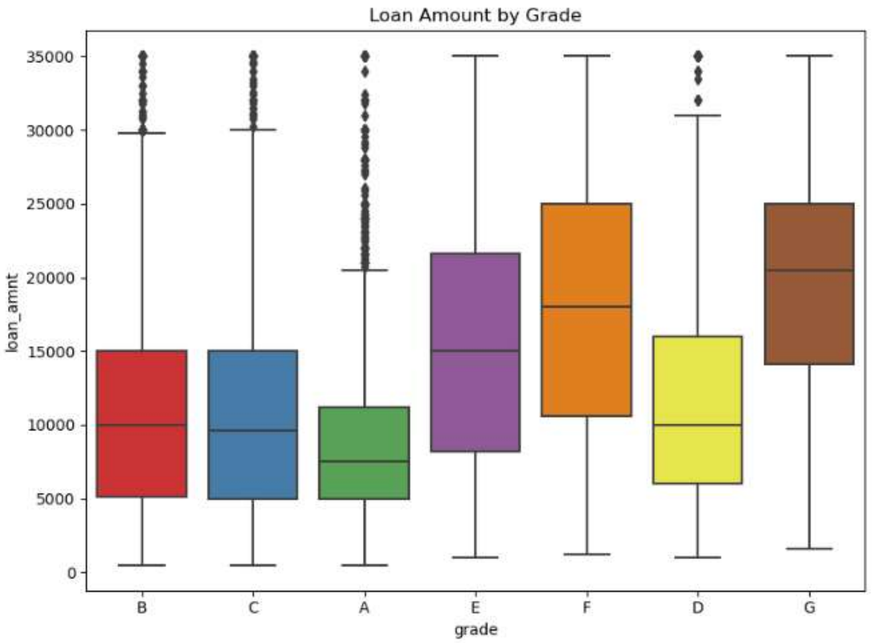
This plot is between Loan Amount vs Verification Status



# Segmented Univariate Analysis

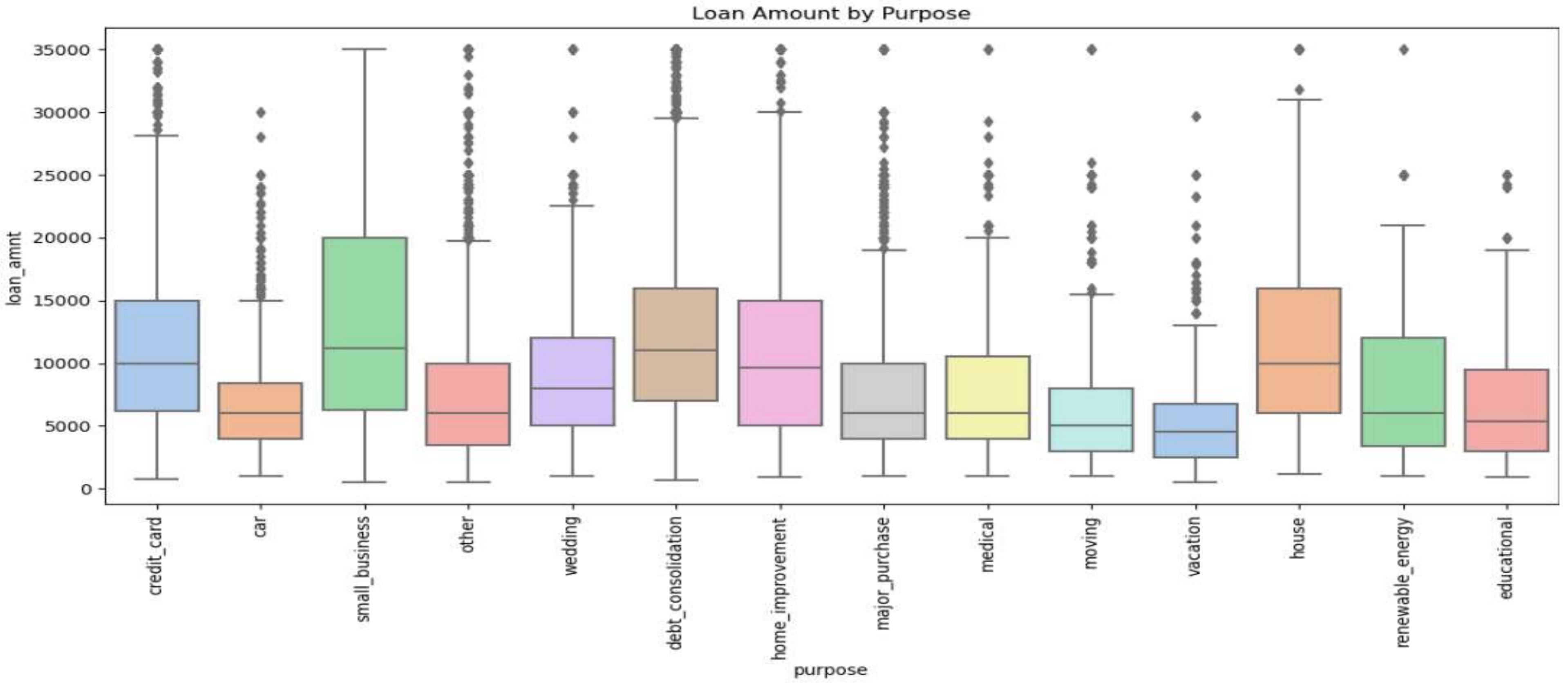
1) This analysis shows Loan Amount v/s Term i.e 60 months of tenure loans are more taken
2) This analysis shows that High Grade has more loan Amount



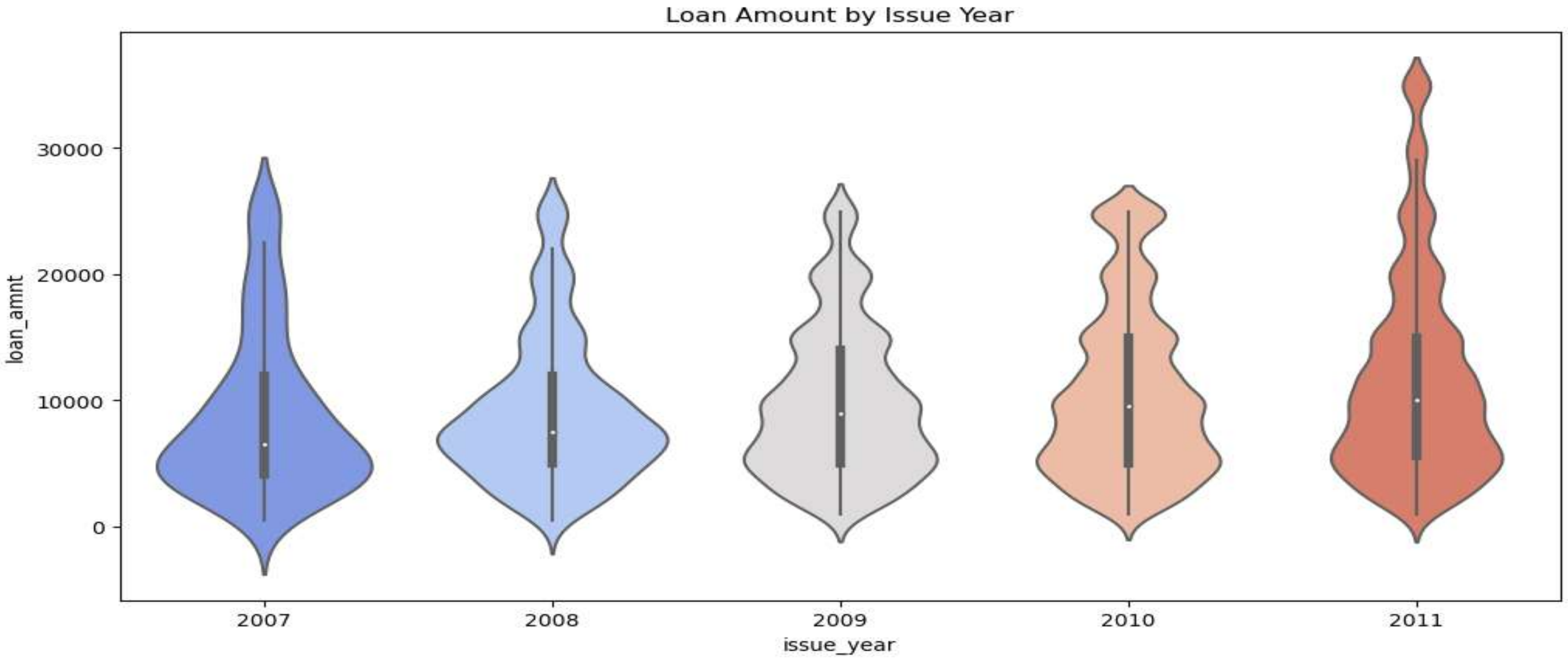
# Segmented Univariate Analysis

This analysis shows that People are taking more loan for creditcard payment, small business, debit consideration & house improvement.

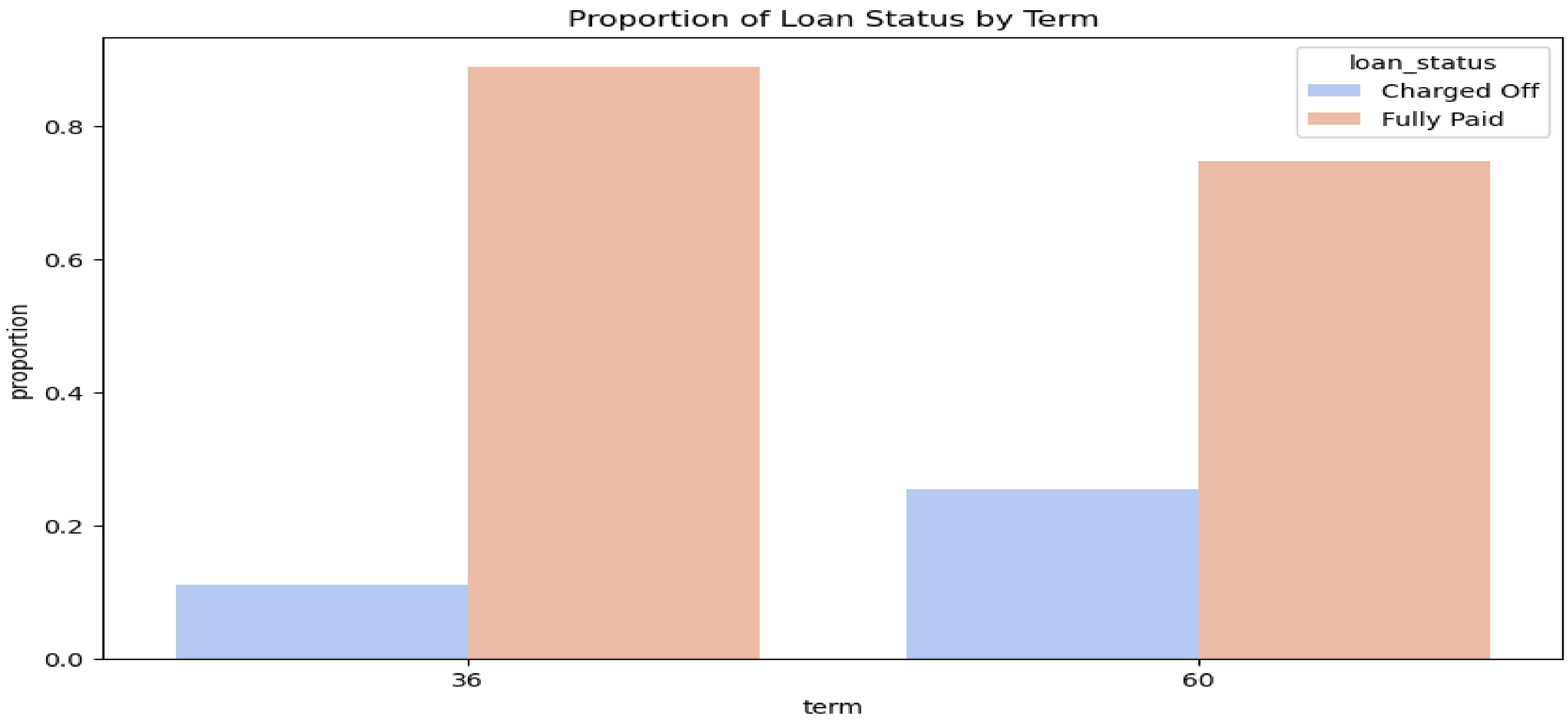


# Bivariate Analysis

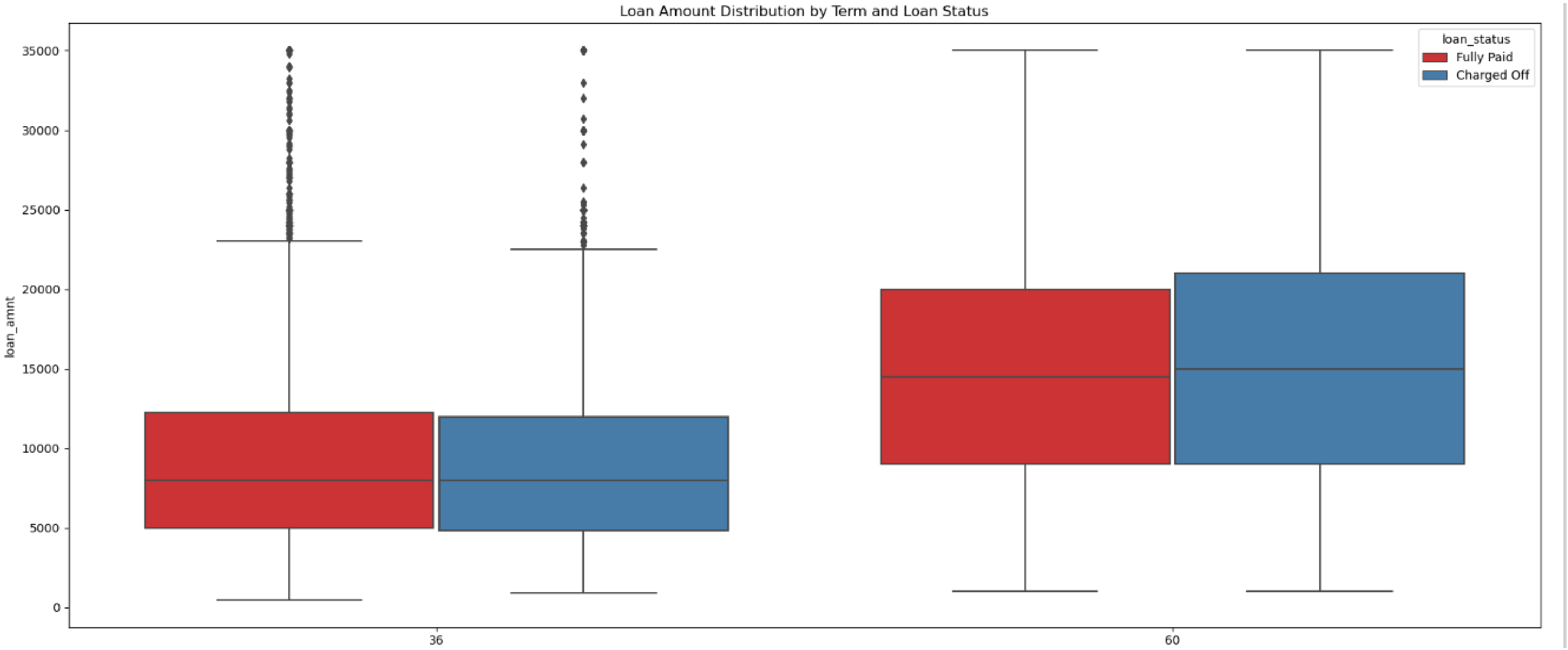
This analysis shows that Loan Amount v/s Issuing Year



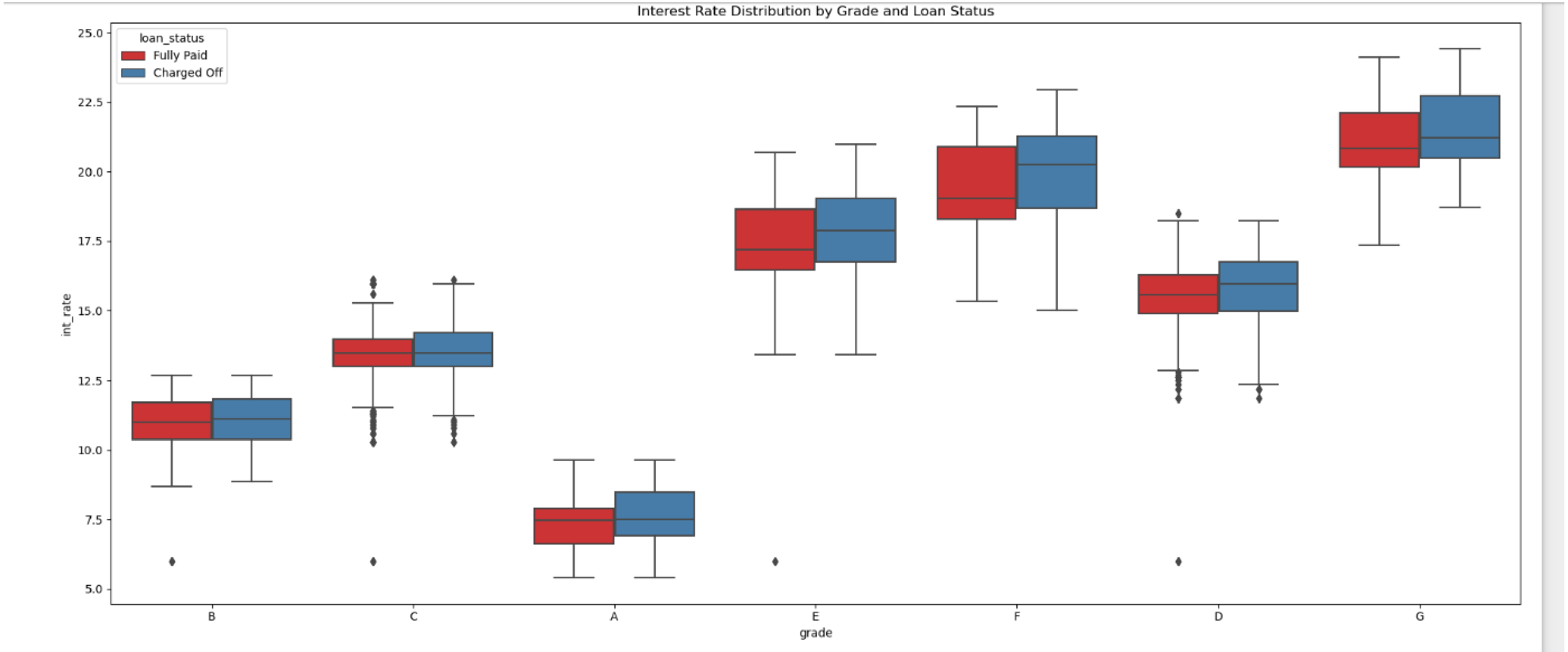
# Analysis on Proportion of Loan Status v/s Term



# Analysis on Loan Amount Distribution By Term v/s Loan Status

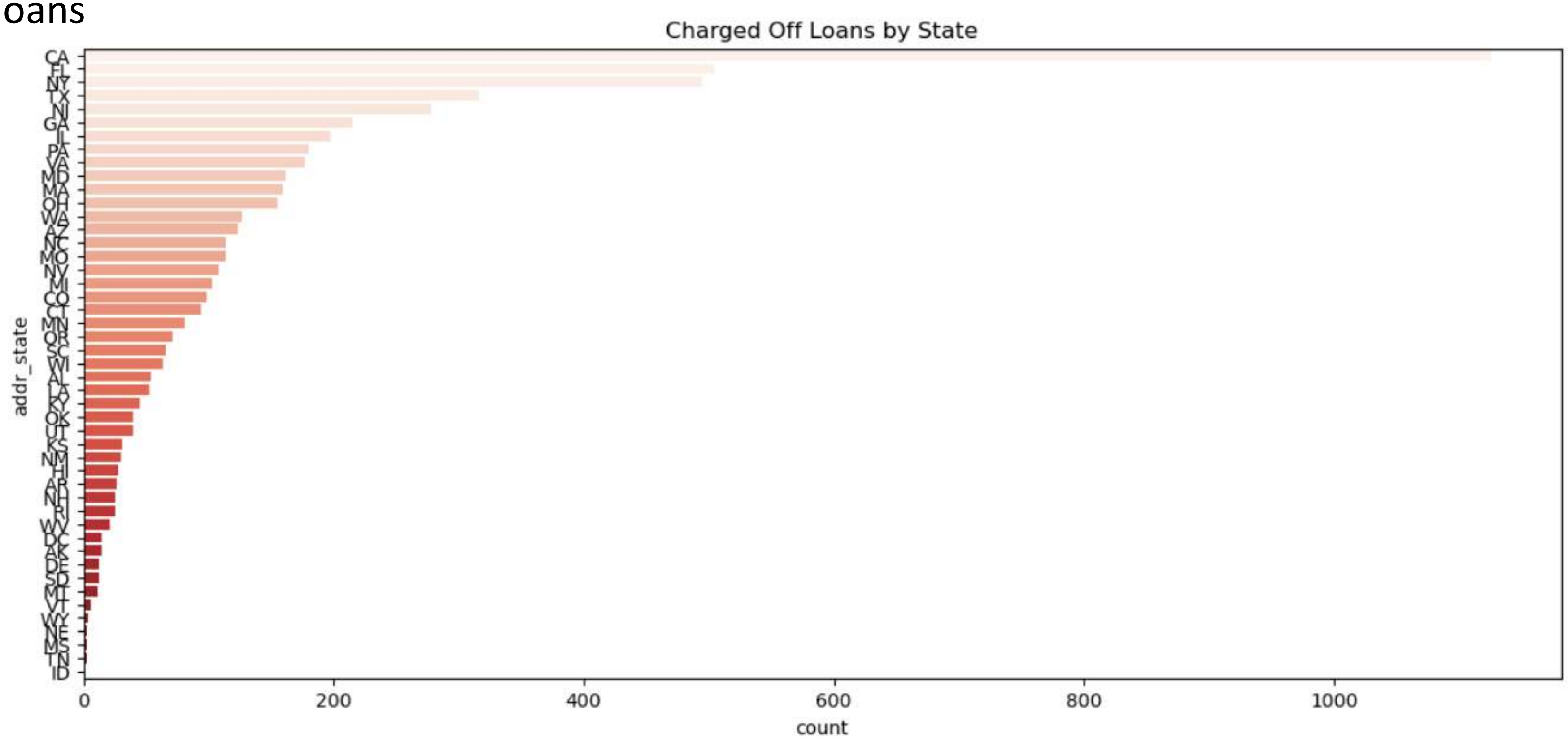


# Analysis on Interest rate By Grade vs Loan Status



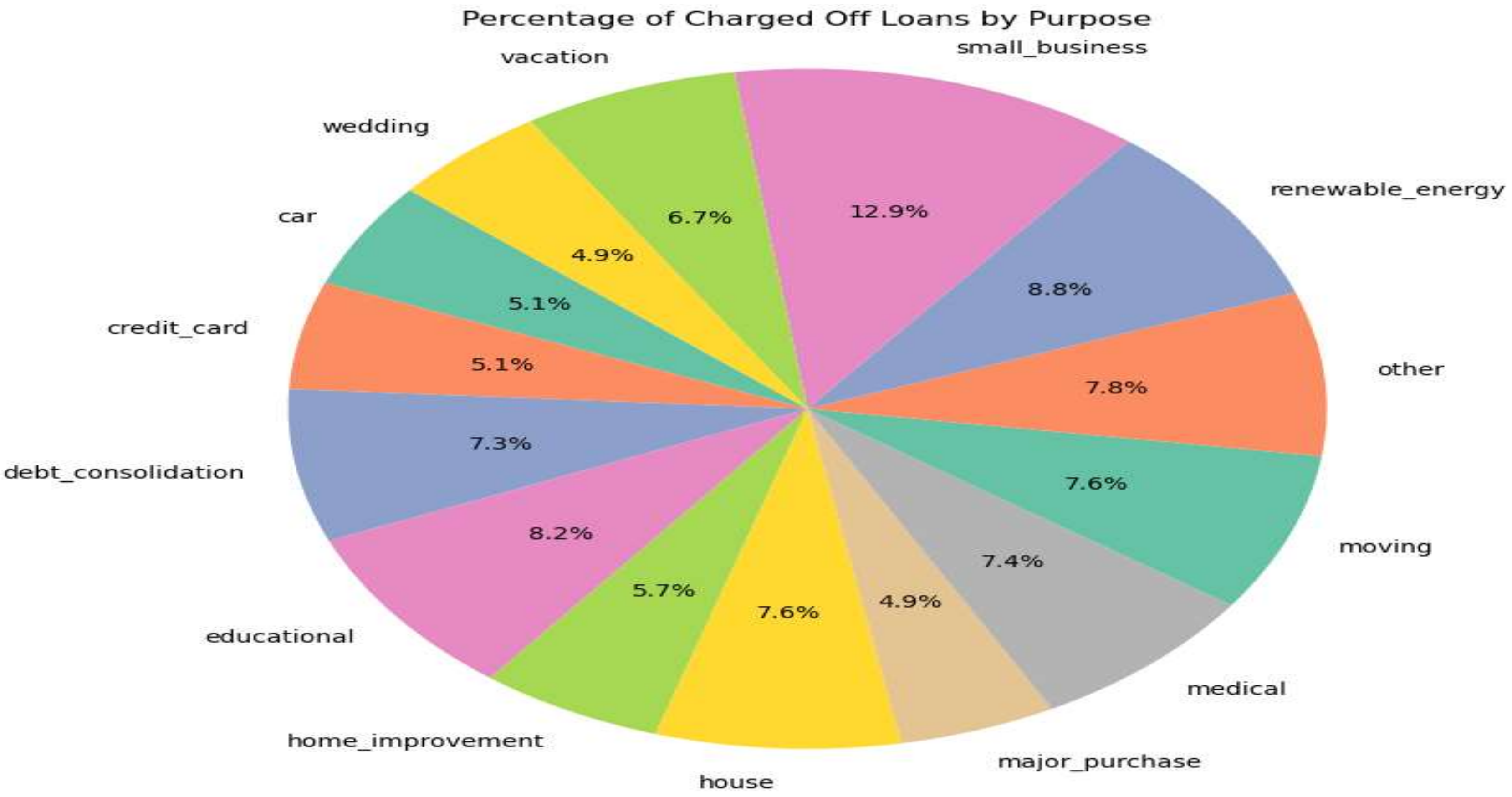
# Analysis on Charged off Loans Vs State

This plot show analysis between the count of charged loans by state i.e CA,FL & NY has more number of charged loans

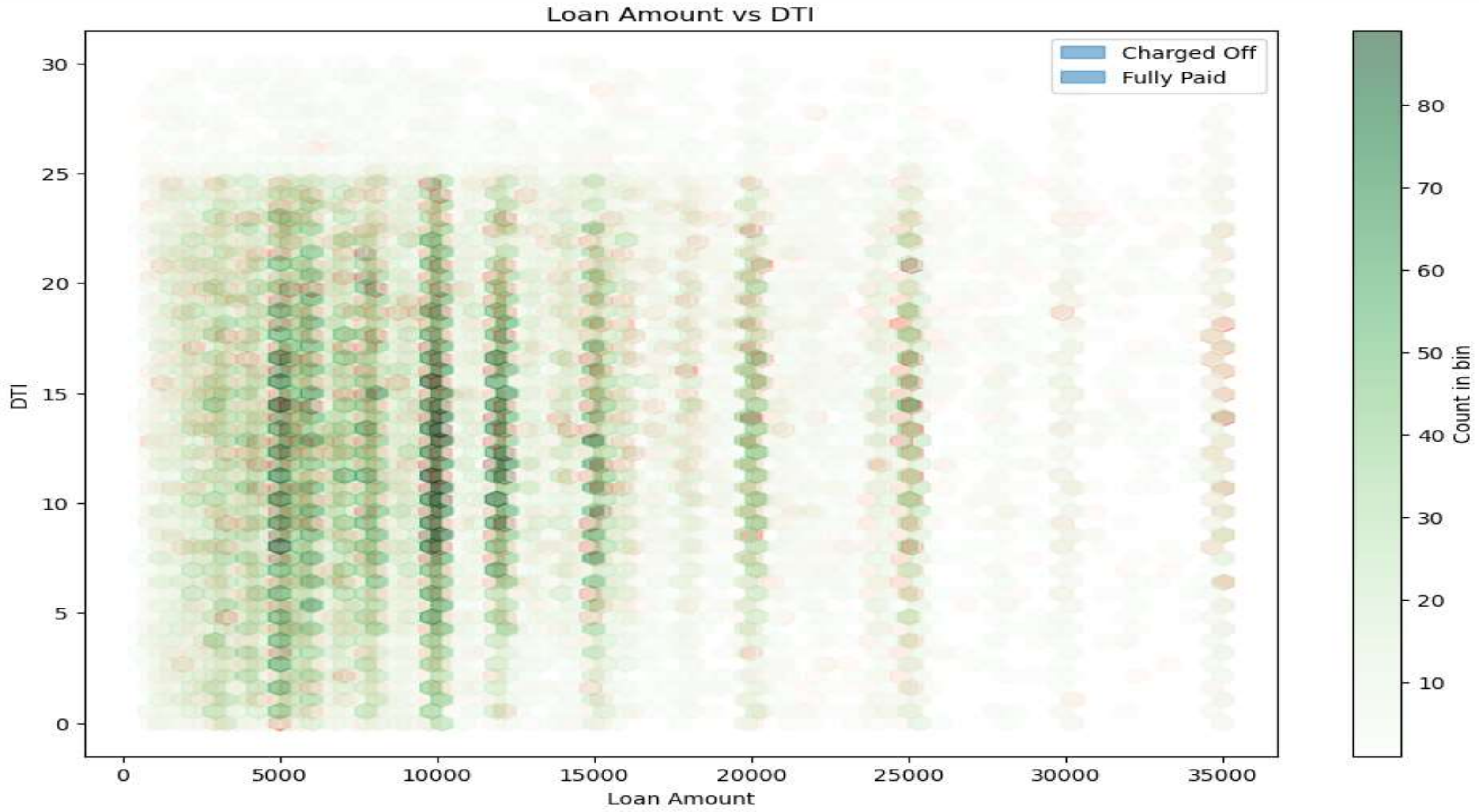


# Analysis on Charged off Loans Vs Purpose

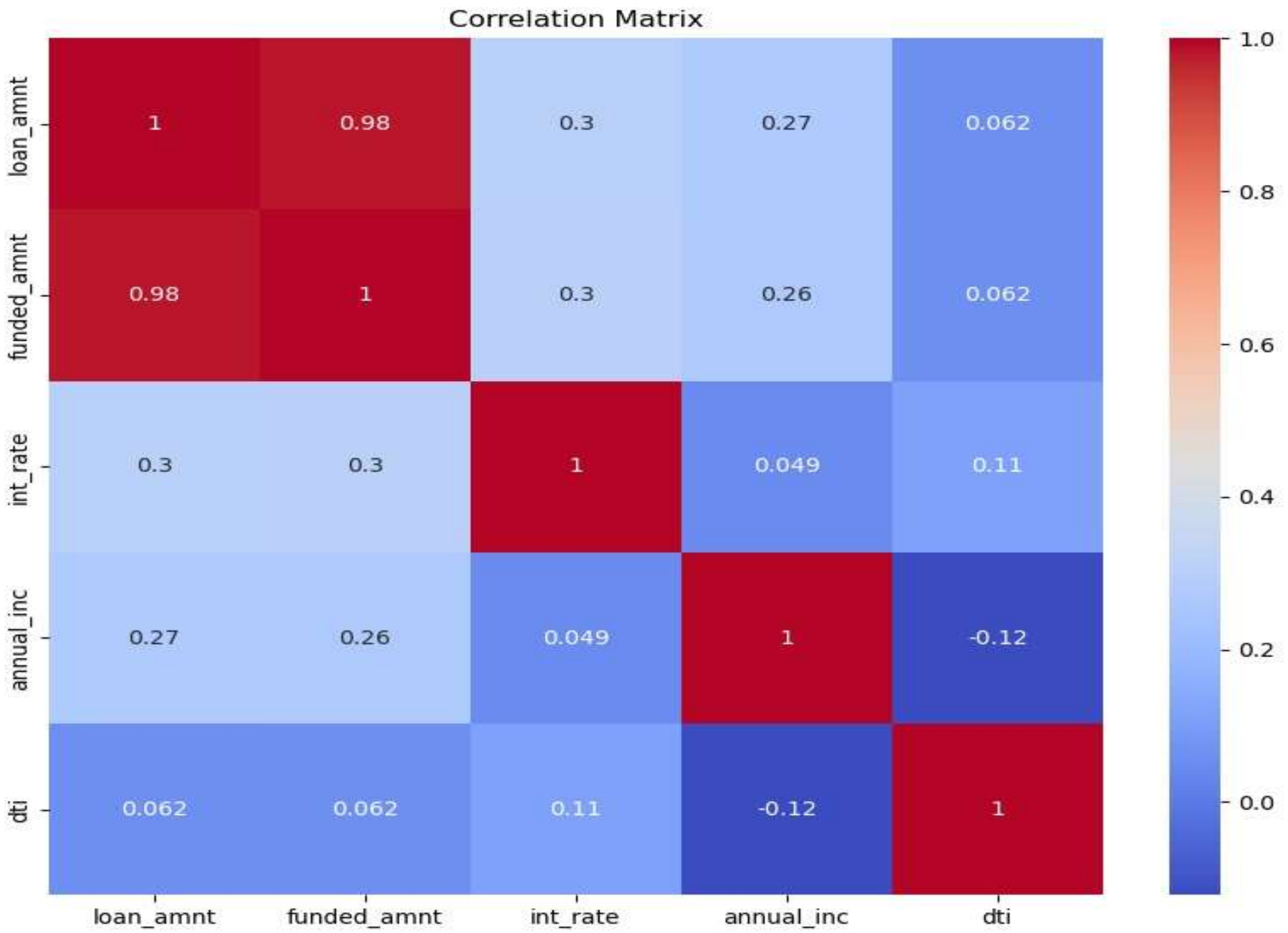
This plot show analysis between Charged off Loans Vs Purpose i.e small business people are more defaulters



# Analysis on Loan Amount Vs DTI



# Correlation analysis




**Inverse Relationships:**

There is a negative correlation between the loan amounts requested (loan_amnt) and the incidences of public record bankruptcies (pub_rec_bankruptcies).

Similarly, the funded amounts (funded_amnt) and annual income exhibit negative correlations with debt-to-income ratio (dti).

**Moderate Associations:**

The size of the loan (loan_amnt) shows moderate positive correlations with the loan duration (term).

The loan duration (term) also moderately correlates with the interest rate charged (int_rate).

**Strong Connections:**

Strong positive correlations exist between the loan amounts (loan amount) and the actual funded amounts (funded amount).

Additionally, the funded amount from investors (funded amnt_inv) demonstrates a robust correlation with the funded amount (funded amount).

# Conclusion

- The analysis provides insights into factors influencing loan defaults. Key observations include:
- Higher loan amounts are associated with higher default risk.
- Interest rates vary significantly across loan grades and verification statuses.
- Certain loan purposes and borrower characteristics correlate with higher default rates.
- Maximum people who takes Loan is 10+ years experienced

# Recommendations

- Based on the findings, recommendations for mitigating default risk include:
- Tighter scrutiny for higher loan amounts.
- Adjusting interest rates based on risk profiles identified.
- Monitoring loans issued during certain months or for specific purposes more closely.
- Tighter scrunity for people with 10+ years of experience