

# LENDING CLUB

## CASE STUDY

# Lending Club Case Study

- Lending Club is US based peer-to-peer consumer finance company founded in 2007.
- Lending Club specialises in lending various types of loans to urban customers.
- LendingClub matches borrowers with investors willing to fund their loans.

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision

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

# Business Objectives

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 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.

To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

# Problem Statement

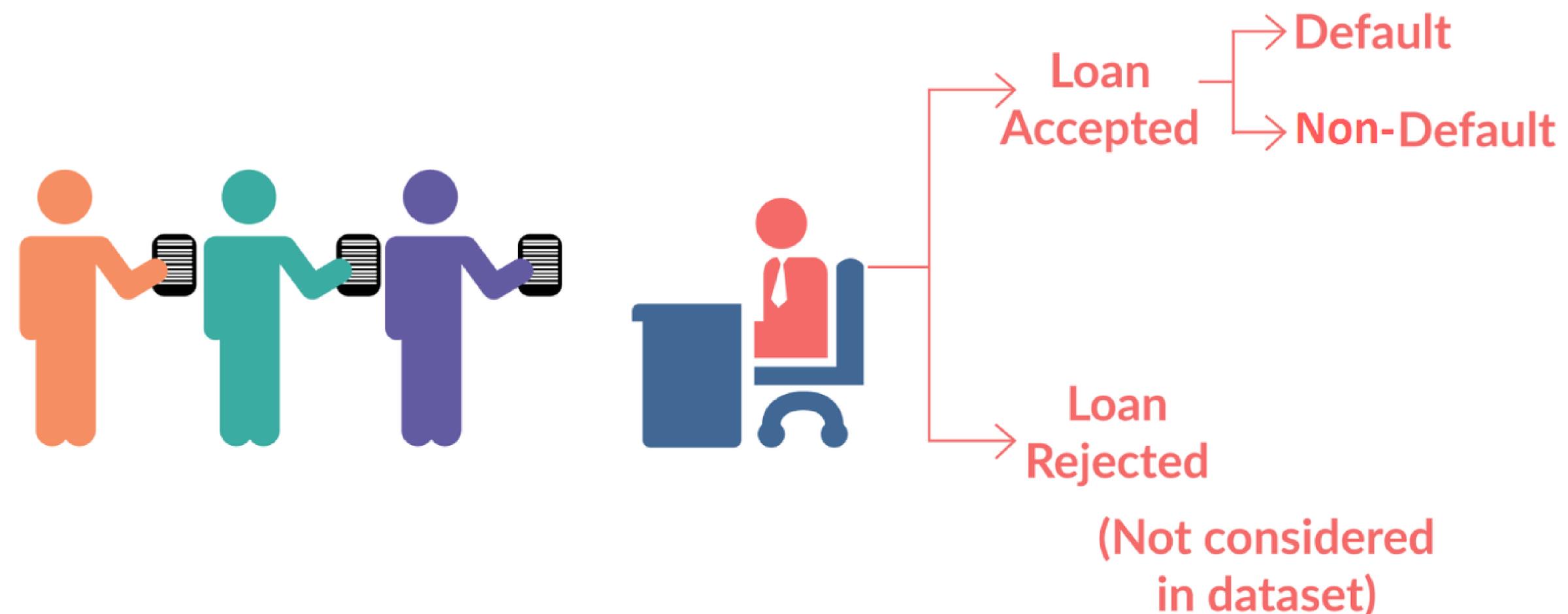
- **Business Understanding**

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)

# LOAN DATASET



# Data Cleaning

What is Data Cleaning ?

Data cleaning is an important step in any data analysis project.

In this case study, the data cleaning process involved the following steps.

Once the data has been cleaned, it is ready for analysis. The analysis of the Lending Club dataset can help to identify patterns that indicate if a person is likely to default on a loan. This information can be used to make decisions about loan approval, loan amount, and interest rate.

Identifying and removing missing values

Exploring and cleaning categorical data

Formatting numerical data

# Data Cleaning:

We identified the rows and columns that have missing values from the given data set.

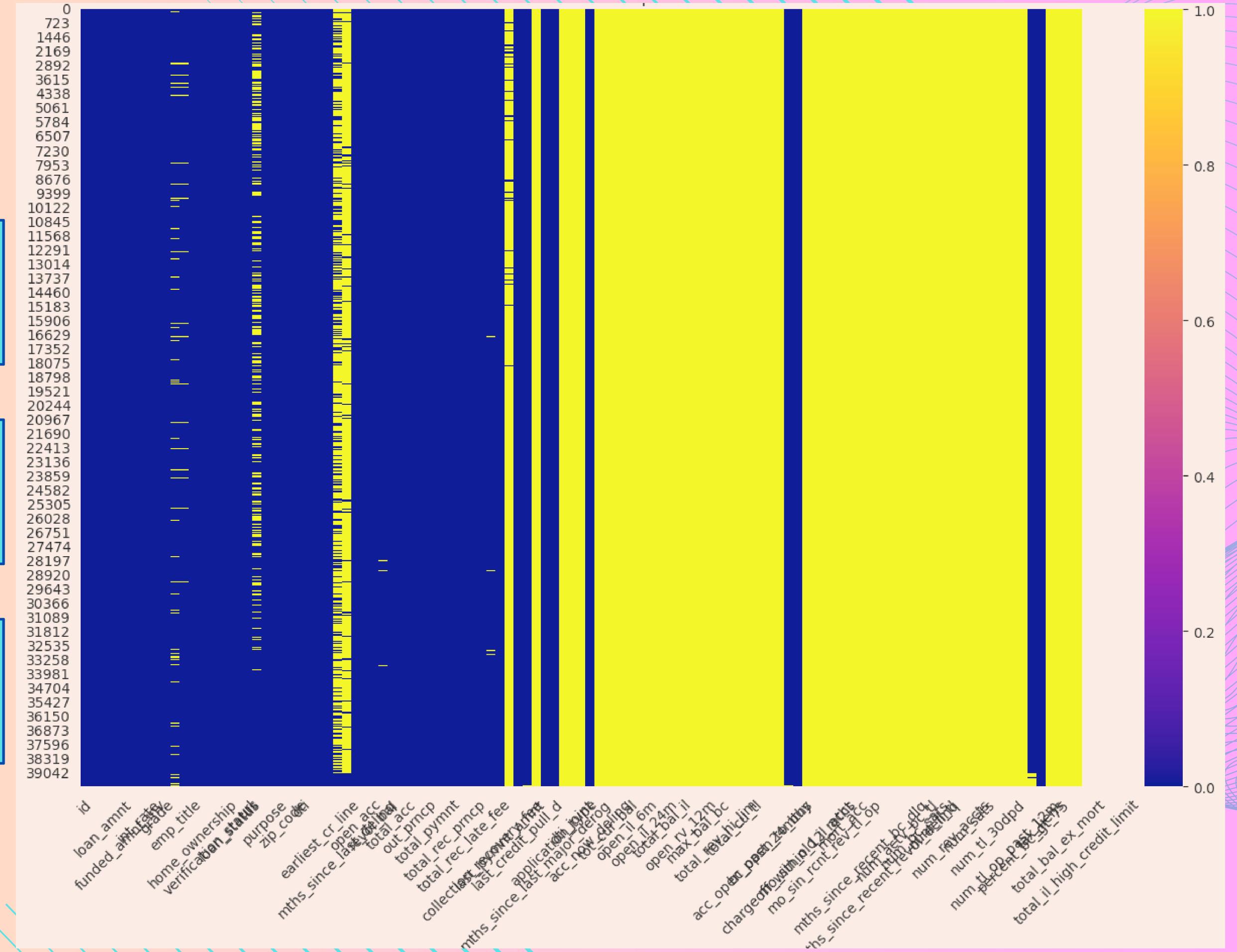
We removed many unnecessary columns and reduced the number of columns from 111 to 30.

For example - few object and we converted it to relevant int and float types

Then we removed those columns and rows.

We transformed and converted few columns to proper data type.

Few derived columns are also created



# Exploratory Data Analysis

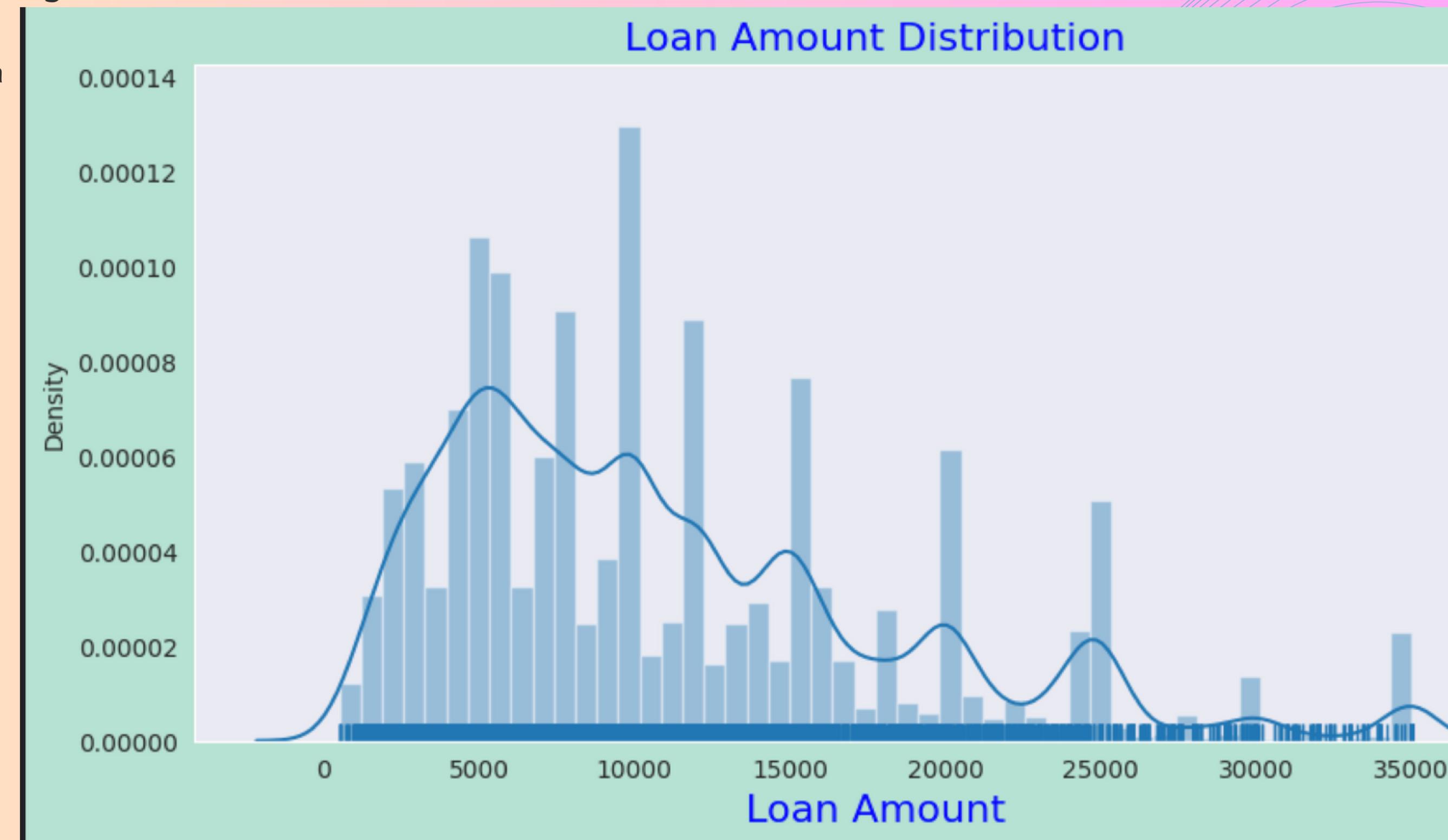
---

**Exploratory Data Analysis (EDA) is an essential step for analyzing data.**

**EDA can help us gain insights into the loan data, identify patterns, discover relationships, and make informed decisions**

# Univariate Analysis

- Univariate analysis is a statistical method that analyzes one variable at a time.
- It is the simplest form of data analysis and is often used as a first step in understanding a dataset.
- Univariate analysis can be used to describe the distribution of a variable, identify outliers, and test hypotheses about the variable

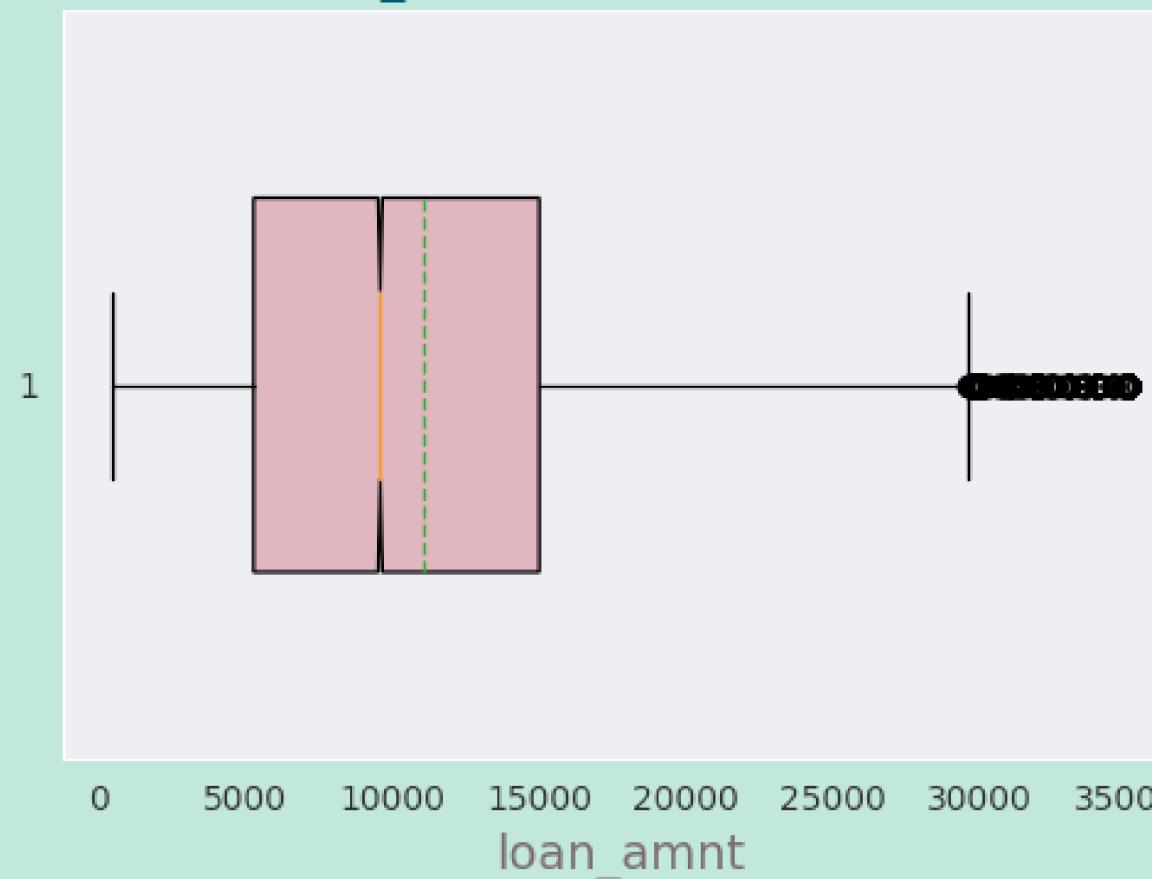


Following is the univariate analysis on continuous variables: **loan\_amount**, **funded\_amount**, **funded\_amnt\_inv**, **int\_rate**, **installment**, **annual\_inc** using Box Plot as shown below

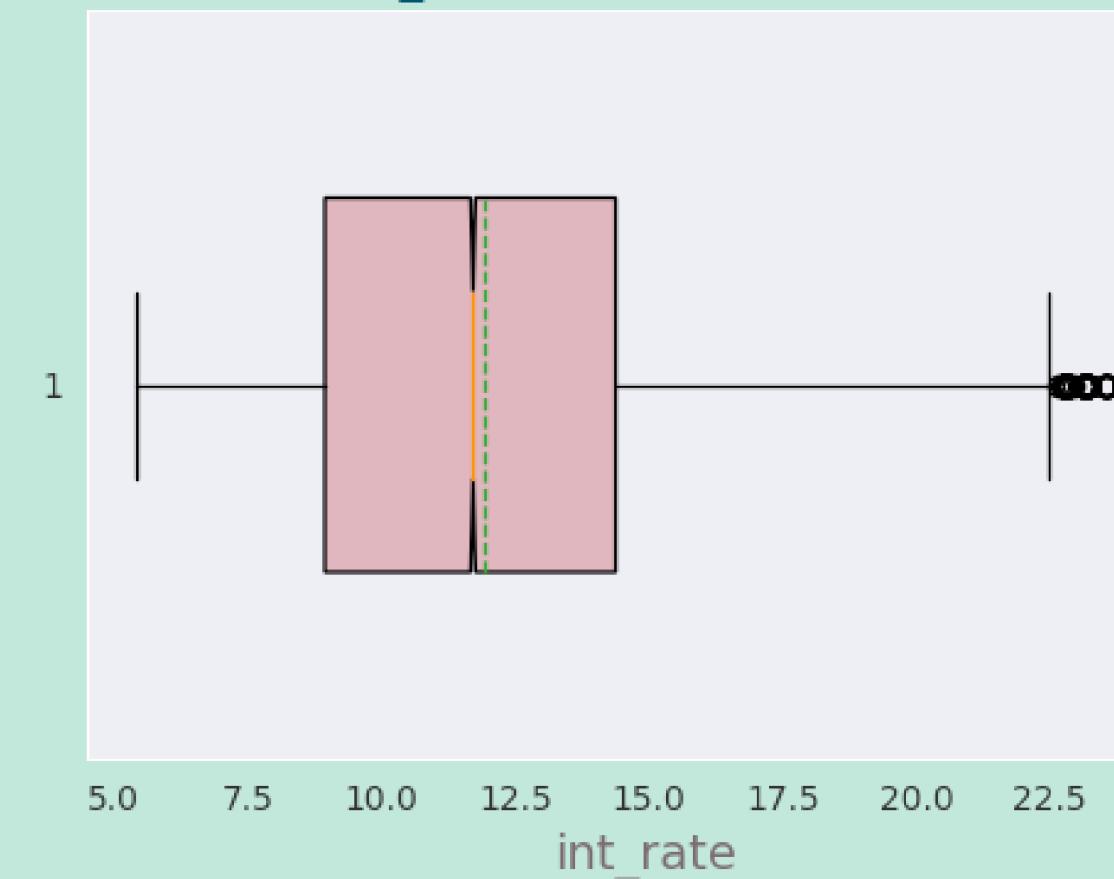
annual\_inc - Distribution Plot



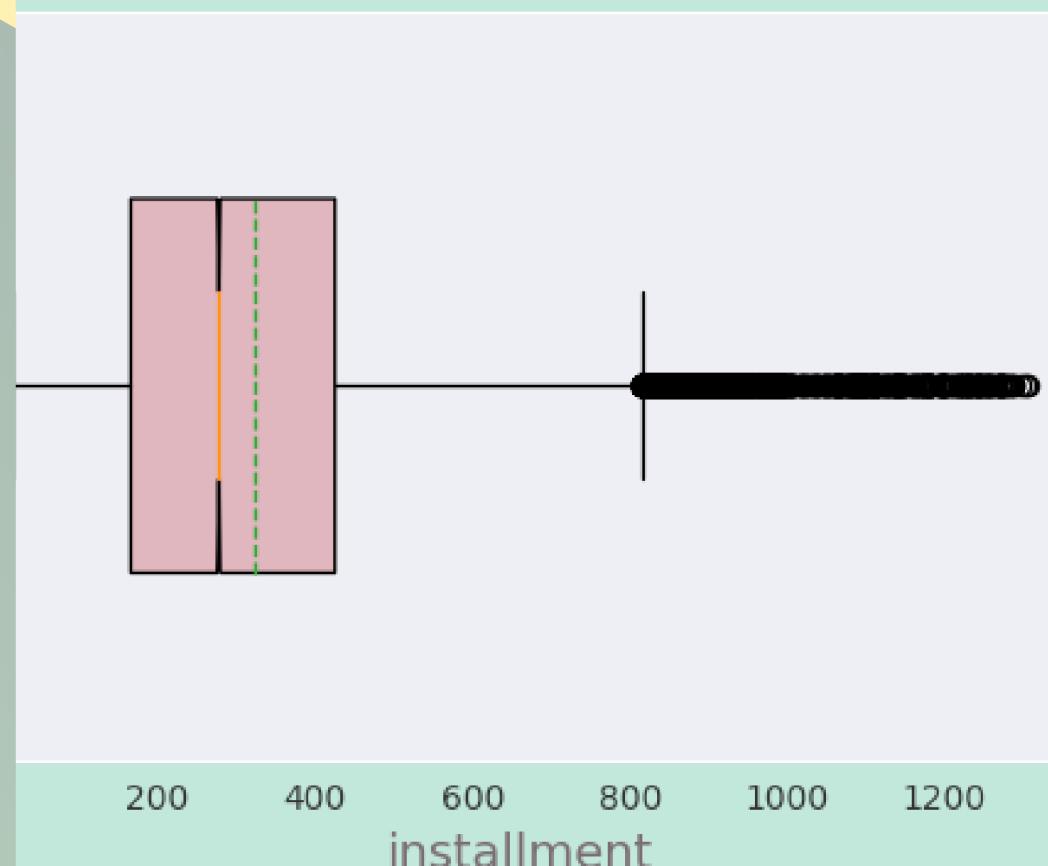
loan\_amnt - Distribution Plot



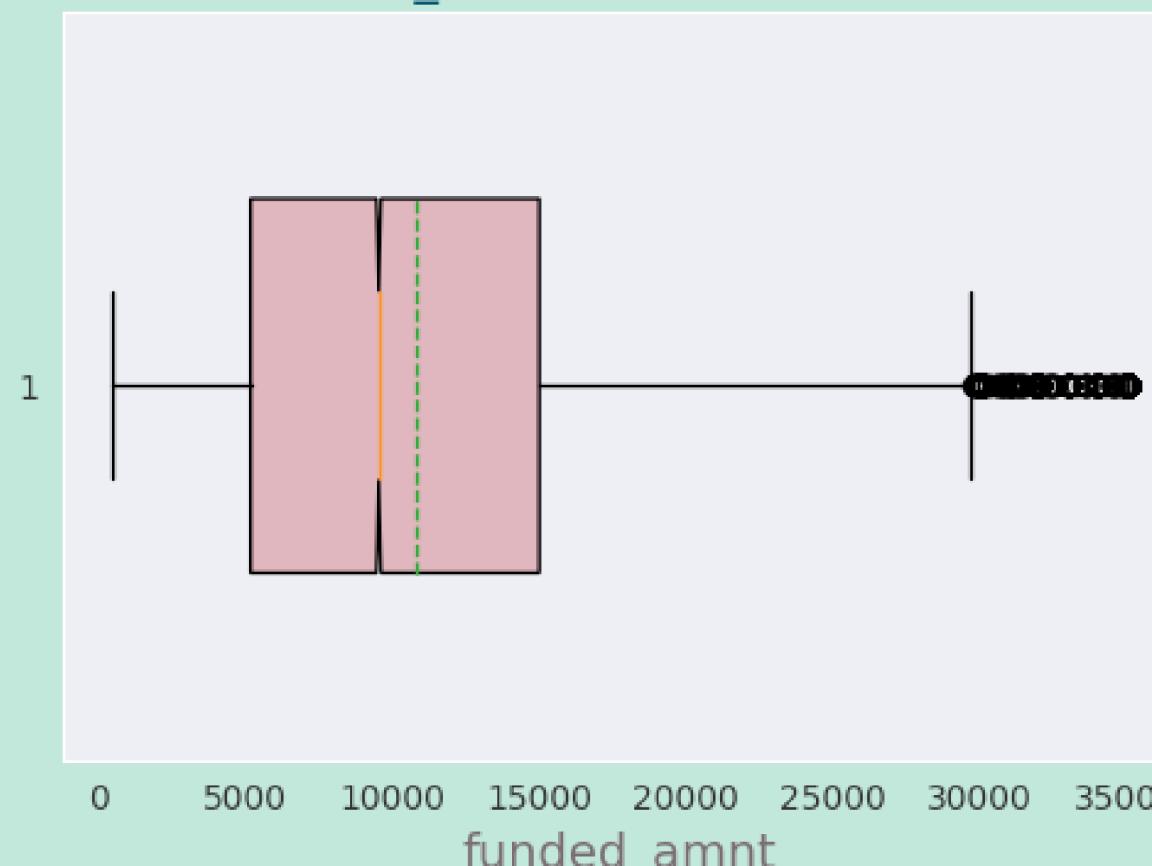
int\_rate - Distribution Plot



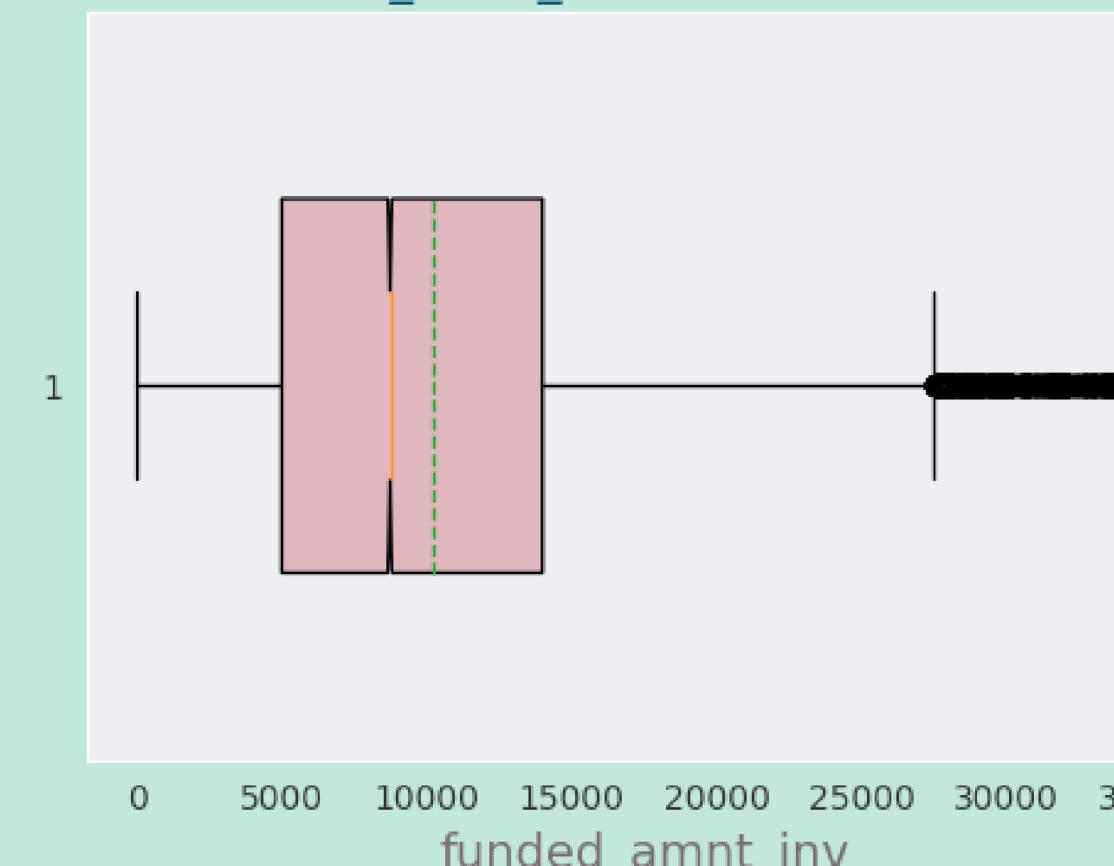
installment - Distribution Plot



funded\_amnt - Distribution Plot



funded\_amnt\_inv - Distribution Plot



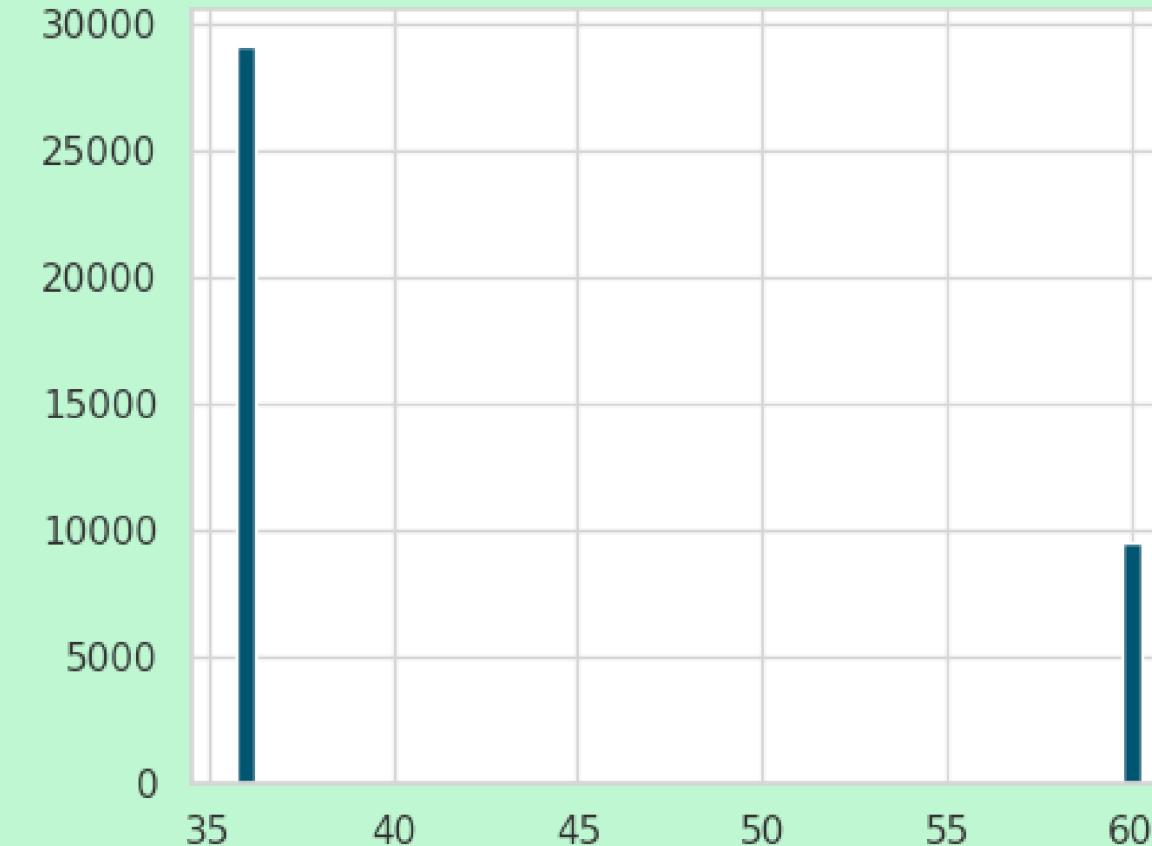
Following is the univariate analysis on categorical variables: emp\_length, grade, sub\_grade, term, loan\_status

## Histogram Representation

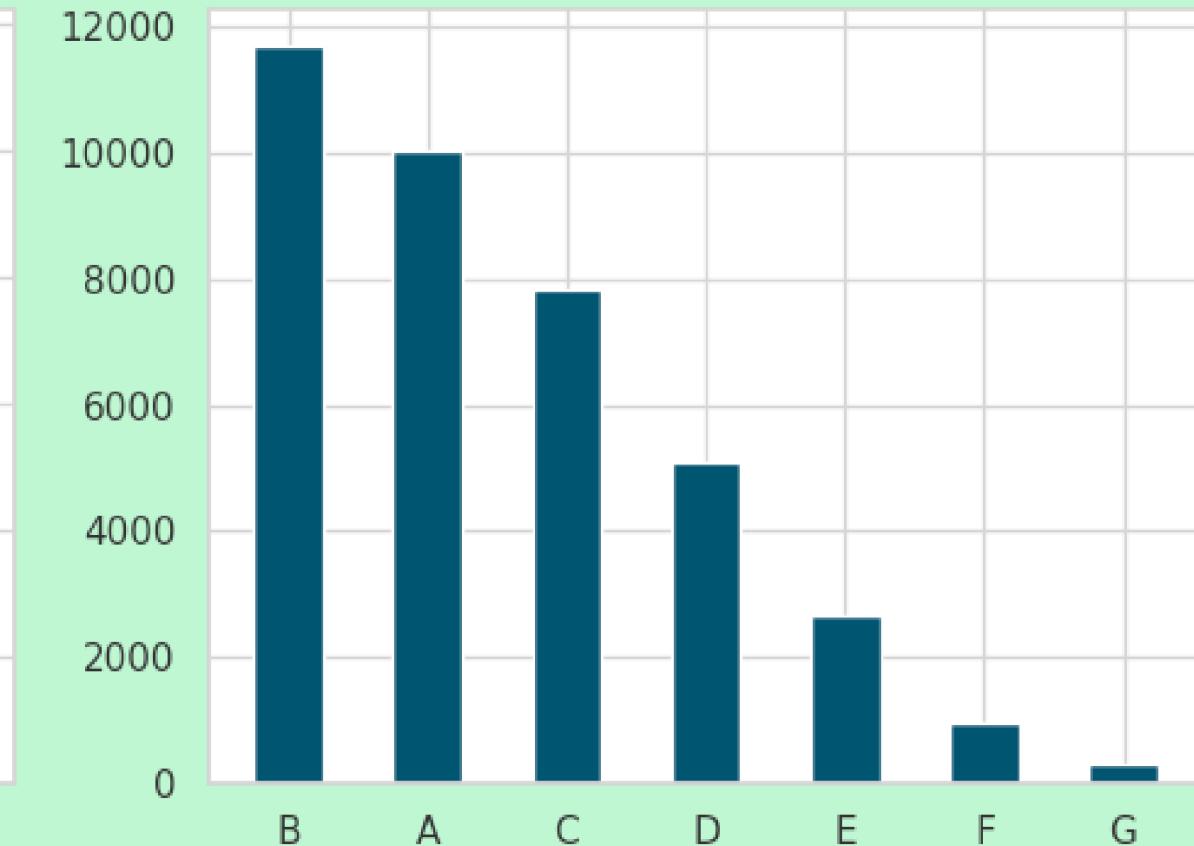
loan\_status - Distribution Plot



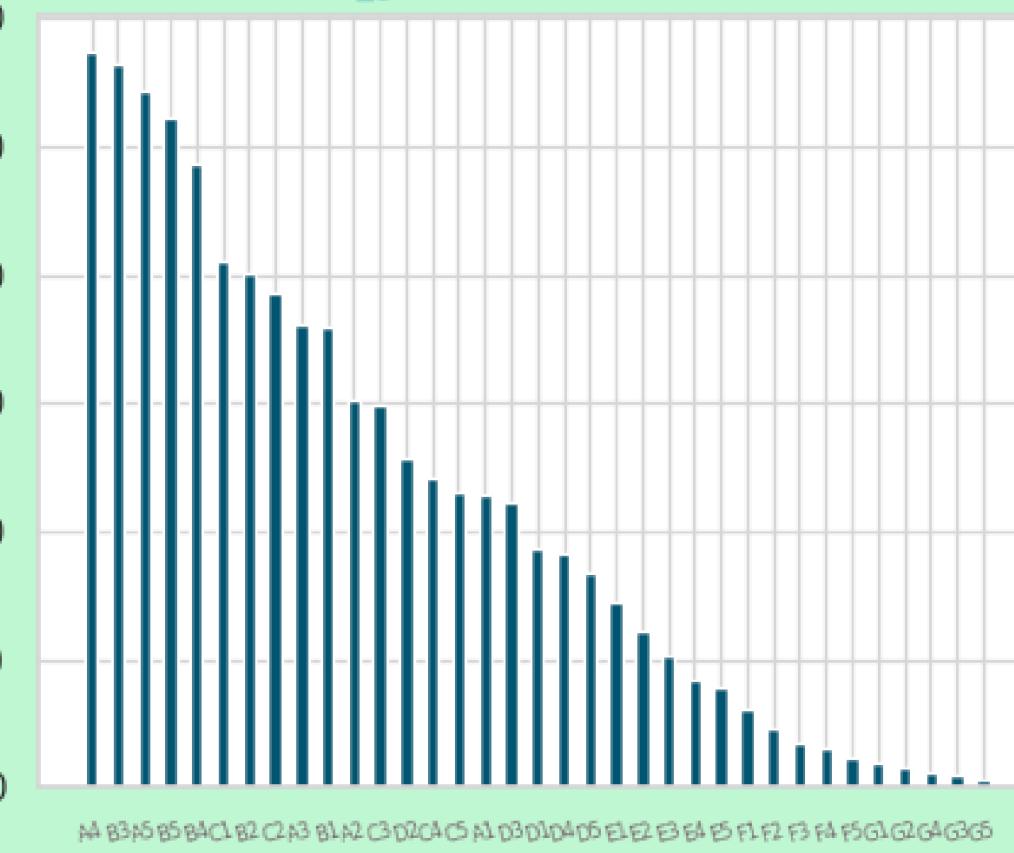
term - Distribution Plot



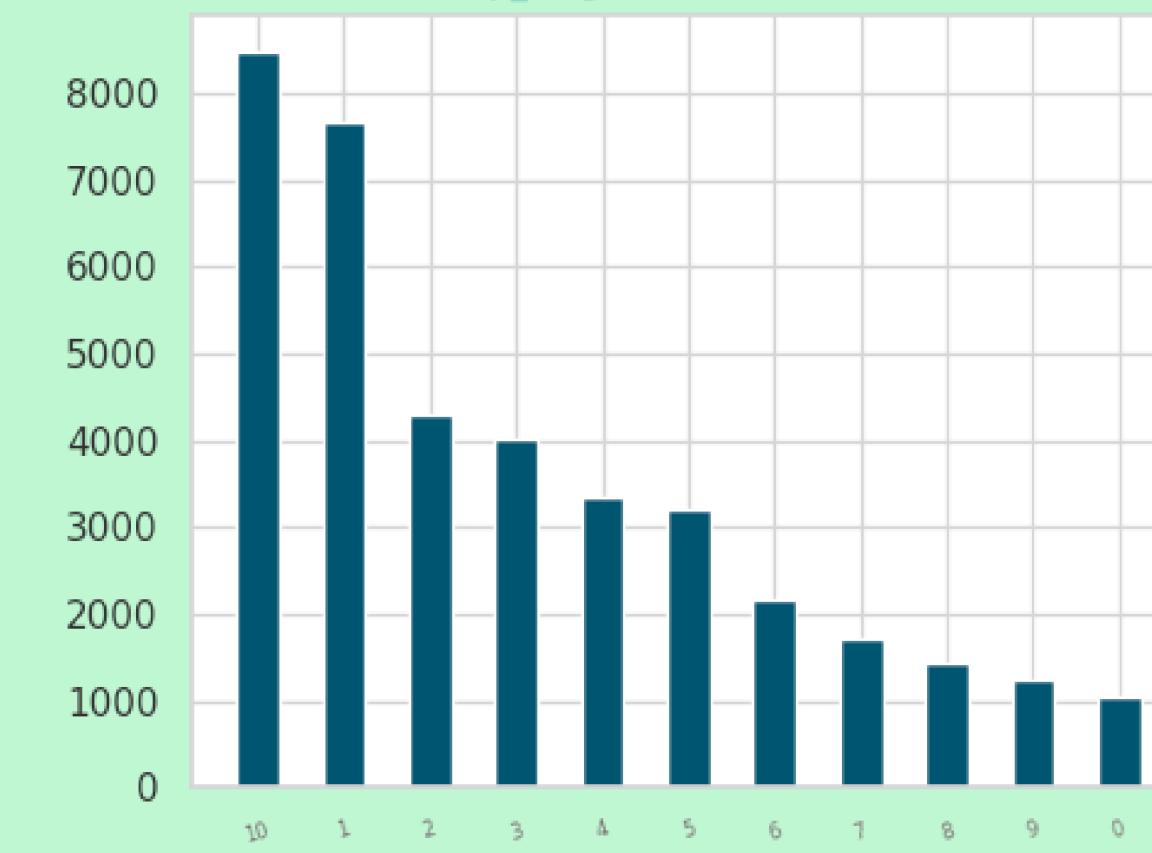
grade - Distribution Plot



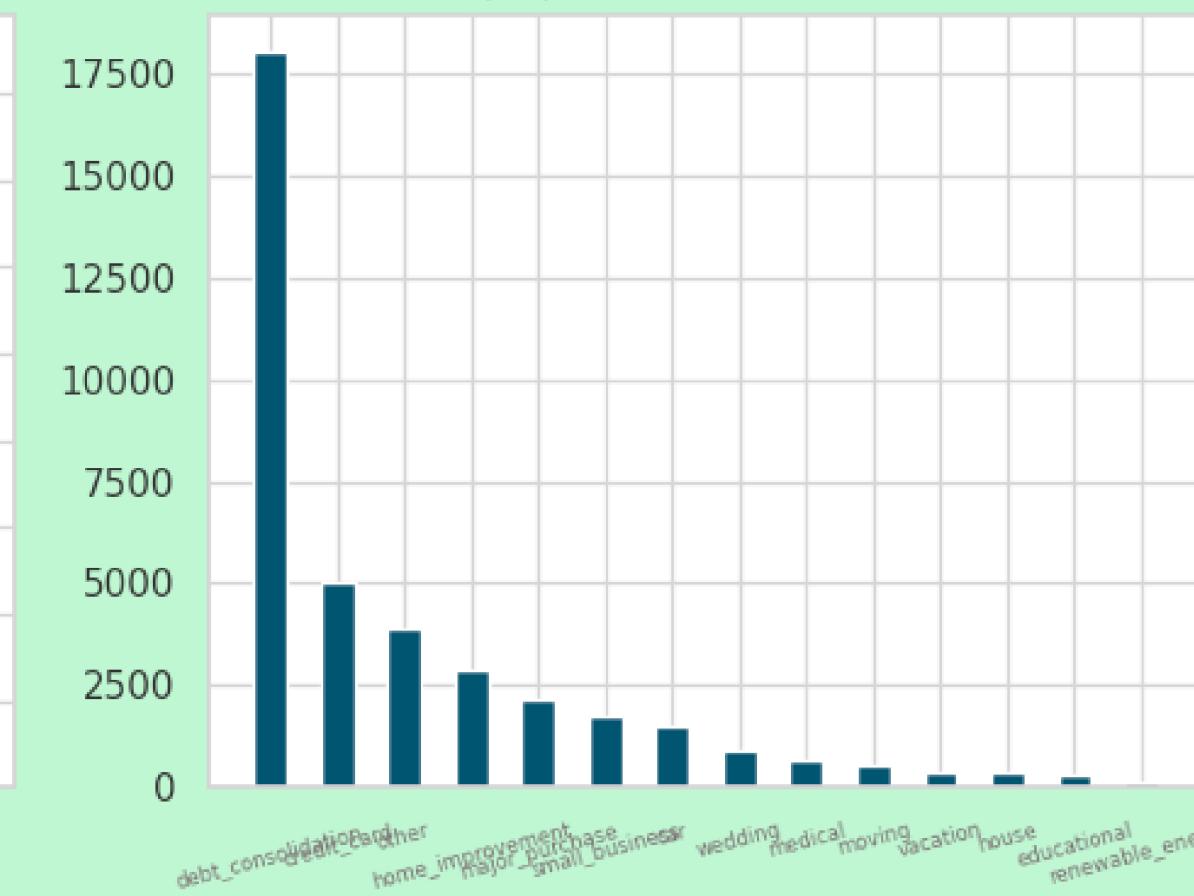
sub\_grade - Distribution Plot



emp\_length - Distribution Plot

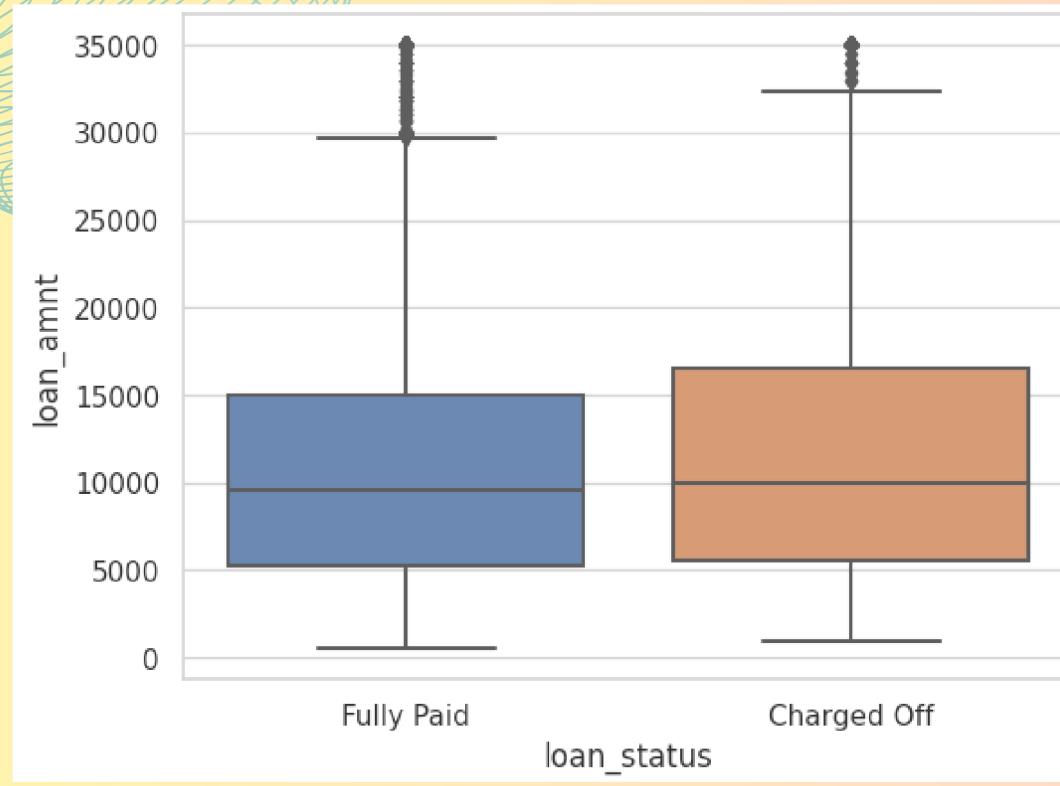


purpose - Distribution Plot

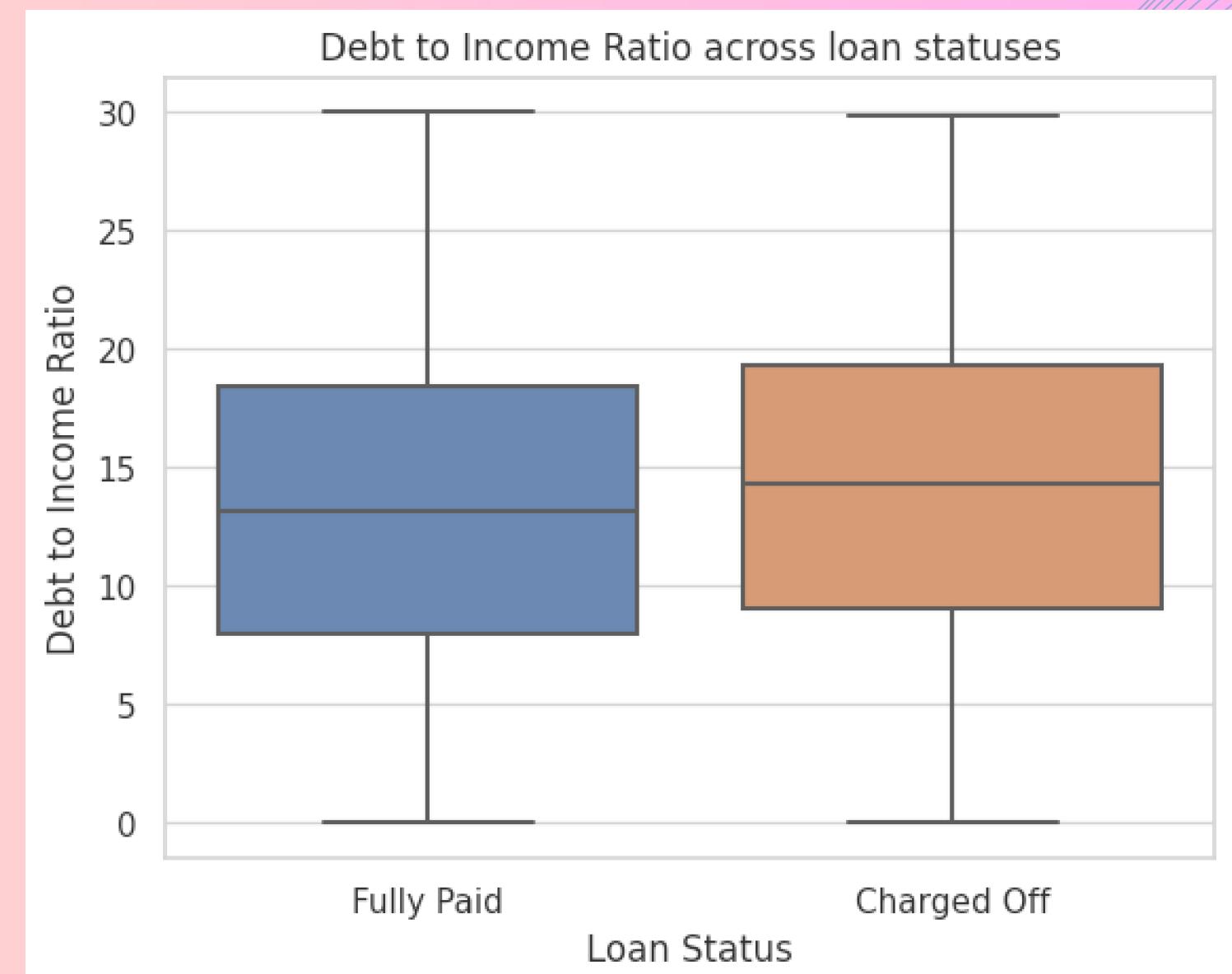


**Segmented Univariate Analysis:** The process typically involves dividing the dataset into distinct segments or subsets based on a specific criterion or grouping variable. Then, an analysis is performed on each segment individually, treating them as separate univariate analyse

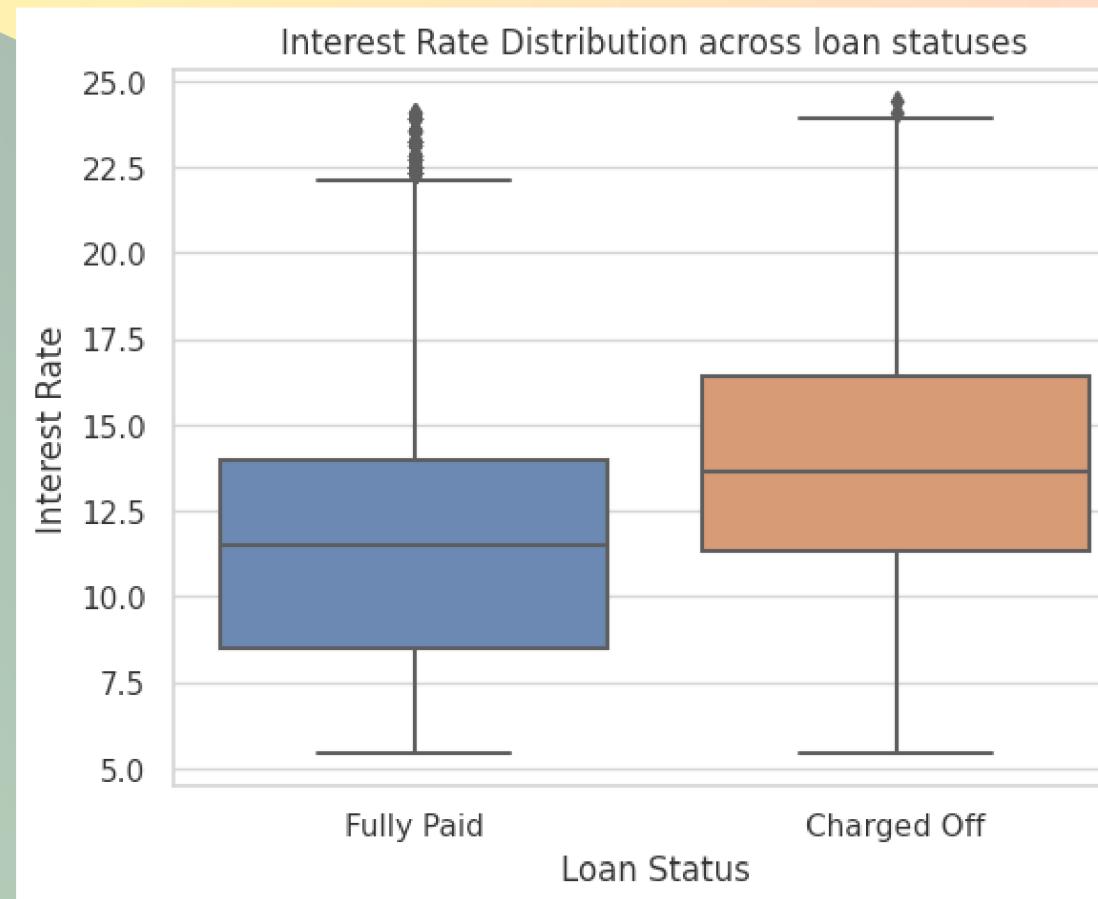
**Box plot representation of loan\_amount vs loan\_status**



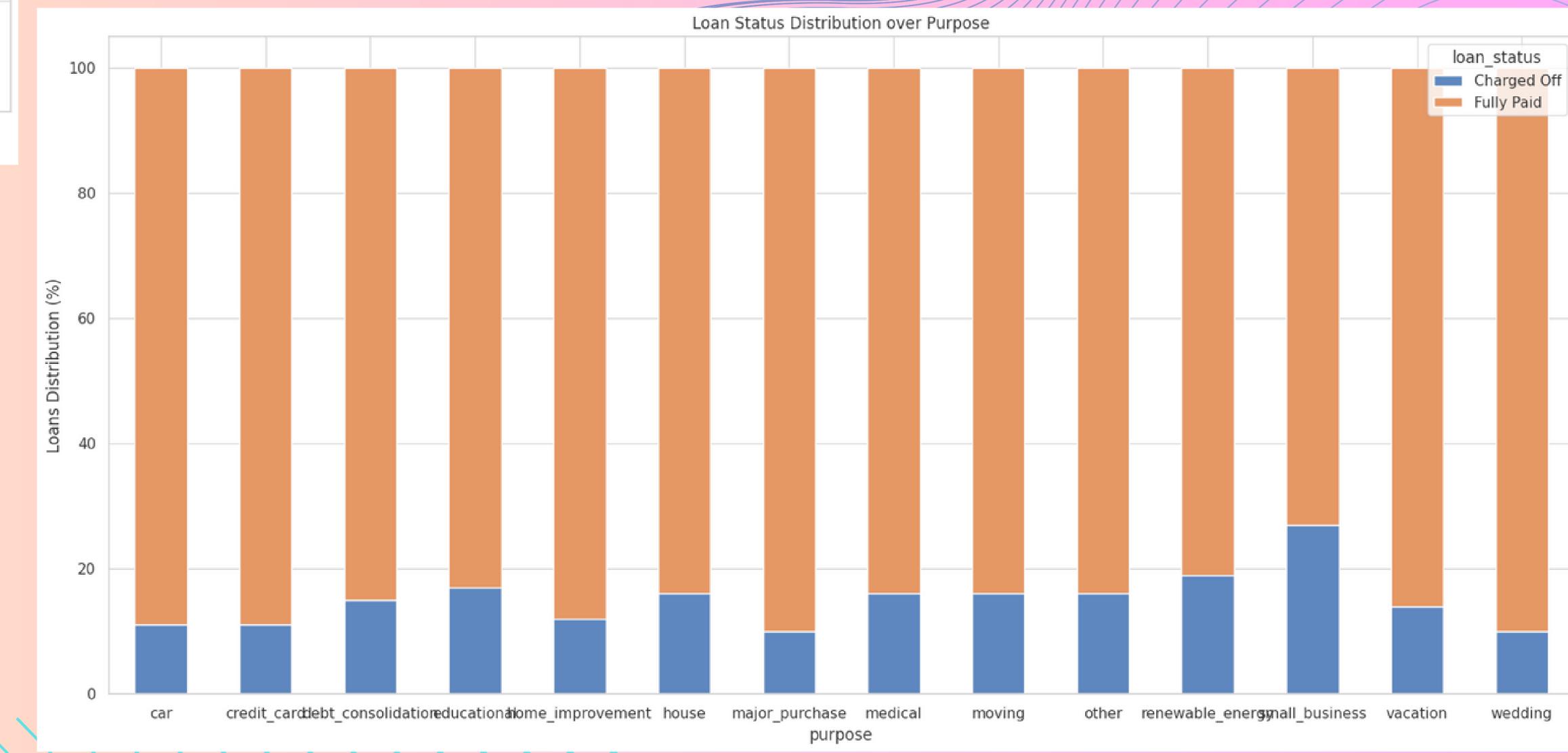
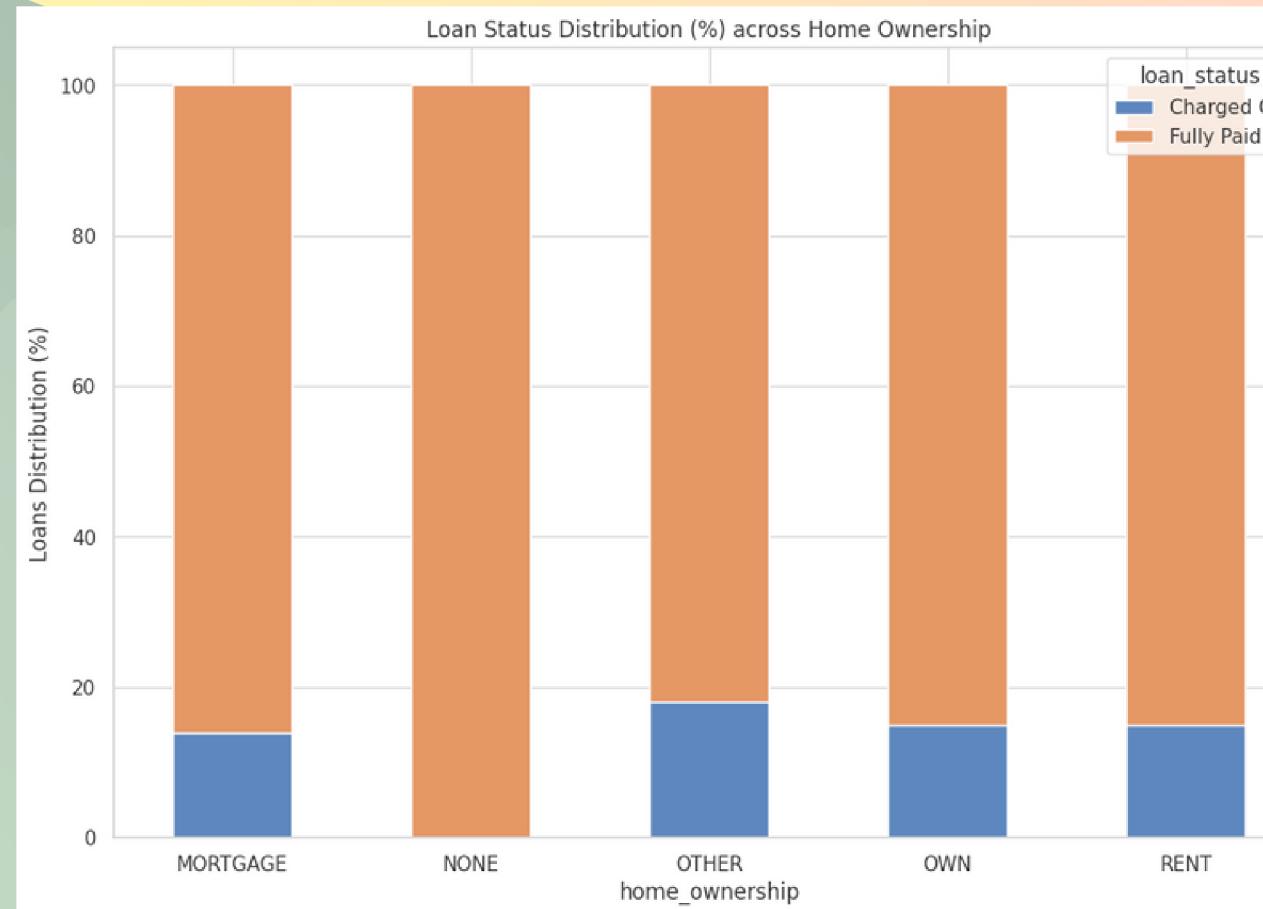
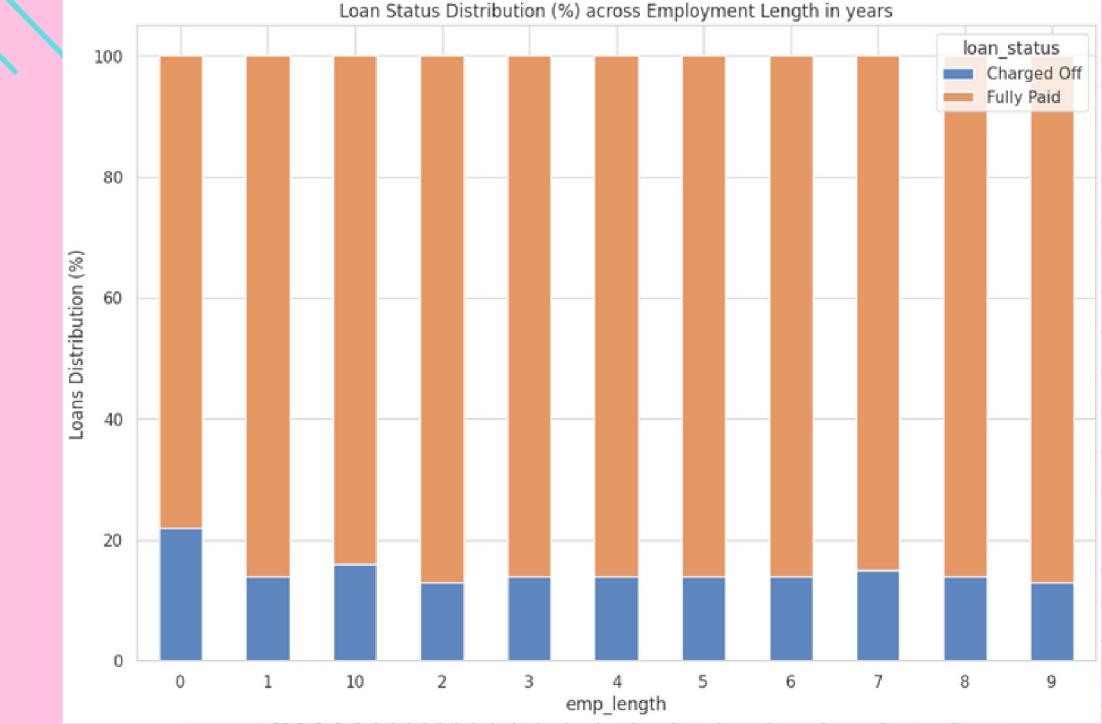
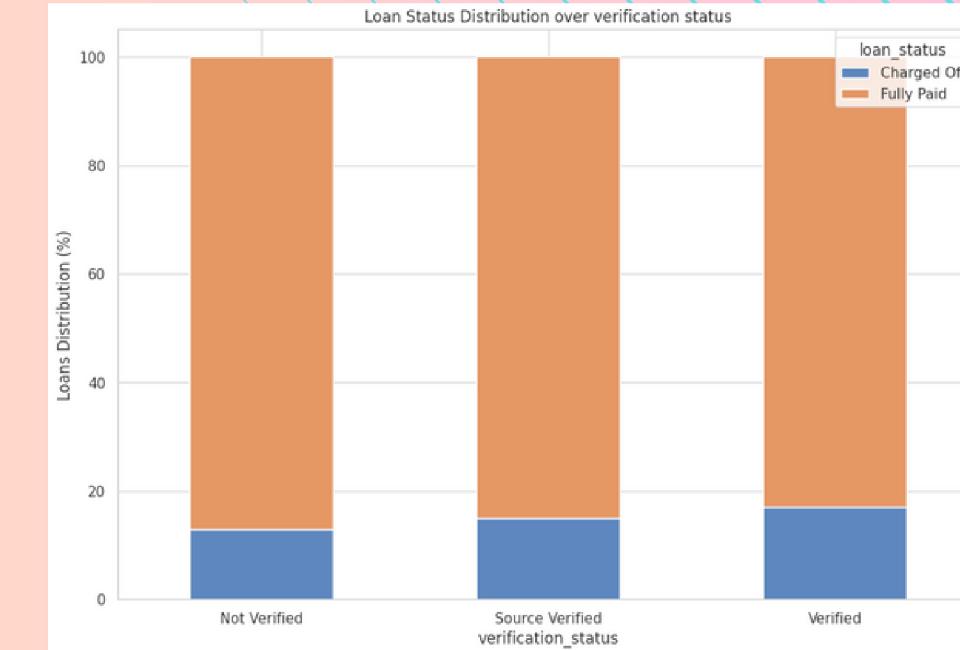
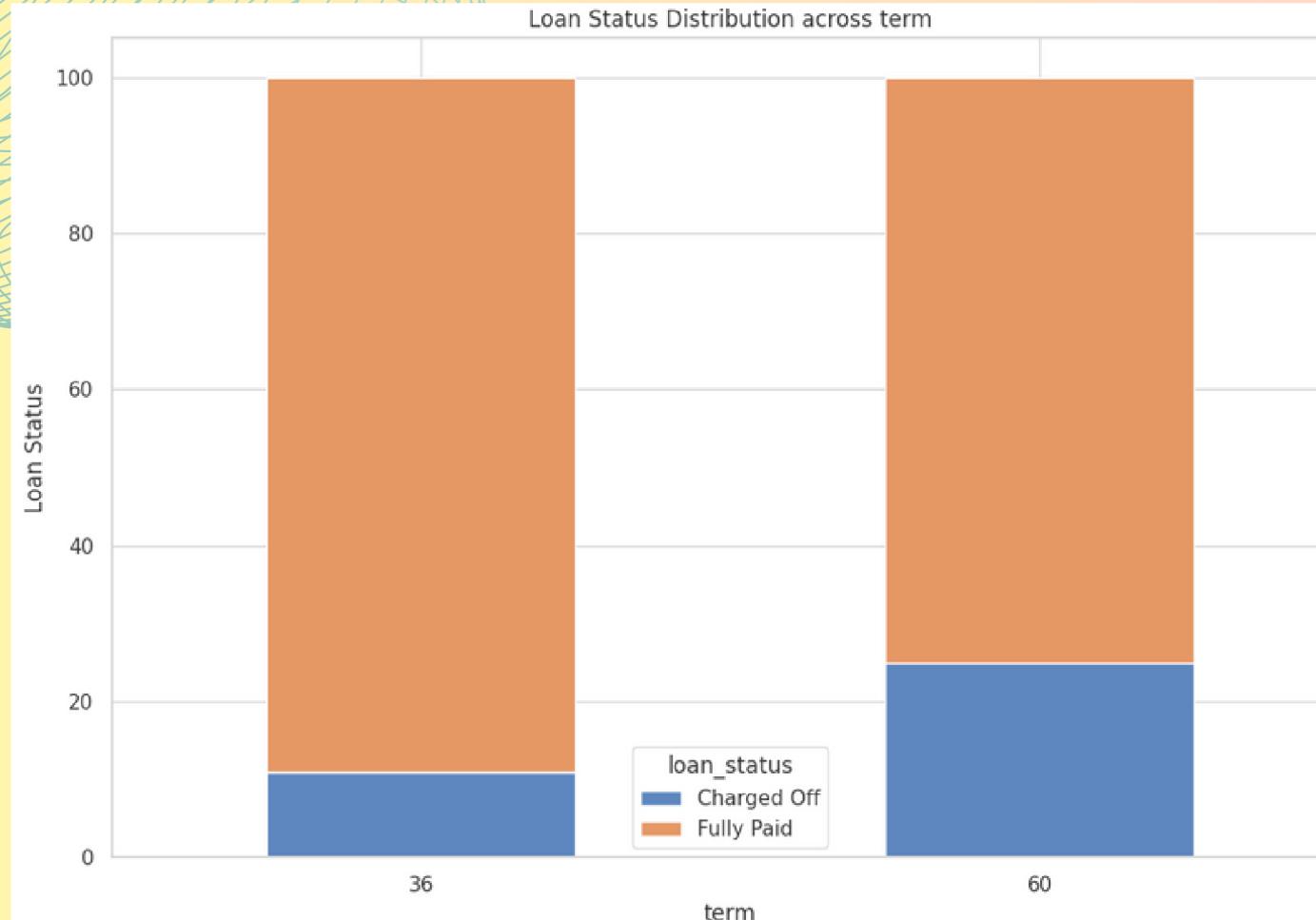
**Box plot representation of debt to income vs loan\_status**



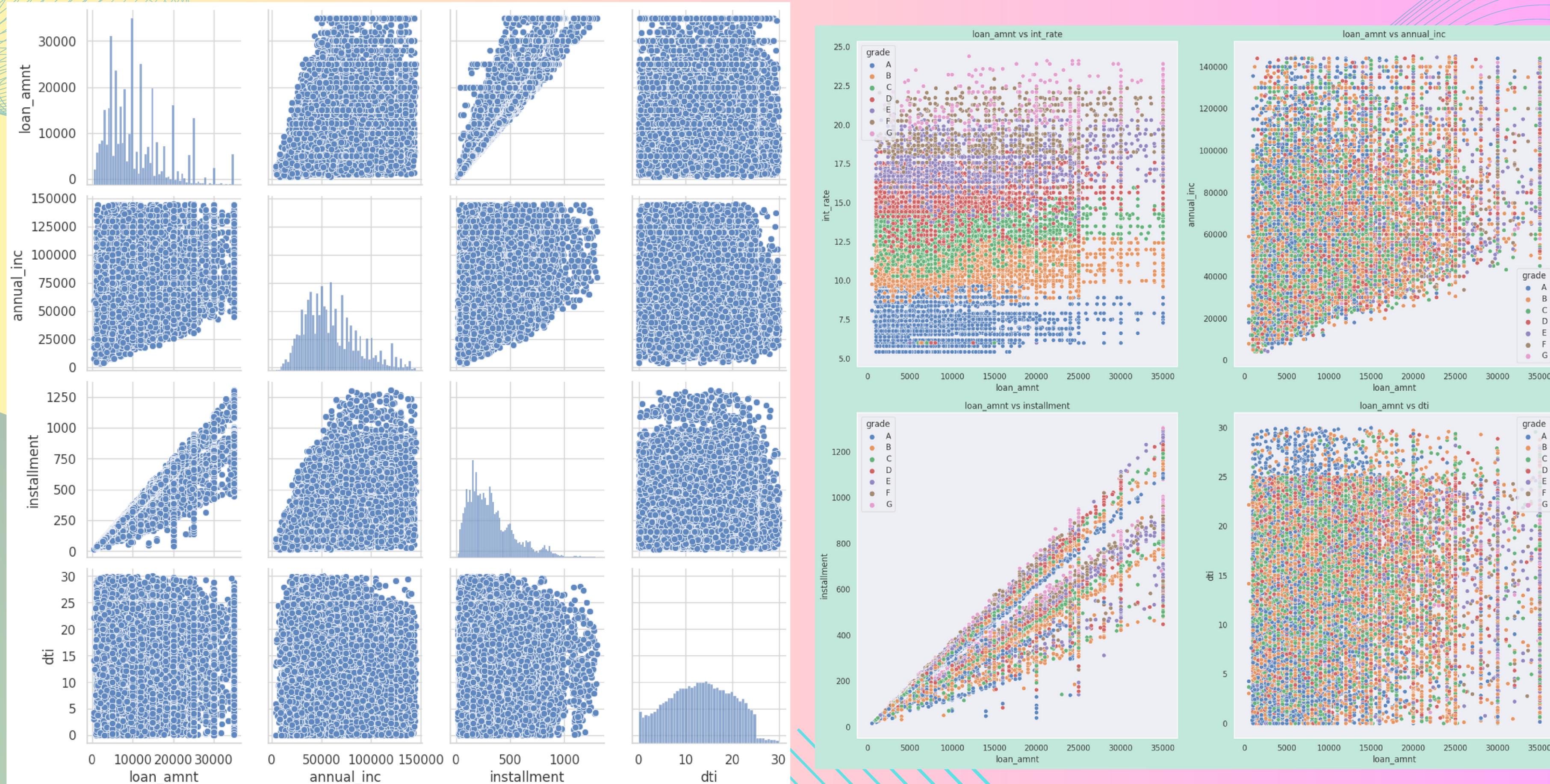
**Box plot representation of interest\_rate vs loan\_status**



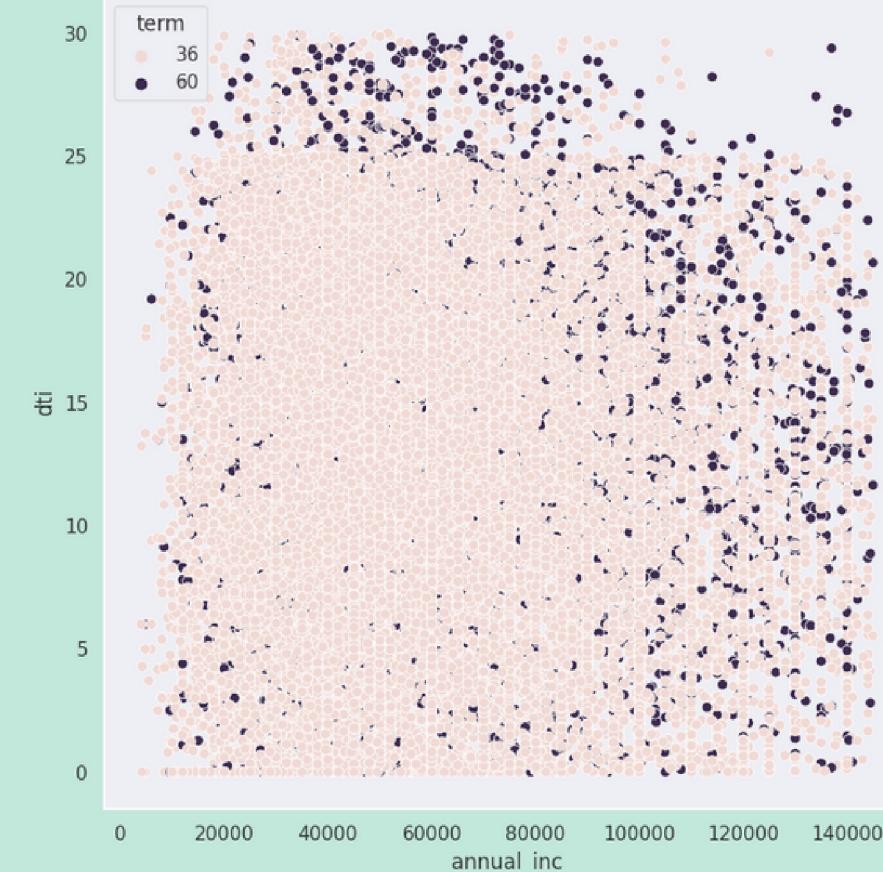
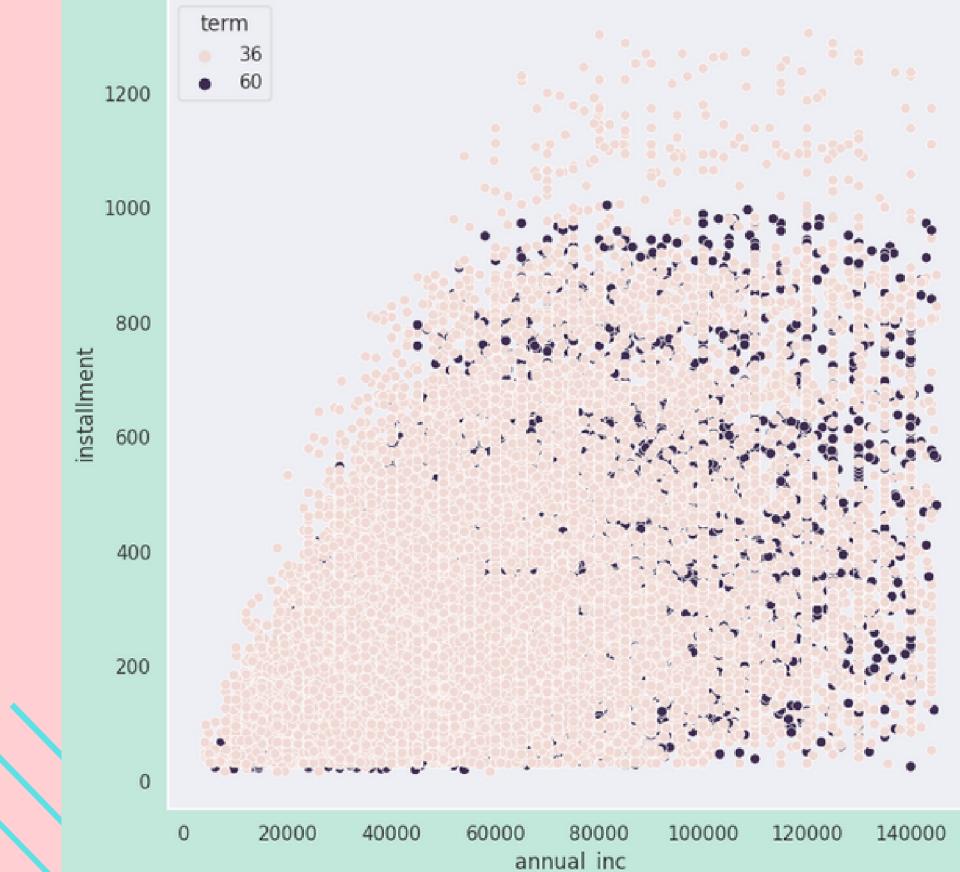
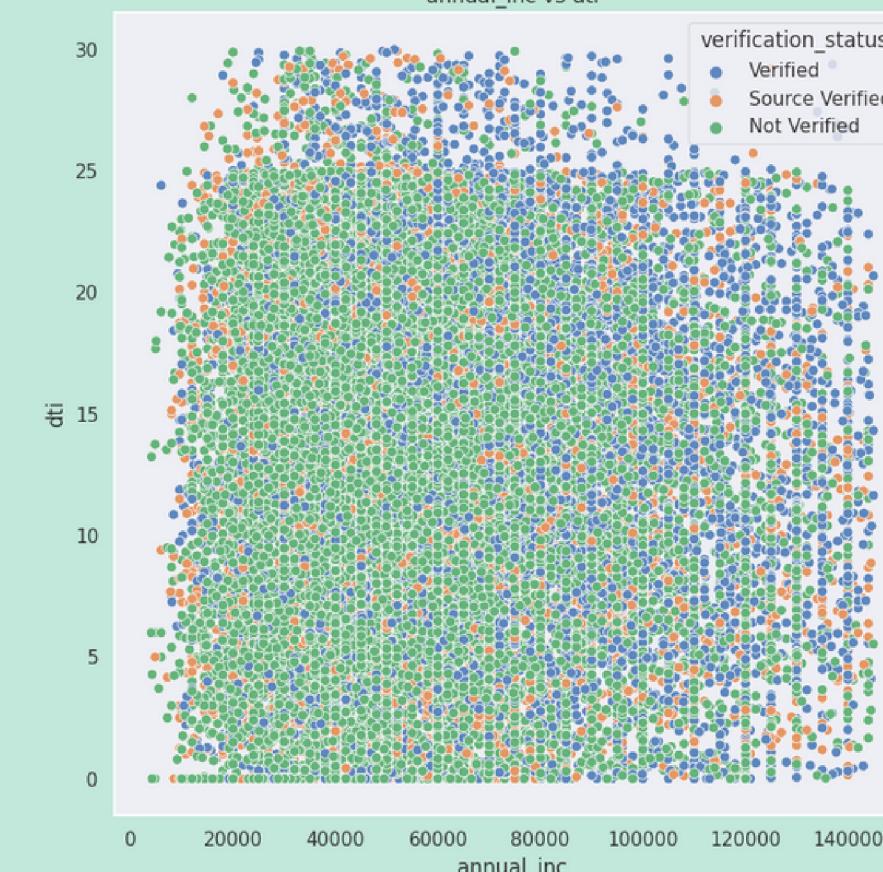
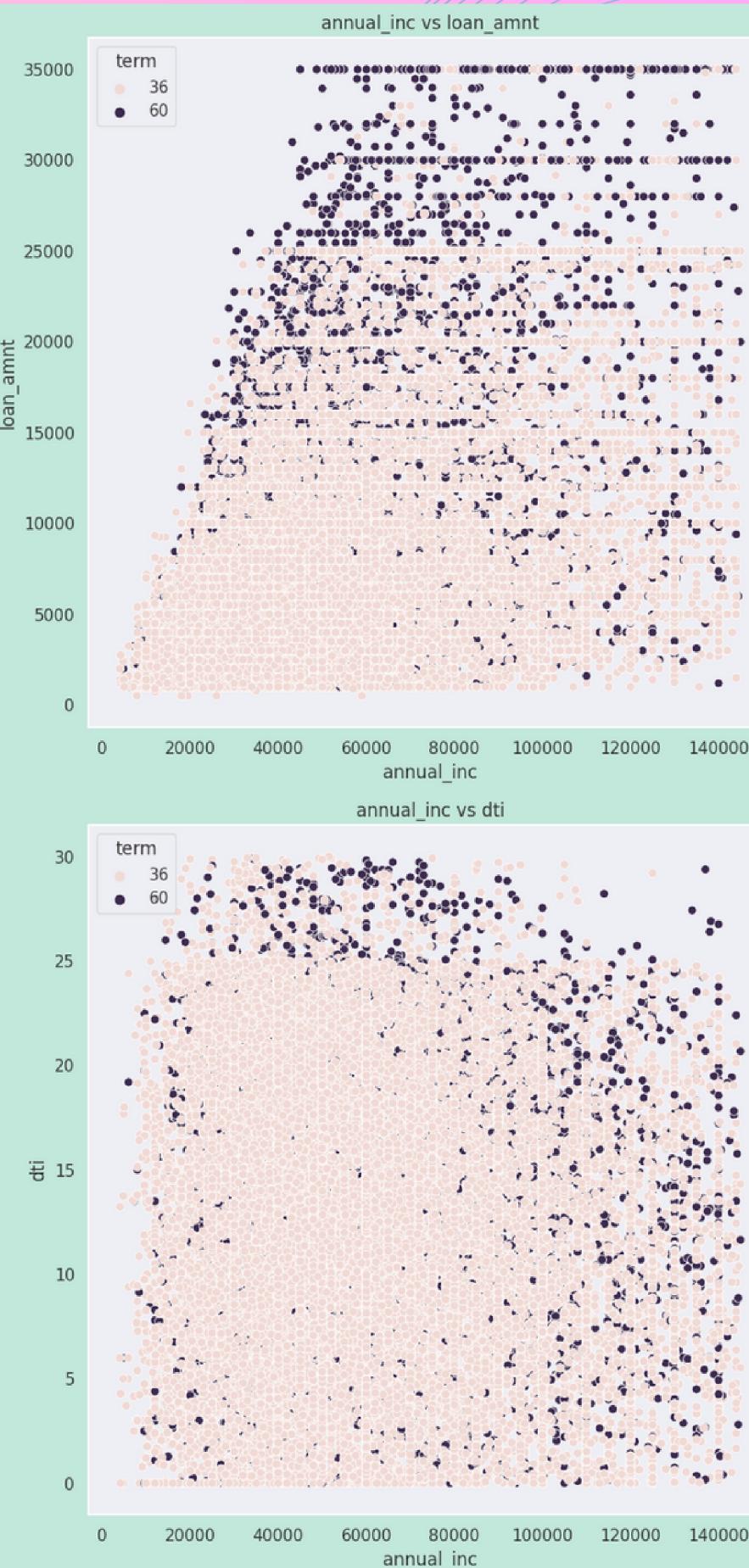
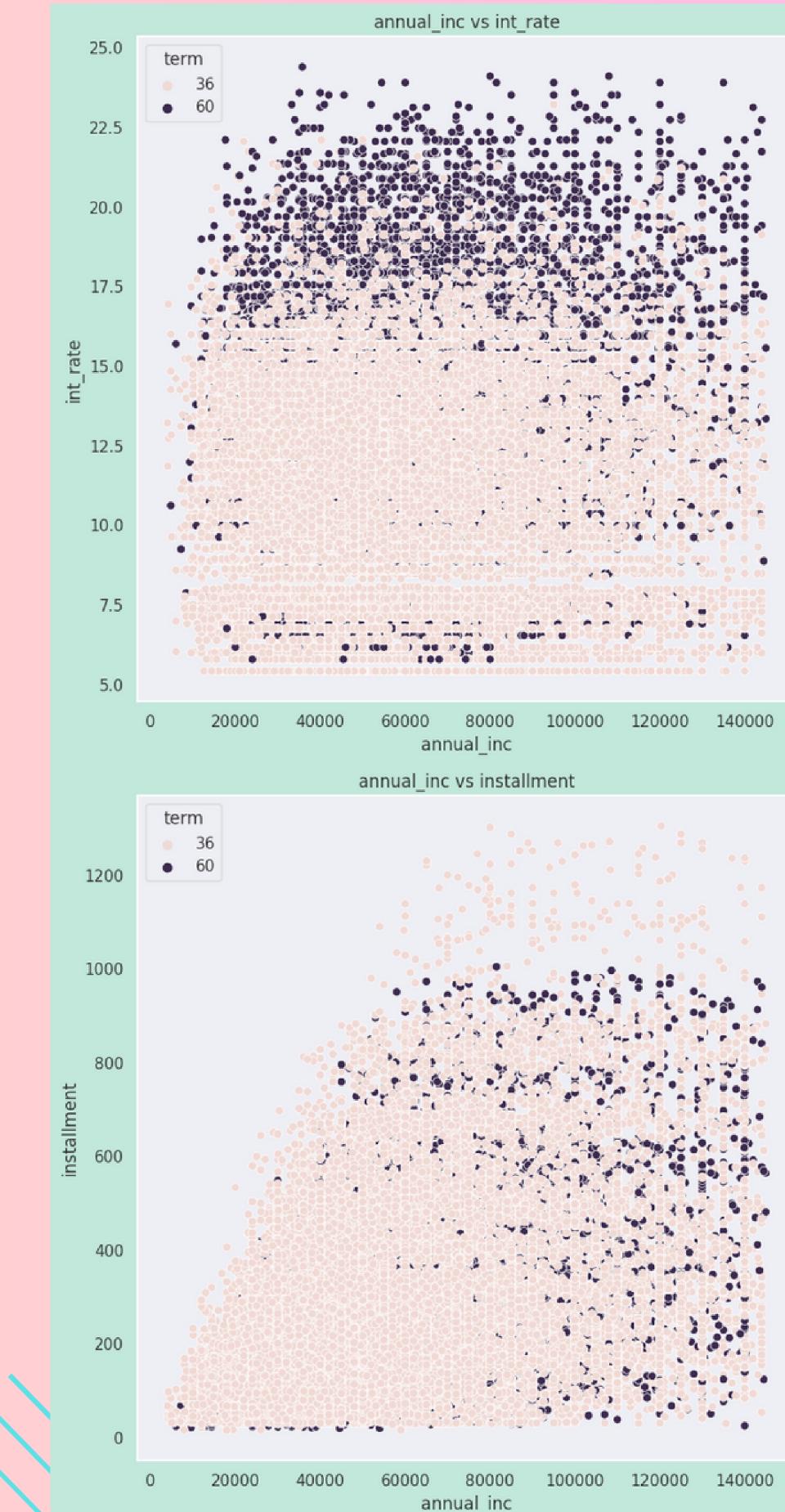
# Following is segmented univariate analysis with loan\_status, verification status, term, emp\_length and purpose



**Bivariate Analysis** : Bivariate analysis is a statistical method that examines the relationship between two variables. It is a simple (two variable) special case of multivariate analysis (where multiple relations between multiple variables are examined simultaneously) **Bivariate analysis between "loan\_amnt" and "int\_rate", "loan\_amnt" and "dti", and "loan\_amnt" and "installment"**



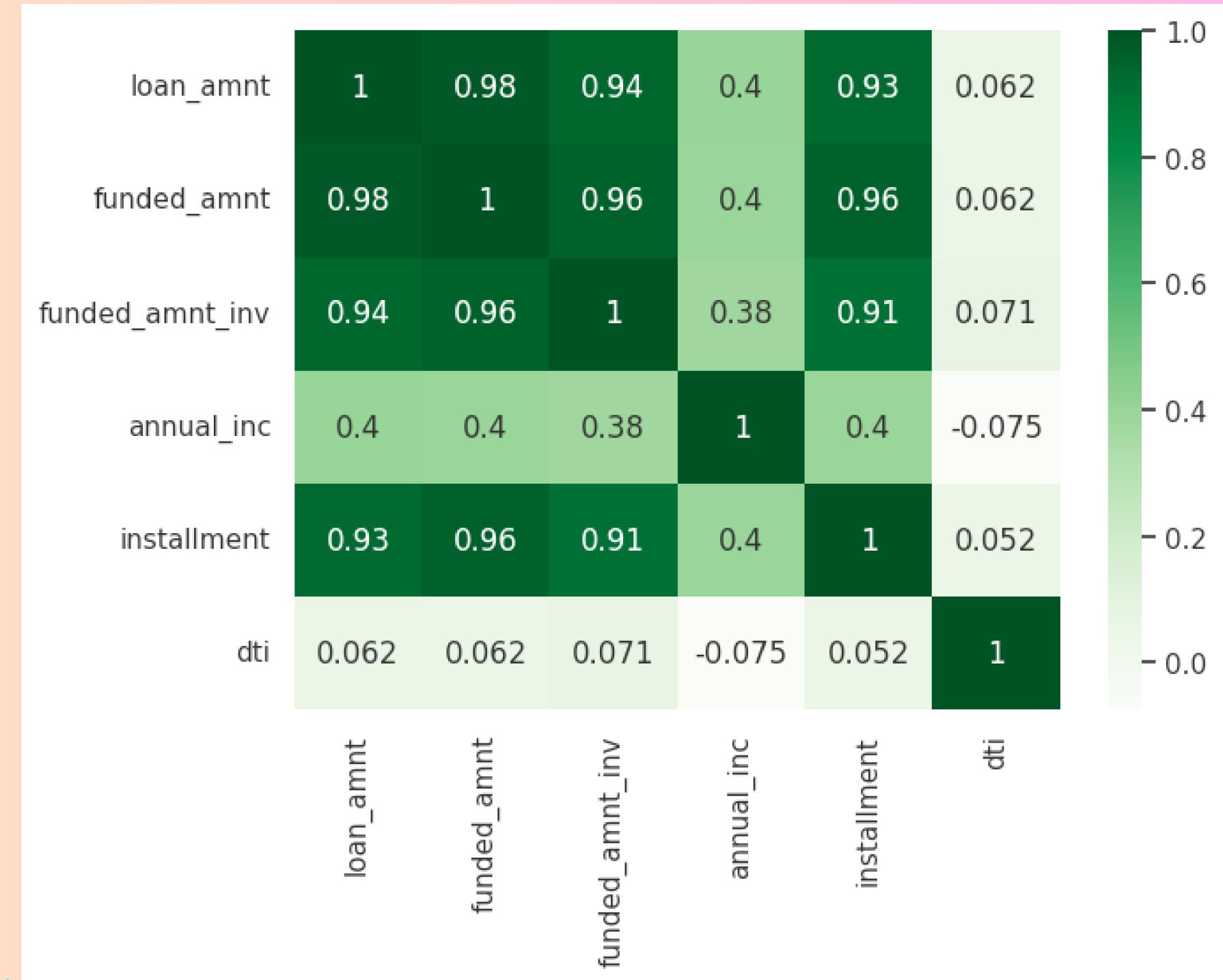
**Bivariate Analysis** : Bivariate analysis is a statistical method that examines the relationship between two variables. It is a simple (two variable) special case of multivariate analysis (where multiple relations between multiple variables are examined simultaneously) **Bivariate analysis between "loan\_amnt" and "int\_rate", "loan\_amnt" and "dti", and "loan\_amnt" and "installment"**



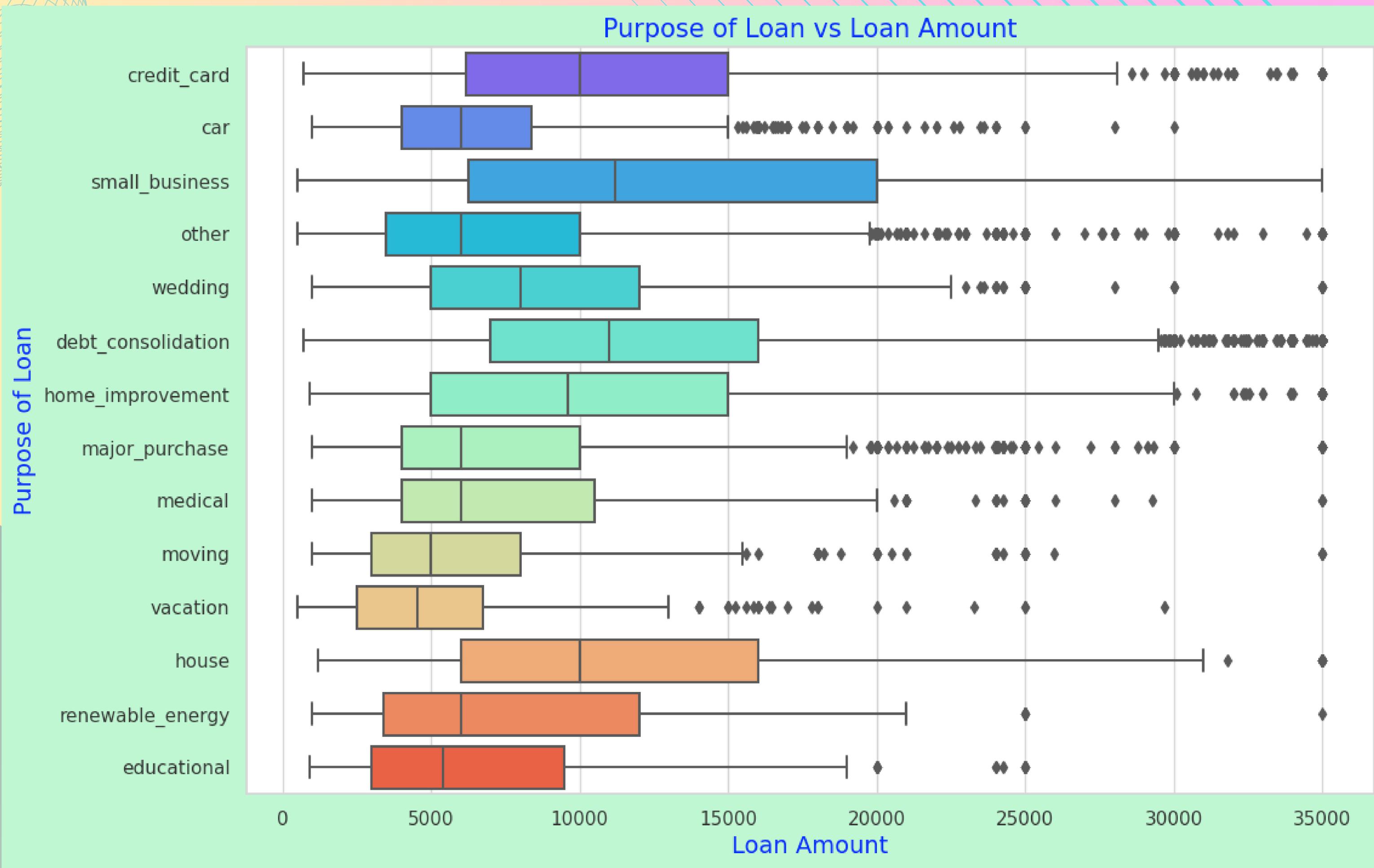
**Bivariate Analysis** : Bivariate analysis is a statistical method that examines the relationship between two variables. It is a simple (two variable) special case of multivariate analysis (where multiple relations between multiple variables are examined simultaneously) **Bivariate analysis between "loan\_amnt" and "int\_rate", "loan\_amnt" and "dti", and "loan\_amnt" and "installment"**



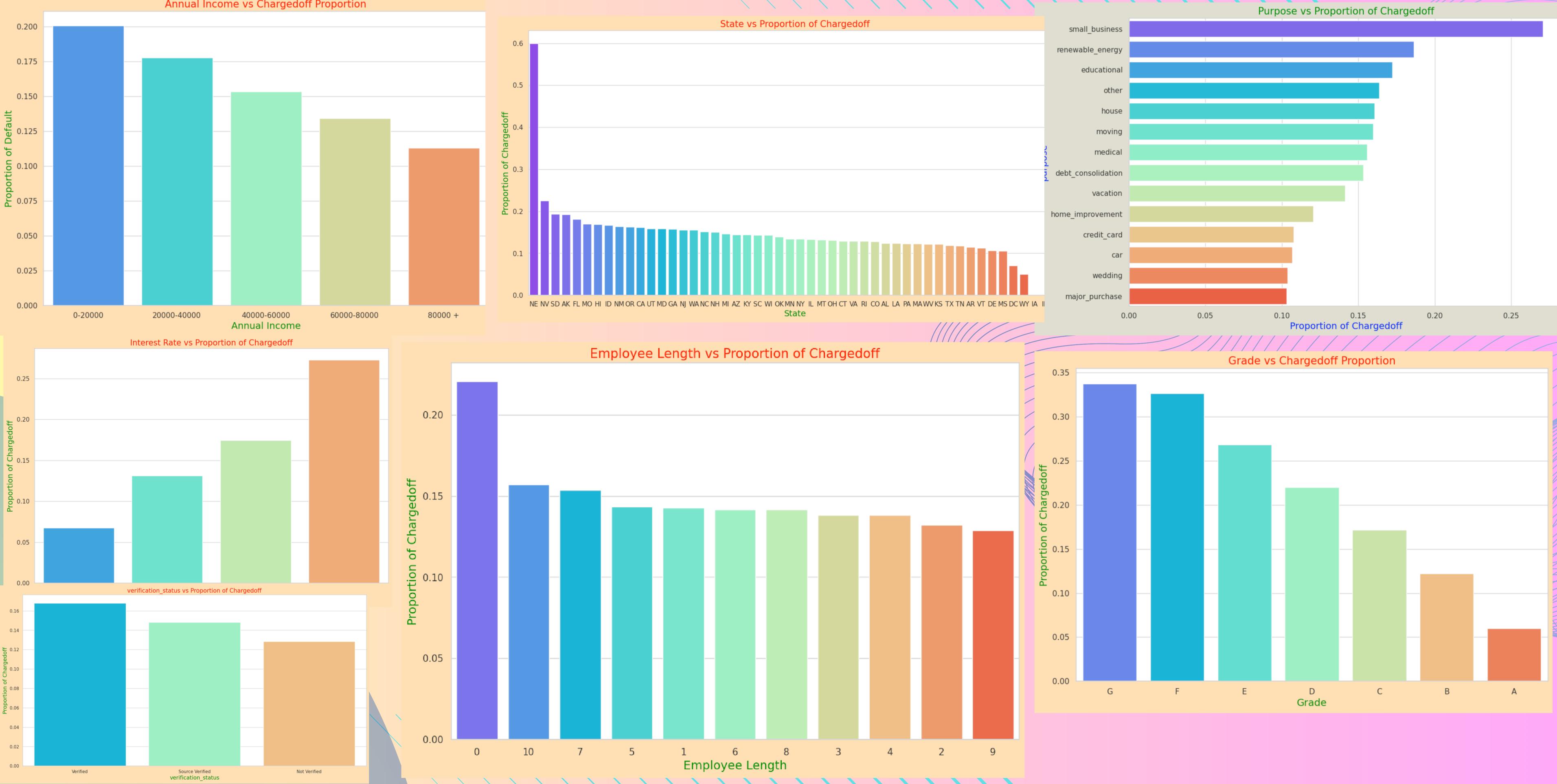
# Bivariate analysis, heatmap of loan amount, annual income, installment and debt to income ratio



# Bivariate Analysis Box Plot Representation, Purpose of loan vs Loan amount

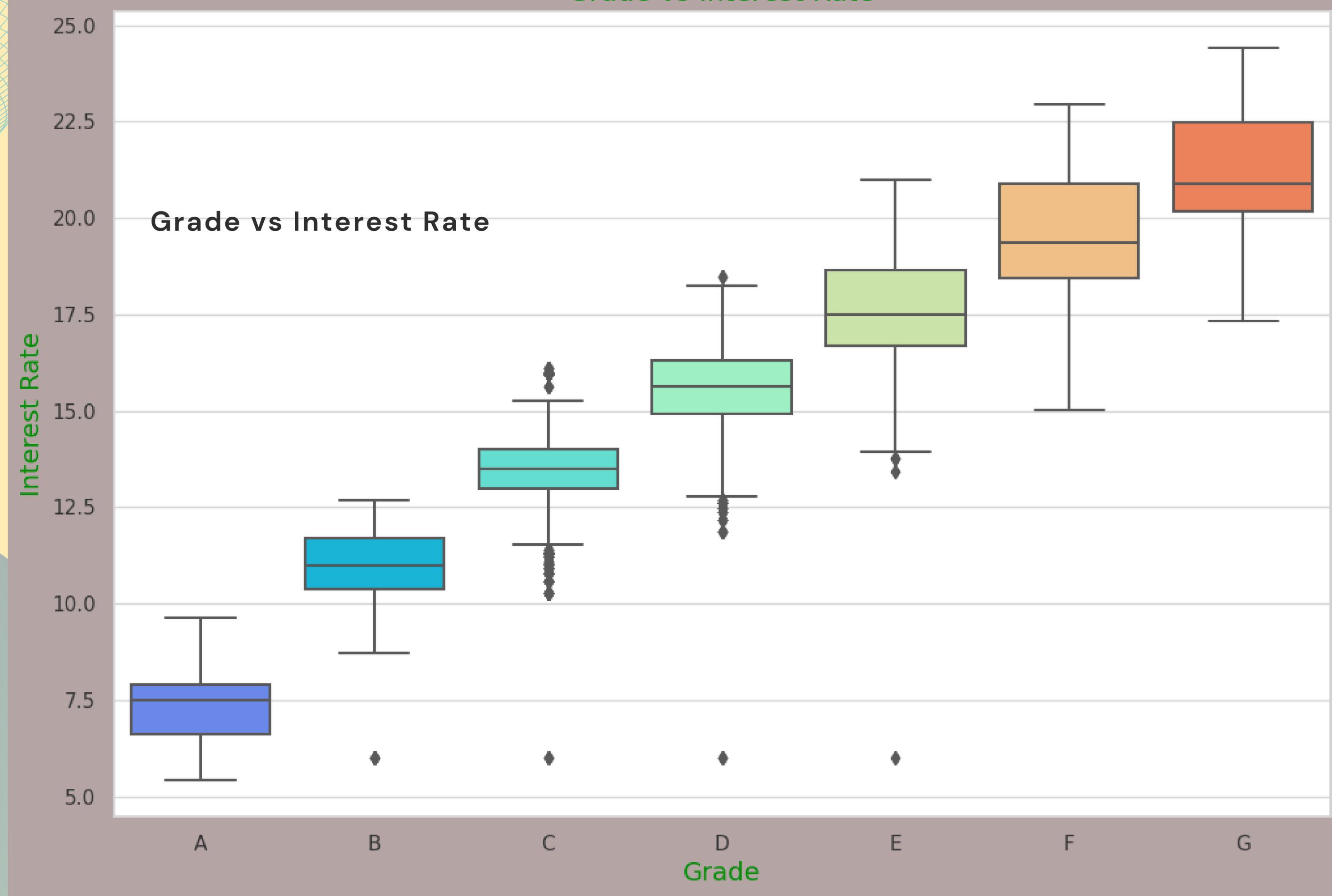


# Bivariate Analysis comparing Proportion of Charged Off with other variables

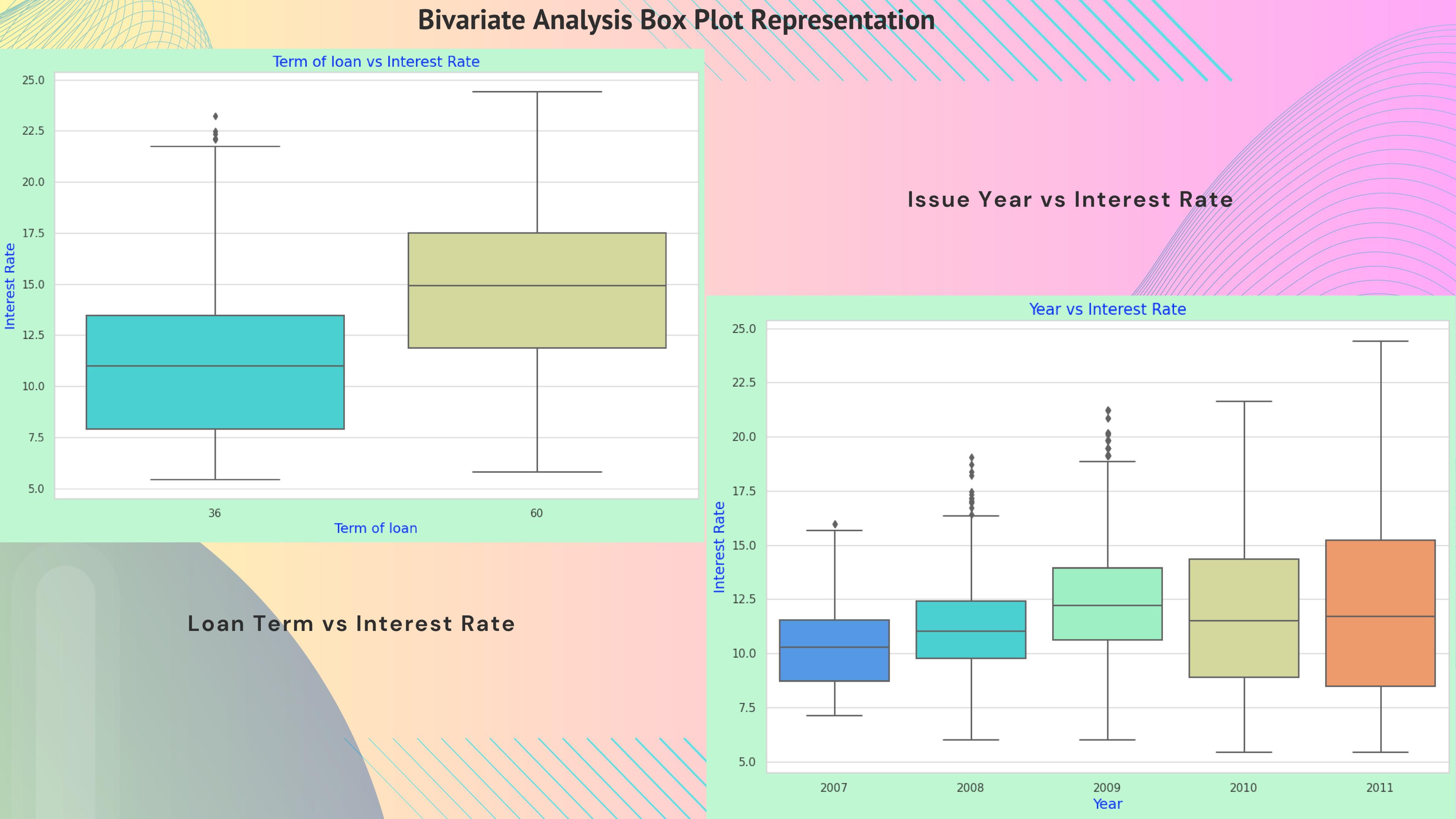


# Bivariate Analysis Box Plot Representation

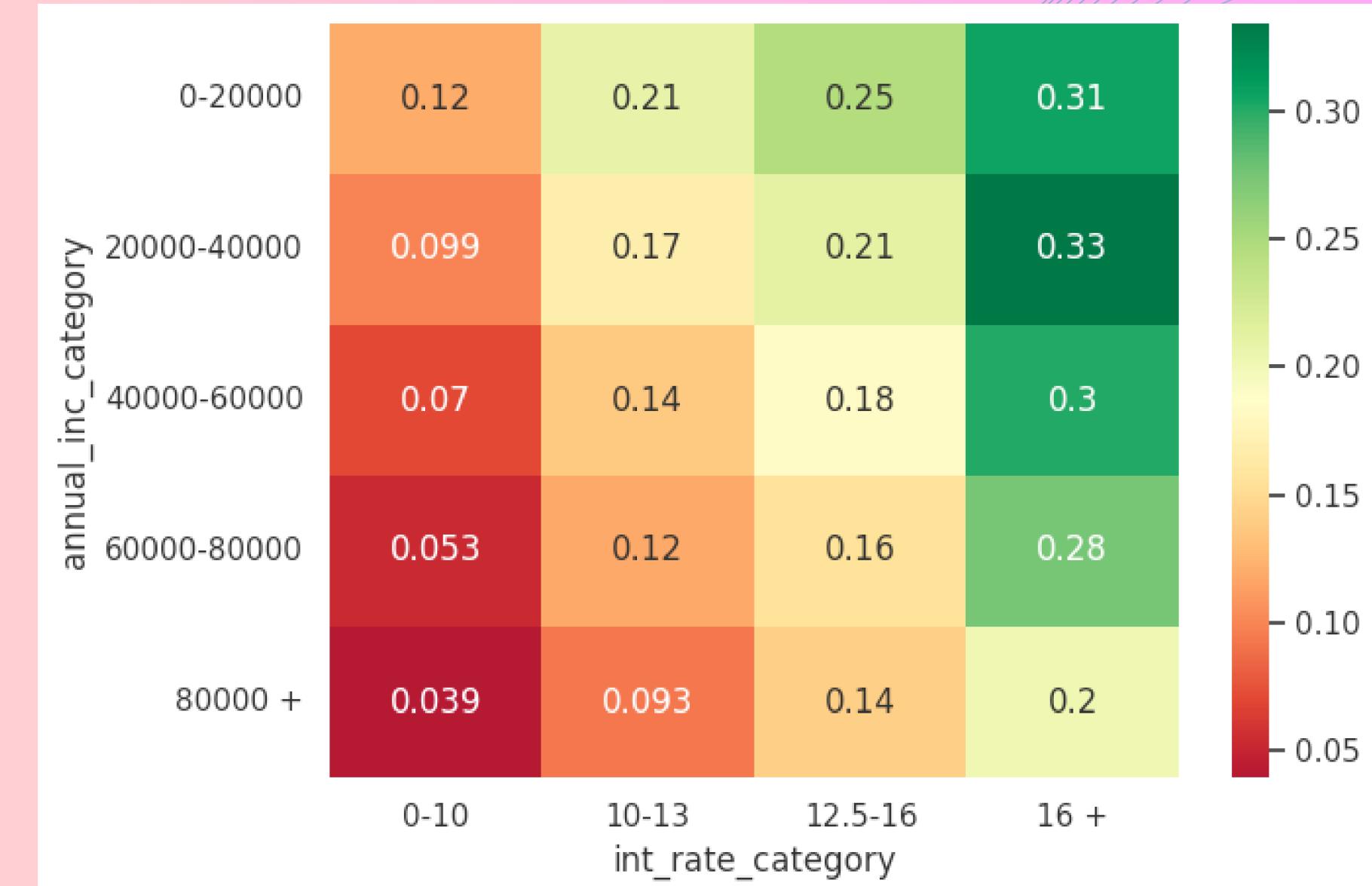
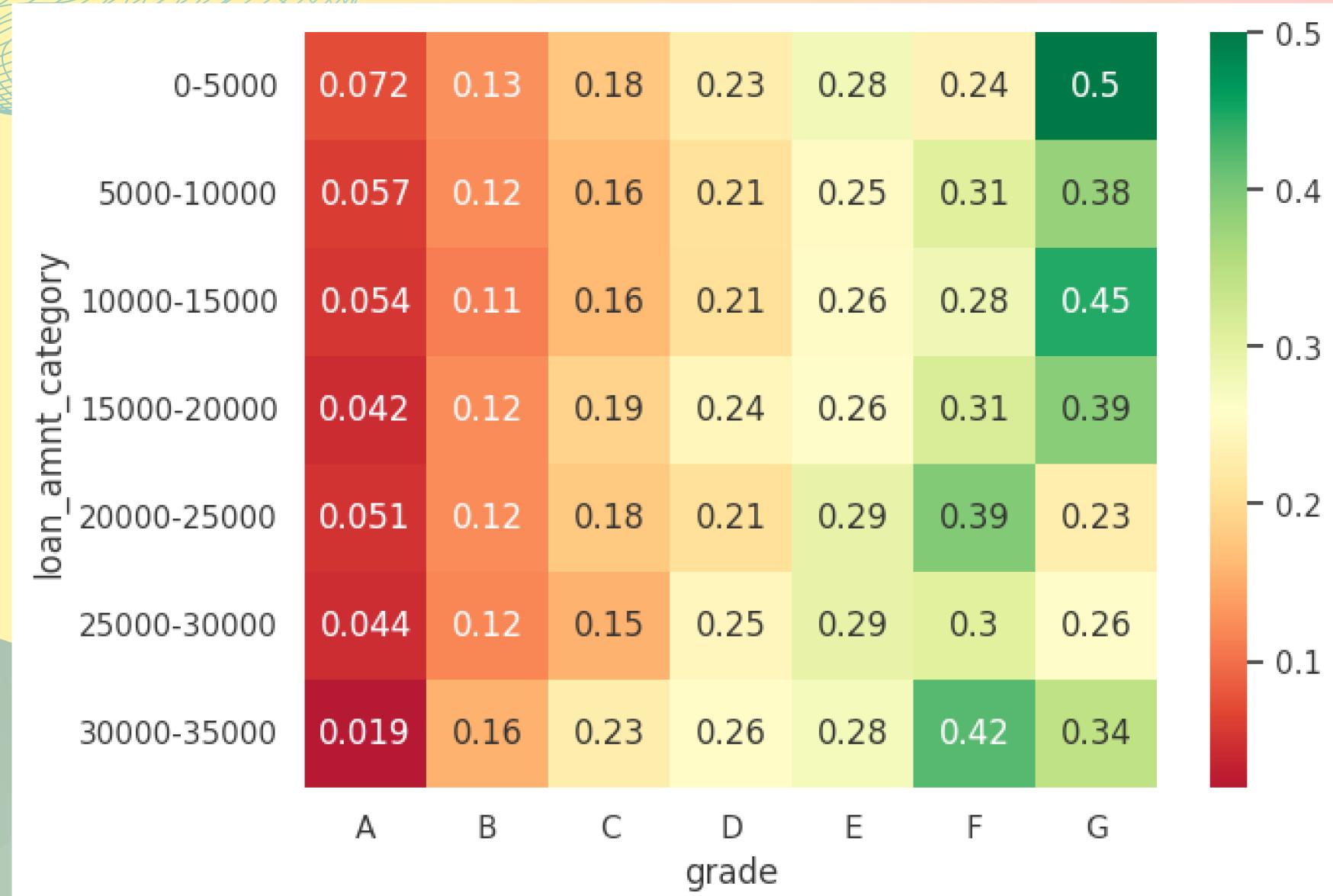
Grade vs Interest Rate



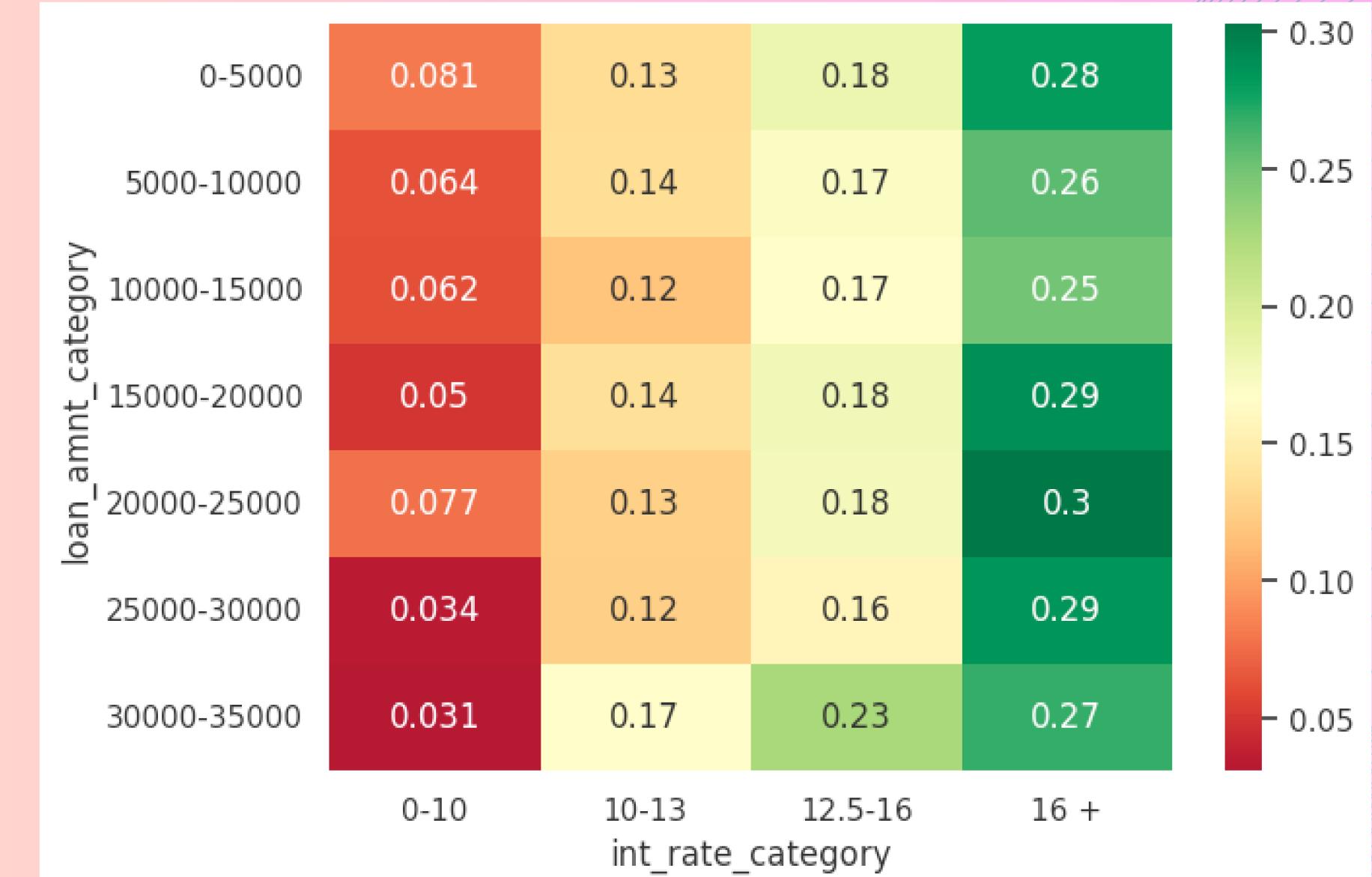
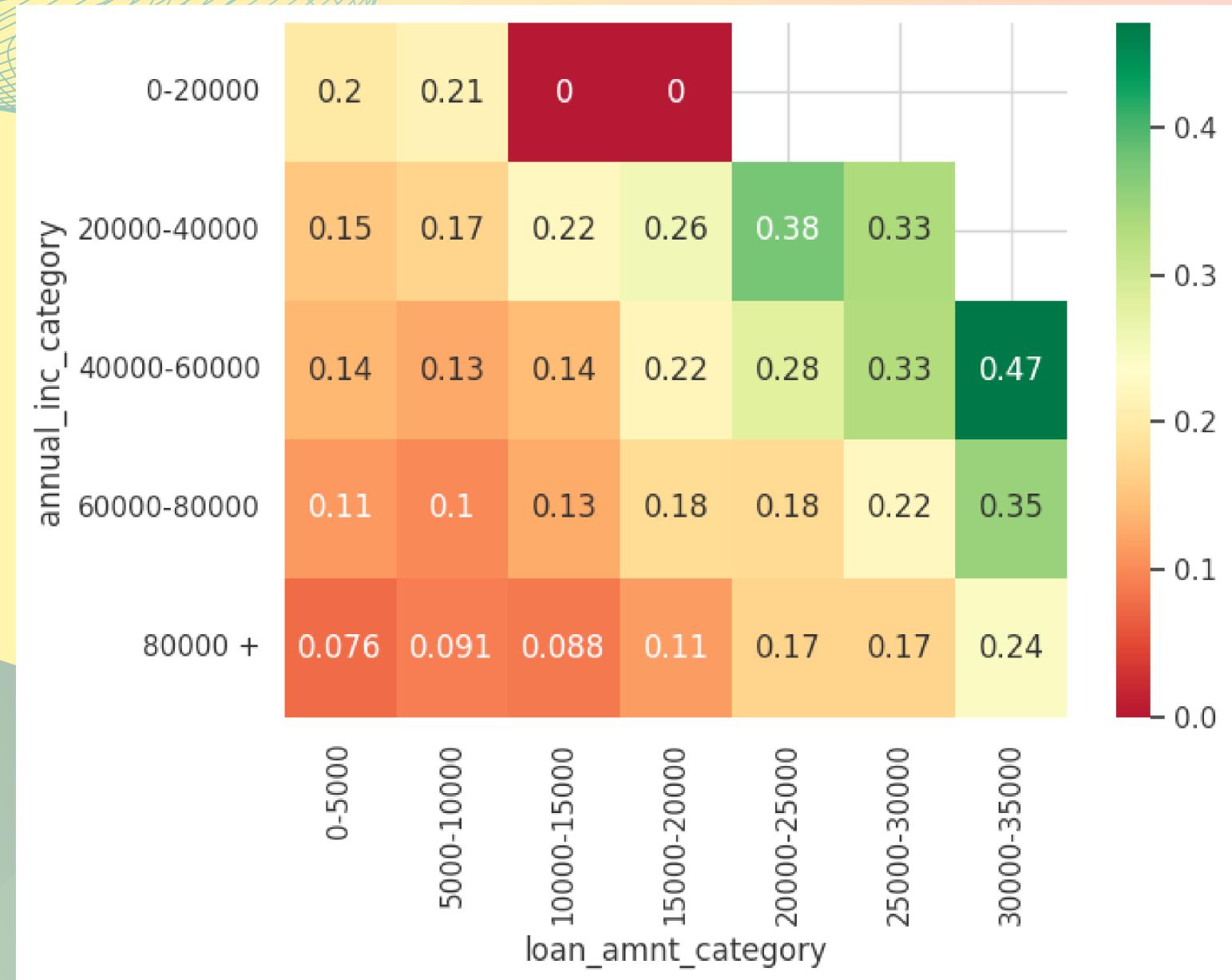
# Bivariate Analysis Box Plot Representation



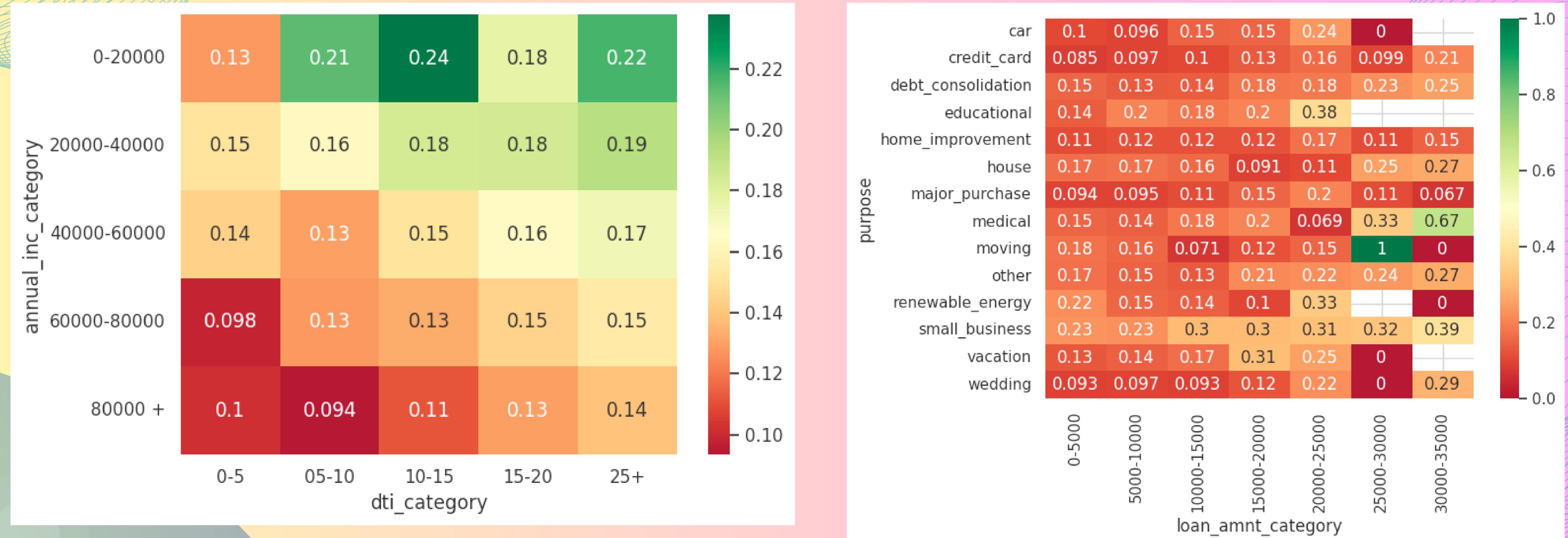
**Multivariate Analysis :** Multivariate analysis is a statistical method that examines the relationships between multiple variables. It is a more complex and powerful tool than bivariate analysis, which only examines the relationship between two variables



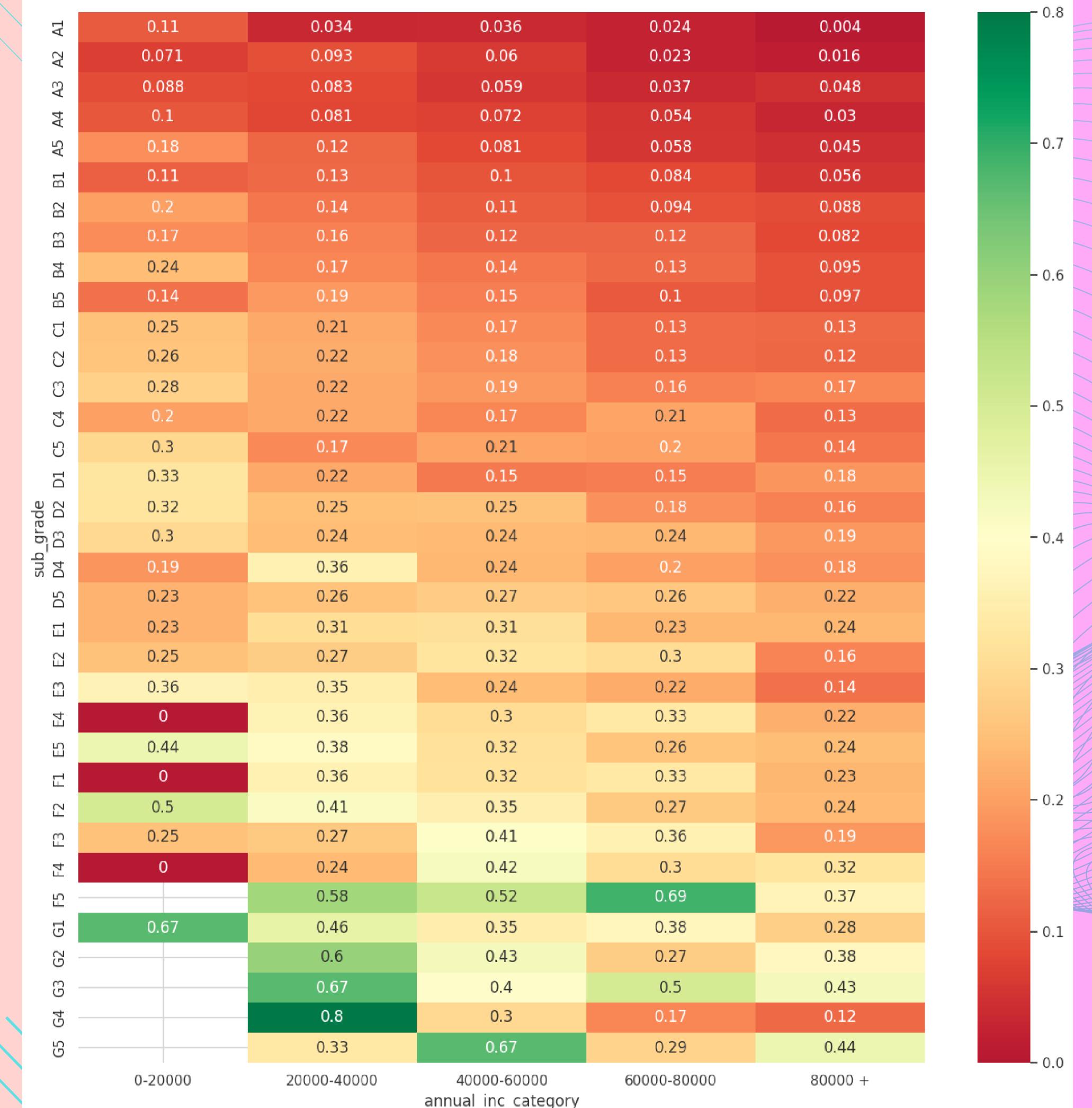
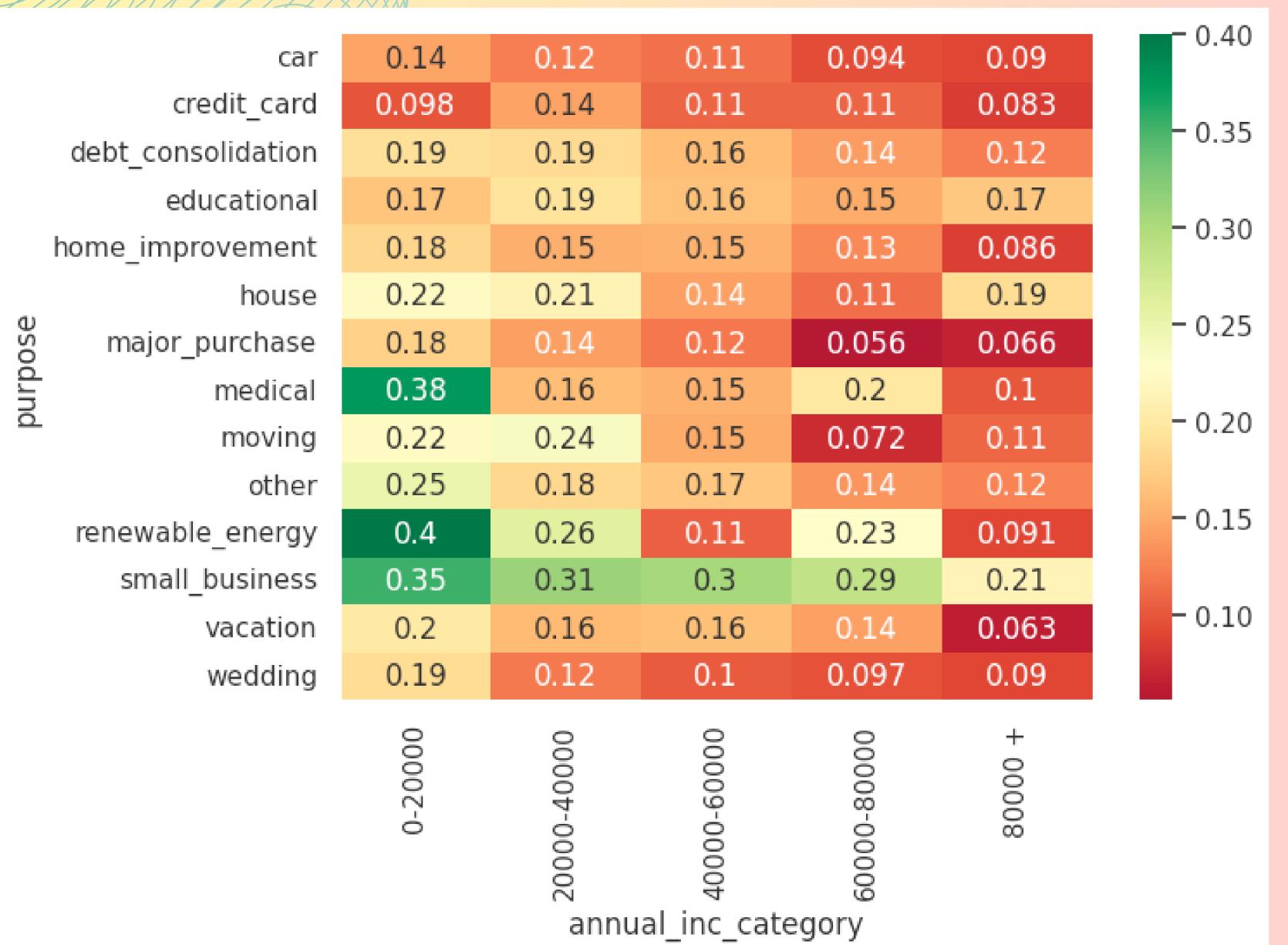
**Multivariate Analysis :** Multivariate analysis is a statistical method that examines the relationships between multiple variables. It is a more complex and powerful tool than bivariate analysis, which only examines the relationship between two variables



**Multivariate Analysis :** Multivariate analysis is a statistical method that examines the relationships between multiple variables. It is a more complex and powerful tool than bivariate analysis, which only examines the relationship between two variables



## Multivariate Analysis :



# Conclusion

## Continuous Variables

- Loans exceeding 15,000 exhibit a higher likelihood of borrower default.
- Funding amounts surpassing 15,000 correlate with elevated default rates.
- Investments exceeding 15,000 demonstrate a higher tendency for default.

### Interest Rate Influence:

A notable increase in default rates accompanies rising interest rates.

### Income and Debt-to-Income Ratio (DTI):

Elevated annual incomes correspond to reduced default rates.  
Conversely, increasing DTI correlates with heightened default rates.

### Influence of Academic Achievement and Income:

Higher incomes coupled with favorable academic performance typically result in fewer defaults.

### Creditworthiness Assessment:

Grades and subgrades predominantly reflect the creditworthiness of borrowers.

# Conclusion

## Categorical Variables

- Loan terms of 60 months exhibit a higher default propensity compared to 36-month terms.
- Default rates escalate as the grade (A, B, C, D, E, F, G) decreases.
- Similarly, default rates rise with decreasing subgrade (A1, A2, B1, B2, etc.).

### **Verification and Borrower Type:**

- Verified borrowers tend to display a higher percentage of defaulted loans.
- Small business borrowers are associated with a heightened default risk.

### **Geographical Influence:**

- The percentage of defaulted loans is notably higher in states like NE and relatively elevated in NV and SD.

Thank  
you!

