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.

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)

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).

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.

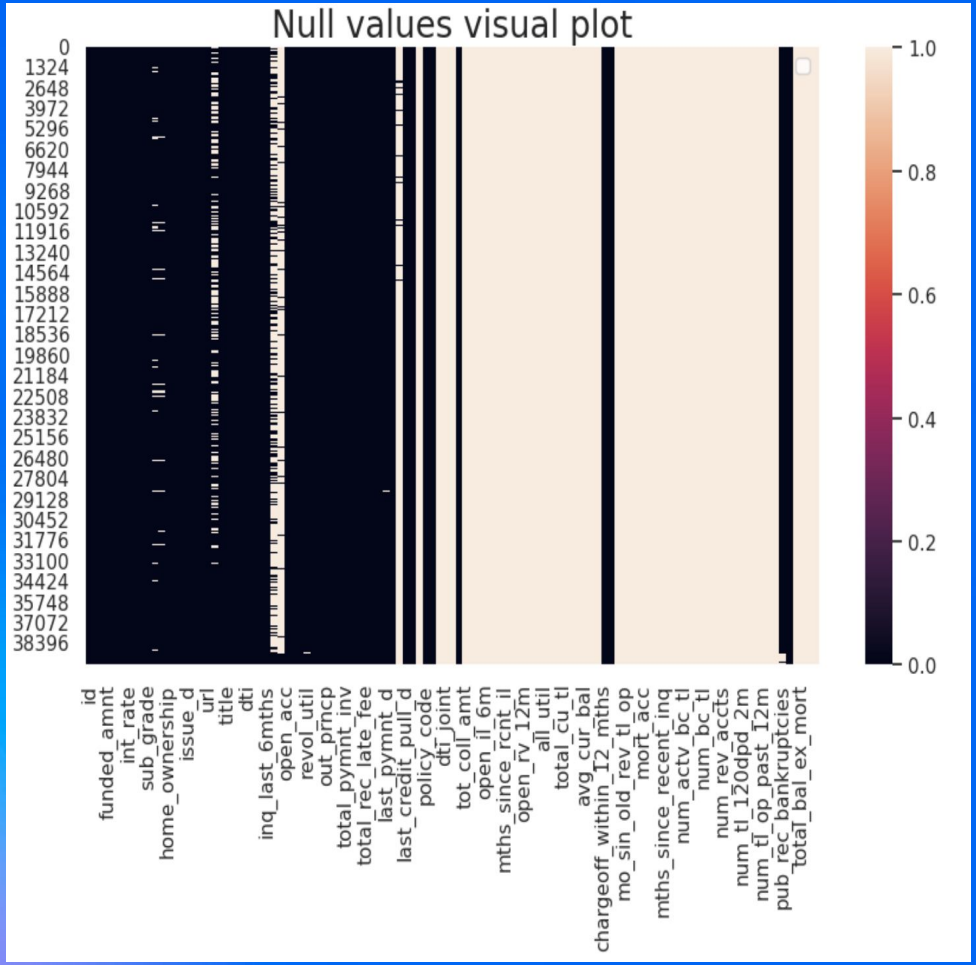
- Identifying and removing missing values
- Exploring and cleaning categorical data
- Formatting numerical data

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.

To be Continued....

Steps we performed in data cleaning:

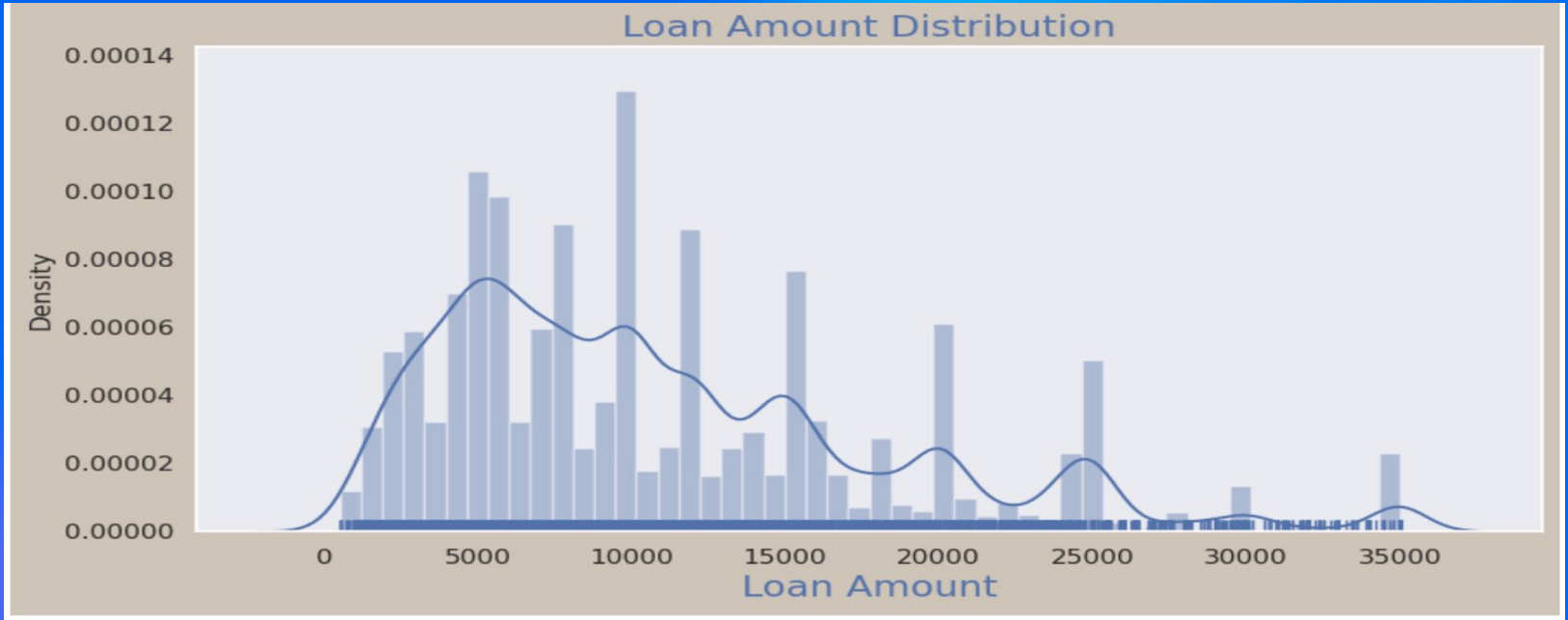
- We identified the rows and columns that have missing values
- Then we removed those columns and rows.
- We removed many unnecessary columns and reduced the number of columns from 111 to 28.
- We formatted and converted many columns to proper data type.
  For example - emp_length was object and we converted it to float.



Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an essential step in und and analyzing data. In the case of a lending club, EDA can help us gain insights into the loan data, identify patterns, discover relationships, and make informed decisions

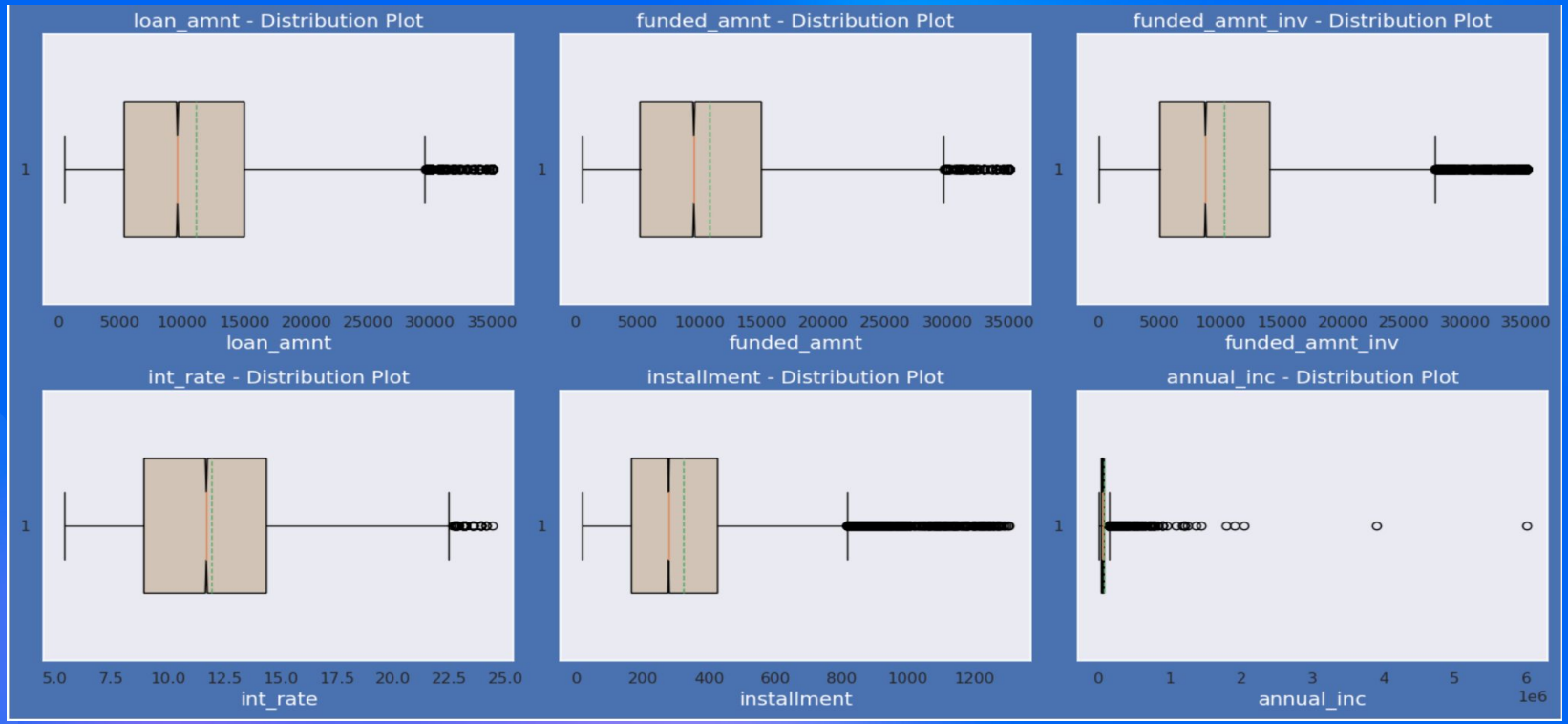
Univariate Analysis : Univariate analysis is a statistical method that analyzes one variable at a time. It is the simplest form of data analysis and is often used as a first step in understanding a dataset. Univariate analysis can be used to describe the distribution of a variable, identify outliers,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



We did univariate analysis both on continuous and categorical variables:

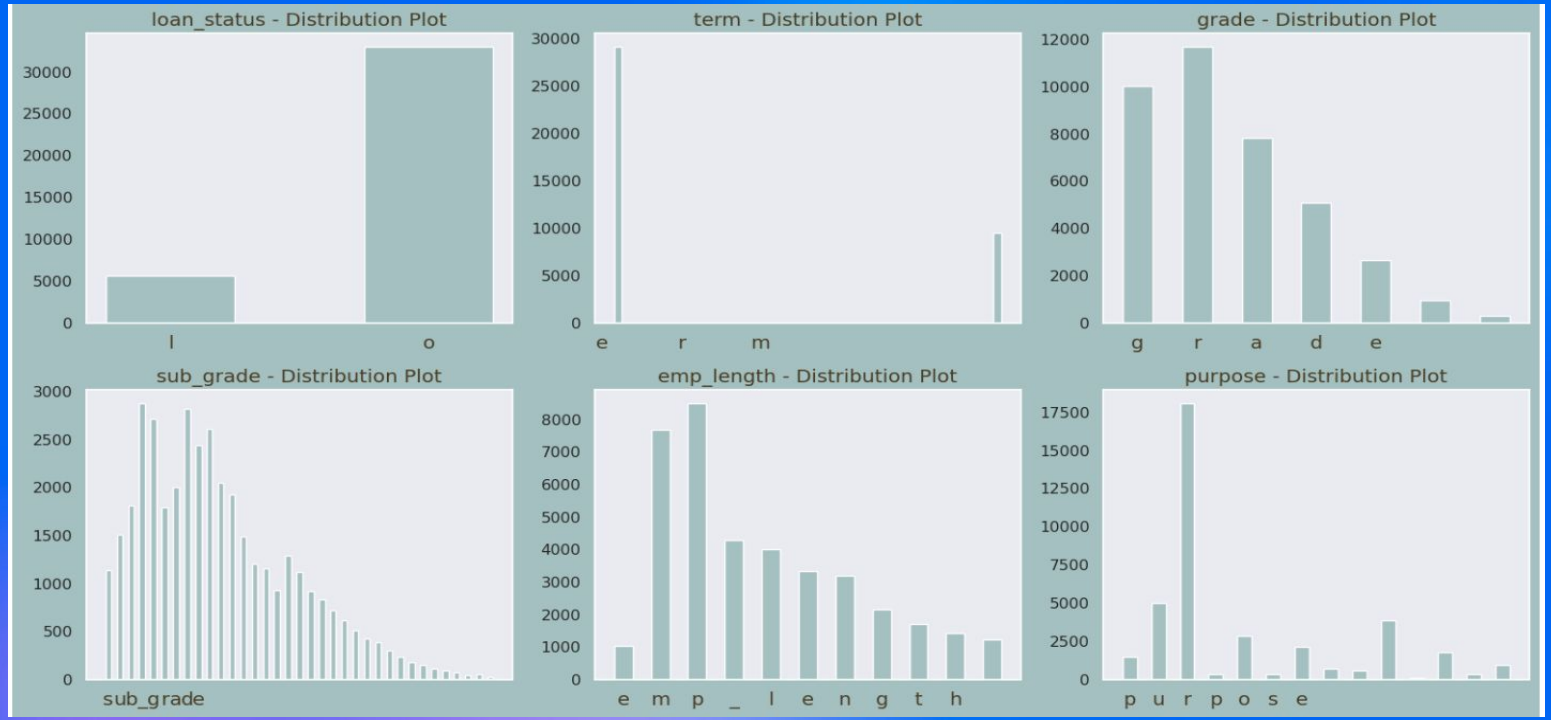
Following is the univariate analysis on continuous variables: **loan_amount**, **funded_amount**, **funded_amnt_inv**, **int_rate**, **installment**, **annual_inc.**

Box Plot Representation



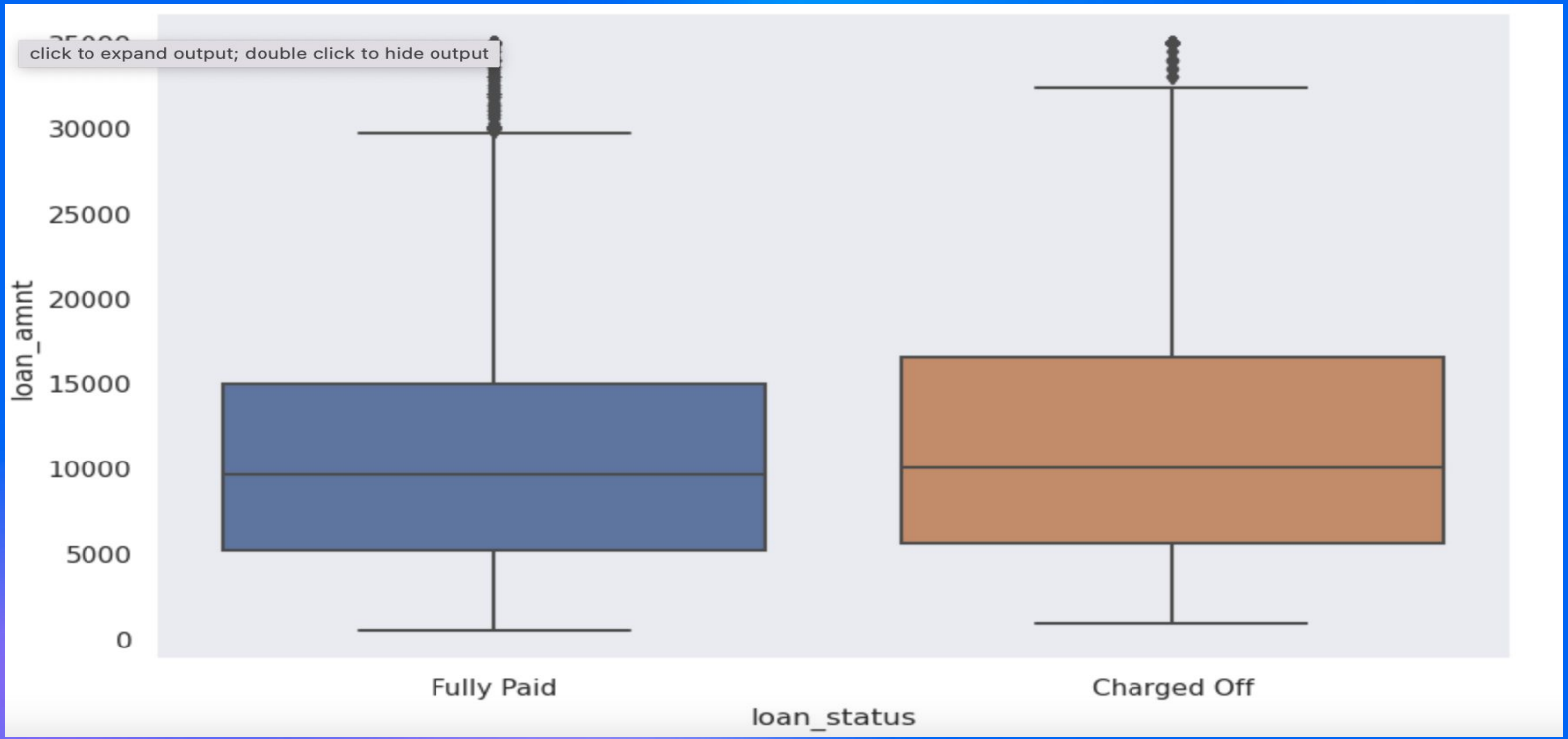
Following is the univariate analysis on **categorical** variables: **emp_length, grade, sub_grade, term, loan_status**

Count Plot Histogram Representation

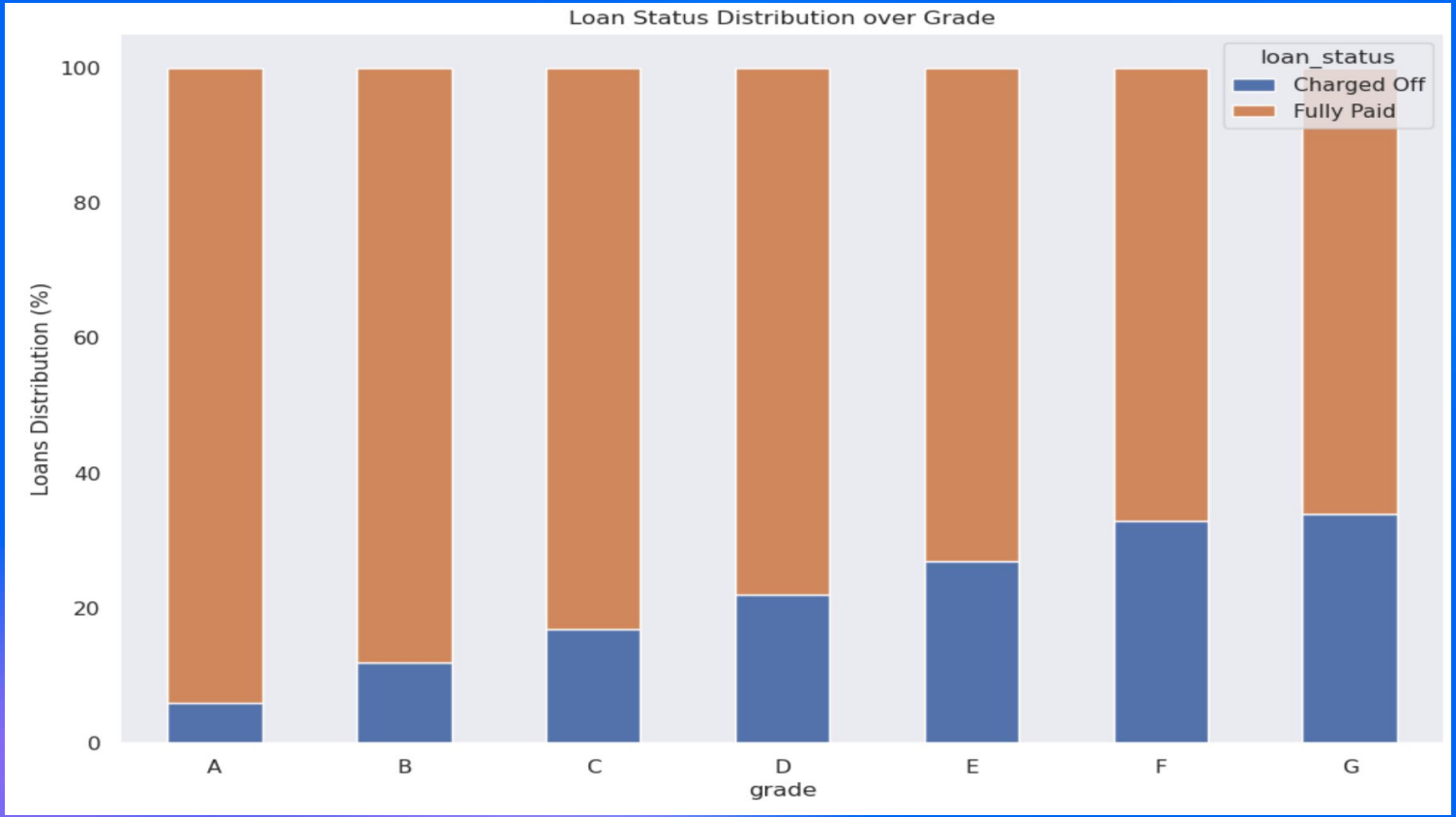


Segmented Univariate Analysis: The process typically involves dividing the dataset into distinct segments or subsets based on a specific criterion or grouping variable. Then, an analysis is performed on each segment individually, treating them as separate univariate analyse

Box plot representation of loan_amount vs loan_status

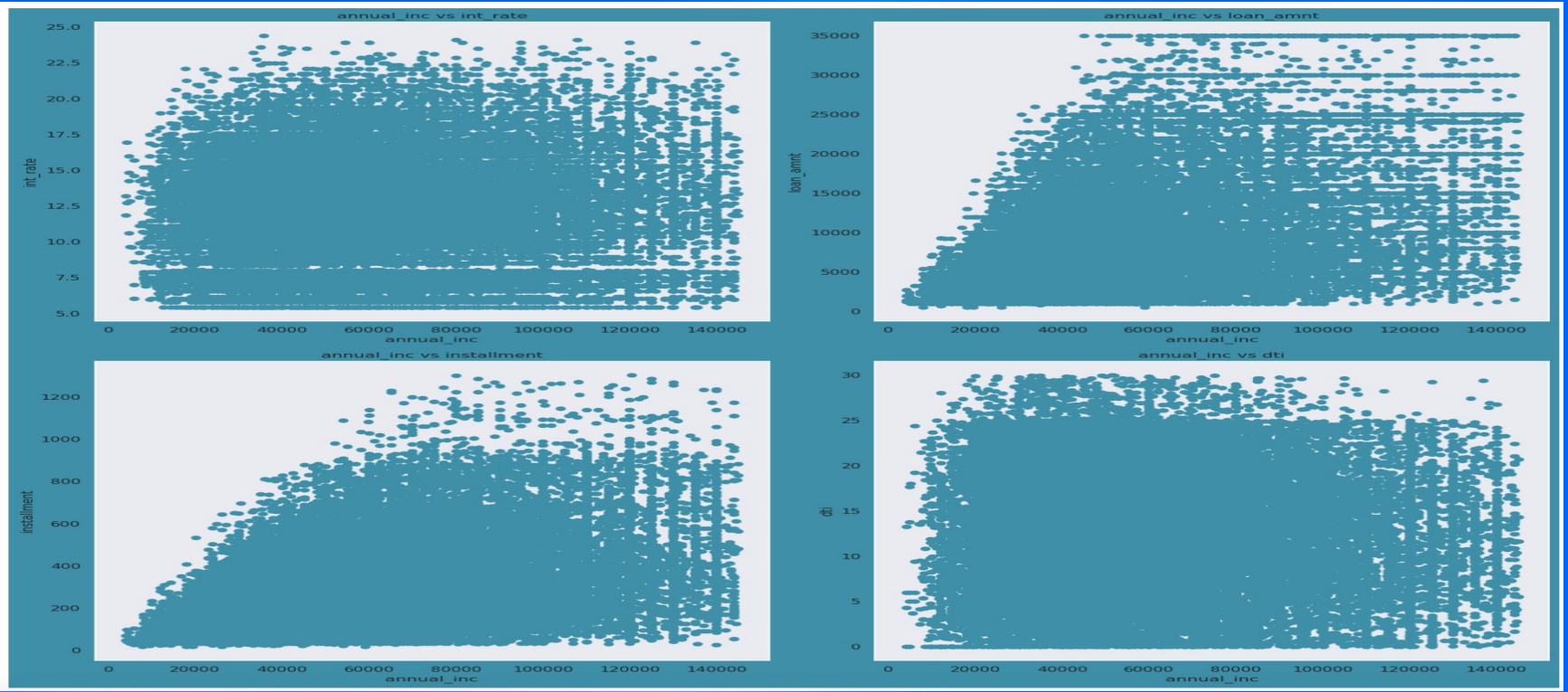


Following is segmented univariate analysis with loan_status distribution over grade

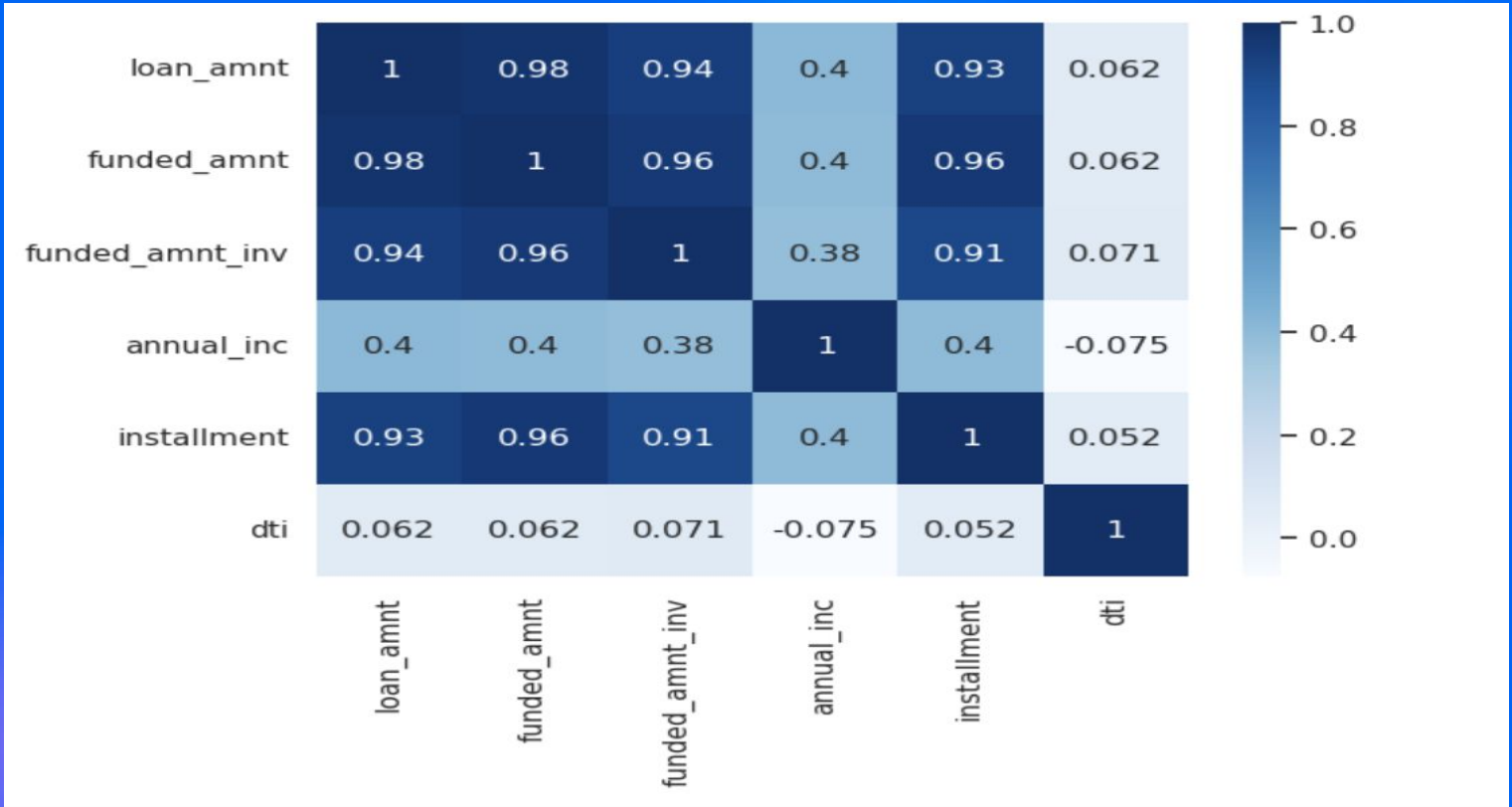


Bivariate Analysis : Bivariate analysis is a statistical method that examines the relationship between two variables. It is a simple (two variable) special case of multivariate analysis (where multiple relations between multiple variables are examined simultaneously)

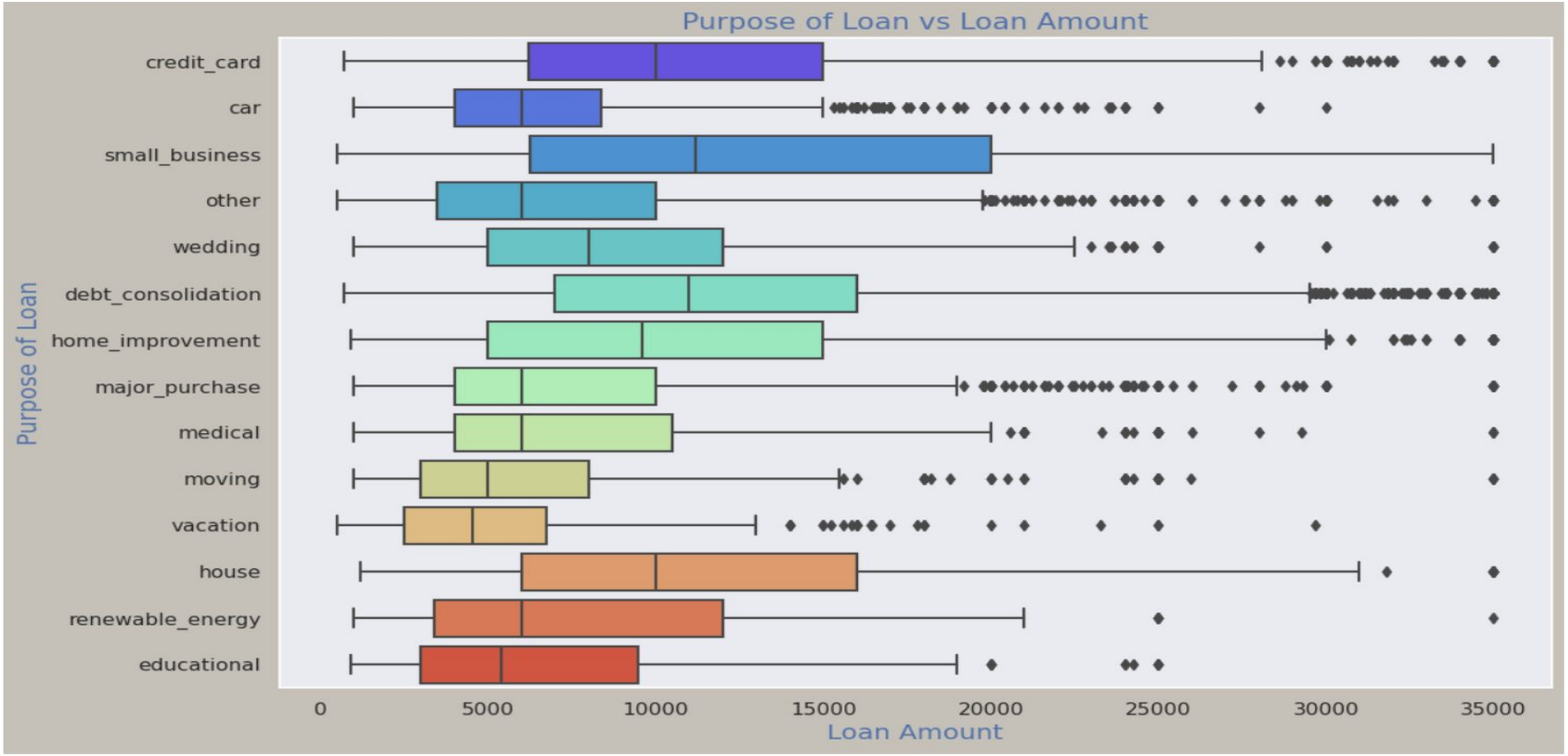
Bivariate analysis between "loan_amnt" and "int_rate", "loan_amnt" and "dti", and "loan_amnt" and "installment"



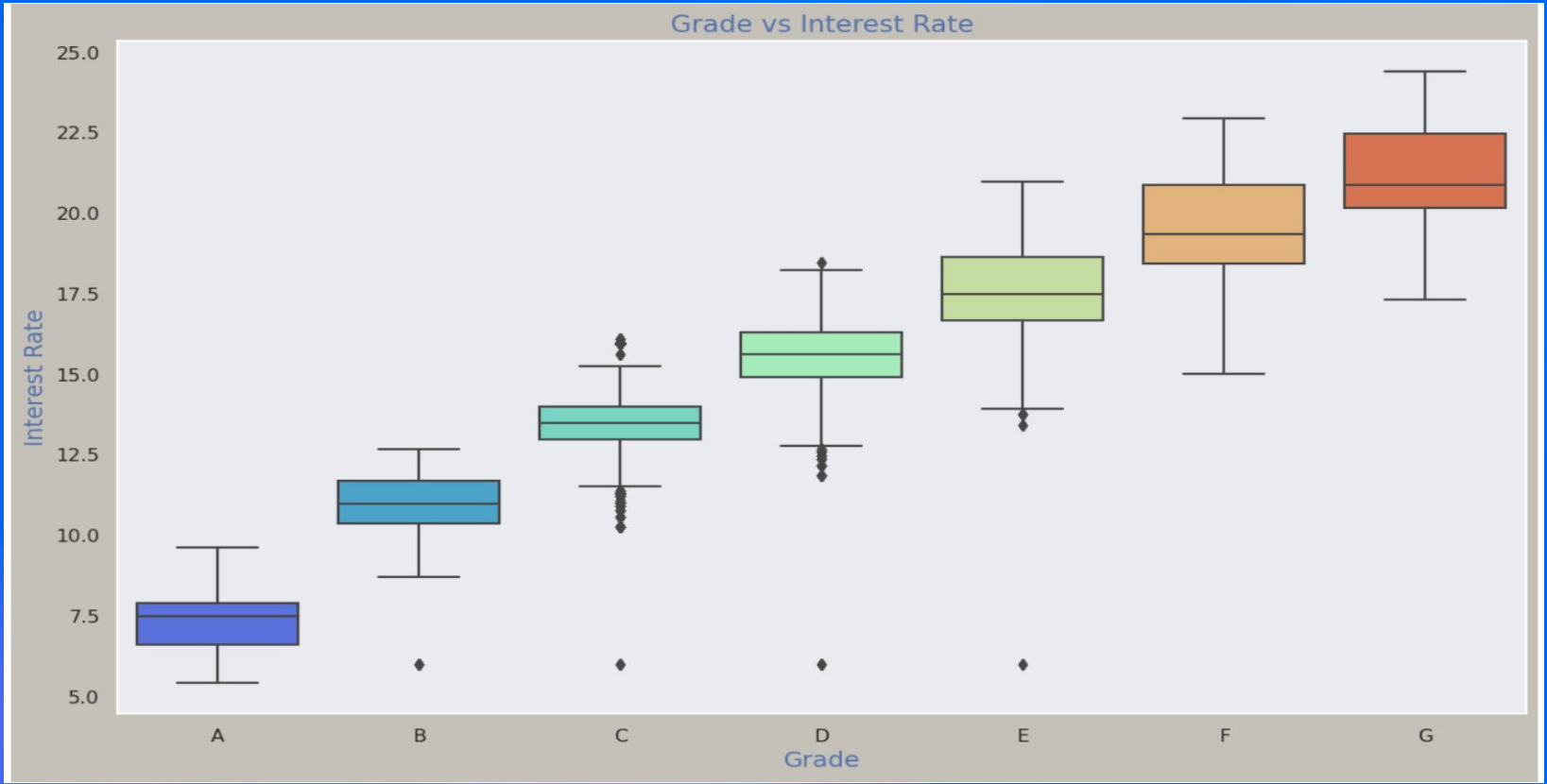
Bivariate analysis, heatmap of loan amount, annual income, installment and debt to income ratio



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

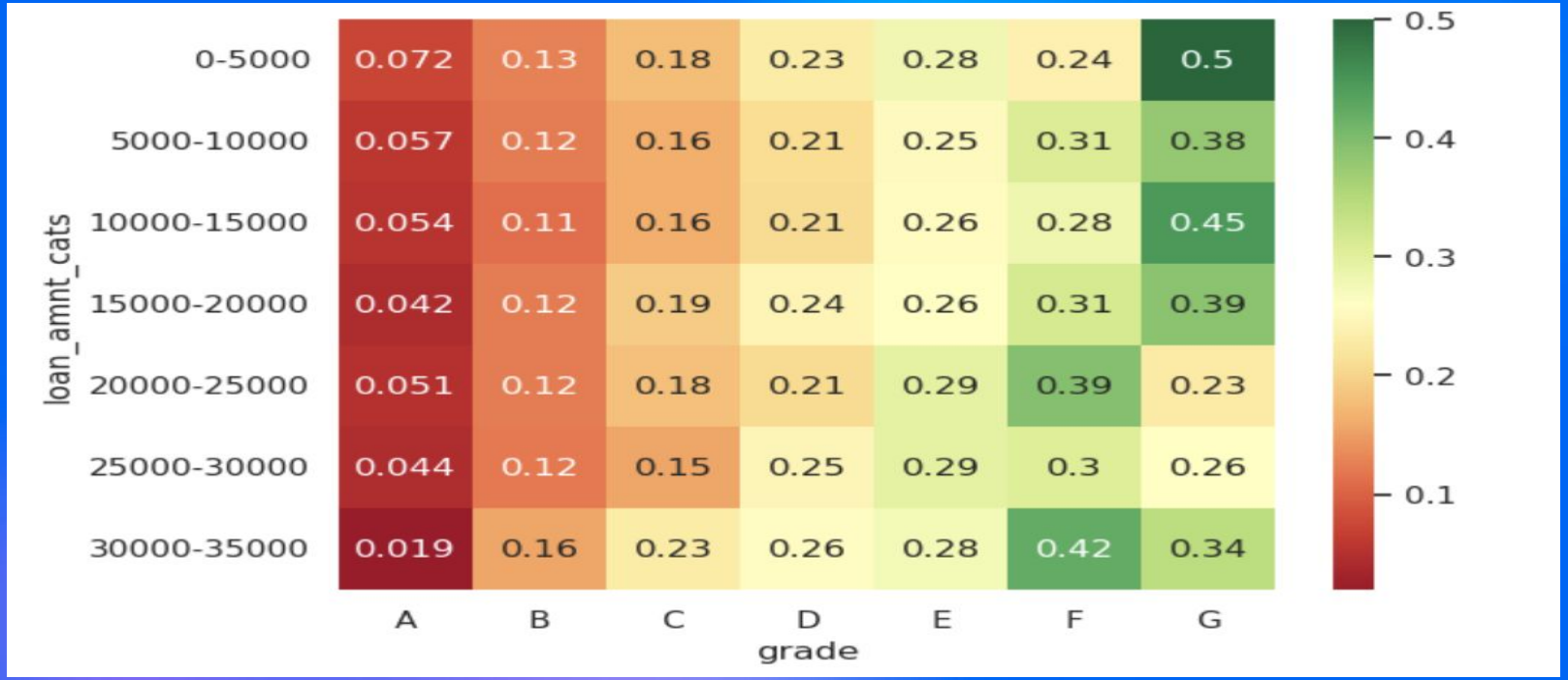


Bivariate Analysis Box Plot Representation, Grade vs Interest Rate

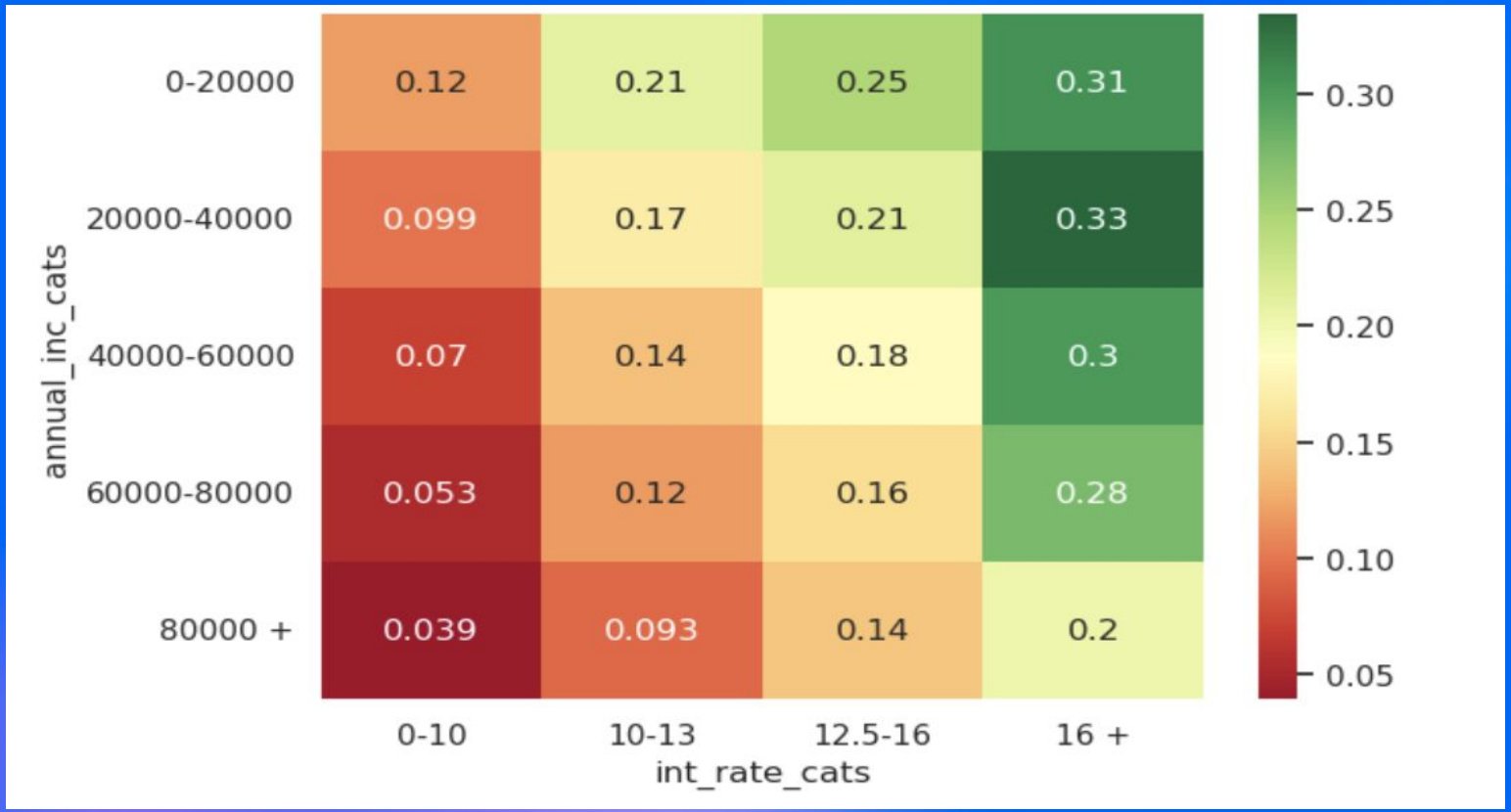


Multivariate Analysis : Multivariate analysis is a statistical method that examines the relationships between multiple variables. It is a more complex and powerful tool than bivariate analysis, which only examines the relationship between two variables

heatmap of grade vs loan_amnt_cats vs loan_status_boolean



Multivariate analysis heatmap of annual_inc_cats vs int_rate_cats vs loan_status_boolean



Conclusion :

Key determinants influencing the likelihood of borrower default:

Continuous Variable

- Loan amount greater than 15000 dollars have higher default rate
- Funded amount greater than 15000 dollars have higher default rate
- Funded amount invested greater than 15000 dollars have higher default rate
- As Interest rate increases the default rate increases steeply
- As the annual income increase the default rate decreases
- As dti increase the default rate increases
- Higher Incomes along with good grades usually leads to less defaulting
- Grades and Sub grades by and large depict creditworthiness of a borrower

Categorical Variable

- 60 months term have a higher default rate than 36 months term
- As the Grade decreases (A B C D E F G) default rate increases
- As the Subgrade decreases (A1 A2 B1 B2.....) default rate increases
- The Percentage of loans defaulted is higher for verified borrowers
- Small business borrowers have high default rate
- The Percentage of loan defaulted is very high for state NE and high for NV and SD

Team

Bhuvnesh Kumar

Pragya Thakur