Project Name : LENDING CLUB LOAN ANALYSIS

Group :

Radha Tanuku and Thejaswini M

Table of Contents

INTRODUCTION:

Problem Statement:

Lending Club is a consumer finance company which specialises in lending various types of loans to urban customers. 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

PROJECT OBJECTIVE

The objective of this project is to analyse Lending Club loan data to identify

- Key risk factors influencing loan approval and default rates
- Generate actionable insights that can enhance the decision-making process for lenders.

DATASET OVERVIEW

The Loan dataset contains customer details with the history of previous loan details including the loan status.

The loan status bifurcate the loans under 3 categories as below:

1. Charge off – where the loans are not paid, but fall in defaulters list
2. Fully Paid- where the loans for all instalments are fully paid
3. Current – Which is in process loan.

Our analysis is more focussed on two classes

1. Charge off
2. Fully Paid

As our aim to understand the dataset /borrowers who are actually the defaulters avoiding the risk of approving loan and loosing business.

ANALYSIS METHODOLOGY

Below are the software and relevant libraries with versions used in this case study:

- Python Version: 3.9.19
- Pandas: 2.2.3
- Numpy: 2.0.2
- Matplotlib: 3.9.2
- Seaborn : 0.13.2

DATA CLEANING

1. **Missing Value Check**

2. **Dropping null values**

Initially, when we picked the dataset there were total 39717 rows and 111 columns. Out of which, we understood that there are 58 columns, which has nulls more than 30%. The columns which are having more nulls have been dropped.. Resulting to 53 total columns.

3. **Missing Value Imputation**

The rows which are numeric, which have Null values have been replaced with median.
The rows which have categorical values have been replaced with mode.

HANDLING INCORRECT DATA TYPES

- For calculating correlative coefficient or Heatmap, it is necessary to convert all data types to integers.

- We have converted all objects into integers, so that we arrive at the right statistics.

DATA CORRECTION

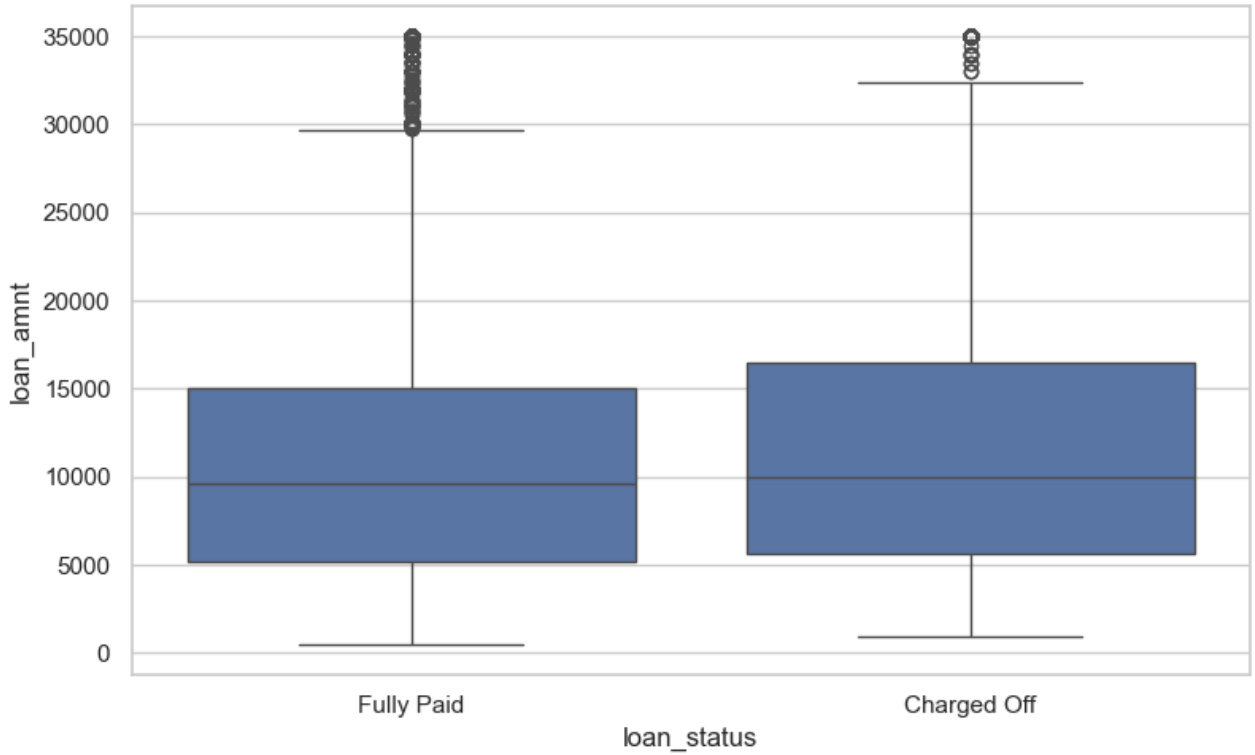
During this process, we have removed the special characters and strings where ever required.

Example: Interest rates have % suffix to the numbers is removed.

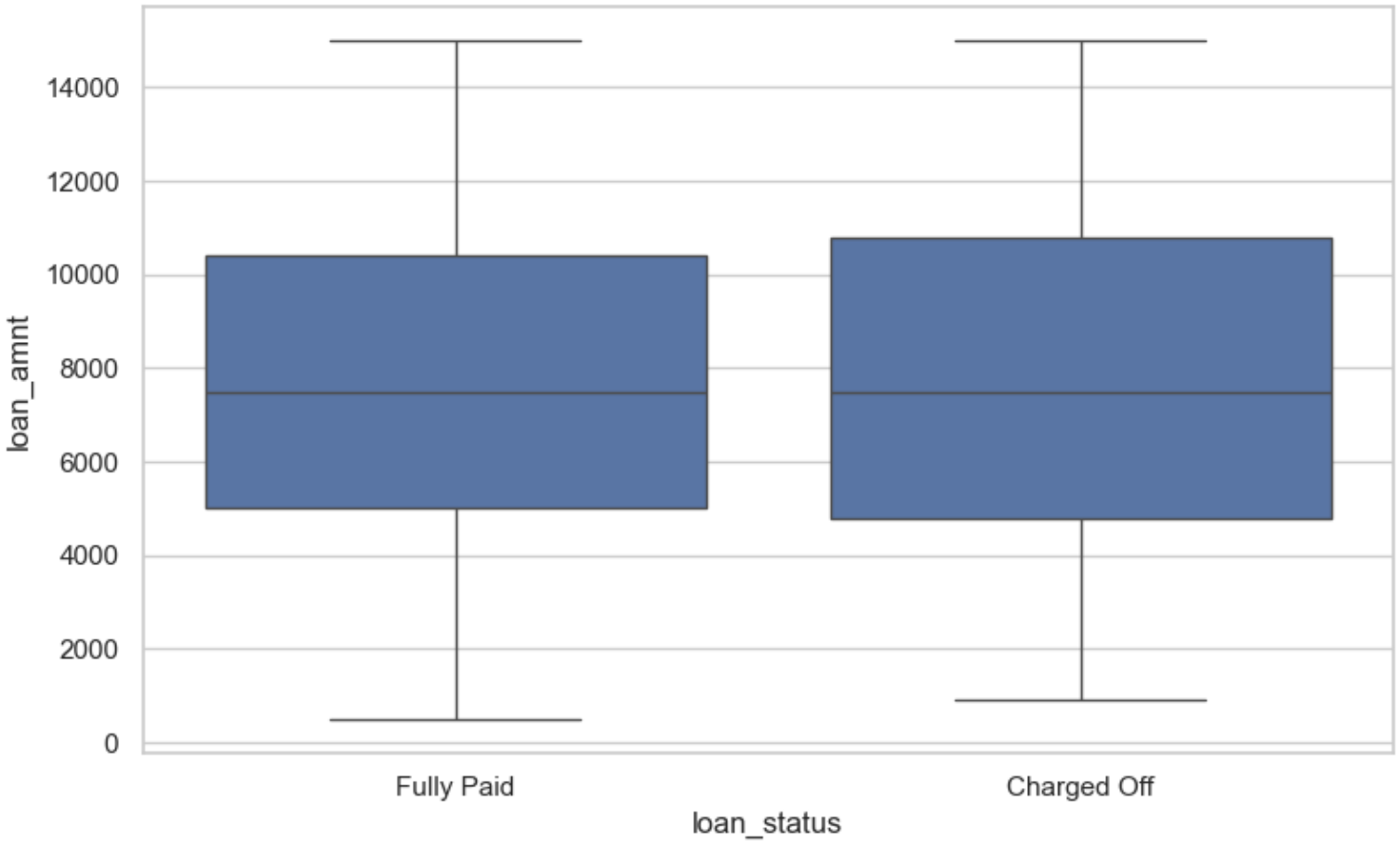
Note: In this process, we also identified that there are 85% of borrowers fully paid the loans and the 15 % of borrowers are the defaulters.

OUTLIERS ANALYSIS USING BOXPLOT

The box plot image below inferred that there is huge difference between mean and median which lead us to limit our data to 75% to remove the outliers



The box plot image below from the filter dataset in the previous slide resulting zero outliers.



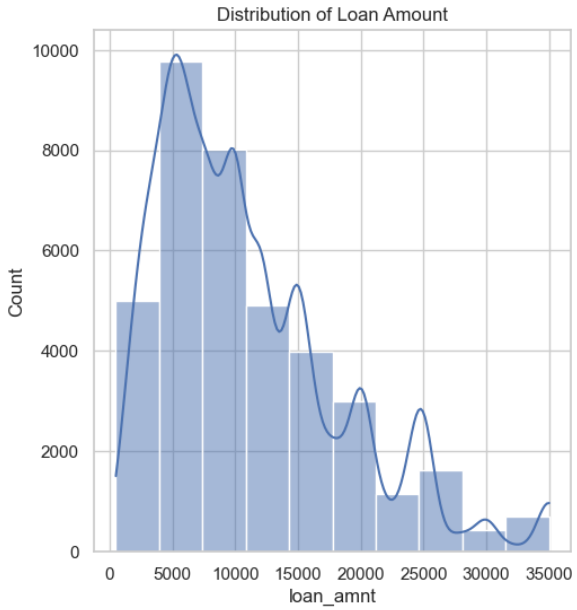
DATA VISUALIZATION

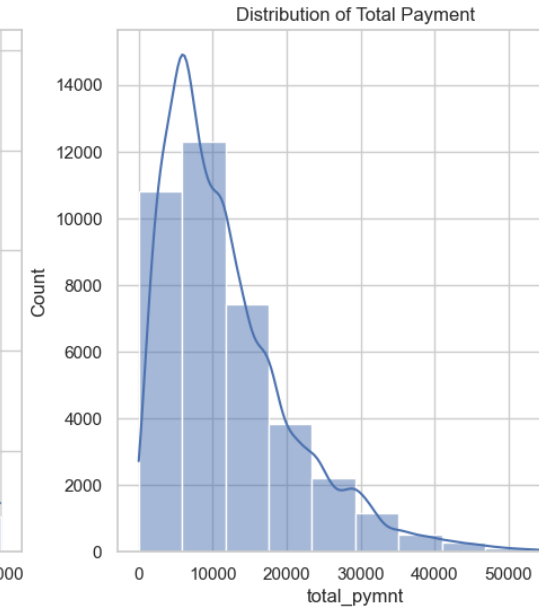
a)Univariate Analysis

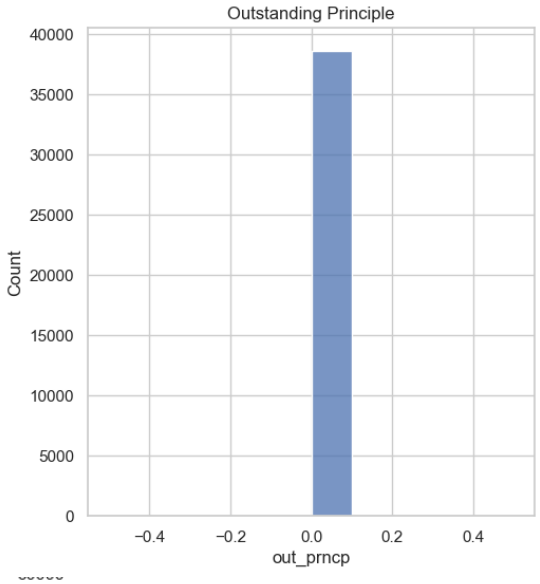
In order to arrive at Univariate Analysis, we picked up each relevant column and created the histograms using subplots. The histograms can be seen for the following attributes and the subplots in the subsequent slides:

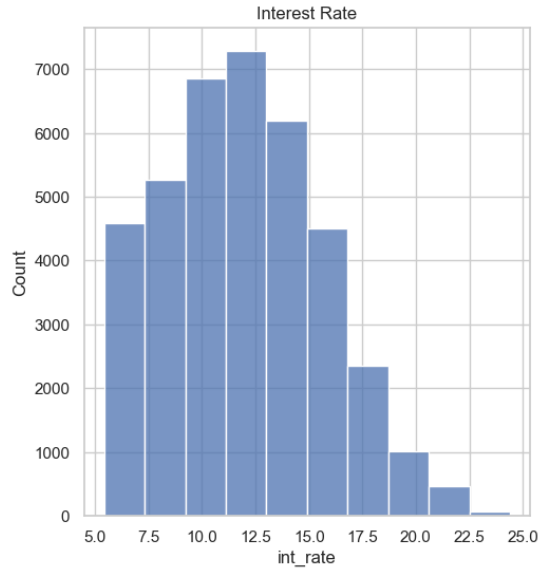
1. loan_status
2. loan_amnt
3. total_paymnt
4. out_prncp
5. int_rate
6. Home_ownership
7. Verification_status
8. grade

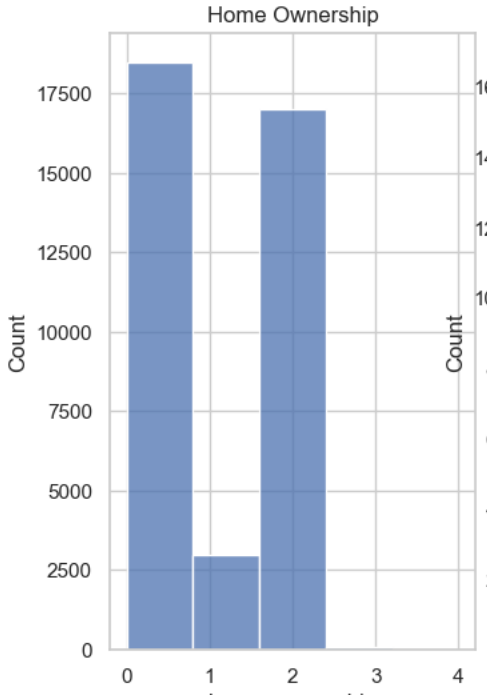
Univariate Analysis

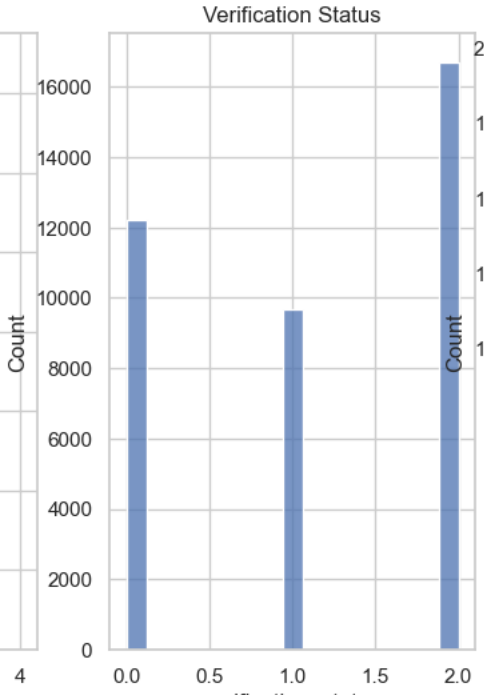


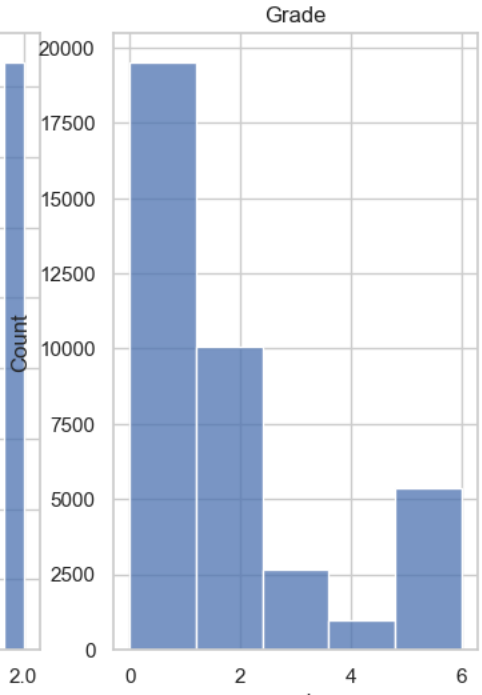








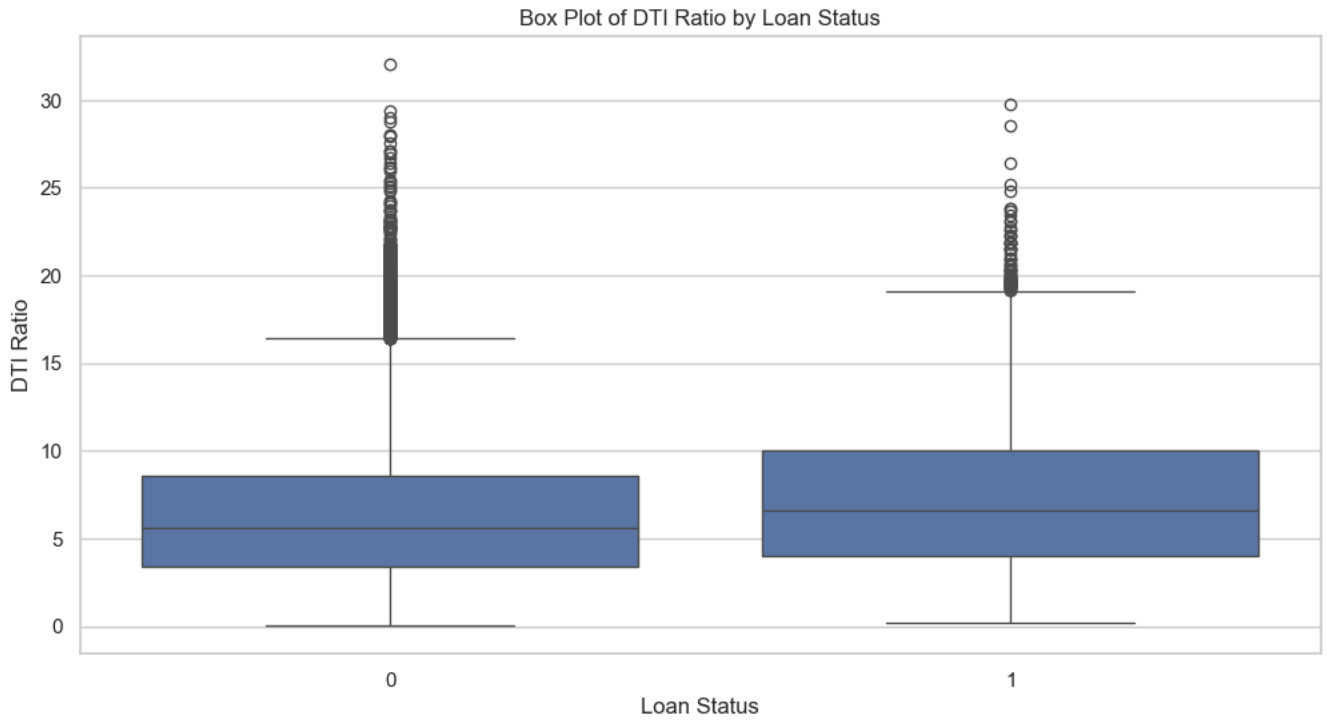




Derived Metrics

- It was important to understand the Debt to Income Ratio (DTI) , which can give some inference on the defaulters.

- We arrived at the new column by calculating the ratio of installments against the derived monthly income.

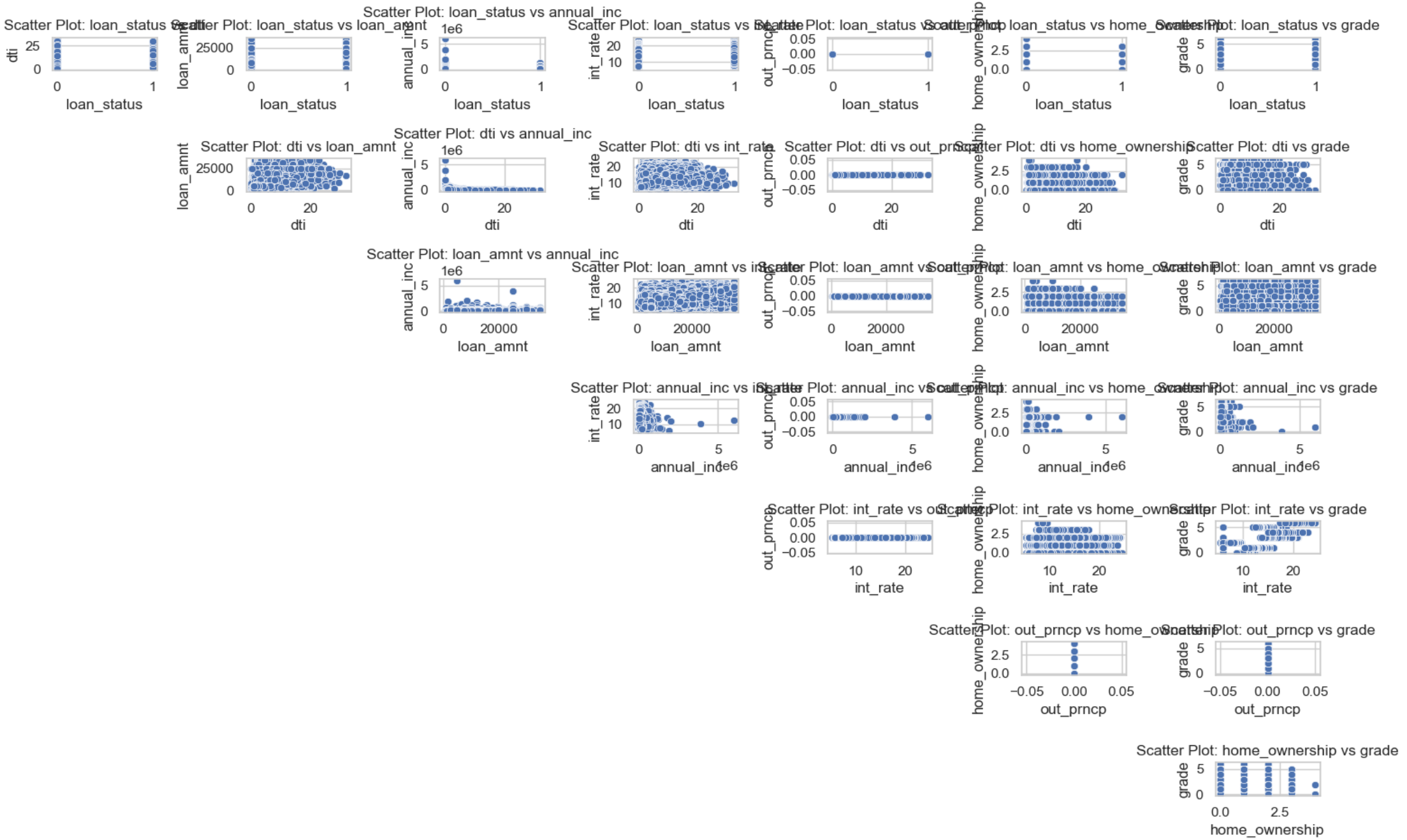


b)Bivariate Analysis

It is important to understand the relationship between the attributes, to know how the attributes are behaving one on the other. Bivariate Analysis, helps us the same. We have considered drawing subplots for the attributes mentioned below.

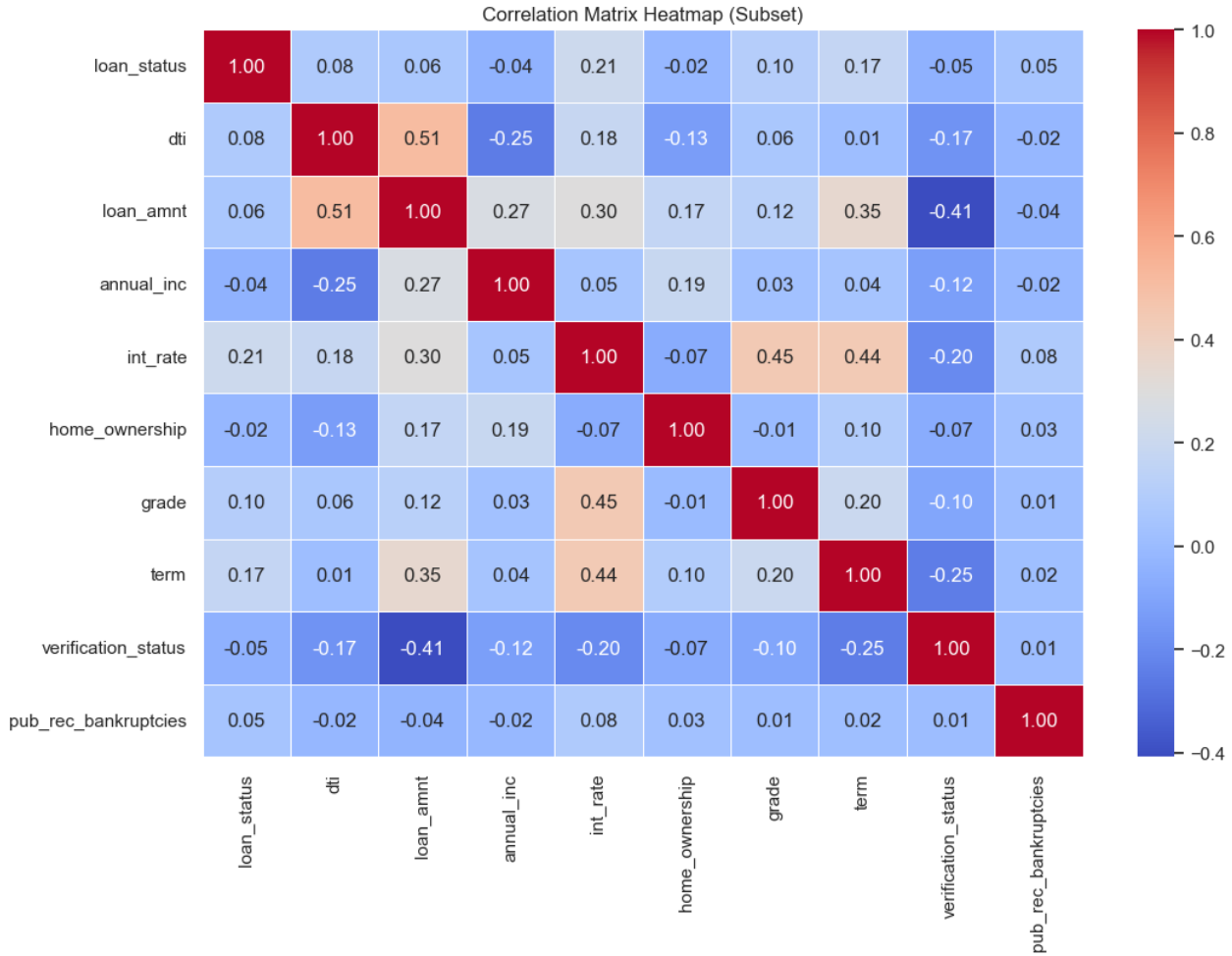
```
‘loan_status‘,'dti', 'loan_amnt', 'annual_inc', 'int_rate',
'out_prncp', 'home_ownership', 'grade'
```

Bivariate Analysis plots



c)Multivariate analysis

Finally, we derived the heatmap to understand how Reds and blues are and shows the correlation between the features. Higher blue color intensity shows low correlation while high red color intensity shows high correlation.



Conclusions

1. **Interest Rate (int_rate):** Higher interest rates are linked to an increased likelihood of default.
2. **Grade:** Lower loan grades (indicating higher risk) are associated with more defaults.
3. **Home ownership (home_ownership)** status has little impact on these decisions
4. **Outstanding Principal (out_prncp):** Higher outstanding principal correlates with a greater risk of default.

Derived metrics:

1. **Debt-to-Income Ratio (dti):** Borrowers with higher debt-to-income ratios are more prone to default.

Result:

- The Finance company need to look at the DTI and if it has higher ratio, then the bank should think before approving loans or increase interest rates to borrower, as it likely a business loss and could be a defaulter.

- Other way, the company should check, if the DTI is low, the loan is likely to be fully paid, which should actually approve the loan to increase the revenue benefit.