EDA ON LENDING CLUB CASE STUDY

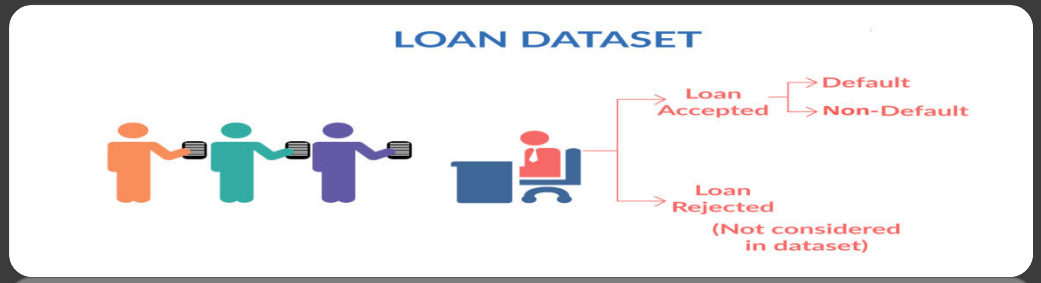
Presented by : Parakh Agarwal

PROBLEM STATEMENT

You work for a **consumer finance company** that specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to decide on 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** for 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

The dataset given contains information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns that indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc.



In this case study, you will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

When a person applies for a loan, two types of decisions 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, 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 can 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 utilize this knowledge for its portfolio and risk assessment.

STEPS PERFORMED

- Data cleaning
- Data standardization
- Univariate Analysis
- Bivariate Analysis

Data Cleaning

- The initial dataset consists of 39717 rows and 111 columns.
- 57 Columns with more than 40 % BLANK/NA/null values were dropped from the dataset
- 10 columns with only 0. and NA values were also dropped.
- 3 columns *id* , *member_id and url* with all unique values were dropped.
- Column application_type and *policy_code* which had only a single unique value was dropped.
- Rows with *loan_status* value as *Current* was also dropped as this had no importance for our analysis.
- The cleaned dataset had 38577 rows and 44 columns.

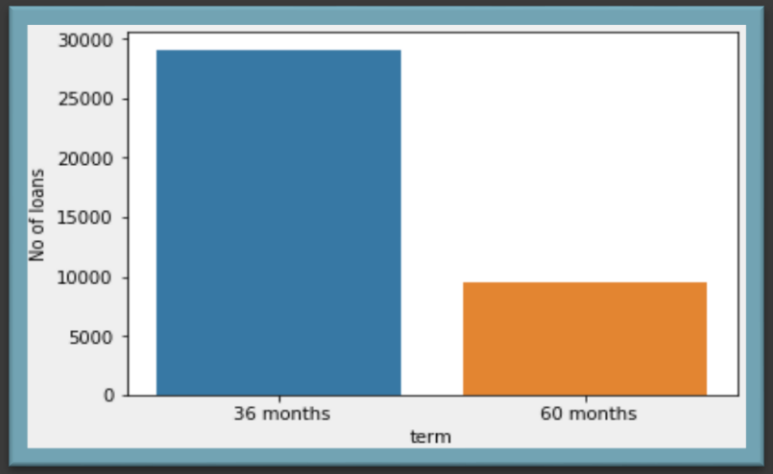
Data Standardization

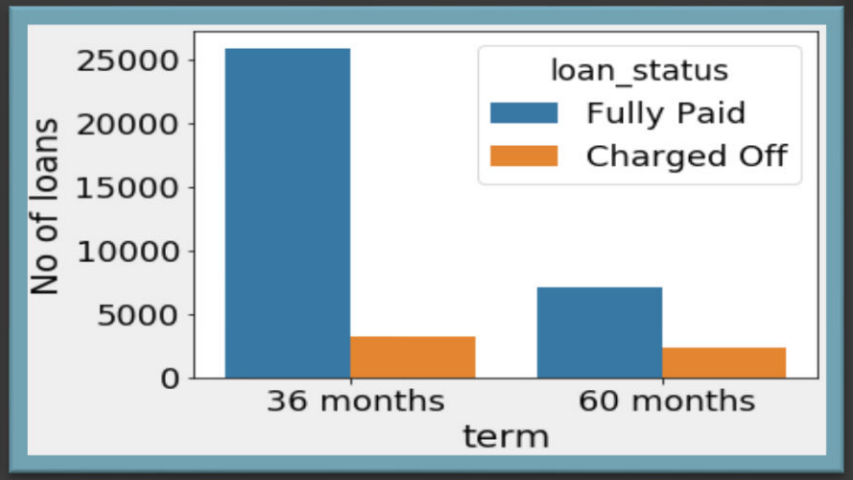
- % character was removed from the int_rate column and renamed as int_rate_percent.
- issue_d column was split into year and month columns.

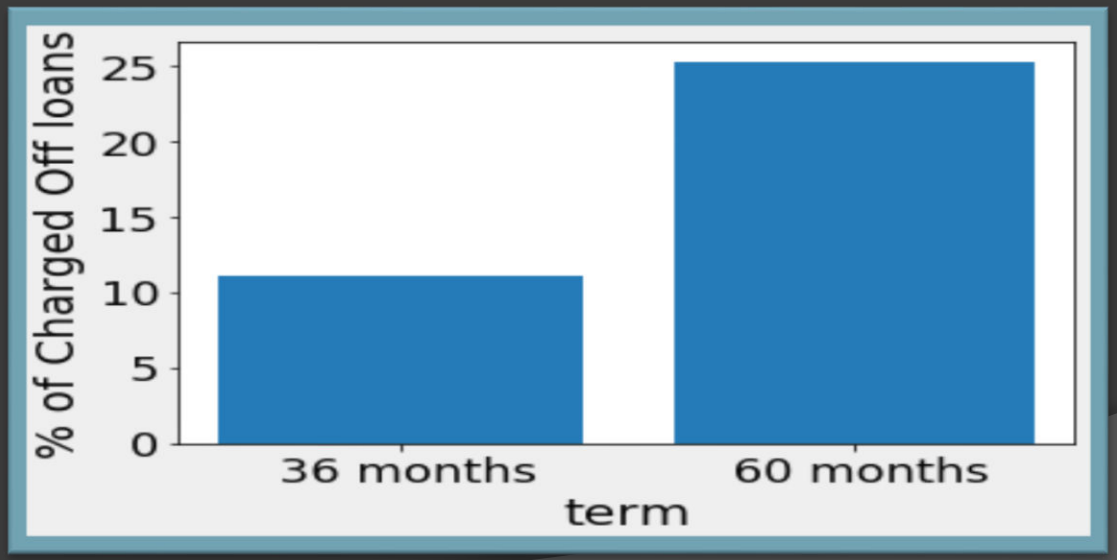
Univariate Analysis

- Univariate analysis was done on columns term, purpose, issue_d loan_amnt , annual_inc ,int_rate , installment , emp_length , home_ownership , verification_status ,dti , year , month to name a few.
- Numerical, Categorical and Derived variables were used in the analysis.
- Analysis was also done across various segments.

Variable: term
Duration for which loan was given

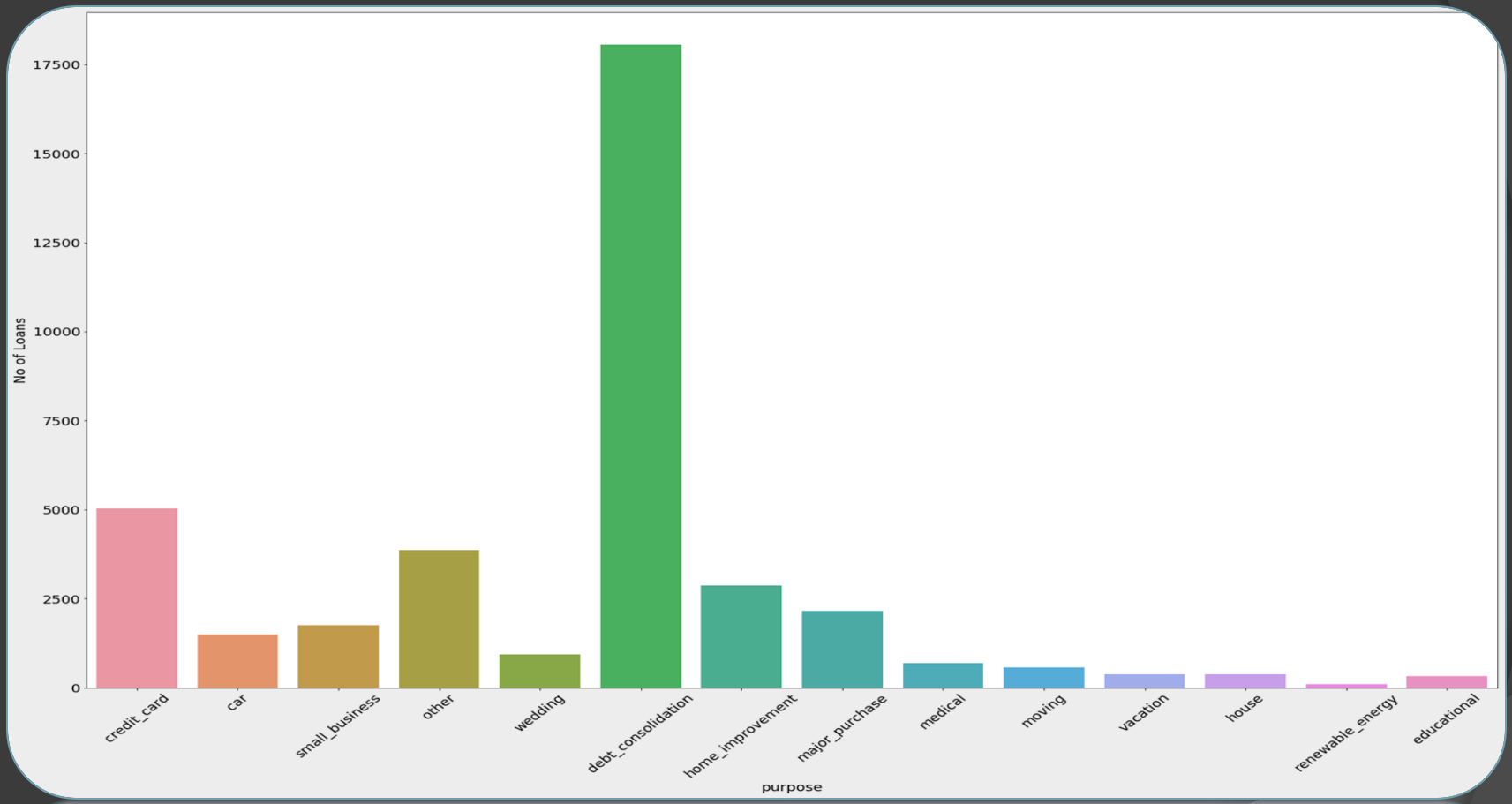


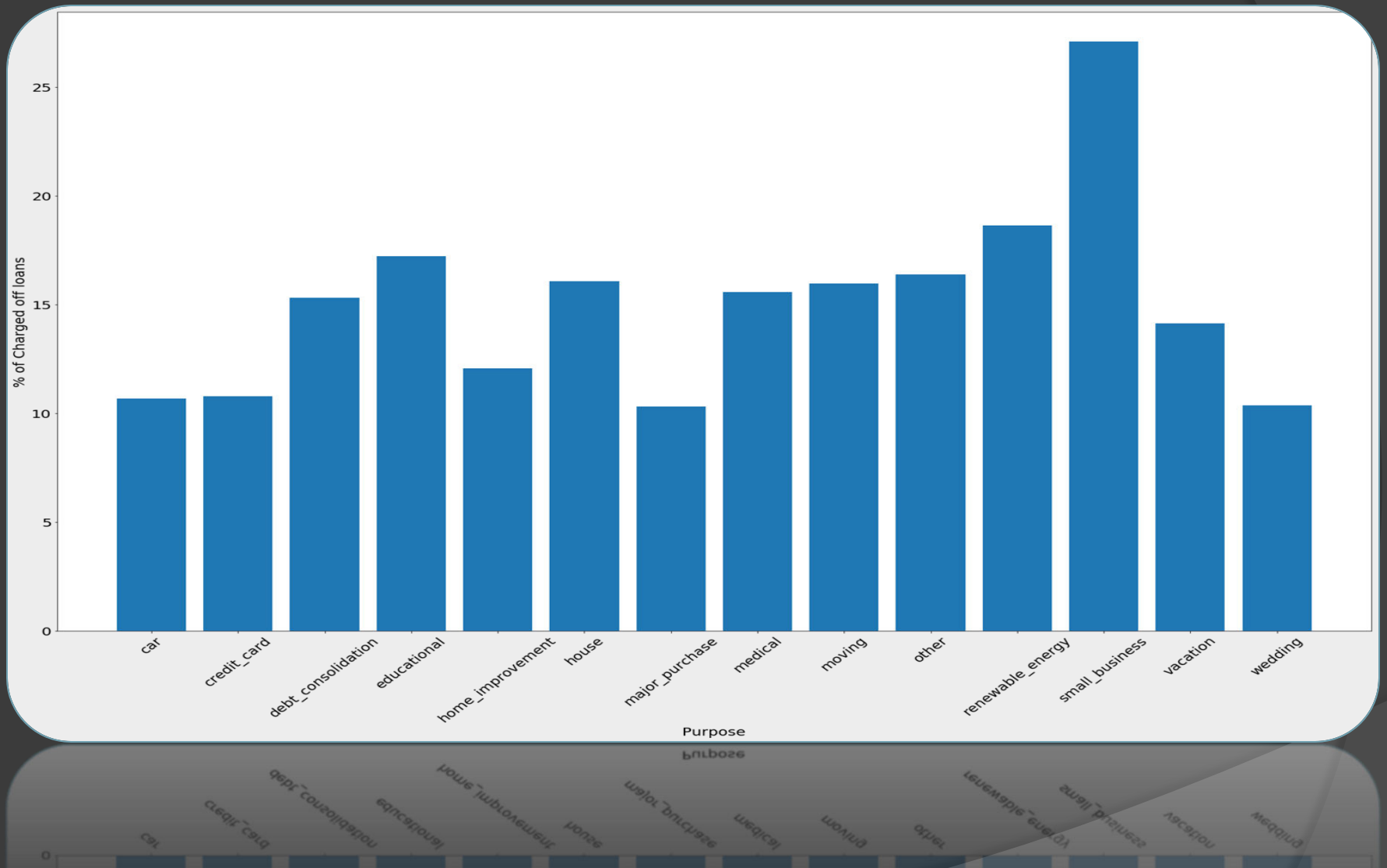




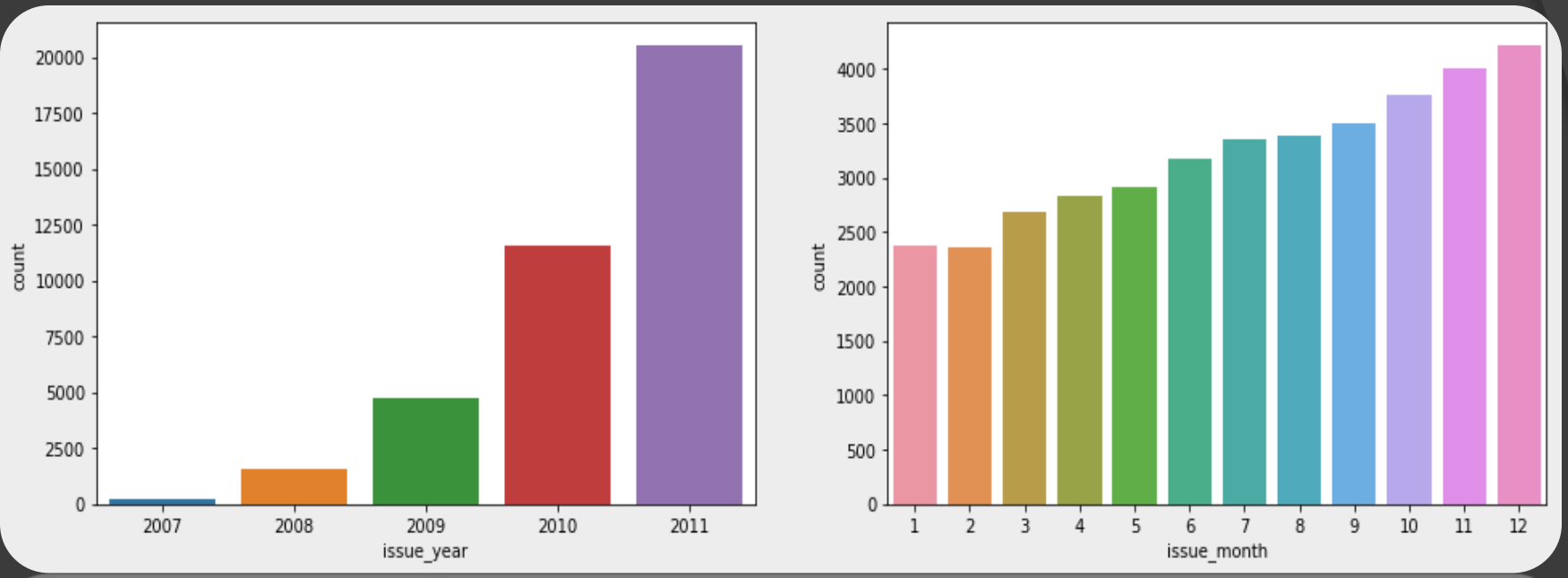
Variable: **Purpose**

The purpose for which the loan was taken

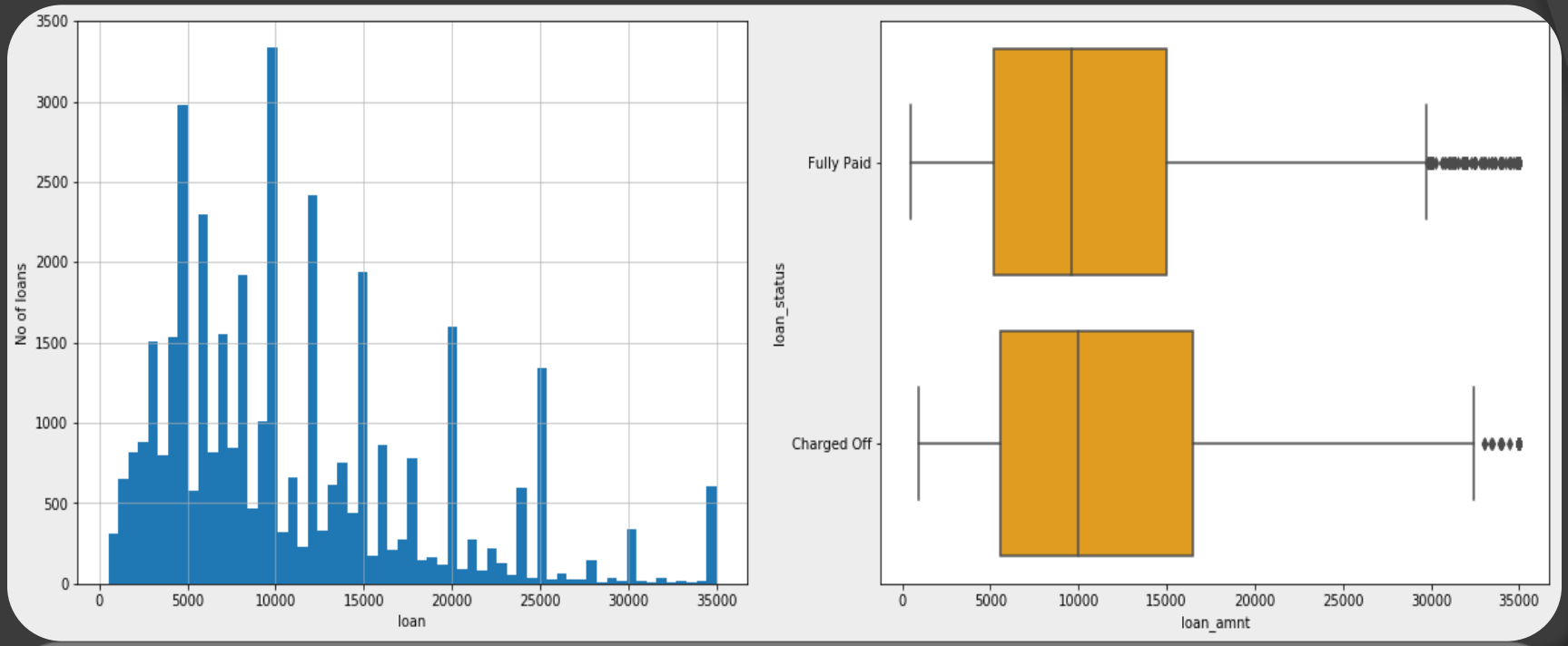


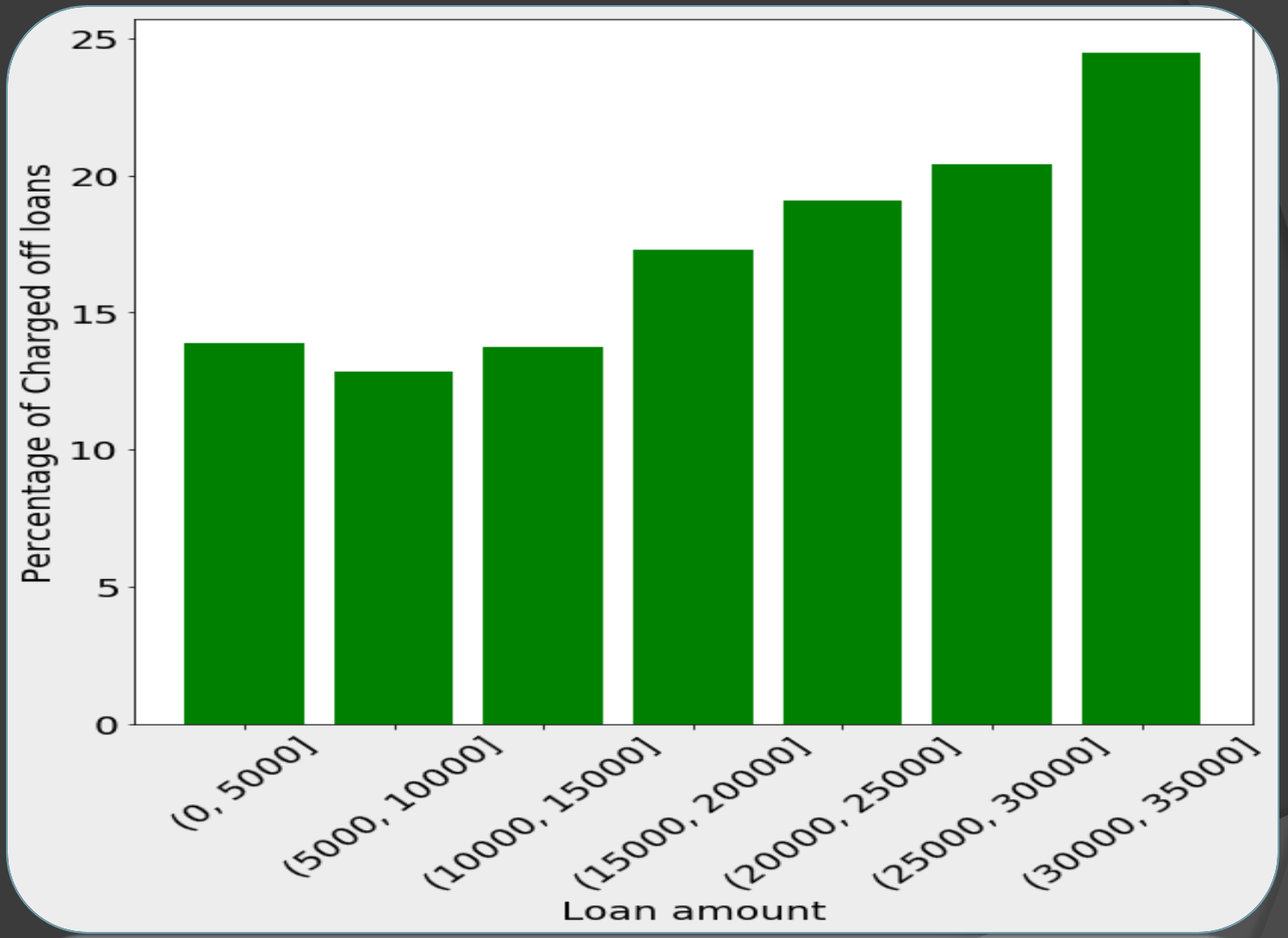

% of Charged off loans
0
5
10
15
20
25
car
credit_card
debt_consolidation
educational
home_improvement
house
major_purchase
medical
moving
other
renewable_energy
small_business
vacation
wedding
Purpose

Variable: Issue_d

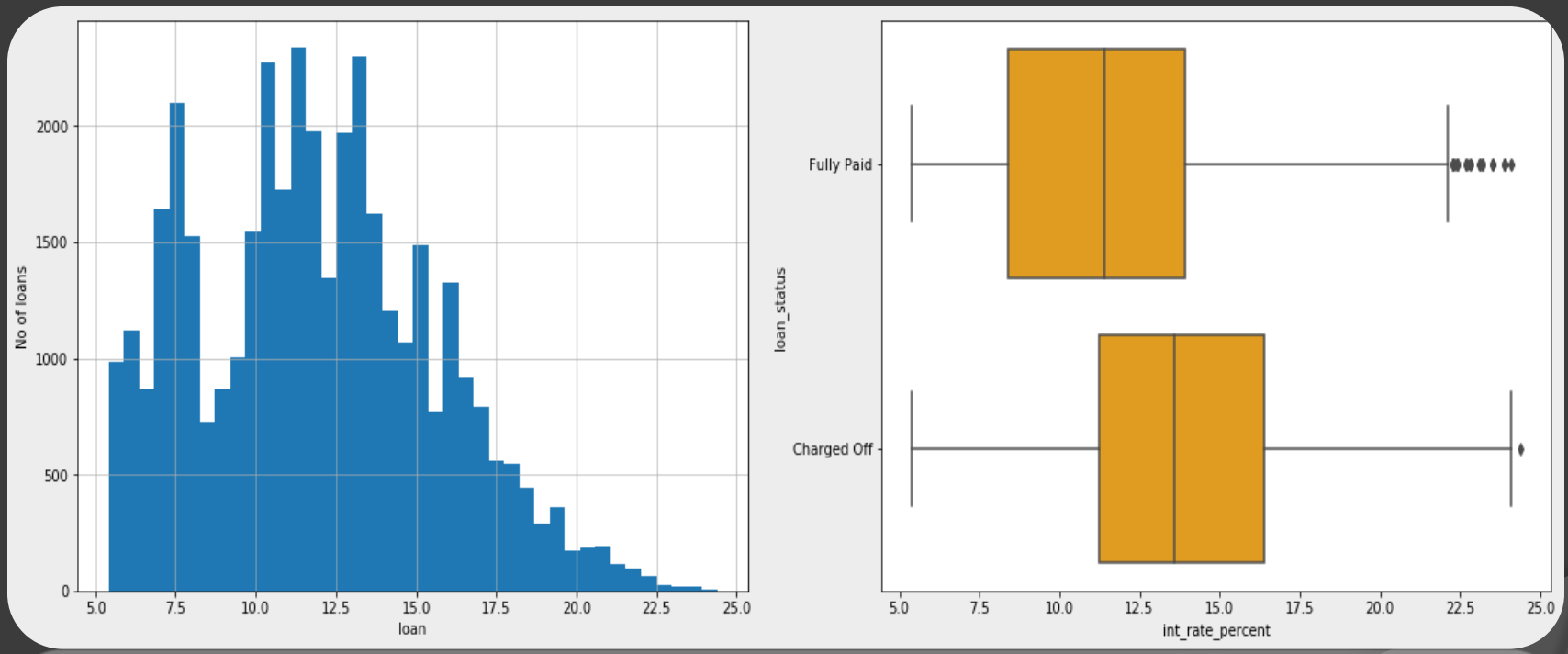


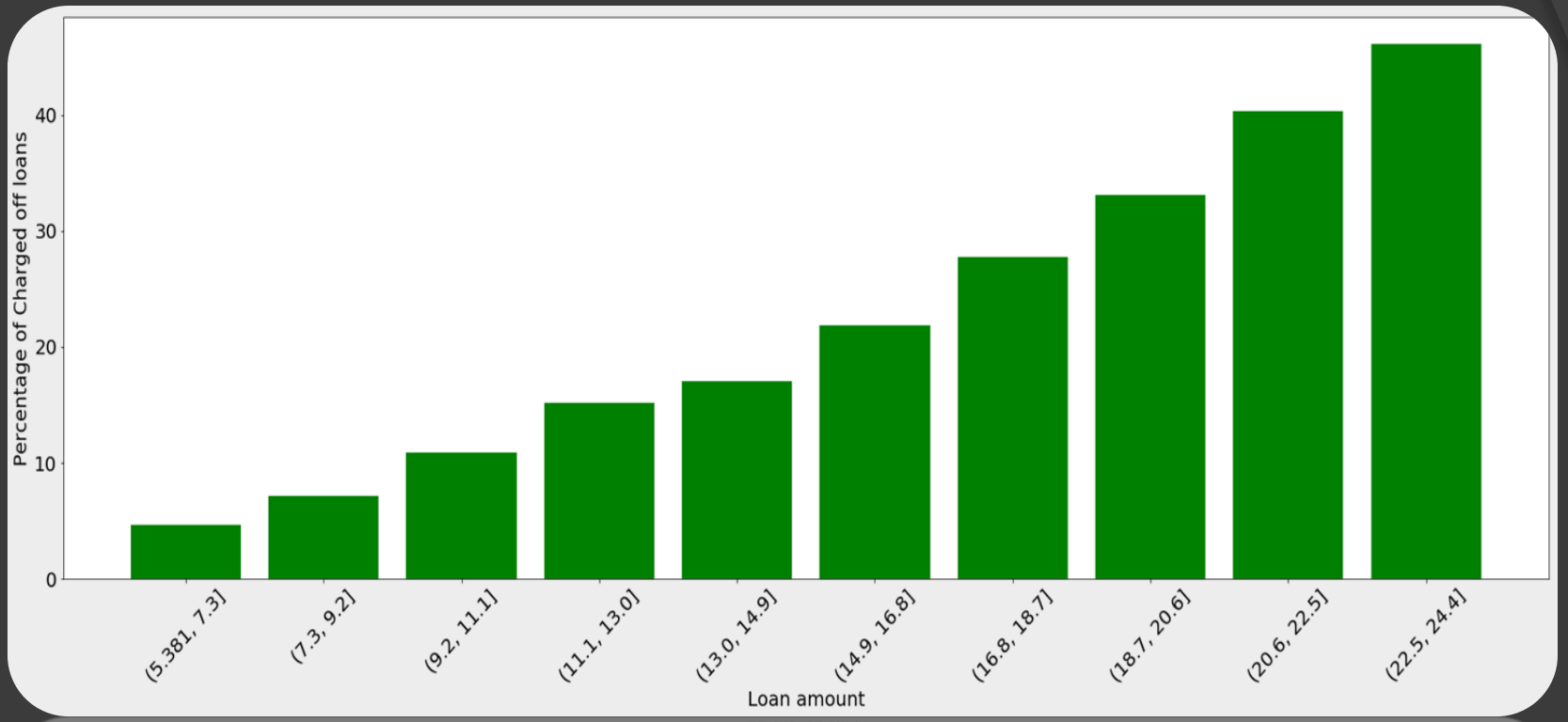
Variable: loan_amt



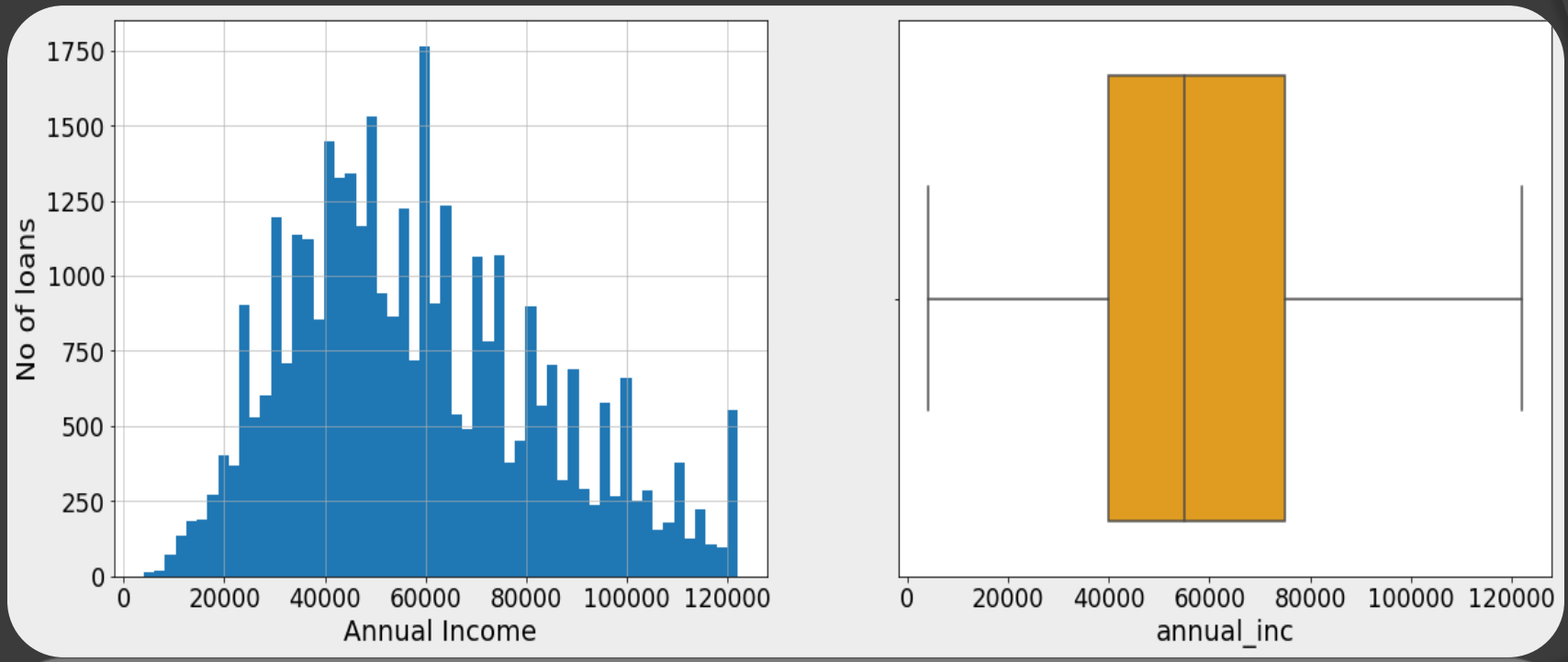

Percentage of Charged off loans
25
20
15
10
5
0
(0, 5000]
(5000, 10000]
(10000, 15000]
(15000, 20000]
(20000, 25000]
(25000, 30000]
(30000, 35000]
Loan amount

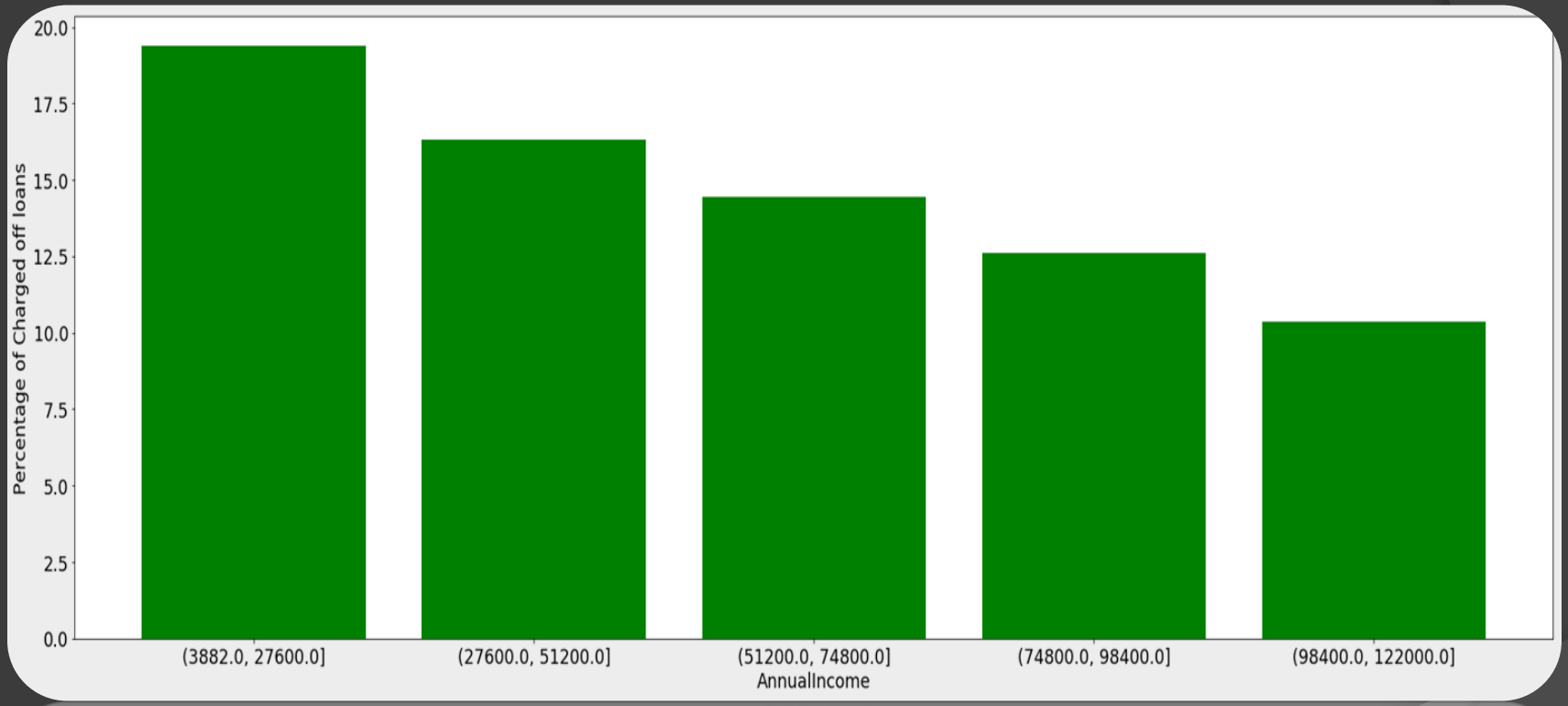
Variable: int_rate_percent
The interest rate at which the loan was borrowed




Percentage of Charged off loans
0
10
20
30
40
(5.381, 7.3]
(7.3, 9.2]
(9.2, 11.1]
(11.1, 13.0]
(13.0, 14.9]
(14.9, 16.8]
(16.8, 18.7]
(18.7, 20.6]
(20.6, 22.5]
(22.5, 24.4]
Loan amount

Variable: annual_inc




20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Percentage of Charged off loans
(3882.0, 27600.0]
(27600.0, 51200.0]
(51200.0, 74800.0]
(74800.0, 98400.0]
(98400.0, 122000.0]
AnnualIncome

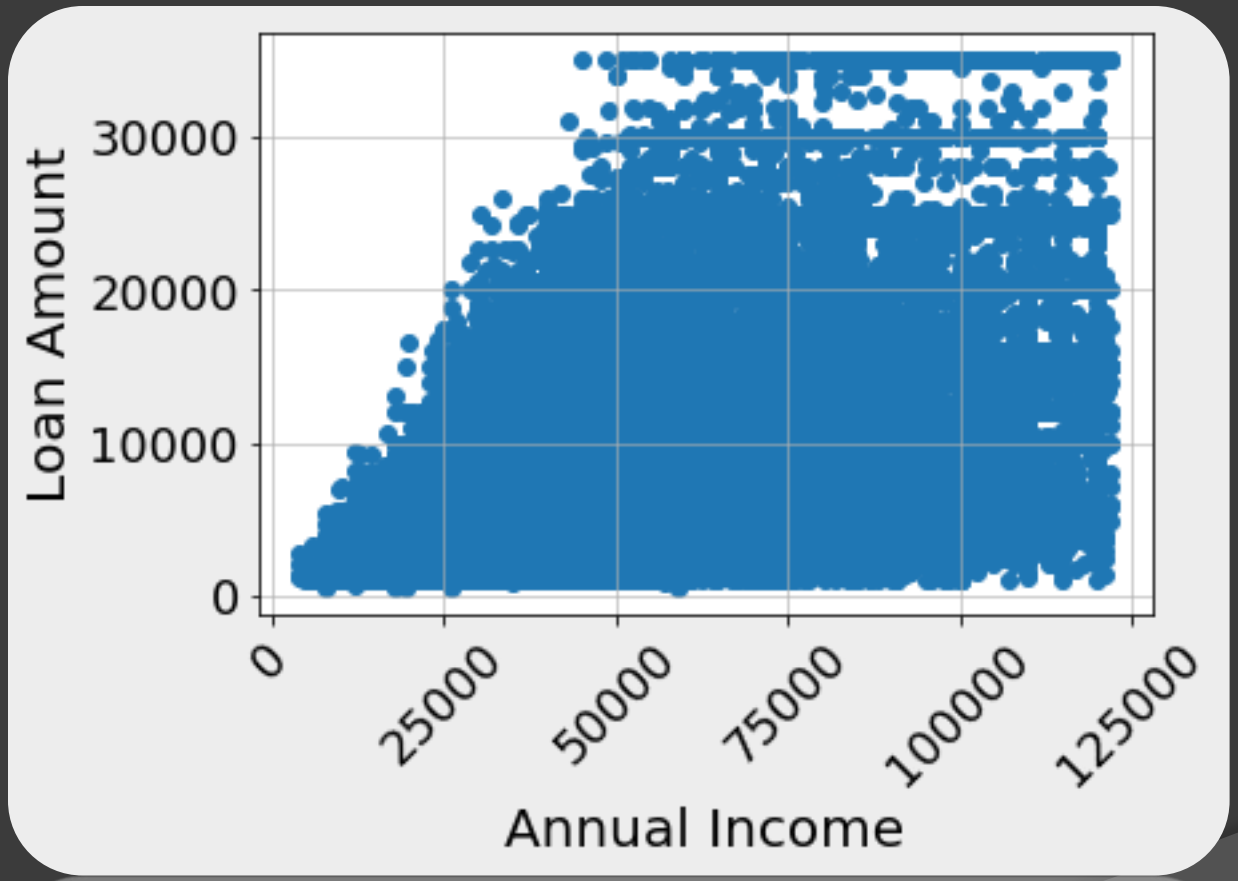
Results and observations:

- Most loans are of term 36 months.
- Loans of tenure 60 months have a higher default rate.
- Most loans were taken for debt consolidation
- Small businesses had the highest default rate among all other purposes.
- A maximum no of loans were taken in the year 2011 and the month of December.
- Borrowers who have higher loan amounts are more likely to default.
- Borrowers who have lower annual incomes are more likely to default.
- Borrowers with higher interest rates are more likely to default.

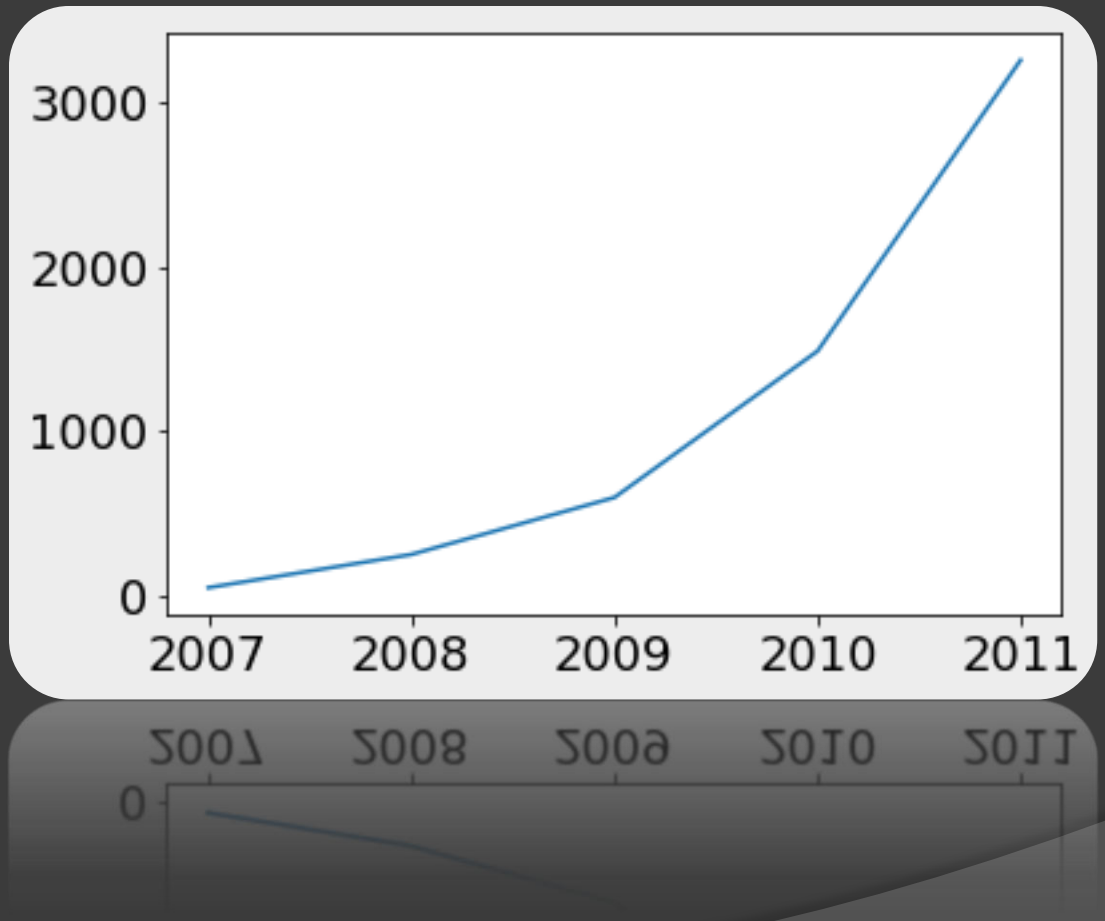
Bivariate analysis :

- Analysis was done for two or more variables.
- Visualizations used were box plots, simple line graphs, heat maps and scatter plots.
- It was done for both categorical and numerical variables.

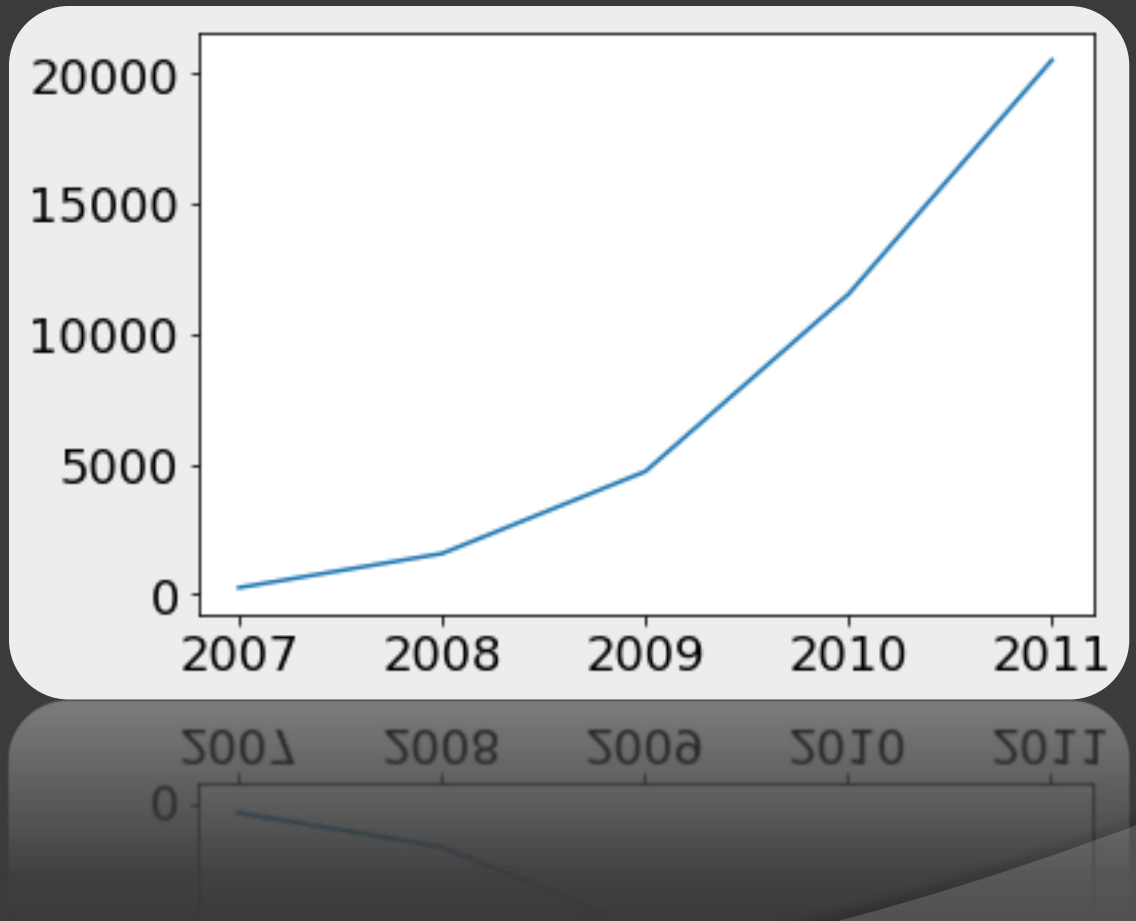
annual_inc vs loan_amnt



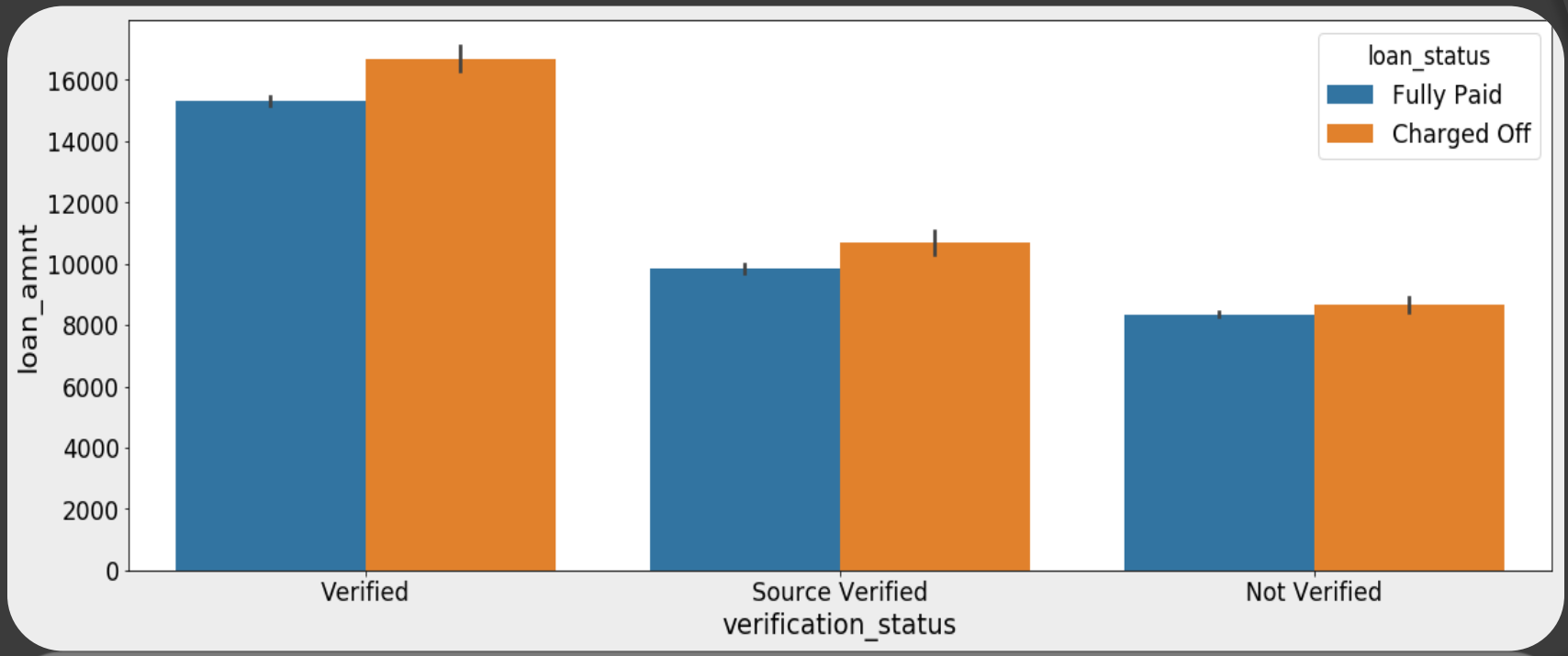
Year vs No. Of loans Charged Off



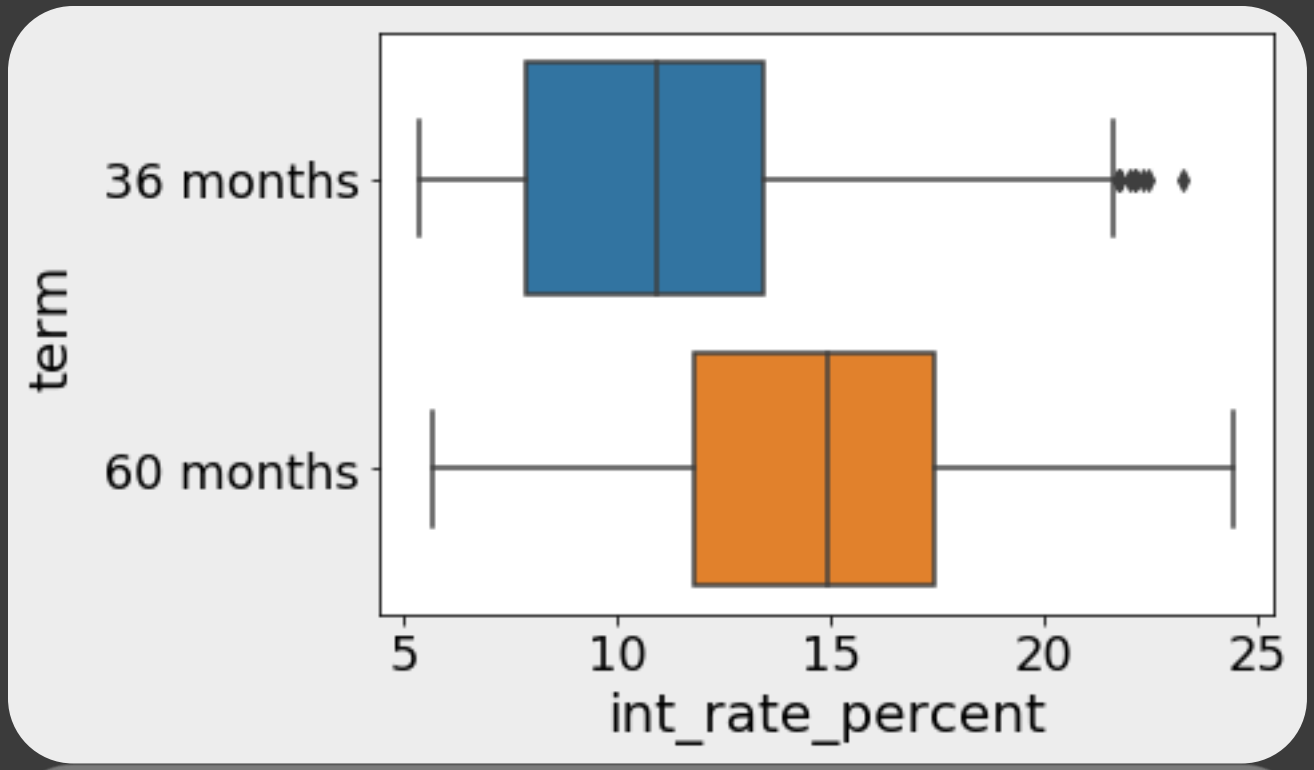
No. of loans over time



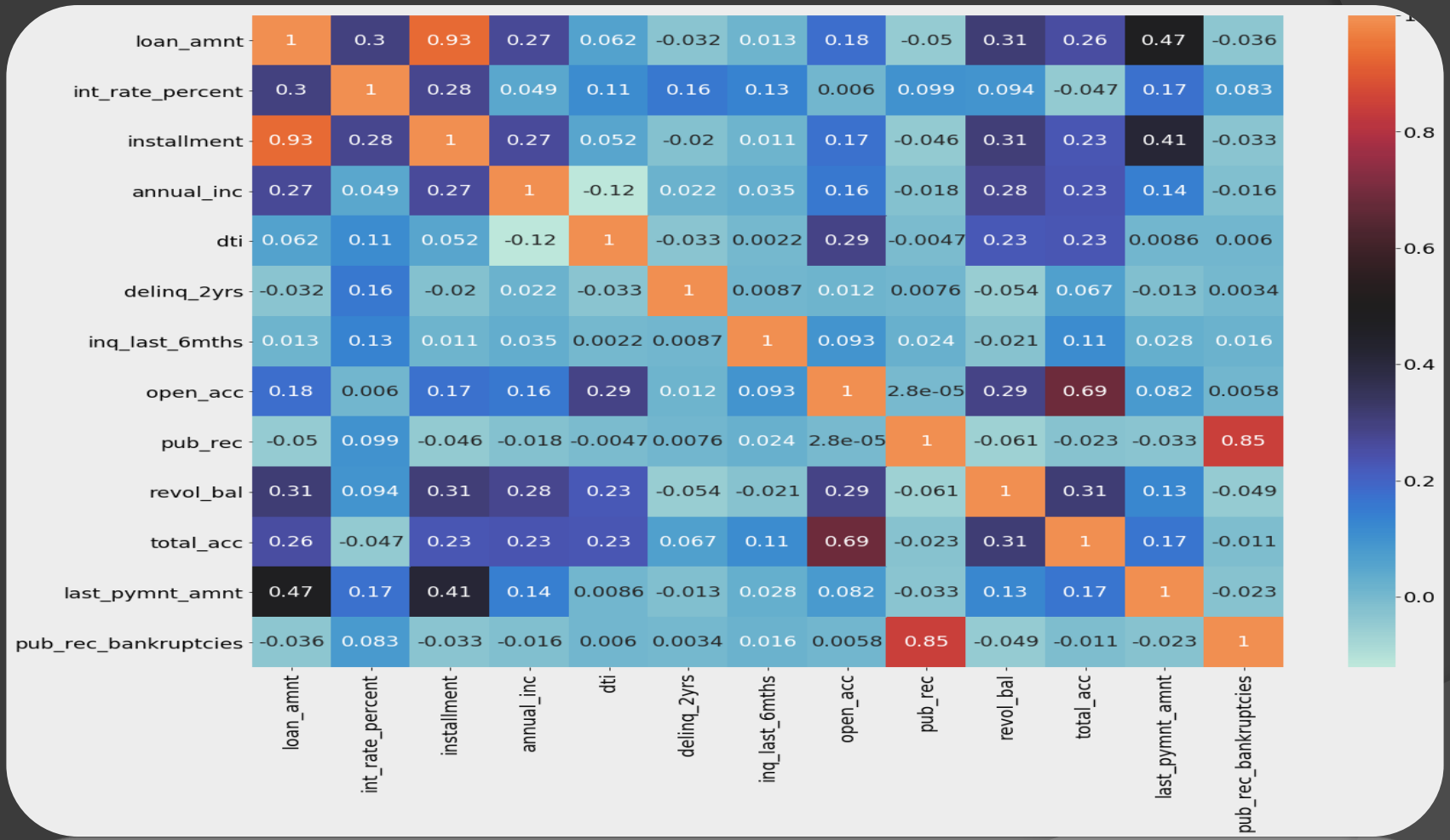
loan_amnt vs verification_status vs loan status



Interest rate and term



Correlation Matrix Heat Map



SUMMARY

- 1)There are people with an income of 50000 or less and a loan amount of 25000 or more. These will be risky loans as we have previously seen as higher loan amount increases the chances of defaulter and higher income reduces it. Hence people with less income and higher loans are very risky.
- Greater the loan amount greater the chance loan being verified
- We already know that larger loans are less in number, but see a higher charge-off rate. This explains why verified loans see a higher rate of default. It's the fact that higher loan amounts are riskier and are also verified more often by Lending Club
- Interest rates are higher for higher terms of loan and a higher loan term means a high chance of being a defaulter.
- Int_rate is correlated to revol_util with an r factor of .47 – This is good, as the company is charging higher interest from the riskier loan.
- loan_amnt is correlated to last_payment_amount with r factor.44, as expected.
- loan_amnt revol_bal are correlated with r factor .35 - This is not good as it suggests that a higher loan amount is being approved to riskier borrowers.

THANK YOU