Lending Club

Risk Analysis

Introduction

Lending club is a **consumer finance company** which specializes 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

Decisioning

- **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

- Lending club 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'.
- The company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default using EDA is the aim of our case study .

Tasks Performed

1. Data Understanding
2. Data Cleaning and Manipulation
3. Data Analysis
    i. Univariate Analysis
    ii. Standard Univariate Analysis
    iii. Bivariate Analysis

Data Analysis and Cleaning

Given Loan.csv file contain 39717 rows and 111 columns

Find and delete invalid rows:

Entire data divided into 3 categories

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 .

Data with loan_status as 'Current' will not help us to predict Defaulters in data set, since we can't determine that, they would come under Fully Paid or Charged-off

So we can remove these rows from the given data set.

Validating amount fields in data

- loan_amnt : loan_amnt is the amount applied by potential borrowers
- funded_amnt : funded_amnt is the amount recommended/approved by Lending Club
- funded_amnt_inv : funded_amnt is the amount recommended/approved by Lending Club

From the above meta data we can conclude as below

Loan_amount > funded_amnt > funded_amnt_inv

- So data which are not satisfying above condition, they will be consider as invalid data.
- If we consider this invalid data, out prediction results will be inconsistent.
- So we are removing these rows as well for further analysis

Find and drop invalid and unnecessary columns:

- id, member_id, url fields are containing unique values.
- They will not helpful in prediction of defaulters. so we can remove those entire columns

- Some of the customer behavior variables are not available at the time of loan application, and thus they cannot be used as predictors for credit approval.

  Eg: pub_rec, delinq_2yrs, earliest_cr_line, inq_last_6mths etc.

  So we can drop these all columns from the data set

- Also some columns contain null for entire dataset, we can drop them as well

  Eg: total_bc_limit, total_il_high_credit_limit, total_bal_ex_mort etc.

- Some columns contain single value in entire dataset, we can drop them also

  Eg: pymnt_plan contains 'n' as value for all rows

Handling missing data:

Drop columns which contains large number of null values in them

- 'desc', 'mths_since_last_delinq','mths_since_last_record','next_pymnt_d' these columns contains huge number of null values compare to others
- So won't help us in our analysis, we can these columns also

Handling **emp_title** missing values:

- Emp_title is a unordered categorical column
- Mode of this column is U.S Army
- Since this value is not generic we can't decide to replace missing values with this value.
- Instead of filling with wrong data we should delete those rows to avoid inconsistence .

Handling **emp_length** missing values:

- Mode of emp_length is '>10 Years'
- We can't replace this value as missing value it may lead to wrong results, so we need to delete to those rows

Handling **title** missing values:

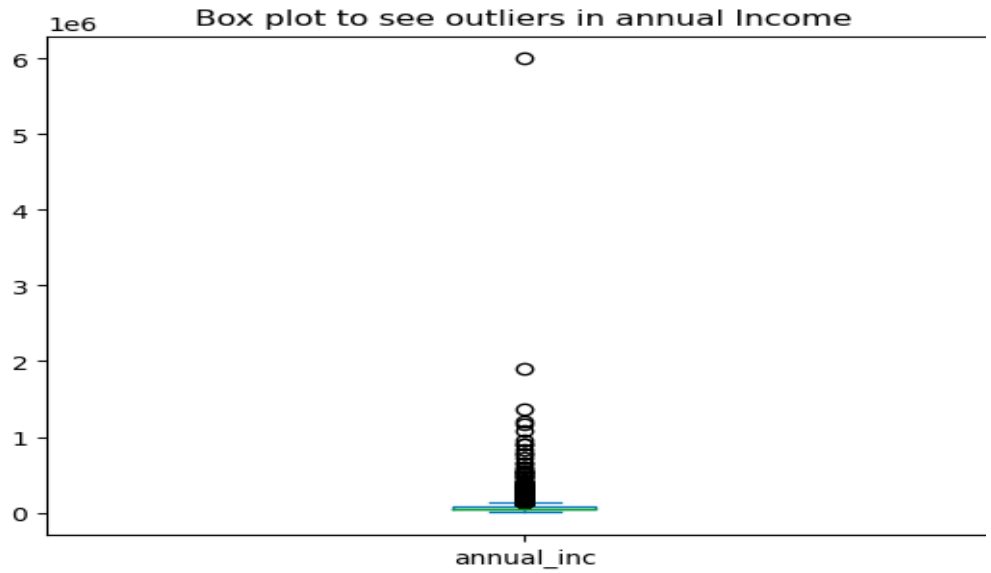
- We are replacing title value with 'Debt Consolidation' value since this value is highest frequent value compared to others
- So if we replace them with these it won't affect the prediction results

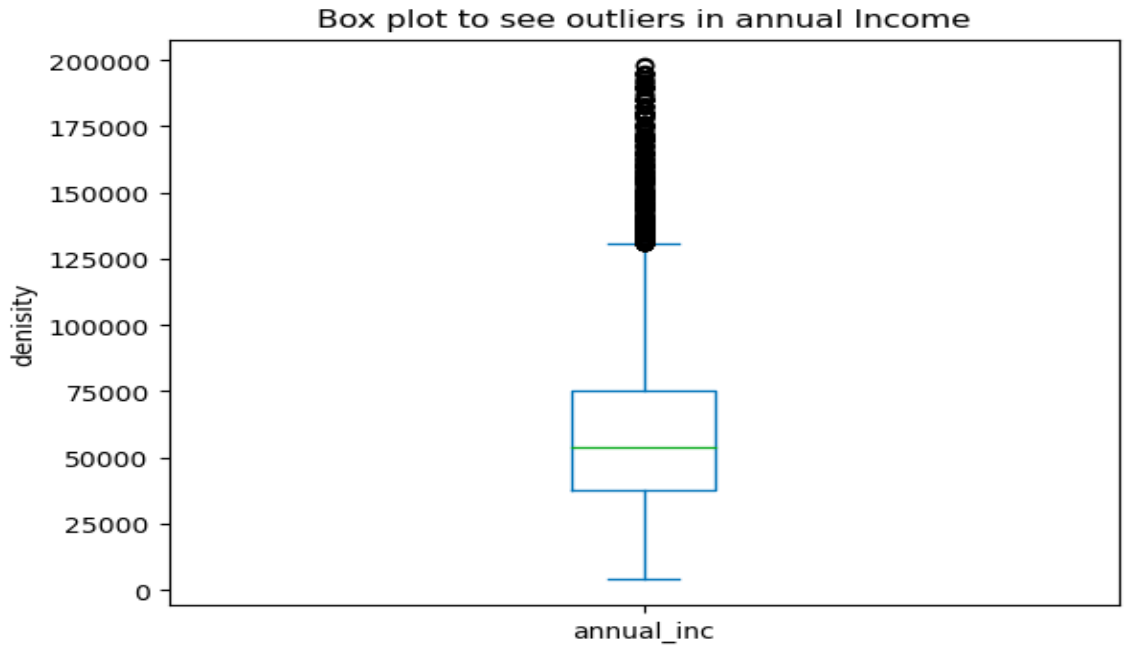
Modifying field values and Deriving new fields for analysis:

- Removed % from **int_rate** values and converted them into float values for analysis
- Removed xx from **zip_code** values and converted to integers
- Removed 'months' string in **term** values and converted into integers
- Derived new columns **issue_month, issue_year** from **issue_d** column
- Derived new columns **loan_amount_bins , int_rate_bins, annual_inc_bins, dti_bins, installment_bins, zip_code_buckets** from **loan_amnt, int_rate, annual_inc, dti, installments, zip_code** respectivly based on binning values

Checking of Outliers

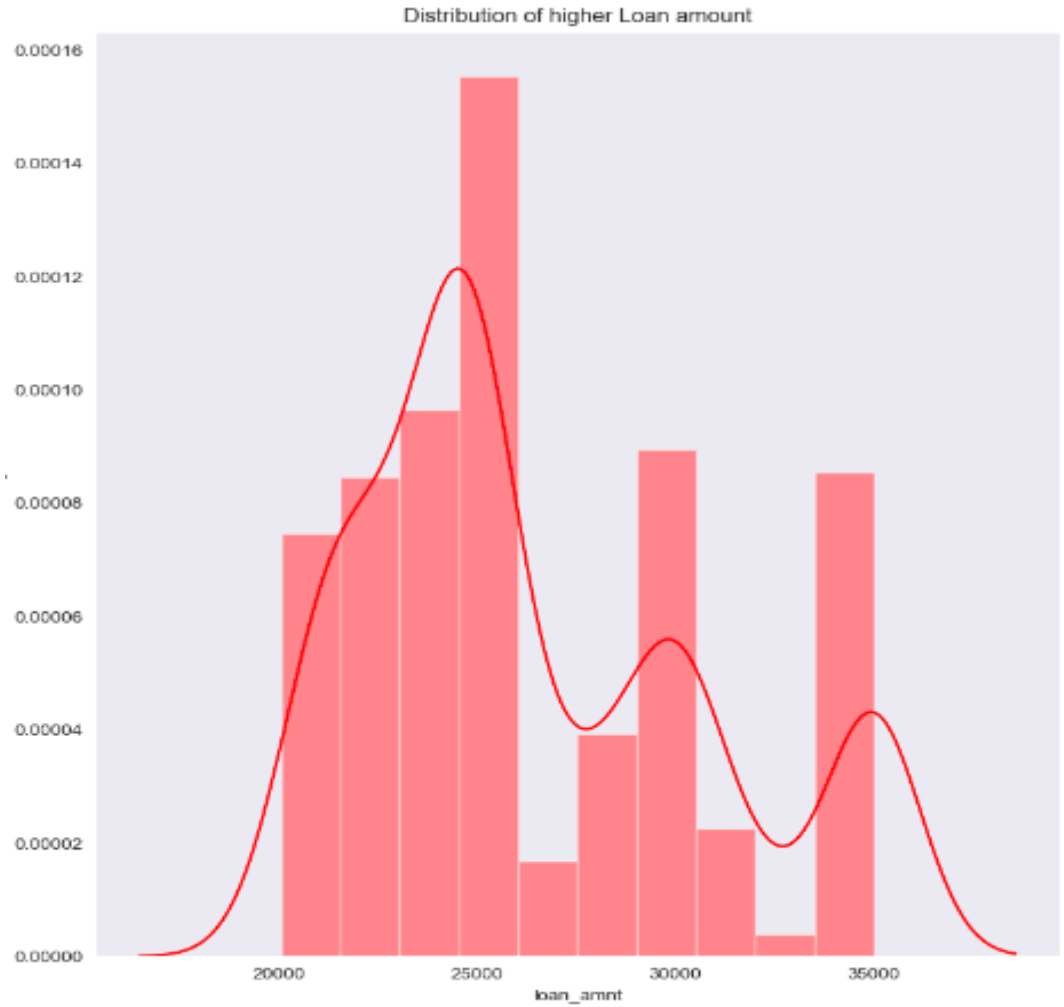
- Annual_inc column has huge outliers which are not continuous to analize defaulter.
- So filtered dataset with annual_inc values which are less than 200000
- There are 2 box plots which are before and after removing outliers from the data of annual_inc

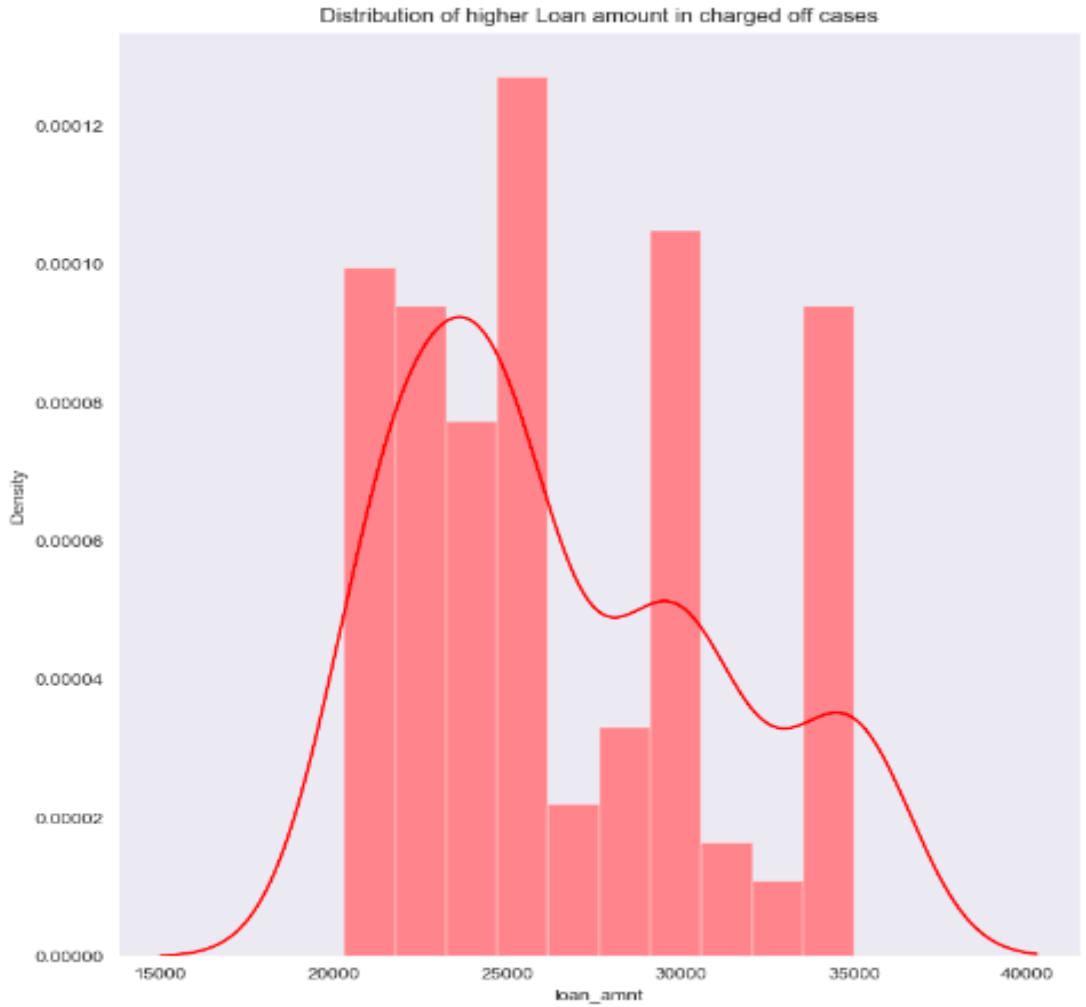




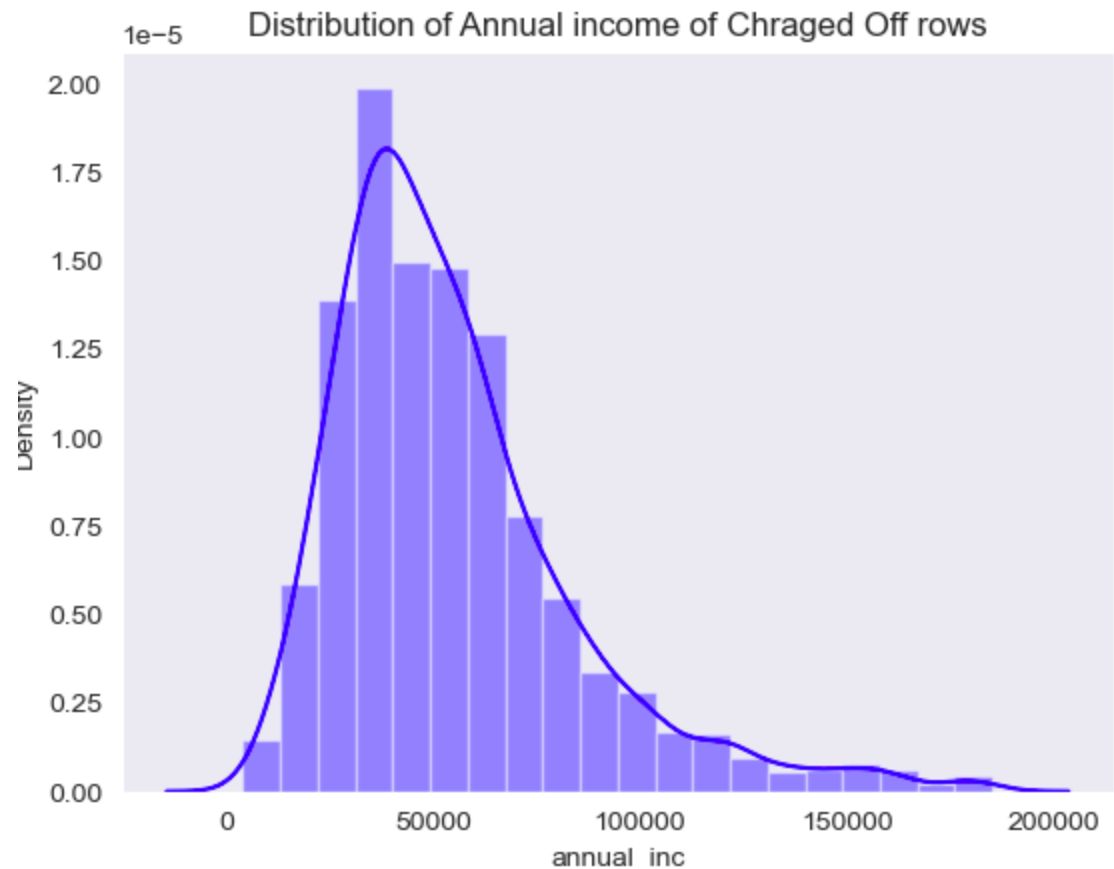
Univariate Analysis:

- Higher loan amount causes more default cases compare to lower loan_amount

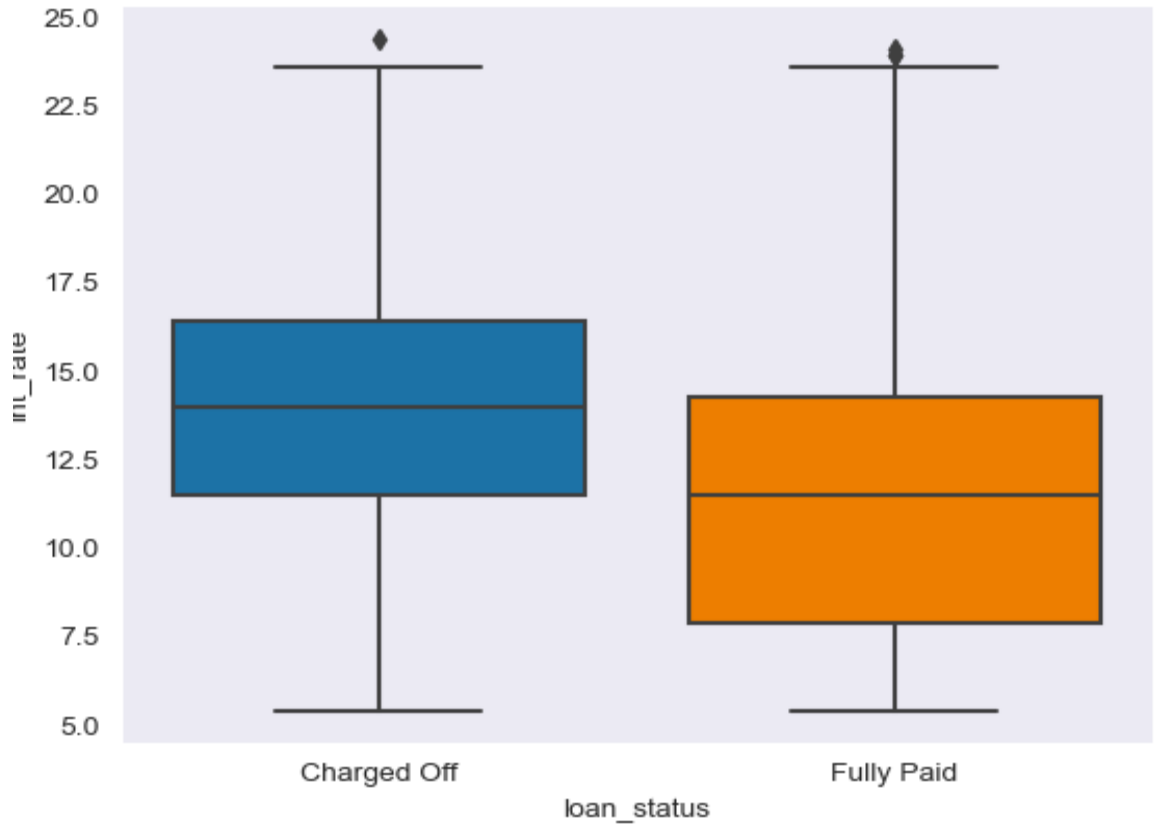




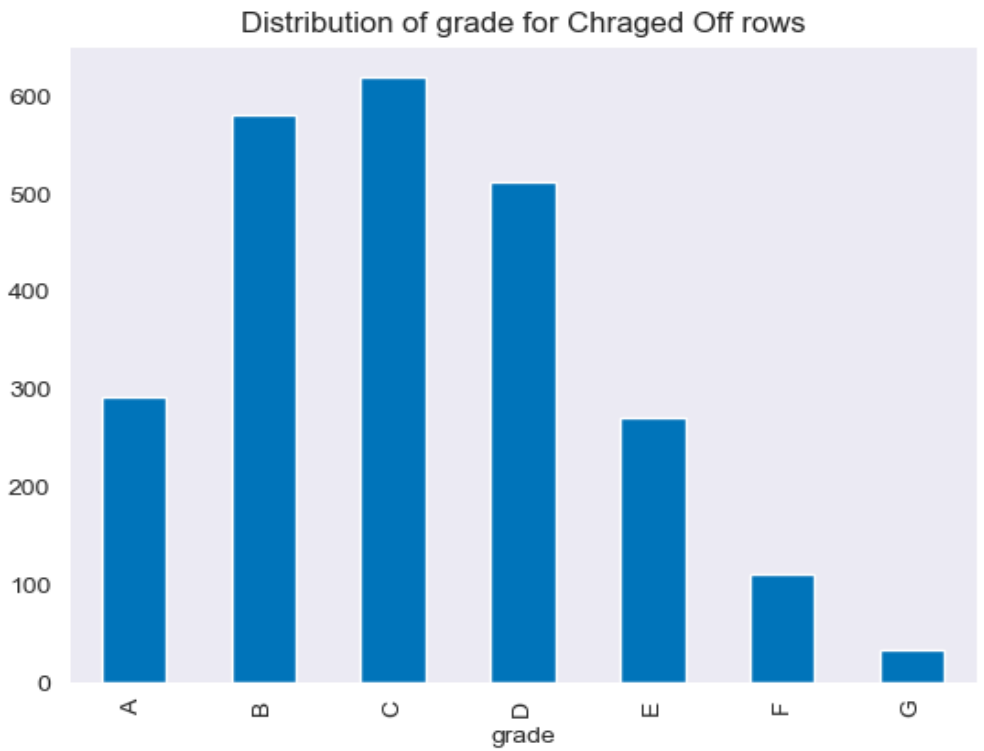
Annual Income between 35000 to 66000 more prone to charging off from their loan



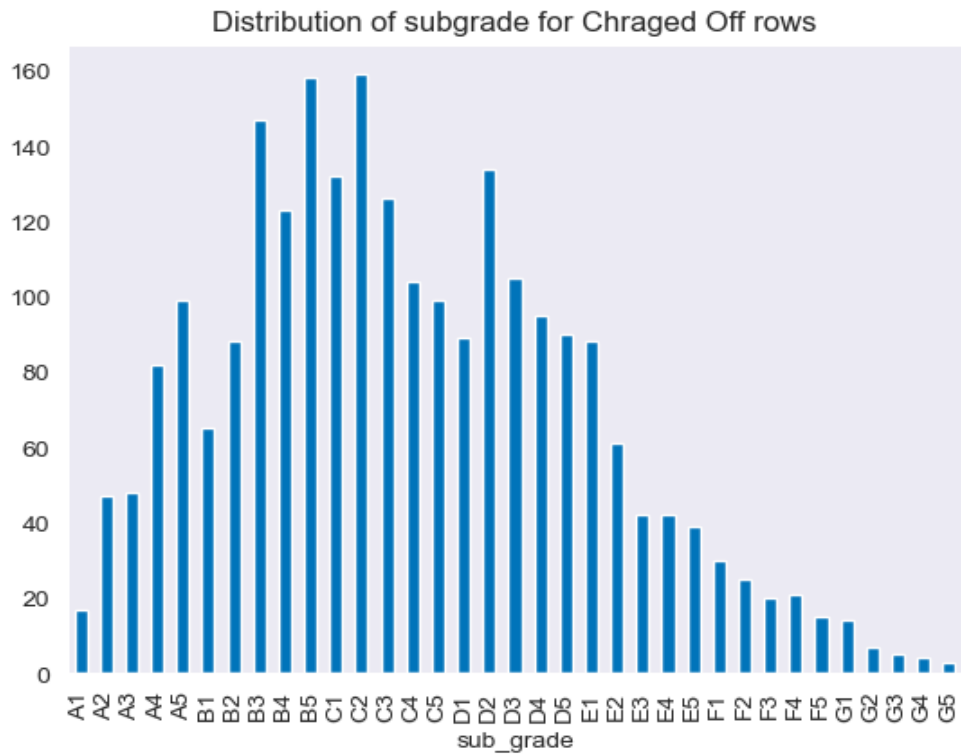
Average Interest Rate value of charged off has higher than fully paid rate



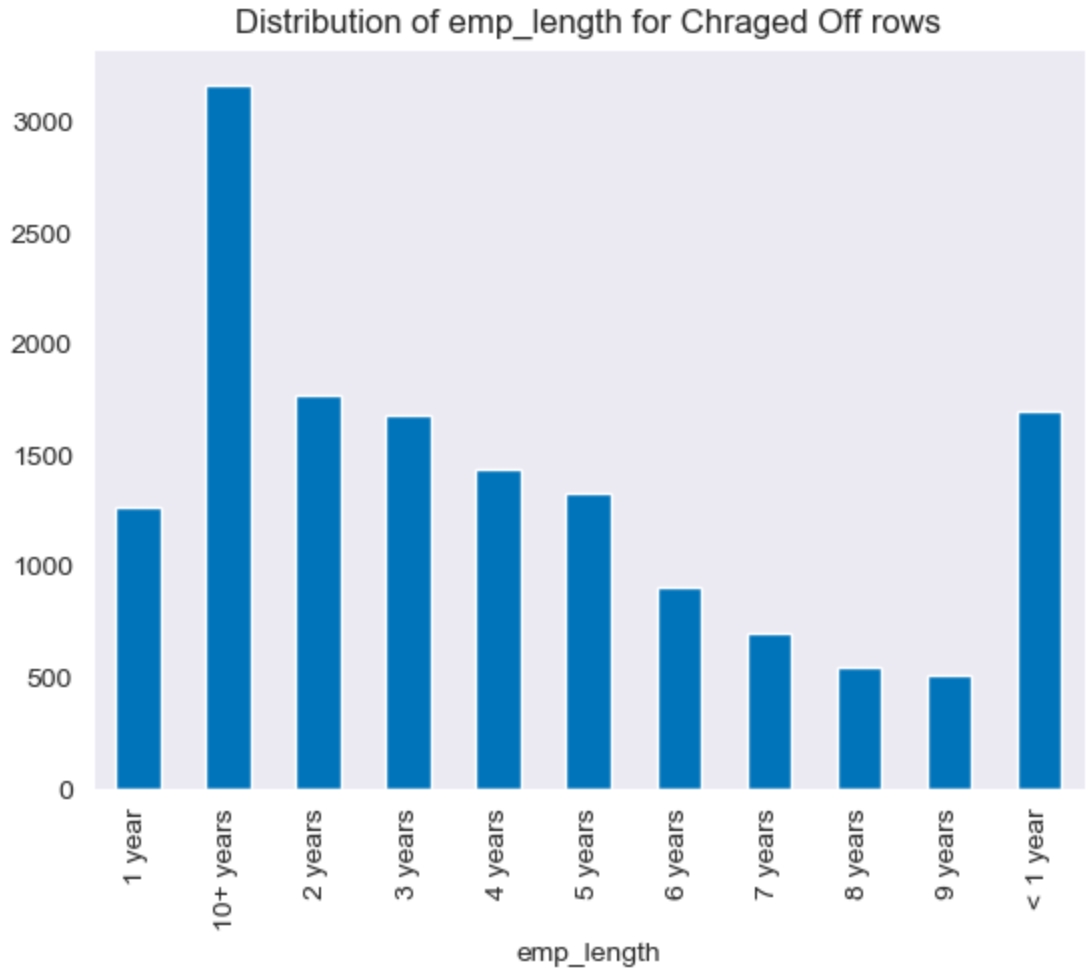
Grade C has high charged off cases



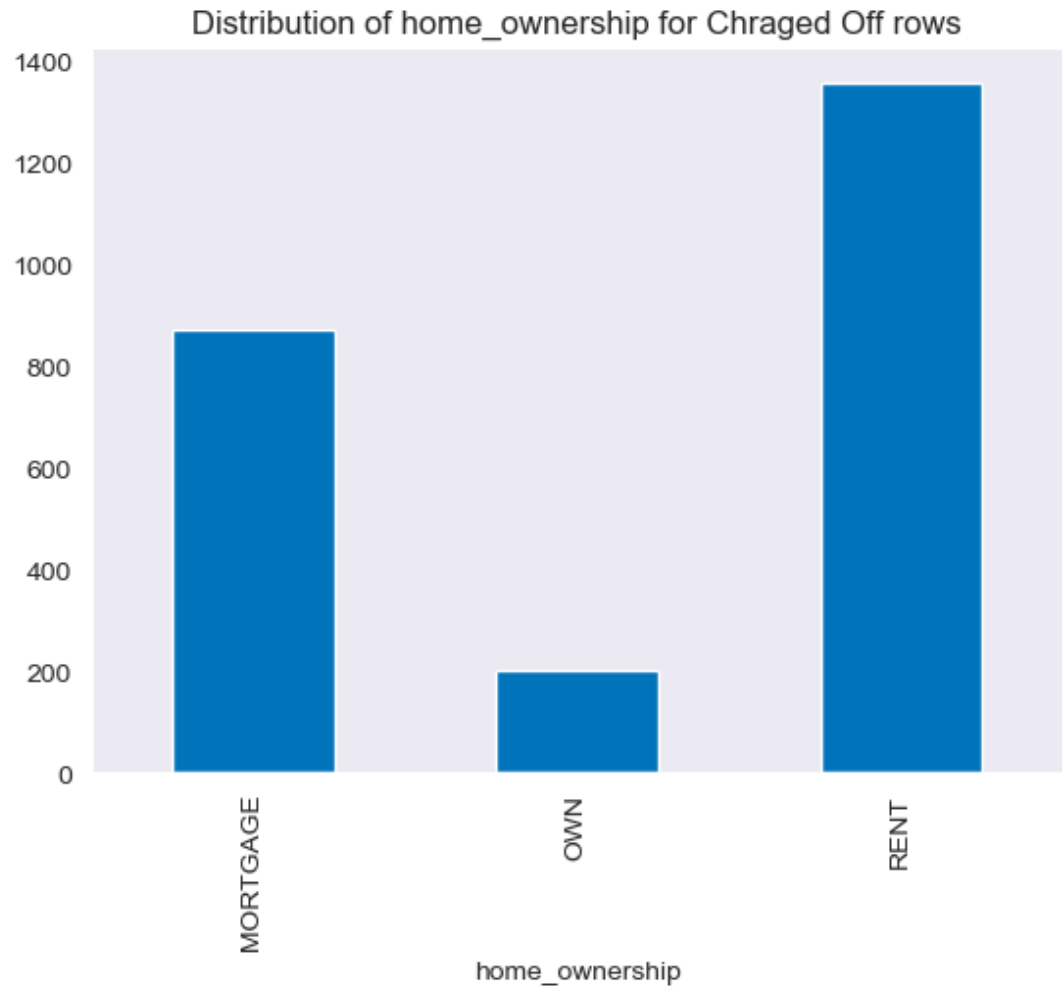
sub_grade C2 has highest charged off cases



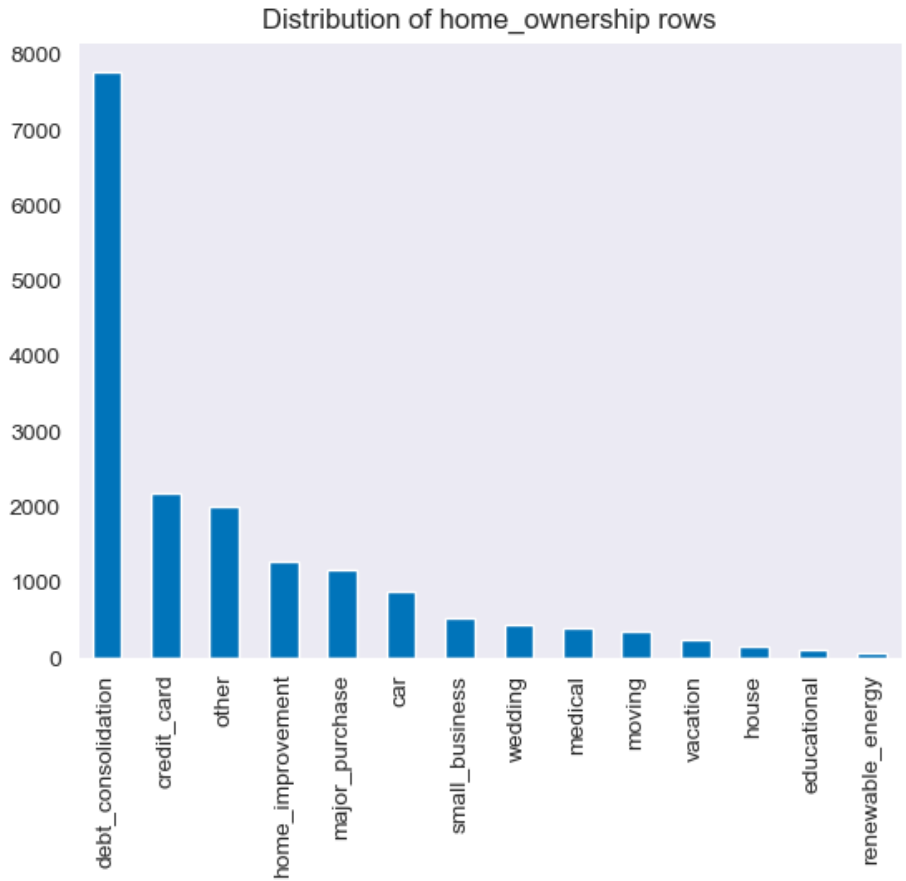
More experienced customers has high prone to charged off

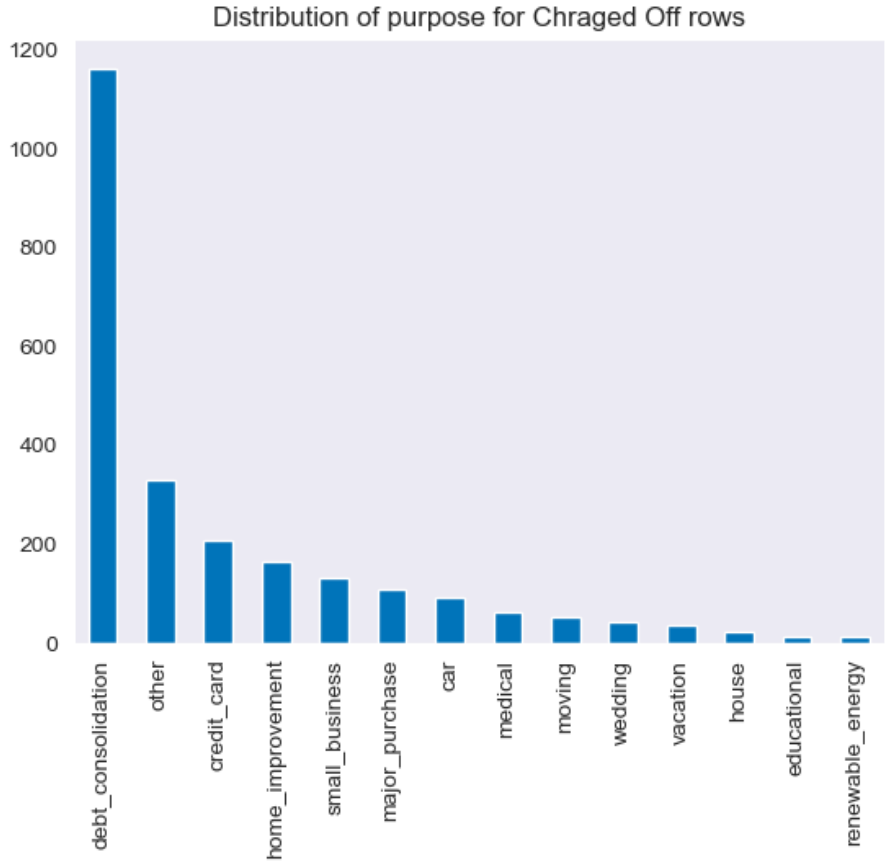


Rented ownership persons have more tends to charged off compare to others



- Debt consolidation purpose has high number of loans
- If we compare both the graphs small_busniess loans high in Charged Off case compared to whole number of loans

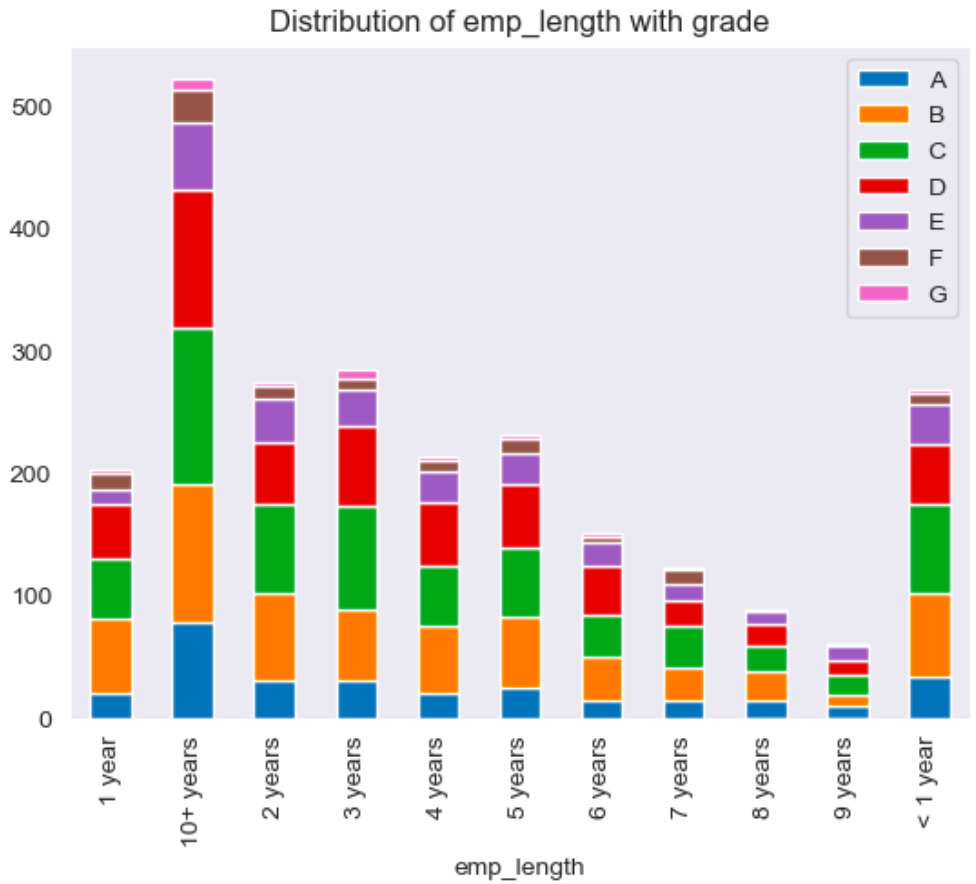


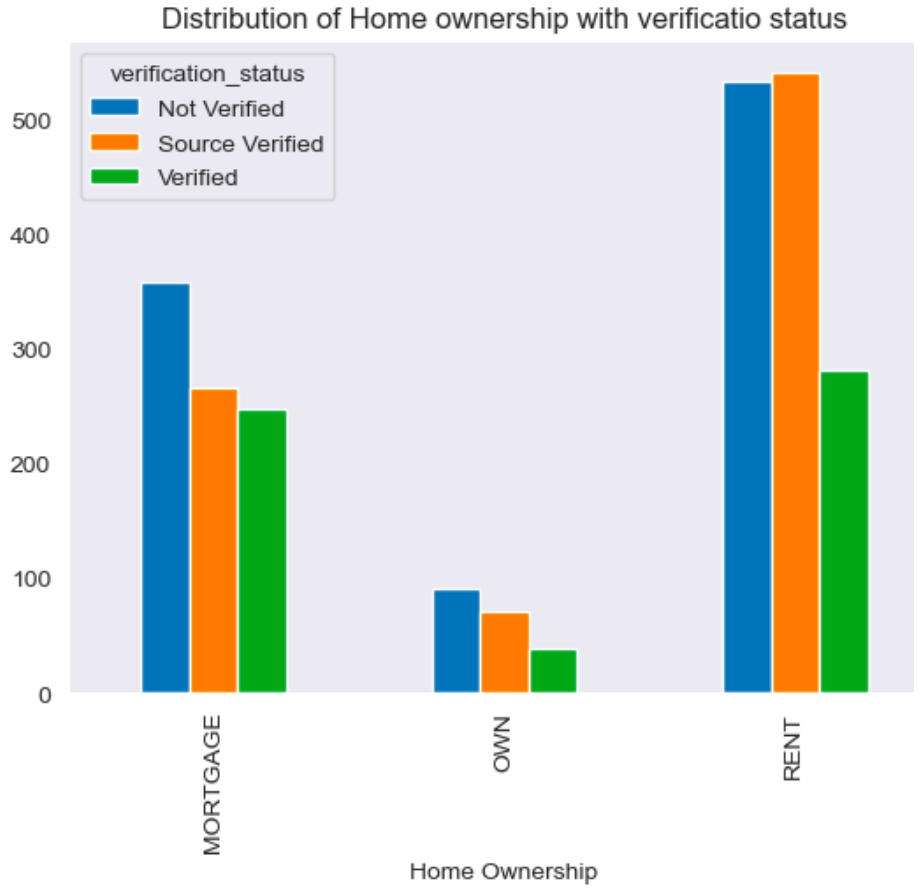


Segmented Univariate Analysis:

Comparitivly 10+years and C grades are highly Charged Off comparatively others.

Rented and Source Verified are more in Charged Off data





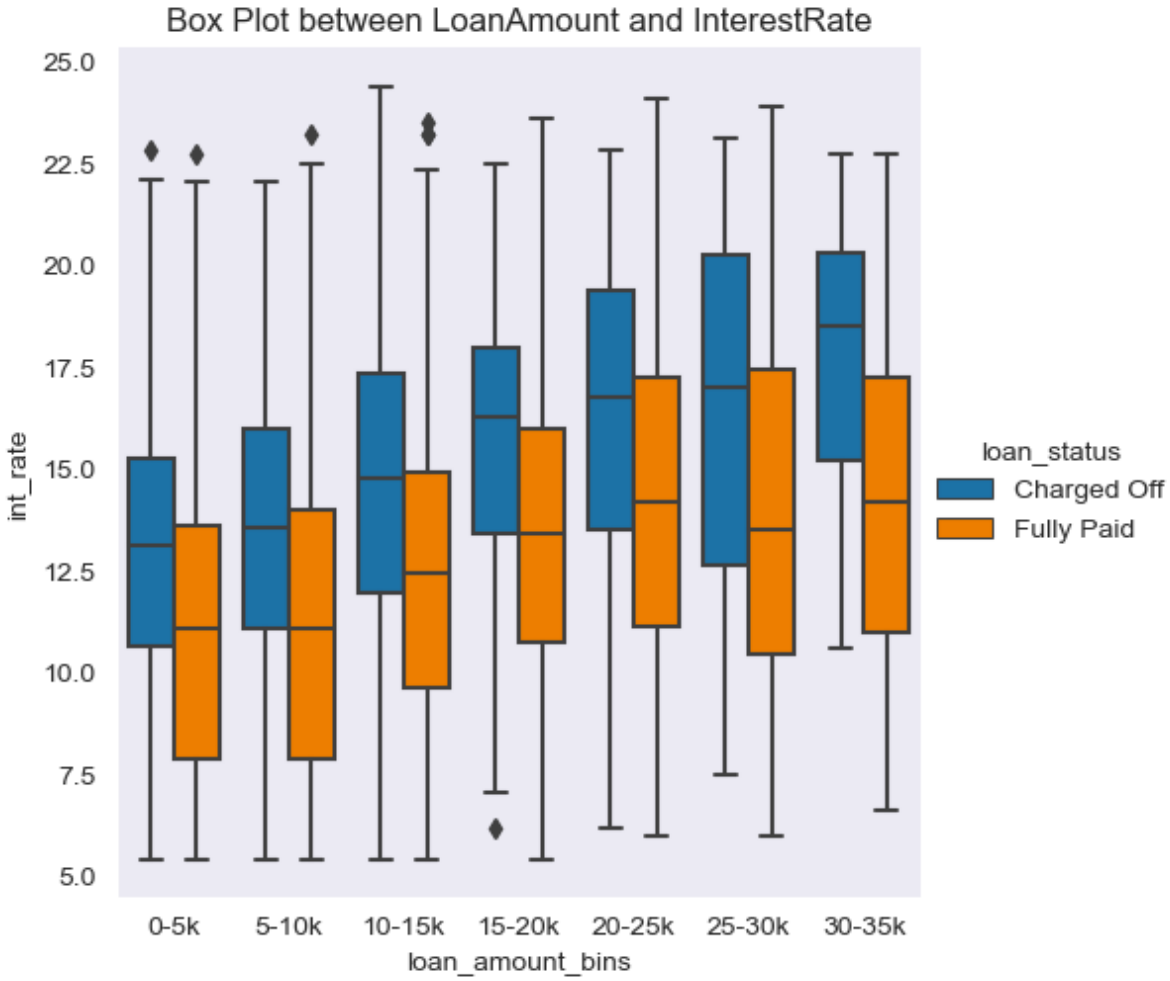
Summary of Univariate Analysis

Below scenario users are more tends toward Charging Off

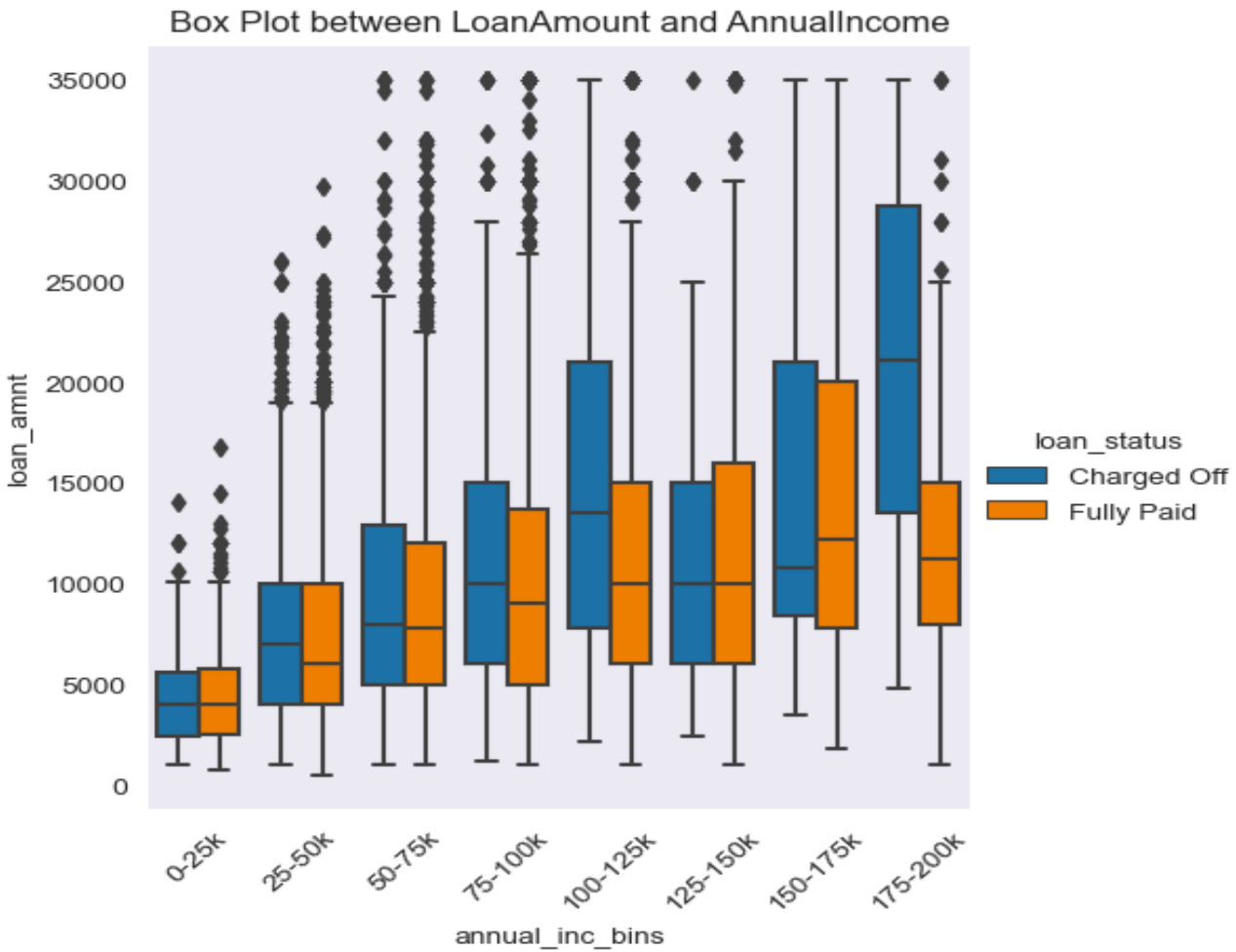
1. Higher loan amount causes more default cases compare to lower loan amount
2. Annual Income between 35000 to 66000 more prone to charging off from their loan
3. Average Interest Rate value of charged off has higher than fully paid rate, Higher interest rates may leads to charged off
4. Grade C has high charged off cases, sub_grade C2 has highest charged off cases
5. More experienced customers(10+ years) has high prone to charged off
6. Rented Source Verified has highest number of Charged Off cases
7. Debt consolidation purpose has high number of loans, If we compare both the graphs small_busniess loans high in Charged Off case compared to whole number of loans.
8. after comparing both the graphs zip_code 900 -999 are taking more number of Chraged Off loans
9. CA state taking more loans and more tends to charged off as well

Bivariate Analysis

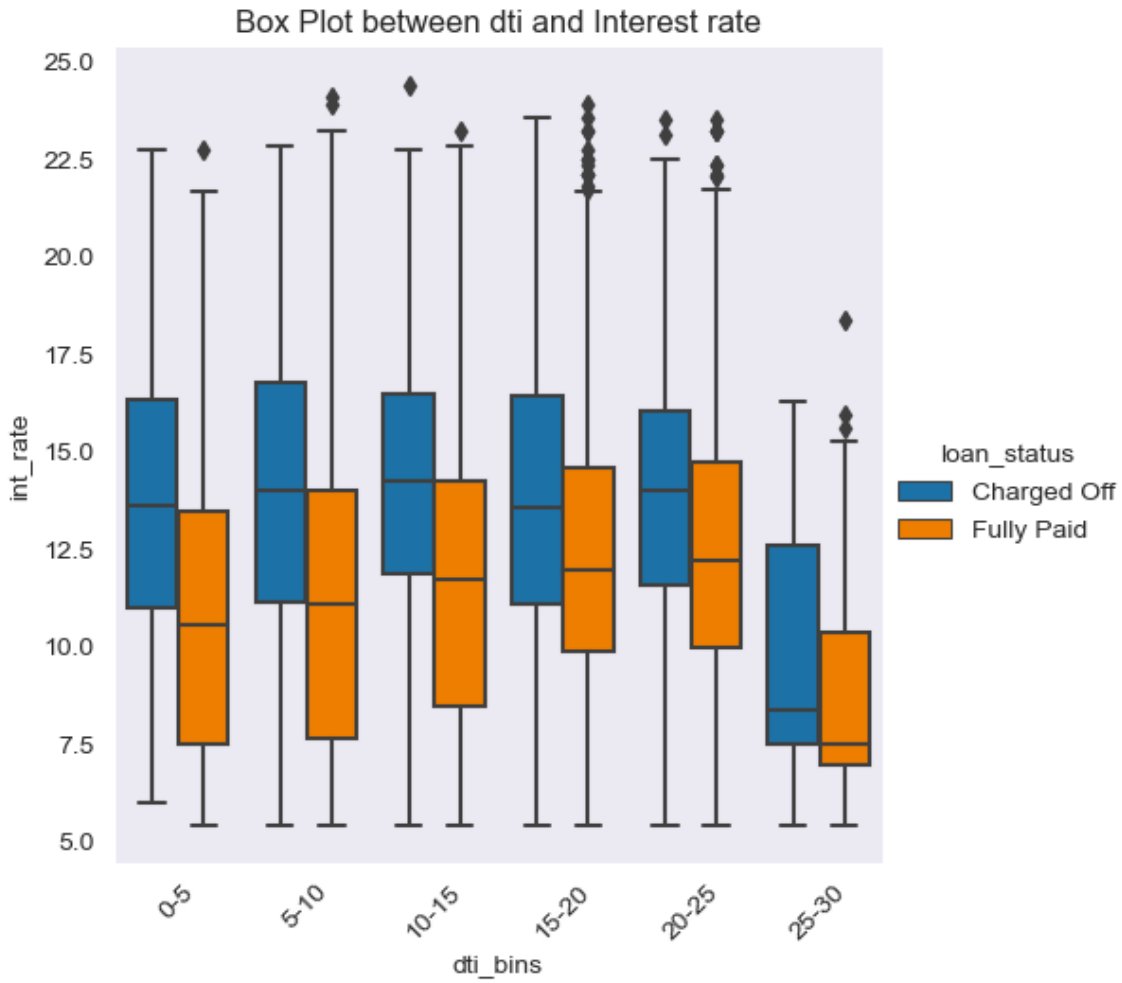
From the box plot we can say that Higher loan amount with higher interest rate leads to Charged Off cases



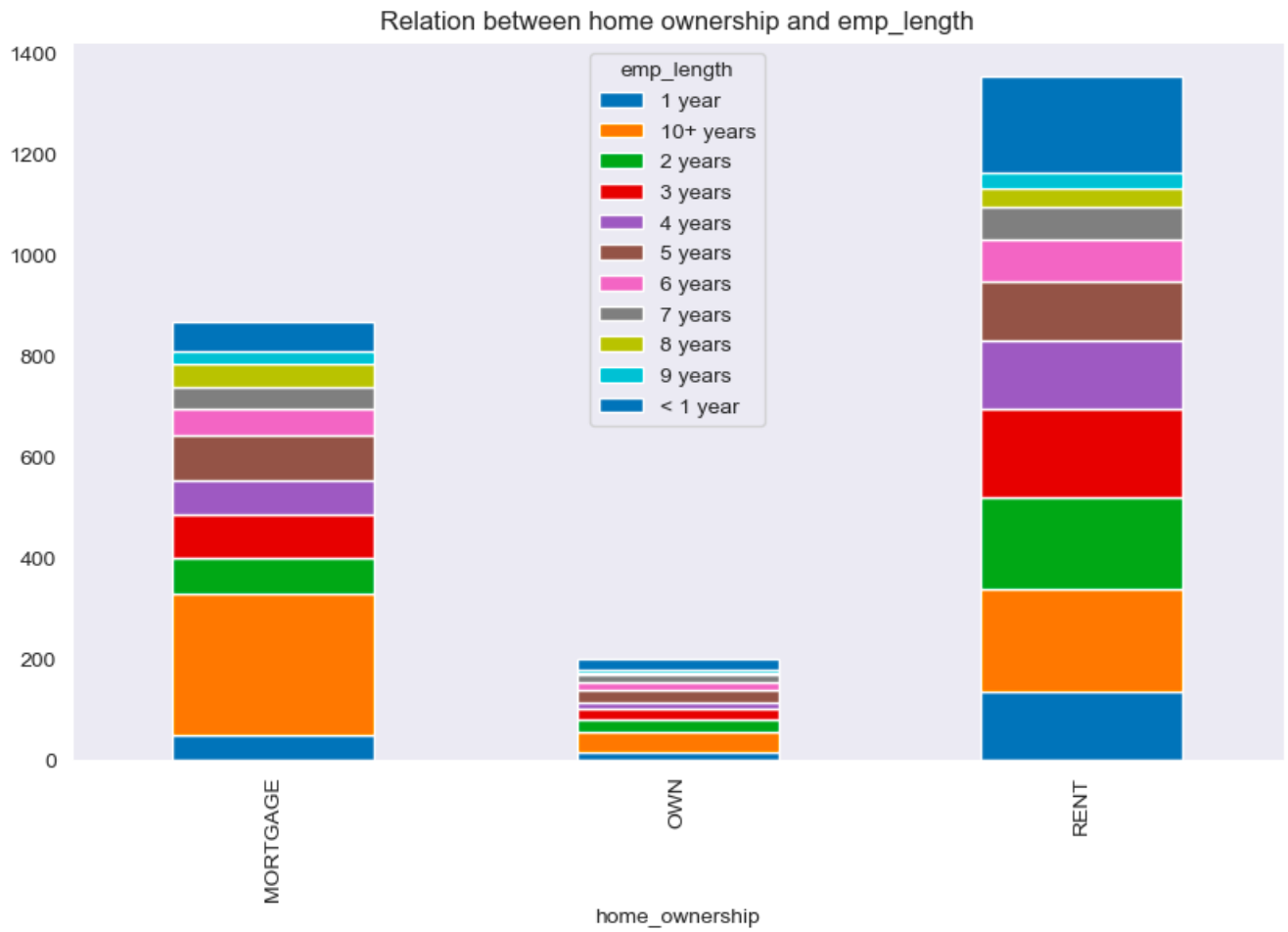
Higher Income with higher loan_amount leads to Default



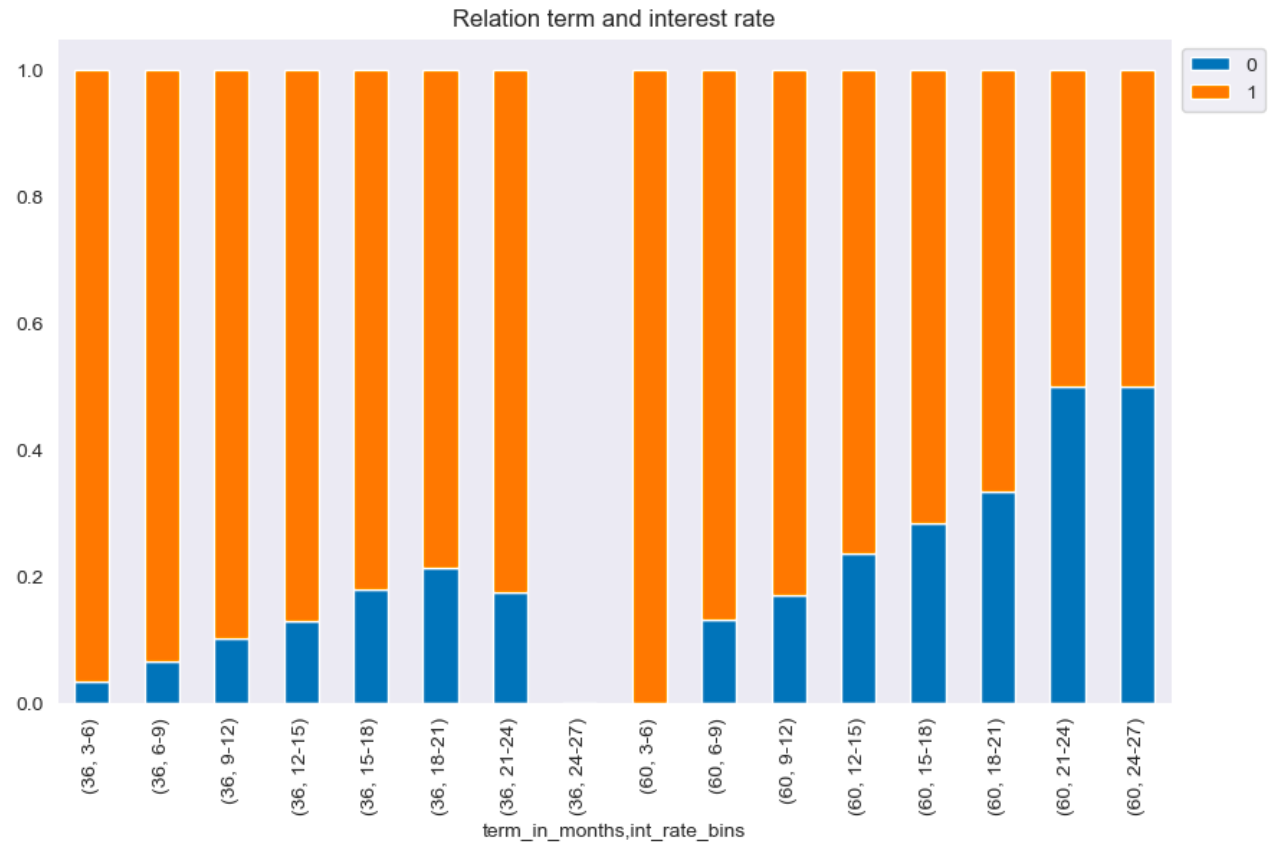
Lower Dti with interest rate between 11-16% leads to Default



More Default cases we can see here in Ownership as Mortgage and emp_length > 10 years



high interest rate with 60 months tenure has high probability of 'Charged Off'



Summary of Bivariate analysis

- Higher loan amount with higher interest rate leads to Charged Off cases
- Higher Income with higher loan amount and annual income leads to Default
- Lower Dti with interest rate between 11-16% leads to Default
- More Default cases we can see here in Ownership as Mortgage and emp length > 10 years
- High interest rate with 60 months tenure has high probability of 'Charged Off'
- Zip Code in 900 -999 range(CA state) highest number of default cases Comparatively high in December month