LENDING CLUB

EDA Case study

Aditi Sharma , S Rajini Parimala

ABSTRACT

Lending Club 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.

AIM:

Identify risky loan applicants, so that 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.

PROBLEM SOLVING METHODOLOGY

DATA CLEANING : In this process, the null valued columns , unnecessary variables are cleaned, and check for null value percentage and remove the respective rows.

DATA UNDERSTANDING: Here we try understanding the data and get an understanding of all the columns and their domain specific uses.

DATA ANALYSIS:

Univariate Analysis : Analysing each column, plotting the distribution of each column
Bivariate Analysis: Analysing the two variable behaviour like term and loan status with respect to loan amount.
Segment Univariate Analysis: Analysing the continuous data columns with respect to the categorical columns.

RECOMMENDATIONS: After analysing all the plots we identify the columns that best contribute to the loan defaulters and hence reducing the loss to the business.

DATA SET GIVEN BY THE COMPANY INCLUDES APPLICANT’S DETAILS SUCH AS

* Applicant’s previous loan details (including the payment details etc.,)

* Applicant’s state, region

*Applicant’s employment details

* Purpose for which applicant has applied for the loan

* Whether the applicant is bankrupted and so on.

DATA CLEANING

As a part of data cleaning, we see that the data set given has many columns that does not play any role in analysing the data, such data is cleaned for better understanding.

Columns that tell about customer's behaviour are cleaned for example,

delinq_2, earliest_cr_line, recoveries, out_prncp, total_acc, last_pumnt_amnt etc.

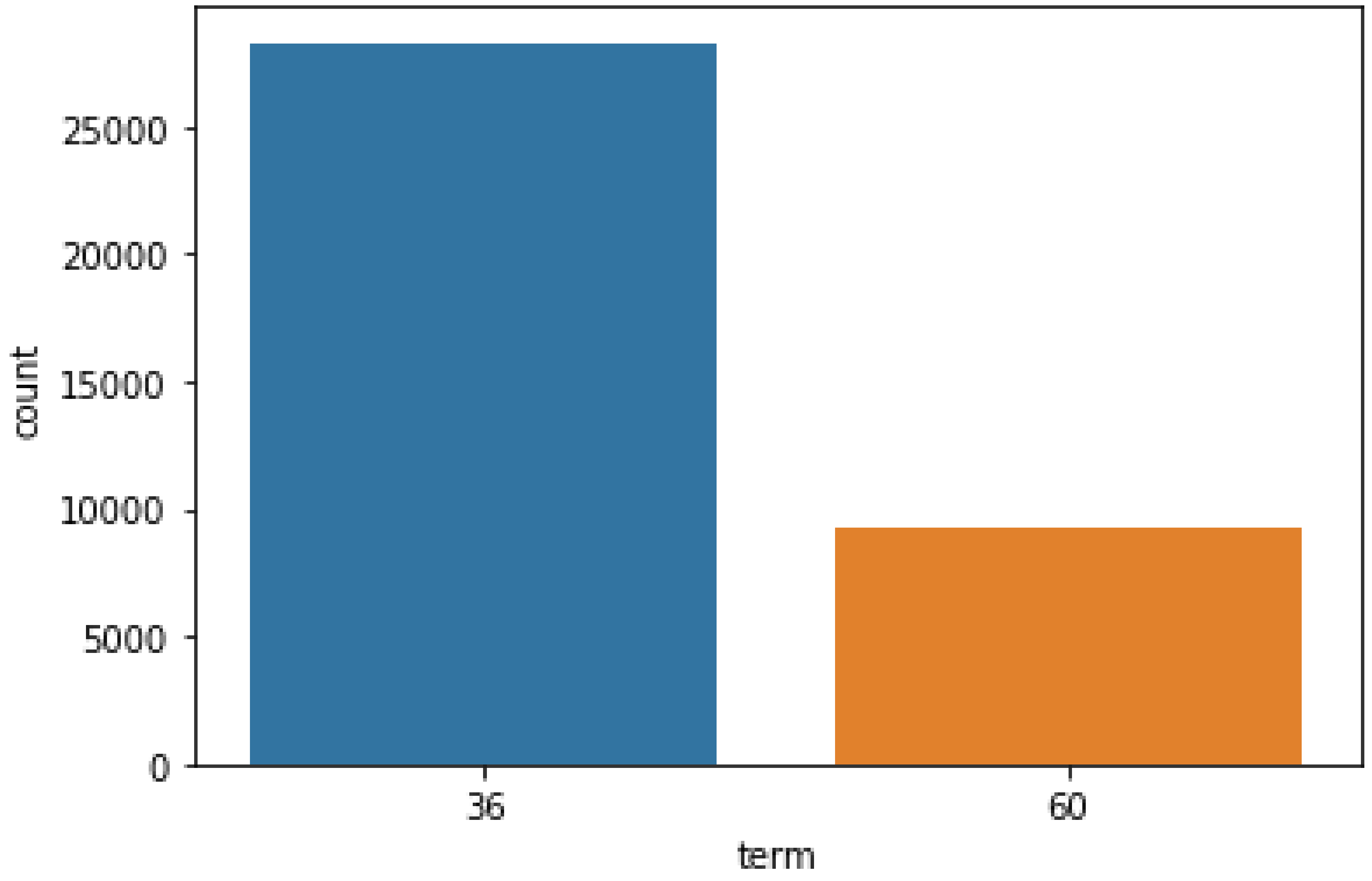
* Most of the columns with highest correlation to the default status are the parameters that are obtained only after the loan is approved .

* So after evaluating the correlation of such columns with the primary details of the customer that can be obtained before the loan is approved , we have selected the columns that have indirect dependence of the default status on them .

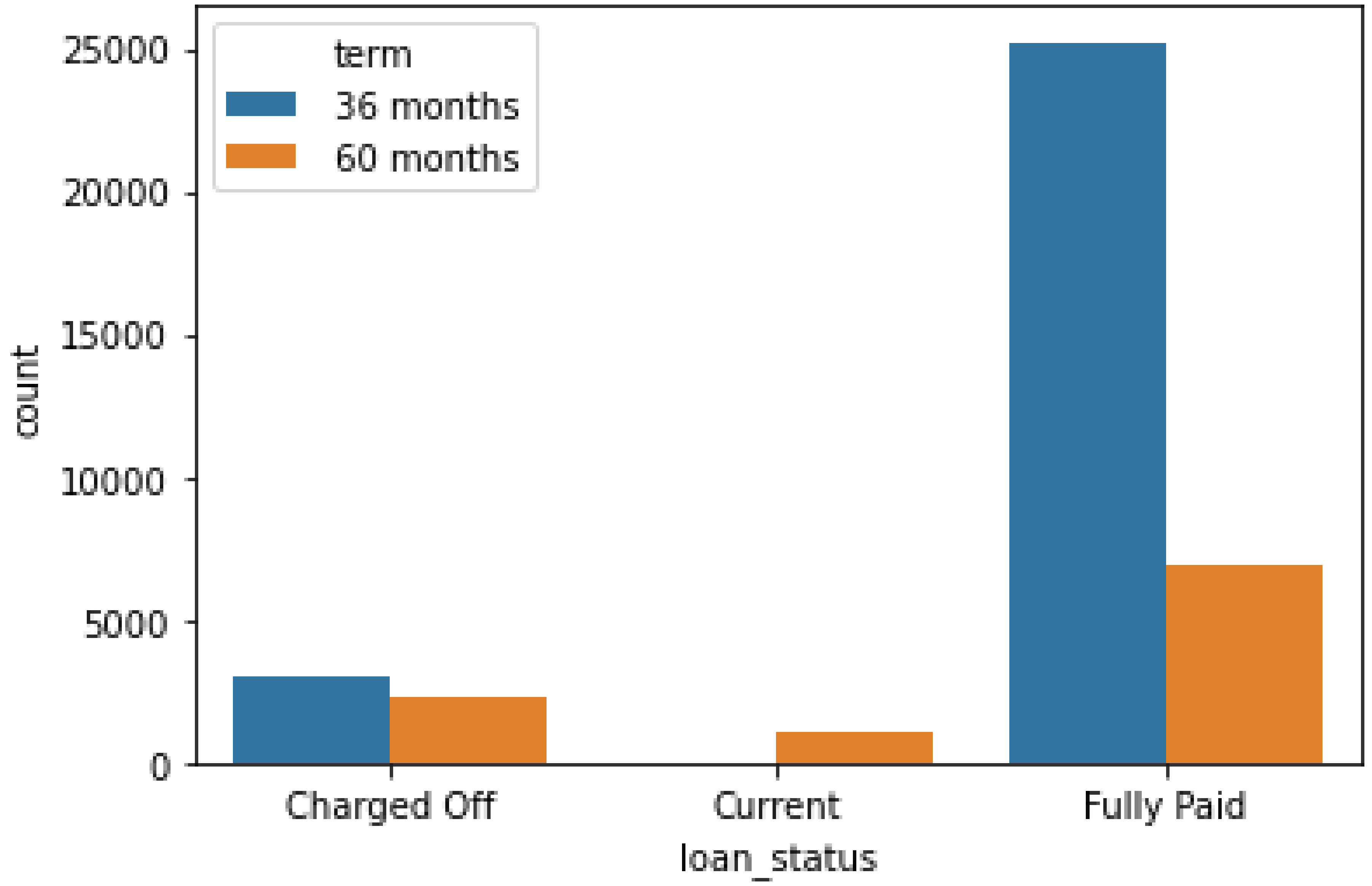
* Such as loan amount , instalments , funded amount and funded amount by the investor .etc.

*   Rows with loan status Current were also removed as they were irrelevant to the purpose .

EXPLORATORY DATA ANALYSIS



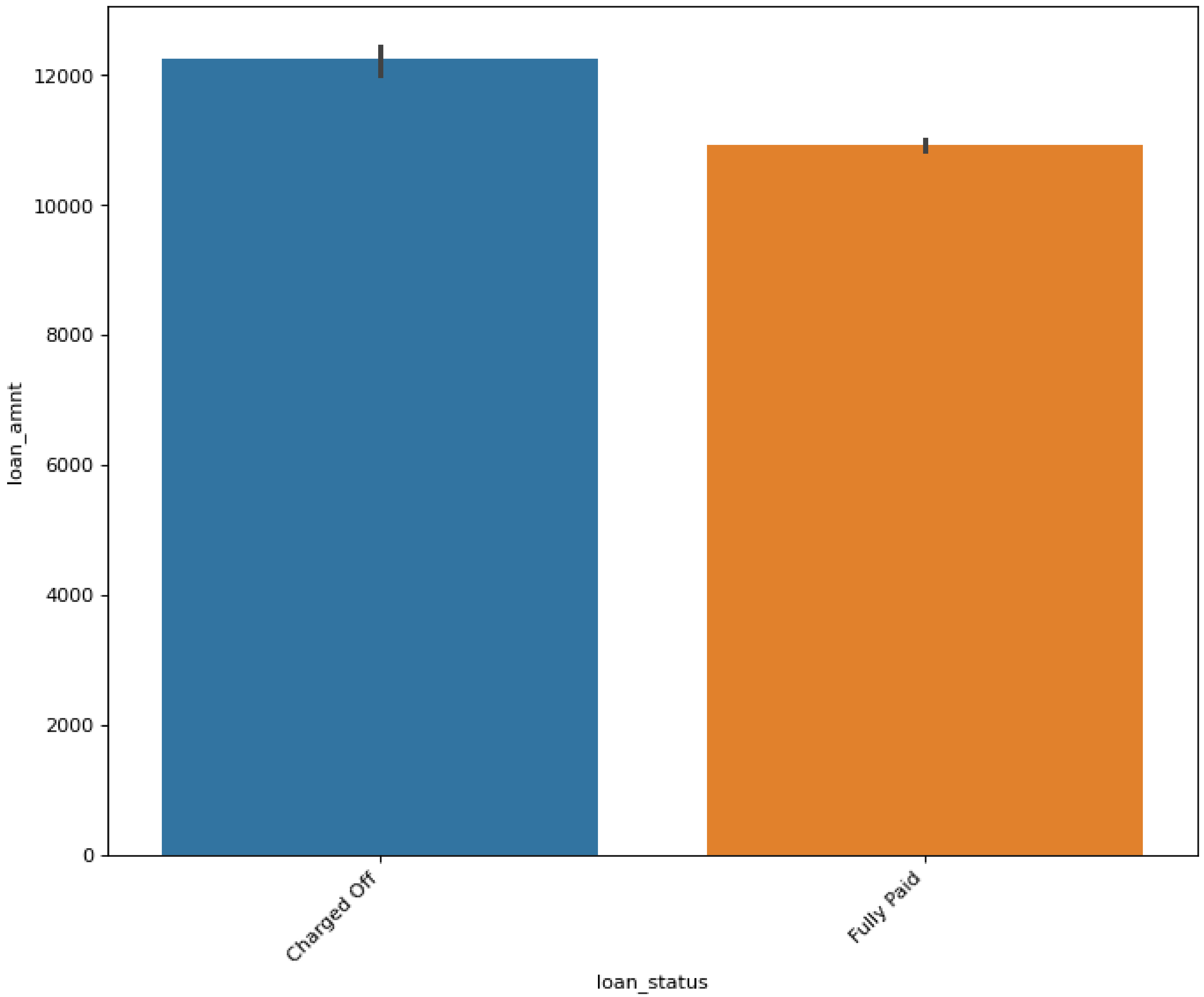
Term Distribution



Loan Status Distribution

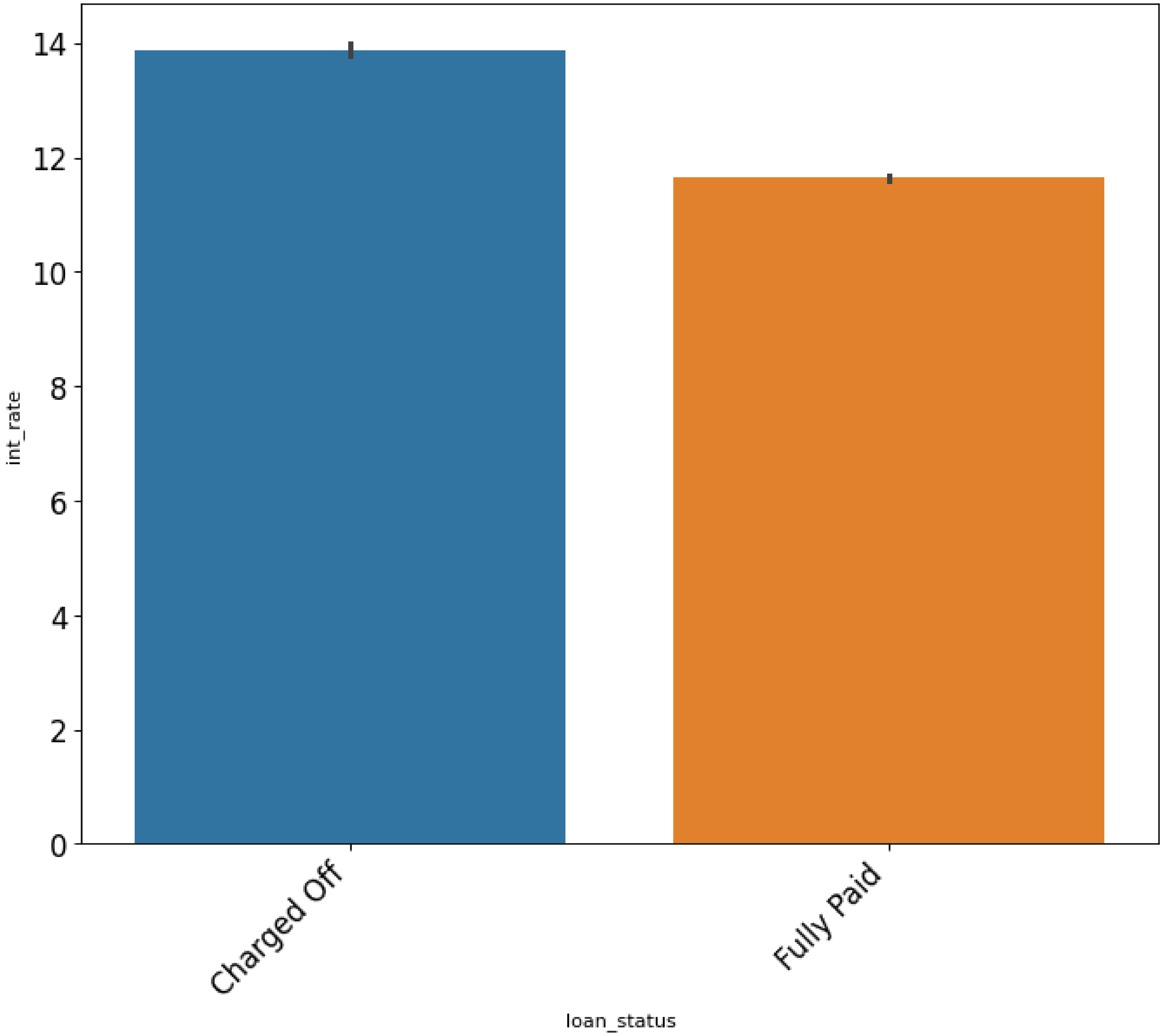
* As per the records, there are two loan terms 36 and 60 months, we observe that almost 75% of the customers have opted for loan term of 36 months.

- Almost 90% of the customers with 36 months term loan are fully paid and about 10% of them are charged off.
- Count of Charged Off customers for both terms are comparable while more than half of the Fully Paid customers belong to term of 36 months



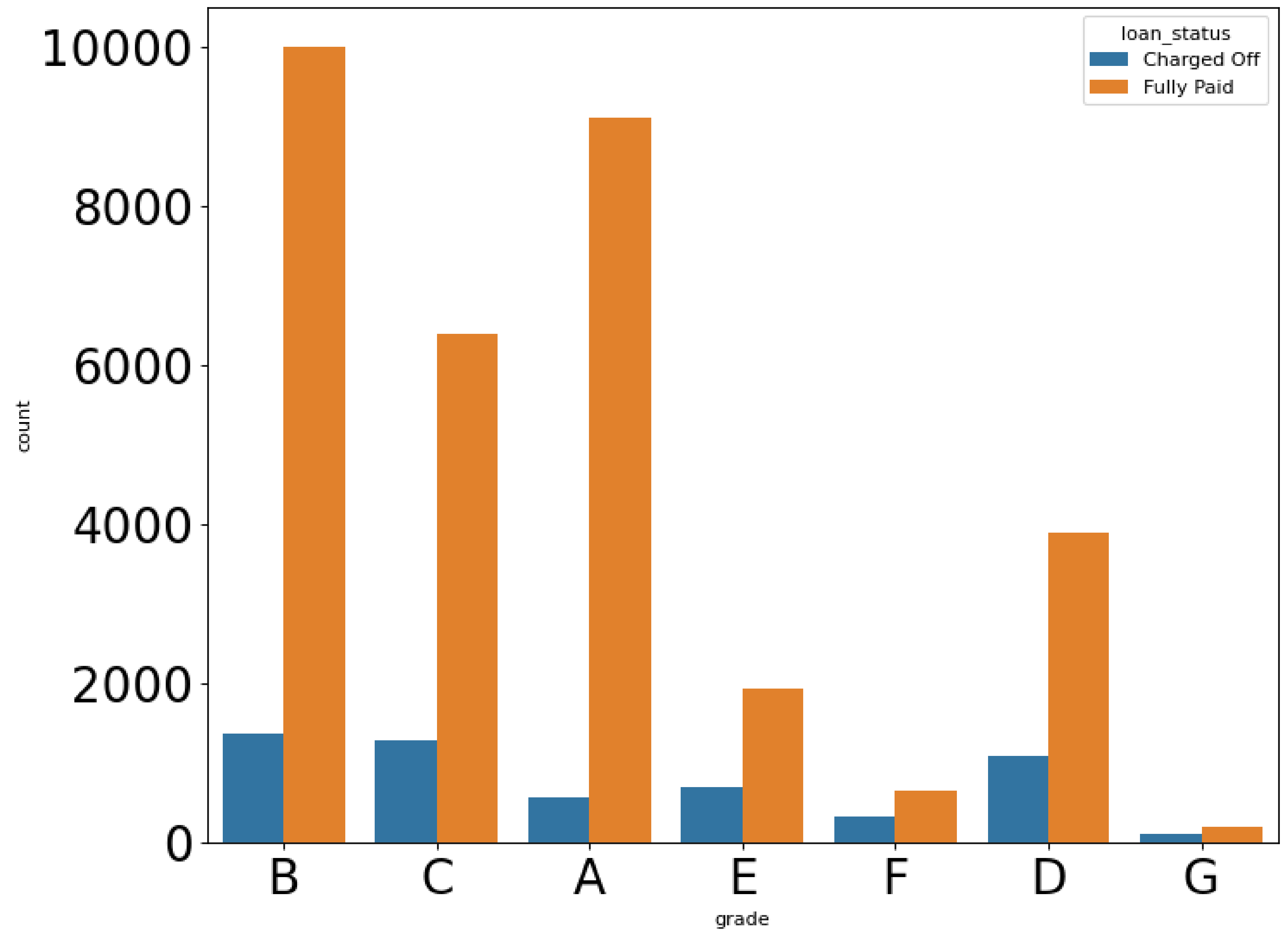
Loan amount Vs Loan status

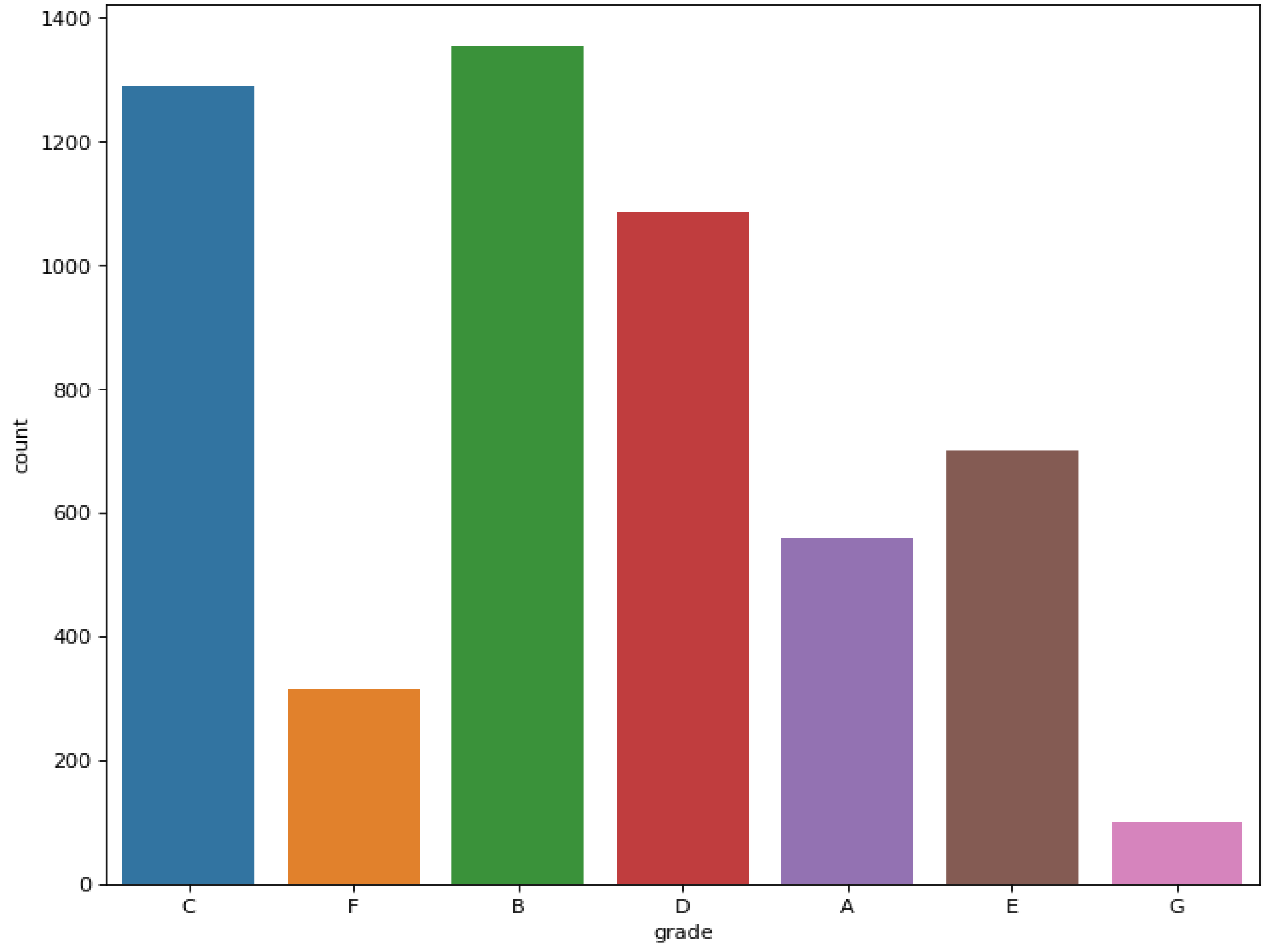
* From the above plot it is understood that higher the loan amount more the charged off and high chance of loan getting defaulted.



Interest Rate

Above plot shows that most of the customers whose data is verified are fully paid. From this we can filter out the loan applicants who are default.



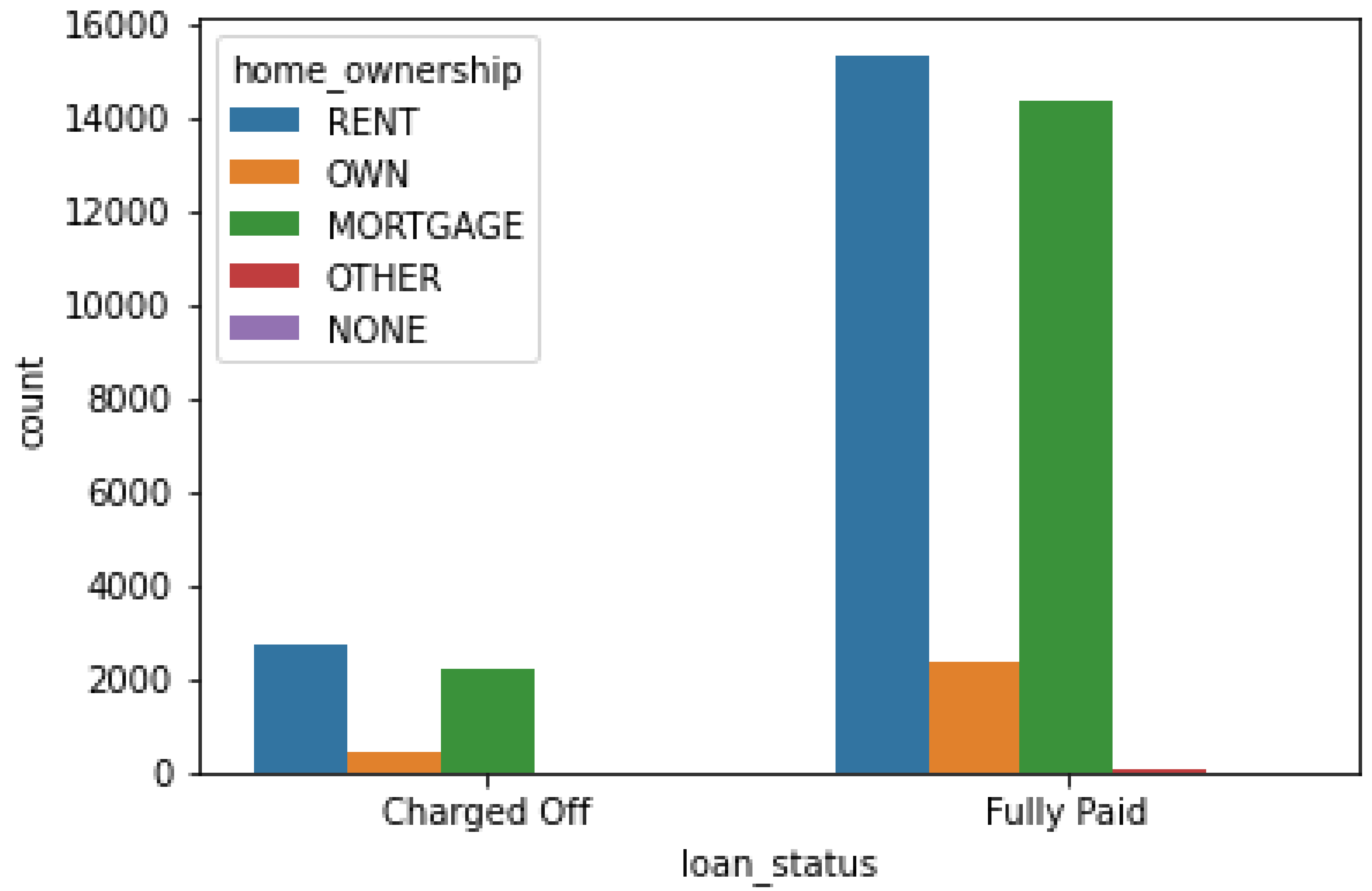


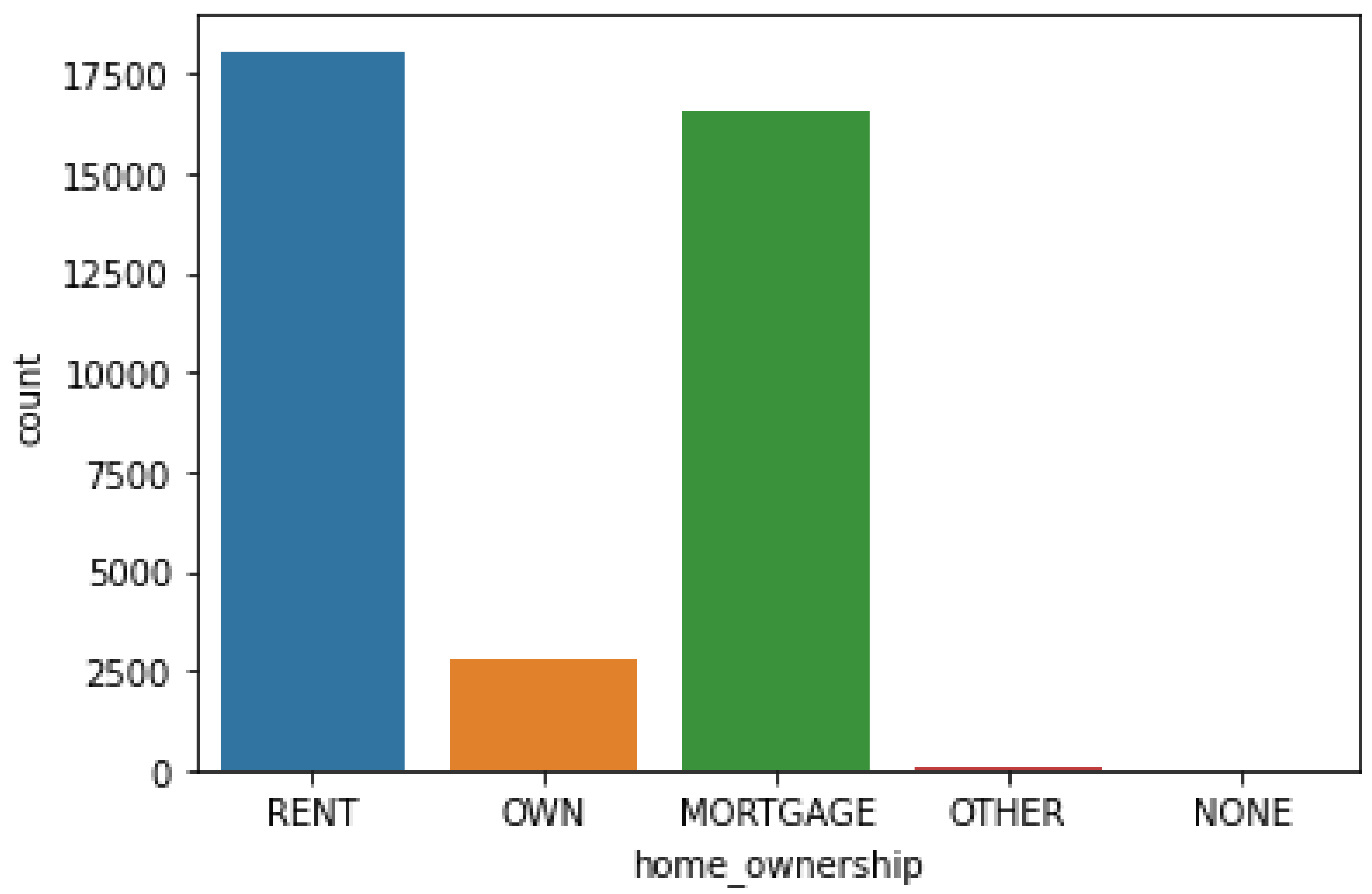
Grade Vs Loan status

* Grades are very good category to tell the borrower probability of defaulting the loan.

* The lower the grade (E,F,G) have higher chances of defaulting the loan than higher ones(A,B)

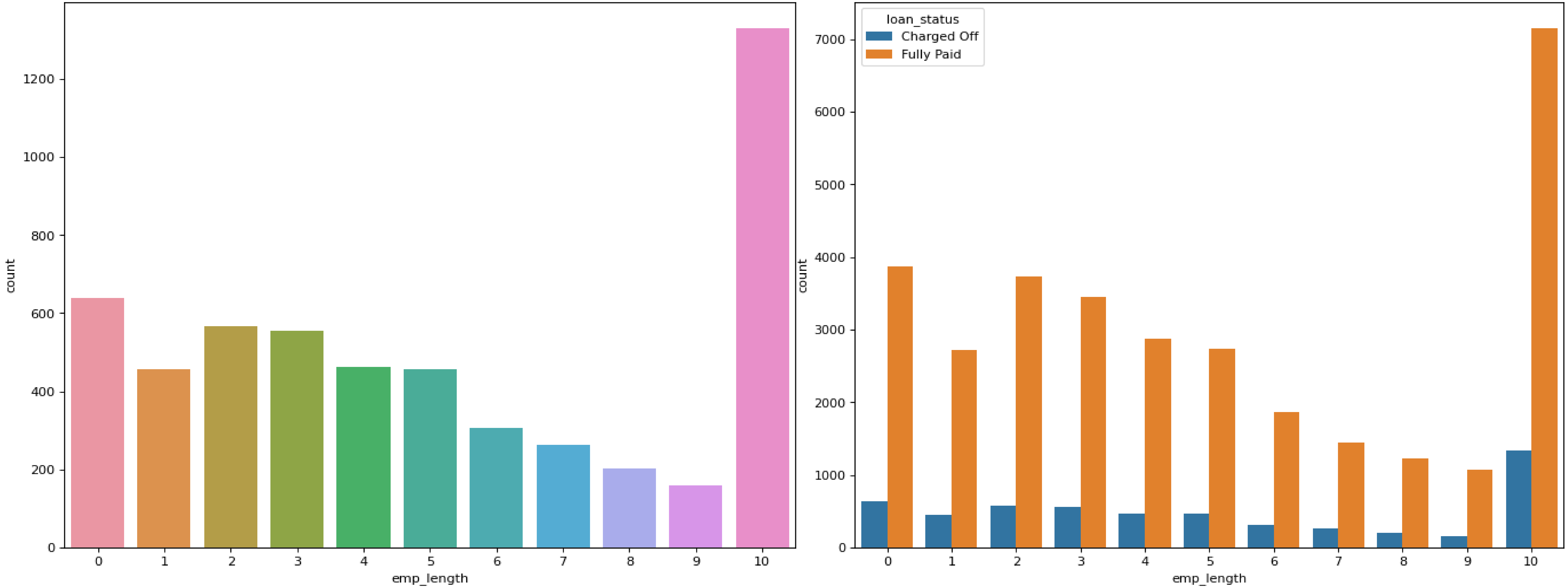
* The lower grades are getting loans for higher interest rates which might be the cause for loan default.





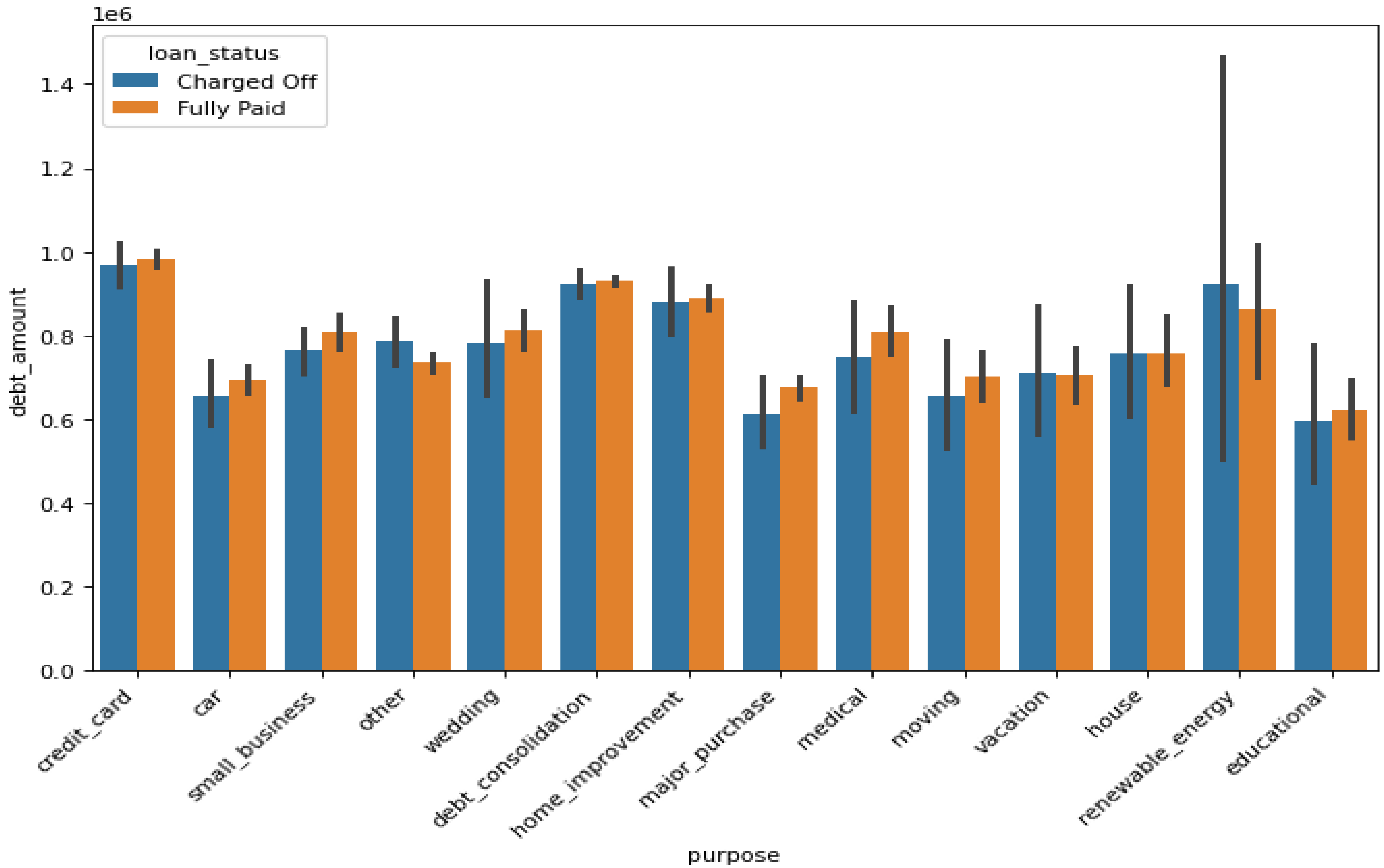
Home ownership Vs Loan status

* Home ownership follows a similar trend in both fully paid and charged off indicating that Home ownership has no major impact on Loan status. Other than fact that the count of loan applicants is more who are paying rent .



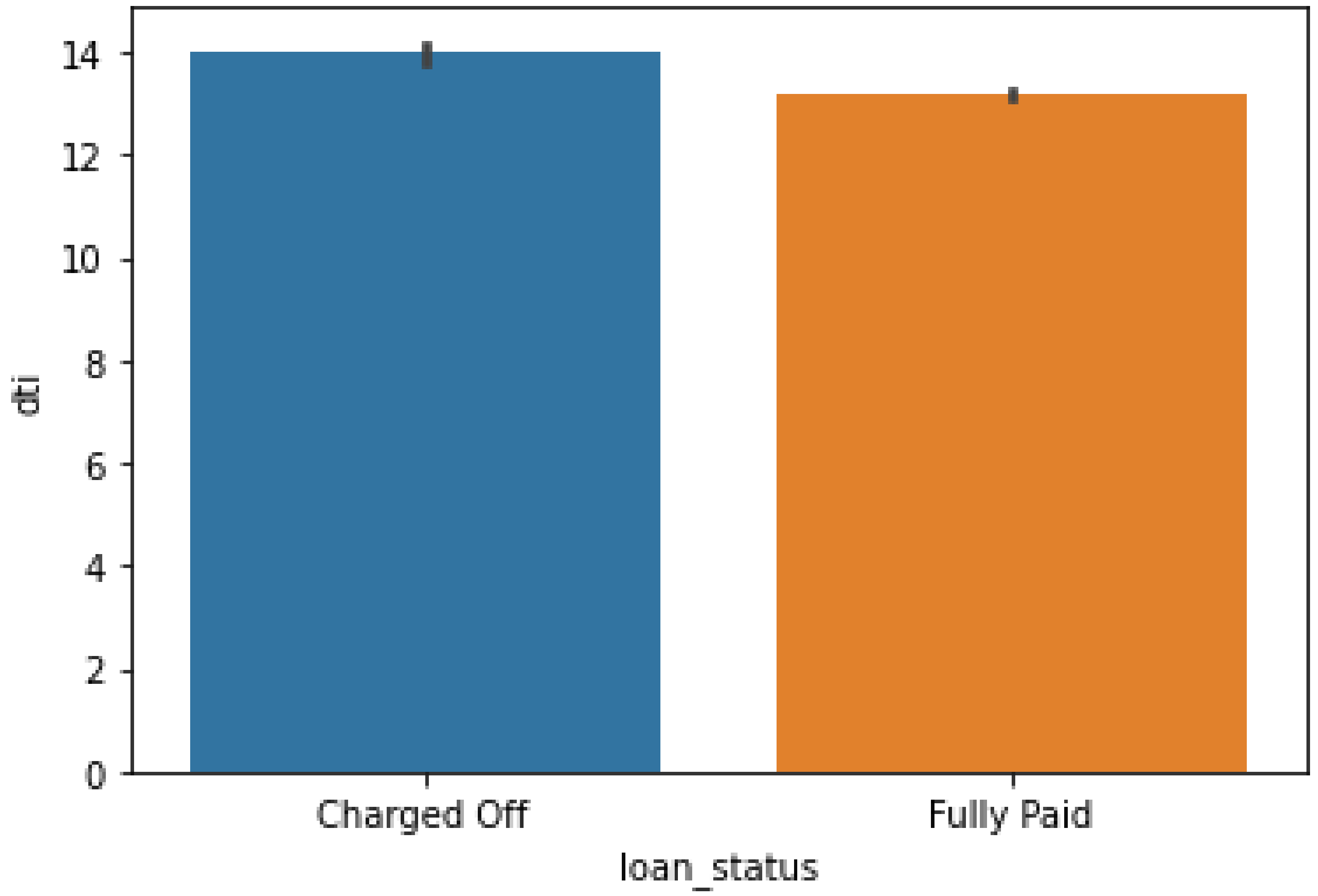
Employee length Vs loan status

. People with experience of 10 years and higher are maximum in count for both the cases of charged off customers and fully paid.

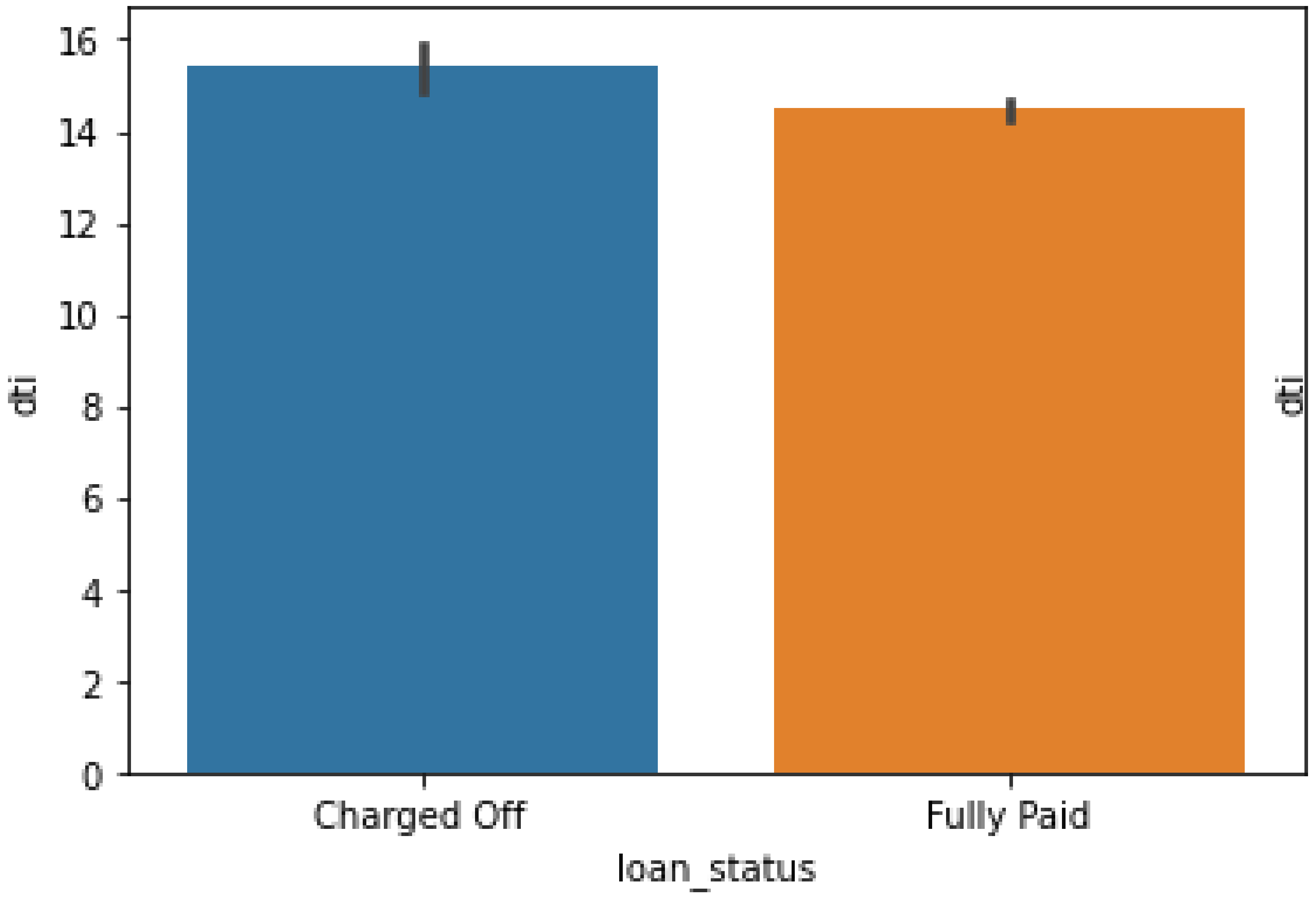


Purpose Vs Loan status

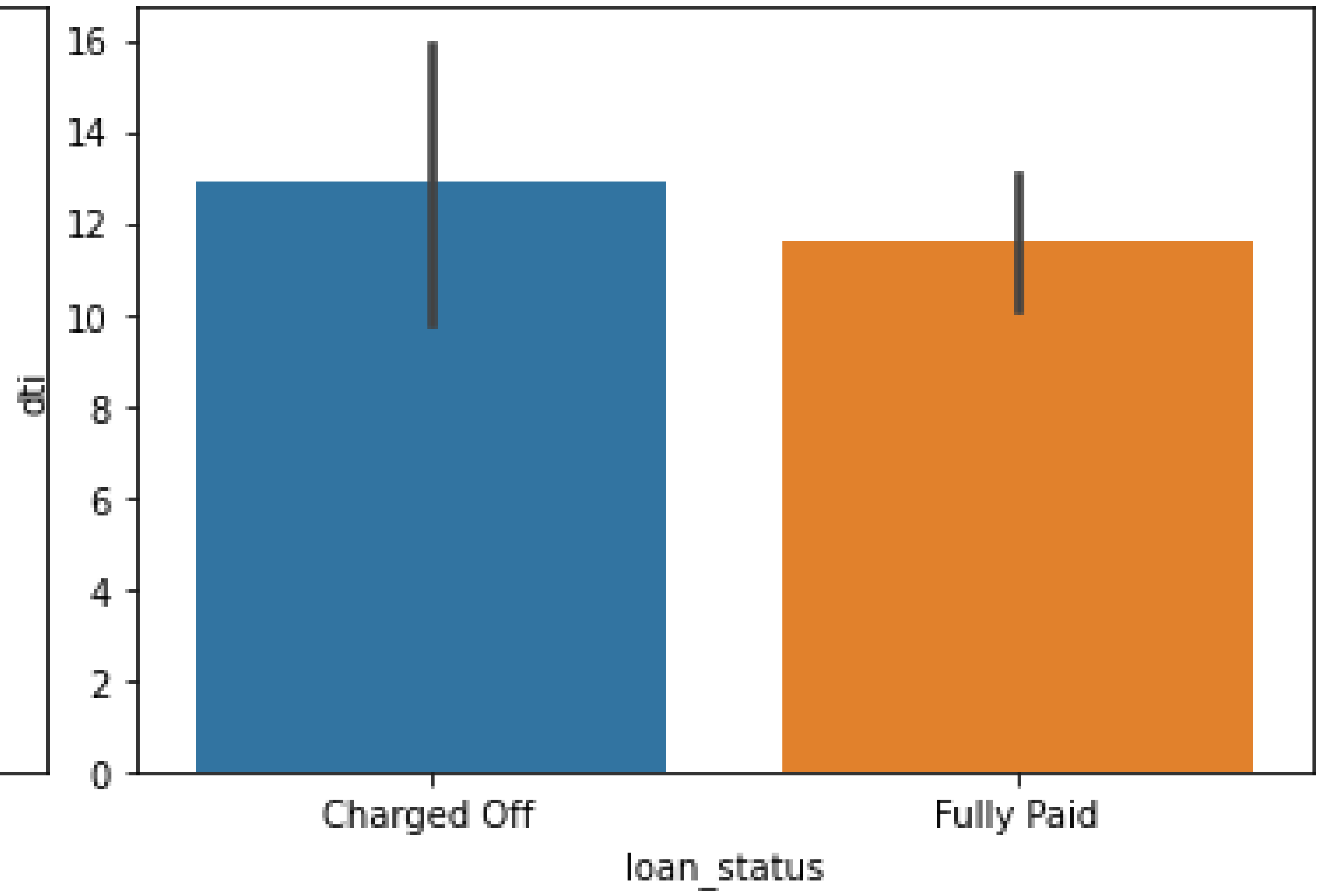
.Customers who applied for loan with the purpose of 'renewable_energy','other'and 'vacation' are more charged off than the ones who fully paid.
. The purpose for which the customers have applied for the loan most are 'credit_card' and 'debt_consolidation'.
. We can conclude that it is safe to approve loans for necessary purposes such as 'small_business','wedding','medical' and 'educational'.



All Purpose



Credit Card



Renewable Energy

Loan status vs dti for loan purpose

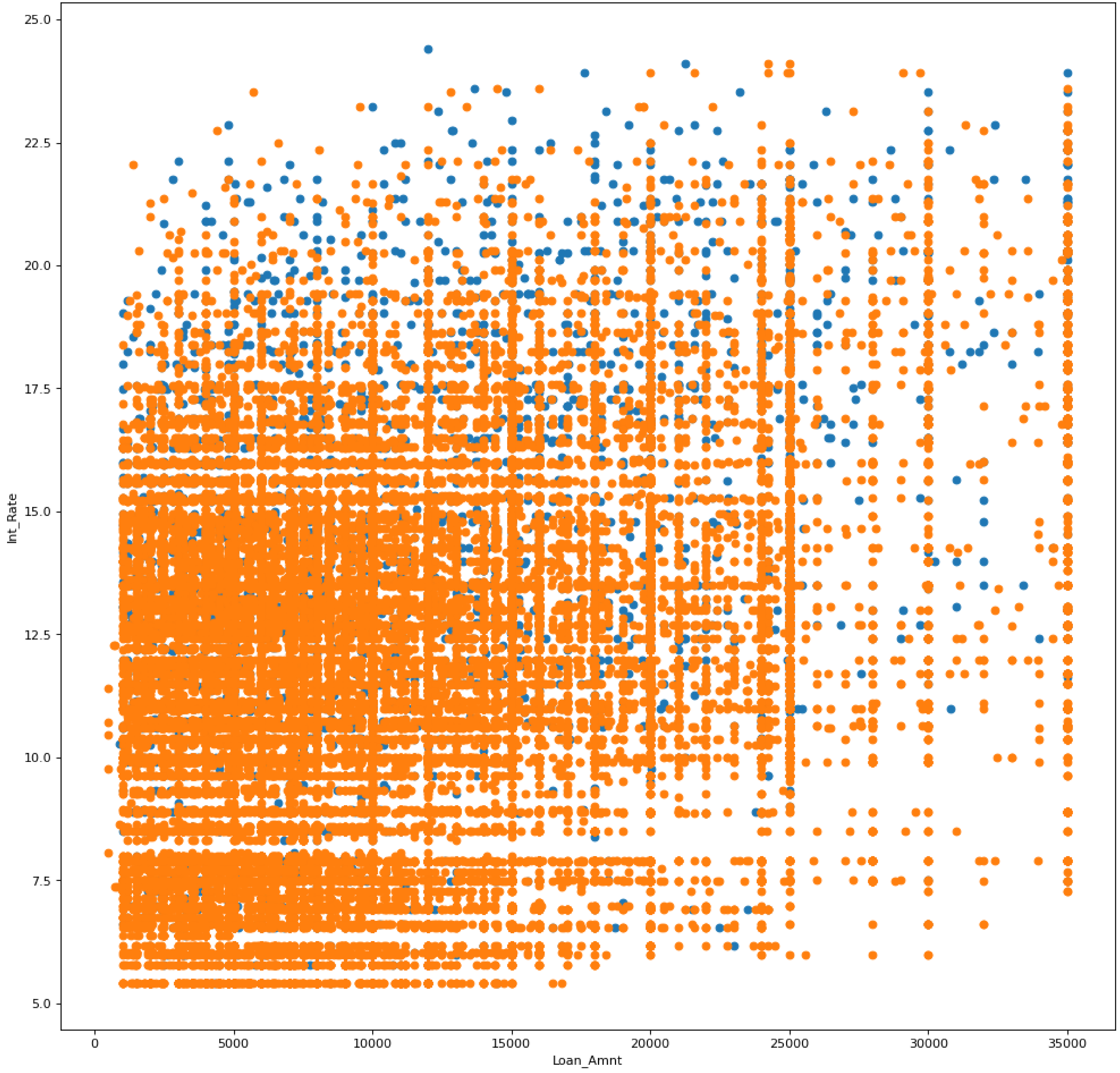
. In general the difference in the average dti for customers who fully paid or charged off is not very significant but from the above graphs we can conclude that with respect to purpose we can distinguish the dti range for which the loans can be approved or discarded .
. It is advisable to not approve loans for customers who have purpose as credit card and dti value greater than 14.48 which is the average dti value for which the customers paid off .
. It is advisable to not approve loans for customers who have purpose as renewable energy and dti value greater than 11.61 which is the average dti value for which the customers paid off .

Loan amount vs interest rate for loan status

. Even if the loan amount is low but the interest rate is too high then the count of charged off customers are more than fully paid.

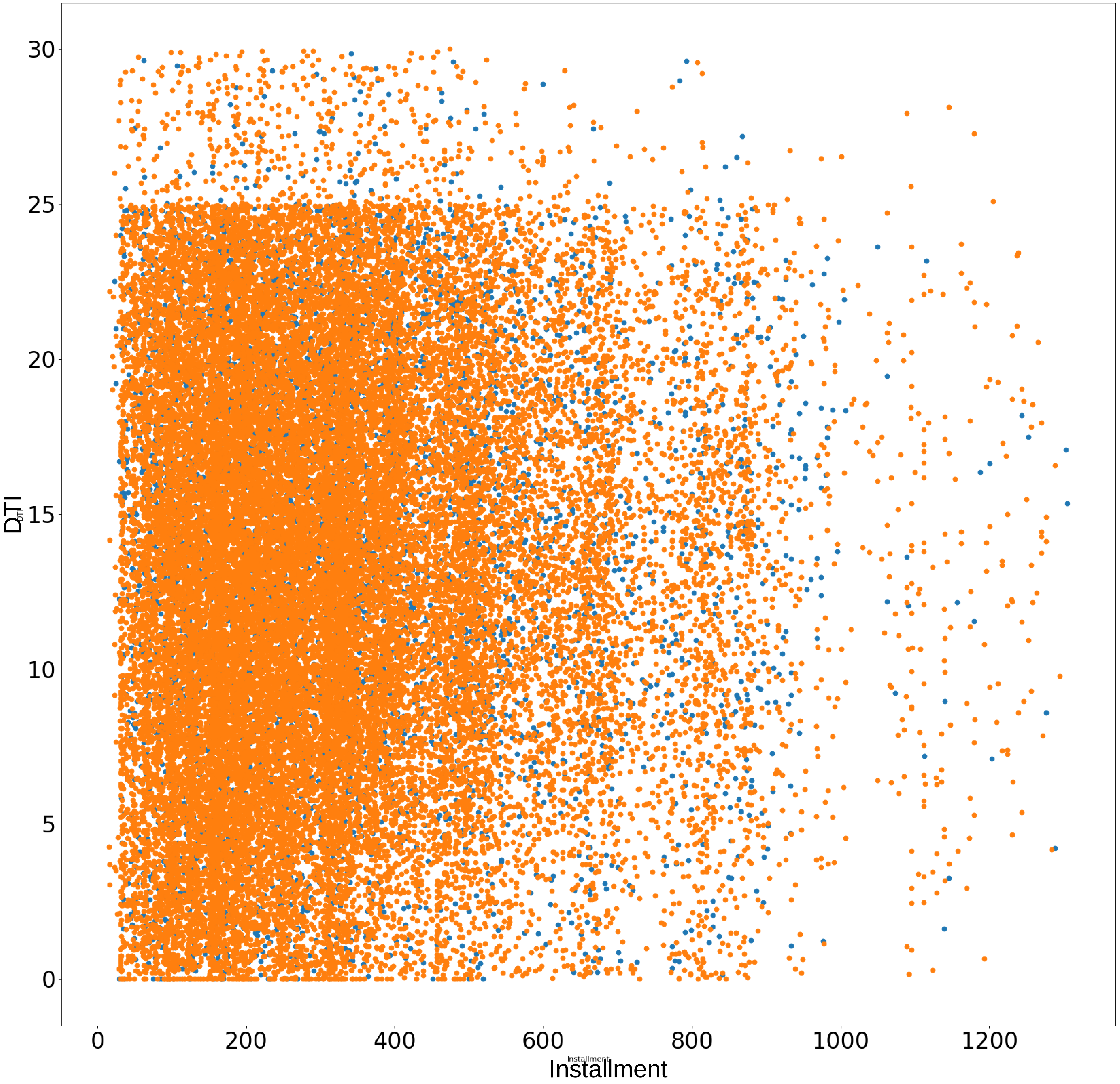
. For loan amounts above 25000, approved applications for loans have decreased significantly.

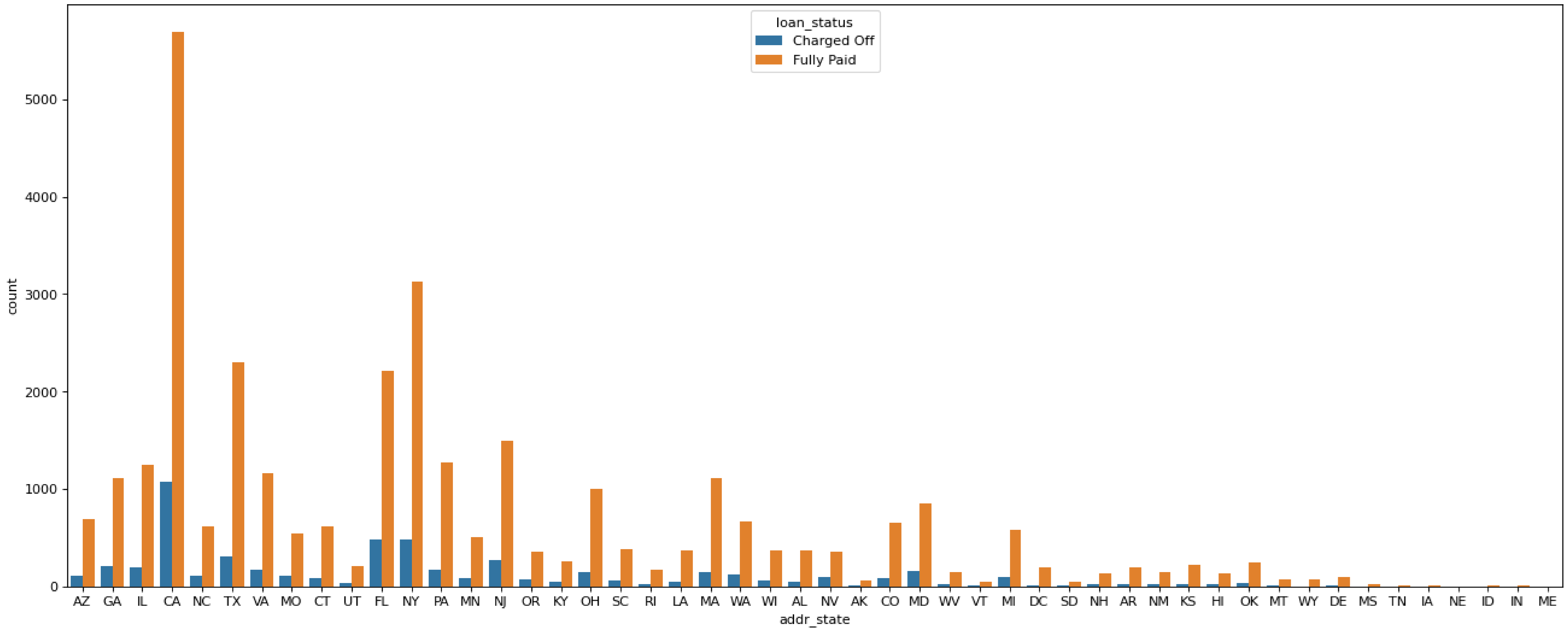
. For loan amounts greater than 25000 and interest rate above 15% people mostly charged off than fully paid .



Loan amount vs interest rate for loan status

.Loan applications have decreased significantly when dti is above 25 as well as when the installment amount increases more than approx. 900. But the charged off count is still comparably dense where the dti increase above 25 but the installment is still low than the charegd of counts in teh region below that dti level.





The loans which are given for CA,FL,TX state borrowers are defaulted more than any other state.

Conclusions

* Higher the loan amount and higher the interest rate then more the loan defaulters.

* Company should reduce the high interest rate for 60 months tenure, there are more chances for loan to get default.

* Grades is one of the good metrics for finding loan defaulters. Company has to look into the information from borrowers before issuing loans to low grade(G to A)

* Purpose also plays an important role in identifying the loan defaulters. Borrowers whose purpose is debt consolidation, credit card repayment etc., are more prone to be defaulters.

* Borrowers who’s house ownership is mortgage have taken higher loans and defaulting the approved loans.