Lending Club Case Study

Group Members:
1.Garga Sen
2.Gaurav Minocha

Problem Statement

Company

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.

Problem

Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). Borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters

Objective

Our Objective is 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.

Analysis Approach

Data Cleaning

- Handling Missing Values
- Removing invalid columns

Data Understanding

- Understanding meaning of columns
- Removing columns not required
- Fixing datatypes of columns
- Deriving new columns from existing columns

Univariate Analysis

- Segregating numerical and categorical columns.
- Analyzing each categorical column and the effect of charged-off loans on these columns.
- Analysing each numerical column and removing outliers.

Segmented Univariate

- Segmenting numerical column on the basis of loan status
- Plotting and Identifying numerical columns which show significant difference in loan status
- Considering only such numerical columns for bivariate analysis

Bivariate Analysis

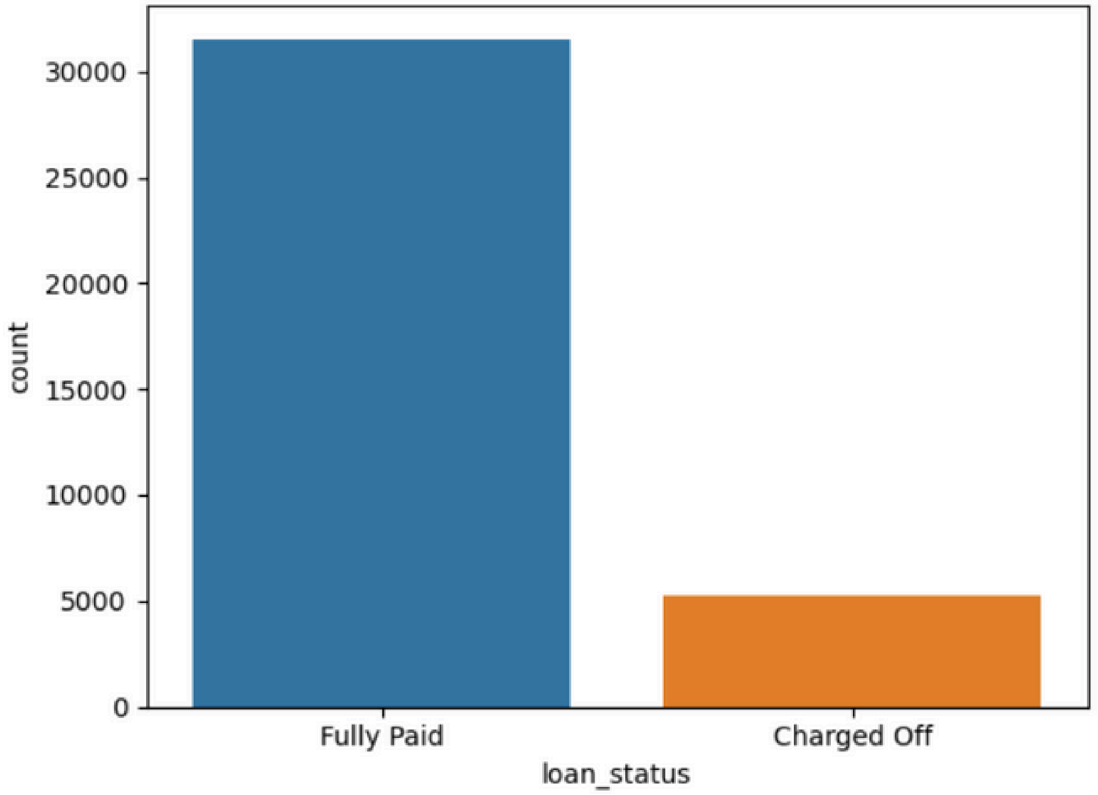
- Bivariate analysis between numerical vs numerical columns and identifying correlation
- Bivariate analysis between numerical vs categorical columns.

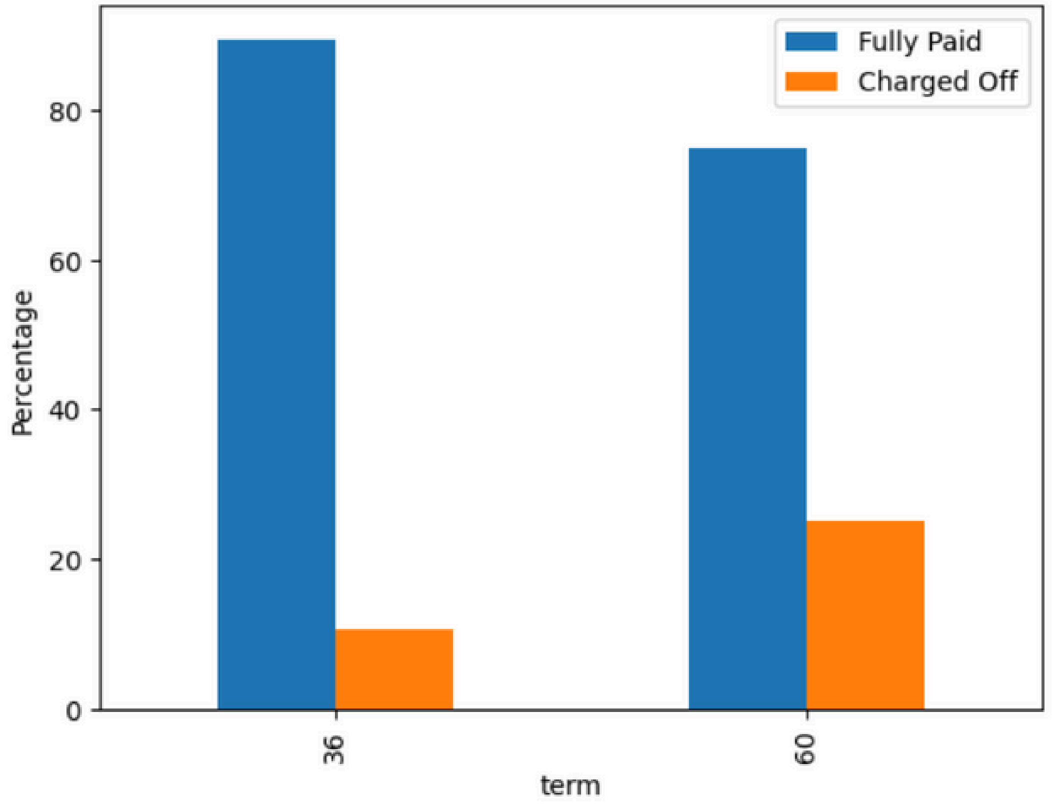
Summary

- Summarizing insights from all the plots
- Noting down the mild and strong drivers behind loan-default based on the analysis and plots.

Univariate Results

```
loan_status
Fully Paid     31534
Charged Off     5266
Name: count, dtype: int64
```





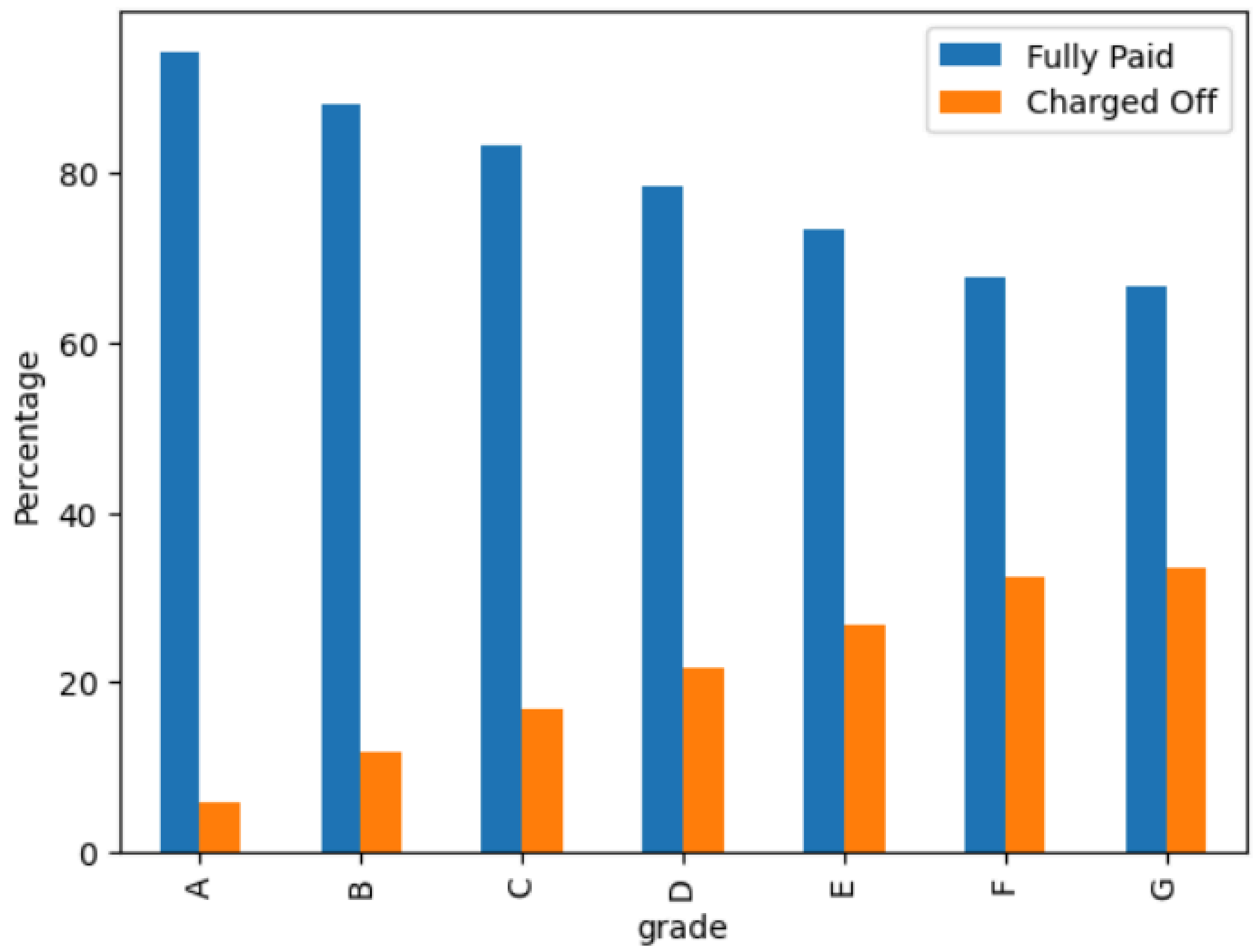
Loan Status

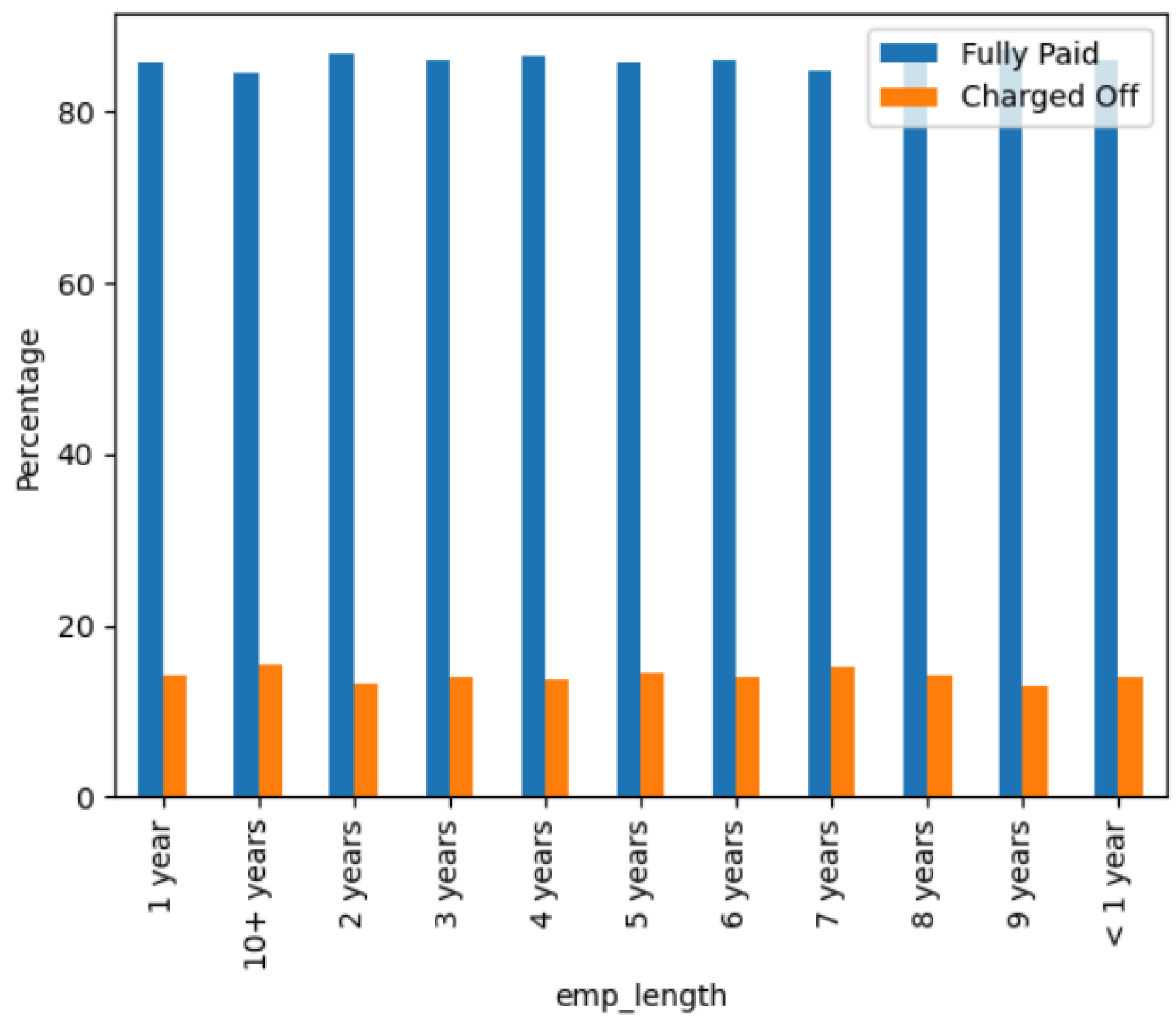
Most of the loans are fully paid however around 15% of the loans are charged off.

Loan Term

Percentage of charged off loans is higher for loan term of 60 months.

Univariate Results





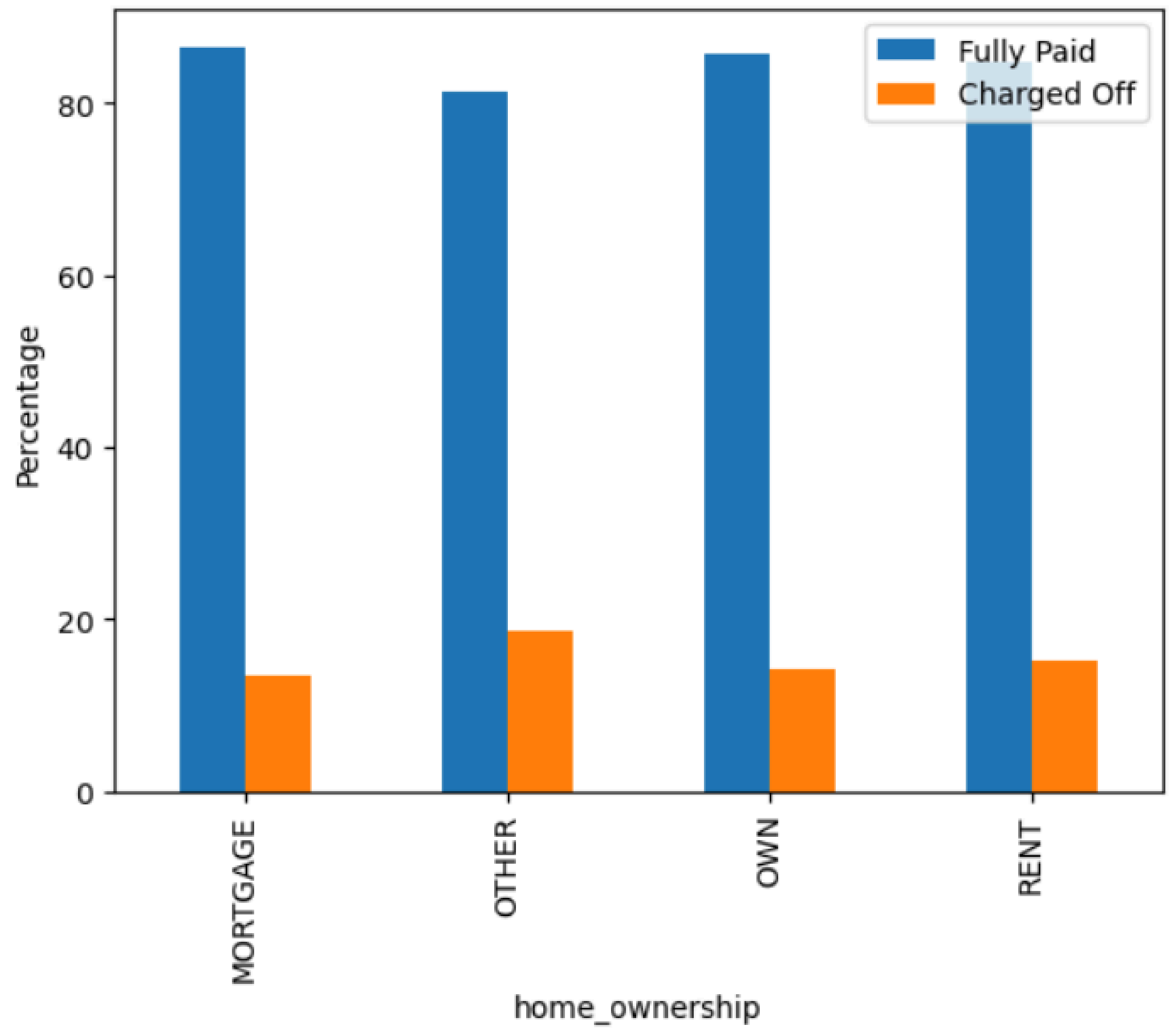
Loan Grade

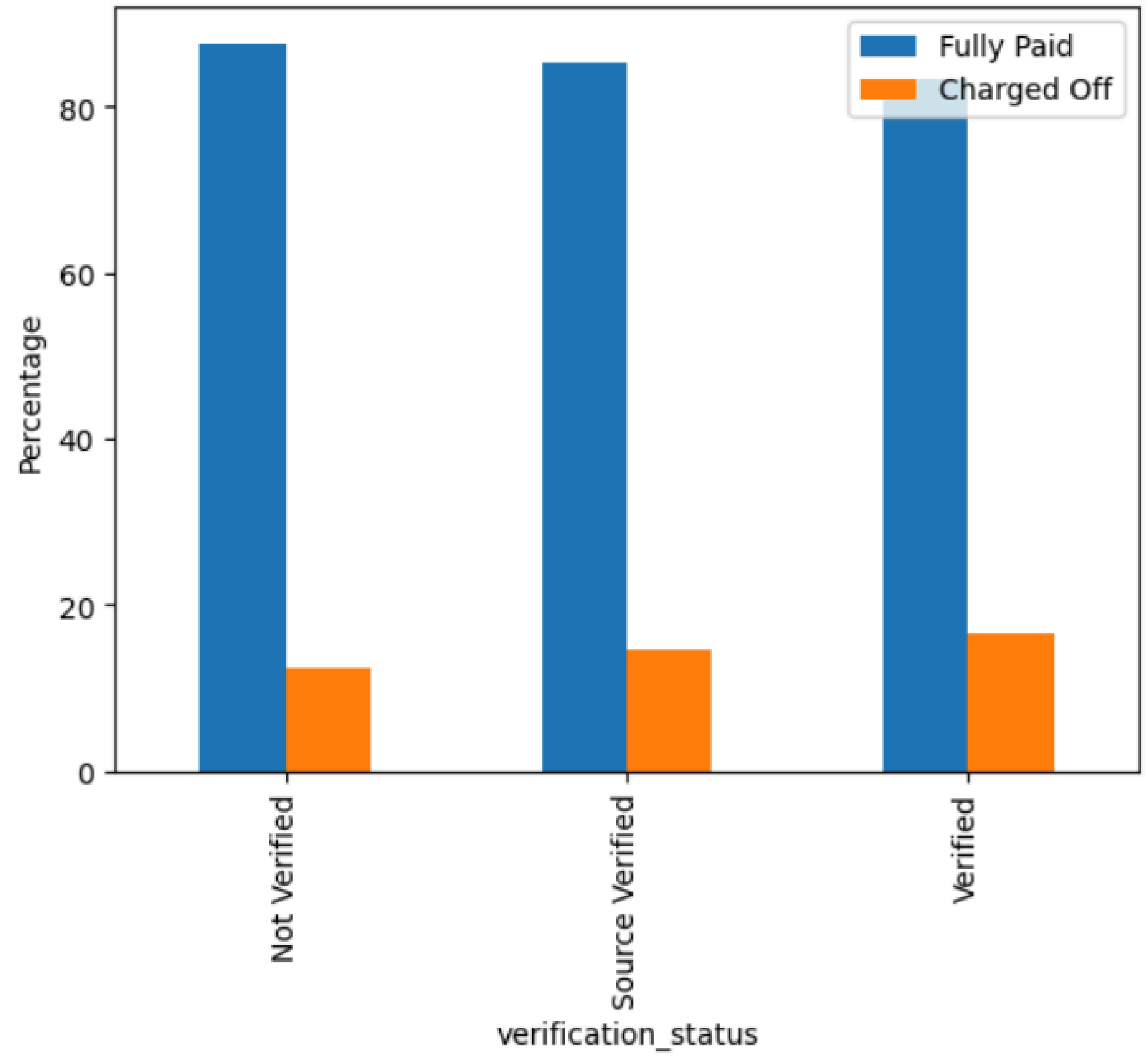
Percentage of charged off loans is higher for lower grades of loans and lower for higher grades of loan

Employment Length

Percentage of charged off loans is less than 20% for each experience level.

Univariate Results





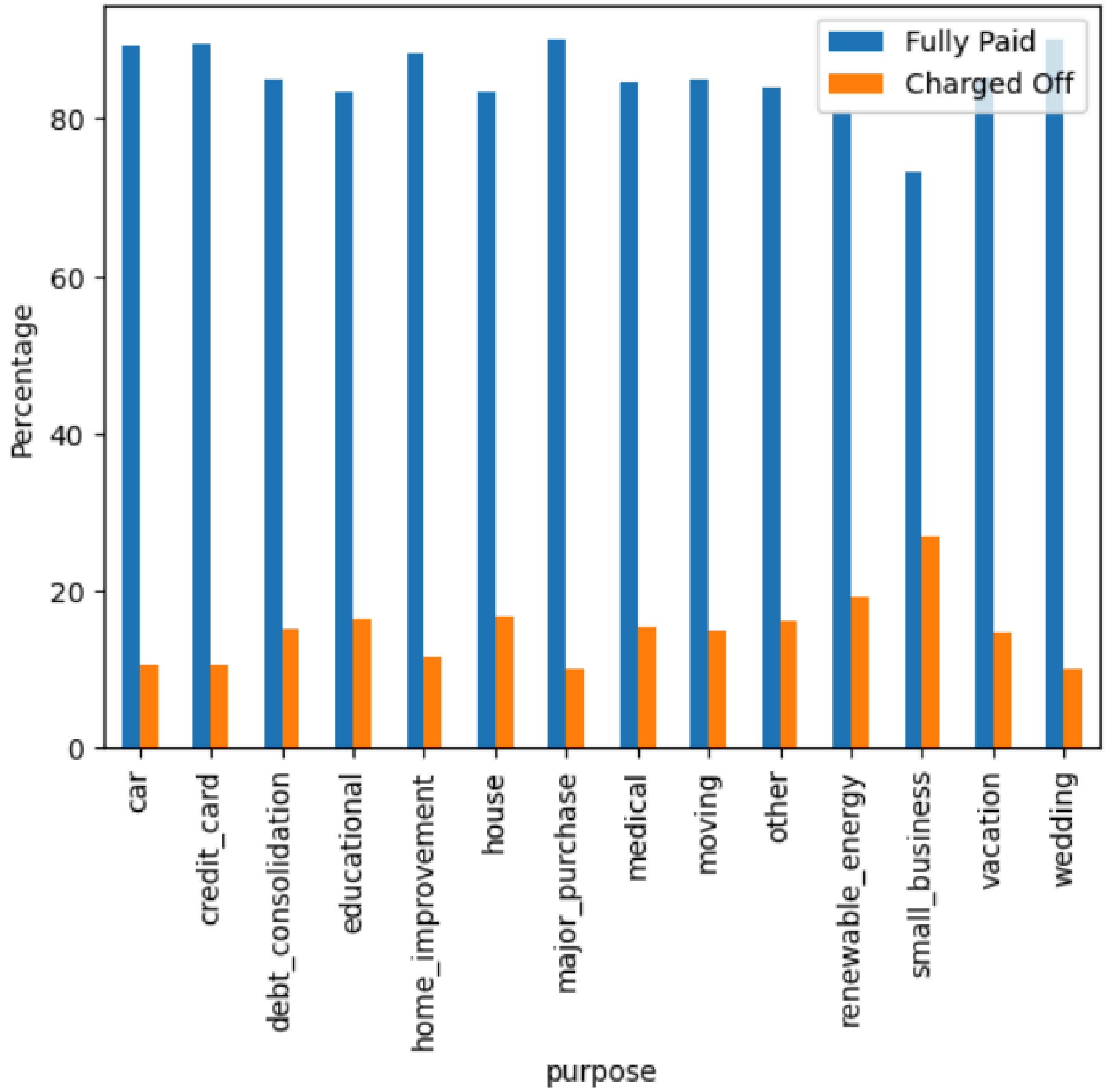
Home-Ownership

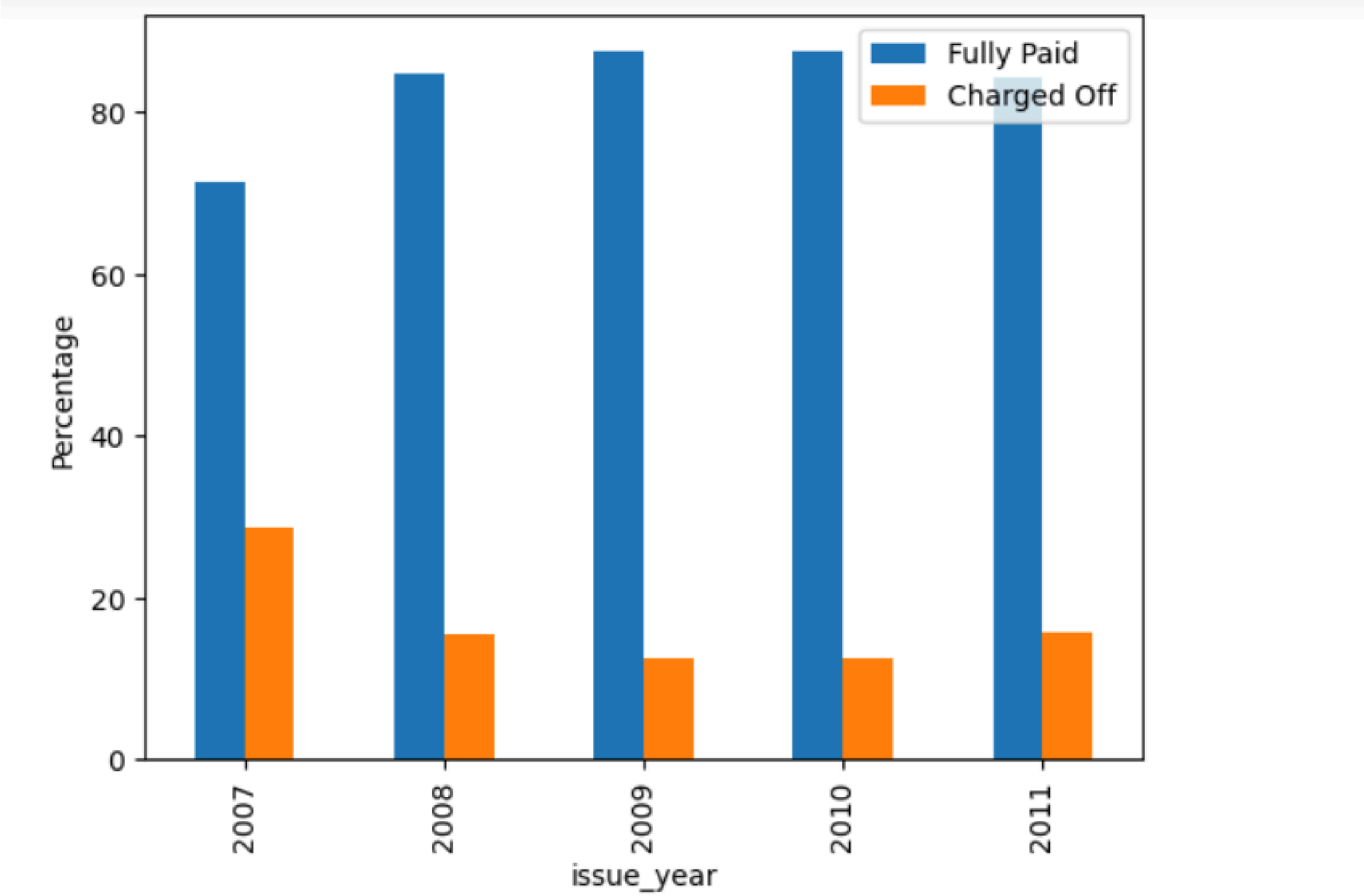
Percentage of charged off loans is more for other and mortgage category followed by rent and own.

Verification Status

Percentage of charged off loans is less than 20% for each verification status.

Univariate Results





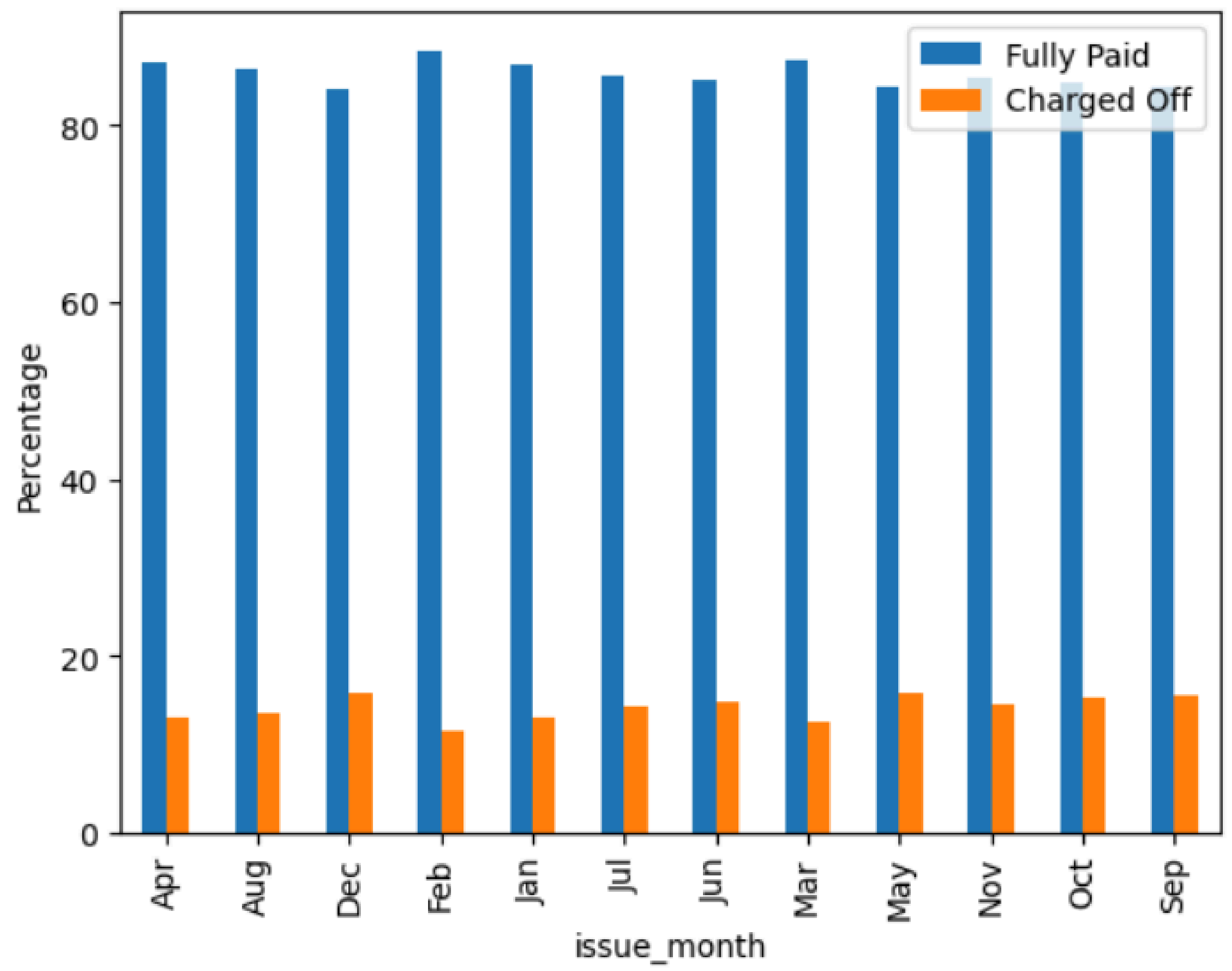
Purpose

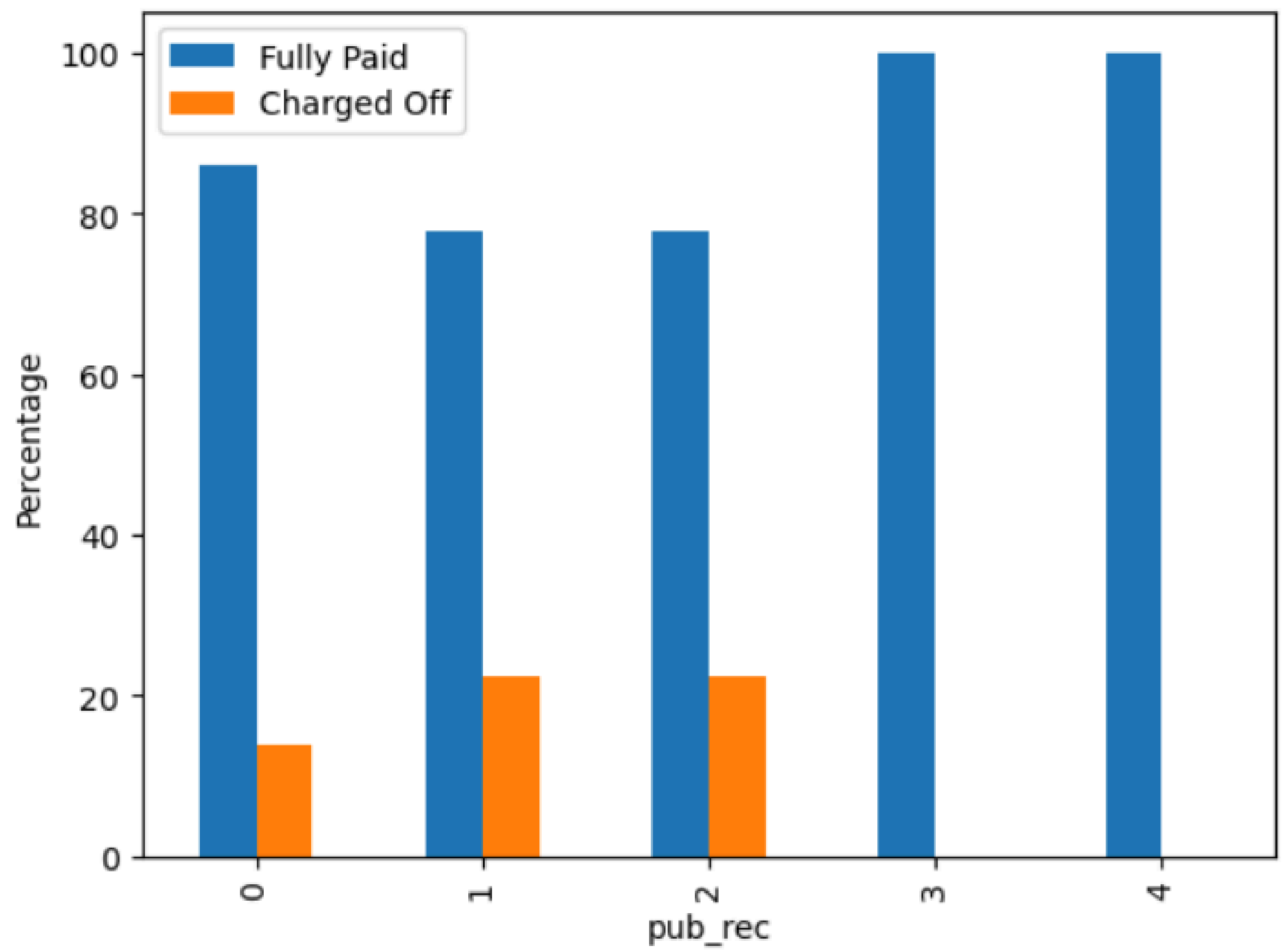
Percentage of charged off loans are high for purposes like small_business, renewable energy, educational, house and debt_consolidation.

Issue Year

Percentage of charged off loans have somewhat decreased over the years but again increased in 2011.

Univariate Results





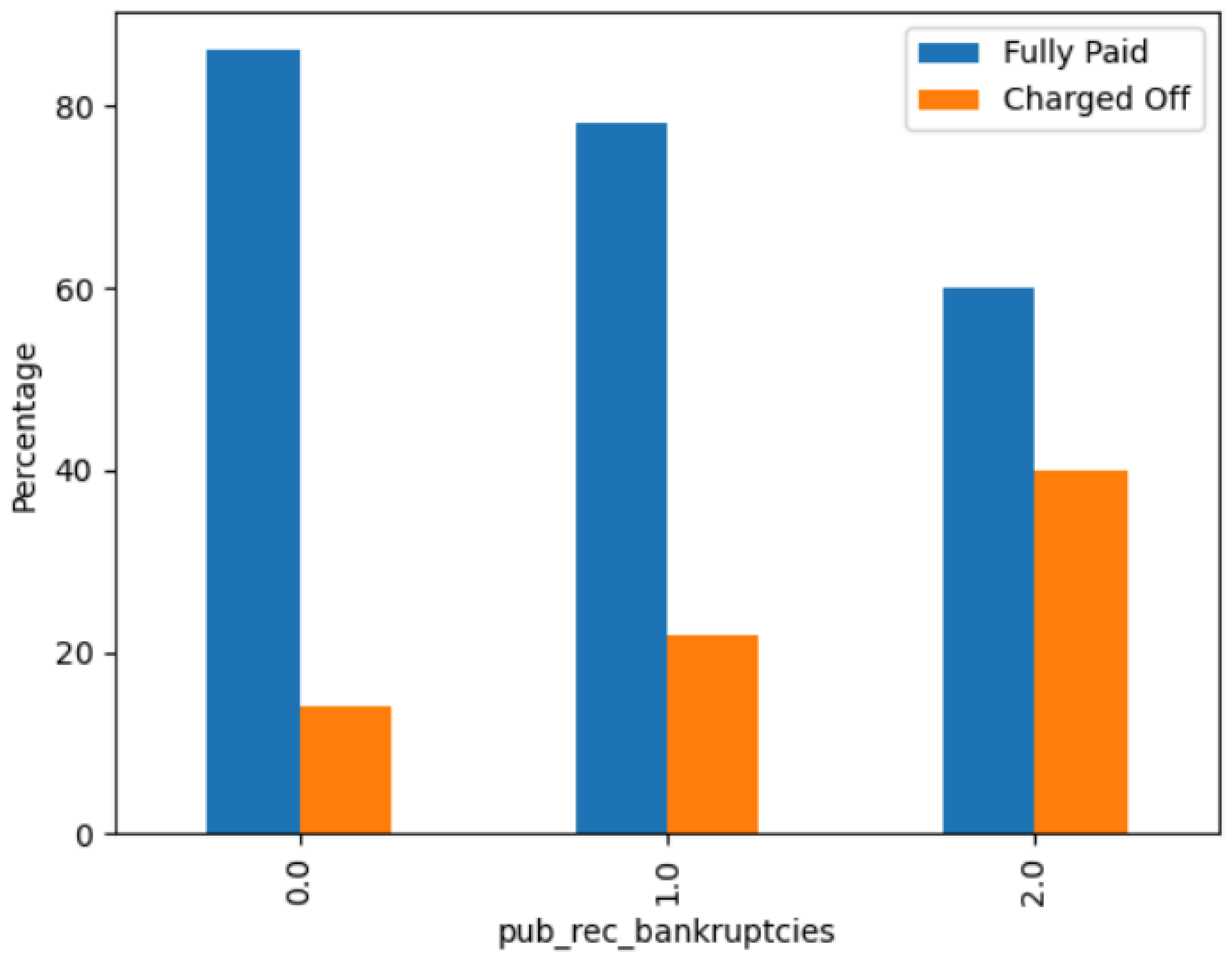
Issue Month

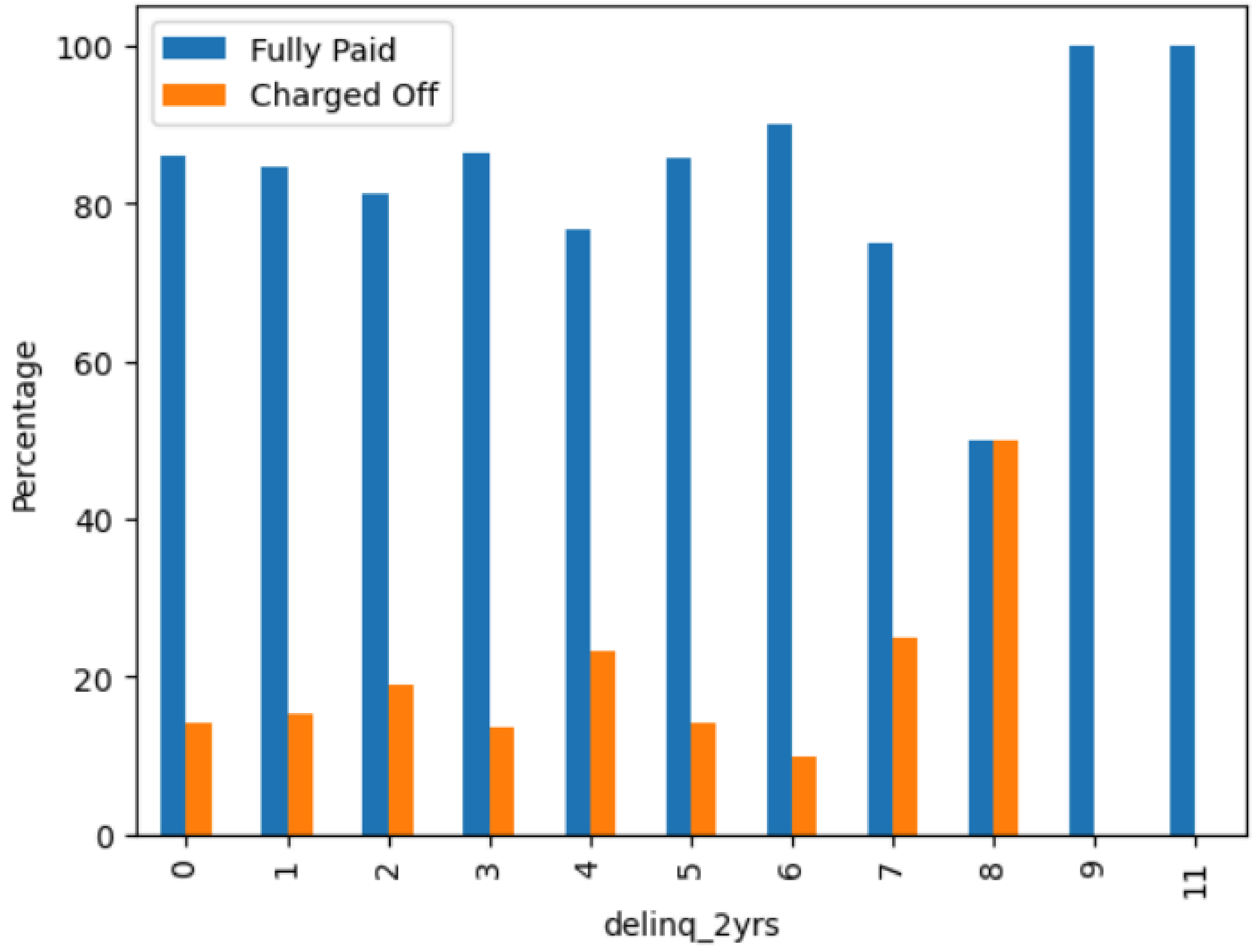
Percentage of charged off loans is higher in December, May and September compared to other months.

Number of derogatory public records

Percentage of charged off loans is higher for people who have 1 or 2 derogatory records.

Univariate Results





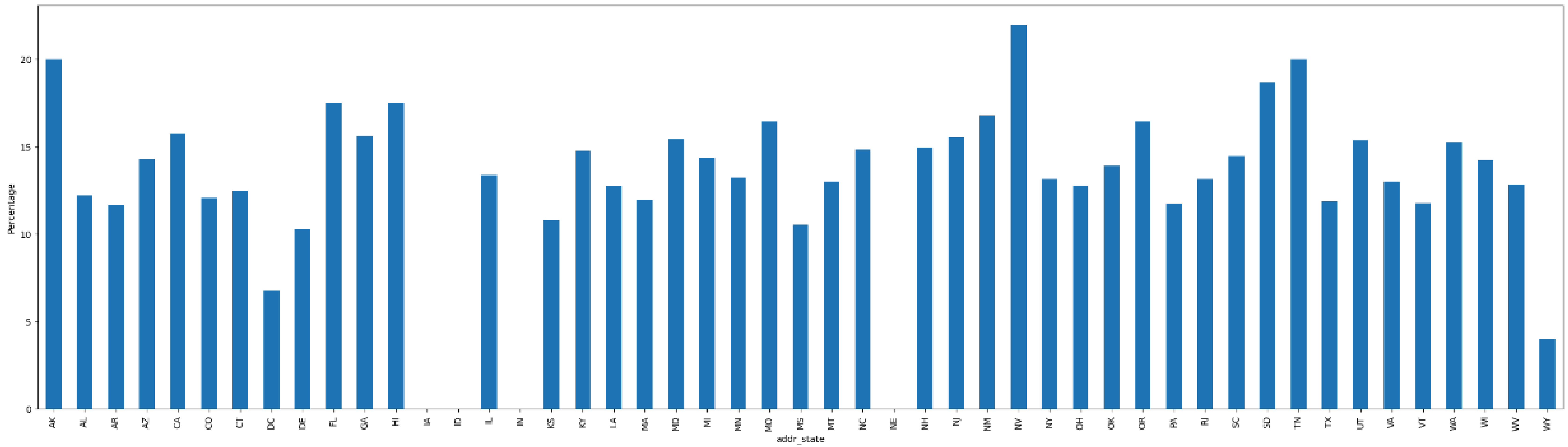
Number of public record bankruptcies

Percentage of charged off loans is higher for people with 1 and 2 bankruptcy records.

Incidences of delinquencies

Percentage of charged off loans follows an overall upward trend with an increase in delinquency record.

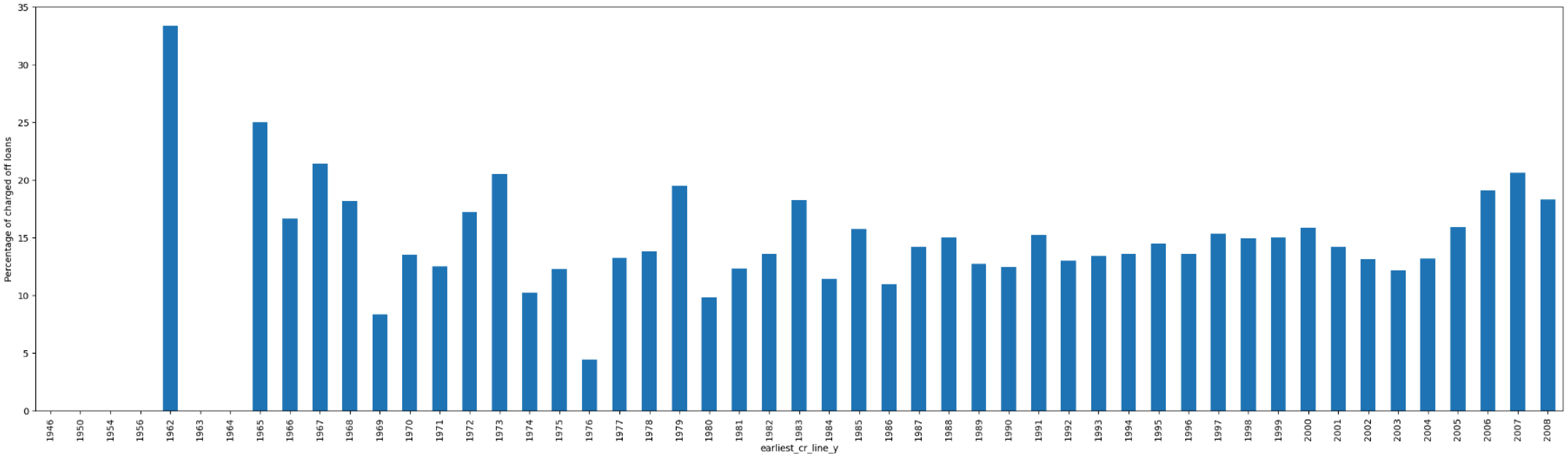
Univariate Results



Address state

Percentage of charged off loans are higher for states having code NV, AK , TN.

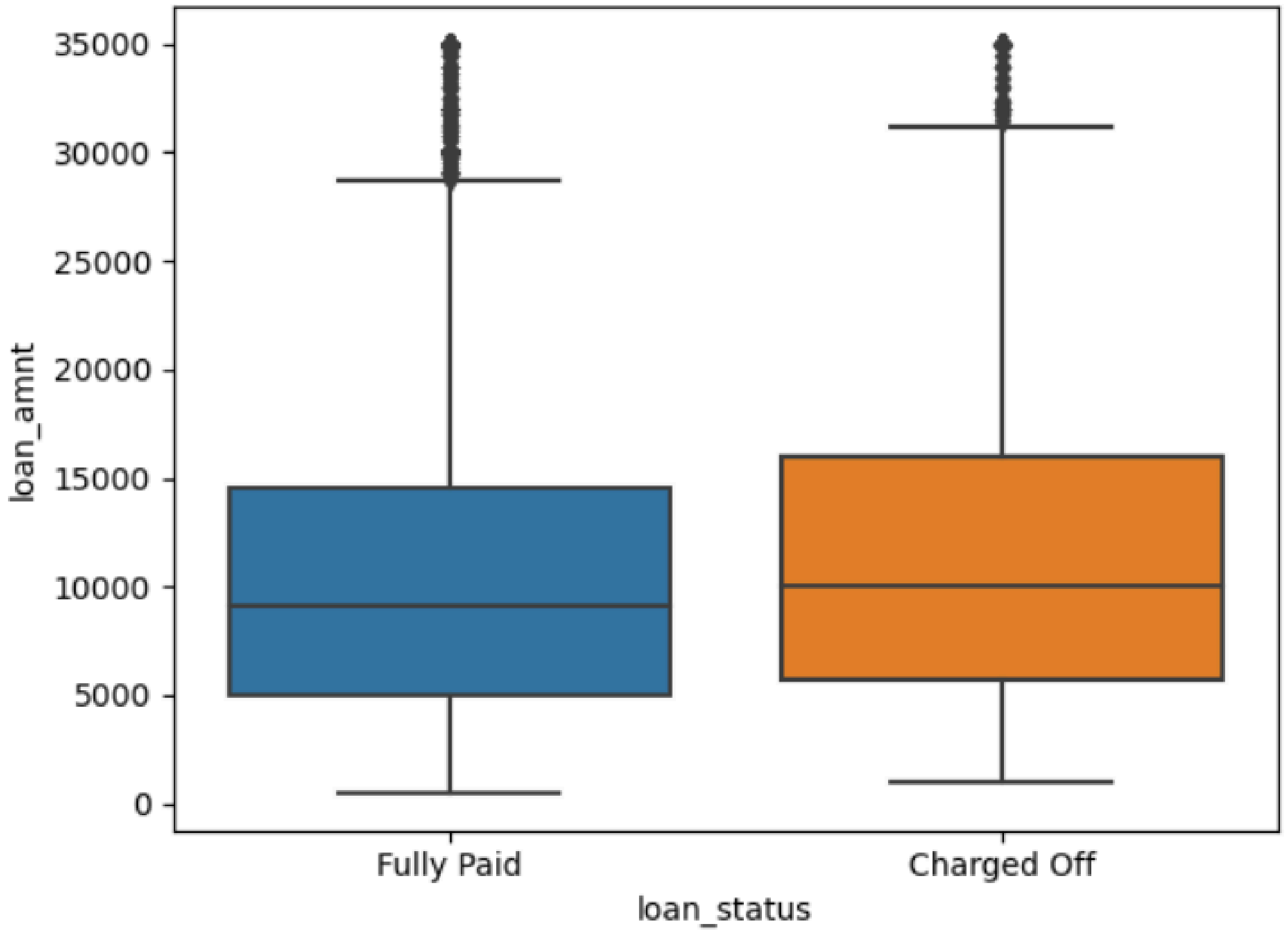
Univariate Results



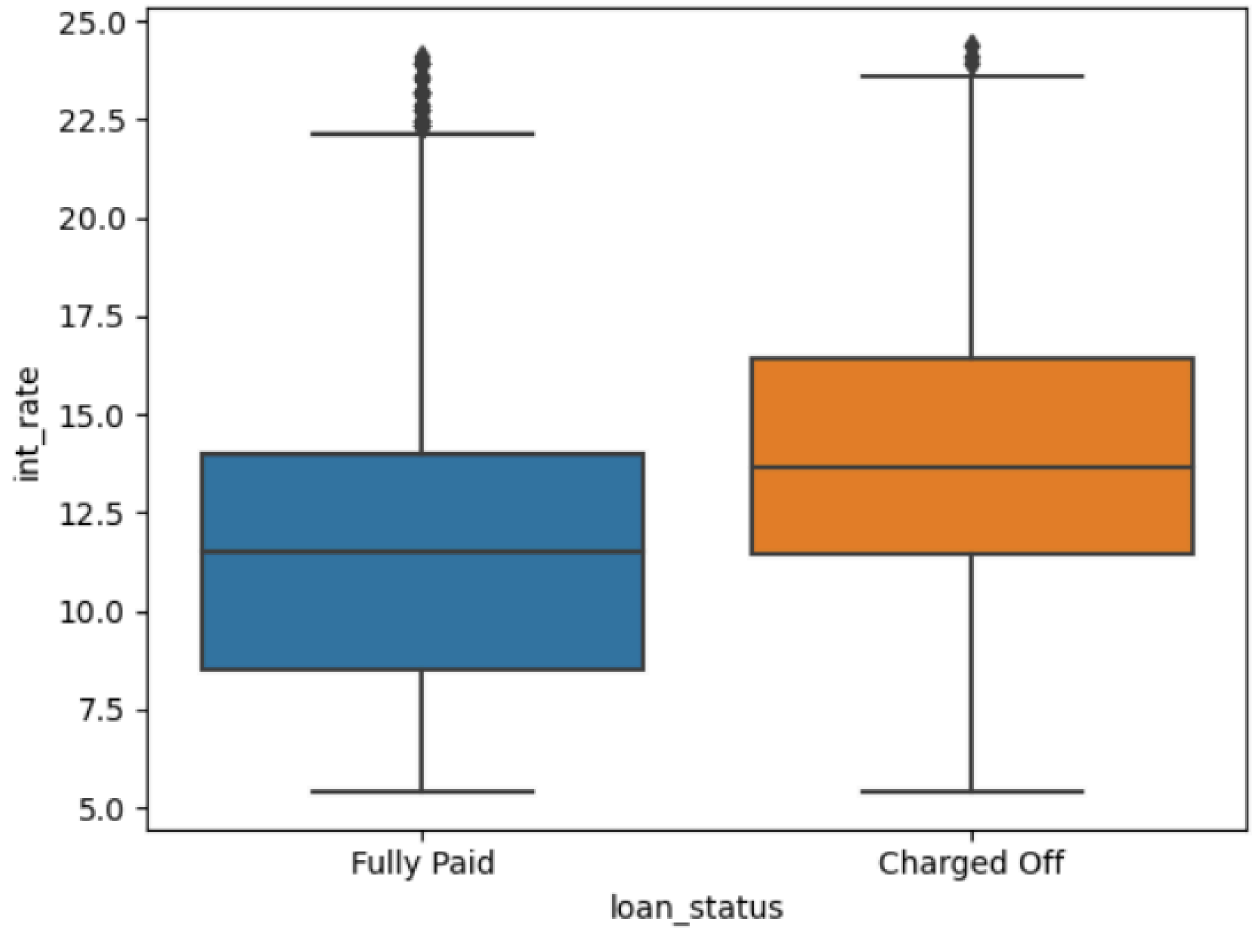
Earliest credit line year

Percentage of charged off loans show a sharp increase for people whose credit line starting year was in range 2003-2007. However highest Percentage of charged off loans is for the credit line year 1962.

Segmented Univariate Results

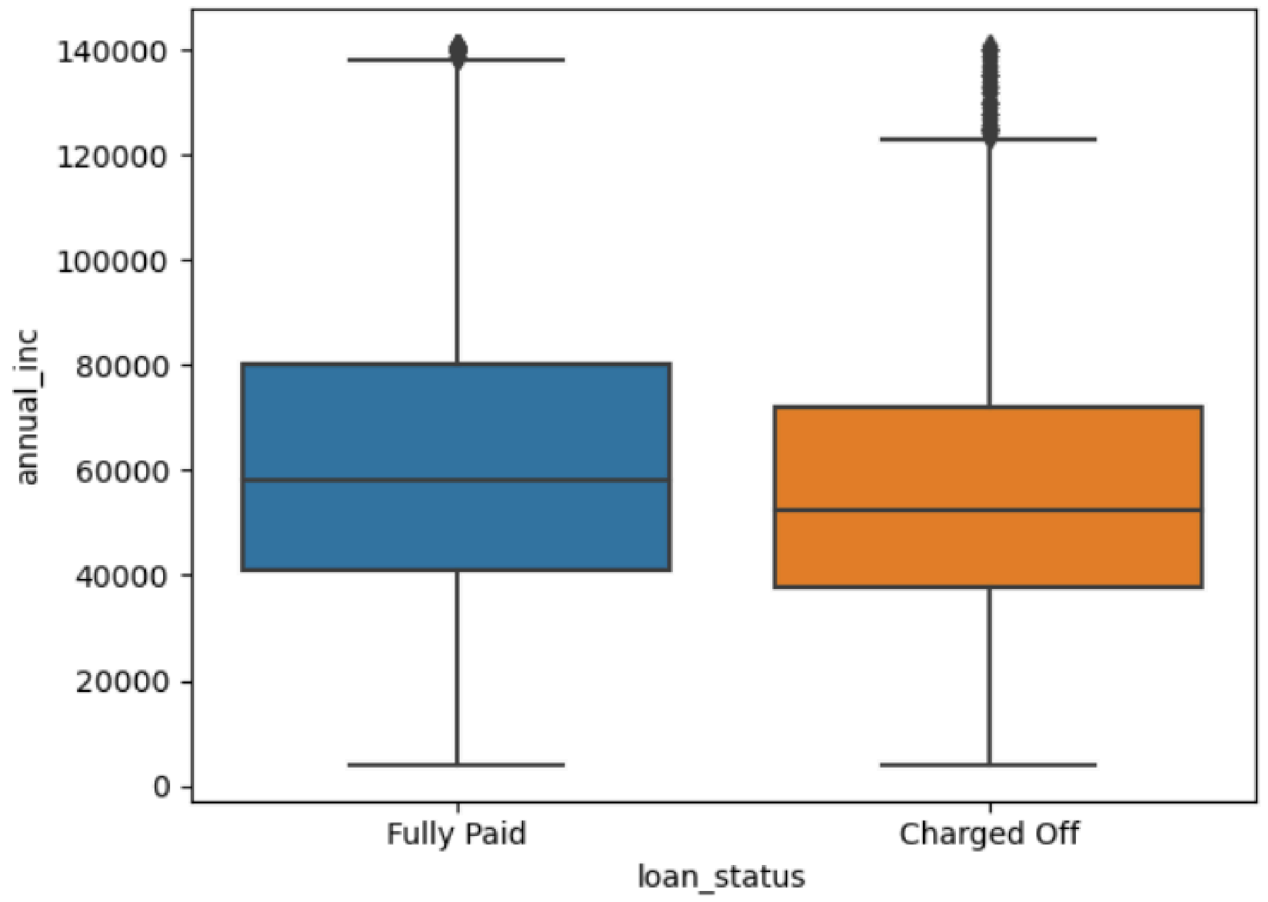


Loan amount, Funded amount and Funded amount by investors are high for people who are defaulters

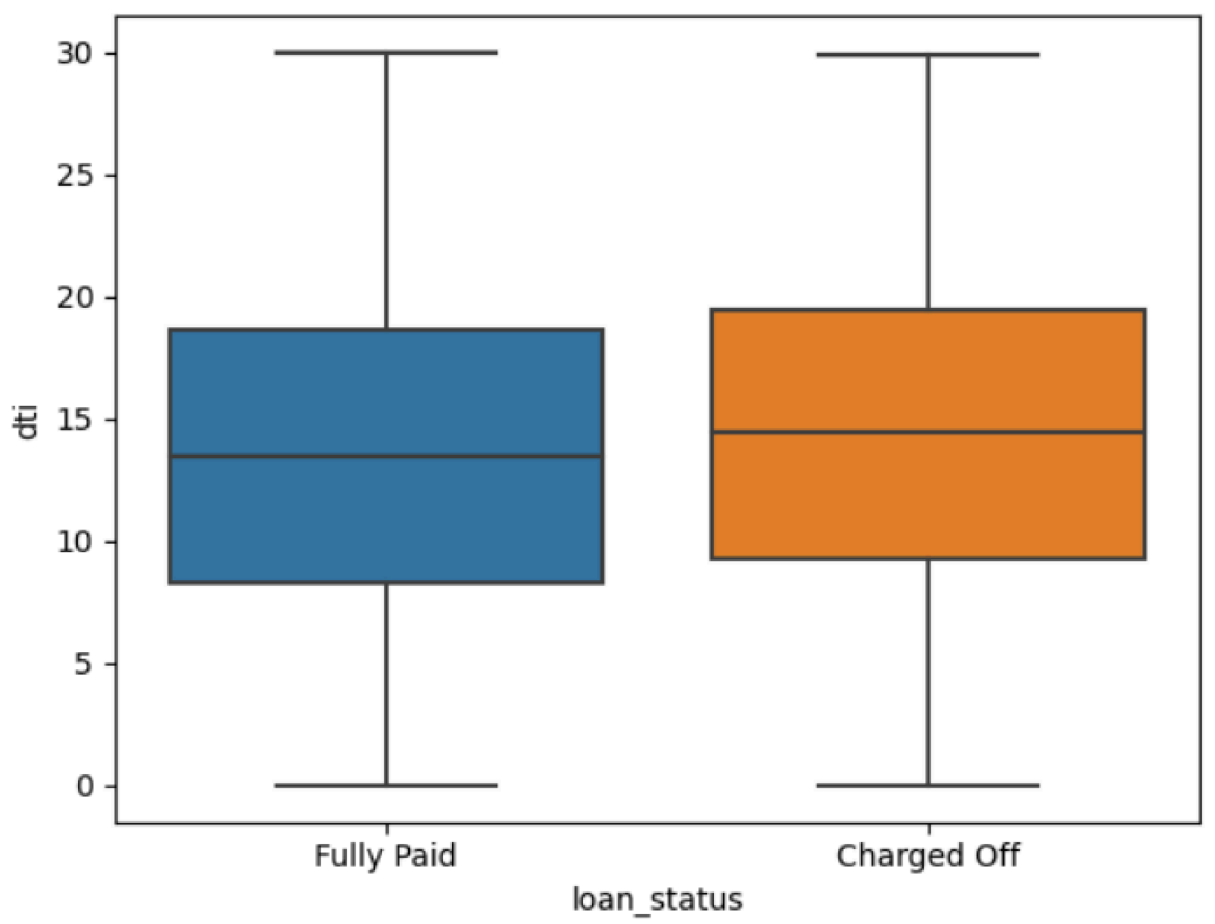


Interest Rate is typically more for the loans which are charged off.

Segmented Univariate Results

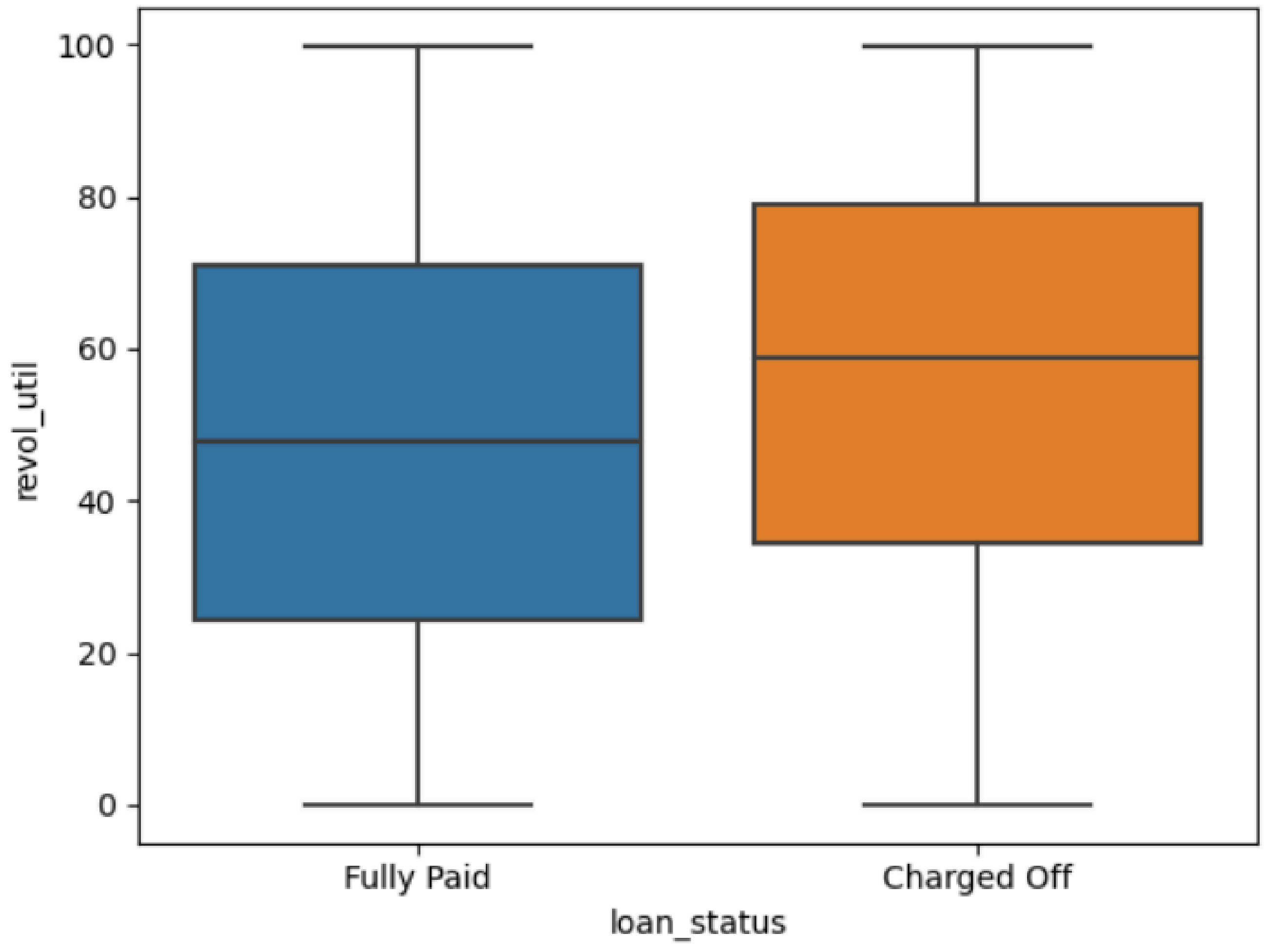


Annual Income is slightly lower for loans which are charged off.



Dti ratio is slightly higher for loans which are charged off.

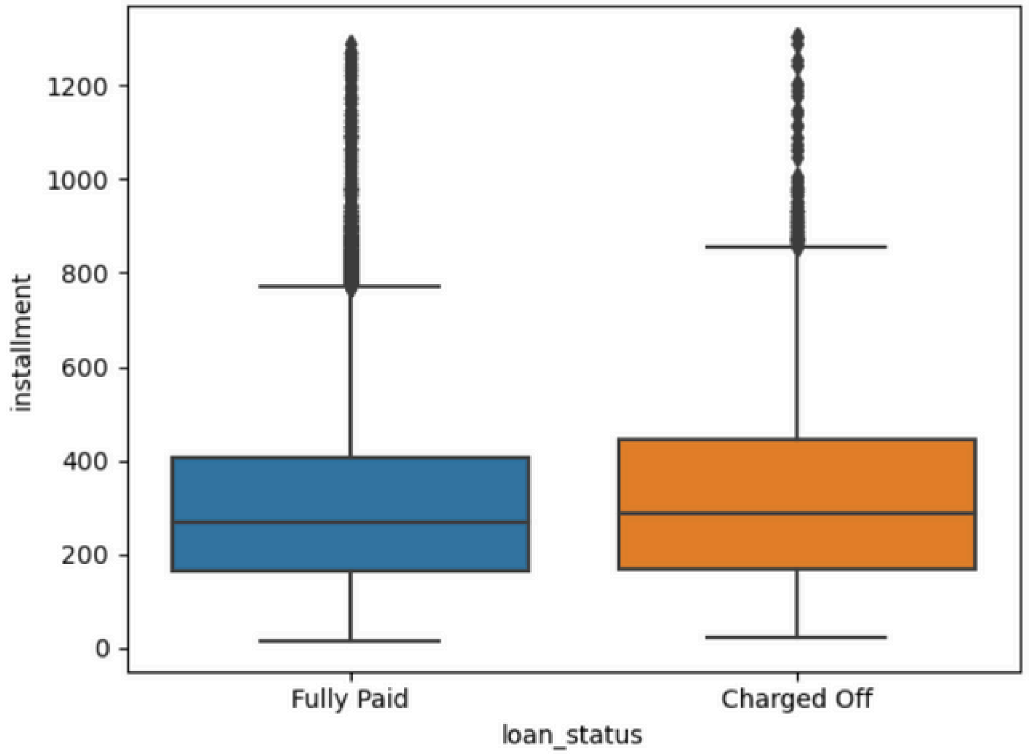
Segmented Univariate Results

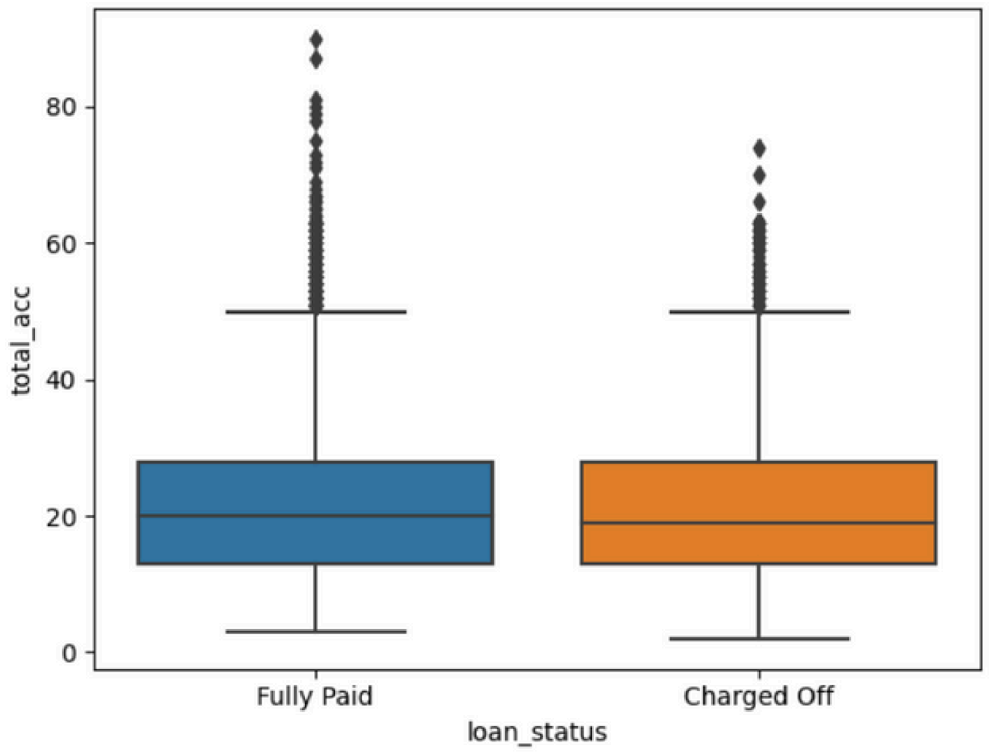


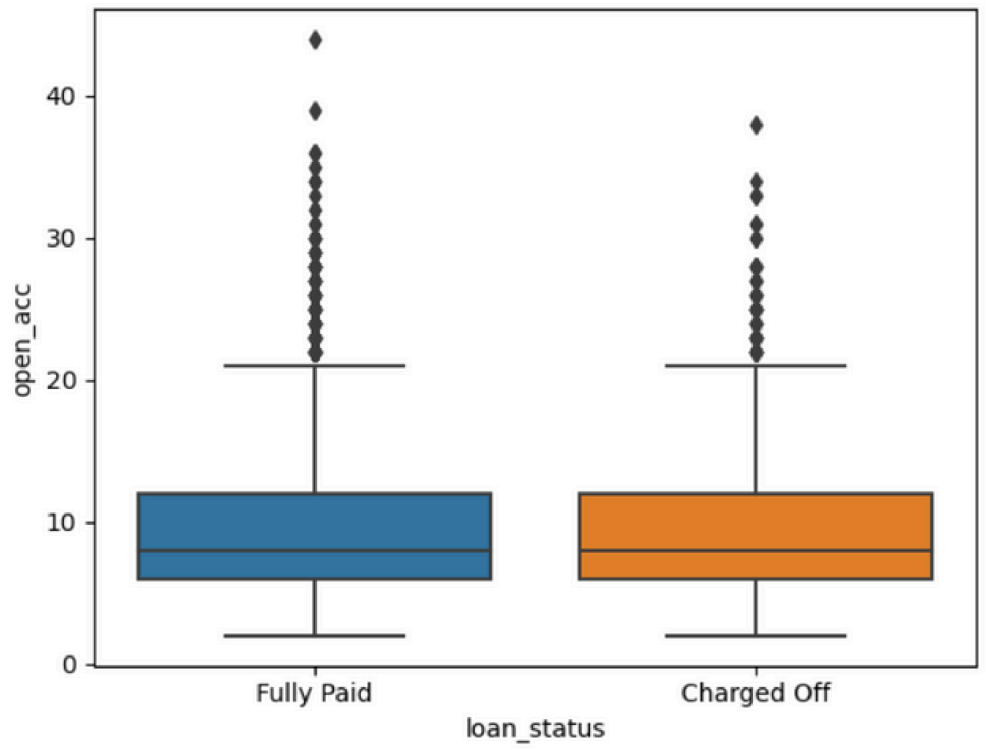
Revolving Utilization is higher for loans which are charged off.

Segmented Univariate Results

Numerical columns : **installment, open accounts and total accounts** does not show much difference for charged off vs fully paid loans.
So we will not be considering them for our bivariate analysis.



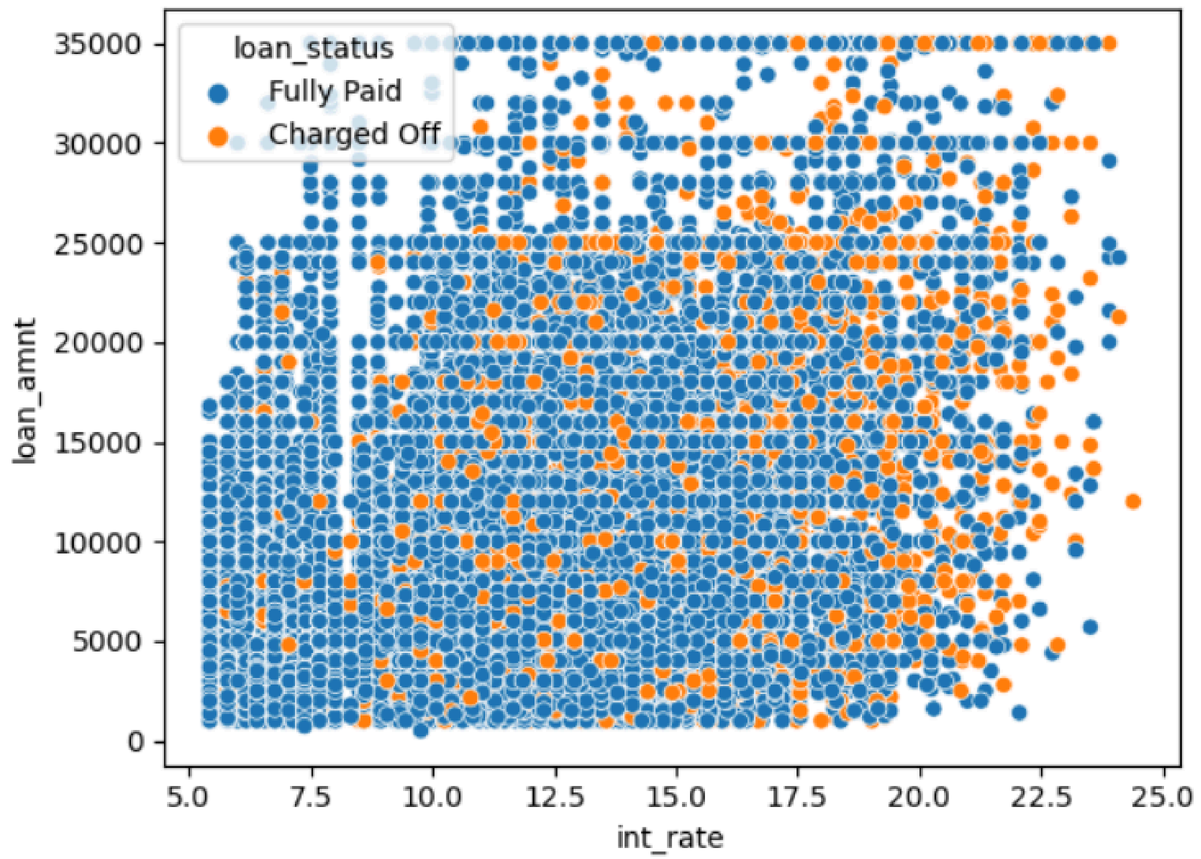




Bivariate Results

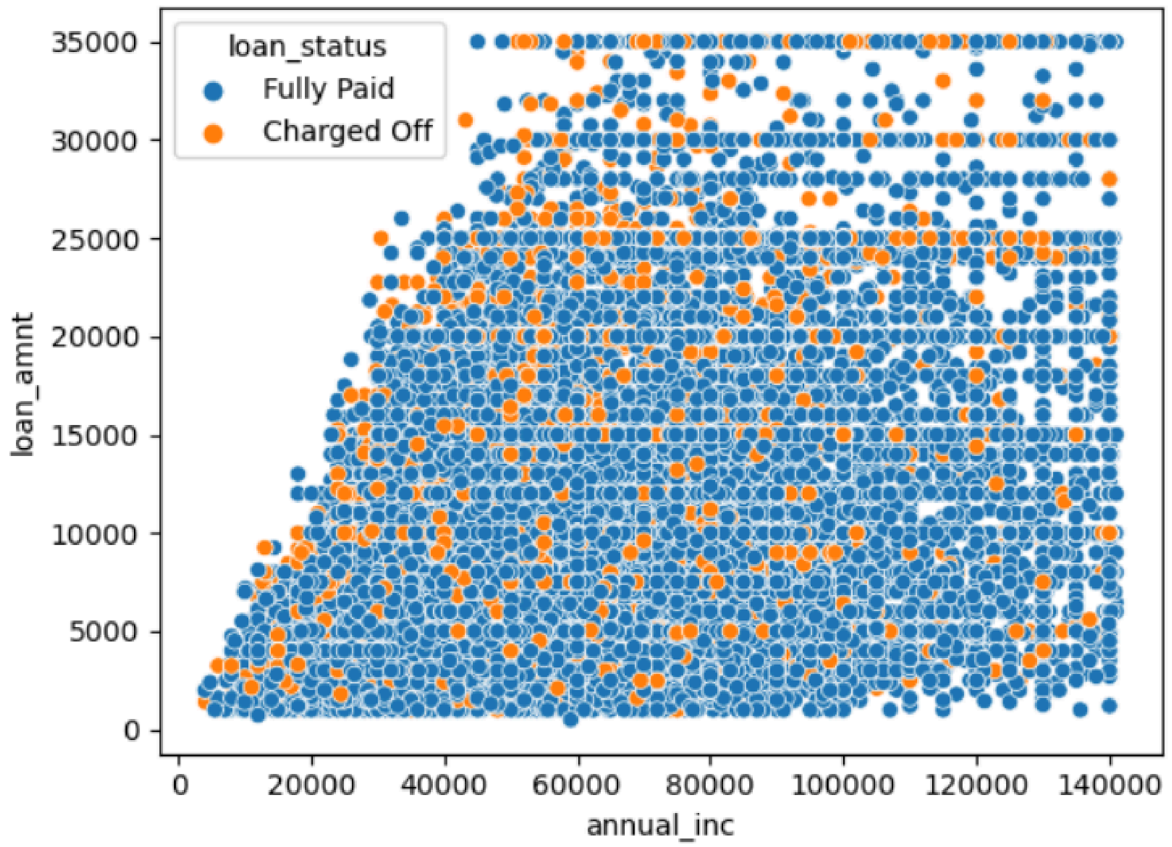
Numerical vs Numerical

```
Overall correlation : 0.28948053891058945
```



Loan amount vs interest rate: There is a small correlation between interest rate and loan amount. Higher the loan amount higher the interest rate. Also there are large number of charged off loans in the high interest high amount range

```
Overall correlation : 0.3998738428954058
```

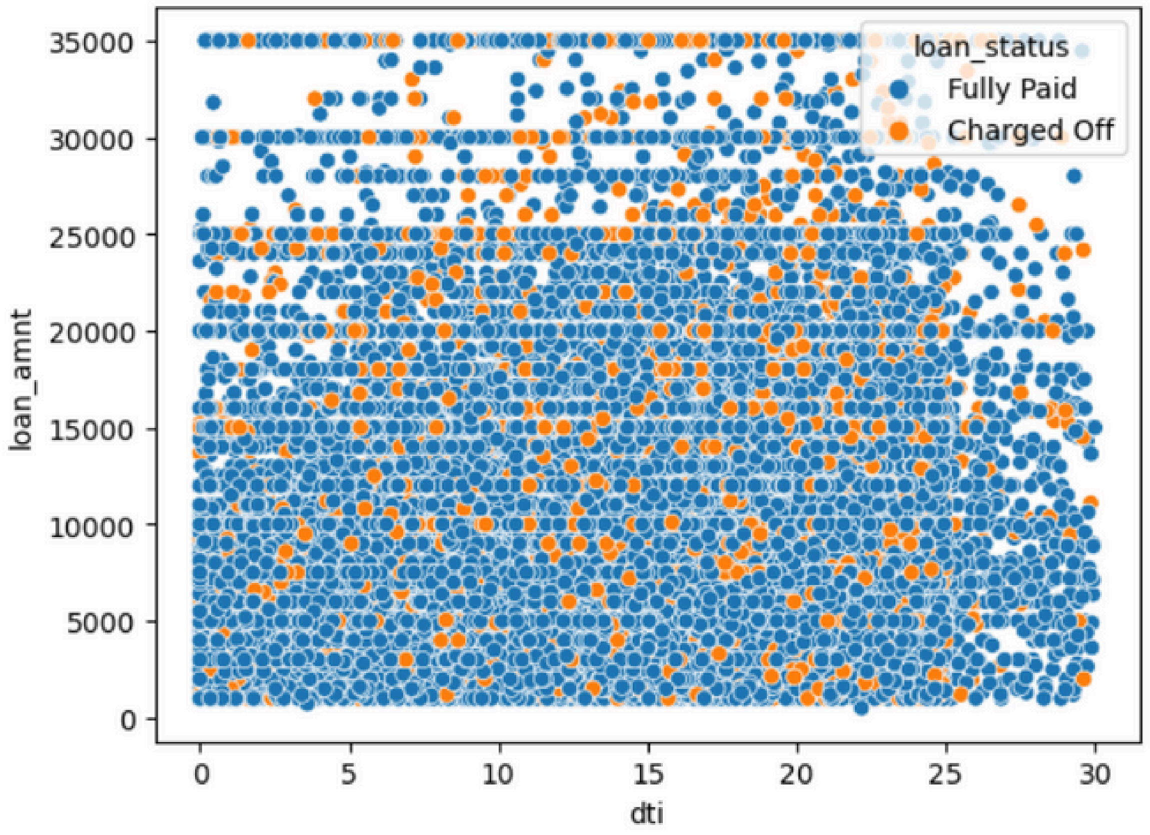


Loan amount vs annual income: In the income range of 0-40000, people with low income takes loan of lower amount.

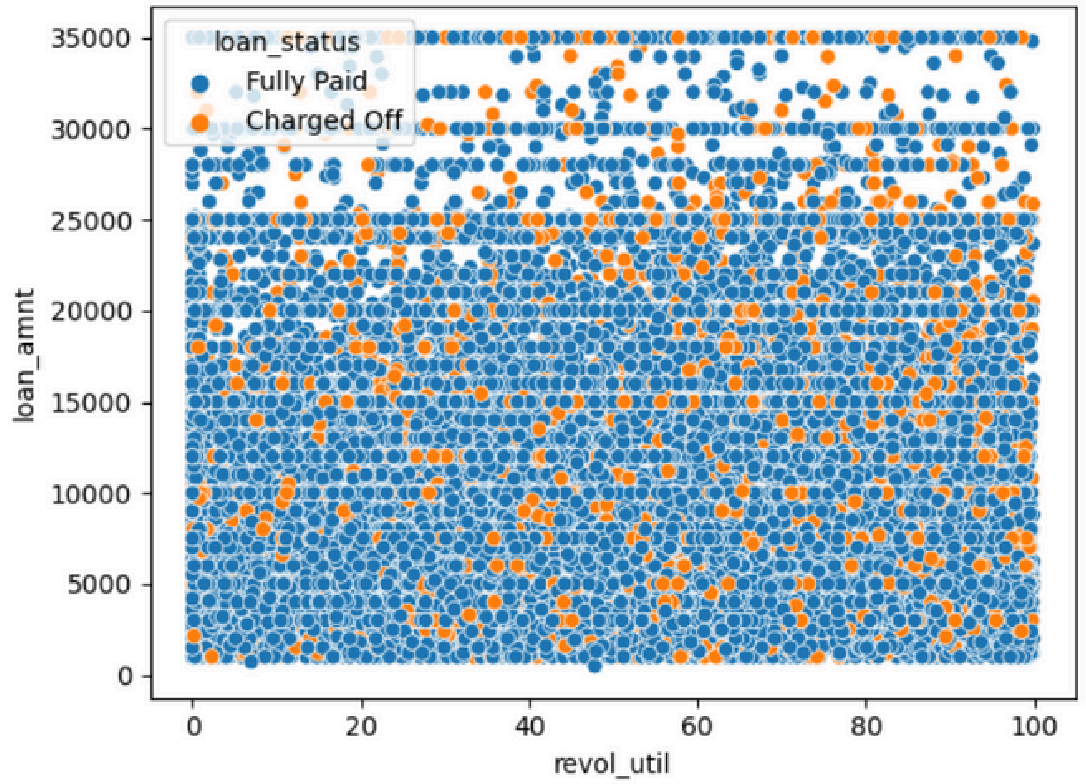
Bivariate Results

Numerical vs Numerical

```
Overall correlation : 0.08879651197941504
```




```
Overall correlation : 0.06795812156743913
```



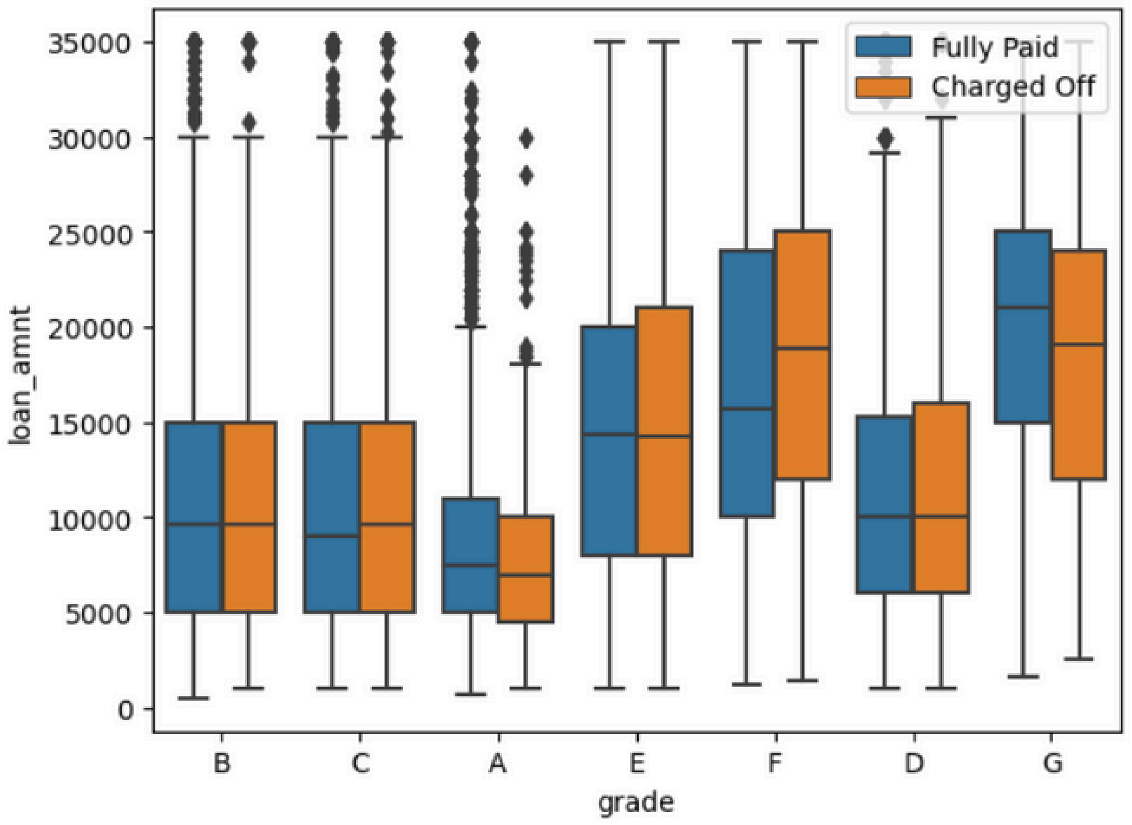
Loan amount vs dti: There is a weak correlation between these two columns. Also we can see people with high dti values are also getting high loan amounts.

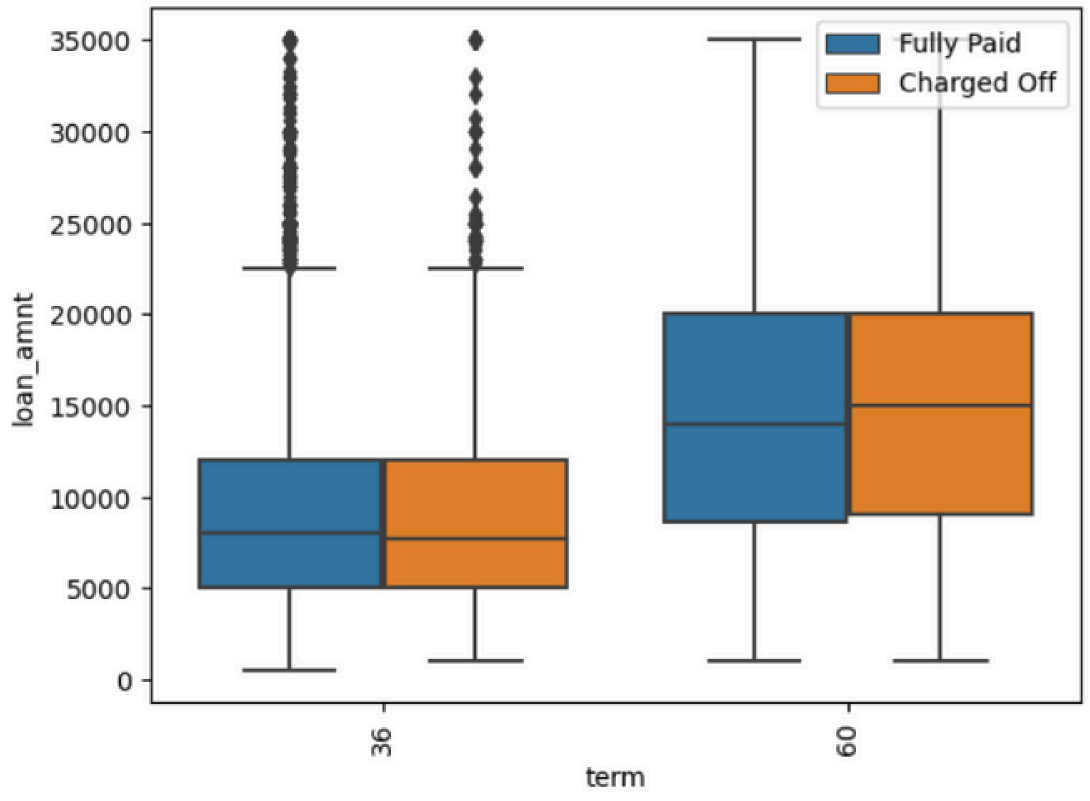
Loan amount vs revolving utilization: There is a weak correlation between these two columns. Also we can seepeople with high utilization values are also getting high loan amounts.

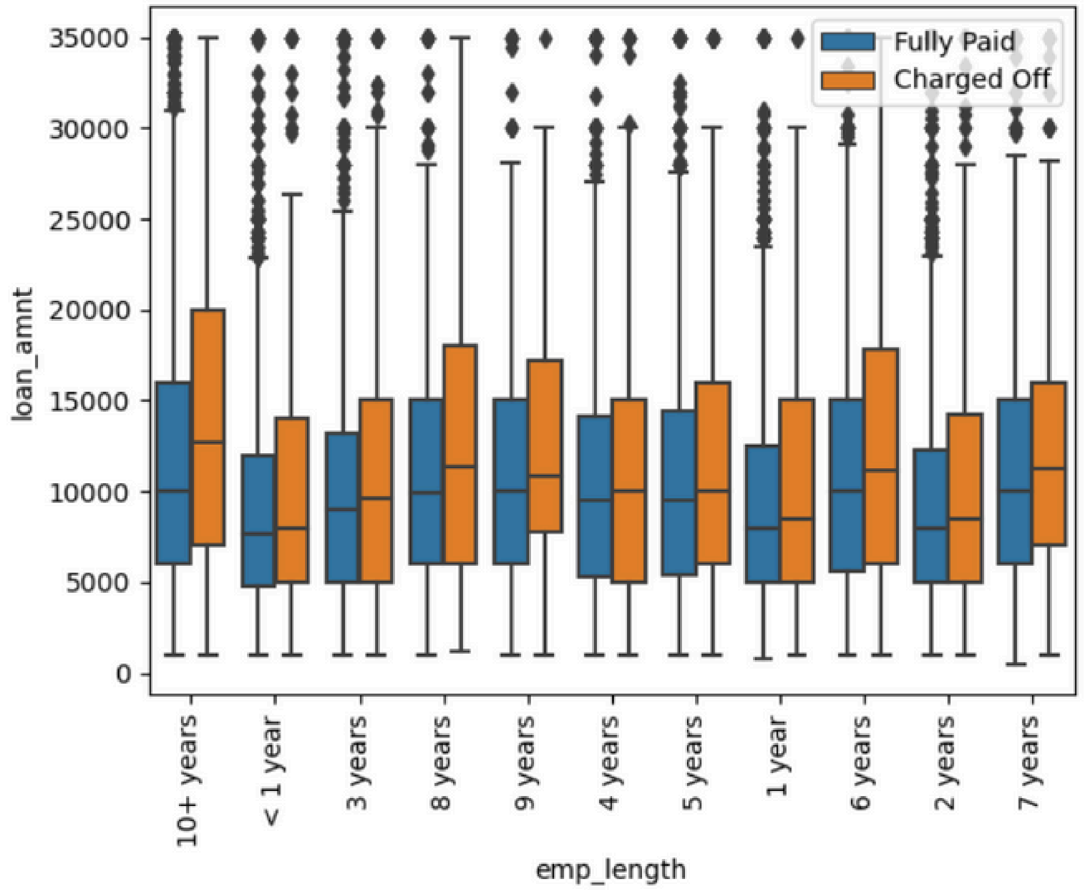
Bivariate Results

Numerical vs Categorical

We will consider loan amount in numerical field because loan amount is directly related to amount of loss/profit based on loan default







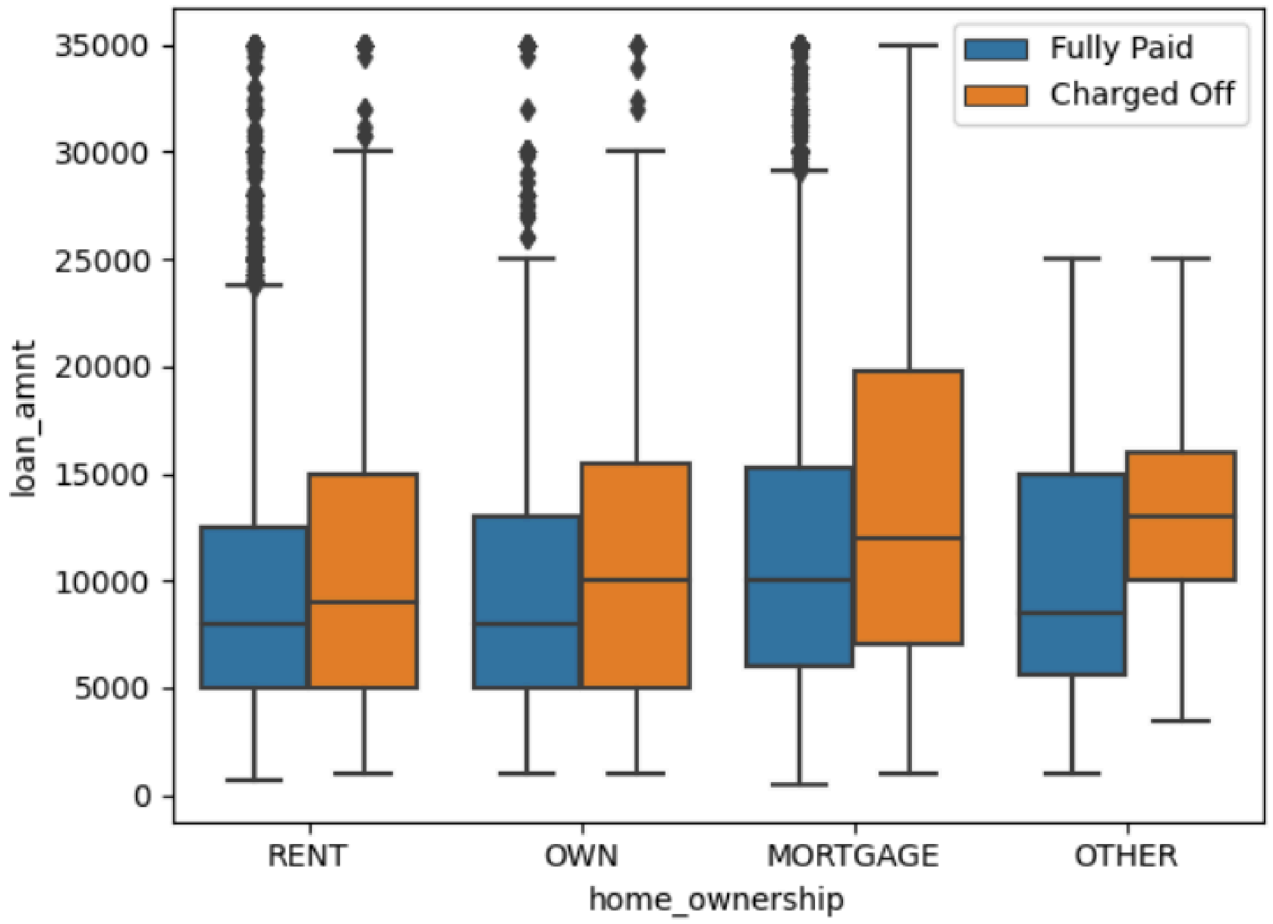
Loan amount vs Grade: Charged off loan amounts are higher for lower grades D, E and F

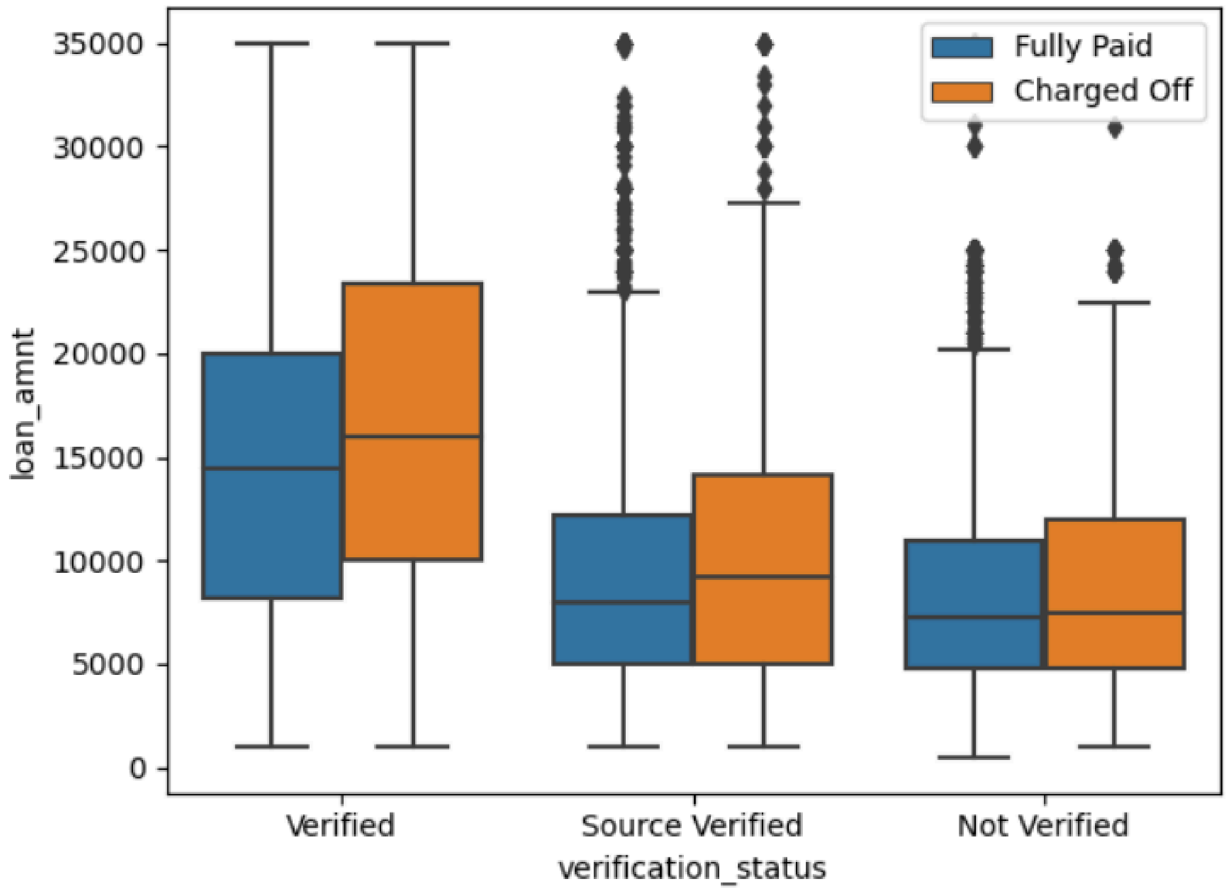
Loan amount vs Term: Loan amounts are overall higher for higher term but there is no significant difference between charged off and fully paid loan amounts for each term

Loan amount vs Employment Length: Loan amounts for higher experience levels are slightly higher but no significant trend observed

Bivariate Results

Numerical vs Categorical



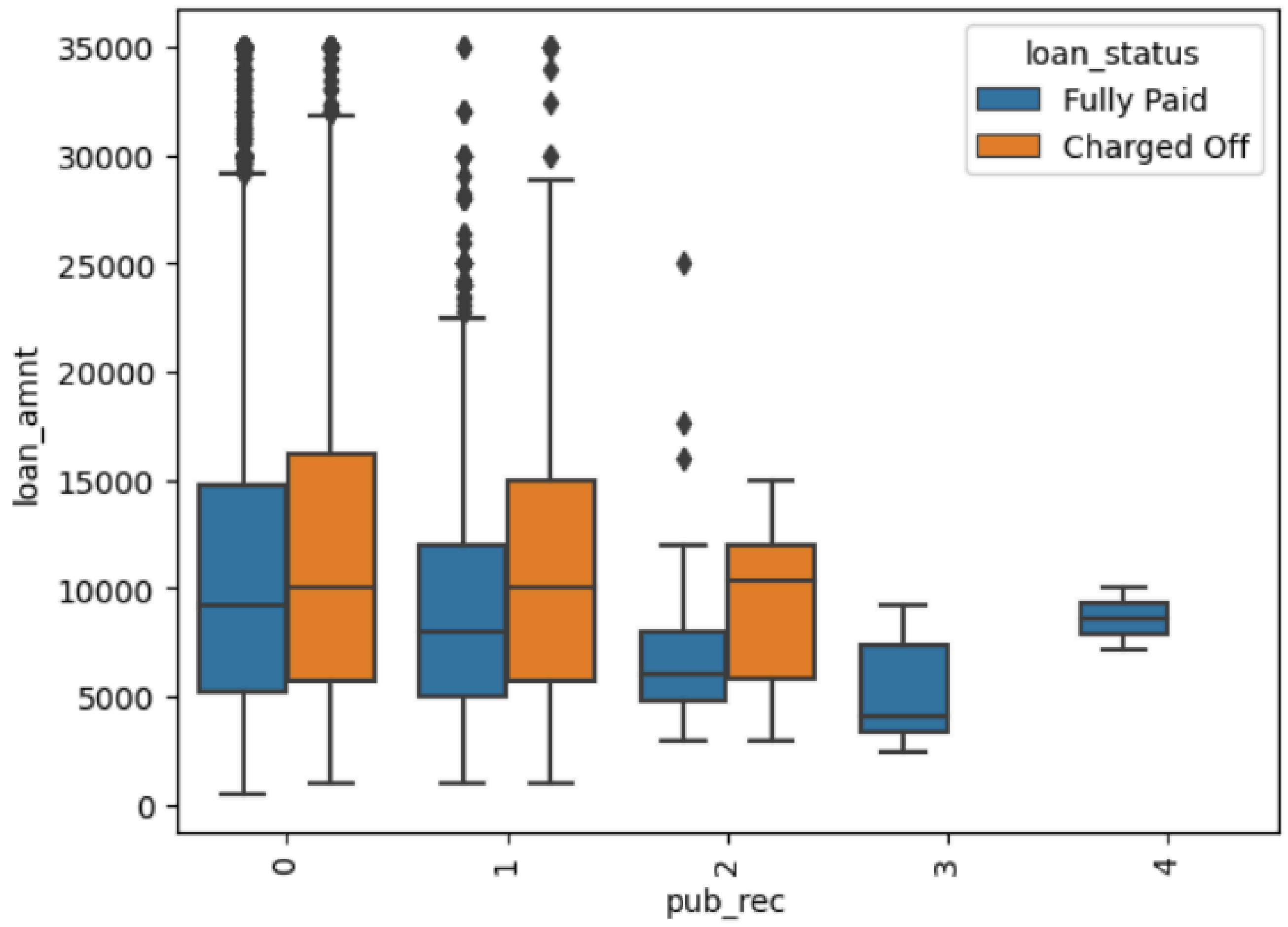


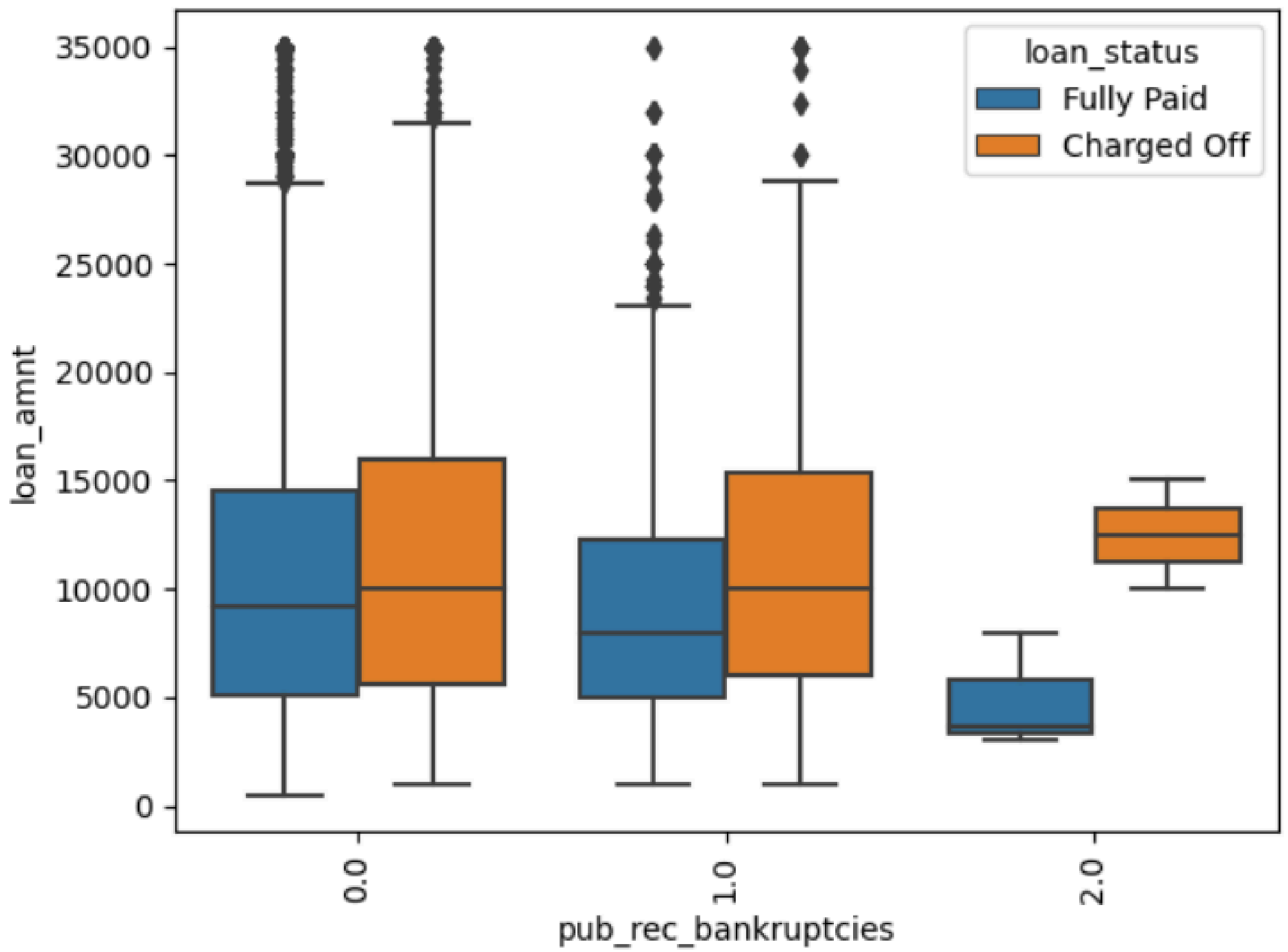
Loan amount vs Homeownership:
Loan amounts are quite higher for Mortgage home and also charged off loans are higher for Mortgage home

Loan amount vs Verification status:
Loan amounts are higher for verified applicants followed by source verified and not verified.

Bivariate Results

Numerical vs Categorical



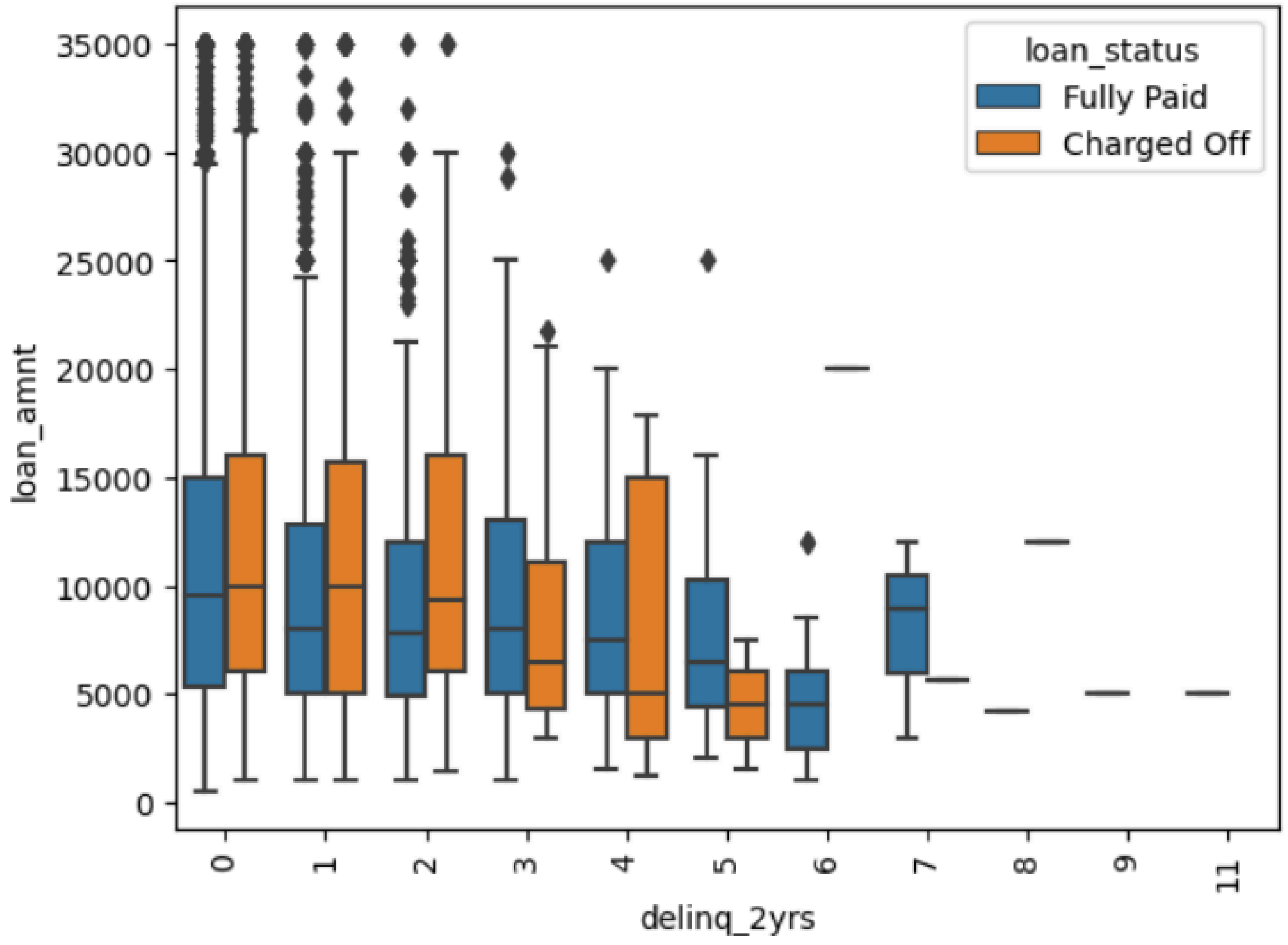


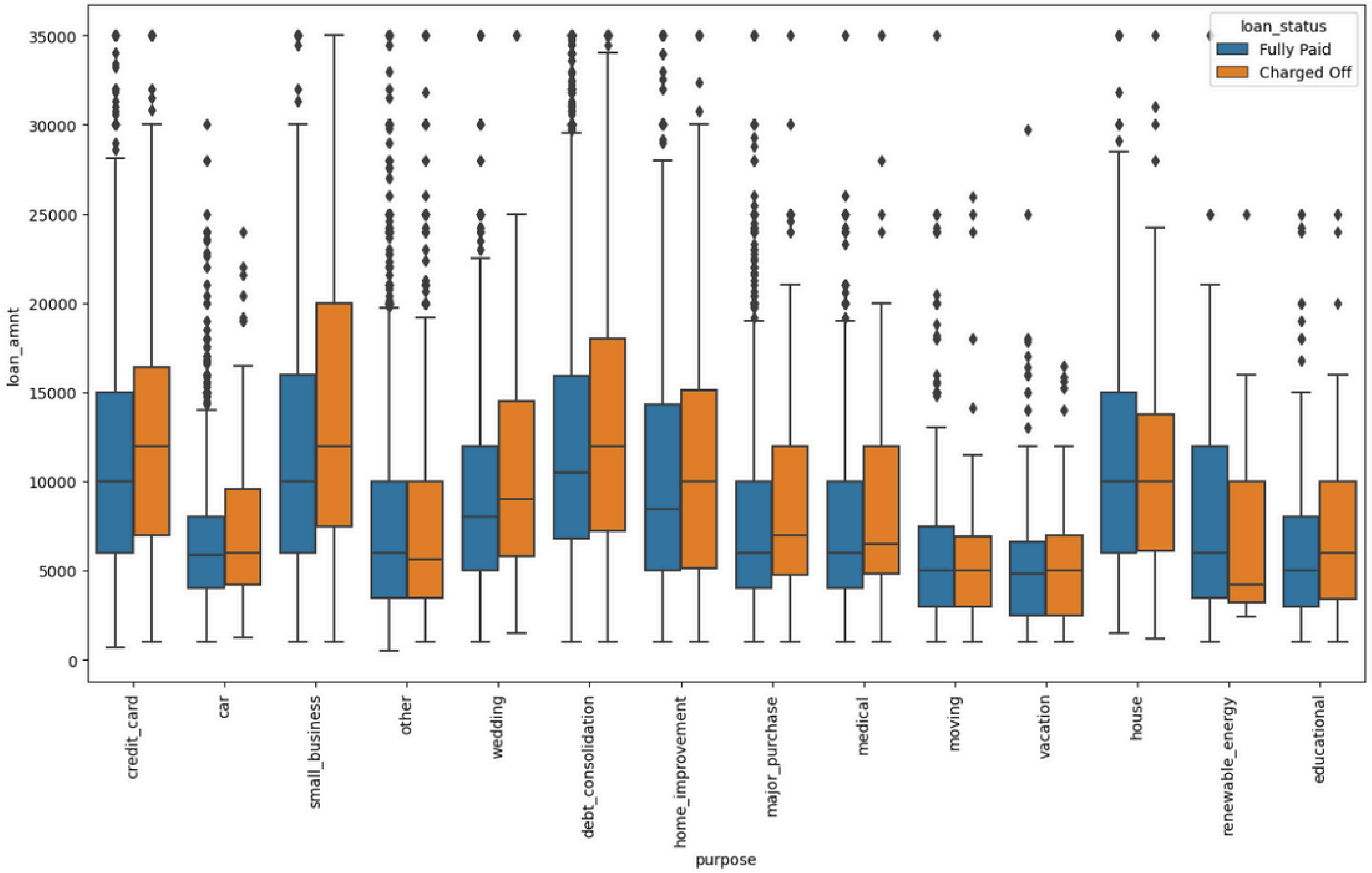
Loan amount vs Public records: Higher loan amounts to people with derogatory public records have been charged off.

Loan amount vs public recorded bankruptcies: Higher loan amounts to people with bankruptcy records have been charged off.

Bivariate Results

Numerical vs Categorical



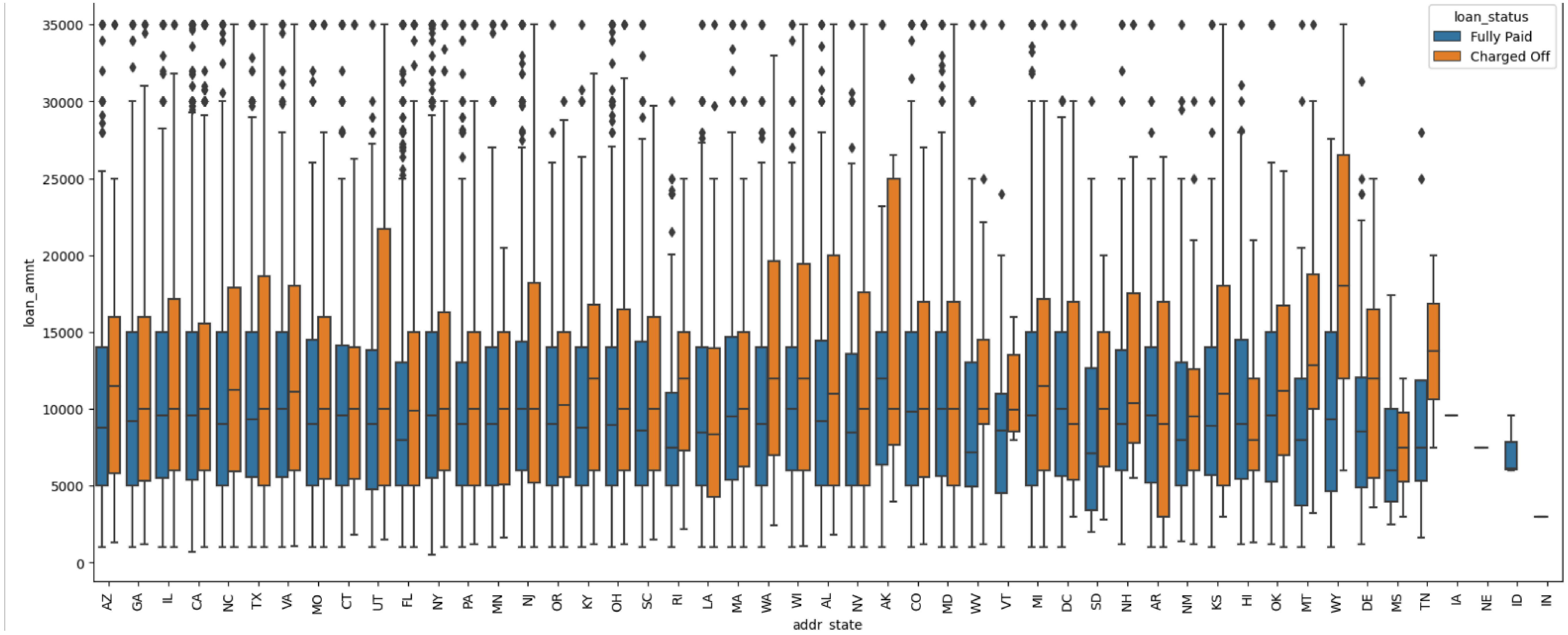


Loan amount vs delinquency: Loan amounts are same for till delinquency record of 4 but charged off loan has increased with increase in delinquency records.

Loan amount vs Puprose: Higher loan amounts have been sanctioned to purposes like small business, debt consolidation, credit card and home improvement and also the charged off amount is higher for these purposes

Bivariate Results

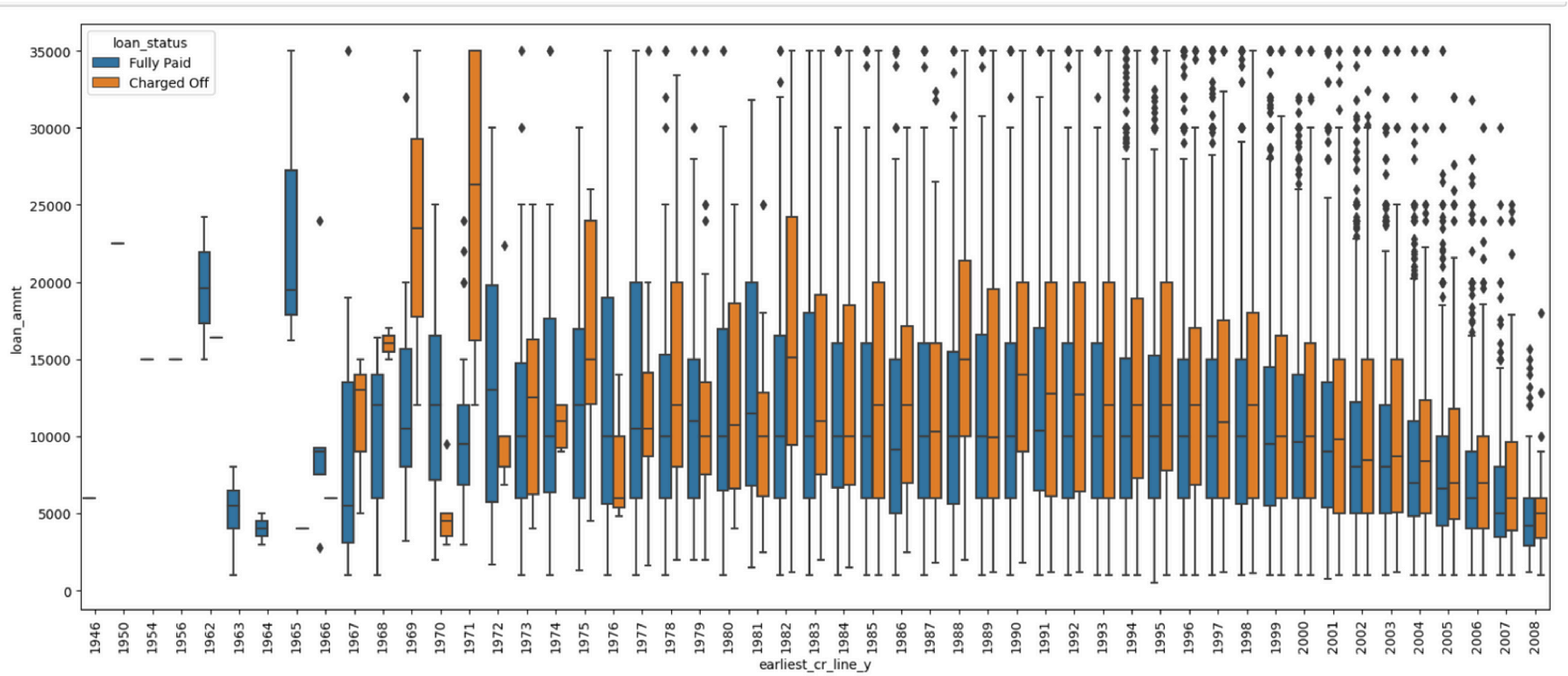
Numerical vs Categorical



Loan amount vs Address State: Higher loan amounts have been sanctioned to address state UT,AK, WY,TN and charged off amount is also higher for these states

Bivariate Results

Numerical vs Categorical



Loan amount vs Earliest Credit line year: Lower loan amounts have been given to newer accounts but charged off loan amount is more for newer accounts compared to fully paid loans.

Summary

From our analysis so far we can consider the below mentioned variables to be indicators of loan-default.

Strong Indicators

- **Loan Grade:** Lower grades(G means lowest A means highest grade of loan) with higher loan amount are very risky and have been shown to become defaulted.
- **Home Ownership :** Higher Loan amounts given to people who have mortgage on their homes are more likely to get charged off.
- **Purpose of the Loan:** Loans taken for purposes like small business, renewable energy, debt consolidation, home improvement, credit card and education have higher chances of getting default.
- **Interest Rate :** Loans with high interest rates are prone to default as primary purposes are debt consolidation and higher interest rates adds to the problem due to which applicants default on the loan amount.
- **Revolving Utilization:** Loans given to applicants with High revolving utilization(>15%) have higher likelihood of default.
- **Derogatory Public records:** Loans given to applicants who have 1-2 number of derogatory records have higher likelihood of default.
- **Public Bankruptcy records:** Loans given to applicants who have 1-2 number of bankruptcy records have higher likelihood of default.

Summary

Mild Indicators

- **Loan Term:** Loan amounts ith 60 months term are somewhat likely to get defaulted.
- **Lower Annual Income :** Loans given to applicants with lower annual income are somewhat likely to get defaulted.
- **DTI :** Loans given to applicants with higher dti are somewhat likely to get defaulted.
- **Address State :** Loans given to applicants who are from NV, SD, AK, FL, TN, WY are somewhat likely to get defaulted.
- **Number of Delinquency in last 2 years:** Loans given to applicants with higher incidences of delinquency have some likelihood of default.
- **Issue Month :** Loans given to applicants on the months of December , September and May have some likelihood of default.