

LENDING CLUB – CASE STUDY

Group facilitators:

Priyavarshini M

Ramkumar P

Submission Date: 13 Sept '23

Table of Contents

- ❑ *Data Understanding*
- ❑ *Data Cleaning and Manipulation*
- ❑ *Univariate Analysis*
- ❑ *Bivariate Analysis*

DATA UNDERSTANDING

- The csv dataset contains 111 columns and 39717 rows (loans) - 5627 charged off loans
- Objective of the case study is to identify markers on default loans, so we will analyse the charged off loans against the performing loans in the dataset.
- **Data quality issues:**
 - Removal of NA values columns
 - Columns irrelevant to the analysis of the objective
 - Columns with missing values (which cannot be populated because of absence of reliable methods/sources to populate these values)

Data Cleaning and Manipulation

- **Data cleaning performed:**

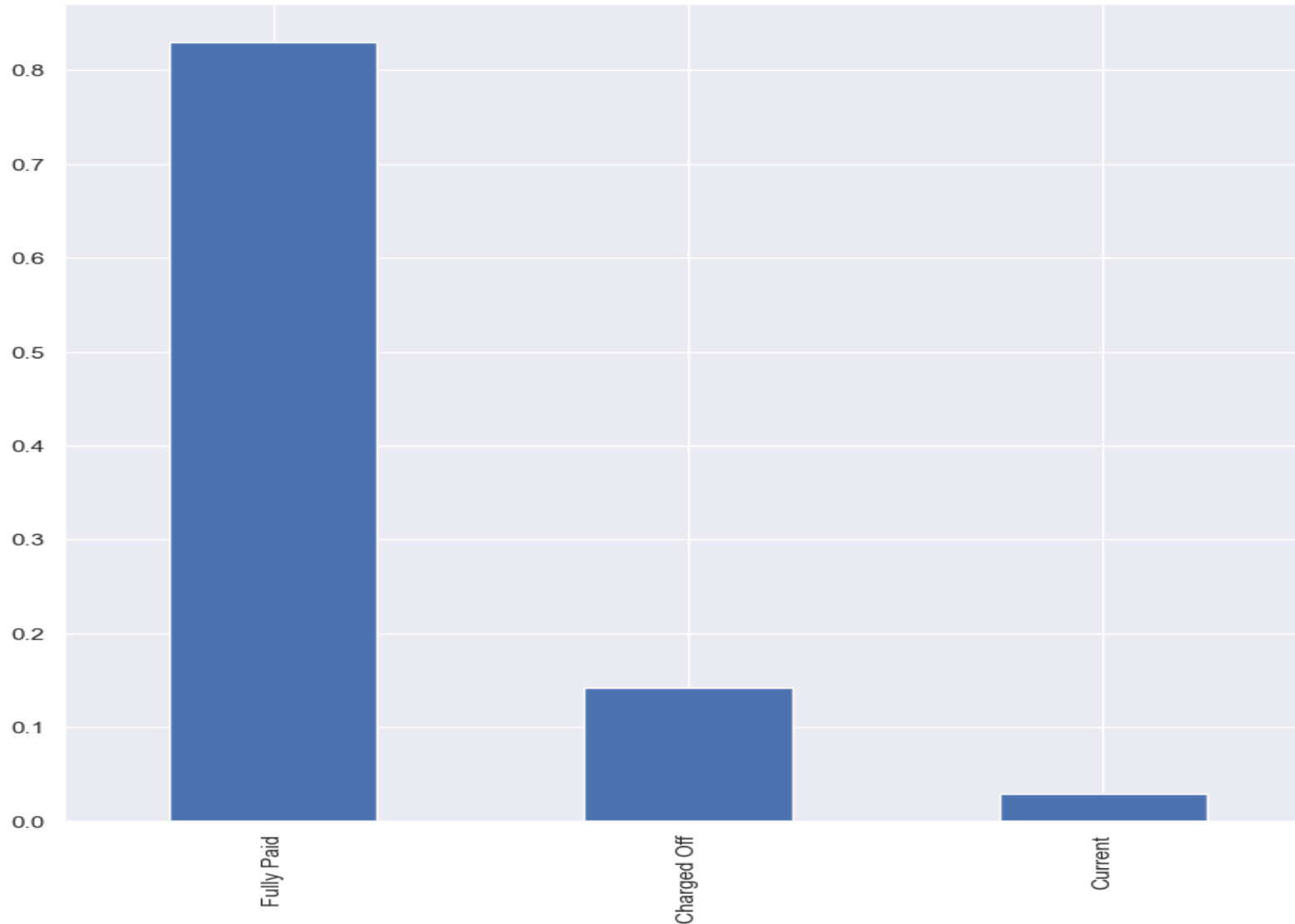
- Removal of NA values columns
- Columns irrelevant to the objective
- Columns with missing values (which cannot be populated because of absence of reliable methods/sources to populate these values)
- No duplicate values found
- Filtering columns by usability

Univariate analysis

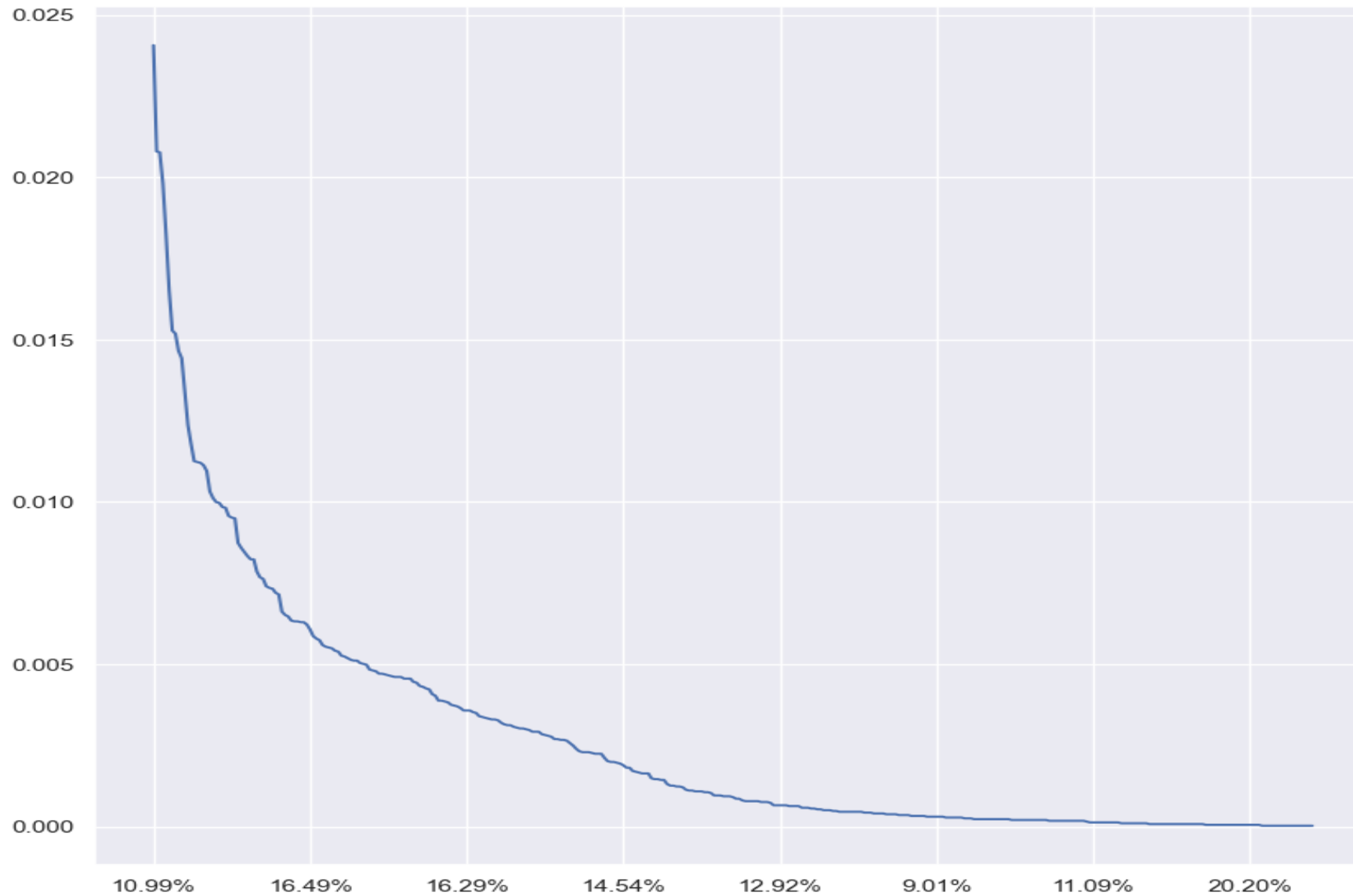
Univariate analysis – Important Driver for default loans

- The following driver variables have been identified:
 - Loan amount
 - House Ownership
 - Interest rate
 - Annual Income
 - Grade
 - Term

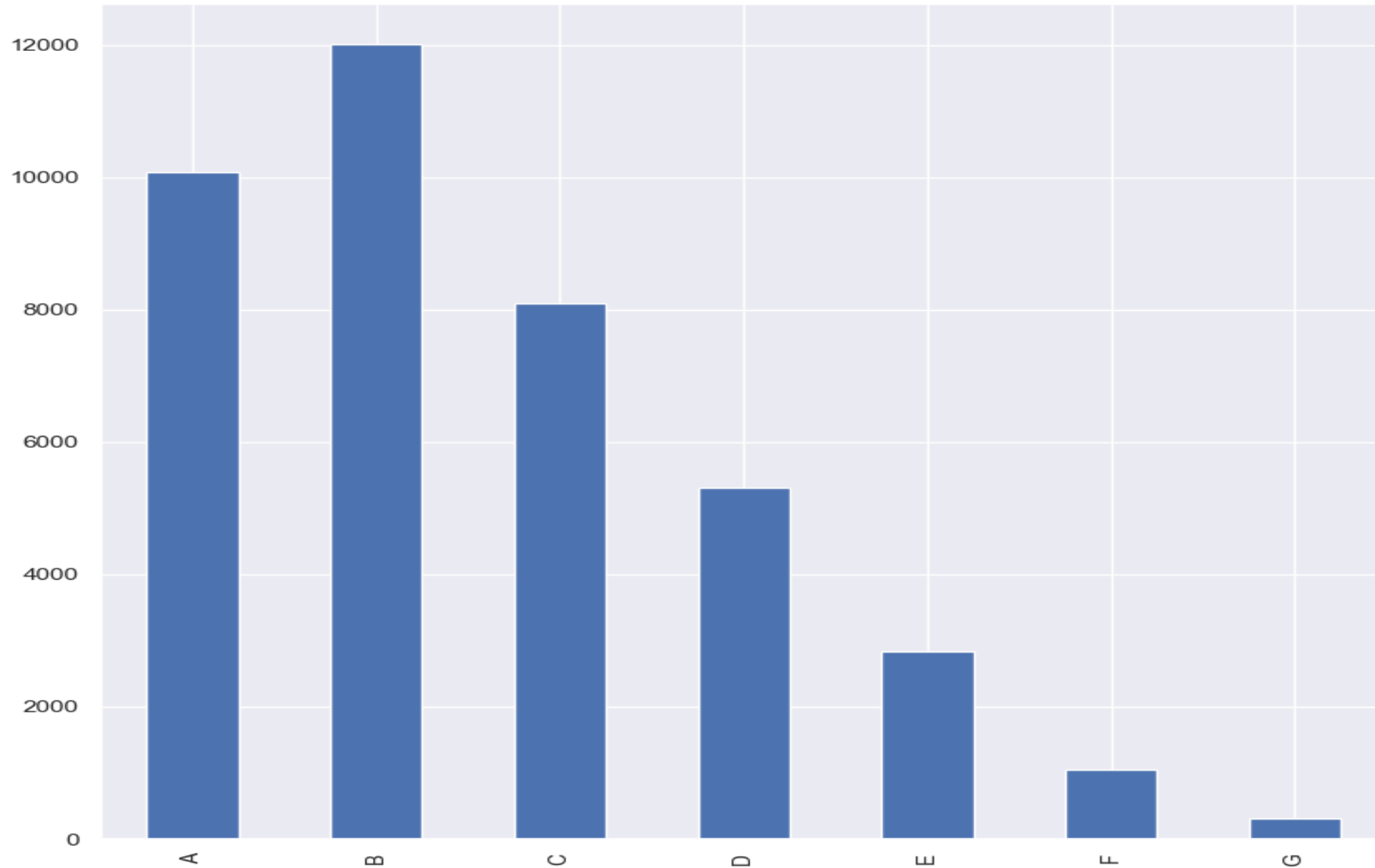
Number of fully paid loans, current loans against default loans
(df['loan_status'].value_counts().head(10) / len(df)).plot.bar()



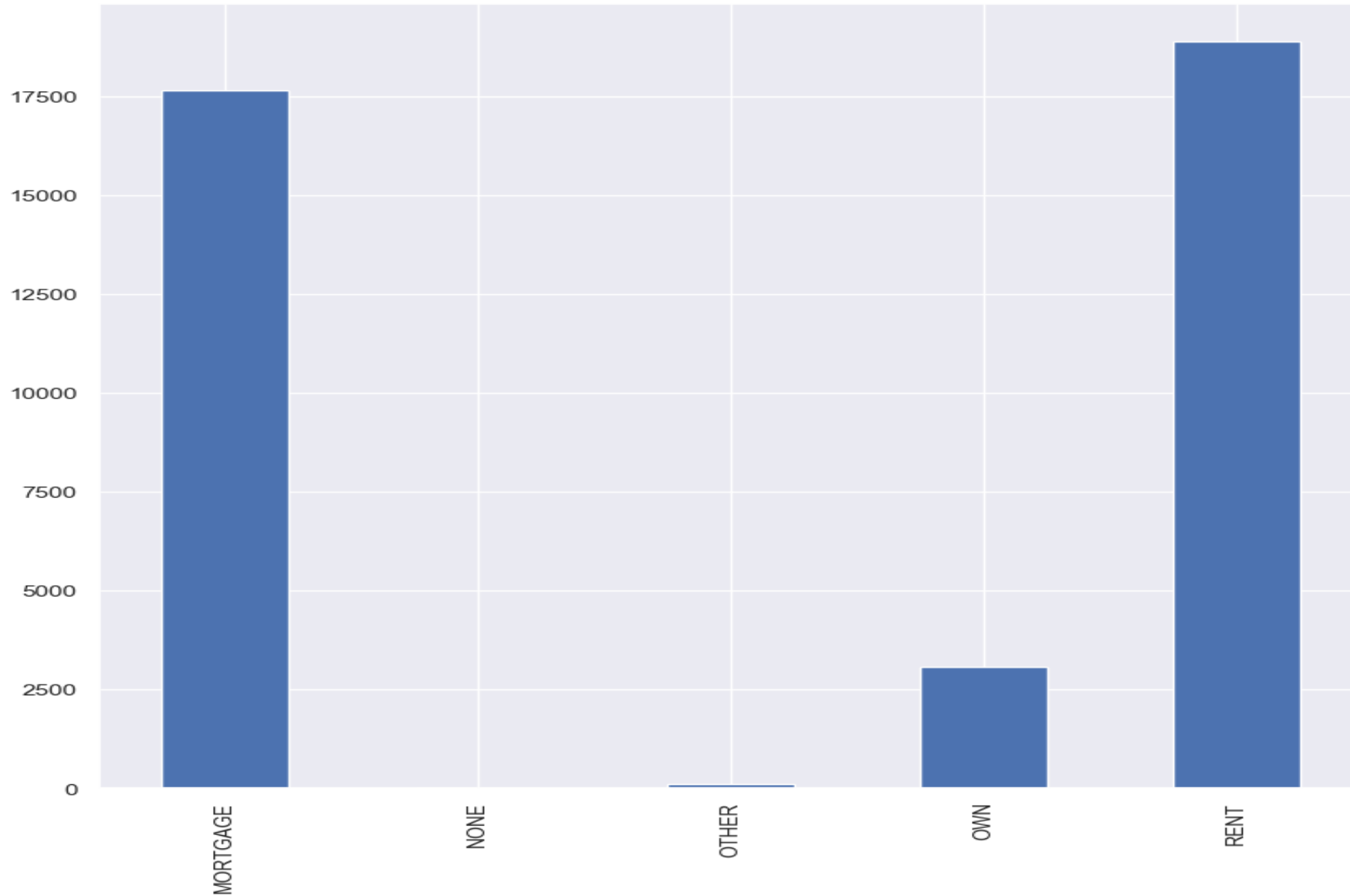
Progression of interest rates for all loans
(df['int_rate'].value_counts() / len(df)).plot.line()



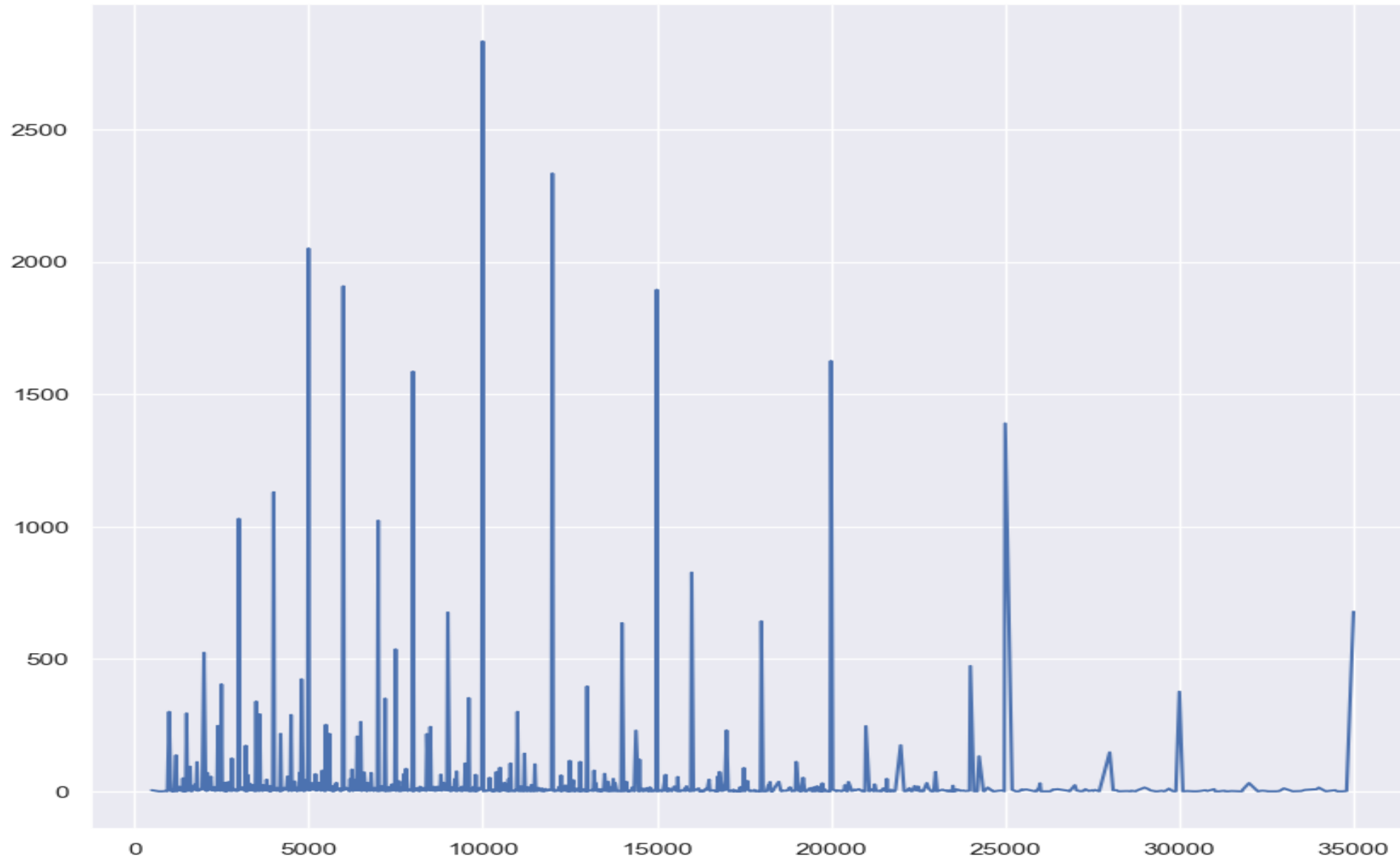
Grade wise distribution of all loans
`df['grade'].value_counts().sort_index().plot.bar()`



Distribution of the house ownership status of the loan availers
`df['home_ownership'].value_counts().sort_index().plot.bar()`

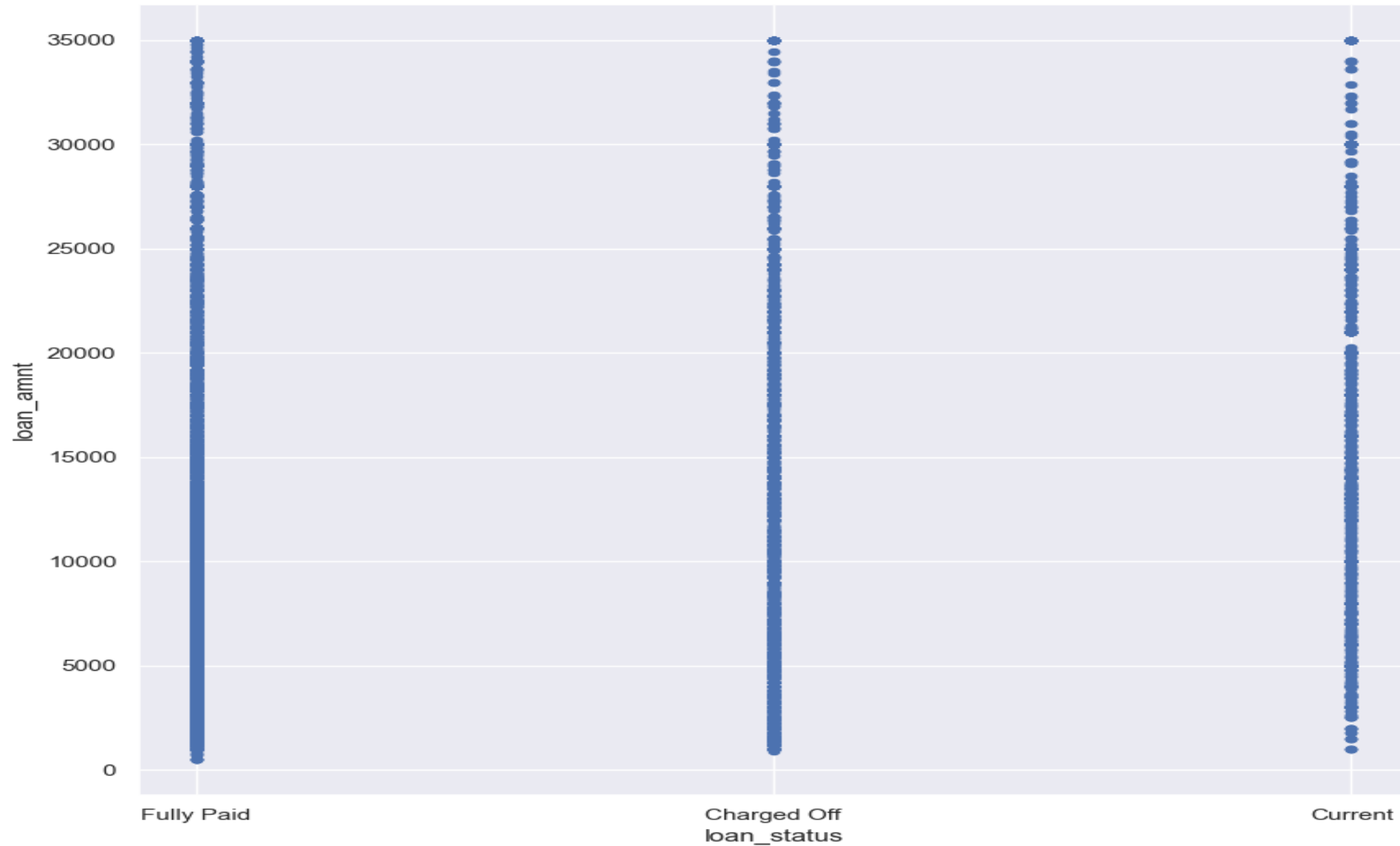


Distribution of the loan amounts in the dataset
`df['loan_amnt'].value_counts().sort_index().plot.line()`



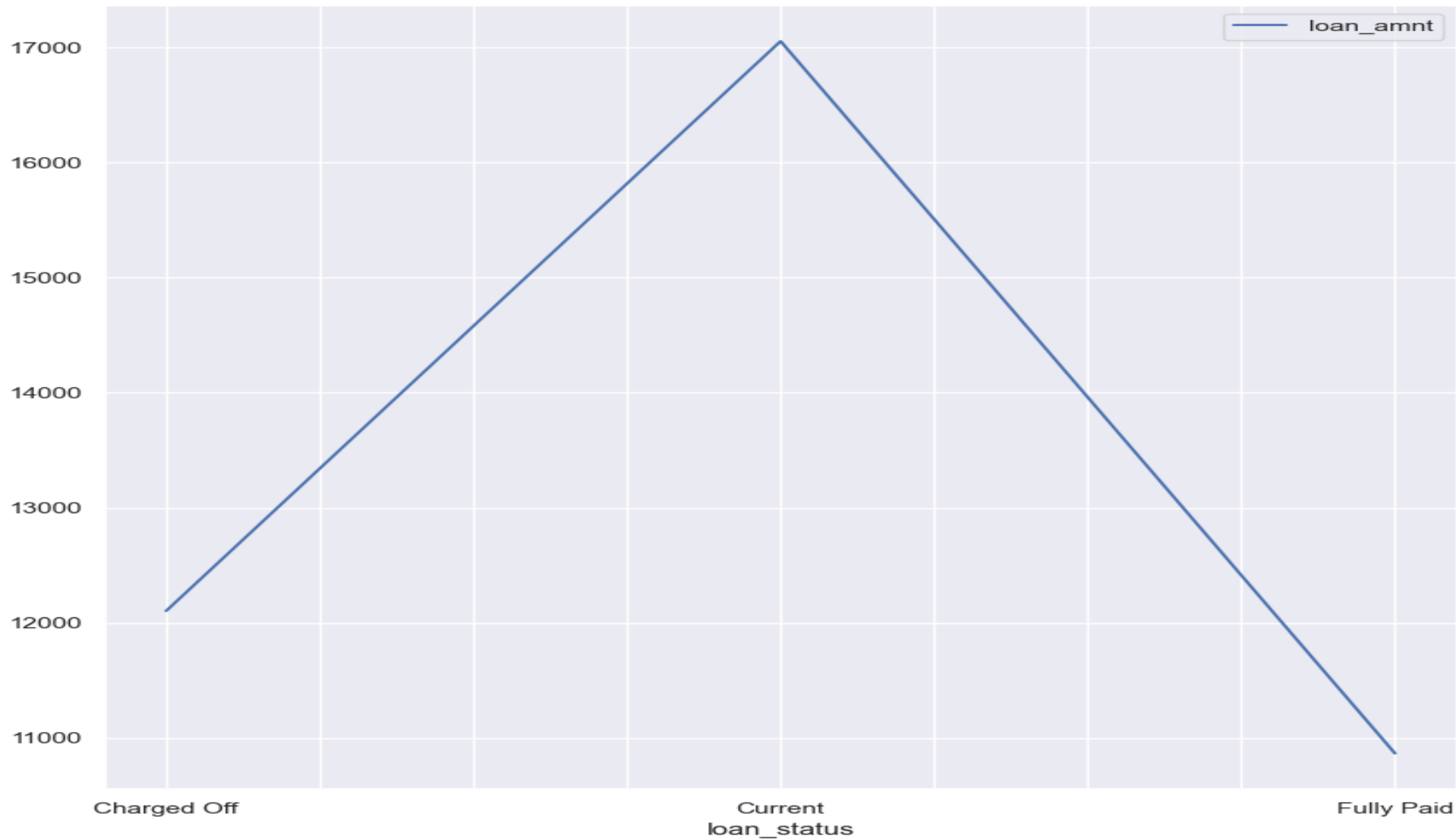
Bivariate Analysis

Default loan numbers reduce as the loan amount increases
`df.plot.scatter(x='loan_status', y='loan_amnt')`

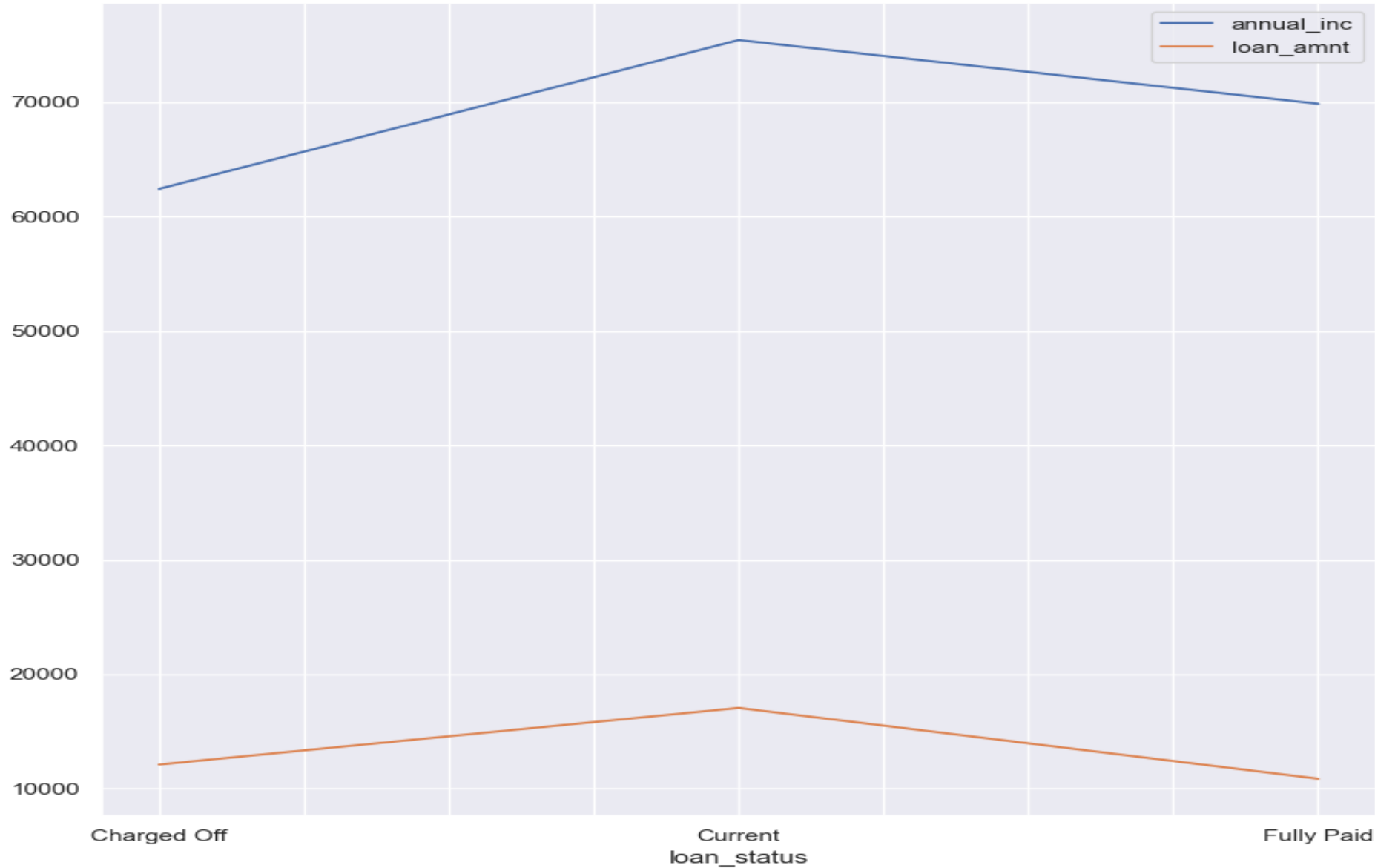


Charged off loans have more occurrence when the loan amount is lesser

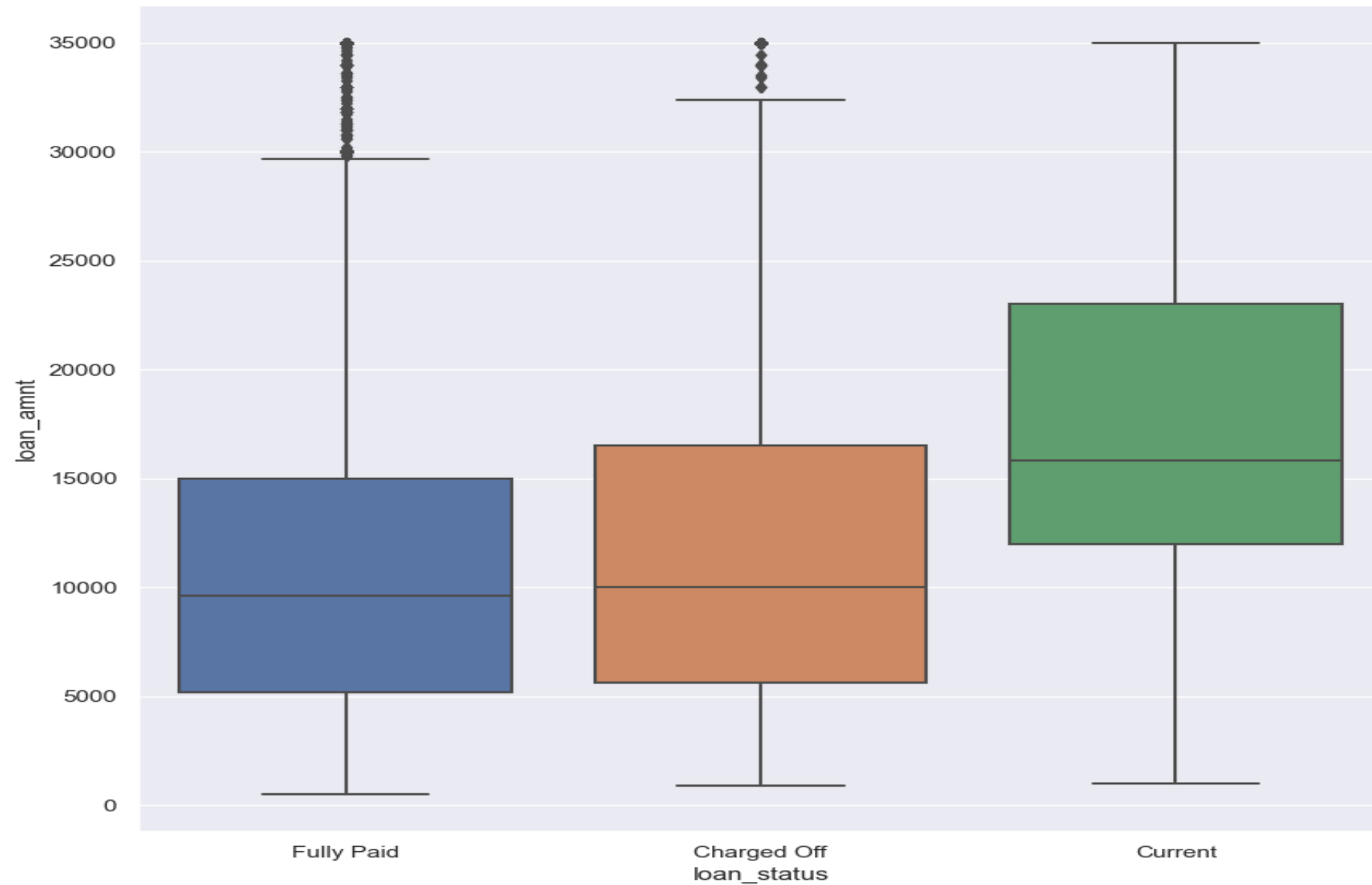
```
dfline = dfnew.groupby('loan_status').mean()[['loan_amnt']]  
dfline.plot.line()
```



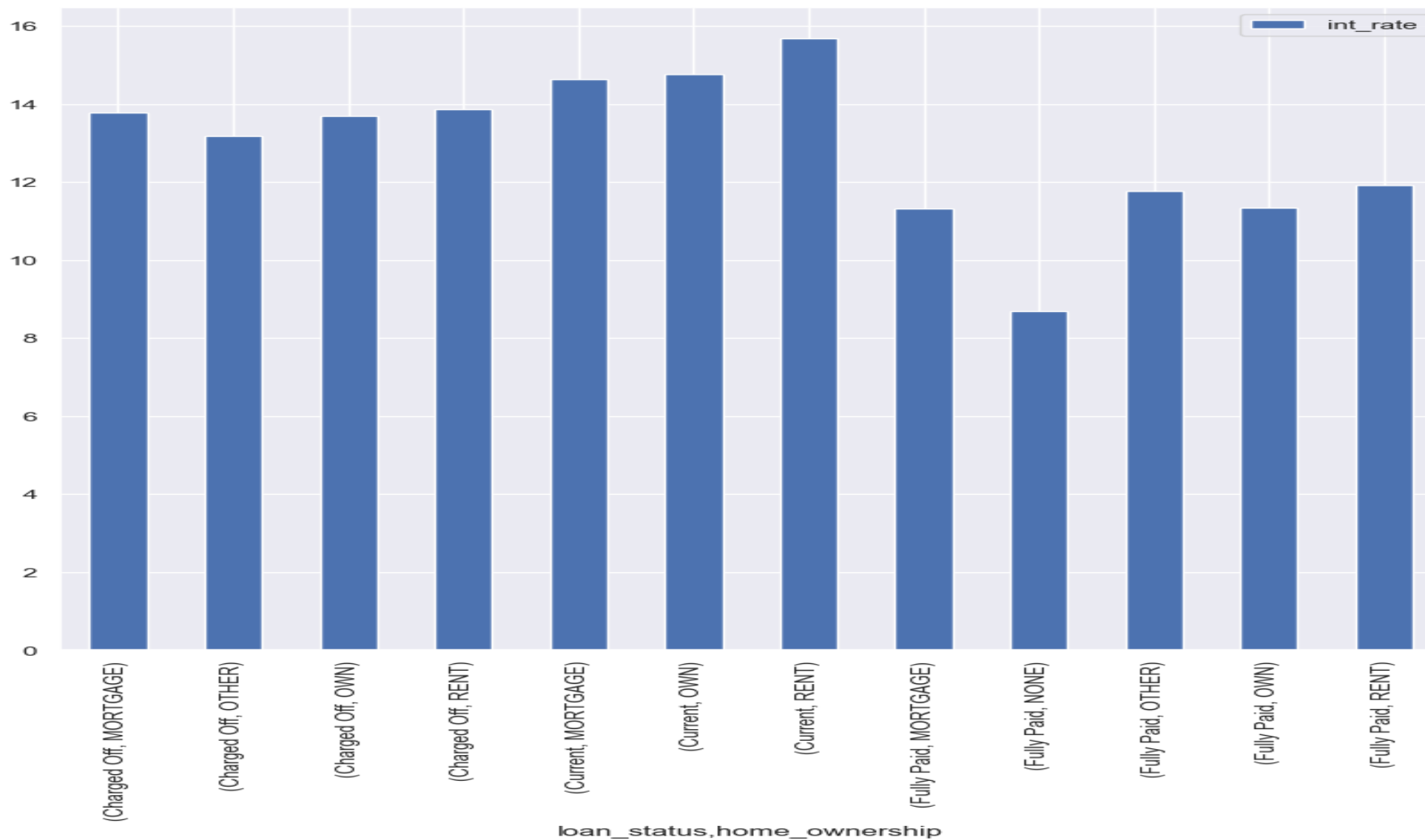
Charged off loans have higher chances of occurrence when annual income is low as well as the loan amount is lower.



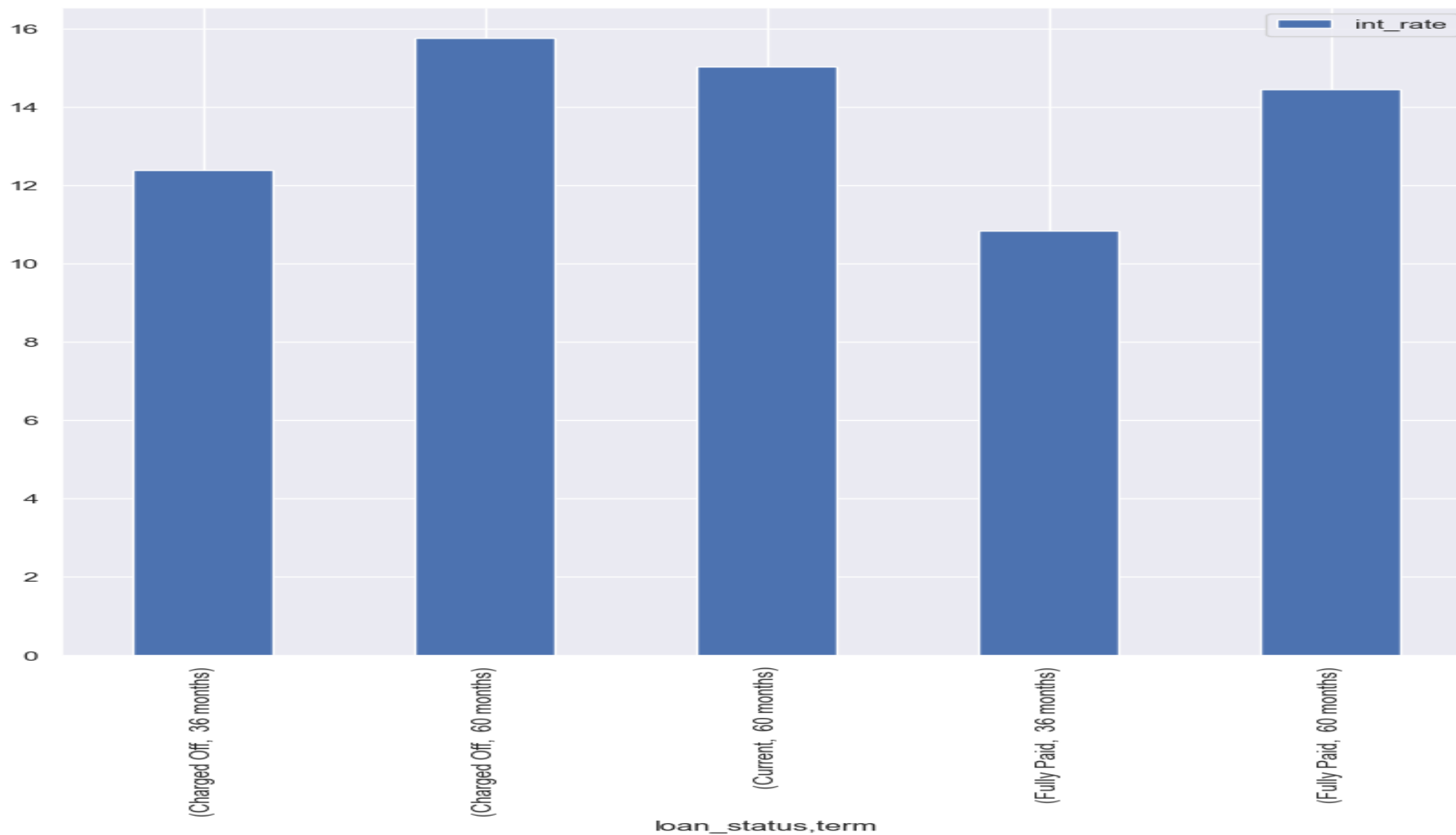
Charged off loans have more occurrence when the loan amount is in this range



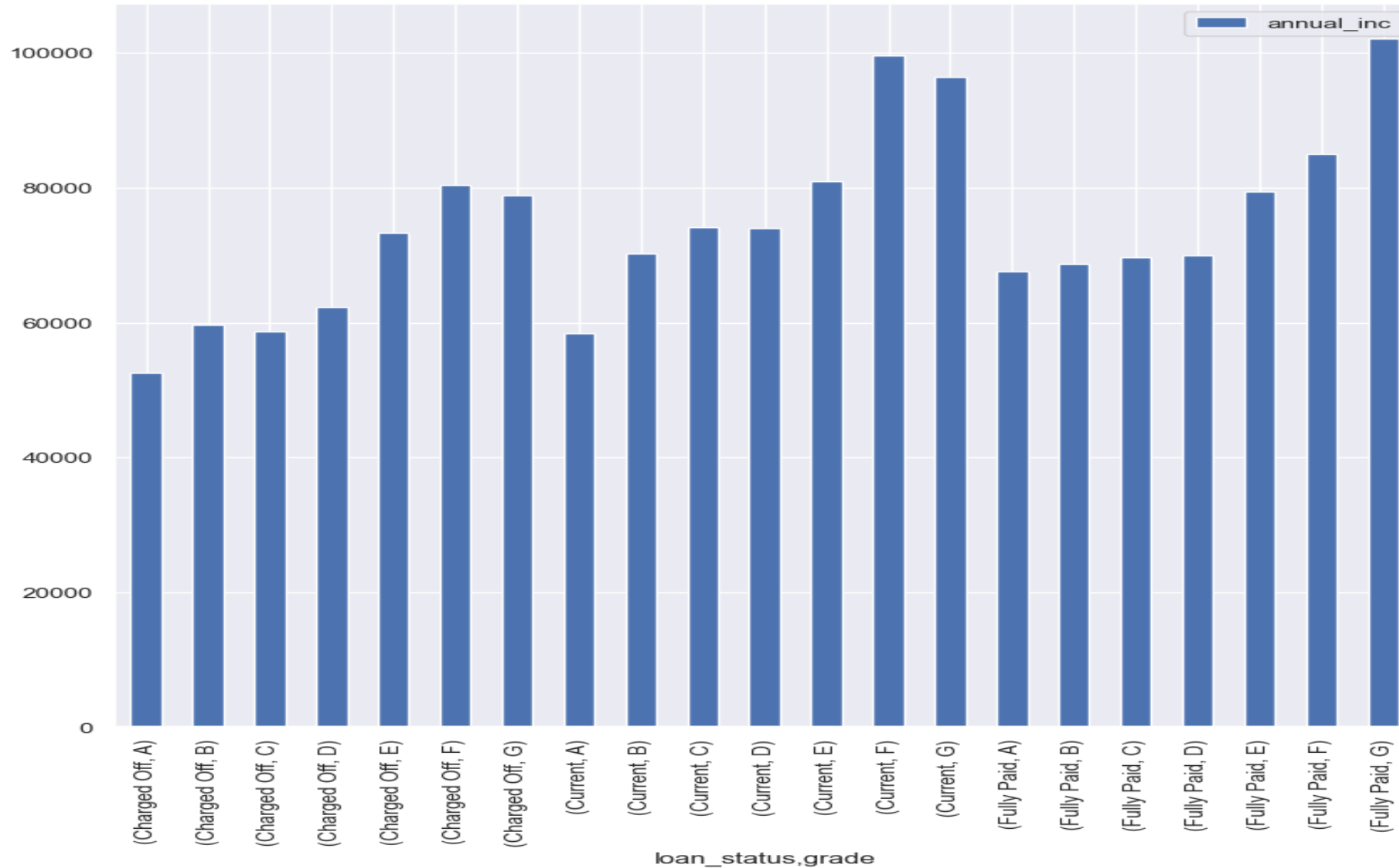
Charged off loans have more occurrence when the home ownership status is in Rent or Mortgaged



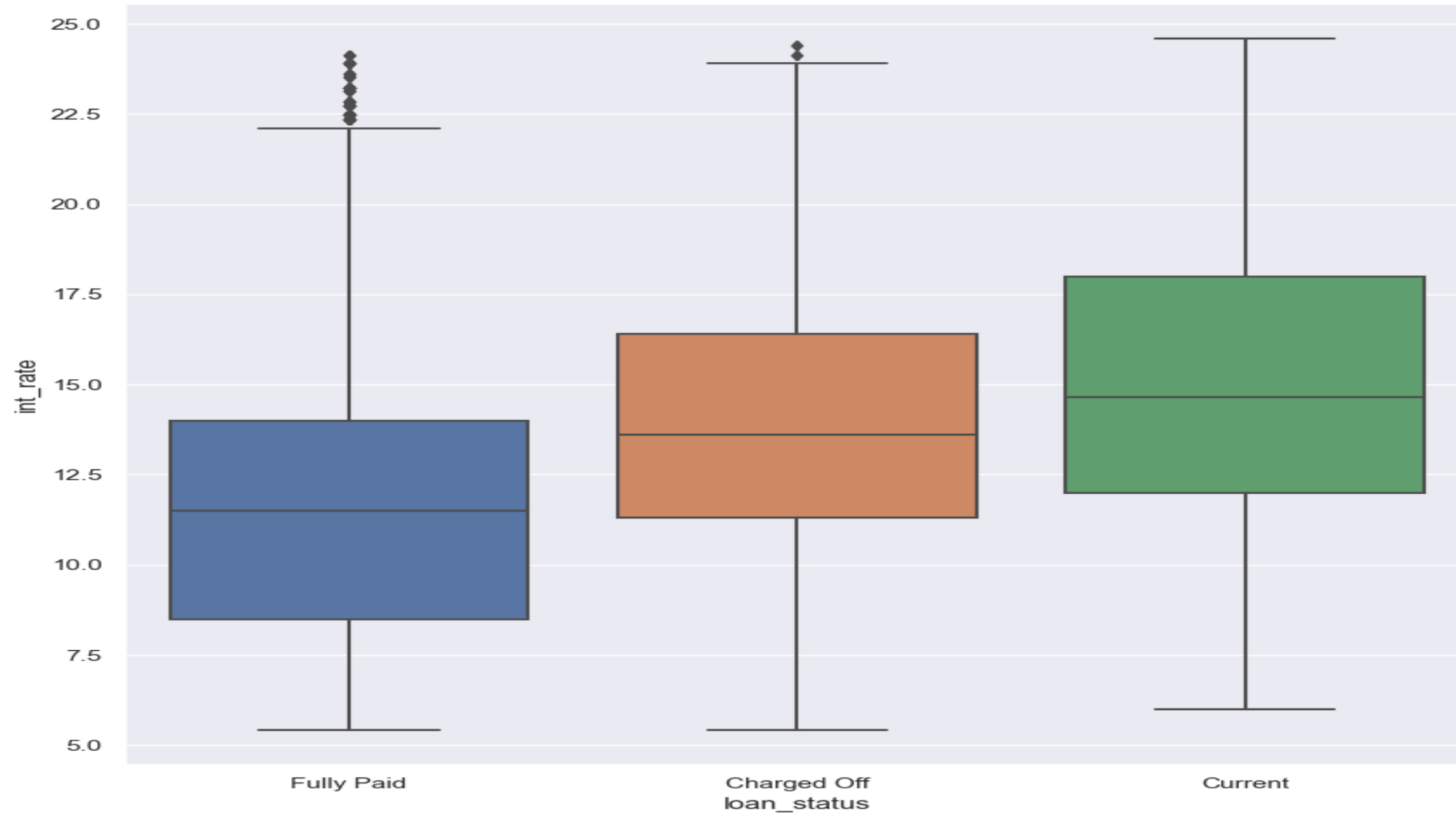
Charged off loans have more occurrence when the term is more



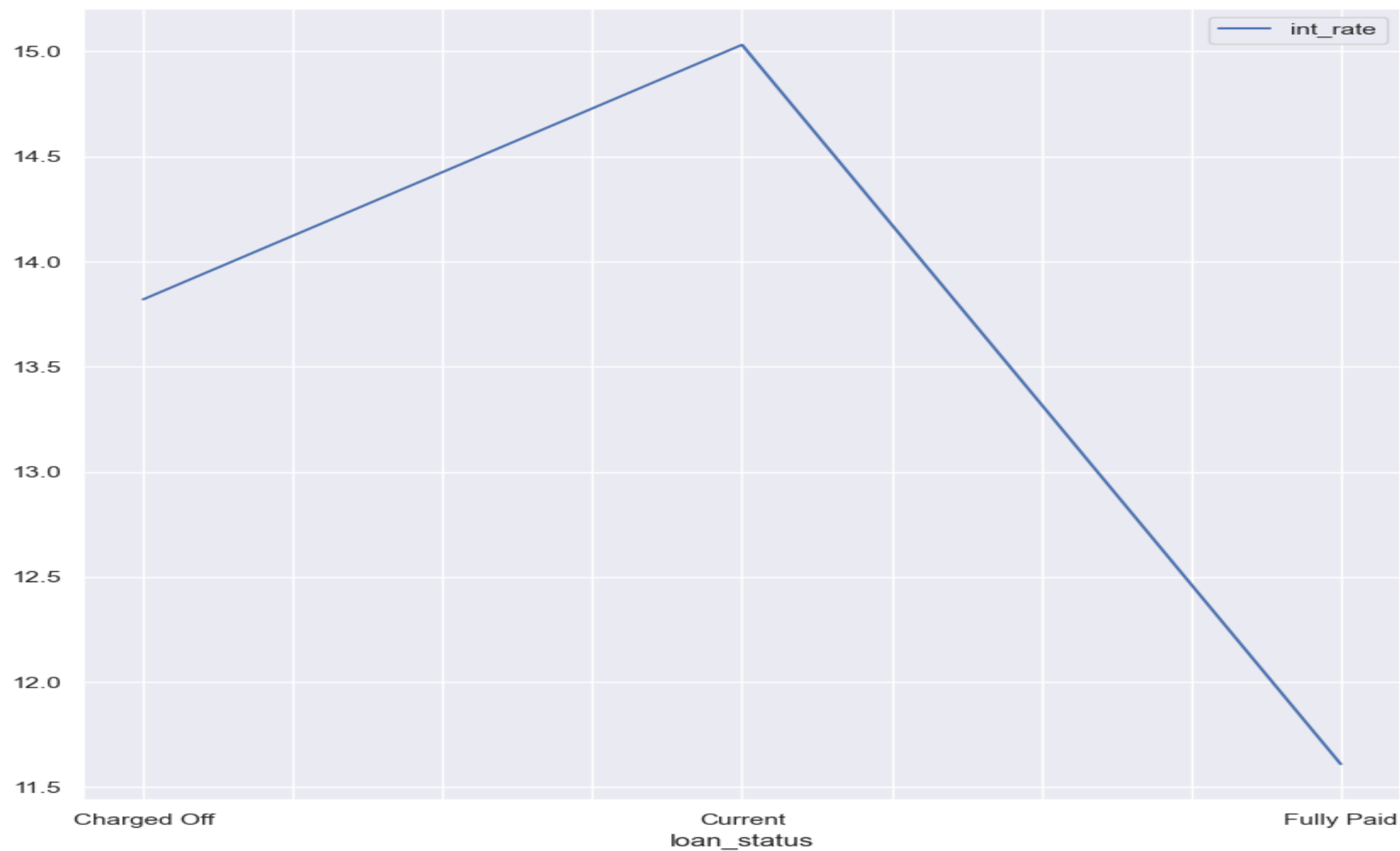
Charged off loans have more occurrence with employment grades E, F, G



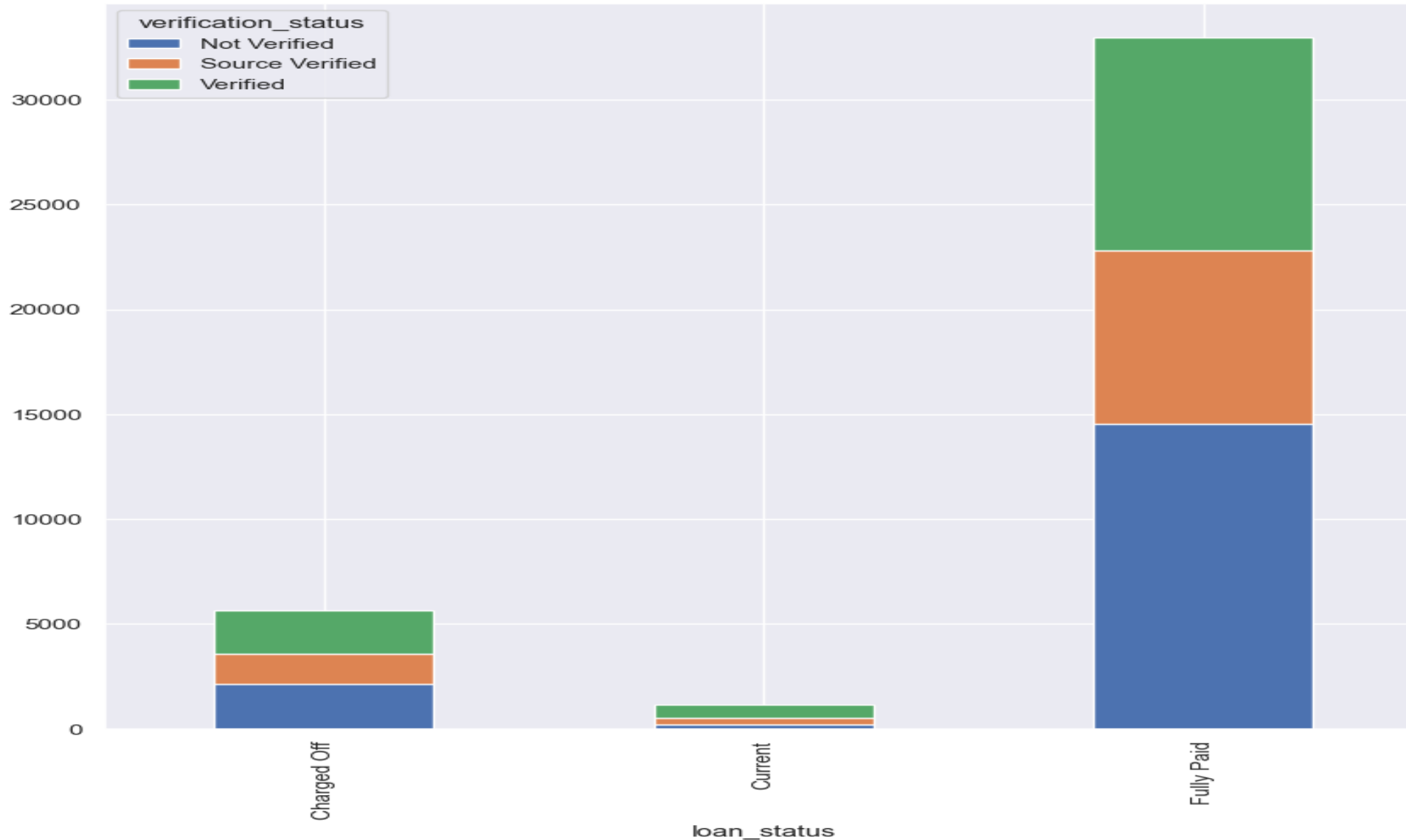
Charged off loans have more occurrence when the interest rate range is 11.25 – 16.25



Charged off loans have more occurrence when the interest rate range is 11.25 – 16.25



Charged off loans have more occurrence when verification status is 'Unverified'. High Risk applications should always be provided loans after complete verification so as to reduce the number of defaulters.



Charged off loans are uniform across grades B, C ,D.

