Group Facilitators
Ramkumar P
Siddharth Rao

Case study – Lending Club

Submission Date: 15/11/2023

One Size Fit Standard Approach - Exploratory Data Analysis

- Business Understanding
- Data Understanding & Summarization
- Data Cleaning and Preparation
- Exploratory Data Analysis
  - Data Exploration
  - Univariate Analysis
  - Bivariate Analysis
  - Multivariate Analysis
  - Data Visualization
- Features Engineering
- Statistical Testing
- Documentation

Business Understanding

Introduction

- The case study involves the application of Exploratory Data Analysis (EDA) and Financial Analysis to address a significant business challenge for a consumer finance company. The primary objective is to identify and understand key patterns and influential factors associated with loan defaults. The study aims to provide insights into Risk Analytics and Financial Services, aiding in the development of effective strategies to mitigate financial risks and enhance overall decision-making processes.

Business Understanding

- The consumer finance company specializes in facilitating various types of loans, including personal, business, and medical procedure financing, through an efficient online loan marketplace. With a focus on providing lower interest rates and a streamlined application process, the company caters primarily to urban customers.
- In the loan approval process, the company faces two primary types of risk:
- Business Loss Risk: Declining a loan application from a creditworthy applicant can result in a loss of potential business for the company.
- Financial Loss Risk: Approving a loan for an applicant likely to default poses the risk of financial loss for the company.
- The company's objective is to utilize comprehensive data analysis techniques to assess applicant profiles and develop a robust risk assessment framework. This approach enables informed decision-making, reducing the likelihood of both business and financial losses.

Loan Application

Loan Acceptance Scenarios:

- When the company approves a loan application, the following scenarios may occur:
- **Fully Paid:** The applicant successfully fulfills the loan obligations by repaying both the principal amount and the interest within the agreed timeframe.
- **Current Status:** The applicant is currently in the process of repaying the loan through regular instalments. These applicants are not considered 'defaulted' as the loan tenure is ongoing.
- **Charged-Off:** The applicant has failed to make timely payments over an extended period, leading to a defaulted status for the loan.

Loan Rejection Scenario:

- In cases where the company rejects a loan application due to the applicant's failure to meet specific requirements, no transactional history is available. Consequently, the company lacks data on these applicants within the dataset.

Credit Loss

- Extending loans to individuals deemed 'risky' constitutes the primary cause of financial loss, commonly referred to as credit loss, for lenders.
- Credit loss represents the financial deficit incurred by the lending institution when borrowers either fail to repay the loan or abscond with the owed funds.
- Put simply, borrowers who default on their loan obligations pose the most substantial financial risk for lenders.
- Within the context of this case, customers classified as 'charged-off' correspond to the category of 'defaulters.'

Business Objective

- The primary business objective is to reduce financial and business loss risks associated with the lending process. This involves both mitigating potential losses due to defaulted loans and ensuring that the company does not miss out on creditworthy applicants.
- EDA is utilized to identify patterns and risk factors related to loan defaults. These insights will inform the company's decision-making processes and risk assessment strategies, aiming to minimize the occurrence of defaulted loans.
- Significance of Risk Assessment and Analytics: In the consumer finance industry, risk assessment is a crucial aspect of maintaining a sustainable and profitable lending business. Understanding the potential risks associated with lending enables the company to make informed decisions about loan approvals, interest rates, and terms.
- Analytics, particularly EDA, plays a vital role in identifying trends, patterns, and risk factors that may contribute to loan defaults. By leveraging data-driven insights, the company can proactively manage its loan portfolio, reduce financial exposure, and optimize its lending strategies to cater to creditworthy customers while minimizing the risk of defaults.
- Through EDA, the goal is to gain a deep understanding of both customer attributes and loan-specific attributes that significantly influence the likelihood of default. This critical analysis aims to minimize the number of risky loans, consequently reducing the overall credit loss incurred.
- Furthermore, the company aims to uncover key patterns that serve as indicators of potential default. These patterns can be instrumental in implementing various risk-mitigating strategies, such as rejecting the loan, adjusting the loan amount, or offering loans to risky applicants at a higher interest rate.
- Additionally, the company endeavours to identify the primary driving factors or driver variables that strongly correlate with loan default. Understanding these factors can aid in optimizing the company's portfolio management and risk assessment procedures.

Key Questions and Objectives

- What are the primary characteristics of customers who have defaulted on their loans?
- Are there any specific trends or patterns in the data that indicate potential risk factors leading to loan defaults?
- How do various loan attributes, such as loan amount, interest rate, and term, impact the likelihood of default?
- Can we identify any correlations between borrower attributes (e.g., employment length, income, home ownership) and the probability of loan default?
- Are there any discernible differences in default rates across different loan categories or purposes?
- How does the company's historical loan data provide insights into effective risk assessment and portfolio management strategies?
- What are the key indicators or driver variables that strongly correlate with loan defaults, and how can this information be used to enhance the risk assessment process?
- Can we derive actionable insights to support decision-making, such as adjusting loan approval criteria, implementing risk-based pricing, or enhancing the company's portfolio management practices?

Data understanding & summarization

Attributes	Values
Csv data set columns(Customer Attributes)	111
Csv Data set Rows(loans)	39717
Charged Off	

Loan Data Set

- The loan dataset provides a comprehensive record of loans issued from 2007 to 2011. It includes a data dictionary explaining the variable meanings. This dataset offers insights into the historical behaviour of loan applicants, helping us understand patterns of loan default and non-default cases.
- The Loan data set we have is a kind of private data since it contains personal and confidential information of loan holders.
- 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 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)

12

Data Quality Checkpoints

- **Completeness Check**: Assess whether all expected data is present and whether there are any missing values that could impact the analysis.
- **Consistency Check**: Ensure that data across different sources or components are consistent and follow uniform formatting and standards.
- **Validity Check**: Verify that the data conforms to predefined formats, standards, and ranges, eliminating any invalid or unrealistic entries.
- **Accuracy Check**: Evaluate the accuracy of data by comparing it with reliable sources or known standards, identifying any potential errors or inconsistencies.
- **Uniqueness Check**: Examine the uniqueness of data points, identifying and resolving any duplicate entries that could skew analysis results.
- **Integrity Check**: Confirm that relationships between different datasets or components are maintained, preserving data integrity throughout the analysis process.
- **Timeliness Check**: Ensure that data is up to date and relevant for the analysis being conducted, avoiding any outdated or irrelevant information.

Data Quality Assessment

Formatting Erros – Ill formatted or unclearly named rows and columns and fixes.

- Delete Incorrect Rows(Entities or Records(loans)) - unnecessary header rows and footer rows
- Delete Summary – Total, Subtotals
- Delete Extra Rows Ex. Column number or Blank rows
- Add Column name if missing(Missing Header)
- Rename Column consistency(Abbreviations, encoded columns)
- Delete Unnecessary columns – unidentified and irrelevant columns
- Split columns for more data
- Merge columns for identifiers Ex. FirstName, LastName, Name state & District
- Align Mis – Aligned column –shifted columns

Missing values – empty or incomplete data and fixes.

Fix filters- Repeated rows(duplicates) and Spelling inconsistencies.

Outliers – Data points that deviate from normal values

- Ex-Extremely High Price or Low Price of product

Data Validation

- Accuracy – Ex
- Consistency – Contra-dictionary and conflicting
- Invalid data Ex N/A

Data Understanding

Columns: column represents a loan and customer attributes associated with the loan application. Each column represents different features or data points related to the loan application.

Rows: Each row corresponds to a particular loan application or borrower's information

Ex

id: An identifier for the loan application

member_id: Member identifier for the loan

loan_amnt: The amount of the loan requested

funded_amnt: The amount of the loan funded

term: The duration of the loan in months

int_rate: The interest rate on the loan

installment: The monthly payment amount

grade: The grade assigned to the loan

sub_grade: The sub-grade assigned to the loan

emp_title: The job title of the applicant

15

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

16

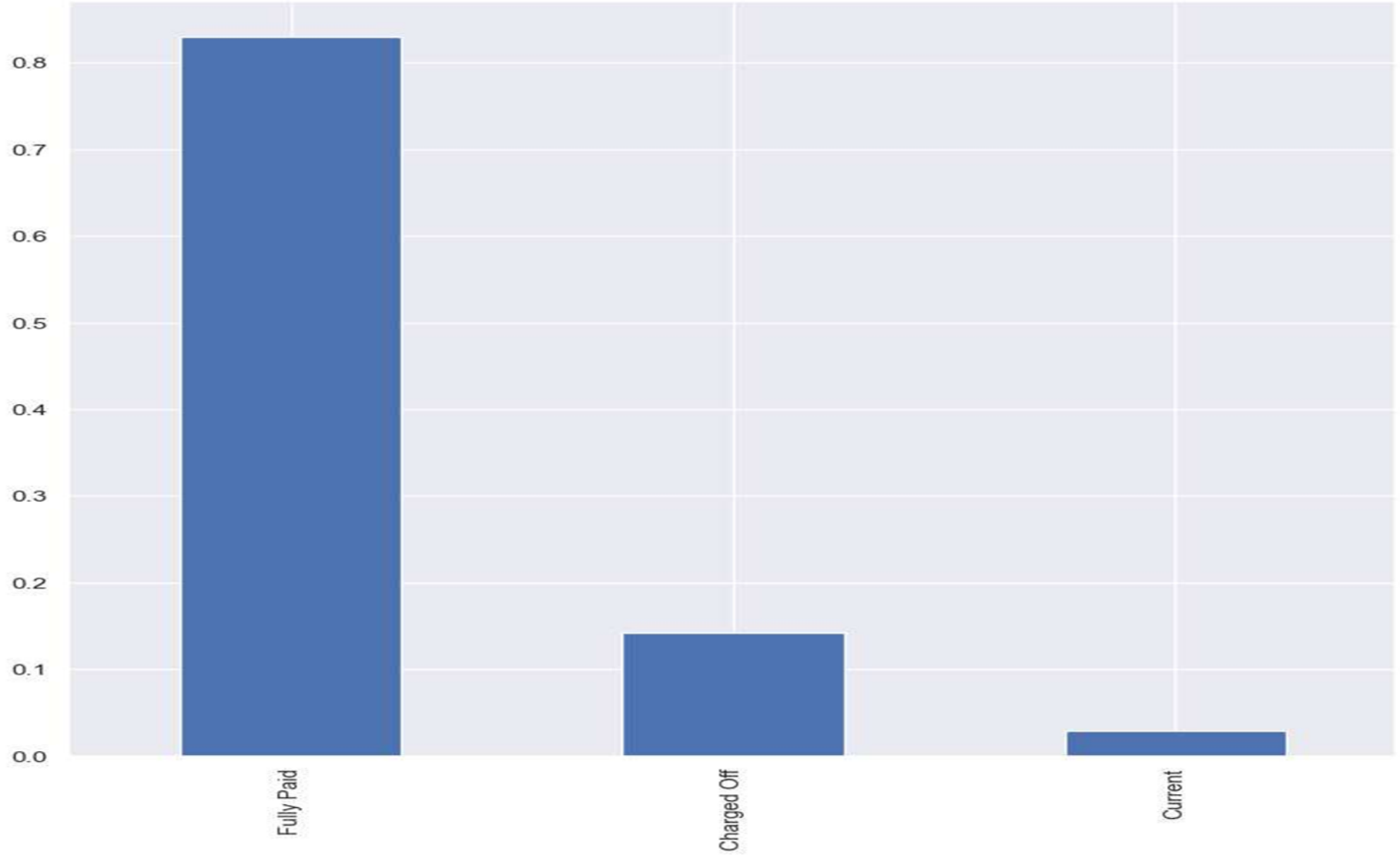
Univariate Analysis

Univariate analysis – Important Driver for default loans

The following driver variables have been identified:

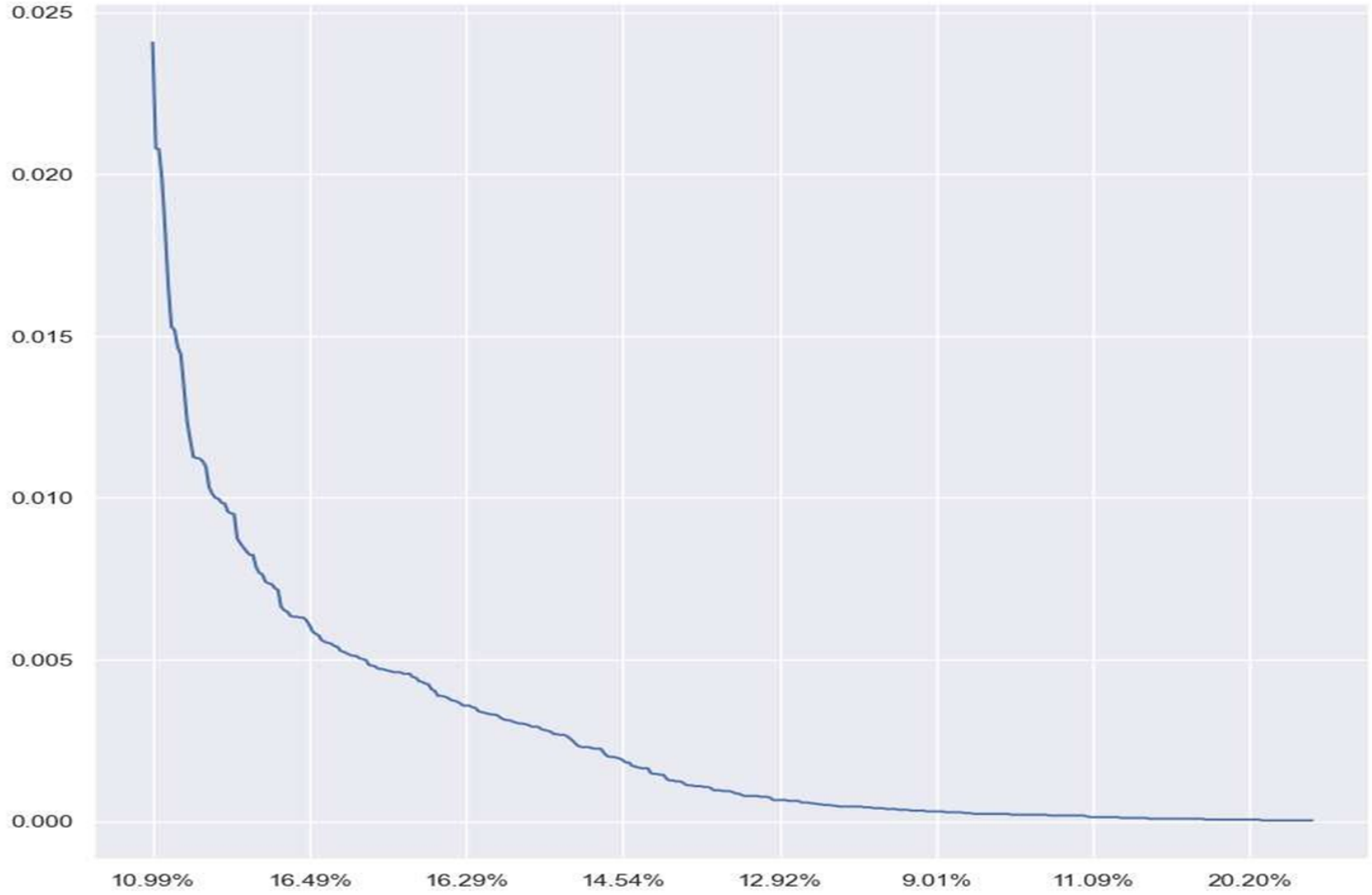
- Loan amount
- House Ownership
- Interest rate
- Annual Income
- Grade
- Term

NUMBER OF FULLY PAID LOANS, CURRENT LOANS AGAINST DEFAULT LOANS (DF['LOAN_STATUS'].VALUE_COUNTS().HEAD(10) / LEN(DF)).PLOT.BAR()



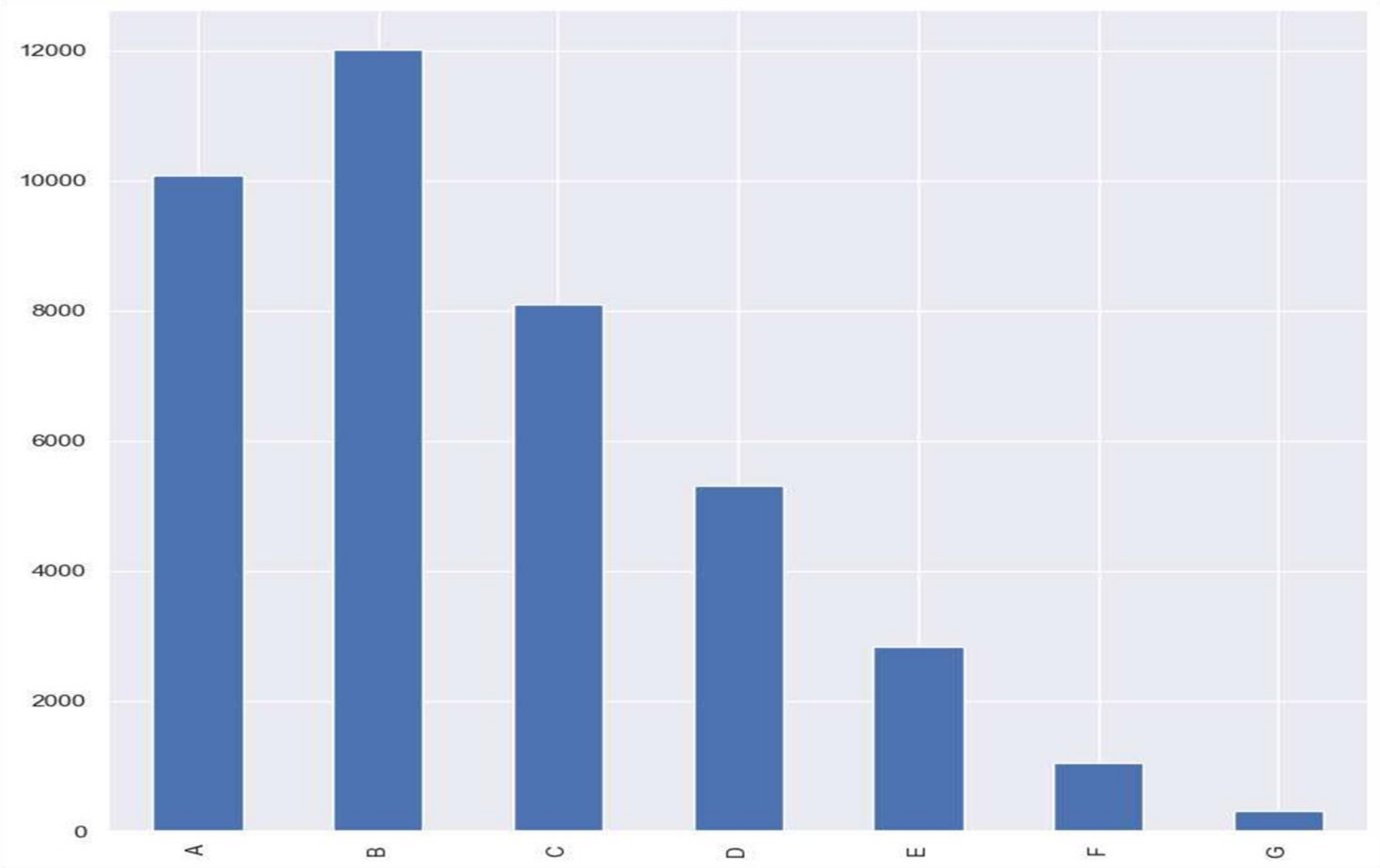
PROGRESSION OF INTEREST RATES FOR ALL LOANS

(DF['INT_RATE'].VALUE_COUNTS() / LEN(DF)).PLOT.LINE()



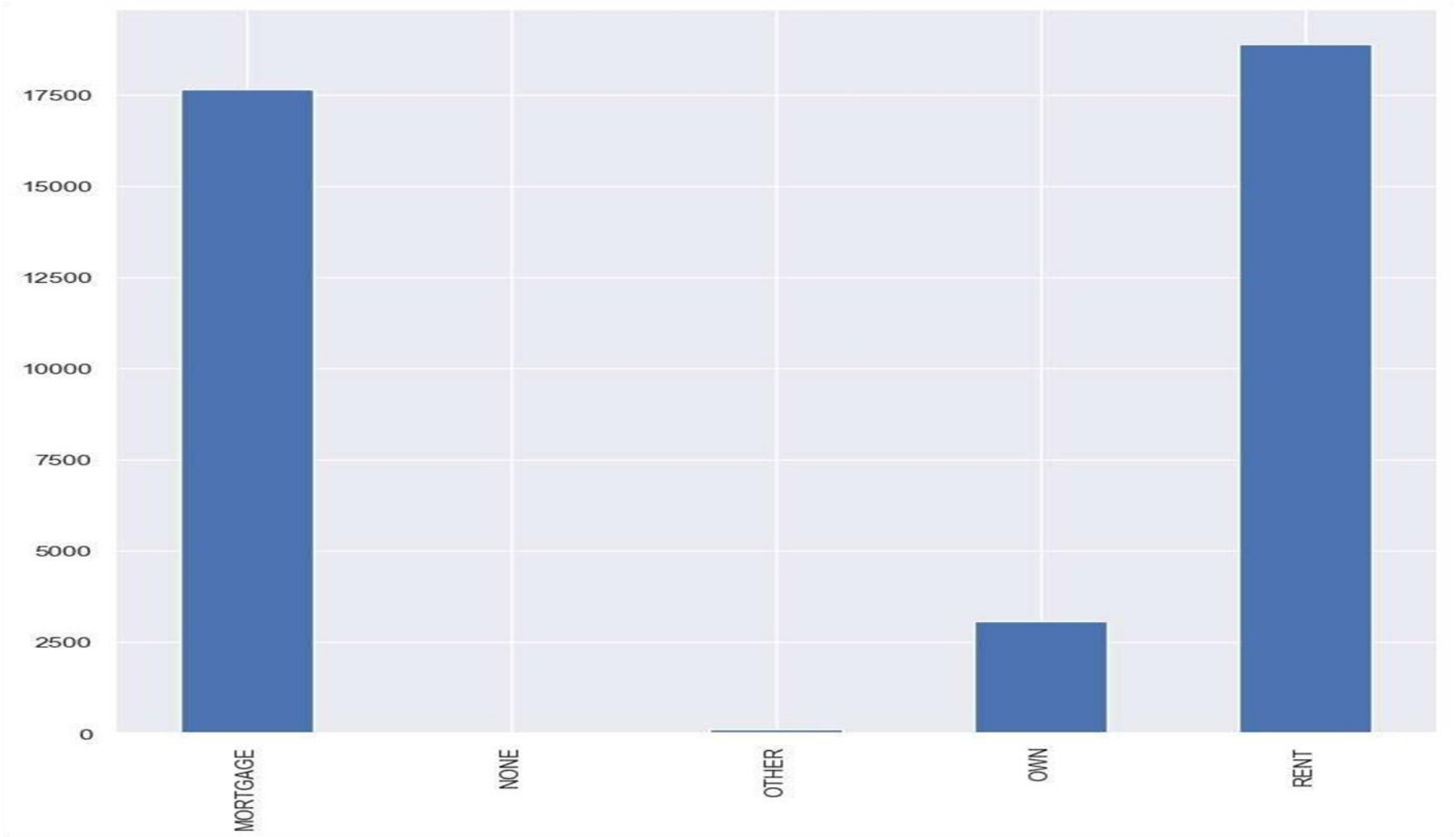
GRADE WISE DISTRIBUTION OF ALL LOANS

DF['GRADE'].VALUE_COUNTS().SORT_INDEX().PLOT.BAR()



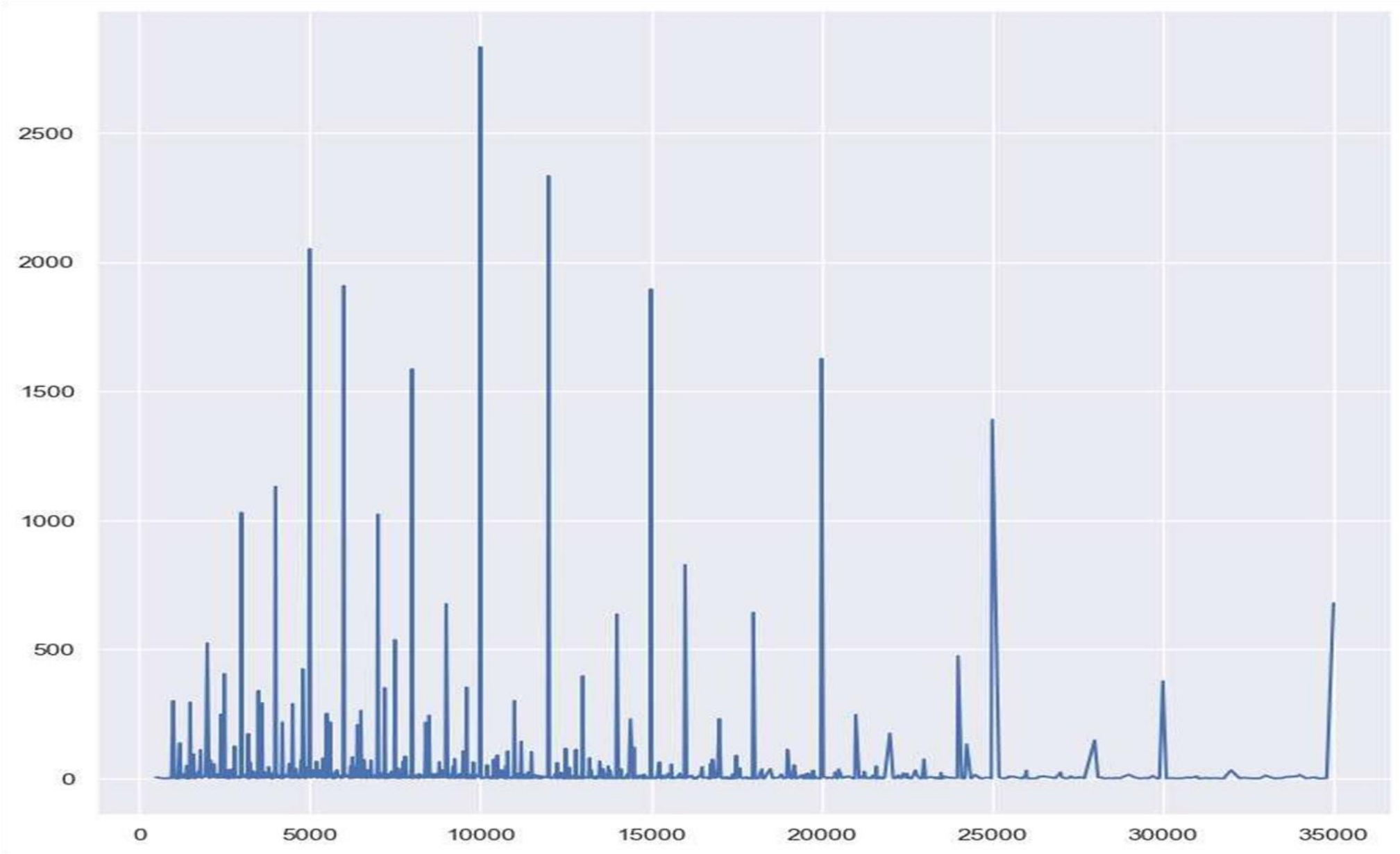
DISTRIBUTION OF THE HOUSE OWNERSHIP STATUS OF THE LOAN AVAILERS

DF['HOME_OWNERSHIP'].VALUE_COUNTS().SORT_INDEX().PLOT.BAR()



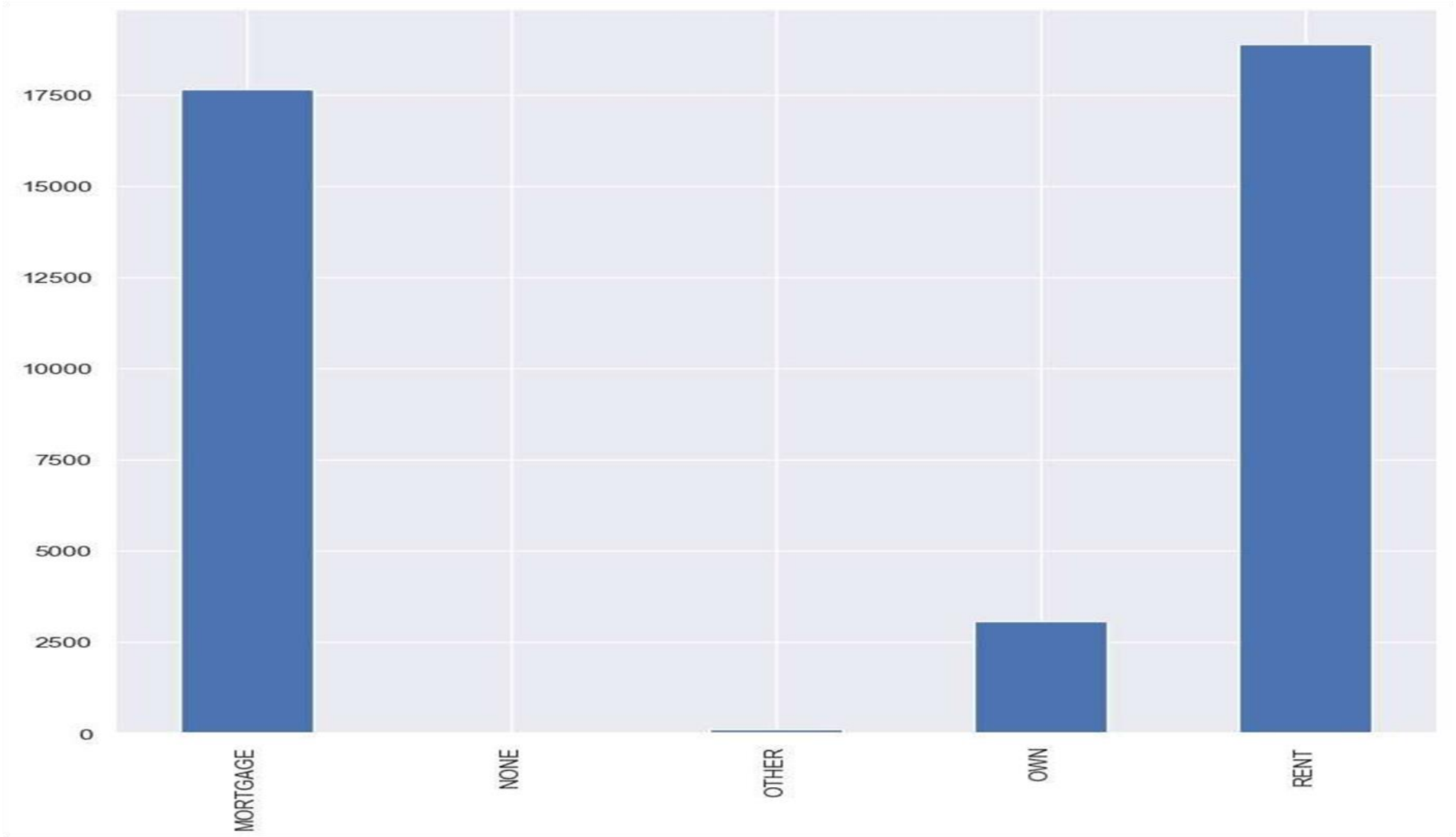
DISTRIBUTION OF THE LOAN AMOUNTS IN THE DATASET

DF['LOAN_AMNT'].VALUE_COUNTS().SORT_INDEX().PLOT.LINE()



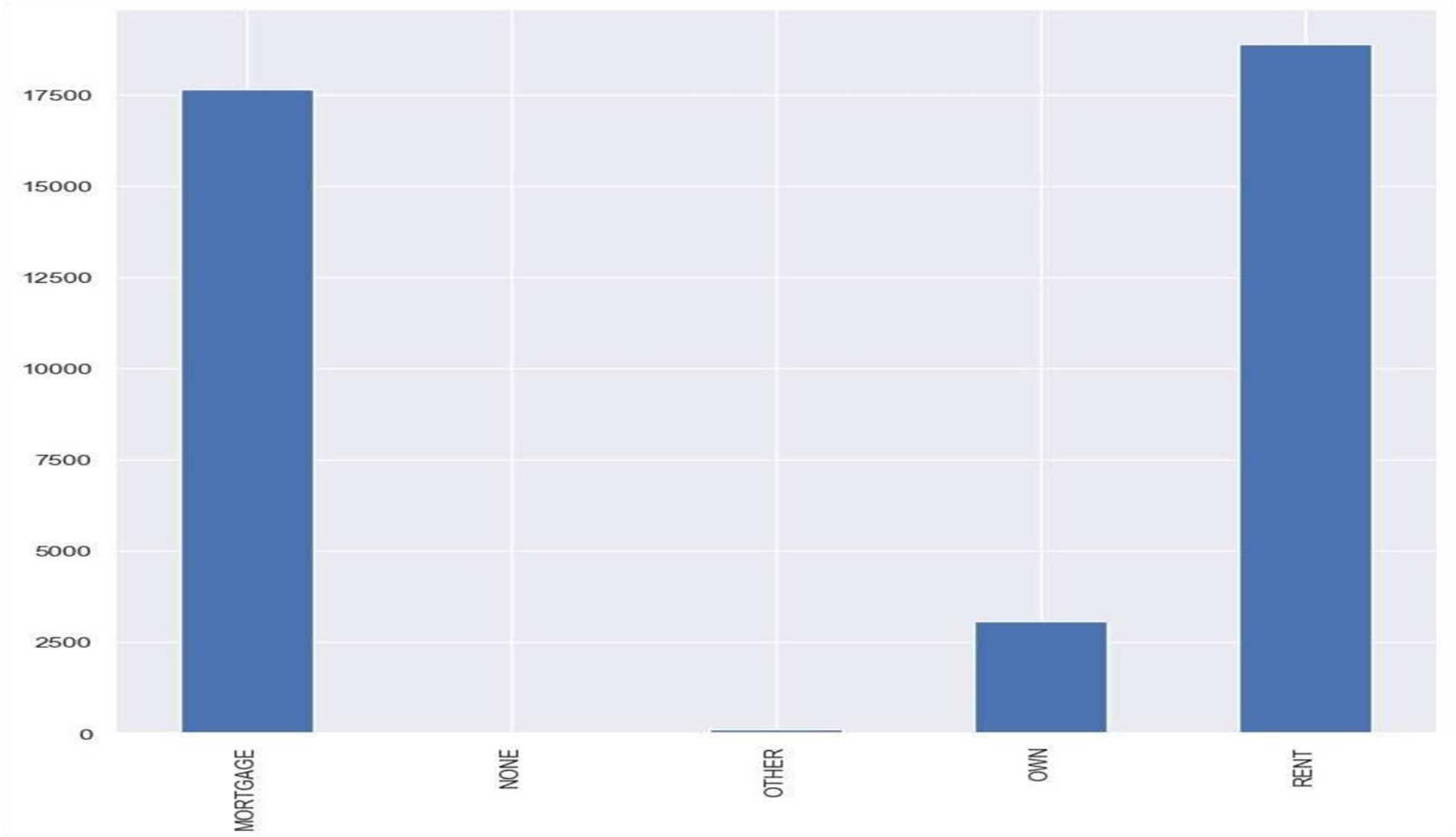
DISTRIBUTION OF THE HOUSE OWNERSHIP STATUS OF THE LOAN AVAILERS

DF['HOME_OWNERSHIP'].VALUE_COUNTS().SORT_INDEX().PLOT.BAR()



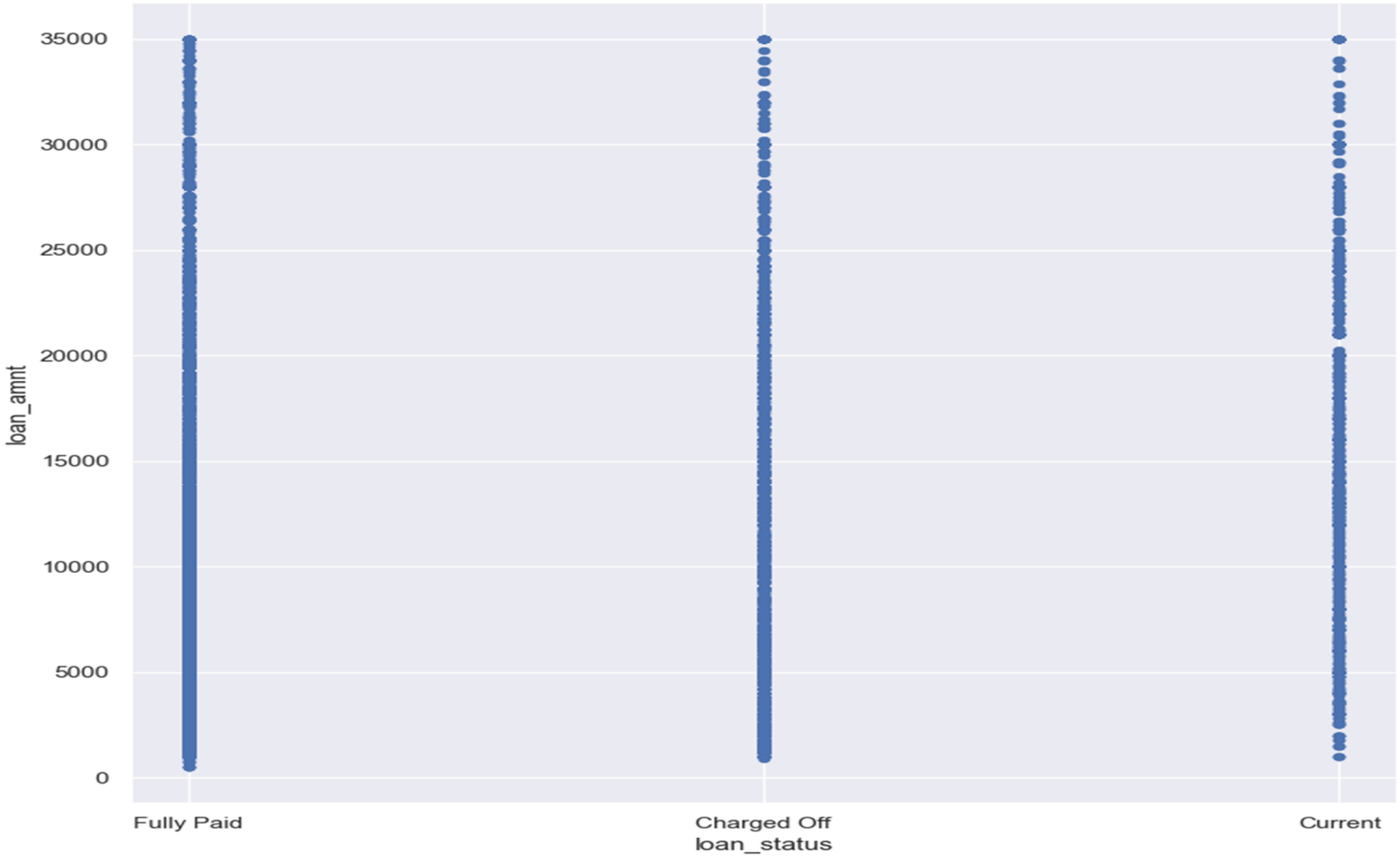
DISTRIBUTION OF THE HOUSE OWNERSHIP STATUS OF THE LOAN AVAILERS

DF['HOME_OWNERSHIP'].VALUE_COUNTS().SORT_INDEX().PLOT.BAR()



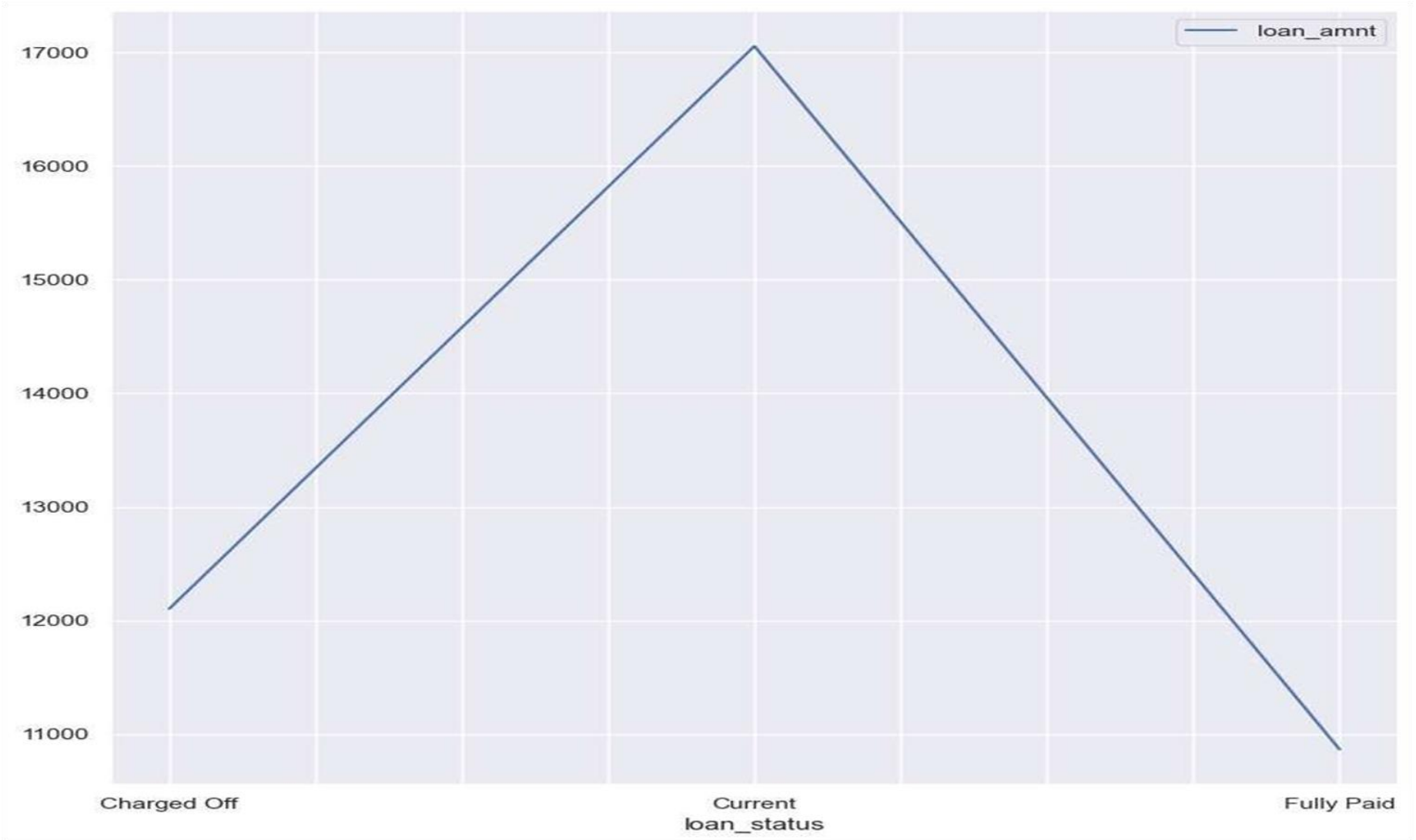
Bi-Variate Analysis

DEFAULT LOAN NUMBERS REDUCE AS THE LOAN AMOUNT INCREASES DF.PLOT.SCATTER(X='LOAN_STATUS', Y='LOAN_AMNT')

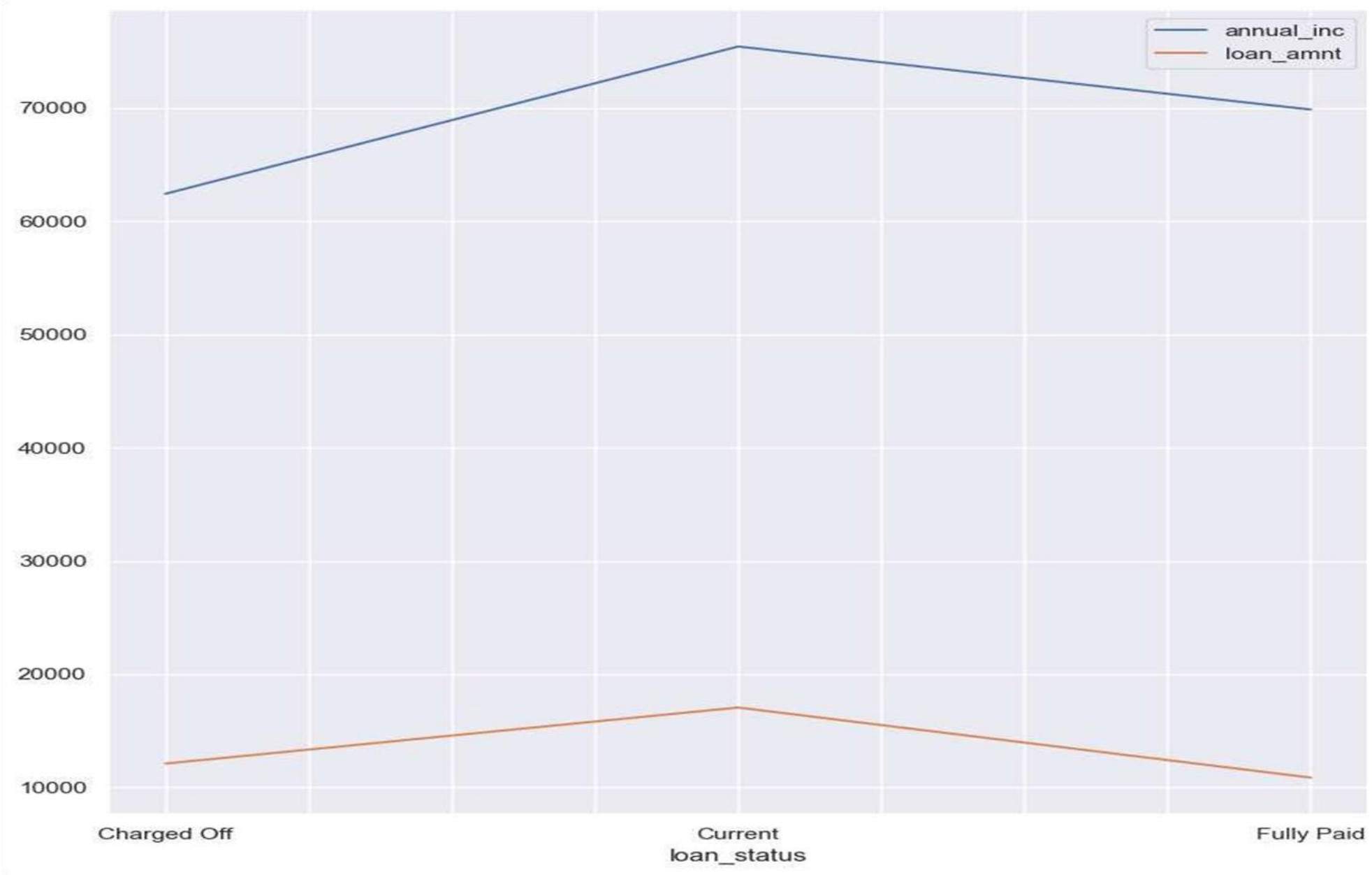


CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE LOAN AMOUNT IS LESSER

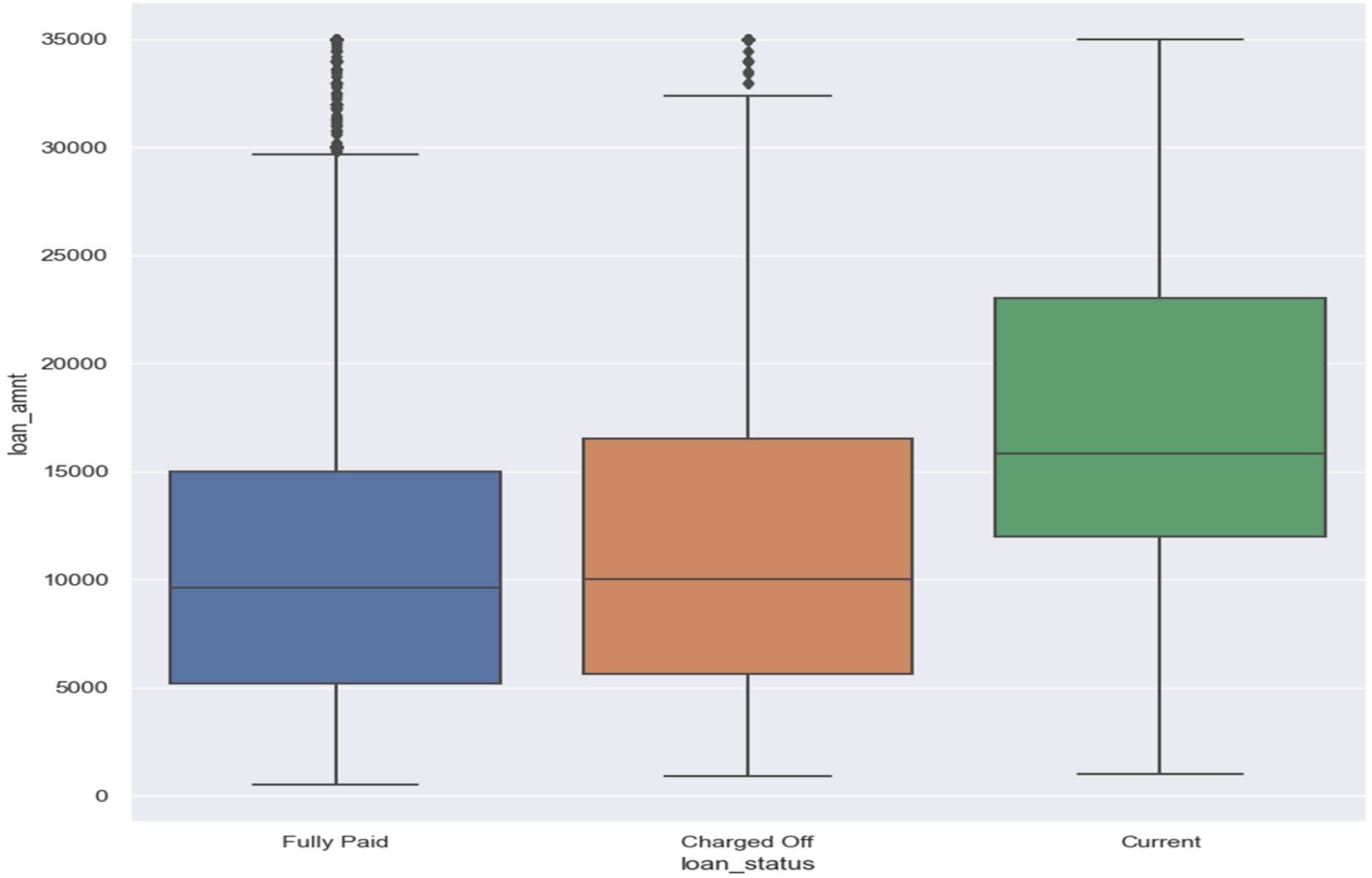

```
DFLINE = DFNEW.GROUPBY('LOAN_STATUS').MEAN()[['LOAN_AMNT']]
DFLINE.PLOT.LINE()
```



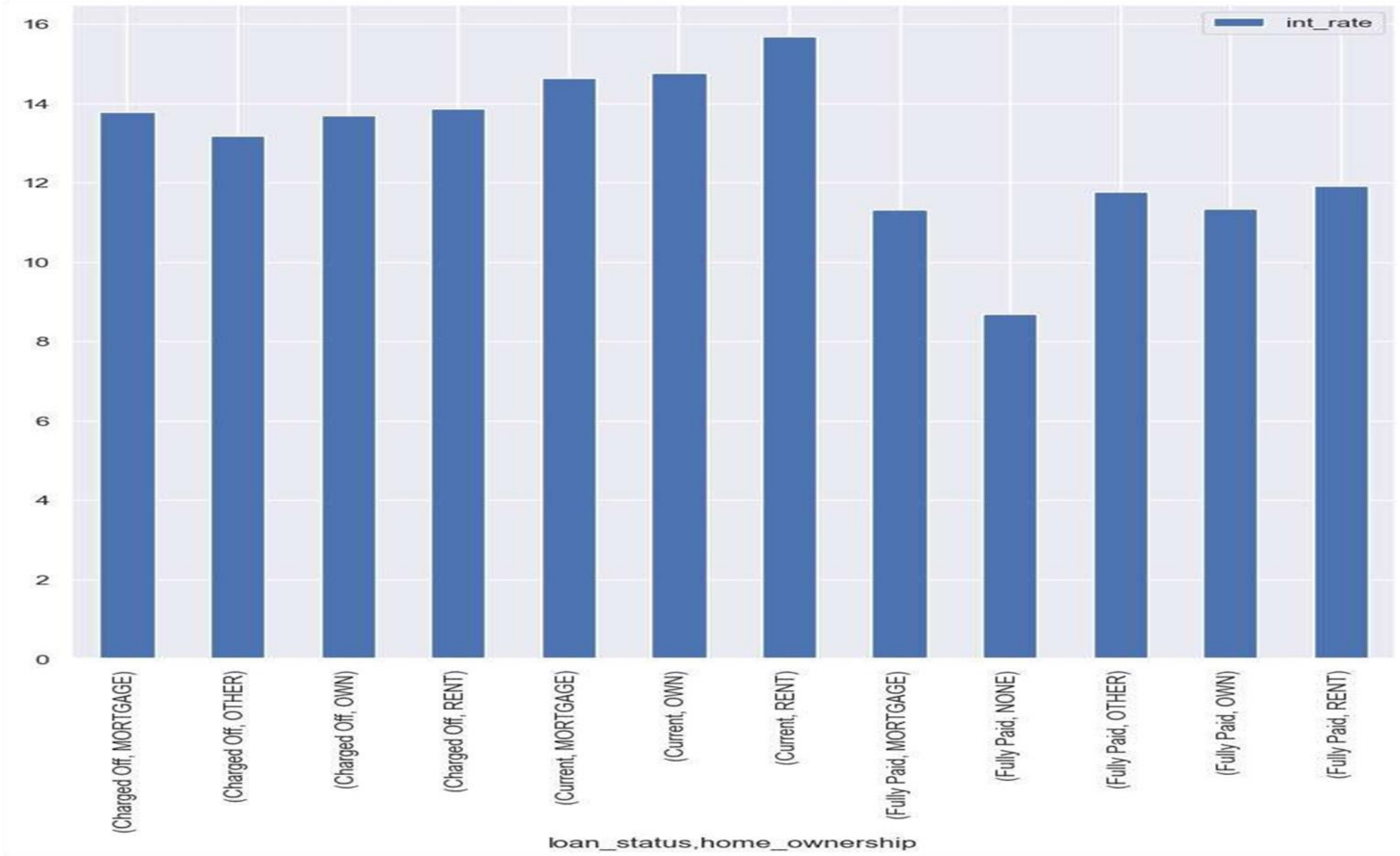
CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE LOAN AMOUNT IS LESSER



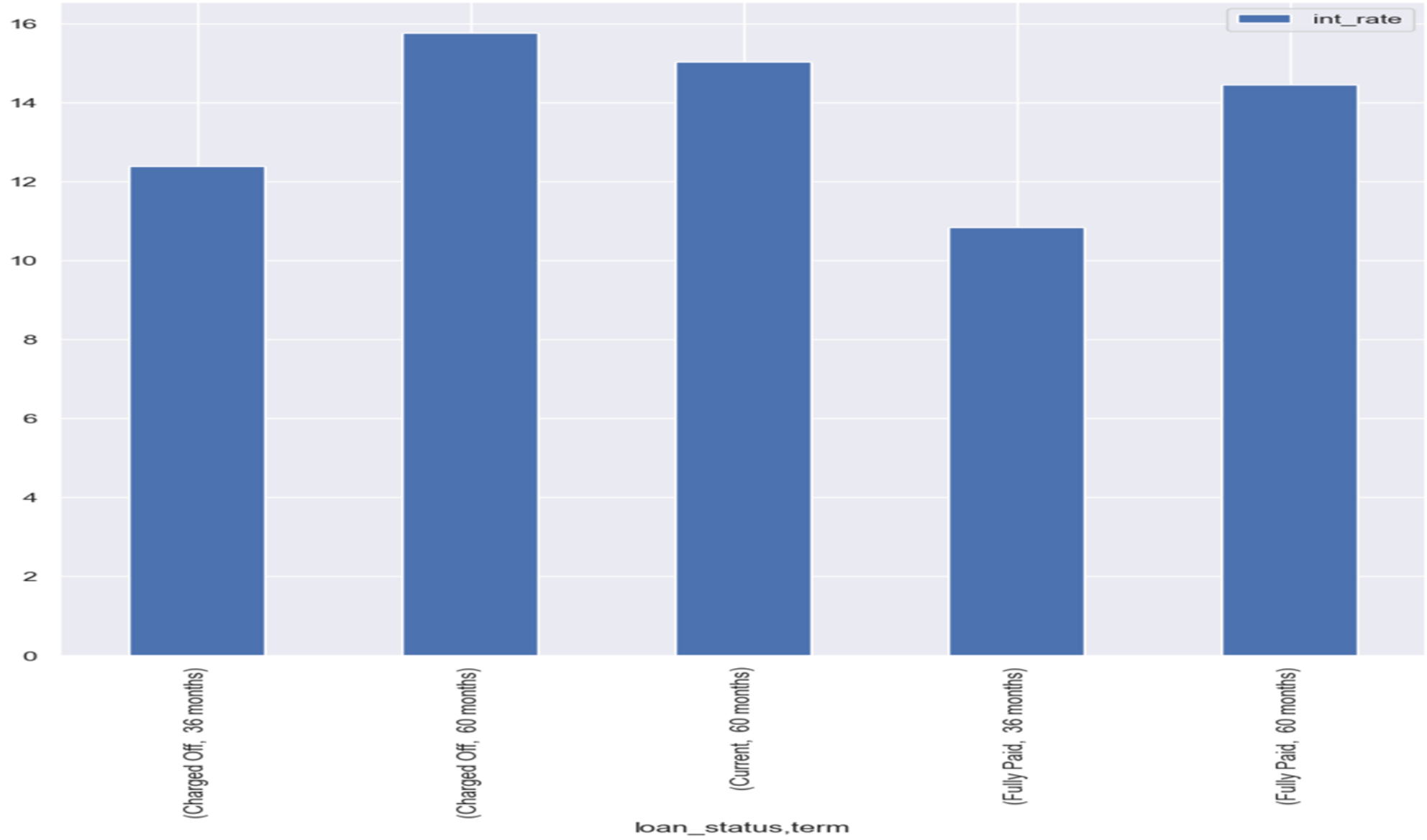
CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE LOAN AMOUNT IS IN THIS RANGE



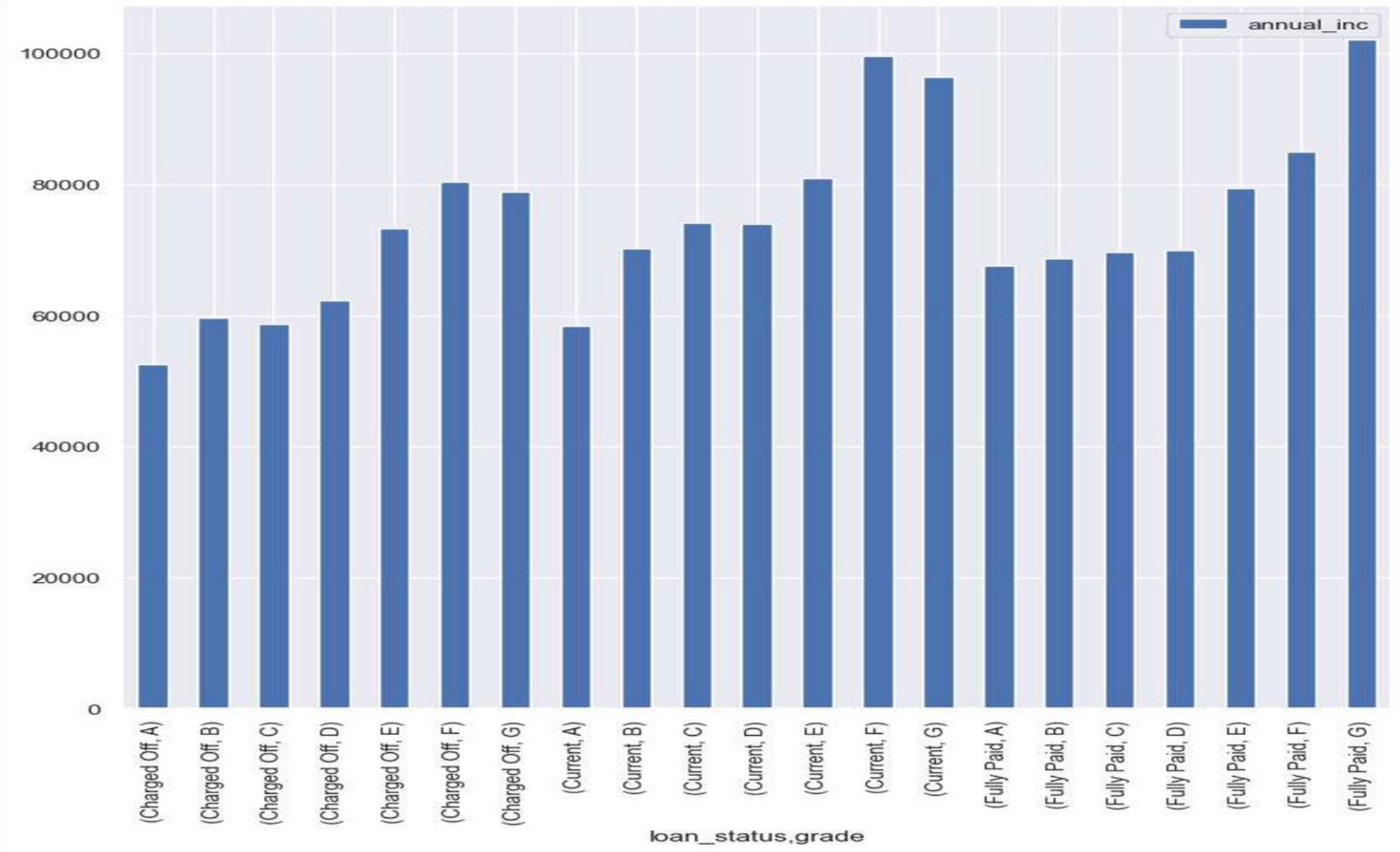
CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE HOME OWNERSHIP STATUS IS IN RENT OR MORTGAGED



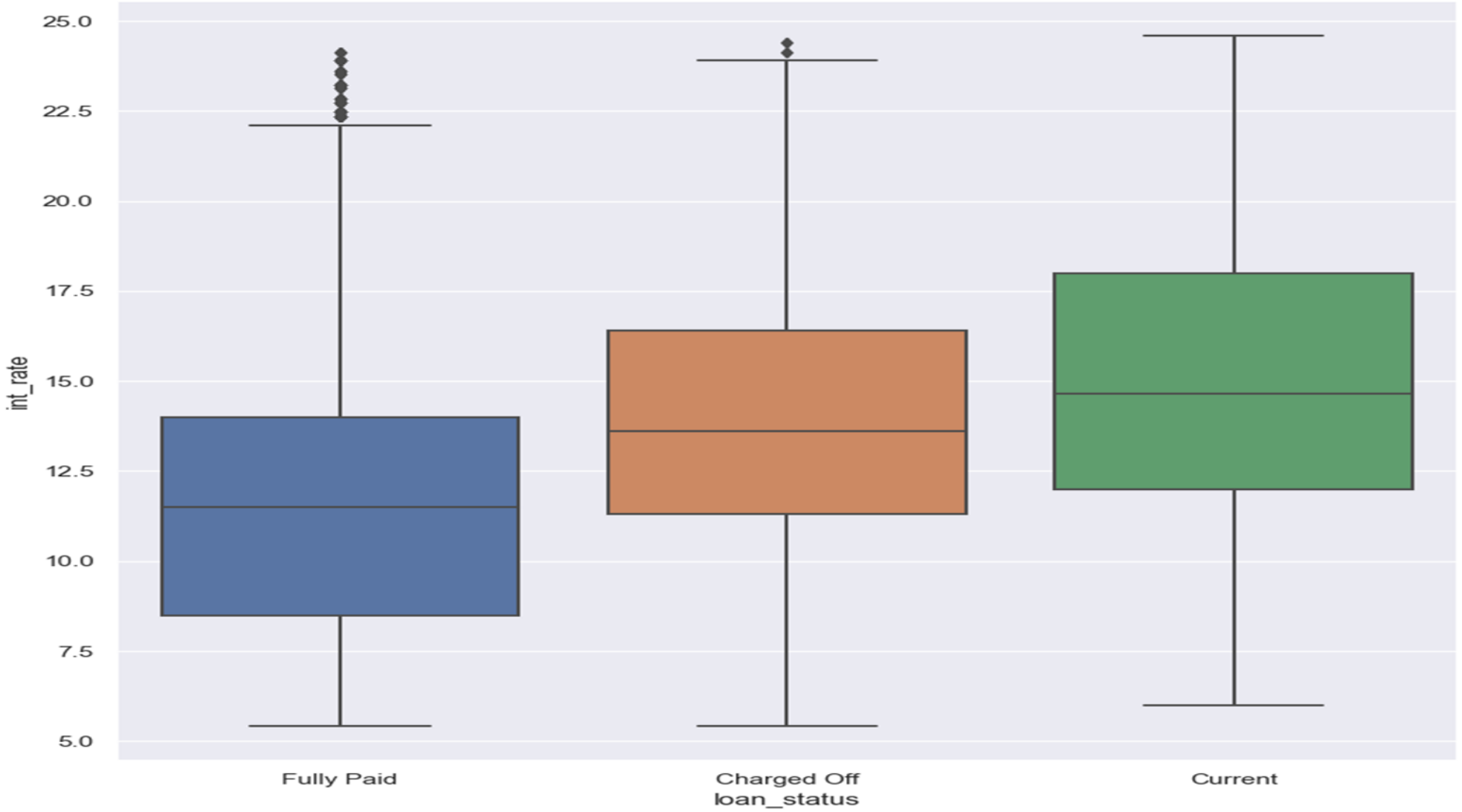
CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE TERM IS MORE



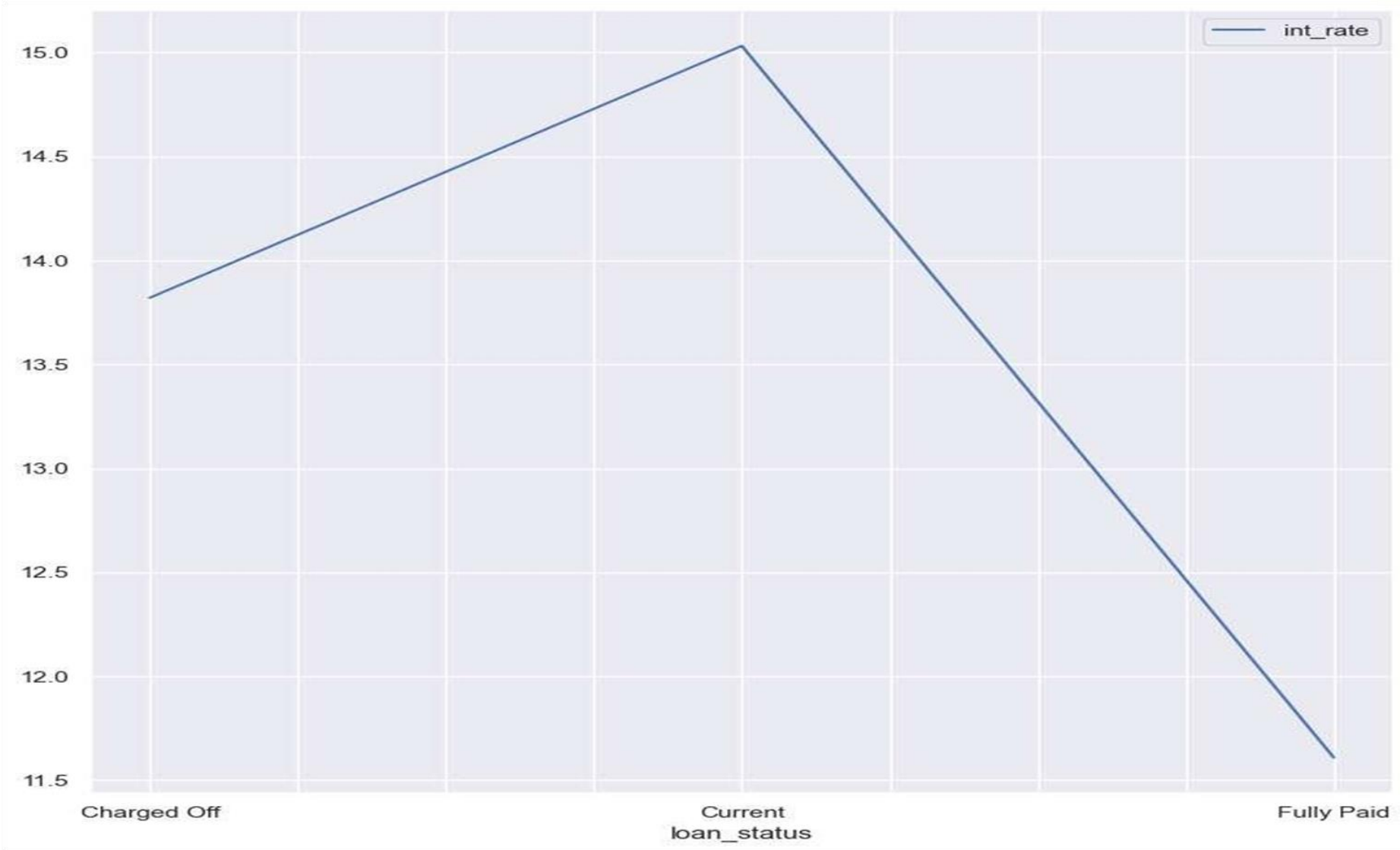
CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE TERM IS MORE



CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE INTEREST RATE RANGE IS 11.25 – 16.25

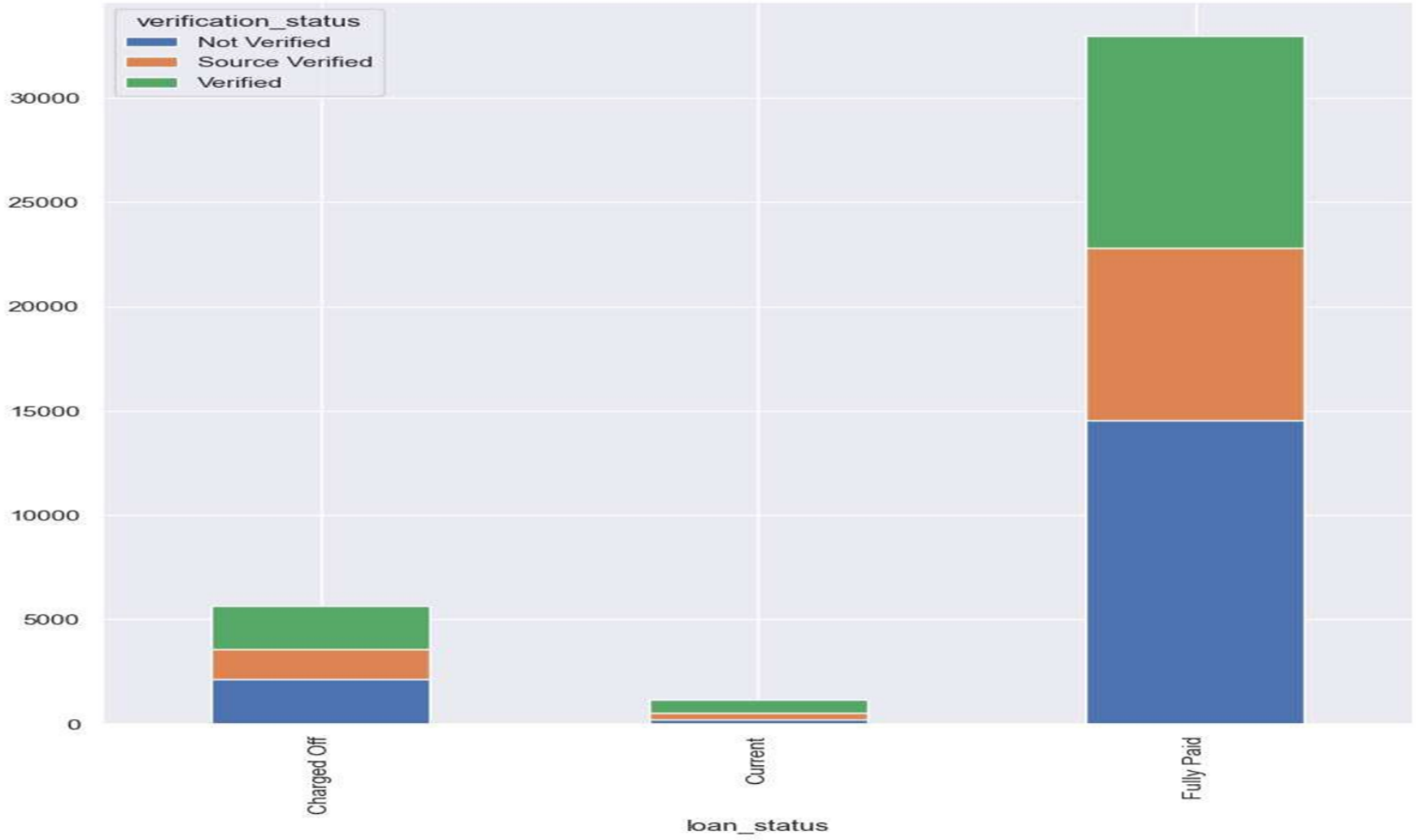


CHARGED OFF LOANS HAVE MORE OCCURRENCE WHEN THE INTEREST RATE RANGE IS 11.25 – 16.25



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.