CASE STUDY

By: Shagufta Naaz Shaikh & Syed Zuhair Abbas

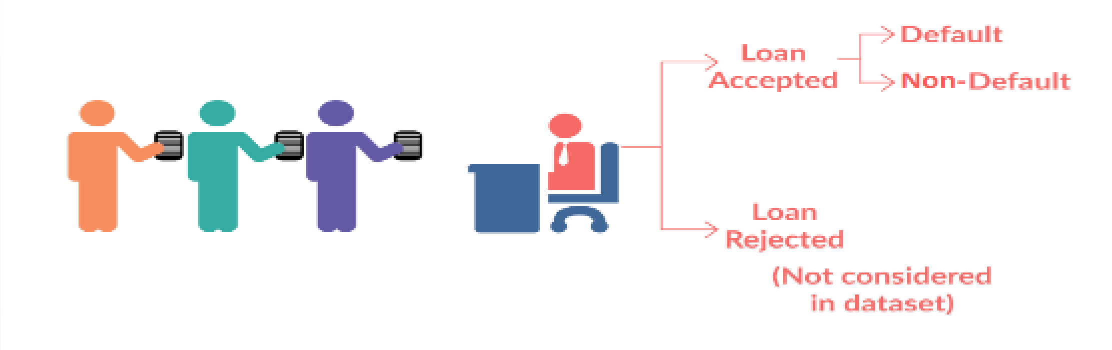
Batch :ML C56

PROBLEM STATEMENT:

- A consumer finance company is lending various types of loan to the urban customer.
- When a company receives a loan application it has to make a decision for loan approval based on applicants profile.
- Two types of risks associated with the Bank’s Decision

1. If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company.
2. 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.

LOAN DATASET



- When a person applies for a loan, there are **two types of decisions** that could be taken by the company:
- **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.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.

- Business Objectives
- This company 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.
- Like most other lending companies, 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'.
- If **one** is able to identify these risky loan applicants, then 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.
- In other words, the company wants 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.

APPROACH:

- EDA is used to understand how consumer attribute and loan attributes influence tendency of default.
- Data Reading.
- Data Cleaning is then most important task, bad quality data can be detrimental to processes and analysis.
- Data Filtering limiting rows and columns one of the most frequent data manipulation operation.
- Data Analysis forms the foundational bedrock upon which data exploration and comprehension are build.

1. **Univariate Analysis:**
    - Ordered and Un ordered
    - Quantitative Variables

2. **Bivariate Analysis** compare more than two variables has more than one dependent variable.
    - Categorical and continuous variables
    - Segmented data
        - LOAN_AMNT_RANGE,
        - EMP_LEN_RANGE
        - INTEREST_RATE_RANGE
    - Correlation Metrics
3. **Derived Metrics**

- Data set taken in Data frame is first analysis on ONLY CHARGE OFF(rslt_df) then Entire DATA FRAME (df2)

ANALYSIS OF VARIABLES:

1. **Categorical variable** are qualitative variables that represent categories or groups.

   eg. home_owner, sub_grade, grade, loan_status, verification_status,year, purpose

   - Ordered Variable: Grade, sub grade , term
   - Un ordered : Loan status, verification status , purpose ,home_owner ship,addr_state

2. **Continuous variable** are quantitative variables that can take an infinite number of values within a given range.

   Quantitative : Apart from above rest like dti, installment, loan_amnt etc . Are continuous vars

3. **Segmented Data**

   - LOAN_AMNT_RANGE,
   - EMP_LEN_RANGE
   - INTEREST_RATE_RANGE

4. **Derived Data** : For the purpose of this case study following data was derived

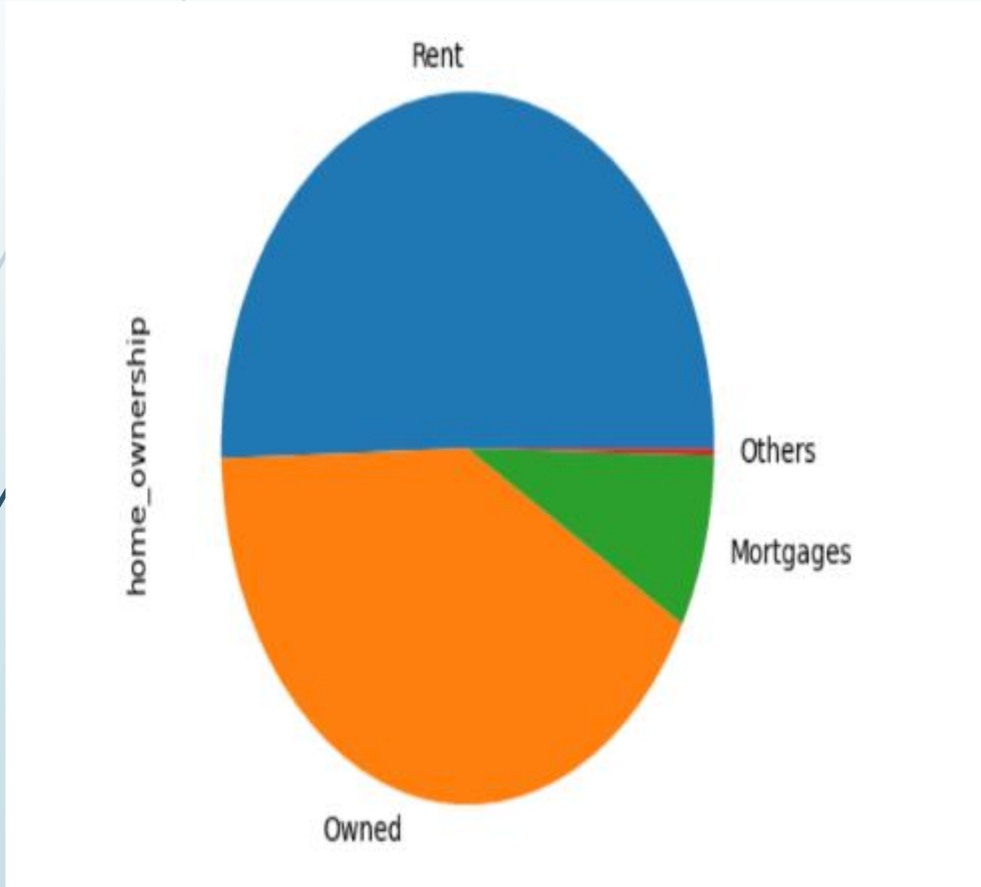
   - CHARGE_OFF_PROPORTION
   - YEAR
   - EMPLOYEE _EXPERIENCE

addr_state	The state provided by the borrower in the loan application
annual_inc	The self-reported annual income provided by the borrower during registration.
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
installment	The monthly payment owed by the borrower if the loan originates.
int_rate	Interest Rate on the loan
issue_d	The month which the loan was funded
last_pymnt_amnt	Last total payment amount received
last_pymnt_d	Last month payment was received
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan_status	Current status of the loan
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
pub_rec_bankruptcies	Number of public record bankruptcies
purpose	A category provided by the borrower for the loan request.
pymnt_plan	Indicates if a payment plan has been put in place for the loan
recoveries	post charge off gross recovery
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub_grade	LC assigned loan subgrade
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

VISUALISING AND SUMMARIS

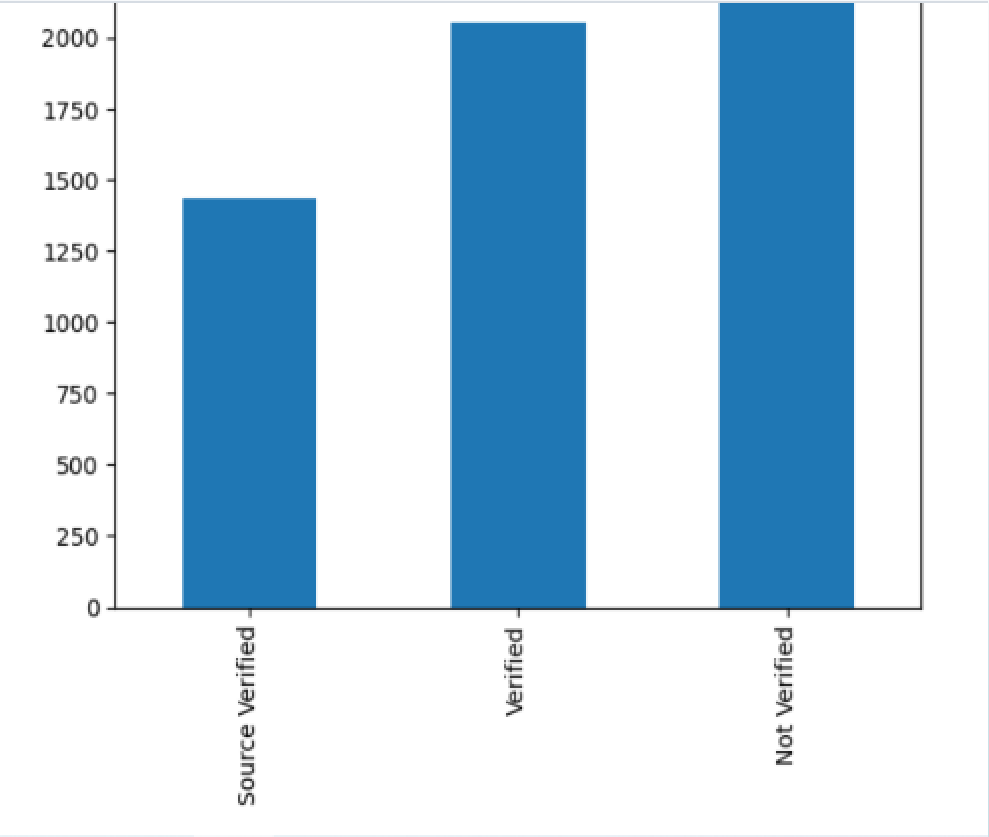
Data Distribution by Ownership

Inference : Customers who are on RENT are more likely to default and hence HIGH risk customer.



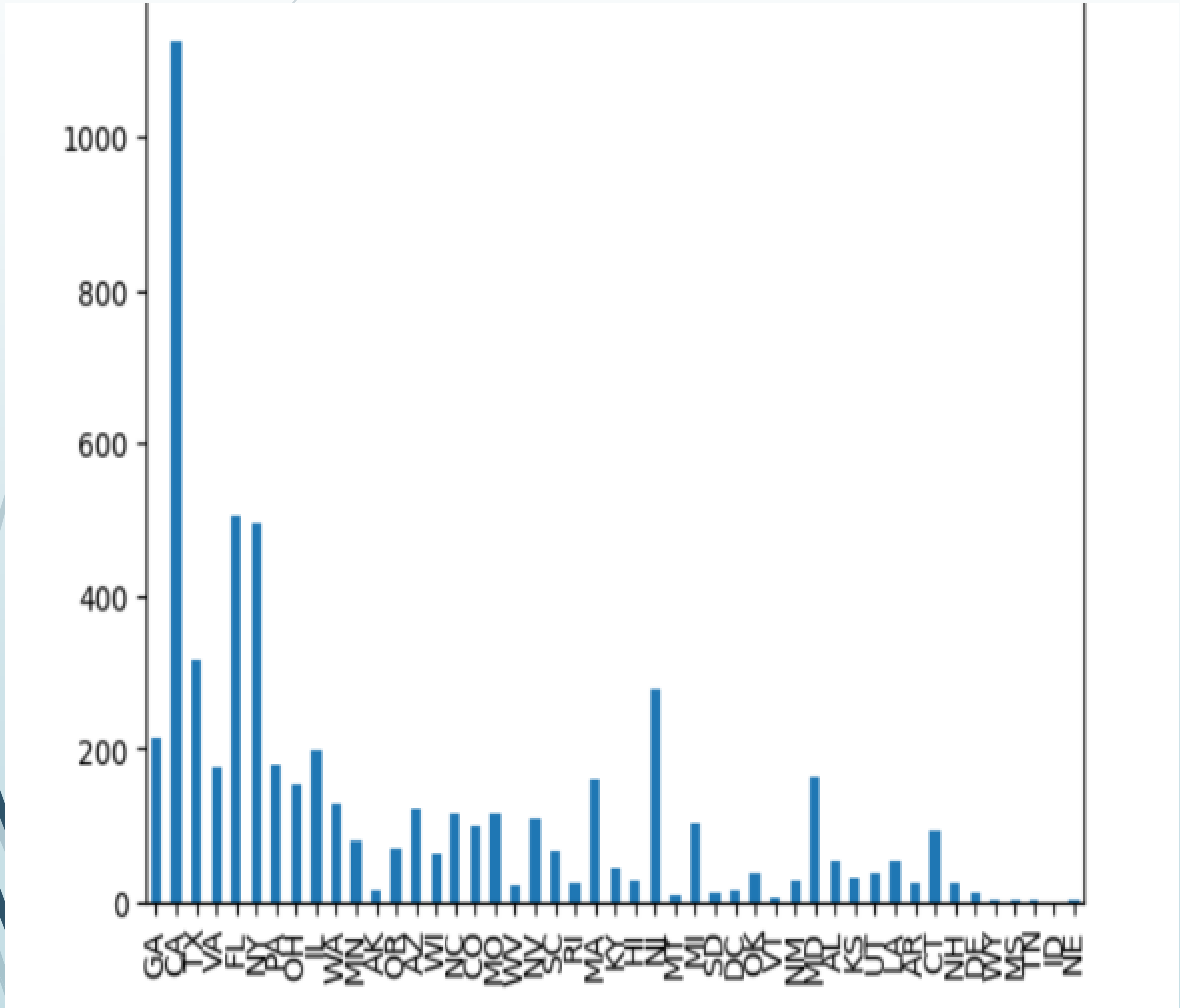
Data distribution by Verification status

Inference : Customers are NOT having proper Verification are more likely to default and hence High risk customers.



Data set is only for charge Off

Data distribution by Address State

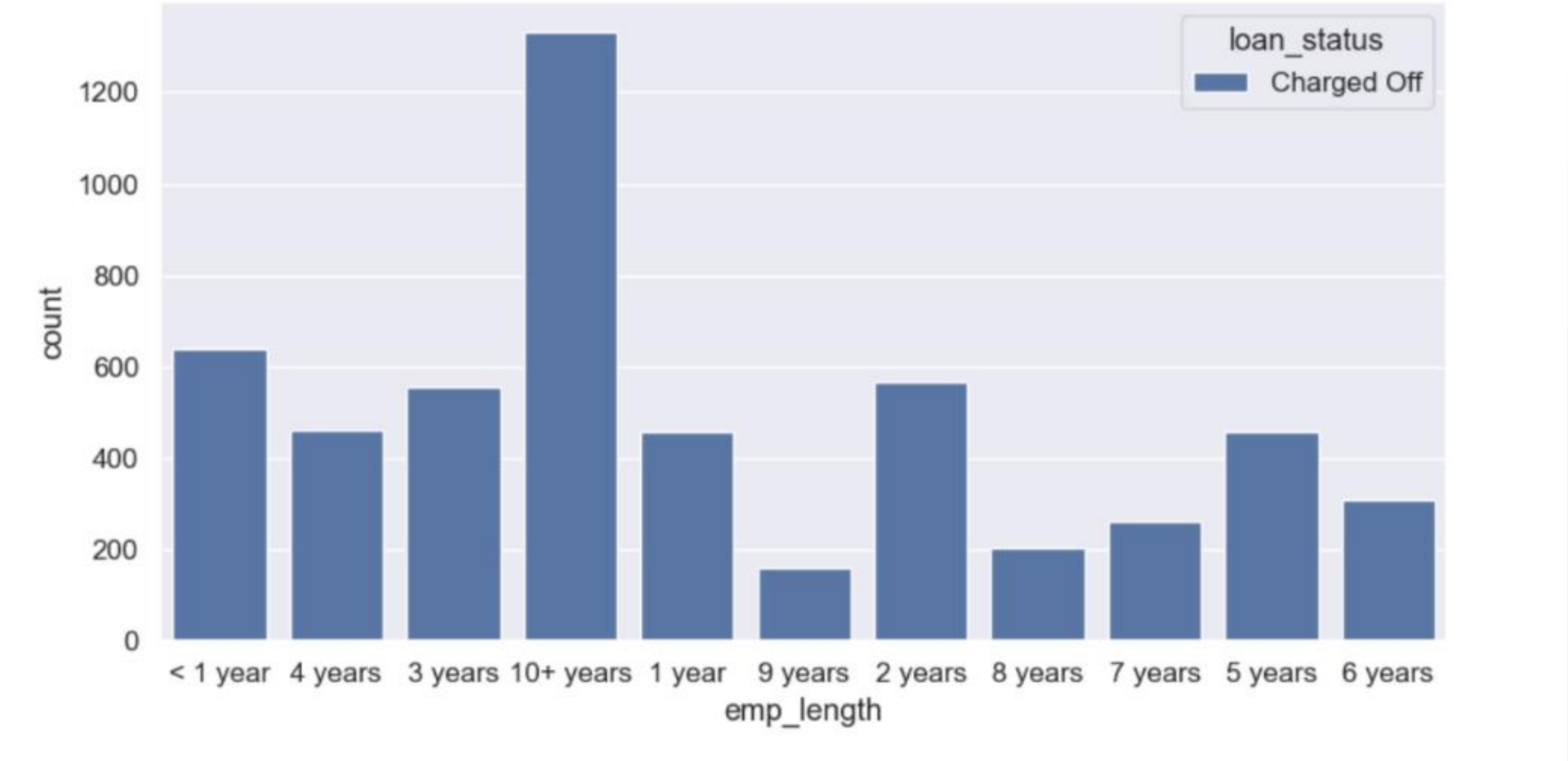


Inference : Maximum charge off are from California that means better checks in CA state and FL or NY.

Data set is only for charge Off

Inference : Maximum charge off are from either less experience or more than 10 years.

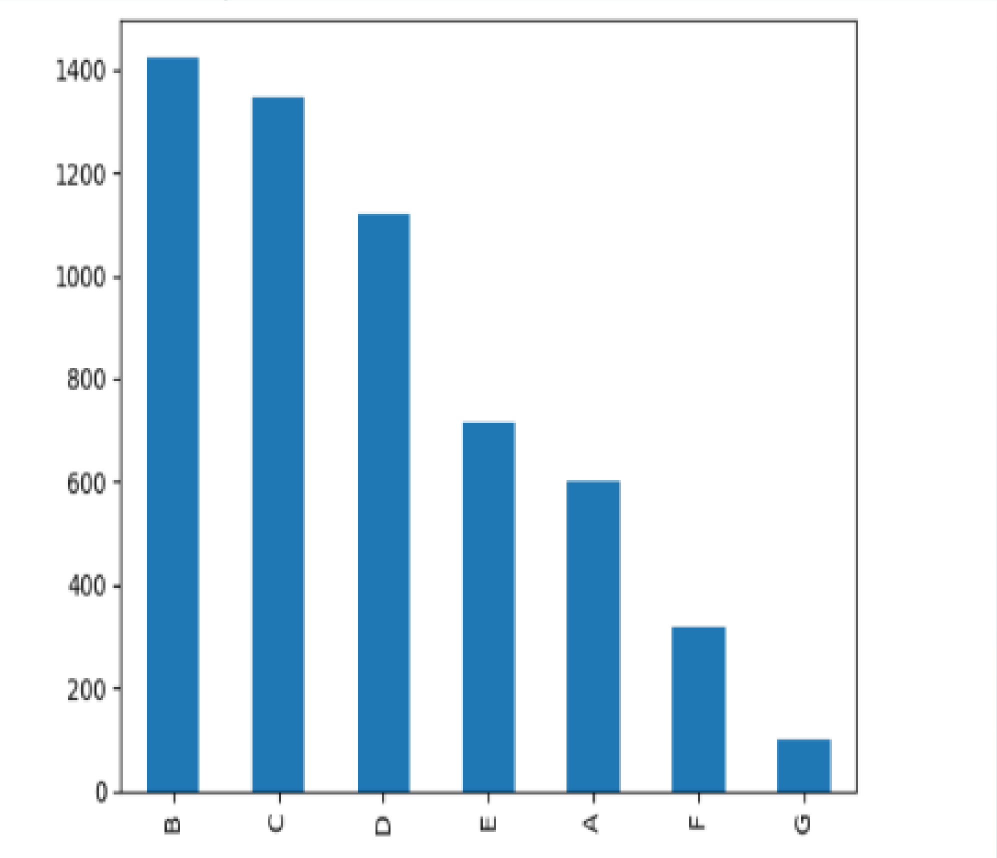
Data set is only for charge Off

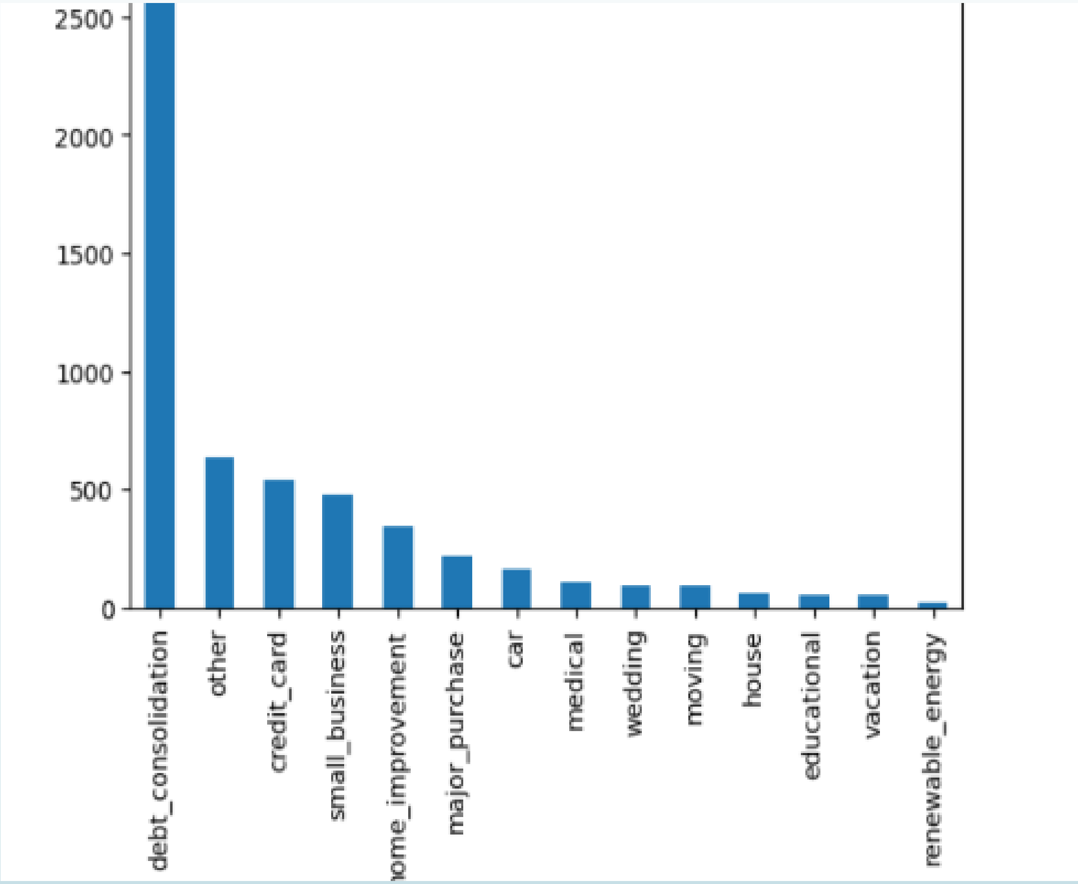


Data distribution by Grade

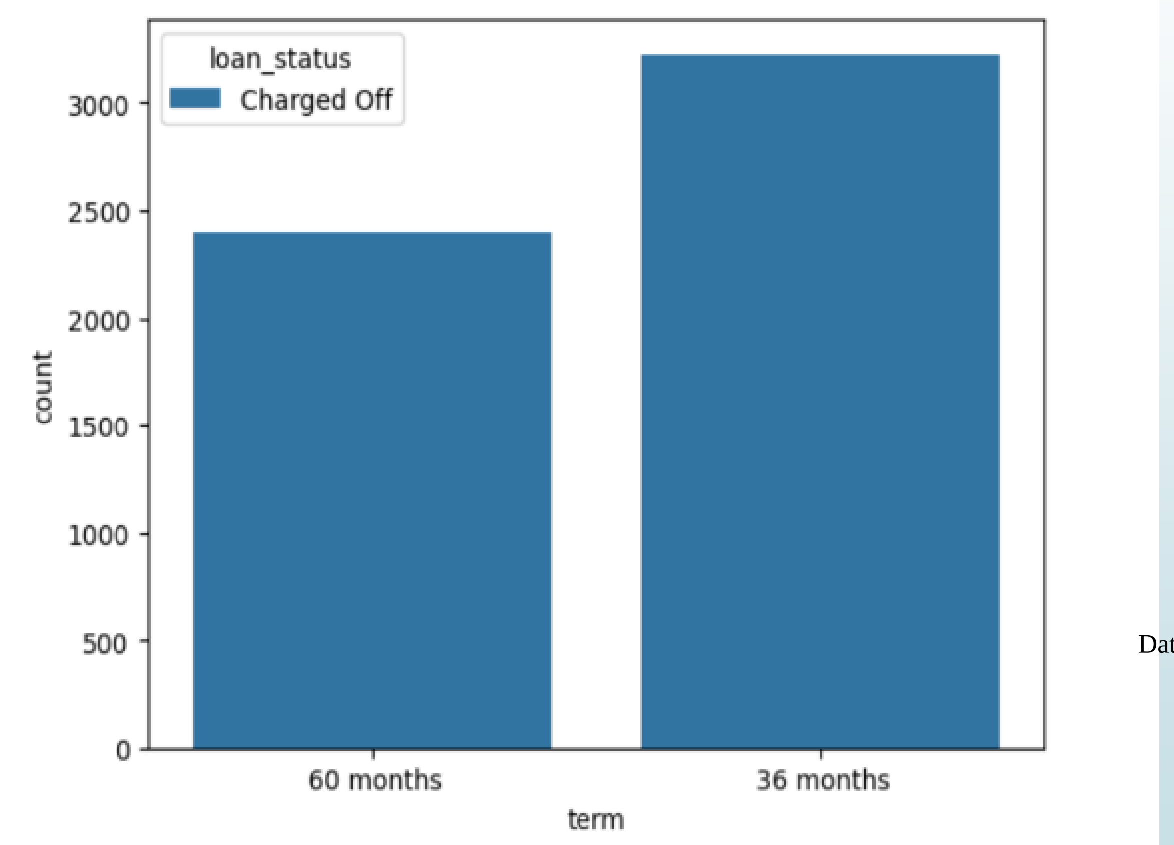
Inference : Not much visibility by this data except that B and C grades are more likely to default within defaulters.

- Inference : Debt consolidation purpose is the main category for charge off.





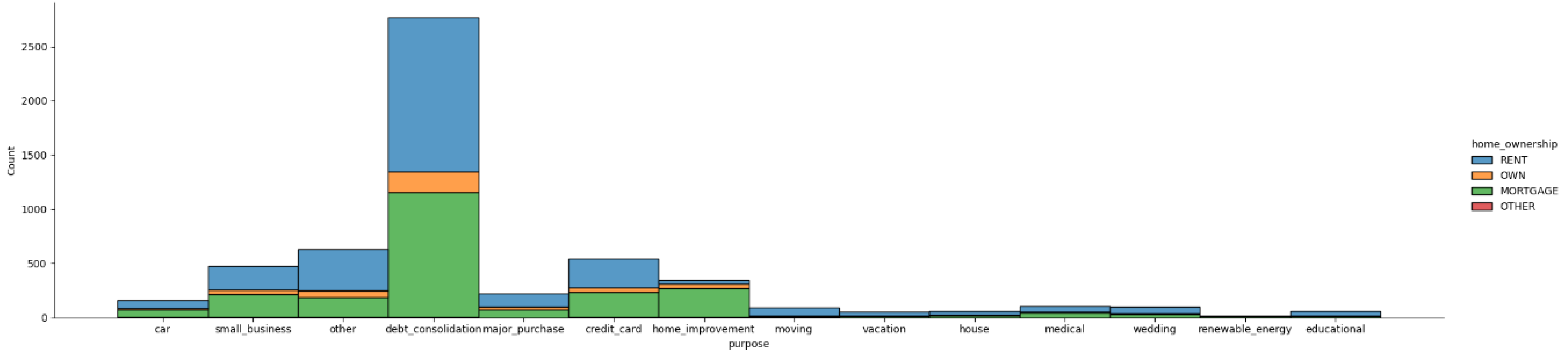
Data set is only for charge Off



Inference: This shows that candidates with less term are more defaulters.

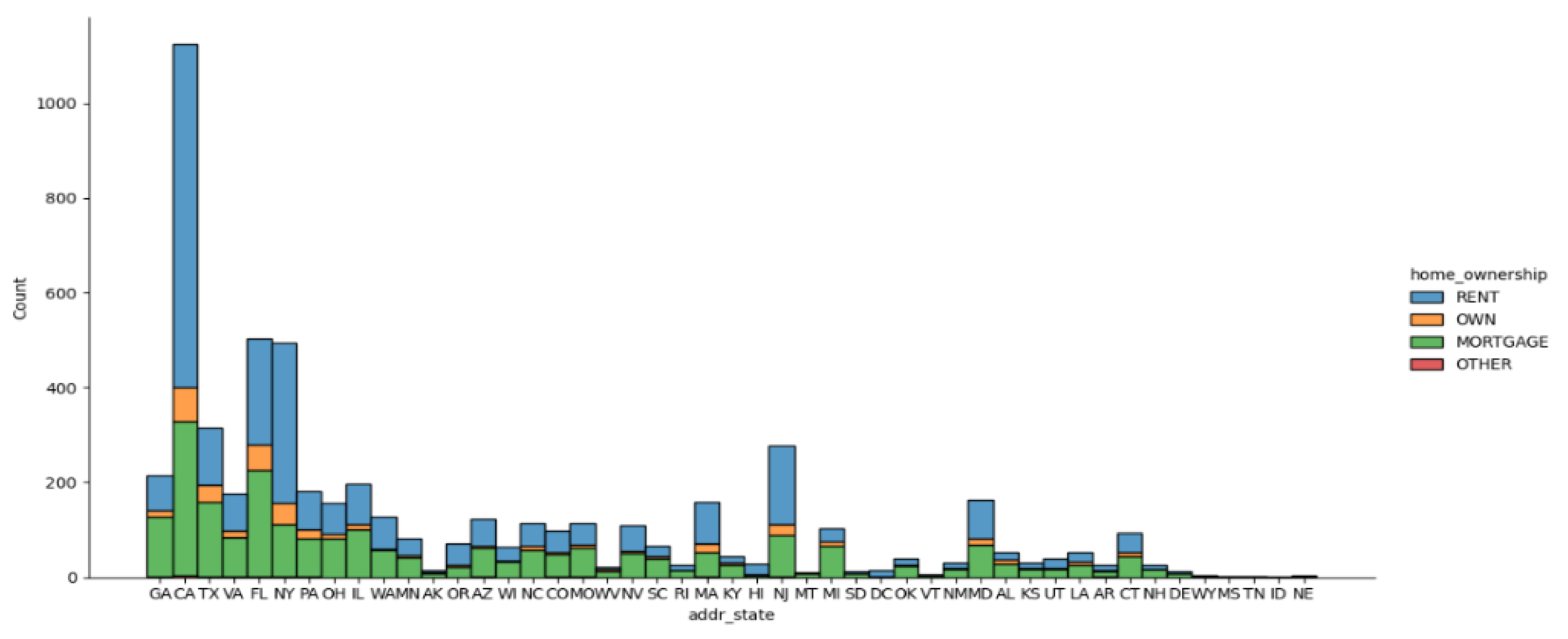
Data set is only for charge Off

Inference : Maximum charge off are from who are renting and have mortages and taken loan for purpose of debt consolidation



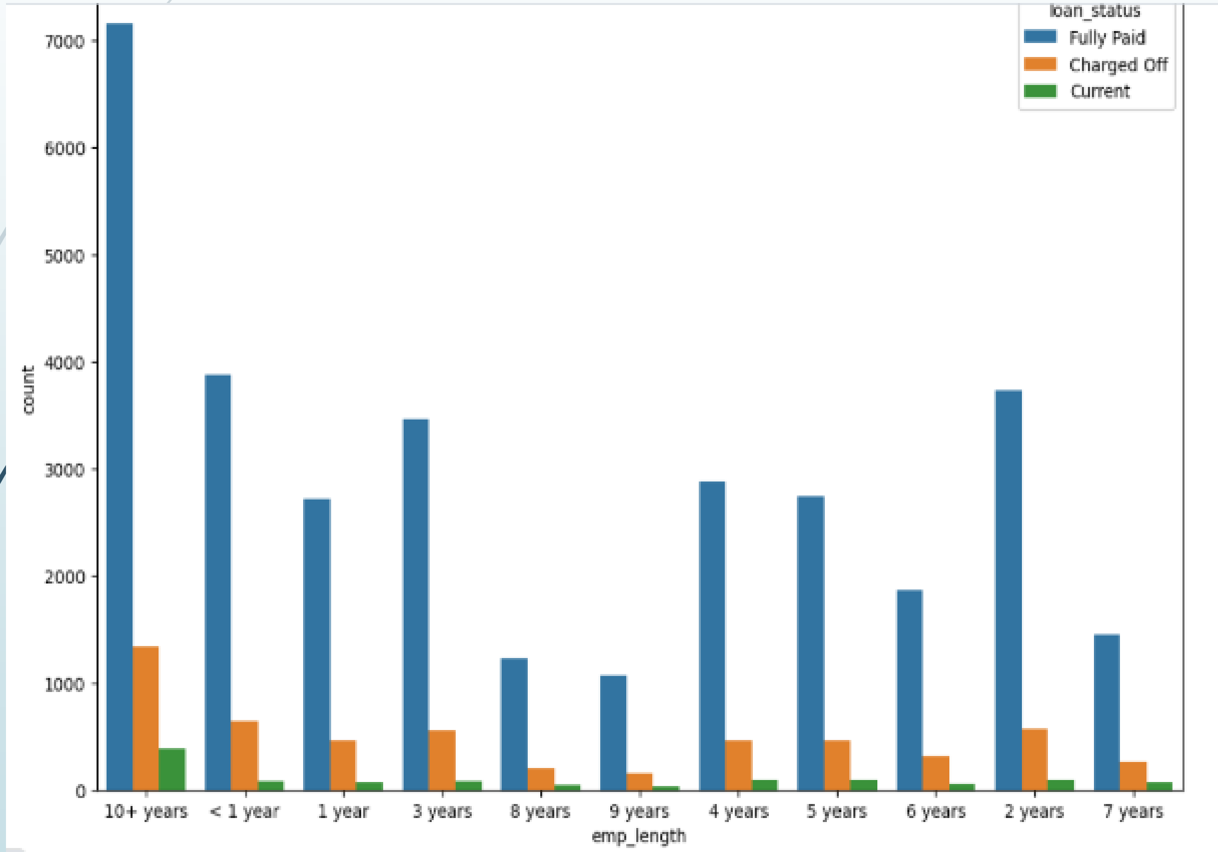
Data set is only for charge Off

Inference: Rented owners and with non verified status are likely to charge off and is prominent in CA.



Data set is complete now including all Rows

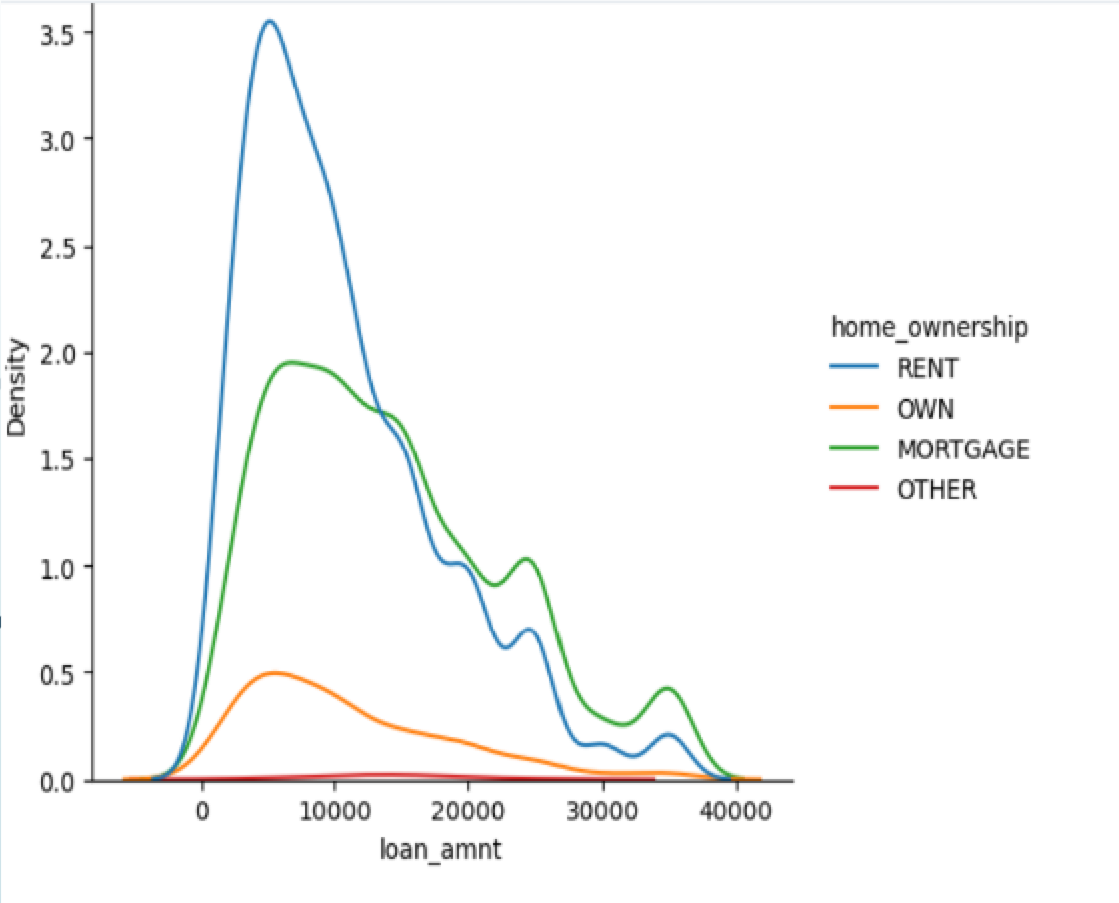
Data distribution by experience of employee.



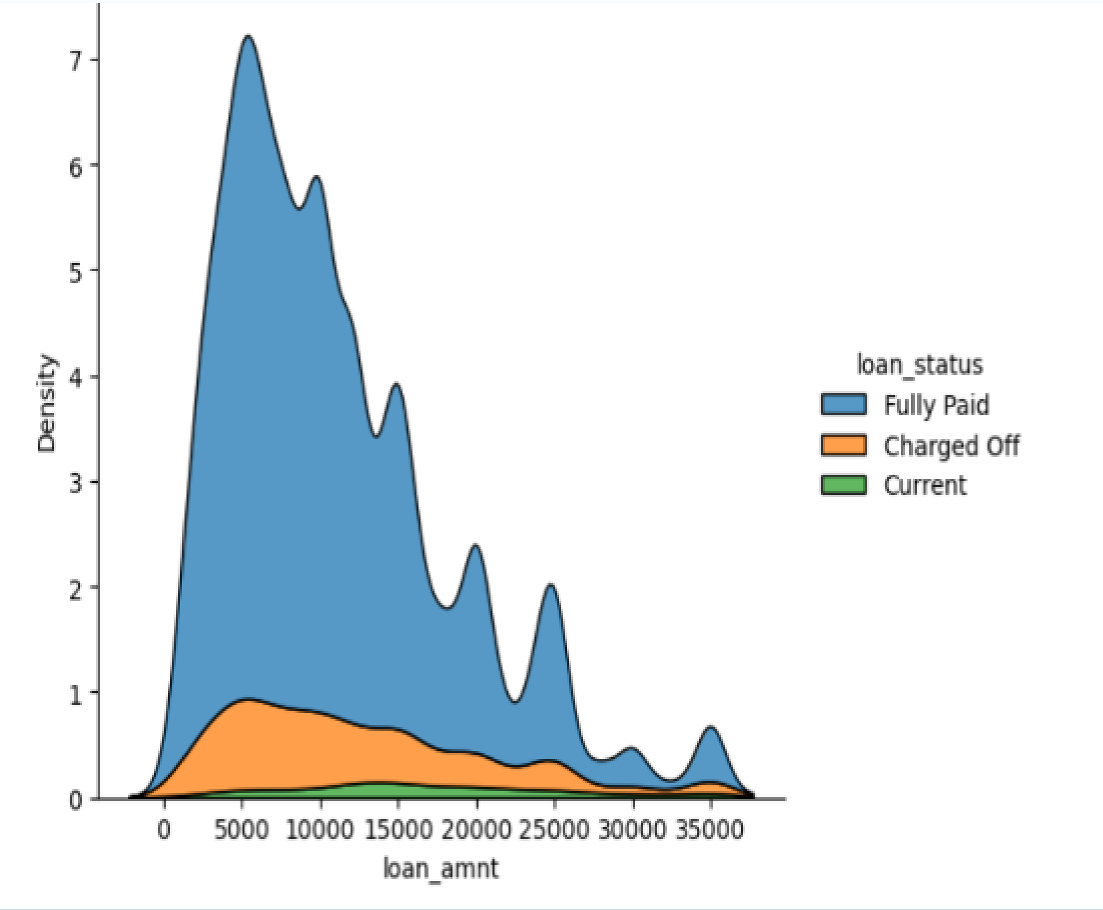
Inference: Ability to pay loan is directly proportional to number of year of service as we can see here more the experience- more he is likely to repay the loan.

Data set is complete now including all Rows

Inference : Loan amount less than 10000 with rent or mortgages have highest defaulters this needs further analysis.

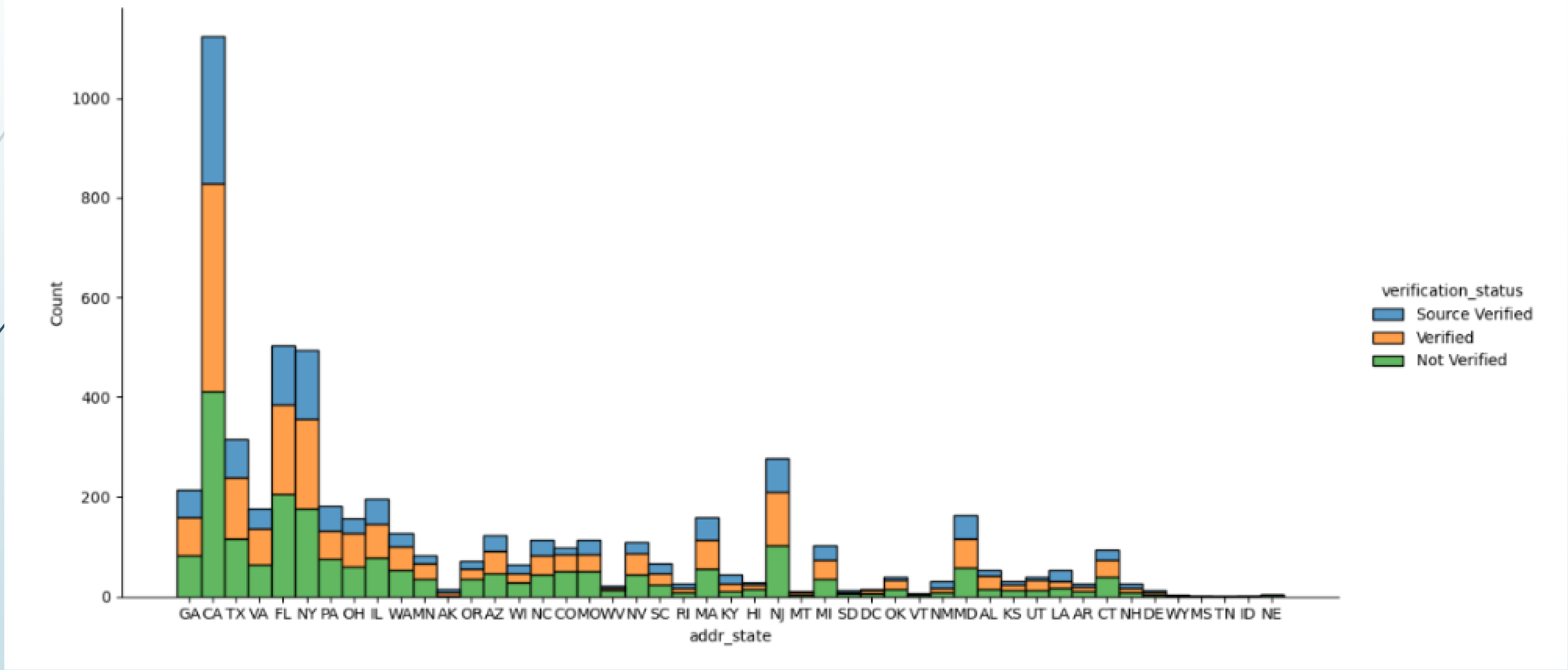


Inference : Loan amount less than 10000 have higher probability for charge off.

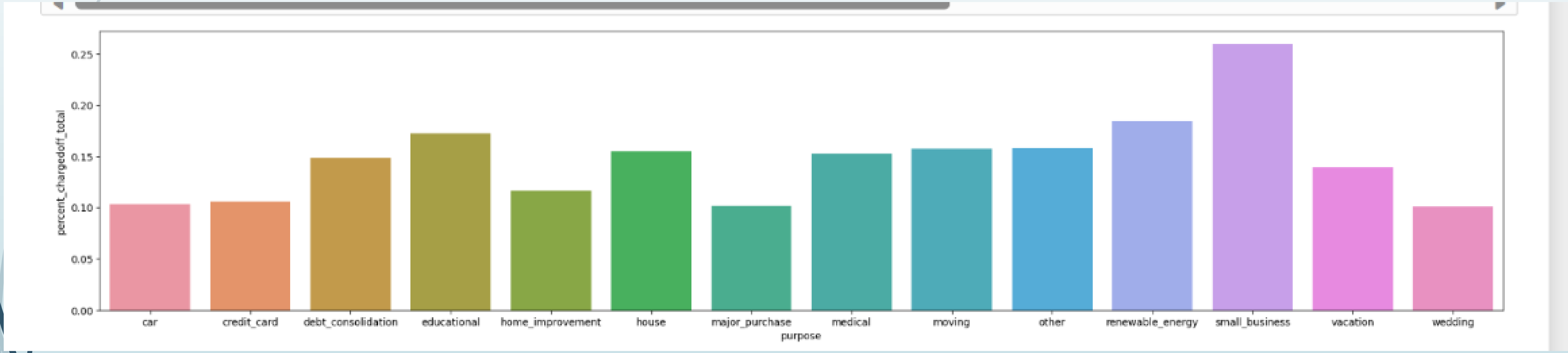


Data set is complete now including all Rows

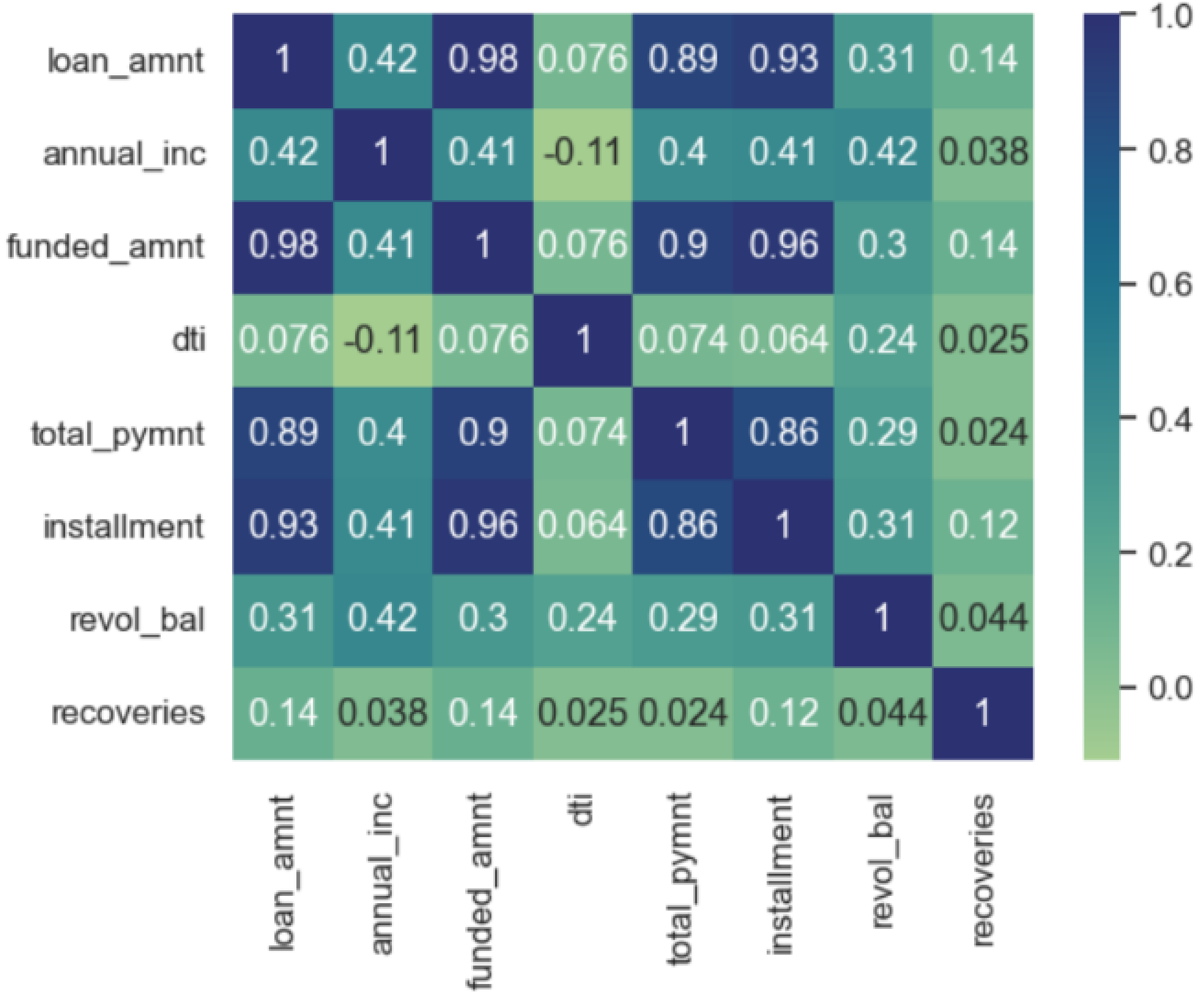
Inference: Rented owners and with non verified status are likely to charge off .



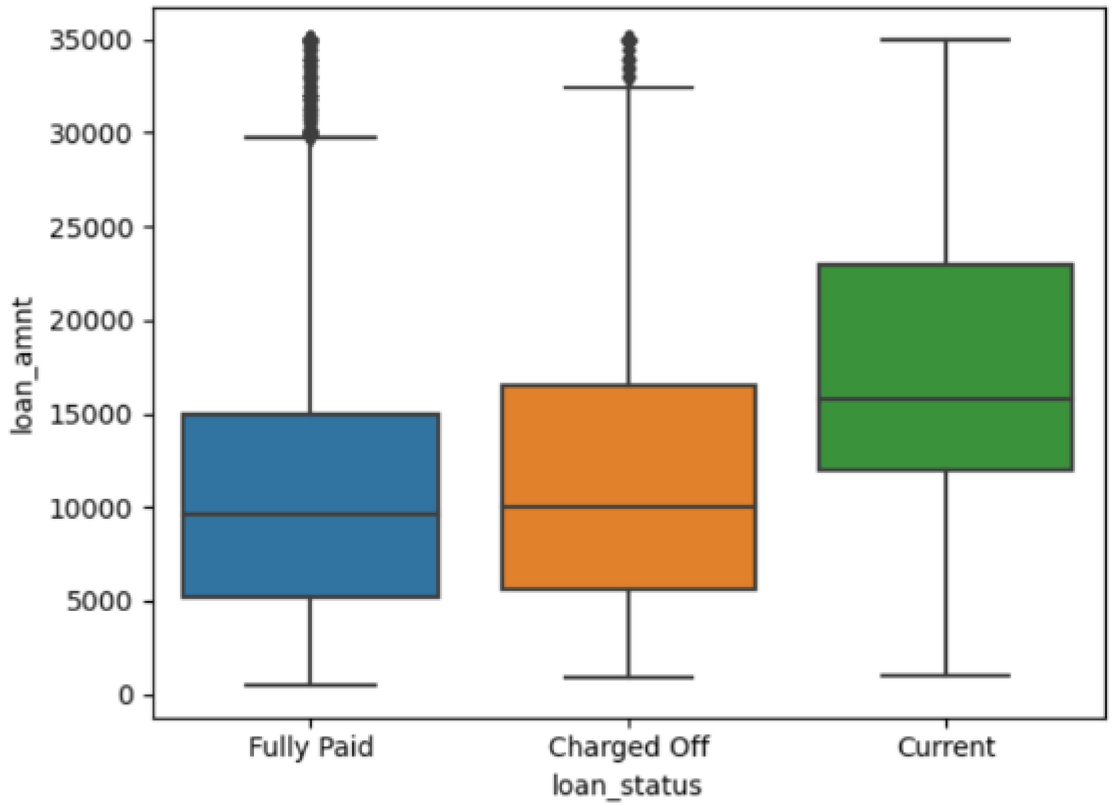
Inference: Data is grouped by purpose and derived value charge off percent , we see that small business and renewable energy are more likely to charge off .



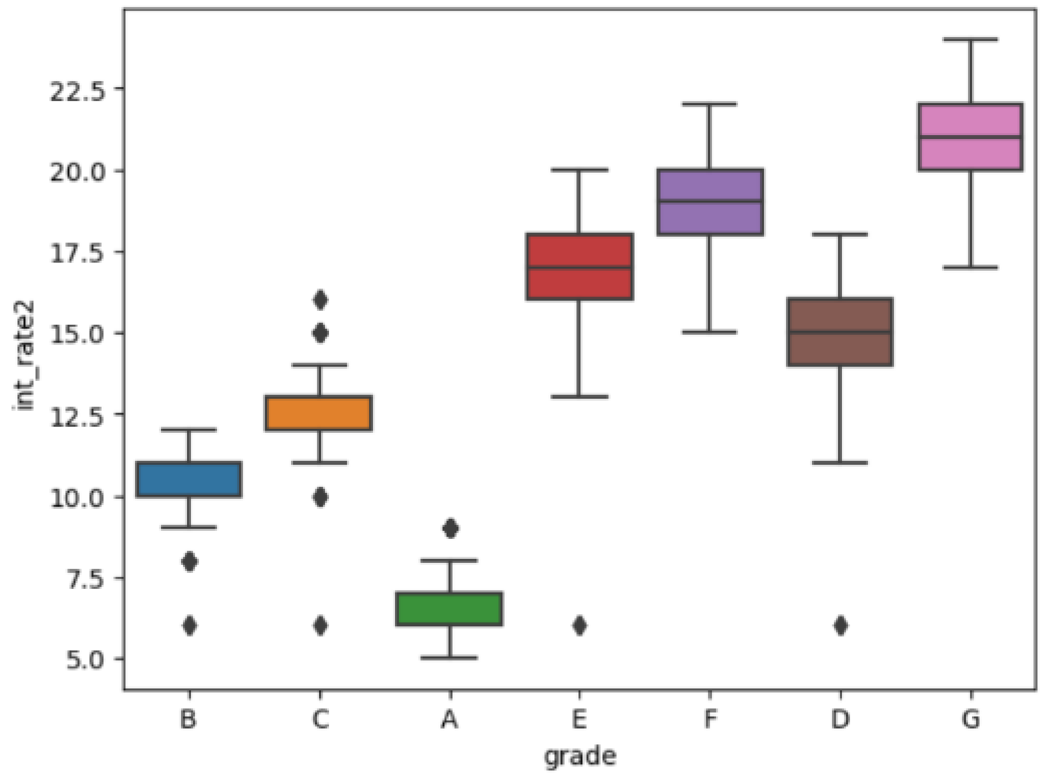
Inference : DTI is negatively correlated with annual inc or loan amount and recoveries is also very weakly correlated with most vars not giving much insight



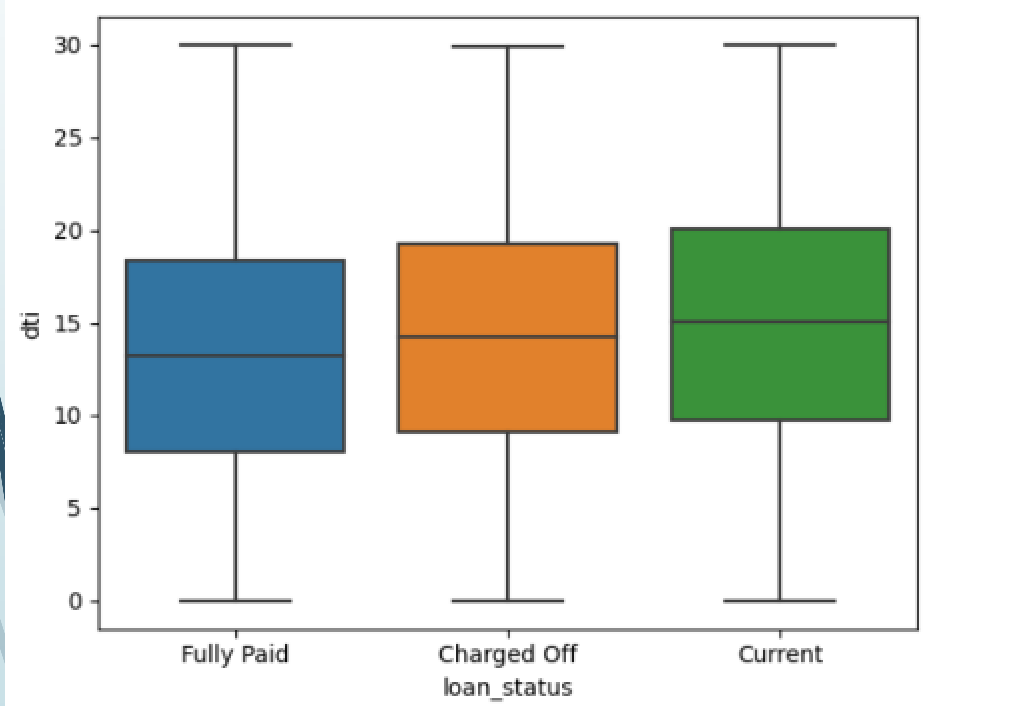
Inference : people having charge off are having median value of 10k



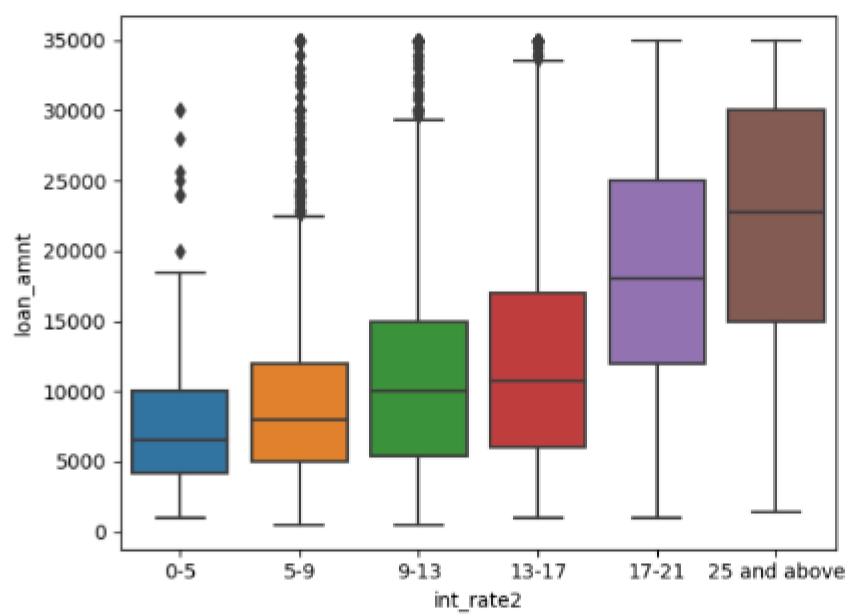
Inference : Grade F , G , E have high interest rates as compared to others



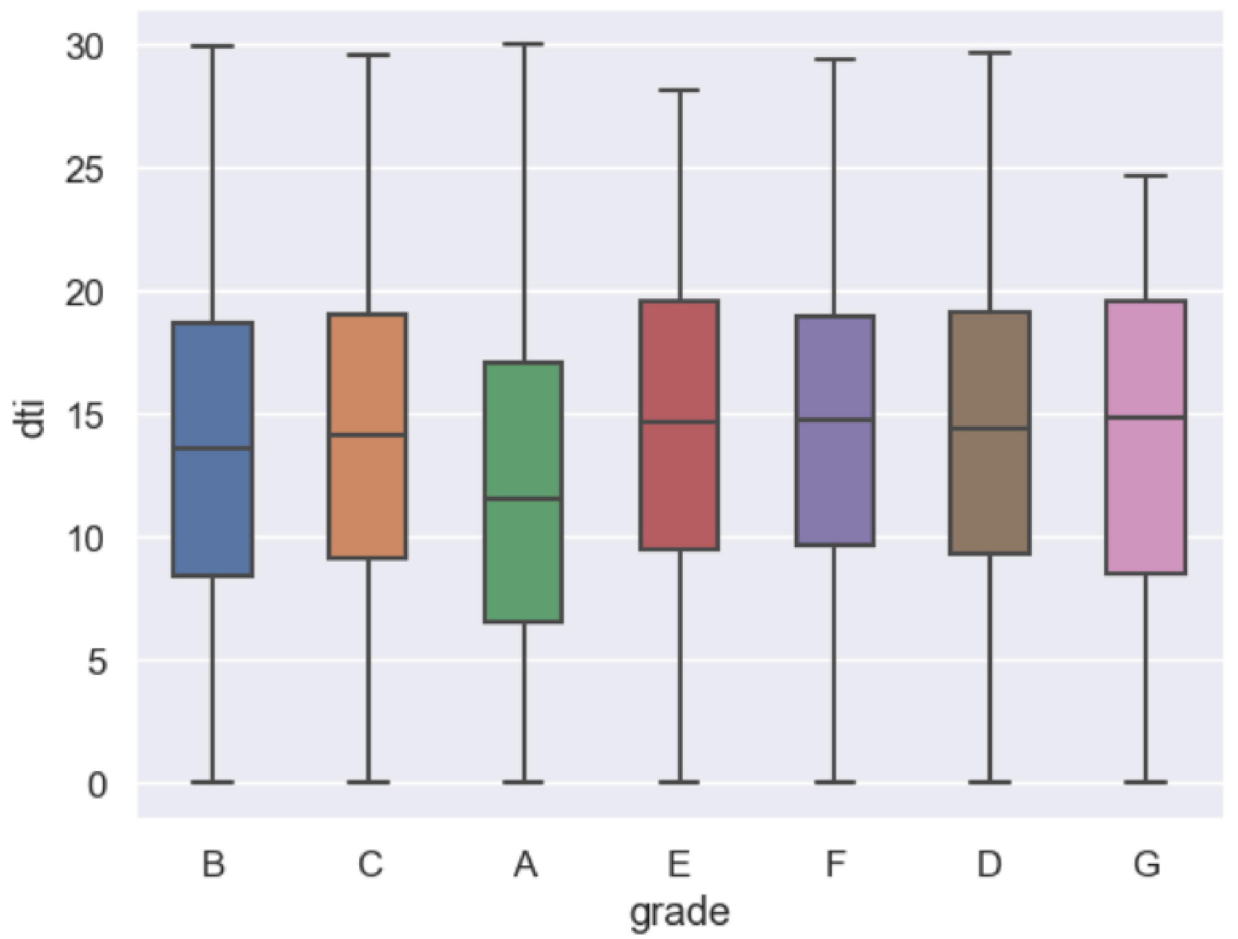
Inference: DTI v/s loan status gives that charge off have higher median than fully paid.



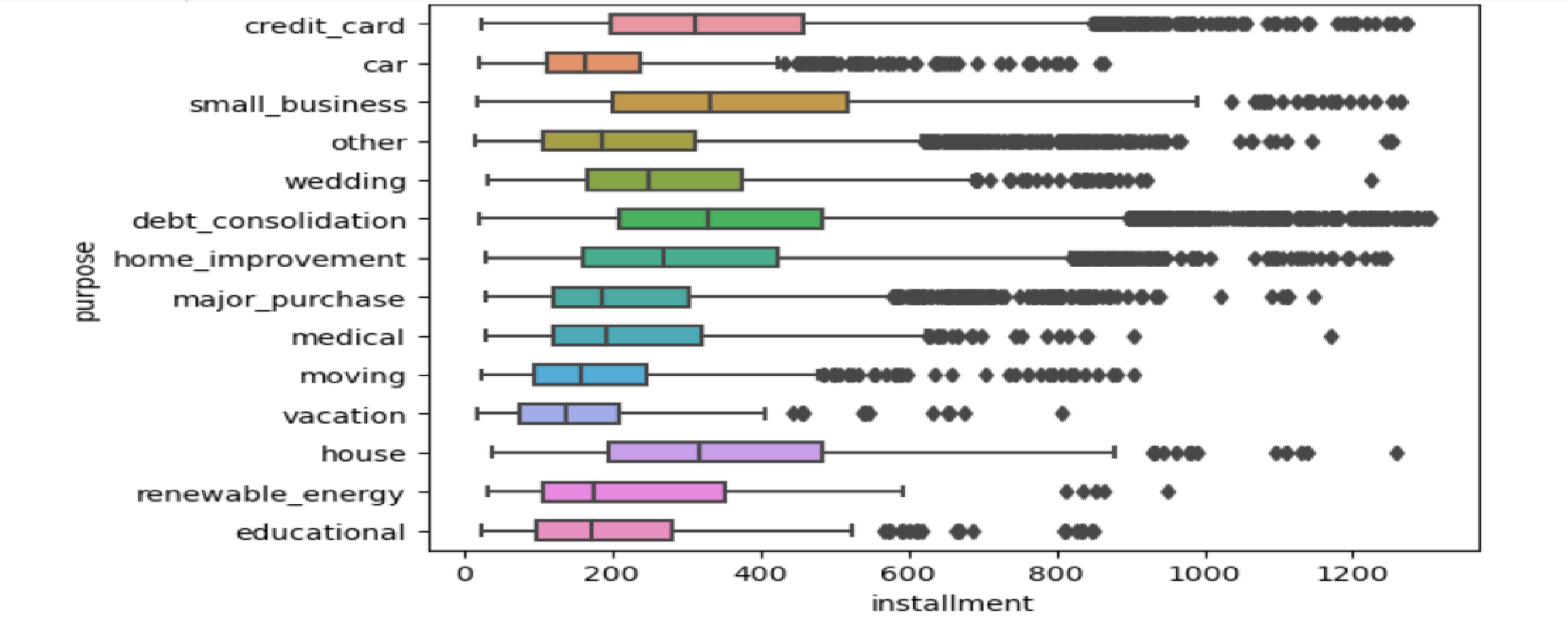
Inference for Interetest rate vs loan amount gives that higher median of loan amount for higher interest rates



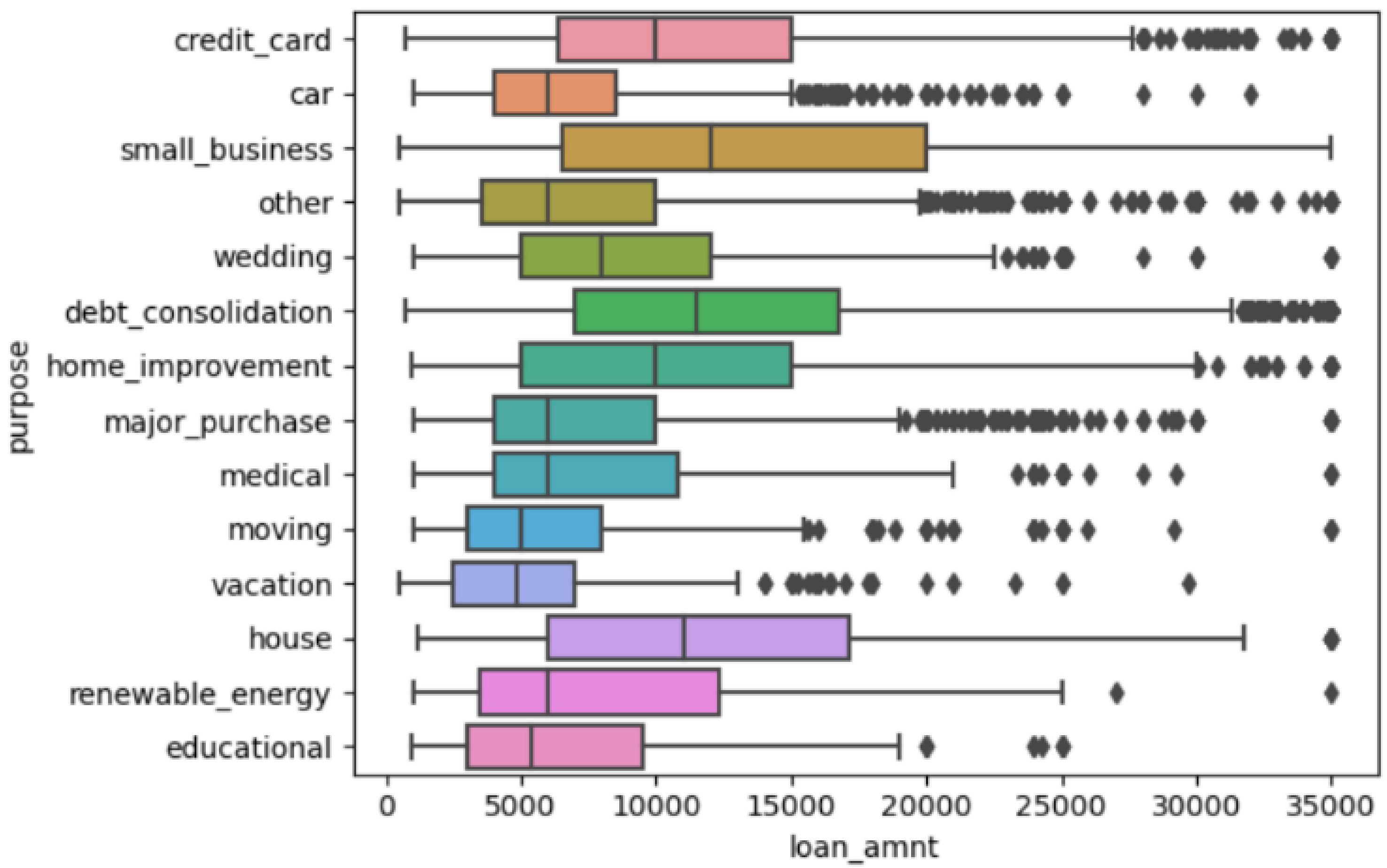
Inference : E,F,G have higher Debt to income ratio



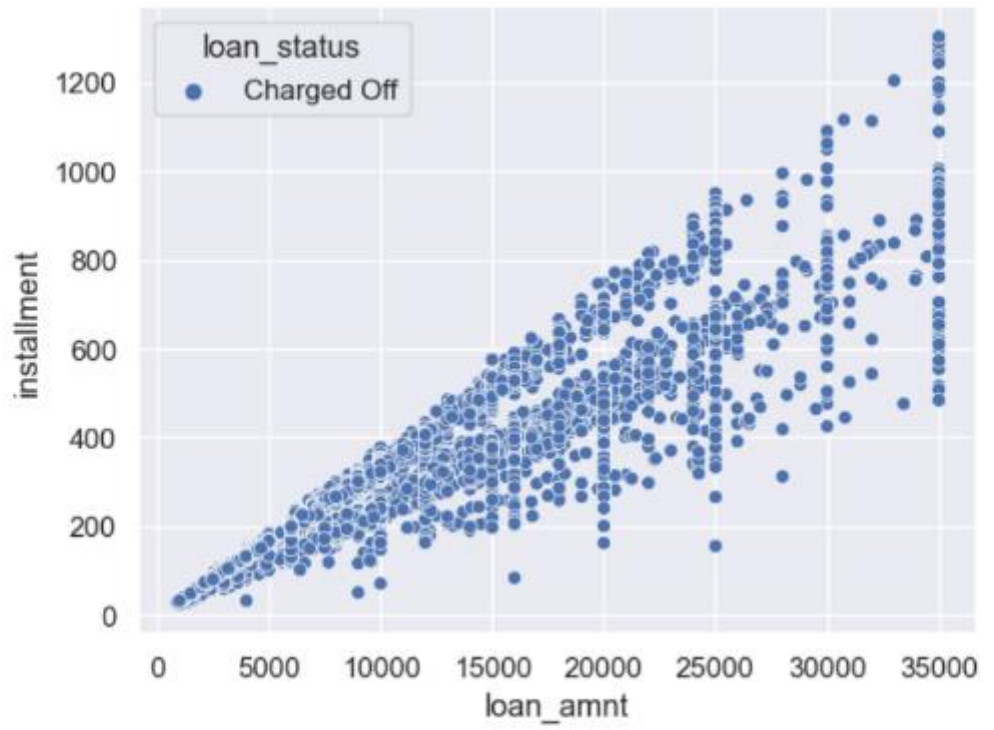
Inference: Installments for credit card , debt consolidation , small business and house are more with regards to others.



Inference : Customers are wanting loan more for the purpose of small business , debt_con ,house,house improvement

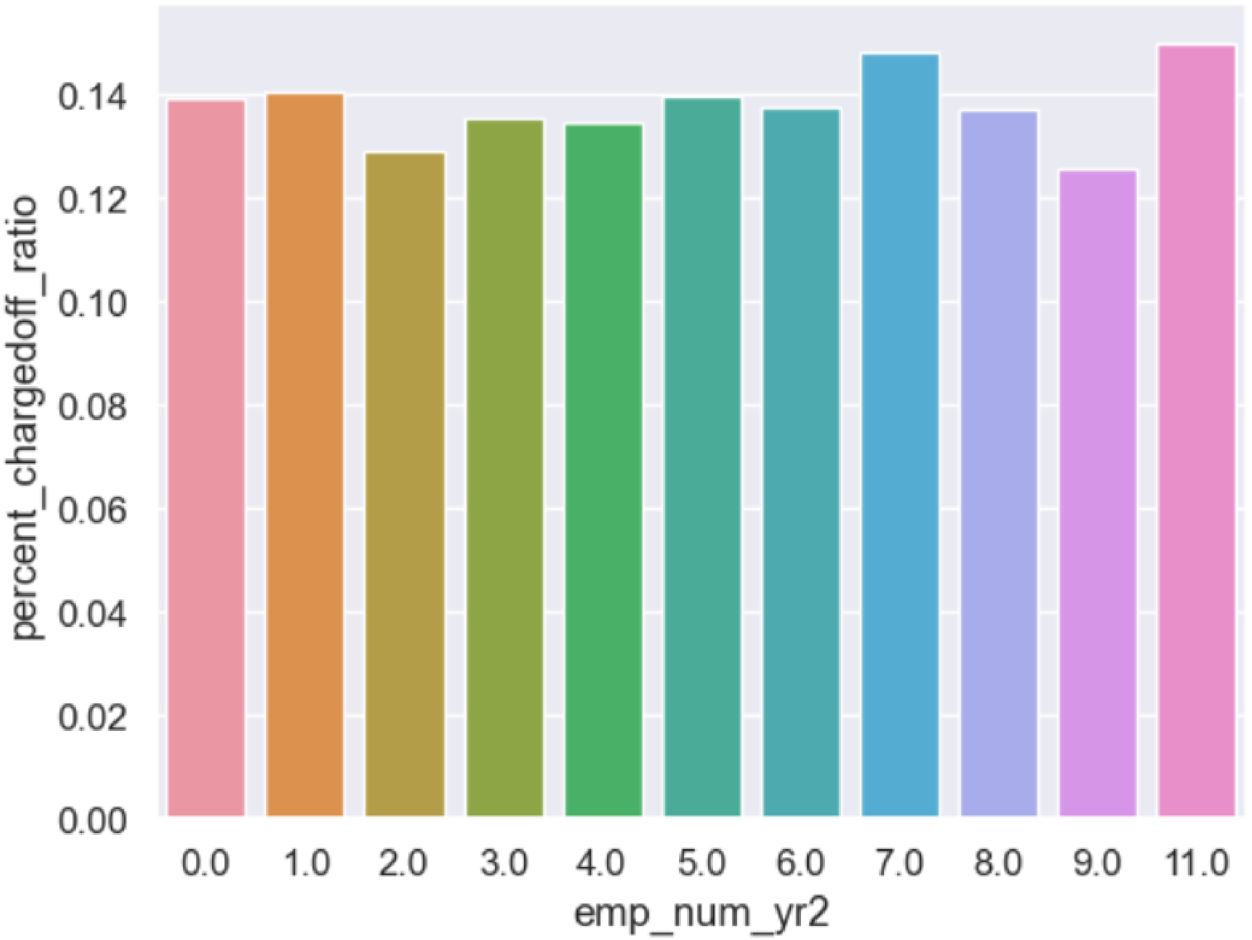


Number of charge off are higher on bigger loan amount and their installment are also high

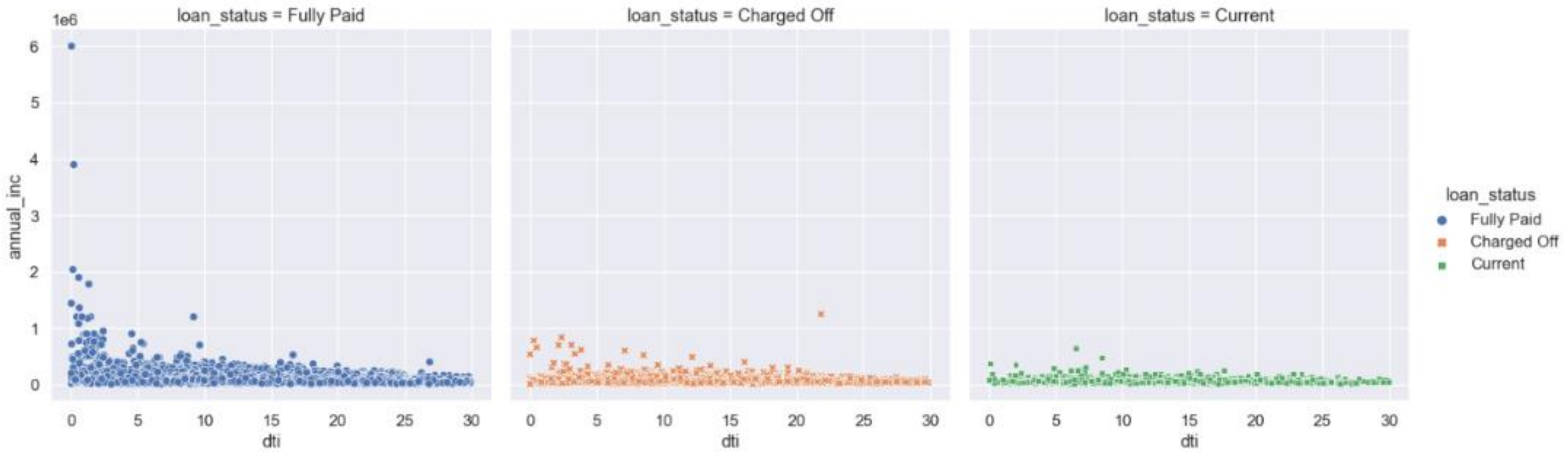


proportion of fully paid are higher than charge off

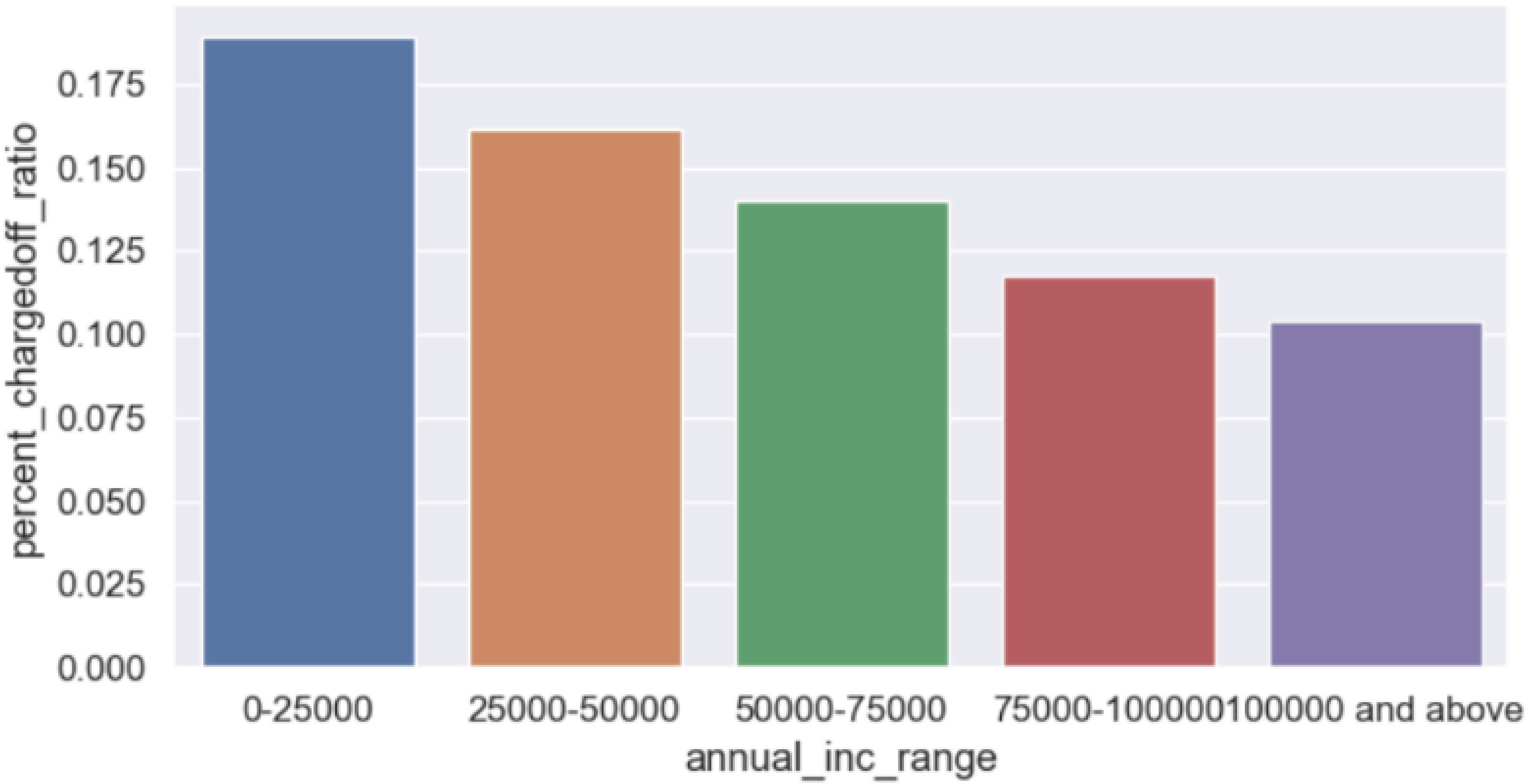
Inference : There are high percentage of people who are un able to pay if their employment experienvce is less especially less than 1 year



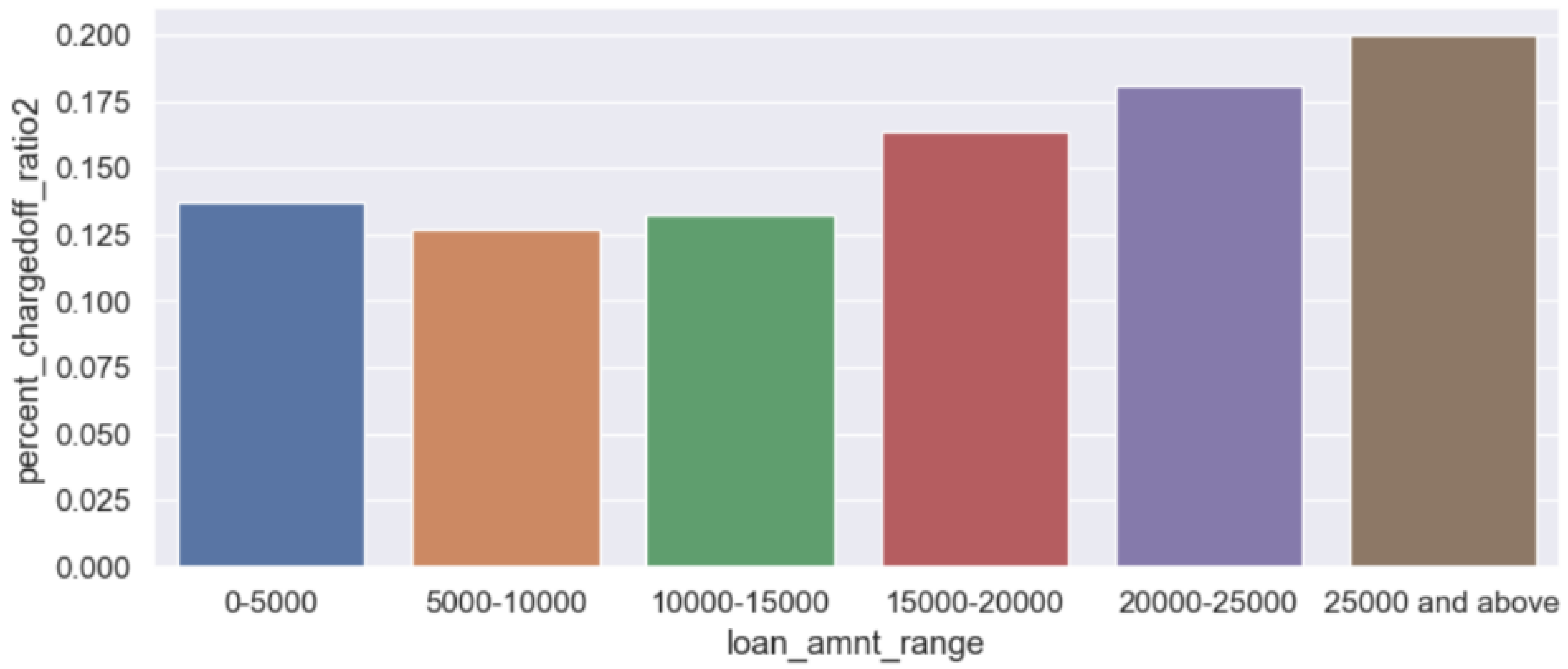
in fully paid their dti is mostly less than 10 while for charge off their dti is mostly > 10 and further dense for 20



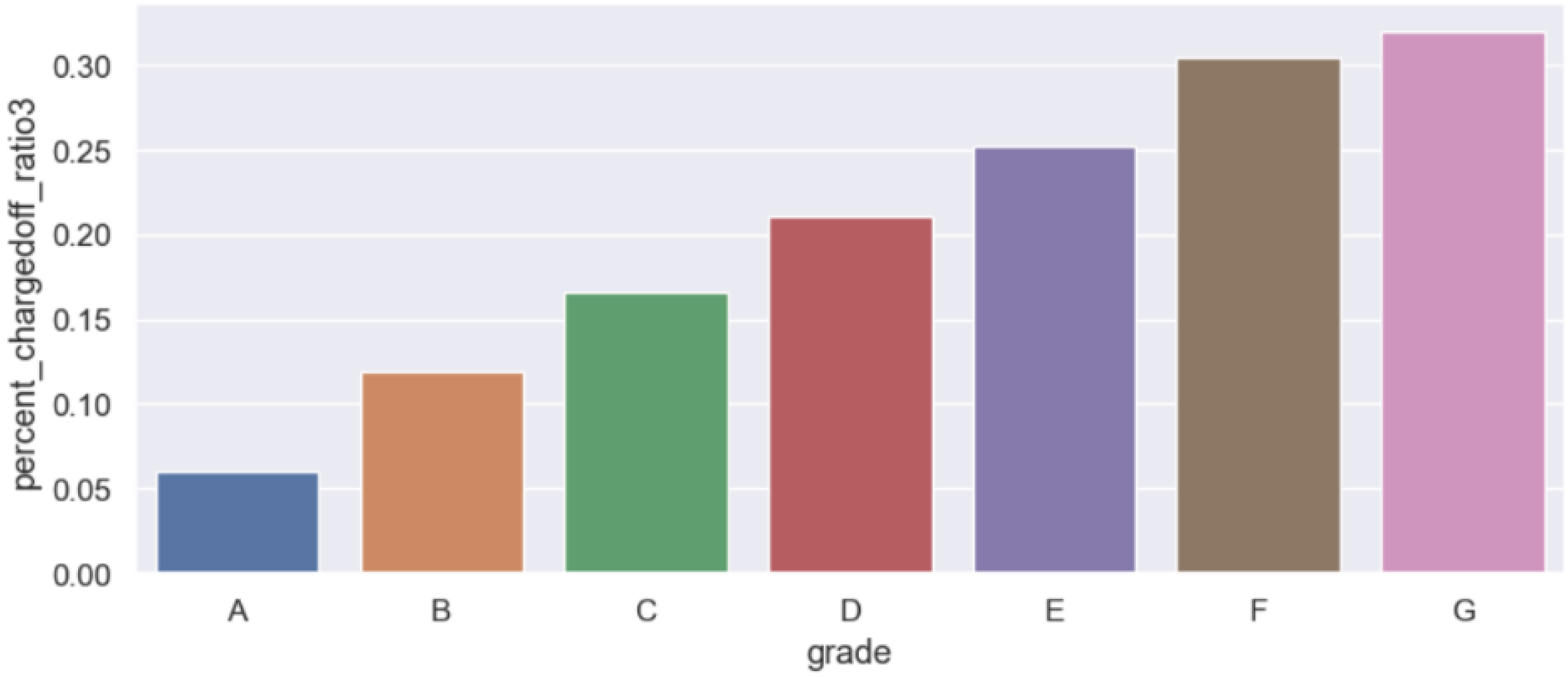
percentage charge off decreases with increase in income



Infernce : percentage charge off increase with increase in loan amount



Inference : Charge off is likely to happen for grade E,F,G grade



Summary of EDA

- Candidates with high loan amount are more likely to charge off
- Grades E,F,G have higher tendency to charge off
- Grade F , G , E have high interest rates as compared to others
- Grade E,F,G have higher Debt to income ratio
- Candidates with high income are less likely to default
- Candidates with high interest rates are likely to default
- Debt to Income ratio (DTI) Is better for Fully paid candidates
- Candidates having less Employee Length(exp) are more likely to not pay the loan
- Charge Off are higher for high installments
- Customers who are on RENT or who are non verified are more likely to default and hence High risk customers
- Candidates are wanting loan more Purpose of home, small business and they take it for less term
- Installments for credit card , debt consolidation , small_business and house are more with regards to others
- Candidates who have taken loan for purpose of small business are more likely to default along with Debt consolidation
- Most percentage of defaulters is in range of less than 10k with rent or mortgages
- Maximum charge off are from California that means better checks in CA state and FL or NY
- DTI is negatively correlated with annual inc or loan amount and recoveries is also very weakly correlated with most vars not giving much insight
- Interest rate vs loan amount gives that higher median of loan amount for higher interest rates