Lending Club Case Study

Group members:

Tanmay Goel

Byjuraj N J

Contents

I. Business Objectives

II. Steps for EDA data:

III. Data understanding for EDA

IV. Data Cleaning and Manipulation

V. Analysis

VI. Conclusion

I. Business objectives:

a. Business Understanding:

Lending Club (LC) is a consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company must make a decision for loan approval based on the applicant's profile.

Two types of risks are associated with the bank's decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- 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

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

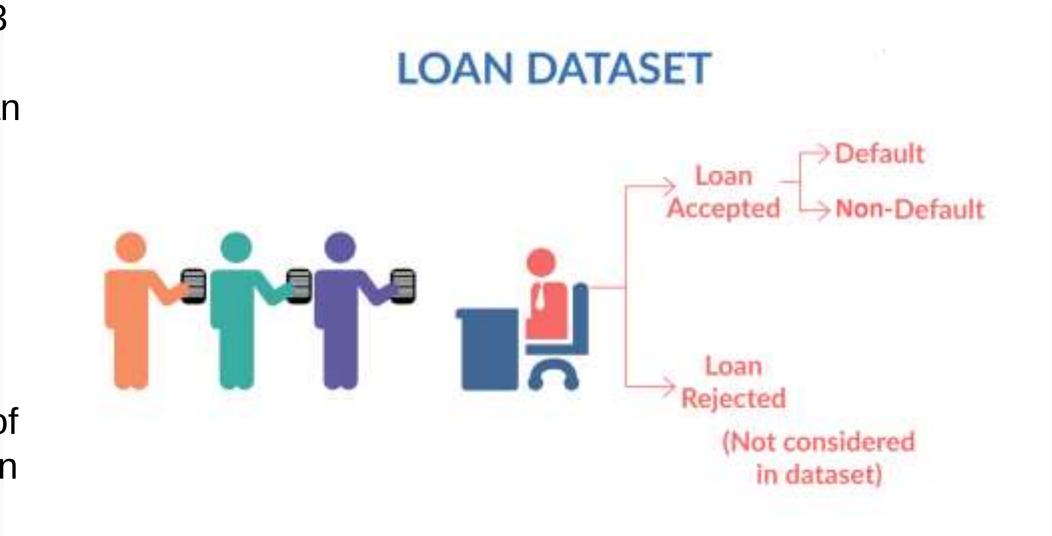
- Loan accepted: company approves the loan
- Loan rejected: company rejects the loan

I. Business objectives:

a. Business Understanding:

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

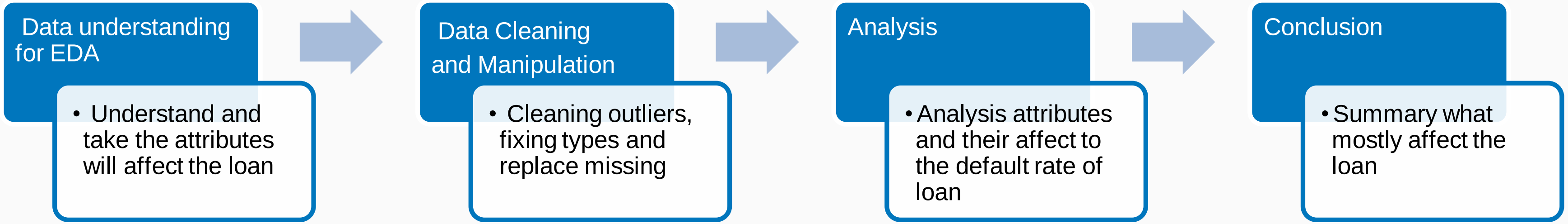


I. Business objectives:

b. The objectives:

Borrowers who **default** cause the largest amount of loss to the lenders
If one can 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.
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 utilize this knowledge for its portfolio and risk assessment.

II. Steps for EDA data:



Data understanding for EDA

- Understand and take the attributes will affect the loan

Data Cleaning and Manipulation

- Cleaning outliers, fixing types and replace missing

Analysis

- Analysis attributes and their affect to the default rate of loan

Conclusion

- Summary what mostly affect the loan

III. Data understanding for EDA:

- Data Understanding refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies
- We begin understanding the data by getting the shape(number of rows and columns) in the data and the datatype of each column
- We then displayed a few lines in the dataset to understand the general composition of the data

III. Data understanding for EDA:

The data have 111 attributes. However, 56 attributes don't have any recorded information, 9 attributes have the same values with each sample. The remain 57 attributes can cluster into 4 group:

- Group 1: 24 attributes are related to customer behavior and not available at the time of loan application or the unique value for customer.
- We remove this group
- Group 2: Consumer attributes when they apply the loan
- Group 3: The loan attributes
- Group 4: Loan status, to determine if a person is default or not. This is the key or our EDA.

III. Data understanding for EDA:

Consumer attributes

Annual income
Employment Length
Employment Title
Home ownership
Past bankruptcies
Earliest credit line open
Address state
Zip code
DTI (loan payment/ income)

Loan attributes

Verification status
Inquiry last 6 months
Issue date
Loan amount
Funded amount total
Funded amount by investor
Grade by LC
Sub grade by LC
Installment
Interest rate

Loan status

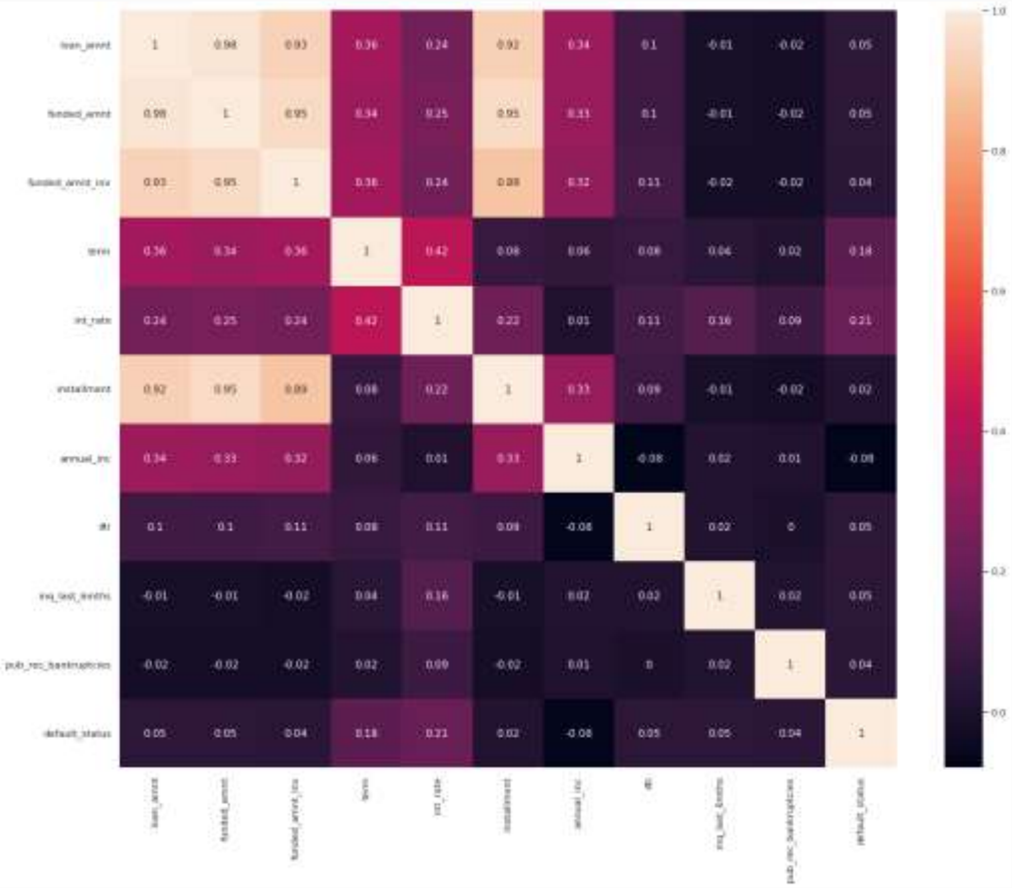
Status of loan

As we see, the number of "Current" in loan status is very small (1k/39k samples). As "Current", it's hard to say if the borrower will likely to be default or not, so we don't use this
For easier to calculus the default rate, we define that if a person is "Charge off" (default), that mean 1, otherwise 0

IV. Data cleaning and manipulation:

- Data cleaning is **the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset**.
- We began by removing the columns having all null values.
- Post this we eliminated columns having single values as it would not be of any use for analysis
- Then we filled missing values in the quantitative variables by their mode values
- We then standardised the values in the rows. For example: the column int_rate was in the form of 10%, We removed the ‘%’ symbol associated with the int_rate so as to produce numeric values that would be helpful for analysis
- We then performed outlier treatment on the data to eliminate abnormal values
- We then examined the columns individually and recorded following observations
    - "desc" has description (text data) which we cannot do anything about it
    - id", "member_id", "url" doesnot affect the loan
    - Certain column are post approval features- delinq_2yrs','earliest_cr_line','last_credit_pull_d','last_pymnt_amnt', etc
- Hence we removed these columns as they do not influence loans status

V. Data analysis:

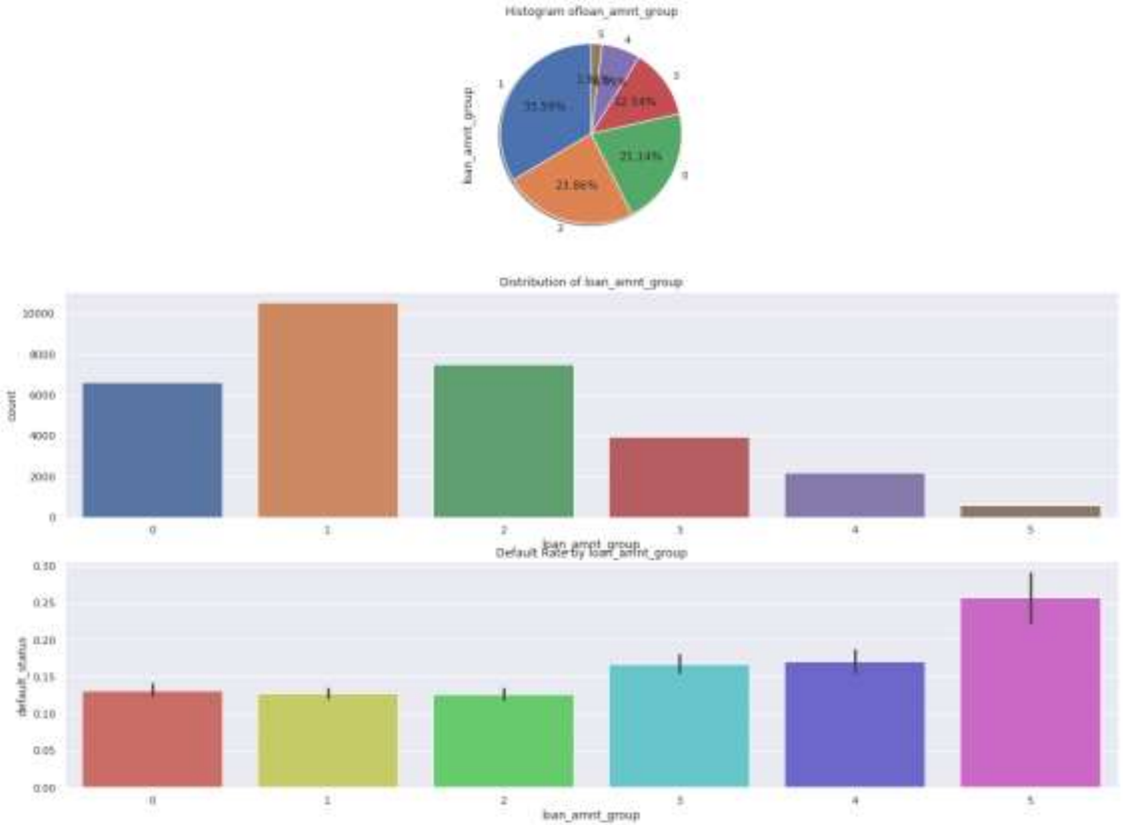


We first look at correlation matrix to reduce the dimension of data to analysis

As we see, 'funded_amnt', 'funded_amnt_inv', 'installment' are highly correlation with "loan_amnt", so we remove them

V. Data analysis:

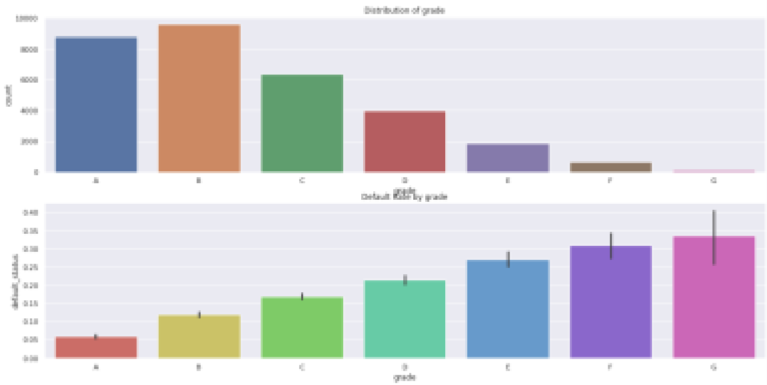
loan_amnt

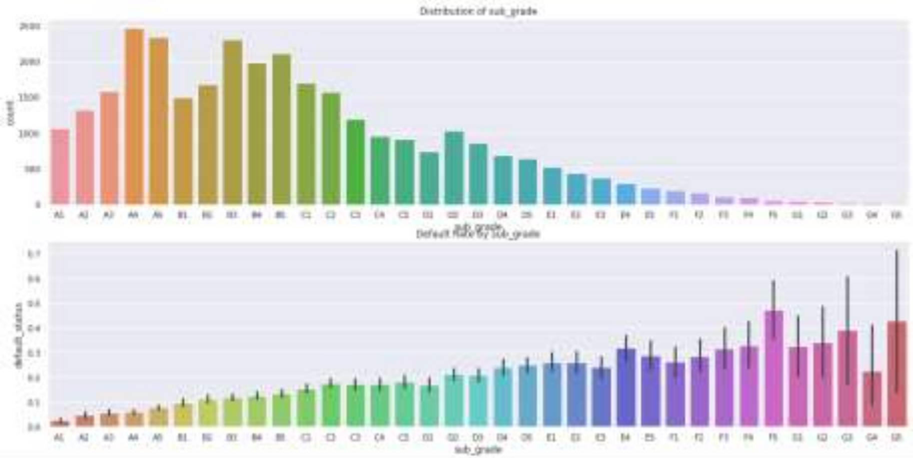


- We derive loan amount to groups; each group have range 5k
- The loan amount mostly in range **5k to 10k**
- The amount of loan affect the default rate, **higher loan amount can get higher rate of default**

V. Data analysis:

grade

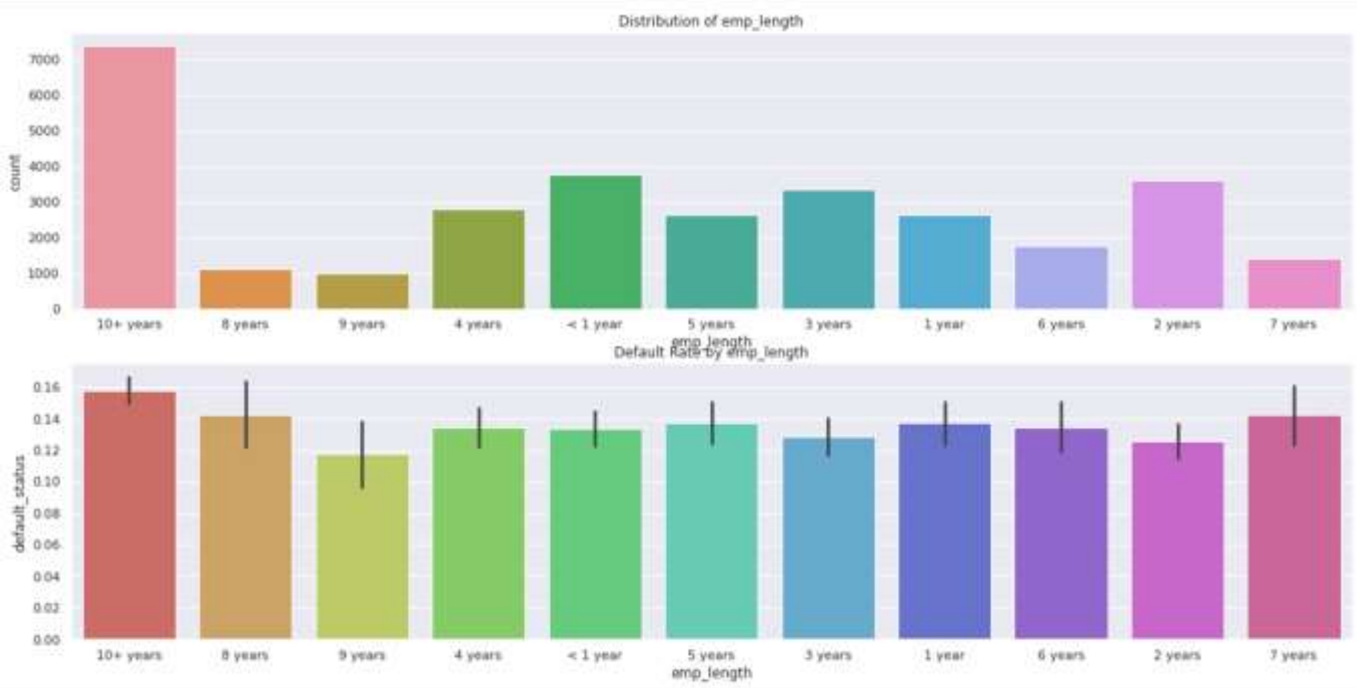




The grading system of LC work well: **higher rank of loan profile, lower rate of default rate**

V. Data analysis:

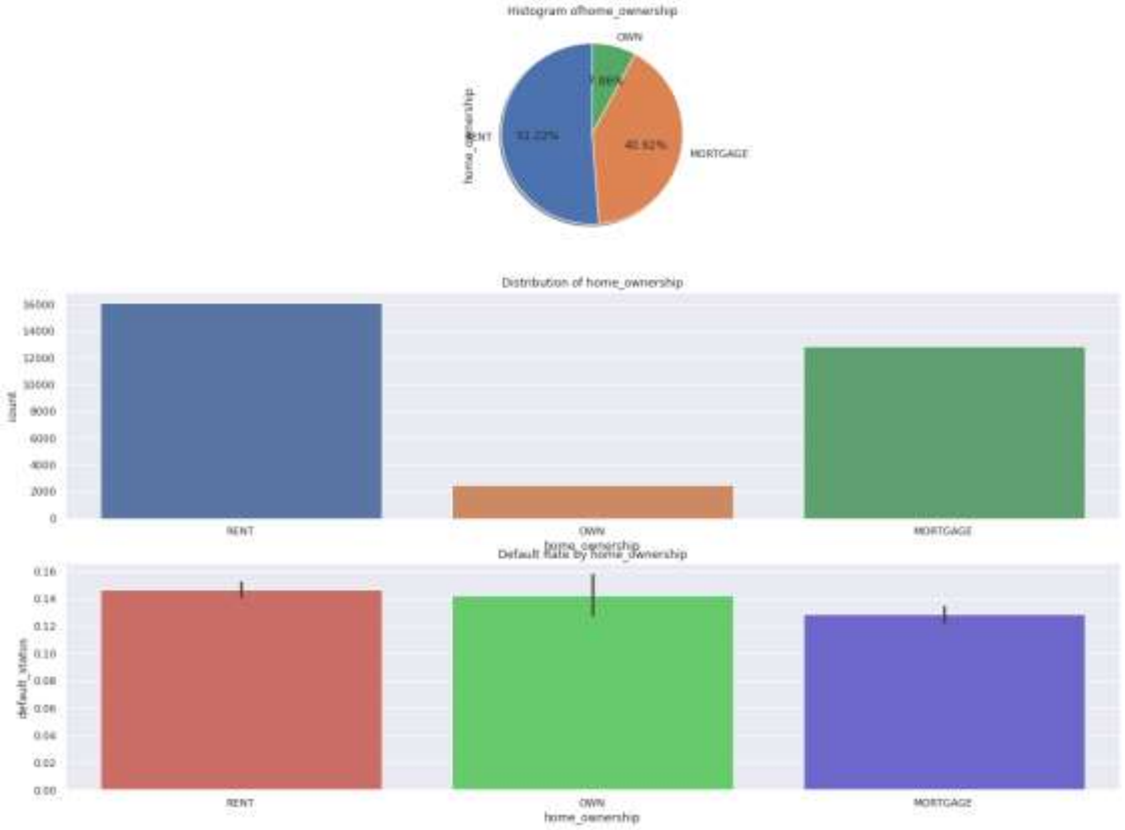
emp_lenght



- '10+ years' mean the average ages will be 30s.
- They has good volume as well has higher default rate

V. Data analysis:

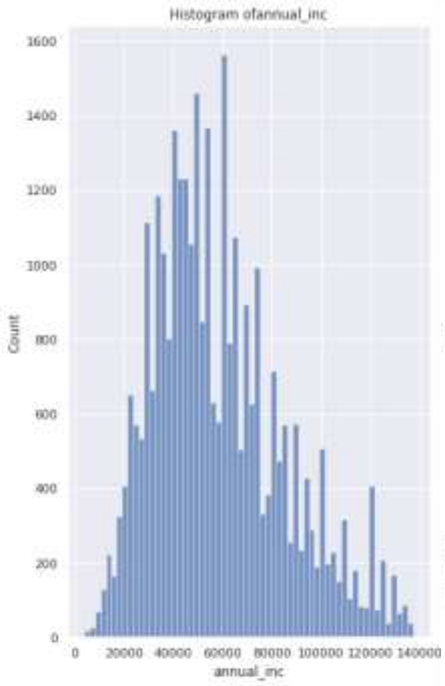
home_ownership

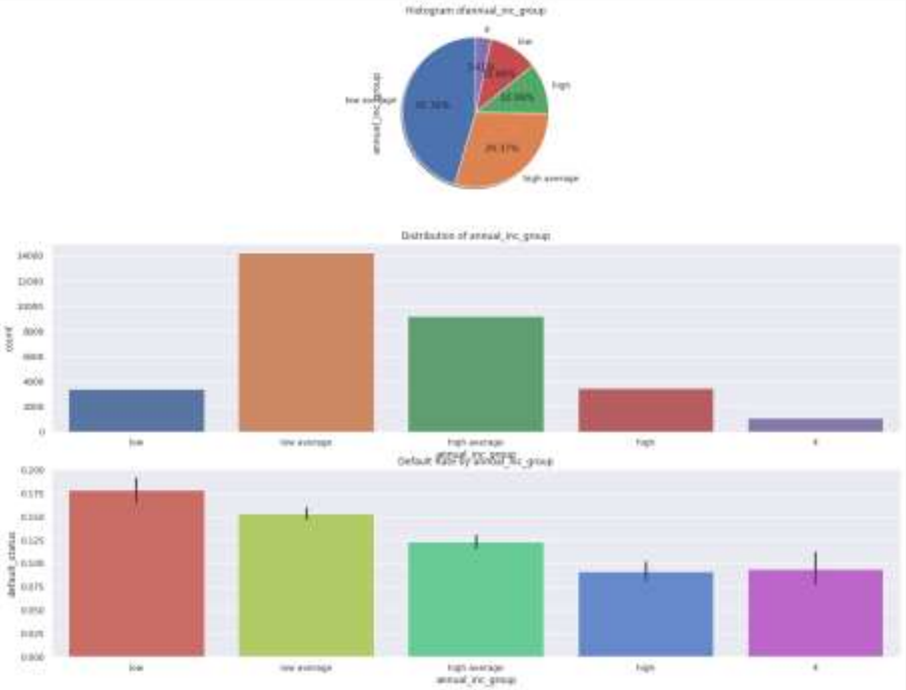


- People who already own home take less loan.
- The home ownership not affect much default or not.
- 'Rent' and 'Own' have similar default rate
- 'Mortgage' group has slightly lower default rate

V. Data analysis:

annual_inc

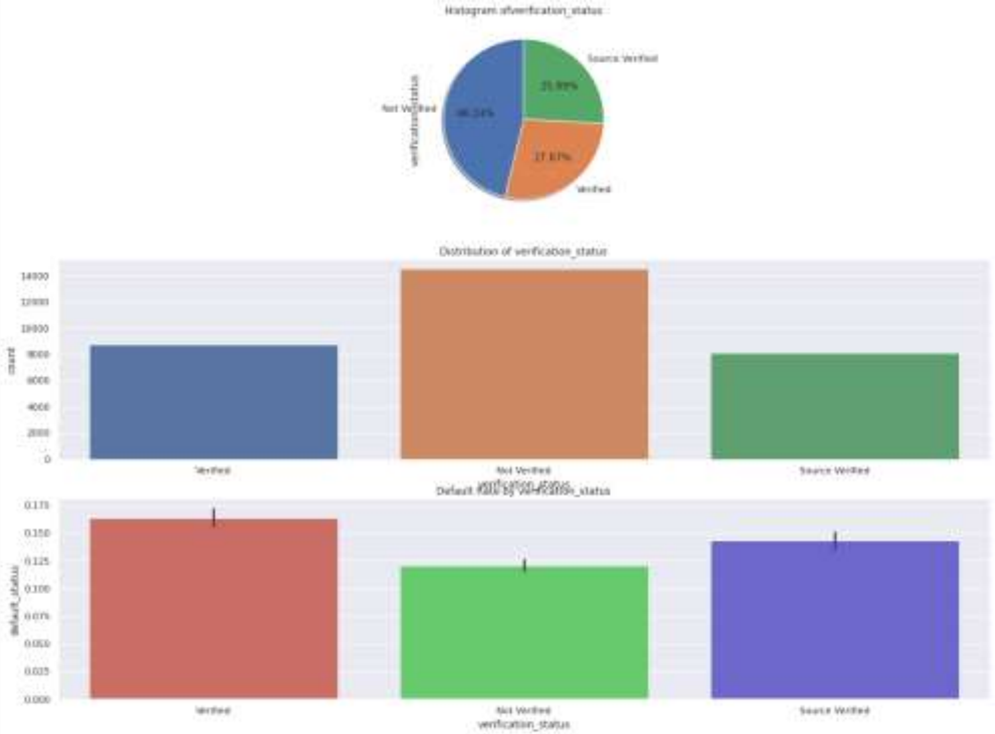




- We derives income to 4 groups: low, low average, high average and high
- Lower incomes have higher default rate
- Most of the loan come from people have low average income(30k-60k)

V. Data analysis:

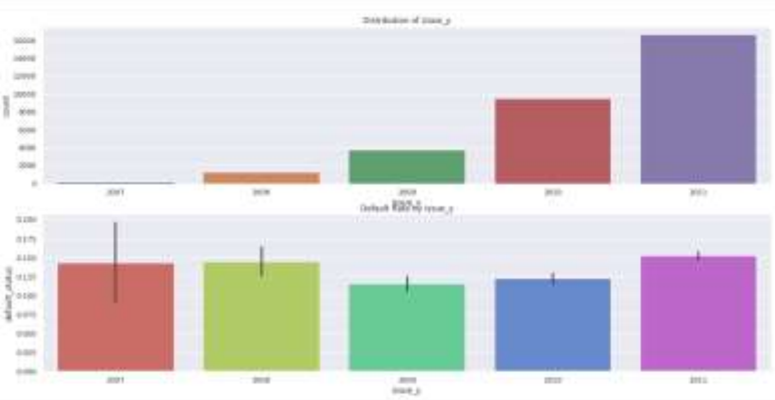
verification_status

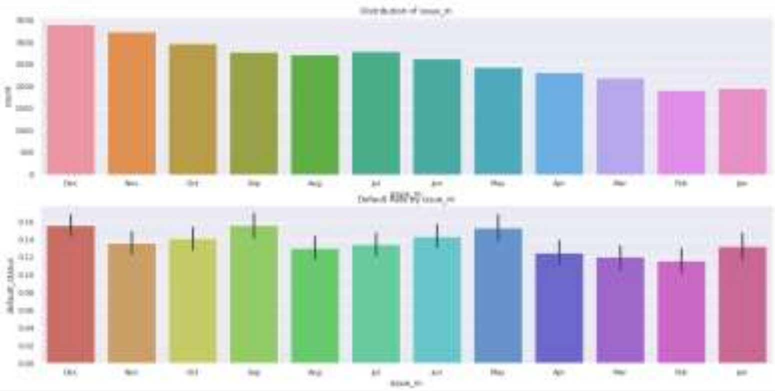


- Strange finding: 'Not verified' has lower default rate than 'Verified', despite it have highest amount of sample.

V. Data analysis:

issue_d

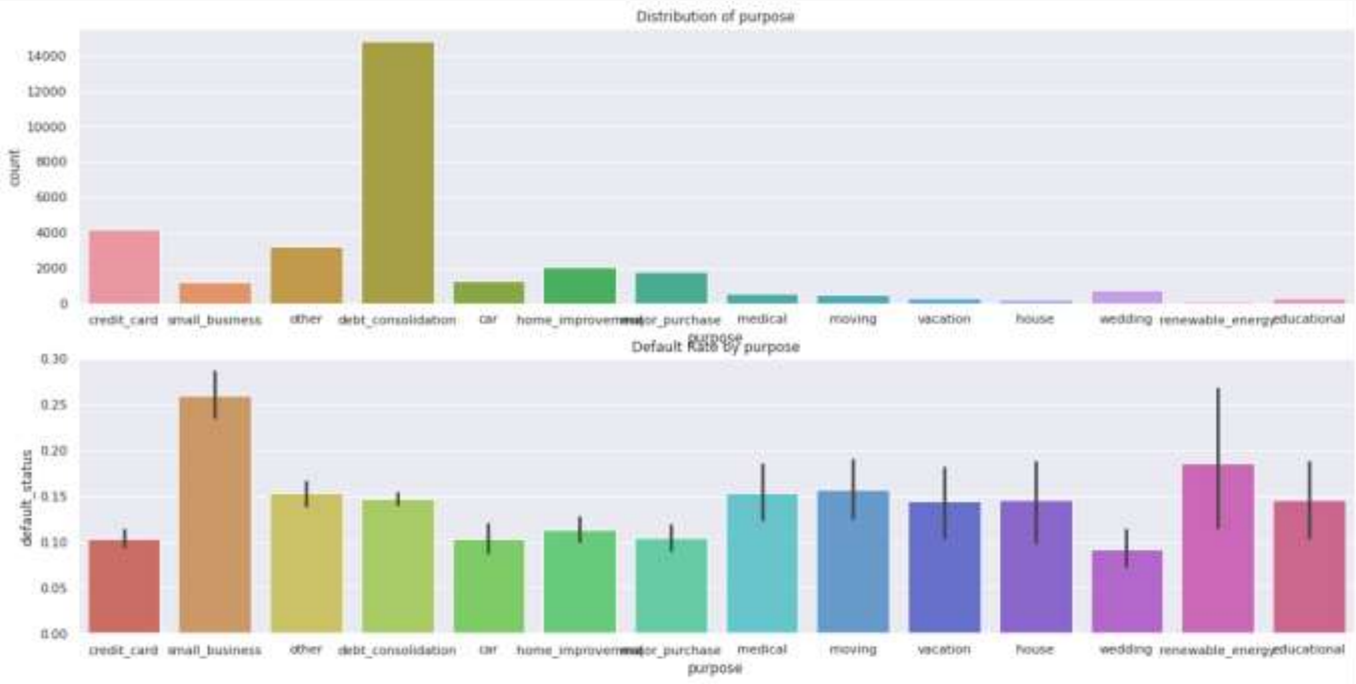




- We derive issue date to year and month
- The number of loans issued go up from 2007 to 2011 and the default rate went down during these year until 2011, where we see spike in default rate.
- 2010 has lower default rate People more likely to have loan in end years

V. Data analysis:

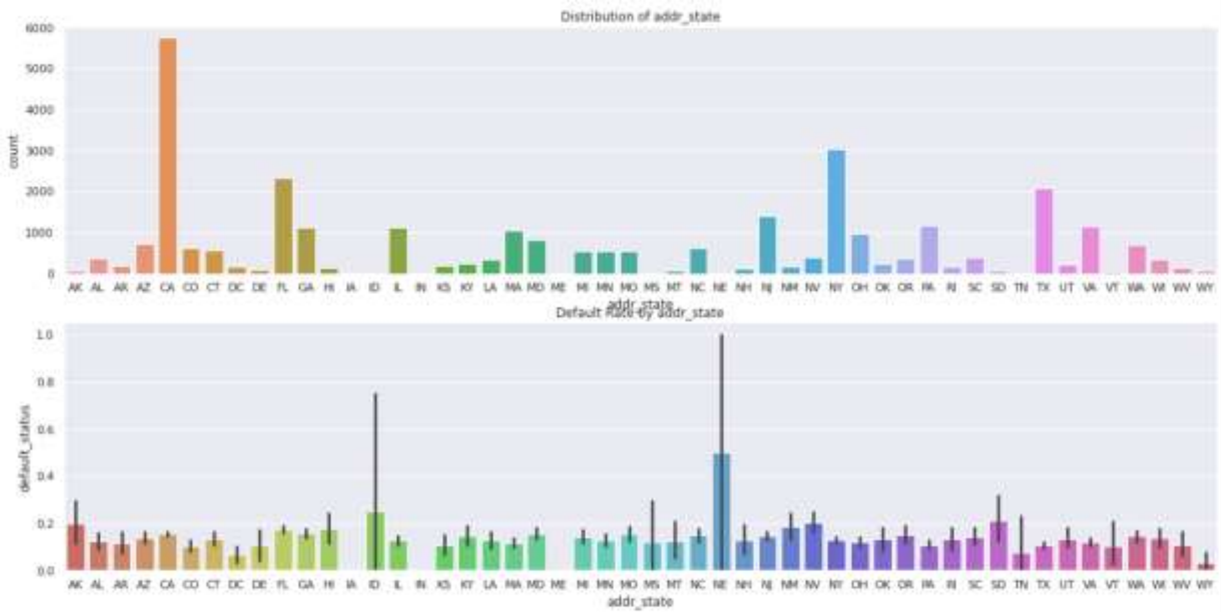
purpose



- **debt_consolidation** has the largest count on aprroved loan.
- **small_business** loans have the highest default rates.
- **wedding** has the lowest default rate.

V. Data analysis:

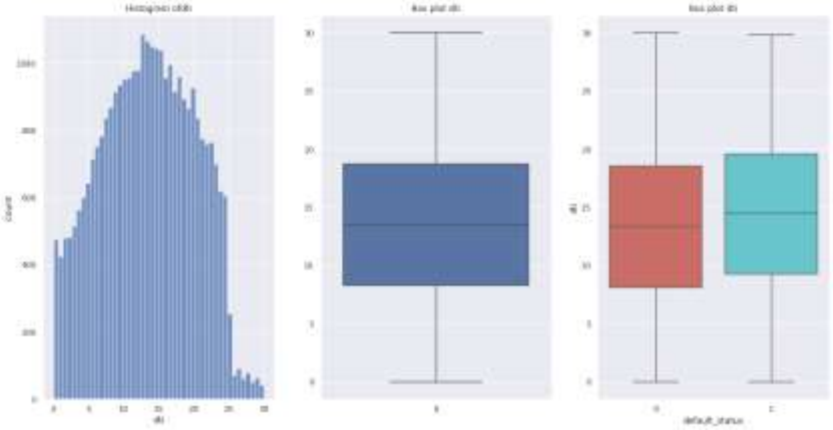
addr_state

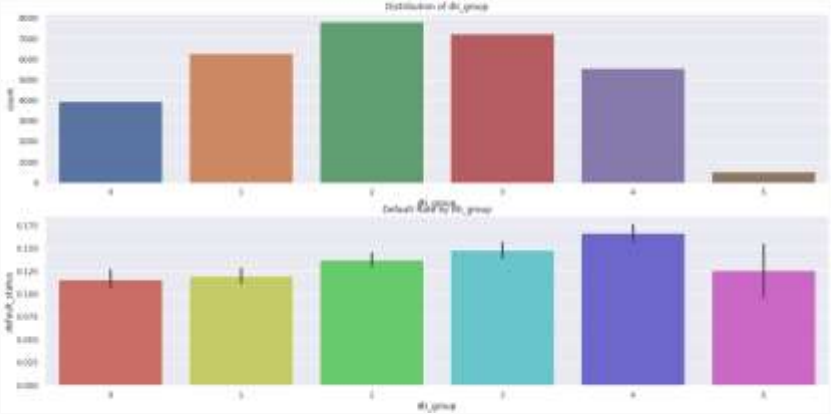


- **CA** have highest loan amount, but very low default rate.
- In other side, **NE, SD** have very few loan amount, but the default rate is very high.

V. Data analysis:

dti

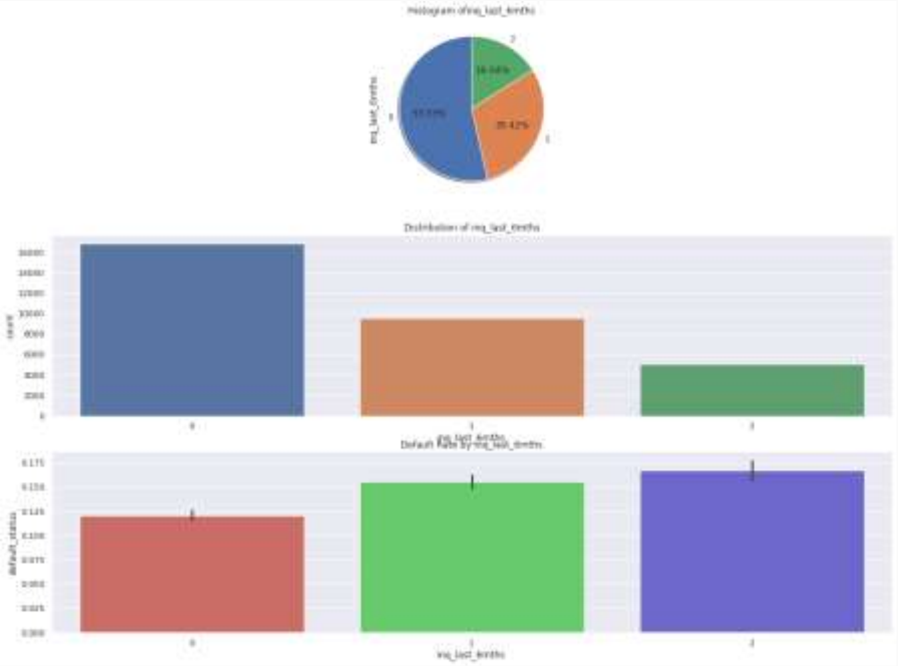




- We derive DTI to range 5% each group
- Most people have their loan amount from 5-20 percent DTI
- Higher the DTI, higher the default rate
  It's risky to offer loan to customers with high DTI.

V. Data analysis:

inq_last_6mths

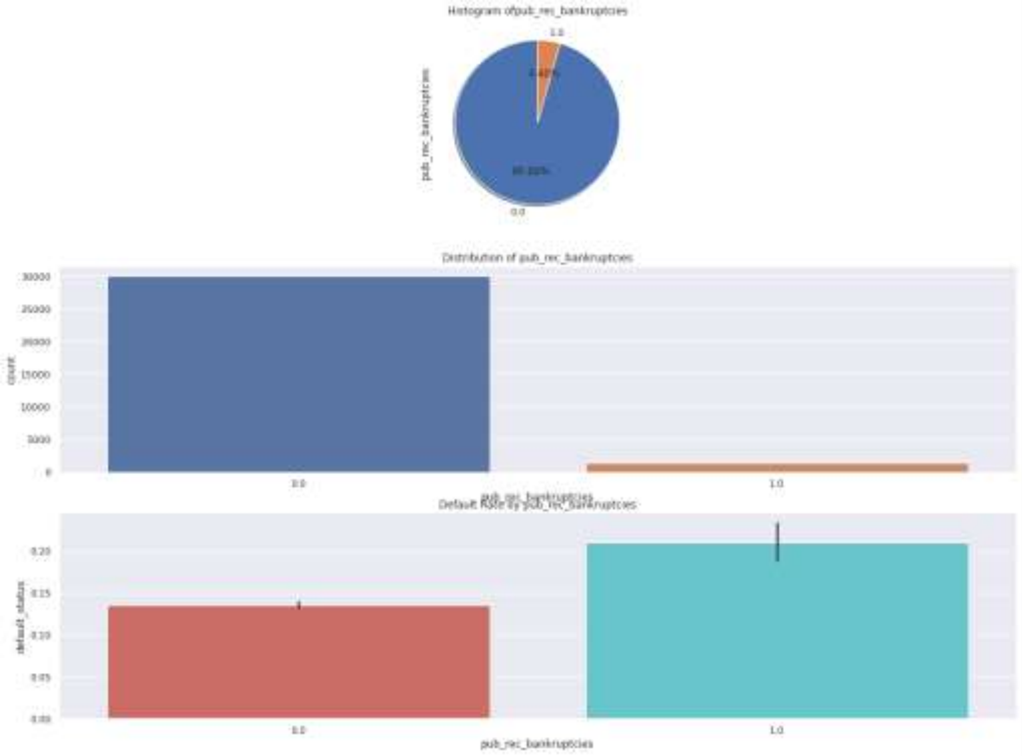


- The people get loan mostly have no inquiry in 6 month, and they have lowest rate of default.

=> This mean the inquiry process of LC perform not good enough

V. Data analysis:

pub_rec_bankruptcies



Higher the bankruptcies, the higher is the default rate

VI. Conclusion:

- Person who don't own house likely to have the loan
- The loan affect by interest rate, higher interest mean higher default rate
- The loan mostly come from the people have annual income in range 30k to 60k.
- When get a loan, there must be look at the amount of loan over their annual income, cause DTI show that if the ratio get higher, the default rate get higher too
- A strange information found from data is the verification status which is not verified get lowest default rate
- Mostly people have loan without any inquiry in 6 month, that mean the LC system need more attention about who could be customer
- Some states like CA are a good market, some like NE are bad
- The loan market is going up on the need of market, cause the number of loan go up every year, and people more likely to have loans in end-year months.