Lending Club Case Study

…

MALLIKARJUNAPPA S SANGANNAVAR
SRIRAMSUBRAMANIYAN NADARAJAN

Problem statement and analysis approach

- Problem statement :

- There is a loan data set from Consumer Finance Company. Need to analyse the customers who defaulted the loan. Come up with consumer and loan attributes influencing the loan defaults. These can be used in deciding the loan for news customer based on these recommendations.

- Risks associated with Bank
  - If loan is rejected for applicant who is likely to repay the loan, results in business loss to the financing company
  - If loan is accepted for an applicant who is not likely repay the loan, may result in loan default hence financial loss to the company.

Problem statement and analysis approach

- Analysis Approach :
  - Understand the data present in data set
  - Check quick view of missing values and decide if any columns or rows can be dropped
  - Analyse missing values and impute missing values with appropriate values
  - Analyse if any constant values in columns or rows. Drop the columns with same values
  - Check outlier in the data and address the outlier
  - Start univariate, bivariate variable analysis to come with factors influencing the loan default.

Problem statement and analysis approach

- Based on analysis output, two types of decision are possible
  - Loan Accepted : In applicant meets all requirements to provide the loan, then loan application will be accepted
  - Loan Rejected : Based on factors influencing default, finance company can reject the loan application

Understand the data and Analysing the missing values

- Original data set size is : Shape of data frame : (39717, 111)
- Printed the number of missing values per column. Based on the output saw many columns were all values are missing. Hence decided to drop columns with more than 90% missing values
- Based on further analysis observed few columns with all zeros and same values like ‘f’ and ‘n’. So we are dropping these columns
- Now again print missing values per columns. There are some columns with very few missing values, hence decided to drop rows where columns have few missing values

Understand the data and Analysing the missing values

- Converted few columns like int_rate, term and zipcode which are of object values to numeric and string type
- Columns like grade, sub_grade, emp_length are also converted from object to valid types
- Split the columns into categorical and continuous for better analysis

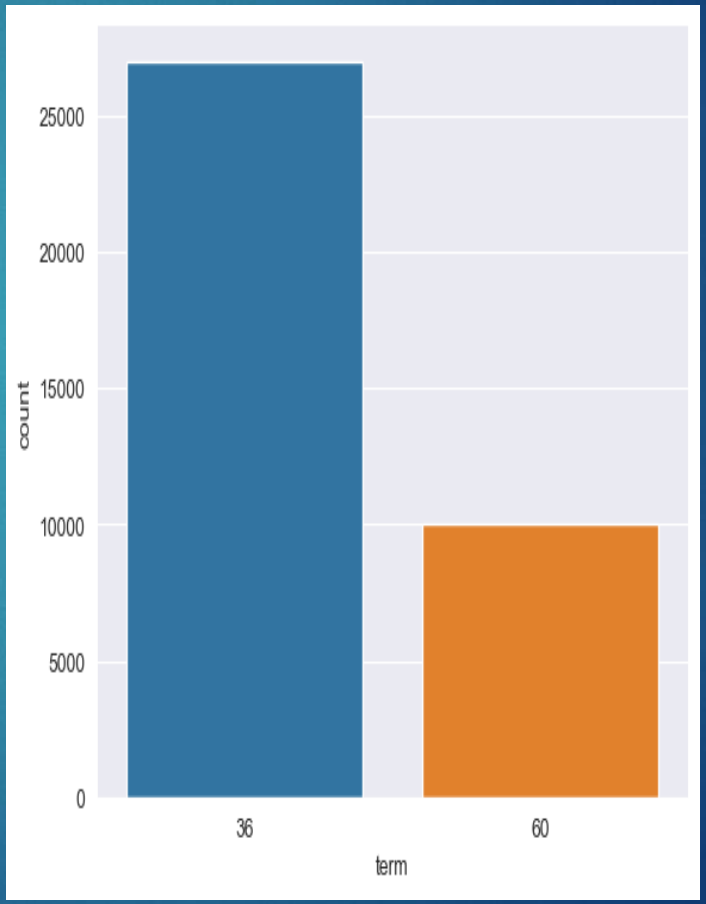
Data set after cleanup of data

- Shape of data frame after data clean up : (37049, 44)
- Shape of fully paid : (30933, 44)
- Shape of charged_off : (5050, 44)
- Number loan defaulters : 5050
- % loan defaulters : 13.630597317066586

Univariate analysis – Term

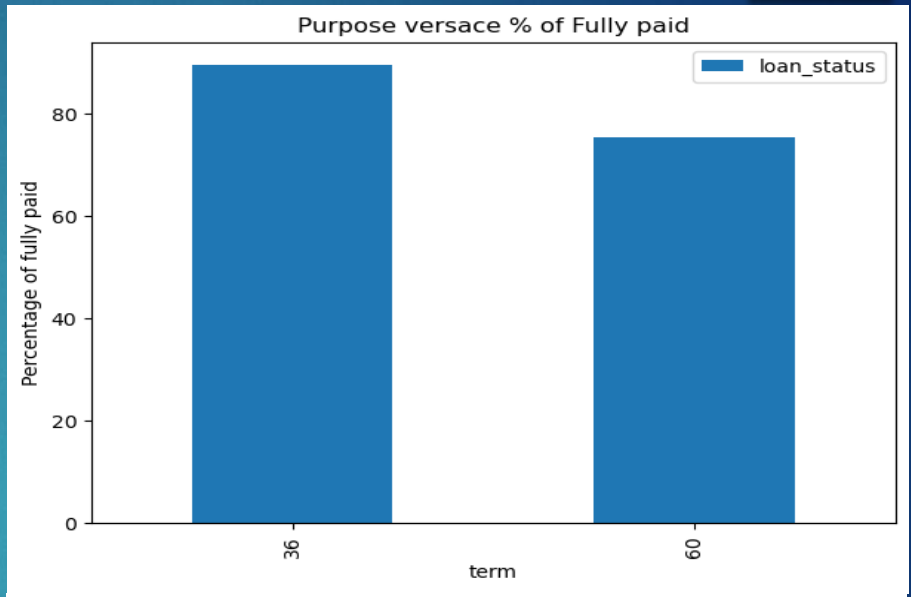
Term

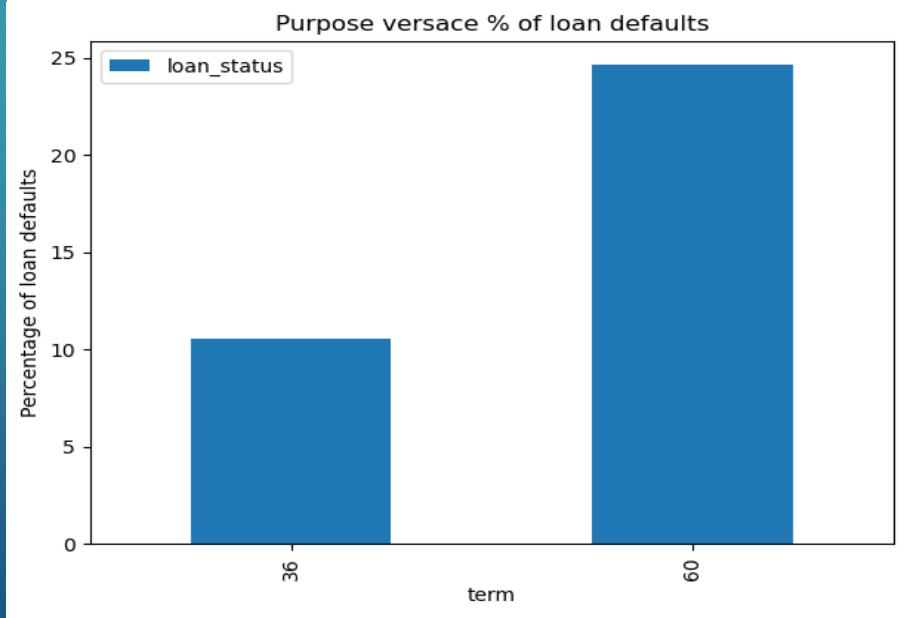
- Loans has been disbursed for 36, 60 months
- More than 25000 loans are disbursed for 36 months and around 10000 loans are given for 60 months
- We can infer that most of the people had opted for loan which has shorter repayment term



Bivariate analysis – Term /Loan Fully Paid, Charged off

- We can infer that Loans given at 36 months has more fully paid customers and less defaulters than the Loans given with 60 months term

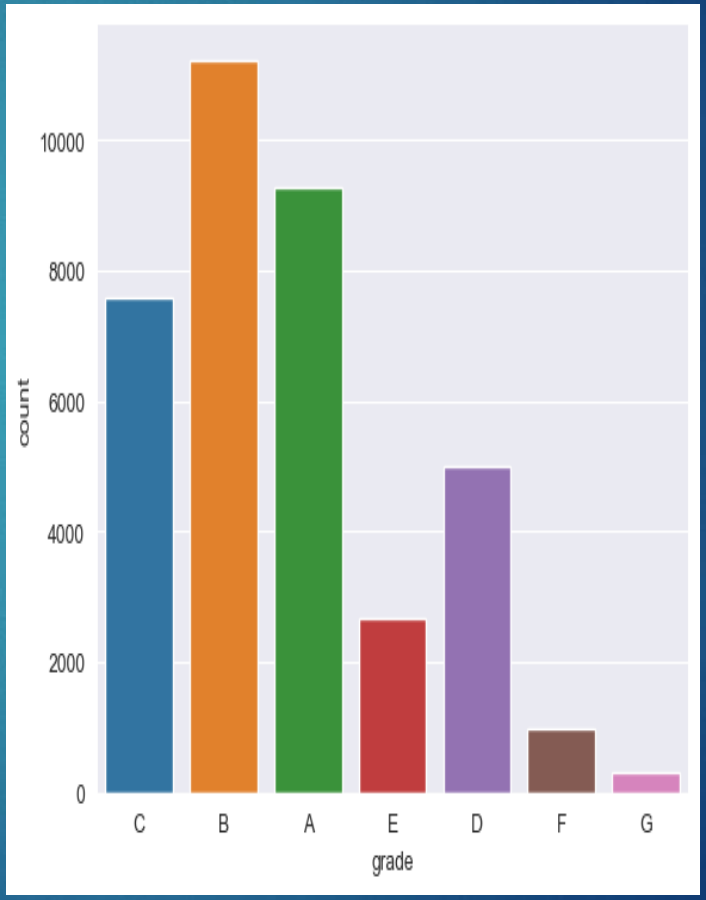




Univariate analysis - Grade

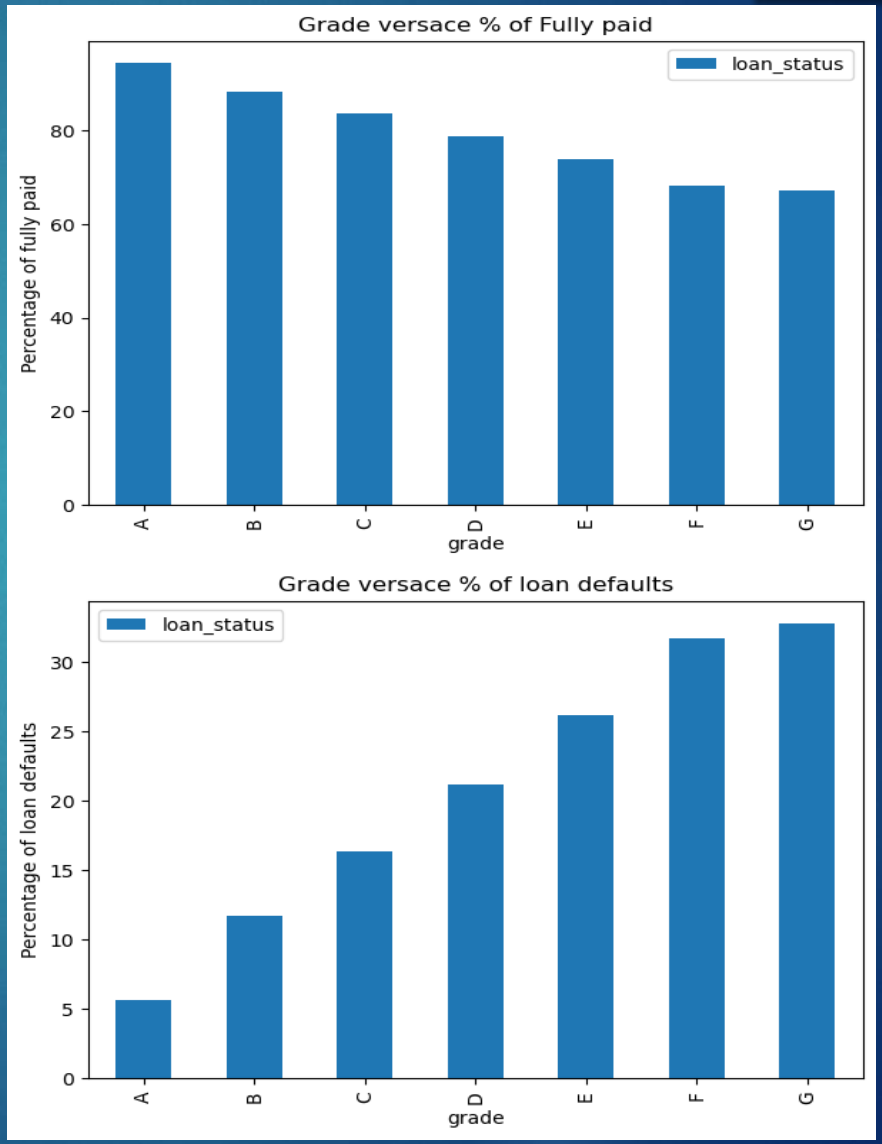
Grade

- We can infer that loans are given more to employees of Grade B than all other Grades
- Employees who are of Grade G are given very less loans than all others



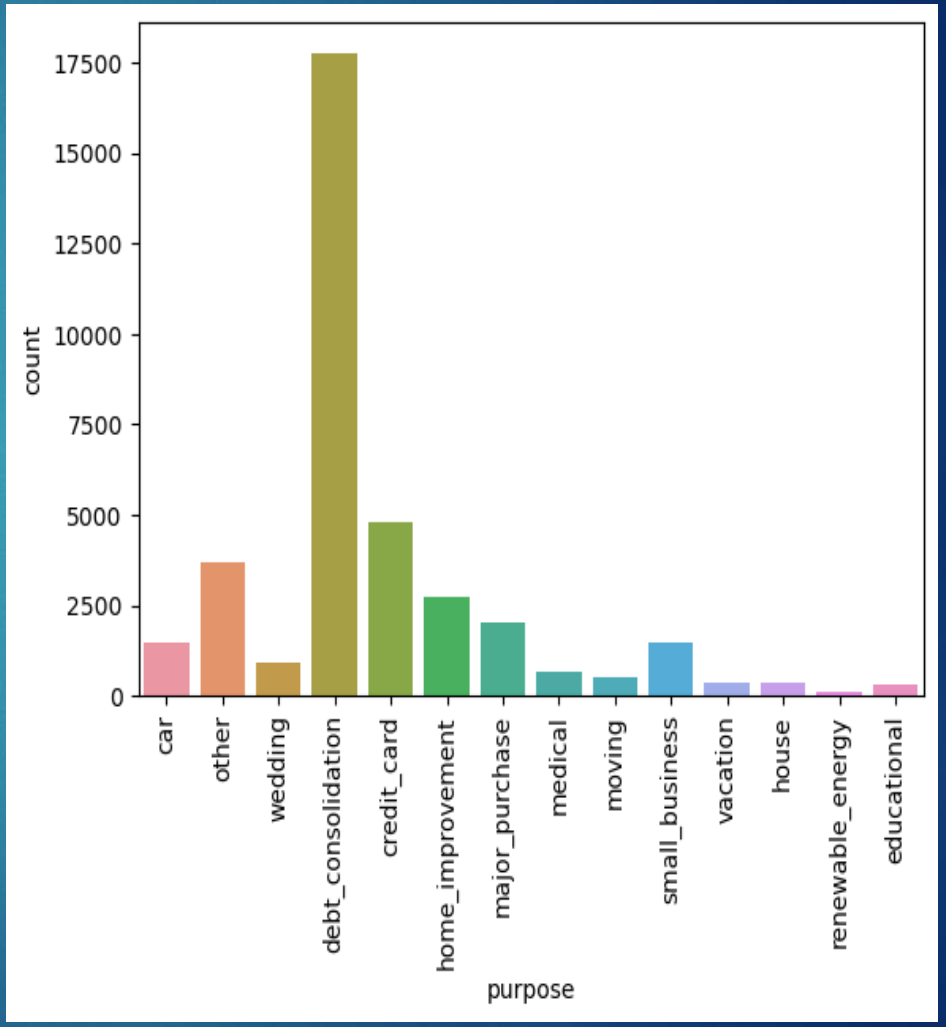
Bivariate analysis – Grade/ Loan Fully Paid, Charged off

- Grade Vs Loans
- We can infer that loans availed by customers who are in Grade A are recovered to full extent.
- While customers who are in Grade G has least loan repayment percentage
- We need to formulate strict disubursal process for customers who are in Grade G, F



Univariate analysis - Purpose

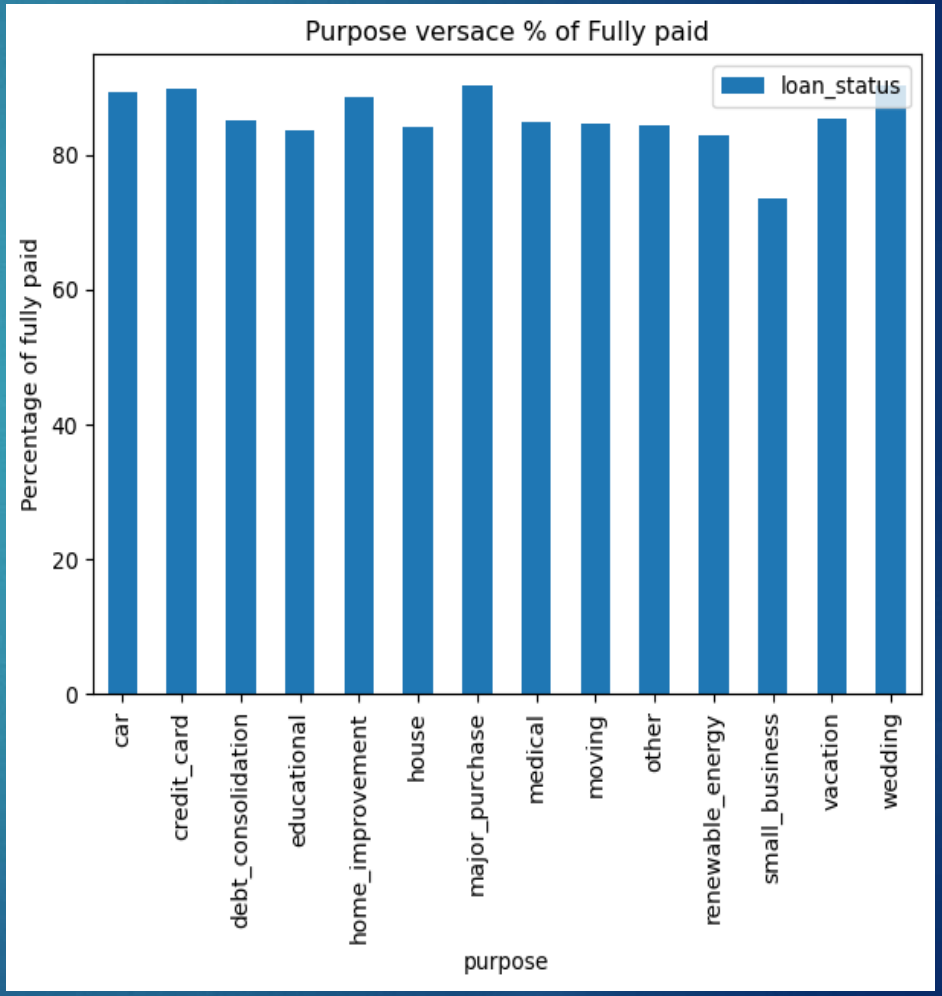
- Purpose:
- We can infer that more loans are given for “Debt-Consolidation” purpose than any other category
- Followed by that “credit_card” and other purpose are the category to which loans are given
- “Renewable_energy” ,educational, housing and vacation are the purpose for which least loans are given



Bivariate analysis – Purpose/ Loan Fully Paid, Charged off

Purpose vs Fully Paid loans

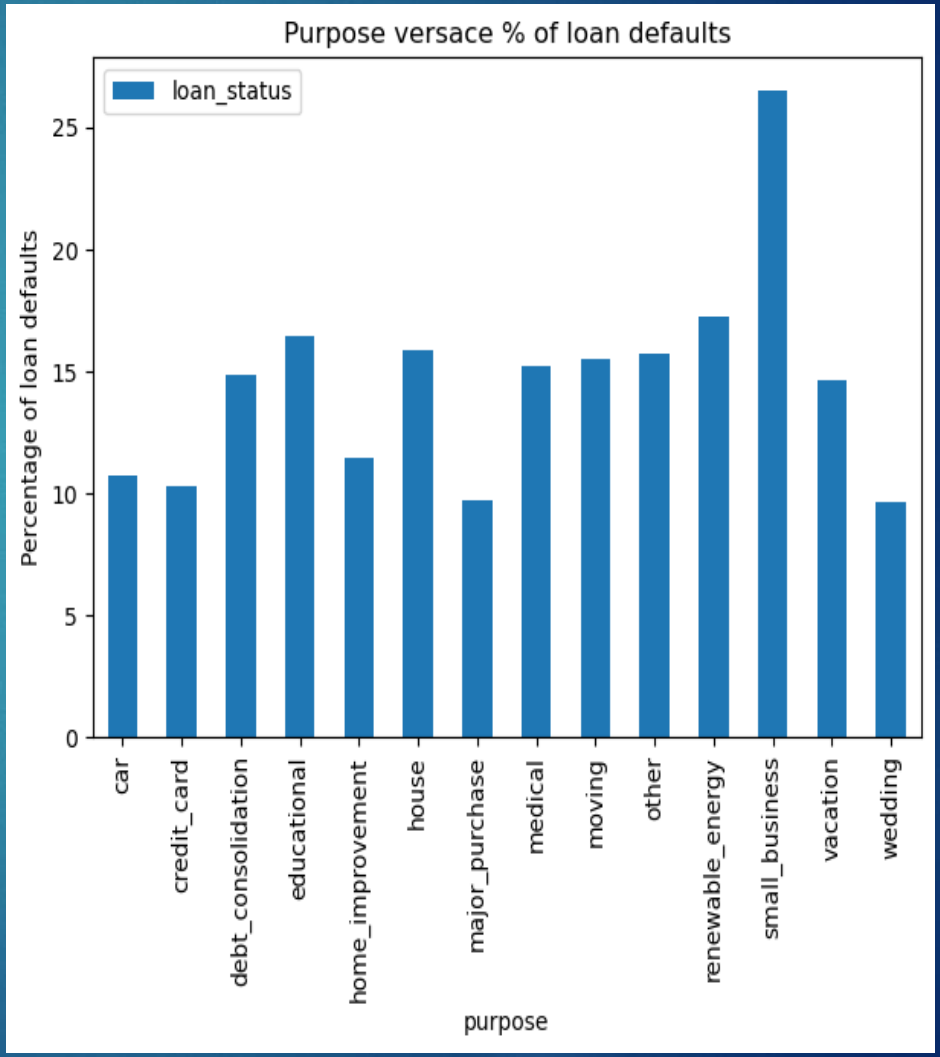
- We can infer that the percentage of loans fully paid are very less for loans availed for small_business category
- So we can do more scrutiny in future when loans are availed for small_business purpose



Bivariate analysis – Purpose / Loan Fully Paid, Charged off

Purpose vs Charged off loans

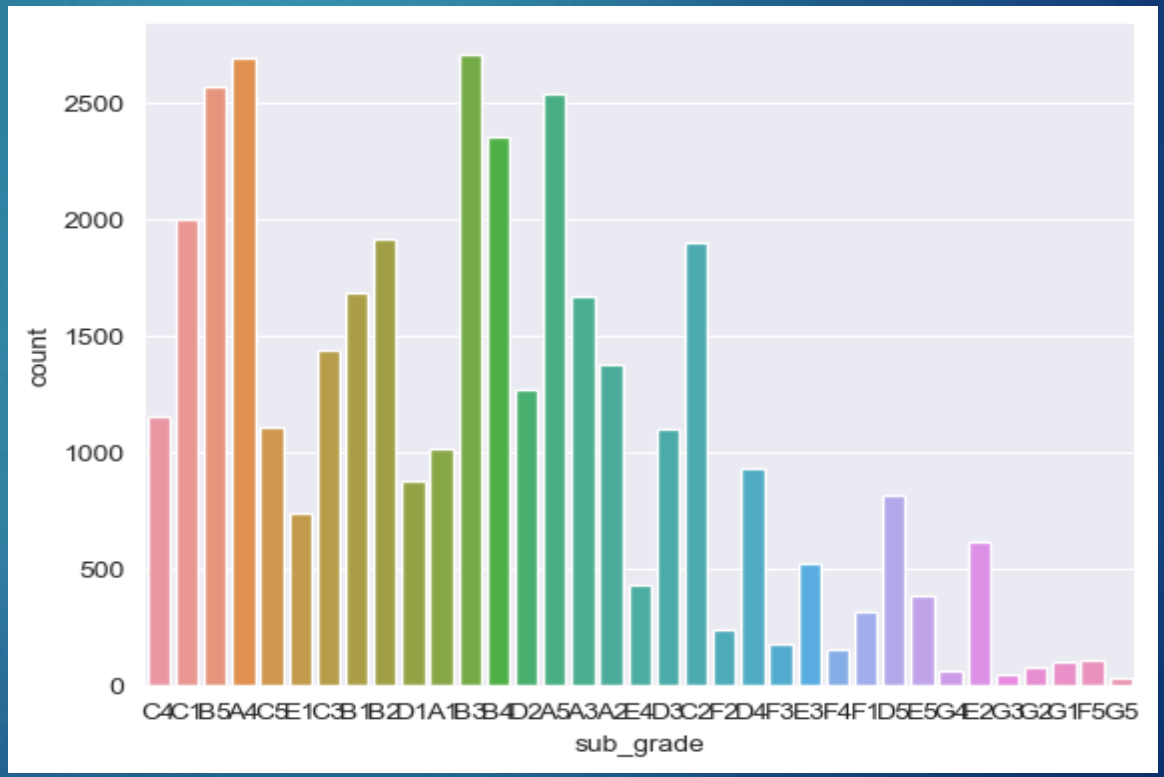
- We can infer that the percentage of loans defaulted are very high for loans availed for small_business category
- Other purpose for which we need more scrutiny while sanctioning the loans are renewable_energy,educational,medical, moving,house



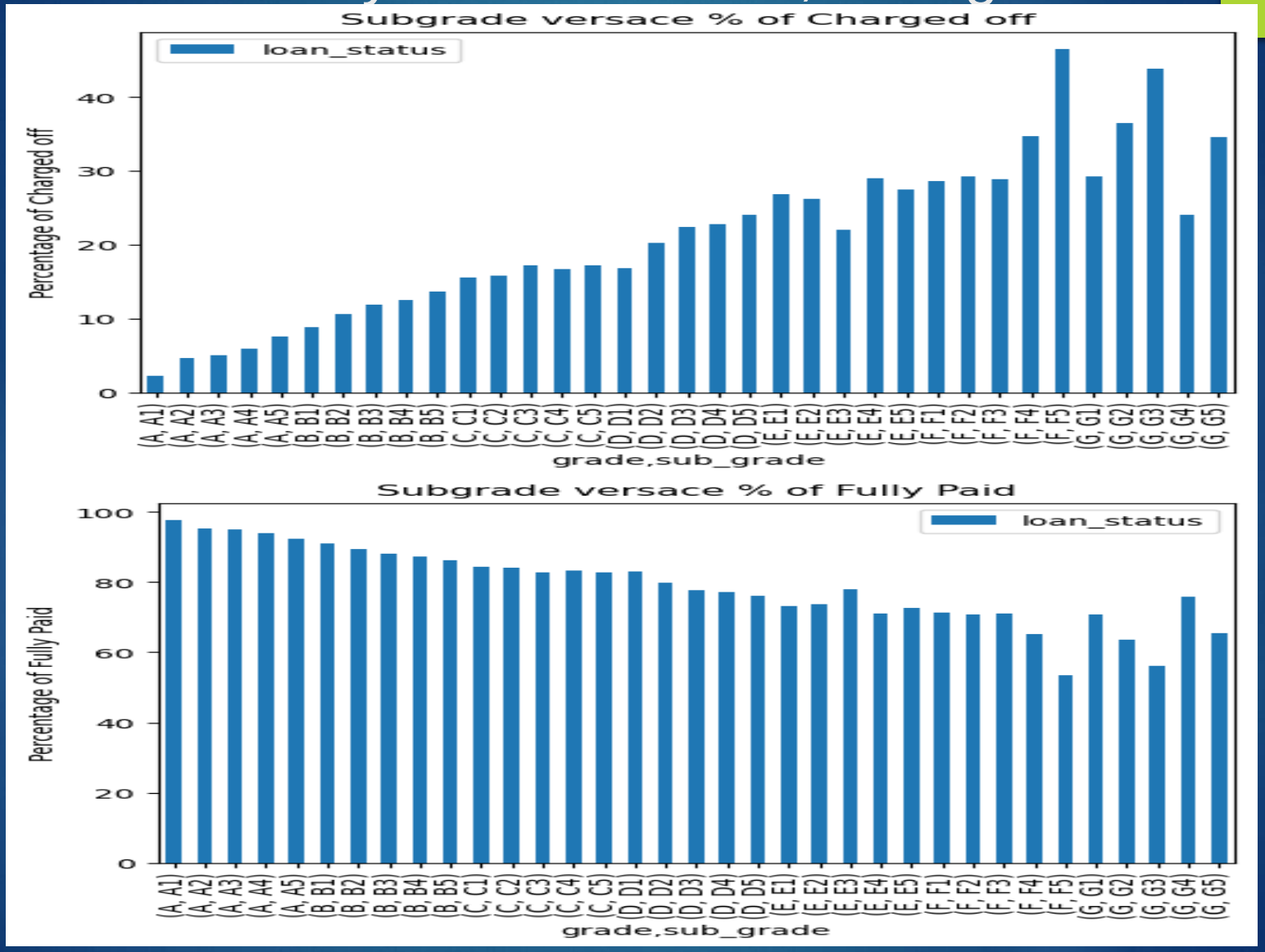
Univariate analysis – Sub Grade

Sub-Grade

- From Sub-Grade analysis also, highest loan acceptance is from sub-grades of Grade – B and A
- Lowest loan acceptance is in sub-grades of G



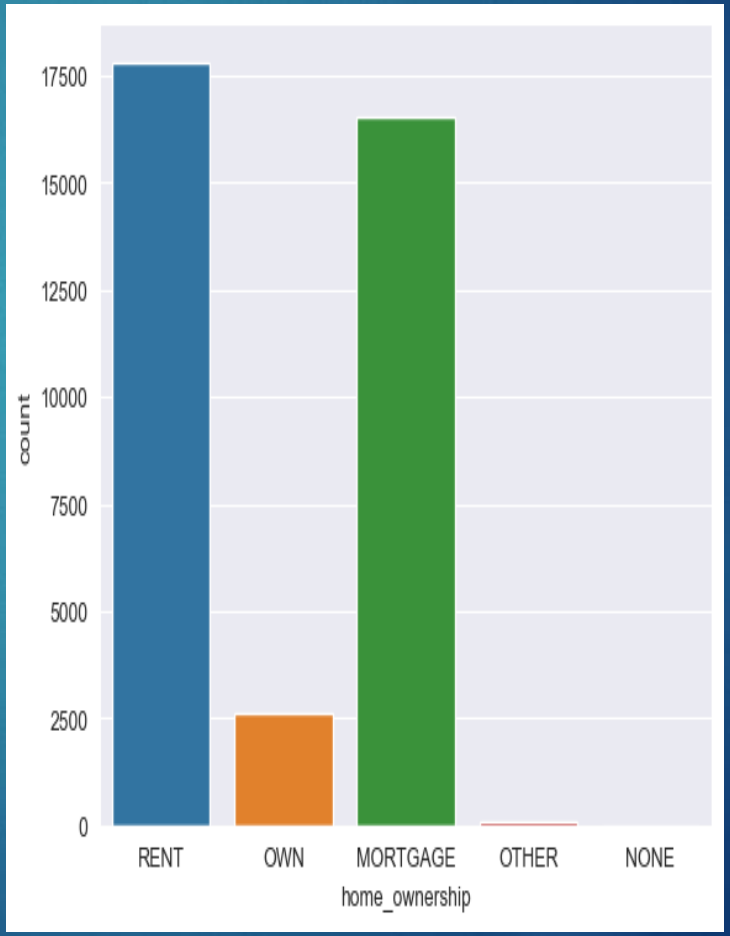
Bivariate Analysis - Loan Paid, Charged off



Univariate analysis – Home Ownership

Home ownership

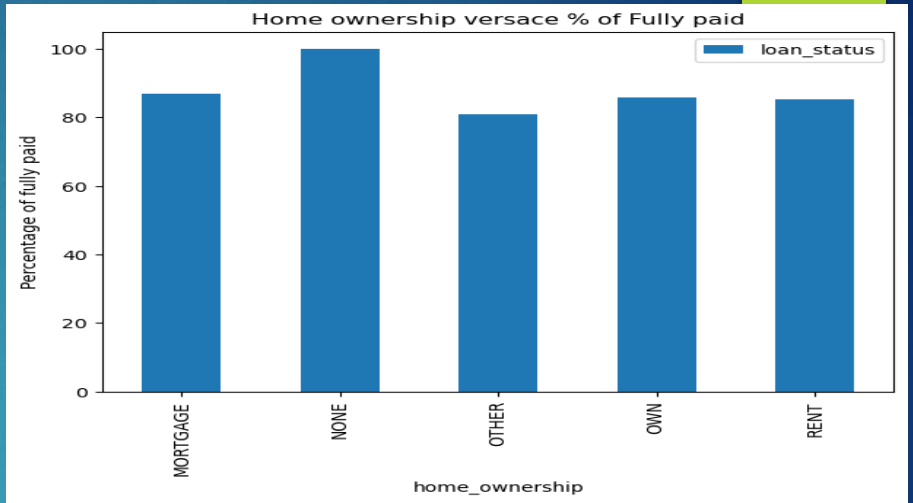
- Maximum loan acceptance is for people staying in rented house
- Second Maximum loan acceptance is for people staying in mortgage house
- Lowest acceptance is for people staying in other category house
- So more importancce should be given for loan defaults from rented house and mortgage house

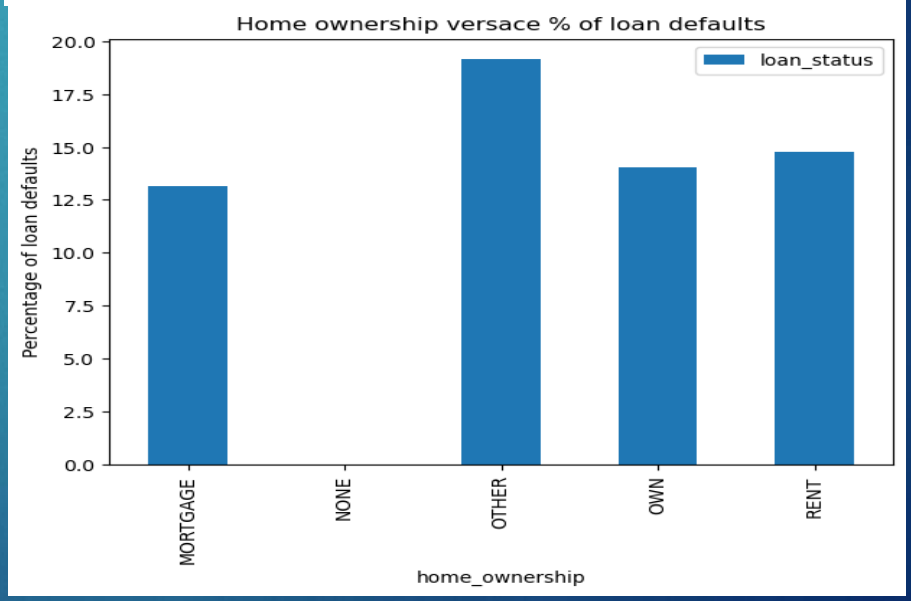


Bivariate analysis – Ownership / Loan Fully Paid, Charged off

Home Ownership vs Loans

- We can infer that loan repayments are high by customers who fall under Mortgage category.
- Followed by customers who are in rent and own house.

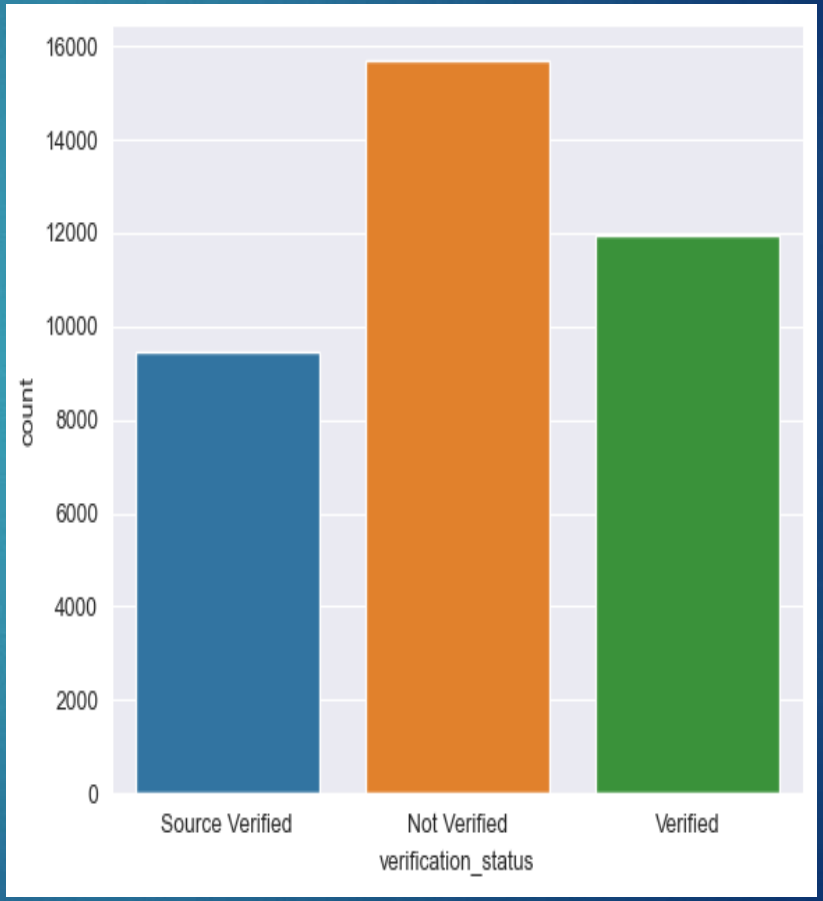




Univariate analysis - Verification

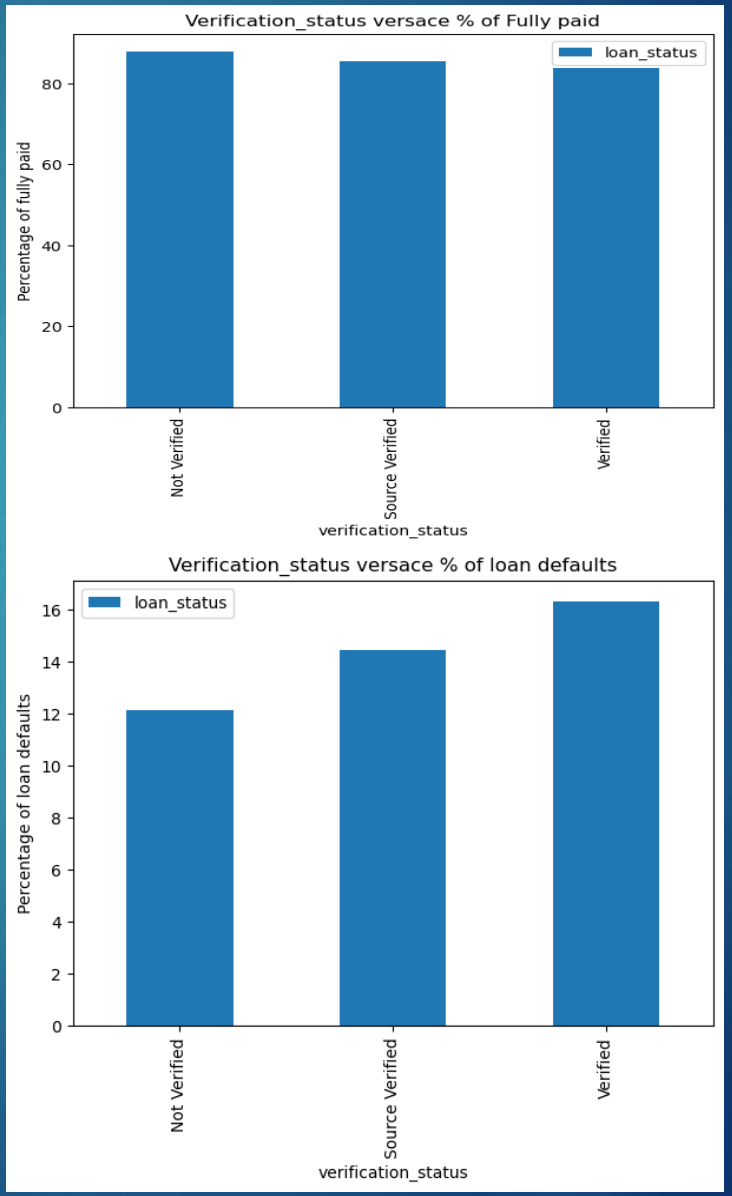
Verification

- Maximum loans are given without verification
- Hence need to check if more loan defaults are from not verified sources



Bivariate analysis – Verification / Loan Fully Paid, Charged off

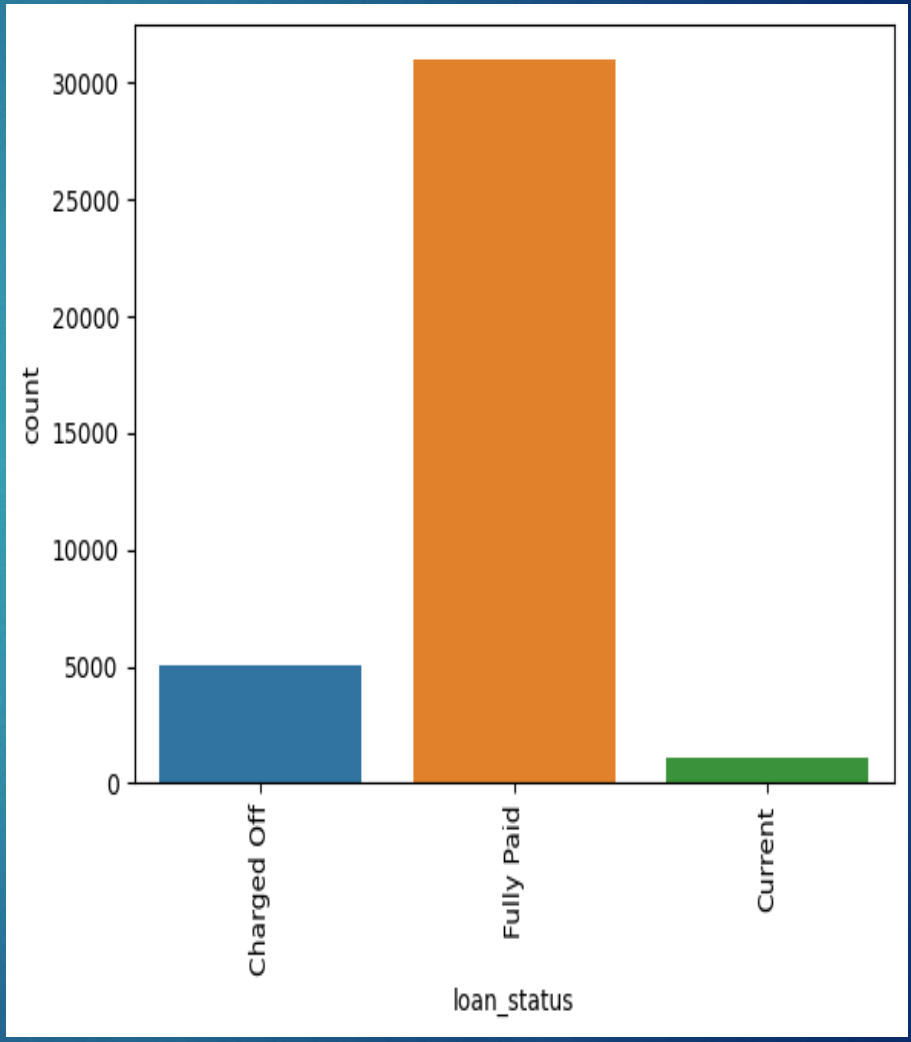
- Verification Status Vs Loans
- Loans which are granted without background Verification are in the higher risks of getting defaulted
- So loans with proper background verification of customers are better for financial institution



Univariate analysis – Loan Status

Loan_Status

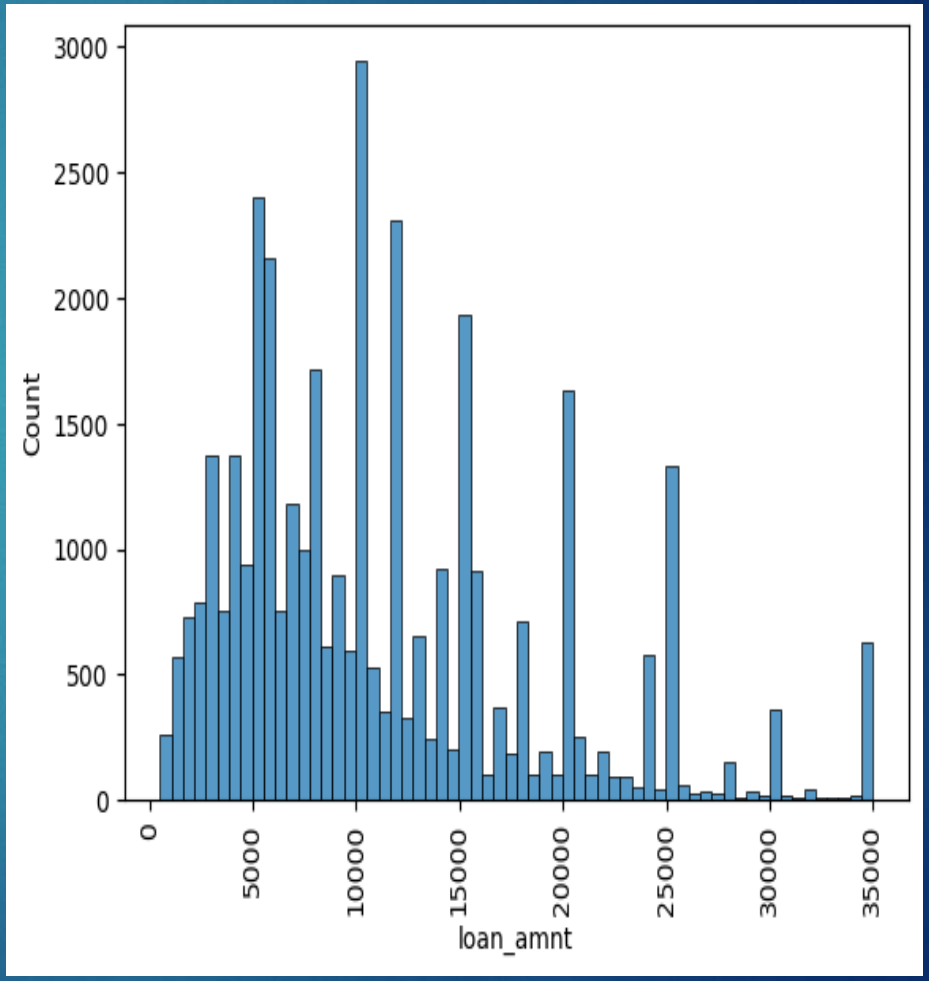
- We can infer that majority of loans are fully paid (30K+), while around 5K loans are charged off



Univariate analysis – Loan Amount

Loan_Amount

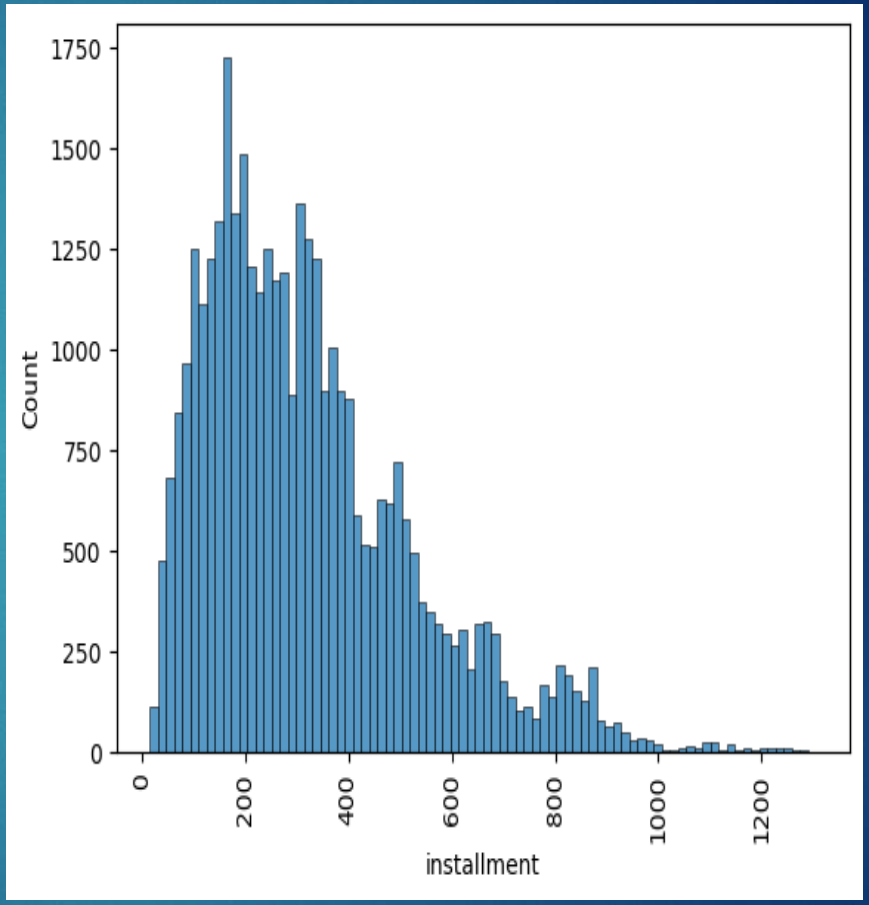
- We can find that loans disbursed in the range of 10K -13K are highest in number



Univariate analysis - Instalment

Instalment

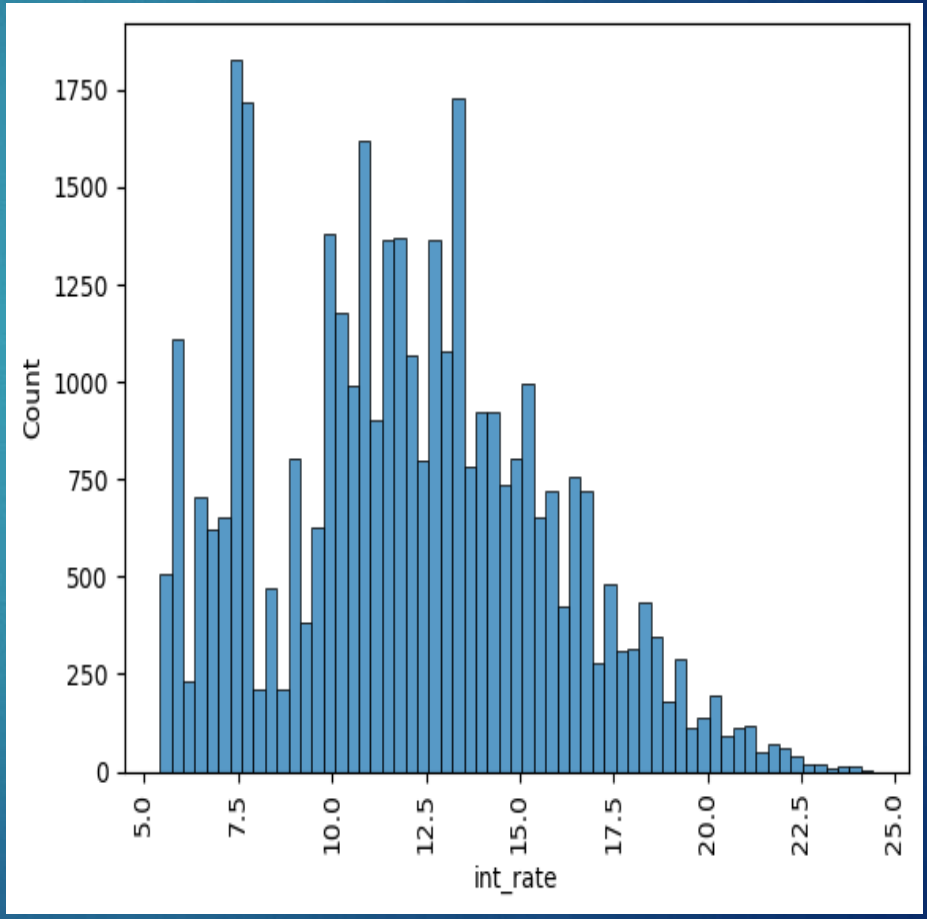
- Maximum loans are processed for 200 instalments
- We can find very less loans for 1200 instalments
- We can infer that people prefer loans which are of shorter instalments



Univariate analysis – Int_Rate

Int_rate

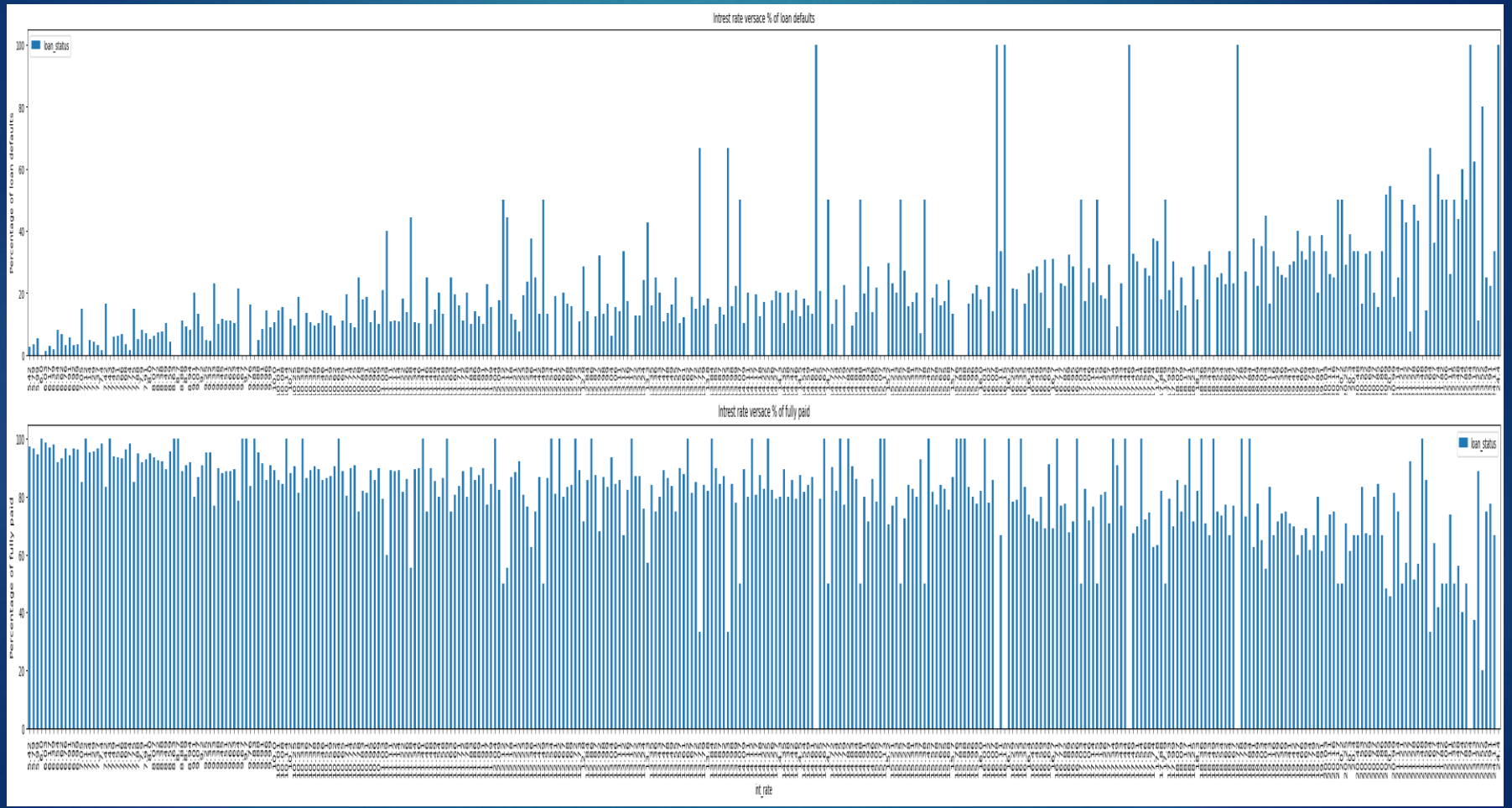
- We can infer that maximum loans are disbursed with an intrest rate in range of 7.5%
- Loan disbursal gradually reduced with higher intrest rate



Bivariate analysis – Int_Rate/ Loan Fully Paid, Charged off

- Interest rate vs Loans
- We can infer that loans which are granted with lower interest rate are re-payed to a greater extent than the loans which are with higher interest rate
- Loans disbursed with optimal interest rates are recovered well
- Loans with high interest rate like 14.62,16.08,16.15,17.46,18.72,24.4 have high defaulting rate
- Loans with low interest rate are having high repayment rates

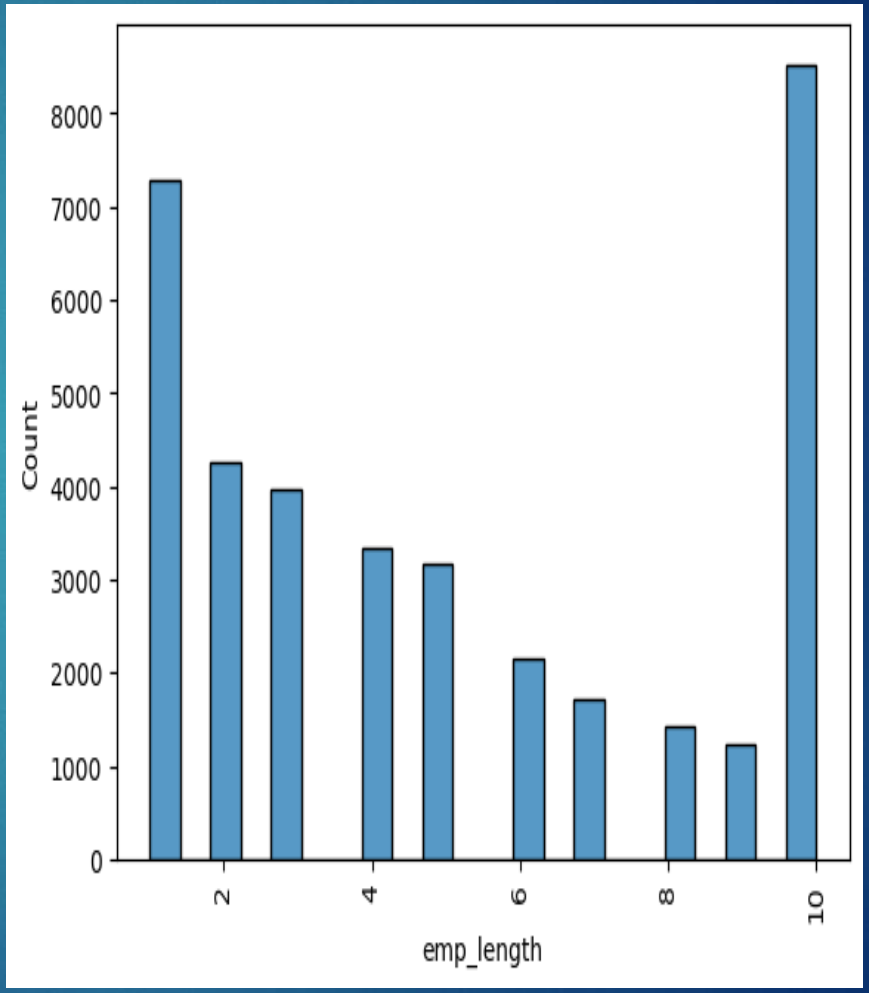
Bivariate analysis –Int_Rate/ Loan Fully Paid, Charged off



Univariate analysis – Employee Tenure

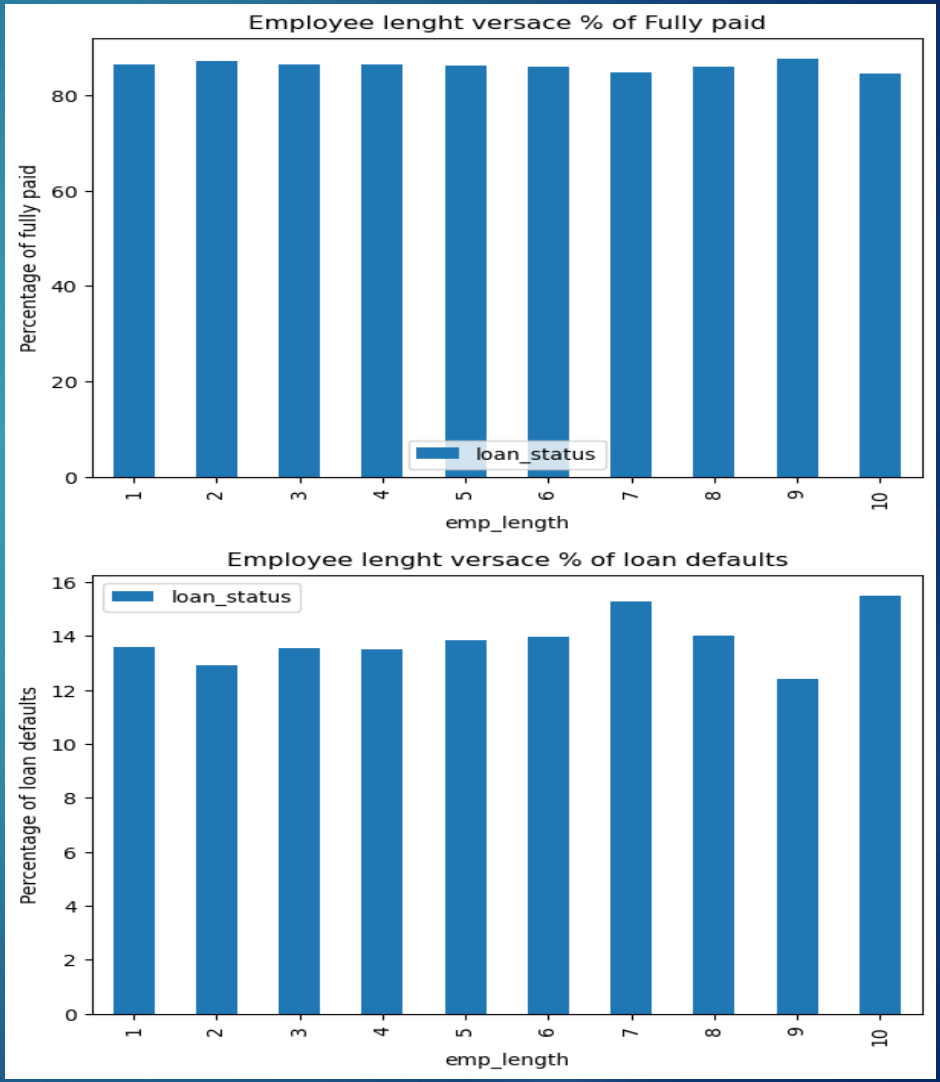
Employee Tenure

- We can find more employees having work experience of 10 years, followed by new joiners
- Loan disbursal pattern gradually decreases with employees who has experience in middle(between 1-10 years



Bivariate analysis /Employee Tenure- Loan Fully Paid, Charged off

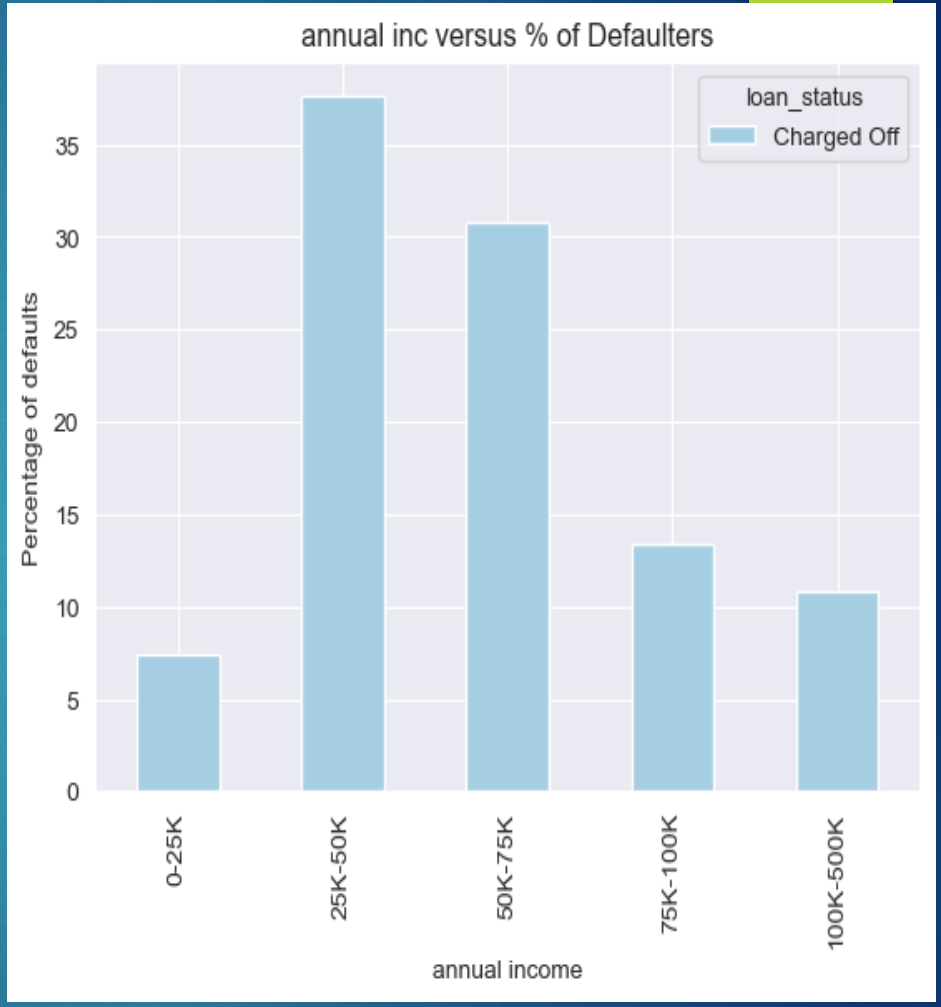
- Loan has been fully repaid by employees having tenure of 9 and 2 years
- We can find more defaulters in those employees having experience of 7 and 10 years



Insights of loan payment pattern through Bivariate analysis

- Annual Income vs Loans

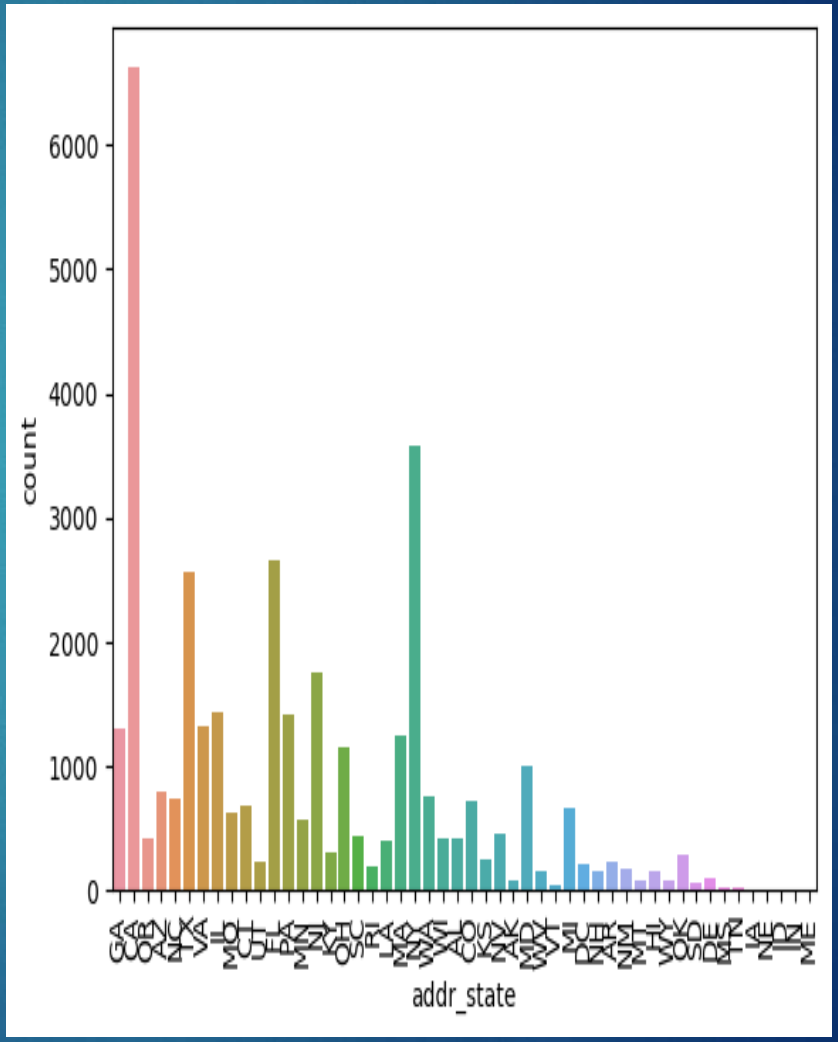
Customers who have annual income in range of 25K-50K , 50K-75K has defaulted more than the customers who have higher salary slab.



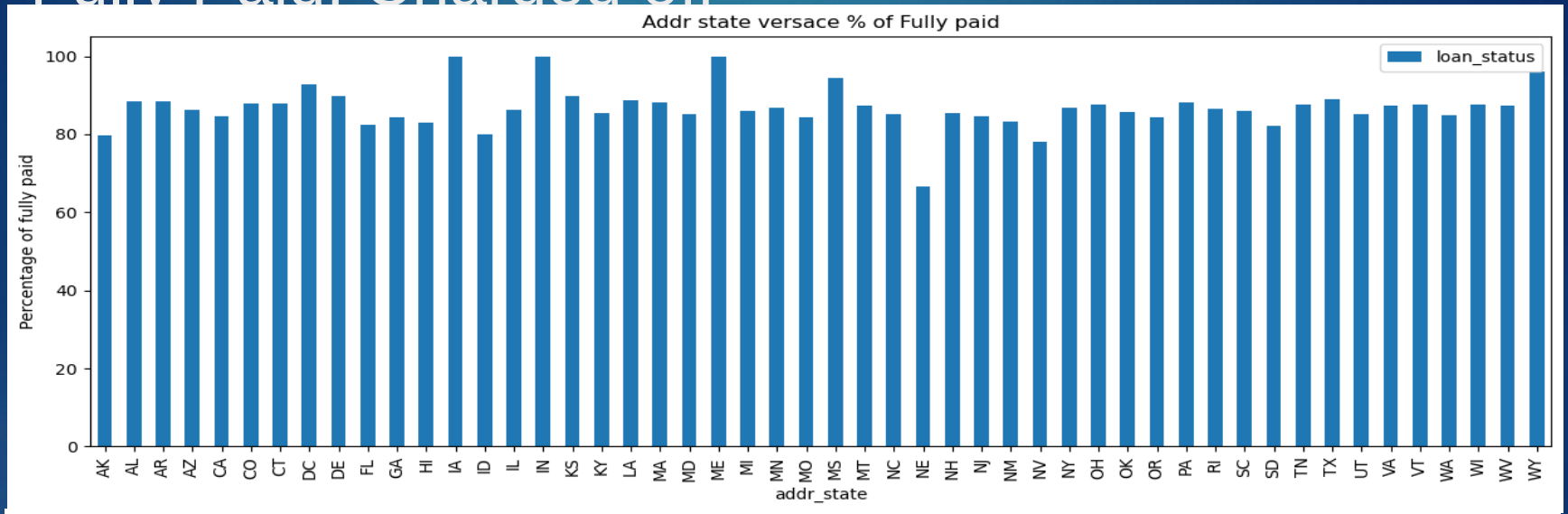
Univariate analysis – Address_State

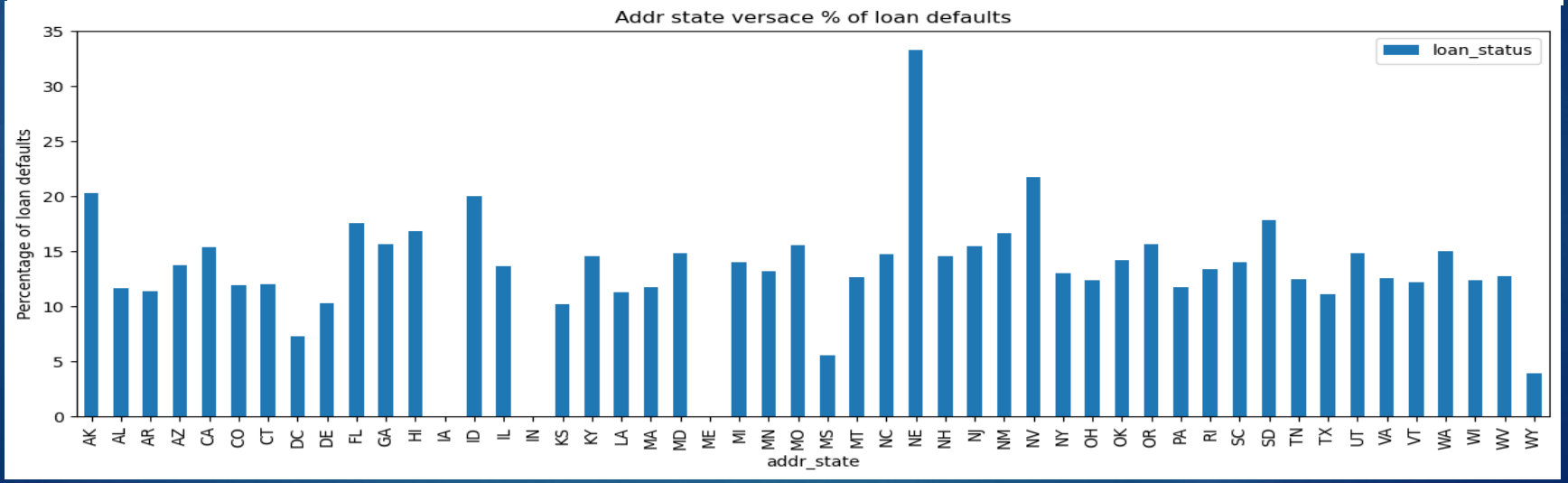
Address_State

- Loans are disbursed more in states like CA,NY,FL,TX.



Bivariate analysis – Address State- Loan Fully Paid, Charged off





Insights of loan disbursal pattern through Bivariate analysis

- We can infer that certain states like IA,IN,ME have high loan repayment
- Some states like NE,ID have high loan defaulters percentage
- We need to take more detailed analysis on states where defaulters percentage is high

Recommendations

- To provide loans which have shorter term and less interest
- To formulate strict loan disubursal process, for employees who are of grade G and F
- Loans sanctioned for small business category, needs more scrutiny
- We can infer that loans sanctioned for employees who have house in mortgage have higher repayments than other categories
- Loans should be sanctioned only after proper verification, as defaults are high in non - verified cases

Recommendations

- Loans with shorter instalment periods, and less interest rates are repaid fully than the longer period
- Loans are mostly repaid by high tenured employess than the employees having experience in middle years
- We can infer that certain states like IA,IN,ME have high loan repayment
- Some states like NE,ID have high loan defaulters percentage
- We need to take more detailed analysis on states where defaulters percentage is high