Lending Club Exploratory Data Analysis

by:

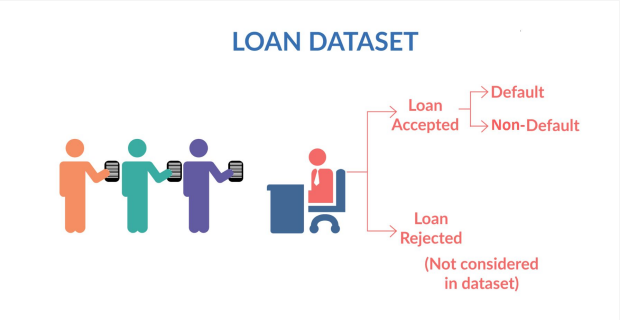
Ajit Gaikwad

Akhil Kumar Garg

Business Overview

A **consumer finance company** which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to 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:** 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 (because the candidate does not meet their requirements etc.). 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 (and thus in this dataset)

Problem Statement

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'.

1. Understand the **driving factors (or driver variables)** behind loan default

2. Identify risky loan applicants using EDA

Our Approach

1. Go through different attributes in Loan data set, use reference of Data dictionary and understand the given data e.g numerical, categorical and null values
2. Import data into dataframe using python
3. Data cleaning:
    a. Identify and remove columns having all null values
    b. Identify percentage of null data in each columns
    c. Remove columns which have higher percentage (36%+) of null data
    d. Identify and remove rows which were having null data. Dropped null data from rows where null data percentage was low
    e. Removed extra string from object columns like emp_lenght, int_rate, term and revol_util and converted to int/float data type
    f. Create derived column IncInstallmentRatio to identify installment ratio on income
    g. Ensure there is no null data in final data set
4. Data Analysis: We performed 2 types of analysis
    - Univariate: We analyzed below columns
        i. Loan Status
        ii. Grade
        iii. Employment Length
        iv. Home Ownership
        v. Verification Status
        vi. Purpose
        vii. Address state
        viii. Loan Term

Our Approach

- Bivariate: We analyzed below columns combinations
  i. Loan status and Employment length
  ii. Loan status and Annual income
  iii. Loan status and interest rate
  iv. Loan status and dti
  v. Loan status and installment
  vi. Loan status and loan amount
  vii. Loan status and IncInstallmentRatio(Income & Installment Ratio)
  viii. Loan status and Annual income (after removing outliers)
  ix. Loans status and inq_last_6mths
  x. Loan status and revol_util
  xi. Loan status and total_acc
  xii. Loan status and grade
  xiii. Loan status and state address
  xiv. Loan status and home ownership
  xv. Loan status and term

Analysis Summary

Univariate Analysis

Loan Status	Total count and percentage of loan customer by their loan status (Fully Paid/Charged Off/Current). Fully Paid: 13982(84%) Most of the customers are in fully paid category Charged : 2152(13%) Current : 540 (3%)
Grade	Total count and percentage of loan customer by Grade (A/B/C/D/E/F/G) Highest customer in Grade B 5144 (31%) Lowest customer in Grade G 118 (1%)
Employment Length	Employees who have employment length 10 and more are largest customer 3959 (24%) second highest are customers who have 1 years experience
Home Ownership	Rent and Mortgage type customer are the largest customer base. Customers who have own house are only 7%.
Verification Status	42% non verified customers are the highest loan takers. High percentage of non verified customer shows company is not having strict policy or measures to verify customer details like address, income sources, credit history. This increases risk of loan defaulters
Purpose	Highest no. of customers are taking loan for debt consolidation. This has direct relation with home ownership, as many customers who have house mortgage may be using this loan for debt consolidation
Address	Plotted most loan taking states in descending order which will eventually help to target customers to increase business.
Loan Term	72% peoples taken loan for 36 month tenure and remaining 28% taken loan for 60 month tenure.

Analysis Summary

Bivariate Analysis

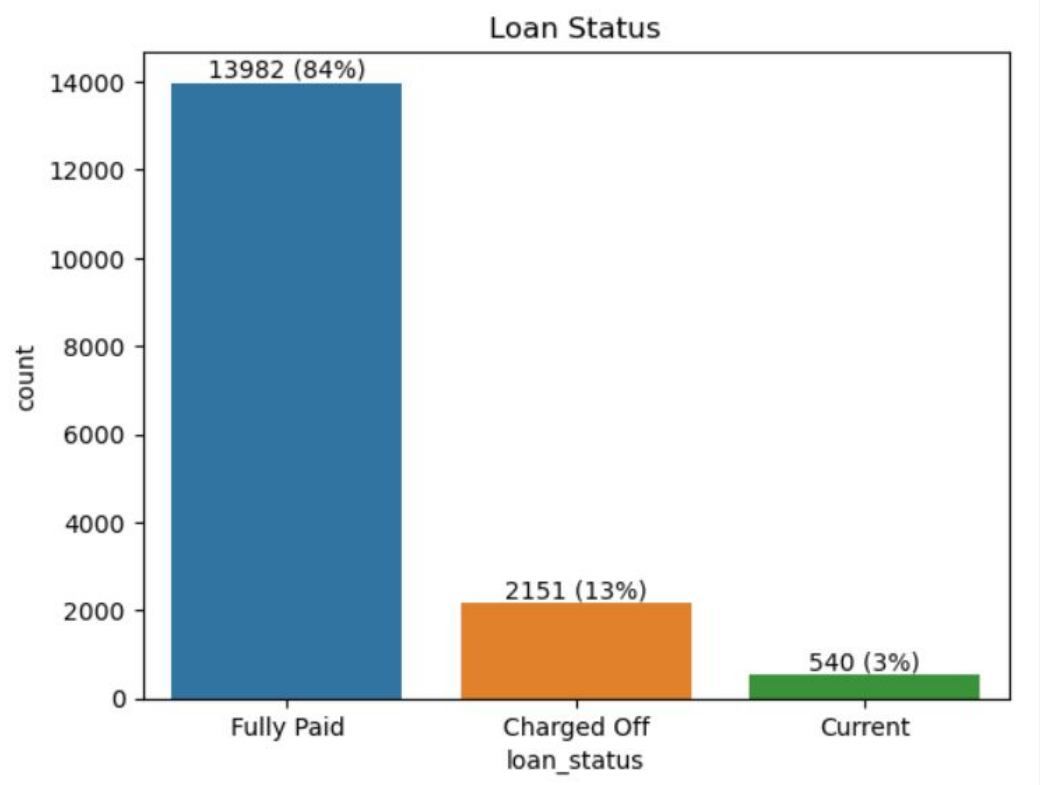
Loan status and Employment length	After plotting box plot using Loan status and employment length we found no correlation between employment length and loan status.
Loan status and Annual income	After plotting box plot we found interquartile range of charged off customers are below compare to fully paid customer. We can infer customers who have low annual income have chances of defaulters
Loan status and interest rate	We plotted box plot using Loan status and interest rate and we found 25th to 75the percentile of charged off customers are on high interest rate. Customers who have high interest rate have high changes of defaulters
Loan status and dti	We plotted box plot using Loan status and dti(debt to income ratio) and we found 25th to 75the percentile of charged off customers are on high dti. Customers who have high dti ratio have high changes of defaulters
Loan status and installment	After plotting box plot using Loan status and installment we found 75th percentile of charged off customer is high which means high installment amount customer have chances of defaulters
Loan status and loan amount	After plotting box plot using Loan status and Loan Amount we found 75th percentile of charged off customer is high which means high loan amount customer have chances of defaulters
Loan status and IncInstallmentRatio(Income & Installment Ratio)	After plotting box plot using Loan status and Income installment ratio we found 25th and 75th percentile of charged off customer is high which means high loan amount customer have chances of defaulters

Analysis Summary

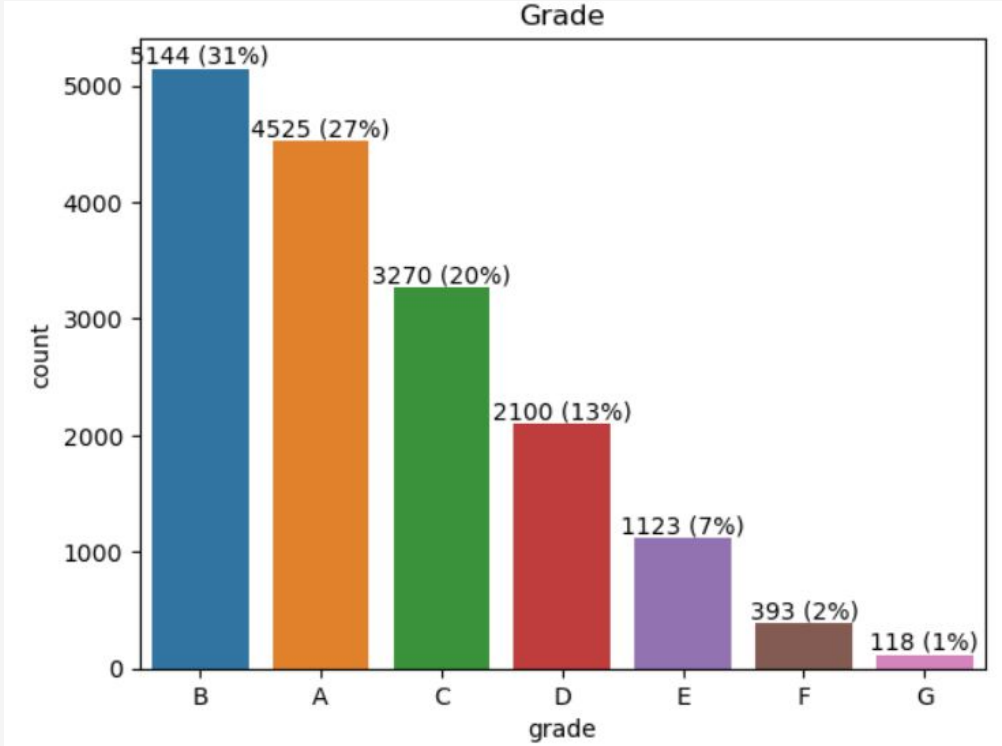
Bivariate Analysis

Loans status and inq_last_6mths	After plotting box plot using Loan status and last 6 months inquiry attribute we can analyse ● if customer enquiries are more than one in last 6 months then there is high chances of charged off
Loan status and revol_util	We plotted box plot using Loan status and revol_util and identified customer who are having more utilization rate compared to borrowed amount have high chances of charged off
Loan status and total_acc	Box plot of Loan status and total account shows plot is more or less same for both changed off and fully paid customers. From the graph we can conclude there is no direct impact of total account on charged off customers
Loan status and Grade	Using Box plot of Loan status and grade we identified B, C & D grades are having high chances of defaulters
Loan status and state address	Box plot for both charged off and fully paid customers is same for loan status vs state address which indicates no direct relationship.
Loan status and home ownership	We have created derive column based on home ownership ("RENT":1,"OWN":2,"MORTGAGE":3,"OTHER":4) From the graph we can conclude there is no significant direct impact of loan status and home ownership
Loan status and term	People who are taking loan for 60 months of tenure are likely to default.

Visualisation – **Univariate Analysis**

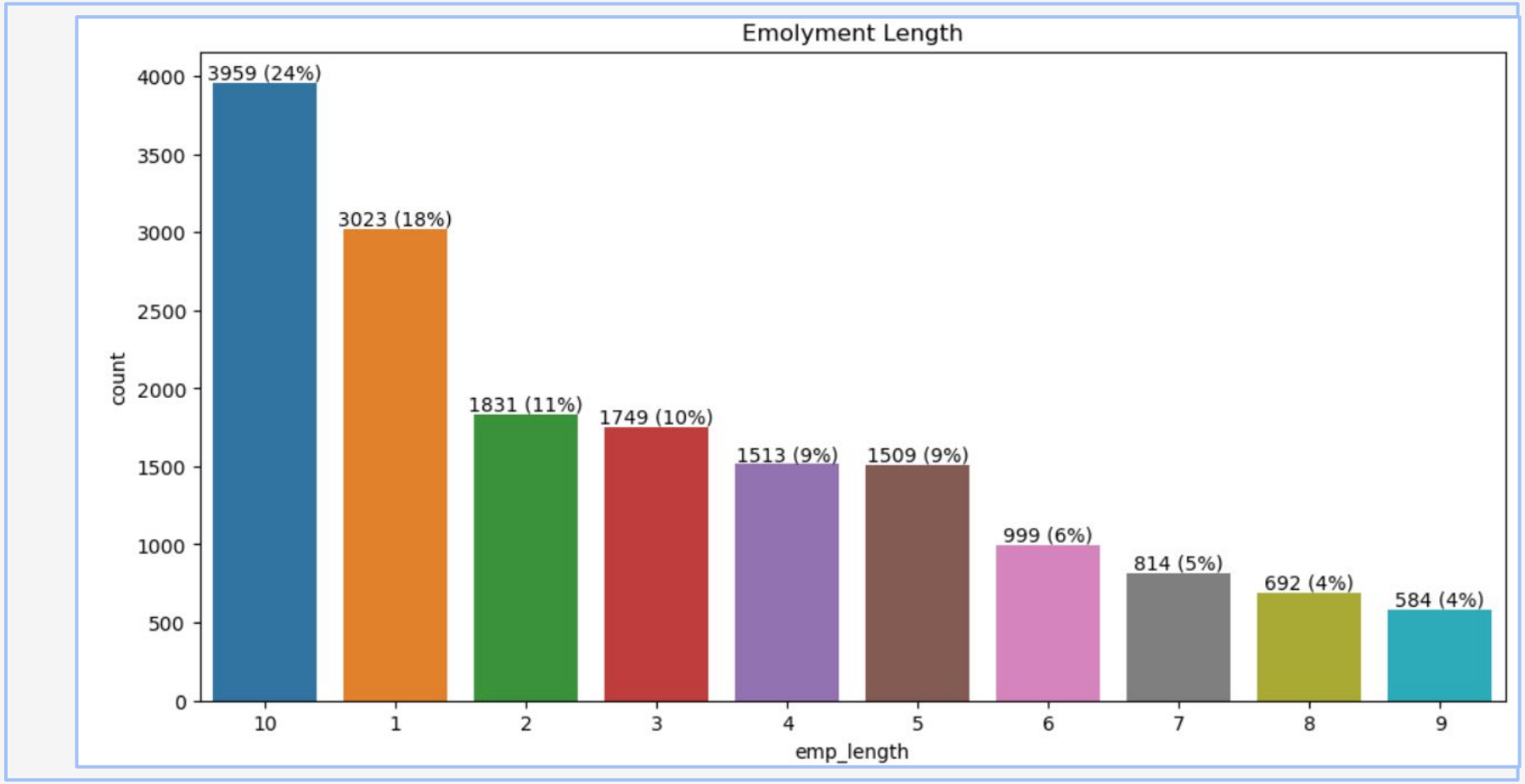


Total count and percentage of loan customer by their loan status (Fully Paid/Charged Off/Current).
Fully Paid: 13982(84%) Most of the customers are in fully paid category
Charged : 2152(13%)
Current : 540 (3%)



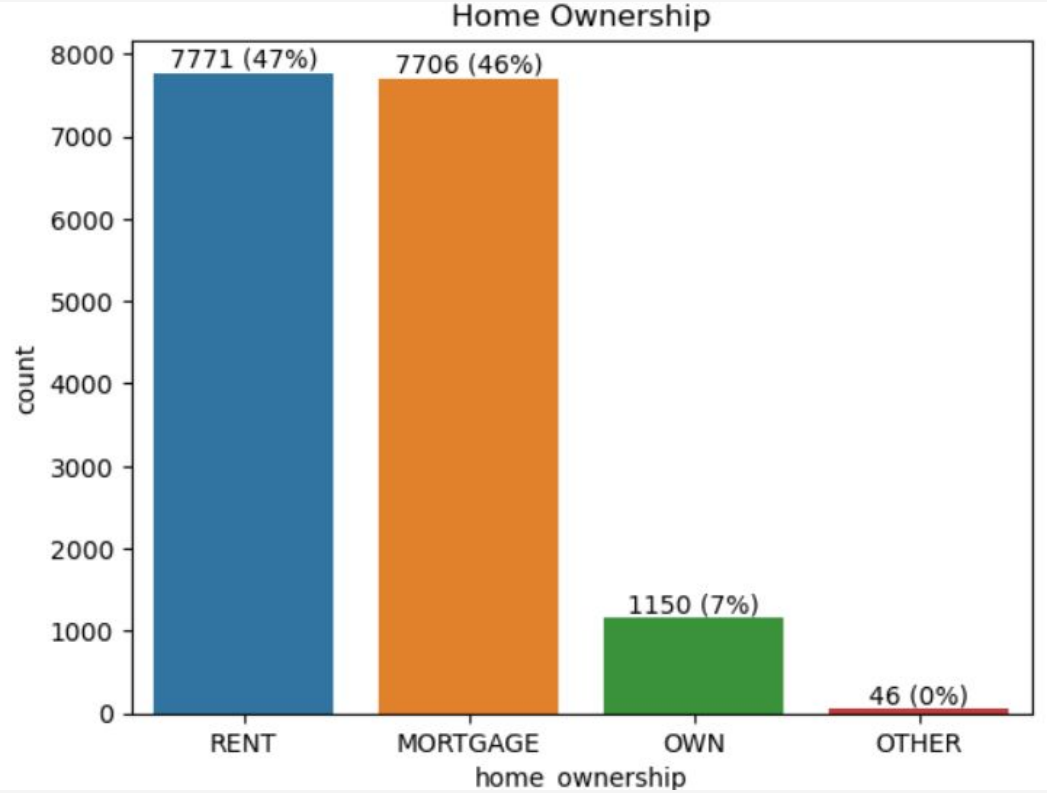
Total count and percentage of loan customer by Grade (A/B/C/D/E/F/G)
Highest customer in Grade B 5144 (31%)
Lowest customer in Grade G 118 (1%)

Visualisation – **Univariate Analysis**

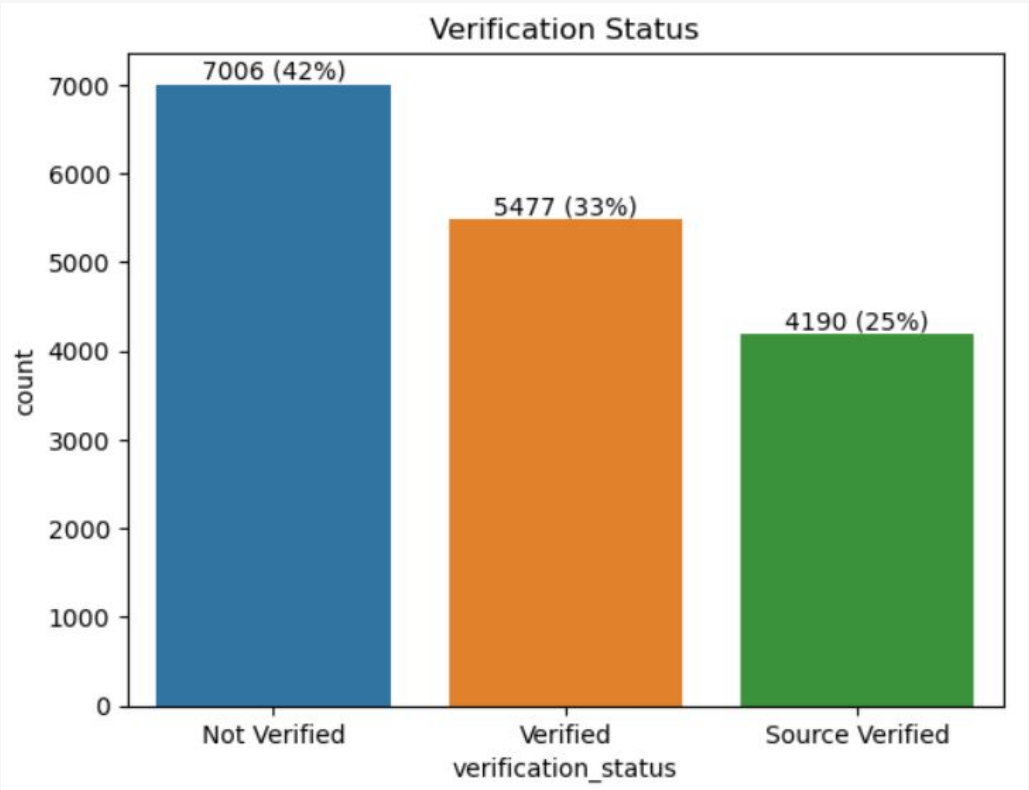


Employees who have employment length 10 and more are largest customer 3959 (24%)
second highest are customers who have 1 years experience

Visualisation – **Univariate Analysis**



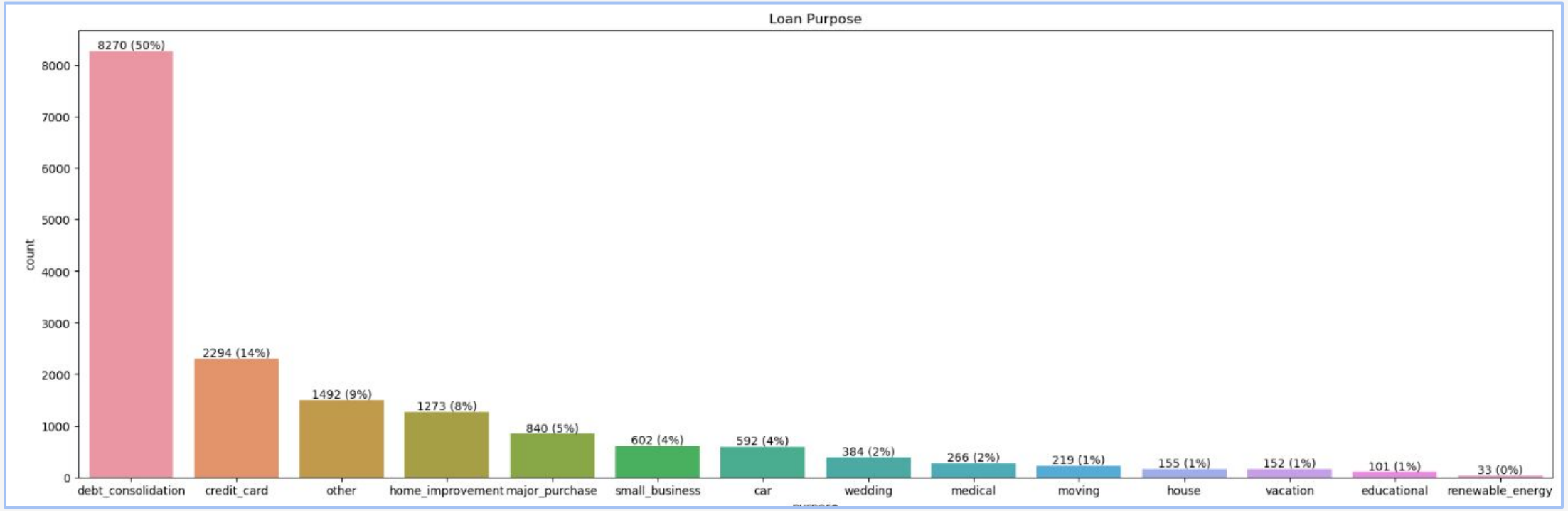
Rent and Mortgage type customer are the largest customer base. Customers who have own house are only 7%.



42% non verified customers are the highest loan takers. High percentage of non verified customer shows company is not having strict policy or measures to verify customer details like address, income sources, credit history. This increases risk of loan defaulters

Visualisation – **Univariate Analysis**

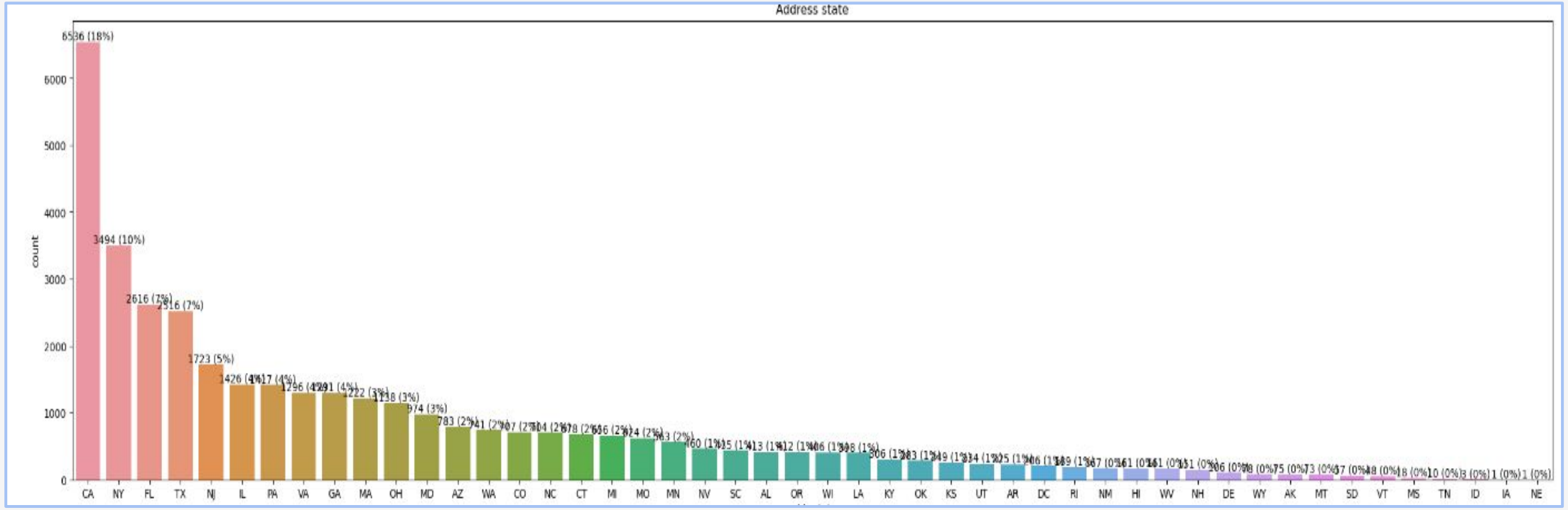
Loan Purpose



Highest no. of customers are taking loan for debt consolidation. This has direct relation with home ownership, as many customers who have house mortgage may be using this loan for debt consolidation

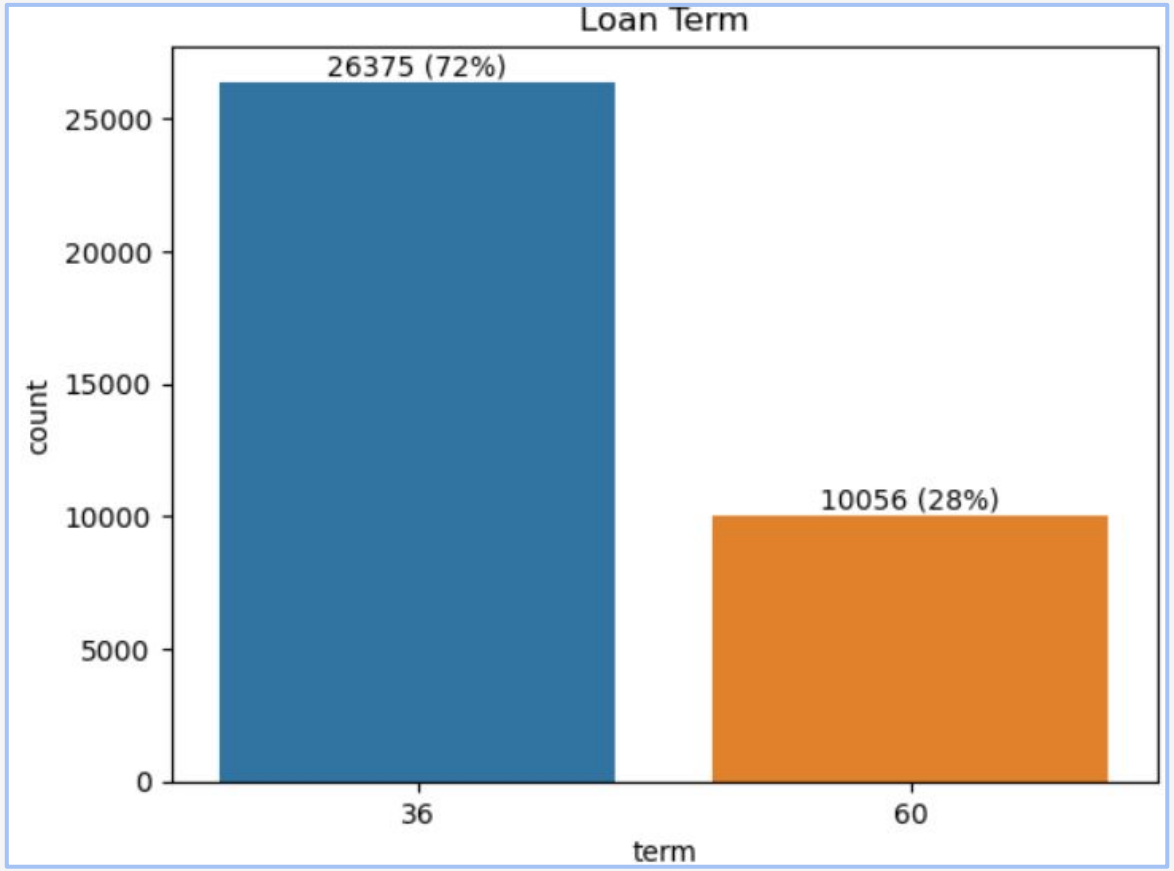
Visualisation – **Univariate Analysis**

Address State



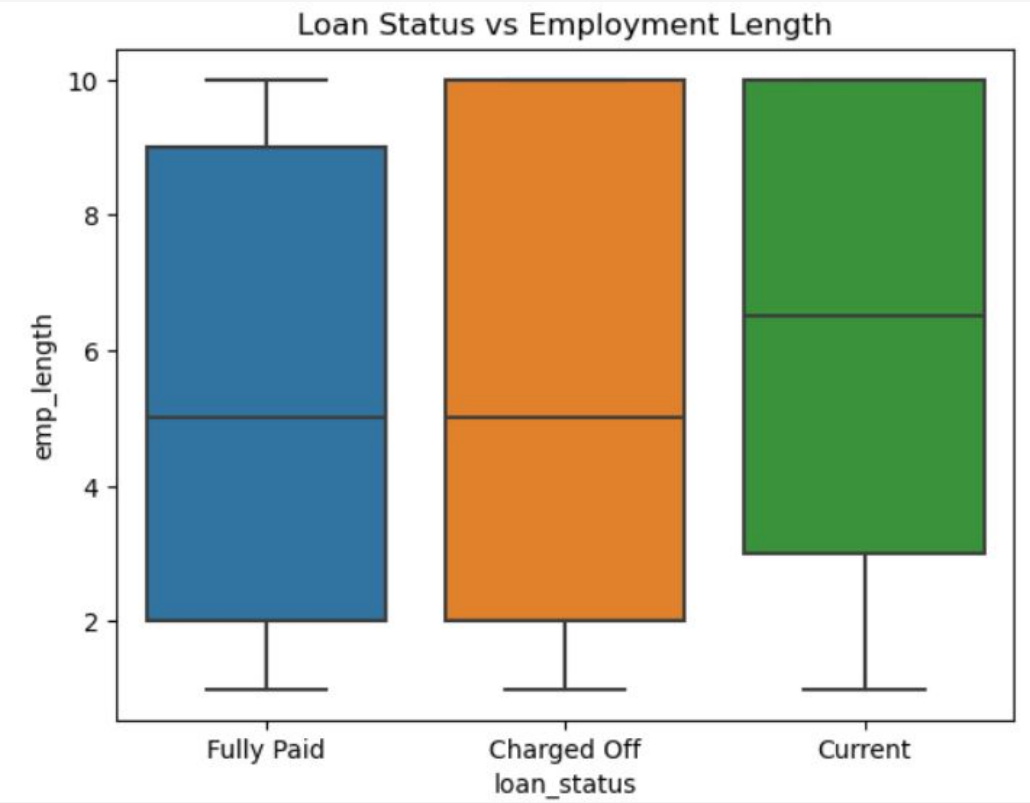
Plotted most loan taking states in descending order which will eventually help to target customers to increase business

Visualisation – **Univariate Analysis**

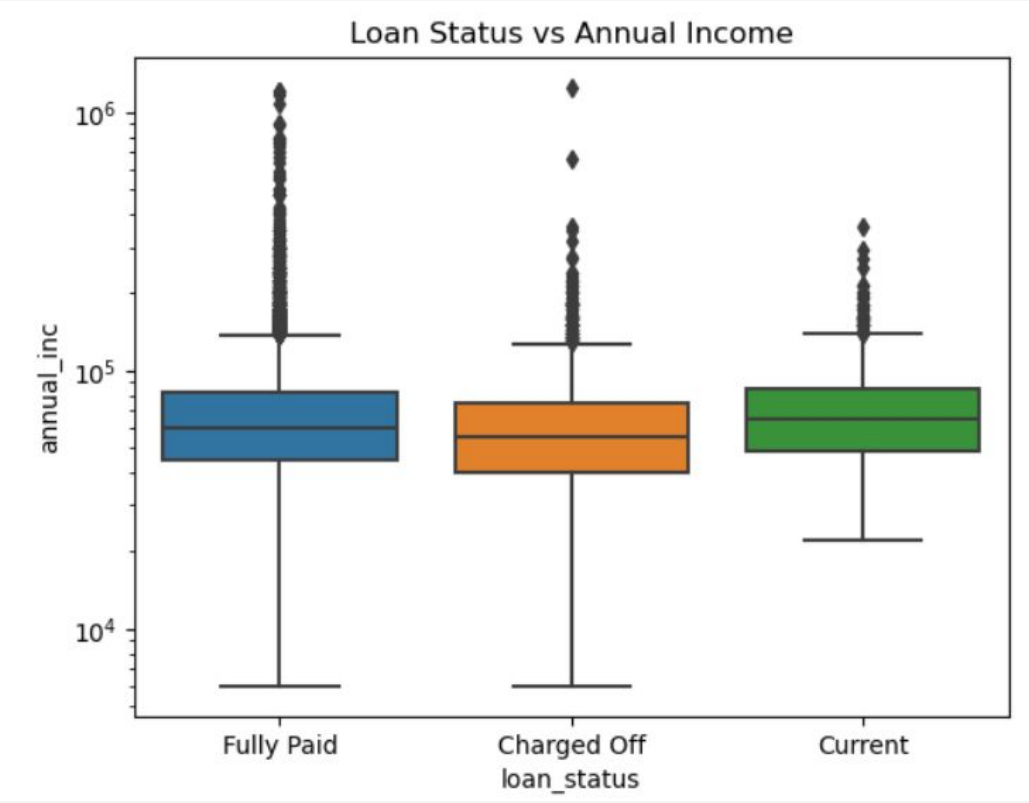


72% peoples taken loan for 36 month tenure and remaining 28% taken loan for 60 month tenure

Visualisation – **Bivariate Analysis**

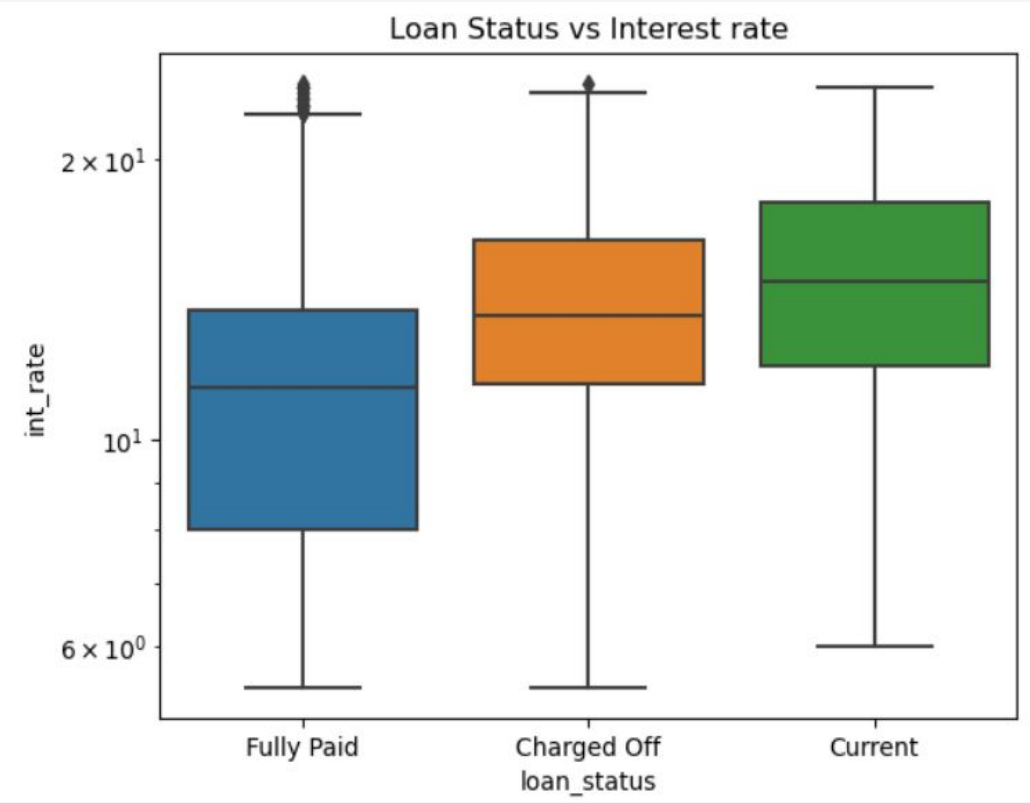


After plotting box plot using Loan status and employment length we found no correlation between employment length and loan status.

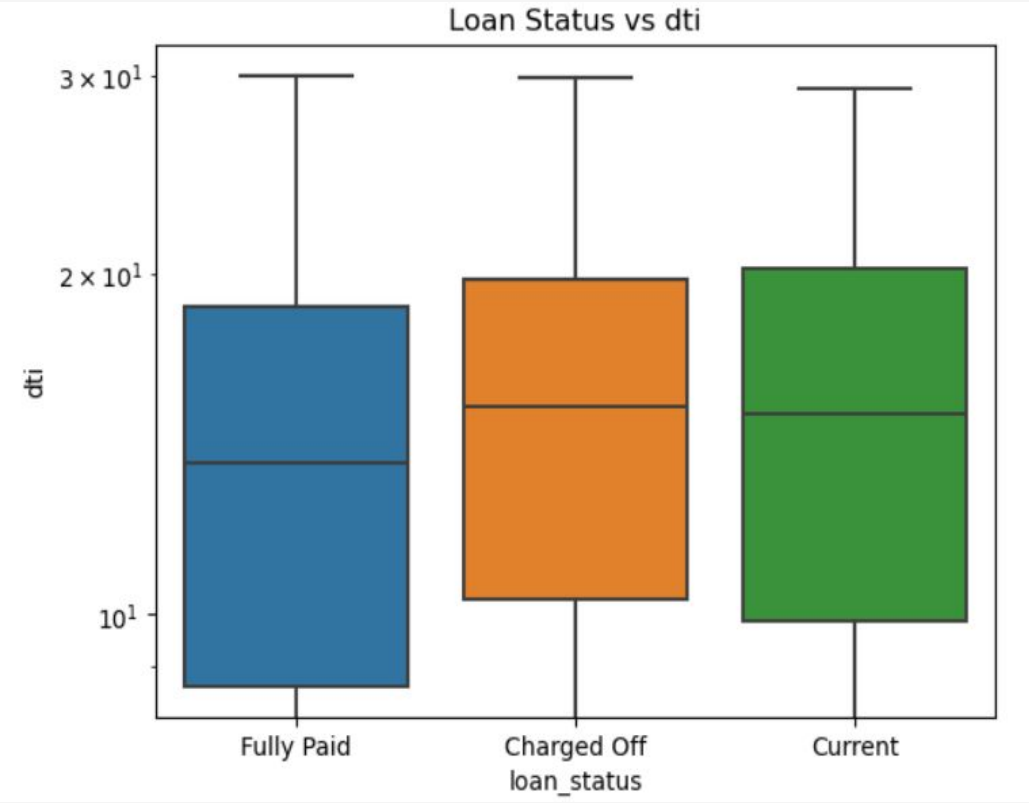


After plotting box plot we found interquartile range of charged off customers are below compare to fully paid customer. We can infer customers who have low annual income have chances of defaulters

Visualisation – **Bivariate Analysis**

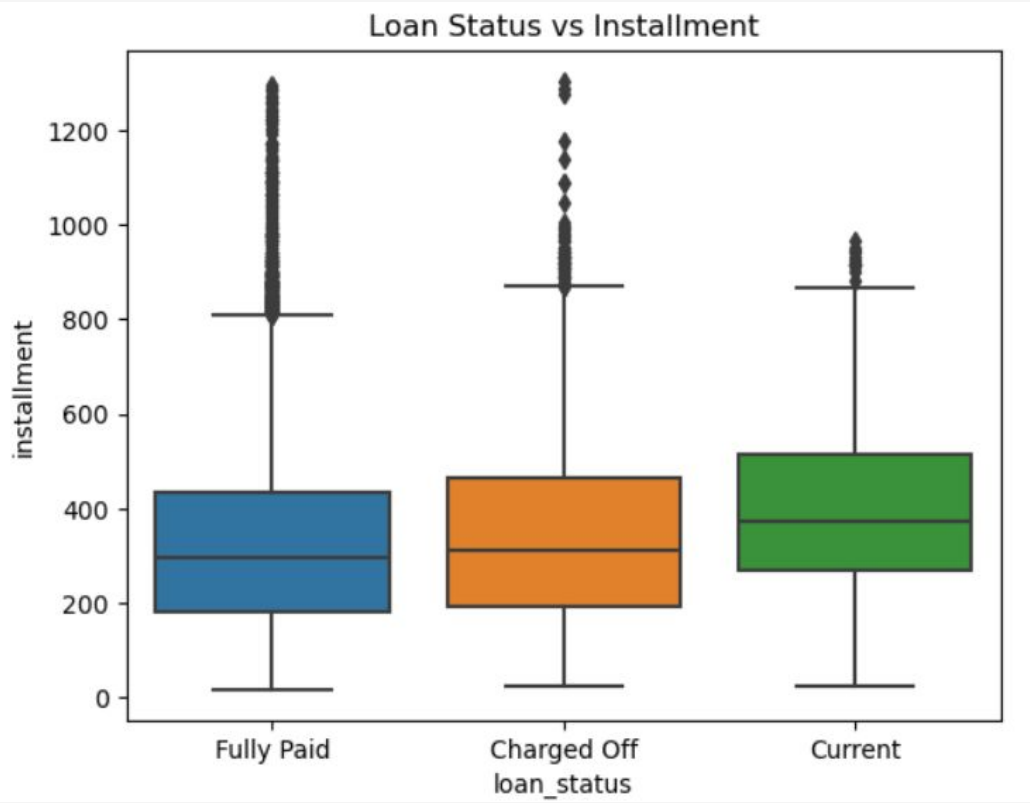


We plotted box plot using Loan status and interest rate and we found 25th to 75the percentile of charged off customers are on high interest rate. Customers who have high interest rate have high changes of defaulters

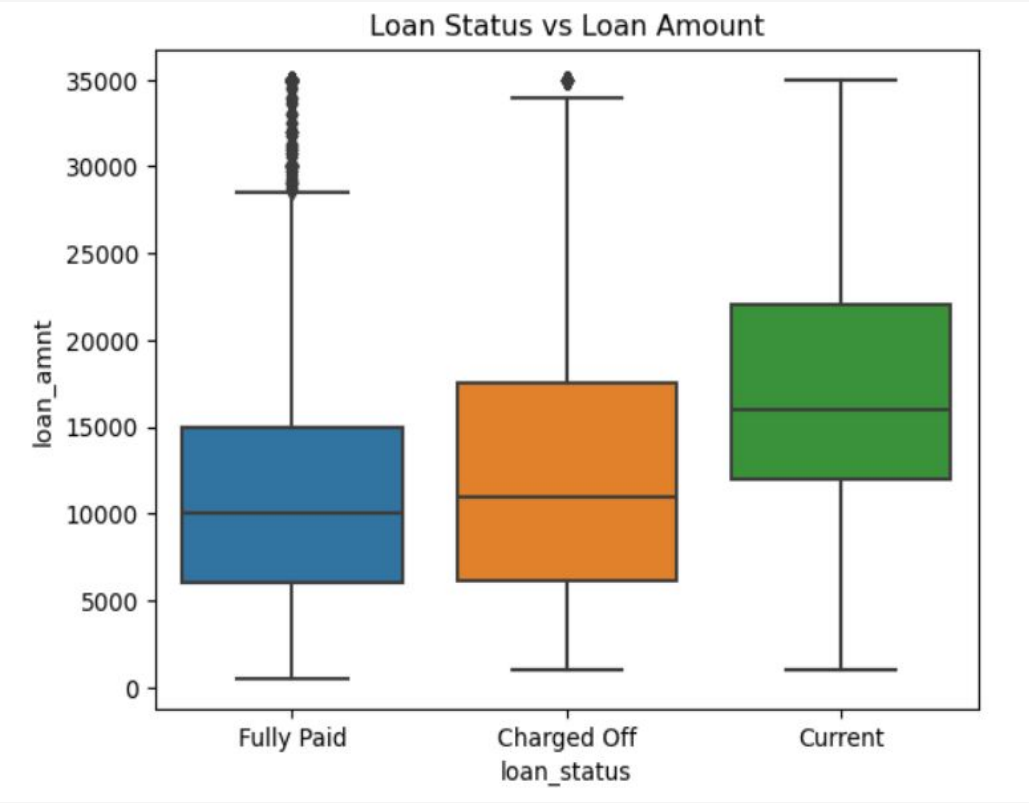


We plotted box plot using Loan status and dti(debt to income ratio) and we found 25th to 75the percentile of charged off customers are on high dti. Customers who have high dti ratio have high changes of defaulters

Visualisation – **Bivariate Analysis**

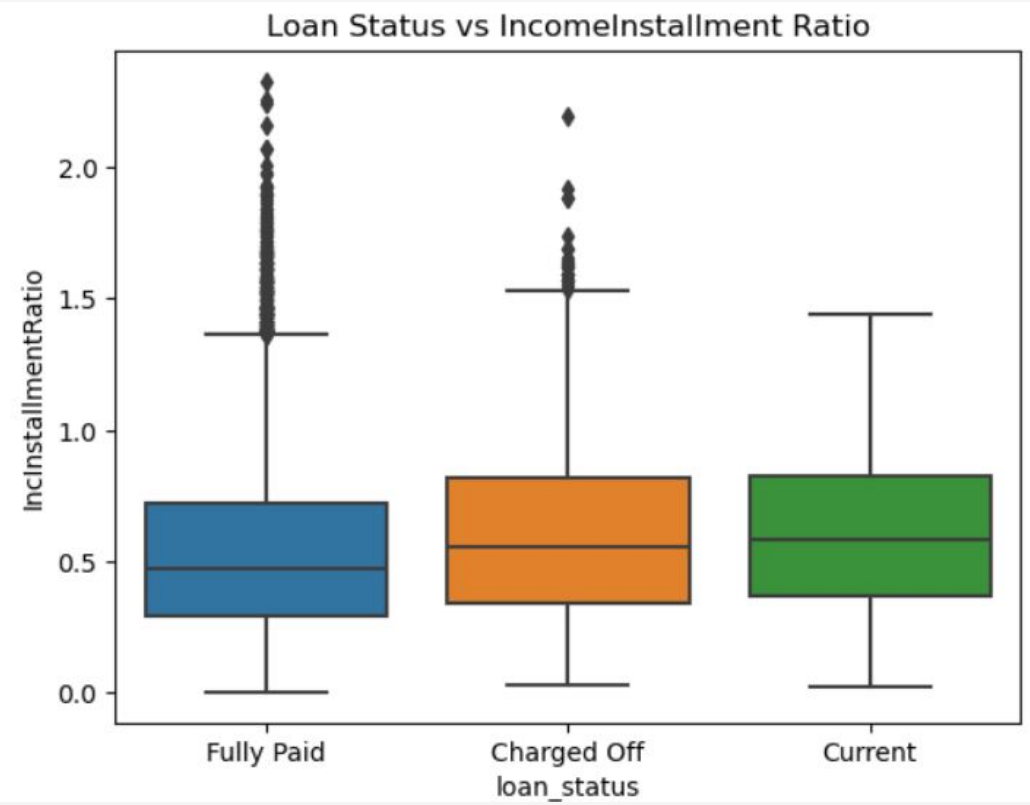


After plotting box plot using Loan status and installment we found 75th percentile of charged off customer is high which means high installment amount customer have chances of defaulters

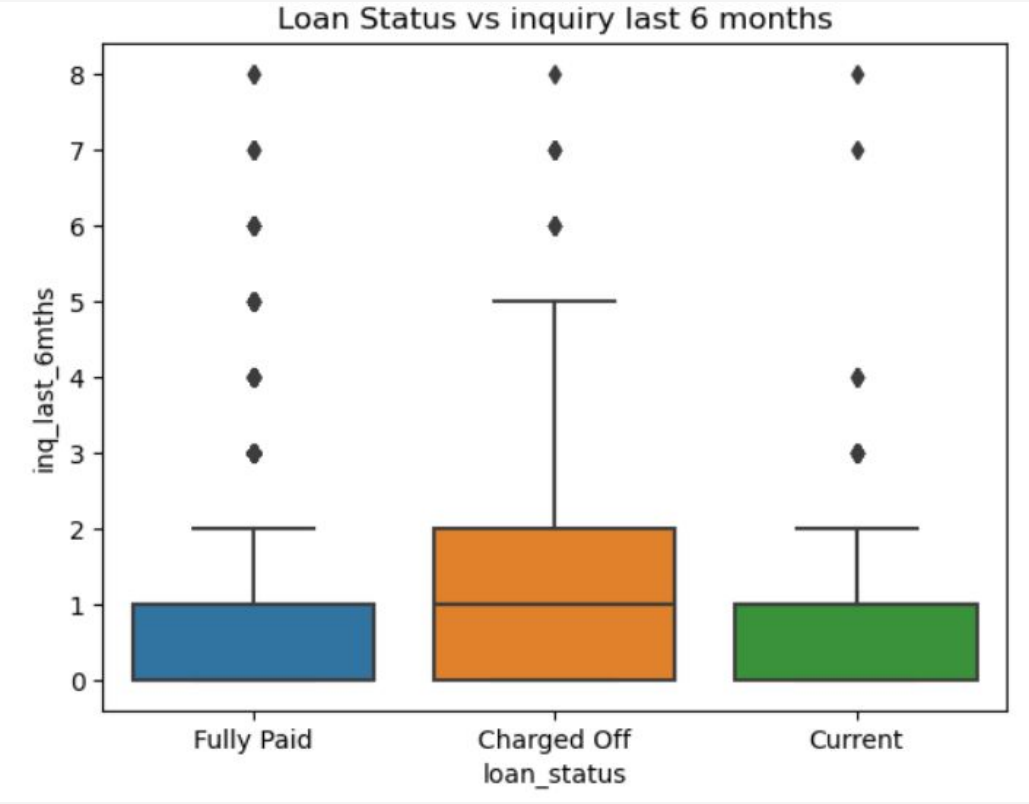


After plotting box plot using Loan status and Loan Amount we found 75th percentile of charged off customer is high which means high loan amount customer have chances of defaulters

Visualisation – **Bivariate Analysis**



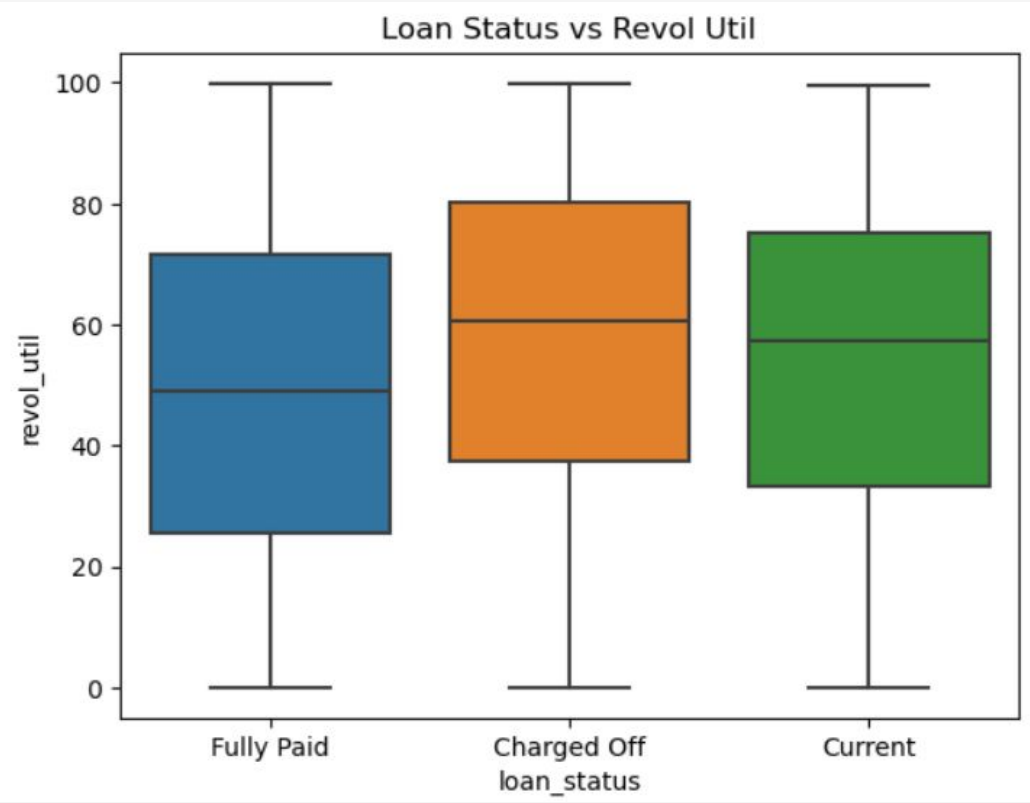
After plotting box plot using Loan status and Income installment ratio we found 25th and 75th percentile of charged off customer is high which means high loan amount customer have chances of defaulters



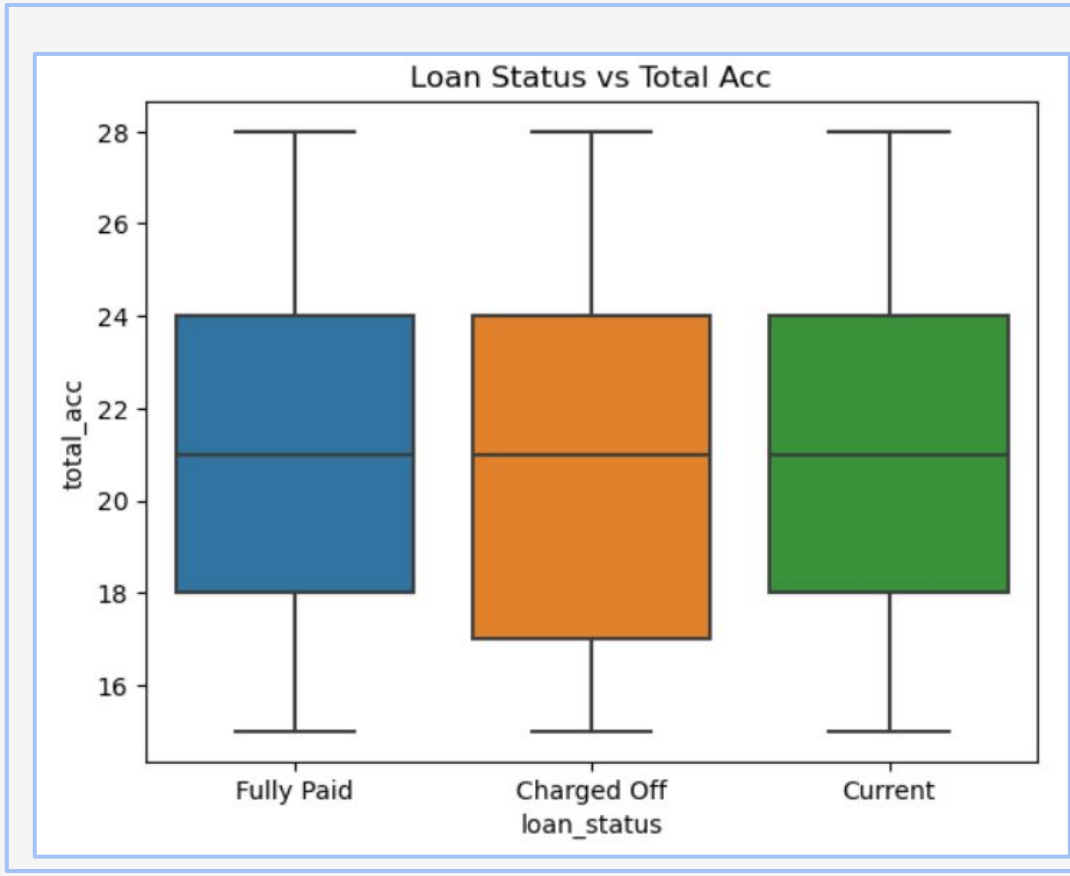
After plotting box plot using Loan status and last 6 months inquiry attribute we can analyse

- if customer enquiries are more than one in last 6 months then there is high chances of charged off

Visualisation – **Bivariate Analysis**

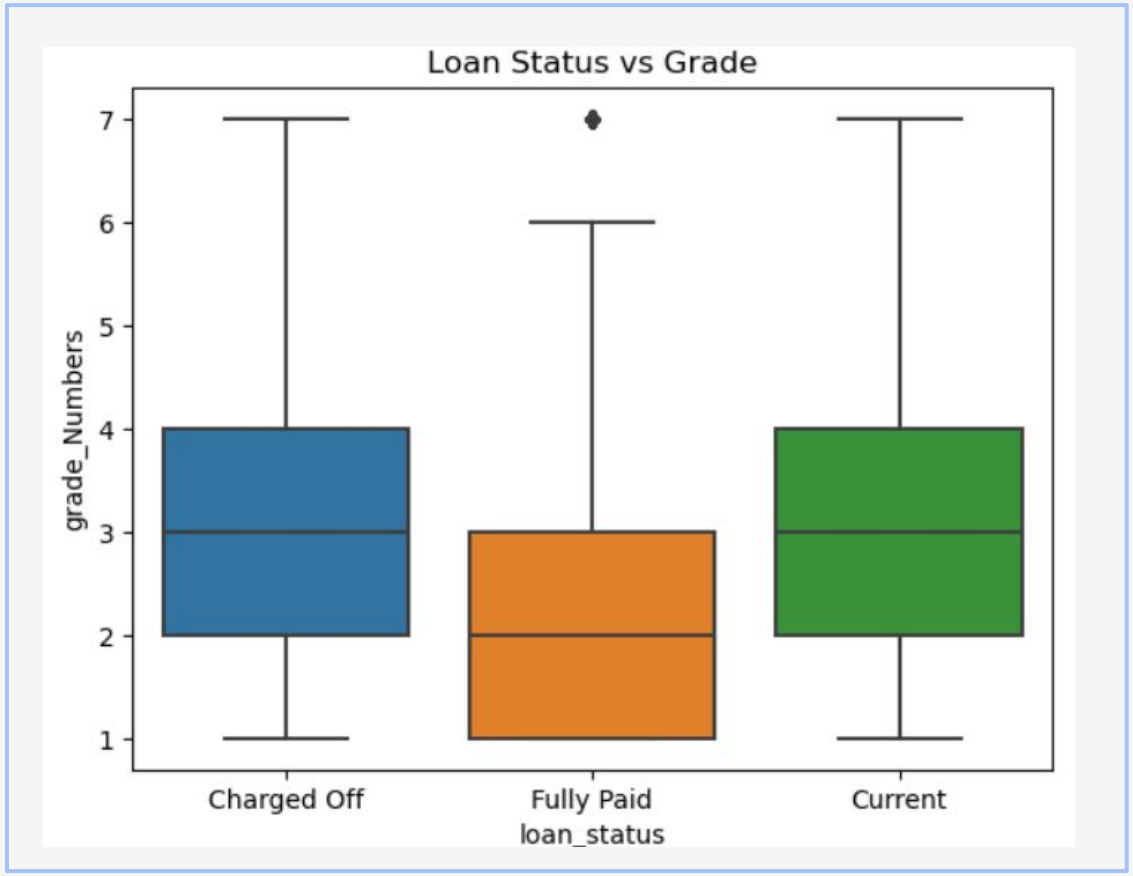


We plotted box plot using Loan status and revol_util and identified customer who are having more utilization rate compared to borrowed amount have high chances of charged off

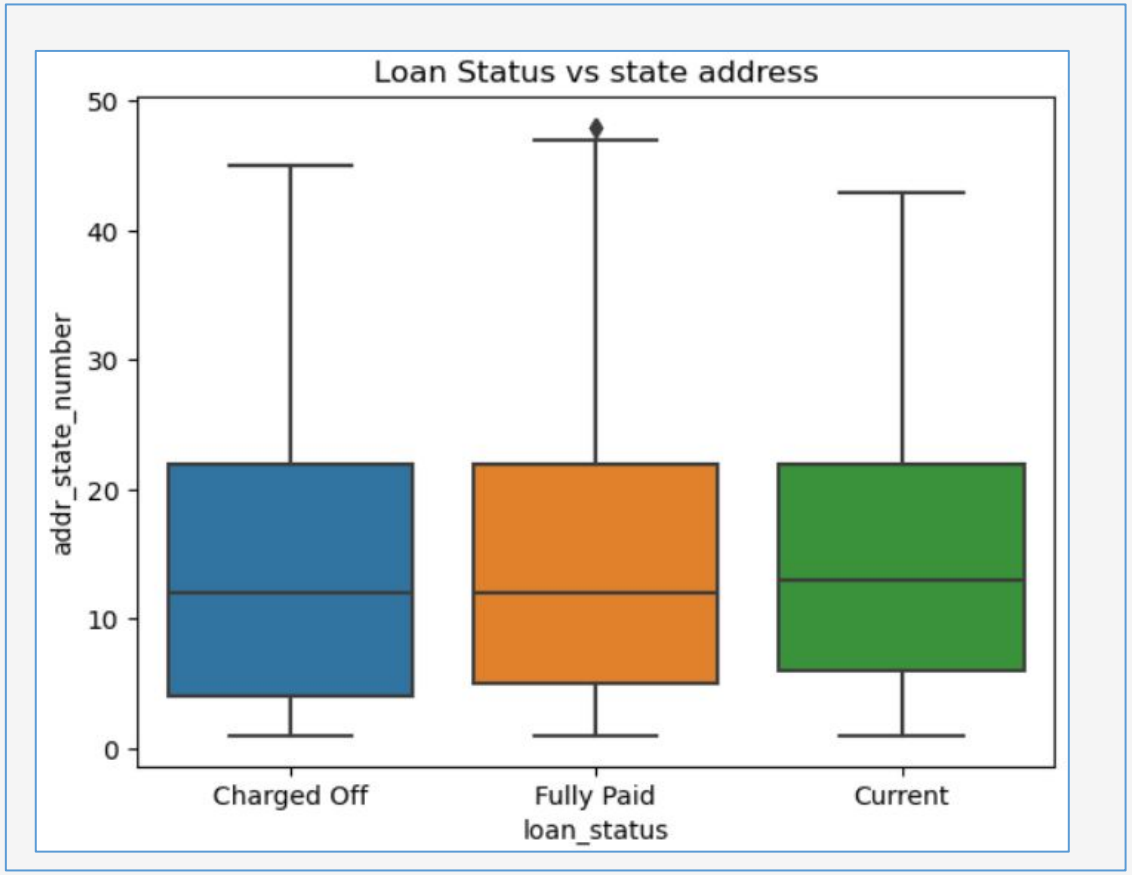


- Box plot of Loan status and total account shows plot is more or less same for both changed off and fully paid customers.
- From the graph we can conclude there is no significant direct impact of total account on charged off customers

Visualisation – **Bivariate Analysis**

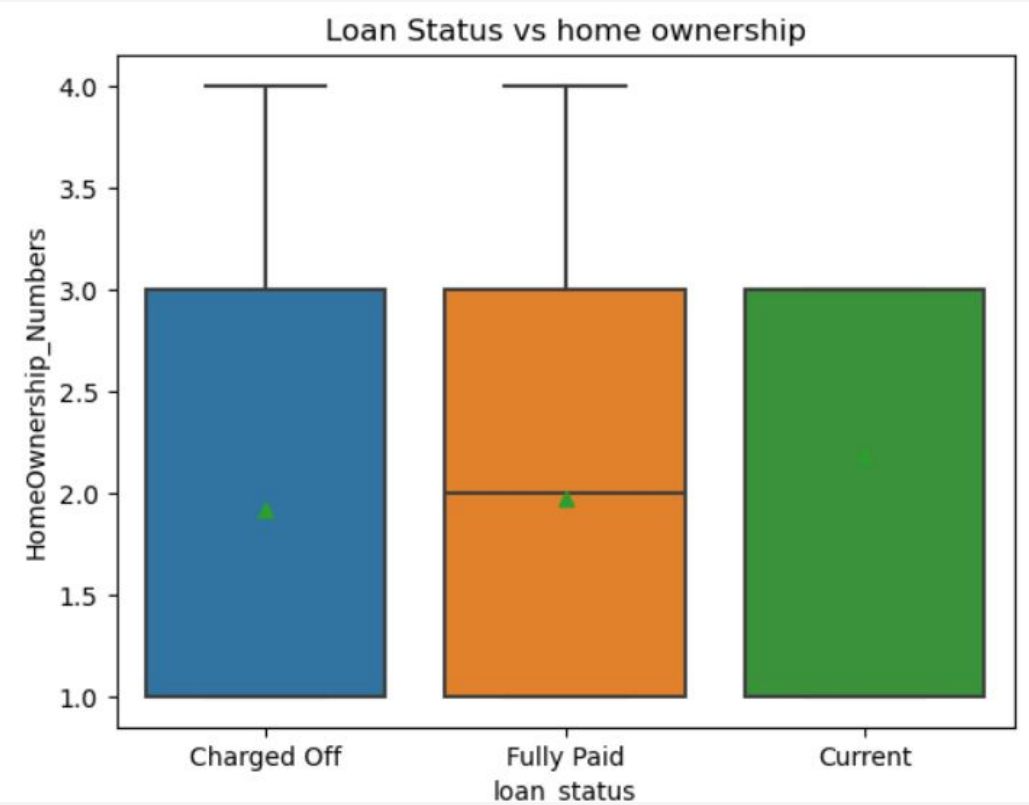


- We have created derived column based on grad ("A":1,"B":2,"C":3,"D":4,"E":5,"F":6,"G":7)
- Using Box plot of Loan status and grade we identified B, C & D grades are having high chances of defaulters

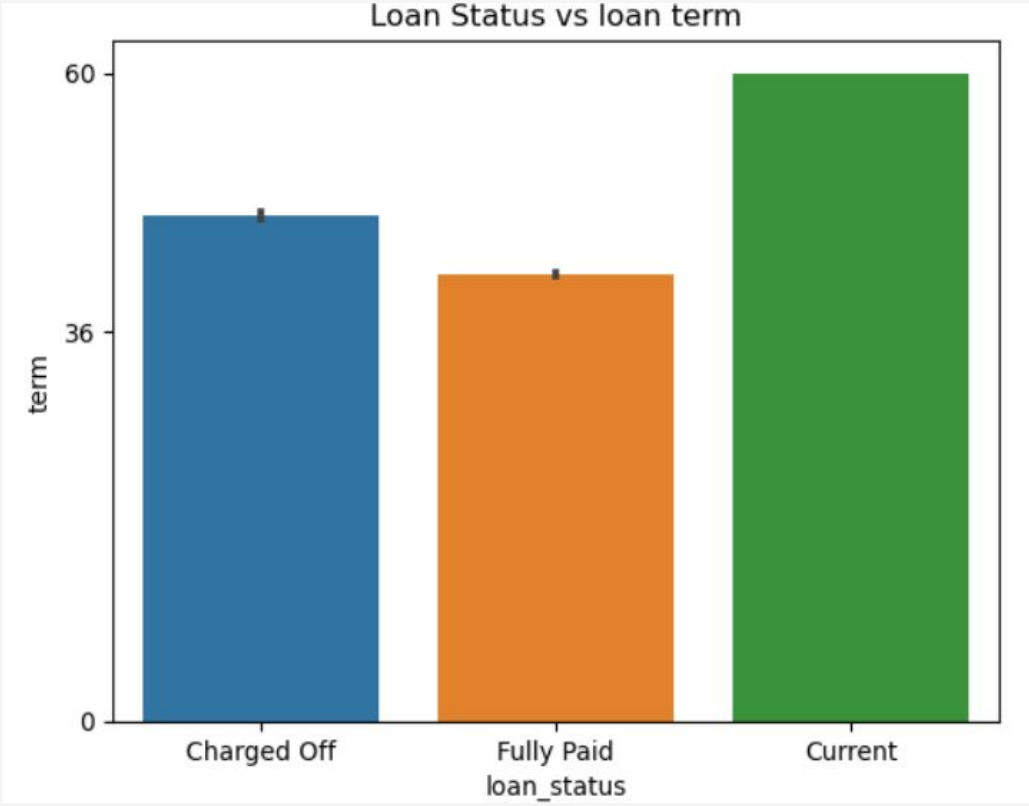


- We have created derived column based on states
- Box plot for both charged off and fully paid customers is same for loan status vs state address which indicates no direct relationship .

Visualisation – **Bivariate Analysis**



- We have created derive column based on home ownership ("RENT":1,"OWN":2,"MORTGAGE":3,"OTHER":4)
- From the graph we can conclude there is no significant direct impact of loan status and home ownership



- People who are taking loan for 60 months of tenure are likely to default.