

Lending Club Case Study

Analysis by

Sreevalli Karanam

Sandhya Girish

Problem Statement

Consumer finance company is specialized in lending various types of loan. Now, there is a need to make a decision for loan approval based on the applicant's profile.

There are 2 types of risks associated with 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

The study has to be done on past data of loan applicants and to find out the various patterns which will help in making the decisions regarding the applicants and take actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

When an applicant is eligible for a loan as per the bank's requirement, there can be 3 possible scenarios :

1. Fully Paid : Applicant has fully paid the loan
2. Current : Applicant is in the process of paying the instalments
3. 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

Now, the provided data should be analysed and find the factors which result into the category of 'Charged-Off' customers or defaulters. These factors will help the bank decide regarding the action taken on each applicant.

Overall approach of analysis

For Exploratory Data Analysis (EDA), it is very important that we should have the correct & concise data which can help us to do the quick analysis and to infer the meaning out of data.

For getting the correct data, it is good if the user analysing the data has a good business knowledge so that decision on certain fields can be taken easily.

For our analysis, we have taken the following steps :

1. Data Cleaning :

- Fixing Rows & Columns : The rows & columns are fixed by renaming the columns, deleting empty & unnecessary columns, splitting columns e.g date column for deriving Month & Year

In the landing club dataset, we have reduced the number of columns from 111 to 28 columns for final analysis

- Missing Values : Fixing rows for missing values
- Standardize Values : By removing outliers, converting % into numeric values for easy calculation, rounding the float variables
- Fixing invalid values
- Filtering data : Removing identical columns, filter by segment, filter by date, Picking only the relevant columns for analysis, Grouping of the data, categorizing the data in certain range

Overall approach of analysis

2. Univariate Analysis :

- Performing the analysis on unordered categorical variables such as 'Loan Purpose', 'Location', 'Address' etc.
- Performing the analysis on ordered categorical variables such as 'Installment', 'Interest Rate', 'Grade', 'Sub-Grade', 'Issued Year', 'Month' etc.
- Performing the analysis on quantitative variables such as 'Loan Amount', 'Funded Amount', 'Annual Income', 'Employee Experience' etc.

3. Segmented Univariate Analysis :

- Segmenting the Annual Income, Interest Rate, DTI, Loan Amount etc and do the univariate analysis

4. Bivariate Analysis :

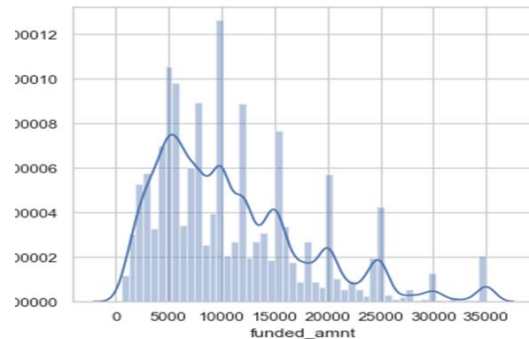
- Continuous Variables
 - If one increases other increases e.g. 'With increasing interest rate/funded amount/grade, Charged-off cases increases', 'With increased annual income, Charged-off cases decreases', etc.

5. Correlation Matrix/Mutivariate Analysis

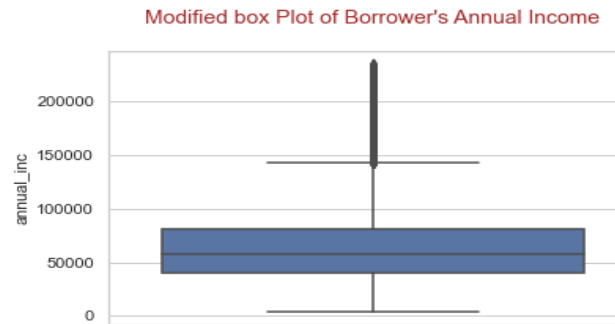
Summary/Conclusion from Univariate Analysis

The summary from the univariate analysis on the behaviour of people who takes majority of the loans is as follows :

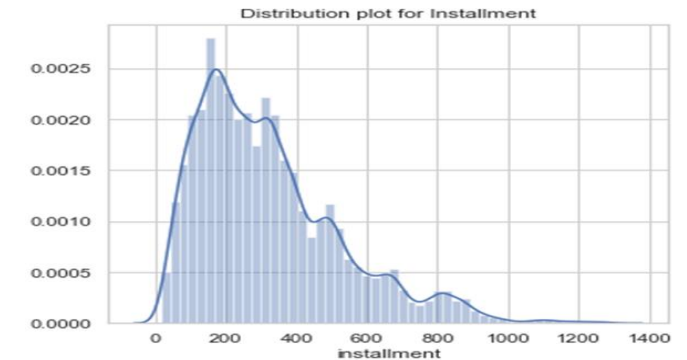
People taking loan amount
between 5000 to 15000



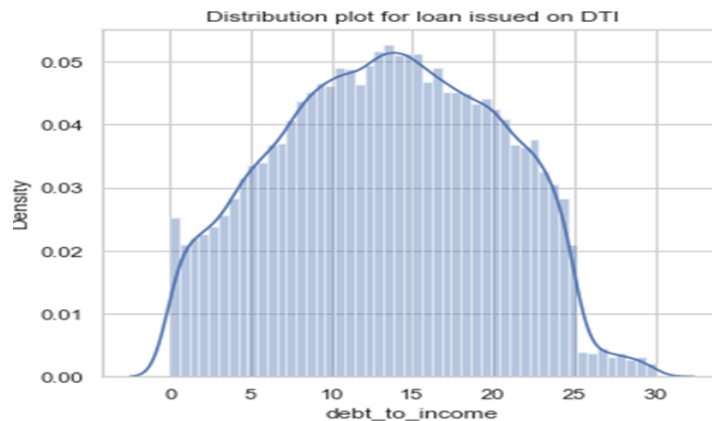
Annual Income lies between
40,000 to 81,000



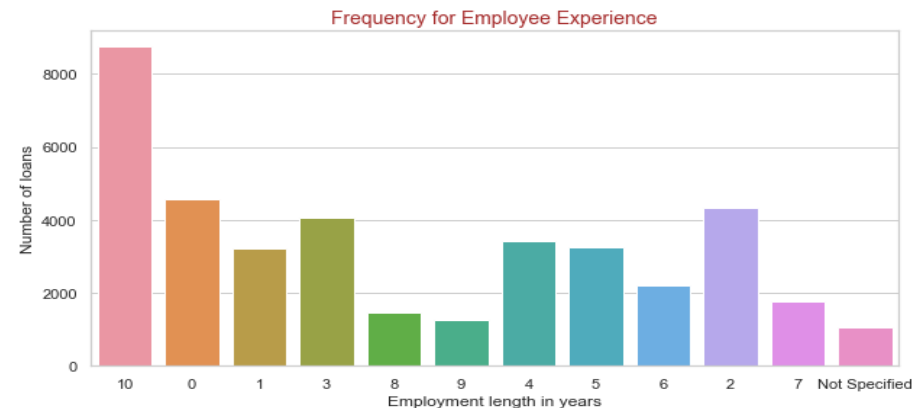
Installment lies between
150 to 450



Debt To Income lies
between 8 to 18



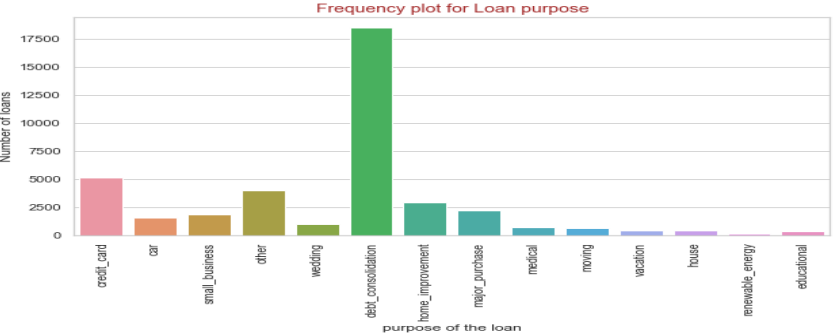
Experience is 10 Years or
more



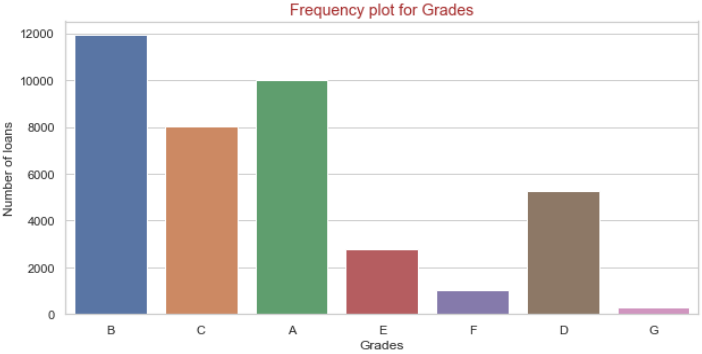
Summary/Conclusion from Univariate Analysis

The summary from the univariate analysis on the behaviour of people who takes majority of the loans is as follows :

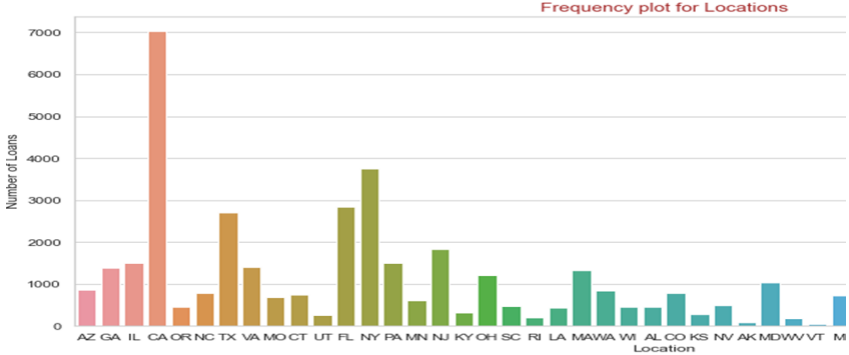
Purpose for taking loan is 'Debt Consolidation'



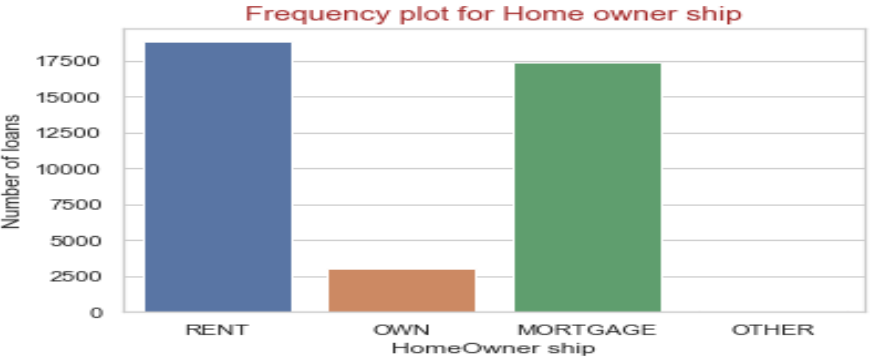
Applicants with Grade B & sub-grade B3



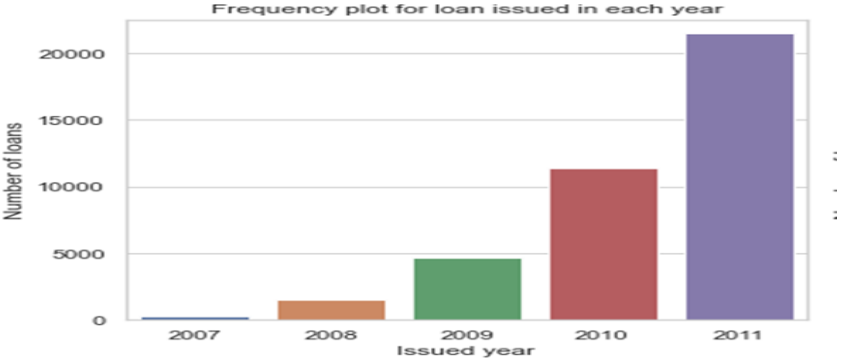
'CA' state has the top-most loan applicants



Loan with ownership as 'Rent' or 'Mortgage'



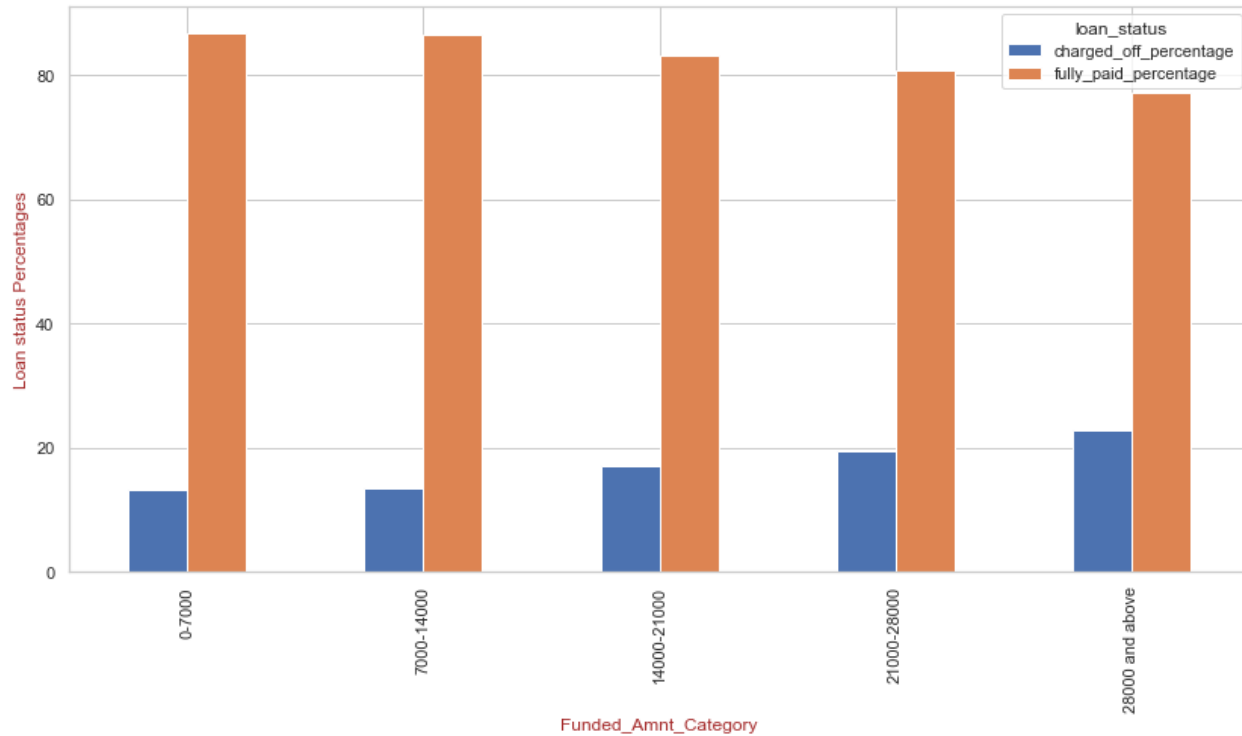
With each passing year, the number of loans is increasing



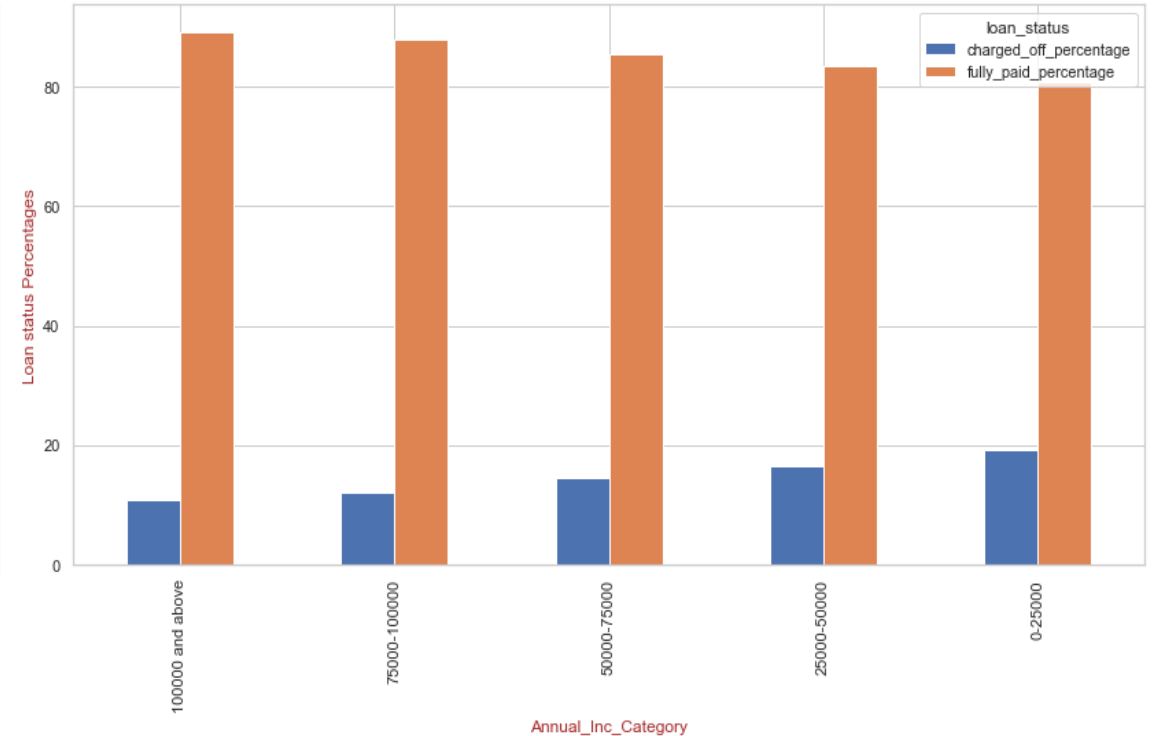
Results of Bivariate Analysis

There is a good insight when 'Charged-Off' value is compared to various fields from the dataset.

Funded_Amnt_Category versus Loan Status



Annual_Inc_Category versus Loan Status

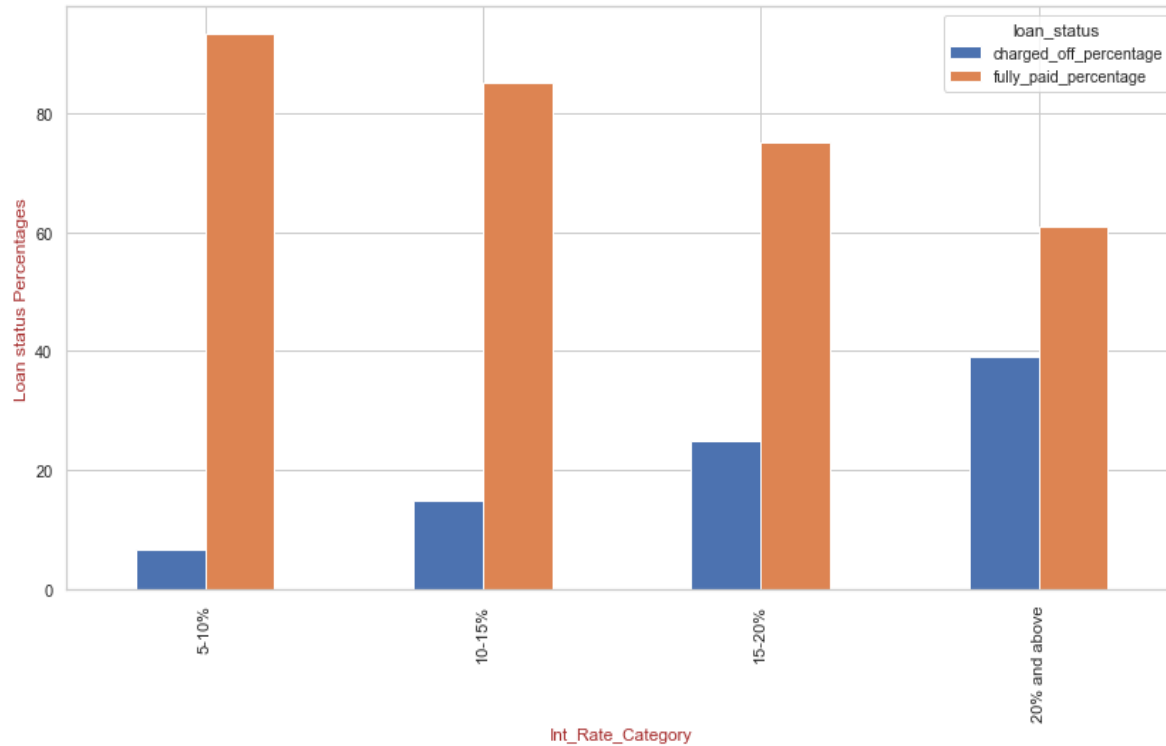


Conclusion :

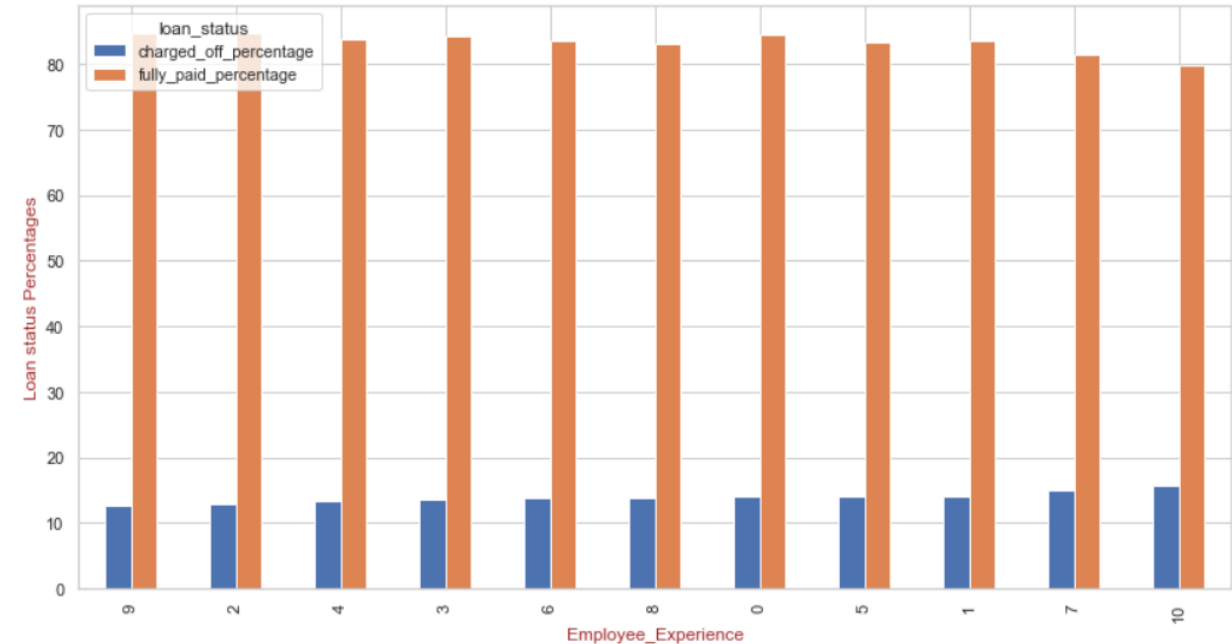
- More the funded amount for loan is, more number of defaulters are observed
- With lesser annual income of an applicant, more defaulters are observed. There is more risk with low income applicants

Results of Bivariate Analysis

Int_Rate_Category versus Loan Status



Employee_Experience versus Loan Status

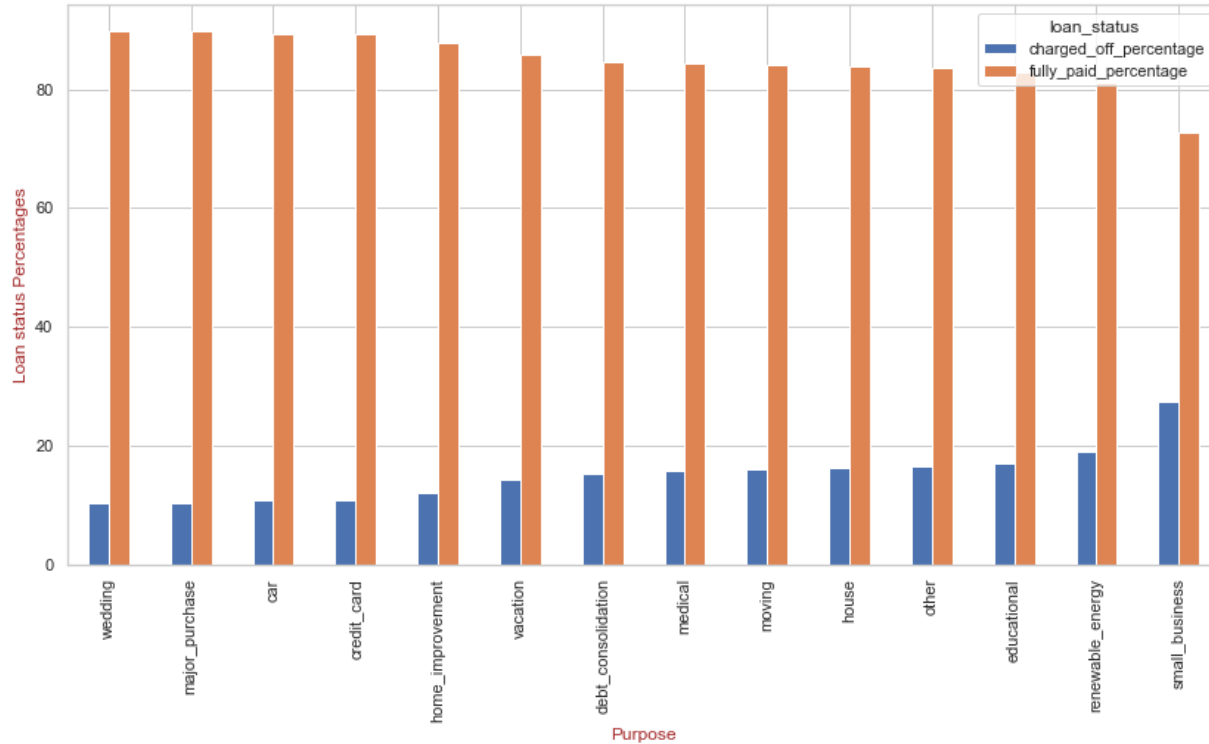


Conclusion :

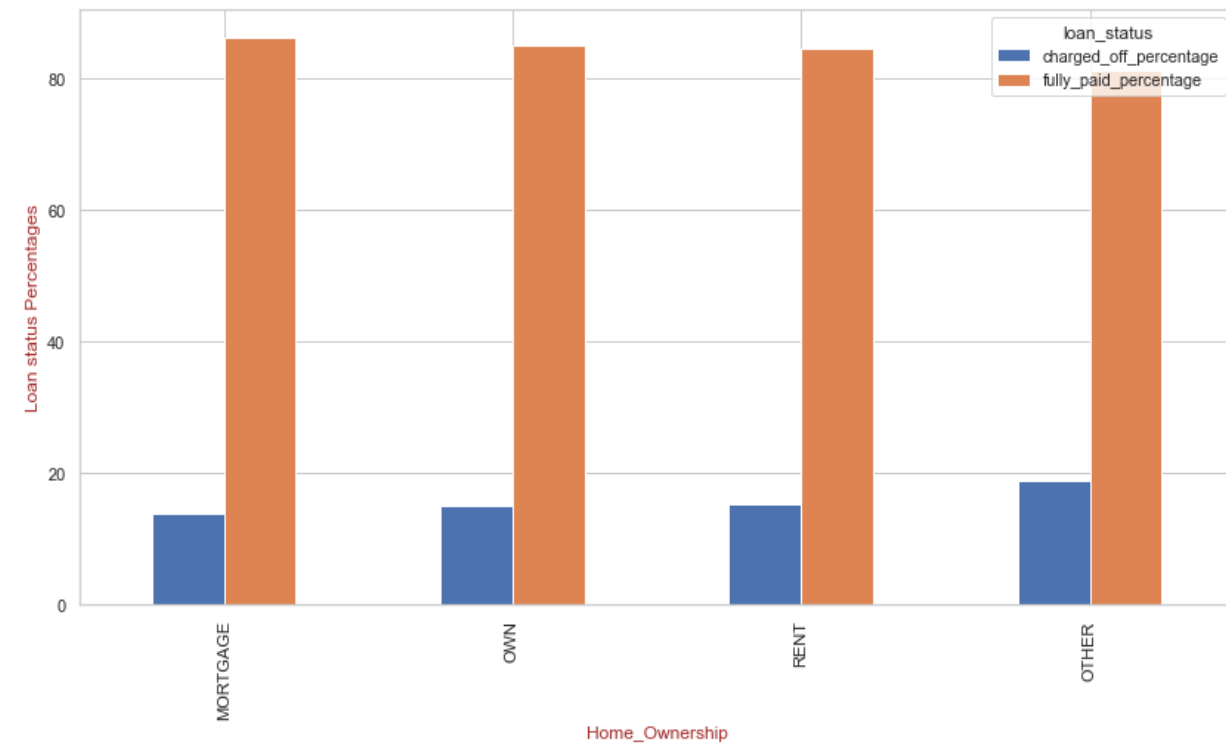
- With more interest rate for loan to an applicant, more defaulters are observed. There is more risk with high interest rate loans
- Employees with 10+ years of experience are likely to default slightly when compared with others

Results of Bivariate Analysis

Purpose versus Loan Status



Home_Ownership versus Loan Status

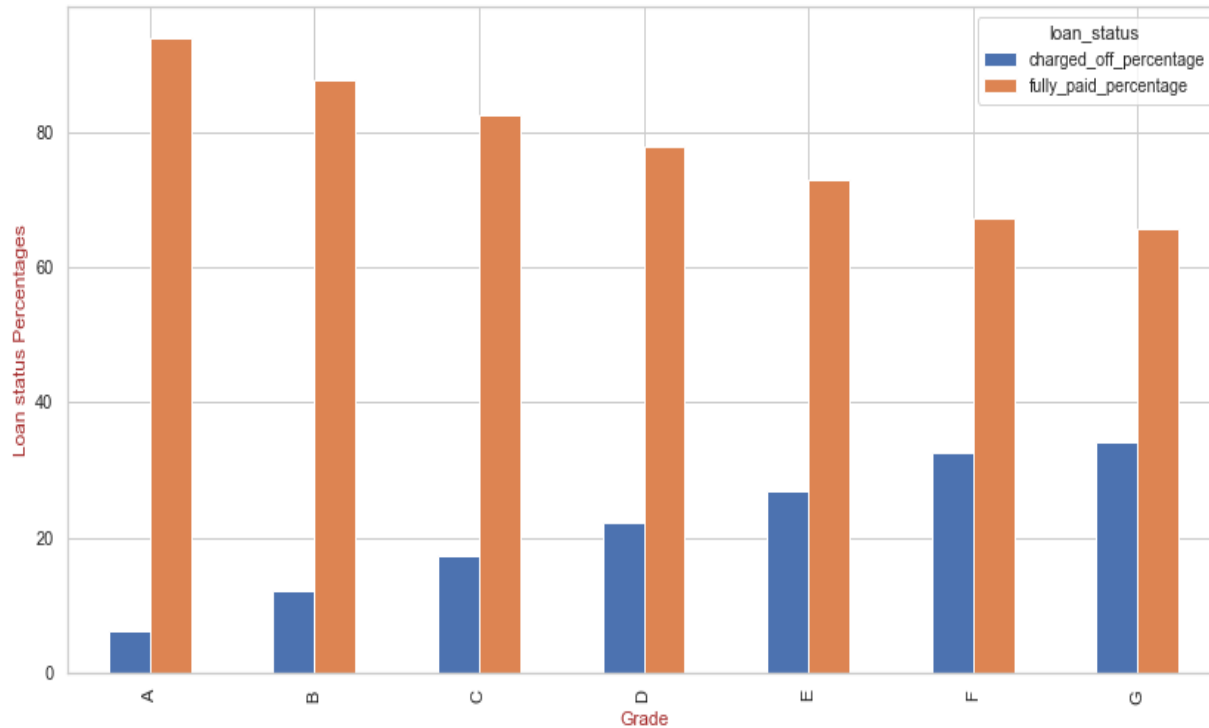


Conclusion :

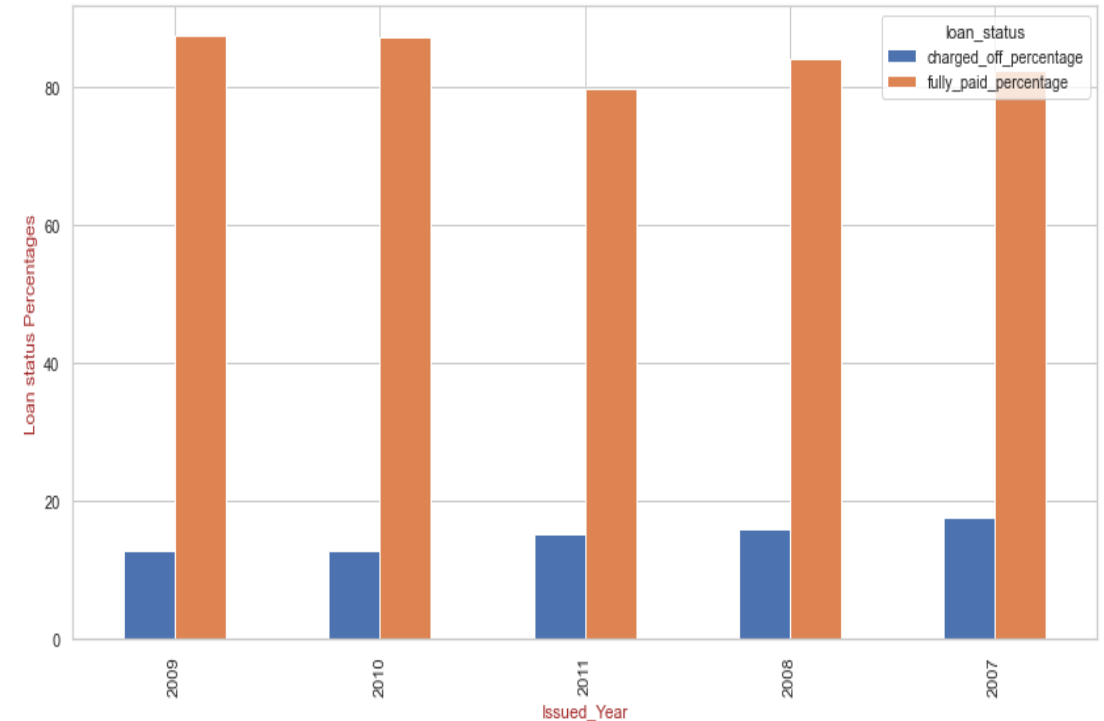
- It is observed that if the loan is taken with a purpose of 'Small Business' result in more number of defaulters. It is followed by 'Renewable_Energy', 'Educational' etc. purposes
- If the ownership category is 'Others', the number of defaulters are slightly more compared to other purposes for loan

Results of Bivariate Analysis

Grade versus Loan Status



Issued_Year versus Loan Status

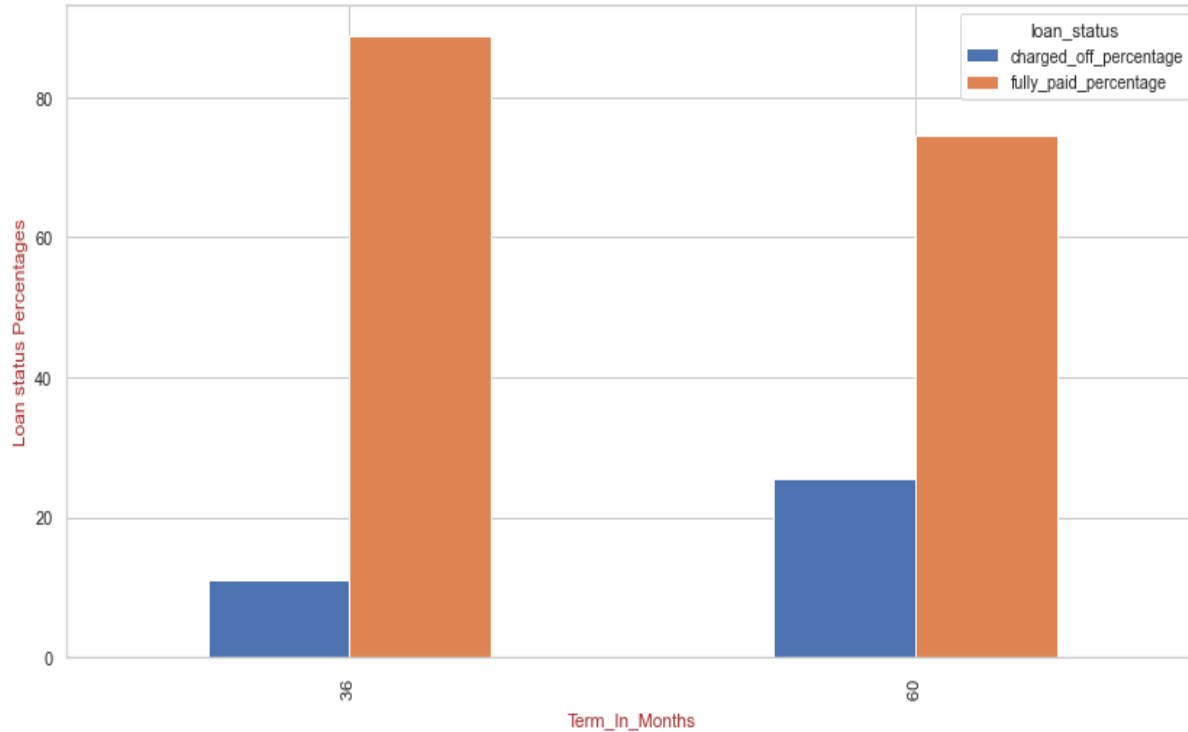


Conclusion :

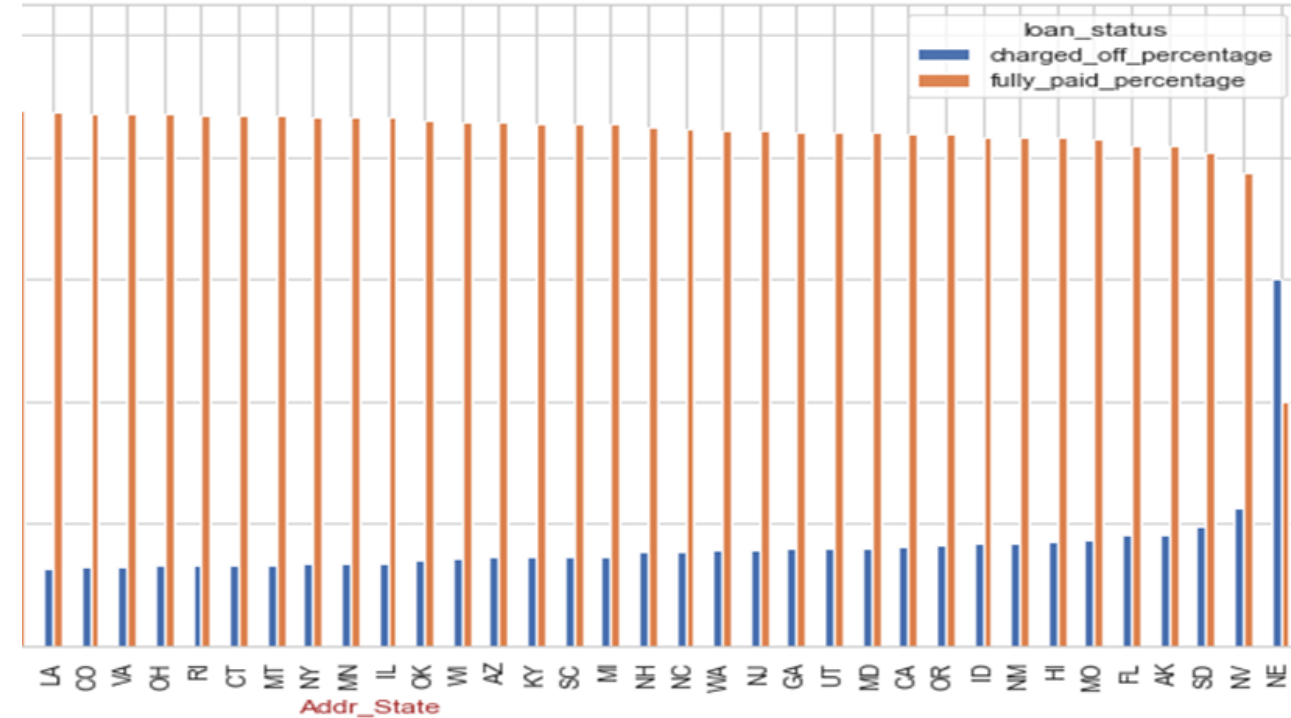
- A grade loans are less charged off whereas G grade loans are highly charged off. When compared to subgrade F5 has higher tendency of default but over all follows as order from A to G
- Charged off percentage slightly increases with year except the year 2007.

Results of Bivariate Analysis

Term_In_Months versus Loan Status



Addr_State versus Loan Status



Conclusion :

- When people take loan with less duration (36 months) are likely to pay complete loan in comparison to people who take loan with more duration (60 months)
- Charged off %'s are more in the following states NE, NV, SD, AK, FL But as per the data NE has only 3 values. So, cannot come to conclusion. But need to check the other states(NV,SD, AK, FL) while processing loan

Final summary/conclusion from the overall Analysis

Recommendations to Lending Club for approving loans to avoid defaults

- High loan amounts has more chances of charged off. Should properly analyze the applicants demanding higher loan i.e; more than 28K+
- More conscious about Annual income. Applicants having annual income range 0-50k has more than 15% chances of default.
- Should consider grades. Applicants with lower grades has high chances to default. In Charged off percentages $A < B < C < D < E < F < G$ (A has lower chances where as G has higher chances for charged off)
- Need to calculate interest rate correctly as higher interest rate has high chances to default. More than 20% of applicants are charged off when the interest rate is 15% or more.
- Loans applied for purpose as Small Business needs to be scrutinised carefully.
- Should accept more loan with term 36months rather than 60months.
- Customer with public bankruptcy history of 1 or more is more likely to get default.