Lending Club Case Study

Using EDA (Exploratory Data Analysis)

By - Seshu Kumar Bangaru

Problem Statement

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

The data given contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

The analysis has to be performed using EDA (Exploratory Data Analysis)

Steps Involving EDA

- Load the data
- Check meta data / sanity check
- Data Cleaning
- Missing value imputation & outlier treatment
- EDA
  - Univariate
  - Segmented Univariated
  - Bivariate

Data Sanity, Cleaning and Imputation

1. Load the loan.csv file
2. Remove all the columns with NAs
3. Columns emp_title, emp_length, desc, title, mths_since_last_delinq,mths_since_last_record,revol_util,last_pymnt_d,next_pymnt_d,last_credit_pull_d,collections_12_mths_ex_med,chargeoff_within_12_mths,pub_rec_bankruptcies,tax_liens
4. Based on the count of NA in each column we can drop following columns desc, mths_since_last_delinq,mths_since_last_record,next_pymnt_d
5. Check if we can fill in the NA values for other columns.
6. We can see that there are 39678 of 0.0 in tax_liens column. So filling the other rows does not give us any advantage. So drop this column
7. Assign pub_rec_bankruptcies with 0.0 for all the pub_rec==0
8. Removing the 20 records which has pub_rec_bankruptcies=NA
9. Removing records which have emp_length as NA
10. Dropping the emp_title as we have large list of unique values
11. Drop the title column as we already have purpose column
12. Drop all the rows for revol_util is NA
13. Drop rows with last_credit_pull_d is NA
14. Drop rows with last_pymnt_d is NA
15. Drop columns url, desc

Cleaning Continued...

Cleaning the data

1. term - remove the "months" string
2. int_rate - remove % character

Univariate Analysis

Analyse few columns individually Segregated into

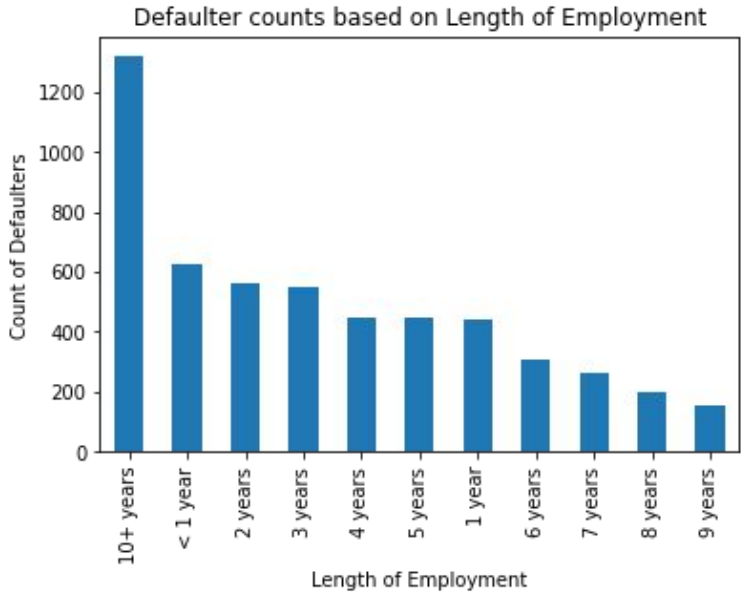
- Categorical
  Ordered - emp_length,issue_d
  Unordered - home_ownership,purpose,application_type
- Quantitative/Numeric - annual_inc, loan_amnt

We have three sub-data sets

- Data without columns which have all NAs/NULLs and not in “Current” loan status
- Data with “Charged Off” Loan Status
- Data with “Fully Paid” Loan Status

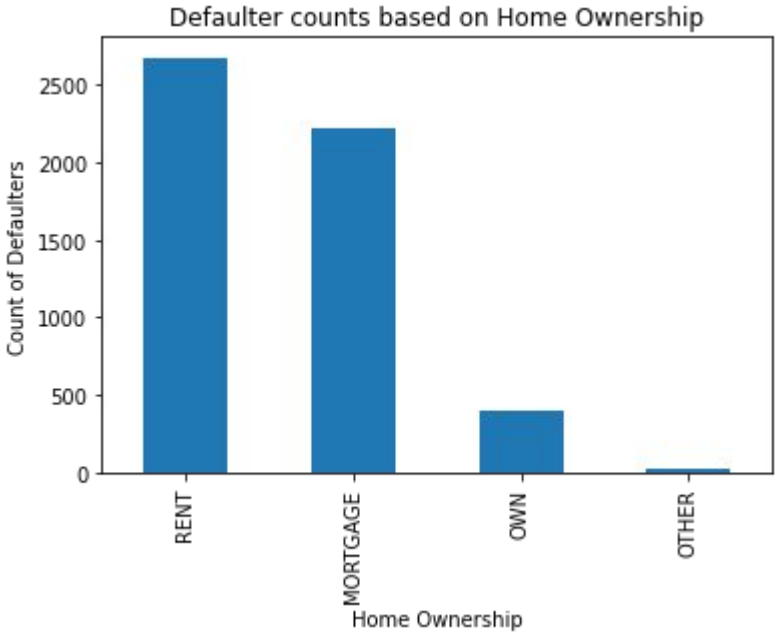
Analysis for emp_length column

The bar chart clearly shows that "10+ Years" has the maximum defaulters. Interesting to see Applicants with 6-9 years are less to default



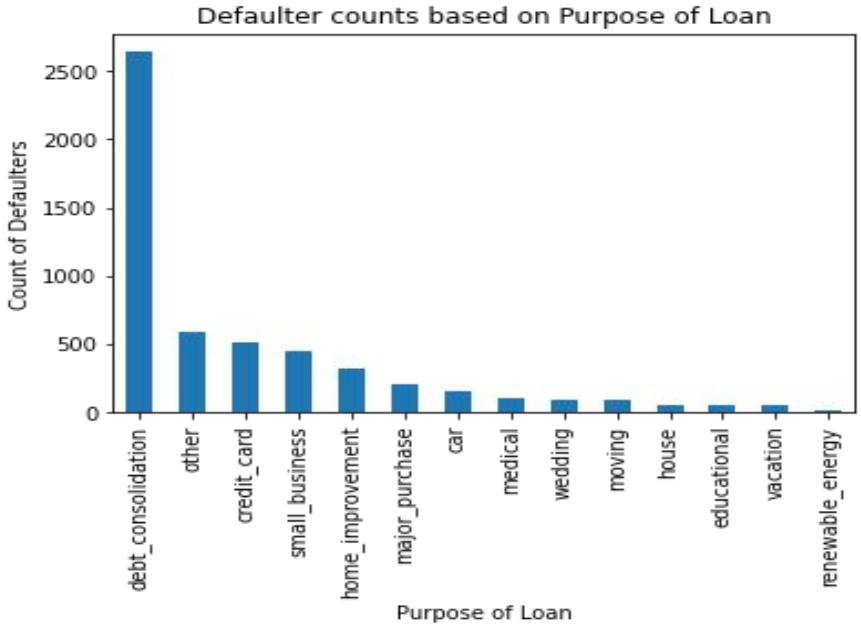
Analysis for home_ownership column

The bar chart shows Applicants who is either in Rented House or with Mortgage are like to default



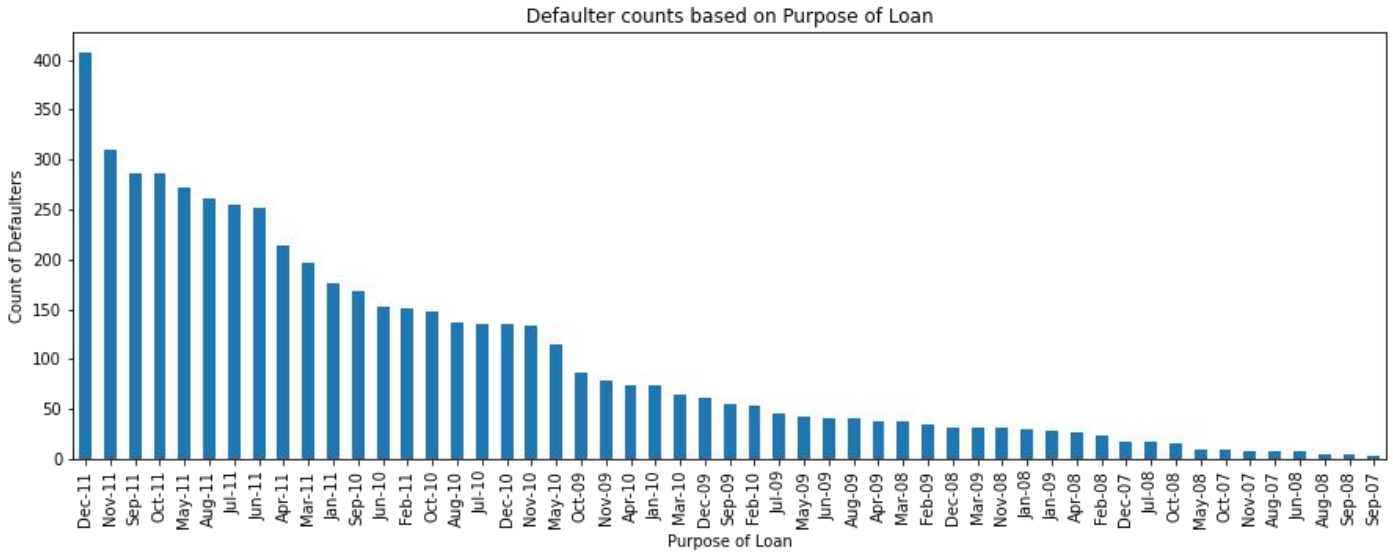
Analysis for purpose column

The bar chart shows that loan with "debt_consolidation" category likely to get defaulted



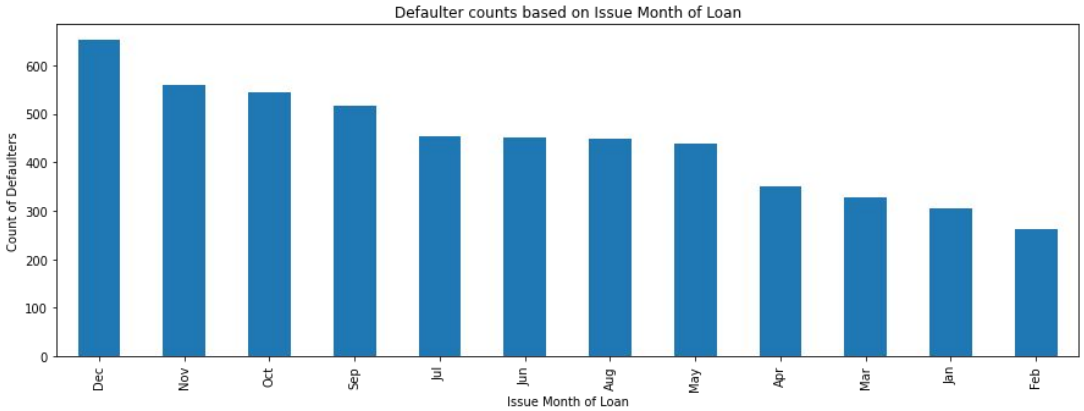
Analysis for issue_d column

In the chart we can see that Dec-11 is the issue month maximum default loans belong to. Lets see the trend of the count based on the Month



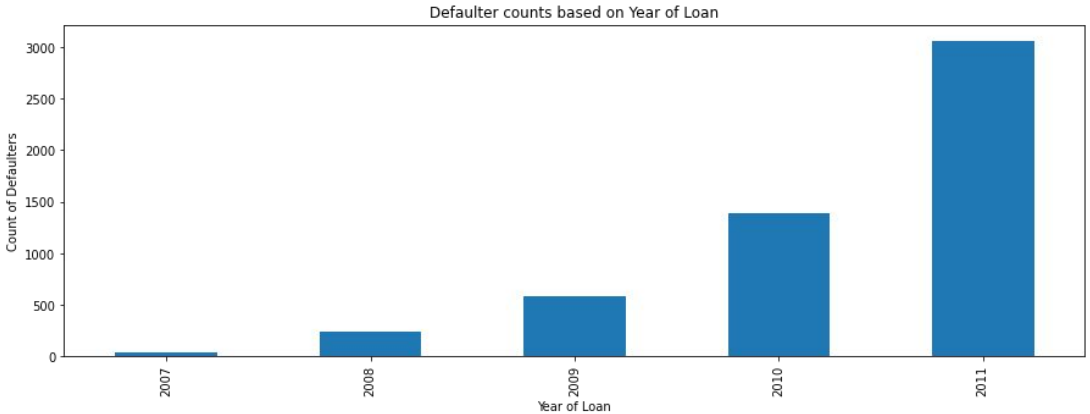
Extracted Month from issue_d column

This chart shows that loans which are issued in Dec are higher when compared to other months



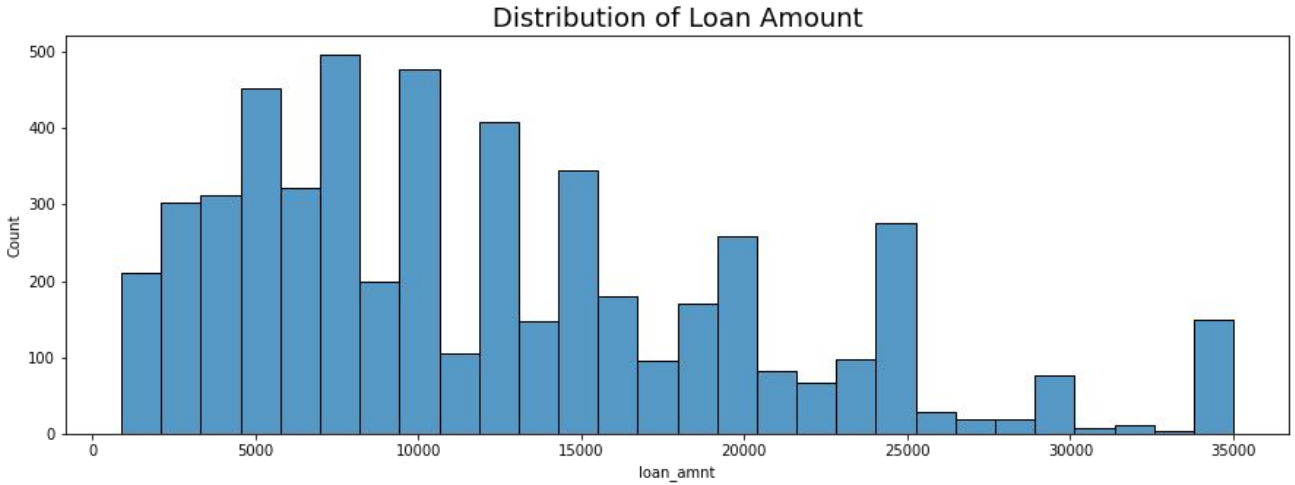
Extracted Year from issue_d column

In the chart we could see that count of defaulters have increased over the years.



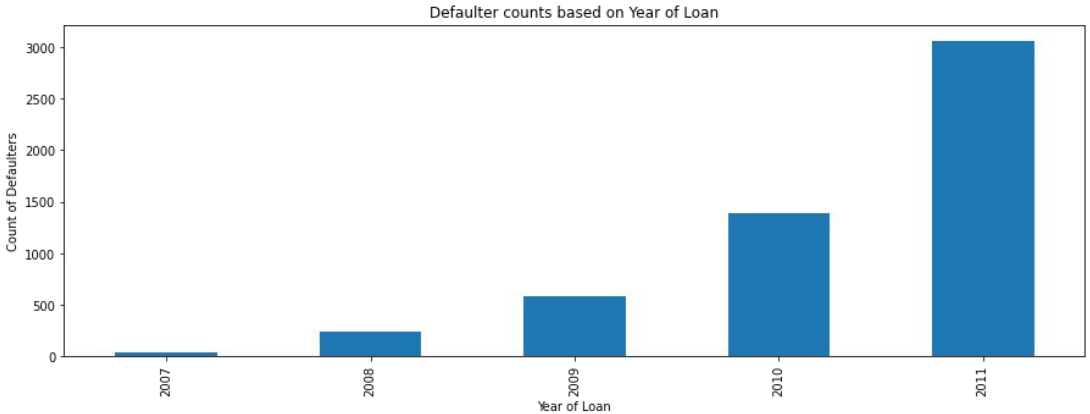
Distribution of Loan Amount

In the chart we can could see that density is at from 900 to 25000. So defaults are more at this range



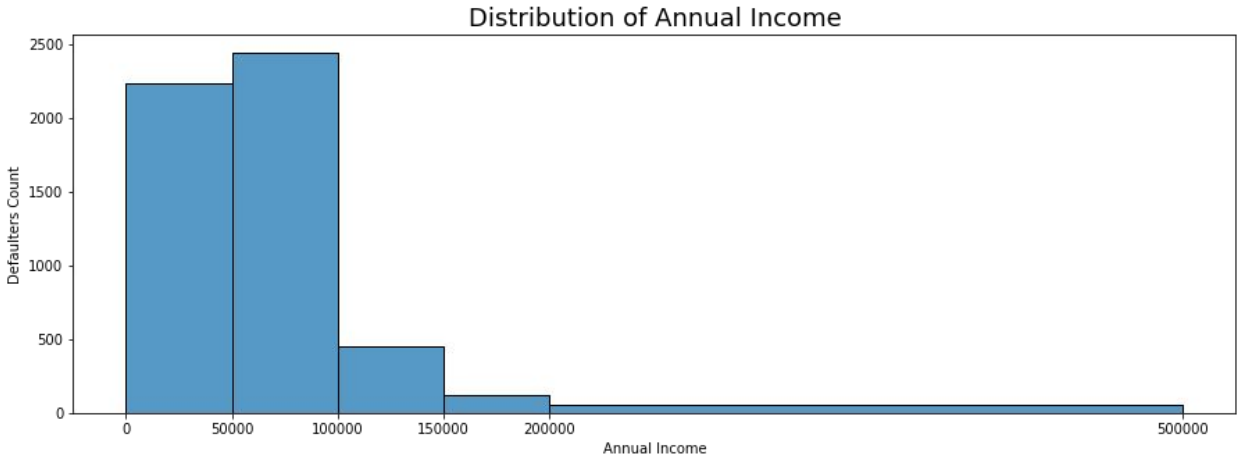
Analysis for issue_d column

In the chart we could see that count of defaulters have increased over the years.



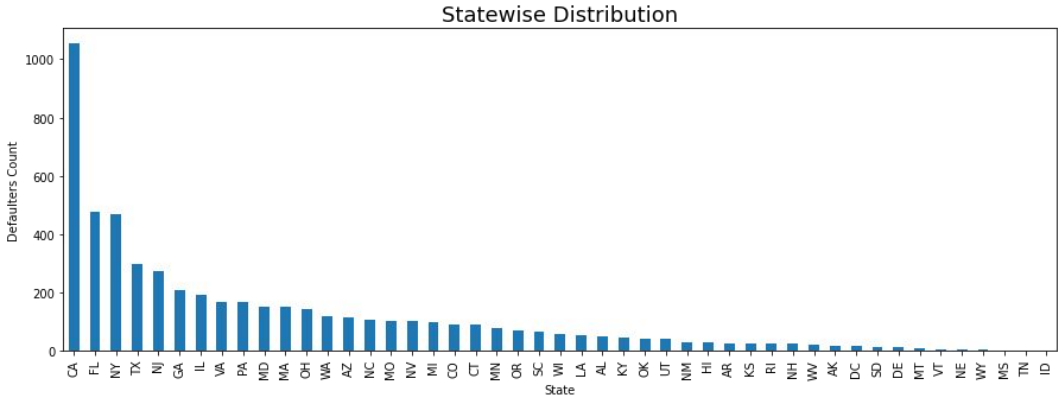
Distribution of Annual Income

Trend shows that majority of the defaulters have Annual income < 100000



Distribution of Statewise

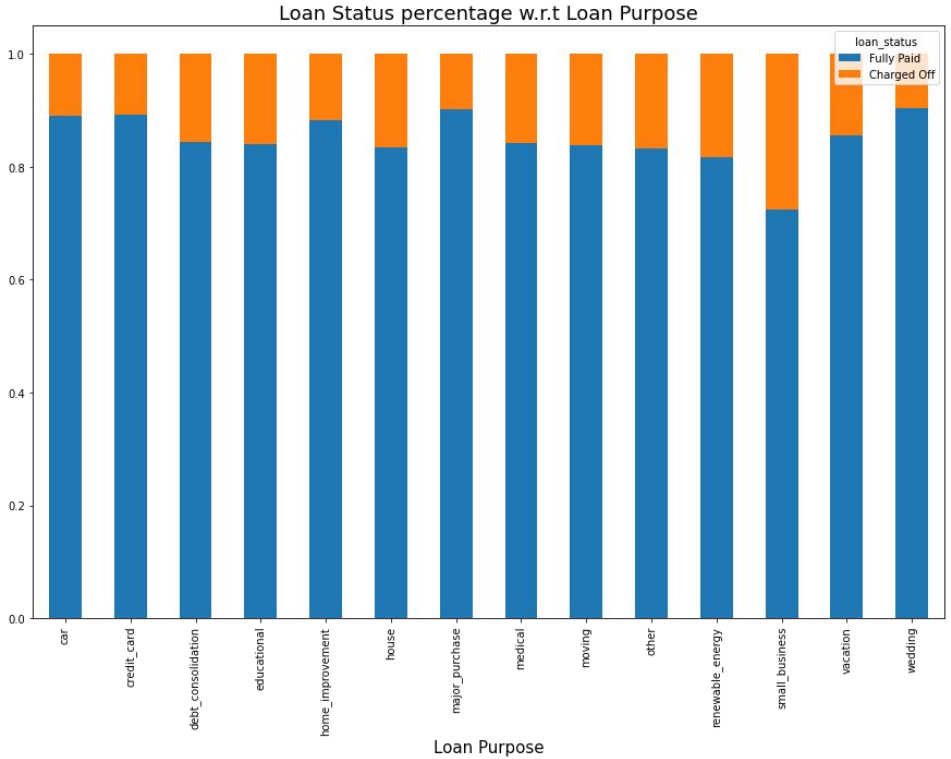
This will give us an idea that majority of the applicants come from CA sate



Segmented Univariate Analysis

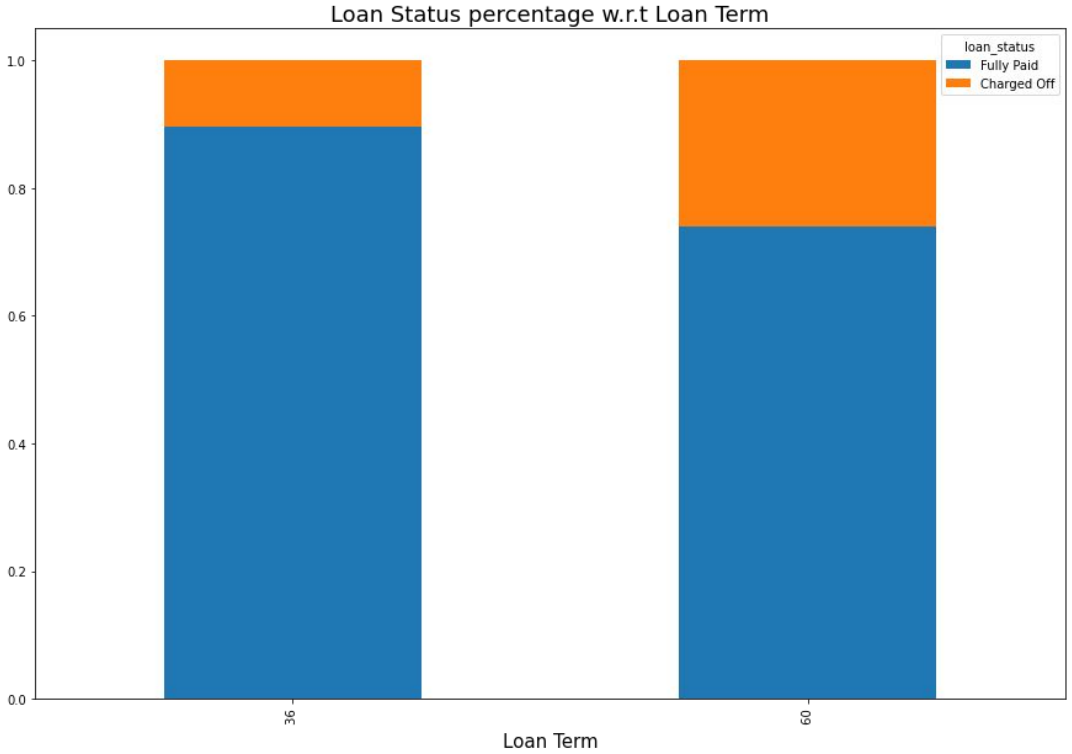
Loan Status percentage w.r.t Loan Purpose

We can see that "small_business" has maximum "Charged Off". Which should have major focus while giving loans



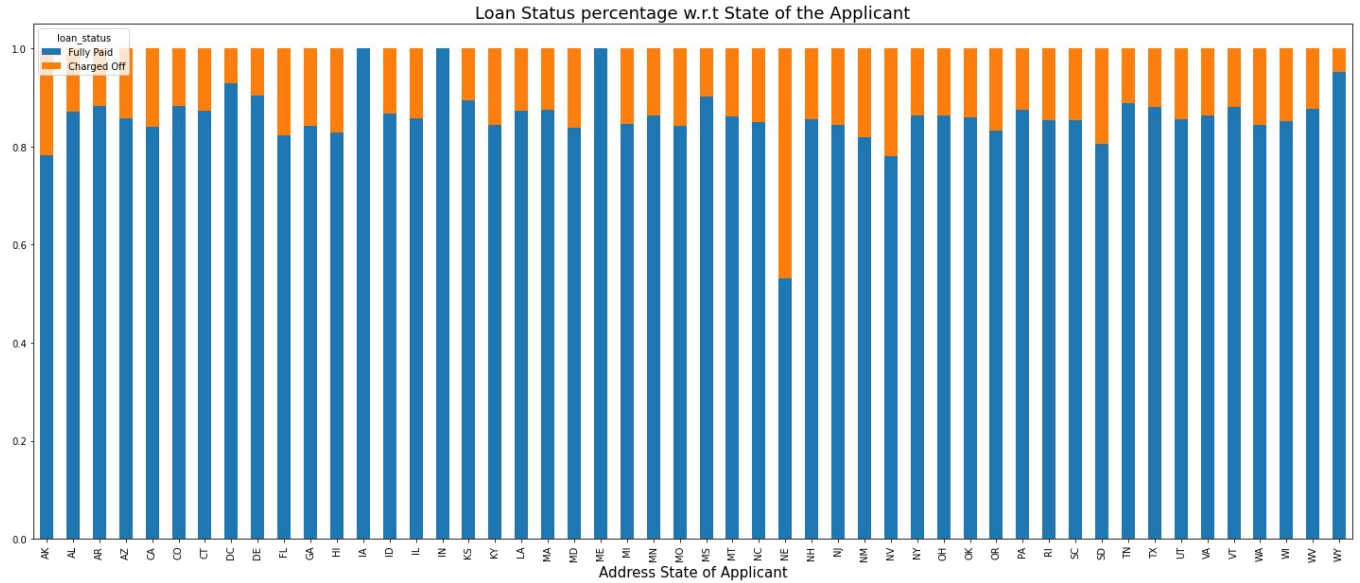
Loan Status percentage w.r.t Loan Term

This shows that loan with 60 terms have maximum "Charged Off"



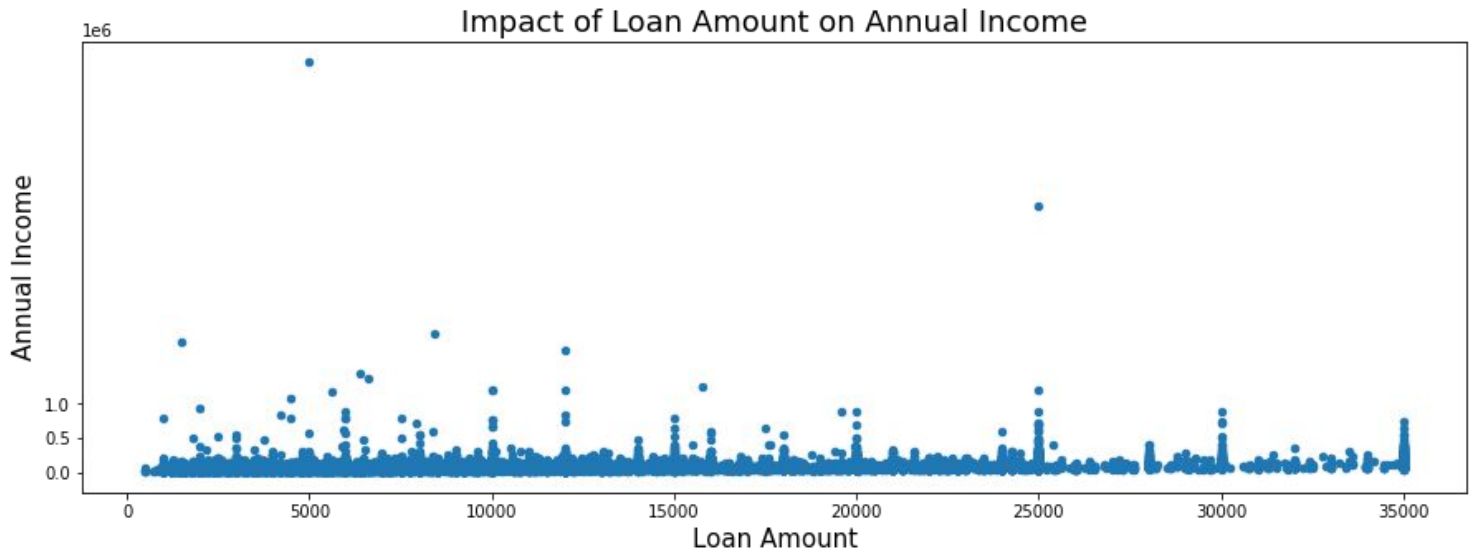
Loan Status percentage w.r.t State of the Applicant

- All Applicants from states "IA"/"IN"/"ME", "Fully Paid"
- Maximum no of "Charged Off" is from "NE"

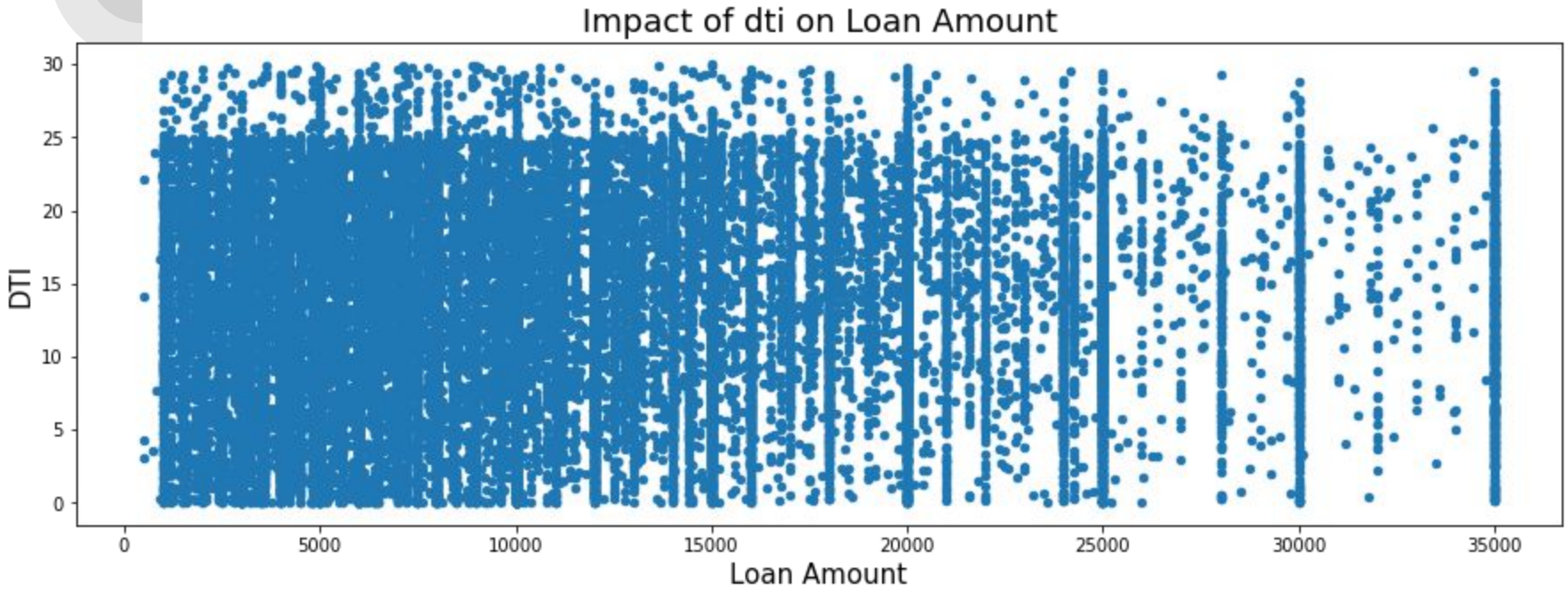


Bivariate Analysis

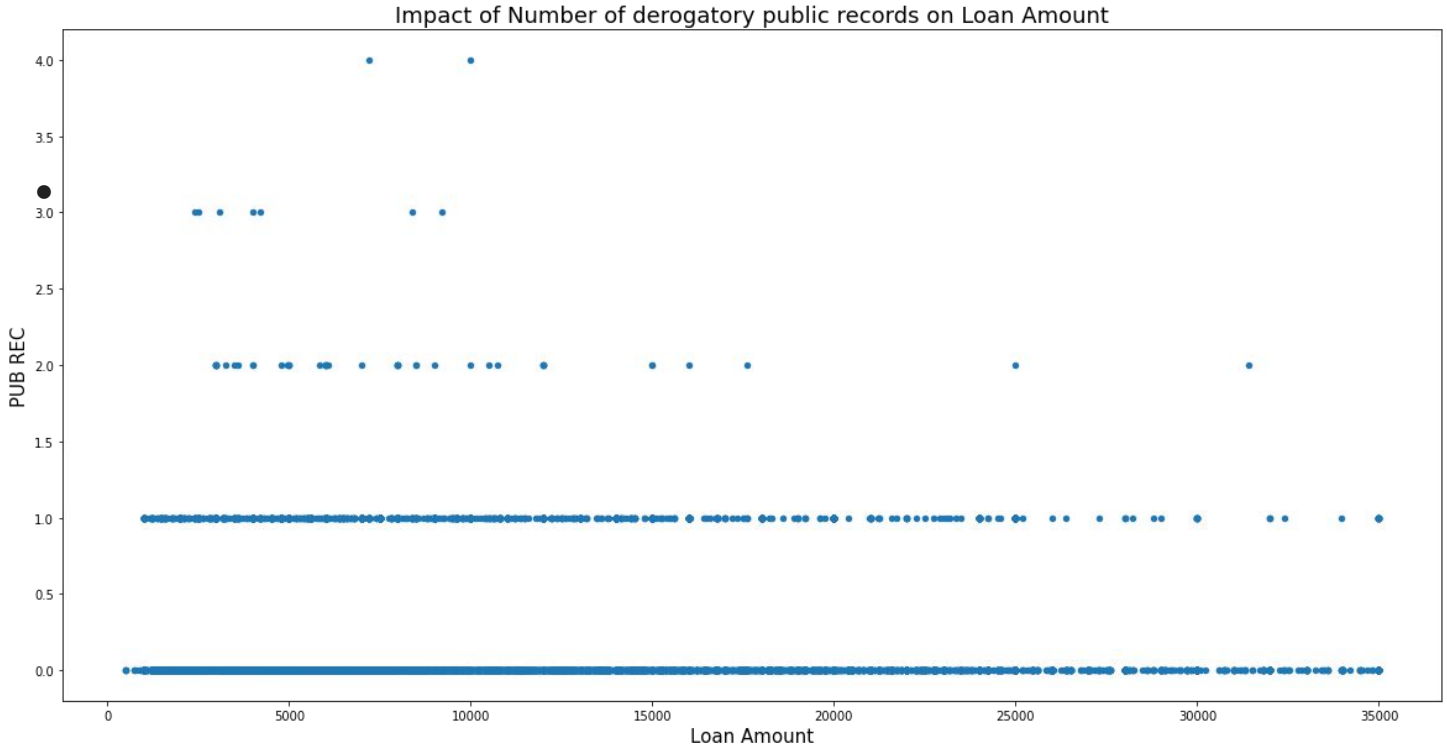
Impact of Loan Amount on Annual Income



Impact of dti on Loan Amount



Impact of Number of derogatory public records on Loan Amount



Conclusion

1. Applicants who is either in Rented House or with Mortgage are like to default. So we can provide the loan with higher rate for these kind of Applicants
2. Small Business Applicants have to be more scrutinized
3. Loan Defaults have been increasing over the years from 2007 - 2011
4. Loans with Higher interest Rates can be provided for the states which have higher default rates