Lending Club Case Study

SUBMISSION

Group Name:

1. Animesh Kalita
2. Roopak Bhardwaj

Overview of what lending club is and how it works:

Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to fund them. The lending club acts as an intermediator who screens the borrowers by assigning a grading score to each applicants using credit and income data that determines the interest rates the borrowers are qualified for, which in turn helps the investors to decide whether to fund their loan or not.

Some of the key purposes of the borrowers for applying the loans are debt consolidation, credit card, improve their homes and major purchases.

Problem Statement:

There are two types of risk associated with a lending company when they decide to give a loan to an applicant.

- 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 financial loss for the company

Business Objective:

The lending company wants to understand the driving factors behind loan default and then utilize this knowledge for its portfolio and risk assessment.

Problem solving methodology – Flow Chart

Business and data understanding:

Understand the business requirement of the problem. Perform basic statistics on the data to know the variables.

Data Cleaning:

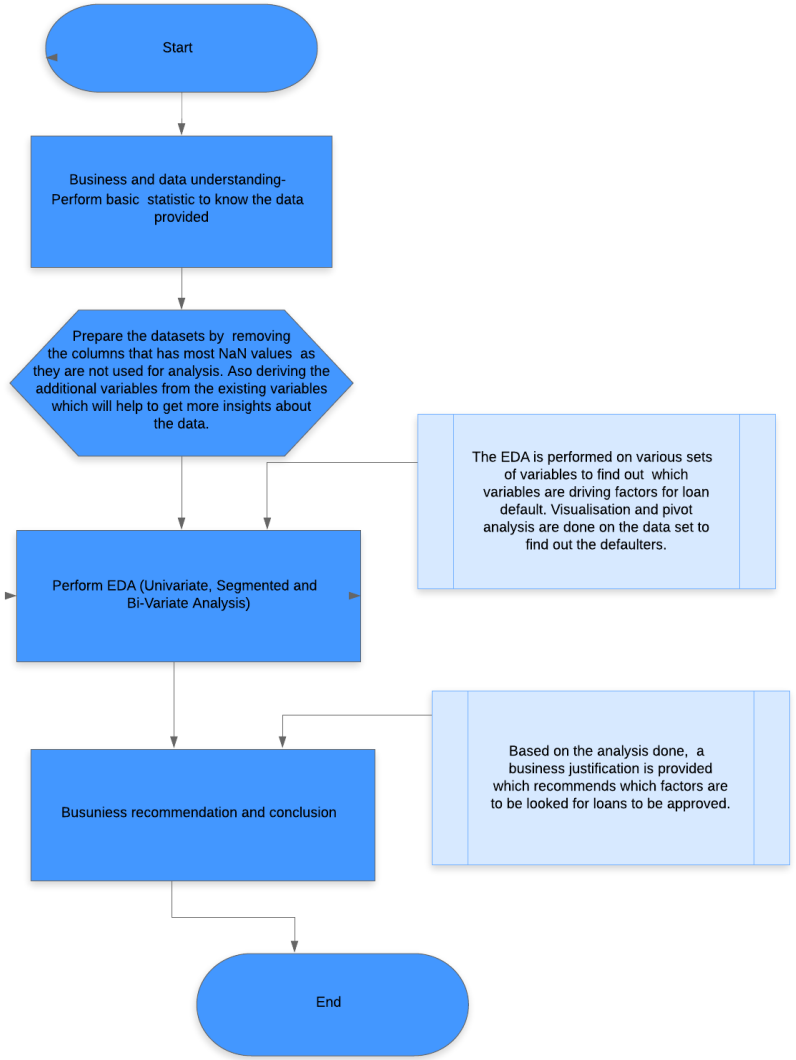
Understand the data , perform data cleaning and missing value treatment and derive additional variables from the existing variables which will help to get more insights about the data.

Perform EDA:

Analysis is done on the data set to find out the defaulters. EDA is performed on various sets of variables to find out which variables are driving factors for loan default and then plot them in graphs for visualization.

Business recommendation and conclusion:

Based on analysis, a business justification is provided to fulfill the business requirement.



Data understanding and cleaning:

We have performed the below process as part of data understanding and cleaning

- We first have a feeling of the data that is presented in the dataset by taking out a subset. Then we have removed all the columns which has missing values as those columns are not required for the analysis.
- We formatted certain columns which we thought were continuous/numeric variables, but they were not represented in the dataset as numeric. We reformatted them accordingly. removeTrailingCharacter and dataTypeConversion are two functions that converts the column values to specific format and type. Such columns are:
  - emp_length : Employment length which is less than 1 year were considered as '0 years' and length which is 10+ years were considered as '10 years'
  - term : This column was changed to numeric by removing the trailing 'month' text (E.g 36 month)
  - int_rate : This column was changed to numeric by removing the trailing '%' text (E.g 10.65%)
  - revol_util : This column was changed to numeric by removing the trailing '%' text (E.g 83.7%)
- We derived two columns namely 'loan_defaulted' and 'issue_year' from the exiting columns 'loan_status' and 'issue_d_year'
  - Loan_defaulted: In the original data set if the value of 'loan_status' is 'Charged Off'; then defaulted is set to 1 and if the loan_status is 'Fully -Paid' and 'Current' then defaulted is set to 0.
  - issue_year : This column is populated with the year when the loan was approved and was derived from 'issue_d'
- We removed columns which has 100% missing values. Then we replaced the null values by MEDIAN if the column/variable is continuous and if it was categorical then it was replaced by MODE
- We Plotted the columns into a heat map to see which columns had still missing values so that they could be imputed

Univariate and Segmented Analysis

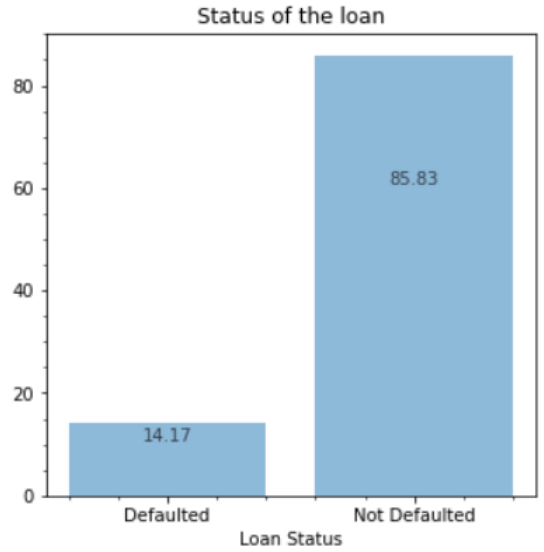
of few important variables

Loan Status:

- From our analysis we have come to know that 83% of the loans is fully paid, 14.17% of loans is defaulted and there is 2.87% of loan which was still in progress. So we can say majority of the loans was fully paid.
- Now we have a column called ‘loan_defaulted’ which says if the status is ‘Fully-Paid’ or ‘Current’ then the value is 0 and if the status is ‘Charged-off’ then its 1. So, the total good loan percentage is 85.83 which includes loans that are currently in progress and 14.17 loans are bad loans.

Loan Status	Percentage(%)
Fully paid	83
Charged-off	14.17
Current	2.87

Charged Off	Percentage(%)
Fully paid	85.83
Charged-off	14.17



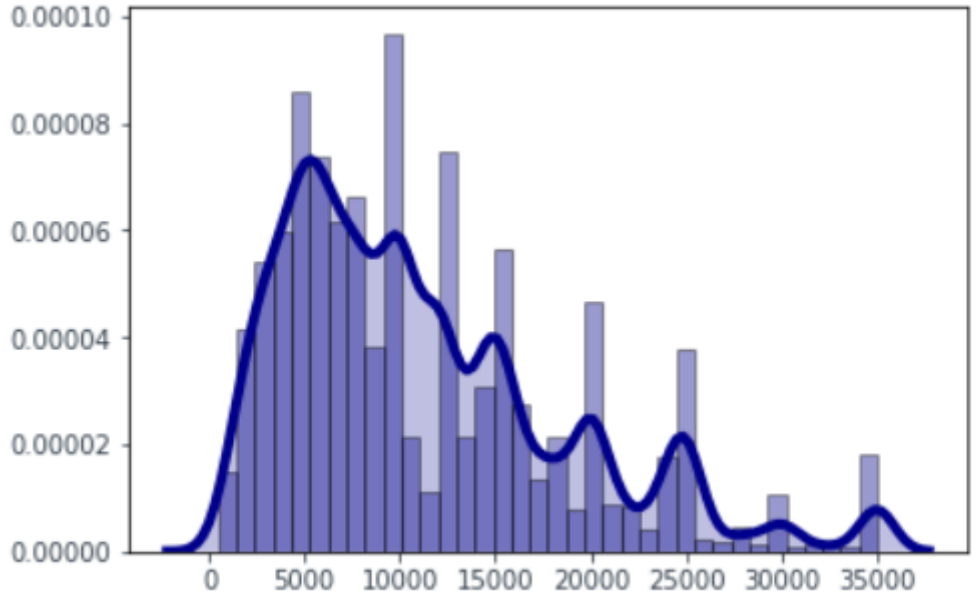
Loan Amount

From our analysis of loan amount we have observed the below facts

- 75% of the loans are below 15000 USD. The distribution of overall applied loan amount is right skewed (mean is greater than the median)
- Spike can be seen around each 5000 boundary(5000,10000,15000,20000,25000,30000 and 35000)

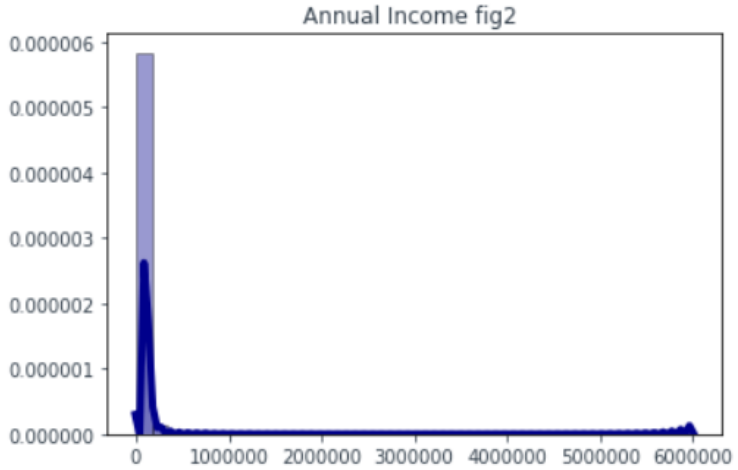
The summary of the loan amount distribution is as below.

Field	values
count	39717.000000
mean	11219.443815
std	7456.670694
min	500.000000
25%	5500.000000
50%	10000.000000
75%	15000.000000
max	35000.000000

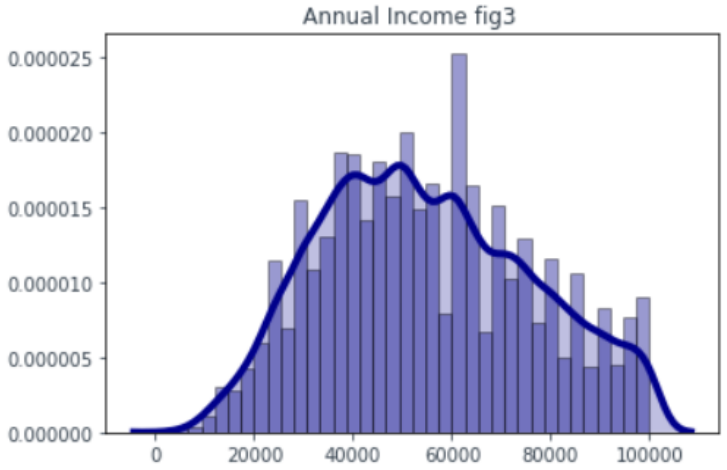


Annual Income

- From our analysis of annual income we have observed the there are outliers in the income and range of income is between 4000 (min) to 6000000(max).
- So, We have plot annual income of the dataset in a dist plot and seen that people have income from 0 to 100000 USD which is 85% of the data.

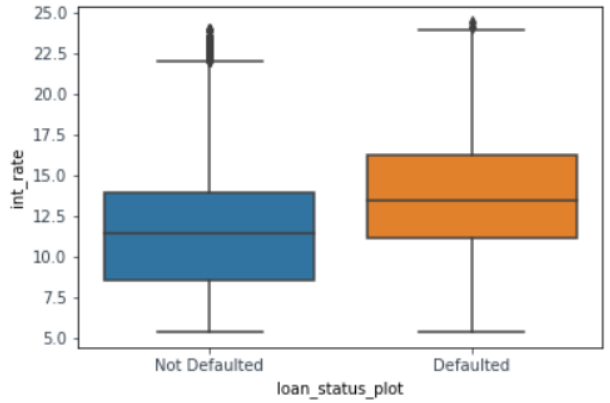


- Finally we have observed that the annual income below 100000 USD follows a normal distribution and spikes can be seen for some annual income value

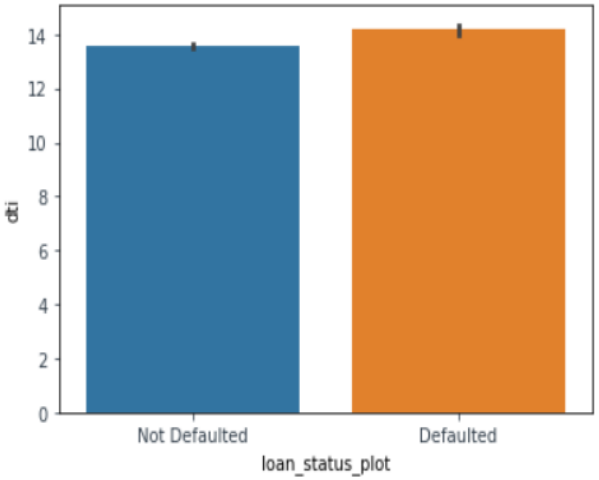


Interest Rate and DTI

- We have gone ahead to check the interest rate and DTI of the applicants. From our analysis of interest rate, we have observed the interest rate varies from 5.42% to 24.40%. Also the interest rate for defaulted loans appear to be higher than that of Not defaulted loans which is normal as the risk increases, the interest rate on loans also increases. We have plotted the dataset in a boxplot based on interest rate and below result can be seen.



- In case of Debt to income ration (DTI), we have seen that high dti translates into higher default rates. We put the dti of the dataset in a bar plot and the results can be seen as shown in the fig.

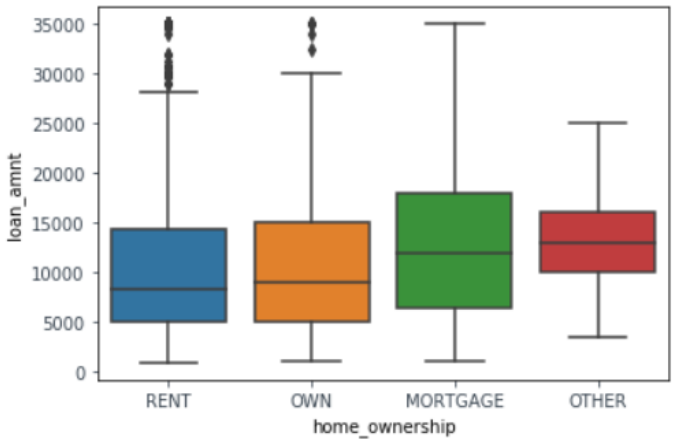


Bi-Variate Analysis
Default loan analysis based on house ownership status

- After our analysis of house ownership based on default loan analysis status, we have concluded that most of the most defaulters are staying in rented house.

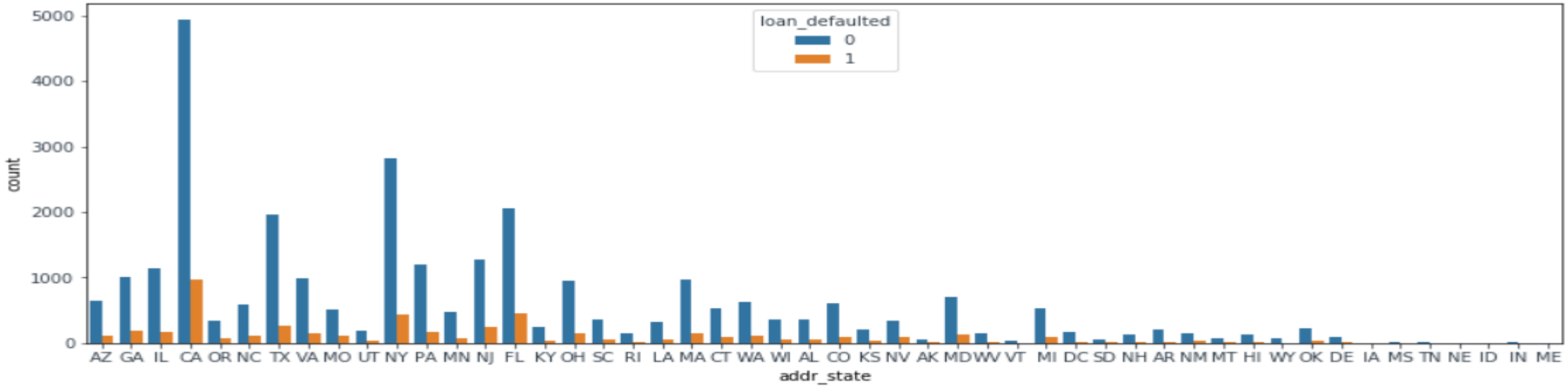
House ownership	Percentage(%)
RENT	52.830189
MORTGAGE	38.629593
OWN	8.202582
OTHER	0.337637

- Also there is a significant amount of high values loans that are defaulted for people who are staying in rented house



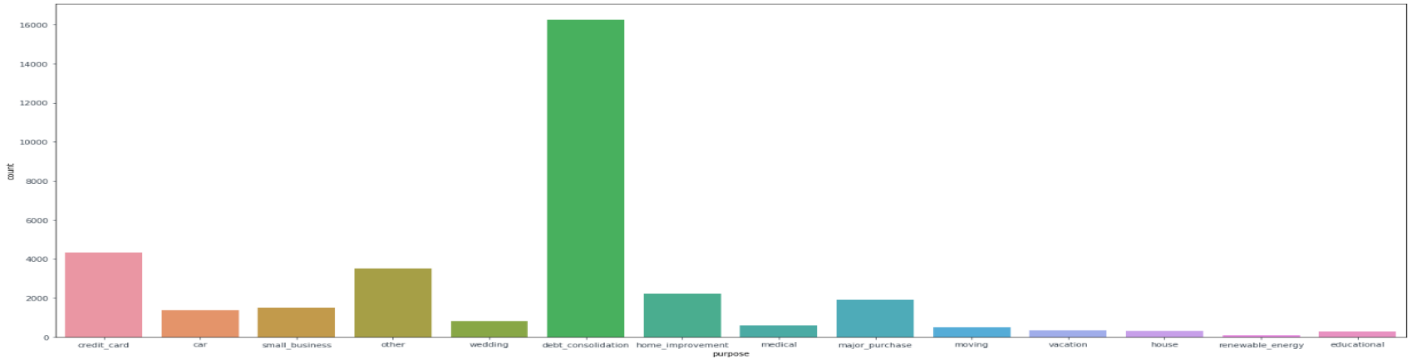
Default loan analysis based on states

- From our analysis we can see two interesting trends:
- Firstly, The states with highest loans are CA, FL and NY and the states with highest defaulters of loans are similarly CA, FL, and NY.
- Secondly, bad loans to good loans ratio is higher in Florida(22.9%) than California(20.26%) and New York(16%). So Florida loans are riskier than other states

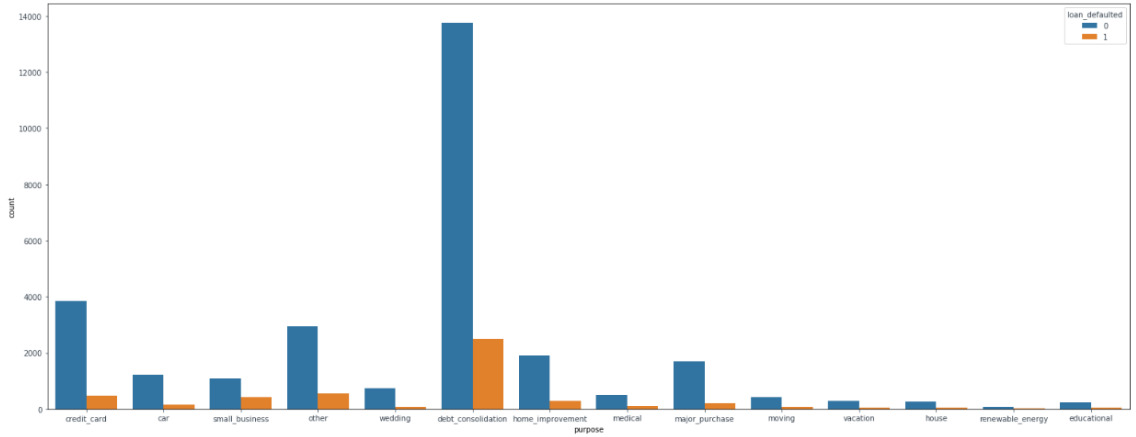


Default loan analysis based on purpose and purpose over interest rate

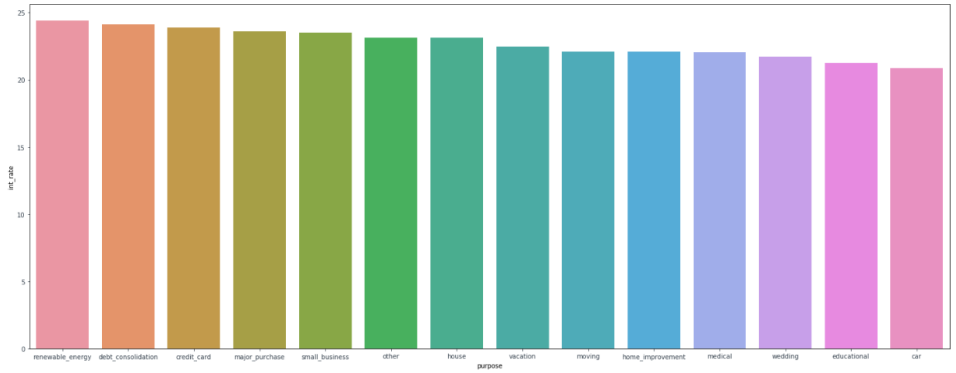
- From our analysis we can see top three application for loans are for the purpose of debt_cpnsolidation, credit_card and home_improvement. (Fig1)



- Top three defaulters similarly debt_consolidation, credit_card and home_improvement.(Fig 2)



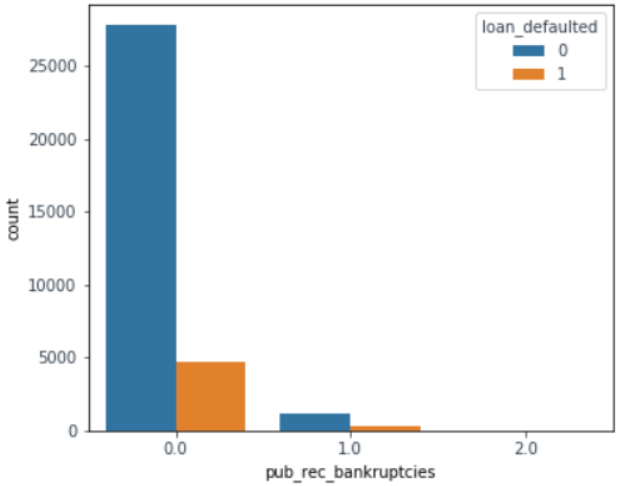
- The top 3 purpose with highest interest rates are renewable_energy, debt_consolidation and credit_card (fig3)



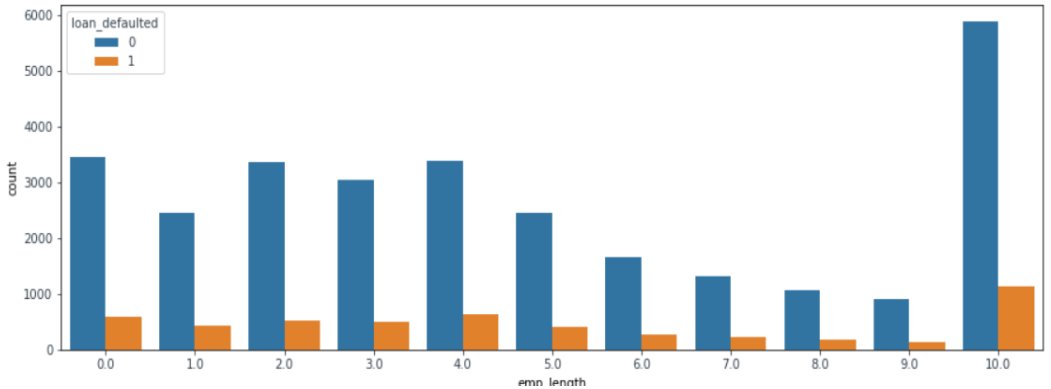
Default loan analysis based bankruptcy and employment length

- Our Observation on bankruptcy is that higher number of loans are allocated to people with 0 bankruptcy record. Less loans are allocated to people with 1 and 2 bankruptcy record. Also there is a high chance that loan will be charged off if bankruptcy record is 1 and more. So higher the bankruptcy higher the loan default rate.

Bankruptcy	Percentage(%)
0.0	93.366435
1.0	6.593843
2.0	0.039722

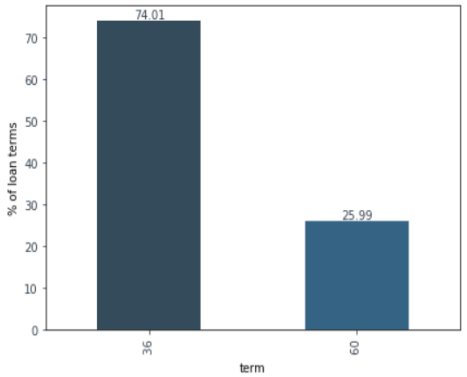


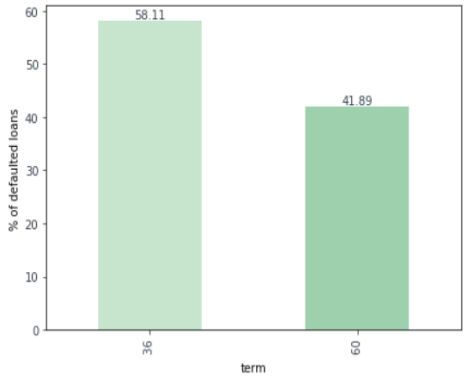
- Observation on employment length is that major amount of loans are given to the people with higher employment length and higher the Employment length, lower the risk of loan getting defaulted.

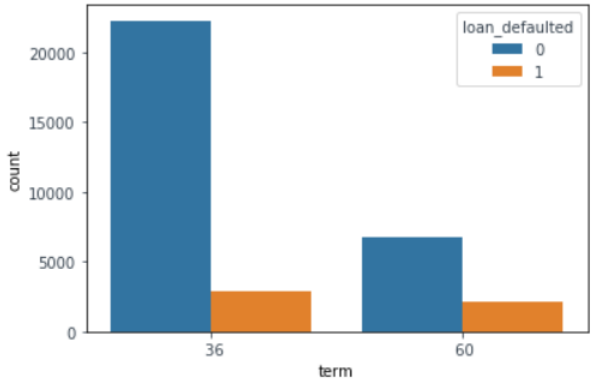


Default loan analysis based loan term and grade.

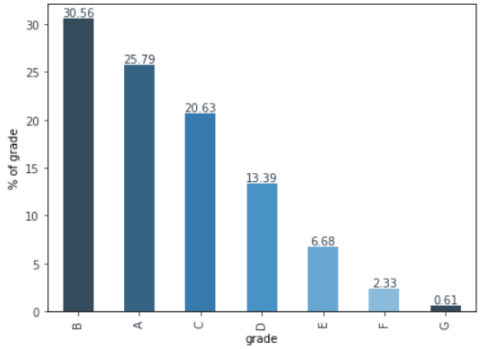
- Our Observation on loan term is that around 75% of the loans are of 36 months duration and 25% loans are of 60 months duration and the longer the duration, the percentage of loan defaulted goes up which can be seen as percentage increases to 41.89% for 60 months duration

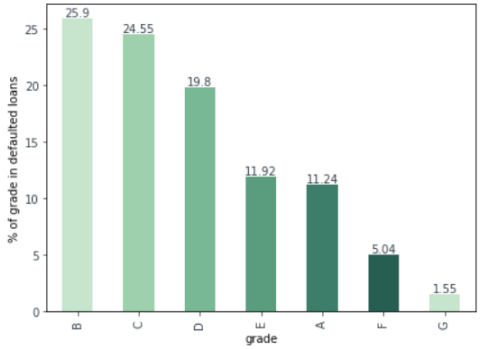


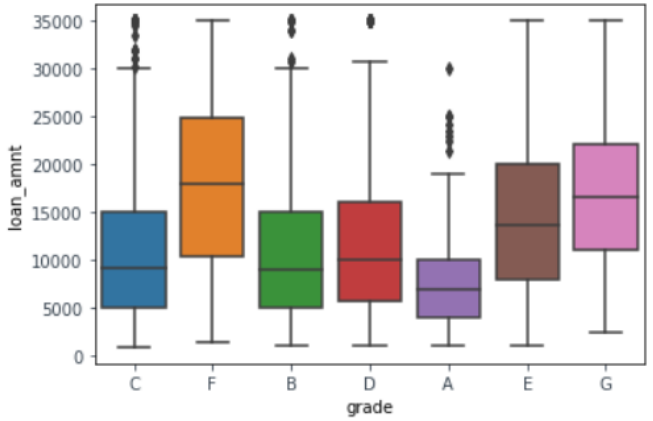




- Observation on grade is that around 50% of the defaulted loans fall under B and C grade. Also grade A is relatively safe and there are high value loans in grade B and C



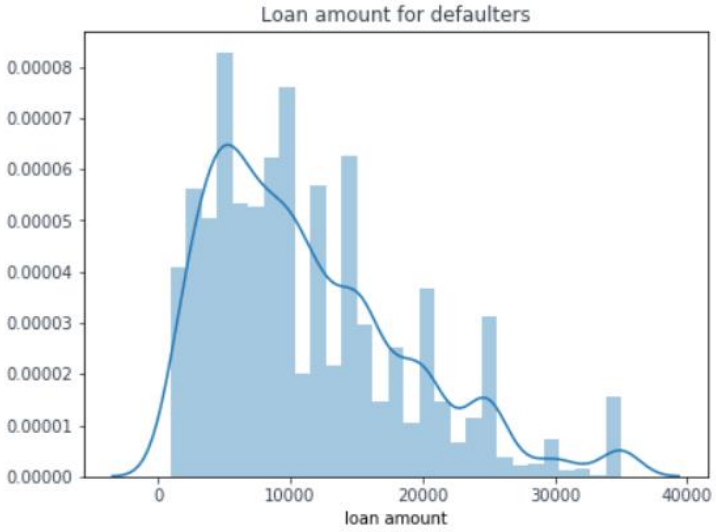


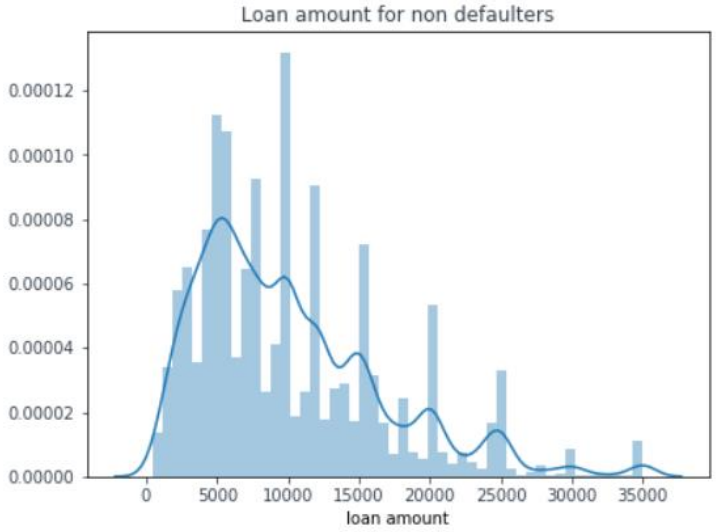


Default loan analysis based on loan amount.

- From our analysis of loan status over loan amount we observed the below facts
  - ❖ Number of loans applied by defaulters are less than loans applied by non defaulters
  - ❖ Average amount of loan applied by defaulters are higher than non defaulters

Statistical Terms	Bad Loans	Good Loans
min	900.000000	500.000000
25%	5000.000000	5000.000000
50%	10000.000000	8800.000000
75%	15250.000000	14000.000000
95%	25000.000000	24000.000000
max	35000.000000	35000.000000





Business Conclusion

- **From our analysis, we have concluded that following set of variables should be considered before loan is approved and the company can strongly utilize this knowledge for its portfolio and risk assessment.**

  - ❖int_rate : Interest rate of defaulters are higher as interest rate increases with risk and higher the chances of default
  - ❖dti: Hight dti translates into high default rate
  - ❖home_ownership: There is a significant amount of high values loans that are defaulted for people who are staying in rented house and rented house is a big factor of loan defaulters.
  - ❖emp_length: Higher the employment length, lower is the risk of loan getting defaulted.
  - ❖purpose: Top three defaulters are debt_consolidation, credit_card and home_improvement. So people with loan application for these purposes are big set of defaulters.
  - ❖pub_rec_bankruptcies: There is a high chance that loan will be charged off if bankrupcy record is 1 and more. So higher the bankrupcy higher the loan default rate.
  - ❖term: The longer the duration, the percentage of loan defaulted goes up which can be seen as percentage increases to 41.89% for 60 months duration
  - ❖grade: Max(50%) people in grade B and C are defaulters.
  - ❖add_state: The states with highest defaulters of loans are from CA, FL, and NY. So its risky to give loan to these three states