Lending Club Case Study

Case Study Group:

- SM Arun Kumar
- Mansi Sharma

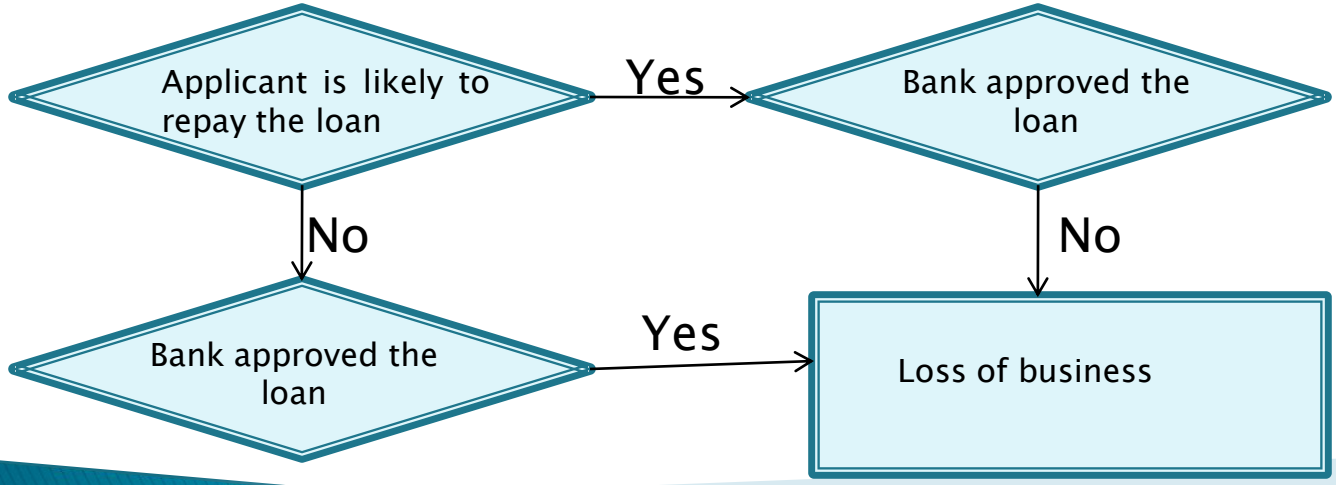
Lending Club Analysis Overview

- Identify the business problem
- Data understanding
- Data Cleaning
- Data Imputing & Pre-Analysis
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Summary
- Suggestions

Identify the business problem

Problem Statement:

- Lending club is a financial services company which specializes in lending various types of loans to urban customers.
- This club approves loan based on the applicant's profile.
- Two types of risks are associated with the bank's decision:



Data understanding

- Dataset contained information containing to the borrower's past credit history and Lending club past history.
- Total dataset contained 39717 records.
- Variables present with the dataset provides ample amount of information which is helpful in analysis.
- We require only those variables which has direct or indirect impact on a borrower to be defaulter.
- We have considered loan_status as a target variable.
- To achieve it, we have prepared the data by selecting variables that would best fit this criteria.

Data Cleaning

1) All the columns which has 0 or NA values are dropped.
2) Following columns which do not aid our analysis are dropped.

Attributes	Dropping reason	Attributes	Dropping reason
URL	It will not add any value	**delinq_amnt**	Data is 0
emp_title	Varying string field, can't be categorized.	**acc_now_delinq**	All borrowers have 0
pymnt_plan	The column has only n which doesn't give any meaning or add value.	**chargeoff_within_12_mths**	Either 0 or NA values
__title__	Loan title provided by the borrower will not help in our analysis much due to varying string.	**__out_prncp, out_prncp_inv__**	>15% of data is 0

Data Cleaning

Attributes	Dropping reason	Attributes	Dropping reason
initial_list_status	It has only type 'f'	**tax_liens**	>30% data is blank
next_pymnt_d	>30% data is blank	**id**	Dropped due to duplicate data with member_id
collections_12_mths_ex_med	Either 0 or NA values	**mths_since_last_record**	> 10% of data is null
policy_code	All the users are with code 1 out of 2 codes code 1 and code 2	**desc**	varying string field and purpose field can be used to categorize on loans
application_type	All applicants are Individual and there are no joint applicants	**member_id**	individual IDs will not help our analysis

Data Cleaning

Following items are modified for analysis:

1. term is in months: Renamed the heading to term_in_ mnt
2. int_rate and revol_util: Removed the % and make to float variable
3. issue_d, earliest_cr_line, last_pymnt_d, last_credit_pull_d: Seggregate it to month/year
4. emp_length: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years. >> change less than 1 year as 0, removed years string, 10+ as 10

Data Imputing & Pre-Analysis

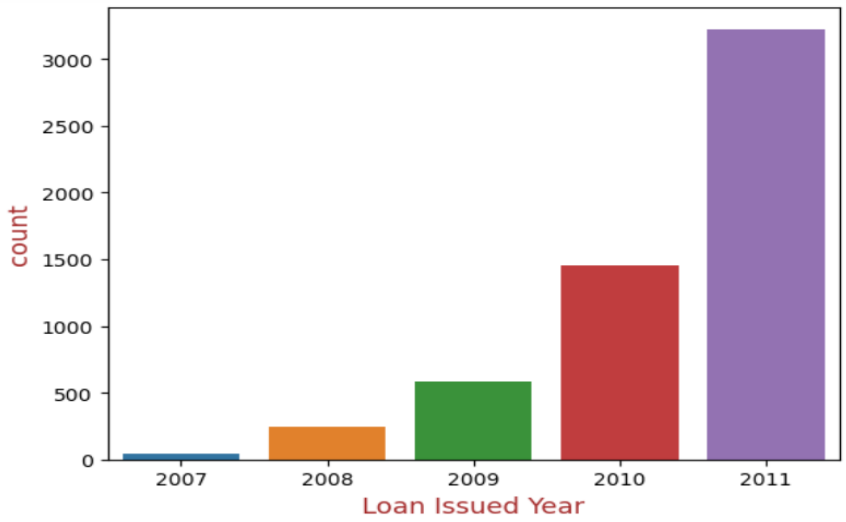
Following items are imputed:

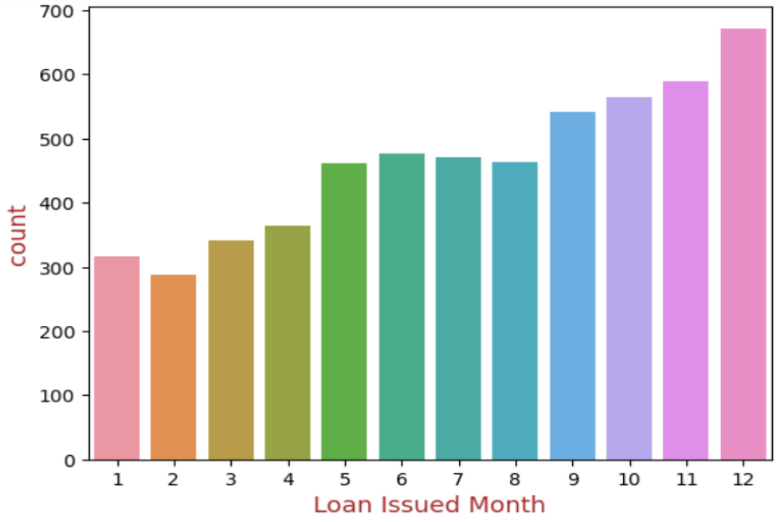
1. HOME OWNERSHIP: NONE is imputed with OTHER.
2. Employee Length: Applicants who have opened accounts before 2000 are imputed with 10 years of experience and after 2000 imputed with 5 years of experience.

Pre-Analysis:

1. It has been observed that there is a correlation with loan amount, funded amount and funded amount investment.
2. Total payment, Total payment to investors, Total recovery Principal amount is also correlated.
3. Loan amount and Total payment has correlation with installment. If the correlated columns will be taken for analysis then the results would be more or less similar.
4. Outliers are removed for annual income.

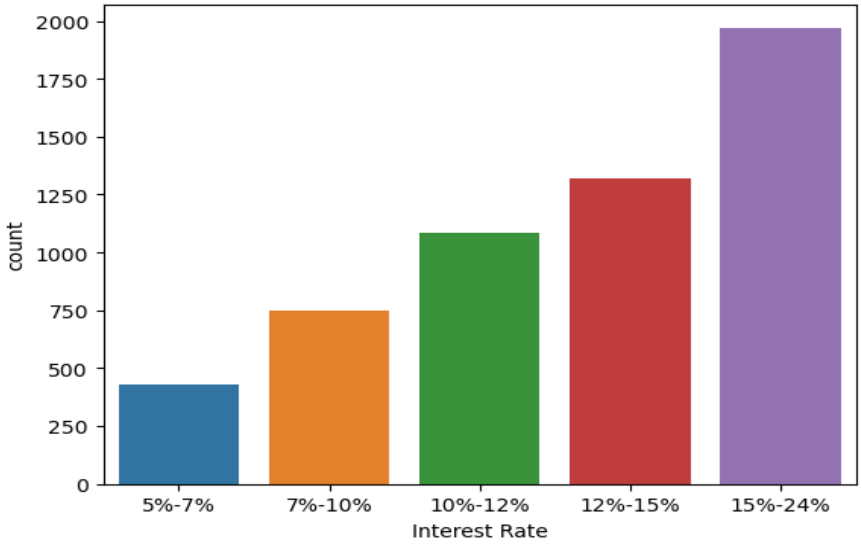
Univariate Analysis

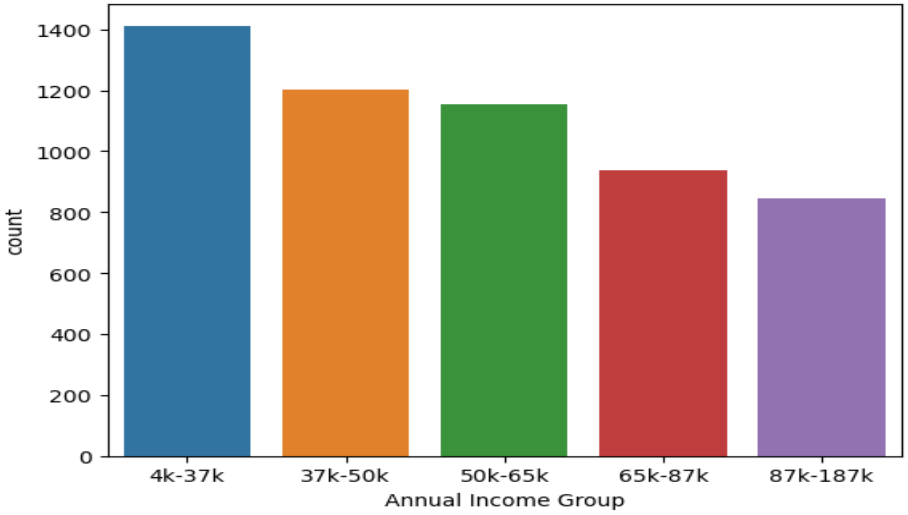




1) Borrowers taken loan in the month of November and December have defaulted more. The probable reasons could be festivals and vacations in these months.
2) Borrowers taken loan in the year 2011 have defaulted more. This could be due to the U.S.A economic conditions prevailing during that time.

Univariate Analysis



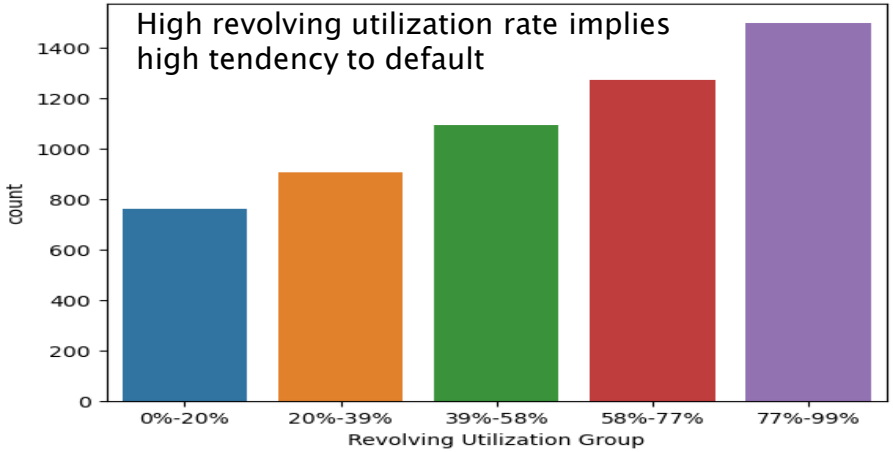


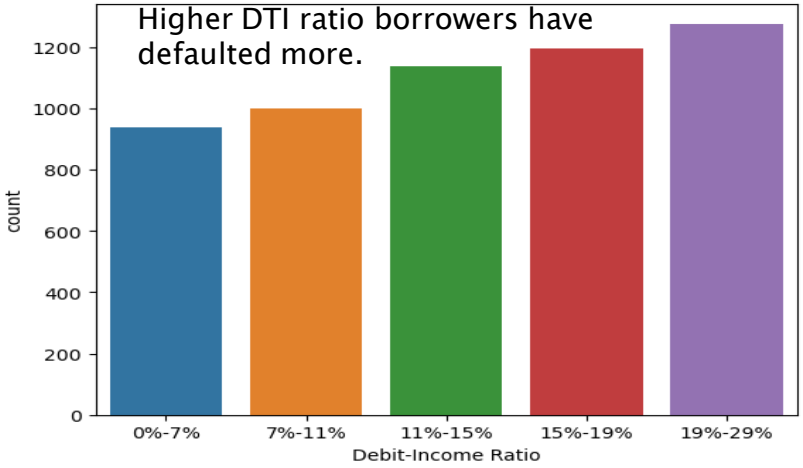
1) Low annual income group applicants have defaulted more.
2) There is a high chance with High interest rate applicants defaulting more.

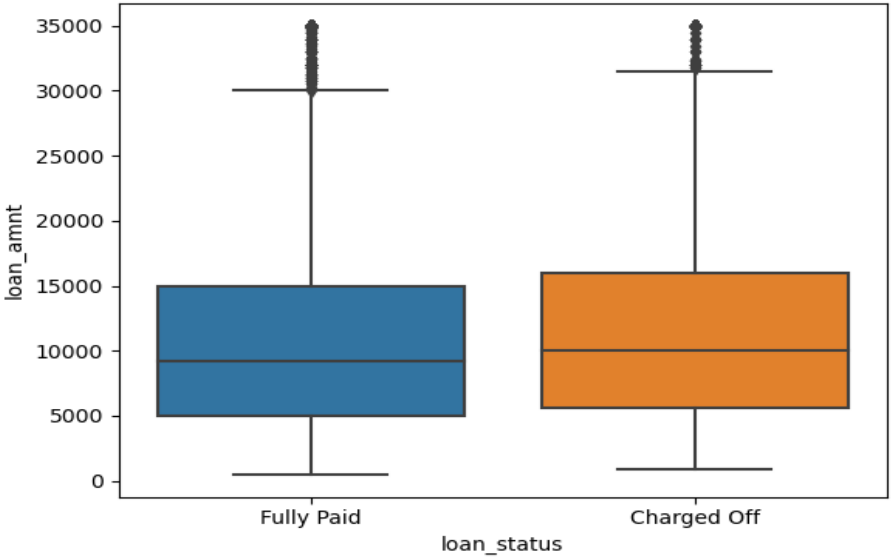
Note1: Univariate analysis graphs can be referred from python jupyter notebook.

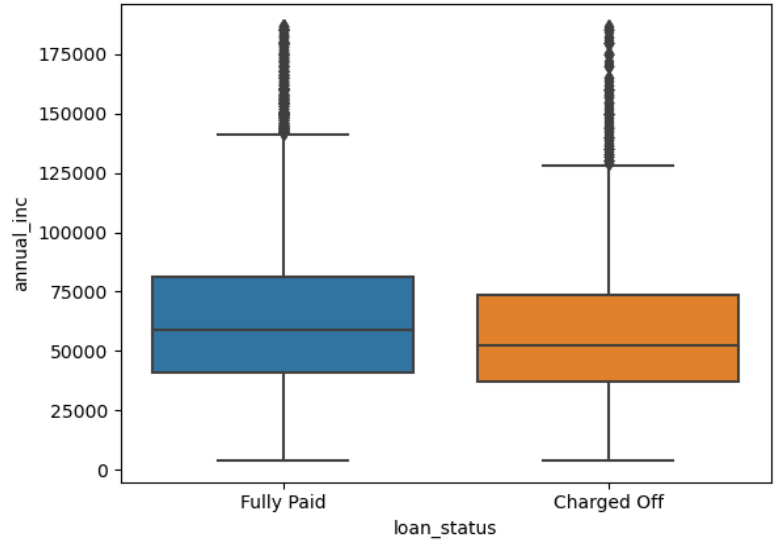
Note2: Here we have added segmented univariate analysis.

Univariate Analysis







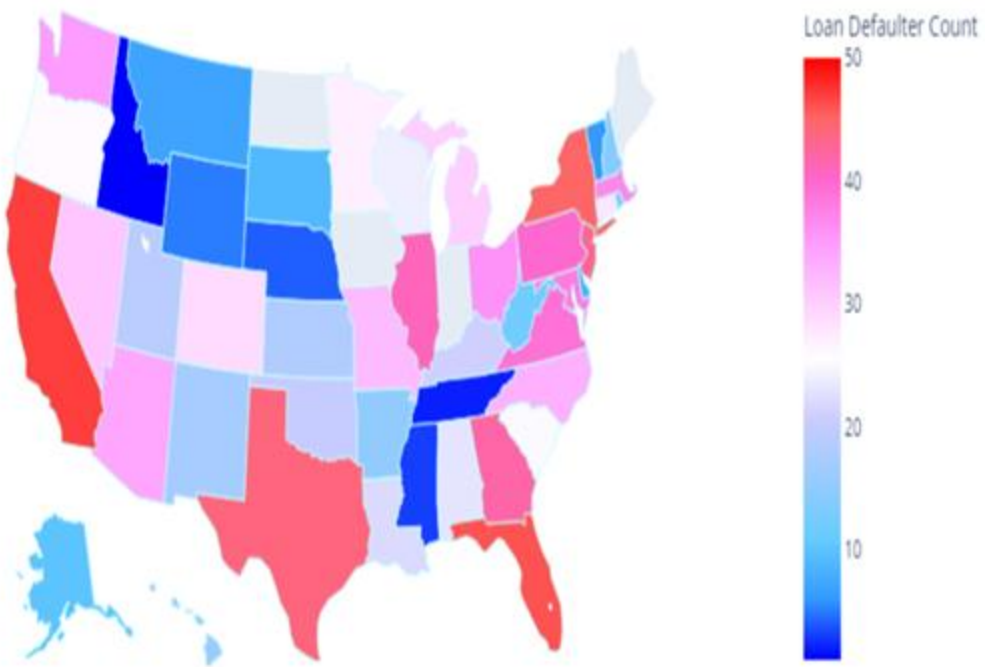


Interquartile range (IQR) for loan amounts tends to be slightly higher for charged-off borrowers compared to fully paid borrowers.

Interquartile range (IQR) for annual salaries is lower for charged-off borrowers compared to fully paid borrowers.

Univariate Analysis

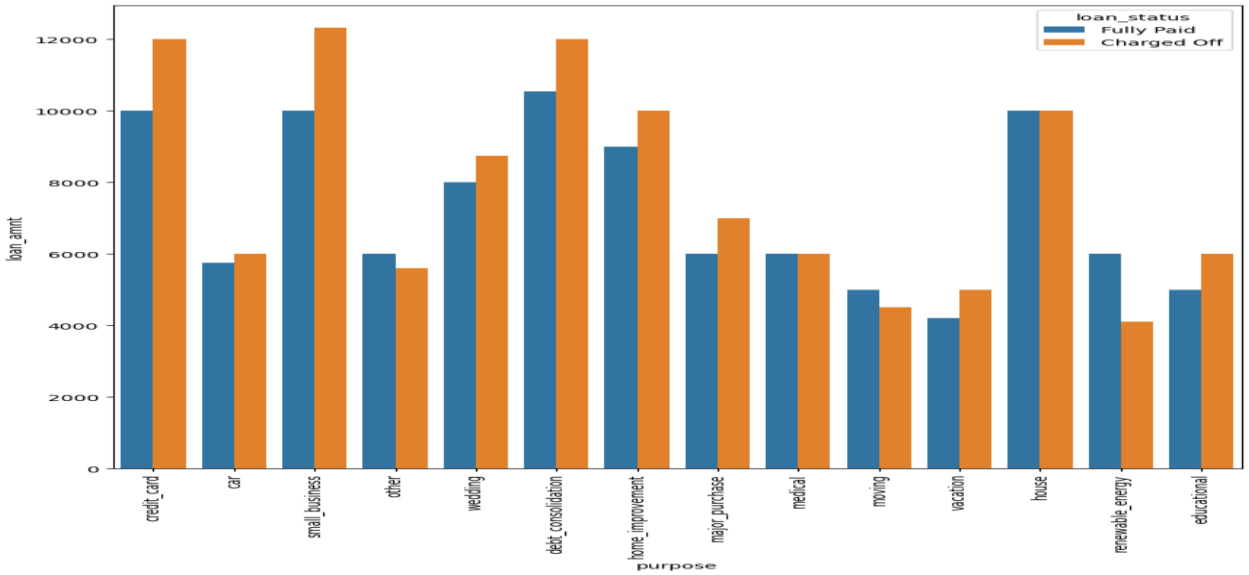
Loan defaulter count v/s Regions



1. More applicants are from California and more defaulters are also from the same region.
2. Regions highlighted with blue have lower defaulter rate.
3. All the States with lower default count are adjacent to each other.
4. It is noteworthy that states located along the border of the USA exhibit a higher default rate.

Bivariate Analysis

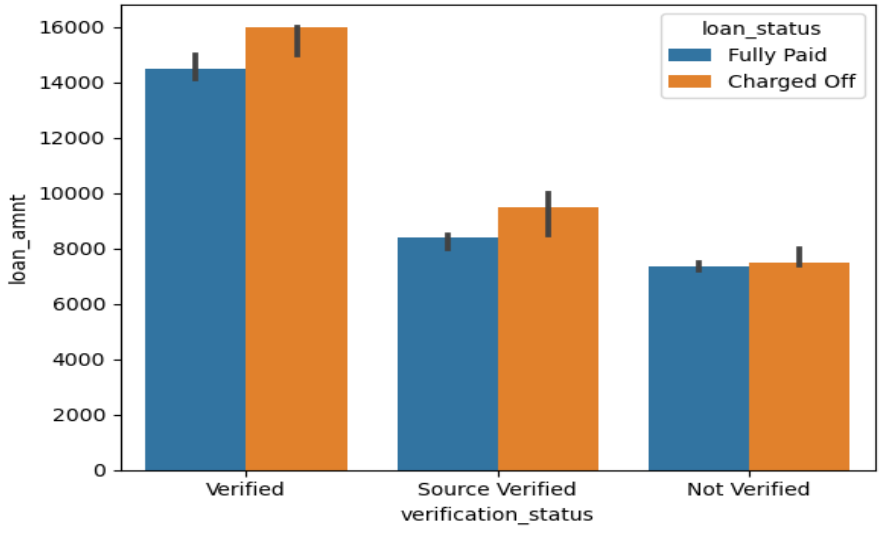
Bar graph of purpose v/s loan_amnt



1. Higher loan amount is sanctioned for following purposes credit card, small business, debt consolidation.
2. High no of defaulters fall in these categories.

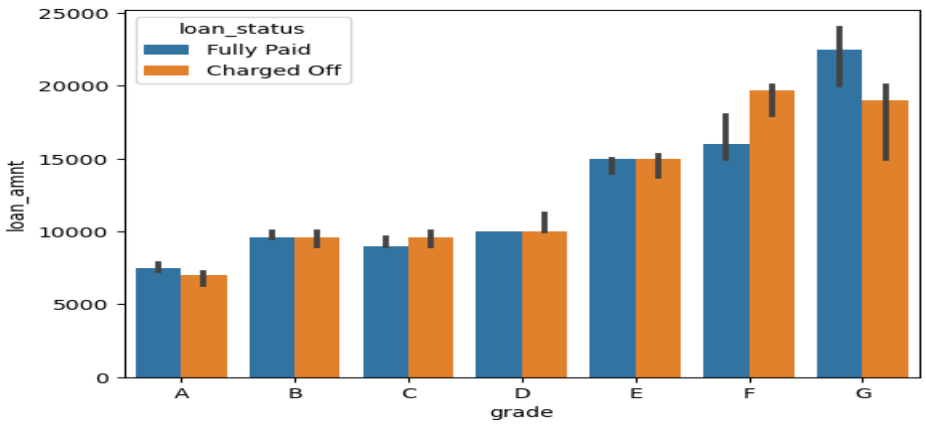
1. Higher loan is approved for loan applicants with verified status.
2. It has been observed that default rate was high for verified loan applicants. So, verification process strictly needs to be amended else there could be higher loss to LC/investors

Bar graph of verification_status v/s loan_amnt

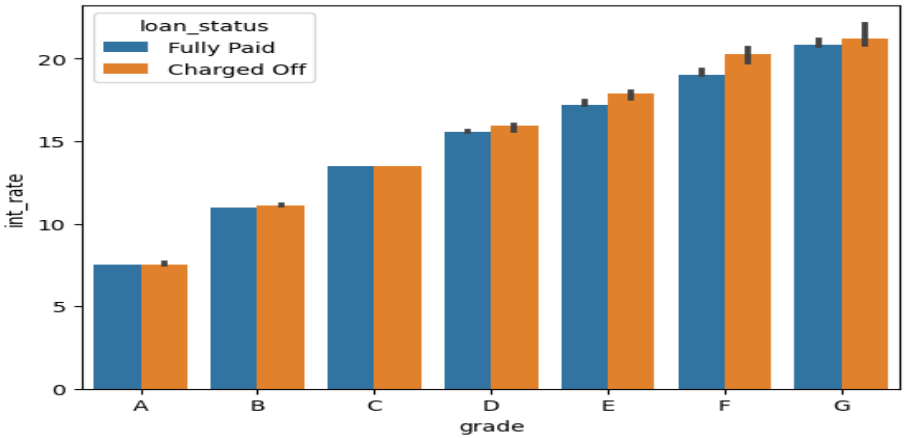


Bivariate Analysis

Bar graph of loan_amnt v/s grade



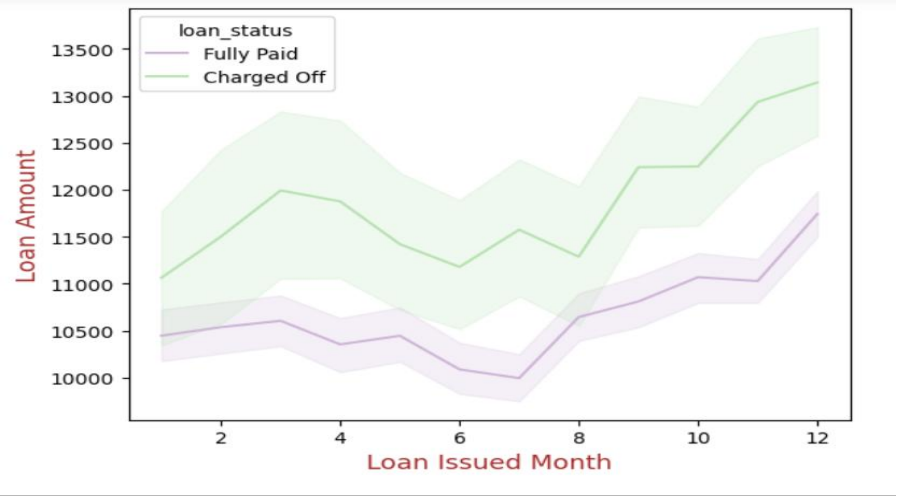
Bar graph of int_rate v/s grade



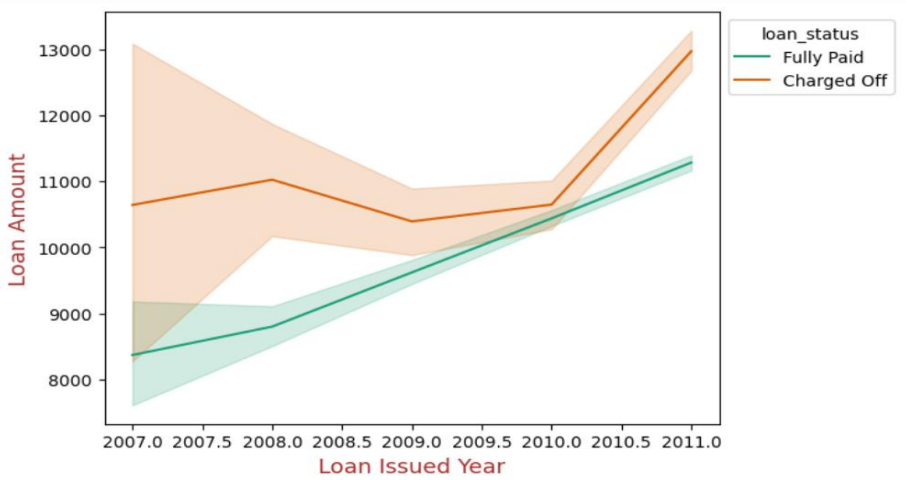
1. Higher loan amount is approved for Grade G applicants.
2. Higher Loan amount implies high interest rate and high interest rate falls for Grade F and G applicants.

Bivariate Analysis

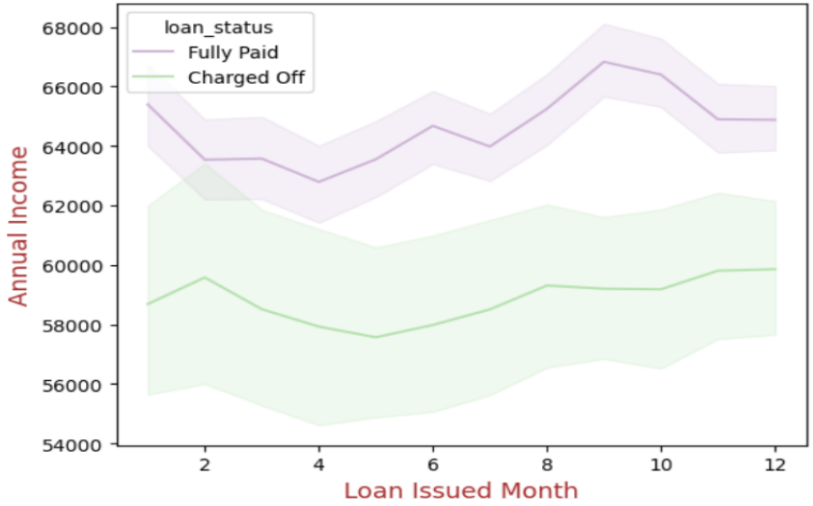
Bar graph of Loan issue month v/s Loan amount



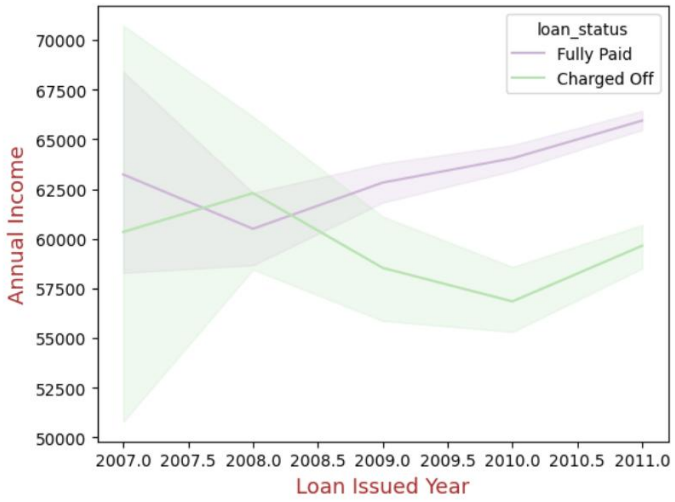
Bar graph of Loan issued year v/s Loan amount



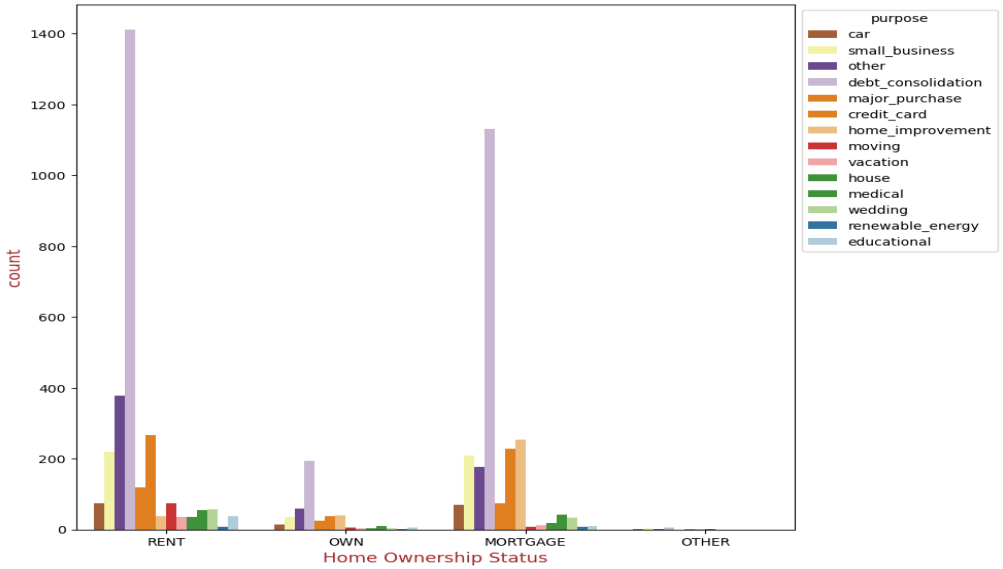
Bar graph of Loan issued month v/s Annual income

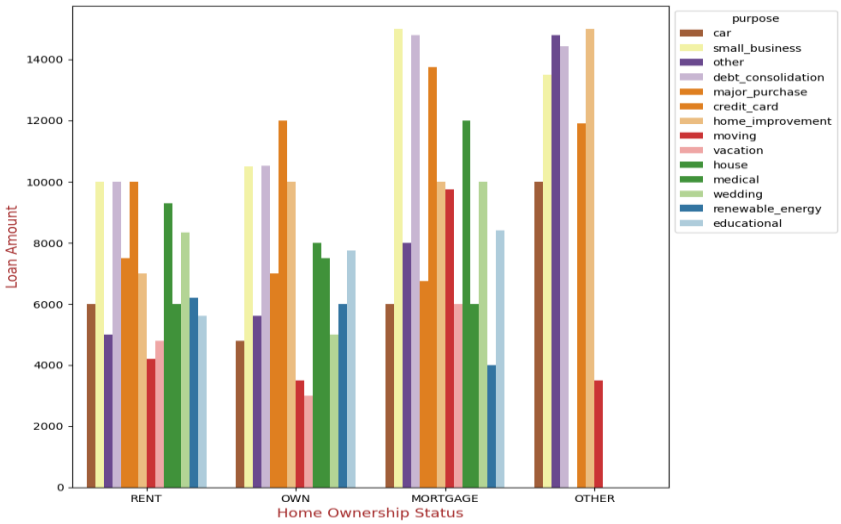


Bar graph of Loan issued year v/s Annual income



Multivariate Analysis

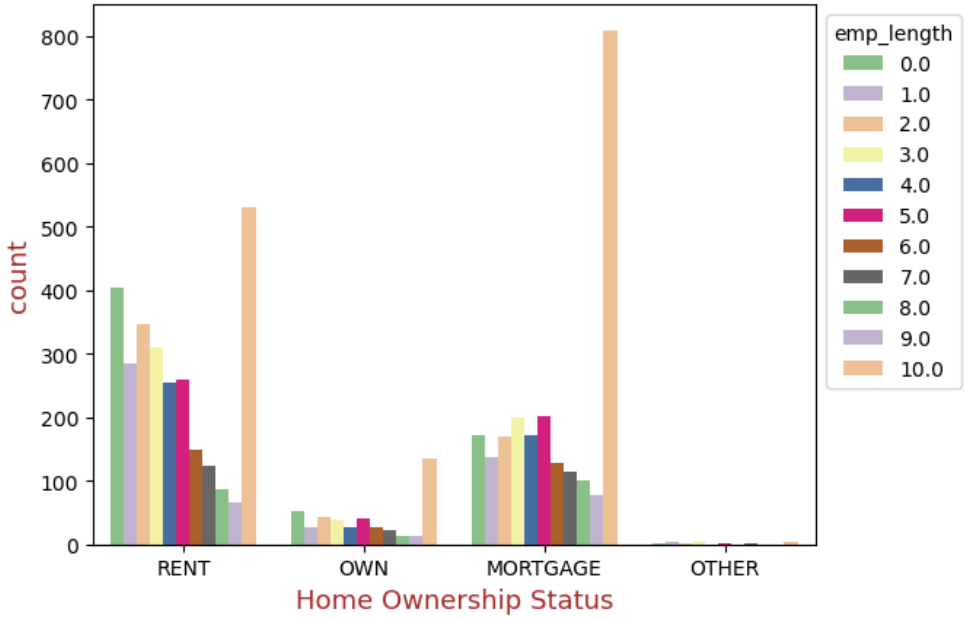


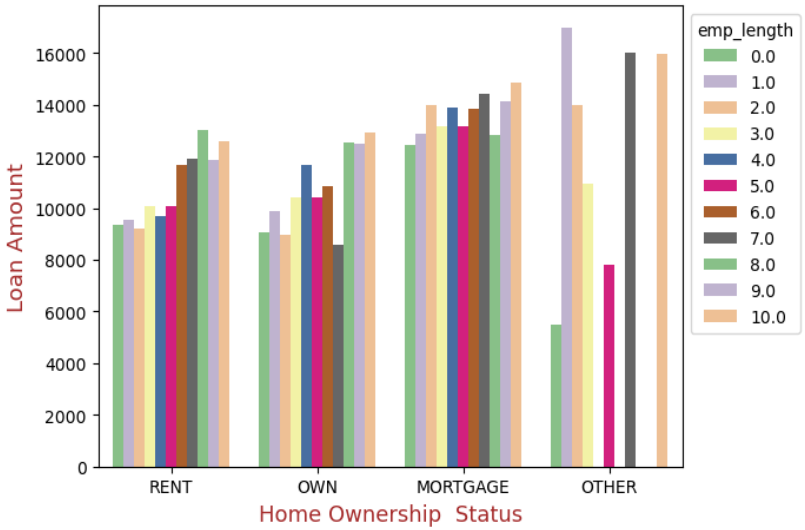


1. Home Ownership v/s frequency plot reveals that loan applicants staying in Rented houses or who have Mortgaged houses are the ones who have applied loan for debt consolidation and they have defaulted more.
2. Home Ownership v/s Loan Amount reveals that loan applicants staying in mortgaged houses applied for a higher loan amount for small business, debt consolidation and credit card.

Multivariate Analysis

Home ownership status v/s employment length v/s charged off

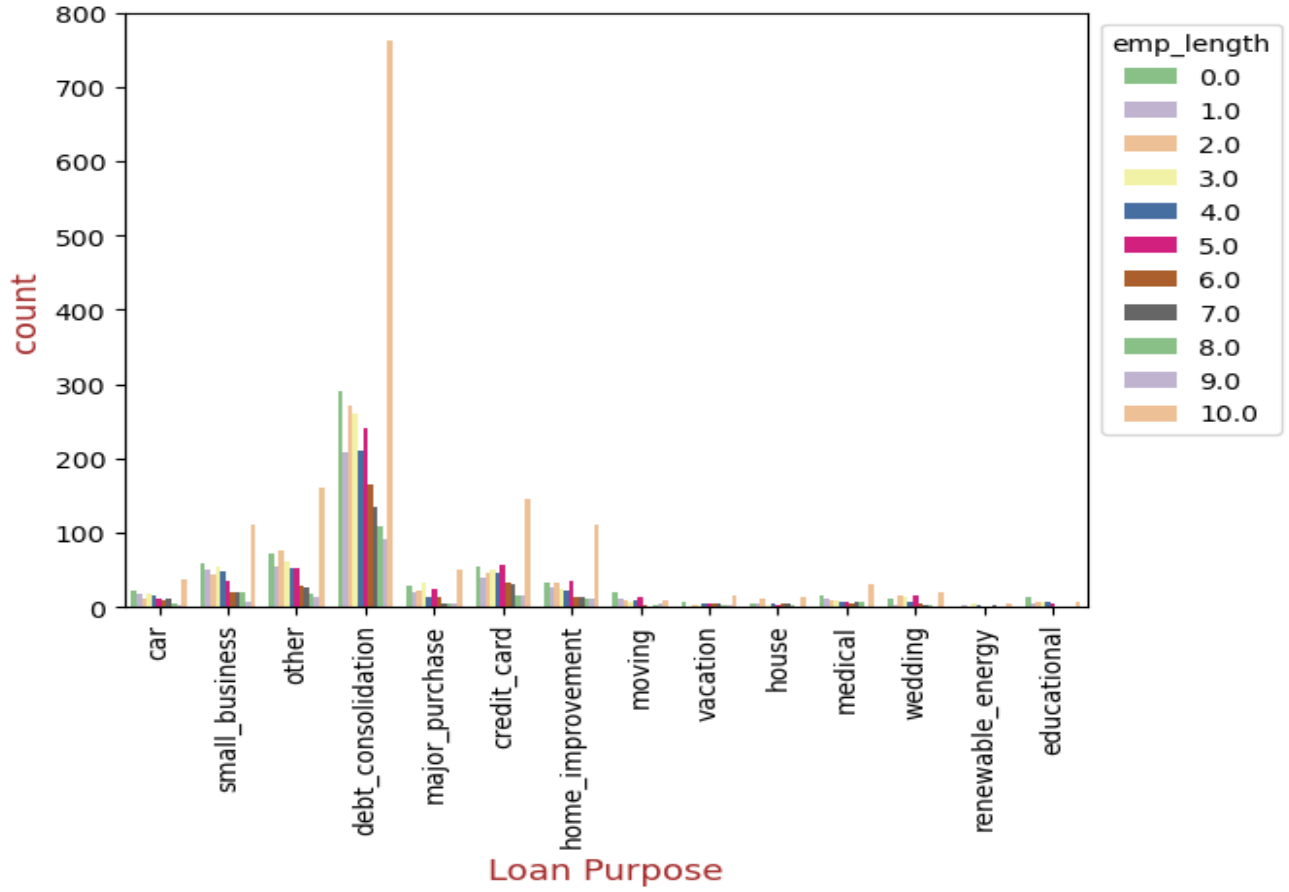




1. Employees who have mortgaged their houses and have 10 or more years of experience have more probability to defaulters.
2. Second highest defaulters are employees having 10 or more years of experience staying in rented house.

Multivariate Analysis

Loan Purpose v/s employment length v/s charged off



1. It shows that employees with experience >= 10 years, have applied loan for debt consolidation are defaulted more.

Summary

Based on the Analysis done on the variables, below mentioned points are concluded :-

- Loan applicants with >= 10 years of experience who have applied for debit consolidation and living in rented and mortgage houses have defaulted more.
- Lower income group applicants have defaulted more.
- Higher loan amount is approved for small business, debit consolidation, home improvement and credit card applicants and defaulters fall in these categories.
- Average interest rate is considerably higher for 60 months loan term than 36 months and there are more no of defaulters with higher interest rate.
- Not verified applicants have defaulted more. Even the verified applicants are the 2nd highest defaulters.
- Borrowers follow under grades B, C and D have defaulted more.
- Borrowers taken loan in the month of November and December have defaulted more. This might be due to festival season and vacations during these months.
- Loans taken in the year 2011 have defaulted more. This could be due to the U.S.A economic conditions prevailing during that time. Also, higher loan amount is taken in year 2011.

Summary

Based on the Analysis done on the variables, below mentioned points are concluded :-

- Annual income is low for charged off candidates and in the year 2008 there is decrement in the annual income of the charged off candidates.
- Borrowers with higher DTI ratio have defaulted more.
- Borrowers with a high revolving utilization rate have shown a higher tendency to default.
- Borrowers paying more late fee can be potential defaulters.
- Last Payment Year wise reveals maximum defaulters paid in 2012 and on contrary last credit pull year wise gives us maximum defaulters pulled credit in 2016. This reveals that defaulters are still borrowing money which could further burden the lenders/LC.

NOTE: For detailed analysis for graphs, refer jupyter python notebook.

Conclusion

Suggestions to Lending Club:

- Loans for Debt consolidation, Small Business, credit card and home improvement applicants should be checked properly or Loans can be provided with lower rate of interest to these applicants as there are more no of defaulters in these categories.
- Loans also taken for purpose as “other” need to be verified properly.
- Default rate was high for verified loan applicants. So, Verification process strictly needs to be amended else there could be higher loss to LC/investors.
- Lower annual income applicants should be avoided for big loan amounts with higher interest rates.
- Loan approvals must be utmost taken care in the months of November and December as there are higher no of defaulters in these months.
- Loans approval for applicants from states located along the border of the USA exhibit a higher default rate and hence applicants from these areas should be properly verified.
- Applicants of work experience >=10 years and have applied loan for debt consolidation and lived in rented or mortgage houses are more prone to defaulters. Hence, these type of applicants should be properly handled.