LENDING CLUB CASE STUDY

Authors: Vineet Kumar
Raghavender B

Problem Statement

Lending Club lends various types of loan to urban customers. On receiving loan application , company has to make a decision to either approve the loan or reject the loan application. If a loan application is rejected and if that applicant would have repaid the loan then it is loss of business. If a loan application is approved and applicant default then it is a financial lost to company.

Lending club wants to analyse and find out the variables/data patterns which impact loan repay capability and want to use those insights to take actions on the loan application like rejecting a high probable default loan application, reducing amount for loan, changing tenure of lloan, charging higher interest rate etc.

As an analyst in the company , we have to do Exploratory data analysis and determine those variables which has clear impact on loan status.

Note: Loan rejected has not been considered in Loan Dataset.

Metadata Description

Meta Data	Loan Data				
Description					
Source	loan.csv				
Format	csv				
Number of rows	39717				
Each row is	Loan information of a particular applicant				
Sampling Method	Data about loans taken between June 2007 to Dec 2011				
Column Name	Type	Description	Missing values		
id	int	A unique LC assigned ID for the loan listing.	0		
member_id	int	A unique LC assigned Id for the borrower member.	0		
loan_amnt	int	Amount applied by the borrower	0		
funded_amnt	int	Total amount committed to the loan	0		
funded_amnt_inv	float	Total amount committed by the investors	0		
term	int	The number of payments on the loan. Values are in months and can be eit	0		
int_rate	int	Interest rate on the Loan	0		
grade	Object	LC assigned loan grade	0		
home_ownership	Object	The home ownership status provided by the borrower during registration.	0		
emp_length	object	Employment length in years. Possible values are between 0 and 10 where	1033		
annual_income	float	The self-reported annual income provided by the borrower during registra	0		
issue_d	datetime	Month which loan was funded	0		
loan_status	object	Current status of Loan	0		
purpose	object	A category provided by the borrower for the loan request.	0		
dti	float	A ratio calculated using the borrower's total monthly debt payments on th	0		
year (Derived column)	int	Year of the loan funded	0		
month(Derived column)	int	Month of the loan funded	0		

Data Cleaning- Fixing Rows and Columns

Fix Columns	Action output
Add Column names if missing	No such scenario in Loan dataset
Rename column consistently	Columns naming does not require any change
Delete unnecessary columns	We deleted 57 columns having only null data
Split columns for more data	We derived Year from date columns for our analysis
Merge columns for identifiers	Not required for this analysis
Align misaligned columns	Not required for this analysis
Delete summary rows: Total, Subtotal rows	Scenario not observed in dataset
Delete incorrect rows: Header rows, Footer rows	Scenario not observed in dataset
Delete extra rows: Column number, indicators, Blank rows, Page No.	Scenario not observed in dataset

Data Cleaning- Fixing missing values

Missing Values columns	Action Output
57 columns with all values null	Removed from dataset for analysis
emp_title – 2459 Null Values , emp_length 1075 Null Values , chargeoff_within_12_mths 56 Nulls, pub_rec_bankruptcies 697 Nulls tax_liens 39 Null values	Small number of Null values compared to total row count 39717. No external data or domain info available to replace it with meaningful values so we are keeping as is into dataset.
desc – 12940 Null Values	While it has significant number of Null rows , We do not plan to consider this column for our analysis to find variables impacting Loan default so we plan to keep this column as is for info purpose.
mths_since_last_delinq – 25682 Null values	Significant Null values greater than 70% so column has been dropped from analysis
mths_since_last_record - 36931 Null Values , next_pymnt_d – 38577 Null Values	Very Significant Null values greater than 90% so columns has been dropped from analysis
Zip_code	It has 0 null values but has been dropped as not much significance to current analysis

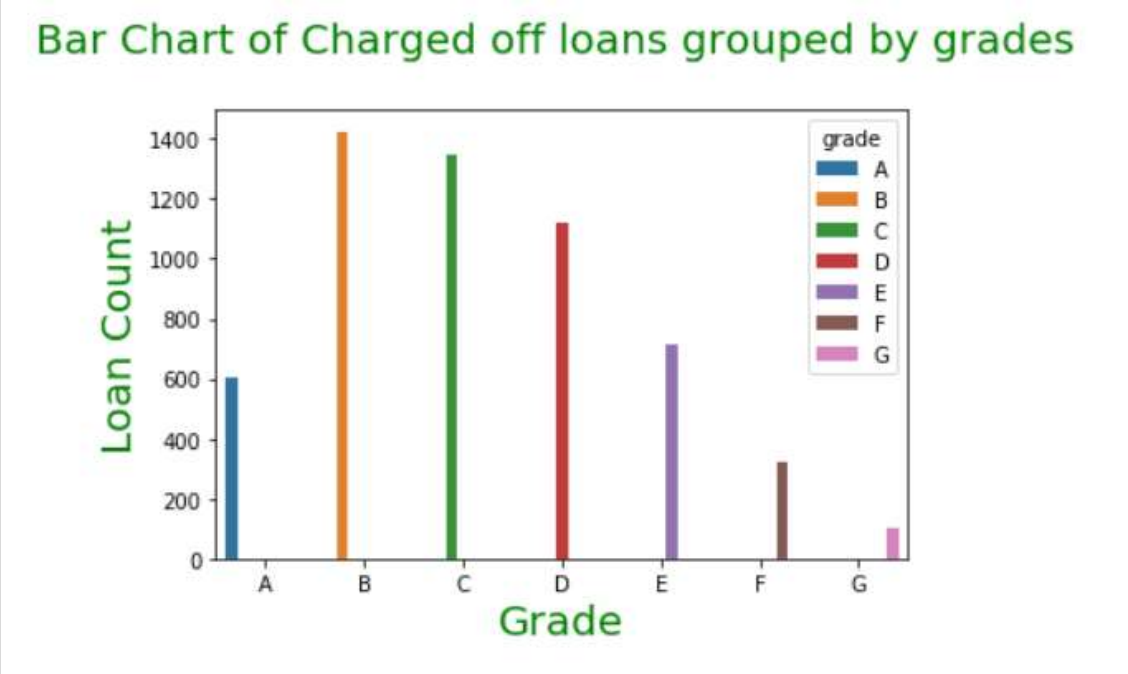
Data Standardization

Standardization	Action Output
Column name updates for better readability	Changing the term column name to term_months and int_rate column name to int_rate_percent
Data Format change	Formatting the data of column term to remove months and removing the % from the column int_rate_percent and revol_util
Data Format change	Changing data types of columns to float and int as per the data
Data Format change	converting data type to date for columns issue_d, earliest_cr_line, last_pymnt_d, last_credit_pull_d

Data Sanity Checks and Removing Invalid values if Any

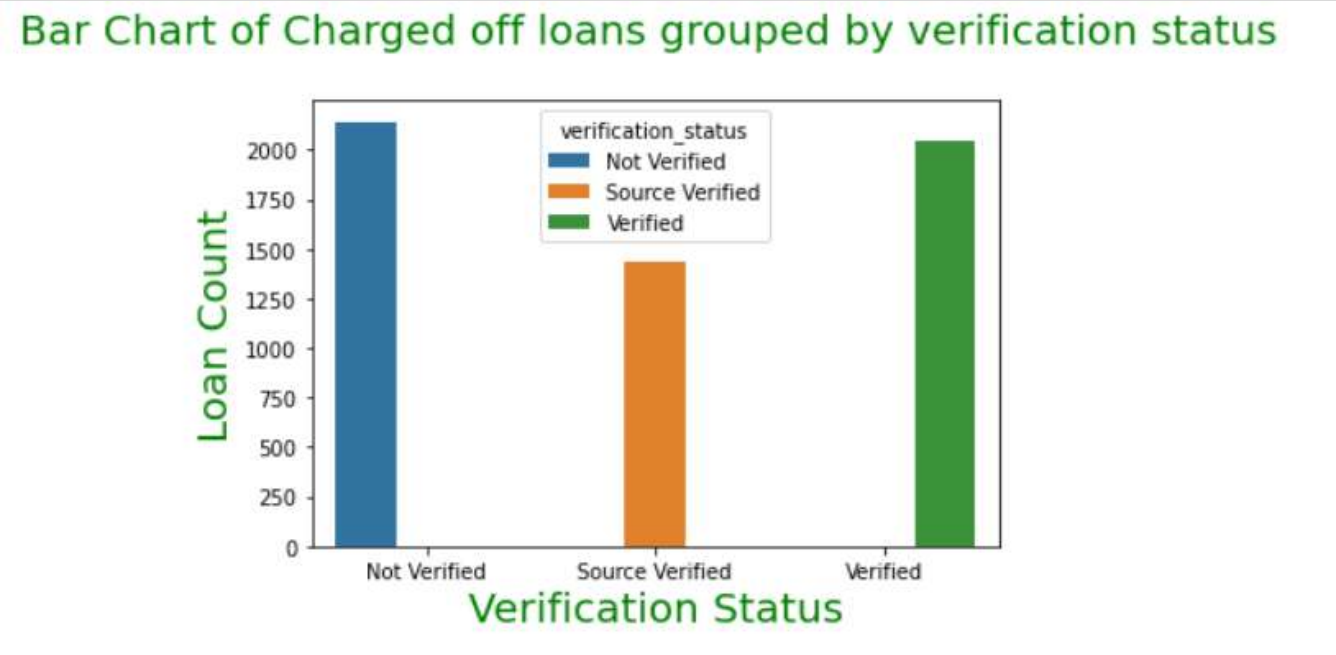
Data Sanity Checks
loan_amnt should be always greater than funded_amnt
loan_amnt should be always greater than funded_amnt_inv
int_rate_percent should be greater than 0 and less than 100

Univariate Analysis (ordered categorical variable)- Loan count distribution over grade for charged off loans



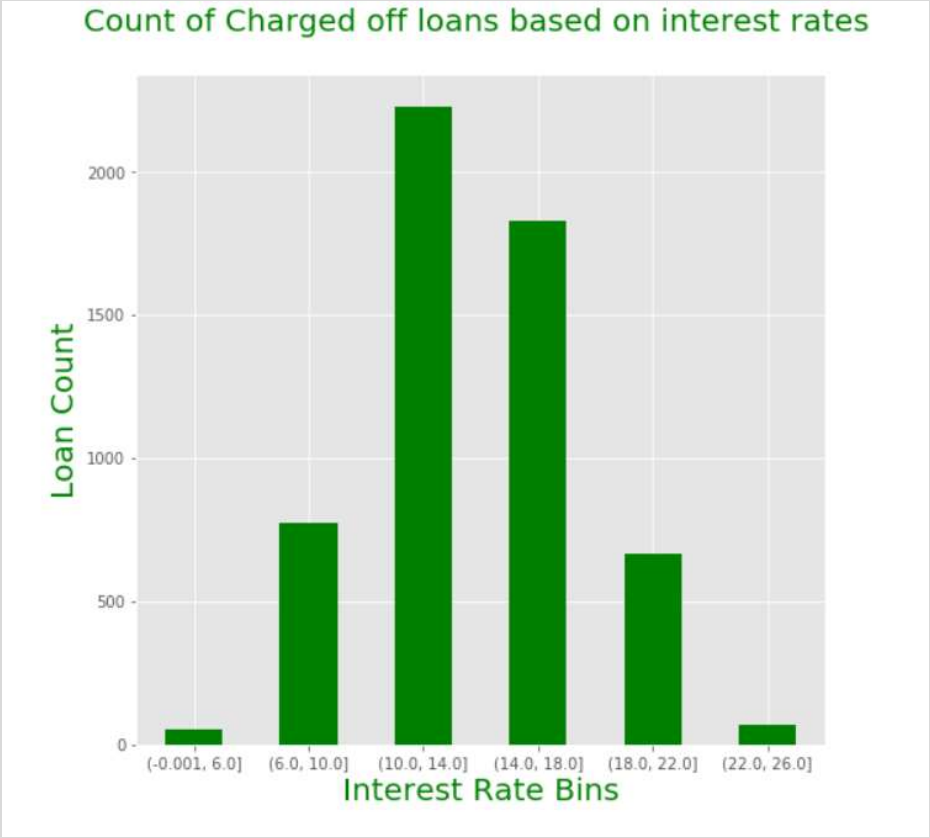
Loans with grade B and C has higher chances of default compared to loans with status A ,F and G

Univariate Analysis (unordered categorical variable)- Loan applications distribution over Verification status for charged off loans



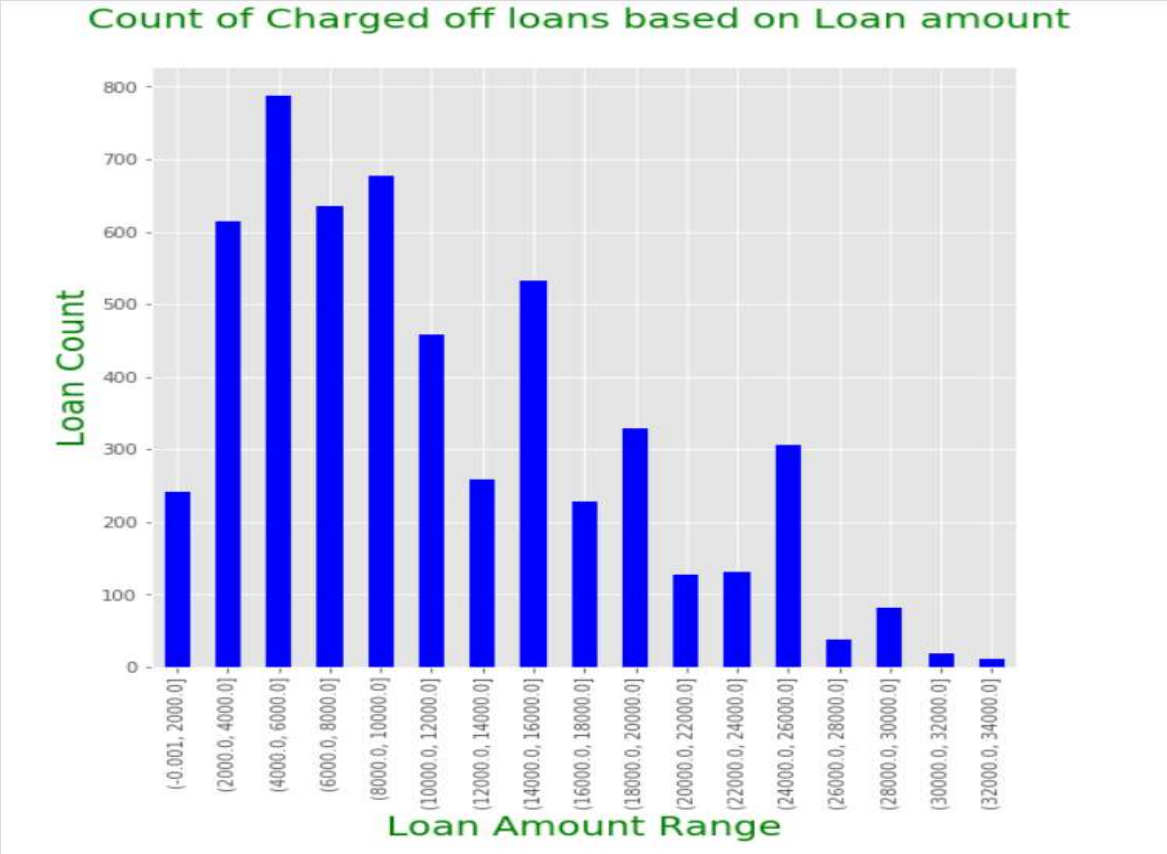
Lending club should give more loans to people whose are source verified

Univariate Analysis (Quantitative variable)- Loan count distribution for Interest rate for defaulters



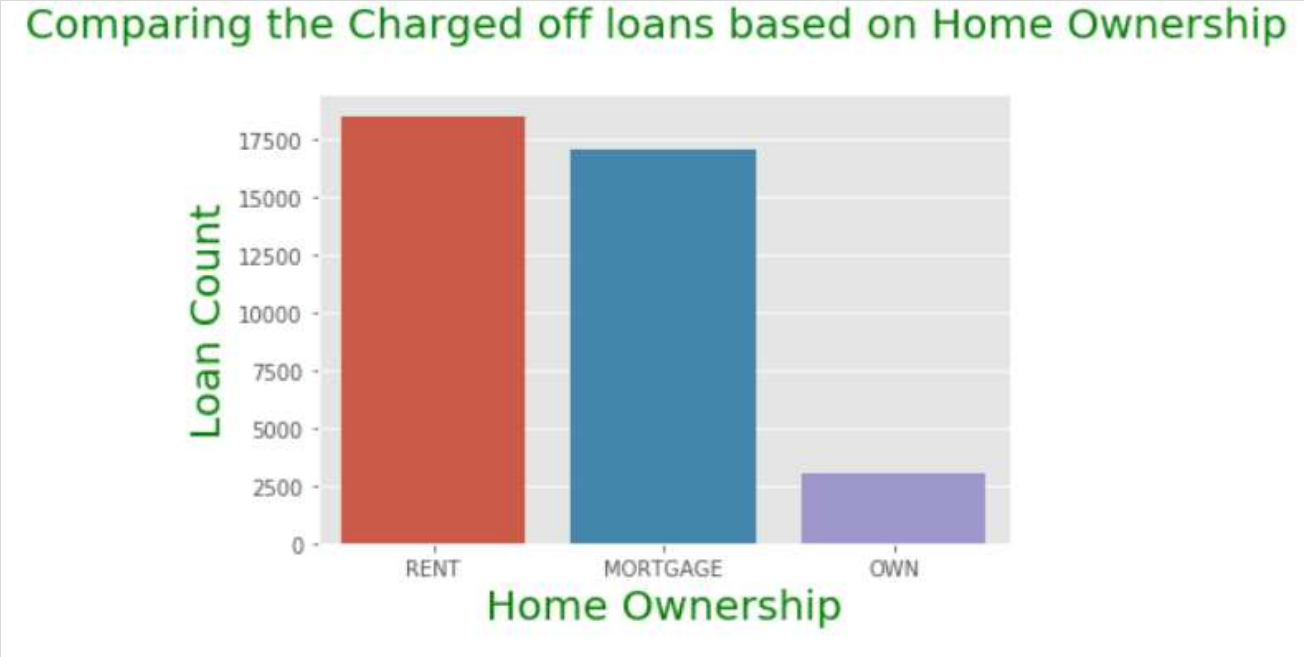
Loans with interest rates between 10% to 18% has highest chance of getting default

Univariate Analysis (Quantitative variable)- Loan count distribution for Loan Amount for defaulters



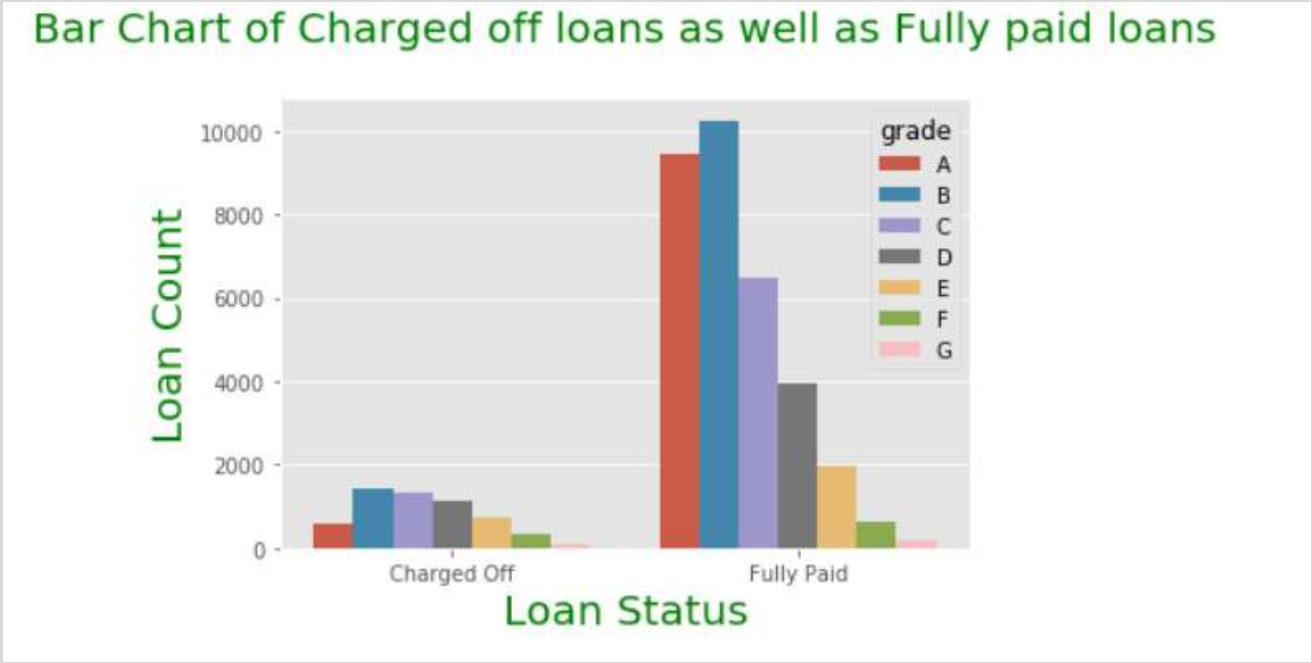
Loans grant with loan amount in range of 2000 -10000 has highest chance of getting charged off

Segmented Analysis - Loan applications distribution over loan status on home_ownership

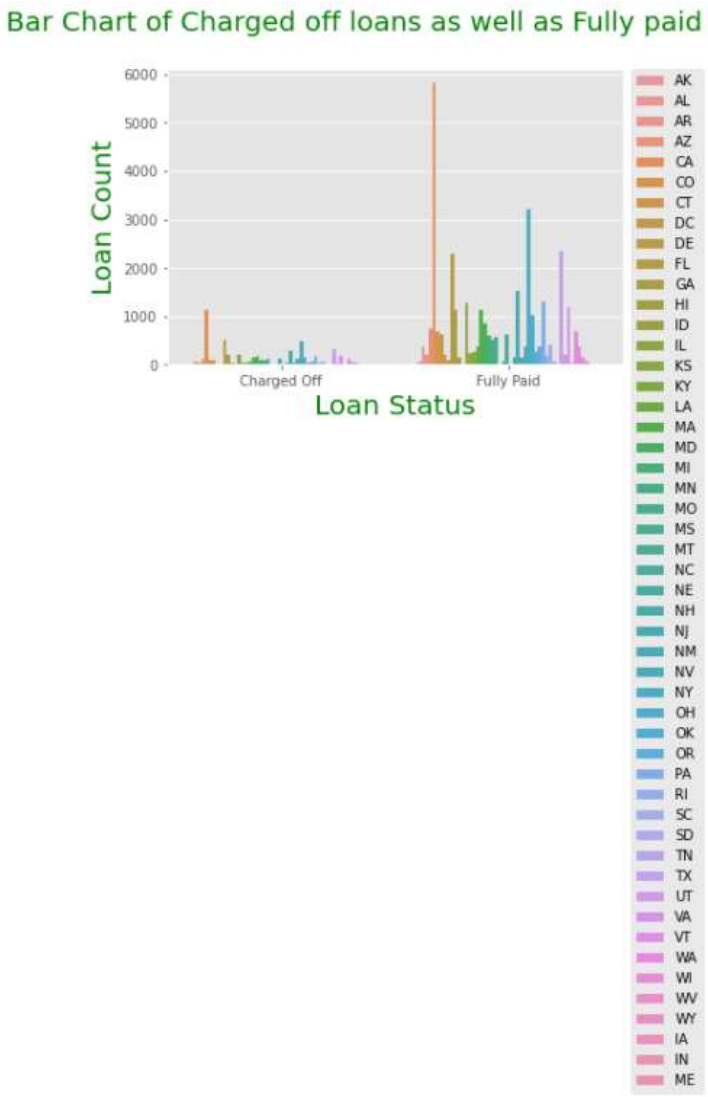


Lending club should give more loans to people who own a house as that is lowest in defaulting loan.

Segmented Analysis - Loan count distribution over loan status on grade



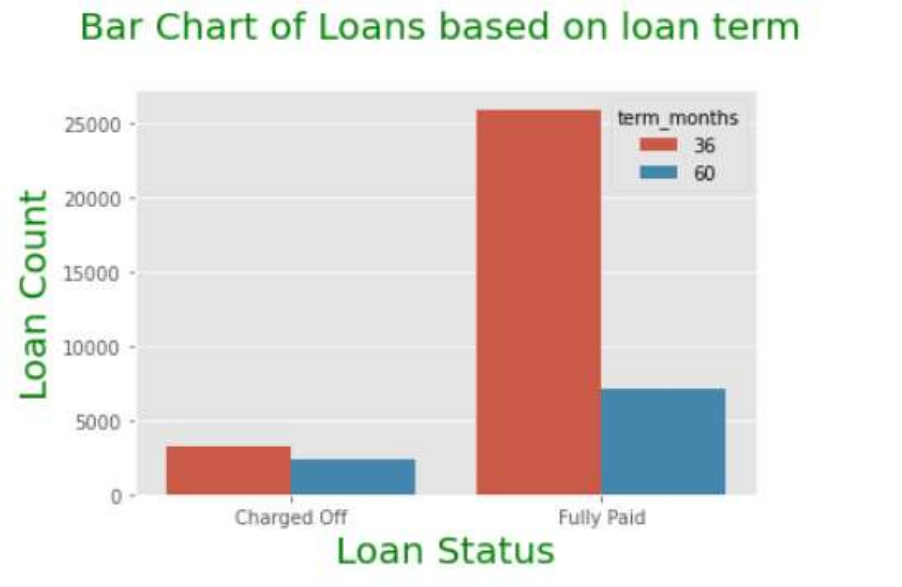
Loans with grade B and C has higher chances of default compared to loans with status A ,F and G



Segmented Analysis - Loan count distribution over loan status on state

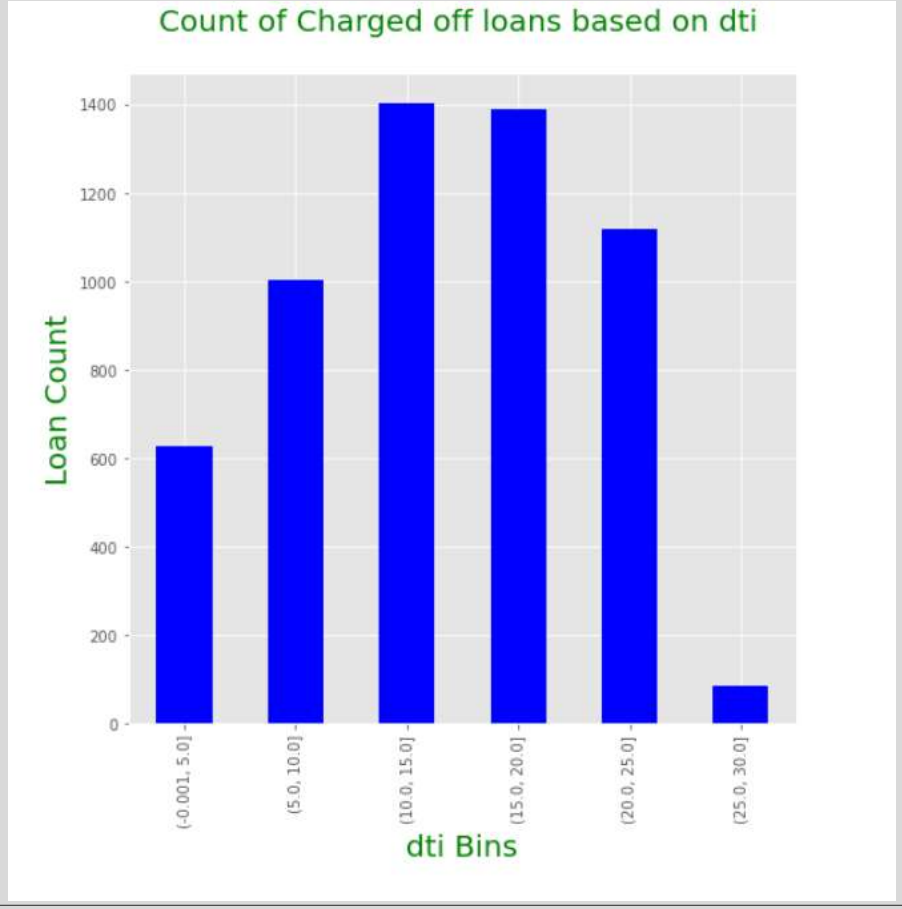
Loan applicants from CA ,CO and CT has higher chance of default may be due to state economy or some other state specific reason to be investigated

Segmented Analysis - Loan count distribution over loan status on loan term



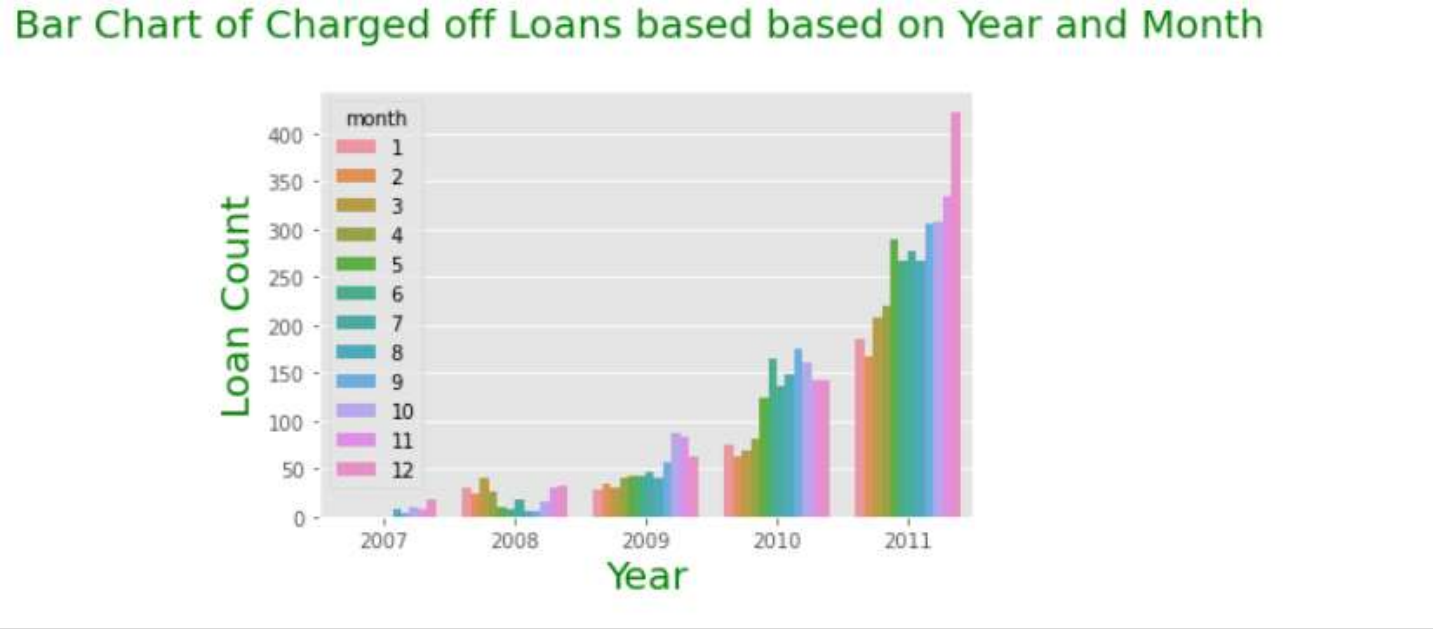
For customer with other variables indicating a possibility of charge off in future, Loan can be granted to applicant for longer duration of 60 month to reduce chances of defaulting on loan

Segmented Analysis - Loan count distribution for charged off loan on dti



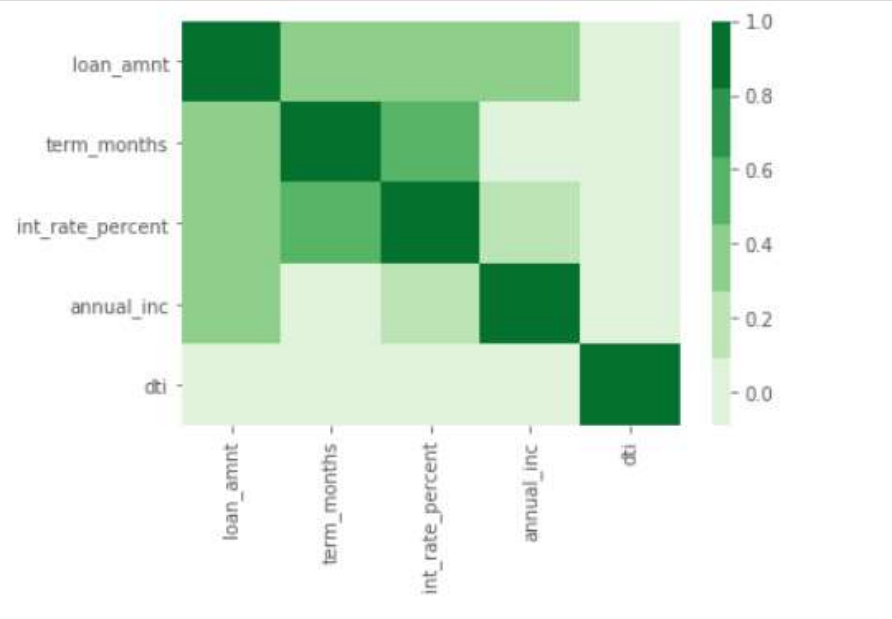
Customers having dti in range of 10 to 20 has higher chance of defaulting their loan. Lending club can use this variable to reject loans or tighten loan conditions.

Derived metrics univariate Analysis - Loan count distribution for charged off loan on Year and Month



Higher number of loans which get defaulted are being given in last 4 months of an year. This could be due to pressure on bankers to meet yearly targets and lesser scrutiny of loan applications. Lending club must plan for more audits and checks in last 4 month of approved loan applications

Bivariate Analysis – Correlation among multiple variables



No Significant info after seeing correlation of multiple variables

Conclusion

- Loans with grade B and C has higher chances of default compared to loans with status A ,F and G
- Lending club should give more loans to people whose are source verified as they have lesser chance of defaulting on their loan.
- Loans with interest rates between 10% to 18% has highest chance of getting charged off. Interest rate should be managed to be out of this range.
- Loans grant with loan amount in range of 2000 -10000 has highest chance of getting charged off
- Lending club should give more loans to people who own a house as they have lowest number of charged off loan.
- Loan applicants from CA ,CO and CT has higher chance of default may be due to state economy or some other state specific reason to be investigated. Lending club should come up with different loan policy for residents of these states.
- For customer with other variables indicating a possibility of charge off in future, Loan can be granted to applicant for longer duration of 60 month to reduce chances of defaulting on loan.
- Customers having dti in range of 10 to 20 has higher chance of defaulting their loan. Lending club can use this variable to reject loans or tighten loan conditions.
- Higher number of loans which get defaulted are being given in last 4 months of an year. This could be due to pressure on bankers to meet yearly targets and lesser scrutiny of loan applications. Lending club must plan for more audits and checks in last 4 month of approved loan applications

Thank You