Lending Club Case Study

Submitted by –
Manikanta Sai Ram Alapati.

End-Goal of the Lending Club Case Study

- **Classify** risky loan applications
- **Conclude** driving factors that are responsible for risky loan applications
- **Suggest** remediation to strengthen their prediction model

Based on which Lending Club can decide whether to reject the loan applications or at flex the interest rates, so that loans repayment doesn't go default !

Our Approach

Data Analysis

Data Analysis

Data Analysis

Data Cleaning

Removing columns having more than 30% of null-values, or with only one data,

Data Understanding

Working with the Data Dictionary and getting knowledge of all the columns and their domain specific uses

Univariate Analysis

Analyzing each column, plotting the distributions of each column.

Segmented Univariate Analysis

Analyzing the continuous data columns with respect to the categorical column

Bivariate Analysis

Analyzing the two variable behavior like term and loan status with respect to loan amount.

Recommendations

Analyzing all plots and recommendations for reducing the loss of business by detecting columns best which contribute to loan defaulters.

Data Cleaning

We followed a 3 step process-

- Following the standard checklist
  - Fixing Rows
  - Fixing Columns
  - Fixing missing values

- Standardise values
- Fixing invalid values
- Filtering Data

After following the steps we transformed an entire dataset from a row to column distribution of (**39717 x 111**) to (**36800 x 31**)!

To make our jobs easier, we segmented the variables/ columns into 3 categories -

01	Customer Demographics	• Employment Length • Annual income • Zip Code • Description
02	Loan information & characteristics	• Loan amount • Funded Amount • Interest rate • Loan Status • Loan Grade
03	Customer Behavior variables	• Delinquency year • Earliest Credit Lines • Revolving Balance • Recoveries • Application Type Loan Purpose

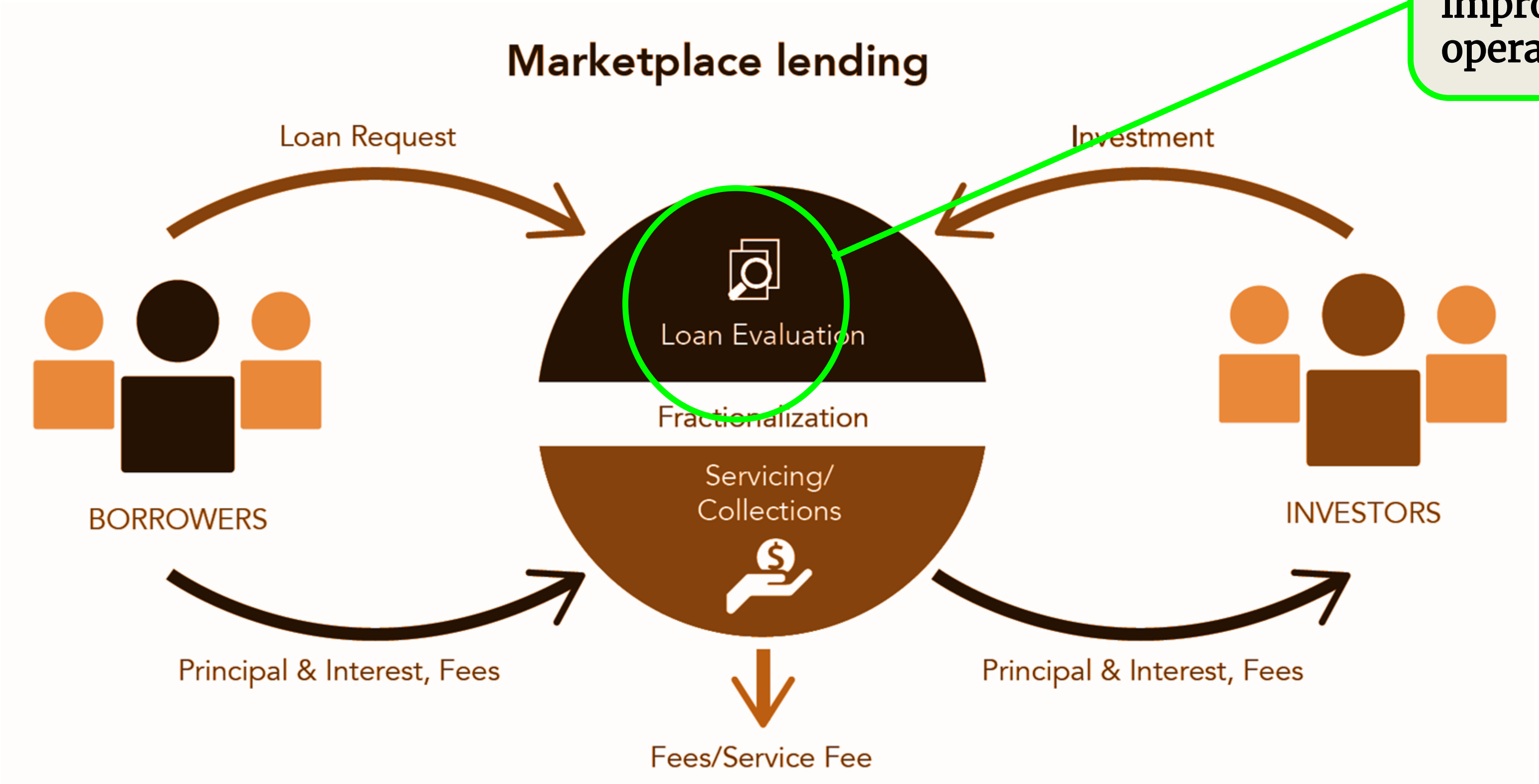
Identifying the key contributors

As our analysis is purely based on EDA, while segmenting the data we ignored few of the variables as we are would require advanced techniques to conclude our analysis. Like-

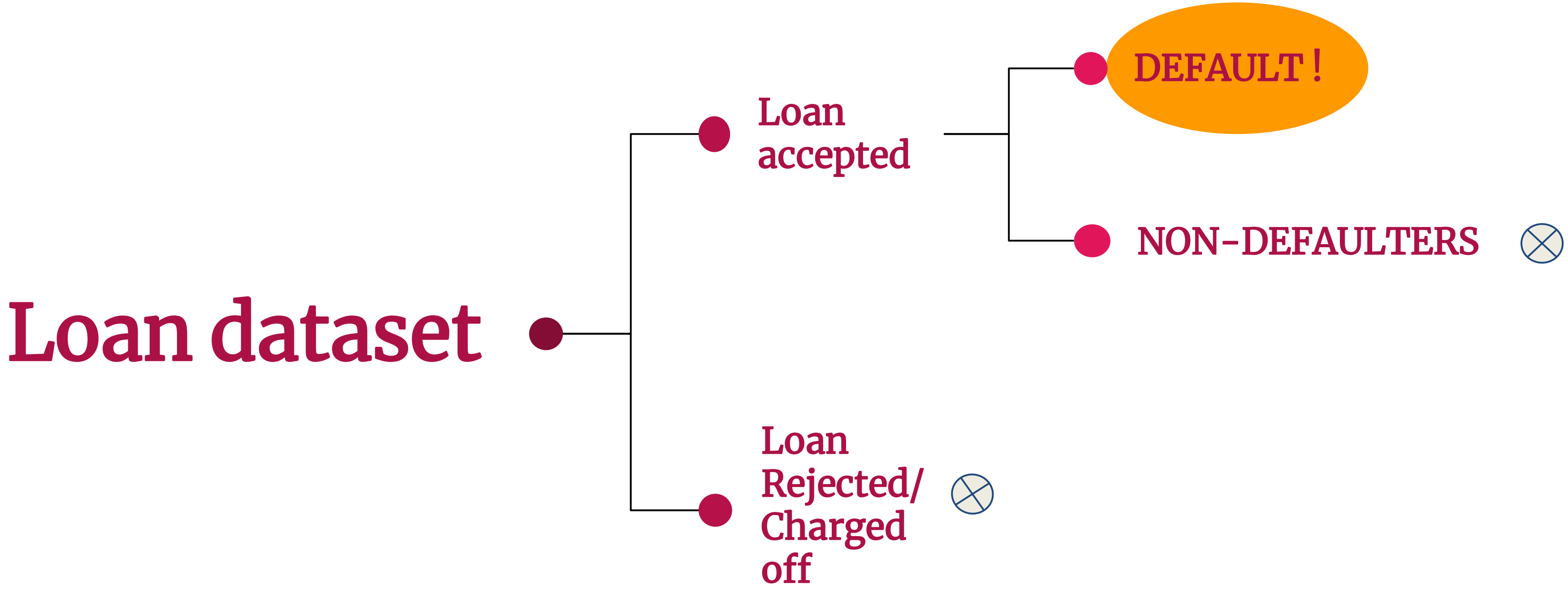
- **Description-** which could be an useful parameter, avoided, as we are yet to learn and apply NLP application
- **Customer behavior variables** - As we are not working on the predictive modeling
- **Variables of current transaction-** Even these variables can’t conclude the risk analysis of the lending club

What we understood by risk analysis for the Lending Club

Our target is to derive insights improve this operating system



How we decided which columns in loan data set qualifies our analysis-

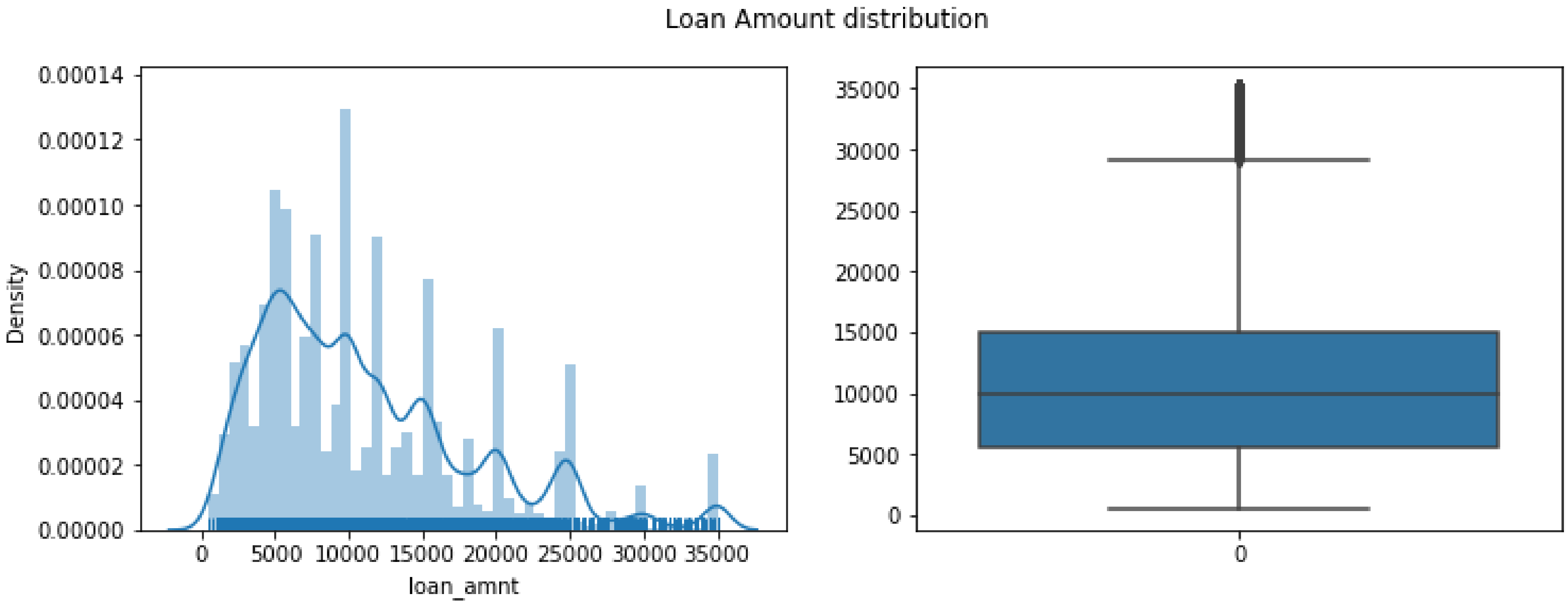


not included in the study, as it disqualifies our conditions to be included in the analysis

In order to conclude the warning signs that the loan can be a defaulter, we proceeded with the following steps of analysis -

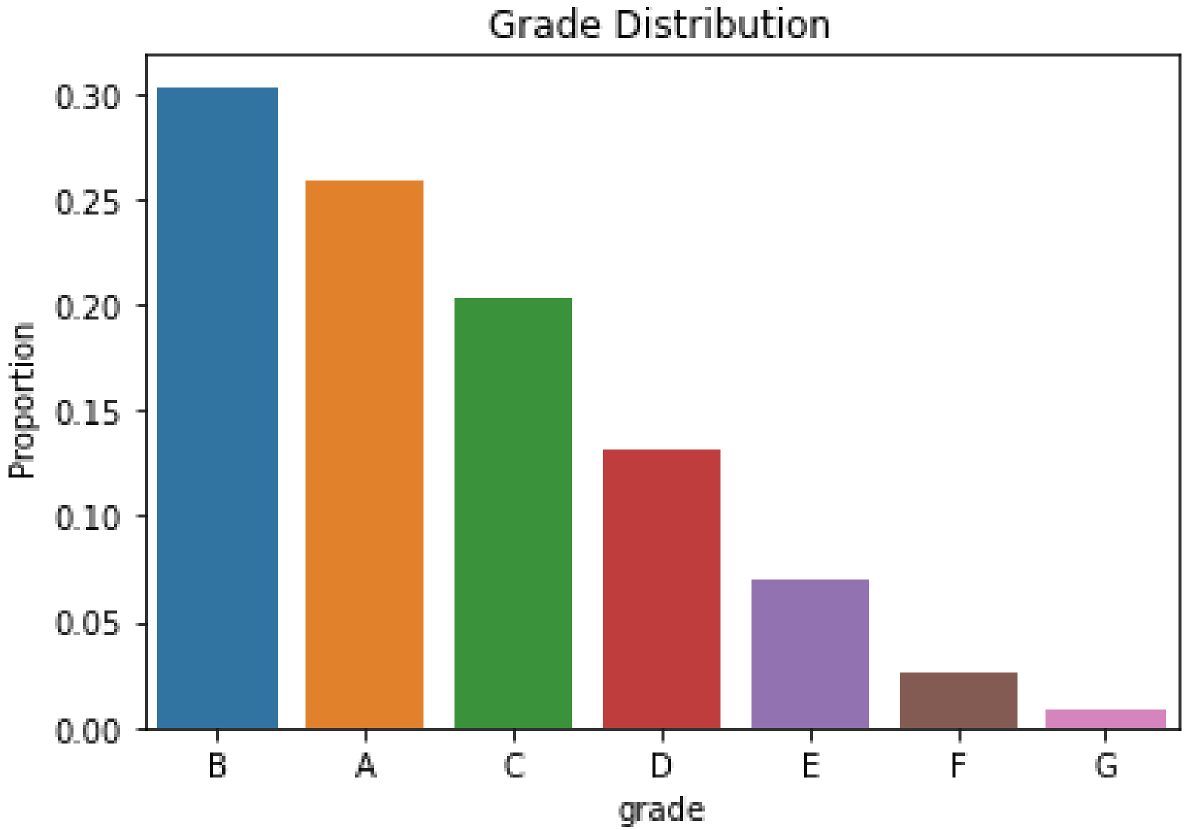
1. **Univariate Analysis-** It helped to visualize the density of the occurring values for the selected variables.

For example: we plotted a distribution and a box plot and that more number of people took loan amount of 10000, and also median of distribution is 10000. And very few people took more than 30000 loan amount.



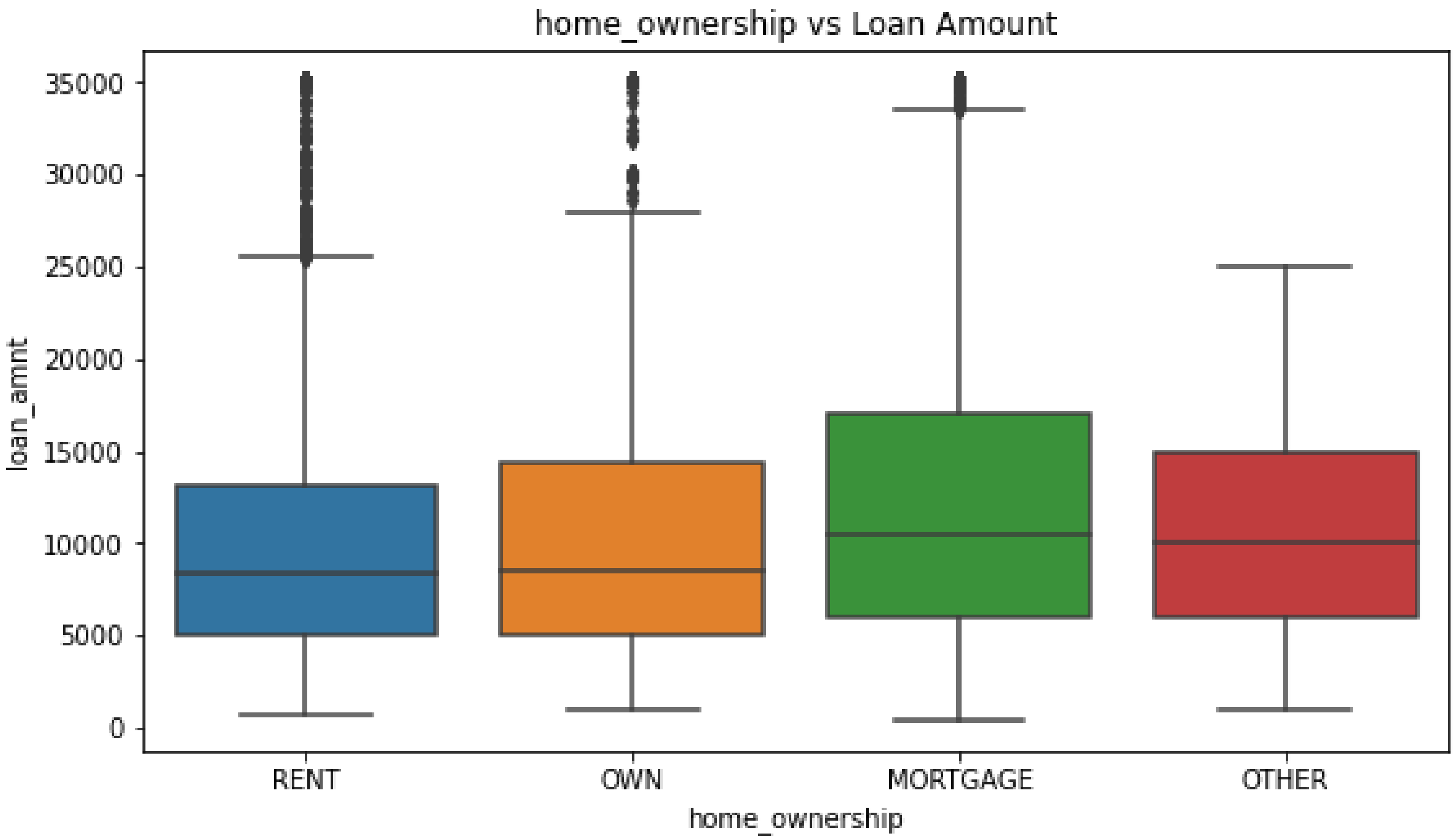
2. **Segmented Univariate Analysis-** It helped to visualize the frequency of bucketed data points and how it varies across the bucketed segments.

For example: To check the frequency of the graded borrowers of Lending Club (LC) we choose a barplot and we observed that most borrowers fall under A and B grades then other grades (as the risk increases with from categories, D, E, F, G)



3. **Bivariate Analysis-** It helps us infer correlation between two variables and concludes simple hypothetical associations.

For example: To check the association between home ownership and loan amount, we used boxplot and concluded that more borrowers are from MORTGAGE and also the median loan amount also high for MORTGAGE owned borrowers.



RESULTS - Prepared a comparative table for ease of understanding

	OBSERVATIONS
Terms vs Loan Amount	Higher amount loans have high tenure i.e, 60 months
Grade vs Loan Amount	Grade 'F' and 'G' have taken max loan amount. As Grades are decreasing the loan amount is increasing.
home_ownership vs Loan Amount	The median loan amount also high for MORTAGE owned borrowers.
verification_status vs Loan Amount	And most of borrowers are verified for borrowing loan >9k
loan_status vs Loan Amount	Charged Off loans have higher amounts than Fully Paid ones.
purpose vs Loan Amount	More loan amount is from small_business and debt_consolidation.
emp_length vs Loan Amount	More borrowers are from 10+ years and least is <1 year
Issued year vs Loan Amount	#The meadian loan amount in each year did not change much but the distribution is more spread as the years increase, which means people have taken different loan amounts in each year. #Few Loan borrowers took higher loan amount in 2008 and 2011 which are plotted as outliers. #Loan borrowers took almost similar amounts in all the months except in December, people have taken higher amounts as distribution is high above median.

NB: Following are the few of the observation, which were inevitable to decide the defaulter's fate

	OBSERVATIONS
Grade vs annual income	Comparatively Annual income is higher for lower grades.
home_ownership vs annual income	The home_ownership status for MORTGAGE has higher income.
verification_status vs annual income	The income source was verfied for most of the borrower's who had higher annual incomes.
loan_status vs annual income	Current status of the loan is Fully paid for most of the borrower's who had higher annual incomes.
purpose vs annual income	A category belonging to renewable_energy, small_business and home_improvement have higher annual income provided by the borrower for the loan request.

dti (Debt to Income Ratio)	OBSERVATIONS
Terms vs dti	dti is bit high for people who got more tenure i.e., 60 months.
Grade vs dti	A' Grade borrowers are having low dti than Other grades. dti should be low for having high repayment percentage.
verification_status vs dti	People in OTHER home_ownership has less dti than others. This is may be bacause other people have mortgage and home loans.
loan_status vs dti	Borrowers with high dti has bit more probability to default
purpose vs dti	People who took loan for credit_card and debt_consolidation purpose has more dti than other purposes.
emp_length vs dti	The dti is pretty much similar for borrowers with all the employment length.

Conclusions

- Lending club should reduce the high interest loans for 60 months tenure, they are prone to loan default.
- Grades are good metric for detecting defaulters. Lending club should examine more information from borrowers before issuing loans to Low grade (G to A).
- Small business loans are defaulted more. Lending club should stop/reduce issuing the loans to them.
- Borrowers with mortgage home ownership are taking higher loans and defaulting the approved loans. Lending club should stop giving loans to this category when loan amount requested is more than 12000.
- People with more number of public derogatory records are having more chance of filing a bankruptcy. Lending club should make sure there are no public derogatory records for borrower.