ANUP SINGH

CHHANDITA CHAVAN

LENDING CLUB CASE STUDY

Abstract

Business Objective

Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

In this business, loan borrowers who 'default' cause the largest amount of loss to the lenders. The company, thus wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Strategy Incorporated

The company wants to focus on the risk factor occurred due to acceptance of a loan application of a borrower.
The risk of default can be predicted by finding patterns between these two attributes:

1. Consumer Attributes

The consumer attributes would signify which features of a consumer act as a risk factor to default

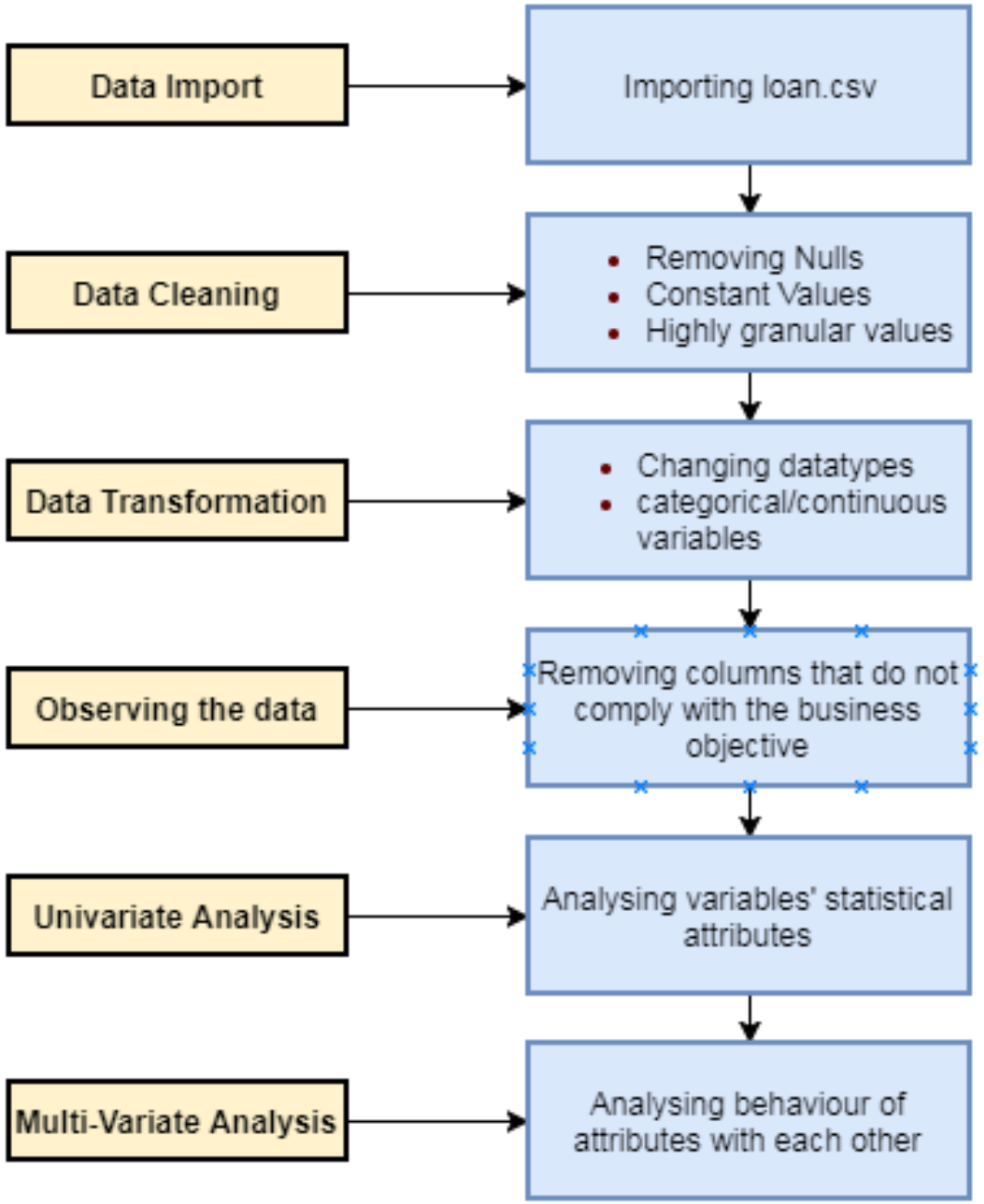
2. Loan Attributes

The loan attributes would signify which features of the loan lent would act as a risk factor to default.

Assumptions

Default / Charged off Loans with other parameters would give better risk insights and patterns.

Approach


Data Import
Importing loan.csv
Data Cleaning
Removing Nulls
Constant Values
Highly granular values
Data Transformation
Changing datatypes
categorical/continuous variables
Observing the data
Removing columns that do not comply with the business objective
Univariate Analysis
Analysing variables' statistical attributes
Multi-Variate Analysis
Analysing behaviour of attributes with each other

Data

The dataset initially has 111 attributes and 39,717 rows in total

The number of relevant attributes were brought down to 20 after performing the following processes:

- Dropping columns with only NULLs
- Dropping rows with high percent of missing values from the columns
- Dropping columns with only one/two constant values
- Dropping columns with too high granularity for analysis

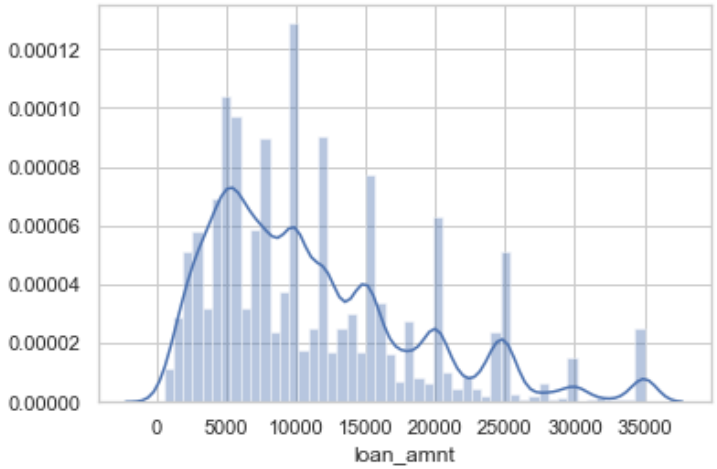
The data has both consumer-based attributes and loan-based attributes.

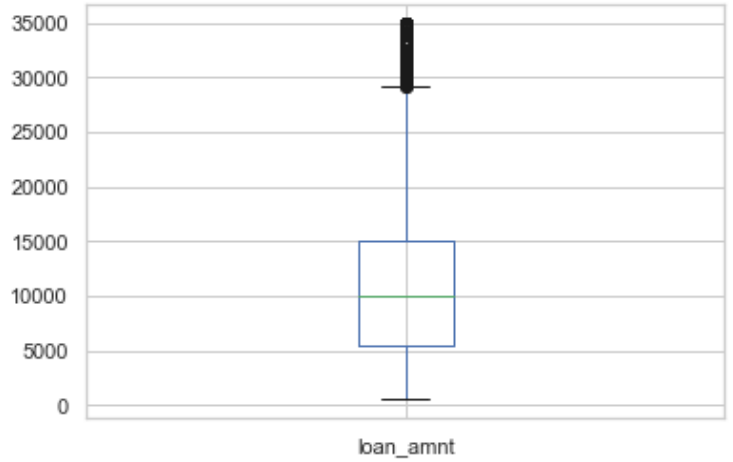
Out of these attributes, a total of 7 continuous variables and 14 categorical variables were finalized. Out of the 14 categorical variables too, some variables are numeric. But due to less number of values in these fields, the numeric variables are treated as categorical variables.

Following variables were highly correlated:
1) loan_amnt, funded_amnt, funded_amnt_inv, installment, total_pymnt, total_pymnt_inv, total_rec_prncp and total_rec_int
2) out_prncp and out_prncp_inv
3) total_acc and open_acc
4) recoveries and collection_recovery_fee

Uni-Variate Analysis

Loan Amount

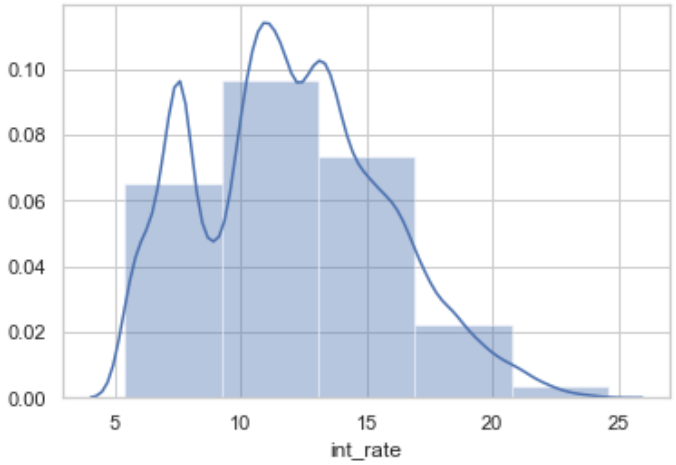


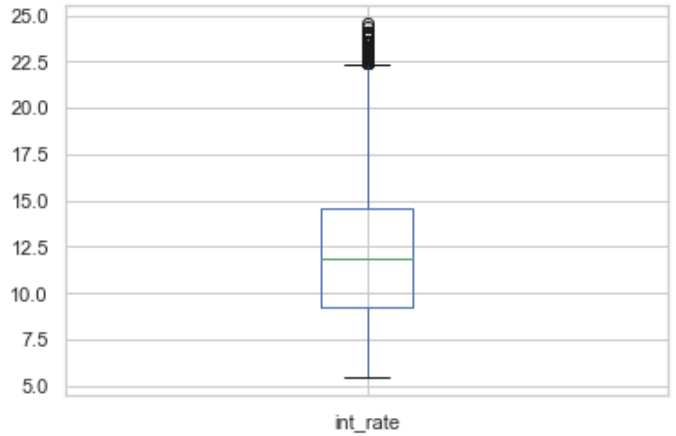


Inferences

- The data is left-skewed as the mean and median do not coincide
- The highest number of loans issues are for 10,000 dollars
- The highest loan amount issued is for 35,000 dollars
- The lowest loan amount issued is for 500 dollars
- The data does not contain any outliers

Interest Rate

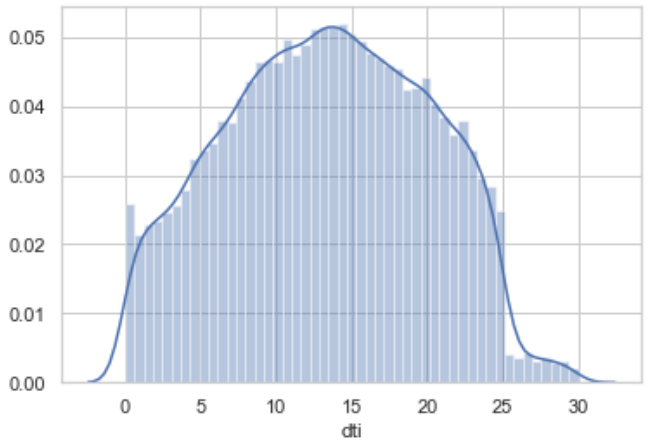


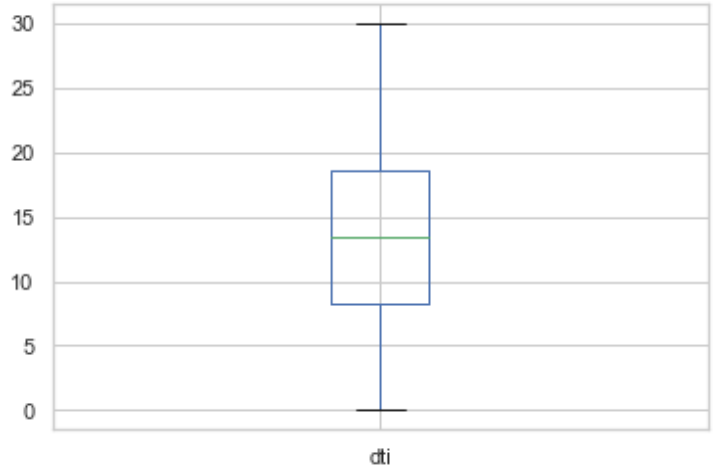


Inferences

- The data is slightly left-skewed as the mean and median do not coincide
- The highest number of interest rates charged are between the range of 10% to 13%
- The highest interest rate charged is 24.59%
- The lowest interest rate charged is 5.42%
- The data does not contain any outliers

Debt to Income Ratio

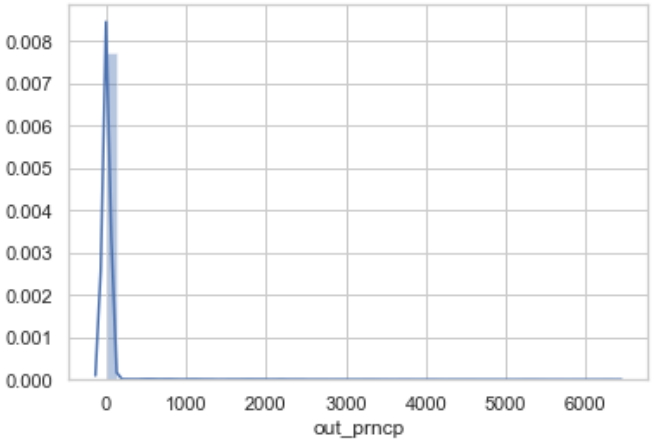




Inferences

- The data is not skewed as the mean and median are nearly equal
- According to investopedia, lenders prefer debt to income ratio less than 36%. In our data, all our data points are lower than 36% which means that borrowers' loans would not be rejected by lenders for the borrowers in this data
- For the data points greater than 25%, the count of debt to income ratio drastically decreases. These borrower's loan must be specifically looked into as their debt per monthly income is greater than others.
- There is a huge spike at 0% debt to income ratio. This could mean that these borrowers for loan are dependents and may be of a young age as they do not have debts to pay.

Outstanding Principal



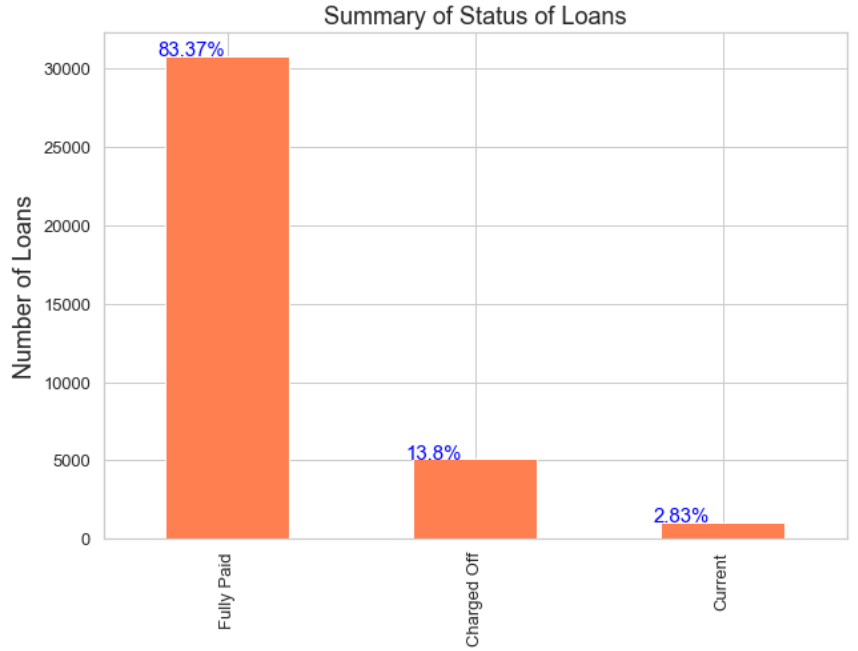
Inferences

We can draw the following conclusions for outstanding principal data:
The data is highly skewed at 0. This basically means that nearly all borrowers have 0 outstanding principal, only a few borrowers are left who are yet to complete their payment of funded amount

Loan Status

Inferences

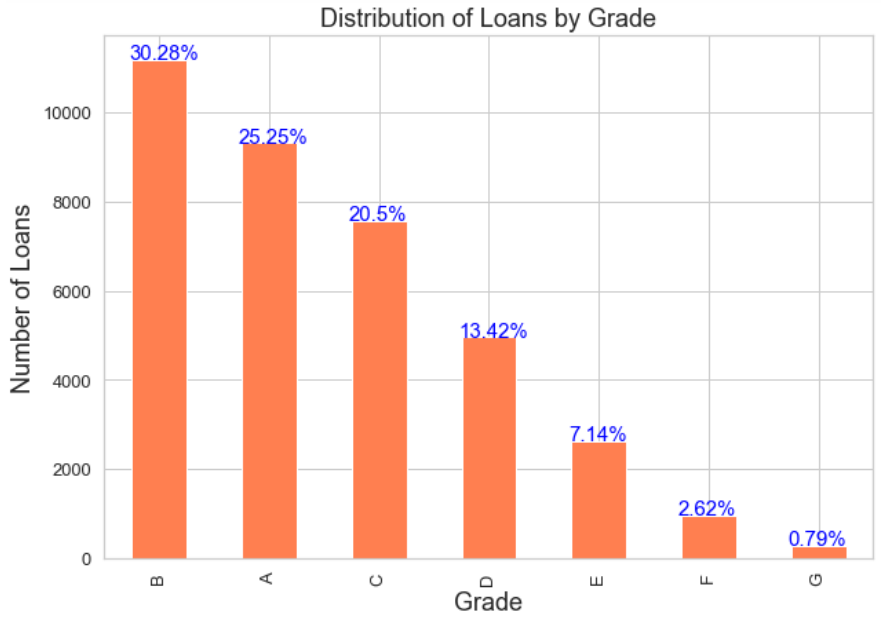
- Charged-Off loans constitute a total of 13.8% of all the loans issued in the data
- Most of the loans are paid-off as the Fully paid loans constitute of 83.37% of the data
- The major focus in bi-variate analysis would be to focus on charged-off loans and determining which factors lead to loan defaults



Grade

Inferences

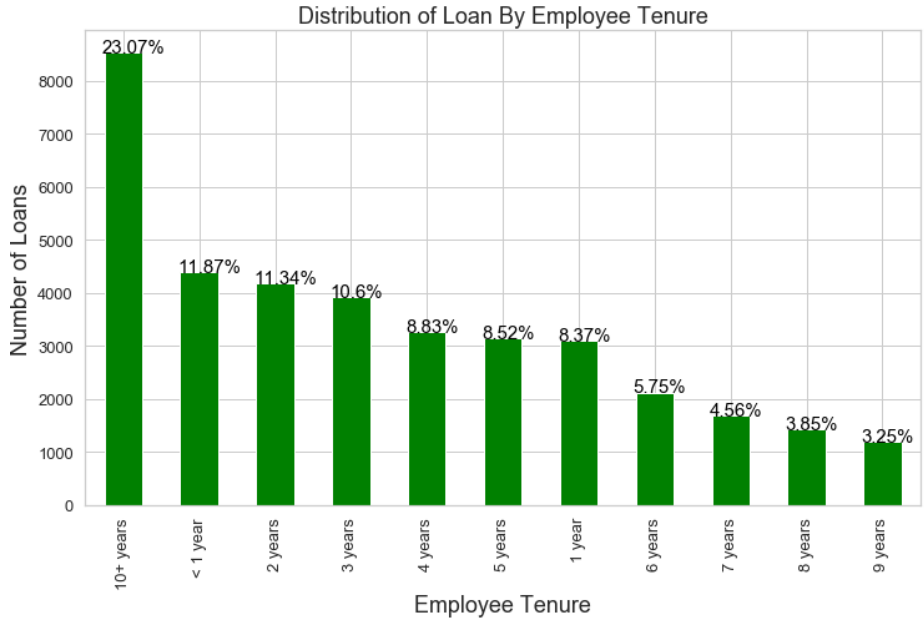
- Highest number of loans in the data are graded 'B'
- Here, Grades A, B and C contribute to nearly 70% of the loans, which infers that nearly 70% of loans have a very low probability to default
- Lowest number of loans in the data are graded 'G'. Considering that grade 'G' signifies the worst loans sanctioned or the loans which would have the highest probability to default, these numbers are quite low.



Employee Tenure

Inferences

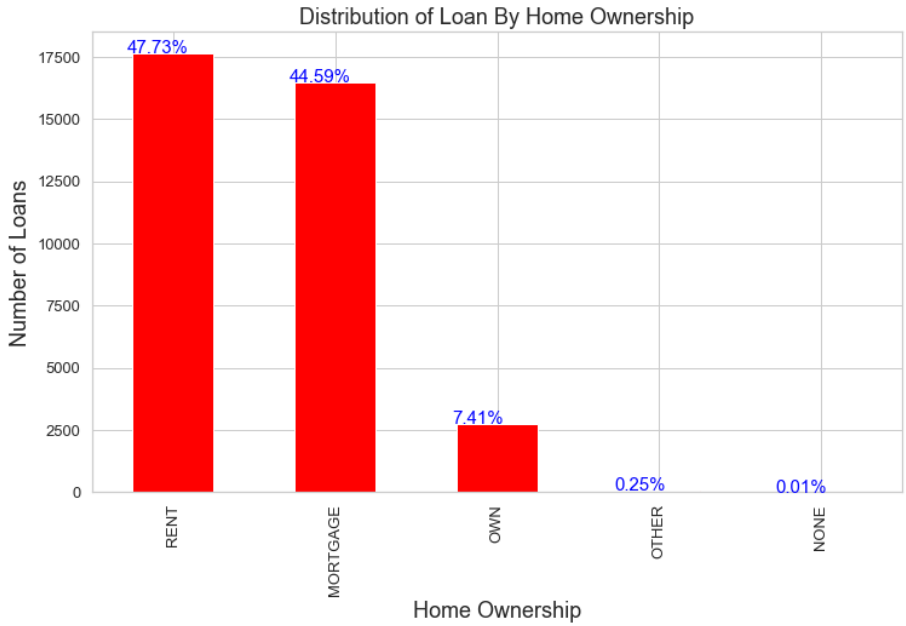
- Its obvious to note that a borrowers with 10+ years of experience would have maximum occurrences in the data. This can be due to two reasons, either because 10+ years covered all employee tenures greater than 10+ years and so the data for this set is large. The other reason could be a borrower with 10+ years of experience would earn comparatively higher to his/her counterparts who have less tenure
- Another interesting observation here is that there is monotonous decrease in number of loans issues after 2 years of employee tenure.
- The above behaviour is not the same when we look into the data for <1 years, 1 year and 2 years. The number of loans sanctioned decreases from <1 year to 1 year and abruptly increases from 1 year to 2 years.

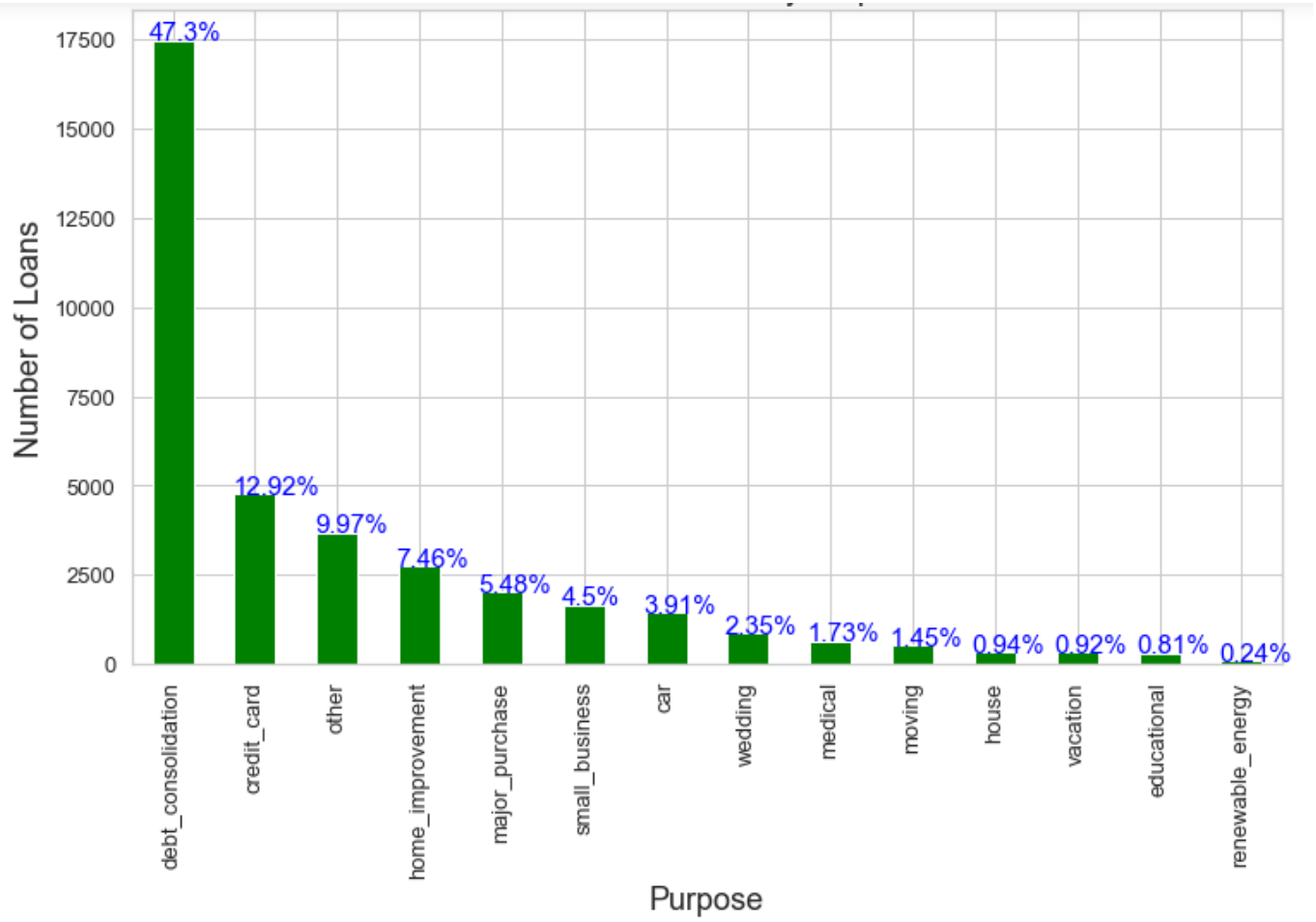


Home Ownership

Inferences

- Nearly 94% of the loan borrowers do not own a house, which makes it more probable for these borrowers to have a greater debt to income ratio every month, which in turn leads to greater probability of default
- There are 7.41% of borrowers who own a house and thus are less probable to default





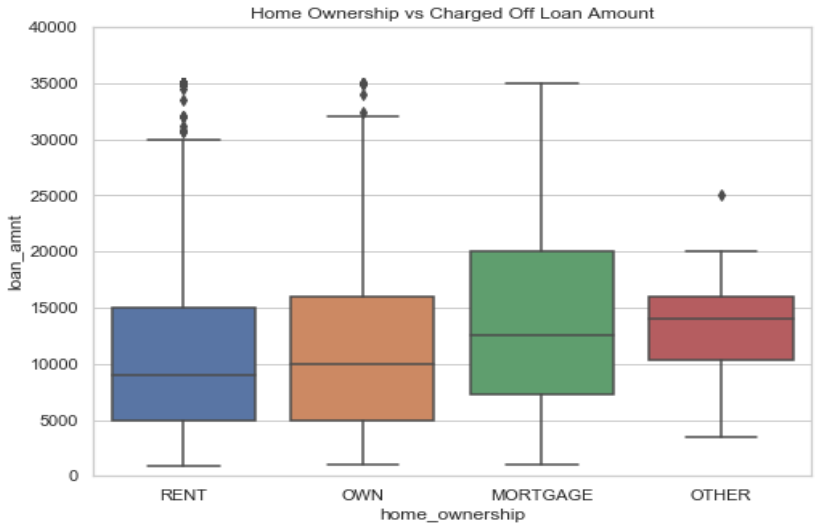
Employee Tenure

Inferences

- Nearly half of the loans are borrowed for the purpose of debt consolidation
- Its interesting to note that, although, most of the borrowers in our dataset do not own a house, loan for buying a house in the dataset is only 0.94%

Bi-Variate Analysis

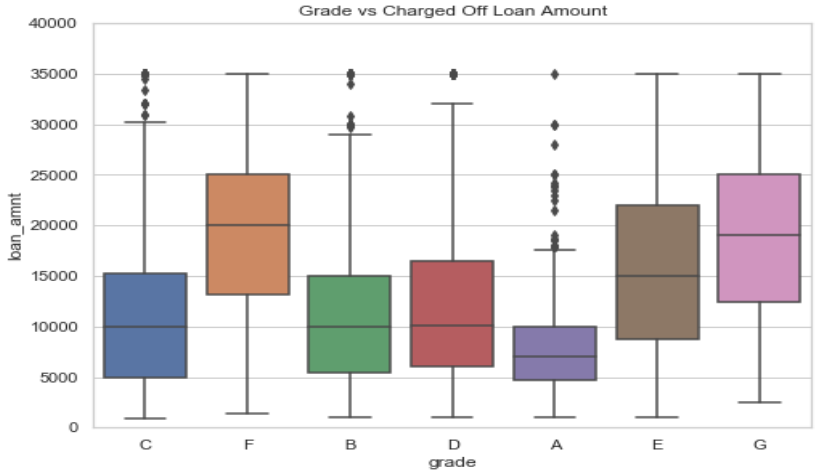
Home Ownership vs Charged Off Loan Amount



Inferences

- A steady increase in the median when we compare home ownership with charged off loan amount.
- Probability of default is higher on loan sanctioned to customers with home ownership - "MORTGAGE".

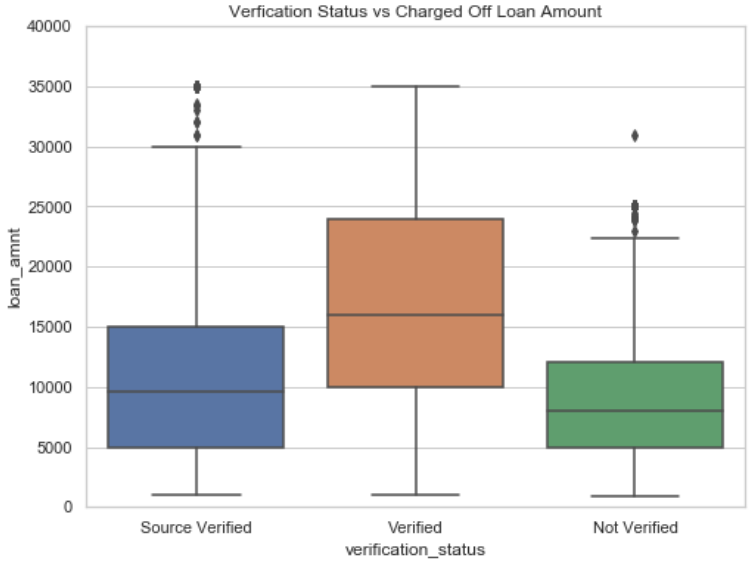
Grade vs Charged Off Loan Amount



Inferences

- Default on Grade F,G and E more prominent.
- There could be elements of customer behavior or we need an in depth analysis further with employee experience.
- The data may contain outliers when compared to employment years of experience.

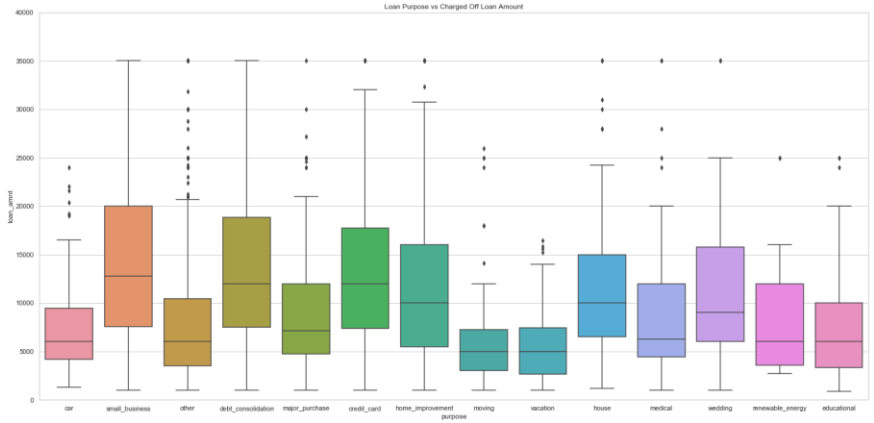
Verification Status vs Loan Amount



Inferences

- The amount of default and loss to the lender happens mainly for verified customers.
- The amount of loss related to verified customers is huge. This seems to be a big risk for lenders in terms of loss.
- Lenders can calculate risk appetite of verified customers on regular basis.

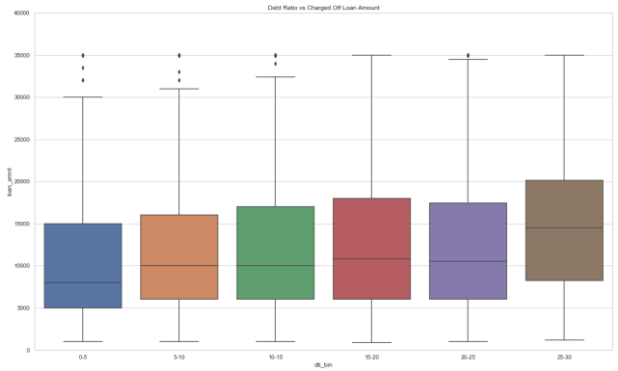
Purpose vs Charged off Loan Amount

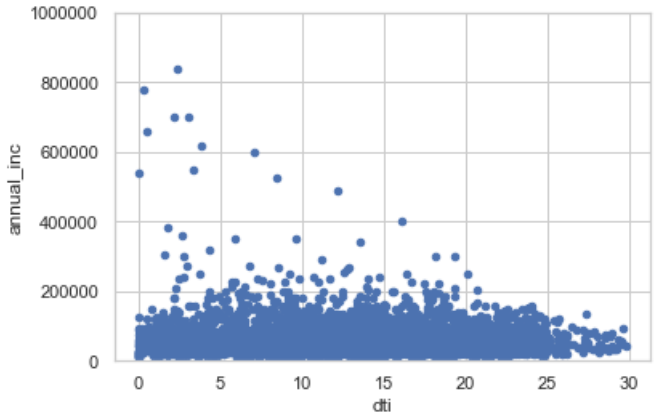


Inferences

- The amount of default and loss to the lender happens mainly for customers who have borrowed for small business, debt consolidation and credit card.
- Lenders should identify the reasons why such customers are not able to repay.
- A regular credit score or history could be monitored for such customers

Debt Ratio vs Charged Off Loan Amount & Debt Ratio vs Annual Income of Charged off Loan

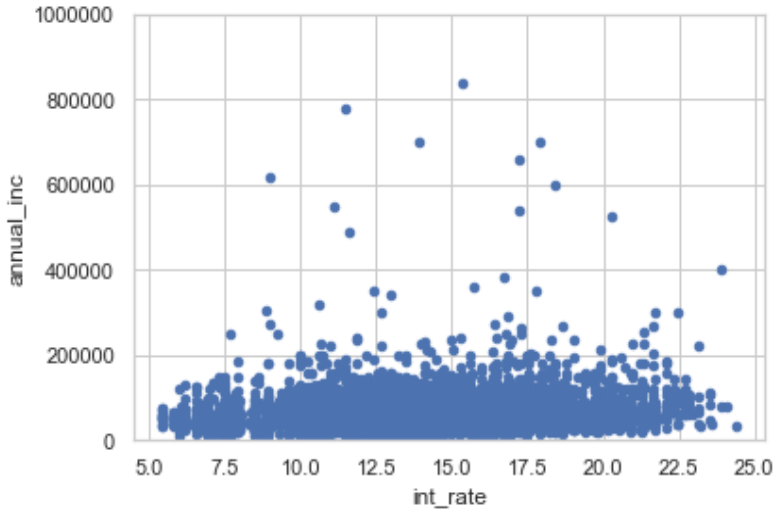




Inferences

- Customers with higher debt ratio will default regularly.
- Scatter plot indicates the debt ratio is primarily higher for low income earners and are quite likely to default.
- Lenders could do carry out due diligence on these factors during verification process.

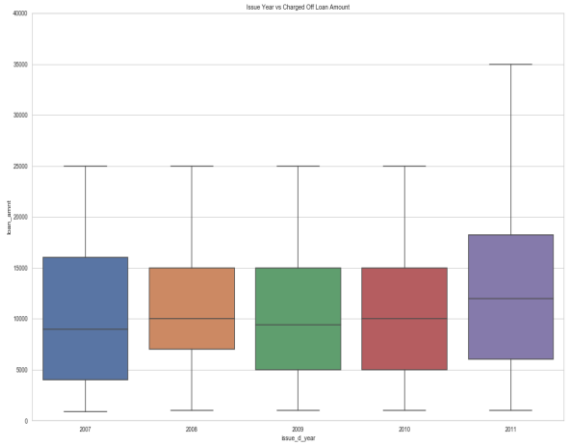
Interest Rate vs Charged Off Loan Amount



Inferences

- Customers with higher income will have less interest rate. E.g. Earning above 200000 USD.
- Scatter plot indicates the higher interest rate is primarily higher for low and mid income earners less than 200000 USD and quite likely to default.
- Lenders could do a regular review may be quarterly or six monthly to track the payments and engage with the customers to find out scenarios like financial situation, medical conditions, bankruptcy etc.

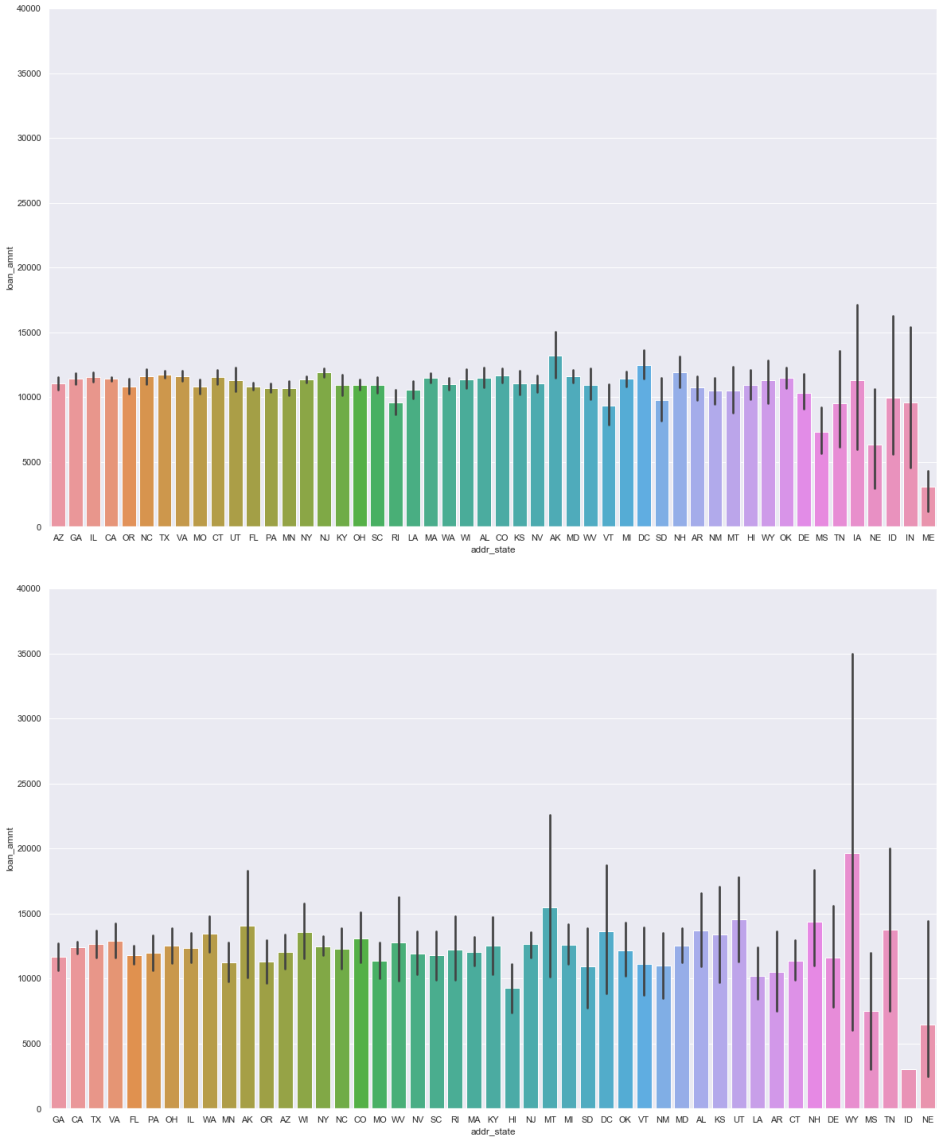
Loan Issue Year vs Annual Income vs Interest Rate vs Pub_Rec vs Charged Off Loan Amount



Inferences

- Customers with higher debt ratio will default regularly.
- Box plot shows default loans were constant from period 2007-2010 and there is a spike in 2011.
- Scatter pairplot indicates the range of interest rates offered to the customers went highest in 2011.
- Scatter pairplot also indicates there is a steep increase in derrogatory public records from 2009-2011.
- Scatter pairplot also indicates there is a pool of derogatory public records atleast between 0 to 1 over the interest rates offered.
- Scatter pairplot also indicates annual income is indirectly proportional to derogatory public records. This derogatory public records increases for annual income earners less than 2,000,000 USD.

State vs Charged Off Loan Amount vs Total Loan Amount



Inferences

- Bar plot shows default loans were massive in WY
- Default loans were substantial across MT, DC, UT, NH, TN, AK, CO, WV.
- This could indicate similar customer patterns or behavior.

Multi-Variate Analysis

Multivariate Analysis: Purpose vs Grade vs Emp_Length vs Charged Off Loan Amount

Inferences

- First Bar plot shows default loans were the most in home_improvement, debt consolidation and credit cards. The default loans is average over 15000 USD across other categories of purpose.
- First Bar plot also indicates the default is highest among customers employed at Grade-'G' and 'F'.
- Second Bar plot indicates the default is maximum among employees/customers with 9 years of work experience with purpose as house and also in mid level and 10+ years.
- Some of the scenarios might be due to already existing house on loan or portfolios and other factors could be marital status, family growth or change in job circumstances.
- Second Bar plot indicates the default is maximum among small business and house and then debt_consolidation, credit_card and major purchases.

a

Conclusion

Final Conclusion

- Customers attributes/variables like purpose, dti, annual income, employee length indicates there is a strong loss incurred due to default loans.
- Lenders are facing a huge loss based on the insights derived from Pub_Rec and Int_Rate. This also has an association with source as verified. It seems the due diligence like strong credit checks and monitoring of payments is a must to avoid default.
- Year 2011 has a massive spike on default primarily due to hike in interest rates.
- Customer behavior with senior level experience should be tracked. Multivariate pattern indicates these customers default despite their income being higher.
- Regular credit monitoring should be done across the customers of states in default.
- The other rationale to above conclusion would be lending should be done in small chunks based on customers credit score with variable interest rates.
- Lenders should calculate frequent risk appetite state wise and also provide a loan which suits on the basis of similar borrower profiles.