upGrad

Lending Club Case study

EDA analysis for a lending company

Ganesh S K, Sunayna Talreja

The problem

Company

Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.

Problem statement

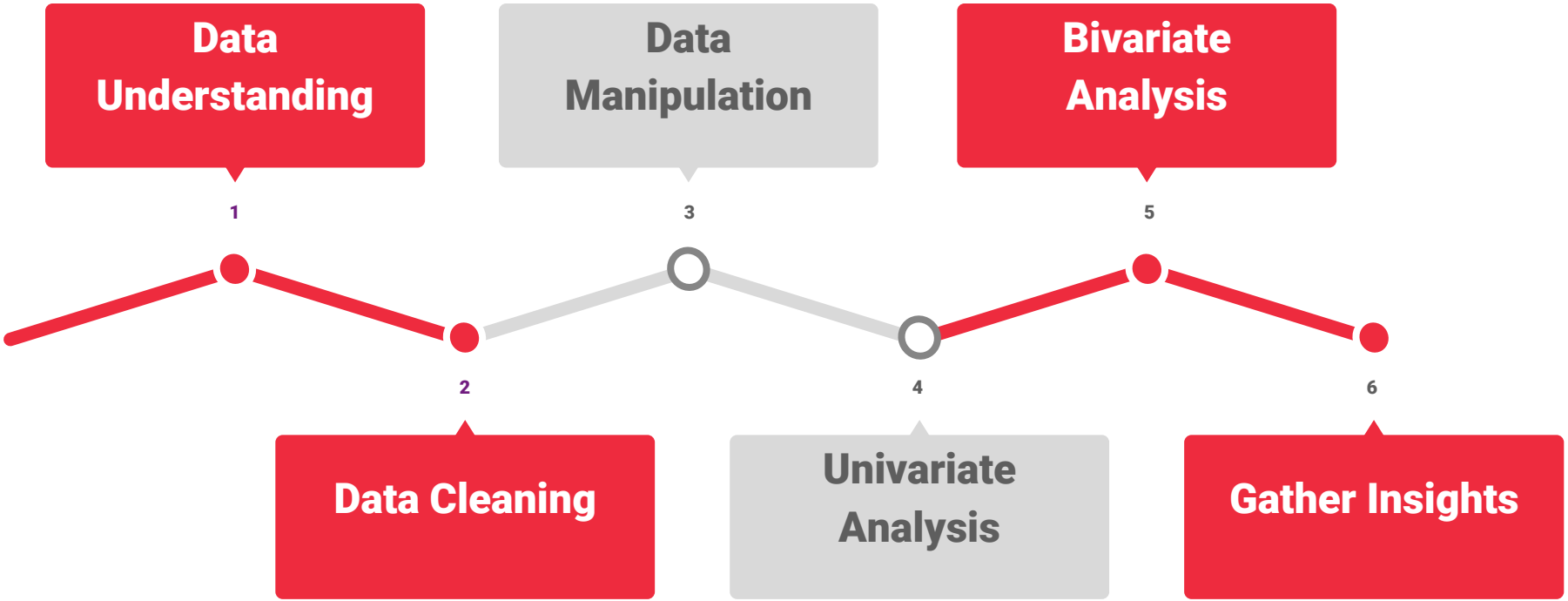
Two types of risks are associated with the bank's decision:

- 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 a financial loss for the company

Goal

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the **variables** which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Our approach



upGrad

Data Understanding

1. Characteristics of Data set
2. Categorical & Continuous data
3. Approach to treating Data

Characteristics of Data set

The Data set contains **39717 rows** and **111 Columns**. The overall schema has following types of data -

1. Customer related information
   Eg: Annual Income, Home ownership, Annual income, Employment Length etc.

1. Loan related information
   Eg: Loan amount, Interest rate, Grade, Sub grade, DTI, Loan date, Verification etc.

1. Customer behavior information
   Eg: Delinq, Credit line, Open account, public records, collection recovery fee, application type , recoveries, last payment date, etc.

Takeaway: With the major goal being identifying variables contributing loan default on a **new customer,** customer behavior variables can be left out as they won't be available during the time of onboarding a new customer

Categorical & Continuous data

The Data set contains **24 Categorical** & **87 Continuous** variables

There are further scopes to refine the data (Please see the XXX section for refined shaped of the table, categorical and continuous variables)

The variables namely Interest rate, Issue date, Employment length etc. are Categorical but can be manipulated to continuous

The variables namely Annual Income, Loan amount etc. are given Continuous but can be treated categorical

Variables like tot_hi_cred_lim, total_bal_ex_mort, total_bc_limit, total_il__high_credit_limit are NULL and can be removed

Takeaway: As next steps, we will clean and manipulate data by removing columns lacking categorization, relevance and value addition

upGrad

Data Cleaning & Manipulation

1. Removing columns
2. Data Imputation
3. Derived metrics

Removing columns

1. Removing columns where 90% of values are null
    a. Eg: tot_hi_cred_,im, total_bal_ex_mort etc.
    b. 56 variables are removed (out of 111)

1. Removing columns where data is homogenous (lack categorization)
    a. Eg: url, policy_code, application_type etc.
    b. 14 variables are removed (out of 55)

1. Removing non deterministic customer behavior related variables
    a. Eg: delinq_2yrs, earliest_cr_line, out_prncp, revolv_util etc.
    b. 19 variable are removed (out of 41)

Takeaway: The data is now refined to have 39717 rows and 22 columns

Data Imputation

1. While the Employee title is categorical, the data is widely diverse and are insignificant. The null values in Employee title can be replaced as “unknown”

1. Public record bankruptcies, while we are unsure how critical the data can be. We see strong correlation with pub_rec (0.84). Imputing the pub_rec_bankruptcies by pub_rec

1. Employee length might be a significant data and any manipulation can lead to severe impact with modelling. Additionally, there is no strong correlation of the data with other variables. So replacing null values with “unknown”

1. Further non deterministic information such as Loan status = “current” will be removed as we are unsure of the final result with ongoing loans

Takeaway: Data imputation have been done on missing rows and the treatments used here are deterministic nature, correlation and marking data as unknown.

Derived metrics

1. Following from Data understanding, here we are deriving new metrics that are transformed from categorical to continuous, continuous to categorical

1. The newly made categorical variables are used as “bins”

1. Employer name is replaced by Employer Title post Sep 2013. However this doesn’t affect our data set much as the data is only available for 2010 and 2011

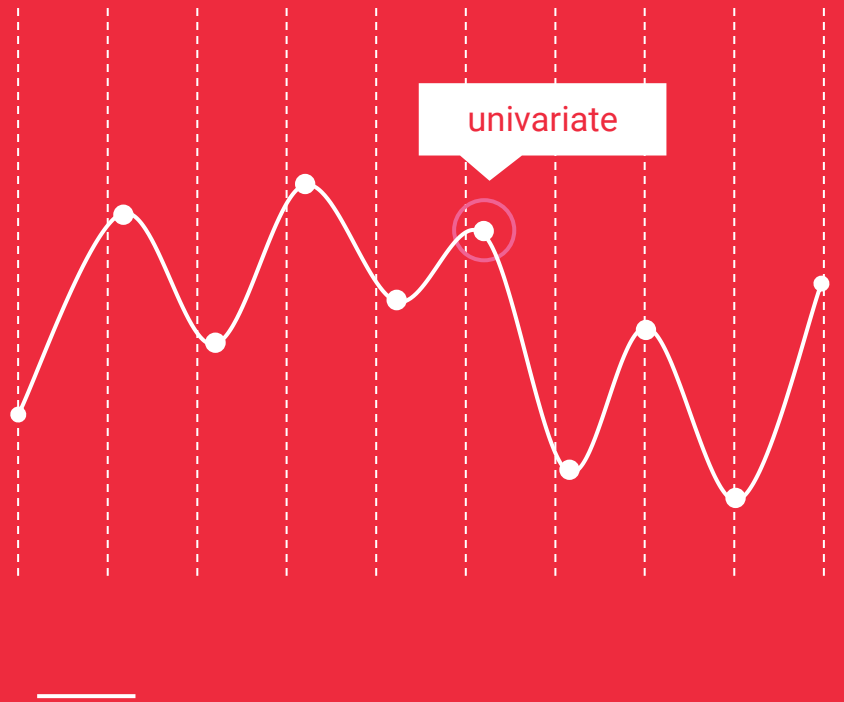
Takeaway: Variables such as annual income, interest, issue date are impacted in this data treatment

Data Analysis

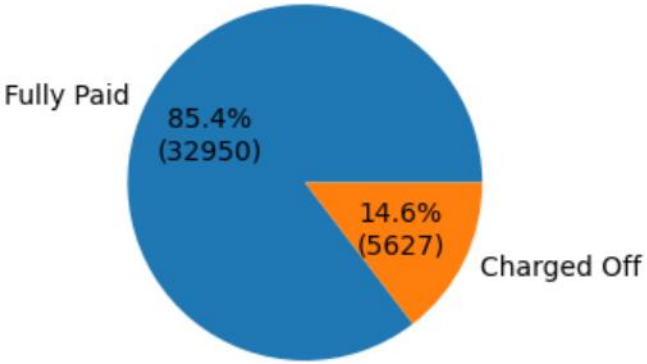
Analysis

- The problem statement of lending club is to analyze how data variables show a pattern which influences the tendency of defaulting.
- During univariate analysis we have created Distribution plots to check out the distribution of all the driver variables and Box plots to detect the Outliers
- Performed the Multivariate analysis to understand how different variables interact with each other.

Univariate Analysis



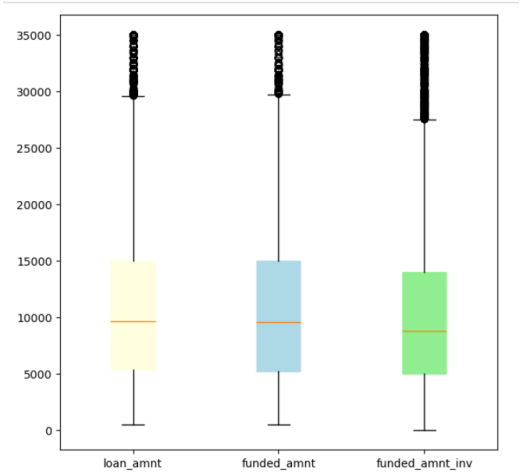
Loan Status

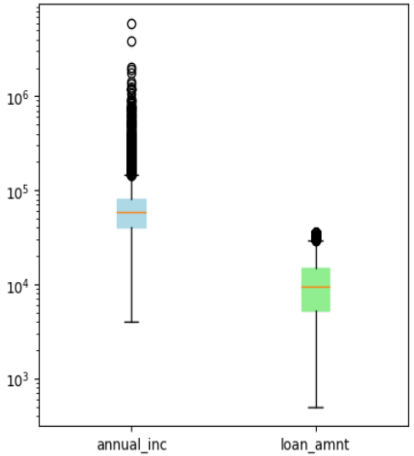


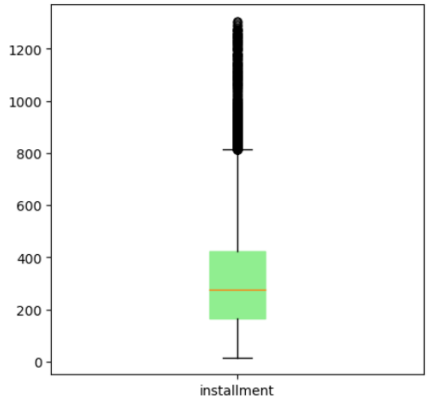
Takeaway:

- We have considered the comparison between paid and charged off, and ignored current as they were irrelevant for the analysis
- Around 85% loans have been fully recovered
- Around 14% of loans are defaulters

Outliers Removal

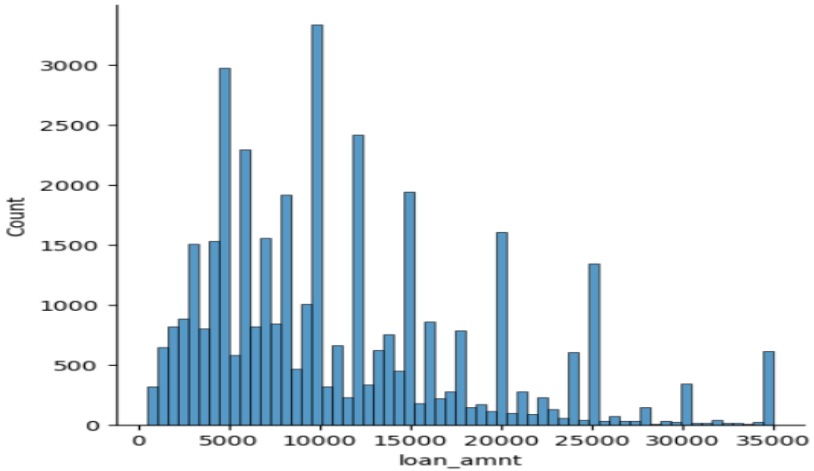






Takeaway:
Outlier analysed for the related variables and values beyond 99.5 percentile are removed.

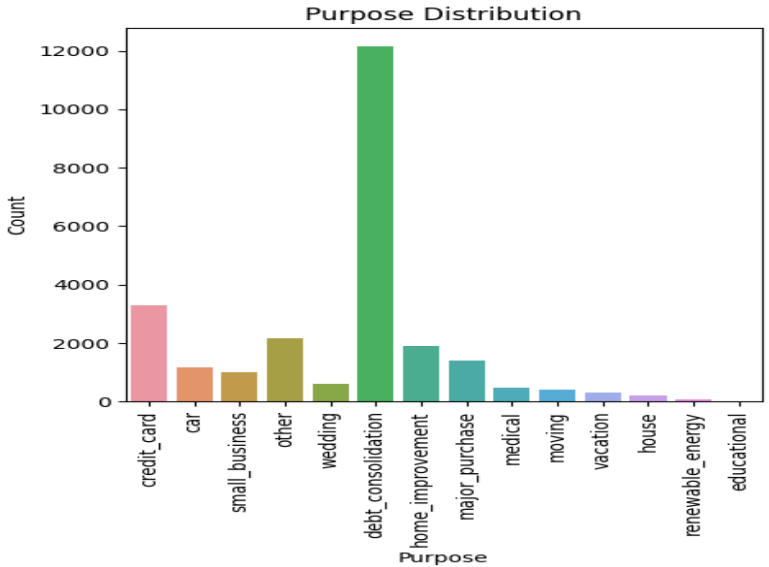
Loan Amount Analysis



Takeaway:

- Distribution of loan amount and most requested loan amount details
- The distribution is left skewed and most of the loan amount is the range of 10K

Purpose Distribution



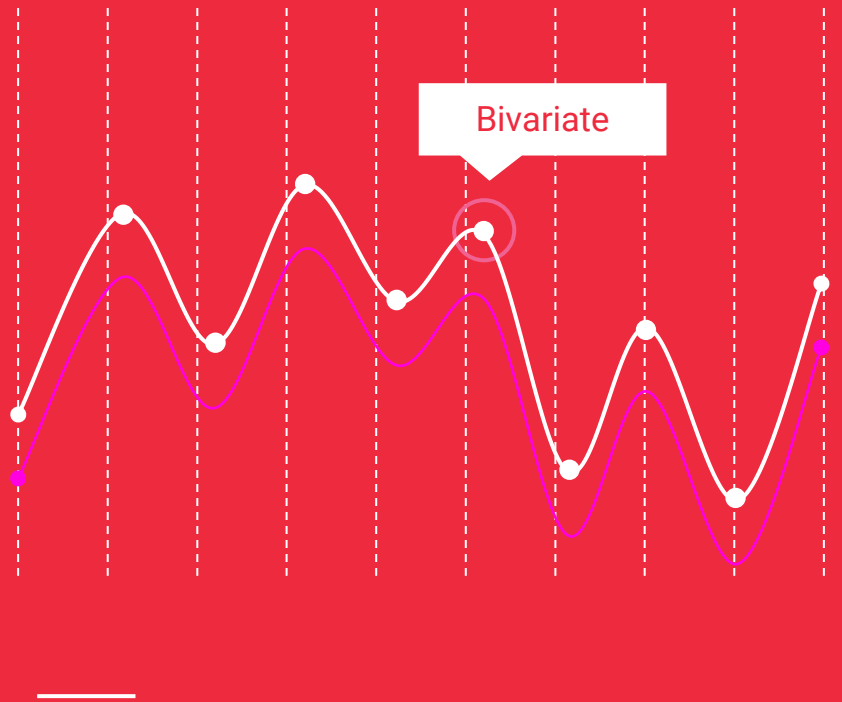
Takeaway: Most of the loans has been on Debt consolidation category followed by credit card (non collateralized debts)

Employment Length Distribution

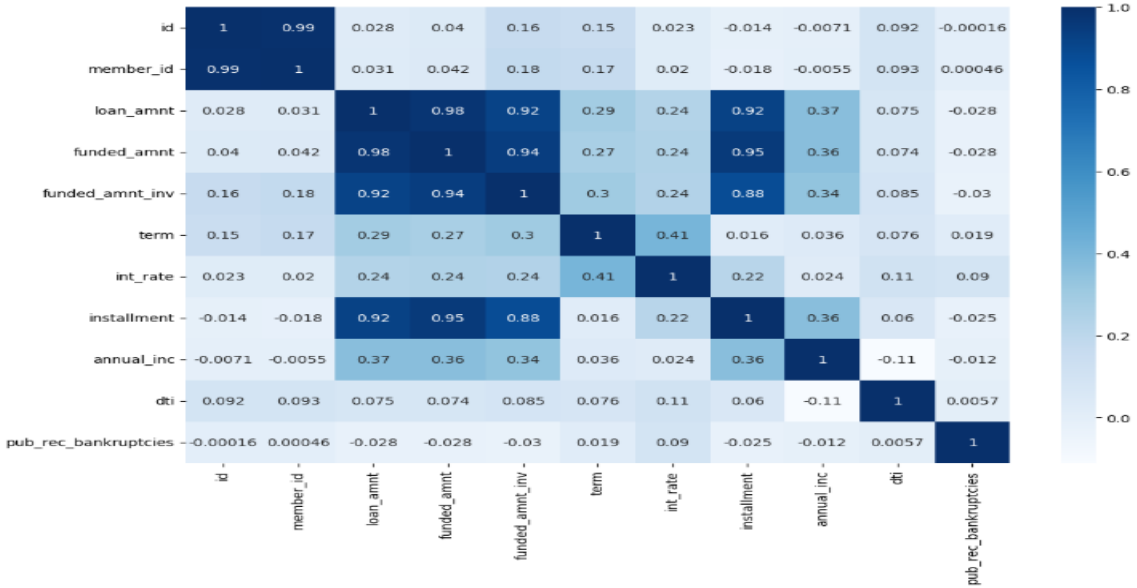
Takeaway:

- Loans has been availed mostly by salaried employees with 10+ year experience.
- In contrary, the next loan takers fall within less than 5 years of experience indicating relationship with applicant life stage

Bivariate Analysis



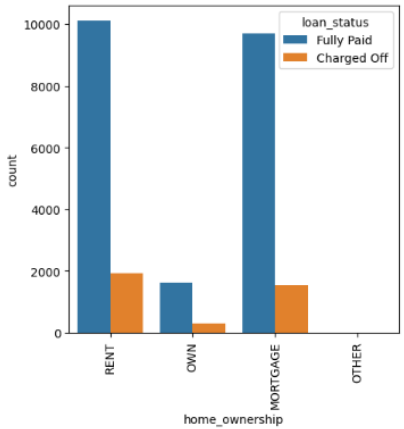
Heatmap

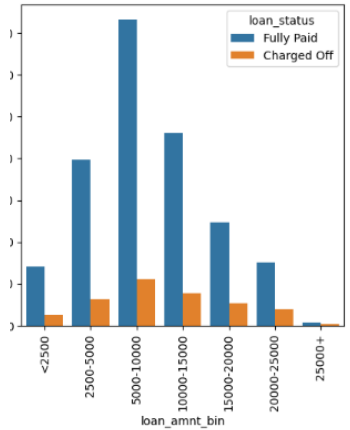


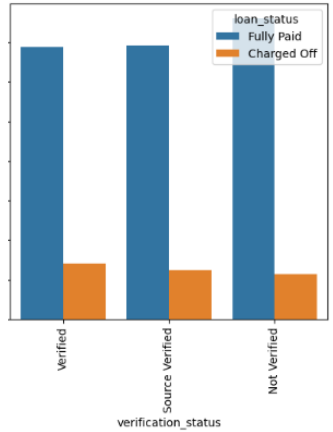
Takeaway:

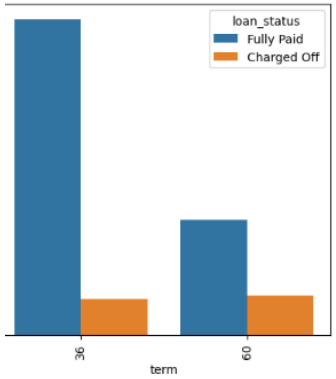
1. Strong correlation between Loan amount, Funded amount and Funded amount by investor as expected. However this insight is not significant
2. Installment also has strong correlation to these factors which is understandable
3. DTI has negative correlation with annual income as they are inversely proportional in nature

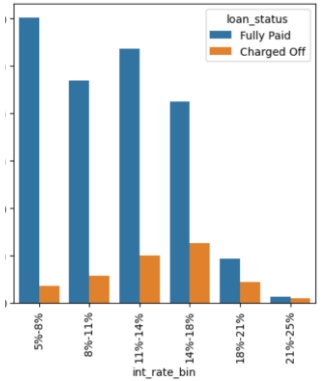
Loan Status Analysis







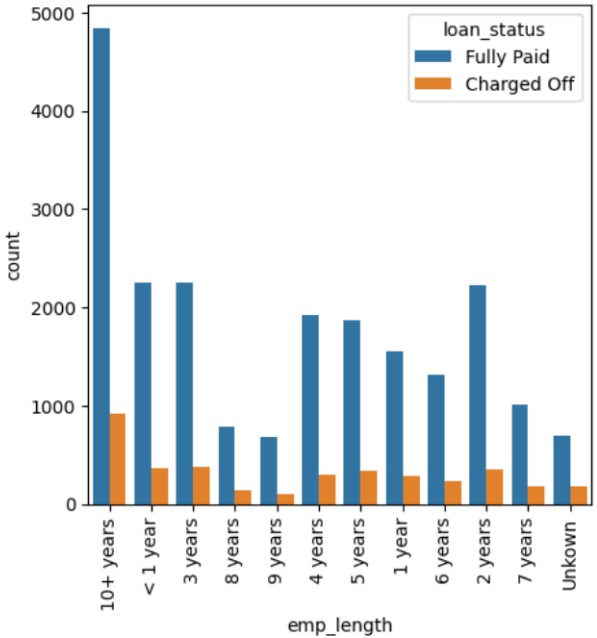




Takeaway:

- The charge off rates are higher with rental and mortgage ownership loanees
- 5K - 10K loan amounts are where the defaulting rate is quite high
- LC verification has been less effective where the charge off rate seems better for unverified profiles over verified and source verified
- Default rate is very higher for long tenure loans over short tenure (36 months) loans
- Default rate is higher for higher interest loans (>11%)

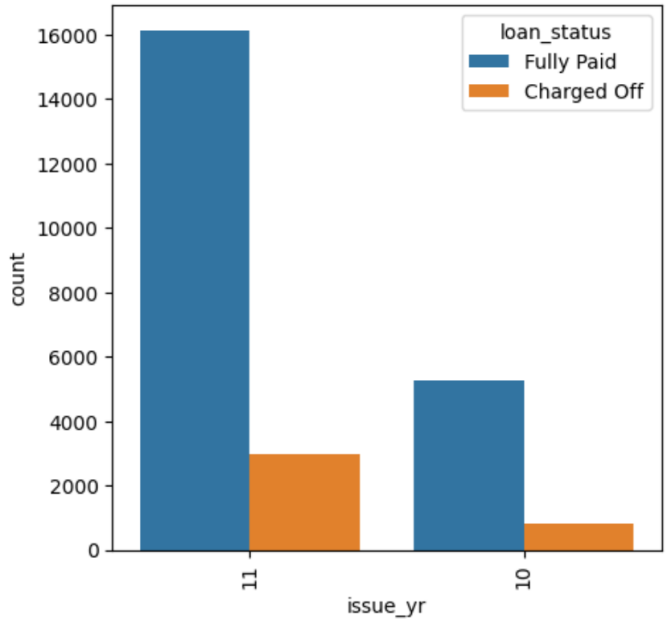
Loan Status vs employment years Analysis

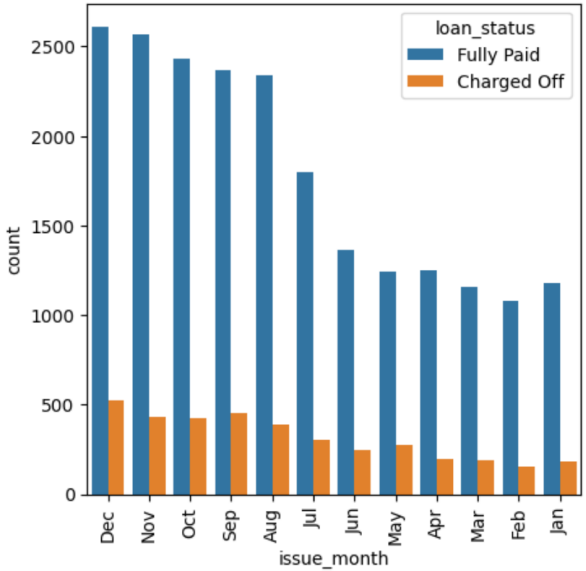


Takeaway:

- The charge off rates and loan requests are higher when employment length is more than 10 years

Loan Status vs Month and Year Analysis

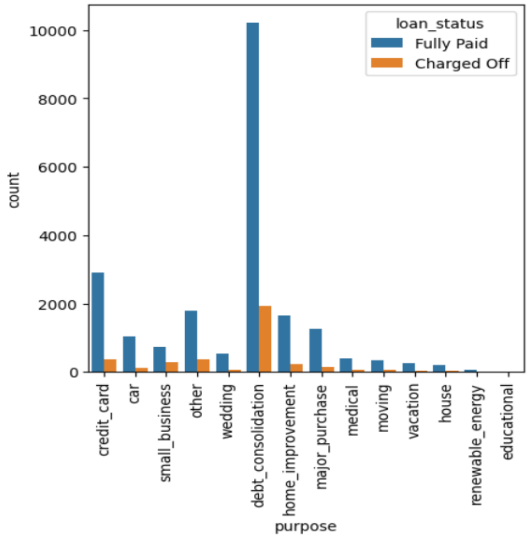


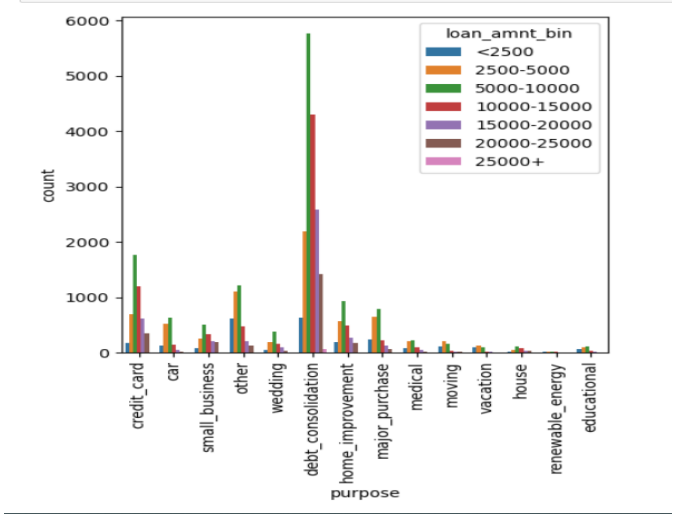


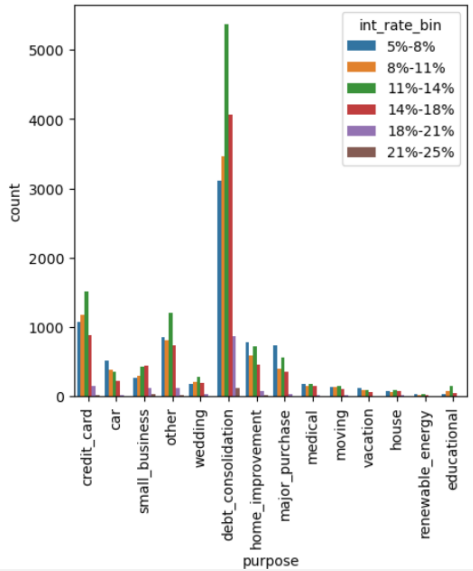
Takeaway:

- The charge off rates and loan requests are higher for year 2011
- The charge off rates and loan requests are higher for months from August

Purpose Analysis



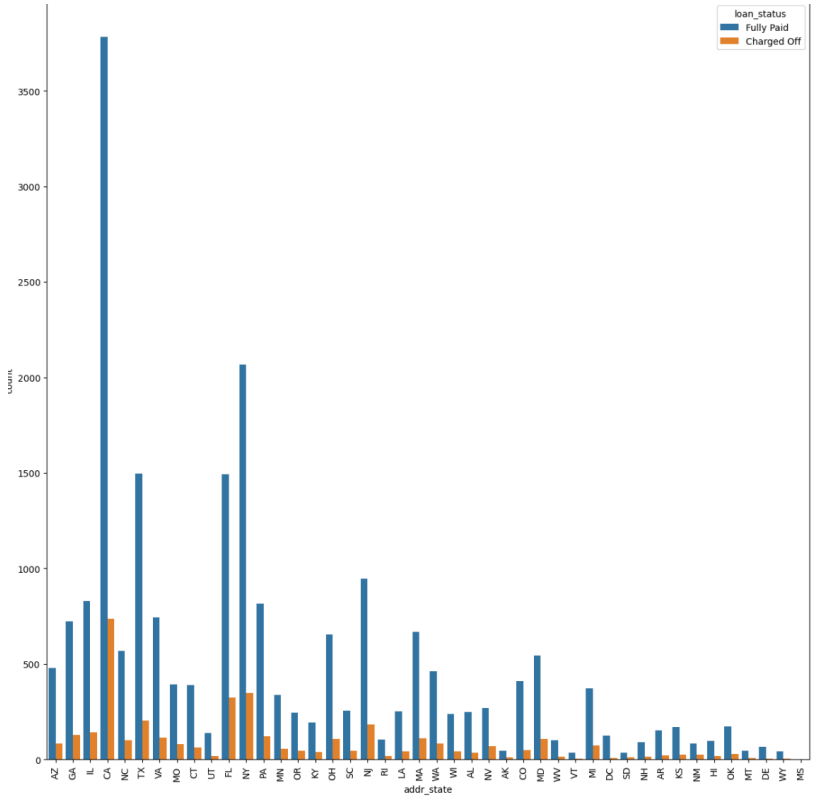




Takeaway:

- The interest rates are fairly higher for non collateralized loan purposes such as credit card, debt consolidation
- For collateral backed loans such as Car, house the interest rates are fairly lower

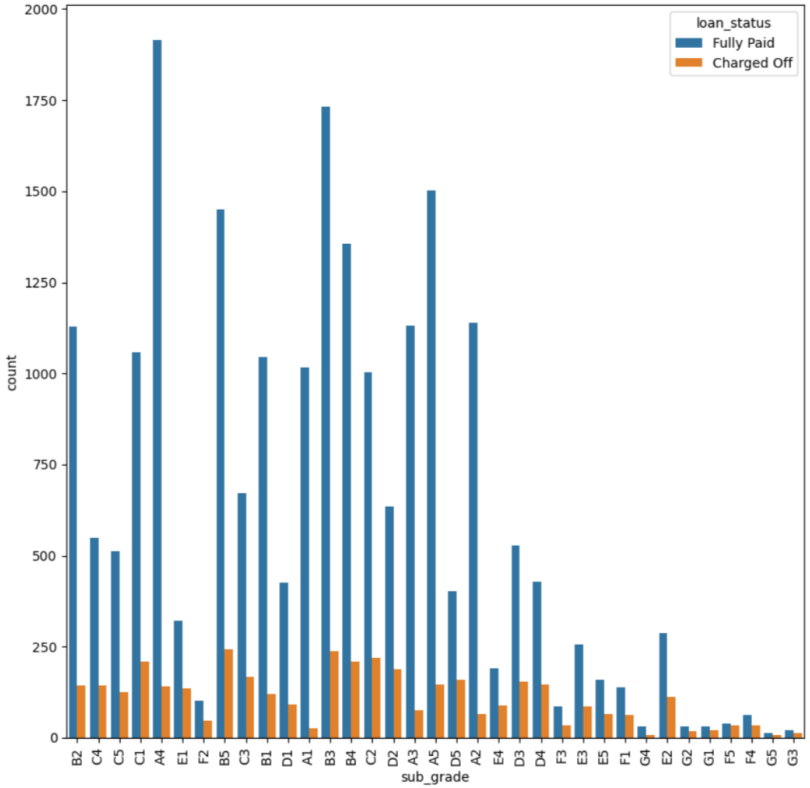
Address vs Loan status



Takeaway:

- Most of the loans are availed in California (CA) and charged offs are higher as well in CA
- However the charged off rate (ie., Loans charged off/Total loans) are higher in Nevada (NV)

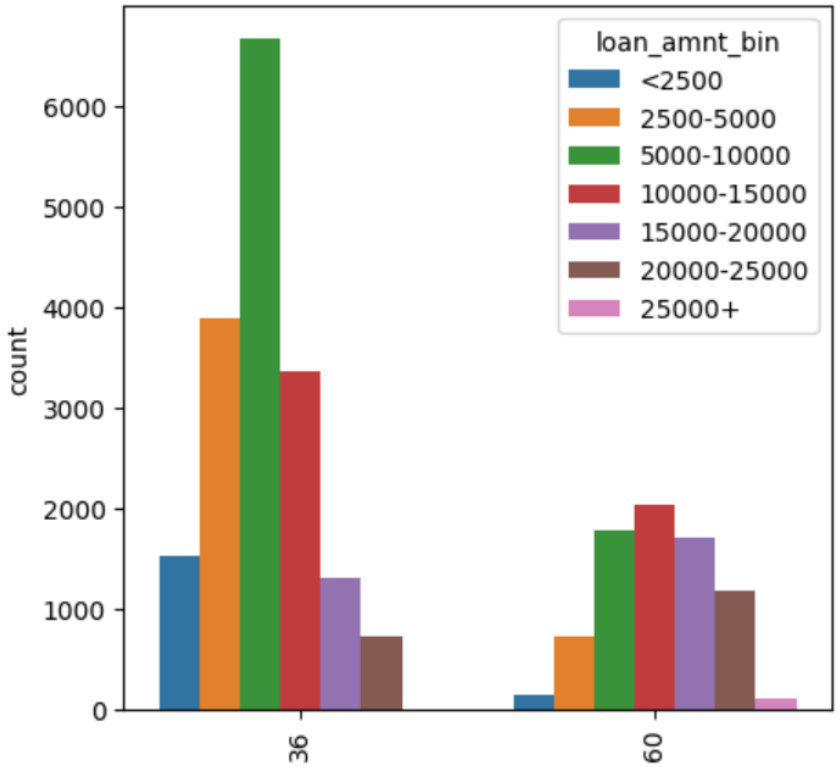
LC grade vs Loan status



Takeaway:

- Proportion of Charge off is very high in E1, E2, E3, F1, F2, F3, G1, G2, G3
- The Default rates are better off in A1, A2, A3, B1, B2, B3

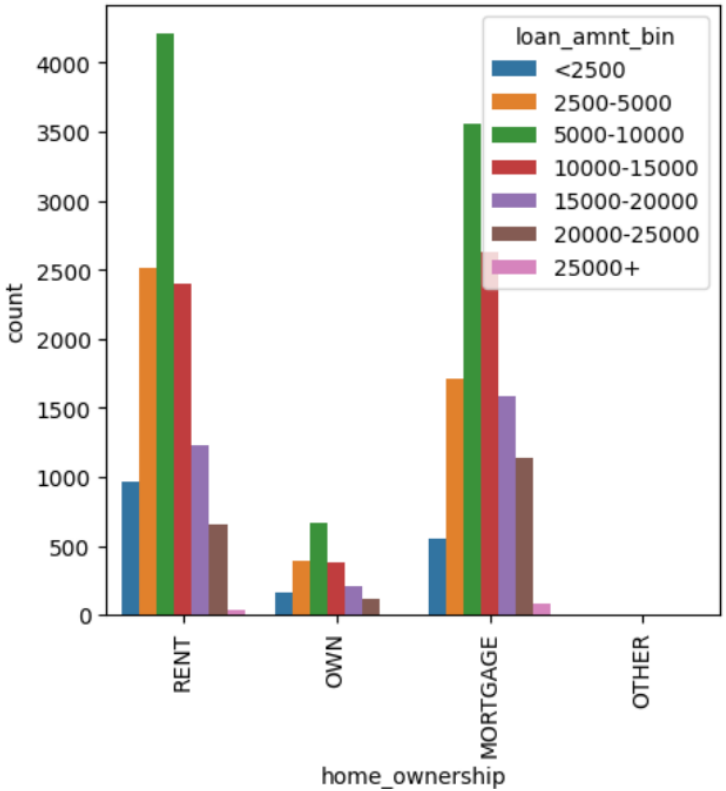
Term vs Loan amount



Takeaway:

- The loan requested amount is more for 36month term and lesser for 60 month term
- The count of amount is also more for range 5k-10k

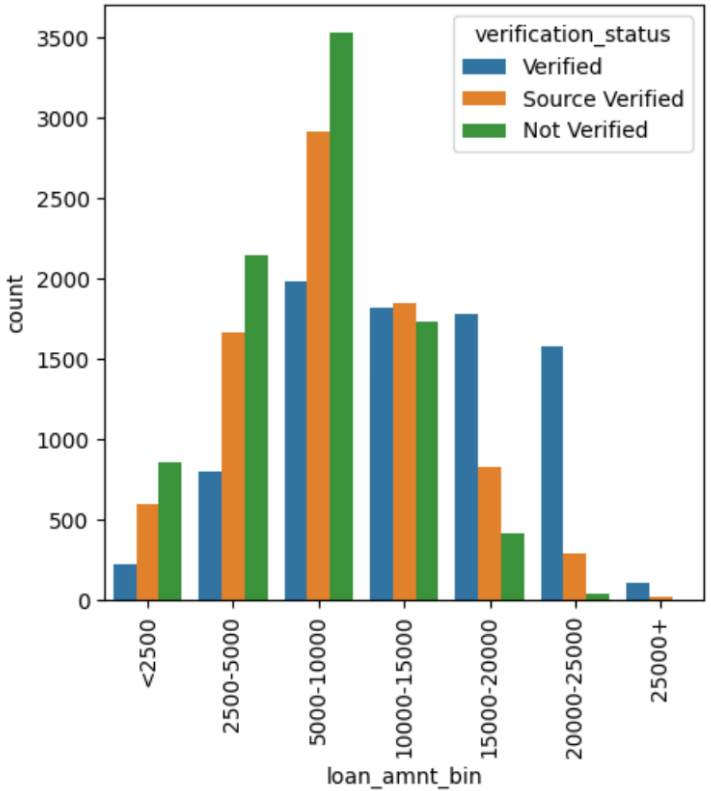
Home ownership status vs Loan amount



Takeaway:

- The loan requested amount is more where ownership status is loan or mortgage
- The count of amount is also more for range 5k-10k

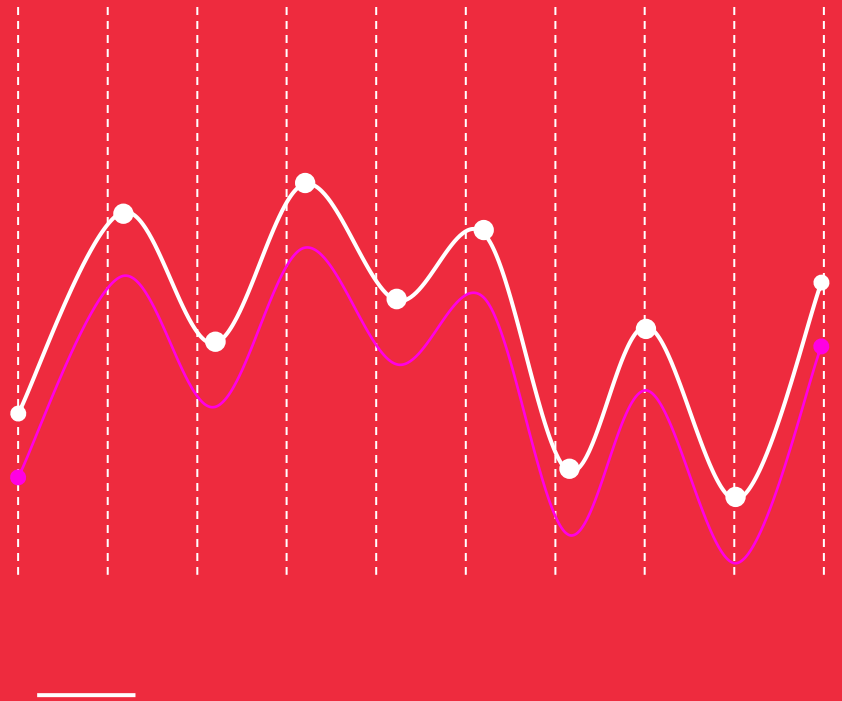
Verification status vs Loan amount



Takeaway:

- The proportion of source verification increases as the loan amount increases.
- However since the charge off rate is more in the 5k - 10k range, more rigor is needed from source verification standpoint

Recommendations to Lending Club



Recommendations to Lending Club

Factors driving **High default rate**

1. Long tenure (60 months)
2. Rental & Mortgage home ownership
3. E,F,G grade and all sub grades in this category
4. >11% interest rate
5. Non-collateralized categories such as credit card, Debt consolidation and High risk categories such as small business

Next Steps for Lending Club

1. Deploy LCs (loan collectors) effectively for user verification. Currently the quality of assessment is sub par
2. Ensure that grading has been done for all loans that fall into high risk category (left section). Grades marked below C are highly prone to get defaulted
3. Manage LC resources effectively by deploying more resources in second half of year as more loans are issued during the same time

Future scope of analysis

For further analysis,

- The final cleaned data can be used up and loan status can be marked as 0 for Charged off and 1 Fully paid
- Longer term data if available can be sourced, cleaned and can be added
- ML models such as Logistic regression, RF or SVM can be applied to build a model that takes up customer information (~22 features) to assess credit worthiness