Lending Club: EDA Case Study

Submitted By:

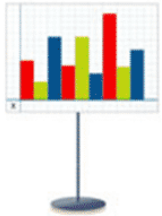
- Charan Hulmani
- Mukul Vashisth

Lending Club: Problem Understanding

How Lending Club Works

Borrowers apply for loans.
Investors open an account.

Borrowers get funded.
Investors build a portfolio.



Borrowers repay automatically.
Investors earn & reinvest.

What is Lending Club?

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.

Objective

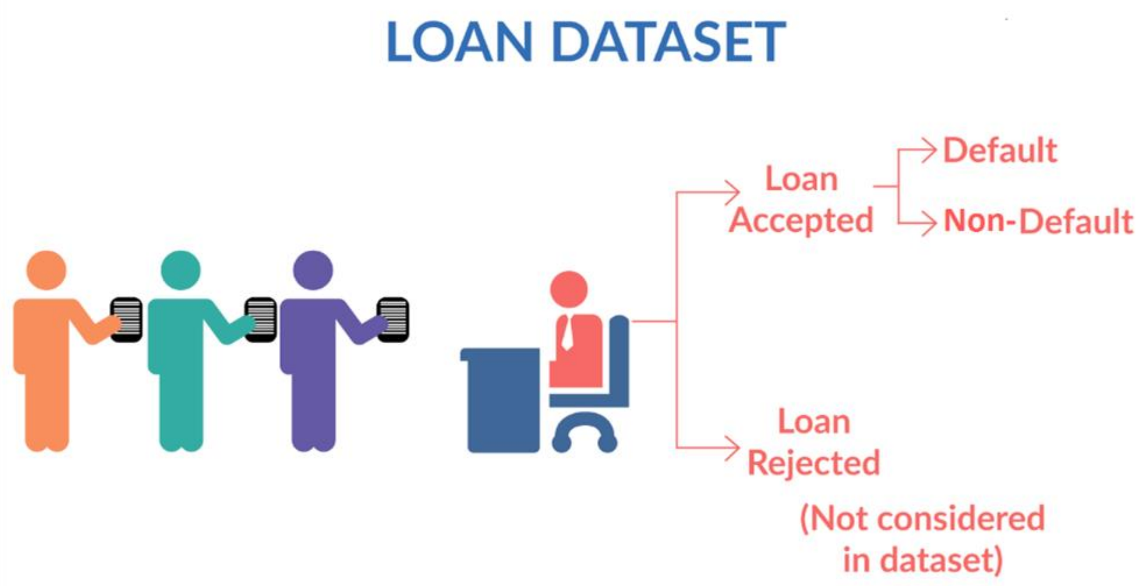
Understand the driving factors/variables behind loan default for risk assessment.

Risks Types: When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. 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

Lending Club: Data Understanding

upGrad



Fully paid: Applicant has fully paid the loan (the principal and the interest rate)

Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

There is an uncertainty for borrowers where the loan status is Current, whether they’re likely to default as time progresses or will they successfully pay-off their loans.

Lending Club: Data Cleaning

upGrad

- Reading and identifying different features of the dataset such as null values and unique counts
- Columns having all null values eliminated from the dataset, along with columns having high % of nulls
- Columns with singular values (single unique value throughout) are of no significance and thus dropped from dataset
- Some columns like description, URL, Zip and Title dropped off before analysing further since they seem to contain no significant information for analysis
- Borrowers who make regular payments towards their loan through EMI’s have their loan status classified as Current. And these shall not be considered for analysis because of the uncertainty revolving around behaviours to default or successful pay-offs.
- Standardized few variables and converted to accurate data type format. Term, Employment length and Interest rate being the columns underlying this situation.
- Outliers identified from the columns Annual Income and Loan Amount and removed using the standard IQR technique.

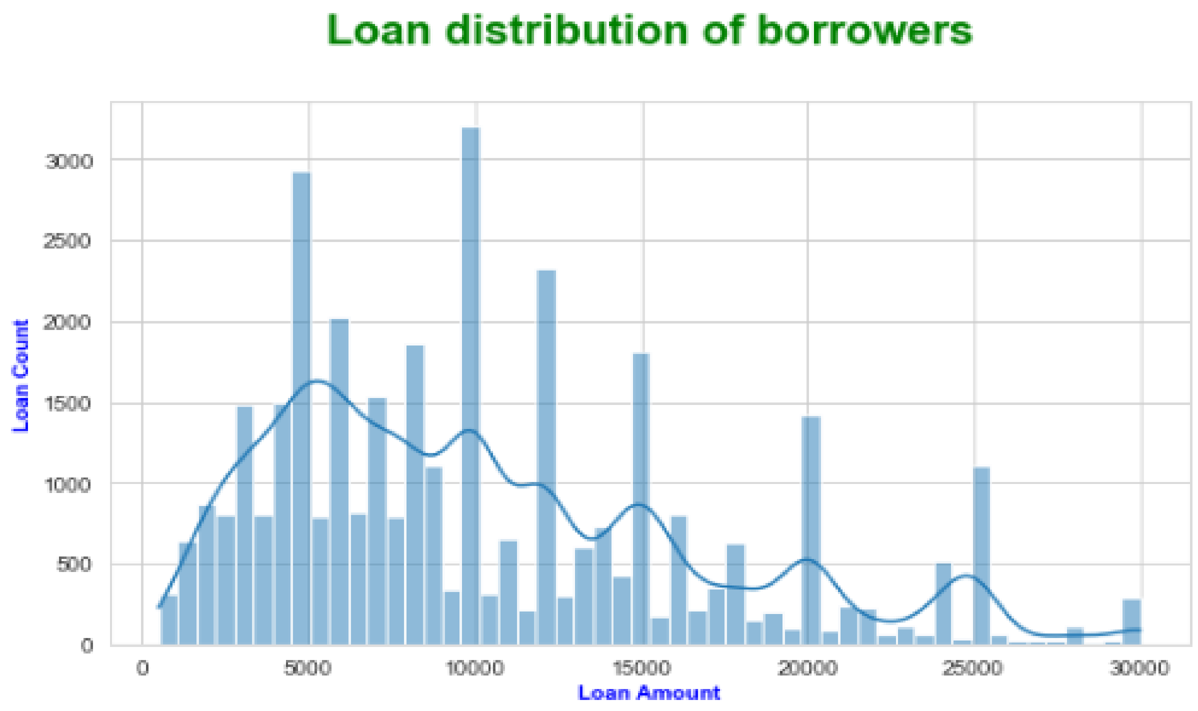
Data Analysis : Univariate

upGrad

Loan Amount

Amount of loan specified by borrower at the time of application

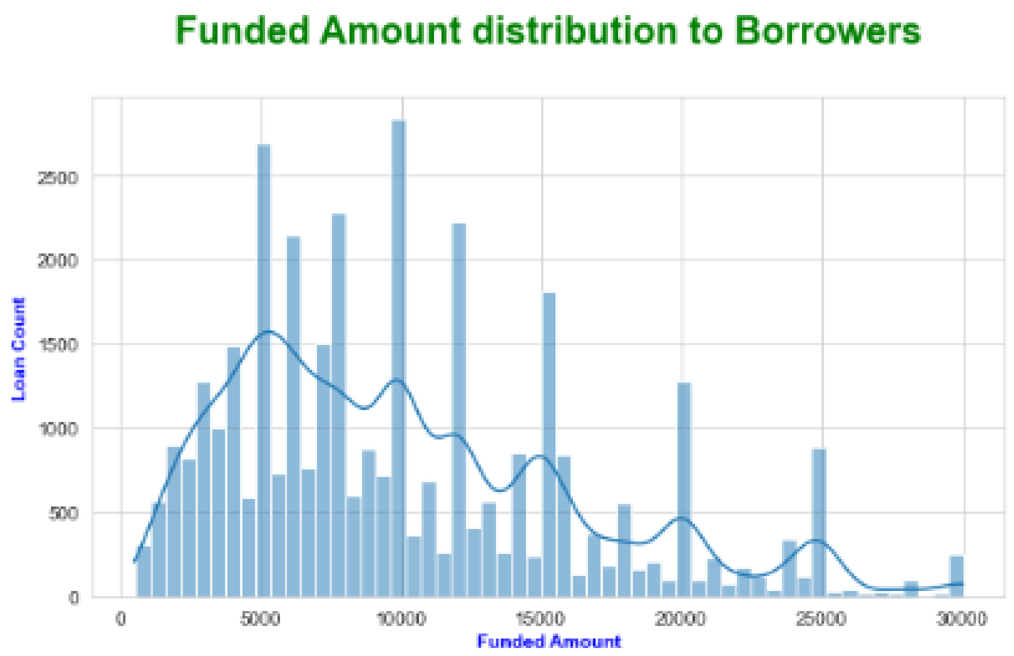
- The number of loan applications is maximum between the range 5k – 15k
- Loan grant requests are at peak for 10k
- Lending club has received most applications within the range 5k-10k

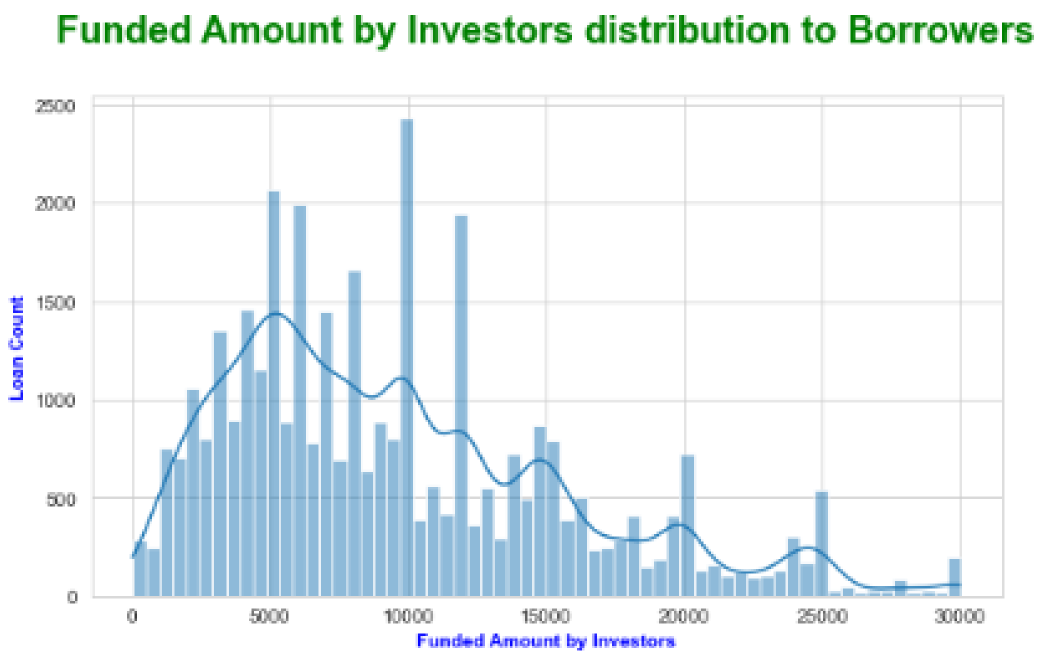


Data Analysis : Univariate

upGrad

Funded Amount and Investor Funded Amount





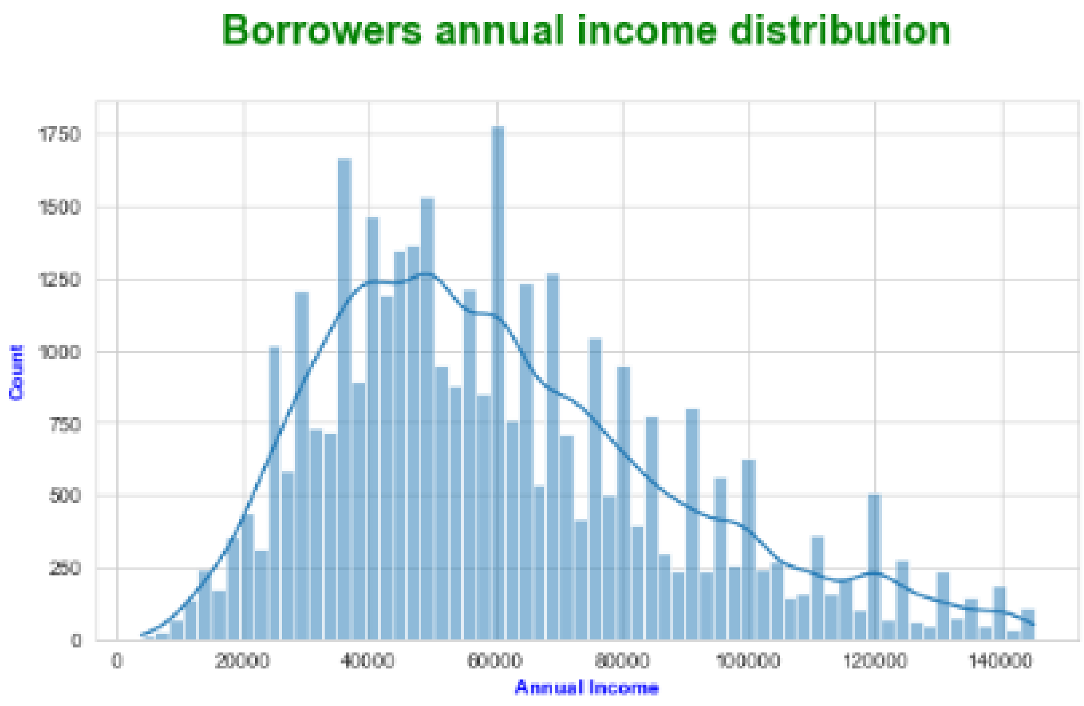
- The difference in Investor funded amount from funded seems to be less for applications below 10k mark
- The investor funded amount shows higher difference with funded amount when the amounts requested are high

Data Analysis : Univariate

upGrad

Annual Income

- The number of loan applications are low where the annual income is high
- Loan requests are high for people earning in the range 20k – 80k
- Maximum number of loan applications are coming from the cluster with annual incomes in range 40k – 50k

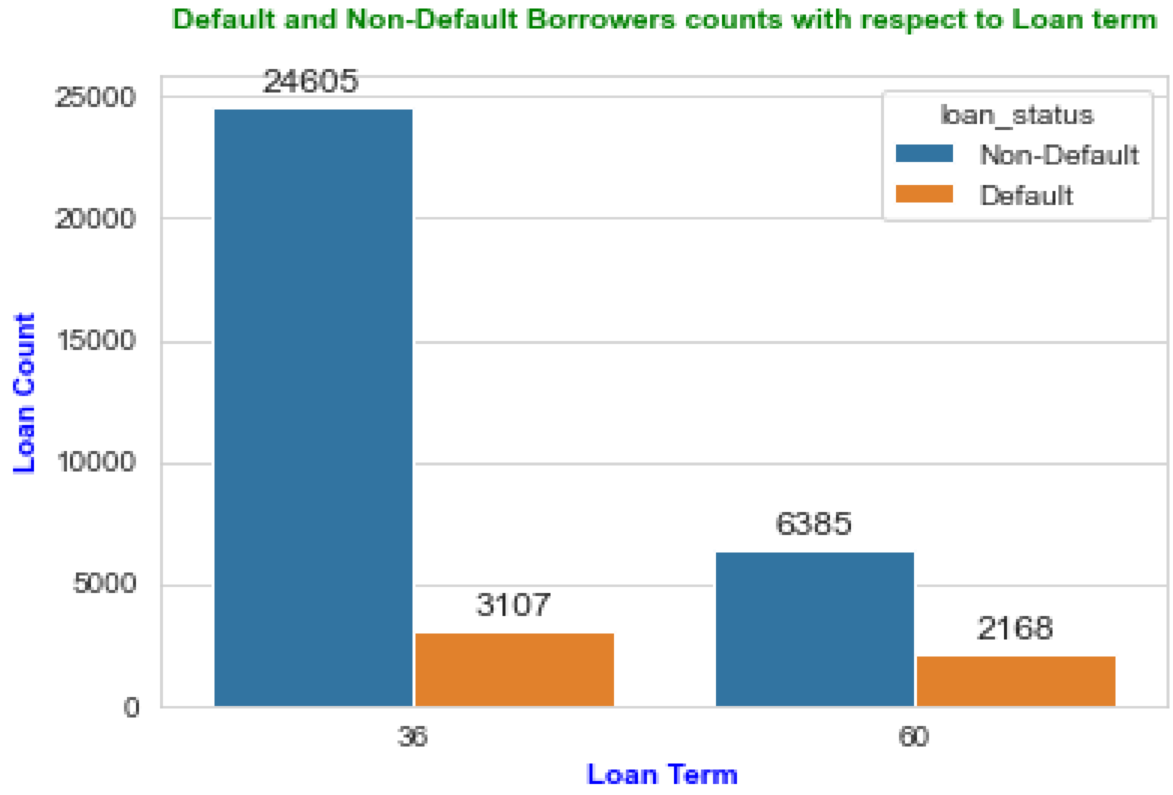


Data Analysis : Segmented Univariate

Term

The number of payments scheduled until loan is paid off successfully. Measured in months

- It can derived from the plot that most number of applicants opt for small tenure of 3 yrs. as compared with 5 yrs.
- Risk in lending loan for borrowers who requests 60months EMI plan is 2.4 times more compared to borrowers who requests 36 months EMI plan.
- The tendency of an applicant to default with subsequent payments is higher for those opting for longer terms.



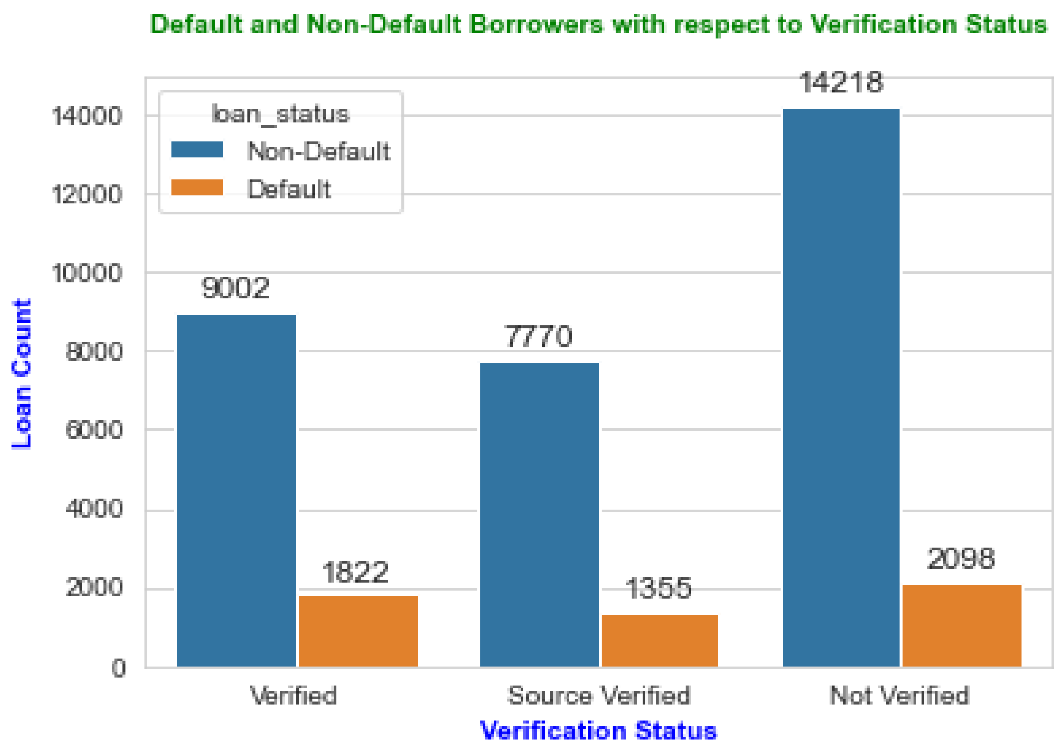
term	%Tot.Default	%Tot.Non-Default	%Total	Ratio	% of Default
36	8.57	67.85	76.42	0.13	11
60	5.98	17.61	23.58	0.34	25

Verification Status

This variable is an indicator whether the source of income for the applicant was verified or not.

i. Income Verified (Pay Slips, Tax Returns)
ii. Income Source Verified (3rd Party checks)
iii. Not Verified

- From the plot, it appears that the risk of lending a loan to borrower increases in order from Not Verified, Source Verified and Verified. More risk involved to lend loan to Verified borrowers than Non Verified borrowers.
- However, we may also realize that income source for a big portion of applicants had not been verified which should ideally be reduced.

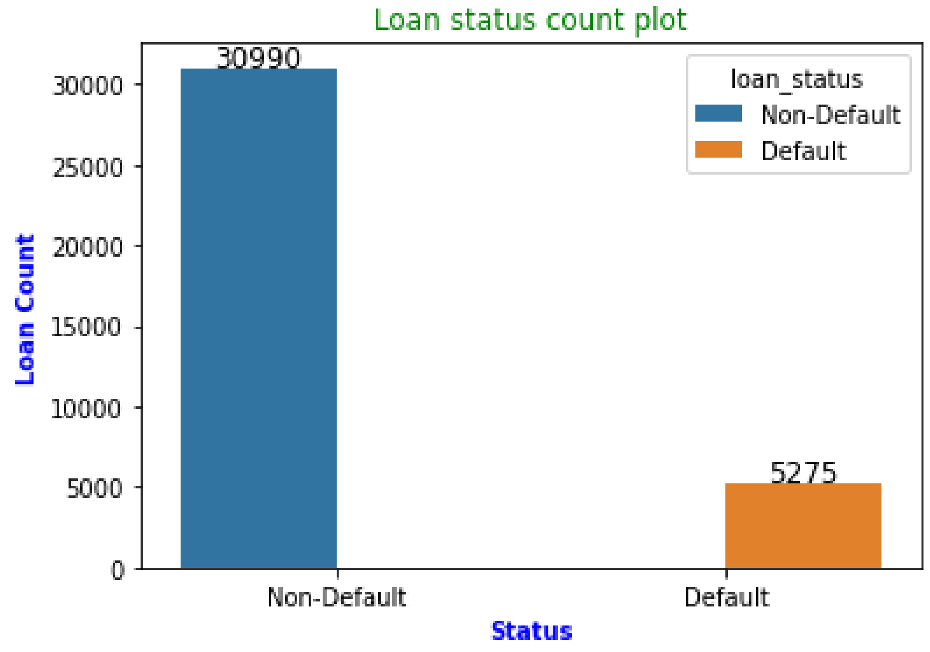


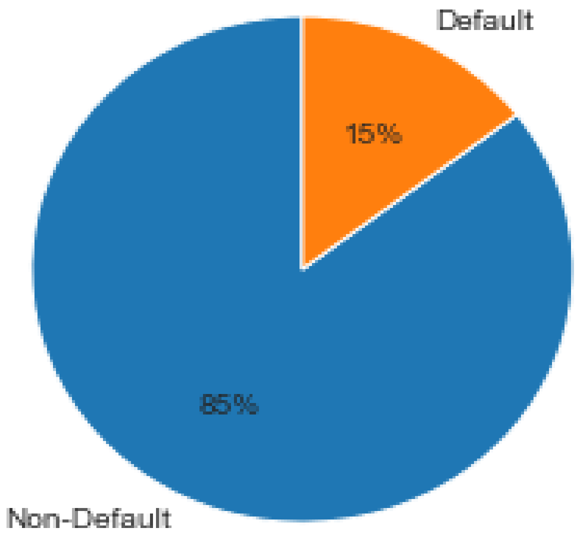
verification_status	%Tot.Default	%Tot.Non-Default	%Total	Ratio
Not Verified	5.79	39.21	44.99	0.15
Source Verified	3.74	21.43	25.16	0.17
Verified	5.02	24.82	29.85	0.20

Data Analysis : Segmented Univariate

upGrad

Loan Status

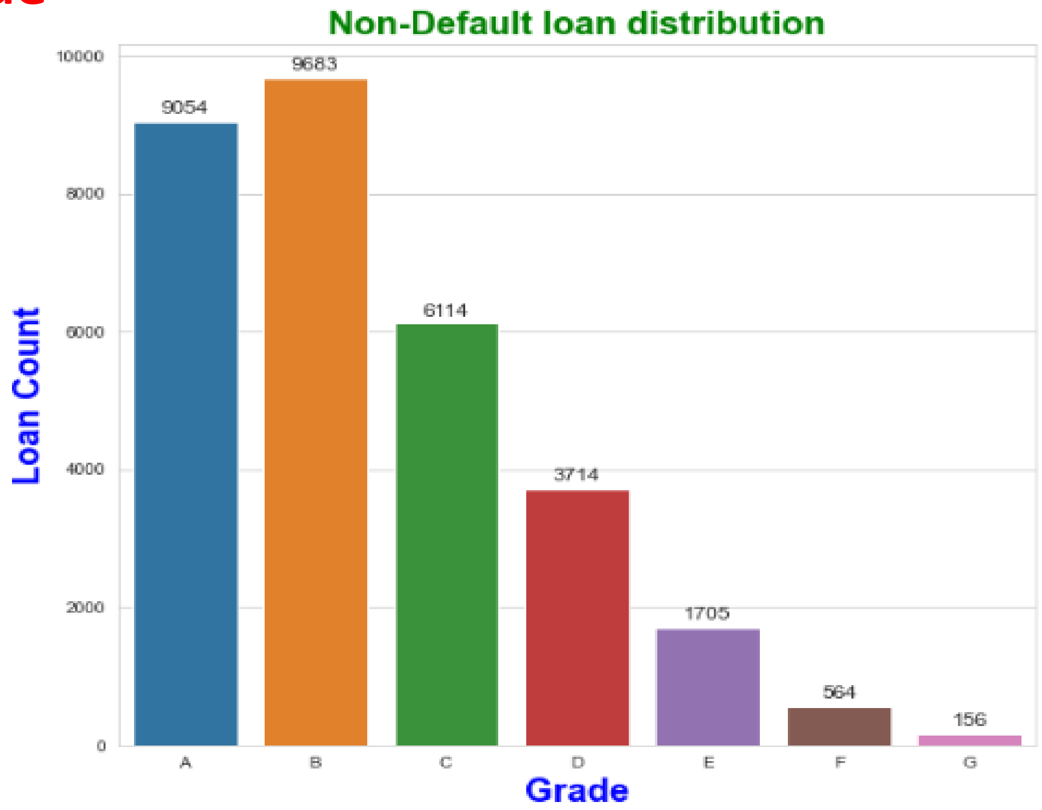


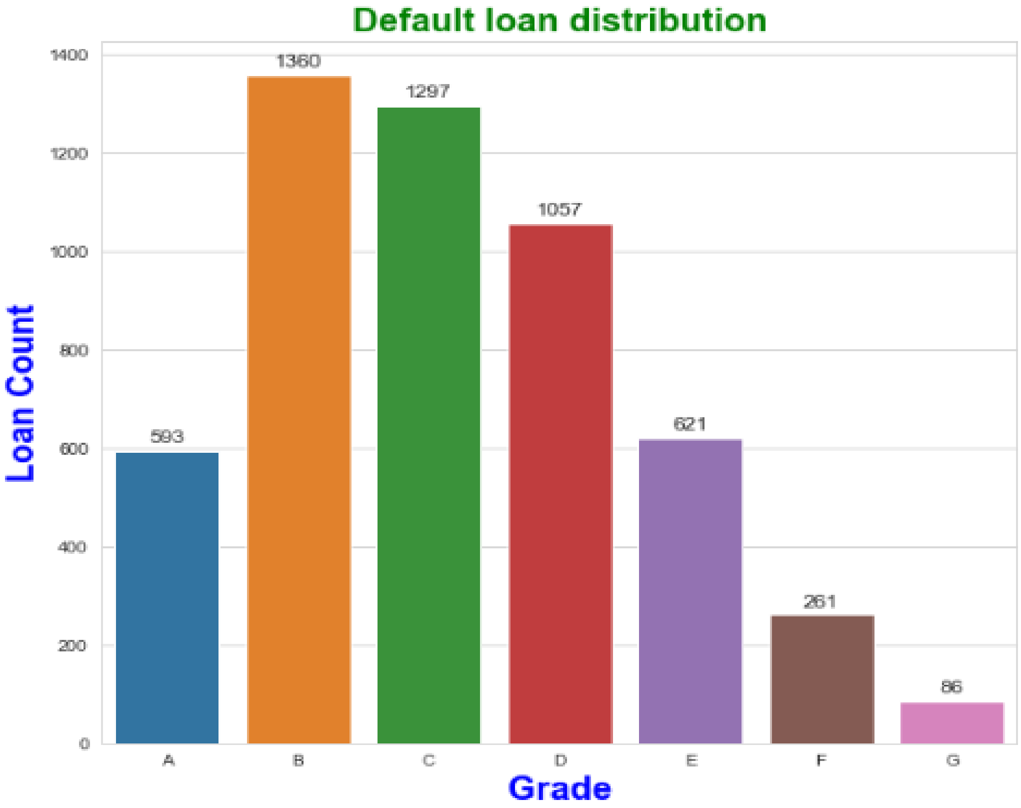


- Loan status represents the current state of loan. It can be either defaulted or successfully paid-off since we have excluded those in current state.
- From the plot:
  - maximum number of applicants are likely to repay their loans successfully than those to default.
  - There’s only 15% chance for a person to default on regular payments

Data Analysis : Segmented Univariate

Grade





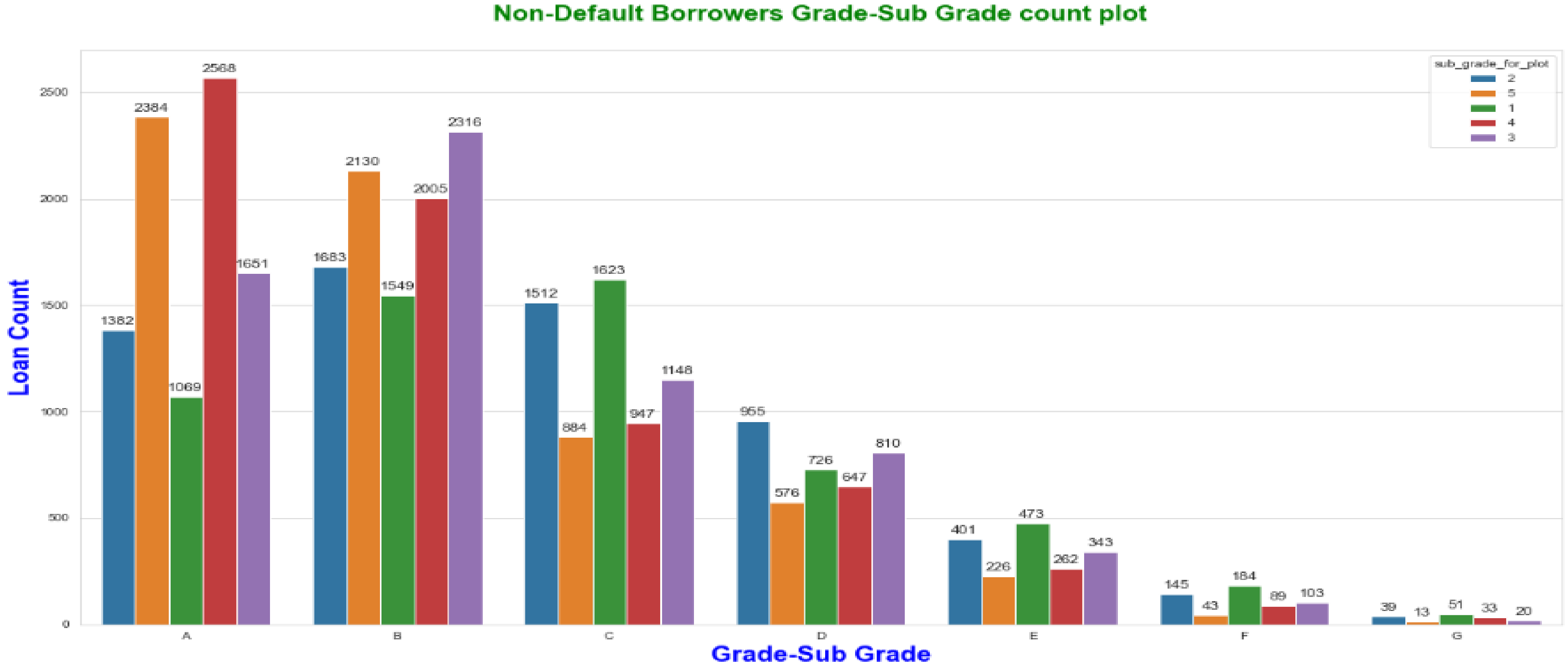
- The maximum number of loans are assigned to Grade B, followed by A, C, D, E,F,G in order.
- From specific grades aligned, the highest probability of default is for Grade G, followed by F, E, D, C, B, A.
- Investments made on applicants from Grade A, B & C are highly profitable with lowest chances of defaulters and maximum number of loans underlying are in these categories. These are followed by grades D, E, F, G.

Data Analysis : Segmented Univariate

upGrad

Sub Grade - Successful Pay-Offs

Sub category against Grade variable decided by the lending club

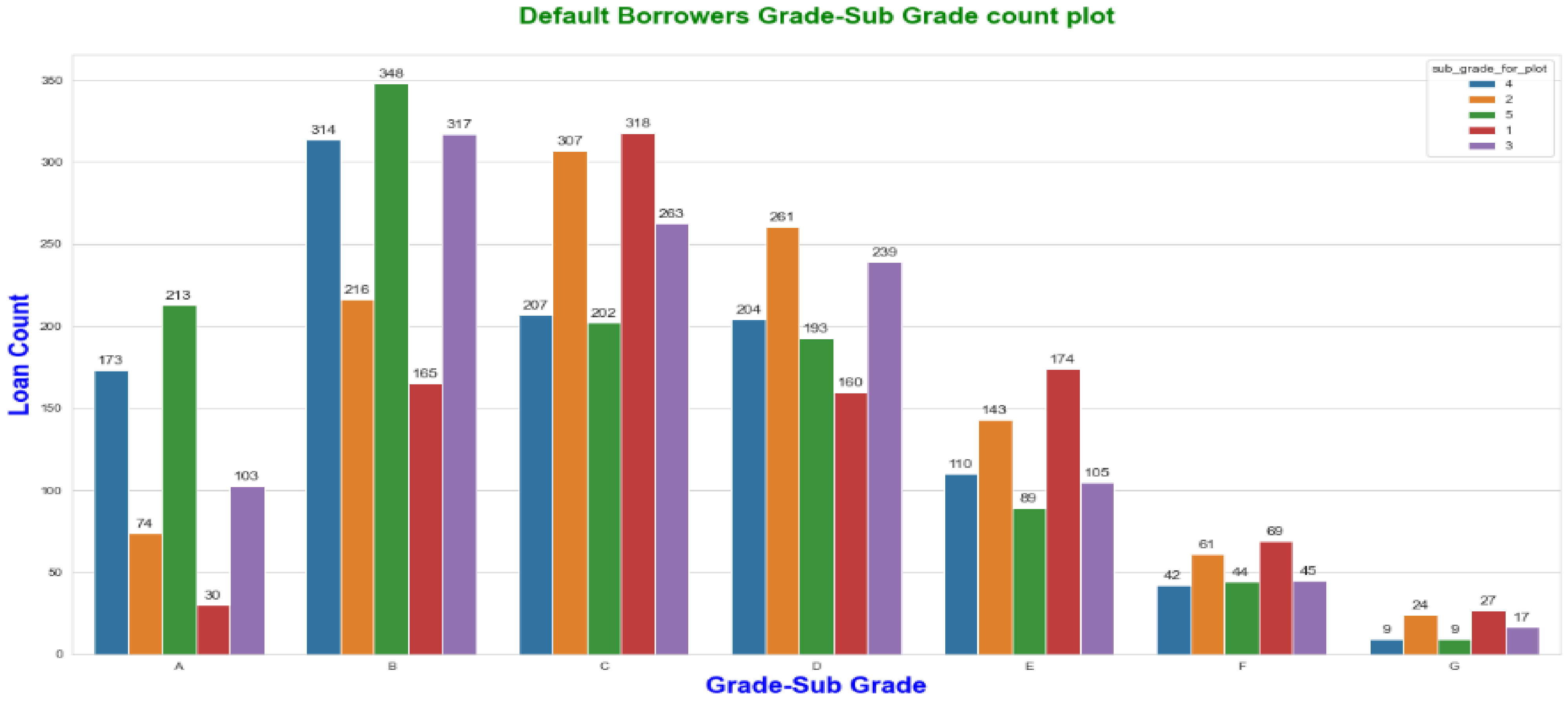


- Sub-Grade 1 & 2 appear to be most reliable across grades

Data Analysis : Segmented Univariate

upGrad

Sub Grade – Defaults



- Grades A, B & D borrowers, risk associated increases in order from sub-grades in 1,2,3,4,5
- Grade C borrowers, risk associated with grades increases in increasing order of sub-grades from 1,2,4,5,3
- Grade E & F borrowers, risk associated with grades increases in increasing order of sub-grades from 3,2,1,5,4

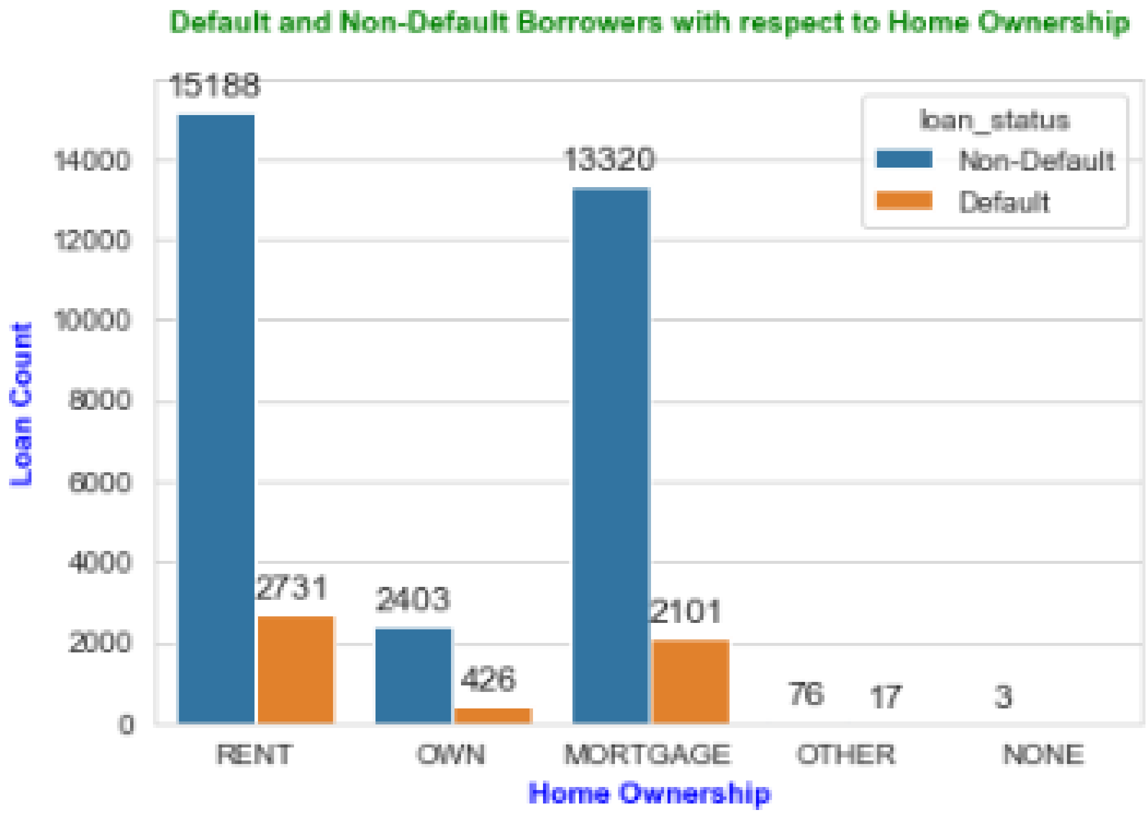
Data Analysis : Segmented Univariate

upGrad

Home Ownership

Represents the ownership status provided by the borrower during registration for their residence.

- Maximum number of borrowers are renting place for their residence or have taken mortgage for their house.
- From the plot, the chances of risk associates with different home ownership status is as:
  - 5 out of 100 borrowers who are in a rented or own house will become defaults
  - 13 out of 100 borrowers who are in a mortgage house will become defaults
  - 18 out of 100 borrowers who are in a other accommodation who gets loan will become defaults
- Risk involved in lending loan to rented and own house is same. Furthermore, risk is comparatively less in lending loan to borrowers in mortgage house.

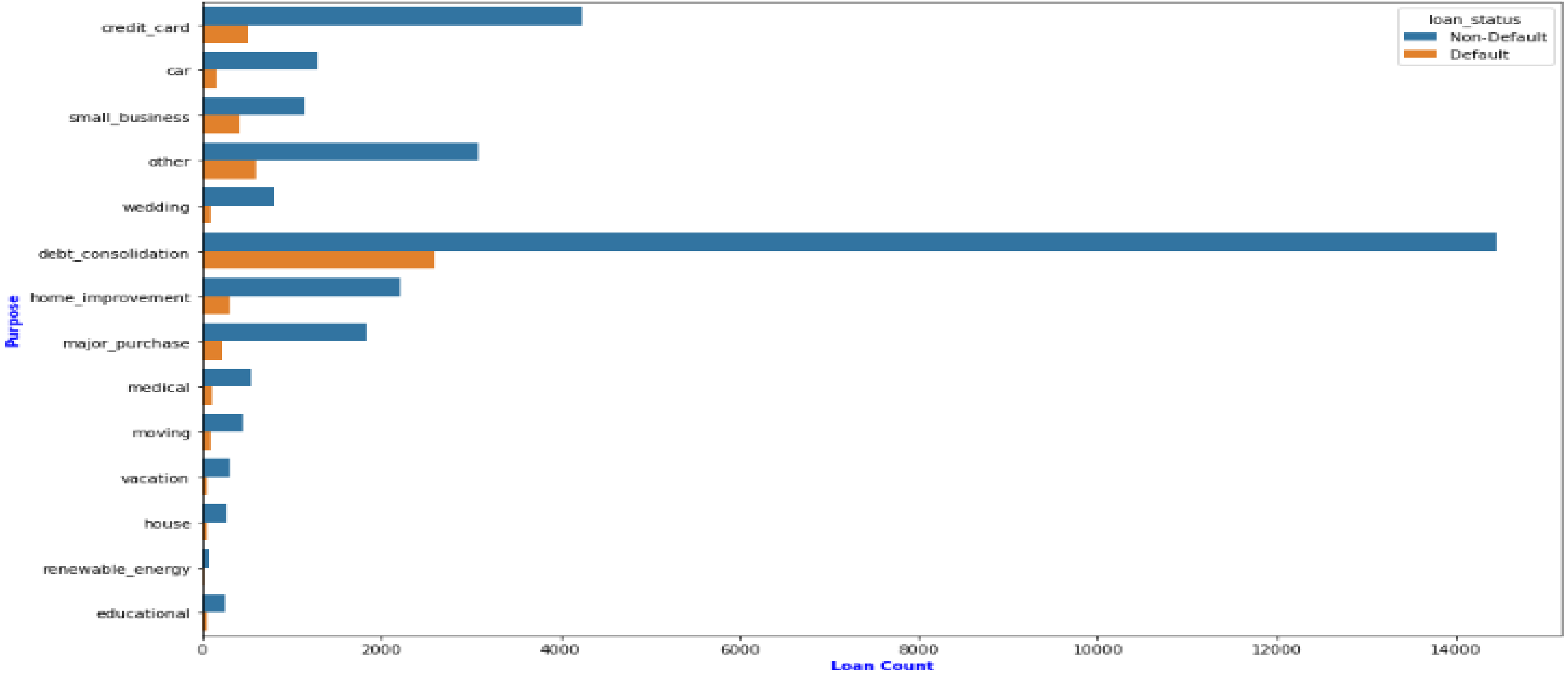


home_ownership	%Tot.Default	%Tot.Non-Default	%Total	Ratio	% of Default
MORTGAGE	5.79	36.73	42.52	0.16	13
OWN	1.17	6.63	7.80	0.18	15
RENT	7.53	41.88	49.41	0.18	15
OTHER	0.05	0.21	0.26	0.22	18

Data Analysis : Segmented Univariate

upGrad

Purpose



- Debt consolidation and credit card make up for maximum number of defaults as well as successful pay-offs
- Risk involved in lending loan to for the purpose of small business is more and is less for wedding.

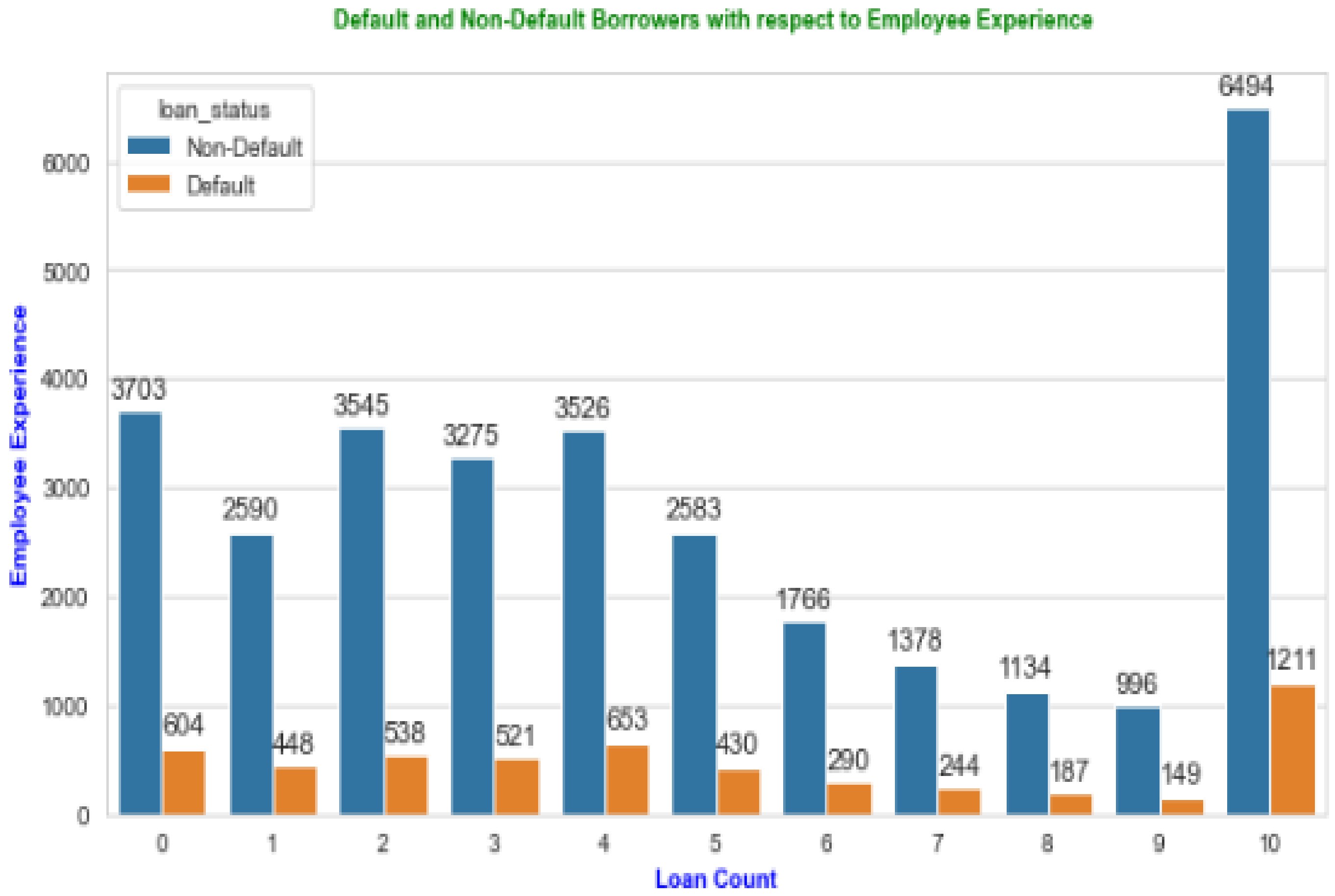
Data Analysis : Segmented Univariate

upGrad

Employment Length

- 9 year Experience employee are paying off the loan pretty good compared to others
- Risk of lending loan increases in the order w.r.t. employee experience as below,
  9, 2, 3, 0, 6, 8, 5, 1, 7, 4, 10
- Those who are having exp. of 10+ yrs. tend to take highest number of loans. And their chances of default are the highest.

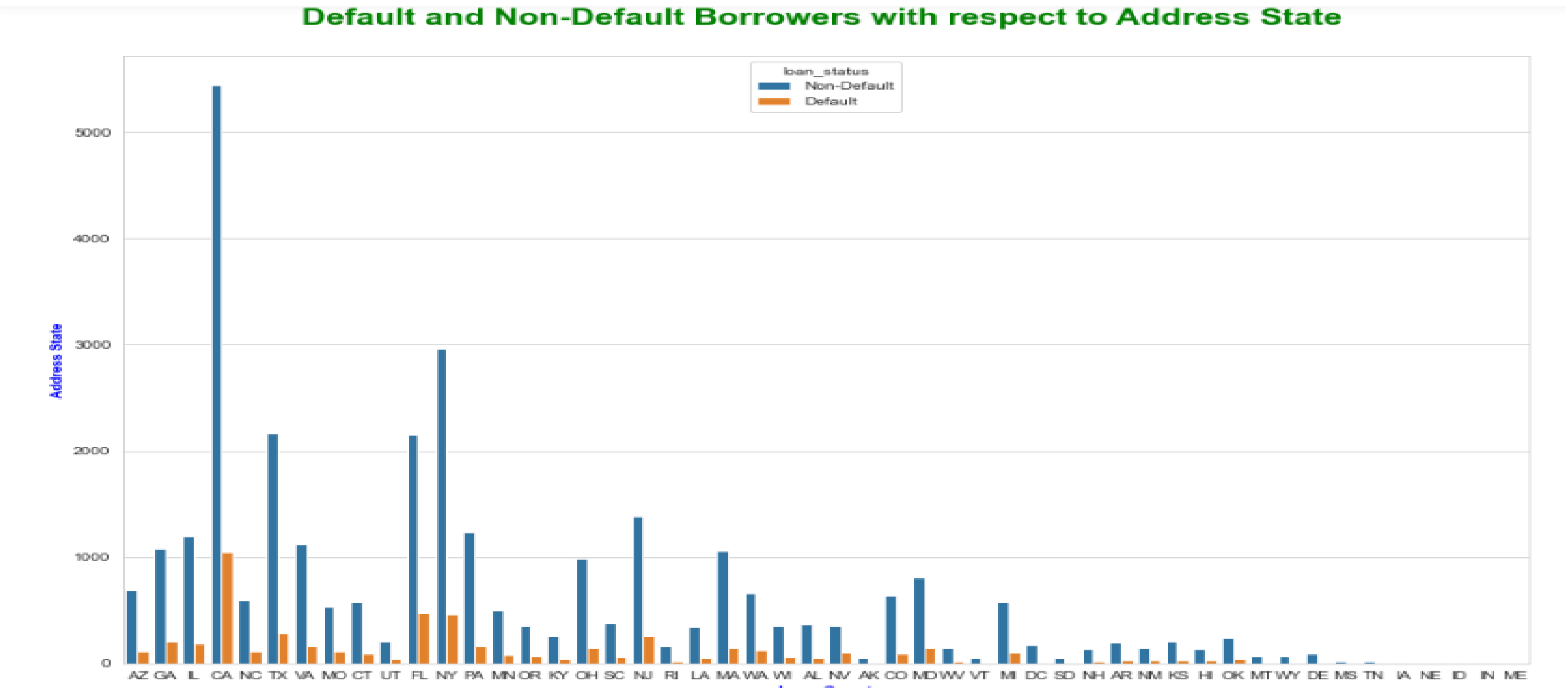
emp_length	%Tot.Default	%Tot.Non-Default	%Total	Ratio	% of Default
9	0.41	2.75	3.16	0.15	13
2	1.48	9.78	11.26	0.15	13
3	1.44	9.03	10.47	0.16	13
0	1.67	10.21	11.88	0.16	14
6	0.80	4.87	5.67	0.16	14
8	0.52	3.13	3.64	0.16	14
5	1.19	7.12	8.31	0.17	14
1	1.24	7.14	8.38	0.17	14
7	0.67	3.80	4.47	0.18	15
4	1.80	9.72	11.52	0.19	15
10	3.34	17.91	21.25	0.19	15



Data Analysis : Segmented Univariate

upGrad

Address State

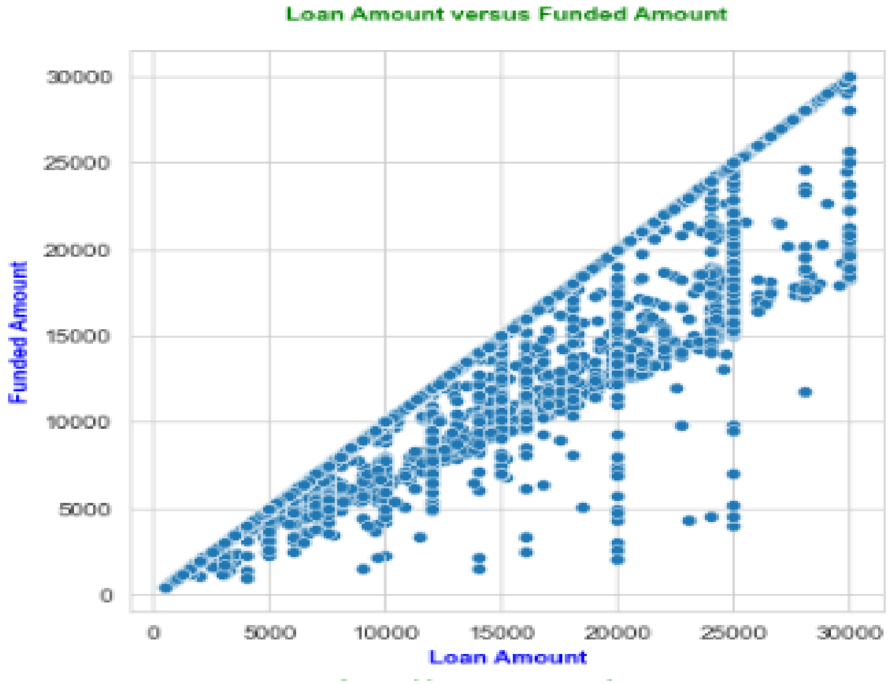


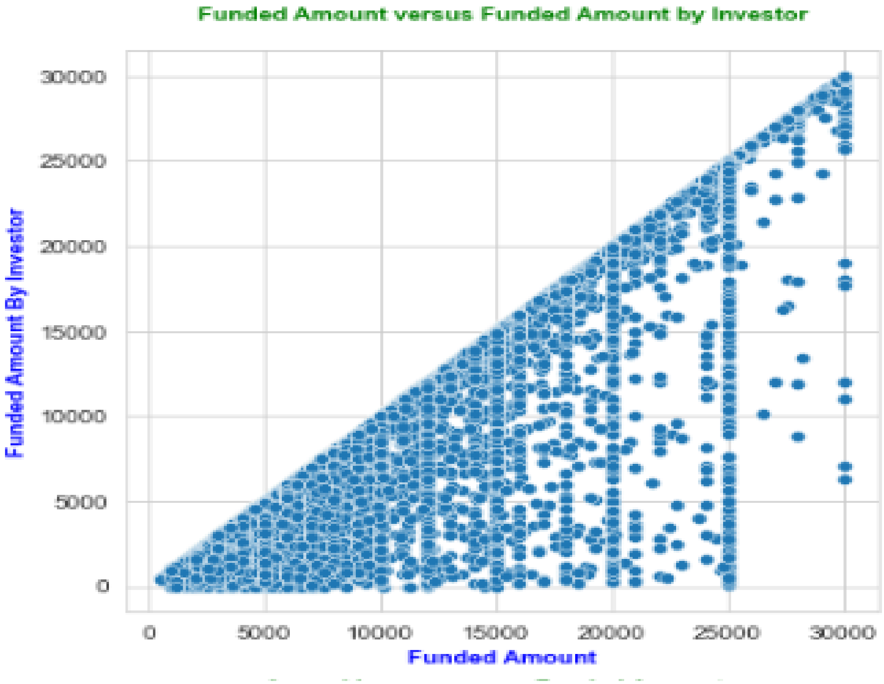
- Wyoming state borrowers are pretty good in paying off the loan. Whereas, the 60% loan borrowers from Nebraska state are not paying off the loan.
- California, New-York, Texas and Florida are the states where people are taking loans from Lending Club
- Risk of lending a loan between these 4 states is in the order from TX, NY, CA, FL

Data Analysis : Bivariate

upGrad

Relationship between Loan Amount, Funded Amount, Investor Funded Amount



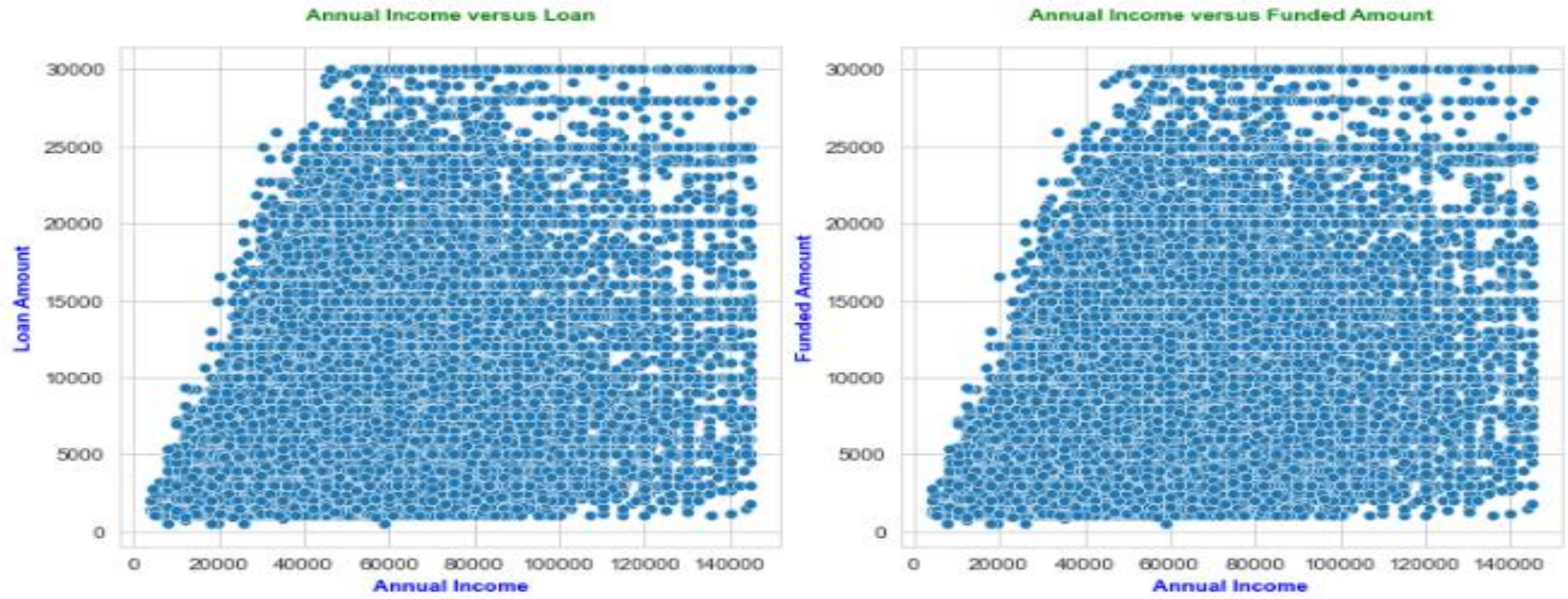


- From the plot between Loan Amount against Funded Amount, the requested loan amount will most likely get approved.
- From the plot of Funded Amount versus Funded Amount by Investor, investors are most likely to lend the funded amount with minor deduction in some cases.

Data Analysis : Bivariate

upGrad

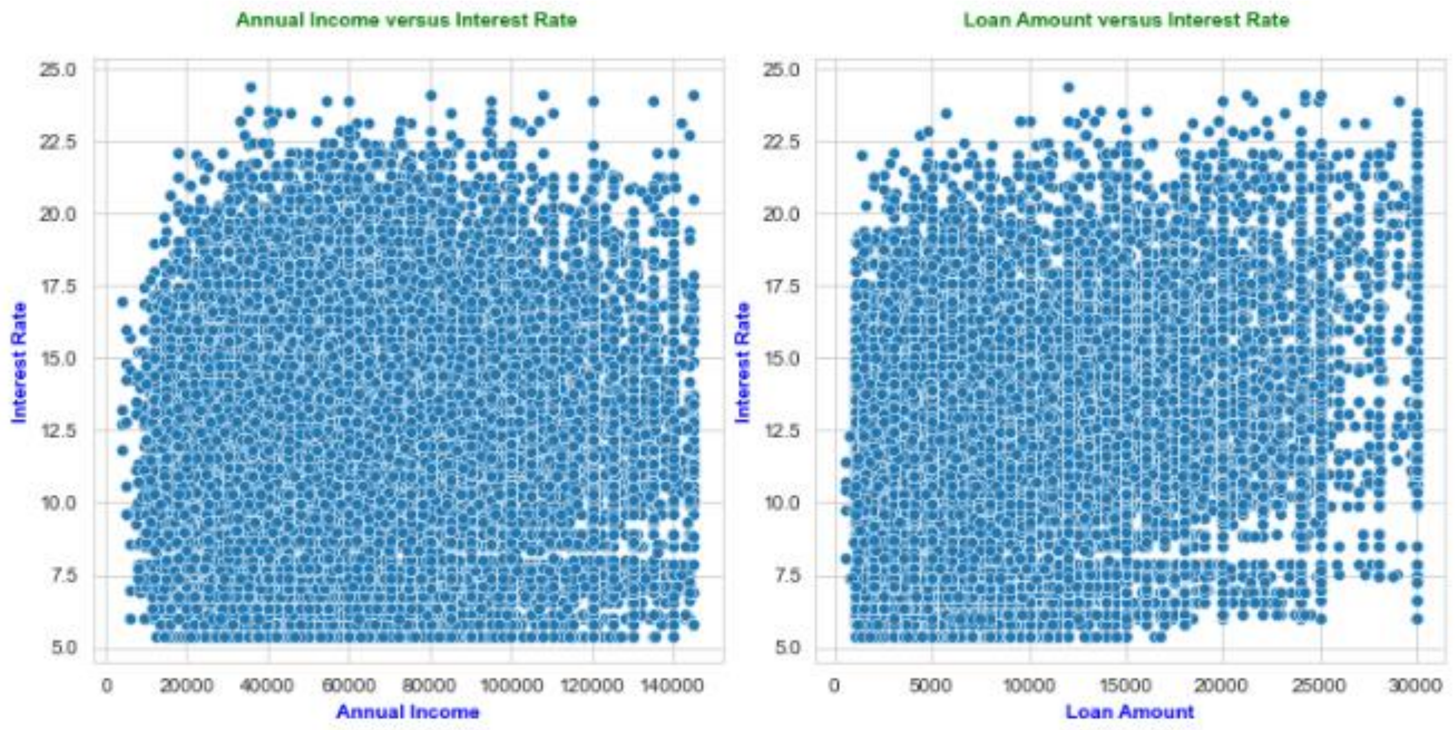
Relationship between Loan Amount, Funded Amount, Annual Income



- Annual income appears to have less influence on loan and funded amounts.
- It look like the offered loan amount for level income employee maxes at $30,000, i.e. limit enforced by the lending club

Data Analysis : Bivariate

Relationship between Loan Amount, Annual Income & Interest Rate

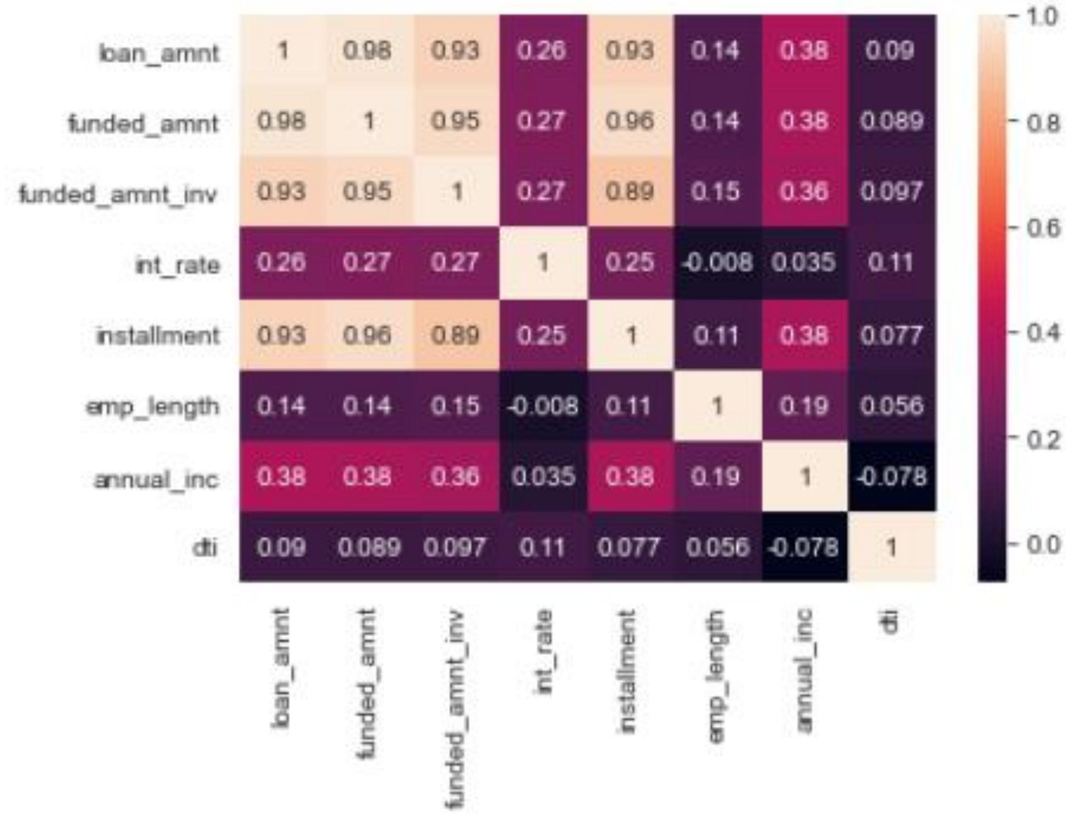


- The random nature of interest rate plots against Annual income and Loan amount implies that it is independent from these factors

Bivariate Correlation Matrix

upGrad

Correlation Matrix



- There is a high positive correlation between loan amount, funded amount, funded amount by investor and installment. This makes sense in general as the loan requested is directly proportional to approved and the actual amount lent to borrowers.
- There is a slight positive correlation between loan amount to annual income, funded amount, funded by investor
- Annual income to dti correlation is negative, which makes sense as the debt increases compared to income when a borrowed loan
- Employment length and interest rate show low correlation score implying they have less influence on other factors.

Recommendations

- Applicants asking longer tenures for repayment have high tendency to default. So, terms can be reduced for some applicants.
- Income verified borrowers have shown high tendency to default. So, new checks might need to be implemented for income verification process.
- Loan was issues to a high number of borrowers without verification. There seems a need to enforce this verification process.
- Grade G borrowers have indicated more probability to turn defaulters. So, they should be investigated more before approving the loans to them
- Sub-Grades 4 & 5 have maximum risk associated for risk. Caution must be taken against these borrowers.
- Borrowers who rent houses or have mortgage have indicated greater default rate.
- Employed years is not a trustworthy factor while lending loans as people with maximum experience have defaulted on maximum occasions.
- Loan applications are maximum where borrowers seek funds between 5k-10k. So, above factors need to tested
- Loans provided to small businesses seem less promising in repaying. Some caution should be taken with such applications

upGrad

Thanks!

Submitted By:

- Charan Hulmani
- Mukul Vashisth