Lending Club Case Study

Tarini Iyengar

emailtarini@gmail.com

Sony Jaiswal

mail4sony@gmail.com

5th July 2023

Agenda

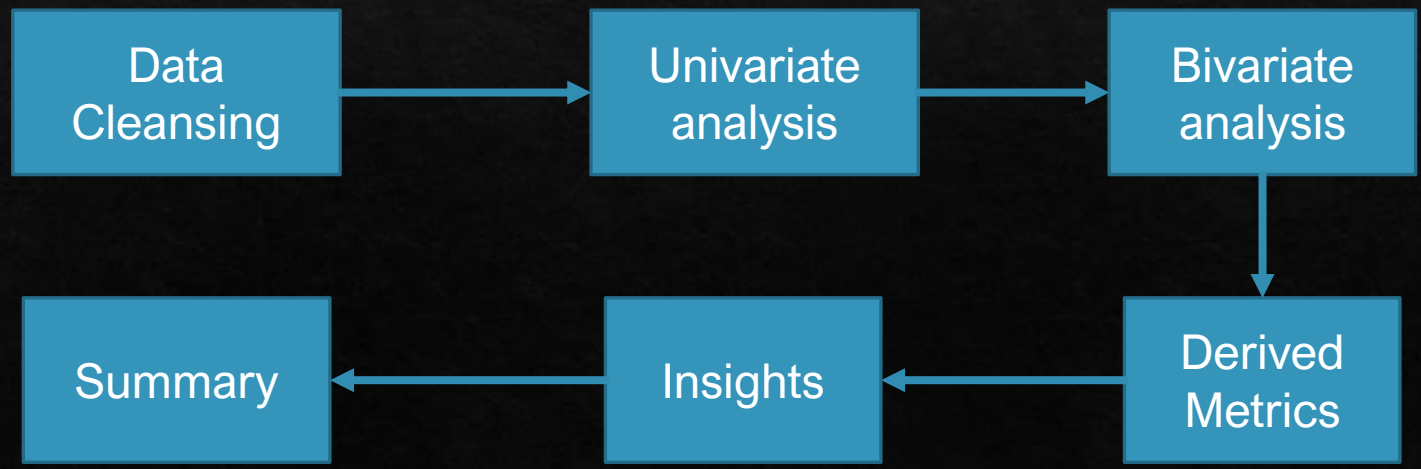
- Problem statement
- Analysis approach
- Data Cleansing
- Data Analysis from
  - Univariate analysis
  - Segmented Univariate Analysis
  - Bivariate Analysis
    - Categorical Bivariate
    - Continuous Bivariate
  - Derived Metrics analysis
- Recommendation for loan officer

Problem Statement

To enable a loan officer to determine credit worthiness of an applicant based on his application and to make a decision on loan approval.

Analysis Approach

- Aim: To find patterns, based on past performance, that indicate credit-worthiness of an applicant. To determine what factors drive
  - loan default and
  - timely payments
- Data Sourcing : as provided in the dataset
-



Data Cleansing

- **Handle Missing Values:** Identified and handled missing / null values in the dataset. Removed rows with entirely null values. We decided not to impute mean in place of any missing values as that is not the most suitable metric. For one column (chargeoff_within_12_mths), we imputed 0 in the place of NA. However, this column was subsequently dropped since there was just a single value 0 in the column. Removed columns that were entirely made up of single values such as 0 or 1 or “INDIVIDUAL” which provided no value add to the insight. We checked for null values across rows but found that to be insignificant.

- **Remove Irrelevant Columns:** Identified and removed columns that are irrelevant to the loan approval decision. For example, columns like "member_id," "title," or "zip_code" that do not provide meaningful insights were dropped. We are retaining the id column as a primary key and to run counts for frequency distribution.

- **Convert Data Types:** Checked if any columns need to be converted to the correct data types. Variables like "int_rate" and "revol_util" converted to different numeric types.

- **Deal with Outliers:** Identified outliers in the dataset. Outliers can affect the analysis and modeling process. However, picking the median instead of average to neglect the impact of outliers. We dropped the currently open loans since that information did not lend itself well to the needed analysis.

List of columns

- At the end of data cleansing, we were left with the following columns:

```
['id', 'loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'term',
'int_rate', 'installment', 'grade', 'sub_grade', 'emp_length',
'home_ownership', 'annual_inc', 'verification_status',
'issue_d', 'loan_status', 'purpose', 'zip_code', 'dti',
'delinq_2yrs', 'earliest_cr_line', 'inq_last_6mths',
'open_acc', 'pub_rec', 'revol_bal', 'revol_util', 'total_acc',
'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp',
'total_rec_int', 'recoveries', 'collection_recovery_fee',
'last_pymnt_d', 'last_pymnt_amnt', 'next_pymnt_d',
'last_credit_pull_d']
```

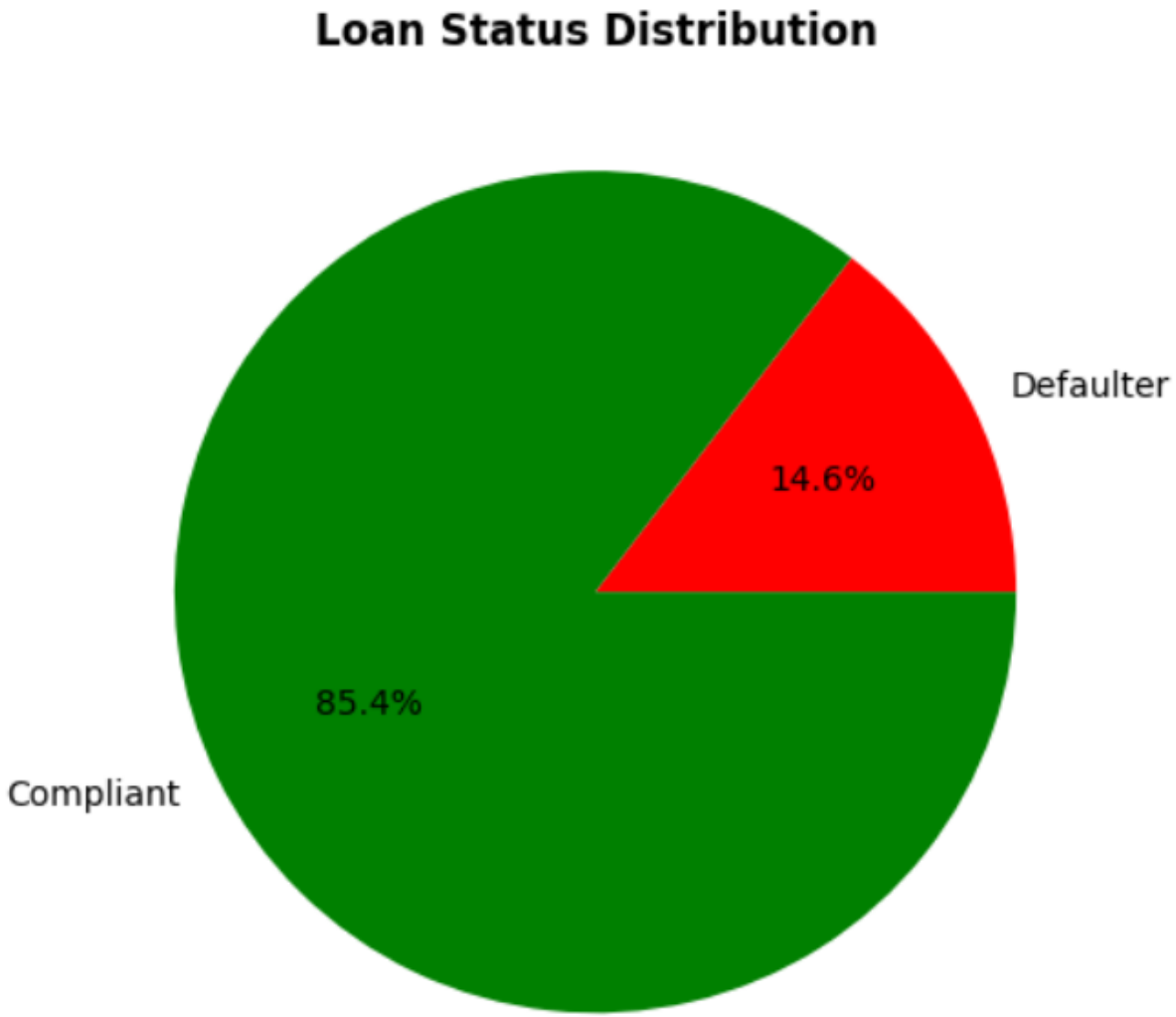
Univariate Analysis

Univariate analysis is a technique that focuses on examining and summarizing the characteristics of a single variable. It involves analyzing and understanding the distribution, central tendency, dispersion, and shape of the variable's values. We explore one variable at a time without considering the relationships or dependencies with other variables.

Univariate Analysis - 1

Loan Status :

After dropping the current loans, we have 85% of compliant loans and 15% bad debts.

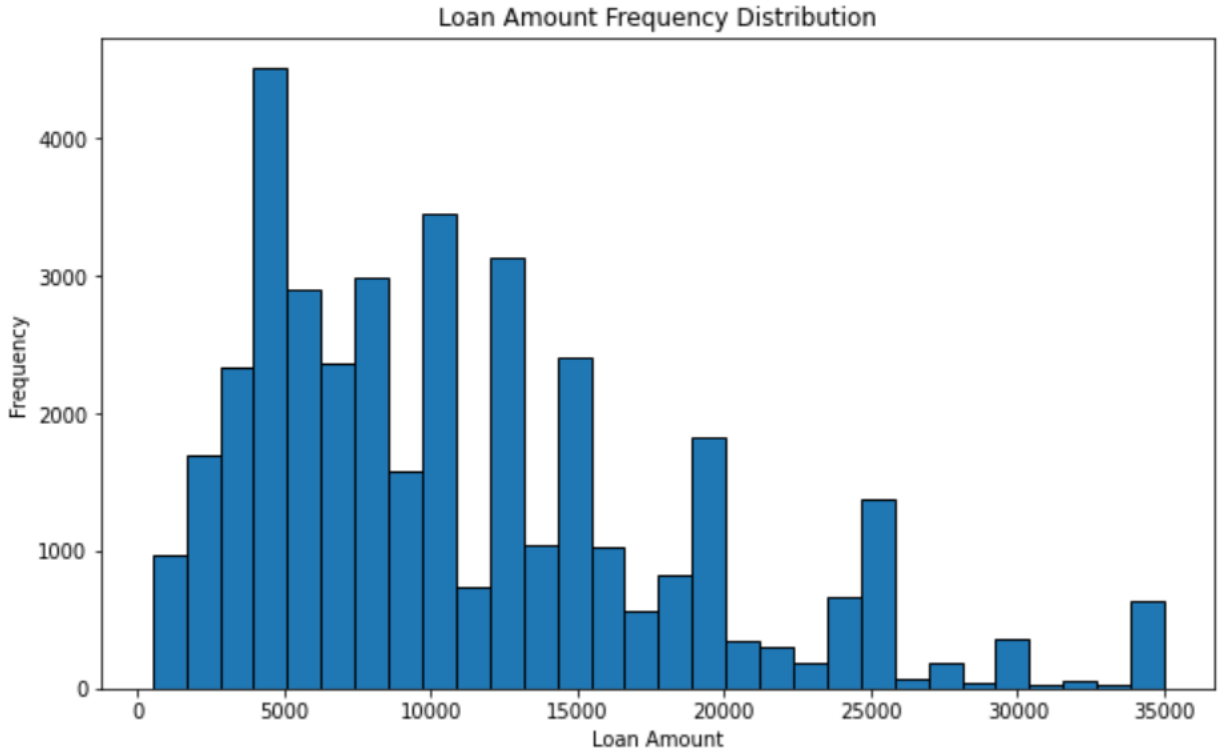


Univariate Analysis - 2

Distribution of loan amount :

we see that majority of the loans are between $5000 - $15000

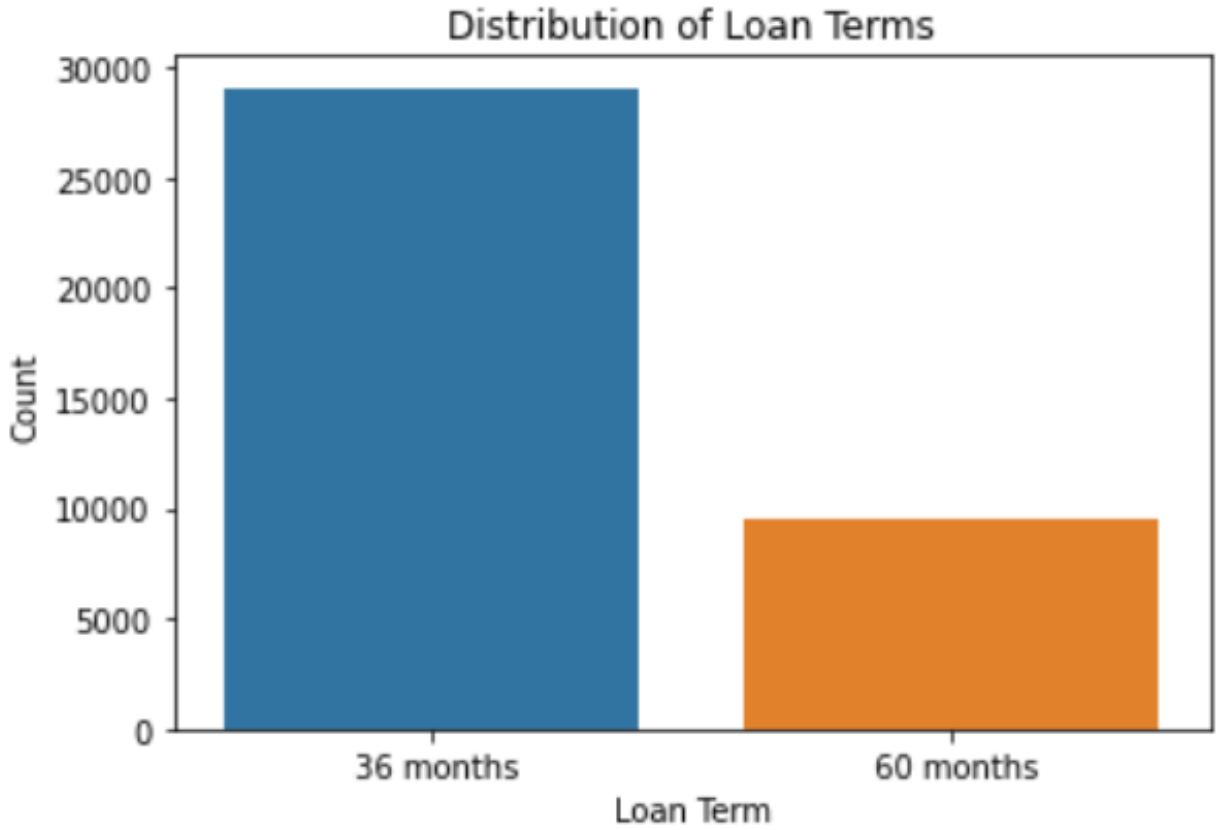
```
Loan Amount Statistics
Mean: $11047.02
Mode: $10000
Minimum: $500.0
25th Percentile: $5300.0
Median: $9600.0
75th Percentile: $15000.0
Maximum: $35000.0
```



Univariate Analysis -3

Loan Term

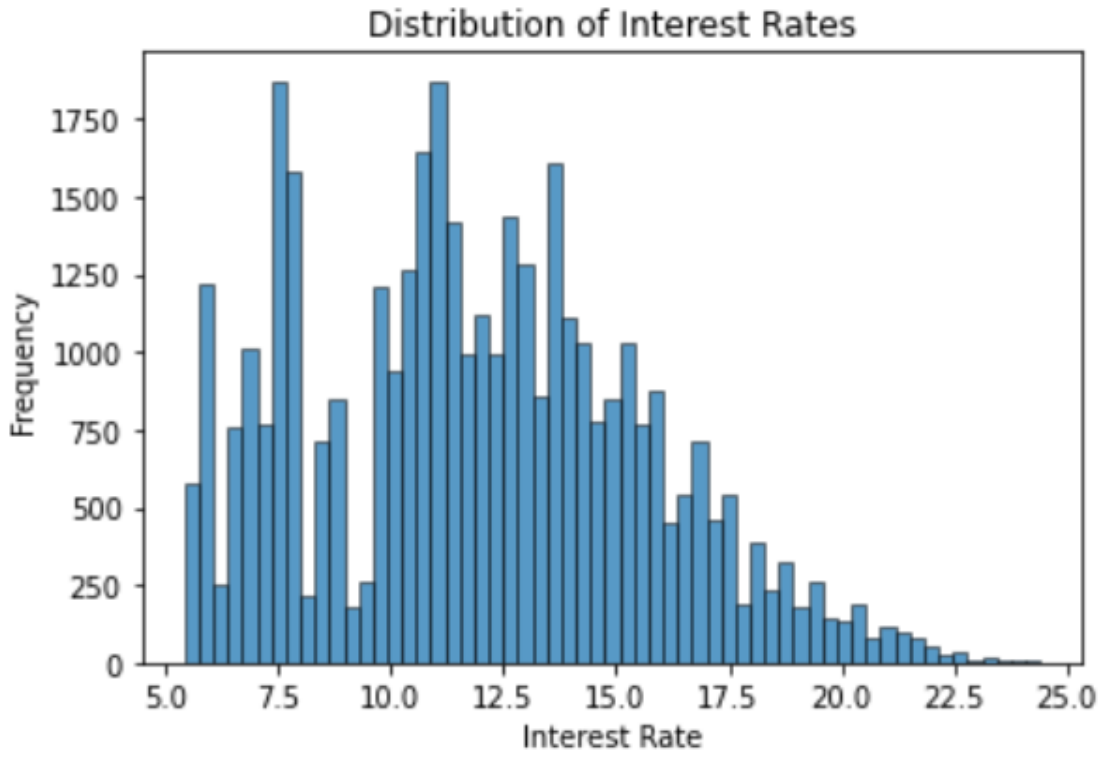
This graph indicates a strong preference amongst Lending Club's borrowers for a shorter duration loan.



Univariate Analysis - 4

Interest Rate distribution

The interest rates exhibit a somewhat normal distribution, with the majority of loans having rates in a specific range.



Segmented Univariate Analysis

Segmented univariate analysis is a technique that involves conducting separate univariate analyses on different segments or subgroups of a dataset based on a specific criterion or variable. The purpose of segmented univariate analysis is to explore and compare the characteristics of a single variable across different segments to identify any patterns, differences, or relationships that may exist.

Segmented Univariate Analysis - 1

Loan Amount by Loan Grade

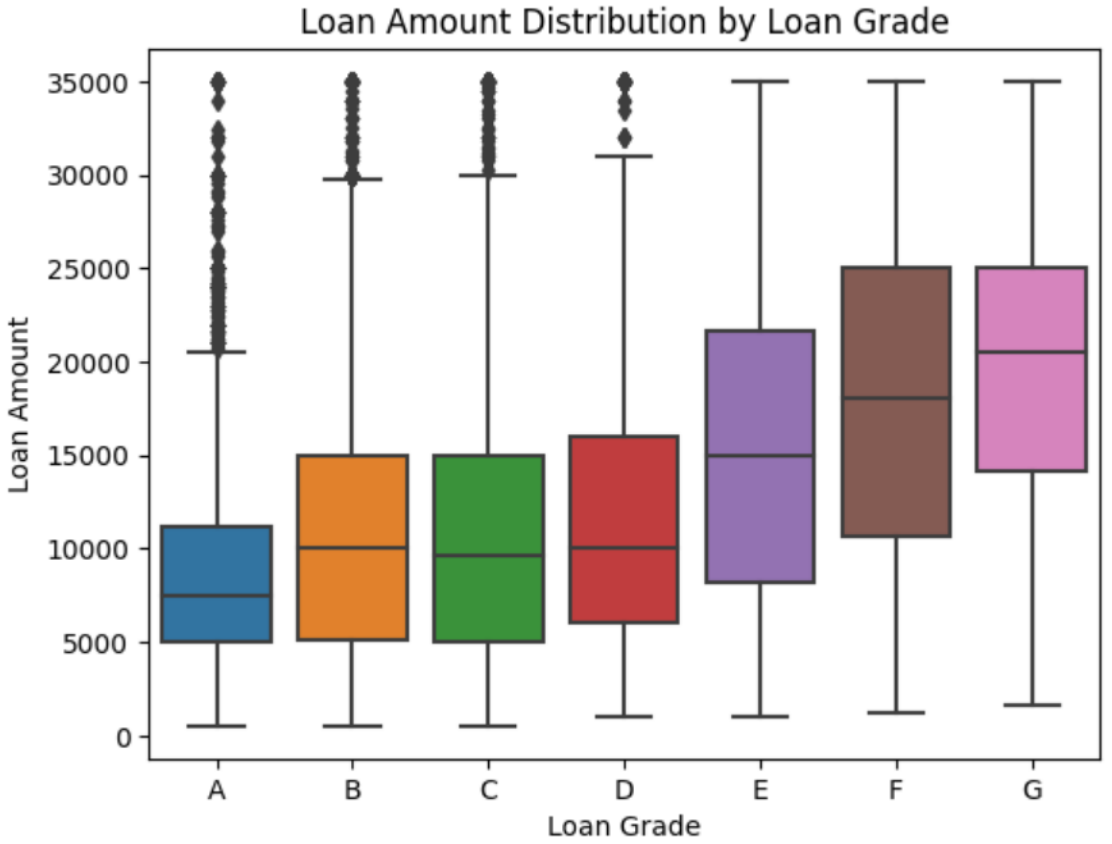
Insight - plot helps to compare the central tendency and spread of loan amounts across different loan grades.

The minimum loan amounts are very similar across grades A,B,C and increase as we move towards G.

The typical loan amounts given keep increasing as we move from A to G.

It appears that Lending Club slots people with lesser credit worthiness towards grade A, and as the credit worthiness improves, it moves them towards grade G.

However, we see a lot many outliers in grade A, some as high as USD 35000 as opposed to a typical loan of ~USD 7000. This is unacceptable and probably a leading driver of default.



Loan Amount Statistics by Loan Grade			
grade	**mean**	**median**	**mode**
A	$ 8,619	$ 7,500	$ 10,000
B	$ 10,935	$ 10,000	$ 10,000
C	$ 10,816	$ 9,600	$ 10,000
D	$ 12,138	$ 10,000	$ 10,000
E	$ 15,680	$ 15,000	$ 20,000
F	$ 18,095	$ 18,000	$ 25,000
G	$ 20,253	$ 20,500	$ 25,000

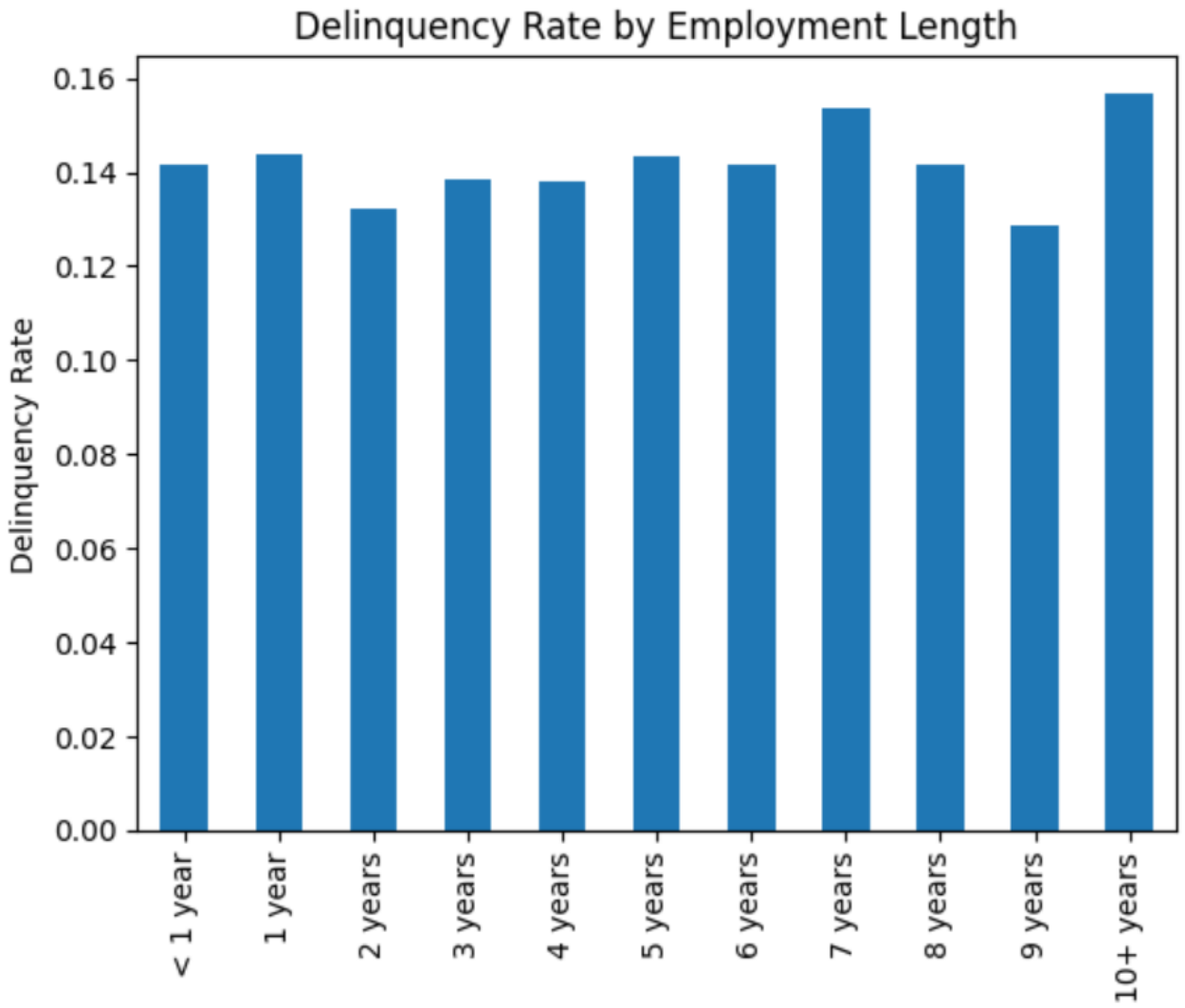
Segmented Univariate Analysis - 2

Delinquency Rate by employment length

Analyzing the delinquency rate by employment length can provide insights into the relationship between employment stability and loan repayment behavior.

Higher delinquency rates in certain employment length categories may indicate a higher risk of default for loans associated with those lengths.

Lenders may consider employment length as a factor in assessing the creditworthiness of loan applicants, as it can provide an indication of stability and ability to repay the loan.




```
Mean Delinquency Rate      : 14%
Median Delinquency Rate    : 14%
Mode Delinquency Rate      : 13%
```

Bivariate Analysis

Bivariate analysis is a statistical analysis technique that involves examining the relationship between two variables. It focuses on understanding how changes in one variable are related to changes in another variable. In bivariate analysis, the variables are usually categorical or continuous, and the goal is to determine whether there is a correlation, association, or dependency between the variables.

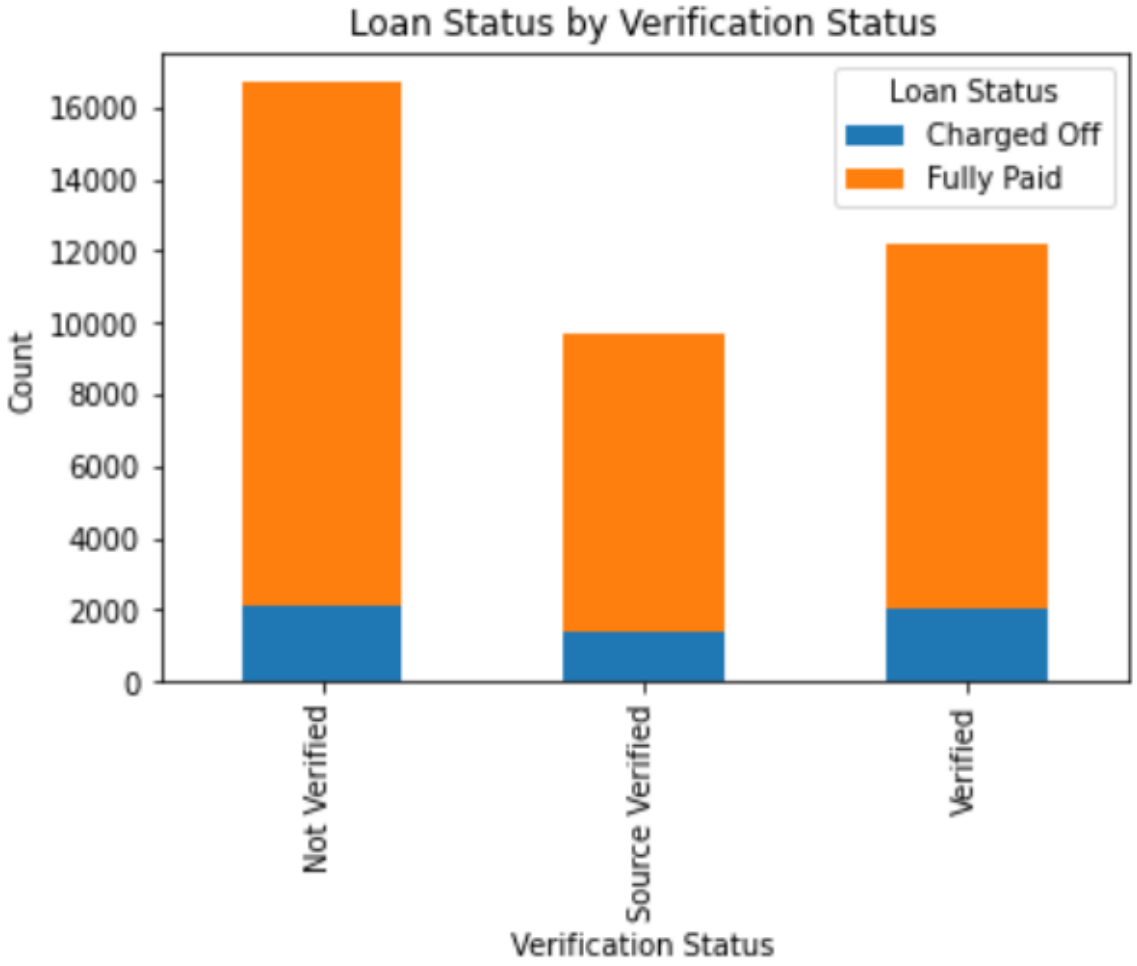
- When we compare two dimensions, we use categorical bivariate analysis
- When we compare two measures, we use continuous bivariate analysis

Categorical bivariate analysis - 1

Loan status by verification status

Insight- Verified loans have a higher likelihood of being fully paid compared to loans with other verification statuses, which could indicate the impact of verification on loan repayment.

However, we notice that in this situation, the verification is not making a difference in the status of the loan as we see similar levels of 'Charged Off' loans across verification status.

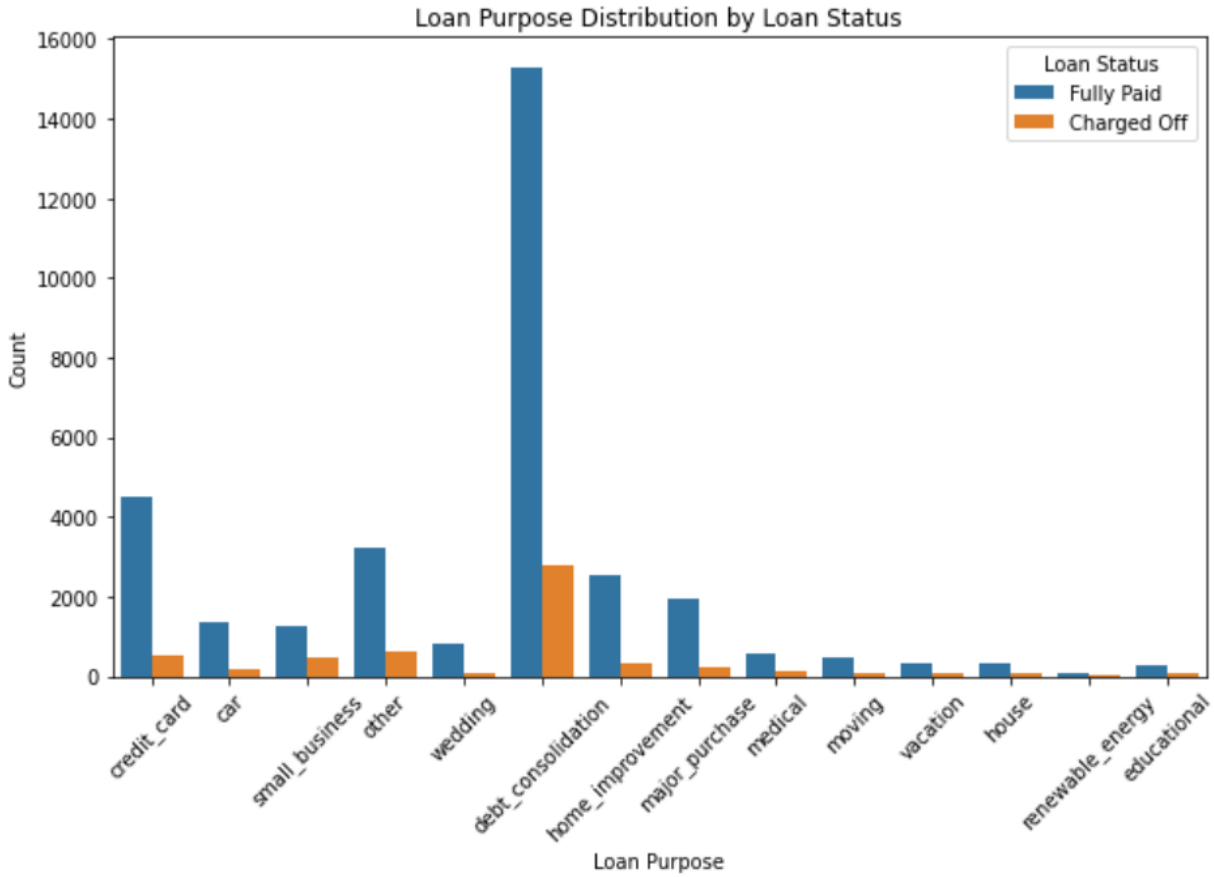


Categorical bivariate analysis - 2

Loan Purpose distribution by loan status

Debt Consolidation seems to be the major reason why borrowers reach out to Lending Club.

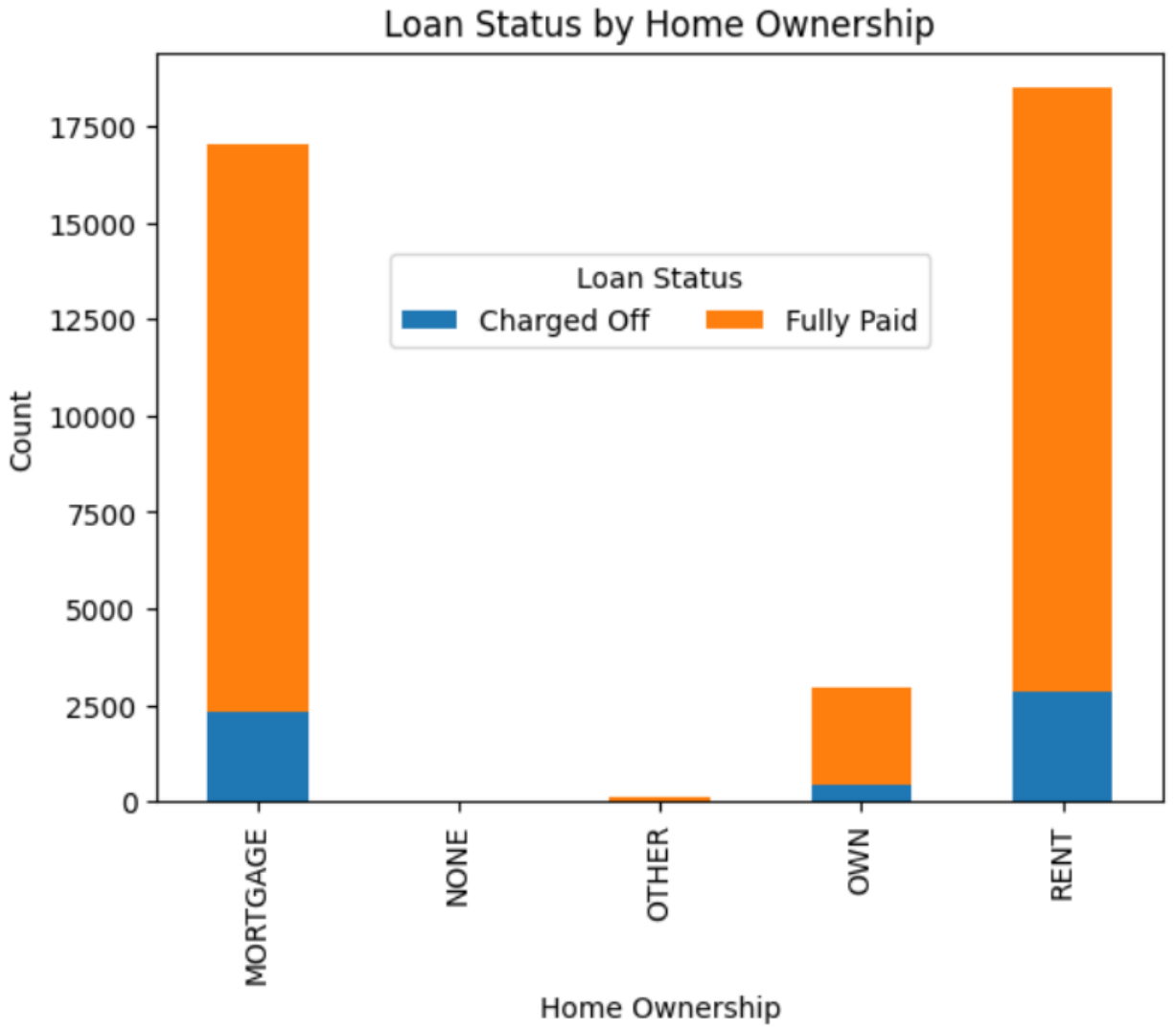
Understandably, that has the highest proportion of charged off loans as well.



Categorical bivariate analysis - 3

Loan Status by Home Ownership

Majority of Lending Club's borrowers are on rent or have mortgaged their properties. However, the ownership status does not seem to have a bearing on the charged off loans, we can not see a distinct pattern.



Continuous Bivariate Analysis - 1

Log-Log plot of Loan Amount vs Annual Income

There appears to be a positive correlation between annual income and loan amount, suggesting that borrowers with higher incomes tend to qualify for larger loans.

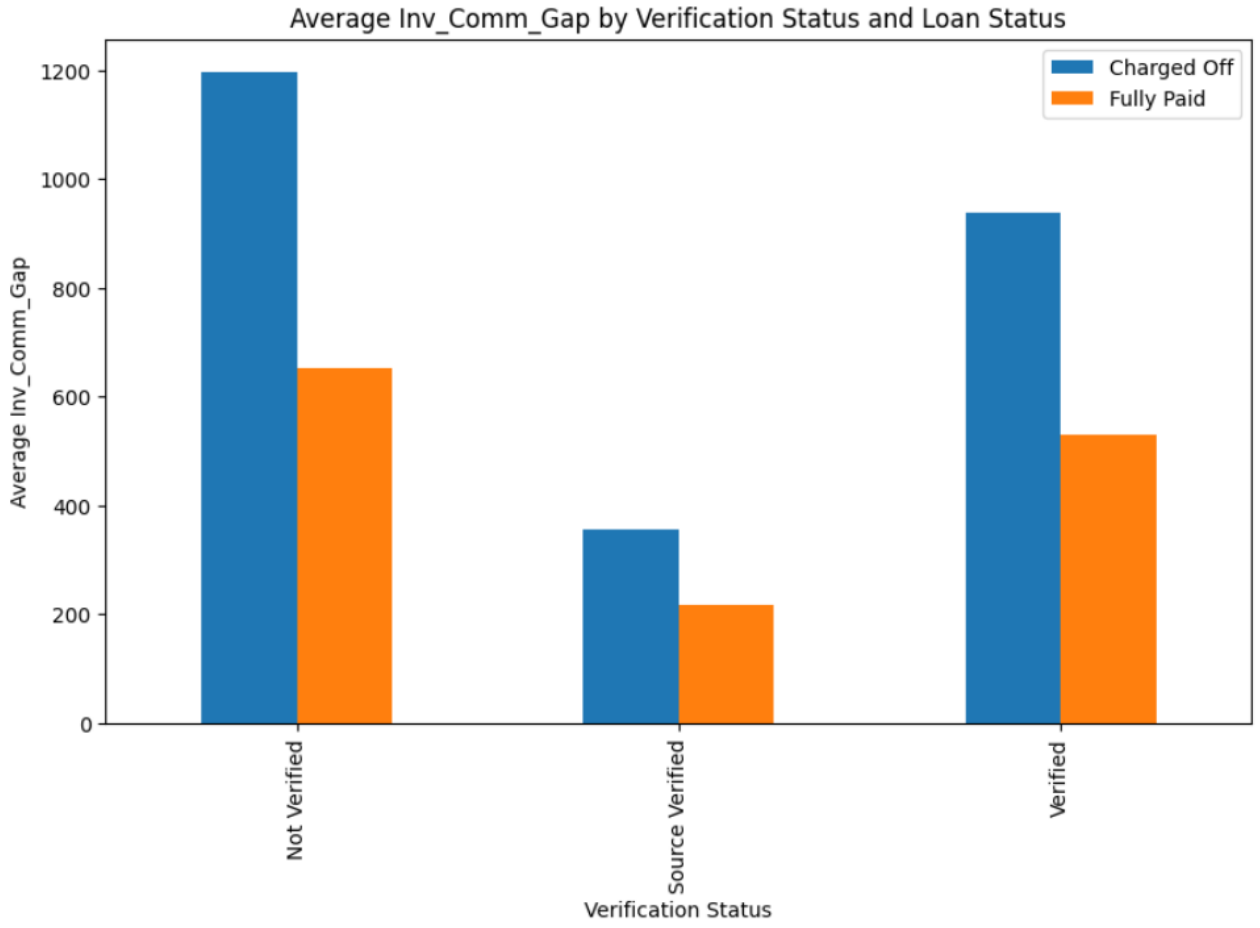
Derived Metrics

We create derived variables based on information available in the data set to fathom more meaningful interpretations.

Derived Metrics - 1

We now create a derived metric "Inv_Comm_Gap" as the difference between "funded_amnt" and "funded_amnt_inv" and analyze it against "verification_status" and "loan_status," . The idea is to see what is the risk exposure Lending Club is holding against different loan status and verification status for those amounts where the investors are not covering for it.

By examining the plot, we can gain insights into the average difference between the funded amount and the invested amount (Inv_Comm_Gap) for different combinations of verification status and loan status. This analysis helps understand how the Inv_Comm_Gap varies based on these factors, providing information about the discrepancy between the funded and invested amounts in different loan scenarios.



Recommendations

- The Loan officer must ensure he verifies all applications
- Stick to the rules of the grade, for instance, do not give loans beyond 75% of the grade value for people on Rent and having Mortgages on Property
- If the Investors are not covering the loan entirely, increase the loan rate by a few percentage points (2-5 pp.).