Lending Club Case Study

Submitted By:
Shruti Mehra
Sayantan Mondal

Contents

- Problem Statement
- Data Description
- Data Cleaning & Pre-processing
- Univariate Analysis
- Bivariate Analysis
- Correlation Analysis
- Recommendations

Problem Statement

Lending Club, a Consumer Finance marketplace specializing in offering a variety of loans to urban customers, faces a critical challenge in managing its loan approval process. When evaluating loan applications, the company must make sound decisions to minimize financial losses, primarily stemming from loans extended to applicants who are considered **“Risky“.**

These financial losses, referred to as **Credit Losses**, occur when borrowers fail to repay their loans or default. In simpler terms, borrowers labeled as **“Charged-Off"** are the ones responsible for the most significant losses to the company.

The primary objective of this exercise is to assist Lending Club in mitigating credit losses. This challenge arises from two potential scenarios:

1. **Identifying applicants likely to repay their loans is crucial, as they can generate profits for the company through interest payments. Rejecting such applicants would result in a loss of potential business.**
2. **On the other hand, approving loans for applicants not likely to repay and at risk of default can lead to substantial financial losses for the company.**

The objective is to pinpoint applicants at risk of defaulting on loans, enabling a reduction in credit losses. This case study aims to achieve this goal through Exploratory Data Analysis (EDA) using the provided dataset.

In essence, 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.

Data Understanding

Lending Club provided us with customer’s historical data. This dataset contained information pertaining to the borrower’s past credit history and Lending Club loan information. The total dataset consisted of over 39717 records and 111 columns, which was sufficient for our team to conduct analysis. Variables present within the dataset provided an ample amount of information which we could use to identify relationships and gauge their effect upon the success or failure of a borrower fulfilling the terms of their loan agreement. It contains the complete loan data for all loans issued through the time period 2007 to 2011.

Primary Attribute

Loan Status: The Principal Attribute of Interest (loan_status). This column consists of three distinct values:

- **Fully-Paid:** Signifies customers who have successfully repaid their loans.
- **Charged-Off:** Indicates customers who have been labeled as "Charged-Off" or have defaulted on their loans.
- **Current:** Represents customers whose loans are presently in progress and, thus, cannot provide conclusive evidence regarding future defaults.

For the purposes of this case study, rows with a "Current" status will be excluded from the analysis.

Data Cleaning & Preprocessing

1. **Loading data from loan CSV:** While loading the dataset, some of the variables had mixed data types so they have to be converted accordingly as per analysis.

2. **Checking for null values in the dataset:** There are many columns with null values. So they had to be dropped as they won't play a role in the analysis of the dataset.

3. **Checking for unique values:** If the column has only a single unique value, it does not make any sense to include it as part of our data analysis. We need to find out those columns and drop them from the dataset. 9 columns had such unique values and they were removed.

4. **Checking for duplicated rows in data:** No duplicate rows were found.

5. **Dropping Records and Columns:**

- Dropped records where **loan_status="Current"** as the loan in progress cannot provide us insights as to whether the borrower is likely to default or not.

- Dropping columns having post approval features like **collection_recovery_fee, delinq_2yrs, desc, earliest_cr_line, emp_title, id, inq_last_6mths, last_credit_pull_d, last_pymnt_amnt, last_pymnt_d, member_id, open_acc, out_prncp, out_prncp_inv, pub_rec, recoveries, revol_bal, revol_util, title, total_acc, total_pymnt, total_pymnt_inv, total_rec_int, total_rec_late_fee, total_rec_prncp, url, zip_code** as these will not contribute to loan pass or fail.

- Dropped columns title,emp_title,desc as it contains textual data.

6. **Data Conversion**:

   Converted columns like debt to income (dti), funded amount (funded_amnt), funded amount investor (funded_amnt_inv) and loan amount (loan_amnt) to float to match the data.

   Converted loan date (issue_d) to DateTime (format: yyyy-mm-dd).

7. **Outlier Treatment:**

   Checked outlier values in continuous columns like loan_amnt, int_rate, annual_inc, dti via box plot. Annual income has outliers. the annual_inc is increasing in exponentially after 95th percentile. Thus removed values after than 95th percentile.

8. **Handling missing values in columns:**

   Replaced missing values of emp_length, public_bnkruptcy column

9. Derived Column:

Derived issue_m , issue_y, issue_q columns from issue_d columns for better analysis across months , year and quarter which will help in making informed business decisions..

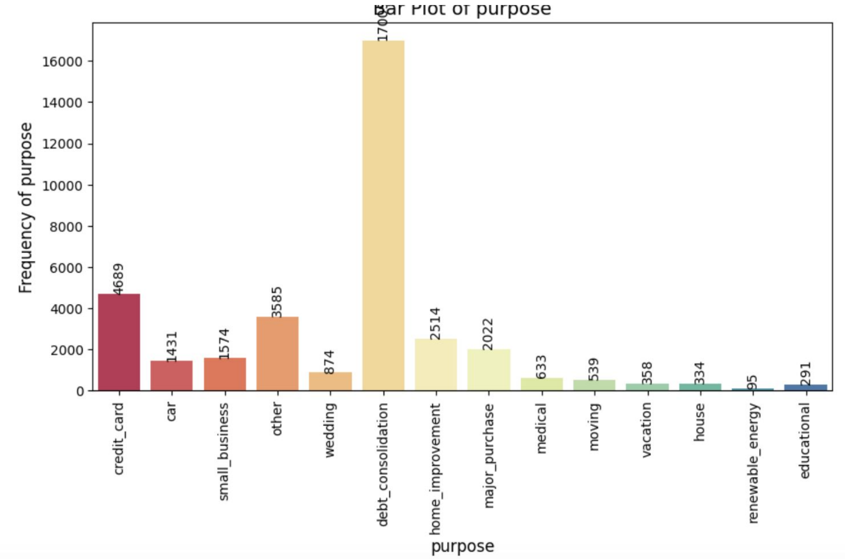
Variables identified for analysis:

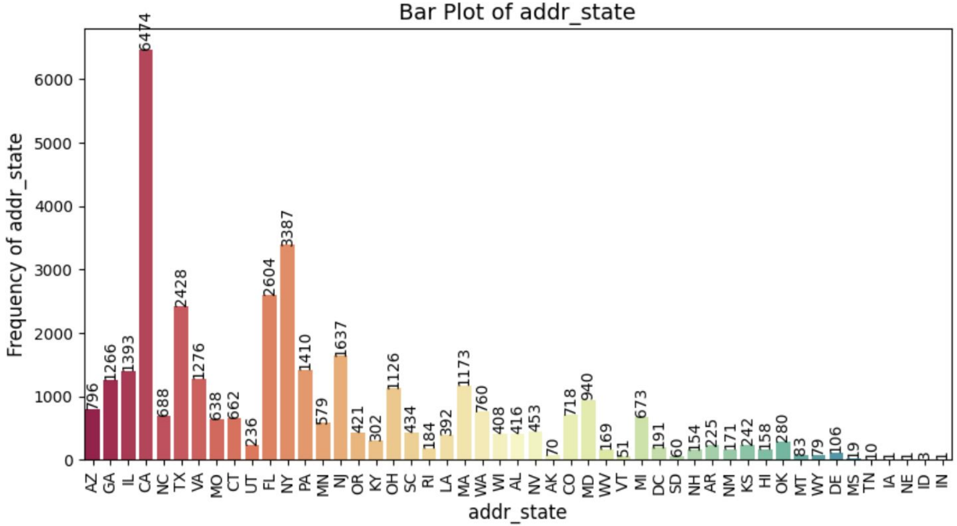
Ordered Categorical Variables: 1. Grade 2. Sub grade 3. Term 4. Employment length (emp_length) 5. Issue year (issue_y) 6. Issue month (issue_m) 7. Issue Quarter (issue_q)

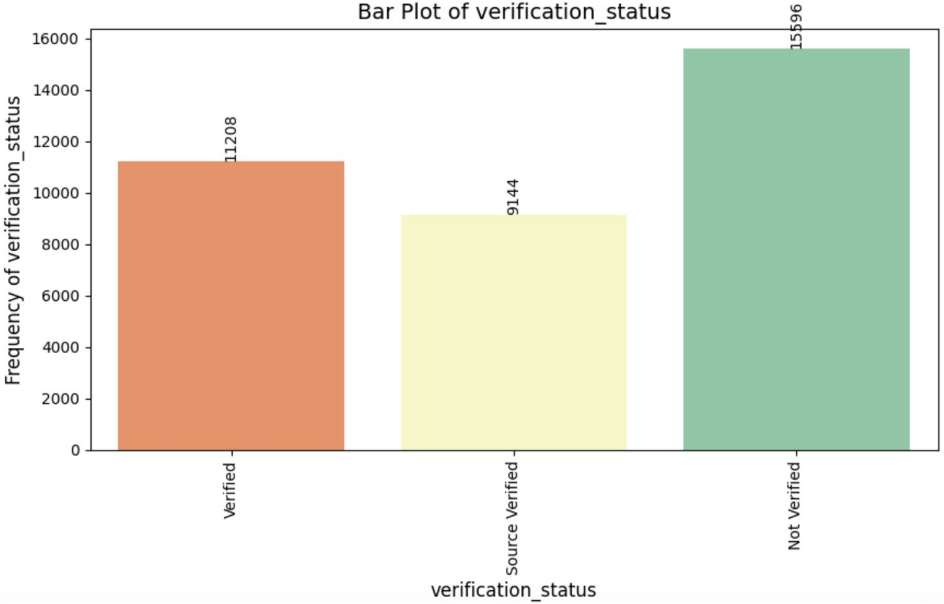
Unordered categorical data 1. Address State (addr_state) 2. Loan purpose (purpose) 3. Home Ownership (home_ownership) 4. Loan status (loan_status) 5. Verification_status

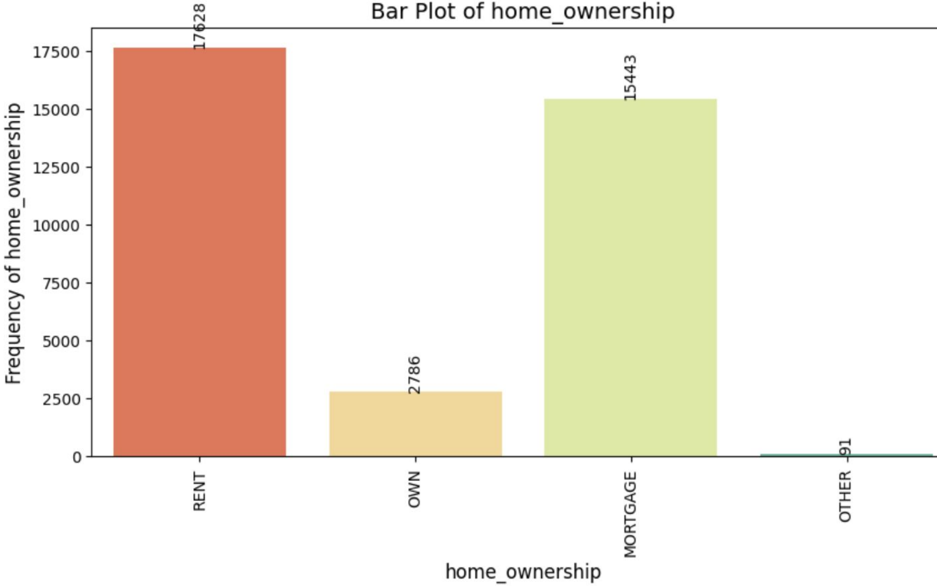
Quantitative variables 1. Loan_amount 2. Funded_amount 3. Funded_amount_inv 4. DTI 5. Annual_income 6. int_rate 7. installment 8. pub_rec_bankruptcies

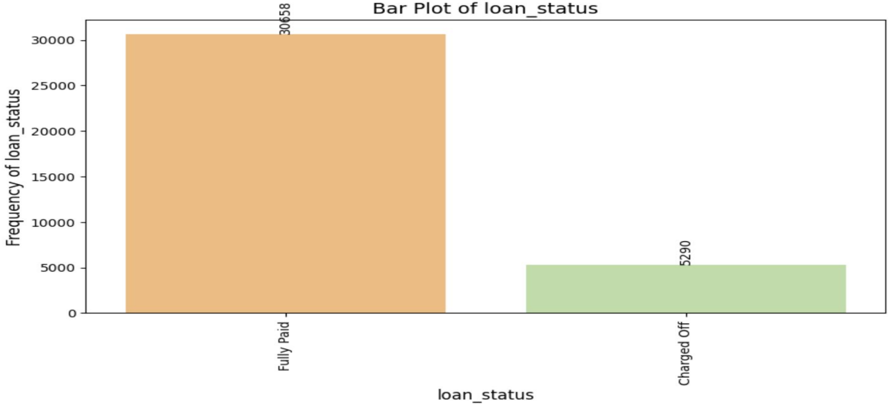
Univariate Analysis

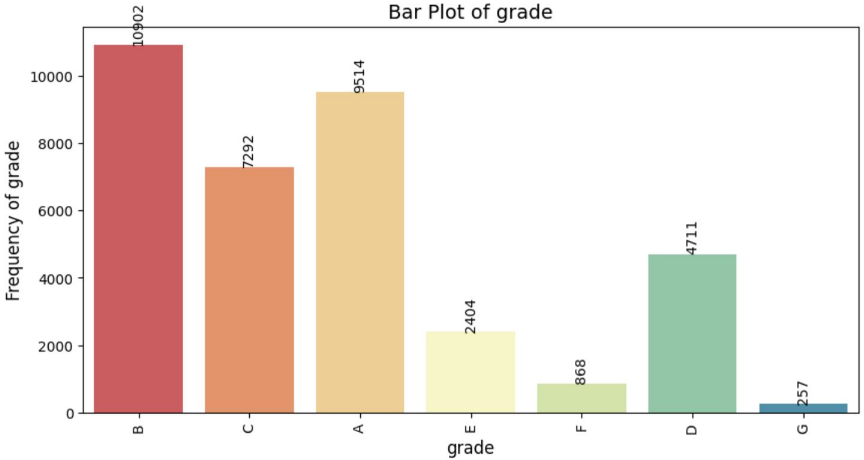




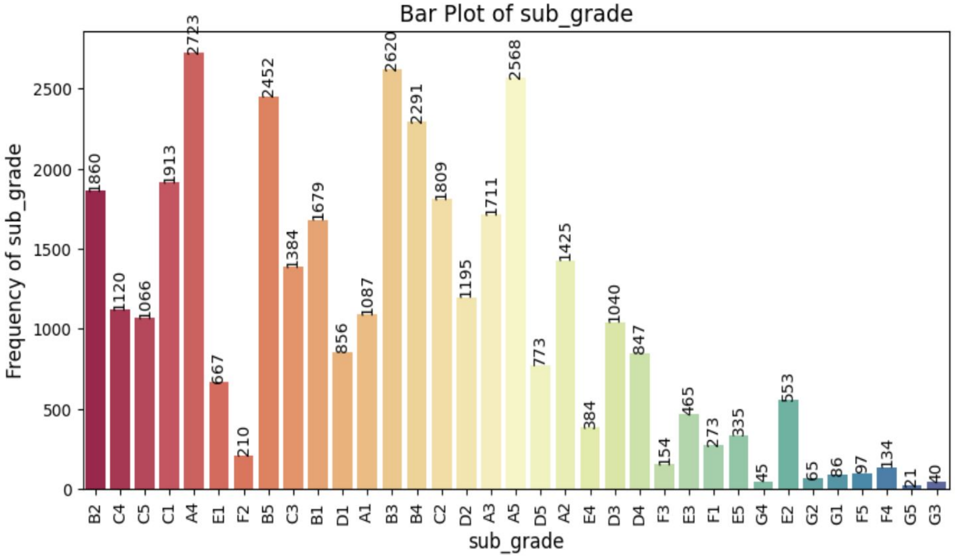

Bar Plot of verification_status
Frequency of verification_status
0
2000
4000
6000
8000
10000
12000
14000
16000
11208
9144
15596
Verified
Source Verified
Not Verified
verification_status


Bar Plot of home_ownership
Frequency of home_ownership
0
2500
5000
7500
10000
12500
15000
17500
17628
2786
15443
91
RENT
OWN
MORTGAGE
OTHER
home_ownership

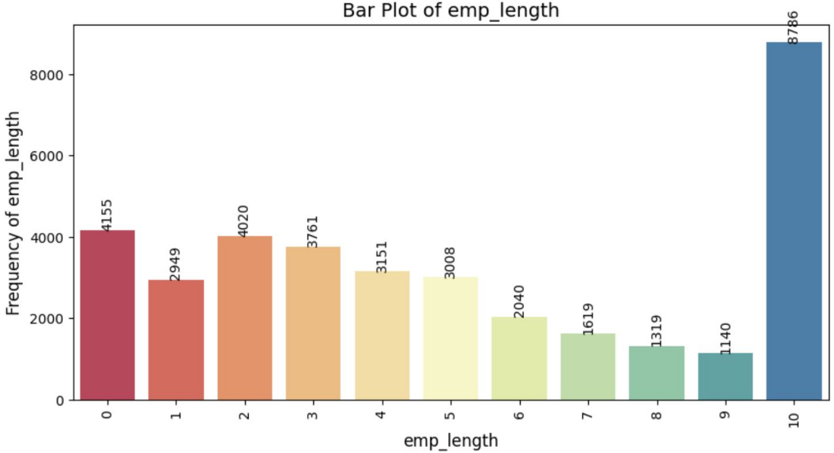

Bar Plot of loan_status
Frequency of loan_status
0
5000
10000
15000
20000
25000
30000
30658
5290
Fully Paid
Charged Off
loan_status



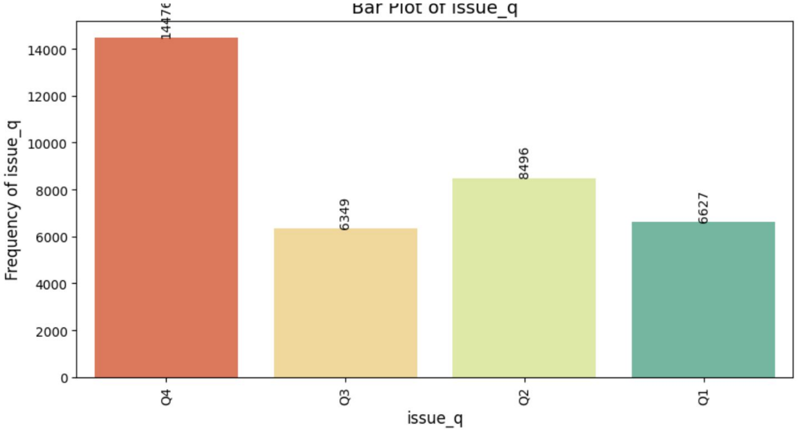
Most borrowers fall under A and B grades as compared to any other grades

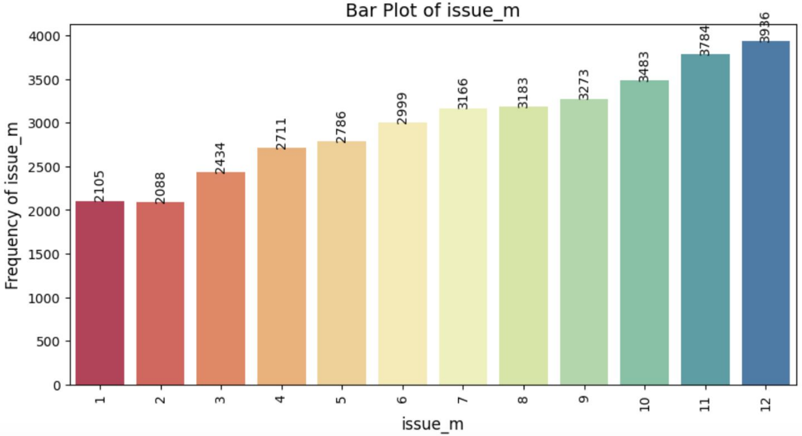


Most borrowers belong to A4, A3, B5 sub grades



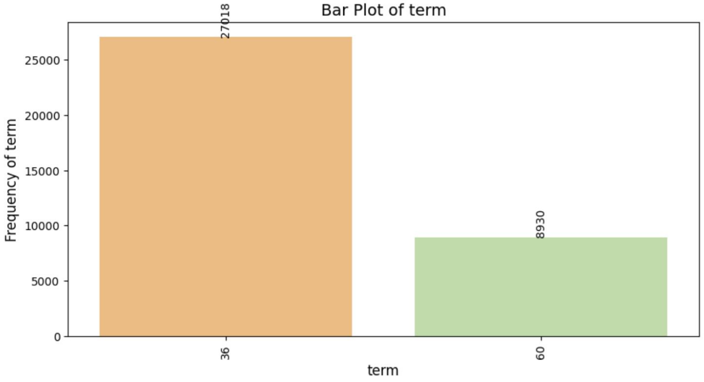
Inference : Most Borrowers have mostly 10+ years of employment length.



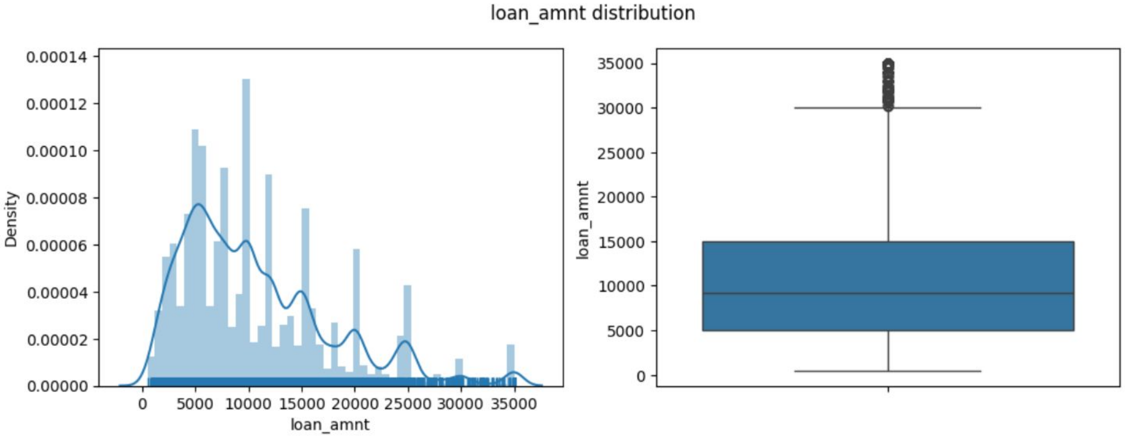


Inference :

- **The number of loans issued has doubled from every year.**
- **There are more loans issued in last 3 months every end of the year i.e., Oct, Nov and Dec.**
- **Quarter Q4 has highest number of loans issued.**

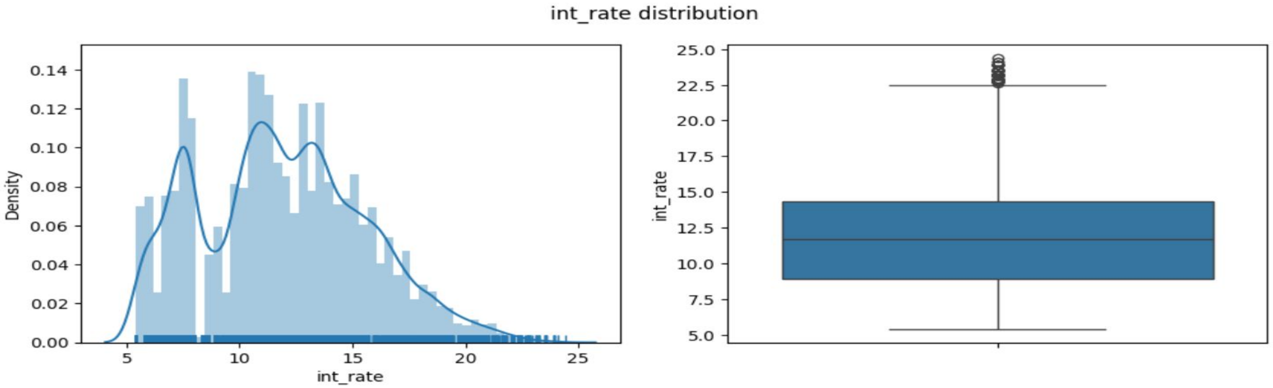


Borrowers have taken loan tenure of 36 months tenure more than 60 months.

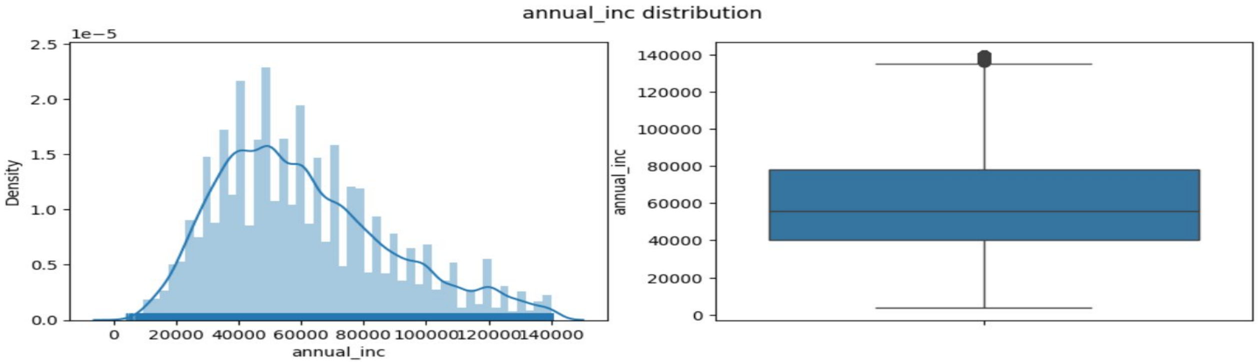


Observation:

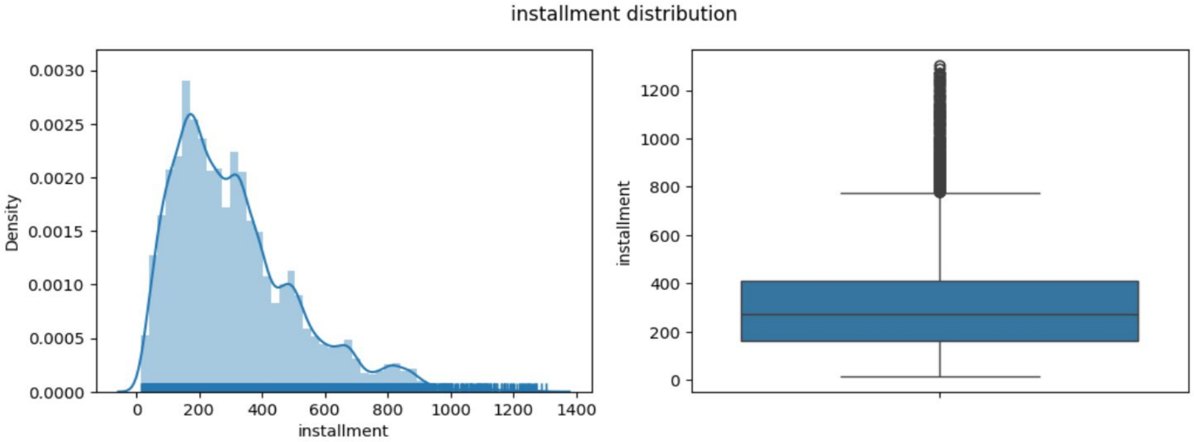
From loan amount data, we can say that most of them have taken loan amount of 5000 to 15000.



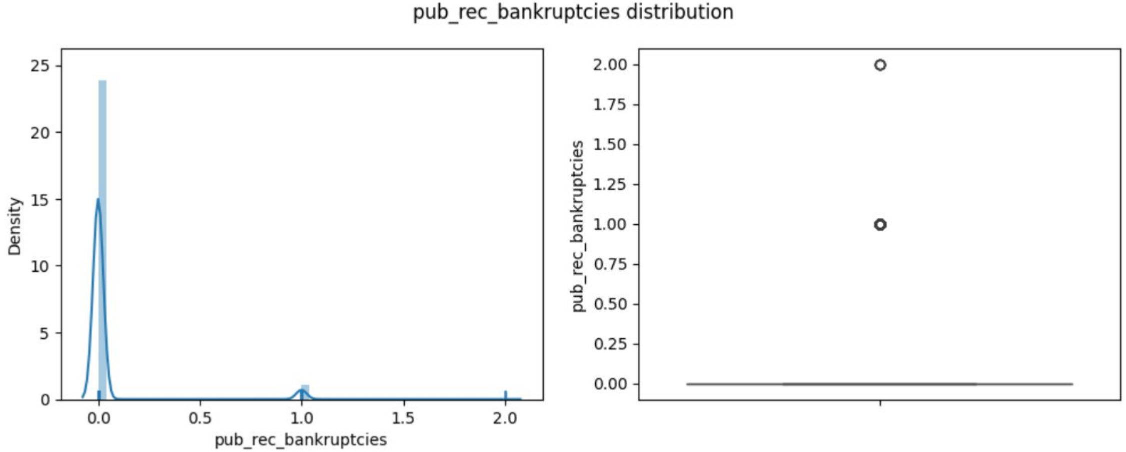
Inference :From interest rate data, we can say that most of the interest rate lies between 9% to 14.5%.



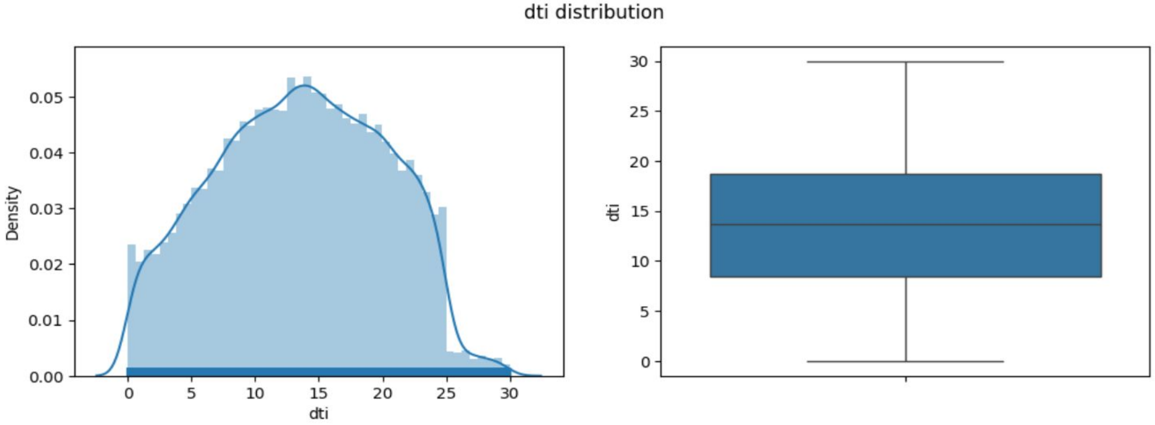
From annual income data, we can say that most of the borrower's annual income are in range of 40k to 60k.



Median value of installment is 280. The installment amount for borrower is betwen 200 - 400 USD



Majority of the loan applicants are in the category of not having an public record of bankruptcies



Loan applicants had very high debt-to-income ratios ranging between 9 to 19.

Observations and Inferences from Univariate Analysis

Ordered Categorical Variables

1. Grade B had the highest number of loan applicants, with a total of 10421 applicants
1. The majority of loan has a term of 36 months compared to 60 months.
2. Majority of borrowers have working experience greater than 10 years.
3. The year 2011 recorded the highest number of loan applications. This shows a positive trend in the number of applicants facing loan defaults over the years. Thus we can say that the loan approval rate is increasing with the time.
4. Most loans were taken during the 4th quarter mostly in October, November ,December.

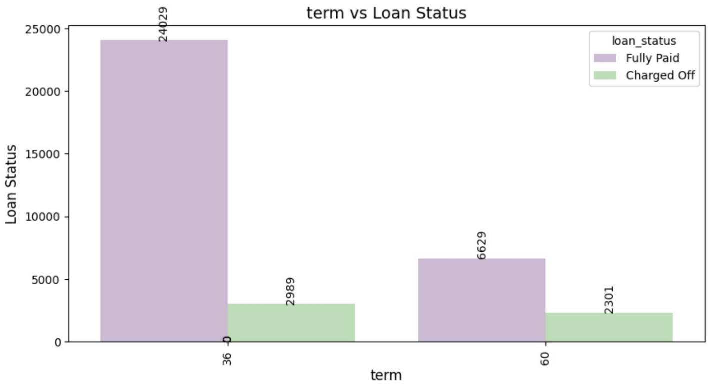
Unordered Categorical Variables

1. California had the highest number of loan applicants followed by New York.
2. Debt consolidation was the primary loan purpose for most loan applicants.
3. The majority of loan participants lived in rented houses or mortgage.
4. A significant number of loan participants were loan defaulters who were unable to clear their loans.
5. About 50% of the borrowers are verified by the company or have source verified.

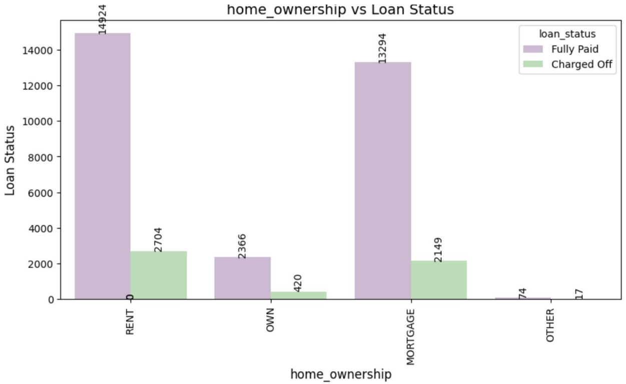
Quantitative variables

1. Most loan applicants had annual salaries less than 40,000 USD
2. Among loan participants, interest rate lies between 13%-17%.
3. Most of the borrowers have taken loan amount between 5000 and 15000.
4. Majority of the borrowers have very large debt compared to the income in the range of 10-15 DTI ratio.
5. It's observed that the majority of borrorowers had monthly installment amounts falling within the range of 160-440 USD.
6. Majority of the borrowers have no record of Public Recorded Bankruptcy.

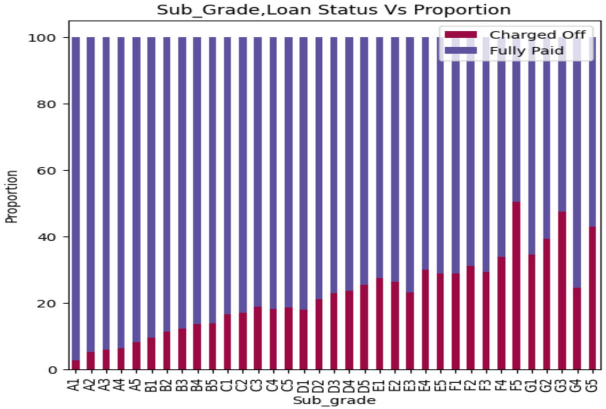
Segmented Univariate Analysis

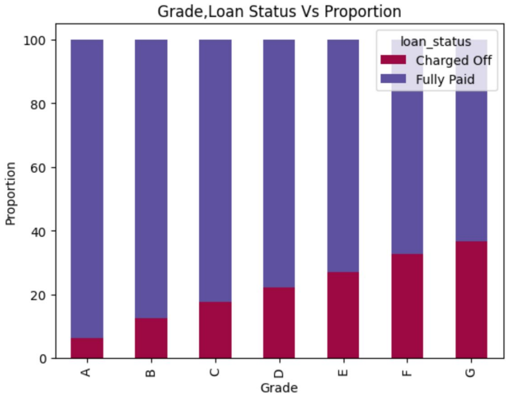


Inference: The 60 month term has higher chance of defaulting than 36 month term whereas the 36 month term has higher chance of fully paid loan.

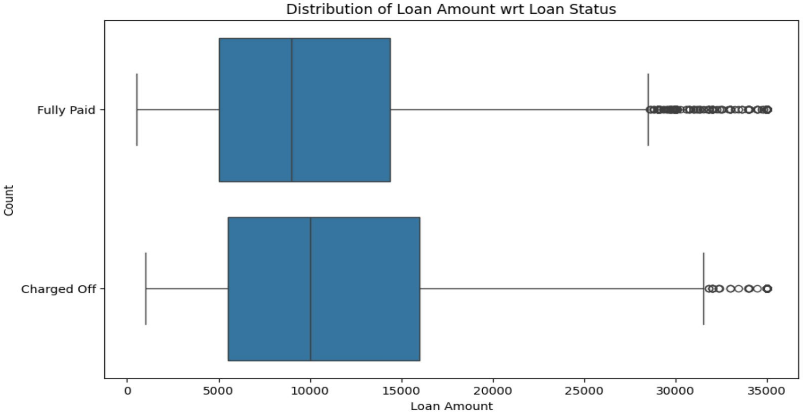


The loan applicants who live in a rented or mortgaged house are more likely to default

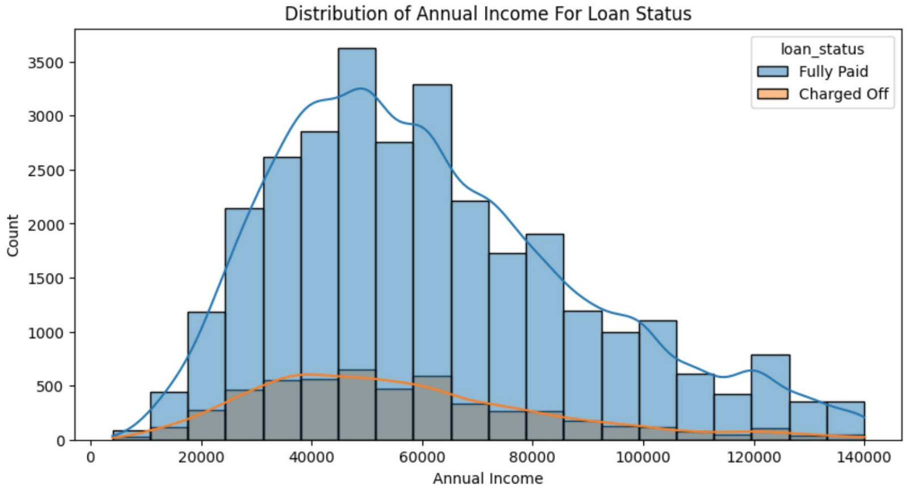




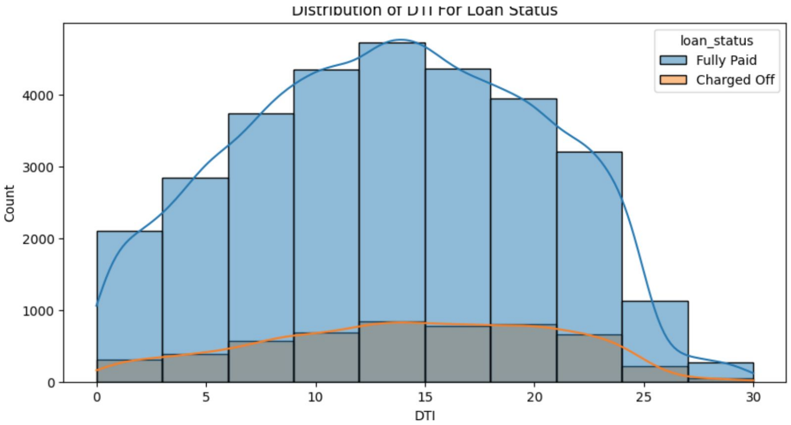
Inference : The loan applicants belonging to Grades B, C and D contribute to most number of "Charged Off" loans. Number of charge off loans increases as we move from grade A to G.



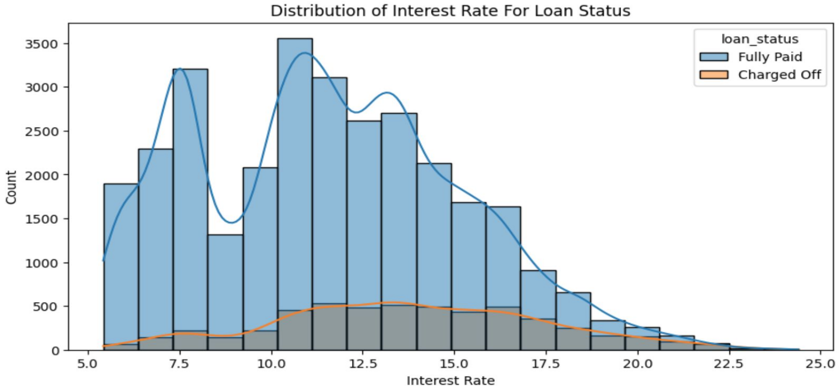
Inference: Large amount of loan has high chances of defaulting



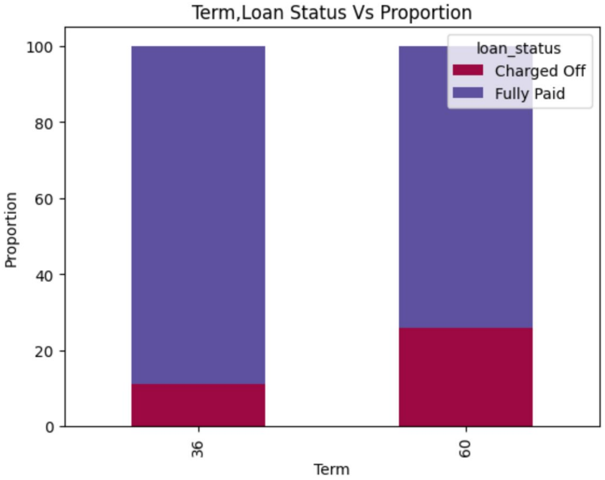
Inference: The loan applicants having annual_income less than 50000 amount are more likely to default.



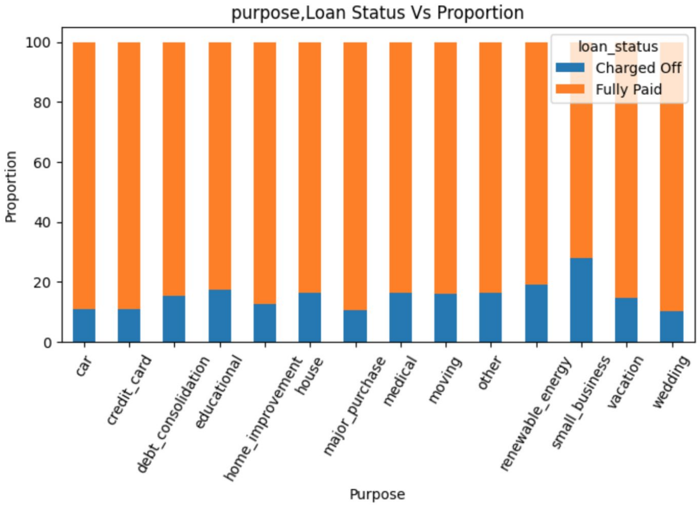
Inference: The Loan Status varies with DTI ratio, we can see that the loans in DTI ratio 10-15 have higher number of defaulted loan but higher dti has higher chance of defaulting.



Inference: As interest rate increases default rate also increases but declines after 17.5 % interest rate.



Inference: The 60 month term has higher chance of defaulting than 36 month term whereas the 36 month term has higher chance of fully paid loan.

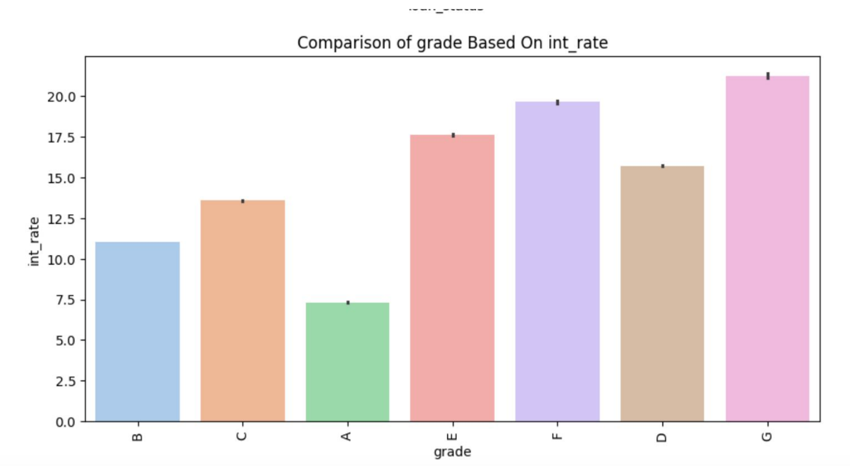


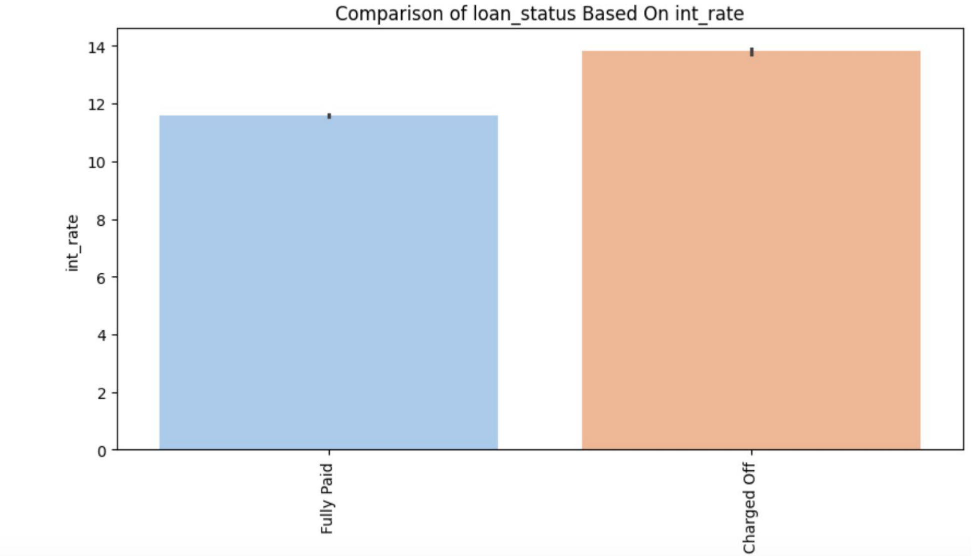
Inference: Debt Consolidation and small business is the most popular loan purpose and has highest number of defaulted loan.

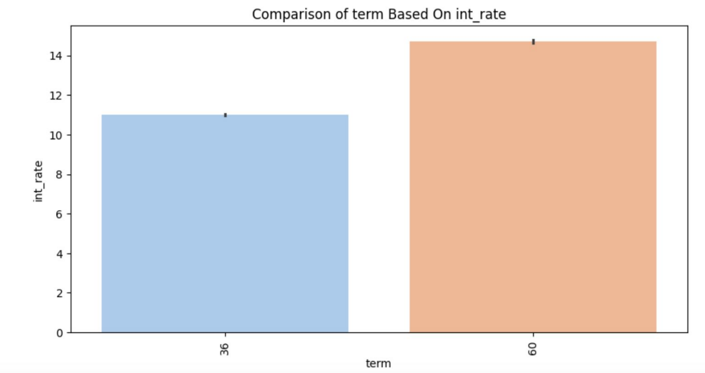
Observations and Inferences from Segmented Univariate Analysis¶

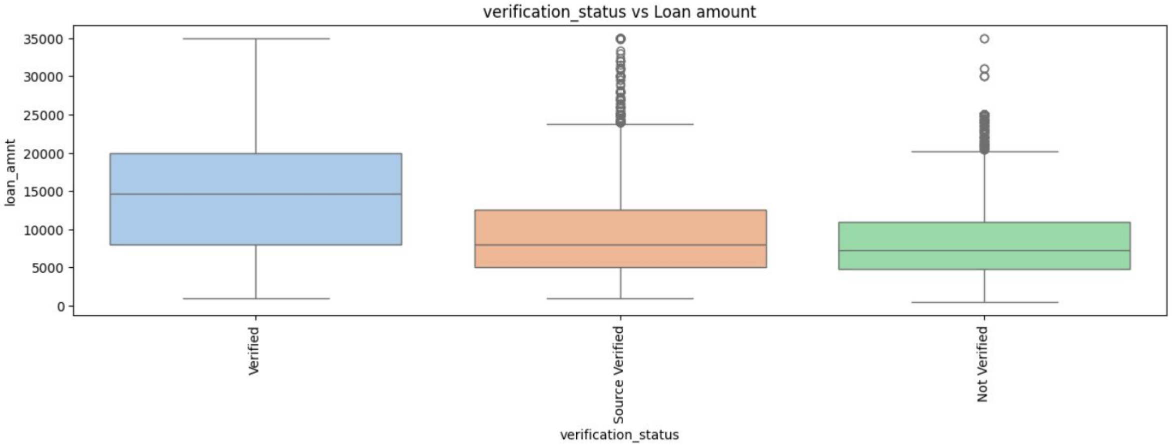
1. Debt Consolidation and small business is the most popular loan purpose and has highest number of fully paid loan and defaulted loan.
2. The loan applicants belonging to Grades B, C and D contribute to most number of "Charged Off" loans
3. The mean and 25% are same for both but we see larger 75% in the defaulted loan which indicate large amount of loan has higher chance of defaulting.
4. The 60 month term has higher chance of defaulting as compared to 36 month term.
5. The loans in 36 month term majorly consist of grade A and B loans whereas the loans in 60 month term mostly consist of grade B, C and D loans.
6. The Loan Status varies with DTI ratio, we can see that the loans in DTI ratio 10-15 have higher number of defaulted loan but higher dti has higher chance of defaulting.
7. The Defaulted loan are lower for the burrowers which own their property compared to on mortgage or rent.
8. Borrowers with less 50000 annual income are more likely to default and higher annual income are less likely to default.
9. The Fully paid loan are increasing exponentially with the time compared to defaulted loan.
10. The default loan amount increases with interest rate and shows are decline aftre 17.5 % interest rate.
11. The Employees with 10+ years of experience are likely to default and have higher chance of fully paying the loan

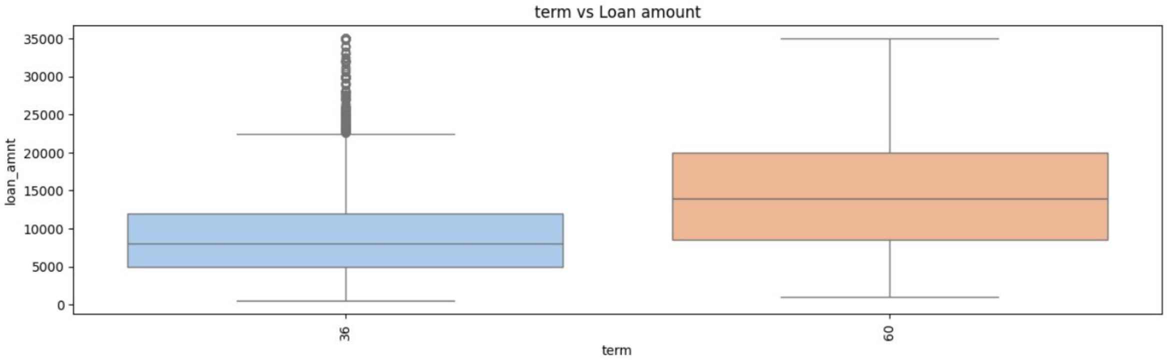
Bivariate Analysis

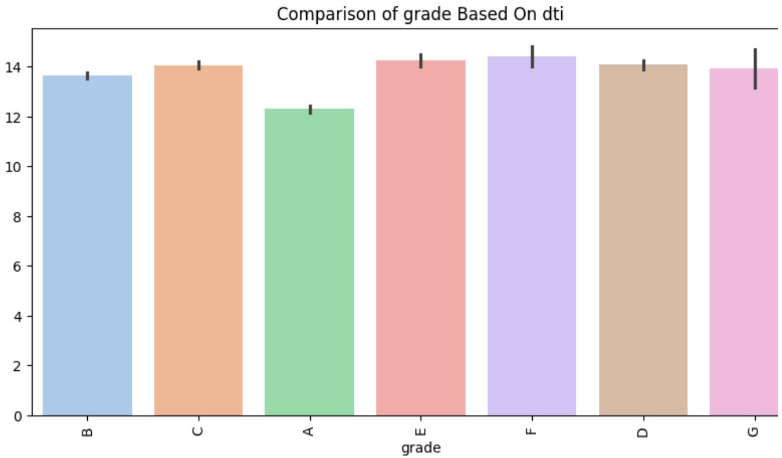




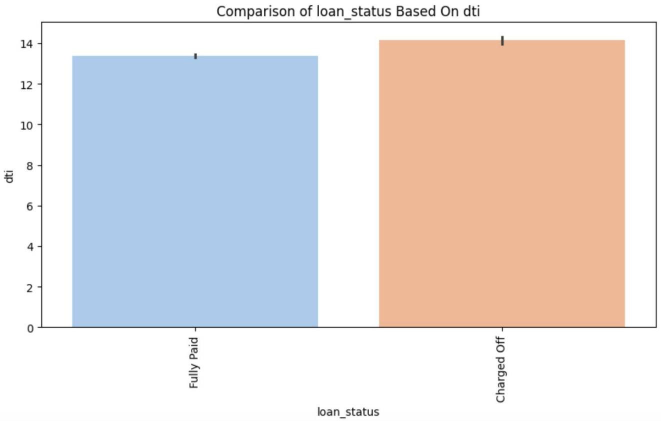



verification_status vs Loan amount
35000
30000
25000
20000
15000
10000
5000
0
loan_amnt
Verified
Source Verified
Not Verified
verification_status

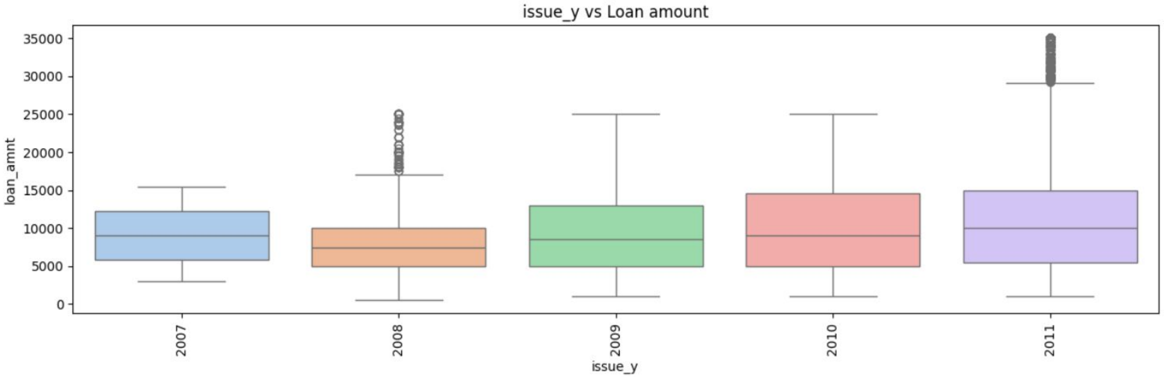

term vs Loan amount
35000
30000
25000
20000
15000
10000
5000
0
loan_amnt
36
60
term



Comparison of grade Based On dti



Comparison of loan_status Based On dti

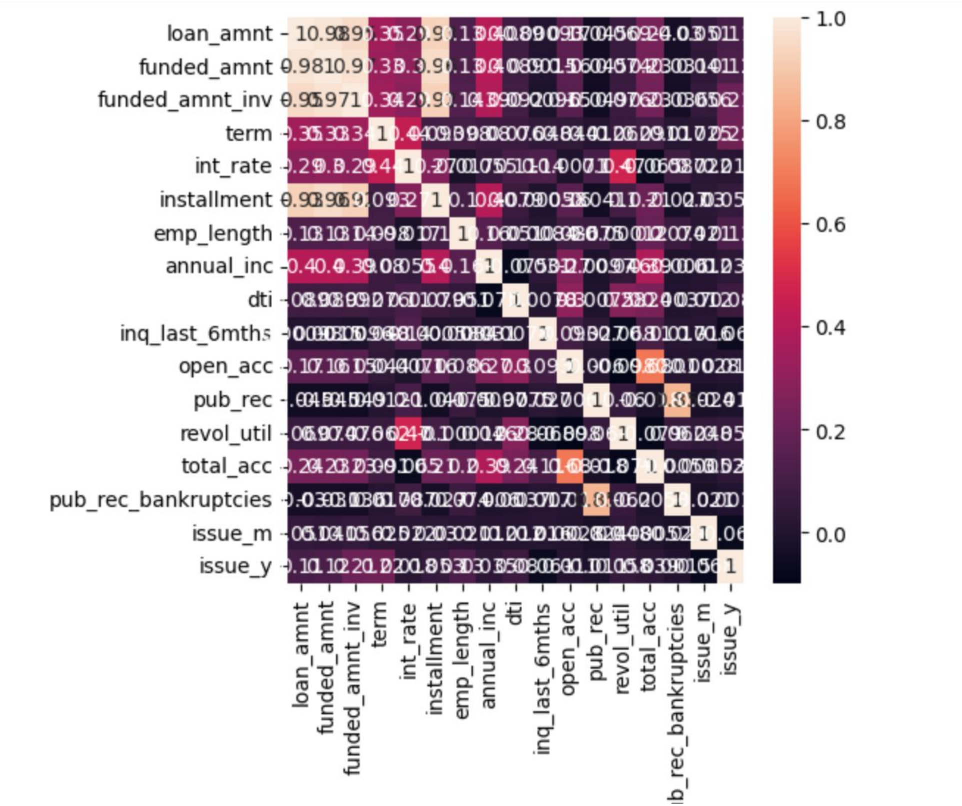


issue_y vs Loan amount

Observations and Inferences from Bivariate Analysis¶

1. Interest rate increases with the grade.
2. Higher the grade more is the risk of default.
3. In term vs interest rate variable, interest rate is less for 36 months tenure and higher for 60 months tenure.
4. The Grade A which is lowest risk also has lowest DTI ratio.Thus higher grade has lower rate of default.
5. Verified borrowers are having high dti ratio.
6. The borrowers are mostly having no record of Public Recorded Bankruptcy are safe choice for loan issue.
7. Higher the loan amount more is the charged off frequency.
8. Verified borrowers are having high dti ratio.
9. In Grade vs loan amount, Grade F & G have taken maximum amount of loan. As the grade decreases amount of loan is increasing.Higher the grade more is the risk of default

Correlation Analysis



Correlation Analysis

Strong Correlation

installment, funded_amnt, loan_amnt, and funded_amnt_inv has strong correlation with each other

Weak Correlation

dti has weak correlation with most of the fields

emp_length has weak correlation with most of the fields

Negative Correlation:

annual_inc and dti is negatively correlated.

pub_rec_bankrupticies has a negative correlation with almost every field

Recommendations:

Major Driving factor which can be used to predict the chance of defaulting and avoiding credit loss:

1. interest_rate
2. annual_income
3. Debt to income ratio (DTI)
4. Grade
5. Verification Status
6. Loan amount
7. Pub_rec_bankruptcies
8. Home ownership
9. Purpose
10. Emp Length
11. Term

Recommendations:

From exploratory data analysis we can conclude that, there is more probability of defaulting when:

- Borrowers who are taking loan for the ‘60 months’ tenure.
- Borrowers having Public Recorded Bankruptcy.
- Borrowers whose loan status is ‘Verified’ and they take high amount of loan with 60 months tenure.
- Borrowers who are having home ownership as ‘Rent’.
- Borrowers whose annual income is low i.e. (0-20000).
- Borrowers who takes loan amount in the range 0 to 14000.
- Borrowers who receive interest at the rate of 15-20%.
- Borrower who takes loan for the purpose of small business or debt consolidation.
- Borrower with least grades and sub_grades like E,F,G which indicates high risk.
- Borrower with very high Debt to Income value.
- Borrower with working experience 10+ years.
- Borrowers who does not belong to large urban cities like CA, NY etc.

Thank You!