Lending Club Case Study

By Alok Abhayajith

Objective of the case study

- The objective of the case study is to understand the pattern of loan defaulting by the borrowers and other business insights for improvement of loan disbursement by using Exploratory Data Analysis technique.
- Goals achieved through the case study:
  - Provides an idea on how EDA can help in solving real business problems.
  - A basic understanding of risk analytics in banking and financial services
  - Usage of techniques like data visualization and pivot tables for getting insights from real time business data.

Business Understanding

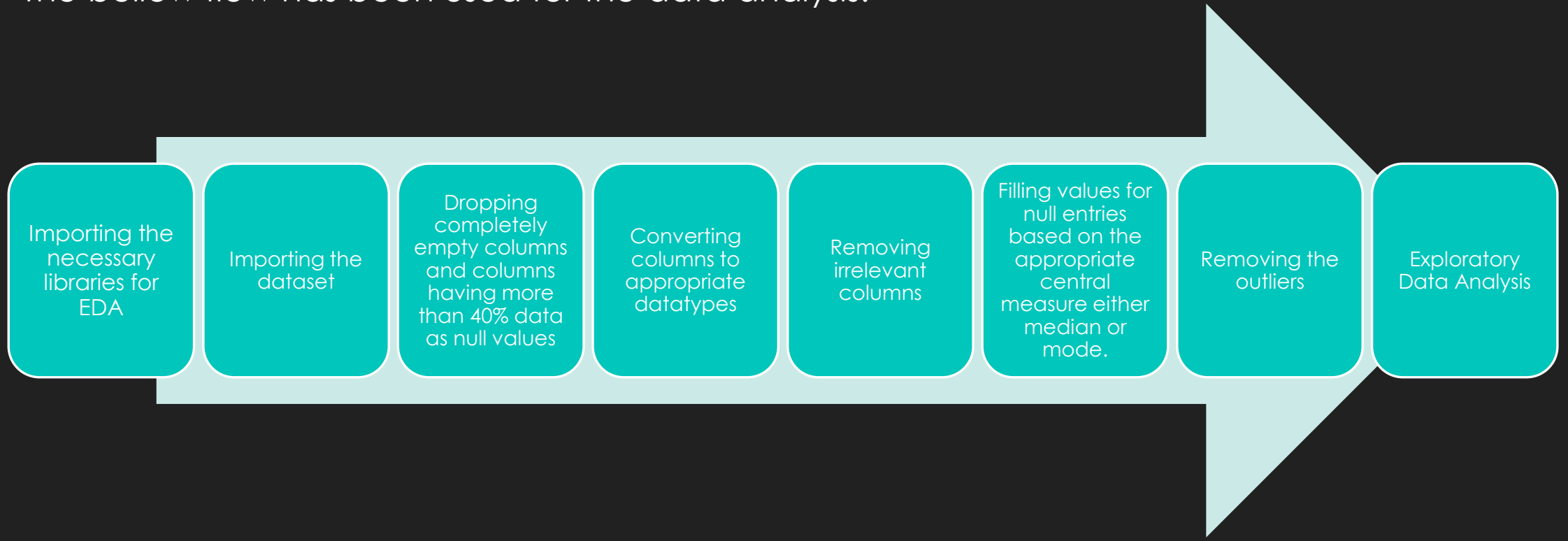
- The aim of the business is to improve the loan disbursement process by being able to identify the following:
  1. Customers who are likely to default
  2. Customers who will pay the loan amount completely
- Following are the important points which a lender uses for analysing a customer:
  1. **Annual Income:** Amount earned by the borrower in a year
  2. **Loan Amount:** Amount requested by the borrower as a loan
  3. **Credit History:** History of the borrower in terms of paying the loans back on time and other factors like amount of open credit lines in the borrower's credit file, any derogatory public records, etc.

Business Objective

- The business objective is to reduce the credit loss by not lending money to risky applicants. For which the following needs to be done:
  1. Driving factors to be identified for loan defaulting by using EDA
  2. Understanding the overall distribution of the loan amount across various categories like house ownership category, annual income, etc.

Process Flow for analysis

The bellow flow has been used for the data analysis:



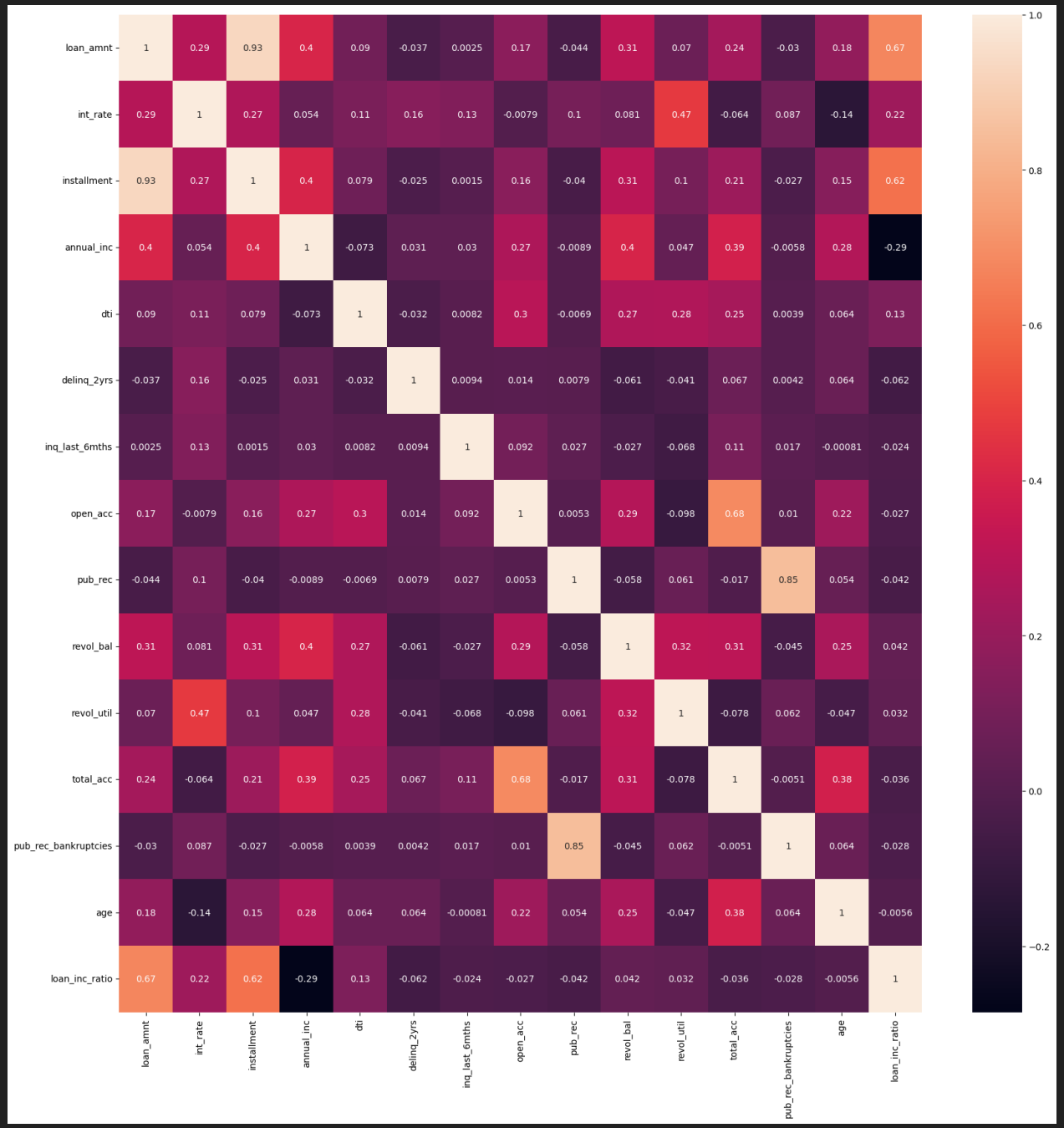
Exploratory Data Analysis

As a part of EDA the following has been done:

1. Correlation checked between columns
2. New metrics (Derived Metrics) have been defined for better understanding based on below 2 methods:
    1. Binning
    2. Derived using columns of the data
3. Univariate Analysis
4. Segmented Univariate Analysis
5. Bivariate Analysis
6. Recommendations based on EDA

Correlation between columns

- Correlation matrix was prepared as shown in the next slide to study the correlation between the variables.
- Following variables were found out to be highly correlated:
  1. loan_amnt & installment
  2. pub_rec_bankruptcies & pub_rec
  3. open_acc & total_acc



1. installment column is highly correlated with the loan_amnt column, hence removing installment column.

2. pub_rec_bankruptcies column is highly correlated with pub_rec, hence removing pub_rec_bankruptcies column.

3. Out of total_acc column and open_acc column, the more relevant column for analysis is open_acc colun as it is the number of open credit lines, which will be more benificial for the analysis.

Derived Metrics

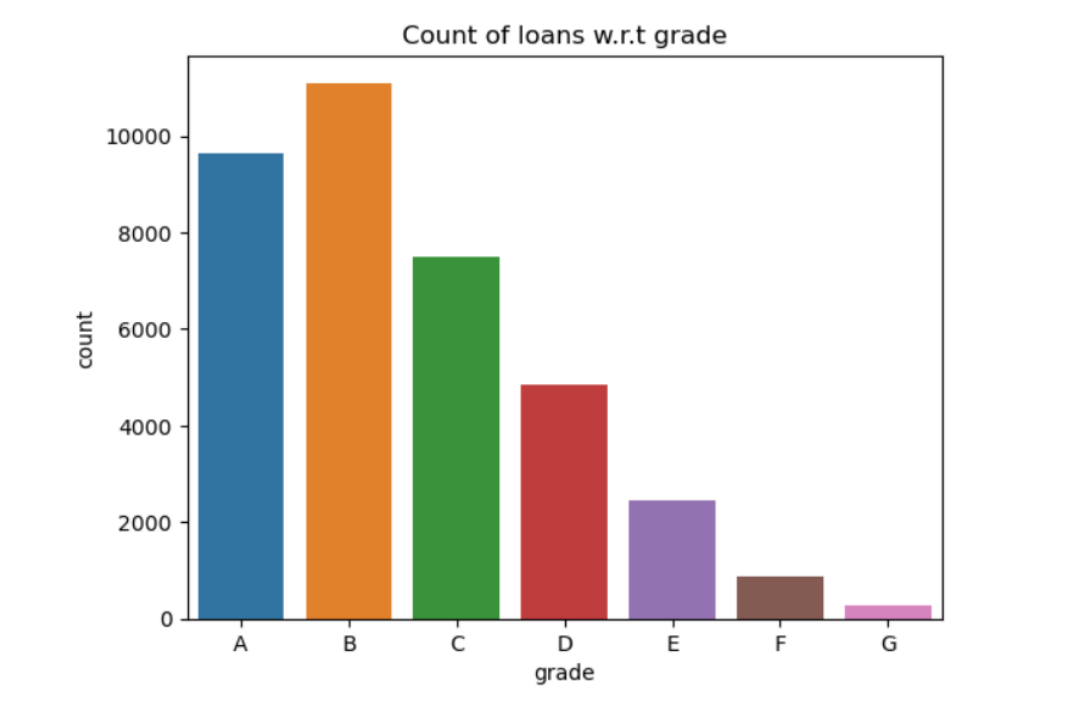
- New metrics created:
  1. age: Days from the issue of the first credit line to the loan issue date
  2. loan_inc_ratio: Loan amount to annual income ratio
- Following binned columns have been created:
  1. loan_amnt has been binned to form loan_amnt_binned column
  2. annual_inc has been binned to form annual_inc_binned column
  3. int_rate column has been binned to form int_rate_binned column
  4. age column has been binned to form age_binned column
  5. revol_bal column has been binned to form revol_bal_binned column
  6. revol_util column has been binned to form revol_util_binned column
  7. dti column has been binned to form dti_binned column
  8. delinq_2yrs column has been binned to form delinq_2yrs_binned column
  9. loan_inc_ratio has been binned to form loan_inc_ratio_binned column

Univariate Analysis

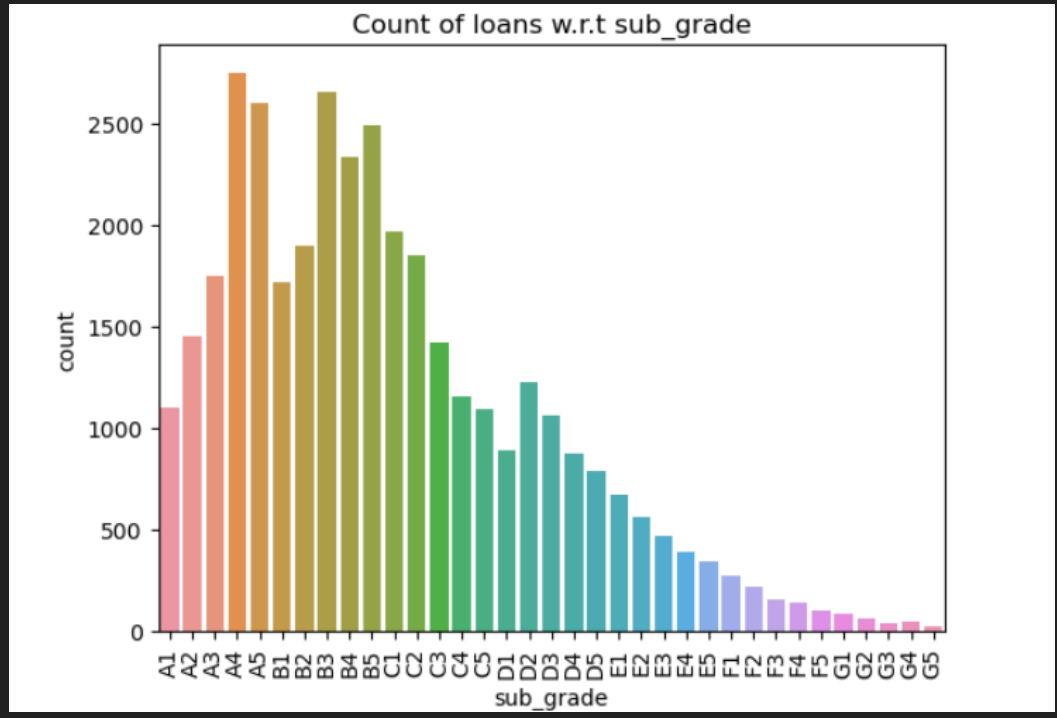
For univariate analysis the distribution of the loan along the following categories has been studied:

- Distribution of loans based on grade
- Distribution of loans based on sub grade
- Distribution of loans based on loan status
- Distribution of loans based on issue year
- Distribution of loans based on issue month
- Distribution of loans based on verification status
- Distribution of loans based on term of loans
- Distribution of loans based on purpose of loans

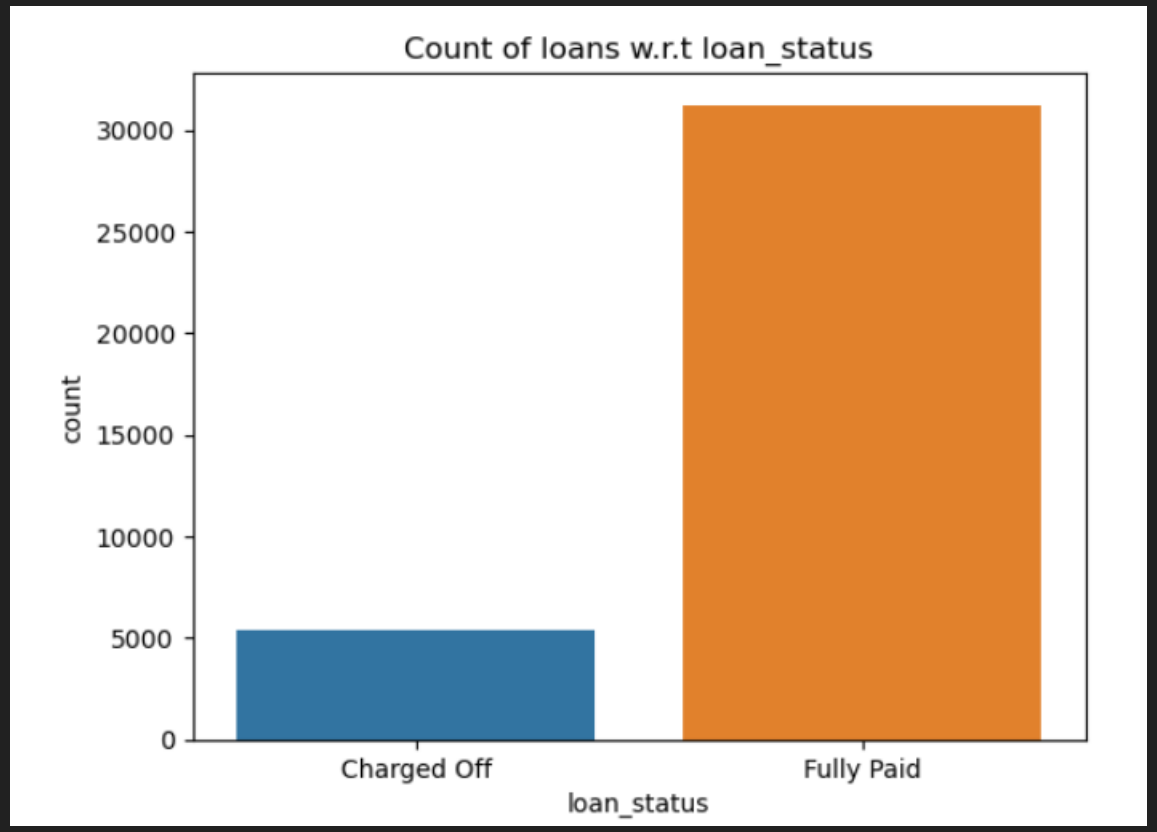
Results of Univariate Analysis

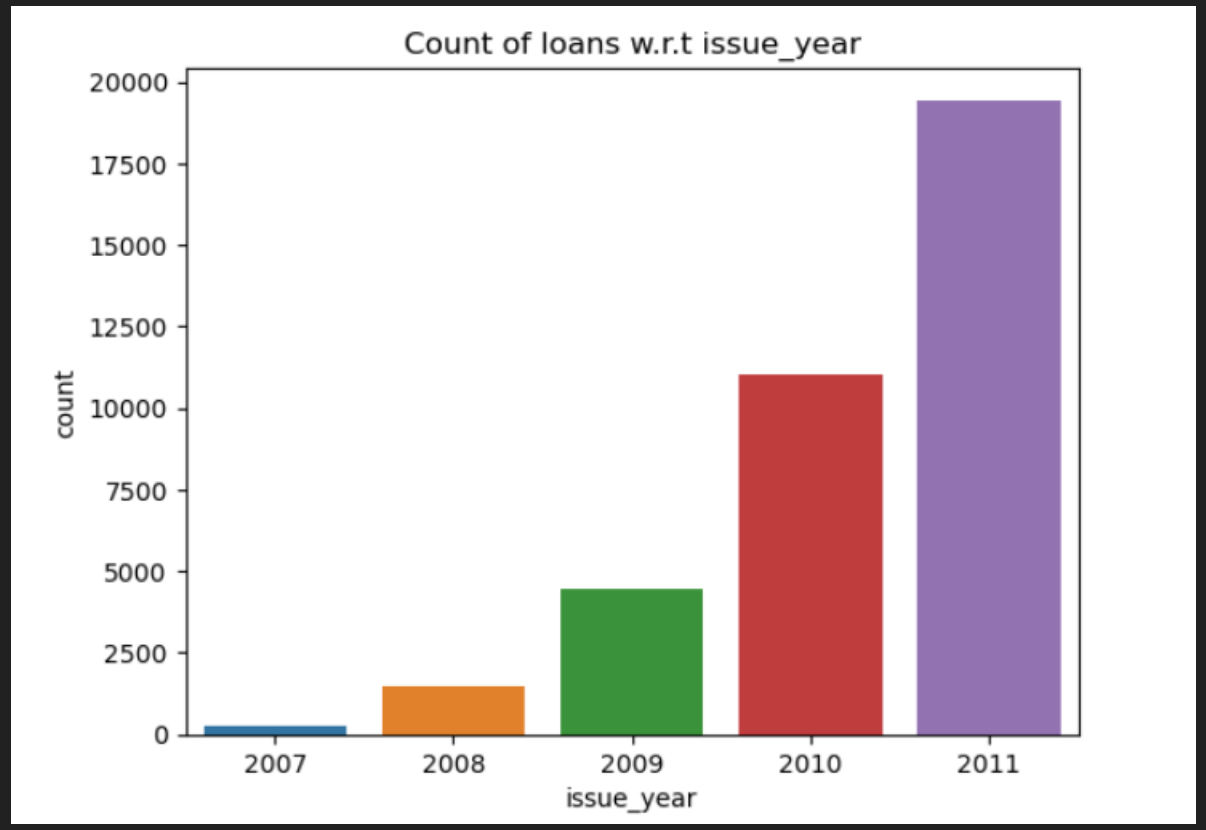


Highest number of loans have been disbursed against grade B while lowest number of loans have been disbursed against grade G.



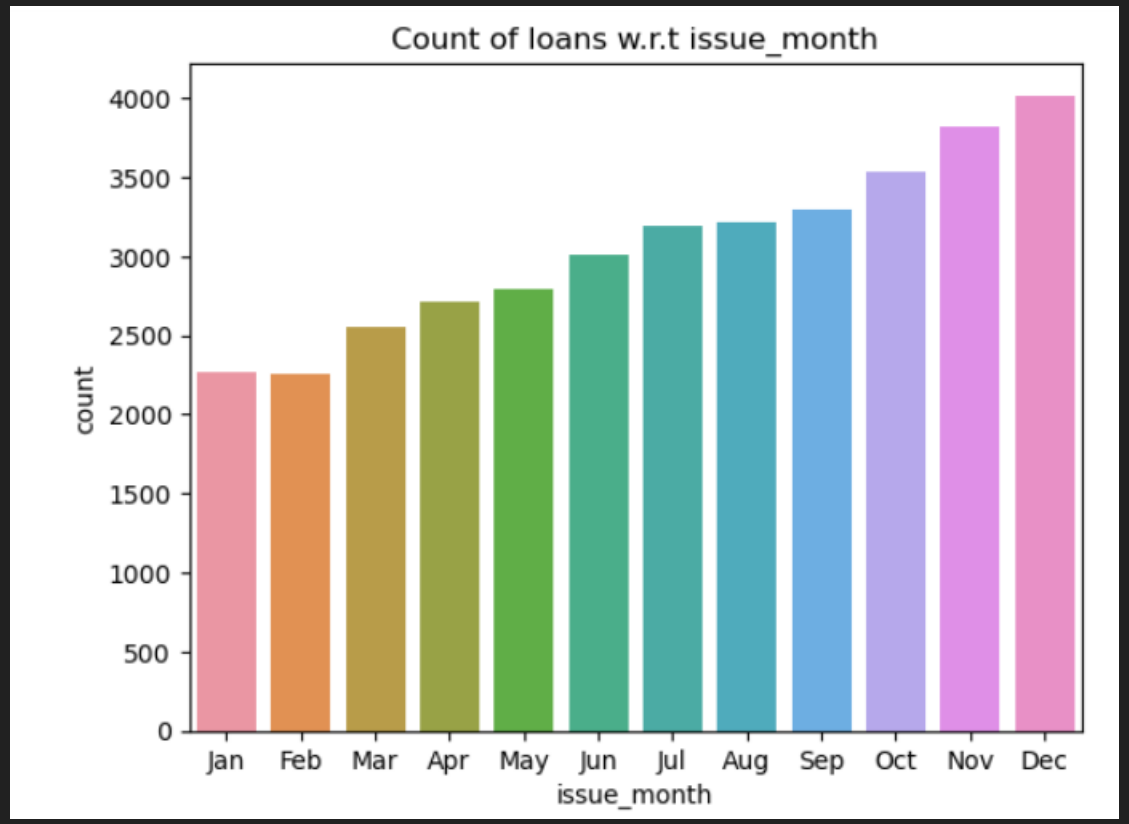
Highest number of loans have been disbursed against sub grade A4 while lowest number of loans have been disbursed against sub grade G5.

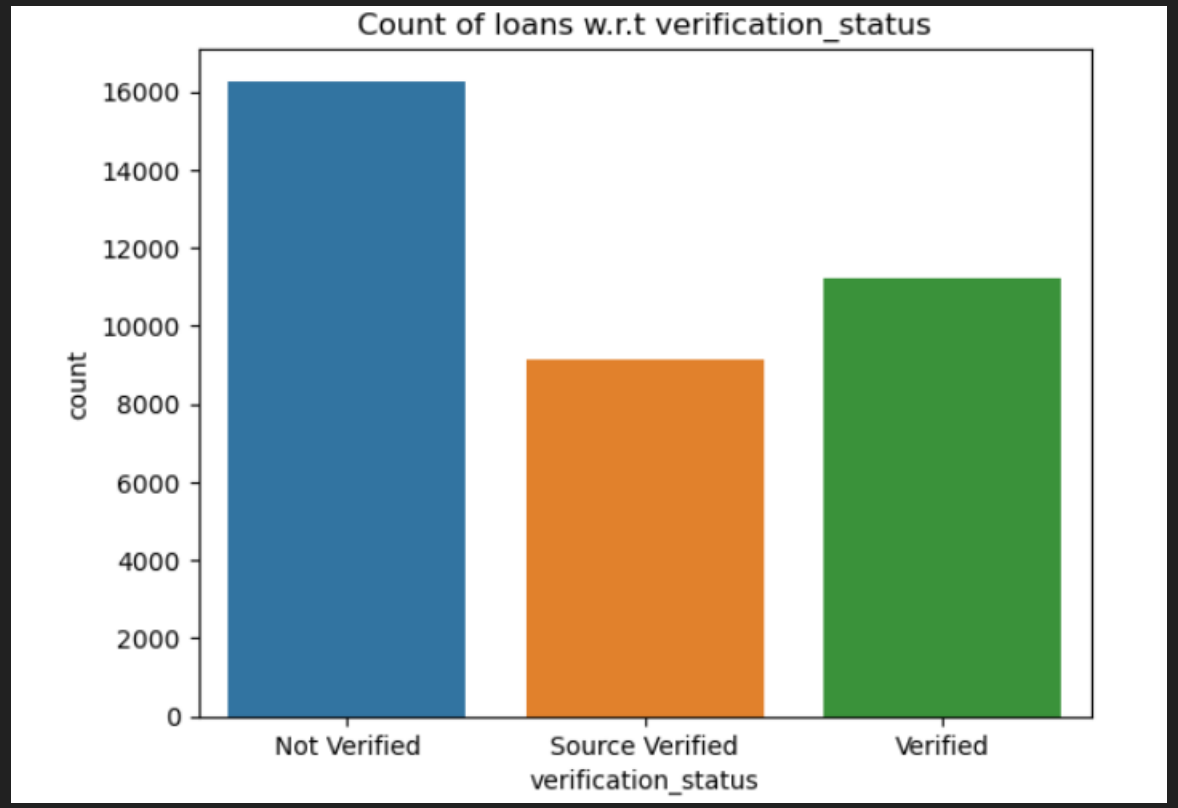




Count of fully paid loans is around 6 times the charged off loans.

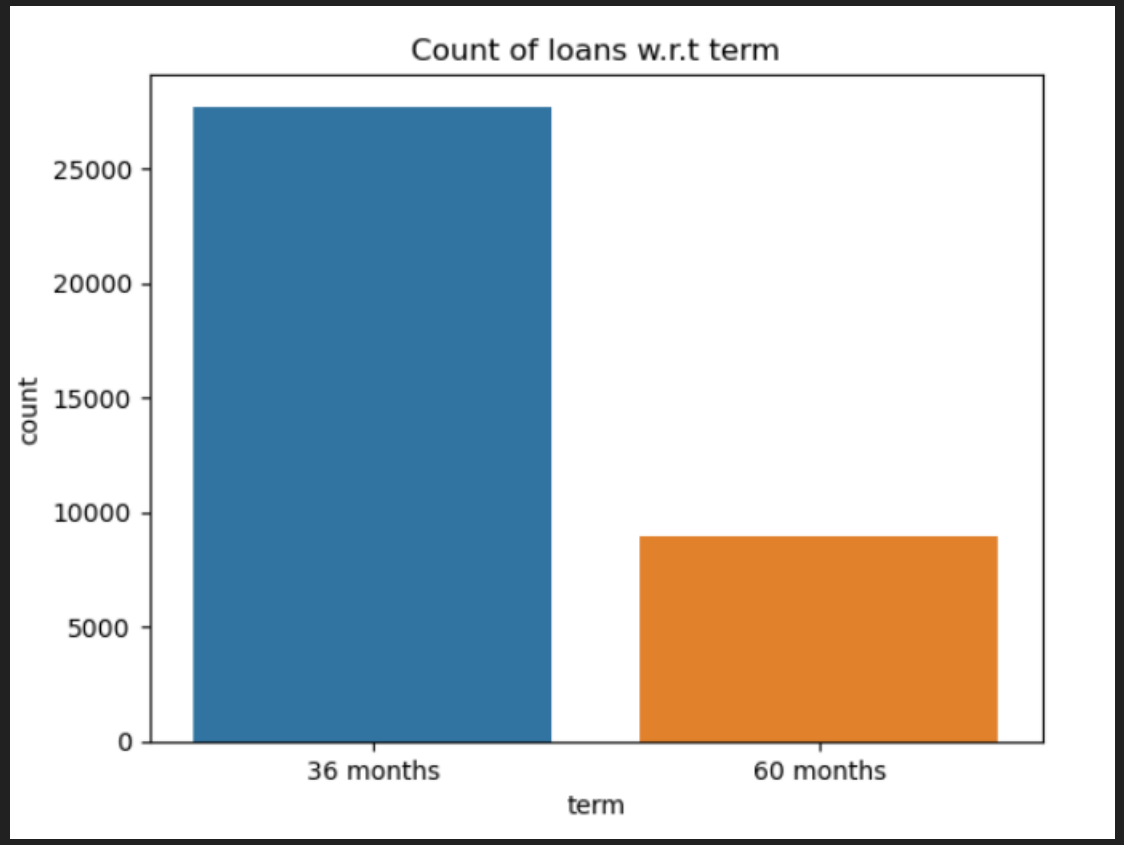
Count of loans disbursed have increased continuously over the years from 2007 to 2011.



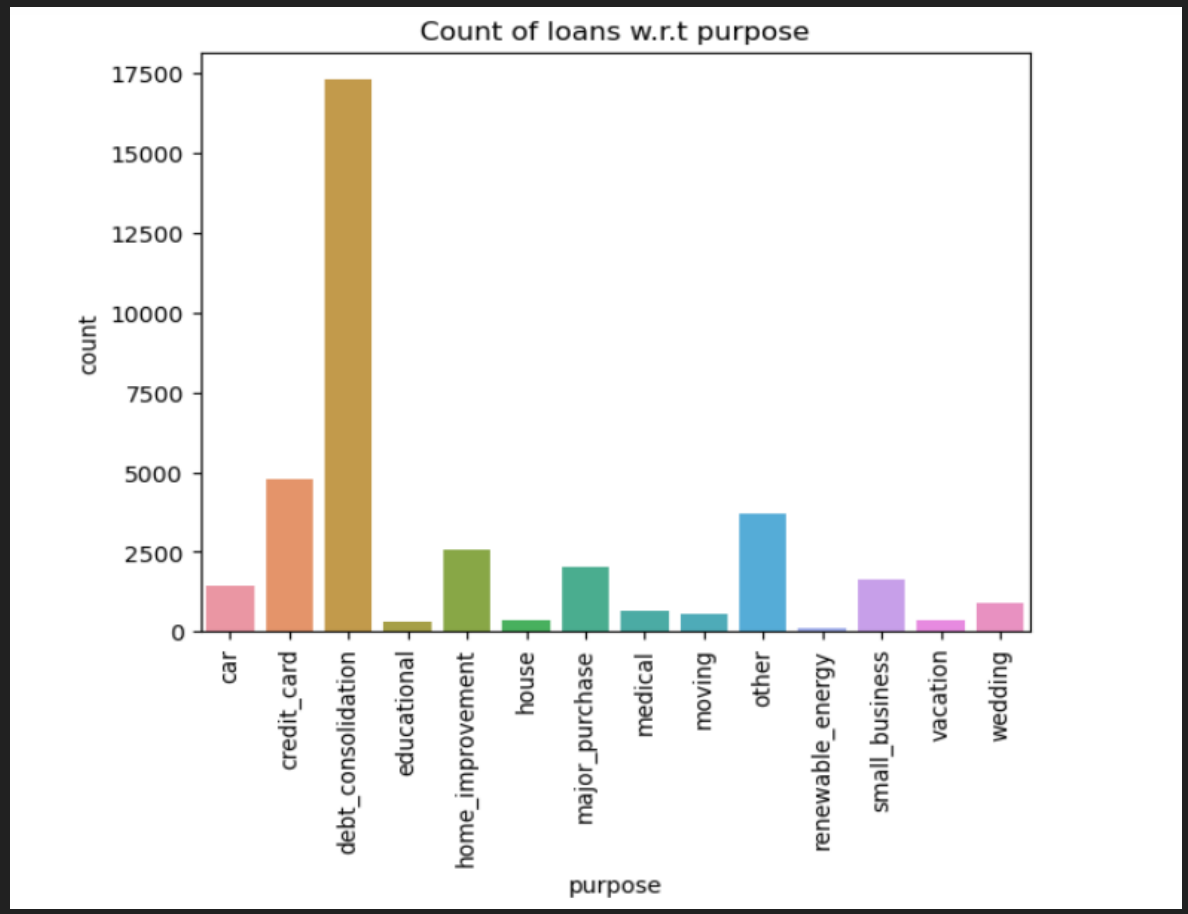


Maximum loans are disbursed in the month of December while minimum number of loans are disbursed in the month of February.

The annual income of the borrower is not verified for maximum loans.



Number of 36 months term loans are around 3 times higher than the number of 60 months term loan.

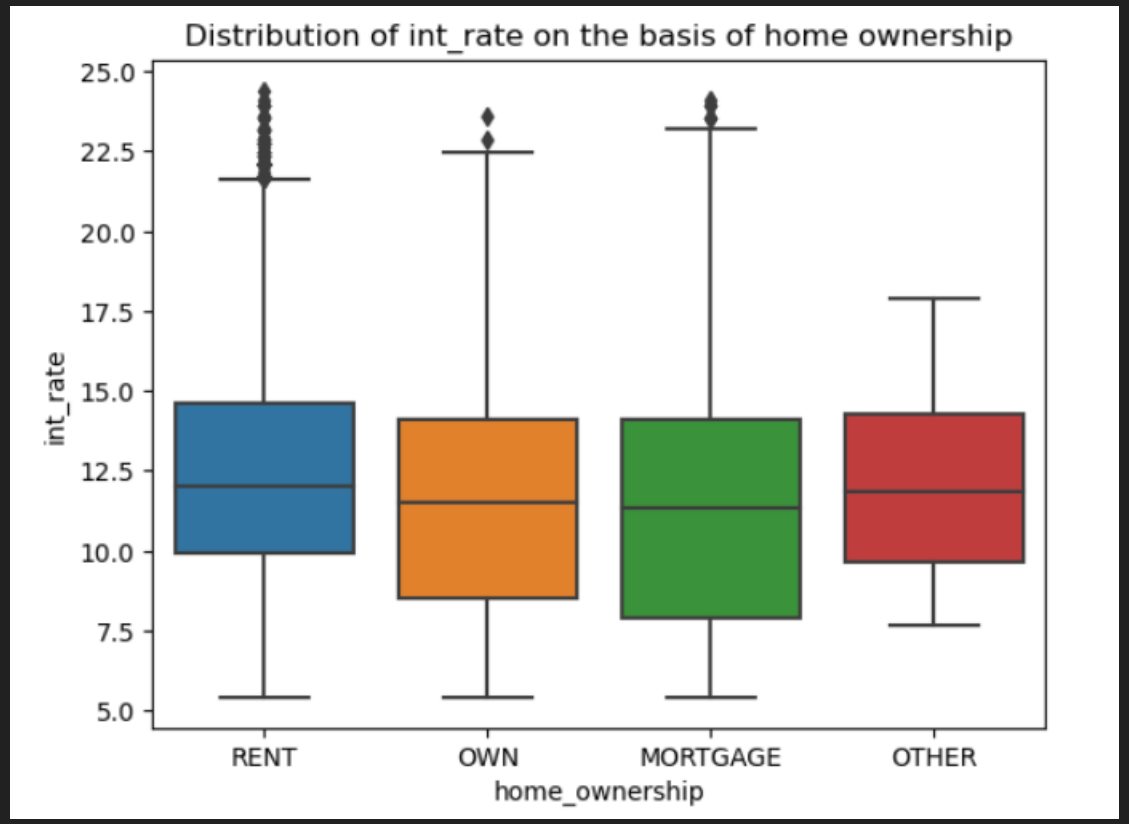


Maximum number of loans are disbursed for debt consolidation purpose and the minimum number of loans are disbursed for renewable energy purpose.

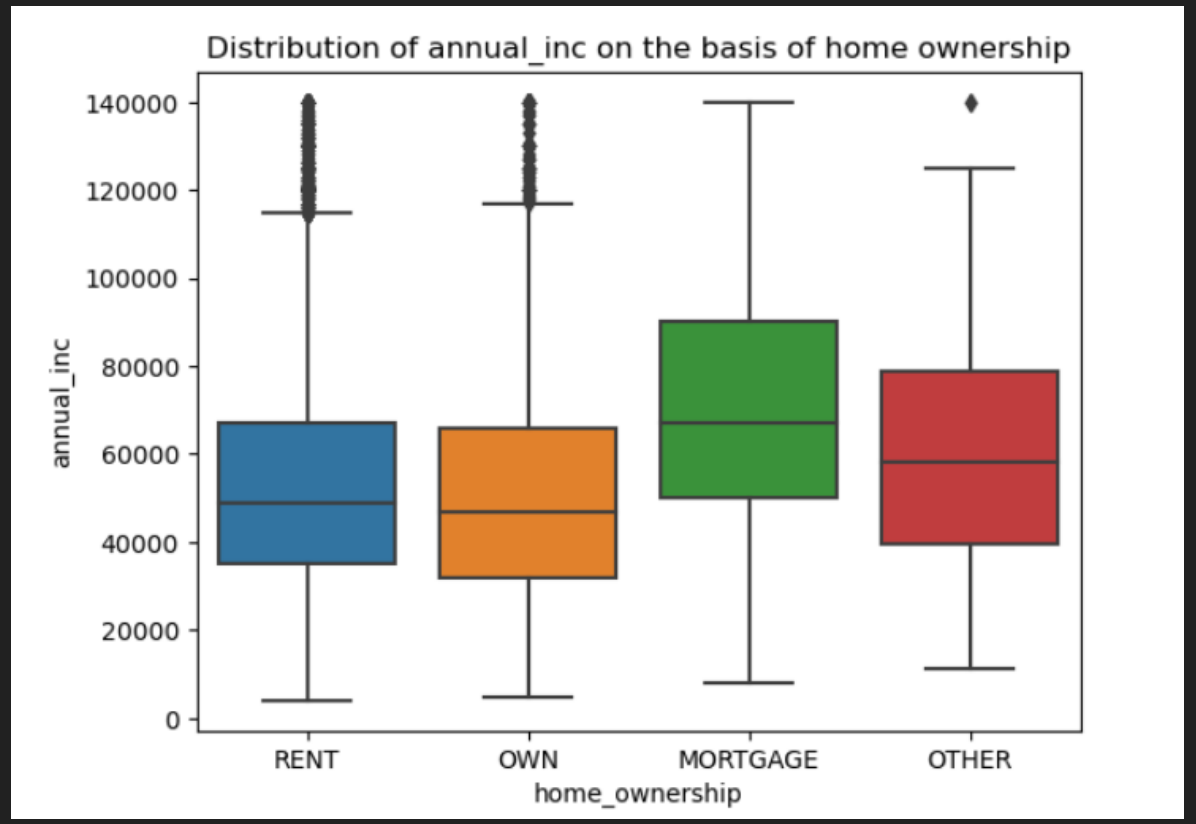
Segmented Univariate Analysis

For segmented univariate analysis the distribution of the loan along the following categories has been studied:

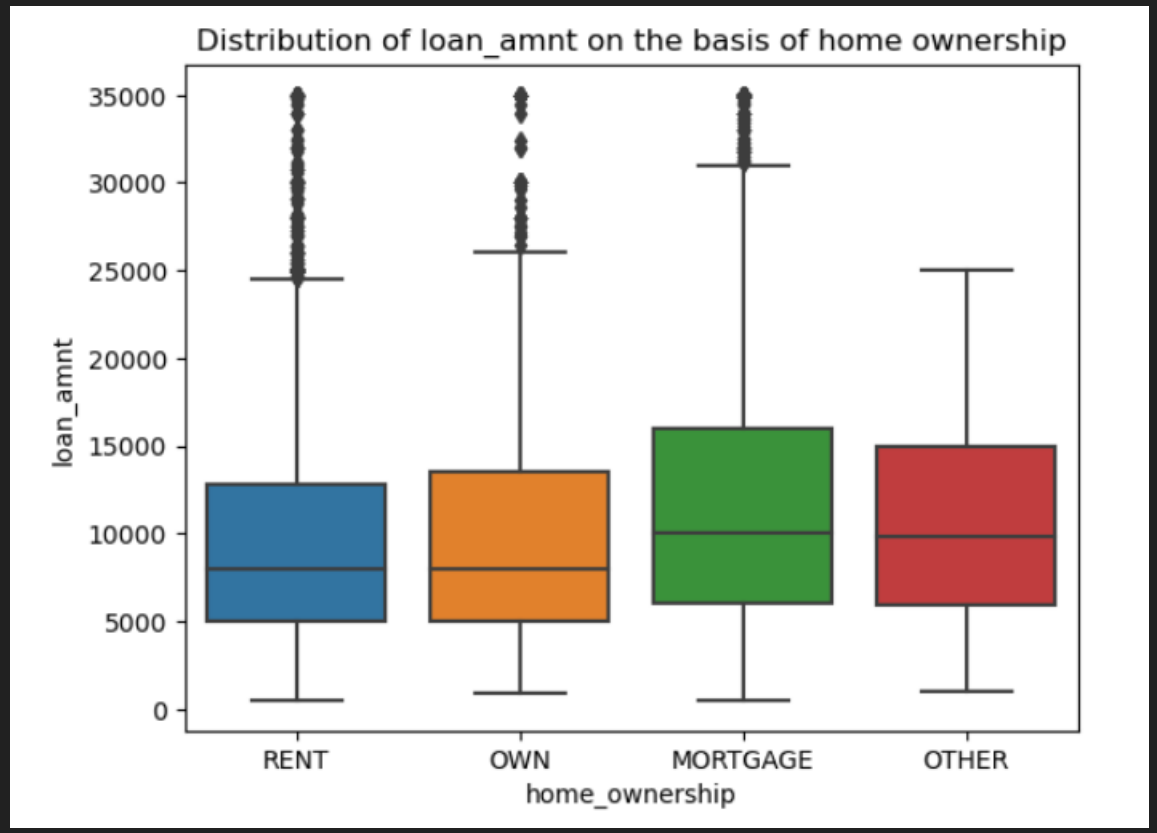
- Segmented univariate analysis of the basis of home ownership
    1. For interest rates
    2. For loan amount
    3. For annual income
- Segmented univariate analysis of the basis of issue month for interest rates
- Segmented univariate analysis of the basis of issue year for interest rates



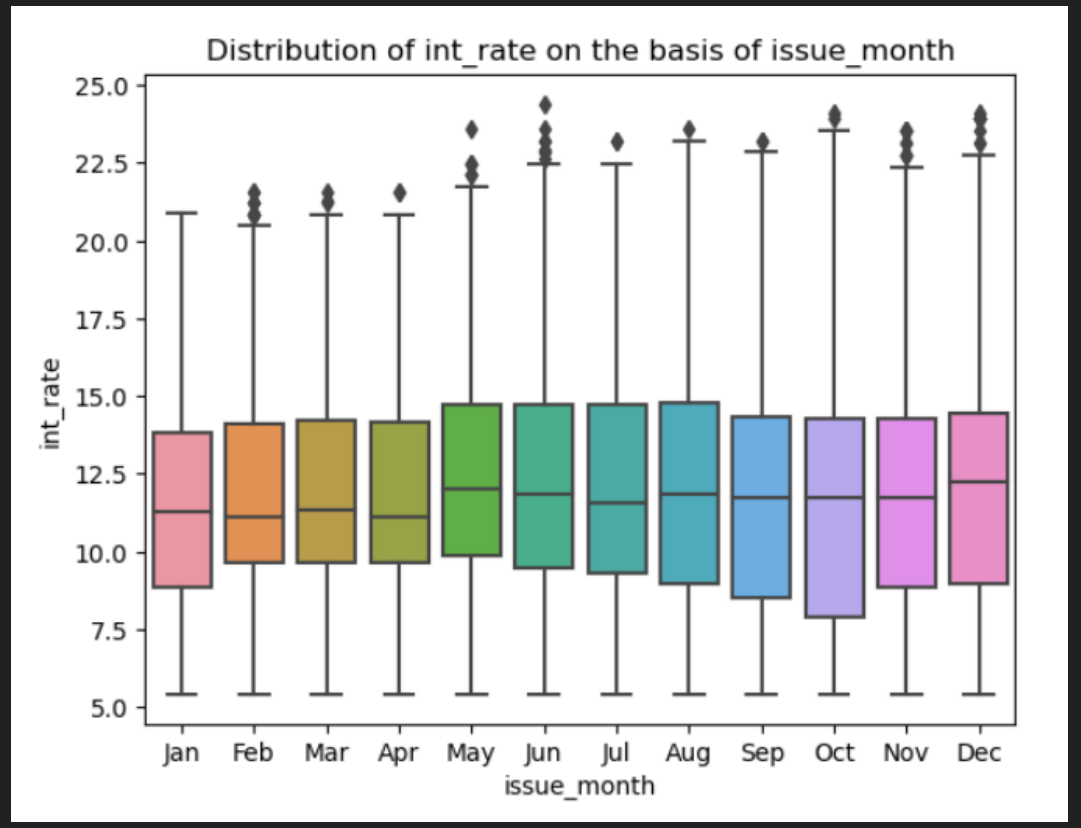
In general borrowers with own homes and mortgaged homes are charged with lower interest rates while those having rent arrangement or any other arrangement are charged higher ineterst rates, as they are more likely to default as they don't have any property to sell in order to pay off the debt.



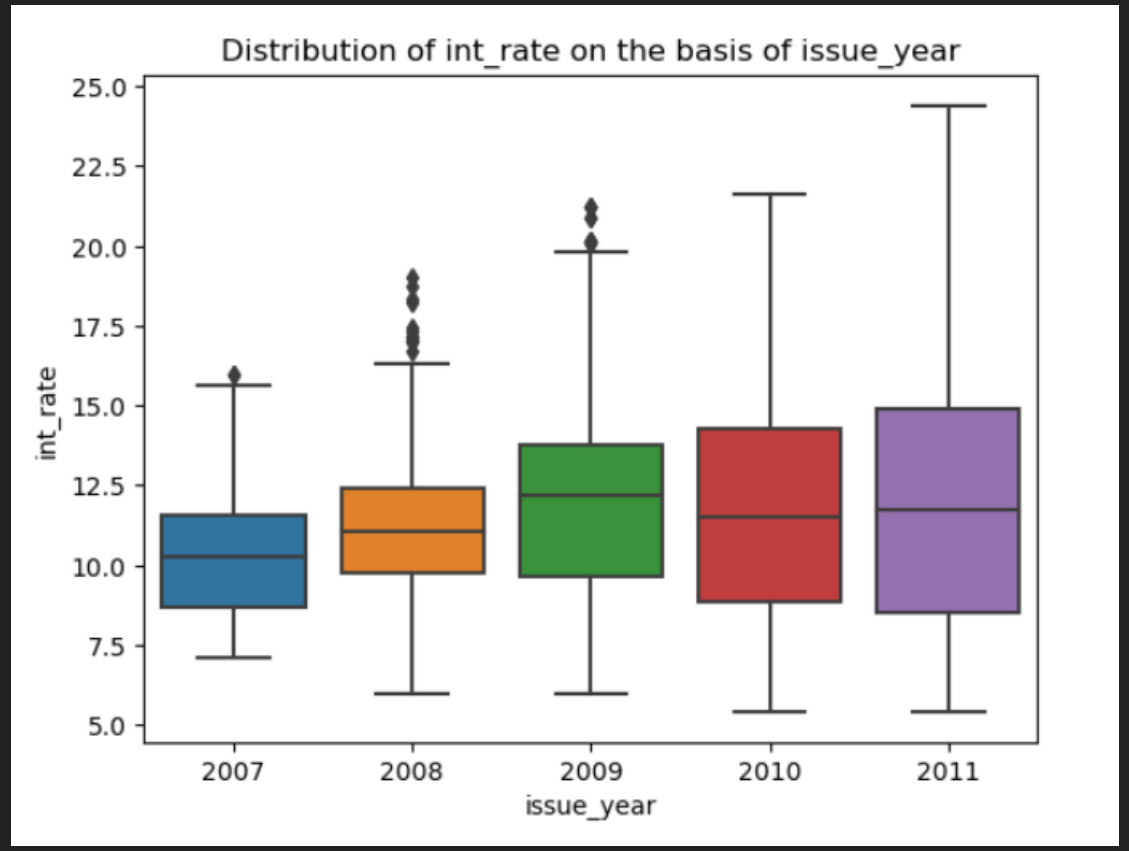
The people with mortgaged homes have higher median salary than any other home ownership category.



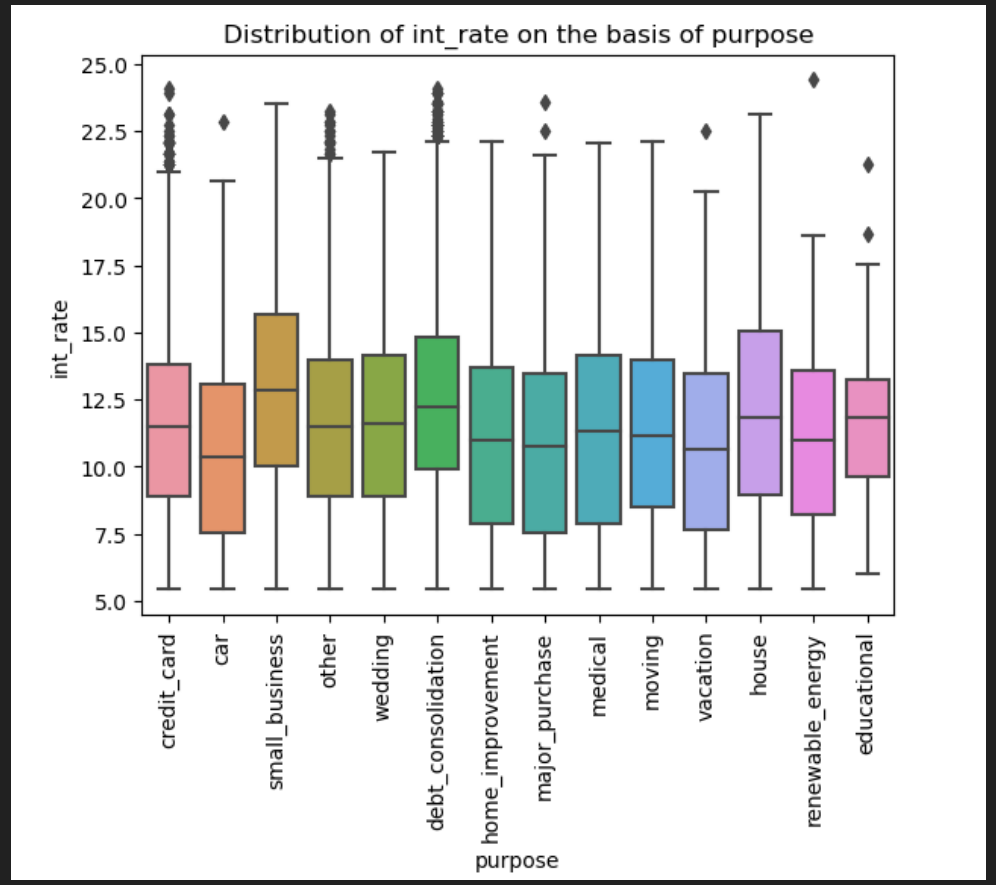
In general borrowers with mortgaged homes and other arrangement tend to take higher amounts as loans as they have higher salaries in comparison to people who own houses or rent houses.



The median interest rate for loans have almost similar distribution across the months, with an increase towards the year end i.e. towards December. This may be because large amount of loans are disbursed at the year end to meet the sales target, and thereby these loans are riskier and hence attract higher interest rates.



The median interest rates have increased throughout the years from 2007 to 2011, thereby indicating growth of the credit industry.

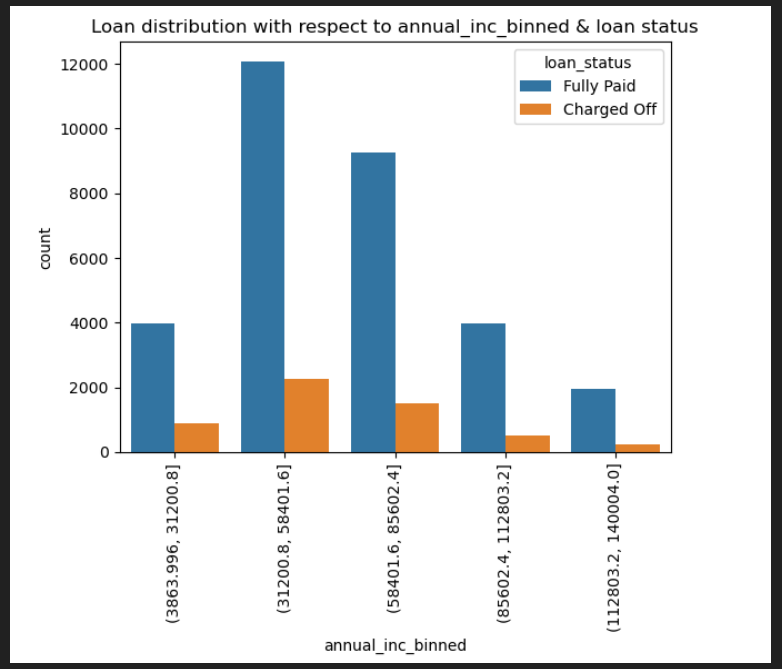


The highest median interest rate is for small businesses loans.

Bivariate Analysis

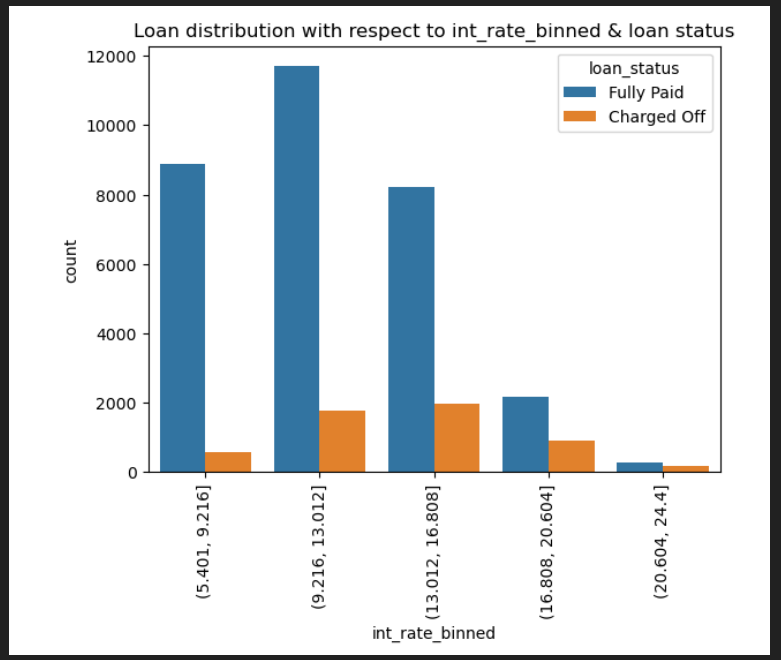
For bivariate analysis the distribution of the loan along the following have been studied:

- Annual income bands & loan status
- Interest rate bands & loan status
- Revolving line utilization rate & loan status
- Revolving balance & loan status
- Purpose & loan status
- Home ownership & loan status
- Verification status & loan status
- Term & loan status
- Public derogatory records & loan status
- Delinquency & loan status
- dti & loan status
- Loan to income ratio and loan status



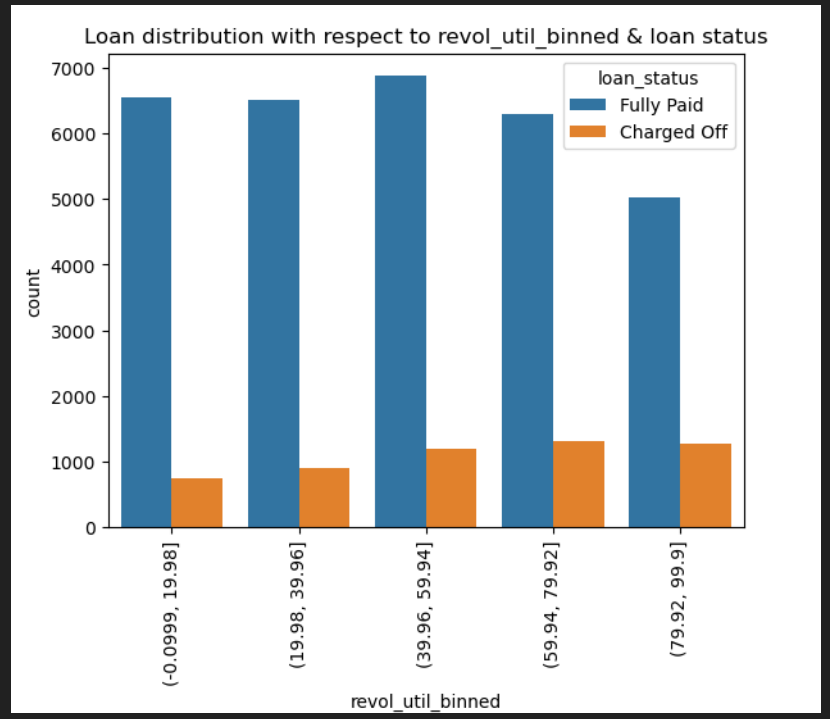
loan_status	Charged Off	Fully Paid	Charge Off %
annual_inc_binned			
(3863.996, 31200.8]	901	3972	18.49
(31200.8, 58401.6]	2276	12080	15.85
(58401.6, 85602.4]	1494	9277	13.87
(85602.4, 112803.2]	502	3962	11.25
(112803.2, 140004.0]	245	1945	11.19

The highest number of loans are disbursed for the borrowers in the income category of 31200 USD to 58402 USD. The borrowers in the income group of 3864 USD to 31200 USD have a higher chance of defaulting.



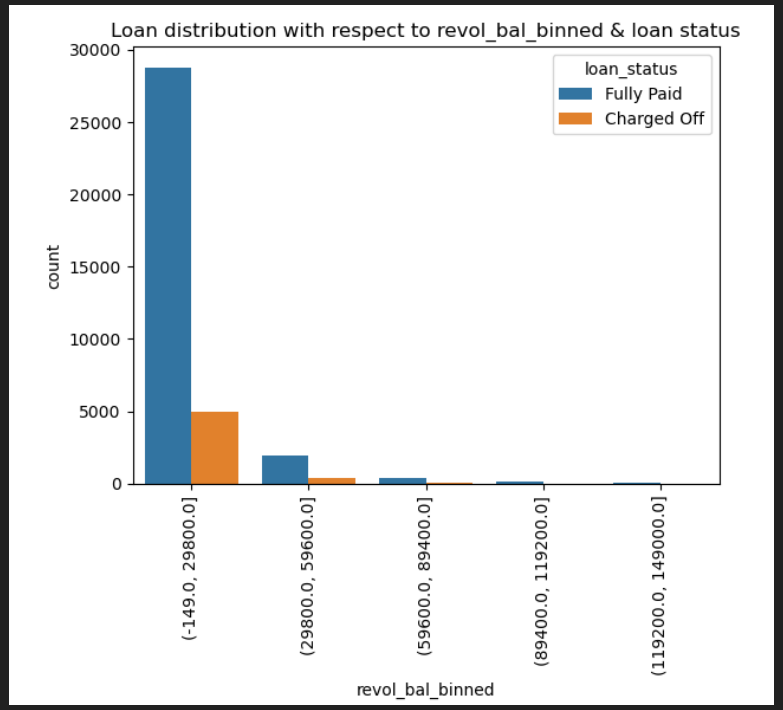
loan_status	Charged Off	Fully Paid	Charge Off %
int_rate_binned			
(5.401, 9.216]	578	8869	6.12
(9.216, 13.012]	1776	11694	13.18
(13.012, 16.808]	1956	8228	19.21
(16.808, 20.604]	917	2178	29.63
(20.604, 24.4]	191	267	41.70

The highest number of loans are disbursed for the interest rate band of 9.2% to 13.1%. For interest rates in the band of 20.6% to 24.4% the default percentage is the highest (41.7%).



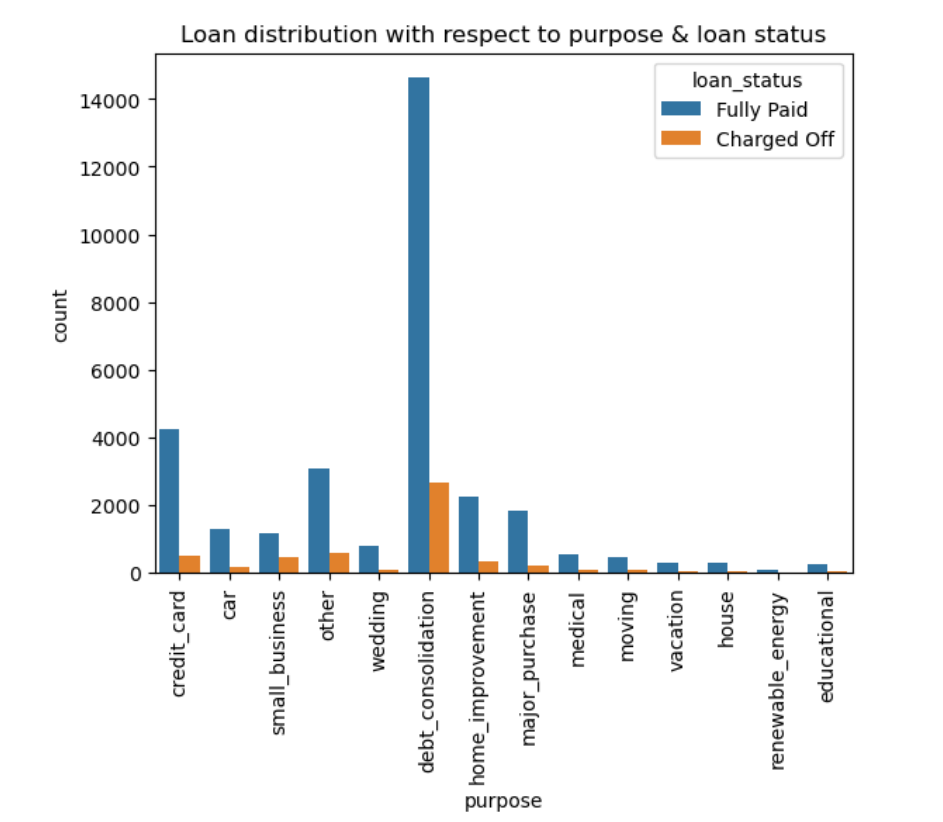
loan_status	Charged Off	Fully Paid	Charge Off %
revol_util_binned			
(-0.0999, 19.98]	741	6546	10.17
(19.98, 39.96]	897	6504	12.12
(39.96, 59.94]	1196	6872	14.82
(59.94, 79.92]	1308	6285	17.23
(79.92, 99.9]	1276	5029	20.24

The borrowers who have revolving utilization in the band of 79.92% to 99.9% are likely to default more than any other bands. It can be concluded that the higher the revolving utiliztion rate it is more likely that the borrower will default.



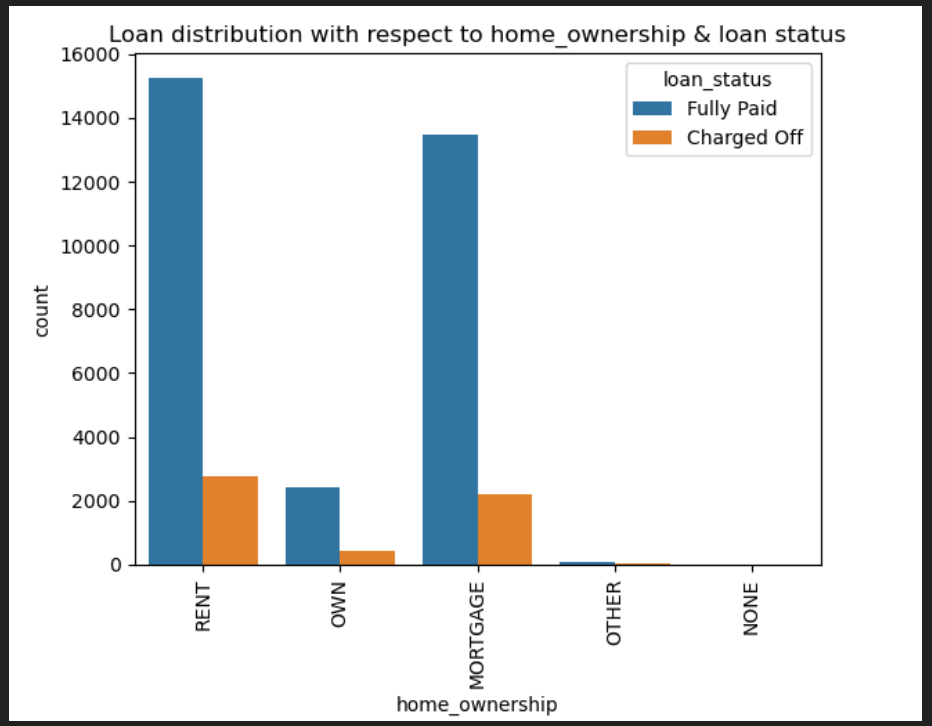
loan_status	Charged Off	Fully Paid	Charge Off %
revol_bal_binned			
(-149.0, 29800.0]	4957	28766	14.70
(29800.0, 59600.0]	387	1969	16.43
(59600.0, 89400.0]	50	362	12.14
(89400.0, 119200.0]	21	111	15.91
(119200.0, 149000.0]	3	28	9.68

In general, higher the revolving balance the lower the likelihood of defaulting.



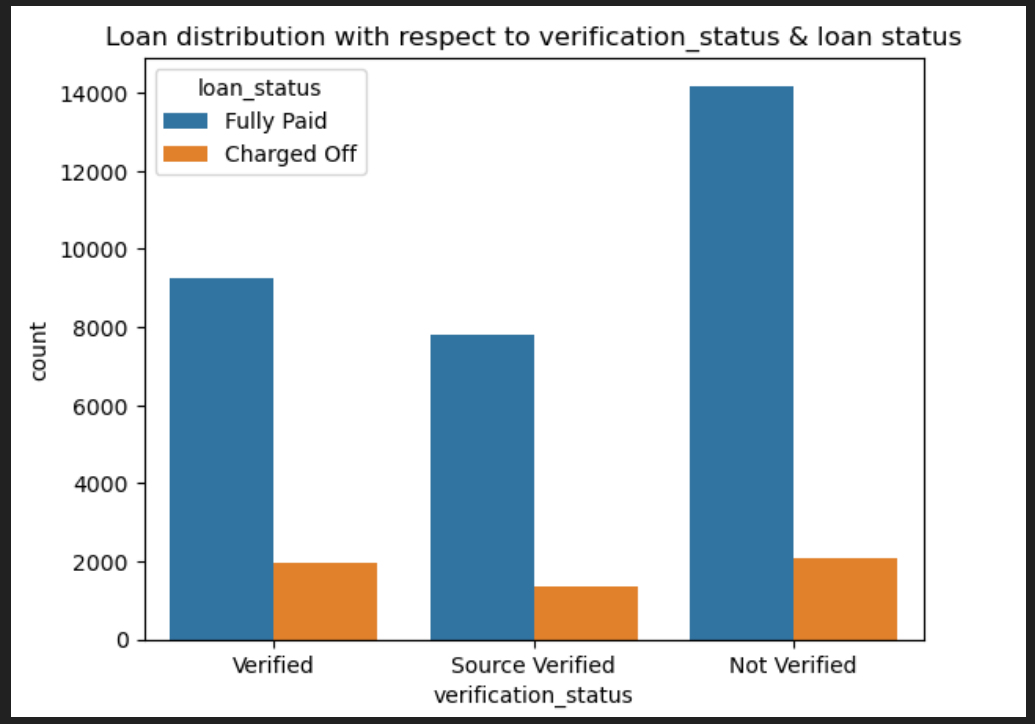
loan_status purpose	Charged Off	Fully Paid	Charge Off %
car	159	1289	10.98
credit_card	519	4264	10.85
debt_consolidation	2677	14628	15.47
educational	54	258	17.31
home_improvement	322	2234	12.60
house	55	283	16.27
major_purchase	217	1829	10.61
medical	104	537	16.22
moving	90	460	16.36
other	605	3094	16.36
renewable_energy	18	77	18.95
small_business	454	1169	27.97
vacation	52	313	14.25
wedding	92	801	10.30

Loans pertaining to small businesses will default more in comparison to loans for any other purpose.



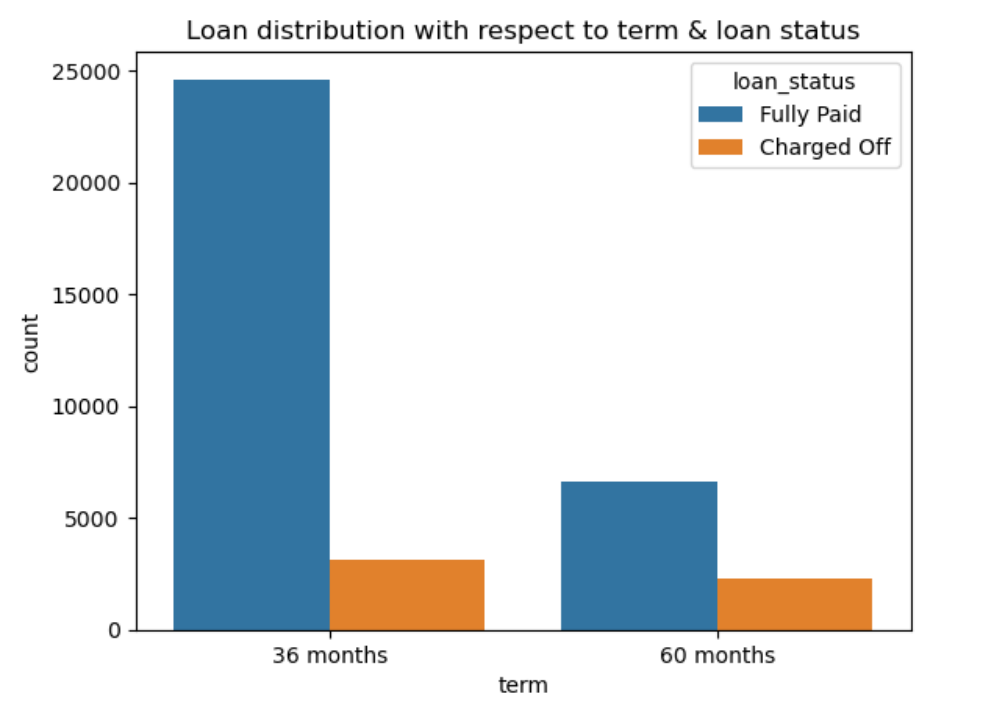
loan_status	Charged Off	Fully Paid	Charge Off %
home_ownership			
MORTGAGE	2194.00	13481.00	14.00
NONE	NaN	3.00	NaN
OTHER	17.00	76.00	18.28
OWN	434.00	2418.00	15.22
RENT	2773.00	15258.00	15.38

People having other housing arrangement are the more likely to default while the people having mortgages houses are less likely to default as they have already mortgaged their houses and hence there will be a pressure on them for repaying the loans.

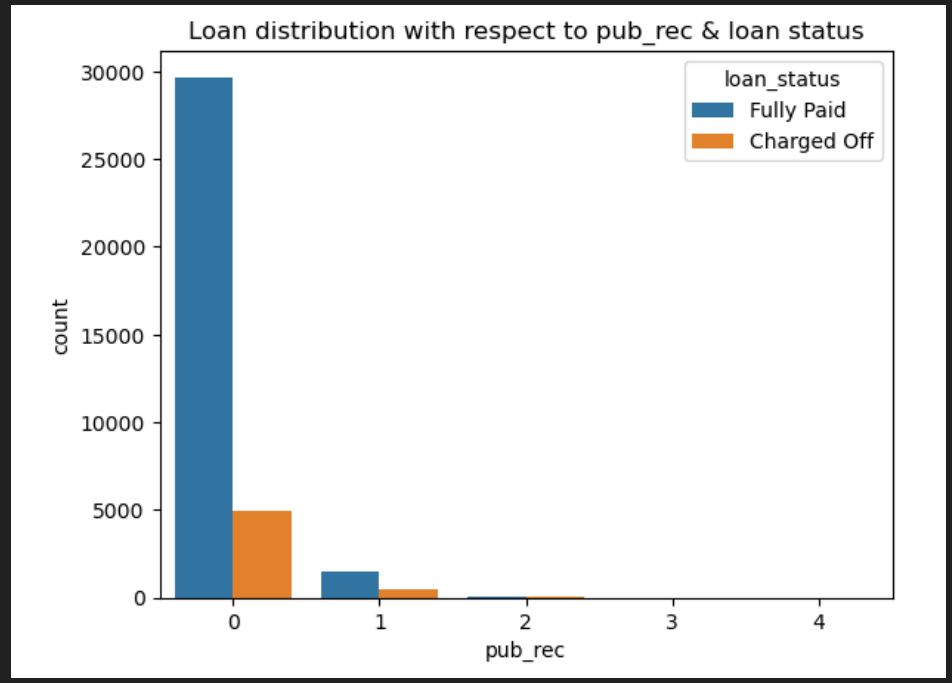


loan_status	Charged Off	Fully Paid	Charge Off %
verification_status			
Not Verified	2096	14184	12.87
Source Verified	1373	7782	15.00
Verified	1949	9270	17.37

Verified loans are more likely to be defaulted as the loans provided to verified income source borrowers are in the higher bands of loan amounts.



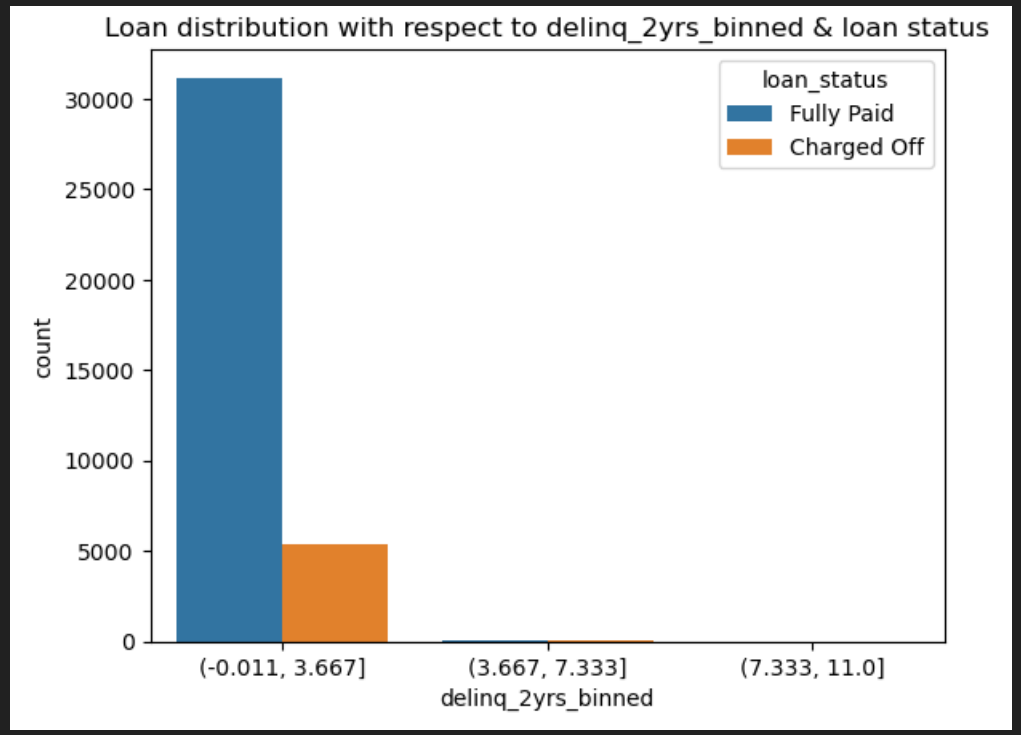
loan_status	Charged Off	Fully Paid	Charge Off %
term			
36 months	3114	24607	11.23
60 months	2304	6629	25.79



loan_status	Charged Off	Fully Paid	Charge Off %
pub_rec			
0	4959.00	29676.00	14.32
1	449.00	1516.00	22.85
2	10.00	36.00	21.74
3	NaN	6.00	NaN
4	NaN	2.00	NaN

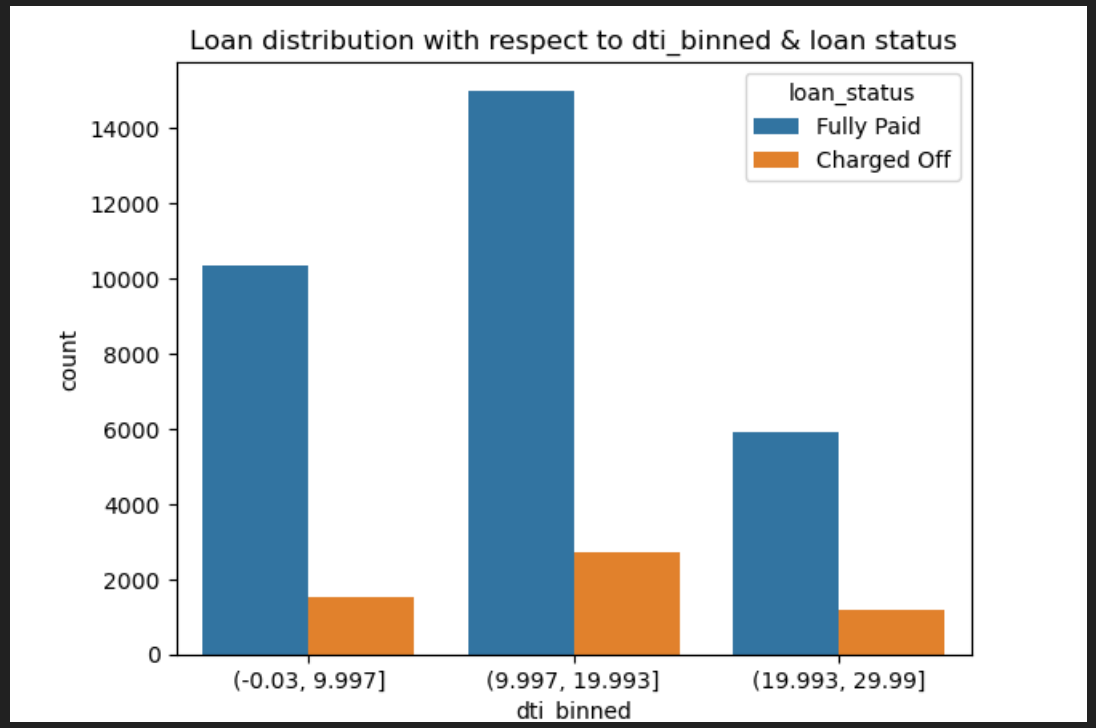
Loans with a 60 month term are more likely to default than the loans with a 36 month term as the majority of the 60 month term loans are in an interest rate band of 13%-24%, while majority of the 36 month loans are in the interest rate band of 5%-13%.

Borrowers having derogatory records are more likely to default.



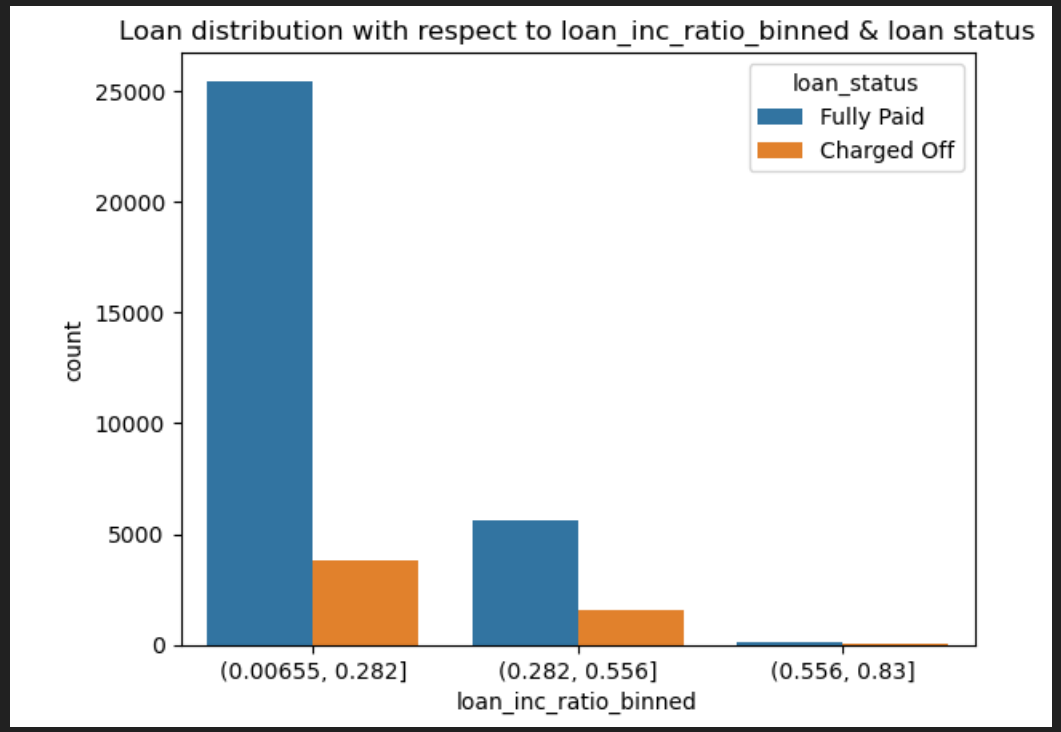
loan_status	Charged Off	Fully Paid	Charge Off %
delinq_2yrs_binned			
(-0.011, 3.667]	5400	31160	14.77
(3.667, 7.333]	17	73	18.89
(7.333, 11.0]	1	3	25.00

Borrowers having higher number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years are more likely to default.



loan_status	Charged Off	Fully Paid	Charge Off %
dti_binned			
(-0.03, 9.997]	1529	10331	12.89
(9.997, 19.993]	2696	15004	15.23
(19.993, 29.99]	1193	5901	16.82

Borrowers having higher dti ratio are more likely to default.



loan_inc_ratio_binned	loan_status Charged Off	Fully Paid	Charge Off %
(0.00655, 0.282]	3796	25484	12.96
(0.282, 0.556]	1558	5610	21.74
(0.556, 0.83]	64	142	31.07

Loans with loan amount to income ratio lesser than 0.282 are less likely to default and loans having higher loan amount to income ratio are more likely to default.

Recommendations

- Borrowers with lower income have a tendency to default and hence either the inetrest rate should be higher for them as they are more likely to default or the loan amount must be lower.
- In general, higher the revolving balance the less likely the borrower will default and hence the loans can be disbursed to such borrowers.
- Small business loans to be provided to borrowers having an income range above 31000 USD as they are more likely to default when provided to borrowers having an income range lower than that.
- Borrowers having higher loan to income ratio are more likely to default in general and hence the interest rate should be higher for such loans and should be provided only when the borrowers have lower revolving utilization rate and higher revolving balance.

Thank You