EDA-UPGRAD

Lending club case study

Problem Statement

- To Identify the driving factors or variables that lead to loan defaults(loan will not be paid by the customer).
- To analyze the factors that are available at the time of loan application and identify whether a loan with default or not

Approach used

- Approach used is EDA Method
    1) Data Understanding
    2) Data Cleaning
    3) Data Analysis
        a) Univariate Analysis
        b) Bivariate Analysis
        c) Segmented Univariate Analysis

Data Understanding

- Dataset provided is the details od applicants at the time of loan application. Either the loan will be fully paid, in-progress or defaulted by the customer. The dataset provided is private datatype.
- It has total 39717 entries with 111 columns
- Some of the priority columns listed out are:-loan amount, interest rate, annual income, purpose and employment etc.
- Loan status is the target column upon which impact of other factors will be analyzed.

Data Cleaning

- Removing all Null Items
- Removing the columns that will not be helpful in analysis
- Removing the columns that will be known after loan application only like
- Converting the datatypes to standard format

Data Analysis

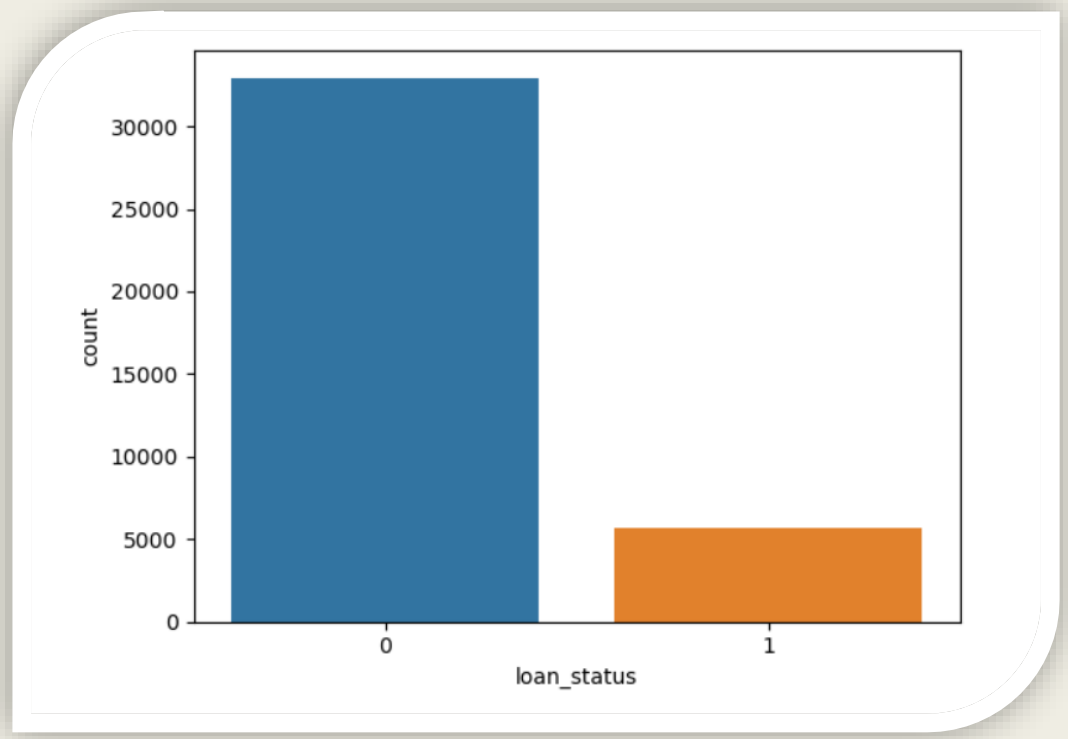
- After Cleaning the data, we need to understand the columns that are required for our Data Analysis.
- We can remove the columns that are not required, or which are not useful for our Data Analysis.
- The main objective of Data Analysis is to analyze whether a person will repay the loan or not at the time of Loan application.
- The column "pymt_plan" is not going to have any impact on the loan status which can be removed.
- Similarly, the column "url" is not going to have any impact on the loan status which can be removed.
- The Columns:- "delinq_2yrs", "earliest_cr_line","inq_last_6mths","open_acc", "pub_rec", "revol_bal", "revol_util", "total_acc",
- "out_prncp", "out_prncp_inv", "total_pymnt", "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "total_rec_late_fee", "recoveries", "collection_recovery_fee", "last_pymnt_d", "last_pymnt_amnt", "last_credit_pull_d", "application_type

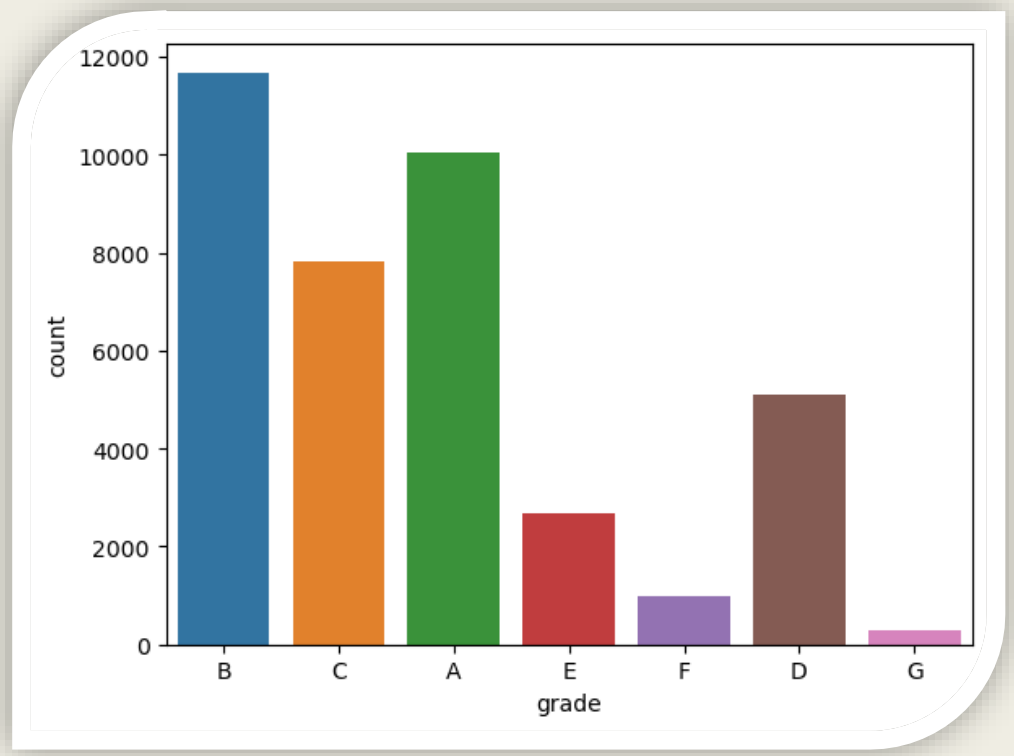
Data Analysis

- Classification of Data columns into Categorical and Numeric
- Classification of Data columns into ordered and unordered

Univariate Analysis of ordered variables

- Analysis of a single column at a time is called Univariate Analysis using countpots





Analysis across loan status

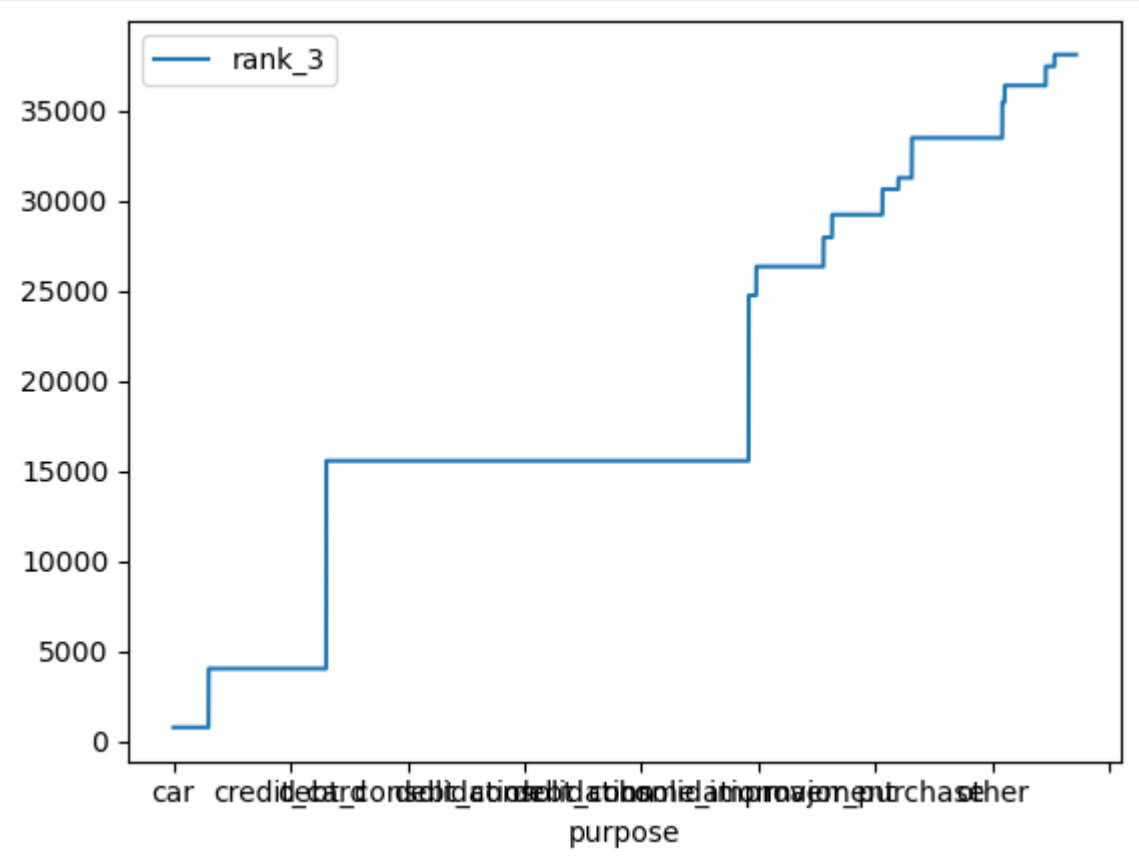
1) Most of the customers that are applying for loan tend get default
2) Default rate is higher on the total loan applications

Analysis across grade

1) Customers get loan maximum with grade B,C,A and D than E,F and G
2) Customers with grade B is maximum and with grade G the minimum

Univariate Analysis of Unordered variables

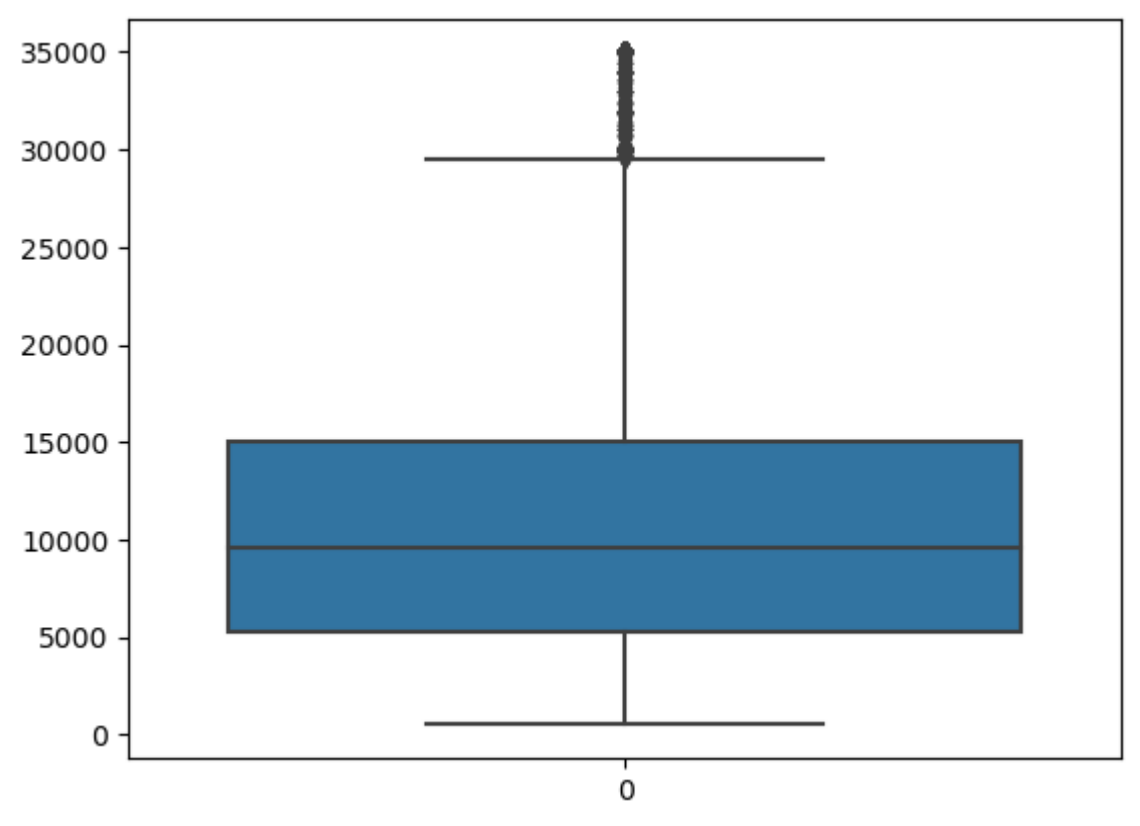
Using Rank Vs Frequency Plot



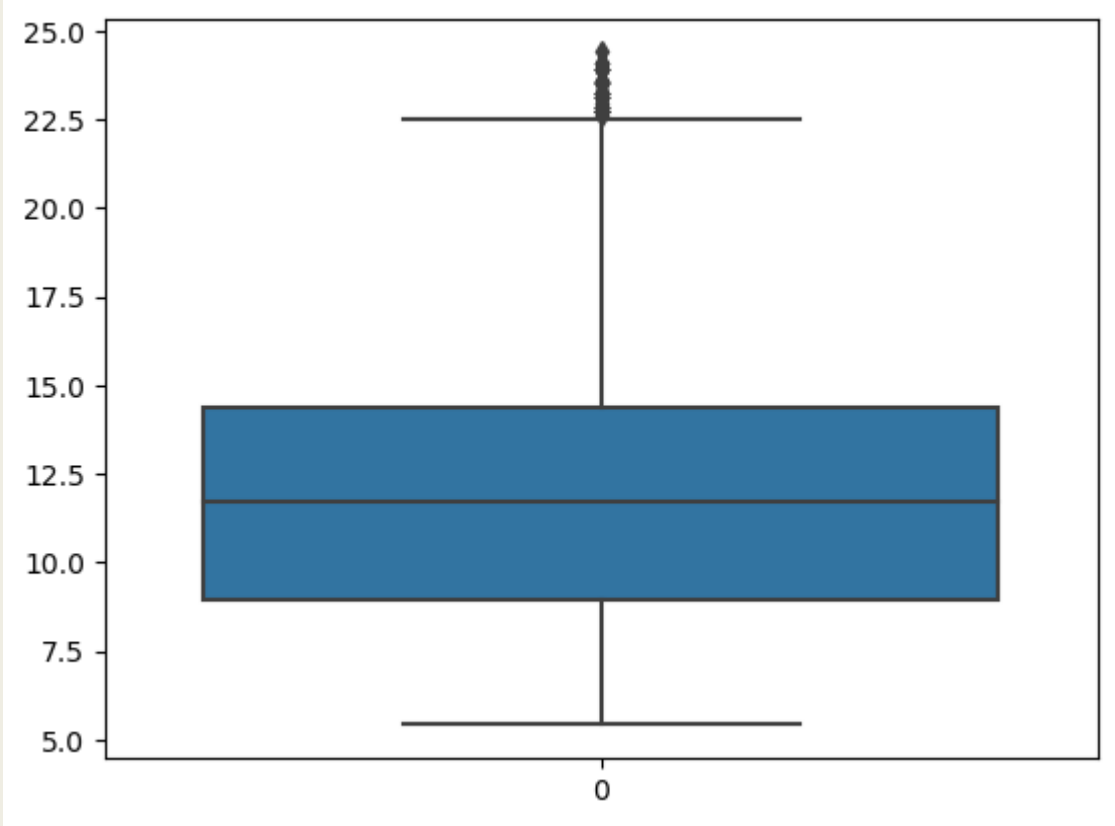
Most customers take loan for consolidation and major purchase

Univariate Analysis of Numeric variables

Using Box plot or Bar graph



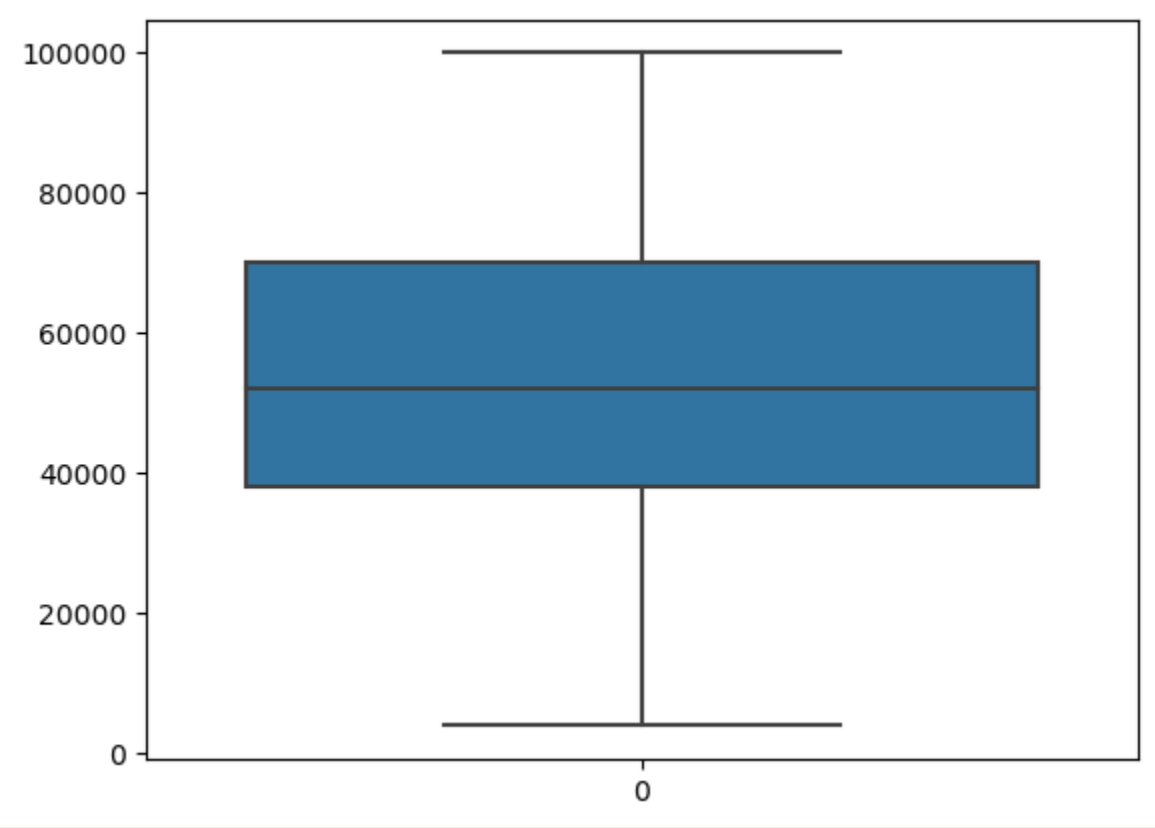
Box plot of Loan amount
1) Average loan amount is around 8000
2) The loan amount contains outliers as there are many data point which are more than 0.75 Percentile



Box plot of interest rate
1) Average interest rate is around 12%
2) The loan amount contains outliers as there are many data point which are more than the 0.75 Percentile

Univariate Analysis of Numeric variables

Using Box plot or Bar graph; removing the outliers wherever necessary

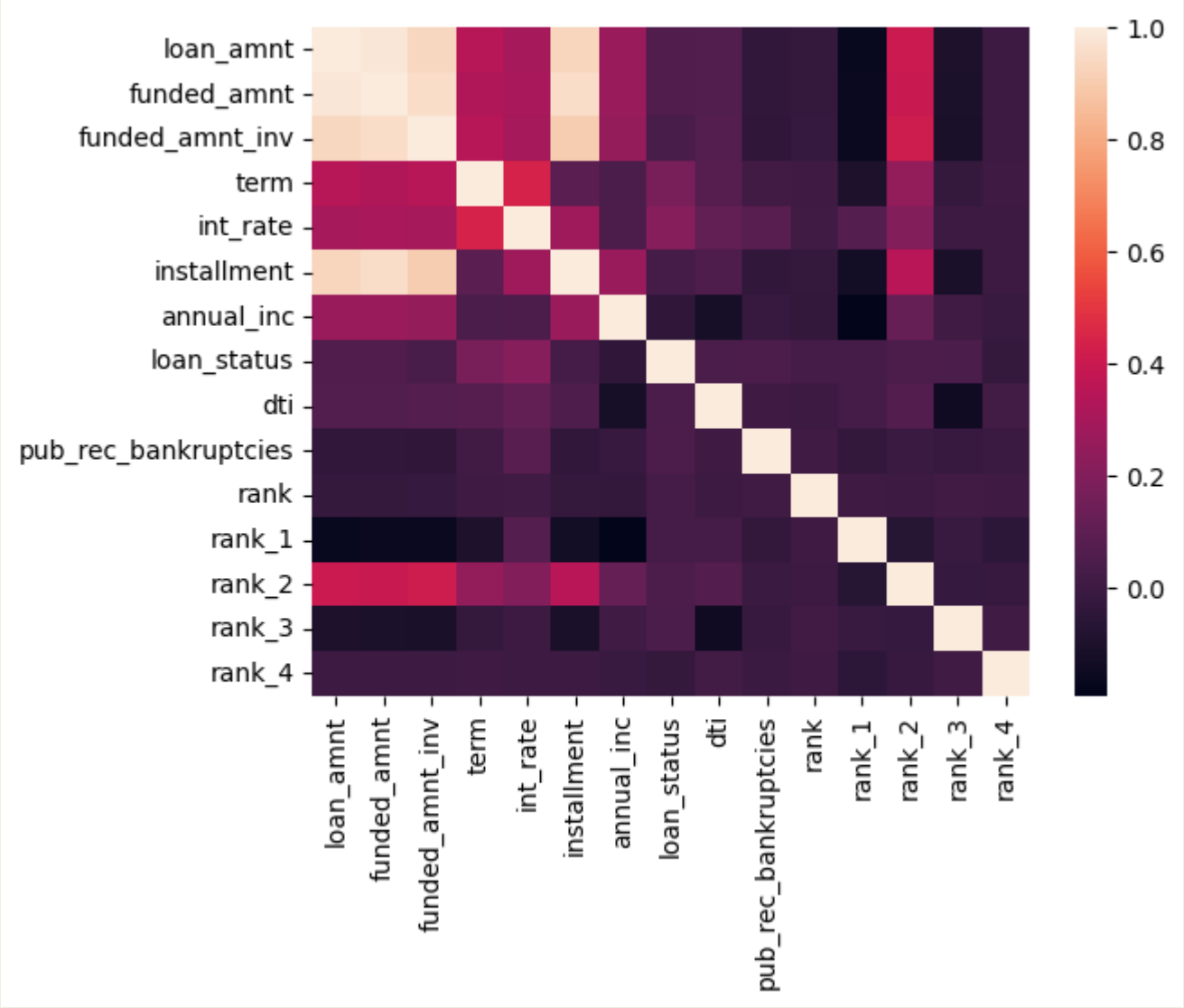


1) Outliers are more in the annual income column removing the same would improve the data analysis
2) After removing the average income is coming to be 50000

Bivariate Analysis of Num-Num variables

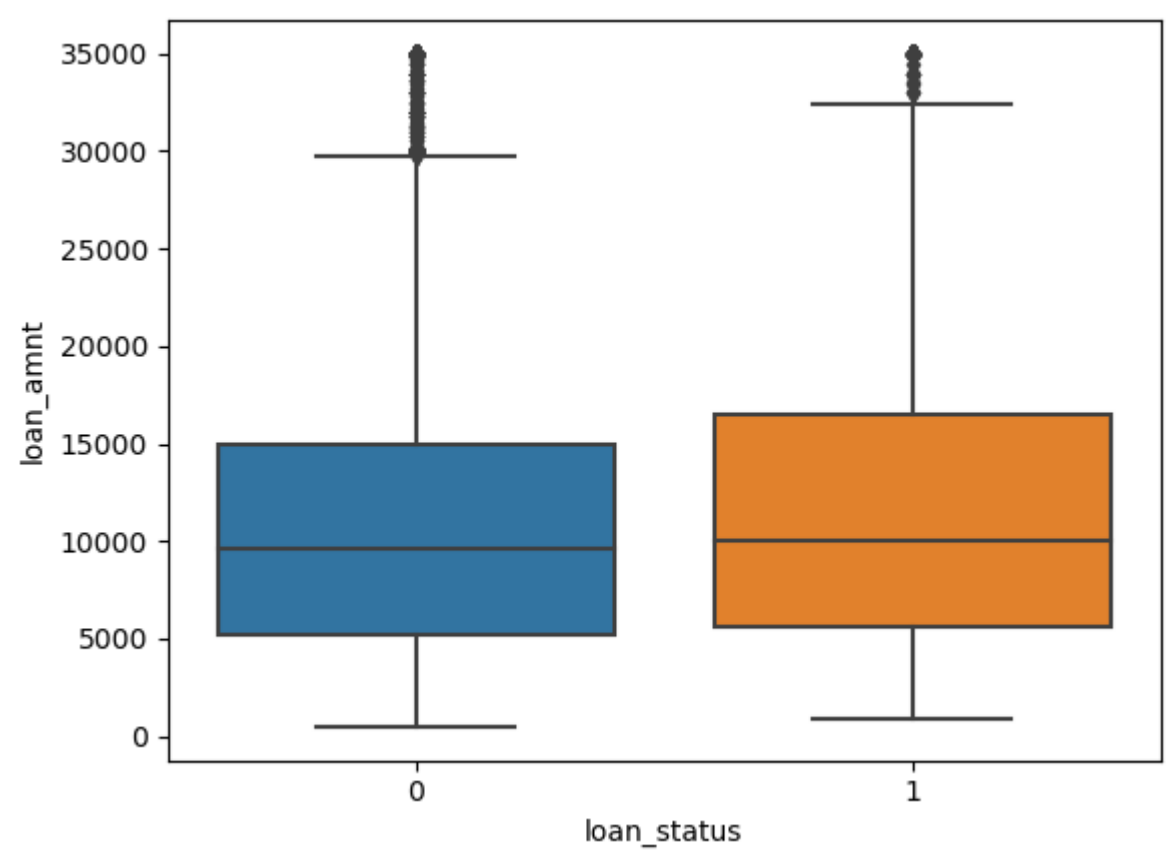
Using Correlation method and heat map

The annual income is highly correlated with the loan status and the least correlation with interest rate
Rank to Rank_4 was derived to understand the frequency

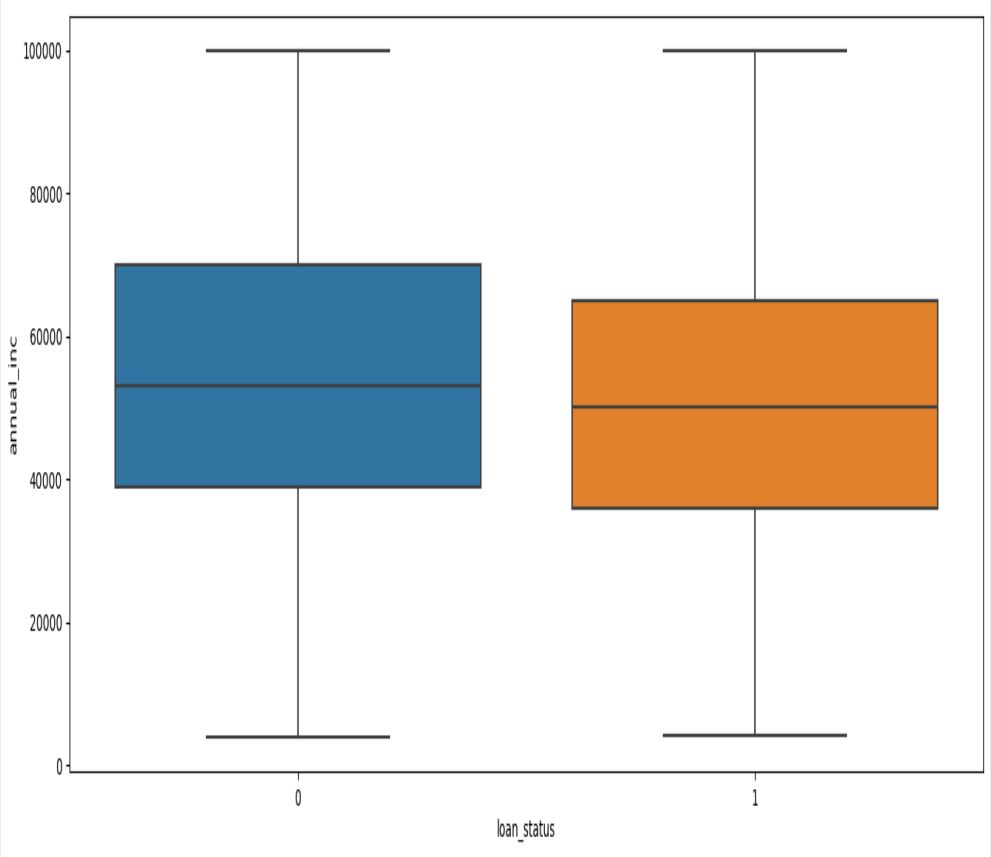


Bivariate Analysis of Num-Cat variables

Using Box Plot



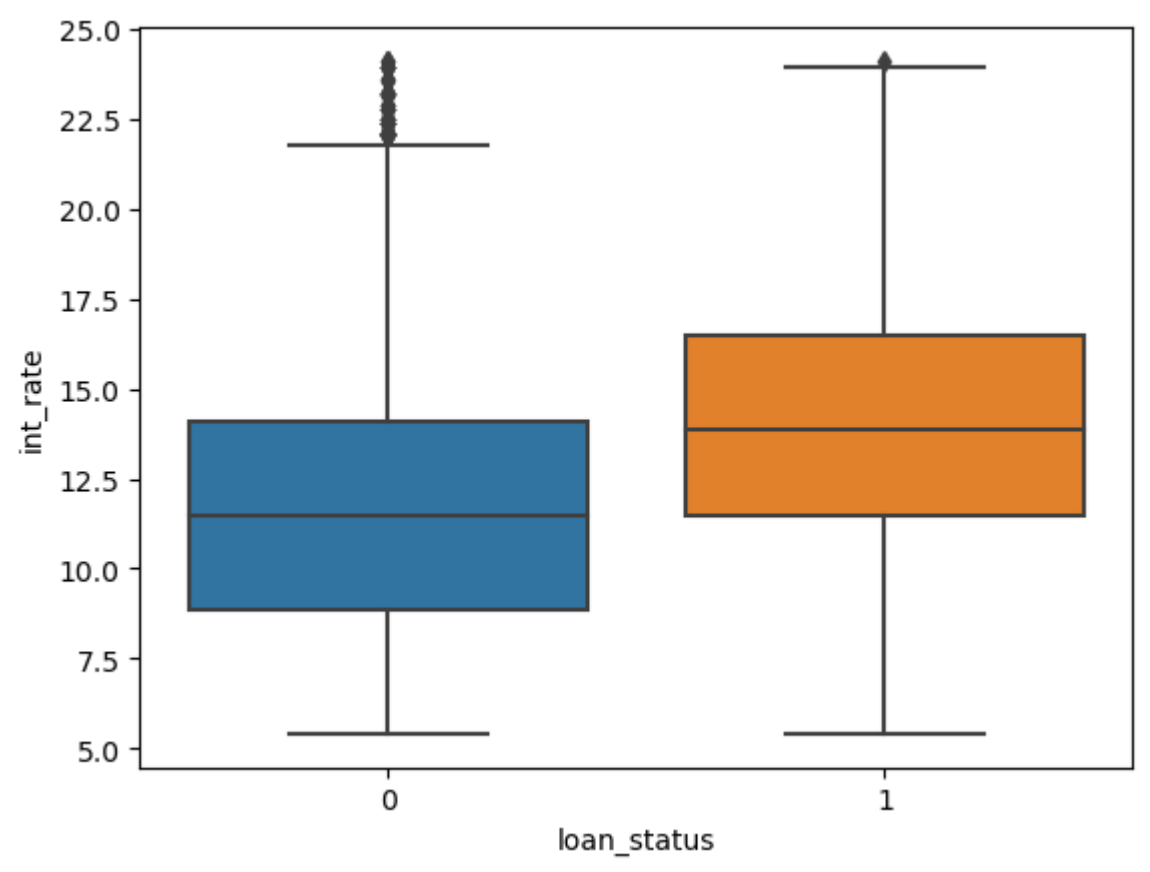
The customers with lesser loan amount tend to default more



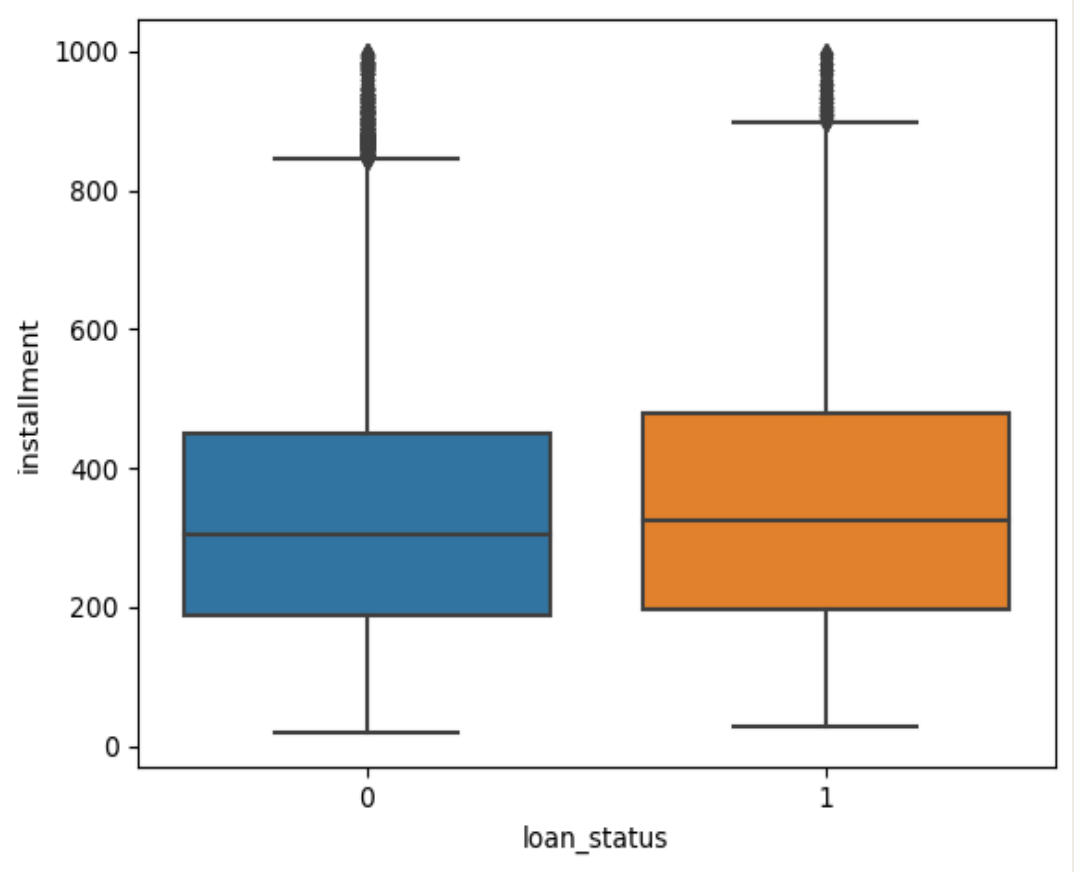
The customers with more annual income tend to default more

Bivariate Analysis of Num-Cat variables

Using Box Plot



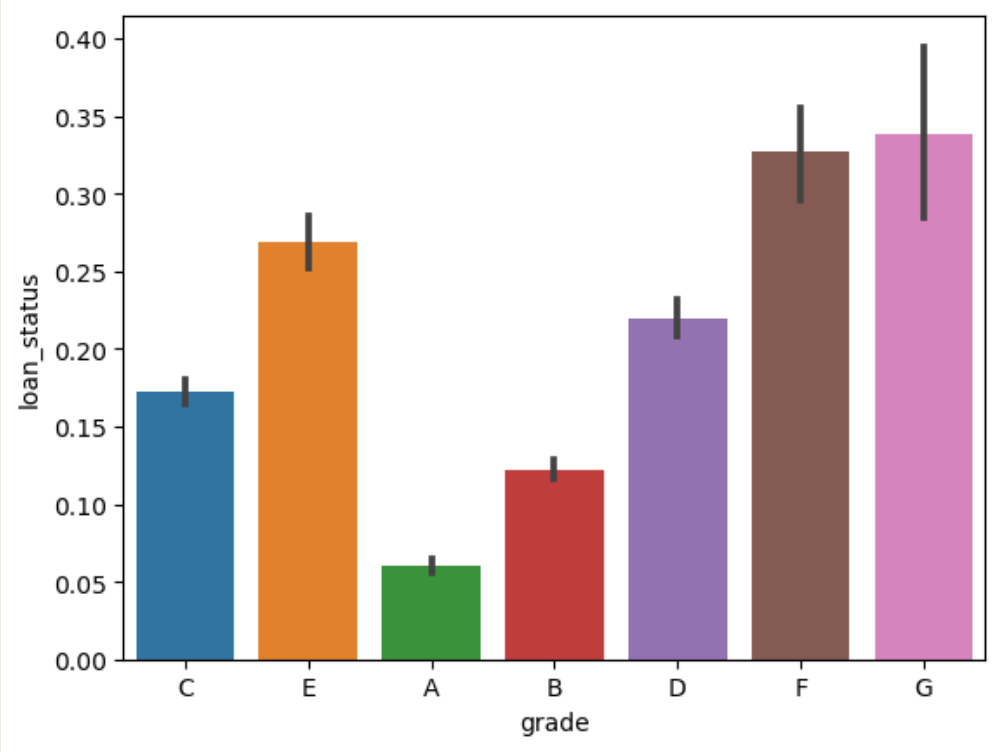
The customers with less interest rate tend to default more



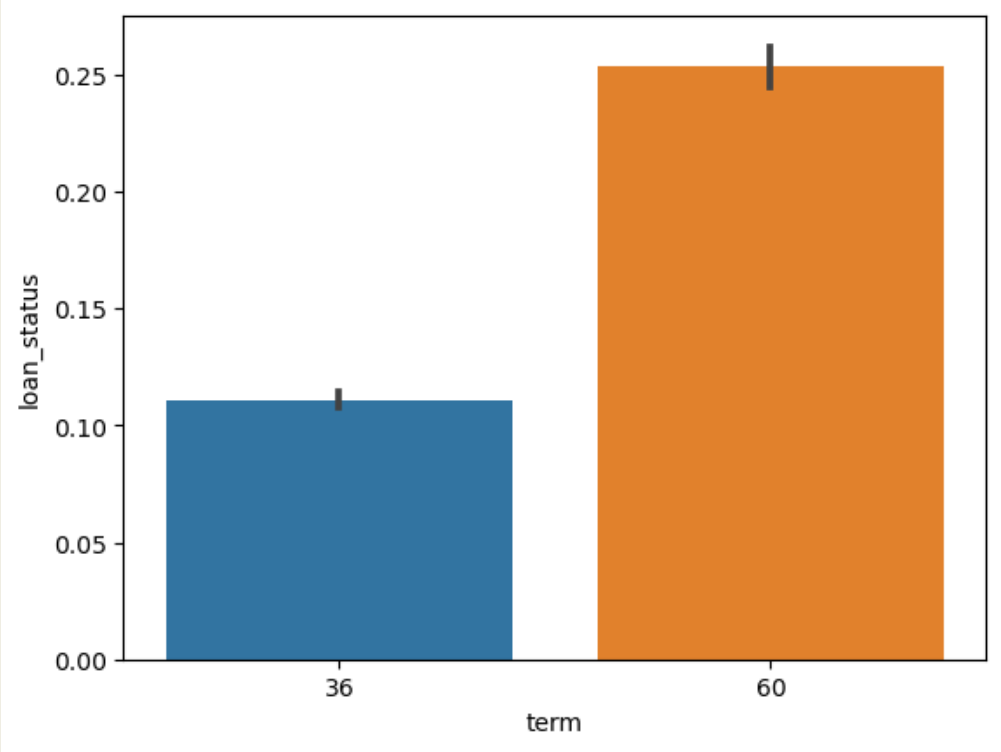
The customers with less installments tend to default more

Bivariate Analysis of Cat-Cat variables

Using Bar Plot



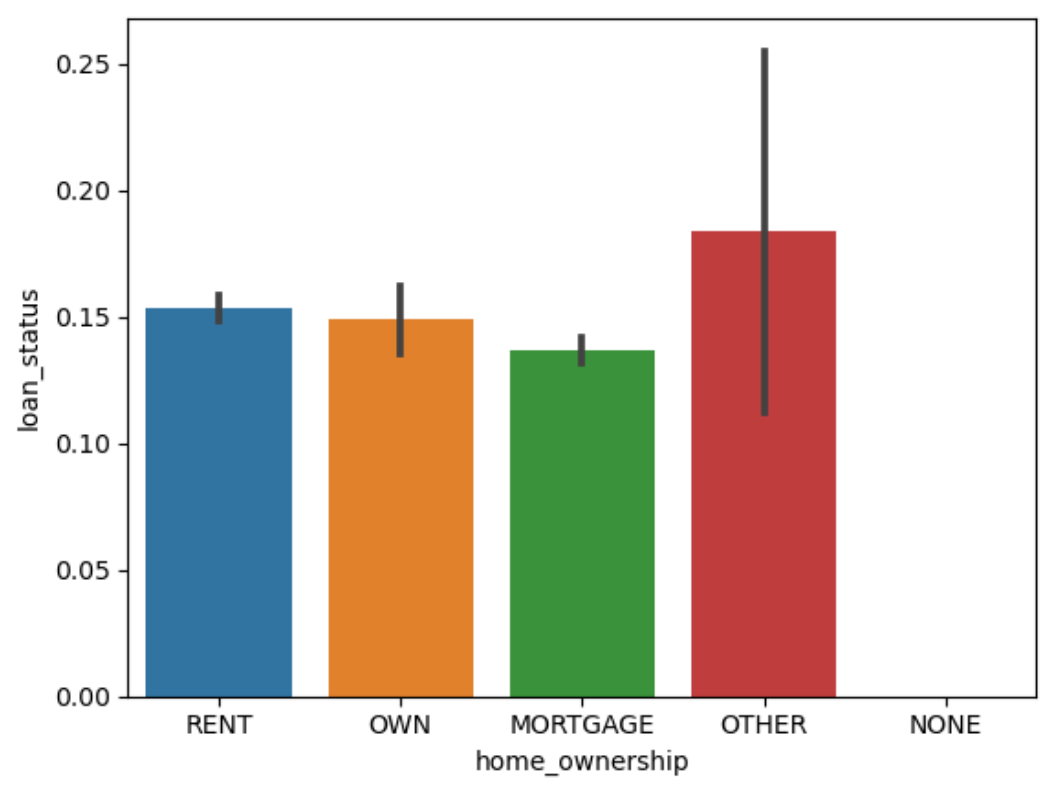
The customers with grade 'G' tend to default more.
The order are as below
Default Rate = G>F>E>D>C>B>A



Conclusion:- The customer with 60 months term period tend to default more than 36 months term period

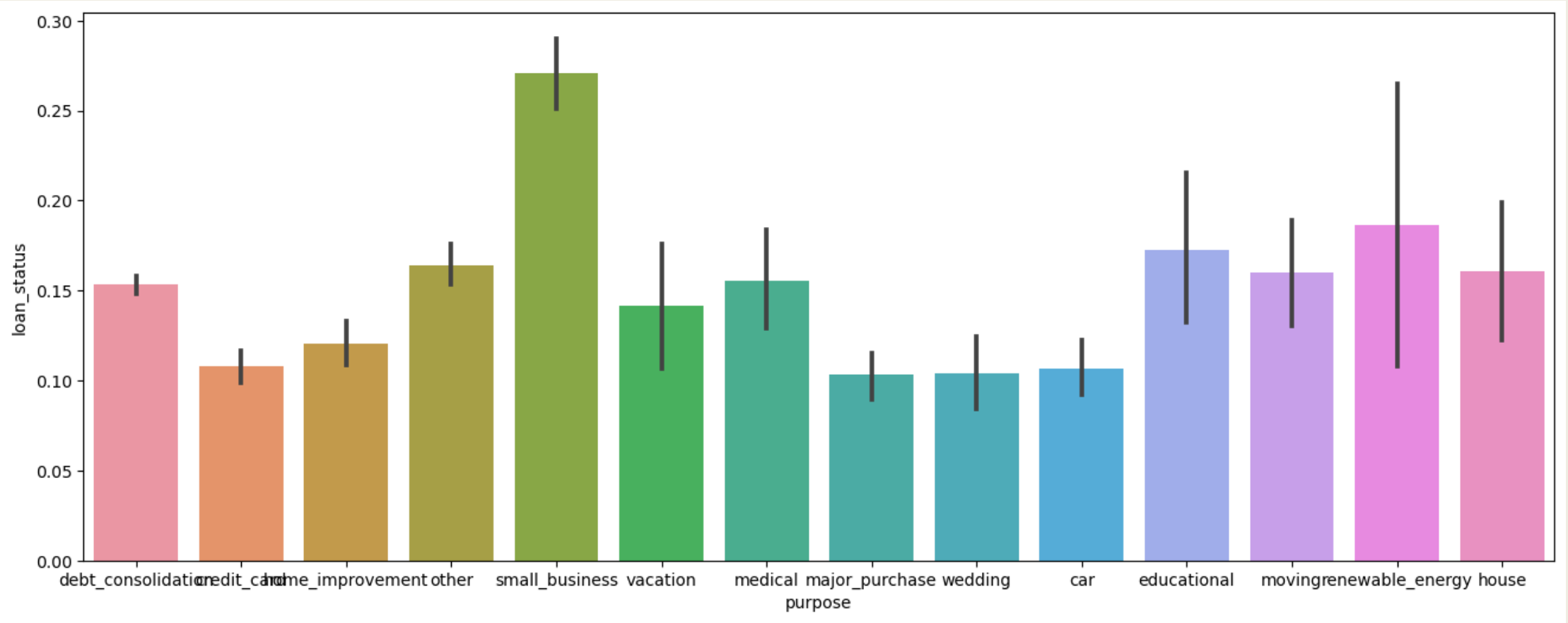
Bivariate Analysis of Cat-Cat variables

Using Bar Plot



Conclusion:- The customer with home ownership classified as other tend to default more.
The order is as follow:- Other>Rent>Own>Mortgage

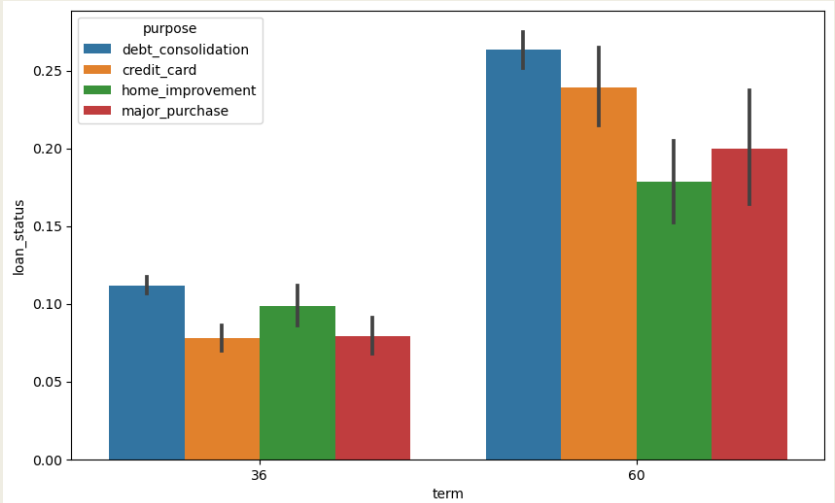
Bivariate Analysis of Cat-Cat variables

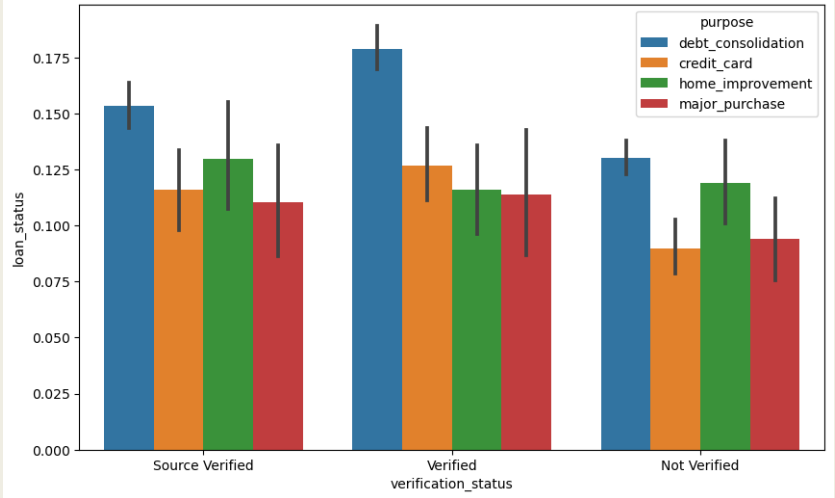


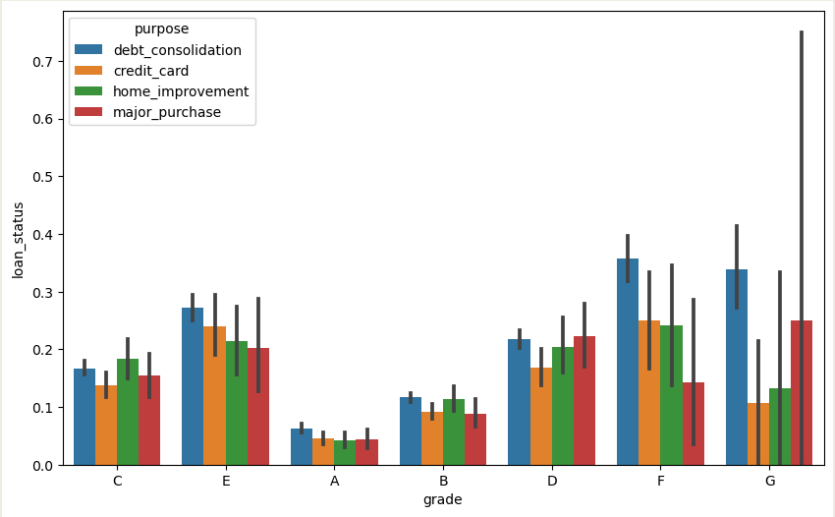
Conclusion:-The loan got for the purpose of small business has the maximum chances for getting default and the least is major purchase

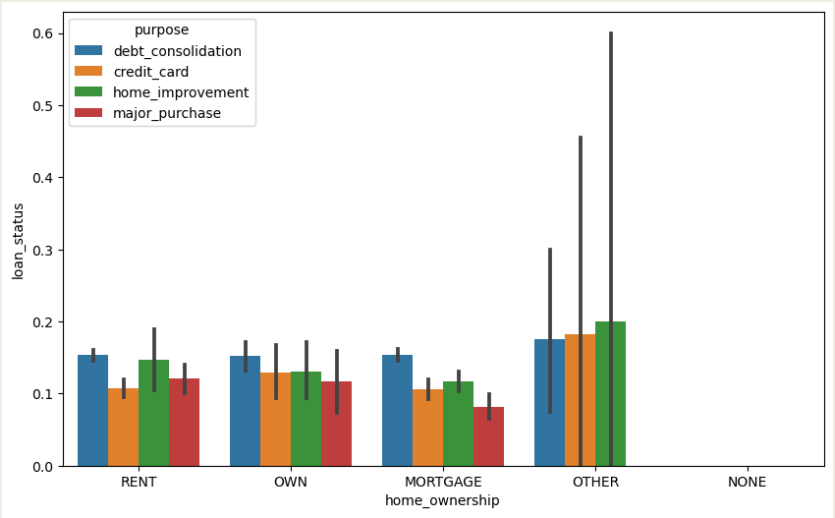
Segmented Univariate Analysis

Filtering the top four purposes and analyzing the default rate across various factors



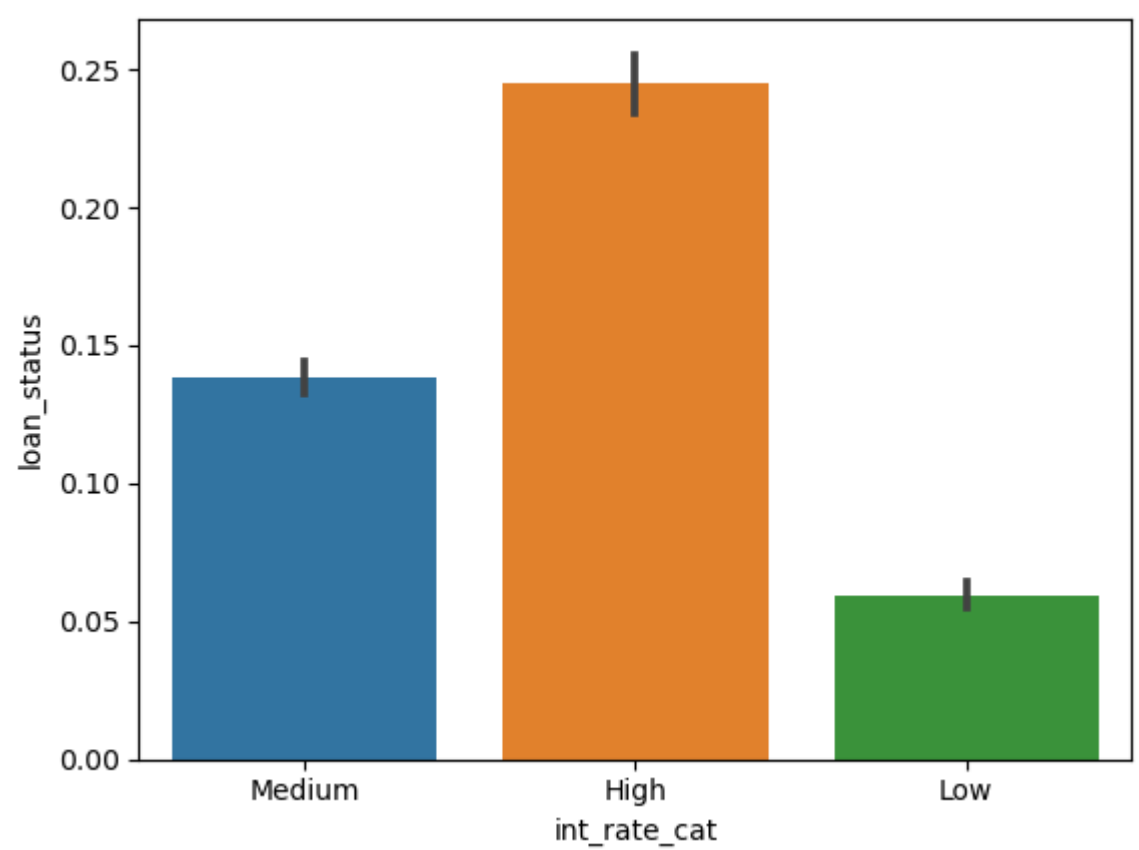


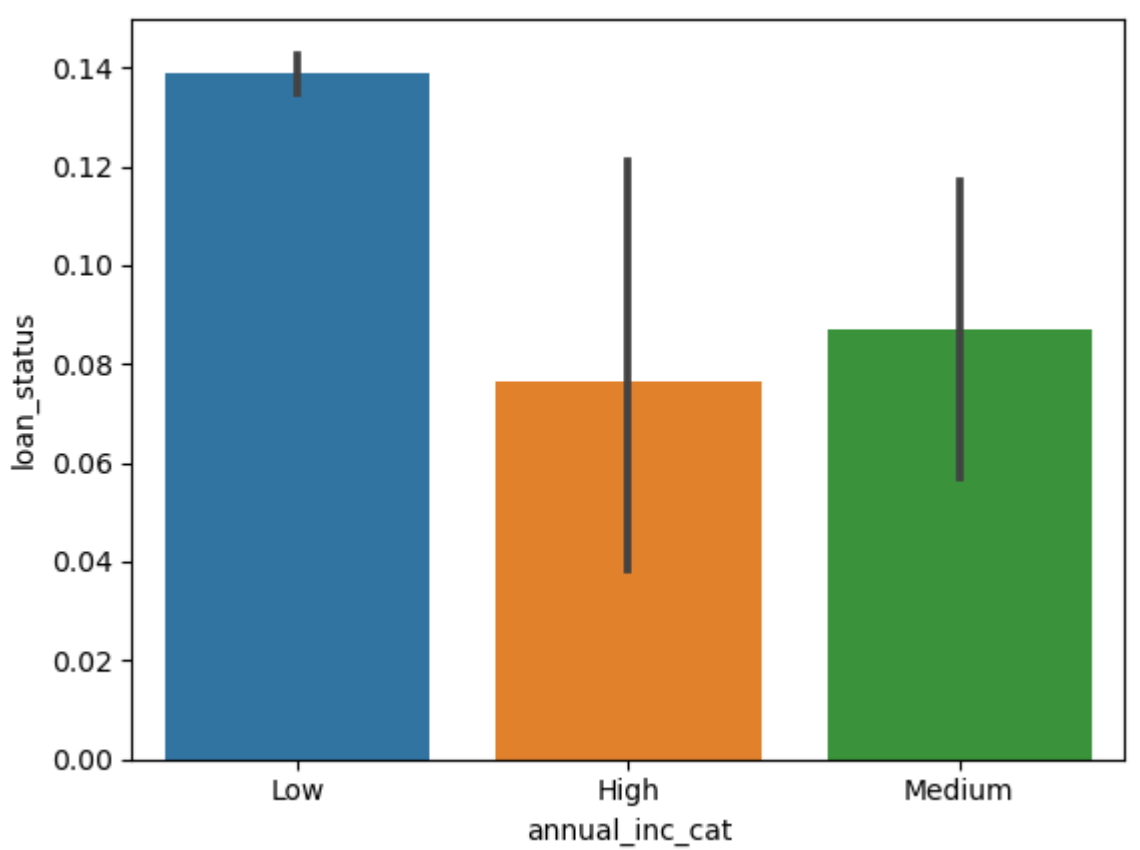




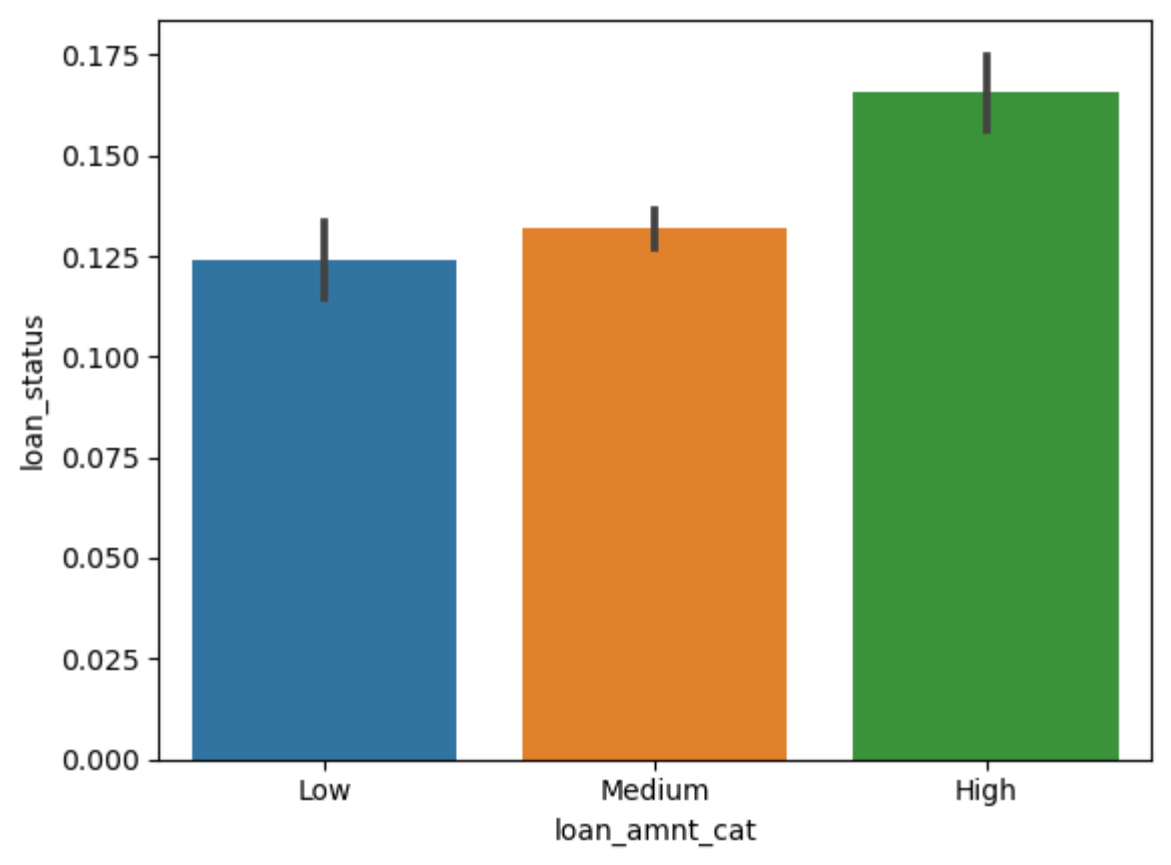
Derived Metrics

Converting Loan Amount, Annual Income, Interest rate and DTI into Categorical data and analyzing.
Analyzing Loan Default rate across month and year bey deriving month and year from Date column

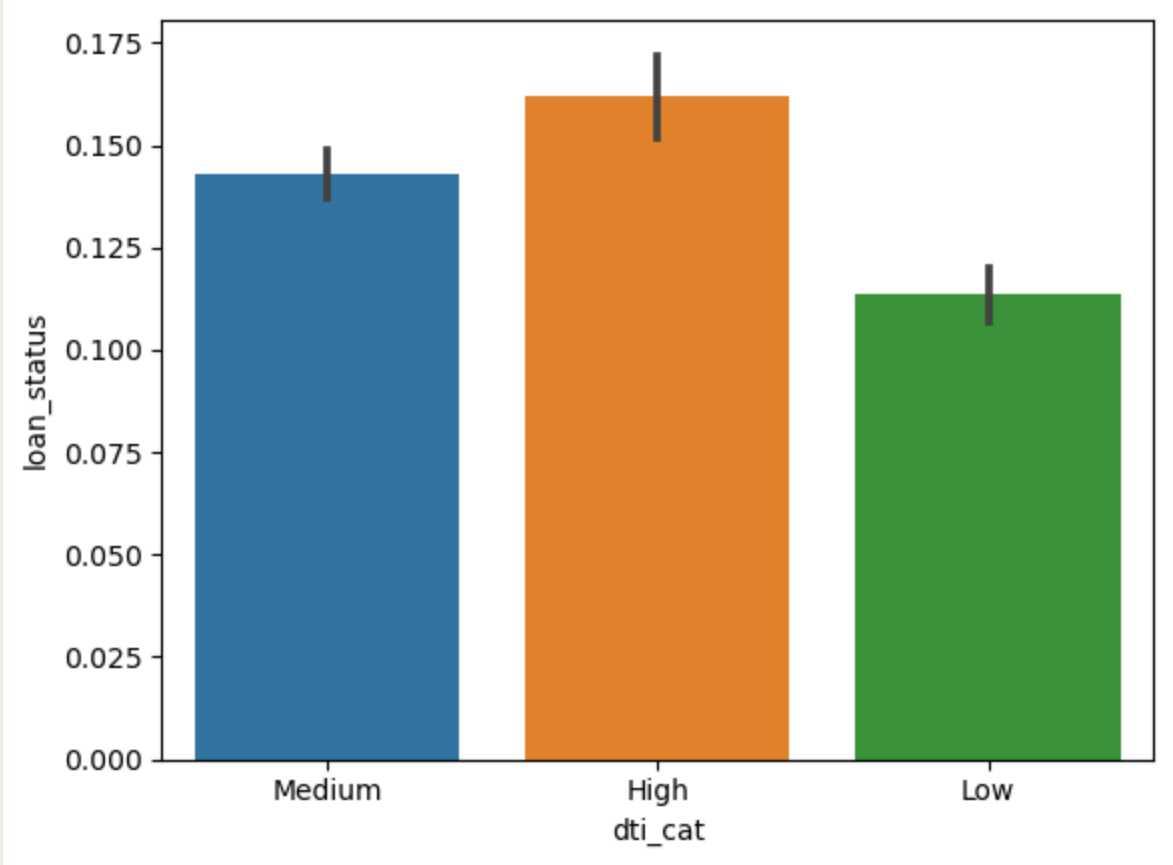




Derived Metrics

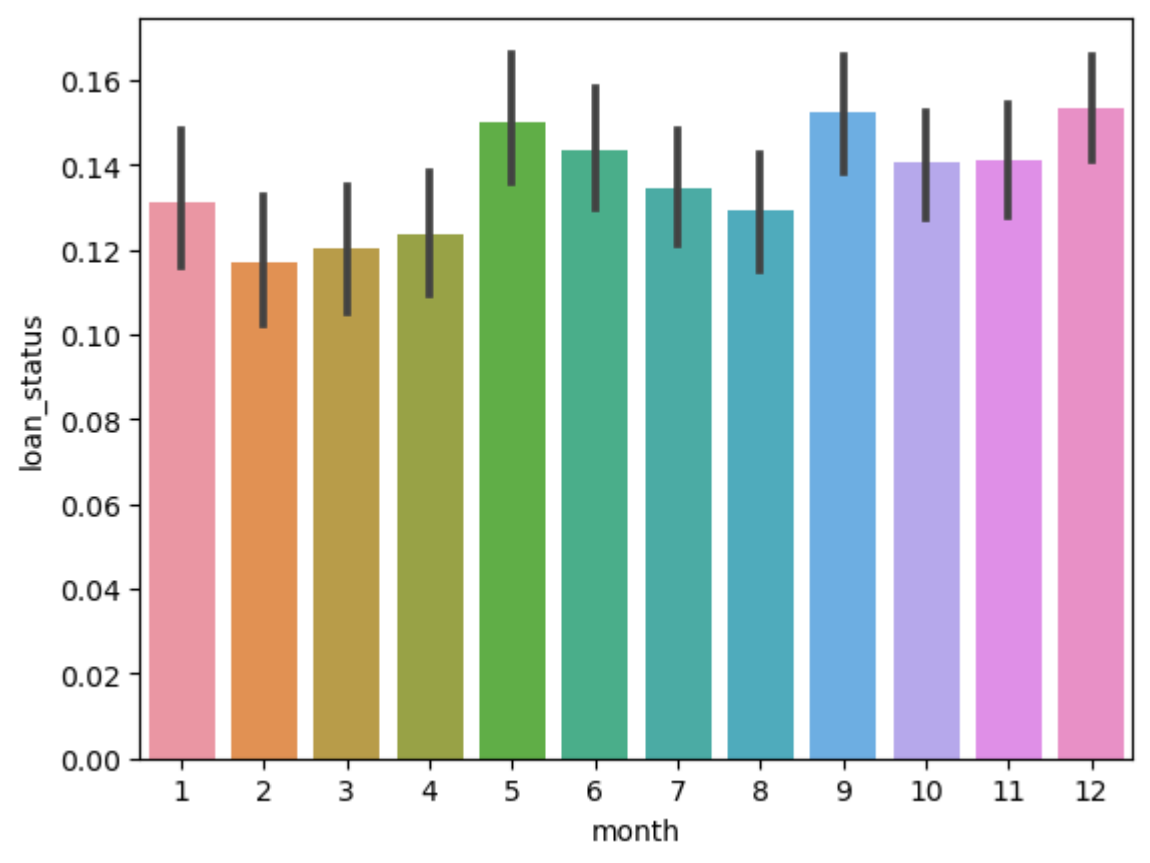


Conclusion:- The loan with higher loan amount tend to default more

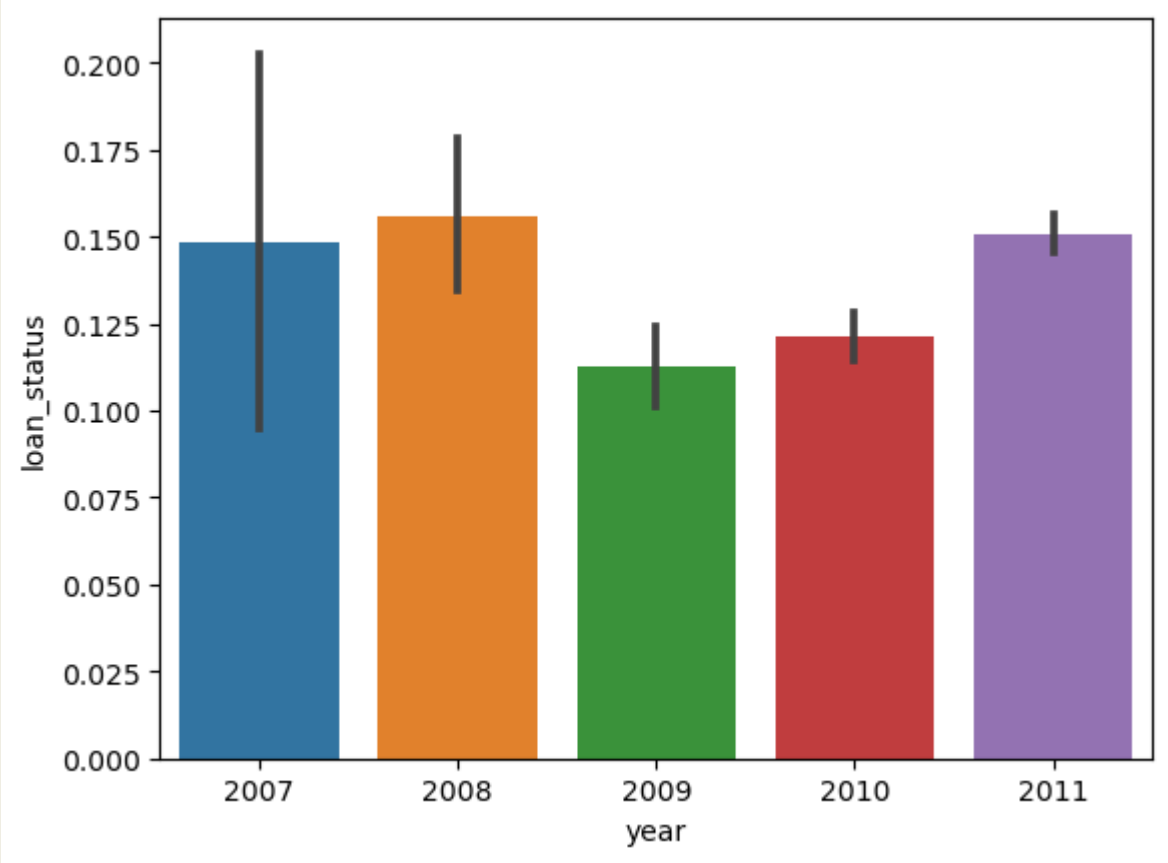


Conclusion:- The loan with high dti tend to default more

Derived Metrics



Conclusion:- The Loans taken in the month of December tend to default more



Conclusion:- The year 2008 has maximum default rates and the year 2009 has the least default rate