Lending Club Case Study

EDA analysis

Predicting Loan Approval and Analyzing Factors Influencing Loan Defaults

- **Background**
- Financial institutions are in the business of lending money, and one of their main challenges is to assess the risk associated with each loan application. A crucial part of this assessment involves predicting whether a loan applicant is likely to default on their loan. By accurately predicting loan defaults, lenders can minimize risk and make informed lending decisions.
- **Objective**
- The study aims to analyze the factors that influence loan defaults and provide insights into the risk factors associated with loan applications.

Begin

- As we start analyzing the data from loan.csv file using python libraries. We import the libraries first, then read the csv file using pandas as data frame.
- We try to find out what columns and row values it has, also shape by using .head(). and .shape() and try to familiarize with the data values.

Cleaning

- We start the EDA by cleaning data. Find if there are missing columns or rows, try to impute or remove them.
- In our case, we had lot of columns that were missing data, hence we removed those columns with total null values.
- Also, there are lot of columns, with single values, we removed them as well.
- Also remove unwanted rows and columns, that doesn’t contribute to our analysis, it is only going be overhead if we do not remove them.

Analysing the data; check for missing/Null values

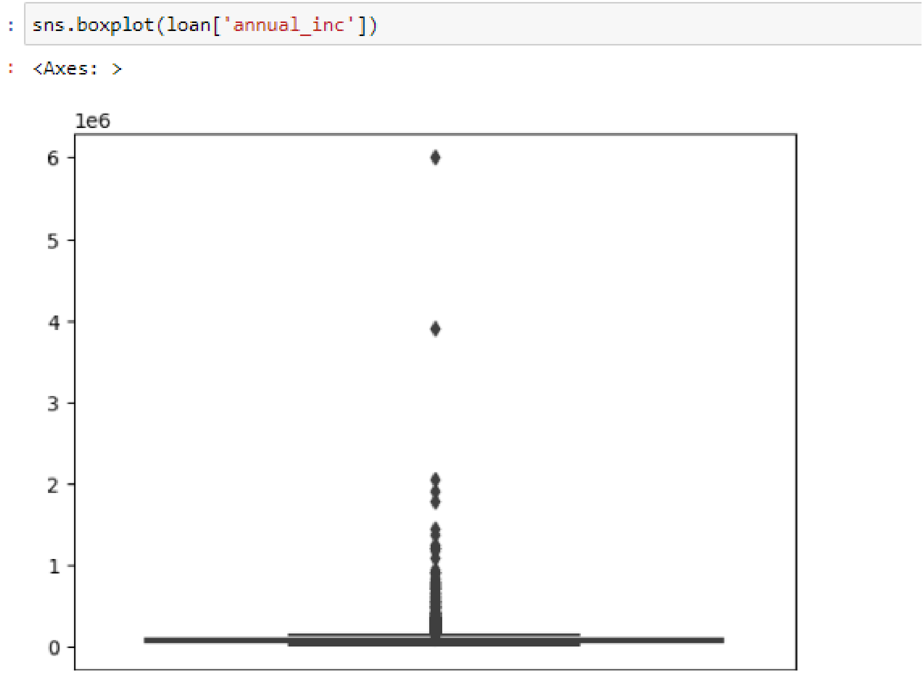
- We then get familiarize with data types, values of individual columns and rows, see which ones are categorical and numerical respectively.
- Like example, loan_status, etc columns can be used as categorical,and int_rate, loan_amnt etc as numerical.
- Find out Missing and Null values in rows columns, either impute or remove them.

Standardization of data

- Analyze and Standarize the data values accordingly.
- Example, int_rate, revol_util are objects but have continuous values, hence they need to changed to ‘int’ types.
- emp_length" --> { (< 1 year) is assumed as 0 and 10+ years is assumed as 10 }
- > Although the datatype of "term" is arguable to be an integer, there are only two values in the whole column and it might as well be declared a categorical variable.

Checking for outliers

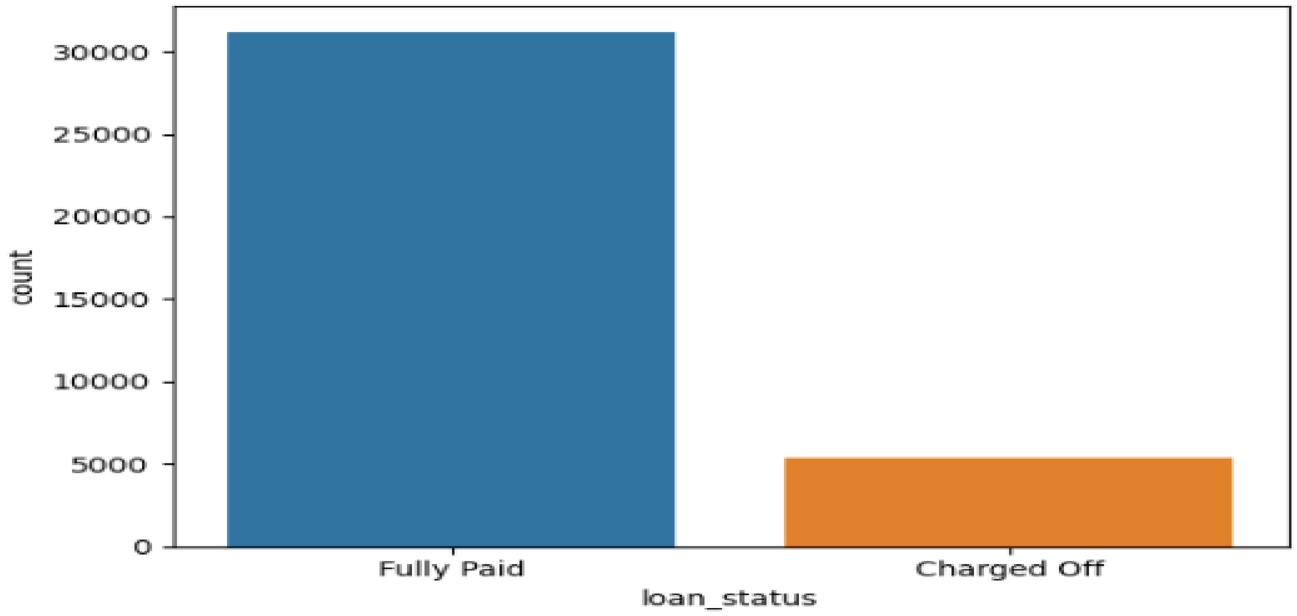
- Check if there are any outliers for numerical columns and see if we can remove them accordingly.
- In our case, we saw some outliers for column ‘annual_inc’, we removed them.



Visualizing Data

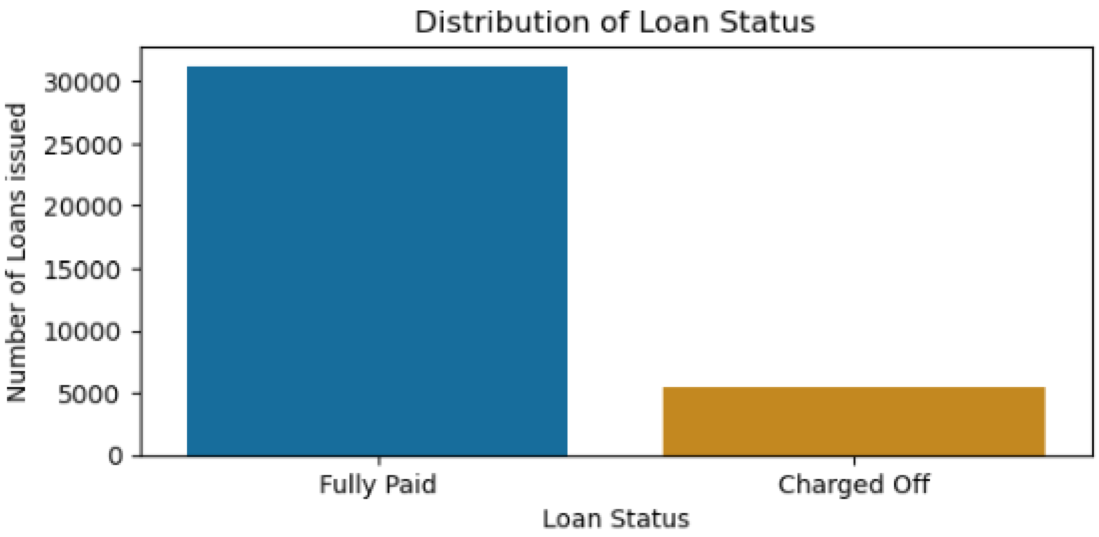
- We being the visualization process, try to plot graphs of various parameters perform univariate, bivariate analysis.

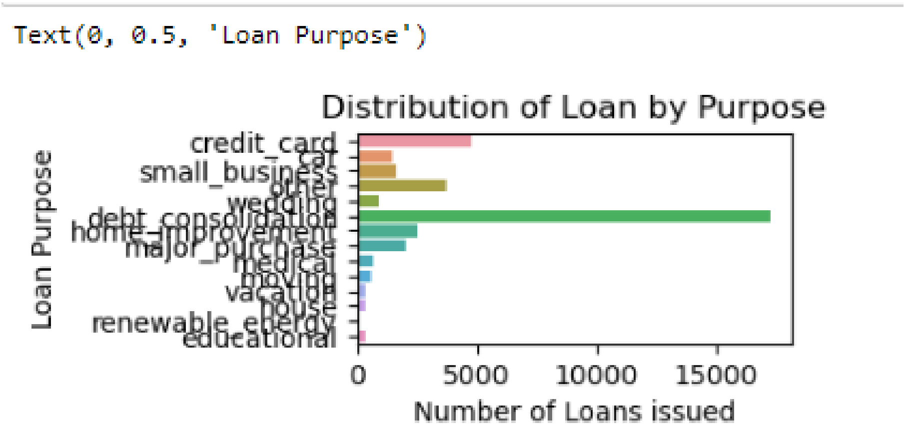
```
[67]: sns.countplot(x='loan_status', data = loan)
t[67]: <Axes: xlabel='loan_status', ylabel='count'>
```



Univariate Analysis

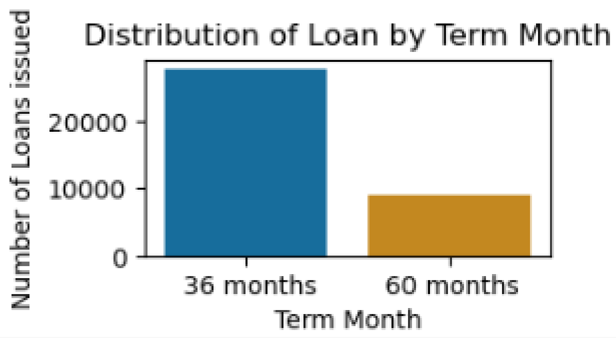
- We perform various univariate Analysis for various columns such as loan_status, ‘purpose’, ‘home_ownership’, ‘term’, ‘verification_status’, etc.

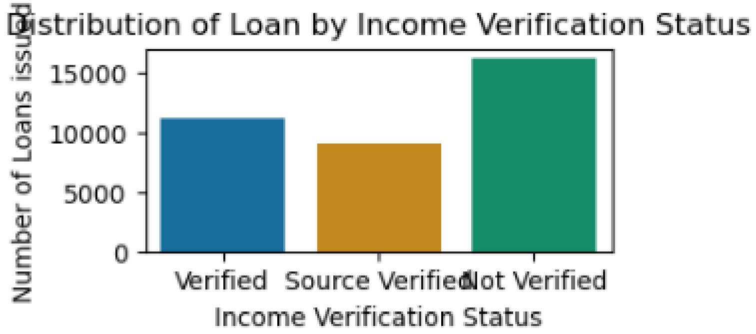


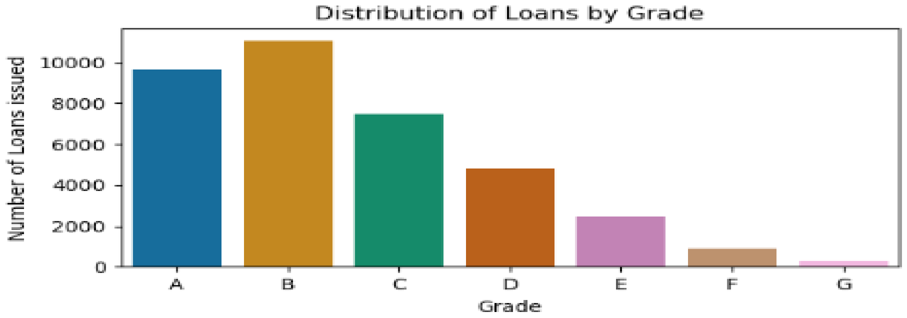


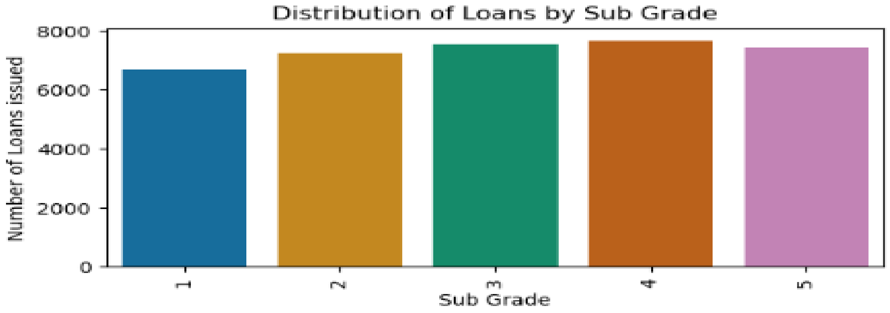
univariate analysis(continued..)

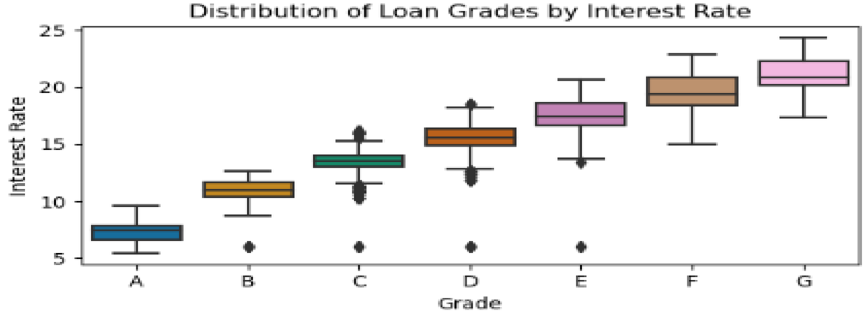
```
Text(0, 0.5, 'Number of Loans issued')
```

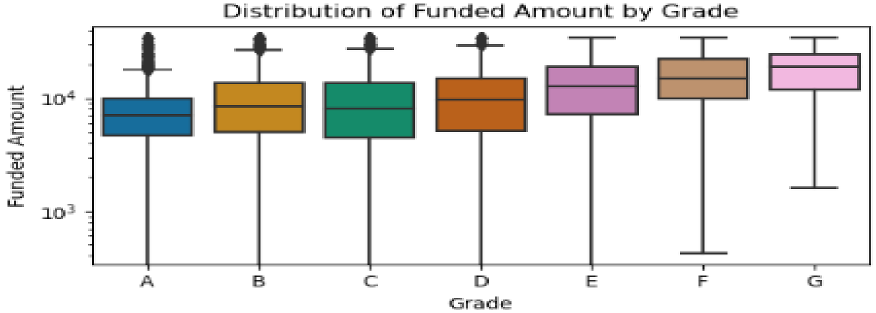












univariate analysis(continued..)

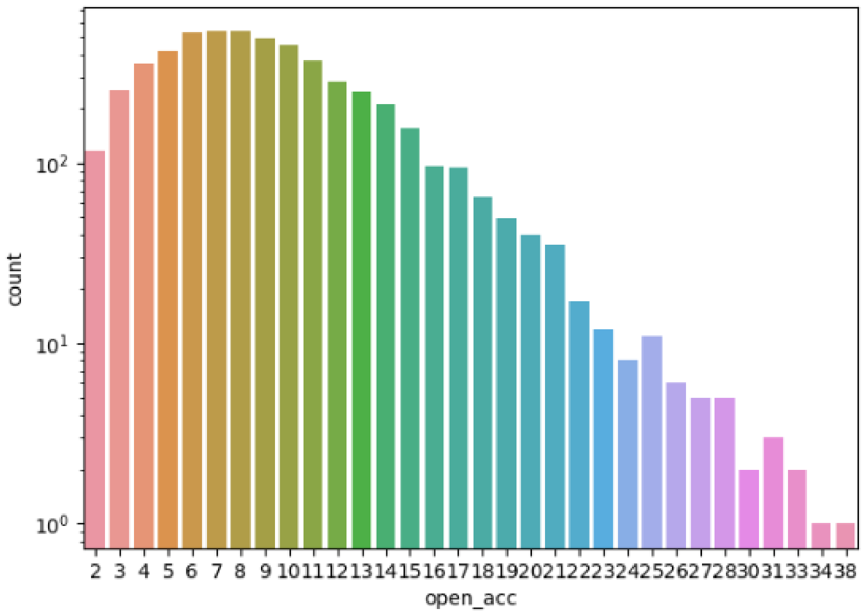
- Observations:
- 1. 85 % are in fully paid status.
- 2. there are more applicants from debt consolidation
- 3. There are more applicants from rented and mortgage
- 4. More number of loans are with 36 month term
- 5. More number loans income verfication status is not verified.
- 6. more number of loans were from B,A and C grade's and least from G grade.
- 7. it shows that A,B,C grade loans have less interest rate and E,F,G have high interest rate.
- 8. it shows that there are high funded amount in A,B,C and D grades.
- 9. The majority of borrowers have been employed for at least 10 years.
- 10. There is a huge number of charged off loans in 2011
- 11. In December month a huge number of loans are issued, probably because of Christmas time

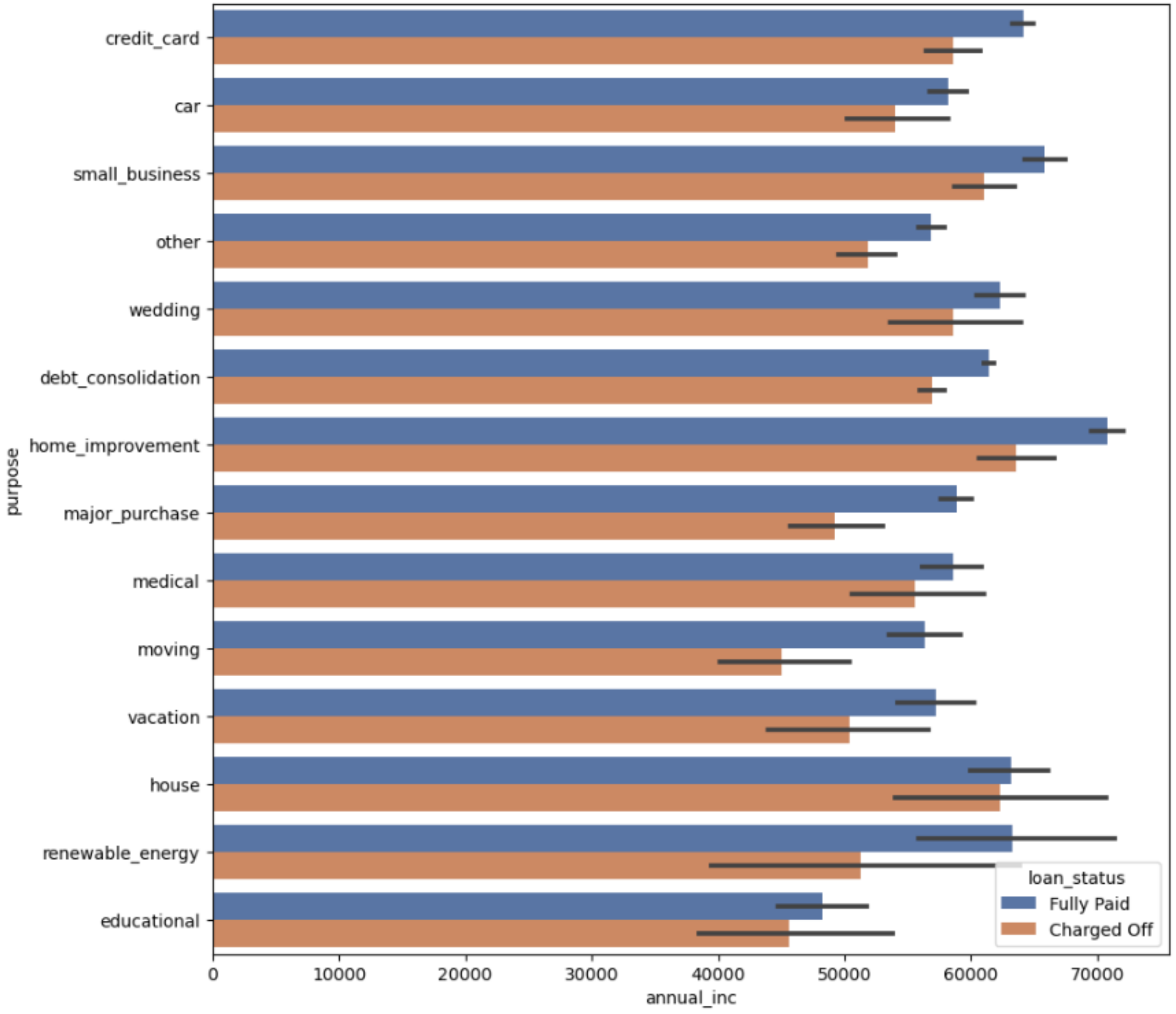
Bivariate Analysis

- Similarly, we plot graphs for bivariate analysis and try to analyze the graphs and find out if any patterns emerges.

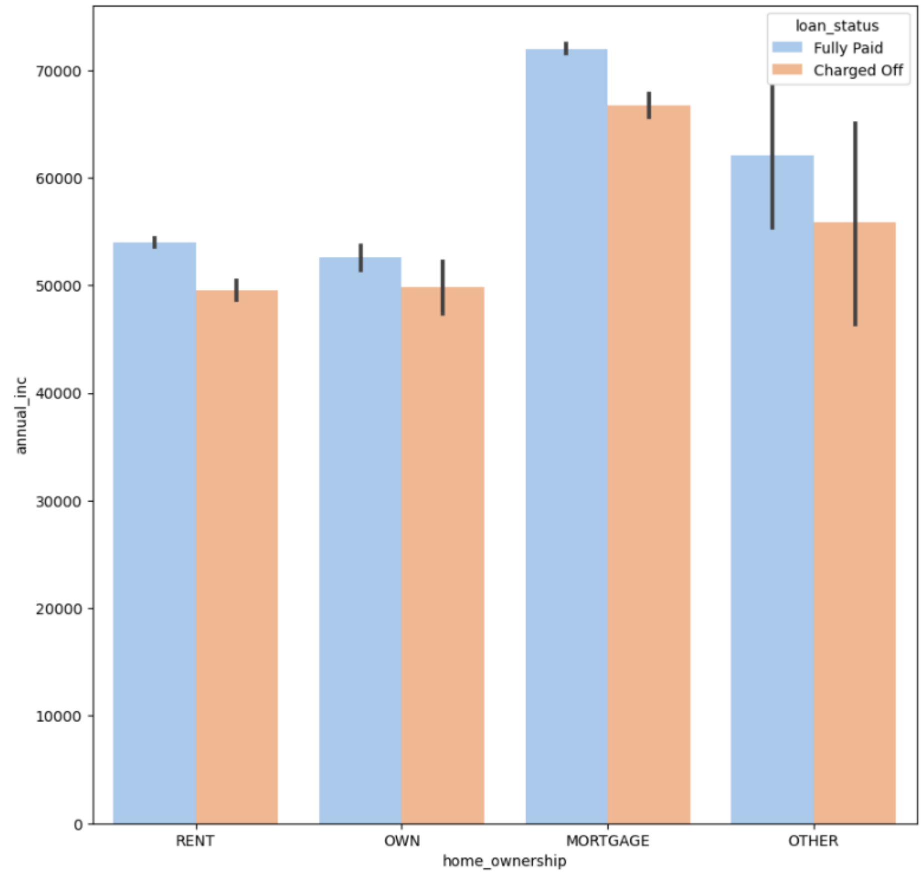
Bivariate Analysis graphs

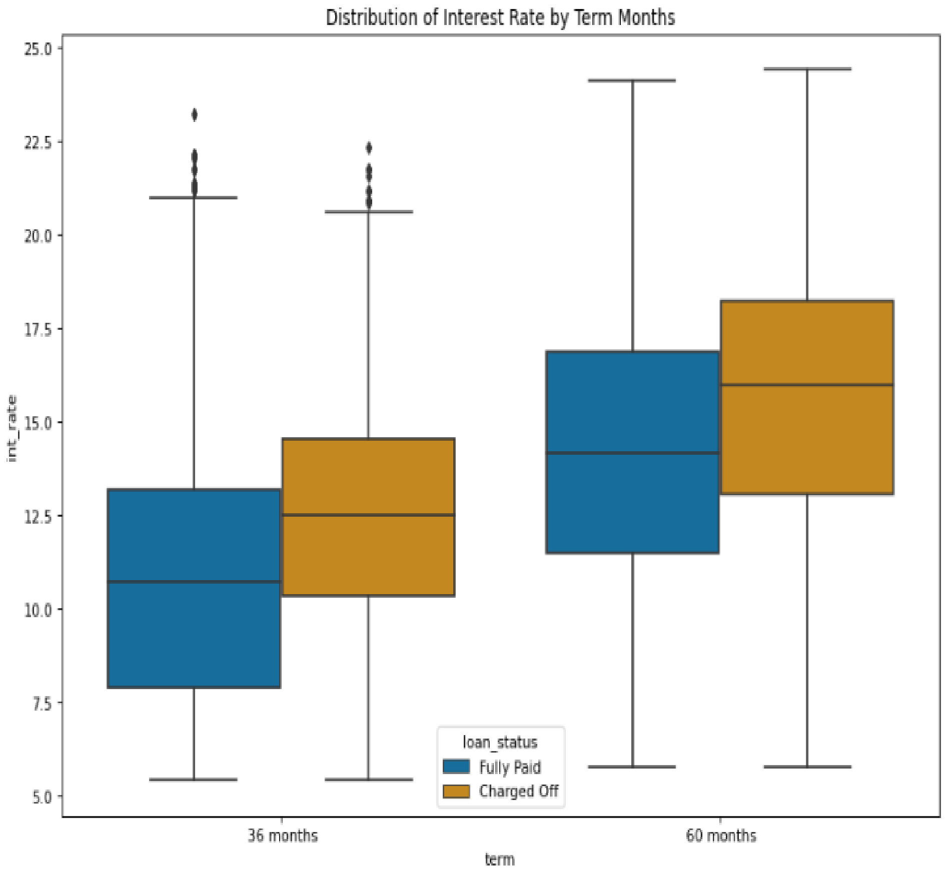
```
02]: <Axes: xlabel='open_acc', ylabel='count'>
```



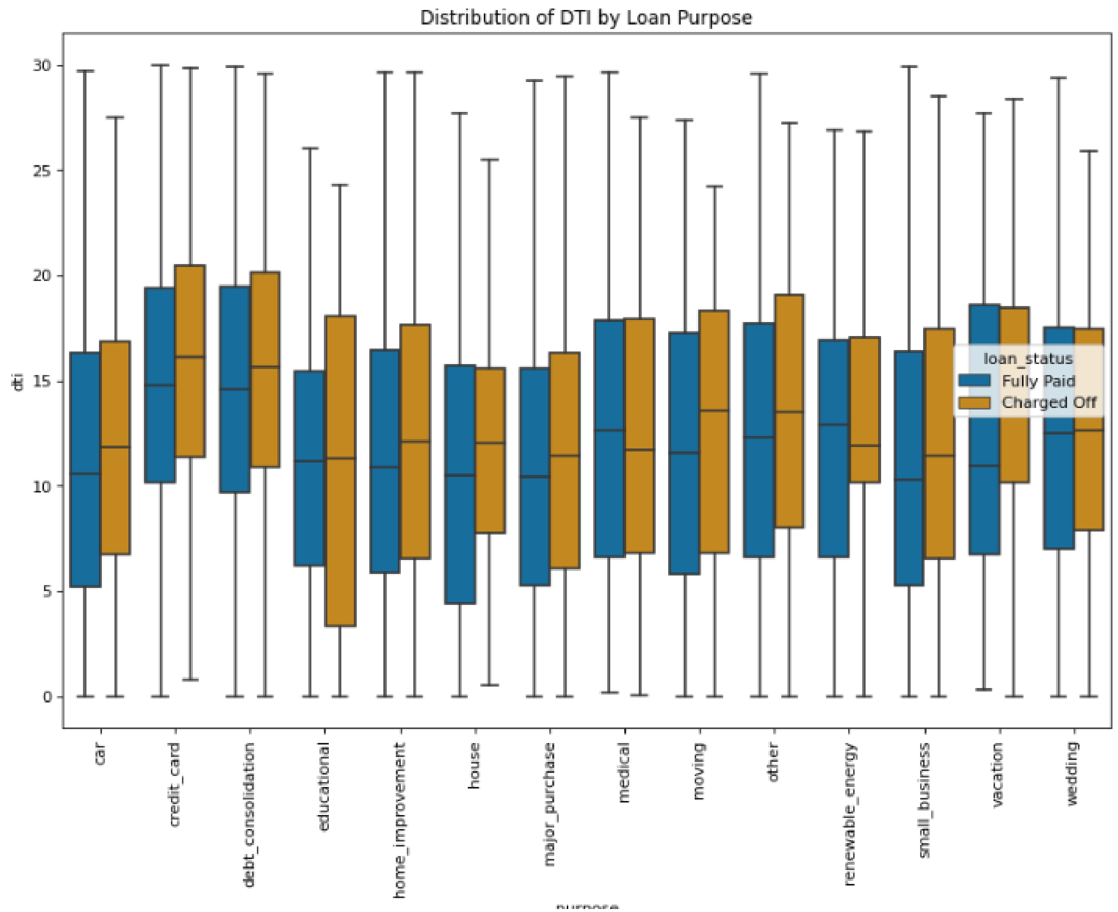


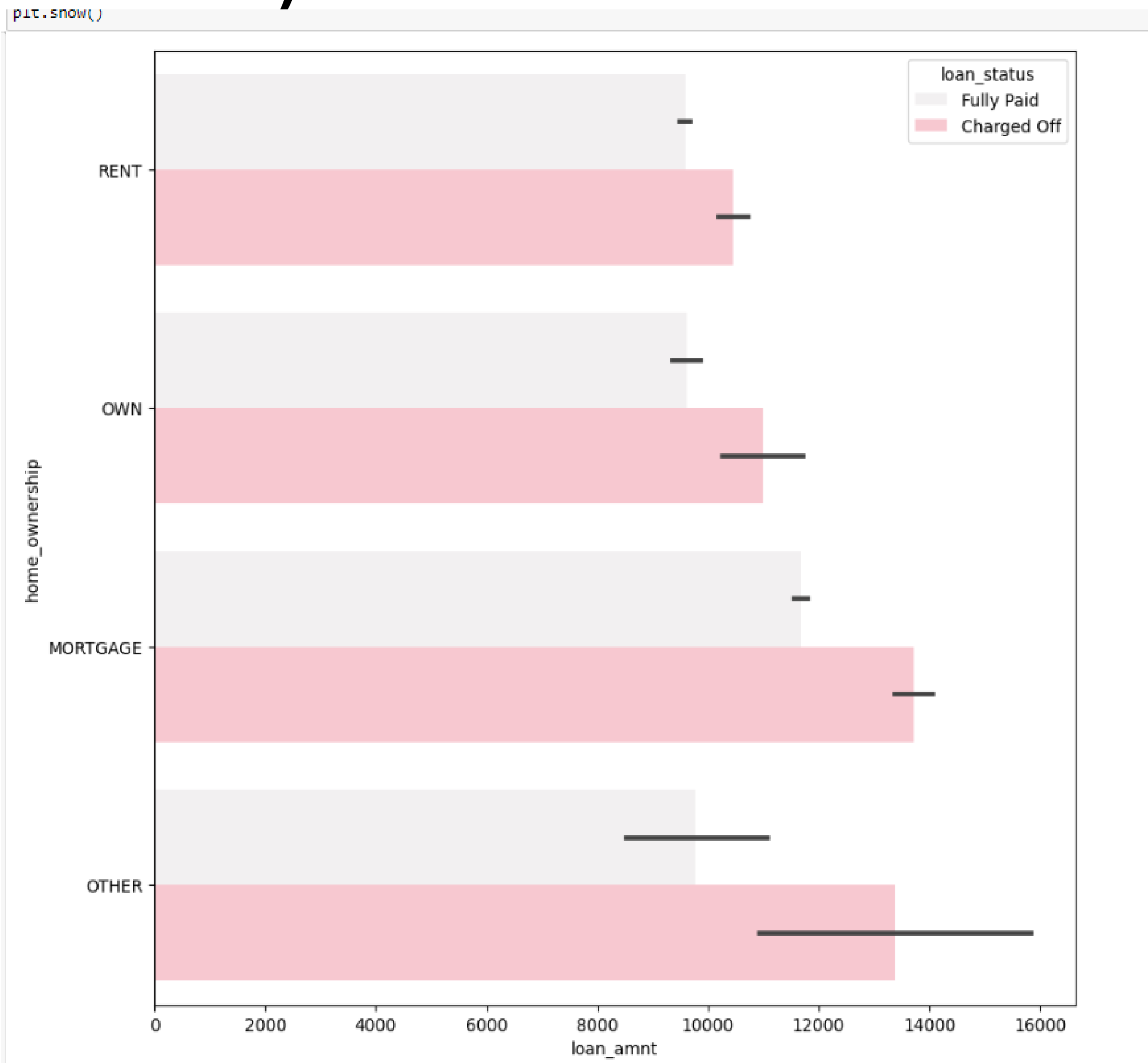
Bivariate Analysis(continued...)





Bivariate Analysis(continued...)



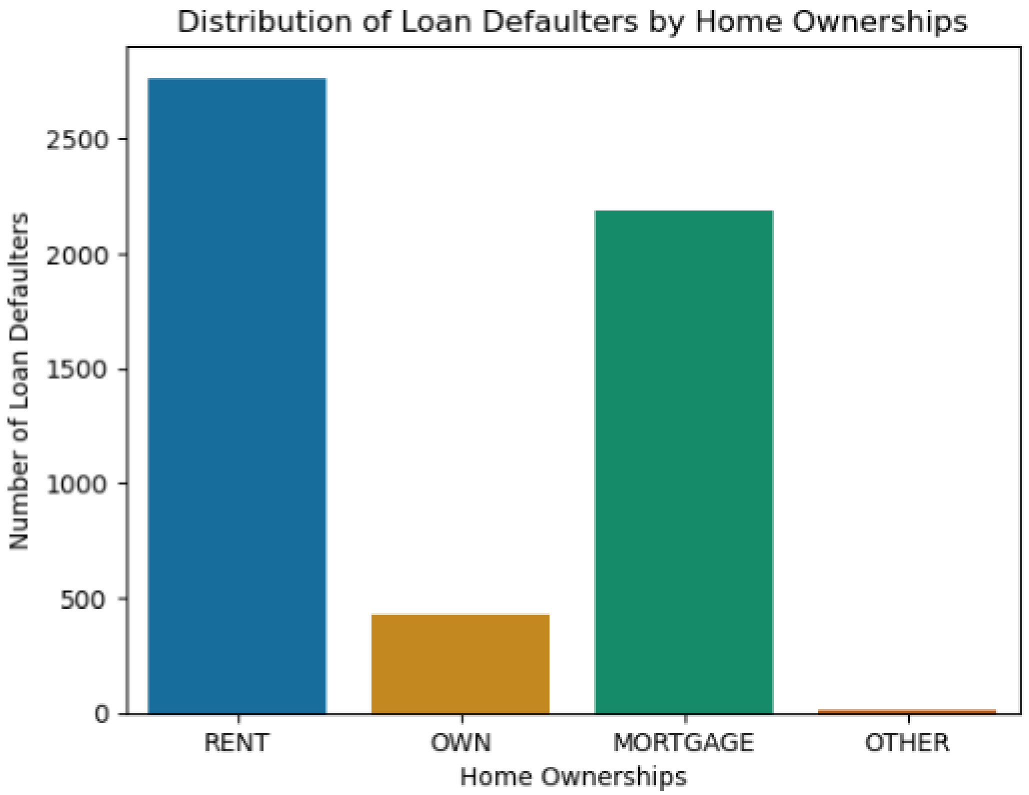


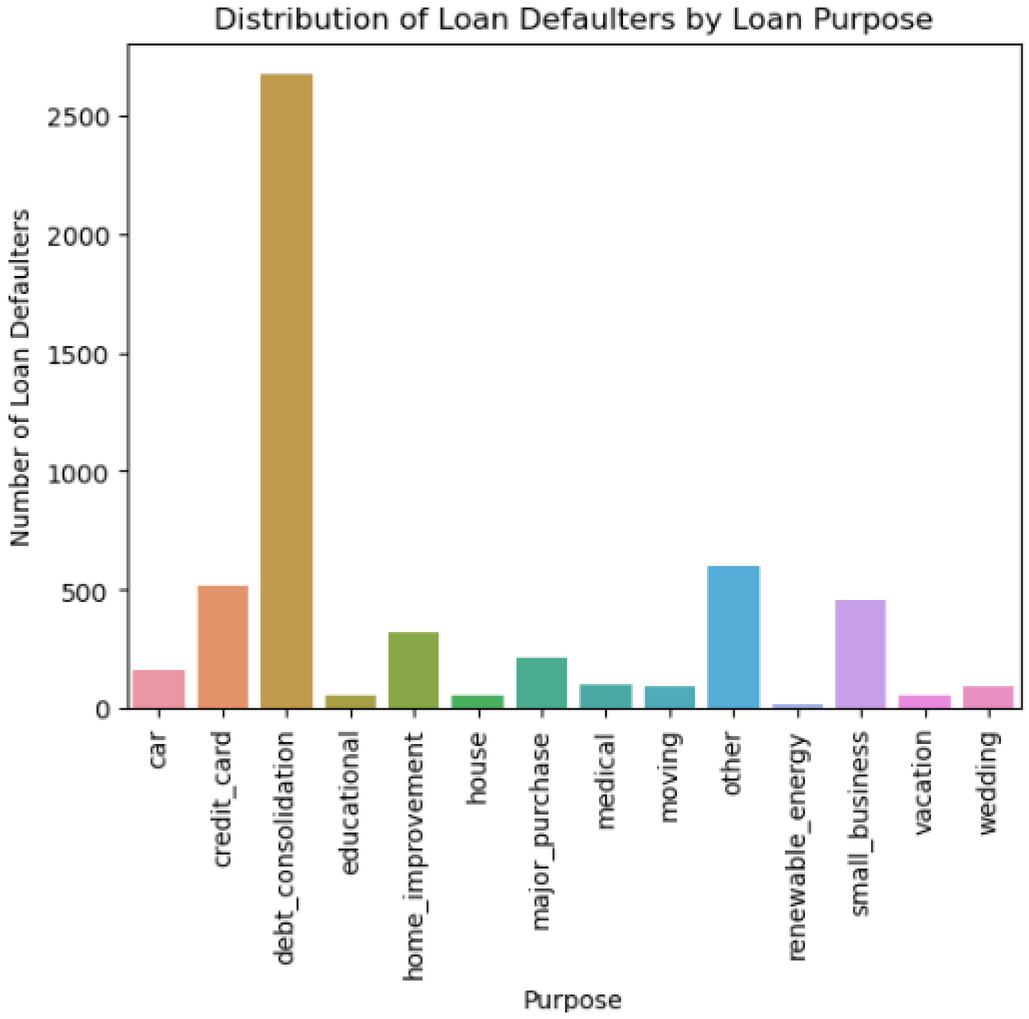
Bivariate Analysis(continued...)

- Observation:

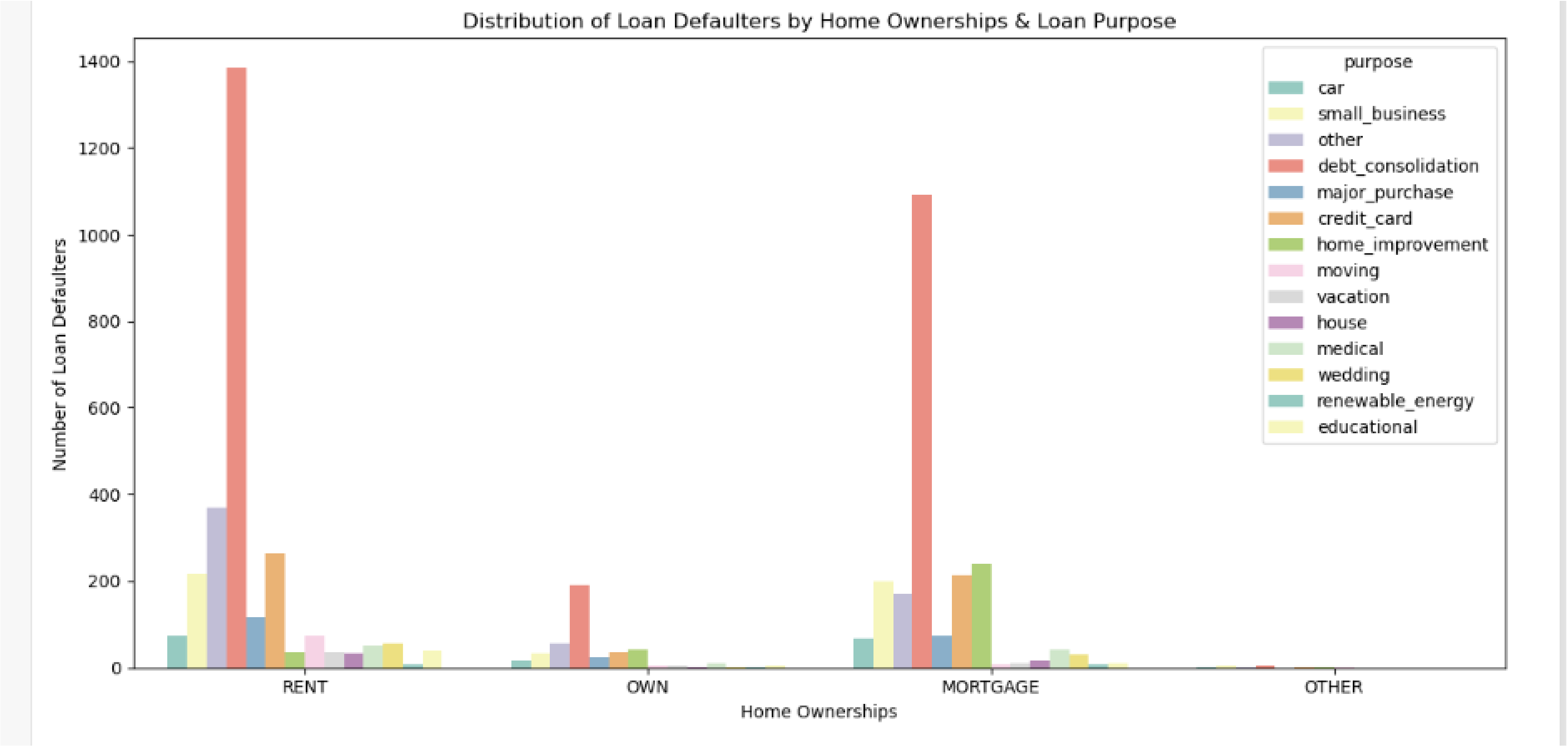
1. Applicants with higher salary mostly applied loans for "home_improvment", "house", "renewable_energy" and "small_business
2. The 60 months term loans have more interest rate.
3. here are more defaulters in both 36, 60 month terms because of high interest rates.
4. Almost in all categories of purpose, defaulter's DTI is high than fully paid borrowers

Analyzing pattern for Loan defaulters

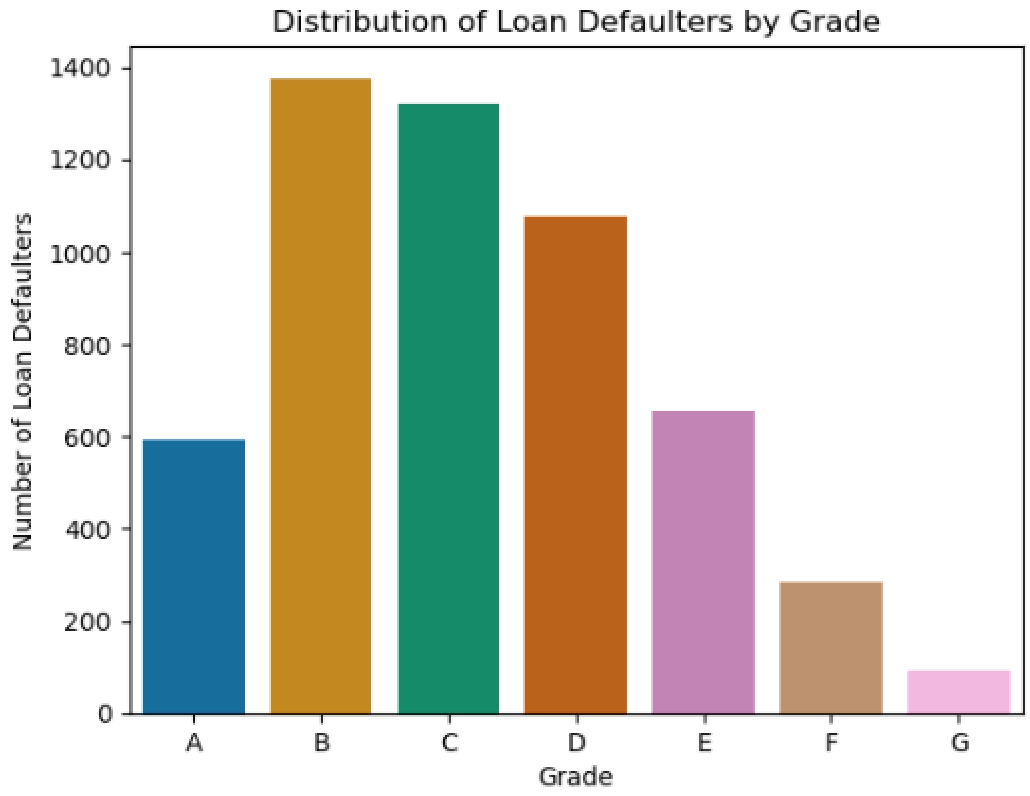




Loan defaulters(Continued...)



Loan defaulters(Continued...)



Loan defaulters(Continued...)

- Analysis and pattern behaviour of loan Defaulters :
- 1. It shows there are more defaulters in RENT and MORTGAGE. let's check it in granular level.
- 2.. From RENT category, there are more defaulters from 'debt_consolidation','other', 'credit_card' and 'small_business’.
- 3.. From MORTGAGE category, there are more defaulters from 'debt_consolidation','home_improvement', 'credit_card' and 'small_business’.
- 4. Overall, one should be carefull with 'debt_consolidation', 'credit_card' and 'small_business' loans when the borrowers dont have own house.
- 5. It shows there are more defaulters in B,C and D grades.
- 6. Grades F,G(more intereste rate grades) are having less defaulters which is a good indicator.
- 7. From all grades, there are more defaulters from 'debt_consolidation', 'others', 'credit_card' and 'small_business' purpose loans.