LENDING CLUB CASE STUDY

(Using Exploratory Data Analysis)

Group:

Arunkumar P
Manojkumar Pandey

Agenda

1. Introduction
2. Problem Statement
3. Data Understanding
4. Data Cleaning and Manipulating
5. Data analysis
6. Presentation and Recommendations

Introduction

- Lending club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
- lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss).
- Main Intention of this assignment is to identify the risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss
- We will be using the Exploratory Data analysis to identify the driving factors behind loan default and other factors which are strong indicators of default.

Problem Statement

The data given for assignment contains the information about past loan applicants and whether they ‘defaulted’ or not. Borrowers who default cause the largest amount of loss to the lenders. Main Intention of this assignment is to identify the risky loan applicants and identify the driving factors behind loan default and other factors which are strong indicators of default.

Data Understanding

Dataset has 39717 records with 111 different variable

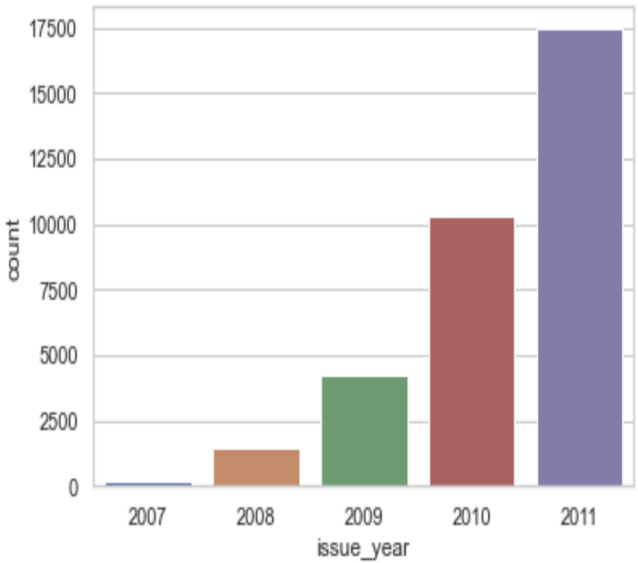
Data has been collected between 2007 to 2011

Dataset has other variables like, loan_ammount, intrest_rate, DTI, Purpose, Employment details and so on.

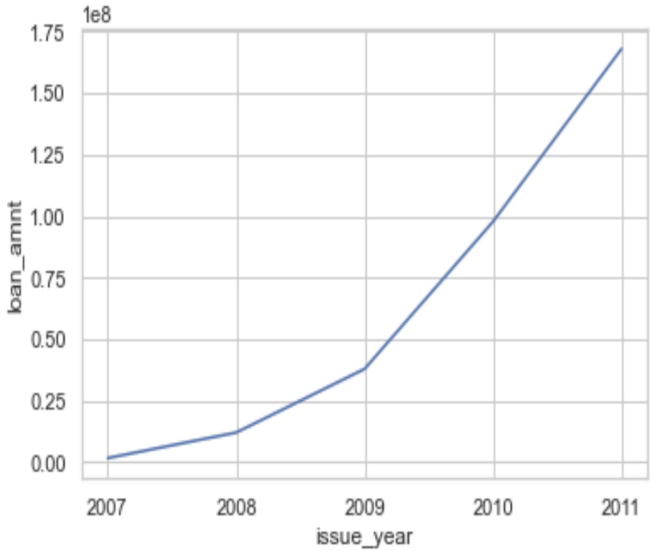
loan_status has been categorised into 3 different category ("Fully Paid","Charged Off", "Current")

Data also has alot to unwanted Variables and 57 columns with >50% of Missing data

Growth of Number of Loans sanctioned

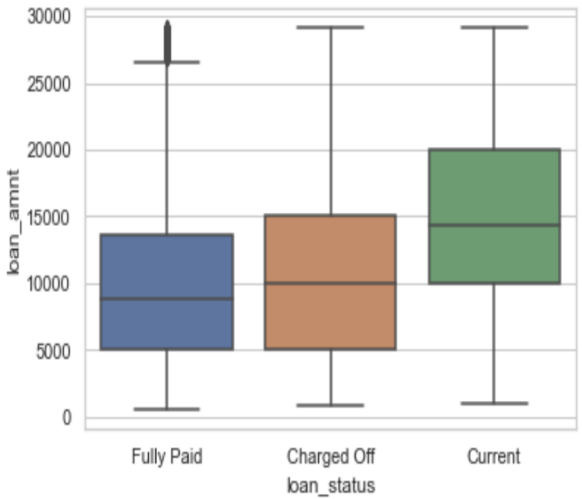


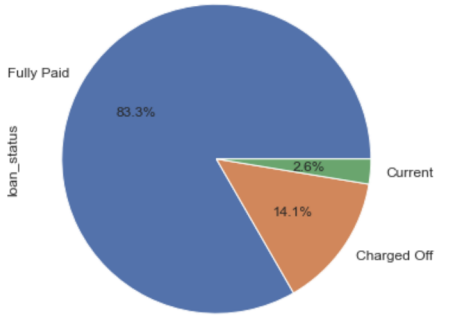
Growth of Loan amount Issued

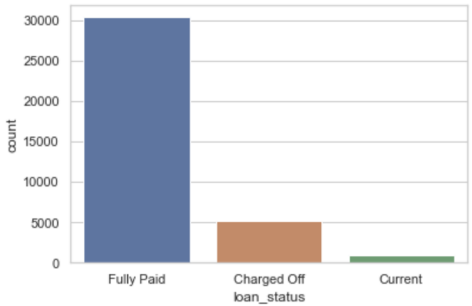


Observation:
Number of issued Loan has increased over past years
Loan ammount also has signficantly increased over the years

Loan Status by volume and percentage







Data Cleaning and Manipulating

- As part of data cleaning removed 57 columns with >50% of Missing data
- Removed unwanted columns by intuition
- Removed unwanted prefixes and suffixes in the data and converted to appropriate datatype
- Manipulated the columns which has object colums with string value to appropriate type (int, float)
- Filled missing values with median values
- identified and removed outliers

Heatmap to Represent the Missing Value

Data manipulate

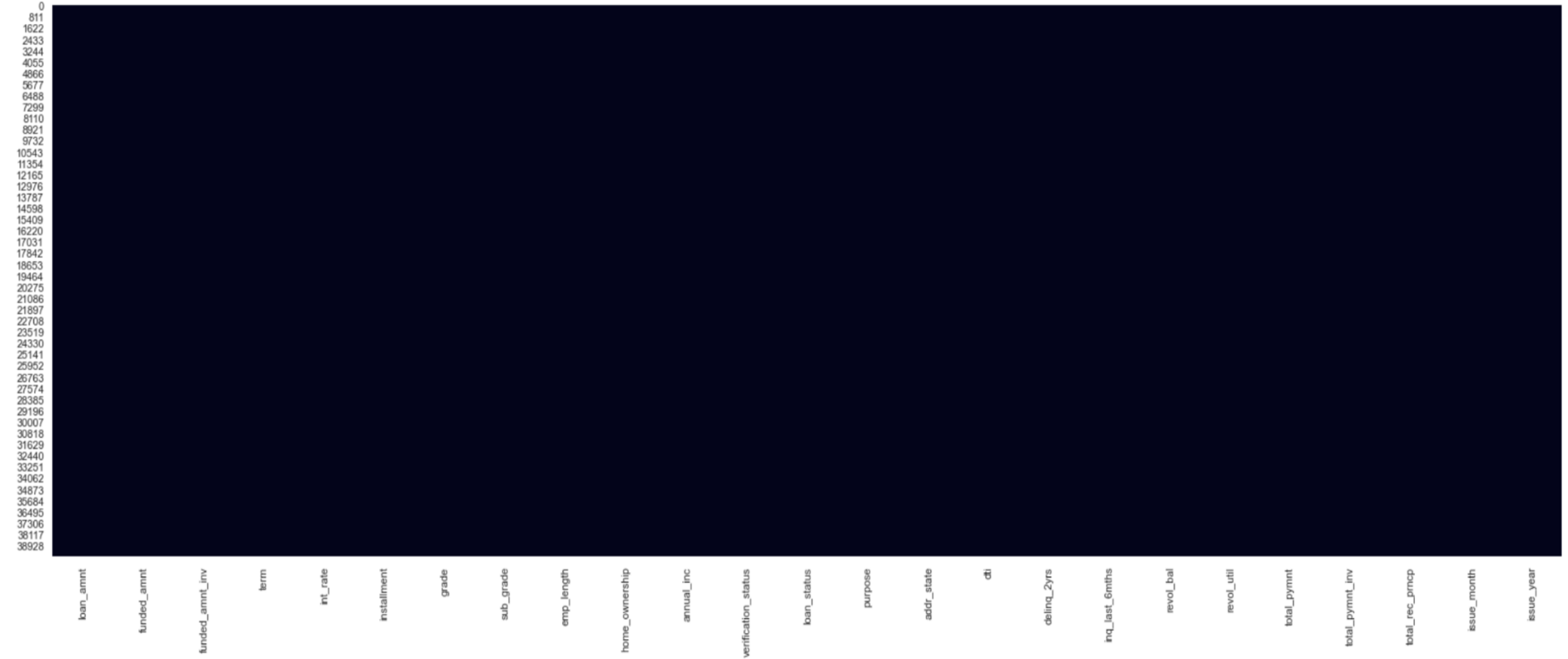
- ❏ Removing strings 'years' and '<' and '+' from emp_length
- ❏ Remove '%' symbol from int_rate and revol_util
- ❏ Fill missing values with median value for revol_util
- ❏ Remove string ' months' from term

```
         term int_rate grade sub_grade emp_length home_ownership  \
0   36 months   10.65%     B        B2  10+ years           RENT
1   60 months   15.27%     C        C4   < 1 year           RENT
2   36 months   15.96%     C        C5  10+ years           RENT
3   36 months   13.49%     C        C1  10+ years           RENT
4   60 months   12.69%     B        B5     1 year           RENT

  verification_status issue_d  loan_status         purpose addr_state  \
0            Verified  Dec-11   Fully Paid     credit_card         AZ
1     Source Verified  Dec-11  Charged Off             car         GA
2        Not Verified  Dec-11   Fully Paid  small_business         IL
3     Source Verified  Dec-11   Fully Paid           other         CA
4     Source Verified  Dec-11      Current           other         OR

  revol_util
0     83.70%
1      9.40%
2     98.50%
3        21%
4     53.90%
```

After Cleaning Dataset



Univariate Data Analysis

Univariate analysis done based on the Variable is categorical or continuous

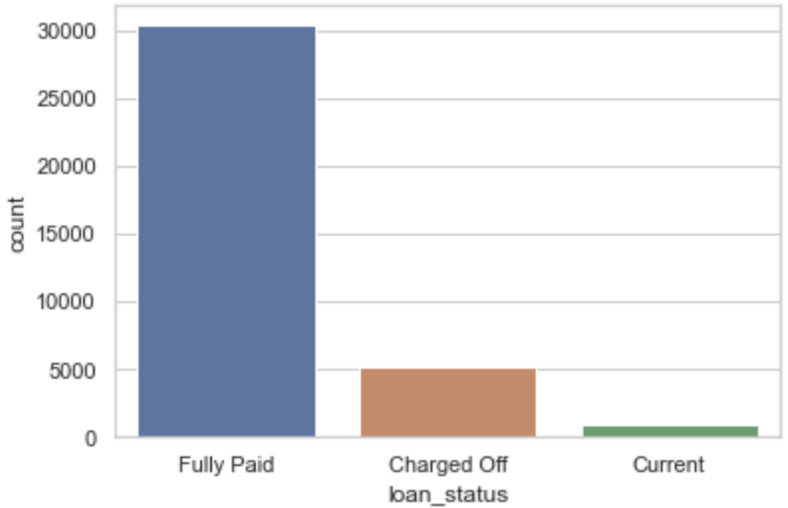
Data is divided into following for the analysis:

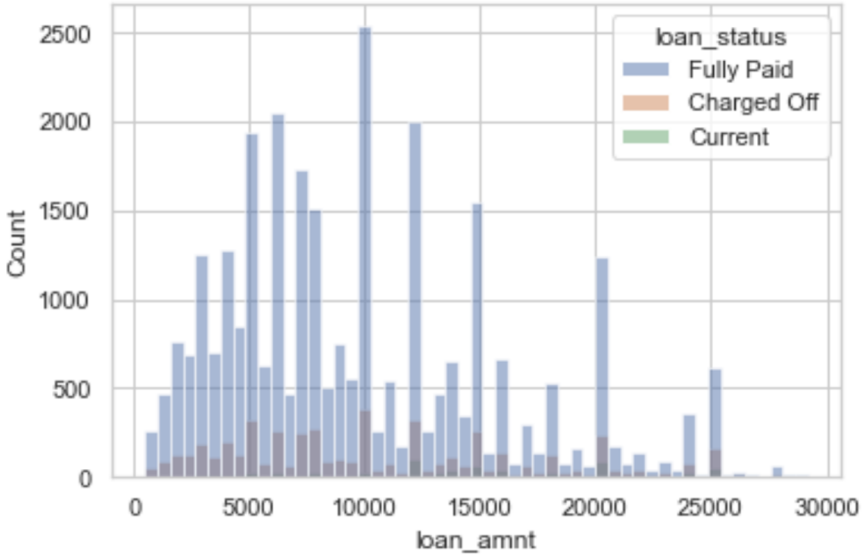
For categorical variable we have used countplot with percentage for every category

For Continuous Variable we have used Box and Histogram to understand the Distribution, min,max, std-deviation, mean values

Categorical	Continuous	
Loan_status Grade subgrade Home_ownership Verification_status perpose issue_year term emp_length	int_rate annual_inc revol_bal revol_util loan_amount dti	Some of the Continuous variables are converted to categorical variable for the analysis Example: Annual_inc is divided into 5 different income category from low,lower-mid,mid,higher-mid,high

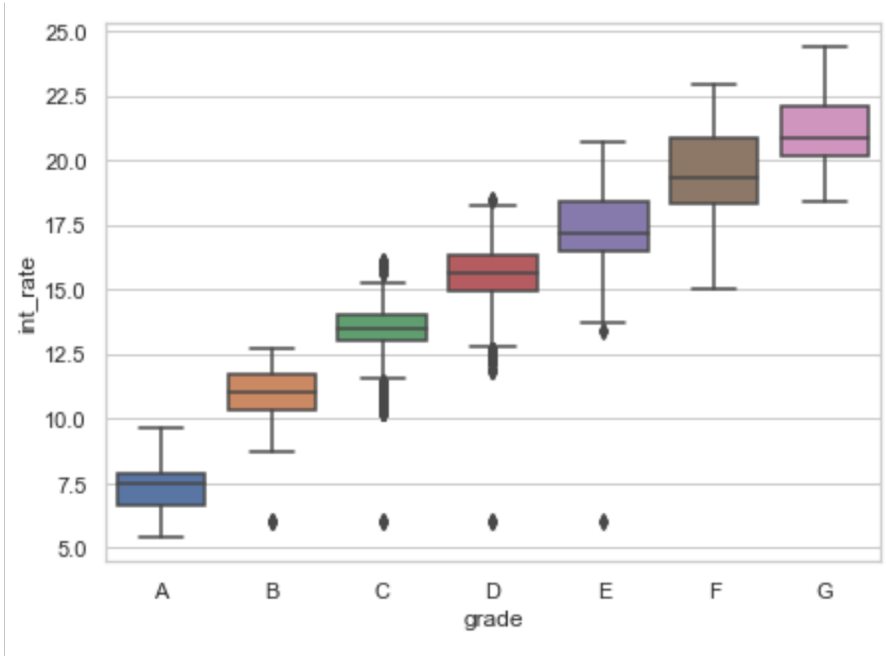
Univariate Data Analysis of Loan_status and Loan_amount

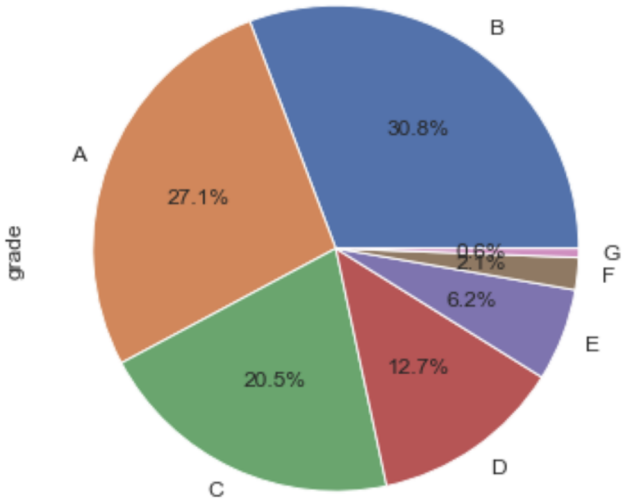




Observation:
Majority of the loan are Fully paid.

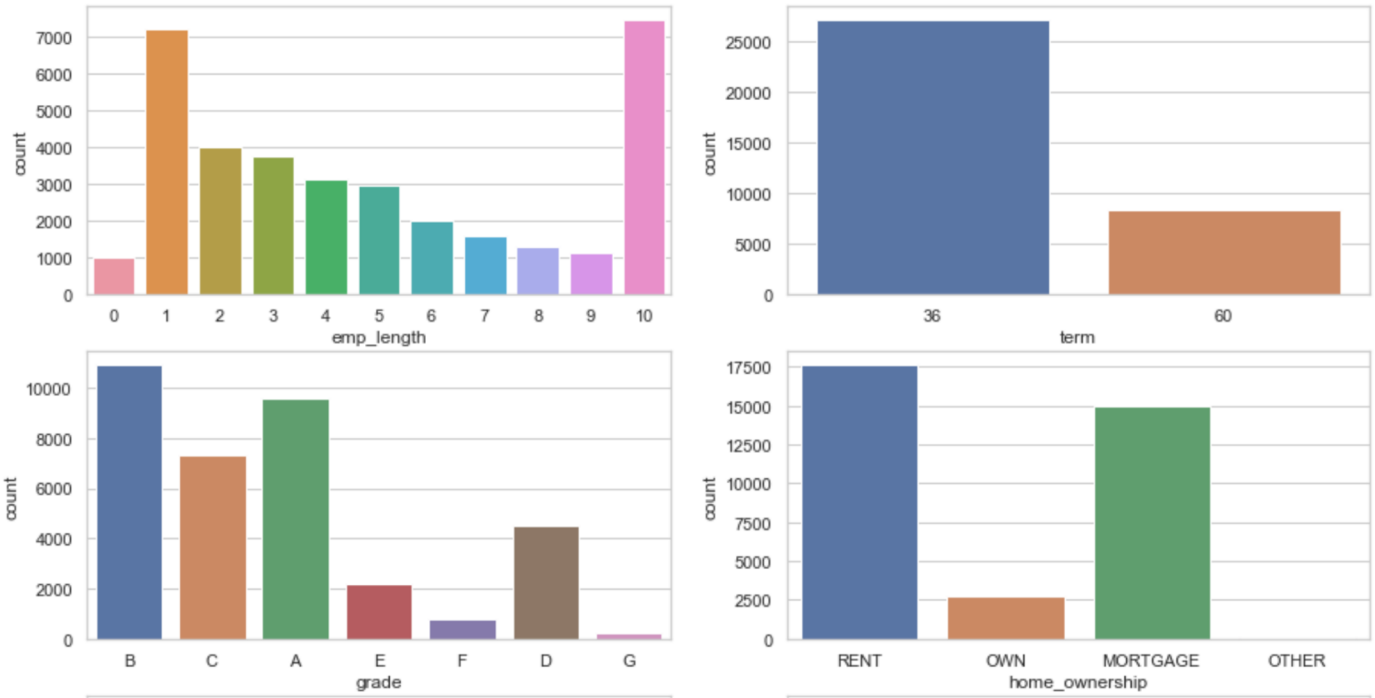
Grade categorised based on Interest rate and majority





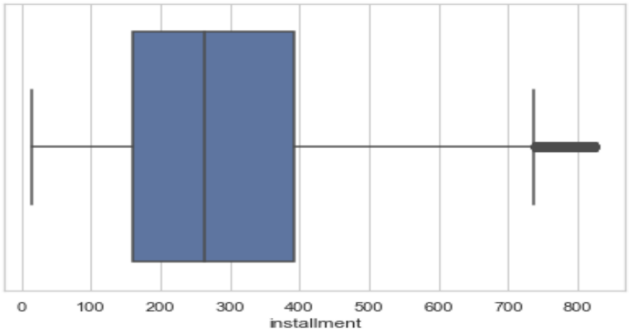
Observation:
Grade is categorised into 7 grade from A-to-G based on the int_rate.. grade 'A' being lowest int_rate
Most of the loans are sanctioned with low grade A,B,C means lower interest rate.

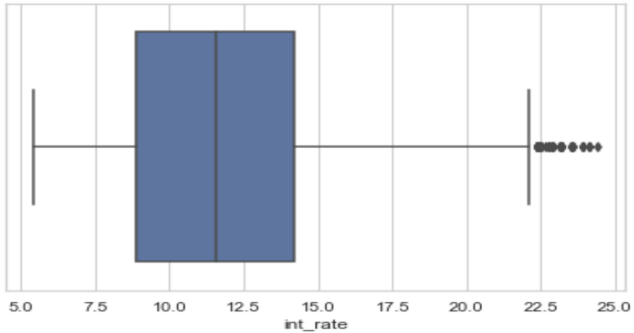
UDA of Emp_length, Term, Grade, House_ownership

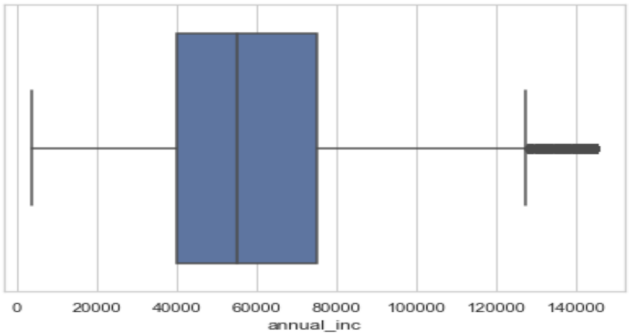


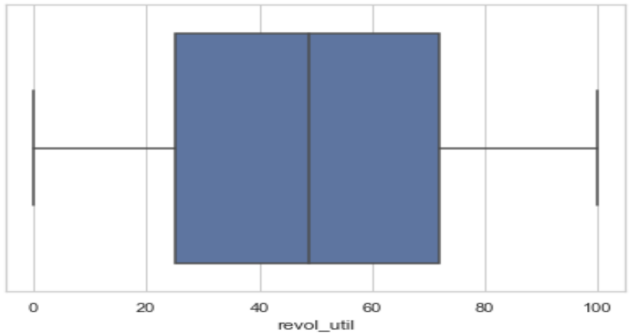
Observation:
Majority of borrowers has more than 10+ years of experience or < 1 year of experience
Majority of borrowers has taken 36 months of term
Most of the loan are sanctioned for lower loan_grade (A being the lowest grade)
Most of borrowers either has Rented or Mortgaged

UDA of installment, int rate, annual income, revol util



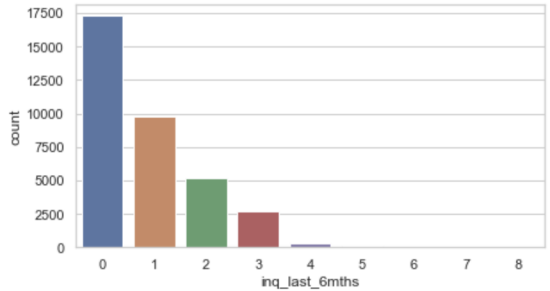


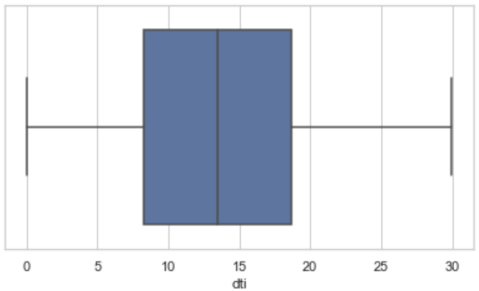


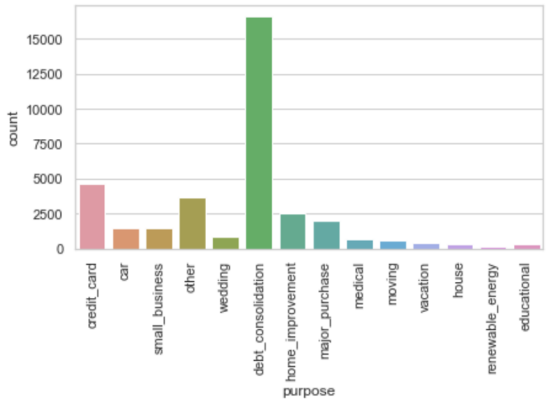


Observation:
Installment has mean of 292.19
int_rate has mean of 11.75
annul_inc has mean of 60027
revol_util has mean of 48

UDA of inq_last_6months, Verification_status, dti, purpose







Observation:
Very few borrowers inquired in last 6 months
Majority of borrowers are not verified
dit has mean of 13.4
Most people borrowed the loan for debt_consolidation, credit_card, home_improvemnt

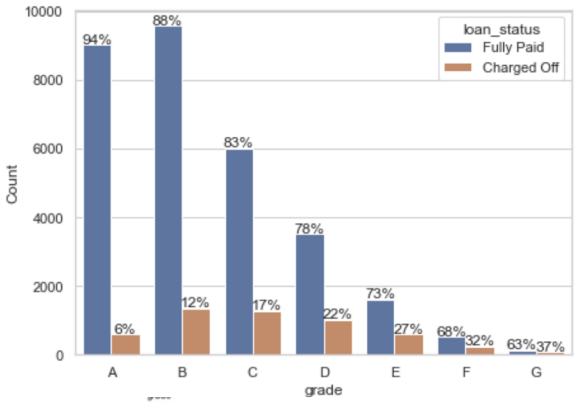
Bivariate Analysis

For understanding the relation between two variables we are calculating the percentage as well per category wise

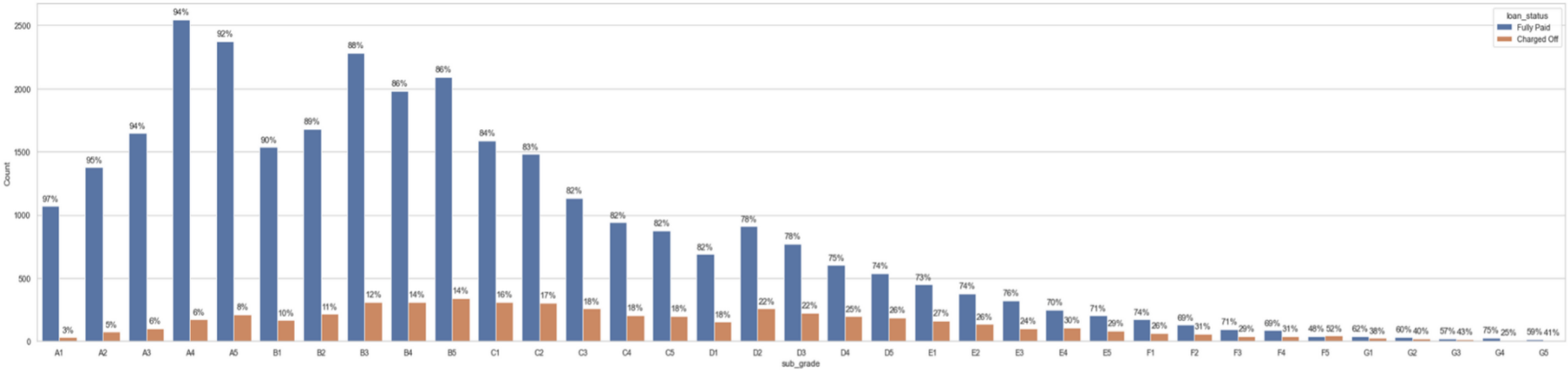
Example: we have two category of loan_status "Fully_Paid" and "Charged_off" (we have dropped "Current" loan_status as they do not have significance)

We also have 7 difference category of grade from A-Z, To understand the relation of each Grade with loan_status, we are computing the percentage of Paid and non-Paid for each Grade. This would help in understadning the behavior of people of every grade.

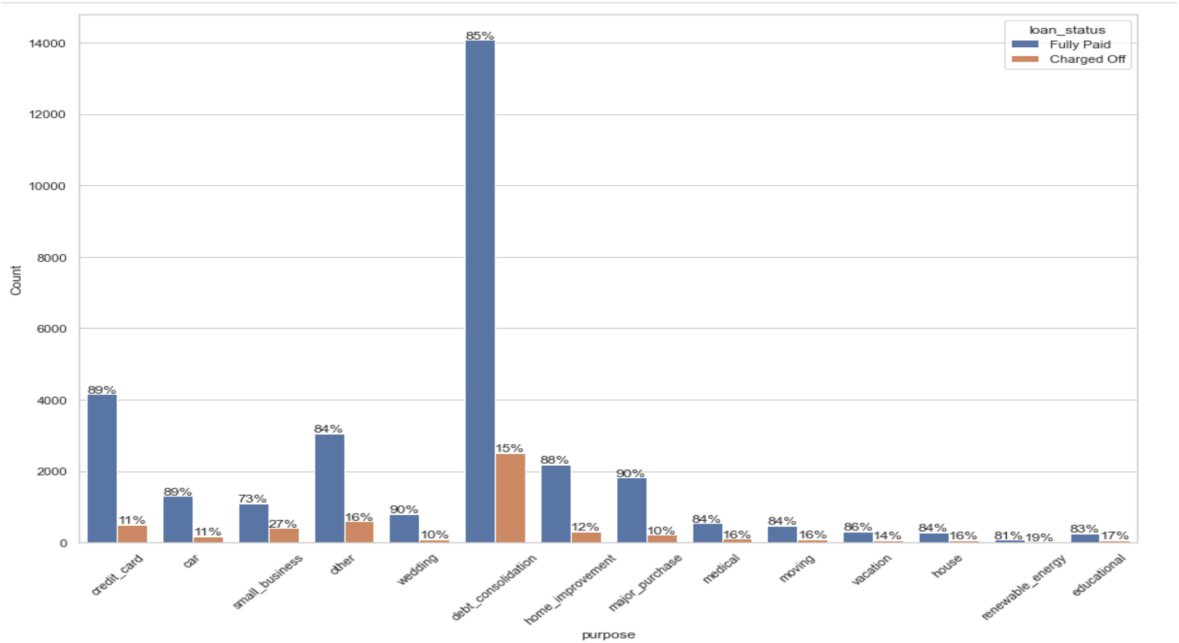
Grade, Subgrade relation to Loan_status



Observation:
Grade-wise percentage is calculated, as we can see in the countplot,
Ratio or percentage of defaulter is increased with the higher grade.
(same observation is seen in subgrade as well)

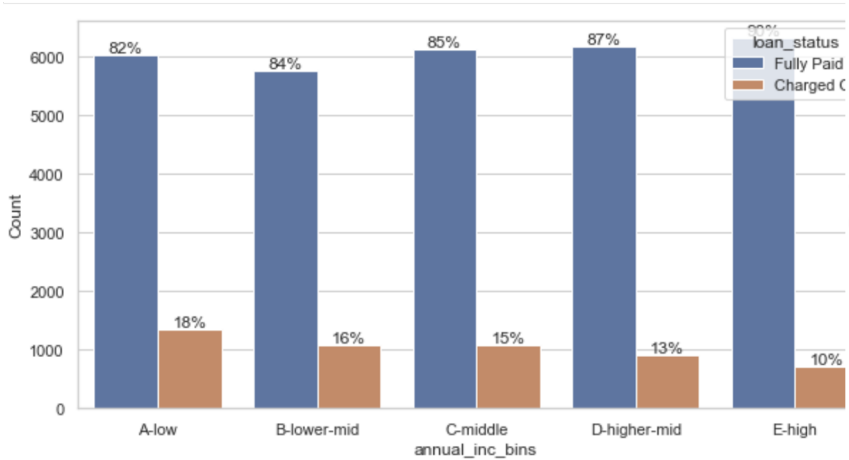


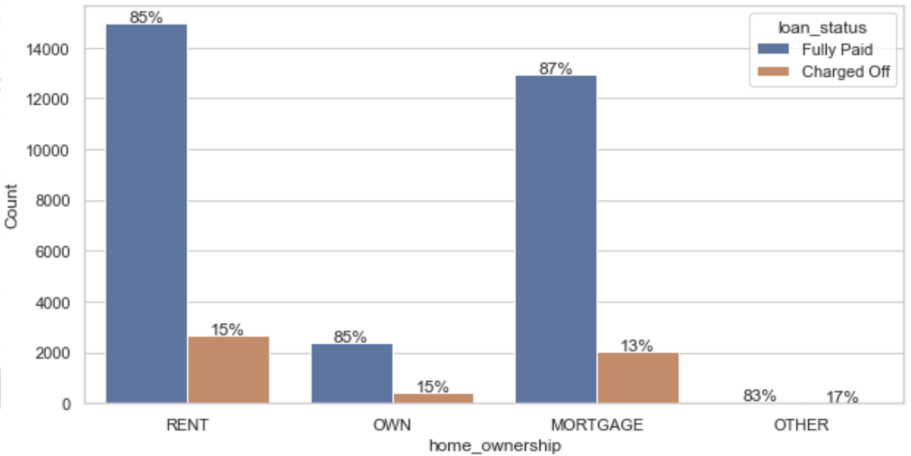
Relation of Purpose to Loan_status



Observation:
Higher percentage of defaulter seen in "small_business (27%)", renewable_energy(19%), Educational(%17%)

Relation of Annual_income and House_ownership to Loan_status



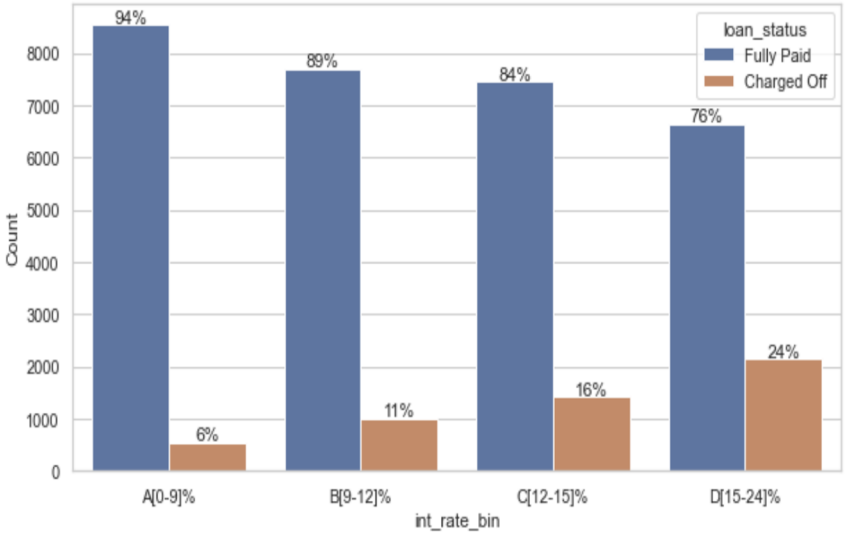


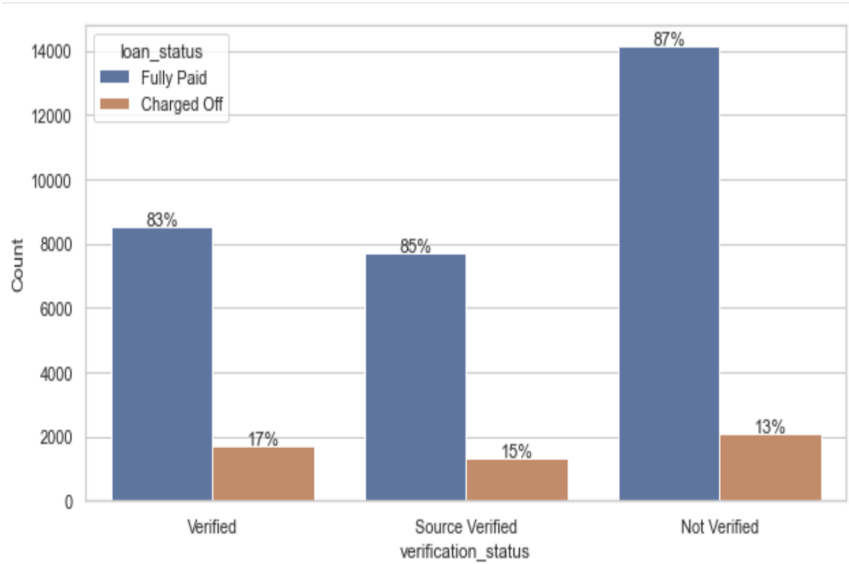
Observation:

Higher percentage of defaulter are more likely in the lower income. as the chart indicate the percentage of defaulter increases with lower income range

Higher percentage of re-paid customers are those who have mortgaged owener-ship

Relation of int_rate and Verification_status to Loan_status



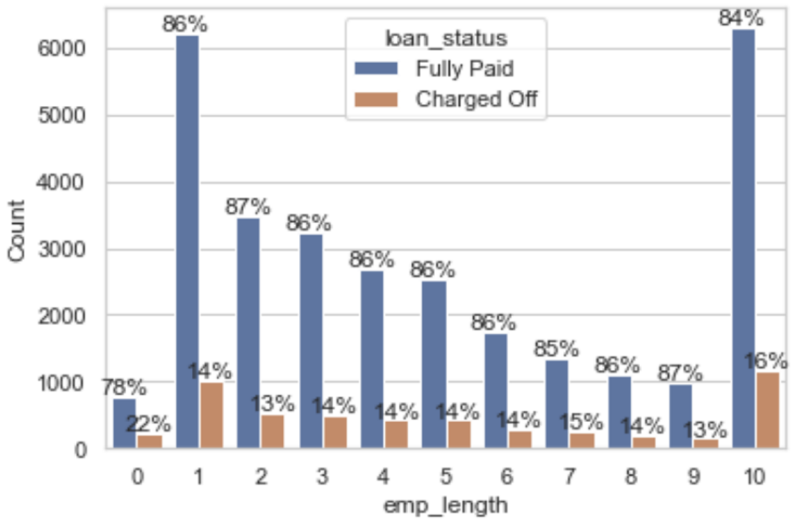


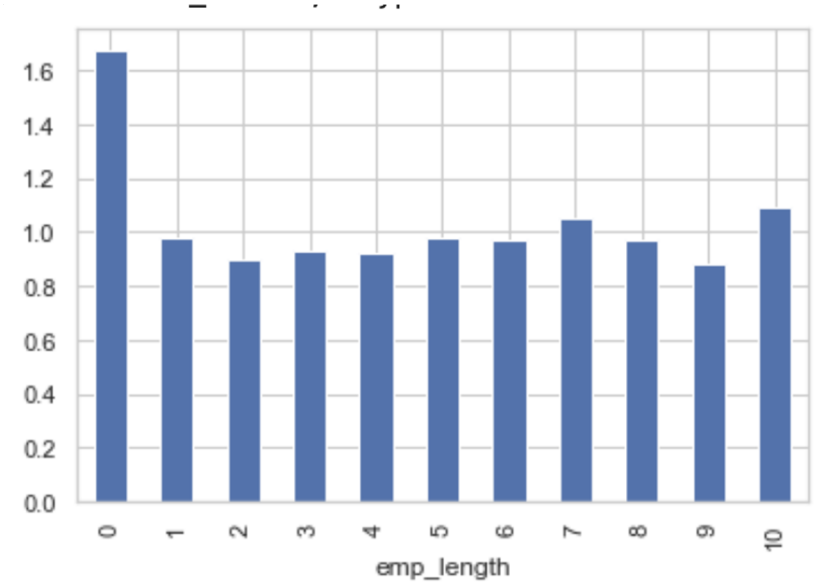
Observation:

The defaulter percentage is higher with high rate of interest

The defaulter percentage Verified borrowers

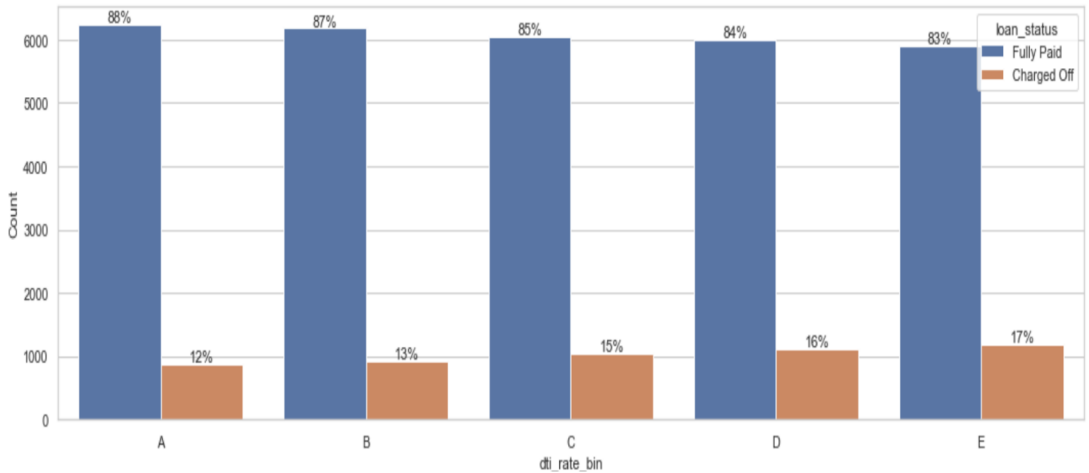
Relation of Emp_length to Loan_status





Observation:
Defaulter percentage is almost same across all Emp_length and does not indicate any specific trend
The plot on the right indicate the ratio to "Fully Paid" to "Charged Off", which indicate almost all has same ratio, There is not significant difference.

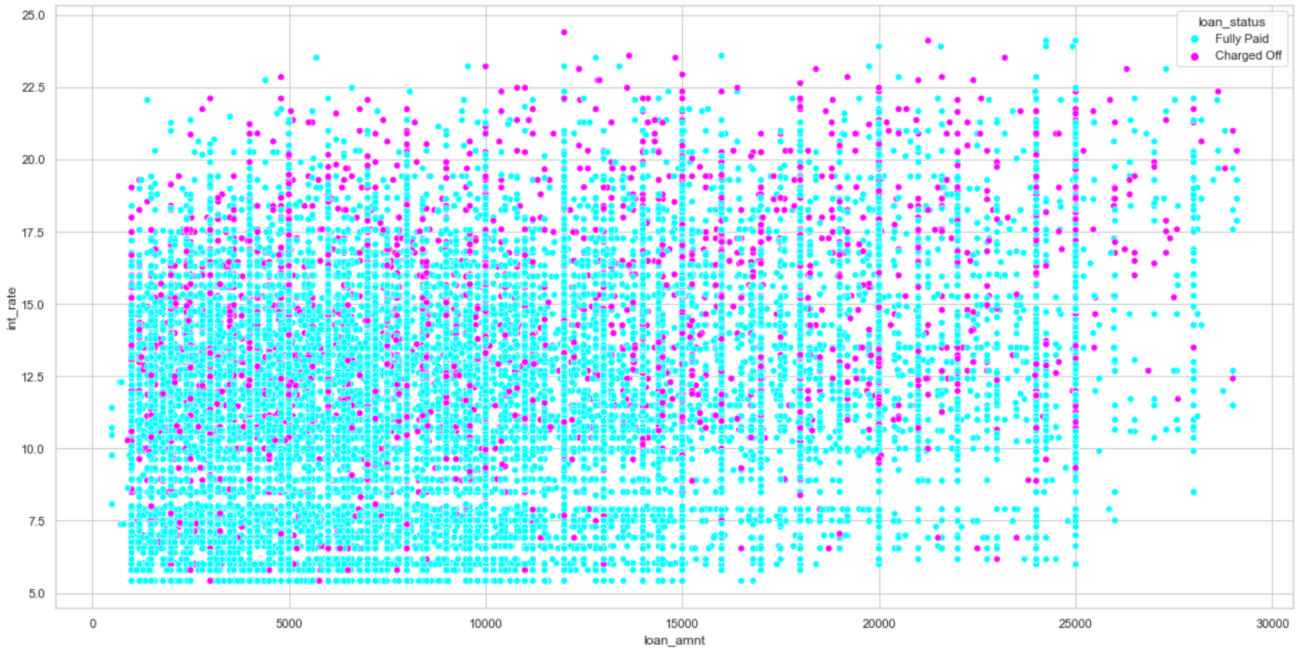
Relation of DTI to Loan_status



Observation:

Higher percentage of defaulter are more likely to have hight dti percentage, as the chart indicate the percetage of defautler increases with Higher dti percentage.

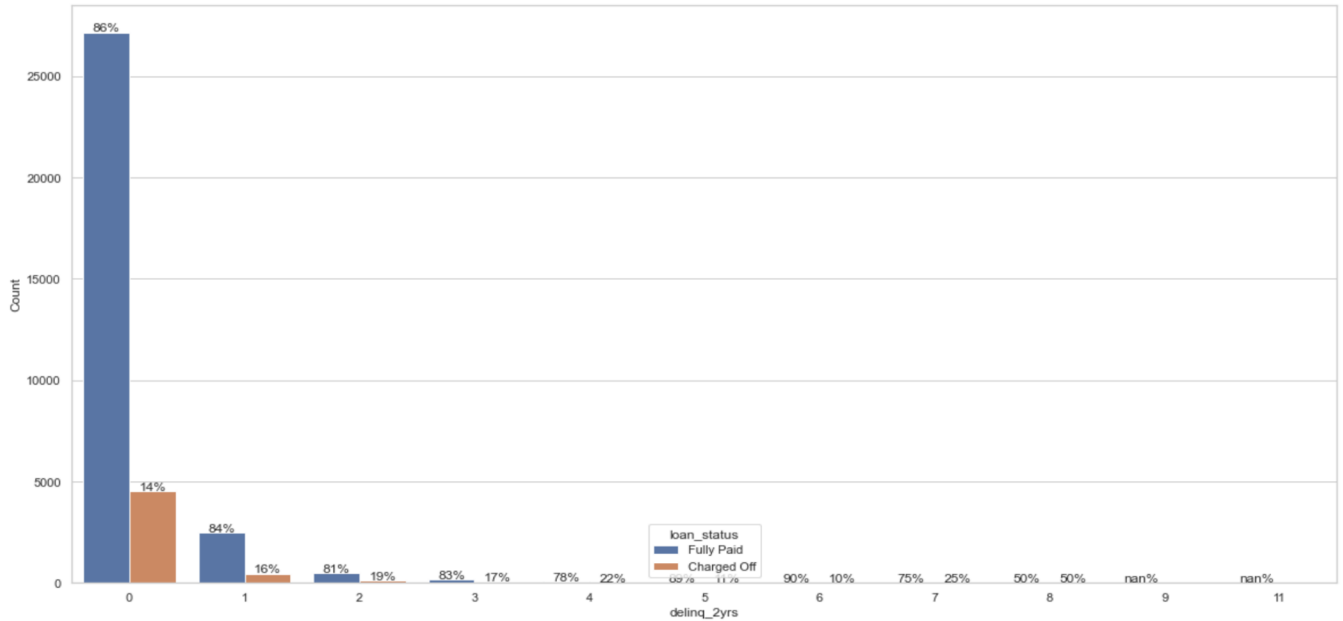
Distibution of of Loan_ammount and int_rate w.r.t Loan_status



Observation:

The Majority of defaulter is seen with high rate of interest

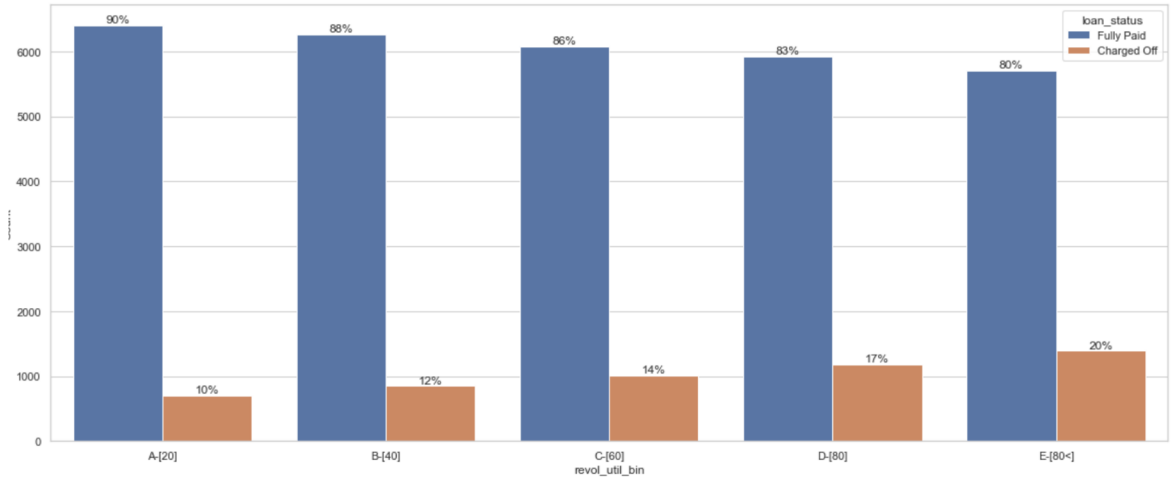
Relation of Delinq_2yrs to Loan_status



Observation:

Percentage of defaulter increasing with value of delinq_2yrs increase

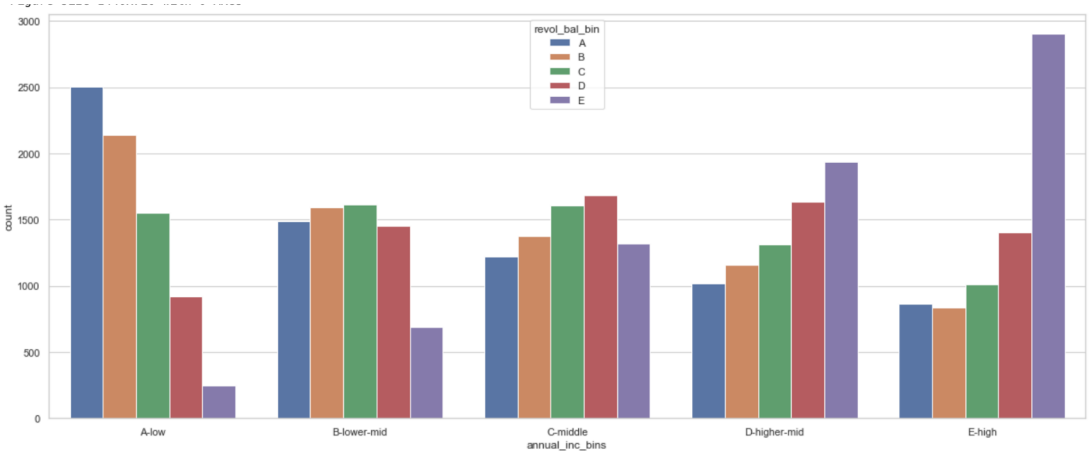
Relation of Revol_util to Loan_status



Observation:

Percentage of defaulter increasing with higher revol_util value

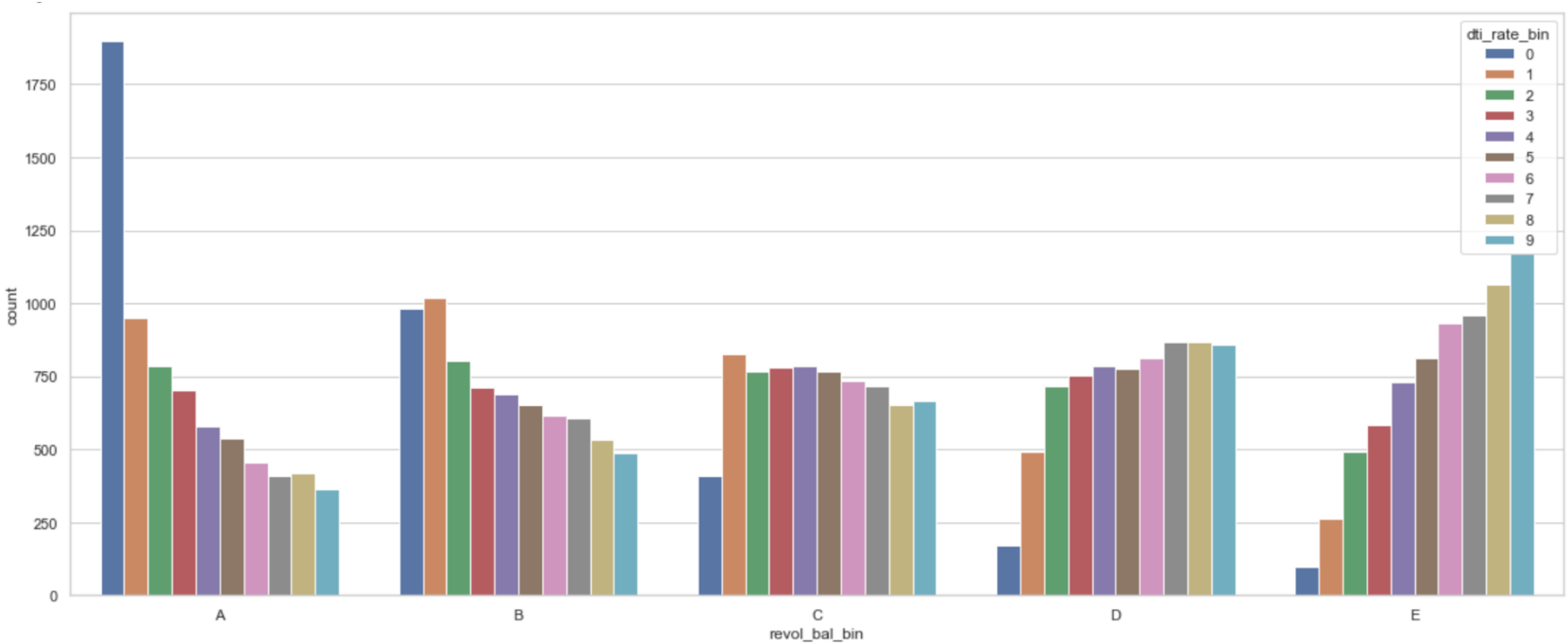
Multivariate Analysis: Relation between Revol_bal, Annual_income and loan_status



Observation:

the revol_bal is increasing as the annual_income is increased

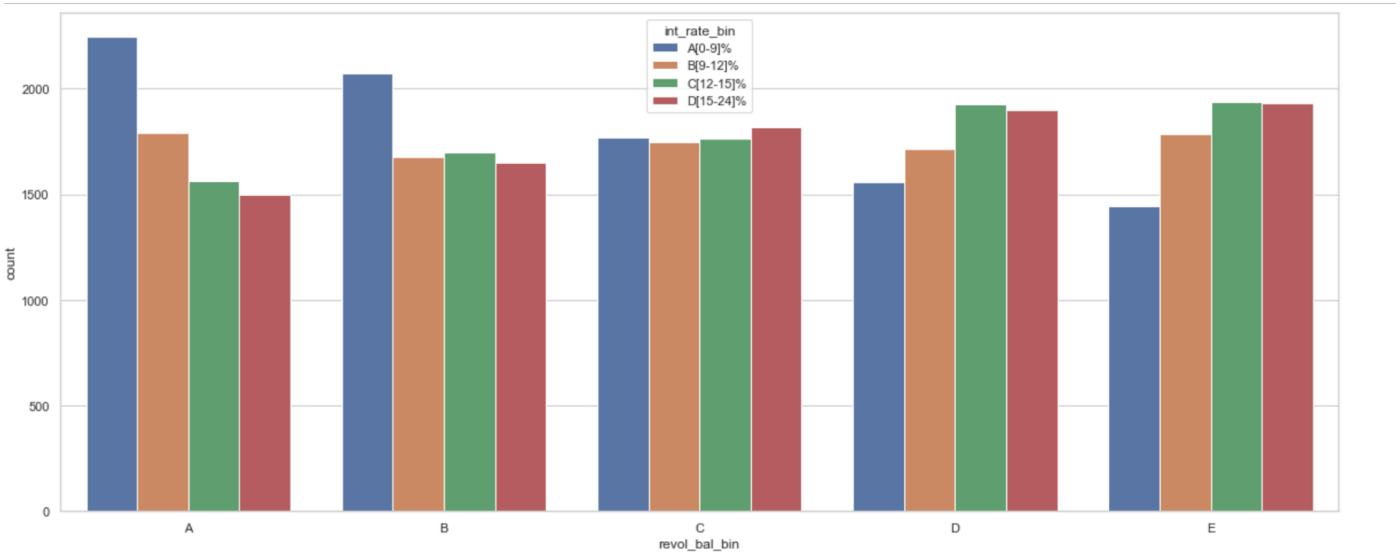
Multivariate Analysis: Relation between dti, revol_bal and loan_status



Observation:

the revol_bal is increases as the dti vlaue is increased

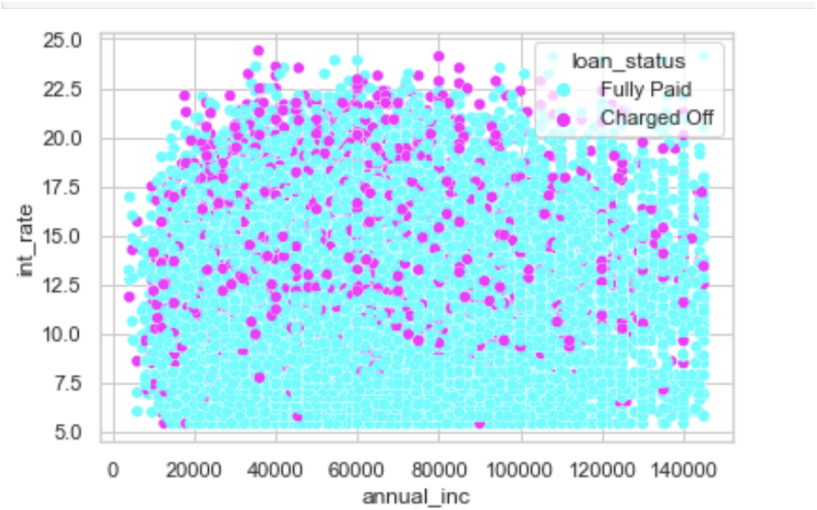
Multivariate Analysis: Relation between dti, revol_bal and loan_status



Observation:

the revol_bal is increases as the Int_rate is increased

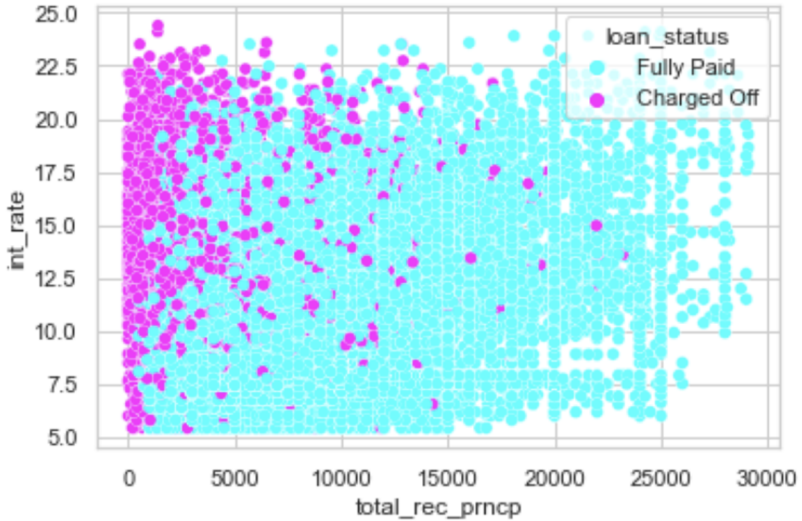
Multivariate Analysis: Relation between dti, revol_bal and loan_status



Observation:

Most defaulter as seen on Higher intrest_rate

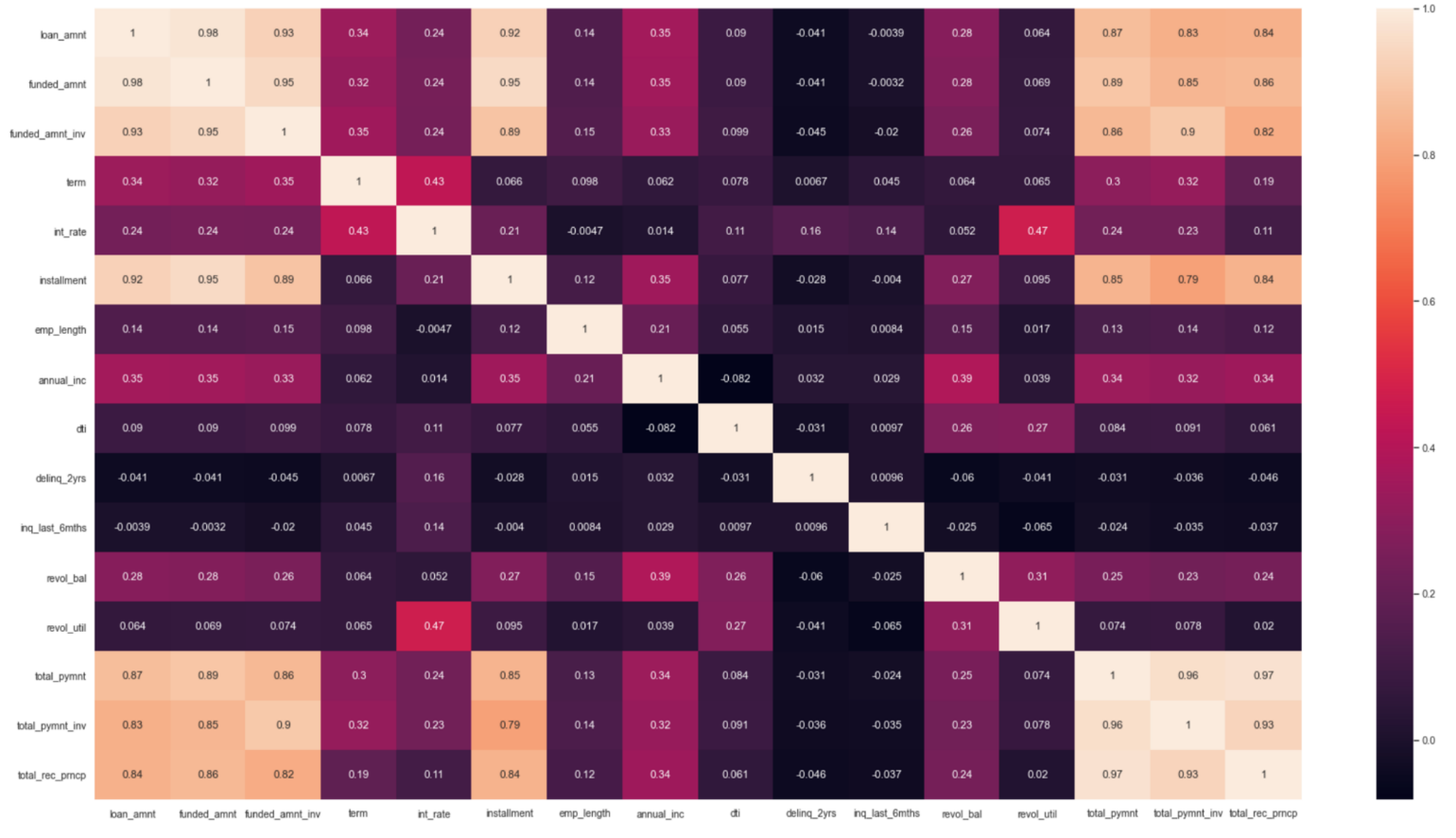
Multivariate Analysis: Relation between dti, revol_bal and loan_status



Observation:

Most defaulter as seen on Higher intrest_rate and Most defaulter are defaulting in the initial repayment itself (higher number of people who have defaulted paid 5000 or less)

Correlation matrix for Multivariate Analysis



Outcome of Correlation matrix analysis

High correlation is observer with the following variables:

Annual_inc (very good indicator for loan eligibility):

positive correlation with:

- loan_amount, total_paymt, total_rec_prncp, installment variable

negative correlation with:

- int_rate, revol_bal, revol_util, dti

int_rate has positive correlation with term, revol_bal, revol_util

Recommendation

- Most people borrowed loan amount between rage of 5000-15000
- People with high DTI and revol_bal are most likely to default. Not recommended to approve the loan for those who has DTI high
- People with low annual income are also more likely to default. also Majority of loan borrower are from lower income category, not approving the loan will also be loss for the company, recommended to consider other parameter as well revol_util and DTI for lower income borrower.
- People borrow loan for the following reason,
- > debt_consolidation, small_business , renewable_energy Educational
- recommended to approve loan If annual_income is more than average or (higher-mid, high)
- Observed Defaulter are high with Higher rate of interest across all income range. Recommended to approve the loan with high interest rate for people who belongs to =annual_income category higher-mid/high