Lending Club Case Study

SUBMISSION

Name:

Syed Inayathullah
Mohamed Sheik Asan Aliyar

Business Understanding

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Help company make a decision for loan approval by deriving insights from the given data. Reduce Credit loss.

PROBLEM STATEMENT

Borrowers can easily access lower interest rate loans through a fast online interface.

Lending loans to risky applicants is the largest source of credit loss

Two types of risks are associated with the bank's decision:

- **Loss of business** - If the applicant is likely to repay the loan, then not approving the loan
- **Financial loss** - If the applicant is not likely to repay the loan, then approving the loan

CASE STUDY OBJECTIVES

Identify risky applicants, defaulters or Charged off customers

- Analyze Historical data and Understand data patterns
- Fully paid, current and charged-off are three possible loan status scenarios

Identify driving factors/variables using EDA

- Understand consumer and loan attributes that are strong indicators for default
- Explore EDA mechanisms for univariate, segmented univariate and bivariate analysis

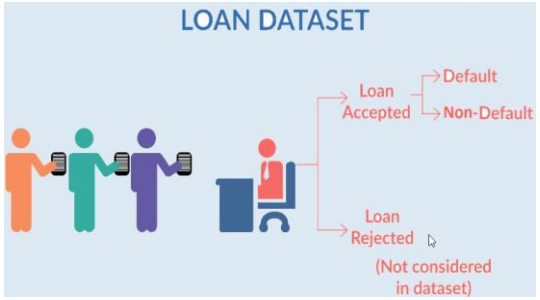
Derive actionable insights - Portfolio and Risk assessment

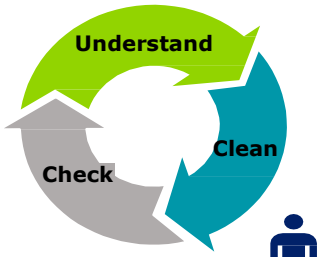
- Deny the loan
- Reduce Amount of loan
- Lending at higher interest rates

Data Analysis Strategy

Investigate data to discover patterns, spot anomalies, check assumptions and derive inferences

Data Understanding and Manipulation





1 Data Sourcing

- Data Extraction
- Data Quality check

- *Do cleaning on first look of data*
- *Remove what is obvious*

2 Data Cleansing

- Remove cols : single type, 0s, NAs
- Clean numeric cols. Remove text like %, months

3 Data Understanding

- Understand Data Dictionary and functional context
- Impute, derive cols
- Remove outliers

- *Analyze data - Iterative process Clean and understand*
- *Take aid from Data dictionary*

Data Analysis and Recommendation

4 Univariate Segmented Analysis

- Data Analysis on single variable
- Describe data and find patterns
- Identify important driver variables

- *Use median, mode, box plots etc.*

5 Bivariate Analysis

- Find relationship between driver variables
- Chose variables that make business sense

- *Appropriate plots to present results*

6 Probable loan defaulters

- Use these inferences to predict.

- *Who are the loan defaulters?*

**Not in scope of case study.*

Driver focus areas ..... **Where this evolves to**

Data Cleansing

Single File, Total Rows : 39717, Total Cols: 111

Drop Cols with all NaN or 0

- Drop cols with only NaN, only one unique record or all the records are unique.
- Removing columns 'tax_liens','chargeoff_within_12_mths' & 'collections_12_mths_ex_med' that has only 0 and NaN values since it is not required in our analysis

Drop other irrelevant cols

- Remove zip_code as we have state_addr
- Remove url, member id, id that are unique for all records
- Remove descriptive cols like emp_title, title, desc
- Remove columns 'delinq_amnt','acc_now_delinq','application_type','policy_code','pymnt_plan', 'initial_list_status' that has only one value thus it is not required in our analysis
- Remove columns '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' which informs about the future and end of loan repayment and thus not required for default loan analysis
- Remove 'mths_since_last_delinq','mths_since_last_record','issue_d','funded_amnt_inv','next_pymnt_d', last_credit_pull_d', 'earliest_cr_line,'inq_last_6mths' columns since it is not required for default loan analysis.

Clean Numeric Cols

- Remove % from from int_rate
- Remove months from term
- Remove symbols > < >= from emp_length
- Remove outliers (> 99%) in annual_inc

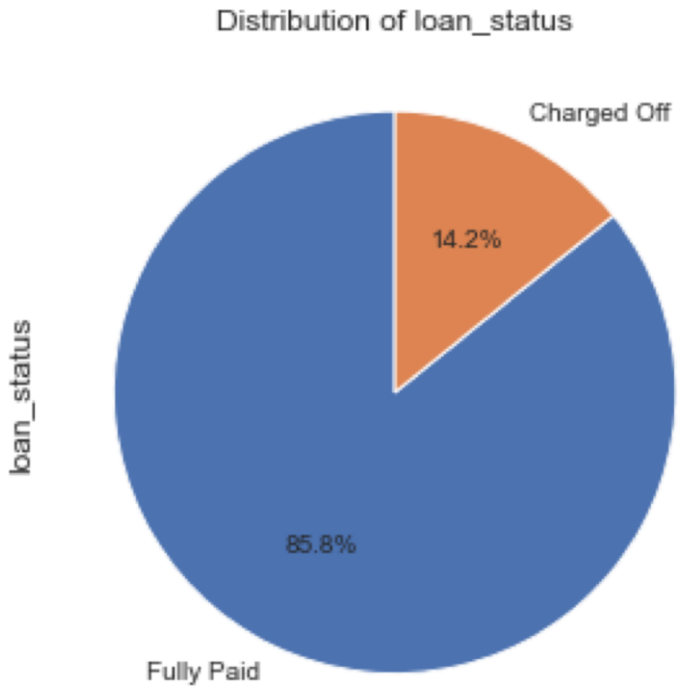
Add Derived Cols

- avail_credit_lines (total_acc – open_acc)
- income_to_funded (annual_inc/funded_amnt)

Distribution of loan status

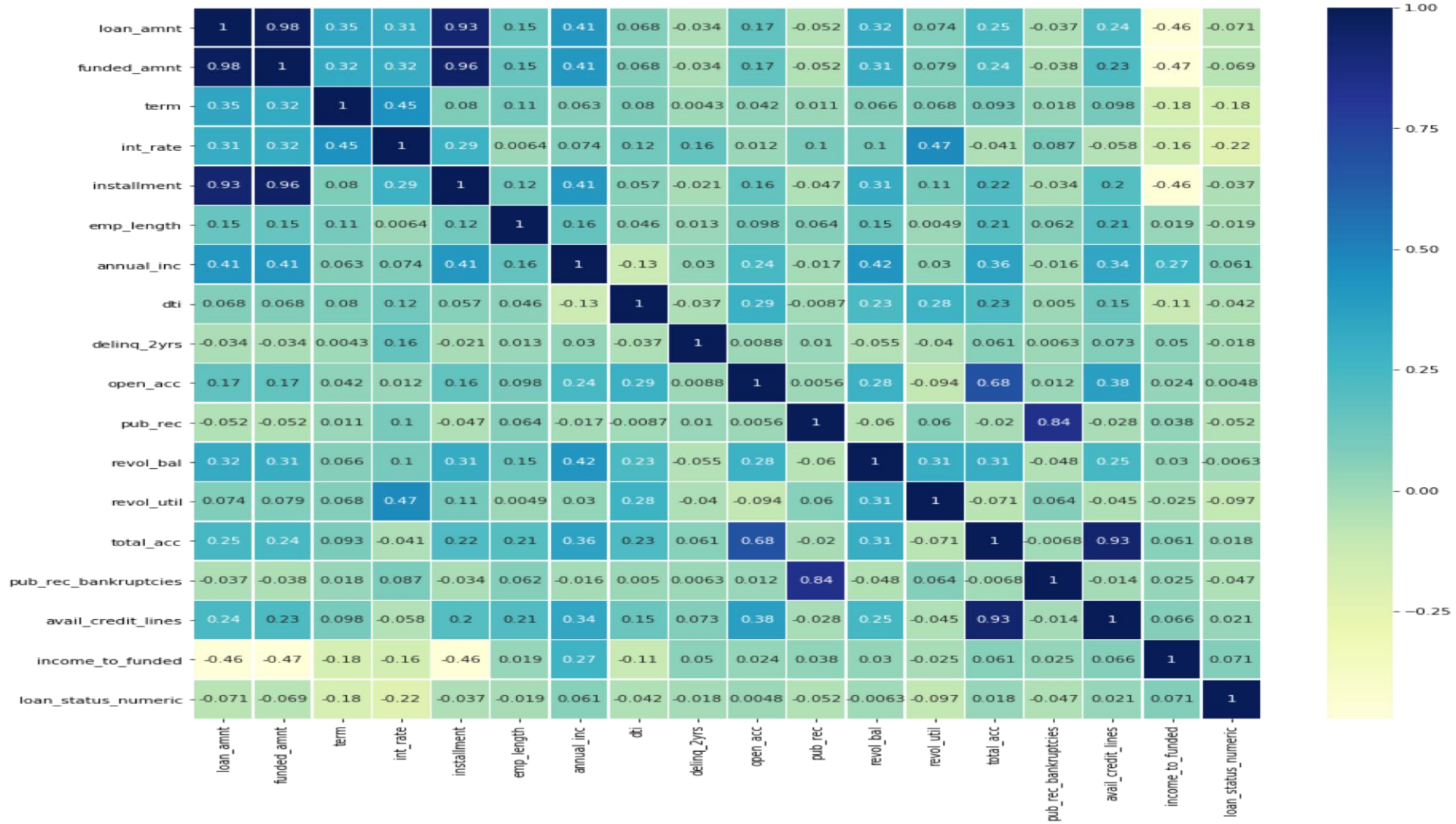
Average 'Charged Off' is 14.2%

Variables which has higher (>> 14.2%) probability of loan default will be influential driver for our analysis.



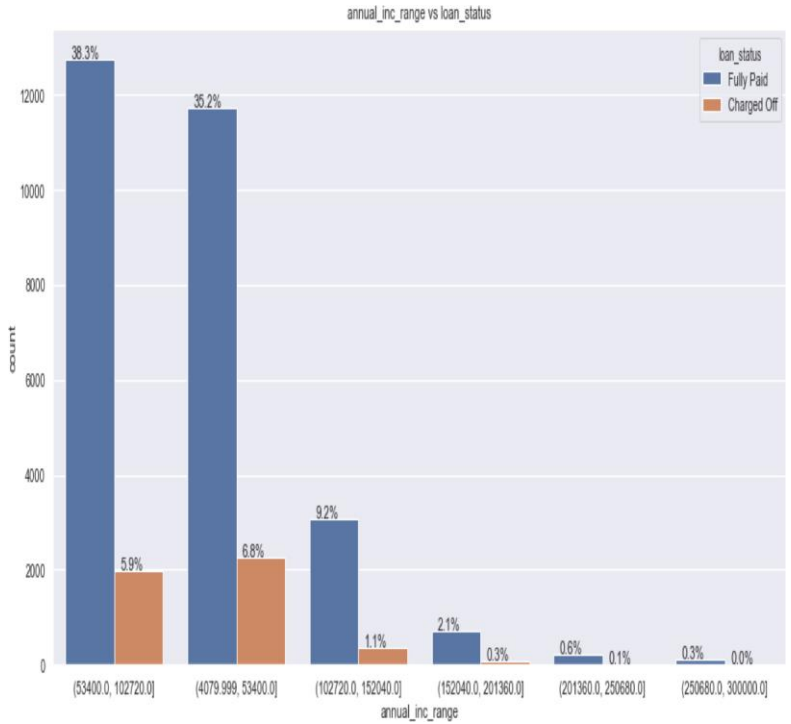
Correlation matrix (Numerical feature vs loan_status)

There is strong correlation between loan_amount vs funded_amnt, installment vs loan_amnt (or funded_amnt), total_acc vs avail_credit_lines, pub_rec vs pub_rec_bankruptencies

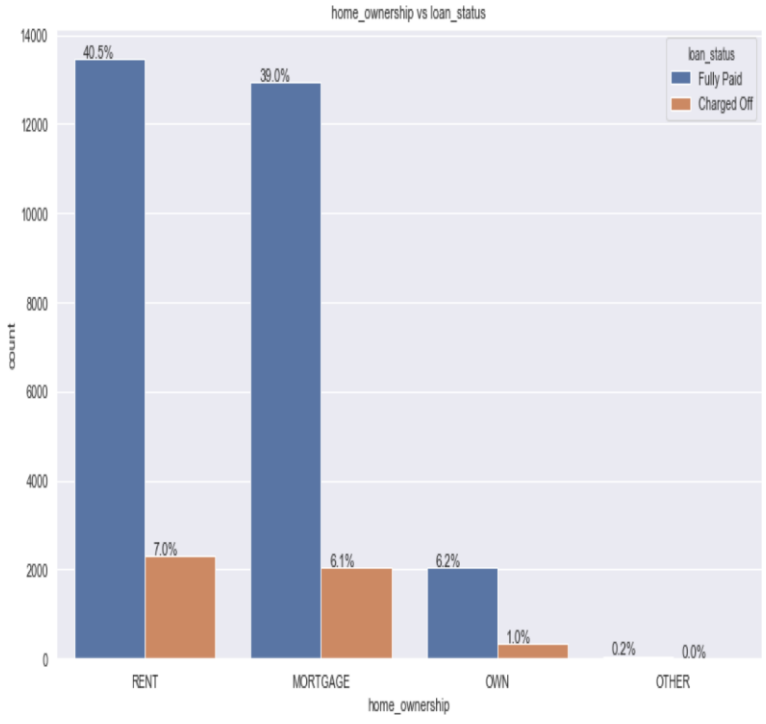


Univariate Analysis - Inferences

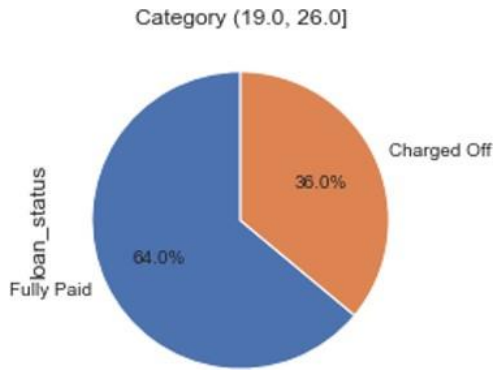
#1 Annual Income

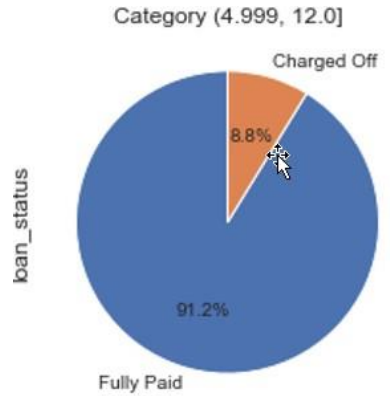


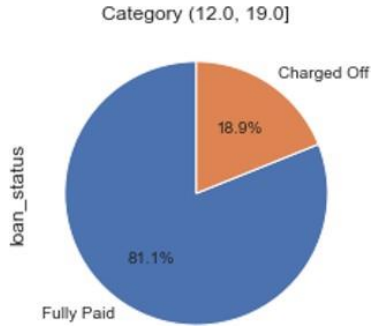
#2 Home Ownership



#3 interest rates







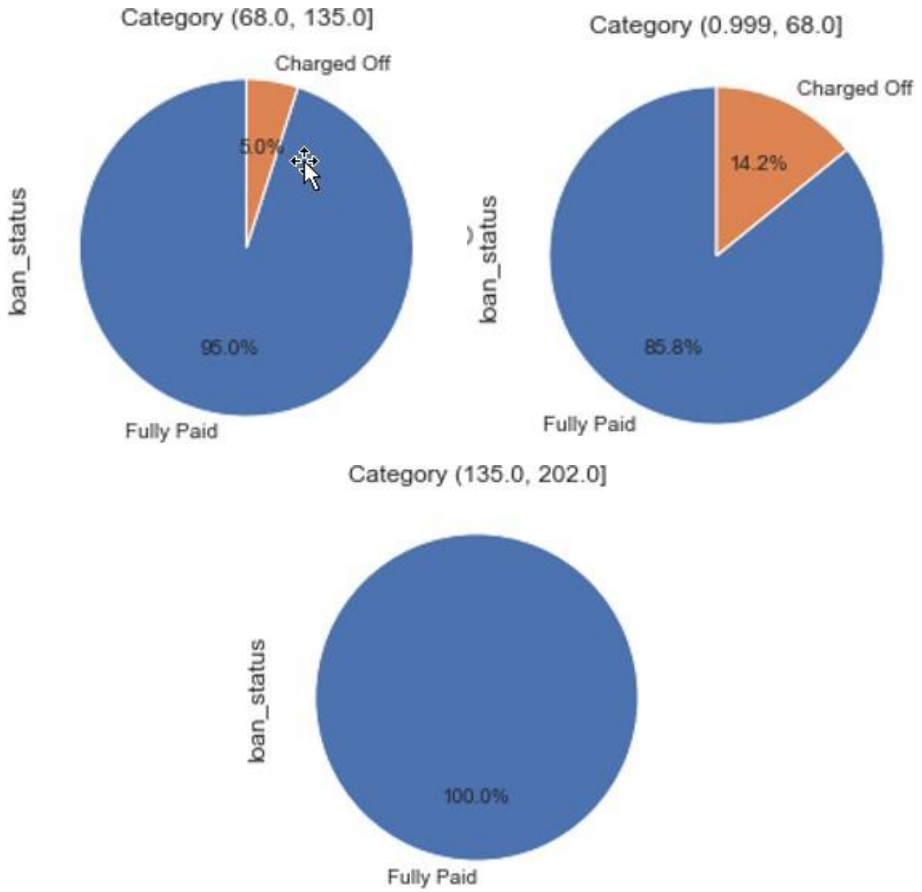
Annual income – Self reported income provided by borrower shows a clear pattern with loan status. Considered for Analysis

Home ownership shows consistent relationship. 0.2% - 0.0% for Other. Considered for Analysis

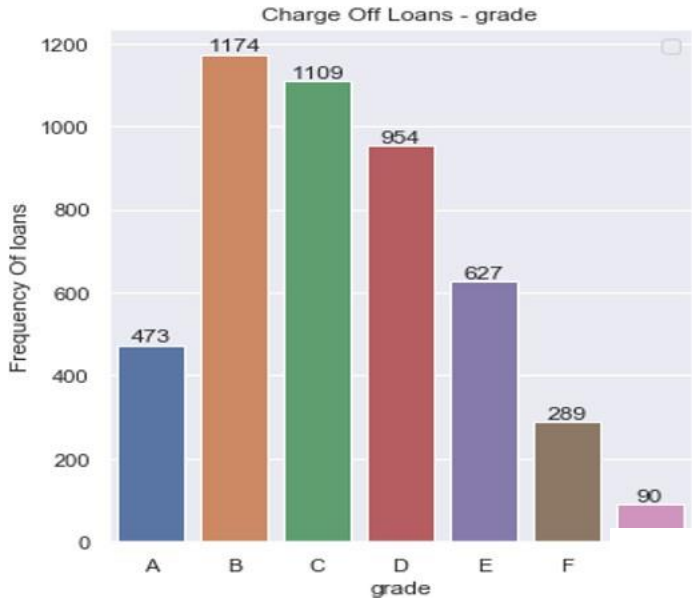
Higher the interest rate, chances of loan default is high. Considered for Analysis

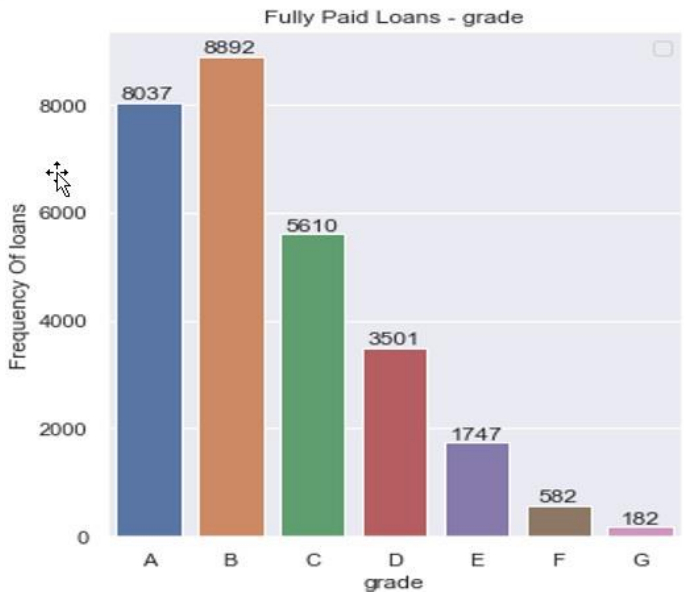
Segmented Univariate Analysis - Inferences

#4 Ratio of Income vs Funded Amount



#5 Grade



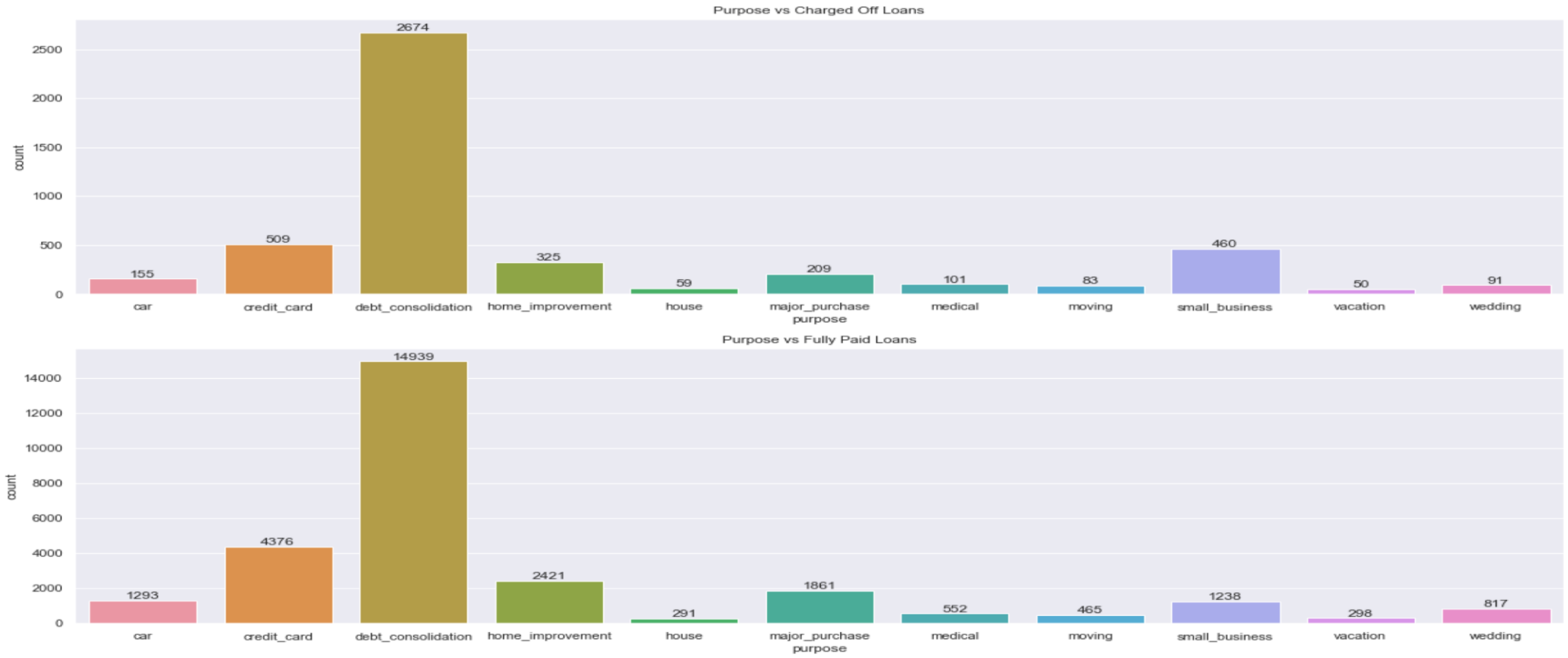


Low ratio between income and funded amount, tends to high loan default.
Dropped funded amount from Analysis.

Grade is assigned in Line of credit. Very Strong relationship emerges for Grades B,C,D.
Considered for analysis.

Univariate Analysis - Inferences

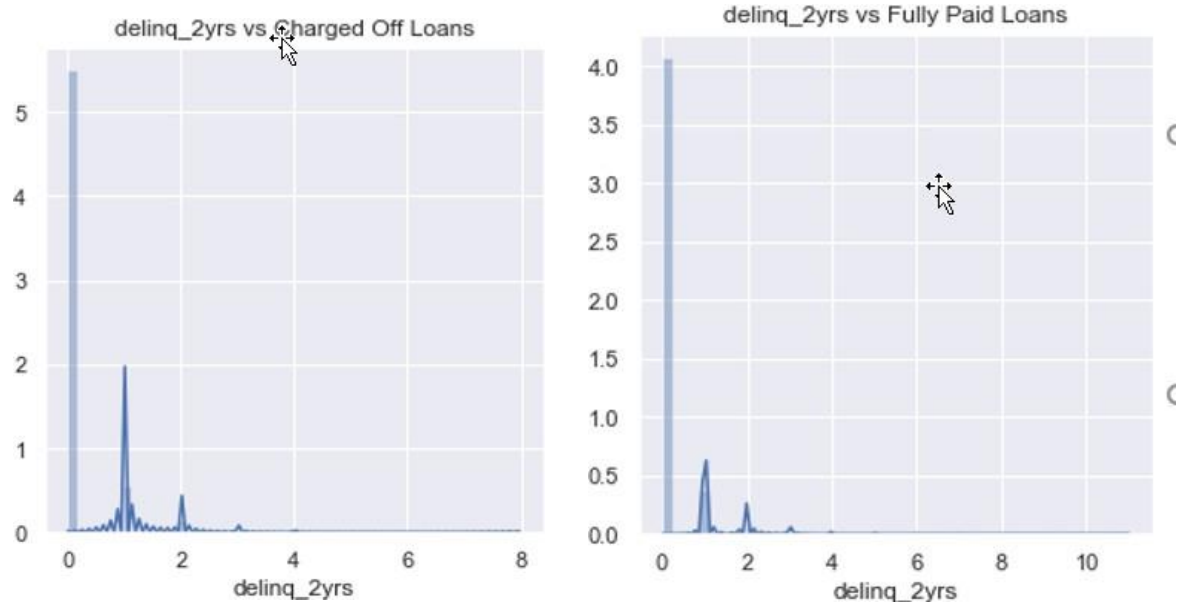
#6 Purpose



Debt consolidations will help to assess the borrower's financial situation which might not be good leading to default . Loan taken for small business has very high chances of failure to repay. Considered for analysis.

Univariate Analysis - Inferences

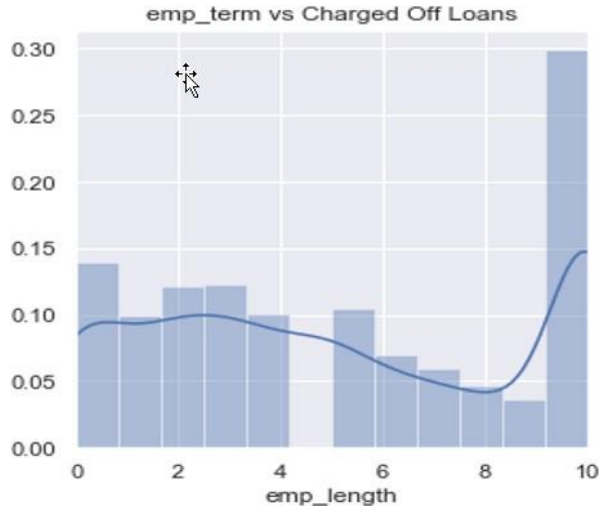
#7 Delinq_2yrs vs Loan Status



delinq_2yrs (which is the number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years) is an important feature to see the trend of default .

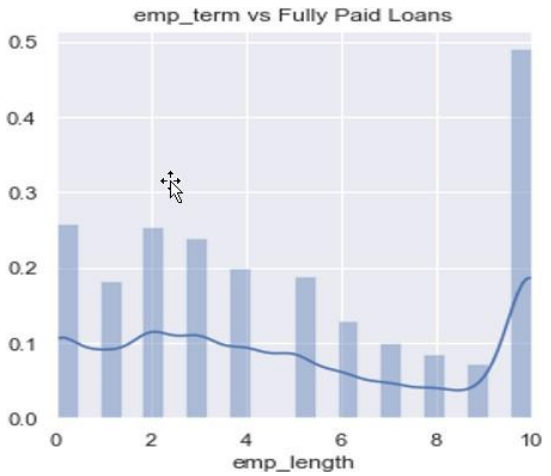
From above plots, for Fully paids loans,delinq_2yrs provides a past trend of the borrower.

#8 Emp_term vs loan_status



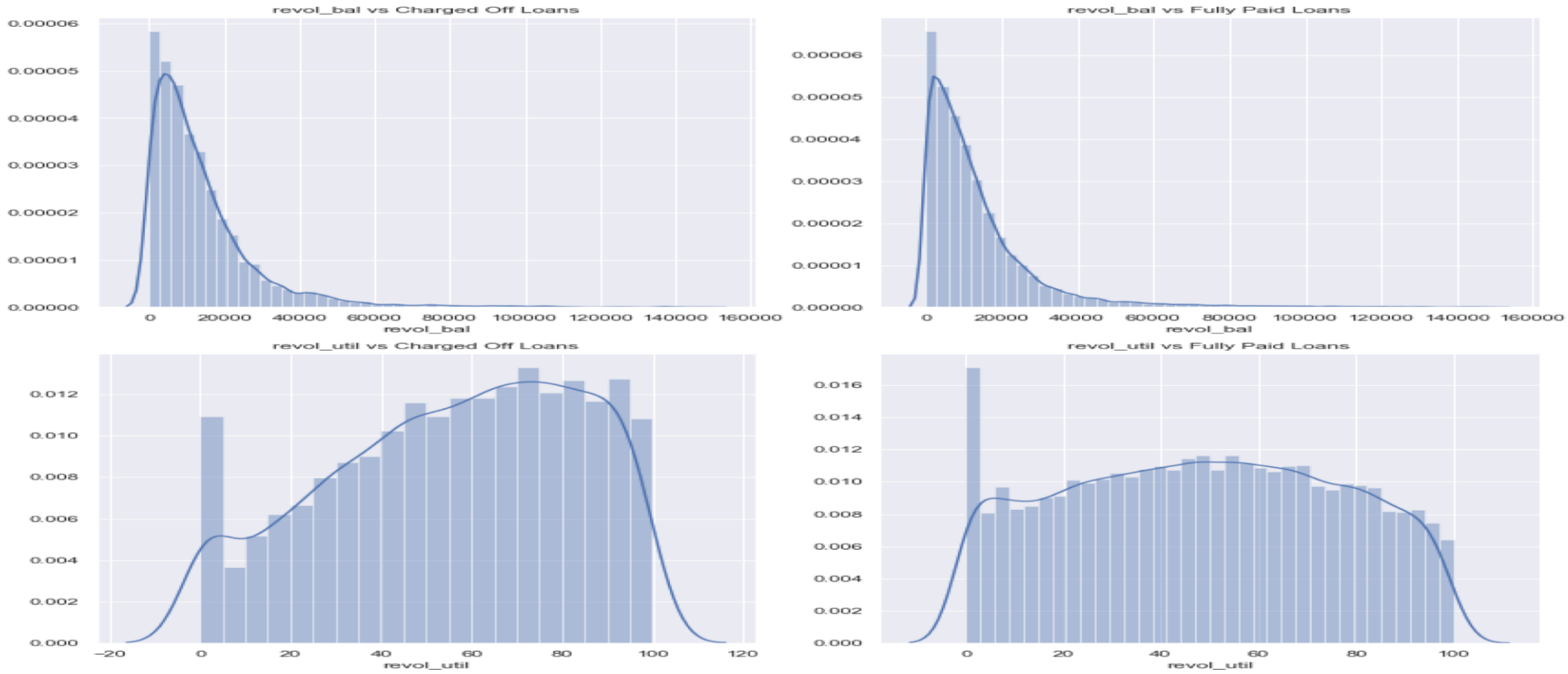
For shorter duration of employment, default is high.

So emp_term too becomes one of the driver variables in predicting loan default (exception is 10+ years)



Univariate Analysis - Inferences

#9 revol_bal ,revol_util w.r.t each loan status

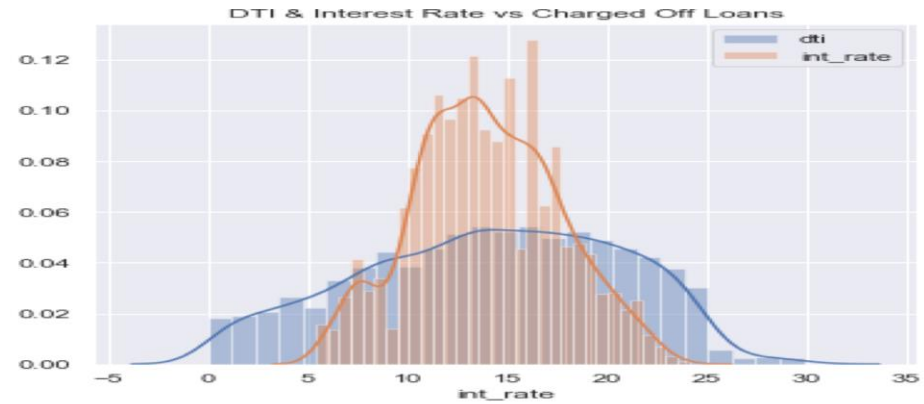


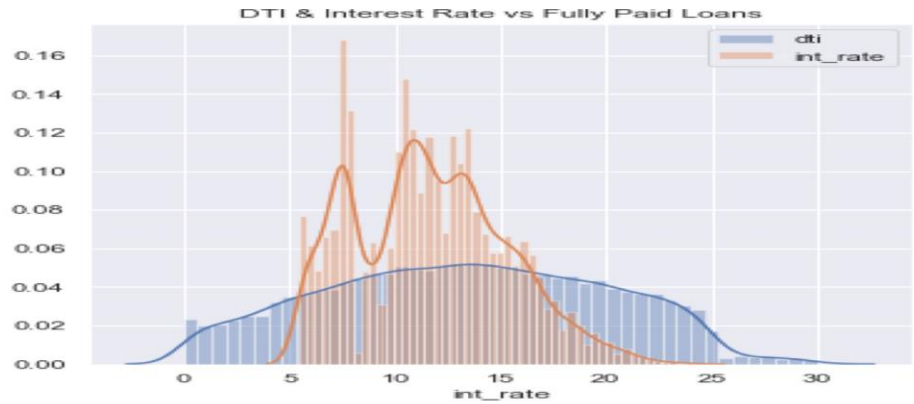
"revol_util"(i.e. the amount of credit the borrower is using relative to all available revolving credit) and "revol_bal" (Total credit revolving balance) is high for charged off compared to fully paid loans.

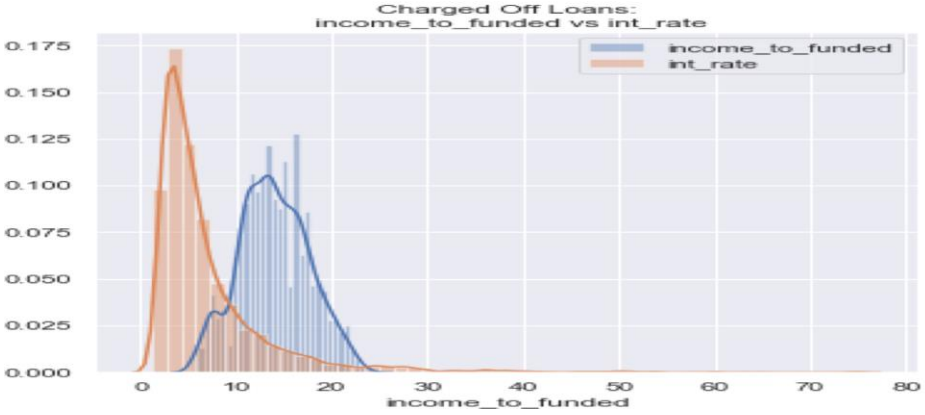
These two variables provides credit information about the borrower.

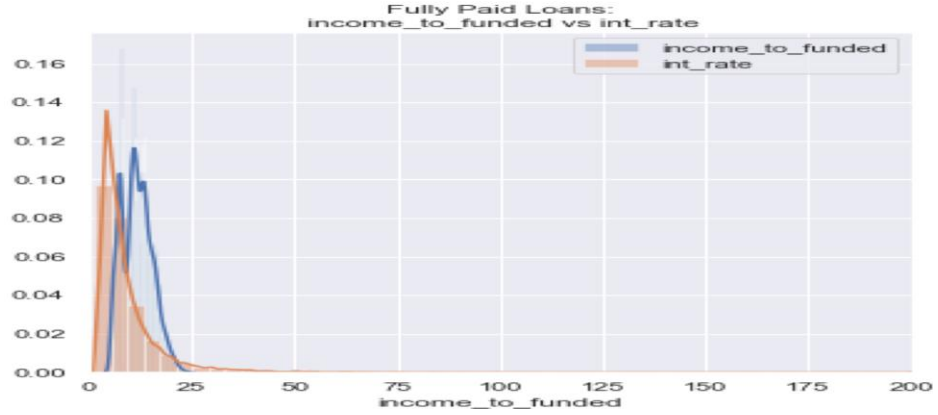
Bivariate Analysis - Inferences

#1 dti and interest rate vs loan status
#2 income_to_funded and int_rate vs loan_status







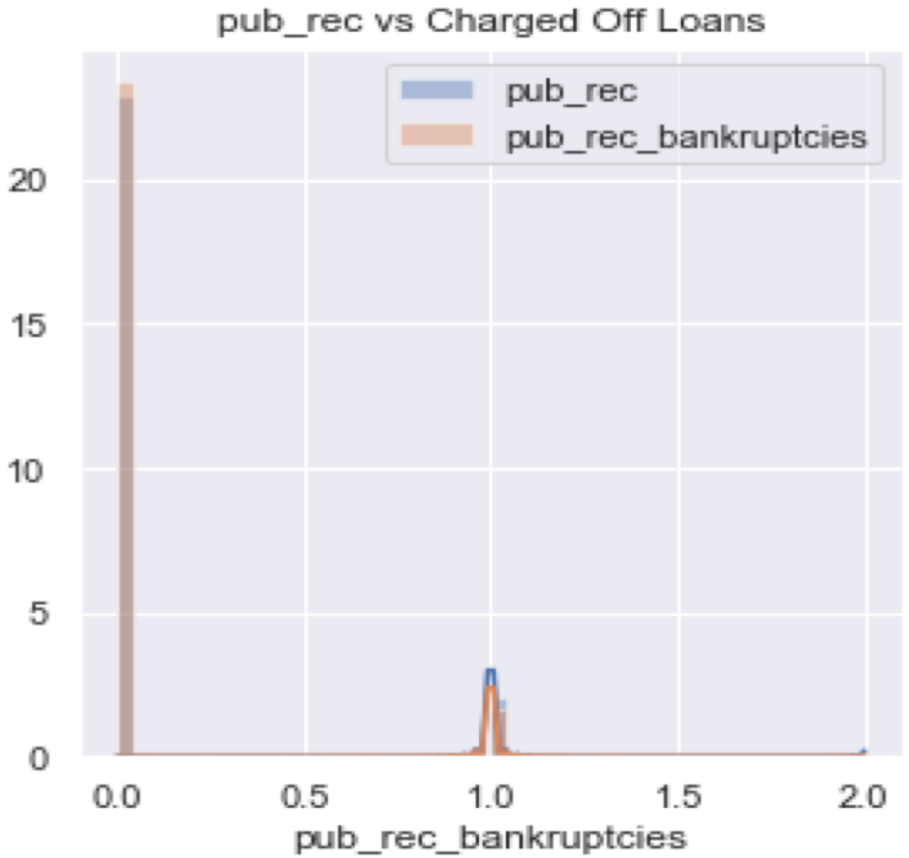


Charged off loans, has high "int_rate" when "dti" is high.

"int_rate" is high when "income_to_funded"(annual_income/funded_amnt)" is high.

Bivariate Analysis - Inferences

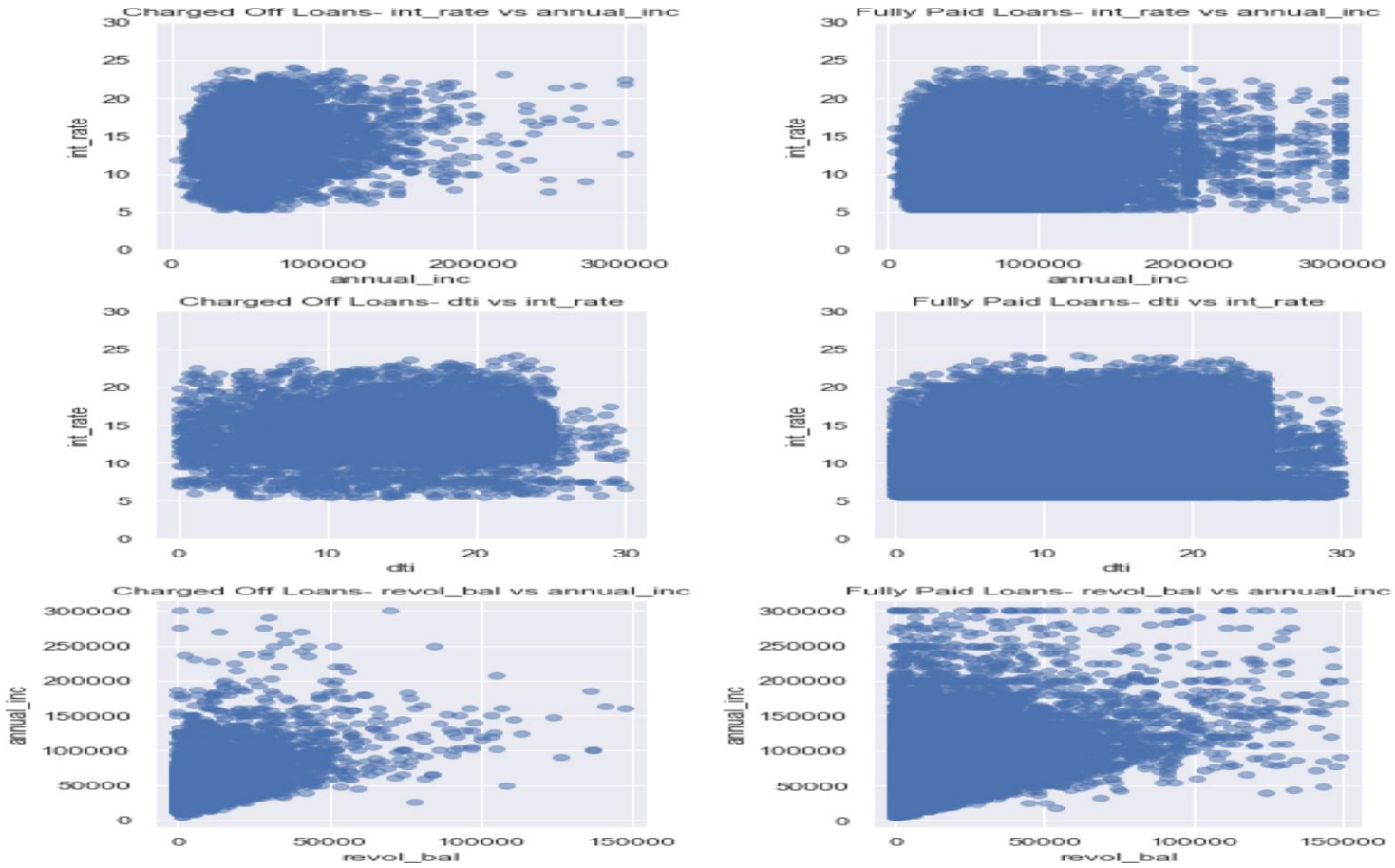
#3 Public record vs Loan Status



"pub_rec" and " pub_rec_bankruptcies" are high in terms of charged off compared to fully paid loans. It gives an idea about the borrower's financial situation which is important for loan approval.

Bivariate Analysis - Inferences

int_rate vs annual_inc ,dti vs int_rate & revol_bal vs annual_inc w.r.t loan_status



"int_rate" is dependent on "annual_inc".

"DTI" and "revol_bal" is dependent on "annual _inc".

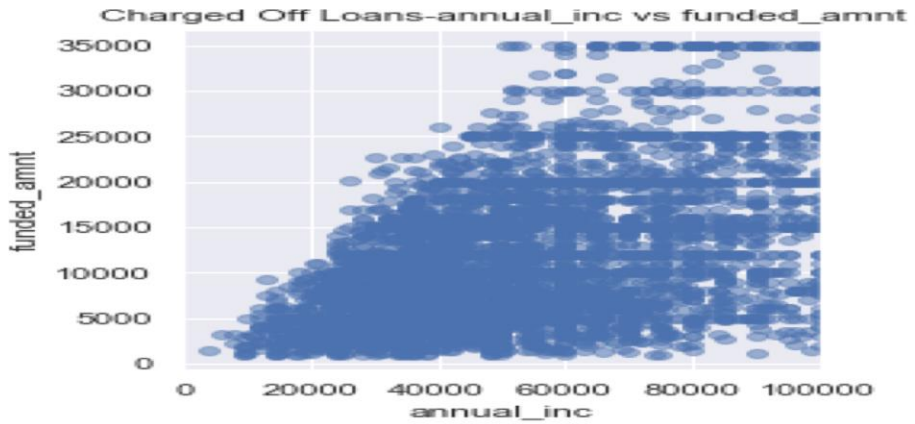
Bivariate Analysis - Inferences

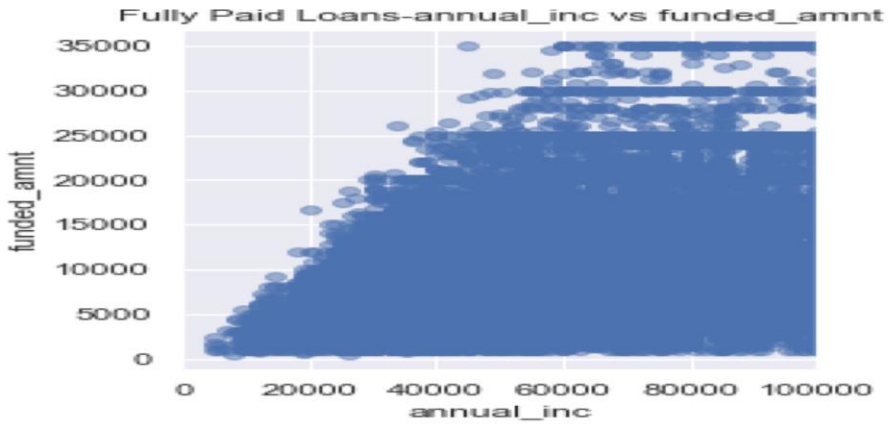
"revol_util" vs "int_rate" w.r.t loan_status

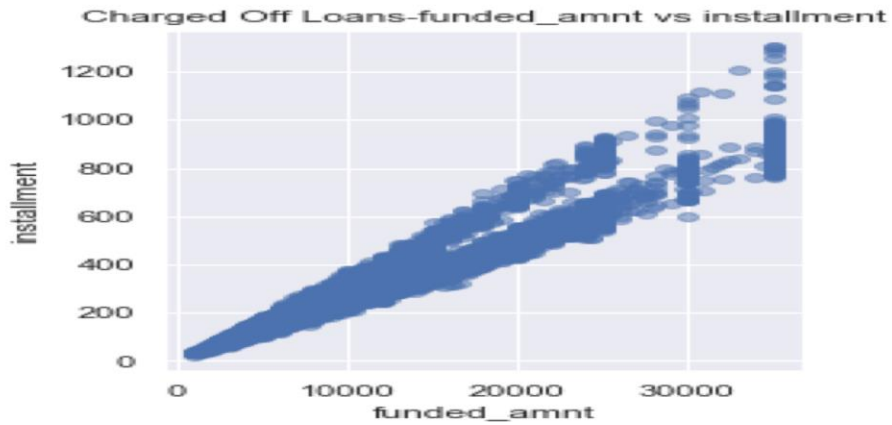
"int_rate" is also dependent on "revol_util".

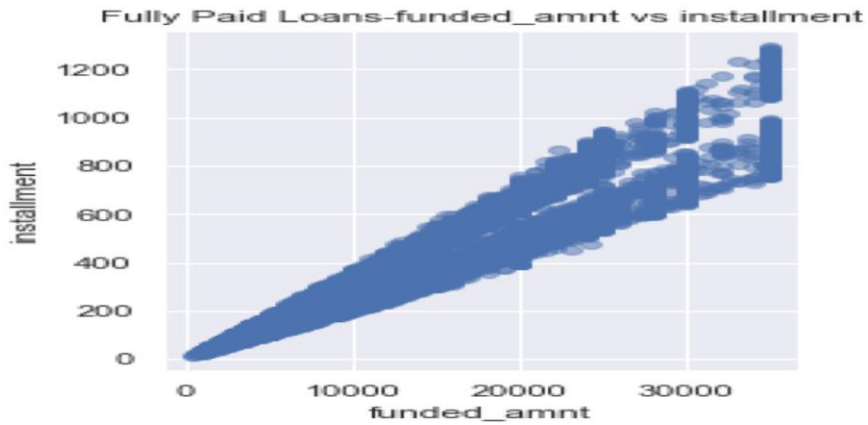
Bivariate Analysis - Inferences

annual_inc vs funded_amnt & funded_amnt vs installment w.r.t loan_status







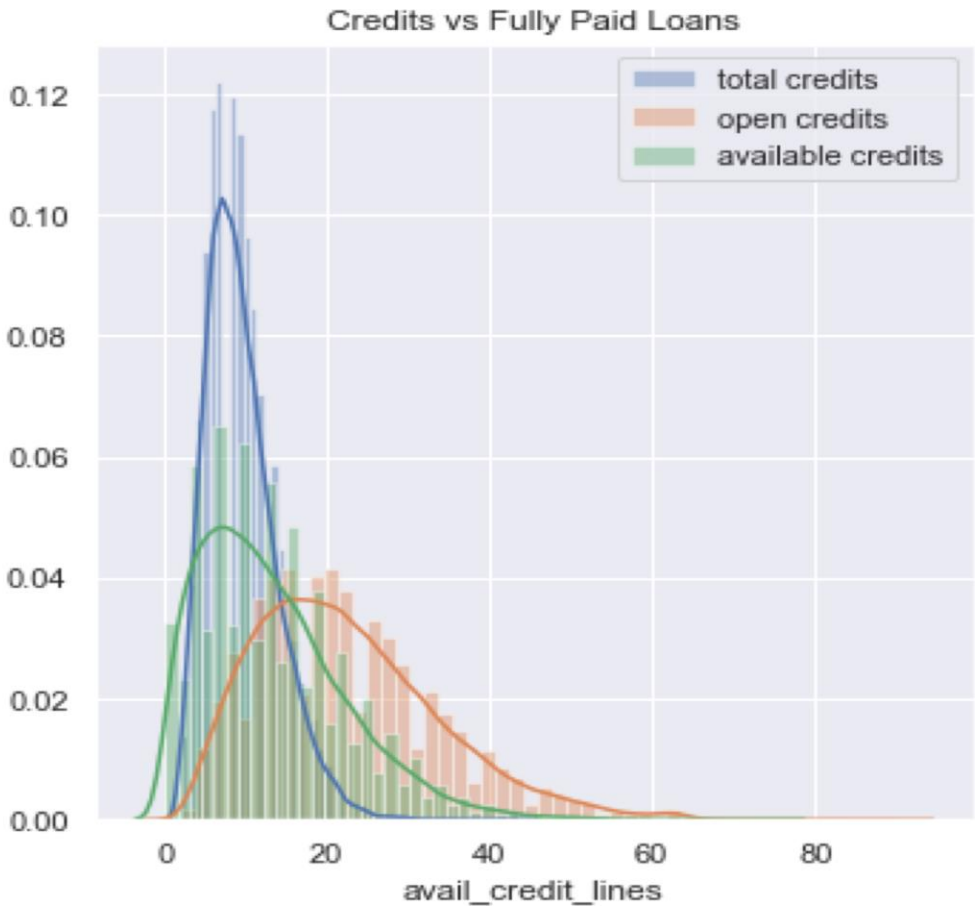


"funded_amnt" is dependent on "annual_inc".

"installment" is dependent on "funded_amount.

Multi-variate Analysis - Inferences

Available Credit Lines



Charged off loans ,has a higher "open_acc"(number of open credit lines) indicating borrower's high expense.

Conclusion - Influential Drivers

As per EDA Analysis, these variables are the primary drivers for loan status which shows strong correlation with the Defaults