LENDING CLUB CASE STUDY

Name:

Sreenath S

N S Chirag

LENDING CLUB CASE STUDY

Lending Club is one of the biggest online consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Hence lending club want to understand key feature that help them to make a decision to approve or reject the loan.

BUSINESS UNDERSTANDING:

There are two types of risks are associated with the bank's decision on loan applications:

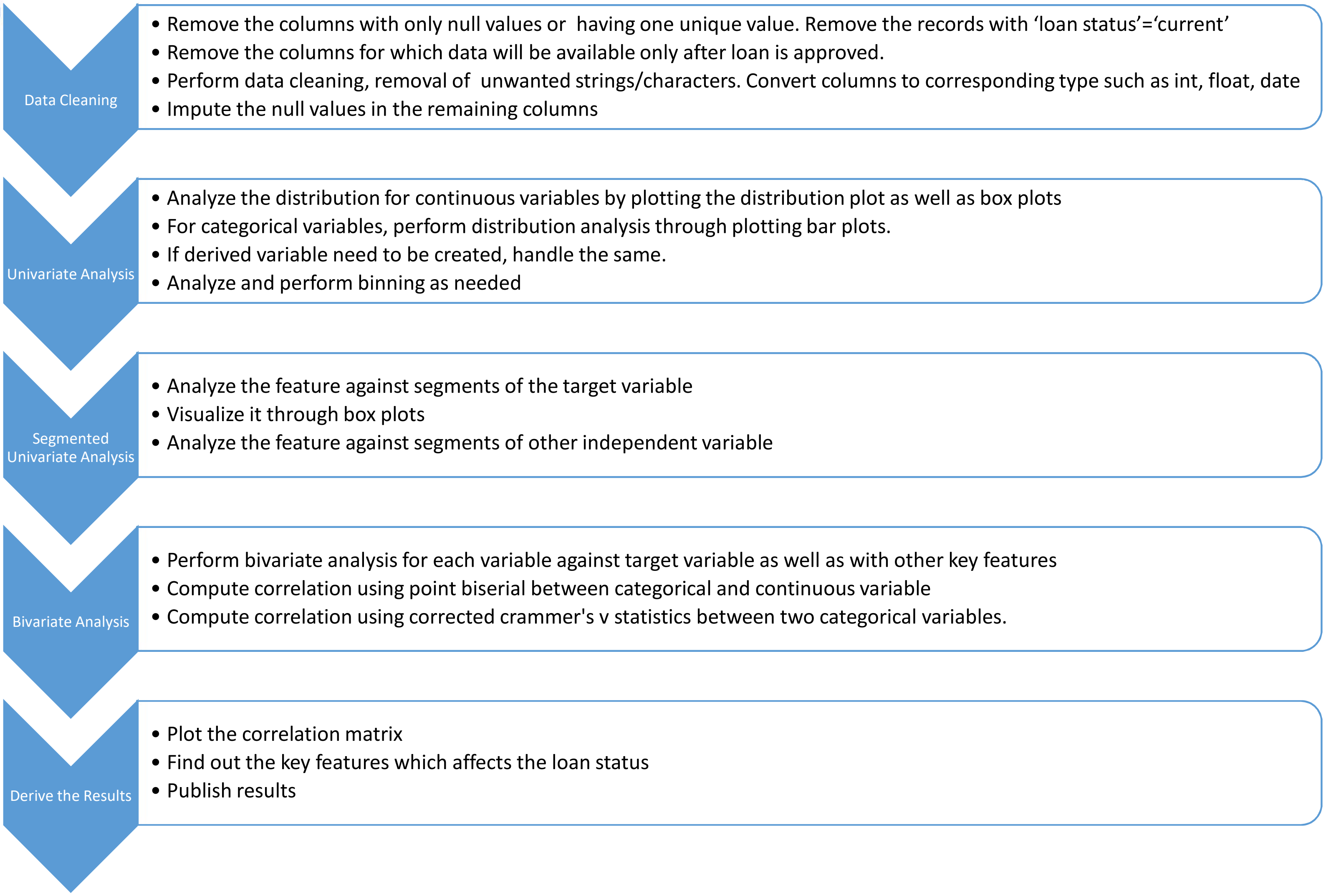
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Lending club want to identify the risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss.

BUSINESS OBJECTIVE OF THE ANALYSIS:

The business objective of this case study is to identify the risky loan applicants using EDA, and then such loans can be reduced thereby cutting down the amount of credit loss. In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

LENDING CLUB EDA- ANALYSIS STEPS



Data Cleaning

- Remove the columns with only null values or having one unique value. Remove the records with ‘loan status’=‘current’
- Remove the columns for which data will be available only after loan is approved.
- Perform data cleaning, removal of unwanted strings/characters. Convert columns to corresponding type such as int, float, date
- Impute the null values in the remaining columns

Univariate Analysis

- Analyze the distribution for continuous variables by plotting the distribution plot as well as box plots
- For categorical variables, perform distribution analysis through plotting bar plots.
- If derived variable need to be created, handle the same.
- Analyze and perform binning as needed

Segmented Univariate Analysis

- Analyze the feature against segments of the target variable
- Visualize it through box plots
- Analyze the feature against segments of other independent variable

Bivariate Analysis

- Perform bivariate analysis for each variable against target variable as well as with other key features
- Compute correlation using point biserial between categorical and continuous variable
- Compute correlation using corrected crammer's v statistics between two categorical variables.

Derive the Results

- Plot the correlation matrix
- Find out the key features which affects the loan status
- Publish results

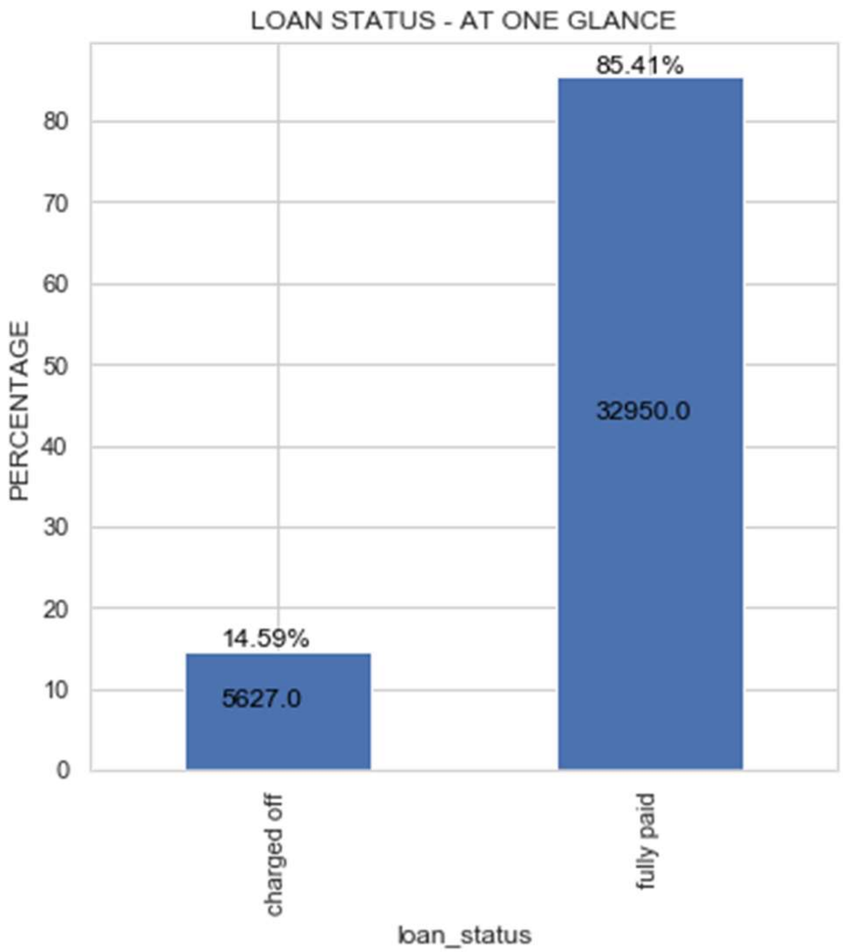
DATA UNDERSTANDING

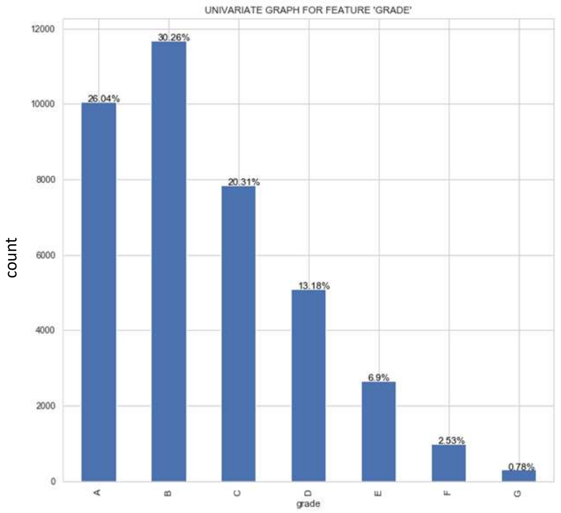
- The data for this analysis are for the period 2007-2011. There were 111 columns and 39717 records.
- Dataset is further tailored to consider only the loans with loan status as “fully paid” and “charged off”.
- Out of 111 feature only the following 24 independent variables are considered for further analysis.

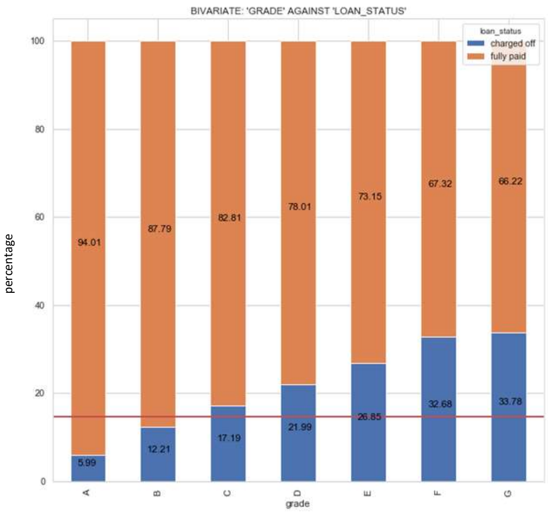
BORROWER’S DEMOGRAPHY	LOAN RELATED DATA	BORROWER’S CREDIT BEHAVIOR
EMPLOYMENT LENGTH	LOAN AMOUNT	DEBIT TO INCOME RATIO*
HOME OWNERSHIP	TERM	DELINQUENCY IN 2 YEARS
ANNUAL INCOME	INTEREST RATE	EARLIEST CREDIT LINE
ADDRESS/STATE	INSTALLMENT	INQUIRIES IN LAST 6 MONTHS
	GRADE	MONTHS SINCE LAST DELINQUENCY
	SUB-GRADE	MONTHS SINCE LAST PUBLIC RECORD
	VERIFICATION STATUS	OPEN ACCOUNTS (FROM ALL DEBIT LINES)
	PURPOSE	TOTAL ACCOUNTS (FROM ALL DEBIT LINES)
		PUBLIC RECORD
		PUBLIC RECORD- BANKRUPTCIES
		REVOLVING BALANCE
		REVOLVING UTIL

Note: Except the Debit To Income Ratio(DTI) all other details in Borrower’s Credit Behavior are available in the Credit Report File prepared by credit agencies/bureaus. Since the debit details are available in credit file and income details will be available as part of loan application, the lenders can calculate DTI at the time of application. Hence all these data are available to lenders through credit report from credit bureaus, or from application at the time of decision making.

DEPENDENT/TARGET FEATURE – 'LOAN STATUS' ANALYSIS





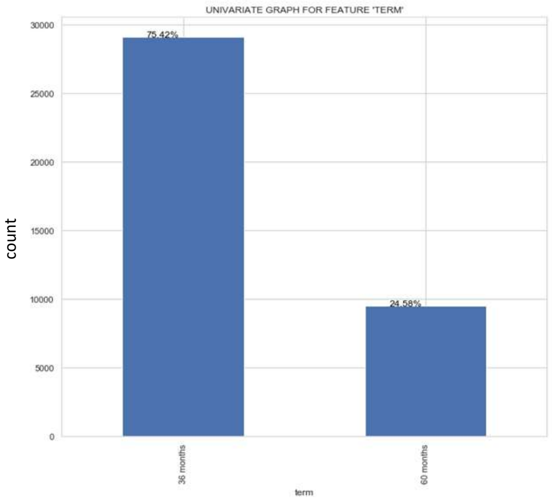


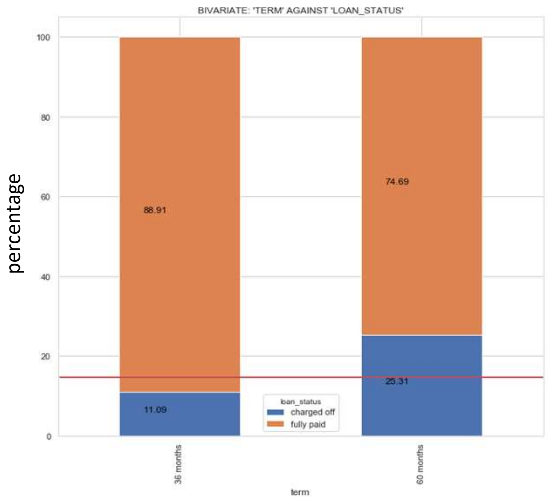
FEATURE ANALYSIS - GRADE

Observation:

- Approximately 77% loans are classified under grade A, B, C
- As grade move towards G, the charge off rate increases. Any loan graded as D, E, F, G has higher risk

The horizontal red line shows the default rate present in the dataset, which is 14.59





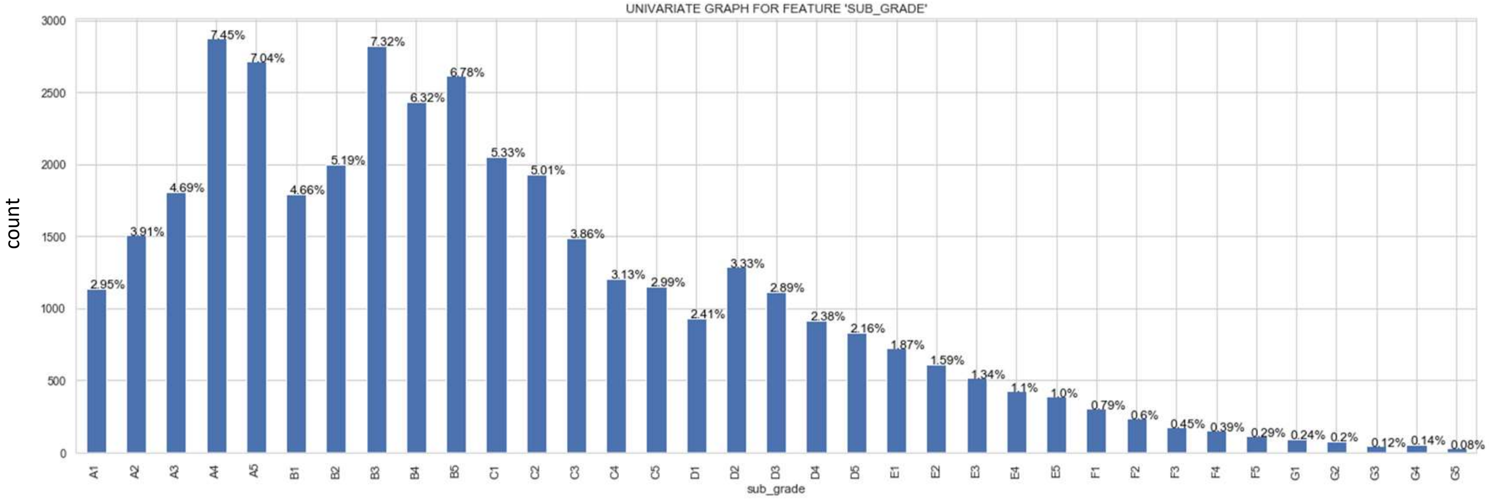
FEATURE ANALYSIS – TERM

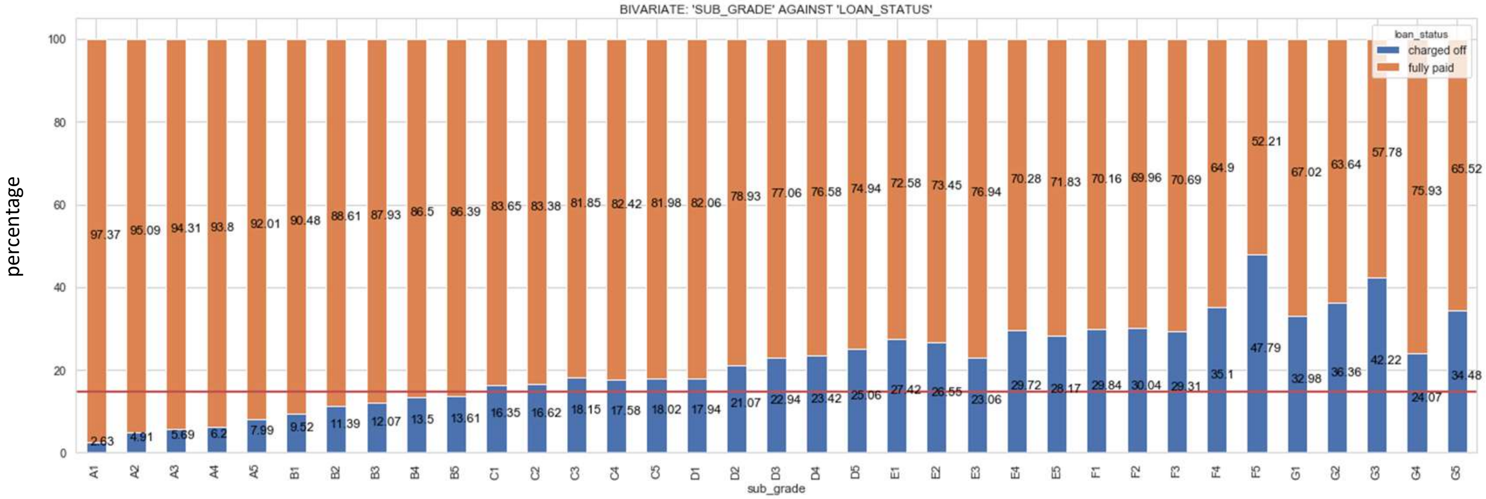
Observation:

- As univariate analysis shows 75.42% are 36 months loan. Only 24.58% loans are having duration 60 months.
- Out of 60 months duration loan, 25.31% of the loans are getting charged off compared to 11.09% for 36 months duration loan

The red line shows the default rate present in the dataset, which is 14.59

FEATURE ANALYSIS – SUB GRADE



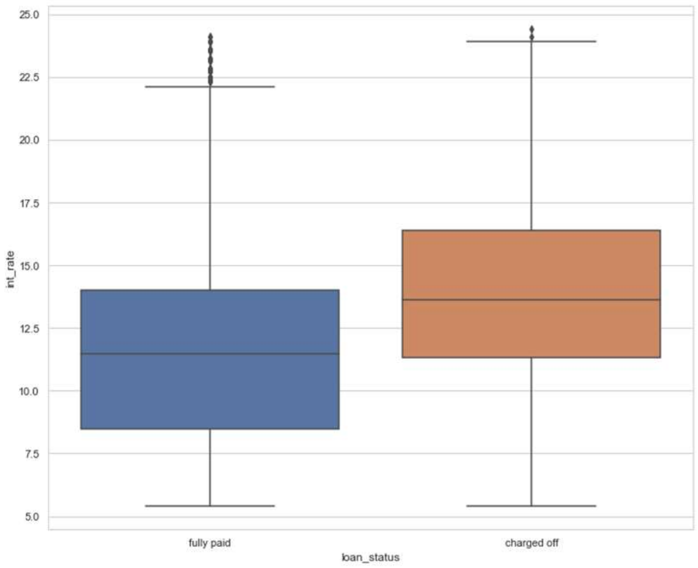


Observation: Sub-grade also follows same pattern as grade. Any sub grade from C1 onwards has higher risk. Risk increases as it graded towards G

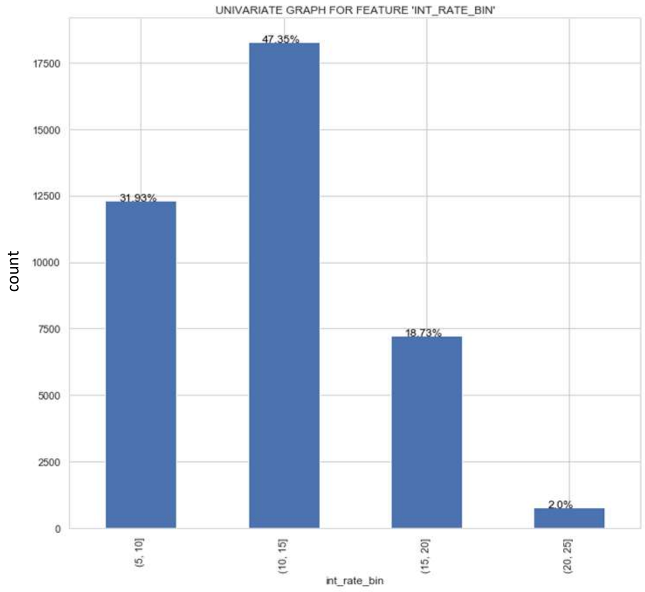
The red line indicates the current charge off rate (14.59%) in the dataset.

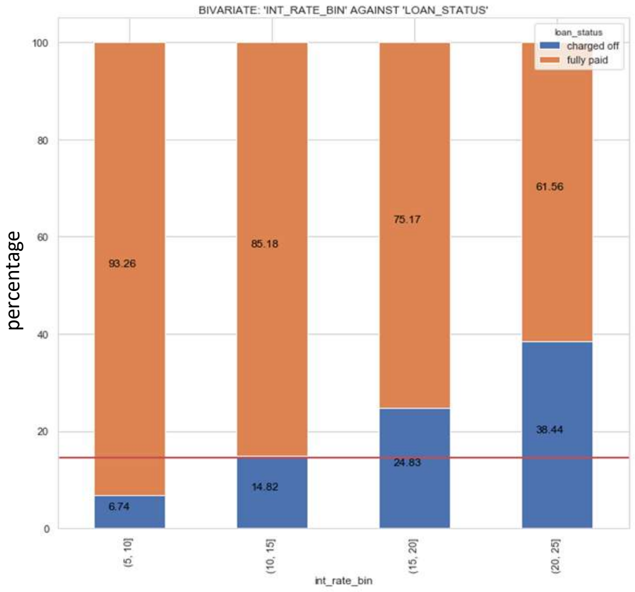
FEATURE ANALYSIS – INTEREST RATE

SEGMENTED UNIVARIATE ANALYSIS



UNIVARIATE AND BIVARIATE ANALYSIS AFTER BINNING THE FEATURE





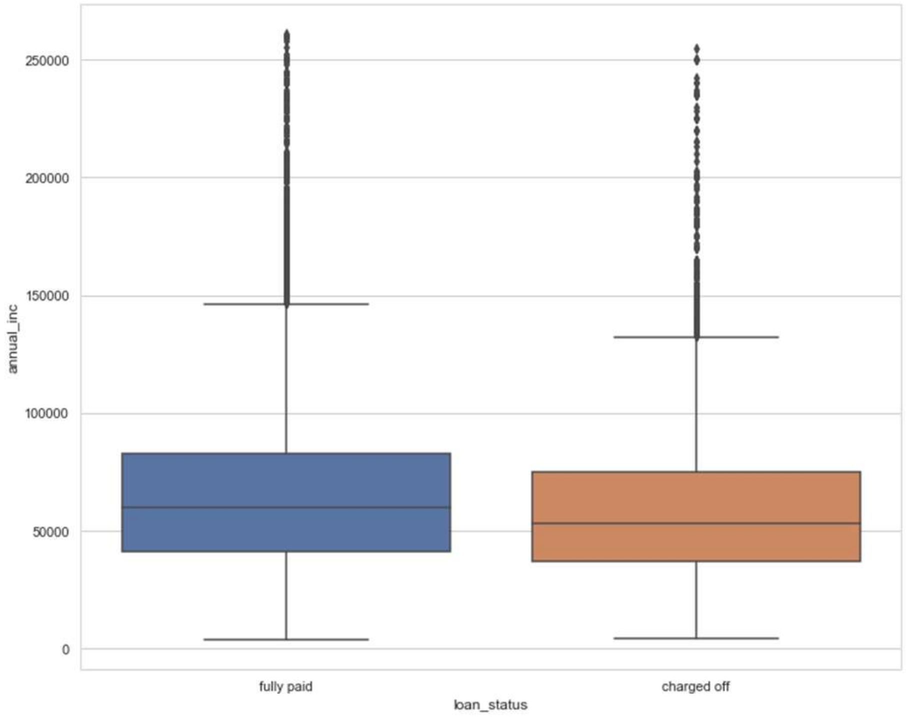
Observation:

- From segmented univariate analysis it is clear that the charge off loans are having higher interest rate
- Loans with interest rate beyond 15% has higher charge off rate and the charge off rate increases as interest rate moves towards 20-25%.

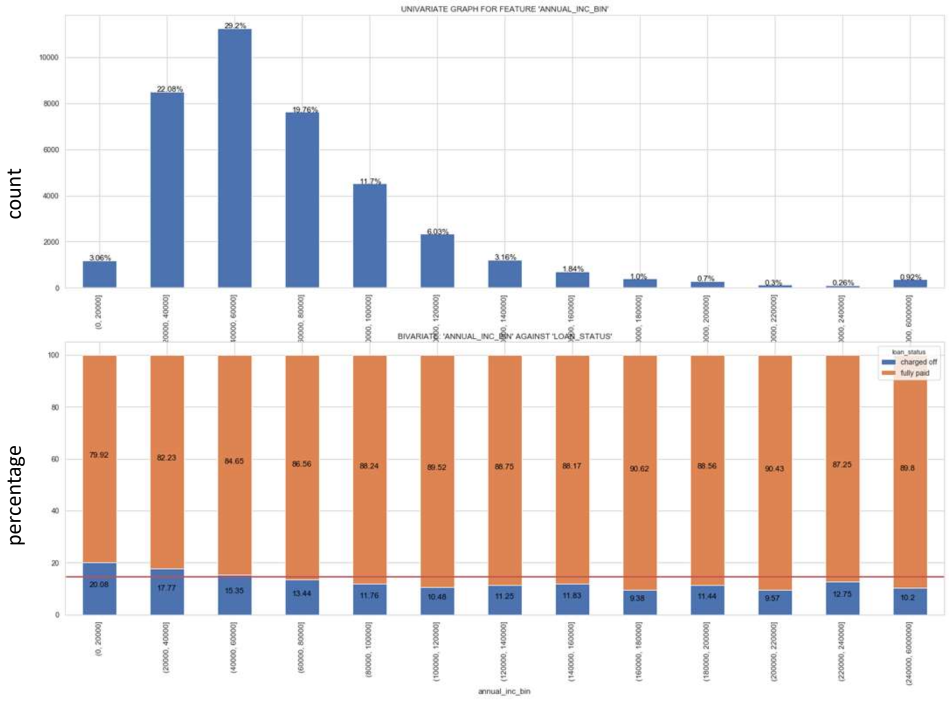
The red line shows the default rate present in the dataset, which is 14.59
For this analysis performed binning on the interest rate variable

FEATURE ANALYSIS – ANNUAL INCOME

SEGMENTED UNIVARIATE ANALYSIS



UNIVARIATE AND BIVARIATE ANALYSIS AFTER BINNING THE FEATURE

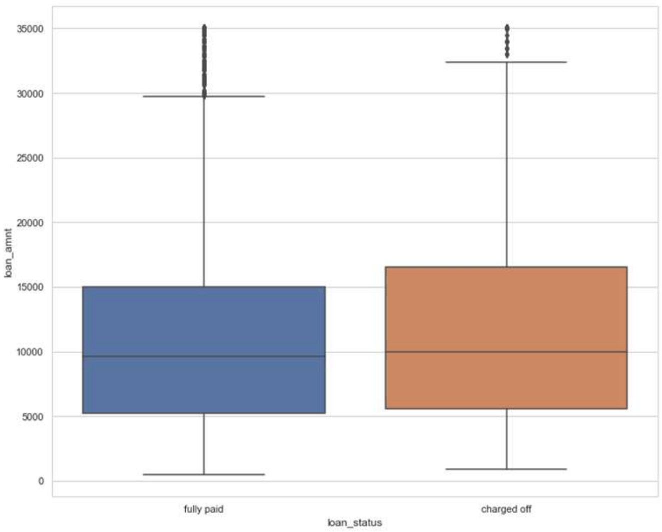


Observation: For charged off loan the annual income is lesser as compared to that of fully paid loans.
Bivariate analysis shows as income decreases the charge off rate increases

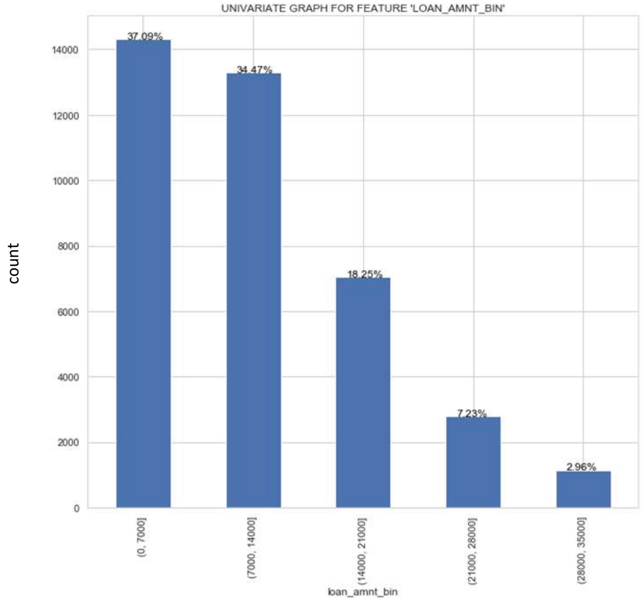
The red horizontal line shows the default rate present in the dataset, which is 14.59%
For this analysis performed binning on the interest rate variable

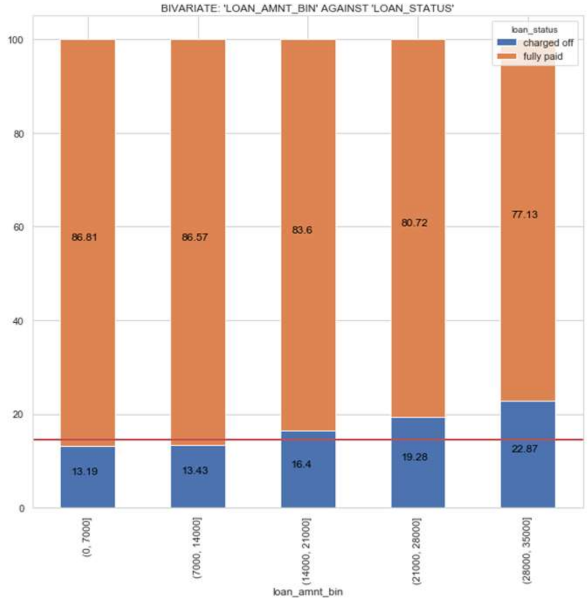
FEATURE ANALYSIS – LOAN AMOUNT

SEGMENTED UNIVARIATE ANALYSIS



UNIVARIATE AND BIVARIATE ANALYSIS AFTER BINNING THE FEATURE


UNIVARIATE GRAPH FOR FEATURE 'LOAN_AMNT_BIN'

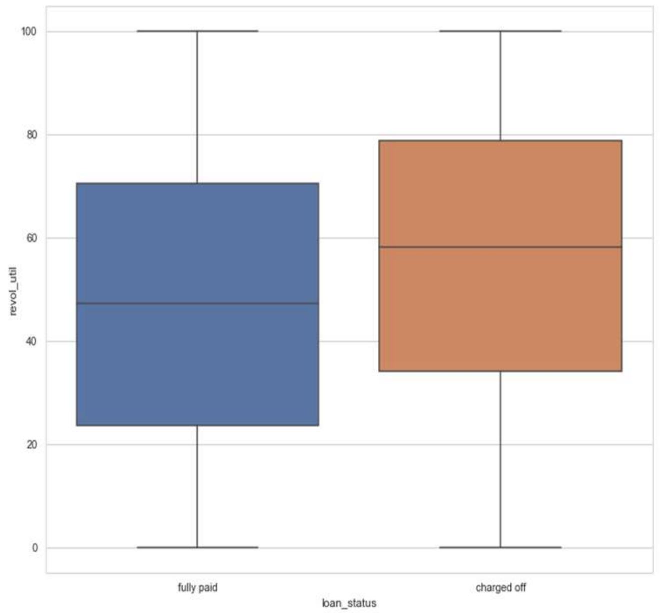

BIVARIATE: 'LOAN_AMNT_BIN' AGAINST 'LOAN_STATUS'

Observation:
It is very clear that as loan amount increases the charge off percentage is also increasing.
For loan amount within 28000 - 35000, the default(charge off) percentage shoots up till 22.87%.

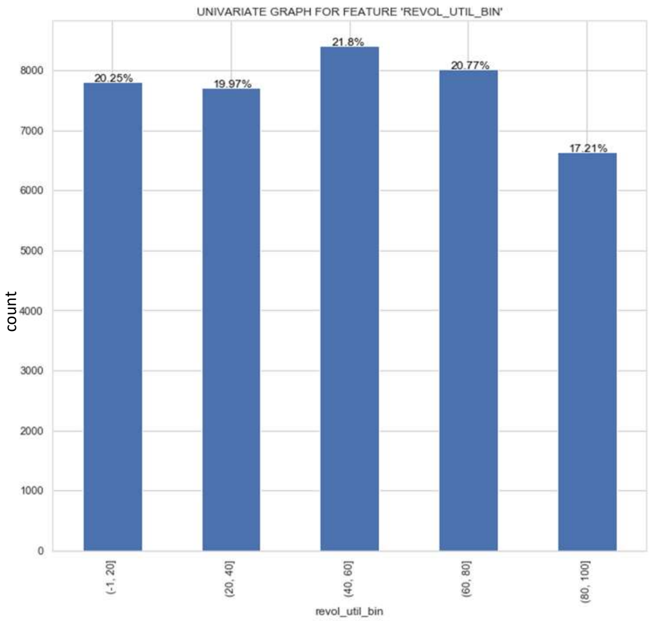
The red horizontal line shows the default rate present in the dataset, which is 14.59%

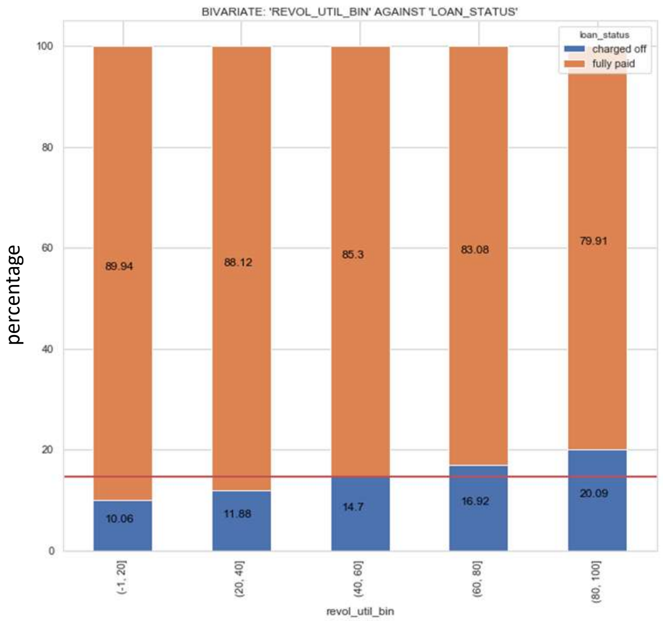
FEATURE ANALYSIS – REVOLVING UTILIZATION

SEGMENTED ANALYSIS



UNIVARIATE AND BIVARIATE ANALYSIS AFTER BINNING THE FEATURE



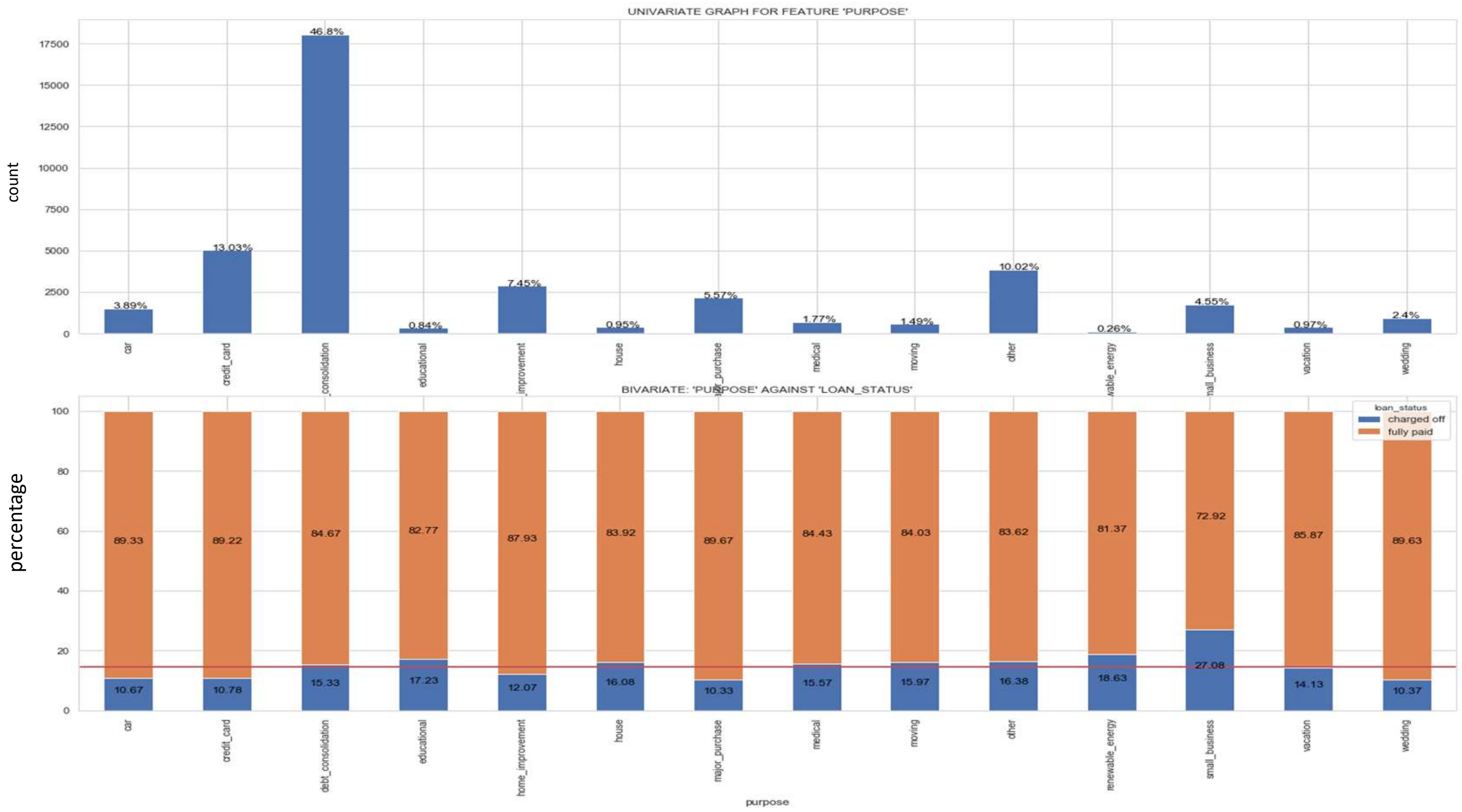


OBSERVATION:

- From segmented univariate analysis it is clear that the charge off loans are having higher revolving utilization
- As revolving utilization increases the charge off % also increases monotonically.
- Loans extended to borrower's with revolving utilization beyond 60 is risky.

The red line shows the default rate present in the dataset, which is 14.59

FEATURE ANALYSIS – PURPOSE

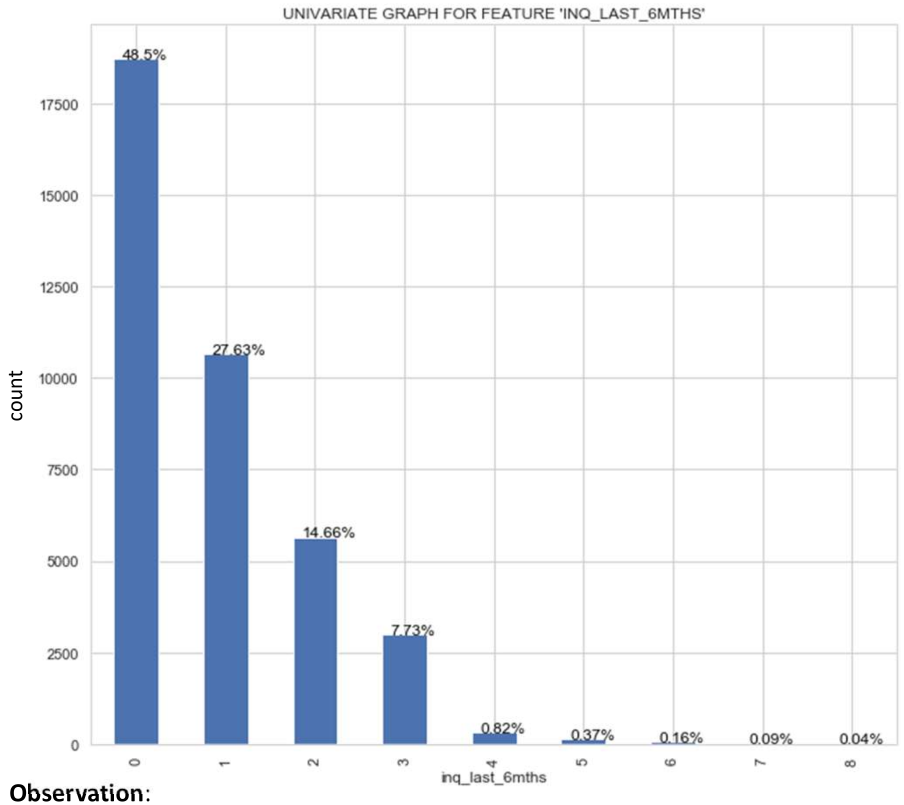


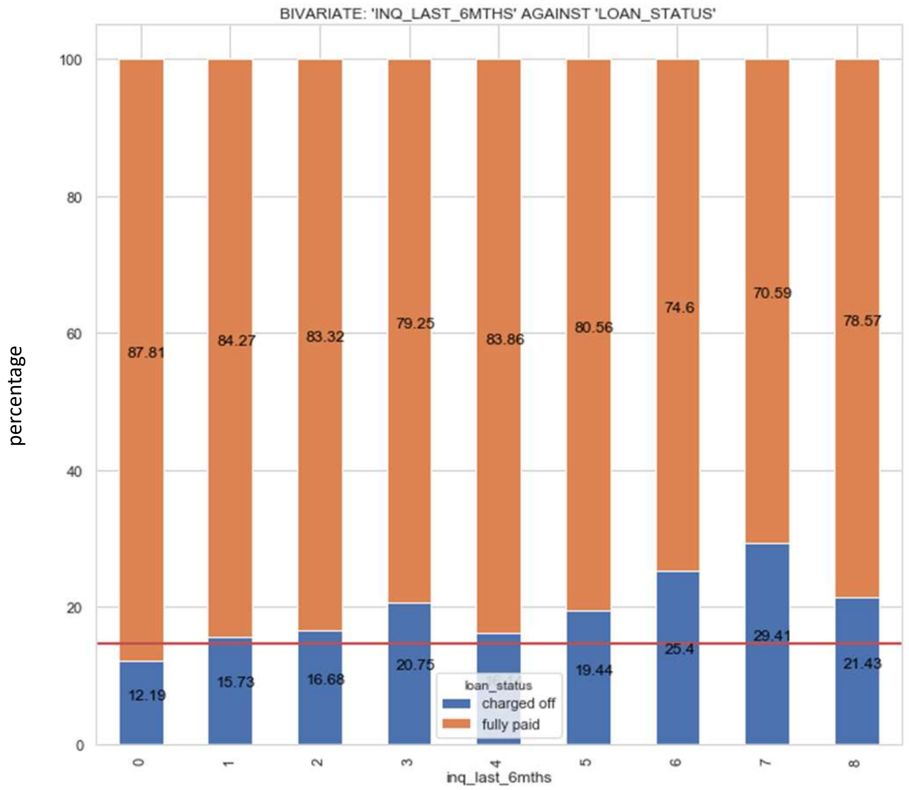
Observation:

- Univariate analysis shows following are top 3 loan purpose: 1. debt_consolidation - 46.80% 2. credit_card - 13.03% 3. other - 10.02%
- Bivariate analysis shows following loans has most default rates above average default rate: 1. small_business - 27.08% 2. renewable_energy - 18.63% 3. educational - 17.23% 4. other - 16.38% 5. house - 16.08% 6. moving - 15.97% 7. medical - 15.57%

- *The red horizontal line shows the default rate present in the dataset, which is 14.59*

FEATURE ANALYSIS – INQUIRIES LAST 6 MONTHS





Observation:

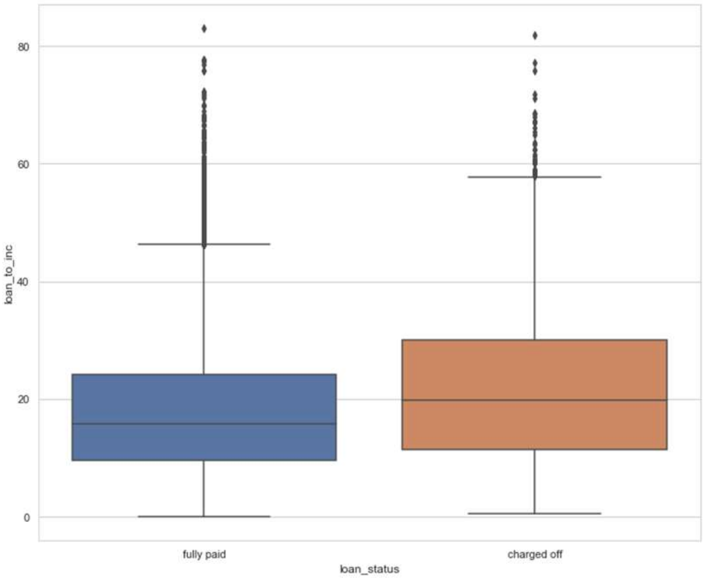
- Univariate analysis shows 48.5% of borrower's has no inquiries in last 6 months as per the credit report created by credit agencies.
- Bivariate analysis shows, beyond 2 credit inquiries in last 6 months, increases the chances of charge off.

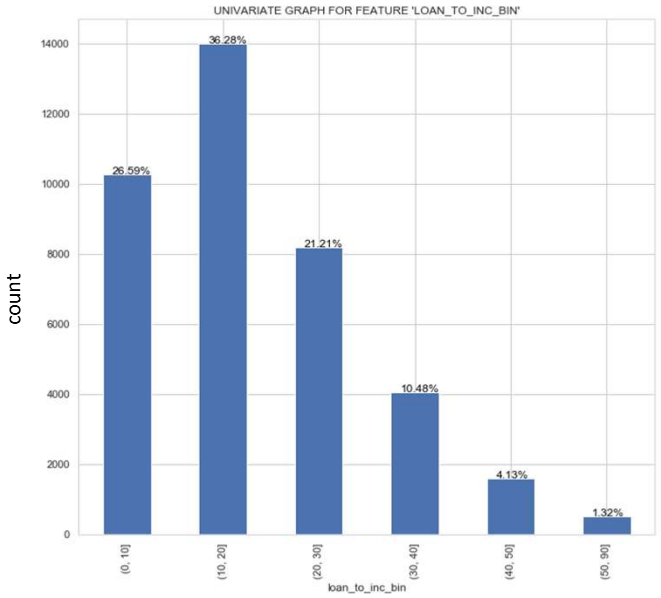
- *The red horizontal line shows the default rate present in the dataset, which is 14.59%*

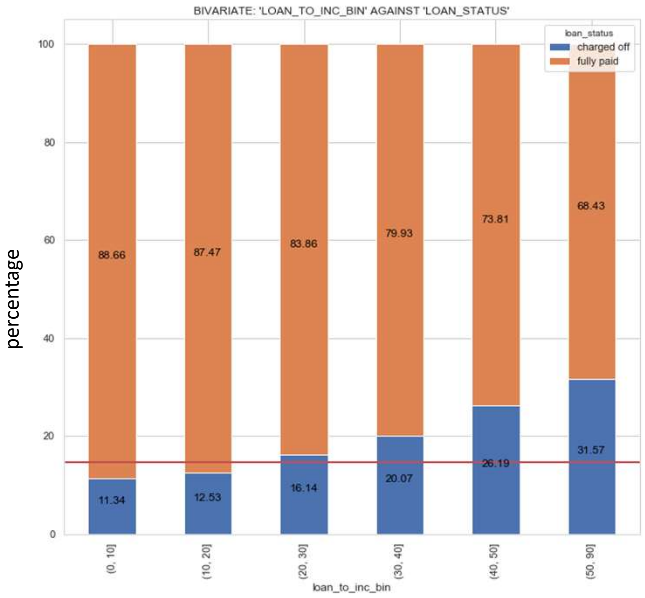
DERIVED FEATURE– LOAN AMOUNT TO ANNUAL INCOME

SEGMENTED UNIVARIATE ANALYIS

UNIVARIATE AND BIVARIATE ANALYIS ON BINNED FEATURE



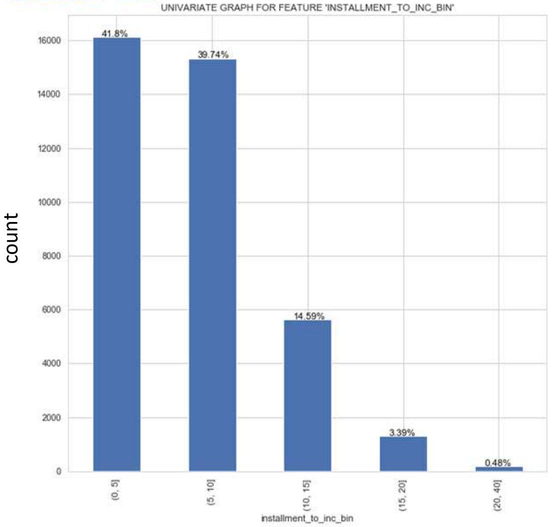


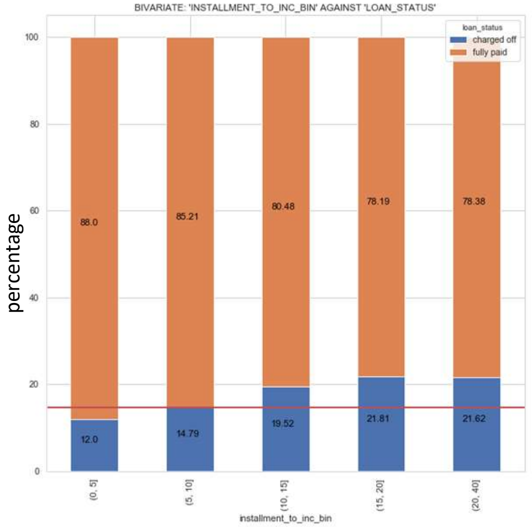


This is a derived variable created as ratio of loan amount to annual income

Observation: If the loan amount to annual income ratio is greater than 30 then there is higher chance for charge off.

The red horizontal line shows the default rate present in the dataset, which is 14.59%





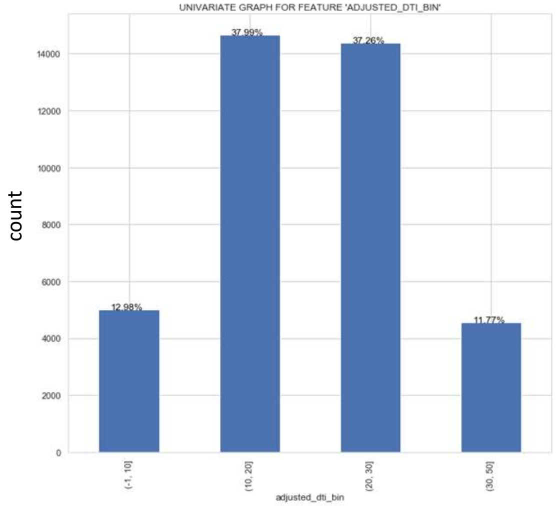
DERIVED FEATURE– INSTALLMENT TO MONTHLY INCOME

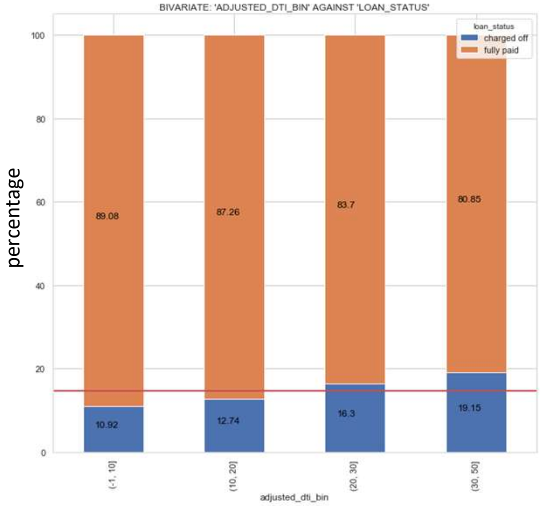
This feature is computed as ratio of monthly installment to monthly income. The Debit to Income Ratio(DTI) does not include the current loan under process. Hence derived a new variable.

Observation

- As installment to monthly income ratio increases the charge off percentage also increases. When this ratio is 20% or more higher the chances of default.

The red line shows the default rate present in the dataset, which is 14.59





DERIVED FEATURE– ADJUSTED DTI

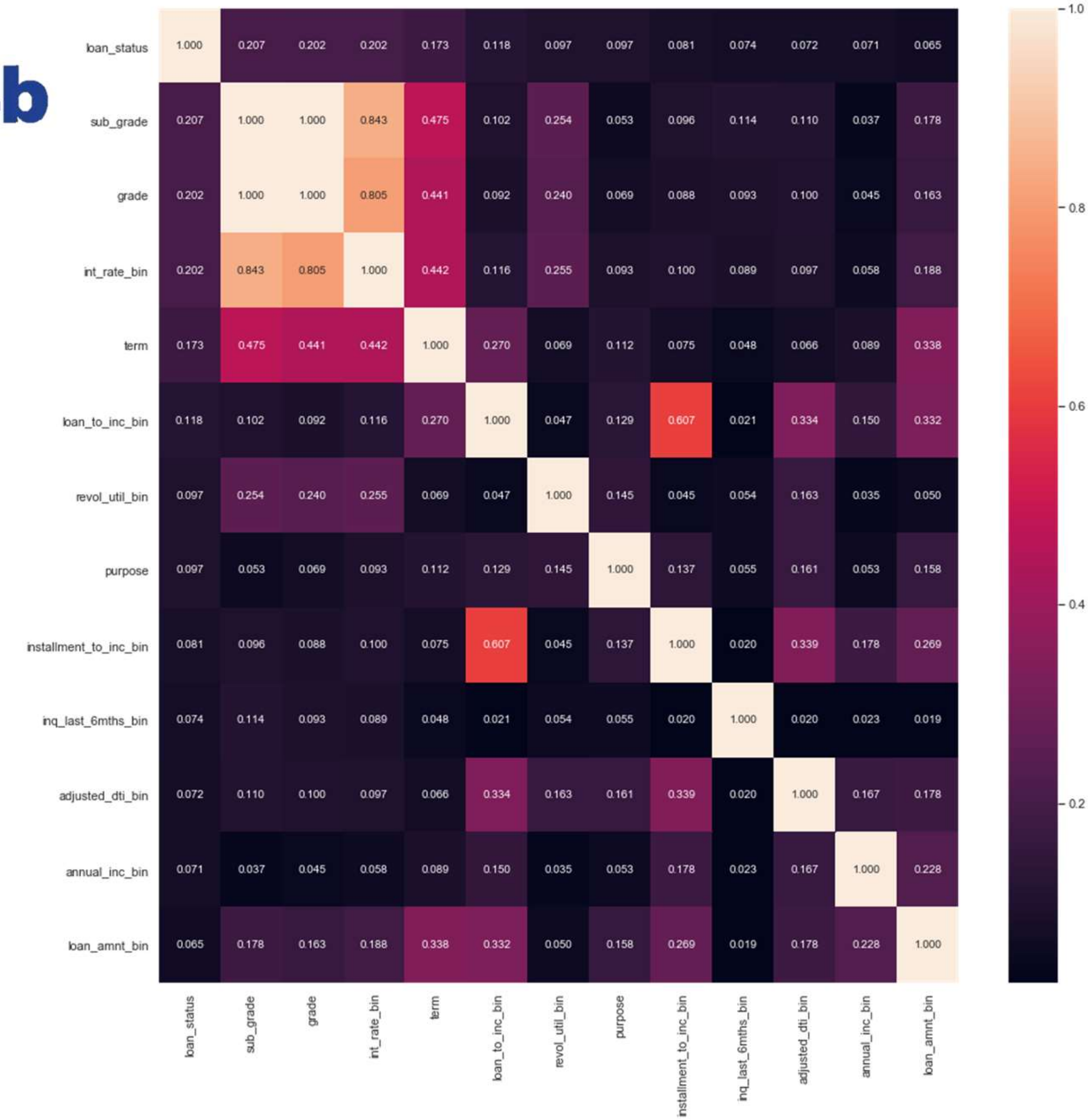
This feature is computed as the sum of DTI + INSTALLMENT TO MONTHLY INCOME RATIO

Observation

- As the Adjusted DTI increases the charge off percentage also increases

The red line shows the default rate present in the dataset, which is 14.59

CORRELATION MATRIX FOR IMPORTANT FEATURES – COMPUTED WITH CORRECTED CRAMER’S V STATISTICS



	loan_status	sub_grade	grade	int_rate_bin	term	loan_to_inc_bin	revol_util_bin	purpose	installment_to_inc_bin	inq_last_6mths_bin	adjusted_dti_bin	annual_inc_bin	loan_amnt_bin
loan_status	1.000	0.207	0.202	0.202	0.173	0.118	0.097	0.097	0.081	0.074	0.072	0.071	0.065
sub_grade	0.207	1.000	1.000	0.843	0.475	0.102	0.254	0.053	0.096	0.114	0.110	0.037	0.178
grade	0.202	1.000	1.000	0.805	0.441	0.092	0.240	0.069	0.088	0.093	0.100	0.045	0.163
int_rate_bin	0.202	0.843	0.805	1.000	0.442	0.116	0.255	0.093	0.100	0.089	0.097	0.058	0.188
term	0.173	0.475	0.441	0.442	1.000	0.270	0.069	0.112	0.075	0.048	0.066	0.089	0.338
loan_to_inc_bin	0.118	0.102	0.092	0.116	0.270	1.000	0.047	0.129	0.607	0.021	0.334	0.150	0.332
revol_util_bin	0.097	0.254	0.240	0.255	0.069	0.047	1.000	0.145	0.045	0.054	0.163	0.035	0.050
purpose	0.097	0.053	0.069	0.093	0.112	0.129	0.145	1.000	0.137	0.055	0.161	0.053	0.158
installment_to_inc_bin	0.081	0.096	0.088	0.100	0.075	0.607	0.045	0.137	1.000	0.020	0.339	0.178	0.269
inq_last_6mths_bin	0.074	0.114	0.093	0.089	0.048	0.021	0.054	0.055	0.020	1.000	0.020	0.023	0.019
adjusted_dti_bin	0.072	0.110	0.100	0.097	0.066	0.334	0.163	0.161	0.339	0.020	1.000	0.167	0.178
annual_inc_bin	0.071	0.037	0.045	0.058	0.089	0.150	0.035	0.053	0.178	0.023	0.167	1.000	0.228
loan_amnt_bin	0.065	0.178	0.163	0.188	0.338	0.332	0.050	0.158	0.269	0.019	0.178	0.228	1.000

Corrected Cramer’s V Statistics is used to compute correlation between categorical variables.

KEY FEATUTES

Following are the key driving features identified during EDA analysis:

- Sub-Grade
- Grade
- Interest Rate
- Term
- Loan amount to Annual Income ratio**
- Revolving Utilization
- Purpose
- Monthly Installment to Monthly Income Ratio**
- Inquiries in Last 6 Months
- Adjusted DTI **

NOTE: Out of this list grade and subgrade are highly correlated hence lending club can consider either of the one.

*The features which are tagged with '**' are derived features. Since these features cover the ratio of annual income to total loan amount, monthly income to installment, these features will cover the impact of income, loan amount, installment, DTI on loan status.*

Apart from the above list, the applicants who hasn't mentioned the employment length (unemployed or self employed) also tends to charge off highly.