LENDING CLUB GROUP CASE STUDY

SUBMITTED BY

Jishna M

Executive Summary

The case study is an attempt using EDA to identify the driving variables which are strong indicators of applicants to be a Loan Defaults. The data set as well as data dictionary were collected and processed with necessary data cleaning steps. Non-impacting features were removed through this process and the rest of the features were analysed using Univariate, Bivariate, Segmented and Multivariate analysis. After each analysis, features those weren't having significant influence of defaulter ratio were removed consecutively. From 115 potential features in the given dataset 7 driver variables that indicates possibility of loan default was identified and they were **annual _inc, funded_amnt_inv, int_rate, term, grade, purpose, and addr_state.**

Business Understanding

The business objective is to identify the risky loan applicants who are potential Defaults , and therefore reducing sanctioning such loans so as to cut down the credit loss. These driving factors can be used for portfolio and risk assessment by the company.

- Two Types of Risks:

  a) Not sanctioning the loan when the applicant can repay, and hence loss of business.

  b) Sanctioning the loan when applicant can’t repay and hence, financial loss.

- The scope of the project is to minimize only second type risk.
- Two Types of Decision taken : a) Loan Accepted and b) Loan Rejected
- Loan Rejected means no transactional history of applicants hence its out of scope for this analysis.
- Three loan statuses: Fully Paid – Applicant fully paid the loan, Charged Off – Applicant has not paid the loan and loan is default, Current – Applicant in process of paying instalment.
- Current status loans are excluded as they can’t provide insight for the defined scope of project.

Data Understanding

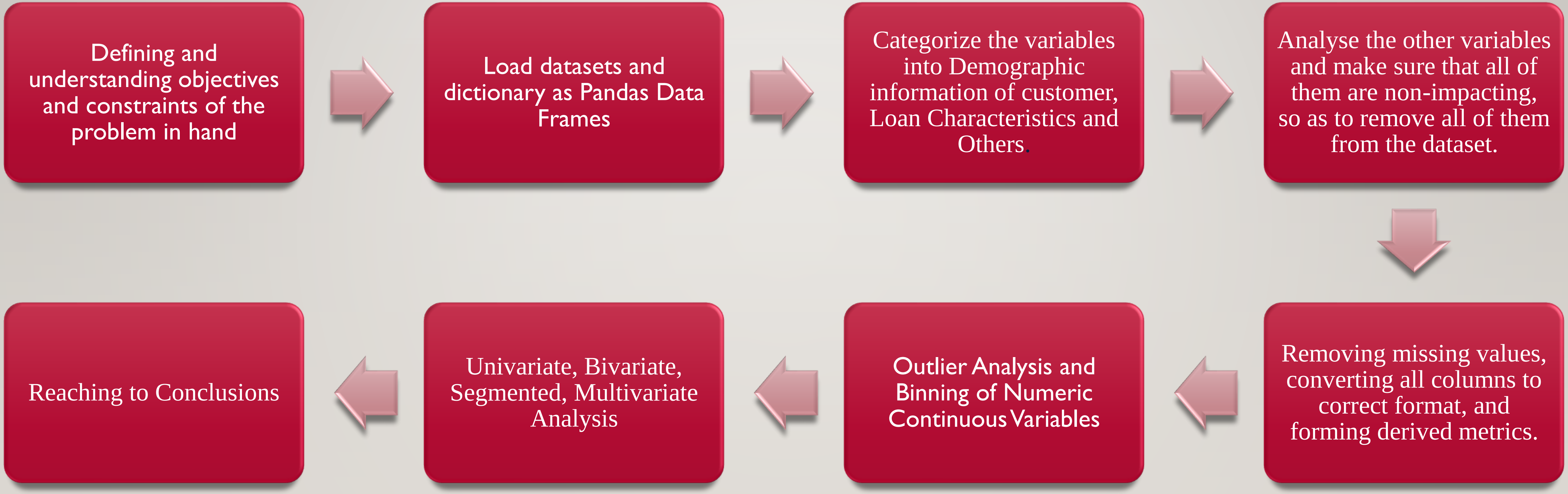
Key Data Inputs:

Dataset	Description	File Type	Size
Loan	Complete loan data for all loans issued through the time period 2007 t0 2011.	csv	39717, 111
Data Dictionary	Describes the meaning of Loan Dataset variables	xlsx	115, 2

Key Data Outputs:

1. Understanding the driving factors(variables) behind loan default
2. Understanding the contribution of each factors(variables) towards loan default

PROBLEM SOLVING STRATEGY



Defining and understanding objectives and constraints of the problem in hand

Load datasets and dictionary as Pandas Data Frames

Categorize the variables into Demographic information of customer, Loan Characteristics and Others.

Analyse the other variables and make sure that all of them are non-impacting, so as to remove all of them from the dataset.

Removing missing values, converting all columns to correct format, and forming derived metrics.

Outlier Analysis and Binning of Numeric Continuous Variables

Univariate, Bivariate, Segmented, Multivariate Analysis

Reaching to Conclusions

Analysis

➢ Data Understanding , Preparation , Cleaning and Binning

1. Checked for Columns with all Values NaN , had single variable , high missing values and ‘0’s, dropped all of them
2. Removed Non Impacting columns(Such as Customer Behavioural Variables, Demographics by referring to Data Dictionary definition for each variable), either they are identifier ,location, descriptive , post approval features , irrelevant to loan approval
3. Checked and assigned correct data type for affected columns, like removed months from ‘Term’ , % from int_rate/revol_util , modified emp_length to int values by removing years , splitted the Issue_d to year and month column
4. Checked Earliest Cr Line Month , discovered years less than 69 exists and cause Pandas to read incorrectly, Year below 69 converted as 20xx , applied function to convert them to 19xx
5. Applied Outlier analysis, removal and Binning on Continuous variable and derived new discrete categorical columns to do a meaningful analysis

Columns with All Values ‘NAN’	Columns with High Null Values and Non Impacting	New Columns Created
mths_since_last_major_derog, annual_inc_joint, dti_joint, bc_util, verification_status_joint, tot_coll_amt, tot_cur_bal, open_acc_6m, open_il_6m, open_il_12m, open_il_24m, mths_since_rcnt_il, total_bal_il, il_util, open_rv_12m, open_rv_24m, max_bal_bc, all_util, total_rev_hi_lim, inq_fi, total_cu_tl, inq_last_12m, acc_open_past_24mths, avg_cur_bal, bc_open_to_buy, mo_sin_old_il_acct, mo_sin_old_rev_tl_op, num_op_rev_tl, num_actv_bc_tl, mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl, mort_acc, mths_since_recent_bc, mths_since_recent_bc_dlq, mths_since_recent_inq, mths_since_recent_revol_delinq, num_accts_ever_120_pd, num_actv_rev_tl, num_bc_sats, num_bc_tl, num_il_tl, num_rev_accts, num_rev_tl_bal_gt_0, num_sats, num_tl_120dpd_2m, num_tl_30dpd, num_tl_90g_dpd_24m, num_tl_op_past_12m, pct_tl_nvr_dlq, percent_bc_gt_75, tot_hi_cred_lim, total_bal_ex_mort, total_bc_limit, total_il_high_credit_limit	emp_title,desc,mths_since_last_delinq,mths_since_last_record,next_pymnt_d’,id,member_id,url,desc,title,zip_code,funded_amnt,delinq_2yrs,total_pymnt,total_pymnt_inv,total_rec_int,total_rec_late_fee,total_rec_prncp,recoveries,collection_recovery_fee,out_prncp,out_prncp_inv, revol_bal,next_pymnt_d,last_pymnt_amnt,last_pymnt_d,chargeoff_within_12_mths,last_credit_pull_d’,pymnt_plan,initial_list_status,collections_12_mths_ex_med,policy_code,acc_now_delinq,delinq_amnt,tax_liens, application_type	**Split Issue_d →** To Issue,_month, Issue_year columns **New Categorical Variable post applying Binning:** Loan_amnt_range , funded_amnt_range, int_rate_range , annual_inc_range ,dti_range , funded_amnt_inv_range ,Installment

Analysis

1. **Univariate Analysis:**
    1. Run through each categorical and continuous variable to analyse the impact to Default count with respect to overall loan status
    2. Select the influential variable based on the positive/negative relation to the Default status
    3. Select the variable infer loan characteristic and help in identifying the default
2. **Bivariate & Segmented Analysis**
    1. Influenced Univariate Variables are combined in to two Variable feature set and analyse the Default count
    2. Compared the Binning Discrete Categorical Variable to the actual Categorical Variable to do segmented analysis to see the relation and pattern towards Loan Default
    3. Compared the actual continuous variable to the actual categorical variable to analyse the spread of the data
    4. Combinations of Variables are identified by applying the analytical sense and had business relations
3. **Segmented Analysis**
    1. Plotting two different features with respect to loan status
    2. Checking whether the combination of features have an influence on defaulter percentage
4. **Multivariate Analysis**
    1. Identified influencing variables after univariate, bivariate analysis are compared to each other to see their influences / relation between them to Loan Default
    2. Applied Correlation on actual continuous variable to see the influence and to derive the final driver variables

Results

Univariate Analysis

Variables that describe something and do not deal with relationships , rather describe something

Analysed Categorical Variables

a) Loan Status -- 14% is Default , 83% is fully paid percentage of customers

b) Purpose – Customer applying for Debt Consolidation, Credit Card , Home Improvement are high in numbers

c) Grade – High volume of Loan Applicant increase from Grade G to A

d) Home Ownership – Mortgage , Rent Customers loan applications are high , compare to Own

e) Term – Higher loan application for lesser term

Analysed Continuous variable by binning into discrete categories and see the pattern.

a) Loan Amount – Customers applying for Medium(5000 - 15000) range are more

b) Funded Amount Inv – Customers in the range Medium(5000 – 15000) is high

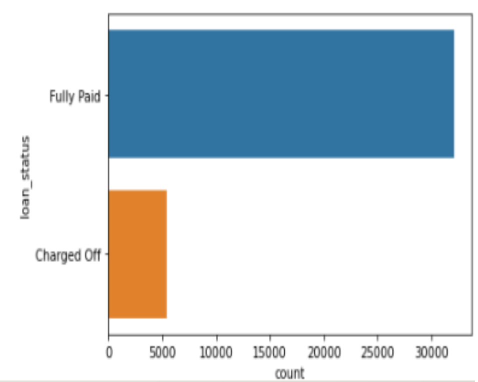
c) Interest Rate – Loan Applicants are in the Medium Range of Int(between 8% to 15%)

d) Annual Inc -- Medium (40000 to 80000) and Very High Category (>80000) customers are high

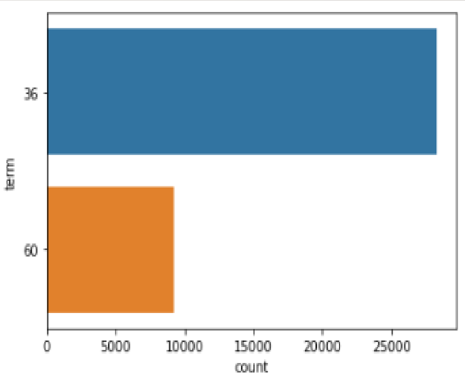
e) Dti – Loan Applications are more for higher dti , but also not much difference to lower dti value

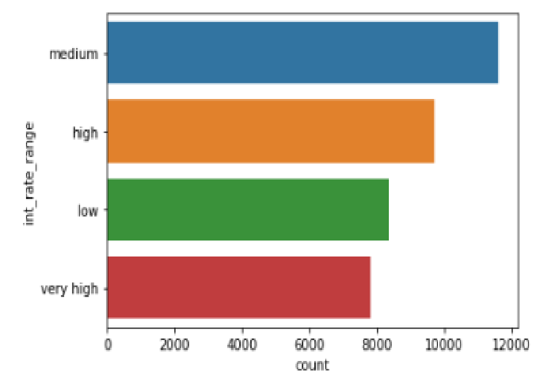
Outcome : By Individual Variable inference, most of the columns showed a pattern with respect to high loan application for specific categories , range of values . It did not show relationship in reference to Default or Loan Status.

Loan Status

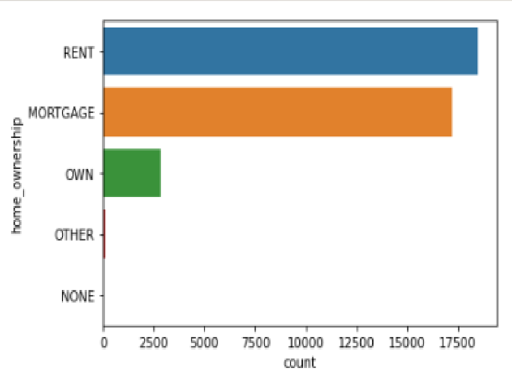


Term

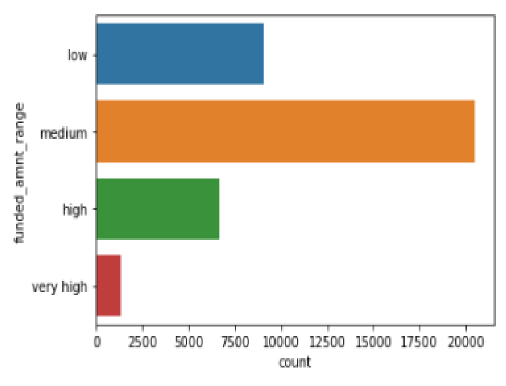




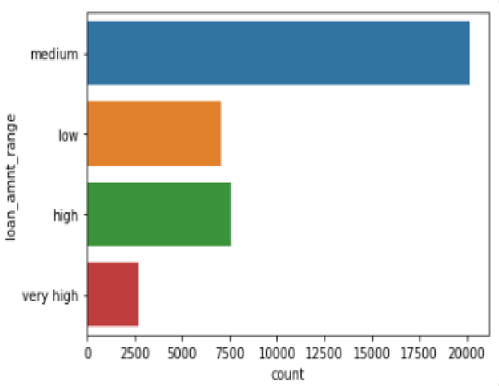
Interest Rate



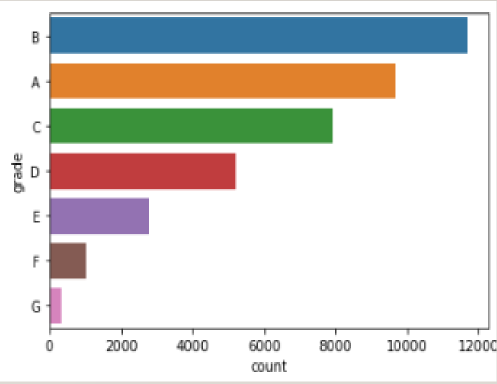
Home Ownership



Funded Amount



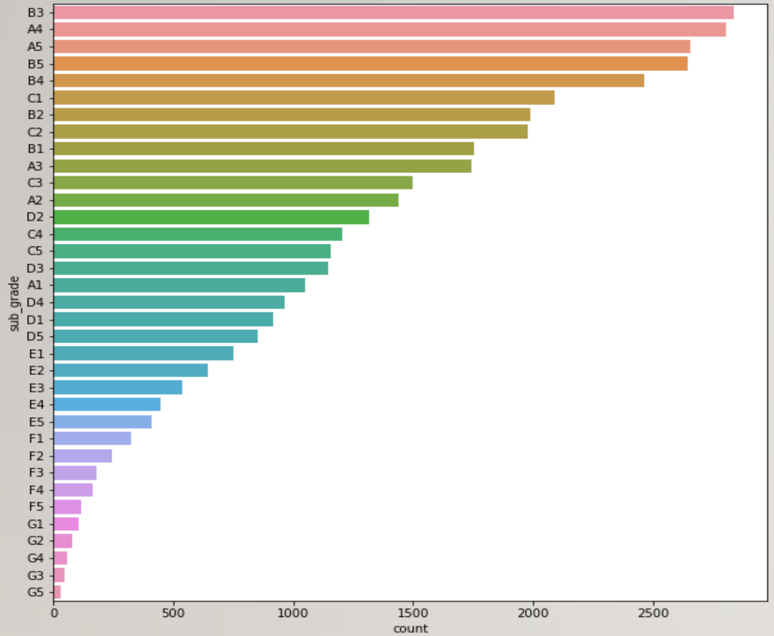
Loan Amount



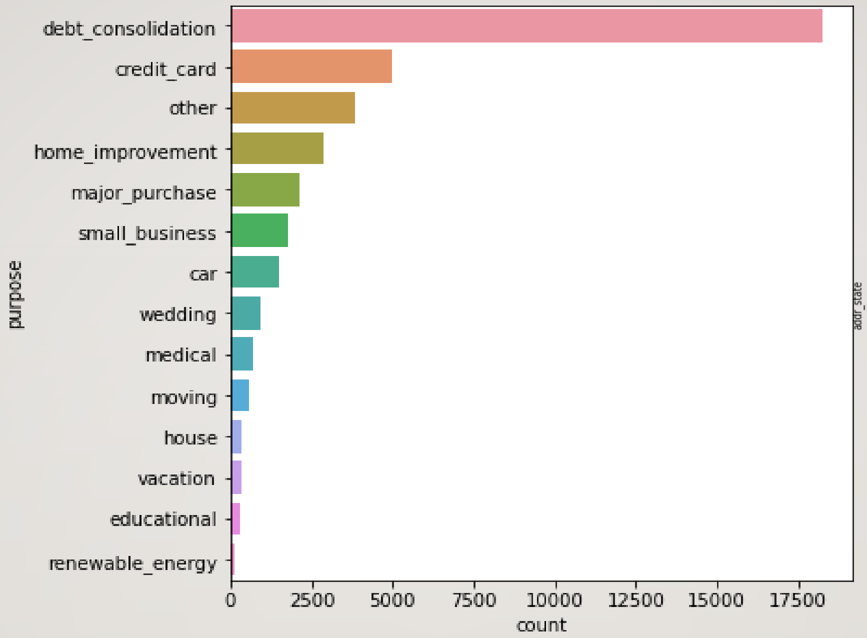
Grade

Univariate Analysis Visualizations

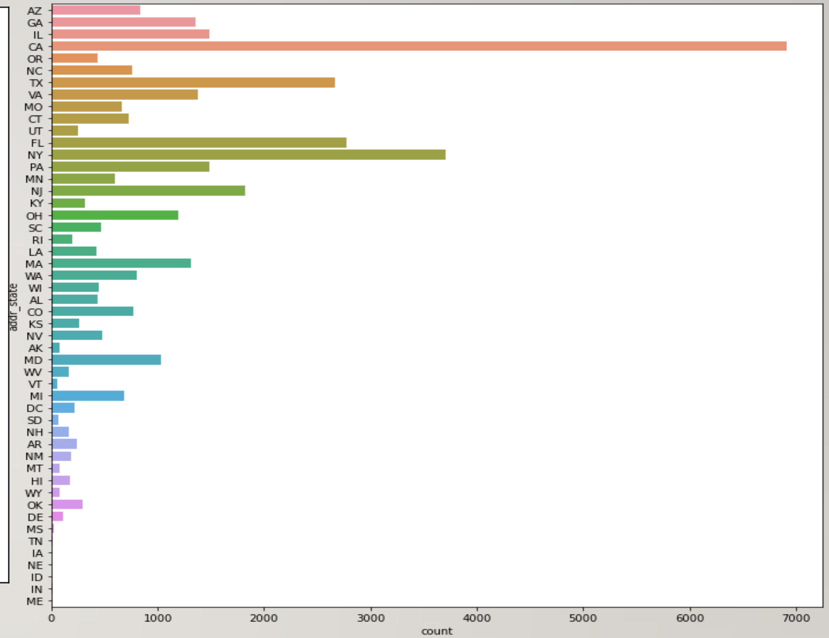
Results



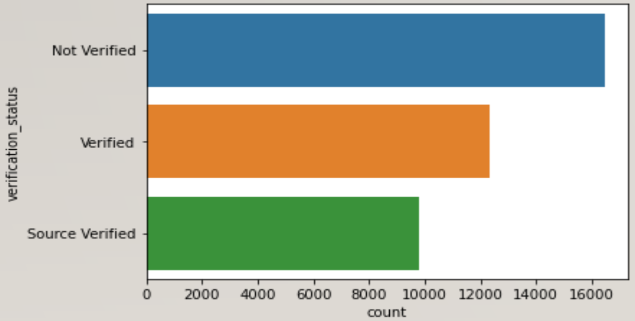
Subgrade



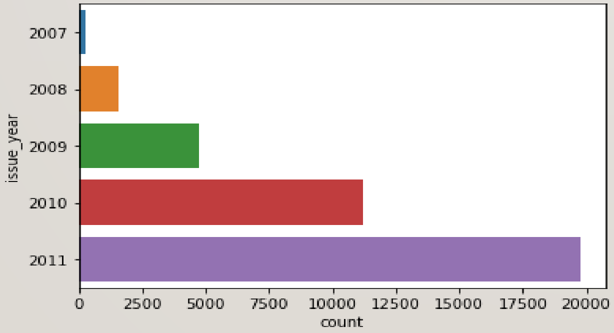
Purpose



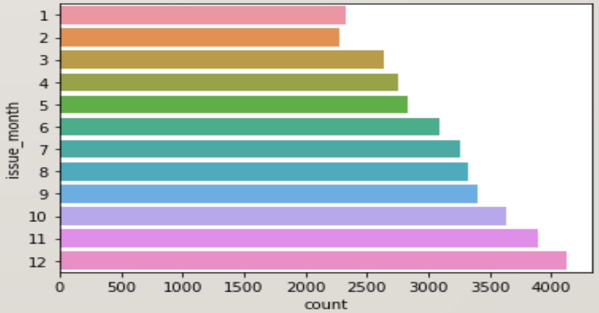
Address State



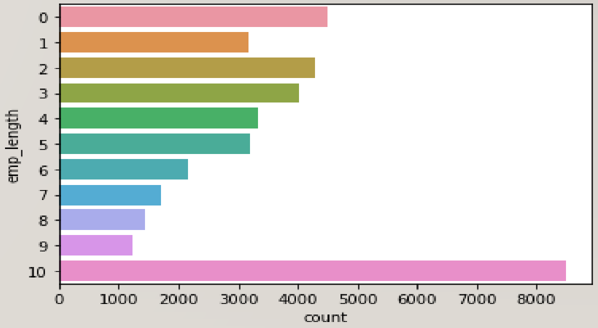
Verification Status



Issue Year



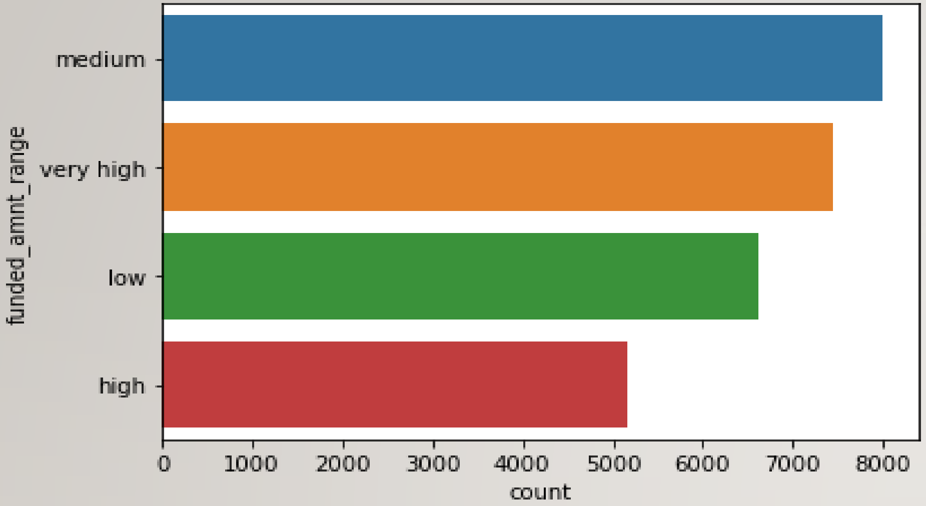
Issue Month



Employee Length

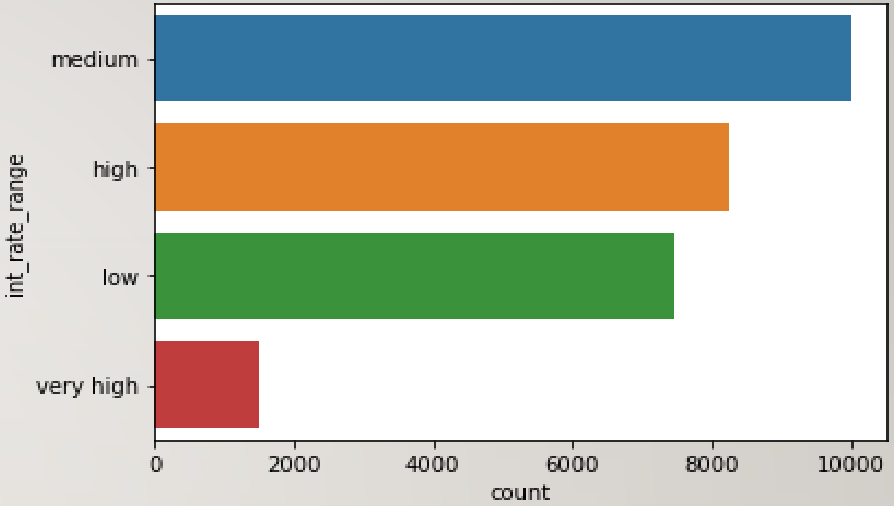
Univariate Analysis Visualizations

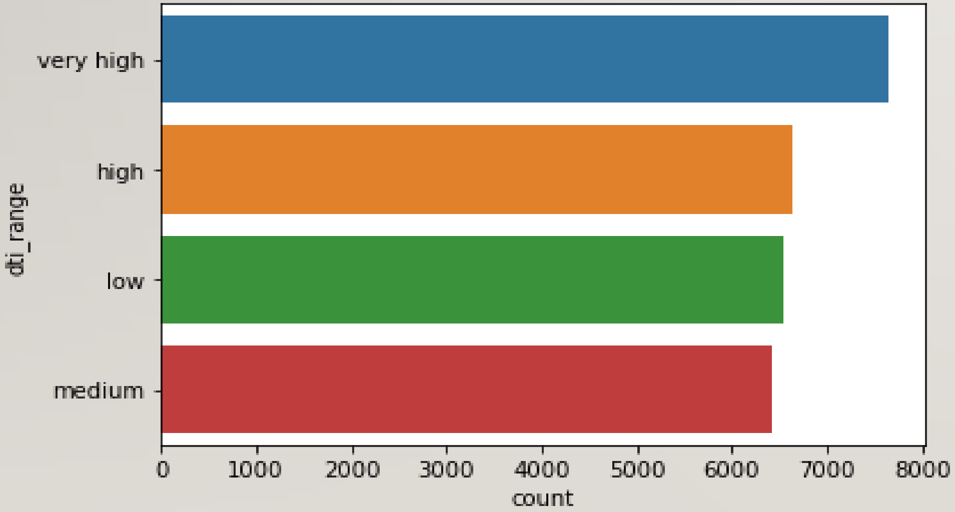
Results



Funded Amount

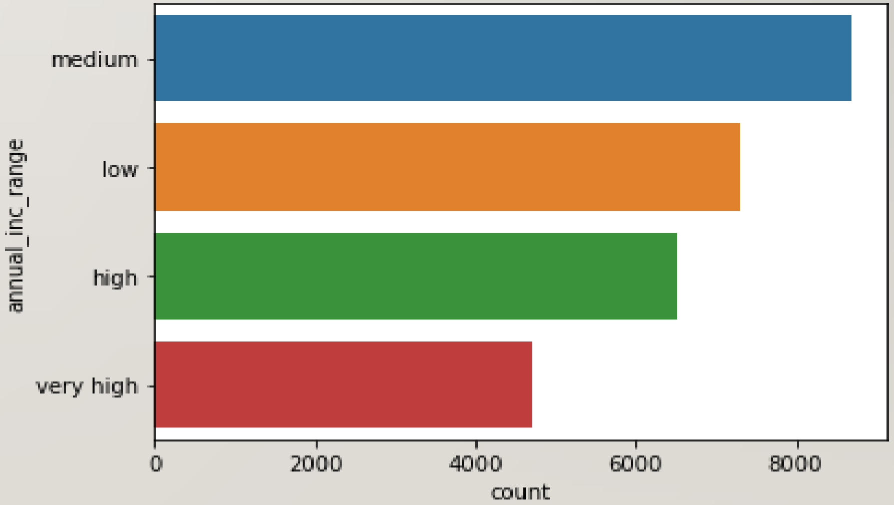
Interest Rate





DTI

Annual Income



Univariate Analysis Visualizations – Binned Continuous Variables

Results

Bivariate Analysis

Variables are compared with respect to Default Loan Status to derive the inference

Analysed Categorical Variables [‘term’,’grade’,’purpose’]

a) Grade –Default increase from A to G, G is grade with lower score as per the Lending Club categorization (assumption : riskiness of the loan) and A be the Superior Customer

b) Term – Longer Term customer default more, they are the risky customers

c) Purpose – Small Business high default the most , then renewable energy and house

Analysed Continuous variable by binning into discrete categories and see the pattern.

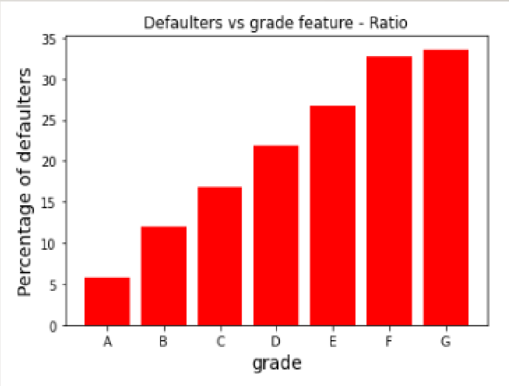
a) Loan Amount – Higher the loan Amount, higher the default rate and risky customers

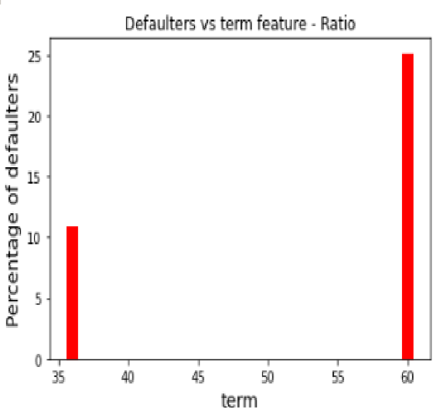
b) Funded Amount Inv – Higher the Amount , higher the default rate and risky customers

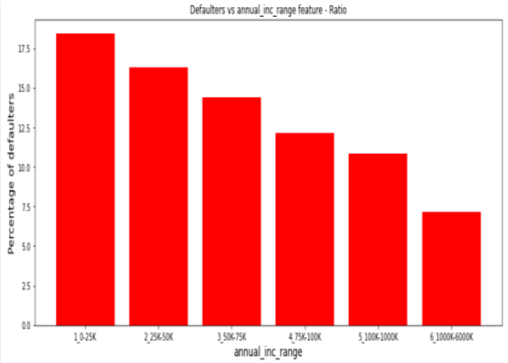
c) Interest Rate – Higher the Int Rate default more as expected

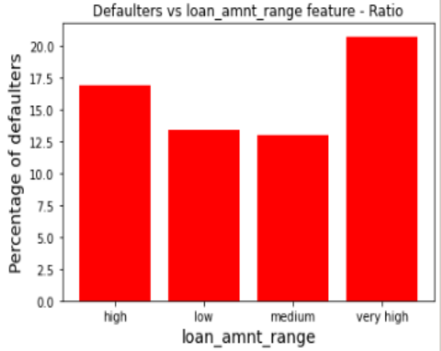
d) Funded Amount Inv , Instalment Loan Amount ,

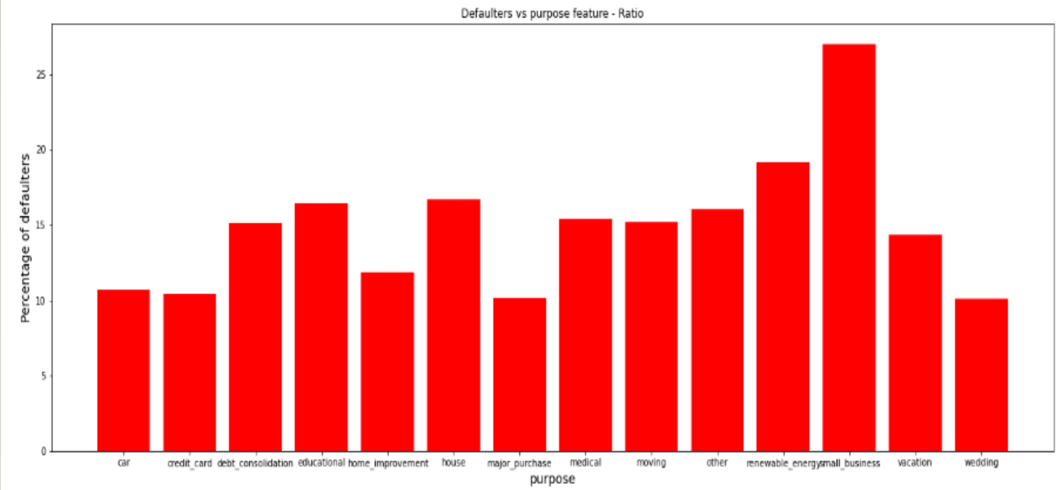
e) Emp Length ,Issue year/month ,earliest_cr_line , total acc , inq_last_6_month ,Verification Status ,home Ownership had flat pattern with respect to Default, hence they are non impacting and removed form further analysis

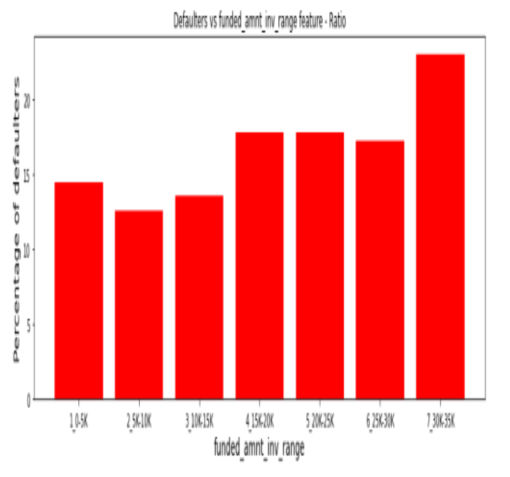


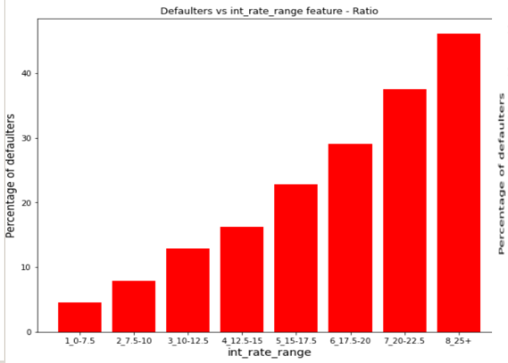


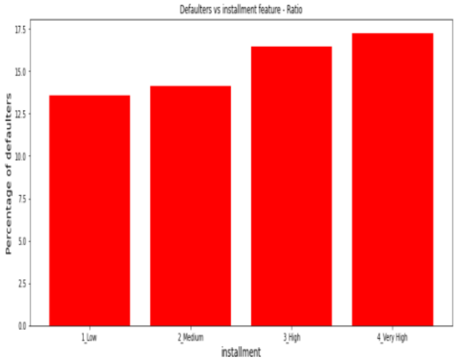






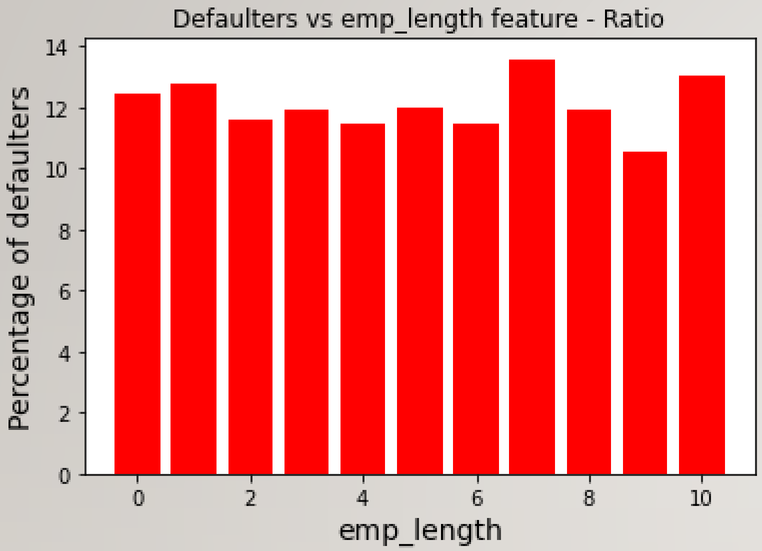


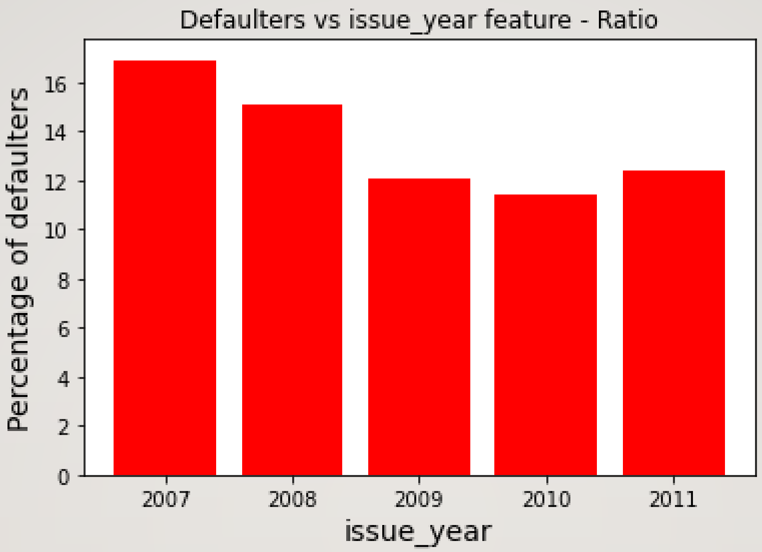


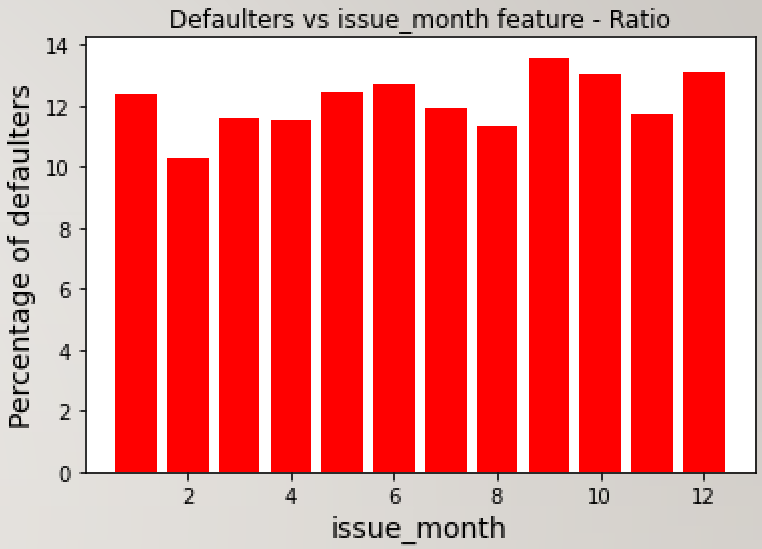


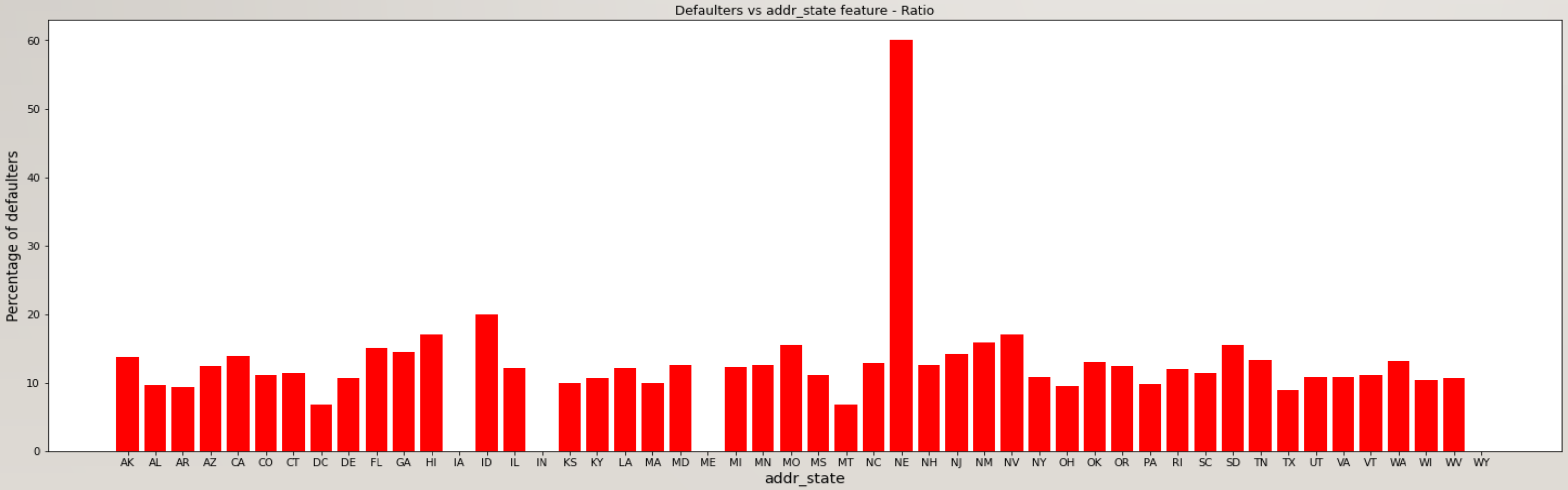
Bivariate Analysis Visualizations

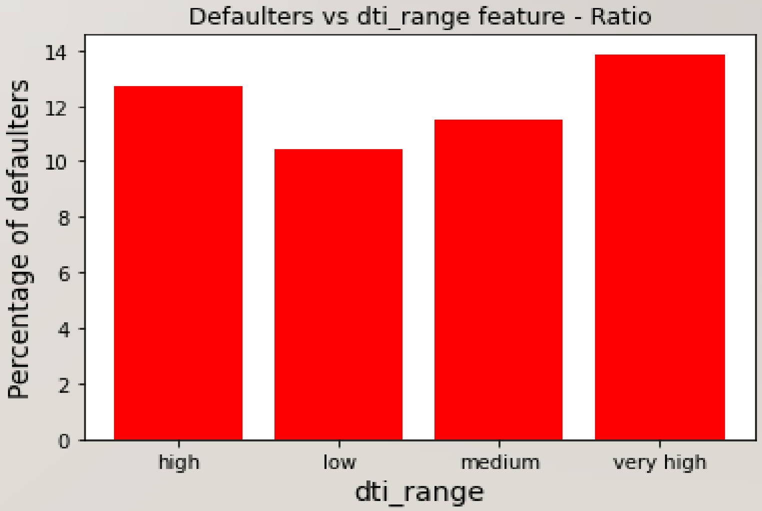
Results


Defaulters vs emp_length feature - Ratio


Defaulters vs issue_year feature - Ratio


Defaulters vs issue_month feature - Ratio


Defaulters vs addr_state feature - Ratio


Defaulters vs dti_range feature - Ratio

Bivariate Analysis Visualizations

Results

Outcome of Bivariate Analysis:

•Since Grade and Sub-Grade have similar trends only one variable is needed for further analysis. Hence considering Grade and Removing Sub-Grade.

•Sice the variables like home_ownership , verification_status , emp_length, issue_year, issue_month, show a flat trend with respect to Defaulter Ratio, it's safe to conclude that they do not influence a loan becoming Default. Thus, removing them from further analysis.

•The variables grade, term, loan_amnt, int_rate, dti show a positive linear relationship with the percentage of Defaulters.

•However, annual_inc shows a negative linear relationship with the percentage of Defaulters.

•Only Small Business purpose has High default percentage.

•Only the addr_state NE has High default percentage.

Features considered for next analysis

grade, term, loan_amnt, int_rate, dti, annual_inc, purpose, addr_state

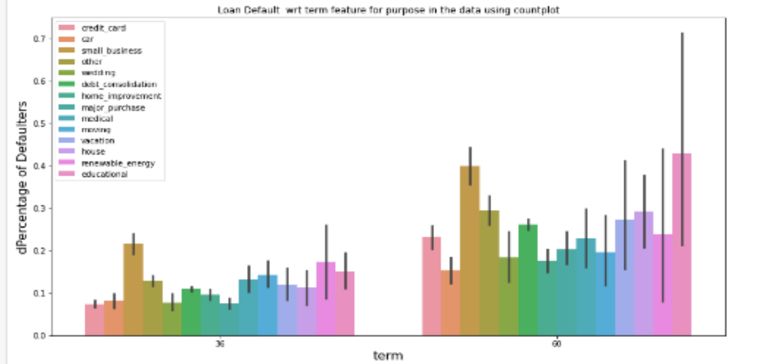
Results

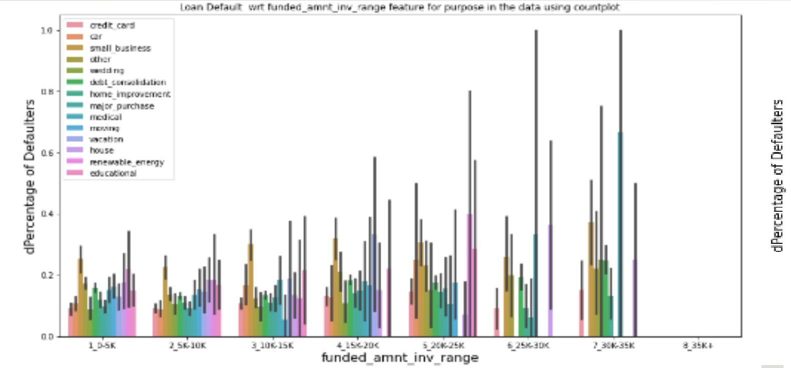
Segmented Analysis:

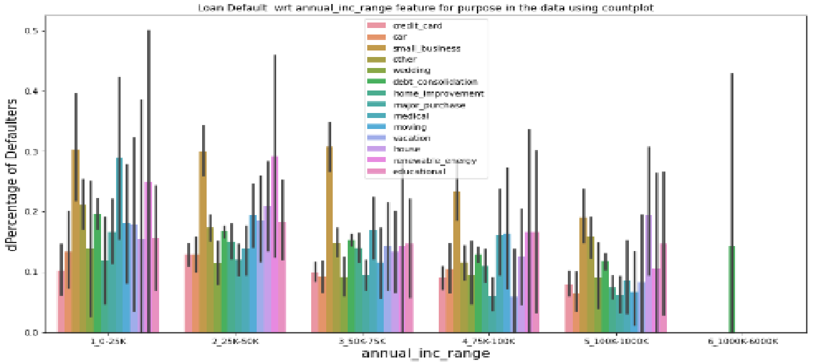
Segmenting the variable based on the categories/range , analyzing their behaviour with other variables in respect to Loan Default.

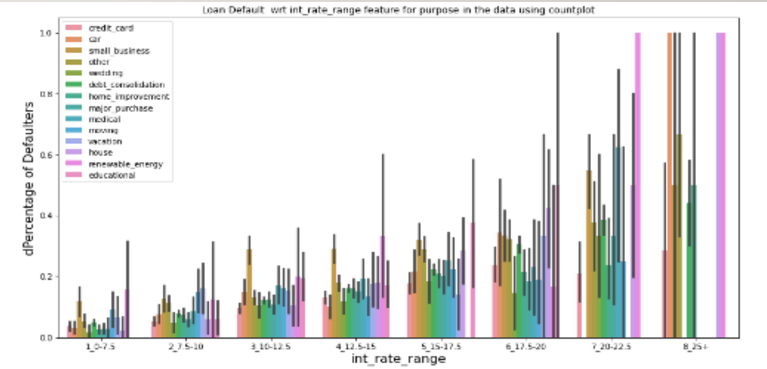
Outcome :

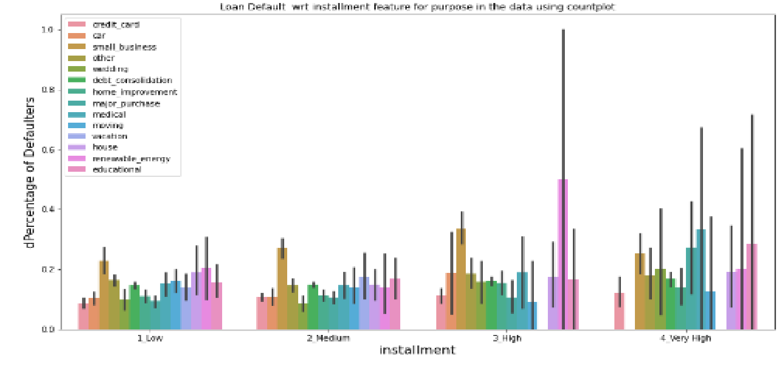
We can conclude that the relationships of variables identified in Bivariate analysis is preserved even when analysed together with other categorical variables

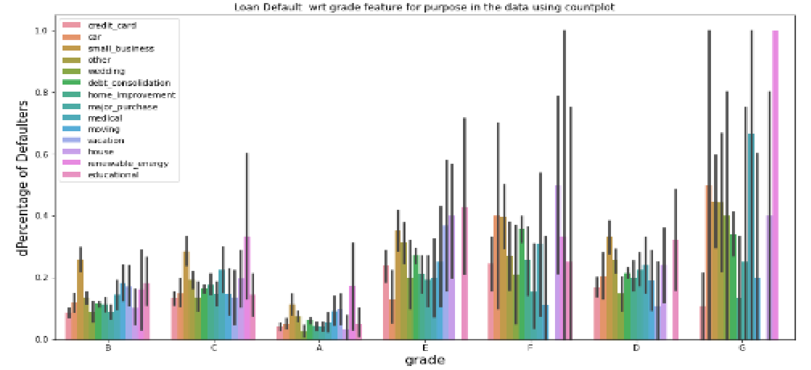


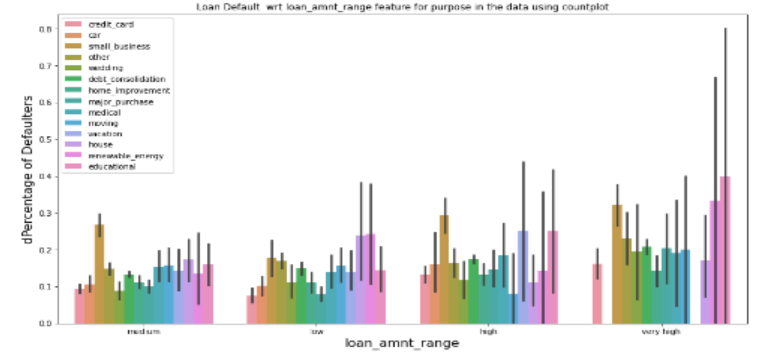






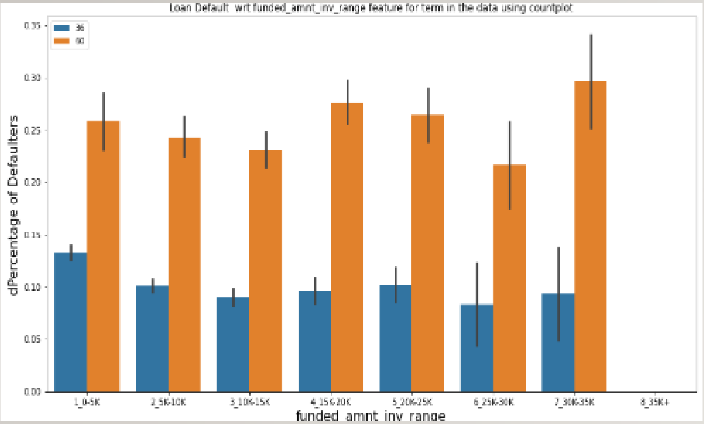


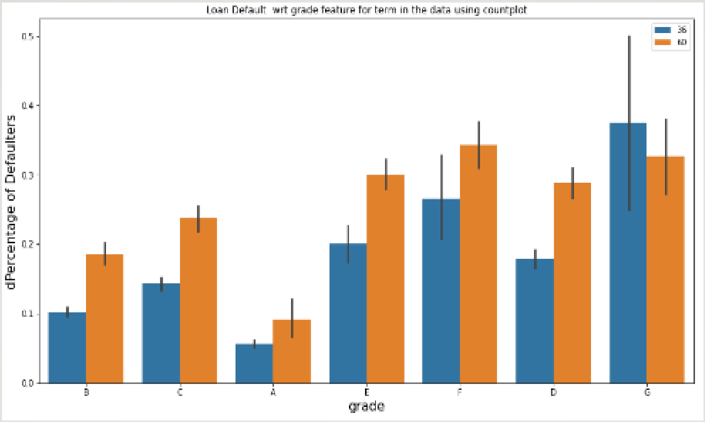


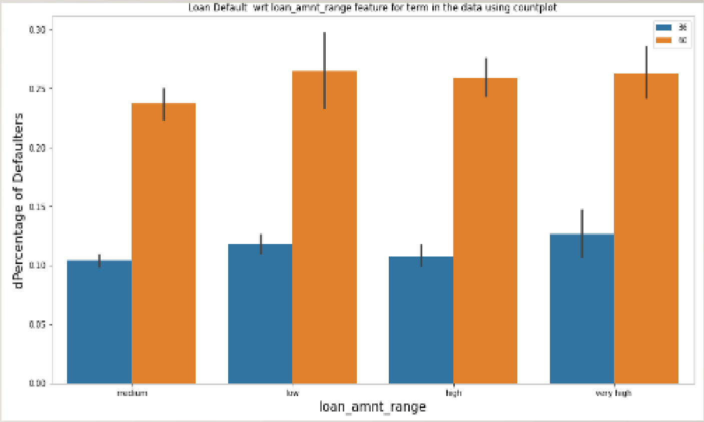


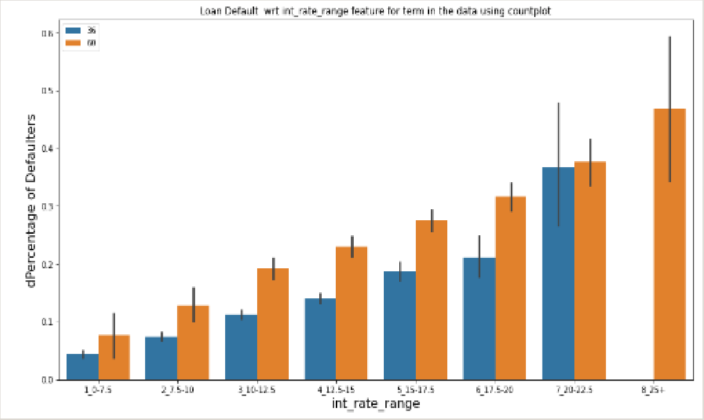
Segmented Analysis Visualizations

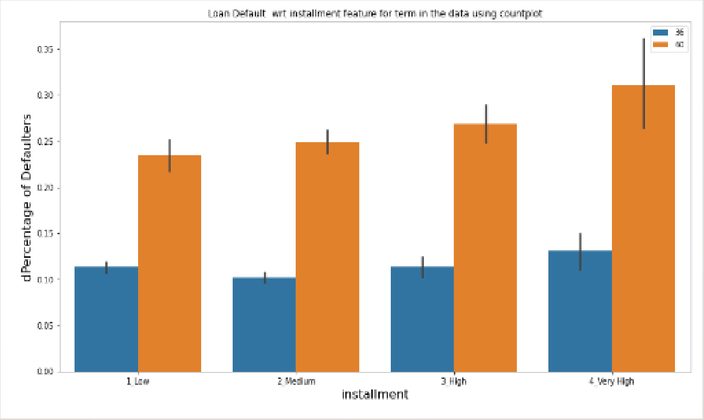
Results

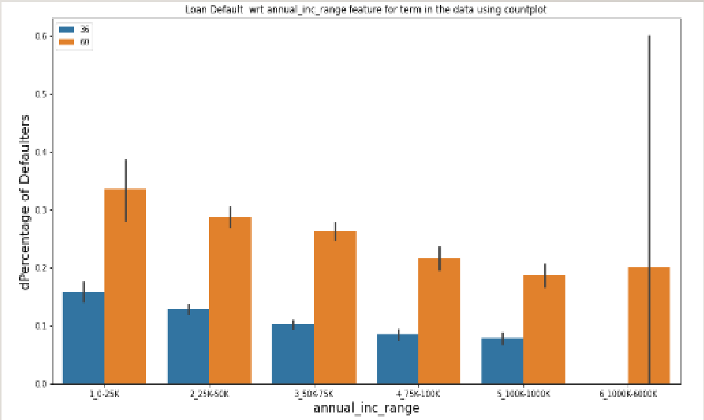






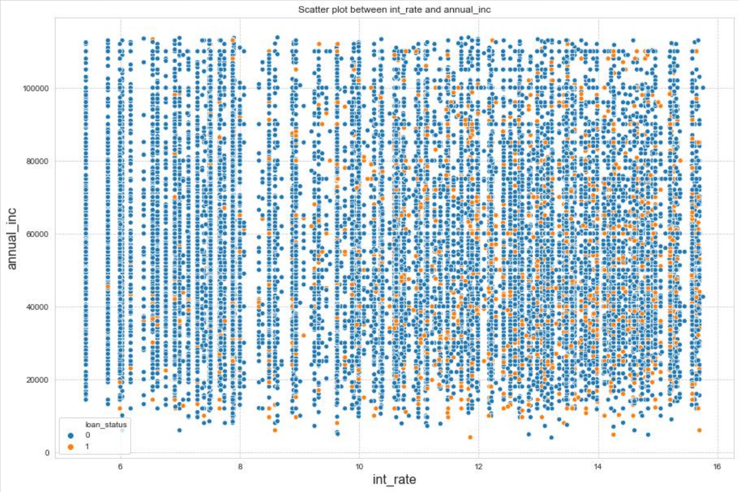


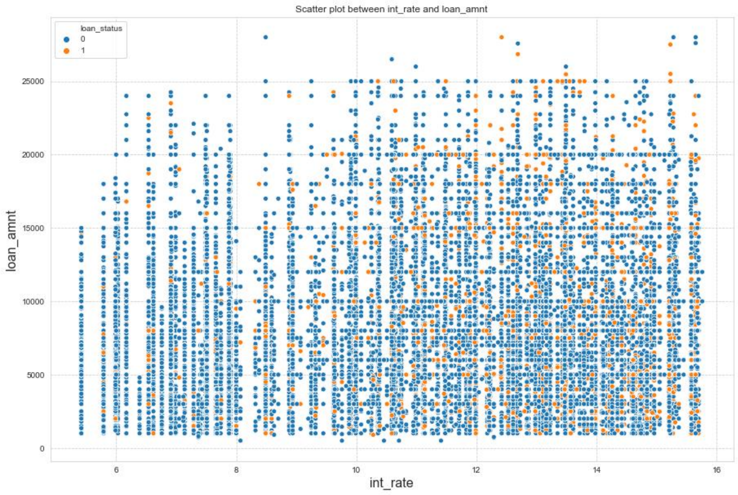


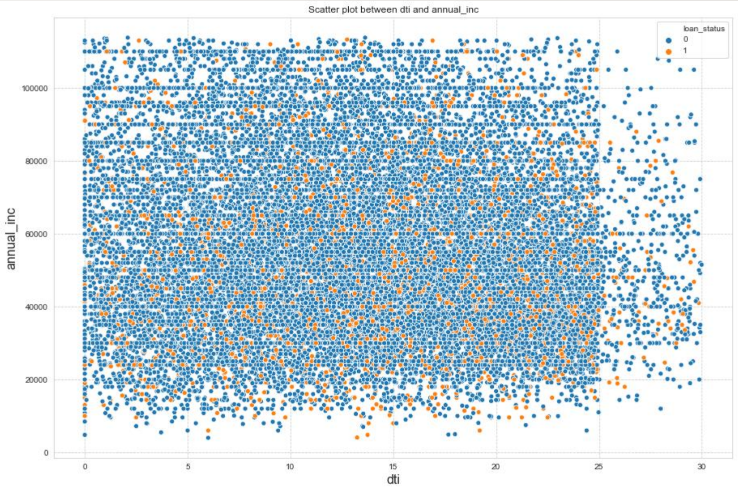


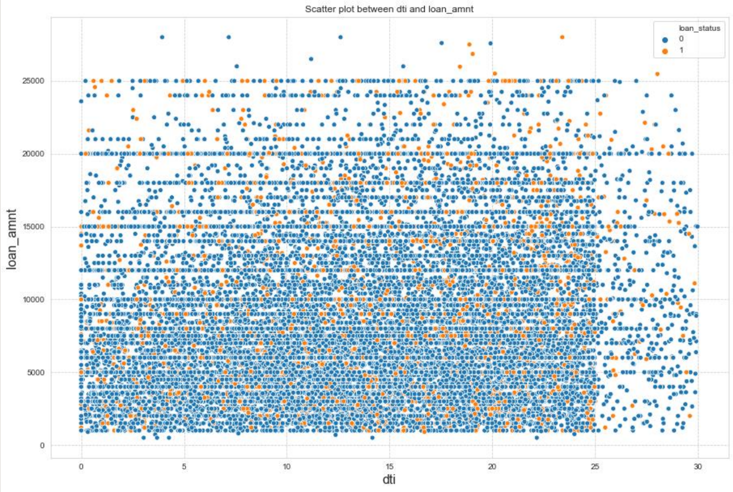
Segmented Analysis Visualizations

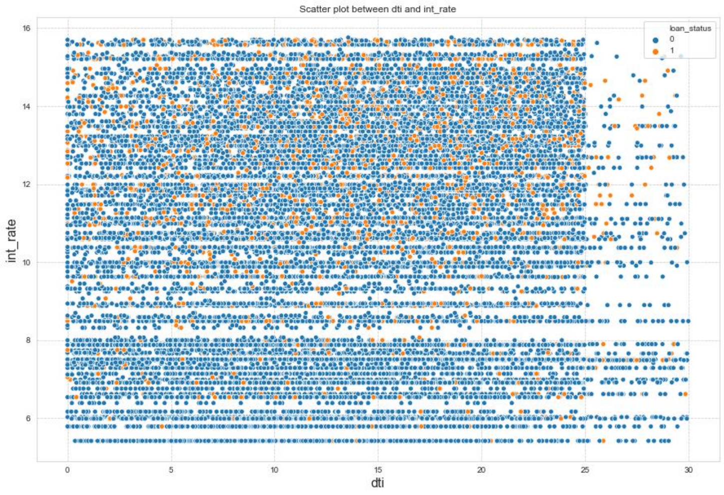
Results

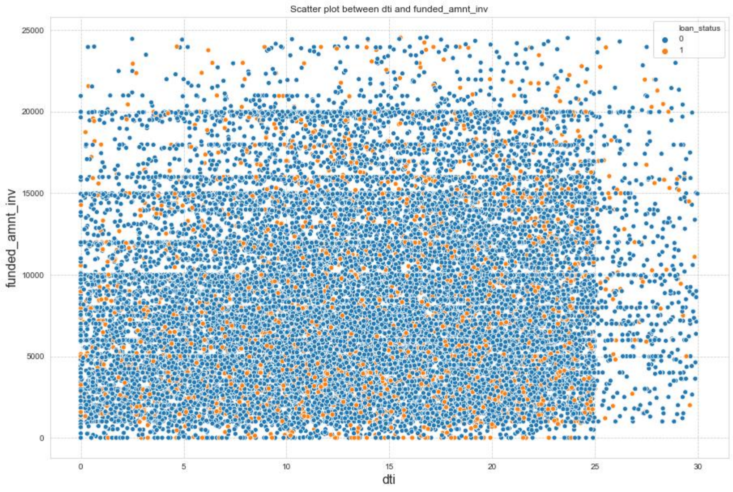










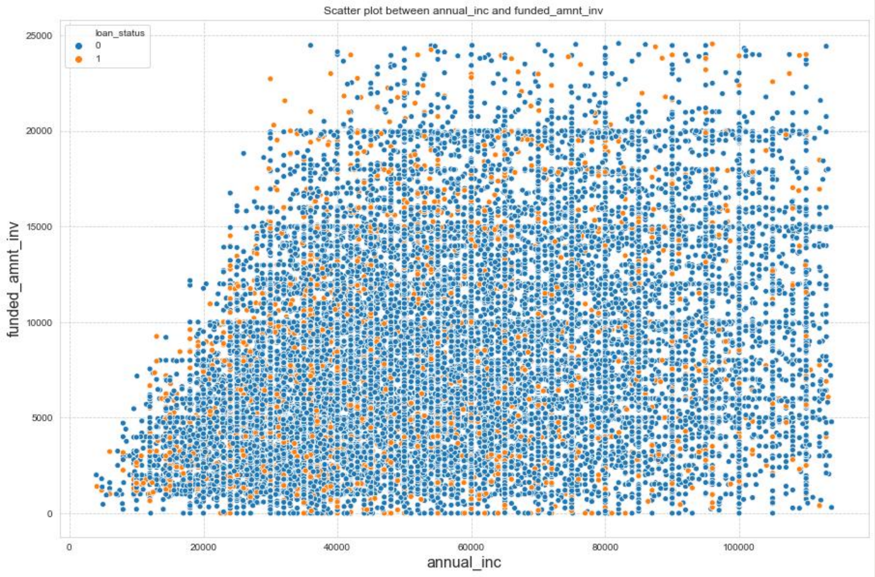


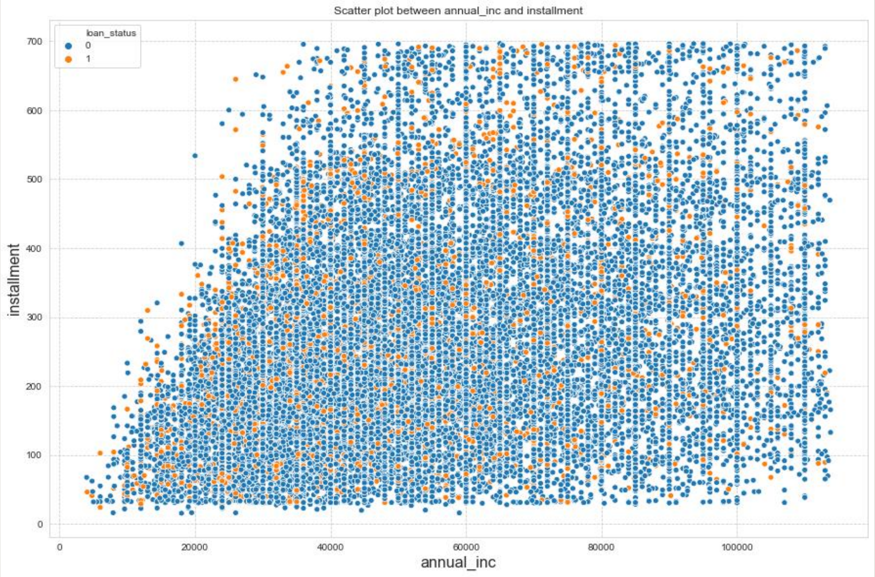
Segmented Analysis Visualizations

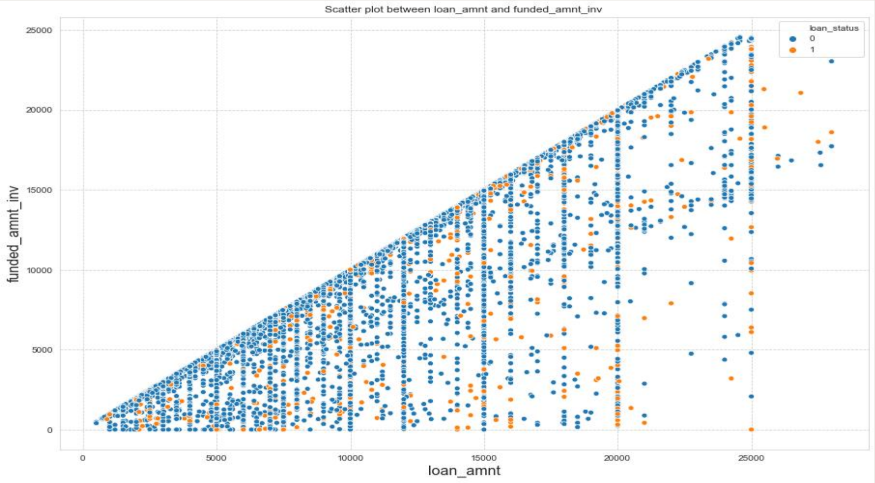
Results

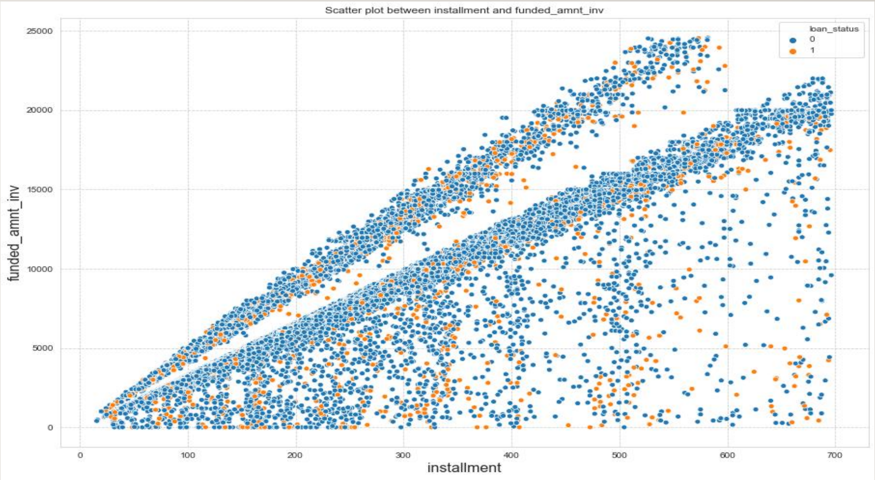
Outcome of Segmented Analysis:

- The variables annual _inc, funded_amnt_inv, int_rate, term, grade, purpose are driving variables in determining defaulters.
- The variables loan_amnt, dti in conjunction with other features don't have an influence on driving variables and hence can be disregarded.

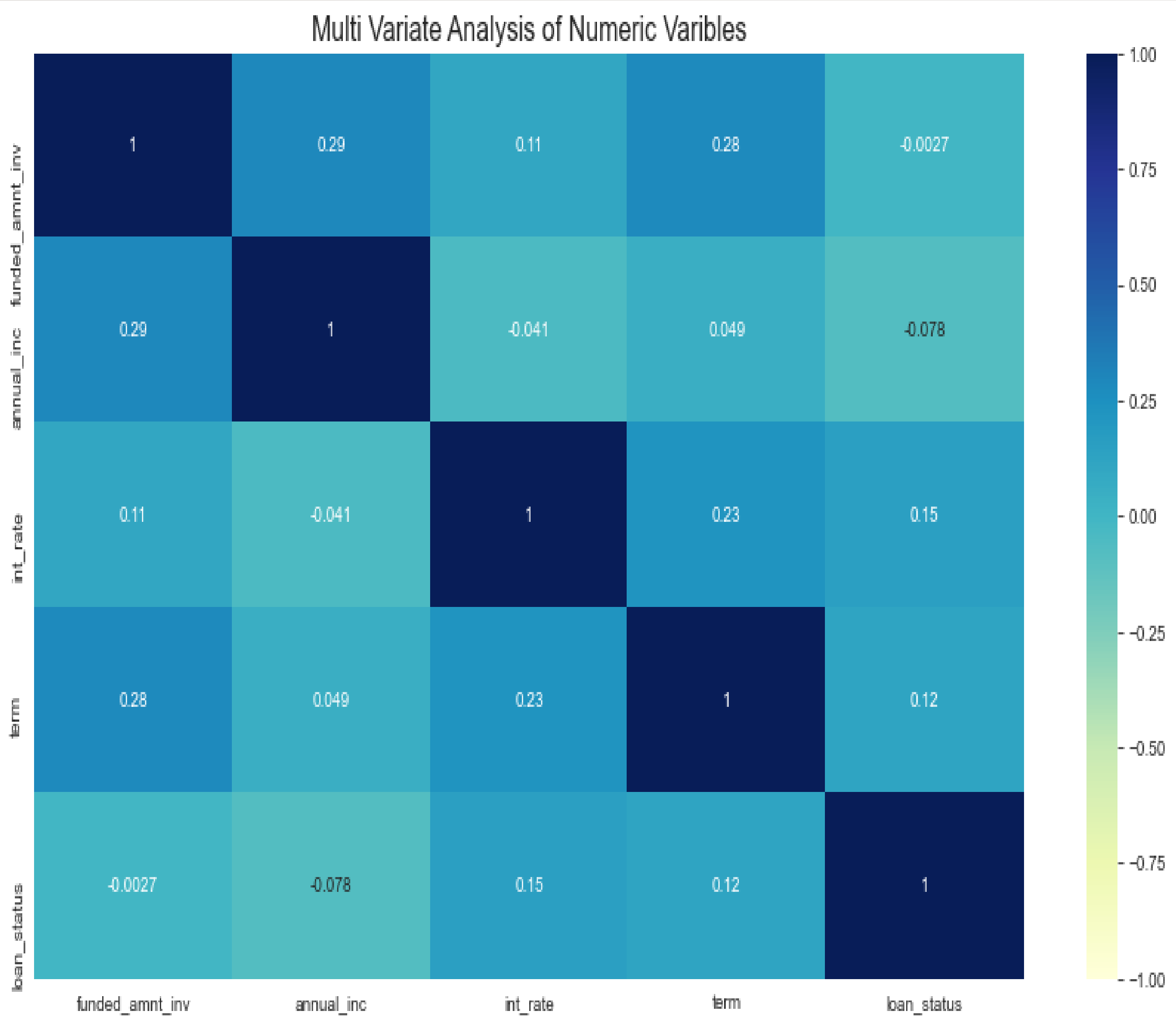

Scatter plot between annual_inc and funded_amnt_inv


Scatter plot between annual_inc and installment


Scatter plot between loan_amnt and funded_amnt_inv


Scatter plot between installment and funded_amnt_inv

Segmented Analysis Visualizations



RESULTS

Multivariate Analysis:

Plotted heat map to analyze all shortlisted variables together to see the influence of them on loan status and relationship between the driver variables.

Outcome :

No strong correlations are detected, hence all variables selected are equally important driving variables

CONCLUSIONS

- **Driver Variables: annual _inc, funded_amnt_inv, int_rate, term, grade, purpose, addr_state**
- Need to manage the above driver variables to reduce influence towards Default and maximize the values to have a potential fully paid customer.
- annual _inc is negatively correlated with defaulter percentage, hence to minimize defaults, appropriate measures should be taken before sanctioning a loan from a user of lower income range.
- funded_amnt_inv is negatively correlated with defaulter percentage. That is, higher the funded_amnt_inv, lower is the chance of loan default.
- int_rate and defaulter percentage are postively correlated, so there ia higher risk of loan becoming default when the interest rate is high.
- Loans given for a longer term has more chance of becoming default.
- Loans of higher grade have higher tendency to be default.
- If the purpose of loan is small business, then stringent measures should be taken to reduce the probability of loan becoming default.
- Also, if the loan applicant is from NE state, strict evaluations needs to be done regarding loan sanctioning, as historically this region has the greatest number of default loans.