Lending club data analysis

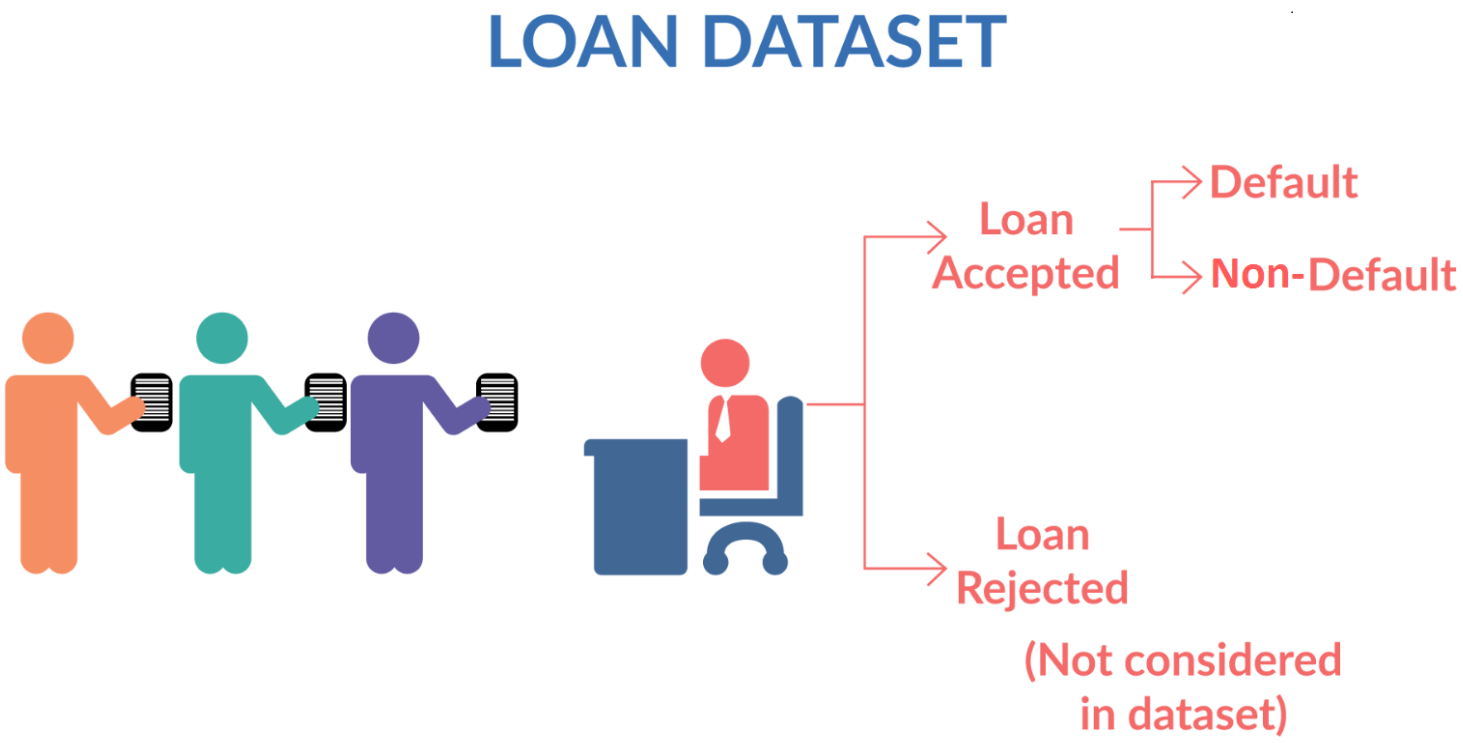
Problem Statement

- Lending club company 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.

- Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

- In this case study we are going to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Loan dataset



Approach

- Data understanding
- Data cleaning
- Data pre-processing
- Handling outliers
- Univariate analysis
- Bivariate Analysis
- Multivariate Analysis

Data understanding

- We have 39717 entries and 111 different columns in our dataset
- Shape - 39717 *111
- Our main aim is to find the important features which are responsible for the loan fault among these 111 features.
- Next step is to clean our data

Data cleaning

- ***Step 1 :*** Check for the null values in the data

  * While checking the null values we found that nearly 58 column contains more than 30% as null values

  * We remove these 58 columns as they don't have enough data to do our analysis.

- ***Step 2 :*** Check for columns with only one class/category

  * We found that 9 columns have only one category for all the entries

  * The columns which has only one class – {pymnt_plan, initial_list_status, collections_12_mths_ex_med, policy_code, application_type, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt, tax_liens }

  * We remove these 9 columns as they don't have any different categories and it doesn't help in our analysis

- ***Step 3 :*** Checking for missing values in other columns and removing those values
- ***Step 4 :*** Drop unwanted columns

  * Personal info columns like
  { id, member_id, emp_title, url , title, zip_code, addr_state } does not provide any valuable information regarding the loan default

  * Behaviour details of the member who applied loan like - {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,last_credit_pull_d, pub_rec_bankruptcies, revol_bal, delinq_2yrs} are removed

At the end of data cleaning we only have 21 columns for our analysis – **['loan_amnt', 'funded_amnt_inv', 'term', 'int_rate', 'installment', 'grade', 'sub_grade', 'emp_length', 'home_ownership', 'annual_inc', 'verification_status', 'issue_d', 'loan_status', 'purpose', 'dti', 'earliest_cr_line', 'inq_last_6mths', 'open_acc', 'pub_rec', 'revol_util', 'total_acc']**

Data Pre-processing

- ***Step 1:*** Check the format and data type in each column of the dataset

  * We found that revol_util, int_rate has % symbol in the entries and is of object type. To format it we remove the % symbol and convert it to float type

  * We found the earliest_cr_line column is of string type (Jan-85) which is actually a date so we convert it into date time format

- ***Step2 :*** Filter the required data

  In loan_status column we have 3 classes

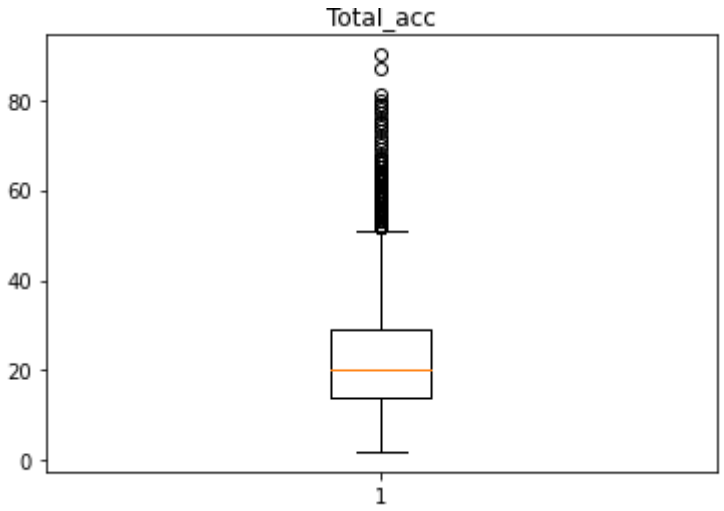
  * Fully Paid –(Non defaulter)
  * Charged Off –(Defaulter)
  * Current

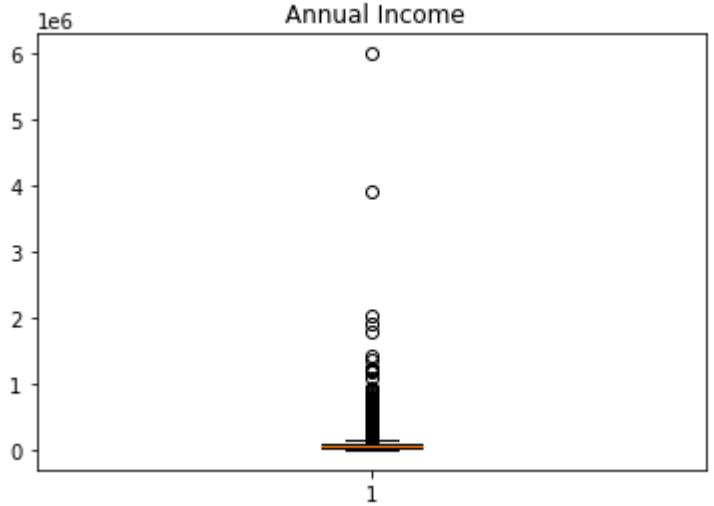
  In this Current represents the persons who are currently paying the loan and we can't categorize as default or non default we remove the data with Current as category

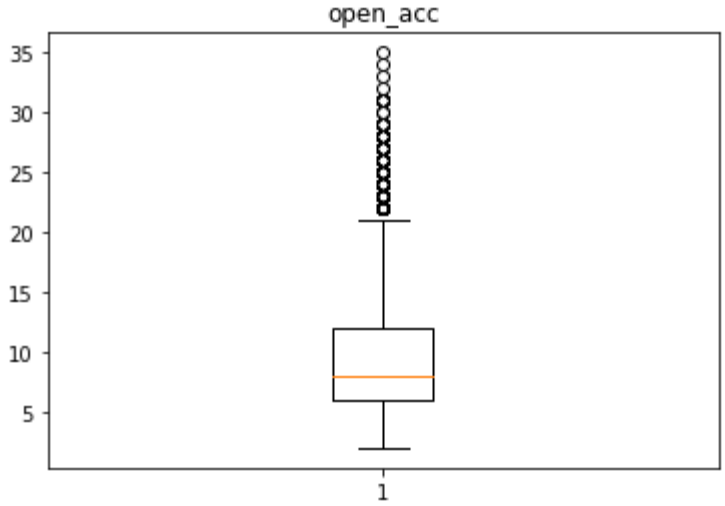
Handling Outlier

- An outlier is a data point that differs significantly from other observations.
- We need to remove the outliers in each columns else it will affect the data distribution of the entire column and will cause drastic change in the mean value
- annual_inc, total_acc, open_acc columns have outliers.
- We need to remove the outliers in these columns for further processing

Boxplot of total_acc, open_acc, annual_inc





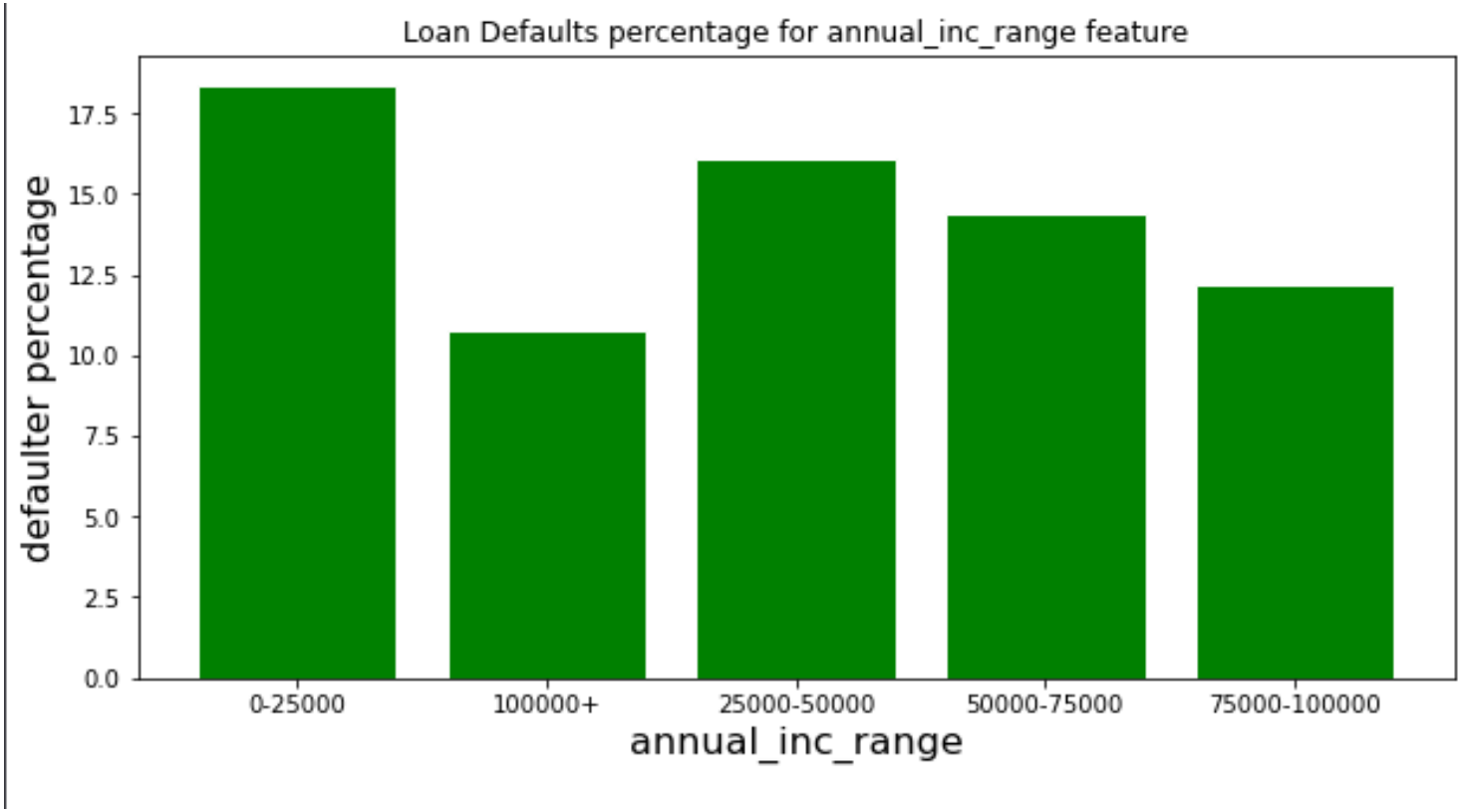


Binning/Grouping

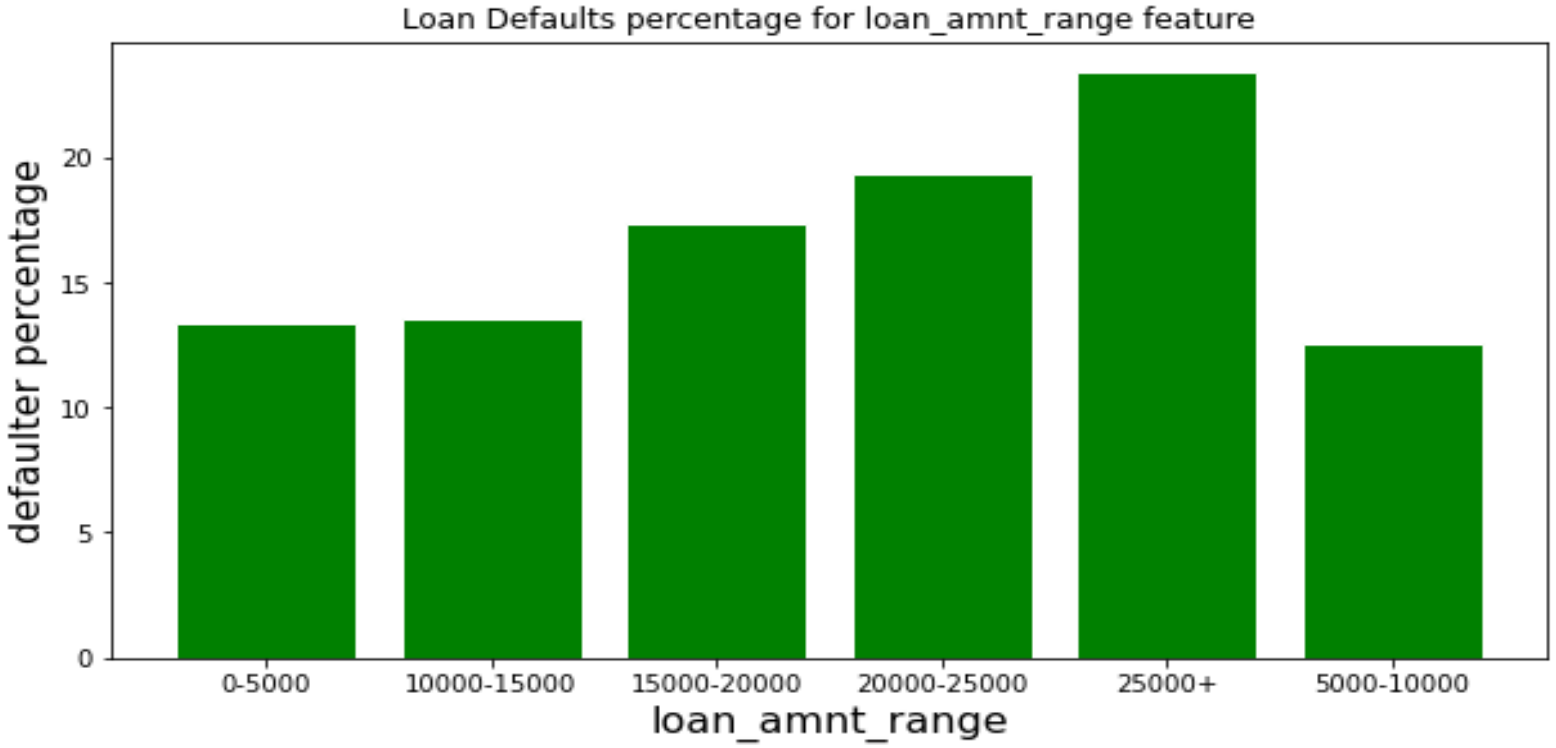
- Loan_amnt column has different values from 5000 to 35000. It would be difficult if we consider this data for analysis. So we will group the data based on the ranges as 0-5000, 5000-10000, 10000-15000, 15000-20000, 20000-25000 and 25000+
- Int_range column also has continuous values from 0 to 25 so we will group them as 0-7.5, 7.5-10, 10-12.5, 12.5-15, 15+
- Similary we will group emp_length column as 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 categories as it will make our analysis simple.
- Annual_inc also varies from 25000 to 200000 so we categorize them as 0-25000', '25000-50000', '50000-75000', '75000-100000', '100000+‘
- Installment also varies from 0 to 1000+ so we group them into four categories '0-200','200-600','600-1000','1000+‘
- Dti also varies from 0 to 25+ and we group them as '0-5%', '5-10%', '10-15%', '15-20%', '20-25%', '25%+’

Univariate analysis

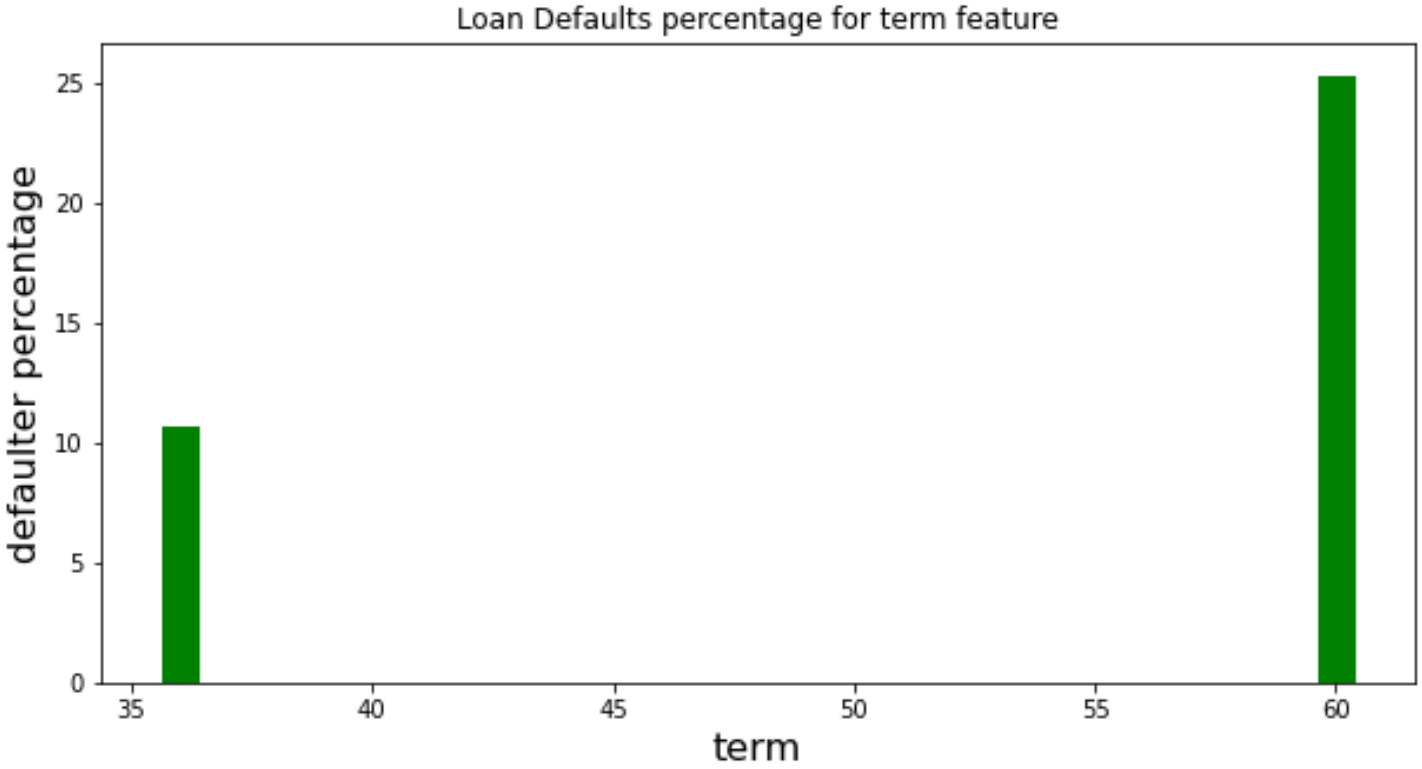
- As the term “univariate” suggests, it deals with analysing variables one at a time.
- It is important to separately understand each variable before moving on to analysing multiple variables together.
- We did analysis on each individual columns with respect to loan default ratio.
- From the analysis we found that loan_amnt, installment, term, purpose, int_rate, dti, grade, sub_grade so will check the if there is any relation between these features.
- Below are the plots for the features which has some patterns during the analysis



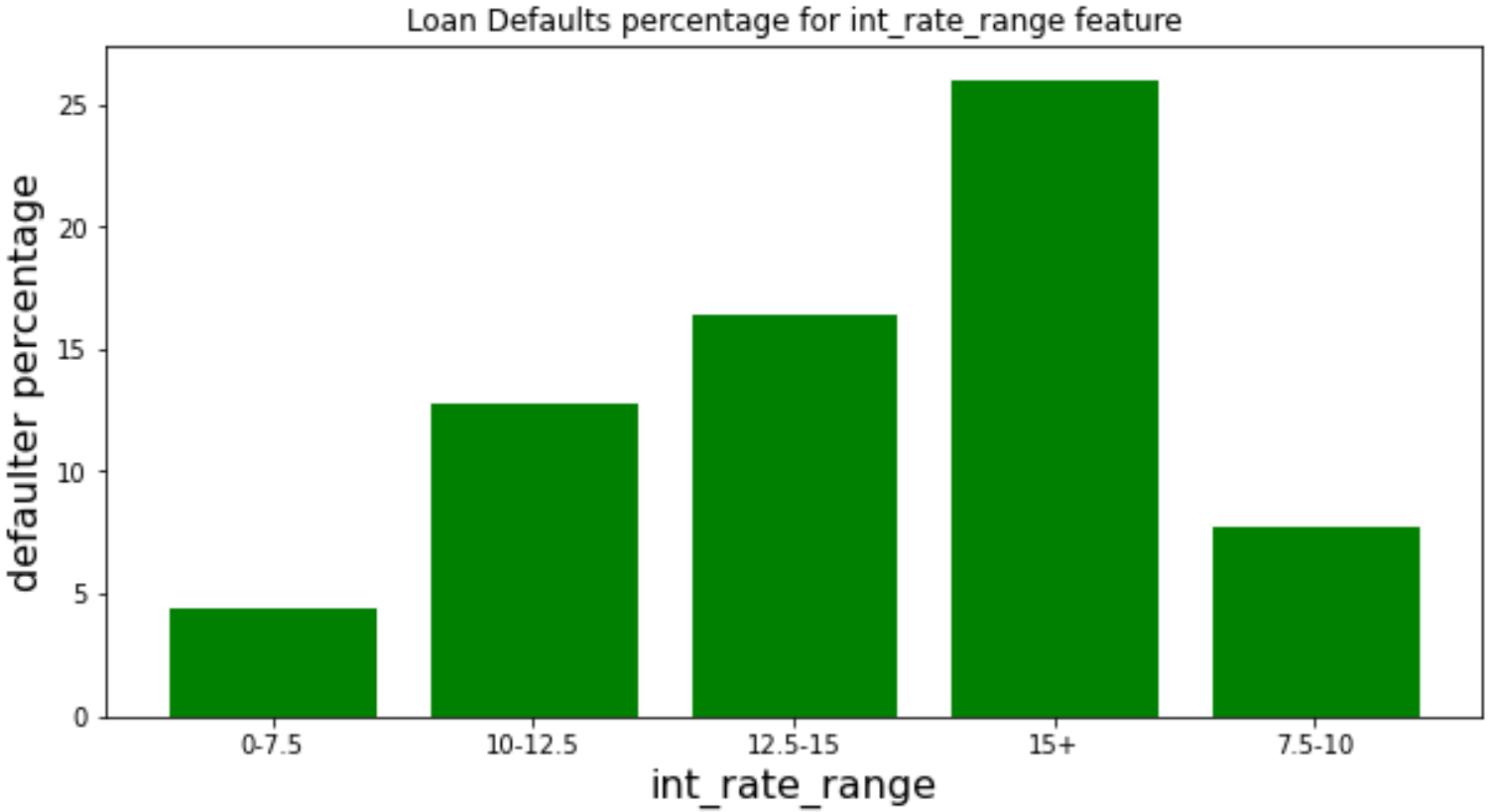
Here we find that lesser the annual income higher the defaulter percentage



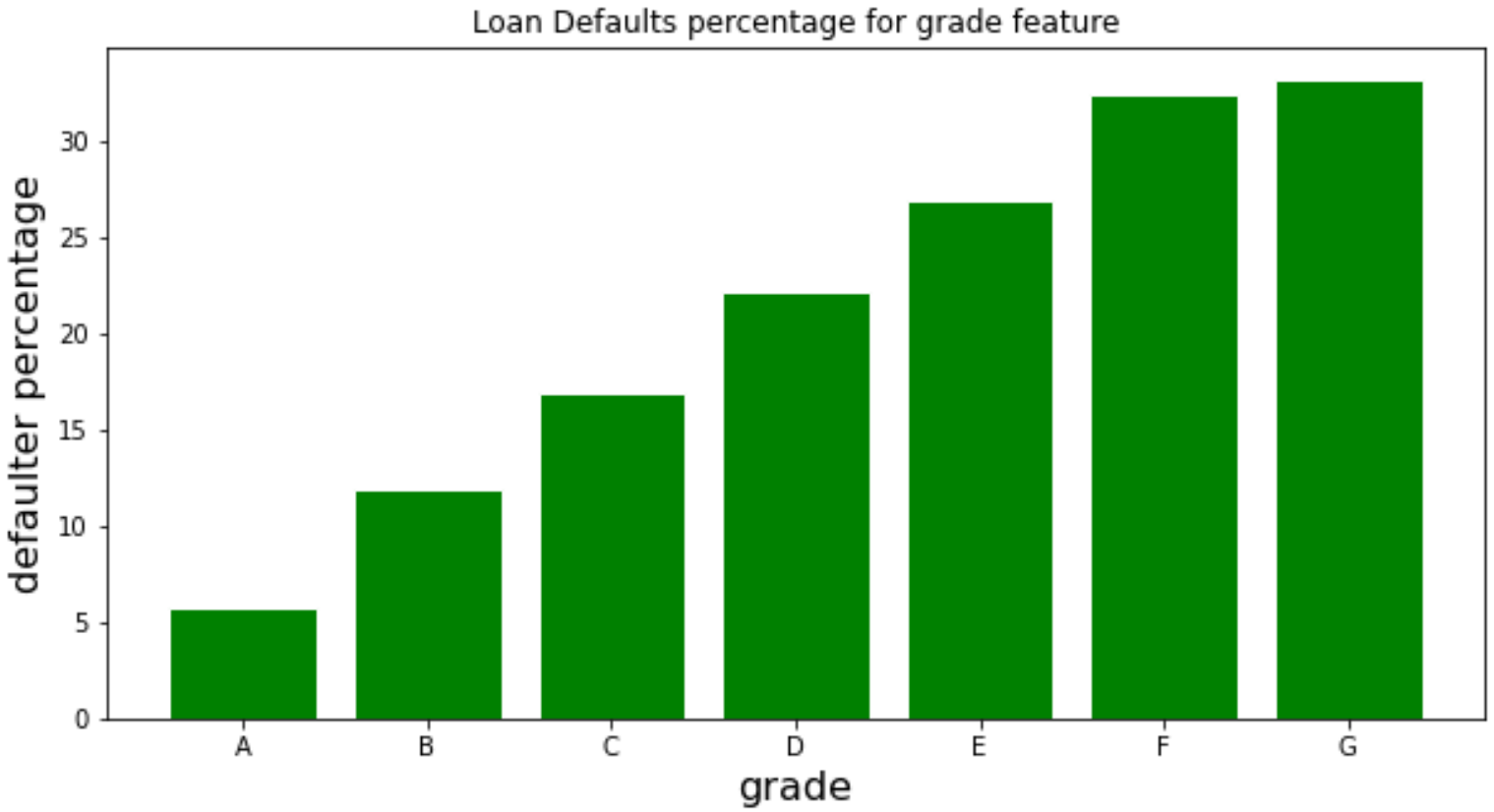
Here we find that higher the loan amount higher the defaulter percentage



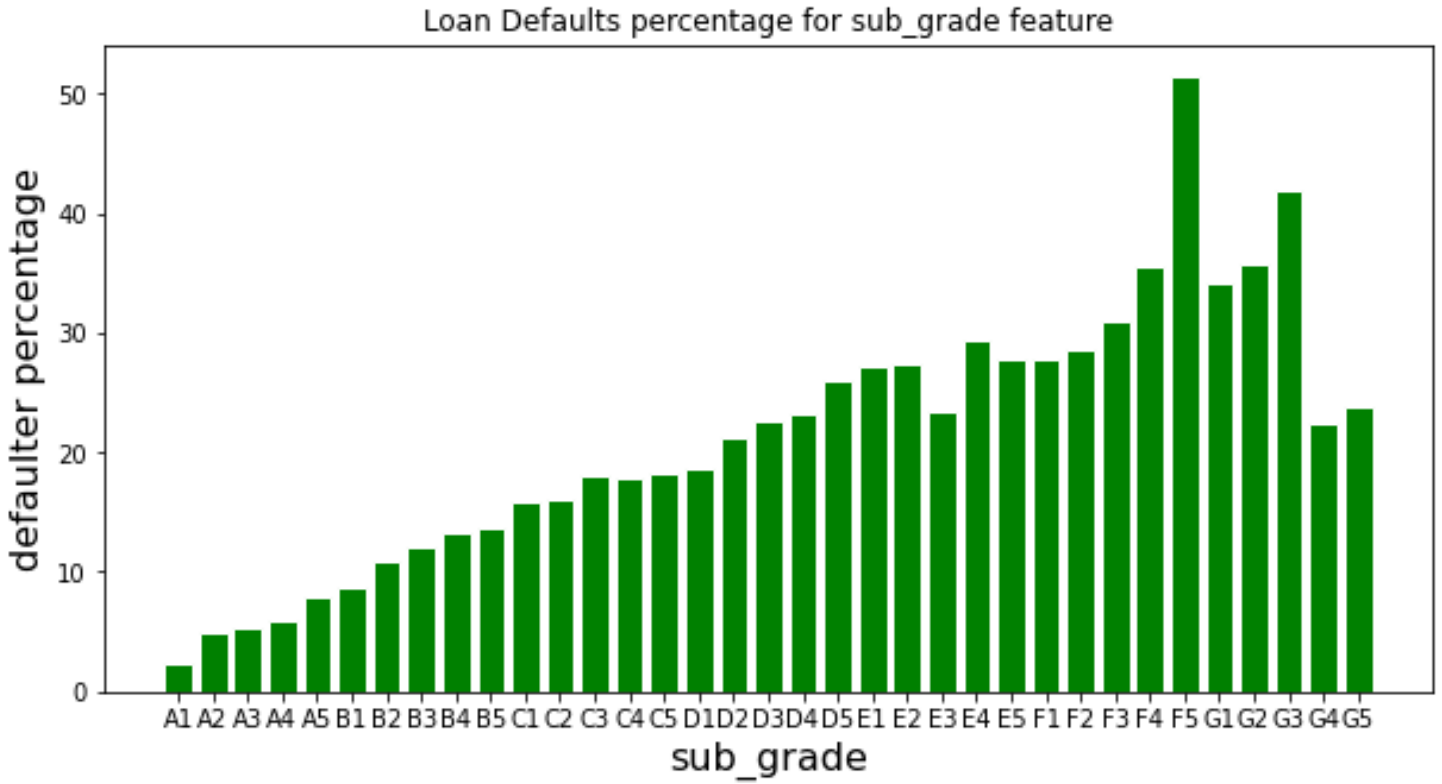
Here we find that increase in term months increases the defaulter percentage



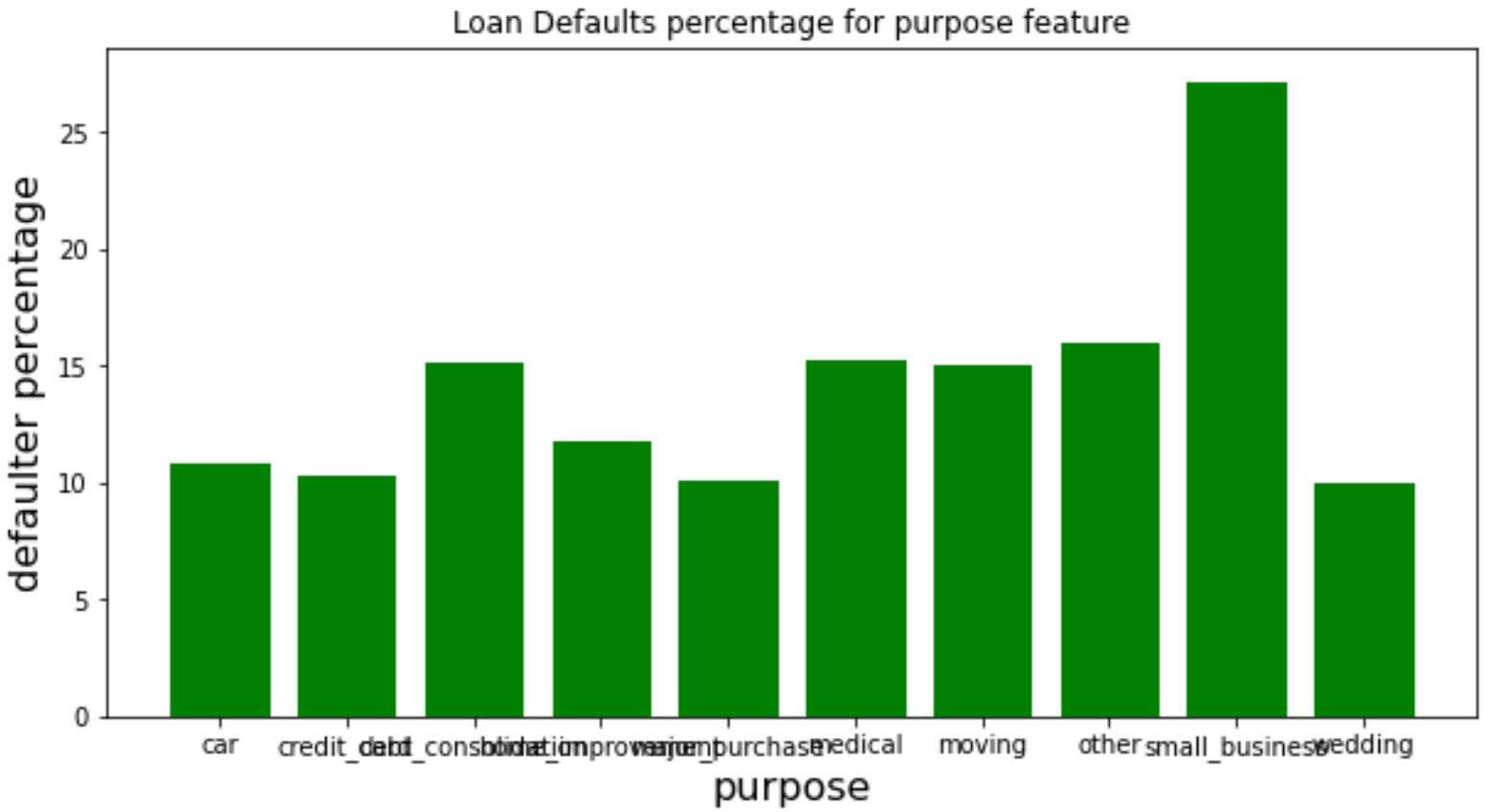
It shows that increase in interest rate increases the defaulter percentage



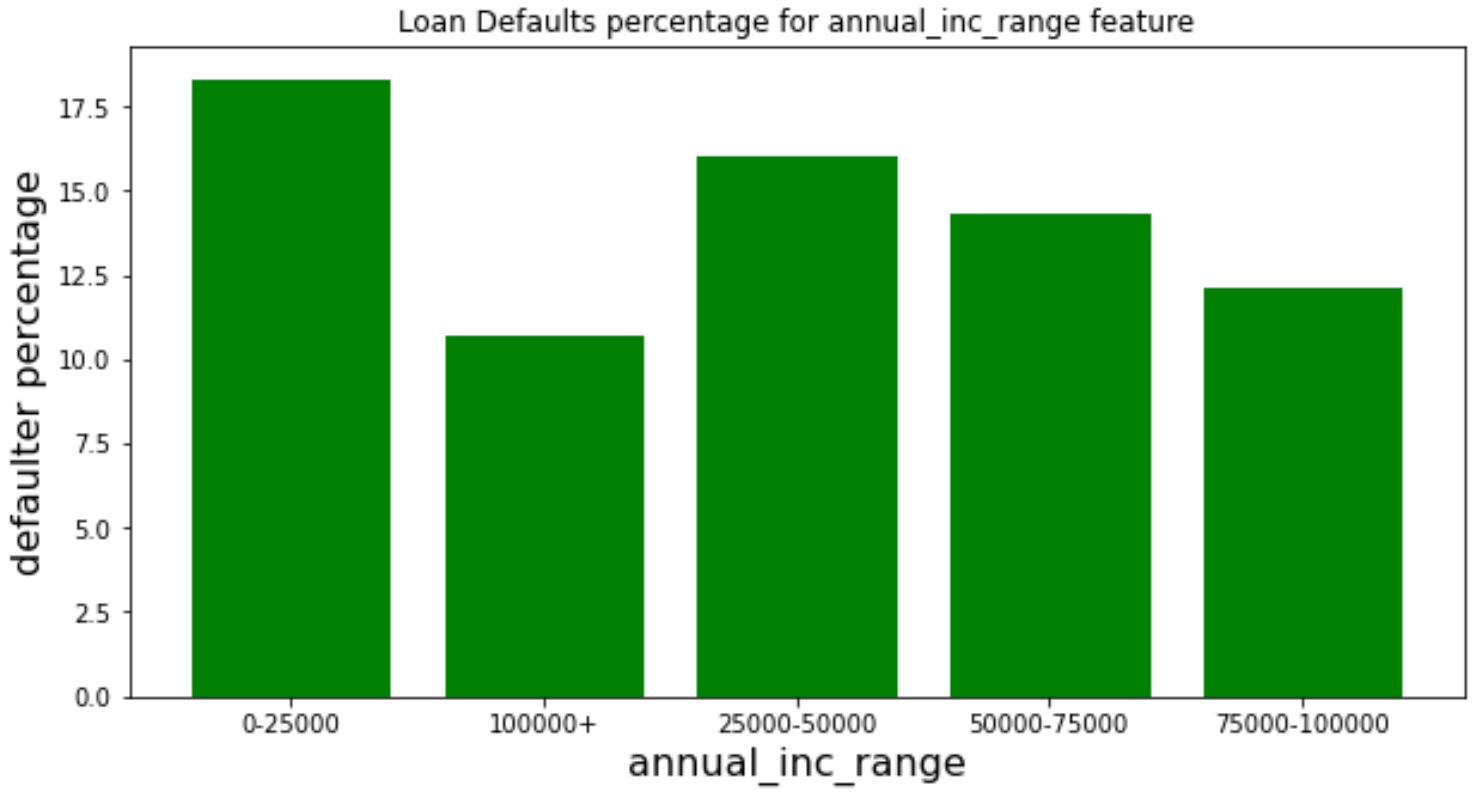
It shows a trend between grade and defaulter ratio.
Grade F and G has high defaulter ratio



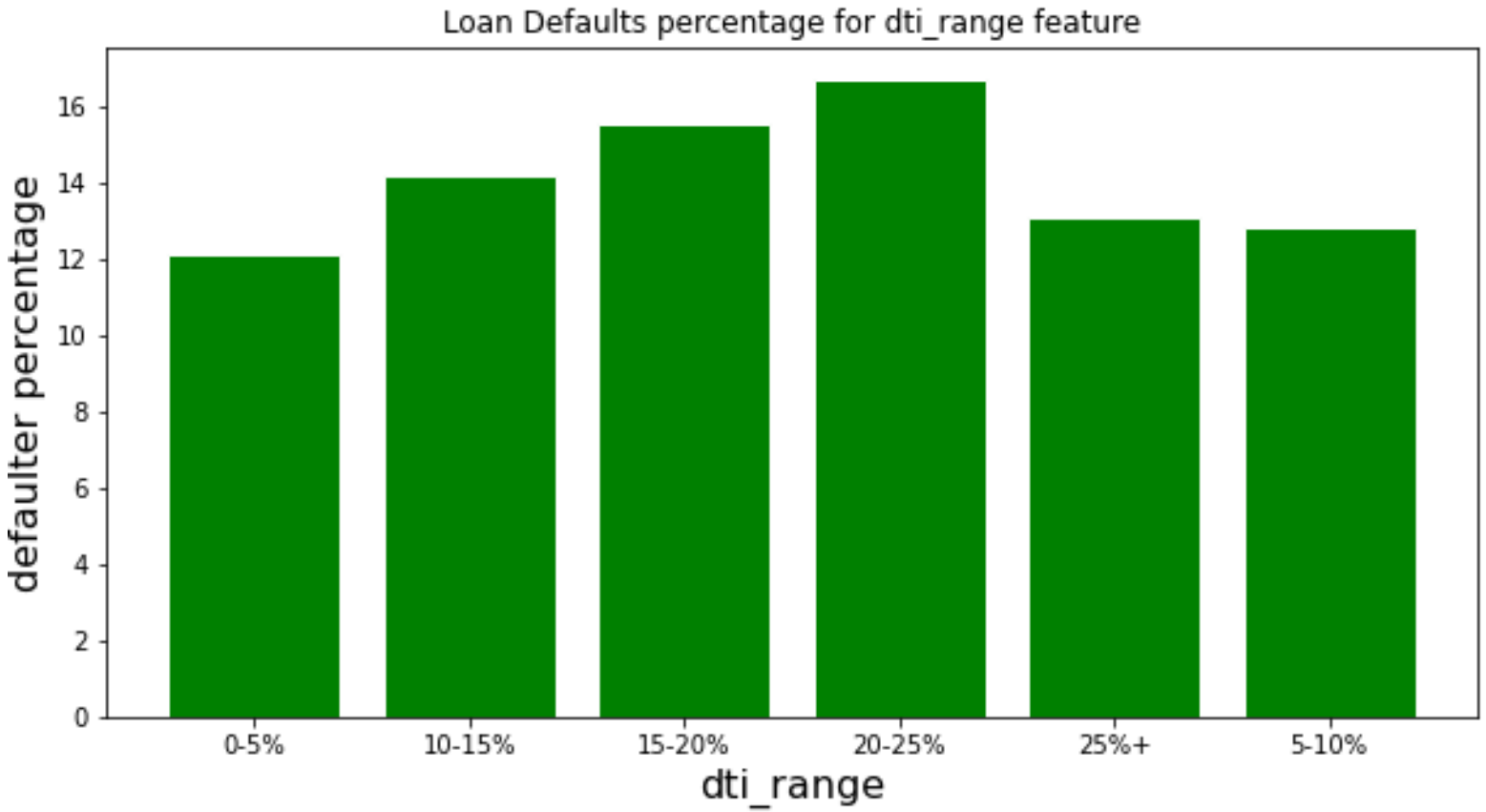
Sub_grade feature also shows the increasing trend for defaulter percentage



We can see there is an increase in defaulter percentage for different class of purpose. Here we can find small_business has high percentage of defaullt ration



Increase in annual income shows a decrease in defaulter percentage



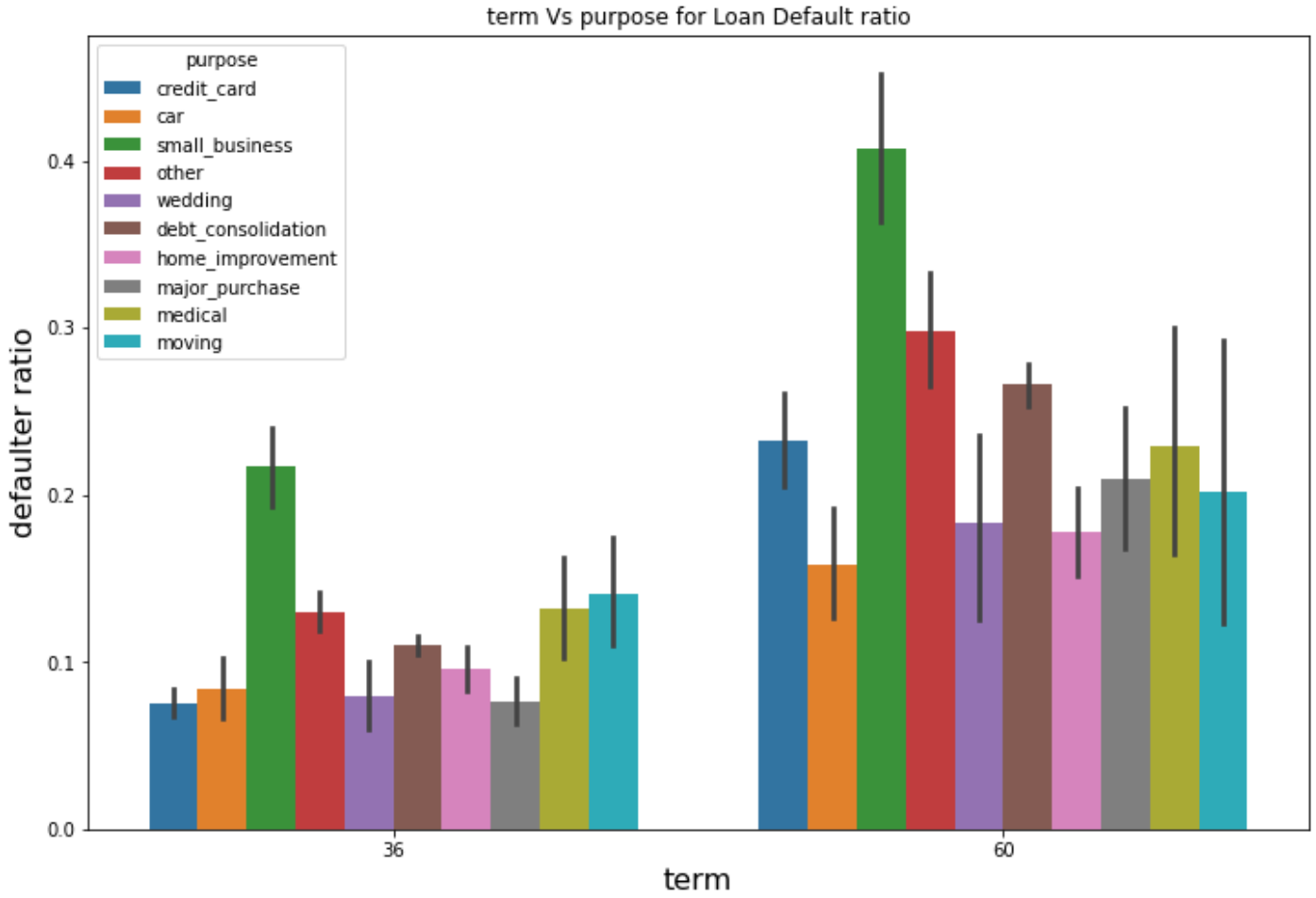
Increase in dti percentage shows increase in defaulter percentage

Bivariate analysis

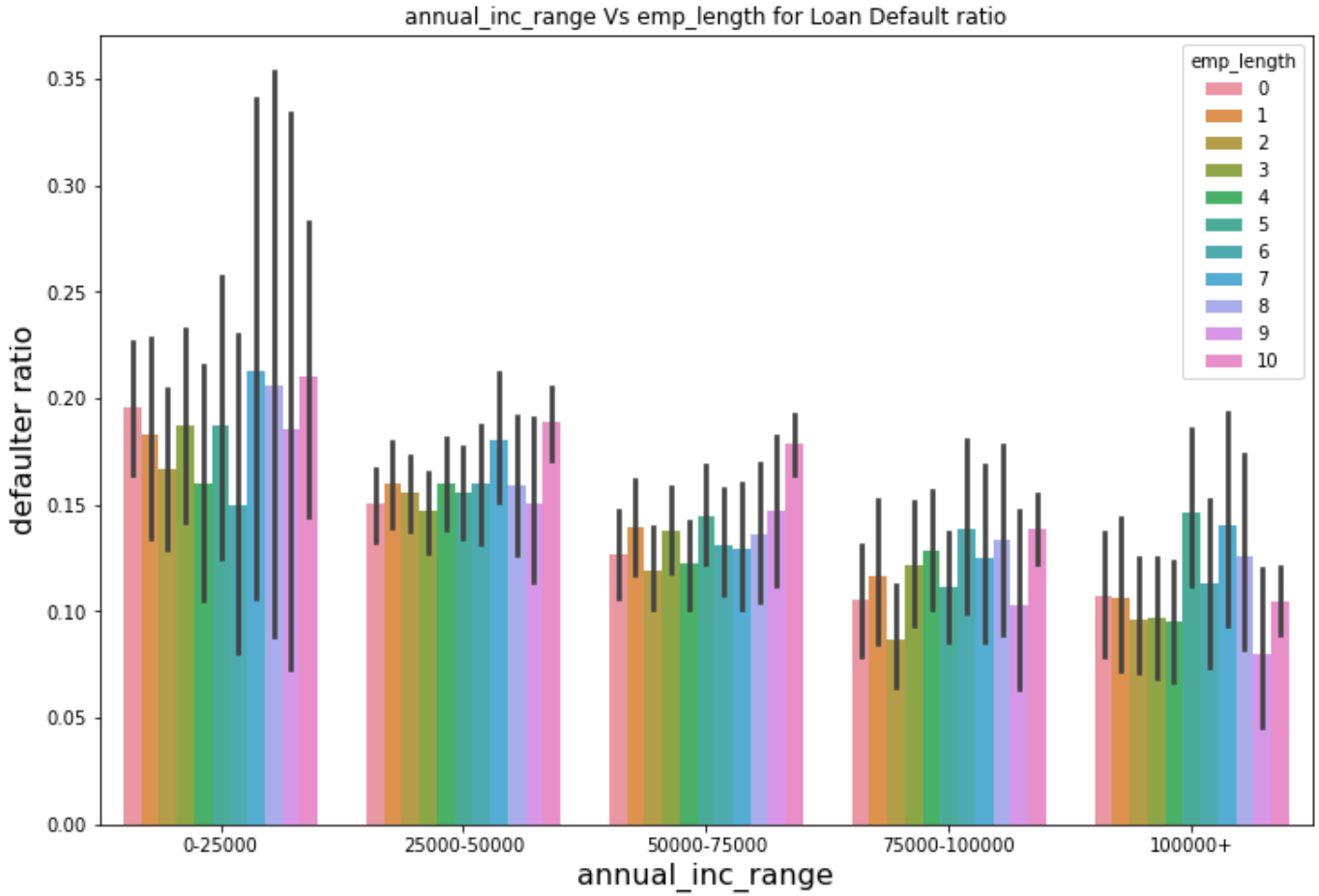
- Bivariate analysis refers to **the analysis of two variables to determine relationships between them**.
- There are 2 types of bivariate analysis

  Bivariate analysis on continuous variable

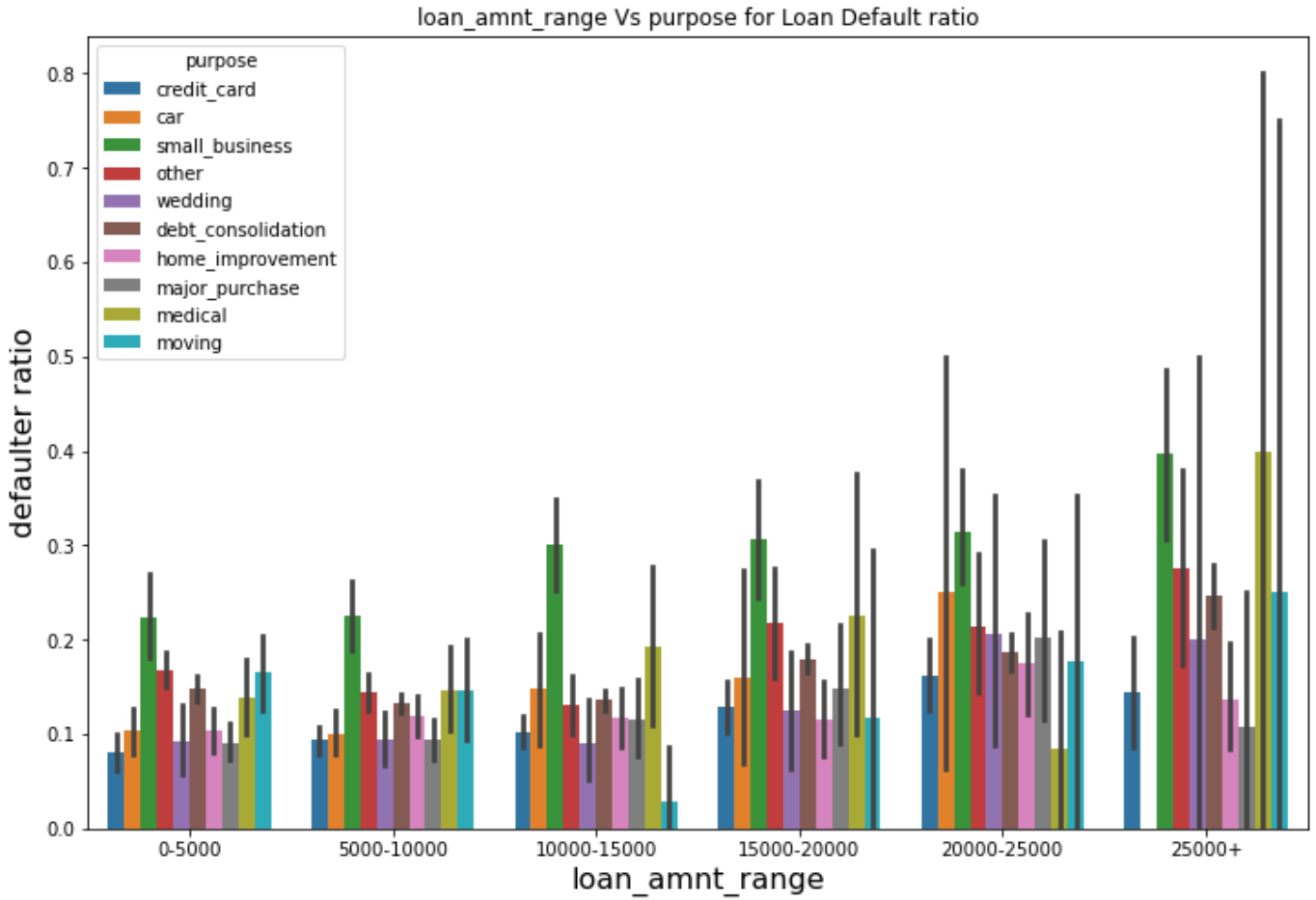
  Bivariate analysis on categorical variable
- From the analysis we found that term, grade, purpose, revol_util, loan_amount, int_rate, annual_inc, installment, funded_amnt_inv columns have some relationship between them
- Following are the graphs for bivariate analysis which shows some relation with respect to the loan status



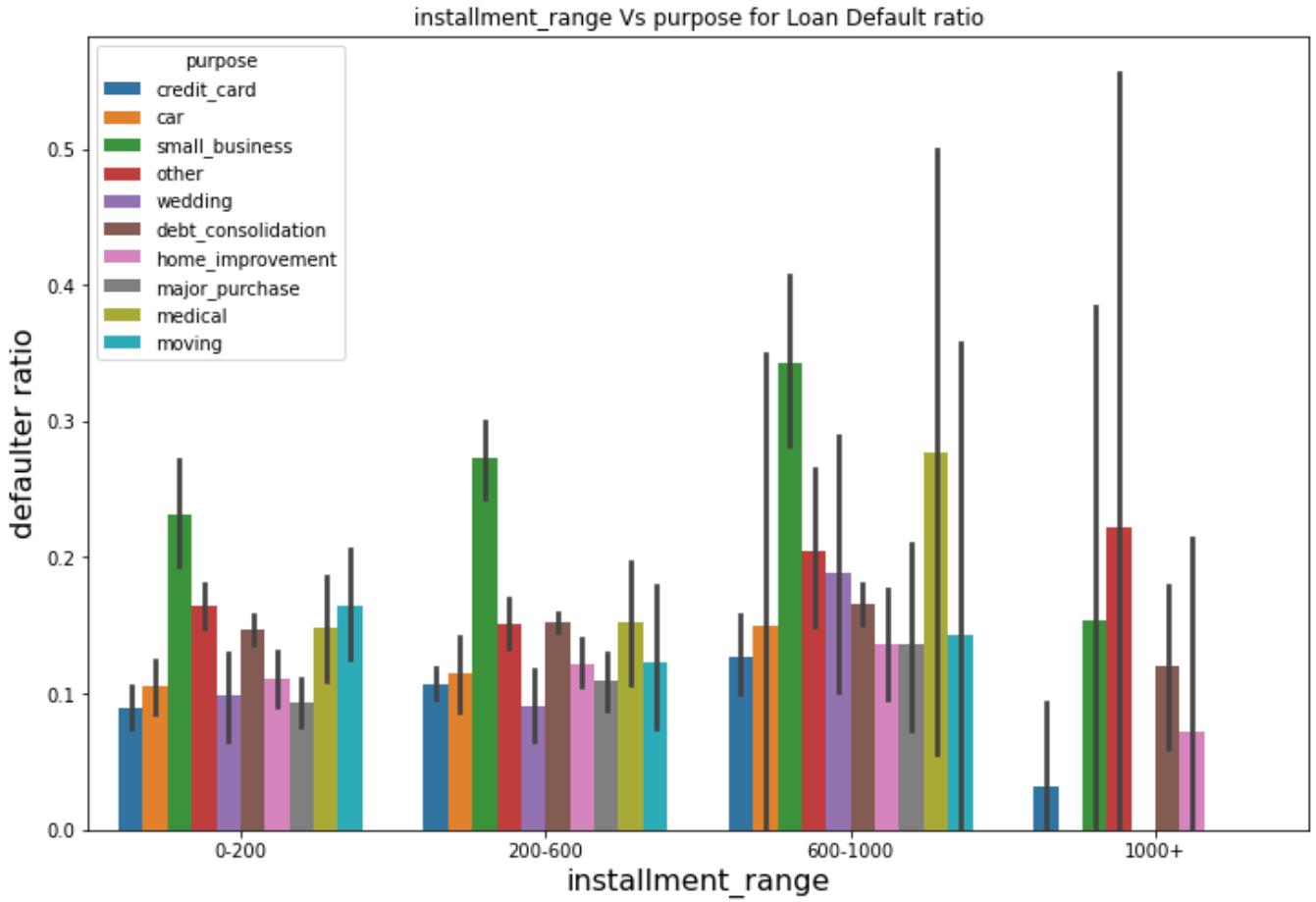
we find that the loan default increases with increase in term months for every different purpose



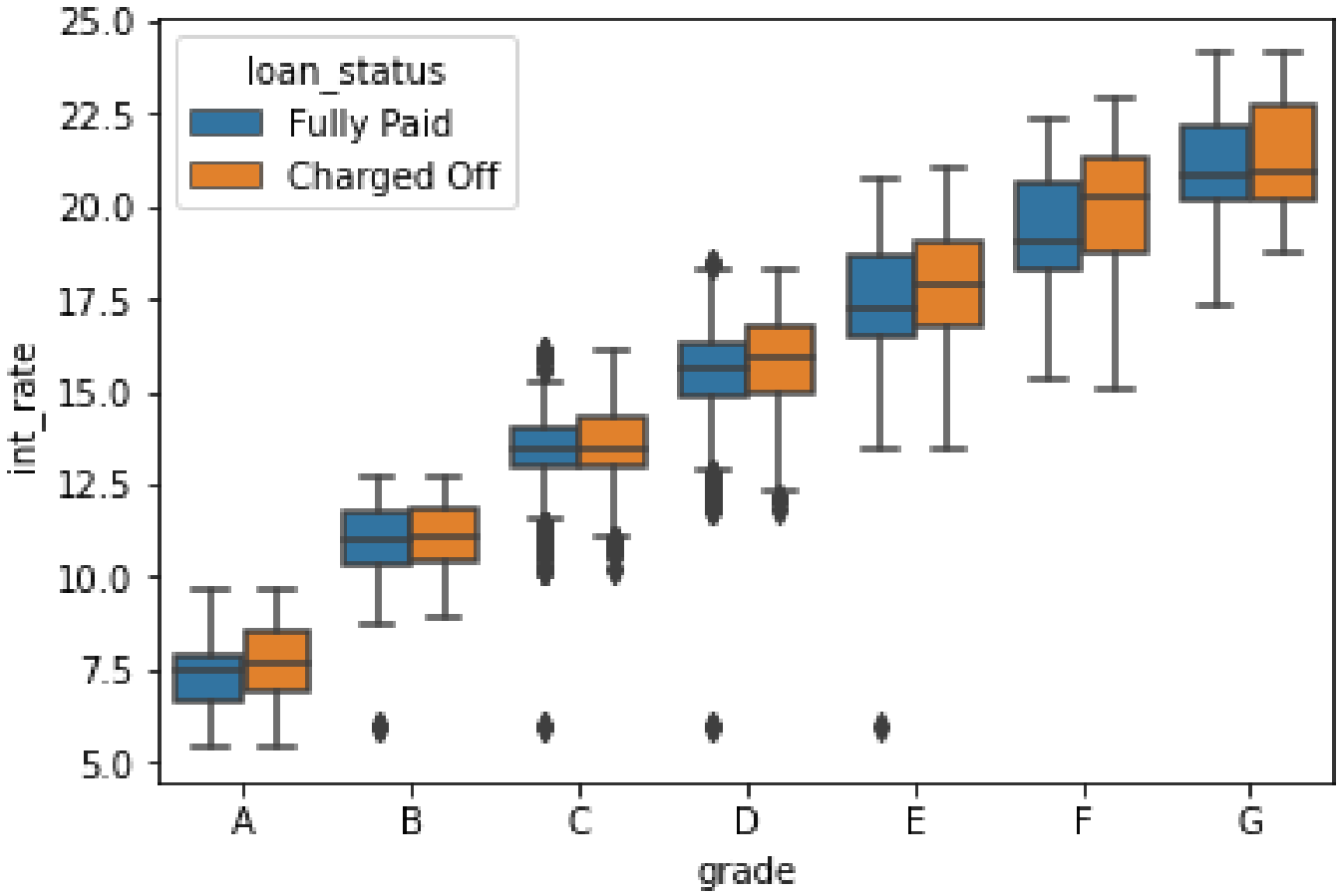
we find that for each purpose the defaulter ratio decreases with increase in annual income



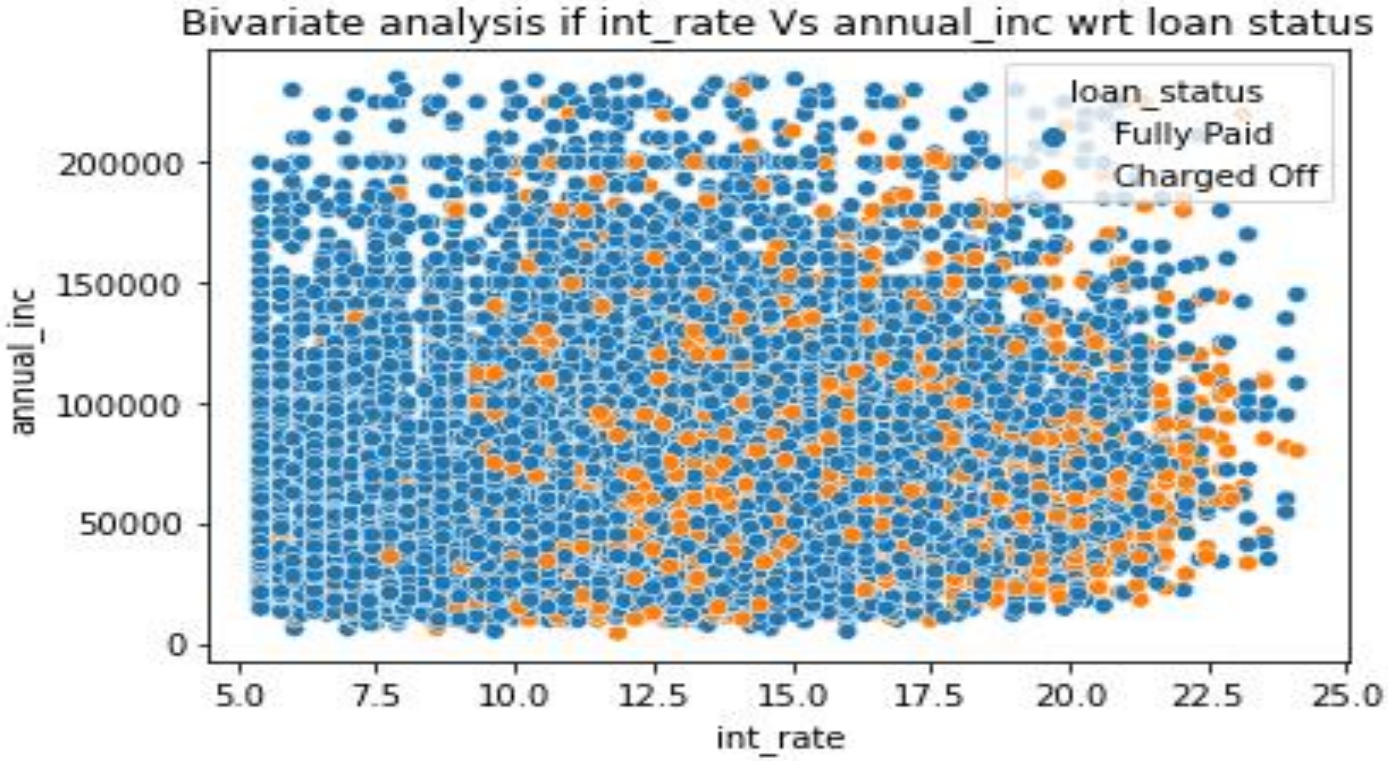
Default rate increases with increase in loan amount for every purpose



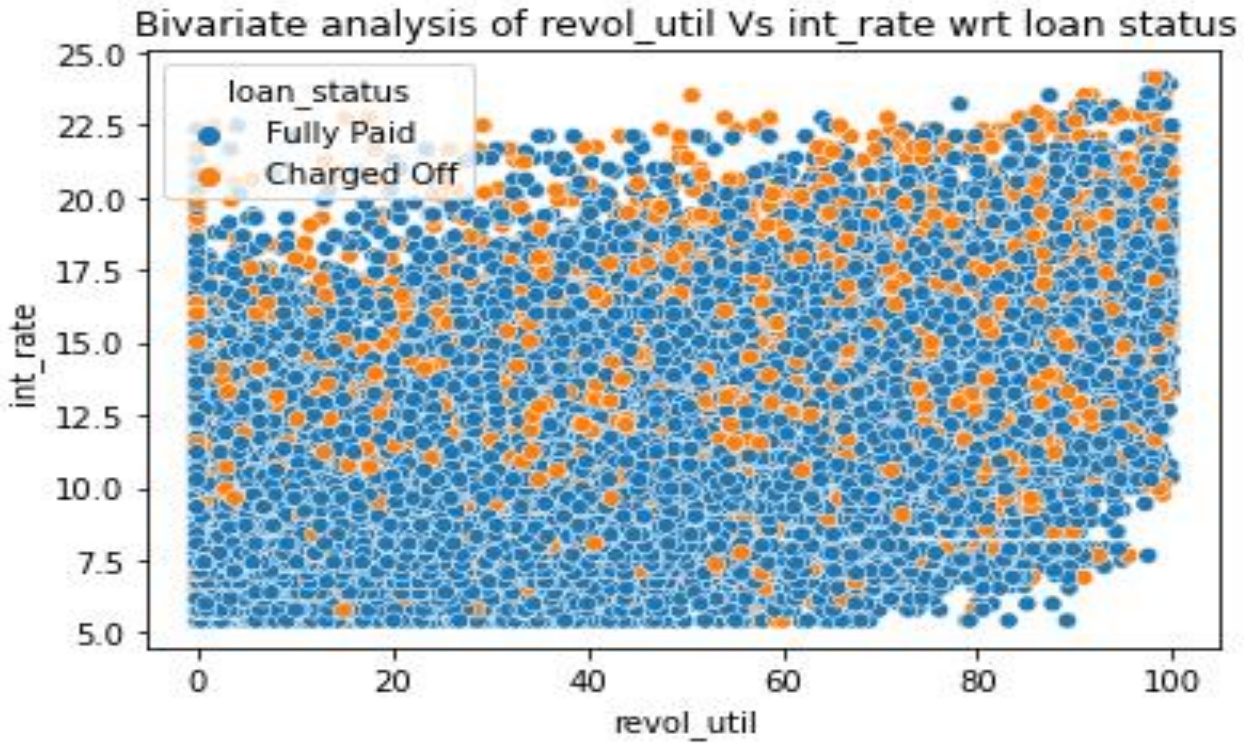
We find loan default for each purpose increases with in crease in installment except for small business .



With increase in int_rate there is an increase in loan default for each grade



we can find that increase in int rate and decrease in an nual income increases default ratio



we can see some relation between revol_util and int_rate as both increases then default ratio also increasing

Multivariate Analysis