Lending Club Case Study

SHRADDHA POWAR

JAY PANCHAL

Problem Statement

Minimise the risk of financial loss while lending money to customer

Two types of risk:

1. Likely to repay loan; not approving will cause loss of business
2. Not likely to repay loan; approving will cause financial loss

Objective

To identify if a person is likely to default. This will help in taking actions to reduce financial loss caused by defaulters

Steps Involved

Load the data

Data Understanding

Data cleaning

Univariate analysis

Segmented Univariate analysis

Bivariate analysis

Multivariate analysis

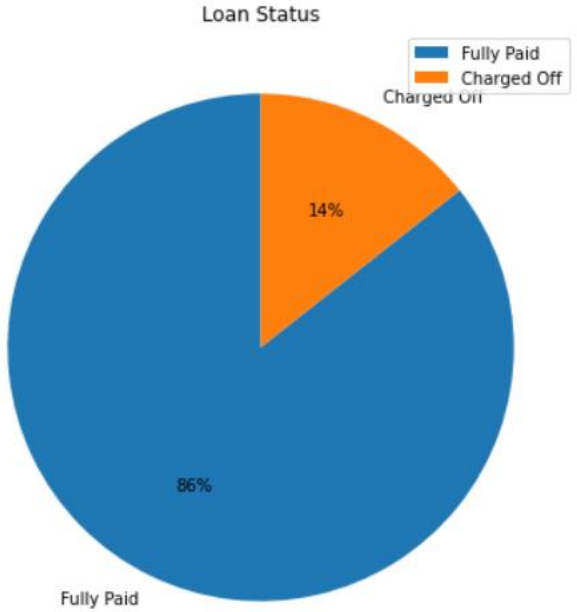
Data cleaning

1. Dropping the columns where:
   - All values are null (***54 columns had all null values***)
   - More than 90% values are null
   - columns having only one value i.e. single value columns
   - Columns such as url, desc, title , etc. which don't add any significant value in analysing whether a customer will default or not
2. Missing value imputation:
   - more than 6% of records in the emp_title column are null so creating new category as "No Title“
   - Fill 0 in place of null values to avoid bias for some columns where null values are less than 5%
3. Cleaning junk characters:
   - emp_lenght has too many junk characters in its values such as (>, <, years, year), we eliminate them to have a standard integer data
4. Updating data type for few columns:
   - emp_length to int , issue_d to datetime

Data cleaning

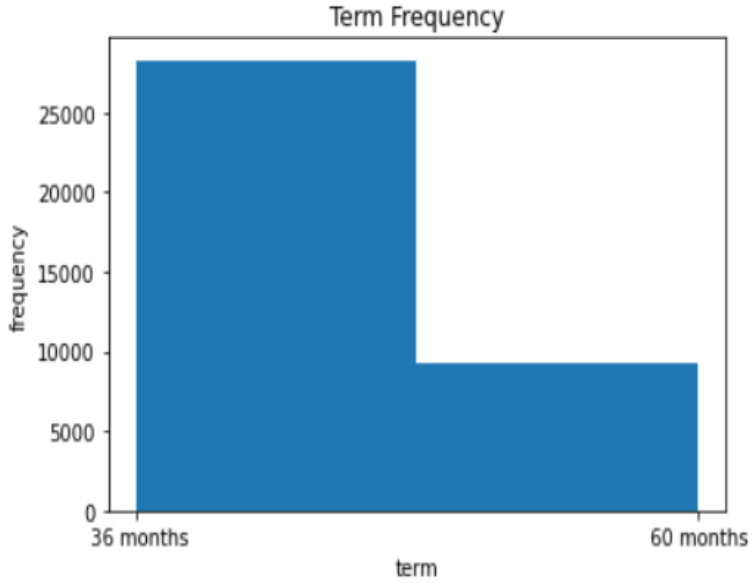
5. Creating reference columns from already existing data:
   - Extracting year and month from issue_d
   - Creating new column loan_status_numeric where we replace Fully paid with 0 and Charged off with 0
   - Categorizing interest rate data into low, medium, high and very high

Univariate analysis

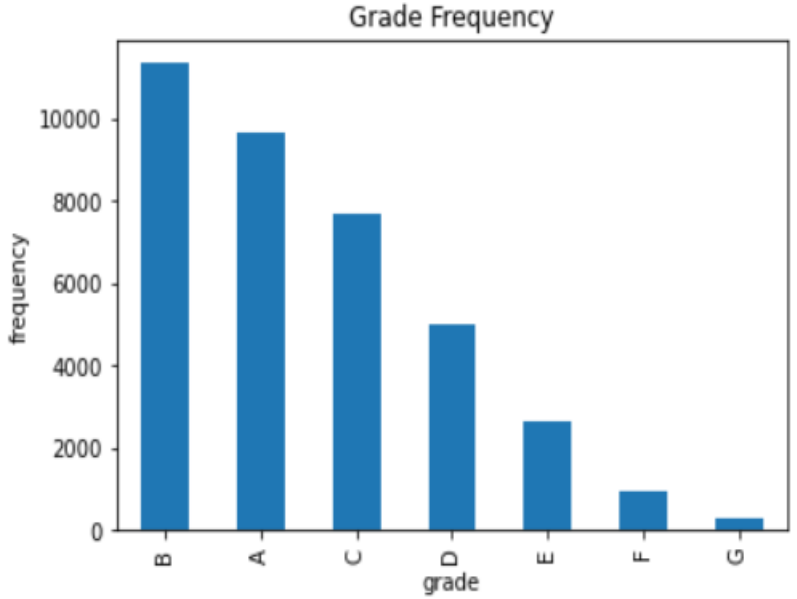


86% customers have fully paid the loan but 14% have been charged off

Default rate is 14%

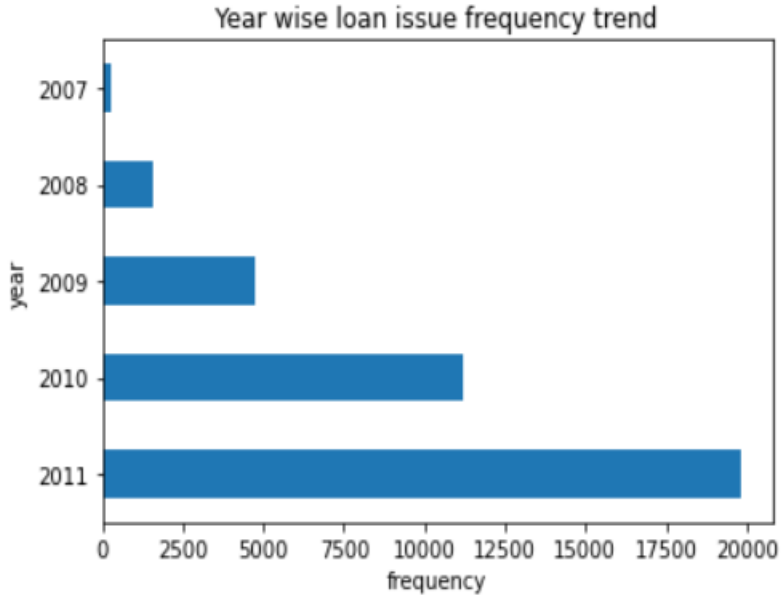


Count of 36months term loan is significantly higher compared to 60months

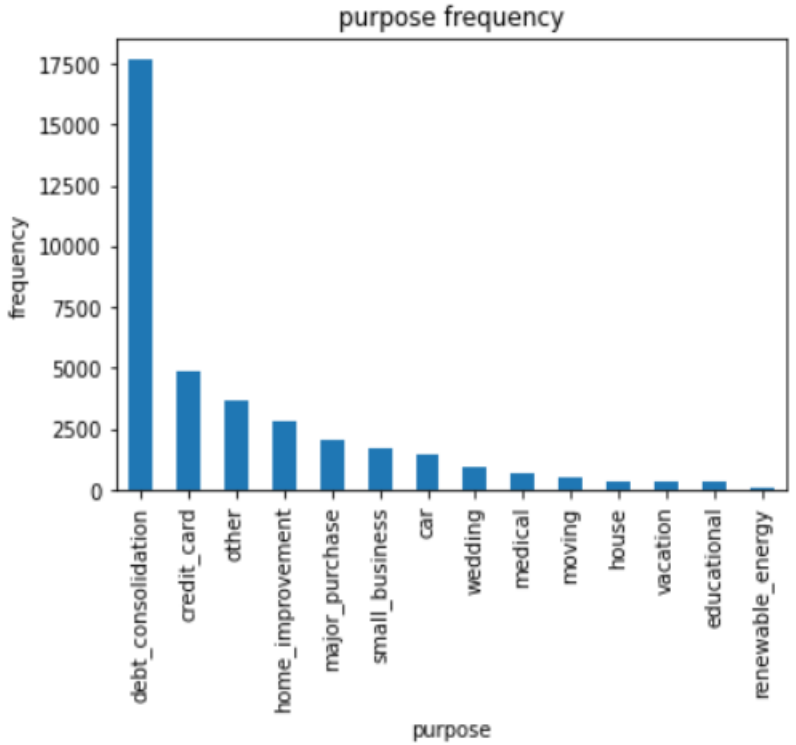


comparatively more number of A, B and C grade customers than others which is good. Higher number of quality customers is good for business.

Univariate analysis

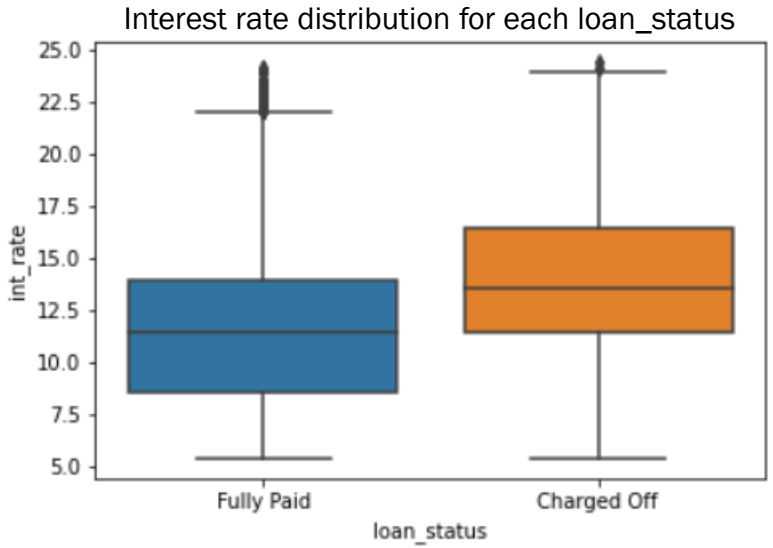


Increasing trend in the loan taken/issued over the years



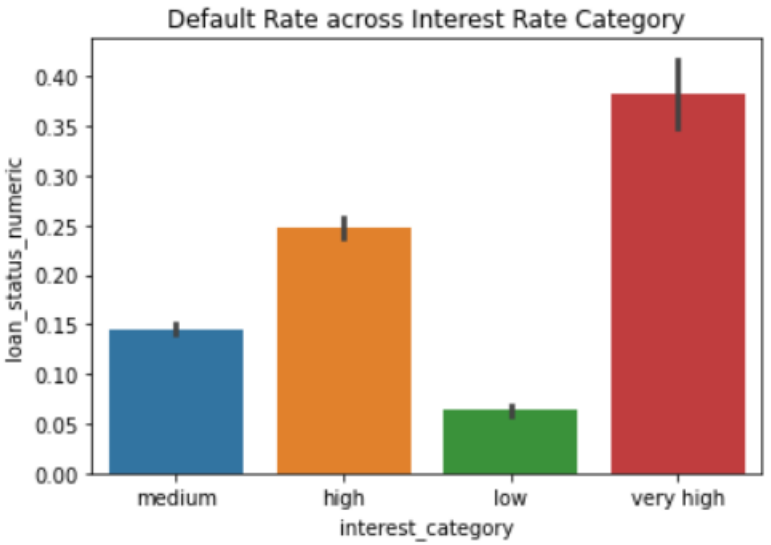
loan is majorly taken for debt consolidation

Segmented Univariate Analysis

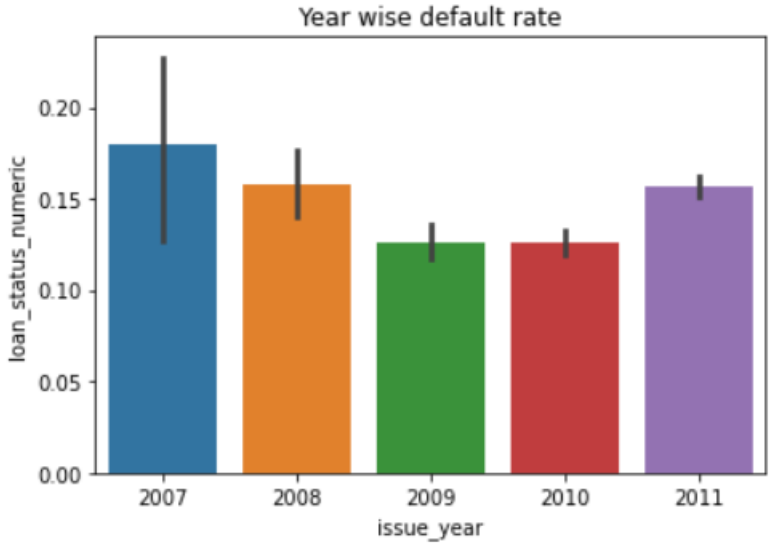


Median of charged off is significantly higher than that of fully paid i.e. when interest rate is high the probability of default is also higher.

Bivariate Analysis

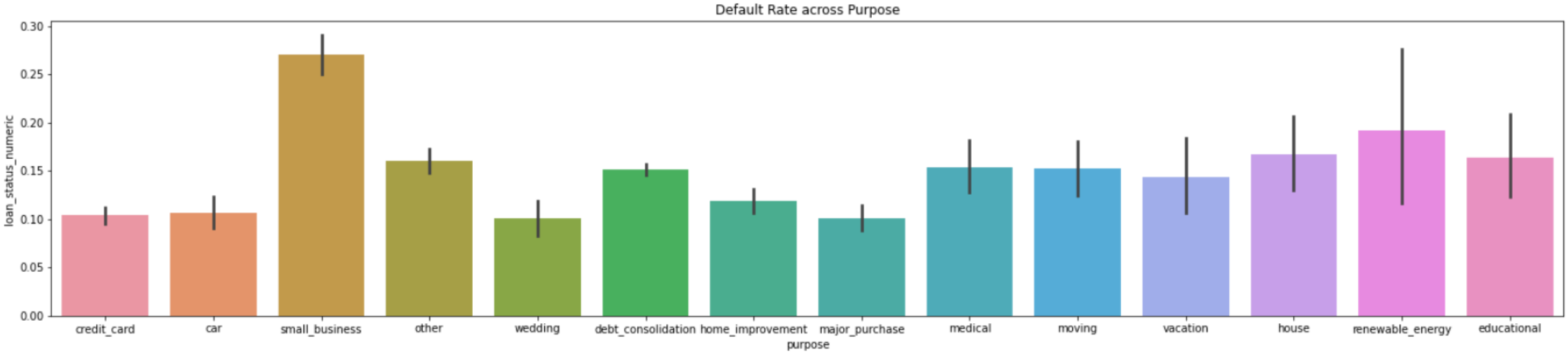


customers whose loans are in very high interest rate category have a comparatively higher tendency to default followed by high, medium and lastly low



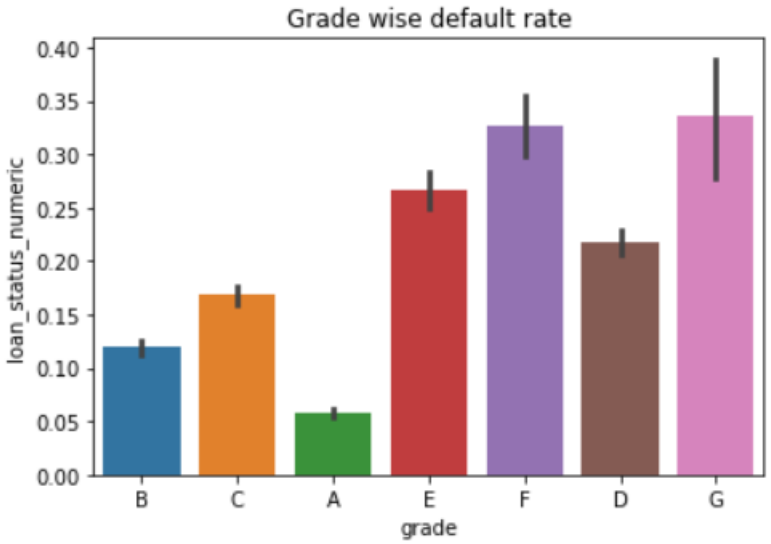
Default rate was max in 2007 after which it kept on dropping till 2009. But from 2010 we see a steady rise in default rate

Bivariate Analysis

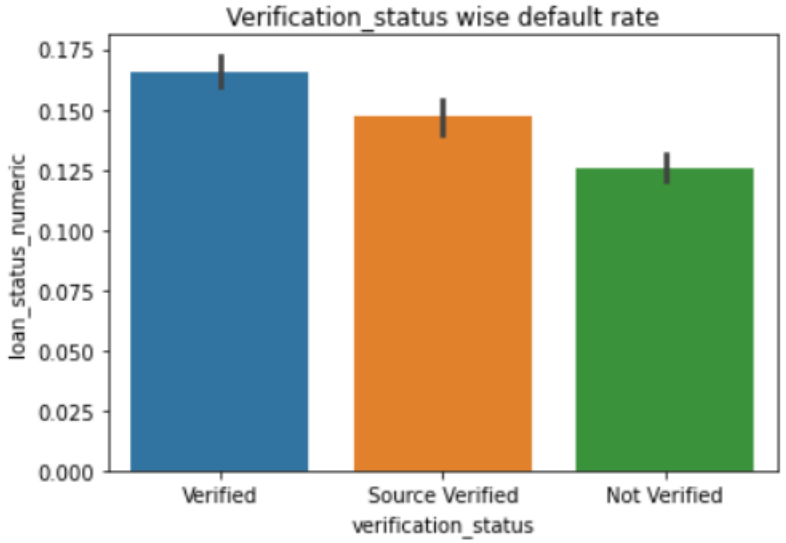


Frequency of small businesses defaulting is very high compared to others
Parameters to give loan to small business should be made more stringent in order to avoid financial loss

Bivariate Analysis

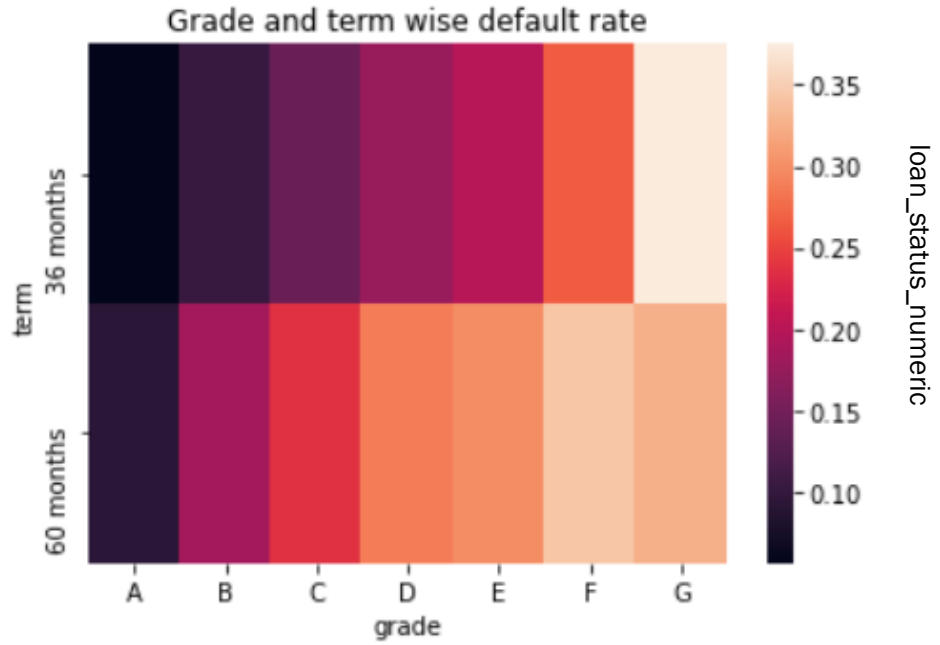


increasing trend in default rate from Grade A to G
E, F, G grade customers are responsible for majority default



default rate for verified customers is significantly more than that for non verified customers

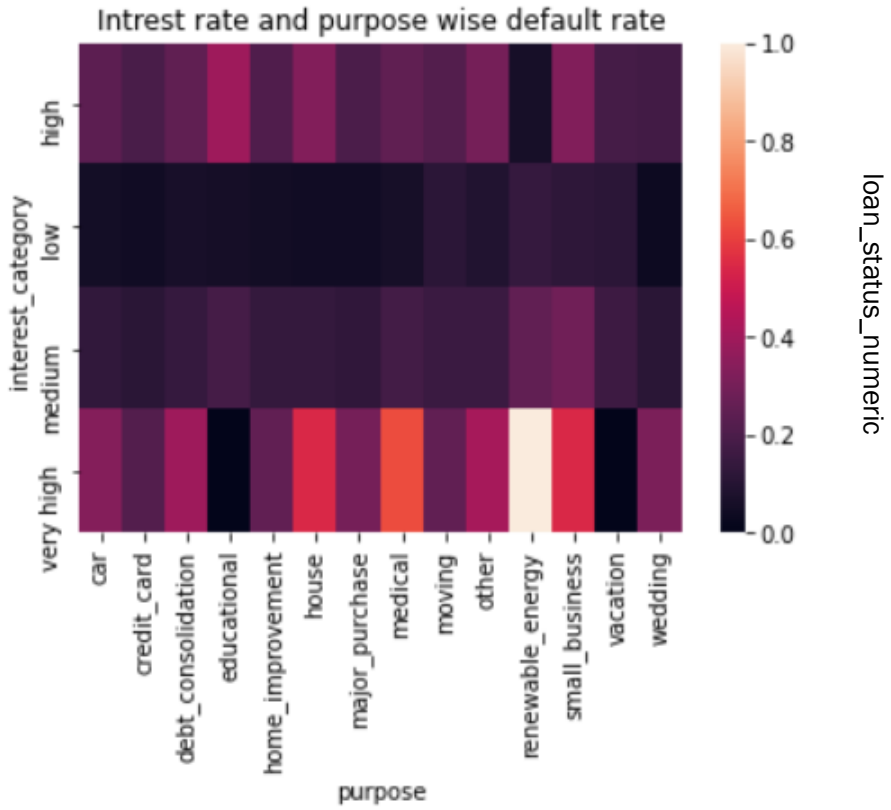
Multivariate Analysis



loan_status	loan_status_numeric
Fully paid	0
Charged off	1

customers whose grade is G default the most in 36month term loan whereas F grade customers default the most in 60months term loan

Multivariate Analysis



loan_status	loan_status_numeric
Fully paid	0
Charged off	1

renewable_energy purpose under very high interest category default the most

Thank You

Shraddha Powar

Jay Panchal