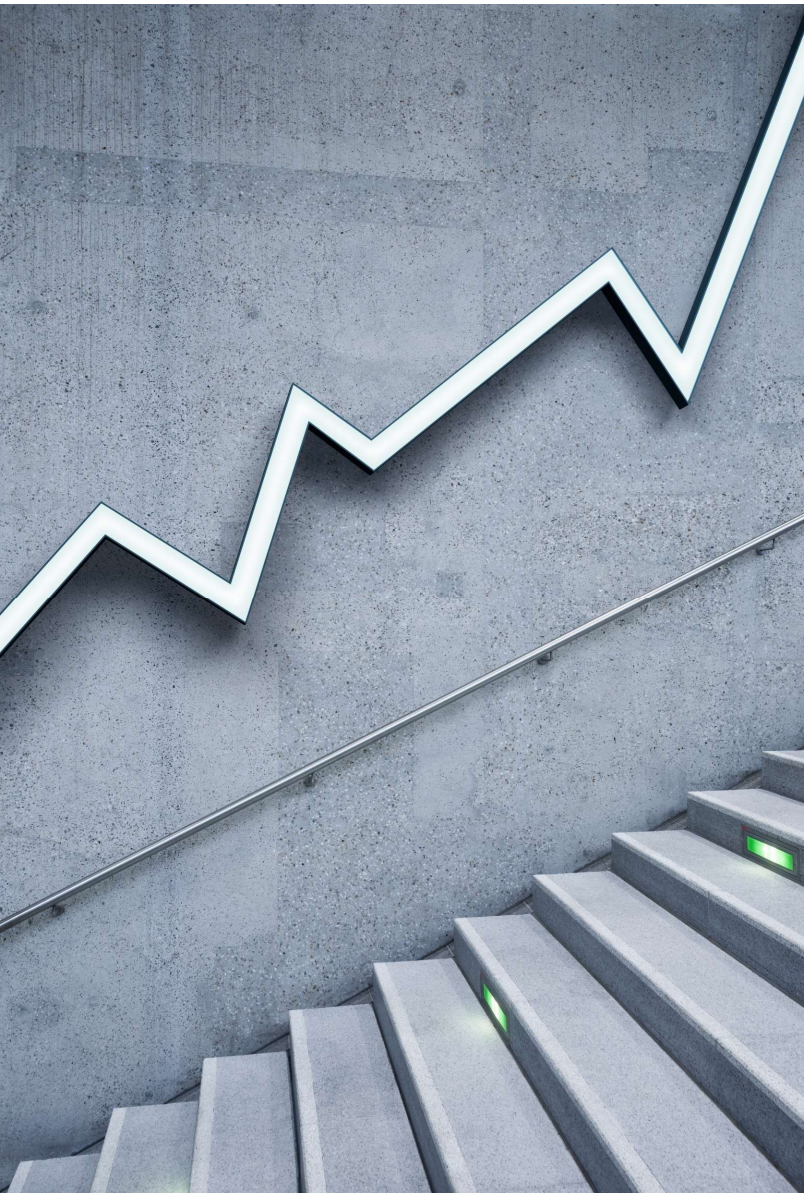


Lending Club Case Study

YUVARAJ DR

Your best quote that reflects your approach… “It’s one small step for man, one giant leap for mankind.”

- NEIL ARMSTRONG

Introduction

The idea about how real business problems are solved using EDA. In this case study, We will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Business Understanding

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

- 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

Business Objectives

- If one is able to identify risky loan applicants(defaulters), then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA.
- 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.

Steps for Enterprise Data Analysis

1. Data understanding
2. Data Cleaning and Manipulation
3. Data analysis
    1. Univariant analysis
    2. Bivariant analysis

1. Data understanding

- **load.csv** contains the complete loan data for all loans issued through the time period 2007 t0 2011.
- **Data_Dictionary.xlsx** the data dictionary which describes the meaning of these variables.
- load datasets from load.csv
- removing current customers as the data driven analysis is better by analysis of fully paid and defaulters

2. Data Cleaning and Manipulation

- List of Columns & NA counts where NA values are more than 30% and remove.
- removing all the columns having more than 30% of missing value. Because it will vary the result accuracy due to null value.
- removing all the rows having more than 10 null value as the details are not worth to consider if its insufficient to describe one's reason.
- let us remove the columns which having non unique values
- Removing unnecessary columns which have description, id and employer title.
- pre processing of data to get good quality of data.

…Continued

- converting columns to there real data type which help for analysis.
- let us consider employee having less than 1 year as 0 year experience and 10+ as 11.
- Let us remove columns contain 75% of data as 0 for numerical variable
- Check for null value and again step by step add relative values or constant.

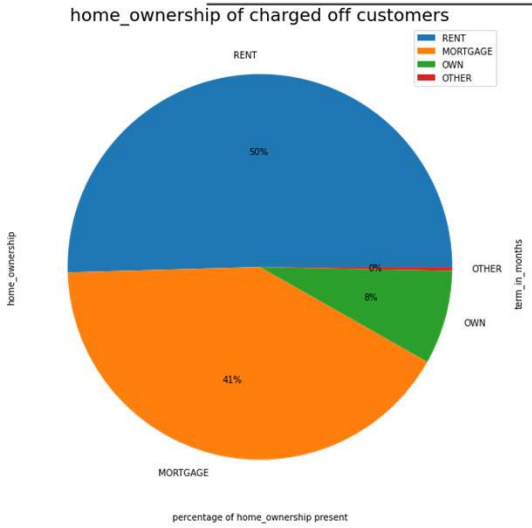
3. Data analysis

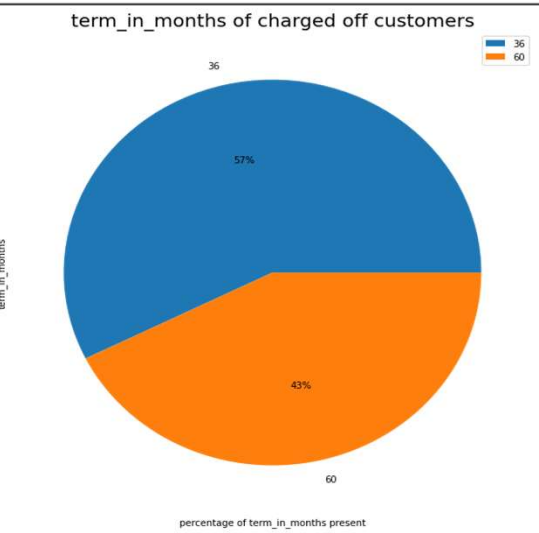
- Get the data driven matrices from the given data sets.
  - Monthly income
  - Monthly income per instalment
  - Monthly income range
  - Instalment range
  - Interest range
  - Loan amount range
  - Experience range
  - Debt to income range

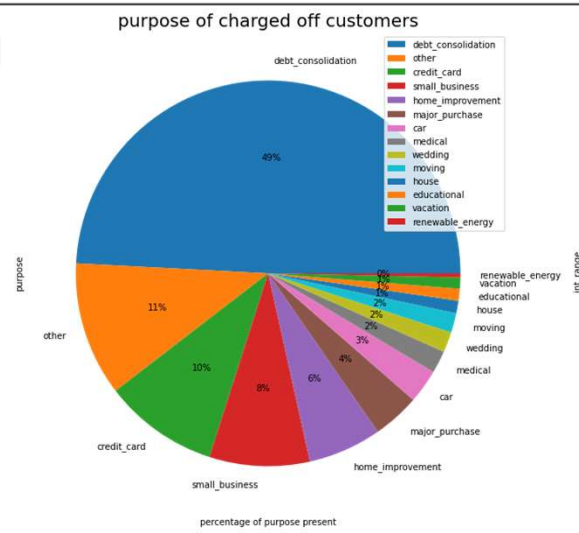
1. Univariate Analysis of Categorical Variable

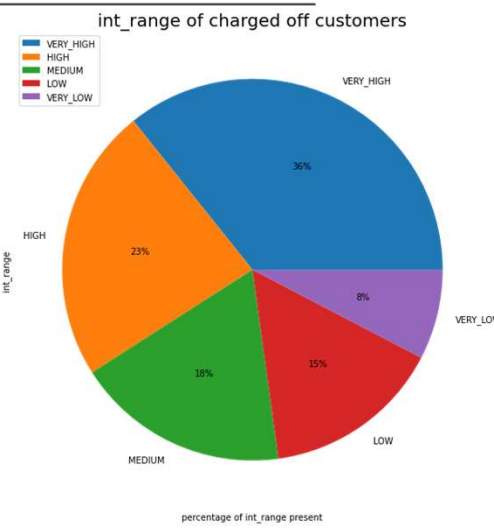
- Beneficial Variable are as fallows:
  - home_ownership: defaulters rate varies
  - term_in_months: defaulters rate decreases
  - Grade: defaulters rate decreases(B>C>D>E>A>F>G), A has to be verified.
  - sub_grade: defaulters rate decreases(B5-G5)
  - Purpose: defaulters rate is varies
  - dti_range: defaulters rate increases
  - Int_range: defaulters rate decreases
  - installment_range: defaulters rate increases
  - monthly_inc_range: defaulters rate
  - Loan_amnt: We can observe some outliers and the first quartile is bigger than third quartile for loan amount which means most of the defaulters clients are from first quartile

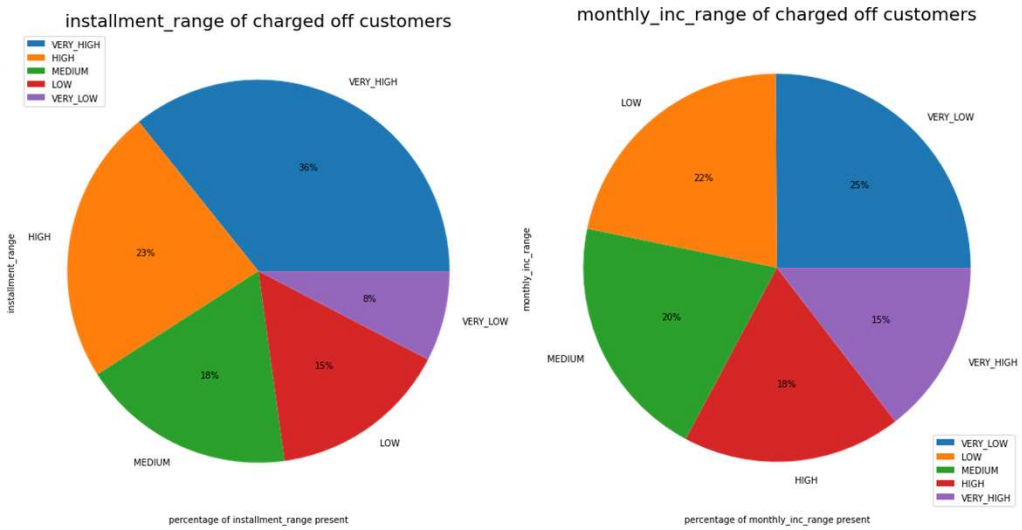
Univariate analysis Graph

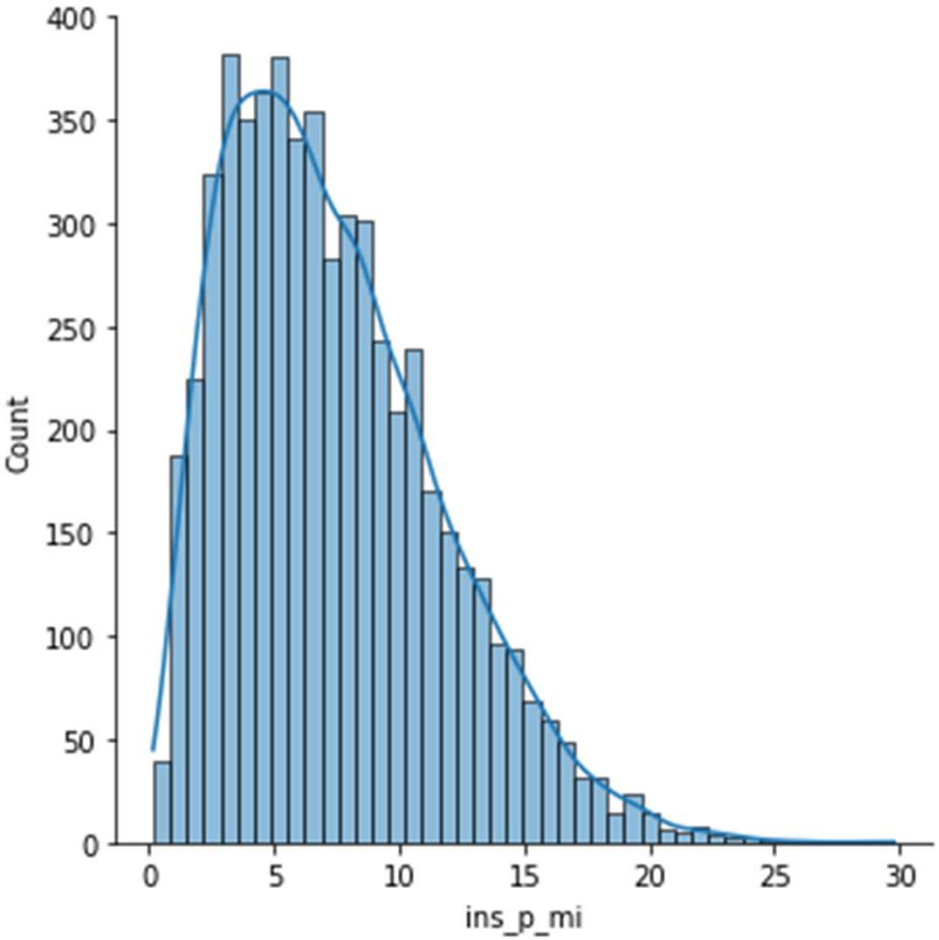










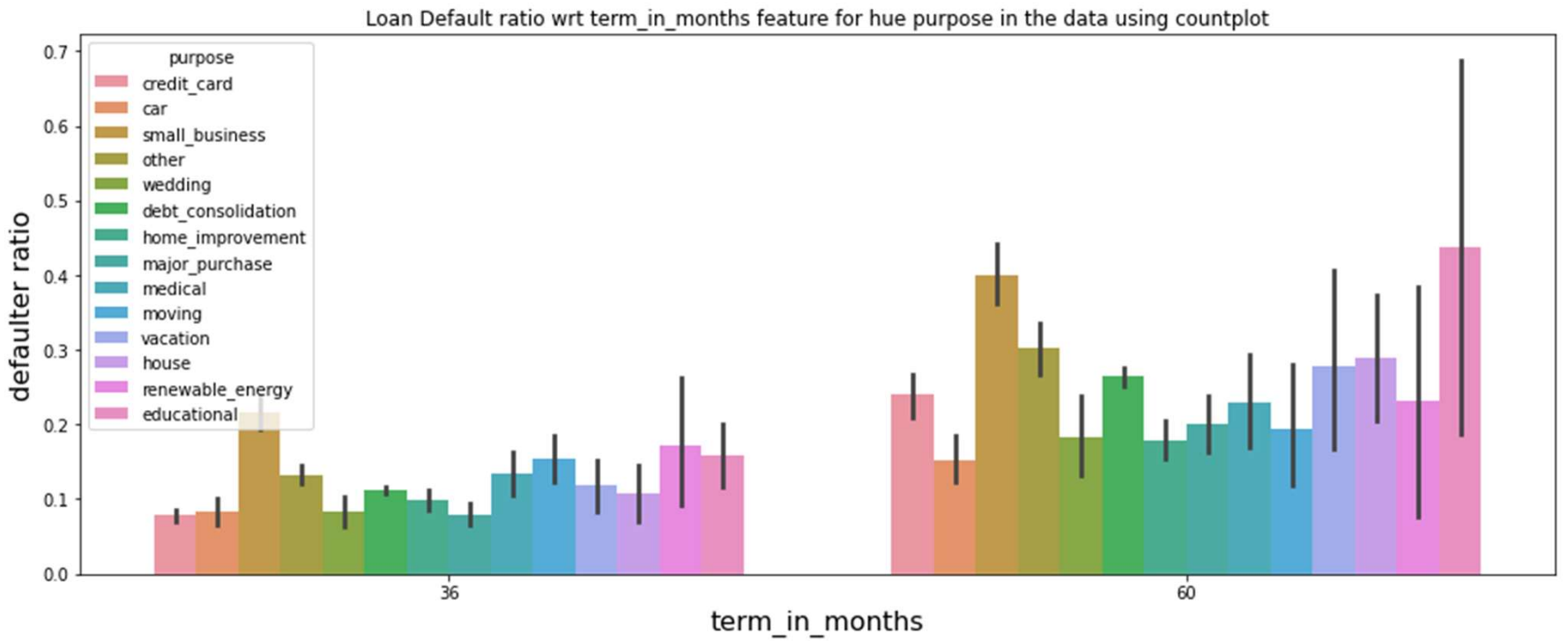


- We can observe instalment per month income(ins_p_mi) is having more count in 0-10 and more in 5.

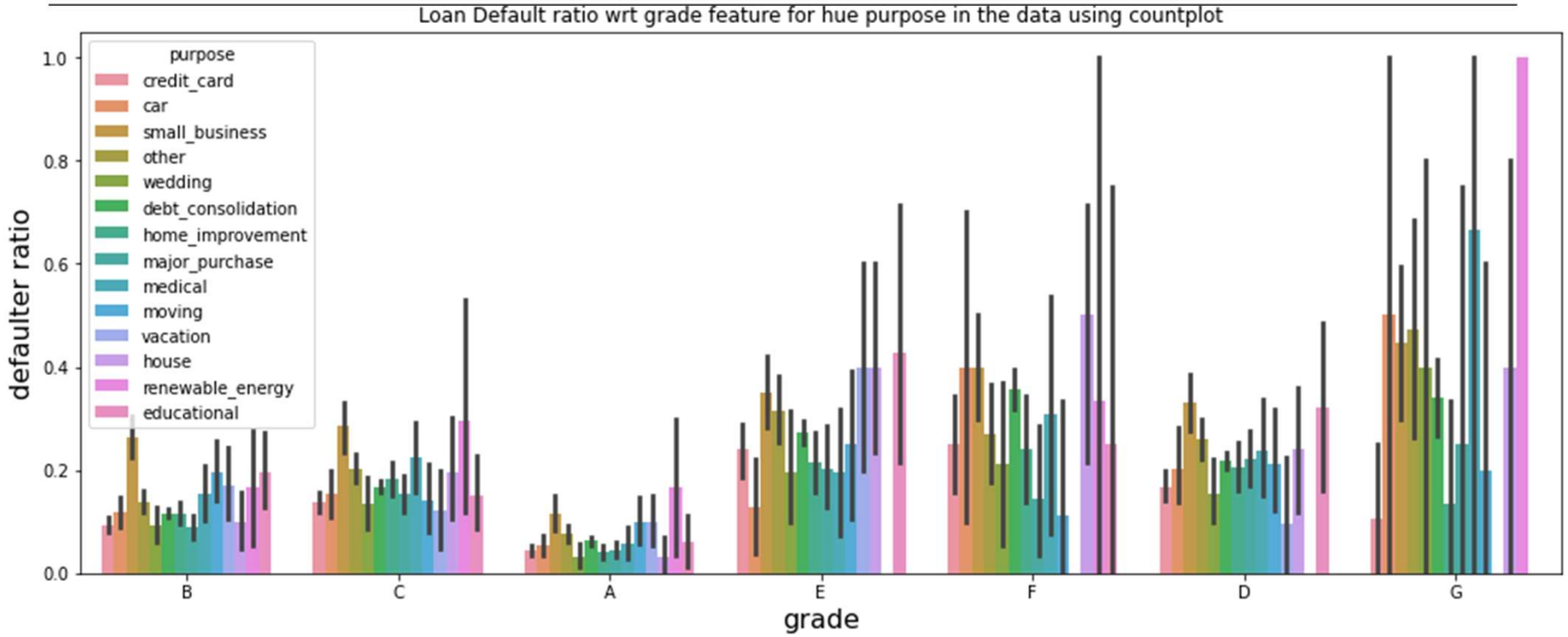
2. Bivariate analysis

- Two variables which are related are as fallows:
  - loan_amnt_range and purpose: default ratio increases for every purpose wrt loan_amnt_range
  - grade and purpose: default ratio increases for every purpose w.r.t grade
  - loan_amnt_range and term_in_months: default ratio increases for every purpose w.r.t loan_amnt_range
  - monthly_inc_range and purpose: default ratio increases for every purpose w.r.t monthly_inc_range
  - term vs purpose: default ration increases for every purpose w.r.t term
  - Instalment_range vs purpose: default ratio increases for every purpose w.r.t installment except for small business

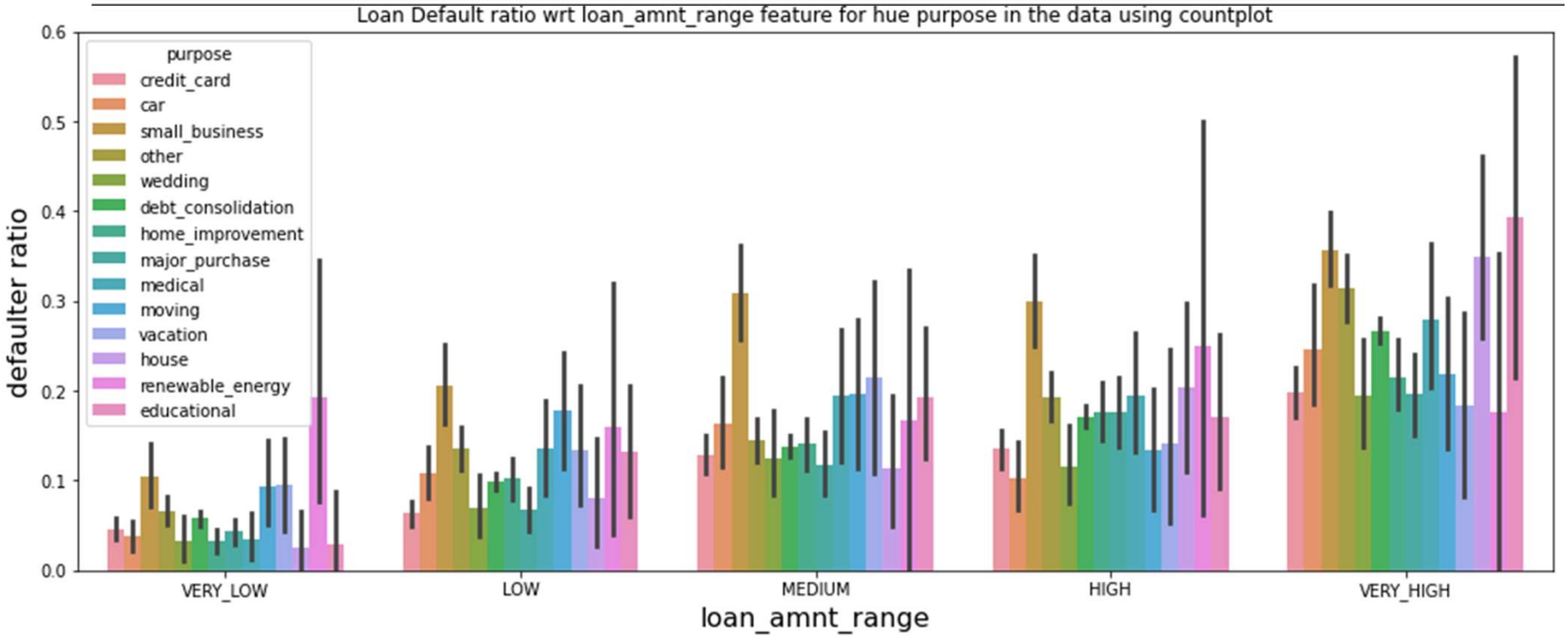
term vs purpose: As we can see straight lines on the plot, default ratio increases for every purpose w.r.t terms
related - Y



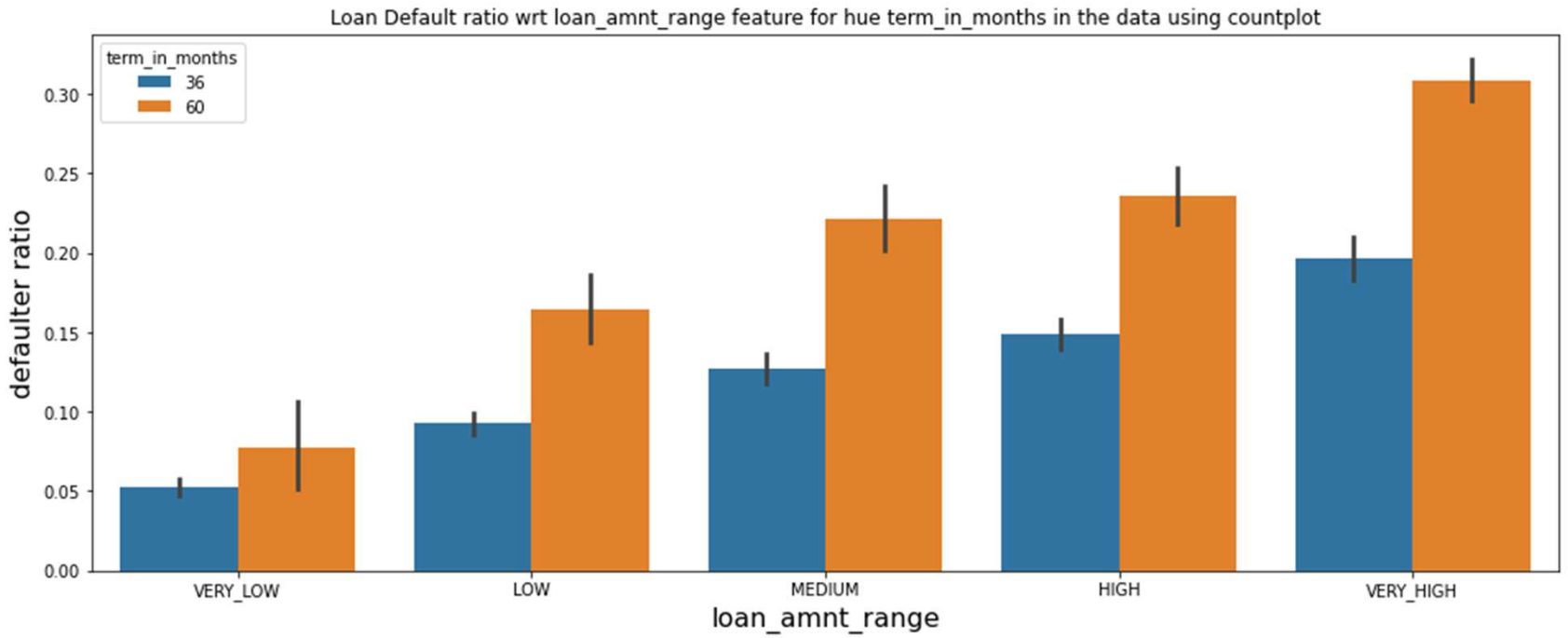
grade and purpose: As we can see straight lines on the plot, default ratio increases for every purpose w.r.t grade
related - Y



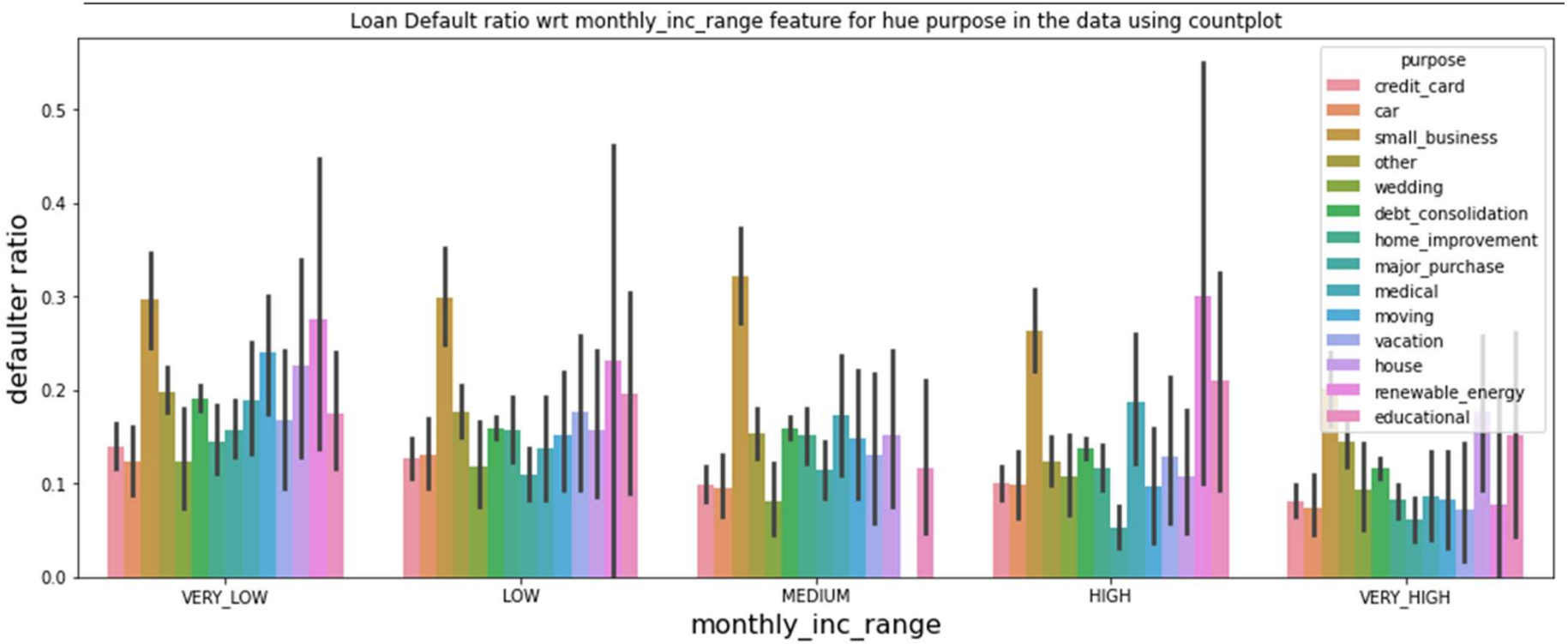
Loan amount range and purpose: As we can see straight lines on the plot, default ratio increases for every purpose w.r.t loan_amnt_range
related - Y



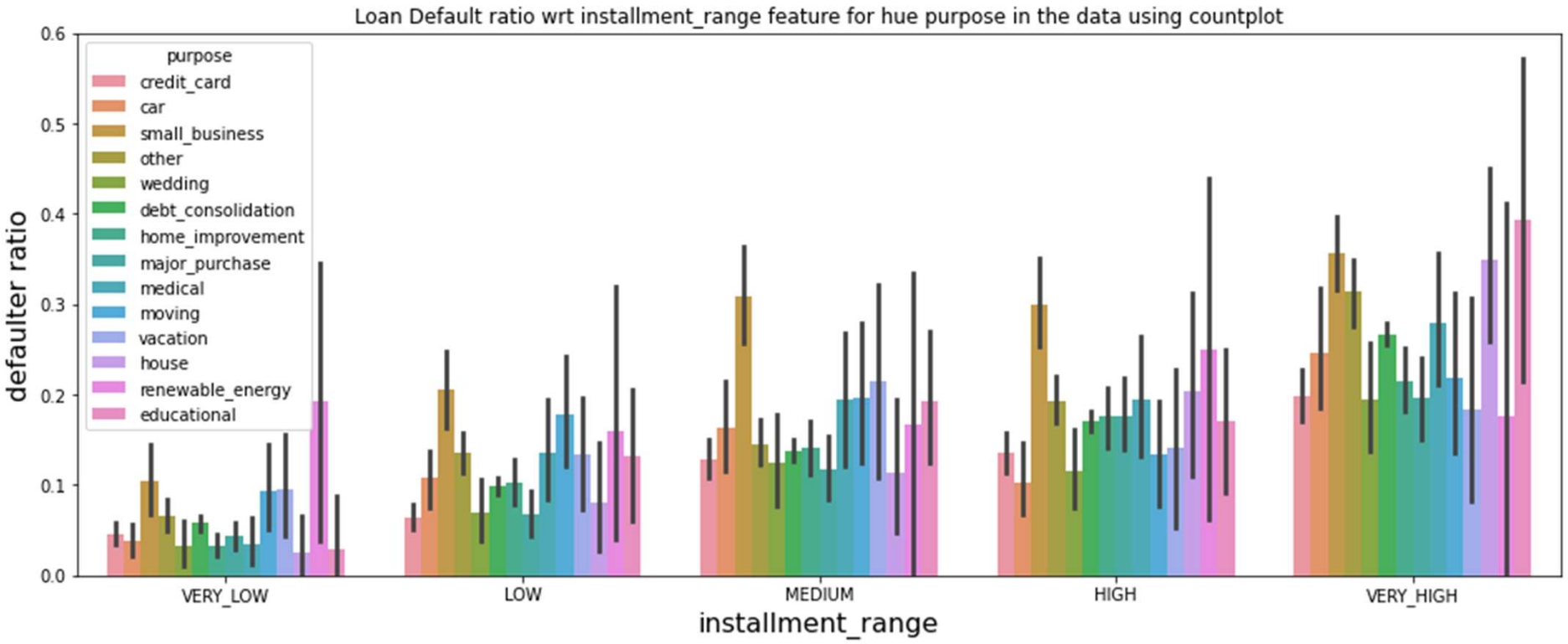
Loan amount range and term in months: As we can see straight lines on the plot, default ratio increases for every purpose w.r.t loan_amnt_range
related - Y



Monthly income vs purpose: As we can see straight lines on the plot, default ratio increases for every purpose w.r.t monthly_inc_range
related - Y

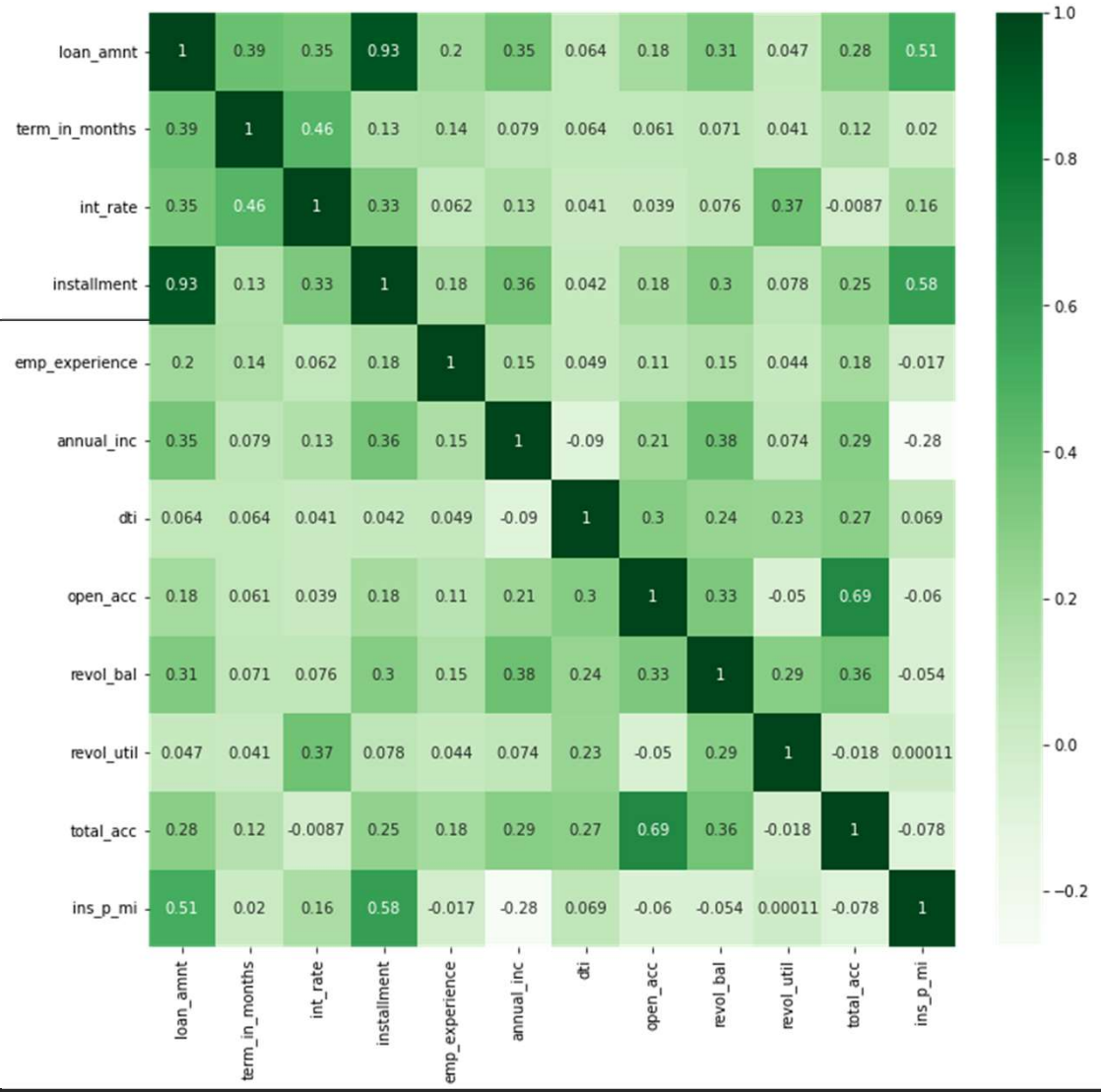


Instalment range and purpose: As we can see straight lines on the plot, default ratio increases for every purpose w.r.t installment except for small business
related - Y



Heat map b/w numerical variable

We can observe there are some number of variables are little co-related in numerical variable.



Conclusion

- From Univariant Analysis the Beneficial Variable we got are as fallows:
  - home_ownership: defaulters rate varies
  - term_in_months: defaulters rate decreases
  - Grade: defaulters rate decreases(B>C>D>E>A>F>G), A has to be verified.
  - sub_grade: defaulters rate decreases(B5-G5)
  - Purpose: defaulters rate is varies
  - dti_range: defaulters rate increases
  - Int_range: defaulters rate decreases
  - installment_range: defaulters rate increases
  - monthly_inc_range: defaulters rate
  - Loan_amnt: We can observe some outliers and the first quartile is bigger than third quartile for loan amount which means most of the defaulters clients are from first quartile

- From Bivariant Analysis Two variables which are related are as fallows W.r.t Defaulters ratio:
  - loan_amnt_range and purpose:
  - grade and purpose:
  - loan_amnt_range and term_in_months:
  - monthly_inc_range and purpose:
  - term vs purpose:
  - Instalment_range vs purpose:

- Through Heat map we observed that there is not that much co-related with each other for numerical variable.
- We can consider the univariant variable and bivariant analysis variable for guessing the Defaulters.
- We can also divide loan according to bivariant variables for guessing Defaulters.