Lending Club Case Study

Background:

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'.

LendingClub wants 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.
1.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.
2.The company can utilize this knowledge for its portfolio and risk assessment.

Business Objective:

To identify variables which are strong indicators of default and potentially use the insights in approval / rejection decision making.

Data Fetching and Cleaning Operations

Analysis Approach:

- We have to analyze the data to identify the driving factors
- To start with the data analysis, the data set had to be cleaned and prepared for the analysis, then analyzed and visualized.
- Clean the dataset using appropriate methods
- Once the data set was cleaned, univariate, bivariate and mulitvariate analysis need to be done
- Visualization techniques were used to draw charts and graphs and then meaningful insights from them

Data Cleaning:

- We have analyzed the past loan data for all loans issued through the time period 2007 to 2011.
- Multiple Columns with more than 70% of null values were dropped.
- Missing values were imputed into respective stat, so that all the NULL and NAN values were removed.
- Descriptive columns, other unnecessary columns are removed from the dataset , so that we don't waste time analyzing those fields

Data Understanding

Types of variables that we need for analysis

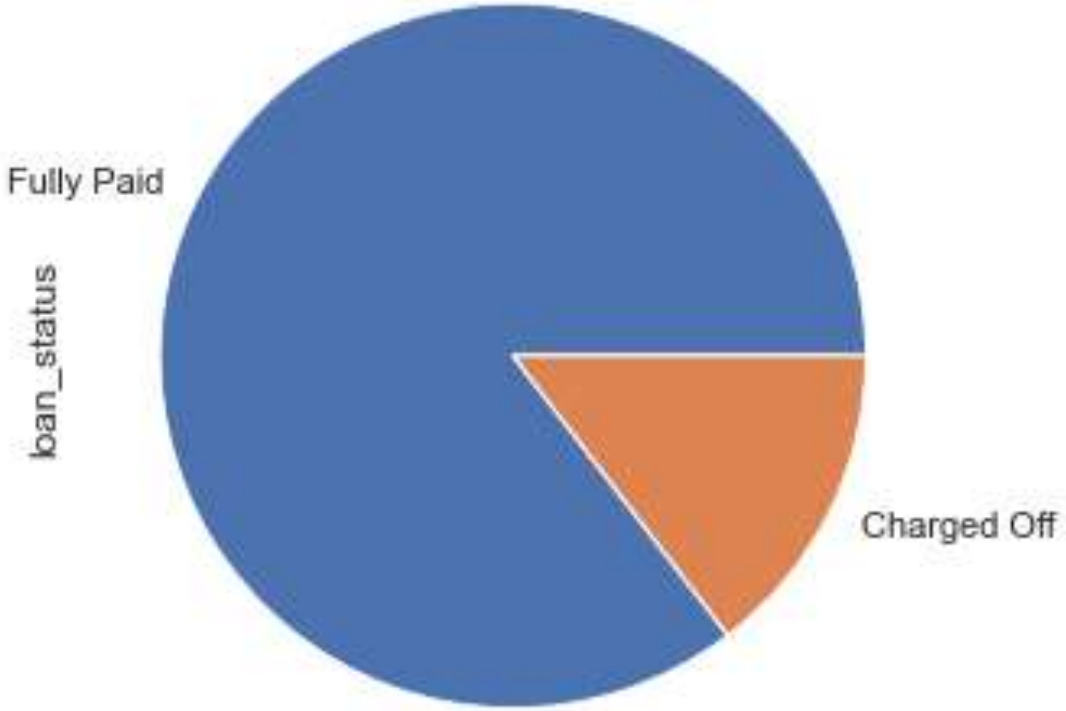
- Categorical Variables
  - Purpose
  - term
  - verification_status
  - home_ownership
  - grade and sub_grade
  - loan_status
- Numeric Variables
  - loan_amnt
  - funded_amnt
  - funded_amnt_inv
  - int_rate
  - emp_length
  - annual_inc
  - issue_d
  - dti
  - installment
  - deliquency values

Data Understanding - Overall Default Rate is 14.6%

loan_status:

Charged Off : 14.586412 %

Fully Paid : 85.413588 %



Conclusion:

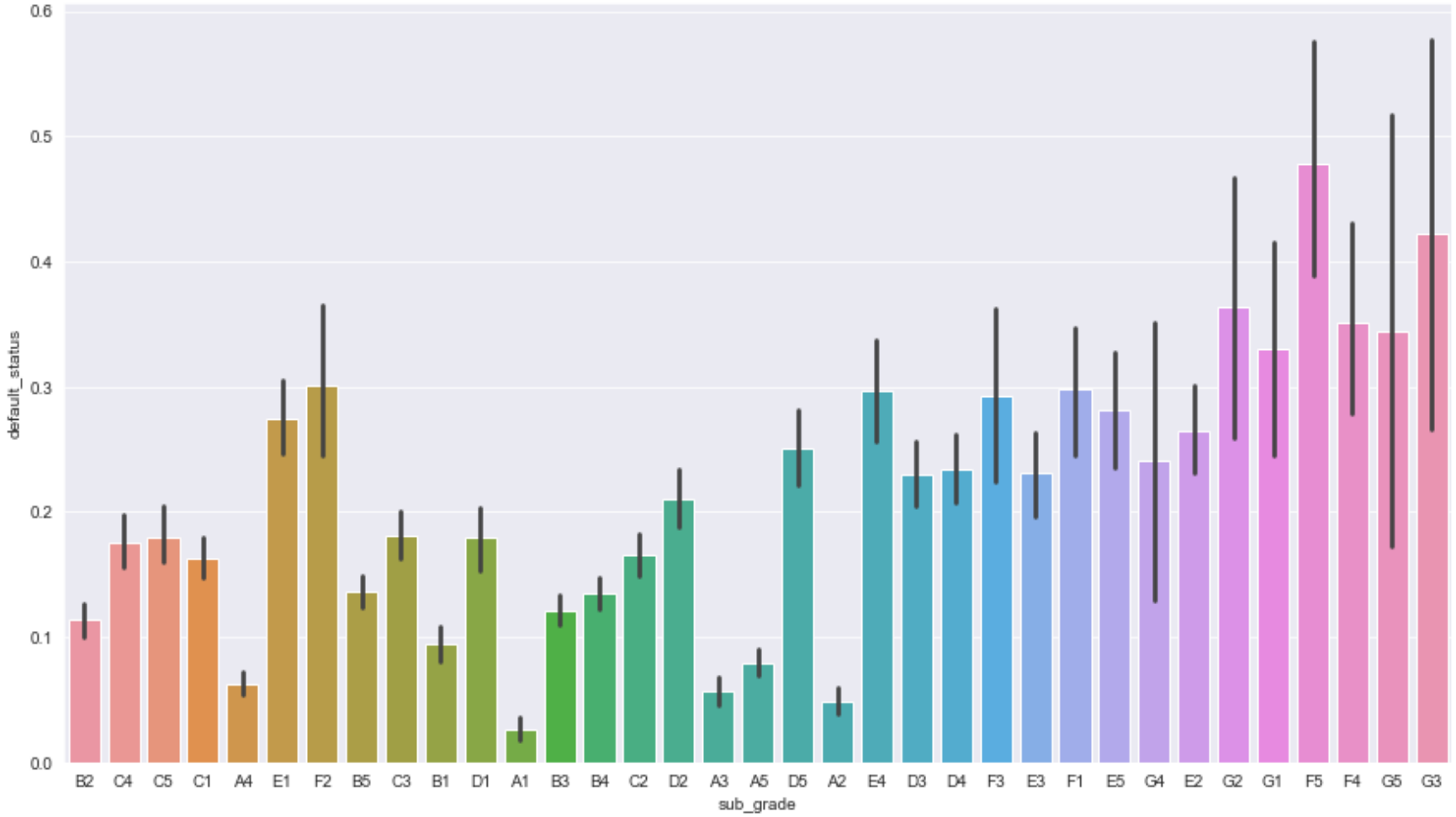
- **Nearly 14.6% of loans are Charged off loans**
- **Now we need to find out the driving factors among the considered columns for this defaulter percentages**

- **In some parts of ppt and code , you can see the following variable for loans default_status(0- Fully paid, 1- charged off)**

UniVariate Analysis

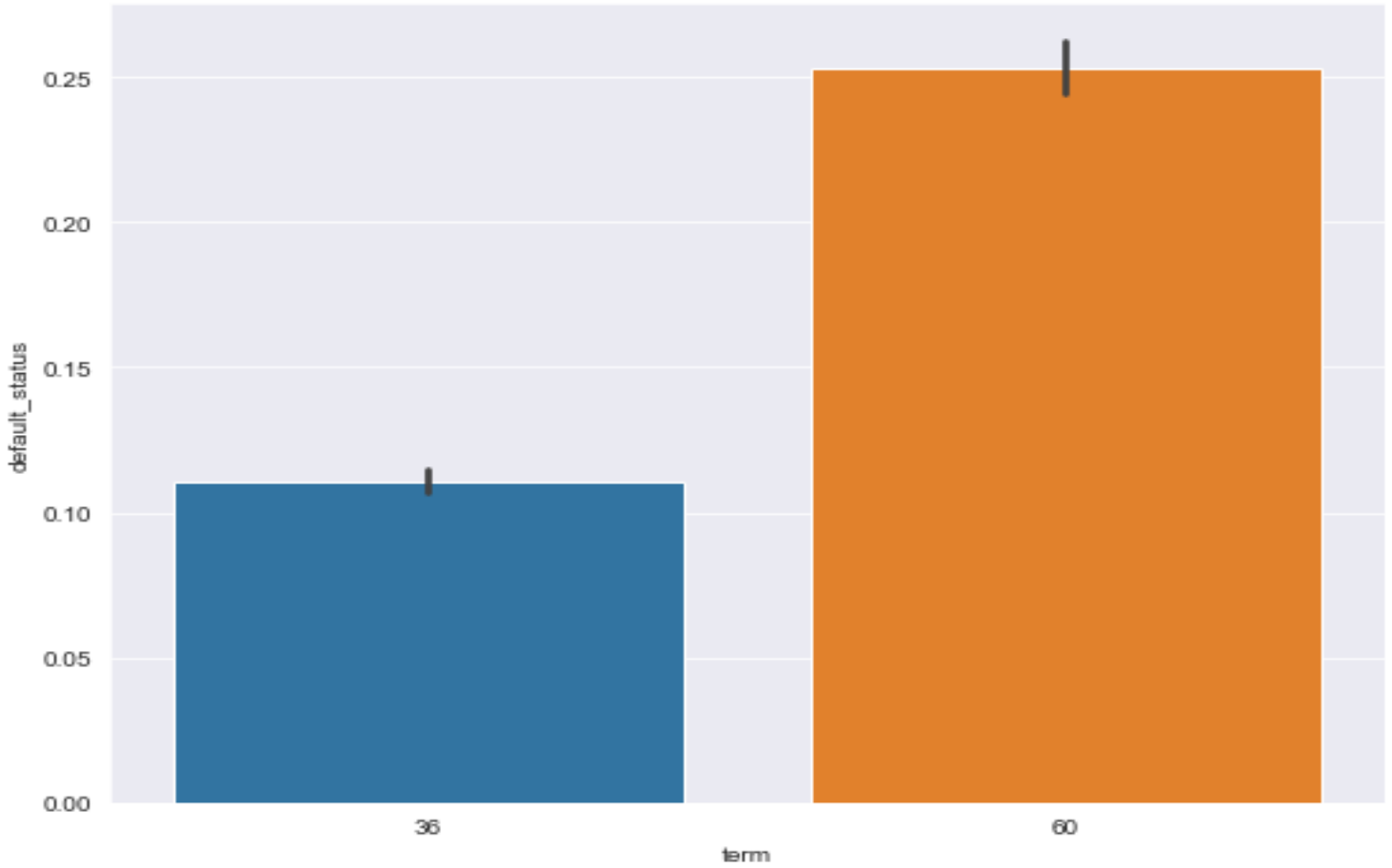
Grade and Sub-Grade

1. F5 grade has highest ratio
2. As whole grade level ,G has Highest defaulters
3. among the G grade G3 sub_grade has highest defaulter ratio.
4. Higher the Grade of the Loan , Higher is the risk of becoming defaulter.



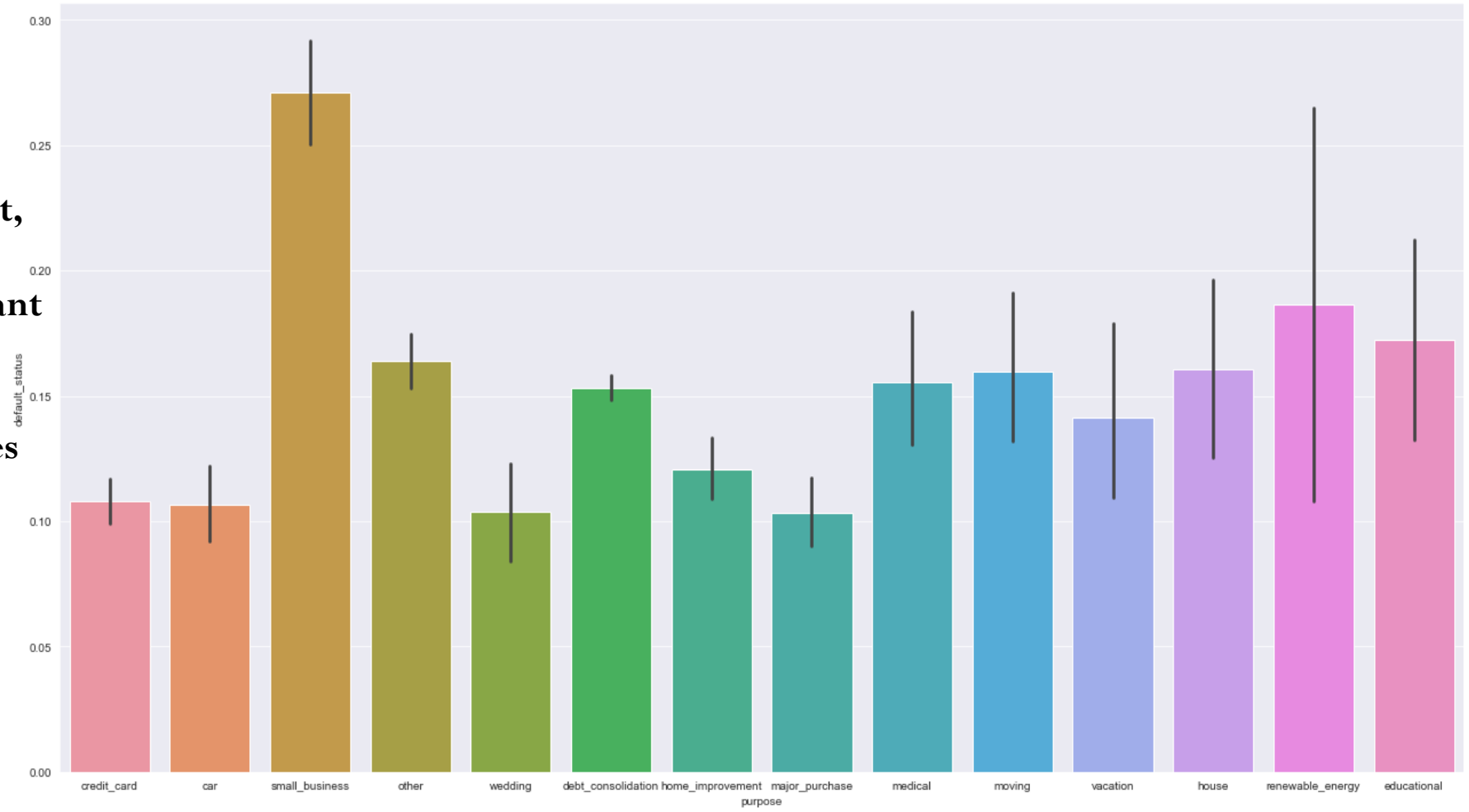
Term of the Loan

The graph suggests that
'Higher the Term of the Loan ,
Higher is the risk of becoming defaulter



Loan Purpose

1. small_business purpose has highest defaulter rate.
2. debt_consolidation, credit_card,home_improvement, other,major_purchase, 'small_business' are the important purposes that we need to consider as they are covering most(~85-90%) of loan purposes



Other Important Univariate Conclusions

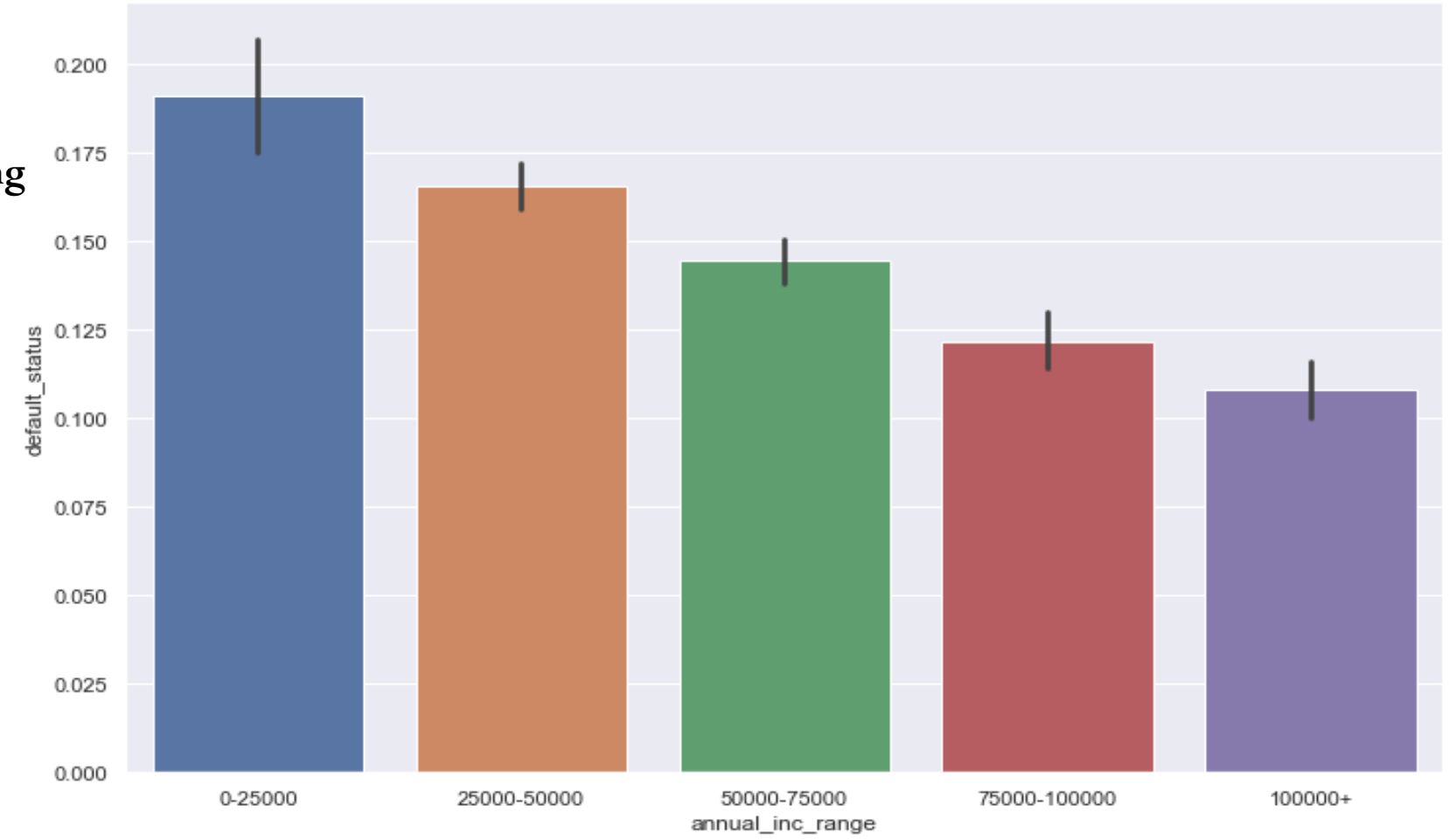
1. Among the verification statuses ,the ‘Verified Status’ Loans have higher Default Ratio when compared to other status defaulter’s ratio
2. The 'other' home ownership customers have higher Default Ratio
3. Year and month of loan issuing didn't help much in analysis

Note:
We observed the individual numerical data graphs for fields [interest_rate,annual_income, dti,loan_amount], we saw that the density and fluctuations in them are specified to particular ranges for respective fields. Though we can get some insights, they might not be of much help, so instead we will do some segmentation on them and do analysis furthur.

Segmented Univariate Analysis
Annual Income Range

Lower the Annual Income range of customer, higher the risk of falling into defaulter state

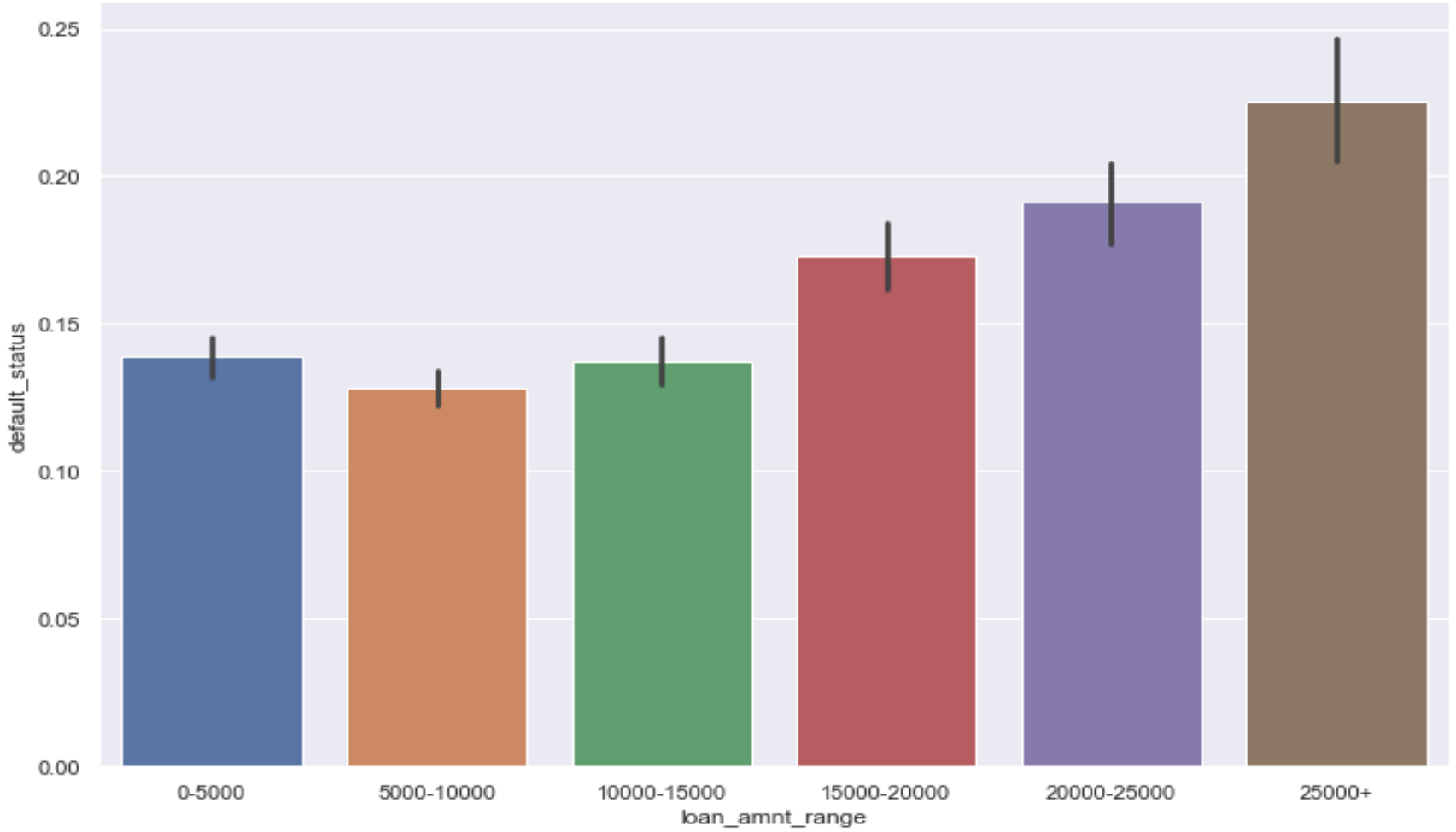
In our Dataset, we have
Minimum Income : 4000$
Maximum Income : 600000$



Loan Amount Range

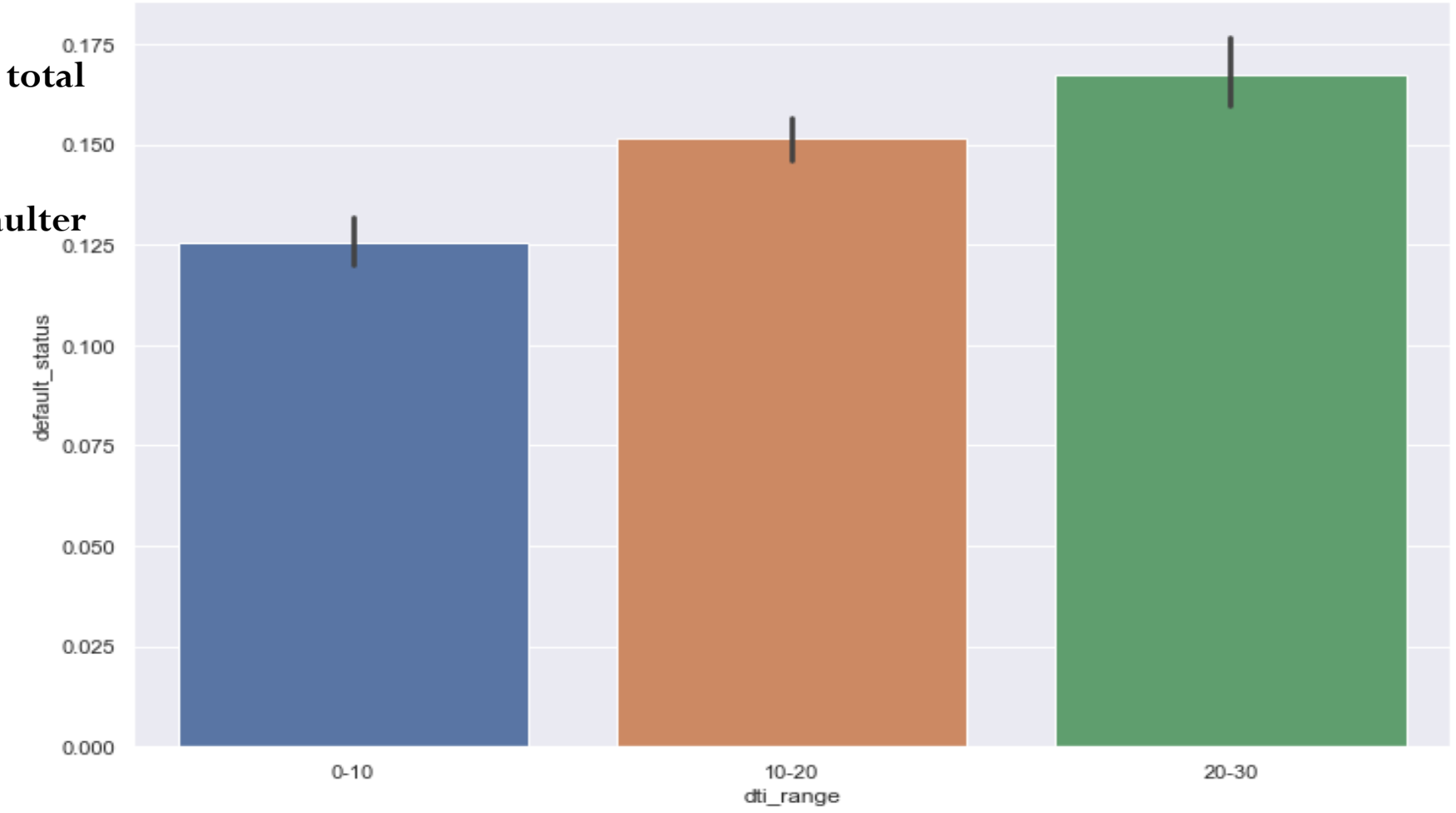
Higher the Loan amount of customer, higher the risk of falling into defaulter state

In our Dataset, we have
Minimum Loan: 500$
Maximum Loan: 35000$



DTI Range

1. dti is defined : existing debt to total income ratio of customer
2. Higher the dti range of person, higher the risk of falling into defaulter state



In our Dataset, we have
Minimum dti : 0.0
Maximum dti : 29.99

Other Important Segmented Univariate Conclusions

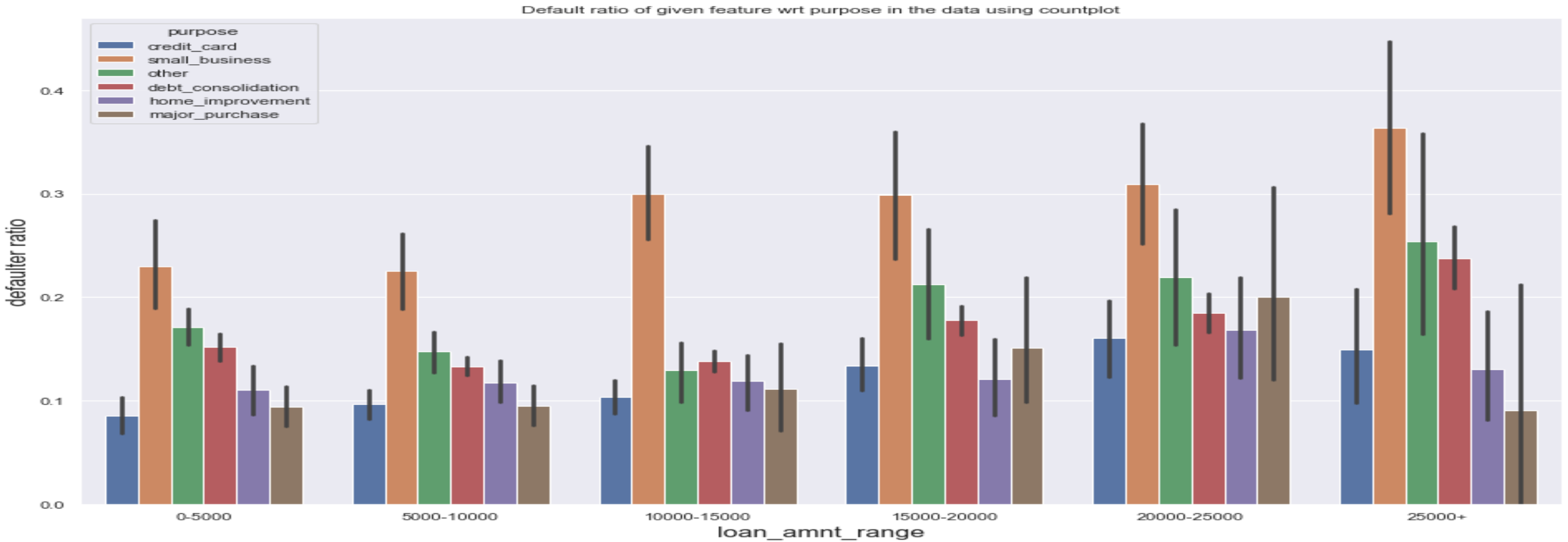
1. Higher the Interest rate for the loan of customer, higher the risk of falling into defaulter state
2. Funded amount Inv is similar to that of Loan amount in terms of trend, so higher the range of funded-amnt_inv, higher the risk of falling into defaulter state
3. Higher the installment amount for a given loan, higher the risk of falling into defaulter state

Note:

For upcoming Bivariate analysis, when ever we do data analysis with respect to purpose, we use debt_consolidation,credit_card,home_improvement,other,major_purchase,'small_business' purposes as they are covering most(~85-90%) of loan purposes because these will be sufficient from Business standpoint to come-up with useful patterns

Bivariate Analysis

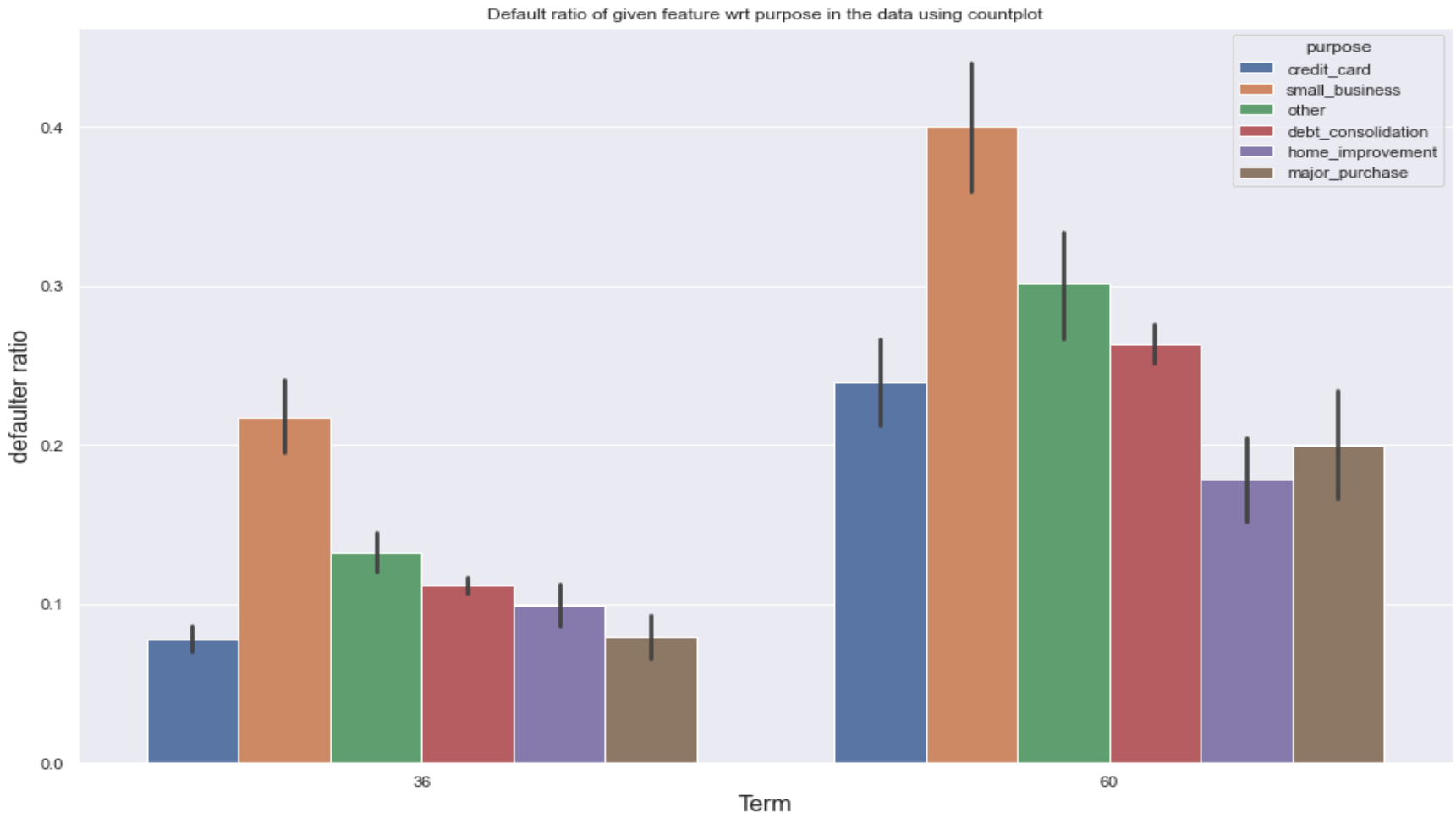
loan_amnt_range wrt purpose



loan_amnt_range wrt purpose , we can observe the similar upward trend for mostly all the purposes in above graph.Thus, defaulter ratio increases with increase of loan amount

Term of loan wrt purpose

Term of loan wrt purpose , we can observe the upward trend for all the purposes . Hence defaulter ratio increases with increase of term period for purposes

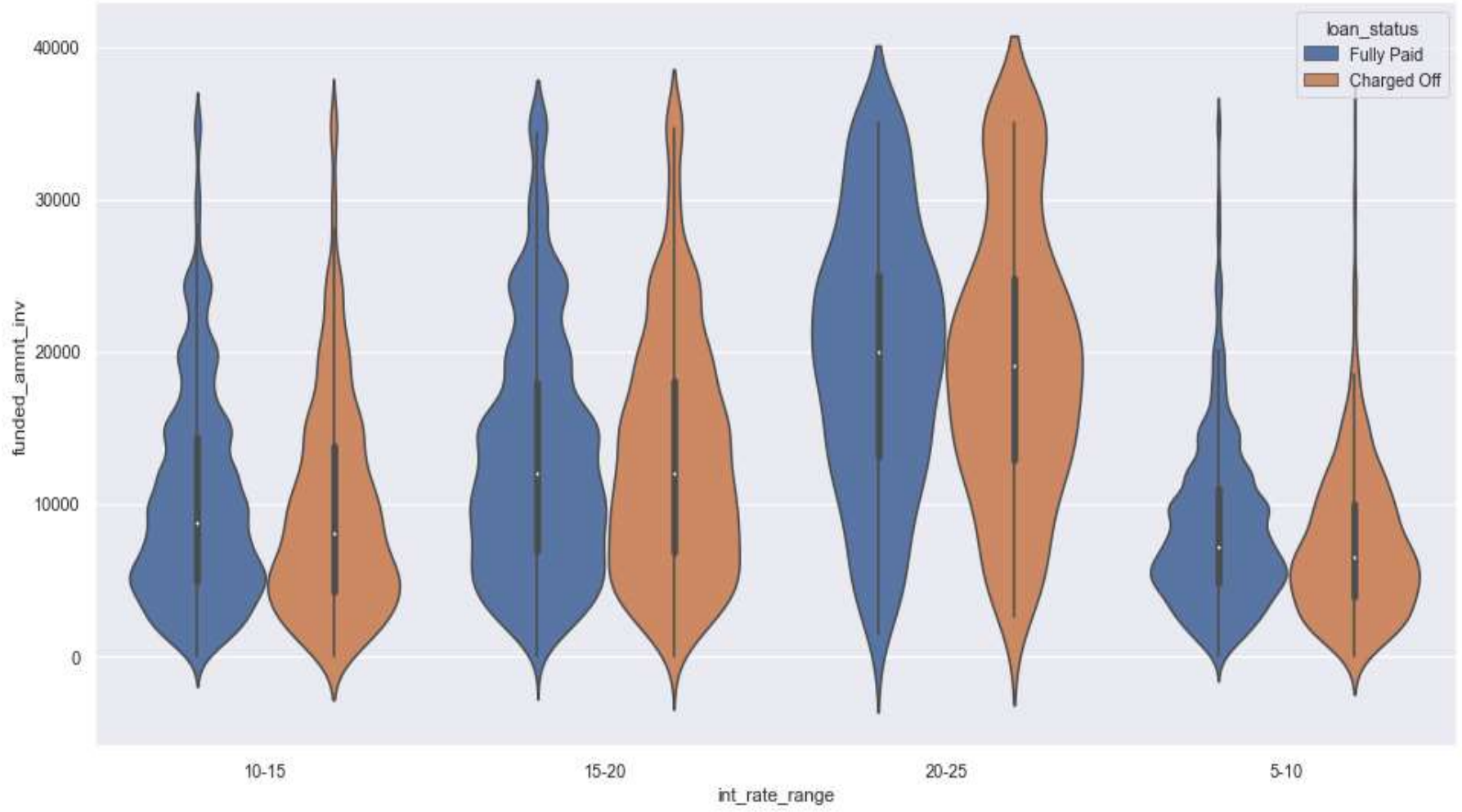


Other Important Observations

1. funded_amnt_inv_range wrt purpose , we can observe the similar or upward trend for mostly all the purposes , defaulter ratio increases with increase of loan amount.
2. annual_inc_range wrt purpose , we can observe the similar or downward trend for mostly all the purposes , defaulter ratio is highest for low income ranges
3. interest range wrt purpose , we can observe the upward trend for all the purposes , defaulter ratio increases with increase of interest rate on given loan.
4. installment_range wrt purpose , we can observe the similar or upward trend for mostly all the purposes , defaulter ratio increases with increase of installment amounts of loan.
5. dti_range wrt purpose , we can observe the upward trend for all the purposes , defaulter ratio increases with increase of dti ratio of loan

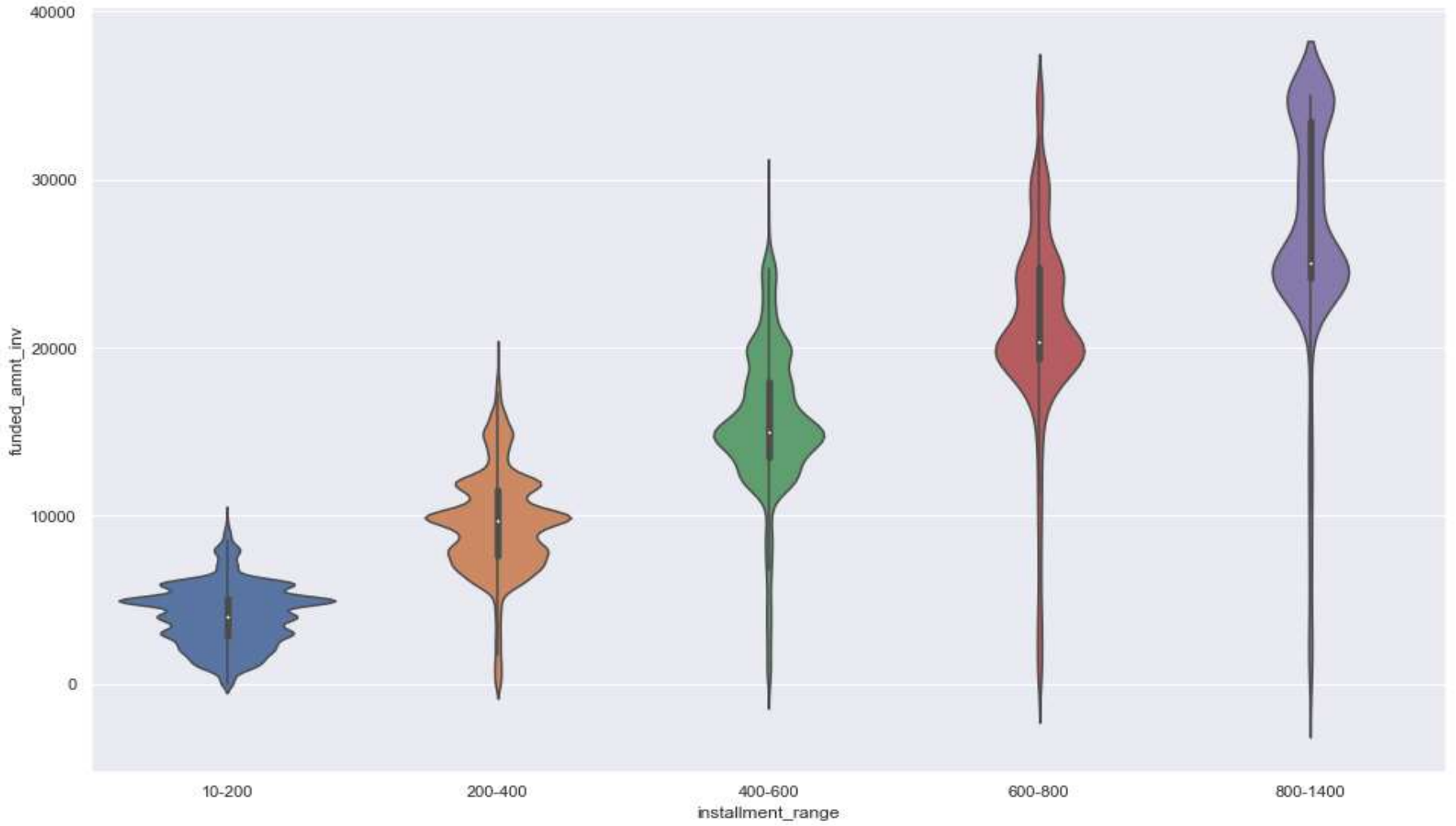
Funded amount wrt Interest Range

Funded amount wrt Interest Range, we can observe the width or distribution from central line Is getting increasing when we Move up interest rate for Funded amounts.



Funded amount wrt instalment Range

We are seeing here a simple straight line relationship among The medians of data indicating High correlation.
Since we know that both have Higher defaulters at higher ends, We need to be careful for this Combinations.



Other Important Observations

1. In Term wrt interest rate ,we can see the defaulter distribution is higher at higher interest rates with respect to term of loan(higher at higher term as well)
2. The above explanation holds true for term wrt funded amount as well

Multivariate Analysis:

•As an attempt for Multivariates, we tried set of 3 or more factors and how this combination effects Defaulter rate. We mentioned an example here with which we can very much of granular level analysis and have better insights over data

Multivariate Analysis

•purpose - small_business wrt installment-range 600-800 and 60 months term have the highest defaulter correlation and followed by other combinations

Since the values are of default_status , the values In the grid is actual distribution of defaulter values

Thank You

praneethvelamuri@gmail.com
kollaneeraja99@gmail.com