Gramener Case Study

Loan Default

How consumer attributes and loan attributes influence the tendency of default.

GROUP MEMBERS

ANKIT KUMAR MISHRA

ROHIT CHOUDHARY

Problem Statement

This 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'.

Objective

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.

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. The company can utilize this knowledge for its portfolio and risk assessment

Data understanding

Data cleaning

Data Analysis

Recommendations

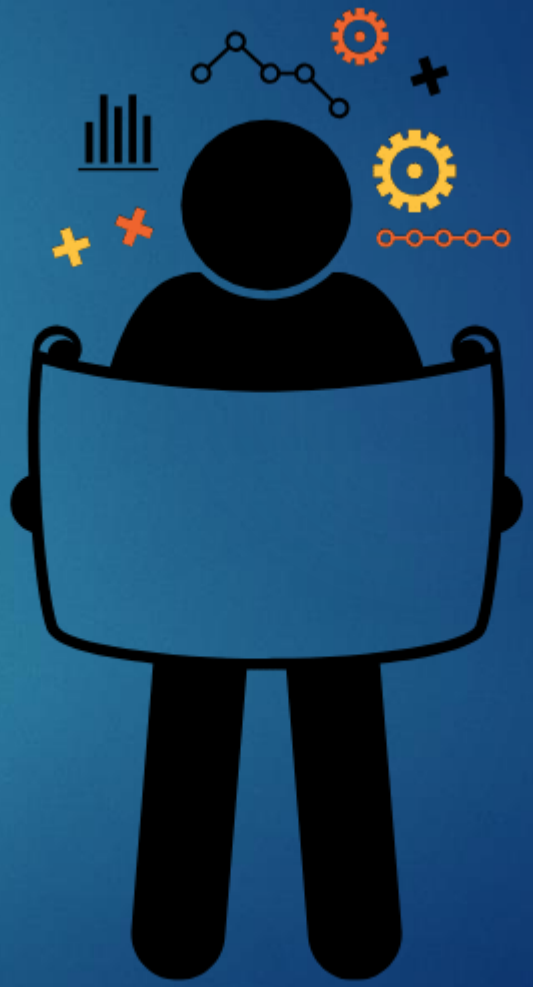
Data understanding

We will look at the first few rows of the data

By using - loan.head()

Looking at all the column names

By using - loan.columns



- Some of the important columns in the dataset are
  - loan_amount
  - funded_amount
  - verification_status
  - emp_length
  - int_rate
  - annual_inc term
  - funded_amount_inv
  - Purpose
  - grade
  - revol_util

Data cleaning

1) Summarize the number of missing values in each column.
2) The percentage of missing values in each column.
3) Removing the columns which have more than 90% missing values.
4) Drop specified label from columns and printing the dimensions.
5) There are now 2 columns having approx 32% and 64% missing values 'desc', 'mths_since_last_delinq' drop them.
6) Find the missing values in rows.
7) Checking whether some rows have more than 5 missing values.
8) The column int_rate is character type, convert it to float.
9) Extract the numeric part from the variable employment length.
10) Print the concise summary of dataframe.

We identified major columns

- loan_amount
- funded_amount
- annual_inc term
- funded_amount_inv
- inq_last_6mths
- open_acc
- verification_status
- installment
- emp_length
- int_rate
- home_ownership
- dti
- purpose
- grade
- revol_util

Data Analysis

UNI-VARIATE ANALYSIS

Univariate analysis is the simplest form of analyzing data. “Uni” means “one” has only one variable. It takes data, summarizes that data and finds patterns in the data.

SEGMENTED UNIVARIATE ANALYSIS

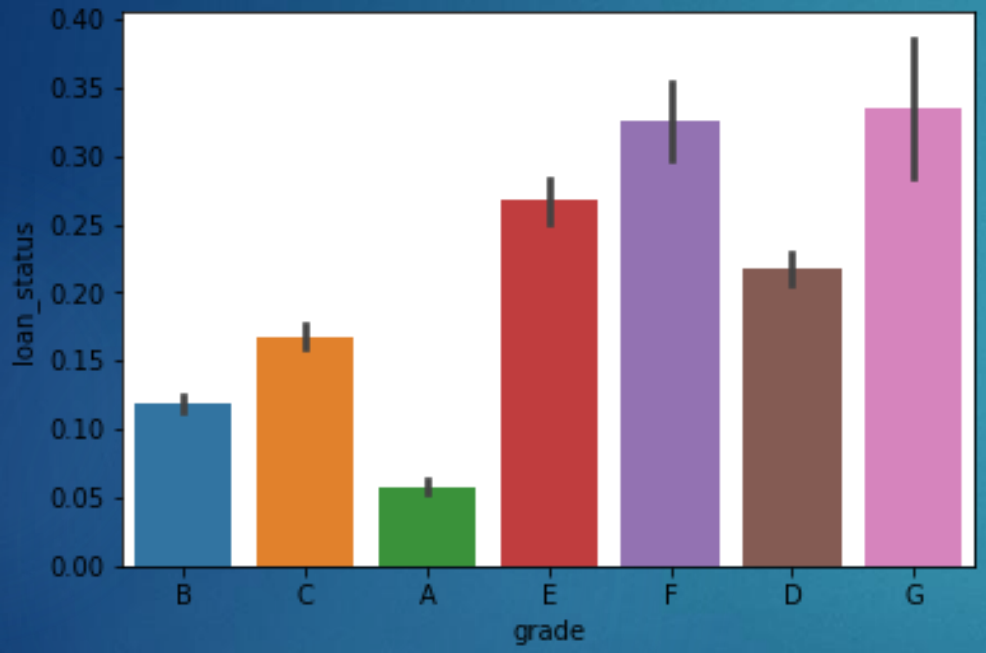
It allows you to compare subsets of data, which is a powerful technique because it helps you understand how a relevant metric varies across different segments. The way we approach this is by to figure out how to segment/group the variable into smaller buckets that we can compare.

DRIVER VARIABLES

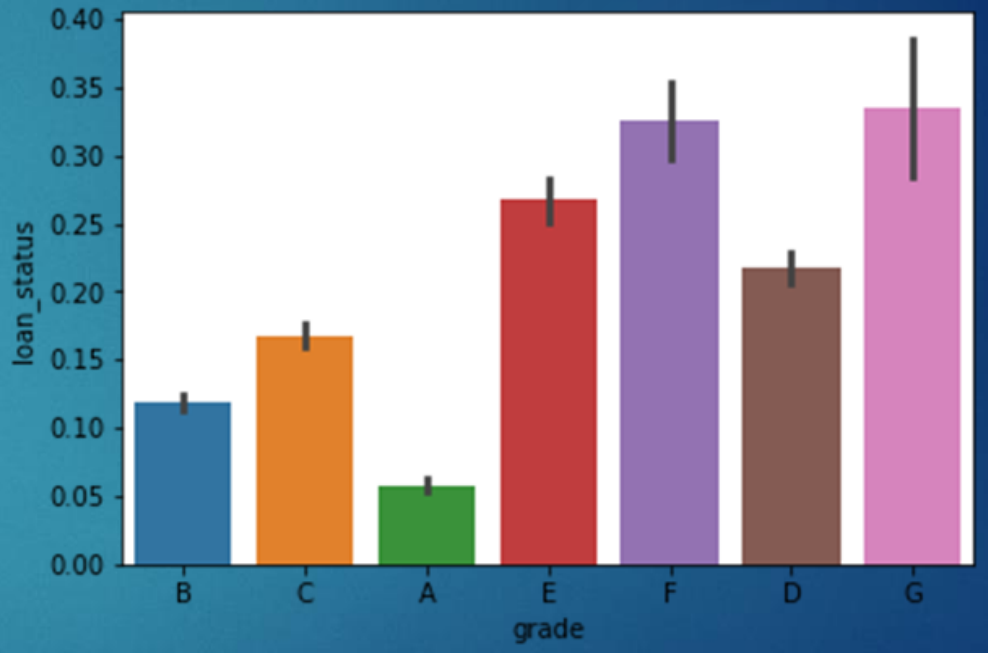
To understand the driving factors i.e. the driver variables, the variables which are strong indicators of default.

Uni-Variate Analysis

Plotting default rates across grade of the loan

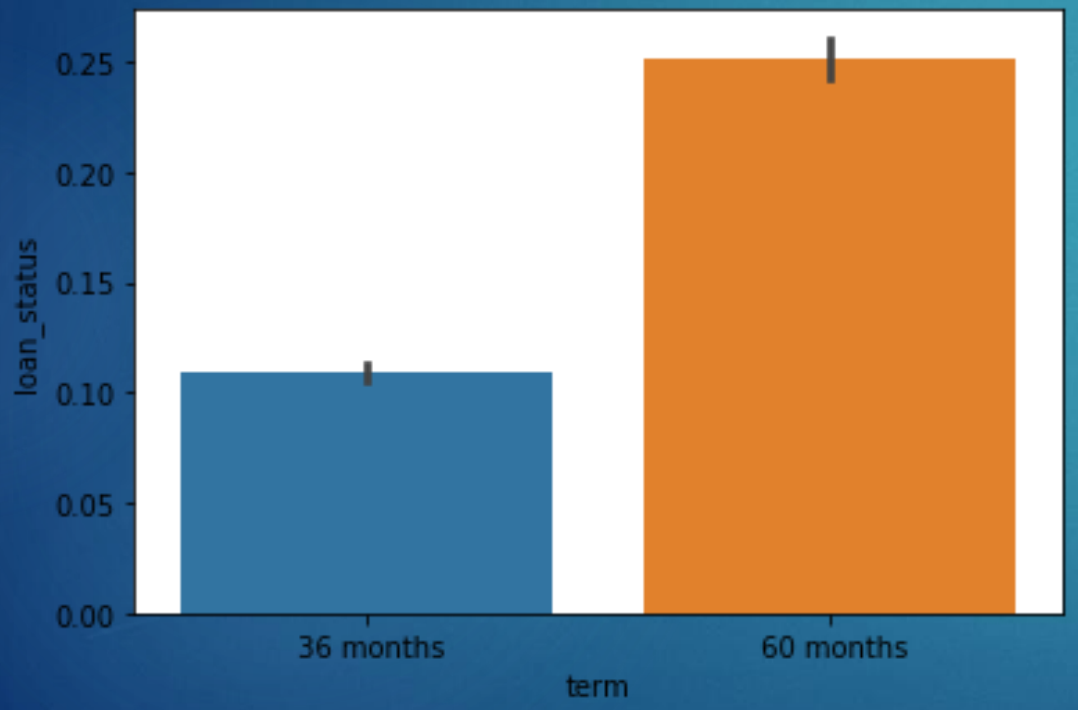


Compare default rates across grade of loan

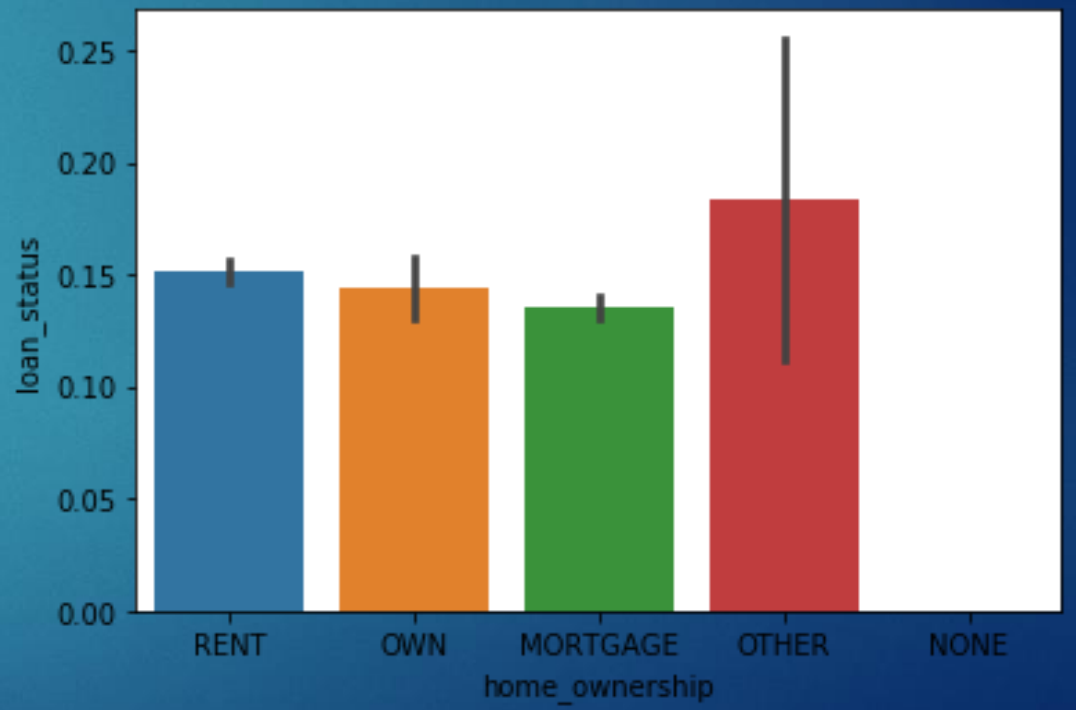


The grade of loan goes from A to G default rate increases. This is expected because the grade is decided by Lending Club based on the riskiness of the loan.

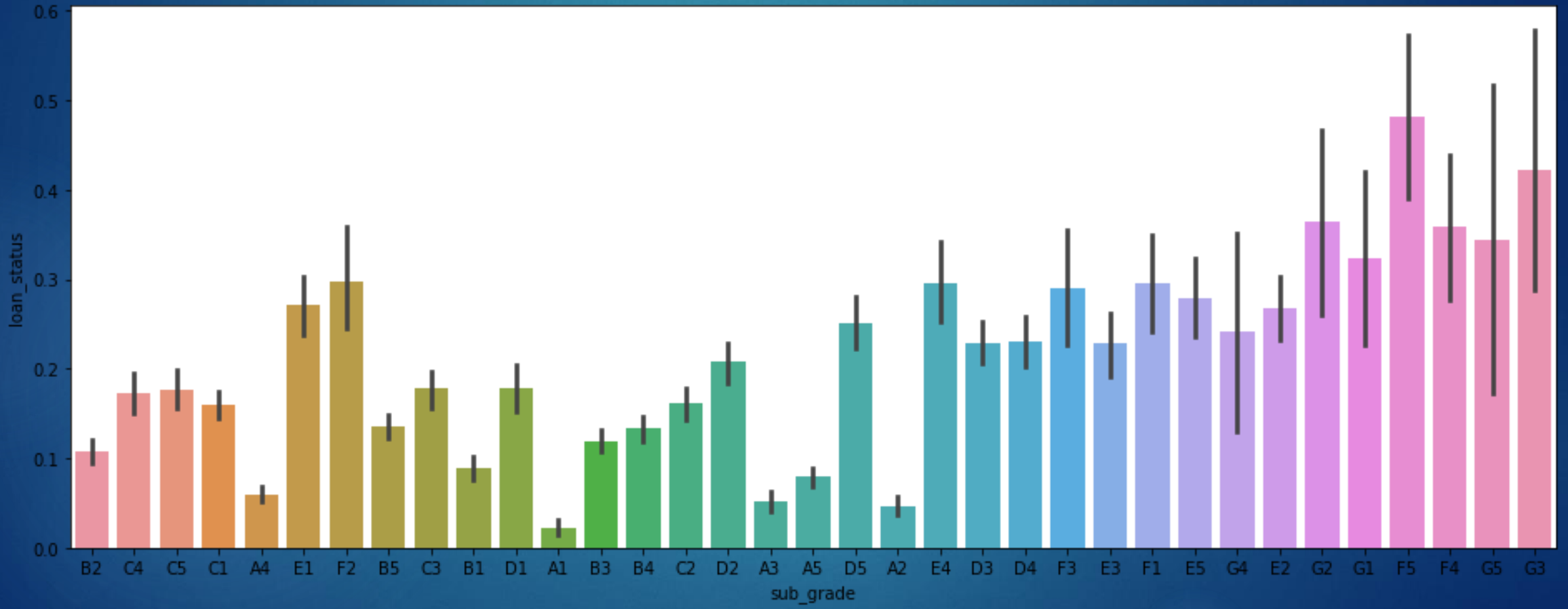
60 months loans default more than 36 months loans



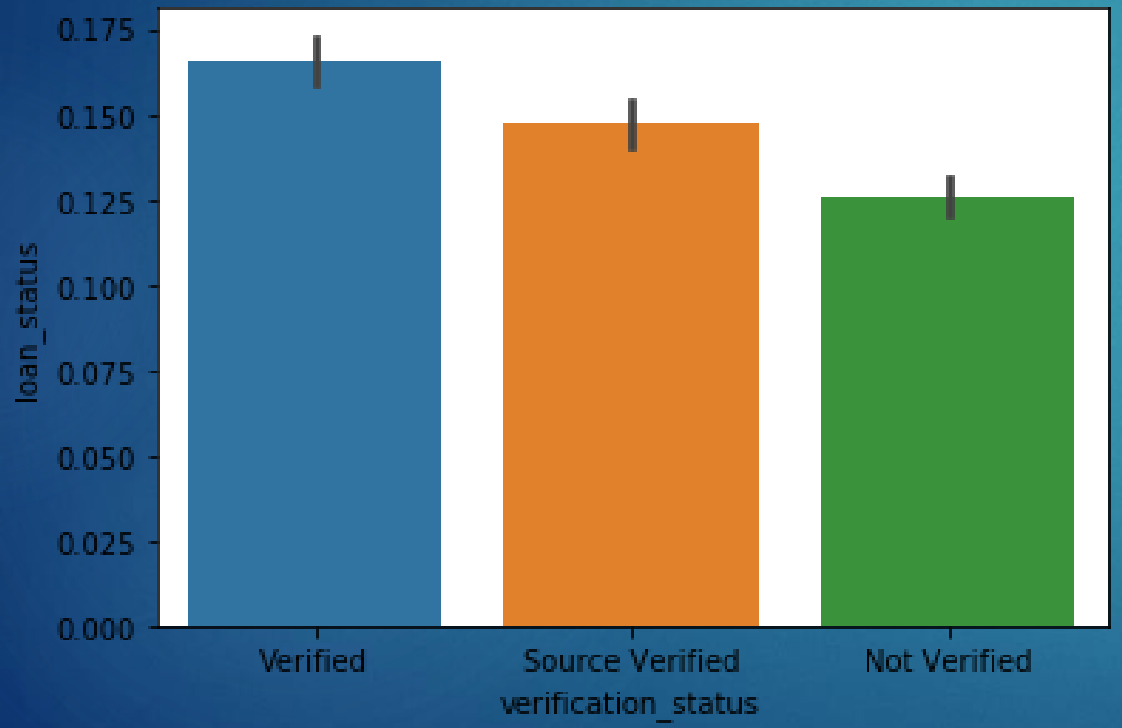
Home ownership: not a great discriminator



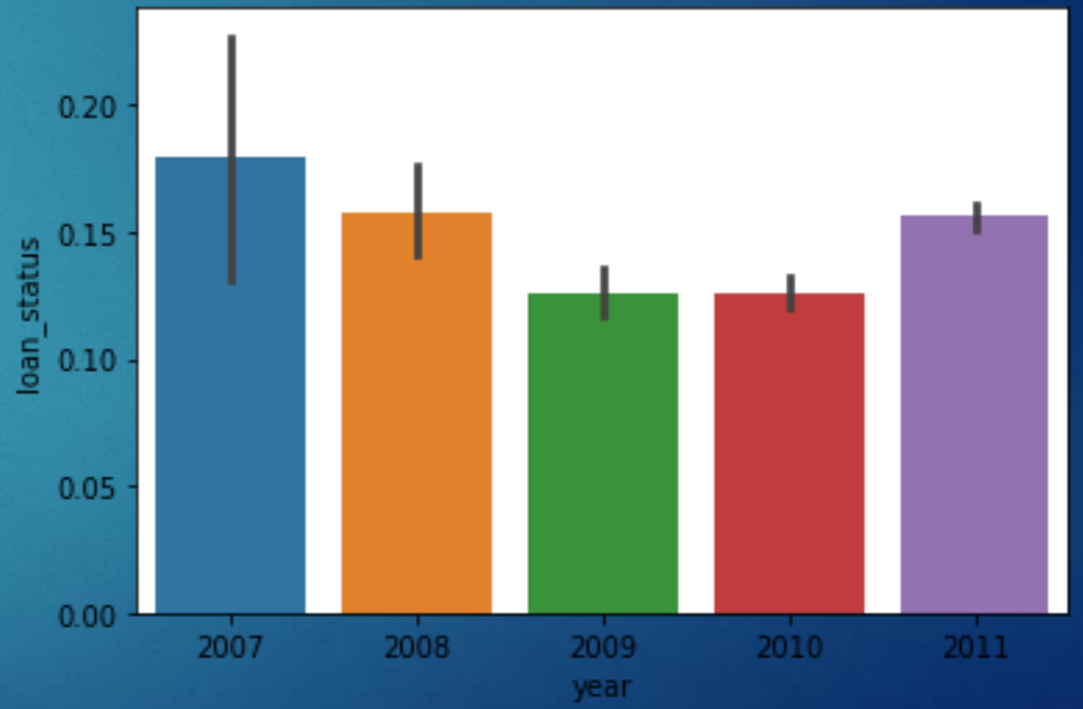
Sub-grade: as expected - A1 is better than A2 better than A3 and so on



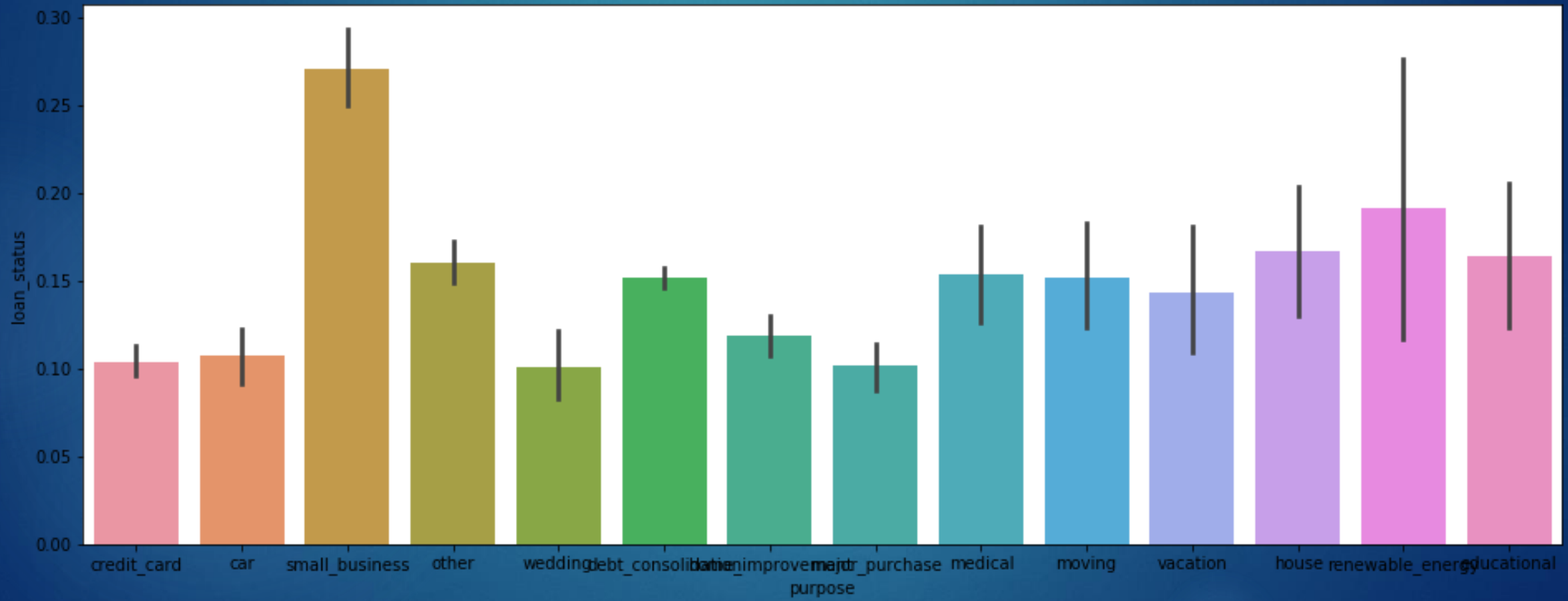
verification status: surprisingly, verified loans default more than not verifiedb



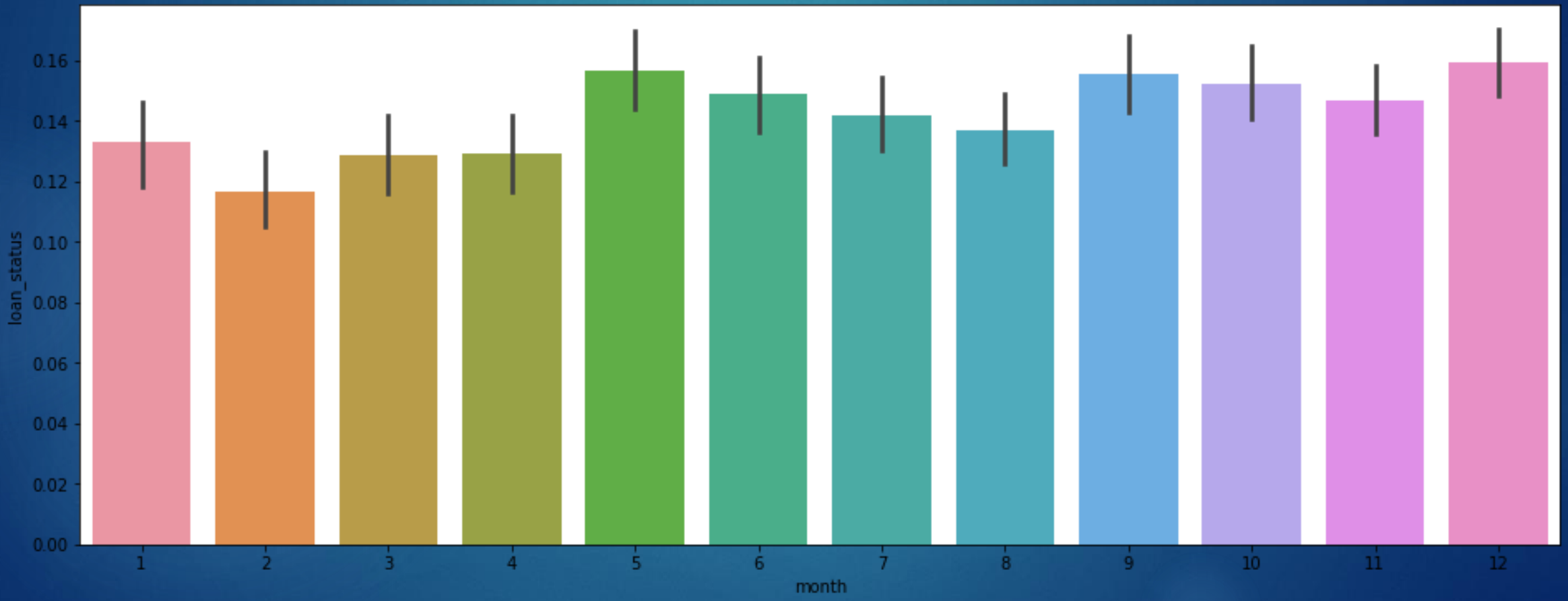
Default rates across years
the default rate had suddenly increased in 2011, in spite of reducing from 2008 till 2010



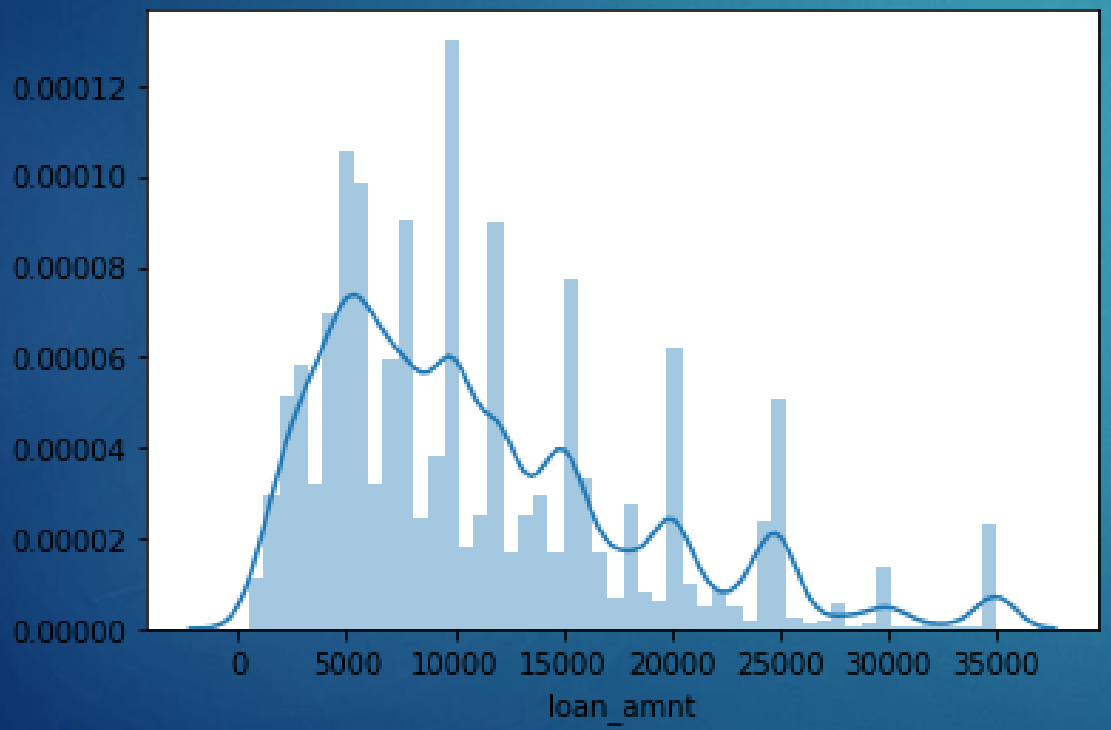
Small business loans default the most, then renewable energy and education



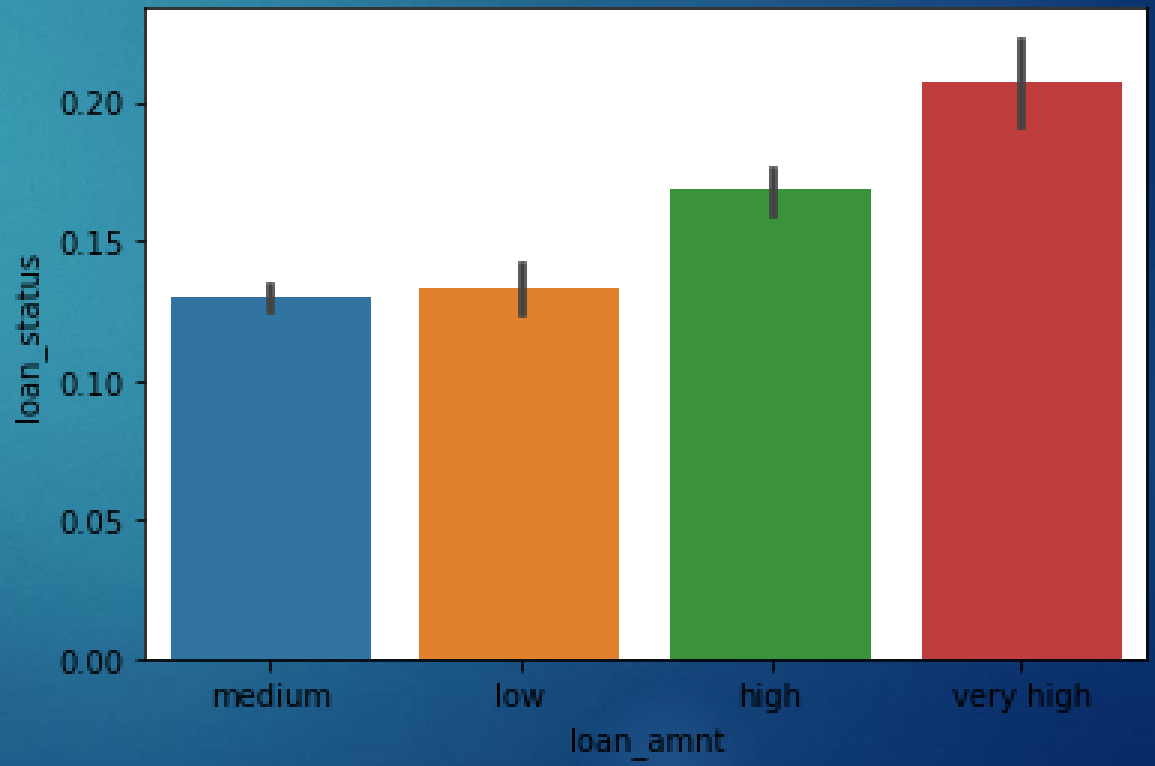
Comparing default rates across months: not much variation across months



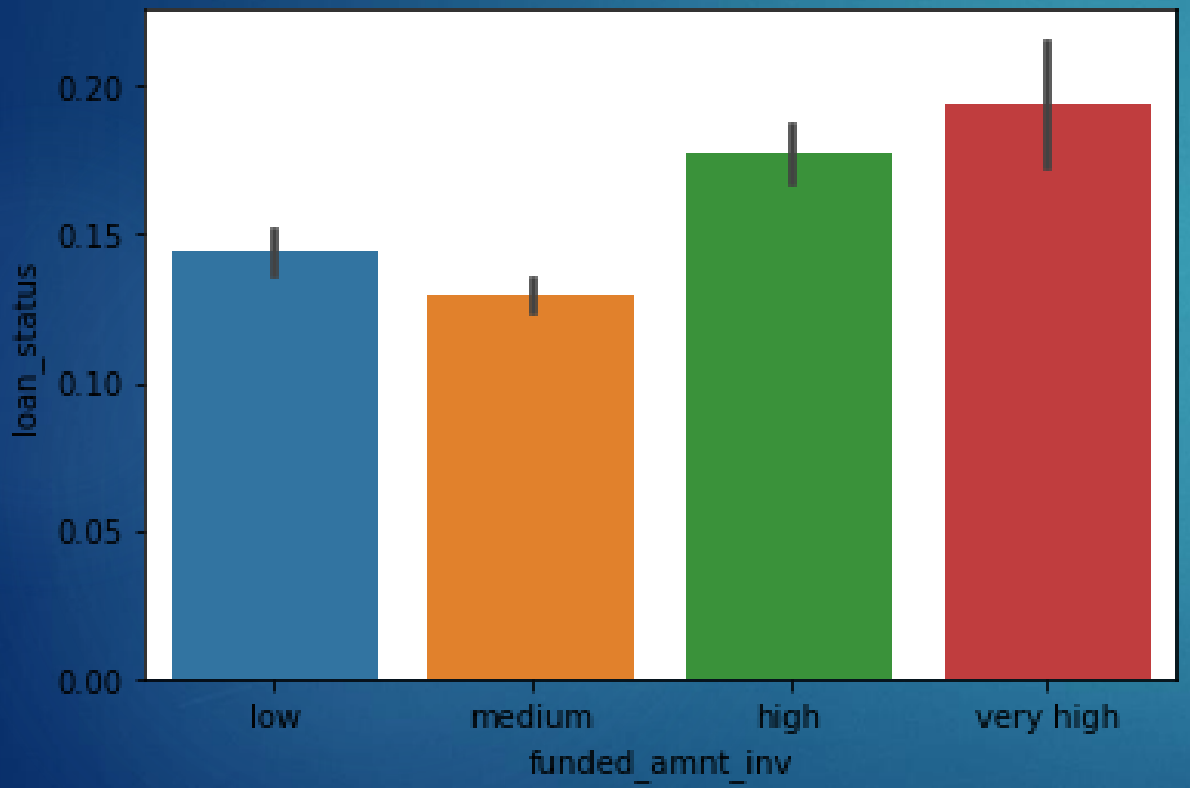
loan amount: the median loan amount is around 10,000



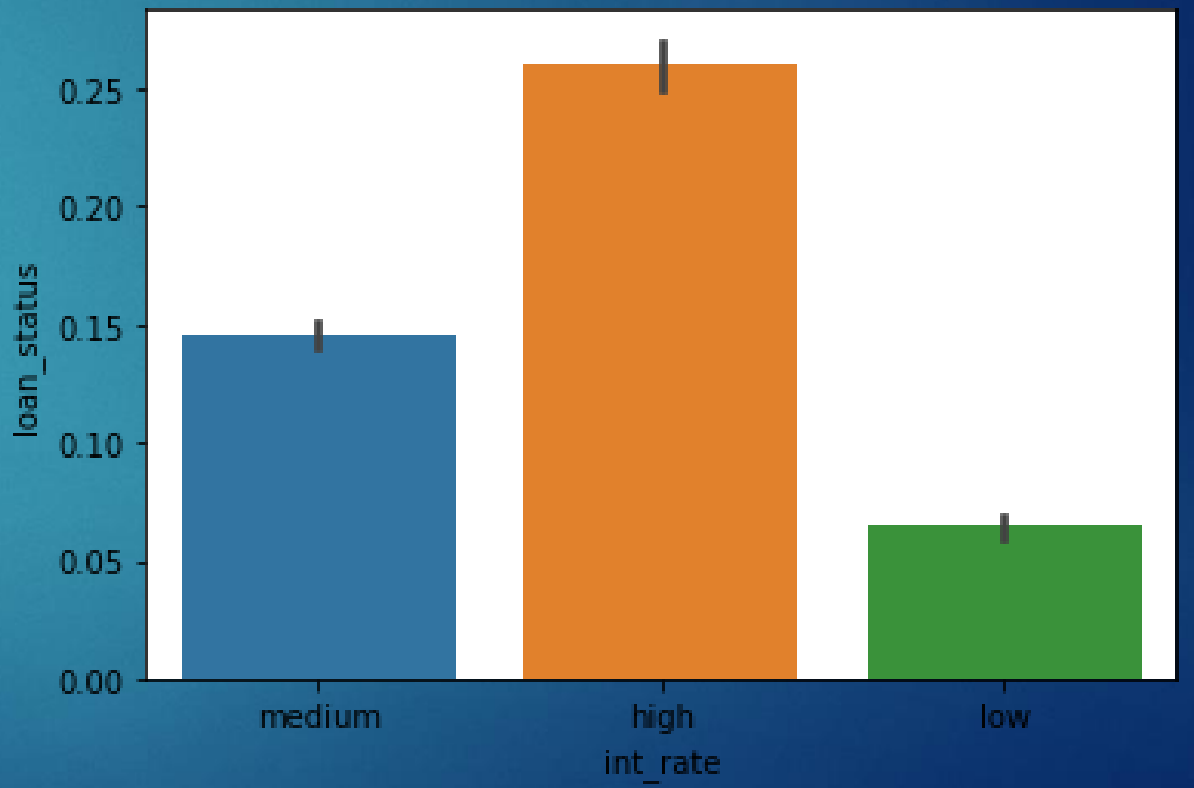
Default rates across loan amount type
Higher the loan amount, higher the default rate



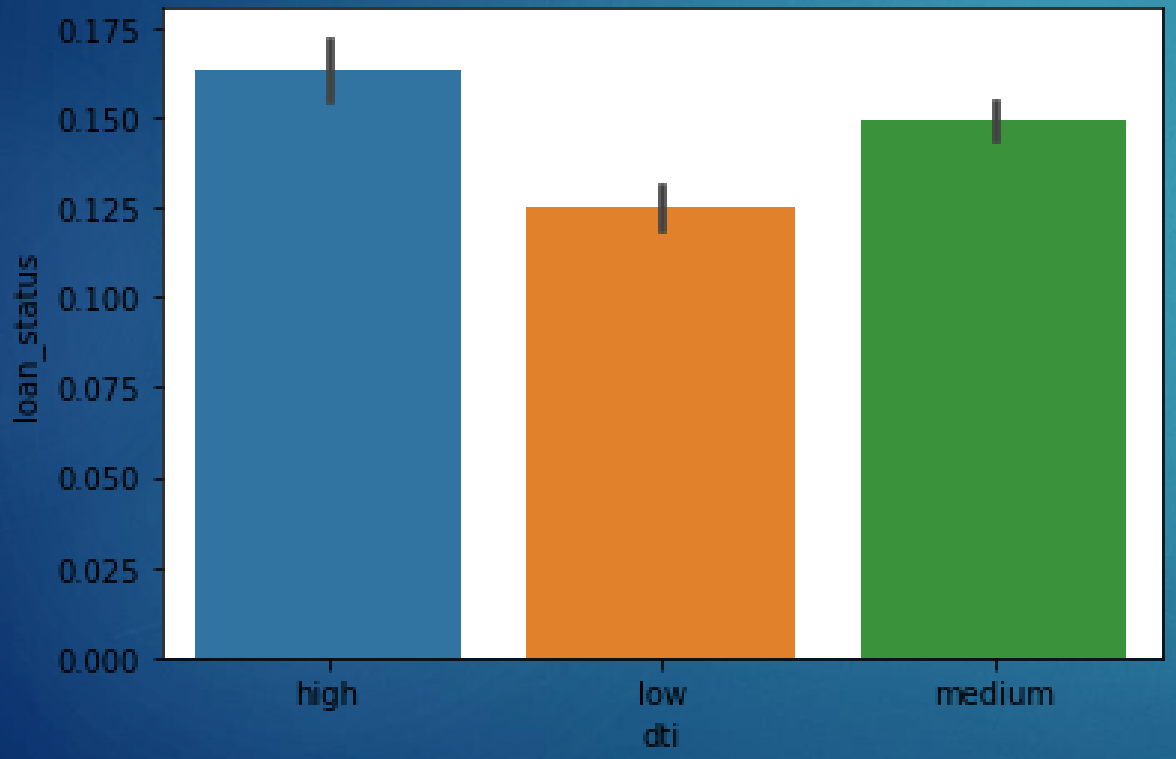
Funded amount invested



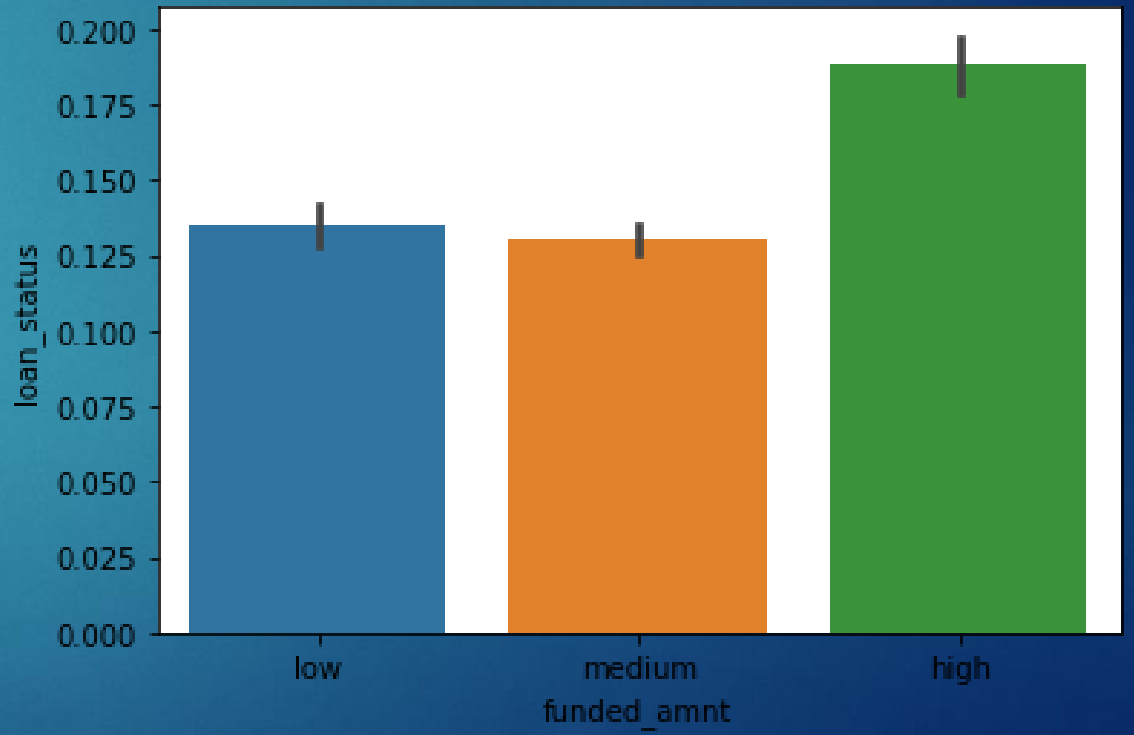
Comparing default rates across rates of interest
High interest rates default more, as expected



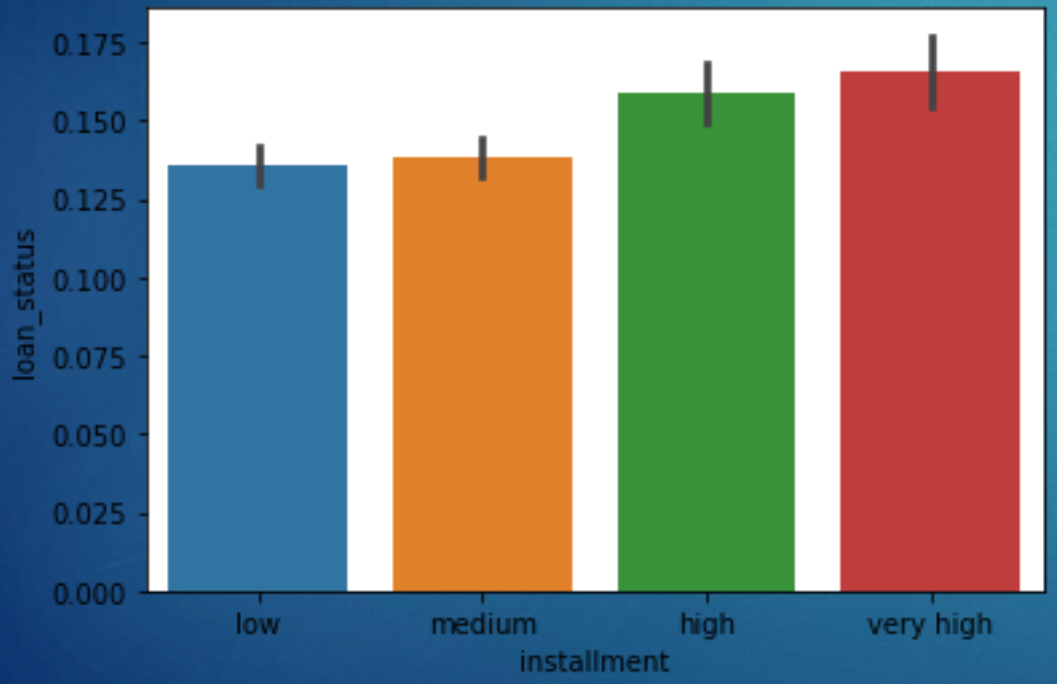
Comparing default rates across debt to income ratio
High dti translates into higher default rates, as expected



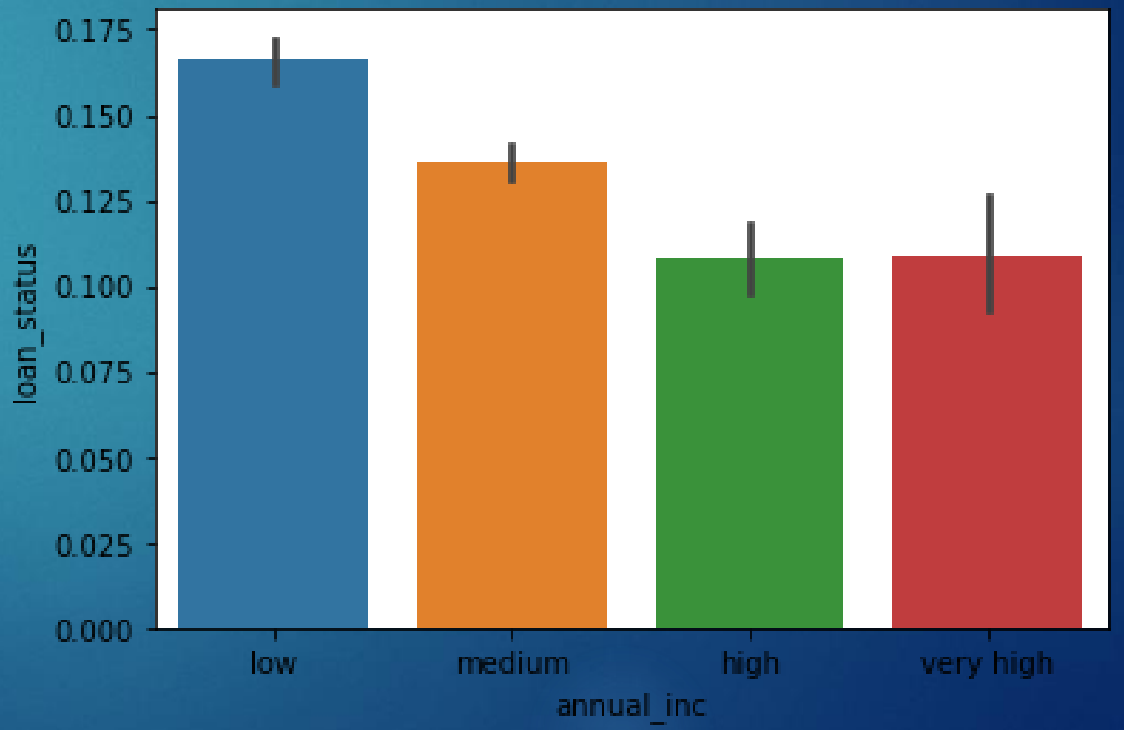
funded_amn



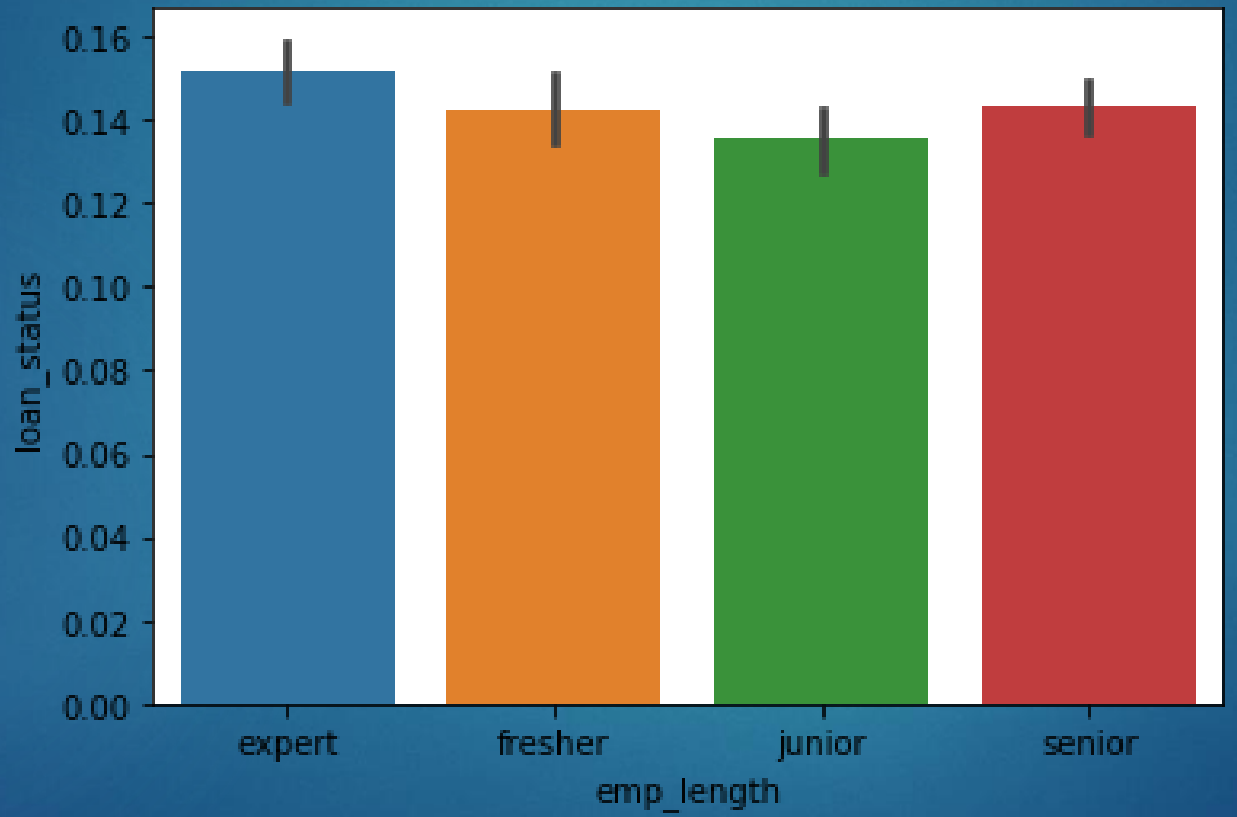
Comparing default rates across installment
The higher the installment amount, the higher the default rate



Annual income and default rate
Lower the annual income, higher the default rate

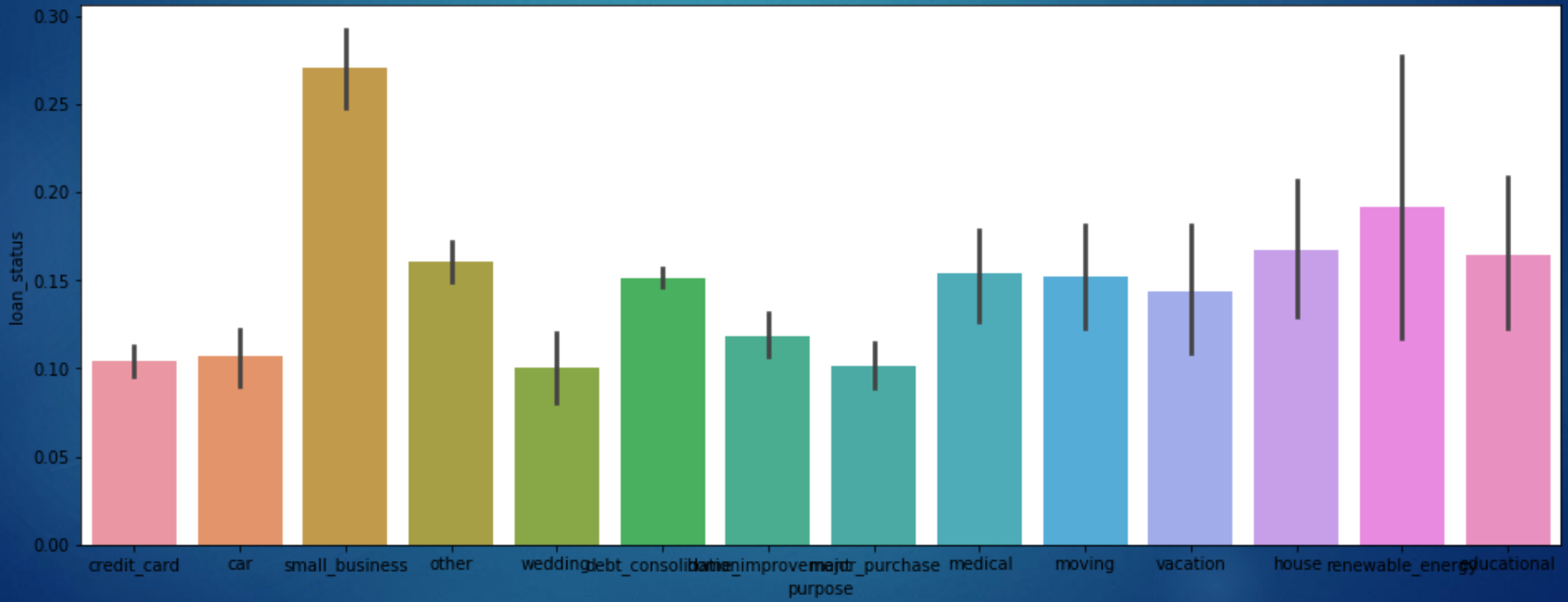


emp_length and default rate
Not much of a predictor of default

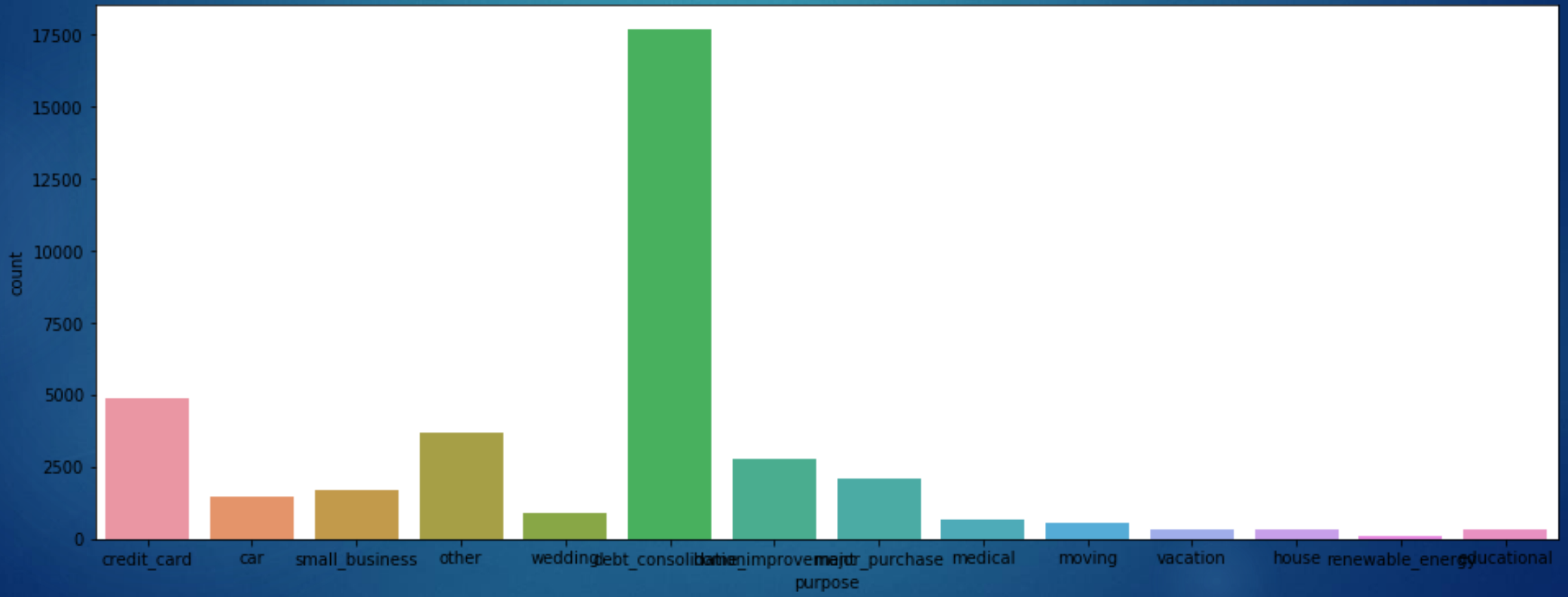


Segmented Univariate Analysis

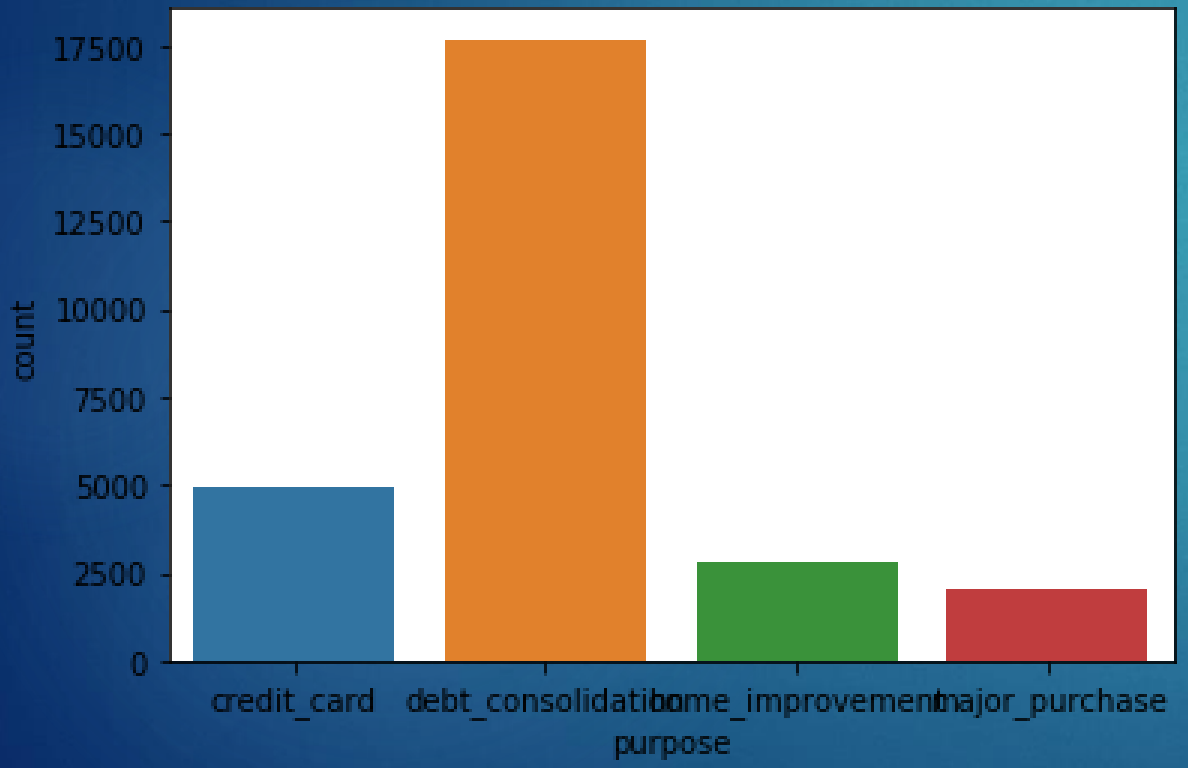
Small business loans default the most, then renewable energy and education



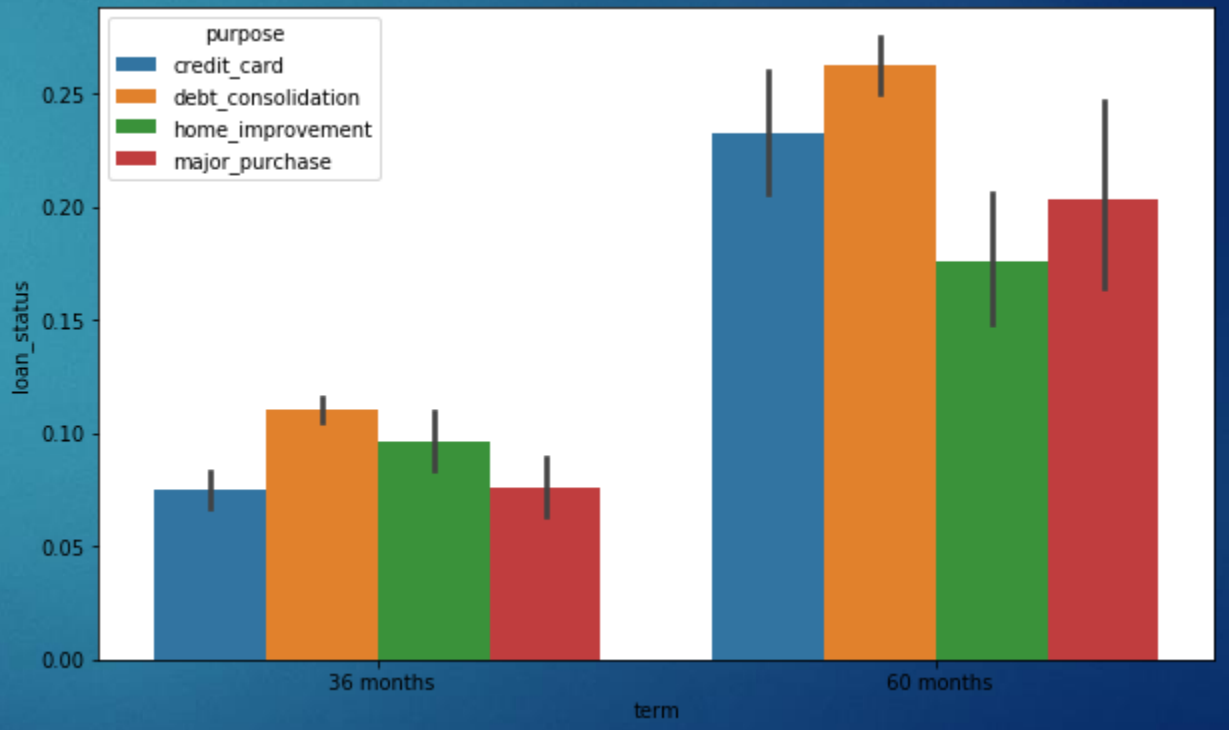
Lets first look at the number of loans for each type (purpose) of the loan
Most loans are debt consolidation (to repay other debts), then credit card, major purchase etc.



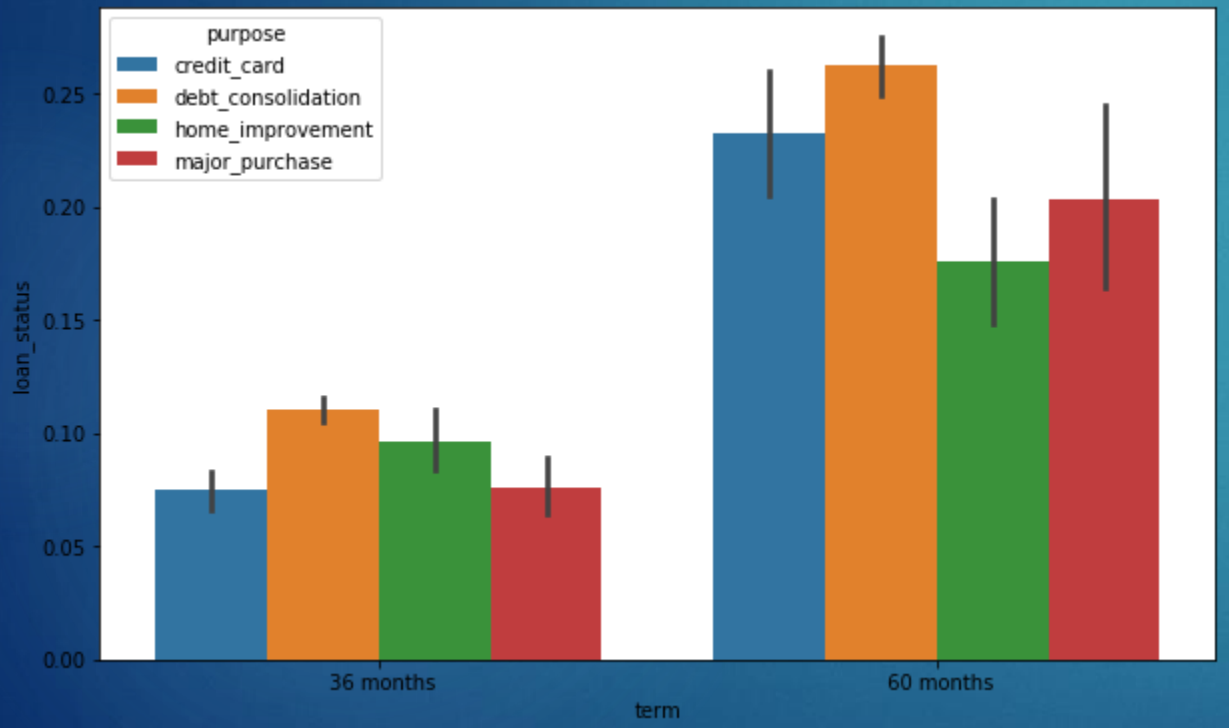
Plotting number of loans by purpose



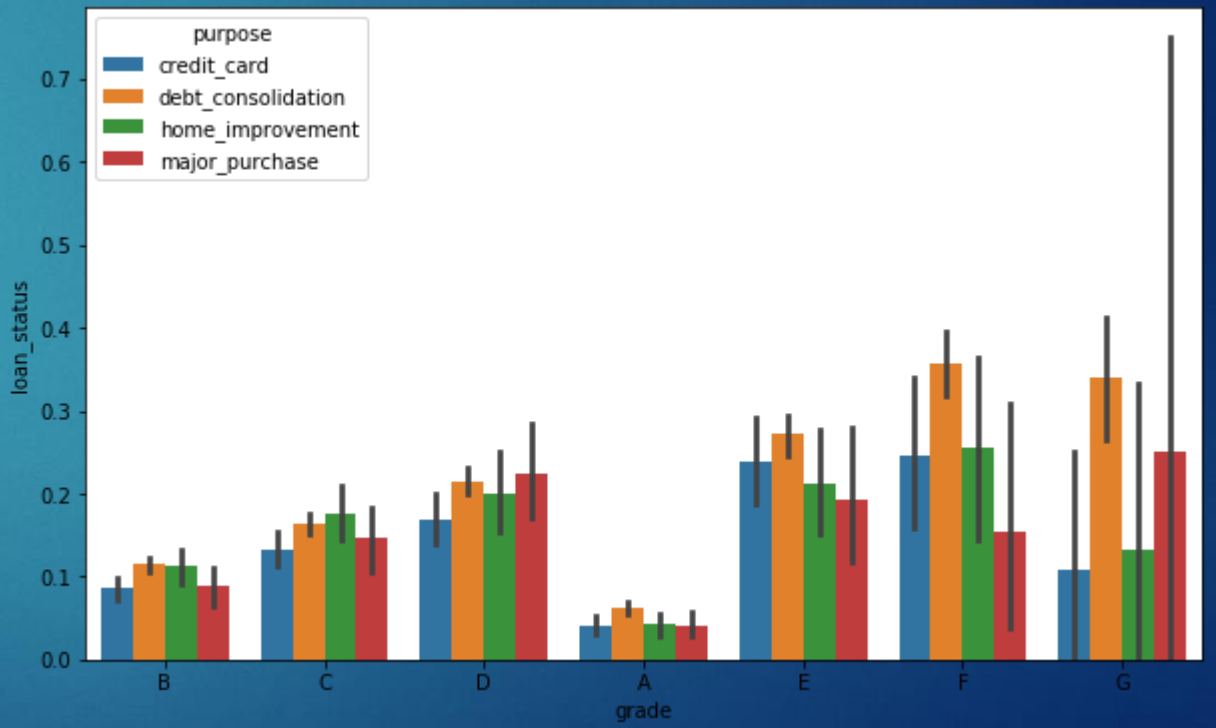
Compare the default rates across two types of categorical variables
Purpose of loan (constant) and another categorical variable (which changes)



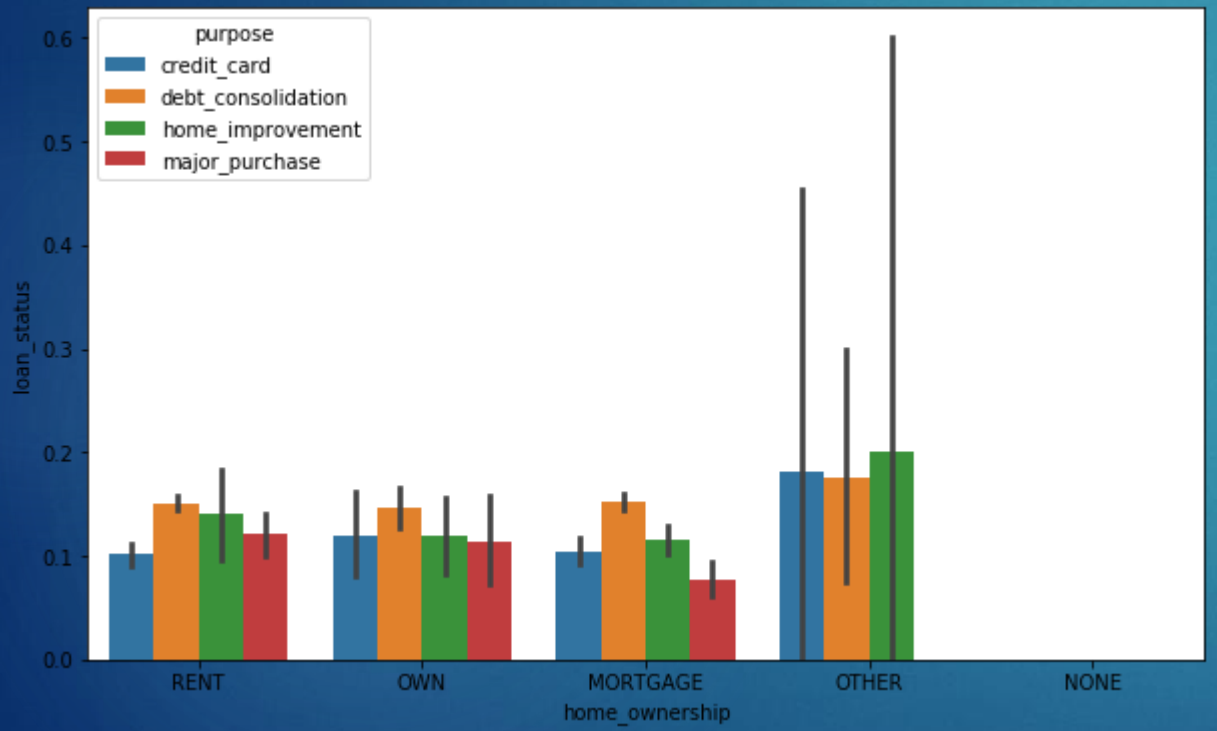
Takes a categorical variable and plots
the default rate
Segmented by purpose



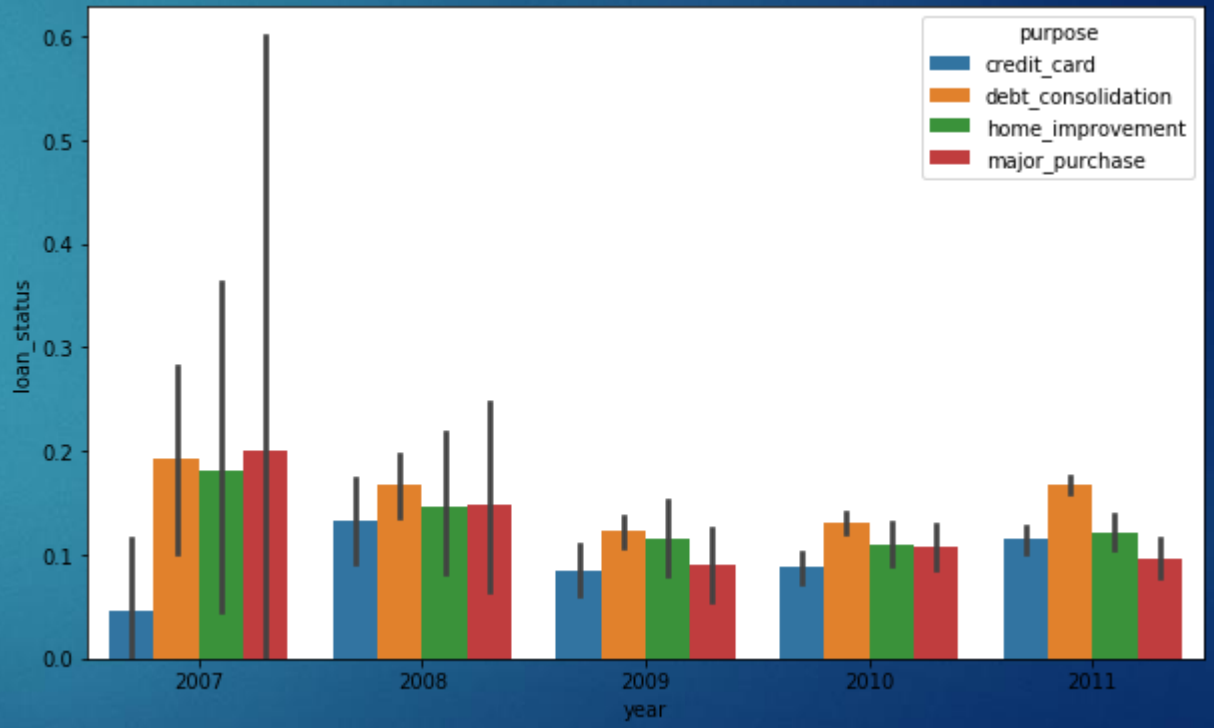
Grade of loan



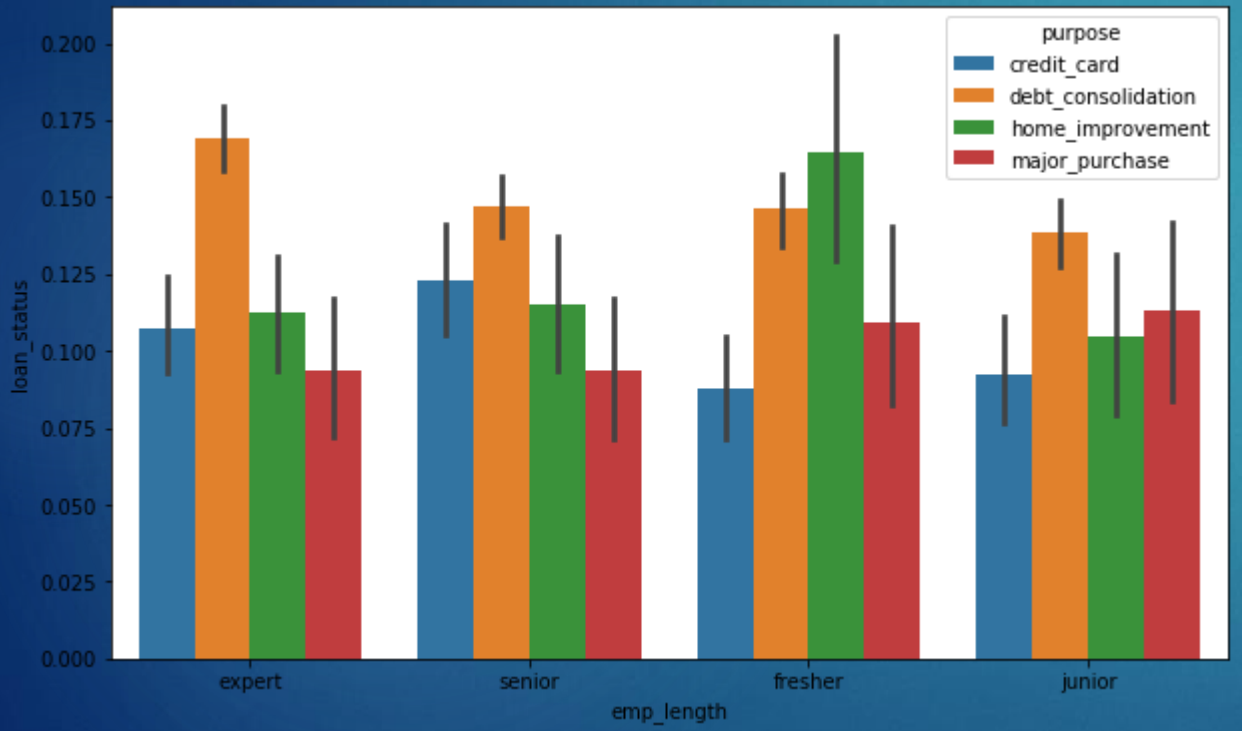
Home ownership



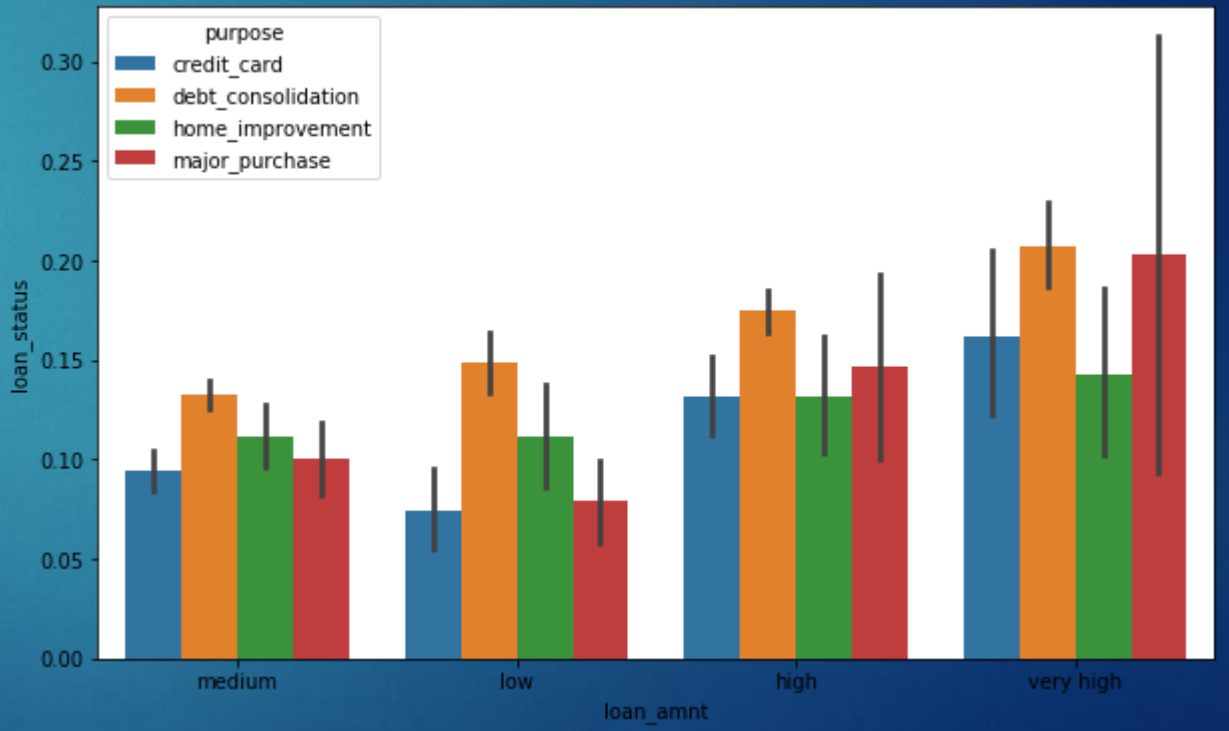
Year



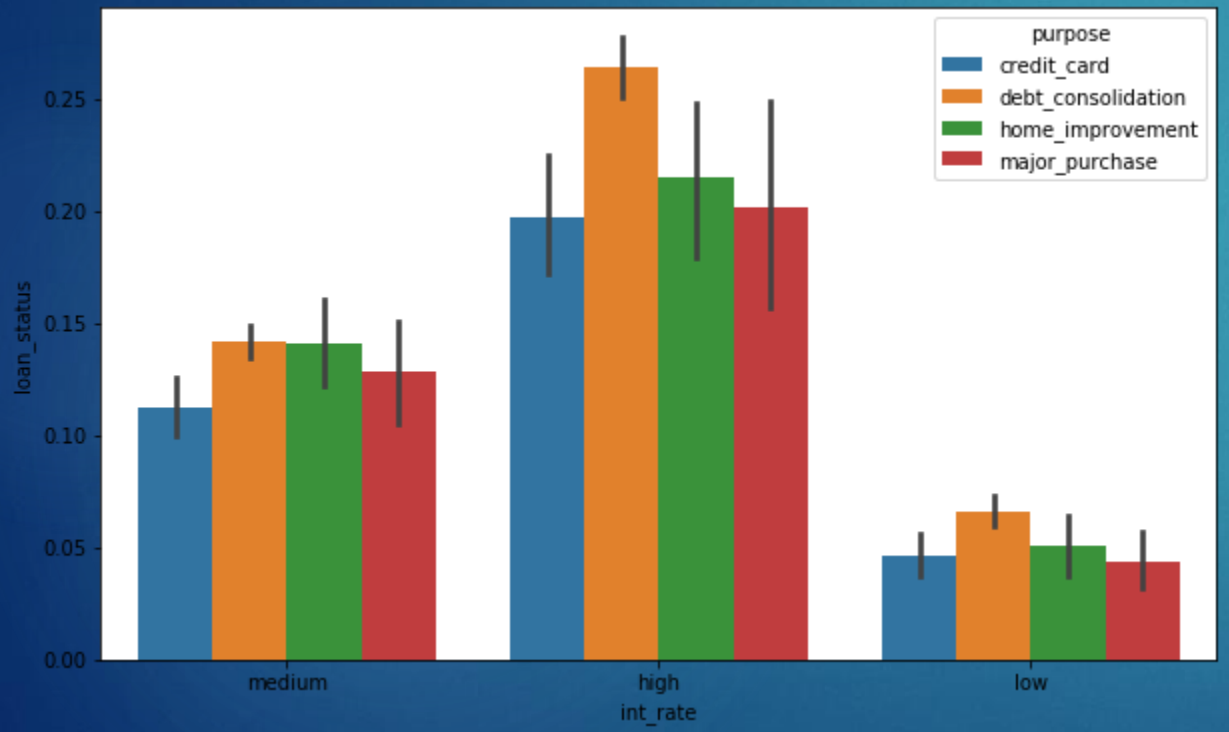
emp_length



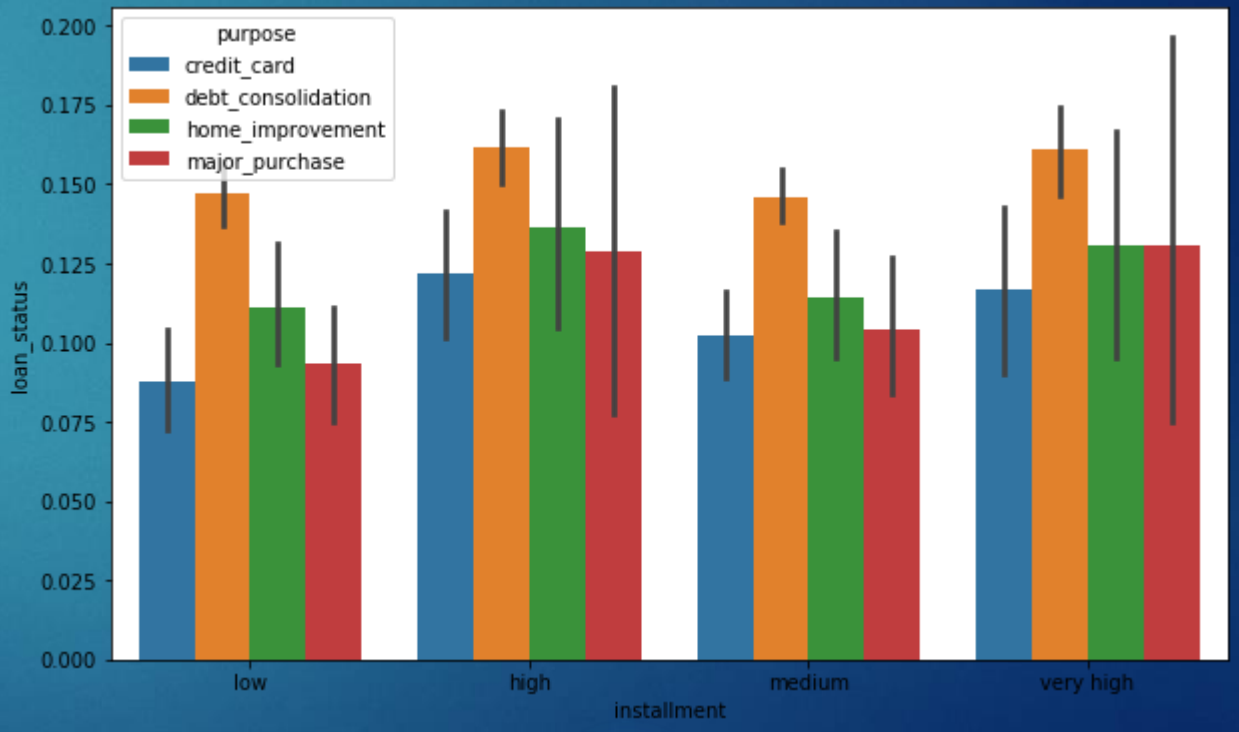
loan_amnt: same trend across loan purposes



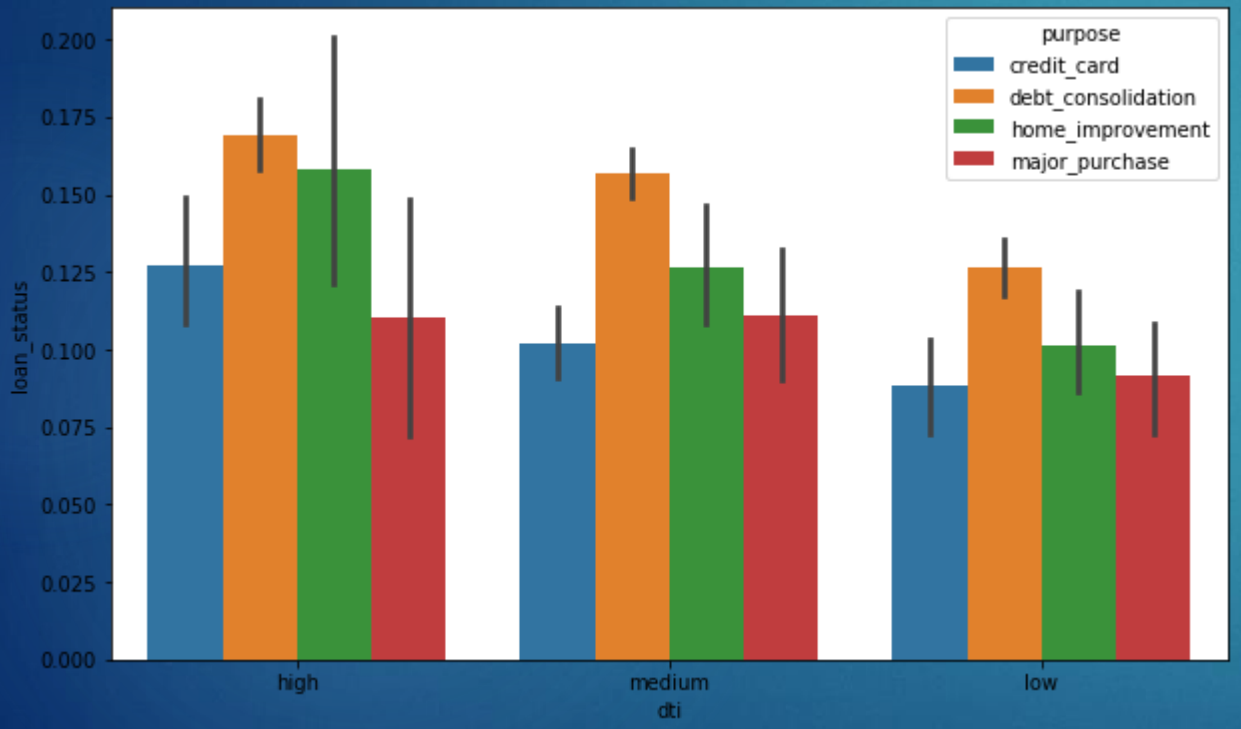
Interest rate



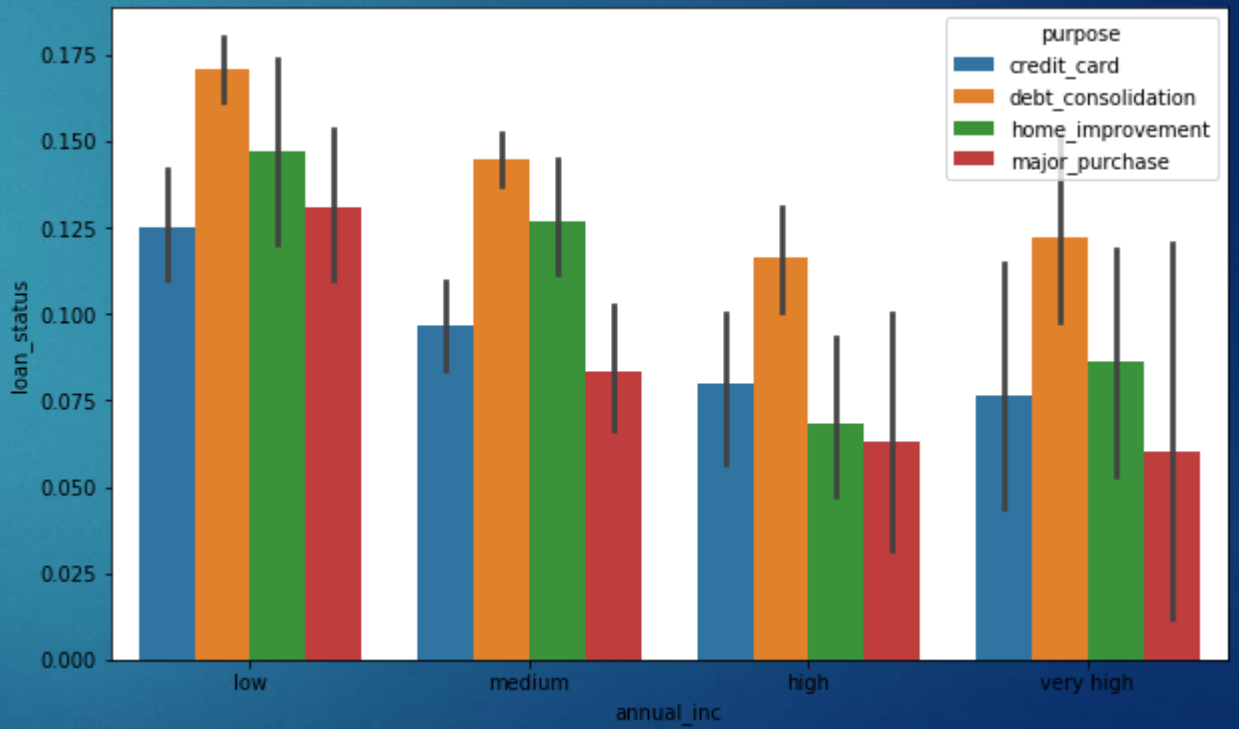
Installment



Debt to income ratio



Annual income



Recommendations

On the basis of our uni and segmented variate analysis we conclude that following factors are the variables which are strong indicators of default so, we must keep our eye on these factors

- loan_amount
- funded_amount
- annual_inc term
- funded_amount_inv
- inq_last_6mths
- open_acc
- verification_status
- emp_length
- Installment
- int_rate
- home_ownership
- Dti
- Purpose
- Grade
- revol_util

Thank You