# Lending Club Case Study

Aloke Kumar Mukherjee (alokkmukherji@gmail.com)
Akshay Kachroo (akshaykachroo2050@gmail.com)

# Background

A lending company facilitates loans for different pupose like personal loans, business loans, medical procedures. Like most other lending companies, lending loans to ‘risky’ applicants is the largest source of financial loss / credit loss. 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'. The number of defaulters can be reduced by identifying risky loan applicants at the time of loan approval.

# Objective

The goal of this project is to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default so that this information can be utilised for risk assessment during loan approval.

# Data Analysis Flow

Data Loading Understanding -> Data Cleaning -> Univariate Analysis -> Segmented Univariate Analysis -> Heatmap Correlation

# Data Understanding

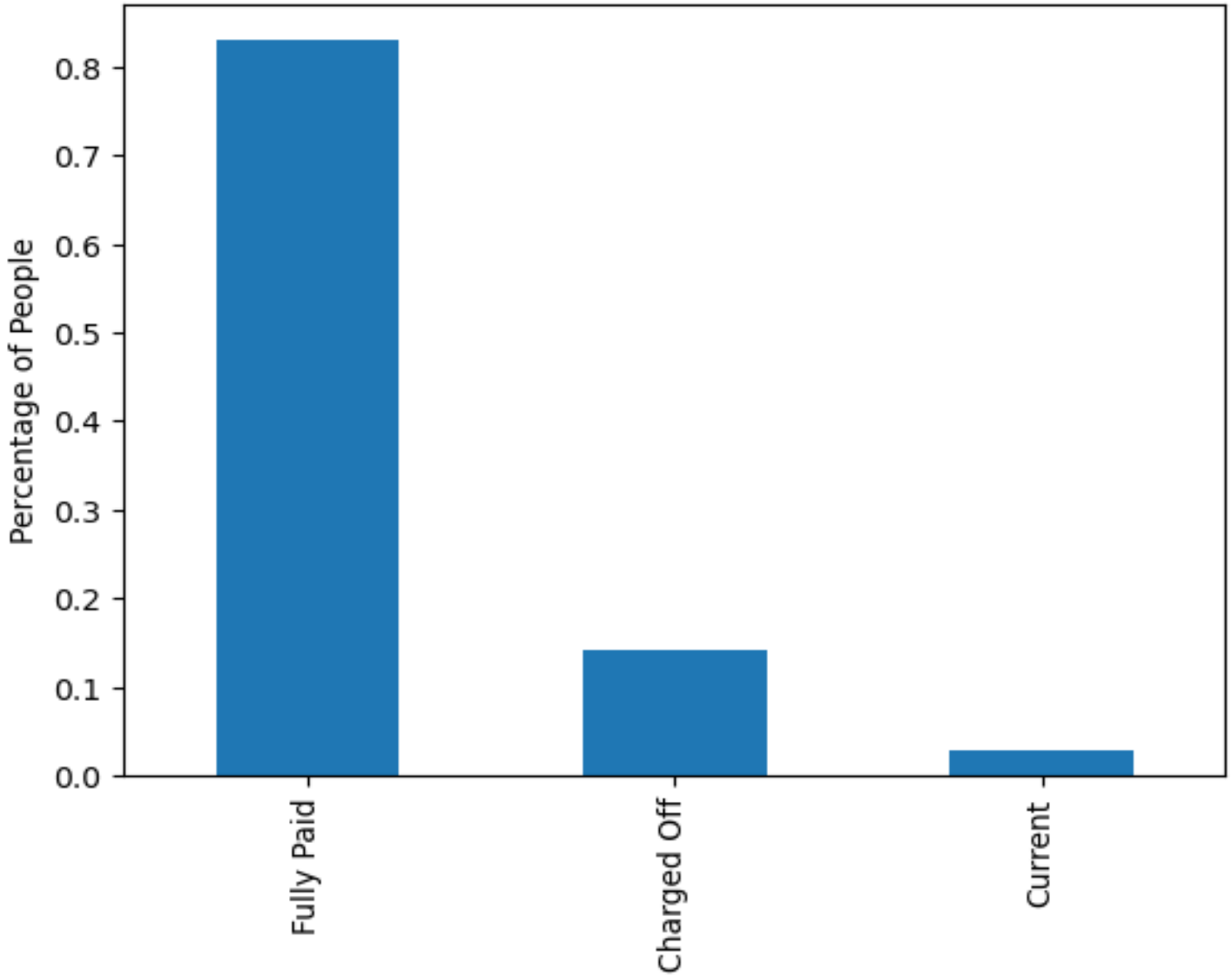
- No. of rows : 39717, No. of columns : 111 present in i/p data set
- No. of columns having no data :  54
- The data present in i/p dataset can be classified in three categories:
  - Customer demographic information (Annual income, No. of employment years etc.)
  - Loan related information (Loan amount, Interest rate, Loan status, Loan grade etc.)
  - Customer behaviour information (Purpose of loan, application type etc.)
- The target variable for the analysis is loan_status which has 3 values: “Charged off”, “Fully Paid”, “Current”

# Data Cleaning

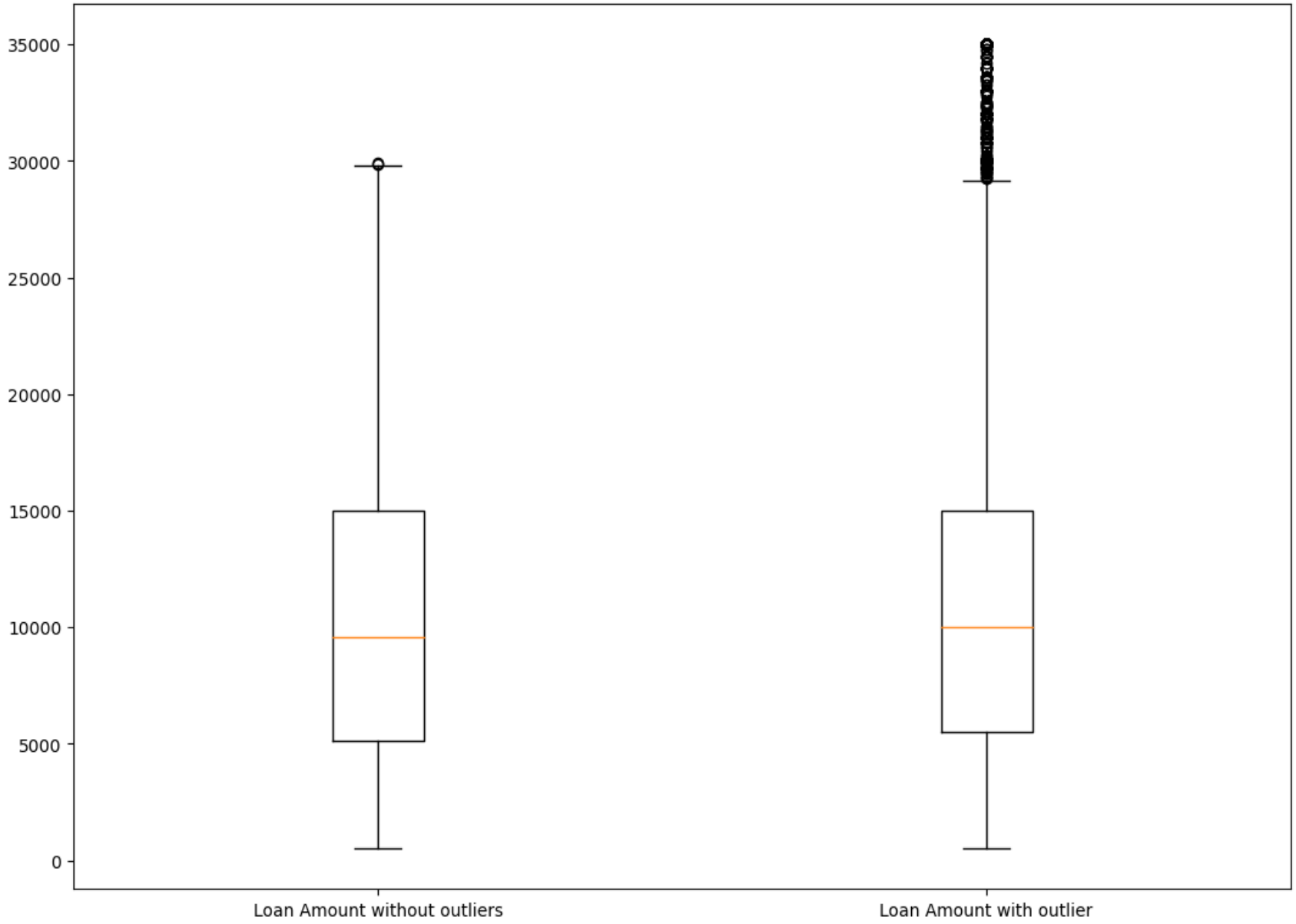
1. Dropping columns having no value
2. Checking for duplicate data
3. Reformatting column value (by changing data type , removing unwanted patterns etc.) suitabe for data analysis
4. Removing outliers
5. Removing columns having same value

The no. of rows and columns reduced to 37398 and 21 respectively, This reduced dataset was used for further analysis

# Data Analysis - Loan Status

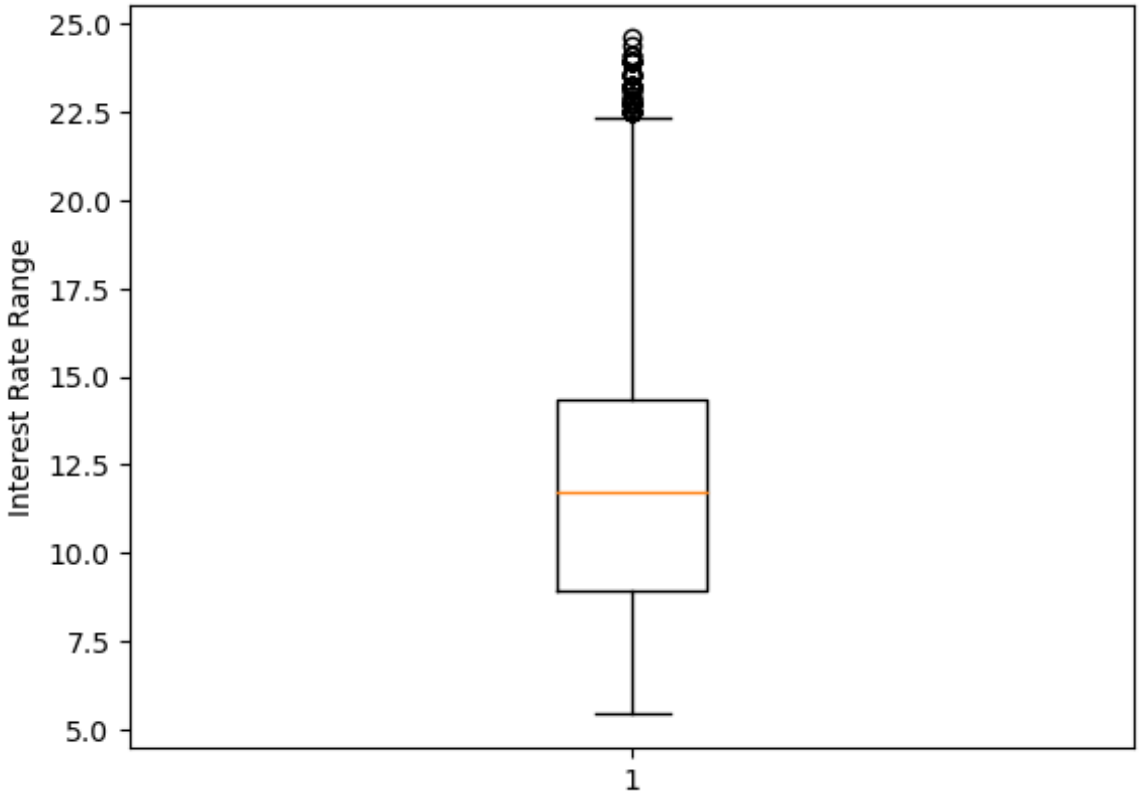


# Data Analysis - Loan Amount

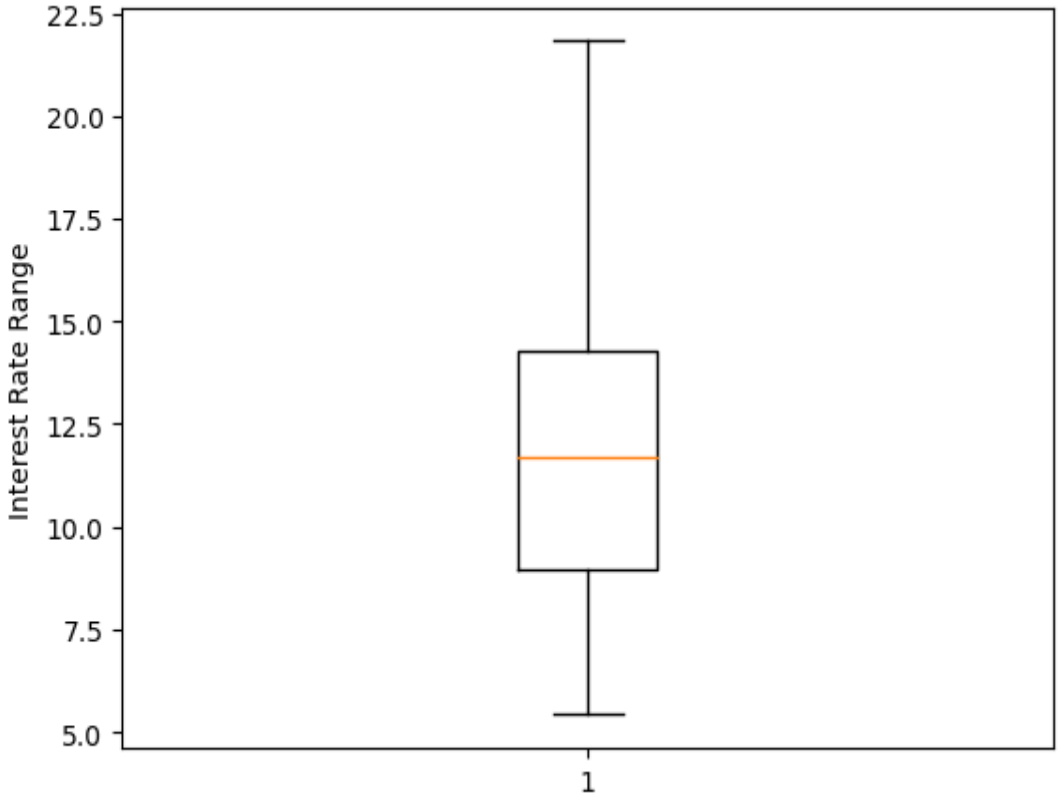


# Data Analysis - Interest Rate

With outlier

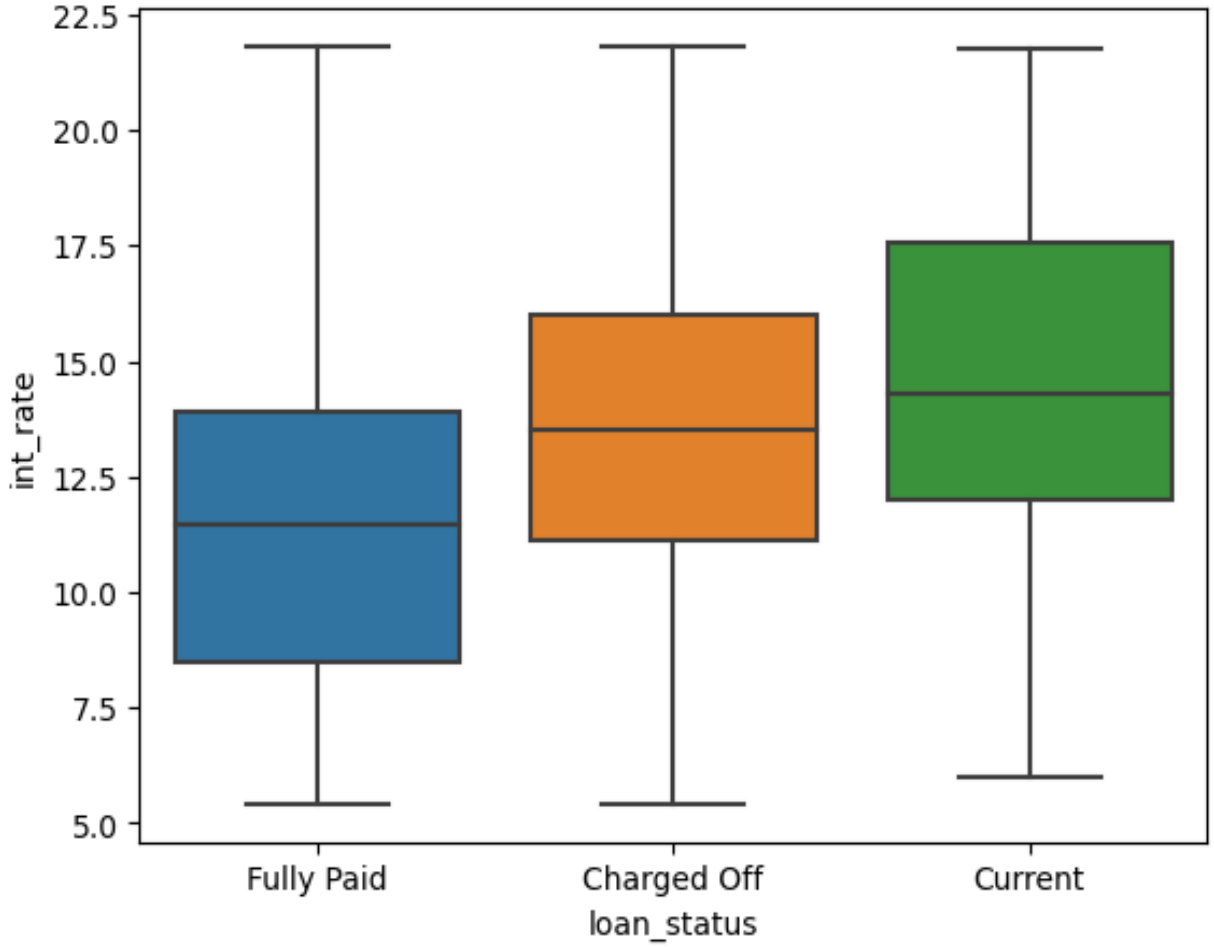



after removing outlier



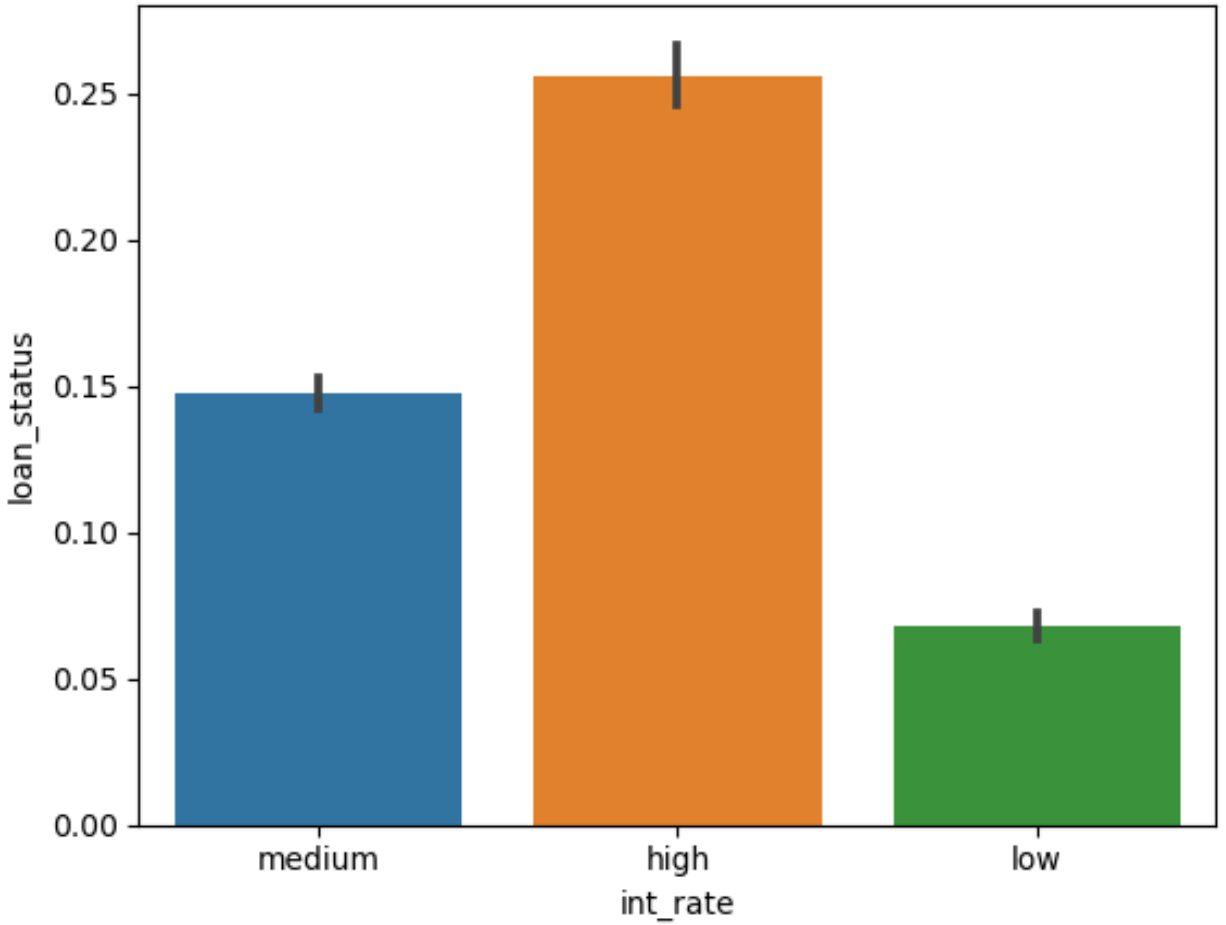
# Data Analysis - Interest Rate & Loan Status

The more chance of default if interest rate is high .. Current means the cases where loan repayment is happening , so it is not under analysis



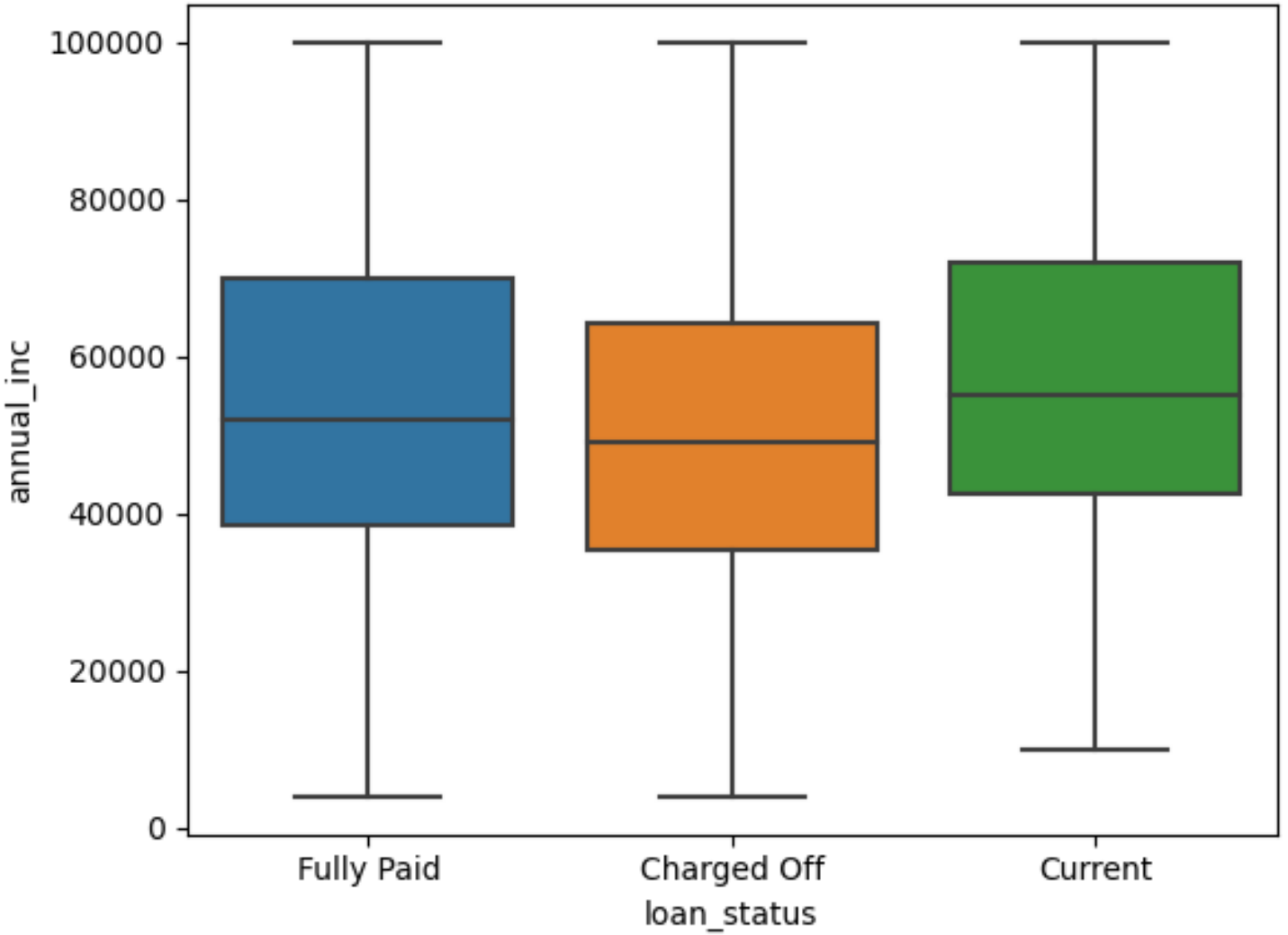
# Data Analysis - Interest Rate & Loan Status (Continued)..

Interest rate is categorized .. More chance of default if interest rate is very high



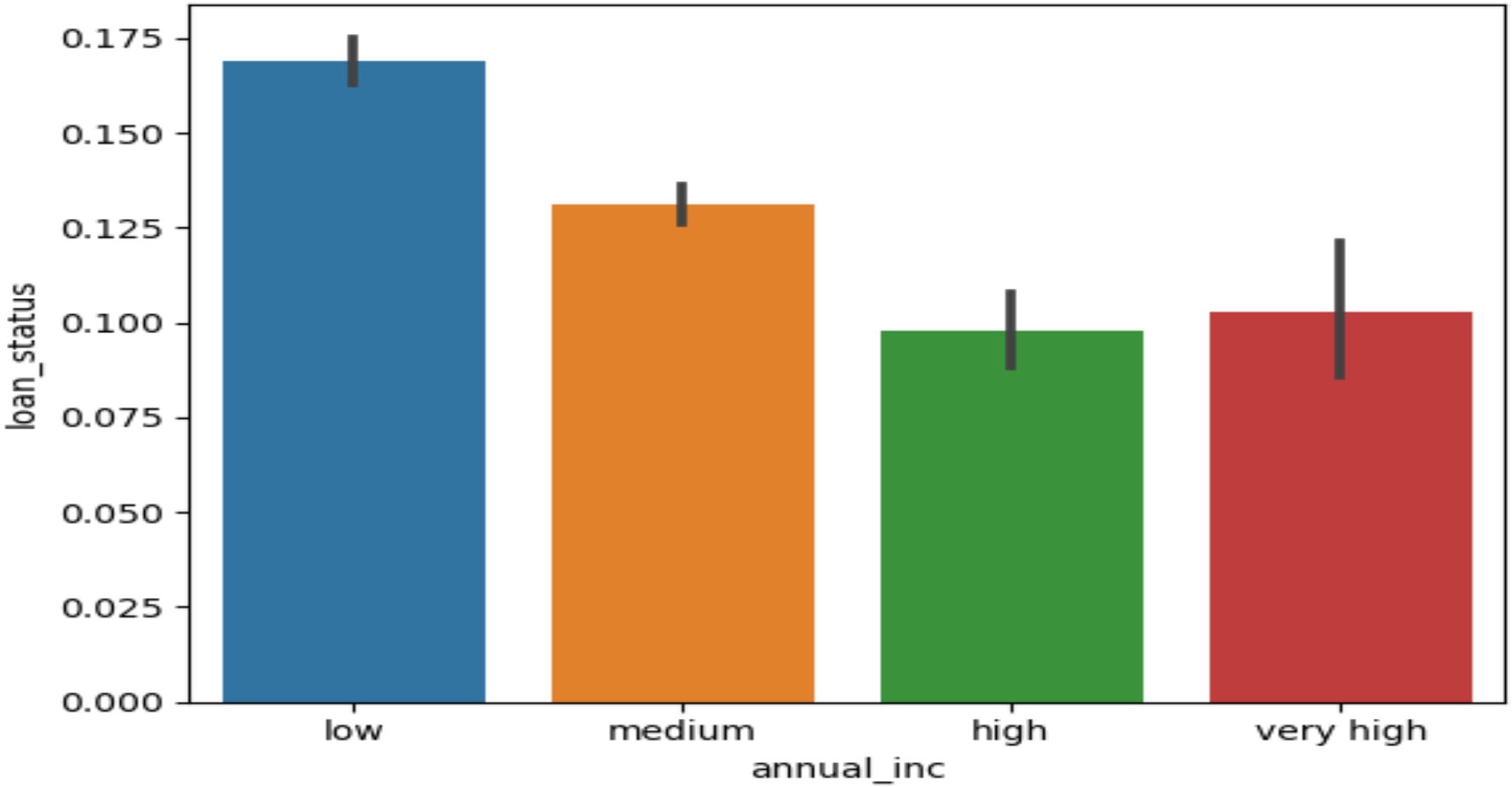
# Data Analysis - Annual Income & Loan Status

The more chance of default if annual income is low .. Current means the cases where loan repayment is happening , so it is not under analysis



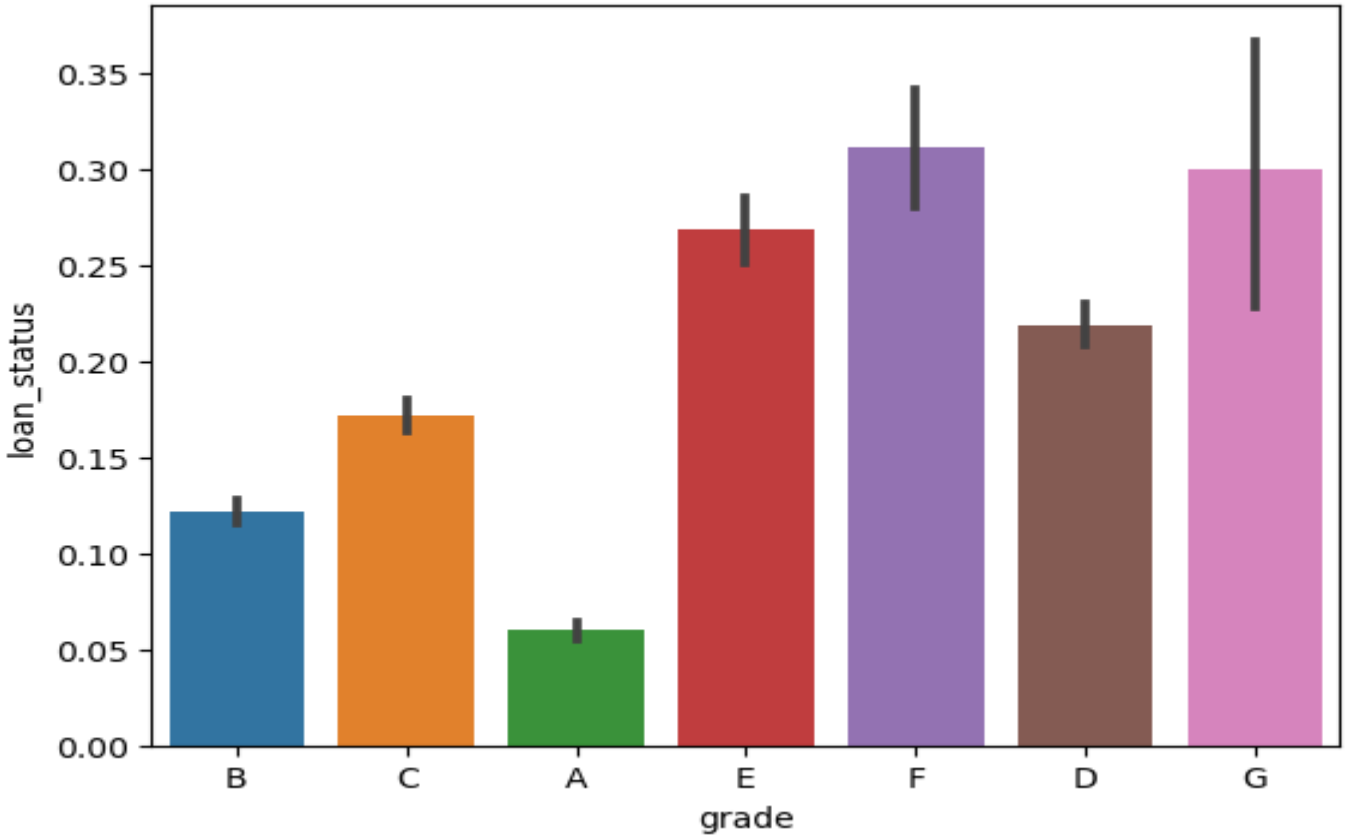
# Data Analysis - Annual Income & Loan Status (Continued) ..

Annual income is categorized .. More chance of default for low income category



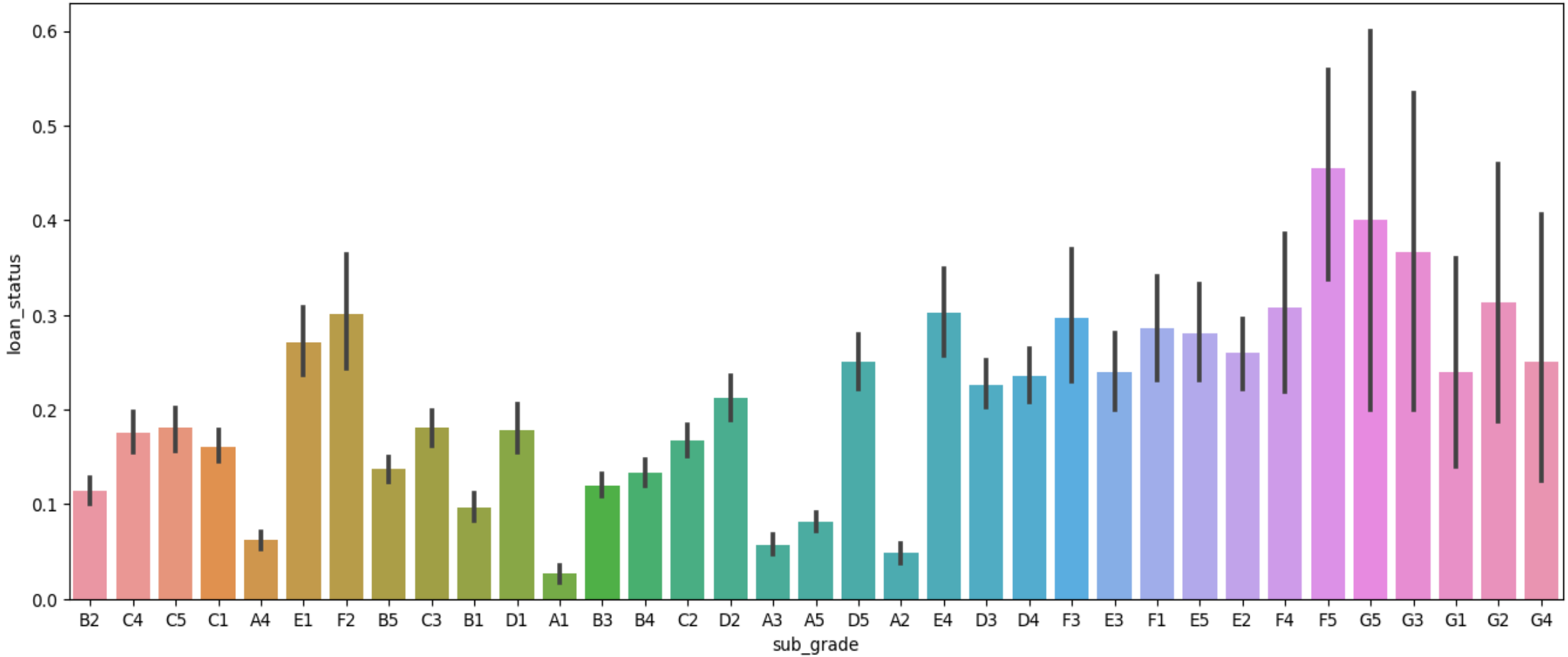
# Data Analysis - Loan Status & Loan Grade

The more chance of default for Grade F



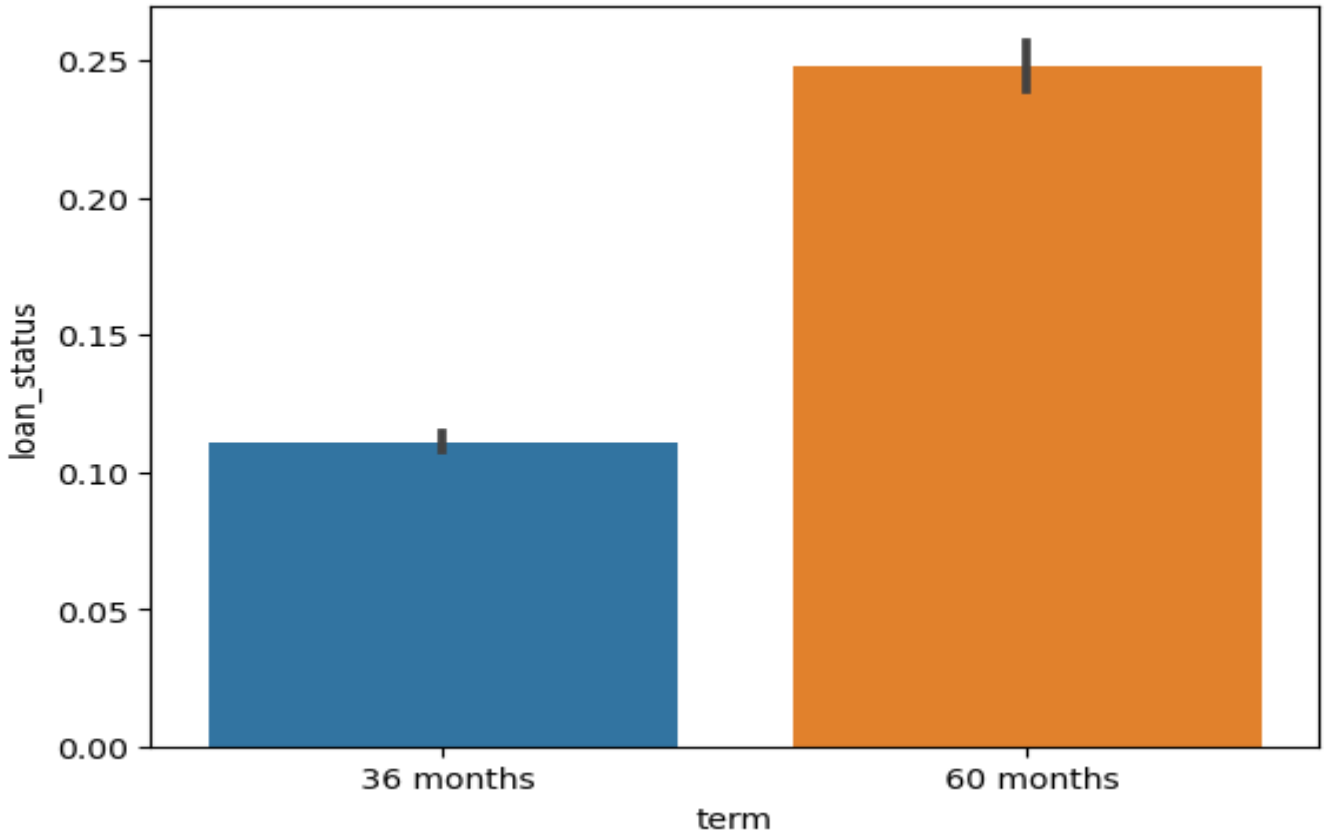
# Data Analysis - Loan Status & Loan Sub Grade

The more chance of default for Sub Grade F5



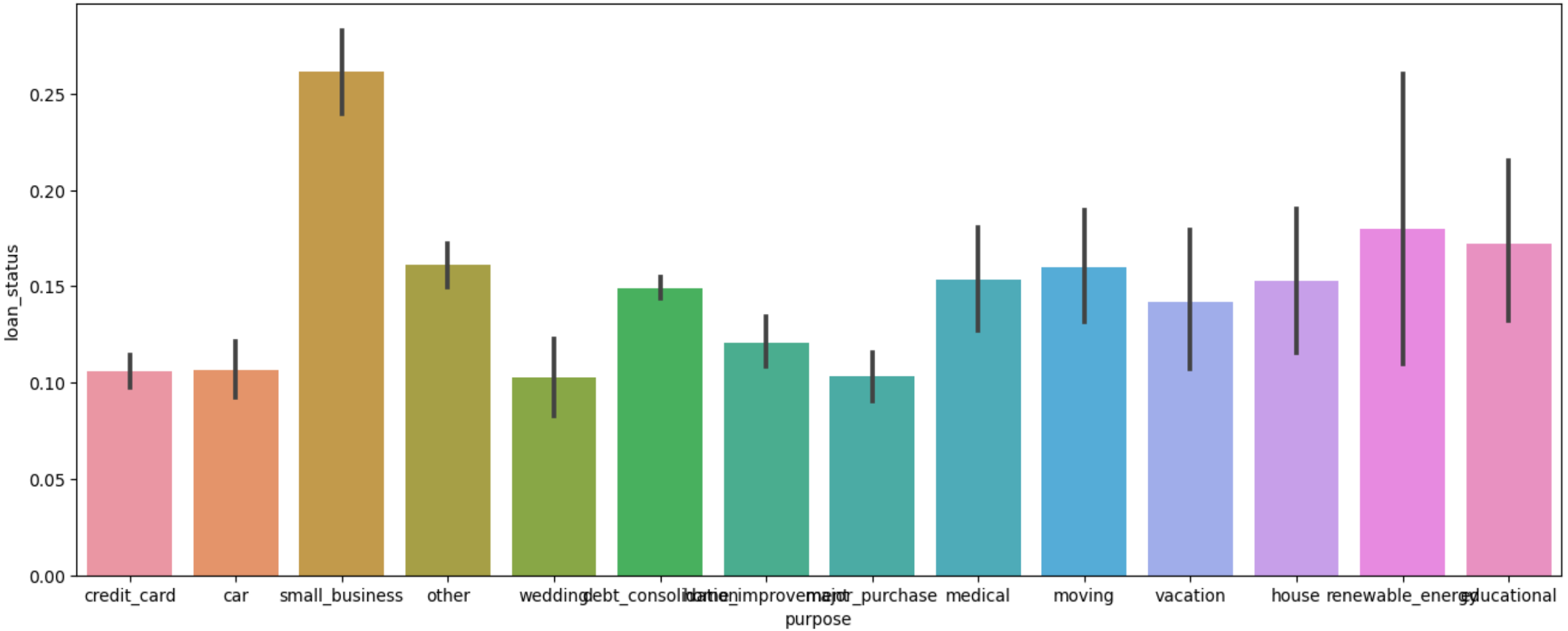
# Data Analysis - Loan Status & Loan Term

The more chance of default if loan term is 60



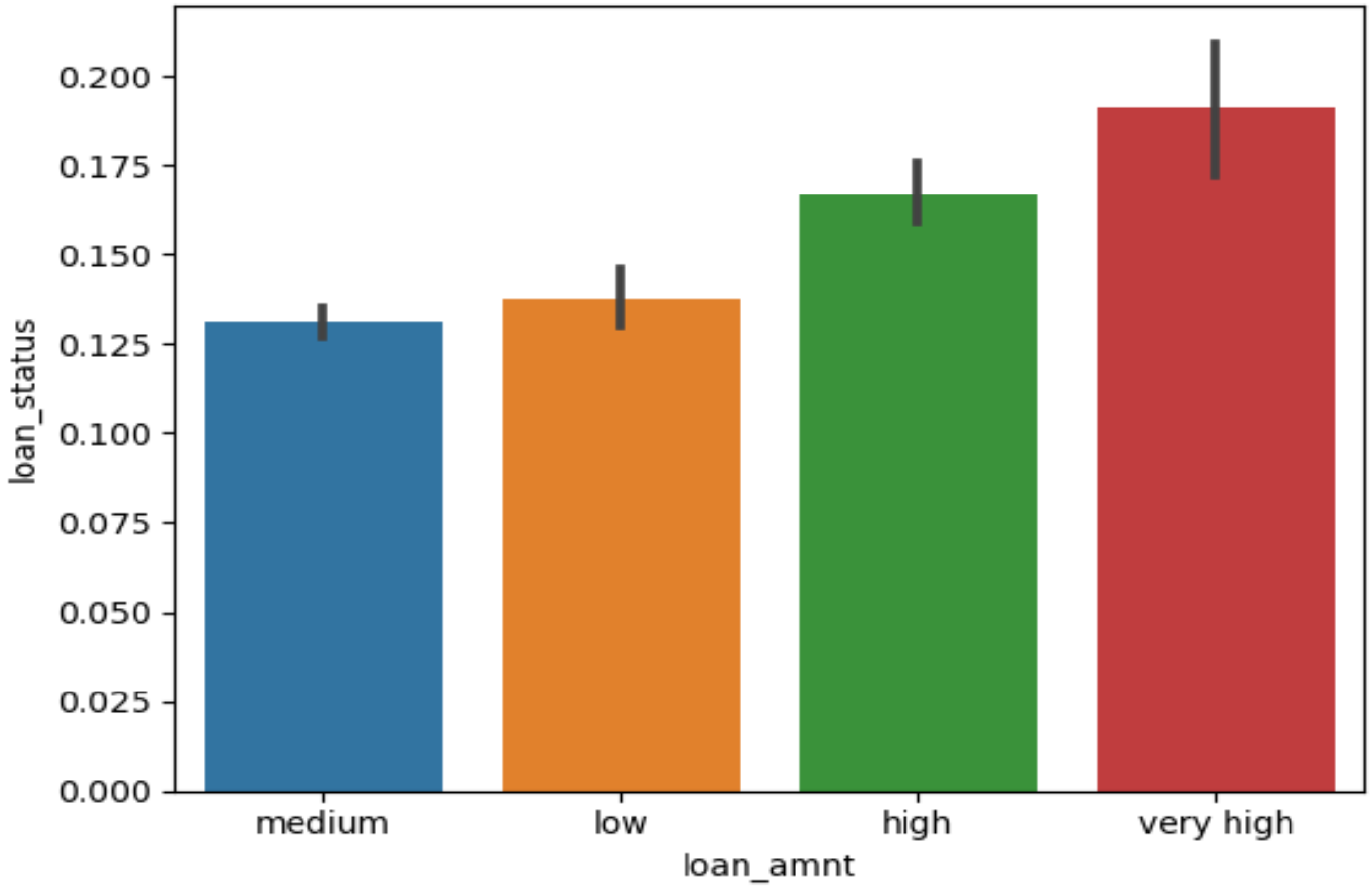
# Data Analysis - Loan Status & Purpose of Loan

The more chance of default if loan is taken for small business purpose



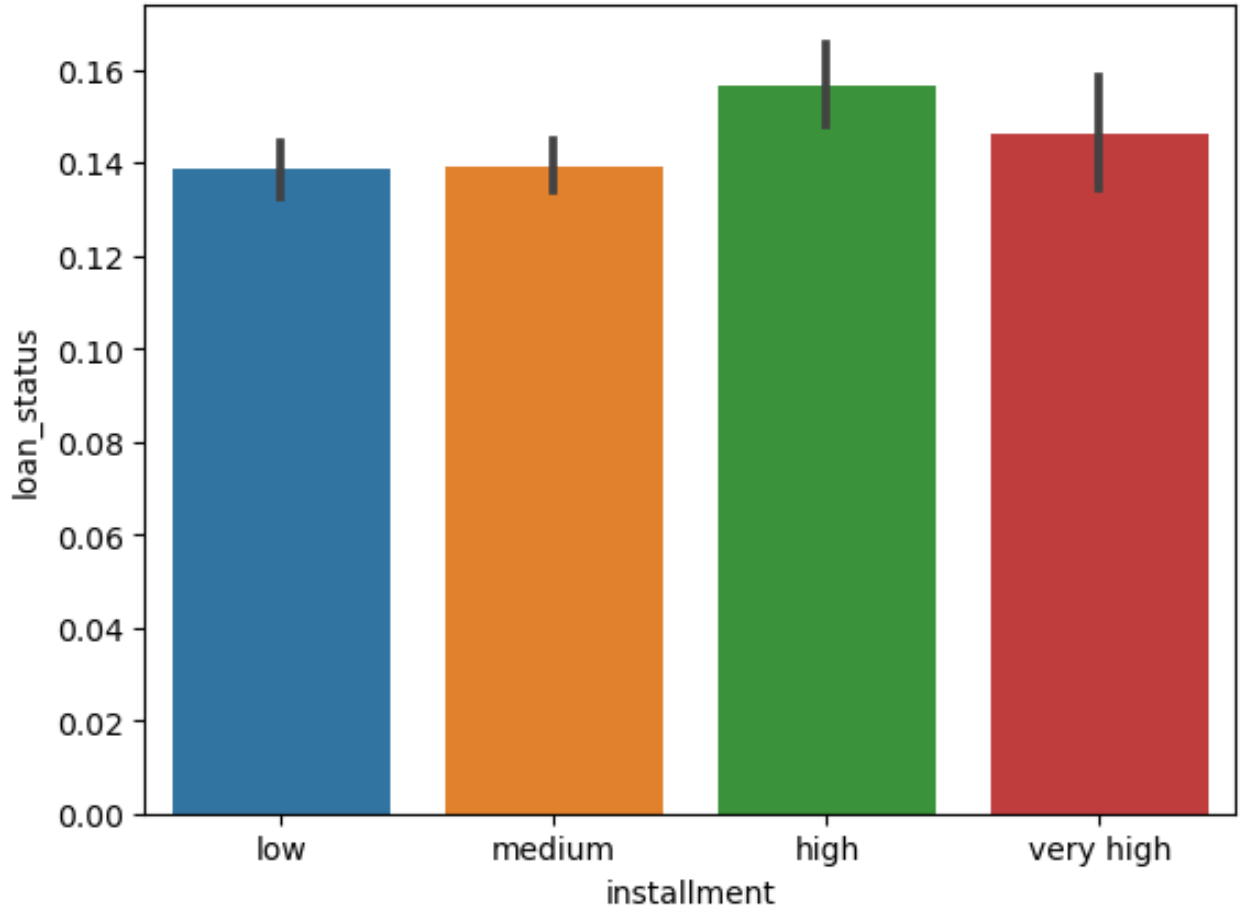
# Data Analysis - Loan Status & Loan Amount

Loan amount is categorized .. More chance of default if loan amount is very high



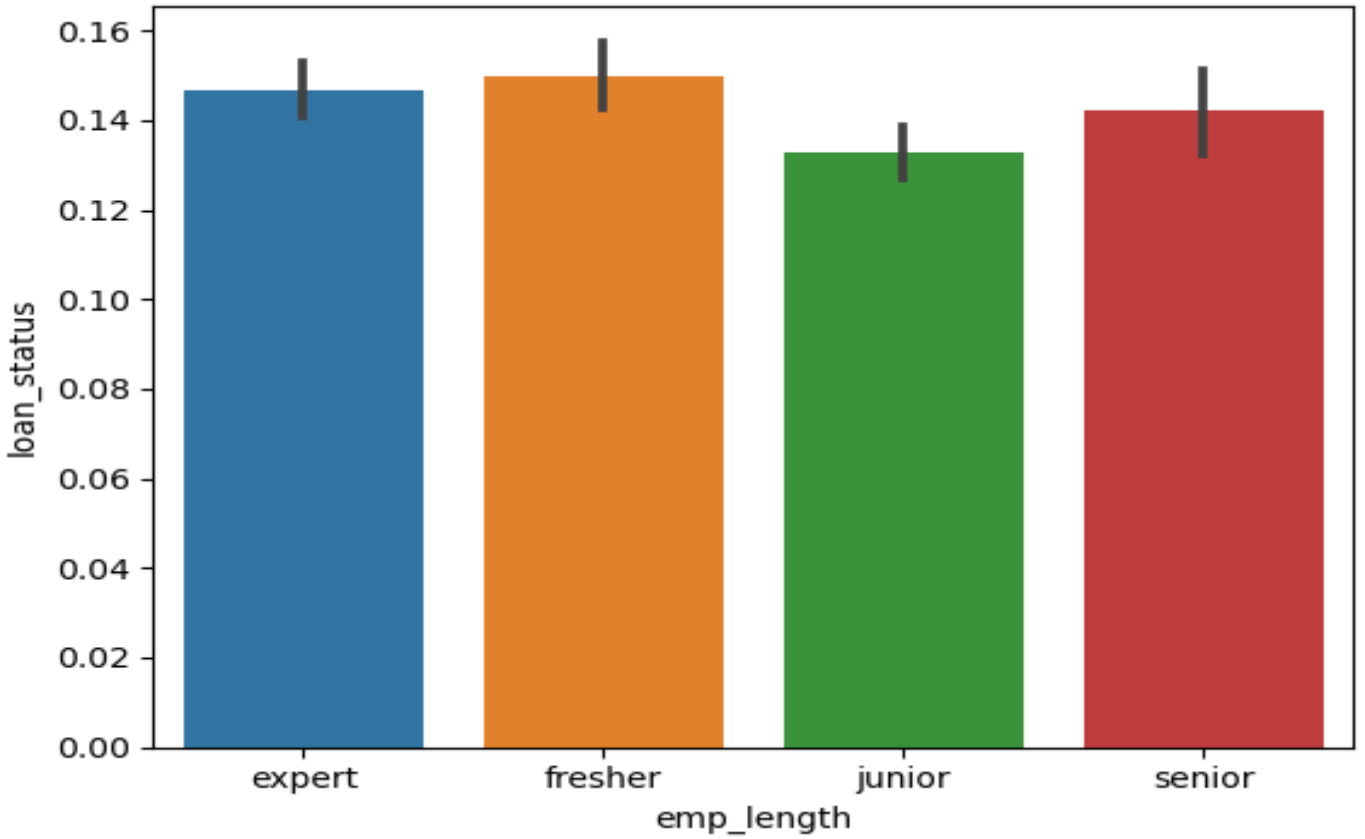
# Data Analysis - Loan Status & Installment

Installment is categorized .. More chance of default if no. of installment is high



# Data Analysis - Loan Status & Employment Length

Employment length is categorized .. More chance of default for the case of fresher



# Heatmap Co-relation Analysis for default category