EDA CASE STUDY

Loan Default Rate

-Ankit Desale

- # Objective :

This company is the largest online loan marketplace facilitating personal loans, business loans, and the financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Like most other lending companies, lending to ‘risky’ applicants is the largest source of financial loss (called credit loss). Credit loss is the amount of money lost by the lender when the borrower refuses to pay or absconds with the money owed. In other words, borrowers who default cause the biggest losses to lenders. In this case, customers labelled as ‘charged-off’ are the ‘defaulters’.

If one is able 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 utilise this knowledge for its portfolio and risk assessment.

- **Purpose:** The aim is to identify patterns indicating that a person is likely to default, which may be used to deny the loan, reduce the loan amount, lend (to risky applicants) at a higher interest rate, etc.

Methodology :

- Data Cleaning :- The Dataset have 39717 rows and 111 columns
  - *Before processing next are certain column which have missing values 100%*

    null values such as 'num_actv_bc_tl','num_bc_tl','num_sats' as they are of no use we need to drop them
  - *Few columns have missing values around 32%,64%,92%,97% missing values name as:*
    'desc','mths_since_last_delinq','mths_since_last_record','next_pymnt_d'

    by dropping these will not affect much on our calculation
  - *Before processing next are certain column which have missing values 100%*

    null values such as 'num_actv_bc_tl','num_bc_tl','num_sats' as they are of no use we need to drop them
  - *There several behaviour variable those which are generated after the loan is approved such as delinques 2 years ,resolving balance which are of no use in*

    our analysis so dropping them will not affect much our calculation

- **Standardizing Values :-**
  - *There is target variable as loan_status. We need to convert the values to a binary form - 0 or 1, 1 indicating that the person has defaulted and 0 otherwise.*
  - *There are some categorical variables which have object we convert them object to categorical type for our calculation.*
  - *There is column issue_d which is store in datatype string we need to convert it into date time format for our calculation.*

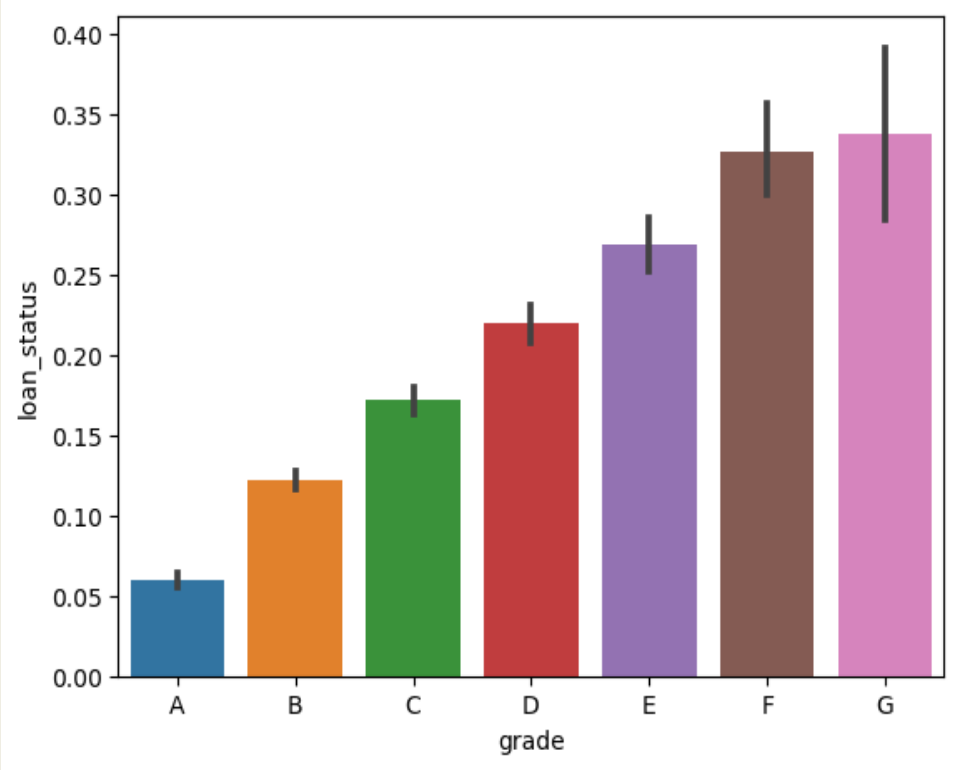
After cleaning the dataset contains 38577 rows and 22 which is much clear now

Data Analysis : The overall default rate is found 0.14

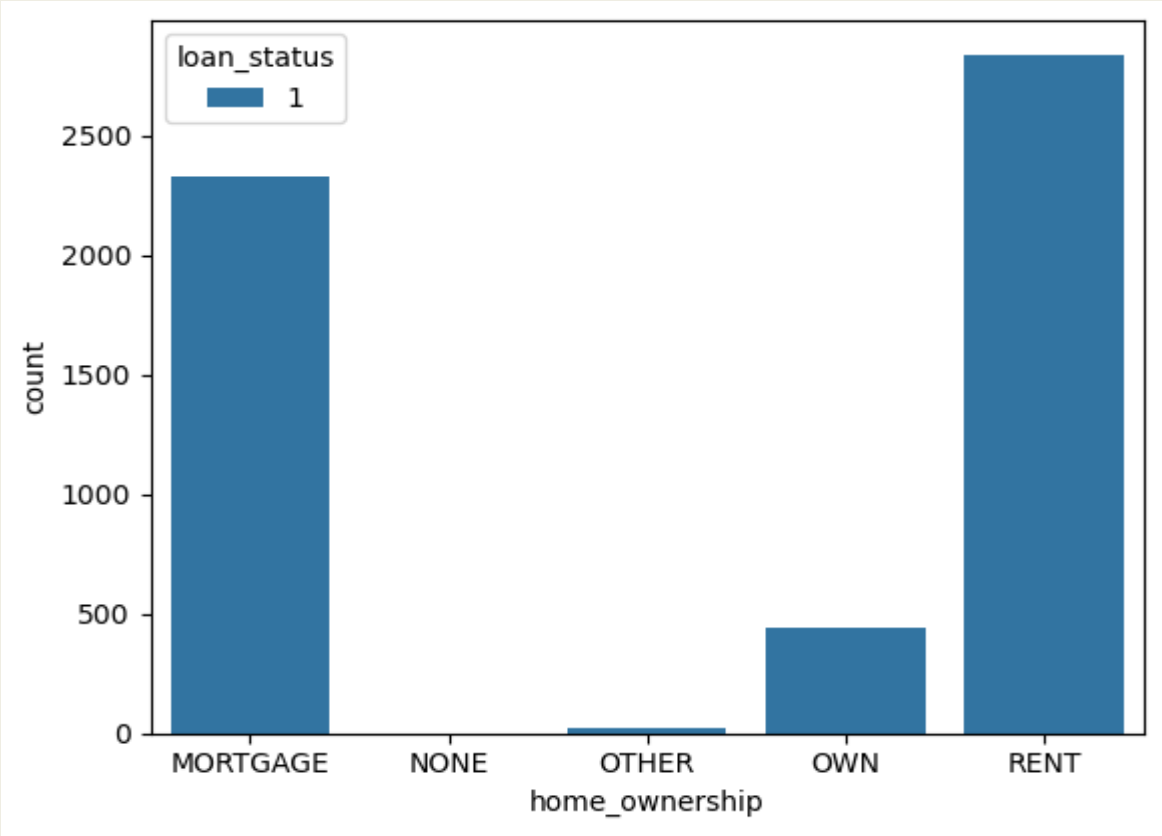
- Univariate Analysis:- As our target variable is loan status let's analyze it with other variable

  a]Grade vs Loan_status : As the grade of loan goes from A to G, the default rate increases.

  this is expected because the grade is decided by the riskiness of the loan.

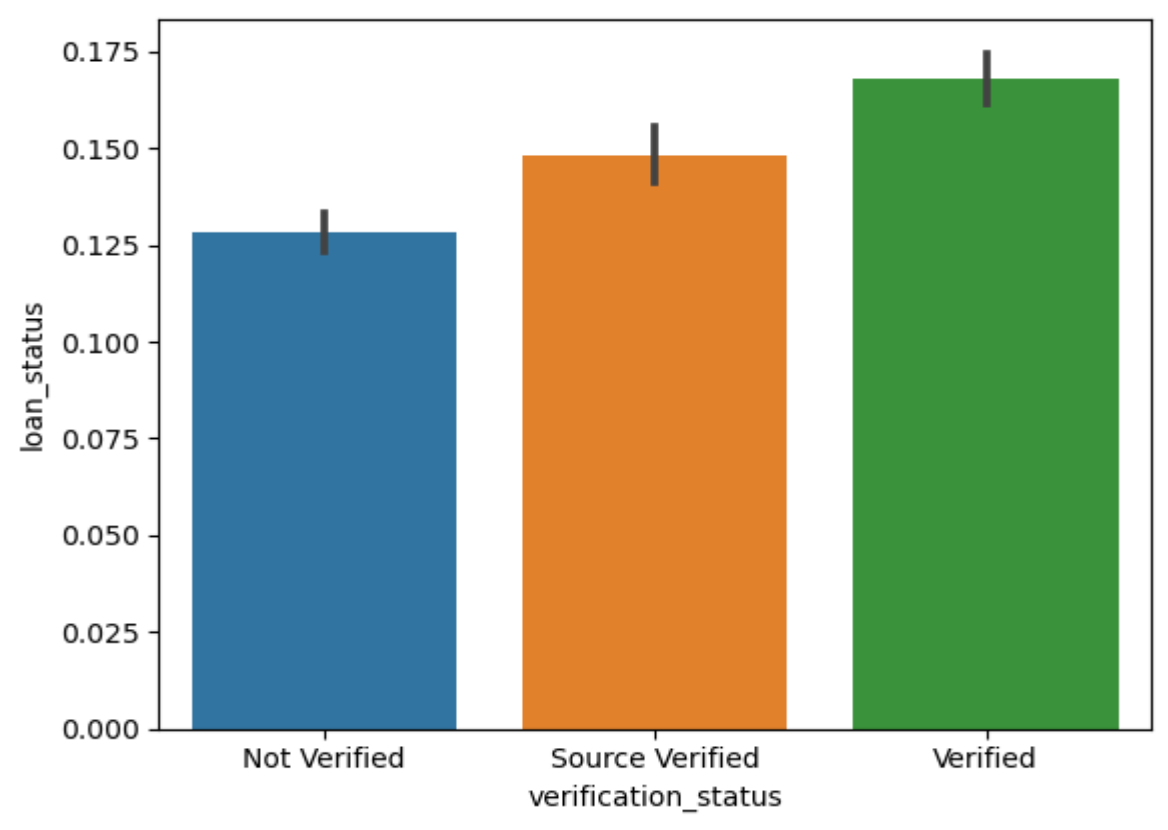


b] Home ownership vs loan Status : Rent and Mortage have higher default rate than who has permanent home ownership



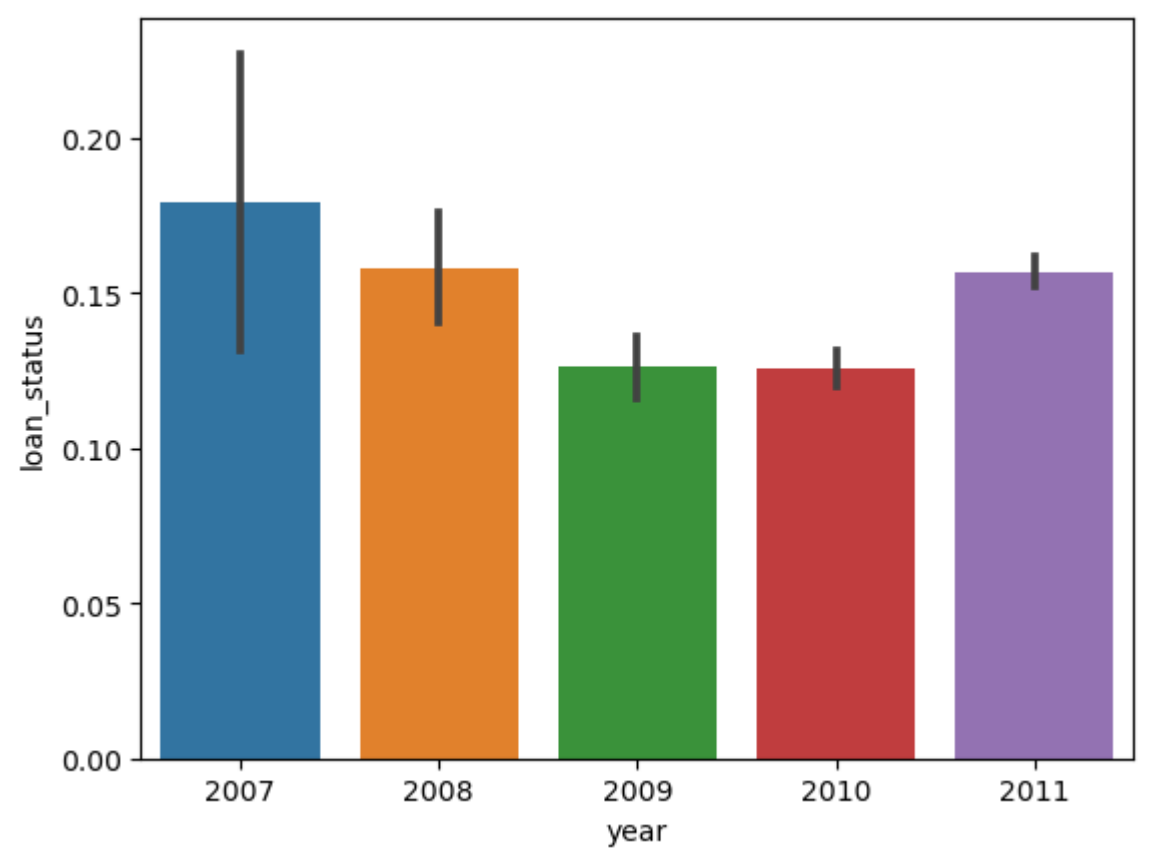
c]Verification status vs loan status:

In this analysis we found that verified have higher default rate than non verified



d] Year vs Loan Status:-

From our analysis it is found that the default rate had suddenly increased in 2011, inspite of reducing from 2008 till 2010



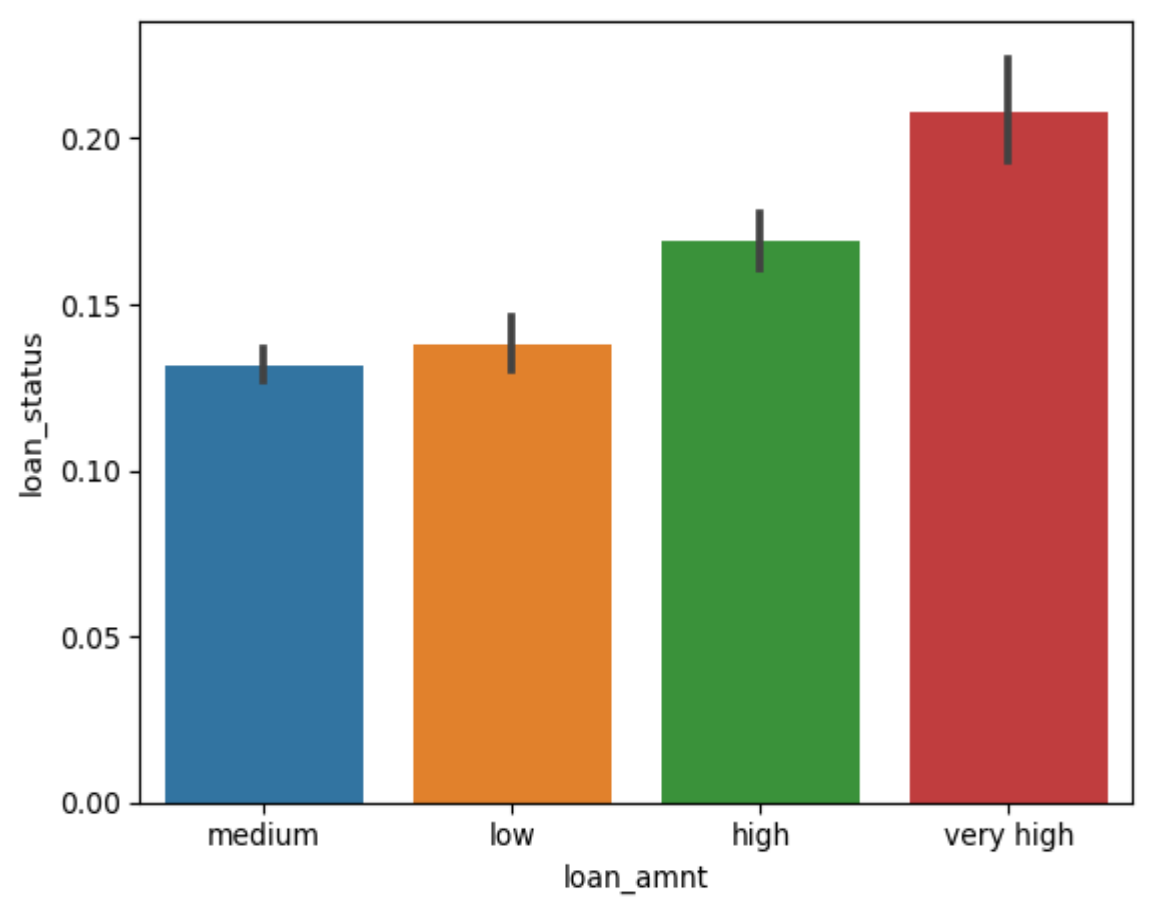
e] Loan amount vs loan status : For this analyzation we have bin the loan amount into high medium and low

Where loan amount < 5000 is low

loan amount between 15000 to 25000 is for medium

loan amount above 25000 is for high level

With this we have analysed that default rate increases with loan amount



f] Interest rate vs loan status :

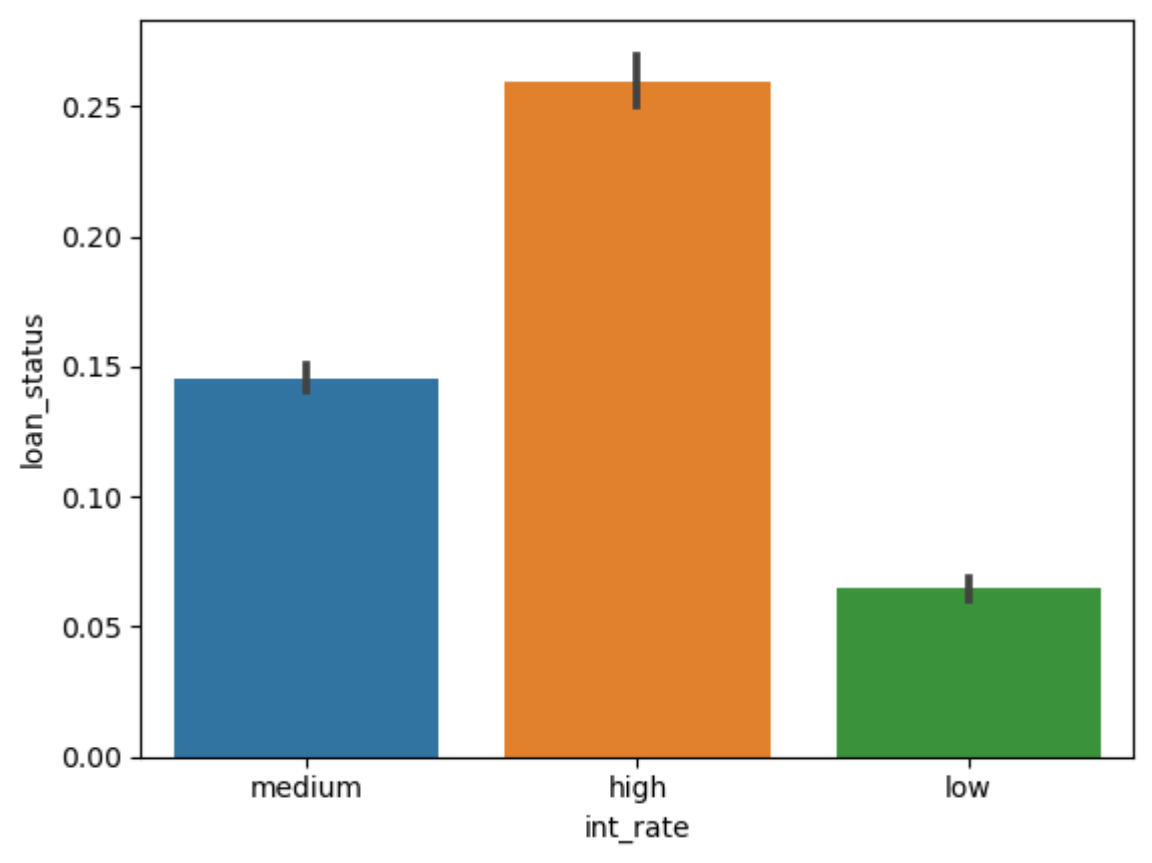
For this analysis we bins interest rate into high, medium and low as follows

i] Interest rate less than 10 is low

ii] Between 10 to 15 is for medium

iii] More than 15 is high

After this got the insight that interest rate is higher for higher default rate



g]Annual income vs loan:-

First we bins the values for our calculation:

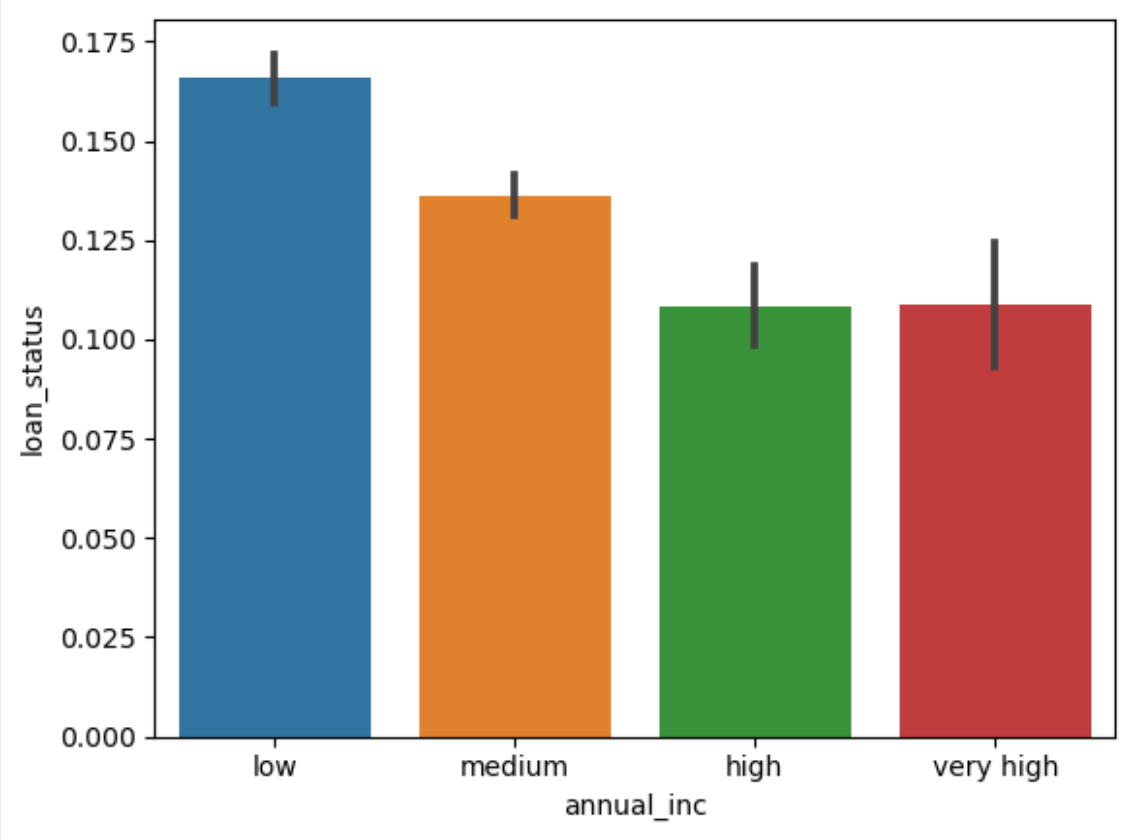
i]Less than 50000 is for low

ii] Between 50000 to 100000 is for medium

iii] Between 100000 to 150000 is for high

iv] More than 150000 is for Very

After analyzation we got the insight that default is high for low income and lower rate for higher income



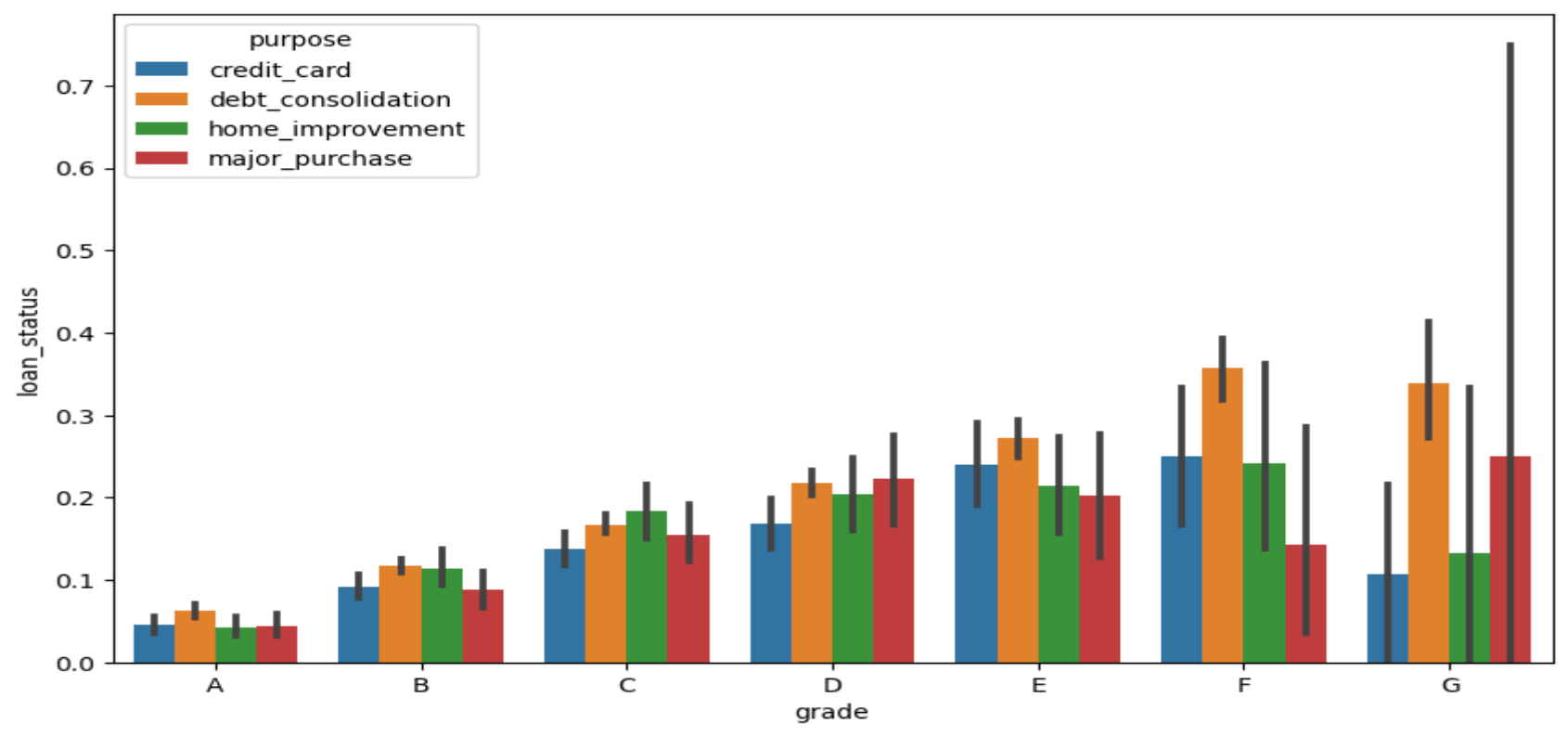
Multivariate Analysis :

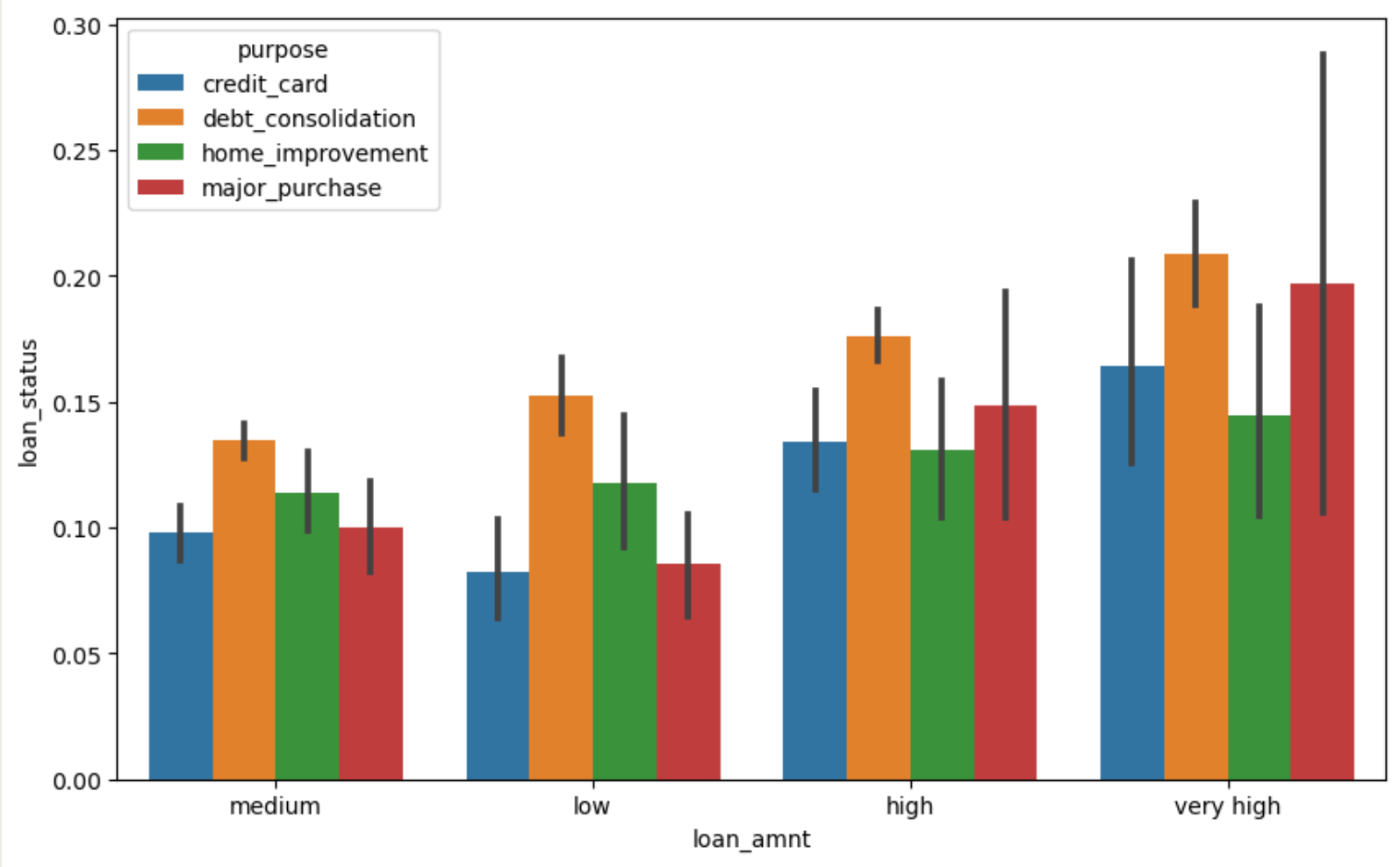
We have now compared the default rates across various variables, and some of the important predictors are purpose of the loan, interest rate, annual income, grade etc.

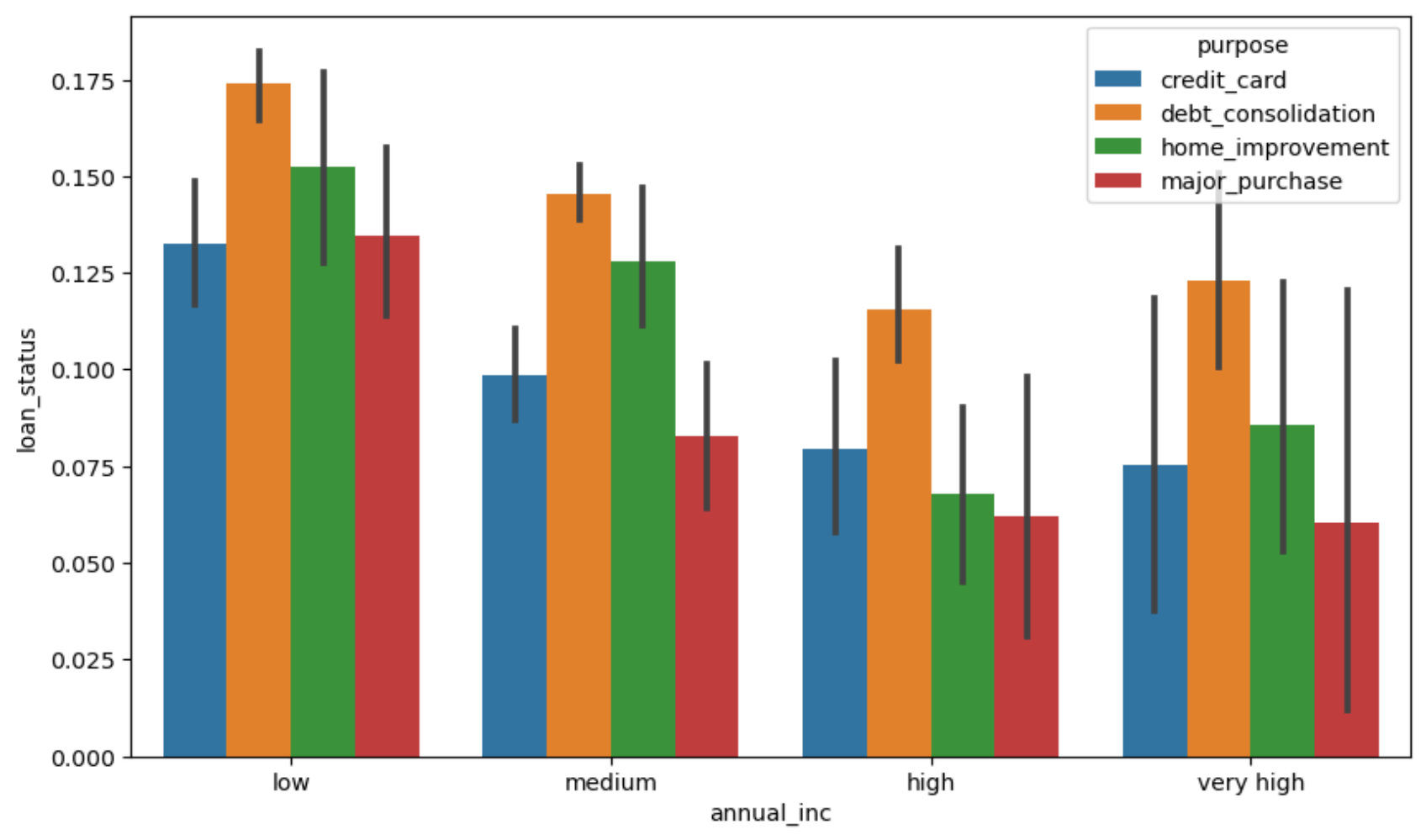
For multivariate analysis we consider the top 4 major purpose applications for analysis which are : debt_consolodation, credit_card,home_improvement , major_purpose

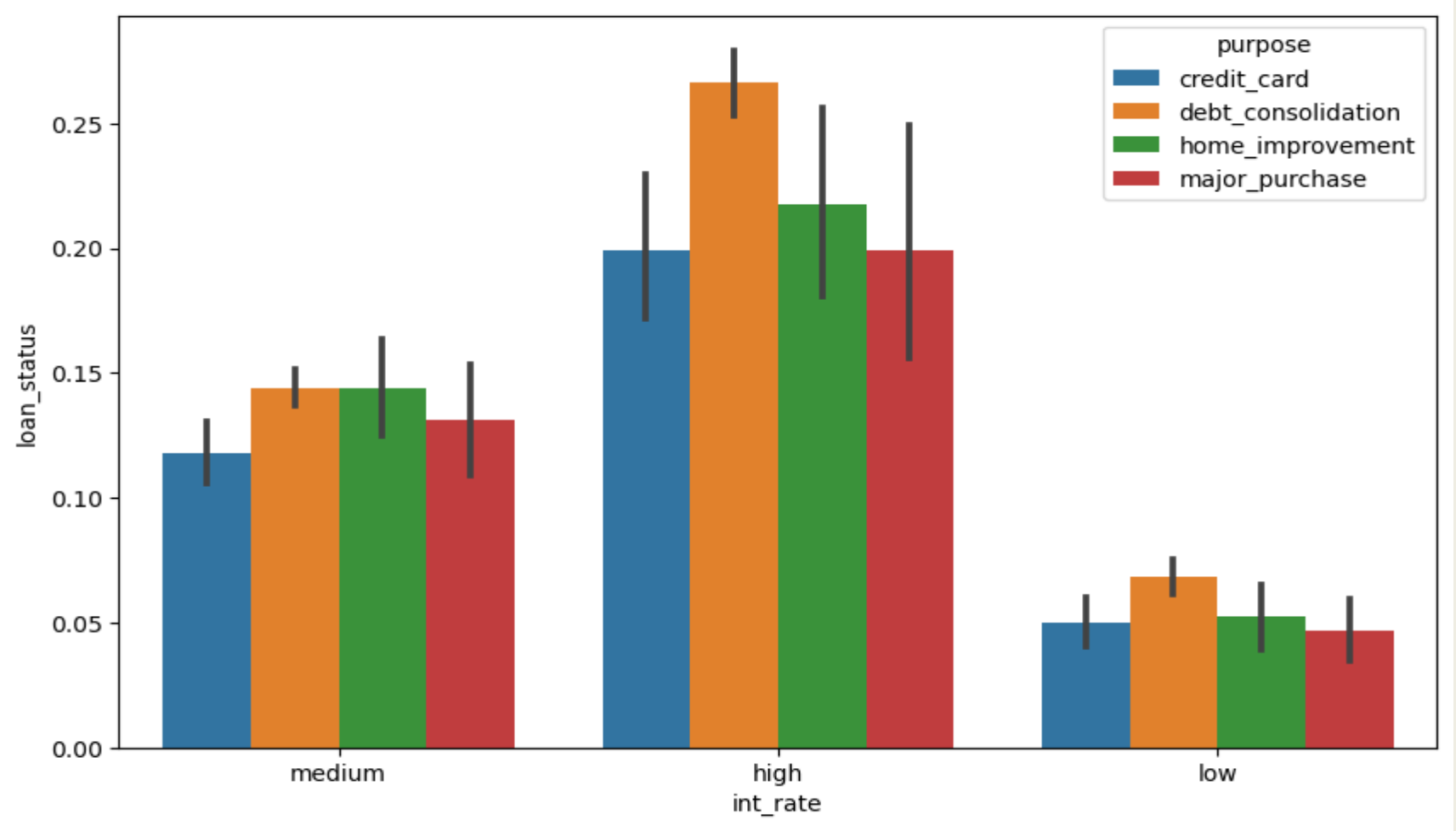
loan status vs purpose vs (grade, loan amount, annual income ,int rate):-

Here got insight that debt consolidation,major purpose have more default rate across each grade




purpose
credit_card
debt_consolidation
home_improvement
major_purchase
0.30
0.25
0.20
0.15
0.10
0.05
0.00
loan_status
medium
low
high
very high
loan_amnt


purpose
credit_card
debt_consolidation
home_improvement
major_purchase
loan_status
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
low
medium
high
very high
annual_inc



In genereal debt consolation and major purpose have high default across Int rate annual inc and loan amount

Key Insights :-

- There is a 6% increase in default rate as you go from high to low annual income
- As the grade of the loan goes from A to G, the default rate increases. This is attributed to the riskiness associated with each grade.
- Individuals with permanent home ownership have a lower default rate compared to those in rental or mortgage situations.
- Default rates tend to increase with higher loan amounts. Loans were categorized into low, medium, and high amounts.
- Debt consolidation and major purposes consistently exhibit higher default rates across different loan attributes, such as interest rate, annual income, and loan amount.